MATTHIAS GERDTS

OPTIMIERUNG

Address of the Author:

Matthias Gerdts

Management Mathematics Group
School of Mathematics
University of Birmingham
Edgbaston, Birmingham, B29 2TT

E-Mail: gerdtsm@maths.bham.ac.uk

WWW: web.mat.bham.ac.uk/M.Gerdts

Preliminary Version: 3. Juni 2008

Vorlesungsplan

Nr.	Datum	Stunden	Seite
1	24.10.2006	2V	1–9
2	27.10.2006	2V	10–28
3	31.10.2006	2V	29–37
4	03.11.2006	2V	37–47
5	07.11.2006	2V	48–54
6	10.11.2006	2V	54–61
7	14.11.2006	2V	61–68
8	17.11.2006	2V	69–74
9	21.11.2006	2V	75–79
10	24.11.2006	2V	80–90
11	28.11.2006	2V	91–97
12	01.12.2006	2V	98–105
13	05.12.2006	2V	105–111
14	08.12.2006	2V	111–119
15	12.12.2006	2V	119–123
16	15.12.2006	2V	123–130 (Ausgleichsprobleme ausgelassen)
17	19.12.2006	2V	139–146
18	22.12.2007	2V	146–151
19	09.01.2007	2V	151–160
20	12.01.2007	2V	161–165
21	16.01.2007	2V	166–171, 175
22	19.01.2007	2V	171–176
23	23.01.2007	2V	177–182
24	26.01.2007	2V	183–185,198–201
25	30.01.2007	2V	202–206
26	02.02.2007	2V	207–216
27	06.02.2007	2V	217–224
28	09.02.2007	2V	224–

Inhaltsverzeichnis

Bezeichnungen

Mit $\|\cdot\| = \|\cdot\|_2$ bezeichnen wir die euklidische Norm im $\mathbb{R}^n$ und mit $\langle x, y\rangle = y^\top x$ das Skalarprodukt. Allgemeiner bezeichnet $\langle x, y\rangle_A = y^\top Ax$ eine Bilinearform mit der Matrix $A \in \mathbb{R}^{n\times n}$.
Sei $f : \mathbb{R}^n \to \mathbb{R}$ und $x = (x_1, \ldots, x_n)^\top \in \mathbb{R}^n$. Der Gradient von f an der Stelle x ist definiert als

$$\nabla f(x) = \left(\frac{\partial f(x)}{\partial x_1}, \ldots, \frac{\partial f(x)}{\partial x_n}\right)^\top.$$

Die Hessematrix von f an der Stelle x ist gegeben durch

$$\nabla^2 f(x) = \begin{pmatrix} \frac{\partial^2 f(x)}{\partial x_1 \partial x_1} & \cdots & \frac{\partial^2 f(x)}{\partial x_1 \partial x_n} \\ \vdots & \ddots & \vdots \\ \frac{\partial^2 f(x)}{\partial x_n \partial x_1} & \cdots & \frac{\partial^2 f(x)}{\partial x_n \partial x_n} \end{pmatrix}.$$

Sei $f : \mathbb{R}^n \to \mathbb{R}^m$ und $x = (x_1, \ldots, x_n)^\top \in \mathbb{R}^n$. Die Jacobimatrix von f an der Stelle x lautet

$$f'(x) = \begin{pmatrix} \frac{\partial f_1(x)}{\partial x_1} & \cdots & \frac{\partial f_1(x)}{\partial x_n} \\ \vdots & \ddots & \vdots \\ \frac{\partial f_m(x)}{\partial x_1} & \cdots & \frac{\partial f_m(x)}{\partial x_n} \end{pmatrix}$$

Speziell gilt $\nabla f(x)^\top = f'(x)$, falls $m = 1$ gilt.
Die Richtungsableitung von $f : \mathbb{R}^n \to \mathbb{R}$ an der Stelle $x \in \mathbb{R}^n$ in Richtung $d \in \mathbb{R}^n$ ist definiert als

$$f'(x; d) = \lim_{\alpha \downarrow 0} \frac{f(x + \alpha d) - f(x)}{\alpha}.$$

Ist f differenzierbar in x, so gilt $f'(x; d) = \nabla f(x)^\top d$.
Die multivariate Taylorentwicklung für eine $p + 1$ fach stetig differenzierbare Funktion $f : \mathbb{R}^n \to \mathbb{R}^m$ in $x \in \mathbb{R}^n$ lautet

$$f(x + h) = \sum_{j=0}^{p} \frac{1}{j!} f^{(j)}(x) \cdot \underbrace{(h, \ldots, h)}_{j-fach} + \mathcal{O}(\|h\|^{p+1})$$

für hinreichend kleines $h \in \mathbb{R}^n$. Hierin ist die j-te Ableitung von f eine j-lineare Abbildung mit

$$f^{(j)}(x) \cdot (h_1, \ldots, h_j) = \sum_{i_1, \ldots, i_j = 1}^{n} \frac{\partial^j f(x)}{\partial x_{i_1} \cdots \partial x_{i_j}} h_{1,i_1} \cdots h_{j,i_j}.$$

Weiterhin gilt die Restglieddarstellung

$$f(x+h) = \sum_{j=0}^{p} \frac{1}{j!} f^{(j)}(x) \cdot \underbrace{(h,\dots,h)}_{j-fach} + \int_0^1 \frac{(1-t)^p}{p!} f^{(p+1)}(x+th) \cdot \underbrace{(h,\dots,h)}_{(p+1)-fach} dt.$$

Speziell erhält man für $m = 1$, $p = 1$ und $p = 2$ die Taylorentwicklungen

$$f(y) = f(x) + \nabla f(x + \xi(y-x))^\top (y-x)$$

mit einer Zwischenstelle $\xi \in [0,1]$ und

$$f(y) = f(x) + \nabla f(x)^\top (y-x) + \frac{1}{2}(y-x)^\top \nabla^2 f(x+\xi(y-x))(y-x)$$

mit einer Zwischenstelle $\xi \in [0,1]$.
Für vektorwertiges $f : \mathbb{R}^n \to \mathbb{R}^m$ gilt der Mittelwertsatz in Integralform

$$f(y) - f(x) = \int_0^1 f'(x + t(y-x))(y-x) dt.$$

Satz 0.0.1 **(Satz über implizite Funktionen)**
Es sei $D \subseteq \mathbb{R}^n \times \mathbb{R}^m$ und $f = (f_1, \dots, f_m) : D \to \mathbb{R}^m$ in einer Umgebung von $(x_0, y_0)^\top \in D$ stetig differenzierbar. Es gelte $f(x_0, y_0) = 0$ und die Jacobimatrix $\partial f / \partial y\big|_{(x_0,y_0)}$ sei invertierbar. Dann gibt es in einer Umgebung U von x_0 eindeutig definierte Funktionen $g_1, \dots, g_m$ mit

(i) $(g_1(x_0), \dots, g_m(x_0))^\top = y_0$;

(ii) *für alle* $x \in U$ *gilt* $f(x, g(x)) = 0$;

(iii) $g_1, \dots, g_m$ *sind stetig differenzierbar in* x_0 *mit*

$$\left.\frac{\partial g}{\partial x}\right|_{(x_0)} = -\left(\left.\frac{\partial f}{\partial y}\right|_{(x_0,y_0)}\right)^{-1} \cdot \left.\frac{\partial f}{\partial x}\right|_{(x_0,y_0)}.$$

Kapitel 1

Problemstellungen und Klassifikation

Optimierungsaufgaben spielen in allen Anwendungsbereichen von Mathematik eine wichtige Rolle. Hauptanwendungsgebiete sind Wirtschaftswissenschaften (Operations Research), Technik und Naturwissenschaften.
In Nocedal, Wright findet man die folgende Formulierung:

> **People optimize:** Airline companies schedule crews and aircraft to minimize cost. Investors seek to create portfolios that avoid risks while achieving a high rate of return. Manufacturers aim for maximizing efficiency in the design and operation of their production processes.
> **Nature optimizes:** Physical systems tend to a state of minimum energy. The molecules in an isolated chemical system react with each other until the total potential energy of their electrons is minimized, Rays of light follow paths that minimize their travel time.

Die Optimierung steht in enger Beziehung zur Modellierung, d.h. Optimierungstechniken werden auf mathematische Modelle angewendet, für die dann gewisse unbekannte Modellparameter oder -funktionen so zu bestimmen sind, dass eine *Zielfunktion* unter vorgegebenen *Nebenbedingungen* minimiert (oder maximiert) wird.
Wir beschränken uns in der Klassifikation von Optimierungsproblemen o.B.d.A. auf Minimierungsprobleme. Ein Maximierungsproblem wird durch Multiplikation der zu maximierenden Funktion mit -1 in ein äquivalentes Minimierungsproblem transformiert.

Allgemeines Optimierungsproblem (OP):

$$\min f(x) \quad \text{unter} \quad x \in X.$$

Darin sei $X \subseteq \mathbb{R}^n$ eine beliebige nichtleere Menge und $f : X \to \mathbb{R}$ eine beliebige Funktion.

Bezeichnungen:

Definition 1.0.2

- *Die zu minimierende Funktion f heißt* **Zielfunktion**.
- *Ein Vektor x heißt* **zulässig** *für (OP), falls $x \in X$ gilt. X heißt* **zulässige Menge** *von (OP).*

- $\hat{x} \in X$ *heißt* **globales Minimum** *von* (OP), *falls*

$$f(\hat{x}) \leq f(x) \quad \forall x \in X. \tag{1.1}$$

  $\hat{x} \in X$ *heißt* **striktes globales Minimum** *von* (OP), *falls in (1.1)* „$<$“ *für alle* $x \in X$, $x \neq \hat{x}$ *gilt.*

- $\hat{x} \in X$ *heißt* **lokales Minimum** *von* (OP), *falls es eine Umgebung*

$$U_\varepsilon(\hat{x}) := \{x \in \mathbb{R}^n \mid \|x - \hat{x}\| < \varepsilon\}$$

  gibt mit

$$f(\hat{x}) \leq f(x) \quad \forall x \in X \cap U_\varepsilon(\hat{x}). \tag{1.2}$$

  $\hat{x} \in X$ *heißt* **striktes lokales Minimum** *von* (OP), *falls in (1.2)* „$<$“ *für alle* $x \in X \cap U_\varepsilon(\hat{x})$, $x \neq \hat{x}$ *gilt.*

Globale bzw. lokale Maxima werden analog definiert. Die Begriffe werden in Abbildung 1.1 erläutert.

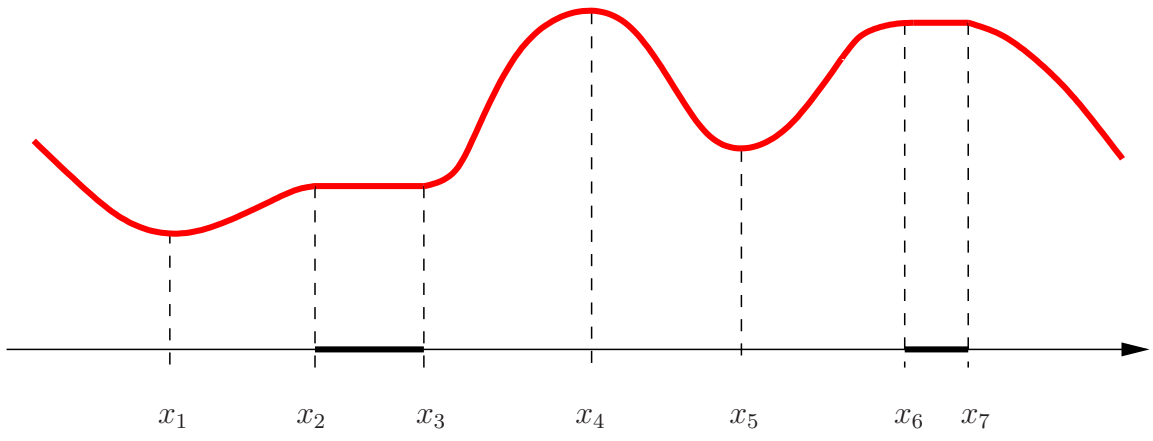


Abbildung 1.1: Lokale und globale Minima und Maxima einer Funktion: x_1: striktes globales Minimum, x_2: lokales Maximum, x_3: lokales Minimum; (x_2, x_3): gleichzeitig lokales Minimum und Maximum, x_4: striktes globales Maximum, x_5: striktes lokales Minimum, x_6, x_7: lokale Maxima, (x_6, x_7): gleichzeitig lokales Minimum und Maximum.

Spezialfälle:

Ein unrestringiertes Problem liegt vor, falls $X = \mathbb{R}^n$ gilt:

Unrestringiertes Optimierungsproblem (UOP):

$$\min f(x) \quad \text{unter} \quad x \in \mathbb{R}^n.$$

Darin sei $f : \mathbb{R}^n \to \mathbb{R}$ eine beliebige Funktion.

Häufig treten konvexe Problem auf:

Konvexes Optimierungsproblem (KOP):

$$\min f(x) \quad \text{unter} \quad x \in X.$$

Darin sei $X \subseteq \mathbb{R}^n$ eine nichtleere konvexe Menge und $f : X \to \mathbb{R}$ eine konvexe Funktion.

Lineare Optimierungsprobleme stellen einen wichtigen Spezialfall konvexer Problem dar:

Lineares Optimierungsproblem (LOP) **in Normalform**:

$$\min c^\top x \quad \text{unter} \quad Ax = b,\ x \geq 0.$$

Darin seien $x, c \in \mathbb{R}^n$ und $b \in \mathbb{R}^m$ Vektoren und $A \in \mathbb{R}^{m \times n}$ eine Matrix.

Häufig wird die Menge X in (OP) durch endlich viele Gleichungen und Ungleichungen beschrieben:

Standard Optimierungsproblem (SOP):

$$\min f(x) \quad \text{unter} \quad x \in X$$

mit

$$X = \{x \in \mathbb{R}^n \mid g_i(x) \leq 0,\ i = 1, \ldots, m,\ h_j(x) = 0,\ j = 1, \ldots, p\}.$$

Darin seien $f : D \to \mathbb{R}$, $g_i : D \to \mathbb{R}$, $i = 1, \ldots, m$ und $h_j : D \to \mathbb{R}$, $j = 1, \ldots, p$ mit $X \subseteq D \subseteq \mathbb{R}^n$ beliebige Funktionen.

Es treten auch Probleme mit diskreten Optimierungsvariablen auf.

Ganzzahliges Optimierungsproblem (GOP):

$$\min f(x) \quad \text{unter} \quad x \in X$$

mit

$$X = \{x \in \mathbb{Z}^n \mid g_i(x) \leq 0,\ i = 1, \ldots, m,\ h_j(x) = 0,\ j = 1, \ldots, p\}.$$

Darin seien $f : \mathbb{Z}^n \to \mathbb{R}$, $g_i : \mathbb{Z}^n \to \mathbb{R}$, $i = 1, \ldots, m$ und $h_j : \mathbb{Z}^n \to \mathbb{R}$, $j = 1, \ldots, p$ beliebige Funktionen.

Bemerkung 1.0.3

- *Mit Ausnahme der Formulierung (LOP) sind alle Problemstellungen formal nichtlinear.*
- *Sind die Funktionen f und g_i, $i = 1, \ldots, m$ konvex und h_j, $j = 1, \ldots, p$ affin linear in (SOP), so ist (SOP) ein konvexes Optimierungsproblem.*
- *Sind die Funktionen f, g_i, $i = 1, \ldots, m$ und h_j, $j = 1, \ldots, p$ in (SOP) affin linear, so liegt ein lineares Optimierungsproblem vor (nicht notwendig in Normalform).*
- *Die Darstellung der verschiedenen Optimierungsprobleme ist nicht vollständig. So gibt es beispielsweise auch unendlichdimensionale Optimierungsprobleme, Vektoroptimierungsprobleme sowie Mischformen der oben dargestellten Problemformen.*

In dieser Vorlesung beschäftigen wir uns mit **unrestringierten Optimierungsproblemen** (UOP) und **Optimierungsproblemen in Standardform** (SOP), wobei die auftretenden Funktionen in der Regel mindestens einmal stetig differenzierbar sind.
Typische Fragestellungen:

i) **Existieren** überhaupt **zulässige Lösungen**?

ii) **Existieren Optimallösungen**?

iii) Ist die Optimallösung **eindeutig** bestimmt?

iv) **Wie hängen die Optimallösungen von den Problemdaten ab**?

v) Welche Eigenschaften besitzen Optimallösungen, mit anderen Worten welche Bedingungen werden von einer Optimallösung **notwendig** erfüllt?

vi) Welche Bedingungen sind **hinreichend** dafür, dass eine zulässige Lösung optimal ist?

vii) Welche Bedingungen sind gleichzeitig notwendig und hinreichend für Optimalität, **charakterisieren** also die Optimallösungen?

viii) Wie gewinnt man aus zulässigen Lösungen Informationen über die Optimallösungen, insbesondere **Einschließungen** für den Optimalwert und **Fehlerabschätzungen** für die Optimallösung?

ix) Welche **konzeptionellen Algorithmen** zur Berechnung einer Optimallösung stehen zur Verfügung?

x) Welche **numerischen Eigenschaften** besitzen diese Algorithmen (Konvergenz, Konvergenzgeschwindigkeit, Stabilität)?

Die Fragestellungen i) – vii) sind überwiegend theoretischer Natur. Aber ohne ihre Beantwortung ist die Behandlung der numerischen Fragestellungen viii) – x) nicht möglich.

1.1 Existenz von Minima

Die Existenz einer Lösung des Minimierungsproblems

$$\min f(x) \quad \text{unter} \quad x \in X$$

mit kompakter Menge $X \subseteq \mathbb{R}^n$ und unterhalbstetiger Funktion f ist durch den folgenden aus der Analysis bekannten Satz gesichert. Eine Funktion $f : D \subseteq \mathbb{R}^n \to \mathbb{R}$ heißt **unterhalbstetig in** $x \in D$, wenn für jede Folge $\{x_i\}$ mit $x_i \to x$ gilt

$$f(x) \le \liminf_{i \to \infty} f(x_i).$$

Beachte, daß stetige Funktionen auch unterhalbstetig sind.

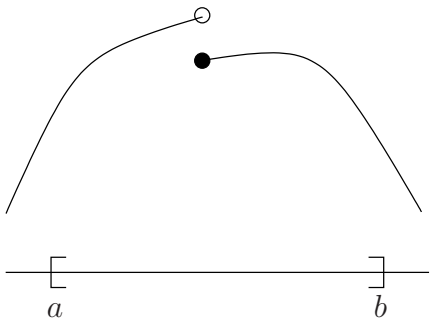


Abbildung 1.2: Eine unterhalbstetige Funktion auf der kompakten Menge $[a, b]$, die ihr Supremum nicht annimmt.

Satz 1.1.1 **(Weierstrass)**

Sei $X \subseteq D \subseteq \mathbb{R}^n$ kompakt und $f : D \to \mathbb{R}$ unterhalbstetig. Dann nimmt f ihr Infimum auf X an.

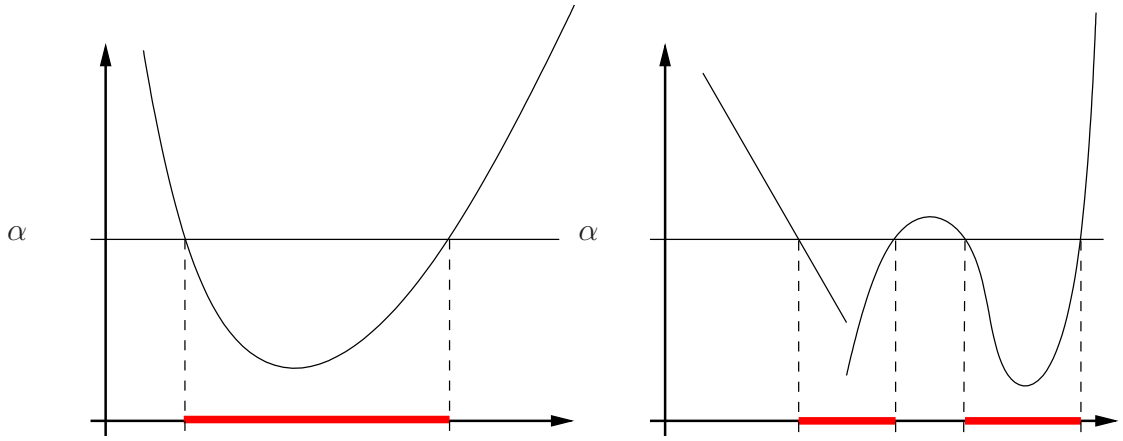
Beweis: Angenommen f ist nicht nach unten beschränkt auf X. Dann gibt es eine Folge $\{x_i\}$ in X mit $f(x_i) \leq -i$. Da X kompakt ist, existiert eine konvergente Teilfolge $\lim_{k\to\infty} x_{i_k} = \hat{x}$ mit $f(x_{i_k}) \leq -i_k$ für alle $k \in \mathbb{N}$. Da f unterhalbstetig ist, folgt $f(\hat{x}) \leq \liminf_{k\to\infty} f(x_{i_k})$. Also ist $f(x_{i_k})$ nach unten durch $f(\hat{x}) \in \mathbb{R}$ beschränkt. Dies widerspricht $f(x_{i_k}) \leq -i_k \to -\infty$. Damit ist f auf X nach unten beschränkt.
Dies wiederum impliziert, daß $\hat{f} = \inf_{x\in X} f(x)$ eine reelle Zahl ist und für jedes $i \in \mathbb{N}$ existiert ein $x_i \in X$ mit $f(x_i) \leq \hat{f} + \frac{1}{i}$. Da X kompakt ist, existiert eine konvergente Teilfolge $x_{i_k} \to \hat{x}$ mit $f(x_{i_k}) \leq \hat{f} + \frac{1}{i_k}$ für alle $k \in \mathbb{N}$. Da f unterhalbstetig ist, folgt $\hat{f} \leq f(\hat{x}) \leq \liminf_{k\to\infty} f(x_{i_k}) \leq \hat{f}$. Also nimmt f ihr Minimum auf X an. □

Dieser Satz läßt sich wie folgt verallgemeinern. Dazu benötigen wir den Begriff der Levelmenge oder Niveaumenge.

Definition 1.1.2 **(Levelmenge, Niveaumenge)**

Die **Levelmenge** *oder* **Niveaumenge** *von $f : D \to \mathbb{R}$, $D \subseteq \mathbb{R}^n$ zum Niveau $\alpha \in \mathbb{R}$ ist definiert als*

$$lev(f,\alpha) = \{x \in D \mid (x,\alpha) \in epi(f)\} = \{x \in D \mid f(x) \leq \alpha\}.$$



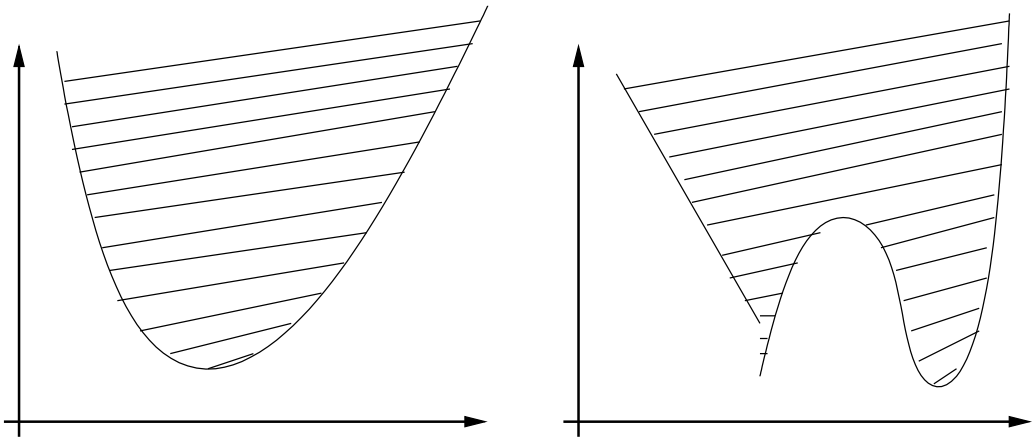
$lev(f,\alpha)$ $\qquad$ $lev(f,\alpha)$

Der Epigraph ist wie üblich definiert:

Definition 1.1.3 **(Epigraph)**

Der **Epigraph** *einer Funktion $f : \mathbb{R}^n \to \mathbb{R}$ ist die Menge*

$$epi(f) = \{(x,r) \in \mathbb{R}^n \times \mathbb{R} \mid f(x) \leq r\}.$$



Dann gilt:

Satz 1.1.4

Sei $X \subseteq D \subseteq \mathbb{R}^n$ und $f : D \to \mathbb{R}$ sei unterhalbstetig auf X. Für ein $w \in X$ sei die Menge

$$lev(f, f(w)) \cap X = \{x \in X \mid f(x) \leq f(w)\}$$

nichtleer und kompakt. Dann gibt es mindestens ein globales Minimum von f auf X.

Beweis: Nach dem Satz von Weierstraß gibt es ein $\hat{x} \in lev(f, f(w)) \cap X$ mit $f(\hat{x}) \leq f(x)$ für alle $x \in lev(f, f(w)) \cap X$. Für $x \in X \backslash (lev(f, f(w)) \cap X) = X \backslash lev(f, f(w))$ gilt $f(x) > f(w) \geq f(\hat{x})$. Damit ist $\hat{x}$ Minimum von f auf X. □

Kapitel 2

Beispiele

Es werden exemplarisch einige typische Optimierungsaufgaben vorgestellt und diskutiert. Weitere interessante Anwendungen sind in Spellucci [Spe93], Bazaraa [BSS93] und Alt [Alt02] zu finden. Darüber hinaus gibt es unzählige Optimierungsprobleme in Industrie, Wirtschaft und Wissenschaft.

2.1 Elementare Funktionen und Höhenlinien

Beispiel 2.1.1 (Funktion von Himmelblau)
Wir wollen die Funktion von Himmelblau

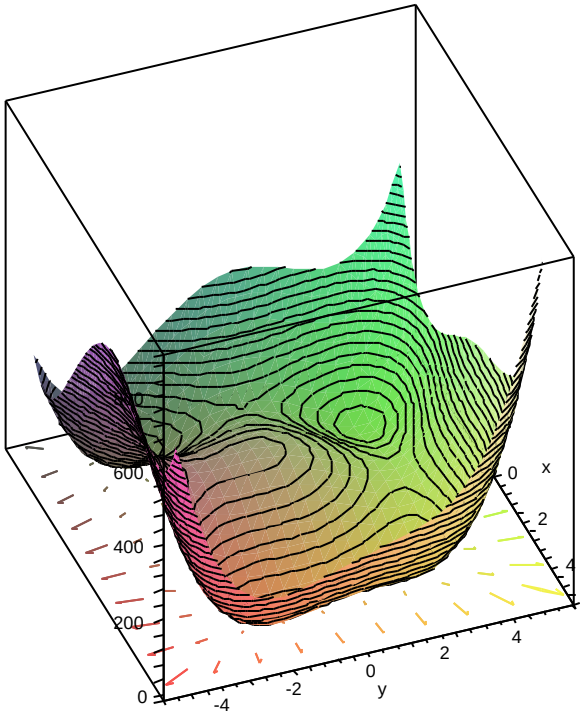
$$f_H(x,y) := (x^2 + y - 11)^2 + (x + y^2 - 7)^2$$

über alle $(x,y)^\top \in \mathbb{R}^2$ *minimieren.*

Um einen besseren Eindruck von der Funktion zu bekommen, stellen wir die Funktion grafisch dar. Dies kann z.B. mit dem Programm Maple *und dem Befehl*

```
plot3d((x^2+y-11)^2+(x+y^2-7)^2,x=-5..5,y=-5..5,axes=boxed,
   style=PATCHCONTOUR,contours = 40,shading=XYZ);
```

geschehen:



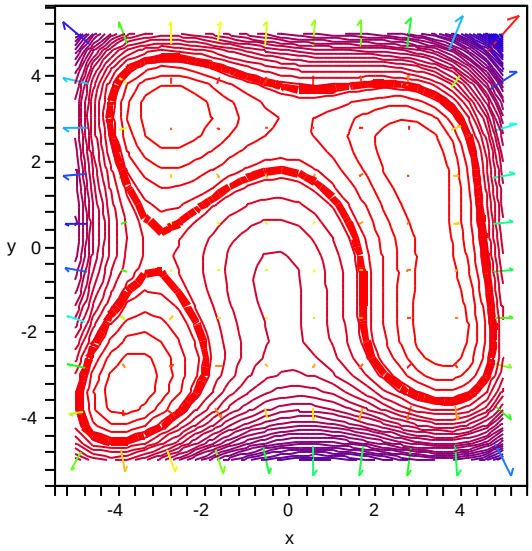
Einen noch besseren Eindruck von den Funktionswerten der Funktion erhalten wir mit dem Befehl

```
contourplot((x^2+y-11)^2+(x+y^2-7)^2,x=-5..5,y=-5..5,contours=80,
   coloring=[red,blue],scaling=constrained,axes=boxed);
```

Dieser Befehl zeichnet die sogenannten **Höhenlinien** *oder* **Niveaulinien** *einer Funktion. Eine* **Höhenlinie zum Niveau** $c \in \mathbb{R}$ *für eine Funktion* $f : \mathbb{R}^n \to \mathbb{R}$ *ist formal definiert als die Menge aller Punkte* $x = (x_1, \ldots, x_n)^\top$*, die* $f(x) = c$ *erfüllen und wird mit* $N_f(c)$ *bezeichnet, also*

$$N_f(c) := \{x \in \mathbb{R}^n \mid f(x) = c\}, \qquad c \in \mathbb{R}.$$

Somit besitzt die Funktion f *entlang einer Höhenlinie immer denselben Funktionswert, ist also entlang einer Höhenlinie konstant. Die Abbildung zeigt die Höhenlinien der Funktion von Himmelblau. Die fett gezeichnete Höhenlinie entspricht dem Niveau* $c = 100$*.*



Die Pfeile in den beiden Grafiken stellen die **Gradienten von** f *in den jeweiligen Punkten dar. Der Gradient einer Funktion* $f : \mathbb{R}^n \to \mathbb{R}$ *an der Stelle* $x = (x_1, \ldots, x_n)^\top \in \mathbb{R}^n$ *ist formal definiert als der Spaltenvektor*

$$\nabla f(x_1, \ldots, x_n) := \begin{pmatrix} \frac{\partial f}{\partial x_1}(x_1, \ldots, x_n) \\ \vdots \\ \frac{\partial f}{\partial x_n}(x_1, \ldots, x_n) \end{pmatrix}.$$

Bekanntlich zeigt der Gradient einer Funktion f *in die* **Richtung des steilsten Anstiegs** *von* f*. Außerdem steht der Gradient* **senkrecht** *auf den Höhenlinien von* f*.*
Für die Funktion von Himmelblau ergibt sich speziell

$$\nabla f_H(x, y) = \begin{pmatrix} \frac{\partial f_H}{\partial x}(x, y) \\ \frac{\partial f_H}{\partial y}(x, y) \end{pmatrix} = \begin{pmatrix} 4x(x^2 + y - 11) + 2(x + y^2 - 7) \\ 2(x^2 + y - 11) + 4y(x + y^2 - 7) \end{pmatrix}.$$

Die Gradienten von f_H *können mit dem folgenden Befehl dargestellt werden:*

```
gradplot( (x^2+y-11)^2+(x+y^2-7)^2,x=-5..5,y=-5..5,grid=[10,10],
   color=(x^2+y-11)^2+(x+y^2-7)^2);
```

Anhand der grafischen Darstellungen läßt sich folgendes ablesen: Die Funktion von Himmelblau besitzt

- *4 lokale Minimalstellen (zugleich global) mit Funktionswert 0*
- *4 Sattelpunkte*
- *ein lokales Maximum in* $(-0.270845, -0.923039)^\top$

Wir werden später sehen, daß der Gradient von f in jedem dieser Punkte gleich dem Nullvektor ist.

Beispiel 2.1.2 **(Funktion von Rosenbrock (Banana-Function))**
Analysieren Sie wie im vorigen Beispiel die Funktion von Rosenbrock (Banana-Function):

$$f_R(x,y) := 100(y-x^2)^2 + (1-x)^2$$

Hinweis:

- *globales Minimum in* $(1,1)^\top$ *mit* $f(1,1) = 0$.

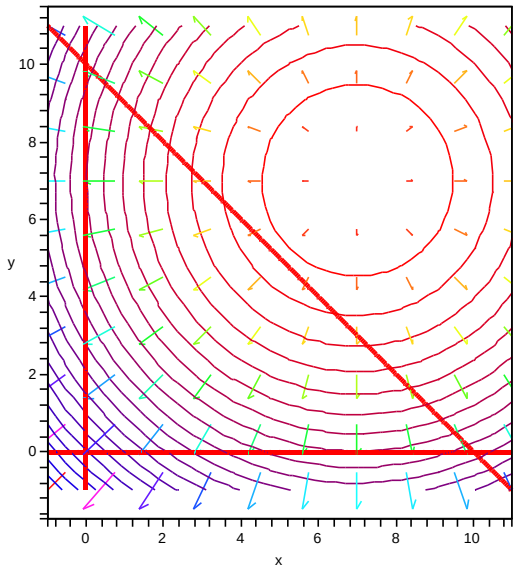
Beispiel 2.1.3 **(Optimierungsproblem mit Nebenbedingungen)**
Betrachte das folgende Standard-Optimierungsproblem:

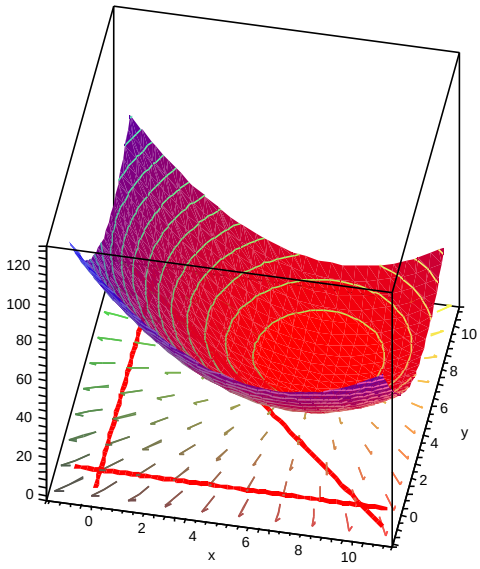
$$\begin{array}{ll} \textit{Minimiere} & f(x,y) = (x-7)^2 + (y-7)^2 \\ \textit{unter} & x,y \in \mathbb{R}, \\ & x,y \geq 0, \\ & x+y \leq 10. \end{array}$$

Zunächst betrachten wir den **zulässigen Bereich** X *des Optimierungsproblems, also alle Punkte* $(x,y)^\top \in \mathbb{R}^2$*, die die Nebenbedingungen*

$$x \geq 0, \quad y \geq 0, \quad x+y \leq 10$$

erfüllen. Der Bereich innerhalb des Dreiecks in der nachfolgenden Abbildung ist der zulässige Bereich (zusätzlich sind die Höhenlinien und die Gradienten der Zielfunktion f eingezeichnet):





Es ist leicht zu sehen, daß die Zielfunktion f im Punkt $(7,7)^\top$ ein globales Minimum mit Funktionswert 0 besitzt. Wegen $7+7>10$ ist dieser Punkt jedoch nicht zulässig (die dritte Nebenbedingung $x+y \le 10$ ist verletzt) und somit keine Lösung des Optimierungsproblems! Die tatsächliche Lösung des Problems liegt im Punkt $(x,y)^\top = (5,5)^\top$ auf dem Rand des zulässigen Bereiches.

