Universität der Bundeswehr München

Optimal Control of Coupled Ordinary and Partial Differential Equations

---

Optimale Steuerung von gekoppelten gewöhnlichen und partiellen Differentialgleichungen

Der Fakultät für Luft- und Raumfahrttechnik

der Universität der Bundeswehr München

zur Erlangung der Lehrbefugnis (*venia legendi*) vorgelegte

Habilitationsschrift

von

Dr. rer. nat. Sven-Joachim Kimmerle

Oktober 2018

Diese Arbeit wurde erstellt am

Institut für Mathematik und Rechneranwendung der
Fakultät für Luft- und Raumfahrttechnik

und am damaligen

Institut für Mathematik und Bauinformatik der
Fakultät für Bau- und Umweltingenieurwesen

an der

Universität der Bundeswehr München.

Anschrift: Sven-Joachim Kimmerle, Werner-Heisenberg-Weg 39, D-85577 Neubiberg
sven-joachim.kimmerle@unibw.de

Datum: eingereicht 4. Oktober 2018, überarbeitet 21. Mai 2019

Zusammenfassung

In einer Vielzahl von Anwendungen trifft man auf mathematische Probleme, in denen sowohl gewöhnliche Differentialgleichungen (ODE) als auch partielle Differentialgleichungen (PDE) auftreten. Oft sind diese Differentialgleichungen vollständig gekoppelt. In der Mathematik der Optimalen Steuerung existieren verschiedenste Resultate zur Optimalen Steuerung von ODE/DAE (Differential-Algebraischen Gleichungen) oder PDE, sofern diese getrennt betrachtet werden können und nicht vollständig gekoppelt sind. Im voll gekoppelten Fall sind nach Kenntnis des Autors bisher nur einige Modellprobleme ausführlich untersucht worden, z.B. ein deformierbarer Satellit [BA86, BA89], das Laserhärten von Stahl [HS97, FHS01, HV03a, HV03b, GNP10], und der Wiedereintritt eines Space Shuttle mit Hyperschallgeschwindigkeit [CPW+09, WRP10, PRWW10, PRWW14, We14]. Beim Problem der Optimalsteuerung von Herz-Defibrillatoren werden PDE lediglich als gekoppelt mit degenerierten PDE (d.h. mit fehlender Ortsableitung) betrachtet [KR15, BK14, BK17].

Auch die Analysis und numerische Simulation von vollgekoppelten Systemen von Differentialgleichungen ist noch nicht vollständig abgeschlossen. Hierzu hatte der Autor in seiner Dissertation [Ki09] bereits voll gekoppelte ODE und PDEs betrachtet, wobei die ODE einen freien Rand modelliert und die PDEs Diffusion und lineare Elastizität beschreiben. Ein Anwendungsbeispiel hierzu ist die Phasenbildung von Galliumarsenid (GaAs)-Tröpfchen [DK10] in GaAs-Wafern, wie sie in der Halbleiter-Herstellung Verwendung finden.

Hauptsächlich wird in dieser Habilitationsschrift die Optimale Steuerung von gekoppelten Differentialgleichungen, einschließlich mathematischer Modellierung, Analysis, numerischen Verfahren und Simulation untersucht. Aufgrund der hochgradigen Komplexität der Problemstellung soll der Fokus vor allem auf eine Vielzahl von neuen Beispielen gelegt werden, die sowohl in Anwendungen wichtig sind, als auch verschiedene Problemklassen repräsentieren. Daher ist diese Arbeit in naheliegender Weise kumulativ gestaltet. Es werden hier exemplarisch betrachtet:

1) Ein Tankfahrzeug,
   dessen Tank über eine Feder-Dämpfung befestigt ist und dessen flüssiger Inhalt durch die Saint-Venant-Gleichungen modelliert wird, während die Fahrzeugdynamik durch die Newtonschen Bewegungsgleichungen bestimmt wird. Bei Bremsvorgängen oder Kurvenfahrten kann es zu einer unerwünschten Interaktion kommen und das Hin- und Herschwappen der Flüssigkeit im Extremfall zum Kippen des Fahrzeugs führen. Die Beschleunigung und im Prinzip der Lenkwinkel des Fahrzeugs sollen so gesteuert werden, dass das Tankfahrzeug schnellstmöglichst eine Strecke bewältigt, ohne dass die Fahrstabilität gefährdet ist.

2) Ein elastisches Kran-Trolley-Last-System,
   in dem die elastische Deformation des Kranauslegers vereinfacht durch die Lineare-Elastizitäts-PDE beschrieben wird. Die Bewegungsgleichungen von Trolley und Last ergeben ein nichtlineares System von ODEs, wobei die Last als zwei- oder dreidimensionales Pendel modelliert wird. Der Transport einer Last mithilfe des Trolleys von einer Anfangs- in eine gewünschte Endposition soll zeit-optimal und sicher gesteuert werden. Ein verwandtes

Problem ist das elastische Brücke-Last-System, bei dem sozusagen der Kranbalken auch an der gegenüberliegenden Seite eingespannt ist.

3) Ein Viertelfahrzeugmodell,
in dem ein linear elastischer Reifen in Straßenkontakt mit einem Feder-Dämpfer-System gekoppelt ist. Durch Steuerung des elektrorheologischen Dämpfers sollen Sicherheit und Komfort des Fahrzeug-Chassis optimiert werden. Hier tritt wiederum die PDE der linearen Elastizität und ein ODE-System für das Feder-Dämpfer-Element auf. Zusätzlich liegt bei diesem Problem eine Komplementaritätsbedingung für den freien Kontaktrand von Reifen und Straße vor.

4) Ein Modell zur Tropfenbildung in Galliumarsenid,
das, wie oben schon angedeutet, die Evolution flüssiger Präzipitate, die bei der abschließenden Wärmebehandlung von GaAs-Kristallen entstehen, beschreibt. In diesem makroskopsichen Modell, ein sogenanntes Mean-Field-Modell, sind die Differentialgleichungen eine hyperbolische Gleichung 1. Ordnung für die Verteilung der Tropfenvolumen und eine ODE für die mittlere Zusammensetzung des Festkörpers, d.h. für das *mean field.* Steuergrößen sind z.B. Anfangszustände oder Temperatur. Hier stellt sich die Frage, ob dieser Prozess geeignet beeinflusst werden kann, um Endprodukte bzgl. gewisser Materialcharakteristika zu optimieren. Man beachte, dass die Zustandsgröße für die Verteilung der Tropfenvolumen ein Maß und keine Funktion im üblichen Sinne ist.

Da wir meist Neuland betreten, steht im Vordergrund, was bei der Optimalsteuerung gekoppelter Differentialgleichungen alles in Anwendungsproblemen auftreten kann, wie man an diese Schwierigkeiten herangeht und numerische Herausforderungen konkret meistert. Eine abschließende Theorie für diese Problemklasse würde den Rahmen dieses Habilitationsvorhabens selbstverständlich sprengen. Der naheliegende Wunsch, ausgehend von diesen Beispielen allgemeingültige generelle Aussagen für die Optimale Steuerung gekoppelter Differentialgleichungen zu treffen, erweist sich schon deswegen als schwierig bis gar nicht möglich, da sich die Beispiele als sehr heterogen erweisen und bereits verschiedene PDEs jeweils eine eigene Theorie und Methoden benötigen. Immerhin kann allgemein gezeigt werden, dass sich bei Betrachtung der adjungierten Probleme bei parabolischen Gleichungen die Kopplungsstruktur umdreht und dass für die Analyse relativ beliebig gekoppelter Zustandsgleichungen sich gut Fixpunktiterationsverfahren anwenden lassen.

Obige Auswahl an gekoppelten Optimalsteuerungs-Problemen wurde bezüglich Modellierung, Theorie (d.h. notwendige Bedingungen 1. Ordnung), numerische Methoden untersucht und insbesondere numerisch simuliert und validiert. In vielen Anwendungen ist im Vorfeld aufgrund der Komplexität eine sorgfältige Reduktion der eigentlichen Problemstellung auf ein mathematisch behandelbares Modell, das aber die wesentlichen Eigenschaften des Problems widerspiegelt, essentiell. Schließlich wurde eine Globalisierungsstrategie für semi-glatte Newtonverfahren entwickelt, die bei zwei der gekoppelten Optimalsteuerungsprobleme erfolgreich eingesetzt werden konnte.

Abstract

In many applications we encounter mathematical problems that exhibit ordinary differential equations (ODE) as well as partial differential equations (PDE). Oftentimes these differential equations are fully coupled. In mathematics of optimal control there exist various results for optimal control of ODE/DAE (differential-algebraic equations), or PDE, as long as they are considered separately and are not fully coupled. To the knowledge of the author, in the fully coupled case only several model problems have been examined thoroughly, e.g. a deformable satellite [BA86, BA89], the laser hardening of steel [HS97, FHS01, HV03a, HV03b, GNP10], and the reentry of a space shuttle with hypersonic speed [CPW+09, WRP10, PRWW10, PRWW14, We14]. For the optimal control of heart defibrillators the PDE are only considered as coupled with degenerated PDE (i.e. with missing spatial derivatives) [KR15, BK14, BK17].

Moreover, the analysis and numerical simulation of fully coupled systems of differential equations has not been completely finished. To this the author had already considered in his PhD thesis [Ki09] fully coupled ODE and PDEs, at which the ODE models a free boundary and the PDEs model diffusion and linear elasticity. An application example for this is the phase formation of gallium arsenide (GaAs) droplets in GaAs wavers that are used in the production of semiconductors.

In this habilitation thesis, we examine mainly the optimal control of coupled differential equations, including mathematical modelling, analysis, numerical methods and simulations. Due to the high complexity of the problem statement the focus shall be set on a multitude of new examples that are important in applications as well as representing various problem classes. Thus this work has been organized cumulatively. Here we consider exemplarily:

1) A tank truck,
   whose tank is fixed by means of a spring-damper element and whose liquid content is modelled by the Saint-Venant equations, whereas the vehicle dynamics are determined by the Newton laws of motion. During the braking operation or driving along curves undesired interactions can occur and sloshing of the fluid from one side to the other may lead, in an extreme case, to a roll over of the vehicle. The acceleration and, in principle, the steering angle of a vehicle should be controlled such that the tank truck travels a certain distance as fast as possible without threatening its driving stability.

2) An elastic crane-trolley-load system,
   at which the elastic deformation of a crane cantilever beam is described in simplification by means of the linear elasticity PDE. The equations of motion of trolley and load yield a nonlinear system of ODEs where the load is modelled as two- or three-dimensional pendulum. The transport of a load by means of a trolley from an initial position to a desired terminal position should be controlled time-optimally and safely. A related problem is the elastic bridge-load-system, at which quasi the crane beam is clamped also at the opposite end.

3) A quarter car model,
at which a linear elastic tyre with road contact is coupled to a tyre-damper system. By control of an electrorheological damper we wish to optimize safety and comfort of the vehicular chassis. Here the PDE of linear elasticity appears again and an ODE system for the spring-damper element. In addition for this problem, we have a complementarity condition for the free road contact between tyre and road.

4) A model for precipitation in gallium arsenide,
that, as indicated above, describes the evolution of liquid precipitates that emerge at a final heat treatment of GaAs crystals. In this macroscopic problem, a so-called mean-field model, the differential equations are a hyperbolic equation of first-order for the distribution of droplet volumes and an ODE for the mean composition of the solid, i.e. for the *mean field.* Control quantities are, e.g., the initial states or the temperature. Here the question arises, whether this process may be manipulated suitably, in order to optimize the final product w.r.t. certain material characteristics. Please note that the state being the distribution of droplet volumes is a measure and no function in the standard sense.

Since we are mainly in uncharted waters, our priority is what may happen in optimal control of coupled differential equations at all in applications, how we approach these difficulties, and how we cope the numerical challenges. A concluding theory for this class of problems is definitely out of the scope of this habilitation project. The obvious wish, to derive general statements for optimal control of coupled differential equations starting from these examples, turns out to be very challenging or even as impossible, since the examples are very heterogeneous and yet different PDEs require different theory and methods. At least we may demonstrate generally that when dealing with adjoint problems for parabolic equations the coupling structure is reverted and that fixed point methods can be applied well for the analysis of pretty arbitrary state equations.

The above selection of coupled optimal control problems has been examined w.r.t. modelling, theory (i.e. necessary first-order conditions), numerical methods, and, in particular, numerically simulated and validated. Due to complexity, in many applications it is essential to perform *a-priori* a careful reduction of the original problem statement to a mathematical treatable model that though reflects the important features of the problem. Finally, we have developed a globalization strategy for semismooth Newton methods that has been applied successfully for two of the coupled control problems.

Contents

Chapter 1

Introduction

Technical systems become more and more complex since recent years. Accordingly, the challenge of complex models and the demand for its accurate simulation and optimization has raised in the last years. Many real-life applications yield mathematical problems that involve ordinary differential equations (ODE) as well as partial differential equations (PDE) that are fully coupled. We will refer to these coupled systems as coupled differential equations, abbreviated by CDE.

Whereas optimal control subject to ordinary differential equations or even differential-algebraic equations (DAE) on one hand and optimal control subject to PDEs on the other is well-established considered alone, the optimal control of CDEs seems to be still in its infancy. Analysis and optimal control of ordinary differential equations is considered as being part of "finite-dimensional" control theory [La03, Ch. 1]. The typical function spaces are Hölder spaces or Sobolev spaces relying on the supremum norm (for details on function spaces see Subsection A.1.3). In case of a PDE as constraint this is considered as part of "infinite-dimensional" control theory. Typical spaces are Sobolev spaces with square-integrable functions and the structure of Hilbert spaces and the concept of Gelfand triples are exploited. Note that in optimization with PDEs mostly linear or semilinear problems are feasible, while in optimal control with ODE fully nonlinear problems are treated in general.

The aim of this work is not a complete and comprehensive treatment of optimal control of coupled differential equations. Due to the complexity described so far this is out of reach here. However, since to the best knowledge of the author only some examples have been examined in this area, we think that this book will provide an important contribution for opening the door to this class of control problems getting more and more important in applications. The focus is not on abstract results for coupled optimal control problems of the considered type. Moreover, we wish to illustrate the importance, complexity and numerical solution of the considered problem class by considering several important applications from modern technologies.

This study is designated for interested readers not only from mathematics, but also from applications, as engineering, computer science, physics, chemistry, economics, life science, and other sciences. We cover modelling issues, like the introduction of an averaging-evaluation operator (see Def. 3.1), algorithmic methods, numerical techniques, and optimal control of CDE. Our approach is to illustrate by considering representative examples, what issues arise in this new class of optimal control problems.

1.1 Coupled Systems of Ordinary and Partial Differential Equations

A differential equation involving only one derivative w.r.t. the same argument is usually referred to as ODE. If in addition, purely algebraic equations are present, we speak of differential-algebraic equations. The orders of the ODE considered here are typically 1 or 2. Note that systems of order n may be rewritten equivalently as first order systems, typically with n times the equations. Note that the contrary is not always possible. We emphasize that we focus on initial value problems involving ODEs here.

In a PDE several partial derivatives may appear. PDEs are commonly classified by the structure of the highest derivatives encoded by a matrix. Depending on the eigenvalues of this matrix, we distinguish

- elliptic equations,
- hyperbolic equations,
- parabolic equations

with different orders (e.g. 1, 2 or 4) that correspond to the order of the highest derivative. In partial differential equations the setting of boundary conditions is crucial as well. Note that a heterogeneous class of problems as partial differential equations requires different theory, depending on type (elliptic, parabolic, hyperbolic) or order, and are treated by different (numerical) methods. However, there is some closer correspondence between elliptic and parabolic PDEs, since a parabolic PDE may be interpreted as a sequence of elliptic problems in the light of a semi-discretization in time. But in general, solutions of parabolic PDEs exhibit a higher regularity as elliptic PDE with the same data and geometry, since the time evolution yields a smoothing effect.

In this study we focus on fully-coupled ordinary and partial differential equations. In particular, the case of coupled systems of ODEs as well as PDEs is included.

The general purpose of optimization with differential equation constraints is to minimize (or maximize) an objective. If the solutions entering the control problem exhibit a structure of being either only solution functions, the so-called states, or functions being controls, then this is called control. If we try to control for a prescribed possibly infinite time horizon without any feedback, this is called optimal control, whereas in (classical) control a system runs for a certain time interval, then depending on observed quantities the control input is chosen that is applied for the next time interval and so on. The control theory of dynamical systems can be grouped into three typical classes depending on the goal:

(i) optimal control,

(ii) controllability, and

(iii) stabilization problems.

In this study, we consider optimal control problems for a finite time horizon and there we focus on optimal control with an open loop (and not on closed loop problems as they appear, e.g., in model predictive control). For control and stability of nonlinear dynamical systems see, e.g. [Ba92, Co07, LT00, Co07], for model predictive control please see [GP17] for instance.

A particular class of optimal control problems are inverse problems. For models arising in inverse problems and the numerical treatment of these typically ill-posed problems see, e.g., [IJ15]. Parameter identification problems, where only a parameter is to be reconstructed, are a simple case of an inverse problem. The numerical approach of choice therefore uses the Tikhonov regularization, being the most powerful and versatile method in this context. We will use the Tikhonov technique in order to obtain projection formulas for the optimal control (see Lemma 2.44) and for our globalization strategy for semismooth Newton (see Section 2.9), where a suitable regularization parameter guarantees descent directions.

Furthermore, we classify by means how the control is exerted to the problem. If the control is applied within the whole of the spatial domain or parts with full measure, e.g., as a right-hand side force term, then this is called a *distributed control*. This a "nice" class yielding bounded control operators [Li68]. However, it may be desirable or only feasible to control the boundary of or points in the domain. This "rough" control is a *boundary control* (on the boundary of the spatial domain or a part of this boundary) or point control (through a Dirac measure or its derivative(!) in the interior of the domain). Usually the control space is larger than the state space and equipped with a weaker topology.

Boundary/point control problems of coupled PDEs are considered in [La03], where the focus is on the structural acoustic problem (coupling a hyperbolic wave equation with an elastic parabolic equation). For some recent results, problems and trends in optimization and control of processes, where optimal control of PDE is applied, see, e.g., [LEG+12]. In [KLST09] optimal control of coupled systems is considered, but the constraints are always partial differential equations. Note that here (coupled) PDEs of different type are considered as well. However, only few results about *combined* ODE-PDE constrained optimal control exist so far, see Sect. 1.2 for details.

As an illustrative introductory example, we choose the optimal control of a moving pendulum.

Example 1.1 *(Introductory Example: Trolley-Load System)*
We consider a load m_L that is fixed to a trolley with mass m_{Tr}. The load may be modelled as rigid pendulum in a 2D plane with length ℓ and it is subject to the gravity acceleration g. The trolley may move (without friction) on an infinite rail and its acceleration force u may be controlled. We consider as states the trolley position q_1 and the angle q_2 between the rigid rope and the perpendicular. At $t = 0$ the initial position $q_{1;0} \in \mathbb{R}$ and initial angle $q_{2;0} \in \mathbb{R}$ are

prescribed and the system is at rest. At the terminal time $t = t_f$ we wish to achieve a given state $[q_{1;f}, q_{2;f}]^\top$. The optimal control problem reads:

Find $u : [0, t_f] \to \mathbb{R}$ such that $\mathcal{J}(u) = \int_0^{t_f} u(t)^2 \, dt$ is minimized,
subject to the constraints, being the ODE of a mathematical pendulum for $[q_1, q_2, v_1, v_2]^\top : [0, t_f] \to \mathbb{R}^4$,

$$\begin{aligned}
\dot{q}_1(t) &= v_1(t), \\
\dot{q}_2(t) &= v_2(t), \\
(m_{Tr} + m_L)\dot{v}_1(t) + m_L \ell \cos(q_2(t))\dot{v}_2(t) &= m_L \ell \sin(q_2(t)) v_2^2(t) + u(t), \\
m_L \cos(q_2(t))\dot{v}_1(t) + m_L \ell \dot{v}_2(t) &= -m_L g \sin(q_2(t)),
\end{aligned}$$

with initial conditions

$$\begin{aligned}
q_1(0) &= q_{1;0}, \\
q_2(0) &= q_{2;0},
\end{aligned}$$

and the terminal constraints

$$\begin{aligned}
q_1(t_f) &= q_{1;f}, \\
q_2(t_f) &= q_{2;f}.
\end{aligned}$$

Note that this problem may as well be considered with variables $\tilde{q} := [q_1 = x_1, x_2, q_2]^\top$ being the coordinates of the load and the angle. This yields an optimal control problem subject to a DAE (with differentiation index 3) where the further constraint $x_1^2 + x_2^2 = \ell^2$ has to be respected, cf. [Ge12, Ex. 1.1.10].

So far this is an optimal control problem subject to an ordinary differential equation. The optimal control of a trolley-load system has been considered in [CG11, CG12]. Including the elastic deformations of the crane beam, described by the linear elasticity PDE, into this model yields in a natural way a fully coupled optimal control problem with ODE-PDE constraints that is discussed here in Section 4.3 in detail.

1.2 Applications with Coupled Ordinary and Partial Differential Equations

In mathematical modelling there exist different approaches. One can start with kinematic models, relying on geometric properties. An alternative are dynamic models, considering forces. The latter could be extended by considering, e.g., momentum balances or conservation laws. If, in addition, we choose proportionality constants in such a way that the first and second law of thermodynamics are fulfilled, this is called a model from first principles. If feasible, we will follow the latter approach.

Coupled systems involving ODE and PDE can be found in many applications in engineering and natural sciences. Recent applications involve, for instance, tracking a bus trajectory (described by an ODE) in a traffic flow (modelled by a scalar hyperbolic conservation law) [MG14], a reaction-diffusion model for tumor invasion where the tumors produce acid that attacks healthy cells [TT16], or the mitochondrial swelling, where the ODE models the evolution of the mitochondrial sub-population and the PDE the spatial calcium propagation [EEO+15]. A further example, among others, is a flexible cantilever structure actuated by piezoelectric macro-fiber composite patches [STK11].

Within the concept of a *digital twin* and the co-simulation of engineering systems the numerical simulation of these systems, in particular complex coupled systems, has found increased interest, see e.g. [HHW18]. Another context where ODE-PDE problems appear are continuous analogues of iterative methods for PDE-constrained optimization problems [BFHK18].

For coupled systems we mention the well-posedness result from the PhD thesis of Kimmerle [Ki09], based on Niethammer [Ni99], that is stated in an abstract form here in Th. 3.4. For further coupled systems that have been considered in optimal control see the next subsection. How the coupled systems considered by the author differ w.r.t. the type of differential equations, the means of coupling (e.g. by boundary conditions (b.c.) or right-hand side (r.h.s.)), the type of coupling (one-sided/full) etc. is shown in Table 1.1.

1.3 Optimal Control Problems with Coupled Ordinary and Partial Differential Equations

To the authors' knowledge, not many optimal control problems for fully coupled ODE-PDE systems have been considered thoroughly. We list the few examples that have been treated so far.

A model for a satellite, whose dynamics are given by a ODE and its flexibility is modelled by the PDE for a beam, has been controlled optimally by Biswas and Ahmed [BA86, BA89]. Furthermore, the optimal control of a gantry crane, where the load is described by a ODE and the deformation by a PDE, has been considered in [Bi04]. Similar evolution problems for flexible structures have been considered from a control theoretic point of view by Zuyev [Zu15]. His focus is on partial stability and controllability of infinite-dimensional mechanical systems, like rotating flexible structures and flexible-link manipulators. The mathematical techniques rely on the Lyapunov function, an approach that we do not pursue in this study, since we confine ourselves mainly to Lagrange function techniques and shortly to the minimum principle.

Chudej *et al.* [CPW+09] consider optimal control for coupling of the heat equation and equations of motions. Similar as in the elastic-crane-trolley-load problem, the coupling from ODE to PDE is achieved by means of a boundary condition, but the controls arise in the ODE system only and the PDE is considered in one space dimension.

In our crane-trolley problem [KGH18a, KGH18b], the controls arise in the ODE as well as

Application	Reference	Type of ODE/DAE	Type of PDE	Means of coupling	Type of coupling
1) Truck-container	[GK15], [KG16], [WGKG18]	Vehicle dynamics	St-Venant equations (1D)	In PDE on Neumann b.c. for h & on r.h.s. for v, in ODE on r.h.s.	Full
2a) Elastic crane-trolley-load	[KGH18a], [KGH18b]	Vehicle/ pendulum dynamics	Linear elasticity (elliptic) (3D)	In PDE on Neumann b.c., in ODE by coefficients & r.h.s. (averaged displacements)	Full
2b) Elastic bridge-load	[Ki16]	Vehicle dynamics	Linear elasticity (elliptic) (2D/3D)	In PDE on Neumann b.c.	One-sided (ODE → PDE)
3) Quarter car model	[KM14]	Spring-damper dynamics	Linear elasticity (elliptic) (2D) & complementarity cond. (Hertz approximation)	In PDE on Dirichlet b.c.	One-sided (Free bdy. → PDE → ODE)
4) Precipitates	[Ki09, DK10], [Ki12, Ki15]	Radii evolution (integro-diff. eq. for measure)	Laplace (3D) (with linear elasticity), finally replaced by ODE	(Time dep. domains)	Full
Nanobubbles	[SKB17, KSB17]	Free boundary (radii evolution)	Advection-diffusion (1D)	Time dep. domains	Full
Nanochannels in PEM	[LBKN11, KBN13], [BKN14, KLNB14]	Free boundary (as DAE not ODE)	PNP-Stokes & linear elasticity (2D/3D)	Shape optimization	Full

Table 1.1:
Overview of the considered examples for coupled ODE-PDE problems and their different features. The last two examples are not discussed in this study in details.
Here PEM abbreviates “polymer electrolyte membrane”, that is typically used for fuel cells. PNP stands for “Poisson-Nernst-Planck”, the PNP system comprises the Poisson equation from electrostatics with the Nernst-Planck equation, being the balance law for charges that may move.

in the boundary condition. Our control acts on the PDE by means of a Neumann boundary control. In our problem we encounter control constraints, too. For a particular class of ODE-PDE-problem, a so-called hypersonic rocket car problem, including the problem of [CPW+09], some new phenomena have been discovered by Pesch *et al.* [PRWW10]. In [WRP10] this OCP is reformulated as a state constrained optimal control problem for PDE and necessary optimality conditions for it are derived. In particular the situations, when state or control hit their constraint, are crucial.

The hypersonic rocket car problem as well as another coupled problem related to molten carbonate fuel cells can be viewed as partial differential algebraic equations (PDAE). Thus the related optimal control problems can be considered in the context of optimal control of PDAE [Ru12].

A model of coupled ODE-PDE systems in the form of monodomain equations arising in the context of heart electrophysiology has been considered by Breiten, Kunisch, and Rund [KR15, BK14, BK17]. Undesired effects like arrhythmia may be modelled as spiral waves and re-entry phenomena. The aim is the stabilization of the system [BK14] around a desired state being free of arrhythmia, similar to an external intervention by a defibrillator. Here a new mathematical phenomenon appears, a finite accumulation point in the spectrum of the corresponding operator, making the null-controllability of coupled ODE-PDE system impossible [BK17]. However, this particular system may be stabilized locally by means of an algebraic Riccati equation for the operator, relying only on the linearized PDE (and not on the ODE at all).

A model for laser hardening of steel, where the phase of steel (i.e. the volume fraction of austenite) is encoded by an ODE and the PDE is a semilinear heat equation, has been considered by Hömberg and Volkwein [HV03a, HV03b] among others. The latent heat required for the phase transition and depending on the austenite phase is a volume term in the heat equation, whereas the temperature enters as a coefficient into the ODE, thus the problem is fully coupled. Note that the phase fraction depends on time and space, correspondingly the ODE is examined in the whole parabolic cylinder. The goal is to track a desired phase fraction and the whole phase transition is controlled by the laser energy that should be kept minimal. For details on the model, see, e.g., [HS97, FHS01]. For a 3D model in [HV03b] necessary optimality conditions are derived that are solved using POD (proper orthogonal decomposition) methods. In the earlier article [HS97] a more enhanced model with five ODEs for different phases of steel is considered and for pointwise state-constraints on the temperature the necessary optimality conditions are derived. The studies [HS97, HV03a, HV03b] use a regularized Heaviside function for the part of the domain where the phase transition takes place. In [GNP10] it is shown that the regularization converges, but they consider only a 2D problem.

In [AKM14] the optimal control of a coupled ODE-PDE problem for supply chains is considered. Here the PDE is for describing the density of the processed parts, whereas the ODE models the queue buffer occupancy. The aim is to minimize the queues and tracking a desired outflow by means of a pointwise control subject to state constraints.

The optimal control of one-dimensional hyperbolic conservation laws coupled to ODEs in time is considered in [BCG10]. This abstract optimal control problem is similar to our truck-container problem, but they allow only for a coupling between the boundary condition of the PDE and the ODE states.

How the optimal control problems for coupled differential equations in this study differ w.r.t. the type of control, objective, constraints, and numerical optimal control methods is denoted in Table 1.2.

1.4 Methods and Results

The numerical methods for simulation and, in particular, for optimal control are crucial for accurate computations. For the ODE suitable Runge-Kutta methods, explicit, semi-explicit or implicit may be considered. In several of our examples, it turns out that the second-order Heun method is helpful and, if we deal with oscillations as in Problem 2, then semi-explicit methods are appropriate, whereas implicit methods yield computational issues. For PDE we work with finite element methods or finite volume methods (inclusive finite difference methods). Similar as for ODE, here typically higher-order methods are advisable, e.g., in Problems 2 and 4 at least quadratic elements are required in order to achieve a certain precision for derivatives in boundary terms that enter into the coupling. Furthermore, when free boundaries are present as in [Ki09, SKB17], special methods have to be applied as transformation techniques and fixed point iterations [Ki09, Ki11, KSB17]. The stability of the scheme may be an issue. For example, for the truck-container problem [GK15, KG16, WGKG18] we end up with factor of up to 30 by that the time grid has to be finer than the space grid. This is due to Courant-Friedrichs-Lewy condition [La06, Sect. 8.2 & 8.3] known for explicit solvers for hyperbolic first-order problems. For instance, higher-order numerical methods for hyperbolic conservation laws coupled with ODEs, covering networks as well, are constructed in [BK16].

Of course, the mentioned fixed point iteration technique, described in details in Section 3.2, may be exploited for basic algorithms as well. But there exist other approaches for coupled problems, too.

If in simulations of coupled systems the sub-systems are simulated apart and only some information is exchanged at fixed time steps, then this method is called a *co-simulation*, see e.g. [AG01]. In some approaches for optimal control problems the Jacobians are transferred between the part systems. This is related to the area of distributed systems that are considered in computer science.

Another possibility could be the freezing of coefficients, i.e. for dependent coefficients the values from the last step enter, whereas other terms are evaluated at the current step. However, within the control problems that are considered so far in literature the subsystems are considered mostly as black-box problems and the particular structure of the coupling is not exploited. This is also one of the goals of the following habilitation thesis. For example, in [WGKG18]

Application	Reference	Type of objective	Type of ctrl.	Constraints	Method	Approach
1) Truck-container	[GK15]	Time-opt. ctrl.	ODE r.h.s.	Ctrl.(/state)	FDTO	Adjoint-based; reduced (shooting)
	[KG16]	Fixed terminal time		Ctrl.	FOTD	Adjoint-based; all-at-once (SSNM)
	[WGKG18]	Time-opt. ctrl.		Ctrl. & state	FDTO	Full discretization all-at-once (SQP)
2a) Elastic crane-trolley-load	[KGH18a] [KGH18b]	a) Time-opt. ctrl. w/o ctrl. cost b) Fixed terminal time Time-opt. ctrl.	ODE r.h.s. Neumann bdy. ctrl.	Ctrl.	FDTO	Sensitivity-based; reduced
2b) Elastic bridge-load	[Ki16]	Fixed terminal time (w/o ctrl. cost)	Neumann bdy. ctrl.	Ctrl.	(FDTO)	“Try-out”
3) Quarter car model	[KM14]	Fixed terminal time	ODE coefficient	Ctrl.	FDTO	Sensitivity-based; reduced
4) Precipitates	[Ki12]	(Fixed terminal time)	Coefficient & initial parameter	Ctrl. & state	FDTO	Sensitivity-based; reduced
Nanochannels in PEM	[KBN13] [BKN14]	Fixed terminal time	Shape optimization	“Ctrl.” (i.e. on shape)	FDTO	Sensitivity-based (shape derivatives); “reduced”

Table 1.2:

Overview of optimal control problems for coupled ODE-PDE and their different features. The last of my examples is not discussed in this study in details.

By “try-out” we man that certain obvious controls are tested. For the abbreviations see the text and Appendix D.

we compute the Jacobians and the Hessians explicitly and the structure of the problem is exploited in the algorithms. By applying a direct solver for sparse matrices the computational precision and efficiency may be increased, however, for large problem sizes an indirect solution method, like Least-Memory-BFGS (Broyden-Fletcher-Goldfarb-Shanno) [LN89] may still be required. We discuss first-discretize-then-optimize approaches vs. first-optimize-then-discretize approaches and Lagrange vs. Pontryagin approaches.

We recall that we do not claim that this study is to some extent exhaustive w.r.t. coupled optimal control problems. Moreover, we would like to show its importance, variety and challenges by different examples. In general, due to the full coupling the number of unknowns to be treated simultaneously are much larger as in optimization with PDE. This illustrates that the simulation and numerical optimal control alone for the examples presented here is non-standard.

As in [PRWW10] we observe in [KGH18a] for optimal control without Tikhonov regularization that the bang-bang principle (for details see Def. 2.45) seems to be violated. However, so far this is only a conjecture and it cannot be excluded that the effect is merely a numerical artifact that would, however, still illustrate the importance of suitable accurate algorithms.

Furthermore, our results comprise the introduction of an evaluation operator, averaging over spatially dependent PDE solutions that enter into the ODE; the reversal of the coupling structure in the adjoint problem; and that it seems to be advisable in FOTD approaches (see Sect. 2.3) to treat ODE as simple PDE and then to apply the theory of optimal control for coupled PDE systems.

1.5 Outline

The outline of this habilitation thesis is due to our approach considering typical examples. We include accepted articles and proceedings, namely [GHK17] in Chapter 2 and [KG16, KGH18a, Ki16, KM14, Ki12] in Chapter 4. The first article is on a globalization strategy for semismooth Newton methods (SSNM) for optimal control of PDE, but it has been developed for optimal control of CDE and applied successfully in this context. The other included papers discuss various real-world examples (see the list in the abstract and at the begin of Chapter 4) for optimal control problems with coupled systems. The selection of the papers has been made in the light to present typical problems with different features and only one article for each problem type for brevity. The content of the proceedings [Ki18] has been included in Sect. 3.3, particularly in Subsect. 3.3.4.

For instance not included are, for the truck-container problem the articles [GK15, WGKG18] and for the second problem the paper [KGH18b] is not presented here. A prerequisite to the last problem, Problem 4, are [Ki09, Ki11, DK10] that have been written within the PhD studies of the author. They deal with modelling and analytical well-posedness. Also related to Problem 4 are the later publications [Ki15] or [SKB17, KSB17] that deal with the evolution or stability, resp., of gaseous precipitates. Here the precipitates are hydrogen nanobubbles, but the focus is on modelling and simulation, yet. In the last paper a coupled ODE-PDE system is derived from a mean field model for interacting bulk and surface nanobubbles, not yet examined as optimal control problem, being similar to the problem considered in [Ki12].

Another type of coupled control problems involving modelling, simulation and shape optimization, in particular, have been treated in [LBKN11, KBN13, BNK11, KLNB14]. The application is a low temperature polymer electrolyte membrane (PEM) for hydrogen fuel cells.

A topic of ongoing research in cooperation with a large automobile manufacturer is the transport of charge carriers (protons and, e.g., platinum ions) in deformable nanochannels fully filled with liquid. This models polymer electrolyte membrane (PEM) under operation and yields a problem of PDE (Poisson equation for the electric field, Nernst-Planck equation for diffusion and transport, Stokes equation for laminar flow) and DAE for the free boundary of the nanochannel. The DAE and PDE are fully coupled since by means of the moving domain the ODE back couples to the PDE. The aim is here the numerical simulation of this complex problem by finite elements and finite volumes and the examination of different geometries as in [BKN14]. A parameter for the shape optimization is given, e.g., by the overall pressure.

In the next chapter we recall the classical theory for optimization and optimal control. In particular we present analytic results and algorithms in function space. We begin with standard results for optimization in Banach spaces in Section 2.2. Starting from Section 2.4 the focus is on the special case of optimal control. Chapter 2 might be skipped by experts and readers familiar with these topics, since we present the theory of coupled optimal control problems in the next chapter. Only in the last section, Sect. 2.9, we present a new result for a globalized algorithm for the semismooth Newton method in Section 2.9.

In Chapter 3, in which we consider the optimal control theory of coupled ODE-PDE , contains our main theoretic contributions to this field. The main result is that either ODEs may be treated as PDEs in a Hilbert space setting that is different to classical ODE theory or the PDE may be considered as an ODE in function spaces, i.e. as an abstract evolution equation. Under reasonable conditions both approaches coincide, however we recommend the “treat ODE as PDE” approach in practice. Chapter 3 provides the theoretical fundament for the following applications.

Chapter 4 summarizes several new examples in engineering and science for this class of coupled optimal control problems that we have solved numerically in particular. For suitable numerical approaches the results from the last two chapters are crucial. In particular we mention our globalized semismooth Newton method, fixed point iterations, and the reverse coupling structure in adjoints.

We close with a discussion in Chapter 5. Further basic results from analysis and functional analysis are recalled in Appendix A for convenience. Appendix B collects some results on numerical methods for PDEs. In Appendix C the connection between Lagrange and Hamilton mechanics and optimization is presented and the last appendix, App. D gives an overview of the used notation and symbols.

Chapter 2

Classical Theory and Methods for Optimization and Optimal Control

In this chapter we revisit the standard theory for optimization and optimal control that is required in Chapter 3, where we adapt this theory for a new framework, being the special case of optimal control with fully coupled ordinary and partial differential equations. We focus mainly on analytic aspects and algorithms in function space, concerning infinite-dimensional problems, and leave partially numerical aspects and the finite-dimensional case aside, since both the latter are covered by a wide variety of books and lecture notes, for instance, [GK99, GK02, LY08, Be10]. After a few general considerations, we start with optimization in Banach spaces (Section 2.2), followed by general optimization algorithms (Section 2.3), and then we restrict ourselves to the important special case of optimal control (Sections 2.4 and 2.5). Then we focus on optimal control of differential algebraic equations (Section 2.6) and on optimal control of partial differential equations (Section 2.7). In the last section we include our original article on a globalization strategy for the semismooth Newton method.

2.1 Optimization, Optimal Control and Control

The general purpose of mathematical optimization is to optimize a certain quantity, called the objective, subject to a model or to drive a system to a certain behaviour, where we assume that the model or the system may be described mathematically. The mathematical model system might be described by certain equations. In this chapter we consider only the situation of ordinary and/or partial differential equations modelling a certain process. The model might be from natural sciences, engineering, economics, or other disciplines, however, the mathematical problem has a similar structure.

In optimal control the underlying model exhibits state variables (or just states), denoted by y, and control variables (or controls), denoted by u. The idea is to choose the optimal control u such that system, with solutions y of the differential equations, behaves optimally in a sense prescribed by the objective.

We consider a real system

$$y = \mathbb{S}(u)$$

that is represented by the model

$$y = \mathcal{S}(u).$$

Clearly, our goal is that $\mathbb{S}$ is in a (here deliberately undefined) sense close to $\mathcal{S}$, but we cannot expect that $\mathbb{S} = \mathcal{S}$ is achieved. We might even encounter the situation of a black-box problem with known in- and output, but without further knowledge of the mechanism $\mathbb{S}$ within.

Furthermore, we have to distinguish between optimal control[1] and control. In control[2], as understood here, systems with a feed-forward control and feedback are considered. In open loop control, the action of the controller is independent of the actual process. This is linked to optimal control, where the optimal control is computed in advance corresponding to optimal solutions, but no feedback is considered. Thus, when the computed optimal control is applied to a real system, it is has to be combined with some feedback control, unless it is known that some stability can be guaranteed. This motivates to consider feedback control as well. In a closed loop approach for control, some output is measured and depending on it, some new input, i.e. a feedback, is computed and applied. However, in many real-world applications the feedback is subject to some significant delay. Within model predictive control (MPC)[3] a time-discretized dynamic model of the process to be controlled is employed in order to predict the behaviour of the process in the near future, i.e. on a finite time horizon. This allows for the computation of an optimal input subject to possible state and/or control constraints.

In this study we consider optimization problems with an open loop (and with an finite time horizon) only. In general, the sought-after optimization variable is denoted by z, i.e. $z = [y, u]$ in the case of optimal control, where y represents a state and u an optimal control.

In this study w.l.o.g. minimization is considered. Then the general task is to minimize an objective

$$\Phi(t_0, t_f, z(t_0), z(t_f)) + \int_{t_0}^{t_f} \phi(t, z(t))\, dt \tag{2.1}$$

with respect to z subject to differential equations complemented with initial and boundary conditions and control and/or state constraints for z. Here t_0 is the initial time, w.l.o.g. we consider mostly $t_0 = 0$, and t_f is a possibly free terminal time.

Φ and ϕ are given sufficiently smooth functions. The first summand, Φ, is called *Mayer term*, the second term is the *Lagrange term*. A combination of both terms is a so-called *Bolza problem*. In principle, a Lagrange term may be expressed as another Mayer term by introducing another state together with an ODE. However, this transformation is not always the best choice, when considering optimization problems.

[1] “Optimalsteuerung” in German.

[2] “Regelung” as well as “Steuerung” (i.e. without feedback) in German.

[3] Also called receding horizon control.

Important tasks in optimization are:

- Existence of optimal solutions,
- Uniqueness of optimal solutions,
- Optimality conditions, and
- Optimization algorithms.

As optimality conditions, in particular the Karush-Kuhn-Tucker (KKT) conditions are important. Most modern optimization algorithms are based on KKT.

Our presentation of the theory for optimization and optimal control follows in particular Gerdts [Ge12] w.r.t. ODE and the books of Tröltzsch [Tr10] and Hinze, Pinnau, Ulbrich, and Ulbrich [HPUU09] w.r.t. PDE. The topics and the extent of the book [HPUU09] are well-selected. Parts of this chapter follow this book, however we present and arrange the results in a different structure.

The outline of this chapter is as follows. At first, in Section 2.2 we consider general optimization problems with some basic structure, then we focus on the case of optimal control, see Section 2.4. Optimality conditions are discussed in Subsections 2.2.2, 2.2.3 and 2.4.2, respectively. Optimization methods are presented in Sections 2.3 for general optimization problems and in 2.5 for optimal control in particular.

Of course, our presentation of optimization theory is far from being complete. Inverse problems may be considered within the class of optimal control problems. Inverse problems have the undesirable feature that they are usually ill-posed. Parameter identification problems are a special case of optimal control or inverse problems, resp. The unknown parameter to be identified is constant in time, it may be treated by introducing another artificial state fulfilling a trivial ODE, then the parameter is determined as a free initial condition. We have a short look at time-optimal control problems that can be considered as parameter identification problems.

Time-Optimal Control and Transformation Techniques

The technique of choice for dealing with problems involving a free terminal time $t_f > 0$, is to consider a time transformation. By using a linear time transformation

$$t(\tau) := t_0 + \tau(t_f - t_0), \quad \tau \in [0, 1],$$

and scaling states and controls, an equivalent problem with scaled time $\tau \in [0, 1]$ is obtained, where the terminal time t_f enters as an unknown parameter. By this means the problem is more suitable for analysis, simulation and optimization. The parameter t_f is incorporated into the objective, commonly as a linear term.

There are many works on time-optimal control of differential equations. For example, the time-optimal control of the heat equation is considered by Kunisch and Wang [KW13]. We consider time-optimal control of coupled problems in Sections 4.2 and 4.3.

Similarly non-autonomous problems may be transformed into autonomous problems, see, e.g., [Ge12, Subsect. 1.2.2]. This is achieved by introduction of an additional state $\tilde{\tau}$ subject to the differential equation

$$\dot{\tilde{\tau}}(t) = 1, \quad \tilde{\tau}(t_0) = t_0$$

yielding that Φ and ϕ in 2.1 do not depend explicitly on t, but a further state has been introduced.

In case of a time-dependent domain, e.g. by means of a free boundary, a standard approach is to transform to a fixed domain, typically the initial domain (see, e.g. [Ki09, Ki11]). For further transformation techniques (transformation of Tschebyscheff problems, L^1-problems, and interior point constraints) yielding a standard problem see [Ge12, Sect. 1.2]. Thus we consider w.l.o.g. the theory of a non-autonomous problem with fixed terminal time, as Problem 2.1.

2.2 Optimization in Banach Spaces

2.2.1 Generic Optimization Problems and Preliminaries

In this subsection our presentation follows mainly [HPUU09, Ch. 2]. We start with the most general form of optimization problem, considered here in this text.

Problem 2.1 *(Optimization Problem in General Form)*
Let Z be a vector space over $\mathbb{K}$. Let $\mathcal{J} : Z \to \mathbb{K}$ be a functional, the so-called objective (objective function or functional). The optimization problem reads:

Find $z \in Z$ such that $\mathcal{J}(z)$ is minimized,
where $z \in Z_{ad}$,
with $\emptyset \neq Z_{ad} \subset Z$.

If $Z = Z_{ad}$ the problem is called unconstrained.

In this study we consider only the case $\mathbb{K} = \mathbb{R}$. Note that w.l.o.g. we consider only minimization problems, since maximization problems can be transformed into equivalent minimization problems.

Definition 2.2 *(Optimality)*
Consider Problem 2.1 for the case that Z is a Banach space.

a) We call $\hat{z} \in Z_{ad}$ locally optimal (minimal) or local optimum (minimum/minimizer), if

$$\mathcal{J}(\hat{z}) \leq \mathcal{J}(z) \quad \forall z \in Z_{ad} \cap V(\hat{z})$$

in some (open) neighbourhood $V(\hat{z})$ of $\hat{z}$.

b) *If the latter holds with strict inequality for $z \neq \hat{z}$, we say $\hat{z}$ is a strict local optimum (minimum).*

c) *If there exists a $V(\hat{z})$ s.t. $Z_{ad} \cap V(\hat{z}) = Z_{ad}$, then we have a global optimum (minimum).*

Assume for the moment that we may split $z = [y, u] \in Z = Y \times U$ as in optimal control, where y is a state and u a control. Note that in a) then we consider a weak local optimum, since we allow for a neighbourhood $V(\hat{y}, \hat{u}) = \{[y, u] \in Z_{ad} \,|\, \|[y, u] - [\hat{y}, \hat{u}]\|_Z < \varepsilon\}$ for some $\varepsilon > 0$. However, a strong optimum means that $\mathcal{J}(\hat{y}, \hat{u}) \leq \mathcal{J}(y, u)$ for all $[y, u] \in Z_{ad}$ such that $\|y - \hat{y}\|_Y < \varepsilon$ (see, e.g. [Ge12, Def. 7.1.3]). Strong optima are weak optima, but the converse is false in general.

In convex optimization problems (i.e. when $\mathcal{J}$ and Z_{ad} are convex), every local optimum is a global optimum. For example, semilinear PDE lead to nonconvex optimization problems, unless the PDE is actually linear. For some examples of nonlinear PDE, we refer to our papers [GHK17] (two semilinear problems) and [KG16] (the Saint Venant equations are yet quasilinear) included here in Sect. 2.9 and Sect. 4.2, respectively.

Lower semi-continuity of $\mathcal{J}$ and compactness of Z_{ad} guarantee the existence of a solution of Problem 2.1 due to the Weierstrass theorem. However, the latter two prerequisites are in general difficult to verify [PT12].

Problem 2.3 *(Standard Optimization Problem (With a Certain Structure of the Constraints)) Let Z, W_G, and W_H be vector spaces (over $\mathbb{R}$) and let $\mathcal{J} : Z \to \mathbb{R}$ be a functional. Furthermore, let $G : Z \to W_G$ and $H \to W_H$ be operators.*

Find $z \in Z$ such that $\mathcal{J}(z)$ is minimized,
where $z \in Z_{ad}$,
subject to the constraints

$$G(z) \in \mathcal{K},$$
$$H(z) = 0_{W_H},$$

where $\emptyset \neq Z_{ad} \subset Z$ is a closed convex set and $\mathcal{K} \subset W_G$ is a closed convex cone (see Def. A.31) with vertex at 0_{W_G}.

Note that we write 0_{W_H} for the zero element of the vector space W_H in this study in Chapters 2, 3, and in the appendices. This also implies that the equation $H(z) = 0_{W_H}$ holds in the sense of W_H. For instance, if Z is a Lebesgue space like L^2, then the equation holds up to sets of L^2-measure zero. We write $0_{\mathbb{R}} = 0$.

Please note that the latter problem is formulated generally, including the infinite-dimensional case. If, e.g., $Z = \mathbb{R}^{n_z}$, $W_G = \mathbb{R}^{n_G}$, and $W_H = \mathbb{R}^{n_H}$, we recover a finite-dimensional problem, where $G(z) \leq 0_{\mathbb{R}^{n_G}}$. Any standard nonlinear program (NLP) in finite-dimensional Euclidean spaces can be treated as a special case of the latter situation.

Remark 2.4 *(Admissible Set; Feasible Set)*
$Z_{ad} \subset Z$ is called here the admissible set. Here G can be interpreted as inequality constraint and H is an equality constraint. Note that there is a certain choice, which constraints are put into Z_{ad} and which are formulated by means of G and H. This choice can be made such that it is most suitable for the analysis, for the numerical solution approach or for the application. The combination of Z_{ad} with both the constraints yields the feasible set

$$\Sigma_{(fs)} := \{z \in Z \,|\, G(z) \in \mathcal{K},\, H(z) = 0,\, z \in Z_{ad}\}. \tag{2.2}$$

Note that we make a distinction between the admissible and the feasible set as e.g. in [HPUU09, p. 53], contrary to [Ge12, Sect. 2.3]. In the admissible set not necessarily all constraints are incorporated, whereas the feasible set includes all constraints of the problem. The different notions are motivated, e.g., by the Def. A.38 of the tangent cone that requires $\Sigma_{(fs)}$ and Z_{ad} at the same time.

An approach for numerical methods is to approximate locally Pb. 2.5 that is nonlinear by

Problem 2.5 *(Convexified Optimization Problem (With a Certain Structure of the Constraints) [GL11, 6.3.1])*
Let $Z_{ad} \subset Z = \mathbb{R}^{n_z}$ and let $\mathcal{J} : Z \to \mathbb{R}$ be a functional. Furthermore, let $G : Z \to \mathbb{R}^{n_G}$ and $H : Z \to \mathbb{R}^{n_H}$ be functions.

Find $\hat{z} \in Z_{ad}$ such that $\mathcal{J}(\hat{z}) + \mathcal{J}(\hat{z})(z - \hat{z})$ is minimized,
where $z \in \hat{z} + \tilde{\mathcal{T}}(Z_{ad}, \hat{z})$,
subject to the constraints

$$\begin{aligned} G(\hat{z}) + G'(\hat{z})(z - \hat{z}) &\le 0_{\mathbb{R}^{n_G}}, \\ H'(\hat{z})(z - \hat{z}) &= 0_{\mathbb{R}^{n_H}}, \end{aligned}$$

where $\tilde{\mathcal{T}}(Z_{ad}, \hat{z}) (\subset \mathcal{T}(Z_{ad}, \hat{z}))$ is a convex partial cone of the tangent cone $\mathcal{T}(Z_{ad}, \hat{z})$ as defined in Def. A.37.

This approximation in finite dimensions is used, e.g., in the local Slater condition [GL11, 6.3.6], or in the sequential quadratic programming (SQP) method (see Sect. 2.3.3), but in the latter case with a quadratic objective. The idea is to prove global first-order necessary optimality conditions for the convexified problem that are local first-order necessary optimality conditions for the original nonlinear problem, for details see, e.g., [GL11, Sect. 6.3]. Note that the last problem is finite-dimensional, but in numerics we always deal with finite dimensions.

For details on conic approximations and separation theorems, as required for a more general derivation of necessary optimality conditions, see [Ge12, Sect. 2.3.2 & 2.3.3].

2.2.2 Necessary Optimality Conditions

Necessary optimality conditions (NOC) provide a criterion that allows to figure out candidates for $\hat{z}$, under the assumption that a minimum $\hat{z}$ exists at all. By sufficient conditions we could ensure that we have actually found a minimum.

If first-order derivatives are involved, then we refer to the NOC as a first-order necessary optimality condition. If derivatives up to second-order appear in the condition, then we call it a second-order necessary optimality condition.

The structure, presupposed in Problem 2.3, allows to derive first-order necessary optimality conditions of the following type, formulated by Fritz John in 1948 [FJ48]. We start with some basic assumptions.

Assumption 2.6 *(Basic Assumptions on Spaces and Regularity)*

a) Z, W_G, and W_H are Banach spaces.

b) $\mathcal{J} : Z \to \mathbb{R}$, $G : Z \to W_G$ are F-differentiable and $H : Z \to W_H$ is continuously F-differentiable, respectively.

Assumption 2.7 *(Basic Assumptions for First-Order Necessary Optimality Conditions)*

a) The closed convex set Z_{ad} has interior points, i.e. $\mathring{Z}_{ad} \neq \emptyset$.

b) The closed convex cone $\mathcal{K}$ (with vertex at 0_{W_G}) has interior points, i.e. $\mathring{\mathcal{K}} \neq \emptyset$.

c) The image of $H'(\hat{z})$ is not a proper dense subset of W_H.

In the following theorem $\mathcal{K}^+$ denotes the positive polar cone of $\mathcal{K}$, see Def. A.36.

Theorem 2.8 *(Fritz John-Conditions (FJ-Conditions))*
Assume $\hat{z}$ is a local minimizer of Problem 2.3. Let Assumptions 2.6 and 2.7 hold, then there exist multipliers $\lambda := [\lambda_0, \lambda_G, \lambda_H] \in \mathbb{R} \times W_G^* \times W_H^*$, $\lambda \neq [0, 0_{W_G}, 0_{W_H}]$ s.t.

$$\lambda_0 \geq 0, \tag{2.3}$$

$$\lambda_G \in \mathcal{K}^+, \tag{2.4}$$

$$\langle \lambda_G, G(\hat{z}) \rangle_{W_G^*, W_G} = 0, \tag{2.5}$$

$$\langle \lambda_0 \mathcal{J}'(\hat{z}), d \rangle_{Z^*, Z} + \langle \lambda_G, G'(\hat{z}) d \rangle_{W_G^*, W_G} + \langle \lambda_H, H'(\hat{z}) d \rangle_{W_H^*, W_H} \geq 0 \quad \forall d \in Z_{ad} - \{\hat{z}\}. \tag{2.6}$$

For a proof of Th. 2.8 see, e.g., [Ge12] or [ZK79]. It relies on the open mapping theorem, stating that a linear, continuous, and surjective operator maps open sets to open sets (see, e.g., [We95, Th. IV.3.3]).

Every point $[z, \lambda]$ with non-trivial λ is called a *Fritz John-point*. If further $\lambda_0 \neq 0$ (w.l.o.g. we may consider $\lambda_0 = 1$ by scaling), then (2.3) – (2.6) are called Karush-Kuhn-Tucker (KKT) conditions and $[z, \lambda]$ is called a *KKT-point*, correspondingly. The minimum principle for optimal control problems may be derived from the FJ-conditions.

We use in this study the notation λ for the multiplier at the optimum. Note that, unless indicated otherwise, we do not write $\hat{\lambda}$, emphasizing that the multiplier is not unique in general.

The following version for necessary optimality conditions [HPUU09, Th. 1.46] requires weaker assumptions, but less structure is obtained.

Theorem 2.9 *(First-Order Necessary Optimality Conditions for General Constraints) Consider the optimization problem with general constraints $z \in Z_{ad}$, i.e. Problem 2.1. Assume $\hat{z}$ is a local minimizer of the problem. Let Z be a Banach space and the non-empty set $Z_{ad} \subset Z$ be closed and convex with a (open) neighbourhood V and let $J : V \to \mathbb{R}$ be G-differentiable at $\hat{z}$. Then there holds the variational inequality*

$$\begin{gathered}\hat{z} \in Z_{ad},\\ \langle \mathcal{J}'(\hat{z}), d\rangle_{Z^*,Z} \geq 0 \quad \forall d \in Z_{ad} - \{\hat{z}\}.\end{gathered}$$

We consider slightly different assumptions than for the Fritz John-conditions for specifying the last theorem in case of structured constraints.

Assumption 2.10 *(Modified Assumptions for First-Order Necessary Optimality Conditions)*

a) The closed convex set Z_{ad} is non-empty and is a subset of Z.

b) The closed convex cone $\mathcal{K}$ is a subset of W_G.

c) The feasible set Σ_{fs} as defined in (2.2) is non-empty.

The following necessary optimality conditions using the tangent cone $\mathcal{T}(\Sigma_{fs}; z)$ of the feasible set at z (see Def. A.37) can be proved.

Theorem 2.11 *(First-Order Necessary Optimality Conditions for Structured Constraints) Assume $\hat{z}$ is a local minimizer of Problem 2.3, but with $z \in \Sigma_{fs}$. Let Assumption 2.6, but with $\mathcal{J}$ being* **continuously** *F-differentiable, and Assumption 2.10 hold, then*

$$\begin{gathered}\hat{z} \in \Sigma_{fs},\\ \langle \mathcal{J}'(\hat{z}), d\rangle_{Z^*,Z} \geq 0 \quad \forall d \in \mathcal{T}(\Sigma_{fs}; \hat{z}).\end{gathered}$$

Proof. We extend here the proof stated for [HPUU09, Th. 1.52] to cover equality constraints as well. For this purpose, we consider

$$\mathcal{G}(z) := [G(z), H(z)]^\top \in \tilde{\mathcal{K}} \tag{2.7}$$

with $\tilde{\mathcal{K}} := \mathcal{K} \times \{0\}$ instead of G and $\mathcal{K}$ there. These constraints and $z \in Z_{ad}$ are summarized by $z \in \Sigma_{fs}$ by definition.

We have to demonstrate the inequality. For any $d \in \mathcal{T}(\Sigma_{fs}; \hat{z})$ there exist according to the definition of a tangent cone sequences $\{z_k\}_{k\in\mathbb{N}} \in \Sigma_{fs}$ and $\{\alpha_k\}_{k\in\mathbb{N}}$ in $\mathbb{R}$ with $\alpha_k > 0$ such that $z_k \to \hat{z}$ and $\alpha_k(z_k - \hat{z}) \to d$. Thus by the continuous F-differentiability of $\mathcal{J}$

$$0 \leq \alpha_k(\mathcal{J}(z_k) - \mathcal{J}(\hat{z})) = \langle \mathcal{J}'(\hat{z}), \alpha_k(z_k - \hat{z})\rangle_{Z^*,Z} + \alpha_k o(\|z_k - \hat{z}\|_Z) \to \langle \mathcal{J}'(\hat{z}), d\rangle_{Z^*,Z}$$

for sufficiently large k. □

In the context of minimization subject to constraints, the Lagrange function appears in a natural way. The Lagrange function is not only used within optimization, but this concept is also well-established in classical mechanics in a slightly different context, see Appendix C for more details.

Definition 2.12 *(Lagrange Function)*
Let $W := \mathbb{R} \times W_G \times W_H$, its dual will turn out to be the space for multipliers. The function $L : Z \times W^* \to \mathbb{R}$,

$$L(z, \lambda_0, \lambda_G, \lambda_H) := \lambda_0 \mathcal{J}(z) + \langle \lambda_G, G(z) \rangle_{W_G^*, W_G} + \langle \lambda_H, H(z) \rangle_{W_H^*, W_H}$$

is called Lagrange function (Lagrangian) for Problem 2.3, i.e. corresponding to the objective $\mathcal{J}$ and the constraints $G \in \mathcal{K}$ and $H = 0_{W_H}$.

Definition 2.13 *(Saddle Point; Lagrange Multiplier)*
Let Z_{ad} be a non-empty, convex set. Any $[\hat{z}, \lambda] \in Z \times W^*$ with $\lambda = [1, \lambda_G, \lambda_H]$, $\lambda_G \in \mathcal{K}^+$, $\lambda_H = 0_{W_H^*}$, satisfying

$$L(\hat{z}, \mu) \leq L(\hat{z}, \lambda) \leq L(z, \lambda) \quad \forall z \in Z_{ad} \, \forall \mu = [\mu_G, \mu_H] \in W^*, \mu_G \in \mathcal{K}^+, \tag{2.8}$$

is a saddle point of the Lagrange function $L(z, \lambda)$.
λ is called a Lagrange multiplier associated to $\hat{z}$.

The existence of saddle points is most easily shown for convex optimization problems. But the notion of Lagrange multipliers is in general not restricted to saddle points.

Note that in order to prove existence of Lagrange multipliers a certain compromise has to be made w.r.t. the choice of Z. On one hand Z has to be small enough, such that differentiability of the nonlinearities can be guaranteed, on the other hand Z should be large enough, such that Z^* is not too large and ensuring a certain regularity of the multipliers.

Further denotations for the Lagrange multiplier λ are adjoints, *adjoint states*, *costates* or, within the context of economics, *shadow prices*. The multipliers represent the marginal costs of violating the constraints. For a physical interpretation of the adjoint, see Appendix C.

Note that in literature, e.g. [HPUU09, Ge12], the multipliers are defined often as λ^* emphasizing that they live in the dual space. Note that we consider the multipliers λ directly as elements of the dual as e.g. in [Tr10].[4]

Remark 2.14 *(Primal and Dual Representations of Adjoints and Gradients)*

a) Using the dual operator $G'(\hat{z})^* : W_G^* \to Z^*$ corresponding to $G'(\hat{z}) : Z \to W_G$, we may write equivalently

$$\langle \lambda_G, G'(\hat{z}) d \rangle_{W_G^*, W_G} = \langle G'(\hat{z})^* \lambda_G, d \rangle_{Z^*, Z}.$$

Analogously, we may rewrite $\langle \lambda_H, H'(\hat{z}) d \rangle_{W_H^*, W_H} = \langle H'(\hat{z})^* \lambda_H, d \rangle_{Z^*, Z}$.

[4] Note that in general the space of adjoints could be a space of measures or exhibit even less structure.

b) Note also that, if Z is a Hilbert space, then a derivative $F'(\hat{z}) : Z \to \mathbb{R}$, i.e. $F'(\hat{z}) \in Z^*$, may be considered by its Riesz representation (see Th. A.8) that is called a gradient and denoted by $\nabla F \in Z$.

The Lagrange function allows to write necessary optimality conditions (2.3) – (2.6) in a more concise formulation. Using Remark 2.14 a), we may replace the optimality condition (2.6) by

$$\langle L'_z(\hat{z}, \lambda_0, \lambda_G, \lambda_H), d\rangle_{Z^*,Z} \geq 0 \quad \forall d \in Z_{ad} - \{\hat{z}\}. \tag{2.9}$$

The main advantage of Lagrange's idea (originally developed for variational problems) is that a problem with restrictions $G(z) \in \mathcal{K}$, that are usually difficult to handle, is replaced by an unrestricted problem. Other approaches following a similar idea, going back to [FM68], are so-called *penalty methods*, where a sequence of penalty terms of the form

$$p^{(k)}(\cdot) = \alpha^{(k)} dist(G(\cdot), \mathcal{K}) \tag{2.10}$$

is added to the objective turning a problem with the constraint $G(z) \in \mathcal{K}$ into an unrestricted minimization problem, but for $\mathcal{J}(z) + p^{(k)}(z)$. Outer penalty methods allow infeasible points for approximating the minimizer of the original problem, while inner penalty methods or *barrier methods* only work with a sequence of points in the interior of the feasible set. For penalty methods, we wish to find a minimizer for finite penalty parameters, shortly *penalties*, $\alpha^{(k)}$. This is the case for so-called *exact penalty functions*.

Definition 2.15 *(ℓ_1-Penalty Function)*
We consider the finite-dimensional case here. For a finite number of n_G inequality constraints G and n_H equality constraints H the penalty function (2.10) w.r.t. the $\|\cdot\|_1$ norm reads

$$p_k(z) = \alpha^{(k)} \left(\sum_{i=1}^{n_G} \max\{G_i(z), 0\} + \sum_{j=1}^{n_H} |H_j(z)| \right).$$

ℓ_1-penalty functions may be proved to be exact, i.e. for a finite parameter $\alpha^{(k)}$ a minimizer of the original problem is obtained. For this and further details on penalty methods, see, e.g., [GL11, Sect. 7.6 & 7.7].

Unfortunately, exact penalty functions are in general non-differentiable. In order to keep differentiability and exactness, the concept of Lagrangians and penalty methods are combined as follows. Note that inequality constraints are eliminated here by the introduction of so-called slack variables, see, e.g., [GL11, Subsect. 7.8.4].

Definition 2.16 *(Augmented Lagrange Function)*
The function

$$L_a(z, \lambda_0, \lambda_H, \alpha) := L(z, \lambda_0, \lambda_H) + \frac{\alpha}{2}\|H(z)\|^2_{W_H}$$

is called augmented Lagrange function (augmented Lagrangian) for Problem 2.3.

Under certain conditions, e.g., on second derivatives of H and second-order sufficient optimality conditions, the exactness of the augmented Lagrangian may be proved [He75, Th. 4.2]. For an iterative algorithm how to determine the multipliers and the penalties sensitively, see Subsection 2.3.2.

In finite-dimensional spaces, Assumption 2.7 c) on the non-density of $image(H'(\hat{z}))$ is satisfied automatically. However, this is not clear in infinite-dimensional spaces, since the image of a linear, continuous operator H is not closed in general. If the target space of H has the structure $W_H = W_H^1 \times \mathbb{R}^{n_\Psi}$ (Ψ representing, e.g., initial and boundary conditions, see (2.60) and 2.79 below), we have the following result (for a proof, see [Ge12, Th. 2.3.29]):

Theorem 2.17 *(Closed Image Space)*
Let $H \in \mathcal{L}(Z, W)$ with $W = W_H \times \mathbb{R}^{n_\Psi}$, W_H a Banach space and $n_\Psi \in \mathbb{N}$. If, for the derivative $H'(\hat{z}) = (T_1, T_2)$ with $T_1 \in \mathcal{L}(Z, W_H)$, T_1 surjective, and $T_2 \in \mathcal{L}(Z, \mathbb{R}^{n_\Psi})$, then $image(H'(\hat{z}))$ is closed in W.

In order to guarantee $\lambda_0 \neq 0$ and the existence of Lagrange multipliers some so-called *constraint qualifications* are required. Usually, $\hat{z}$ is involved in the CQ and, thus, the CQ cannot be analyzed *a priori* without further knowledge on $\hat{z}$. For further details on constraint qualifications we refer to Appendix A.2.2).

In the following we consider $L(z, \lambda_G, \lambda_H) := L(z, 1, \lambda_G, \lambda_H)$ as Lagrange function without change of notation.

Theorem 2.18 *(KKT-Conditions)*
Assume $\hat{z}$ is a local minimizer of Problem 2.3. If Assumptions 2.6, 2.7, and A.40 (the Robinson CQ) are fulfilled, there exist Lagrange multipliers $\lambda := [\lambda_G, \lambda_H] \in W_G^* \times W_H^*$.

a) Moreover λ fulfils the KKT-system

$$G(\hat{z}) \in \mathcal{K}, \tag{2.11}$$

$$\langle \lambda_G, k \rangle_{W_G^*, W_G} \geq 0 \quad \forall k \in \mathcal{K}, \tag{2.12}$$

$$\langle \lambda_G, G(\hat{z}) \rangle_{W_G^*, W_G} = 0, \tag{2.13}$$

$$H(\hat{z}) = 0_{W_H}, \tag{2.14}$$

$$\hat{z} \in Z_{ad}, \tag{2.15}$$

$$\langle L'_z(\hat{z}, \lambda_G, \lambda_H), d \rangle_{Z^*, Z} \geq 0 \quad \forall d \in Z_{ad} - \{\hat{z}\}. \tag{2.16}$$

The combination of (2.11) – (2.13) is a so-called complementarity condition. In finite dimensions, (2.11) reads $G(\hat{z}) \leq 0$, thus (2.12) yields $\lambda_G \geq 0$.

b) Without inequality constraints the KKT-system can be written as

$$\tilde{\mathcal{G}}(\hat{z}, \lambda) := \begin{bmatrix} H(\hat{z}) \\ \mathcal{J}'(\hat{z}) + H'(\hat{z})^* \lambda_H \end{bmatrix} = \begin{bmatrix} 0_{W_H} \\ 0 \end{bmatrix}, \quad \hat{z} \in Z_{ad}.$$

c) *Consider a finite-dimensional optimization problem, i.e. let $Z \subset \mathbb{R}^{n_z}$, $W_G = \mathbb{R}^{n_G}$, and $W_H = \mathbb{R}^{n_H}$. Moreover, if the stronger constraint qualification LICQ (Assumption A.45) holds, then (2.16) yields*

$$\nabla_z L(\hat{z}, \lambda) = 0_{\mathbb{R}^{n_z}}$$

and λ is uniquely determined.

Proof.

a) This follows from Theorem 2.8 using $\lambda_0 \neq 0$. Furthermore, for the reformulation of the KKT-conditions, we have used Def. 2.12, Remark 2.14 a) (yielding (2.9)), the cone structure is exploited, and the original constraints are recalled in addition.

b) This is just the special case, when no inequality constraints are present.

c) See [Ge12, Coroll. 2.3.39].

□

For other constraint qualifications implying the Robinson CQ, we refer to Appendix A.2.2.

A similar result, but with a more complicated structure replacing (2.12) & (2.13), may be obtained, if we require only $\mathcal{K}$ to be a closed convex set without a cone structure [BS98].

(2.11) – (2.13) can be reformulated by means of a NCP[5] function yielding a nonsmooth equation. A useful NCP function is the Fischer-Burmeister function, whose square is differentiable, see, e.g., [GH11].

Here we understand as KKT-conditions the conditions following from the Fritz John-conditions together with the original constraints that follow trivially, but have to be fulfilled necessarily as well. This is different to many references.

Remark 2.19 *(Exact and Formal Lagrange Method) [Tr10, Sect. 2.10]*
Here we have derived the necessary optimality conditions by the so-called exact Lagrange method. This method requires to state the function spaces suitably, to figure out the right differentiability of the operators, to derive the adjoint equations rigorously, and other challenges in functional analysis have to be tackled.

Another approach is the formal Lagrange method. It means to put up the adjoint equation in a way, where, e.g., differential operators are considered formally, all adjoints are assumed to be actually functions, and all appearing functions are supposed to be square-integrable. In general this allows for a more direct derivation of the adjoint equation, that may be justified mathematically rigorously afterwards. In particular for complicated problems this approach might help to establish some formal results at first.

In general necessary optimality conditions are not sufficient and further information is required. If $\mathcal{J}$ is convex, the inequality constraints are continuously differentiable convex functions, and

[5] nonlinear complementarity problem

the equality constraints are affine-linear, then the KKT-conditions are not only necessary, but sufficient as well, see, e.g., [Ge12, Th. 2.3.41].

2.2.3 Sufficient Optimality Conditions

Briefly we discuss sufficient optimality conditions. We focus on second-order sufficient optimality conditions for equality constrained problems, following Tröltzsch [Tr10, Lemma 6.4] and Maurer and Zowe [MZ79, Ma81], resp. If the second-order conditions are formulated by means of the Lagrange function, more structure may be observed.

Lemma 2.20 *(Second-Order Sufficient Optimality Conditions for Pure Equality Constraints) In addition to the Assumptions 2.6 & 2.7 for Problem 2.3, in which here G is neglected and inequality constraints enter by means of Z_{ad}, we assume the Zowe-Kurcyusz CQ (Assumpt. A.41) and that $\mathcal{J} : Z \to \mathbb{R}$ and $H : Z \to W_H$ are twice continuously F-differentiable.*

Let $[\hat{z}, \lambda]$ fulfil the first-order necessary optimality conditions in Th. 2.18 a). If $d \in Z_{ad}$ and if for some $\varepsilon > 0$

$$L''_{zz}(\hat{z}, \lambda)[d, d] = \mathcal{J}''(\hat{z})[d, d] + \langle \lambda, H''(\hat{z})[d, d] \rangle_{W^*,W} \geq \varepsilon \|d\|_Z^2 \quad \forall d \in C_{Z_{ad}}(\hat{z}) \tag{2.17}$$

such that

$$\langle H'(\hat{z}), d \rangle_{Z^*,Z} = 0,$$

then $\hat{z}$ is locally optimal for Problem 2.3 with equality constraints only.

The conical hull $C_{Z_{ad}}$, that enters here and is related to the Zowe-Kurcyusz CQ (Assumption A.41), is defined in Def. A.33. For details on the notation of second-order F-derivatives, using a bilinear form like $H''(\cdot)[d, d]$ here, see, e.g., [Tr10, Sect. 4.9].

The latter result may be generalized to inequality constraints, if in addition (i) strict complementarity holds, i.e. for the associated multipliers $\lambda_G <_K 0$, and (ii) $\langle G'(\hat{z}), z \rangle_{Z^*,Z} = 0$, then the second-order sufficient optimality condition (2.17) holds as well.

Lemma 2.21 *(Second-Order Sufficient Optimality Conditions in Finite Dimensions [BZ82], [BSS93, Th. 4.4.2])*
In addition to the Assumptions 2.6 & 2.7 for Problem 2.3, we assume the Zowe-Kurcyusz CQ (Assumpt. A.41). We consider the finite-dimensional case with $Z = \mathbb{R}^{n_z}$, $W_G = \mathbb{R}^{n_G}$, and $W_H = \mathbb{R}^{n_H}$. Furthermore, let $\mathcal{J} : \mathbb{R}^{n_z} \to \mathbb{R}$, $G : \mathbb{R}^{n_z} \to \mathbb{R}^{n_G}$, and $H : \mathbb{R}^{n_z} \to \mathbb{R}^{n_H}$ be twice continuously F-differentiable.

Let $[\hat{z}, \lambda]$ fulfil the first-order necessary optimality condition (2.16). For $z \in \mathbb{R}^{n_z}$ let

$$L''_{zz}(\hat{z}, \lambda)[d, d] > 0 \quad \forall d \in C_0(\hat{z}), \tag{2.18}$$

where $C_0(\hat{z})$ denotes the critical cone (see Def. A.35), then there exists a neighbourhood $V(\hat{z})$ of $\hat{z}$ and some $\alpha > 0$ such that

$$\mathcal{J}(z) \geq \mathcal{J}(\hat{z}) + \alpha \|z - \hat{z}\|^2,$$

i.e. $\hat{z}$ is strictly locally optimal for Problem 2.3.

A second-order necessary optimality condition on the critical cone is presented in [GK02, Th. 2.54] for $Z = \mathbb{R}^{n_z}$. For problems with inequality constraints only, first-order sufficient optimality conditions may derived [MZ79].

Note that in case of unrestricted optimization problems, we recover the first- and second-order optimality conditions known from standard calculus.

2.3 Function Space Methods in Optimization

Optimization methods are iterative algorithms for finding minimizers(/maximizers) for the underlying optimization problems. Usually, we are satisfied, if the method can be proved to converge to stationary points. Note that stationary points, i.e. points that fulfil necessary optimality conditions, are only candidates for minimizers(/maximizers) and might be saddle points as well. The primary goal is to show fast local convergence, to demonstrate global convergence might be out of reach. All fast algorithms rely to some extent on the Newton method for optimization problems in Banach spaces.

Differential equations are posed in infinite-dimensional spaces and so the corresponding optimization problems involving differential equations are posed in a functional analytic setting in general. However, infinite dimensions are not suitable for numerical algorithms, since analytic solutions are in most cases not available. Thus at some point we have to "discretize" in order to obtain a finite-dimensional problem.

When we have to solve a general optimization problem, we may discretize directly, and then optimize the discretized equations in a second step. This strategy is a so-called *first-discretize-then-optimize (FDTO) approach*. On the other hand we may optimize within the function space setting, obtain certain optimality conditions (see Subsections 2.2.2 and 2.2.3) and then discretize these conditions to obtain a numerical optimal solution. This gives raise to a *first-optimize-then-discretize (FOTD) approach*.

For the different optimization methods see Figure 2.1.

On a lower level "optimization" and "discretization" may commute, but in general this is not the case and slight deviations in the discretized systems may turn up. For numerical schemes in optimal control, where FOTD and FDTO do commute, see, e.g., [AF12]. FOTD algorithms work in the same spaces as the problems are stated and the structure of the problem may be exploited. Hence no discretization error is introduced at this stage, contrary to methods in which we discretize firstly. When realized numerically, function space methods, yield mesh independence results and error estimators in addition.

In the next two subsections, we consider the gradient method in function spaces, the multiplier-penalty method, and Newton-type methods.

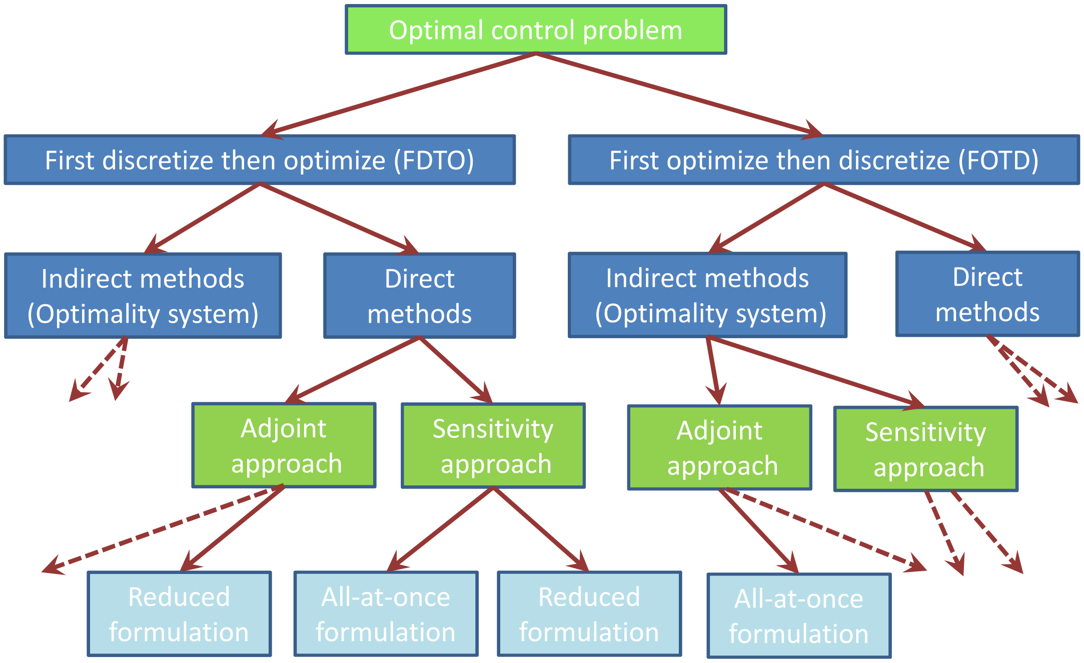

Figure 2.1: Different optimization methods. In case of optimal control problems, there is a choice between sensitivity- vs. adjoint-based approaches, for details see Subsect. 2.5.1. If we may solve the state equation uniquely, then a reduced approach (see Subsect. 2.4.3) as well as an all-at-once approach may be pursued. The direct FDTO approach in all-at-once formulation is called full discretization or collocation, the direct FDTO approach in reduced formulation is the direct shooting method.

2.3.1 Descent Methods

The gradient method (also method of steepest descent) is one of the most standard methods for solving optimization problems. Since in finite dimensions it is obvious that the gradient is perpendicular on contour lines, the negative gradient determines the direction of steepest descent. It is intuitive to follow the descent, until a stationary point has been reached, at which within a certain numerical tolerance, no further descent happens. Usually a line search strategy is applied in order to obtain suitable step lenghts and to reduce a zig-zagging behaviour (though it cannot be avoided completely, see the gradient method). This method is straightforward in $\mathbb{R}^d$ or any space with finite dimensions, but the definition of a suitable gradient in a function space (that is in general infinite-dimensional) is a subtle question. Note that we assume in this subsection always that $\mathcal{J}$ is G-differentiable (cf. [Tr10, Subsect. 2.12.1]) and Z is a Banach space.

Unconstrained Optimization

In order to avoid projections at first, we start with the unconstrained case, i.e. $Z_{ad} = Z$. Then the first-order optimality condition states that a local minimum (/maximum) $\hat{z} \in Z$ satisfies

$$\mathcal{J}'(\hat{z}) = 0_{Z^*}$$

that is a standard result in calculus.

We present a general descent method, that in this context turns out to be globally convergent. Let $z^{(k)}$ denote the variable at step k of the algorithm, $k \in \mathbb{N}_0$. For step k we introduce a descent direction $s^{(k)} \in Z$ as

$$\langle \mathcal{J}'(z^{(k)}), s^{(k)} \rangle_{Z^*,Z} < 0. \tag{2.19}$$

In the Banach space setting, there is no straightforward way to compute descent directions. The derivative of $\mathcal{J}$ lives in the dual space,

$$\mathcal{J}'(z^{(k)})(z) = \langle \mathcal{J}'(z^{(k)}), z \rangle_{Z^*,Z},$$

i.e. $\mathcal{J}'(z^{(k)}) \in Z^*$, but it is in general not in Z.

If Z is a Hilbert space, then by the Riesz representation theorem (Th. A.8) we may identify Z^* by Z and there exists for every $\hat{z}$ a unique element $\gamma(\hat{z}) \in Z$ such that

$$\mathcal{J}'(z^{(k)})(z) = (\gamma(z^{(k)}), z)_Z.$$

We write accordingly $\nabla \mathcal{J}(z^{(k)}) := \gamma(z^{(k)})$ and the negative gradient is a well-defined search direction in Z. Indeed, the choice of wish, $s^{(k)} = -\nabla \mathcal{J}(z^{(k)})$, yields directly

$$\langle \mathcal{J}'(z^{(k)}), s^{(k)} \rangle_{Z^*,Z} = -\|\nabla J(z^{(k)})\|_Z^2 < 0. \tag{2.20}$$

If Z is not a Hilbert space, e.g., $Z = L^\infty(\Omega)$, there does not a hold a representation theorem nor is the dual space really useful. However, in normed spaces a so-called metric gradient may be introduced [GT72]. For this reason, we usually assume Z to be a Hilbert space in the following.

For determining possible minimizers, we consider the following algorithm including an adaptive step size rule.

Algorithm 2.22 *(General Descent Method)*

1.) Choose a starting point $z^{(0)} \in Z$. Set $k := 0$.

2.) If $\mathcal{J}'(z^{(k)}) = 0$, then STOP.

3.) Compute a descent direction $s^{(k)} \in Z$ s.t. $\langle \mathcal{J}'(z^{(k)}), s^{(k)} \rangle_{Z^*,Z} < 0$.

4.) Determine a step size $\alpha^{(k)} > 0$ s.t. $\mathcal{J}(z^{(k)} + \alpha^{(k)} s^{(k)}) < \mathcal{J}(z^{(k)})$.

5.) Update $z^{(k+1)} := z^{(k)} + \alpha^{(k)} s^{(k)}$.

6.) Set $k := k + 1$ and return to 2.).

The decrease of $\mathcal{J}$ by means of a descent direction (2.20) can be arbitrarily small. Note that the slope of $\mathcal{J}$ at $z^{(k)}$ in direction $s^{(k)}$ is given by

$$\frac{d}{dt}\mathcal{J}\left(z^{(k)} + t\frac{s^{(k)}}{\|s^{(k)}\|_Z}\right)\Bigg|_{t=0} = \frac{\langle \mathcal{J}'(z^{(k)}), s^{(k)}\rangle_{Z^*,Z}}{\|s^{(k)}\|_Z}.$$

This motivates the following definition, loosely speaking it links small decreases to small steepest slopes and to small step sizes simultaneously.

Definition 2.23 *(Admissible Search Directions and Step Sizes)*

a) For an admissible search direction, we require that the search direction $s^{(k)}$ fulfils

$$\frac{\langle \mathcal{J}'(z^{(k)}), s^{(k)}\rangle_{Z^*,Z}}{\|s^{(k)}\|_Z} \overset{k\to\infty}{\longrightarrow} 0 \quad \Longrightarrow \quad \|\mathcal{J}'(z^{(k)})\|_{Z^*} \overset{k\to\infty}{\longrightarrow} 0.$$

b) An admissible step size $\alpha^{(k)}$ means that

$$\begin{aligned} &\mathcal{J}(z^{(k)} + \alpha^{(k)}s^{(k)}) - \mathcal{J}(z^{(k)}) < 0 \quad \text{and} \\ &\mathcal{J}(z^{(k)} + \alpha^{(k)}s^{(k)}) - \mathcal{J}(z^{(k)}) \overset{k\to\infty}{\to} 0 \quad \Longrightarrow \quad \frac{\langle \mathcal{J}'(z^{(k)}), s^{(k)}\rangle_{Z^*,Z}}{\|s^{(k)}\|_Z} \overset{k\to\infty}{\to} 0. \end{aligned}$$

If Z is a Hilbert space and we consider $s^{(k)} = -\nabla\mathcal{J}(z^{(k)})$, then the descent method is called the *gradient method* (or *method of steepest descent*). The latter can be seen by the Cauchy-Schwarz inequality. Note that the **normalized** negative gradient might be considered as search direction as well.

If the *angle condition*

$$\langle \mathcal{J}'(z^{(k)}), s^{(k)}\rangle_{Z^*,Z} \leq -\eta\|\mathcal{J}'(z^{(k)})\|_{Z^*}\|s^{(k)}\|_Z$$

for some $\eta \in (0,1)$ is satisfied, then the search direction $s^{(k)}$ is admissible.

As an example for a step size rule we consider the Armijo (or backtracking) line search, that allows to generate admissible step sizes under certain conditions.

Algorithm 2.24 *(Armijo Rule)*
Let $s^{(k)}$ be a descent direction w.r.t. $\mathcal{J}$ at $z^{(k)}$. Furthermore $\beta_A \in (0,1)$ and $\sigma_A \in (0,1/2)$.

We determine $\alpha^{(k)} = \beta_A^{(k)}$ for the minimal k s.t.

$$\Theta(z^{(k)} + \beta_A^{(k)}) \leq \Theta(z^{(k)}) + \sigma_A\beta^{(k)}\langle\Theta'(z^{(k)}), s^{(k)}\rangle_{Z^*,Z}.$$

The functional $\Theta : Z \to \mathbb{R}$ is a so-called *merit function*. The most straightforward choice for a merit function is considering $\mathcal{J}(z) + \alpha\left(dist\{G(z), \mathcal{K}\} + \|H(z)\|_{W_H}\right)$ for a parameter $\alpha > 0$ where *dist* means the distance of a point to a set. Note that in finite dimensions we find $dist\{G(z), \mathcal{K}\} = \max\{G(z), 0\}$. Another typical choice for a merit function is $\Theta(z) = \|f(z)\|_W^2/2$, since minimizing Θ is equivalent to a zero of a vector-valued function $f : Z \to W$, that encodes, e.g., optimality conditions.

It is easy to see, that Armijo step sizes exist [HPUU09, Lemma 2.2], if $\mathcal{J}$ is uniformly continuous in some neighbourhood. Typically values used in applications are $\beta_A = 0.5$ or $= 0.9$ and $\sigma_A = 0.5$.

Lemma 2.25 *(Admissible Armijo Step Sizes)*
Let $\Theta : Z \to \mathbb{R}$ be a merit function with Θ' being uniformly continuous on

$$V_\rho(z, s; \Theta) := \{z + s \,|\, \Theta(z) \leq \Theta(z^{(0)}), \|s\|_Z \leq \rho\} \tag{2.21}$$

for some radius $\rho > 0$. Let $z^{(k)}$ and $s^{(k)}$ be generated by Algorithm 2.22 for $\mathcal{J} = \Theta$, where we require the norm of the search direction to be bounded from below by

$$\|s^{(k)}\|_Z \geq \phi\left(-\frac{\langle \Theta'(z^{(k)}), s^{(k)}\rangle_{Z^*,Z}}{\|s^{(k)}\|_Z}\right),$$

$\phi : \mathbb{R}_0^+ \to \mathbb{R}_0^+$ being monotonically increasing and strictly positive on $\mathbb{R}^+$. Then the step sizes $\alpha^{(k)}$, chosen by Algorithm 2.24, are admissible.

For a proof see, e.g., [HPUU09, Lemma 2.3], where $\mathcal{J}$ may be replaced by a more general Θ.

Note that there exist many other step size rules, e.g. bisection or Wolfe-Powell.

Under suitable assumptions, we may prove that every accumulation point of $z^{(k)}$ is a stationary point, i.e., we have global convergence:

Theorem 2.26 *(Global Convergence of General Descent Method)*
We assume Z to be a Banach space. Let $z^{(k)}$, $s^{(k)}$ and $\alpha^{(k)}$, the latter both admissible, be generated by Algorithm 2.22. If $\mathcal{J} : Z \to \mathbb{R}$ is continuously F-differentiable and the generated sequence $\{\mathcal{J}(z^{(k)})\}$ is bounded from below, then

$$\lim_{k\to\infty} \mathcal{J}'(z^{(k)}) = 0.$$

For a proof, see [HPUU09, Th. 2.2].

The steepest descent may turn out to be inefficient in the sense that many iterations in Algorithm 2.22 are needed. In general, the step size cannot be determined analytically. Furthermore, if the state equation is nonlinear, an iterative solver, e.g. of Newton-type has to be employed anyways. In practice it is important in the sense that, if the Newton method or another method does not work, then it makes sense that the algorithm switches to the steepest descent method.

Optimization on Closed Convex Sets

In addition to the last subsubsection we consider simply constrained problems

$$\min\ f(z) \text{ s.t. } z \in Z_{ad}, \tag{2.22}$$

where we restrict the variable z to

$$Z_{ad} = \{z \in Z \,|\, z_{min}(x) \leq z(x) \leq z_{max}(x) \quad \text{a.e. on } \Omega\} \tag{2.23}$$

where z_{min}, $z_{max} \in L^\infty(\Omega)$, $\Omega \subset \mathbb{R}^d$ an open set. Let Z be a Hilbert space with partial order, for ease of presentation we consider $Z = L^2(\Omega)$. We may compute the projection $P_{Z_{ad}} : Z \to Z_{ad}$ pointwise by means of $\tilde{P} : \mathbb{R} \to \mathbb{R}$:

$$P_{Z_{ad}}(z)(x) := \tilde{P}_{[z_{min}(x), z_{max}(x)]}(z(x)) = \max(z_{min}(x), \min(z(x), z_{max}(x))). \tag{2.24}$$

Thus for $Z = L^2(\Omega)$ and box constraints Z_{ad}, the last KKT-condition that is a variational inequality may be reformulated.

Lemma 2.27 *(Reformulation of Variational Inequality by Projection onto Box Constraints) Let $\lambda_0 = 1$, Z be a Hilbert space with partial order and we identify $Z^* = Z$, such that $\nabla f(z)$ is the Riesz representation of $f'(z)$. Furthermore, we have box constraints for z, i.e. $Z_{ad} = \{z \in Z \,|\, z_{min} \leq z \leq z_{max}\}$ with $z_{min} \leq z_{max}$ and we write for the pointwise Euclidean projection (as in (2.24)) onto the admissible set*

$$P_{Z_{ad}}(z) := \max\{z_{min}, \min\{z, z_{max}\}\}.$$

Then the necessary optimality condition corresponding to (2.22) is a variational inequality. It is equivalent to

$$\hat{z} = P_{Z_{ad}}(\hat{z} - \tilde{\gamma}\nabla f(\hat{z})) \quad \forall \tilde{\gamma} > 0. \tag{2.25}$$

If we consider $f = L$, then this yields directly that (2.9) may be written equivalently

$$\hat{z} = P_{Z_{ad}}(\hat{z} - \tilde{\gamma}\nabla L(\hat{z})) \quad \forall \tilde{\gamma} > 0. \tag{2.26}$$

In case of a reduced optimal control problem, where formally $z = u$, (2.23) may be interpreted as box constraints for the control. This will be exploited further in Lemma 2.52 as well as in Lemma 2.44 for the all-at-once approach. For the following algorithm $\tilde{\gamma} > 0$ will be fixed, w.l.o.g. we set here $\tilde{\gamma} = 1$.

We distinguish between two types of methods, an *admissible method* (a *feasible method*[6]), i.e. $z^{(k)} \in Z_{ad}$ for all k and an *inadmissible method* (*unfeasible method*), where we require only convergence to an admissible (feasible) solution. Furthermore, we have the choice between performing a line search first and then to project vs. to project and then do a line search, i.e. a line search along the projected path

$$\{P_{Z_{ad}}(z^{(k)} + \alpha s^{(k)}) \,|\, \alpha \geq 0\}.$$

The first idea is not efficient, yielding possibly small or zero step sizes. The matter with the second ansatz is, that not any descent direction works, i.e. a descent direction might yield an ascent along the projected path. For a detailed discussion of the second approach in finite dimensions we refer to [Ke99]. We focus on the following admissible modification of Algorithm 2.22 for the projected path.

[6] Admissible method corresponds to our notion of an admissible set (Remark 2.4), but the notion feasible method is common in literature.

Algorithm 2.28 *(Projected Descent Method)*

1.) *Choose a starting point* $z^{(0)} \in Z_{ad}$. *Set* $k := 0$.

2.) *If* $\mathcal{J}'(z^{(k)}) = 0$, *then STOP.*

3.) *Compute a descent direction* $s^{(k)} \in Z$ *s.t.* $\langle \mathcal{J}'(z^{(k)}), s^{(k)} \rangle_{Z^*,Z} < 0$.

4.) *Determine a step size* $\alpha^{(k)} > 0$ *s.t.* $\mathcal{J}(P_{Z_{ad}}(z^{(k)} + \alpha^{(k)} s^{(k)})) < \mathcal{J}(z^{(k)})$.

5.) *Update* $z^{(k+1)} := P_{Z_{ad}}(z^{(k)} + \alpha^{(k)} s^{(k)})$.

6.) *Set* $k := k + 1$ *and return to 2.).*

A step size rule for Step 4.) is the projected Armijo rule that is well-defined, see, e.g. [HPUU09, 2.2.2.1]. We may replace the objective by a merit function in the context of a projected Armijo line search as well. Again global convergence can be proved.

Theorem 2.29 *(Global Convergence of Projected Descent Method)*
We assume Z *to be a Hilbert space and* Z_{ad} *is non-empty, closed, and convex. Let* $z^{(k)}$, $s^{(k)}$ *and* $\alpha^{(k)}$, *the latter both admissible, be generated by Algorithm 2.28 with* $k \in \mathbb{N}_0$. *If* $\mathcal{J} : Z \to \mathbb{R}$ *is continuously F-differentiable, the generated sequence* $\{\mathcal{J}(z^{(k)})\}$ *is bounded from below, and let for some* $\tilde{\alpha} > 0$ *the gradient* $\nabla \mathcal{J}$ *be Hölder continuous of* $\tilde{\alpha}$*-order on* $V_\rho(z, s; \mathcal{J})$ *as defined in (2.21), then*

$$\lim_{k\to\infty} \|z^{(k)} - P_{Z_{ad}}(z^{(k)} - \nabla \mathcal{J}(z^{(k)}))\|_Z = 0.$$

For a proof, compare [HPUU09, Th. 2.4].

2.3.2 Multiplier-Penalty Method

A method for unconstrained problems is the multiplier-penalty method, where the augmented Lagrange function, see Def. 2.16, is minimized. We recall that inequality constraints may be treated by the introduction of slack variables. An algorithm for determining penalty parameters $\alpha^{(k)}$ is

Algorithm 2.30 *(Multiplier-Penalty Method)*

1.) *Choose starting points* $z^{(0)}$, $\lambda^{(0)}$, *a starting penalty* $\alpha^{(0)} > 0$, *and a parameter* $\sigma_M \in (0, 1)$. *Set* $k := 0$.

2.) *If* $[z^{(k)}, \lambda^{(k)}]$ *is a KKT-point of the originally restricted problem, then STOP.*

3.) *Update multipliers by* $\lambda^{(k+1)} := \lambda^{(k)} + \alpha^{(k)} H(z^{(k+1)})$.

4.) *If* $\|H(z^{(k+1)}\|_{W_H} \geq \sigma_M \|H(z^{(k+1)}\|_{W_H}$, *increase* $\alpha^{(k+1)}$, *else* $\alpha^{(k+1)} := \alpha^{(k)}$.

5.) *Set* $k := k + 1$ *and return to 2.).*

Further details for an efficient implementation, that can be in this case quite tricky, see [KS03a, KS03b].

2.3.3 Newton-Type Methods

Variational equations and complementarity conditions may be rewritten as nonsmooth equations. Generalizations of the Newton method are suitable for this case. However, we have to distinguish between the finite-dimensional setting, presented, e.g., in [HPUU09, Sect. 2.3], and the function space setting that is infinite-dimensional. In this book we focus on the infinite-dimensional case.

Lagrange-Newton Method

KKT-systems may be reformulated as possibly nonsmooth operator equations of the form

$$f(z) = 0_W,$$

for $f : Z \to W$, Z, W Banach spaces, see, e.g., Eq. (2.25). Applying the Newton method to the KKT-system is called the *Lagrange-Newton method*, one very powerful approach for solving equality constrained optimization problems.

Algorithm 2.31 *(Local Generalized Newton Method)*

1.) Choose a starting point $z^{(0)} \in Z$. Set $k := 0$.

2.) If a stopping criterion is satisfied, then STOP.

3.) Choose an invertible operator $M^{(k)} \in \mathcal{L}(Z, W)$.

4.) Compute a search direction $s^{(k)} \in Z$ by solving $M^{(k)} s^{(k)} = -f(z^{(k)})$.

5.) Update $z^{(k+1)} = z^{(k)} + s^{(k)}$.

6.) Set $k := k + 1$ and return to 2.).

For the analysis of the latter algorithm a certain smallness assumption,

$$\|[M^{(k)}]^{-1}(f(\hat{x} + d^{(k)}) - f(\hat{x})) - d^{(k)}\|_W = o(\|d^{(k)}\|_Z) \quad \text{as } \|d^{(k)}\|_Z \to 0 \tag{2.27}$$

is required. As common, we split this assumption into two parts.

Assumption 2.32 *(Regularity Condition for the Local Generalized Newton Method)*
All operators $M^{(k)}$ are continuously invertible, i.e.

$$\|[M^{(k)}]^{-1}\|_{\mathcal{L}(W,Z)} \leq Const \quad \forall k \in \mathbb{N}_0. \tag{2.28}$$

Assumption 2.33 *(Approximation Condition for the Local Generalized Newton Method)*
Let the distance $d^{(k)}$ to the solution be small. If f' is Hölder continuous of $\tilde{\alpha}$-order, where $\tilde{\alpha} \geq 0$, then the solution is approximated s.t.

$$\|f(\hat{x} + d^{(k)}) - f(\hat{x}) - M^{(k)} d^{(k)}\|_W = O(\|d^{(k)}\|^{1+\tilde{\alpha}}) \quad \textit{as } \|d^{(k)}\|_Z \to 0. \tag{2.29}$$

Note that $\tilde{\alpha} = 0$ corresponds just to the case that f' is continuous.

For the local generalized Newton method we have the following convergence result [HPUU09, Th. 2.9].

Theorem 2.34 *(Fast Convergence of the Local Generalized Newton Method)*
Let Z, W be Banach spaces and $f : Z \to W$. Let $z^{(k)}$ be generated by Algorithm 2.31 in order to approximate solutions $\hat{z}$ of the operator equation $f(z) = 0_W$. For a starting point $z^{(0)}$ sufficiently close to $\hat{z}$, there holds:

a) If Assumption 2.32 holds, then $z^{(k)} \to \hat{z}$ linearly.

b) If Assumption 2.33 holds and f' is $\tilde{\alpha}$-order Hölder continuous, then $z^{(k)} \to \hat{z}$ superlinearly with order $1 + \tilde{\alpha}$. If $\tilde{\alpha} = 0$ this is simple superlinear convergence.

For globalized Newton methods see [Be99, Ke99] and for its superlinear convergence see [KS94].

If f is continuously F-differentiable, then we may choose $M^{(k)} = f'(z^{(k)})$ and Algorithm 2.31 becomes the classical Newton method. In this case, Assumption 2.32 reduces to

$$\|f'(z^{(k)})^{-1}\|_{\mathcal{L}(W,Z)} \leq Const \quad \forall k \in \mathbb{N}_0,$$

that is guaranteed by the continuity of f' at $\hat{z}$ and that $f' \in \mathcal{L}(Z, W)$ is continuously invertible. Assumption 2.33 follows from the F-differentiability of f at $\hat{x}$ for $\tilde{\alpha} = 0$. Thus we have locally superlinear convergence.

In case of quasi-Newton methods we consider an approximation or perturbation, resp., of the Jacobian, i.e. $M^{(k)} \approx f'(z^{(k)})$. The reason for this might be, e.g., that the Jacobian is expensive to compute or due to insufficient accuracy in the implementation. If the Dennis-Moré condition (see [DM74, DM77, DS83] and for the infinite-dimensional case [DR14])

$$\lim_{k\to\infty} \frac{\|(M^{(k)} - f'(z^{(k)}))s^{(k)}\|_Z}{\|s^{(k)}\|_Z} = 0$$

is satisfied, than the approximation or perturbation does not destroy the superlinear convergence.

If f is semismooth, see Def. A.29 e), then we may choose $M^{(k)} \in \partial^* f(z^{(k)})$ as an arbitrary invertible element of the subdifferential of f and Algorithm 2.31 is the *local semismooth Newton method*. In this case we obtain locally superlinear convergence as well.

The common *primal-dual active set strategy* is a special case in the class of semismooth Newton methods [BIK99, HIK03]. The primal-dual active set strategy is widely used for restricted optimization problems involving PDE, since the weak formulation of an elliptic PDE yields a quadratic optimization problem.

A globalization strategy for the semismooth Newton method is derived in our article [GHK17] for Hilbert spaces Z and $W = Z^*$. The local semismooth Newton method is equipped with an Armijo line search w.r.t. the merit function $\Theta(z) := \|f(z)\|_{Z^*}^2/2$. The proof is given under reasonable assumptions and the result is illustrated by applications in optimal control. The article is presented in the published version at the end of this chapter.

Josephy-Newton Method and SQP Method

We consider $f : Z \to \mathbb{R}$, Z a Hilbert space where we identify Z^* with Z. Another idea for solving a simply constrained problem as (2.22) is to consider instead of (2.25) the variational inequality

$$z \in Z_{ad}, \quad \nabla f(z)^*(\zeta - z) \geq 0 \quad \forall \zeta \in Z_{ad}$$

and then to linearize ∇f about the current iterate, cf. [HPUU09, 2.3.2 & 2.3.3]. This procedure is generalized to variational inequalities with a Fréchet-differentiable $F : Z \to Z$ of the general type

$$(F(z), \zeta - z)_Z \geq 0 \quad \forall \zeta \in Z_{ad}.$$

Linearization around the current iterate $z^{(k)} \in Z_{ad}$ yields another variational inequality

$$(F(z^{(k)}) + F'(z^{(k)})(z - z^{(k)}), \zeta - z)_Z \geq 0 \quad \forall \zeta \in Z_{ad}.$$

Its solution is the next iterate $z^{(k+1)} := z \in Z_{ad}$.

Algorithm 2.35 *(Josephy-Newton Method)*

1.) Choose a starting point $z^{(0)}$. Set $k := 0$.

2.) If $z^{(k+1)} = z^{(k)}$ within the numerical tolerance, then STOP.

3.) Compute the solution $z^{(k+1)}$ of

$$(F(z^{(k)}) + F'(z^{(k)})(z^{(k+1)} - z^{(k)}), \zeta - z^{(k+1)})_Z \geq 0 \quad \forall \zeta \in Z_{ad} \tag{2.30}$$

that is closest to $z^{(k)}$.

4.) Set $k := k + 1$ and return to 2.).

The variational inequality (2.30) is called *strongly regular*, if there exist open neighbourhoods $V_0 \subset Z$ of 0 and $V_{\hat{z}} \subset Z$ of $\hat{z}$ and a Lipschitz continuous function $V_0 \ni p \mapsto z(p) \in V_{\hat{z}}$ such that $z = z(p)$ is the unique solution in $V_{\hat{z}}$ of the parametrized variational inequality $(F(\hat{z}) + F'(\hat{z})(z - \hat{z}) - p, \zeta - z))_Z \geq 0$ for all $\zeta \in Z_{ad}$. It can be demonstrated that strong regularity implies local superlinear convergence.

In the case $Z_{ad} = Z$, the Josephy-Newton method turns out to be the Newton method and, moreover, then the invertibility of $F'(\hat{z})$ is equivalent to the notion of strong regularity.

In the case $F = \nabla f$, the variational inequality (2.30) is the first-order necessary optimality condition of the problem

$$\min_{z \in Z} \left\{ (\nabla f(z^{(k)}), z - z^{(k)})_Z + \frac{1}{2}(z - z^{(k)}, f''(z^{(k)})(z - z^{(k)}))_Z \right\} \text{ s.t. } z \in Z_{ad}. \tag{2.31}$$

We have a quadratic objective. In this case Algorithm 2.35 is called *sequential quadratic programming (SQP)* according to the definition in [HPUU09, Algorithm 2.7] or [Tr10, 4.11.2]. If the Hessian $f''(z^{(k)}) = \nabla^2 f(z^{(k)})$ is positive definite, the objective in the quadratic subproblem (2.31) is convex, and we have a unique minimizer.