Beispiel 2.1.4 (Portfoliooptimierung)

Wir betrachten das Beispiel einer Portfoliooptimierungsaufgabe nach Markowitz. Gegeben seien $j = 1,\dots,n$ mögliche Anlagen (z.B. Aktien, Fonds, Optionen, Wertpapiere). Jede Anlage wirft im nächsten Zeitintervall einen Gewinn (oder Verlust) R_j ab. Leider ist R_j in der Regel nicht bekannt, sondern zufällig verteilt. Um einerseits den Gewinn zu maximieren und andererseits das Risiko eines Verlusts zu minimieren, wird die Anlagesumme zu Anteilen x_j auf die Anlagen $j = 1,\dots,n$ verteilt und die anteiligen Anlagen werden in einem Portfolio zusammengefaßt.

Die Aufgabe eines Portfoliomanagers besteht in der optimalen Zusammensetzung eines solchen Portfolios, d.h. die Anteile x_j der jeweiligen Anlagen, die in das Portfolio übernommen werden sollen, müssen in einem gewissen Sinne optimal bestimmt werden. Ein mögliches Ziel ist es, den erwarteten Gewinn

$$E(R) = \sum_{j=1}^n x_j E(R_j), \qquad R = \sum_{j=1}^n x_j R_j$$

zu maximieren (E bezeichnet den Erwartungswert). Jedoch ist ein hoher Gewinn in der Regel nur mit riskanten Anlagen möglich, so daß auch das Risiko eines Verlusts steigt. Als Maß für das Risiko kann die Varianz des Gewinns dienen:

$$Var(R) = E(R - E(R))^2 = E\left(\sum_{j=1}^n x_j (R_j - E(R_j))\right)^2$$

Ein Kompromiss zwischen hohem Gewinn und geringem Risiko kann durch Lösen des folgenden Optimierungsproblems erreicht werden:

$$\begin{aligned} \min \quad & -\sum_{j=1}^n x_j E(R_j) + \alpha E\left(\sum_{j=1}^n x_j (R_j - E(R_j))\right)^2 \\ \text{unter} \quad & \sum_{j=1}^n x_j = 1,\ x_j \ge 0,\ j = 1,\dots,n. \end{aligned}$$

Hierin bezeichnet $\alpha > 0$ einen Gewichtungsparameter, mit dem die Risikobereitschaft gesteuert werden kann. Mit $\alpha = 0$ wird der Varianzterm in der Zielfunktion eliminiert, so daß nur noch der Gewinn maximiert wird. Dies entspricht einer hohen Risikobereitschaft. Mit wachsendem α wird der Varianzterm stärker gewichtet und die Risikobereitschaft sinkt.

2.2 Exakte L1-Penalty-Funktion

Ein Ansatz zur Lösung von Optimierungsproblemen in Standardform (SOP) besteht darin, das Problem (SOP) auf ein unrestringiertes Optimierungsproblem (UOP) zurückzuführen. Dazu wird eine geeignete Hilfsfunktion definiert:

$$l_1(x;\alpha) = f(x) + \alpha \sum_{i=1}^{m} \max\{0, g_i(x)\} + \alpha \sum_{j=1}^{p} |h_j(x)|. \tag{2.1}$$

Die Funktion $l_1(x;\alpha)$ heißt **exakte l_1-Penaltyfunktion**. Sie bestraft das Verlassen des zulässigen Bereichs. Die Bezeichnung exakt folgt aus der Tatsache, daß für alle hinreichend großen $\alpha > 0$ und unter geeigneten Bedingungen an die Restriktionen g_i und h_j die lokalen Minimalstellen von (SOP) auch lokale Minimalstellen von l_1 bzgl. x sind.
Anstatt das Problem (SOP) zu lösen, wird daher die Funktion l_1 bzgl. $x \in \mathbb{R}^n$ mit einem geeigneten Parameter $\alpha > 0$ minimiert. Allerdings ist $l_1(x;\alpha)$ i.a. **nicht differenzierbar bzgl.** x, selbst wenn g_i und h_j differenzierbar sind.

Beispiel 2.2.1

Die Abbildung 2.1 zeigt die l_1-Penaltyfunktion für das Problem

$$\begin{aligned} f(x,y) &= (x-2)^2 + (y-3)^2, \\ h(x,y) &= y + \frac{x}{2} - \frac{1}{2}, \\ g_1(x,y) &= y + 2x^2 - 2, \\ g_2(x,y) &= x^2 - y - 1, \end{aligned}$$

für verschiedene Werte von α.

Bemerkung 2.2.2

Alternativ können auch die exakte l_∞-Penaltyfunktion

$$l_\infty(x;\alpha) = f(x) + \alpha \max\{0, g_1(x), \ldots, g_m(x), |h_1(x)|, \ldots, |h_p(x)|\},$$

die exakte l_2-Penaltyfunktion

$$l_2(x;\alpha) = f(x) + \alpha \left(\sum_{i=1}^{m} \left(\max\{0, g_i(x)\}\right)^2 + \sum_{j=1}^{p} h_j(x)^2 \right)^{1/2}$$

oder allgemein die exakte l_q-Penaltyfunktion

$$l_q(x;\alpha) = f(x) + \alpha \left(\sum_{i=1}^{m} \left(\max\{0, g_i(x)\}\right)^q + \sum_{j=1}^{p} h_j(x)^q \right)^{1/q}$$

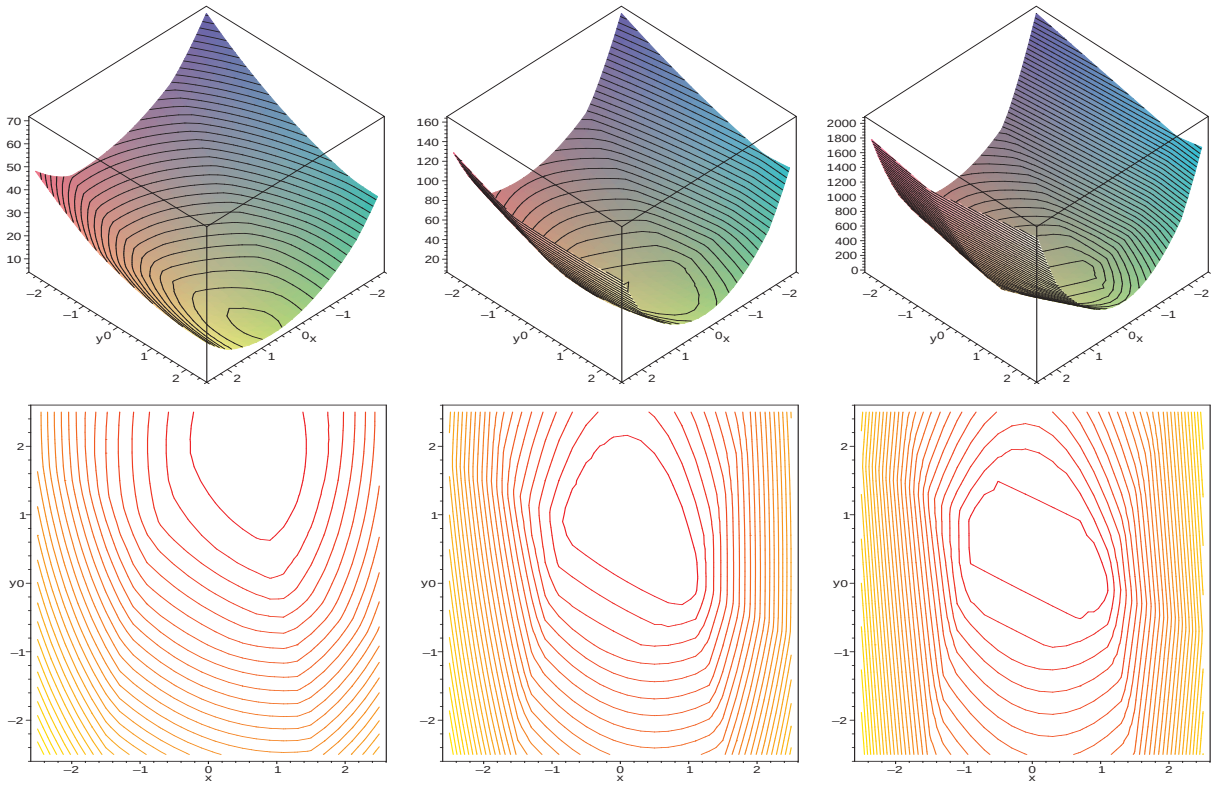


Abbildung 2.1: 3D-Darstellung (oben) und Höhenlinien (unten) der l_1-Penaltyfunktion für $\alpha = 1$ (links), $\alpha = 28/5$ (mitte) und $\alpha = 100$ (rechts).

für $1 \le q < \infty$ *verwendet werden.*

Bei den sogenannten **Penalty-Verfahren** kommen auch nicht exakte Penaltyfunktionen zum Einsatz, etwa

$$l(x;\alpha) = f(x) + \frac{\alpha}{2}\|h(x)\|^2 + \frac{\alpha}{2}\sum_{i=1}^{m}\left(\max\{0, g_i(x)\}\right)^2.$$

Diese Verfahren setzen lediglich die Stetigkeit von f, g_i und h_j voraus und benötigen keine Restriktionsqualifikationen. I.A. ist l dann nicht differenzierbar, so daß Verfahren der nichtdifferenzierbaren Optimierung zur Minimierung von l bzgl. x benötigt werden.

2.3 Wirtschaftswissenschaftliche Aufgabenstellungen

(siehe [BM68])

Es soll eine bestimmte Menge M eines Gutes gekauft werden. Dazu werden Angebote von n Lieferanten eingeholt, wobei kein Lieferant die gewünschte Menge alleine liefern kann, sondern höchstens M_i Einheiten des Gutes. Die Preise $f_i(x_i)$ des i-ten Lieferanten sind i.a. abhängig von der Liefermenge x_i und beinhalten Mengenrabatte. Oft sind die Funktionen

f_i lediglich monoton wachsend und nichtlinear, aber nicht differenzierbar. Die Aufgabe des Käufers ist, die Gesamtkosten zu minimieren:

$$\begin{aligned} \min_x \quad & \sum_{i=1}^n f_i(x_i) \\ \text{unter} \quad & \sum_{i=1}^n x_i = M, \\ & 0 \le x_i \le M_i,\ i = 1, 2, \dots, n. \end{aligned}$$

2.4 Ausgleichs- und Parameteridentifizierungsprobleme

Ein Experiment liefert die Meßpunkte (t_i, y_i), $i = 1, \dots, m$. Der dem Experiment zu Grunde liegende Vorgang werde durch die Funktion $f(t, p)$ modelliert, die einen funktionalen Zusammenhang zwischen den Meßstellen t_i und den Meßwerten y_i herstellt. Allerdings hängt die Funktion auch noch vom unbekannten Parameter $p \in \mathbb{R}^{n_p}$ ab. In der Praxis sind die Meßwerte verrauscht bzw. fehlerbehaftet, so daß es in der Regel keinen Parameter p gibt, der die Meßpunkte exakt reproduziert. Daher wird versucht, die Meßpunkte so gut wie möglich zu approximieren, indem

$$\frac{1}{2} \sum_{i=1}^m \left(y_i - f(t_i, p)\right)^2$$

bezüglich p minimiert wird. Häufig sind zusätzlich noch Nebenbedingungen an den Parameter p gegeben.

Das resultierende Optimierungsproblem ist ein spezielles Least-Squares Problem. Ein allgemeines Least-Squares Problem lautet

Least-Squares-Problem:
Finde $x \in \mathbb{R}^n$, so daß

$$\frac{1}{2}\|\Phi(x)\|_2^2 = \frac{1}{2} \sum_{i=1}^q \Phi_i(x)^2$$

minimal wird unter den Nebenbedingungen

$$\begin{aligned} g_i(x) &\le 0, \quad i = 1, \dots, m, \\ h_j(x) &= 0, \quad j = 1, \dots, p. \end{aligned}$$

Beispiel 2.4.1

Die Meßpunkte

t_i	y_i
−0.7882416043	0.396878358
−0.6056336413	0.418410056
−0.3976460600	0.627676951
−0.2144255029	0.821174784
−0.0107919623	0.962155739
0.1997798535	1.303597193
0.3741472164	1.362401309
0.5955672872	1.470902326
0.7899671852	1.528415842
0.9997213026	1.510113124

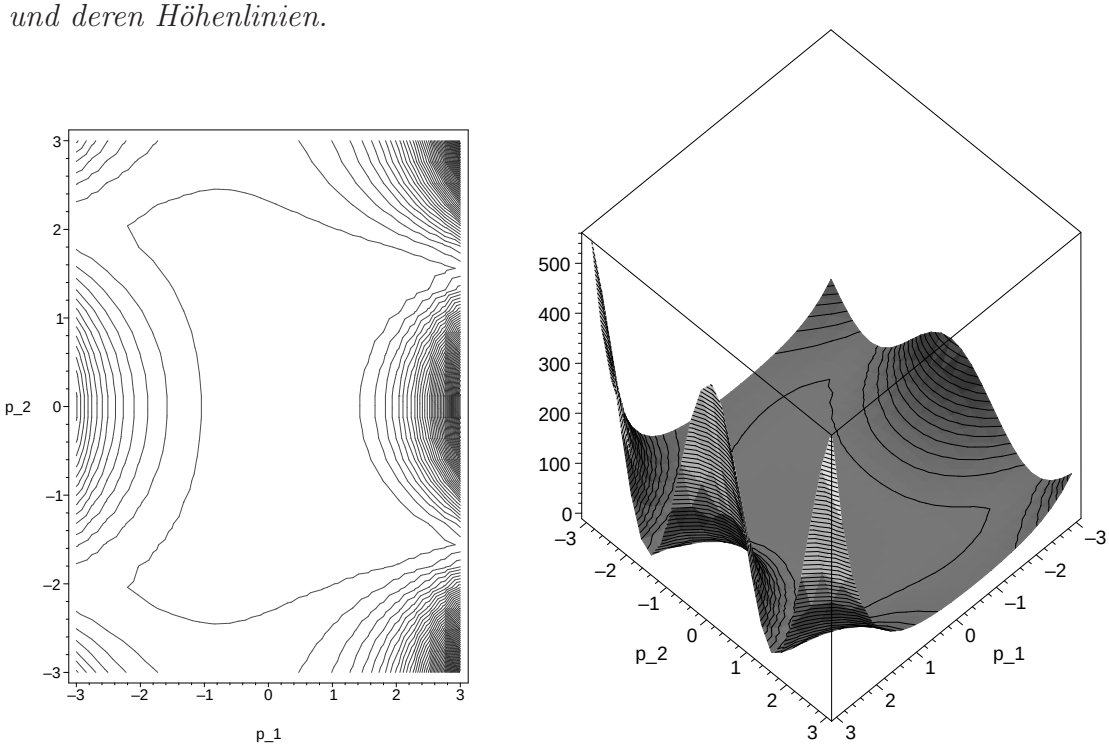
sollen durch die Funktion

$$f(t, p_1, p_2) = \exp(p_1 t)\cos(p_2 t)$$

wiedergegeben werden.
Die Abbildungen zeigen die zu minimierende Funktion

$$g(p_1, p_2) = \frac{1}{2}\sum_{i=1}^{10}\left(y_i - f(t_i, p_1, p_2)\right)^2$$

und deren Höhenlinien.



Die Fehlerfunktion zeigt, daß es Bereiche mit sehr steilen Flanken und sehr flache Täler gibt.

Anwendung eines SQP-Verfahrens liefert das folgende Resultat.

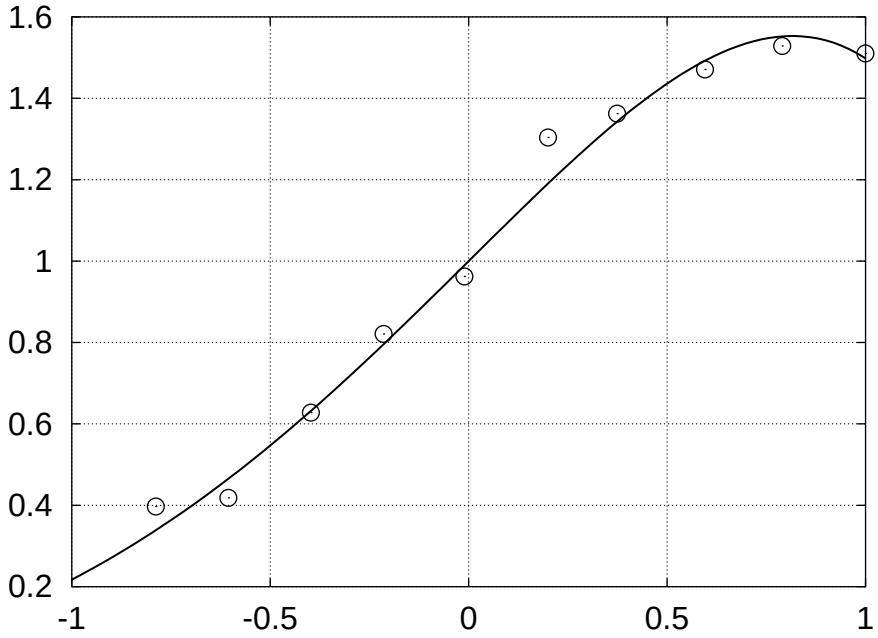
```
---------------------------------------- SQP VERSION 1.1 (C) Matthias Gerdts, University of Bayreuth, 2004 ----------------------------------------

      NUMBER OF VARIABLES               :          2
      NUMBER OF CONSTRAINTS             :          0
      METHOD                            :  SEQUENTIAL QUADRATIC PROGRAMMING (SQP)
      MERIT FUNCTION                    :  AUGMENTED LAGRANGIAN
      MULTIPLIER UPDATE RULE            :  SCHITTKOWSKI OR
      OPTIMALITY TOLERANCE              :  0.149E-07
      FEASIBILITY TOLERANCE             :  0.100E-11
      LINE SEARCH PARAMETER             :  SIGMA=  0.100E+00 BETA=  0.900E+00
      MAXIMUM NUMBER OF ITERATIONS      :      10000
      INFINITY                          :  0.100E+21
      ROUNDOFF TOLERANCE                :  0.300E-12
      REAL WORK SPACE PROVIDED          :       5338  NEEDED :          50
      INTEGER WORK SPACE PROVIDED       :         82  NEEDED :           6

-------------------------------------------------------------------------------------------------------------------------------------------------
 ITER    QPIT     ALPHA                OBJ                NB          KKT         PEN         |D|         DELTA        RDELTA         F/G
-------------------------------------------------------------------------------------------------------------------------------------------------
   0       0    0.0000E+00    0.6954883660193656E+01    0.0000E+00    0.4961E+01    0.0000E+00    0.0000E+00    0.0000E+00    0.1000E+01     1/   0    i
   1       1    0.4305E+00    0.4967522120517841E+01    0.0000E+00    0.6195E+01    0.0000E+00    0.5608E+01    0.0000E+00    0.1000E+01    10/   0
   2       1    0.1000E+01    0.3151863029012962E+00    0.0000E+00    0.2952E+01    0.0000E+00    0.2958E+01    0.0000E+00    0.1000E+01    11/   0
   3       1    0.2059E+00    0.1828168391624115E+00    0.0000E+00    0.1755E+01    0.0000E+00    0.1609E+01    0.0000E+00    0.1000E+01    27/   0
   4       1    0.1000E+01    0.1275778706056140E-01    0.0000E+00    0.1589E+00    0.0000E+00    0.2781E+00    0.0000E+00    0.1000E+01    28/   0
   5       1    0.1000E+01    0.1158079716928466E-01    0.0000E+00    0.7956E-01    0.0000E+00    0.8876E-02    0.0000E+00    0.1000E+01    29/   0
   6       1    0.1000E+01    0.1096782064438834E-01    0.0000E+00    0.1468E-01    0.0000E+00    0.1593E-01    0.0000E+00    0.1000E+01    30/   0
   7       1    0.1000E+01    0.1074391932113511E-01    0.0000E+00    0.1579E-01    0.0000E+00    0.1675E-01    0.0000E+00    0.1000E+01    31/   0
   8       1    0.1000E+01    0.1067396922629297E-01    0.0000E+00    0.3616E-02    0.0000E+00    0.1358E-01    0.0000E+00    0.1000E+01    32/   0
   9       1    0.1000E+01    0.1067272182372846E-01    0.0000E+00    0.7891E-03    0.0000E+00    0.9467E-03    0.0000E+00    0.1000E+01    33/   0
  10       1    0.1000E+01    0.1067267346498150E-01    0.0000E+00    0.5059E-04    0.0000E+00    0.1286E-03    0.0000E+00    0.1000E+01    34/   0
  11       1    0.1000E+01    0.1067267301930280E-01    0.0000E+00    0.1533E-05    0.0000E+00    0.3190E-04    0.0000E+00    0.1000E+01    35/   0
  12       1    0.1000E+01    0.1067267301842276E-01    0.0000E+00    0.7265E-07    0.0000E+00    0.1620E-05    0.0000E+00    0.1000E+01    36/   0
  13       1    0.1000E+01    0.1067267301842218E-01    0.0000E+00    0.9460E-09    0.0000E+00    0.2440E-07    0.0000E+00    0.1000E+01    37/   0
-------------------------------------------------------------------------------------------------------------------------------------------------
KKT CONDITIONS SATISFIED (IER= 0)!
SOLUTION:

OBJ =        0.1067267301842218E-01
KKT =        0.9460120590228484E-09
CON =        0.0000000000000000E+00
X   =
             0.9656009650544685E+00
            -0.9636591123058328E+00
```

Schließlich ergibt sich die Funktion f aus den identifizierten Parametern:



2.5 Klassifikationsprobleme und Support-Vector-Machine

(siehe O. Mangasarian: A Finite Newton Method for Classification Problems, Report 01-11, Optimization Methods and Software 17, 2002. Weitere Artikel und Details: http://www.cs.wisc.edu/∼olvi)

Gegeben sei das folgende **Klassifikationsproblem**: Es sollen m Punkte $x_i \in \mathbb{R}^n$, $i = 1, \ldots, m$ in die Klassen $+1$ oder -1 klassifiziert werden, siehe Abbildung 2.2. Die Punkte werden in der Matrix $A \in \mathbb{R}^{m \times n}$ zusammengefaßt, d.h. die i-te Zeile von A entspricht dem Punkt x_i.
Die Klassifikation erfolgt mit Hilfe einer Diagonalmatrix $D \in \mathbb{R}^{m \times m}$, die auf ihrer Diagonalen nur die Werte $+1$ oder -1 annimmt. Dabei bedeutet $d_{ii} = +1$, daß Vektor x_i zur Klasse $+1$ gehört. Entsprechend bedeutet $d_{ii} = -1$, daß Vektor x_i zur Klasse -1 gehört. Das Klassifikationsproblem wird mit Hilfe der **Support-Vector-Machine** gelöst. Diese bestimmt eine trennende Hyperebene

$$w^\top x = \gamma$$

und sogenannte Randhyperebenen

$$\begin{aligned} w^\top x &= \gamma + 1, \\ w^\top x &= \gamma - 1, \end{aligned}$$

mit Abstand $2/\|(w, \gamma)\|$, vgl. Abbildung 2.2, die die Vektoren der jeweiligen Klassen (zumindest theoretisch) trennen. Diese Randhyperebenen sollen die Vektoren der jeweiligen Klassen trennen, d.h.

$$\begin{aligned} w^\top x_i - \gamma &\geq +1, \text{ falls } x_i \text{ zur Klasse } d_{ii} = +1 \text{ gehört}, \\ w^\top x_i - \gamma &\leq -1, \text{ falls } x_i \text{ zur Klasse } d_{ii} = -1 \text{ gehört}. \end{aligned}$$

In der Praxis wird es i.a. nicht möglich sein, die Punkte tatsächlich zu trennen. Daher werden Fehler $y_i \geq 0$ zugelassen:

$$\begin{aligned} w^\top x_i - \gamma + y_i &\geq +1, \text{ falls } d_{ii} = +1, \\ w^\top x_i - \gamma - y_i &\leq -1, \text{ falls } d_{ii} = -1. \end{aligned}$$

Insgesamt läßt sich das Problem als quadratisches Optimierungsproblem formulieren, in dem der Abstand der Randhyperebenen maximiert bzw. $\|(w, \gamma)\|$ minimiert wird:

$$\begin{array}{ll} \min_{(w,\gamma,y) \in \mathbb{R}^{n+1+m}} & \frac{\nu}{2} y^\top y + \frac{1}{2}(w^\top w + \gamma^2), \\ \text{unter} & D\,(Aw - e\gamma) + y \geq e,\ y \geq 0, \end{array}$$

wobei ν einen festen Gewichtungsparameter bezeichnet und $e = (1, \ldots, 1)^\top$ ist.
Da m in der Praxis durchaus Werte jenseits der Million annehmen kann (der Autor O. Mangasarian hat sogar einen Test mit einer Milliarde Punkten durchgeführt), wird alternativ das äquivalente einmal stetig differenzierbare Problem

$$\min_{(w,\gamma) \in \mathbb{R}^{n+1}} \frac{\nu}{2} \| \max\{0, e - D\,(Aw - e\gamma)\} \|^2 + \frac{1}{2}(w^\top w + \gamma^2)$$

betrachtet und mit Hilfe eines nichtglatten Newtonverfahrens gelöst.

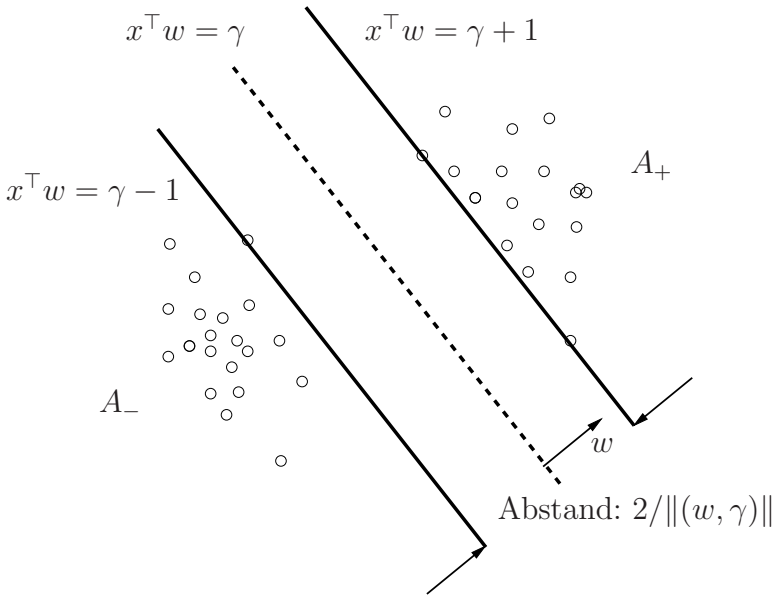


Abbildung 2.2: Trennende Hyperebenen und Support-Vector-Machine zur Klassifizierung von Daten.

Die Support-Vector-Machine hat u.a. Anwendungen in der Brustkrebstherapie (Klassifizierung der Patienten in solche, die von einer Behandlung mit Chemotherapie profitieren würden bzw. in solche, die nicht profitieren würden.).

2.6 Diskretisierte Optimalsteuerungsprobleme

Sei $[t_0, t_f] \subset \mathbb{R}$ ein nichtleeres und beschränktes Intervall mit festen Zeitpunkten $t_0 < t_f$. Seien stetig differenzierbare Abbildungen

$$\begin{aligned}
\varphi &: \mathbb{R}^{n_x} \times \mathbb{R}^{n_x} \to \mathbb{R}, \\
f_0 &: [t_0, t_f] \times \mathbb{R}^{n_x} \times \mathbb{R}^{n_u} \to \mathbb{R}, \\
f &: [t_0, t_f] \times \mathbb{R}^{n_x} \times \mathbb{R}^{n_u} \to \mathbb{R}^{n_x}, \\
\psi &: \mathbb{R}^{n_x} \times \mathbb{R}^{n_x} \to \mathbb{R}^{n_\psi}, \\
c &: [t_0, t_f] \times \mathbb{R}^{n_x} \times \mathbb{R}^{n_u} \to \mathbb{R}^{n_c}, \\
s &: [t_0, t_f] \times \mathbb{R}^{n_x} \to \mathbb{R}^{n_s}
\end{aligned}$$

gegeben. Betrachte das folgende Optimalsteuerungsproblem.

Finde Funktionen $x(\cdot)$ und $u(\cdot)$, so daß die Zielfunktion

$$\varphi(x(t_0), x(t_f)) + \int_{t_0}^{t_f} f_0(t, x(t), u(t))dt$$

minimiert wird unter Beachtung der **Differentialgleichung**

$$\dot{x}(t) = f(t, x(t), u(t)) \qquad \text{f.ü. in } [t_0, t_f],$$

der **Randbedingungen**

$$\psi(x(t_0), x(t_f)) = 0,$$

und der **Zustandsbeschränkungen**

$$c(t, x(t), u(t)) \leq 0 \qquad \text{f.ü. in } [t_0, t_f].$$

Diskretisierung der Differentialgleichungen auf einem Gitter

$$t_i = t_0 + ih, \quad i = 0, 1, \ldots, N, \; h = \frac{t_f - t_0}{N}$$

mit dem Eulerverfahren gemäß

$$x_{i+1} \;=\; x_i + hf(t_i, x_i, u_i), \quad i = 0, 1, \ldots, N-1$$

wobei $x_i \approx x(t_i)$ und $u_i \approx u(t_i)$ Approximationen des Zustands bzw. der Steuerung darstellen und Approximation der Zielfunktion gemäß

$$\int_{t_0}^{t_f} f_0(t, x(t), u(t))dt \approx h \sum_{i=0}^{N-1} f_0(t_i, x_i, u_i)$$

liefert das endlichdimensionale, nichtlineare Optimierungsproblem

$$\begin{array}{lll}
\text{Minimiere} & \varphi(x_0, x_N) + h \displaystyle\sum_{i=0}^{N-1} f_0(t_i, x_i, u_i) & \\
\text{bezüglich} & x_i, \; i = 0, 1, \ldots, N-1, & \\
& u_i, \; i = 0, 1, \ldots, N-1 & \\
\text{unter} & x_{i+1} - x_i - hf(t_i, x_i, u_i) \;=\; 0, & i = 0, 1, \ldots, N-1, \\
& c(t_i, x_i, u_i) \;\leq\; 0, & i = 0, 1, \ldots, N, \\
& \psi(x_0, x_N) \;=\; 0. &
\end{array}$$

Beispiel 2.6.1 (Minimum Energy)

Gegeben sei folgendes Optimalsteuerungsproblem (vgl. Bryson/Ho [BH75], S.120, Sec. 3.11,

Ex. 2): Minimiere

$$\frac{1}{2}\int_0^1 u(t)^2 dt$$

unter den Differentialgleichungsnebenbedingungen

$$\begin{aligned} \dot{x}_1(t) &= x_2(t), \\ \dot{x}_2(t) &= u(t), \end{aligned}$$

den Randbedingungen

$$x_1(0) = x_1(1) = 0, \qquad x_2(0) = -x_2(1) = 1,$$

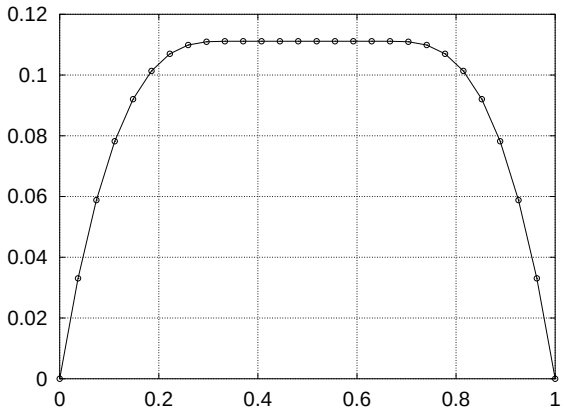
und der Zustandsbeschränkung

$$x_1(t) - \frac{1}{9} \leq 0.$$

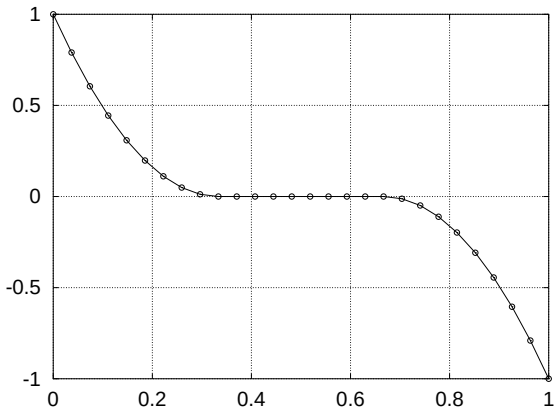
Die Lösung dieses Optimalsteuerungsproblems beschreibt die Form eines an beiden Enden eingespannten Balkens unter einer Last (x_1 beschreibt die Form des Balkens, u ist ein Mass für die Krümmung der Kurve).

Als numerische Lösung des mit $N = 27$ diskretisierten Problems ergeben sich die folgenden Funktionen (genau genommen wurde das klassische Runge-Kutta Verfahren anstatt des Eulerverfahrens zur Diskretisierung der Differentialgleichungen verwendet):

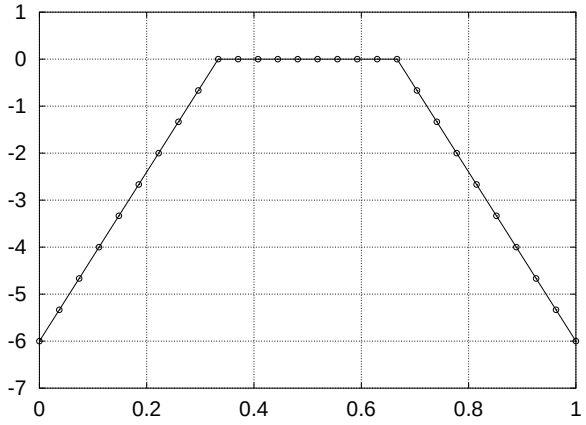
Zustand $x_1(t)$:



Zustand $x_2(t)$:



Steuerung $u(t)$:



2.7 Bestimmung eines zulässigen Punkts

Wir betrachten ein Beispiel aus der Biochemie. Gegeben sind n Proteine und ein Peptid P. Die n Proteine $1, \ldots, n$ reagieren mit dem Peptid und bilden Komplexe $1P, \ldots, nP$ mit Konzentrationen $[1P], \ldots, [nP]$. Die totalen Konzentrationen $[P]_t, [1]_t, \ldots, [n]_t$ und die Reaktionsgeschwindigkeiten $k_1, \ldots, k_n$ sind bekannt.
Es bestehen die folgenden Zusammenhänge:

$$\begin{aligned} [P]_f &= [P]_t - ([1P] + ... + [nP]), \\ [i]_f &= [i]_t - [iP], \qquad i = 1, ..., n, \\ [iP] \cdot k_i &= [i]_f \cdot [P]_f, \qquad i = 1, ..., n \end{aligned}$$

Gesucht sind die Konzentrationen der Komplexe $[iP]$, $i = 1, ..., n$, und die Konzentration $[P]_f$.
Einsetzen der obigen Beziehungen führen auf das nichtlineare Gleichungssystem

$$[iP] \cdot \left(k_i + [P]_t - \sum_{j=1}^{n} [jP] \right) - [i]_t \cdot \left([P]_t - \sum_{j=1}^{n} [jP] \right)) = 0, \qquad i = 1, ..., n.$$

Dieses nichtlineare System für $[iP]$, $i = 1, \ldots, n$ kann nicht analytisch gelöst werden und besitzt i.a. mehrere Lösungen. Wir interessieren uns aber nur für nicht-negative Konzentrationen. Daher wird eine nicht-negative Lösung $0 \leq [iP] \leq [i]_t 0$, $i = 1, \ldots, n$, gesucht, die das obige nichtlineare Gleichungssystem löst und $[P]_f \geq 0$ erfüllt.
Dieses Problem kann als degeneriertes Optimierungsproblem aufgefasst werden, indem die Zielfunktion konstant auf Null gesetzt wird. Damit entsteht ein Optimierungsproblem, welches zum Ziel hat, einen zulässigen Punkt zu bestimmen. Der zulässige Bereich besteht aus nichtlinearen Gleichungen und linearen Ungleichungen.

Kapitel 3

Unrestringierte Optimierung

Gegenstand dieses Kapitels sind die theoretische Untersuchung und die Entwicklung numerischer Verfahren für das

unrestringierte Optimierungsproblem (UOP):

$$\min f(x) \quad \text{unter} \quad x \in \mathbb{R}^n.$$

Darin sei $f : \mathbb{R}^n \to \mathbb{R}$ eine beliebige Funktion.
Wir werden **notwendige Bedingungen** und **hinreichende Bedingungen** für ein lokales Minimum entwickeln.

- **Notwendige Bedingungen** sind Bedingungen, die ein lokales Minimum $\hat{x}$ zwangsläufig erfüllt. Bei der Herleitung wird daher stets vorausgesetzt, daß bereits ein lokales Minimum $\hat{x}$ bekannt ist. Allgemein werden wir sehen, daß die Bedingung $\nabla f(\hat{x}) = 0$ für unrestringierte Optimierungsprobleme eine notwendige Bedingung darstellt. Sie ist aber nicht hinreichend.

- Im Gegensatz zu den notwendigen Bedingungen ist bei **hinreichenden Bedingungen** a priori nicht bekannt, ob es sich bei einem Kandidaten $\hat{x}$ um ein lokales Minimum handelt oder nicht. Hinreichende Bedingungen sind Bedingungen, mit denen entschieden werden kann, ob es sich bei $\hat{x}$ um ein lokales Minimum handelt oder nicht. Später wird gezeigt, daß die Bedingung „$H_f(\hat{x})$ positiv definit“, wobei $H_f(\hat{x})$ die Hessematrix von f in $\hat{x}$ bezeichnet, in der unrestringierten Optimierung eine hinreichende Bedingung für ein Minimum ist. Sie ist aber nicht notwendig.

- Ideal sind Bedingungen, die sowohl notwendig als auch hinreichend sind. Derartige Bedingungen sind aber nur für spezielle Optimierungsprobleme, etwa konvexe Probleme, bekannt.

Anschließend werden numerische Verfahren diskutiert und analysiert. Die meisten der vorzustellenden Verfahren versuchen, die notwendigen Bedingungen zu erfüllen.

Zur Erinnerung:

- Der **Gradient** einer Funktion $f : \mathbb{R}^n \to \mathbb{R}$ an der Stelle $x = (x_1, \dots, x_n)^\top \in \mathbb{R}^n$ ist formal definiert als der Spaltenvektor

$$\nabla f(x_1, \dots, x_n) := \begin{pmatrix} \frac{\partial f}{\partial x_1}(x_1, \dots, x_n) \\ \vdots \\ \frac{\partial f}{\partial x_n}(x_1, \dots, x_n) \end{pmatrix}.$$

  Im Spezialfall $n = 1$ ist der Gradient gerade die erste Ableitung von f. Bekanntlich zeigt der Gradient einer Funktion f in die **Richtung des steilsten Anstiegs** von f. Außerdem steht der Gradient **senkrecht** auf den Höhenlinien von f.

- Die **Hessematrix** von f an der Stelle x ist gegeben durch

$$H_f(x) = \begin{pmatrix} \frac{\partial^2 f(x)}{\partial x_1 \partial x_1} & \cdots & \frac{\partial^2 f(x)}{\partial x_1 \partial x_n} \\ \vdots & \ddots & \vdots \\ \frac{\partial^2 f(x)}{\partial x_n \partial x_1} & \cdots & \frac{\partial^2 f(x)}{\partial x_n \partial x_n} \end{pmatrix}.$$

  Häufig schreibt man auch $\nabla^2 f(x)$ anstatt $H_f(x)$.

  Im Spezialfall $n = 1$ ist die Hessematrix gerade die zweite Ableitung von f. Die Hessematrix beschreibt anschaulich die lokale Krümmung einer Funktion.

Die folgenden Betrachtungen basieren auf der lokalen Approximierbarkeit von f in $\hat{x}$. Nach dem **Satz von Taylor** (**Taylorentwicklung**) gilt:

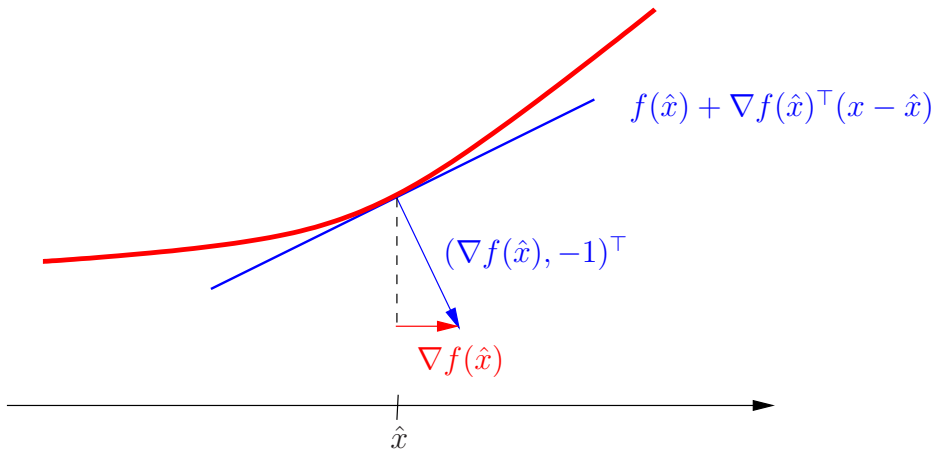
- Ist f stetig differenzierbar, so gilt

$$f(x) = f(\hat{x}) + \nabla f(\hat{x})^\top (x - \hat{x}) + o(\|x - \hat{x}\|)$$

  mit $\lim_{x \to \hat{x}} \frac{o(\|x-\hat{x}\|)}{\|x-\hat{x}\|} = 0$.

  Somit kann f in einer Umgebung von $\hat{x}$ approximiert werden durch eine affin lineare Funktion:

$$f(x) \approx f(\hat{x}) + \nabla f(\hat{x})^\top (x - \hat{x}).$$



- Ist f zweimal stetig differenzierbar, so gilt

$$f(x) = f(\hat{x}) + \nabla f(\hat{x})^\top (x-\hat{x}) + \frac{1}{2}(x-\hat{x})^\top H_f(\hat{x})(x-\hat{x}) + o(\|x-\hat{x}\|^2)$$

mit $\lim_{x\to\hat{x}} \frac{o(\|x-\hat{x}\|^2)}{\|x-\hat{x}\|^2} = 0$.

Somit kann f in einer Umgebung von $\hat{x}$ approximiert werden durch eine quadratische Funktion:

$$f(x) \approx f(\hat{x}) + \nabla f(\hat{x})^\top (x-\hat{x}) + \frac{1}{2}(x-\hat{x})^\top H_f(\hat{x})(x-\hat{x}). \tag{3.1}$$

3.1 Notwendige Bedingungen

Satz 3.1.1 (Notwendige Bedingung erster Ordnung)
Sei $D \subseteq \mathbb{R}^n$ offen, $f : D \to \mathbb{R}$ stetig differenzierbar und $\hat{x} \in D$ ein lokales Minimum von f. Dann gilt

$$\nabla f(\hat{x}) = 0.$$

Beweis: Die Funktion $\varphi(t) = f(\hat{x} - t\nabla f(\hat{x}))$ ist für kleines $|t|$, $t \in \mathbb{R}$ stetig differenzierbar mit

$$\varphi'(0) = -f'(\hat{x})\nabla f(\hat{x}) = -\|\nabla f(\hat{x})\|^2.$$

Also gilt $\varphi'(0) < 0$ falls $\nabla f(\hat{x}) \neq 0$.
Wir nehmen nun an, daß $\nabla f(\hat{x}) \neq 0$ gilt. Für die Richtungsableitung in Richtung $d = -\nabla f(\hat{x})$ gilt dann

$$f'(\hat{x}; d) = \lim_{\alpha \downarrow 0} \frac{f(\hat{x} - \alpha\nabla f(\hat{x})) - f(\hat{x})}{\alpha} = -\nabla f(\hat{x})^\top \nabla f(\hat{x}) = -\|\nabla f(\hat{x})\|^2 < 0.$$

Also gibt es ein $\bar{\alpha} > 0$ mit $\hat{x} + \alpha d \in D$ und

$$\frac{f(\hat{x} - \alpha\nabla f(\hat{x})) - f(\hat{x})}{\alpha} < 0$$

für alle $\alpha \in (0, \bar{\alpha}]$. Folglich gilt

$$f(\hat{x} - \alpha\nabla f(\hat{x})) < f(\hat{x})$$

für alle $\alpha \in (0, \bar{\alpha}]$ im Widerspruch zur lokalen Minimalität von $\hat{x}$. □

Definition 3.1.2 (stationärer Punkt)
Jeder Punkt x mit $\nabla f(x) = 0$ heißt **stationärer Punkt von** f.

Die Bedingung $\nabla f(\hat{x}) = 0$ ist nur notwendig, aber nicht hinreichend.

Beispiel 3.1.3

Betrachte die Funktionen $f(x) = x^2$ *und* $g(x) = -x^2$.
In $\hat{x} = 0$ *gilt* $f'(\hat{x}) = 0$ *und* $g'(\hat{x}) = 0$. *Allerdings besitzt* f *in* $\hat{x} = 0$ *ein globales Minimum, während* g *dort ein globales Maximum besitzt.*

Wir wollen eine weitere notwendige Bedingung herleiten. Dazu greifen wir auf die Taylorentwicklung (3.1) zurück. Ist $\hat{x}$ ein stationärer Punkt von f, so gilt $\nabla f(\hat{x}) = 0$ und die Taylorentwicklung in (3.1) reduziert sich zu

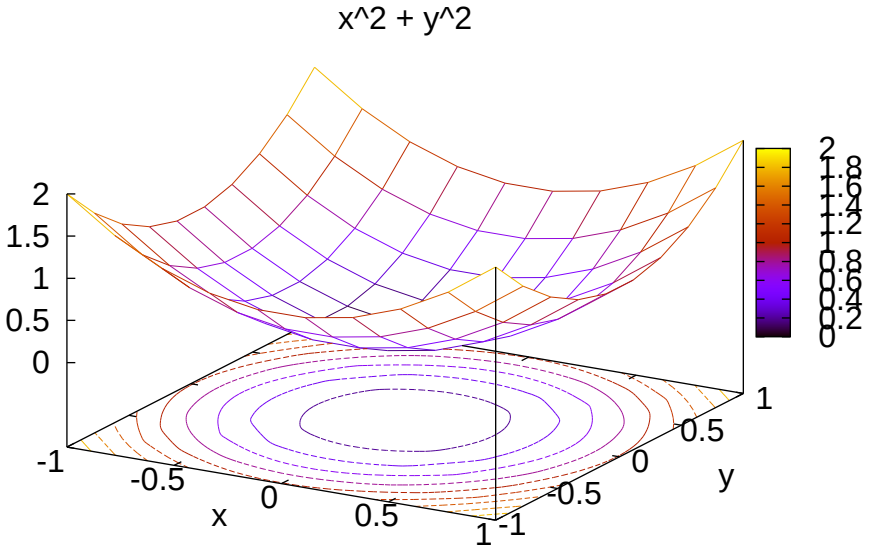
$$f(x) \approx f(\hat{x}) + \frac{1}{2}(x - \hat{x})^\top H_f(\hat{x})(x - \hat{x}).$$

Damit verhält sich f lokal genauso wie die quadratische Funktion

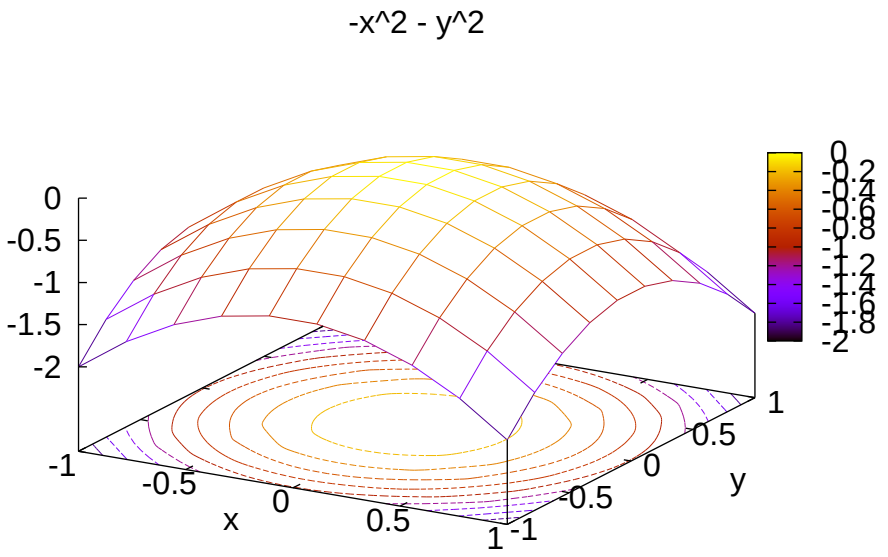
$$q(x) = f(\hat{x}) + \frac{1}{2}(x - \hat{x})^\top H_f(\hat{x})(x - \hat{x}).$$

Wir veranschaulichen mögliche Fälle im $\mathbb{R}^2$:

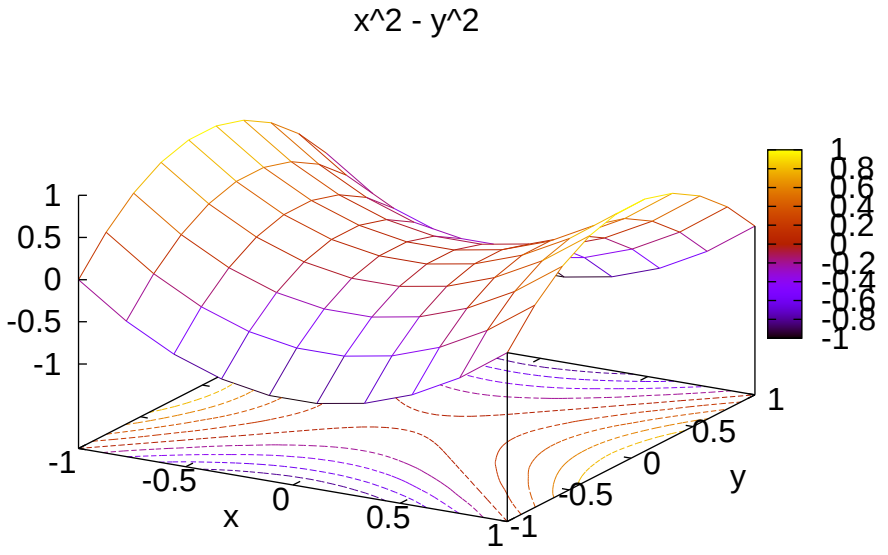
- $f(x,y) = x^2 + y^2$, $H_f(\hat{x}) = \begin{pmatrix} 2 & 0 \\ 0 & 2 \end{pmatrix}$ positiv definit:



- $f(x,y) = -x^2 - y^2$, $H_f(\hat{x}) = \begin{pmatrix} -2 & 0 \\ 0 & -2 \end{pmatrix}$ negativ definit:



- $f(x,y) = x^2 - y^2$, $H_f(\hat{x}) = \begin{pmatrix} 2 & 0 \\ 0 & -2 \end{pmatrix}$ indefinit:



Satz 3.1.4 **(Notwendige Bedingung zweiter Ordnung)**

Sei $D \subseteq \mathbb{R}^n$ offen, $f : D \to \mathbb{R}$ zweimal stetig differenzierbar und $\hat{x} \in D$ ein lokales Minimum von f. Dann ist die Hessematrix $\nabla^2 f(\hat{x})$ positiv semidefinit.

Beweis: Wäre $\nabla^2 f(\hat{x})$ nicht positiv semidefinit, so gäbe es ein $d \in \mathbb{R}^n$, $d \neq 0$ mit

$d^\top \nabla^2 f(\hat{x}) d < 0$. Mit dem Satz von Taylor folgt unter Ausnutzung von $\nabla f(\hat{x}) = 0$ die Beziehung

$$f(\hat{x} + td) = f(\hat{x}) + \frac{1}{2} t^2 d^\top \nabla^2 f(\hat{x} + \xi_t d) d$$

mit einem $\xi_t \in (0, t)$. Für hinreichend kleines $t > 0$ folgt aus der Stetigkeit von $\nabla^2 f(\hat{x})$, daß auch $d^\top \nabla^2 f(\hat{x} + \xi_t d) d < 0$ gilt. Somit folgt $f(\hat{x} + td) < f(\hat{x})$ für alle hinreichend kleinen Werte $t > 0$. Dies ist ein Widerspruch zur lokalen Minimalität von $\hat{x}$. □

Beispiel 3.1.5

Betrachte wieder die Funktionen $f(x) = x^2$ und $g(x) = -x^2$. In beiden Fällen ist $\hat{x} = 0$ ein stationärer Punkt.

Wegen $f''(0) = 2 > 0$ erfüllt f die notwendige Bedingung zweiter Ordnung.

Wegen $g''(0) = -2 < 0$ erfüllt g die notwendige Bedingung zweiter Ordnung nicht, d.h. $\hat{x} = 0$ ist kein lokales Minimum der Funktion g.

Das folgende Beispiel zeigt, daß die notwendigen Bedingung zweiter Ordnung ebenfalls nicht hinreichend für lokale Optimalität ist.

Beispiel 3.1.6

Die Funktion $f(x) = x^3$ erfüllt ebenfalls $f'(0) = 0$ und sogar $f''(0) = 0$, ist also positiv semidefinit. Somit sind beide notwendigen Bedingungen erfüllt. Jedoch besitzt sie weder ein lokales noch globales Minimum oder Maximum in $x = 0$.

Beispiel 3.1.7

Betrachte $f(x_1, x_2) := x_1^2 - x_2^4$. Es gilt

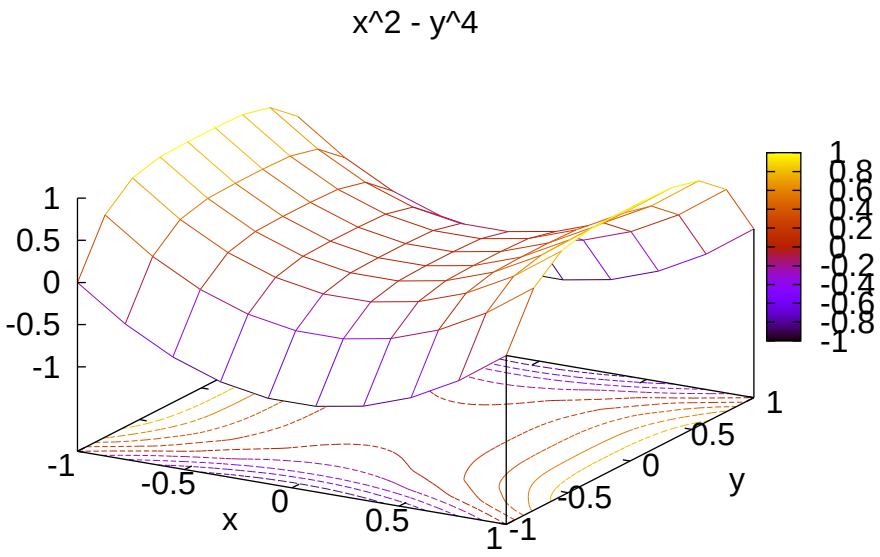
$$\nabla f(x_1, x_2) = \begin{pmatrix} 2x_1 \\ -4x_2^3 \end{pmatrix}.$$

Daher ist $\hat{x} = (0, 0)^\top$ der einzige stationäre Punkt von f.

Weiter gilt

$$H_f(x_1, x_2) = \begin{pmatrix} 2 & 0 \\ 0 & -12x_2^2 \end{pmatrix} \qquad \text{bzw.} \qquad H_f(0, 0) = \begin{pmatrix} 2 & 0 \\ 0 & 0 \end{pmatrix}$$

Die Hessematrix ist positiv semidefinit, d.h. $\hat{x}$ erfüllt die notwendige Bedingung zweiter Ordnung. Allerdings ist $\hat{x}$ kein lokales Minimum, sondern ein Sattelpunkt.



Fazit:
Bestimmt man alle Punkte, die die notwendigen Bedingungen $\nabla f(x) = 0$ und $\nabla^2 f(x)$ positiv semidefinit erfüllen, so sind diese Punkte lediglich Kandidaten für ein (lokales) Minimum.

Wir werden später erkennen, daß die meisten numerischen Verfahren Folgen erzeugen, die gegen einen stationären Punkt konvergieren bzw. deren Häufungspunkte stationäre Punkte sind.

3.2 Hinreichende Bedingungen

Um entscheiden zu können, ob ein Punkt $\hat{x}$, der die notwendigen Bedingungen erfüllt, tatsächlich ein (lokales) Minimum ist, werden hinreichende Bedingungen benötigt.

Satz 3.2.1
Sei $D \subseteq \mathbb{R}^n$ offen, $f : D \to \mathbb{R}$ zweimal stetig differenzierbar und $\hat{x} \in D$ ein Punkt mit $\nabla f(\hat{x}) = 0$ und positiv definiter Hessematrix $\nabla^2 f(\hat{x})$. Dann ist $\hat{x}$ ein striktes lokales Minimum von f auf D.

Beweis: Taylorentwicklung von f um $\hat{x}$ liefert

$$f(\hat{x}+d) = f(\hat{x}) + \underbrace{\nabla f(\hat{x})^\top d}_{=0} + \frac{1}{2} d^\top \nabla^2 f(\hat{x}+\xi d) d$$

mit $\xi \in (0,1)$. Da $\nabla^2 f(\hat{x})$ positiv definit ist, gibt es ein $\alpha > 0$, so daß für alle $d \in \mathbb{R}^n$ gilt

$$d^\top \nabla^2 f(\hat{x}) d \geq \alpha d^\top d.^1$$

Damit folgt mit der Cauchy-Schwarzschen Ungleichung[2] die Beziehung

$$\begin{aligned} f(\hat{x}+d) &= f(\hat{x}) + \frac{1}{2} d^\top \nabla^2 f(\hat{x}+\xi d) d \\ &= f(\hat{x}) + \frac{1}{2} d^\top \nabla^2 f(\hat{x}) d + \frac{1}{2} d^\top \left(\nabla^2 f(\hat{x}+\xi d) - \nabla^2 f(\hat{x}) \right) d \\ &\geq f(\hat{x}) + \frac{1}{2} \left(\alpha - \|\nabla^2 f(\hat{x}+\xi d) - \nabla^2 f(\hat{x})\| \right) \|d\|^2. \end{aligned}$$

Aufgrund der Stetigkeit von $\nabla^2 f(\hat{x})$ gilt $\|\nabla^2 f(\hat{x}+\xi d) - \nabla^2 f(\hat{x})\| \to 0$ für $d \to 0$. Es folgt $f(\hat{x}+d) > f(\hat{x})$ für alle hinreichend kleinen Vektoren $d \neq 0$ und somit ist $\hat{x}$ striktes lokales Minimum von f. □

Bemerkung 3.2.2

- *Die hinreichende Bedingung ist i.a. nicht notwendig. Ein Gegenbeispiel liefert die Funktion $f(x_1, x_2) := x_1^2 + x_2^4$. Hier ist $x = 0$ ein striktes globales Minimum, die Hesse-Matrix ist jedoch nicht positiv definit.*

- *Numerisch kann die positive Definitheit einer Matrix überprüft werden, indem ihre Eigenwerte berechnet werden, etwa mit dem QR-Verfahren. Da die Hessematrix einer zweimal stetig differenzierbaren Funktion symmetrisch ist, sind sämtliche Eigenwerte reell. Sind sämtliche Eigenwerte positiv (negativ), so ist die Matrix positiv (negativ) definit. Sind die Eigenwerte ≥ 0 (≤ 0), so ist die Matrix positiv (negativ) semidefinit. Treten sowohl negative als auch positive Eigenwerte auf, so ist die Matrix indefinit. In diesem Fall ist $\hat{x}$ ein* **Sattelpunkt**, *d.h. in jeder Umgebung von $\hat{x}$ gibt es Punkte x_1, x_2 mit $f(x_1) < f(\hat{x}) < f(x_2)$.*

- *Da unter den Voraussetzungen des Satzes die Hesse-Matrix von f auch in einer ganzen Umgebung $K_\varepsilon(\hat{x})$ positiv definit ist, folgt die Existenz einer (von ε abhängigen) Konstanten $\mu > 0$ mit*

$$f(\hat{x}+d) - f(\hat{x}) \;=\; (1/2)\, d^\top \nabla^2 f(\tilde{x})\, d \;\geq\; (1/2)\, \mu \,\|\, d \,\|_2^2$$

  für alle $d \in K_\varepsilon(0)$. Die Funktion f wächst also bei $\hat{x}$ wenigstens **quadratisch** *mit $\|x - \hat{x}\|$ an.*

[1] Diese Behauptung ist nur für endlichdimensionale Vektorräume richtig. Sie basiert auf der Kompaktheit der Einheitskugel in endlichdimensionalen Vektorräumen. In unendlichdimensionalen Vektorräumen gilt dies nicht mehr.

[2] Cauchy-Schwarzsche Ungleichung: $|\langle a, b\rangle| = |a^\top b| \leq \|a\| \cdot \|b\|$, $a, b \in \mathbb{R}^n$

Beispiel 3.2.3

Für die Funktion

$$f(x,y) := y^2(x-1) + x^2(x+1)$$

berechnen wir den Gradienten

$$\nabla f(x,y) = \begin{pmatrix} y^2 + 3x^2 + 2x \\ 2y(x-1) \end{pmatrix}$$

und die Hessematrix

$$\nabla^2 f(x,y) = \begin{pmatrix} 6x+2 & 2y \\ 2y & 2(x-1) \end{pmatrix}.$$

Aus $\nabla f(x,y) = 0$ *ergeben sich somit die stationären Punkte* $(x^0,y^0) = (0,0)$ *und* $(x^1,y^1) = (-2/3, 0)$.

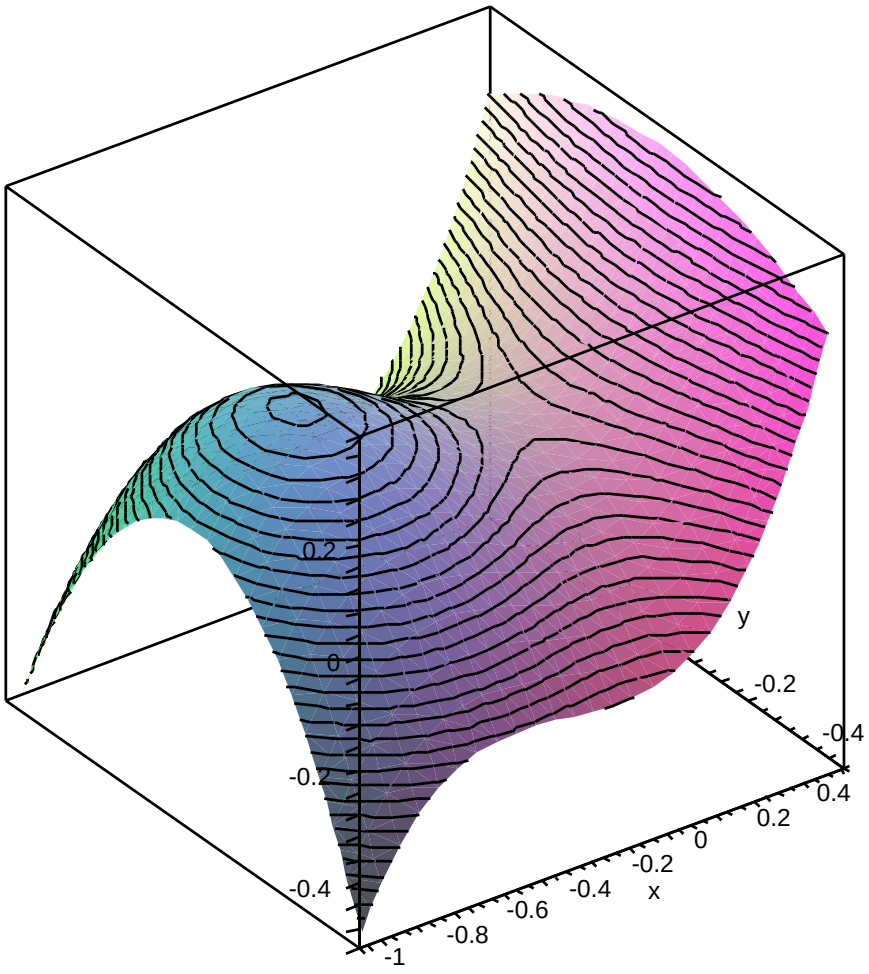
Für die zugehörigen Hesse-Matrizen erhält man

$$\nabla^2 f(x^0,y^0) = \begin{pmatrix} 2 & 0 \\ 0 & -2 \end{pmatrix} \textit{ indefinit } \Rightarrow \begin{pmatrix} x^0 \\ y^0 \end{pmatrix} \textit{ Sattelpunkt}$$

$$\nabla^2 f(x^1,y^1) = \begin{pmatrix} -2 & 0 \\ 0 & -\frac{10}{3} \end{pmatrix} \textit{ negativ definit } \Rightarrow \begin{pmatrix} x^1 \\ y^1 \end{pmatrix} \textit{ striktes lokales Maximum}$$

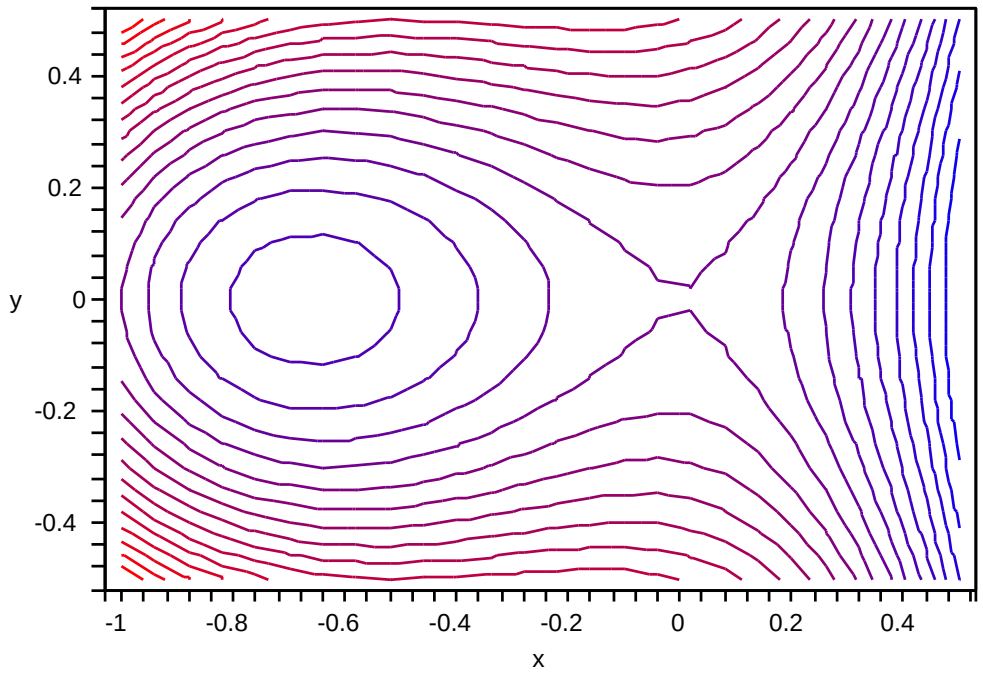
MAPLE-*Befehl*

```
plot3d(y^2*(x-1)+x^2*(x+1),x=-1..0.5,y=-0.5..0.5,axes=boxed,
       style=PATCHCONTOUR,contours = 40,shading=XYZ):
```



MAPLE-*Befehl*

```
contourplot(y^2*(x-1)+x^2*(x+1),x=-1..0.5,y=-0.5..0.5,contours=20,
            coloring=[red,blue],scaling=constrained,axes=boxed):
```

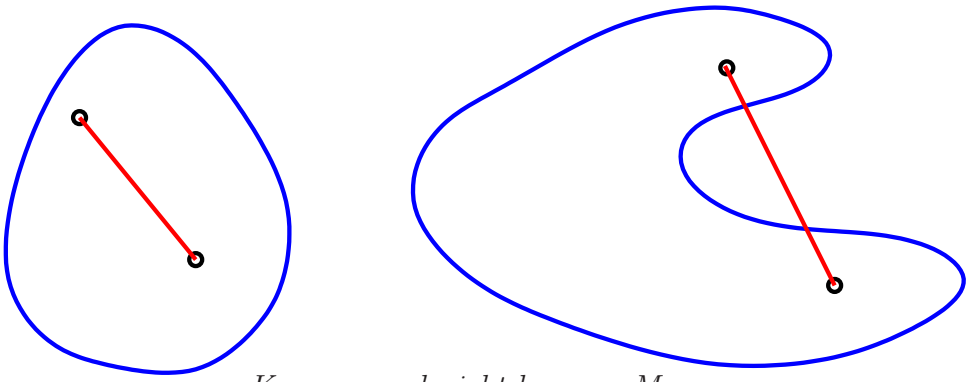


3.3 Konvexe Funktionen

Definition 3.3.1

Eine Menge $X \subseteq \mathbb{R}^n$ heißt **konvex**, *falls mit zwei Punkten stets auch deren gesamte Verbindungsstrecke zu X gehört, also*

$$\forall\, x, y \in X : \forall\, \lambda \in [0,1] : \; x + \lambda\,(y - x) \in X.$$



Konvexe und nicht konvexe Menge

Bemerkung 3.3.2

- *Beliebiger Durchschnitt konvexer Mengen ist konvex.*
- *Zu jeder Menge $X \subseteq \mathbb{R}^n$ gibt es eine (eindeutig bestimmte) kleinste konvexe Menge $co(X)$, die diese umfasst. Diese heißt die* **konvexe Hülle** *von X.*

  Sie besitzt die folgenden Darstellungen:

$$\begin{aligned} co(X) &= \bigcap\{K \subset \mathbb{R}^n \mid X \subset K \wedge K \text{ konvex}\} \\ &= \left\{ \sum_{i=1}^{m} \lambda_i x_i \mid x_i \in X,\ \lambda_i \geq 0,\ \sum_{i=1}^{m} \lambda_i = 1 \right\} \end{aligned}$$

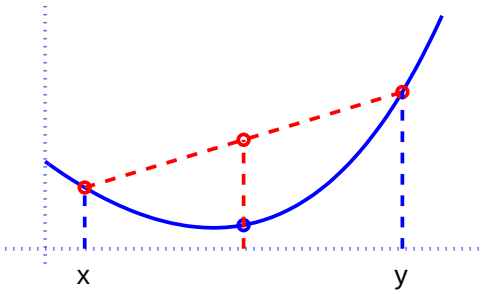
Definition 3.3.3

- *Sei $X \subseteq \mathbb{R}^n$ konvex. Eine Funktion $f : \mathbb{R}^n \to \mathbb{R}$ heißt* **konvex** *auf X, falls*

$$\forall\, x, y \in X : \forall\, \lambda \in [0,1] : \ f(x + \lambda\,(y - x)) \ \leq \ f(x) \ + \ \lambda\,(f(y) - f(x)).$$

- *Die Funktion f heißt* **strikt konvex** *auf X, falls*

$$\forall\, x \neq y \in X : \forall\, \lambda \in]0,1[: \ f(x + \lambda\,(y - x)) \ < \ f(x) \ + \ \lambda\,(f(y) - f(x)).$$



Satz 3.3.4

(a) Ist $f : \mathbb{R}^n \to \mathbb{R}$ konvex auf einer nichtleeren, konvexen Menge X, so ist die Menge der globalen Minima konvex. Ist f sogar strikt konvex, so gibt es höchstens ein globales Minimum von f über X (Eindeutigkeit).