We wish to solve Problem 2.3 in case that $G(z) \in \mathcal{K}$ simplifies to the inequality constraint $G(z) \leq 0_{W_G}$ and for Hilbert spaces Z, W. Considering $\xi := [z, \lambda]$, $F(\xi) = -\nabla_z L(\xi)$ for the Lagrangian $L(\xi)$ as in Def. (2.12) and linearizing around $\xi^{(k)} := [z^{(k)}, \lambda^{(k)}]$ yields the subproblem:

$$\begin{aligned} \nabla_z L(\xi^{(k)}) + L''_{z\xi}(\xi^{(k)})(\xi - \xi^{(k)}) &\geq 0, \\ \lambda_G &\geq 0_{W_G^*}, \\ (G(z^{(k)}) + G'(z^{(k)})(z - z^{(k)}), \zeta - \lambda_G)_Z &\leq 0 \qquad\qquad \forall \zeta \geq 0_{W_G^*}, \\ H(z^{(k)}) + H'(z^{(k)})(z - z^{(k)}) &= 0_{W_H}. \end{aligned}$$

Using that $(G(z^{(k)}) + G'(z^{(k)})(z - z^{(k)}), \zeta - \lambda_G)_Z \leq 0$ for all $\zeta \geq 0_{W_G^*}$ iff $(G(z^{(k)}) + G'(z^{(k)})(z - z^{(k)}), \lambda_G)_Z = 0$, we check that this is equivalent to the KKT-conditions (Th. 2.18) of the following quadratic program (with linearized constraints):

Find $z \in Z$ such that

$$(\nabla \mathcal{J}(z^{(k)}), z - z^{(k)})_Z + \frac{1}{2}(z - z^{(k)}, L''_{zz}(z^{(k)}, \lambda^{(k)})(z - z^{(k)}))_Z \tag{2.32}$$

is minimized,
where $z \in Z_{ad}$,
subject to the constraints

$$\begin{aligned} G(z^{(k)}) + G'(z^{(k)})(z - z^{(k)}) &\leq 0_{W_G}, \\ H(z^{(k)}) + H'(z^{(k)})(z - z^{(k)}) &= 0_{W_H}. \end{aligned}$$

Note that in literature the latter is also used for the definition of the SQP method.

If the Hessian $L''_{zz}(z^{(k)})$ is positive definite, then (2.32) is uniquely solvable. In order to proof local superlinear convergence, a certain regularity of $\mathcal{J}$ is required, see, e.g., [Al02]. For further details and references on sufficient conditions see [Tr10, Subsect. 4.11.2].

For the extension of the Lagrange-Newton method and the SQP method to Banach spaces see [Al90, Al91].

2.4 Optimal Control in Banach Spaces

Optimal control problems are a particular, but important case of optimization problems. In this situation, the problem has a certain structure and we distinguish between states y that solve a differential equation (or other (in)equality constraints) and control (design) variables u that influence the system in an optimal sense. The variable z to be optimized thus is split as $z = [y, u]$. Consequently, we consider the space $Z = Y \times U$.

2.4.1 Generic Optimal Control Problems and Existence of Optimal Controls

We reformulate Problem 2.3 as follows

Problem 2.36 *(Standard Optimal Control Problem)*
Let Y, U, W_G, and W_H be vector spaces (over $\mathbb{R}$) and let $\mathcal{J} : Y \times U \to \mathbb{R}$ be a functional. Furthermore, let $G : Y \to W_G$ and $H : Y \times U \to W_H$ be operators.

Find $[y, u] \in Y \times U$ such that $\mathcal{J}(y, u)$ is minimized,
where $y \in Y$, $u \in U_{ad} \subset U$,
subject to the constraints

$$\begin{aligned} G(y) &\in \mathcal{K}, \\ H(y, u) &= 0_{W_H}, \end{aligned}$$

where $\mathcal{K} \subset W_G$ is a closed convex cone.

We set

$$\begin{aligned} \mathcal{G} : \begin{bmatrix} y \\ u \end{bmatrix} \in Z = Y \times U \mapsto \begin{bmatrix} G(y) \\ H(y,u) \end{bmatrix} &\in W_G \times W_H, \\ \tilde{\mathcal{K}} &:= \mathcal{K} \times \{0\}, \\ Z_{ad} &:= Y \times U_{ad}. \end{aligned}$$

Thus the dual cone to $\tilde{\mathcal{K}}$ is $\tilde{\mathcal{K}}^+ = \mathcal{K}^+ \times W_H^*$.

Let Y_{ad} and U_{ad} denote the sets of admissible states and controls, respectively. Sometimes it is useful to put the inequality constraints into the subsets Y_{ad} and U_{ad} of spaces Y and U, respectively. The set of admissible states Y_{ad} represents the so-called *state constraints* and U_{ad} the *control constraints*. Note that in Problem 2.36, we have not restricted the states $y \in Y_{ad}$. Furthermore, there the inequality constraints depend only on the states and not on the control.

Hence, we may choose to consider only equality constraints. In this case $H(z) = 0_{W_H}$ (with $H : Z \to W_H, z \mapsto H(z)$) is replaced by the state equation $E(y, u) = 0_W$ with $E : Y \times U \to W, [y, u] \mapsto E(y, u)$. Then Problem 2.36 turns into the problem:

Problem 2.37 *(Standard Optimal Control Problem with Equality Constraints)*
Let Y, U, and W be vector spaces (over $\mathbb{R}$) and let $\mathcal{J} : Y \times U \to \mathbb{R}$ be a functional. Furthermore, let $E : Y \times U \to W$ be an operator.

Find $[y, u] \in Y \times U$ such that $\mathcal{J}(y, u)$ is minimized,
where $y \in Y_{ad} \subset Y$, $u \in U_{ad} \subset U$,
subject to the constraint

$$E(y, u) = 0_W.$$

Note that nonlinear problems are included in this setting.

The optimality from Def. 2.2 translates into

Definition 2.38 *(Optimal State and Optimal Control)*
Consider Problem 2.37 and let U be a Banach space. Assume there exists a unique state $y(u) \in Y_{ad}$ for any given control.

a) We call $\hat{u} \in U_{ad}$ a (locally) optimal control, if

$$\mathcal{J}(y(\hat{u}), \hat{u}) \leq \mathcal{J}(y(u), u) \quad \forall u \in U_{ad} \cap V(\hat{u})$$

in some neighbourhood $V(\hat{u})$ of $\hat{u}$.

b) If there exists a $V(\hat{u})$ s.t. $U_{ad} \cap V(\hat{u}) = U_{ad}$, then we have a globally optimal control.

c) $\hat{y} := y(\hat{u}) \in Y_{ad}$ is called the associated optimal state.

Note that we write also $\hat{y}(\hat{u})$ for the associated optimal state to underline that we follow a reduced approach.

For proving the existence of optimal controls, we follow mainly [HPUU09, Sect. 1.5.2] and we make the following Assumptions:

Assumption 2.39 *(Basic Assumptions on Spaces and Regularity for Existence of Optimal Controls)*

a) Y and U are reflexive Banach spaces and W a Banach space.

b) $U_{ad}(\subset U)$ are non-empty, convex, bounded, and closed.

c) The functional $\mathcal{J} : Y \times U \to \mathbb{R}$ is sequentially weakly lower semicontinuous and the operator $E : Y \times U \to W$ is weakly continuous, respectively.

Assumption 2.40 *(Further Assumptions for Existence of Optimal Controls)*

a) The non-empty, closed convex set Y_{ad} has a feasible point for Problem 2.37.

b) The equation $E(y, u) = 0_W$ has a bounded control-to-state operator (solution operator) $S : U_{ad} \to Y, u \mapsto S(u) := y(u)$.

Theorem 2.41 *(Existence of Optimal Controls (with Equality Constraints))*
Let Assumptions 2.39 and 2.40 hold, then the optimal control problem, Problem 2.37, has a solution $[\hat{y}, \hat{u}]$.

We compare with our Assumptions 2.6 and 2.7 considered for the necessary optimality conditions of a general optimization problem. Here we assume additionally that the Banach spaces Y and U are reflexive. Note that no F-differentiability of $\mathcal{J}$ and E is required here so far, but we need assumptions on the control-to-state operator. The existence of interior points is dropped, but the existence of a feasible point in Y_{ad} is required. This implies that the feasible set Σ_{fs} is non-empty.

For checking the Assumption on E in 2.39 c), it is useful to consider compact embeddings $Y \overset{c}{\hookrightarrow} \tilde{Y}$, that convert weak into strong convergence. We remark that this setting is typical for optimal control subject to PDE, see Section 2.7 for further examples and further details.

2.4.2 First-Order Necessary Optimality Conditions for Optimal Control

Here we turn back to Problem 2.36, where both inequality constraints $G(y) \in \mathcal{K}$ and equality constraints $H(y,u) = 0_{W_H}$ are present. We derive necessary optimality conditions for optimal control in a so-called *all-at-once approach* (or *full approach*), where we solve for states and controls simultaneously. We have introduced the Lagrange function in Definition 2.12. In the setting here, it reads

$$L(y,u,\lambda_G,\lambda_H) := \mathcal{J}(y,u) + \langle \lambda_G, G(y) \rangle_{W_G^*,W_G} + \langle \lambda_H, H(y,u) \rangle_{W_H^*,W_H}. \tag{2.33}$$

Assumption 2.42 *(Basic Assumptions on Spaces and Regularity for First-Order NOC in Optimal Control)*

a) Y, U, W_G, and W_H are Banach spaces.

b) $U_{ad} (\subset U)$ are non-empty, convex, and closed.

c) $\mathcal{J} : Y \times U \to \mathbb{R}$, $G : Y \to W_G$, and $H : Y \times U \to W_H$ are continuously F-differentiable.

This corresponds to Assumption 2.6, but we require the continuity of the F-derivatives of $\mathcal{J}$ and G in addition.

The necessary optimality conditions in optimal control (above see Th. 2.18 for the general case in optimization, below see Th. 2.73 for the special case of optimal control with PDE) in an all-at-once formulation read

Theorem 2.43 *(KKT-Conditions for Optimal Control Problem (with General Constraints))* *Assume $\hat{z}$ is a local minimizer of Problem 2.36. If Assumptions 2.42 and 2.7, translated to the case $Z = Y \times U$, hold, and if the Robinson CQ (Assumpt. A.40) is fulfilled at $[\hat{y}, \hat{u}]$, reading here*

$$\begin{bmatrix} 0_{W_G} \\ 0_{W_H} \end{bmatrix} \in int \left\{ \begin{bmatrix} G(\hat{y}) + G'(\hat{y})y - k \\ H'_y(\hat{y},\hat{u})y + H'_u(\hat{y},\hat{u})(u - \hat{u}) \end{bmatrix} \middle| \; [y,u] \in Y \times U_{ad}, k \in \mathcal{K} \right\},$$

then there exist Lagrange multipliers $\lambda = [\lambda_G, \lambda_H] \in W^* := W_G^* \times W_H^*$ such that λ fulfils the KKT-system

$$G(\hat{y}) \in \mathcal{K}, \tag{2.34}$$

$$\lambda_G \in \mathcal{K}^+, \tag{2.35}$$

$$\langle \lambda_G, G(\hat{y}) \rangle_{W_G^*, W_G} = 0, \tag{2.36}$$

$$H(\hat{y}, \hat{u}) = 0_{W_H}, \tag{2.37}$$

$$\hat{u} \in U_{ad}, \tag{2.38}$$

$$L'_y(\hat{y}, \hat{u}, \lambda) = 0_{Y^*}, \tag{2.39}$$

$$\langle L'_u(\hat{y}, \hat{u}, \lambda), d_u \rangle_{U^*, U} \geq 0 \quad \forall d_u \in U_{ad} - \{\hat{u}\}. \tag{2.40}$$

The equations (2.34) – (2.37) are the so-called *primal equations*, the next-to-last equation is the *adjoint equation* and the last equation is the *optimality (condition)*.

Proof. This follows by exploiting Theorem 2.18 for $z = [y, u]$ and the special case $G(y, u) = G(y)$ (using, furthermore, Eq. (2.33) for the Lagrangian). □

Objectives, where the control cost term

$$\frac{\alpha}{2} \|u\|_U^2 \tag{2.41}$$

is present, are easier to handle, e.g., yielding an explicit formula for the control as an projection of an adjoint, see (2.26). Thus sometimes this term is added artificially with a small regularization parameter $\alpha > 0$. This is called *Tikhonov regularization*. α is called *Tikhonov parameter*.

In case of box constraints for the control, the variational inequality may be expressed more concisely according to Lemma 2.27.

Lemma 2.44 *(Reformulation of Optimality Inequality by Projection onto Box Control Constraints)*
Let U be a Hilbert space with a partial order (e.g. $U = L^2(\Omega)$ for an open set $\Omega \subset \mathbb{R}^d$) and we identify $U^* = U$, such that $\nabla L(y, u, \lambda)$ is the Riesz representation of $L'(y, u, \lambda)$. Furthermore, we consider box constraints for the control, i.e.

$$U_{ad} = \{u \in U \,|\, u_{min} \leq u \leq u_{max}\} \text{ with } u_{min} \leq u_{max} \tag{2.42}$$

and we write for the pointwise Euclidean projection (as in (2.24)) onto the admissible controls

$$P_{U_{ad}}(u) = \max\{u_{min}, \min\{u, u_{max}\}\}.$$

Then (2.40) is equivalent to

$$\hat{u} = P_{U_{ad}}(\hat{u} - \tilde{\gamma} \nabla L(\hat{y}, \hat{u}, \lambda)) \quad \forall \tilde{\gamma} > 0. \tag{2.43}$$

Assume we have the structure $L(y, u, \lambda) = \alpha u + \tilde{B}(y, \lambda)$ with $\alpha > 0$, then setting $\tilde{\gamma} = 1/\alpha$ allows for the elimination of $\hat{u}$ in (2.43).

If the control does not enter into the objective, e.g. if the control cost term (2.41 is not present ($\alpha = 0$), an optimal control of so-called bang-bang type may be expected. By a bang-bang control we mean control functions that almost everywhere take only values on ∂U_{ad}.

Definition 2.45 *(Bang-Bang Property (for Linear-Quadratic Problems in Time and Space), cf. [KW13, Th. 2.1])*
Let $\hat{t}_f < \infty$ be the optimal final time, $\Omega_{t_f} := (0, t_f) \times \Omega$, $U = L^\infty(\Omega_{t_f})$, and $\hat{u}$ be the optimal distributed control for a linear PDE problem in time and space under box constraints for the control, i.e. $|u| \leq M$ for almost all $[t, x] \in \Omega_{t_f}$ and a linear-quadratic objective. For example, in [KW13] the linear heat equation is considered. The optimal control is said to be of bang-bang type, iff

$$|\hat{u}(t, x)| = M \quad \textit{for a.a. } [t, x] \in \Omega_{t_f}.$$

For bang-bang control for elliptic PDE see [MM00, MM01]. Bang-bang controls exist for optimal control problems with DAE as well, see, e.g., [Ge12, Sect. 7.1.1]. We remark that in the context of sparsity and L^1-terms in the objective, so-called zero-bang controls are observed.

Lemma 2.46 *(Reformulation of Optimality Inequality in Case of No Control Costs)*
Let the necessary optimality conditions from Th. 2.43 with equality constraints $E(y, u) = 0$ only hold. We consider box constraints for the control as in (2.42) where u_{min} and u_{max} are essentially bounded functions. If $\mathcal{J}'_u = 0_{U^*}$, then (2.40) reduces to

$$\langle E'_u(\hat{y}, \hat{u})^* \lambda, d_u \rangle_{U^*, U} \geq 0 \quad \forall d_u \in U_{ad} - \{\hat{u}\}.$$

For instance, we consider the distributed control of the quadratic tracking type problem subject to a Poisson PDE on Ω with homogeneous Dirichlet b.c. where $U = L^2(\Omega)$, $\Omega \subset \mathbb{R}^d$ a bounded Lipschitz domain, see [Tr10, Lemma 2.26] for further details. We assume that $E'_u(\hat{y}, \hat{u})^* \lambda$ is well-defined pointwise. Then the last inequality yields

$$\hat{u}(x) = \begin{cases} u_{min}(x) & \textit{if } E'_u(\hat{y}(x), \hat{u}(x))^* \lambda(x) > 0, \\ \in [u_{min}(x), u_{max}(x)] & \textit{if } E'_u(\hat{y}(x), \hat{u}(x))^* \lambda(x) = 0 \textit{ on some subset } I \subset \Omega \textit{ with } |I| > 0, \\ u_{max}(x) & \textit{if } E'_u(\hat{y}(x), \hat{u}(x))^* \lambda(x) < 0. \end{cases}$$

Thus if $|I| = 0$, then the optimal control is of bang-bang type. If $|I|$ we call u a bang-bang control with singular arc.

2.4.3 Reduced Optimal Control Problems

In the following we consider only equality constraints and write again E instead of H (cf. the lead text for Pb. 2.37). The basic assumptions, Assumption 2.39, simplify in this context.

Assumption 2.47 *(Basic Assumptions for Existence of a Control-to-State Operator)*

a) The equation $E(y, u) = 0_W$ has for each $u \in U$ a unique solution $y(u) \in Y$.

b) $E'_y(y(u),u) \in \mathcal{L}(Y,W)$ *is continuously invertible.*

The Assumption 2.40 b) is here replaced by the stronger Assumption 2.47 a), requiring uniqueness of the corresponding states y. This guarantees the existence and uniqueness of a bounded control-to-state operator $S : U \to Y, u \mapsto S(u) := y(u)$. By Assumption 2.47 b) the implicit function theorem (Th. A.30) guarantees that $y(u)$ is continuously F-differentiable w.r.t. u. $S'(u) = y'(u)$ may be obtained from differentiation of the state equation w.r.t. u,

$$E'_y(y(u),u)y'(u) + E'_u(y(u),u) = 0_{U^*}. \tag{2.44}$$

We see directly that a reduced problem is helpful in case of equality constraints only. Then the reduced approach is an alternative to the all-at-once approach (see Subsect. 2.4.2).

We insert $y(u)$ in Problem 2.37 and, writing $\tilde{\mathcal{J}}(u) := \mathcal{J}(y(u),u)$ for the reduced objective and $\tilde{U}_{ad} := \{u \in U \mid [y(u),u] \in Z_{ad} = Y_{ad} \times U_{ad}\}$, we obtain

Problem 2.48 *(Reduced Optimal Control Problem)*
Let U and W be vector spaces, and let $\tilde{\mathcal{J}} : U \to \mathbb{R}$ be a functional.

Find $u \in U$ such that $\tilde{\mathcal{J}}(u)$ is minimized,
where $u \in \tilde{U}_{ad} \subset U_{ad} \subset U$,
subject to the constraint

$$E(y(u),u) = 0_W.$$

For ease of presentation, we write again U_{ad} instead of $\tilde{U}_{ad}$ in the next section.

Example 2.49 *(Linear-Quadratic Optimal Control Problem)*
Let H, U be Hilbert spaces and Y, W be Banach spaces.

Find $[y,u] \in Y \times U$ such that the quadratic objective

$$\mathcal{J}(y,u) = \frac{1}{2}\|R_H y - y_{H,ref}\|_H^2 + \frac{\alpha}{2}\|u\|_U^2$$

is minimized,
where $u \in \tilde{U}_{ad} \subset U$,
subject to the constraint

$$Ay - Bu = f.$$

The objective consists of a tracking type term, where $y_{H,ref} := R_H y_{ref} \in H$ should be tracked, and of control costs, where $\alpha \geq 0$. We assume that the given right-hand side of the PDE has the regularity $f \in W$. For the operators we assume $A \in \mathcal{L}(Y,W)$, $B \in \mathcal{L}(U,W)$ and $R_H \in \mathcal{L}(Y,H)$. Typically, we have $Y \subset H$ and R_H embeds from the state space Y into the Hilbert space H. A Hilbert space is required for a quadratic objective term. We set

$$E(y,u) = Ay - Bu - f.$$

Then the Assumption 2.39 is satisfied, if $\tilde{U}_{ad}$ is convex, closed, and bounded. Furthermore, we assume that $E'_y(y,u) = A$ has a bounded inverse. In this case the control-to-state operator reads

$$\mathcal{S} : U \to Y, u \mapsto A^{-1}(Bu + f).$$

We find as reduced objective

$$\tilde{\mathcal{J}}(u) = \frac{1}{2}\|R_H A^{-1}(Bu + f) - y_{H,ref}\|_H^2 + \frac{\alpha}{2}\|u\|_U^2.$$

Note that the existence of optimal controls, provided by Theorem 2.41, follows here directly under Assumptions 2.39 and 2.47.

We formulate the necessary optimality conditions for optimal control of the reduced problem, Pb. 2.48.

Theorem 2.50 *(First-Order Necessary Optimality Conditions for Reduced Optimal Control Problem (with Control Constraints))*
Assume $\hat{u}$ is a local minimizer of Problem 2.48. We assume that only control constraints are implied, i.e. $Y_{ad} = Y$. Let Assumption 2.39 hold. Furthermore, let V be a neighbourhood of $\tilde{U}_{ad}$ and for each $u \in V$ instead of $u \in U$ let Assumption 2.47 hold. Then there holds

$$\hat{u} \in \tilde{U}_{ad},$$
$$\langle \tilde{\mathcal{J}}'(\hat{u}), d_u \rangle_{U^*,U} \geq 0 \quad \forall d_u \in \tilde{U}_{ad} - \{\hat{u}\}.$$

Proof. Assumption 2.47 implies that the control-to-state operator $S : V \to Y, u \mapsto y(u)$ is continuously F-differentiable by the implicit function theorem (Th. A.30). Then we apply Th. 2.9. □

We may use the adjoints in order to calculate explicitly $\tilde{\mathcal{J}}'$ (see [HPUU09, Sect. 1.6]). For any $\lambda \in W^*$, we have $\tilde{\mathcal{J}}(u) = L(y(u), u, \lambda)$ since the state equation holds according to Assumption 2.47 a). Differentiation w.r.t. u (in direction $d \in U$) yields

$$\langle \tilde{\mathcal{J}}'(u), d \rangle_{U^*,U} = \langle L'_y(y(u), u, \lambda), y'(u)d \rangle_{Y^*,Y} + \langle L'_u(y(u), u, \lambda), d \rangle_{U^*,U}.$$

We assume that $E'_y(y(u), u)^*$ is well-defined. $L'_y(y(u), u, \lambda) = 0$ is equivalent to the adjoint equation

$$E'_y(y(u), u)^*\lambda = -\mathcal{J}'_y(y(u), u), \tag{2.45}$$

that yields $\lambda(u)$ as solution of this linear system due to Assumpt. 2.47 b). Moreover, by means of (2.45) the term $y'(u)$ drops out (see the discussion in Subsect. 2.5.1).

Using the so-called adjoint representation for $\tilde{\mathcal{J}}'$,

$$\tilde{\mathcal{J}}'(u) = E'_u(y(u), u)^*\lambda(u) + \mathcal{J}'_u(y(u), u)$$

we may rewrite Th. 2.50:

Theorem 2.51 *(KKT-Conditions for Reduced Optimal Control Problem with Control Constraints: Variational Form)*
Assume $[\hat{y}(\hat{u}), \hat{u}]$ is a local minimizer of Problem 2.48. If Assumption 2.39 and Assumption 2.47 for all $u \in V$, V a neighbourhood of $\tilde{U}_{ad}$, are fulfilled, and only control constraints are implied, then there exist Lagrange multipliers $\lambda \in W^*$ such that λ fulfils the KKT-system

$$\langle w, E(\hat{y}(\hat{u}), \hat{u})\rangle_{W^*,W} = 0 \quad \forall w \in W^*, \tag{2.46}$$

$$\langle L'_y(\hat{y}(\hat{u}), \hat{u}, \lambda), v\rangle_{Y^*,Y} = 0 \quad \forall v \in Y, \tag{2.47}$$

$$\hat{u} \in \tilde{U}_{ad}, \tag{2.48}$$

$$\langle L'_u(\hat{y}(\hat{u}), \hat{u}, \lambda), d_u\rangle_{U^*,U} \geq 0 \quad \forall d_u \in \tilde{U}_{ad} - \{\hat{u}\}. \tag{2.49}$$

Note that (2.46) is trivial. However, we state it here in order to emphasize that a PDE is usually considered in variational (weak) form.

Proof. The statement follows from Theorem 2.50 and Remark 2.14 a). More precisely, the first equation of the KKT-system is the state equation being identical to $L'_\lambda(\hat{y}(\hat{u}), \hat{u}, \lambda) = E(\hat{y}(\hat{u}), \hat{u}) = 0_W$. The second equation defines the adjoint since $L'_y(\hat{y}(\hat{u}), \hat{u}, \lambda) = \mathcal{J}'_y(\hat{y}(\hat{u}), \hat{u}) + \langle \lambda, E'_y(\hat{y}(\hat{u}), \hat{u})\rangle_{W^*,W} = 0$ according to (2.45). Finally, this adjoint equation implies $\tilde{\mathcal{J}}'_u(u) = L'_u(y(u), u, \lambda(u))$, thus from Th. 2.50 follows the last equation of this theorem. □

In case of a reduced approach we may rewrite Lemma 2.44:

Lemma 2.52 *(Reformulation of Reduced Optimality Inequality by Projection onto Box Control Constraints)*
Let U be a Hilbert space with partial order and we identify $U^* = U$, such that $\nabla\tilde{\mathcal{J}}(u)$ is the Riesz representation of $\tilde{\mathcal{J}}'(u)$. Furthermore, we consider the box constraints (2.42) for the control.

Then (2.49) is equivalent to

$$\hat{u} = P_{\tilde{U}_{ad}}(\hat{u} - \tilde{\gamma}\nabla\tilde{\mathcal{J}}(\hat{u})) \quad \forall \tilde{\gamma} > 0. \tag{2.50}$$

For other equivalent reformulations of (2.49) in this context, see, e.g., [HPUU09, Lemma 1.12 (i) & (ii)].

Example 2.53 *(KKT-Conditions for Reduced Linear-Quadratic Optimal Control Problem)*
We continue with Example 2.49 and exploit the KKT-conditions. The Lagrange function reads

$$L(y, u, \lambda) = \frac{1}{2}\|R_H y - y_{H,ref}\|_H^2 + \frac{\alpha}{2}\|u\|_U^2 + \langle \lambda, Ay - Bu - f\rangle_{W^*,W}.$$

As Riesz representations we work with $H^* = H$ and $U^* = U$. Thus

$$\mathcal{J}'_y(y, u) = (R_H y - y_{H,ref}, R_H \cdot)_H = \langle R_H^*(R_H y - y_{H,ref}), \cdot\rangle_{Y^*,Y} = R_H^*(R_H y - y_{H,ref}),$$
$$\mathcal{J}'_u(y, u) = \alpha(u, \cdot)_U = \alpha u.$$

According to the latter theorem, Th. 2.51, there exists a Lagrange multiplier $\lambda \in W^*$ such that at the optimal solution $[\hat{y}(\hat{u}), \hat{u}] \in Y \times U$ the λ fulfils the KKT-system (in variational form)

$$\begin{aligned} \langle w, A\hat{y} - B\hat{u} - f\rangle_{W^*,W} &= 0 \quad \forall w \in W^*, \\ \langle A^*\lambda + R_H^*(R_H\hat{y} - y_{H,ref}), v\rangle_{Y^*,Y} &= 0 \quad \forall v \in Y, \\ \hat{u} &\in U_{ad}, \\ \langle \alpha\hat{u} - B^*\lambda, d_u\rangle_{U^*,U} &\geq 0 \quad \forall d_u \in \tilde{U}_{ad} - \{\hat{u}\}. \end{aligned}$$

The first equation is the state equation itself, the second is the adjoint state equation yielding λ, and the last equation is the optimality, that can be simplified, if $\alpha > 0$, by Lemma 2.44 as

$$\hat{u} = P_{\tilde{U}_{ad}}\left(\frac{1}{\alpha}B^*\lambda\right) \tag{2.51}$$

by choosing $\tilde{\gamma} = 1/\alpha$. This is an explicit formula linking the optimal control to the multiplier, thus it remains to solve the coupled state-multiplier system. In the case $\alpha = 0$, if the preliminaries of Lemma 2.46 (among other things $B^*\lambda$ is defined pointwise for $x \in \Omega$ and u_{min}, u_{max} are essentially bounded) hold, then this lemma yields

$$\hat{u}(x) = \begin{cases} u_{min}(x) & \text{if } (B^*\lambda)(x) < 0, \\ \in [u_{min}(x), u_{max}(x)] & \text{if } (B^*\lambda)(x) = 0 \text{ (could be a subset with measure zero only)}, \\ u_{max}(x) & \text{if } (B^*\lambda)(x) > 0. \end{cases}$$

In reduced form the adjoint equation reads

$$\langle A^*\lambda, v\rangle_{Y^*,Y} = -\langle R_H^*(R_HA^{-1}(B\hat{u} + f) - y_{H,ref}), v\rangle_{Y^*,Y} \quad \forall v \in Y,$$

or with $Q_H := R_HA^{-1}$

$$\lambda(u) = -Q_H^*(Q_H(B\hat{u} + f) - y_{H,ref}) \in Y^*.$$

Then the optimal control $\hat{u}$ can be determined by the optimality condition (2.49).

Note that due to the structure $L_u(y, u, \lambda) = \alpha u - B\lambda$ with $\alpha > 0$, the projection formula (2.51) transfers the possibly higher regularity of the adjoint (that may be proven for a suitable PDE) to the control.

Remark 2.54 *(Sign Convention)*
In principle, we could multiply an inequality or equality constraint by -1. If $E(y, u) = 0_W$ represents a differential equation, then our convention is to choose the sign of E just that the principal part of the differential operator has the "right", i.e. positive, sign. For instance, if we consider a parabolic PDE like the heat equation, then $E(y, u) = \partial_t y - \Delta_x y - f$, where f is the source term that might be subject to control. If we have the Poisson equation, then $E(y, u) = -\Delta_x y - f$ is the consistent choice of sign.

If we chose the sign differently (e.g., corresponding to [CRT18]), then the adjoint equation would have the other sign as well and, since we would encounter the other sign in (2.49) and *thus in the projection formula for the control as well, we obtain finally the same sign for the control as before.*

This can be illustrated easily by Example 2.53. The question of sign turns up in the all-at-once approach as well as in the reduced approach.

The reduced problem allows for a sensitivity-based approach and an adjoint-based approach for computing the derivative of $\tilde{\mathcal{J}}$, see Subsection 2.5.1.

2.5 Function Space Methods for Optimal Control

In this section, we assume that the state equation admits for every control $u \in U$ a unique solution $y(u)$ - this means that the differential equation (that might be a system as well) is well-posed. Moreover, let Assumpt. 2.47 hold. We consider the reduced optimal control problem, Pb. 2.48,

$$\min_{u \in U} \tilde{\mathcal{J}}(u), \tag{2.52}$$

where $\tilde{\mathcal{J}} : U \to \mathbb{R}$ is the reduced objective.

Assumption 2.55 *(Basic Assumption for Function Space Methods for Optimal Control)*
We assume that $\tilde{\mathcal{J}}$ are **twice** *continuously F-differentiable. Moreover, let the Assumptions of Th. 2.50 or Th. 2.51, resp., hold, i.e. the KKT-conditions are valid.*

2.5.1 Calculation of Derivatives of the Objective

The gradient method has been introduced in Subsection 2.3.1 as classical descent method. Actually, not the whole gradient but only the derivative in direction of the search direction is required. In optimal control there are two nearby approaches for computing the directional derivative of an objective.

Sensitivity Approach

For the sensitivity-based approach for computing the derivative of the reduced objective, shortly just *sensitivity approach*, we follow the presentation in [HPUU09, Subsect. 1.6.1]. The chain rule yields for the directional derivative of the reduced objective

$$\delta_{u;d}\tilde{\mathcal{J}}(u) := \langle \tilde{\mathcal{J}}'(u), d\rangle_{U^*,U} = \langle \mathcal{J}'_y(y(u),u), y'(u)d\rangle_{Y^*,Y} + \langle \mathcal{J}'_u(y(u),u), d\rangle_{U^*,U}, \tag{2.53}$$

where $d \in U$ is a direction. The directional derivatives $\delta_{u;d}y(u) = y'(u)d$ are called *sensitivities*. They can be computed by solving the linear equation

$$E'_y(y(u),u)\,\delta_{u;d}y(u) = -E'_u(y(u),u)d$$

that follows from (2.44).

This approach is computationally expensive if the whole $\tilde{\mathcal{J}}'(u)$ is required. All sensitivities $\delta_{u;b}y(u)$ have to be computed for all basis elements b in B, supposed B is a basis of U. Thus the effort is proportional to the dimension of U.

Adjoint Approach

The adjoint-based approach for computing the derivative of the reduced objective, shortly just *adjoint approach*, can be derived in a natural way from the Lagrange function, see the derivation of (2.45). Rewriting (2.53), we find

$$\tilde{\mathcal{J}}'(u) = y'(u)^* \mathcal{J}_y'(y(u), u) + \mathcal{J}_u'(y(u), u).$$

Thus not $y'(u)$ is actually required, but only $y'(u)^* \mathcal{J}_y'(y(u), u) \in U^*$. According to the adjoint equation (2.45), we finally have

$$\tilde{\mathcal{J}}'(u) = E_u'(y(u), u)^* \lambda(u) + \mathcal{J}_u'(y(u), u).$$

In general the adjoint approach is computationally cheaper as the sensitivity approach unless many constraints are applied.

Note that a sensitivity-based and an adjoint-based approach can be pursued for FDTO approaches as well [Ge12, Sect. 5.3]. In Fig. 2.1, for FDTO the all-at-once approach is called a *full discretization (collocation)*, whereas the reduced approach corresponds to a *direct shooting* method. For instance, the latter approach has been exploited in [GK15].

Second-Order Derivatives

For second-order derivatives of $\tilde{\mathcal{J}}$, assuming that $\mathcal{J}$ and E are twice continuously F-differentiable, we obtain

$$\tilde{\mathcal{J}}''(u) = T(u)^* L_{zz}''(y(u), u, \lambda(u)) T(u) = L_{zz}''(y(u), u, \lambda(u))[T(u), T(u)] \tag{2.54}$$

with

$$T(u) := \begin{bmatrix} y'(u) \\ Id_U \end{bmatrix} \in \mathcal{L}(U, Y \times U), \quad L_{zz}'' := \begin{bmatrix} L_{yy}'' & L_{yu}'' \\ L_{uy}'' & L_{uu}'' \end{bmatrix}.$$

Here it has been exploited that due to the definition of the adjoint the term $y''(u)$ drops out (see, e.g., [HPUU09, Subsect. 1.6.5] for details).

Example 2.56 *(Derivatives for Reduced Linear-Quadratic Optimal Control Problem)*
We revisit Example 2.53, abbreviating $Q_H = R_H A^{-1}$, $Q := Q_H B$, and obtain

$$\tilde{\mathcal{J}}'(u) = (\alpha Id_U + Q^* Q) u + Q^* (y_{H,ref} - Q_H f). \tag{2.55}$$

Otherwise, according to (2.54) we find

$$T(u) := \begin{bmatrix} A^{-1} B \\ Id_U \end{bmatrix}, \quad L_{zz}'' = \begin{bmatrix} R_H^* R_H & 0 \\ 0 & \alpha Id_U \end{bmatrix}$$

and thus, as we may check directly by (2.55),

$$\tilde{\mathcal{J}}''(u) = B^* A^{-*} R_H^* R_H A^{-1} B + \alpha Id_U = \alpha Id_U + Q^* Q.$$

2.5.2 Newton-Type Methods

We consider the iterative solution of the Newton equation

$$\tilde{\mathcal{J}}''(u^{(k)})s^{(k)} = -\tilde{\mathcal{J}}'(u^{(k)}).$$

Let s denote a step for an update, then not the whole Hessian $\tilde{\mathcal{J}}''(u)$, but only operator-vector products of the type $\tilde{\mathcal{J}}''(u)s$ are required. These can be computed efficiently as follows.

Algorithm 2.57 *(Efficient Computation of the Hessian of the Lagrange Function)*

1.) Solve $E'_y(y(u),u)(T(u)_1 s) = -E'_u(y(u),u)s$ for the sensitivity $T(u)_1 s = y'(u)s = \delta_{u;s}y(u)$.

2.) Compute $\xi = L''_{zz}(y(u),u,\lambda(u))(T(u)s)$.

3.) Solve $E'_y(y(u),u)^*\lambda(u) = -\xi_1$ for the adjoint $\lambda(u)$.

4.) Compute $\xi_3 = y'(u)^*\xi_1$ by $\xi_3 = E'_u(y(u),u)^*\lambda(u)$.

5.) Compute $\tilde{\mathcal{J}}''(u)s = \xi_2 + \xi_3$.

The first step requires one solve of the linearized state equation and the third step one solve of an adjoint equation.

Semismooth Newton Methods for Optimal Control

We apply the semismooth Newton method introduced in Subsection 2.3.3 to the KKT-conditions of the optimal control problem, Problem 2.36. In order to obtain the semismoothness of the KKT-system, the idea is to get a smoothing operator inside the projection. Therefore additional structure is required, either a so-called two-norm gap or a smoothing step as in [GH11] is required.

Here we follow the first approach. Let Ω be non-empty, open, bounded and $U = L^2(\Omega)$ is a Hilbert space that is identified with U^*. We consider box constraints $U_{ad} = \{u \in U \,|\, u_{min} \leq u(x) \leq u_{max} \text{ a.e. in } \Omega\}$ (with $u_{\min} < u_{max}$) for the control and we exploit (2.40) by introducing similar to (2.24) and (2.26) an operator $\Pi : Y \times U \times W \to U$ to be defined pointwise as

$$\Pi(y,u,\lambda)(x) := \pi(y(x),u(x),\lambda(x)) := u(x) - \tilde{P}_{U_{ad}}(u(x) - \tilde{\gamma}\nabla_u L(y(x),u(x),\lambda(x))). \tag{2.56}$$

The KKT-system from Th. 2.43 may be written in the following form

$$f(y,u,\lambda) = \begin{bmatrix} L'_y(y,u,\lambda) \\ \Pi(y,u,\lambda) \\ E(y,u) \end{bmatrix} = 0_{Y^* \times U \times W}.$$

Assumption 2.58 *(Assumption for Semismooth KKT-System [Ul01, Assumpt. 5.20])*

a) $E : Y \times U \to W$ and $\mathcal{J} : Y \times U \to \mathbb{R}$ are twice continuously F-differentiable.

b) L'_u *has the structure* $L'_u(y,u,\lambda) = \alpha u + \tilde{B}(y,u,\lambda)$ *and there exist* $\alpha > 0$ *and* $p > 2$ *such that*

(i) $\tilde{B} : Y \times U \times W^* \to U$ *is continuously F-differentiable and*

(ii) The operator $(y,u,\lambda) \in Y \times U \times W^* \mapsto \tilde{B}(y,u,\lambda) \in L^p(\Omega)$ *is locally Lipschitz-continuous.*

For example, the part b) of this assumption is fulfilled for linear-quadratic optimal control problems (Examples 2.49 and 2.53) for $W = H^1(\Omega)$, Ω open bounded, and $B = Id_U$. Then we have $\tilde{B}(y,u,\lambda) = -B^*\lambda$ and $\tilde{\gamma}$ is the reciprocal of the Tikhonov parameter α.

A generalized differential is not given naturally (see for instance the discussion following [GHK17, Def. 2.1]), we consider the set-valued mapping

$$\partial_C : Y \times U \times W^* \rightrightarrows \mathcal{L}(Y \times U \times, W^*, Y^* \times U \times W^{**})$$

with the differential

$$\begin{aligned} \partial_C f := \{ & M \in \mathcal{L}(Y \times U \times W^*, Y^* \times U \times W^{**}) : \\ & M(y,u,\lambda) = \begin{bmatrix} L''_{yy}(y,u,\lambda) & L''_{yu}(y,u,\lambda) & E'_y(y,u)^* \\ \tilde{\gamma} D \tilde{B}'_y(y,u,\lambda) & Id_U + \tilde{\gamma} D \tilde{B}'_u(y,u,\lambda) & \tilde{\gamma} D \tilde{B}'_\lambda(y,u,\lambda) \\ E'_y(y,u) & E'_u(y,u) & 0_{W^{**}} \end{bmatrix}, \\ & D \in L^\infty(\Omega), D(x) \in \partial_C P_{U_{ad}}(-\tilde{\gamma}\tilde{B}(y,u,\lambda)(x)) \, \forall x \in \Omega \Big\}. \end{aligned}$$

This differential is motivated by Qi's C-subdifferential in finite dimensions.

If Assumption 2.58 holds, then the projection $P_{U_{ad}} : L^p(\Omega) \to U_{ad} \subset L^2(\Omega)$, $p > 2$, maps between spaces with a norm gap. Then $P_{U_{ad}}$ and thus the superposition operator Π is $\partial_C P_{U_{ad}}$-semismooth and, furthermore, f is locally Lipschitz continuous and $\partial_C f$-semismooth [Ul01, Th. 5.21]. We assume to start with an initial control $u^{(0)} \in U_{ad}$. Thus the local semismooth Newton method, i.e. Algorithm 2.31 with $M^{(k)} \in \partial_C f(y^{(k)}, u^{(k)}, \lambda^{(k)})$, may be applied. For further details and results for this application see, e.g., [Ul11, GHK17]. Note that the latter reference is embedded at the end of this chapter.

The semismooth Newton method may be applied to a reduced optimal control problem as well. We mention that there might be numerical issues due to ill-conditioned Newton matrices.

SQP Method Applied to Optimal Control

In case of a reduced optimal control problem as Problem 2.48 the SQP method introduced in Subsection 2.3.3 can be applied as well. We consider the Banach space U, the admissible set of controls $U_{ad} \subset U$, $\tilde{\mathcal{J}}$ is the reduced objective, and here the Lagrangian reduces to $\tilde{\mathcal{J}}$. We choose $u^{(0)} \in U_{ad}$.

We distinguish between the classical SQP method as it is defined in literature (again commonly called the Newton method) and the SQP method following [Tr10, Subsect. 4.11.2] that we present here. We recall that $S : y \to u$ is the control-to-state operator corresponding to $E(y,u) = 0_W$.

In the classical approach for SQP, once $u^{(k+1)}$ is computed a new state $y^{(k+1)}$ is calculated as the solution $y^{(k+1)} = S(u^{(k+1)})$ of the possibly nonlinear state equation $E(y^{(k+1)}, u^{(k+1)}) = 0_W$, e.g., by the classical Newton method. This additional effort may be avoided by using the linearized state equation

$$y^{(k+1)} = y^{(k)} + S'(u^{(k)})(u^{(k+1)} - u^{(k)}).$$

Of course, in case of a linear state equation, both approaches coincide.

For concrete examples applying the SQP method to optimal control of PDE, see, e.g., [Tr10, Subsect. 4.11.2].

2.5.3 Optimization Methods for State Constraints

Up to this point we have considered methods mainly for optimization methods without constraints or with control constraints. The topic of Newton-type methods for state constrained problems is subject of ongoing research and recent advances have been made. However, we would like to discuss briefly the different approaches only, following [HPUU09, Sect. 2.7].

In SQP methods the state constraints enter in linearized form and then the issues due to the state constraints appear in the subproblems. However, both second-order optimality theory and proving fast local convergence of SQP methods are challenging in presence of state constraints. Recent results in this direction can be found in [CRT08, HMR10].

Semismooth Newton methods for optimal control problems with state constraints exhibit principal issues since these methods rely on pointwise formulations. In general, the multiplier associated to a state constraint is only a regular Borel measure, i.e. $\mu \in \mathcal{M}(\Omega)$, and the complementarity condition between the state and μ cannot be understood pointwise.

Another approach is the so-called *Lavrentiev regularization*, where the state constraint of the type

$$y \leq y_{max} \tag{2.57}$$

is replaced by

$$y + \varepsilon u \leq y_{max}$$

for a sufficiently small $\varepsilon > 0$, supposing $Y = U$. Introducing a new control $u_\varepsilon := y + \varepsilon u$, then $u_\varepsilon \leq y_{max}$ and, thus, a typical optimal control problem is transferred into a control-constrained optimal control problem as considered so far. However, after the regularization the control cost term reads

$$\frac{\alpha}{2\varepsilon^2} \|u_\varepsilon - y\|_U^2.$$

The hope is that under suitable assumptions, see [MPT07], the regularized solution converges strongly as $\varepsilon \downarrow 0$, requiring a suitable choice of the Tikhonov parameter α.

In the *Moreau-Yosida regularization* the state constraint (2.57) is incorporated as a penalty term, yielding the regularized version of an optimal control problem without inequality constraints:

Find $[y,u] \in Y \times U$ such that $\mathcal{J}(y,u) + \frac{1}{2\gamma}\| \max\{0, \sigma + \gamma(y - y_{max})\}\|^2_{L^2(\Omega)}$ is minimized,
where $y \in Y$, $u \in U$,
subject to the equality constraint

$$E(y,u) = 0_W,$$

with a penalty parameter $\gamma > 0$ and a shift parameter function $\sigma \in \{f \in L^2(\Omega) \,|\, f \geq 0_{L^2(\Omega)}\}$.

For this problem the KKT-conditions may be formulated as in Subsection 2.4.2, we may apply a semismooth Newton method and, finally, we let γ tend to infinity. This approach is analyzed, e.g, in [HK06a, HK06b].

Finally, we would like to mention that there exist interior point methods that are well adapted for optimization problems in Banach spaces as well, see, e.g., [SW08, UU09, WGS08].

2.6 Optimal Control of Ordinary Differential Equations and Differential Algebraic Equations

In this section we follow mainly the presentation as in [Ge05]. On optimal control of DAE see the books [Ge12, Wa72] for instance, but also the paper [ILWW18] that links the control of DAE to the control of ODE.

In order to distinguish between ODE and PDE in the next chapter on coupled ODE-PDE systems, we write q for ODE states and y for the (PDE) states in the following. In DAE the state may be split into differential states q_1, subject to an differential equation, and algebraic states q_2, subject to algebraic equations.

2.6.1 Optimal Control of an Index-1 DAE

The general form of a typical optimal control problem for a DAE is

Problem 2.59 *(DAE Optimal Control Problem in Standard Form)*
Let the time interval $I = [t_0, t_f] \subset \mathbb{R}$ be non-empty with a fixed $t_f < \infty$.

Find $[q_1, q_2, u]$ such that the objective

$$\mathcal{J}(q_1, q_2, u) = \Phi(q_1(t_0), q_1(t_f)) + \int_{t_0}^{t_f} \phi(t, q_1(t), q_2(t), u(t))\, dt$$

is minimized,
where $[q_1, q_2, u] \in Z := Y_1 \times Y_2 \times U := [W^{1,\infty}]^{n_{q,1}} \times [L^\infty]^{n_{q_2}} \times [L^\infty]^{n_u}$

subject to

$$\begin{aligned}\dot{q}_1(t) - f_1(t, q_1, q_2, u) &= 0_{W_{H_1}}, && (2.58)\\ f_2(t, q_1, q_2, u) &= 0_{W_{H_2}}, && (2.59)\\ \Psi(q_1(t_0), q_1(t_f)) &= 0_{W_{H_3}}, && (2.60)\\ G(t, q_1, q_2, u) &\leq 0_{W_G}, && (2.61)\\ u(t) &\in \tilde{U}, \qquad \textit{a.e. in } I && (2.62)\end{aligned}$$

for suitable Banach spaces $W_H = W_{H_1} \times W_{H_2} \times W_{H_3}$ and W_G, in which the equations are considered, to be specified. As common we set for the equality constraints

$$H(t, q_1, q_2, u) = \begin{bmatrix} f_1(t, q_1, q_2, u) - \dot{q}_1 \\ f_2(t, q_1, q_2, u) \\ -\Psi(q_1(t_0), q_1(t_f)) \end{bmatrix}.$$

For the admissible controls we consider the set $U = \{u \in [L^\infty(I)]^{n_u} \mid u(t) \in \tilde{U}$ for a.a. $t \in I\}$ for $\tilde{U} \subset \mathbb{R}^{n_u}$.

Here we encounter a semi-explicit DAE with the general constraint (2.61) that may be mixed control-state constraints or include pure state constraints also.

In principle we could introduce parameters p to be identified in the latter problem. However, p could be substituted by a further ODE state, then the solution for p appears as a free initial condition.

We work with the functions $\Phi : \mathbb{R}^{n_{q_1}} \times \mathbb{R}^{n_{q_1}} \to \mathbb{R}$, $\phi : I \times \mathbb{R}^{n_{q_1}} \times \mathbb{R}^{n_{q_2}} \times \mathbb{R}^{n_u} \to \mathbb{R}$, $f_i : I \times \mathbb{R}^{n_{q_1}} \times \mathbb{R}^{n_{q_2}} \times \mathbb{R}^{n_u} \to \mathbb{R}^{n_{q_1}}$, $i = 1, 2$, $\Psi : \mathbb{R}^{n_{q_1}} \times \mathbb{R}^{n_{q_1}} \to \mathbb{R}^{n_\Psi}$, and $G : I \times \mathbb{R}^{n_{q_1}} \times \mathbb{R}^{n_{q_2}} \times \mathbb{R}^{n_u} \to \mathbb{R}^{n_m}$.

Moreover, we drop the mixed control-state constraints and the pure state constraints in the following and set

$$\begin{aligned} Z &= [W^{1,\infty}(I)]^{n_{q_1}} \times [L^\infty(I)]^{n_{q_2}} \times [L^\infty(I)]^{n_u},\\ W = W_H &= [L^\infty(I)]^{n_{q_1}} \times [L^\infty(I)]^{n_{q_2}} \times \mathbb{R}^{n_\Psi},\\ U &= [L^\infty(I)]^{n_u},\\ U_{ad} &= \{u \in L^\infty(I)]^{n_u} \mid u_{min,i}(t) \leq u_i(t) \leq u_{max,i}(t), i = 1, \dots, n_u, \text{ a.e. in } I\},\\ Z_{ad} &= [W^{1,\infty}(I)]^{n_{q_1}} \times [L^\infty(I)]^{n_{q_2}} \times U_{ad}. \end{aligned}$$

We presume the involved functions to be sufficiently smooth:

Assumption 2.60 *(Smoothness Assumptions for DAE Problems [Ge12, Assumpt. 2.2.8])*

a) Φ and Ψ are continuously differentiable w.r.t. all arguments.

b) For a sufficiently large convex compact neighbourhood V of $\hat{z} := [\hat{q}_1, \hat{q}_2, \hat{u}] \in [W^{1,\infty}(I)]^{n_{q_1}} \times [L^\infty(I)]^{n_{q_2}} \times [L^\infty(I)]^{n_u}$

   (i) The maps ϕ, f_1, f_2, and G, are measurable (as maps) in t for all $[q_1, q_2, u] \in V$ and continuously differentiable in $[q_1, q_2, u]$ uniformly for all $t \in I$, and

(ii) The derivatives of ϕ, f_1, f_2, and G w.r.t. q_1, q_2, and u are bounded in $I \times V$.

Note that no explicit continuity w.r.t. t is required, only measurability. The latter assumption implies the F-differentiability of $\mathcal{J}$, G, and H.

Since this study is on CDE, we discuss mainly semi-explicit index-1 DAE systems that occur often, e.g., in vehicle simulation [Ge03] or process engineering [Hi97]. Note that multibody systems as in Example C.5 could be considered as DAE of index 3, but often simpler models are available, see e.g. Example 1.1 that may be reformulated as a pure ODE system. Index-1 DAE are characterized by the following postulation.

Assumption 2.61 *(Index-1 DAE)*
The matrix

$$M(t) := f'_{2;q_2}(t, \hat{q}_1(t), \hat{q}_2(t), \hat{u}(t)) \tag{2.63}$$

is non singular almost everywhere in I and
$M^{-1}(t)$ is essentially bounded in I.
This is equivalent to Pb. 2.59 exhibiting a index-1 DAE.

Note that this assumption allows to solve the algebraic equation for q_2, obtaining an expression for the algebraic variable q_2 as a function of t, q_1, and u. For the definition of the (perturbation) index of a DAE and a discussion on other definitions of indices of a DAE see [Ge12, p. 24].

For applying Th. 2.8, we write $z = [q_1, q_2, u]$ and the local minimizer is $\hat{z} = [\hat{q}_1, \hat{q}_2, u]$. Using Assumption 2.60, Assumption 2.6 can be verified directly. Here Assumption 2.7 b) is not applicable. We use Theorem 2.17 for verifying Assumption 2.7 c): we prove that T_1 defined by

$$T_1(\hat{z})(z) := \begin{bmatrix} H'_1(\hat{z})(z) \\ H'_2(\hat{z})(z) \end{bmatrix} = \begin{bmatrix} f'_{1;q_1}(t,\hat{q},\hat{u})q_1 + f'_{1;q_2}(t,\hat{q},\hat{u})q_2 + f'_{1;u}(t,\hat{q},\hat{u})u - \dot{q}_1 \\ f'_{2;q_1}(t,\hat{q},\hat{u})q_1 + f'_{2;q_2}(t,\hat{q},\hat{u})q_2 + f'_{2;u}(t,\hat{q},\hat{u})u \end{bmatrix}$$

is surjective, see, e.g., [Ge12, Lemma 3.1.4] under Assumption (2.61). Note that, if we considered a space of continuous functions as target for H_2, then the image of H'_2 is no proper dense subset of this space and we could not prove the non-density by this means.

Due to the Fritz John-conditions, Th. 2.8, we have demonstrated

Theorem 2.62 *(NOC of Fritz John-type for Index-1 DAE Optimal Control Problem (with Pure Equality Constraints))*
Assume $\hat{z} = [\hat{q}, \hat{u}]$ is a local minimizer of Problem 2.59 that we consider without (2.61). If the Assumptions 2.7 a) (reading here $u_{min,i} < u_{max,i}$ for all $i = 1, \ldots, n_u$), 2.60, and 2.61 hold, then there exist multipliers $\lambda := [\lambda_0, \lambda_H] = [\lambda_0, \lambda_1, \lambda_2, \lambda_\Psi] \in \mathbb{R}^+_0 \times W^*$, $\lambda \neq [0, 0_W]$, s.t.

$$\lambda_0 \geq 0, \tag{2.64}$$

$$H(\hat{z}) = 0_W, \tag{2.65}$$

$$\langle \lambda_0 \mathcal{J}'(\hat{z}), d \rangle_{Z^*,Z} + \langle \lambda_H, H'(\hat{z})d \rangle_{W^*,W} \geq 0 \qquad \forall d := [q_1, q_2, u] \in Z_{ad} - \{\hat{z}\}. \tag{2.66}$$

The multiplier $\lambda_H = [\lambda_1, \lambda_2, \lambda_\Psi]$ belongs to a dual of state spaces, namely $W^* = [L^\infty(I)^*]^{n_{q_1}} \times [L^\infty(I)^*]^{n_{q_2}} \times \mathbb{R}^{n_\Psi}$, where the first two subspaces do not allow for suitable representations. However, by exploiting the variational inequality (2.66) and using solution formulae that are available for DAE of index 1, we may prove a certain representation of the multiplier [Ge05].

Lemma 2.63 *(Representations of Adjoints)*
Let Assumptions 2.7 a), 2.60, and 2.61 hold, then there exist $\Lambda_1 \in W_1 = [W^{1,\infty}(I)]^{n_{q_1}}$ and $\Lambda_2 \in W_2 = [L^\infty(I)]^{n_{q_2}}$, s.t.

$$\langle \lambda_1, \phi_1 \rangle_{W_1^*, W_1} = -\int_{t_0}^{t_f} \Lambda_1(t)^\top \phi_1(t)\, dt \qquad \forall \phi_1 \in [L^\infty(I)]^{n_{q_1}},$$
$$\langle \lambda_2, \phi_2 \rangle_{W_2^*, W_2} = -\int_{t_0}^{t_f} \Lambda_2(t)^\top \phi_2(t)\, dt \qquad \forall \phi_2 \in [L^\infty(I)]^{n_{q_2}},$$

for λ_1, λ_2 as given by Th. 2.62.

This shows that our multipliers exhibit actually more regularity than being elements of the dual space only.

2.6.2 Minimum Principles

Necessary optimality conditions for optimal control of ODE are known as *minimum principles* or *maximum principles*. The first proofs of these go back to Pontryagin *et al.* [PBGM64] and Hestenes [He66].

We consider *local minimum principles*, i.e. the NOC are here interpreted w.r.t. a local minimum of the Hamilton function By *global minimum principles* we mean that the NOC w.r.t. a global minimum of the Hamilton function are considered. If a Hamilton functions was to be maximized, the obtained NOC would be called a maximum principle.

The importance of local minimum principles is due to the fact, that the discrete adjoints corresponding to the discretized local minimum principle approximate the adjoints of the NOC [Ge12, Sect. 5.4]. This relation does not hold for global minimum principles, unless strong conditions are additionally satisfied.

Definition 2.64 *(Hamilton Function for DAE Optimal Control)*
The function

$$\mathcal{H}(t, q_1, q_2, u, \lambda_0, \lambda_1, \lambda_2) := \lambda_0 \phi(t, q_1, q_2, u) + \lambda_1 f_1(t, q_1, q_2, u) + \lambda_2 f_2(t, q_1, q_2, u)$$

is called the Hamilton function (Hamiltonian) corresponding to Problem 2.59 without (2.61), i.e. the optimal control problem for semi-explicit index-1 DAE, in which mixed-control state constraints and pure state constraints are excluded.

From Theorem 2.62 we may prove

Theorem 2.65 *(Local Minimum Principle for DAE Optimal Control)*
Assume $\hat{z} = [\hat{q}, \hat{u}]$ is a local minimizer of Problem 2.59. Let the Assumptions 2.7 a), 2.60, and 2.61 hold. Then there exist multipliers $\lambda := [\lambda_0, \lambda_1, \lambda_2, \sigma] \in \mathbb{R}_0^+ \times W_1^* \times W_2^* \times \mathbb{R}^{n_\Psi}$, $\lambda \neq [0, 0_{W_1^*}, 0_{W_2^*}, 0_{\mathbb{R}^{n_\Psi}}]$, s.t.

a)

$$\lambda_0 \geq 0,$$

b) (Adjoint differential equations)

$$\dot{\lambda}_1(t) = -\mathcal{H}'_{q_1}(t, \hat{q}_1(t), \hat{q}_2(t), \hat{u}(t), \lambda_0, \lambda_1(t), \lambda_2(t)) \qquad \text{a.e. in } I, \tag{2.67}$$

$$0 = \mathcal{H}'_{q_2}(t, \hat{q}_1(t), \hat{q}_2(t), \hat{u}(t), \lambda_0, \lambda_1(t), \lambda_2(t)) \qquad \text{a.e. in } I, \tag{2.68}$$

c) (Transversality conditions)

$$\lambda_f(t_0) = -(\lambda_0 \Phi'_{q_1(t_0)}(\hat{q}_1(t_0), \hat{q}_1(t_f)) + \sigma^\top \Psi'_{q_1(t_0)}(\hat{q}_1(t_0), \hat{q}_1(t_f))),$$
$$\lambda_f(t_f) = \lambda_0 \Phi'_{q_1(t_f)}(\hat{q}_1(t_0), \hat{q}_1(t_f)) + \sigma^\top \Psi'_{q_1(t_f)}(\hat{q}_1(t_0), \hat{q}_1(t_f)),$$

d) (Stationarity of the Hamilton function)

$$\langle \mathcal{H}'_u(t, \hat{q}_1(t), \hat{q}_2(t), \hat{u}(t), \lambda_0, \lambda_1(t), \lambda_2(t)), u - \hat{u}(t) \rangle_{U^*,U} \geq 0 \quad \text{a.e. in } I \; \forall u \in \tilde{U}.$$

(2.67) & 2.68) are an index-1 DAE for the differential variable λ_1 and the algebraic variable λ_2.

In order to guarantee that $\lambda_0 \neq 0$, a constraint qualification is required. We translate RCQ, see Appendix A.2.2 for details, into the situation of Problem 2.59 without inequality constraints. Note that here RCQ is equivalent to the Mangasarian-Fromowitz condition (MFCQ).

The surjectivity of $H'(\hat{q}_1, \hat{q}_2, \hat{u})$ follows by Assumption 2.61 and by

$$rank(\Psi'_{q_1(t_0)} F(t_0) + \Psi'_{q_1(t_f)} F(t_f)) = n_\Psi, \tag{2.69}$$

where F denotes the fundamental solution of the ODE $\dot{F} = \tilde{A} F$, $F(t_0) = Id_{n_{q_1}}$ in I, here $\tilde{A}(t) = f'_{1;q_1}(t, \hat{q}(t), \hat{u}(t)) - f'_{1;q_2}(t, \hat{q}(t), \hat{u}(t)) M(t)^{-1} f'_{2;q_1}(t, \hat{q}(t), \hat{u}(t))$, see [Ge05, Lemma 6.1]. The two latter assumptions yield what is called *complete controllability of linearized dynamics* within control theory.

Theorem 2.66 *(Regularity of DAE Optimal Control)*
Let Assumptions 2.7 a), 2.60, 2.61 and Eq. (2.69) hold. If there exist $\tilde{q}_1 \in [W^{1,\infty}(I)]^{n_{q_1}}$, $\tilde{q}_2 \in [L^\infty]^{n_{q_2}}$, and $\tilde{d}_u \in int(U_{ad} - \hat{u})$ fulfilling

$$\dot{\tilde{q}}_1 = f'_{1;q_1}(t, \hat{q}, \hat{u}) \tilde{q}_1 + f'_{1;q_2}(t, \hat{q}, \hat{u}) \tilde{q}_2 + f'_{1;u}(t, \hat{q}, \hat{u}) \tilde{d}_u \qquad \text{a.e. in } I, \tag{2.70}$$

$$0_{\mathbb{R}^{n_{q_2}}} = f'_{2;q_1}(t, \hat{q}, \hat{u}) \tilde{q}_1 + f'_{2;q_2}(t, \hat{q}, \hat{u}) \tilde{q}_2 + f'_{2;u}(t, \hat{q}, \hat{u}) \tilde{d}_u \qquad \text{a.e. in } I, \tag{2.71}$$

$$0_{\mathbb{R}^{n_{q_1}}} = \Psi'_{q_1(t_0)} \tilde{q}_1(t_0) + \Psi'_{q_1(t_f)} \tilde{q}_1(t_f), \tag{2.72}$$

then Theorem 2.62 and Theorem 2.65 hold with $\lambda_0 = 1$.

Proof. The surjectivity yields Assumption 2.7 c). Under the above assumptions, (2.69) and (2.70) – (2.72) yield MFCQ that is equivalent to RCQ (as required in Th. 2.18)), since no cone constraints are present. □

The numerical methods discussed in Sect. 2.3 and Sect. 2.5 could be applied to the particular case of optimal control with DAE as well unless the methods are restricted to Hilbert spaces. The natural function spaces for ODE and DAE, $W^{1,\infty}$ and L^∞, are no Hilbert spaces.

The gradient method w.r.t. typical spaces for DAE is presented in [Ge12, Sect. 8.1] for optimal control of DAE with index 1. For details of the Lagrange-Newton method applied to typical spaces for DAE see [Ge12, Sect. 8.2].

2.7 Optimal Control of Partial Differential Equations

In this section we reconsider the results from Section 2.4, when the constraints are partial differential equations. In principle, variational inequalities could be considered as constraints as well, see, e.g., [BZ99]. In the following we present only the case of a PDE $E(y, u) = 0_W$ as a special case of an equality constraint $H(y, u) = 0_{W_H}$, whereas we do not consider inequality constraints like $G(y) \leq 0_{W_G}$ and thus no variational inequalities here. We follow an abstract approach as pursued in [IK08, Sect. 1.5].

2.7.1 Optimal Control of PDE without Inequality Constraints

Problem 2.67 *(Optimal Control Problem with PDE)*
Let Y, U, and W be vector spaces, $\tilde{Y} \subset Y$, and let $\mathcal{J} : Y \times U \to \mathbb{R}$ be a functional. Furthermore, let $E : \tilde{Y} \times U \to W$ be an operator, typically representing a PDE.

Find $[y, u] \in Y \times U$ such that $\mathcal{J}(y, u)$ is minimized,
where $y \in Y$, $u \in U_{ad} \subset U$,
subject to the constraints

$$E(y, u) = 0_W. \tag{2.73}$$

Usually, it is helpful to consider the partial differential equation (system) $E(y, u) = 0_W$ in an abstract operator formulation. The latter being typically the PDE in weak formulation (variational formulation), since this requires least regularity of the states, plus, e.g., initial conditions, if applicable. If the PDE is considered in the strong formulation, the corresponding spaces have to be chosen suitably, e.g. $W = H^1(\Omega)$ and $W^* = H^1(\Omega)^*$ for the Laplace/Poisson problem.

Example 2.68 *(Formulations of an Elliptic Second-Order PDE Problem)*
a) Let $\Omega \in \mathbb{R}^d$, $d \geq 1$, be open bounded with Lipschitz boundary[7]*. The boundary part with*

[7] Note that in 1D (i.e. $d = 1$) a single point is a $C^{k-1,1}$, $k \geq 1$, (Lipschitz) boundary by definition in our study. This is different to [Tr10], but simplifies the formulation. Anyways in 1D an open bounded interval Ω is assumed.

Neumann boundary conditions is $\Gamma_N \subset \partial\Omega$ and the boundary part with Dirichlet boundary conditions is $\Gamma_D := \partial\Omega \setminus \overline{\Gamma}_N$.

The classical formulation is: for given $f \in C^0(\overline{\Omega})$ and $g \in C^0(\overline{\Gamma}_N)$, we wish to find a $y \in C_{ell}(\Omega, \Gamma_N, \Gamma_D) := C^2(\Omega) \cap C^1(\Omega \cup \Gamma_N) \cap C^0(\overline{\Omega})$ s.t.

$$\begin{aligned} L_0 y &= f, && \text{in } \Omega, \\ -A\partial_\nu y + c_\Gamma(x,y) + d_\Gamma &= g, && \text{on } \Gamma_N, \\ y &= 0, && \text{on } \Gamma_D, \end{aligned}$$

with a linear second-order partial differential operator (in divergence form)

$$L_0 y := -Ay + \sum_{i=1}^{d} b_i(x) y'_{x_i} + c_0(x,y) + d_0, \tag{2.74}$$

where

$$Ay := \sum_{i,j=1}^{d} (a_{ij}(x) y'_{x_i})'_{x_j}. \tag{2.75}$$

b) For brevity we consider here the case $\Gamma = \Gamma_D$ and $c_0(x,y) = c_0(x)y$. The corresponding weak formulation of this initial-boundary value problem reads

$$a(y,v) = \langle f, v \rangle_{H^{-1},H^1_0} \quad \forall v \in H^1_0(\Omega), \tag{2.76}$$

where the bilinear form associated to (2.74) is

$$\begin{aligned} &a : H^1_0(\Omega) \times H^1_0(\Omega) \to \mathbb{R}, \\ &[y,v] \mapsto a(y,v) := \int_\Omega \sum_{i,j=1}^{d} (a_{ij}(x) y'_{x_i}) v'_{x_j} + \sum_{i=1}^{d} b_i(x) y'_{x_i} v + (c_0(x) + d_0) yv \, dx. \end{aligned}$$

For a bounded and coercive bilinear form a and $f \in H^{-1}(\Omega)$, this weak formulation admits a unique solution $y \in H^1_0(\Omega)$. Furthermore, $H^1_0(\Omega)$ embeds into $H^{1/2}(\Gamma)$, thus prescribing a Dirichlet boundary condition makes here indeed sense. Please see Subsection A.1.3 in the appendix for details.

Let the coefficients be bounded, i.e. a_{ij}, b_i, c_0, $d_0 \in L^\infty(\Omega)$. The weak formulation is equivalent to the operator formulation, i.e., a bounded linear operator

$$\begin{aligned} E : H^1_0(\Omega) &\to H^{-1}(\Omega), \\ y &\mapsto L_0 y \end{aligned}$$

is defined in the sense that

$$Ey = f \quad \Longleftrightarrow \quad (2.78) \text{ holds.}$$

The existence of a unique solution of the weak formulation yields that the operator E ("$= L_0$ & hom. Dirichlet b.c.") has a bounded inverse.

Example 2.69 *(Formulations of a Parabolic Second-Order PDE Problem)*
a) Let $\Omega \in \mathbb{R}^d$, $d \geq 1$, be open bounded with Lipschitz boundary[8] *and let $I = (0, t_f)$ with $t_f > 0$ be a fixed time interval. The time-space cylinder is denoted by $\Omega_{t_f} = (0, t_f) \times \Omega$ and its spatial boundary by $\Sigma_{t_f} = (0, t_f) \times \partial\Omega$. Let the boundary part with Neumann boundary conditions be $\Sigma_N \subset \Sigma_{t_f}$ and the boundary part with Dirichlet boundary conditions $\Sigma_D := \Sigma_{t_f} \setminus \overline{\Sigma}_N$. We consider on Ω_{t_f} the general case of a semilinear parabolic (second-order) PDE.*

The classical formulation is: for given $f \in C^0(\overline{\Omega}_{t_f})$, $g \in C^0(\overline{\Sigma}_N)$, and $y_0 \in C^0(\overline{\Omega})$, we wish to find a $y \in C_{par}(I, \Omega, \Gamma_N, \Gamma_D) := C^1(I; C_{ell}(\Omega, \Gamma_N, \Gamma_D)) \cap C^0(\overline{I}; C_{ell}(\Omega, \Gamma_N, \Gamma_D))$ s.t.

$$\begin{aligned} y_t' + L_1 y &= f, && \text{in } \Omega_{t_f}, \\ -A\partial_\nu y + c_\Sigma(t, x, y) &= g, && \text{on } \Sigma_N \subset \Sigma_{t_f} \\ y &= 0, && \text{on } \Sigma_D := \Sigma_{t_f} \setminus \overline{\Sigma}_N, \\ y(0, \cdot) &= y_0 && \text{on } \Omega, \end{aligned}$$

with a second-order partial differential operator (in divergence form)

$$L_1 y := -Ay + \sum_{i=1}^{d} b_i(t, x) y'_{x_i} + c_0(t, x, y) \tag{2.77}$$

with A as defined in (2.75) where the a_{ij} may depend additionally on time, too.

b) Here we consider $\Sigma_{t_f} = \Sigma_D$ and the case of a linear PDE by assuming $c_0(t, x, y) = c_0(t, x)y$. The corresponding weak formulation of this initial-boundary value problem reads

$$\langle y_t'(t), v\rangle_{H^{-1}, H_0^1} + a(y(t), v; t) = \langle f(t), v\rangle_{H^{-1}, H_0^1} \quad \forall v \in H_0^1(\Omega)\ \forall t \in \overline{I}, \tag{2.78}$$

$$y(0, \cdot) = y_0 \quad \text{on } \Omega, \tag{2.79}$$

where the bilinear form associated to (2.77) is

$$\begin{aligned} &a : H_0^1(\Omega) \times H_0^1(\Omega) \times \overline{I} \to \mathbb{R}, \\ &[y, v, t] \mapsto a(y, v, t) := \int_\Omega \sum_{i,j=1}^{d} (a_{ij}(t, x) y'_{x_i}) v'_{x_j} + \sum_{i=1}^{d} b_i(t, x) y'_{x_i} v + c_0(t, x) y v \, dx. \end{aligned}$$

For a bilinear form a, being bounded and coercive for almost all t, $f \in L^2(0, t_f; H^{-1}(\Omega))$, and $y_0 \in L^2(\Omega)$, this weak formulation admits a unique solution $y \in W(I; L^2, H_0^1)$, where $W(I; L^2, H_0^1)$ is defined as in Def. A.17. Furthermore, $W(; L^2, H_0^1)$ embeds compactly into $C^0(\overline{I}; L^2(\Omega))$, thus prescribing an initial condition makes here sense.

Let the coefficients be bounded, i.e. $a_{ij}, b_i, c_0 \in L^\infty(\Omega_{t_f})$.The weak formulation is equivalent to the operator formulation, i.e., a bounded linear operator

$$\begin{aligned} E : W(I; L^2, H_0^1) &\to L^2(I; H^{-1}(\Omega)) \times L^2(\Omega), \\ y &\mapsto \begin{bmatrix} y_t' + L_1 y \\ y(0, \cdot) \end{bmatrix} \end{aligned}$$

[8] For $d = 1$ see the footnote in Ex. 2.68 a).

is defined in the sense that

$$Ey = \begin{bmatrix} f \\ y_0 \end{bmatrix} \quad \Longleftrightarrow \quad (2.78) \wedge (2.79) \textit{ holds.}$$

The existence of a unique solution of the weak formulation yields that the operator E ("$= \partial_t + L_1$ & hom. Dirichlet b.c. & initial condition") has a bounded inverse.

The Lagrange function, defined in Def. 2.12, corresponding to Problem 2.67, reads

$$L(y, u, \lambda) := \mathcal{J}(y, u) + \langle \lambda, E(y, u) \rangle_{W^*,W},$$

where we just write $\lambda = \lambda_H$ and the equality constraint H is just the PDE complemented with initial conditions.

Here Assumption 2.42 is replaced by

Assumption 2.70 *(Basic Assumptions on Spaces and Regularity for Optimal Control with PDE)*

a) Y, U, and W are Hilbert spaces and $\tilde{Y}$ be a Banach space densely embedded in Y.

b) $U_{ad}(\subset U)$ is a non-empty, closed convex subset.

c) $\mathcal{J}$ is F-differentiable in a neighbourhood (w.r.t. the $Y \times U$ topology) of $[\hat{y}, \hat{u}]$. This Fréchet derivative $\mathcal{J}'$ is locally Lipschitz continuous.

d) E is assumed to be F-differentiable at $[\hat{y}, \hat{u}]$, in particular $E'_y(\hat{y}, \hat{u}) \in \mathcal{L}(\tilde{Y}, W)$. According to Th. A.5 the operator $E'_y : \tilde{Y} \to W$ may be extended uniquely to a densely defined operator $\tilde{G}$ with domain in Y.

Assumption 2.71 *(Assumptions for Weakly Singular Optimal Control Problems with PDE [IK08, Sect. 1.5])*

a) We assume that the adjoint operator

$$\tilde{G}^* : D(\tilde{G}^*) \subset W \to Y$$

is densely defined and thus it is necessarily closable. The closed operator is denoted with the same letter for ease of presentation.

b) Furthermore, we need the regularity assumption

$$\mathcal{J}'_y(\hat{y}, \hat{u}) \in \mathit{Range}(\tilde{G}^*),$$

implying the existence of a solution $\lambda \in D(\tilde{G}^*)$ of the adjoint equation

$$\mathcal{J}'_y(\hat{y}, \hat{u}) + \tilde{G}^* \lambda = 0.$$

c) *There exists a dense subset $U_{ad}^D \subset U_{ad}$, such that for every $u \in U_{ad}^D$ there exists a $\tau_u > 0$ implying the existence of $y(\tau) \in \tilde{Y}$ for all $\tau \in [0, \tau_u]$ such that*

$$E(y(\tau), \hat{u} + \tau(u - \hat{u})) = 0_W,$$
$$\text{and } \lim_{\tau \to 0+} \frac{\|y(\tau) - \hat{y}\|_Y^2}{\tau} = 0.$$

d) *For every $u \in U_{ad}^D$ and $y(\cdot)$ as introduced in c), E is directionally differentiable at every element of the plane*

$$\{(\hat{y} + \sigma(y(t) - \hat{y}), \hat{u} + \sigma\tau(u - \hat{u})) \mid \sigma \in [0,1], \tau \in [0, \tau_u]\}$$

w.r.t. all directions $(\hat{y}, \hat{u}) \in \tilde{Y} \times U$ and, moreover,

$$\left\langle \lambda, \int_0^1 \left[E'(\hat{y} + \sigma(y(\tau) - \hat{y}), \hat{u} + \sigma\tau(u - \hat{u})) - E'(\hat{y}, \hat{u})\right](y(\tau) - \hat{y}, \tau(u - \hat{u}))\, d\sigma \right\rangle_{W^*,W}$$

converges to 0 for $\tau \to 0+$.

Remark 2.72 *(Simplifications and Comments for Assumption 2.71)*

a) *If $E : \tilde{Y} \times U \to W$ is F-differentiable w.r.t. y with locally Lipschitz derivative and if Assumption 2.71 c) holds for Y instead of $\tilde{Y}$, then Assumption 2.71 d) follows automatically from c).*

b) *Please note that we do not require that $E'(\hat{y}, \hat{u}) : \tilde{Y} \times U \to W$ is surjective.*

c) *Analogously, we do not require that $E'(\hat{y}, \hat{u})$ is well-defined everywhere on $Y \times U$.*

d) *Typically, we think of $\tilde{Y} = Y \cap L^\infty(Q)$, where Y is a function space over $Q = \Omega$ or $= \Omega_{t_f}$ being a Hilbert space. This motivates to allow for the spaces Y and $\tilde{Y}$ with $\tilde{Y} \subsetneq Y$.*

These assumptions allow to prove the existence of a Lagrange multiplier w.r.t. the equality constraints.

Theorem 2.73 *(KKT-Conditions for Optimal Control with PDE)*
Assume $[\hat{y}, \hat{u}] \in Y \times U$ is a local minimizer of Problem 2.67. If Assumptions 2.70 and 2.71 hold, and if the state equation has a unique solution in a neighbourhood V with $U_{ad} \subset V \subset U$, where $E_y'(y(u), u) \in \mathcal{L}(Y, Z)$ has a bounded inverse for all $u \in V$, then there exists a Lagrange multiplier $\lambda \in W^*$ that fulfils the KKT-system

$$E(\hat{y}, \hat{u}) = 0_W,$$
$$\hat{u} \in U_{ad},$$
$$\mathcal{J}_y'(\hat{y}, \hat{u}) + \tilde{G}^*\lambda = 0_{Y^*},$$
$$\langle \mathcal{J}_u'(\hat{y}, \hat{u}) + E_u'(\hat{y}, \hat{u})^*\lambda, d_u \rangle_{U^*,U} \geq 0 \quad \forall d_u \in U_{ad} - \{\hat{u}\}.$$

For a proof of Theorem 2.73 see [IK08, Sect. 1.5]. For the applicability of the Assumptions 1) – 4), see [IK08, Examples 1.18 – 1.20, 1.22, 1.23].

The last but one equation is the equation for the adjoint λ that reads in the case $Y = \tilde{Y}$

$$\mathcal{J}'_y(y,u) + E'_y(y,u)^*\lambda = 0_{Y^*}.$$

Note that Th. 2.73 does not yield the uniqueness of the Lagrange multiplier. In the following this theorem is only used for Th. 3.6 that covers coupled ODE and parabolic PDE. The latter theorem is merely stated for completeness. However, we discuss its applicability to the truck-container example that is treated diffferently. For coupled CDE problems involving semilinear parabolic PDE and ODE, we follow a different approach relying on the implicit function theorem, see [CRT18], that requires weaker assumptions, but is restricted to the spatial dimensions $d = 1, 2, 3$. Note that the examples in [HPUU09, Ch. 1] exhibit a Hilbert space structure, where we may choose $Y = \tilde{Y}$. Thus Th. 2.51 can be applied directly and the particular setting of Th. 2.73 is not exploited.

Note that Th. 2.51 cannot be applied in general to our coupled CDE problems. We illustrate this by the following example that demonstrates some limits of applying the general optimization theory in Banach spaces to PDE. It has been thankfully brought to the authors' attention by F. Tröltzsch:

Example 2.74 *(Bratu Problem)*
We consider a special case of Example 2.68 for a semilinear elliptic PDE. Let $\Omega \subset \mathbb{R}^d$, $d \geq 1$. For given $u \in L^p(\Omega)$ for some $p \in [1, \infty]$, find y in a suitable state space Y, s.t.

$$\begin{aligned} -\Delta_x y + \exp(y) &= u, && \text{in } \Omega \\ y &= 0 && \text{on } \partial\Omega, \end{aligned}$$

i.e.

$$E : Y \times U \to W,\ [y,u] \mapsto E(y,u) = -\Delta_x y + \exp(y) - u = 0_W,$$

where we encode in Y the homogeneous Dirichlet boundary conditions for y.

At first glance we would choose $H^1_0(\Omega) \cap L^\infty(\Omega)$ for the state space Y, but the image space for the equation is $W = H^{-1}(\Omega)$, even for $u \in L^\infty(\Omega)$. E is F-differentiable from $H^1_0(\Omega) \cap L^\infty(\Omega)$ to W, but E'_y is no isomorphism from $H^1_0(\Omega) \cap L^\infty(\Omega) \to H^{-1}(\Omega)$ for $d \neq 1$.

This issue can be resolved by considering $Y = \{y \in H^1_0(\Omega) \,|\, -\Delta_x y \in L^p(\Omega)\}$, $p > d/2$ which is a Banach space with the corresponding norm. Moreover, there holds $Y \hookrightarrow C^0(\overline{\Omega})$. We have to choose $U = L^p(\Omega)$ such that $p > d/2$. Now $E : Y \times U \to W$ is continuously F-differentiable and $E_y : Y \to W$ is an isomophism, thus the implicit function theorem is applicable and the control-to-state operator $S : U \to Y$ is differentiable on $U = L^p(\Omega)$ to Y.

Then the problem is linearized and formally extended to $H^1_0(\Omega) \times L^2(\Omega)$ that allows for working with dual spaces. We find $E'_y : H^1_0 \to H^{-1}$ and the F-derivative in the direction $w \in W = H^1_0(\Omega)$ at $[\hat{y}, \hat{u}] \in Y \times U$ is given by

$$\langle E'_y(\hat{y},\hat{u})y, w\rangle_{W^*,W} = (\nabla y, \nabla w)_{L^2(\Omega)} + (\exp(\hat{y})y, w)_{L^2(\Omega)} \quad \forall w \in W.$$

In case of an optimal control problem as Problem 2.67, we would finally find as corresponding adjoint equation

$$-\Delta_x \lambda + \exp(\hat{y})\lambda = -\mathcal{J}_y'(\hat{y}, \hat{u})$$

that holds in sense of Y^*.

We compare with [IK08, Ex. 1.19] where the Bratu problem with distributed control $u \in U = L^2(\Omega)$ *for* $d = 1, 2, 3$ *is discussed. We consider only the case of pure Dirichlet b.c. They work with* $Y = H_0^1(\Omega)$ *and* $\tilde{Y} = H_0^1(\Omega) \cap L^\infty(\Omega)$ *in contrast and* $W = H^{-1}(\Omega)$. *In this case Assumption 2.71 can be verified using among other things that the exponential function is pointwise Lipschitz. Thus Theorem 2.73 is applicable. Note that for* $d \leq 3$ *we have* $p > 3/2$, *thus* $p = 2$ *is feasible as in the antecedent approach.*

We remark that this type of PDE has been examined thoroughly in the context of a shape optimization problem by the author in [BKN14].

As a special feature of optimal control with PDE, the underlying Banach spaces are typically Sobolev spaces or even Hilbert spaces, like L^2 or $H^1 := W^{1,2}$. Often the following situation is encountered that separable Hilbert spaces H, V form a Gelfand triple $V \overset{cd}{\hookrightarrow} H \overset{cd}{\hookrightarrow} V^*$ with continuous and dense embeddings (cf. Def. A.17). These structures may be exploited.

Remark 2.75 *(Constraint Qualifications and Optimal Control of PDE)*
In optimal control subject to PDE we consider here in this section no inequality constraints for the states. In this case it can be shown that the Robinson constraint qualification, then reading for surjective $E_y'(\hat{y}, \hat{u}) \in \mathcal{L}(Y, W)$

$$E_y'(\hat{y}, \hat{u})(y - \hat{y}) + E_u'(\hat{y}, \hat{u})(u - \hat{u}) = 0_W, \quad y \in Y, u \in U_{ad},$$

is satisfied and we obtain the KKT-conditions instead of the Fritz John-conditions [HPUU09, Remark 1.22].

We annotate that for semilinear PDE the existence of a bounded inverse $E_y'(\hat{y}, \hat{u})$ *often may be demonstrated, using a Nemytskii operator* Φ *from the state to the nonlinearity. For instance, if* $\Omega \subset \mathbb{R}^d$, $d \in \{1; 2; 3\}$, *a bounded Lipschitz domain and* $E(y, u) = -\Delta_x y + y^3 - u$ *(together with homogeneous Neumann b.c.), then* $\Phi : V = H^1(\Omega) \to L^2(\Omega)$, $y \mapsto y^3$ *is F-differentiable in* $H^1(\Omega)$. *For details, see [Tr10, 6.1.3].*

Theory and methods for optimal control of PDE depend on the type of PDE and on the type of control. Here we discuss only elliptic and parabolic PDE. For the optimal control of hyperbolic equations of first-order we refer, e.g., to [Ul02, Ul03]. For hyperbolic equations of second-order see [Zu05, GGP08, Kr11] among many others. Note that the viscosity solution of a hyperbolic equation of first-order is determined by solving a parabolic problem.

If the control enters by a source term within the domain, this case is called distributed control, or if it acts by means of a Neumann or Dirichlet boundary, it is called boundary control.

We present some specific examples (see [HPUU09]).

Example 2.76 *(Neumann Boundary Control of a Linear Elliptic PDE)*
Let $\Omega \subset \mathbb{R}^d$ be a bounded Lipschitz domain. Furthermore, let $y_{ref} \in L^2(\Omega)$ be a reference state, $\alpha > 0$, $d_0 \in L^\infty(\Omega)$, $d_0 > 0$, u_{min}, $u_{max} \in L^\infty(\partial\Omega)$, and $f \in H^1(\Omega)^*$.

Find $y \in \tilde{Y} = Y = V = H^1(\Omega)$, $u \in U = L^2(\partial\Omega)$ such that

$$\mathcal{J}(y,u) = \frac{1}{2}\|y - y_{ref}\|^2_{L^2(\Omega)} + \frac{\alpha}{2}\|u\|^2_{L^2(\partial\Omega)}$$

is minimized,
subject to the constraints

$$\begin{aligned} -\Delta_x y + d_0 y &= f && \textit{in } \Omega, \\ \partial_\nu y &= u && \textit{on } \partial\Omega, \\ u &\in [u_{min}, u_{max}] && \textit{on } \partial\Omega, \end{aligned}$$

where ∂_ν denotes the normal derivative.

The weak formulation of this linear elliptic PDE reads[9]

$$\int_\Omega \nabla y \cdot \nabla v + d_0 y\, v\, dx = \int_\Omega f v\, dx + \int_{\partial\Omega} uv\, d\sigma(x) \quad \forall v \in H^1(\Omega)$$

or equivalently, in operator formulation,

$$Ay = Bu + f,$$

introducing (using the trace embedding $L^2(\partial\Omega) \hookrightarrow H^1(\Omega)$, see Th.A.21)

$$\begin{aligned} &A \in \mathcal{L}(H^1(\Omega), H^1(\Omega)^*), && \langle Ay, v\rangle_{H^1(\Omega),H^1(\Omega)^*} := \int_\Omega \nabla y \cdot \nabla v + d_0 y\, v\, dx, \\ &B \in \mathcal{L}(L^2(\partial\Omega), H^1(\Omega)^*), && \langle Bu, v\rangle_{H^1(\Omega),H^1(\Omega)^*} := \int_{\partial\Omega} uv\, d\sigma(x). \end{aligned}$$

This yields that A is self-adjoint and that $B^*v = v|_{\partial\Omega}$. Note that we have $W = Y^* = H^1(\Omega)^*$. We set according to our standard notation

$$E(y,u) = Ay - Bu - f,$$

being continuous and F-differentiable.

This constitutes a linear-quadratic optimal control problem as in Example 2.49, where $U_{ad} = \{u \in L^2(\partial\Omega) \,|\, u_{min}(x) \le u(x) \le u_{max}(x)\}$, $H = L^2(\Omega)$, $R_H = Id_{L^2(\Omega)}$, and $y_{H,ref} = y_{ref}$. The necessary optimality conditions yield as in Example 2.53,

$$\langle A^*\lambda + \hat{y} - y_{ref}, v\rangle_{H^1(\Omega)^*,H^1(\Omega)} = 0 \quad \forall v \in H^1(\Omega), \qquad \hat{u} = P_{U_{ad}}\left(\frac{1}{\alpha}B^*\lambda\right).$$

[9] The surface measure on $\Gamma = \partial\Omega$ is denoted here by $d\sigma$.

using a pointwise projection as in (2.43) with $\tilde{\gamma} = 1/\alpha$. Written as a PDE the adjoint equation is

$$\begin{aligned} -\Delta_x \lambda + d_0 \lambda &= -(\hat{y} - y_{ref}) && \text{in } \Omega, \\ \partial_\nu \lambda &= 0_{L^2(\partial\Omega)} && \text{on } \partial\Omega, \end{aligned}$$

with

$$\hat{u} = P_{U_{ad}} \left(\frac{1}{\alpha} \lambda|_{\partial\Omega} \right).$$

Since this implies that the adjoint has the regularity $H^1(\Omega)$, the trace $\lambda|_{\partial\Omega} \in H^{1/2}(\partial\Omega)$ is actually well-defined.

Moreover, using that the Sobolev embedding $H^{1/2}(\partial\Omega) \hookrightarrow L^p(\partial\Omega)$ actually holds for $p > 2$ as well, we find a norm gap as required for the theory of semismooth Newton systems. Consequently, we reconsider $B^* \in \mathcal{L}(H^1(\Omega), L^p(\Omega))$. We choose a subdifferential $D \in L^\infty(\partial\Omega)$ with $D(x) \in \partial_C P_{U_{ad}}(\frac{1}{\alpha} B^* \lambda^{(k)})$ for almost all $x \in \partial\Omega$. Then we find the semismooth Newton system

$$\begin{bmatrix} Id_{H^1(\Omega)^*} & 0_{L^2(\partial\Omega)} & A^* \\ 0_{H^1(\Omega)^*} & Id_{L^2(\partial\Omega)} & -\frac{D^{(k)}}{\alpha} B^* \\ A & -B & 0_{H^1(\Omega)} \end{bmatrix} s^{(k)} = - \begin{bmatrix} y^{(k)} - y_{ref} + A^* \lambda^{(k)} \\ u^{(k)} - P_{U_{ad}}(\frac{1}{\alpha} B^* \lambda^{(k)}) \\ Ay^{(k)} - Bu^{(k)} - f \end{bmatrix}.$$

Assumption 2.77 *(Assumption for Semilinear Parabolic PDE)*
We consider Example 2.69 a) where c_0 and c_Γ may be nonlinear in y.

The function $c_0(t, x, y) : \Omega_{t_f} \times \mathbb{R} \to \mathbb{R}$ is measurable w.r.t. $[t, x] \in \Omega_{t_f}$ for any fixed $y \in \mathbb{R}$ and almost everywhere in Ω_{t_f} it is monotone increasing, locally bounded and locally Lipschitz continuous w.r.t. y.

The function $c_\Sigma(t, x, y) : \Sigma_N \times \mathbb{R} \to \mathbb{R}$ is measurable w.r.t. $[t, x] \in \Sigma_N$ for any fixed $y \in \mathbb{R}$ and almost everywhere in Σ_N it is monotone increasing, locally bounded and locally Lipschitz continuous w.r.t. y.

For details on the definition of local boundedness and local Lipschitz continuity see [Tr10, Assumpt. 5.4].

Example 2.78 *(Distributed Control of a Semilinear Parabolic PDE)*
Let the assumptions from Example 2.69 a) hold and for the nonlinearities c_0 and c_Γ let Assumption 2.77 hold. Contrary to Ex. 2.69 b) we consider $\Sigma_{t_f} = \Sigma_N$. Note that here $a_{ij} = \delta_{ij}$ yielding the Laplace operator, i.e. $A = \Delta_x$, and that $b \equiv 0$.

Furthermore, let $y_{ref} \in W(I; L^2, H^1)$, $\alpha > 0$, $y_0 \in C^0(\overline{\Omega})$ and $g \in L^s(\Sigma_{t_f})$, $s > d + 1$.

The set of controls is $U_{ad} = \{u \in L^\infty(\Omega_{t_f}) \,|\, u_{min}(t, x) \le u(t, x) \le u_{max}(t, x)$ for a.a. $[t, x] \in \Omega_{t_f}\}$, where u_{min}, $u_{max} \in L^\infty(\Omega_{t_f})$.

Find $y \in \tilde{Y} = Y = W(0, t_f; L^2, H^1)$, $u \in U = L^r(\Omega_{t_f})$, $r > d/2 + 1$ and $r \ge 2$, such that

$$\mathcal{J}(y, u) = \frac{1}{2} \|y - y_{ref}\|^2_{W(I; L^2, H^1)} + \frac{\alpha}{2} \|u\|^2_{L^2(\Omega_{t_f})}$$

is minimized,
subject to the constraints

$$\begin{aligned} y'_t - \Delta_x y + c_0(t,x,y) &= u && \text{in } \Omega_{t_f}, \\ \partial_\nu y + c_\Sigma(t,x,y) &= g && \text{on } \Sigma_{t_f}, \\ y(0,\cdot) &= y_0 && \text{in } \Omega, \\ u &\in [u_{min}, u_{max}] && \text{in } \Omega_{t_f}. \end{aligned}$$

Note that here the Gelfand triple is $V = H^1(\Omega)$, $H = L^2(\Omega)$, and $V^* = H^1(\Omega)^*$.

2.7.2 Optimal Control of a Parabolic PDE

We recall a well-known existence and uniqueness result for a semilinear parabolic PDE , see, e.g., [Tr10, Lemma 5.3].

Theorem 2.79 *(Existence and Uniqueness for a Parabolic PDE (L^2-theory))*
Let the assumptions in Example 2.78 hold for $r = s = 2$, d arbitrary, and let in addition $c_0(t,x,y)$ and $c_\Gamma(t,x,y)$ be uniformly bounded and globally Lipschitz w.r.t. y for almost all $[x,t] \in \Omega_{t_f}$ and Σ_N, respectively. For any $u \in L^2(\Omega_{t_f})$, $g \in L^2(\Sigma_N)$, and $y_0 \in L^2(\Omega)$, the initial-boundary value problem from Example 2.69 has a unique weak solution $y \in W(I; L^2, H^1)$.

The standard existence and uniqueness of weak solutions for parabolic PDE has been extended in the context of optimal control by Casas [Ca97] and Raymond and Zidani [RZ99], see Th. A.28. We continue with the optimal control problem in Example 2.78.

Theorem 2.80 *(Existence of Optimal Controls for OCP subject to a Parabolic PDE)*
Let the assumptions in Example 2.78 hold. Then for the distributed and the boundary control problem, where $u_\Sigma = g$ is another optimal control, there exists at least one optimal pair $[\hat{u}, \hat{u}_\Sigma]$ and an optimal state $\hat{y}$.

For a proof see, e.g., [Tr10, Th. 5.7].

Theorem 2.81 *(NOC for Distributed OCP subject to a Semilinear Parabolic Problem)*
We consider Example 2.78 (with Neumann b.c.) and let the assumptions there hold. The weak formulation of this semilinear parabolic PDE reads [Tr10, Sect. 5.1], find $y \in W(I; L^2, H^1) \cap L^\infty(\Omega_{t_f})$ such that[10]

$$\begin{aligned} &\int_{\Omega_{t_f}} -y v'_t + \nabla y \cdot \nabla v + c_0(t,x,y) v \, dx \, dt + \int_{\Sigma_{t_f}} c_\Sigma(t,x,y) v \, d\sigma(x), dt \\ &= \int_{\Omega_{t_f}} uv \, dx \, dt + \int_{\Sigma_{t_f}} gv \, d\sigma(x) \, dt + \int_\Omega y_0 v \, dx \quad \forall v \in \{v \in H^1(\Omega_{t_f}) \,|\, v(t_f, x) = 0\}. \end{aligned}$$

[10] The surface measure on Σ_{t_f} is denoted here by $d\sigma$.

It can be proved [Tr10, Th. 5.5], [RZ99] (see also Th. A.28) that under the above assumptions together with $y_0 \in C^0(\overline{\Omega_{t_f}})$, there exists a unique $y \in W(I; L^2, H^1) \cap C^0(\overline{\Omega_{t_f}})$. Furthermore, we have the embedding $W(I; L^2, H^1) \hookrightarrow C^0(\bar{I}; L^2(\Omega))$.

The equivalent operator formulation is

$$\begin{aligned} Ay &= Bu + Cg + Dy_0, \\ y(0, \cdot) &= y_0, \end{aligned}$$

where we have introduced

$$A : Y \to Y^*, \quad \langle Ay, v\rangle_{Y,Y^*} := \int_{\Omega_{t_f}} -yv'_t + \nabla y \cdot \nabla v + c_0(t, x, y)v \, dx \, dt + \int_{\Sigma_{t_f}} c_\Sigma(t, x, y)v \, d\sigma(x) \, dt,$$

and

$$\begin{aligned} &B \in \mathcal{L}(L^r(\Omega_{t_f}), L^{r'}(\Omega_{t_f})), && \langle Bu, v\rangle_{L^r(\Omega_{t_f}),L^{r'}(\Omega_{t_f})} := \int_{\Omega_{t_f}} uv \, dx \, dt, \\ &C \in \mathcal{L}(L^s(\Sigma_{t_f}), L^{s'}(\Sigma_{t_f})), && \langle Cg, v\rangle_{L^s(\Sigma_{t_f}),L^{s'}(\Sigma_{t_f})} := \int_{\Sigma_{t_f}} gv \, d\sigma(x) \, dt, \\ &D \in \mathcal{L}(L^2(\Omega), L^2(\Omega)), && \langle Dy_0, v\rangle_{L^2(\Omega),L^2(\Omega)} := \int_\Omega y_0 v \, dx. \end{aligned}$$

Note that we have $W = Y^* = W(I; L^2, (H^1)^*)$. We set according to our standard notation

$$E(y, u) = [Ay - Bu - Cg - Dy_0, y(0, \cdot) - y_0]^\top.$$

This operator is continuous and F-differentiable.

If we assume further c_0, c_Σ to be twice differentiable w.r.t. y, the necessary optimality conditions yield

$$\begin{aligned} \langle A^*\lambda + \hat{y} - y_{ref}, v\rangle_{W(I;L^2,(H^1)^*),W(I;L^2,H^1)} &= 0 \quad \forall v \in W(I; L^2, H^1), \\ \hat{u} &= P_{[u_{min},u_{max}]}\left(\frac{1}{\alpha}\lambda\right) \end{aligned}$$

using a pointwise projection as in (2.43).

Written as a PDE the adjoint equation is

$$\begin{aligned} -\lambda'_t - \Delta_x\lambda + c'_{0;y}(t, x, \hat{y})\lambda &= -(\hat{y} - y_{ref}) && \text{in } \Omega_{t_f}, \\ \partial_\nu\lambda + c'_{\Sigma;y}(t, x, \hat{y})\lambda &= 0_{L^2(\Sigma_{t_f})} && \text{on } \Sigma_{t_f}, \\ \lambda(t_f, x) &= 0_{L^2(\Omega)} && \text{in } \Omega. \end{aligned}$$

This implies that the adjoint actually has the regularity $W(I; L^2, H^1) \cap C^0(\overline{\Omega_{t_f}})$ as well.

Remark 2.82 *(Adjoint Corresponding to Neumann Boundary Conditions or Initial Conditions)*
Note that, in general, it turns out that the adjoints corresponding to a Neumann boundary condition of a PDE coincide with the adjoint introduced for the partial differential equation itself

[Tr10, Sect. 3.1]. The same reference illustrates that we may include initial conditions into the Lagrange function as well and the corresponding adjoint fits to the adjoint of the PDE. An analogous result holds for incorporating the initial condition of an ODE into the Lagrangian.

For a bang-bang example for parabolic PDE see [Tr10, Subsect. 3.2.4] and [Tr84a, Tr84b].

2.8 Comparison of Lagrange and Pontryagin Approach

The Pontryagin minimum(/maximum) principle avoids the differentiation w.r.t. the control and is valid under natural assumptions without any assumption on convexity, as it is used often for the exact Lagrange approach [Tr10, Sect. 4.8]. Typically the Pontryagin minimum principle is applied in the context of optimal control with ODE. However, an approach using Pontryagin's minimum principle is also used in optimal control of partial differential equations.

Pontryagin Principle Applied to Optimal Control with PDE

Under natural assumptions it can be expected that optimal controls for problems subject to semilinear elliptic and to semilinear parabolic PDE satisfy the minimum principle. The first results [Wo76, Wo77] were obtained for semilinear parabolic PDE, and then for semilinear elliptic PDE [BC91]. Recent extensions, in particular for state constraints, can be found in [Ca86, BC95] for elliptic and in [Ca97, RZ99] for semilinear parabolic PDE.

In the latter article by Raymond and Zidani the controls might turn out to be unbounded. Let $W(I) := W(I; L^2, H^1) = \{y \in L^2(I; H^1(\Omega)) \,|\, y_t' \in L^2(I; H^1(\Omega)^*)\}$, see Def. A.17. They use the existence and uniqueness of a PDE solution in $W(I) \cap L^\infty(\Omega_{t_f})$ and that the setting for the distributed, boundary and initial controls are subsets of function spaces of the type $\{u \in L^q(\Omega_{t_f}) \,|\, u(t,x) \in U_K$ for a.e. $[t,x] \in \Omega_{t_f}\}$, where U_K is a non-empty, closed subset of $\mathbb{R}$. We recall that $I = (0, t_f)$. They define a distributed Hamilton function, a boundary Hamilton function, and an initial Hamilton function and show that they each obey a minimum principle of Pontryagin's type. An application of their result is the optimal control with pointwise state constraints.

Lagrange Approach Applied to Optimal Control with ODE

The formal and the exact Lagrange method could be applied to ODE as well. We reconsider Problem 2.59 but without algebraic states q_2, rewriting $q = q_1$ for the differential states, $f = f_1$ and $H_2 = -\Psi$ consequently. The Lagrange function reads

$$L(q, \lambda, u) = \lambda_0 \mathcal{J}(q, u) + \langle \lambda_{H_1}, \dot{q} - f \rangle_{W_{H_1}^*, W_{H_1}} - (\lambda_{H_2}, \Psi)_{\mathbb{R}^{n_\Psi}} + \langle \lambda_G, G \rangle_{W_G^*, W_G}$$

and we obtain the necessary optimality conditions as in Th. 2.62. The question remains what are suitable spaces Y, U, and W in this setting. Neglecting the inequality constraints, according to Sect. 2.6 we work with $Y = [W^{1,\infty}(I)]^{n_q}$, $W_H = [L^\infty(I)]^{n_q} \times \mathbb{R}^{n_\Psi}$, $U = [L^\infty(I)]^{n_u}$. We have demonstrated *a posteriori* in Lemma 2.63 that the adjoints exhibit a higher regularity than just being elements of duals, this yields here $\lambda_{H_1} \in W^*_{H_1}$. As it turns out in the next chapter, we may also consider $Y = [L^2(I)]^{n_q}$, $U = [L^2(I)]^{n_u}$, and $W_H = Y^* \times \mathbb{R}^{n_\Psi}$, where the dual Y^* is identified with Y itself.

2.9 Article: Globalized Semismooth Newton Method for Operator Equations in Hilbert Spaces

We close this chapter with our article, published as [GHK17]. In this paper we present a globalization strategy for the semismooth Newton method for solving the operator equation $f(z) = 0_{Z^*}$ in a Hilbert space Z. This is achieved by equipping the local semismooth Newton method with an Armijo line search w.r.t. the merit function $\Theta(z) := \|f(z)\|^2_{Z^*}/2$. The result is applied to optimal control of semilinear elliptic PDE, where the control costs 2.41 enter the objective with a weight[11] $\alpha > 0$. In this setting it turns out that, if we may choose the Tikhonov parameter $\alpha = 1/\tilde{\gamma}$ ($\tilde{\gamma}$ as given in Eq. (2.26)) sufficiently large, then the search direction of the Newton method is always a descent direction and we have transition from our globalized algorithm to the fast local Newton method. As far to the knowledge of the author from an algorithmic point of view, this is the most efficient globalization strategy for Newton methods applied to optimal control of semilinear elliptic differential equations. By a semi-discretization in time this may be applied to parabolic differential equations as well, see [KG16].

[11] (α corresponds to λ in the notation of the paper.)

JOURNAL OF INDUSTRIAL AND
MANAGEMENT OPTIMIZATION

doi:10.3934/jimo.2016003

LINE SEARCH GLOBALIZATION OF A SEMISMOOTH NEWTON METHOD FOR OPERATOR EQUATIONS IN HILBERT SPACES WITH APPLICATIONS IN OPTIMAL CONTROL

Matthias Gerdts, Stefan Horn and Sven-Joachim Kimmerle*

Institut für Mathematik und Rechneranwendung (LRT-1)
Universität der Bundeswehr München
Werner-Heisenberg-Weg 39
85577 Neubiberg/München, Germany

(Communicated by Liqun Qi)

Abstract. We consider the numerical solution of nonlinear and nonsmooth operator equations in Hilbert spaces. A semismooth Newton method is used for search direction generation. The operator equation is solved by a globalized semismooth Newton method that is equipped with an Armijo linesearch using a semismooth merit function. We prove that an accumulation point of the globalized algorithm is a solution and transition to fast local convergence under a directional Hadamard-like continuity assumption on the Newton matrix. In particular, no auxiliary descent directions or smoothing steps are required. Finally, we apply this method to a control-constrained and also to a regularized state-constrained optimal control problem subject to partial differential equations.

1. **Introduction.** In this article, we consider the problem of finding a $\bar{z} \in Z$ with

$$f(\bar{z}) = 0, \tag{1}$$

where Z is a Hilbert space, Z^* denotes its topological dual, and $f : Z \to Z^*$ is a locally Lipschitz continuous operator. In particular, we allow f to be nonlinear and nonsmooth. Problems of this type arise, e.g., from the reformulation of the necessary optimality conditions in control-constrained optimal control subject to partial differential equations (PDE), see Section 4.

Semismooth Newton methods for operator equations of the type of (1) have been examined in [4, 10, 15, 16, 18, 19, 21, 22, 23, 24]. A trust-region globalization of this method was presented in [22], global convergence is proved for a primal-dual active set method in finite dimensions in [15, 16]. A primal-dual active set strategy and semismooth Newton methods may lead to identical algorithms [8]. The application of semismooth Newton methods to optimal control of PDE was considered, e.g., in [9, 10, 12, 13, 20, 22, 23, 24]. For variationally discretized control-constrained optimal control problems subject to elliptic differential equations a globalization

2010 *Mathematics Subject Classification.* Primary: 49J20, 49J52, 49M15; Secondary: 90C56, 65K10, 65N12.

Key words and phrases. Semismooth Newton method, line search globalization, superlinear convergence, optimal control of partial differential equations, control constraints, state constraints.