(b) Jedes lokale Minimum einer konvexen Funktion f über X (konvex) ist zugleich global. Ist f konvex und stetig differenzierbar, so ist jeder stationäre Punkt bereits ein globales Minimum von f über X.

Beweis:

(a) Aus $f(x) = f(y) = \inf f(X),\ x, y \in X$ und $\lambda \in [0,1]$ folgt

$$f(x + \lambda\,(y - x)) \ \leq\ f(x) + \lambda\,(f(y) - f(x)) \ =\ \inf f(X),$$

d.h. auch $x + \lambda\,(y - x)$ ist ein globales Minimum von f über X.

Da in der obigen Ungleichung also Gleichheit gilt, folgt für strikt konvexes f somit $x = y$ und damit die behauptete Eindeutigkeit.

(b) Nehmen wir an, $\hat{x}$ sei ein lokales, aber kein globales Minimum. Dann gibt es ein $y \in X$ mit $f(y) < f(\hat{x})$. Für $\lambda \in]0,1]$ folgt

$$f(\hat{x} + \lambda(y - \hat{x})) \ \leq\ f(\hat{x}) + \lambda(f(y) - f(\hat{x})) \ <\ f(\hat{x}) + \lambda(f(\hat{x}) - f(\hat{x})) \ =\ f(\hat{x}).$$

Damit gibt es in jeder Umgebung von $\hat{x}$ Punkte mit kleinerem Zielfunktionswert. Widerspruch!

Zur zweiten Aussage betrachten wir zu $y \in X$ die Hilfsfunktion

$$\Psi(\lambda) \ :=\ f(\hat{x}) + \lambda(f(y) - f(\hat{x})) \ -\ f(\hat{x} + \lambda(y - \hat{x})) \ \geq\ 0.$$

Ψ ist stetig differenzierbar und nicht negativ auf dem Intervall $[0,1]$. Ferner gilt $\Psi(0) = 0$. Hiermit folgt

$$\Psi'(0) \ =\ (f(y) - f(\hat{x})) \ -\ \nabla f(\hat{x})^\top (y - \hat{x}) \ \geq\ 0.$$

Gilt also $\nabla f(\hat{x}) = 0$, so ist $f(y) - f(\hat{x}) \geq 0$ für alle $y \in X$. Damit ist gezeigt, dass $\hat{x}$ ein globales Minimum von f über X ist.

□

Die Definition der Konvexität einer Funktion f eignet sich meist nicht so gut zur konkreten Überprüfung dieser Eigenschaft. Da f jedoch meist auch gewisse Glattheitseigenschaften besitzt, lassen sich die folgenden Charakterisierungen der Konvexität verwenden.

Satz 3.3.5

(a) Für $f \in C^1(\mathbb{R}^n)$ *und* $X \subseteq \mathbb{R}^n$ *konvex und nichtleer gilt:*

$$f \text{ konvex auf } X \iff \forall\, x, y \in X:\ f(y) \geq f(x) + \nabla f(x)^\top (y - x)$$

Ferner ist f *genau dann strikt konvex, wenn die obige Ungleichung für* $x \neq y$ *strikt erfüllt ist.*

(b) Für $f \in C^2(\mathbb{R}^n)$ *und* $X \subseteq \mathbb{R}^n$ *offen, konvex und nichtleer gilt:*

$$f \text{ konvex auf } X \iff \forall\, x \in X:\ \nabla^2 f(x) \text{ positiv semidefinit}$$

Ferner: Ist die Hesse-Matrix $\nabla^2 f(x)$ *sogar positiv definit auf* X*, so ist* f *strikt konvex.*

Beweis:

(a) (i) $\Rightarrow$: Der Beweis erfolgt wie der zu Satz 3.3.4 (b). Die Funktion

$$\Psi(\lambda) := f(x) + \lambda(f(y) - f(x)) - f(x + \lambda(y - x)) \geq 0,\ \ \lambda \in [0, 1],$$

ist stetig differenzierbar mit $\Psi(0) = 0$. Daher ist auch

$$\Psi'(0) = f(y) - f(x) - \nabla f(x)^\top (y - x) \geq 0.$$

(i) $\Leftarrow$: Mit $\bar{x} := x + \lambda(y - x)$ gelten nach Voraussetzung

$$\begin{aligned} f(x) &\geq f(\bar{x}) + \nabla f(\bar{x})^\top (x - \bar{x}) \\ f(y) &\geq f(\bar{x}) + \nabla f(\bar{x})^\top (y - \bar{x}) \end{aligned}$$

Multipliziert man die erste Ungleichung mit $(1 - \lambda)$, die zweite mit λ und addiert, so ergibt sich die Behauptung

$$(1 - \lambda)\, f(x) + \lambda\, f(y) \geq f(\bar{x}) + 0 = f(x + \lambda(y - x)).$$

(ii) $\Rightarrow$: Für $x \neq y$ sei $z := (1/2)(x + y) = x + (1/2)(y - x)$. Mit (i) und der strikten Konvexität folgt dann:

$$\begin{aligned} \nabla f(x)^\top (y - x) &= 2\, \nabla f(x)^\top (z - x) \leq 2\, (f(z) - f(x)) \\ &< 2\, (f(x) + (1/2)(f(y) - f(x)) - f(x)) = f(y) - f(x) \end{aligned}$$

(ii) $\Leftarrow$: Ganz analog zum Beweis von (i) $\Leftarrow$.

(b) (i) $\Rightarrow$: Für $x, y \in X$ gelten nach (a) die folgenden Ungleichungen

$$\begin{aligned} f(x) - f(y) &\geq \nabla f(y)^\top (x-y) \\ f(y) - f(x) &\geq \nabla f(x)^\top (y-x) \end{aligned}$$

Addition dieser Ungleichungen ergibt $(\nabla f(y) - \nabla f(x))^\top (y-x) \geq 0$, d.h. der Gradient ∇f ist *monoton* auf X.

Nun ist ∇f stetig differenzierbar. Daher folgt für $d \in \mathbb{R}^n$

$$\begin{aligned} d^\top \nabla^2 f(x)\, d &= d^\top \lim_{t \downarrow 0} \tfrac{\nabla f(x+td) - \nabla f(x)}{t} \\ &= \lim_{t \downarrow 0} \tfrac{(\nabla f(x+td) - \nabla f(x))^T (t\, d)}{t^2} \geq 0, \end{aligned}$$

also die positive Semidefinitheit der Hesse-Matrix. Man beachte, dass mit $x \in X$ (offen!) auch $x + td \in X$ gilt für alle hinreichend kleinen $t > 0$.

(i) $\Leftarrow$: Nach dem Taylorschen Satz gilt für $x, y \in X$:

$$f(y) = f(x) + \nabla f(x)^\top (y-x) + (1/2)(y-x)^\top \nabla^2 f(x + \theta(y-x))\,(y-x)$$

mit einem (von x, y abhängigen) $\theta \in]0,1[$.

Ist die Hesse-Matrix nun positiv semidefinit, so folgt

$$f(y) \geq f(x) + \nabla f(x)^\top (y-x)$$

und damit die Konvexität von f über X nach (a).

(ii): Aus der positiven Definitheit der Hesse-Matrix folgt analog für $x \neq y$:

$$f(y) > f(x) + \nabla f(x)^\top (y-x)$$

und damit die strikte Konvexität von f über X.

□

Beispiel 3.3.6

Eine quadratische Funktion $f : \mathbb{R}^n \to \mathbb{R}$ *hat die Form*

$$f(x) = \frac{1}{2}\, x^\top A\, x + b^\top x + c$$

wobei $A \in \mathbb{R}^{n \times n}$ *eine* symmetrische *Matrix ist,* $b \in \mathbb{R}^n$, $c \in \mathbb{R}$.
f *ist offenbar eine* C^∞*-Funktion mit*

$$\nabla f(x) = A\, x + b, \qquad \nabla^2 f(x) = A.$$

Damit ist f genau dann konvex, wenn A positiv semidefinit ist. Ist A sogar positiv definit, so ist f strikt konvex und besitzt ein eindeutig bestimmtes striktes, globales Minimum, das sich über das lineare Gleichungssystem

$$\nabla f(\hat{x}) \;=\; A\,\hat{x} \,+\, b \;=\; 0$$

berechnen lässt.

3.4 Das Verfahren von Nelder und Mead

Wir folgen der Darstellung von W. Alt [Alt02], welche auf dem Originalartikel [NM65] basiert.

In der Praxis erfreut sich das Verfahren von Nelder und Mead großer Beliebtheit, da es keinerlei Voraussetzungen an die zu minimierende Funktion f stellt. Insbesondere werden zur Durchführung des Algorithmus lediglich Funktionswerte von f benötigt. Allerdings gibt es auch keine allgemeinen Konvergenzaussagen und keine Garantie, daß das Verfahren tatsächlich ein (zumindest lokales) Minimium von f liefert. Für spezielle Funktionen und niedrige Raumdimensionen werden in [LRWW98] Konvergenzresultate erzielt. Jedoch sind diese Resultate nicht allgemein gültig, da in [McK98] Beispiele konstruiert werden, für die das Verfahren gegen nichtstationäre Punkte konvergiert. Es gibt neuerdings jedoch auch eine konvergente Modifikation des Verfahrens, vgl. [PCB02].

Das Verfahren basiert auf der Konstruktion von Simplizes.

Definition 3.4.1 **(Simplex)**
Seien $x^0, x^1, \ldots, x^n \in \mathbb{R}^n$ gegebene Vektoren, wobei die Vektoren $x^i - x^0$, $i = 1, \ldots, n$ linear unabhängig seien. Die konvexe Hülle

$$S = \left\{ x = \sum_{i=0}^{n} \lambda_i x^i \mid \lambda_i \geq 0,\ i = 0, 1, \ldots, n,\ \sum_{i=0}^{n} \lambda_i = 1 \right\}$$

dieser $n+1$ Punkte im $\mathbb{R}^n$ heißt **(n-dimensionales) Simplex mit Ecken** $x^0, x^1, \ldots, x^n$.

Zum Start des Verfahrens wird ein Simplex S_0 vorgegeben. Jeder Iterationsschritt des Verfahrens besteht aus folgenden Aktionen:

- Bestimme zum aktuellen Simplex S_k mit den Ecken $x^0, x^1, \ldots, x^n$ die Ecke x^m mit dem größten Funktionswert:

$$f(x^m) = \max\{f(x^0), f(x^1), \ldots, f(x^n)\}.$$

- Berechne einen Punkt y mit einem kleineren Funktionswert $f(y) < f(x^m)$ und ersetze x^m durch den neuen Punkt y. Damit erhält man einen neuen Simplex S_{k+1} zu den Punkten y und x^i, $i = 0, 1, \ldots, n$, $i \neq m$.

Für $j = 0, 1, \ldots, n$ sind die **Schwerpunkte der Ecken bzgl.** x^j durch

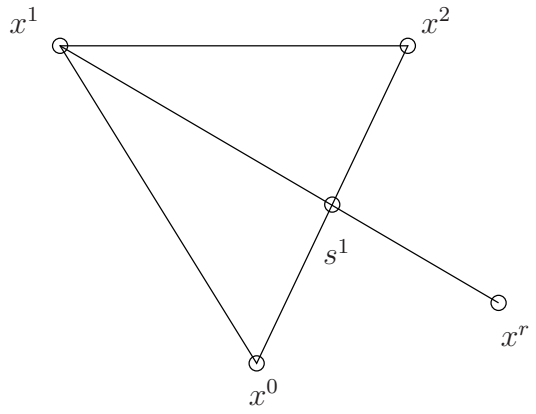
$$s^j = \frac{1}{n} \sum_{\substack{i=0 \\ i \neq j}}^{n} x^i$$

gegeben.

Zur Bestimmung des neuen Punktes y werden drei Konstruktionsprinzipien verwendet:

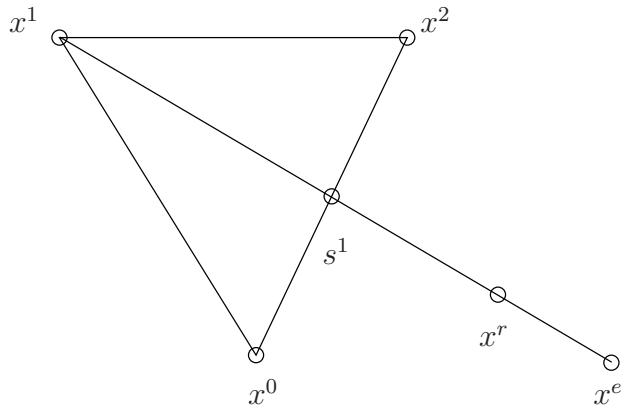
(i) **Reflektion**
Ein neuer Punkt x^r wird durch Reflektion der Ecke x^j am Schwerpunkt s^j bestimmt:

$$x^r = s^j + \gamma(s^j - x^j), \qquad 0 < \gamma \leq 1.$$



(ii) **Expansion**
Der Punkt x^e wird über x^r hinaus weiter nach aussen in Richtung $s^j - x^j$ bzw. $x^r - s^j$ verschoben:

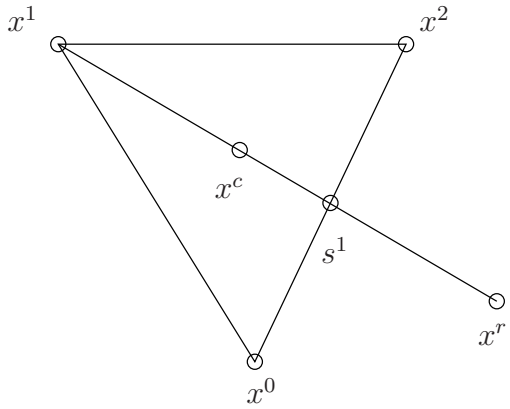
$$x^e = s^j + \beta(x^r - s^j), \qquad \beta > 1.$$



(iii) **Kontraktion**

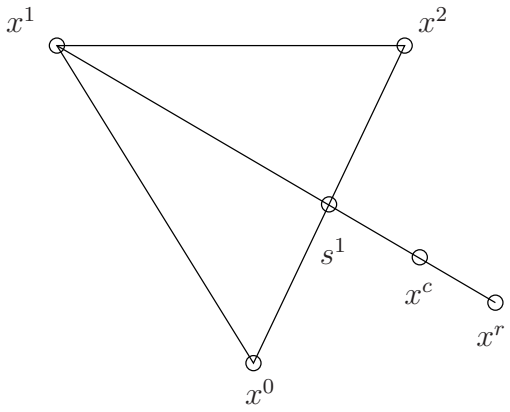
– **Innere partielle Kontraktion**
Der Punkt x^c wird zwischen x^j und s^j verschoben:

$$x^c = s^j + \alpha(x^j - s^j), \qquad 0 < \alpha < 1.$$



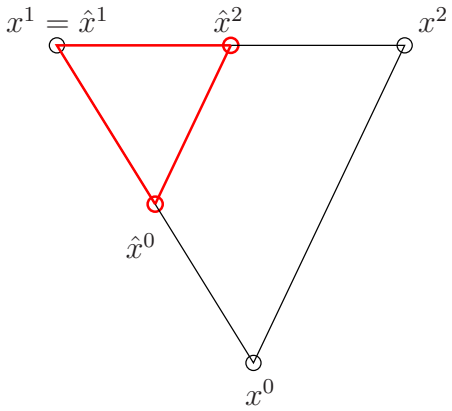
– **Äussere partielle Kontraktion**
Der Punkt x^c wird zwischen s^j und x^r verschoben:

$$x^c = s^j + \alpha(x^r - s^j), \qquad 0 < \alpha < 1.$$



– **Totale Kontraktion**
Die Punkte x^i, $i = 0, 1, \ldots, n$ mit $i \neq j$ werden durch die Mittelpunkte der Strecken von x^j nach x^i ersetzt:

$$\hat{x}^i = x^i + \frac{1}{2}(x^j - x^i) = \frac{1}{2}(x^j + x^i)$$



Basierend auf diesen Konstruktionsprinzipien lautet das
Verfahren von Nelder und Mead:

(0) Gegeben seien Parameter $0 < \alpha < 1$, $\beta > 1$ und $0 < \gamma \leq 1$, sowie ein Startpunkt $x^{(0,0)} \in \mathbb{R}^n$. Setze $k = 0$.

(i) Bestimme die Eckpunkte des Startsimplex S_0:

$$x^{(0,i)} = x^{(0,0)} + e_i, \qquad i = 1, 2, \ldots, n.$$

(e_i bezeichnet den i-ten Einheitsvektor)

(ii) Bestimme einen Punkt $x^{(k,m)}$ mit

$$f(x^{(k,m)}) = \max\{f(x^{(k,0)}), \ldots, f(x^{(k,n)})\},$$

einen Punkt $x^{(k,l)}$ mit

$$f(x^{(k,l)}) = \min\{f(x^{(k,0)}), \dots, f(x^{(k,n)})\},$$

sowie den Schwerpunkt bzgl. $x^{(k,m)}$

$$s^{(k,m)} = \frac{1}{n} \sum_{\substack{i=0 \\ i \neq m}}^{n} x^{(k,i)}.$$

(iii) Berechne den Punkt

$$x^r = s^{(k,m)} + \gamma(s^{(k,m)} - x^{(k,m)})$$

durch Reflektion von $x^{(k,m)}$ an $s^{(k,m)}$.

(iv) (a) Falls $f(x^r) < f(x^{(k,l)})$ gilt, ist x^r neuer minimaler Punkt. Versuche durch Expansion einen noch besseren Wert zu erhalten. Berechne

$$x^e = s^{(k,m)} + \beta(x^r - s^{(k,m)})$$

und ersetze $x^{(k,m)}$ durch den besten der beiden Punkte x^r und x^e:

$$x^{(k+1,m)} = \begin{cases} x^e, & \text{falls } f(x^e) < f(x^r), \\ x^r, & \text{falls } f(x^r) \leq f(x^e). \end{cases}$$

(b) Falls

$$f(x^{(k,l)}) \leq f(x^r) \leq \max\{f(x^{(k,j)}) \mid j \neq m\}$$

gilt, setze $x^{(k+1,m)} = x^r$. Hier ist x^r höchstens schlechter als $x^{(k,l)}$ und in der Regel besser als $x^{(k,m)}$.

(c) Falls

$$f(x^r) > \max\{f(x^{(k,j)}) \mid j \neq m\}$$

gilt, unterscheide folgende Fälle:

I. Falls $f(x^r) > f(x^{(k,m)})$ gilt, so ist x^r eine Verschlechterung und es ist möglicherweise ratsam, den Simplex nicht zu verlassen. Führe innere partielle Kontraktion durch und berechne

$$x^c = s^{(k,m)} + \alpha(x^{(k,m)} - s^{(k,m)}).$$

II. Falls $f(x^r) < f(x^{(k,m)})$ gilt, so ist x^r zumindest besser als $x^{(k,m)}$, allerdings schlechter als alle anderen Eckpunkte. Es ist möglicherweise ratsam, näher am Simplex zu suchen. Führe äussere partielle Kontraktion durch und berechne

$$x^c = s^{(k,m)} + \alpha(x^r - s^{(k,m)}).$$

Ist $f(x^c) < f(x^{(k,m)})$, setze $x^{(k+1,m)} = x^c$. Andernfalls haben wir trotz aller Versuche keine Verbesserung erreichen können und führen daher eine totale Kontraktion bzgl. des momentan besten Punktes $x^{(k,l)}$ durch und setzen

$$x^{(k+1,i)} = \frac{1}{2}(x^{(k,i)} + x^{(k,l)}), \qquad i \neq l.$$

(v) Setze $k = k + 1$ und gehe zu (ii).

Abbruchkriterien:

- Nelder und Mead schlagen folgendes Abbruchkriterium vor: Die Standardabweichung der Funktionswerte an den Simplexecken soll kleiner als eine vorgegebene Toleranz sein, d.h.

$$\left(\frac{1}{n+1} \sum_{i=0}^{n} \left(f(x^{(k,i)}) - \bar{f}_k \right)^2 \right)^{1/2} < tol,$$

  wobei

$$\bar{f}_k = \frac{1}{n+1} \sum_{j=0}^{n} f(x^{(k,j)})$$

  den Mittelwert der Funktionswerte an den Simplexecken bezeichnet. Dieses Kriterium wird auch in der NAG-Version `E04CCF` des Verfahrens verwendet.

- In der MATLAB-Implementation `fminsearch` des Verfahrens wird das Verfahren abgebrochen, sobald der Durchmesser des Simplex S_k kleiner als eine gegebene Toleranz ist.

Bemerkung 3.4.2

Der Punkt $x^{(k,l)}$ kann als aktuelle Iterierte betrachtet werden, da dieser Punkt den minimalen Funktionswert unter allen Ecken des aktuellen Simplex S_k liefert. Per Konstruktion gilt zumindest

$$f(x^{(k+1,l_{k+1})}) \leq f(x^{(k,l_k)}), \qquad k = 0, 1, \ldots,$$

wobei l_k den minimalen Funktionswert in Iteration k bezeichnet.
Für die Konstanten haben sich in zahlreichen Anwendungen die Werte $0.4 \leq \alpha \leq 0.6$, $2 \leq \beta \leq 3$ und $\gamma = 1$ bewährt.

Beispiel 3.4.3

Wir wollen die in Beispiel 2.2.1 betrachtete exakte l_1-Penaltyfunktion mit $\alpha = 100$ minimieren, d.h. unsere Zielfunktion ist durch

$$l_1(x, y; \alpha) = f(x, y) + \alpha |h(x, y)| + \alpha \max\{0, g_1(x, y)\} + \alpha \max\{0, g_2(x, y)\}$$

mit

$$\begin{aligned} f(x,y) &= (x-2)^2+(y-3)^2, \\ h(x,y) &= y+\frac{x}{2}-\frac{1}{2}, \\ g_1(x,y) &= y+2x^2-2, \\ g_2(x,y) &= x^2-y-1, \end{aligned}$$

gegeben. Die Optimallösung lautet

$$x=\frac{3}{5}, \quad y=\frac{1}{5}, \quad f(x,y)=\frac{49}{5}.$$

Das Programm `fminsearch` *der* MATLAB*-Optimization Toolbox, welches eine Implementation des Verfahrens von Nelder und Mead darstellt, findet die Optimallösung ausgehend vom Startwert* $(0,0)^\top$ *tatsächlich. Aufruf von*

```
[x,val,exitflag,output] = fminsearch(@l1penalty,[0,0],...
                    optimset('TolX',1e-12,'Display','iter'))
```

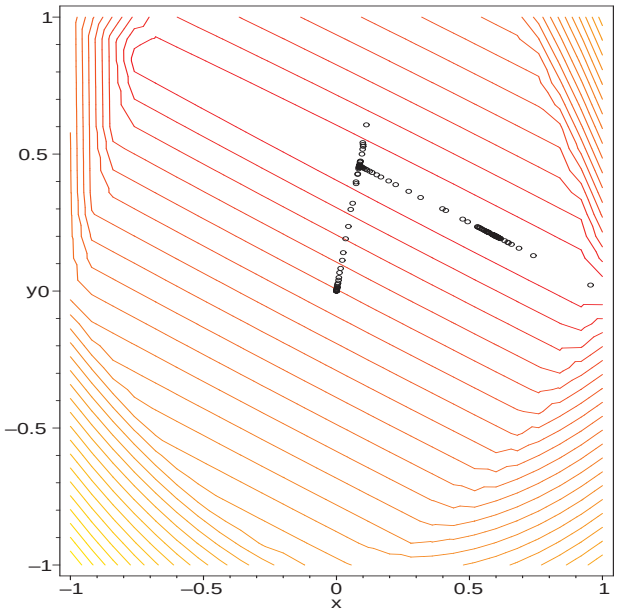
wobei die Funktion in `l1penalty.m` *gemäß*

```
function f = l1penalty(x)
  alpha=100;
  f = (x(1)-2)^2 + (x(2)-3)^2 + alpha*abs(x(2)+x(1)/2-0.5) ...
                              + alpha*max(0,x(2)+2*x(1)^2-2) ...
                              + alpha*max(0,x(1)^2-x(2)-1);
```

gegeben ist, liefert die Ausgabe:

```
Optimization terminated successfully:
 the current x satisfies the termination criteria using OPTIONS.TolX of
 1.000000e-12 and F(X) satisfies the convergence criteria using
 OPTIONS.TolFun of 1.000000e-04
x =
  0.6000    0.2000
val =
  9.8000
exitflag =
   1
output =
  iterations: 180
   funcCount: 366
   algorithm: 'Nelder-Mead simplex direct search'
```

Die Abbildung veranschaulicht diejenigen Punkte, an denen Funktionsauswertungen vorgenommen wurden:



3.5 Allgemeine Abstiegsverfahren

Wir konstruieren ein allgemeines Konzept zur Minimierung einer stetig differenzierbaren Funktion $f : \mathbb{R}^n \to \mathbb{R}$. Nahezu alle später diskutierten Verfahren basieren auf diesem Konzept.

Der folgende Algorithmus basiert auf der Verwendung von Abstiegsrichtungen. Eine Abstiegsrichtung d von f in x liegt anschaulich vor, wenn es entlang der Richtung d im Punkt x (lokal) „bergab“ geht. Eine formale Definition ist gegeben durch

Definition 3.5.1 **(Abstiegsrichtung)**

Seien $f : \mathbb{R}^n \to \mathbb{R}$ und $x \in \mathbb{R}^n$. $d \in \mathbb{R}^n$ heißt **Abstiegsrichtung von f in x,** *falls es ein $\bar{\alpha} > 0$ gibt mit*

$$f(x + \alpha d) < f(x) \qquad \forall 0 < \alpha \leq \bar{\alpha}.$$

Eine hinreichende Bedingung für eine Abstiegsrichtung liefert der folgende Hilfssatz.

Hilfssatz 3.5.2

Sei $f : \mathbb{R}^n \to \mathbb{R}$ stetig differenzierbar in x. d ist eine Abstiegsrichtung von f in x, wenn $f'(x; d) = \nabla f(x)^\top d < 0$ gilt.

Beweis: Aus

$$f'(x; d) = \lim_{\alpha \downarrow 0} \frac{f(x + \alpha d) - f(x)}{\alpha} = \nabla f(x)^\top d < 0$$

folgt die Existenz eines $\bar{\alpha} > 0$ mit $f(x + \alpha d) - f(x) < 0$ für alle $0 < \alpha \leq \bar{\alpha}$, d.h. d ist Abstiegsrichtung. □

Die Bedingung $\nabla f(x)^\top d < 0$ bedeutet geometrisch, daß der Winkel zwischen Gradient und Abstiegsrichtung zwischen 90° und 270° liegt. Dies läßt sich aus der für Vektoren $a, b \in \mathbb{R}^n$ allgemein gültigen Beziehung

$$\cos \angle (a, b) = \frac{a^\top b}{\|a\| \cdot \|b\|} = \frac{\langle a, b \rangle}{\|a\| \cdot \|b\|}$$

ableiten.

Bemerkung 3.5.3

Die Bedingung in Hilfssatz 3.5.2 ist nur hinreichend für eine Abstiegsrichtung aber nicht notwendig. Betrachte z.B. ein striktes lokales Maximum $\hat{x}$ mit $\nabla f(\hat{x}) = 0$. Für alle Richtungen $d \in \mathbb{R}^n$ gilt dann $f'(\hat{x}; d) = \nabla f(\hat{x})^\top d = 0$. Andererseits ist in einem strikten lokalen Maximum jede Richtung $d \in \mathbb{R}^n$ eine Abstiegsrichtung.

Mit diesen Begriffen lautet das allgemeine **Abstiegsverfahren mit Schrittweitensteuerung** zur Minimierung von f:

Algorithmus: Abstiegsverfahren

(i) Bestimme einen Startpunkt $x^{[0]} \in \mathbb{R}^n$ und setze $i = 0$.

(ii) Falls ein Abbruchkriterium erfüllt ist, STOP.

(iii) Berechne eine Abstiegsrichtung $d^{[i]}$ und eine Schrittweite $\alpha_i > 0$, so daß

$$f(x^{[i]} + \alpha_i \cdot d^{[i]}) < f(x^{[i]})$$

gilt und setze $x^{[i+1]} = x^{[i]} + \alpha_i \cdot d^{[i]}$.

(iv) Setze $i := i + 1$ und gehe zu (ii).

Natürlich ist dieser Algorithmus lediglich von konzeptioneller Art, da die wesentlichen Komponenten (Bestimmung der Abstiegsrichtung, der Schrittweite und geeigneter Abbruchkriterien) noch nicht näher beschrieben wurden und sehr viel Freiraum lassen.
Im Vorgriff auf später seien einige Beispiele für Suchrichtungen genannt:

- Beim **Gradientenverfahren** oder **Verfahren des steilsten Abstiegs** wird die Richtung $d^{[i]} := -\nabla f(x^{[i]})$ gewählt. Bekanntlich zeigt der Gradient einer Funktion in Richtung des steilsten Anstiegs und somit zeigt der negative Gradient in Richtung des steilsten Abstiegs. Diese naheliegende Wahl der Suchrichtung muss aber nicht die beste sein. Wegen $\nabla f(x^{[i]})^\top d^{[i]} = -\|\nabla f(x^{[i]})\|^2 < 0$ ist $d^{[i]}$ eine Abstiegsrichtung, falls $\nabla f(x^{[i]}) \neq 0$ gilt.

- Beim **Newtonverfahren** wird die Richtung $d^{[i]} := -\left(\nabla^2 f(x^{[i]})\right)^{-1} \cdot \nabla f(x^{[i]})$ verwendet. Hierbei wird allerdings die Hessematrix $\nabla^2 f$ von f benötigt. Ist diese positiv definit, so ist $d^{[i]}$ eine Abstiegsrichtung, falls noch $\nabla f(x^{[i]}) \neq 0$ gilt.

- Beim **Quasi-Newtonverfahren** wird die Richtung $d^{[i]} := -B_i \cdot \nabla f(x^{[i]})$ verwendet. Hierbei ist B_i eine geeignete positiv definite Matrix. Wegen

$$\nabla f(x^{[i]})^\top d^{[i]} = -\nabla f(x^{[i]})^\top B_i \nabla f(x^{[i]}) < 0$$

ist $d^{[i]}$ eine Abstiegsrichtung, falls $\nabla f(x^{[i]}) \neq 0$ gilt.

Satz 3.5.4

Ist $B \in \mathbb{R}^{n\times n}$ symmetrisch und positiv definit, so minimiert die (skalierte und normierte) Gradientenrichtung

$$d := -\frac{B^{-1}\nabla f(x)}{\|B^{-1}\nabla f(x)\|_B}$$

den Anstieg $\nabla f(x)^\top d$ über alle Richtungen $d \in \mathbb{R}^n$ mit $\|d\|_B = 1$.
Dabei sei $\nabla f(x) \neq 0$ und die (skalierte) Norm gegeben durch $\|x\|_B := (x^\top B\, x)^{1/2}$.

Beweis: (für $B = I_n$)
Mit der Cauchy-Schwarzschen Ungleichung folgt für $\|d\|_2 = 1$:

$$|\nabla f(x)^\top d| \;\leq\; \|\nabla f(x)\|_2 \, \|d\|_2 \;=\; \|\nabla f(x)\|_2.$$

Diese Schranke wird gerade für $d = \pm\nabla f(x)/\|\nabla f(x)\|_2$ angenommen. □

Beispiel 3.5.5
Wir verwenden das (unskalierte) Gradientenverfahren zur Minimierung der Funktion

$$f(x_1, x_2) := x_1^2 + 10\, x_2^2.$$

Die Bestimmung der Schrittweite erfolge mit exakter Liniensuche, d.h.

$$\alpha := \mathit{argmin}\{f(x + \alpha\, d) : \alpha > 0\}.$$

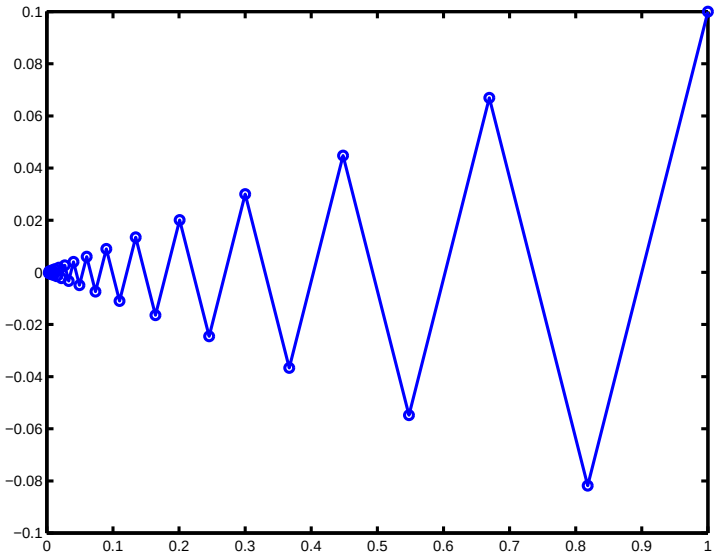
Wir erhalten

$$\begin{aligned}
\nabla f(x_1, x_2) &= (2x_1, 20x_2)^\top \\
d &= -(2x_1, 20x_2)^\top \\
\varphi(\alpha) &= (x_1 + \alpha d_1)^2 \;+\; 10\,(x_2 + \alpha d_2)^2 \\
\varphi'(\alpha) &= 2\,(x_1 + \alpha d_1)\, d_1 \;+\; 20\,(x_2 + \alpha d_2)\, d_2 \;=\; 0(!)
\end{aligned}$$

Damit:

$$\alpha \;:=\; -\,\frac{x_1 d_1 + 10\, x_2 d_2}{d_1^2 + 10 d_2^2}, \quad x_i^{neu} := x_i + \alpha\, d_i,\ (i = 1, 2).$$

Mit dem Startvektor $x^{[0]} := (1, 0.1)^\top$ benötigt das Verfahren 63 Iterationen um das Abbruchkriterium $\|\nabla f(x)\|_2 \leq 10^{-5}$ zu erfüllen.



Würden wir dagegen die in Satz 3.5.4 angegebene skalierte Gradientenrichtung verwenden mit einer geschickt gewählten Matrix

$$B = \begin{pmatrix} 1 & 0 \\ 0 & 10 \end{pmatrix},$$

so würde sich ergeben

$$d = -B^{-1}\nabla f(x) = -\begin{pmatrix} 1 & 0 \\ 0 & 1/10 \end{pmatrix}\begin{pmatrix} 2\,x_1 \\ 20\,x_2 \end{pmatrix} = -2\,x,$$

sowie (bei exakter Liniensuche) $\alpha = 0.5$, so dass der erste Iterationsschritt bereits die exakte Lösung $\hat{x} = 0$ liefert.

Um einen Konvergenzbeweis für den obigen Algorithmus führen zu können, benötigt man Verfahren, die **effiziente Schrittweiten** bestimmen. Daß eine beliebige Wahl von Schrittweiten α_i nicht zum Erfolg führt, zeigt folgendes

Beispiel 3.5.6

$f(x) = x^2$, $d^{[i]} := -f'(x^{[i]}) = -2x^{[i]}$. Sei $x^{[0]} := 1$ und

$$\alpha_i := \frac{1}{2^{i+3}}.$$

Insgesamt erhält man

$$x^{[i+1]} = x^{[i]} - 2\alpha_i x^{[i]} = x^{[i]} - \frac{1}{2^{i+2}}x^{[i]} = x^{[i]}\left(1 - \frac{1}{2^{i+2}}\right).$$

Mit $x^{[0]} = 1 > 0$ folgt $0 < x^{[i]} < 1$ für alle i und es gilt

$$x^{[0]} - x^{[i]} = \sum_{j=0}^{i-1} x^{[j]} - x^{[j+1]} = \sum_{j=0}^{i-1} \frac{1}{2^{j+2}}x^{[j]} \le \sum_{j=0}^{i-1} \frac{1}{2^{j+2}} \le \frac{1}{2}.$$

Also ist $|x^{[0]} - x^{[i]}| = |1 - x^{[i]}| \leq 1/2$. *Somit konvergiert* $x^{[i]}$ **nicht** *gegen* 0.

Das Beispiel zeigt, daß die Schrittweiten α_i zusätzliche Eigenschaften besitzen müssen, um Konvergenz gegen stationäre Punkte beweisen zu können.
Wir verlangen, daß durch eine geeignete Schrittweitenwahl $\alpha(x,d)$ der Funktionswert $f(x+\alpha d)$ in ausreichendem Maße verkleinert wird. Genauer: Wir suchen eine Aussage darüber, um wieviel man $f(x)$ in Richtung $x+\alpha d$ (d vorgegebene Abstiegsrichtung) verkleinern kann.
Wir nehmen an, dass $f \in C^2(\mathbb{R}^n, \mathbb{R})$ und dass die Niveaumenge $\mathrm{lev}(f, f(x^{[0]}))$ zum Startvektor $x^{[0]}$ kompakt ist. $x = x^{[k]}$ sei eine Iterierte aus $\mathrm{lev}(f, f(x^{[0]}))$. Da die Hesse-Matrix $\nabla^2 f(x)$ stetig ist, gibt es eine Konstante $C > 0$ mit

$$\|\nabla^2 f(x)\|_2 \;\leq\; C, \quad \forall\, x \in \mathrm{lev}(f, f(x^{[0]})).$$

Mittels Taylor-Entwicklung folgt

$$\begin{aligned} f(x+\alpha d) &= f(x) \,+\, \alpha\, \nabla f(x)^\top d \,+\, (\alpha^2/2)\, d^\top \nabla^2 f(z)\, d \\ &\leq f(x) \,+\, \alpha\, \nabla f(x)^\top d \,+\, (\alpha^2/2)\, C\, \|d\|_2^2 \,. \end{aligned}$$

Dabei ist $z = x + \Theta\,\alpha\, d$ eine Zwischenstelle, $0 < \Theta < 1$. Die Abschätzung gilt für alle $\alpha > 0$ mit $x + [0,\alpha]\, d \subseteq \mathrm{lev}(f, f(x^{[0]}))$.
Nun ist die rechte Seite dieser Abschätzung ein Polynom $p(\alpha)$ zweiten Grades in α, das in

$$\alpha^* \;:=\; -\frac{\nabla f(x)^\top d}{C\, \|d\|_2^2} \;>\; 0$$

ein striktes globales Minimum besitzt.
Ist $\hat\alpha$ die (eindeutig bestimmte) maximale Schrittweite mit $\forall \alpha \in [0,\hat\alpha] : \; x + \alpha\, d \in \mathrm{lev}(f, f(x^{[0]}))$, so folgt

$$p(\hat\alpha) \;\geq\; f(x + \hat\alpha\, d) \;=\; f(x^{[0]}) \;\geq\; f(x) \;=\; p(0).$$

Wegen $p'(0) < 0$ folgt hieraus, dass α^* im offenen Intervall $]0,\hat\alpha[$ liegt, insbesondere also zulässig ist. Mit der Abschätzung folgt damit

$$f(x+\alpha^* d) \;\leq\; p(\alpha^*) \;=\; f(x) \,-\, \frac{1}{2\,C} \left(\frac{\nabla f(x)^\top d}{\|d\|} \right)^2 .$$

Diese Abschätzung zeigt, dass ein Mindestabstieg der obigen Gestalt unter den genannten Voraussetzungen möglich ist. Da man aber andererseits die Schranke C für die Hesse-Matrix i.a. nicht kennt, definieren wir etwas vorsichtiger:

Definition 3.5.7 **(Effiziente Schrittweiten)**
Eine Schrittweitenstrategie (also eine Abbildung, die zu x *und* d *eine oder mehrere Schrittweiten* $\alpha = \alpha(x,d) > 0$ *bestimmt) heißt* **effizient**, *wenn es zu* $x^{[0]} \in \mathbb{R}^n$ *ein* $C > 0$ *gibt*

mit

$$f(x+\alpha d) \le f(x) - C\left(\frac{\nabla f(x)^\top d}{\|d\|}\right)^2$$

für alle $x \in lev(f, f(x^{[0]}))$ *und* $d \in \mathbb{R}^n$ *mit* $\nabla f(x)^\top d < 0$ *und alle von der Schrittweitenstrategie gelieferten Schrittweiten.*

Eine effiziente Schrittweite führt also zu einer hinreichend großen Abnahme des Funktionswerts beim Übergang von $x^{[i]}$ zu $x^{[i+1]}$.

Bemerkung 3.5.8

Unter den obigen Voraussetzungen ($f \in C^2$, Niveaumenge kompakt) ist die exakte Schrittweitenstrategie $\alpha := argmin\{f(x+\alpha\, d) : \alpha > 0\}$ *effizient.*

Selbst effiziente Schrittweiten können die Konvergenz noch nicht erzwingen, denn es könnte noch der Fall $\nabla f(x^{[i]})^\top d^{[i]} = 0$ (die Suchrichtung verläuft tangential zu den Höhenlinien von f) bzw. $\nabla f(x^{[i]})^\top d^{[i]} \to 0$ mit $d^{[i]} \neq 0$ eintreten, so daß kein bzw. ein zu geringer Abstieg erfolgt. Um dieses unerwünschte Verhalten auszuschließen, werden nur Suchrichtungen $d^{[i]}$ verwendet, die der folgenden Winkelbedingung genügen:

Definition 3.5.9 (Winkelbedingung)

Sei $f : \mathbb{R}^n \to \mathbb{R}$ *stetig differenzierbar. Die Folgen* $\{x^{[i]}\}$ *und* $\{d^{[i]}\}$ *erfüllen die* **Winkelbedingung**, *falls es eine Konstante* $c_2 > 0$ *gibt mit*

$$-\frac{\nabla f(x^{[i]})^\top d^{[i]}}{\|\nabla f(x^{[i]})\| \cdot \|d^{[i]}\|} \ge c_2$$

für alle i.

Anschaulich besagt die Winkelbedingung, daß der Winkel zwischen $-\nabla f(x^{[i]})$ und $d^{[i]}$ immer zwischen 90° und −90° liegt und dabei **gleichmäßig** (also unabhängig von i) von 90° und −90° wegbleibt, vgl. Abbildung 3.1. Insbesondere ist $d^{[i]}$ dann auch Abstiegsrichtung.

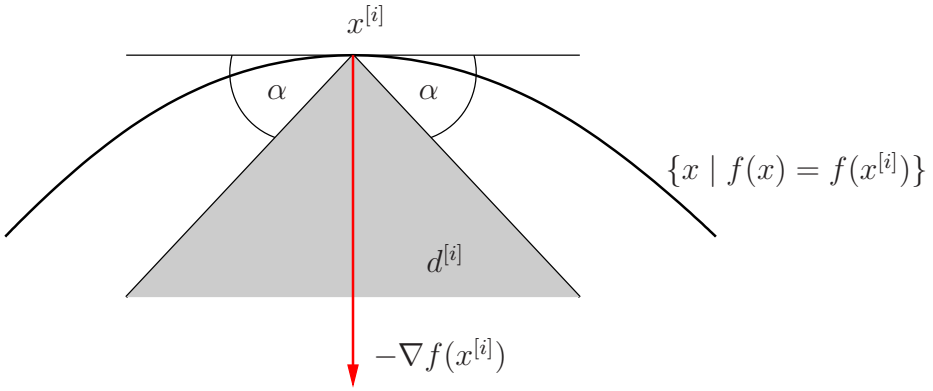


Abbildung 3.1: Winkelbedingung für die Suchrichtungen.

Mit diesen Hilfsmitteln kann man folgenden Konvergenzsatz beweisen, wobei wir implizit davon ausgehen, daß das Abstiegsverfahren nicht abbricht:

Satz 3.5.10 (Konvergenzsatz für das allgemeine Abstiegsverfahren)
Sei $f : \mathbb{R}^n \to \mathbb{R}$ stetig differenzierbar und die durch den Algorithmus erzeugten Folgen $\{x^{[i]}\}$ und $\{d^{[i]}\}$ erfüllen die Winkelbedingung. Die Schrittweiten $\alpha_i > 0$ seien effizient für alle i. Dann ist jeder Häufungspunkt der Folge $\{x^{[i]}\}$ ein stationärer Punkt von f.

Beweis: Da jedes $\alpha_i > 0$ effizient ist, existiert eine Konstante $c_1 > 0$ mit

$$f(x^{[i+1]}) = f(x^{[i]} + \alpha_i d^{[i]}) \leq f(x^{[i]}) - c_1 \left(\frac{\nabla f(x^{[i]})^\top d^{[i]}}{\|d^{[i]}\|} \right)^2$$

für alle i. Aus der Winkelbedingung folgt somit

$$f(x^{[i+1]}) \leq f(x^{[i]}) - \kappa \|\nabla f(x^{[i]})\|^2 < f(x^{[i]}) \tag{3.2}$$

mit $\kappa = c_1 \cdot c_2^2 > 0$. Sei nun x^* ein Häufungspunkt der Folge $\{x^{[i]}\}$. Da $\{f(x^{[i]})\}$ monoton fällt und für eine Teilfolge gegen $f(x^*)$ konvergiert, konvergiert bereits die gesamte Folge $\{f(x^{[i]})\}$ gegen $f(x^*)$. Insbesondere folgt

$$f(x^{[i+1]}) - f(x^{[i]}) \to 0.$$

Wegen (3.2) folgt $\|\nabla f(x^{[i]})\| \to 0$. Jeder Häufungspunkt von $\{x^{[i]}\}$ ist somit ein stationärer Punkt. □

3.6 Abbruchkriterien

Wir klären die Frage nach geeigneten Abbruchkriterien für das allgemeine Abstiegsverfahren. Dazu gibt der Benutzer eine relative Genauigkeitsschranke $\varepsilon \approx 10^{-r}$ vor, wobei r die Anzahl der gewünschten gültigen Dezimalstellen der numerischen Lösung sei. Da numerische Rechnungen in der Regel rundungsfehlerbehaftet sind, ist es sinnlos, ε kleiner als die relative Maschinengenauigkeit ε_{mach} zu wählen. Für doppelt genaue Gleitpunktzahlen gilt $\varepsilon_{mach} = 2^{-52} \approx 2.22 \cdot 10^{-16}$.
Da die Bedingung $\nabla f(\hat{x}) = 0$ notwendig für ein Minimum ist, andererseits in der numerischen Praxis aber so gut wie niemals von den Iterierten $x^{[i]}$ erfüllt wird, ist

$$\|\nabla f(x^{[i]})\| \leq \varepsilon$$

eine sinnvolle Abbruchbedingung. Allerdings ist diese Bedingung nicht invariant bezüglich der Skalierung von f. Denn durch Multiplikation von f mit einer hinreichend kleinen positiven Zahl ist diese Bedingung nahezu immer erfüllbar. Andererseits ist es nicht sinnvoll, das Abstiegsverfahren fortzuführen, da auf Grund von Rundungsfehlern kein besseres Ergebnis zu erwarten ist. Daher sollte dieses Abbruchkriterium mit einer Warnung versehen

werden. Ebenso ist es sinnvoll, eine maximale Iterationszahl i_{max} vorzuschreiben und das Abstiegsverfahren abzubrechen, sobald

$$i \geq i_{max}$$

gilt. Dies verhindert zu lange Iterationen und sollte ebenfalls mit einer Warnung versehen werden.
Gill et al. [GMW81] schlagen folgende Abbruchkriterien vor, die nicht nur die absoluten Größen $\nabla f(x^{[i]})$ bzw. $\|x^{[i-1]} - x^{[i]}\|$ bzw. $f(x^{[i-1]}) - f(x^{[i]})$ überprüfen, sondern diese durch zusätzliche Faktoren $(1 + \ldots)$ in Relation setzen zur Größe der Funktionswerte $|f(x^{[i]})|$ bzw. der Iterierten $\|x^{[i]}\|$:

- $f(x^{[i-1]}) - f(x^{[i]}) \leq \varepsilon \cdot \big(1 + |f(x^{[i]})|\big)$
- $\|x^{[i-1]} - x^{[i]}\| \leq \sqrt{\varepsilon} \cdot \big(1 + \|x^{[i]}\|\big)$
- $\|\nabla f(x^{[i]})\| \leq \sqrt[3]{\varepsilon} \cdot \big(1 + |f(x^{[i]})|\big)$

Für sehr kleine Werte $|f(x^{[i]})|$ bzw. $\|x^{[i]}\|$ gehen diese relativen Abfragen in absolute Kriterien über. Sind diese Kriterien erfüllt, so kann das Verfahren mit Erfolg beendet werden. Die Wurzel $\sqrt{\varepsilon}$ in der zweiten Bedingung erklärt sich durch Taylorentwicklung in der Nähe des Lösungspunktes (dort gilt $\nabla f(x^{[i]}) \approx 0$):

$$f(x^{[i-1]}) \approx f(x^{[i]}) + \mathcal{O}(\|x^{[i-1]} - x^{[i]}\|^2).$$

Die dritte Wurzel in der dritten Bedingung stellt eine Abschwächung der theoretisch begründbaren zweiten Wurzel dar. Die Verwendung der zweiten Wurzel stellt sich in der Praxis als zu restriktiv heraus.

3.7 Schrittweitenstrategien

Im folgenden wird vorausgesetzt, daß wir im Punkt $x^{[i]}$ bereits eine Richtung $d^{[i]}$ mit

$$\nabla f(x^{[i]})^\top d^{[i]} < 0$$

gefunden haben. Nach Hilfssatz 3.5.2 ist $d^{[i]}$ somit eine Abstiegsrichtung von f in $x^{[i]}$. Um das allgemeine Abstiegsverfahren durchführen zu können, muß also nur noch die **Schrittweite** α_i bestimmt werden.
Zur Bestimmung der Schrittweite genügt es, für $\alpha \geq 0$ die Funktion $\varphi : \mathbb{R} \to \mathbb{R}$ mit

$$\varphi(\alpha) := f(x^{[i]} + \alpha \cdot d^{[i]})$$

zu betrachten. Aus

$$\varphi'(0) = \nabla f(x^{[i]})^\top d^{[i]} < 0$$

folgt

$$\varphi(\alpha) < \varphi(0)$$

für alle $0 < \alpha \leq \bar{\alpha}$ mit einem $\bar{\alpha} > 0$, vgl. Abbildung 3.2.

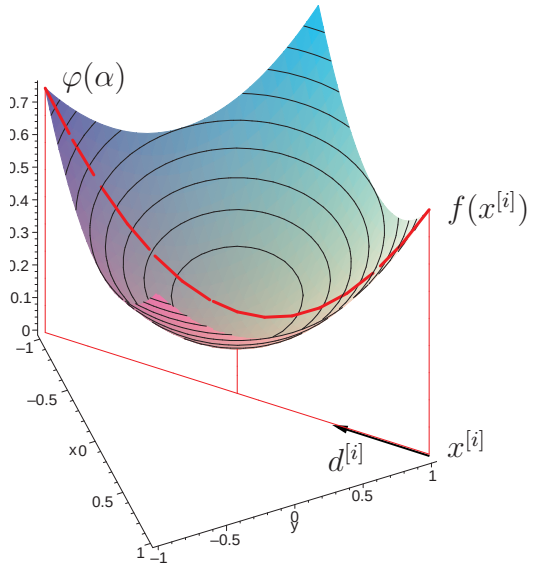


Abbildung 3.2: Eindimensionale Liniensuche ausgehend von $x^{[i]}$ in Richtung $d^{[i]}$.

Wir werden nun einige gebräuchliche Schrittweitenstrategien diskutieren, wovon einige im Hinblick auf den Konvergenzsatz 3.5.10 effizient sind.

3.7.1 Armijo-Regel

Zu vorgegebenen Zahlen $\beta \in (0,1)$ und $\sigma \in (0,1)$ bestimme

$$\alpha_i := \max\{\beta^j \mid j = 0, 1, 2, \ldots\},$$

so daß

$$\varphi(\alpha_i) \leq \varphi(0) + \sigma \cdot \alpha_i \cdot \varphi'(0) \tag{3.3}$$

gilt. Wegen $\varphi'(0) < 0$ garantiert die Armijo-Regel, daß die Zielfunktionswerte $f(x^{[i]})$, $i = 0, 1, \ldots$ streng monoton fallen. Die Realisierung der Armijo-Regel ist einfach und kann mit dem folgenden Algorithmus erfolgen:

Algorithmus: Armijo-Regel

(i) Setze $\alpha := 1$.

(ii) Falls die Bedingung

$$\varphi(\alpha) \leq \varphi(0) + \sigma \cdot \alpha \cdot \varphi'(0)$$

erfüllt ist, setze $\alpha_i := \alpha$ und beende das Verfahren.
Andernfalls gehe zu (iii).

(iii) Setze $\alpha := \beta \cdot \alpha$ und gehe zu (ii).

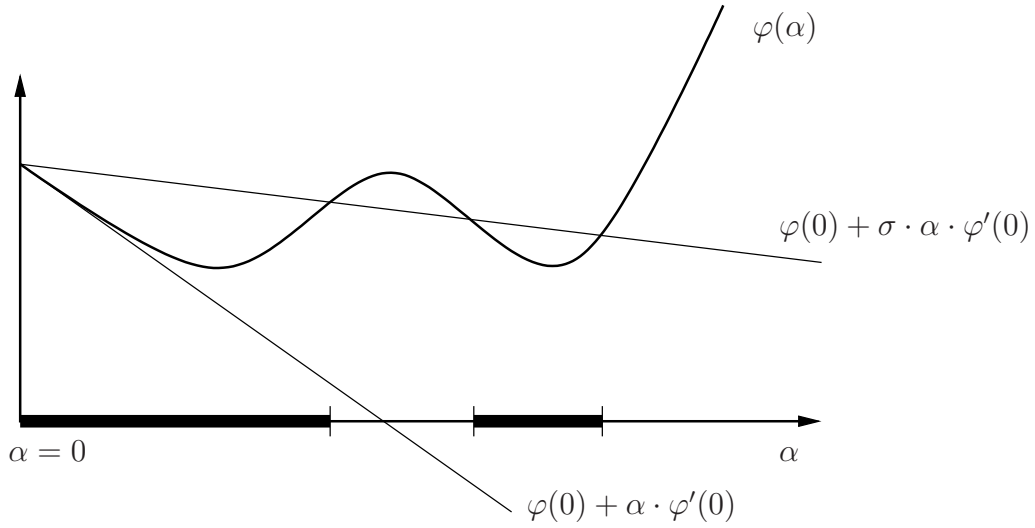


Abbildung 3.3: Schrittweiten, die der Armijoregel genügen.

Die fett eingezeichneten Intervalle in Abbildung 3.3 erfüllen die Armijo-Bedingung (3.3). Die Armijo-Regel ist wohldefiniert, denn es gilt

Satz 3.7.1

Sei f stetig differenzierbar. Zu $x, d \in \mathbb{R}^n$ mit $\nabla f(x)^\top d < 0$ und $\beta \in (0,1)$,$\sigma \in (0,1)$ existiert ein Index $j \in \mathbb{N}$ mit

$$f(x + \beta^j \cdot d) \leq f(x) + \sigma \cdot \beta^j \cdot \nabla f(x)^\top d.$$

Beweis: Annahme: Für alle $j \in \mathbb{N}$ gilt

$$f(x + \beta^j \cdot d) > f(x) + \sigma \cdot \beta^j \cdot \nabla f(x)^\top d.$$

Dann ist auch

$$\frac{f(x+\beta^j \cdot d) - f(x)}{\beta^j} > \sigma \cdot \nabla f(x)^\top d.$$

Für $j \to \infty$ folgt wegen der Differenzierbarkeit von f

$$\nabla f(x)^\top d \geq \sigma \cdot \nabla f(x)^\top d.$$

Wegen $\sigma \in (0,1)$ folgt $\nabla f(x)^\top d \geq 0$ im Widerspruch zu $\nabla f(x)^\top d < 0$. □

Bemerkung 3.7.2

(i) *In der Praxis liefern die Werte $\sigma = 0.01$ und $\beta = 0.9$ häufig gute Ergebnisse. Genauere Untersuchungen finden sich bei Spellucci [Spe93] und Alt [Alt02].*

(ii) *Die Armijo-Regel ist i.a. nicht effizient.*

(iii) *Es gibt eine Variante der Armijo-Regel, die als skalierte Armijo-Regel bezeichnet wird. Für ein $s > 0$ wird die Schrittweite hierbei durch*

$$\alpha_i := \max\{s \cdot \beta^j \mid j = 0, 1, 2, \ldots\}$$

mit

$$\varphi(\alpha_i) \leq \varphi(0) + \sigma \cdot \alpha_i \cdot \varphi'(0)$$

festgelegt. Satz 3.7.1 läßt sich hierfür analog beweisen. Bei geeigneter Wahl des Skalierungsparameters s ist diese Variante sogar effizient, siehe Geiger und Kanzow [GK99], Aufgabe 5.3, S. 42.

(iv) *Häufig kann man beobachten, daß die Armijobedingung gegen Ende des Abstiegsalgorithmus bereits im ersten Versuch erfüllt ist. Dies ist jedoch zu Beginn des Abstiegsalgorithmus unerwünscht, da ein zu schnelles Akzeptieren der Schrittweite oft zu zu kleinen Schrittweiten führt. Dieses Verhalten vermeidet die* **Armijoregel mit Aufweitung**. *Hier wird eine von der Armijoregel im ersten Schritt akzeptierte Schrittweite so lange mit $1/\beta$ multipliziert („aufgeweitet“) bis die Armijobedingung (3.3) nicht mehr erfüllt ist. Der zugehörige Algorithmus lautet:*

Algorithmus: Armijo mit Aufweitung

(i) Gegeben seien Zahlen $\beta \in (0,1)$ *und* $\sigma \in (0,1)$. *Setze* $\alpha := 1$.

(ii) Falls (3.3) nicht erfüllt ist, verfahre wie bei der Armijoregel.
Andernfalls gehe zu (iii).

(iii) Setze $\alpha_i := \alpha$. *Setze dann* $\alpha := \frac{\alpha}{\beta}$ *(Aufweitung) und gehe zu (iv).*

(iv) Ist (3.3) erfüllt, gehe zu (iii).
Andernfalls beende das Verfahren mit Schrittweite α_i.

Die Armijo-Regel mit Aufweitung ist wohldefiniert, falls f und damit φ nach unten beschränkt ist. Mit einer zusätzlichen Lipschitzbedingung an ∇f ist sie sogar effizient, siehe Geiger und Kanzow [GK99], Aufgabe 5.4, S. 43.

3.7.2 Wolfe-Powell-Regel und strenge Wolfe-Powell-Regel

Wolfe-Powell-Regel:
Zu vorgegebenen Zahlen $\sigma \in (0, 1/2)$ und $\varrho \in [\sigma, 1)$ bestimme $\alpha > 0$, so daß

$$\begin{aligned} \varphi(\alpha) &\leq \varphi(0) + \sigma \cdot \alpha \cdot \varphi'(0) && (3.4) \\ \varphi'(\alpha) &\geq \varrho \cdot \varphi'(0). && (3.5) \end{aligned}$$

gilt.

Strenge Wolfe-Powell-Regel:
Zu vorgegebenen Zahlen $\sigma \in (0, 1/2)$ und $\varrho \in [\sigma, 1)$ bestimme $\alpha > 0$, so daß

$$\begin{aligned} \varphi(\alpha) &\leq \varphi(0) + \sigma \cdot \alpha \cdot \varphi'(0) && (3.6) \\ |\varphi'(\alpha)| &\leq -\varrho \cdot \varphi'(0). && (3.7) \end{aligned}$$

gilt.

Die jeweils erste Bedingung ist aus der Armijoregel bekannt. Die Bedingung (3.5) besagt, daß φ in α nicht mehr so stark fällt, wie für $\alpha = 0$, bzw. sogar schon wieder ansteigt. Die strenge Bedingung (3.7) besagt, daß die Funktion φ nicht zu stark steigen darf. Abbildung 3.4 verdeutlicht die beiden Regeln.

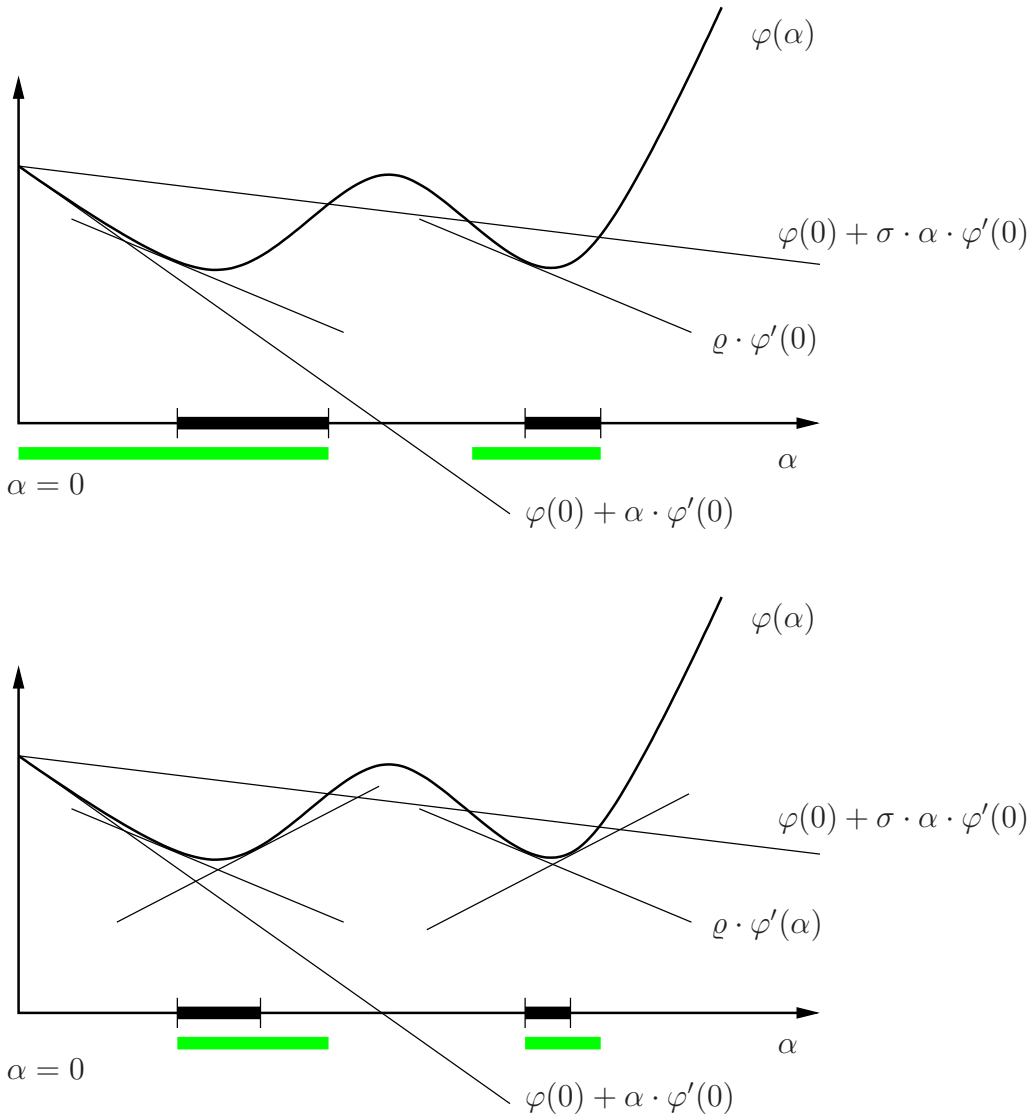


Abbildung 3.4: Die schwarzen Bereiche kennzeichnen die Schrittweiten, die der Wolfe-Powell-Regel (oben) bzw. der strengen Wolfe-Powell-Regel (unten) genügen. Die grünen Bereiche kennzeichnen die Schrittweiten, die der Armijoregel genügen.

Zunächst stellt sich die Frage nach der Wohldefiniertheit der Regeln. Da die strenge Wolfe-Powell-Regel eine Verschärfung der Wolfe-Powell-Regel darstellt, genügt es, die Wohldefiniertheit der strengen Wolfe-Powell-Regel zu zeigen. Die Wohldefiniertheit der Wolfe-Powell-Regel folgt dann automatisch.

Satz 3.7.3 **(Existenz)**

Sei $f : \mathbb{R}^n \to \mathbb{R}$ *stetig differenzierbar und nach unten beschränkt,* $x^{[0]} \in \mathbb{R}^n$ *und* $\sigma \in (0, 1/2)$, $\varrho \in [\sigma, 1)$. *Weiter seien*

$$x \in lev(f, f(x^{[0]})) = \{y \in \mathbb{R}^n \mid f(y) \leq f(x^{[0]})\}$$

und $d \in \mathbb{R}^n$ mit $\nabla f(x)^\top d < 0$. Dann gibt es Schrittweiten α, die den strengen Wolfe-Powell-Bedingungen (3.6)-(3.7) genügen.