* Corresponding author.

strategy was developed by Hinze and Vierling [12, 13]. Discretized nonlinear optimal control problems subject to control and state constraints can be solved by globally convergent nonsmooth Newton methods, see for instance [7]. It has been demonstrated that a norm gap between range and domain of a projection operator is required for semismoothness, see e.g. [21, Th. 3.3], [22, Remark 3.34] or [24, Ch. 4]. In [21] the semismooth Newton method is revisited from a different point of view involving semismooth superposition operators and replacing Lipschitz continuity by weaker grow conditions.

We extend the local semismooth Newton method, see M. Ulbrich [22], by a globalization strategy in Hilbert spaces from [14]. We prove that if our global algorithm produces an accumulation point, then we have global convergence to a zero of f, see our main results Theorem 3.4 and Theorem 3.5, that, to our best knowledge, have not been proved so far. As in [22] we do not require a strict complementarity assumption for these results.

Our article is organized as follows. The concept of a semismooth Newton method is briefly presented in Section 2. In Section 3 we prove our main result. This globalization result is applied to optimal control problems with control constraints in Sect. 4, using an all-at-once approach. We close with numerical examples for control constraints and for state constraints in Sect. 5.

2. Semismooth Newton method.

For applying a Newton method on (1) an appropriate substitute for the Fréchet-derivative f' has to be found. Therefore we work with the following abstract semismoothness concept, cf. Def. 3.1 in [22]:

Definition 2.1 ($\partial^* f$-Semismoothness)**.** Let $f : Z \supset V \to Z^*$ be defined on an open subset V of Z. Further, let be given a set-valued mapping $\partial^* f : V \rightrightarrows \boldsymbol{\mathcal{L}}(Z, Z^*)$. f is called

(a) *$\partial^* f$-semismooth* at $z \in V$ if f is continuous near z and

$$\sup_{M \in \partial^* f(z+s)} \|f(z+s) - f(z) - M\,s\|_{Z^*} = o(\|s\|_Z) \quad \text{as } \|s\|_Z \to 0,$$

$$\text{i.e. } \forall \varepsilon > 0\, \exists \delta > 0 : \|s\|_Z < \delta \;\Rightarrow\; \|f(z+s) - f(z) - M\,s\|_{Z^*} \le \varepsilon\, \|s\|_Z \quad \forall M \in \partial^* f(z+s).$$

(b) *α-order $\partial^* f$-semismooth* at $z \in V$, $0 < \alpha \le 1$, if f is continuous near z and

$$\sup_{M \in \partial^* f(z+s)} \|f(z+s) - f(z) - M\,s\|_{Z^*} = \boldsymbol{\mathcal{O}}(\|s\|_Z^{1+\alpha}) \quad \text{as } \|s\|_Z \to 0.$$

$$\text{i.e. } \exists \varepsilon > 0\, \exists \delta > 0 : \|s\|_Z < \delta \;\Rightarrow\; \|f(z+s) - f(z) - M\,s\|_{Z^*} \le \varepsilon\, \|s\|_Z^{1+\alpha} \quad \forall M \in \partial^* f(z+s).$$

(c) *uniformly $\partial^* f$-semismooth* for all $z \in V$, if f is continuous in V and

$$\forall \varepsilon > 0\, \exists \delta > 0\, \forall z \in V : \|s\|_Z < \delta \;\Rightarrow\; \|f(z+s) - f(z) - M\,s\|_{Z^*} \le \varepsilon\, \|s\|_Z$$

$\forall M \in \partial^* f(z+s)$, i.e. δ does not depend on z.

The multifunction $\partial^* f$ is called *generalized differential of f*.

Please note that the uniform semismoothness of f in part (c) of Definition 2.1 is only required later on in the particular case that a (sub)sequence of stepsizes in the globalized Newton method tends to zero, compare Part b) of Theorem 3.4. However, there exist projectors s.t. uniform semismoothness implies Fréchet differentiability. For this and a discussion of non-uniform semismoothness of projectors see [21, Sect. 3]. Note that a different definition of uniform semismoothness is used by

Correa and Joffre [5, Remark 6.3] in which the uniformity refers to a family of functions. For a uniform semismoothness within the context of mesh independence see [24, Ch. 6].

It is easy to prove that the direct product of (α-order) semismooth operators is (α-order) semismooth itself with respect to the direct product of the generalized differentials of its components. An equivalent relation holds for the composition of semismooth operators. For details see [22]. For $Z = \mathbb{R}^n$ thanks to Rademacher's theorem such a substitute for locally Lipschitz continuous functions is given by Clarke's generalized Jacobian. In infinite dimensional spaces such a generalized differential is not given naturally and has to be defined separately, see e.g. [6, 22].

A semismooth Newton method for equation (1) is given by

Algorithm 2.2 (Local semismooth Newton method)**.**

(i) Choose a starting point $z_0 \in Z$ and set $k = 0$.

(ii) Is a stopping criterion satisfied: STOP.

(iii) Choose an arbitrary $M(z_k) \in \partial^* f(z_k)$ and compute the search direction $s_k \in Z$ by solving

$$M(z_k)\, s_k = -f(z_k). \tag{2}$$

(iv) Set $z_{k+1} = z_k + s_k$, $k := k + 1$ and goto (ii).

Here we need some regularity condition for the operators $M(z_k) \in \partial^* f(z_k)$ similar as in [22, Assumption 3.11 (i)].

Assumption 2.3 (Uniform non-singularity of the Newton matrix)**.** *The operators $M(z_k) \in \partial^* f(z_k)$ are continuously invertible elements of $\boldsymbol{\mathcal{L}}(Z, Z^*)$ and there exists a constant $C_{M^{-1}} > 0$ with*

$$\left\| M(z_k)^{-1} \right\|_{\boldsymbol{\mathcal{L}}(Z^*,Z)} \leq C_{M^{-1}} \quad \forall k \in \mathbb{N}.$$

This assumption is a standard assumption for Newton methods and can be verified specifically for various examples e.g. [9, 10, 17]. In a finite dimensional setting, this assumption will be satisfied close to a zero $\bar{z}$ of f, if all elements M of $\partial^* f(\bar{z})$ are non-singular, where $\partial^* f$ denotes the compact-valued and upper semicontinuous subdifferentials of Bouligand, Clarke or Qi, compare [23, Remark 2.8].

For our globalization result we need Assumption 3.1, see below, for the subdifferentials. Furthermore we need a certain structure of the superposition operator arising within the context of the application to optimal control, see Assumption 4.2 below, that avoids a smoothing step in Algorithm 2.2 required in [6] within this context.

We recall the local convergence of Algorithm 2.2 that has been shown in [6] and [22].

Theorem 2.4 (Local convergence of Algorithm 2.2)**.** *Let $V \subset Z$ be open and $f : V \to Z^*$. Let Assumption 2.3 hold. Assume $\bar{z} \in V$ is a zero of (1), then the following assertions hold:*

(a) If f is $\partial^* f$-semismooth in $\bar{z}$, then there exists a constant $\delta > 0$, such that for all $z_0 \in \bar{z} + \delta B_Z$ Algorithm 2.2 generates a sequence $\{z_k\} \subset V$ that converges superlinearly to $\bar{z}$.

(b) If f in (a) is α-order $\partial^* f$-semismooth in $\bar{z}$, $0 < \alpha \leq 1$, the rate of convergence is $1 + \alpha$.

(c) If $f(z_k) \neq 0$ for all k, the sequence of residuals converges superlinearly, i.e.

$$\lim_{k\to\infty} \frac{\|f(z_{k+1})\|_{Z^*}}{\|f(z_k)\|_{Z^*}} = 0.$$

3. **Globalization.** To construct a globally convergent Newton method Algorithm 2.2 is extended by an Armijo line search. An appropriate merit function $\Theta : Z \to \mathbb{R}$ with $\Theta(z) \to 0$ for $\|f(z)\|_{Z^*} \to 0$ is given by

$$\Theta(z) := \frac{1}{2} \|f(z)\|^2_{Z^*} . \tag{3}$$

Due to the fact that Z is a Hilbert space one can prove the semismoothness of (3). To this end, let $\langle \cdot, \cdot \rangle_{Z^*,Z}$ denote the dual pairing between Z^* and Z while $(\cdot, \cdot)_Z$ is the inner product on Z. We need

Assumption 3.1 (Directional Hadamard-like continuity of the Newton matrix)**.**

(a) *For any $k \in \mathbb{N}$ there exist $M_\rho := M(z_k + \rho s_k) \in \partial^* f(z_k + \rho s_k)$ for every ρ sufficiently small with a $\tilde{\varepsilon} \in [0, 1)$ such that*

$$\lim_{\rho\downarrow 0} \|(M_\rho - M(z_k))s_k\|_{Z^*} = \tilde{\varepsilon}\|f(z_k)\|_{Z^*}.$$

In particular, this assumption is fulfilled in the case $\tilde{\varepsilon} = 0$, corresponding to a directional Hadamard-like continuity of M.

(b) For $k \to \infty$, $\alpha_k \to 0$ and a subsequence of $\{z_k\}$ converging to $\bar{z}$ there exists $\bar{M} \in \partial^* f(\bar{z})$ with

$$\lim_{k\to\infty} \|(M_k - \bar{M})s_k\|_{Z^*} = 0$$

for $M_k \in \partial^* f(z_k + \alpha_k s_k)$.

The conditions in Assumption 3.1 have some similarity with the definition of Hadamard differentiability [1, Def. 2.45] along curves. In our context, however, these conditions ensure the continuity or the smallness of jumps of certain elements of the subdifferentials at least in the directions s_k. Assumptions (a) and (b) are satisfied generically at all points of differentiability of f. At points of non-differentiability of f the assumptions do impose restrictions on the choice of the elements from the subdifferentials and enforce a continuous selection. There might be cases where such a selection is not possible, because the search direction s_k may be obtained with some $M(z_k)$ and points into a direction where M_ρ is bounded away from $M(z_k)$. In the latter case we need at least bounds on the jump of M along a search direction.

However, we are able to monitor this situation algorithmically since Assumption 3.1 (a) will be exploited in the subsequent lemma to prove that the search direction is a direction of descent. Hence, if the search direction turns out to be a direction of ascent, then Assumption 3.1 (a) will be violated. In our numerical experiments that, however, rely on a discretized version of Algorithm 3.3 we haven't encountered this situation, though.

Lemma 3.2 (Semismoothness of Θ)**.** *Let $f : Z \to Z^*$ be (α-order) $\partial^* f$-semismooth. Furthermore, let $H : Z^* \to \mathbb{R}$ be given by $H(\cdot) := \frac{1}{2} \|\cdot\|^2_{Z^*} = \frac{1}{2} (\cdot, \cdot)_{Z^*}$, where $(\cdot, \cdot)_{Z^*}$ denotes the inner product in Z^*. Then the following assertions hold:*

(a) H is Fréchet-differentiable on Z^* with $H'(f)h = (f, h)_{Z^*}$ for $f : Z \to Z^*$,$h \in Z^*$.

(b) *The merit function $\Theta : Z \to \mathbb{R}$ in (3) is (α-order) $\partial^*\Theta$-semismooth, where the generalized differential $\partial^*\Theta := \{H'\} \circ \partial^* f : Z \rightrightarrows \boldsymbol{\mathcal{L}}(Z, \mathbb{R})$ is defined by*

$$\partial^*\Theta(z) = (\{H'\} \circ \partial^* f)\,(z) = \{H'(f(z))\, M(z) : M(z) \in \partial^* f(z)\}\,.$$

(c) *Let $\{z_k\} \subset Z$ be a sequence generated by Algorithm 2.2 and let Assumption 2.3 hold. For any s_k being defined by $M(z_k)\, s_k = -f(z_k)$ and the associated $V(z_k) = H'(f(z_k))M(z_k) \in \partial^*\Theta(z_k)$, it holds that*

$$\langle V(z_k), s_k\rangle_{Z^*,Z} = -\, \|f(z_k)\|^2_{Z^*} = -2\,\Theta(z_k). \tag{4}$$

Furthermore, if Assumption 3.1 (a) holds, then s_k is a descent direction of Θ in z_k unless $\Theta(z_k) = 0$.

Proof. (a) As common we identify the dual space Z^{**} with Z^* by means of the Riesz representation. We find

$$\frac{H(f+h) - H(f) - (f,h)_{Z^*}}{\|h\|_{Z^*}} = \frac{\frac{1}{2}\,(f+h, f+h)_{Z^*} - \frac{1}{2}\,(f,f)_{Z^*} - (f,h)_{Z^*}}{\|h\|_{Z^*}}$$
$$= \frac{1}{2}\|h\|_{Z^*}.$$

Taking the limit $\|h\|_{Z^*} \to 0$ yields the assertion.

(b) Due to the (α-order) $\partial^* f$-semismoothness and local Lipschitz continuity of f, the validity of assertion (a) and the boundedness of $H'(f)$, all requirements of Proposition 3.7 in [22] are satisfied and Θ is (α-order) $\partial^*\Theta$-semismooth. Thus assertion (b) holds.

(c) With (a) we find

$$\langle V(z_k), s_k\rangle_{Z^*,Z} = H'(f(z_k))(M(z_k)s_k) = (f(z_k), M(z_k)s_k)_{Z^*} = -\|f(z_k)\|^2_{Z^*}$$

and thus $\langle V(z_k), s_k\rangle_{Z^*,Z} = -2\Theta(z_k)$.

The $\partial^*\Theta$-semismoothness of Θ according to (b) implies

$$\begin{aligned}
0 &= \limsup_{\rho\downarrow 0} \left| \frac{\Theta(z_k + \rho s_k) - \Theta(z_k)}{\rho} - \langle V(z_k + \rho s_k), s_k\rangle_{Z^*,Z} \right| \\
&\geq \limsup_{\rho\downarrow 0} \Bigg| \left| \frac{\Theta(z_k + \rho s_k) - \Theta(z_k)}{\rho} - \langle V(z_k), s_k\rangle_{Z^*,Z} \right| \\
&\qquad - |\langle V(z_k), s_k\rangle_{Z^*,Z} - \langle V(z_k + \rho s_k), s_k\rangle_{Z^*,Z}| \Bigg| \\
&= \limsup_{\rho\downarrow 0} \Bigg| \left| \frac{\Theta(z_k + \rho s_k) - \Theta(z_k)}{\rho} + 2\Theta(z_k) \right| \\
&\qquad - |\langle V(z_k), s_k\rangle_{Z^*,Z} - \langle V(z_k + \rho s_k), s_k\rangle_{Z^*,Z}| \Bigg|.
\end{aligned}$$

This relation holds for arbitrary choices of $M(z_k + \rho s_k) \in \partial^* f(z_k + \rho s_k)$ in $V(z_k + \rho s_k) = H'(f(z_k + \rho s_k)M(z_k + \rho s_k)$.

According to Assumption 3.1 (a) there exists $M_\rho := M(z_k + \rho s_k) \in \partial^* f(z_k + \rho s_k)$ such that

$$\lim_{\rho\downarrow 0} \|(M_\rho - M(z_k))s_k\|_{Z^*} = \tilde{\varepsilon}\|f(z_k)\|_{Z^*}.$$

Then, together with the continuity of H' we find

$$\begin{aligned}
&|\langle V(z_k), s_k\rangle_{Z^*,Z} - \langle V(z_k+\rho s_k), s_k\rangle_{Z^*,Z}| \\
&= \Big|(H'(f(z_k)), M(z_k)s_k)_{Z^*,Z^*} - (H'(f(z_k+\rho s_k)), M(z_k+\rho s_k)s_k)_{Z^*,Z^*}\Big| \\
&\leq \Big|(H'(f(z_k)), M(z_k)s_k)_{Z^*,Z^*} - (H'(f(z_k+\rho s_k)), M(z_k)s_k)_{Z^*,Z^*}\Big| \\
&\quad + \Big|(H'(f(z_k+\rho s_k)), M(z_k)s_k)_{Z^*,Z^*} - (H'(f(z_k+\rho s_k)), M(z_k+\rho s_k)s_k)_{Z^*,Z^*}\Big| \\
&\leq \|H'(f(z_k)) - H'(f(z_k+\rho s_k))\|_{\mathcal{L}(Z,Z^*)} \cdot \|f(z_k)\|_{Z^*} \\
&\quad + \|H'(f(z_k+\rho s_k))\|_{\mathcal{L}(Z,Z^*)} \cdot \|(M(z_k) - M_\rho)s_k\|_{Z^*} \\
&\to \|f(z_k)\|_{Z^*}^2 \tilde{\varepsilon} \qquad \text{as } \rho \downarrow 0.
\end{aligned}$$

Hence,

$$\lim_{\rho\downarrow 0}\left|\frac{\Theta(z_k+\rho s_k)-\Theta(z_k)}{\rho} + 2\Theta(z_k)\right| = 2\tilde{\varepsilon}\,\Theta(z_k)$$

and since $\tilde{\varepsilon} < 1$, s_k is a direction of descent of Θ at z_k unless $\Theta(z_k) = 0$.

□

Hence with $f(z_k) \neq 0$ the Armijo line search in the following algorithm is well defined. The descent property of s_k w.r.t. $\Theta(z_k)$ follows particularly from the construction of the merit function Θ.

Algorithm 3.3 (Global semismooth Newton method)**.**

(i) Choose $z_0 \in Z$, $\beta \in (0,1)$, $\sigma \in (0, 1/2)$ and set $k = 0$.

(ii) Is a stopping criterion satisfied: STOP.

(iii) Choose an arbitrary $M(z_k) \in \partial^* f(z_k)$ and compute the search direction $s_k \in Z$ by solving

$$M(z_k)\, s_k = -f(z_k).$$

(iv) Calculate $V(z_k) = H'(f(z_k))\, M(z_k) \in \partial^*\Theta(z_k)$, find the smallest $i_k \in \mathbb{N}_0$ with

$$\Theta(z_k + \beta^{i_k}\, s_k) \leq \Theta(z_k) + \sigma\, \beta^{i_k}\, \langle V(z_k), s_k\rangle_{Z^*,Z} \tag{5}$$

and set $\alpha_k = \beta^{i_k}$.

(v) Set $z_{k+1} = z_k + \alpha_k\, s_k$, $k := k+1$, and goto (ii).

Using (4) the inequality in (5) can equivalently be written as

$$\Theta(z_k + \beta^{i_k}\, s_k) \leq \left(1 - 2\,\sigma\,\beta^{i_k}\right)\Theta(z_k).$$

Hence $\langle V_k, s_k\rangle_{Z^*,Z}$ has not to be computed explicitly. Global convergence of Algorithm 3.3 can be shown by adapting Theorem 4.2 in [6] that has been derived within the frame of optimal control of ordinary differential equations.

Theorem 3.4 (Global convergence of Algorithm 3.3)**.** *Let f be semismooth. Let $\bar{z} \in Z$ be an accumulation point of the sequence $\{z_k\}$ generated by Algorithm 3.3 and let Assumption 2.3 hold.*

(a) If Assumption 3.1 (a) holds and

$$\underline{\alpha} := \liminf_{j\to\infty} \alpha_{k_j} > 0,$$

then $\bar{z}$ is a zero of f.

(b) If $\underline{\alpha} = 0$, if Assumption 3.1 holds and if f is locally uniformly semismooth in a neighbourhood of $\bar{z}$, then $\bar{z}$ is a zero of f.

For clarification, we would like to mention already here that the locally uniform semismoothness and Assumption 3.1 (b), are not fulfilled by the numerical examples in Sect. 5. However, it turns out that the pathological case (b) of the theorem, in which the Armijo step α_k tends to zero, can be ruled out by the following Theorem 3.5 under Assumption 3.1 (a) and is not met within both our examples.

Proof. We adapt the argumentation of Theorem 4.2 in [6]. Therefore let $\{z_{k_j}\}_{j\in\mathbb{N}}$ be a subsequence with $z_{k_j} \to \bar{z}$ and $f(z_{k_j}) \neq 0$. Then, $\langle V(z_{k_j}), s_{k_j}\rangle_{Z^*,Z} = -2\,\Theta(z_{k_j}) = -\|f(z_{k_j})\|^2_{Z^*} < 0$. According to Lemma 3.2 (c), s_{k_j} is a descent direction of Θ at z_{k_j} and the line search is well defined. There are two cases:

(a) Assume $\underline{\alpha} > 0$. We have

$$0 \leq \Theta(z_{k_{j+1}}) \leq \Theta(z_{k_j+1}) \leq \Theta(z_{k_j}) + \sigma\,\alpha_{k_j}\langle V(z_{k_j}), s_{k_j}\rangle_{Z^*,Z} = (1 - 2\,\sigma\,\alpha_{k_j})\,\Theta(z_{k_j}).$$

With $\sigma \in (0, 1/2)$ and $\underline{\alpha} \leq \alpha_{k_j} \leq 1$ it follows that $0 < 1 - 2\,\sigma\,\alpha_{k_j} \leq 1 - 2\,\sigma\,\underline{\alpha} < 1$. Repeated application yields

$$0 \leq \Theta(z_{k_j}) \leq (1 - 2\,\sigma\,\underline{\alpha})^j\,\Theta(z_{k_0}) \to 0.$$

By the continuity of f, $\bar{z}$ is a zero of f.

(b) Assume that there is a subsequence $\{z_k\}_{k\in J}$, $J \subseteq \{k_j : j \in \mathbb{N}\}$ with $\alpha_k \to 0$, $k \in J$. The sequence $\{s_k\}$ is bounded since $\{M(z_k)^{-1}\}$ is bounded and

$$0 \leq \|s_k\|_Z = \left\|M(z_k)^{-1} f(z_k)\right\|_Z \leq C_{M^{-1}}\,\|f(z_k)\|_{Z^*} \leq C_{M^{-1}}\,\|f(z_0)\|_{Z^*}\,.$$

Z is a Hilbert space and thus reflexive. According to the Eberlein-Smulian theorem, there exists a weakly convergent subsequence $\{s_k\}$, $k \in I \subseteq J$. Hence, there exists some $\bar{s} \in Z$ such that for every $V(\bar{z}) \in Z^*$ we have

$$\langle V(\bar{z}), s_k\rangle_{Z^*,Z} \to \langle V(\bar{z}), \bar{s}\rangle_{Z^*,Z} \quad \text{as } k \to \infty. \tag{6}$$

We get

$$\begin{aligned}
&\left|\frac{\Theta(z_{k+1}) - \Theta(z_k)}{\alpha_k} - \langle V(\bar{z}), \bar{s}\rangle_{Z^*,Z}\right| \\
&\quad\leq \left|\frac{\Theta(z_{k+1}) - \Theta(z_k)}{\alpha_k} - \langle V(z_{k+1}), s_k\rangle_{Z^*,Z}\right| \\
&\qquad + |\langle V(z_{k+1}), s_k\rangle_{Z^*,Z} - \langle V(\bar{z}), \bar{s}\rangle_{Z^*,Z}| \\
&\quad\leq \frac{1}{\alpha_k}\,|\Theta(z_{k+1}) - \Theta(z_k) - \langle V(z_{k+1}), \alpha_k s_k\rangle_{Z^*,Z}| \\
&\qquad + |\langle V(z_{k+1}), s_k\rangle_{Z^*,Z} - \langle V(\bar{z}), s_k\rangle_{Z^*,Z}| \\
&\qquad + |\langle V(\bar{z}), s_k\rangle_{Z^*,Z} - \langle V(\bar{z}), \bar{s}\rangle_{Z^*,Z}|\,.
\end{aligned}$$

The first term vanishes owing to the locally uniform semismoothness of Θ, the last term vanishes owing to (6) as $k \to \infty$. For the remaining term we obtain with $\bar{M}$, as defined in Assumption 3.1 (b), the following estimate

$$\begin{aligned}
&|\langle V(z_{k+1}), s_k\rangle_{Z^*,Z} - \langle V(\bar{z}), s_k\rangle_{Z^*,Z}| \\
&\quad = \left|H'(f(z_{k+1}))(M(z_{k+1})s_k) - H'(f(\bar{z}))(\bar{M}s_k)\right| \\
&\quad = \left|(f(z_{k+1}), M(z_{k+1})s_k)_{Z^*} - \left(f(\bar{z}), \bar{M}s_k\right)_{Z^*}\right| \\
&\quad \le |(f(z_{k+1}), M(z_{k+1})s_k)_{Z^*} - (f(\bar{z}), M(z_{k+1})s_k)_{Z^*}| \\
&\qquad + \left|(f(\bar{z}), M(z_{k+1})s_k)_{Z^*} - \left(f(\bar{z}), \bar{M}s_k\right)_{Z^*}\right| \\
&\quad \le \|f(z_{k+1}) - f(\bar{z})\|_{Z^*} \cdot \|M(z_{k+1})\|_{\mathcal{L}(Z,Z^*)} \cdot \|s_k\|_{Z^*} \\
&\qquad + \|f(\bar{z})\|_{Z^*} \cdot \|(M(z_{k+1}) - \bar{M})s_k\|_{Z^*}.
\end{aligned}$$

The first term vanishes owing to the continuity of f and the boundedness of $\{M(z_k)\}$ and $\{s_k\}$. The second term vanishes because of Assumption 3.1 (b). In summary we have shown that

$$\lim_{k\to\infty, k\in I} \frac{\Theta(z_{k+1}) - \Theta(z_k)}{\alpha_k} = \langle V(\bar{z}), \bar{s}\rangle_{Z^*,Z}.$$

The line search in step (iv) of Algorithm 3.3 yields

$$\frac{\Theta(z_k + \alpha_k\, s_k) - \Theta(z_k)}{\alpha_k} \le \sigma\, \langle V(z_k), s_k\rangle_{Z^*,Z} < \frac{\Theta(z_k + \frac{\alpha_k}{\beta}\, s_k) - \Theta(z_k)}{\frac{\alpha_k}{\beta}}.$$

Passing to the limit and exploiting the previous considerations yields

$$\sigma\, \langle V(\bar{z}), \bar{s}\rangle_{Z^*,Z} = \langle V(\bar{z}), \bar{s}\rangle_{Z^*,Z}.$$

Since $\sigma \in (0, 1/2)$ this only holds for $\langle V(\bar{z}), \bar{s}\rangle_{Z^*,Z} = 0$. Thus, we have shown

$$-\|f(z_k)\|_{Z^*}^2 = \langle V(z_k), s_k\rangle_{Z^*,Z} \to \langle V(\bar{z}), \bar{s}\rangle_{Z^*,Z} = 0.$$

By the continuity of f, $\bar{z}$ is a zero of f.

□

Now we can prove the transition of Algorithm 3.3 to fast local convergence, provided it converges at all. We present the proof here in details in order to show that we do not need Assumption 3.1 (b) and the locally uniform semismoothness in Definition 2.1 (c).

Theorem 3.5 (Transition to fast local convergence)**.** *Let Assumptions 2.3 and 3.1 (a) hold. Let $\{z_k\}$ be a sequence generated by Algorithm 3.3 with $f(z_k) \neq 0$ for all $k \in \mathbb{N}_0$. Furthermore, let $\bar{z}$ be an accumulation point of $\{z_k\}$ and a zero of f.*

If f is $\partial^* f$-semismooth the sequence $\{z_k\}$ converges superlinearly to $\bar{z}$ and α_k finally becomes 1. If f is α-order $\partial^* f$-semismooth the rate of convergence is $1 + \alpha$.

The idea of the proof is as follows. According to the local convergence result (Theorem 2.4), we know that the local semismooth Newton method (with $\alpha_k = 1$) converges superlinearly in some neighbourhood of a zero. Now, since $\bar{z}$ is an accumulation point of the sequence $\{z_k\}$ there is a subsequence converging to $\bar{z}$, which is assumed to be a zero. For k sufficiently large, z_k is in a neighbourhood of $\bar{z}$, where we have superlinear convergence of the local method. For one step of the local method we show that $\alpha_k = 1$ satisfies the Armijo condition. Hence, the globalized method at z_k will accept $\alpha_k = 1$ and from that point on the sequences of the local method and the globalized method coincide.

Proof. Let $\varepsilon > 0$ be arbitrary. The superlinear convergence result of the local algorithm (Theorem 2.4) with f being semismooth implies the existence of a $\delta > 0$ such that for all $\|z_k - \bar{z}\|_Z \leq \delta$ we have

$$\|z_k + s_k - \bar{z}\|_Z \leq \varepsilon \, \|z_k - \bar{z}\|_Z. \tag{7}$$

Since $f(\bar{z}) = 0$ the Lipschitz continuity of f yields

$$\|f(z_k + s_k)\|_{Z^*} \leq L \, \|z_k + s_k - \bar{z}\|_Z$$

with a constant $L > 0$. Combining this estimate with (7) we have

$$\|f(z_k + s_k)\|_{Z^*} \leq L \, \varepsilon \, \|z_k - \bar{z}\|_Z. \tag{8}$$

On the other hand with Assumption 2.3

$$\|s_k\|_Z = \|M_k^{-1}\|_{\mathcal{L}(Z^*,Z)} \, \|f(z_k)\|_{Z^*} \leq C_{M^{-1}} \|f(z_k)\|_{Z^*},$$

and moreover

$$\|z_k - \bar{z}\|_Z \leq \|z_k - (z_k + s_k)\|_Z + \|z_k + s_k - \bar{z}\|_Z \leq C_{M^{-1}} \|f(z_k)\|_{Z^*} + \varepsilon \|z_k - \bar{z}\|_Z.$$

This yields

$$\|z_k - \bar{z}\|_Z \leq \frac{C_{M^{-1}}}{1 - \varepsilon} \, \|f(z_k)\|_{Z^*}. \tag{9}$$

Together with (8) we have the estimate

$$\|f(z_k + s_k)\|_{Z^*} \leq L\varepsilon \frac{C_{M^{-1}}}{1 - \varepsilon} \, \|f(z_k)\|_{Z^*}.$$

If we choose

$$\varepsilon \leq \frac{\sqrt{1 - 2\sigma}}{LC_{M^{-1}} + \sqrt{1 - 2\sigma}} < 1,$$

then

$$\|f(z_k + s_k)\|_{Z^*} \leq \sqrt{1 - 2\sigma} \, \|f(z_k)\|_{Z^*},$$

i.e. $\alpha_k = 1$ is accepted. With (7) and $\varepsilon < 1$ we can repeat the argument and obtain convergence of the whole sequence. If f is $\partial^* f$-semismooth the convergence is superlinear according to Theorem 2.4 (a). If f is α-order $\partial^* f$-semismooth the rate of convergence is $1 + \alpha$ according to Th. 2.4 (b). □

Thereby we have proved the transition to fast local convergence of Algorithm 3.3 requiring only Assumption 3.1 (a) on the directional Hadamard-like continuity of the Newton matrix and a standard assumption on its uniform non-singularity.

4. Application to optimal control.

In control-constrained optimal control subject to PDE constraints the reformulation of the necessary optimality conditions leads to problems of the type of (1). For Hilbert spaces Y, W, U and the feasible set $U_{\text{ad}} \subset U$ such an optimal control problem, in general, is given as

$$\min_{(y,u) \in Y \times U} J(y,u) \quad \text{s.t.} \quad E(y,u) = 0, \quad u \in U_{ad} \qquad \text{(P1)}$$

with objective functional $J : Y \times U \to \mathbb{R}$ and state equation $E : Y \times U \to W^*$. Here y denotes the state while u is the control. Semismooth Newton methods for optimal control problems in the general setting of (P1) are investigated in [22], for more special cases see also [9, 12, 13, 20, 24]. Globalization strategies for semismooth Newton methods applied to a reduced formulation of (P1) are investigated, e.g., in [12, 13, 22]. Although Algorithm 3.3 would be applicable to this reduced problem as well, we want to solve the full Karush-Kuhn-Tucker (KKT) system related to (P1). In [22] this is referred to as the all-at-once approach.

As usual the Lagrange function $L : Y \times U \times W \to \mathbb{R}$ is defined as

$$L(y,u,w) = J(y,u) + \langle E(y,u), w\rangle_{W^*,W}. \tag{10}$$

Then the KKT-system is given by

Lemma 4.1 (cf. Corollary 1.3 in [11])**.** *Let $(\bar{y},\bar{u})$ be a solution of (P1). Furthermore, let $J : Y \times U \to \mathbb{R}$ and $E : Y \times U \to W^*$ be continuously Fréchet-differentiable. Then there exists a Lagrange multiplier $\bar{w} \in W$, such that the following optimality conditions hold:*

$$L_w(\bar{y},\bar{u},\bar{w}) = E(\bar{y},\bar{u}) = 0, \tag{11}$$

$$L_y(\bar{y},\bar{u},\bar{w}) = J_y(\bar{y},\bar{u}) + E_y(\bar{y},\bar{u})^*\bar{w} = 0, \tag{12}$$

$$\begin{aligned}\bar{u} \in U_{ad},\ &\langle L_u(\bar{y},\bar{u},\bar{w}), u - \bar{u}\rangle_{U^*,U}\\ &= \langle J_u(\bar{y},\bar{u}) + E_u(\bar{y},\bar{u})^*\bar{w}, u - \bar{u}\rangle_{U^*,U} \geq 0\ \forall u \in U_{ad}.\end{aligned} \tag{13}$$

We consider the Euclidean projection $P_{U_{ad}}$ onto the set of admissible controls and introduce the continuous function $\pi : \mathbb{R}\times\mathbb{R}\times\mathbb{R} \to \mathbb{R}, \xi = (\xi_1,\xi_2,\xi_3) \mapsto \pi(\xi) = \xi_2 - P_{U_{ad}}(\xi_2 - \gamma L_u(\xi_1,\xi_2,\xi_3))$. With the superposition operator $\Pi : Y \times U \times W \to U$, given as

$$\begin{aligned}\Pi(y,u,w)(x) &:= \pi(y(x),u(x),w(x))\\ &= u(x) - P_{U_{ad}}\left(u(x) - \gamma\, L_u(y(x),u(x),w(x))\right), \quad \gamma > 0,\end{aligned} \tag{14}$$

the variational inequality in (13) can be rewritten [22, Prop. 5.8] as

$$\Pi(y,u,w) = 0.$$

Thus, with $z := (u,v,w)$ and

$$f(z) := \begin{bmatrix} L_y(y,u,w) \\ \Pi(y,u,w) \\ E(y,u) \end{bmatrix} \tag{15}$$

the solution of (P1) is equivalent to finding a $\bar{z} := (\bar{y},\bar{u},\bar{w}) \in Z := Y \times U \times W$ with

$$f(\bar{z}) = 0.$$

For further analysis let $\Omega \subset \mathbb{R}^n$ be an open, bounded, measurable set with measure $\mu(\Omega) > 0$. We define the set of control functions as $U := L^2(\Omega)$. For simplicity we consider pointwise convex controls by using

$$U_{ad} := \left\{u \in L^2(\Omega) : a \leq u(x) \leq b \text{ a.e. in } \Omega;\ a,b \in \mathbb{R}\right\}$$

as the feasible set. The pointwise projection $P_{U_{ad}} : \mathbb{R} \to \mathbb{R},\ \xi \mapsto \max\{a, \min\{\xi, b\}\}$ is (1-order) semismooth [22, Ex. 5.23]. Note that $P_{U_{ad}}$ is not locally uniformly semismooth in any neighbourhood of a resp. b. In order to establish semismoothness of the superposition operator Π we follow [22, Assumption 5.20] and assume the following:

Assumption 4.2. *The following conditions hold:*

(a) *$E : Y \times L^2(\Omega) \to W^*$ and $J : Y \times L^2(\Omega) \to \mathbb{R}$ are twice continuously Fréchet-differentiable.*

(b) *L_u has the form $L_u(y,u,w) = \lambda u + G(y,u,w)$ and there exist $\lambda > 0$ and $p > 2$, such that*
   - (i) *$G : Y \times L^2(\Omega) \times W \to L^2(\Omega)$ is continuously Fréchet-differentiable.*
   - (ii) *The operator $(y,u,w) \in Y \times L^2(\Omega) \times W \mapsto G(y,u,w) \in L^p(\Omega)$ is locally Lipschitz-continuous.*

k	α_k	$\lVert f(z_k)\rVert_{Z^*}$	$\lVert s_k\rVert_Z$
0	–	5.43111E-02	–
1	9.76563E-04	5.43015E-02	5.00085E+00
2	3.12500E-02	5.36304E-02	1.82556E+00
3	5.00000E-01	2.91839E-02	1.55585E+00
4	6.25000E-02	2.75202E-02	3.87423E-01
⋮			
16	0.25000E+00	1.65715E-02	2.48095E-02
17	0.50000E+00	1.38976E-02	1.28644E-02
18	1.00000E+00	1.24060E-02	6.81858E-03
19	1.00000E+00	9.44693E-03	1.63072E-03
20	1.00000E+00	5.60965E-06	4.47294E-05
21	1.00000E+00	2.27743E-15	1.57318E-11

Table 1. Iteration history for the solution of problem (P2) for $h = 1/256$. Step size α_k, norm $\lVert f(z_k)\rVert_{Z^*}$ and norm of the search direction $\lVert s_k\rVert_Z$ for the k-th iterate. These numerical results exhibit the superlinear convergence.

As already mentioned in Section 2, a generalized differential of (15) is not given naturally. Motivated by the sum and chain rule in finite dimensions in [22], the set-valued mapping $\partial_C f : Y \times L^2(\Omega) \times W \rightrightarrows \boldsymbol{\mathcal{L}}(Y \times L^2(\Omega) \times W, Y^* \times L^2(\Omega) \times W^*)$ with

$$
\begin{aligned}
\partial_C f :=\Big\{ & M \in \boldsymbol{\mathcal{L}}(Y \times L^2(\Omega) \times W, Y^* \times L^2(\Omega) \times W^*) : \\
& M(y,u,w) = \begin{bmatrix} L_{yy}(y,u,w) & L_{yu}(y,u,w) & E_y(y,u)^* \\ \gamma D\, G_y(y,u,w) & I + \gamma D\, G_u(y,u,w) & \gamma D\, G_w(y,u,w) \\ E_y(y,u) & E_u(y,u) & 0 \end{bmatrix}, \\
& D \in L^\infty(\Omega), D(x) \in \partial_C P_{U_{ad}}(-\gamma G(y,u,w)(x)) \text{ in } \Omega \Big\} \qquad (16)
\end{aligned}
$$

is used as a generalized differential for (15). Here the subscript "C" emphasizes the close relation to Qi's C-subdifferential in finite dimensions.

Theorem 4.3 (Semismoothness of f)**.** *Let Assumption 4.2 hold and choose γ in (14) as $\gamma = 1/\lambda$. Then $f : Y \times L^2(\Omega) \times W \to Y^* \times L^2(\Omega) \times W^*$ in (15) is locally Lipschitz continuous and $\partial_C f$-semismooth.*

For a proof of Th. 4.3 the reader is referred to [22, Th. 5.21]. Therefore f as in (15) meets the requirements needed for Algorithms 2.2 and 3.3. The merit function in (3) for problem (P1) is thus given as

$$
\Theta(z) := \frac{1}{2}\,\lVert f(z)\rVert^2_{Z^*} = \frac{1}{2}\lVert L_y(y,u,w)\rVert^2_{Y^*} + \frac{1}{2}\,\lVert \Pi(y,u,w)\rVert^2_{L^2(\Omega)} + \frac{1}{2}\,\lVert E(y,u)\rVert^2_{W^*}\,. \tag{17}
$$

5. **Numerical examples.** In this section we start with a numerical example for the application of Algorithm 3.3 to problems of the type of (P1).

5.1. **Semilinear elliptic PDE with control constraints.** For simplicity we restrict ourselves at first to the following control-constrained optimal control problem subject to a semilinear elliptic PDE that was taken from [10, Ch. 6] and is motivated by an application in superconductivity:

$$\begin{aligned} &\text{Minimize} && J(y,u) := \frac{1}{2}\int_\Omega (y(x)-y_d(x))^2\,\mathrm{d}x + \frac{\lambda}{2}\int_\Omega u(x)^2\,\mathrm{d}x \\ &\text{w. r. t.} && (y,u) \in H_0^1(\Omega)\times L^2(\Omega) \\ &\text{subject to} && E(y,u) := -\Delta y + y^3 + y - u = 0 \quad \text{in } \Omega := (0,1)^2, \\ & && u \in U_{ad} := \left\{ u \in L^2(\Omega) : -4 \le u(x) \le 0 \ \text{ a.e. in } \Omega \right\}. \end{aligned} \tag{P2}$$

The desired state is $y_d := \frac{1}{6}\sin(2\pi x_1)\sin(2\pi x_2)\exp(2x_1)$, the Tikhonov parameter is chosen as $\lambda := 10^{-3}$. We convince ourselves that Assumption 4.2 is fulfilled for this example, where $Y = W = H_0^1(\Omega)$, $Y^* = W^* = H^{-1}(\Omega)$ and $L_u(y,u,w) = \lambda u - w$ yielding that $G(y,u,w) = -w$ has the required regularity. Hence $Z^* = H^{-1}(\Omega)\times L^2(\Omega)\times H^{-1}(\Omega)$.

It is not clear, how to check *a priori* whether Assumption 2.3 and Assumption 3.1 (a) hold in general in an infinite dimensional function space.

However, for both our examples the well-posedness of the linear operator M and the optimal control problem in a function space setting can be shown by means of the Lax-Milgram theorem (see [10, Sect. 4 & 5] and [9, Sect. 3], respectively), exploiting a sufficient second-order condition [3] for the semilinear state equation. Otherwise, we can see within our numerics whether Assumption 2.3 holds.

We verify Assumption 3.1 (a) for our undiscretized example, assuming that the Tikhonov parameter λ is sufficiently large. In the matrix M we have only the non-differentiable component $\gamma DG_w(y,u,w)$ (see (16)), in which in our example $G(y,u,w) = -w$. Thus, exploiting that E_y depends continuously on z in our example and $\alpha_k \le 1$, we estimate

$$\begin{aligned} &\lim_{\rho\downarrow 0} \|(M_\rho - M(z_k))s_k\|_{Z^*} \\ &\le \gamma \lim_{\rho\downarrow 0} \|\partial_C P_{U_{ad}}(\gamma(w_k+\rho s_{k,3})) - \partial_C P_{U_{ad}}(\gamma w_k)\|_{\mathcal{L}(Z,Z^*)}\|s_{k,3}\|_Z \\ &\le \frac{1}{\lambda} C_{M^{-1}}\|f_k\|_{Z^*}, \end{aligned}$$

where we use (2) in the last estimate. In case of $\|\partial_C P_{U_{ad}}(\gamma(w_k+\rho s_{k,3})) - \partial_C P_{U_{ad}}(\gamma w_k)\|_{\mathcal{L}(Z,Z^*)} = 0$, Assumption 3.1 (a) holds immediately with $\tilde{\varepsilon} = 0$. However, once the subdifferential $M(z_k)$ has been fixed at a point of non-differentiability, $\partial_C P_{U_{ad}}(\gamma w_k) \in (0,1]$ (or $\in [0,1)$ alternatively), it may happen that M_ρ can only be chosen such that $\partial_C P_{U_{ad}}(\gamma w_k) = 0$ (or $= 1$ otherwise) at a point of differentiability. In the latter situation $\|\partial_C P_{U_{ad}}(\gamma(w_k+\rho s_{k,3})) - \partial_C P_{U_{ad}}(\gamma w_k)\|_{\mathcal{L}(Z,Z^*)} \in (0,1]$ and this does not vanish as ρ tends to zero from above. However the constant $C_{M^{-1}}$ is uniform and for sufficiently large $\lambda = 1/\gamma$, we may guarantee Assumption 3.1 (a). This estimate is also supported by numerical evidence, where we observe always descent directions, unless we reduce λ significantly.

For the approximation of the state as well as the control, piecewise linear finite elements are used. In this context, let $\mathcal{T}_h$ be a triangulation of Ω for which the usual regularity assumptions hold. Furthermore, let $h := \max_{T\in\mathcal{T}_h}\operatorname{diam}(T)$ denote the size of the mesh. For details we refer to [2]. Analogously for a finite element discretization the well-posedness of the Newton step and the discretized problem

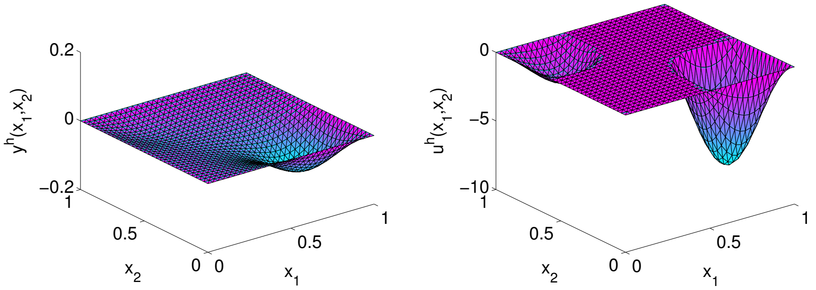

FIGURE 1. Discrete solution of (P2) for $h = 1/64$. Left-hand side: Optimal state $y^h(x_1, x_2)$ on x_3 axis vs. x_1 and x_2. Right-hand side: Optimal control $u^h(x_1, x_2)$ on x_3 axis vs. x_1 and x_2.

can be derived for both examples [10, 9], together with a mesh independence result for the local algorithm under the further assumption that the set where the strict complementarity is violated has measure zero. As a starting point we use $z_0 = 0$. Our stopping condition is $\|f(z_k)\|_{Z^*} \leq 10^{-10}$. Table 1 shows the iteration history for $h = 1/256$. In iterations 1 to 17 the step sizes are chosen with $\alpha_k < 1$ due to the Armijo line search. As predicted by Theorem 3.5, the acceptance of $\alpha_k = 1$ by the Armijo line search leads to the transition of Algorithm 3.3 into Algorithm 2.2. The expected superlinear convergence is confirmed by the evolution of the values $\|f(z_k)\|_{Z^*}$ and $\|s_k\|_Z$. We observe that $\underline{\alpha}$ is uniformly bounded below and corresponding to Th. 3.4 (a) and Th. 3.5 the numerical accumulation point is a solution. Figure 1 shows the discrete solution of the optimal state y^h and the optimal control u^h for $h = 1/64$.

5.2. **Semilinear elliptic PDE with state constraints.** Problem (P2) has illustrated the performance of our algorithm. However, problems of the type of (P2) have been solved so far, see e.g. [10, 12, 13]. Hence we close with an example for a semilinear elliptic PDE with state constraints [9, Ch. 6] where, to our knowledge, previous globalization strategies cannot be applied:

$$\begin{aligned}
&\text{Minimize} && J(y,u) := \frac{1}{2}\,\|y - y_d\|^2_{L^2(\Omega)} + \frac{\lambda}{2}\,\|u\|^2_{L^2(\Omega)} \\
&\text{w. r. t.} && (y,u) \in H^1_0(\Omega) \times L^2(\Omega) \\
&\text{subject to} && E(y,u) := -\Delta y + y^3\,\exp(10\,y) + y - u = 0 \;\text{ in } \Omega := (0,1)^2, \\
&&& -10^{-2} \leq y(x) \leq 0 \;\text{ a.e. in } \Omega.
\end{aligned} \tag{P3}$$

We would like to track the state $y_d := \frac{1}{2}\cos(\pi\,x_1)\,\cos(\pi\,x_2)\,\exp(x_1)$. As Tikhonov parameter we consider $\lambda := 10^{-4}$. We use a Lavrentiev regularization of the state-constraint in (P3)

$$-10^{-2} \leq \epsilon\,u + y \leq 0,$$

where $\epsilon > 0$ is a sufficiently small parameter, yielding a mixed control-state constraint.

Note that we consider the same objective functional as in (P2), but in addition to the regularized state constraints, we allow for a term $y^3\exp(10y)$ instead of y^3

k	α	$\|f(z_k)\|_{Z^*}$	$\|s_k\|_{Z^*}$
0	–	7.59736E+05	–
1	1.00000E+00	1.14024E+05	1.93458E+03
2	1.00000E+00	3.61620E+04	7.83427E+02
3	1.00000E+00	1.59280E+04	1.62132E+03
$\vdots$			
9	2.50000E-01	3.03640E-02	1.48894E-01
10	1.00000E+00	9.69843E-03	3.23249E-02
11	1.00000E+00	2.42234E-05	9.90030E-06
12	1.00000E+00	3.15754E-06	2.56947E-07
13	1.00000E+00	1.14583E-07	1.59876E-09
14	1.00000E+00	1.70426E-13	5.17916e-13

TABLE 2. Iteration history for the solution of problem (P3) for $h = 1/128$. Step size α_k, norm $\|f(z_k)\|_{Z^*}$ and norm of the search direction $\|s_k\|_Z$ for the k-th iterate. We observe transition to local superlinear convergence.

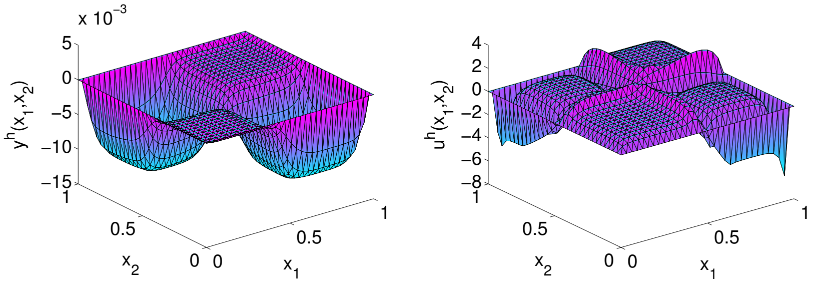

FIGURE 2. Discrete solution of (P3) for $h = 1/32$. Left-hand side: Optimal state $y^h(x_1, x_2)$ on x_3 axis vs. x_1 and x_2. Right-hand side: Optimal control $u^h(x_1, x_2)$ on x_3 axis vs. x_1 and x_2.

in the PDE. However, the mixed control-state constraint yields a slightly different setting as in Sect. 4. Concerning the validation of Assumption 2.3 and Assumption 3.1 (a) see the discussion for the first example. However, with respect to Assumption 3.1 (a) we have to keep in mind that the Lavrentiev regularization parameter ϵ and the Tikhonov parameter λ should not be chosen independently.

Again, we work with the stopping condition $\|f(z_k)\|_{Z^*} \leq 10^{-10}$. For numerical results for $\epsilon = 10^{-3}$ as Lavrentiev regularization parameter, see Table 2 and Fig. 2.

Acknowledgments. The authors acknowledge gratefully the useful suggestions and comments of Prof. M. Ulbrich, Prof. A. Schiela, and an unknown referee.

REFERENCES

[1] J. F. Bonnans and A. Shapiro, *Perturbation Analysis of Optimization Problems*, Springer Series in Operations Research, Springer, New York, 2000.

[2] E. Casas, Using piecewise linear functions in the numerical approximation of semilinear elliptic control problems, *Adv. Comput. Math.*, **26** (2007), 137–153.

[3] E. Casas and F. Tröltzsch, Second order optimality conditions and their role in PDE control, *Jahresber. Dtsch. Math.-Ver.*, **117** (2015), 3–44.

[4] X. Chen, Z. Nashed and L. Qi, Smoothing methods and semismooth methods for nondifferentiable operator equations, *SIAM J. Numer. Anal.*, **38** (2000), 1200–1216.

[5] R. Correa and A. Joffre, Tangentially continuous directional derivatives in nonsmooth analysis, *J. Optim. Theory Appl.*, **61** (1989), 1–21.

[6] M. Gerdts, Global convergence of a nonsmooth Newton's method for control-state constrained optimal control problems, *SIAM J. Optim.*, **19** (2008), 326–350; M. Gerdts and B. Hüpping, Erratum: Global convergence of a nonsmooth Newton's method for control-state constrained optimal control problems, Technical report, Universität der Bundeswehr München, Neubiberg (2011). Available online: `http://www.unibw.de/lrt1/gerdts/forschung/publikationen/erratum-siam-19-1-2008-326-350-full.pdf`.

[7] M. Gerdts and M. Kunkel, A nonsmooth Newton's method for discretized optimal control problems with state and control constraints, *J. Ind. Manag. Opt.*, **4** (2008), 247–270.

[8] M. Hintermüller, K. Ito and K. Kunisch, The primal-dual active set strategy as a semismooth Newton method, *SIAM J. Optim.*, **13** (2003), 865–888.

[9] M. Hintermüller, F. Tröltzsch and I. Yousept, Mesh-independence of semismooth Newton methods for Lavrentiev-regularized state constrained nonlinear optimal control problems, *Numer. Math.*, **108** (2008), 571–603.

[10] M. Hintermüller and M. Ulbrich, A mesh-independence result for semismooth Newton methods, *Math. Program., Ser. B*, **101** (2004), 151–184.

[11] M. Hinze, R. Pinnau, M. Ulbrich and S. Ulbrich, Optimization with PDE Constraints, Math. Modelling: Theory and Applications, Vol. 23, Springer, New York, 2009.

[12] M. Hinze and M. Vierling, The semi-smooth Newton method for variationally discretized control constrained elliptic optimal control problems; implementation, convergence and globalization, *Optim. Methods Softw.*, **27** (2012), 933–950.

[13] M. Hinze and M. Vierling, A globalized semi-smooth Newton method for variational discretization of control constrained elliptic optimal control problems, in *Constrained Optimization and Optimal Control for Partial Differential Equations* (eds. G. Leugering et al.), Int. Ser. Numer. Math., 160, Birkhäuser/Springer, Basel, 2012, 171–182.

[14] S. Horn, *Fixpunktiterationsverfahren für PDE-restringierte Optimalsteuerungsverfahren*, Master thesis, Universität der Bundeswehr München, Neubiberg, 2012.

[15] K. Ito and K. Kunisch, Applications of semi-smooth Newton methods to variational inequalities, in *Control of Coupled Partial Differential Equations* (eds. K. Kunisch, G. Leugering, J. Sprekels and F. Tröltzsch), Internat. Ser. Numer. Math., 155, Birkhäuser, Basel, 2007, 175–192.

[16] K. Ito and K. Kunisch, On a semi-smooth Newton method and its globalization, *Math. Program., Ser. A*, **118** (2009), 347–370.

[17] A. Kröner, K. Kunisch and B. Vexler, Semismooth Newton methods for optimal control of the wave equation with control constraints, *SIAM J. Control Optim.*, **49** (2011), 830–858.

[18] B. Kummer, Newton's method for non-differentiable functions, in *Advances in Mathematical Optimization* (eds. J. Guddat, et al.), Math. Res., 45, Akademie-Verlag, Berlin, 1988, 114–125.

[19] B. Kummer, Newton's method based on generalized derivatives for nonsmooth functions: Convergence analysis, in *Advances in Optimization (Lambrecht 1991)* (eds. W. Oettli and D. Pallaschke), Lecture Notes in Econom. and Math. Systems, 382, Springer, Berlin, 1992, 171–194.

[20] A. Rösch and D. Wachsmuth, Semi-smooth Newton's method for an optimal control problem with control and mixed control-state constraints, *Optim. Methods Softw.*, **26** (2011), 169–186.

[21] A. Schiela, A simplified approach to semismooth Newton methods in function space, *SIAM J. Optim.*, **19** (2008), 1417–1432.

[22] M. Ulbrich, *Nonsmooth Newton-like Methods for Variational Inequalities and Constrained Optimization Problems in Function Spaces*, Habilitation thesis, Technical University Munich, München, 2001.

[23] M. Ulbrich, Semismooth Newton methods for operator equations in function spaces, *SIAM J. Optim.*, **13** (2003), 805–842.

[24] M. Ulbrich, *Semismooth Newton Methods for Variational Inequalities and Constrained Optimization Problems in Function Spaces*, MOS-SIAM Series on Optimization, 11, SIAM/MOS, Philadelphia, 2011.

Received May 2014; 1st revision April 2015; 2nd revision August 2015.

E-mail address: matthias.gerdts@unibw.de
E-mail address: stefanhorn87@aol.com
E-mail address: sven-joachim.kimmerle@unibw.de

Chapter 3

Optimal Control of Coupled Ordinary and Partial Differential Equations

In this chapter we bring together the results known separately for ordinary and partial differential equations, respectively. Of course, we cannot expect better results for optimal control of coupled ODE and PDE than for optimal control of ODE and for optimal control of PDE alone.

This new contribution provides the fundament for the numerical solution of real-world problems in the next chapter. We consider the approach to treat ODEs like PDEs in a Hilbert space setting versus to treat PDEs as ODEs in function spaces. Though under reasonable assumptions both approaches coincide, it turns out in applications (see Chapter 4) to prefer the strategy "treat ODE as PDE" for conceptual and numerical reasons.

3.1 Modelling of Coupling and Averaging-Evaluation Operators

In a fully coupled problem a PDE solution $y_1(t,x)$ (in time and space) may enter directly into the ODE (in time), but y_1 depends on x and thus the ODE has x as a parameter. In many contexts this makes no sense from a modelling point of view, see, e.g., the elastic crane-trolley-load example in Section 4.3 or the quarter car model in Section 4.4. Furthermore from a technical point of view, we are not interested in a family of ODEs that have the spatial point x as parameter here. Thus we consider here the case that some spatial average over the PDE solution y_1 or some point evaluation of y_1 enters the ODE for y_2.

As before we consider the time interval $I = (0, t_f)$ with $t_f > 0$ and for the spatial coordinate $\Omega \subset \mathbb{R}^d$, $d \in \mathbb{N}$, open, bounded.

Definition 3.1 *(Averaging-Evaluation Operator)*
Let Y_1 be a Banach space and $y_1 \in Y_1$ such that $y_1 : \Omega_{t_f} = I \times \Omega \to \mathbb{R}^{n_{y_1}}$. We call a linear operator $\mathcal{E} : Y_1 \to \mathcal{L}(I, \mathbb{R}^{n_{y_1}})$ with $y_1(t,x) \mapsto (\mathcal{E}y_1)(t)$, a (spatial) averaging-evaluation operator.

Example 3.2 *(Averaging Operator)*
Consider a subset $A \subset \Omega$ or $A \subset \partial\Omega$ with a corresponding measure dA on A. We denote $|A| := \int_A 1\, dA(x)$ and assume $|A| > 0$. Let $Y_1 = L^p(\Omega_{t_f}) \cap L^\infty(I, L^p(A))$ with $1 \leq p \leq \infty$. The operator

$$(\mathcal{E}y_1)(t) := \fint_A y_1(t,x)\, dA(x) := \frac{1}{|A|} \int_A y_1(t,x)\, dA(x) \tag{3.1}$$

averages y_1 over x on A. For instance, if $\Omega \subset \mathbb{R}^3$ and $A \subset \partial\Omega$, then dA is a surface measure.

Averaging operators appear in our model for the elastic crane-trolley-load problem and in the quarter car model that may be rewritten such that an averaging operator over a boundary part is used, see Eq. (4.10).

Note that this operator is different to the so-called Steklov averaging technique [OR79, §8, 2.], since we do not consider an operator averaging locally around every space point.

Typically, an averaging operator has a smoothing property. It can be approximated by a convolution with a mollifier, see [OR11, Sect. 2.10], yielding an infinitely smooth approximation.

Example 3.3 *(Evaluation Operator)*
Let $Y_1 \hookrightarrow L^p(I, C^0(\Omega_{t_f}))$, $1 \leq p \leq \infty$. Another operator in the sense of Def. 3.1 is obtained by a linear combination of evaluations of y_1 at certain points $a, b \in \Omega$, e.g.

$$(\mathcal{E}y_1)(t) := y_1(t,b) - y_1(t,a). \tag{3.2}$$

This operator may also be interpreted as a generalized version of a trace operator, see Th. A.21.

This operator appears naturally in the truck-container example, see Section 4.2, where $a = 0$ and $b = L$ are boundary points and we evaluate at boundary values. In the quarter car model an evaluation operator is introduced by Eq. (4.9) as the restriction to a boundary part.

First of all we ask the question what is known on the analytic well-posedness of coupled ODE-PDE systems.

3.2 Analytic Results for Coupled Ordinary and Partial Differential Equations

We start with a quite general strategy for local existence and uniqueness of fully coupled systems due to the Banach fixed point theorem (Th. A.3). It has been applied to coupled systems: in [Ki09, Ki11] to a parabolic, an elliptic and a free boundary evolution equation, in [KG16] to a parabolic system and an ODE system, and in [KGH18a] to an elliptic system and an ODE system. In [Ki09, Ki11] the local result is extended to global well-posedness using uniform bounds that can be derived for this specific problem. Here we focus on the case of a parabolic PDE.

Theorem 3.4 *(A Local Existence and Uniqueness Result for Coupled ODE-PDE Systems)*
Let $Y = Y_1 \times Y_2$ be the state space, Y_i a separable (real) Hilbert space, and $u_i \in U_i$, U_i a (real)

Hilbert space, $i = 1, 2$. We assume the PDE to be parabolic of second order and a Gelfand triple structure $V \overset{cd}{\hookrightarrow} H \overset{cd}{\hookrightarrow} V^*$ (see Def. A.17) with $V = Y_1$ and H a separable Hilbert space. The control space is $U = U_1 \times U_2$. We consider an operator $E_1 : Y \times U_1 \to W_1$ for a parabolic PDE , living on $\Omega_{t_f} = (0, t_f) \times \Omega$, $\Omega \subset \mathbb{R}^d$ open, bounded, and $E_2 : Y \times U_2 \to W_2$ for a first-order ODE with a time evolution on $(0, t_f)$ (where $t_f > 0$), i.e.

$$E_1(y, u_1) = \dot{y}_1 - F_1(y, u_1) = 0_{W_1}, \tag{3.3}$$

$$E_2(y, u_2) = \dot{y}_2 - F_2(y, u_2) = 0_{W_2}, \tag{3.4}$$

where $\dot{y}_i \in W_i$, $i = 1, 2$, for the time derivative of the states and F_i, $i = 1, 2$, is an operator $F_i : Y \times U_2 \to W_2$. We suppose this is complemented with initial and boundary data that fulfil sufficient regularity assumptions that are encoded in the operators E_1, E_2 and in the state space.

We make the strong assumption $(*)$ that (i) the PDE for itself is well-defined for given ODE state $y_2 \in \tilde{Y}_2$ and for every $u_1 \in U_1$ and (ii) the ODE for itself is well-defined for given PDE state $y_1 \in \tilde{Y}_1$ and for every $u_2 \in U_2$. Here $\tilde{Y} = \tilde{Y}_1 \times \tilde{Y}_2$ with $\tilde{Y}_i \subset Y_i$, $i = 1, 2$, are suitable subspaces, typically we find $\tilde{Y}_1 = Y_1 \cap L^\infty(\Omega_{t_f})$ and $\tilde{Y}_2 = Y_2 \cap L^\infty(I)$. Assume (3.3) and (3.4) yield the estimates

$$\|y_1\|_{\tilde{Y}_1} \le c_1 \left(\|y_{1;0}\|_{H_1} + t_f^\kappa \|y_2\|_{\tilde{Y}_2} + \|u_1\|_{U_1} \right), \tag{3.5}$$

$$\|y_2\|_{\tilde{Y}_2} \le c_2 \left(\|y_{2;0}\|_{H,2} + \|y_1\|_{\tilde{Y}_1} + \|u_2\|_{U_2} \right), \tag{3.6}$$

with $\kappa > 0$ and some constants c_i, $i = 1, 2$, independent of t_f. Note that the last two assumptions are related to the semilinearity of F_1 and F_2, i.e. $\dot{y}$ does not enter into F_1 and neither $\dot{y}$ nor $y''_{1,x}$ does enter into F_2.

Moreover, for two states $y_i^{(k)}$, $i, k = 1, 2$ with given different right-hand side states $\tilde{y}_i^{(k)}$ but identical $y_{i;0}$ and u_i we assume that the equations (3.3) and (3.4) yield the following estimates of Lipschitz-type

$$\|y_1^{(1)} - y_1^{(2)}\|_{\tilde{Y}_1} \le c_1^L t_f^\kappa \|y_2^{(1)} - y_2^{(2)}\|_{\tilde{Y}_2}, \tag{3.7}$$

$$\|y_2^{(1)} - y_2^{(2)}\|_{\tilde{Y}_2} \le c_2^L \|y_1^{(1)} - y_1^{(2)}\|_{\tilde{Y}_1} \tag{3.8}$$

with some constants c_i^L, $i = 1, 2$, independent of t_f.

If $u_1 \in U_1$ (typically a control) and initial values $y_{1;0} \in H$ and $y_{2;0} \in \mathbb{R}^n$ are given, then for sufficiently small terminal times $t_f > 0$ there exists a unique solution of the coupled system (3.3) & (3.4) in the space $\tilde{Y}$.

Note that the strong assumption $(*)$ on the well-posedness of the ODE and PDE, resp., and the validity of (3.5) – (3.8) has to be checked for every problem. Thus this theorem is of limited practical use, but it exhibits the general coupling structure that is exploited in several applications within this study. At the end of this section we discuss examples, where this strategy has been applied.

We remark that in our notation $y_{1;0}$ denotes the prescribed initial value for the PDE state y_1 at time $t = 0$, whereas $y_{1,1}$ would be the first component of a vector-valued PDE state.

Derivatives are denoted like $y'_{1;t}$, being the partial time derivative of y_1.

Proof. The assumptions and the estimates (3.5) & (3.6) say that the PDE has a solution in $\tilde{Y}_1$ for given suitable data of the ODE as well as the ODE has a solution in $\tilde{Y}_2$ for given suitable data of the PDE. Thus both uncoupled solution operators $\check{E}_1^{-1} : \tilde{Y}_2 \times U \to \tilde{Y}_1, [y_2, u_1] \mapsto y_1$ and $\check{E}_2^{-1} : \tilde{Y}_1 \times U_2 \to \tilde{Y}_2, [y_1, u_2] \mapsto y_2$ are well-defined but only since we have replaced Y_i by $\tilde{Y}_i$, $i = 1, 2$.

Indeed, by plugging in the solution of the second differential equation into the first (or vice versa) successively, starting, e.g., with the initial data $y_i^{(0)} = y_{i;0}$, $i = 1, 2$, this defines the fixed point iteration[1]

$$M := \begin{bmatrix} y_1^{(k)} \\ y_2^{(k)} \end{bmatrix} \to \begin{bmatrix} y_1^{(k+1)} \\ y_2^{(k+1)} \end{bmatrix} = \begin{bmatrix} \check{E}_1^{-1}(\check{E}_2^{-1}(y_1^{(k)}, u_2), u_1) \\ \check{E}_2^{-1}(y_1^{(k)}, u_2) \end{bmatrix}, \quad k \in \mathbb{N}. \tag{3.9}$$

In order to apply the Banach fixed point theorem we have to show that the mapping

$$M : \tilde{Y} \to \tilde{Y},$$

is actually a self-mapping on $\tilde{Y}$ and that this mapping is strictly contractive.

We consider the differences $y_\Delta^{(k+1)} := y_1^{(k+1)} - y_1^{(k)}$, $k \in \mathbb{N}$, that solve the coupled equations with zero initial and boundary data. Applying (3.7) and (3.8) to the iteration defined by (3.9), we find the estimate

$$\|y_\Delta^{(k+1)}\|_{\tilde{Y}_1} \leq c_1^L c_2^L t_f^\kappa \|y_\Delta^{(k)}\|_{\tilde{Y}_1}. \tag{3.10}$$

W.l.o.g. $c_1^L c_2^L \neq 0$, otherwise there is nothing to prove. Together with the corresponding estimate for $y_2^{(k+1)} - y_2^{(k)}$ that is obtained analogously, choosing $t_f < 1/(c_1^L c_2^L)^{1/\kappa}$ gives the desired strict contraction.

When considering only a solution $y_1^{(k)}$ instead of differences $y_\Delta^{(k)}$ of the solution in iteration k, from (3.5) and (3.6) follows

$$\|y_1^{(k+1)}\|_{\tilde{Y}_1} \leq c_1 c_2 t_f^\kappa \|y_1^{(k)}\|_{\tilde{Y}_1},$$

that proves together with the analogous estimate for y_2 the self-mapping. The Banach fixed point theorem yields not only the existence of a fixed point, due to the construction of the fixed point iteration M the uniqueness follows directly. For further details on the technique of proof see [Ni99] or [Ki09]. □

This theorem may be adapted to elliptic PDE or coupled elliptic and parabolic PDE. Note that the derivation of the presupposed estimates can be quite tricky for certain problems, see, for instance, [KG16, Sect. 3] for the truck-container problem or [KGH18a, Subsect. 3.3] for the elliptic elastic crane-trolley-load problem. Typically, we may use the following crucial estimate for a first-order ODE of the type $\dot{y}_2 = F_2(y, u_2)$ on the time interval $(0, t_f)$ in order to obtain the factor t_f^κ in the second estimate. Let $Y_2 = W^{1,p}(0, t_f)$, $1 < p < \infty$, thus due to the embedding

[1] Note that this is similar to the classical Picard iteration known for ODE in integral form.

in 1D the typical candidate for the space $\tilde{Y}_2 = Y_2 \cap L^\infty(0,t_f)$ allows for the choice $\tilde{Y}_2 = Y_2$. Moreover, by the Hölder inequality

$$|y_2(t) - y_{2;0}| = \left| \int_0^t \dot{y}_2 \, dt \right| \leq t^\kappa \, \|\dot{y}_2\|_{L^{p'}(0,t_f)} \leq t_f^\kappa \, \|\dot{y}_2\|_{L^{p'}(0,t_f)}, \tag{3.11}$$

for any $\kappa = 1/p < 1$, where $p' = p/(p-1) = 1/(1-\kappa)$ is the dual exponent to p. We set $p = 2$ yielding $Y_2 = H^1(0,t_f)$. The question remains to ensure $\|y_1\|_{Y_1} \leq c_1(\|y_{1;0}\|_H + \|y_2 - y_{2;0}\|_{L^2(0,t_f)} + \|u_1\|_{U_1})$ for (3.3), then we exploit (3.11) on the right-hand side.

In addition, typically in this situation we expect to find

$$\|\dot{y}_2\|_{L^2(0,t_f)} = \|F_2(y,u_2)\|_{L^2(0,t_f)} \leq c_2 \left(\|y_1\|_{Y_1} + \|y_2\|_{L^2(0,t_f)} + \|u_2\|_{U_2} \right)$$

for the ODE, then the term with y_2 on the right-hand side may be estimated by (3.11) and absorbed due to the smallness of t_f and (3.6) follows.

Another tool for deriving estimates of the type (3.5) – (3.8) for evolution equations is the Gronwall inequality (Lemma A.11).

Though t_f might be arbitrarily small, in many specific problems we may either demonstrate global bounds on the states or choose suitable controls in U_{ad} such that no blow up may happen in finite time. Then we can extend the time interval successively again and again and obtain global existence and uniqueness of the states up to a prescribed terminal time t_f. However, this cannot be always the case, since for certain ODEs, e.g., a blow up in finite time is generic and bounds on the control might prevent us from avoiding the blow up.

In case of certain averaging-evaluation operators, e.g. as in [KGH18a], another "internal" fixed point iteration over the averaging procedure (see, e.g., [KGH18a, Th. 3.2]) might be in order before applying our "external" fixed point iteration (3.9). Here E_1 is an elliptic PDE where time enters as a parameter.

Furthermore, our fixed point strategy is applied for an existence and uniqueness result, [KG16, Th. 3.2], for the coupled ODE-PDE problem arising in the truck-container problem (see Sect. 4.2 for a reprint of [KG16]). Using the notation of this work, in this example we have the state spaces

$$Y = W(I; L^2, H^1) \times W(I; L^2, H_0^1) \times [H^2(0,1)]^2 \times [H^1(0,1)]^2$$

and

$$\begin{aligned}\tilde{Y} = {} & [W(I; L^2, H^1) \cap L^\infty(0,1; H^1(0,L))] \\ & \times [W(I; L^2, H_0^1) \cap L^\infty(0,1; H_0^1(0,L))] \times [H^2(0,1)]^2 \times [H^1(0,1)]^2,\end{aligned}$$

the control spaces $U = U_1 = L^2(0,1)$ and $U_{ad} \subset U$, the latter containing standard box constraints for the control. The Gelfand triple reads V^*, H, V with $V = Y_1$ and $H = L^2(0,L)$. Note that $I = (0,t_f)$ and $\Omega = (0,L)$. The operators E_i, $i = 1,2$, correspond to the Saint-Venant equations [KG16, (2.3) – (2.8)] that are parabolic and the Newton dynamics [KG16, (2.9) –

(2.11)]. For instance, the estimates [KG16, (3.3) & (3.4)] are the crucial estimate corresponding to (3.10) for the ODE.

Another example (not included in this work) for demonstrating existence and uniqueness by our fixed-point strategy can be found in [Ki09, Th. 4.7], where a parabolic PDE, an elliptic PDE and a single ODE for the time evolution of R are fully coupled. In this GaAs droplet problem the crucial estimate follows from the so-called Stefan condition for a free radius R. In this case the existence and uniqueness may be extended globally in time [Ki09, Th. 4.8], since uniform bounds in t_f are available.

3.3 Combined Results for Optimal Control of Coupled ODE-PDE Systems

We combine the abstract approaches presented in Sections 2.6 and 2.7 and adapt this for coupled ODE-PDE systems. It turns out that it makes sense to treat ODEs as PDEs, yielding a (possibly fully) coupled PDE system, but with some peculiarities. On the other hand PDEs might be treated as ODEs in suitable function spaces, see Subsection 3.3.3. In this chapter we denote by $y = [y_1, y_2]$ PDE states y_1 and ODE states $y_2 (= q)$ together. The main idea is to consider ODEs, subject to initial value problems, in spaces similar to the typical spaces, used for PDE-constrained optimization with parabolic PDE or elliptic PDE of first-order. However, we annotate that it is not straightforward, how to consider DAE as PDE (and *vice versa*) in general. It might be an option to establish a relation between an algebraic equation and a degenerated elliptic PDE of second-order.[2]

3.3.1 Treat ODEs as PDEs

Typically, we think of $Y = Y_1 \times Y_2$ being a Hilbert space (separable) and $\tilde{Y} = Y \cap L^\infty$ and consider controls $U_1 = L^2$ in space and/or time and $U_2 = L^2$ in time only, and $U_{ad} = U_{ad,1} \times U_{ad,2}$ representing box constraints ($U_{ad,i}$, $i = 1, 2$ defined analogously as in (2.42)). In particular, $\tilde{Y}$ is densely embedded in Y. Our problem reads

Problem 3.5 *(Constrained Optimal Control Problem for Coupled ODE-PDE Systems) Solve*

$$\begin{aligned} &\min_{[y,u] \in Y \times U} \mathcal{J}(y, u) \\ \textit{subject to } &E_1(y, u_1) = 0_{W_1}, \\ &E_2(y, u_2) = 0_{W_2}, \\ &u \in U_{ad} \subset U = U_1 \times U_2. \end{aligned} \tag{3.12}$$

[2] Note that for the (Navier-)Stokes system the incompressibility equation (i.e. the PDE for the pressure) may be interpreted as an algebraic equation as well.

with $E_1 : \tilde{Y} \times U_1 \to W_1$ and $E_2 : \tilde{Y} \times U_2 \to W_2$. Let Y, U, $W = W_1 \times W_2$ be Hilbert spaces and the Banach spaces $\tilde{Y}_i$ are densely embedded in Y_i, $i = 1, 2$. U_{ad} is a closed convex subset of U. The objective $\mathcal{J} : Y \times U \to \mathbb{R}$, is assumed to be F-differentiable in a neighbourhood (w.r.t. the $Y \times U$ topology) of $[\hat{y}, \hat{u}]$. This Fréchet derivative $\mathcal{J}'$ is assumed to be locally Lipschitz continuous.

The system of differential equations is denoted by the abstract operator equations

$$E(y, u) = \begin{bmatrix} E_1(y, u_1) \\ E_2(y, u_2) \end{bmatrix} = 0_W, \tag{3.13}$$

for $E : \tilde{Y} \times U \to W$, where E_1 represents a PDE and E_2 represents an ODE that may be fully coupled. Note that E and E_i, $i = 1, 2$, are considered on $\tilde{Y}$ and not on Y. E is assumed to be F-differentiable at $[\hat{y}, \hat{u}]$, in particular $E'_y(\hat{y}, \hat{u}) \in \mathcal{L}(\tilde{Y}, W)$. Note that E'_y might be extended to a densely defined operator $\tilde{G} = [\tilde{G}_1, \tilde{G}_2]^\top$ with domain in $Y = Y_1 \times Y_2$.

Note that F-differentiability is quite restrictive for differential equations. For optimization with PDE usually semilinear equations are considered, whereas for optimization with ODE nonlinear equations are common. For fully nonlinear ODE F-differentiability in spaces like L^2 or H^1 cannot be guaranteed in general. Concerning objectives, we have that linear-quadratic functionals in Hilbert spaces are F-differentiable.

Again, we assume that Assumption 2.71 holds. Please note that we do not require that $E'(\hat{y}, \hat{u}) : \tilde{Y} \times U \to W$ is surjective.

Without considering a specific structure of the ODE-PDE system, we can state for the coupled optimal control problem only the same result as for a general PDE. Only here we follow the approach from Ito and Kunisch [IK08], see Th. 2.73. Again, under these assumptions we may prove the existence of a Lagrange multiplier[3] w.r.t. the equality constraints:

Theorem 3.6 *(First-Order Necessary Optimality Conditions for Coupled ODE-PDE Systems) Let $[\hat{y}, \hat{u}]$ be a local minimizer of Problem 3.5 and let the assumptions there and Assumption 2.71 hold, then there exists a Lagrange multiplier $\lambda \in W_1^* \times W_2^*$ that fulfils*

$$\begin{aligned} E_1(\hat{y}, \hat{u}_1) &= 0_{W_1}, \\ E_2(\hat{y}, \hat{u}_2) &= 0_{W_2}, \\ \tilde{G}_1^* \lambda_1 + \mathcal{J}'_{y_1}(\hat{y}, \hat{u}) &= 0_{Y_1^*}, \\ \tilde{G}_2^* \lambda_2 + \mathcal{J}'_{y_2}(\hat{y}, \hat{u}) &= 0_{Y_2^*}, \\ \left\langle E'_{u_1}(\hat{y}, \hat{u}_1)^* \lambda_1 + \mathcal{J}'_{u_1}(\hat{y}, \hat{u}_1), u_1 - \hat{u}_1 \right\rangle_{U_1^*, U_1} &\geq 0 \qquad \forall u_1 \in U_{ad,1}, \\ \left\langle E'_{u_2}(\hat{y}, \hat{u}_2)^* \lambda_2 + \mathcal{J}'_{u_2}(\hat{y}, \hat{u}_2), u_2 - \hat{u}_2 \right\rangle_{U_2^*, U_2} &\geq 0 \qquad \forall u_2 \in U_{ad,2}. \end{aligned}$$

The proof follows directly from Theorem 2.73, if the solution space of the coupled system can be considered as a product in the framework $\tilde{Y}_i \subsetneq Y_i$, $i = 1, 2$, with dense embeddings, whereas

[3] Note that in general the space W^* might be a space of measure or even without this structure. Then equations involving multipliers cannot be exploited sensibly.

Y_i are Hilbert spaces.

We check the applicability of Assumption 2.71 a) – d) for one of our examples. We consider here the truck-container example [GK15, KG16]. For this problem the latter theorem, Th. 3.6, may be applied, though the proof in our paper [KG16] is slightly different. Indeed, here we work in contrary to Section 4.2 for the time interval $I = (0, 1)$ and the space interval $\Omega = (0, L)$ with

$$\begin{aligned}
Y_1 &= W(I; L^2, H^1) \times W(I; L^2, H_0^1) \times H^1(\Omega) \times H_0^1(\Omega),\\
Y_2 &= [H^2(I)]^2 \times [H^1(I)]^2 \times \mathbb{R}^4,\\
\tilde{Y} &= Y_1 \cap [L^\infty(I; H^1(\Omega)) \times L^\infty(I; H_0^1(\Omega)) \times L^\infty(\Omega) \times L^\infty(\Omega)] \times Y_2,
\end{aligned}$$

$U = L^2(I)$, and U_{ad} represents box constraints for the control (see (2.42)). Note that in 1D we have the embedding $H^1 \hookrightarrow L^\infty$. In particular, $\tilde{Y}_1 = Y_1 \cap [[L^\infty(\Omega_{t_f})^2] \times [L^\infty(\Omega)]^2$ is densely embedded in Y_1, being a separable Hilbert space, and thus $\tilde{Y} \overset{d}{\hookrightarrow} Y$.

We have to check that the adjoint solution is in the domain $D(\tilde{G}^*)$ and that E is F-differentiable with locally Lipschitz derivative (implying c) ⇒ d) in Assumpt. 2.71). In addition we need for the solvability of the state equations, that the solution depends $C^{0,1/2}$-continuous on u for sufficiently small perturbations τ_u (part c) of Assumpt. 2.71). Indeed, the adjoint system is linear and exhibits at least the same regularity as the nonlinear system for the states and E is F-differentiable w.r.t. $[y, u]$ with Lipschitz continuous derivative. Furthermore, the state equation has at least one solution and this solution depends Lipschitz continuously on u.

3.3.2 Optimal Control of Coupled ODE (Treated as PDE) and Parabolic PDE

We annotate that for the following problem we use known results for the more general situation of optimal control of reaction-diffusion systems [Ry16, CRT18], where Neumann b.c. (in space) for the PDEs are considered and distributed controls of the PDEs or controls of the r.h.s. of the ODEs, resp., are considered. In the case of Dirichlet boundary conditions (for the PDE) they could be incorporated into function spaces in principle.

Problem 3.7 *(Optimal Distributed Control of Coupled ODE (Treated as PDE) and Parabolic PDE)*
For the notation concerning the geometry and the parabolic PDE we refer to Example 2.69 . We consider PDE states $y_1 : \Omega_{t_f} \to \mathbb{R}^{n_{y_1}}$, ODE states $y_2 : I \to \mathbb{R}^{n_{y_2}}$, the distributed control $u_1 : \Omega_{t_f} \to \mathbb{R}^{u_1}$ for the PDE, and the control $u_2 : I \to \mathbb{R}^{n_{u_2}}$ for the right-hand side of the ODE that might be interpreted formally as a distributed control as well. We set $n_y = n_{y_1} + n_{y_2}$ and $n_u = n_{u_1} + n_{u_2}$.
For the operators in the PDE let $A_1 \in L^\infty(\Omega_{t_f}, \mathbb{R}^{n_{y_1} \times n_{y_1}})$, $B_1 \in L^\infty(\Omega_{t_f}, \mathbb{R}^{n_{y_1} \times n_{u_1}})$, let $C_1 : \Omega_{t_f} \times \mathbb{R}^{n_y} \to \mathbb{R}^{n_{y_1}}$ have Carathéodory component functions and for the initial data $y_{1;0} : \Omega \to \mathbb{R}^{n_{y_1}}$. For the ODE let $A_2 \in L^\infty(I, \mathbb{R}^{n_{y_2} \times n_{y_2}})$, $B_2 \in L^\infty(I, \mathbb{R}^{n_{y_2} \times n_{u_2}})$, let $C_2 : I \times \mathbb{R}^{n_y} \to \mathbb{R}^{n_{y_2}}$ have Carathéodory component functions, $y_{2;0} \in \mathbb{R}^{n_{y_2}}$, and $\mathcal{E}$ is an averaging-evaluation operator

as introduced in Def. 3.1.

We wish to find $[y,u] \in Y \times U$ *such that*

$$\mathcal{J}(y,u) = \Phi(y(0), y(t_f)) + \int_0^{t_f} \int_\Omega \phi_1(t,x,y,u_1)\, dx\, dt + \int_0^{t_f} \phi_2(t, \mathcal{E}y_1, y_2, u_2)\, dt,$$

where $\Phi : \mathbb{R}^{2n_y} \to \mathbb{R}$, $\phi_1 : \Omega_{t_f} \times \mathbb{R}^{n_y} \times \mathbb{R}^{n_{u_1}} \to \mathbb{R}$, *and* $\phi_2 : I \times \mathbb{R}^{n_y} \times \mathbb{R}^{n_{u_2}} \to \mathbb{R}$, *is minimized w.r.t.* $y \in Y$, $u \in U_{ad} := U_{ad,1} \times U_{ad,2} \subset U = U_1 \times U_2$ *with*

$$\begin{aligned} U_{ad,1} &:= \{u_1 \in U_1 \,|\, u_{1,min} \le u_1(t,x) \le u_{1,max}, \ \text{a.e. in } \Omega_{t_f}\}, \\ U_{ad,2} &:= \{u_2 \in U_2 \,|\, u_{2,min} \le u_2(t) \le u_{2,max}, \ \text{a.e. in } I\}, \end{aligned}$$

subject to the constraints

$$\begin{aligned} y'_{1;t} - k\Delta_x y_1 + A_1(t,x)y_1 &= B_1(t,x)u_1 + C_1(t,x,y) && \text{in } \Omega_{t_f}, \\ \partial_\nu y_1 &= 0 && \text{on } \Sigma_{t_f}, \\ y_1(0,\cdot) &= y_{1;0} && \text{on } \Omega, \\ y'_{2;t} + A_2(t)y_2 &= B_2(t)u_2 + C_2(t, \mathcal{E}y_1, y_2) && \text{in } I, \\ y_2(0) &= y_{2;0}. \end{aligned}$$

We introduce operators $E_1 : Y_1 \times U_1 \to W_1 \times L^2(\Omega)$ *and* $E_2 : Y_2 \times U_2 \to W_2 \times \mathbb{R}^{n_{y_2}}$ *by*

$$E_1 : y \times u_1 \mapsto \begin{bmatrix} k(\nabla w, \nabla y_1)_H + \langle w, y'_{1;t} + A_1(t,x)y_1 - B_1(t,x)u_1 - C_1(t,x,y)\rangle_{W_1^*,W_1} \\ y_1(0,\cdot) - y_{1;0} \end{bmatrix}, \quad (3.14)$$

$$E_2 : y \times u_2 \mapsto \begin{bmatrix} y'_{2;t} + A_2(t)y_2 - B_2(t)u_2 - C_2(t, \mathcal{E}y_1, y_2) \\ y_2(0) - y_{2;0} \end{bmatrix}, \quad (3.15)$$

where the PDE holds in weak formulation for all $w \in W_1^*$ *and the ODE holds for almost all* $t \in I$.

Here we consider $W_1 := W(I)^*$ *where* $W(I) := W(I;H,V) = \{y_1 \in [L^2(I;H^1(\Omega))]^{n_{y_1}} \,|\, y'_{1;t} \in [L^2(I;H^1(\Omega)^*)]^{n_{y_1}}\}$ *with the Hilbert spaces* $H := [L^2(\Omega)]^{n_{y_1}}$ *and* $V = [H^1(\Omega)]^{n_{y_1}}$, *and* $W_2 := \mathbb{R}^{n_{y_2}}$. *We set* $Y = W(I) \times [H^1(I)]^{n_{y_2}}$. *A suitable state space is* $\tilde{Y} = [W(I) \cap [L^\infty(\Omega_{t_f})]^{n_{y_1}}] \times Y_2$. *This is a Banach space endowed with the norm*

$$\begin{aligned} \|y\|_{\tilde{Y}} := &\left(\sum_{j=1}^{n_{y_1}} \left(\|y_{1,j}\|^2_{L^2(I;H^1(\Omega))} + \|y'_{1,j;t}\|^2_{L^2(I;H^1(\Omega)^*)} \right) + \sum_{j=1}^{n_{y_2}} \|y_{2;j}\|^2_{H^1(I)} \right)^{1/2} \\ &+ \max_{j=1,\dots,n_{y_1}} \|y_{1;j}\|_{L^\infty(\Omega_{t_f})}. \end{aligned}$$

Note that $H^1(I) \hookrightarrow L^\infty(I)$ *since the time interval* I *is 1D.*

A standard example of this type of problem is the FitzHugh-Nagumo system for modelling neurons.

Example 3.8 *(FitzHugh-Nagumo System)*
Let $n_{y_1} = n_{y_2} = 1 = n_{u_1}$, $n_{u_2} = 0$, $k = 1$, $A_1 = 0$, $A_2 = \gamma_0$, and $B_1 = 1$. Let R be a cubic polynomial with derivative bounded from below. As coupling matrices consider here $C_1(t, x, y_1, y_2) = -R(y_1) - \gamma_1 y_2$ and $C_2(t, x, y_1, y_2) = \gamma_2 y_1 - \gamma_3$ with real numbers γ_i, $i = 1, \ldots, 3$. Under these assumptions, Problem 3.7 is the dedimensionalized form of the FitzHugh-Nagumo system. It serves as a model to describe the activation of neurons. Note that no spatial average over y_1 enters in C_2, yielding another ODE for every space point x. Thus, in this example, the ODE is actually a degenerated PDE.

This example exhibits a variety of solutions, including impulses in 1D, turning spirals in 2D, and scroll rings in 3D. For further analytic and optimal control results for this example see [CRT18] and the references therein.

Assumption 3.9 *(Assumptions for Existence and Uniqueness of States in Coupled ODE and Parabolic PDE)*
Let $\Omega \in \mathbb{R}^d$, $d = 1, 2, 3$, be a bounded domain with Lipschitz boundary[4] *and let $I = (0, t_f)$ with $t_f > 0$ fixed. Again, the parabolic cylinder is denoted by $\Omega_{t_f} = (0, t_f) \times \Omega$.*

For the operators introduced in Pb. 3.7 there holds $C_1(\cdot, 0, 0) \in [L^{\check{p}_1}(I; L^{\check{q}_1}(\Omega))]^{n_{y_1}}$ with $\check{p}_1, \check{q}_1 \in [2, \infty]$, $1/\check{p}_1 + d/(2\check{q}_1) < 1$ and $C_2(\cdot, \mathcal{E}0, 0) \in [L^{\check{p}_2}(I)]^{n_{y_2}}$ with $\check{p}_2 \in [2, \infty]$. For both operators C_i, $i = 1, 2$, we require that they are of class C^1 w.r.t. y and for the averaging-evaluation operator to be of class C^1 w.r.t. y_1.

We assume that for fixed y_j, $j = 1, 2$, there exist constants C_{N_i} and for all $M_i > 0$ constants C_{M_i}, $i = 1, 2, 3$, such that

$$\begin{aligned}
C'_{1;y_j}(t, x, y) &\leq C_{N_1} && \text{for a.a. } [t, x] \in \Omega_{t_f}, \\
\forall M_1 > 0 : |C'_{1;y_j}(t, x, y)| &\leq C_{M_1} && \text{for a.a. } [t, x] \in \Omega_{t_f}, \forall |y_j| \leq M_1, \\
C'_{2;y_j}(t, \mathcal{E}y_1, y_2) &\leq C_{N_2} && \text{for a.a. } t \in I, \\
\forall M_2 > 0 : |C'_{2;y_j}(t, \mathcal{E}y_1, y_2)| &\leq C_{M_2} && \text{for a.a. } t \in I, \forall |y_j| \leq M_2, \\
\mathcal{E}'_{y_1}(y_1) &\leq C_{N_3} && \text{for a.a. } t \in I, \\
\forall M_3 > 0 : |\mathcal{E}'_{y_1}(y_1)| &\leq C_{M_3} && \text{for a.a. } t \in I, \forall |y_1| \leq M_3.
\end{aligned}$$

Note that the latter two assumptions on $\mathcal{E}$ are fulfilled for any averaging operator $(\mathcal{E}y_1)(t) = ⨍_A y_1(t, x) dA(x)$, if A has positive measure, and for any evaluation operator $\mathcal{E}y_1 = y_1|_{x=a}$, if the function value at a is bounded.

The growth conditions of the latter assumption may be weakened, a polynomial growth in y and certain monotonicity conditions are actually sufficient, see [CRT18, Remark 2.1].

Instead of using the theory developed above for Problem 3.7, we work here with tailored results for the optimal control of reaction-diffusion systems [Ry16, CRT18].

[4] For $d = 1$ see the footnote in Ex. 2.68 a).

Theorem 3.10 *(Existence and Uniqueness for Coupled ODE and Parabolic PDE)*
Under Assumption 3.9 and if $y_0 \in [L^\infty(\Omega)]^{n_{y1}} \times \mathbb{R}^{n_{y2}}$, Problem 3.7 has a unique solution $y \in \tilde{Y}$ for every $u \in U := [L^{\tilde{p}_1}(I; L^{\tilde{q}_1}(\Omega))]^{n_{u1}} \times [L^{\tilde{p}_2}(I)]^{n_{u2}}$, where $\tilde{p}_1, \tilde{q}_1 \in [2, \infty]$ with $1/\tilde{p}_1 + d/(2\tilde{q}_1) < 1$ and $\tilde{p}_2 \in [2, \infty]$.

There holds the estimate

$$\|y\|_{\tilde{Y}} \leq c \left(\|y_{1;0}\|_{[L^\infty(\Omega)]^{n_{y1}}} + \|y_{2;0}\| + \|u\|_U + \|C_1(\cdot, 0, 0)\|_{[L^{\check{p}_1}(I; L^{\check{q}_1}(\Omega))]^{n_{y1}}} + \|C_2(\cdot, \mathcal{E}0, 0)\|_{[L^{\check{p}_2}(I)]^{n_{y2}}} \right)$$

with an constant c being independent from u.

If, in addition $y_{1;0} \in [C^0(\Omega)]^{n_{y1}}$, then we have $y \in [C^0(\Omega_{t_f})]^{n_{y1}} \times [C^0(I)]^{n_{y2}}$.

Proof. The proof given in [CRT18, Th. 2.1] relies on the Schauder fixed point theorem. It is translated to our special case, where $\mathcal{E}y_1$ enters into C_2. □

Note that our general results from Sect. 3.2 guarantee only existence and uniqueness local in time and thus the existence of optimal controls, following Subsect. 3.3.1, hold only for sufficiently short time intervals. However, a proof relying on the Banach fixed point theorem is constructive, whereas the Schauder fixed point theorem used in [CRT18] is not.

In the last theorem we have considered the elliptic operator $\mathcal{A}_1 := -k\Delta_x + A_1$. However, the proof may be extended to general elliptic operators $\mathcal{A}_1$ with essentially bounded coefficients.

Moreover, this result allows to define a control-to-state operator, that turns out to be differentiable under certain conditions. For details see [CRT18, Th. 2.2].

Assumption 3.11 *(Assumptions on the Objective for Optimal Distributed Control of Coupled ODE (Treated as PDE) and Parabolic PDE)*
We assume a linear-quadratic objective

$$\begin{aligned} \mathcal{J}(y, u) = {} & \frac{1}{2} \int_\Omega |R_{H,1;f} y_1(t_f, x) - y_{ref,1;f}(x)|^2 \, dx + \frac{1}{2} |R_{H,2;f} y_2(t_f) - y_{ref,2;f}|^2 \\ & + \frac{1}{2} \int_{\Omega_{t_f}} |R_{H,1} y_1(t, x) - y_{ref,1}(t, x)|^2 \, dx \, dt + \frac{1}{2} \int_I |R_{H,2} y_2(t) - y_{ref,2}(t)|^2 \, dt \\ & + \frac{\alpha_1}{2} \int_{\Omega_{t_f}} |u_1(t, x)|^2 \, dx \, dt + \frac{\alpha_2}{2} \int_I |u_2(t)|^2 \, dt. \end{aligned}$$

Let $\alpha_i \geq 0$, $i = 1, 2$, and consider $R_{H,1;f} \in L^\infty(\Omega; \mathbb{R}^{n_{y1} \times \tilde{n}_{y_i}})$, $R_{H,1} \in L^\infty(\Omega_{t_f}; \mathbb{R}^{n_{y1} \times \tilde{n}_{y_i}})$, $R_{H,2;f} \in \mathbb{R}^{n_{y2} \times \tilde{n}_{y2}}$, and $R_{H,2} \in L^\infty(I; \mathbb{R}^{n_{y2} \times \tilde{n}_{y2}})$. Furthermore, let $y_{ref,1;f} \in L^2(\Omega; \mathbb{R}^{\tilde{n}_{y1}})$, $y_{ref,1} \in L^2(\Omega_{t_f}; \mathbb{R}^{\tilde{n}_{y1}})$, and $y_{ref,2;f} \in \mathbb{R}^{\tilde{n}_{y2}}$, $y_{ref,2} \in L^2(I; \mathbb{R}^{\tilde{n}_{y2}})$.

We consider box constraints for the control as in Problem 3.7 with $-\infty < u_{i,min} < u_{i,max} < \infty$, $i = 1, 2$, the latter ensuring that U_{ad} is bounded and non-empty, if $U \neq \emptyset$.

Under the latter assumption the reduced objective is continuously F-differentiable w.r.t. the control.

We have according to Def. 2.12 as Lagrange function

$$\begin{aligned}L(y,u,\lambda) = \mathcal{J}(y,u) &+ \langle \lambda_1, E_1(y,u_1)\rangle_{W_1^*,W_1} + \langle \lambda_{1b}, y_1(0,\cdot) - y_{1;0}\rangle_{W_{1b}^*,W_{1b}} \\ &+ \langle \lambda_2, E_2(y,u_2)\rangle_{W_2^*,W_2} + \lambda_{2b}^\top (y_2(0) - y_{2;0}),\end{aligned}$$

where $W = W(I)^* \times [L^2(\Omega)]^{n_{y_1}} \times [H^1(I)^*]^{n_{y_2}} \times \mathbb{R}^{n_{y_2}}$. Correspondingly, e.g. $W_1 = W(I)^*$, $W_{1b} = [L^2(\Omega)]^{n_{y_1}}$, $W_2 = [H^1(I)^*]^{n_{y_2}}$, and $W_{2b} = \mathbb{R}^{n_{y_2}}$ are the factors of the product space W. It turns out, see also Remark 2.82, that the multipliers λ_{1b} and λ_{2b} are proportional to λ_1 and λ_2, resp., and thus may be ignored in the following. For the adjoint corresponding to the Neumann boundary condition we make an analogous observation. Hence we work with $W = W(I)^* \times [H^1(I)^*]^{n_{y_2}}$.

Theorem 3.12 *(First-Order Necessary Optimality Conditions for Optimal Distributed Control with Coupled ODE (Treated as PDE) and Parabolic PDE)*
Assume $[\hat{y},\hat{u}]$ is a minimizer of the optimal control problem, Problem 3.7, whereby we recall $Y = W(I) \times [H^1(I)]^{n_{y_2}}$ and $W = Y^*$. Under Assumption 3.9 (in particular $\Omega \subset \mathbb{R}^d$, $d = 1,2,3$) and under Assumption 3.11 with[5] $\tilde{p}_1 = \tilde{p}_2 = \hat{p}_1 = \hat{p}_2$, there exist adjoints $[\lambda_1, \lambda_2] \in W_1^* \times W_2^*$, s.t. the first-order optimality conditions, i.e. the state equations

$$\begin{aligned}\hat{y}_{1;t}' - k\Delta_x \hat{y}_1 + A_1(t,x)\hat{y}_1 - B_1(t,x)\hat{u}_1 - C_1(t,x,\hat{y}) &= 0_{W_1} && \text{in } \Omega_{t_f}, \\ \partial_\nu \hat{y}_1 &= 0_{[L^2(\Sigma_{t_f})]^{n_{y_1}}} && \text{on } \Sigma_{t_f}, \\ \hat{y}_1(0,\cdot) &= y_{1;0} && \text{a.e. on } \Omega, \\ \hat{y}_{2;t}' + A_2(t)\hat{y}_2 - B_2(t)\hat{u}_2 - C_2(t,\mathcal{E}\hat{y}_1,\hat{y}_2) &= 0_{W_2} && \text{in } I, \\ \hat{y}_2(0) &= y_{2;0},\end{aligned}$$

the adjoint equations,

$$\begin{aligned}-\lambda_{1;t}' - k\Delta_x\lambda_1 &+ (A_1(t,x)^* - C_{1;y_1}'(t,x,\hat{y})^*)\lambda_1 \\ &- \mathcal{E}_{y_1}'(\hat{y}_1)^* C_{2;\mathcal{E}y_1}'(t,\mathcal{E}\hat{y}_1,\hat{y}_2)^*\lambda_2 = -\phi_{1;y_1}'(t,x,\hat{y},\hat{u}_1) \\ &\qquad - \phi_{2;\mathcal{E}y_1}'(t,\mathcal{E}\hat{y}_1,\hat{y}_2,\hat{u}_2)\mathcal{E}_{y_1}'(\hat{y}_1) && \text{a.e. in } \Omega_{t_f}, \\ \partial_\nu \lambda_1 &= 0_{[L^2(\Sigma_{t_f})]^{n_{y_1}}} && \text{on } \Sigma_{t_f}, \\ \lambda_1(t_f,\cdot) &= -\Phi_{y_1(t_f)}'(\hat{y}(0),\hat{y}(t_f)) && \text{a.e. on } \Omega, \\ -\lambda_{2;t}' &+ (A_2(t)^* - C_{2;y_2}'(t,\mathcal{E}\hat{y}_1,\hat{y}_2)^*)\lambda_2 \\ &- \int_\Omega C_{1;y_2}'(t,x,\hat{y})^*\lambda_1\,dx = -\int_\Omega \phi_{1;y_2}'(t,x,\hat{y},\hat{u}_1)\,dx \\ &\qquad - \phi_{2;y_2}'(t,\mathcal{E}\hat{y}_1,\hat{y}_2,\hat{u}_2) && \text{a.e. in } I, \\ \lambda_2(t_f) &= -\Phi_{y_2(t_f)}'(\hat{y}(0),\hat{y}(t_f)),\end{aligned}$$

[5] This is just for ease of presentation of the idea of proof.

and the optimality

$$\begin{aligned}
\langle -B_1^*\lambda_1 + \phi'_{1;u_1}(\cdot, \hat{y}, \hat{u}_1), u_1 - \hat{u}_1\rangle_{U_1^*, U_1} &\geq 0 && \forall u_1 \in U_{ad,1},\\
\langle -B_2^*\lambda_2 + \phi'_{2;u_2}(\cdot, \mathcal{E}\hat{y}_1, \hat{y}_2, \hat{u}_2), u_2 - \hat{u}_2\rangle_{U_2^*, U_2} &\geq 0 && \forall u_2 \in U_{ad,2},
\end{aligned}$$

hold.

Let $\mathcal{E}$ be given by the averaging operator from Ex. 3.2 with a d-dimensional subset A of $\Omega \subset \mathbb{R}^d$. If we exploit the specific objective (Assumption 3.11) and let furthermore $\alpha_i > 0$, $i = 1, 2$, and U be a Hilbert, then the equations read for the adjoints

$$\begin{aligned}
-\lambda'_{1;t} - k\Delta_x\lambda_1 + (A_1(t,x) - C'_{1;y_1}(t,x,\hat{y}))^*\lambda_1&\\
-\frac{1}{|A|}C'_{2;\fint_A \hat{y}_1, \hat{y}_2}(t, \fint_A \hat{y}_1, \hat{y}_2)^*\lambda_2 &= -R^*_{H,1}(R_{H,1}\hat{y}_1 - y_{ref,1}(t,x)) && \text{a.e. in } \Omega_{t_f},\\
\partial_\nu\lambda_1 &= 0_{[L^2(\Sigma_{t_f})]^{n_{y_1}}} && \text{on } \Sigma_{t_f},\\
\lambda_1(t_f, x) &= -R^*_{H,1;f}(R_{H,1;f}\hat{y}_1(t_f, x)\\
&\qquad - y_{ref,1;f}(x)) && \text{on } \Omega,\\
-\lambda'_{2;t} + (A_2(t) - C'_{2;y_2}(t, \fint_A \hat{y}_1, \hat{y}_2))^*\lambda_2&\\
-\int_\Omega C'_{1;y_1}(t,x,\hat{y})^*\lambda_1\, dx &= -R^*_{H,2}(R_{H,2}\hat{y}_2 - y_{ref,2}(t)) && \text{a.e. in } I,\\
\lambda_2(t_f) &= -R^*_{H,2;f}(R_{H,2;f}\hat{y}_2(t_f) - y_{ref,2;f}),
\end{aligned}$$

and for the controls

$$\begin{aligned}
\hat{u}_1(t,x) &= P_{U_{ad,1}}\left(\frac{1}{\alpha_1}B_1^*(t,x)\lambda_1(t,x)\right) && \text{a.e. in } \Omega_{t_f},\\
\hat{u}_2(t) &= P_{U_{ad,2}}\left(\frac{1}{\alpha_2}B_2^*(t)\lambda_2(t)\right) && \text{a.e. in } I.
\end{aligned}$$

Moreover, then the adjoint problem shows that we have here the higher regularity $\lambda \in Y$.

For a proof, we follow closely [CRT18, Th. 2.2 & Corollary 2.1 & Th. 2.3] or the slightly different approach in [Ry16, Sect. 1.2]. An averaging-evaluation operator $\mathcal{E}$ fulfilling the assumptions does not complicate[6] the proof given there. For semilinear parabolic PDE we encounter similar issues as for semilinear elliptic PDE as illustrated in Example 2.74.

The idea is to consider ODE as PDE and for the moment the spaces $\hat{Y} = \{y \in [W(I)]^{n_y} \mid \partial_t y - k\Delta_x y \in [L^{\tilde{p}_1}(I; L^{\tilde{q}_1}(\Omega_{t_f})]^{n_y}\}$ (that is isomorph to Y where we differ between PDE and ODE) and $U = [L^2(\Omega_{t_f})]^{n_u}$. In fact, $\hat{Y}$ is a Banach space with a suitable norm. We exploit that $\hat{Y} \simeq W^{1,\tilde{p}_1}(I; W^{2,\tilde{q}_1}(\Omega)) \hookrightarrow L^\infty(\Omega_{t_f})$ under the premise $1/\tilde{p}_1 + d/(2\tilde{q}_1) < 1$, where $\tilde{p}_1$, $\tilde{q}_1 \in [2, \infty]$.

Now the linearized coupled ODE-PDE problem and the corresponding operator E can be extended onto $\hat{Y} \times U = [W(I)]^{n_y} \times [L^2(\Omega_{t_f})]^{n_u}$. The reason for this is that in the case of exponent 2 in time and space, the arising dual spaces are suitable.

[6] Note that, for instance, an evaluation operator appears in a slightly different, more direct proof of NOC for the truck-container problem [KG16].

This allows for showing the F-differentiability of $E : \hat{Y} \times U \to W$ and that $E'_y : \hat{Y} \to U$ is an isomorphism (consider the linearized coupled problem and exploit exponent 2), thus we may apply the implicit function theorem. Then we can derive the adjoint equations, particularly the variational inequality yielding the optimality condition. Finally, the higher regularity of the adjoint problem that is generically linear can be exploited.