Beweis: Sei $\varphi(\alpha) = f(x + \alpha d)$ und $\psi(\alpha) = f(x) + \sigma \cdot \alpha \cdot \nabla f(x)^\top d$. Zu zeigen ist, daß ein $\alpha > 0$ existiert mit

$$\varphi(\alpha) \le \psi(\alpha), \qquad |\varphi'(\alpha)| \le -\varrho\varphi'(0).$$

Wegen $\varphi'(0) < \psi'(0) < 0$ und der Stetigkeit von f und ∇f gilt $\varphi(\alpha) < \psi(\alpha)$ für alle hinreichend kleinen $\alpha > 0$. Da einerseits $\psi(\alpha) \to -\infty$ gilt und andererseits f und somit auch φ nach unten beschränkt sind, gibt es nach dem Zwischenwertsatz (mindestens) ein $\alpha > 0$ mit $\varphi(\alpha) = \psi(\alpha)$. Sei α^* das kleinste $\alpha > 0$ mit dieser Eigenschaft. Insbesondere gilt dann $\varphi'(\alpha^*) \ge \psi'(\alpha^*)$. Wir unterscheiden zwei Fälle:

(i) **Fall 1:** $\varphi'(\alpha^*) < 0$.
Dann gilt

$$|\varphi'(\alpha^*)| = -\varphi'(\alpha^*) \le -\psi'(\alpha^*) = -\sigma\nabla f(x)^\top d = -\sigma\varphi'(0) \overset{\sigma\le\varrho}{\le} -\varrho\varphi'(0).$$

Also erfüllt α^* die Bedingungen (3.6)-(3.7).

(ii) **Fall 2:** $\varphi'(\alpha^*) \ge 0$.
Wegen $\varphi'(0) < 0$ existiert ein $\hat\alpha \in (0, \alpha^*]$ mit $\varphi'(\hat\alpha) = 0$. Aus $\hat\alpha \le \alpha^*$ und der Definition von α^* folgt die Bedingung $\varphi(\hat\alpha) \le \psi(\hat\alpha)$. Die Bedingung $|\varphi'(\hat\alpha)| \le -\varrho\varphi'(0)$ ist wegen $\varphi'(\hat\alpha) = 0$ automatisch erfüllt. Also erfüllt $\hat\alpha$ die Bedingungen (3.6)-(3.7).

In beiden Fällen existiert also eine Schrittweite, die die strenge Wolfe-Powell-Regel erfüllt. Daraus folgt die Behauptung. □

Nun stellen wir die Frage nach der Effizienz der Regeln. Da die Wolfe-Powell-Regel eine Abschwächung der strengen Wolfe-Powell-Regel darstellt (also mehr Schrittweiten akzeptiert), genügt es, die Effizienz der Wolfe-Powell-Regel zu zeigen. Die Effizienz der strengen Wolfe-Powell-Regel (also der kleineren Schrittweitenmenge) folgt dann automatisch.

Satz 3.7.4 (Effizienz)
Sei $f : \mathbb{R}^n \to \mathbb{R}$ stetig differenzierbar und nach unten beschränkt, $x^{[0]} \in \mathbb{R}^n$ und $\sigma \in (0, 1/2)$, $\varrho \in [\sigma, 1)$. Weiter seien

$$x \in lev(f, f(x^{[0]})) = \{y \in \mathbb{R}^n \mid f(y) \le f(x^{[0]})\}$$

und $d \in \mathbb{R}^n$ mit $\nabla f(x)^\top d < 0$. Außerdem sei der Gradient $\nabla f(x)$ auf der Niveaumenge $lev(f, f(x^{[0]}))$ lipschitzstetig mit Lipschitzkonstante L. Dann gilt

$$f(x + \alpha d) \le f(x) - \frac{(1-\varrho)\sigma}{L}\left(\frac{\nabla f(x)^\top d}{\|d\|}\right)^2.$$

Beweis: Sei α Schrittweite mit (3.4)-(3.5) (die Existenz ist nach Satz 3.7.3 gesichert). Dann gilt $f(x+\alpha d) \le f(x)$ und insbesondere $x+\alpha d \in \operatorname{lev}(f, f(x^{[0]}))$. Die Wolfe-Powell-Bedingung (3.5) impliziert

$$(\varrho - 1)\nabla f(x)^\top d \le (\nabla f(x+\alpha d) - \nabla f(x))^\top d.$$

Die Cauchy-Schwarzsche Ungleichung zusammen mit der Lipschitzstetigkeit von ∇f liefert

$$(\varrho - 1)\nabla f(x)^\top d \le \|\nabla f(x+\alpha d) - \nabla f(x)\| \cdot \|d\| \le L\alpha\|d\|^2.$$

Es folgt

$$\alpha \ge \frac{(\varrho - 1)\nabla f(x)^\top d}{L\|d\|^2}$$

und mit (3.4) schließlich

$$f(x+\alpha d) \le f(x) + \sigma \cdot \alpha \cdot \nabla f(x)^\top d \le f(x) - \frac{(1-\varrho)\sigma}{L}\left(\frac{\nabla f(x)^\top d}{\|d\|}\right)^2.$$

Damit ist die Effizienz nachgewiesen. □

Abschließend widmen wir uns der Realisierung der Wolfe-Powell-Regel (3.4)-(3.5) und formulieren einen 2-Phasen-Algorithmus. Die strenge Wolfe-Powell-Regel kann auf ähnliche Art realisiert werden, vgl. Geiger und Kanzow [GK99], 6.3.

Algorithmus: Wolfe-Powell-Schrittweite

(i) Wähle Parameter $\tau_1, \tau_2 \in (0, 1/2]$, $\gamma > 1$ und eine Startschrittweite $\alpha > 0$.

(ii) Falls (3.4) für α nicht erfüllt ist, gehe zu (iii).
Falls (3.4) für α erfüllt ist, aber (3.5) für α nicht erfüllt ist, setze $\alpha := \gamma \cdot \alpha$ und gehe zu (ii).
Falls (3.4) und (3.5) für α erfüllt sind, setze $\alpha_i = \alpha$ und beende das Verfahren.

(iii) Setze $a := 0$ und $b := \alpha$.

(iv) Wähle $\alpha \in [a + \tau_1(b-a), b - \tau_2(b-a)]$.

(v) Falls (3.4) für α nicht erfüllt ist, setze $b := \alpha$ und gehe zu (iv).
Falls (3.4) für α erfüllt ist, aber (3.5) für α nicht erfüllt ist, setze $a := \alpha$ und gehe zu (iv).
Falls (3.4) und (3.5) für α erfüllt sind, setze $\alpha_i = \alpha$ und beende das Verfahren.

Erläuterungen:

- In (ii) findet die **Expansionsphase** statt: Das Intervall $[0, \alpha]$ wird solange vergrößert, bis (3.4) nicht mehr gilt. Der Beweis von Satz 3.7.3 zeigt, daß es in diesem Intervall dann zulässige Schrittweiten geben muß (beachte die Definition von α^*).

- In der **Kontraktionsphase** (iv), (v) wird das Intervall $[a, b]$ schrittweise verkleinert, wobei stets $0 \le a < b$ und

$$\varphi(a) \le \varphi(0) + \sigma \cdot a \cdot \varphi'(0), \quad \varphi'(a) < \varrho \cdot \varphi'(0), \quad \varphi(b) > \varphi(0) + \sigma \cdot b \cdot \varphi'(0)$$

  gelten. Fordert man noch zusätzlich, daß $\sigma < \varrho$ gilt, so zeigt untenstehender Hilfssatz 3.7.5, daß das Intervall $[a, b]$ unter diesen Bedingungen stets ein ganzes Intervall mit zulässigen Schrittweiten enthält. Da die Intervalllänge $b - a$ gegen Null konvergiert, bricht die Schleife nach endlich vielen Schritten mit einer zulässigen Schrittweite ab.

- Die Schrittweite α in (iv) kann durch Interpolation bestimmt werden. Dazu bestimmt man das kubische Hermite-Interpolationspolynom p, welches durch die Da-

ten

$$p(a) = \varphi(a), \quad p'(a) = \varphi'(a), \quad p(b) = \varphi(b), \quad p'(b) = \varphi'(b)$$

eindeutig bestimmt ist. Das lokale Minimum $\hat{\alpha}$ kann (falls existent) explizit berechnet werden. Liegt $\hat{\alpha}$ im Intervall $[a + \tau_1(b-a), b - \tau_2(b-a)]$, so wähle $\alpha = \hat{\alpha}$. Andernfalls wähle $\alpha = \frac{1}{2}(a + b + (\tau_1 - \tau_2)(b-a))$ (Mittelpunkt des Intervalls). Möchte man die Berechnung des Gradienten $\varphi'(b)$ vermeiden, so kann alternativ das quadratische Interpolationspolynom q zu den Bedingungen

$$q(a) = \varphi(a), \quad q'(a) = \varphi'(a), \quad q(b) = \varphi(b)$$

verwendet werden.

Hilfssatz 3.7.5

Seien $\sigma < \varrho$ und $\varphi'(0) < 0$. Ist $[a,b]$ mit $0 \le a < b$ ein Intervall mit den Eigenschaften

$$\varphi(a) \le \varphi(0) + \sigma \cdot a \cdot \varphi'(0), \quad \varphi(b) \ge \varphi(0) + \sigma \cdot b \cdot \varphi'(0), \quad \varphi'(a) < \sigma\varphi'(0), \tag{3.8}$$

so enthält $[a,b]$ einen Punkt $\bar{\alpha} > 0$ mit

$$\varphi(\bar{\alpha}) < \varphi(0) + \sigma \cdot \bar{\alpha} \cdot \varphi'(0), \quad \varphi'(\bar{\alpha}) = \sigma\varphi'(0).$$

Außerdem ist $\bar{\alpha}$ innerer Punkt eines Intervalls I, so daß für alle $\alpha \in I$ die Wolfe-Powell-Bedingungen (3.4)-(3.5) gelten.

Beweis: Sei $\bar{\alpha}$ ein globales Minimum von

$$\psi(\alpha) := \varphi(\alpha) - \varphi(0) - \sigma \cdot \alpha \cdot \varphi'(0)$$

auf $[a,b]$. Wegen (3.8) ist $\bar{\alpha}$ innerer Punkt von $[a,b]$. Notwendig gilt dann $\psi'(\bar{\alpha}) = 0$ und $\psi(\bar{\alpha}) < \psi(a) \le 0$. Aus $\psi(\bar{\alpha}) < 0$ und $\psi'(\bar{\alpha}) = 0$ und $\sigma < \varrho$ folgt die Existenz eines Intervalls I mit $\bar{\alpha}$ als innerem Punkt, so daß für alle $\alpha \in I$ gilt:

$$\psi(\alpha) \le 0, \quad \psi'(\alpha) \ge (\varrho - \sigma)\varphi'(0),$$

was äquivalent ist mit $\psi(\alpha) \le 0$, $\varphi'(\alpha) \ge \varrho\varphi'(0)$. □

3.7.3 Goldstein-Regel

Zu vorgegebener Zahl $\sigma \in (0, 1/2)$ bestimme $\alpha > 0$, so daß

$$\varphi(0) + (1-\sigma) \cdot \alpha \cdot \varphi'(0) \le \varphi(\alpha) \le \varphi(0) + \sigma \cdot \alpha \cdot \varphi'(0) \tag{3.9}$$

gilt. Ziel dieser Regel ist es, ähnlich wie bei den Wolfe-Powell-Regeln, zu kleine Schrittweiten zu vermeiden.

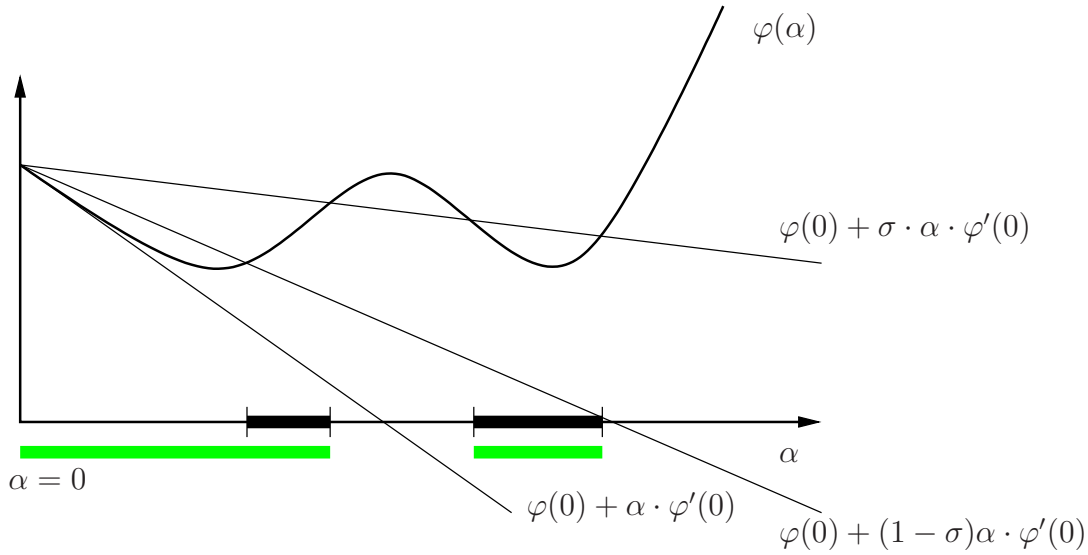


Abbildung 3.5: Die schwarzen Bereiche kennzeichnen die Schrittweiten, die der Goldsteinregel genügen. Die grünen Bereiche kennzeichnen die Schrittweiten, die der Armijoregel genügen.

Die fett eingezeichneten grünen Intervalle in Abbildung 3.5 erfüllen die Armijo-Bedingung (3.3). Die fett eingezeichneten schwarzen Intervalle in Abbildung 3.5 erfüllen die Goldstein-Bedingung (3.9). Zu kleine Schrittweiten werden abgeschnitten.

Die Sätze 3.7.3 und 3.7.4 gelten entsprechend, d.h. unter den dortigen Voraussetzungen ist die Goldstein-Regel wohldefiniert und effizient.

Zur Realisierung der Goldstein-Regel kann der folgende Algorithmus verwendet werden:

Algorithmus: Goldstein-Regel

(i) Wähle einen Parameter $\gamma > 1$ und setze $a_0 := 0$, $b_0 > 0$, $i = 0$.

(ii) Falls b_i die Bedingung (3.9) erfüllt, setze $\alpha_i = b_i$ und beende das Verfahren.
Falls $\varphi(b_i) < \varphi(0) + (1-\sigma) b_i \varphi'(0)$ gilt, setze $a_{i+1} := b_i$, $b_{i+1} := \gamma \cdot b_i$, $i := i+1$ und gehe zu (ii).
Falls $\varphi(b_i) > \varphi(0) + \sigma b_i \varphi'(0)$ gilt, gehe zu (iii).

(iii) Falls $\alpha_i = (a_i + b_i)/2$ die Bedingung (3.9) erfüllt, beende das Verfahren mit Schrittweite α_i.
Falls $\varphi(\alpha_i) < \varphi(0) + (1-\sigma)\alpha_i \varphi'(0)$ gilt, setze $a_{i+1} := \alpha_i$, $b_{i+1} := b_i$, $i := i+1$ und gehe zu (iii).
Falls $\varphi(\alpha_i) > \varphi(0) + \sigma \alpha_i \varphi'(0)$ gilt, setze $a_{i+1} := a_i$, $b_{i+1} := \alpha_i$, $i := i+1$ und gehe zu (iii).

Bemerkung 3.7.6

- *Ist f stetig differenzierbar und nach unten beschränkt und gilt $\nabla f(x)^\top d < 0$, $x, d \in \mathbb{R}^n$, so bricht der Algorithmus nach endlich vielen Schritten mit einer Goldstein-Schrittweite ab.*

- *Ein Nachteil der Goldstein-Regel ist, daß das tatsächliche Minimum von φ mitunter abgeschnitten wird.*

- *Alternativ zur Halbierung des Intervalls $[a_i, b_i]$ in (iii) können wiederum Interpolationstechniken verwendet werden, z.B. indem die Punkte $(a_i, \varphi(a_i))$ und $(b_i, \varphi(b_i))$ linear interpoliert werden, die Schnittpunkte ξ_1, ξ_2 mit den Geraden $\varphi(0) + (1-\sigma)\alpha\varphi'(0)$ und $\varphi(0) + \sigma\alpha\varphi'(0)$ berechnet werden und $\alpha_i = (\xi_1 + \xi_2)/2$ verwendet wird.*

- *Bei Verwendung der Goldstein-Regel in einem Abstiegsverfahren kann die vorherige Schrittweite als das b_0 in (i) verwendet werden. Bei Verwendung der Goldstein-Regel in Newton- oder Quasi-Newton-Verfahren sollte allerdings stets die Schrittweite $\alpha = 1$ ($\rightarrow$ quadratische Konvergenz) zuerst getestet werden.*

3.7.4 Exakte Minimierung

Die Schrittweite $\alpha_i > 0$ wird durch eindimensionale Minimierung der Funktion φ erhalten:

$$\alpha_i := arg \min_{\alpha > 0} \varphi(\alpha)$$

Existenz:
Ist f stetig und ist die Niveaumenge

$$\text{lev}(f, f(x^{[0]}) = \{x \in \mathbb{R}^n \mid f(x) \leq f(x^{[0]})\}$$

beschränkt, so besitzt die stetige Funktion φ auf der kompakten Menge

$$\{\alpha \geq 0 \mid \varphi(\alpha) \leq \varphi(0)\}$$

ein Minimum.

Variante: Zusatzforderung $\alpha_i \in [0, \bar{\alpha}]$.
Ist f stetig, so nimmt φ auf dem Kompaktum $[0, \bar{\alpha}]$ ihr Minimum an. Diese Regel ist also wohldefiniert – auch für unbeschränkte Niveaumengen.

Allerdings läßt sich diese Regel nur in Spezialfällen realisieren, etwa für quadratische Funtionen f. Im allgemeinen kann die Bestimmung des Minimums nur iterativ und approximativ erfolgen (u.U. mit der Golden-Section Search, Fibonacci-Search, ...).
Das Verfahren der exakten Minimierung liefert unter geeigneten Voraussetzungen eine effiziente Schrittweite.
Die exakte Minimierung wird häufig für theoretische Untersuchungen verwendet.

3.8 Gradientenverfahren

Wir untersuchen hier das Gradientenverfahren bzw. das Verfahren steilsten Abstiegs zur (unrestringierten) Minimierung einer Funktion $f \in C^1(\mathbb{R}^n, \mathbb{R})$. Das Gradientenverfahren verwendet die Suchrichtung $d^{[i]} = -\nabla f(x^{[i]})$. Da die exakte Liniensuche i.a. zu aufwendig ist, ersetzen wir diese durch die Schrittweitenbestimmung nach Armijo mit Parametern $\sigma, \beta \in (0, 1)$.
Als Abbruchkriterium verwenden wir die Bedingung $\|\nabla f(x^{[i]})\| \leq \varepsilon$ mit $\varepsilon > 0$. Für die folgende Konvergenzuntersuchung setzen wir jedoch $\varepsilon = 0$. Die Schrittweitenstrategie nach Armijo erfüllt i.a. nicht die Effizienzbedingung, so dass der allgemeine Konvergenzsatz nicht ohne zusätzliche Voraussetzungen angewendet werden kann.

Satz 3.8.1

Wird durch das Gradientenverfahren für ein $f \in C^1(\mathbb{R}^n, \mathbb{R})$, $x^{[0]} \in \mathbb{R}^n$ und $\varepsilon = 0$ eine nicht abbrechende Folge $\{x^{[k]}\}$ definiert, so gilt für jeden Häufungspunkt x^* dieser Folge $\nabla f(x^*) = 0$.

Beweis: (indirekt)
Wir nehmen an, es gäbe eine Teilfolge $\{x^{[k_j]}\}$ von $\{x^{[k]}\}$ mit $x^{[k_j]} \to x^*$ und $\nabla f(x^*) \neq 0$.

Da die Folge $\{f(x^{[k]})\}$ nach Konstruktion streng monoton fällt, folgt aus der Stetigkeit von f, dass $f(x^{[k]}) \to f(x^*)$ konvergiert. Somit gilt nach Wahl der Schrittweite

$$\sigma\,\alpha_k\,\|\nabla f(x^{[k]})\|^2 \;=\; -\sigma\,\alpha_k\,\nabla f(x^{[k]})^\top d^{[k]} \;\le\; f(x^{[k]}) - f(x^{[k+1]}) \;\to\; 0$$

Speziell für die Teilfolge $k = k_j$ folgt somit aus $\nabla f(x^*) \neq 0$, dass

$$\alpha_{k_j} \to 0 \quad (j \to \infty).$$

Für hinreichend große j und $k = k_j$ kann daher $\alpha_k < 1$ angenommen werden. Aus der Schrittweitensteuerung nach Armijo ergibt sich nun:

$$f(x^{[k]} + \beta^{\ell_k - 1}\,d^{[k]}) \;>\; f(x^{[k]}) \;+\; \sigma\,\beta^{\ell_k-1}\,\nabla f(x^{[k]})^\top d^{[k]},$$

oder umgeformt

$$\frac{f(x^{[k]} + \beta^{\ell_k-1}\,d^{[k]}) \;-\; f(x^{[k]})}{\beta^{\ell_k-1}} \;>\; \sigma\,\nabla f(x^{[k]})^\top d^{[k]}$$

Der Mittelwertsatz liefert nun mit einem Zwischenpunkt $z^{[k]} = x^{[k]} + \Theta_k\,\beta^{\ell_k-1}\,d^{[k]}$, $\Theta_k \in]0,1[$: $\nabla f(z^{[k]})^\top d^{[k]} > \sigma\,\nabla f(x^{[k]})^\top d^{[k]}$.
Für $j \to \infty$ gilt nun $\beta^{\ell_k-1} = \alpha_k/\beta \to 0$ und $d^{[k]} \to -\nabla f(x^*) \neq 0$. Damit folgt auch $z^{[k]} \to x^*$ und somit im Grenzwert

$$-\,\|\nabla f(x^*)\|^2 \;\ge\; -\sigma\,\|\nabla f(x^*)\|^2,$$

was der Voraussetzung $0 < \sigma < 1$ widerspricht. □

3.8.1 Quadratische Zielfunktion

Zur Untersuchung der Konvergenzgeschwindigkeit des Gradientenverfahrens beschränken wir uns auf den Fall einer quadratischen Zielfunktion

$$f(x) = \frac{1}{2}(x - x^*)^\top A (x - x^*), \tag{3.10}$$

wobei $A \in \mathbb{R}^{n\times n}$ symmetrisch und positiv definit sei.

Bemerkung 3.8.2

Eine **beliebige** *quadratische Funktion*

$$\tilde f(x) = \frac{1}{2}x^\top A x + b^\top x + c$$

mit symmetrischer und positiv definiter Matrix A, lässt sich (formal) stets in die Form (3.10) transformieren. Dazu sei x^* Lösung des linearen Gleichungssystems $\nabla \tilde f(x) = Ax + b = 0$.

Wir formen nun um

$$\begin{aligned}\tilde{f}(x) &= \frac{1}{2}(x-x^*)^\top A(x-x^*) + (x^*)^\top Ax - \frac{1}{2}(x^*)^\top Ax^* + (-Ax^*)^\top x + c \\ &= \frac{1}{2}(x-x^*)^\top A(x-x^*) + c - \frac{1}{2}(x^*)^\top Ax^*.\end{aligned}$$

Die beiden Zielfunktionen $\tilde{f}$ und f unterscheiden sich also nur durch eine additive Konstante, $\tilde{f}(x) = f(x) + \tilde{f}(x^*)$.

Für die quadratische Zielfunktion f lässt sich nun die **exakte** Schrittweite explizit angeben. Wir haben:

$$\begin{aligned}f(x) &= \frac{1}{2}(x-x^*)^\top A(x-x^*) \\ g(x) &:= \nabla f(x) = A(x-x^*)\end{aligned}$$

Damit folgt

$$\begin{aligned}\frac{d}{d\alpha}\left(f(x-\alpha g)\right) &= \left\{A(x-\alpha g-x^*)\right\}^\top (-g) \\ &= -(x-x^*)^\top Ag + \alpha(g^\top Ag) \\ &= -g^\top g + \alpha(g^\top Ag)\end{aligned}$$

und für die optimale Schrittweite (exakte Liniensuche) ergibt sich

$$\alpha_k := \frac{g_k^\top g_k}{g_k^\top A g_k}, \quad g_k := \nabla f(x^{[k]})$$

Im Gradientenverfahren wird also die Armijo-Regel ersetzt durch die exakte Liniensuche. Setzt man nun $x^{[k+1]}$ in f ein, so erhält man nach etwas Rechnung die folgende Relation für den Abstieg der Zielfunktion:

$$f(x^{[k+1]}) = \left(1 - \frac{(g_k^\top g_k)^2}{(g_k^\top A g_k)(g_k^\top A^{-1} g_k)}\right) f(x^{[k]}). \tag{3.11}$$

Zur Abschätzung des Verkleinerungsfaktors von f aus der obigen Relation verwenden wir nun die so genannte Kantorowitsch–Ungleichung.

Satz 3.8.3 (Kantorowitsch–Ungleichung)
Sei $A \in \mathbb{R}^{n\times n}$ eine symmetrische und positiv definite Matrix mit Eigenwerten $0 < \lambda_1 \leq \lambda_2 \leq \ldots \leq \lambda_n$. Dann gilt für alle $x \in \mathbb{R}^n$:

$$\frac{(x^\top x)^2}{(x^\top Ax)\,(x^\top A^{-1}x)} \;\geq\; 4\,\frac{\lambda_1\,\lambda_n}{(\lambda_1 \,+\, \lambda_n)^2}\,, \qquad \textit{für } x \neq 0.$$

Fasst man nun (3.11) und Satz 3.8.3 zusammen, so ergibt sich die folgende Aussage über die Konvergenzgeschwindigkeit des Gradientenverfahrens

Satz 3.8.4 (Konvergenzgeschwindigkeit)

Wendet man auf eine quadratische Zielfunktion

$$f(x) = \frac{1}{2}x^\top Ax + b^\top x + c$$

mit $A \in \mathbb{R}^{n\times n}$ symmetrisch und positiv definit, das Gradientenverfahren mit exakter Schrittweitenbestimmung an, so gilt für alle $k \in \mathbb{N}$:

$$(f(x^{[k+1]}) - f(x^*)) \le \left(\frac{\lambda_{\max} - \lambda_{\min}}{\lambda_{\max} + \lambda_{\min}}\right)^2 (f(x^{[k]}) - f(x^*)).$$

Dabei bezeichnet $\lambda_{\max}$ bzw. $\lambda_{\min}$ den größten bzw. kleinsten Eigenwert der Hesse–Matrix $A = \nabla^2 f(x^*)$.

Beweis: Nach Bemerkung 3.8.2 gilt $f(x) = \frac{1}{2}(x-x^*)^\top A(x-x^*) + f(x^*)$, also folgt mit (3.11) auch

$$\begin{aligned} f(x^{[k+1]}) - f(x^*) &= \left(1 - \frac{(g_k^\top g_k)^2}{(g_k^\top A g_k)(g_k^\top A^{-1} g_k)}\right)(f(x^{[k]}) - f(x^*)) \\ &\le \left(1 - 4\frac{\lambda_1\lambda_n}{(\lambda_1+\lambda_n)^2}\right)(f(x^{[k]}) - f(x^*)) \\ &= \left(\frac{\lambda_n - \lambda_1}{\lambda_n + \lambda_1}\right)^2 (f(x^{[k]}) - f(x^*)). \end{aligned}$$

□

Bemerkung 3.8.5

- *Die Abschätzung in Satz 3.8.4 wird für kritische Beispiele mit Gleichheit erfüllt.*
- *Satz 3.8.4 gilt sinngemäß auch für skalierte negative Gradientenrichtungen der Form $d = -B^{-1}\nabla f(x)$, B symmetrisch und positiv definit. Dabei bezeichnet nun $\lambda_{\max}$ bzw. $\lambda_{\min}$ den größten bzw. kleinsten Eigenwert von $B^{-1}A$. Man beachte, daß $B^{-1}A$ ähnlich ist zur Matrix $B^{-1/2}AB^{1/2}$ und daß diese Matrix symmetrisch und positiv definit ist.*
- *$\kappa := \lambda_{\max}/\lambda_{\min}$ ist die spektrale Konditionszahl der Matrix A. Nach Satz 3.8.4 gilt*

  $$(f(x^{[k+1]}) - f(x^*)) \le \left(\frac{\kappa - 1}{\kappa + 1}\right)^2 (f(x^{[k]}) - f(x^*)),$$

  d.h. die Konvergenz ist umso langsamer, je größer die Kondition κ ist. Für $\kappa = 100$ ergibt sich z.B. der Verkleinerungsfaktor ≈ 0.96.
- *Die Aussage des Satzes 3.8.4 läßt sich lokal auch auf den Fall nichtquadratischer Funktionen $f \in C^2(\mathbb{R}^n, \mathbb{R})$ übertragen (vgl. Luenberger, 1973). Dabei ist $\lambda_{\max}$ bzw. $\lambda_{\min}$ der größte bzw. kleinste Eigenwert der (als symmetrisch und positiv definit vorausgesetzten) Hesse–Matrix $\nabla f(x^*)$ im Lösungspunkt.*

Beispiel 3.8.6 (Luenberger)
Wir betrachten die quadratische Zielfunktion $f(x) = 0.5x^\top Ax - b^\top x$ *mit*

$$A = \begin{pmatrix} 0.78 & -0.02 & -0.12 & -0.14 \\ -0.02 & 0.86 & -0.04 & 0.06 \\ -0.12 & -0.04 & 0.72 & -0.08 \\ -0.14 & 0.06 & -0.08 & 0.74 \end{pmatrix}, \quad b = \begin{pmatrix} 0.76 \\ 0.08 \\ 1.12 \\ 0.68 \end{pmatrix}.$$

Die Matrix A *ist symmetrisch und diagonaldominant und damit auch positiv definit. Beispielsweise mit der MATLAB–Routine eig findet man* $\lambda_{\min} \approx 0.52$ *und* $\lambda_{\max} \approx 0.94$. *Damit wird* $\kappa \approx 1.8$ *und*

$$\left(\frac{\lambda_{\max} - \lambda_{\min}}{\lambda_{\max} + \lambda_{\min}}\right)^2 \approx 0.081.$$

Mit den Startvektor $x^{[0]} := 0$ *wird das Abbruchkriterium* $\|\nabla f(x^{[k]})\| \le 10^{-5}$ *bereits nach 8 Iterationen erfüllt. Für die numerische Lösung findet man*

$$x^{[8]} \approx (1.53496, 0.12201, 1.97515, 1.41295)^\top.$$

Beweis: Beweis der Kantorowitsch–Ungleichung:
Zu zeigen ist:

$$\forall\, x \in \mathbb{R}^n \setminus \{0\} : \quad \frac{x^T x}{(x^T Ax)(x^T A^{-1} x)} \;\ge\; \frac{4\lambda_1 \lambda_n}{(\lambda_1 + \lambda_n)^2}.$$

Dabei seien $0 < \lambda_1 \le \ldots \le \lambda_n$ die Eigenwerte der symmetrischen und positiv definiten Matrix A.
Es sei (u_j) eine Orthonormalbasis aus zugehörigen Eigenvektoren von A. Mit der Darstellung $x = \sum_{i=1}^n \xi_i u_i$, $\xi_i \in \mathbb{R}$, $\sum_{i=1}^n \xi_i^2 > 0$ folgt dann

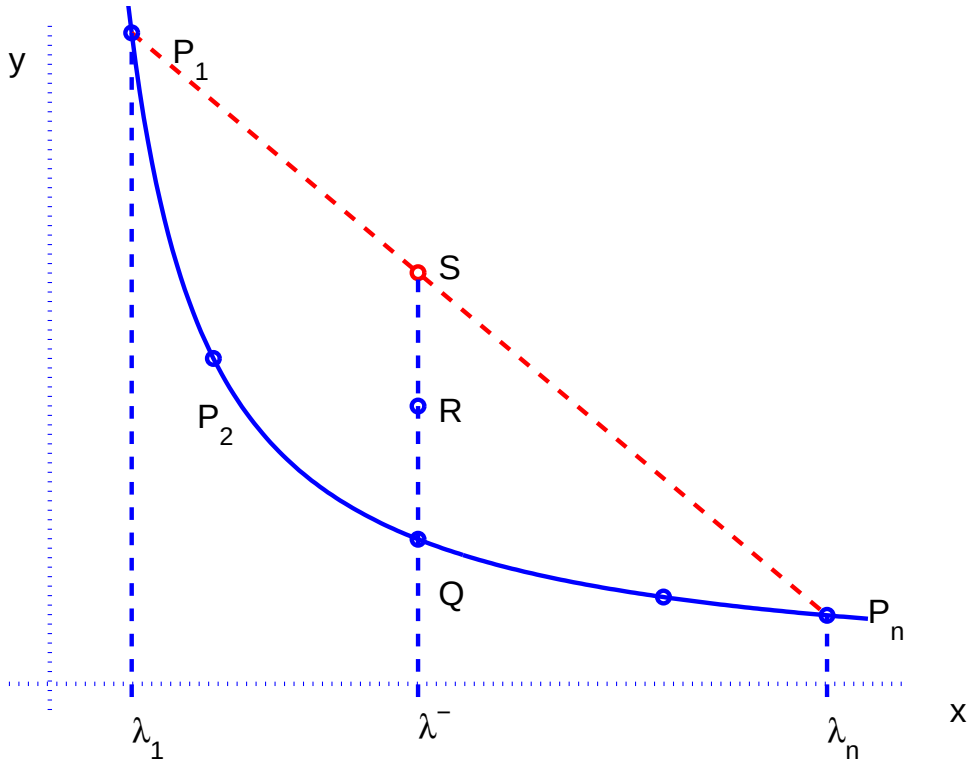
$$\begin{aligned} F(x) \;&:=\; \frac{x^\top x}{(x^\top Ax)(x^\top A^{-1}x)} = \frac{(\sum \xi_j^2)^2}{(\sum \lambda_i \xi_i^2)(\sum \frac{1}{\lambda_i}\xi_i^2)} \\ &=\; \frac{1}{(\sum \lambda_i \frac{\xi_i^2}{\sum \xi_j^2})(\sum \frac{1}{\lambda_i}\frac{\xi_i^2}{\sum \xi_j^2})} \\ &=\; \frac{(\sum \lambda_i \gamma_i)^{-1}}{(\sum \lambda_i^{-1}\gamma_i)}, \quad \text{mit} \quad \gamma_i := \xi_i^2 / \sum \xi_j^2. \end{aligned}$$

Die γ_i erfüllen die Voraussetzung der Koeffizienten einer Konvexkombination $\gamma_i \ge 0$, $\sum \gamma_i = 1$.

Betrachtet man die folgenden Punkte in der Ebene

$$P_i := (\lambda_i, \frac{1}{\lambda_i}), \quad Q := (\sum \gamma_i\lambda_i, \frac{1}{\sum \gamma_i\lambda_i}), \quad R := (\sum \gamma_i\lambda_i, \sum \gamma_i\frac{1}{\lambda_i}),$$

so stellt man fest:
Die Punkte $P_1, \ldots, P_n$ und Q liegen auf dem Hyperbelast $y = 1/x,\ x > 0$.
Die x–Koordinate $\bar\lambda := \sum \gamma_i\lambda_i$ von Q ist einen Konvexkombination der $\lambda_1, \ldots, \lambda_n$; daher gilt $\lambda_1 \le \bar\lambda \le \lambda_n$.
R hat die gleiche x–Koordinate wir Q und ist eine Konvexkombination von $P_1, \ldots, P_n$. Deshalb muss R in der konvexen Hülle von $P_1, \ldots, P_n$ liegen, also oberhalb von Q und unterhalb der Sekante zwischen P_1 und P_n.



Die Sekantengleichung (Zweipunkteform der Geradengleichung) lautet

$$\frac{y - 1/\lambda_1}{x - \lambda_1} = \frac{1/\lambda_n - 1/\lambda_1}{\lambda_n - \lambda_1} \quad \text{oder} \quad y = \frac{\lambda_1 + \lambda_n - x}{\lambda_1\lambda_n}.$$

Damit folgt $\sum \gamma_i\frac{1}{\lambda_i} \le \frac{\lambda_1 + \lambda_n - \bar\lambda}{\lambda_1\lambda_n}$ und somit

$$F(x) = \frac{\bar\lambda^{-1}}{\sum \gamma_i\lambda_i^{-1}} \ge \frac{\lambda_1\lambda_n}{\bar\lambda(\lambda_1 + \lambda_n - \bar\lambda)} \ge \min\{\frac{\lambda_1\lambda_n}{\mu(\lambda_1 + \lambda_n - \mu)} : \lambda_1 \le \mu \le \lambda_n\}.$$

Das Minimum wird in $\mu = (\lambda_1 + \lambda_n)/2$ angenommen (der Nenner ist eine nach unten geöffnete Parabel) und somit folgt

$$f(x) \ge \frac{4\lambda_1\lambda_n}{(\lambda_1 + \lambda_n)^2}$$

□

3.9 Newton-Verfahren

Ein generelles Prinzip zur Bestimmung einer Abstiegsrichtung besteht darin, die Funktion f lokal in der Nähe der Iterierten $x^{[i]}$ durch eine Approximation $\hat{f}(y)$ zu ersetzen und diese dann zu minimieren.

Beispiel 3.9.1 (Lineare Approximation)
Linearisierung von f in x ergibt die Funktion

$$\hat{f}(y) = f(x) + \nabla f(x)^\top (y - x).$$

Allerdings besitzt diese lineare Funktion i.a. kein Minimum auf $\mathbb{R}^n$. Daher schränkt man zusätzlich die Werte von y ein, indem man fordert

$$\|y - x\| \leq 1.$$

Wir haben bereits gesehen, daß die Lösung $\hat{y}$ des Minimierungsproblems

$$\hat{f}(y) \to \min \qquad \textit{unter} \qquad \|y - x\| \leq 1$$

gerade auf die (normierte) Suchrichtung $d = \hat{y} - x = -\nabla f(x)/\|\nabla f(x)\|$ führt. Es entsteht also das Gradientenverfahren.

Es ist nun naheliegend, f quadratisch zu approximieren, um ein weiteres Verfahren zu erhalten.

Beispiel 3.9.2 (Quadratische Approximation)
f wird lokal durch

$$\hat{f}(y) = f(x) + \nabla f(x)^\top (y - x) + \frac{1}{2}(y - x)^\top \nabla^2 f(x)(y - x)$$

approximiert. Ist die Hessematrix $\nabla^2 f(x)$ positiv definit, so besitzt $\hat{f}$ ein eindeutig bestimmtes Minimum $\hat{y}$, welches durch das lineare Gleichungssystem

$$\nabla \hat{f}(\hat{y}) = \nabla f(x) + \nabla^2 f(x)(\hat{y} - x) = 0$$

bestimmt ist. Minimierung der quadratischen Approximation $\hat{f}(y)$ führt also auf die Suchrichtung

$$d = \hat{y} - x = -\left(\nabla^2 f(x)\right)^{-1} \nabla f(x). \tag{3.12}$$

Wegen

$$\nabla f(x)^\top d = -\nabla f(x)^\top \left(\nabla^2 f(x)\right)^{-1} \nabla f(x) < 0$$

für $\nabla f(x) \neq 0$ ist d eine Abstiegsrichtung von f in x.

Beispiel 3.9.2 beweist

Hilfssatz 3.9.3

Für eine zweimal stetig differenzierbare Funktion $f : \mathbb{R}^n \to \mathbb{R}$ führt das Prinzip der lokalen Minimierung mit quadratischem Modell auf die Richtung

$$\nabla^2 f(x)d = -\nabla f(x). \tag{3.13}$$

Ist die Hessematrix $\nabla^2 f(x)$ positiv definit, so ist d eine Abstiegsrichtung.

Die Annahme, daß die Hessematrix $\nabla^2 f(x)$ positiv definit ist – zumindest für alle vom Abstiegsverfahren erzeugten Punkte –, ist durchaus restriktiv und keinesfalls immer erfüllt. Allerdings werden wir sehen, daß das entstandene Verfahren auch sinnvoll ist, wenn $\nabla^2 f(x)$ nicht positiv definit ist (dann ist es i.a. allerdings kein Abstiegsverfahren mehr). Eine andere Motivation zur Herleitung des Verfahrens besteht nämlich darin, die notwendigen Bedingungen für ein unrestringiertes, lokales Minimum zu erfüllen. Bekanntlich erfüllt ein lokales Minimum $\hat{x}$ von f notwendigerweise

$$\nabla f(\hat{x}) = 0. \tag{3.14}$$

Dies ist ein **nichtlineares Gleichungssystem** für die Funktion $g(x) := \nabla f(x)$. Anwendung des aus der Numerik bekannten Newton-Verfahrens (falls $f \in C^2(\mathbb{R}^n, \mathbb{R})$!) liefert die Iterationsvorschrift

$$\begin{aligned} g'(x^{[i]})\Delta x^{[i]} &= -g(x^{[i]}), \\ x^{[i+1]} &= x^{[i]} + \Delta x^{[i]}, \qquad i = 0, 1, 2, \ldots. \end{aligned}$$

$g'(\cdot)$ bezeichnet die Jacobi-Matrix von g. Die erste Gleichung lautet insbesondere

$$\nabla^2 f(x^{[i]})\Delta x^{[i]} = -\nabla f(x^{[i]}).$$

Ist nun die Hessematrix $\nabla^2 f(x^{[i]})$ regulär, so ist die **Newton-Richtung** $\Delta x^{[i]}$ für das nichtlineare Gleichungssystem (3.14) gegeben durch

$$\Delta x^{[i]} = -\left(\nabla^2 f(x^{[i]})\right)^{-1} \nabla f(x^{[i]}).$$

Ein Vergleich mit (3.12) zeigt, daß beide Verfahren – obwohl vom Ansatz her unterschiedlich – auf dieselbe Suchrichtung $d = \Delta x$ führen. Beide Verfahren heißen daher **Newtonverfahren**.

Wir widmen uns nun der Konvergenz des **lokalen Newtonverfahrens**, welches sich dadurch auszeichnet, daß stets die Schrittweite $\alpha_i = 1$ gewählt wird.

Algorithmus: lokales Newtonverfahren

(i) Wähle einen Startvektor $x^{[0]} \in \mathbb{R}^n$ und setze $i = 0$.

(ii) Falls ein Abbruchkriterium erfüllt ist, STOP.

(iii) Berechne (falls möglich) die Suchrichtung $d^{[i]}$ als Lösung des Gleichungssystems (3.13) mit $x = x^{[i]}$, setze $x^{[i+1]} = x^{[i]} + d^{[i]}$, $i := i + 1$ und gehe zu (ii).

Definition 3.9.4 **(lineare, superlineare und quadratische Konvergenz)**
Sei $\{x^{[i]}\}$ eine Folge.

(a) $\{x^{[i]}\}$ konvergiert **linear** *gegen $\hat{x}$, falls es ein $0 < c < 1$ gibt mit*

$$\|x^{[i+1]} - \hat{x}\| \leq c \cdot \|x^{[i]} - \hat{x}\|$$

für alle hinreichend großen i.

(b) $\{x^{[i]}\}$ konvergiert **superlinear** *gegen $\hat{x}$, falls es eine Nullfolge $\{c_i\}$, $c_i \downarrow 0$ gibt mit*

$$\|x^{[i+1]} - \hat{x}\| \leq c_i \cdot \|x^{[i]} - \hat{x}\| \qquad \forall i \in \mathbb{N},$$

d.h.

$$\|x^{[i+1]} - \hat{x}\| = o(\|x^{[i]} - \hat{x}\|).$$

(c) $\{x^{[i]}\}$ konvergiert **von der Ordnung** $p \geq 2$ *gegen $\hat{x}$, falls es ein $c > 0$ gibt mit*

$$\|x^{[i+1]} - \hat{x}\| \leq c \cdot \|x^{[i]} - \hat{x}\|^p \qquad \forall i \in \mathbb{N},$$

d.h.

$$\|x^{[i+1]} - \hat{x}\| = \mathcal{O}(\|x^{[i]} - \hat{x}\|^p).$$

Im Fall $p = 2$ heißt die Folge **quadratisch** *konvergent.*

Wir wollen die superlineare bzw. quadratische Konvergenz des Newtonverfahrens beweisen. Zur Abkürzung setzen wir $g(x) = \nabla f(x)$, $g'(x) = \nabla^2 f(x)$.
Nun interpretieren wir die Iterationsvorschrift des lokalen Newtonverfahrens als Fixpunktiteration

$$x^{[i+1]} = G(x^{[i]}), \qquad i = 0, 1, 2, \ldots \tag{3.15}$$

für die Fixpunktfunktion

$$G(x) := x - (g'(x))^{-1} g(x). \tag{3.16}$$

Wegen $g(\hat{x}) = \nabla f(\hat{x}) = 0$ ist $\hat{x}$ Fixpunkt von G.
Wir benötigen folgende Hilfsätze:

Hilfssatz 3.9.5

Sei $\hat{x}$ Fixpunkt von G. Es existiere ein $r > 0$ und eine Umgebung

$$U_r(\hat{x}) := \{x \in \mathbb{R}^n \mid \|x - \hat{x}\| \leq r\},$$

sowie eine Konstante $k < 1$ mit

$$\|G(x) - \hat{x}\| \leq k\|x - \hat{x}\|, \qquad \forall x \in U_r(\hat{x})$$

(Kontraktionsbedingung).
Dann gilt: Für jeden Startvektor $x^{[0]} \in U_r(\hat{x})$ erfüllt die durch (3.15) definierte Folge $x^{[i]} \in U_r(\hat{x})$ für $i = 1, 2, \ldots$ und es gilt

$$\lim_{i\to\infty} x^{[i]} = \hat{x}.$$

Die Konvergenz ist lokal (mindestens) linear und $\hat{x}$ ist der einzige Fixpunkt von G in $U_r(\hat{x})$.

Beweis: Wähle $x^{[0]} \in U_r(\hat{x})$ beliebig. Dann gilt

$$\|x^{[1]} - \hat{x}\| = \|G(x^{[0]}) - \hat{x}\| \leq k\|x^{[0]} - \hat{x}\|,$$

also $x^{[1]} \in U_r(\hat{x})$. Induktiv folgt

$$\|x^{[i]} - \hat{x}\| = \|G(x^{[i-1]}) - \hat{x}\| \leq k\|x^{[i-1]} - \hat{x}\| \leq \ldots \leq k^i\|x^{[0]} - \hat{x}\| \overset{i\to\infty}{\to} 0.$$

Insbesondere ist die Iteration wohldefiniert und $x^{[i]} \in U_r(\hat{x})$ für alle $i \in \mathbb{N}$. Ebenso ergibt sich die lineare Konvergenz aus der Abschätzung.
Zur Eindeutigkeit: Ist $y \in U_r(\hat{x})$ ein beliebiger Fixpunkt von G, dann ist $\|\hat{x} - y\| = \|G(\hat{x}) - G(y)\| \leq k\|\hat{x} - y\|$ mit $k < 1$ und daher $y = \hat{x}$. □

Hilfssatz 3.9.6

$G : D \subseteq \mathbb{R}^n \to \mathbb{R}^n$ besitze einen Fixpunkt $\hat{x}$ im Inneren von D und sei differenzierbar in $\hat{x}$ mit

$$G'(\hat{x}) = 0.$$

Dann existiert ein $r > 0$, so daß die Fixpunktiteration (3.15) für alle Startwerte $x^{[0]} \in U_r(\hat{x}) \subseteq D$ lokal superlinear gegen $\hat{x}$ konvergiert.
Gilt zusätzlich für Konstanten $\alpha \geq 0$ und $p > 1$ noch

$$\|G(x) - G(\hat{x})\| \leq \alpha\|x - \hat{x}\|^p \qquad \forall x \in U_r(\hat{x}),$$

so ist die Konvergenzordnung lokal um $\hat{x}$ (mindestens) gleich p.

Beweis:

(i) **Konvergenz:**

Da G differenzierbar ist in $\hat{x}$, existiert zu jedem $\varepsilon > 0$ ein $r > 0$ mit

$$\frac{\|G(x) - G(\hat{x}) - G'(\hat{x})(x - \hat{x})\|}{\|x - \hat{x}\|} \leq \varepsilon \qquad \forall x \in U_r(\hat{x}),\ x \neq \hat{x}.$$

Wegen $G'(\hat{x}) = 0$ folgt daraus

$$\|G(x) - G(\hat{x})\| = \|G(x) - \hat{x}\| \leq \varepsilon \|x - \hat{x}\| \qquad \forall x \in U_r(\hat{x}).$$

Wähle nun $\varepsilon < 1$ und wende Hilfssatz 3.9.5 an.

(ii) **Superlineare Konvergenz:**

Sei nun $\{x^{[i]}\}$ eine Iterationsfolge mit

$$\lim_{i \to \infty} x^{[i]} = \hat{x}.$$

Ist $x^{[i_0]} = \hat{x}$ für ein $i_0 \in \mathbb{N}$, so gilt $x^{[i_0+1]} = G(x^{[i_0]}) = G(\hat{x}) = \hat{x}$ und somit $x^{[i]} = \hat{x}$ für alle $i \geq i_0$. Wir können also ohne Beschränkung der Allgemeinheit annehmen, daß $x^{[i]} \neq \hat{x}$ für alle $i \in \mathbb{N}$ gilt. Wegen $G'(\hat{x}) = 0$ und der Differenzierbarkeit von G in $\hat{x}$ folgt dann

$$\lim_{i \to \infty} \frac{\|x^{[i+1]} - \hat{x}\|}{\|x^{[i]} - \hat{x}\|} = \lim_{i \to \infty} \frac{\|G(x^{[i]}) - G(\hat{x}) - G'(\hat{x})(x^{[i]} - \hat{x})\|}{\|x^{[i]} - \hat{x}\|} = 0,$$

d.h. die Folge ist superlinear konvergent.

(iii) **Konvergenzordnung p:**

Wähle $\tilde{r} = \left(\frac{1}{\alpha}\right)^{1/(p-1)} > 0$. Für alle $x \in U_{\tilde{r}}(\hat{x})$ gilt $\|x - \hat{x}\| \leq \left(\frac{1}{\alpha}\right)^{1/(p-1)}$. Dann gilt für $x^{[i]} \in U_{\tilde{r}}(\hat{x})$ die Beziehung

$$\|x^{[i+1]} - \hat{x}\| = \|G(x^{[i]}) - G(\hat{x})\| \leq \alpha \|x^{[i]} - \hat{x}\|^p \leq \|x^{[i]} - \hat{x}\| \leq \tilde{r},$$

d.h. die Folge $\{x^{[i]}\}$ bleibt in $U_{\tilde{r}}(\hat{x})$. Der Rest folgt aus der Voraussetzung.

□

Der folgende Hilfssatz stellt fest, daß $G'(\hat{x}) = 0$ gilt.

Hilfssatz 3.9.7

Es sei $g : \mathbb{R}^n \to \mathbb{R}^n$ stetig differenzierbar in der Nullstelle $\hat{x}$ und $g'(\hat{x})$ sei invertierbar. Dann existiert ein $r > 0$, so daß die Abbildung (3.16) für alle $x \in U_r(\hat{x})$ wohldefiniert ist und G ist differenzierbar in $\hat{x}$ mit

$$G'(\hat{x}) = I - (g'(\hat{x}))^{-1} g'(\hat{x}) = 0.$$

Beweis: Übung □

Mit diesen Vorarbeiten können wir die Konvergenz des Newtonverfahrens beweisen.

Satz 3.9.8 (Konvergenzsatz für das Newtonverfahren)
Es sei $D \subseteq \mathbb{R}^n$ offen und $f : D \to \mathbb{R}$ besitze ein (lokales) Minimum $\hat{x}$ in D. f sei zweimal stetig differenzierbar auf D und $\nabla^2 f(\hat{x})$ sei invertierbar.
Dann gibt es ein $r > 0$, so daß das lokale Newton-Verfahren für alle Startwerte $x^{[0]} \in U_r(\hat{x})$ wohldefiniert ist und die Folge $\{x^{[i]}\}$ konvergiert superlinear gegen $\hat{x}$ (lokal superlineare Konvergenz). $\hat{x}$ ist das einzige lokale Minimum in $U_r(\hat{x})$.
Erfüllt $\nabla^2 f$ zusätzlich noch die Lipschitz-Bedingung

$$\|\nabla^2 f(x) - \nabla^2 f(y)\| \le L \cdot \|x - y\| \qquad \forall x, y \in D,$$

so ist die Folge $\{x^{[i]}\}$ sogar quadratisch konvergent (lokal quadratische Konvergenz).

Beweis: Die superlineare Konvergenz und die Eindeutigkeit folgen aus den Hilfsätzen 3.9.5, 3.9.6 und 3.9.7. Zu zeigen bleibt die quadratische Konvergenz. Nach Hilfssatz 3.9.6 genügt es,

$$\|G(x) - G(\hat{x})\| \le K\|x - \hat{x}\|^2 \qquad \forall x \in U_r(\hat{x})$$

zu zeigen. Anwendung des Mittelwertsatzes in Integralform und Ausnutzung der Lipschitz-Bedingung liefert

$$\begin{aligned}
\|G(x) - G(\hat{x})\| &= \|x - \hat{x} - g'(x)^{-1} g(x)\| \\
&= \|g'(x)^{-1} \left(g(\hat{x}) - g(x) - g'(x)(\hat{x} - x)\right)\| \\
&\le \|g'(x)^{-1}\| \cdot \left\| \int_0^1 g'(x + t(\hat{x} - x))(\hat{x} - x)dt - g'(x)(\hat{x} - x) \right\| \\
&= \|g'(x)^{-1}\| \cdot \left\| \int_0^1 \left(g'(x + t(\hat{x} - x)) - g'(x)\right) \cdot (\hat{x} - x)dt \right\| \\
&\le \|g'(x)^{-1}\| \cdot \int_0^1 Lt\|\hat{x} - x\|^2 dt \\
&= \|g'(x)^{-1}\| \cdot L \cdot \|\hat{x} - x\|^2 \cdot \int_0^1 t dt \\
&= \|g'(x)^{-1}\| \cdot \frac{L}{2} \cdot \|\hat{x} - x\|^2.
\end{aligned}$$

Da $g'(\hat{x})$ invertierbar ist, existiert auch $g'(x)^{-1}$ in einer Umgebung von $\hat{x}$ und ist dort beschränkt, vgl. Hilfssatz 3.9.9 weiter unten. Damit folgt auch die quadratische Konvergenz. □

Hilfssatz 3.9.9

Es sei $f : \mathbb{R}^n \to \mathbb{R}$ zweimal stetig differenzierbar und $\nabla^2 f(\hat{x})$ sei invertierbar für $\hat{x} \in \mathbb{R}^n$. Dann existiert ein $r > 0$, so daß auch $\nabla^2 f(x)$ für alle $x \in U_r(\hat{x})$ invertierbar ist. Weiter existiert eine Konstante $C > 0$ mit

$$\|\nabla^2 f(x)^{-1}\| \leq C \qquad \forall x \in U_r(\hat{x}).$$

Beweis: Da $\nabla^2 f$ in $\hat{x}$ stetig ist, existiert $r > 0$ mit

$$\|\nabla^2 f(\hat{x}) - \nabla^2 f(x)\| \leq \frac{1}{2\|\nabla^2 f(\hat{x})^{-1}\|} \qquad \forall x \in U_r(\hat{x}).$$

Also ist

$$\|I - \nabla^2 f(\hat{x})^{-1}\nabla^2 f(x)\| \leq \|\nabla^2 f(\hat{x})^{-1}\| \cdot \|\nabla^2 f(\hat{x}) - \nabla^2 f(x)\| \leq \frac{1}{2}$$

für alle $x \in U_r(\hat{x})$. Aus der Numerik ist folgendes bekannt: Sind $A, B \in \mathbb{R}^{n\times n}$ Matrizen mit $\|I - BA\| < 1$, so sind A und B invertierbar und es gilt die Abschätzung

$$\|A^{-1}\| \leq \frac{\|B\|}{1 - \|I - BA\|}.$$

Ausnutzung dieser Aussage liefert in unserem Fall, daß auch $\nabla^2 f(x)$ invertierbar ist für alle $x \in U_r(\hat{x})$ mit

$$\|\nabla^2 f(x)^{-1}\| \leq \frac{\|\nabla^2 f(\hat{x})^{-1}\|}{1 - \underbrace{\|I - \nabla^2 f(\hat{x})^{-1}\nabla^2 f(x)\|}_{\leq 1/2}} \leq 2\|\nabla^2 f(\hat{x})^{-1}\| =: C.$$

□

Bemerkung 3.9.10

- *Das Newtonverfahren läßt sich also entweder als* **Nullstellenverfahren** *zur Bestimmung eines stationären Punktes oder als* **Abstiegsverfahren** *(bei positiv definiter Hessematrix) interpretieren, bei dem stets mit der Schrittweite $\alpha_i = \alpha = 1$ gearbeitet wird.*

  Einen ähnlichen Zusammenhang werden wir später auch bei SQP-Verfahren und Lagrange-Newton-Verfahren erkennen.

- *Nach den obigen Erläuterungen liefert das Newtonverfahren – falls konvergent – lediglich stationäre Punkte. Es könnte daher auch gegen ein Maximum oder einen Sattelpunkt konvergieren.*

- *Eine dritte Motivation der Newton-Richtung liefert das Minimierungsproblem*

$$\nabla f(x)^\top d \to \min \qquad \textit{unter} \qquad \|d\|_A = \sqrt{d^\top A d} = 1,$$

wobei $A = \nabla^2 f(x)$ positiv definit sei. Beachte, daß $\|\cdot\|_A$ dann eine Norm ist. Es wird also der steilste Abstieg in dieser Norm gesucht. Es zeigt sich, daß die Lösung $\hat{d}$ gerade die (normierte) Newton-Richtung ist:

$$\hat{d} = -\frac{A^{-1}\nabla f(x)}{\|A^{-1}\nabla f(x)\|_A}.$$

- *Über Satz 3.9.8 hinaus lassen sich weitere Konvergenzresultate erzielen, wenn $\nabla^2 f$ nur approximativ durch eine Matrix $A(x)$ bestimmt wird, etwa durch finite Differenzen Approximationen oder indem $\nabla^2 f(x^{[i]})$ durch $\nabla^2 f(x^{[0]})$ ersetzt wird. Um die superlineare Konvergenzgeschwindigkeit zu erhalten, wird man verlangen müssen, daß $A(x) \overset{x\to\hat{x}}{\to} \nabla^2 f(\hat{x})$ gilt. Beispiel 3.9.11 verdeutlicht einige Varianten numerisch.*

Beispiel 3.9.11 **(Funktion von Himmelblau)**
Wir minimieren

$$f(x,y) = (x^2 + y - 11)^2 + (x + y^2 - 7)^2$$

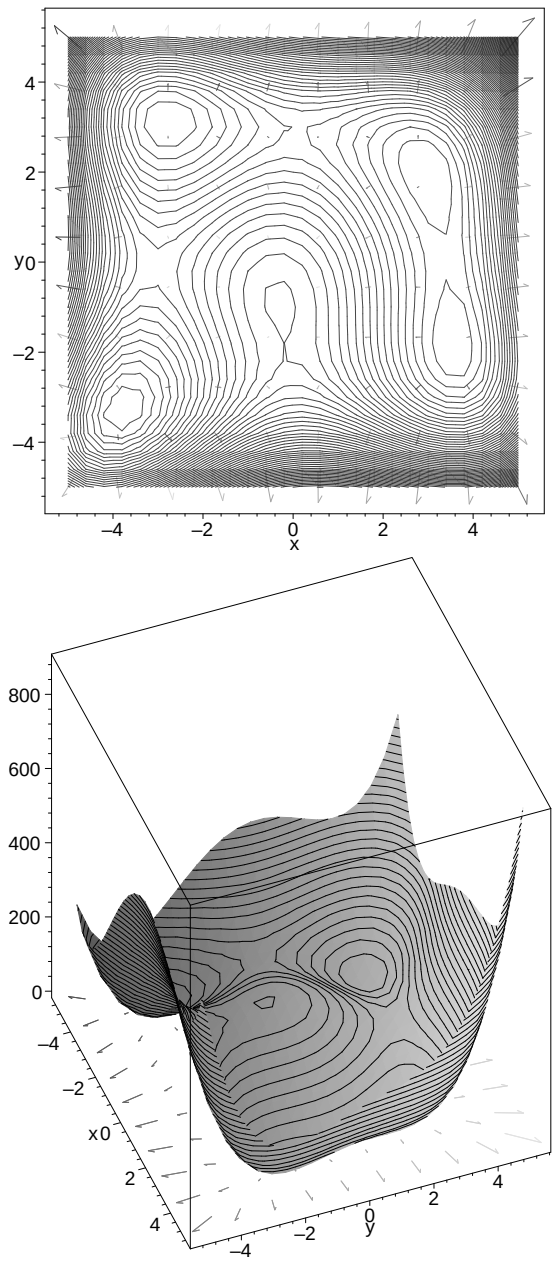
mit dem Newtonverfahren. Es gilt

$$\begin{aligned}
\nabla f(x,y) &= \begin{pmatrix} 4x(x^2+y-11) + 2(x+y^2-7) \\ 2(x^2+y-11) + 4y(x+y^2-7) \end{pmatrix}, \\
\nabla^2 f(x,y) &= \begin{pmatrix} 12x^2 + 4y - 42 & 4(x+y) \\ 4(x+y) & 4x + 12y^2 - 26 \end{pmatrix}.
\end{aligned}$$

- *4 lokale Minimalstellen (zugleich global) mit Funktionswert 0; darunter der Punkt $(3,2)^\top$.*

- *4 Sattelpunkte*

- *ein lokales Maximum in $(-0.270845, -0.923039)^\top$*



Wir diskutieren einige Varianten des Newtonverfahrens. Die Iterationsvorschrift für einen Startwert $x^{[0]}$ *und* $g(x) = \nabla f(x)$ *lautet*

$$\begin{aligned} A(x^{[i]})d^{[i]} &= -g(x^{[i]}), \\ x^{[i+1]} &= x^{[i]} + d^{[i]}, \quad i = 0, 1, 2, \ldots. \end{aligned}$$

Die Matrix $A(\cdot) \in \mathbb{R}^{n \times n}$ *wird dabei auf die folgenden Arten berechnet:*

- $A(x) := \nabla^2 f(x)$ *(klassisches Newtonverfahren);*

- *finite Differenzen-Approximation der Hessematrix* $\nabla^2 f(x)$*, d.h.* $A(x) = (A_{jk}(x))_{j,k=1,\dots,n}$ *ist gegeben durch*

$$A_{jk}(x) := \frac{g_j(x + h_i e_k) - g_j(x)}{h_i}, \quad j,k = 1,\dots,n,$$

wobei e_k *den k-ten Einheitsvektor bezeichnet. Die Schrittweite* h_i *sei durch* $h_i = \frac{0.1}{i+1}$ *gegeben, wobei* i *den Iterationsindex des Newtonverfahrens bezeichnet. Beachte, daß* $h_i \to 0$ *für* $i \to \infty$ *gilt.*

- *vereinfachtes Newtonverfahren mit* $A(x^{[i]}) := \nabla^2 f(x^{[0]})$ *für* $i = 0, 1, 2, \dots$.

Als Abbruchkriterium des Newtonverfahrens verwenden wir jeweils $\|\nabla f(x,y)\| \le 10^{-13}$ *und als Startpunkt* $(x^{[0]}, y^{[0]}) = (4, 2.5)$.
Das klassische Newtonverfahren liefert das folgende Ergebnis:

ITER	X(1)	X(2)	||GRAD||	||DX||	P=1	P=2
0	0.4000000E+01	0.2500000E+01	0.1351240E+03	0.0000000E+00	0.0000000E+00	0.0000000E+00
1	0.3281417E+01	0.2056664E+01	0.2617493E+02	0.8443389E+00	0.2567589E+00	0.2296521E+00
2	0.3035131E+01	0.1988137E+01	0.2424694E+01	0.2556422E+00	0.1291689E+00	0.4499638E+00
3	0.3000634E+01	0.1999744E+01	0.4203618E-01	0.3639711E-01	0.1844451E-01	0.4974264E+00
4	0.3000000E+01	0.2000000E+01	0.1406811E-04	0.6837008E-03	0.3217815E-03	0.4704954E+00
5	0.3000000E+01	0.2000000E+01	0.1500787E-11	0.2200729E-06	0.1109167E-06	0.5039996E+00
6	0.3000000E+01	0.2000000E+01	0.0000000E+00	0.2440976E-13	0.0000000E+00	0.0000000E+00

Die letzten beiden Spalten testen die Folge $(x^{[i]}, y^{[i]})^\top$, $i = 0, 1, 2, \dots$ *auf lineare (P=1 strebt gegen Wert* $\neq 0$*, P=2 explodiert), superlineare (P=1 strebt gegen* 0*, P=2 explodiert) und quadratische Konvergenz (P=1 strebt gegen* 0*, P=2 strebt gegen Wert* $\neq 0$*).*
Die Rechnungen zeigen eine quadratische Konvergenz des klassischen Newtonverfahrens gegen das Minimum $(3,2)^\top$.
Das Newtonverfahren mit finiten Differenzen liefert das folgende Ergebnis:

ITER	X(1)	X(2)	||GRAD||	||DX||	P=1	P=2
0	0.4000000E+01	0.2500000E+01	0.1351240E+03	0.0000000E+00	0.0000000E+00	0.0000000E+00
1	0.3300753E+01	0.2071139E+01	0.2853753E+02	0.8202854E+00	0.2764246E+00	0.2472417E+00
2	0.3044647E+01	0.1988635E+01	0.3188243E+01	0.2690673E+00	0.1490712E+00	0.4823498E+00
3	0.3001880E+01	0.1998866E+01	0.1165668E+00	0.4397377E-01	0.4765407E-01	0.1034366E+01
4	0.3000033E+01	0.1999965E+01	0.1813021E-02	0.2149084E-02	0.2198280E-01	0.1001284E+02
5	0.3000000E+01	0.1999999E+01	0.2684597E-04	0.4737814E-04	0.1856229E-01	0.3846122E+03
6	0.3000000E+01	0.2000000E+01	0.3878641E-06	0.8817947E-06	0.1574341E-01	0.1757351E+05
7	0.3000000E+01	0.2000000E+01	0.5091715E-08	0.1391334E-07	0.1351821E-01	0.9584736E+06
8	0.3000000E+01	0.2000000E+01	0.5957654E-10	0.1884045E-09	0.1182730E-01	0.6203365E+08
9	0.3000000E+01	0.2000000E+01	0.6269307E-12	0.2231274E-11	0.1056873E-01	0.4686828E+10
10	0.3000000E+01	0.2000000E+01	0.8758808E-14	0.2360669E-13	0.9316950E-02	0.3909375E+12

Die Rechnungen zeigen eine superlineare Konvergenz des Newtonverfahrens mit finiten Differenzen gegen das Minimum $(3,2)^\top$.
Das vereinfachte Newtonverfahren liefert das folgende Ergebnis:

ITER	X(1)	X(2)	||GRAD||	||DX||	P=1	P=2
0	0.4000000E+01	0.2500000E+01	0.1351240E+03	0.0000000E+00	0.0000000E+00	0.0000000E+00
1	0.3281417E+01	0.2056664E+01	0.2617493E+02	0.8443389E+00	0.2567589E+00	0.2296521E+00
3	0.3071606E+01	0.1980562E+01	0.5151481E+01	0.6534609E-01	0.5470076E+00	0.4032719E+01
5	0.3022629E+01	0.1985541E+01	0.1403370E+01	0.1757236E-01	0.6120523E+00	0.1394972E+02
7	0.3007754E+01	0.1993240E+01	0.4467921E+00	0.6376896E-02	0.6191372E+00	0.3726386E+02
9	0.3002760E+01	0.1997232E+01	0.1540965E+00	0.2446909E-02	0.6154210E+00	0.9688894E+02
11	0.3001001E+01	0.1998926E+01	0.5514621E-01	0.9306565E-03	0.6121533E+00	0.2552127E+03
13	0.3000366E+01	0.1999593E+01	0.2006926E-01	0.3497360E-03	0.6102365E+00	0.6801452E+03
15	0.3000135E+01	0.1999848E+01	0.7361719E-02	0.1304451E-03	0.6092050E+00	0.1825138E+04
17	0.3000050E+01	0.1999943E+01	0.2710854E-02	0.4845351E-04	0.6086619E+00	0.4915933E+04
19	0.3000018E+01	0.1999979E+01	0.1000183E-02	0.1795865E-04	0.6083767E+00	0.1326686E+05
21	0.3000007E+01	0.1999992E+01	0.3693921E-03	0.6648483E-05	0.6082266E+00	0.3584082E+05
23	0.3000003E+01	0.1999997E+01	0.1364969E-03	0.2459848E-05	0.6081473E+00	0.9687759E+05
25	0.3000001E+01	0.1999999E+01	0.5045188E-04	0.9098193E-06	0.6081053E+00	0.2619347E+06
...						
61	0.3000000E+01	0.2000000E+01	0.8472333E-12	0.1511068E-13	0.6121321E+00	0.1587489E+14
63	0.3000000E+01	0.2000000E+01	0.3142520E-12	0.5625244E-14	0.6151352E+00	0.4292624E+14
65	0.3000000E+01	0.2000000E+01	0.1217558E-12	0.2090737E-14	0.6172799E+00	0.1154325E+15
66	0.3000000E+01	0.2000000E+01	0.7240833E-13	0.1287557E-14	0.6201737E+00	0.1878784E+15

Sogar das vereinfachte Newtonverfahren konvergiert, allerding nur linear.