An alternative proof strategy could rely on the result of Ito and Kunisch, Th. 2.73, where certain assumptions would have to be checked. Note that there it has to be distinguished between Y and $\tilde{Y}$ that implies to work with $W(I) \times [H^1(I)]^{n_{y_2}}$ instead of $\big(W(I) \cap [L^\infty(\Omega_{t_f})]^{n_{y_1}}\big) \times [H^1(I)]^{n_{y_2}}$, which has no useful dual space.

However, for unbounded $u_{min,i}$ or $u_{max,i}$, in particular the case without control constraints, no bounded control exists, unless $\alpha_1 > 0$ and more restrictive assumptions on U_{ad}, e.g. $U \in [L^\infty(I; L^2(\Omega))]^{n_u}$, are fulfilled [CRT18, Th. 2.4]. Then the control-to-state operator might be not continuous, making the existence of a control not very useful in practice.

For optimal control problems of this type with state constraints, we refer to [CRT18, Ch. 3].

3.3.3 Treat Parabolic PDE as DAE in Function Spaces

For the approach to consider a PDE as an ODE in a function space it is required that the PDE is an evolution equation. If the PDE is, e.g., elliptic, we can still consider the PDE as an algebraic equation in a function space. In our situation the relevant function spaces are Banach spaces.

Strongly continuous semigroups provide solutions of linear ODEs in Banach spaces with constant coefficients. ODEs in Banach spaces arise from partial differential equations as well as from delay differential equations.

We consider a Gelfand triple $V \overset{cd}{\hookrightarrow} H \overset{cd}{\hookrightarrow} V^*$ as defined in Def. A.17. In this study we restrict our focus on the formal computations for a parabolic PDE interpreted as an ODE in a Banach space.

Problem 3.13 *(Abstract Parabolic Evolution Equation)*
Let $a : V \times V \to \mathbb{R}, [w, v, t] \mapsto a(w, v, t)$ be a measurable mapping w.r.t. t for all w, $v \in V$ and a coercive, continuous bilinear form for almost all $t \in I := (0, t_f)$.

Find $y \in W(I) := W(I; H, V)$ s.t.

$$\langle y'_t(t), v\rangle_{V^*,V} + a(y(t), v; t) - \langle f(t), v\rangle_{V^*,V} = 0 \quad \forall v \in V \text{ for a.a. } t \in I := (0, t_f) \tag{3.16}$$

$$y(0) = y_0, \tag{3.17}$$

for given $f \in L^2(I; V^*)$, $y_0 \in H$.

In this problem boundary conditions are included in the function spaces H and V.

We consider the distributed control of the PDE, i.e. $f(t) = Bu(t)$ with $u \in U$, U a Hilbert space, and $B \in \mathcal{L}(U, L^2(I; V^*))$.

We revisit Problem 3.7 and reformulate it when we treat PDE as ODE in function spaces.

Problem 3.14 *(Optimal Distributed Control of Coupled ODE and Parabolic PDE (Treated as ODE in Function Spaces))*
We reconsider Problem 3.7. In addition we introduce a bilinear form $a(w, v; t) = k(\nabla w, \nabla v)_H$ for $k > 0$ as in Pb. 3.13. Contrary to Pb. 3.7 here we consider $A_1 \in \mathcal{L}(Y_1; L^2(I, V^*))$, $B_1 \in \mathcal{L}(U; L^2(I, V^*))$, $C_1 \in \mathcal{L}(Y; L^2(I, V^*))$ as time-dependent operators in Banach spaces. Accordingly, we drop in our notation the x-dependency of the operators A_1, B_1, and C_1. Furthermore, let $y_{1;0} \in H$, $u \in \tilde{U}_{ad,1} \times \tilde{U}_{ad,2} \subset U = U_1 \times U_2$ with $U_1 = V$, $U_2 = \mathbb{R}$,

$$\begin{aligned} \tilde{U}_{ad,1} &:= \{\tilde{u}_1 \in V \,|\, u_{min,1} \leq \tilde{u}_1 \leq u_{max,1}\}, \\ \tilde{U}_{ad,2} &:= [u_{2,min}, u_{2,max}]. \end{aligned}$$

We rewrite the differential equation constraints in Pb. 3.7 equivalently as

$$\begin{aligned} \langle y'_{1;t}, v_1\rangle_{V^*,V} + k(\nabla y_1, \nabla v_1)_H & \\ +\langle A_1(t)y_1 - B_1(t)u_1 - C_1(t, y), v\rangle_{V^*,V} &= 0 && \forall v_1 \in V \text{ a.e. in } I, \\ (y_1(0, \cdot), v_0)_H &= (y_{1;0}, v_0)_H && \forall v_0 \in H, \\ y'_{2;t} + A_2(t)y_2 - B_2(t)u_2 - C_2(t, \mathcal{E}y_1, y_2) &= 0 && \text{a.e. in } I, \\ y_2(0) &= y_{2;0}. \end{aligned}$$

Let the Robinson constraint qualification hold. Alternatively, we may use Th. 2.66 on the regularity of DAE problems that translates here to an ODE in function space and to a standard ODE. Thus we assume $\lambda_0 = 1$ in the remaining subsection and the Hamilton function as defined in Def. 2.64 for DAE optimal control reads in this context, where $W_1 = V$,

$$\begin{aligned} \mathcal{H}(t, y(t), u(t), \lambda(t)) = \Phi(y(0), y(t_f)) &+ \int_\Omega \phi_1(t, x, y(t), u_1(t))\, dx + \phi_2(t, \mathcal{E}y_1(t, \cdot), y_2(t), u_2(t)) \\ &+ \langle \lambda_1(t), -k\Delta_x y_1(t) + A_1(t)y_1(t) - B_1(t)u_1(t) - C_1(t, y(t))\rangle_{W_1^*,W_1} \\ &+ \lambda_2(t)^\top (A_2(t)y_2(t) - B_2(t)u_2(t) - C_2(t, \mathcal{E}y_1(t), y_2(t))) \quad \text{for a.a. } t \in I \end{aligned} \tag{3.18}$$

and we have the necessary first-order optimality conditions.

Theorem 3.15 *(First-Order Necessary Optimality Conditions for Optimal Distributed Control with Coupled ODE and Parabolic PDE (Treated as ODE in a Function Space))*
Assume $\hat{z} = [\hat{y}_1, \hat{y}_2, \hat{u}_1, \hat{u}_2]$ is a local minimizer of Problem 3.14. Let Assumptions 2.7 a) and

2.60 hold. In addition to the state equations holding at $[\hat{y}, \hat{u}]$, there follows the adjoint PDE

$$\begin{aligned}
\langle \lambda'_{1;t}(t), v_1 \rangle_{V^*,V} &= k(\nabla \lambda_1(t), \nabla v_1)_H \\
&\quad + \langle (A_1(t)^* - C'_{1;y_1}(t)^*)\lambda_1(t), v_1 \rangle_{V^*,V} \\
&\quad - \langle C'_{2;\mathcal{E}y_1}(t)^* \lambda_2(t), \mathcal{E}'_{y_1}(\hat{y}_1) v_1 \rangle_{V^*,V} \\
&\quad + \langle \phi'_{1;y_1}(t, \cdot, \hat{y}(t), \hat{u}_1(t)), v_1 \rangle_{V^*,V} \\
&\quad + \langle \phi'_{2;\mathcal{E}y_1}(t, \mathcal{E}\hat{y}_1(t,\cdot), \hat{y}_2(t), \hat{u}_2(t)), \mathcal{E}'_{y_1}(\hat{y}_1) v_1 \rangle_{V^*,V} \qquad && \forall v_1 \in V \ \textit{a.e. in } I, \\
(\lambda_1(t_f), v_f)_H &= -(\Phi'_{y_1(t_f)}(\hat{y}(0), \hat{y}(t_f)), v_f)_H && \forall v_f \in H,
\end{aligned}$$

the adjoint ODE

$$\begin{aligned}
\lambda'_{2;t}(t) &= (A_2(t)^* - C'_{2;y_2}(t)^*)\lambda_2 - \int_\Omega C'_{1;y_2}(t)^* \lambda_1 \, dx \\
&\quad + \int_\Omega \phi'_{1;y_2}(t, x, \hat{y}(t), \hat{u}_1(t)) \, dx \\
&\quad + \phi'_{2;y_2}(t, \mathcal{E}\hat{y}_1(t,\cdot), \hat{y}_2(t), \hat{u}_2(t)) \qquad && \textit{a.e. in } I, \\
\lambda_2(t_f) &= -\Phi'_{y_2(t_f)}(\hat{y}(0), \hat{y}(t_f)),
\end{aligned}$$

and the optimality

$$\begin{aligned}
0 &\leq \langle -B_1^* \lambda_1(t) + \phi'_{1;u_1}(t, \cdot, \hat{y}(t), \hat{u}_1(t)), u_1 - \hat{u}_1(t) \rangle_{V^*,V} \qquad && \forall u_1 \in \tilde{U}_{ad,1} \ \textit{a.e. in } I, \\
0 &\leq (-B_2^* \lambda_2(t) + \phi'_{2;u_2}(t, \mathcal{E}\hat{y}_1(t,\cdot), \hat{y}_2(t), \hat{u}_2(t)))^\top (u_2 - \hat{u}_2(t)) && \forall u_2 \in \tilde{U}_{ad,2} \ \textit{a.e. in } I.
\end{aligned}$$

Finally we conclude $\lambda_1 \in W(I; H; V)$ and $\lambda_2 \in W^{1,\infty}(I)$ exploiting the higher regularity for both adjoints (see also Lemma 2.63).

Proof. We apply Theorem 2.65 that translates analogously into the situation of the coupled ODE-PDE problem. □

Let $\tilde{u}_i \in \tilde{U}_{ad,i} \subset \tilde{U}_i$, $i = 1, 2$, denote the controls from the last approach. For comparison, we extend these controls into the time interval I by setting $U_i = \{v(t) \in \tilde{U} \,|\, t \in I\}$, $i = 1, 2$, $u_1(t) = \{\tilde{u}_1(x)(t) \,|\, t \in I\}$ and $u_2(t) = \{\tilde{u}_2(t)\}$. The necessary optimality equations derived in Th. 3.15 are pointwise in time.

Using (i) the Gelfand structure and (ii) that the bilinear form $a(\cdot, \cdot, t)$ is coercive and bounded, it follows that $t \to a(v_1, v_2, t) \in L^1(I)$ for any $v_1, v_2 \in L^2(I; V)$. Hence, in case $\tilde{p}_1 = \tilde{p}_2 = \tilde{q}_1 = 2$ and if the coupling operators C_1, C_2 are suitably bounded, it can be shown that the latter necessary optimality conditions are equivalent to the NOCs stated in Th. 3.12, cf. [HPUU09, 1.3.2.4].

3.3.4 Reverse Coupling Structure

Here we present in details the results from our paper [Ki18]. Concerning the notation we differ between the ODE and PDE states y_1 and y_2 and the components, like $y_{1,i}$, of these states.

Moreover, here we introduce the combined state-adjoint variables, by writing

$$z_{(\cdot)} = [y_{1,\cdot}, y_{2,\cdot}, \lambda_{1,\cdot}, \lambda_{2,\cdot}]^\top \in \mathbb{R}^{2n_y} \tag{3.19}$$

and

$$E_{(\cdot)}(z) := [E_{1,\cdot}(y,u), E_{2,\cdot}(y,u)] \in \mathbb{R}^{2n_y}.$$

Note the difference between z (the optimization variable) and $z_{(\cdot)}$.

Definition 3.16 *(Coupling Matrix and Coupling Arrow)*
We assume C_i and thus E_i, $i = 1, 2$, to be linear w.r.t. the states. The coupling matrix $K \in \mathbb{R}^{n_y \times n_y}$ of state equations is defined by

$$K_{kl} := E'_{(k);z_{(l)}}(1 - \delta_{kl}), \quad k, l = 1, \ldots, n_y.$$

Note that by construction we have $K_{ll} = 0$, $l = 1, \ldots, n_y$.

If an entry K_{kl}, $k \neq l$, is non-zero this means that state k depends on state l. In general (including states and adjoints) we denote this coupling shortly by a coupling arrow, i.e. by

$$z_{(k)} \rightsquigarrow z_{(l)}, \quad k, l = 1, \ldots, 2n_y.$$

For example, in our Problem 3.7 we have $n_y = n_{y_1} + n_{y_2}$ and in case of a linear C the coupling structure is encoded in

$$K := \begin{bmatrix} 0_{n_{y_1} \times n_{y_1}} & C'_{1;y_2} \\ C'_{2;\mathcal{E}y_1}\mathcal{E}'_{y_1} & 0_{n_{y_2} \times n_{y_2}} \end{bmatrix} \in \mathbb{R}^{n_y \times n_y}.$$

Theorem 3.17 *(Reverse Coupling Structure in the Adjoint CDE Problem)*
We consider Problem 3.7. For the geometry as in Assumption 3.9 and let Assumption 3.11 with $\alpha_i > 0$, $i = 1, 2$, hold. Furthermore, for simplicity we assume for the control space U that $\tilde{p}_1 = 2$, $\tilde{q}_1 = 2$, and $\tilde{p}_2 = 2$.

The coupling matrix of the adjoint system is $\check{K}^*$, i.e. the adjoint matrix of the coupling matrix K of the linearized state equation system, where a further operator for integrating over the spatial domain is applied to the adjoint ODE in addition. Thus $z_{(n_y+l)} \rightsquigarrow z_{(n_y+k)}$ in the adjoint system iff $z_{(k)} \rightsquigarrow z_{(l)}$ in the state system for $k \neq l$, $k, l = 1, \ldots, n_y$.

Proof. We introduce the diagonal projection operators

$$\Pi_i \lambda_i := -B_i P_{U_{ad},i}\left(\frac{1}{\alpha_i} B_i^* \lambda_j\right), \quad i = 1, 2,$$

where the projection is taken w.r.t. the L^2-norm. Due to the optimality conditions from Th. 3.12 we may eliminate the controls by Lemma 2.27, inserting $B_i \hat{u}_i = \Pi_i \lambda_i$, $i = 1, 2$, into the state equations. We set

$$\tilde{N}z := z_{(n_y+\cdot)} := diag(\Pi_1 \lambda_{1,\cdot}; \Pi_2 \lambda_{2,\cdot}) \in \mathbb{R}^{n_y \times n_y}.$$

Furthermore, we abbreviate the diagonal operator $D := E'_{(k);z_{(k)}}$, $k = 1, \ldots, n_y$, and introduce

$$C_0(t,x) := [C_1(t,x,y) - C'_{1;y}(t,x,y)y, C_2(t,\mathcal{E}y_1,y_2) - C'_{2;y}(t,\mathcal{E}y_1,y_2)y]^\top.$$

We introduce

$$\mathcal{A} := \begin{bmatrix} D & 0_{n_y\times n_y} \\ 0_{n_y\times n_y} & D^* \end{bmatrix} \in \mathbb{R}^{2n_y\times 2n_y},\ \mathcal{C} := \begin{bmatrix} K & \tilde{N} \\ 0_{n_y\times n_y} & \check{K}^* \end{bmatrix} \in \mathbb{R}^{2n_y\times 2n_y},\ f := \begin{bmatrix} C_0 \\ -\mathcal{J}'_y(\hat{y},\hat{u}) \end{bmatrix} \in \mathbb{R}^{2n_y}.$$

In detail, here we have

$$\check{K}^*\lambda := \begin{bmatrix} 0_{n_{y_1}\times n_{y_1}} & C'_{2;\mathcal{E}y_1}\mathcal{E}'_{y_1}\lambda_2 \\ \int_\Omega C'_{1;y_2}\lambda_1\,dx & 0_{n_{y_2}\times n_{y_2}} \end{bmatrix}.$$

By this means we rewrite the coupled state-adjoint system that follows from the necessary optimality system from Th. 3.12 as

$$\mathcal{A}\hat{z} + \mathcal{C}\hat{z} = f \quad \text{in } W \times W^*$$

where here $\hat{z}_{(\cdot)} = [\hat{y}_{1,\cdot}, \hat{y}_{2,\cdot}, \lambda_{1,\cdot}, \lambda_{2,\cdot}]^\top$ corresponding to (3.19) at the optimum. Note that the terms of $\mathcal{J}'_y$ that are linear in y could be incorporated in the lower left quarter block of the matrix $\mathcal{C}$ as well. Since $\mathcal{A}$ is a diagonal operator, the coupling matrix of the adjoint system is the lower right quarter block of the matrix $\mathcal{C}$. Note that the integrations over Ω appearing in $\check{K}^*$ do not change the coupling structure. □

In the example of the truck-container problem, see Sect. 4.2, where the viscosity solution of the Saint-Venant equations in 1D (i.e. two parabolic PDE of second-order) are fully coupled with Newton dynamics (i.e. two ODE of first-order) we observe this reverse coupling structure [KG16, Remark 4.2]. This reversal of the coupling structure should be preserved in the discretization and can be exploited for effective computations.

3.3.5 Numerical Optimal Control Methods for Coupled Ordinary and Partial Differential Equations

In numerical methods for optimal control problems with coupled ODE and PDE the availability of efficient algorithms is even more crucial than for optimal control with PDE alone. In this context we refer to our paper [GHK17] on a globalization strategy for semismooth Newton methods that is applied in the example of a truck transporting a fluid container in [KG16] yielding a certain speed-up compared to our first paper on this model [GK15], but for fixed terminal times. By structure-exploiting SQP methods, see [WGKG18], a further speed-up is achieved.

3.4 A Simple Example for a Fully Coupled ODE-PDE Problem

To keep things simple we consider a problem that is 1D in space ($d = 1$), e.g., $\Omega = (0, L)$ for some given length $L > 0$. The time interval is again $I = (0, t_f)$ with some fixed terminal time t_f. Again we write $\Omega_{t_f} = I \times \Omega$ for the parabolic cylinder. We consider a single parabolic PDE with state $y_1 : \Omega_{t_f} \to \mathbb{R}$ and the ODE system is 2D written as a first-order system. The ODE state is $y_2 : I \to \mathbb{R}$. Here we write $y = [y_1, y_2]$ for the states. If $v = \dot{y}_2$ is required explicitly in this context, then we write y_3 for v.

For the PDE we focus here on the heat equation (with Neumann b.c. in space)

$$\begin{aligned} y'_{1;t}(t,x) - k y''_{1;xx}(t,x) &= -A_{12}(t,x) y_2(t) + B_1 u_1(t,x) + C_{0,1}(t,x), && \forall (t,x) \in \Omega_{t_f}, && (3.20)\\ y'_{1;x}(t,0) = y'_{1;x}(t,L) &= 0, && \forall t \in I, && (3.21)\\ y_1(0,x) &= y_{1;0}(x), && \forall x \in \Omega, && (3.22) \end{aligned}$$

with the diffusion coefficient $k > 0$, suitable operators $A_{12} : \Omega_{t_f} \to \mathbb{R}$, $B_1 : \mathbb{R} \to \mathbb{R}$ to be defined, and a distributed control $u_1 : \Omega_{t_f} \to \mathbb{R}$. Furthermore, we assume $C_{0,1} : \Omega_{t_f} \to \mathbb{R}$ and an initial value $y_{1;0} : \Omega \to \mathbb{R}$ to be given.

Let the ODE be given by the Newton dynamics (without damping and restoring forces)

$$\begin{aligned} & y''_2(t) = -A_{21}(t) \fint_A y_1(t,x)\, dA(x) + B_2 u_2(t) + C_{0,2}(t) \quad \forall t \in I, && (3.23)\\ & y_2(0) = y_{2;0} := q_0, && (3.24)\\ & y'_2(0) = y_{2;1} := v_0, && (3.25) \end{aligned}$$

where $A_{21} : I \to \mathbb{R}$, $B_2 : \mathbb{R} \to \mathbb{R}$ are some suitable operators to be precised yet, $u_2 : I \to \mathbb{R}$ is the control force for the ODE, $C_{0,2} : I \to \mathbb{R}$ is a given force term, and q_0, $v_0 \in \mathbb{R}$ are given initial conditions for position and velocity, respectively.

Note that the indices of the operator A_{12} signify that it couples y_2 to y_1, whereas A_{21} maps y_1 into the y_2 equation.

Since we do not wish to consider a family of ODEs that have the spatial point x as parameter, we consider here again the case that some spatial average over the PDE solution y_1 enters the ODE for y_2 (cf. Def. 3.1 & Ex. 3.2), where $A \subset \Omega$ is an arbitrary interval here.

In the following we consider this specific coupled ODE-PDE model problem for optimal control. We follow the approach to treat ODE as PDE and compare with the other approach, to treat PDE as ODE in function spaces, at the end of this subsection.

Problem 3.18 *(Model Problem for Optimal Control of Coupled ODE and PDE)*
Let the prerequisites in Problem 3.7 hold, where $\Omega = (0, L)$ for $L > 0$, $n_{y_1} = n_{y_2} = 1$ and $n_{u_1} = n_{u_2} = 1$.

Thus $Y_1 = W(I)$, $Y_2 = H^2(I)$, $U_1 = L^{\tilde{p}_1}(I, L^{\tilde{q}_1}(\Omega))$, $U_2 = L^{\tilde{p}_2}(I)$ with $\tilde{p}_1, \tilde{q}_1, \tilde{p}_2 \in [2, \infty]$ s.t. $1/\tilde{p}_1 + 1/(2\tilde{q}_1) < 1$, $W_1 = Y_1^*$, and $W_2 = Y_2^*$ if $\tilde{p}_2 < \infty$ or $W_2 = L^\infty(I)$ if $\tilde{p}_2 = \infty$. We set $Y = Y_1 \times Y_2$, $U = U_1 \times U_2$, and $W = W_1 \times W_2$.

We assume $A_1 \equiv 0_{W_1}$, $A_2 \equiv 0_{W_2}$, B_1 an embedding from U_1 to $W_1 = W(I)^*$, B_2 is an embedding from U_2 to $H^2(I)^*$ ($\tilde{p}_2 < \infty$) or $L^\infty(I)$ ($\tilde{p}_2 = \infty$), $C_1(t,x,y) \equiv -A_{12}(t,x) y_2 + C_{0,1}(t,x)$,

$$(\mathcal{E} y_1)(t) = \fint_A y_1(t,x)\, dA(x), \qquad (3.26)$$

and $C_2(t, \mathcal{E} y_1, y_2) \equiv -A_{21}(t) \fint_A y_1 \, dA(x) + C_{0,2}(t)$.

The given tracking states are $y_{ref,1} \in L^2(\Omega)$ and $y_{ref,2} \in \mathbb{R}$.

We wish to find $[y, u] \in Y \times U$ such that
the objective, consisting of tracking type terms and control costs only,

$$\begin{aligned}\mathcal{J}(y,u) = \frac{1}{2}\int_0^{t_f}\int_\Omega |y_1(t,x) - y_{ref,1}(x)|^2\,dx\,dt + \frac{1}{2}\int_0^{t_f} |y_2(t) - y_{ref,2}|^2\,dt \\ + \frac{\alpha_1}{2}\int_0^{t_f}\int_\Omega |u_1(t,x)|^2\,dx\,dt + \frac{\alpha_2}{2}\int_0^{t_f} |u_2(t)|^2\,dt \qquad (3.27)\end{aligned}$$

with suitable weights $\alpha_i \geq 0$, not all being zero, is minimized,
where for the controls we consider box constraints, i.e. $u_{min,i} \leq u_i \leq u_{max,i}$ (with $-\infty < u_{min,i} < u_{max,i} < \infty$), $i = 1, 2$, for almost all $[t, x]$ or t, respectively,
subject to the constraints (3.20) – (3.25).

We annotate that in principle we could also consider tracking term functions $y_{ref,1}(t, x)$ and $y_{ref,2}(t)$ depending on time as well, cf. Assumpt. 3.11.

Note that we rewrite the second-order ODE as two first-order ODEs in the following. Then we have $n_{y_1} = 1$, $n_{y_2} = 2$, and $n_{u_1} = n_{u_2} = 1$ and the spaces are adjusted suitably.

This could be interpreted as a version of the rocket car problem [WRP10, PRWW10, PRWW14], where the heating up of the car is modelled in some sense. The Neumann b.c. (3.21) represent a boundary, where heat may be released freely. Then u_1 is the active cooling of the heat shield of the car and u_2 is the acceleration of the car. However, state constraints, e.g. an upper bound for the temperature y_1 is reasonable, are not yet considered in our problem. Another interpretation is in the light of the truck-container problem (see Sect. 4.2), where we have two parabolic equations after adding an artificial viscosity, one subject to Dirichlet and one subject to Neumann boundary conditions. For numerical examples we refer to our included paper [KG16] on the truck-container problem.

Our model problem can be considered as a special case of the more general optimal control problem for reaction diffusion systems, treated in [Ry16, CRT18], see Subsection 3.3.2. We recall the assumptions and the notation in Problem 3.7.

Note that $U_{ad,i}$, $i = 1, 2$, and U_{ad} are convex, closed, bounded, and non-empty. Furthermore,

$$E'_y(y,u) = \begin{bmatrix} \partial_t - k\partial_{xx} & A_{12} \\ A_{21} & \partial_{tt} \end{bmatrix},$$

and $E'_{1;y_1} = [E'_y]_{1,1}$ and $E'_{2;y_2} = [E'_y]_{2,2}$ have a bounded inverse.

Neglecting the (full) coupling for the moment (i.e. $A_{12} = 0_{W_1}$, $A_{21} = 0_{W_2}$), we have linear-quadratic optimal control problems (see Examples 2.49 and 2.53) for the ODE and the PDE, respectively.

Example 3.19 *(Model Problem - Uncoupled Case)*
We consider the uncoupled version of Problem 3.18 for $\tilde{p}_1 = \tilde{q}_1 = \tilde{p}_2 = 2$. Let $\alpha_1 > 0$ and

$\alpha_2 > 0$. Thus here we consider $U_1 = L^2(I; L^2(\Omega)) = U_1^*$, $Y_1 = W(I)$, $W_1 = Y_1^*$, and

$$U_{ad,1} = \{u_1 \in U_1 \,|\, u_{min,1} \le u_1(t,x) \le u_{max,1} \quad \text{for a.a. } [t,x] \in \Omega_{t_f}\}$$

that is convex, closed, and bounded. The operators are $A_{11} \in \mathcal{L}(Y_1, W_1)$,

$$A_{11} y_1 := y'_{1,t}(t) - k y''_{1,xx},$$

and $B_1 \in \mathcal{L}(U_1, W_1)$,

$$B_1 u_1 = u_1|_{W_1}$$

in the PDE. We remark that B_1 is a compact embedding operator. The right-hand side is $C_{0,1} \in W_1$.

This PDE has a unique solution for all times, the general solution of the PDE subproblem reads

$$y_1(t,x) = \int_0^L G_{X22}(t,x,t;0,\tilde{x}) y_{1;0}(\tilde{x})\, d\tilde{x} - \int_0^t \int_0^L G_{X22}(t,x;\tilde{t},\tilde{x}) \left(B_1 u_1(\tilde{t},\tilde{x}) + C_{0,1}(\tilde{t},\tilde{x})\right)\, d\tilde{x}\, d\tilde{t},$$

where the corresponding fundamental solution for the heat equation, also called heat kernel, on $\Omega = (0,L)$ with homogeneous Neumann boundary conditions on $\Sigma_{t_f} = I \times \{0; L\}$ is the Green's function number X22 [CBH+11]

$$\begin{aligned} G_{X22}(t,x;\tilde{t},\tilde{x}) &= \frac{1}{2\sqrt{\pi k (t-\tilde{t})}} \sum_{n=-\infty}^{\infty} \left(\exp\left(-\frac{(2nL + x - \tilde{x})^2}{4k(t-\tilde{t})}\right) + \exp\left(-\frac{(2nL + x + \tilde{x})^2}{4k(t-\tilde{t})}\right)\right) \\ &= \frac{1}{L}\left(1 + 2\sum_{m=1}^{\infty} \exp\left(-\frac{m^2\pi^2 k(t-\tilde{t})}{L^2}\right) \cos\left(m\pi\frac{x}{L}\right)\cos\left(m\pi\frac{\tilde{x}}{L}\right)\right). \end{aligned}$$

Note that the latter two (Fourier) series have different convergence properties, for small $t - \tilde{t}$ the first representation is preferable, for large the second one. By using Green's functions it may be avoided to use the concept of weak solutions. This Green's function is non-negative and symmetric w.r.t. x and $\tilde{x}$. It has a weak singularity at $[x = \tilde{x}, t = 0]$ [Tr10, Subsect. 3.2.2].

We could also consider homogeneous Dirichlet b.c. at $x = 0$ and $x = L$ for our parabolic PDE, requiring the Greens function X11 [CBH+11] (with two representations with suitable convergence for small or large time differences $t - \tilde{t}$, resp.) instead of X22:

$$\begin{aligned} G_{X11}(t,x;\tilde{t},\tilde{x}) &= \frac{1}{2\sqrt{\pi k (t-\tilde{t})}} \sum_{n=-\infty}^{\infty} \left(\exp\left(-\frac{(2nL + x - \tilde{x})^2}{4k(t-\tilde{t})}\right) - \exp\left(-\frac{(2nL + x + \tilde{x})^2}{4k(t-\tilde{t})}\right)\right) \\ &= \frac{2}{L}\sum_{m=1}^{\infty} \exp\left(-\frac{m^2\pi^2 k(t-\tilde{t})}{L^2}\right) \sin\left(m\pi\frac{x}{L}\right)\sin\left(m\pi\frac{\tilde{x}}{L}\right). \end{aligned}$$

A_{11} has a bounded inverse. The uncoupled control-to-state operator reads

$$\begin{aligned} &\mathcal{S}_1 : U_1 \to Y_1, u_1 \mapsto A_{11}^{-1}(B_1 u_1 + C_{0,1}) \\ &\quad := \int_0^L G_{X22}(t,x;0,\tilde{x}) y_{1,0}(\tilde{x})\, d\tilde{x} - \int_0^t \int_0^L G_{X22}(t,x;\tilde{t},\tilde{x}) \left(B_1 u_1(\tilde{t},\tilde{x}) + C_{0,1}(\tilde{x},\tilde{t})\right)\, d\tilde{x}\, d\tilde{t} \end{aligned} \tag{3.28}$$

and it may demonstrated that, e.g., $S : L^2(Q) \to C^0(I; L^2(\Omega))$ for $y_0 \in L^2(\Omega)$, though G_{X22} exhibits a weak singularity [Tr84b]. The corresponding part of the Lagrange function $L(y, u, \lambda) = L_1(y_1, u_1, \lambda_1) + L_2(y_2, u_2, \lambda_2)$ (that is additive for uncoupled problems) is

$$L_1(y_1, u_1, \lambda_1) := \frac{1}{2}\|y_1 - y_{ref,1}\|^2_{L^2(I;L^2(A))} + \frac{\alpha_1}{2}\|u_1\|^2_{L^2(I;L^2(A))} + \langle \lambda_1, A_{11}y_1 - B_1u_1 - C_{0,1}\rangle_{W_1^*,W_1}.$$

Note that the homogeneous Neumann b.c. are included in the weak formulation of the PDE.

There exists a Lagrange multiplier $\lambda_1 \in W_1^*$ such that at the optimal solution $[\hat{y}_1(\hat{u}_1), \hat{u}_1] \in Y_1 \times U_1$ the adjoint λ_1 fulfils the KKT-system

$$\begin{aligned} \langle w_1, A_{11}\hat{y}_1 - B_1\hat{u}_1 - C_{0,1}\rangle_{W_1^*,W_1} &= 0 && \forall w_1 \in W_1^*, \\ \langle A_{11}\lambda_1 + \hat{y}_1 - y_{ref,1}, v_1\rangle_{Y_1^*,Y_1} &= 0 && \forall v_1 \in Y_1, \end{aligned}$$
$$\hat{u}_1 = P_{U_{ad,1}}\left(\frac{1}{\alpha_1}B_1^*\lambda_1\right).$$

Note that the Laplace operator is self-adjoint. We write the adjoint explicitly,

$$\begin{aligned} -\lambda'_{1,t}(t,x) - k\lambda''_{1,xx}(t,x) &= -\hat{y}_1(t,x) + y_{ref,1}(x) && \forall [t,x] \in \Omega_{t_f}, \\ \lambda'_{1;x}(t,x) &= 0 && \forall x \in \Sigma_{t_f}, \\ \lambda_1(t_f, x) &= 0 && \forall x \in \Omega, \end{aligned}$$

yielding the solution formula

$$\lambda_1(t,x) = -\int_t^{t_f}\int_0^L G_{X22}(t - t_f, x; \tilde{t} - t_f, \tilde{x})(\hat{y}_1(\tilde{t}, \tilde{x}) - y_{ref,1}(\tilde{x}))\, d\tilde{x}\, d\tilde{t},$$

where the arguments of the heat kernel are adjusted for solving backward in time or for a diffusion coefficient $-k < 0$, resp. Thus we have the state

$$\begin{aligned} \hat{y}_1(t,x) = \int_0^L G_{X22}(t,x;0,\tilde{x})y_{1;0}(\tilde{x})\, d\tilde{x} - \int_0^t\int_0^L G_{X22}(t,x;\tilde{t},\tilde{x}) \times \Big(C_{0,1}(\tilde{t},\tilde{x}) + B_1 \times \\ \times P_{U_{ad,1}}\left(-\frac{1}{\alpha_1}B_1^*\int_{\tilde{t}}^{t_f}\int_0^L G_{X22}(\tilde{t} - t_f, \tilde{x}; \check{t} - t_f, \check{x})(\hat{y}_1(\check{t},\check{x}) - y_{ref,1}(\check{x}))\, d\check{x}\, d\check{t}\right)\Big)\, d\tilde{x}\, d\tilde{t}. \end{aligned}$$

Given a suitable control, the well-posedness for a single parabolic PDE and its optimal control is a standard result (see Subsect. 2.7.2). Since a convolution with a kernel like G_{X22} has a smoothing property, the last equation may serve to construct a fixed point iteration for solving the optimal control problem.

For the ODE, the operators are $A_{22} \in \mathcal{L}(Y_2, W_2)$, $A_{22}y_2 := y_2''(t)$, and $B_2 \in \mathcal{L}(U_2, W_2)$, $B_2u_2 = u_2|_{W_2}$. We remark that B_2 is a compact embedding operator. The right-hand side is $C_{0,2} \in W_2$.

This ODE has a unique solution for sufficiently small times, obtained by integrating twice as

$$y_2(t) = -\int_0^t\int_0^s B_2u_2(\tau) + C_{0,2}(\tau)\, d\tau\, ds - tv_0 - q_0. \tag{3.29}$$

Thus A_{22} has a bounded inverse. The uncoupled control-to-state operator reads

$$\mathcal{S}_2 : U_2 \to Y_2, \quad u_2 \mapsto A_{22}^{-1}(B_2u_2 + C_{0,2}) := -\int_0^t \int_0^s B_2u_2(\tau) + C_{0,2}(\tau)\, d\tau\, ds - tv_0 - q_0$$

and the part of the Lagrange function

$$L_2(y_2, u_2, \lambda_2) = \frac{1}{2}\|y_2 - y_{ref,2}\|^2_{H^1(I)} + \frac{\alpha_2}{2}\|u_2\|^2_{L^2(I)} + \langle \lambda_2, A_{22}y_2 - B_2u_2 - C_{0,2}\rangle_{W_2^*,W_2}.$$

There exists a Lagrange multiplier $\lambda_2 \in W_2^*$ such that at the optimal solution $[\hat{y}_2(\hat{u}_2), \hat{u}_2] \in Y_2 \times U_2$ the λ_2 fulfils the KKT-system

$$\begin{aligned}
\langle w_2, A_{22}\hat{y}_2 - B_2\hat{u}_2 - C_{0,2}\rangle_{W_2^*,W_2} &= 0 & \forall w_2 \in W_2^*,\\
\langle A_{22}^*\lambda_2 + \hat{y}_2 - y_{ref,2}, v_2\rangle_{Y_2^*,Y_2} &= 0 & \forall v_2 \in Y_2,\\
\hat{u}_2 &= P_{U_{ad,2}}\left(\frac{1}{\alpha_2}B_2^*\lambda_2\right).
\end{aligned}$$

We write the adjoint explicitly, by considering $v_2 \in \{\tilde{v} \in Y_2 \,|\, \tilde{v}(0) = \tilde{v}'(0) = 0\}$,

$$\begin{aligned}
\lambda_2''(t) &= -\hat{y}_2(t) + y_{ref,2}(t) & \forall t \in I,\\
\lambda_2(t_f) &= 0,\\
\lambda_2'(t_f) &= 0,
\end{aligned}$$

yielding the solution formula by integrating twice

$$\lambda_2(t) = -\int_0^t \int_0^s \hat{y}_2 - y_{ref,2}\, d\tau\, ds.$$

We note the higher regularity of the adjoint, i.e. $\lambda_2 \in Y_2$, too. Thus we have for the state the equation

$$\hat{y}_2(t) = -\int_0^t \int_0^s B_2P_{U_{ad,2}}\left(-\frac{1}{\alpha_2}B_2^*\int_0^{\tilde{t}}\int_0^{\tilde{s}} \hat{y}_2 - y_{ref,2}\, d\tilde{\tau}\, d\tilde{s}\right) + C_{0,2}(\tau)\, d\tau\, ds - tv_0 - q_0$$

to be solved.

For a parabolic PDE there hold the Theorems 2.80, and 2.81, discussed above. Note that we work here with a spatial domain that is one-dimensional, thus we have the (compact) embedding $H^1(\Omega) \hookrightarrow L^\infty(\Omega)$. Thus the approach to consider $\tilde{Y} = Y \cap L^\infty(\Omega_{t_f})$ as in [IK08] coincides with the standard approach in Sect. 2.7.

Treat ODE as PDE Yielding an OCP for a Coupled PDE System

Here we consider the ODE of the coupled model problem as an elliptic PDE of second-order. We consider the state spaces as before. For the boundedness of the coupling operators A_{12} and A_{21} we make the following assumption, in line with Assumption 3.9.

Assumption 3.20 *(Assumptions for Existence and Uniqueness for Optimal Control of Model Problem)*
For the coupling operators there holds $A_{12} \in L^{\check{p}_1}(I; L^{\check{q}_1}(\Omega))$ with $\check{p}_1, \check{q}_1 \in [2, \infty]$, $1/\check{p}_1 + 1/(2\check{q}_1) < 1$ and $A_{21} \in L^{\check{p}_2}(I)$ with $\check{p}_2 \in [2, \infty]$. Both latter operators are linear and we require that they are of class C^1 w.r.t. y_j and that

$$\begin{aligned} A_{12}(t,x) &\geq C_{N_1} && \text{for a.a. } [t,x] \in \Omega_{t_f}, \\ A_{21}(t) &\geq C_{N_2} && \text{for a.a. } t \in I. \end{aligned}$$

Note that the averaging-evaluation operator as defined in (3.26) fulfils Assumption 3.9, if $|A| > 0$.

We pursue a fixed point strategy that is motivated by Sect. 3.2.

Example 3.21 *(Model Problem - Coupled Case)*
Let Assumption 3.20 hold. We continue with Example 3.19. W.l.o.g. let $v_0 = 0$, $C_{0,1} \equiv 0_{W_1}$ and $C_{0,2} \equiv 0_{W_2}$. Plugging in the ODE solution y_2, of the type (3.29) but with our coupling term, into the PDE following (3.28) yields

$$\begin{aligned} y_1(t,x) = &\int_0^L G_{X22}(t,x;0,\tilde{x}) y_{1;0}(\tilde{x})\, d\tilde{x} - \int_0^t \int_0^L G_{X22}(t,x;\tilde{t},\tilde{x}) \times \\ &\times \left(A_{12}(\tilde{t},\tilde{x}) \int_0^{\tilde{t}} \int_0^s A_{21}(\tau) \fint_A y_1(\tau, X)\, dA(X) - B_2 u_2(\tau)\, d\tau\, ds + q_0 + B_1 u_1(\tilde{t},\tilde{x}) \right) d\tilde{x}\, d\tilde{t}. \end{aligned}$$

We encounter the structure

$$y_1(t,x) = \textit{terms with init. values} + \int_0^t \int_0^L G_{X22}(t,x;\tilde{t},\tilde{x}) \left((\tilde{\mathcal{A}}_1 y_1)(\tilde{t},\tilde{x}) + (\tilde{\mathcal{B}}_1 u)(\tilde{t},\tilde{x}) \right) d\tilde{x}\, d\tilde{t} \tag{3.30}$$

with

$$\begin{aligned} (\tilde{\mathcal{A}}_1 y_1)(\tilde{t},\tilde{x}) &= -A_{12}(\tilde{t},\tilde{x}) \int_0^{\tilde{t}} \int_0^s A_{21}(\tau) \fint_A y_1(\tau, X)\, dA(X)\, d\tau\, ds, \\ (\tilde{\mathcal{B}}_1 u)(\tilde{t},\tilde{x}) &= -A_{12}(\tilde{t},\tilde{x}) \int_0^{\tilde{t}} \int_0^s B_2 u_2(\tau)\, d\tau\, ds - B_1 u_1(\tilde{t},\tilde{x}). \end{aligned}$$

For sufficiently small times t, we may prove that (3.30) yields a strict contraction in the Banach space Y_1 provided the given effective control under application as encoded in the operator $\tilde{\mathcal{B}}_1$ is bounded.

On the other hand, plugging in the PDE solution into the ODE yields

$$\begin{aligned} y_2''(t) - (\tilde{\mathcal{A}}_2 y_2)(t) &= (\tilde{\mathcal{B}}_2 u)(t) + \textit{term with initial value } y_{1;0} && \forall t \in I, \\ y_2(0) &= y_{2;0} = q_0, \\ y_2'(0) &= 0, \end{aligned}$$

or

$$y_2(t) = \int_0^t \int_0^s (\tilde{\mathcal{A}}_2 y_2)(\tau) + (\tilde{\mathcal{B}}_2 u)(\tau)\, d\tau\, ds + \textit{term with initial values} \quad \forall t \in I,$$

with

$$\begin{aligned}
(\tilde{\mathcal{A}}_2 y_2)(t) &:= -A_{21}(t) \fint_A \int_0^t \int_0^L G_{X22}(t,x;\tilde{t},\tilde{x}) A_{12}(\tilde{t},\tilde{x})\, d\tilde{x}\, y_2(\tilde{t})\, d\tilde{t}\, dx,\\
(\tilde{\mathcal{B}}_2 u)(t) &= -A_{21}(t) \fint_A \int_0^t \int_0^L G_{X22}(t,x;\tilde{t},\tilde{x}) B_1 u_1(\tilde{t},\tilde{x})\, d\tilde{x}\, d\tilde{t}\, dx - B_2 u_2(t).
\end{aligned}$$

Assuming that we may interchange the order of integration, we may rearrange

$$\begin{aligned}
(\tilde{\mathcal{A}}_2 y_2)(t) &= -A_{21}(t) \int_0^t \int_0^L \tilde{G}_{X22}(t;\tilde{t},\tilde{x})\, A_{12}(\tilde{t},\tilde{x})\, d\tilde{x}\, y_2(\tilde{t})\, d\tilde{t},\\
(\tilde{\mathcal{B}}_2 u)(t) &= -A_{21}(t) \int_0^t \int_0^L \tilde{G}_{X22}(t;\tilde{t},\tilde{x}) B_1 u_1(\tilde{t},\tilde{x})\, d\tilde{x}\, d\tilde{t} - B_2 u_2(t),
\end{aligned}$$

where we set

$$\tilde{G}_{X22}(t;\tilde{t},\tilde{x}) := \fint_A G_{X22}(t,x;\tilde{t},\tilde{x})\, dx.$$

Note that by a formal integration in the case $A = \Omega$ this simplifies to $\tilde{G}_{X22}(t;\tilde{t},\tilde{x}) = 1$. Again, a solution can be obtained by a fixed point approach.

We abbreviate

$$\breve{A}_{21}(t) y_1 := A_{21}(t) \mathcal{E} y_1 = A_{21}(t) \fint_A y_1(t,x)\, dA(x).$$

We rewrite the differential equations in our model problem, Problem 3.18, as a system:

$$\begin{aligned}
&\partial_t \begin{bmatrix} y_1(t,x) \\ q(t) \\ v(t) \end{bmatrix} - k\partial_{xx} \begin{bmatrix} y_1(t,x) \\ q(t) \\ v(t) \end{bmatrix} + \begin{bmatrix} 0 & A_{12}(t,x) & 0 \\ 0 & 0 & -1 \\ \breve{A}_{21}(t) & 0 & 0 \end{bmatrix} \begin{bmatrix} y_1(t,x) \\ q(t) \\ v(t) \end{bmatrix}\\
&\quad = \begin{bmatrix} B_1 u_1(t,x) \\ 0 \\ B_2 u_2(t) \end{bmatrix} + \begin{bmatrix} C_{0,1}(t,x) \\ 0 \\ C_{0,2}(t) \end{bmatrix} \qquad \forall [t,x] \in \Omega_{t_f},\\
&\begin{bmatrix} y(0,x) \\ q(0) \\ v(0) \end{bmatrix} = \begin{bmatrix} y_0(x) \\ q_0 \\ v_0 \end{bmatrix} \qquad \forall x \in \Omega, \text{ and } \begin{bmatrix} \partial_x y(t,0) \\ \partial_x y(t,L) \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix} \qquad \forall t \in I.
\end{aligned}$$

We recall that we use the notation $y = [y_1, y_2, y_3]^\top := [y_1, y_2 = q, v]^\top$, $y_0 = [y_{1;0}, y_{2;0} = q_0, v_0]^\top$, and $C_0 = [C_{0,1}(t,x), 0, C_{0,2}(t)]^\top$, $y_{ref} = [y_{ref,1}, y_{ref,2} = q_{H,ref}, 0]^\top$ (note that the latter value is required later). Moreover, we consider the linear operator A with a block diagonal structure, namely

$$Ay(t,x) := \begin{bmatrix} \partial_t - k\partial_{xx} & 0 & 0 \\ 0 & \partial_t - k\partial_{xx} & -1 \\ 0 & 0 & \partial_t - k\partial_{xx} \end{bmatrix} y(t,x) = \begin{bmatrix} \partial_t - k\partial_{xx} & 0 & 0 \\ 0 & \partial_t & -1 \\ 0 & 0 & \partial_t \end{bmatrix} \begin{bmatrix} y_1(t,x) \\ q(t) \\ v(t) \end{bmatrix},$$

as coupling matrix we find

$$K(t,x) := \begin{bmatrix} 0 & A_{12}(t,x) & 0 \\ 0 & 0 & 0 \\ \check{A}_{21}(t) & 0 & 0 \end{bmatrix},$$

and

$$Bu(t,x) = [B_1 u_1(t,x), 0, B_2 u_2(t)]^\top.$$

Then we write the system in the form

$$\begin{aligned} Ay(t,x) + Ky(t,x) &= Bu(t,x) + C_0(t,x) && \text{for a.a. } [t,x] \in \Omega_{t_f}, \\ \partial_x y(t,0) = \partial_x y(t,L) &= 0 && \text{for a.a. } t \in I, \\ y(0,x) &= y_0(x) && \text{for a.a. } x \in \Omega. \end{aligned}$$

We have the following result following from Subsection 3.3.2 that has been proven for the more general situation of optimal control of reaction-diffusion systems (see Th. 3.10) . For our coupled model problem we have $d = 1$, $n_{y_1} = 2$, $n_{y_2} = 1$, thus $n_y = 3$, and $n_u = 2$ in Assumption 3.9. For the operators there holds $A \in [L^\infty(\Omega_{t_f})]^{3\times 3}$, $B \in [L^\infty(\Omega_{t_f})]^{3\times 2}$, $K \in [L^\infty(\Omega_{t_f})]^{3\times 3}$, and $C_0 \in [L^{\check{p}}(I; L^{\check{q}}(\Omega))]^3$ with $\check{p}, \check{q} \in [2, \infty]$, $1/\check{p} + 1/(2\check{q}) < 1$.

Theorem 3.22 *(Existence and Uniqueness for Model Problem)*
Under Assumption 3.20 for general $\tilde{p} := \tilde{p}_1 = \tilde{p}_2$, $\tilde{q} := \tilde{q}_1$ and if $y_0 \in L^\infty(\Omega) \times \mathbb{R}^{2n_{y_2}}$, Problem 3.18 has a unique solution $y \in \tilde{Y} := [W(I) \cap L^\infty(\Omega_{t_f})] \times [L^{\tilde{p}}(I)]^2$ for every $u \in U := L^{\tilde{p}}(I; L^{\tilde{q}}(\Omega)) \times [L^{\tilde{p}}(I)]^2$, where $\tilde{p}$, $\tilde{q} \in [2, \infty]$ with $1/\tilde{p} + 1/(2\tilde{q}) < 1$.

There holds the estimate

$$\|y\|_{\tilde{Y}} \leq c \left(\|y_{1;0}\|_{L^\infty(\Omega)} + \|y_{2;0}\| + \|u\|_U + \|C_{0,1}\|_{L^{\check{p}}(I;L^{\check{q}}(\Omega))} + \|C_{0,2}\|_{[L^{\check{p}}(I)]^2} \right)$$

with an constant c being independent from u.

If, in addition $y_{1;0} \in C^0(\Omega)$, then we have $y \in C^0(\Omega_{t_f}) \times [C^0(I)]^2$.

Again, there exists an optimal control $u \in U$ and we may define a control-to-state map $S : U \to Y, u \to y$, that turns out to be suitably differentiable under certain conditions.

For Pb. 3.18 the formal Lagrange function reads with $W = W(I) \times [H^1(I)]^2$

$$\begin{aligned} L(y,u,\lambda) = \mathcal{J}(y,u) &+ \langle \lambda_{(a)}, Ay + Ky - Bu - C_0 \rangle_{W^*,W} + (\lambda_b, y(0,\cdot) - y_0)_{(L^2(\Omega))} \\ &+ \lambda_c^\top \partial_x y(t,0) - \lambda_c^\top \partial_x y(t,L). \end{aligned}$$

In line with Remark 2.82, it can be shown that we may choose $\lambda_c = \lambda_{(a)}$. The λ_b-term in the Lagrange function may be integrated into the weak formulation of the PDE.

We find according to Theorem 3.12 for the case $d = 1$, yielding $H^1 \hookrightarrow L^\infty$ and thus $Y = \tilde{Y}$,

Theorem 3.23 *(First-Order Necessary Optimality Conditions for Model Problem)* *Assume $[\hat{y}, \hat{u}]$ is a minimizer of the optimal control problem, Problem 3.18. Let the assumptions stated in Pb. 3.18 and Assumption 3.20 hold with $\tilde{p}_1 = \tilde{p}_2 = \hat{p}_1 = \hat{p}_2$. We write α for the diagonal matrix with $[\alpha_1, 0, \alpha_2]$ on the diagonal. Then there exist adjoints $\lambda = [\lambda_1, \lambda_2, \lambda_3]^\top \in W_1^* \times W_2^* \times H^1(I)^*$ s.t. the first-order necessary optimality conditions*

$$\begin{aligned} \langle w, (A+K)\hat{y} - B\hat{u} - C_0\rangle_{W^*,W} &= 0 && \forall w \in W^*, \\ \langle (A^* + \check{K}^*)\lambda + \hat{y} - y_{ref}, v\rangle_{Y^*,Y} &= 0 && \forall v \in Y, \\ \hat{u} &= P_{U_{ad}}\left(\alpha^{-1} B^* \lambda\right) && \textit{for a.a. } [t,x] \in \Omega_{t_f}, \end{aligned}$$

hold. For the definition of $\check{K}^*$ see Th. 3.17. Due to the regularity of solutions of the adjoint problem, we deduce $\lambda \in Y$.

Explicitly, we may rewrite the adjoint equations

$$\begin{aligned} -\lambda'_{1;t} - k\lambda''_{1;xx} + \frac{1}{|A|}\chi_A A_{21}^* \lambda_3 &= -\hat{y}_1 + y_{ref,1} && \text{a.e. in } \Omega_{t_f}, \\ \partial_\nu \lambda_1 &= 0_{L^2(I)} && \text{on } \Sigma_{t_f}, \\ \lambda_1(t_f, \cdot) &= 0_{L^2(\Omega)} && \text{on } \Omega, \\ -\lambda'_{2;t} + \check{A}_{12}^* \lambda_1 &= -\hat{y}_2 + y_{ref,2} && \text{a.e. in } I, \\ \lambda_2(t_f) &= 0, && \\ -\lambda'_{3;t} - \lambda_2 &= 0 && \text{a.e. in } I, \\ \lambda_3(t_f) &= 0, && \end{aligned}$$

and the formulas for the control

$$\hat{u}_1 = P_{U_{1,ad}}\left(\frac{1}{\alpha_1} B_1^* \lambda_1\right) \qquad \text{a.e. in } \Omega_{t_f}, \tag{3.31}$$

$$\hat{u}_2 = P_{U_{2,ad}}\left(\frac{1}{\alpha_2} B_2^* \lambda_2\right) \qquad \text{a.e. in } I, \tag{3.32}$$

where $\check{A}_{21}(t)^* = \mathcal{E}^* A_{21}(t)^* = (1/|A|)\chi_A(x) A_{21}(t)^*$ for the averaging operator $(\mathcal{E}y_1)(t)$ from Eq. 3.26 has been exploited. As usual χ_A denotes the characteristic function of the set A, being one in A and zero otherwise. Moreover, we have $\check{A}_{12}(t)^* := \fint_\Omega A_{12}(t,x)^* \lambda_1 \, dx$.

Remark 3.24 *(Combination of Bang-Bang Controls with Singular Arcs and Smooth Controls)* *For our coupled problems with two controls weighted with $\alpha_1 \geq 0$ and $\alpha_2 \geq 0$, the following situations may occur:*

a) $\alpha_1 > 0$, $\alpha_2 > 0$: Two smooth controls (3.31) and (3.32) as in the latter example.

b) $\alpha_1 = 0 = \alpha_2$: Two bang-zero-bang controls with singular arcs, $i = 1, 2$, of the type (analogously as in Lemma 2.46)

$$\hat{u}_i(t,x) = \begin{cases} u_{min,i}(t,x) & \textit{if } (B_i^*\lambda)_i(t,x) < 0, \\ \in [u_{min,i}(t,x), u_{max,i}(t,x)] & \textit{if } (B_i^*\lambda)_i(t,x) = 0 \textit{ on } \omega \subset \Omega_{t_f} \textit{ with } |\omega| > 0, \\ u_{max,i}(t,x) & \textit{if } (B_i^*\lambda)_i(t,x) > 0, \end{cases}$$

if $B_i^*\lambda$ is defined pointwise. (Note that for $i = 1$ the spatial argument x drops out here.)

c) *$\alpha_1 > 0$, $\alpha_2 = 0$ or $\alpha_1 = 0$, $\alpha_2 > 0$: One smooth control and one bang-bang control with singular arc. This might yield interesting phenomena.*

We wish to solve the system from Th. 3.23 and we proceed very similar as in Subsection 3.3.4. We eliminate the controls by inserting them into the state equations and set

$$z(t,x) = [y(t,x), \lambda(t,x)]^\top = [y_1(t,x), q(t), v(t), \lambda_1(t,x), \lambda_2(t), \lambda_3(t)]^\top.$$

We work with a slightly different notation (compared to Subsect. 3.3.4). In order to discuss the stability we work with $A_D = \partial_t - k\partial_{xx}$ being a pure diagonal operator and the complementary matrix $K_D := A + K - A_D$. Thus

$$K_D = \begin{bmatrix} 0 & A_{12} & 0 \\ 0 & 0 & -1 \\ \check{A}_{21} & 0 & 0 \end{bmatrix}, \quad \tilde{N}\lambda = \begin{bmatrix} -B_1 P_{U_{ad,1}}\left(\frac{1}{\alpha_1}B_1^*\lambda_1\right) & 0 & 0 \\ 0 & 0 & 0 \\ 0 & -B_2 P_{U_{ad,2}}\left(\frac{1}{\alpha_2}B_2^*\lambda_2\right) & 0 \end{bmatrix},$$

$$\tilde{M} := \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 0 \end{bmatrix}, \quad \mathcal{C} = \begin{bmatrix} K_D & \tilde{N} \\ \tilde{M} & \check{K}_D^* \end{bmatrix}, \quad \mathcal{I} := \begin{bmatrix} Id_3 & 0_{3\times 3} \\ 0_{3\times 3} & -Id_3 \end{bmatrix},$$

and

$$\tilde{\mathcal{A}} = \partial_t \mathcal{I} - k\partial_{xx}, \quad f(t,x) := [C_{0,1}(t,x), 0, C_{0,2}(t), y_{ref,1}(x), y_{ref,2}(x), 0]^\top,$$

where the dual $\check{K}_D^*$ (with averaging in the adjoint ODE) is defined as in Th. 3.17. Now the necessary optimality conditions yield the coupled state-adjoint system for optimal $\hat{z}$

$$\mathcal{A}\hat{z} + \mathcal{C}\hat{z} = \partial_t(\mathcal{I}\hat{z}) - k\partial_{xx} z + \mathcal{C}\hat{z} = f(t,x) \quad \text{for a.a. } [t,x] \in \Omega_{t_f},$$

that may be solved as well by using the analytic solutions as in Example 3.21 above.

Note that the order of the position and velocity equation has changed in the adjoint equation system. Using exactly the same notation as in Subsect. 3.3.4, we observe again a reversal of the coupling structure as described in Th. 3.17.

If we have $A = \Omega$, $|\Omega| = 1$, $A_{12} = a_{12} \in \mathbb{R}$ and $A_{21} = a_{21} \in \mathbb{R}$, introducing a kind of combined coupling factor $\gamma = a_{12}a_{21} \in \mathbb{R}$, the eigenvalues of $\check{K}_D^*$ are $\{-\gamma^{1/3}, \gamma^{1/3}(1 + i\sqrt{3})/2, \gamma^{1/3}(-1 + i\sqrt{3})/2\}$. Thus we note that the adjoint system for given y is unstable, iff $|\gamma| > 1$.

Treat PDE as ODE Yielding an OCP for a Coupled ODE System in Function Spaces

We apply the results from Section 3.3.3 to our coupled model problem. For our objective (3.27) we have the correspondence

$$\begin{aligned} \Phi(y(0), y(t_f)) &\equiv 0, \\ \phi_1(t,x,y,u) &= \int_\Omega \frac{1}{2}|y_1(t,x) - y_{ref,1}(x)|^2 + \frac{\alpha_1}{2}|u_1(t,x)|^2\,dx, \\ \phi_2(t, \fint_A y_1, y_2, u) &= \frac{1}{2}|y_2(t) - y_{ref,2}|^2 + \frac{\alpha_2}{2}|u_2(t)|^2, \end{aligned}$$

and we identify the right-hand sides of the differential equations. We assume $y_{ref,1} \in L^2(\Omega_{t_f})$ and $y_{ref,2} \in L^2(I) \times \mathbb{R}$.

The second-order ODE is treated as a system of two linear ODE. For the parabolic PDE with initial conditions in H and right-hand side in $L^2(I;V^*)$, we consider again a Gelfand triple $V \overset{cd}{\hookrightarrow} H \overset{cd}{\hookrightarrow} V^*$ where

$$H = L^2(\Omega), \quad V = H^1(\Omega).$$

The non-homogeneous initial conditions $y_1(0,x) = y_{1;0}(x)$ for all $x \in \Omega$ and $q(0) = q_0$ may be incorporated by a substitution into the differential equation. Thus w.l.o.g. we assume here $y_{1;0}(x) \equiv 0$ and $q_0 = 0$. Here we consider the following spaces for states and differential equations (the adjoints live in the corresponding duals)

$$\begin{aligned} Y_1 &= \{y_1 \in W(I) \,|\, y_1(0,x) = 0 \ \forall x \in \Omega\}, \\ Y_2 &= \{q \in H^2(I) \,|\, q(0) = 0\}, \quad Y_3 = \{v \in H^1(I) \,|\, v(0) = 0\}, \\ Y &= Y_1 \times Y_2 \times Y_3, \quad W = Y^*. \end{aligned}$$

The control spaces and admissible sets for controls for "Treat ODE as PDE" and "Treat PDE as ODE", resp., are

$$\begin{aligned} \tilde{U}_1 &= L^2(\Omega), \quad \tilde{U}_2 = \mathbb{R}, \\ U_1 &= L^2(I;L^2(\Omega)), \quad U_2 = L^2(I), \quad U = U_1 \times U_2, \\ \tilde{U}_{ad,1} &= \{u_1 \in \tilde{U}_1 \,|\, u_{1,min} \leq u_1 \leq u_{1,max} \text{ for a.a. } x \in \Omega\}, \\ \tilde{U}_{ad,2} &= \{u_2 \in \tilde{U}_2 \,|\, u_{2,min} \leq u_2 \leq u_{2,max}\}, \quad \tilde{U}_{ad} := \tilde{U}_{ad,1} \times \tilde{U}_{ad,2}, \\ U_{ad,1} &= \{u_1 \in U_1 \,|\, u_{1,min} \leq u_1 \leq u_{1,max} \text{ for a.a. } [t,x] \in \Omega_{t_f}\}, \\ U_{ad,2} &= \{u_2 \in U_2 \,|\, u_{2,min} \leq u_2 \leq u_{2,max} \text{ for a.a. } t \in I\}, \quad U_{ad} := U_{ad,1} \times U_{ad,2}. \end{aligned}$$

Note that the Neumann boundary conditions are included in the weak formulation and that $V^* = H^1(\Omega)^*$. As usual we identify $H^* = H$. We recall that we have operators $A \in \mathcal{L}(Y_1, Y_1^*) \times \mathcal{L}(Y_2, Y_2^*)$ and $B \in \mathcal{L}(U_1, H) \times \mathcal{L}(U_2, \mathbb{R})$.

In this context the Hamilton function from (3.18) for coupled differential equations, introducing (as in the last subsubsection) a further state $y_3 = v$ with the corresponding adjoint λ_3 in order to rewrite the second-order ODE in the desired form, reads

$$\begin{aligned} \mathcal{H}(t,y,u,\lambda) = \frac{1}{2}\int_\Omega |y_1(t,x) - y_{ref,1}(x)|^2\,dx + \frac{1}{2}|y_2(t) - y_{ref,2}|^2 + \frac{\alpha_1}{2}\int_\Omega |u_1(t,x)|^2\,dx \\ + \frac{\alpha_2}{2}|u_2(t)|^2 + \langle \lambda_1(t,\cdot), -ky''_{1,xx}(t,\cdot) + A_{12}(t,\cdot)y_2(t) - B_1u_1(t,\cdot) - C_{0,1}(t,\cdot)\rangle_{V^*,V} \\ - \lambda_2(t)^* y_3(t) + \lambda_3(t)^* \left(\check{A}_{21}(t)y_1(t,x) - B_2u_2(t) - C_{0,2}(t)\right) \\ \text{for a.a. } t \in I. \end{aligned}$$

Here we exploited that we have according to Th. 3.22 a bounded inverse and thus we may suppose that the Robinson constraint qualification holds in this setting, i.e. $\lambda_0 = 1$ may be assumed.

For the minimizer $[\hat{y}, \hat{u}]$ we have the necessary optimality conditions as in Th. 3.15, yielding

$$\begin{aligned}
\langle \lambda'_{1,t}(t), v_1 \rangle_{V^*,V} = k(\lambda'_{1,x}(t), v'_{1,x})_H + \frac{1}{|A|} A^*_{21}(t)\lambda_3(t)\, \langle \chi_A, v_1 \rangle_{V^*,V} & \\
+ \langle \hat{y}_1(t) - y_{ref,1}, v_1 \rangle_{V^*,V} & \qquad \forall v_1 \in V \text{ a.e. in } I, \quad (3.33) \\
\lambda_1(t_f, \cdot) = 0_H, & \qquad (3.34)
\end{aligned}$$

$$\begin{aligned}
\lambda'_{2,t}(t) &= \langle A^*_{12}(t)\lambda_1(t,\cdot), 1 \rangle_{V^*,V} + \hat{y}_2(t) - y_{ref,2} && \text{a.e. in } I, && (3.35) \\
\lambda_2(t_f) &= 0, && && (3.36) \\
\lambda'_{3;t}(t) &= -\lambda_2(t) && \text{a.e. in } I, && (3.37) \\
\lambda_3(t_f) &= 0, && && (3.38)
\end{aligned}$$

and

$$\begin{aligned}
0 &\leq \langle -B^*_1\lambda_1(t) + \alpha_1\hat{u}_1(t), u_1 - \hat{u}_1 \rangle_{U^*_1,U_1} && \forall u_1 \in \tilde{U}_{ad,1} \text{ a.e. in } I, && (3.39) \\
0 &\leq (-B^*_2\lambda_2(t) + \alpha_2\hat{u}_2(t))^\top (u_2 - \hat{u}_2(t)) && \forall u_2 \in \tilde{U}_{ad,2} \text{ a.e. in } I, && (3.40)
\end{aligned}$$

where $\lambda \in W^*$.
Again, we check that the coupling structure of the state equations is reversed in the adjoint problem.

Remark 3.25 *(Treat ODE as PDE vs. Treat PDE as ODE in Function Spaces)*
The question remains, what is the difference between the approaches relying on the Lagrangian or on the Hamiltonian, respectively. In the latter setting, the ODE in function spaces approach, we obtain pointwise necessary optimality conditions, but possibly in different weaker spaces. Under some assumptions, i.e. the Gelfand triple structure (as in Def. A.17 for evolution problems), a coercive continuous bilinear form, and essentially bounded A_{12}, A_{21}, both formulations are again equivalent [HPUU09, 1.3.2.4].

However, comparing the two approaches "Treat ODE as PDE" and "Treat PDE as ODE (in function spaces)", we suggest to consider preferably ODE as PDE. Our advice is based on numerical experience, as for the truck-container problem (see Sect. 4.2) and, moreover, it seems technically to be more intuitive, e.g., when applying our semismooth Newton method (see Sect. 2.9). Note that the approach "Treat PDE as ODE" is only available for PDE that are evolution equations, other PDE can be treated as DAE though.

Note that without taking a mean over A in the ODE and considering an ODE for every space point $x \in \Omega$ instead, we end up with similar necessary optimality conditions.

Remark 3.26 *(First-Order Necessary Optimality Conditions for Model Problem Without Averaging Operator)*
If we consider directly $y_1(t,x)$ instead of the average $\fint_A y_1(t,x)\, dA(x)$ and consider an ODE for

every space point $x \in \Omega$, the adjoint equation (3.33) reads

$$\begin{aligned}\langle \lambda'_{1,t}(t), v_1 \rangle_{V^*,V} &= k(\lambda'_{1,x}(t), v'_{1,x})_H + \langle A_{21}(t)^* \lambda_3(t), v_1 \rangle_{V^*,V} \\ &\quad + \langle \hat{y}_1(t) - y_{ref,1}, v_1 \rangle_{V^*,V} \qquad \forall v_1 \in V \ \textit{a.e. in } I.\end{aligned}$$

If $|A| = 1$ and $A = \Omega$ both approaches yield the same adjoint equations.

Chapter 4

Applications in Engineering and Science for Optimal Control of Coupled ODE-PDE Systems

In this chapter we discuss several applications in engineering and science, where optimal control problems of coupled systems appear. The presented real-world problems differ in particular w.r.t. the coupling structure and the underlying type of PDE. These examples shall illustrate the variety and complexity of optimal control problems of this class and the issues arising additionally in modelling. On one hand we have to emulate the important features of a real-world problem and on the other hand we wish the model to fit in a certain mathematical framework and numerically feasible methods should be available (for validation and optimization), too.

Challenging issues w.r.t. modelling (averaging-evaluation operators), analysis (well-posedness of coupled systems; fixed-point iterations), and numerics (e.g. dealing with non-differentiability) appear for these problems. Interesting phenomena, like the reverse coupling structure or the question, whether the bang-bang principle is violated or not, that is not met in optimal control of ODE or optimal control of PDE, may be observed. Both latter issues are not treated in literature to the best knowledge of the author.

In this part of the study we connect the new theory of Sect. 2.9 on our globalized semismooth Newton method and of the last chapter with modelling and numerics for real-world problems.

4.1 Classification of the Presented Real-World Examples

In this section we give an overview of the examples in this chapter and we classify these examples w.r.t. the type of optimization problem; the underlying differential equations; the type of coupling; the type of constraints; the applied optimization approach; and the used discretization methods.

From the examples treated by myself, we focus on the problems:

1) Truck-container problem, see Section 4.2,

2) Elastic structure-load problem, see Section 4.3,

3) Elastic tyre-damper system with road contact (quarter car model), see Section 4.4,

4) Nanodroplets/-bubbles evolution, see Section 4.5.

Problem 2 is split into two related problems, Problem 2a, being the elastic crane-trolley-load problem, and Problem 2b, the elastic bridge-load problem. For the description of the other examples, we refer to the introductory and the outlook chapter. The selected problems exhibit different features of coupled optimal control problems and cover different topics in optimization.

We have included the originally published versions of our articles. Please note that the notation may vary, that is also due to the application context or the journal that suggests a certain standard notation. Important changes in the notation are indicated in the introductory text of each subsection.

Moreover, in this chapter we will discuss the validity of the assumptions in theory, as discussed in Chapter 3, and further prerequisites and requirements.

We file the listed problems by characteristic features in the following. For a first classification of these and related problems, please see Tables 1.1 and 1.2.

At first we consider the type of optimization problem. All problems are optimal control problems or parameter identification problems. Note that the free terminal time is treated as a parameter to be identified as well.

Problem 1 is considered as a time-optimal control problem in [GK15, WGKG18] and as an optimal control problem with fixed terminal time in [KG16]. Except for the possibly present free terminal time term, the objective consists of tracking-type terms, the control effort and possibly penalty terms for terminal conditions. Concerning penalty terms, it may help to consider the penalties such that the structure of an augmented Lagrangian is obtained, see [KG16]. The case of neglecting tracking-type terms and control effort is considered in [GK15] as well.

In [KGH18a, KGH18b] Problem 2 is considered for a free terminal time, while in [Ki16] and in Problem 3 the terminal time is fixed. For the elastic crane-trolley-load system we have kinetic energy terms and control costs in the objective, in addition to the free terminal time term. Terminal constraints are treated as penalties by considering the augmented Lagrange function. In [KGH18a] we consider the case without control costs and free terminal time as well as a fixed terminal time with control costs, whereas in [KGH18b] we have a free terminal time, a non-zero weight for the kinetic energy and for the control effort. For the elastic bridge-load system the objective is to minimize the maximal absolute displacement.

In Problem 3 the objective is the linear combination of a comfort term (the chassis acceleration), a spring robustness term (the difference of chassis and tyre displacement), and a safety term (loosely speaking the contact between tyre and road). The first and the third terms are of tracking-type, while the second might be interpreted as a penalty term.

Problem 4 is an optimal control problem combined with another parameter identification for a fixed terminal time. Here the objective is to minimize a linear combination of the control costs and, at the terminal time t_f, the total number and total volume as well as the deviation from the mean volume. The latter terms are tracking-type terms at t_f.

Furthermore, we may classify w.r.t. the present differential equations. The type of PDE or of system of PDEs might be parabolic, elliptic, hyperbolic or even something else, each with different orders of the PDE and the PDE might be linear, semilinear, quasilinear, or fully nonlinear. ODE or systems of ODE may be sorted w.r.t. order and linear behaviour as well. In principle DAEs or PDAEs (with methods dependent on the index) might appear as well. Algebraic equations may arise, e.g, by constraining forces or free boundary conditions.

In Problem 1 we have a system of two quasilinear hyperbolic PDE of first-order (in 1D) and it may be treated as a system of two semilinear (second-order) parabolic PDE after a regularization by an artificial viscosity. The ODEs here are given by Newton dynamics and are considered taken for itself (i.e. w/o coupling) as a system of 2 linear differential equations of first-order.

For the elastic structure in Problem 2 we consider as PDE either the Lame system, an elliptic linear system of 3 PDE of second-order, or the plate equation of fourth-order, that is a linear elliptic PDE. For the elastic crane, the differential equations of trolley and load are given by Newton dynamics and are linear for the trolley itself, but nonlinear for the load, that is an differential equation for a pendulum. For the elastic bridge-load problem, a simple linear ODE is assumed that is eliminated actually in the modelling process.

In Problem 3 we consider the elliptic Lame system in 2D as PDE (of second order) that is coupled to a linear ODE of second order. Furthermore, this is coupled to a free boundary condition that constitutes in principle a complementarity condition for the PDE, but within the so-called Hertz approximation it may eliminated by algebraic equations.

The last problem, Problem 4, is of different type, since we have a hyperbolic first-order PDE for a measure and a single algebraic equation (of index one) for the only time-dependent so-called mean field.

The type of coupling might be fully or one-sided, the latter allowing for solving one differential equation and then inserting the solution into the other that yields an important reduction of complexity. In all problems considered here the coupling is non-trivial, except for the elastic structure-load problem, and the coupling may introduce further nonlinear dependencies. Furthermore, we distinguish between the type of control, for a PDE typically the boundary control (on a Neumann, Dirichlet, Robin, or radiation boundary) and the distributed control are considered. A moving boundary condition that is controlled may also arise, see, e.g., our Problem 2. An ODE may be controlled by a force term that corresponds to a distributed control. In addition, differential equations may be controlled by coefficients as well.

In Problem 1, the force term of the ODE may be controlled. On one hand a Dirichlet boundary value for the PDE state h, i.e. the fluid level at the container boundary (this fluid level is not given explicitly by a boundary condition for the PDE), enters into the ODE for the acceleration. On the other hand the ODE states enter by means of a Neumann boundary into the PDE for h and in a force term into the PDE for the horizontal fluid velocity v as well. Thus we have full coupling in the truck-container problem.

In the elastic crane trolley load problem the control enters into the trolley ODE as a force term and into the Neumann boundary condition of the PDE as well. The ODE for the load is fully coupled to the trolley ODE. The solution of the trolley ODE determines the moving Neumann boundary condition that enters the PDE, while averages over derivatives of the PDE solution enter the ODE system as coefficients and in the force term. For the elastic bridge-load problem, we have a one-sided coupling from the position of the loads, being for instance trucks, to the PDE solution, i.e. the displacement. Here the goal is to find an optimal strategy for the number of trucks, its driving direction, and, in particular, its distance, while not exceeding a maximal displacement of the bridge.

In Problem 3 the damping coefficient in the ODE may be controlled. By means of the spring-damper force the ODE solution is coupled to the Neumann b.c. and to the Dirichlet b.c. at the wheel rim base as well, yielding the mechanical displacement field at this boundary and the shift y of the rim that enters into the ODE as well. If we exploit the approximation by the Hertzian stress, then the free contact boundary of the tyre with the road may be stated explicitly and we can replace the PDE by another ordinary differential for y that corresponds to a spring as substitute model for the tyre.

In the optimal control of mean field problems, we may control a parameter, the initial mass, and the temperature that enters into several coefficient functions in the PDE for the measure and into the algebraic equation for the mean field that is differentiated for further analysis. The mean field enters into the advection coefficient of the PDE and the measure into the algebraic equation for the mean field, thus we have a full coupling.

Furthermore we classify by the type of constraints, control constraints, state constraints, or mixed constraints. For Problem 1 the scenarios considered for [GK15, Fig. 1–3] and in [KG16] box constraints for the control are prescribed, whereas in the considered scenarios the state constraints do not get active. In the scenarios for [GK15, Fig. 4] and in [WGKG18] we consider box constraints for the control and box constraints for the state h, being the fluid level, simultaneously. For the elastic crane-trolley-load problem [KGH18a, KGH18b] we consider again box constraints for the control and the state constraints may be ignored safely since they do not get active in the considered parameter ranges. Again in Problem 3 and Problem 4 box constraints for the control and for the parameter, resp., are considered. In Problem 4 we require pure state constraints in order to guarantee the non-negativity of droplet volumes.

For Problem 1 a first-discretize-then-optimize approach relying on the Lagrange formalism is considered in [GK15, WGKG18]. In [GK15] an adjoint-based gradient method for the reduced problem is exploited, while in [WGKG18] a SQP method (with structure exploitation) based on the Hessian is applied to an all at-once approach. In [KG16] the optimal control problem is considered in an all-at-once approach. Here a first-discretize-then-optimize approach relying again on the Lagrange formalism is pursued in order to derive necessary first-order optimality conditions that are semi-discretized in time and then solved by the globalized semismooth Newton method [GHK17].

For the elastic crane-trolley-load problem [KGH18a, KGH18b] the reduced optimal control problem is tackled by means of a first-discretize-then-optimize approach. The projected gradient is calculated by means of a sensitivity-based approach since the adjoint equations have a complicated form due to averages over derivatives. Here for Newton-type methods we do not have the required smoothness due to moving boundary conditions implying measure-valued F-derivatives for the control-to-state operator [KGH18b].

In Problem 3 we pursue a first-discretize-then-optimize approach and consider a sensitivity-based, projected gradient method (without and with BFGS update) for an reduced objective.

In the last Problem we consider a customized first-discretize-then-optimize approach, where at first a multi-scale approach is used for model reduction. The PDE for the measure is replaced by a large number of ODEs coupled to the mean field. This reduction method may be interpreted as the special case of initial data of Dirac type and yields a so-called mean field model. This reduced model is optimized by a projected gradient method using a sensitivity-based approach.

Note that the Pontryagin minimum principle is not exploited in these examples at all. The elastic bridge-load problem is of different type anyways since the control is examined by a try-out approach.

Concerning the discretization of the differential equations and the implementation, in Problem 1 we apply a Lax-Friedrich scheme [La06, Sect. 8.1] to the Saint-Venant system and an explicit Euler for the ODE. This is implemented in the solver sqpfiltertoolbox from the software package OCPID-DAE1 [Ge10] in [GK15], in [KG16] the commercial software MATLAB is used, and in [WGKG18] the commercial solver SNOPT as well as sqpfiltertoolbox are considered.

For Problem 2 we use the finite element method (FEM) for the elasticity equation in the structure and, in case of the crane, the finite difference method (FDM), the Heun method to be precise, for the ODE system. Since we need second-order derivatives on some boundary parts of the crane beam, we work with quadratic Lagrange elements. This has been implemented in the open software package FEniCS [LMW12] and, in case of the bridge, in MATLAB, too.

For the elastic-tyre-damper road problem, Problem 3, we use the finite element method for the PDE in the tyre and a time-stepping scheme for the spring-damper ODE. Using the quadratic Lagrange elements $CG_2(=\mathbb{P}_2)$ and the explicit Heun method this has been implemented in FEniCS.

Problem 4 has been implemented in OCODE 1.5, a software code by Matthias Gerdts later included in the package OCPID-DAE1 [Ge10], using central differences as updates for the states and trying out different solvers for differential algebraic equations (Runge-Kutta with fixed suitably small step sizes, DASSL). Here the numerical conservation of mass is crucial.

Note that there are further types of coupled optimization problems, e.g., in optimal design the control enters as diffusion coefficient of a partial differential equation (see e.g. [HPUU09, Sect. 1.1.5]). Instead of a single PDE, optimal design problems may be subject to a coupled system of differential equations as well, the ODE might denote, e.g., a phase field.

In our model problem, Problem 3.18 we have considered a simple structure, i.e. a linear ODE and PDE with full coupling. We start with the truck-container problem that is similar in its structure, but exhibits nonlinear terms in the PDE and a coupling structure involving boundary conditions. By means of this problem we may illustrate again the challenges and phenomena of optimal control of coupled differential equations, e.g., the reversal of the coupling structure in the adjoint system. In Problem 2, some numerical results [KGH18a] seem to underline the conjecture from Pesch *et al.* [PRWW10] that the bang-bang principle known from classical optimal control might be violated for coupled optimal control problems. However, we cannot exclude that this is due to numerical artifacts.

Only in the following sections, Sections 4.2 and 4.3, we underline vectors and matrices, e.g., $\underline{v}$ or $\underline{\underline{A}}$.

4.2 Article: Optimal Control of a Truck Carrying a Fluid Container

In this article the optimal control problem for the truck-container model is considered. At the beginning the model for the truck-container problem is derived shortly. The fluid (e.g. water) in the container is modelled by the Saint-Venant equations in 1D (or 2D might be interesting as well) and the vehicle dynamics in 1D (here 2D or 3D are possible) of the truck by the ODE due to Newton. The container may move in the cargo bay, since it is loosely fixed to the truck that is modelled by a spring-damper element. The PDE states are the vertical fluid level h and the horizontal fluid velocity v and the ODE states are positions d_X and velocities $\dot{d}_X$ of the truck ($X = tr$) and the water container ($X = w$), respectively. We may control the acceleration/deceleration of the truck, i.e. the force term in the ODE for $\dot{d}_{tr}$. We have a full coupling, since

$$\text{Dirichlet boundary value } h \rightsquigarrow \dot{d}_w(\text{force term})$$

by the momentum of the container exerted on the truck and there holds

$$\underline{d}, \dot{\underline{d}} \rightsquigarrow v \text{ (force term) and } \underline{d}, \dot{\underline{d}} \rightsquigarrow \text{Neumann boundary value } h_x,$$

due to the spring-damper force acting on the container. A tracking-type objective with a Tikhonov regularization is minimized subject to this fully coupled differential equation system, consisting of a system of hyperbolic first-order equations that are quasilinear and two ODE. By introducing an artificial viscosity the PDEs may be considered as a system of semilinear parabolic equations of second-order. Furthermore, we consider box constraints for the control and box constraints for the states. We follow a first-optimize-then-discretize strategy in an all-at-once-approach, relying on the Lagrange method. The existence of optimal controls and the necessary optimality conditions of first-order are deduced rigorously. The St-Venant equations are discretized by the Lax-Friedrich scheme, where the artificial viscosity appears naturally, see, e.g., [Kr97, Vr98] for details. This is an explicit method and requires that a CFL condition is met in order to be stable, for the ODE we use the explicit Euler. Then the KKT-system is semi-discretized in time and the numerical computation of the control is performed by a semismooth Newton method with a suitable globalization strategy implemented in MATLAB.

In the following subsection we recapitulate the Saint-Venant equations for convenience.

4.2.1 Saint-Venant Equations

The Saint-Venant equations (shallow water equations) read in general form including friction [BC16, Sect. 1.4]

$$\partial_t h + \partial_x (hv) = 0, \tag{4.1}$$

$$\partial_t v + \partial_x \left(\frac{v^2}{2} + gh \right) + \left(c\frac{v^2}{h} - gB'_x - f \right) = 0. \tag{4.2}$$

The first equation is the mass balance, the second the momentum balance. The unknowns are $h(t,x)$, the height of the fluid level (fluid depth), and $v(t,x)$, the horizontal fluid velocity (here 1D). More precisely, v is the horizontal fluid velocity averaged over a vertical column of fluid. c is a constant friction coefficient, $B(x)$ denotes a given bottom profile, and $f(t)$ is the external acceleration (e.g. due to the spring-damper-element, fixing the container). (4.2) simplifies in case of a constant slope $B'_x = S_b$.

The 1D Saint-Venant equations constitute a one-dimensional hyperbolic balance law[1] of the form $\partial_t \underline{y}(t,x) + \partial_x \underline{f}(\underline{y}(t,x)) + \underline{g}(t,x,\underline{y}(t,x)) = \underline{0}$, where y has values in $\mathcal{Y}$, $\mathcal{Y}$ a connected open subset of $\mathbb{R}^d$. Transport processes of various kinds (of electrical energy, the flow of fluids in open channels/gas pipelines, light propagation in optical fibers, road traffic, etc.) may be typically represented by hyperbolic partial differential equations. If the transport process of interest exhibits a dominant coordinate dimension, the dynamics may be modelled by one-dimensional hyperbolic balance laws. [BC16, Preface].

The St-Venant system, written with

$$\underline{y} = \begin{bmatrix} h \\ v \end{bmatrix}, \quad F(\underline{y}) = \begin{bmatrix} v & h \\ g & v \end{bmatrix}, \quad \underline{g} = \begin{bmatrix} 0 \\ cv^2/h - gB'_x - f \end{bmatrix},$$

reads in matrix-vector-notation

$$\partial_t \underline{y} + F(\underline{y}) \partial_x \underline{y} + \underline{g}(\underline{y}) = 0.$$

[1] A conservation law is a special case of a balance law, where no "force" term g enters the equation.

The Froude number is defined as

$$Fr := \frac{v(t,x)}{\sqrt{gh(t,x)}}$$

and describes the relation between the fluid velocity and the phase velocity (of a long wave). The flow is called *subcritical (fluvial)*, if $Fr < 1$. Then the flow may be characterized as slow, but deep, and no hydraulic jumps can appear. Furthermore, then the system is indeed hyperbolic that can be checked by computing the corresponding eigenvalues of F.

We prescribe initial conditions $h(0,x) = h_0(x)$, $v(0,x) = v_0(x)$ for given initial fluid level h_0 and fluid velocity v_0. The hyperbolic PDE needs two boundary conditions, on each end, to be closed, e.g.:

- if the "pool" is closed, with "pumps" at both ends:

  $$h(t,0)v(t,0) = U_0(t), \tag{4.3}$$

  $$h(t,L)v(t,L) = U_0(t). \tag{4.4}$$

  The flow rates U_0 and U_L are given or control variables.

- tunable hydraulic gates (as in irrigation or navigable canals with a flow direction from $x = 0$ to $x = L$) in standard hydraulic models:

  for overflow gates:

  $$\begin{aligned} h(t,0)v(t,0) &= c_g(Z_0(t) - U_0(t))^{3/2}, \\ h(t,L)v(t,L) &= c_g(h(t,L)) - U_0(t))^{3/2}, \end{aligned}$$

  for underflow gates (sluices):

  $$\begin{aligned} h(t,0)v(t,0) &= c_g U_0(t)\sqrt{Z_0(t) - h(t,0)}, \\ h(t,L)v(t,L) &= c_g U_L(t)\sqrt{h(t,L) - Z_L(t)}. \end{aligned}$$

  Again U_0 and U_L are possible controls (the elevations or apertures, respectively), Z_0 and Z_L (fluid level outside the pool at the other side of the gate) might be considered as disturbance inputs. Here c_g is a constant "discharge" coefficient for all types of gates.

For the truck-container-system the two boundary conditions read

$$\begin{aligned} \partial_x h(t,0) &= -B_x - \frac{1}{g}f(t), \\ \partial_x h(t,L) &= -B_x - \frac{1}{g}f(t), \\ v(0) &= 0, \\ v(L) &= 0. \end{aligned}$$

The latter two b.c. correspond to (4.3) & (4.4) in case of a closed pool without pumps. Note that we are here in the non-characteristic case.

Our optimal control problem is more general than the problems covered by Borsche and others [BCG10], since we allow for a distributed control in the hyperbolic conservation laws as well.

In [DPR99, Co07] the optimal control of a container that is fixed to a moving conveyor belt is treated. This model exhibits a one-sided coupling only, whereas we consider a spring-damper element that re-couples the container to the truck.

Court et al. [CKP18] consider the distributed optimal control of hyperbolic conservation laws without any coupling to an ODE, but with a geometric parameter, being a point in the domain, to be determined in addition to the optimal control. The objective has a terminal cost term and the goal is to maximize the height of a shallow water wave at this unknown point.

The derivation of the truck-container model, involving the Saint-Venant equations as PDE and the vehicle dynamics as ODE can be found in details in [GK15]. For shortness here only our original paper [KG16] is depicted, wherein the existence of optimal control and the necessary optimality conditions are demonstrated. The globalized semismooth Newton method referred to in this paper can be found in [GHK17] that is presented at the end of Chapter 3 in full length.

We shortly comment on the paper [WGKG18] and on the work in progress [KW18]. In the first paper we follow an all-at-once-approach and the numerical optimal control by a first-discretize-then-optimize approach is performed using the SQP method. The efficiency of the computations is increased by using exact first- and second-order derivative information, i.e. the structure of Jacobian and Hessian is exploited, and further techniques, like primal or dual regularization. In the upcoming study [KW18] the truck-container model is reformulated for a 2D fluid level, yielding two momentum balances for two horizontal fluid velocities, v and w, and the truck dynamics are in 3D, modelling a drive on a (given) route through a real landscape. Interesting would be to simulate and control a steep twisted mountain road. Motivated by the optimal solution found in [GK15, KG16, WGKG18] a model predictive control will be considered. The idea is to exploit the turnpike behaviour of the control that is suggested by our numerical computations. For the study of turnpike solutions in infinite-dimensional control systems see, e.g., [Za00].

4.2.2 Comparison with Abstract Theory

We compare our publication [KG16] with our results in Chapter 3. In [KG16], where after a scaling to a fixed time interval $(0,1)$, the spaces for the states read

$$Y = L^2(0,1;H^1(0,L)) \times L^2(0,1;H_0^1(0,L)) \times [H^2(0,1)]^2 \times [H^1(0,1)]^2$$

and

$$\begin{aligned}\tilde{Y} &= [H^1(0,1;L^2(0,L)) \cap L^2(0,1;H^2(0,L)) \cap L^\infty(0,1;H^1(0,L))] \\ &\quad \times [H^1(0,1;L^2(0,L)) \cap L^2(0,1;H^2(0,L)) \cap L^\infty(0,1;H_0^1(0,L))] \times [H^2(0,1)]^2 \times [H^1(0,1)]^2.\end{aligned}$$

In the latter space higher regularity in the truck-container example is exploited, this is not possible in the general setting of CDE presented in this study. The control space $U = L^2(0,1)$

is equipped with standard box constraints, i.e. $U_{ad} \subset U$ is a closed convex subset of U. We find the structure of a Gelfand triple, having $H = L^2(0, L)$ and $V = Y$.

The existence and uniqueness of solutions in Y for the coupled state equations (under basic regularity assumptions on the data) is proven analogously to Th. 3.4 for sufficiently small terminal times. The approach to consider ODEs as PDEs is pursued. For the existence of optimal controls we may apply again Th. 2.41, for the necessary optimality conditions of first-order, please see Th. 3.6.

We compare formally the scaled problem for the states

$$\underline{y} = [h, v, d_\Delta = d_{tr} - d_w, d_w, v_\Delta = v_{tr} - v_w, v_w]^\top$$

(with a slightly different coupling structure) in [KG16] with the setting of Problem 3.7, changing the notation of the paper, if it differs, i.e., T becomes t_f. Note that this here is a problem with a nonlinear PDE, non-zero boundary conditions for h, and homogeneous Dirichlet b.c. for v in contrast to Problem 3.7. We have $n_{y_1} = 2$, $n_{y_2} = 4$, $n_{u_1} = 0$ (i.e. no direct control of the PDE system), $n_{u_2} = 1$, and $k = \varepsilon$, being the (fixed) artificial viscosity. We identify the operators as (rewriting the Neumann b.c. for h as Dirac distributions)

$$A_1 = t_f \begin{bmatrix} v\partial_x & h\partial_x \\ g\partial_x & v\partial_x \end{bmatrix}, \quad C_1 = -\eta(\tilde{c}d_\Delta + \tilde{k}v_\Delta) \begin{bmatrix} \frac{1}{g}\mathcal{E}(h) \\ t_f v \end{bmatrix},$$

$$A_2 = t_f \begin{bmatrix} 0 & 0 & -1 & 0 \\ 0 & 0 & 0 & -1 \\ -\tilde{c} & 0 & -\tilde{k} & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix}, \quad B_2 = \begin{bmatrix} 0 \\ 0 \\ \frac{1}{m_{tr}} Id_U \\ 0 \end{bmatrix}, \quad C_2 = -t_f \frac{g}{L}\mathcal{E}(h) \begin{bmatrix} 0 \\ 0 \\ 1 \\ 1 \end{bmatrix}.$$

The evaluation operator $\mathcal{E}(h) = h(t, \cdot)|_0^L$ in this problems evaluates h at $x = 0$ and $x = L$. Here w.l.o.g. $\bar{d} \equiv 0$ for the offset and $B \equiv 0$ for the bottom profile of the container. B_1 does not appear.

Note that the nonlinearities in the PDE yield that A_1 depends on y_1 as well. However, in the analysis this may be handled by some standard embedding theorems.

In the objective we have the penalty terms of an augmented Lagrange function,

$$\Phi = \sum_{I \in \{\Delta, w\}} \left(\frac{\alpha_4}{2} |d_I(1) - d_I^{(T)}|^2 + \sigma_4(d_I(1) - d_I^{(T)}) + \frac{\alpha_5}{2} |v_I(1) - v_I^{(T)}|^2 + \sigma_5(v_I(1) - v_I^{(T)}) \right),$$

and

$$\phi_1 = \frac{\alpha_1}{2} |h(t, x) - h_d(x)|^2 + \frac{\alpha_3}{2} v(t, x)^2$$

for tracking a fluid level $h_d \in L^2(0, L)$ and zero horizontal velocity, and for the control effort

$$\phi_2 = \frac{\alpha_2}{2} u(t)^2.$$

Furthermore, for our comparison here we ignore the state constraints on h assuming that they do not get active (that is realistic for certain data), since the result in Subsection 3.3.2 is not formulated for state or mixed state-control constraints.

We have $\mathcal{E}_h' = \delta_x(\cdot)|_0^L$, where δ_x is the Dirac distribution. Thus

$$\begin{aligned} C_{1;y_{1,1}} &= -\eta(\tilde{c}d_\Delta + \tilde{k}v_\Delta)\begin{bmatrix} \delta_x(\cdot)|_0^L/g \\ 0 \end{bmatrix}, \\ C_{1;y_{1,2}} &= -\eta(\tilde{c}d_\Delta + \tilde{k}v_\Delta)\begin{bmatrix} 0 \\ t_f \end{bmatrix}. \end{aligned}$$

Since additionally $C_{1;y_{2,1}} = -\eta\tilde{c}[\mathcal{E}(h)/g, t_f]^\top$, $C_{1;y_{2,3}} = -\eta\tilde{k}[\mathcal{E}(h)/g, t_f]^\top$, and $C_{1;y_{2,\cdot}}$ is zero otherwise and since $C_{2;y_{j,k}} = -t_f\frac{g}{L}\delta_x(\cdot)|_0^L[0,0,1,1]^\top$ for $j = k = 1$ and zero otherwise, we see that Assumption 3.9 is fulfilled. However, Assumpt. 3.11 does not cover the additional terms in the augmented Lagrange function. (The weights α_i, $i = 1,3,4,5$, that do not appear in Assumpt. 3.11 could be included there in principle.)

In a neighbourhood of the solution, the objective is F-differentiable with locally Lipschitz derivative as well as the PDE operator E is F-differentiable. For a further discussion see also the end of Subsection 3.3.1.

Volume 1, Number 2, 2016, 1–

NECESSARY OPTIMALITY CONDITIONS AND A SEMI-SMOOTH NEWTON APPROACH FOR AN OPTIMAL CONTROL PROBLEM OF A COUPLED SYSTEM OF SAINT-VENANT EQUATIONS AND ORDINARY DIFFERENTIAL EQUATIONS

SVEN-JOACHIM KIMMERLE* AND MATTHIAS GERDTS

Abstract. In this article necessary optimality conditions for Saint-Venant equations coupled to ordinary differential equations (ODE) are derived rigorously. The Saint-Venant equations are first-order hyperbolic partial differential equations (PDE) and model here the fluid in a container that is moved by a truck that is subject to Newton's law of motion. The acceleration of the truck may be controlled.

We describe the mathematical model and the corresponding tracking-type optimal control problem. First we prove existence and uniqueness for the coupled ODE-PDE problem locally in time. For sufficiently small times, we derive the first-order necessary optimality conditions in the corresponding function spaces. Furthermore we prove existence of optimal controls.

The optimality system is formulated in the setting of a semi-smooth operator equation in Hilbert spaces which we solve numerically by a semi-smooth Newton method. We close with a numerical example for a typical driving maneuver.

1. Introduction

We consider a container with a fluid that is subject to the Saint-Venant equations, that model fluid flow in shallow water. The container is mounted on a vehicle that is subject to Newton's law of motion. Our optimal control problem (OCP), see Subsection 2.1, is to minimize tracking type functionals for the fluid height and the horizontal fluid velocity and to minimize the control costs, subject to the coupled ODE-PDE system and box constraints for the control. As a prototype example we consider a truck with a fluid container as load, which are coupled by a spring-damper element. We derive the corresponding model in Section 2 in detail. Our problem exhibits partial differential equations, i.e. the Saint-Venant equations, that are fully coupled to ordinary differential equations, given by Newton's law of motion. We may control the acceleration of the truck, that by means of the spring-damper force, acts as an indirect Neumann boundary control on the fluid height and as a distributed control on the fluid velocity. Conversely, the momentum of the moving fluid in the

2010 *Mathematics Subject Classification.* 35Q35, 49K15, 49K20, 49J15, 49J20, 49J52, 49M15, 90C46, 35L04, 49N90, 76B15, 90C56, 65K10.

Key words and phrases. ODE-PDE constrained optimization, Saint-Venant equations, coupled ODE-PDE system, necessary optimality conditions, semi-smooth Newton method, adjoint optimal control, hyperbolic conservation law, vehicle dynamics, numerical optimal control.

* Corresponding author.

container affects the motion of the truck. The scaled problem is summarized in Subsection 2.2.

Optimal control problems for Saint-Venant equations without coupling to ODE have been considered in [3, 28], for instance. However, they consider distributed controls of the fluid level equation and only control costs. A similar control problem with a tracking-type objective for a given fluid level profile h_d, where a container is accelerated directly (representing, e.g., a container fixed on a moving horizontal conveyor) is considered by Coron *et al.* [6]. Global boundary controllability of the Saint-Venant equations between steady states is demonstrated by Gugat and Leugering [12]. Moreover, they have considered the controllability of the Saint-Venant equations in the situation of sloped canals with friction [13]. In their result, it turns out to be crucial that the considered terminal time is not too small. The latter is due to the finite propagation velocity. We note that we do not consider such a Dirichlet boundary control for h in our model.

Coupled systems involving ODEs as well as PDEs and their control have been considered only for particular examples. Hömberg *et al.* [8, 17, 18] and Gupta *et al.* [14] consider a model for laser hardening of steel, involving the heat equation and a differential equation describing phase transitions. In another example, in a gallium-arsenide crystal the phase transition of arsenic-rich droplets is modelled by an ODE for the free boundary and by the quasi-linear diffusion equation. This is further coupled to the PDE of linear elasticity, modelling mechanical stresses within the crystal. For this model and its well-posedness see [21], for the optimal control of a resulting macroscopic model see [22]. Pesch *et al.* [5, 30, 34] consider a hypersonic rocket car subject to driving dynamics and to the heat equation with state constraints on the temperature. This represents a simplified model for the re-entry of a spacecraft into atmosphere. In [23] the optimal control of a quarter car model by an electrorheological damper is considered. Here the behavior of the elastic tyre is modelled by the PDE of linear elasticity and the spring-damper element is subject to an ODE. In addition, the latter example involves a complementarity condition modelling the free road contact. In [24] an elastic crane beam subject to linear elasticity coupled to the dynamics of a pendulum, modelling the crane trolley and the applied load, is studied. To the knowledge of the authors, no analytic results have been derived for our particular kind of coupled ODE-PDE problem so far.

Our problem has been stated and solved numerically by a first-discretize-then-optimize (FDTO) approach in [11]. In contrast, in this paper we follow a first-optimize-then-discretize (FOTD) approach. Here we consider a so-called all-at-once approach, i.e. we solve for the states and the control simultaneously. We do not replace the states by the control-to-state operator as for a reduced objective that depends only on the control. First, in Section 3 we prove existence and uniqueness for the coupled state equation, that is not standard, and then apply an abstract result for the existence of optimal controls (Theorem 4.1). In Section 4 we derive analytically the necessary optimality conditions (NOC), including the adjoint differential equations, by a Lagrangian based approach. Furthermore, from the NOC we deduce the existence of the optimal control. Our problem has in common with the general situation in [16, Ch. 1], that a Tikhonov regularization is considered and that we have a tracking-type part of the objective and control box constraints.

In contrast, our problem exhibits a coupled system involving also ODEs and, in addition, terminal conditions in the objective.

We solve the NOC numerically by a semi-smooth Newton method, see Subsection 4.4. For a safety breaking maneuver of the truck-load system we present numerical results in Subsection 5.2. We close with a short discussion in Section 6.

2. Mathematical model

A typical example for Saint-Venant equations coupled to ODEs, corresponding to Newton's law of motion, is a moving truck with a fluid container as load that is not fixed permanently (see Fig. 1). We recall the model derived in [11, Sect. 1]. We consider a finite time interval $[0, T]$ with a terminal time T, that is not considered as a free parameter in this study. For ease of presentation the truck may move in one dimension only. The truck and the container are considered in a fixed coordinate system $(X_1, X_2) \in \mathbb{R}^2$. The horizontal position of the truck is represented by d_{tr} and the container is located at the x-coordinate d_w, the corresponding velocities are the time derivatives $v_{tr} := \dot{d}_{tr}$ and $v_w := \dot{d}_w$. The container has length L, height H (here defined different as in [11]), and a given (continuously differentiable) bottom profile $B(x)$, $0 \le x \le L$. The container could move in the horizontal direction and its mounting to the truck frame is modelled by a linear spring-damper element with damping coefficient k and spring rate c. We consider a moving coordinate system $(x_1, x_2) \in [0, L] \times [0, H]$ for the container. For keeping notation short, we write $x = x_1$. The height of the fluid in the container is represented by $h(t, x)$ and the fluid velocity in x-direction by $v(t, x)$. As domain for the fluid we introduce $Q := (0, T) \times (0, L)$ with the spatial boundary $\Gamma := (0, T) \times \{0, L\}$. From the geometry of the container, we see directly the natural state constraints

$$B(x) \le h(t, x) + B(x) \le H \quad \forall (t, x) \in \overline{Q}.$$

In addition, we require for our model

$$0 < \underline{h} \le h(t, x) \le \overline{h} < H - B(x) \quad \forall (t, x) \in \overline{Q}, \tag{2.1}$$

since we do not want to deal with issues modelling contacts (when the bottom runs dry or the fluid spills over) in this study. The fluid is assumed to be an incompressible Newtonian fluid (like water). The mass of the truck is denoted by m_{tr} and the mass of the container by m_w. The whole system may be controlled by the acceleration $u(t)$ of the truck.

We have for the force that acts between the truck and the container

$$F(d_{tr}, d_w, \dot{d}_{tr}, \dot{d}_w) := c(d_{tr} - d_w + \bar{d}) + k(\dot{d}_{tr} - \dot{d}_w). \tag{2.2}$$

For the offset $\bar{d} := d_w(0) - d_{tr}(0)$ the fluid container is at rest initially. If we assume $d_w^{(T)} = d_{tr}^{(T)} + \bar{d} = 0$, the container rests for the terminal time, too. The fluid is subject to the one-dimensional Saint-Venant equations

$$h_t + (hv)_x = 0, \qquad (t, x) \in Q, \tag{2.3}$$

$$v_t + \left(\frac{1}{2}v^2 + gh\right)_x = -gB_x - \frac{1}{m_w} F(d_{tr}, d_w, \dot{d}_{tr}, \dot{d}_w), \qquad (t, x) \in Q. \tag{2.4}$$

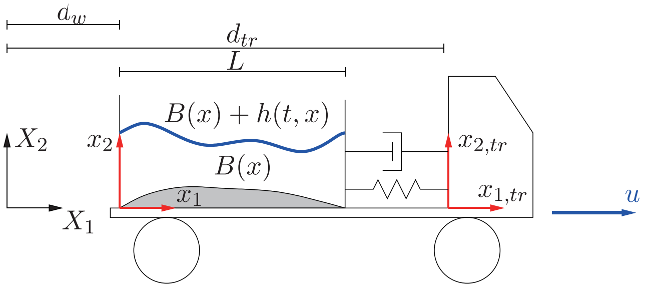

FIGURE 1. Truck with a fluid container as load, illustration of geometric quantities.

where we have exploited $h \neq 0$ in order to derive (2.4) from the conservation of linear momentum (see [11, Sect. 1]) for details). These PDEs are complemented by the initial and boundary conditions

$$h(0,x) = h^{(0)}(x), \qquad x \in [0,L], \tag{2.5}$$

$$v(0,x) = v^{(0)}(x), \qquad x \in [0,L], \tag{2.6}$$

$$h_x = -B_x - \frac{1}{gm_w} F(d_{tr}, d_w, \dot{d}_{tr}, \dot{d}_w), \qquad (t,x) \in \Gamma, \tag{2.7}$$

$$v = 0, \qquad (t,x) \in \Gamma, \tag{2.8}$$

where $h^{(0)}$ and $v^{(0)}$ are given functions. The Neumann boundary condition (2.7) follows from (2.4) and (2.8), see [11, Sect. 1] about the details.

The truck and the container observe Newton's law of motion yielding

$$m_{tr}\ddot{d}_{tr} = u - F(d_{tr}, d_w, \dot{d}_{tr}, \dot{d}_w), \qquad t \in [0,T], \tag{2.9}$$

$$m_w \ddot{d}_w = -\frac{m_w g}{L} \left[h(t,\cdot) + B\right]_0^L \qquad t \in [0,T]. \tag{2.10}$$

We set as initial conditions

$$d_{tr}(0) = d_{tr}^{(0)}, \qquad \dot{d}_{tr}(0) = v_{tr}^{(0)}, \qquad d_w(0) = d_w^{(0)}, \qquad \dot{d}_w(0) = v_w^{(0)}, \tag{2.11}$$

where $d_{tr}^{(0)}$, $v_{tr}^{(0)}$, $d_w^{(0)}$, and $v_w^{(0)}$ are given numbers.

2.1. **Optimal control problem.** Motivated by the coupling force (2.2), we simplify the equation system by replacing d_{tr} by the horizontal distance between truck and container position

$$d_\Delta := d_{tr} - d_w + \bar{d}.$$

We consider the velocities $v_\Delta = \dot{d}_\Delta$ and $v_w = \dot{d}_w$ as independent variables, such that only first order time-derivatives remain in our problem. The PDE states are (h, v) and the ODE states $(d_\Delta, d_w, v_\Delta, v_w)$. We refer to all state variables by $y = (h, v, d_\Delta, d_w, v_\Delta, v_w)$. Consistently, we abbreviate for the initial conditions $y^{(0)} := (h^{(0)}, v^{(0)}, d_\Delta^{(0)}, d_w^{(0)}, v_\Delta^{(0)}, v_w^{(0)})$. Moreover, we write $[\xi]_T := \xi(T) - \xi^{(T)}$ for the difference of a function ξ at terminal time T to a given terminal value $\xi^{(T)}$.

For suitable weights α_1, $\alpha_3 \geq 0$, α_2, α_4, $\alpha_5 > 0$ and σ_4, $\sigma_5 \in \mathbb{R}$, we would like to minimize the objective function

$$\begin{aligned}J(y,u) := \frac{\alpha_1}{2}\int_Q |h(t,x) - h_d(x)|^2\,dx\,dt + \frac{\alpha_2}{2}\int_0^T u(t)^2\,dt \\ + \frac{\alpha_3}{2}\int_Q v(t,x)^2\,dx\,dt + \sum_{I\in\{\Delta;w\}}\left(\frac{\alpha_4}{2}[d_I]_T^2 + \sigma_4[d_I]_T\right) \\ + \sum_{I\in\{\Delta;w\}}\left(\frac{\alpha_5}{2}[v_I]_T^2 + \sigma_5[v_I]_T\right),\end{aligned} \tag{2.12}$$

modelling a tracking type term for a given fluid level h_d, the control effort, the kinetic energy of the fluid, and penalties for achieving given terminal positions $d_\Delta^{(T)}$, $d_w^{(T)}$ and velocities $v_\Delta^{(T)}$, $v_w^{(T)}$ of the truck-container distance and the container, respectively. The control u, being the acceleration of the truck, is subject to the control constraints

$$u_{min} \leq u \leq u_{max}, \tag{2.13}$$

representing that an infinite acceleration is technically not realizable. Note that a negative acceleration corresponds to braking. For the height of the fluid level, we have to require the state constraints (2.1) in principle. However, it turns out in our numerics that they never get active, so we may ignore the state constraints here. For a numerical study incorporating these state constraints into a FDTO approach and the resulting minor effects, see [11, Sect. 4].

The optimal control problem (OCP) reads

$$\min_{Y\times U} J(y,u), \tag{2.14}$$

subject to the system (2.3) - (2.11) and the control constraints (2.13). We work with the state spaces

$$Y_1 := L^2(0,T;H^1(0,L)) \times L^2(0,T;H_0^1(0,L)) \tag{2.15}$$

for the PDE states and

$$Y = Y_1 \times [H^2(0,T)]^2 \times [H^1(0,T)]^2. \tag{2.16}$$

for all states. Since we can prove further regularity of the states (see Section 3), we introduce a space $\tilde{Y}$ for all states in (3.1). The control space is $U = L^2(0,T)$.

Our OCP exhibits an indirect Neumann boundary control for the PDE (2.3) and an indirect distributed control for the PDE (2.4), where the control is acting by means of the force F. This coupling force F is determined by an ODE system, that is controlled directly in (2.9). A back coupling takes place via the Dirichlet boundary values of h entering into the ODE (2.10).

2.2. **Artificial viscosity and rescaled problem.** We regularize the hyperbolic equations (2.3) and (2.4) by introducing an artificial viscosity ε, $0 < \varepsilon \ll 1$. The motivation for this is that hyperbolic conservation laws exhibit non-unique solutions. The physically correct solution (satisfying the entropy principle) is selected [25, Example 2.2.6/Th. 3.3.28] by considering the regularized system, that is semilinear parabolic and has a unique solution, in the limit of vanishing viscosity ε.

Furthermore, by this regularization we avoid to deal with shocks and rare-faction waves, that are typically encountered for these hyperbolic conservation laws [7, Subsection 11.3.2]. It turns out that for a given ε, T has to be chosen sufficiently small for our analytic existence and uniqueness results (Th. 3.2), locally in time, to hold. Thus there is a trade-off between a good approximation of the original Saint-Venant equations, i.e. $\varepsilon \to 0$, and the validity of the well-posedness of our model in time. The regularized solutions for fixed ε are again denoted by h and v.

In order to formulate our coupled system in a simpler form, in particular more accessible for numerics, we perform some scalings. We introduce a new time $\tilde{t} \in (0,1)$ by $t = T\tilde{t}$. In this study, the dedimensionalization of time simplifies the existence and uniqueness proof for the states and, moreover, turns out to have numerical advantages. Coherently, we write $\tilde{Q} := (0,1) \times (0,L)$ and $\tilde{\Gamma} := (0,1) \times \{0,L\}$. We introduce the mass ratio $\eta = m_{tr}/m_w$ and, in addition, scale $\tilde{u} := u/m_{tr}$, $\tilde{c} := c/m_{tr}$, $\tilde{k} := k/m_{tr}$. Finally, we write for the scaled counter-force on the spring-damper system

$$F_s(d_\Delta, v_\Delta) := -\frac{1}{g}(\tilde{c}d_\Delta + \tilde{k}v_\Delta) \tag{2.17}$$

and for the force on the container due to gravity of the fluid

$$F_f(h) := -\frac{g}{L}\left[h(t,\cdot) + B(\cdot)\right]_0^L . \tag{2.18}$$

For ease of presentation we abbreviate

$$F_c(d_\Delta, v_\Delta) = -B_x + \eta F_s(d_\Delta, v_\Delta). \tag{2.19}$$

This leads to the following initial-boundary value problem, following from (2.3) – (2.11),

$$h_t + T(hv)_x - \varepsilon T h_{xx} = 0, \qquad (t,x) \in \tilde{Q}, \tag{2.20}$$

$$h_x = F_c(d_\Delta, v_\Delta), \qquad (t,x) \in \tilde{\Gamma}, \tag{2.21}$$

$$h(0,x) = h^{(0)}(x), \qquad x \in [0,L], \tag{2.22}$$

$$v_t + T\left(\frac{1}{2}v^2 + gh\right)_x - \varepsilon T v_{xx} = TgF_c(d_\Delta, v_\Delta), \qquad (t,x) \in \tilde{Q}, \tag{2.23}$$

$$v = 0, \qquad (t,x) \in \tilde{\Gamma}, \tag{2.24}$$

$$v(0,x) = v^{(0)}(x), \qquad x \in [0,L], \tag{2.25}$$

$$\dot{d}_\Delta = Tv_\Delta, \qquad t \in [0,1], \tag{2.26}$$

$$d_\Delta(0) = d_\Delta^{(0)} := 0, \tag{2.27}$$

$$\dot{d}_w = Tv_w, \qquad t \in [0,1], \tag{2.28}$$

$$d_w(0) = d_w^{(0)}, \tag{2.29}$$

$$\dot{v}_\Delta = T\tilde{u} + TgF_s(d_\Delta, v_\Delta) + TF_f(h), \qquad t \in [0,1], \tag{2.30}$$

$$v_\Delta(0) = v_\Delta^{(0)} := v_{tr}^{(0)} - v_w^{(0)}, \tag{2.31}$$

$$\dot{v}_w = TF_f(h) \qquad t \in [0,1], \tag{2.32}$$

$$v_w(0) = v_w^{(0)}. \tag{2.33}$$

Note that on the right-hand-side of (2.30), the (time-scaled) effective acceleration, i.e. control plus acceleration due to fluid motion in the container, appears (see [11, Sect. 1] for details). For ease of notation, we drop the tildes on $\tilde{t}$, $\tilde{Q}$, $\tilde{\Gamma}$, $\tilde{u}$, $\tilde{c}$, and $\tilde{k}$ in the following.

3. Existence and uniqueness of states

(2.20) and (2.23) are semi-linear parabolic equations that are fully coupled to the ODEs (2.26), (2.28), (2.30), and (2.32). We start by considering existence and uniqueness for this coupled problem that is non-standard assuming a given control $u \in U = L^2(0,1)$ and a given terminal time $T > 0$.

An ODE can be identified as a special case of a PDE (elliptic 1st order), see for instance [19, Subsect. 2.6]. We consider the ODE-PDE system as PDE system. As state spaces we work here with

$$\begin{aligned}\tilde{Y}_1 :=&[H^1(0,1;L^2(0,L))]^2 \cap [L^2(0,1;H^2(0,L))]^2 \\ &\cap [L^\infty(0,1;H^1(0,L)) \times L^\infty(0,1;H_0^1(0,L))] \subset Y_1\end{aligned}$$

for the PDE states and

$$\tilde{Y} = \tilde{Y}_1 \times [H^2(0,1)]^2 \times [H^1(0,1)]^2 \subset Y \tag{3.1}$$

for all states. Both the state spaces Y_1 and Y (here to be understood with $T \equiv 1$) and the control space U are separable Hilbert spaces, while $\tilde{Y}_1$ and $\tilde{Y}$ are not. Note that we have the structure of a Gelfand triple $Y \subset U = U^* \subset Y^*$ that will be exploited for the derivation of optimality conditions below. For Bochner spaces we use established abbreviations like $L^2H^1 :=L^2(0,1;H^1(0,L))$ or $L^2L^2 := L^2(Q)$ in the following.

For $\varepsilon > 0$ the regularized system for h and v is semi-linear parabolic and mathematically well-posed in $\tilde{Y}_1$ for right-hand sides in L^2 and initial data in H_0^1, see [7, Subsect. 7.1.3] for a proof in case of homogeneous Dirichlet boundary conditions. As a preliminary we need

Lemma 3.1 (Estimate for the trace with explicit constant for 1d time-space intervals)**.** *For a function $h \in L^2(0,1;H^1(0,L))$ there holds*

$$\|h\|^2_{L^2(\Gamma)} \le \max\{4/L; L/2\}\, \|h\|^2_{L^2H^1}.$$

The proof is straightforward and relies on Hölder's inequality, the fundamental theorem of calculus and the Cauchy-Schwarz inequality. For details see, e.g., the proof of the similar result [20, Th. II.1 b)].

Now we may prove

Theorem 3.2 (Local existence and uniqueness of the coupled ODE-PDE system for a given control)**.** *For $u \in L^2(0,1)$, $h^{(0)} \in H^1(0,L)$, $v^{(0)} \in H_0^1(0,L)$, $B \in C^1(0,L)$, $\varepsilon \in \mathbb{R}^+$, and sufficiently small times $T > 0$, there exists a unique solution $(h, v, d_\Delta, d_w, v_\Delta, v_w)$ of the coupled system (2.20) – (2.33) s.t. $(h,v)^\top \in \tilde{Y}_1$, $d_\Delta, d_w \in H^2(0,1)$, and $v_\Delta, v_w \in H^1(0,1)$.*

The idea of proof relies on the Banach fixed point theorem and an estimate yielding a factor proportional to $\sqrt{T}$ in the contraction constant. This method has been used e.g. by Niethammer [29] in case of an ODE, that results from a free boundary, coupled to the Laplace PDE. A similar proof is given in [21] for a coupled ODE-PDE problem involving a free boundary, a quasi-linear diffusion PDE and linear elasticity.

Proof. We are going to apply the Banach fixed point theorem in the space

$$\mathcal{M} = \{(h,v)^\top \in Y_1, d_\Delta, d_w \in H^2(0,1), v_\Delta, v_w \in H^1(0,1) \mid \|v\|_{L^\infty L^\infty} \le \kappa\},$$

where κ is a fixed arbitrary positive number. At first we determine *a priori* estimates. In the following we use frequently the Young and Hölder inequalities in order to compensate certain terms from the right-hand sides. We test equation (2.20) by h, using the boundary conditions and Lemma 3.1, and find

$$\begin{aligned}&\sup_{t\in(0,1)} \frac{1}{2}\|h(t)\|^2_{L^2(0,L)} + \frac{\varepsilon}{4}T\|h_x\|^2_{L^2L^2}\\ &\le \frac{1}{2}\|h^{(0)}\|^2_{L^2(0,L)} + T\left(\frac{\kappa^2}{2\varepsilon} + 2\varepsilon C_\Gamma\right)\|h\|^2_{L^2L^2} + TL\|F_c\|^2_{L^2(0,1)}\end{aligned}$$

where $C_\Gamma := \max\{4/L, L/2\}$ is the constant appearing in Lemma 3.1. We multiply by 2 and by Gronwall's inequality this yields

$$\begin{aligned}&\sup_{t\in(0,1)} \|h(t)\|^2_{L^2(0,L)} + \frac{\varepsilon}{2}T\|h_x\|^2_{L^2L^2}\\ &\le \left(\|h^{(0)}\|^2_{L^2(0,L)} + 2TL\|F_c\|^2_{L^2(0,1)}\right)\exp\left(\left(\frac{\kappa^2}{\varepsilon} + 4\varepsilon C_\Gamma\right)T\right).\end{aligned}$$

For fixed ε and $T \le C_h$ with a sufficiently small constant C_h depending on ε, this gives an H^1 estimate on h:

$$\|h\|^2_{L^2H^1} \le \frac{2}{\varepsilon}\|h^{(0)}\|^2_{L^2(0,L)} + 4TL\|F_c\|^2_{L^2(0,1)}.$$

We turn to estimates for v,

$$\begin{aligned}&\sup_{t\in(0,1)} \|v(t)\|^2_{L^2(0,L)} + \varepsilon T\|v_x\|^2_{L^2L^2}\\ &\le \|v^{(0)}\|^2_{L^2(0,L)} + \frac{T}{\varepsilon}g^2\|h\|^2_{L^2L^2} + \|v\|^2_{L^2L^2} + T^2L^2g^2\|F_c\|^2_{L^2(0,1)}.\end{aligned}$$

Again by compensating terms and Gronwall, we have

$$\begin{aligned}&\sup_{t\in(0,1)} \|v(t)\|^2_{L^2(0,L)} + \varepsilon T\|v_x\|^2_{L^2L^2}\\ &\le \left(\|v^{(0)}\|^2_{L^2(0,L)} + \frac{Tg^2}{\varepsilon}\|h\|^2_{L^2L^2} + T^2L^2g^2\|F_c\|^2_{L^2(0,1)}\right)(1 + \exp(1)).\end{aligned}$$

Now we test the equation (2.23) for v with v_t and get

$$\begin{aligned}&\|v_t\|^2_{L^2L^2} + 2\varepsilon T\sup_{t\in(0,1)} \|v_x(t)\|^2_{L^2(0,L)}\\ &\le 2\varepsilon T\|v_x^{(0)}\|^2_{L^2(0,L)} + 4T^2\|v_x\|_{L^2L^2} + 4T^2g^2\|h_x\|^2_{L^2L^2} + 4T^2L^2g^2\|F_c\|^2_{L^2(0,1)}.\end{aligned}$$

By Gronwall

$$\begin{aligned}&\|v_t\|^2_{L^2L^2} + 2\varepsilon T \sup_{t\in(0,1)} \|v_x(t)\|^2_{L^2(0,L)}\\ &\quad\le \left(2\varepsilon T\|v_x^{(0)}\|^2_{L^2(0,L)} + 4T^2g^2\|h_x\|^2_{L^2L^2} + 4T^2L^2g^2\|F_c\|^2_{L^2(0,1)}\right)\exp\left(\frac{2}{\varepsilon}T\right),\end{aligned}$$

Due to the embedding $H^1 \hookrightarrow L^\infty$ in 1d we get with a nonnegative constant C_L from the last estimate

$$\|v\|^2_{L^\infty L^\infty} \le C_L\left(\|v_x^{(0)}\|^2_{L^2(0,L)} + \frac{2}{\varepsilon}Tg^2\left(L^2\|F_c\|^2_{L^2(0,1)} + \|h_x\|^2_{L^2L^2}\right)\right)\exp\left(\frac{2}{\varepsilon}T\right).$$

We will see below that the F_c term is dominated by a factor T. This shows with a $\kappa > \max\{\sqrt{2C_L}\|v_x^{(0)}\|_{L^2(0,L)}, \delta\}$, $0 < \delta < 1$ that we have for sufficiently small $T < \varepsilon\delta^2$ that there is a map from $\mathcal{M}$ into itself. The contraction property of the fixed point map follows analogously. The estimate for the coupling force term reads

$$\|F_c\|^2_{L^2(0,1)} \le C_B + 3\frac{\eta}{g}\left(c\|d_\Delta\|^2_{L^2(0,1)} + k\|v_\Delta\|^2_{L^2(0,1)}\right), \tag{3.2}$$

where C_B is a (nonnegative) constant depending on B_x. For the ODEs we have the estimates

$$\|d_I\|^2_{L^2(0,1)} \le 2T|d_I^{(0)}|^2 + 2T^2\|v_I\|^2_{L^2(0,L)}, \quad I \in \{\Delta, w\},$$

$$\|v_\Delta\|^2_{L^2(0,1)} \le 2T|v_\Delta^{(0)}|^2 + 6T^2\left(\|u\|^2_{L^2(0,1)} + \frac{g}{\eta}\|F_c\|^2_{L^2(0,1)} + \frac{gC_\Gamma}{L}\|h\|^2_{L^2H^1}\right) \tag{3.3}$$

and

$$\|v_w\|^2_{L^2(0,1)} \le 2T|v_w^{(0)}|^2 + 2T^2\frac{gC_\Gamma}{L}\|h\|^2_{L^2H^1}. \tag{3.4}$$

In particular, a factor T^2 enters into (3.3). Thus for sufficiently small T the term with $\|h\|_{L^2H^1}$ entering $\|F_c\|_{L^2(0,1)}$ via (3.2) yields a map into $Y_{1,1}$, being the first component of $\mathcal{M}$. This yields directly $h \in L^\infty L^2 \cap L^2H^1$, thus $v_\Delta, v_w \in H^1(0,1)$ (by the ODEs), $d_\Delta, d_w \in H^2(0,1)$ (again by the ODEs) and finally $v \in L^\infty H^1 \cap H^1L^2$. Thus we have a map into $\mathcal{M}$. From the v-PDE we see that further $v_{xx} \in L^2L^2$, thus $v \in L^2H^2$, too.

So far $h \in L^\infty L^2 \cap L^2H^1$. In order to decide whether $h \in L^\infty H^1 \cap H^1L^2$, we consider another fixed point argument. Assume $\|v_x\|_{L^\infty L^\infty} \le K$. We test (2.20) with $h_t\phi$, where we choose a sufficiently smooth ϕ with compact support in $(0, L)$ s.t. no boundary terms contribute,

$$\begin{aligned}&\|h_t\|^2_{L^2L^2} + \varepsilon T \sup_{t\in(0,1)} \|h_x(t)\|^2_{L^2(0,L)}\\ &\quad\le \varepsilon T\|h_x^{(0)}\|^2_{L^2(0,L)} + K^2T^2\|h\|_{L^2L^2} + \kappa^2T^2\|h_x\|_{L^2L^2}.\end{aligned}$$

Again by a Gronwall argument

$$\begin{aligned}&\|h_t\|^2_{L^2L^2} + \varepsilon T \sup_{t\in(0,1)} \|h_x(t)\|^2_{L^2(0,L)}\\ &\quad\le \left(\varepsilon T\|h_x^{(0)}\|^2_{L^2(0,L)} + K^2T^2\|h\|_{L^2L^2}\right)\exp\left(\frac{\kappa^2}{\varepsilon}T\right)\end{aligned}$$

we get $h \in H^1L^2 \cap L^\infty H^1$ in the interior. Thus from the PDE for h, we see that further $h_{xx} \in L^2L^2$, thus $h \in L^2H^2$, too. This justifies $K < \infty$ by the v-PDE and the Banach fixed point theorem. For the second h-estimate including the boundary, assume at first $F_c = 0$, proceed as above and then add the H^1C^0-function F_s.

By Banach's fixed point theorem the solution is unique. □

We observe that the estimates for h_x, v_x, h_t, and v_t depend critically on ε, even for arbitrary small times T, while the estimates on h and v, do not. Since $T > 0$ has to be sufficiently small we have existence and uniqueness only locally in time. Due to our objective function we expect that we may assume that the control u is determined s.t. there is no blow up for finite times and we can extend the local solution to a solution for any finite time T. Higher regularity results similar as in [7, Subsect. 7.1.3, Th. 6] are not needed in the following and therefore omitted here. Note that in one (time) dimension, we have the embeddings $H^1 \hookrightarrow C^{0,1/2}$ and $H^2 \hookrightarrow C^{1,1/2}$.

4. Necessary optimality conditions and existence of an optimal control

The control effort serves as well as a Tikhonov regularization, therefore we require $\alpha_2 > 0$. We emphasize that for our proof of existence of optimal controls, see Th. 4.1, it is crucial to treat the terminal conditions for the positions d_Δ, d_w and the velocities v_Δ, v_w as penalties.

We minimize (2.14), subject to the regularized ODE-PDE system (2.20) – (2.33) with all boundary and initial conditions and the point-wise control constraints, following from (2.13),

$$u \in U_{ad} := \{u \in L^2(0,1) \mid u_{min} \le u \le u_{max}\}.$$

The adjoints that are introduced in the Subsection 4.1 live in the space W, that turns out in our example that it can be identified with Y, furthermore let

$$\begin{aligned}\Xi &= Y^* \times H^1(0,L) \times H_0^1(0,L) \times \mathbb{R}^4 \\ &= L^2(H^1)^* \times L^2H^{-1} \times [(H^2)^*]^2 \times [(H^1)^*]^2 \times H^1(0,L) \times H_0^1(0,L) \times \mathbb{R}^4,\end{aligned}$$

the latter six factor spaces representing initial conditions. Then the weak formulation of the differential equations yields a bounded operator

$$e : (y,u) \in Y \times U \mapsto \begin{pmatrix} E(y,u) \\ y(0) - y^{(0)} \end{pmatrix} \in \Xi \tag{4.1}$$

where the operator representing the PDE-ODE system is of the following form

$$E(y,u) := \begin{pmatrix} h_t + A_1(y) \\ v_t + A_2(y) \\ \dot{d}_\Delta - v_\Delta \\ \dot{d}_w - v_w \\ \dot{v}_\Delta + A_5 y + B_5 u \\ \dot{v}_w + A_6 y \end{pmatrix} \in W^*$$

with nonlinear operators A_l, $l = 1, 2$ and linear operators B_5, A_l, $l = 5, 6$.

Theorem 4.1 (Existence of optimal controls)**.** *For sufficiently small $T > 0$, there exists an optimal solution $(\hat{y}, \hat{u})$ of our optimal control problem.*

Proof. U_{ad} is a convex, bounded, and closed subset of U. Obviously, the embedding $H^1(0, L) \to L^p(0, L)$ is compact for any $1 < p < \infty$. Thus, weak convergence in Y_1 implies strong convergence in $[L^2L^4]^2 \times [H^2(0, 1)]^2 \times [H^1(0, 1)]^2$, thus h^2, v^2 have a spatial L^2 regularity and the nonlinear terms $h_x v$, hv_x and vv_x multiplied with a test function may be bounded in L^2L^2. Therefore, with $[L^2(Q)]^* = L^2(Q)$ and $[L^2(Q)]^2 \subset Y_1^*$, the map $E : Y \times U \to W^*$ is continuous under weak convergence. Moreover, the state equation $E(y, u) = 0$ has a bounded control-to-state operator $S : u \in U_{ad} \mapsto y = S(u) \in Y$, see our local existence and uniqueness result Th. 3.2 for the states. The considered objective J is sequentially weakly lower semi-continuous. Now we may apply [16, Th. 1.45]. □

However, in order to compute optimal controls it is favorable to solve numerically necessary optimality conditions, see Subsection 4.3.

4.1. **Lagrangian based approach.** We start with formal Lagrange techniques as in [32, Kap. 3.1]. It would also be possible to follow a Hamiltonian approach in order to derive NOCs by Pontryagin's minimum principle, see e.g. [31] for the optimal control of a nonlinear parabolic equations.

The Lagrange function $\mathcal{L} : Y \times U \times W \to \mathbb{R}$ is defined as the objective J coupled to the weak formulation of the PDE-ODE constraints by Lagrange multipliers λ,

$$\mathcal{L}(y, u, \lambda) := J(y, u) + \langle \lambda, E(y, u)\rangle_{W,W^*}\,. \tag{4.2}$$

Here the multipliers are functions $\lambda = (\lambda_1, \lambda_2, \lambda_3, \lambda_4, \lambda_5, \lambda_6)^\top$ and are the so-called adjoints. We insert the modified version (reflecting the scaling in Subsection 2.2) of the objective (2.12) and the weak formulation of the differential equations into (4.2)

$$\begin{aligned}\mathcal{L}(y, u, \lambda) = {} & \frac{\alpha_1 T}{2}\int_Q |h - h_d|^2\,dx\,dt + \frac{\alpha_2 T}{2}\int_0^1 u^2\,dt + \frac{\alpha_3 T}{2}\int_Q v^2\,dt \\ & + \sum_{I\in\{\Delta,w\}} \Big(\frac{\alpha_4}{2}[d_I]_1^2 + \sigma_4[d_I]_1 + \frac{\alpha_5}{2}[v_I]_1^2 + \sigma_5[v_I]_1\Big) + \varepsilon T\int_0^1 F_c[\lambda_1]_0^L\,dt \\ & - \int_Q h_t\lambda_1 - T(hv - \varepsilon h_x)\lambda_{1,x}\,dx\,dt - T\int_0^1 [(gh - \varepsilon v_x)\lambda_2]_0^L\,dt \\ & - \int_Q (v_t - TgF_c)\lambda_2 - T\Big(\frac{v^2}{2} + gh - \varepsilon v_x\Big)\lambda_{2,x}\,dx\,dt \\ & - \int_0^1 (\dot{d}_\Delta - Tv_\Delta)\lambda_3 + (\dot{d}_w - Tv_w)\lambda_4 + (\dot{v}_\Delta - T(u + gF_s + F_f))\lambda_5\,dt \\ & - \int_0^1 (\dot{v}_w - TF_f)\lambda_6\,dt.\end{aligned} \tag{4.3}$$

Let $\hat{y}$, $\hat{u}$, and λ be a solution of the optimal control problem. We expect the necessary optimality conditions

$$\langle \mathcal{L}_y(\hat{y}, \hat{u}, \lambda), y\rangle_{Y^*,Y} = 0 \qquad \forall y \text{ with } y(0) = 0, \tag{4.4}$$

$$\langle \mathcal{L}_u(\hat{y}, \hat{u}, \lambda), u - \hat{u} \rangle_{U^*,U} \geq 0 \qquad \forall u \in U_{ad}. \tag{4.5}$$

They are derived rigorously below, see in Th. 4.3. The derivative of the Lagrange function w.r.t. the states yields after integrations by parts for all time and space derivatives of states (using in particular (2.17) – (2.19))

$$\begin{aligned}
\langle \mathcal{L}_y(\hat{y}, \hat{u}, \lambda), y \rangle_{Y^*,Y} &= \int_Q h \left(\alpha_1 T(\hat{h} - h_d) + \lambda_{1,t} + T\hat{v}\lambda_{1,x} + \varepsilon T \lambda_{1,xx} \right) dx\, dt \\
&- \int_0^L (h\lambda_1)(1,\cdot)\, dx - T \int_0^1 \varepsilon [h\lambda_{1,x} + gh\lambda_2]_0^L\, dt + Tg \int_Q h\lambda_{2,x}\, dx\, dt \\
&- T\frac{g}{L} \int_0^1 [h]_0^L (\lambda_5 + \lambda_6)\, dt + \int_Q v \left(\alpha_3 T\hat{v} + \lambda_{2,t} + T\hat{v}\lambda_{2,x} + \varepsilon T \lambda_{2,xx} \right) dx\, dt \\
&- \int_0^L (v\lambda_2)(1,\cdot)\, dx + \varepsilon T \int_0^1 [v_x \lambda_2]_0^L\, dt + T \int_Q \hat{h} v \lambda_{1,x}\, dx\, dt + \int_0^1 d_\Delta \dot{\lambda}_3\, dt \\
&+ (\alpha_4 [\hat{d}_\Delta]_1 + \sigma_4 - \lambda_3(1)) d_\Delta(1) - Tc \int_0^1 d_\Delta \left(\eta \left(\frac{\varepsilon}{g} [\lambda_1]_0^L + \int_0^L \lambda_2\, dx \right) + \lambda_5 \right) dt \\
&+ (\alpha_4 [\hat{d}_w]_1 + \sigma_4 - \lambda_4(1)) d_w(1) + \int_0^1 d_w \dot{\lambda}_4\, dt + (\alpha_5 [\hat{v}_\Delta]_1 + \sigma_5 - \lambda_5(1))\, v_\Delta(1) \\
&+ \int_0^1 v_\Delta \dot{\lambda}_5\, dt - T \int_0^1 v_\Delta \left(k \left(\eta \left(\frac{\varepsilon}{g} [\lambda_1]_0^L + \int_0^L \lambda_2 \right) + \lambda_5 \right) - \lambda_3 \right) dt \\
&+ (\alpha_5 [\hat{v}_w]_1 + \sigma_5 - \lambda_6(1))\, v_w(1) + \int_0^1 v_w \dot{\lambda}_6\, dt + T \int_0^1 v_w \lambda_4\, dt.
\end{aligned}$$

From (4.5) we obtain

$$\langle \mathcal{L}_u(\hat{y}, \hat{u}, \lambda), u - \hat{u} \rangle_{U^*,U} = T \int_0^1 (\alpha_2 \hat{u} + \lambda_5)(u - \hat{u})\, dt \geq 0 \quad \forall u \in U_{ad}.$$

Here we have the structure $\mathcal{L}_u(y, u, \lambda) = \tilde{\mu} u + G(y, u, \lambda)$ with $\tilde{\mu} = \alpha_2 > 0$ and $G(y, u, \lambda) = \lambda_5$ continuously Fréchet-differentiable from $Y \times L^2(0,1) \times W \to L^2(0,1)$ and locally Lipschitz-continuous from $Y \times L^2(0,1) \times W \to L^p(0,1)$, $p > 2$, as required in [10, Assumpt. 4.2 (b)]. $U = L^2(0,1)$ is a Hilbert space and $U_{ad} \subset U$ is nonempty, closed, and convex. By means of a superposition operator $\Pi : Y \times U \times W \to U$,

$$\Pi(y, u, \lambda)(t, x) := u(t) - P_{U_{ad}}(u(t) - \alpha_2^{-1} \mathcal{L}_u(y(t,x), u(t), \lambda(t,x))), \tag{4.6}$$

where $P_{U_{ad}}$ is the Euclidean projection onto U_{ad}, we may rewrite [16, Corollary 1.2]) the variational inequality (4.5) as

$$\hat{u} = P_{U_{ad}} \left(-\frac{1}{\alpha_2} \lambda_5 \right). \tag{4.7}$$

As a check, by

$$\langle \mathcal{L}_\lambda(\hat{y}, \hat{u}, \lambda), \lambda \rangle_{W^*,W} = 0 \tag{4.8}$$

we recover the differential equations and its Neumann boundary conditions. Here we apply the fundamental lemma of calculus of variations (also known as Du Bois-Reymond lemma) in order to deduce that the integrands themselves are zero.

4.2. **Adjoint differential equations.** The so far only formally derived NOC (4.4) yield the adjoint system for λ. We would have the adjoint PDE

$$\lambda_{1,t} + T\hat{v}\lambda_{1,x} + \varepsilon T\lambda_{1,xx} + Tg\lambda_{2,x} = -\alpha_1 T(\hat{h} - h_d), \qquad (t,x) \in Q, \tag{4.9}$$

$$\lambda_{1,x} = -\frac{g}{\varepsilon L}(\lambda_5 + \lambda_6), \qquad (t,x) \in \Gamma, \tag{4.10}$$

$$\lambda_1(1,x) = 0, \qquad x \in [0,L], \tag{4.11}$$

$$\lambda_{2,t} + T\hat{v}\lambda_{2,x} + \varepsilon T\lambda_{2,xx} + T\hat{h}\lambda_{1,x} = -T\alpha_3\hat{v}, \qquad (t,x) \in Q, \tag{4.12}$$

$$\lambda_2 = 0, \qquad (t,x) \in \Gamma, \tag{4.13}$$

$$\lambda_2(1,x) = 0, \qquad x \in [0,L], \tag{4.14}$$

and the adjoint ODE

$$\dot{\lambda}_3 = Tc\left(\eta\left(\frac{\varepsilon}{g}[\lambda_1]_0^L + \int_0^L \lambda_2\,dx\right) + \lambda_5\right), \qquad t \in (0,1), \tag{4.15}$$

$$\lambda_3(1) = \alpha_4\left(\hat{d}_\Delta(1) - d_\Delta^{(T)}\right) + \sigma_4, \tag{4.16}$$

$$\lambda_4 = \alpha_4\left(\hat{d}_w(1) - d_w^{(T)}\right) + \sigma_4, \qquad t \in (0,1], \tag{4.17}$$

$$\dot{\lambda}_5 = -T\lambda_3 + Tk\left(\eta\left(\frac{\varepsilon}{g}[\lambda_1]_0^L + \int_0^L \lambda_2\,dx\right) + \lambda_5\right), \qquad t \in (0,1), \tag{4.18}$$

$$\lambda_5(1) = \alpha_5\left(\hat{v}_\Delta(1) - v_\Delta^{(T)}\right) + \sigma_5, \tag{4.19}$$

$$\lambda_6 = T\lambda_4(1-t) + \alpha_5\left(\hat{v}_w(1) - v_w^{(T)}\right) + \sigma_5, \qquad t \in (0,1], \tag{4.20}$$

where we have exploited that $\dot{\lambda}_4 = 0$ and that we may integrate $\dot{\lambda}_6$ in time. We abbreviate the adjoint terminal conditions by

$$\lambda^{(1)} := -(0, 0, \alpha_4[\hat{d}_\Delta]_1 + \sigma_4, \alpha_4[\hat{d}_w]_1 + \sigma_4, \alpha_5[\hat{v}_\Delta]_1 + \sigma_5, \alpha_5[\hat{v}_w]_1 + \sigma_5)^\top.$$

Remark 4.2 (Coupling structure)**.** In the ODE-PDE system we have a control that acts on the ODE system (state v_Δ). All ODE states enter the PDE or the boundary conditions for the states h and v, that are fully coupled. Finally, h enters into the ODE states v_Δ and v_w.

In the adjoint system the adjoints λ_5 and λ_6 (corresponding to v_Δ and v_w) enter via the Neumann boundary condition for λ_1 (corresponding to h), fully coupled with λ_2 (corresponding to v) and all PDE adjoints (λ_1 and λ_2) enter into the ODE for the adjoints λ_3 and λ_5. The adjoint λ_5 (corresponding to v_Δ) determines the control. We notice a reversed coupling in the adjoint system.

4.3. **Derivation of necessary optimality conditions.** In order to prove the necessary optimality conditions we combine the concept and notation of [32, Kap. 5.5] and [16, Sect. 1.7], who have demonstrated this for the Neumann boundary control as well as for the distributed control in case of a single parabolic PDE.

Theorem 4.3 (First-order necessary optimality conditions)**.** *Let $T > 0$ be a sufficiently small time. Then for our optimal control problem (2.14) subject to (2.20) – (2.33) and (2.13), the necessary optimality conditions (4.4), (4.7) and (4.8) hold,* *i.e. an optimal solution $(\hat{y}, \hat{u})$ fulfills: i) the adjoint system (4.9) – (4.20), the ii) optimality condition (4.7), and iii) the PDE-ODE system in the weak form (4.8).*

Proof. We consider box constraints and U_{ad} is a closed, convex, nonempty subset of an open Banach space U. Furthermore, according to Th. 3.2 there exists a control-to-state operator $S : U \to Y, u \mapsto S(u)$ such that the reduced objective $\mathcal{J}(u) = J(S(u), u)$ is well-defined on an open neighborhood V of U_{ad} and Gâteaux differentiable around $\hat{u}$. Thus we may apply [16, Th. 1.46] yielding (4.7).

We consider the operator for the state equation $E(y, u) = 0$, $E : Y \times U \to W^*$ in (4.1). Since we consider control constraints, [16, Th. 1.48, Corollary 1.3] shows that it suffices to require only

1) Continuous F-differentiability of $J : Y \times U \to \mathbb{R}$ and $E : Y \times U \to W^*$,
2) Unique solvability of the state equation in $V \subset U$, and
3) $E_y(y(u), u) \in \mathcal{L}(Y, W^*)$ has a bounded inverse for all $u \in V \supset U_{ad}$.

We check:

1) The statement follows from the imbedding $Y \hookrightarrow [C^0([0,1]; L^2(0,L))]^2 \times [C^1([0,1])]^2 \times [C^0([0,1])]^2$. The terminal conditions on d_Δ, d_w, v_Δ, and v_w are well-defined, too.
2) Let $S : V \to Y$ denote the control-to-state operator (solution operator) of the differential equation system. This control-to-state operator S is well-defined, since for every u we have a solution, see Th. 3.2.
3) The linearized problem in strong form is obtained by linearizing (2.20) – (2.33)

$$\begin{aligned}
\tilde{h}_t + T(\hat{h}\tilde{v} + \tilde{h}\hat{v})_x - \varepsilon T\tilde{h}_{xx} &= 0, && (t,x) \in Q,\\
\varepsilon T\tilde{h}_x &= \varepsilon T\eta F_s(\tilde{d}_\Delta, \tilde{v}_\Delta), && (t,x) \in \Gamma,\\
\tilde{h}(0,x) &= 0, && x \in [0,L],\\
\tilde{v}_t + (T\hat{v}\tilde{v} + g\tilde{h})_x - \varepsilon T\tilde{v}_{xx} &= Tg\eta F_s(\tilde{d}_\Delta, \tilde{v}_\Delta), && (t,x) \in Q,\\
\tilde{v} &= 0 && (t,x) \in \Gamma,\\
\tilde{v}(0,x) &= 0, && x \in [0,L]
\end{aligned}$$

and

$$\begin{aligned}
\dot{\tilde{d}}_\Delta &= T\tilde{v}_\Delta, && t \in (0,1), && \tilde{d}_\Delta(0) = 0,\\
\dot{\tilde{d}}_w &= T\tilde{v}_w, && t \in (0,1), && \tilde{d}_w(0) = 0,\\
\dot{\tilde{v}}_\Delta - TgF_s(\tilde{d}_\Delta, \tilde{v}_\Delta) &= Tu - T\frac{g}{L}[\tilde{h}(t,\cdot)]_0^L, && t \in (0,1), && \tilde{v}_\Delta(0) = 0,\\
\dot{\tilde{v}}_w &= -T\frac{g}{L}[\tilde{h}(t,\cdot)]_0^L, && t \in (0,1), && \tilde{v}_w(0) = 0.
\end{aligned}$$

Here we have scaled the Neumann boundary condition for h by a factor of εT corresponding to the conormal derivative. Note that for well-balanced equations in our numerics, we scale the Neumann boundary condition for λ_1 analogously by εT. $\langle E_y, \tilde{y}\rangle_{Y^*,Y}$ is the linearized problem in weak form, that follows from integrating by parts.

For the linearized problem we have the same structure of estimates as for the full problem. The linearized problem has a unique solution for every u and every initial data $h^{(0)} \in H^1(0,L)$, $v^{(0)} \in H_0^1(0,L)$, and $(d_\Delta^{(0)}, d_w^{(0)}, v_\Delta^{(0)}, v_w^{(0)})^\top \in \mathbb{R}^4$. Thus the linearized control-to-state operator $\tilde{S}(\hat{u})$ is a well-defined unique linear continuous (bounded) affine operator and, in particular, is surjective (not a proper dense subset of the image). It has a bounded inverse, according to the theorem of the inverse mapping [1, Satz 5.8] (for a suitable neighborhood $V \supset U_{ad}$).

□

The adjoint system has the same structure as the linearized original ODE-PDE problem and possesses hence a unique solution. Note that the adjoint equations are solved backwards in time. Consistently, we have terminal conditions in the parabolic PDEs for λ_1 and λ_2 and the Laplacian operator has the opposite sign in the adjoint PDE (compared to the state PDE).

Using the same ingredients (Hölder/Young inequalities, trace theorem, Gronwall inequality, compensation for sufficiently small T) as for state equations, we get the following regularity from estimates for the adjoint PDEs:

$$(\lambda_1, \lambda_2)^\top \in \tilde{Y}_1 \subset Y_1, \quad \lambda \in \tilde{Y} \subset Y \simeq W, \tag{4.21}$$

indeed, showing the higher regularity for λ.

Remark 4.4 (Convexity and direct approach)**.** For a convex problem the necessary conditions would be also sufficient. Note that due to the Saint-Venant equations the problem is in general not convex.

For a semi-linear parabolic equation with an objective convex w.r.t. u, the necessary optimality conditions are proven in [32, Satz 5.8/Satz 5.15] under reasonable assumptions. The proofs use, among other things, the convexity and closedness of U_{ad}, the lower semi-continuity of the objective, a Hilbert space structure, and the solvability of the adjoint equation. The necessary optimality conditions are derived directly.

We could also follow a direct approach with a reduced objective as in [16].

4.4. **Semi-smooth Newton method in a Hilbert space.** We introduce $\bar{Z} = Y \times U \times W$ and aggregate all unknowns in

$$z = (y, u, \lambda)^\top = (h, v, d_\Delta, d_w, v_\Delta, v_w, u, \lambda_1, \lambda_2, \lambda_3, \lambda_4, \lambda_5, \lambda_6)^\top \in \bar{Z}.$$

The necessary optimality conditions (4.4), (4.7), and (4.8) yield the following non-smooth system:

$$f(z) = \begin{pmatrix} \mathcal{L}_\lambda(z) \\ \Pi(z) \\ \mathcal{L}_y(z) \end{pmatrix} \overset{!}{=} 0, \tag{4.22}$$

where $\mathcal{L}_\lambda(z)$ represents the PDE-ODE system for the states and $\mathcal{L}_y(z)$ yields the adjoint PDE-ODE system.

In order to continue analogously as in [10], where elliptic problems are considered, we semi-discretize in time. First semi-discretizing in time and then discretizing in space is a standard technique and is e.g. applied by Hinze *et al.* in [15, Section 2] to optimal flow. We divide the time interval $[0, 1]$ by $N+1$ time steps with constant increment $\Delta t = 1/N$. For ease of presentation we keep a similar notation as in the continuous case. Let $Z = H^1(0,L) \times H_0^1(0,L) \times \mathbb{R}^4 \times U \times H^1(0,L) \times H_0^1(0,L) \times \mathbb{R}^4$ and

$$z := (y^0, u^0, \lambda^0, \dots, y^N, u^N, \lambda^N)^\top \in Z^{N+1}.$$

Then the time-discretized PDE-ODE system is of the following form

$$y^i = y^{i-1} + \Delta t\, \Phi(y^i, u^i), \qquad i = 1, \dots, N, \tag{4.23}$$

with initial conditions $y^0 = \phi := (h^{(0)}, v^{(0)}, d_\Delta^{(0)}, d_w^{(0)}, v_\Delta^{(0)}, v_w^{(0)})$. Analogously, the adjoint system semi-discretized in time reads

$$\lambda^i = \lambda^{i+1} + \Delta t\, \Psi(y^i, u^i, \lambda^{i+1}), \qquad i = 0, \dots, N-1, \tag{4.24}$$

with terminal conditions $\lambda^N = \psi := \lambda^{(1)}$. Note that the state equations are solved forward in time by an implicit Euler method, while the adjoint equations are solved backward in time, see e.g. [4, Sect. 3.3]. Note that the coupled forward/adjoint system can be interpreted as a Hamiltonian system [2, Sect. 2.2] and that it is computed here by a symplectic method [9, Th. 5.3.3]. In particular, the discrete adjoint backward update is explicit w.r.t. λ and implicit w.r.t. y.

As a possibility to speed up our numerics, we could try a more elaborated scheme of Crank-Nicolson type. For evolution equations where the spatial differential operator is self-adjoint it has been demonstrated that FDTO and FOTD approaches commute for certain variants of Crank-Nicolson schemes [2], but it is not clear how to apply this framework to our problem. Alternatively, we could start with a semi-discretization in space, too.

From (4.22) this yields the equation for the vector $f \in [Z^*]^{N+1}$

$$f = (f_0, \dots, f_i, \dots, f_N)^\top = 0 \tag{4.25}$$

where

$$f_0 = \begin{pmatrix} -y^0 + \phi \\ u^0 - P_{[u_{min}, u_{max}]}(-\lambda_5^0/\alpha_2) \\ -(\lambda^0 - \lambda^1) + \Delta t\, \Psi(y^0, u^0, \lambda^1) \end{pmatrix},$$

$$f_i = \begin{pmatrix} -(y^i - y^{i-1}) + \Delta t\, \Phi(y^i, u^i) \\ u^i - P_{[u_{min}, u_{max}]}(-\lambda_5^i/\alpha_2) \\ -(\lambda^i - \lambda^{i+1}) + \Delta t\, \Psi(y^i, u^i, \lambda^{i+1}) \end{pmatrix}, \quad i = 1, \dots, N-1,$$

and

$$f_N = \begin{pmatrix} -(y^N - y^{N-1}) + \Delta t\, \Phi(y^N, u^N) \\ u^N - P_{[u_{min}, u_{max}]}(-\lambda_5^N/\alpha_2) \\ \lambda^N - \psi(y^N) \end{pmatrix}.$$

Let $\mathcal{L}^i$ denote the semi-discretized Lagrange function. The assumptions on the structure of $\mathcal{L}_u^i$ [10, Assumpt. 4.2], yielding the semi-smoothness of f_i, and on the uniform invertibility of the Newton matrices M_i [10, Assumpt. 2.3], due to the Lax-Milgram theorem are fulfilled. Furthermore, we assume that the Tikhonov parameter α_2 is sufficiently large [10, Assumpt. 3.1 (a)], such that a descent direction w.r.t. the merit function $\Theta = \|f\|_{[Z^*]^{N+1}}^2$ is always obtained. Thus we may apply the globalized semi-smooth Newton method derived in [10] to compute a zero of f.

Let I_Y denote the identity operator in the Banach space Y. The Newton matrix M is a block-tridiagonal matrix of the form

$$M := \begin{pmatrix} M_1 & R_1 & & & \\ L_2 & M_2 & R_2 & & \\ & \ddots & \ddots & \ddots & \\ & & L_{N-1} & M_{N-1} & R_{N-1} \\ & & & L_N & M_N \end{pmatrix} \tag{4.26}$$

with

$$\begin{aligned} M_i &\in \partial_C f_i, && i = 1, \dots, N, \\ L_i &:= diag(I_Y, 0, 0), && i = 2, \dots, N, \\ R_i &:= diag(0, 0, I_W), && i = 1, \dots, N-1. \end{aligned}$$

The set $\partial_C f_i$ consists of all matrices $M_i \in L(Z, Z^*)$, $i = 0, \dots, N$. M_i has the structure

$$M_i(y,u,\lambda) = \begin{pmatrix} \Delta t\, \Phi_y(y,u) - I_Y & \Delta t\, E_u^i(y,u) & 0 \\ 0 & 1 & (0,0,0,0,D,0) \\ \mathcal{L}_{y,y}^i(y,u,\lambda) & \mathcal{L}_{y,u}^i(y,u,\lambda) & \mathcal{L}_{y,\lambda}^i(y,u,\lambda) - I_W \end{pmatrix}, \tag{4.27}$$

where the generalized differential $D \in L^\infty(0,1)$ is chosen such that

$$D(t) \in \partial_C P_{[u_{min}, u_{max}]}(-\lambda_5(t)/\alpha_2), \quad t \in (0,1). \tag{4.28}$$

The subdifferential $\partial_C P_{[u_{min}, u_{max}]}$ takes its values in $\{\{0\}, [0,1], \{1\}\}$, whereupon at a non-differentiability point we may choose a fixed value in the interval $[0,1]$.

Here we have matrix blocks of the following structure, for the PDE-ODE system (incorporating the boundary conditions for the h-PDE into the first/last components of the corresponding block suitably)

$$\Phi_y(y,u) =$$

$$\left(\begin{array}{ll|c|ccc} & & & Tc\eta\frac{\varepsilon}{g} & Tk\eta\frac{\varepsilon}{g} & \\ \multicolumn{2}{c|}{-T(\hat{v}\partial_x + \hat{v}_x) + \varepsilon T\partial_{xx}} & -T(\hat{h}\partial_x + \hat{h}_x) & & & \\ & & & -Tc\eta\frac{\varepsilon}{g} & -Tk\eta\frac{\varepsilon}{g} & \\ \hline \multicolumn{2}{c|}{-Tg\partial_x} & -T\hat{v}\partial_x + \varepsilon T\partial_{xx} & -Tc\eta & -Tk\eta & \\ \hline & & & & T & \\ & & & & & T \\ \frac{Tg}{L} & \frac{-Tg}{L} & & -Tc & -Tk & \\ \frac{Tg}{L} & \frac{-Tg}{L} & & & & \end{array} \right),$$

and for the transposed adjoint system

$$\mathcal{L}_{y,\lambda}^i(y,u,\lambda) = \Delta t\, \Psi_\lambda(y,u)^\top =$$

$$\Delta t \left(\begin{array}{ll|c|ccc} & & & Tc\eta\frac{\varepsilon}{g} & Tk\eta\frac{\varepsilon}{g} & \\ \multicolumn{2}{c|}{T\hat{v}\partial_x + \varepsilon T\partial_{xx}} & T\hat{h}\partial_x & & & \\ & & & -Tc\eta\frac{\varepsilon}{g} & -Tk\eta\frac{\varepsilon}{g} & \\ \hline \multicolumn{2}{c|}{Tg\partial_x} & T\hat{v}\partial_x + \varepsilon T\partial_{xx} & -Tc\eta & -Tk\eta & \\ \hline & & & & T & \\ & & & & & T \\ \frac{Tg}{L} & \frac{-Tg}{L} & & -Tc & -Tk & \\ \frac{Tg}{L} & \frac{-Tg}{L} & & & & \end{array} \right).$$

Furthermore, $E_u^i = (0,0,0,0,T,0)^\top$, and for the blocks corresponding to the objective we have

$$\mathcal{L}^i_{y,y}(y,u,\lambda) = \Delta t \begin{pmatrix} T\alpha_1 & T\lambda_{1,x} & & & & \\ T\lambda_{1,x} & T\alpha_3 + T\lambda_{2,x} & & & & \\ & & \alpha_4\delta_1 & & & \\ & & & \alpha_4\delta_1 & & \\ & & & & \alpha_5\delta_1 & \\ & & & & & \alpha_5\delta_1 \end{pmatrix}$$

(δ_1 here denoting the Dirac distribution that is 1 for the terminal time $t = 1$ and 0 otherwise) and $\mathcal{L}_{y,u}(y,u,\lambda)$ vanishes. This yields set valued mappings $\partial_C f_i : Z \rightrightarrows L(Z, Z^*)$ where the map has values in the set of all M fulfilling (4.27). Note that the the subscript "C" is due to the close relation in finite dimensions to Qi's C-subdifferential.

The semi-smooth Newton method with a suitable globalization strategy [10, Algorithm 3.3], there applied to semi-linear elliptic equations, is here adapted to the semi-discretization in time of semi-linear parabolic equations.

Algorithm 4.5 (Global semi-smooth Newton method)**.**

(i) Set $k = 0$, define $z_0 := z^{(0)} \in Z^{N+1}$, and choose $\beta \in (0,1)$, $\sigma \in (0, 1/2)$.

(ii) If $\|f\|_{[Z^*]^{N+1}} < tol$, then stop.

(iii) For fixed $M(z_k)$, $M_i(z_k) \in \partial_C f_i(z_k)$ for all $i = 0, \ldots, N$ compute the search direction s_k by solving

$$M(z_k) s_k = -f(z_k)$$

(iv) Determine the smallest $i_k \in \mathbb{N}_0$ such that

$$\Theta(z_k + \beta^{i_k} s_k) \leq (1 - 2\sigma\beta^{i_k})\Theta(z_k)$$

and set $\tilde{\beta}_k := \beta^{i_k}$.

(v) Update $z_{k+1} := z_k + \tilde{\beta}_k s_k$ and $k := k + 1$. Goto (ii).

In Step (iv) the step-size $\tilde{\beta}_k$ is determined by an Armijo line-search, relying in particular on the merit function $\Theta(z_k)$ and that the gradient of $f(z_k)$ applied to s_k is $-2\Theta(z_k)$.

According to [10, Th. 3.4, Th. 3.5] we have

Theorem 4.6 (Accumulation points are global solutions)**.** *For α_2 sufficiently large we have that any accumulation point $\bar{z}$ of a sequence $\{z_k\}_{k\in\mathbb{N}_0}$ (with $f(z_k) \neq 0$ for all $k \in \mathbb{N}_0$) generated by Algorithm 4.5 is a zero. The sequence converges super-linearly to $\bar{z}$ in a suitable neighborhood of $\bar{z}$.*

5. Numerical methods

The discretized optimal control problem can be solved by a gradient-based optimization procedure like SQP (as in [11]), or, the discretized NOC by a semi-smooth Newton method as we do in this study.

5.1. **Fully discretized problem.** We discretize the space interval $[0, L]$ equidistant with $\Delta x = L/M$, yielding $x_j := j\Delta x$, $j = 0, \dots, M$. We introduce

$$z^{\Delta x,i} := (h_1^i, \dots, h_{M-1}^i, v_1^i, \dots, v_{M-1}^i, d_\Delta^i, d_w^i, v_\Delta^i, v_w^i)^\top, \quad i = 0, \dots, N.$$

The boundary values $h_0^i, h_M^i, v_0^i, v_M^i$, $i = 0, \dots, N$, are determined directly and, thus, are not included in the solution vectors $z^{\Delta x,i}$. The vectors in case of the full discretization are indicated by an upper index Δx. For ease of presentation we state the discretized problem for the case $B \equiv 0$.

Note that we discretize the flux terms in the Saint-Venant equations as in the Lax-Friedrichs scheme, e.g.,

$$(h\partial_x v + \partial_x hv)(x_j) = (hv)_x(x_j) \approx \frac{h_{j+1}v_{j+1} - h_{j-1}v_{j-1}}{2\Delta x}.$$

Furthermore, in consistence with the Lax-Friedrichs scheme we have set

$$\varepsilon = \frac{1}{2}\frac{(\Delta x)^2}{T\,\Delta t}$$

for the artificial viscosity. With these two considerations, for instance (2.20) is approximated by

$$\frac{h_j^{i+1} - h_j^i}{\Delta t} \approx -\frac{T}{2\Delta x}\left(h_{j+1}^i v_{j+1}^i - h_{j-1}^i v_{j-1}^i\right) + \frac{1}{2\Delta t}\left(h_{j+1}^i - 2h_j^i + h_{j-1}^i\right).$$

The Lax-Friedrichs scheme is an explicit method that is first order in time and second order in space. For the convergence, and hence the stability, of an explicit scheme the Courant-Friedrichs-Levy (CFL) condition

$$V\frac{T\Delta t}{\Delta x} < 1 \tag{5.1}$$

is necessary and sufficient [27, Sect. 8.3]. Here $V = \max_{t,x}\{v \pm \sqrt{gh}\}$ denotes the so-called group velocity at which information is exchanged within the numerical grid. But V cannot be determined *a priori* and is part of the numerical solution. Thus the CFL number $VT\Delta t/\Delta x$ has to be checked in the numerical results *a posteriori*.

Analogously as in [11] time integrals are approximated by first order Riemann sums and space integrals by the second order trapezoidal rule. This is consistent with the Lax-Friedrichs scheme that is first order in time and second order in space. Here the PDE-ODE (4.23) are fully discretized by

$$y^i = y^{i-1} + \Delta t\, \Phi^{\Delta x,i}, \qquad i = 1, \dots, N,$$

where $\Phi^{\Delta x,i}$ is here the space discretization of $\Phi(y^i, u^i)$ $(i = 1, \dots, N)$, together with the boundary conditions that follow by the method of undetermined coefficients accurately in second order,

$$h_0^i = \frac{4}{3}h_1^i - \frac{1}{3}h_2^i - \frac{2}{3}\Delta x F_{c,0}^i, \qquad v_0^i = 0,$$

$$h_M^i = \frac{4}{3}h_{M-1}^i - \frac{1}{3}h_{M-2}^i + \frac{2}{3}\Delta x F_{c,M}^i, \qquad v_M^i = 0,$$

where $F_{c,j}^i$ is the discretization of (2.19) at time step t_i at the grid point x_j. The initial conditions $y^0 \in \mathbb{R}^{2(M+1)}$ enter by

$$y^0 = \phi(y^{(0)}) := \left(\begin{array}{ccc|ccc|cccc} h_1^0 & \dots & h_{M-1}^0 & v_1^0 & \dots & v_{M-1}^0 & d_{tr}^0 & d_w^0 & v_{tr}^0 & v_w^0 \end{array} \right)^\top .$$

From (4.7) we obtain for the discretized control directly

$$u^i = P_{[u_{min},u_{max}]}(-\lambda_5^i/\alpha_2), \quad i = 0, \dots, N.$$

By discretizing the NOC in time and space, the obtained FOTD-adjoints λ are different to the FDTO adjoints. We write

$$\lambda^i = (\lambda_{1,1}^i, \dots, \lambda_{1,M-1}^i, \lambda_{2,1}^i, \dots, v_{2,M-1}^i, \lambda_3^i, \lambda_4^i, \lambda_5^i, \lambda_6^i)^\top.$$

From (4.24) we find for the fully discretized adjoint equation

$$\lambda^i = \lambda^{i+1} + \Delta t\, \Psi^{\Delta x,i}, \qquad i = 0, \dots, N-1,$$

where $\Psi^{\Delta x,i}$ is here the space discretization of $\Psi(y^i, u^i, \lambda^{i+1})$ with boundary conditions

$$\begin{aligned} \lambda_{1,0}^i &= \frac{4}{3}\lambda_{1,1}^i - \frac{1}{3}\lambda_{1,2}^i + \frac{4}{3}\frac{T\Delta t}{\Delta x}\frac{g}{L}\left(\lambda_5^i + \lambda_6^i\right), & \lambda_{2,0}^i &= 0, \\ \lambda_{1,M}^i &= \frac{4}{3}\lambda_{1,M-1}^i - \frac{1}{3}\lambda_{1,M-2}^i - \frac{4}{3}\frac{T\Delta t}{\Delta x}\frac{g}{L}\left(\lambda_5^i + \lambda_6^i\right), & \lambda_{2,M}^i &= 0, \end{aligned}$$

and terminal conditions $\lambda^N = \psi(y^N)$. Note that in the initial guess z_0 we set $z_{Nm+2M-1} = d_\Delta^{(T)}$ and $z_{Nm+2M} = d_w^{(T)}$. We abbreviate the number of vector components at each time step by $m = 2(M+1) + 1 + 2(M+1) = 4M+5$. Let

$$z^{\Delta x} := (y^0, u^0, \lambda^0, \dots, y^N, u^N, \lambda^N)^\top \in \mathbb{R}^{(N+1)\times m}$$

and $f_i^{\Delta x}(z^{\Delta x}) \in \mathbb{R}^m$, $i = 0, \dots, N$, are vectors for each time step with

$$f_i^{\Delta x} = \begin{pmatrix} -(y^i - y^{i-1}) + \Delta t\, \Phi^{\Delta x,i} \\ \pi(u^i, \lambda_5^i) \\ -(\lambda^i - \lambda^{i+1}) + \Delta t\, \Psi^{\Delta x,i} \end{pmatrix}, \quad i = 1, \dots, N-1,$$

but with the two exceptions

$$\begin{aligned} (f_0^{\Delta x})_{\text{1st line}} &= -y^0 + \phi, \\ (f_N^{\Delta x})_{\text{3rd line}} &= \lambda^N - \psi(y^N). \end{aligned}$$

Then it remains to solve

$$f^{\Delta x} = (f_0^{\Delta x}, \dots, f_i^{\Delta x}, \dots, f_N^{\Delta x})^\top = 0$$

by a semi-smooth Newton method. The Newton matrix $M_i^{\Delta x}$ is the space-discretized version of (4.26), where the entries are the following matrices

$$\begin{aligned} M_i^{\Delta x} &\in \partial_C f_i^{\Delta x} \quad i = 1, \dots, N, \\ L_i^{\Delta x} &:= diag(Id_{2(M+1)}, 0, 0_{2(M+1)\times 2(M+1)}), \quad i = 2, \dots, N, \\ R_i^{\Delta x} &:= diag(0_{2(M+1)\times 2(M+1)}, 0, Id_{2(M+1)}), \quad i = 1, \dots, N-1, \end{aligned}$$

and the subdifferential $\partial_C f_i^{\Delta x}$ consists of all matrices in $\mathbb{R}^{m\times m}$ of the form (following from (4.27))

$$\begin{pmatrix} \Delta t\, \Phi_{y^i}^{\Delta x,i} - Id_{2(M+1)} & (0_{2M}, T, 0)^\top & 0 \\ 0 & 1 & (0_{2M}, D, 0)^\top \\ (\mathcal{L}_{y^i,y^i})^{\Delta x,i} & 0 & \Delta t\, \Psi_{\lambda^i}^{\Delta x,i-1} - Id_{2(M+1)} \end{pmatrix}.$$

with $D \in [L^\infty(0,1)]$, $\quad D(t) \in \partial_C P_{[u_{min},u_{max}]}(-\lambda_5^i/\alpha_2)$, whereupon we have modified Newton matrices in the cases $i = 0$ and $i = N$.

Our Algorithm 4.5 for the time-discretized situation is additionally equipped with a standard expansion strategy in the Armijo line-search. This expanded Armijo line-search is efficient [26, §5, Satz 1] and allows for a significant speed-up in the numerical computation of the optimal control.

5.2. **Numerical results.** The Algorithm 4.5 has been implemented in MATLAB R2015b. The $[Z^*]^{N+1}$ norm entering in the stopping criterion in step (ii) of the algorithm is discretized using again the trapezoidal rule for the spatial integrals. As parameters in this algorithm we work with $\beta = 0.9$, $\sigma = 0.001$, and $tol = 10^{-6}$.

We consider two examples, the first scenario corresponding to an optimal braking maneuver as considered in [11] and a second scenario with different parameters and weights. The following data is underlying both examples. We work with the values $d_w(0) = -5$, $d_\Delta(1) = 0$ (corresponding to $d_{tr}(1) = 100$), $d_w(1) = 95$, $v_{tr}(0) = 10$, $v_w(0) = 10$, and the parameters in Table 1. By definition of the offset $\bar{d}$, we have $d_\Delta(0) = 0$. d_Δ, d_w are measured in m, v_Δ, v_w in m/s.

Table 1. Parameters (unscaled)

Parameter	Value	Unit	Description
L	4	[m]	length of fluid container
b	1	[m]	width of fluid container
$h^{(0)}$	1	[m]	initial height of fluid level
ρ	1000	[kg/m^3]	density of fluid (water)
m_{tr}	2000	[kg]	mass of truck
m_w	$\rho\, b\, h^{(0)} L$	[kg]	mass of fluid container
c	40000	[N/m]	spring force constant
k	10000	[Ns/m]	damper force constant
g	9.81	[N/kg]	earth acceleration

Feasible values for the terminal time T (in s) are taken from [11]. The control u is considered between the bounds $u_{min} = -20000/m_{tr}$ and $u_{max} = 2000/m_{tr}$. We start with $u^{(0)} = u_{max}/2 = const$. Furthermore we set $h_j^i \equiv 1$, $i = 0, \ldots, N$, $j = 1, \ldots, M-1$. The other values of $z^{\Delta x,i}$, $i = 0, \ldots N$, unless they are determined by initial values are set to zero at the start of the Newton method.

5.2.1. *Example 1, as in [11].* We consider here the situation $B \equiv 0$. For the control problem we work with the weights $\alpha_1 = 1$, $\alpha_2 = 0.01/m_{tr}^2$, $\alpha_3 = 0$, $\alpha_4 = 10^3$, $\sigma_4 = -10^{-4}$, $\alpha_5 = 100$, and $\sigma_5 = -10^{-5}$.

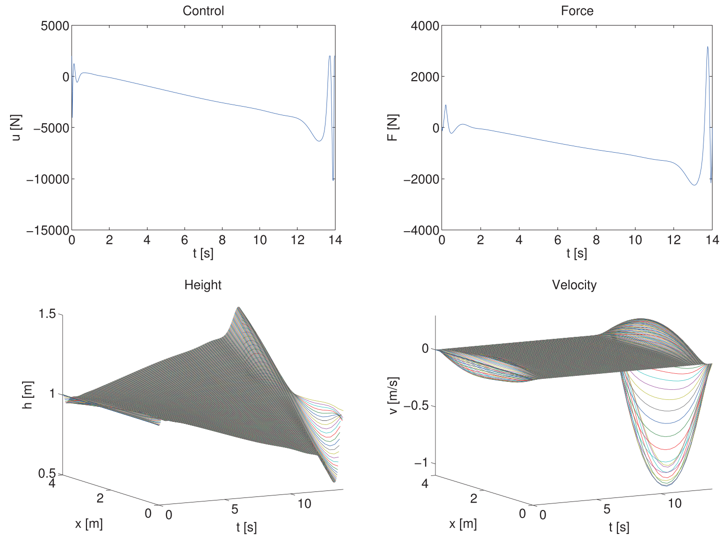

Figure 2. Unscaled computed optimal control $m_{tr}u$ vs. time t (top left), unscaled computed spring-damper force $m_{tr}F$ vs. time t (top right), computed fluid height h vs. time and space (t, x) (bottom left), and computed horizontal fluid velocity v vs. time and space (t, x) for a safety braking maneuver. We observe an excitation of h and v shortly before the end of the braking maneuver. This is reflected in the control, that has a general behavior turning from almost maximal acceleration to maximal deceleration, by some counteractions at the begin and at the end. The coupling force has a qualitatively similar behavior as the approximate control.

For the space discretization we consider $M = 20$ and a factor of 30 for the time discretization, yielding $N = 600$. Here the CFL number (and thus the factor 30) is suggested by the numerical results in [11, Subsection 3.3] and the CFL condition (5.1), that depends itself on the numerical solution, is verified *a posteriori*. As in [11] we find for the artificial viscosity $\varepsilon \approx 0.85714286$ for this example.

For a safety breaking maneuver, i.e. with $T = 14$, the numerical optimal control u, the spring-damper force, the vertical fluid level, and the horizontal fluid velocity are depicted in Figure 2. Our algorithm requires about 30 Newton iterations and yields a feasibility of the terminal constraints smaller than 10^{-7}.

5.2.2. *Example 2.* Now we consider the situation $B = -0.05\sin(\pi x/L)$. For the control problem we work with the weights $\alpha_1 = 5$, $\alpha_2 = 0.01/m_{tr}^2$, $\alpha_3 = 0$, $\alpha_4 = 10^3$, $\sigma_4 = 10^{-4}$, $\alpha_5 = 10$, $\sigma_5 = 10^{-6}$. For the discretization we consider again $M = 20$ and $N = 600$. Again, the CFL condition is checked numerically. Consequently, the numerical viscosity ε has the same value as in Example 1.

For a safety breaking maneuver, i.e. with $T = 14$, the numerical optimal control u, the spring-damper force, the vertical fluid level, and the horizontal fluid velocity

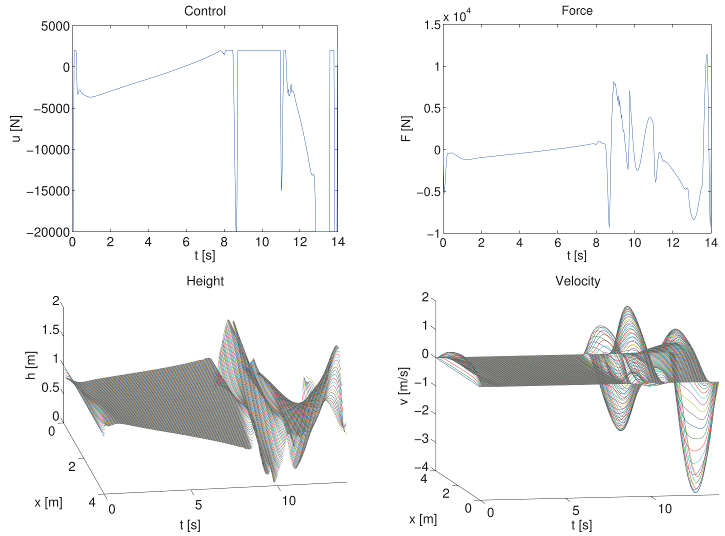

Figure 3. Unscaled computed optimal control $m_{tr}u$ vs. time t (top left), unscaled computed spring-damper force $m_{tr}F$ vs. time t (top right), computed fluid height h vs. time and space (t, x) (bottom left), and computed horizontal fluid velocity v vs. time and space (t, x) for a safety braking maneuver. We observe an excitation of h and v in the second half of the braking maneuver. This is reflected in the control, that has a general behavior reaching from maximal acceleration to maximal deceleration. In contrast to Example 1 we find oscillations that might be due to the slope of the container bottom B.

are depicted in Figure 3. Our algorithm requires about 80 Newton iterations and yields a feasibility of the terminal constraints smaller than 10^{-6}.

6. Conclusion and outlook

We compare with the results obtained in Gerdts *et al.* [11] by a FDTO approach, but for a free terminal time T with a further contribution $\alpha_0 T$ in the objective (with $\alpha_0 > 0$). The obtained numerical results in Example 1 are almost identical. Note that in [11] and in Example 1 different weights are considered as in Example 2. The mild oscillatory behavior of the spring-damper force in Example 2 might be explained by a swinging regime of the spring-damper element and the wave character of h and v, describing shallow water waves.

The initial guess $(y^0, u^0, \lambda^0)^\top$ turns out to be crucial for the performance of our algorithm. In the first Newton iterations we observe with our method the theoretically predicted super-linear convergence. However, for the last iterations this fast convergence is not always observed due to issues with the numerical precision. In particular, our algorithm terminates with less than 100 iterations, while the FDTO approach in [11] requires up to about 3900 iterations of the SQP method. Furthermore in [11] no reformulation introducing d_Δ is exploited. A FDTO optimization approach [33], taking into account the particular structure of the problem, features super-linear convergence, but for a certain range of parameters only.

The reason for considering a FOTD approach like Algorithm 4.5 is, in addition to theoretical insight, that a faster convergence, i.e. less iterations and computing times, are obtained by discretizing in the second step, not before the optimization. However, the numerical precision does not outperform our first approach. In the FDTO ansatz the Courant-Friedrichs-Levy (CFL) condition, that is required for numerical stability of the Lax-Friedrichs scheme, leads to a time discretization finer than the space discretization by a factor of 30. For the computing times of numerical optimal control this is unfavorable, but we meet again this issue in our FOTD approach.

As next step we study further the convergence properties of the global Newton method and how they could be improved by exploiting the structure of the problem. It could be interesting to consider free terminal times. This case would require to adapt our techniques to the non-linearities in T. Furthermore, more simulations for a variation of different parameter sets are of interest. In the near future, we will extend our model to the situation, where the truck moves on the surface of a three-dimensional landscape together with simulating the fluid by the 2d Saint-Venant equations. We might also think of a truck with a semitrailer, involving the drive dynamics both of the drawing vehicle and of the semitrailer with the fluid container.

References

[1] H. W. Alt, *Lineare Funktionalanalysis,* Springer, Berlin-Heidelberg, 3rd ed.,1999.

[2] T. Apel and T. Flaig, *Crank-Nicolson schemes for optimal control problems with evolution equations*, SIAM J. Numer. Anal. **50** (2012), 1484–1512.

[3] A. Bermudez, *Some applications of optimal control theory of distributed systems*, ESAIM Control Opt. Calc. Var. **8** (2002), 195–218.

[4] A. Borzi and V. Schulz, *Computational Optimization of Systems Governed by Partial Differential Equations*, Computational Science and Engineering 8, Society for Industrial and Applied Mathematics (SIAM), Philadelphia, 2012.

[5] K. Chudej, H. J. Pesch, M. Wächter, G. Sachs and F. Le Bras, *Instationary heat-constrained trajectory optimization of a hypersonic space vehicle by ODE-PDE-constrained optimal control*, in: Variational analysis and aerospace engineering, G. Buttazzo, A. Frediani (eds.), Springer Optimization and Its Applications 33, Springer, New York (2009), 127–144.

[6] J.-M. Coron, *Control and Nonlinearity*, Mathematical Surveys and Monographs 136, American Mathematical Society (AMS), Providence, RI, 2007.

[7] L. C. Evans, *Partial Differential equations*, Graduate Studies in Mathematics 19, American Mathematical Society (AMS), Providence, RI, 2nd ed., 2010.

[8] J. Fuhrmann, D. Hömberg and J. Sokolowski, *Modeling, simulation and control of laser heat treatments*, in *Optimal Control of Complex Structures*, K.-H. Hoffmann, I. Lasiecka, G. Leugering, J. Sprekels and F. Tröltzsch (eds.), ISNM International Series of Numerical Mathematics 139, Springer, Heidelberg, 2002, pp. 71–82.

[9] M. Gerdts, *Optimal control of ODEs and DAEs*, DeGruyter, Berlin, 2012.

[10] M. Gerdts, S. Horn and S.-J. Kimmerle, *Line search globalization of a semismooth Newton method for operator equations in Hilbert spaces with applications in optimal control*, to appear in J. Ind. Manag. Optim., 2016, doi 10.3994/jimo.2016003.

[11] M. Gerdts and S.-J. Kimmerle, *Numerical optimal control of a coupled ODE-PDE model of a truck with a fluid basin*, Dyn. Syst. Differ. Equ., AIMS Proceedings **2015** (2015), 515–524.

[12] M. Gugat and G. Leugering, *Global boundary controllability of the De St. Venant equations between steady states*, Ann. Inst. H. Poincaré Anal. Non Linéaire **20** (2003), 1–11.

[13] M. Gugat and G. Leugering, *Global boundary controllability of the Saint-Venant system for sloped canals with friction*, Ann. Inst. H. Poincaré Anal. Non Linéaire **26** (2009), 257–270.

[14] N. Gupta, N. Nataraj and A. K. Pani, *On the optimal control problem of laser surface hardening*, Int. J. Numer. Anal. Model. **7** (2010), 667–680.

[15] M. Hinze, M. Köster, and S. Turek, *A space–time multigrid method for optimal flow control*, in: *Constrained Optimization and Optimal Control for Partial Differential Equations*, G. Leugering, S. Engell, A. Griewank, M. Hinze, R. Rannacher, V. Schulz, M. Ulbrich, S. Ulbrich (eds.), ISNM International Series of Numerical Mathematics 160, Birkhäuser/Springer, Basel, 2012, pp. 147–170.

[16] M. Hinze, R. Pinnau, M. Ulbrich and S. Ulbrich, *Optimization with PDE constraints*, Mathematical Modelling: Theory and Applications 23, Springer, 2009.

[17] D. Hömberg and J. Sokolowski, *Optimal control of laser hardening*, Adv. Math. Sci. **8** (1998), 911–928.

[18] D. Hömberg and S. Volkwein, *Control of laser surface hardening by a reduced-order approach using proper orthogonal decomposition*, Math. Comput. Model. **38** (2003), 1003–1028.

[19] K. Ito and K. Kunisch, *Lagrange multiplier approach to variational problems and applications*, Advances in Design and Control 15, Society for Industrial and Applied Mathematics (SIAM), Philadelphia, 2008.

[20] S.-J. Kimmerle, *Homogenisierung alternierender Randbedingungen bei der Poissongleichung*, diploma thesis, in German with an abstract in English, Ruprecht-Karls-Universität Heidelberg, 2004.

[21] S.-J. Kimmerle, *Well-posedness of a coupled quasilinear parabolic and elliptic free boundary problem from a model for precipitation in crystalline solids*, preprint (2010).

[22] S.-J. Kimmerle, *Optimal control of mean field models for phase transitions*, in: 7th Vienna International Conference on Mathematical Modelling, Vienna, Austria, February 14–17, 2012, IFAC Mathematical Modelling **7** (2013), 1107–1111.

[23] S.-J. Kimmerle and R. Moritz, *Optimal Control of an Elastic Tyre-Damper System with Road Contact*, Proc. Apply. Math. Mech. **14** (2014), 875–876.

[24] S.-J. Kimmerle, M. Gerdts and R. Herzog, *Optimal control of an elastic crane-trolley-load system – A case study for optimal control of coupled ODE-PDE systems*, preprint, Universität der Bundeswehr München, Neubiberg, 2015. Electronic version with supplementary for download on `http://www.unibw.de/sven-joachim.kimmerle`.

[25] D. Kroener, *Numerical Schemes for Conservation Laws*, Wiley, Chichester / Teubner, Stuttgart, 1997.

[26] P. Kosmol, *Methoden zur numerischen Behandlung nichtlinearer Gleichungen und Optimierungsaufgaben*, Teubner Studienbücher Mathematik, B. G. Teubner, Stuttgart, 2nd, revised ed., 1993.

[27] P. D. Lax, *Hyperbolic Partial Differential Equations*, with an appendix by C. S. Morawetz, Courant Lecture Notes in Mathematics 14, New York University, Courant Institute of Mathematical Sciences, New York / American Mathematical Society (AMS), Providence, RI, 2006.

[28] A. Martínez, C. Rodríguez, and M. E. Vázquez-Méndez, *Theoretical and numerical analysis of an optimal control problem related to wastewater treatment*, SIAM J. Control Optim. **38** (2000), 1534–1553.

[29] B. Niethammer, *Derivation of the LSW theory for Ostwald ripening by homogenization methods*, Arch. Ration. Mech. Anal. **147** (1999), 119–178.

[30] H. J. Pesch, A. Rund, W. Von Wahl and S. Wendl, *On some new phenomena in state-constrained optimal control if ODEs as well as PDEs are involved*, Control Cybern. **39** (2010), 647–660.

[31] J. P. Raymond, H. Zidani, *Hamiltonian Pontryagin's principles for control problems governed by semilinear parabolic equations*, Appl. Math. Optim. **39** (1999), 143–177.

[32] F. Tröltzsch, *Optimale Steuerung partieller Differentialgleichungen. Theorie, Verfahren und Anwendungen*, Vieweg, Wiesbaden, 1st ed., 2005.

[33] J.-H. Webert, *Structure-exploiting optimization algorithms for an optimal control problem with coupled hyperbolic and ordinary differential equation constraints*, Master's thesis, Universität der Bundeswehr München, Neubiberg, 2015.

[34] S. Wendl, A. Rund and H. J. Pesch: *On a State-Constrained PDE Optimal Control Problem arising from ODE-PDE Optimal Control*, in: Recent Advances in Optimization and its Applications in Engineering, M. Diehl, F. Glineur and W. Michiels (eds.), Springer, Berlin/Heidelberg, 2010, pp. 429–438.

Manuscript received
revised

S.-J. Kimmerle
Universität der Bundeswehr München, Werner-Heisenberg-Weg 39, 85577 Neubiberg/München, Germany
E-mail address: sven-joachim.kimmerle@unibw.de

M. Gerdts
Universität der Bundeswehr München, Werner-Heisenberg-Weg 39, 85577 Neubiberg/München, Germany
E-mail address: matthias.gerdts@unibw.de

4.3 Articles: Optimal Control of Elastic Structure-Load Systems

In the following two papers [KGH18a] and [Ki16] an optimal control problem subject to a linear elliptic PDE of second-order, the Navier-Lamé system, coupled to nonlinear second-order ODEs is considered. In the first article the elastic beam-like structure represents an elastic crane and the structure is fixed at one end. If we fix the beam-like structure on both ends this models an elastic bridge. A special case of an elastic bridge-load system is considered in the second paper, where the coupling is one-sided, and the ODE solution, representing the motion of vehicles over the bridge, may be prescribed in principle. The motion of the vehicles is subject to a control. In the problem for the elastic crane the ODE models a trolley at position $q_1 = x_T$ that transports a load with angle $q_2 = \alpha$ and we have fully coupled differential equations. We may control the acceleration of the trolley.

Note that the classical problem of a trolley-load system (without elasticity), see Example 1.1 in the introduction, has been considered as optimal control problem in [CG11, CG12].

In [KGH18a, Subsect. 2.2–2.4] the differential equations are derived by the Lagrange mechanism as known in mechanics (see Appendix C). The mechanical displacement field u (considered here in undeformed coordinates) enters the ODE for the multibody system for trolley and load. On one hand it is more realistic to model the contact between the trolley and the beam by an area Γ_C instead of a point and point loads are an technical issue in PDEs, on the other hand we wish to consider a single ODE for the trolley's centre of mass. This motivates the introduction of averages of derivatives of u over the contact area Γ_C, denoted by $\bar{u}$. Note that Γ_C depends also on $q_1(t)$. We observe that

$$\begin{aligned} &\underline{q} \rightsquigarrow \text{Neumann boundary condition for } \underline{u}, \\ &\underline{\bar{u}} \rightsquigarrow \underline{q} \text{ (mass matrix \& force term).} \end{aligned}$$

Furthermore, we have another coupling between $\bar{u}$ and u in the static elasticity PDE, yielding a semilinear problem unless $\bar{u}$ is considered as an independent variable subject to an algebraic equation (its definition as a mean).

The control U enters into the Neumann boundary term of the PDE as well as into the force term of the ODE. The objective is a linear combination of the possibly free terminal time, the kinetic energy, possibly the control effort, and penalty terms for the control, that are handled by an augmented Lagrangian approach. We consider standard box constraints for the control, whereas state constraints, that are required in principle by the model, may be safely neglected since they never get active for the considered data.

The coupling structure is more complicated than in the truck-container example due to these averages over derivatives and the formal adjoint structure involves integro-differential equations that seem to be more difficult to handle for optimal control. For this reason a FDTO approach based on sensitivities is pursued. We consider the reduced objective, i.e. a direct shooting method. The discretized optimal control problem is solved by a projected gradient method with BFGS update. Due to non-smoothness of the F-derivative of the moving boundary condition $\underline{\tilde{g}}$ at the contact area Γ_C w.r.t. the trolley position q_1 to be controlled (see [KGH18b] for details) a Newton-type method relying on the exact Hessian is not available. The PDE is solved by the finite element method, using quadratic Lagrange elements, and the ODE by means of the explicit Heun method of second-order. The numerical optimal control has been computed by means of the open software package FEniCS [LMW12].

The elastic bridge-load problem has been included here mainly for illustration purposes. In comparison to the elastic crane-trolley-load problem the coupling is not two-sided, the objective is to reduce the maximum of the absolute displacement $\|\underline{u}\|$ of the bridge, and the control is the shape of the moving Neumann boundary condition (representing vehicles with various loads and driving at different distances between each other). Due to the multilateral type of control, the behaviour of the system is studied for different given controls only. For validation the model is considered once more when the 3D beam has been replaced by a flat 2D plate, i.e. the Navier-Lame system is replaced by the plate equation that is an elliptic forth-order PDE, but again fully coupled to $\underline{\bar{u}}$ terms. The latter reduction allows for a comparison with standard formulas from engineering literature for the maximum of $\|\underline{u}\|$. For the numerical solution of the plate equation a so-called hybrid finite element method has been employed. This has been implemented in MATLAB.

A further reduction might be to model the elastic structure by a cantilever beam that is 1D and prescribed by a curve $u(x_1)$. Here u corresponds to the vertical displacement u_3 in the 2D or 3D problem. A cantilever beam is subject to the Euler-Bernoulli equation

$$\frac{d^2}{dx_1^2} M(u) = f, \quad M = EI\, u''_{1,1}(x_1) \quad \text{in } (0, \ell_2),$$

where M is called the bending moment and F is a distributed load. EI is the given flexural rigidity that is actually the product of the elastic Young modulus E with I, being the geometrical moment of inertia (second moment of area) of the beam's cross section, calculated w.r.t. the corresponding axis passing through the center of mass and perpendicular to the load. This leads to another optimal control problem with less complexity since the PDE is 1D. But it is not clear if significant coupling effects at the contact area Γ_C of the trolley might be neglected as in this simplification.

In the extended version [KGH17] further technical details for the modelling and the optimal control algorithm for Problem 2a are added. In the paper [KGH18b], not included here, the generalization of the model in [KGH18a], living in a 2D plane, to 3D including rotations of the crane is considered. In addition, the model is extended to include damping and moments of inertia in the multibody system. Furthermore, it contains a general result on the F-differentiability of the control-to-state operator w.r.t. the control of the trolley of the 2D and 3D problem. Note that the elastic bridge-load problem is naturally restricted to a 2D plane.

Finally, we discuss the relevant assumptions from Chapter 3 for the elastic crane-trolley-load problem. For the well-posedness of the coupled state equations for given control $u \in U_{ad} :=$

$\{\tilde{u} \in \mathcal{U} := W^{1,\infty}(0,t_f) \,|\, u_{min} \leq \tilde{u} \leq u_{max}\}$ with $u_{min} < u_{max}$, we have the states $\underline{y} = [u, q_1, q_2]^\top$ and the corresponding space

$$Y = L^\infty(0,t_f; H^2(\Omega)) \times [H^2(0,t_f)]^2$$

within the notation[2] of this study. By smoothing the characteristic function for the contact area Γ_C, where a non-zero boundary condition holds, the spatial regularity of the displacement field may be increased further. Note that in order to avoid technical difficulties with corner singularities these are smoothed out. The solvability of the coupled state equations for sufficiently short times t_f and sufficiently small $|\Gamma_C|$ is proved in [KGH18a, Th. 3.2] using an approach similar to Theorem 3.4.

The control space $\mathcal{U}$ is isomorphic to $C^{0,1}([0,t_f])$ and embeds into $L^2(0,t_f)$ and U_{ad} is a closed convex subset of $\mathcal{U}$. The idea to prove the existence of optimal controls, see [KGH18b, Sect. 3], is to consider the control-to-state operator $S : \mathcal{U} \to Y, U \mapsto \underline{y}$ and to show its F-differentiability w.r.t. U. In particular, the F-differentiability of the moving boundary condition w.r.t. q_1 is non-standard. For this task we could consider either the two ODEs as an elliptic PDE system of four first-order equations or to consider the elliptic PDE as an DAE in suitable function spaces. Note that the model problem, Problem 3.7, is restricted to parabolic ODE (here of second-order). In the following article the averaging-evaluation operator reads

$$\begin{aligned}\underline{\mathcal{E}}(\underline{u}) &:= \underline{\bar{u}}(t) \\ &:= \frac{1}{|\Gamma_C|} \int_{\Gamma_C(q_1(t))} [\partial_1 u_1(t,\underline{x}), \partial_1 u_3(t,\underline{x}), D_1 u_1(t,\underline{x}), D_1 u_3(t,\underline{x}), \partial_3 u_1(t,\underline{x}), \partial_3 u_3(t,\underline{x})]^\top \, d\underline{x},\end{aligned}$$

where the average is taken over each component and we abbreviate $D_1 := \sum_{i=1}^3 \partial_{i,1}$. In [KGH18b] the average is taken over further derivatives as well, which is due to the 3D situation.

In [KGH18a] the bang-bang principle for optimal control of PDE without control efforts seems to be violated. A similar phenomenon has been observed in [PRWW10] as well. But, it is not clear yet, whether this is to due to terminal constraints and the coupling to u or this is only a numerical artifact.

[2] Furthermore please note that in these papers [KGH18a, KGH17, KGH18b, Ki16] U denotes the control, u the mechanical displacement field, and $\mathcal{U}$ the control space.

MATHEMATICAL AND COMPUTER MODELLING OF DYNAMICAL SYSTEMS, 2017
https://doi.org/10.1080/13873954.2017.1405046

Optimal control of an elastic crane-trolley-load system - a case study for optimal control of coupled ODE-PDE systems

S.-J. Kimmerle [a], M. Gerdts [b] and R. Herzog [c]

[a]Institut für Mathematik und Bauinformatik, Universität der Bundeswehr München, Neubiberg/München, Germany; [b]Institut für Mathematik und Rechneranwendung, Universität der Bundeswehr München, Neubiberg/München, Germany; [c]Fakultät für Mathematik, Technische Universität Chemnitz, Chemnitz, Germany

ABSTRACT

We present a mathematical model of a crane-trolley-load model, where the crane beam is subject to the partial differential equation (PDE) of static linear elasticity and the motion of the load is described by the dynamics of a pendulum that is fixed to a trolley moving along the crane beam. The resulting problem serves as a case study for optimal control of fully coupled partial and ordinary differential equations (ODEs). This particular type of coupled systems arises from many applications involving mechanical multi-body systems. We motivate the coupled ODE-PDE model, show its analytical well-posedness locally in time and examine the corresponding optimal control problem numerically by means of a projected gradient method with Broyden-Fletcher-Goldfarb-Shanno (BFGS) update.

ARTICLE HISTORY
Received 8 May 2017
Accepted 9 November 2017

KEYWORDS
Crane-trolley-load system; ODE-PDE-constrained optimization; optimal control of ODE; coupled ODE-PDE system; projected gradient method; FEM for moving boundary conditions

1. Introduction

In this article we consider a crane model, where the crane arm is modelled by an elastic beam Ω fixed at one end to the crane tower. The load is represented as a mathematical pendulum that is fixed to a trolley which moves along the crane beam. The mechanical displacement $\boldsymbol{u}$ of the crane beam is determined by the elliptic partial differential equation of static linear elasticity. The trolley and load states $\boldsymbol{q} = (q_1, q_2)$, that is, the position of the trolley and the angle of the load are subject to an ordinary differential equation. It turns out that all differential equations are fully coupled with one another. The goal is to transport the load by means of the trolley along the crane beam from a given initial position $\boldsymbol{q}^0$ and initial velocity $\boldsymbol{v}^0$ to a designated terminal position $\boldsymbol{q}^f$ with terminal velocity $\boldsymbol{v}^f$, while minimizing vibrations as well as the total time T. Here, the control is given by the acceleration force U of the trolley. For the relevance of this optimal control problem (OCP) in engineering, see [1,2]. The optimal control of a gantry crane has been examined by Biswas [3] who considers additionally the displacement of a non-rigid cable fixing the load, while the rails are considered as unflexible. This different model yields a coupled ordinary differential equation (ODE)-partial differential equation (PDE) system as well, but with a one-dimensional PDE.

For the geometry of our model, we refer to Figure 1. Our OCP is to find

$$\min_{U,T,\boldsymbol{q}} J(\boldsymbol{q}, \dot{\boldsymbol{q}}, U, T) \tag{1}$$

with the objective

CONTACT S.-J. Kimmerle ✉ sven-joachim.kimmerle@unibw.de

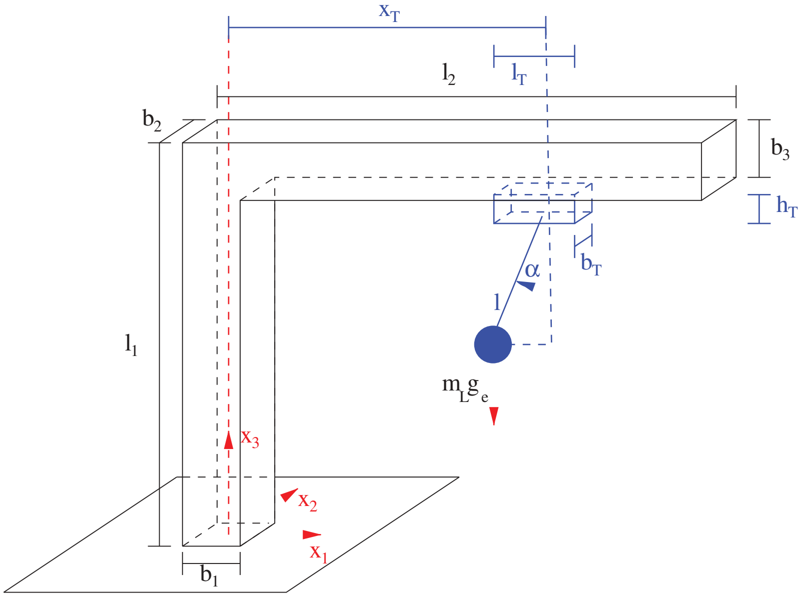

Figure 1. Configuration of the elastic crane (within Lagrangian coordinates).

$$
\begin{aligned}
J(\boldsymbol{q},\dot{\boldsymbol{q}},U,T) = {} & \nu_1 T + \frac{\nu_2}{2}\|\dot{q}_2\|^2_{L^2(0,T)} + \frac{\nu_3}{2}\|U\|^2_{L^2(0,T)} \\
& + \sum_{i=0}^{1}\frac{\nu_{4+i}}{2}|q_i(T)-q_i^f|^2 + \sum_{i=0}^{1}\frac{\nu_{6+i}}{2}|\dot{q}_i(T)|^2
\end{aligned} \tag{2}
$$

subject to the elliptic PDE

$$
-\operatorname{div}\sigma(\boldsymbol{u}) = \boldsymbol{H} \quad \text{in } \Omega \times [0,T] \tag{3}
$$

and the ODE system

$$
\mathbf{M}(\boldsymbol{q},\bar{\boldsymbol{u}})\ddot{\boldsymbol{q}} = \boldsymbol{F}(\boldsymbol{q},\dot{\boldsymbol{q}},\bar{\boldsymbol{u}},U) \quad \text{in } [0,T], \tag{4}
$$

and the PDE and ODE are completed by boundary and initial conditions, respectively. Here, $\boldsymbol{\nu} \in \mathbb{R}^7$ denotes a non-zero vector of non-negative weights. The state ***q(t)*** denotes the vector of generalized coordinates of the rigid bodies at time t (see Section 2.3). The right hand side $\boldsymbol{H}$ encodes the gravitational forces due to the weight of the beam itself, $\mathbf{M}$ the mass matrix and $\boldsymbol{F}$ comprises generalized Coriolis forces and external forces. We mention as a particular feature of our model that only mean values $\bar{\boldsymbol{u}}$ of the beam's displacement $\boldsymbol{u}$, averaged over the surface $\Gamma_C(q_1) = q_1 + \Gamma_C(q_1^0)$ connecting the trolley to the beam (for the precise definition see Figure 1 and Section 2.1), enter into the ODE (Equation (4)). The differential equations are completed by boundary and initial conditions. Time-optimal control is realized by scaling to a fixed time interval, yielding the total time T only as a control parameter. The scaled full problem is stated precisely in Section 2.

Our problem class cannot be solved directly by standard software owing to the strong coupling of the constraints. In this study, we prove the local-in-time well-posedness of the coupled dynamical problem and present an algorithm for solving the OCP, based on a first-discretize-then-optimize (FDTO) approach. The non-standard Algorithms 4.1–4.4 are designed specifically for our problem and are based on a projected gradient method without and with Broyden-Fletcher-Goldfarb-Shanno (BFGS) update, respectively. We emphasize that it is not clear whether Newton-based methods as in Chen and Gerdts [4,5] could be applied in our situation. On the one hand, it is not obvious whether our problem exhibits the smoothness needed for Newton methods, and on the other hand, this approach requires the solution of the necessary optimality conditions that turns out to be challenging. Finally we present numerical results for the optimal control of the crane-trolley-load system. The open-source finite element software FEniCS [6] is employed to solve the PDEs of linear elasticity. Special attention has to be given to the elements at surfaces where forces are applied, that is, at the free boundary between moving trolley and crane beam. For ease of presentation, we focus in this article on a crane that does not rotate.

In the modelling of multi-body systems we often encounter the following situation. We find ODEs representing the interactions between the centres of mass, algebraic equations from constraining forces and elliptic PDEs modelling mechanical deformations within the bodies. Here, we focus on the crane-trolley-load system serving as a case study. Other applications include, for example, (a) a quarter-car model, where an elastic wheel-tyre-damper system with free road contact is controlled [7], (b) the heat-optimal ascent and re-entry into atmosphere of a hypersonic spacecraft [8] and (c) a truck or plane with a tank filled with fluid [9,10]. In the latter example, the PDEs are the St-Venant equations yielding the height and velocity of the fluid. However, it should be emphasized that the class of OCPs subject to ODE and PDE constraints is heterogeneous, since different types of PDE require already different theories and methods. Furthermore, the methods for this problem class depend on the particular coupling structure between ODE and PDE.

On optimal control of PDE and optimal control of differential-algebraic equations (DAE) alone there exists a wide variety of results and numerical approaches; see, for example, the overview article [11] or the textbooks [12–15] for control of PDEs and [16,17] for optimal control of ODE and DAE, respectively.

However, only few results about *combined* ODE-PDE constrained optimal control exist so far. Biswas et al. [18,19] examine control and optimal control of a large flexible space structure, e.g. a satellite, subject to PDE modelling vibrations and to ODE describing the dynamics. Chudej et al. [8] consider optimal control for coupling of the heat equation and equations of motions. Similar as in our situation, the coupling from ODE to PDE is effected by means of a boundary condition, but the controls arise in the ODE system only and the PDE is considered in one space dimension. In our problem, the controls arise in the ODE as well as in the boundary condition. Our control acts on the PDE by means of a Neumann boundary control. In our problem, we encounter control constraints too. For a particular class of ODE-PDE-problem, a so-called hypersonic rocket car problem, including the problem of [8], some new phenomena have been discovered by Pesch et al. [20]. In [21], this OCP is reformulated as a state-constrained OCP for PDE and necessary optimality conditions for it are derived.

Our study is organized as follows. In Section 2, we derive the mathematical model. The local-in-time well-posedness of our model is shown in Section 3. Due to the complexity of the resulting coupled ODE-PDE problem, we apply here the direct FDTO method, and follow a sensitivity-based approach. The discretization for the numerical simulation, the projected gradient method and the quasi-Newton method BFGS for the optimal control are described in Section 4. In Section 5, numerical results are presented. Finally, we close with a discussion of our results and give, in particular, an outlook on open questions and future work.

2. Mathematical model of the elastic crane

2.1. *Geometry*

For the geometry of a crane in the undeformed configuration, see Figure 1. In this study, we consider no rotation of the crane beam, yielding that the pendulum is restricted to the plane defined by $x_2 \equiv 0$. The full 3d model with a rotating crane is presented in the upcoming study [22].

The deformations of the crane beam are here considered in a full 3d model avoiding the issue, how to replace the crane beam by a lower dimensional model sensitively. The domain

$$\Omega = \{\boldsymbol{x} = (x_1, x_2, x_3) \in \mathbb{R}^3 \,|\, b_1/2 \le x_1 \le l_2 - b_1/2,\ |x_2| \le b_2/2,\ l_1 - b_3 \le x_3 \le l_1\},$$

where $b_1 < l_2$, $b_3 < l_1$, describes the undeformed extension arm of the crane, which is considered elastic and attached to the rigid vertical bar. On $\Gamma_D = \partial\Omega \cap \{\boldsymbol{x} \in \mathbb{R}^3 \,|\, x_1 = b_1/2\}$ the extension arm is fixed.

A force is applied on the trolley at its center of mass, located at

$$\boldsymbol{r}_{\mathbf{M}}(t) = \left(x_{\mathrm{T}}(t),\ 0,\ l_1 - b_3 - \frac{h_{\mathrm{T}}}{2}\right)^{\top} \tag{5}$$

in the undeformed state. This load is a pendulum of length l with a point mass m_{L}. The trolley, with mass m_{T}, may move on rails, modelled by translating the x_1-position $x_{\mathrm{T}}(t) \in ((b_1 + l_T)/2, l_2 - (b_1 + l_T)/2)$ of its center of mass. The application of a force U, e.g. by means of a neglectable cable along the beam, at the trolley's center of gravity allows to control the trolley. Alternatively, if the trolley was controlled by a motor within the trolley, moving a wheel that is in contact with rails at the crane beam, this would yield the same model, but with a slightly different area of support Γ_{C}, now being the wheel's area of support. On the boundary $\Gamma_{\mathrm{C}}(x_{\mathrm{T}}) = \{\boldsymbol{x} \in \partial\Omega \mid x_{\mathrm{T}}(t) - l_{\mathrm{T}}/2 \le x_1 \le x_{\mathrm{T}}(t) + l_{\mathrm{T}}/2,\ |x_2| \le b_{\mathrm{T}}/2,\ x_3 = l_1 - b_3\} = \Gamma_{\mathrm{C}}(x_{\mathrm{T}}(0)) + x_{\mathrm{T}}\,\boldsymbol{e}_1$ varying with time, the trolley exerts a force on the extension arm. As usual $\boldsymbol{e}_1$ denotes the unit vector in x_1-direction. In particular, the definition of Γ_{C} implies that the surface area of Γ_{C}, denoted by $|\Gamma_{\mathrm{C}}|$, remains constant. The application of a moment U_{M} at the bearing of the crane beam allows to rotate the crane by an angle β. We assume $\alpha \in (-\pi/2, \pi/2)$.

2.2. *Strains and stresses in the beam*

Within the domain Ω of the crane we aim to solve for the mechanical displacement field. For a realistic crane we may assume small displacement gradients and, hence, model the elastic deformations within the structure of the crane by linear elasticity. As usual in linear elasticity, we do not distingiush between the *reference* (undeformed) configuration Ω (in Lagrangian coordinates $\boldsymbol{x}$) and the deformed configuration $\Omega' \subset \mathbb{R}^3$ (in Eulerian coordinates $\hat{\boldsymbol{x}}$). As reference configuration we consider the crane's extension arm in the absence of strains or stresses. For the interaction between the deformed beam and the trolley, the deformation of the beam is not neglected in our study. In this context, we consider u in the reference configuration. For the deformed configuration the formula (19) would look different, though leading to the same results within the approximation of small displacement gradients. The deformation depends on time through the control. The system is considered on the compact time interval $[0, T] \subset \mathbb{R}$ and thus the mechanical displacement field reads $\boldsymbol{u} : \Omega \times [0, T] \to \Omega'$, $(\boldsymbol{x}, t) \mapsto \boldsymbol{u}(\boldsymbol{x}, t) = \hat{\boldsymbol{x}}(\boldsymbol{x}, t) - \boldsymbol{x}$. The symmetrized strain associated with the displacement field $\boldsymbol{u}$ is $\epsilon(\boldsymbol{u}) = (\nabla\boldsymbol{u} + \nabla\boldsymbol{u}^{\top})/2$ and as a constitutive assumption we work with the Cauchy stress tensor

$$\sigma(\boldsymbol{u}) = \lambda\, \mathrm{trace}\,(\epsilon(\boldsymbol{u}))\mathbf{1} + 2\,\mu\,\epsilon(\boldsymbol{u}), \tag{6}$$

where $\mathbf{1}$ denotes the unit matrix and $\mu > 0$, $\lambda > -2\mu/3$ are the Lamé constants scaled by $1/(l m_{\mathrm{L}})$. For our purposes it suffices to consider $\lambda > 0$. We assume that we may neglect the deformation of the trolley, since typically $h_{\mathrm{T}} \ll b_3$ holds.

Terms of higher order in $\|\nabla\boldsymbol{u}\|$ are neglected, since $\|\nabla\boldsymbol{u}\| \ll 1$. (This is emphasized in the following by the $\approx$ symbol.) For more details see [23, Ch. 3].

2.3. *Governing equations for the trolley and the load*

We introduce the generalized coordinates $\boldsymbol{q} = (x_{\mathrm{T}}, \alpha)^{\top}$, see Figure 1. The trolley's center of gravity at time t is given by

$$\boldsymbol{r}_{\mathrm{T}}(q_1, \boldsymbol{u}, t) = \boldsymbol{r}_{\mathrm{M}}(q_1) + \boldsymbol{u}(r_{\mathrm{M}}(q_1), t), \tag{7}$$

where $\boldsymbol{r}_{\mathrm{M}}$ is defined by (5). The position of the load, considered as point mass, at time t is

$$\boldsymbol{r}_{\mathrm{L}}(\boldsymbol{q}, \boldsymbol{u}, t) = \boldsymbol{r}_{\mathrm{T}}(q_1, \boldsymbol{u}, t) + \begin{pmatrix} l \sin q_2 \\ 0 \\ -l \cos q_2 \end{pmatrix},$$

where $q_2 = \alpha$ is positive for a counterclockwise rotation around the x_2-axis, see Figure 1. The kinetic energy of the mechanical system written for the generalized coordinates $\boldsymbol{q} = (x_{\mathrm{T}}, \alpha)^{\top}$ and scaled by $1/(l m_{\mathrm{L}})$ is given by

$$\mathcal{T}(\boldsymbol{q}, \dot{\boldsymbol{q}}, \boldsymbol{u}, t) = \frac{1}{2 l m_{\mathrm{L}}} \left(m_{\mathrm{T}} \| \dot{\boldsymbol{r}}_{\mathrm{T}}(q_1, \boldsymbol{u}, t) \|^2 + m_{\mathrm{L}} \| \dot{\boldsymbol{r}}_{\mathrm{L}}(\boldsymbol{q}, \boldsymbol{u}, t) \|^2 \right), \tag{8}$$

where we neglect the moment of inertia of the trolley and the load. Furthermore, we have neglected velocities

$$\dot{\boldsymbol{u}} \approx \boldsymbol{0} \tag{9}$$

in (8), which is motivated by a short dimensional analysis (see Subsection 2.4) and is consistent as we will see in our numerics (see, e.g. Figure 3, bottom left). For brevity, we set $m := (m_{\mathrm{T}} + m_{\mathrm{L}})/(l m_{\mathrm{L}})$. The scaled, generalized potential $\mathcal{V}$, uniquely determined up to a constant, is $\mathcal{V}(\boldsymbol{q}, U) = -U q_1 - g_{\mathrm{e}} \cos q_2$, where U is the control (divided by $l m_{\mathrm{L}}$) and g_{e} is the gravity acceleration. From $\boldsymbol{E} = -\nabla_{\boldsymbol{q}} \mathcal{V}$ we obtain the vector of applied generalized forces

$$\boldsymbol{E}(\boldsymbol{q}) = \begin{pmatrix} U \\ -g_{\mathrm{e}} \sin q_2 \end{pmatrix}, \tag{10}$$

that is, the control acting along the deformed rail of the trolley and gravitation acting on the deformed system. The Lagrange function is $\mathcal{L}\left(\boldsymbol{q}, \dot{\boldsymbol{q}}, \boldsymbol{u}, U\right) = \mathcal{T}(\boldsymbol{q}, \dot{\boldsymbol{q}}, \boldsymbol{u}) - \mathcal{V}(\boldsymbol{q}, U)$.

By means of the Euler-Lagrange equation $\frac{\mathrm{d}}{\mathrm{d}t} \nabla_{\dot{\boldsymbol{q}}} \mathcal{L} = \nabla_{\boldsymbol{q}} \mathcal{L}$ for a given control we derive the equations of motion within our approximation of small displacement gradients. Neglecting $\dot{\boldsymbol{u}}$, we obtain from (7) the time derivative

$$\dot{\boldsymbol{r}}_{\mathrm{T}} = \mathbf{F}_{\mathrm{D}}(\nabla \boldsymbol{u})^{\mathrm{T}} \dot{q}_1 \boldsymbol{e}_1,$$

where $\mathbf{F}_{\mathrm{D}}(\nabla \boldsymbol{u}) = \mathbf{1} + \nabla \boldsymbol{u}$ is the deformation gradient. In our approximation of small displacement gradients we have

$$F_{D,11}^2(\nabla \boldsymbol{u}) + F_{D,21}^2(\nabla \boldsymbol{u}) + F_{D,31}^2(\nabla \boldsymbol{u}) = 1 + 2\partial_1 u_1 + \sum_{i=1}^{3} |\partial_i u_1|^2 \approx 1 + 2\partial_1 u_1.$$

Furthermore, $\dot{\boldsymbol{r}}_{\mathrm{L}} = \dot{\boldsymbol{r}}_{\mathrm{T}} - l(\cos q_2, 0, \sin q_2)^{\top} \dot{q}_2$ and the scaled kinetic energy reads

$$\begin{aligned} \mathcal{T}(\boldsymbol{q}, \dot{\boldsymbol{q}}, \boldsymbol{u}, t) = {} & \frac{m}{2}(1 + 2\partial_1 u_1(\boldsymbol{r}_{\mathrm{M}}(q_1), t))\dot{q}_1^2 + \frac{l}{2}\dot{q}_2^2 \\ & + ((1 + \partial_1 u_1(\boldsymbol{r}_{\mathrm{M}}(q_1), t)) \cos q_2 + \partial_1 u_3(\boldsymbol{r}_{\mathrm{M}}(q_1), t) \sin q_2)\dot{q}_1 \dot{q}_2. \end{aligned}$$

For brevity, we introduce the notation $\partial_i := \partial_{x_i}$ and the operator D_1 defined by

$$D_1 u_i := \partial^2_{1,1} u_i + \partial^2_{2,1} u_i + \partial^2_{3,1} u_i, \quad i = 1, 2, 3, \tag{11}$$

and abbreviate

$$\Phi_1(q_2, \mathbf{D}^2\boldsymbol{u}) = D_1 u_1 \cos q_2 + D_1 u_3 \sin q_2,$$

$$\Phi_2(q_2, \mathbf{D}\boldsymbol{u}) = -(1+\partial_1 u_1) \sin q_2 + \partial_1 u_3 \cos q_2.$$

We compute

$$\partial_{\boldsymbol{q}} \mathcal{T}(\boldsymbol{q}, \dot{\boldsymbol{q}}, \boldsymbol{u}) = \begin{pmatrix} m\mathrm{D}_1 u_1 \dot q_1^2 + \Phi_1(q_2, \mathbf{D}^2\boldsymbol{u}) \dot q_1 \dot q_2 \\ \Phi_2(q_2, \mathbf{D}\boldsymbol{u}) \dot q_1 \dot q_2 \end{pmatrix}$$

and

$$\partial_{\dot{\boldsymbol{q}},\boldsymbol{q}} \mathcal{T}\, \dot{\boldsymbol{q}} = \begin{pmatrix} 2mD_1 u_1 \dot q_1^2 + \Phi_1(q_2, \mathbf{D}^2\boldsymbol{u}) \dot q_1 \dot q_2 + \Phi_2(q_2, \mathbf{D}\boldsymbol{u}) \dot q_2^2 \\ \Phi_1(q_2, \mathbf{D}^2\boldsymbol{u}) \dot q_1^2 + \Phi_2(q_2, \mathbf{D}\boldsymbol{u}) \dot q_1 \dot q_2 \end{pmatrix}.$$

Thus the generalized Coriolis forces are

$$\boldsymbol{G}(\boldsymbol{q}, \dot{\boldsymbol{q}}, \boldsymbol{u}) = \partial_{\boldsymbol{q}} \mathcal{T}(\boldsymbol{q}, \dot{\boldsymbol{q}}, \boldsymbol{u}) - \partial^2_{\dot{\boldsymbol{q}},\boldsymbol{q}} \mathcal{T}(\boldsymbol{q}, \dot{\boldsymbol{q}}, \boldsymbol{u}) \dot{\boldsymbol{q}} = \begin{pmatrix} -mD_1 u_1 \dot q_1^2 - \Phi_2(q_2, \mathbf{D}\boldsymbol{u}) \dot q_2^2 \\ -\Phi_1(q_2, \mathbf{D}^2\boldsymbol{u}) \dot q_1^2 \end{pmatrix}.$$

This yields as equation of motion for $\boldsymbol{q}$,

$$\tilde{\mathbf{M}}(\boldsymbol{q}, \boldsymbol{u}) \ddot{\boldsymbol{q}} = \boldsymbol{G}(\boldsymbol{q}, \dot{\boldsymbol{q}}, \boldsymbol{u}) + \boldsymbol{E}(\boldsymbol{q}, U) \tag{12}$$

where the symmetric and positive definite mass matrix reads

$$\tilde{\mathbf{M}}(\boldsymbol{q}, \boldsymbol{u}) = \begin{pmatrix} m(1 + 2\partial_1 u_1) & (1 + \partial_1 u_1) \cos q_2 + \partial_1 u_3 \sin q_2 \\ * & l \end{pmatrix}, \tag{13}$$

and where the generalized Coriolis forces are

$$\boldsymbol{G}(\boldsymbol{q}, \dot{\boldsymbol{q}}, \boldsymbol{u}) = \begin{pmatrix} -mD_1 u_1 \dot q_1^2 + ((1 + \partial_1 u_1) \sin q_2 - \partial_1 u_3 \cos q_2) \dot q_2^2) \\ -(D_1 u_1 \cos q_2 + D_1 u_3 \sin q_2) \dot q_1^2. \end{pmatrix}. \tag{14}$$

The ODE system for itself has a unique solution locally in time, since $\tilde{\mathbf{M}}$ is invertible for $\|\nabla \boldsymbol{u}\| \ll 1$.

2.4. *Governing equations for the crane beam*

For the domain of the crane's extension arm we consider the standard model of linear elasticity [23]. We do not observe significant vibrations in the beam in our simulations. Elastic waves within the crane beam may be safely neglected due to different typical time scales, see [22] for the full elastodynamical problem for the displacement field and its dimensional analysis. Indeed, the speed of shear waves is $\sqrt{l\mu/\rho} \approx 4.73$ km/s and the speed of longitudinal waves, i.e. the speed of sound, is of the same order. This elastostatic problem, where time enters as a parameter, reads

$$-\operatorname{div} \sigma(\boldsymbol{u}) = \boldsymbol{H} \qquad \text{in } \Omega \times [0, T], \tag{15}$$

$$\boldsymbol{u} = \mathbf{0} \qquad \text{on } \Gamma_{\mathrm{D}} \times [0, T], \tag{16}$$

$$-\sigma(\boldsymbol{u}).\boldsymbol{n} - \tilde{\boldsymbol{g}}(\boldsymbol{q}, \boldsymbol{u}, U) = \mathbf{0} \qquad \text{on } \Gamma_{\mathrm{N}} \times [0, T]. \tag{17}$$

where $\tilde{\boldsymbol{g}} : \Gamma_\mathrm{N} \times [0, T] \to \mathbb{R}^3$ is a boundary force (scaled by $1/(l m_\mathrm{L})$), Γ_D is the part of $\partial\Omega$ where the Dirichlet boundary condition (b.c.) holds and $\Gamma_\mathrm{N} := \partial\Omega \backslash \overline{\Gamma_\mathrm{D}}$ the boundary with the Neumann b.c. The term

$$\boldsymbol{H} = -\rho g \boldsymbol{e}_3 \tag{18}$$

in (15) models the gravity of the crane cantilever. Here ρ is the mass density divided by m_L and $g := g_\mathrm{e}/l$ is the reduced gravity acceleration.

Note that we work within the approximation of small displacement gradients and, thus, for consistency the linear deformation of the crane beam has to be modelled for the trolley-load system as well, only higher order terms in $\nabla\boldsymbol{u}$ may be safely neglected. According to (10), the scaled forces $\boldsymbol{E}$ applied to the crane beam read in Cartesian coordinates

$$\boldsymbol{E}_c(\boldsymbol{q}, U) = \begin{pmatrix} U + g \sin q_2 \\ 0 \\ -g \cos q_2 \end{pmatrix}$$

acting in the *deformed* configuration. We assume that this force is realized by a constant pressure acting on the contact surface Γ_C between the trolley and the beam. In order to transform a surface integral from the deformed to the reference configuration we use [23, Th. 1.7–1], also called Nanson's formula. The deformation gradient $\mathbf{F}_\mathrm{D}(\nabla\boldsymbol{u}) = \mathbf{1} + \nabla\boldsymbol{u}$ is invertible for $\|\nabla\boldsymbol{u}\| \ll 1$. Note that $\partial_2 u_i \approx 0$, $i = 1, 2, 3$, on Γ_C. In our situation the scaled gravity and control forces by the load and the trolley onto the crane beam yield $\boldsymbol{g}_0 : \mathbb{R} \times \mathbb{R}^3 \times \mathbb{R} \to \mathbb{R}^3$ by

$$\det(\mathbf{F}_\mathrm{D}(\nabla\boldsymbol{u})^{-1})\,\mathbf{F}_\mathrm{D}^\top(\nabla\boldsymbol{u})(\boldsymbol{E}_\mathrm{c}(\boldsymbol{q}, U) - \eta g \boldsymbol{e}_3)$$

$$\approx ((1 - \mathrm{trace}(\nabla\boldsymbol{u}))\,\mathbf{1} + \nabla\boldsymbol{u}^\top)\,(U + g\sin q_2, 0, -g(\cos q_2 + \eta))^\top =: \boldsymbol{g_0}(\boldsymbol{q}, \boldsymbol{u}, U),$$

where $\eta = m_\mathrm{T}/m_\mathrm{L}$ abbreviates the trolley/load mass ratio. This pressure is distributed on the surface $\partial\Omega \cap \{x_3 = l_1 - b_3\}$ as follows

$$\tilde{\boldsymbol{g}}(\boldsymbol{x}, \boldsymbol{q}(t), \boldsymbol{u}(\boldsymbol{x}, t), U(t)) := \begin{cases} \frac{1}{|\Gamma_\mathrm{C}|}\boldsymbol{g}_0(\boldsymbol{q}(t), \boldsymbol{u}(\boldsymbol{x}, t), U(t)); & \boldsymbol{x} \in \Gamma_\mathrm{C}(q_1(t)), \\ \mathbf{0} \qquad ; & \text{otherwise.} \end{cases} \tag{19}$$

The elliptic PDE problem (15 – 17) for $\boldsymbol{u}$ depends on the trolley/load states $\boldsymbol{q}$ by means of the contact pressure $\tilde{\boldsymbol{g}}$, while the ODE system (12) for $\boldsymbol{q}$ is coupled as it depends on derivatives of $\boldsymbol{u}$. Since the important coupling effect that we would like to consider takes place on a small part Γ_C of the surface, we introduce the mean values

$$\bar{\boldsymbol{u}}(t) := \begin{pmatrix} \bar{u}_1(t) \\ \bar{u}_2(t) \\ \bar{u}_3(t) \\ \bar{u}_4(t) \\ \bar{u}_5(t) \\ \bar{u}_6(t) \end{pmatrix} = \begin{pmatrix} \frac{1}{|\Gamma_\mathrm{C}|}\int_{\Gamma_\mathrm{C}(q_1(t))} \partial_1 u_1(\boldsymbol{x}, t)\,\mathrm{d}\boldsymbol{x} \\ \frac{1}{|\Gamma_\mathrm{C}|}\int_{\Gamma_\mathrm{C}(q_1(t))} \partial_1 u_3(\boldsymbol{x}, t)\,\mathrm{d}\boldsymbol{x} \\ \frac{1}{|\Gamma_\mathrm{C}|}\int_{\Gamma_\mathrm{C}(q_1(t))} D_1 u_1(\boldsymbol{x}, t)\,\mathrm{d}\boldsymbol{x} \\ \frac{1}{|\Gamma_\mathrm{C}|}\int_{\Gamma_\mathrm{C}(q_1(t))} D_1 u_3(\boldsymbol{x}, t)\,\mathrm{d}\boldsymbol{x} \\ \frac{1}{|\Gamma_\mathrm{C}|}\int_{\Gamma_\mathrm{C}(q_1(t))} \partial_3 u_1(\boldsymbol{x}, t)\,\mathrm{d}\boldsymbol{x} \\ \frac{1}{|\Gamma_\mathrm{C}|}\int_{\Gamma_\mathrm{C}(q_1(t))} \partial_3 u_3(\boldsymbol{x}, t)\,\mathrm{d}\boldsymbol{x} \end{pmatrix}, \tag{20}$$

in the coupling terms of the ODE and, consistently, in $\tilde{\boldsymbol{g}}$. For the definition of the operator D_1 see (11). This also implies that we do not have to solve for every point in space $\boldsymbol{x}$ another ODE system. We will consider $\bar{\boldsymbol{u}}$ as an independent state variable in the following.

2.5. Objective function

Our aim is to transport a load at rest ($\boldsymbol{v}^0 = \boldsymbol{0}$) from an initial state $\boldsymbol{q}^0$ to a terminal position q_1^f with angle $q_2^f = 0$, where the load should be at rest, that is, $\boldsymbol{v}^f = \boldsymbol{0}$. We would like to achieve this in minimal time, while also minimizing the swing of the load. Now the objective function (2), consisting of a time-minimal term, a (kinetic) energy-minimal term, possibly a regularization term, and terms penalizing the violation of terminal conditions, reads:

$$\begin{aligned} J(\boldsymbol{q}, \dot{\boldsymbol{q}}, \boldsymbol{U}, T) = \nu_1 T + \frac{\nu_2}{2}\|\dot{q}_2\|^2_{L^2(0,T)} + \frac{\nu_3}{2}\|U\|^2_{L^2(0,T)} + \frac{\nu_4}{2}|q_1(T) - q_1^f|^2 \\ + \frac{\nu_5}{2}|q_2(T)|^2 + \frac{\nu_6}{2}|\dot{q}_1(T)|^2 + \frac{\nu_7}{2}|\dot{q}_2(T)|^2. \end{aligned} \tag{21}$$

Different choices for the weights $\nu_1 > 0$, $\nu_j \geq 0$, $j = 2, \dots, 7$ are discussed in Section 5. Except for the first term, the objective function J exhibits only quadratic terms and J is positive. Note that the displacements $\boldsymbol{u}$ or $\bar{\boldsymbol{u}}$ do not enter explicitly into the objective function. However, since $\bar{\boldsymbol{u}}$ enters into the ODE for $\boldsymbol{q}$ by means of the mass matrix $\boldsymbol{M}$ and the right-hand side $\boldsymbol{F}$, finding an optimal control U implies that vibrations of $\bar{\boldsymbol{u}}$ are damped out as well.

2.6 OCP for 2d crane

We solve the second-order ODE (12) for $\boldsymbol{q}$ as a system of coupled ODEs for $\boldsymbol{q}$ and $\boldsymbol{v} := \dot{\boldsymbol{q}}$, where $\boldsymbol{v}$ is considered as an independent state.

Our OCP reads: Find states $\boldsymbol{q} \in [C^2([0,T])]^2$, $\boldsymbol{v} \in [C^1([0,T])]^2$, $\boldsymbol{u} \in C^0([0,T]; H^1(\Omega; \mathbb{R}^3))$, a control $U \in L^\infty(0,T)$ and a parameter $T \geq T_{\min} > 0$ (without loss of generality $T_{\min}$ arbitrarily small) such that the reduced cost function

$$\mathcal{F}(U,T) = J(\boldsymbol{q}(U,T), \boldsymbol{v}(U,T), U, T) \tag{22}$$

is minimized under the following constraints:

- the ODE system

$$\mathbf{M}(\boldsymbol{q}, \bar{\boldsymbol{u}})\,\dot{\boldsymbol{v}} = \boldsymbol{F}(\boldsymbol{q}, \boldsymbol{v}, \bar{\boldsymbol{u}}, U) \tag{23}$$

$$\dot{\boldsymbol{q}} = \boldsymbol{v} \qquad\qquad \text{in } (0,T), \tag{24}$$

resulting from (12) having employed the mean values $\bar{\boldsymbol{u}}$ in (13), yielding

$$\mathbf{M}(\boldsymbol{q}, \bar{\boldsymbol{u}}) := \begin{pmatrix} m(1+2\bar{u}_1) & (1+\bar{u}_1)\cos q_2 + \bar{u}_2 \sin q_2 \\ (1+\bar{u}_1)\cos q_2 + \bar{u}_2 \sin q_2 & l \end{pmatrix}, \tag{25}$$

and analgously in (14) and in (10), yielding

$$\boldsymbol{F}(\boldsymbol{q}, \boldsymbol{v}, \bar{\boldsymbol{u}}, U) := \begin{pmatrix} -m\,\bar{u}_3 v_1^2 + ((1+\bar{u}_1)\sin q_2 - \bar{u}_2 \cos q_2)v_2^2 + U \\ -(\bar{u}_3 \cos q_2 + \bar{u}_4 \sin q_2)v_1^2 - g_e \sin q_2 \end{pmatrix}, \tag{26}$$

together with initial conditions

$$\boldsymbol{q}(0) = \boldsymbol{q}^0, \quad \boldsymbol{v}(0) = \boldsymbol{0}, \tag{27}$$

- the PDE (15 – 17)

$$-\operatorname{div}\sigma(\boldsymbol{u}) = \boldsymbol{H} \qquad \text{in}\,\Omega\times[0,T], \tag{28}$$

$$\boldsymbol{u} = \boldsymbol{0} \qquad \text{on}\ \Gamma_{\mathrm{D}}\times[0,T], \tag{29}$$

$$-\sigma(\boldsymbol{u}).\boldsymbol{n} = \boldsymbol{R}(\boldsymbol{q},\bar{\boldsymbol{u}},U) \qquad \text{on}\ \Gamma_{\mathrm{N}}\times[0,T], \tag{30}$$

with the gravity term $\boldsymbol{H}$ from (18) and with averaging over Γ_{C} in (19) yielding

$$\boldsymbol{R}(\boldsymbol{q},\bar{\boldsymbol{u}},U) \;:=\; \begin{cases} \frac{1}{|\Gamma_{\mathrm{C}}|}\boldsymbol{R}_0(\boldsymbol{q},\bar{\boldsymbol{u}},U); & \boldsymbol{x}\in\Gamma_{\mathrm{C}}(q_1), \\ 0 \qquad ; & \text{otherwise}, \end{cases} \tag{31}$$

where

$$\boldsymbol{R}_0(\boldsymbol{q},\bar{\boldsymbol{u}},U) \;:=\; \begin{pmatrix} (1-\bar{u}_6)(U+g\sin q_2)-\bar{u}_2 g(\cos q_2+\eta) \\ 0 \\ \bar{u}_5(U+g\sin q_2)-(1-\bar{u}_1)g(\cos q_2+\eta) \end{pmatrix}, \tag{32}$$

- the state equation (20) for $\bar{\boldsymbol{u}}$, and
- the control constraints

$$U_{\min} \le U(t) \le U_{\max} \quad \text{point-wise for all } t\in[0,T]. \tag{33}$$

In this paper we consider a smoothed version $\chi^{\varepsilon}_{\Gamma_{\mathrm{C}}}$, see (47), of the characteristic function $\chi_{\Gamma_{\mathrm{C}}}$ (being 1 on the set Γ_{C} and 0 otherwise) entering in (31). The solvability of this coupled problem for a given control is discussed in Section 3. We check that in case of neglected elastic deformations we recover in (23 – 27) and (33) the standard ODE-problem for an inelastic overhead gantry crane [4,5]. As in the model in [4,5] we consider no damping term and, furthermore, neglect the moment of inertia.

The terminal conditions $\boldsymbol{q}(T)=\boldsymbol{q}^f$ and $\boldsymbol{v}(T)=\boldsymbol{0}$ are realized approximately by the penalty terms corresponding to the weights ν_i, $i=4,\ldots 7$, within the objective function (21). We could also require state constraints, as $(b_1+l_{\mathrm{T}})/2 < q_1 < l_2-(b_1+l_{\mathrm{T}})/2$, modelling that neither the trolley touches the crane tower nor that it jumps off the rails at the end of the crane beam, and $|q_2| < q_2^{max}$, reflecting that the pendant cord may not be tight for angles larger than a certain q_2^{max}. State constraints are neglected as in [4,5], since we see in our numerical experiments that for typical initial data a control is found such that the state constraints are safely guaranteed.

3. Local-in-time existence and uniqueness of the coupled states

3.1. *Smoothed computational domain*

Let us consider a sufficiently small $T>0$. We emphasize that, without loss of generality, Ω has a C^3 boundary, that could be obtained by smoothing out the corners, since corners are not relevant for the trolley-load system. We introduce $\Omega_{\tilde{\varepsilon}} = \{\boldsymbol{x}\in\Omega\,|\,x_1 > b_1/2+\tilde{\varepsilon}\}$ where $\tilde{\varepsilon}\in(0,\min\{(l_2-b_1)/2, b_2/2, b_3/2\})$ may be chosen arbitrarily small. In order to avoid technical regularity issues, we focus in this section on

$$\Omega^* := \Omega_{\tilde{\varepsilon}} \cup \mathcal{B}_{\tilde{\varepsilon}}(\{\boldsymbol{x}\in\Omega\,|\,x_1 = b_1/2+\tilde{\varepsilon}, |x_2|\le b_2/2-\tilde{\varepsilon}, l_1-b_3+\tilde{\varepsilon}\le x_3\le l_1-\tilde{\varepsilon}\})$$

where $\mathcal{B}_{\tilde{\varepsilon}}(S) = \cup_{\boldsymbol{x}\in S} B_{\tilde{\varepsilon}}(\boldsymbol{x})$ is a usual $\tilde{\varepsilon}$-neighborhood of a closed set S. Ω^* is constructed smoothly such that the contact lines $\overline{\Gamma_{\mathrm{D}}}\cap\overline{\Gamma_{\mathrm{N}}}$ between Dirichlet and Neumann boundary, where singularities might occur (compare [24, Sect. 3]), are taken out of Ω. Let $\boldsymbol{u}^*$ denote for the moment the solution on Ω^*. On $B^* := \partial\Omega^*\backslash\partial\Omega$ we prescribe $a^*\boldsymbol{u}+(1-a^*)\sigma(\boldsymbol{u}^*).\boldsymbol{n}=\boldsymbol{0}$ as boundary condition, where a^* is a smooth spatial function on B^* such that $a^*(\boldsymbol{x})=1$ for $x_1=b_1/2$ and $a^*(\boldsymbol{x})=0$ for $x_3=l_1-b_3$. According to [24, Sect. 3] and the references therein, $\boldsymbol{u}$ can be decomposed into a regular and singular parts. For the intersection $\{\boldsymbol{x} \in \Omega \,|\, x_1 = b_1/2, |x_2| = b_2/2 \,\text{or}\, |x_3 - l_1 - b_3/2| = b_3/2\}$ between homogeneous Dirchlet and Neumann b.c., the usual regularity results hold for the regular part of $\boldsymbol{u}$, while the singular parts (smooth except for the singularity) are bounded by $\tilde{r}^{-\tilde{m}}$, $\tilde{r}$ being the distance to the respective singularity, where $0 < \tilde{m} < 3$ depends on the interior angle and the Lamé parameters [25, Sect. 2]. Thus the norm of the whole singular part is arbitrarily small on Ω^* for suitable $\tilde{\varepsilon}$. Furthermore, our numerical simulations do not exhibit a singular behaviour in the corners or at the edges or in a neighborhood of these, thereby reconfirming that the corner and edge smoothing is only an auxiliary construction in order to avoid technical details in the proof concerning the interplay between crane and trolley.

Furthermore, we assume that the prescribed control $U \in L^\infty(0, T)$. For suitable data U, ρ, η, and g and sufficiently small T we may assume $\|\nabla \boldsymbol{u}\|_{L^2(\Omega)}, |\bar{u}_j| \ll 1$, $j = 1, 2, 5, 6$, for all $t \in [0, T]$, meaning that the assumption of small displacement gradients is justified (see also our numerics in Sect. 5). The latter implies *inter alia* $1 - 2\bar{u}_1 > 0$. As usual we denote, e.g. $H^k = W^{k,2}$ or $L^pH^k = L^p(0, T; H^k(\Omega^*; \mathbb{R}^3))$ for the Bochner space with the norm $\|\boldsymbol{f}\|^p_{L^pH^k} = \int_0^T \|\boldsymbol{f}(t)\|^p_{H^k(\Omega^*;\mathbb{R}^3)} \mathrm{d}t$.

3.2. *Standard results for uncoupled differential equations*

We summarize standard results for the ODE system and the PDE considered as stand alone equations. For given $\bar{u}_j \in C^0([0, T])$, $j = 1, 2, 3, 4$, $U \in C^0([0, T])$ there exists a unique solution $\boldsymbol{q} \in [C^2([0, T])]^2$ of the ODE system for sufficiently small T by the theorem of Picard-Lindelöf, since $\mathbf{M}^{-\mathbf{1}}\boldsymbol{F}$ is Lipschitz in $\boldsymbol{q}$ and $\boldsymbol{v} = \dot{\boldsymbol{q}}$. For given $\bar{u}_j \in L^\infty(0, T)$, $j = 1, 2, 3, 4$, $U \in L^\infty(0, T)$ there exists a unique solution $\boldsymbol{q} \in [W^{2,\infty}(0, T)]^2$ (that is, C^2 for almost all t) of the ODE system for sufficiently small T. We have the standard estimate for ODE

$$\|q_j\|_{H^2(0,T)} \leq C\|(\mathbf{M}^{-1}\boldsymbol{F})(\boldsymbol{q}, \dot{\boldsymbol{q}}, \bar{\boldsymbol{u}}, U)\|_{L^2(0,T)}.$$

Let $\lambda, \mu > 0$ in (6). For given $\boldsymbol{q} \in [W^{\theta,\kappa}(0, T)]^2$, $U \in L^\kappa(0, T)$ and $\bar{u}_j \in L^\kappa(0, T)$, $j = 1, 2, 5, 6$, there exists a unique solution $\boldsymbol{u} \in L^\kappa(0, T; W^{1,p}(\Omega))$ of the PDE problem for $0 \leq \theta < \infty$, $6/5 \leq p < \infty$, and $1 \leq \kappa \leq \infty$ [23, Sect.6.3], where the time regularity carries over from the data. For almost all t, there holds the estimate

$$\|\boldsymbol{u}(\cdot, t)\|_{H^1(\Omega)} \leq c(\|\boldsymbol{R}^\varepsilon(\boldsymbol{q}, \bar{\boldsymbol{u}}, U)\|_{L^2(\Gamma_\mathrm{N})} + 1), \tag{34}$$

where

$$\boldsymbol{R}^\varepsilon(\boldsymbol{q}, \bar{\boldsymbol{u}}, U) := \frac{1}{|\Gamma_\mathrm{C}|}\boldsymbol{R}_0(\boldsymbol{q}, \bar{\boldsymbol{u}}, U) \;\; \chi^\varepsilon_{\Gamma_\mathrm{C}}(\boldsymbol{x}) \tag{35}$$

is a smoothed version of $\boldsymbol{R}$ defined in (31), $\varepsilon > 0$ being another sufficiently small smoothing parameter. Note that the estimate (34) holds on Ω (as well as for $\boldsymbol{u}^*$ on Ω^* with the corresponding boundary conditions) and, therefore, $\boldsymbol{u} \in H^{1/2}(B^*)$. Now testing on Ω^* (instead of Ω) yields an additional integral term $\|a^*\boldsymbol{u}\|_{L^2(B^*)}$ in the estimates for $\boldsymbol{u}^*$.

We need the regularity result from [23, Sect. 6.3] combined with [24, Sect. 3] that for $\partial\Omega^* \in C^3$ we have $\boldsymbol{u}^* \in W^{3,p}(\Omega^*)$ for any $p \geq 6/5$, since $\boldsymbol{R}^\varepsilon$ is $W^{1-1/r,r}(\Gamma_\mathrm{N})$, $6/5 \leq r \leq \infty$. Thus together with the Sobolev embedding for $W^{3,p}(\Omega^*)$ for $p > 3/(1-\delta)$, $0 < \delta < 1$ [26, Th. 10.13, 2)], we get

$$\boldsymbol{u}^* \in L^\kappa(0, T; W^{3,p}(\Omega^*)) \mapsto L^\kappa(0, T; C^{2,\delta}(\overline{\Omega^*})). \tag{36}$$

We consider only Ω^* in the following and, for ease of notation, we write again $\boldsymbol{u}$ instead of $\boldsymbol{u}^*$. (36) shows that the derivatives entering in the definition (20) of $\bar{\boldsymbol{u}}$ are well-defined. We remark that using mean values is crucial to obtain the stated regularity for $\boldsymbol{u}$ and $\bar{\boldsymbol{u}}$.

3.3. *Local-in-time existence for coupled differential equations*

We cannot expect an existence and uniqueness result for an arbitrary right hand side $\mathbf{M}^{-1}\boldsymbol{F}$ in the ODE (23). We will need the following local Lipschitz estimate.

Lemma 3.1 (Estimate for the right-hand-side of the ODE) *Within our approximation* $\|\nabla \boldsymbol{u}\| \ll 1$, *there holds for two different state pairs* $(\boldsymbol{q}^{(i)}, \bar{\boldsymbol{u}}^{(i)})$, $i = 1, 2$,

$$\begin{aligned} &\|(\mathbf{M}^{-1}\boldsymbol{F})(\boldsymbol{q}^{(1)}, \dot{\boldsymbol{q}}^{(1)}, \bar{\boldsymbol{u}}^{(1)}, U) - (\mathbf{M}^{-1}\boldsymbol{F})(\boldsymbol{q}^{(2)}, \dot{\boldsymbol{q}}^{(2)}, \bar{\boldsymbol{u}}^{(2)}, U)\|_{L^2(0,T)} \\ &\le \mathrm{Const}(k, K)\Big(\|\boldsymbol{q}^{(1)} - \boldsymbol{q}^{(2)}\|_{H^1(0,T)} + \|\bar{\boldsymbol{u}}^{(1)} - \bar{\boldsymbol{u}}^{(2)}\|_{L^2(0,T)}\Big), \end{aligned} \tag{37}$$

when $\sup_j \| \dot{q}_j^{(i)}\|_{L^\infty(0,T)} \le k$ *and* $\sup_l \|\bar{u}_l^{(i)}\|_{L^\infty(0,T)} \le K$ *for all* $i = 1, 2$.

Proof. The proof relies on computing $\mathbf{M}^{-1}$ explicitly by Cramer's rule. For further details, see [27, App. A]. Then we exploit that for the quadratic terms

$$\begin{aligned} \|(\dot{q}_j^{(1)})^2 - (\dot{q}_j^{(2)})^2\|_{L^2(0,T)} &\le \|(\dot{q}_j^{(1)} + \dot{q}_j^{(2)})(\dot{q}_j^{(1)} - \dot{q}_j^{(2)})\|_{L^2(0,T)} \\ &\le 2k\|\dot{q}_j^{(1)} - \dot{q}_j^{(2)}\|_{L^2(0,T)} \end{aligned}$$

that for the mixed terms

$$\begin{aligned} \|\bar{u}_l^{(1)}\dot{q}_j^{(1)} - \bar{u}_l^{(2)}\dot{q}_j^{(2)}\|_{L^2(0,T)} &\le \|\bar{u}_l^{(1)}\dot{q}_j^{(1)} - \bar{u}_l^{(1)}\dot{q}_j^{(2)} + \bar{u}_l^{(1)}\dot{q}_j^{(2)} - \bar{u}_l^{(2)}\dot{q}_j^{(2)}\|_{L^2(0,T)} \\ &\le K\|\dot{q}_j^{(1)} - \dot{q}_j^{(2)}\|_{L^2(0,T)} + k\, \| \,\bar{u}_l^{(1)} - \bar{u}_l^{(2)}\|_{L^2(0,T)}, \end{aligned}$$

and that $|\sin q_1|, |\cos q_2| \le 1$ and $\sin$, $\cos$ are Lipschitz, where $j = 1, 2$, $l = 1, 2, 3, 4$. □

Theorem 3.2 (Local-in-time well-posedness of the dynamics) *Let* Ω^* *be a* C^3 *domain and let* $\lambda, \mu > 0$ *be the Lamé parameters. Suppose that* $U \in L^\infty(0, \infty)$ *and that for* $\mathbf{M}^{-1}\boldsymbol{F}$ *the estimate (37) holds. Then for each pair* (k, K) *of positive numbers, there exists* $T > 0$ *and an area of the contact surface* $|\Gamma_C| > 0$, *such that the coupled ODE-PDE problem (23 – 32), with (20) and the smoothing (35), has a unique solution* $\boldsymbol{q} \in [H^2(0, T)]^2$ *and* $\boldsymbol{u} \in L^\infty(0, T; W^{3,p}(\Omega^*))$ *for any* $p > 3$.

Our strategy is inspired by Algorithm 4.1: We solve alternately for $\boldsymbol{u}$ and the mean values $\bar{\boldsymbol{u}}$, then the result is used for the right-hand side of the ODE. The ODE solution $\boldsymbol{q}$ is inserted into the PDE and a fixed point iteration is invoked. The idea of proof, relying on the Banach fixed point theorem and the estimate (38) yielding a factor proportional to $\sqrt{T}$ in the contraction constant, has been described for example by Niethammer [28] for a coupled ODE-Laplace PDE problem. A similar proof as needed for our problem is given in [29] for a coupled problem consisting of a single ODE for a free boundary, a quasilinear diffusion PDE, and the PDE of linear elasticity. Both cited proofs consider free boundary problems with a time-dependent domain and require beforehand a transformation to a fixed domain that is not needed here.

Proof. We would like to apply the Banach fixed point theorem in the space

$$\mathcal{M} = \mathcal{M}_T^k \times \mathcal{M}_T^K,$$

where

$$\mathcal{M}_T^k := \left\{ \boldsymbol{q} \in [H^2(0, T)]^2 \mid \sup_j \|\dot{q}_j\|_{L^\infty(0,T)} \le k \right\},$$

$$\mathcal{M}_T^K := \left\{ \boldsymbol{u} \in L^\infty H^2 \mid \sup_l \|\bar{u}_l\|_{L^\infty(0,T)} \le K \right\},$$

to the map

$$G : \mathcal{M} \to [H^2(0,T)]^2 \times L^\infty H^2,$$

$$(\boldsymbol{q}, \boldsymbol{u}) \mapsto (\boldsymbol{q}^+, \boldsymbol{u}^+) := (\mathcal{L}_1(\mathbf{M}^{-1}\boldsymbol{F})(\boldsymbol{q}, \dot{\boldsymbol{q}}, \bar{\boldsymbol{u}}^+, U),\ \mathcal{L}_2\boldsymbol{R}^\varepsilon(\boldsymbol{q}, \bar{\boldsymbol{u}}, U)),$$

where the operators $\mathcal{L}_1 : L^2(0,T) \to H^2(0,T)$ and $\mathcal{L}_2 : W^{2,\infty}(\Gamma_N) \to H^2(\Omega)$ map onto the solution of the ODE with right-hand side $\mathbf{M}^{-1}\boldsymbol{F}$ and onto the solution of the elasticity problem with right-hand side $\boldsymbol{R}^\varepsilon$ on Γ_N, respectively.

I. Strict contraction:
We consider two pairs of time trajectories $(\boldsymbol{q}^{(1)}, \boldsymbol{u}^{(1)})$ and $(\boldsymbol{q}^{(2)}, \boldsymbol{u}^{(2)})$. Define $\boldsymbol{q}_\Delta = \boldsymbol{q}^{(1)} - \boldsymbol{q}^{(2)}$ and $\boldsymbol{u}_\Delta = \boldsymbol{u}^{(1)} - \boldsymbol{u}^{(2)}$. The main ingredient is the following Poincaré inequality:

$$\|q_j - q_j^0\|_{L^\infty(0,T)} = \left\| \int_0^T \dot{q}_j \, \mathrm{d}t \right\|_{L^\infty(0,T)} \le \sqrt{T} \|\dot{q}_j\|_{L^2(0,T)} \tag{38}$$

that follows by applying Hölder's inequality. Thus

$$\|q_j - q_j^0\|_{L^2(0,T)} = \sqrt{\int_0^T (q_j - q_j^0)^2 \, \mathrm{d}t} \le \sqrt{T} \|q_j - q_j^0\|_{L^\infty(0,T)},$$

$$\|q_j - q_j^0\|_{H^1(0,T)} \le \sqrt{T^2+1} \|\dot{q}_j\|_{L^2(0,T)}.$$

This procedure is repeated:

$$\|\dot{q}_j\|_{L^2(0,T)} = \sqrt{\int_0^T \dot{q}_j^2 \, \mathrm{d}t} \le \sqrt{T} \|\dot{q}_j\|_{L^\infty(0,T)} \le T \|\ddot{q}_j\|_{L^2(0,T)}.$$

Thus

$$\begin{aligned} \|q_j - q_j^0\|_{H^1(0,T)} &\le T\sqrt{T^2+1} \|\ddot{q}_j\|_{L^2(0,T)}, \\ \|q_j - q_j^0\|_{H^2(0,T)} &\le \sqrt{T^4+T^2+1} \|\ddot{q}_j\|_{L^2(0,T)}. \end{aligned} \tag{39}$$

We write $\boldsymbol{q}_\Delta^+ = \boldsymbol{q}^{(1),+} - \boldsymbol{q}^{(2),+}$ and $\boldsymbol{u}_\Delta^+ = \boldsymbol{u}^{(1),+} - \boldsymbol{u}^{(2),+}$. Applying this to the map $\mathcal{G}$ yields

$$\|q_{\Delta,j}^+\|_{H^2(0,T)} \le \sqrt{T^4+T^2+1} \times$$

$$\times \|(\mathbf{M}^{-1}\boldsymbol{F})(\boldsymbol{q}^{(1)}, \dot{\boldsymbol{q}}^{(1)}, \bar{\boldsymbol{u}}^{(1),+}, U) - (\mathbf{M}^{-1}\boldsymbol{F})(\boldsymbol{q}^{(2)}, \dot{\boldsymbol{q}}^{(2)}, \bar{\boldsymbol{u}}^{(2),+}, U)\|_{L^2(0,T)},$$

and together with Lemma 3.1 we have

$$\|q_{\Delta,j}^+\|_{H^2(0,T)} \le \sqrt{T^4+T^2+1}\ \mathrm{Const}(k,K)\ \left(\|\ \boldsymbol{q}_\Delta\|_{H^1(0,T)} + \|\bar{\boldsymbol{u}}_\Delta^+\|_{L^2(0,T)}\right). \tag{40}$$

For the PDE we estimate for fixed ε and $\tilde{\varepsilon}$

$$\|\boldsymbol{R}^\varepsilon(\boldsymbol{q}^{(1)}, \bar{\boldsymbol{u}}^{(1)}, U) - \boldsymbol{R}^\varepsilon(\boldsymbol{q}^{(2)}, \bar{\boldsymbol{u}}^{(2)}, U)\|_{L^2(\Gamma_N)}$$

$$\le \mathrm{const}\,(k,K)\,|\Gamma_C|^{\frac{1}{2}} \left(\sum_{j=1,2,5,6} |\bar{u}_j^{(1)} - \bar{\mu}_j^{(2)}| + |\boldsymbol{q}^{(1)} - \boldsymbol{q}^{(2)}| \right)$$

where, for instance, $\bar{\mu}_1^{(2)} = |\Gamma_{\mathrm{C}}|^{-1}\int_{\Gamma_{\mathrm{C}}(q_1^{(1)})} \partial_1 u_1{}^{(2)}\,\mathrm{d}\boldsymbol{x}$. The other $\bar{\mu}_j$ are defined analogously. By using $\bar{u}_1 = |\Gamma_{\mathrm{C}}|^{-1}\int_{\Gamma_{\mathrm{C}}} \partial_1 u_1\,\mathrm{d}\boldsymbol{x}$, the Hölder inequality and the trace theorem [30, Th. 5.22] in 3d

$$\begin{aligned}
|\Gamma_{\mathrm{C}}(q_1^{(1)})|^{\frac{1}{2}}|\bar{u}_1^{(1)} - \bar{\mu}_1^{(2)}| &\le |\Gamma_{\mathrm{C}}|^{-\frac{1}{2}}\int_{\Gamma_{\mathrm{C}}(q_1^{(1)})} |\partial_1 u_1^{(1)} - \partial_1 u_1^{(2)}|\,\mathrm{d}\boldsymbol{x} \\
&\le |\Gamma_{\mathrm{C}}|^{-\frac{1}{2}+\frac{5}{6}}\,\|\nabla\boldsymbol{u}^{(1)} - \nabla\boldsymbol{u}^{(2)}\|_{L^6(\Gamma_{\mathrm{C}}(q_1^{(1)}))} \\
&\le |\Gamma_{\mathrm{C}}|^{\frac{1}{3}} C(\Omega^*)\,\|\boldsymbol{u}^{(1)} - \boldsymbol{u}^{(2)}\|_{H^2(\Omega^*;\mathbb{R}^3)}, \qquad (41)
\end{aligned}$$

where the constant does not depend on $|\Gamma_{\mathrm{C}}|$. Analogously the proof follows for $j = 2, 5, 6$. Using the estimate corresponding to (34) for the difference $\boldsymbol{u}_\Delta^+$ and for $\partial_i \boldsymbol{u}_\Delta^+$, $i = 1, 2, 3$, we end up with

$$\|\boldsymbol{u}_\Delta^+\|_{L^\infty H^2} \le \tilde{C}(k, K, \Omega^*)\Big(|\Gamma_{\mathrm{C}}|^{\frac{1}{3}}\|\boldsymbol{u}_\Delta\|_{L^\infty H^2} + \|\boldsymbol{q}_\Delta\|_{L^\infty(0,T)}\Big). \qquad (42)$$

We use the regularity result (36) for $\boldsymbol{u}$ in order to get that $\bar{u}_3, \bar{u}_4$ are well-defined, then we combine the estimates (38) and (42), yielding

$$\|\boldsymbol{u}_\Delta^+\|_{L^\infty H^2} \le \hat{C}(k, K, \Omega^*)\Big(|\Gamma_{\mathrm{C}}|^{\frac{1}{3}}\|\boldsymbol{u}_\Delta\|_{L^\infty H^2} + \sqrt{T}\,\|\boldsymbol{q}_\Delta\|_{H^2(0,T)}\Big).$$

This estimate is inserted into (40) and we use (39) and (41), yielding

$$\begin{aligned}
\|\boldsymbol{q}_\Delta^+\|_{H^2(0,T)} \le \sqrt{T^4 + T^2 + 1}\,\breve{C}(k, K, \Omega^*) \;\times \\
\times \Big(\Big(T\sqrt{T^2+1} + |\Gamma_{\mathrm{C}}|^{-\frac{1}{6}}\sqrt{T}\Big)\,\|\boldsymbol{q}_\Delta\|_{H^2(0,T)} + |\Gamma_{\mathrm{C}}|^{\frac{1}{6}}\|\boldsymbol{u}_\Delta\|_{L^\infty H^2}\Big).
\end{aligned}$$

Now choosing $|\Gamma_{\mathrm{C}}|$ and T (depending on $|\Gamma_{\mathrm{C}}|$) sufficiently small guarantees that $\mathcal{G}$ is a strict contraction. From the strict contraction we also get directly that the local-in-time solution is unique.

II. Self-mapping:

In order to check that $\mathcal{G}$ maps $\mathcal{M}$ into itself, we use that the estimates from Part I carry over. So the self-mapping property follows analogously for sufficiently small T and $|\Gamma_{\mathrm{C}}|$.

This shows that there exists a unique solution **(*q*, *u*)** with the stated regularity. From the estimates in Part I of the proof, we see that the solution depends continuously on the data. □

Since we do not know from the last lemma whether the time T guaranteed by the fixed-point method is larger then the total time obtained by the optimal control, global existence of a solution for our problem is not guaranteed and cannot be expected for arbitrary data and control. However, when minimizing our objective function we may hope that the control U is determined in such a way that no blow up for q_1 or q_2 may happen in finite time.

4. Discretized problem and optimal control method

Due to the complicated model we follow here a FDTO approach for solving our OCP. Furthermore we work with a sensitivity-based approach since the adjoint optimality system cannot be derived by standard methods. The reason for this is the averaging over $\boldsymbol{u}$ that appears in the integral equation (20).

Since in our numerical example we consider $\nu_1 > 0$ in (21), we perform a time transformation onto a fixed time interval, mapping $t \in [0, T]$ to $\tau \in [0, 1]$. This provides an easy way to determine the control parameter T in the sequel. According to the time transformation, time derivatives have to be scaled with the factor T. We consider from now on time-transformed functions but denote them again with the same symbol in order to keep the notation simple.

4.1. *Time integration*

The PDE problem for $\boldsymbol{u}$ and $\bar{\boldsymbol{u}}$ is solved by a finite element method (FEM). In order to avoid locking effects [31, Ch. VI, §3] we use Lagrange $\mathbb{P}_2$-elements and a vertically refined layered mesh. The spatially discretized problem may be considered as a semi-explicit DAE of index 1 with $\boldsymbol{u}$ being the algebraic variable. We solve in time by means of the explicit Heun method with respect to $\boldsymbol{q} = (x_{\mathrm{T}}, \alpha)^{\top}$ and $\boldsymbol{v} = \dot{\boldsymbol{q}}$. This second-order method allows to solve the 2-dimensional ODE system accurately enough.

By dividing the time interval $[0, 1]$ into $N \in \mathbb{N}$ intervals of length $h := 1/N$, we define the time steps $\tau_k := kh$, $k = 0, \dots, N$. We abbreviate $\boldsymbol{q}_{(k)} := \boldsymbol{q}(\tau_k)$, $\boldsymbol{v}_{(k)} := \boldsymbol{v}(\tau_k)$, $\bar{\boldsymbol{u}}_{(k)} := \bar{\boldsymbol{u}}(\tau_k)$, $\boldsymbol{u}_{(k)}(\cdot) := \boldsymbol{u}(\cdot, \tau_k)$, $U_{(k)} = U(\tau_k)$, and for the predictor step of the Heun method we introduce the values $\tilde{\boldsymbol{q}}_{(k)}$ and $\tilde{\boldsymbol{v}}_{(k)}$, $k = 1, \dots, N$. Furthermore we write, for instance,

$$\mathbf{M}_{(k)} = \mathbf{M}(\boldsymbol{q}_{(k)}, \bar{\boldsymbol{u}}_{(k)}), \quad \boldsymbol{F}_{(k)} = \boldsymbol{F}(\boldsymbol{q}_{(k)}, \boldsymbol{v}_{(k)}, \bar{\boldsymbol{u}}_{(k)}, U_{(k)}),$$

$$\tilde{\mathbf{M}}_{(k+1)} = \mathbf{M}(\tilde{\boldsymbol{q}}_{(k+1)}, \bar{\boldsymbol{u}}_{(k)}), \quad \tilde{\boldsymbol{F}}_{(k+1)} = \boldsymbol{F}(\tilde{\boldsymbol{q}}_{(k+1)}, \tilde{\boldsymbol{v}}_{(k+1)}, \bar{\boldsymbol{u}}_{(k)}, U_{(k)}),$$

$$\boldsymbol{R}_{(k)} = \boldsymbol{R}^{\varepsilon}(\boldsymbol{q}_{(k)}, \bar{\boldsymbol{u}}_{(k)}, U_{(k)}).$$

The discretized version of the set of admissible controls is

$$\mathcal{U}_{ad} := \{\boldsymbol{V} \in \mathbb{R}^{N+1} \,|\, V_{(j)} \in U_{ad} := [U_{\min}, U_{\max}]\ \forall j = 0, \dots, N\}.$$

Thanks to Theorem 3.2 we may expect that the following algorithm, including a fixed-point iteration for $\boldsymbol{u}$ and $\bar{\boldsymbol{u}}$, works out well.

Algorithm 4.1 (Simulation procedure with Heun method in time)

(0) Init: Let $T = T^{(0)}$ be given, $k := 0$, $\bar{\boldsymbol{u}}_{-1} \equiv 0$. Let control input $U_{(\cdot)} \in \mathcal{U}_{ad}$ and initial values $\boldsymbol{q}_0 = \boldsymbol{q}^0$ and $\boldsymbol{v}_0 = \boldsymbol{v}^0$ be given.

(1) (0) $\bar{\boldsymbol{u}}_{(k)} = \bar{\boldsymbol{u}}_{(k-1)}$.

(i) At time τ_k solve

$$\begin{aligned} -\operatorname{div} \sigma(\boldsymbol{u}_{(k)}) &= \boldsymbol{H} && \text{in } \Omega \times \{\tau_k\}, \\ \boldsymbol{u}_{(k)} &= \boldsymbol{0} && \text{on } \Gamma_{\mathrm{D}} \times \{\tau_k\}, \\ -(\boldsymbol{u}_{(k)}).\boldsymbol{n} &= \boldsymbol{R}_{(k)} && \text{on } \Gamma_{\mathrm{N}} \times \{\tau_k\}. \end{aligned}$$

(ii) $\bar{\boldsymbol{u}}^{old}_{(k)} := \bar{\boldsymbol{u}}_{(k)}$. Compute $\bar{\boldsymbol{u}}_{(k)}$.

(iii) If $\|\bar{\boldsymbol{u}}_{(k)} - \bar{\boldsymbol{u}}^{old}_{(k)}\| > err_0$ for a suitable norm and given error tolerance err_0 go to (i).

(2) Set

$$\mathbf{M}_{(k)} \tilde{\boldsymbol{v}}_{(k+1)} = \mathbf{M}_{(k)} \boldsymbol{v}_{(k)} + hT\boldsymbol{F}_{(k)}, \tag{43}$$

$$\tilde{\boldsymbol{q}}_{(k+1)} = \boldsymbol{q}_{(k)} + hT\boldsymbol{v}_{(k)}, \tag{44}$$

$$\tilde{\mathbf{M}}_{(k+1)} \boldsymbol{v}_{(k+1)} = \tilde{\mathbf{M}}_{(k+1)} \frac{1}{2} (\boldsymbol{v}_{(k)} + \tilde{\boldsymbol{v}}_{(k+1)}) + \frac{hT}{2} \tilde{\boldsymbol{F}}_{(k+1)}, \tag{45}$$

$$\boldsymbol{q}_{(k+1)} = \frac{1}{2} \left(\boldsymbol{q}_{(k)} + \tilde{\boldsymbol{q}}_{(k+1)} \right) + \frac{hT}{2} \tilde{\boldsymbol{v}}_{(k+1)}. \tag{46}$$

(3) Set $\tau_{k+1} = \tau_k + h$, $k := k + 1$.
If $\tau_k < 1$ go to (1), otherwise Stop.

In order to avoid technical issues with the size of finite elements, we replace the characteristic function χ_{Γ_C}, appearing in the definition (31) of $\boldsymbol{R}$, by an approximation with a version $\tilde{\chi}^\varepsilon_{\Gamma_\mathrm{C}}$ smoothed in x_1-direction, where $\varepsilon > 0$ is a small, but fixed parameter:

$$\tilde{\chi}^\varepsilon_{\Gamma_\mathrm{C}}(\boldsymbol{x}) := 1 - \tanh^2\big(((x_1 - q_1)/\delta_1)^L\big) \tag{47}$$

with $L \in \mathbb{N}$, $\delta_1 > 0$ such that

$$|\chi_{\Gamma_\mathrm{C}}(\boldsymbol{x}) - \tilde{\chi}^\varepsilon_{\Gamma_\mathrm{C}}(\boldsymbol{x})| \leq \begin{cases} 1/2; & \text{for } |x_1 - q_1 \mp l_\mathrm{T}/2| < \varepsilon, \\ \varepsilon \quad ; & \text{otherwise.} \end{cases}$$

Abbreviating $A_1 := \ln(\operatorname{atanh}(\sqrt{1-\varepsilon}))$, $A_2 := \ln(\operatorname{atanh}(\sqrt{\varepsilon}))$ and $B_{1/2} = \ln(l_\mathrm{T}/2 \pm \varepsilon)$, we have $L = \lceil (A_1 - A_2)/(2(B_1 - B_2)) + 1 \rceil$ ($\lceil\cdot\rceil$ denoting the ceiling function) and $\delta_1 = \exp((A_1B_2 - A_2B_1)/(A_1 - A_2))$. For our numerics we work with $\varepsilon = 0.1$, yielding $L = 4$ and $\delta_1 \approx 0.314$, whereas we do not consider any corner and edge smoothing for the simulations.

For small $\eta = m_\mathrm{T}/m_\mathrm{L}$, which is realistic in applications, the matrix $\boldsymbol{M}$ is ill-conditioned. But this issue may be overcome by a left preconditioning of the equations (43) and (45), respectively, with the matrix $\mathbf{M}(\boldsymbol{q} = \boldsymbol{q}_{pc}, \bar{\boldsymbol{u}} = \bar{\boldsymbol{u}}_{pc})^{-1}$, where, for example, $\boldsymbol{q}_{pc} = (x^0_\mathrm{T}, -1/(2\pi))^\top$ and $\bar{\boldsymbol{u}}_{pc} = (0, -0.1, 0, 0, 0, 0)^\top$.

We remark that in Algorithm 4.1, Step (1)(o), we could set alternatively $\bar{\boldsymbol{u}}_{(k)} \equiv 0$, but the choice above turns out to yield fewer iterations in Step (1).

4.2. *Projected gradient method with BFGS update*

We consider a projected gradient method [11], Algorithm 2.2] using a sensitivity-based approach (cf [14., p. 58] within the context of Banach spaces). We prefer to project onto the set of admissible controls within the line search, instead of projecting first and then performing a line search. We discretize the integrals appearing in the reduced objective function by the trapezoidal rule, according to our second-order time integration. Then we calculate a discretized reduced gradient (cf [17., Ch. 4] for a full discretization approach).

Note that due to our modelling ansatz $\boldsymbol{u}$ and $\bar{\boldsymbol{u}}$ do not enter the reduced objective function. Thus our OCP comes down to finding a time-optimal parameter T and control U of a discretized OCP for a DAE with index 1 that may be solved with respect to $\boldsymbol{u}$. It is an open question whether for this class of problems we may expect the existence of optimal controls (U, T). The objective exhibits a quadratic term in U and additional nonlinear terms with respect to T and U.

We use the notation $\boldsymbol{q}^{(n)}_{(k)} = \boldsymbol{q}(U^{(n)}, T^{(n)})(t_k)$, $\mathcal{F}^{(n)} = \mathcal{F}(U^{(n)}, T^{(n)})$ and $J^{(n)} = J(\boldsymbol{q}^{(n)}, \boldsymbol{v}^{(n)}, U^{(n)}, T^{(n)})$ to indicate the dependence on the control $(U^{(n)}, T^{(n)})$ in the optimization iteration $n \in \mathbb{N}$. For brevity we write $S_{(i)} = 1$ for $i = 1, \ldots, N-1$ and $S_{(i)} = 1/2$ for $i = 0$ or N. The discretized time-scaled version of the objective function (21) reads according to the trapezoidal rule

$$\begin{aligned} \mathcal{F}^{(n)} = \nu_1 T^{(n)} + \frac{\nu_2}{2} h T^{(n)} \sum_{k=0}^{N} S_{(i)} \left| v^{(n)}_{(k),2} \right|^2 + \frac{\nu_3}{2} h T^{(n)} \sum_{i=0}^{N} S_{(i)} \left| U^{(n)}_{(i)} \right|^2 \\ + \frac{\nu_4}{2} \left| q^{(n)}_{(N),1} - q^f_1 \right|^2 + \frac{\nu_5}{2} \left| q^{(n)}_{(N),2} \right|^2 + \frac{\nu_6}{2} \left| v^{(n)}_{(N),1} \right|^2 + \frac{\nu_7}{2} \left| v^{(n)}_{(N),2} \right|^2 \end{aligned} \tag{48}$$

As usual we write $\delta_{U^{(n)}_{(i)}}\boldsymbol{q}^{(n)}_{(i)}$ for the sensitivity of the discretized states $\boldsymbol{q}$ with respect to the discretized control U (at time step i and optimization iteration n). Other sensitivities are defined analogously.

Algorithm 4.2 (Projected gradient method with BFGS update, sensitivity-based approach)

(i) Initialize $U^{(0)}_{(k)} \in U_{ad}$, $k = 0, \dots N$, $T^{(0)} \geq T_{\min}$, $\mathbf{H}^{(0)} = 1$, $n = 0$.

(ii) For $k = 0, \dots, N-1$ solve the state equations for $\boldsymbol{q}^{(n)}_{(k+1)}, \boldsymbol{u}^{(n)}_{(k)}, \bar{\boldsymbol{u}}^{(n)}_{(k)}$ by Algorithm 4.1, given a control $U^{(n)}_{(k)}$ and a parameter $T^{(n)}$.

(iii) For $k = 0, \dots, N-1$ compute the sensitivities for $\boldsymbol{u}^{(n)}_{(k)}$, $\bar{\boldsymbol{u}}^{(n)}_{(k)}$ by iterations, then solve the sensitivity equations for $\boldsymbol{q}^{(n)}_{(k+1)}$, $\boldsymbol{v}^{(n)}_{(k+1)}$, the latter are,

$$\begin{aligned}\mathbf{M}^{(n)}_{(k)}\delta_{U^{(n)}_{(i)}}\boldsymbol{v}^{(n)}_{(k+1)} = \boldsymbol{M}^{(n)}_{(k)}\delta_{U^{(n)}_{(i)}}\boldsymbol{v}^{(n)}_{(k)} + hT^{(n)}\delta_{U^{(n)}_{(i)}}\boldsymbol{F}^{(n)}_{(k)} \\ -\delta_{U^{(n)}_{i}}\mathbf{M}^{(n)}_{(k)}\left(\boldsymbol{v}^{(n)}_{(k+1)} - \boldsymbol{v}^{(n)}_{(k)}\right), \quad i = 0, \dots, N,\end{aligned} \tag{49}$$

$$\begin{aligned}\mathbf{M}^{(n)}_{(k)}\delta_{T^{(n)}}\boldsymbol{v}^{(n)}_{(k+1)} = \mathbf{M}^{(n)}_{(k)}\delta_{T^{(n)}}\boldsymbol{v}^{(n)}_{(k)} + h\left(T^{(n)}\delta_{T^{(n)}}\boldsymbol{F}^{(n)}_{(k)} + \boldsymbol{F}^{(n)}_{(k)}\right) \\ - \delta_{T^{(n)}}\mathbf{M}^{(n)}_{(k)}\left(\boldsymbol{v}^{(n)}_{(k+1)} - \boldsymbol{v}^{(n)}_{(k)}\right),\end{aligned} \tag{50}$$

and set

$$\delta_{U^{(n)}_{(i)}}\boldsymbol{q}^{(n)}_{(k+1)} = \delta_{U^{(n)}_{(i)}}\boldsymbol{q}^{(n)}_{(k)} + hT^{(n)}\delta_{U^{(n)}_{(i)}}\boldsymbol{v}^{(n)}_{(k)}, \quad i = 0, \dots, N, \tag{51}$$

$$\delta_{T^{(n)}}\boldsymbol{q}^{(n)}_{(k+1)} = \delta_{T^{(n)}}\boldsymbol{q}^{(n)}_{(k)} + h\left(T^{(n)}\delta_{T^{(n)}}\boldsymbol{v}^{(n)}_{(k)} + \boldsymbol{v}^{(n)}_{(k)}\right). \tag{52}$$

(iv) Determine $\boldsymbol{D}^{(n)}$ as quasi-Newton direction, solving

$$\mathbf{H}^{(n)}\boldsymbol{D}^{(n)} = -\delta_{(U^{(n)},T^{(n)})}\mathcal{F}^{(n)},$$

where the approximated Hessian $\mathbf{H}^{(n)}$ is determined by the modified BFGS update [32] and the anti-gradient of $\mathcal{F}$ with respect to the scalar product in $L^2(0,T) \times \mathbb{R}$ is calculated using the sensitivities:

$$\begin{aligned}\delta_{U^{(n)}_{(i)}}\mathcal{F}^{(n)} = \boldsymbol{\delta}_{\boldsymbol{q}^{(n)}}J^{(n)} \cdot \delta_{U^{(n)}_{(i)}}\boldsymbol{q}^{(n)} + \boldsymbol{\delta}_{\boldsymbol{v}^{(n)}}J^{(n)} \cdot \delta_{U^{(n)}_{(i)}}\boldsymbol{v}^{(n)} + \delta_{U^{(n)}_{(i)}}J^{(n)}, \\ i = 0, \dots, N, \\ \delta_{T^{(n)}}\mathcal{F}^{(n)} = \boldsymbol{\delta}_{\boldsymbol{q}^{(n)}}J^{(n)} \cdot \delta_{T^{(n)}}\boldsymbol{q}^{(n)} + \boldsymbol{\delta}_{\boldsymbol{v}^{(n)}}J^{(n)} \cdot \delta_{T^{(n)}}\boldsymbol{v}^{(n)} + \delta_{T^{(n)}}J^{(n)}.\end{aligned}$$

(v) If a stopping condition is fulfilled, then Stop.

(vi) Determine the step size $s^{(n)}$ by an Armijo line search (see Algorithm 4.3)

$$\mathcal{F}(P_A((U^{(n)}, T^{(n)})^{\mathrm{T}} + s^{(n)}\boldsymbol{D}^{(n)})) = \min_{s \in (0,1]} \mathcal{F}(P_A((U^{(n)}, T^{(n)})^{\mathrm{T}} + s\boldsymbol{D}^{(n)})),$$

where P_A is the Euclidean projection onto the set $A := U_{ad} \times \{T \geq T_{\min}\}$.

(vii) Update control and parameter $(U^{(n+1)}, T^{(n+1)})^{\top} = P_A((U^{(n)}, T^{(n)})^{\top} + s^{(n)}\boldsymbol{D}^{(n)})$.

(viii) Set $n := n + 1$ and go to Step (i).

For the ease of presentation we have stated the last algorithm with Euler steps in (49 – 52), instead of Heun steps. Actually, we implemented a Heun method being consist with Algorithm 4.1. For further information and the formulas for derivatives of $\boldsymbol{F}^{(n)}_{(k)}$ and $\mathbf{M}^{(n)}_{(k)}$ with respect to controls and for the derivatives $\delta_{(\boldsymbol{q}^{(n)},\boldsymbol{v}^{(n)},U^{(n)},T^{(n)})}J^{(n)}$, see [27, Appendix B]. Since $\bar{\boldsymbol{u}}$ and $\boldsymbol{u}$ do not appear in the objective function, the sensitivities for $\bar{\boldsymbol{u}}$ only enter in $\delta_{U^{(n)}_{(i)}}\boldsymbol{F}^{(n)}_{(k)}$, $\delta_{T^{(n)}}\boldsymbol{F}^{(n)}_{(k)}$, $\delta_{U^{(n)}_{(i)}}\mathbf{M}^{(n)}_{(k)}$, and $\delta_{T^{(n)}}\mathbf{M}^{(n)}_{(k)}$.

The Armijo line search in Step (vi) of the Algorithm 4.2 is performed as follows

Algorithm 4.3 (Projected Armijo line search)

(i) Let $\beta_A \in (0,1)$, $\sigma_A \in (0,1/2)$, $l \in \mathbb{N}^*$ and let $\phi(s) = \mathcal{F}(P_A((U^{(n)},T^{(n)}) + s\boldsymbol{D}^{(n)}))$ be given.

(ii) Find maximal $s = \beta_A^l$ such that

$$\phi(s) \le \phi(0) + \sigma_A s \phi'(0) \tag{53}$$

with $\phi'(0) = \boldsymbol{\delta}_{(U^{(n)},T^{(n)})}\mathcal{F}^{(n)} \cdot \tilde{\boldsymbol{D}}^{(n)}$, where for $i = 0,\dots,N$

$$\tilde{\boldsymbol{D}}^{(n)}_{(i)} = \begin{cases} 0 \quad ; \quad U^{(n)}_{(i)} \notin U_{ad} = [U_{\min}, U_{\max}] \text{ or} \\ \qquad \left(U^{(n)}_{(i)} = U_{\max} \wedge D^{(n)}_{(i)} \ge 0\right) \text{or} \left(U^{(n)}_{(i)} = U_{\min} \wedge D^{(n)}_{(i)} \le 0\right), \\ D^{(n)}_{(i)}; \quad \text{else,} \end{cases}$$

and

$$\tilde{\boldsymbol{D}}^{(n)}_{(N+1)} = \begin{cases} 0 \quad ; & T^{(n)} < T_{\min} \text{ or } (T^{(n)} = T_{\min} \wedge D^{(n)}_{(N+1)} \le 0), \\ D^{(n)}_{(N+1)}; & \qquad\qquad \text{else.} \end{cases}$$

(iii) Set $s^{(n)} := s$.

For the choice $\beta_A = 0.9$ and $\sigma_A = 10^{-4}$ we observe a good performance of the line search within our numerical experiments. For the Armijo line search it turns out to be crucial to solve for the new states, when computing $\phi(s)$ in the Armijo condition (53), to sufficient precision, requiring a finite element (FE) solution of the PDE for every β_A^l. Otherwise the algorithm might terminate since an admissible step size cannot be found.

Our reduced gradient reads,

$$\begin{aligned} \delta_{U^{(n)}_{(i)}}\mathcal{F}^{(n)} = {} & \nu_2 hT^{(n)} \sum_{k=i+1}^{N} S_{(k)} v^{(n)}_{(k),2}\, \delta_{U_{(i)}} v^{(n)}_{(k),2} + \nu_3 hT^{(n)} S_{(i)} U^{(n)}_{(i)} \\ & + \nu_4 (q^{(n)}_{(N),1} - q^f_1)\, \delta_{U_{(i)}} q^{(n)}_{(N),1} + \nu_5 q^{(n)}_{(N),2}\, \delta_{U_{(i)}} q^{(n)}_{(N),2} \\ & + \nu_6 v^{(n)}_{(N),1}\, \delta_{U_{(i)}} v^{(n)}_{(N),1} + \nu_7 v^{(n)}_{(N),2}\, \delta_{U_{(i)}} v^{(n)}_{(N),2} \end{aligned}$$

for the components $i = 0,\dots,N$, and

$$\begin{aligned} \delta_{T^{(n)}}\mathcal{F}^{(n)} = {} & \nu_1 + \frac{\nu_2}{2} h \sum_{k=0}^{N} S_{(k)} |v^{(n)}_{(k),2}|^2 + \nu_2 hT^{(n)} \sum_{k=1}^{N} S_{(k)} v^{(n)}_{(k),2}\, \delta_{T^{(n)}} v^{(n)}_{(k),2} \\ & + \frac{\nu_3}{2} h \sum_{k=0}^{N} S_{(k)} |U^{(n)}_{(k)}|^2 + \nu_4 (q^{(n)}_{(N),1} - q^f_1)\, \delta_{T^{(n)}} q^{(n)}_{(N),1} + \nu_5 q^{(n)}_{(N),2}\, \delta_{T^{(n)}} q^{(n)}_{(N),2} \\ & + \nu_6 v^{(n)}_{(N),1}\, \delta_{T^{(n)}} v^{(n)}_{(N),1} + \nu_7 v^{(n)}_{(N),2}\, \delta_{T^{(n)}} v^{(n)}_{(N),2} \end{aligned}$$

Clearly, for $i \ge k$ we find $\delta_{U^{(n)}_{(i)}}\boldsymbol{q}^{(n)}_{(k)} = 0$ and $\delta_{U^{(n)}_{(i)}}\boldsymbol{v}^{(n)}_{(k)} = 0$. Thus we can simplify $\delta_{U^{(n)}_{(N)}}\mathcal{F}^{(n)} = \nu_3(hT^{(n)}/2)U^{(n)}_{(N)}$. We conclude that if the control cost parameter $\nu_3 = 0$, then $U^{(n)}_{(N)}$ is arbitrary. For uniqueness, we set $U^{(n)}_{(N)} = 0$.

For small trolley/load mass ratios $\eta = m_{\mathrm{T}}/m_{\mathrm{L}}$, our ODE system turns out to be stiff and a semi-explicit method [33] is applied by using on the right-hand side of (44) $\tilde{\boldsymbol{v}}_{(k+1)}$ instead of $\boldsymbol{v}_{(k)}$ and on the right-hand side of (46) $\boldsymbol{v}_{(k+1)}$ instead of $\tilde{\boldsymbol{v}}_{(k+1)}$.

A suitable algorithm for determining the weights ν_i, $i = 4, \dots, 7$, is crucial in order to avoid a breakdown of the Armijo line search and to obtain feasible computing times. Within this framework our objective function can be interpreted as an exact penalty function [34, Sect. 5.4]. The optimal penalty weights ν_i, $i = 4, \dots, 7$ are unknown, but we may expect that we approximate the optimal weights numerically sufficiently well.

Algorithm 4.4 (Penalty method)

(i) Set initial weights $\nu_i = \nu_i^0$, $i = 4, \dots, 7$.

(ii) Run Algorithm 4.2.

(iii) If $\left|q_{(N),j}^{(n)} - q_j^f|,|v_{(N),j}^{(n)}\right| < err_1$ for $j = 1, 2$ and a given error tolerance err_1, then Stop.

(iv) Increase the weights $\nu_j := \tilde{\rho}\,\nu_j$ (where $\tilde{\rho} > 1$) for indices j corresponding to violated terminal conditions. Go to (ii).

Typically we consider the factor $\tilde{\rho} = 10$. We remark that the initial value of ν_4 has to be chosen such that the trolley remains within the feasible area of the crane beam for the first run of Algorithm 4.2. We start with $\nu_4^0 = 1/(q_1^f - q_1^0)^2$, $\nu_5^0 = 10\,\nu_4^0$, $\nu_6^0 = 5000\,\nu_7^0$, and $\nu_7^0 = 1/(T^{(0)})^2$.

5. Numerical results

We solve our OCP by means of Algorithm 4.2. This algorithm has been implemented in the open-source software package FEniCS v1.4 (API Python 2.7.3 with PETSc v3.2 as linear algebra package). It has been executed on a workstation, equipped with Intel(R) Xeon(R) CPU E5640 @2.67GHz × 16 processors and a memory of 23.6 GiB, under Ubuntu Linux. For data considered in our simulations, see Table 1.

In the following we present the results for a mesh that has been refined adaptively on $\{\Gamma_{\mathrm{C}}(q_1)\,|\,q_1^0 \le q_1 \le q_1^f\}$ beforehand by solving an auxiliary Poisson problem and refining recursively cells with a residual error larger than 10^{-4}. This procedure yielded about 10 500 vertices and 43 200 3d-cells. We consider $N = 100$ time steps on the normalized time interval $[0, 1]$, the error tolerance $err_0 = 2.0 \cdot 10^{-5}$ for the $\boldsymbol{u}$-iteration, the error tolerance $err_1 = 5.0 \cdot 10^{-2}$ for the terminal conditions, and the relative error tolerance $err = 10^{-8}$ for the optimization. As stopping condition we work with

Table 1. Typical data for a large crane (see, e.g. [35],) and further used parameters.

Description	Symbol	Value	Unit
Width crane beam	b_2	0.80	m
Height crane beam	$b_3 = b_2$	0.80	m
Length crane beam	l_2	45.80	m
Width trolley	b_{T}	0.80	m
Height trolley	h_{T}	0.10	m
Length trolley	l_{T}	0.60	m
Mass trolley	m_{T}	150	kg
Mass load	m_{L}	3340	kg
Length pendulum	l	17.5	m
Scaled maximal accelerating force	$U_{\max}$	0.006	s^{-2}
Scaled minimal accelerating force	$U_{\min}$	− 0.006	s^{-2}
Scaled mass density crane beam	ρ	0.0104	m^{-3}
Scaled Lamé parameter 1	λ	$1.76 \cdot 10^6$	$\mathrm{N\ kg^{-1}\ m^{-3}}$
Scaled Lamé parameter 2	μ	$1.33 \cdot 10^6$	$\mathrm{N\ kg^{-1}\ m^{-3}}$
Standard gravity earth	g_{e}	9.81	$\mathrm{N\ kg^{-2}}$
Initial angle trolley	q_2^0	0	rad
Terminal angle trolley	q_2^f	0	rad
Initial velocities	$\boldsymbol{v}^0$	$\mathbf{0} = (0, 0)^\top$	$(\mathrm{m\ s^{-1},\ rad\ s^{-1}})^\top$
Terminal velocities	$\boldsymbol{v}^{\mathrm{f}}$	$\mathbf{0} = (0, 0)^\top$	$(\mathrm{m\ s^{-1},\ rad\ s^{-1}})^\top$

$$\frac{\mathcal{F}^{(n)} - \mathcal{F}^{(n+1)}}{1 + \mathcal{F}^{(n+1)}} < err. \tag{54}$$

As initial guess we start with

$$U_0(t) = \begin{cases} U_{\max} & ; \quad 0 \le t < 0.4, \\ -U_{\max} & ; \quad 0.4 \le t < 0.8, \\ 0 & ; \quad 0.8 \le t \le 1, \end{cases}$$

for the control, motivated by the intuitive strategy to (i) accelerate, then (ii) brake, and (iii) wait until the system swings out. As initial guess for the total time we take $T^{(0)} = 18.5$ [s].

For the weights v we consider $v_1 = 10/T^{(0)}$, $v_2 = 500$ and we focus here on the situation $v_3 = 0$, though the algorithm clearly can handle $v_3 > 0$ as well. According to Algorithm 4.4, a suitable choice for the weights v_i, $i = 4, \ldots, 7$ is obtained by successively increasing the weights according to violated terminal conditions and restarting with the control and final time determined so far, until the terminal conditions are fulfilled to sufficient accuracy. It turns out to be crucial that the weights are scaled such that the reduced objective function is approximately of order 1.

Table 2. Numerical results for the number of optimization iterations, violation of terminal conditions and final time $T^{(n)}$ for each stage of the penalty method.

Stage	# It.	$\left\lvert q^{(n)}_{(N),1} - q^f_1 \right\rvert$	$\left\lvert q^{(n)}_{(N),2} \right\rvert$	$\left\lvert v^{(n)}_{(N),1} \right\rvert$	$\left\lvert v^{(n)}_{(N),2} \right\rvert$	$T^{(n)}$
1	18	0.4710181849	0.0342489553	0.0112566293	0.1530321823	18.49935459
2	2	0.4710181849	0.0342489553	0.0112566293	0.1530321823	18.49935459
3	184	0.0564336136	0.0166066948	0.0018213617	0.1608312973	18.49840484
4	2	0.0564336136	0.0166066948	0.0018213616	0.1608312973	18.49840484
5	998	0.0475493891	0.0451017689	0.0046039452	0.0354990300	18.49836286

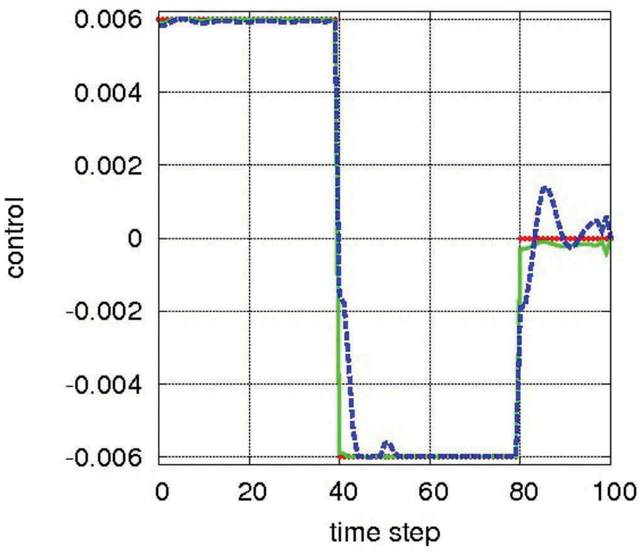

Figure 2. Initial control (red dotted line) versus time steps, the computed control after 1 stage (green continuous line) and the computed optimal control (blue dashed line) after 5 stages. The objective $\mathcal{F}$ is dimensionless and the control U is given in s^{-2}.

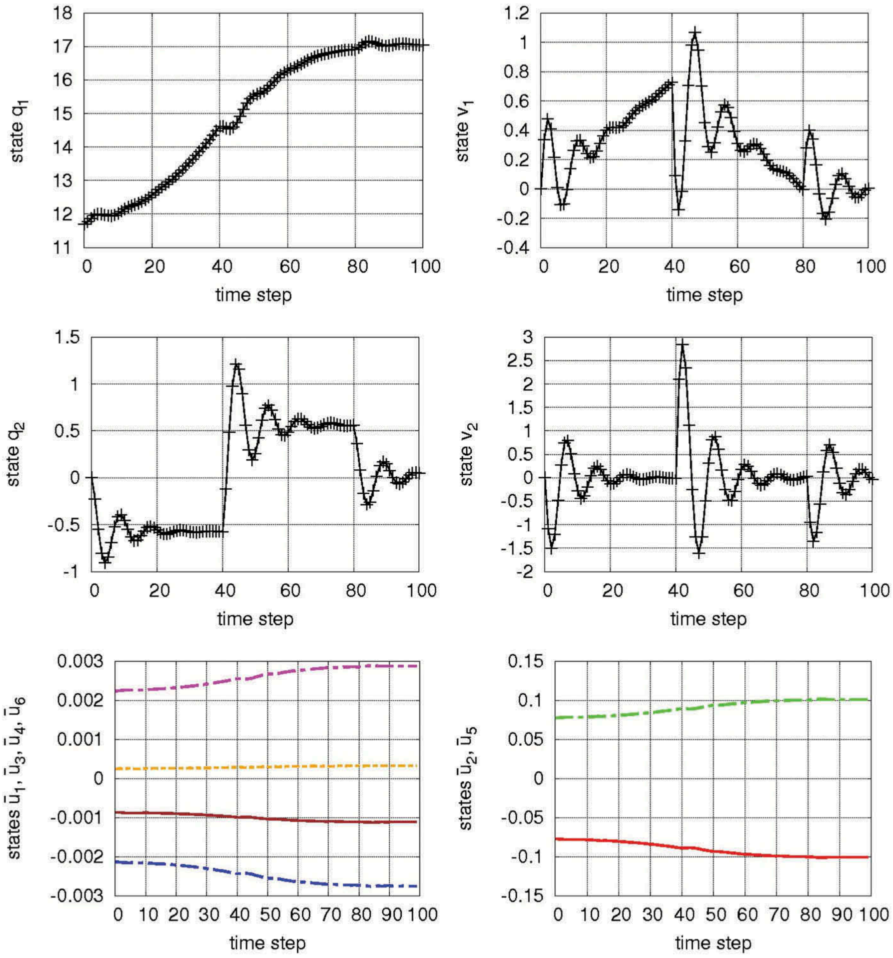

Figure 3. States $\boldsymbol{q} = (x_{\mathrm{T}}, \alpha)^{\top}$ (top/center left), $\boldsymbol{v} = \dot{\boldsymbol{q}}$ (top/center right), $\bar{\boldsymbol{u}}$ (bottom left/right) versus time step number on the interval $[0, T]$ with $T = 18.49836$ [s]; bottom left: $\bar{u}_1$ brown continuous line, $\bar{u}_3$ magenta upper dashed line, $\bar{u}_4$ blue lower dashed line, $\bar{u}_6$ orange central finely dashed line; bottom right: $\bar{u}_2$ red continuous line; $\bar{u}_5$ green dashed line. The quantities q_1, v_1 are given in m, q_2, v_2 in degree, $\bar{u}_j$, $j = 1, 2, 5, 6$, are dimensionless and $\bar{u}_3$, $\bar{u}_4$ are given in m^{-1}.

5.1. *Numerical time-optimal control*

In a first study we consider a projected gradient method as obtained from Algorithm 4.2 by fixing $\mathbf{H}^{(n)} = \mathbf{H}^{(0)}$. Table 2 shows the stages of the penalty method and the violation of terminal conditions. The optimization of the final time T is illustrated in Table 2. Here we consider as initial position of the trolley $q_1^0 = 11.7$ [m] and as terminal position $q_1^f = 17.0$ [m]. The resulting control and states are depicted in Figure 2 and in Figure 3, where the final penalty weights are $\nu_4 = 10^4 \nu_4^0$, $\nu_5 = \nu_5^0$, $\nu_6 = 10^2 \nu_6^0$, and $\nu_7 = 10^4 \nu_7^0$.

From this simulation, we find as optimal total time $T = 18.49836$ [s]. In Figure 2 we see that the initial guess of bang-bang type is quite good already, but the obtained optimal control for the final values of the weights is not of bang-bang type, though we omit a regularization term for the control by setting $\nu_3 = 0$. If we start with an arbitrary initial guess, then the computing times are increased and not for any initial guess convergence is obtained. By Figure 3 we convince ourselves that the terminal conditions $q_1^0 = 17.0, q_2^f = v_1^f = v_2^f = 0.00$ are respected within good approximation. In Figure 3, bottom left, we see $\bar{u}_2 + \bar{u}_5 = |\Gamma_{\mathrm{C}}|^{-1} \int_{\Gamma_{\mathrm{C}}} \partial_1 u_3 + \partial_3 u_1 \approx 0$, this reflects the observation that the (orthogonal) shear stress $\tau_{13} = \tau_{31} = \mu(\partial_1 u_3 + \partial_3 u_1)$

vanishes on the trolley-beam contact surface, representing a mainly isotropic compression or tension on the surface. Figure 3, bottom right, shows $\bar{u}_3 + \bar{u}_4 \approx 0$, corresponding to $\partial_1 \text{trace}(\nabla \boldsymbol{u}) \approx 0$, that is, the average pressure on Γ_C is constant in x_1-direction. Keep in mind that displacements in x_2-direction are very small due to the non-rotating crane.

By increasing the fineness of the mesh, we checked that our choice of Lagrange $\mathbb{P}_2$-elements and the applied spatial discretization avoid the well-known locking effects for Timoshenko beams [31, Ch. VI, §3]. For a mesh with about twice the number of cells, we obtain almost identical results. By the penalty method, Algorithm 4.4, we are able to obtain a lower value of the objective $\mathcal{F}^{(n)}$ than by solving directly for a fixed choice of weights. Furthermore, using the Euler method for time discretization requires several thousands of iterations while by the Heun method our algorithm terminates within several hundreds of iterations.

5.2. *Numerical time-optimal control using a modified BFGS update*

It turns out that time-optimal control of our problem works faster, when applying a projected quasi-Newton method relying on the BFGS update. However, decreasing optimality and feasibility tolerances further for time-optimal control, yields Armijo steps close to the computing precision and long computing times. Therefore we focus on the case of a fixed terminal time in the following.

5.3. *Numerical optimal control for fixed terminal time*

Our numerical optimal control presented in Subsection 5.1 turns out to require large computing times for a decreased feasibility tolerance err_1. We examine this phenomenon by considering the optimal control of the problem but now with $v_1 = 0$, $v_3 = 0.1$, a fixed terminal time $T = 19.0$ [s], and $N = 500$ time steps. Our results are depicted in Figure 4 for same data as in Subsection 5.1, except for $err_1 = 10^{-8}$, and absolute error tolerance $err_0 = 10^{-7}$ for the optimality, where as stopping criterion we use

$$\left|T^{(n+1)} - T^{(n)}\right| + \|U^{(n+1)} - U^{(n)}\|_\infty < err_0$$

instead of (54). Contrary to the last subsection, we consider $q_1^0 = 12.0$ [m] and $q_1^f = 18.0$ [m], in order to demonstrate that our algorithm does not work only for particular initial and terminal conditions of the trolley. Since the computed control and states show a qualitative behavior similar to Figure 2 and Figure 3, we omit them here.

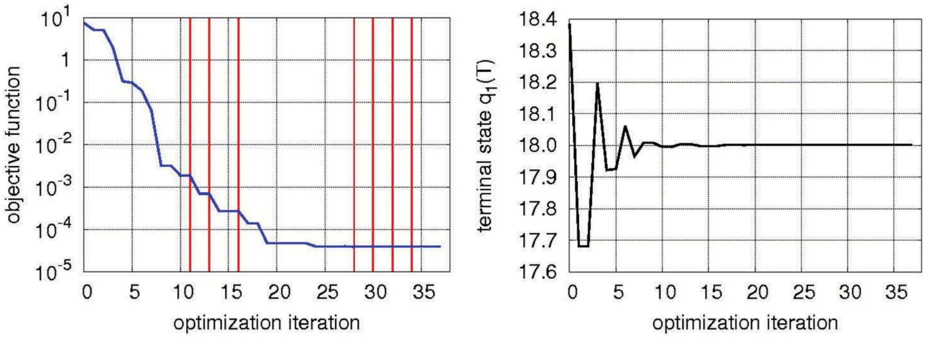

Figure 4. Reduced objective function $\mathcal{F}(U, T)$ (left), and convergence of $q_1(T)$ vs. q_1^f (right) for fixed terminal time T. On the left-hand side we show the decrease of $\mathcal{F}$ over all 8 stages (separated by vertical lines) of the penalty method yielding 37 optimization iterations in total. $\mathcal{F}$ is normalized to the initial value or to the last value of the preceeding stage of the penalty method, respectively. On the right-hand side we see the approach to the terminal trolley position q_1^f vs. the optimization iteration. The objective $\mathcal{F}$ is dimensionless and the terminal trolley position $q_1(T)$ is given in m.

We see that our combined Algorithms 4.2 and 4.4 perform more accurately in case of a fixed terminal time. This suggests that the convergence issues in Subsection 5.1 are due to the nonlinear time-optimal control. Other explanations for this observation could be the non-existence of a time-optimal control or that a further time grid refinement, turning out to be a computational challenge, is required in case of a free terminal time. Typically the precise resolution of switching points of controls is numerically expensive.

5.4. *Observations from simulations*

For the solution of the linear 3d FE systems of moderate size (43 200 cells, 21 100 degrees of freedom) for this initial study, it turned out to be sufficient to employ a direct method which re-uses the factorization of the stiffness matrix.

From further simulations we noticed the following coupling effects and dependencies:

- For heavier crane beams, that is, larger values of the mass density ρ, the trolley position q_1 moves faster to the free end and it may happen that the mass matrix $\mathbf{M}$ becomes singular.
- The longer the pendulum, that is, the larger the value of l, the smaller $\max_\Omega \|\boldsymbol{u}\|$. Then the faster q_1, the more q_2 deviates from 0 and $\det(\mathbf{M})$ tends to zero.
- For a crane beam of half the length, $\max_\Omega \|\boldsymbol{u}\|$ becomes smaller and $\det(\mathbf{M})$ is very close to 1.
- The typical ratio of trolley and load masses is $\eta = m_\mathrm{T}/m_\mathrm{L} \approx 0.05$. For $\eta \to \infty$ while the scaled total mass m remains constant, the values of $\dot{q}_1$ decrease and $\max_\Omega \|\boldsymbol{u}\|$ becomes smaller. In the opposite case, for lower values of η such as 0.01, q_1 is faster and larger angles q_2 appear.
- The inclusion of mechanical displacements has a significant impact on the speed of the trolley, for example, considering $\bar{\boldsymbol{u}}$ terms yields a slowdown of the trolley of about 10% compared to an inelastic ($\bar{\boldsymbol{u}} \equiv \mathbf{0}$) trolley system.
- Adding a damping term to the ODEs for the trolley and for the load, respectively, yields that the amplitudes of the oscillations decay faster, but the qualitative effect is neglectable on the optimal control of the whole coupled system (see [22]).

In particular, the latter observation underlines the importance of taking elastic deformations into account, represented by the terms involving $\bar{\boldsymbol{u}}$ in the ODE.

5.5. *Discussion*

With the beam fixed on both sides, that is, homogeneous Dirichlet b.c. for $\boldsymbol{u}$ on Γ_D as well as on $\Gamma_{\mathrm{D},r} = \partial\Omega \cap \{x_1 = l_2 - b_1/2\}$, our situation can be thought of as an overhead gantry crane as used frequently in environments as different as high rise racks or seaports. For $m_\mathrm{L} \to 0$ one might think of a train or truck ('trolley') on a very long beam bridge ('elastic structure') that might easily tend to vibrate, see [36] for the model and simulations. The considerations of this article could be easily applied to these elastic crane bridge-trolley-load or elastic bridge-vehicle situations.

Our results of Section 5.4 underline the necessity to incorporate elastic deformations into the standard trolley-load system. We compare with the optimal controls obtained by Chen and Gerdts [4,5] for such a trolley-load system without elasticity. They have applied smoothed Newton methods for the optimal control. Most data is of same order as in our case, but they simulate for a significantly larger mass ratio $\eta = 0.6$ and for the case $v_1 = v_2 = 0$ and $v_3 > 0$. The numerical results resemble our figures, but we observe more oscillations after a change in the control. This might be due to mechanical effects and to the lower value for η. In particular, when no control costs enter the objective function, we do not necessarily end up with a bang-bang control (see Figure 2). It is not clear so far whether the 'bang-bang-principle', known for optimal control of ODE and of semilinear elliptic PDE apart, does not hold for this kind of ODE-PDE-constrained OCP or whether our observation is only a numerical artefact. A possible explanation for this new phenomenon might be that the crane beam constitutes an infinite-dimensional system of states. From the work of Pesch *et al.* [8,20,21], we see that they encounter also controls that are not of bang-bang type. Please note that our situation differs from the Neumann boundary control of semilinear elliptic PDE, where a bang-bang result holds [37], since we consider a control in time acting by a shift on the Neumann boundary condition.

We give some reflections over coupled ODE-PDE control problems. Although this class of problems has many real-life applications, only few results exist as discussed in the introduction. This study shows a typical case study for this class of OCP, yielding a richer variety of controls than OCP with ODE, DAE or PDE constraints alone. We remark that the coupled ODE-PDE system could be regarded as a partial differential algebraic equation system (PDAE) with differential time index 1 and differential space index infinity [38]. However, the literature on PDAE, even on basic definitions, seems to be small.

6. Conclusion and outlook

We have formulated an OCP for a coupled elastic crane-trolley-load system, proved analytically the local-in-time well-posedness of the coupled ODE-PDE problem, presented a solution algorithm and computed a first numerical optimal control for typical data. The following challenges had to be overcome:

- the complexity of the model and the involved scales,
- a special algorithm had to be developed for this non-standard problem, involving ODE *and* PDE constraints and possibly a lack of differentiability of $\boldsymbol{u}$ with respect to q_1,
- the trolley-load ODE system resembles a double pendulum system, i) exponential reaction on perturbations and ii) possibly chaotic behaviour,
- the solution times for the PDE (yielding a large number of degrees of freedom) are an issue.

However, the computation of the numerical optimal control could still be improved with respect to computing times and resolution (the latter could be improved, e.g. by use of suitable variational integrators), in particular in the case of time-optimal control. Provided an improved algorithm, the OCP of a crane rotating as well could be tackled numerically.

Finally, we would like to close with an outlook. It might be useful to employ automatic differentiation for the calculation of sensitivities. Work-in-progress includes (i) modelling and numerics of the pendulum in 3d, (ii) possibly faster and more accurate first-optimize-then-discretize methods, and (iii) to examine whether we might guarantee terminal conditions by Newton-type methods instead of penalty techniques. However, it is not obvious whether a Newton method can be applied to our problem as in Chen and Gerdts [4,5]. This depends on the smoothness of the necessary optimality conditions that is not clear, since we need the Fréchet differentiabilty of $\boldsymbol{u}$ with respect to moving boundary conditions. The corresponding Fréchet derivative turns out to involve line measures, see [22]. Furthermore, it cannot be guaranteed that a Newton-based method is at all faster than a projected gradient method combined with exact penalty techniques, because the coupling of the problem might yield a dense Hessian matrix. In particular, open questions comprise theoretical results for ODE-PDE constrained optimal control.

Disclosure statement

No potential conflict of interest was reported by the authors.

References

[1] E.P. Hofer, O. Sawodny, H. Aschemann, S. Lahres, and A. Hartmann, *Verfahren zur Bahnregelung von Kranen und Vorrichtung zum bahngenauen Verfahren einer Last*, German patent application No. 199 07 989.7, 24.02.1999.

[2] H. Aschemann, *Passivity-based trajectory control of an overhead crane by interconnection and damping assignment*, in *Motion and Vibration Control*, H. Ulbrich and L. Ginzinger, eds., Springer Science + Business Media B.V., Dordrecht, 2009, pp. 21–30.

[3] S.K. Biswas, *Optimal control of gantry crane for minimum payload oscillations*, in *International Conference on Dynamic Systems and Applications*, Atlanta, Proc. of Dynam. Systems Appl. 4 (2004), G.S. Ladde, N.G. Medhin, and M. Sambandham, eds., May 20–24 2003.

[4] J. Chen and M. Gerdts, *Smoothing technique of nonsmooth Newton methods for control-state constrained optimal control problems, SIAM J*, Numer. Anal. 50 (2012), pp. 1982–2011. doi:10.1137/110822177

[5] J. Chen and M. Gerdts, *Numerical solution of control-state constrained optimal control problems with an inexact smoothing Newton method*, IMA J. Numer. Anal 31 (2011), pp. 1598–1624. doi:10.1093/imanum/drq023

[6] A. Logg, K.-A. Mardal, and G.N. Wells (eds.), *Automated Solution of Differential Equations by the Finite Element Method, the FEniCS Book*, Springer, Heidelberg, 2012.

[7] S.-J. Kimmerle and R. Moritz, *Optimal control of an elastic tyre-damper system with road contact,* Proc. Apply. Math. Mech. 14 (2014), pp. 875–876.

[8] K. Chudej, H.J. Pesch, M. Wächter, G. Sachs, and F. Le Bras, *Instationary heat-constrained trajectory optimization of a hypersonic space vehicle by ODE-PDE-constrained optimal control*, in *Variational Analysis and Aerospace Engineering*, G. Buttazzo and A. Frediani, eds., Vol. 33, Springer Optimization and Its Applications, New York, 2009, pp. 127–144.

[9] J.-M. Coron, *Control and Nonlinearity*, AMS, Providence, 2009.

[10] M. Gerdts and S.-J. Kimmerle, *Numerical optimal control of a coupled ODE-PDE model of a truck with a fluid basin*. Discrete Contin. Dyn. Syst. Suppl. 2015 (2015), pp. 515–524.

[11] R. Herzog and K. Kunisch, *Algorithms for PDE-constrained optimization*, in *GAMM-Mitteilungen*, M. Hinze and V. Schulz, eds., Vol. 33, Wiley-VCH Verlag, Weinheim, 2010, pp. 163–176.

[12] I. Lasiecka and R. Triggiani, *Control theory for partial differential equations: continuous and approximation theories, Vols I & II.*, in *Encyclopedia of Mathematics and Its Applications* Vol. 74, Cambridge University Press, Cambridge, 2000.

[13] J.L. Lions, *Contrôle Optimal De Systeèmes Gouvernés Par Des Équations Aux Dérivées Partielles*, Dunod and Gauthier-Villars, Paris, 1968.

[14] M. Hinze, R. Pinnau, M. Ulbrich, and S. Ulbrich, *Optimization with PDE Constraints*, Mathematical Modelling: Theory and Applications Vol. 23, Springer, New York, 2009.

[15] F. Tröltzsch, *Optimal Control of Partial Differential Equations, Theory, Methods and Applications*, Graduate Studies in Mathematics Vol. 112, AMS, Providence, 2010.

[16] J.T. Betts, *Practical Methods for Optimal Control and Estimation Using Nonlinear Programming*, 2nd, SIAM, Philadelphia, 2010. Advances in Design and Control Vol. 19

[17] M. Gerdts, *Optimal Control of ODEs and DAEs*, De Gruyter, Berlin, 2012.

[18] S.K. Biswas and N.U. Ahmed, *Stabilization of a class of hybrid systems arising in flexible spacecraft, J*, Optim. Theory Appl. 50 (1986), pp. 83–108. doi:10.1007/BF00938479

[19] S.K. Biswas and N.U. Ahmed, *Optimal control of large space structures governed by a coupled system of ordinary and partial differential equations,* Math. Control Signals Syst. 2 (1989), pp. 1–18. doi:10.1007/BF02551358

[20] H.J. Pesch, A. Rund, W. Von Wahl, and S. Wendl, *On some new phenomena in state-constrained optimal control if ODEs as well as PDEs are involved*, Contr. Cybern. 39 (2010), pp. 647–660.

[21] S. Wendl, A. Rund, and H.J. Pesch, *On a state-constrained PDE optimal control problem arising from ODE-PDE optimal control*, in *Recent Advances in Optimization and Its Applications in Engineering*, M. Diehl, F. Glineur, and W. Michiels, eds., Springer, Berlin/Heidelberg, 2010, pp. 429–438.

[22] S.-J. Kimmerle, M. Gerdts, and R. Herzog, *An optimal control problem for a rotating elastic crane-trolley-load system*, preprint, submitted 2017. Available at http://www.unibw.de/sven-joachim.kimmerle. 'Preprints'.

[23] P.-G. Ciarlet, *Mathematical Elasticity Vol. 1: Three Dimensional Elasticity*, Elsevier, Amsterdam, 1998. Studies in Mathematics and its Applications Vol. 20

[24] S. Nicaise, *About the Lamé system in a polygonal or a polyhedral domain and a coupled problem between the Lamé system and the plate equation. I: regularity of the solutions*, Annali della Scuola Normale Superiore di Pisa, Classe di Scienze 4^{e} série 19 (1992), pp. 327–361.

[25] T. Apel, A.-M. Sändig, and J.R. Whiteman, *Graded mesh refinement and error estimates for finite element solutions of elliptic boundary value problems in non-smooth domains*, Math. Methods Appl. Sci. 19 (1996), pp. 63–85. doi:10.1002/(SICI)1099-1476(19960110)19:1<63::AID-MMA764>3.0.CO;2-S

[26] H.W. Alt, *Linear Functional Analysis*, Springer, Berlin/Heidelberg, 2016.

[27] S.-J. Kimmerle, M. Gerdts, and R. Herzog, *Optimal control of an elastic crane-trolley-load system. A case study for optimal control of coupled ODE-PDE systems*, online version with an additional appendix. Available at http://www.unibw.de/sven-joachim.kimmerle. 'Refereed publications'.

[28] B. Niethammer, *Derivation of the LSW theory for Ostwald ripening by homogenization methods*, Arch. Ration. Mech. Anal. 147 (1999), pp. 119–178. doi:10.1007/s002050050147

[29] S.-J. Kimmerle, *Well-Posedness of a Coupled Quasilinear Parabolic and Elliptic Free Boundary Problem from a Model for Precipitation in Crystalline Solids*, preprint, Universität der Bundeswehr München, Neubiberg, 2011.

[30] R. Adams and J. Fournier, *Sobolev Spaces*, 2nd ed., Academic Press, New York, 2003.

[31] D. Braess, *Finite Elements. Theory, Fast Solvers, and Applications in Solid Mechanics*, 3rd ed., Cambridge University Press, Canbridge, UK, 2007.

[32] M.J.D. Powell, *A fast algorithm for nonlinearily constrained optimization calculation*, in *Numerical Analysis*, G.A. Watson, ed., Springer, New York, 1978. Lecture Notes in Mathematics Vol. 630.

[33] M. Arnold, B. Burgermeister, C. Führer, G. Hippmann, and G. Rill, *Numerical methods in vehicle system dynamics: state of the art and current developments*, Vehicle Syst. Dyn (2011), pp. 1159–1207. doi:10.1080/00423114.2011.582953

[34] C. Geiger and C. Kanzow, *Theorie Und Numerik Restringierter Optimierungsaufgaben*, Springer, Berlin, 2002.

[35] Liebherr-International AG, *Liebherr Turmdrehkran/Tower Crane 71 EC*, Technical Data sheet, Liebherr-Werk, Biberach, 2004.

[36] S.-J. Kimmerle, *Modelling, simulation and optimization of an elastic structure under moving loads*, Proc. Appl. Math. Mech. 16 (2016), pp. 697–698. doi:10.1002/pamm.v16.1

[37] H. Maurer and H.D. Mittelmann, *Optimization techniques for solving elliptic control problems with control and state constraints: Part 1. boundary control*, Comput. Optim. Appl. 16 (2000), pp. 29–55. doi:10.1023/A:1008725519350

[38] W.S. Martinson and P.I. Barton, *A differentiation index for partial differential-algebraic equations*, SIAM J. Sci. Comput. 21 (1999), pp. 2295–2315. doi:10.1137/S1064827598332229

PAMM · Proc. Appl. Math. Mech. **16**, 697 – 698 (2016) / **DOI** 10.1002/pamm.201610337

Modelling, simulation and optimization of an elastic structure under moving loads

Sven-Joachim Kimmerle[1,*]

[1] Universität der Bundeswehr München, Institut für Mathematik und Bauinformatik, Werner-Heisenberg-Weg 39,
85577 Neubiberg, Germany

We consider an elastic structure that is subject to moving loads representing e.g. heavy trucks on a bridge or a trolley on a crane beam. A model for the quasi-static mechanical behaviour of the structure is derived, yielding a coupled problem involving partial differential equations (PDE) and ordinary differential equations (ODE). The problem is simulated numerically and validated by comparison with a standard formula used in engineering. We derive an optimal policy for passing over potentially fragile bridges. In general, our problem class leads to optimal control problems subject to coupled ODE and PDE.

The elastic structure is modelled by a single solid thin cuboid i.e. a beam in three dimensions. By choosing suitable boundary conditions (b.c.) this serves as a simple model either for a bridge, fixed on both ends of the solid beam, or for a crane, fixed only at one side. The bridge could be passed over by vehicles that are modelled as several area loads. In case of a crane, the goal is to transport a load from an initial to a terminal position, where the load is fixed to a moving trolley that runs along the crane beam and that is modelled as well as an area load. An objective could be to pass the bridge or to move the trolley in minimal time, while the structure should not be damaged, e.g. the elastic deflection is to be minimized. A motivation for the bridge problem is the damage caused mainly by heavy trucks to road bridges and overcrossings. Heavy traffic is one of the reasons for increasing road maintenance costs. An application for the crane model is that it serves as a first step to the challenging control and optimal control for pontoon cranes that would require enhanced models. Both situations share some mathematical aspects that are discussed in this short paper together. The investigation of the simple model for the bridge has been started in [1,4]. The optimal control problem for an elastic crane with a moving load is considered in [2] in details.

As geometry we consider the undeformed reference configuration $\Omega = \{x \in \mathbb{R}^3 \,|\, 0 \leq x_1 \leq \ell, |x_2| \leq b, |x_3| \leq h\}$ in space and the time interval is $[0,T]$. In this short study the terminal time $T > 0$ is fixed. The mechanical displacement field $u : \Omega \to \mathbb{R}^3$ is considered in (undeformed) Lagrangian coordinates. The other states q correspond to the centres of mass of moving loads. The time-dependent control $U : [0,T] \to \mathbb{R}$ enters the state equations for the loads. U is subject to control constraints, i.e. $U(t) \in [U_{min}, U_{max}]$ for all times $t \in [0,T]$ for given $U_{min} < 0 < U_{max}$, modelling maximal deceleration and acceleration, respectively. For a bridge the structure is clamped at both ends, $\Gamma_D = \{x \in \partial\Omega \,|\, x_1 = 0 \vee x_1 = \ell\}$, while for a crane the beam is clamped at one end only, i.e. $\Gamma_D = \{x \in \partial\Omega \,|\, x_1 = 0\}$. The remaining boundary $\Gamma_N := \partial\Omega \setminus \overline{\Gamma_D}$ is subject to Neumann boundary conditions. In linear elasticity, i.e. under the assumption of small displacement gradients $\|u\| \ll 1$, the mechanical strain reads $\epsilon(\nabla u) = (\nabla u + \nabla u^\top)/2$. The stress is modelled by the Cauchy stress tensor $\sigma(\nabla u) = E\tilde{\nu}/((1+\tilde{\nu})(1-2\tilde{\nu}))\,\mathrm{trace}(\nabla u)\, Id_3 + E/(1+\tilde{\nu})\,\epsilon(\nabla u)$, depending on the Young modulus $E > 0$ and the Poisson number $\tilde{\nu} \in (-1, 1/2)$. This yields the following quasi-static PDE completed with Dirichlet and Neumann boundary conditions:

$$-\mathrm{div}\,\sigma(\nabla u) = H \qquad \text{in } \Omega \times [0,T], \tag{1}$$

$$u = 0 \qquad \text{on } \Gamma_D \times [0,T], \qquad\qquad -\sigma(\nabla u).\nu - G(q,\bar{u},U) = 0 \qquad \text{on } \Gamma_N \times [0,T]. \tag{2}$$

Here $H = -\rho g e_3$ is a volume force (due to the dead load of the bridge/crane beam with density ρ, g being the gravity acceleration) and ν denotes the outer normal. $G : [0,T] \times \Gamma_N \to \mathbb{R}^3$ is a boundary force modelling the total area forces of N loads with mass m_i at position q_i $(i = 1, \ldots, N)$ and has the following structure $G(q(t), \bar{u}(t), U(t), x) = \sum_{i=1}^{n} \chi_{\Gamma_C^i(q_i(t))}(x)\, G_0(q_i(t), \bar{u}(t), U(t), m_i)$, where $\chi_{\Gamma_C^i(q_i(t))} : \Gamma_N \to \{0;1\}$ is the characteristic function of the contact area Γ_C^i, $i = 1, \ldots, n$ $(n \leq N)$, where the area force G_0 is applied. The contact area is shifting with time: $\Gamma_C^i(q_i(t)) := \Gamma_C^{i,0} + q_i(t)$. Note that in G and G_0 a vector of mean values $\bar{u}$, averaging spatially the functions u, ∇u, and $D^2 u$ over the contact area Γ_C^i enters. The ODE states $q : [0,T] \to \mathbb{R}^N$ are subject to the Newton law of motion $M(q)\,\ddot{q} = F(q, \dot{q}, \bar{u}, U)$ for all times $t \in [0,T]$, where M is a given mass matrix and F is a generalized force, combining Coriolis and external forces, the latter depending on the control U. Since it is reasonable to consider a model, where the ODEs do not depend additionally on x, this motivates that the dependence of F on u is modelled by averaging over u on Γ_C^i and considering a dependence of F on $\bar{u}$ instead. Terminal conditions for the ODE states could be added as quadratic penalty terms to the objective $J(U, u, q, \dot{q})$ that is to be minimized w.r.t. U.

The PDE problem (1) & (2) is solved by the finite element method, while the ODEs are discretized by an Euler method. In order to avoid locking effects (typical for thin beam-like structures) and to represent the second order derivatives $D^2 u$, we

[*] Corresponding author: e-mail sven-joachim.kimmerle@unibw-muenchen.de, phone +49 89 6004 3082, fax +49 89 6004 4136

consider quadratic Lagrange elements. The mesh is refined on the part of Γ_N, where the loads may possibly move. In order to avoid technical problems, whether a finite cell belongs to Γ_C^i or not, the characteristic functions $\chi_{\Gamma_C^i}$ are approximated smoothly. The dependence of the b.c. (2).2 on mean values $\bar{u}$ over u is solved by an inner fixed point loop. An outer loop, being another fixed point iteration, is applied for solving the possibly fully coupled ODE-PDE system. For details for the algorithm see in case of the crane [2, Sect. 3 & 4].

The algorithms for the bridge as well as for the crane problem have been implemented in the free, open source finite element library FEniCS. Here we present the simple case of trucks, passing with constant velocity over the whole bridge. This situation has been simulated for a full bridge (3-dimensional beam) and optimized by comparing various scenarios (number of trucks, distance, opposing traffic) in [4]. Here the simulations are validated by increasing the polynomial degree of trial functions and comparing maximal displacements of the bridge. In this case the maximum is attained at the centre of the bridge. Varying parameters, we observe that the width b of bridge has almost no impact on the maximal displacement, while length ℓ and height h do. This behaviour is due to the planar moment of inertia $I_3 = bh^3/12$ [5, §17]. For the simulation of two trucks following each other at two different distances, see Fig. 1. The opposing traffic of trucks exhibits even more critical effects.

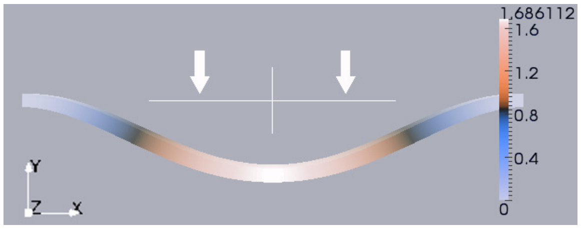

Fig. 1: Deformed beam (scaled), colored according to absolut deformation $\|u\|$, in the case of two trucks following each other with a distance of 12 m, at $t = T/2 = 20$ s, when the maximal deflection is attained at the centre of the bridge. Here $\ell = 100$ m, $b = 6$ m, $h = 3$ m, $E = 2.1 \cdot 10^6$ N/m^2, $\tilde{\nu} = 0.3$, $\rho = 7850$ kg/m^2, and $G_0 = -10^7\, e_3$ N/m^2 acts for both trucks on an area $\Gamma_C^0 = 3\text{m}\times 18\text{m}$. The two centres of mass are indicated by arrows, cf. [4, Abb. 23].

However, the simulation of a full 3-dimensional model of the bridge involves many unknowns due to the 3-dimensional mesh. This motivates to reduce the model by replacing the full bridge structure by a suitable plate model. For modelling the bridge as a Kirchhoff plate and the corresponding simulations see [1]. Then the structure is reduced to a 2-dimensional geometry $\tilde{\Omega} = \{x \in \Omega \,|\, x_3 = 0\}$. The resulting PDE in $\tilde{\Omega} \times [0, T]$ for the 2-dimensional displacement w is the Kirchhoff plate equation $k\, \Delta^2 w = \tilde{G}$, where k is the rigidity constant of the plate and $\tilde{G}$ represents the original effects of volume and area forces, H and G. It is an elliptic forth-order PDE and it is equipped with two suitable b.c., depending on whether the plate is fully clamped or simply supported at the left- and right-hand-sides, respectively. The implementation uses a discontinuous Galerkin method using a penalty term in order to guarantee continuity over element boundaries [6]. The simulation [1] is validated by comparing the maximal deflection in vertical direction with a standard formula [5, §20]. For the optimal control of the elastic crane-trolley-load system by means of a sensitivity-based first-discretize-then-optimize approach, see [2].

Our simulations suggest as an optimal policy for driving over the bridge, to respect a certain minimal distance between two heavy vehicles and, in particular, to avoid opposite traffic. This shows that our simple model for an elastic bridge verifies statements well-known in civil engineering. However, so far our models use the quasi-static PDE of linear elasticity and cannot prescribe a swinging behavior due to elastic waves that might be relevant e.g. for large suspension bridges. The latter would require second-order time derivatives of u in (1), leading to a hyperbolic PDE instead of an elliptic PDE. The overall aim is an optimal control for passing over potentially fragile bridges. This leads to an optimal control problem subject to coupled ODE and PDE as considered for the elastic crane-trolley-load system in [2]. For an efficient computation of this full optimal control problem for the bridge, it could be helpful to study, whether the bridge model could be reduced further to a 1-dimensional Euler-Bernoulli beam. So far only a few studies exist for optimal control of fully coupled ODE-PDE, see e.g. the optimal braking of a truck with a fluid container [3] and the references therein.

References

[1] H. Händly, Simulation des Befahrens einer elastischen Brücke durch schwere Fahrzeuge. Master thesis (Universität der Bundeswehr München, Neubiberg, 2015).

[2] S.-J. Kimmerle, M. Gerdts, and R. Herzog, Optimal control of an elastic crane-trolley-load system - A case study for optimal control of coupled ODE-PDE systems. Preprint (Universität der Bundeswehr München, Neubiberg, 2016).

[3] S.-J. Kimmerle and M. Gerdts, Necessary optimality conditions and a semi-smooth Newton approach for an optimal control problem of a coupled system of Saint-Venant equations and ordinary differential equations, Pure Appl. Funct. Anal. **1**, 231–256 (2016).

[4] P. Nixdorf, Simulation einer Brücke unter quasi-statischer Belastung. Report of pre-master thesis project (Universität der Bundeswehr München, Neubiberg, 2015).

[5] L. D. Landau, E. M. Lifshitz, Theory of Elasticity, vol. 7 of Course of Theoretical Physics, 3rd ed. (Butterworth-Heinemann, Burlington, MA, 1986).

[6] G. N. Wells and N. T. Dung, A C^0 discontinuous Galerkin formulation for Kirchhoff plates, Comput. Methods Appl. Mech. Engrg. **196**, 3370–3380 (2007).

4.4 Article: Optimal Control of an Elastic Tyre-Damper System with Road Contact

Here we present a proceedings article [KM14] that considers the optimal control of a quarter car model yielding an elastic tyre-damper system with road contact. The main feature of this problem is the presence of a free contact boundary. Let $0 < r < R < \infty$ denote the interior and exterior radius of the undeformed tyre that are given and $\Omega = \{r < (x_1^2 + x_2^2)^{1/2} < R\} \times (0, t_f)$. The terminal time $t_f > 0$ is fixed. The 2D model consists of the linear elasticity equation (elliptic second-order) for the mechanical displacement u (w.r.t. the undeformed coordinate system) in an elastic tyre and a second-order ODE for the displacement z of the chassis mounted on the tyre by a spring-damper element. The ODE follows by the Newton laws. Furthermore, the displacement y at the wheel rim base $\Gamma_i = \{(x_1^2 + x_2^2)^{1/2} = r\} \times (0, t_f)$ has to be determined as well. The elastic tyre has a free contact with the road that is modelled by a complementarity condition at the boundary $\Gamma_e = \{(x_1^2 + x_2^2)^{1/2} = R\} \times (0, t_f)$ of the PDE problem. The PDE problem for itself is a Signorini problem. We may control proactively the damping coefficient D of the electrorheologic damper within certain bounds $D_{min} \leq D \leq D_{max}$. Thus the control enters as a coefficient in front of the velocity terms in the ODE for z.

We assume that the road profile is known. In practice this may be realized, e.g., by a windshield camera or suitable sensors at the front of the car. Note that we neglect friction, horizontal motion of the tyre, and the more complex nonlinear elastic behaviour of the tyre, since this is the first approach to this kind of problem including the full tyre to the best knowledge of the author. Other studies consider a more simplified problem, where the elastic tyre is replaced completely by one, e.g. [Ca93, RS05, MG14], or multiple springs [BG16]. The advantage of our approach is mainly that it provides an explicit formula for the spring coefficient that is required in those models, where the tyre is just replaced by a spring. Finally, we would like to mention that we do not allow for actuator forces (requiring another auxiliary motor) as used in some real-world applications.

In the full model we have the following coupling

$$y, z \text{ (spring-damper force)} \rightsquigarrow \text{Neumann b.c. } \sigma(\nabla u).\nu \text{ (on } \Gamma_i), \tag{4.5}$$

$$\text{Dirichlet boundary value } u \text{ (on } \Gamma_i) \rightsquigarrow y \rightsquigarrow z, \tag{4.6}$$

where ν denotes the normal vector here. Alternatively, we could model the coupling such that

$$y, z \text{ (spring-damper force)} \rightsquigarrow \text{Dirichlet b.c. for } u \text{ on } \Gamma_i, \tag{4.7}$$

$$\text{Neumann boundary value } \sigma(\nabla u).\nu \text{ (on } \Gamma_i) \rightsquigarrow y \rightsquigarrow z. \tag{4.8}$$

This model is reduced by the so-called Hertzian stress approximation, that provides an explicit algebraic equation for the width of the contact boundary and the tyre displacement at the active contact boundary. By a further assumption this yields an explicit expression for the displacement

y at the tire rim base. Having checked numerically the applicability of the Hertzian stress contact formula in our case, this allows the reduction of the model such that a one-sided coupling from the free boundary to the ODE remains:

$$\text{Free boundary contact width } a \rightsquigarrow \text{Dirichlet b.c.} \quad \text{for } u, y,$$
$$y \rightsquigarrow z.$$

Note that it is required only once to solve for the PDE after the optimization algorithm, since the Hertzian stress approximation decouples the problem.

In the included paper this model is optimally controlled in a way such that the comfort (the acceleration $\ddot{z}$ of the chassis), the spring robustness (difference $z - y$ between chassis and tyre displacement) and the safety (related to the contact width a of the tyre with the road) are minimized. It is consistent to fix the terminal time, since the velocity of the quarter car is prescribed. Again the second-order ODE system is considered as two linear ODE of first-order for the simulation and the optimal control. A sensitivity-based FDTO reduced approach is applied. The projected gradient method equipped with an Armijo line search (see Algorithm 2.28 with the negative gradient as descent direction) is used. We may reuse the algorithm and code as developed for the elastic crane-trolley load problem.

Here again we apply again the finite element method for the elasticity PDE using quadratic Lagrange elements CG_2, the Heun method (of second-order) for the ODE is exploited, and this algorithm is implemented in FEniCS [LMW12]. Note that the obtained numerical optimal control is similar to a bang-bang type (no control effort is minimized) and depends significantly on the weights chosen in the objective.

We put the full optimal control problem into the setting of the theory developed in Chapter 3. The state space for $[u, y, z]^\top$ is the Banach space

$$L^\infty(0, t_f; H^2(0, t_f)) \times [H^2(0, t_f)]^2,$$

the control space for D is

$$U = L^\infty(0, t_f),$$

and the subset of admissible controls is

$$U_{ad} = \{D \in U \,|\, D_{min} \leq D \leq D_{max}\}.$$

Thus we consider box constraints for the control and U_{ad} is a closed convex subset of U. We could think about imposing state constraints in order to ensure that the tyre rim never hits the road (what does not happen for realistic data) and that the contact width a remains strictly positive, since the loss of contact is tried to be avoided in reality. However, in our approach the latter is achieved by means of a sufficiently large weight in the objective for the safety term.

Working with (4.5) & (4.6), in this problem the averaging-evaluation operator (see Def. 3.1) appears naturally as

$$\mathcal{E}u(t, x) := u(t, x)|_{\Gamma_i} = y(t), \tag{4.9}$$

since the elastic tyre rim Γ_i is non-deformable in good approximation. Whereas using the coupling structure (4.7) & (4.8) would yield the equation

$$y(t) = \mathcal{E}_{(\Gamma_i)} u(t,x) := \fint_{\Gamma_i} u(t,x)\, d\Gamma_i(x) \tag{4.10}$$

in the problem.

The next steps include the well-posedness of the one-sided and the fully coupled problem and the examination, whether an adjoint-based approach is feasible. In contrast to the elastic crane-trolley-load problem, here $\mathcal{E}$ does not yield a complicated structure and we do not have to deal with a moving boundary condition.

The full coupling has been also simulated and compared with the Hertzian stress approximation [Mo13], using quadratic finite elements. The fully coupled CDE could be considered (as a constraint) in an optimal control problem in principle as well, though the computational effort is much higher, whereas the precision is restricted by the approximation of the contact boundary by finite elements anyways.

PAMM · Proc. Appl. Math. Mech. **14**, 875 – 876 (2014) / **DOI** 10.1002/pamm.201410418

Optimal Control of an Elastic Tyre-Damper System with Road Contact

Sven-Joachim Kimmerle[1,*] and **Roman Moritz**[1]

[1] Universität der Bundeswehr München, Institut für Mathematik und Rechneranwendung, Werner-Heisenberg-Weg 39, 85577 Neubiberg, Germany

We study an elastic tyre with a wheel rim that is suspended at the chassis of a car by means of a spring-damper element. This quarter car model may be controlled by varying the damping constant of the electrorheological damper. Our mathematical model yields a coupled ODE-PDE problem with a free boundary at the tyre-road contact. In this study we approximate the tyre by the Hertz contact stress formula. The resulting optimal control problem with control constraints is solved numerically.

1 Modelling of a quarter car and differential equations

In modern cars a dynamic control of the vehicle suspension is employed for better safety and comfort. The road profile is monitored, e.g. by a stereo camera, and the suspension is actively controlled. We focus on a proactive control of an electrorheological damper where the damping constant $D \in J = [D_{min}, D_{max}]$ may be varied. In our model we do not use additional actuator forces for the control. We consider an elastic tyre-damper system where the wheel rim is connected to the car chassis by a spring-damper element. The tyre has a contact boundary with the road where it is deformed, the free boundary and the deformation depending on the weight of the car and the elastic forces. For simplicity, we consider a 2d cross-section of the quarter car model. The geometry of the model is depicted in Fig. 1, the time-interval is $(0, T)$.

The system is modelled by an ordinary differential equation (ODE) for the spring-damper-system, the stationary elasticity equation for the tyre deformation, and a complementarity condition for the free boundary with the road. Newton's law of motion yields for the (relative) displacement z of the spring from its rest position (e.g. [1, (4.1)])

$$m\ddot{z} = -D(\dot{z} - \dot{y}) - k(z - y) \ \text{ in } (0, T), \qquad z(t = 0) = \dot{z}(t = 0) = 0 \tag{1}$$

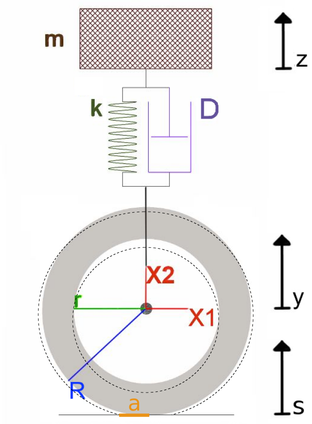

Fig. 1: Geometry of the quarter car model with free road contact. The elastic tyre is described in the undeformed configuration (solid grey).

where y is the (relative) displacement of the rim, m the quarter mass of a car chassis, and k the spring constant. As an approximation the elastic tyre is modelled in linear elasticity yielding the partial differential equation (PDE) for the displacement $u(x, t)$ in Lagrangian coordinates (undeformed configuration) with boundary conditions (b. c.):

$$-\operatorname{div}\sigma(u) = -\rho g e_2 \qquad \text{in } \{r < (x_1^2 + x_2^2)^{1/2} < R\} \times [0, T], \tag{2}$$

$$u = y(t)\, e_2 \ \text{ or } \ -\sigma(u).n = -F(t)/(\pi r b) e_2 \qquad \text{on } \{(x_1^2 + x_2^2)^{1/2} = r\} \times [0, T], \tag{3}$$

$$0 \leq -n \cdot \sigma(u).n \perp ((s(x_1, t) - R + x_2)\, e_2 - u) \cdot n \geq 0 \qquad \text{on } \{(x_1^2 + x_2^2)^{1/2} = R\} \times [0, T], \tag{4}$$

$$\tau \cdot \sigma(u).n = 0 \qquad \text{on } \{(x_1^2 + x_2^2)^{1/2} = R\} \times [0, T]. \tag{5}$$

This is a Signorini problem. Here σ denotes the Cauchy stress tensor, linear in ∇u, depending on the Young modulus E and the Poisson ratio ν. ρ prescribes the mass density within the tyre and g the gravitational acceleration. Eq. $(3)_1$ models that the displacement u is continuous at the wheel rim where the (stiff) wheel rim is shifted by the unknown y, while the second b.c. models the balance with the force F, exerted by the rim, per area (b the width of the tyre). In general both b.c. in (3) have to be satisfied, one b.c. enters the PDE problem for u, the other determines y. Eq. (4) & (5) encodes that there is either a negligible outer pressure and a positive gap between the deformed tyre and a given road profile s, or there is a positive contact pressure and contact between deformed tyre and road. n denotes the outer normal of the tyre, τ the tangential.

The symmetric free boundary is determined approximately by the Hertz formula (plain strain) [4] yielding $a = \sqrt{Rd}$ with the maximal penetration depth $d = -\frac{4}{\pi}\frac{1-\nu^2}{Eb}F$ under the static force $F = -(m + m_T)g + k(z - y)$ where m_T the mass of tyre and rim. On the contact boundary $\Gamma_c := \partial B_R(0) \cap \{|x_1| < a(t), x_2 < -r\}$, according to Hertz for the normal pressure $-e_2 \cdot \sigma(u).e_2 = (\frac{E}{1-\nu^2}\frac{F}{bR}((x_1/a)^2 - 1))^{1/2} \geq 0$, since typically $F \ll 0$. The Hertz approximation is justified numerically by

[*] Corresponding author: e-mail sven-joachim.kimmerle@unibw.de, phone +49 89 6004 3082, fax +49 89 6004 2136

solving the complementarity conditions iteratively [3]. The approximation motivates to replace (4) & (5) by the explicit b.c.

$$-\sigma(u).n = 0 \text{ on } (\partial B_R(0) \setminus \Gamma_c)\times(0,T), \qquad u = \left(s(x_1,t) - R + (R^2 - x_1^2)^{1/2}\right) e_2 \text{ on } \Gamma_c\times(0,T). \tag{6}$$

If we model the tyre as a second spring [1, (4.2)], then we have $m_T\ddot{y} = -(m+m_T)g + k(z-y) + D(\dot{z}-\dot{y}) - k_T(y - s(0,\cdot))$, where we introduce k_T as a "spring" constant for the tyre. As a first approach, the vertical acceleration and the damping term are neglected, motivated by $m_T \ll m$ and $D \ll k$. In $(3)_1$, we approximate $u_2|_{\partial B_r(0)} - s(0,\cdot) \approx -d$ which is reasonable for small strains. We identify $k_T = \frac{\pi}{4}\frac{Eb}{1-\nu^2}$ and we end up with the explicit formula $y = \frac{1}{k_T+k}(-(m+m_T)g + kz + k_T s(0,\cdot))$.

2 Optimal control problem and numerical results

The goal is to minimize the objective $I(D) = \frac{1}{2}\int_0^T \alpha_1|\ddot{z}(D)|^2 + \alpha_2|z(D)-y(D)|^2 + \alpha_3|y(D)-s(0,\cdot)|^2$ where the parameters α_k, $k = 1, 2, 3$, are positive. Here we consider the states u, z, and y as functions of the control D. Minimizing the first term in the reduced objective function corresponds to an increase in the comfort, the second term accounts for the spring robustness (a spring cannot be contracted/extended without limits), and the last "safety" term models the change of distance between rim and road (neither the grip i.e. the contact area should go to zero nor the chassis should touch the road) [1, 4.3.1.].

As a first approach we solve the optimal control problem to minimize $I(D)$ subject to (1), (2), $(3)_1$, (6), and $D \in J$ by a projected gradient method with Armijo line search. The gradient is computed by a sensitivity based approach. Within our approximation u does not enter into the reduced cost functional and there is no need to solve a sensitivity PDE for $u_{,D}$. Besides $y_{,D} = \gamma z_{,D}$ where $\gamma = k/(k_T + k)$. For the sensitivity $z_{,D}$ we have the ODE $m\ddot{z}_{,D} = -(1-\gamma)\dot{z} - D(1-\gamma)\dot{z}_{,D} - k(1-\gamma)z_{,D}$. The ODE for z and $z_{,D}$ are discretized by Heun's method and the PDE for u may be solved by FEM. This is implemented in the open software FEniCS. As a stopping condition we work with $(I(D)^{(i)} - I(D)^{(i+1)})/(1 + |I(D)^{(i+1)}|) < err$ where the index (i) indicates the optimization iteration. For numerical results, in case of an initial control $D^{(0)} = 1500$, weights $\alpha_1 = 1/4$, $\alpha_2 = 10^2$, $\alpha_3 = 10^4$, $err = 10^{-13}$, and $2 \cdot 10^4$ time steps, see Fig. 2. The given road profile is $s(x_1,t) = 0.25(1 + \cos(2(x_1+t)))\,\chi_{[\pi/2,3\pi/2]}(x_1+t)$, χ a characteristic function. Realistic data for a midsized car has been used. We observe an antagonistic behaviour between optimal comfort and safety as in [5]. Comfort yields high oscillations of D, while safety requires high accelerations $\ddot{z}$.

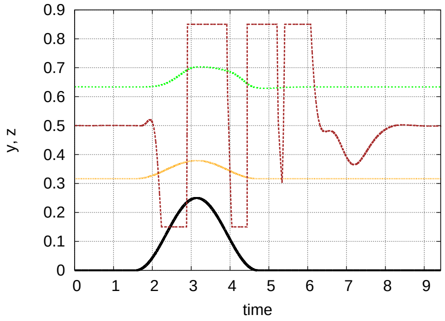

Fig. 2: Computed optimal control D (brown dashed line) for a compromise between comfort and safety, $T = 3\pi$. Relative displacement z of the chassis (shifted by $2R$, green upper dotted line) and y of the rim (shifted by R, orange lower dotted line) together with cosinusoidal road profile $s(0,t)$ modelling a speed bump (black continuous line).

The approach to optimize the vehicle dynamics by means of optimal control and not only by feedback control seems to be new within this field [5]. In many other models the elastic tyre is replaced at once by a spring [1, 2, 5] or several springs (e.g. software CDTire by ITWM Kaiserslautern). These models might, as well as our model, allow for a real-time control which is out of sight for a fully ODE-PDE model. However, in these analogous models the "tyre spring" parameters have to be fitted, that are given directly in our model. Compared to [3], where another approximation $F \approx -(m + m_T)g$ is used, we obtain slight corrections. Well-posedness of this problem will be discussed in an upcoming study. A future goal is to work without the Hertz approximation and to consider the full coupling between u and y. This situation raises also interesting analytical questions. Finally, it might be interesting to incorporate friction and non-linear elasticity into our model.

Acknowledgements The authors would like to thank professor Matthias Gerdts for his valuable advice.

References

[1] B. Cai, Neural Networks, Fuzzy Logic, and Optimal Control for Vehicle Active Systems with Four-Wheel Steering and Active Suspension, Fortschrittsberichte VDI Reihe 12 Nr. 196 (VDI Verlag, Düsseldorf, 1993).
[2] J. Michael and M. Gerdts, Optimal Control in Proactive Chassis Dynamics - A fixed step size time-stepping scheme for complementarity problems. In: Progress in Industrial Mathematics at ECMI 2012, Mathematics in Industry, Vol. 19, Springer, 2014. To appear.
[3] R. Moritz, Simulation und Optimierung eines Reifens in Fahrbahnkontakt. Master thesis (Universität der Bundeswehr München, Neubiberg, 2013).
[4] V. L. Popov, Contact Mechanics and Friction (Springer, Heidelberg, 2010).
[5] U. Rettig, and O. von Stryk, Optimal and Robust Damping Control for Semi-Active Vehicle Suspension. In: Proc. 5th EUROMECH Nonlinear Dynamics Conference (ENOC 2005), Eindhoven, The Netherlands (2005) pp. paper no. 20-316.

4.5 Article: Optimal Control of Phase Transitions

The last article [Ki12] that is included in this work is on a model with CDE, where a state is a measure or switching conditions appear, respectively. We consider a generalization of the classical Lifshitz-Slyozov-Wagner (LSW) model for an unknown measure[3] $\nu : (t, V) \to \mathbb{R}_0^+$ describing, e.g., a droplet (bubble, particle) volume distribution that evolves with time and that is subject to an algebraic equation that enforces the mass (or volume) conservation. The PDE for the measure is a transport equation and hyperbolic of first-order, the measure has to be non-negative in the sense that $\nu(t, V) \in \{0\} \cup (V_{min}, \infty)$, where V_{min} represents a physical minimal volume for a droplet to behave still as a liquid and not as a gas or molecule. Commonly, the algebraic equation is differentiated w.r.t. time, yielding a first-order ODE for the so-called mean field $\bar{V}$ that is a mean volume dominating the solution far away from droplets.

We observe the coupling structure

$$\begin{aligned} \nu &\rightsquigarrow \overline{V} \text{ (as coefficient)},\\ \overline{V} &\rightsquigarrow \nu \text{ (as coefficient)}. \end{aligned}$$

We wish to optimally control certain moments of the distribution with measure ν. Here the zeroth and first moment at t_f are interesting modelling the total number and the total volume of droplets at the terminal time. Apart from the control effort, we add a variance term, i.e. the quadratic deviation from the mean value of $\nu(t_f, \cdot)$, to the objective. In this paper, the terminal time t_f is fixed. We may control the initial mass u_0 that may be interpreted as a parameter to be identified, i.e. chosen optimally in this context, and the temperature $u_1(t)$ of the system that enter both as coefficients at various places in the equations. u_0 and u_1 are subject to box constraints.

We semi-discretize the single PDE for the measure by considering initial data that consists of a fixed number $\mathcal{N}_0$ of initial droplets with volume V_i^0, i.e. a sum of Dirac data that is normalized. The latter is a so-called Mullins-Sekerka (MS) model. Under certain periodicity assumptions it may be justified rigorously that LSW is the homogenization limit of the MS model [Ni99, NO01]. In our discretization the PDE for the measure is replaced by $\mathcal{N}_0$ ordinary differential equations of first-order for V_i, $i = 1, \dots, \mathcal{N}_0$, that are nonlinear and fully coupled to each other, but the coupling takes place only via the mean field equation. Over time droplets may vanish (corresponding to the fact that the LSW partial differential equation has only leaving characteristics) and nucleation is not allowed in this model. Thus the droplet number $\mathcal{N}$ is another time-dependent state.

The coupling structure of the semi-discretized problem reads

$$\begin{aligned} V_i &\rightsquigarrow \overline{V} && \text{(as coefficient) } \forall i = 1, \dots, \mathcal{N},\\ V_i &\rightsquigarrow \mathcal{N} && \text{(vanishing droplets) } \forall i = 1, \dots, \mathcal{N}, \end{aligned}$$

[3] Note that the measure is denoted ν_t in the paper [Ki12].

$$
\begin{aligned}
&\mathcal{N} \rightsquigarrow \overline{V} && (\text{“evaluation operator”}),\\
&\overline{V} \rightsquigarrow V_i \quad \forall i = 1, \ldots, \mathcal{N} && (\text{as coefficient}).
\end{aligned}
$$

The vanishing droplets are a numerical challenge, since the times when droplets vanish have to be computed with sufficient precision. Since for long time intervals no droplet vanishes, adaptive time-schemes are crucial for an efficient solution of the state equation.

For the numerical optimal control we consider a full discretization approach relying on sensitivities. For the time-discretization of the states central differences are used. For the numerical solution it is preferable to differentiate the algebraic equation, i.e. working with the mean field ODE. The numerical solution of the mean field ODE has been computed by a Runge-Kutta integrator with fixed, but sufficiently small time stepping as well as by the DASSL solver [BCP96]. Note that the numerical conservation of mass with time is crucial for a reasonable solution.
Our algorithm has been implemented in OCODE 1.5 (part of the package OCPID-DAE1 [Ge10]) that exploits the direct shooting method and internally uses the SQP method for the solution of the discretized system. We observe that for u_0 the smallest admissible value is optimal and a bang-bang type for the control u_1 is obtained. This corresponds to the situation of classical DAE optimal control.

Finally, in the included paper the stability of the semi-discretized mean field problem is discussed as well. Note that in this paper a problem involving nanodroplets is considered. However, this can be as well adapted to our problem for bulk and surface nanobubbles as simulated in [Ki15]. Furthermore, many other types of precipitates may be considered as well.

In the original problem the state space for $[\nu, \overline{V}]^\top$ reads

$$Y = C^0_{weak}([t_0, t_f], C^0_0(0, \infty)^*) \times C^0([t_0, t_f]; \mathbb{R})$$

where C^0_{weak} indicates a weakly continuous map $t \to \nu$. We have the pure state constraints

$$
\begin{aligned}
\nu_t(V) &\geq 0 && \forall V \in \mathbb{R}^+_0 \, \forall t \in (t_0, t_f),\\
\overline{V} &\geq 0 && \forall t \in (t_0, t_f).
\end{aligned}
$$

In the semi-discretized problem for the characteristics V_i, $i = 0, \ldots, \mathcal{N}_0$, the droplet number $\mathcal{N}$, and the mean field, we have

$$Y = [H^1(t_0, t_f)]^{\mathcal{N}_0} \times BV(t_0, t_f; \mathbb{N}_0) \times H^1(t_0, t_f).$$

Here BV denotes functions of bounded variation. This is subject to the state constraints

$$
\begin{aligned}
&V_i \geq V_{min} && \forall i = 1, \ldots, \mathcal{N} \, \forall t \in (t_0, t_f),\\
&\overline{V} \geq 0 && \forall t \in (t_0, t_f).
\end{aligned}
$$

We consider box constraints for the initial mass and the temperature control. Thus U_{ad} is a closed convex subset of $U = \mathbb{R}^+ \times L^\infty(t_0, t_f; \mathbb{R}^+)$.[4] For well-posedness of the classical LSW

[4] Mathematically, we may consider $\mathbb{R} \times L^\infty(0, t_f; \mathbb{R})$ as well, but in physics this makes no sense.

equations for given control, see [NP00, NP01]. In this problem the evaluation operator is just the integration over the measure. In the semi-discretized problem this becomes just the sum over the still existing droplets and no special treatment of the evaluation operator is required.

Since this problem is very complex, it has been presented here only for illustration of what challenges may occur in applications for coupled differential equations. An analytic theory for this problem alone is out of scope for this study. Note that the switching conditions, i.e. changing technically from an ODE to the trivial algebraic equations $V_i = 0$ for further times, once droplet i has vanished, is a non-trivial issue for optimal control. Furthermore, problems of this type might be extended to the case where switching costs (see, e.g., [Ge12, Subsect. 7.3.1]) are present in the objective. The optimal control of semilinear parabolic PDE involving switching times is covered in [CKP17].

Open tasks are to include a free terminal time, to establish efficient algorithms for long-time behaviour (see [CG01] for this challenge in the context of the classical LSW equations), and to solve the optimal control problem with an adjoint approach. Due to the switching points the latter is non-standard.

Optimal Control of Mean Field Models for Phase Transitions

Sven-Joachim Kimmerle *

* *University of the Federal Armed Forces Munich,*
Institute of Mathematics and Computer Applications (LRT 1),
85577 Neubiberg/München, Werner-Heisenberg-Weg 39, Germany
(e-mail: sven-joachim.kimmerle@unibw.de)

Abstract: Various models prescribe precipitation due to phase transitions. On a macroscopic level the well-known Lifshitz-Slyozov-Wagner (LSW) models and its discrete analogons, so-called mean field models, prescribe the size evolution of precipitates for two-phase systems. For industrial tasks it is desirable to control the resulting distribution of droplet volume. While there are optimal control results for phase-field models and for nonlinear hyperbolic conservation laws, it seems that control problems for LSW equations and mean field models, including measure-valued solutions or switching conditions, have not been considered so far. We formulate the model for this important new control problem and present first numerical results.

Keywords: Optimal control, Process control, Nonlinear control systems, End point control, Partial differential equations, Characteristic curves, Differential algebraic equations (DAE), Mathematical models, Phase transition, LSW equation

1. MODELS FOR PHASE TRANSITIONS

Phase transitions are an important phenomenon in material science. On the one hand phase transitions may be exploited in order to design a requested material, on the other hand they may destroy desirable properties of designed materials. For example, the industrial production process of semi-insulating gallium arsenide (GaAs) requires at the end some additional heat treatment at high temperatures in order to ameliorate the quality of the semi-insulator.

One of the challenges is the necessity to guarantee a mean mole fraction of arsenic (As) in the GaAs wafer of $X_0 = 0.500082$ within high accuracy, in order to have the desired semi-insulating behaviour. During this final heating process unwanted liquid droplets precipitate in the solid crystal due to misfits and due to supersaturation. These precipitates influence negatively mechanical and semi-insulating properties of the crystal. Their elimination, if possible, is a crucial point for the production of semi-insulators.

For the modelling of phase transitions various types of models have been suggested. Sharp-interface models and phase-field models, where the interface is smeared out in the latter, capture the spatial structure of a phase transition, while macroscopic models, like the LSW model (Lifshitz and Slyozov (1961); Wagner (1961)) or the mean field model, and BD models (Becker and Döring (1935)) do not. Sharp-interface models, phase-field models and macroscopic models are continuous diffusion models while the BD model is an atomic nucleation model. Macroscopic models may be justified rigorously as homogenization limits of sharp-interface models or of BD models for small droplet volume fraction.

We consider a mathematical model that describes the evolution of the precipitates including surface tension and bulk stresses on a macroscopic scale. It is obtained by homogenization of a sharp-interface model, derived from thermodynamical principles in Kimmerle (2009). We examine the corresponding control problem. While results exist for the optimal control of phase-field models, e.g. for the Allen-Cahn equation (Farshbaf-Shaker (2011a,b)) or Cahn-Hilliard equation (Hintermüller and Wegner (2011)), the control of a macroscopic model has not been considered so far as known by the author. Instead of the well-established LSW model our model comprises the microstructure of the crystal within the diffusion process, mechanical deformations within linear elasticity, and the fact that droplets with only a few atoms do not behave like a liquid. The latter is modelled by the introduction of a minimal droplet volume $V_{min} > 0$. Our model is a realistic model for phase transitions between liquid and solid including linear elasticity and is not only restricted in its applicability to semi-insulating GaAs.

We focus on the homogeneous version of this generalized LSW model, corresponding to the dilute scaling of droplet volume fraction. In the homogeneous LSW model the bulk is in quasi-static diffusion equilibrium. Different regimes for the motion of free boundaries, as volume-diffusion-controlled or interface-reaction-controlled interface motion can be considered. We assume homogeneous precipitates, i.e. the mass density and concentrations within the precipitates are constant.

2. CONTROL PROBLEM

We look for regimes, where for large times either only a few droplets, as small as possible, survive or where a homogeneous distribution $\nu_t(V)$ of the droplet volume V

may be achieved at a given final time t_f. This motivates the cost function

$$J(u_1,\nu_{t_f}) = \frac{\alpha_0}{2}\|u_1\|^2_{L^2(t_0,t_f)} + \int\limits_{V_{min}}^{\infty} (\alpha_1 + \alpha_2 V)\, d\nu_{t_f}(V)$$

$$+\frac{\alpha_3}{2}\int\limits_{V_{min}}^{\infty} \left| V \int\limits_{V_{min}}^{\infty} d\nu_{t_f}(S) - \int\limits_{V_{min}}^{\infty} S d\nu_{t_f}(S) \right|^2 d\nu_{t_f}(V) \quad (1)$$

where the non-negative weights α_k, $0 \leq k \leq 3$, with $\sum_k \alpha_k > 0$, may be chosen basically as needed within the application. Note, that the integral terms are evaluated at the end point t_f and only the control cost depends on the whole time interval (t_0, t_f). A natural parameter control is provided by physical quantities like temperature or pressure. An initial control is provided by the total mass M and the total arsenic $N_1 = MX_0/m_1$, m_1 the molar mass of As, that are conserved, and, in principle, by the initial volume distribution of droplets $\nu_0(V)$. Since the pressure yields only a slight correction (Kimmerle (2009)), and it is not clear whether it is technically possible to influence precisely the initial distribution of droplets, we focus on the total mass $u_0 = M$ and the temperature difference u_1 as control.

2.1 Volume-diffusion-controlled regime

While droplets are parametrized by V, the bulk is parametrized by a fictive volume $\overline{V}$ corresponding to the mean field of the chemical potential. It is convenient to consider the density $\nu_t(V)$ of droplet volume that is the ratio of droplets with volume between V_1 and V_2 at a fixed time t w.r.t. the total droplet volume at the initial time. The evolution of this time-dependent non-negative measure $\nu_t(V)$ is prescribed by the LSW equation

$$\partial_t \nu_t + a(\overline{V}, V, u_1)\, \partial_V \nu_t = 0 \text{ in } (V_{min}, \infty), \text{ a.e. in } [t_0, t_f], (2)$$

while droplets smaller than V_{min} do not exist, yielding

$$\nu_t(V) = 0 \text{ in } [0, V_{min}], \text{ in } [t_0, t_f]. \quad (3)$$

The term a expresses the speed how fast the volume of a droplet changes and is determined by the so-called Stefan condition, modelling the balance of mass/substance at the interface. It reads for the volume-diffusion-controlled regime

$$a(\overline{V}, V, u_1) = V^{1/3}\frac{\mu_I(\overline{V}, u_1) - \mu_I(V, u_1)}{\mathbb{X}(V, u_1)} \quad (4)$$

with a strictly positive function $\mathbb{X}$, monotone decreasing in V. Here $\mu_I(V, u_1)$ is the chemical potential of a precipitate, strictly monotone decreasing in the volume, with parameter u_1. From a it turns out that the chemical potential $\overline{\mu} = \mu_I(\overline{V}, u_1)$ is associated to the fictive volume $\overline{V}$ and droplets with chemical potential $\mu_I(V_i, u_1)$ smaller as $\overline{\mu}$ grow, while droplets with a larger chemical potential shrink. The function

$$\mathbb{X}(V,u) = \frac{u_1}{B^D}\left[\frac{\rho_S(\overline{V},u_1)}{\eta_S(\overline{V},u_1)}\eta_L(V,u_1) - \rho_L(V,u_1) + \left(\frac{\rho_S(\overline{V},u_1)}{\eta_S(\overline{V},u_1)}\partial_V \eta_L(V,u_1) - \partial_V \rho_L(V,u_1)\right) V\right] \quad (5)$$

results from the continuity of the flux of mass and substance over the interface. Note that the mass densities/concentrations ρ_L/η_L are evaluated on the liquid side of the interface, while ρ_S and η_S are evaluated on the solid side. B^D is a positive constant related to the mobility in the volume-diffusion-controlled regime.

Eq. (3) together with (4) implies that we switch from an ODE to the equation $\nu_t(V) = 0$ once a shrinking droplet reaches V_{min}.

The corresponding initial condition is

$$\nu(t_0, V) = \nu_0(V). \quad (6)$$

This is coupled to the global conservation of mass/substance yielding an algebraic equation for $\overline{V}$,

$$\overline{V} = \zeta\left(\frac{M - \int_{V_{min}}^{\infty} \rho_L(V,u_1)\, V d\nu_t(V)}{MX_0 - m_1 \int_{V_{min}}^{\infty} \eta_L(V,u_1)\, V d\nu_t(V)}, u_1\right) \quad (7)$$
$$\text{in } [t_0, t_f],$$

with a nonlinear, strictly monotone function ζ, defined by

$$\zeta^{-1}(\cdot, u_1) = \frac{\eta_S(\cdot, u_1)}{\rho_S(\cdot, u_1)}. \quad (8)$$

The index of the algebraic equation (7) is 1.

Let $\mathcal{C} := C_0^0(0,\infty)$ and $\mathcal{C}'$ denote its dual space. Our control problem is to find states

$$\{\nu_t, \overline{V}\} \in C^0_{weak}([t_0, t_f], \mathcal{C}') \times C^0([t_0, t_f], \mathbf{R}), \quad (9)$$

an initial control parameter

$$u_0 \in \mathbf{R}^+, \quad (10)$$

and a control

$$u_1 \in L^\infty([t_0, t_f], \mathbf{R}^+), \quad (11)$$

s.t. the cost functional J, given in (1), is minimized under respect of the initial value problem for our DAE system (2) – (8), the pure state constraints,

$$\begin{array}{ll} a)\ \nu_t(V) \geq 0 & \forall V \in \mathbf{R}_0^+\ \forall t \in [t_0, t_f], \\ b)\ \overline{V} \geq 0 & \forall t \in [t_0, t_f], \end{array} \quad (12)$$

and box constraints for the controls

$$\begin{array}{ll} u_{min,0} \leq u_0 \leq u_{max,0}, & \\ u_{min,1} \leq u_1(t) \leq u_{max,1} & \forall t \in [t_0, t_f], \end{array} \quad (13)$$

where $u_{min,j}$, $u_{max,j}$, $j = 0, 1$, are given strictly positive bounds.

2.2 Interface-reaction-controlled regime

In case of the interface-reaction-controlled regime we have again (2), (3), (6) – (13). We replace (4) by

$$a(\overline{V}, V, u_1) = V^{4/3}\frac{\mu_I(\overline{V}, u_1) - \mu_I(V, u_1)}{\mathbb{Z}(V, \overline{V}, u_1)} \quad (14)$$

where the local conservation of mass and substance at the interface between solid and a liquid droplet is encoded in the strictly positive function

$$\mathbb{Z}(V,\overline{V},u_1) = \frac{u_1}{B^I}\rho_L(\overline{V},V,u_1)^{1/2}\Bigg[(\tilde{\mu}-1) \\ -\Bigg((\tilde{\mu}-1)\frac{\partial_V \rho_L(\overline{V},V,u_1)}{\rho_L(\overline{V},V,u_1)} + \tilde{\mu}\frac{\partial_V \eta_L(\overline{V},V,u_1)}{\eta_L(\overline{V},V,u_1)} \\ -\frac{\partial_V\left(\rho_L(\overline{V},V,u_1) - m_1\eta_L(\overline{V},V,u_1)\right)}{\rho_L(\overline{V},V,u_1) - m_1\eta_L(\overline{V},V,u_1)}\Bigg)V\Bigg], \quad (15)$$

that is monotone decreasing in $\overline{V}$. $B^I > 0$ is a constant linked to the mobility in this regime and $\tilde{\mu}$ is the quotient of the molar mass of gallium and arsenide. In this regime the dynamics of the droplet volume distribution are driven by the jump between the chemical potentials between solid and liquid phase, while in the volume-diffusion-controlled regime it can be demonstrated that the volume evolution is related to the jump of the averaged normal derivative of the chemical potential. For further details of the mathematical modelling for the DAE system see Kimmerle (2009).

Summarized, the optimal control problem is to find the states (9) and the controls (10) and (11), minimizing (1) and satisfying the equations (2), (3), (6), and (7), under the constraints (12) and (13).

3. NUMERICAL SOLUTION

3.1 Mean field model

Numerically, we solve our problem by the method of characteristics. We consider a special case of our DAE system, where we assume a specific initial condition

$$\nu_0(V) = \frac{1}{\mathcal{N}_0}\sum_{i=1}^{\mathcal{N}_0}\delta_{V_i^0}(V), \quad (16)$$

i.e. we have initially a discrete finite number $\mathcal{N}_0$ of distinct volumes

$$V_i(t_0) = V_i^0 \quad \forall i \in \{1;...;\mathcal{N}_0\}. \quad (17)$$

This mean field model amounts to solving our original LSW equation for a finite number $\mathcal{N}_0$ of characteristics.

We introduce another unknown $\mathcal{N}(t)$ encoding the number of droplets at time t with $V > V_{min}$. Precipitates below V_{min} vanish. We introduce t_j as the first time when $V_j \leq V_{min}$. If a precipitate never disappears, i.e. $V_j > V_{min}$ for all times, then we set $t_j = \infty$. For ease of notation we assume w.l.o.g. $V_1 \geq V_2 \geq .. \geq V_{\mathcal{N}_0-1} > V_{\mathcal{N}_0}$, thus it turns out that $V_{\mathcal{N}_0}$ vanishes first and V_1 remains as last droplet.

In this situation our control problem simplifies for both regimes to the following evolution of precipitates

$$\begin{aligned} \partial_t V_i &= a(\overline{V},V_i,u) \text{ in } [t_0,t_f] \setminus \cup_{1\leq j\leq\mathcal{N}_0}\{t_j\}, \\ V_i(t+) &= V_i(t-) \quad\;\; \text{in } (\cup_{1\leq j\leq\mathcal{N}_0}\{t_j\}) \cap [t_0,t_f], \end{aligned} \quad (18)$$

for $V_i > V_{min}$, while

$$V_i = 0 \text{ in } [t_0,t_f], \quad (19)$$

for $V_i \leq V_{min}$. Here we keep record of droplets below V_{min}, contrary to our original model, in order not to change the number of states with time which turns out to be more suitable for numerics. The ODE system is completed by the initial condition (17) and the global conservation of mass/substance

$$\overline{V} = \zeta\left(\frac{M - \frac{1}{\mathcal{N}_0}\sum_{i=1}^{\mathcal{N}(t)}\rho_L(V_i,u_1)\,V_i}{MX_0 - m_1\frac{1}{\mathcal{N}_0}\sum_{i=1}^{\mathcal{N}(t)}\eta_L(V_i,u_1)\,V_i}, u_1\right) \quad (20)$$
$$\text{in } [t_0,t_f].$$

The state constraints read

$$\{V_i\}_{1\leq i\leq\mathcal{N}}, \overline{V} \geq 0 \quad (21)$$

and we have the constraints (13) for the controls. With (16) the cost function (1) reads

$$J(u_1,\nu_{t_f}) = \frac{\alpha_0}{2}\|u_1\|^2_{L^2(t_0,t_f)} + \frac{\alpha_1\mathcal{N}(t_f)}{\mathcal{N}_0} + \frac{\alpha_2}{\mathcal{N}_0}\sum_{i=1}^{\mathcal{N}_0}V_i(t_f) \\ +\frac{\alpha_3}{2}\frac{1}{\mathcal{N}_0}\sum_{i=1}^{\mathcal{N}_0}\left|V_i(t_f)\frac{\mathcal{N}_{t_f}}{\mathcal{N}_0} - \frac{1}{\mathcal{N}_0}\sum_{j=1}^{\mathcal{N}_0}V_j(t_f)\right|^2. \quad (22)$$

The DAE system (17) - (20) is called *mean field model* and (22), (17) - (21), (13) is referred to as *mean field optimal control problem.*

3.2 Numerical methods

We follow a direct method, meaning optimization of the time-discretized version of the mean field optimal control problem. We decompose the time interval s.t. $t_0 = t^{[0]} < t^{[1]} < ... < t^{[k]} < ... < t^{[n_d]} = t_f$, $1 \leq k \leq n_d$. The time discretization has to be chosen sensitively w.r.t. times t_j when droplets vanish. The discretized optimal control problem, resulting for the mean field model, is to find states $V_i(t^{[k]}) \in \mathbf{R}^+$, $1 \leq i \leq \mathcal{N}_0$, and controls $u_0 \in \mathbf{R}^+$ and $u_1(t^{[k]}) \in \mathbf{R}^+$, minimizing (22) and satisfying the discretized equations and constraints (17) – (21) and (13).

Under reasonable assumptions on the data $\mathbb{X}$ or $\mathbb{Z}$, and $\mu_I, \zeta, \rho_L, \eta_L$, $u_{min,\cdot}$, and $u_{max,\cdot}$ we may solve the resulting discretized control problem with jumps in the states and its derivatives, both occurring whenever a droplet disappears. The numerical results rely on data for GaAs, summarized in Dreyer and Kimmerle (2009). Note that the main influence of the control enters within (5) or (15). We study in the following the numerical solution in case of the volume-diffusion-controlled regime.

For the update of the states we use central differences. Our optimization algorithm follows a sensitivity-based approach. This has been implemented in OCPID-DAE1 V1.1, a software code developed by Gerdts (2011), that uses the SQP method as optimizer. In doing so we have two possibilities for the implementation of the DAE. Eq. (20) has index 1 and if we suppose a suitable initial condition

$$\overline{V}(t = t_0) = \overline{V}_0, \quad (23)$$

s.t. $\overline{V}_0$ fulfils (20), then we may replace the algebraic equation by an ODE. This explicit ODE for $\overline{V}$ is obtained by differentiation of (20) w.r.t. time, and reads

$$\partial_t\overline{V} = \frac{-\frac{1}{\mathcal{N}_0}\sum_{i=1}^{\mathcal{N}}V_i^{1/3}(\mu_I(\overline{V},u_1) - \mu_I(V_i,u_1))}{\mathcal{X}(\overline{V},u_1)(MX_0 - m_1\frac{1}{\mathcal{N}_0}\sum_{i=1}^{\mathcal{N}(t)}\eta_L(V_i,u_1)\,V_i)} \quad (24)$$
$$\text{in } [t_0,t_f] \setminus \cup_{1\leq j\leq\mathcal{N}_0}\{t_j\},$$

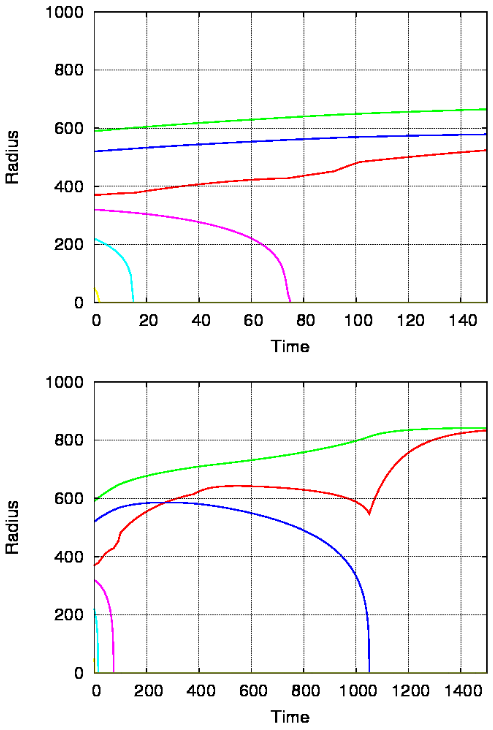

Fig. 1. Evolution of 5 droplet radii with initial radii 50 (yellow), 220 (cyan), 320 (magenta), 520 (blue), 590 (green), together with the bulk mean field radius (red) $[10^{-9}\,\mathrm{m}]$ vs. time $[1\,\mathrm{s}]$. Upper figure: short-time behaviour (up to 150 s), lower figure: long-time behaviour (up to 1500 s).

where

$$\mathcal{X}(\overline{V},u_1) = \frac{1}{\eta_S(\overline{V},u_1)} \times \left(\frac{\rho_S(\overline{V},u_1)}{\eta_S(\overline{V},u_1)} \partial_{\overline{V}} \eta_S(\overline{V},u_1) - \partial_{\overline{V}} \rho_S(\overline{V},u_1) \right). \tag{25}$$

After the times t_j when droplets vanish we use (20) in order to determine $\overline{V}(t_j+)$.

We solve our problem by means of

(i) a Runge-Kutta integrator with fixed but suitably small time step size, where the algebraic equation (20) has been replaced by an ODE for $\overline{V}$,

(ii) using the DASSL solver, see Brenan et al. (1996),
   a) keeping the algebraic equation (19),
   b) replacing the algebraic equation (19) by the ODE $\partial_t V_i = 0$ for $V_i \leq V_{min}$ with the initial condition $V_j(t_j+) = 0$.

The algorithms (i), (ii)a) and (ii)b) solve the original problem. However, the algorithms (i) and (ii)b) turn out to run more reliably for a large set of initial conditions, while (ii)a) exhibits occasionally problems determining the control at switching points. A critical point with (i) is, that the time step has to be chosen very small for certain data, since the times, when droplets vanish, have to be located as precisely as possible in order to avoid propagated errors. The algorithms based on DASSL turn out to be even more sensitive to the choice of too large time steps. Hence we present our results obtained by algorithm (i) in the following.

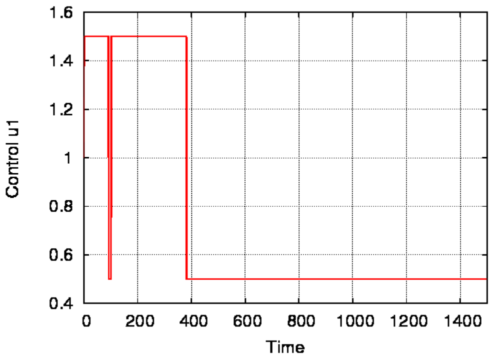

Fig. 2. Control by temperature $[10^2\,\mathrm{K}]$ vs. time $[1\,\mathrm{s}]$

3.3 Numerical results

We examine the different contributions to the cost function and the two controls and discuss their impact on the solution. For better illustration, we present our results for radii $r_i = 3/(4\pi)V_i^{1/3}$, corresponding to the special case of spherical precipitates. In case of $\alpha_2 = 1$, $\alpha_0 = \alpha_1 = \alpha_3 = 0$ in the cost function, Fig. 1 shows the time evolution for 5 droplet radii, together with the fictive bulk radius $\overline{r} = 3/(4\pi)\overline{V}^{1/3}$, and Fig. 2 presents the corresponding control. The initial control parameter turns out to be $u_0 = u_{min,0}$. In Fig. 3 we give the evolution of the volume fraction, that enters into the α_2-term of the cost function, but at the final time.

Further numerics suggest a control of bang-bang type for u_1 for relatively small α_0 and that the α_3-term, representing the deviation from the mean droplet volume within the cost functional, has no impact for large times t_f since the mean field $\overline{V}$ represents an unstable stationary point for the V_i. We conjecture that the α_2-term is the most controllable contribution within the cost functional. Furthermore, for sufficiently small time steps our numerical results seem to be mesh independent.

4. THEORETICAL ASPECTS AND OPEN QUESTIONS

Besides further numerical tasks, like more efficient algorithms for long-time behaviour e.g. by a suitable finite-volume discretization for LSW as in Carillo and Goudon (2001), also from a theoretical point of view the above stated optimal control problem exhibits very interesting aspects. The analysis of the classical LSW model without control has been treated in Niethammer and Pego (2000, 2001). Testing the LSW equation (2) yields a nonlocal hyperbolic conservation law in the dual space $\mathcal{C}$. Within optimal control theory there are several results

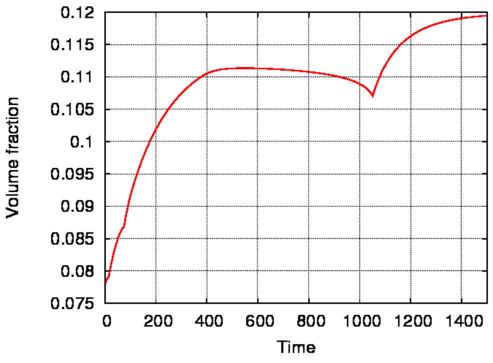

Fig. 3. Volume fraction of precipitates $\frac{1}{\mathcal{N}_0}\sum_{i=1}^{\mathcal{N}_0} V_i(t)$ $[10^{-18}\,\mathrm{m}^3]$ vs. time t $[1\,\mathrm{s}]$

(e.g. Colombo et al. (2011); Coron et al. (2010); Shang and Wang (2011); Gugat et al. (2006); Jacquet et al. (2006)), for a hyperbolic first-order equation, but as far as known by the author neither non-local conservation laws involving measures as states nor systems with switching between an ODE and an algebraic equation have been considered so far.

We summarize the main distinct features of our problem. We have a measure-valued solution (LSW) or switching conditions (Mean field model), the droplet volume is not bounded from above, but we do not observe shocks as they might occur typically due to nonlinearities of the flux function.

An issue for our optimal control problem is that it depends on the control, how many droplets vanish within a prescribed final time t_f. An adjoint based approach for this hybrid optimal control is an open question where the difficulty is due to the switching conditions. Finally, it would be interesting to consider the control of the inhomogeneous LSW equation where the equation depends weakly on the spatial structure, too.

The theoretical issues of the optimal control of LSW-type models are an important question and are work in progress. Problems of this type have applications also within a wider frame, e.g. highly re-entrant manufacturing systems (Coron et al. (2010)), traffic flow (Benzoni-Gavage et al. (2006)), two phase flow, gas dynamics, or aerospace dynamics.

ACKNOWLEDGEMENTS

The author would like to thank M. Gerdts for valuable comments regarding OCPID-DAE1.

REFERENCES

Becker, R. and Döring, W. (1935). Kinetische Behandlung der Keimbildung in übersättigten Dämpfen. *Ann. Physik*, 24, 719–752.

Benzoni-Gavage, S., Colombo, R.M., and Gwiazda, P. (2006). Measure valued solutions to conservation laws motivated by traffic modelling. *Proc. R. Soc. A*, 462(2070), 1791–1803.

Brenan, K.E., Campbell, S.L., and Petzold, L.R. (1996). *Numerical Solution of Initial-Value Problems in Differential-Algebraic Equations*. Number 14 in Classics in Applied Mathematics. SIAM, Philadelphia, PA.

Carillo, J.A. and Goudon, T. (2001). A numerical study on large-time asymptotics of the Lifshitz-Slyozov system. *INRIA*, 4287. Rapport de recherche.

Colombo, R.M., Herty, M., and Mercier, M. (2011). Control of the continuity equation with a non local flow. *ESAIM COCV*, 17(2), 353–379.

Coron, J.M., Kawski, M., and Wang, Z. (2010). Analysis of a conservation law modeling a highly re-entrant manufacturing system. *Discrete Contin. Dyn. Syst.-B*, 14(4), 1337–1359.

Dreyer, W. and Kimmerle, S.J. (2009). Mean field diffusion models for precipitation in crystalline GaAs including surface tension and bulk stresses. *Preprint Series of Weierstrass Institute for Applied Analysis and Stochastics, Berlin*, 1475.

Farshbaf-Shaker, M.H. (2011a). A penalty approach to optimal control of Allen-Cahn variational inequalities: MPEC-view. Preprint University of Regensburg No. 6/11.

Farshbaf-Shaker, M.H. (2011b). A relaxation approach to vector-valued Allen-Cahn MPEC problems. Preprint University of Regensburg No. 27/11.

Gerdts, M. (2011). OCPID-DAE1, Optimal control and parameter identification with differential-algebraic equations of index 1. User's guide (Online documentation).

Gugat, M., Herty, M., Klar, A., and Leugering, G. (2006). Conservation law constrained optimization based upon front-tracking. *ESAIM: M2AN*, 40(5), 939–960.

Hintermüller, M. and Wegner, D. (2011). Distributed optimal control of the Cahn-Hilliard system including the case of a double-obstacle homogeneous free energy density. Matheon-Preprint, to appear in SIAM J. Control and Optimization.

Jacquet, D., Krstic, M., and Canudas de Wit, C. (2006). Optimal control of scalar one-dimensional conservation laws. In *Proceedings of the 2006 American Control Conference, Minneapolis, Minnesota, USA, June 14–16, 2006*, 5213–5218.

Kimmerle, S.J. (2009). *Macroscopic diffusion models for precipitation in crystalline gallium arsenide*. Ph.D. thesis, Humboldt-Universität zu Berlin, Berlin.

Lifshitz, I.M. and Slyozov, V.V. (1961). The kinetics of precipitation from supersaturated solid solutions. *J. Phys. Chem. Solids*, 19(1/2), 35–50.

Niethammer, B. and Pego, R.L. (2000). On the initial-value problem in the Lifshitz-Slyozov-Wagner theory of Ostwald ripening. *SIAM J. Math. Anal.*, 31(3), 467–485.

Niethammer, B. and Pego, R.L. (2001). The LSW model for domain coarsening: Asymptotic behavior for conserved total mass. *J. Stat. Phys.*, 104(5/6), 1113–1143.

Shang, P. and Wang, Z. (2011). Analysis and control of a scalar conservation law modeling a highly re-entrant manufacturing system. *J. Differential Equations*, 250, 949–982.

Wagner, C. (1961). Theorie der Alterung von Niederschlägen durch Umlösen (Ostwald-Reifung). *Z. Elektrochem.*, 65(7/8), 581–594.

Chapter 5

Classification, Conclusion, and Outlook

We classify the four application examples in the next section w.r.t. the type of differential equations and compare with other examples in literature. In Section 5.2 we summarize our results in modelling, analysis, simulation, and optimal control. The last section provides an overview on near future projects and open tasks.

5.1 Comparison with Other Results on Coupled Ordinary and Partial Differential Equations

In the latter chapter we have considered four typical examples for optimal control with coupled (ordinary and partial) differential equations. We have discussed with different detailedness problems out of the following classes:

1a) ODEs and elliptic equations of second order (Problems 2 & 3),

1b) ODEs and elliptic equations of forth order (Problem 2b, case of plate equation),

2a) ODEs and hyperbolic first-order equations (Problems 1 & 4)[1],

2b) ODEs and hyperbolic second-order equations (mentioned in Problem 2a & Problem 2b),

3) ODEs and parabolic equations of second-order (Problem 1).

For further features of the discussed four coupled problems and other of my works, please see also Table 1.1 that illustrates which example fits to which situation. For the optimal control problems see Table 1.2.

We classify w.r.t. the type of differential equations all the examples for optimal control with CDE, as far as they are known to the author. The optimal control of a flexible satellite

[1] In Pb. 1 the PDE is treated finally as a parabolic equation of second order by the concept of viscosity solutions.

by Biswas and Ahmed [BA86, BA89] is subject to ODE and the beam equation, belonging to Class 1b). In the same class Biswas considered the optimal control of a gantry overhead crane [Bi04], where the crane beam is modelled as a 1D beam. In Class 3 we have the heat-optimal entry of a spacecraft into atmosphere [CPW+09] and the related toy problem of a hypersonic rocket car [WRP10, PRWW10, PRWW14, We14]. We annotate that further details on the modelling in [CPW+09] are given in [WC12]. The problem of laser surface hardening of steel [HS97, FHS01, HV03a, HV03b, GNP10], where an ODE for the phase transition and the heat equation are present, belongs also to Class 3. More complicated is the situation in [KR15, BK14, BK17] where the monodomain equation of cardiac electrophysiology is studied in the context of optimized defibrillation. Here the coupled system is the FitzHugh-Nagumo model (Example 3.8) that is related to Class 3.

Further CDE that I have considered are arsenide-rich droplets within gallium arsenide crystals (Classes 1a & 3) [Ki09, DK10, Ki11], the evolution of hydrogen nanobubbles in electrolysis (Class 3) [SKB17, KSB17], and the electrokinetic flow in deformable PEM nanochannels [LBKN11, BNK11, KBN13, KLNB14] that fits best to Class 1. The latter has been considered as a shape optimization problem in [BKN14]. Note that free boundary problems are related to shape optimization.

Of course, there are many other coupled problems that could be studied as optimal control problems as well, e.g., the problem of optimal trajectories for ferries crossing rivers or narrows, where an ODE is coupled to flow equations. The latter problem would correspond to another class.

5.2 Conclusion

5.2.1 Modelling Issues and Analytic Results

Since we wish neither to consider different ODEs for each space points nor this is realistic for our applications, we have introduced a so-called averaging-evaluation operator $\mathcal{E}$ in Def. 3.1. This point of view is a new concept to the best knowledge of the author. Certain smoothness properties of $\mathcal{E}$, e.g. in case of an integral-type operator, are essential for the further analysis of coupled systems.

The well-posedness (existence, uniqueness, and continuous dependence on data) of coupled ODE-PDE systems alone is in general not clear. For evolution problems, for sufficiently small times we may guarantee a strict contraction and apply the Banach fixed point theorem. This is the result of Th. 3.4 that may be applied directly or adopted to the truck-container problem (Problem 1), the elastic crane-trolley-load problem (Problem 2a) and the GaAs droplet problem. We expect that for suitable controls, we may extend the well-posedness local in time to a global well-posedness result. However, if the optimal control is, e.g., subject to control constraints that are too restrictive, this might not always be possible.

Concerning the first-optimize-then-discretize approach, there are the possibilities to treat the

ODEs like PDEs working in Hilbert spaces or to treat PDEs as DAEs in function spaces using spaces with functions and derivatives that are both essentially bounded. Though in theory these approaches may commute in certain cases, it turns out in our examples that it is favourable to treat ODEs like PDEs in spaces like L^2 or H^1.

Furthermore, we have demonstrated in Th. 3.17 that the coupling structure is reversed in the adjoint equations. This is a new observation that is relevant for CDE with a non-trivial coupling structure. This reversal is demonstrated explicitly for the truck-container problem, Problem 1, in the original and in the semi-discretized problem for a suitable discretization. [KG16]. For accuracy, it makes sense to exploit the latter observation when choosing a numerical scheme.

Finally, we mention that for FOTD approaches, the (formal) Lagrangian approach, common in optimization with PDE, (see, e.g., [Tr10, HPUU09]) competes against the Pontryagin minimum principle, e.g. [RZ99, Ge12], the latter more common for optimal control of DAE. Here we can say that the formal Lagrangian approach requires less strict assumptions, whereas the Pontryagin approaches allows for pointwise minimum principles. However, we encounter situations where both approaches coincide, see, e.g., the discussion of Pb. 3.18 at the end of Chapter 3.

5.2.2 Numerical Challenges in Simulation and Optimal Control

From our applications we see that suitable algorithms are required for coupled differential equations.

We observe that the adjoint equations involve mean values (Problems 1 & 2a), integro-differential equations (Problem 2a), and difficult coupling (Problems 1, 2 & 4). For the elastic crane-trolley-load problem, the elastic tyre-damper system (quarter car model), and the mean field droplet problem, it is not straightforward at all to formulate adjoint equations that guarantee to be solvable.

In addition terminal conditions may cause numerical issues. For the truck-container and the elastic crane-trolley-load problem we consider penalty methods, exploiting the concept of the augmented Lagrange function, though this may provoke an increase in computing times. However, it is not clear how terminal constraints could be addressed with a significant precision alternatively.

Another issue is the parameter sensitivity of the complex models. In particular in Problems 2 and 4 this is a challenge.

We illustrate the numerical complexity of the problems by the size of the discretized problems. Problem 1 is only 1D in space. A spatial discretization of $M = 20$ [KG16] or $M = 60$ [GK15] yields $N = 600$ or $N = 1500$ time steps due to the Courant-Friedrichs-Lewy condition. We observe in the numerics that it is required to consider at least $N = C_{CFL}M$, where $C_{CFL} \approx 25 - 30$, for a stable scheme. Thus the discretized grid is of size $C_{CFL}M^2 = 12\,000$ [KG16] or $= 90\,000$ [GK15] for a single simulation only. In [WGKG18] for the case $M = 20$ the number of unknown variables is $n_z = 28\,247$ and the number of constraints is $n_c = 27\,650$, whereas for $M = 50$ it is stated that $n_z = 160\,607$ and $n_c = 159\,110$. The complexity will increase significantly when going to a 2D Saint-Venant problem and 3d truck dynamics.

For the elastic crane-trolley-load problem, in [KGH18a] the FE system has about 3 200 cells (tetrahedrons) corresponding to about 21 100 degrees of freedom and we consider 100 – 500 time steps. For the numerical optimal control in [KGH18b] we consider about 5 400 cells but only 50 time steps. Concerning elastic bridge-load systems, for the 3D beam model we consider 5 400 cells and for the 2D plate equation model only 200 elements are required (on 20 m), whereas 40 time steps are sufficient.
For the quarter car model we have to work with even up to 61 503 nodes and 121 428 elements for justifying the Hertz stress approximation. However, using this approximation the numerical solution of the PDE is required only once. For the numerical ODE solution typically up to 20 000 time steps are considered.

Finally, for the mean field model between 5 characteristics [DK10, Ki12] and 100 characteristics [Ki15] (there for bulk and surface nanobubbles together) are considered for the PDE approximation. The time steps are determined adaptively and its number depends thus on the solution and varies a lot (up to several thousands).

In the simulation papers on nanobubbles [SKB17, KSB17] for the steady-state model a locally refined mesh of 3 021 or 720 784 FE, resp., is considered. This is due to the solution type that is delicately changing near the interface in this application.

Note that in the illustrative examples in [GHK17] only $32^2 = 1\,024$ elements or $64^2 = 4\,096$ elements, resp., are considered.

In order to reduce the computing times, model reduction techniques are important. For the quarter car model the Hertzian stress approximation is exploited, whereas for the GaAs droplets (as well as for the nanochannels [KBN13, KLNB14]) asymptotic analysis and homogenization techniques are used to derive macroscopic equations that are easier to handle, but still exhibit the main features of the problem. In particular, fast accurate algorithms are important, e.g., for Problem 1 we have applied our newly developed globalized semismooth Newton method and we have considered structure-exploiting SQP methods, using in particular exact information on derivatives. We observe that in many of the examples discussed in Chapter 4 we require higher-order methods, e.g. quadratic Lagrange elements and the Heun method, in order to obtain at least reasonable simulation results. Note that for PDE we have employed FDM (Lax-Friedrich in Problem 1, Runge-Kutta in Problem 4) as well as FEM (Problems 2 & 3).

5.3 Open Tasks

For illustration of what may happen in optimal control of CDE, we refer to the FitzHugh-Nagumo system treated by Casas, Ryll, and Tröltzsch [Ry16, CRT18] and by Breiten, Kunisch, and Rund [BK14, KR15, BK17]. In [Ry16, CRT18] interesting optimal solutions (like turning spirals and scroll rings depending on the dimension) are observed. Breiten *et al.* encounter new phenomenon, like the loss of null-controllability. A similar observation is made by Zuyev for the example of a rigid body with a thin elastic plate, for which it is shown that neither partial stability nor controllability is possible [Zu15, Ch. 6].

For our coupled model problem, Pb. 3.18, we have not discussed state constraints here. The case of state constraints might exhibit some other interesting phenomena. For Problem 2a we make the conjecture, that the bang-bang principle (cf. Def. 2.45) might be violated. The latter phenomenon has been already observed in [PRWW10]. However, this observation is so far only a conjecture and it might be only a numerical artifact that, however, should be taken care of. This has to be studied further.

Schemes, where FDTO and FOTD commute, allow to benefit of both the advantages of FDTO, like fewer assumptions on regularity, and FOTD, like mesh independence. A near future goal is to consider the commuting schemes of Crank-Nicolson type derived in [AF12] to the truck-container problem.

For the truck-container problem, Problem 1, the idea is to extend our model to the case, where the truck moves on a curve or on a bounded surface strip embedded in 3D (modelling a road in a landscape) together with a fluid plane that is described by the 2D Saint-Venant equations. It is also of interest to model a truck with a semitrailer, involving the drive dynamics of both the drawing vehicle and the semitrailer with the fluid container. Furthermore, we could examine as second control the damping coefficient of the spring-damper element (like the electrorheologic damper in Problem 3). The numerical techniques examined in [WGKG18] could be refined further. In particular, an adaptive time grid (refined at the left and at the right end of the container) makes sense for this problem.

For the elastic crane-trolley load problem discussed in Section 4.3 a self-evident project would be to replace the beam modelled in 3D by a plate equation in 2D or a cantilever beam in 1D and then to reconsider the optimal control problem. The latter should be compared with the results in [BA86, BA89]. For a 1D beam, on one hand, however, coupling effects at the contact area between trolley and beam are not included, on the other hand the 1D case might be interesting, also for Problem 2b, since the F-differentiability of the moving boundary conditions w.r.t. the trolley position (see [KGH18b]) could be treated by Fourier series then. Furthermore, this problem then might open the door for a FOTD approach of this problem. Since an adjoint-based approach might be out of reach here, a task is to use automatic differentiation for the calculation of sensitivities. From a modelling point of view it would be interesting to include the control of the rope length, too.

For the quarter car model, Problem 3, the well-posedness is one of the next logical steps. Interesting analytical questions, related to optimal control of variational inequalities, appear in the full model without the Hertzian approximation of the contact stress. Concerning the optimal control problem, it should be examined whether an adjoint-based approach is possible and efficient, since the averaging-evaluation operator $\mathcal{E}$ doesn't implicate a complicated structure contrary to Problem 2. The fully coupled optimal control problem is to be simulated allowing for a comparison with the Hertzian approximation.

A problem of the type of Problem 4 might be extended to an optimal control problem with switching costs. There might be some connection from switching costs to the disappearance of shrinking droplets. The mean field model exhibits also some similarities with optimal control of transport equations as they appear in applications like traffic, see e.g. [BKKM14]. However, in our example we do not encounter shocks that appear naturally in traffic as congestions or due to traffic lights. An extension of our optimal control problem to general hyperbolic conservation laws might be thus of interest. Another open problem is to examine the adjoint-based approach for Problem 4 that is non-standard due to the vanishing droplets.

There are further examples for coupled differential equations, see the introduction in Chapter 1, that may be considered as optimal control problems as well. For instance, the optimal control of the full re-entry problem of a spacecraft into atmosphere is the topic of a Munich Aerospace project with Prof. M. Gerdts, Prof. C. Mundt, and MSc. B. Pablos at Bundeswehr University Munich.

Finally, the big task remains to close or at least diminish the gap between theory and the applications, though this gap is related to the complexity of the coupled real-world problems. The next step is here the study of sufficient conditions for optimal control with coupled differential equations. The general controllability of coupled systems could be examined as well.

As discussed in Section 3.3 we may consider ODE/DAE as PDE/PDAE and *vice versa* and apply the established techniques for optimal control with ODE/DAE or PDE, resp. The question remains if the theory of PDAE that is still in progress may help to consider CDE in a coherent way. In particular, it has to be decided how to define the index of the algebraic equation in PDAE [MB99]. For linear time-invariant PDAE results have been obtained by Reis et *al.* [Re06, RT05].

One open question for optimal control of coupled ODE and PDE problems is, of course, the issue with state constraints, that is still open for optimal control of PDE, see Subsection 2.5.3. An interesting task would be to apply our globalized semismooth Newton algorithm [GHK17] to a Moreau-Yosida regularization of an elliptic optimal control problem with state constraints, where standard algorithms have troubles.

As an outlook we ask what are the differences and what are the commons of our coupled and of distributed systems, the latter being a whole subbranch of theoretical computer sciences. This might open the door for faster algorithms for coupled problems. Finally, the most important task will be the adaption, combination, and new invention of efficient algorithms for coupled differential equations allowing for reasonable computing times for applications. However, the real-time optimization might be for certain coupled systems (as this is already the case for many optimal control problems with PDE) out of reach, even when exploiting snapshots precomputed by proper orthogonal decomposition methods or reduced basis methods.

Bibliography

[Al16] H.W. Alt: *Linear Functional Analysis.* Springer, Berlin/Heidelberg, 2016.

[Al90] W. Alt: *The Lagrange-Newton method for infinite dimensional optimization problems.* Numer. Funct. Anal. Optim. 11 (1990), 201–224.

[Al91] W. Alt: *Sequential quadratic programming in Banach spaces.* In W. Oettli and D. Pallaschke (eds.): *Advances in Optimization*, pp. 281–301, Springer, Berlin, 1991.

[Al02] W. Alt: *Nichtlineare Optimierung.* Vieweg, Braunschweig/Wiesbaden, 2002.

[AF12] T. Apel and T. Flaig: *Crank-Nicolson schemes for optimal control problems with evolution equations.* SIAM J. Numer. Anal. 50 (2012), 1484–1512.

[AG01] M. Arnold and M. Günther: *Preconditioned dynamic iteration for coupled differential-algebraic systems.* BIT Numer. Math. 41 (2001), 1–25.

[BG16] M. Bäcker and A. Gallrein: *CDTire R4.0.0, Benutzerhandbuch.* Fraunhofer-Institut für Techno- und Wirtschaftsmathematik (ITWM), Kaiserslauten, 2016.

[Ba92] V. Barbu: *Analysis and Control of Nonlinear Infinite Dimensional Systems.* Academic Press, San Diego, 1992.

[BC16] G. Bastin, J.-M. Coron: *Stability and Boundary Stabilization of 1-D Hyperbolic Systems.* Progress in Nonlinear Differential Equations and Their Applications: Subseries in Control 88, Birkhäuser, Switzerland, 2016.

[BSS93] M.S. Bazaraa, H.D. Sherali, and C.M. Shetty: *Nonlinear Programming: Theory and Algorithms.* John Wiley & Sons, 1993, 2nd ed.

[B08] M. Bebendorf: *Hierarchical Matrices.* Lect. Notes Comput. Sci. Eng. 63, Springer, Berlin, 2008.

[BZ82] A. Ben-Zal and J. Zowe: *A unified theory of first and second order conditions for extremum problems in topological vector spaces.* In M. Guignard (ed.): *Optimality and Stability in Mathematical Programming*, pp. 39–76, Mathematical Programming Studies 19, Springer, Berlin/Heidelberg, 1982.

[BKN14] P. Berg, S.-J. Kimmerle, and A. Novruzi: *Modeling, shape analysis and computation of the equilibrium pore shape near a PEM-PEM intersection.* J. Math. Anal. Appl. 410 (2014), 241–256.

[BNK11] P. Berg, A. Novruzi, and S.-J. Kimmerle: *Mathematical Modelling of Ionomer-/Ionomer and Ionomer/Catalyst Layer Inferfaces - Iononer/Ionomer Interface*

Model. Final report of an industrial research project sponsored by Toyota Motor Corporation, UOIT, Oshawa, 2011.

[BIK99] M. Bergounioux, K. Ito, and K. Kunisch: *Primal-dual strategy for constrained optimal control problems.* SIAM J. Control Optim. 37 (1999), 1176–1194.

[BZ99] M. Bergounioux and H. Zidani: *Pontryagin maximum principles for optimal control of variational inequalities.* SIAM J. Control Optim. 37 (1999), 1273-1290.

[Be99] D.P. Bertsekas: *Nonlinear Programming.* Athena Scientific, 1999, 2nd ed.

[Be10] J.T. Betts: *Practical Methods for Optimal Control and Estimation Using Nonlinear Programming.* Adv. Des. Control, SIAM, Philadelphia, PA, 2010, 2nd ed.

[BA86] S.K. Biswas and N.U. Ahmed: *Stabilization of a class of hybrid systems arising in flexible spacecraft.* J. Optim. Theory Appl. 50 (1986), 83–108.

[BA89] S.K. Biswas and N.U. Ahmed: *Optimal control of large space structures governed by a coupled system of ordinary and partial differential equations.* Math. Control Signals Systems 2 (1989), 1–18.

[Bi04] S.K. Biswas: *Optimal control of gantry crane for minimum payload oscillations.* In: International Conference on Dynamic Systems and Applications, Atlanta, May 2024, 2003, Proc. Dynam. Systems Appl. 4 (2004).

[BFHK18] R. Boiger, A. Fiedler, J. Hasenauer, and B. Kaltenbacher: *Continuous analogue to iterative optimization for PDE-constrained inverse problems.* Inverse Probl. Sci. Eng. 27 (2019), 710–734.

[BC91] F. Bonnans and E. Casas: *Une principe de Pontryagine por le contrôle des systèmes semilinéaires elliptiques.* J. Differential Equations 90 (1991), 288–303.

[BC95] F. Bonnans and E. Casas: *An extension of Pontryagin's principle for state-constrained optimal control of semilinear elliptic equations and variational inequalities.* SIAM J. Control Optim. 33 (1995), 274–298.

[BS98] J.F. Bonnans and A. Shapiro: *Optimization problems with perturbations: a guided tour.* SIAM Rev. 40 (1998), 228–264.

[BS00] J.F. Bonnans and A. Shapiro: *Perturbation Analysis of Optimization Problems.* Springer Series in Operations Research, Springer, New York, 2000.

[BCG10] R. Borsche, R.M. Colombo, and M. Garavello: *On the coupling of systems of hyperbolic conservation laws with ordinary differential equations.* Nonlinearity 23 (2010), 2749–2770.

[BK16] R. Borsche and J. Kall: *High order numerical methods for networks of hyperbolic conservation laws coupled with ODEs and lumped parameter models.* J. Comput. Phys. 327 (2016), 678–699.

[BKKM14] R. Borsche, A. Klar, S. Kühn, and A. Meurer: *Coupling traffic flow networks to pedestrian motion.* Math. Models Methods Appl. Sci. 24 (2014), 221–244.

[BS12] A. Borzi and V. Schulz: *Computational Optimization of Systems Governed by Partial Differential Equations.* Series on Computational Science & Engineering 8, SIAM, Philadelphia, PA, 2012.

[Br07] D. Braess: *Finite Elements. Theory, Fast Solvers, and Applications in Solid Mechanics.* Cambridge University Press, Cambridge, UK, 2007, 3rd ed.

[BK14] T. Breiten and K. Kunisch: *Riccati-based feedback control of the monodomain equations with the FitzHugh-Nagumo model.* SIAM J. Contr. Opt. 52 (2014) 4057–4081.

[BK17] T. Breiten and K. Kunisch: *Compensator design for the monodomain equations with FitzHugh-Nagumo model.* ESAIM: Control, Optimisation and Calculus of Variations 23 (2017) 241–262.

[BCP96] K.E. Brenan, S.L. Campbell, and L.R. Petzold: *Numerical Solution of Initial-Value Problems in Differential-Algebraic Equations.* Classics Appl. Math. 14, SIAM, Philadelphia, PA, 1996.

[Ca93] B. Cai: *Neural Networks, Fuzzy Logic, and Optimal Control for Vehicle Active Systems with Four-Wheel Steering and Active Suspension.* Fortschrittsberichte VDI 12(196), VDI Verlag, Düsseldorf, 1993.

[CG01] J.A. Carillo, and T. Goudon: *A numerical study on large-time asymptotics of the Lifshitz-Slyozov system.* Rapport de recherche, INRIA 4287, 2001.

[Ca86] E. Casas: *Control of an elliptic problem with pointwise state constraints.* SIAM J. Control Optim. 4 (1986), 1309–1322.

[Ca97] E. Casas: *Pontryagin's principle for state-constrained boundary control problems of semilinear parabolic equations.* SIAM J. Contr. Optim. 35 (1997), 1297–1327.

[CRT08] E. Casas, J.C. de los Reyes, and F. Tröltzsch: *Sufficient second-order optimality theory conditions for semilinear control problems with pointwise state constraints.* SIAM J. Optim. 19 (2) (2008), 616–643.

[CRT18] E. Casas, C. Ryll, and F. Tröltzsch: *Optimal control of a class of reaction-diffusion systems.* Comput. Optim. Appl. 70 (2018), 677–707.

[Ci78] P.-G. Ciarlet: *The Finite Element Method for Elliptic Problems.* Stud. Math. Appl. 4, North-Holland, New York, 1978.

[Ci98] P.-G. Ciarlet: *Mathematical Elasticity, vol. 1: Three Dimensional Elasticity.* Stud. Math. Appl. 20, Elsevier, Amsterdam, 1998.

[CG11] J. Chen and M. Gerdts: *Numerical solution of control-state constrained optimal control problems with an inexact smoothing Newton method.* IMA J. Numer. Anal. 31 (2011), 1598–1624.

[CG12] J. Chen and M. Gerdts: *Smoothing technique of nonsmooth Newton methods for control-state constrained optimal control problems.* SIAM J. Numer. Anal. 50 (2012), 1982–2011.

[CPW+09] K. Chudej, H. J. Pesch, M. Wächter, G. Sachs, and F. Le Bras: *Instationary heat-constrained trajectory optimization of a hypersonic space vehicle by ODE-PDE-constrained optimal control.* In: G. Buttazzo and A. Frediani (eds.): *Variational Analysis and Aerospace Engineering*, pp. 127–144, Springer Optim. Appl. 33, Springer, New York, 2009.

[CBH+11] K.D. Cole, J.V. Beck, A. Haji-Sheikh, and B. Litkouhi: *Heat Conduction Using Green's Functions.* Series in Computational Methods and Physical Processes in Mechanics and Thermal Sciences, CRC Press, Boca Raton, 2011, 2nd ed.

[Co07] J.-M. Coron: *Control and Nonlinearity.* Math. Surveys Monogr. 136. Amer. Math. Soc., Providence, RI, 2007.

[CKP17] S. Court, K. Kunisch, and L. Pfeiffer: *Hybrid optimal control problems for a class of semilinear parabolic equations.* Discrete Contin. Dyn. Syst. Ser. S 11 (2018), 1031–1060.

[CKP18] S. Court, K. Kunisch, and L. Pfeiffer: *Optimal control problem for systems of conservation laws, with geometric parameter, and application to the Shallow-Water Equations.* Preprint, arXiv:1801.07205.

[AKM14] C. D'Apice, P.I. Kogut, and R. Manzo: *On relaxation of state constrained optimal control problem for a PDE-ODE model of supply chains.* Networks Heterogeneous Networks 9 (2014), 501–518.

[DM74] J.E. Dennis and J.J. Moré: *A characterization of superlinear convergence and its application to quasi-Newton methods.* Math. Comp. 28 (1974), 549–560.

[DM77] J.E. Dennis and J.J. Moré: *Quasi-Newton methods, methods, motivation and theory.* SIAM Rev. 19 (1977), 46–89.

[DS83] J.E. Dennis and R.B. Schnabel: *Numerical Methods for Unconstrained Optimization and Nonlinear Equations.* Prentice Hall, Englewood Cliffs, 1983.

[DR14] A.L. Dontchev and R.T. Rockafellar: *Implicit Functions and Solution Mappings. A View from Variational Analysis.* Springer Ser. Oper. Res. Financ. Eng., Springer, New York, 2014, 2nd ed.

[DPR99] F. Dubois, N. Petit, and P. Rochon: *Motion Planning and Nonlinear Simulations for a Tank Containing a Fluid.* In Proc. of the 5th European Control Conf. (ECC 99), Karlsruhe, 31.08.-03.09.1999, Düsseldorf, 1999.

[DK10] W. Dreyer and S.-J. Kimmerle: *Mean field diffusion models for precipitation in crystalline GaAs including surface tension and bulk stresses.* Preprint No. 1475, Weierstrass Institute for Applied Analysis and Stochastics, Berlin, 2009 / Preprint No. 698, DFG Research Center MATHEON - Mathematics for Key Technologies, Berlin, 2010.

[EEO+15] S. Eisenhofer, M.A. Efendiev, M. Ôtani, S. Schulz, and H. Zischka: *On an ODE-PDE coupling model of the mitochondrial swelling process.* Discrete Contin. Dyn. Syst. Ser. B 20 (2015), 1031–1057.

[EG04] A. Ern and J.-L. Guermond: *Theory and Practice of Finite Elements.* Springer, Berlin, 2004.

[Ev10] L.C. Evans: *Partial Differential Equations.* Grad. Stud. Math. 19, Amer. Math. Soc., Providence, RI, 2010, 2nd ed.

[FM68] A.V. Fiacco and G.P. McCormick: *Nonlinear Programming: Sequential Unconstrained Minimization Techniques.* John Wiley & Sons, New York/London/Sydney, 1968.

[FHS01] J. Fuhrmann, D. Hömberg, and J. Sokolowski: *Modeling, simulation and control of laser heat treatments.* In K.H. Hoffmann, I. Lasiecka, G. Leugering, J. Sprekels, and F. Tröltzsch (eds.): *Optimal Control of Complex Structures*, pp. 71–82, Internat. Ser. Numer. Math. 139, Birkhäuser, Basel, 2001.

[GKW14] M.J. Gander, F. Kwok, and G. Wanner: *Constrained optimization: from Lagrangian mechanics to optimal control and PDE constraints.* In: R. Hoppe (ed.): *Optimization with PDE Constraints: ESF Networking Program 'OPTPDE'*, Lecture Notes in Science and Engineering 101, Springer Intern. Publishing, Switzerland, 2014.

[GK99] C. Geiger and C. Kanzow: *Numerische Verfahren zur Lösung unrestringierter Optimierungsaufgaben.* Springer, Berlin, 1999

[GK02] C. Geiger and C. Kanzow: *Theorie und Numerik restringierter Optimierungsaufgaben.* Springer, Berlin, 2002.

[Ge03] M. Gerdts: *A moving horizon technique in vehicle simulation.* Z. Angew. Math. Mech. 83 (2003), 147–162.

[Ge05] M. Gerdts: *Local minimum principle for optimal control problems subject to index one differential-algebraic equations.* Technical report, Universiät Hamburg, 2005.

[Ge10] M. Gerdts: *OCPID-DAE1, Optimal Control and Parameter Identification with Differential-Algebraic Equations of Index 1.* Users Guide (Online Documentation), Universität der Bundeswehr München, Neubiberg/München, 2010. Available at `http://www.optimal-control.de/index.php/software`.

[Ge12] M. Gerdts: *Optimal Control of ODEs and DAEs.* DeGruyter, Berlin, 2012.

[GGP08] M. Gerdts, G. Greif, and H.J. Pesch: *Numerical optimal control of the wave equation: optimal boundary control of a string to rest in finite time.* Math. Comput. Simulation 79 (2008), 1020–1032. Part of the special issue: 5th Vienna International Conference on Mathematical Modelling/Workshop on Scientific Computing in Electronic Engineering of the 2006 International Conference on Computational Science/Structural Dynamical Systems: Computational Aspects, edited by I. Troch, F. Breitenecker, Y. Li, N. del Buono, L. Lopez, and T. Politi.

[GHK17] M. Gerdts, S. Horn, and S.-J. Kimmerle: *Line search globalization of a semismooth Newton method for operator equations in Hilbert spaces with applications in optimal control.* J. Ind. Manag. Optim. 13 (2017), 47–62.

[GH11] M. Gerdts and B. Hüpping: *Erratum: Global convergence of a nonsmooth Newton's method for control-state constrained optimal control problems (SIAM J. Optim., 19 (2008), 326–350).* Technical report, Universität der Bundeswehr München, Neubiberg (2011).

[GK15] M. Gerdts and S.-J. Kimmerle: *Numerical optimal control of a coupled ODE-PDE* *model of a truck with a fluid basin.* Discrete Contin. Dyn. Syst. 2015 (2015), 515–524.

[GL11] M. Gerdts and F. Lempio: *Mathematische Optimierungverfahren des Operations Research.* DeGruyter, Berlin, 2011.

[GT72] M. Golomb and R.A. Tapia: *The metric gradient in normed linear spaces.* Numer. Math. 20 (1972), 115–124.

[GR94] C. Großmann and H.-G. Roos: *Numerik partieller Differentialgleichungen.* Teubner-Studienbücher, B.G. Teubner, Stuttgart, 1994, 2. Aufl.

[GP17] L. Grüne and J. Pannek: *Nonlinear Model Predictive Control. Theory and Algorithms.* Springer, London, 2017, 2nd ed.

[GNP10] N. Gupta, N. Nataraj, and A.K. Pani: *On the optimal control problem of laser surface hardening.* Int. J. Numer. Anal. Model. 7 (2010), 667–680.

[H85] W. Hackbusch: *Multi-Grid Methods and Applications.* Springer, Berlin 1985.

[HLG06] E. Hairer, C. Lubich, and G. Wanner: *Geometric Numerical Integration. Structure-Preserving Algorithms for Differential Equations.* Springer Ser. Comput. Math. 31, Springer, Berlin/Heidelberg, 2006.

[HNW93] E. Hairer, S.P. Nørsett, and G. Wanner: *Solving Ordinary Differential Equations I. Nonstiff Problems.* Springer Ser. Comput. Math. 8, Springer, Berlin/Heidelberg, 1993.

[HW96] E. Hairer and G. Wanner: *Solving Ordinary Differential Equations II. Stiff and Differential-Algebraic Problems.* Springer Ser. Comput. Math. 14, Springer, Berlin/Heidelberg, 1996.

[HHW18] D. Hartmann, M. Herz, and U. Wever: *Model order reduction a key technology for digital twins.* In: W. Keiper, A. Milde, and S. Volkwein (eds.): *Reduced-Order Modeling (ROM) for Simulation and Optimization – Powerful Algorithms as Key Enablers for Scientific Computing*, Springer, Cham, 2018.

[HMR10] R. Herzog (Griesse), N. Metla, and A. Rösch: *Local quadratic convergence of SQP for elliptic optimal control problems with nonlinear mixed control-state constraints.* Control Cybernet. 39 (2010), 717–738.

[HV04] R. Herzog (Griese) and S. Volkwein: *A semi-smooth Newton method for optimal boundary control of a nonlinear reaction-diffusion system.* In: Proceedings of the Sixteenth International Symposium on Mathematical Theory of Networks and Systems (MTNS), Leuven, Belgium, July 5-9, 2004.

[He66] M.R. Hestenes: *Calculus of Variations and Optimal Control Theory.* John Wiley & Sons, New York, 1966.

[He75] M.R. Hestenes: *Optimization Theory - The finite dimensional case.* John Wiley & Sons, New York, 1975.

[Hi97] H. Hinsberger: *Ein direktes Mehrzielverfahren zur Lösung von Optimalsteuerungsproblemen mit großen, differential-algebraischen Gleichungssystmen und Anwendungen aus der Verfahrenstechnik.* PhD thesis, Institut für Mathematik, Technische

Universität Clausthal, 1997.

[HIK03] M. Hintermüller, K. Ito, and K. Kunisch: *The primal-dual active set strategy as a semismooth Newton method.* SIAM J. Optim. 13(3) (2003), 865–888.

[HK06a] M. Hintermüller and K. Kunisch: *Feasible and noninterior path-following in constrained minimization with low multiplier regularity.* SIAM J. Control Optim. 45 (2006), 1198–1221.

[HK06b] M. Hintermüller and K. Kunisch: *Path-following methods for a class of constrained minimization problems in function space.* SIAM J. Optim. 17 (2006), 159–187.

[HPUU09] M. Hinze, R. Pinnau, M. Ulbrich, and S. Ulbrich: *Optimization with PDE constraints.* Math. Model. Theory Appl. 23, Springer, Dordrecht, 2009.

[HS97] D. Hömberg and J. Sokolowski: *Optimal control of laser hardening.* Adv. Math. Sci. Appl. 8 (1998), 911–928.

[HV03a] D. Hömberg and S. Volkwein: *Laser surface hardening using proper orthogonal decomposition for a three-dimensional example.* Proc. Appl. Math. Mech. 3 (2003), 1–4.

[HV03b] D. Hömberg and S. Volkwein: *Control of Laser Surface Hardening by a Reduced-order Approach using Proper Orthogonal Decomposition.* Math. Comp. Modelling 38 (2003), 1003–1028.

[ILWW18] A. Ilchmann, L. Leben, J. Witschel, and K. Worthmann: *Optimal control of differential-algebraic equations from an ordinary differential equation perspective.* Optim. Contr. Appl. Meth. 40 2019, 351–366 .

[IJ15] K. Ito and B. Jin: *Inverse Problems. Tikhonov Theory and Algorithms.* Ser. Appl. Math. 22, World Sci. Publ., Singapore, 2015.

[IK08] K. Ito and K. Kunisch: *Lagrange Multiplier Approach to Variational Problems and Applications.* Adv. Des. Control 15, SIAM, Philadelphia, PA, 2008.

[J00] W. Jäger: *Skript zur Analysis I - III (WS 1998/99 – SS 2000.* Ruprecht-Karls-Universität Heidelberg, Heidelberg, 2000.

[FJ48] F. John: *Extremum problems with inequalities as subsidiary conditions.* In: K. Friedrichs, O. Neugebauer, J.J. Stoker (ed.): *Studies and Essays. Courant Anniversary Volume*, pp. 187–204, Wiley, 1948.

[Ke99] C.T. Kelley: *Iterative Methods For Optimization.* SIAM, Philadelphia, PA., 1999.

[KS94] C.T. Kelley and E.W. Sachs: *Multilevel algorithms for constrained compact fixed point problems.* SIAM J. Sci. Comput. 15 (1994), 645–667.

[Ki09] S.-J. Kimmerle: *Macroscopic Diffusion Models for Precipitation in Crystalline Gallium Arsenide - Modelling, Analysis and Simulation.* PhD thesis, Humboldt-Universität zu Berlin, Berlin, 2009.

[Ki11] S.-J. Kimmerle: *Well-posedness of a coupled quasilinear parabolic and elliptic free boundary problem from a model for precipitation in crystalline solids.* Preprint, Universität der Bundeswehr München, Neubiberg/München, 2011.

[Ki12] S.-J. Kimmerle: *Optimal Control of Mean Field Models for Phase Transitions.* In: 7th Vienna International Conference on Mathematical Modelling, Vienna, Austria, February 14-17, 2012, IFAC Proceedings Volumes 45/2 (2012), 1107–1111.

[Ki15] S.-J. Kimmerle: *Modelling, Simulation and Stability of Free Surface and Bulk Nanobubbles in Hydrogen Electrolysis.* In: 8th Vienna International Conference on Mathematical Modelling, Vienna, Austria, February 18-20, 2015, IFAC-PapersOnLine 48/1 (2015), 621–626.

[Ki16] S.-J. Kimmerle: *Modelling, Simulation and Optimization of an Elastic Structure under Moving Loads.* Proc. Appl. Math. Mech. 16 (2016), 697–698.

[Ki17] S.-J. Kimmerle: *Finite Elemente. Skript zur Vorlesung.* (In German.) Universität der Bundeswehr München, Neubiberg/München, 2017, 3rd ed.

[KG16] S.-J. Kimmerle, M. Gerdts: *Necessary optimality conditions and a semi-smooth Newton approach for an optimal control problem of a coupled system of Saint-Venant equations and ordinary differential equations.* Pure Appl. Funct. Anal. 1 (2016), 231–256.

[KGH17] S.-J. Kimmerle, M. Gerdts, and R. Herzog: *Optimal control of an elastic crane-trolley-load system - a case study for optimal control of coupled ODE-PDE systems. Preprint version with two additional appendices.* Available for download at http://www.unibw.de/sven-joachim.kimmerle, Universität der Bundeswehr München, Neubiberg/München, 2017.

[KGH18a] S.-J. Kimmerle, M. Gerdts, and R. Herzog: *Optimal control of an elastic crane-trolley-load system - a case study for optimal control of coupled ODE-PDE systems.* Math. Comput. Model. Dyn. Syst. 24 (2018), 182–206.

[KGH18b] S.-J. Kimmerle, M. Gerdts, and R. Herzog: *An optimal control problem for a rotating elastic crane-trolley-load system.* In: 9th Vienna International Conference on Mathematical Modelling, Vienna, Austria, February 21-23, 2018, IFAC-PapersOnLine 51/2 (2018), 272–277.

[KBN13] S.-J. Kimmerle, P. Berg, and A. Novruzi: *An electrohydrodynamic equilibrium shape problem for polymer electrolyte membranes in fuel cells.* In: System Modeling and Optimization, 25th IFIP TC 7 Conference on System Modeling and Optimization (CSMO 2011), Berlin, Germany, September 12-16, 2011, Revised Selected Papers, Springer, Heidelberg. IFIP Adv. Inf. Commun. Technol. 391 (2013), 387–396.

[KLNB14] S.-J. Kimmerle, K. Ladipo, A. Novruzi, and P. Berg: *Contact resistance at PEM-PEM interfaces: charged fluid flow between nanochannels.* ECS Transactions 59 (2014), 145–159.

[KM14] S.-J. Kimmerle and R. Moritz: *Optimal control of an elastic tyre-damper system with road contact.* Proc. Appl. Math. Mech. 14 (2014), 875–876.

[KSB17] S.-J. Kimmerle, K. Sverdrup, and P. Berg: *Dynamic equilibrium of a coupled ODE-PDE problem for surface nanobubbles.* Proc. Appl. Math. Mech. 17 (2017), 843–844.

[KW18] S.-J. Kimmerle and K. Worthmann: *Model predictive control of a coupled ODE-PDE model of a truck with a fluid basin.* Preprint, Universität der Bundeswehr München, Neubiberg/München, 2018, in preparation.

[Ki18] S.-J. Kimmerle: *Coupling structure in an optimal control problem with PDEs and ODEs.* Proc. Appl. Math. Mech. 18 (2018), 1–2.

[KA00] P. Knabner and L. Angermann: *Numerik partieller Differentialgleichungen.* Springer, Berlin, 2000.

[KS03a] M. Kočvara and M. Stingl: *Pennon: A code for convex nonlinear and semidefinite programming.* Optim. Methods Softw. 18(3) (2003), 317–333.

[KS03b] M. Kočvara and M. Stingl: *PENNON: A generalized augmented Lagrangian method for semidefinite programming.* In G. Di Pillo et al. (ed.): *High performance algorithms and software for nonlinear optimization*, pp. 303-321, selected lectures presented at the workshop, Erice, Italy, June 30–July 8, 2001, Appl. Optim. 82, Kluwer Academic Publishers, Boston, MA, 2003.

[Kr11] A. Kröner: *Numerical Methods for Control of Second Order Hyperbolic Equations.* PhD thesis, TU Munich, München, 2011.

[Kr13] A. Kröner: *Semi-smooth Newton methods for optimal control of the dynamic Lamé system with control constraints.* Numer. Funct. Anal. Optim. 34 (2013), 741–769.

[Kr97] D. Kröner: *Numerical Schemes for Conservation Laws.* Wiley, Chichester / Teubner, Stuttgart, 1997.

[KLST09] K. Kunisch, G. Leugering, J. Sprekels, and F. Tröltzsch (eds.): *Optimal Control of Coupled Systems of Partial Differential Equations.* Internat. Ser. Numer. Math. 158, Birkhäuser, Basel, 2009.

[KR15] K. Kunisch and A. Rund: *Time optimal control of the monodomain model in cardiac electrophysiology.* IMA J. Appl. Math. 80 (2015), 1664–1683.

[KW13] K. Kunisch and L. Wang: *Time optimal control of the heat equation with pointwise control constraints.* ESAIM Control Optim. Calc. Var. 19 (2013) 460–485.

[LBKN11] K. Ladipo, P. Berg, S.-J. Kimmerle, and A. Novruzi: *Effects of radially-dependent parameters on proton transport in polymer electrolyte nanopores.* J. Chem. Phys. 134 (2011), 074103-1–12.

[La03] I. Lasiecka: *Mathematical Control Theory of Coupled PDEs.* Appl. Mech. Rev.56(2003), B3.

[LT00] I. Lasiecka and R. Triggiani, *Control Theory for Partial Differential Equations.* Cambridge University Press, Cambridge, 2000.

[La06] P. D. Lax: *Hyperbolic Partial Differential Equations.* Courant Lect. Notes Math. 14. Amer. Math. Soc., Providence, RI, 2006.

[LEG+12] G. Leugering, S. Engell, A. Griewank, M. Hinze, R. Rannacher, V. Schulz, M. Ulbrich, and S. Ulbrich (eds.): *Constrained Optimization and Optimal Control for Partial Differential Equations.* Internat. Ser. Numer. Math. 160, Birkhäuser/Springer, Basel, 2012.

[Li68] J.L. Lions: *Contrôle optimal de systèmes gouvernés par des équations aux dérivées partielles.* Dunod, Paris, 1968.

[LM72] J.L. Lions and E. Magenes: *Nonhomogeneous Boundary Value Problems and Applications.* Vol. 1–3. Springer, Berlin, 1972.

[LN89] D. Liu and J. Nocedal: *On the limited memory BFGS method for large scale optimization.* Math. Programm. 45 (1989), 503–528.

[LMW12] A. Logg, K.A. Mardal, and G.N. Wells (eds.): *Automated Solution of Differential Equations by the Finite Element Method, The FEniCS Book.* Springer, Heidelberg, 2012.

[LY08] D.G. Luenberger and Y. Ye: *Linear and Nonlinear Programming.* Springer, New York, 2008, 3rd ed.

[MB99] W.S. Martinson and P.I. Barton: *A differentiation index for partial differential-algebraic equations.* SIAM J. Sci. Comput. 21 (1999), 2295–2315.

[Ma81] H. Maurer: *First and second order sufficient optimality conditions in mathematical programming and optimal control.* Math. Program. Study 14 (1981), 163–177.

[MM00] H. Maurer and H.D. Mittelmann: *Optimization techniques for solving elliptic control problems with control and state constraints. Part 1: boundary control.* Comput. Optim. Appl. 16 (2000), 29–55.

[MM01] H. Maurer and H.D. Mittelmann: *Optimization techniques for solving elliptic control problems with control and state constraints. Part 2: distributed control.* Comput. Optim. Appl. 18 (2001), 141–160.

[MZ79] H. Maurer and J. Zowe: *First- and second-order conditions in infinite-dimensional programming problems.* Math. Program. 16 (1979), 98–110.

[MPT07] C. Meyer, U. Prüfert, and F. Tröltzsch: *On two numerical methods for state-constrained elliptic control problems.* Optim. Methods Softw. 22 (2007), 871–899.

[MG14] J. Michael and M. Gerdts: *Optimal control in proactive chassis dynamics - A fixed step size time-stepping scheme for complementarity problems.* In M. Fontes, M. Günther, and N. Marheineke (eds.): *Progress in Industrial Mathematics at ECMI 2012*, pp. 271–276, Math. Ind. 19, Springer International Publishing, Switzerland, 2014.

[MG14] M.L.D. Monache and P. Goatin: *A front tracking method for a strongly coupled PDE-ODE system with moving density constraints in traffic flow.* Discrete Contin. Dyn. Syst. Ser. S 7 (2014), 435–447.

[Mo13] R. Moritz: *Simulation und Optimierung eines Reifens in Fahrbahnkontakt.* Master thesis. Universität der Bundeswehr München, Neubiberg/München, 2013.

[Ni99] B. Niethammer: *Derivation of the LSW theory for Ostwald ripening by homogenization methods.* Arch. Rational Mech. Anal. 147 (1999), 119–178.

[NO01] B. Niethammer and F. Otto: *Ostwald Ripening: The screening length revisited.*

Calc. Var. Partial Differential Equations 13 (2001), 33–68.

[NP00] B. Niethammer and R.L. Pego: *On the initial-value problem in the Lifshitz-Slyozov-Wagner theory of Ostwald ripening.* SIAM J. Math. Anal. 31 (2000), 467–485.

[NP01] B. Niethammer and R.L. Pego: *The LSW model for domain coarsening: Asymptotic behavior for conserved total mass.* J. Stat. Phys. 104 (2001), 11131143.

[NW06] J. Nocedal and S. Wright: *Numerical Optimization.* Springer Series in Operations Research and Financial Engineering, Springer, New York, 2006.

[OR79] L.A. Oganesyan and L.A. Rukhovets: *Variational-difference Methods for the Solution of Elliptic Equations. (In Russian.)* Izd. Akad. Nauk Armyanskoi SSR, Jerevan, 1979.

[OR11] J.T. Oden and J.N. Reddy: *An Introduction to the Mathematical Theory of Finite Elements.* Dover Publications, Mineola, NY, 2011, 2nd ed.

[PT12] P. Pedregal and J. Tiago: *Existence results for optimal control problems with some special nonlinear dependence on state and control.* SIAM J. Contr. Optim. 48 (2009), 415–437.

[PB94] H.J. Pesch and R. Bulirsch: The maximum principle, Bellman's equation, and Carathéodory's work. J. Optim. Theory Appl. 80 (1994), 199–225.

[PRWW10] H.J. Pesch, A. Rund, W. Von Wahl, and S. Wendl: *On some new phenomena in state-constrained optimal control if ODEs as well as PDEs are involved*, Control Cybern. 39 (2010) 647–660.

[PRWW14] H.J. Pesch, A. Rund, W. Von Wahl, and S. Wendl: *On a Prototype Class of ODE-PDE State-constrained Optimal Control Problems Part 1: Analysis of the State-unconstrained Problems*, Preprint, Universität Bayreuth, 2014.

[PBGM64] L.S. Pontryagin, V.G. Boltyanskij, R.V. Gamkrelidze, and E.F. Mischenko: *Mathematische Theorie optimaler Prozesse.* Oldenbourg-Verlag, München, 1964.

[RZ99] J.P. Raymond and H. Zidani: *Hamiltonian Pontryagin's Principles for Control Problems Governed by Semilinear Parabolic Equations.* Appl. Math. Optim. 39 (1999), 143–177.

[Re06] T. Reis: *Systems Theoretic Aspects of PDAEs and Applications to Electrical Circuits.* PhD thesis, Institut für Mathematik, Technische Universität Kaiserslautern, 2006.

[RT05] T. Reis and C. Tischendorf: *Frequency domain methods and decoupling of linear infinite dimensional differential algebraic systems.* J. Evol. Equ. 5 (2005), 357–385.

[RS05] U. Rettig and O. von Stryk: *Optimal and Robust Damping Control for Semi-Active Vehicle Suspension.* In: Proc. 5th EUROMECH Nonlinear Dynamics Conference (ENOC 2005), Eindhoven, The Netherlands (2005), pp. paper no. 20-316.

[Ro76] M.S. Robinson: *Stability theory for systems of inequalities in nonlinear programming, part II: differentiable nonlinear systems.* SIAM J. Num. Anal. 13 (1976), 497–513.

[Ru12] A. Rund: *Beiträge zur Optimalen Steuerung partiell-differential algebriascher Gle-*

ichungen. PhD thesis, Universität Bayreuth, Bayreuth, 2012.

[Ry16] C. Ryll: *Optimal control of patterns in some reaction-diffusion systems.* PhD thesis, Technical University of Berlin, 2016. Available at `https://doi.org/10.14279/depositonce-5712`.

[SW08] A. Schiela and M. Weiser: *Superlinear convergence of the control reduced interior point method for PDE constrained optimization.* Comput. Optim. Appl. 39(3) (2008), 369–393.

[STK11] J. Schröck, T. Meurer, and A. Kugi: *Control of a flexible beam actuated by macro-fiber composite patches: I. Modeling and feedforward trajectory control.* Smart Mater. Struct. 20 (2011) 015015-1–7.

[SW97] H.J. Sussmann and J.C. Willems: *300 Years of Optimal Control: From the Brachystochrone to the Maximum Principle.* In: IEEE Control Systems (June 1997), pp. 32–44.

[SKB17] K. Sverdrup, S.-J. Kimmerle, and P. Berg: *Computational investigation of the stability and dissolution of nanobubbles.* Appl. Math. Model. 49 (2017), 199–219.

[TT16] *Nonlinear stability of a heterogeneous state in a PDE-ODE model for acid-mediated tumor invasion.* Math. Biosciences Engrg. 13 (2016), 193–207.

[Tr10] F. Tröltzsch: *Optimal Control of Partial Differential Equations. Theory, Methods and Applications.* Graduate Studies in Mathematics 112, AMS, Providence, Rhode Island, 2010.

[Tr84a] F. Tröltzsch: *The generalized bang-bang principle and the numerical solution of a parabolic boundary-control problem with constraints on the control and the state.* Z. Angew. Math. Mech. 64 (1984), 551–557.

[Tr84b] F. Tröltzsch: *Optimality Conditions for Parabolic Control Problems and Applications.* Teubner-Texte zur Mathematik 62, Teubner, Leipzig, 1984.

[Ul01] M. Ulbrich: *Nonsmooth Newton-like Methods for Variational Inequalities and Constrained Optimization Problems in Function Spaces.* Habilitation thesis, Technical University Munich, München, 2001.

[Ul11] M. Ulbrich: *Semismooth Newton methods for Variational Inequalities and Constrained Optimization Problems in Function Spaces.* MOS-SIAM Series in Optimization 11, SIAM/MOS, Philadelphia, PA, 2011.

[UU09] M. Ulbrich and S. Ulbrich: *Primal-dual interior point methods for PDE-constrained.* Math. Prog. 117 (2009), 435–485.

[Ul02] S. Ulbrich: *A sensitivity and adjoint calculus for discontinuous solutions of hyperbolic conservation laws with source terms.* SIAM J. Control Opt. 41 (2002), 740–797.

[Ul03] S. Ulbrich: *Adjoint-based derivative computations for the optimal control of discontinuous solutions of hyperbolic conservation laws.* Syst. Control Lett. 48 (2003), 309–324.

[Vr98] C.B. Vreugdenhil: *Numerical Methods for Shallow-Water Flow.* Kluwer Academic

Publishers, Dordrecht, reprinted ed. 1998.

[Wa72] J. Warga: *Optimal Control of Differential and Functional Equations.* Academic Press, New York/London, 1972.

[WGKG18] J.-H. Webert, P.E. Gill, S.-J. Kimmerle, and M. Gerdts: *A study of structure-exploiting SQP algorithms for an optimal control problem with coupled hyperbolic and ordinary differential equation constraints.* Discrete Contin. Dyn. Syst. Ser. S 11 (2018), 1259–1282.

[WGS08] M. Weiser, T. Gänzler, and A. Schiela: *A control reduced primal interior point method for a class of control constrained optimal control problems.* Comput. Optim. Appl. 41 (2008), 127–145.

[WRP10] S. Wendl, A. Rund, and H.J. Pesch: *On a state-constrained PDE optimal control problem arising from ODE-PDE optimal control* In: M. Diehl, F. Glineur, and W. Michiels (eds.): *Recent Advances in Optimization and its Applications in Engineering*, pp. 429–438, Springer, Berlin/Heidelberg, 2010.

[We14] S. Wendl: *On a Prototype of an Optimal Control Problem Governed by Ordinary and Partial Differential Equations.* PhD thesis, Universität Bayreuth, Bayreuth, 2014.

[WC12] M. Witzgall and K. Chudej: *Flight Path Optimization subject to Instationary Heat Constraints.* In: 7th Vienna International Conference on Mathematical Modelling, Vienna, Austria, February 14–17, 2012, IFAC Proceedings Volumes 45/2 (2012), 1141–1146.

[We95] D. Werner: *Funktionalanalysis.* Springer, Berlin-Heidelberg-New York, 1995.

[Wo76] L. v. Wolfersdorf: *Optimal control for processes governed by mildly nonlinear differential equations of parabolic type I.* Z. Angew. Math. Mech. 56 (1976), 531–538.

[Wo77] L. v. Wolfersdorf: *Optimal control for processes governed by mildly nonlinear differential equations of parabolic type II.* Z. Angew. Math. Mech. 57 (1977), 11–17.

[Za15] L. Zajíček: *Hadamard differentiability via Gâteaux differentiability.* Proc. Amer. Math. Soc. 143 (2015), 279–288.

[Za00] A. Zaslavski: *Existence and structure of optimal solutions of infinite-dimensional control problems.* Appl. Math. Optim. 42 (2000), 291–313.

[Ze86] E. Zeidler: *Nonlinear Functional Analysis and Its Applications. I, Fixed-Point Theorems.* Springer, Berlin, 1986.

[ZK79] J. Zowe and S. Kurcyusz: *Regularity and stability for the mathematical programming problem in Banach spaces.* Appl. Math. Optim. 5 (1979), 49–62.

[Zu05] E. Zuazua: *Propagation, observation, and control of waves approximated by finite difference methods.* SIAM Rev. 47 (2005), 197–243.

[Zu15] A. Zuyev: *Partial Stabilization and Control of Distributed Parameter Systems with Elastic Elements.* Lecture Notes in Control and Inform. Sci. 458, Springer, Cham, 2015.

Appendix A

Mathematical Toolbox

A.1 Some Basics from Analysis

A.1.1 Functional Analysis

For convenience of the reader, we collect here some basic results from functional analysis, see, e.g. [Al16, HPUU09].

Definition A.1 *(Banach Space)*
A complete normed vector space X, equipped with a norm $\|\cdot\|_X$, is called a Banach space $(X, \|\cdot\|_X)$. If the usual norm $\|\cdot\|_X$ is considered for X, we drop it in our notation and write only X.

We denote 0_X for the zero element of a Banach space unless this is clear from the context.

Any norm induces $\|\cdot\|$ a metric d by means of $d(x^{(1)}, x^{(2)}) := \|x^{(1)} - x^{(2)}\|$, where $x^{(i)} \in X$ for $i = 1, 2$.

Definition A.2 *(Separable Banach Space)*
A Banach space X is separable, if it contains a countable dense subset: there exists a countable subset $S = \{x^{(i)} \in X \,|\, i \in \mathbb{N}\} \subset X$ such that

$$\forall x \in X \,\forall \varepsilon > 0 \,\exists s \in S : \|x - s\|_X < \varepsilon \quad \textit{i.e. } S \textit{ is dense in } X.$$

Theorem A.3 *(Banach Fixed Point Theorem)*
Let (X, d) be a complete metric space with metric d.
If the self-mapping $f : X \to X$ is a strict contraction, i.e.

$$\exists C \in (0,1) : d(f(x), f(\tilde{x})) \leq C d(x, \tilde{x}) \quad \forall x, \tilde{x} \in X,$$

then the sequence $\{x^{(i)}\}$, $i \in \mathbb{N}_0$, generated by the fixed point iteration

$$x^{(i+1)} := f(x^{(i)}), \quad x^{(0)} := x_0 \in X \textit{ (arbitrary)}$$

converges to the unique fixed point $\hat{x}$ of f, i.e. $\hat{x} = f(\hat{x}) \in X$.

Note that this theorem exhibits a constructive procedure how to compute $\hat{x}$.

Proof. By complete induction we obtain by the definition of the sequence

$$d(x^{(m)}, x^{(n)}) \leq \sum_{j=n}^{m} d(x^{(j+1)}, x^{(j)}) \quad \forall m > n,\ m, n \in \mathbb{N}_0.$$

Exploiting the strict contraction and the geometric series we have

$$d(x^{(m)}, x^{(n)}) \leq \sum_{j=n}^{m} c^j d(x^{(1)}, x^{(0)}) \leq \frac{c^n}{1-c} d(x^{(1)}, x^{(0)}).$$

Thus $\{x^{(i)}\}$ is a Cauchy sequence that implies the convergence of the sequence, since X is complete. This implies convergence of f by means of

$$d(f(x^{(i)}, f(\tilde{x})) \leq c\, d(x^{(i)}, \tilde{x}) \overset{i\to\infty}{\longrightarrow} 0.$$

This provides an error estimate for the fixed point iteration, too.

For the uniqueness, consider a different fixed point $\check{x}$ and we obtain

$$d(\hat{x}, \check{x}) = d(f(\hat{x}), f(\check{x})) \leq C d(\hat{x}, \check{x}) \Leftrightarrow d(\hat{x}, \check{x}) = 0 \Leftrightarrow \hat{x} = \check{x}.$$

□

Definition A.4 *(Linear Operator)*
Let X, Y be normed real vector spaces with corresponding norms $\|\cdot\|_X$ and $\|\cdot\|_Y$.

a) A map $A : X \to Y$ is a linear operator, if there holds

$$A(\alpha_1 x_1 + \alpha_2 x_2) = \alpha_1 A x_1 + \alpha_2 A x_2 \quad \forall x_1, x_2 \in X,\ \alpha_1, \alpha_2 \in \mathbb{R}.$$

b) The space of all bounded linear operators $A : X \to Y$, equipped with the operator norm

$$\|A\|_{X,Y} := \sup_{\|x\|_X = 1} \|Ax\|_Y,$$

is denoted by $\mathcal{L}(X, Y)$.

$\mathcal{L}(X, Y)$ is a normed space. If Y is a Banach space, then $\mathcal{L}(X, Y)$ is a Banach space.

Theorem A.5 *(Densely Defined Linear Operator)*
Let X be a normed space, Y be a Banach space (i.e. a complete normed vector space) and let $V \subset X$ be a dense subspace equipped with the same norm as X. Then for any $A \in \mathcal{L}(V, Y)$, there exists a unique extension $\tilde{A} \in \mathcal{L}(V, Y)$ s.t. $\tilde{A}|_V = A$. Then $\tilde{A}$ is called a densely defined linear operator.[1]

[1] Densely defined operators arise if one would like to apply an operation to a larger class of objects than those for which this makes sense *a priori*.

Definition A.6 *(Linear Functional; Dual Space)*
Let X be a Banach space. A linear operator $X \to Y$ mapping into $Y = \mathbb{R}$ is called linear functional on X.

The space $X^* := \mathcal{L}(X, \mathbb{R})$ of linear functionals on X is called the (topological) dual space of X.

Let $x \in X$, $\xi \in X^*$, then we write

$$\langle \mathcal{J}_X \xi, x \rangle_{X^*,X} := \xi\, x := \xi(x).$$

The notation $\langle \cdot, \cdot \rangle_{X^*,X}$ is called the dual pairing between X^* and X. Here J_X is the canonical isomorphism from $\mathcal{L}(X, \mathbb{R})$ into the space of continuous linear functionals on X with respect to the dual pairing (cf. [IK08, p.7]).

In finite dimensions the isomorphism J_X corresponds to a transposition. The use of the notation J_X will be omitted in this study for ease of presentation.

Definition A.7 *(Reflexive Banach Space; Bidual Space)*
A Banach space X is called reflexive, if the mapping $X \to X^{**} := (X^*)^*$, $x \mapsto \langle \cdot, x \rangle_{X^*,X}$ is surjective, i.e.:

$$\forall x^{**} \in X^{**}\ \exists x \in X \text{ s.t. } \langle x^{**}, x^* \rangle_{X^{**},X^*} = \langle x^*, x \rangle_{X^*,X} \quad \forall x^* \in X^*.$$

X^{**} is called bidual space of X.

Theorem A.8 *(Riesz Representation Theorem)*
Let H be a Hilbert space. For every $v \in H$ the linear functional defined by

$$\langle u^*, u \rangle_{H^*,H} := (v, u)_H \quad \forall u \in H$$

is in H^* and for any $u^* \in H$ there exists a unique $v \in H$ such that

$$\langle u^*, u \rangle_{H^*,H} = (v, u)_H \quad \forall u \in H$$

Moreover, $\|u^*\|_{H^*} = \|v\|_H$.

This implies that every Hilbert space is reflexive.

Definition A.9 *(Dual Operator)*
Let X, Y be Banach spaces, then the dual operator A^* to an operator $A \in \mathcal{L}(X, Y)$ is defined by

$$\langle A^* \xi, x \rangle_{X^*,X} = \langle \xi, Ax \rangle_{Y^*,Y} \quad \forall \xi \in Y^*, x \in X.$$

Consequently, $A^* \in \mathcal{L}(Y^*, X^*)$ and for the operator norm $\|A^*\|_{Y^*,X^*} = \|A\|_{X,Y}$.

If $X = \mathbb{R}^m$, $Y = \mathbb{R}^n$, $m, n \in \mathbb{N}$, then $A \in \mathbb{R}^{n \times m}$ is a matrix and $A^* = A^\top \in \mathbb{R}^{m \times n}$ is the transpose.

A.1.2 Ordinary Differential Equations

In this subsection we provide a few useful tools for the analysis of ordinary differential equations.

Hölder Spaces

Definition A.10 *(Hölder Spaces)*
Let Ω open and bounded, $r \in \mathbb{N}$ and $0 < \alpha \leq 1$. The α-Hölder seminorm is

$$[u]_{C^{0,\alpha}(\overline{\Omega})} := \sup_{x \neq y; x,y \in \overline{\Omega}} \left\{ \frac{|u(x) - u(y)|}{|x-y|^\alpha} \right\}. \tag{A.1}$$

The Hölder spaces

$$C^{r,\alpha}(\overline{\Omega}) := \left\{ u \in C^r(\overline{\Omega}) \,\middle|\, \sum_{|\gamma| \leq r} \sup_{x \in \overline{\Omega}} |\partial^\gamma u(x)| + \sum_{|\gamma| = r} [\partial^\gamma u]_{C^{0,\alpha}(\overline{\Omega})} < \infty \right\} \tag{A.2}$$

contain functions that are r-times continuously differentiable with r-th partial derivatives that are Hölder continuous with exponent α.

Special cases:

- $C^{r,0}(\overline{\Omega}) := C^r(\overline{\Omega})$ *is the space of r-times continuously differentiable functions.*
- $(C(\overline{\Omega}) :=)\, C^0(\overline{\Omega}) = C^{0,0}(\overline{\Omega})$ *is the space of continuous functions.*
- $C^{0,1}(\overline{\Omega})$ *is the space of Lipschitz continuous functions. These are exactly the functions that are continuously differentiable almost everywhere (according to the Rademacher theorem). Note that $C^{0,1} \neq C^{1,0} = C^1$.*

Hölder spaces are Banach spaces.

Well-Posedness Results

We state the *Gronwall inequality* in the following formulation [J00, Kap. 5, 2.12]:

Lemma A.11 *(Gronwall)*
Let $g \in C^0([0,t_f);\mathbb{R}_0^+)$, $h : [0,t_f) \to \mathbb{R}^+$ be integrable, and $C \in \mathbb{R}_0^+$. If

$$g(t) \leq C + \int_0^t h(\tau) g(\tau)\, d\tau \quad \forall t \in [0,t_f),$$

then the growth of the function g may be limited exponentially, i.e.

$$g(t) \leq C \exp\left(\int_0^t h(\tau)\, d\tau \right) \quad \forall t \in [0,t_f).$$

Proof.
W.l.o.g. $g \not\equiv 0$, otherwise the statement is trivial. For all $t \in [0, t_f)$ we set

$$f(t) := \ln\left(C + \int_0^t h(\tau)g(\tau)\, d\tau\right).$$

By the premise there follows for the derivative

$$f'(t) = \frac{h(t)g(t)}{C + \int_0^t h(\tau)g(\tau)\, d\tau} \leq h(t).$$

We derive from the premise

$$g(t) \leq C + \int_0^t h(\tau)g(\tau)\, d\tau = \exp(f(t)) = \exp\left(f(0) + \int_0^t f'(\tau)\, d\tau\right) \leq C \exp\left(\int_0^t h(\tau)\, d\tau\right)$$

by applying the fundamental theorem of calculus in the third step and the estimate on f' in the last step. □

By means of the Gronwall lemma we may state the following local existence (meaning existence local in time) and uniqueness result for the initial value problem for differential equations.

Theorem A.12 *(Uniqueness for Initial Value Problems)*
We consider the initial value problem

$$y'(t) = f(t, y(t)) \qquad \forall t \in (0, t_f), \tag{A.3}$$

$$y(0) = y_0, \tag{A.4}$$

with given $y_0 \in \mathbb{R}^d$ and right-hand side $f : [0, t_f) \times \mathbb{R}^d \to \mathbb{R}^d$, $t \times y \mapsto y$.
Let $f \in C^0([0, t_f] \times \mathbb{R}^d; \mathbb{R}^d)$ and Lipschitz continuous w.r.t. y, such that

$$\|f(t, y_1) - f(t, y_2)\| \leq h(t)\|y_1 - y_2\| \quad \forall t \in [0, t_f]\, \forall y_1, y_2 \in \mathbb{R}^d$$

with $h : [0, t_f] \to \mathbb{R}^+$. Then there exists at most one solution $y \in C^0([0, t_f); \mathbb{R}^d)$ of the initial value problem (A.3) & (A.4).

Proof.
Let y_i, $i = 1, 2$, be solutions of this initial value problem and set $y_\Delta := y_1 - y_2$. This difference solves the following initial value problem

$$\begin{aligned} y_\Delta'(t) &= f(t, y_1(t)) - f(t, y_2(t)) \qquad && \forall t \in (0, t_f), \\ y_\Delta(0) &= 0 \in \mathbb{R}^d, \end{aligned}$$

We integrate the ODE, use an inequality for the modulus, and exploit the Lipschitz continuity of f w.r.t. y:

$$\begin{aligned} \|y_\Delta(t)\| &= \left\| y_\Delta(0) + \int_0^t f(\tau, y_1(\tau)) - f(\tau, y_2(\tau))\, d\tau \right\| \\ &\leq \int_0^t \|f(\tau, y_1(\tau)) - f(\tau, y_2(\tau))\|\, d\tau \\ &\leq \int_0^t h(\tau)\|y_\Delta(\tau)\|\, d\tau. \end{aligned}$$

By means of the Gronwall inequality, this implies $y_\Delta \equiv 0$ for all $t \in [0, t_f)$. □

However, it is not clear, whether a solution local in time of the initial value problem exists at all. If we require the continuity of f, then by the Peano theorem we obtain existence of at least one solution.

For the definition of the index of differential algebraic equations and for tools for considering DAE, we refer to [Ge12] and the references therein.

A.1.3 Partial Differential Equations

In the following we recall some basic results for partial differential equations. Please find proofs and further details, e.g., in the textbooks of Alt [Al16] and Evans [Ev10].

If all derivatives of highest order appear only linearly (but this has not to be the case for the function or for derivatives of lower order), then the PDE is called *semilinear*.

If the coefficient functions for the highest derivatives may depend additionally on the unknown function and on lower derivatives, then the PDE is called *quasilinear*.

Sobolev Spaces

In this subsection let $\Omega \subset \mathbb{R}^d$. If Ω is open, non-empty, and connected, Ω is called a *domain*.

For PDEs the concept of weak solutions turns out to be more suitable than classical solutions:

Remark A.13 *(Weak Solution)*
It turns out that it makes sense in general to look for the solution of the Poisson problem with homogeneous Dirichlet boundary conditions in a larger function space as $H_0^1(\Omega)$.

If we have found a so-called weak solution (or variational solution) in $V = H_0^1(\Omega)$, then we may prove under certain assumptions on the data (as $f \in L^2(\Omega)$ and either (i) C^1-boundary $\partial\Omega$ or (ii) Ω convex & polygonal) that actually $u \in H^2(\Omega)$ holds. By means of the following embedding theorems this implies that $u \in C^2(\overline{\Omega})$ (the reversal is not true in general), thus u is also a classical solution.

Weak solutions open the door for an approach using variational methods, i.e. we can multiply the Poisson equation with an arbitrary test function $\phi \in W$,

$$\int_\Omega -\Delta_x u(x)\, \phi(x)\, dx = \int_\Omega f(x)\, \phi(x)\, dx \quad \forall \phi \in W \tag{A.5}$$

and we may vary this arbitrary function ϕ. For ease of presentation we choose $W = V$ here.

We recall that classical solutions live in Hölder spaces, see Def. A.10. For our purposes the suitable spaces for weak solutions are Sobolev spaces. For the definition of Sobolev spaces, we consider at first

Definition A.14 *(Weak Derivatives)*
Let Ω be open. The weak derivative $D^\gamma f$ (for a multi index γ) of a function $f : \Omega \to \mathbb{R}$ is defined by means of partial integration as

$$\int_\Omega D^\gamma f \cdot \zeta := (-1)^{|\gamma|} \int_\Omega f \cdot \partial^\gamma \zeta \quad \forall \zeta \in C_0^\infty(\Omega). \tag{A.6}$$

The weak derivative is uniquely determined.

Definition A.15 *(Sobolev Spaces)*
Let Ω open, $k \in \mathbb{N}_0$, $1 \leq p \leq \infty$. The Sobolev norm is

$$\|u\|_{W^{k,p}(\Omega)} := \begin{cases} \left(\sum_{|\gamma|\leq k} \int_\Omega |D^\gamma u|^p\right)^{1/p}, & 1 \leq p < \infty, \\ \sum_{|\gamma|\leq k} \textit{ess sup}_{x\in\Omega} |D^\gamma u|, & p = \infty. \end{cases} \tag{A.7}$$

The Sobolev spaces are

$$W^{k,p}(\Omega) := \left\{ u : \Omega \to \mathbb{R} \textit{ measurable } | D^\gamma u \textit{ fulfils (A.6) } \forall |\gamma| \leq k \,\wedge\, \|u\|_{W^{k,p}(\Omega)} < \infty \right\}. \tag{A.8}$$

Special cases:

- *$H^k(\Omega) := W^{k,2}(\Omega)$ are Hilbert spaces.*
- *$L^p(\Omega) := W^{0,p}(\Omega)$ are Banach spaces containing the functions, whose absolute value to the power of p is Lebesgue integrable.*

In this study we consider only the cases $1 \leq p \leq \infty$.

Example A.16 *(Important Sobolev Spaces)*

a) $L^2(\Omega)$, the space of square-integrable functions.

b) $H^1(\Omega)$, the space of square-integrable functions with square-integrable weak derivatives of first-order.

c) $L^\infty(\Omega)$, the space of essentially bounded functions, equipped with the supremum norm.

Sobolev spaces are Banach spaces.

Definition A.17 *(Gelfand Triple; Function Space for Evolution Problems)*
If for separable Hilbert spaces H and V

$$V \overset{cd}{\hookrightarrow} H \simeq H^* \overset{cd}{\hookrightarrow} V^*$$

holds where both embeddings are continuous and dense, then we call V, H, V^* a Gelfand triple. Here the dual H^* is identified with H.

For a Gelfand triple V, H, V^* we may introduce the function space

$$W(0,t_f) := W(0,t_f;H,V) := \left\{ y \in L^2(0,t_f;V) \,\middle|\, y'_t \in L^2(0,t_f;V^*) \right\}$$

that appears naturally for parabolic problems with initial condition in H and right-hand side in $L^2(0,t_f;V^*)$ (see Pb. 3.13).

Theorem A.18 *(Embedding Theorem for Sobolev and Hölder Spaces)*
Let Ω open and bounded with Lipschitz boundary[2] *$\partial\Omega$.*

a) *Embeddings between Sobolev spaces*

If

$$k_1 - \frac{d}{p_1} \geq k_2 - \frac{d}{p_2}, \quad k_1 \geq k_2,\, k_1, k_2 \in \mathbb{N}_0,\, 1 \leq p_1, p_2 \leq \infty, \tag{A.9}$$

then

$$W^{k_1,p_1}(\Omega) \subset W^{k_2,p_2}(\Omega). \tag{A.10}$$

b) *Embeddings between Hölder spaces*

If

$$r_1 + \alpha_1 > r_2 + \alpha_2, \quad r_1, r_2 \in \mathbb{N}_0,\, \alpha_1, \alpha_2 \in [0,1], \tag{A.11}$$

then

$$C^{r_1,\alpha_1}(\overline{\Omega}) \subset C^{r_2,\alpha_2}(\overline{\Omega}). \tag{A.12}$$

(In the case $r_1 = 0$ the assumption on $\partial\Omega$ may be dropped.)

c) *Embeddings from Sobolev spaces into Hölder spaces*

If

$$k - \frac{d}{p} > r + \alpha, \quad 1 \leq p < \infty,\, k, r \in \mathbb{N}_0, \alpha \in (0,1] \;\; (\alpha \neq 0!), \tag{A.13}$$

then

$$W^{k,p}(\Omega) \subset C^{r,\alpha}(\overline{\Omega}). \tag{A.14}$$

If

$$p = \infty,\, 1 \leq k \in \mathbb{N}, \tag{A.15}$$

then

$$W^{k,\infty}(\Omega) \cong C^{k-1,1}(\overline{\Omega}), \tag{A.16}$$

where a function in $W^{k,p}$ may be identified by its Lipschitz continuous representative (in $C^{k-1,1}$) that is almost everywhere identical to the function itself.

Here we have combined several results that are proved in [Al16].

Remark A.19 *(Embeddings and Estimates)*
For the embedding, we write $X \subset Y$, denoted also as $X \hookrightarrow Y$. This is equivalent that there exists a constant $C > 0$ such that the inequality $\|u\|_Y \leq C\|u\|_X$ for all $u \in X$ holds.

In particular, we find for any dimension

$$C^{0,1}(\Omega) \subset L^\infty(\Omega) \tag{A.17}$$

and that for $d = 1$

$$H^1(\Omega) \subset L^\infty(\Omega), \quad H^1(\Omega) \subset C^0(\Omega). \tag{A.18}$$

[2] See the footnote in Ex. 2.68 a).

Remark A.20 *(Piecewise Regularity in H^1)*
Let Ω be open and bounded. If a function is bounded, continuous, and piecewise continuously differentiable, then it belongs to $H^1(\Omega)$.

Since a function u from a Sobolev space is not continuous in general and only differentiable almost everywhere, we have to ensure that boundary values of u are well-defined.

Theorem A.21 *(Trace Theorem) [Ci98, Th. 6.1-7]; [Tr10, Th. 7.2]*
Let $\Omega \subset \mathbb{R}^d$ open and bounded with boundary[3] *$\partial\Omega$ of class $C^{p-1,1}$ and let $1 \leq p < \infty$, $k \in \mathbb{N}$, then*

$$u \in W^{k,p}(\Omega) \Longrightarrow u|_{\partial\Omega} \in W^{k-1/p,p}(\partial\Omega). \tag{A.19}$$

Equivalently, we may rewrite the latter as an embedding operator, the so-called trace operator,

$$Tr : W^{k,p}(\Omega) \to W^{k-1/p,p}(\partial\Omega). \tag{A.20}$$

Loosely speaking, if the function is considered on the boundary $\partial\Omega$, i.e. the *trace* $Tr\, u = u|_{\partial\Omega}$ of the function, then we loose a $1/p$-order of derivative.

This may be combined with the embedding theorems, Th. A.18). Then we may characterize the trace mapping as follows :

a) If $p\,k < d$, then

$$u \in W^{k,p}(\Omega) \Longrightarrow u|_{\partial\Omega} \in L^r(\partial\Omega) \quad \text{for all } 1 \leq r \leq \frac{(d-1)p}{d-p\,k}.$$

b) If $p\,k = d$, then

$$u \in W^{k,p}(\Omega) \Longrightarrow u|_{\partial\Omega} \in L^r(\partial\Omega) \quad \text{for all } 1 \leq r < \infty.$$

c) If $p\,k > d$, then

$$W^{k,p}(\Omega) \subset C^0(\overline{\Omega}), \text{ also } u|_{\partial\Omega} \in C^0(\partial\Omega).$$

In particular we have $u \in H^1(\Omega) \subset H^{1/2}(\partial\Omega) \subset L^2(\partial\Omega)$. Thus we are able to define

Definition A.22 *(Sobolev Spaces With Homogeneous Dirichlet Boundary Values)*

$$W_0^{k,p}(\Omega) := \left\{ u \in W^{k,p}(\Omega) \mid u|_{\partial\Omega} = 0 \right\} := \left\{ u \in W^{k,p}(\Omega) | Tr\, u = 0 \right\}. \tag{A.21}$$

In particular we have $H_0^1(\Omega) = W_0^{1,2}(\Omega)$.

We define $H^{-1}(\Omega) := (H_0^1(\Omega))^*$ as the dual space of $H_0^1(\Omega)$. (Note this is no Sobolev space.)

Furthermore, a trace operator $u|_{\Gamma_0}$ may be defined by restriction on measurable subsets $\Gamma_0 \subset \partial\Omega$.

[3] For $d = 1$ see the footnote in Ex. 2.68 a).

Remark A.23 *(Approximation Properties of Sobolev Spaces)*
For a smooth boundary $\partial\Omega$ we may prove that

$$W_0^{k,p}(\Omega) = \overline{C_0^\infty(\Omega)}^{W^{k,p}}, \tag{A.22}$$

i.e. the closure (w.r.t. the $W^{k,p}$ norm) of the space of infinitely continuously differentiable functions with compact support is just the Sobolev space $W_0^{k,p}(\Omega)$. Analogously there holds

$$W^{k,p}(\Omega) = \overline{C^\infty(\overline{\Omega})}^{W^{k,p}}. \tag{A.23}$$

These observations motivate that we may work in variational problems with test functions of a Sobolev space instead of test functions $C_0^\infty(\Omega)$.

There holds the following important estimate.

Lemma A.24 *(Poincaré-Friedrichs Inequality)*
For $u \in H_0^1(\Omega)$ we have

$$\|u\|_{L^2(\Omega)} \le C\|\nabla u\|_{L^2(\Omega)} \tag{A.24}$$

with a constant $C > 0$. (The constant depends only on Ω and the dimension d.)

This inequality may be weakened in the sense that it suffices that u is fixed only on a part of the boundary with non-zero measure or that the mean value $(1/|\Omega|)\int_\Omega u = c = const$ prescribed, whereas u may have any boundary values. However, this inequality does not hold for any function in $H^1(\Omega)$.

In particular, this implies that the standard H^1-norm is in $H_0^1(\Omega)$ equivalent to the norm $[[u]]_{H^1(\Omega)} := \|\nabla u\|_{L^2(\Omega)}$. However, in general $[[u]]_{H^k(\Omega)} = (\sum_{|\gamma|=k} \|D^\gamma u\|^2_{L^2(\Omega)})^{1/2}$, $k \in \mathbb{N}$, is a seminorm, i.e. $[[u]]_{H^k(\Omega)} = 0$ does not necessarily imply $u = 0$ almost everywhere.

If it is evident from the context, we do not indicate the domain with norms. For instance, we write $\|u\|_{W^{2,3}} := \|u\|_{W^{2,3}(\Omega)}$.

Well-Posedness Results

Finally, we recall the standard results for existence and uniqueness of solutions.

Definition A.25 *(Ellipticity (Coercivity))*
Let $\Omega \subset \mathbb{R}^d$ be open and let $a_{ij}(x) \in L^\infty(\Omega)$ be coefficient functions. If there exists a constant $C > 0$ s.t.

$$\sum_{i,j=1}^d a_{ij}(x)\xi_i\xi_j \ge C\|\xi\|^2 \quad \textit{for almost all } x \in \Omega \textit{ and for a.a. } \xi \in \mathbb{R}^d,$$

and $c_0 \ge 0$, then the operator $-\sum_{i,j=1}^d (a_{ij}y'_{x_i})'_{y_{x_j}} + c_0 y$ is called uniformly elliptic (coercive). The associated bilinear form

$$a : V \times V \to \mathbb{R}, \quad [y,\eta] \mapsto a(y,\eta) := \int_\Omega \left(\sum_{i,j=1}^d a_{ij}y'_{x_i}\eta'_{x_j} + c_0 y\eta \right) dx$$

is called <u>*V-elliptic (V-coercive)*</u>*, too.*

Theorem A.26 *(Lax-Milgram)*
Let V be a Hilbert space and

$$a : V \times V \to \mathbb{R}$$

a (not necessarily symmetric) bilinear form, that fulfils

$$a(v,v) \geq \alpha_0 \|v\|_V^2 \qquad V\textit{-ellipticity,} \tag{A.25}$$

$$|a(v,w)| \leq \alpha_1 \|v\|_V \|w\|_V \qquad V\textit{-boundedness,} \tag{A.26}$$

for any $v, w \in V$, with $0 < \alpha_0 \leq \alpha_1 < \infty$.

Then for any $f \in V^*$ the equation

$$a(v,w) = \langle f, w \rangle_{V^*,V} \tag{A.27}$$

has a unique solution $v \in V$.

Furthermore, there holds the estimate $\|v\|_V \leq 1/\alpha_0 \, \|f\|_{V^*}$.

Proof. By means of the Riesz representation theorem, see Th. A.8. □

Theorem A.27 *(Existence and Uniqueness for Semilinear Elliptic Equations)*
We consider Example 2.68 where part b) is adapted to the case of pure Neumann b.c., that is $\Gamma = \Gamma_N$. Let $\Omega \subset \mathbb{R}^d$ a bounded Lipschitz domain, let A an elliptic differential operator of the form (2.75) with bounded and measurable coefficient functions a_{ij} that are symmetric and uniformly elliptic (i.e. (A.25) holds for almost all x where $V = H^1(\Omega)$), and let $d_0 \in L^\infty(\Omega; \mathbb{R}_0^+)$ and $d_\Gamma \in L^\infty(\Gamma_N; \mathbb{R}_0^+)$ such that $\|d_0\|_{L^\infty(\Omega)} + \|d_\Gamma\|_{L^\infty(\partial\Omega)} > 0$. Furthermore, let $c_0 : \Omega \times \mathbb{R} \to \mathbb{R}$ and $c_\Gamma : \partial\Omega \times \mathbb{R} \to \mathbb{R}$ continuous and monotone increasing in y for almost all $x \in \Omega$ and $x \in \partial\Omega$, respectively, and $c_0(\cdot, y) \in L^\infty(\Omega)$ and $c_\Gamma(\cdot, y) \in L^\infty(\partial\Omega)$ for all $y \in \mathbb{R}$. If $f \in L^r(\Omega)$ for $r > d/2$ and $g \in L^s(\partial\Omega)$ for $s > d-1$, $d \geq 2$, then there exists a unique solution $y \in H^1(\Omega) \cap C^0(\overline{\Omega})$. Note that in the case $d = 1$, we require just $g \in \mathbb{R}$ on the two isolated points of $\partial\Omega$.
Furthermore, we have the estimate

$$\|y\|_{H^1(\Omega)} + \|y\|_{C^0(\overline{\Omega})} \leq C \left(\|f - c_0(\cdot, 0)\|_{L^r(\Omega)} + \|g - c_\Gamma(\cdot, 0)\|_{L^s(\partial\Omega)} \right)$$

with a constant C not depending on f, g, d_0, d_Γ, c_0, and c_Γ.

For a proof the theory of maximal monotone operators is combined with a technique by Stampacchia, see, e.g., [Tr10]. Similar results hold for pure Dirichlet b.c. as well.

Theorem A.28 *(Existence and Uniqueness for Semilinear Parabolic Equations)*
We consider Example 2.69 for the case $a_{ij}(t,x) = a_{ij}(x)$ and $b \equiv 0$. Let $\Omega \subset \mathbb{R}^d$ be a bounded domain with $C^{1,1}$-boundary, we adapt Ex. 2.69 b) for the case $\Sigma_{t_f} = \Sigma_N$ (i.e. pure Neumann b.c.), and let $A := \sum_{i,j=1}^d (a_{ij}(x) y'_{x_i})'_{x_j}$ be a symmetric, uniformly elliptic operator, i.e. (A.25) holds for almost all x (where $V = H^1(\Omega)$), with essentially bounded coefficients, i.e. $a_{ij} \in L^\infty(\Omega)$.

Moreover, let c_0 and c_Σ fulfil the Assumptions 2.77 and be essentially bounded for fixed $y \in \mathbb{R}^d$ and let $c_0(t,x,\cdot)$ and $c_\Sigma(t,x,\cdot)$ be continuous and locally Lipschitz continuous uniformly for a.a. $[t,x] \in \Omega_{t_f}$ and $[t,x] \in \Sigma_{t_f}$, resp. If $f \in L^r(\Omega_{t_f})$ for $r > d/2+1$, $g \in L^s(\Sigma_N)$ for $s > d+1$, and $y_0 \in C^0(\overline{\Omega})$, then there exists a unique weak solution $y \in W(I; L^2, H^1) \cap C^0(\overline{\Omega})$.

Furthermore, we have the estimate

$$\|y\|_{W(I)} + \|y\|_{C^0(\overline{\Omega}_{t_f})} \leq C \left(\|f - c_0(\cdot,\cdot,0)\|_{L^r(\Omega_{t_f})} + \|g - c_\Sigma(\cdot,\cdot,0)\|_{L^s(\Sigma_N)} + \|y_0\|_{C^0(\overline{\Omega})} \right)$$

with a constant C not depending on f, g, c_0, c_Σ, and y_0.

For a proof see [Ca97, RZ99].

Hyperbolic equations of first-order, like in the truck-St-Venant example, are not solved directly in this manuscript, since we regularize by an artificial viscosity in order to avoid technical difficulties with shocks and rarefaction waves. Thus we do not recall existence and uniqueness results for this kind of PDE as well as for hyperbolic equations of second-order.

A.1.4 Differentiability in Banach Spaces

We would like to extent the concept of derivatives to operators between Banach spaces.

Definition A.29 *(Differentiability in Banach Spaces)*
Let $E : V \to Y$ be an arbitrary operator, where V is a non-empty open subset of X and X, Y are Banach spaces.

a) E is said to be directionally differentiable at $x \in V$ in direction $d \in V$, if the limit

$$E'(x;d) := \lim_{t \downarrow 0} \frac{E(x+td) - E(x)}{t}$$

exists. $E'(x;d)$ is called directional derivative of E at x in direction d.

a2) E is called Hadamard (directional) differentiable (H-differentiable) at $x \in V$ in direction $d \in X$,

- *if there exists a map $E' \in \mathcal{L}(X,Y)$ such that*

$$E'(x;d) := \lim_{d_i \to d, t_i \downarrow 0} \frac{E(x+t_i d_i) - E(x)}{t_i}$$

exists and

- *E' is linear in d [BS00, Def. 2.45].*

Note that the existence of this limit implies the continuity of E', but in general not its linearity.

b) E is called Gâteaux differentiable (G-differentiable) at $x \in V$,

- *if E is directionally differentiable at x and*
- *if for $E'(x) : X \ni d \mapsto E'(x; d) \in Y$ holds that $E'(x) \in \mathcal{L}(X, Y)$.*

c) E is called Fréchet differentiable (F-differentiable) or totally differentiable at $x \in V$,

- *if E is Gâteaux differentiable at x and*
- *if*

$$\lim_{\|d\|_X \to 0} \frac{\|E(x+d) - E(x) - E'(x)d\|_Y}{\|d\|_X} = 0.$$

Note that F-differentiability of E at x implies continuity of E at x.

d) E is called Newton (or slant) differentiable (N-differentiable) at $x \in V$,

- *if for each $\varepsilon > 0$ there exists a neighbourhood $\tilde{V}(x)$ of x, such that for each $x + d \in \tilde{V}(x)$ there exists a mapping $E'(x+d) : \tilde{V}(x) \to \mathcal{L}(X, Y)$, such that*

$$\|E(x+d) - E(x) - E'(x+d)d\|_Y \leq \varepsilon \|d\|_X.$$

The family $\{E'(\xi) \,|\, \xi \in \tilde{V}(\xi)\}$ is called the Newton (or slant) derivative (N-derivative) of E at x [IK08].

e) E is called $\partial^* E$-semismooth at $x \in V$,

- *if E is continuous near x and*
- $\sup_{M \in \partial^* E(x+d)} \|E(x+d) - E(x) - Md\|_Y = o(\|d\|_X)$ *as* $\|d\|_X \to 0$,

where $\partial^* E : V \rightrightarrows \mathcal{L}(X, Y)$ is a given set-valued mapping. For further details see [GHK17, Def. 2.1] and the comments there.

f) E is called Bouligand differentiable (B-differentiable) at $x \in V$,

- *if E is directionally differentiable at x and*
- *there holds*

$$\lim_{\|d\|_X \to 0} \frac{E(x+d) - E(x) - E'(x; d)}{\|d\|_X} = 0_Y.$$

If E exhibits one of the above differentiabilities at all x in $V \subset X$, then we say that E is correspondingly differentiable in the set V. For example, if E is F-differentiable at all x in V, then E is said to be F-differentiable in V.

For an alternative definition of Newton (slant) differentiability, please see [Kr13, Def. 2.1 & Remark 2.2].

We have the following implications, see, e.g., [BS00, IK08, Za15], between the different derivatives in Banach spaces:

- If E' is continuous, then F-differentiability implies H-differentiability. If X is finite-dimensional, then the converse is also true.

- H-differentiability of E implies the Gâteaux differentiability (and both derivatives coincide). The converse is true, if E is Lipschitz continuous.

- G-differentiability of E together with continuity of E' implies F-differentiability of E. For continuously differentiable functions G- and F-differentiability are equivalent.

- In finite dimensions, F-, G- and H-differentiability are equivalent. Note that in infinite dimensions there exist convex continuous functions that are G- and H-differentiable, but not F-differentiable [BS00, Ex. 2.50].

- N-differentiability does not imply F-differentiability in general. For instance in infinite dimensions max- and min-functions are not F-differentiable, but N-differentiable, see, e.g., [HV04].

- If E is N-differentiable at x and $\lim_{t\downarrow 0} E'(x+td)d$ exists uniformly in $\|d\|_X = 1$, then E is semismooth at x.

- N-differentiability at x implies directional differentiability at x if and only if $\lim_{t\downarrow 0} E'(x+td)d$ exists for all $d \in X$. Then $E'(x;d) = \lim_{t\downarrow 0} E'(x+td)d$ [IK08, Lemma 8.11 (1)].

- N-differentiability at x implies B-differentiability at x if and only if $\lim_{t\downarrow 0} E'(x+td)d$ exists uniformly for all $\|d\|_X = 1$. Then E is semismooth [IK08, Lemma 8.11 (2)].

In particular we have that linear-quadratic functionals in Hilbert spaces are F-differentiable. A common situation in optimization is that we deal with a state equation $E(y,u) = 0$ with an operator $E : Y \times U \to W$ that is continuously F-differentiable and, moreover E'_y has a bounded inverse. Then the following standard result yields the local existence of a map $u \mapsto y(u)$.

Theorem A.30 *(Implicit Function Theorem)*
Let Y, U, W be Banach spaces, let $V \subset Y \times U$ be an open set, and let $F : V \to W$ be a continuously F-differentiable map. If $F(\hat{y},\hat{u}) = 0$ for $[\hat{y},\hat{u}] \in V$ such that $F'_y(\hat{y},\hat{u}) \in \mathcal{L}(Y,W)$ has a bounded inverse, then there exists a neighbourhood $V_y(\hat{y}) \times V_u(\hat{u}) \subset V$ of $[\hat{y},\hat{u}]$ and a unique continuous function $H : V_u(\hat{u}) \to Y$ such that

(i) $H(\hat{u}) = \hat{y}$,

(ii) for all $u \in V_u(\hat{u})$ there exists exactly one $y := H(u) \in V_y(\hat{y})$ such that $F(y,u) = 0$.

Furthermore, the mapping $H : V_u(\hat{u}) \to Y$ is F-differentiable with

$$H'(u) = F'_y(H(u),u)^{-1} F'_u(H(u),u).$$

The proof, see, e.g., [Ze86, Th. 4B], relies among other things on the Banach fixed point theorem (Th. A.3).

A.2 Cones and Constraint Qualifications

A.2.1 Cones

In optimization theory inequality constraints are dealt with suitably by cones. We refer, e.g., to [Ge12] for the following definitions.

Definition A.31 *(Convex Cone)*
A subset $\mathcal{K}$ of a vector space W, such that

$$k \in \mathcal{K} \quad \Longrightarrow \quad \alpha k \in \mathcal{K} \quad \forall \alpha \in \mathbb{R}_0^+$$

is called cone with vertex at 0_W, i.e. the origin of W. If $\mathcal{K}$ is convex, then it is a convex cone.

A partial ordering in W is induced by any convex cone $\mathcal{K} \subset W$.

Definition A.32 *(Partial Ordering in Cones)*
We use the notation $w \geq_{\mathcal{K}} 0$ if and only if $w \in \mathcal{K}$.
We write $w >_{\mathcal{K}} 0$ if and only if $w \in \overset{\circ}{\mathcal{K}}$.
We have $w \leq_{\mathcal{K}} 0$ if and only if $-w \in \mathcal{K}$.
Analogously, we introduce $<_{\mathcal{K}}$.

In order to derive useful second-order optimality conditions, we have to select reasonable search directions. These directions are called critical directions and they have the structure of a cone. This motivates the following two definitions.

Definition A.33 *(Conical Hull)*
Suppose $\zeta \in \Sigma$, Σ a non-empty convex set. The set

$$C_\Sigma(\zeta) := \{\gamma(z - \zeta) \,|\, \gamma \geq 0, z \in \Sigma\}$$

is called conical hull to Σ at ζ.

The critical cone is so to speak the set of ambiguous directions. For simplicity we provide the definition only for an equality constrained optimal control problem without Tikhonov term in the objective. The latter yields a solution of bang-bang type.

For a more general definition of the critical cone in infinite dimensions we refer to the works of Maurer [MZ79, Ma81].

Definition A.34 *(Critical Cone)*
Let the setting of Lemma 2.46 hold. We recall we consider only equality constraints $E(y, u) = 0$ in Ω and we assume $U = L^2(\Omega)$, $\Omega \subset \mathbb{R}^d$ be a bounded Lipschitz domain. We assume that a minimizer $[\hat{y}, \hat{u}]$ exists and that $E'_u(\hat{y}, \hat{u})^*$ is well-defined pointwise.

We define the set of active inequalities

$$A_0(\hat{u}) = \left\{x \in \Omega \,|\, |E'_u(\hat{y}(x), \hat{u}(x))^* \lambda(x)| > 0\right\},$$

where u is defined either by u_{min} or u_{max} uniquely. Then the <u>*critical cone*</u> $C_0(\hat{u})$ *is the set of all* $d \in L^\infty(\Omega)$ *such that*

$$d(x) \begin{cases} = 0 & \text{if } x \in A_0(\hat{u}), \\ \geq 0 & \text{if } x \notin A_0(\hat{u}) \text{ and } \hat{u}(x) = u_{min}(x), \\ \leq 0 & \text{if } x \notin A_0(\hat{u}) \text{ and } \hat{u}(x) = u_{max}(x). \end{cases}$$

d may be chosen arbitrarily on the inactive set $\{x \in \Omega \,|\, u_{min}(x) < \hat{u}(x) < u_{max}(x)\}$.

Definition A.35 *(Critical Cone in Finite Dimensions)*
We consider the general optimization problem, Problem 2.3, but in the finite-dimensional case $Z = \mathbb{R}^{n_z}$. *We define the set of active inequality constraints*

$$A(z) = \{i = 1, \ldots, n_G \,|\, G_i(z) = 0\}.$$

The corresponding <u>*critical cone*</u> *of directions is the set*

$$C_0(\hat{z}) := \left\{ d \in Z \;\middle|\; \begin{array}{lll} G'_{i,z}(\hat{z})d & \leq 0 & \text{if } i \in A(\hat{z}), \mu_i = 0, \\ G'_{i,z}(\hat{z})d & = 0 & \text{if } i \in A(\hat{z}), \mu_i > 0, \\ H'_{j,z}(z)d & = 0 & \text{if } j = 1, \ldots, n_H. \end{array} \right\}.$$

For further interpretation of the critical cone in finite dimensions see [Ge12, Sect. 6.1].

Definition A.36 *(Polar Cones)*
Let $\mathcal{K}$ be a cone with vertex 0_W. The <u>*positive polar cone*</u> *of $\mathcal{K}$ is given by*

$$\mathcal{K}^+ := \{w \in W^* \,|\, \langle w, k\rangle_{W^*,W} \geq 0 \;\; \forall k \in \mathcal{K}\}$$

and the <u>*negative polar cone*</u> *of $\mathcal{K}$ by*

$$\mathcal{K}^- := \{w \in W^* \,|\, \langle w, k\rangle_{W^*,W} \leq 0 \;\; \forall k \in \mathcal{K}\}.$$

Polar cones are non-empty, closed convex cones.

A polar cone is also called dual cone or conjugated cone.

The set of feasible directions in optimization has the structure of a cone:

Definition A.37 *(Tangent Cone)*
Let $\Sigma \subset Z$ be non-empty, Z a Banach space. The <u>*tangent cone*</u> *of Σ at $z \in \Sigma$ is defined by*

$$\mathcal{T}(\Sigma; z) := \{d \in Z \,|\, \exists \{\alpha_k\}_{k\in\mathbb{N}} > 0 \text{ with } \alpha_k \overset{k\to\infty}{\to} \infty, \exists \{z_k\}_{k\in\mathbb{N}} \in \Sigma \text{ with } \lim_{k\to\infty} z_k = z$$
$$\text{s.t. } \lim_{k\to\infty} \alpha_k(z_k - z) = d\}.$$

Typically $\Sigma(\subset Z_{ad} \subset Z)$ is the feasible set as defined in (2.2). We define in this case

Definition A.38 *(Linearized (Tangent) Cone) [HPUU09, 1.7.3.2]*
Let the feasible set $\Sigma \subset Z_{ad} \subset Z$ be non-empty, let $\mathcal{G}$ encode inequality and equality constraints, let $\mathcal{K}$ be a convex cone and let Z_{ad} be the set of admissible states. The <u>*linearized (tangent) cone*</u> $T_{lin}(\Sigma, \mathcal{G}, \mathcal{K}, Z_{ad}; z)$ *at $z \in \Sigma$ is the set*

$$T_{lin}(\Sigma, \mathcal{G}, \mathcal{K}, Z_{ad}; z) := \{\alpha d \,|\, \alpha > 0,\, d \in Z,\, \mathcal{G}(z) + \mathcal{G}'(z)d \in \mathcal{K},\, z + d \in Z_{ad}\}.$$

A.2.2 Constraint Qualifications

In order to guarantee the existence of Lagrange multipliers some regularity conditions, called constraint qualifications (CQs), have to be presumed. Most constraint qualifications may be interpreted geometrically as a statement on the non-separability of two convex sets within the space of constraints. In general, the local optimum $\hat{z} \in Z$ is involved in the CQ and, thus, the CQ cannot be checked *a priori*.

We start with one of the weakest constraint qualifications and we use the notation of Section 2.2. $\mathcal{G}$ summarizes inequality and equality constraints as in (2.7). Furthermore, we need the set of feasible directions that is the tangent cone (see Def. A.37) of the feasible set. The linearized cone, Def. A.38, allows for a less complicated representation of the tangent cone, if we have the

Assumption A.39 *(Abadie Constraint Qualification)*
The (Abadie) constraint qualification (ACQ), cf. [Ge12, Remark 2.3.40], reads

$$\mathcal{T}_{lin}(\Sigma, \mathcal{G}, \mathcal{K}, Z_{ad}; \hat{z}) \subset \mathcal{T}(\Sigma; \hat{z}),$$

where we use the notation from Appendix A.2.1.

Since the linearized cone is always a superset of the tangent cone, this implies that the tangent cone and the linearized cone coincide at $\hat{z}$. In general, it may be difficult to verify ACQ without going back to other constraint qualifications. But if all restrictions are affine, then ACQ is always fulfilled.

Assumption A.40 *(Robinson Constraint Qualification [Ge12, Def. 2.3.32])*
The Robinson constraint qualification (RCQ) or Robinson regularity condition is fulfilled at $\hat{z}$, if

$$\begin{bmatrix} 0_{W_G} \\ 0_{W_H} \end{bmatrix} \in int \left\{ \begin{bmatrix} G(\hat{z}) + G'(\hat{z})(z - \hat{z}) - k \\ H'(\hat{z})(z - \hat{z}) \end{bmatrix} \middle| \, z \in Z_{ad}, k \in \mathcal{K} \right\}.$$

The following rather general constraint qualification provides the existence of Lagrange multipliers in our setting from Subsection 2.2.2.

Assumption A.41 *(Zowe-Kurcyusz Constraint Qualification [ZK79])*
Let Z_{ad} be convex. The Zowe-Kurcyusz constraint qualification (ZKCQ) is fulfilled at $\hat{z}$, if

$$\mathcal{G}'(\hat{z}) C_{Z_{ad}}(\hat{z}) + C_{\mathcal{K}}(\mathcal{G}(\hat{z})) = W,$$

where $C_{Z_{ad}}(\hat{z})$ and $C_{\mathcal{K}}(\mathcal{G}\hat{z})$ are conical hulls (see Def. A.33) w.r.t. Z_{ad} at $\hat{z}$ and $\mathcal{K} := \{w \in W \,|\, w \geq 0 \text{ a.e.}\}$ at $-\mathcal{G}(\hat{z})$, resp.

It is equivalent to the Zowe-Kurcyusz CQ to require that

$$\alpha \mathcal{G}'(\hat{z})(z - \hat{z}) + \beta(v - \mathcal{G}(\hat{z})) = \omega$$

is solvable for any $\omega \in W$ with a $z \in Z_{ad}$, $v \in \mathcal{K}$, and $\alpha, \beta \geq 0$.

In case of pure equality constraints where $Z_{ad} = Z$ the latter CQ becomes the surjectivity of $\mathcal{G}'(\hat{z}) = H'(\hat{z}) : Z \to W$. In case of pure inequality constraints and if $\mathcal{G}(\hat{z}) >_{\mathcal{K}} 0$, this constraint qualification is not active at all.

Assumption A.42 *(Slater Condition [HPUU09, (1.132)])*
For convex optimization problems the Slater condition or constraint qualification (SCQ) reads: there exists a feasible point, that is strictly feasible w.r.t. the inequality constraints:

$$\exists z \in Z \;\; s.t. \;\; \mathcal{G}(z) \in \overset{\circ}{\mathcal{K}} \; \wedge \; z \in Z_{ad}.$$

For $Z_{ad} \subset Z$, closed and convex , $\mathcal{K} \subset W$ a closed convex cone, and $\mathcal{G} : Z \to W$ convex, this implies RCQ for all $\tilde{z} \subset Z_{ad}$ with $\mathcal{G}(\tilde{z}) \in \mathcal{K}$.

The Slater condition fails for pure equality constraints, but in this case the existence of Lagrange multipliers may be guaranteed by other arguments.

Assumption A.43 *(Linearized Slater Condition [Tr10, (6.18)])*
The linearized Slater condition or constraint qualification (LSCQ) is fulfilled at $\hat{z}$ if

$$\exists z \in Z_{ad} : \mathcal{G}(\hat{z}) + \mathcal{G}'(\hat{z})(z - \hat{z}) >_{\mathcal{K}} 0.$$

Here $>_{\mathcal{K}}$ is to be understood as defined in Def. A.32.

Note that in the finite-dimensional case a similar CQ, the local Slater condition, exists (see, e.g., [GL11, 6.3.6]), that is related to the tangent cone.

LSCQ is sufficient for the Zowe-Kurcyusz condition, see, e.g., [Tr10, Subsect. 6.1.2].

For optimal control problems, we have the following: let $\hat{z} = [\hat{y}, \hat{u}] \in \Sigma_{fs}$. If $H_y'(\hat{z}) \in \mathcal{L}(Y, W_H^*)$ is surjective, $G(\hat{z}) = G(\hat{y})$ and if there exists $u \in U_{ad}$ and $y \in Y$ with

$$\begin{aligned} G(\hat{y}) + G'(\hat{y})(y - \hat{y}) &\in \overset{\circ}{\mathcal{K}}, \\ [H_y'(\hat{z}), H_u'(\hat{z})]^\top [y - \hat{y}, u - \hat{u}] &= 0_{W_H}, \end{aligned}$$

the linearized Slater condition implies the Robinson CQ as well [HPUU09, Lemma 1.14].

Assumption A.44 *(Mangasarian-Fromowitz constraint qualification [Ge12, Coroll. 2.3.35])*
The Mangasarian-Fromowitz constraint qualification (MFCQ) reads:

a) Let $H'(\hat{z})$ be surjective and

b) let there exist a $\hat{d} \in int(Z_{ad} - \hat{z})$ with

$$\begin{aligned} &G'(\hat{z})(\hat{d}) \in int(\mathcal{K} - \{G(\hat{z})\}), \\ &H'(\hat{z})(\hat{d}) = 0_{W_H}, \end{aligned}$$

then RCQ holds.

For standard optimization problems as in Problem 2.1, MFCQ and RCQ coincide.

Assumption A.45 *(Linear Independence Constraint Qualification [Ge12, Coroll. 2.3.34]) The linear independence constraint qualification (LICQ) or also called surjectivity constraint qualification is fulfilled at $\hat{z}$, if $\hat{z} \in \mathring{Z}_{ad}$ and if*

$$T : Z \to W = W_G \times W_H, T(\hat{z}) := [G'(\hat{z}), H'(\hat{z})]^\top$$

is surjective.

The latter CQ states equivalently that the gradients of the active inequality constraints and the gradients of the equality constraints are linearly independent at $\hat{z}$.

The following implications hold in infinite dimensions:

$$\begin{array}{rcl}
LICQ \overset{\text{[GL11, 6.3.16]}}{\Longrightarrow} MFCQ \overset{\text{[Ge12, Coroll. 2.3.35]}}{\Longrightarrow} & & RCQ \overset{\text{[Ro76, Th. 1, Coroll. 2]}}{\Longrightarrow} ACQ, \\
MFCQ \overset{\text{Pb. 2.1}}{\Longleftrightarrow} & & RCQ \\
\text{SCQ} \overset{\text{convex pb. [HPUU09, 1.7.3.2]}}{\Longrightarrow} & & RCQ, \\
\text{LSCQ} \overset{\text{[Tr10, Subsect. 6.1.2]}}{\Longrightarrow} ZKCQ \overset{\text{[ZK79]}}{\Longrightarrow} & & RCQ.
\end{array}$$

Generally, it makes sense to imply weaker constraint qualifications, in order to obtain stronger optimality conditions.

There exist further constraint qualifications. For completeness we list these for finite-dimensional problems in an informal way:

Constant positive-linear dependence constraint qualification (CPLD): for every subset of gradients of the active inequality constraints and gradients of the equality constraints at $\hat{z}$ there holds: If a positive-linear dependence holds at $\hat{z}$, then there exists a positive-linear dependence within a neighbourhood of $\hat{z}$.

We have the implication $MFCQ \Longrightarrow CPLD$ in finite dimensions.

Constant rank constraint qualification (CRCQ): for every subset of gradients of the active inequality constraints and gradients of the equality constraints the rank is constant within a neighbourhood of $\hat{z}$.

Note that there holds $LICQ \Longrightarrow CRCQ \Longrightarrow CPLD$ in finite dimensions. However, MFCQ is not equivalent to CRCQ, nor weaker, nor stronger.

Nondegenerate constraint qualification (NDCQ): at $\hat{z}$ the rank of the Jacobian of the equality constraints and the inequality constraints that are actually zero (said to be *binding*) is as large as it can be.

We note that redundant equations should be eliminated before the NDCQ is checked. The NDCQ is quite popular in numerical codes.

Appendix B

Numerical Methods for Partial Differential Equations

B.1 Numerical Discretization

Any numerical approach requires finally a discretization, since the computers of today can handle only a finite number of variables. We introduce some basic notation on discretizations.

For iterates, in abstract algorithms in function space as well as in the fully discretized case, we use an upper index in brackets for the iteration step. For instance, $z^{(k)}$, $k = 0, 1, \ldots$, denotes the variable z at iteration k. The initial condition reads $z^{(0)} := z_0$, where z_0 is given. We wish to achieve a good approximation

$$z^{(k)} \approx \hat{z}(t^{(k)}),$$

where here $\hat{z}$ denotes the exact (typically unknown) solution.

A time interval $I = (t_0, t_f)$ is typically discretized by $\mathcal{N} + 1$ time points

$$t^{(k)} := t^{(k-1)} + h^{(k)},\ k = 1, \ldots \mathcal{N}, \quad t^{(0)} := t_0,$$

such that $t^{(\mathcal{N})} = t_f$. For an equidistant time grid, we have $h^{(k)} = h = (t_f - t_0)/\mathcal{N}$ for all $k = 1, \ldots, \mathcal{N}$. Since our approximate solution $z^{(k)}$ depends on the fineness $h := \max_k h^{(k)}$ of the discretization or the number $\mathcal{N}$, resp., this may be indicated by writing

$$z_h^{(k)} = z_{\mathcal{N}}^{(k)} = z^{(k)},$$

if necessary. The corresponding function value $f(t^{(k)}, z^{(k)})$ is denoted by

$$f_{\mathcal{N}}^{(k)} = f^{(k)} := f(t^{(k)}, z^{(k)}) \approx f(t^{(k)}, \hat{z}(t^{(k)})).$$

Unless mentioned otherwise, we work with equidistant time grids.

Similarly, we discretize in space. However, in dimensions larger than 1 the indexing of the space points is more complicated. For finite differences, it is obvious how to count the space points, but, for finite element methods (see Appendix B.3) or finite volume methods (e.g., [KA00]) this can be quite tricky. For instance, by the Cuthill-McKee algorithm a suitable re-numbering may be generated in order to reduce the bandwidth of the obtained FEM matrices.

In problems that live in time and space, we use upper indices for the time grid and lower indices for the space grid. For example, for the Saint-Venant equations in 1D we write for the discretized fluid level

$$h_{(j)}^{(i)} \approx h(t^{(i)}, x^{(j)}), \quad i = 1, \dots, \mathcal{M},\, j = 1, \dots, \mathcal{N}.$$

Note that it is often helpful to choose a different discretization for states and controls [Ge12, Subsect. 5.1.3].

For numerical methods for ODE and DAE, we refer to [HNW93, HW96, HLG06] for instance. In the next sections we give a short overview of the finite element method, summarizing the main results from the lecture note [Ki17]. For the finite volume method (FVM) and the finite difference method (FDM) that is an important special case of FVM, we refer, e.g., to the textbook [KA00].

B.2 Galerkin Methods

For the underlying theory on PDE see Appendix A.1.3 for a short presentation or, e.g., [LM72, Ev10].

We consider a general variational problem

$$\text{Find } u \in V: \quad a(u, v) = L(v) \quad \forall v \in W, \tag{B.1}$$

where V, W are vector spaces, $a : V \times W \to \mathbb{R}$ is a bilinear form, and $L : V \to \mathbb{R}$ a linear form. If $V = W$ and a is positive (i.e. $a(v, v) \geq 0 \, \forall v \in V$), symmetric, then (B.1) is equivalent to the minimization problem

$$\text{Minimize } J(v) = \frac{1}{2} a(v, v) - L(v) \text{ in } V. \tag{B.2}$$

A *Petrov-Galerkin method* relies on the following ansatz for a discretization. V is replaced by a finite-dimensional space V_h, the *solution or trial space*. Analogously, W is replaced by W_h, the *test space*. Instead of (B.1) we solve

$$\text{Find } u_h \in V_h: \quad a(u_h, v_h) = L(v_h) \quad \forall v_h \in W_h. \tag{B.3}$$

If $V = W$ and a is symmetric, this method is called the *(Bubnov-) Galerkin method*.

The *Ritz or Ritz-Galerkin method* is the ansatz, where instead of (B.2) the finite-dimensional minimization problem

$$\text{Minimize } J(v_h) = \frac{1}{2} a(v_h, v_h) - L(v_h) \text{ in } V_h, \tag{B.4}$$

is solved. For a positive (i.e. $a(v, v) \geq 0$), symmetric bilinear form the Ritz method is equivalent to the Galerkin method. Please note that the denomination of the different Galerkin-type methods is not uniform in literature.

In contrast to finite difference methods, for all Galerkin-type methods only the spaces V and W are discretized, but not the differential operator. In the following we consider only the case $W = V$, unless mentioned otherwise.

We are interested in the quality of the approximation that is provided by the numerical solution relying on a Galerkin method.

An *a priori estimate* is an estimate that may be derived without knowing any solution before. In contrary, an *a posteriori estimate* relies on the availability of a computed numerical (or known analytical) solution.

Lemma B.1 *(Stability of Galerkin Method)*
Let V be a Hilbert spaces and $a : V \times V \to \mathbb{R}$ be a bounded and V-elliptic bilinear form (with ellipticity constant α_0). Regardless of the choice of a subspace $V_h \subset V$ there holds for a solution u_h of (B.3) that

$$\|u_h\|_V \leq \frac{1}{\alpha_0} \|f\|_{V^*}. \tag{B.5}$$

Proof. Since V_h is a closed subset of V and itself a Hilbert, the boundedness and coercivity of a holds in V_h as well as in V for the same constants.
Testing with the solution, i.e. inserting $v_h = u_h$ in (B.3) yields

$$a(u_h, u_h) = \langle f, u_h \rangle, \tag{B.6}$$

and by means of the V-ellipticity and a standard estimate for linear forms we obtain

$$\alpha_0 \|u_h\|_V^2 \leq \|f\|_{V^*} \|u_h\|_V. \tag{B.7}$$

If $\|u_h\|_V = 0$, then we are done, otherwise the statement of the lemma follows by cancellation of $\|u_h\|$. □

Thus the stability of the solution of problem (B.1) is guaranteed directly by the Galerkin method (B.3).

Lemma B.2 *(Céa)*
Let V be a Hilbert space and $a : V \times V \to \mathbb{R}$ a V-bounded and V-elliptic bilinear form (bounded with constant α_1 and with ellipticity constant α_0). If further u or u_h are solutions of the variational problem V or $V_h \subset V$, respectively, (i.e. of (B.1) or (B.3), then there holds

$$\|u - u_h\|_V \leq \frac{\alpha_1}{\alpha_0} \inf_{v_h \in V_h} \|u - v_h\|_V. \tag{B.8}$$

Proof. The solutions u or u_h, resp., are defined as

$$a(u,v) = L(v) \quad \forall v \in V, \tag{B.9}$$

$$a(u_h,v) = L(v) \quad \forall v \in V_h. \tag{B.10}$$

Since $V_h \subset V$, Equation (B.9) holds for all $v \in V_h$. By subtracting (B.10) from (B.9) where $v \in V_h$:

$$a(u - u_h, v) = 0 \quad \forall v \in V_h. \tag{B.11}$$

Now let $v_h \in V_h$ be arbitrary, we insert $v = v_h - u_h \in V_h$ into (B.11) and obtain

$$a(u - u_h, v_h - u_h) = 0 \quad \forall v_h \in V_h. \tag{B.12}$$

This allows for the estimate

$$a(u - u_h, u - u_h) = a(u - u_h, u - v_h) + a(u - u_h, v_h - u_h) \tag{B.13}$$

$$\leq \alpha_1 \|u - u_h\|_V \|u - v_h\|_V + 0, \tag{B.14}$$

exploiting the bilinearity, the boundedness and (B.11).

On the other hand the V-ellipticity implies

$$\alpha_0 \|u - u_h\|_V^2 \leq a(u - u_h, u - u_h). \tag{B.15}$$

If $\|u - u_h\|_V = 0$, then (B.8) holds. Otherwise we combine the latter two estimates and by cancellation

$$\alpha_0 \|u - u_h\|_V \leq \alpha_1 \|u - v_h\|_V \quad \forall v_h \in V_h. \tag{B.16}$$

□

Eq. (B.11) is the so-called *Galerkin orthogonality*. For a symmetric it says, that the approximation error $u - u_h$ is orthogonal to V_h.

The importance of the *Céa lemma* is that the a-priori error of the numerical solution can be determined within the space of approximation functions.

Thus we should choose function spaces that allow for a good approximation of the solution u. For instance, considering polynomial spaces, it depends on the smoothness of the solution, how well we may approximate.

For boundary value problems typically the regularity decreases near the boundary. Thus we cannot expect to obtain a better approximation by increasing again and again the polynomial degrees. Here it makes more sense to consider piecewise polynomials and to increase the accuracy by refining the mesh. This is the concept of hp-methods that we do not wish to discuss here with further details.

The finite element method is a specific Galerkin method. The bounded domain Ω is decomposed into simple compact subsets (cells or elements) T_i, $1, \ldots, M$ with $\overset{\circ}{T} := int(T) \neq \emptyset$. For instance, triangles ($d = 2$) or tetrahedrons ($d = 3$) are used, but also quadrangles ($d = 2$) or cuboids ($d = 3$) are common. The decomposition (mesh or grid) $\mathcal{T}_h$, in the case of triangles we call this a triangulation, has to be admissible in a certain way.

B.3 Finite Element Method

For simplicity we formulate these postulations for $d = 2$ and $d = 3$:

Definition B.3 *(Admissible Mesh in 2 Dimensions)*
Let Ω be a polygonal domain that may be decomposed into triangles and quadrangles.

Then a mesh $\mathcal{T}_h = \{T_1, T_2, \ldots, T_M\}$ with triangles and quadrangles T_i is called admissible (or geometrically conform), if:

(i) $\overline{\Omega} = \cup_{i=1}^{M} T_i$, *i.e. the mesh covers exactly the closure of the domain.*

(ii) If $T_i \cap T_j$, $i \neq j$, consists of a single point, then this point is a corner point of T_i and of T_j, i.e. no hanging nodes appear.

(iii) If $T_i \cap T_j$, $i \neq j$, consists of more than one point, then $T_i \cap T_j$ is an edge of T_i and of T_j.

(ii) and (iii) mean that two different elements intersect at most in one corner or on one edge.

Definition B.4 *(Admissible Mesh in 3 Dimensions)*
Let Ω be a polyedric domain that may be decomposed into tetrahedrons and cuboids.

Then a mesh $\mathcal{T}_h = \{T_1, T_2, \ldots, T_M\}$ with tetrahedrons and cuboids T_i is called admissible (or geometrically conform), if:

(i) $\overline{\Omega} = \cup_{i=1}^{M} T_i$, *i.e. the mesh covers exactly the closure of the domain.*

(ii) If $T_i \cap T_j$, $i \neq j$, consists of a single point, then this point is a corner point of T_i and of T_j, i.e. no hanging nodes.

(iii) If $T_i \cap T_j$, $i \neq j$, consists of more than one point, then $T_i \cap T_j$ is
an edge of T_i and of T_j
or a face of T_i and of T_j,
i.e. no hanging nodes.

The index h at $\mathcal{T}_h$ designates that every element has a diameter smaller than $2h$ (2D) or $3h$ (3D), resp.

The parts of the boundary ∂T_i that lie within in a hyperplane are called facets. In 2D these are edges or in 3D these are faces.

Remark B.5 *(Curvilinear Boundaries)*
Due to the assumption that elements T_i are polygons or polyhedrons, we may cover exactly only

domains Ω that are bounded polygonally (2D) or by polygons (3D). An Ω of this type is called a domain with <u>*P-boundary*</u>*. The exhaustion of an Ω with curvilinear boundaries may be improved by the introduction of curvilinear elements.*

We look for basis functions ϕ_i, $1, \dots, \mathcal{N}$, such that as many as possible entries in the resulting stiffness matrix $A_{ij} := a(\phi_j, \phi_i)$ vanish, yielding a linear equation system that is "easily" solvable. This may be realized by basis functions with support as small as possible. Then the supports of different basis functions intersect only in the direct neighbourhood and A is sparse, meaning only a few entries are non-zero.

Example B.6 *(FEM Yields Linear Equation System)*
Let the nodal functions ϕ_i, $i = 1, \dots, \mathcal{N}$, be a global basis of V_h such that $\phi_i = \delta_{x_i}(x)$ for all nodes x_i, $i = 1, \dots, \mathcal{N}$. For the ansatz

$$u_h(x) = \sum_{i=1}^{\mathcal{N}} u_i \phi(x),$$

the problem (B.3) (in case $W = V$) becomes the linear system

$$Au = b, \tag{B.17}$$

where

$$\begin{aligned} A_{ij} &= a(\phi_j, \phi_i), && i, j = 1, \dots, \mathcal{N}, \\ b_i &= L(\phi_i), && i = 1, \dots, \mathcal{N}. \end{aligned}$$

In order to obtain a sparse A, the nodal functions may be chosen piecewise smooth such that the support of any ϕ_i is the union of the cells that intersect with x_i only. These nodal functions with hat-like shape have finite support. The supports overlap only for nodes that are neighbours, the latter depending on the precise geometry of the mesh.

Please note that it is often practice to designate the functions with finite support as finite elements (FE), too, whereas the subsets T_i are called just elements. If we speak of C^k-elements or linear elements, for instance, then the corresponding functions are to be meant.

Except for being admissible, the choice of the mesh $\mathcal{T}_h$ is arbitrary and can be adapted to the underlying problem, e.g. in case of framework constructions.

Definition B.7 *(Abstract Definition of a Finite Element According to Ciarlet [Ci78])*
A finite element is a triplet (T, Π, Σ) with the following properties:

(i) T *is a compact polyhedron in* $\mathbb{R}^d$ *with* $\mathring{T} \neq \emptyset$.

(ii) Π *is a subspace of* $C^0(T; \mathbb{R}^m)$ *for some* $m \in \mathbb{N}$ *with finite dimension* $s \geq 1$.

(iii) Σ *is a set of* s *linear independent functionals on* Π.
Any $p \in \Pi$ *is determined by the values of the* s *functionals in* Σ.

T is called element (or cell).

The functions in the space Π are called *(local) trial functions*. Typically, Π consists of polynomials, however, piecewise polynomials are used as well, e.g. for the HCT-element.

The elements $\sigma_i \in \Sigma$ are called *(local) degrees of freedom* (or nodal variables). Since these functionals correspond typically to function values and derivatives at certain points in T, we call (iii) generalized interpolation conditions. Note that the functionals in Σ are of the type of evaluation operators in Ex. 3.3.

A FE is called *Lagrange element*, if all degrees of freedom are function evaluations in points (cf. the Lagrange interpolation problem). We speak of a *Hermite element*, if derivatives of a function are evaluated as well.

Lagrange elements for triangles (2D)/tetrahedron (3D) with polynomials of maximal total degree q are denoted by $\mathbb{P}_q$. In case of quadrangles (2D)/cuboids (3D) with maximal degree q in each variable we use the symbol $\mathbb{Q}_q$. For intervals (1D) both types of Lagrange elements coincide, i.e. $\mathbb{P}_q = \mathbb{Q}_q$.

In principle Π and Σ depend on the element T. This can be emphasized by the notation (T, Π_T, Σ_T) for a $T \in \mathcal{T}_h$. The set of finite elements $(T, \Pi_T, \Sigma_T)_{T\in\mathcal{T}_h}$ is called a FE complex. Furthermore, we introduce the global interpolation domain $V_{\mathcal{T}} := \prod_{T\in\mathcal{T}} V_T$ and $\Pi_{\mathcal{T}} := \prod_{T\in\mathcal{T}} \Pi_T$, being the approximation space for the FE complex.

Definition B.8 *(FE-Space; Global Degrees of Freedom)*
Let $\{T, \Pi_T, \Sigma_T\}_{T\in\mathcal{T}}$ be a FE complex on a mesh $\mathcal{T}$, let $\Pi_{\mathcal{T}}$ be the corresponding approximation space, and let $\Sigma_{\mathcal{T}} := \cup_{T\in\mathcal{T}}\Sigma_T$.

(i) Let $\sigma \in \Sigma_T$ for some $T \in \mathcal{T}$, then $\{\Sigma^1, ..., \Sigma^S\}$ designates a partition of $\Sigma_{\mathcal{T}}$, i.e. a decomposition of $\Sigma_{\mathcal{T}}$ into non-empty disjoint subsets (equivalence classes).

(ii) The subspace

$$V_h := \left\{ v \in \Pi_{\mathcal{T}} \,\middle|\, \sigma, \tilde{\sigma} \in \Sigma^i \text{ for some } i \in \{1, ..., S\} \Leftrightarrow \tilde{\sigma}(v) = \sigma(v) \right\}$$

of $\Pi_{\mathcal{T}}$ is called a finite element-space (FE-space) or global trial space.

(iii) The representatives $\{\sigma_1, ..., \sigma_S\}$ of the equivalence classes $\{\Sigma^1, ..., \Sigma^S\}$ are called global degrees of freedom.

The elements $\phi_1, ..., \phi_S \in V_h$ with $\sigma_i(\phi_j) = \delta_{ij}$ are called *global trial functions* and provide a *(global) nodal basis* of V_h.

Definition B.9 *(Conformal Approximation)*
For a general variational problem (B.3) in a Hilbert space V, the vector space V_h is called

V-conformal, if $V_h \subset V$.
Then the discrete problem (B.3) is called a conformal approximation.

In general the space Π is not in $H^1(\Omega)$, and we have not always a H^1-conformal approximation.

Suitable interpolation error estimates help us to decide how small h has to be chosen at least, in order to guarantee a given precision of the numerical solution. This will depend on the type of finite element and on the intrinsic regularity of the PDE problem.
We recall the Céa lemma, Lemma (B.2), and estimate the right-hand side further

$$\|u - u_h\|_V \leq C\|u - \pi_h u\|_V. \tag{B.18}$$

The idea is choose a manageable global interpolation in V_h for the projection $\pi_h : V \to V_h$.

Let $I_T : V_T \to \Pi_T$ be an interpolation operator corresponding to a single element T. We consider a modified global interpolation operator I_{V_h} that guarantees

$$[I_{V_h} v]\,\big|_T = I_T(v|_T). \tag{B.19}$$

For this purpose, let

$$\tilde{V_\mathcal{T}} = \big\{ v \in V_\mathcal{T} \,\big|\, \sigma, \tilde{\sigma} \in \Sigma^i, i \in \{1, \dots, S\} \;\Rightarrow\; \tilde{\sigma}(v) = \sigma(v) \big\} \tag{B.20}$$

be the subspace of $V_\mathcal{T}$, in which global basis functions for coincident degrees of freedom are identified. Then the FE-space interpolation reads

$$\tilde{V_\mathcal{T}} \ni v \mapsto I_{V_h} v = \sum_{i=1}^{\mathcal{N}} \sigma_i(v)\phi_i \in V_h. \tag{B.21}$$

This guarantees (B.19).

We introduce the FE-space interpolation into (B.18) as projection, i.e. setting $\pi_h = I_{V_h}$, where we presuppose in addition that u is sufficiently smooth, such that $u \in \tilde{V_\mathcal{T}}$.

We observe the typical structure of the error estimate

$$\|u - u_h\| \leq ch^p|u|, \tag{B.22}$$

where $|\cdot|$ is a (semi-)norm that considers higher derivatives as $\|\cdot\|$ does. The maximal order of convergence depends on the used polynomial degree $\mathcal{P}_m$ and on the regularity of the solution, i.e. on $H^{m+1}(\Omega)$.

We consider the order of convergence for $\mathbb{P}_q$- or $\mathbb{Q}_q$-elements, resp. We make the following assumption.

Assumption B.10 *(Assumption on FE-Spaces)*
In this subsection we assume in general that (T, Π, Σ) is a finite element with local interpolation space V_T, whereas

(i) the space of polynomials of degree m on the element $\mathcal{P}_m(T)$ is a subset of Π,

(ii) $V_T = H^{m+1}(\mathring{T})$,

with $m \in \mathbb{N}$.

Let $d = 2, 3$, $u \in H^{m+1}(\Omega)$, $m+1 \geq 2$, then the order of convergence $m = q$ is optimal w.r.t. the H^1-norm ($k = 1$). Hermite elements, e.g., payoff only for high regularity of the solution, since $m + 1 = 2, 3, 4$ in 1D and $m + 1 = 3, 4$ in 2D/3D is allowed.

For global error estimates we need a certain uniformity of the mesh. We consider the 2D case. Let h_T or h denote the maximal diameter of a cell or all cells in a mesh, resp. Furthermore, we need the maximal diameter ρ_T of circles that may be inscribed in a cell T.

Definition B.11 *(Uniform Mesh)*
Let $\Upsilon := \{\mathcal{T}_h\}_{h\searrow 0}$ be a family of meshes with the property that for all $h_n \searrow 0$ a mesh $\mathcal{T}_{h_n}$ exists in this family.

(i) Such a family $\Upsilon := \{\mathcal{T}_h\}_{h\searrow 0}$ is called shape regular, if there exists a constant $\gamma > 0$, such that

$$\frac{h_T}{\rho_T} = \gamma_T \leq \gamma \quad \forall T \in \mathcal{T}_h \quad \forall \mathcal{T}_h \in \Upsilon. \tag{B.23}$$

(ii) The family Υ is called (quasi-)uniform, if there exists a constant $\gamma > 0$, such that

$$\frac{h}{\rho_T} \leq \gamma \quad \forall T \in \mathcal{T}_h \quad \forall \mathcal{T}_h \in \Upsilon \tag{B.24}$$

or equivalently, Υ is shape-regular and there exists a $c > 0$ such that

$$\frac{h}{h_T} \leq c \quad \forall T \in \mathcal{T}_h \forall \mathcal{T}_h \in \Upsilon \tag{B.25}$$

The latter means that the ratio between the largest and the smallest diameter remains bounded over all meshes in the family.

Note that we follow [EG04] for the definition of quasi-uniformity. However, note that in some literature ii) it is called quasi-uniform.

We state three standard results on estimates, see, e.g., [Ki17] and the references therein. The proofs rely on the Bramble-Hilbert lemma.

Lemma B.12 *(Global Interpolation Error Estimates for Lagrange Elements in 2D)*
Let $\mathcal{T}_h$ be a family of shape-regular meshes and we assume a conformal FE-approximation. There holds for a:

Lagrange triangular element $\mathbb{P}_1$ *and* $u \in H^2(\Omega)$

$$\|u - u_h\|_{0,\Omega} \leq ch^2 |u|_{2,\Omega}, \quad |u - u_h|_{1,\Omega} \leq ch|u|_{2,\Omega}, \tag{B.26}$$

Lagrange triangular element $\mathbb{P}_2$ *and* $u \in H^3(\Omega)$

$$\|u - u_h\|_{0,\Omega} \leq ch^3|u|_{3,\Omega}, \quad |u - u_h|_{1,\Omega} \leq ch^2|u|_{3,\Omega}, \quad |u - u_h|_{2,\Omega} \leq ch|u|_{3,\Omega}, \tag{B.27}$$

Lagrange triangular element $\mathbb{P}_2$ *and* $u \in H^2(\Omega)$

$$\|u - u_h\|_{0,\Omega} \leq ch^2|u|_{2,\Omega}, \quad |u - u_h|_{1,\Omega} \leq ch|u|_{2,\Omega}, \tag{B.28}$$

Lagrange quadrilateral element $\mathbb{Q}_1$ *and* $u \in H^2(\Omega)$

$$\|u - u_h\|_{0,\Omega} \leq ch^2|u|_{2,\Omega}, \quad |u - u_h|_{1,\Omega} \leq ch|u|_{2,\Omega} \tag{B.29}$$

(i.e. estimates as with $\mathbb{P}_1$*).*

Remark B.13 *(L^2-Error Estimate)*
By means of the Céa lemma we obtain for an elliptic boundary value problem of second-order an error estimate w.r.t. the $H^1(\Omega)$ norm. However, the simple estimate

$$\|u - u_h\|_{L^2(\Omega)} \leq \|u - u_h\|_{H^1(\Omega)} \leq ch^{\mu-1}|u|_{\mu,\Omega} \tag{B.30}$$

doesn't yield the best order of the error ($k = 1$, $\mu := m + 1$). The "correct" L^2 error estimate

$$\|u - u_h\|_{L^2(\Omega)} \leq ch\|u - u_h\|_{H^1(\Omega)} \leq ch^{\mu}|u|_{\mu,\Omega} \tag{B.31}$$

is proven by the Aubin-Nitsche lemma relying on a duality argument (the so-called "Nitsche trick"). Here we refer to [Br07].

For the L^∞ norm we may prove

Lemma B.14 *(Global L^∞-Error Estimate)*
For $\mathbb{P}_k$ elements there holds for $k = 1$

$$\|u - u_h\|_{L^\infty(\Omega)} \leq ch^2|\ln(h)|\,|u|_{W^{2,\infty}(\Omega)}, \tag{B.32}$$

and for $k \geq 2$

$$\|u - u_h\|_{L^\infty(\Omega)} \leq ch^{k+1}\,|u|_{W^{k+1,\infty}(\Omega)}. \tag{B.33}$$

Remark B.15 *(Numerical Linear Algebra)*
An important issue in FEM is the efficient numerical solution of the arising linear equation systems of type (B.17). Since direct methods have limitations in the system size, for large-scale systems iterative methods are considered. The numerical accuracy and convergence depends on the condition of the matrix A that may be influenced suitably by preconditioning. Among various methods the following are very feasible:

- *conjugated gradient methods with good preconditioning,*
- *multigrid methods [H85],*
- *fast Fourier transformation, see, e.g., [GR94, pp. 250-255],*
- *H matrices [B08].*

For an overview of further numerical methods in optimization not discussed in this work, see, e.g., [GK99, GK02] in finite dimensions and [NW06, LY08, Be10, BS12] in the general case.

Appendix C

Lagrange and Hamilton Approach in Classical Mechanics

In this appendix we comment on the close relation of the notions of Lagrange and Hamilton functions that appear in physics on one hand and in optimization and control theory on the other hand, see also, e.g., [GKW14]. For a discussion on the history of optimal control and its notion, see, e.g., [PB94, SW97].

Definition C.1 *(Generalized Coordinates and Holonomic Constraints)*
Let $I \subset \mathbb{R}$ represent a time interval. For each mass point, we transform the given coordinates $[x_1, \ldots, x_{n_x}]$, e.g. points for positions and angles for the orientation, to generalized coordinates $[q_1, \ldots, q_{n_q}]$ that are more suitable for the problem. Let the vector q denote the generalized coordinates.

In most general form, a constraint $H(t, q, \dot{q}) = 0$, where $H : I \times \mathbb{R}^{2n_q} \to \mathbb{R}^{n_H}$, is said to be holonomic, if its dependence on the velocity coordinates $\dot{q} := q'(t)$ may be eliminated by integration. Therefrom, w.l.o.g., we may restrict us to the case $H(t, q) = 0$ in the following. For holonomic constraints it is always possible to find canonical generalized coordinates q, s.t. all forces may be expressed as gradients of generalized potentials.

However, in principle, non-holonomic constraints could partly be considered by the following techniques as well.

Problem C.2 *(System of Mass Points)*
We consider on a time interval I a certain multibody system with mass points. The system is subject to holonomic, independent constraints $H(t, q) = 0$, where the positions and orientations of the mass points are given by the vector q of canonical generalized coordinates. We wish to determine the equations of motion for q and $\dot{q}$ and, then, we would like to solve for q as a function of t.

The kinetic energy of this system is denoted by $\mathcal{E}_{kin}$, its generalized potential (potential energy) by $\mathcal{E}_{pot}$, and its total energy by $\mathcal{E} = \mathcal{E}_{kin} + \mathcal{E}_{pot}$.

Definition C.3 *(Lagrange Function in Classical Mechanics)*
The function $\tilde{L} : I \times \mathbb{R}^{2n_q} \to \mathbb{R}$

$$\tilde{L}(t, q, \dot{q}) := \mathcal{E}_{kin}(t, q, \dot{q}) - \mathcal{E}_{pot}(t, q)$$

is called Lagrange function (Lagrangian) w.r.t. Problem C.2 (with holonomic constraints).

The equations of motion for this system are obtained by the *Lagrange equations (of the second kind)*, customarily written as

$$\frac{d}{dt}\frac{\partial \tilde{L}(t, q, \dot{q})}{\partial \dot{q}_i} - \frac{\partial \tilde{L}(t, q, \dot{q})}{\partial q_i} = 0 \quad \forall i = 1, \ldots, n_q$$

or in our notation reading

$$\frac{d}{dt}\tilde{L}'_{\dot{q}}(t, q, \dot{q}) - \tilde{L}'_q(t, q, \dot{q}) = 0_{\mathbb{R}^{n_q}}. \tag{C.1}$$

Note that often the Lagrange equations of the second kind are stated for the special case $\tilde{L} \equiv \mathcal{E}_{kin}$.

For completeness, in case of non-holonomic constraints $\tilde{H} : I \times \mathbb{R}^{2d} \to \mathbb{R}^{n_{\tilde{H}}}$ and for Cartesian coordinates x, the *Lagrange equations of the first kind* read

$$m\ddot{x} = F_a + \lambda^\top \tilde{H}'_x + \mu^\top \tilde{H}'_{\dot{x}}, \tag{C.2}$$

where $F_a : I \to \mathbb{R}^{n_x}$ denotes active forces and the multipliers are λ, $\mu \in \mathbb{R}^{n_{\tilde{H}}}$. The last two terms are the total constraining force $F_c : I \times \mathbb{R}^{2d} \to \mathbb{R}^{n_H}$, that can be expressed as a linear combination of gradients of $\tilde{H}$.

The relation to the Lagrange function L as used in optimization subject to constraints, see Def. 2.12, becomes obvious when we set here

$$L(t, q, \dot{q}) := \tilde{L}(t, q, \dot{q}) + \lambda^\top H(t, q),$$

yielding that (C.1) is formally equivalent to

$$\frac{d}{dt}L'_{\dot{q}}(t, q, \dot{q}) - L'_q(t, q, \dot{q}) = 0,$$

being the Lagrange equation of the second kind, but w.r.t. L. The constraining force naturally appears as a product of Lagrange multipliers (adjoints) and constraint equations.

Conjugated to q, we introduce the *generalized momenta*, given by the vector $p := L'_{\dot{q}}$.

Definition C.4 *(Hamilton Function in Classical Mechanics)*
The Legendre transformation, where $\dot{q}$ *is replaced as independent variable by* p*, applied to the Lagrange function to Problem C.2, i.e.,*

$$\tilde{\mathcal{H}}(t, q, p) := \dot{q}(t, q, p)^\top p - \tilde{L}(t, q, \dot{q}(t, q, p))$$

is called the Hamilton function (Hamiltonian) corresponding to Problem C.2.

If scleronomic constraints (i.e. no explicit time dependence of the constraints) are implied, $\tilde{\mathcal{H}}$ corresponds to the total energy of the system, $\mathcal{E}_{kin} + \mathcal{E}_{pot}$, considered as a function of generalized position q and momentum p.

According to the Hamilton principle, the action

$$\int_{t_0}^{t_f} L(t, q, \dot{q})\, dt \tag{C.3}$$

has a stationary point, corresponding to the fact that a physical process tends to a stationary point - usually we think of a minimum or maximum, resp., but not of a saddle point (that is also instable, but possible as well). The Euler-Lagrange equations (of the second kind) may be derived from the Hamilton principle as well. In this setting the equations of motion are obtained as

$$\begin{aligned} \dot{q} &= \tilde{\mathcal{H}}'_p, \\ \dot{p} &= -\tilde{\mathcal{H}}'_q. \end{aligned}$$

Here we observe some similarity to optimal control theory of DAE, where the states q and adjoints λ solve (for Problem C.2, i.e. without constraints on u)

$$\begin{aligned} \dot{q} &= \mathcal{H}'_\lambda, \\ \dot{\lambda} &= -\mathcal{H}'_q, \\ 0 &= \mathcal{H}'_u. \end{aligned}$$

The additional latter equation represents the stationarity w.r.t. u of the Hamilton function $\mathcal{H}(t, q, \lambda, u)$, depending here on u as well. In this setting momenta may be interpreted as adjoint and *vice versa*.

Please note that for Problem C.2 the Hamilton function $\mathcal{H}$, as defined in Def. 2.64 within the context of optimal control subject to an ODE $\dot{q} = f(t, q, u)$,

$$\mathcal{H}(t, q, u, \lambda) = \lambda_0 \phi(t, q, u) + \lambda_f^\top f(t, q, u)$$

is of slightly different character than $\tilde{\mathcal{H}}$. If we assume $\lambda_0 = -1$ for a formal equivalence, then λ_f mimics the role of p as conjugated variable. We introduce the new state $\tilde{q} := [q, v] := [q, \dot{q}]$ by adding the velocities as independent states. The holonomic constraint reads $H(t, \tilde{q}, u) = f(t, q, u) - v = 0$. The Legendre transformation is performed from $v = \dot{q}$, being equal to the r.h.s. f of the corresponding differential equation, to λ_f corresponding to p. Here $\mathcal{J}$ is the action (C.3) and has the dimension energy times time, and ϕ is the Lagrange function L. The function values L, $\tilde{L}$, $\mathcal{H}$, and $\tilde{\mathcal{H}}$ have usually the dimension of energy. If some constraint qualification is satisfied, we may assume w.l.o.g. $\lambda_0 = -1$ corresponding to maximization of the Hamilton function, if the Lagrange function is to be minimized; for a minimization of the Hamilton function, $\lambda_0 = 1$ would be appropriate, if the Lagrange function is to be minimized. In the latter case $p = -\lambda_f$ and $H = v - f$ can be set for consistency [Tr10, Subsect. 4.8.2].

In optimal control theory the objective $\mathcal{J}$ is minimized. In some situations the kinetic energy $\propto \mathcal{E}_{kin}$ or the potential energy $\mathcal{E}_{pot}$ times a weighting factor enter as a summand into $\mathcal{J}$. However, $\mathcal{J}$ can be chosen more general.

Further formulations of classical mechanics rely on Newton's laws of motion or the Hamilton-Jacobi equation, both equivalent to Lagrangian and Hamiltonian mechanics discussed above. In control theory the Hamilton-Jacobi-Bellman equation is the counterpart to the Hamilton-Jacobi equation in classical mechanics.

We close with an example for a multibody system from mechanics. For the notation see also Section 2.6.1 though it is restricted to index-1 DAE. A typical multibody system has a structure of DAE form.

Example C.5 *(Multibody System)*

$$M(q)\ddot{q}(t) = f_1(t, q, \dot{q}, u) + \lambda(t)^\top f'_{2;q}(t, q) \tag{C.4}$$

$$0_{\mathbb{R}^{n_{q2}}} = f_2(t, q), \tag{C.5}$$

where q is the ODE state and u the control. f_1 is the vector of applied forces, torques and Coriolis forces, whereas (C.5) represents (holonomic) constraints, here stated as a constraint on the position level. The algebraic variable, here λ, plays the role of a constraining force in physics and appears here as a multiplier. The symmetric and positive definite matrix M is the so-called mass matrix. From the kinetic energy of the multibody system, (C.4) may be derived by means of the Euler-Lagrange equations.

By multiplication with M^{-1} from the left, this system exhibits the form of a so-called Hessenberg DAE of order 3. For example, rewriting Example 1.1 as a DAE yields this structure.

Appendix D

Acronyms

Here we give a compilation about the most important symbols, abbreviations, nomenclature, physical constants and material data that we use. Please note that we would prefer to stay in line with some standard notations, so some ambiguities are inavoidable, e.g., p might denote a parameter to be identified, the momentum, the pressure or the exponent of functions in Lebesgue spaces. In literature p denotes often the adjoint, a nomenclature not used here. However, it should be clear from the context which definition is applicable in each case.

Symbol	**Declaration**
Abbreviations	
w.r.t.	with respect to
s.t.	such that
i.e.	*ita est* (latin), that is
w.l.o.g.	without loss of generality
b.c.	boundary condition
ODE	ordinary differential equation
PDE	partial differential equation
DAE	differential algebraic equation
PDAE	partial differential algebraic equation
CDE	coupled differential equations, i.e. coupled ordinary and partial differential equations
FDM	finite difference method
FVM	finite volume method
FEM	finite element method

Symbol	Declaration
Abbreviations (continued)	
LP	linear program
NLP	nonlinear program
SQP	sequential quadratic programming
SSNM	semismooth Newton method
MPC	model predictive control
CQ	constraint qualification
BFGS	Broyden-Fletcher-Goldfarb-Shanno (formula/update/method)
DFP	Davidon-Fletcher-Powell (formula/update/method)
KKT	Karush-Kuhn-Tucker (condition/point)
PNP	Poisson-Nernst-Planck (equation)
PEM	polymer electrolyte membrane
GaAs	gallium arsenide

Symbol	Declaration
Geometry	
$\mathcal{S} \subset \mathcal{R}$	a subset $\mathcal{S}$ of $\mathcal{R}$, also denoted $\mathcal{S} \subseteq \mathcal{R}$
$\mathcal{S} \subsetneq \mathcal{R}$	a proper subset $\mathcal{S}$ of $\mathcal{R}$, i.e. with $\mathcal{S} \neq \mathcal{R}$, also denoted $\mathcal{S} \subseteq \mathcal{R}$ in literature
$\mathring{\mathcal{S}} = int(\mathcal{S})$	interior of a set $\mathcal{S}$
$\emptyset$	empty set
Ω	an open bounded set or a domain (i.e. an open, non-empty, connected set) in a topological space
t	$\in \mathbb{R}_0^+$, time
$t_f = \mathcal{T}$	terminal time (final time), i.e. maximal t, until which a model is considered
d	$\in \mathbb{N} \setminus \{0\}$, the spatial dimension
x	Eulerian coordinates of $\Omega(t) \subset \mathbb{R}^d$
$\mathbf{X}$	Lagrangian coordinates of $\Omega(0) \subset \mathbb{R}^d$
e_i	unit vectors, $i = 1, \ldots, d$
ν	outer normal vector
τ_l	tangential vectors,$l = 1, \ldots, d-1$
k_M	$:= -\text{div}\ \nu$, mean curvature of a surface

Vectors are only underlined, e.g., $\underline{e}_1$, in some examples of Chapter 4.

Symbol	Declaration
Spaces	
$\|\cdot\|$	norm; the Euclidean norm unless stated otherwise
X	a normed vector space, equipped with a norm $\|\cdot\|_X$; typically a Banach space
X^*	(topological) dual space of X
0_X	Zero element of (Banach) space X (could be, e.g., the zero function)
$d(\cdot,\cdot)$	$:= \|\cdot - \cdot\|$, metric induced by norm
$\mathcal{L}(X,Y)$	space of linear operators, defined on X mapping into Y
Z	space of optimization variables, i.e. $Z = Y \times U$
Z_{ad}	admissible set of optimization variables, see Remark 2.4
$\Sigma_{(fs)}$	feasible set of optimization variables, see Remark 2.4
Y	space of states
Y_{ad}	subset of admissible states
W	space where the state equation $E(z) = 0$ is solved
W^*	space of adjoints, being the dual of W
U	space of controls
U_{ad}	subset of admissible controls
$C^{r,\alpha}$	Hölder spaces, $r \in \mathbb{N}_0$, $0 \leq \alpha \leq 1$
L^p	Lebesgue spaces, consisting of functions that are integrable to the power $1 \leq p \leq \infty$
$W^{k,p}$	Sobolev spaces, consisting of functions that are $0 \leq k < \infty$ times weakly differentiable, $1 \leq p \leq \infty$
H^k	$:= W^{k,2}$, Sobolev spaces that are Hilbert spaces, $0 \leq k < \infty$
α	Tikhonov regularization parameter for control costs or penalty parameter
$\tilde{\gamma}$	parameter entering in projection formulation of variational inequalities, typically $\tilde{\gamma} = 1/\alpha$
$X \hookrightarrow Y$	embedding from space X into space Y, in literature also denoted as $X \subset Y$
$X \overset{c}{\hookrightarrow} Y$	compact embedding from space X into space Y, in literature also denoted as $X \subset\subset Y$
$X \overset{d}{\hookrightarrow} Y$	dense embedding from space X into space Y
$y_1 \rightsquigarrow y_2$	Variable y_2 is coupled to y_1, i.e. y_1 influences y_2

Symbol	**Declaration**
General operators	
$f[t] = f(t, q(t), u(t))$	Example for Nemitsky type-operator applied to a function f with several variables depending on t
$f'_t = \partial_t f$	partial derivative of function f w.r.t. t
$f'_x = [f'_{x_1}, \dots, f'_{x_d}]$	partial derivative of function f w.r.t. vector x, usually written as a row vector
$f'_{i;t}$	partial derivative of component i of vector-valued function f w.r.t. t
f_0	initial value for scalar function
$f_{i;0}$	initial value for component i of vector-valued function f
E'_u	partial derivative of operator E w.r.t. function u
E'	total derivative of operator E w.r.t. all arguments
$\dot{f}(= \frac{d}{dt}f)$	total derivative of function f w.r.t. t
∇	$:= [\partial_{x_1}, \partial_{x_2}, \dots, \partial_{x_d}]$, the Nabla operator in dimension d
$\Delta_x (= \Delta)$	$:= \sum_{i=1}^d \partial_{x_i}^2$, the Laplace operator
$A^\top$	transposition of a matrix A
A^*	dual operator of A
$\mathcal{E}$	averaging-evaluation operator, see Def. 3.1
tr	$:= \sum_{i=1}^d A^{ii}$, trace of a $d \times d$ matrix A^{ij}, $1 \leq i, j \leq d$
$\mathbb{S}$	operator representing a real system
$\mathcal{S}$	operator representing the model system for $\mathbb{S}$

Symbol	**Declaration**
Natural constants	
R_G	$= 8.3145\,\mathrm{J\,mol^{-1}\,K^{-1}}$ universal gas constant
N_A	$= 6.0221 \cdot 10^{23}\,\mathrm{mol^{-1}}$ Avogadro's constant

Symbol	**Declaration**
Variables and some dependent quantities	
q	ODE states
y	(PDE) states
u	controls
z	$:= [y, u]$, optimization variable
$\lambda (= p)$	adjoints / multipliers
$z_0 := z(t_0)$	value of optimization variable at given initial time $t_0 \in \mathbb{R}$
$z_f := z(t_f)$	value of optimization variable at possibly free terminal time $t_f \in \mathbb{R}$
$\hat{z}$	local optimal (minimizing) optimization variable
$\hat{y}$	local optimal (minimizing) states
$\hat{u}$	local optimal (minimizing) controls
$\mathcal{J}$	objective as a functional of y and u
$\tilde{\mathcal{J}}$	$:= \mathcal{J}(y(u), u)$, reduced objective as a functional of u
G	Inequality constraints
H	Equality constraints
$\mathcal{G}$	Combined inequality and equality constraints
ϕ	integrand in objective
$\mathcal{E}_{kin}$	kinetic energy
$\mathcal{E}_{pot}$	potential energy
L	Lagrange function (Lagrangian)
$\mathcal{H}$	Hamilton function (Hamiltonian)
E	operator representing the system of differential equations (of any type), with values in W
S	control-to-state (or solution) operator, resp., $S : U \to Y, u \mapsto y(u)$ (defined implicitly by $E(y, u) = 0$)
P_Σ	projection operator on set Σ ($\tilde{P}_\Sigma$ pointwise Euclidean projection)
$\mathcal{K}$	convex cone
$\mathcal{K}^+$	positive polar (or dual) cone
$\mathcal{K}^-$	negative polar (or dual) cone
M	mass matrix, unless stated otherwise
σ_S	Cauchy stress tensor
p	momentum or pressure

Erklärung

Hiermit erkläre ich bzw. versichere ich an Eides statt, dass ich die schriftliche Habilitationsleistung selbst verfasst habe, mich keiner fremden Hilfe bedient habe und die Herkunft der verwendeten und zitierten Materialien ordnungsgemäß kenntlich gemacht habe.

Neubiberg, den 4. Oktober 2018