3.9.1 Globalisierung des Newton-Verfahrens

Wie Satz 3.9.8 erkennen läßt, ist das lokale Newtonverfahren i.a. nicht global (d.h. für jeden Startpunkt $x^{[0]}$) konvergent. Es gibt verschiedene Möglichkeiten, um das lokale Newton-Verfahren durch Zusatzmaßnahmen zu einem global konvergenten Verfahren zu erweitern. Um die quadratische Konvergenz des lokalen Newtonverfahrens zu sichern, hat man dabei darauf zu achten, daß das globale Newtonverfahren in der Nähe der Lösung in das lokale Verfahren (also mit Schrittweite $\alpha_i = 1$) übergeht.
Geiger und Kanzow [GK99] schlagen folgende globale Variante vor:

Algorithmus: globales Newtonverfahren

(i) Wähle einen Startvektor $x^{[0]} \in \mathbb{R}^n$, Konstanten $\sigma \in (0, 1/2)$, $\beta \in (0, 1)$, $\varrho > 0$, $p > 2$ und setze $i = 0$.

(ii) Falls ein Abbruchkriterium erfüllt ist, STOP.

(iii) Berechne (falls möglich) die Suchrichtung $d^{[i]}$ als Lösung von (3.13) mit $x = x^{[i]}$. Besitzt das Gleichungssystem keine Lösung oder ist die Bedingung

$$\nabla f(x^{[i]})^\top d^{[i]} \leq -\varrho \|d^{[i]}\|^p$$

nicht erfüllt, so setze $d^{[i]} = -\nabla f(x^{[i]})$.

(iv) Bestimme eine Schrittweite $\alpha_i > 0$ mit dem Armijo-Verfahren.

(v) Setze $x^{[i+1]} = x^{[i]} + \alpha_i d^{[i]}$, $i := i + 1$ und gehe zu (ii).

Unter der Voraussetzung, daß f zweimal stetig differenzierbar ist, zeigen Geiger und Kanzow [GK99] auf S. 87, daß jeder Häufungspunkt der Folge $\{x^{[i]}\}$ ein stationärer Punkt von f ist. Darüber hinaus läßt sich der folgende Konvergenzsatz beweisen, den wir hier ohne Beweis aus Geiger und Kanzow [GK99], S. 92 zitieren.

Satz 3.9.12 (globale Konvergenz)
Sei $f : \mathbb{R}^n \to \mathbb{R}$ zweimal stetig differenzierbar und $\{x^{[i]}\}$ eine durch das globale Newtonverfahren erzeugte Folge. Für einen Häufungspunkt $\hat{x}$ der Folge mit positiv definiter Hessematrix $\nabla^2 f(\hat{x})$ gelten:

(a) Die gesamte Folge $\{x^{[i]}\}$ konvergiert gegen $\hat{x}$ und $\hat{x}$ ist striktes lokales Minimum von f.

(b) Für hinreichend großes $i \in \mathbb{N}$ ist $d^{[i]}$ stets Newton-Richtung und die Schrittweite erfüllt $\alpha_i = 1$.

(c) $\{x^{[i]}\}$ konvergiert superlinear gegen $\hat{x}$.

(d) Ist $\nabla^2 f$ sogar (lokal) Lipschitz-stetig, so konvergiert $\{x^{[i]}\}$ quadratisch gegen $\hat{x}$.

Anstatt eines Beweises wollen wir motivieren, warum bzw. ob die Schrittweite $\alpha_i = 1$ gegen Ende akzeptiert wird. Wegen der vorausgesetzten positiven Definitheit der Hessematrix $\nabla^2 f$ im stationären Punkt $\hat{x}$ verhält sich f lokal wie eine quadratische Funktion. Wir beschränken uns daher auf die Untersuchung der quadratischen Funktion

$$f(x) = \frac{1}{2} x^\top A x, \qquad A \in \mathbb{R}^{n \times n} \text{ symmetrisch, positiv definit.}$$

Als Newton-Richtung in x ergibt sich

$$Ad = -Ax \quad \Rightarrow \quad d = -x.$$

Als neue Iterierte erhält man

$$x_{neu} = x + \alpha d = (1 - \alpha)x.$$

Das Minimum wird in $\hat{x} = 0$ angenommen, folglich gilt

$$\frac{\|x^{[i+1]} - \hat{x}\|}{\|x^{[i]} - \hat{x}\|} = |1 - \alpha_i|.$$

Also: Die Konvergenz ist genau dann **superlinear**, wenn $\alpha_i \to 1$ für $i \to \infty$.
Wir untersuchen die bisherigen Schrittweitenstrategien auf die Wahl $\alpha_i = 1$ bzw. $\alpha_i \to 1$ für $i \to \infty$:

$$\begin{aligned} \varphi(\alpha) &= f(x + \alpha d) = (1 - \alpha)^2 \varphi(0), \\ \varphi'(0) &= -2(1 - \alpha)\varphi(0). \end{aligned}$$

Damit folgt:

- Armijo:

$$\begin{aligned} & \varphi(\alpha) \le \varphi(0) + \sigma \cdot \alpha \cdot \varphi'(0) \\ \Leftrightarrow \quad & (1 - \alpha)^2 \varphi(0) \le \varphi(0) + \sigma \cdot \alpha(-2\varphi(0)) \\ \Leftrightarrow \quad & (1 - \alpha)^2 \le 1 - 2\sigma\alpha \\ \overset{\alpha \ge 0}{\Leftrightarrow} \quad & \alpha \le 2(1 - \sigma) \end{aligned}$$

- Goldstein:

$$\begin{aligned} & \varphi(\alpha) \geq \varphi(0) + (1-\sigma)\cdot\alpha\cdot\varphi'(0) \\ \Leftrightarrow\quad & (1-\alpha)^2\varphi(0) \geq \varphi(0) + (1-\sigma)\cdot\alpha(-2\varphi(0)) \\ \Leftrightarrow\quad & (1-\alpha)^2 \geq 1 - 2(1-\sigma)\alpha \\ \overset{\alpha\geq 0}{\Leftrightarrow}\quad & \alpha \geq 2\sigma \end{aligned}$$

- Wolfe-Powell:

$$\begin{aligned} & \varphi'(\alpha) \geq \varrho\varphi'(0) \\ \Leftrightarrow\quad & -2(1-\alpha)\varphi(0) \geq -2\varrho\varphi(0) \\ \Leftrightarrow\quad & \alpha \geq 1-\varrho \end{aligned}$$

- Strenge Wolfe-Powell:

$$\begin{aligned} & |\varphi'(\alpha)| \leq -\varrho\varphi'(0) \\ \Leftrightarrow\quad & -2\varrho\varphi(0) \leq -2(1-\alpha)\varphi(0) \leq 2\varrho\varphi(0) \\ \Leftrightarrow\quad & 1-\varrho \leq \alpha \leq 1+\varrho \end{aligned}$$

Daraus schließen wir:
Armijo-Bedingung:
Die Bedingung $\alpha \leq 2(1-\sigma)$ wird wegen $0 < \sigma < 1/2$ von $\alpha = 1$ erfüllt.

Goldstein-Bedingung:
Die Bedingungen $2\sigma \leq \alpha \leq 2(1-\sigma)$ werden wegen $0 < \sigma < 1/2$ von $\alpha = 1$ erfüllt.

Wolfe-Powell-Bedingung:
Die Bedingungen $1-\varrho \leq \alpha \leq 2(1-\sigma)$ werden wegen $0 < \sigma < 1/2$ und $\sigma < \varrho < 1$ von $\alpha = 1$ erfüllt.

Strenge Wolfe-Powell-Bedingung:
Die Bedingungen $1-\varrho \leq \alpha \leq \min\{1+\varrho, 2(1-\sigma)\}$ werden wegen $0 < \sigma < 1/2$ und $\sigma < \varrho < 1$ von $\alpha = 1$ erfüllt.

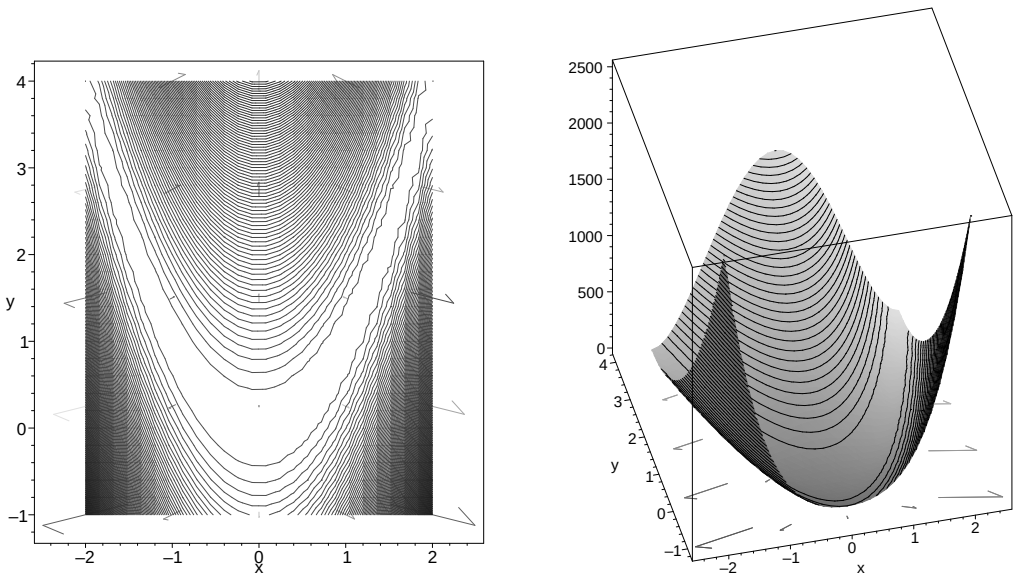
Fazit:
Alle Schrittweitenstrategien ermöglichen die Wahl $\alpha_i = 1$. Allerdings muß man durch zusätzliche Maßnahmen im Algorithmus dafür sorgen, daß auch tatsächlich $\alpha_i = 1$ gewählt wird.

Beispiel 3.9.13 **(Funktion von Rosenbrock)**

Wir minimieren

$$f(x,y) = 100(y-x^2)^2 + (1-x)^2$$

mit dem globalen Newtonverfahren. f hat ein globales Minimum in $(1,1)^\top$ mit Funktionswert 0.



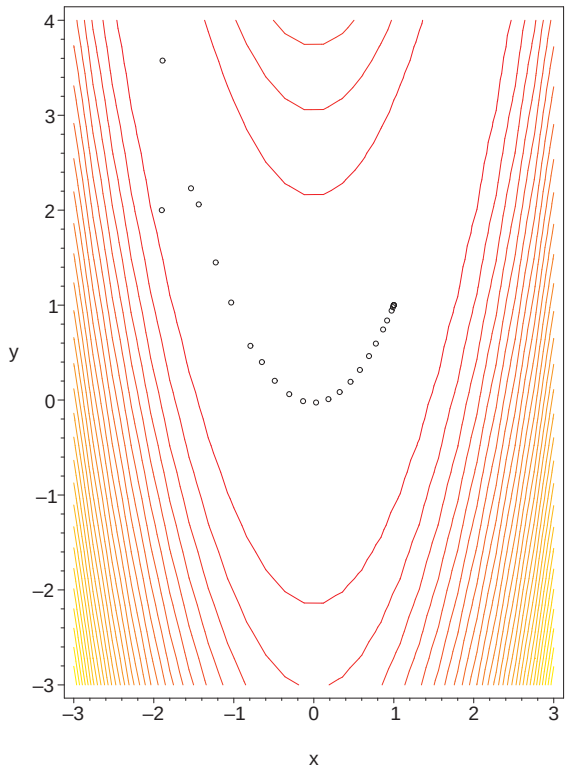
Die Hessematrix und deren Eigenwerte im Lösungspunkt lauten

$$\nabla^2 f(1,1) = \begin{pmatrix} 802 & -400 \\ -400 & 200 \end{pmatrix}, \qquad \lambda_1 \approx 1001.6,\ \lambda_2 \approx 0.3994.$$

Darüber hinaus ist die Funktion (als Polynom) beliebig oft differenzierbar. Alle Voraussetzungen des Konvergenzsatzes für das globale Newtonverfahren sind erfüllt. Das globale Newtonverfahren mit den Parametern $\sigma = 0.01$, $\beta = 0.5$, $\varrho = 0.01$, $p = 3$, $\varepsilon = 10^{-13}$ und Abbruchkriterium $\|\nabla f(x)\| \le \varepsilon$ liefert für den Startpunkt $x^{[0]} = (-1.9, 2)^\top$ das folgende Ergebnis:

ITER	X(1)	X(2)	||GRAD||	||DX||	P=1	P=2	ALPHA	F
0	-0.1900000E+01	0.2000000E+01	0.1270869E+04	0.0000000E+00	0.0000000E+00	0.0000000E+00	0.0000000E+00	0.2676200E+03
1	-0.1891022E+01	0.3575882E+01	0.5843040E+01	0.1575908E+01	0.1262269E+01	0.4114880E+00	0.1000000E+01	0.8358007E+01
2	-0.1535378E+01	0.2230831E+01	0.8657596E+02	0.1113020E+02	0.7278599E+00	0.1879754E+00	0.1250000E+00	0.8029711E+01
3	-0.1439014E+01	0.2061477E+01	0.1039035E+02	0.1948509E+00	0.9438102E+00	0.3348806E+00	0.1000000E+01	0.5957414E+01
4	-0.1225603E+01	0.1449594E+01	0.3196743E+02	0.2592127E+01	0.8535986E+00	0.3209034E+00	0.2500000E+00	0.5229027E+01
5	-0.1032102E+01	0.1027792E+01	0.2090905E+02	0.4640683E+00	0.8950620E+00	0.3942032E+00	0.1000000E+01	0.4269634E+01
6	-0.7927080E+00	0.5710765E+00	0.2459178E+02	0.5156536E+00	0.9070084E+00	0.4462983E+00	0.1000000E+01	0.3542240E+01
7	-0.6488530E+00	0.4003159E+00	0.9606086E+01	0.2232789E+00	0.9518327E+00	0.5163726E+00	0.1000000E+01	0.2761541E+01
8	-0.4884230E+00	0.2024721E+00	0.1235395E+02	0.5094308E+00	0.9624431E+00	0.5485510E+00	0.5000000E+00	0.2345615E+01
9	-0.3072831E+00	0.6161129E-01	0.9340990E+01	0.2294633E+00	0.9529718E+00	0.5643479E+00	0.1000000E+01	0.1816650E+01
10	-0.1344153E+00	-0.1181580E-01	0.7123217E+01	0.1878159E+00	0.9446169E+00	0.5870059E+00	0.1000000E+01	0.1376199E+01
11	0.2818625E-01	-0.2564481E-01	0.5537978E+01	0.1631886E+00	0.9295046E+00	0.6114805E+00	0.1000000E+01	0.1014325E+01
12	0.1827404E+00	0.9507067E-02	0.4778695E+01	0.1585012E+00	0.9088423E+00	0.6432327E+00	0.1000000E+01	0.7249721E+00
13	0.3241985E+00	0.8509427E-01	0.4190771E+01	0.1603865E+00	0.8857638E+00	0.6897774E+00	0.1000000E+01	0.4967493E+00
14	0.4593025E+00	0.1927057E+00	0.4299926E+01	0.1727232E+00	0.8542343E+00	0.7510176E+00	0.1000000E+01	0.3256713E+00
15	0.5755661E+00	0.3177591E+00	0.3525688E+01	0.1707500E+00	0.8269453E+00	0.8510849E+00	0.1000000E+01	0.1984157E+00
16	0.6901713E+00	0.4632021E+00	0.3992294E+01	0.1851702E+00	0.7713783E+00	0.9600343E+00	0.1000000E+01	0.1132449E+00
17	0.7755972E+00	0.5942535E+00	0.2329187E+01	0.1564355E+00	0.7480969E+00	0.1207007E+01	0.1000000E+01	0.5568209E-01
18	0.8668357E+00	0.7430797E+00	0.3104276E+01	0.1745672E+00	0.6241119E+00	0.1346035E+01	0.1000000E+01	0.2466239E-01
19	0.9168056E+00	0.8380355E+00	0.9004797E+00	0.1073013E+00	0.6292140E+00	0.2174353E+01	0.1000000E+01	0.7544806E-02
20	0.9722908E+00	0.9422709E+00	0.1297329E+01	0.1180832E+00	0.3516811E+00	0.1931445E+01	0.1000000E+01	0.1715585E-02
21	0.9894405E+00	0.9786985E+00	0.1119780E+00	0.4026267E-01	0.3712850E+00	0.5798178E+01	0.1000000E+01	0.1201525E-03
22	0.9994134E+00	0.9987276E+00	0.4341178E-01	0.2237466E-01	0.5893078E-01	0.2478671E+01	0.1000000E+01	0.1333310E-05
23	0.9999886E+00	0.9999768E+00	0.1278973E-03	0.1375217E-02	0.1847117E-01	0.1318345E+02	0.1000000E+01	0.1418487E-09
24	0.1000000E+01	0.1000000E+01	0.5718384E-07	0.2587793E-04	0.6996609E-04	0.2703508E+01	0.1000000E+01	0.2286153E-17
25	0.1000000E+01	0.1000000E+01	0.0000000E+00	0.1810704E-08	0.0000000E+00	0.0000000E+00	0.1000000E+01	0.0000000E+00

Man sieht sehr schön die quadratische Konvergenz in den letzten 4 Iterationen. Davor liegt nur lineare Konvergenz vor. Die folgende Abbildung zeigt die Lage der Iterationspunkte $x^{[i]}$.



3.9.2 Modifikation der Newton-Richtung

Welche Möglichkeiten gibt es, die Newton-Richtung abzuändern, wenn diese keine Abstiegsrichtung ist? Fletcher [Fle03] beschreibt auf den Seiten 47/48 folgende Ansätze.

(a) **Ansatz von Goldstein und Prince:**

Teste, ob die Newton-Richtung $d = -\nabla^2 f(x)^{-1} \nabla f(x)$ die Winkelbedingung

$$-\frac{\nabla f(x)^\top d}{\|\nabla f(x)\| \cdot \|d\|} \geq \nu$$

für ein fest vorgegebenes $\nu > 0$ erfüllt. Ist dies nicht der Fall, so wähle $d = -\nabla f(x)$ als Suchrichtung.

Kombiniert man dieses Verfahren mit einer effizienten Schrittweite, so konvergiert das Abstiegsverfahren nach Satz 3.5.10.

Nachteil: Eine häufige Gradientenwahl zerstört i.a. die superlineare (oder quadratische) Konvergenz.

(b) **Ansatz von Levenberg, Marquardt:** (vgl. Trust-Region-Verfahren)

Ist $\nabla^2 f(x)$ nicht positiv definit, so bestimme ein $\lambda > 0$, so daß die Matrix $\nabla^2 f(x) + \lambda I$ positiv definit ist und berechne die Suchrichtung d aus

$$\big(\nabla^2 f(x) + \lambda I\big)\, d = -\nabla f(x).$$

Beachte, daß die Eigenwerte der symmetrischen Matrix $\nabla^2 f(x) + \lambda I$ durch Spektralverschiebung der Eigenwerte von $\nabla^2 f(x)$ gegeben sind. Für hinreichend großes $\lambda > 0$ werden also alle Eigenwerte ins Positive verschoben.

(c) **Ansatz von Murray, Hebden, Gill, Picken:**

Die folgende Idee basiert auf dem Ansatz von Levenberg und Marquardt und beschreibt ein modifiziertes Cholesky-Verfahren, falls $A = \nabla^2 f(x)$ nicht positiv definit ist.

Zur Erinnerung: Die Cholesky-Zerlegung einer Matrix A existiert genau dann, wenn A symmetrisch und positiv definit ist. Sie lautet

$$A = L \cdot L^\top$$

mit einer linken unteren Dreicksmatrix L. Ein Gleichungssystem $Ax = L \cdot L^\top x = b$ kann dann mittels Vorwärts-Ruckwärtssubstitution gelöst werden (d.h. löse erst $Ly = b$ und dann $L^\top x = y$).

Das Verfahren lautet im Detail:

```
for k = 1 : N
    l_kk = sqrt(a_kk − Σ_{j=1}^{k−1} l_kj^2)
    for i = k + 1 : N
        l_ik = (a_ik − Σ_{j=1}^{k−1} l_ij l_kj) / l_kk
    end
end
```

Ist nun A nicht positiv definit, so wird der obige Algorithmus mit $l_{kk} \le 0$ abbrechen. Um dieses zu vermeiden, wird der Algorithmus mit Hilfe einer Konstanten $\mu > 0$ wie folgt modifiziert:

```
for k = 1 : N
    l_kk = { sqrt(a_kk − Σ_{j=1}^{k−1} l_kj^2),   falls a_kk − Σ_{j=1}^{k−1} l_kj^2 > μ,
           { sqrt(μ),                             sonst
    for i = k + 1 : N
        l_ik = (a_ik − Σ_{j=1}^{k−1} l_ij l_kj) / l_kk
    end
end
```

Dieser Algorithmus läuft durch, auch wenn A nicht positiv definit ist. Allerdings wird nicht mehr A zerlegt, sondern es gilt

$$A + D = L \cdot L^\top,$$

wobei D eine Diagonalmatrix mit nicht-negativen Diagonalelementen ist, vgl. Geiger und Kanzow [GK99], S. 95.

Da die Diagonalelemente von L per Konstruktion positiv sind, ist L und damit $A+D$ positiv definit. Berechnet man nun eine Suchrichtung d als Lösung des modifizierten Gleichungssystems

$$L \cdot L^\top d = -\nabla f(x),$$

so ist d zumindest eine Abstiegsrichtung.

Vorteil: Man verspricht sich von diesem Verfahren (und kann dieses numerisch oft beobachten), daß die Wahl von $d = -\nabla f(x)$ im globalen Newton-Verfahren weniger häufig erfolgt.

3.9.3 Modifikationen an Sattelpunkten

Die bisherigen Abstiegsverfahren bekommen Probleme in der Nähe von Sattelpunkten oder lokalen Maxima. Dort gilt bekanntlich ebenfalls $\nabla f(x) = 0$ und damit $\nabla f(x)^\top d = 0$ für alle $d \in \mathbb{R}^n$. Andererseits gibt es in beiden Fällen Abstiegsrichtungen gemäß Definition 3.5.1.
Da diese Abstiegsrichtungen nicht über den Gradienten von f charakterisiert werden können, ist es naheliegend, die Krümmung der Funktion f ins Spiel zu bringen. Ein Maß für die Krümmung ist die zweite Ableitung von f, also die Hessematrix.
Wir setzen voraus, daß die Hessematrix von f in einem Sattelpunkt bzw. in einem Maximum mindestens einen negativen Eigenwert besitzt. Es gibt daher eine Richtung d mit $d^\top \nabla^2 f(x) d < 0$.

Definition 3.9.14

$d \in \mathbb{R}^n$, $d \neq 0$ *mit* $d^\top \nabla^2 f(x) d < 0$ *heißt* **Richtung negativer Krümmung von** f **in** x.

Für eine Richtung d negativer Krümmung in einem stationären Punkt x gilt nach Taylorentwicklung

$$f(x + \alpha d) = f(x) + \alpha \underbrace{\nabla f(x)}_{=0} d + \frac{\alpha^2}{2} \underbrace{d^\top \nabla^2 f(x) d}_{<0} + o(\alpha^2).$$

Somit existiert ein $\bar{\alpha} > 0$ mit

$$f(x + \alpha d) < f(x) \qquad \forall \alpha \in (0, \bar{\alpha}].$$

d ist also Abstiegsrichtung.

Zur Berechnung einer Richtung negativer Krümmung schlagen Fiacco und McCormick [FM90] auf S. 167 folgende Modifikation des Newtonverfahrens vor.

Zunächst versucht man, die Cholesky-Zerlegung von $\nabla^2 f(x)$ zu erzeugen. Gelingt dies, so wähle die Newtonrichtung $d = -\nabla^2 f(x)^{-1} \nabla f(x)$.

Andernfalls, da $\nabla^2 f(x)$ symmetrisch ist, ist $\nabla^2 f(x)$ diagonalisierbar, d.h. es gibt eine Zerlegung

$$\nabla^2 f(x) = T \cdot \Lambda \cdot T^{-1},$$

wobei T orthogonal (d.h. $T^\top = T^{-1}$) und $\Lambda = \operatorname{diag}(\lambda_1, \ldots, \lambda_n)$ Diagonalmatrix ist. T enthält eine Basis aus Eigenvektoren und Λ die Eigenwerte von $\nabla^2 f(x)$. Existiert ein negativer Eigenwert, so enthält Λ auch negative Diagonalelemente $\lambda_i < 0$.

Löse nun das lineare Geichungssystem

$$T^{-1} d = a, \qquad a = (a_1, \ldots, a_n)^\top,\ a_i = \begin{cases} 1, & \text{falls } \lambda_i \le 0, \\ 0, & \text{sonst.} \end{cases} \tag{3.17}$$

Man prüft leicht nach, daß gilt

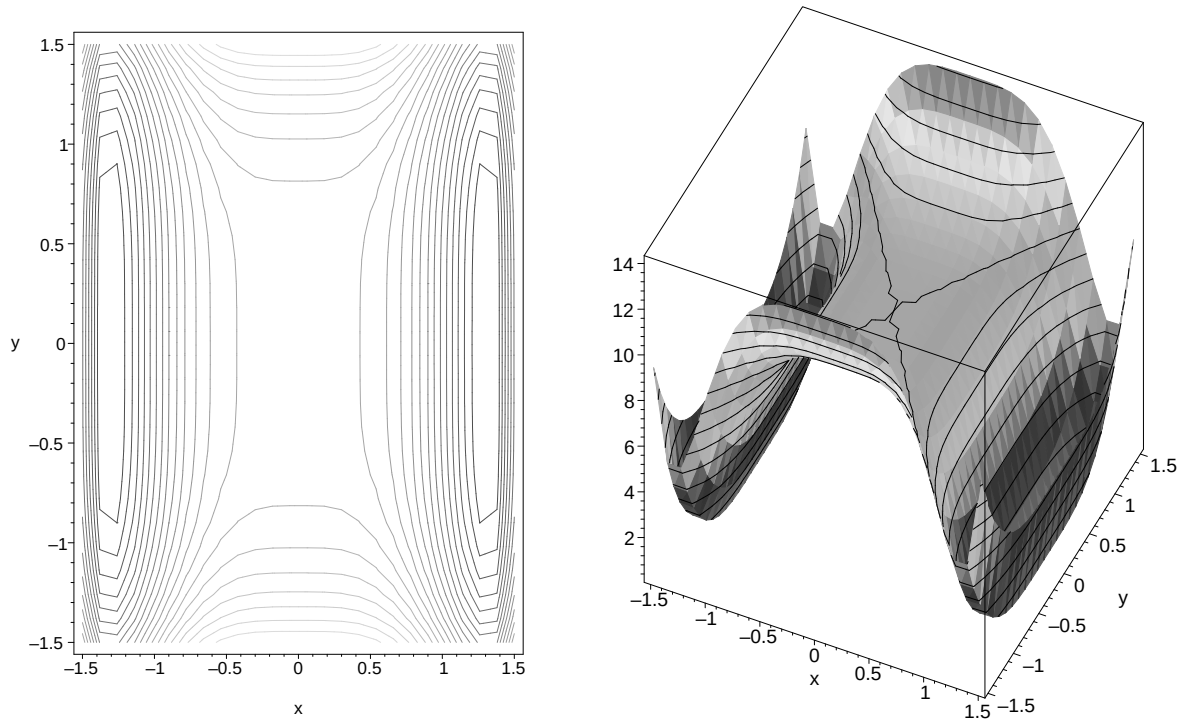
$$d^\top \nabla^2 f(x) d = d^\top T \Lambda T^{-1} d = (T^{-1} d)^\top \Lambda T^{-1} d = a^\top \Lambda a = \sum_{\lambda_j \le 0} \lambda_j < 0.$$

Somit ist das so bestimmte d tatsächlich eine Richtung negativer Krümmung und man kann das Abstiegsverfahren fortsetzen.

Beispiel 3.9.15 **(vgl. Fiacco und McCormick [FM90], S. 167/168)**
Die Funktionen

$$\begin{aligned} f(x) &= 100(x_2 - x_1^2)^2 + (1 - x_1)^2 + 90(x_4 - x_3^2)^2 + (1 - x_3)^2 + 10.1\left((x_2 - 1)^2 + (x_4 - 1)^2\right) \\ &\quad +19.8(x_2 - 1)(x_4 - 1), \\ g(x) &= (x_1^4 - 3)^2 + x_2^4 \end{aligned}$$

sind Testfunktionen für numerische Verfahren. Sie besitzen Sattelpunkte, in denen Verfahren häufig stecken bleiben, vgl. Abbildung für die Funktion g.



Bemerkung 3.9.16

Es ist ratsam, bereits in der Nähe eines Sattelpunktes Richtungen negativer Krümmung zu berücksichtigen, etwa in der Form

$$x^{[i+1]} = x^{[i]} + \alpha d_N^{[i]} + \alpha^2 d_K^{[i]},$$

wobei $d_N^{[i]}$ die Newton-Richtung und $d_K^{[i]}$ eine Richtung negativer Krümmung bezeichnen. Die Richtung $d_K^{[i]}$ ist durch d aus (3.17) gegeben, falls $\nabla f(x^{[i]})^\top d \leq 0$ gilt, andernfalls durch $-d$.

3.9.4 Berechnung von Ableitungen

In der Regel sind die in Anwendungen auftretenden Funktionen zu komplex, um deren Ableitungen analytisch angeben zu können. Die folgenden Varianten können häufig alternativ genutzt werden:

- Programme wie MAPLE oder MATLAB (Symbolic Toolbox) erlauben es, Ableitungen **symbolisch** zu berechnen.

  In MAPLE stehen hierfür der Befehl `diff` bzw. der Operator `D` zur Verfügung. Zur Berechnung des Gradienten bzw. der Hessematrix dienen die Befehle `grad` bzw. `hessian`.

  Beispiele:

```
f := (x^2+y-11)^2 + (x+y^2-7)^2;
```

```
diff(f,x);    % partielle Ableitung nach x
diff(f,y);    % partielle Ableitung nach y

g := (x,y) -> (x^2+y-11)^2 + (x+y^2-7)^2;
D[1](g);      % partielle Ableitung nach x
D[2](g);      % partielle Ableitung nach y

with(linalg):
grad(f,vector([x,y]));
grad(g(x,y),vector([x,y]));

hessian(f,vector([x,y]));
hessian(g(x,y),vector([x,y]));
```

- Die Auswertung einer von x abhängigen Funktion liege in Form eines Fortran- oder C-Programms vor. Beim **automatischen Differenzieren** wird das komplette Programm (d.h. die Anweisungsfolge des Programms) algorithmisch differenziert. Man erhält wieder ein Fortran- oder C-Programm, welches dann z.B. den Gradienten von f abhängig von den Eingabedaten x liefert. Im wesentlichen basiert das automatische Differenzieren auf der Anwendung von Produkt- und Kettenregel, gekoppelt mit einem Parser, der Fortran- oder C-Programme interpretieren kann.

  Weitere Hinweise und Software finden sich unter `www.autodiff.org`.

- Bei der **numerischen Differentiation** werden geeignete finite-Differenzen-Schemata zur approximativen Berechnung von Ableitungen verwendet, etwa

$$\frac{\partial f}{\partial x_i}(x) \approx \frac{f(x+he_i)-f(x)}{h},$$

  wobei e_i den i-ten Einheitsvektor bezeichnet. Das folgende Fortran-Programm approximiert die Jacobimatrix einer Funktion $f : \mathbb{R}^n \to \mathbb{R}^m$:

```
      CALL F( X,C )
      DO 10,J=1,N
         TEMP=X(J)
         H=MAX(EPS,EPS*DABS(TEMP))
         DEL=1.0D0/H
         X(J)=X(J)+H
         CALL F( X,W )
         X(J)=TEMP
```

```
        DO 20,I=1,M
           JAC(I,J)=(W(I)-C(I))*DEL
20      CONTINUE
10    CONTINUE
```

Es sei jedoch bemerkt, daß die Berechnung zweiter Ableitungen mittels finiter Differenzen schlecht konditioniert ist.

3.10 Quasi-Newton-Verfahren

Das Newtonverfahren hat im wesentlichen zwei Nachteile:

- Die Hessematrix von f wird benötigt. In der Praxis ist diese in der Regel nicht bekannt und nur schwer zu bestimmen.
- Die Newton-Richtung ist i.a. keine Abstiegsrichtung.

Andererseits besitzt das Newtonverfahren sehr gute (lokale) Konvergenzeigenschaften (mindestens superlinear!).
Quasi-Newton-Verfahren (**Variable-Metrik-Verfahren**) versuchen nun, die Nachteile des Newtonverfahrens zu vermeiden und dabei die Konvergenzgeschwindigkeit zu retten.
Die Idee der Quasi-Newton-Verfahren ist es, die Hessematrix $\nabla^2 f(x^{[i]})$ durch eine Matrix

$$H_i \approx \nabla^2 f(x^{[i]})$$

bzw. die Inverse Hessematrix $\nabla^2 f(x^{[i]})^{-1}$ durch eine Matrix

$$B_i \approx \nabla^2 f(x^{[i]})^{-1}$$

zu ersetzen. Wesentlich dabei ist, daß für H_i bzw. B_i von Schritt zu Schritt ein einfach zu berechnender **Update** durchgeführt wird. Wie wir später sehen werden, werden üblicherweise **Update-Formeln** der Form

$$H_{i+1} = \Phi(H_i, d^{[i]}, y^{[i]}) \qquad \text{bzw.} \qquad B_{i+1} = \Psi(B_i, d^{[i]}, y^{[i]})$$

mit $d^{[i]} = x^{[i+1]} - x^{[i]}$ und $y^{[i]} = \nabla f(x^{[i+1]}) - \nabla f(x^{[i]})$ verwendet. Dabei werden H_i bzw. B_i i.a. so konstruiert, daß sie **symmetrisch** und **positiv definit** sind (Abstiegsrichtung!). Dies führt dann zu sehr effizienten Verfahren mit guten Konvergenzeigenschaften.
Es gibt dabei zwei Klassen von Quasi-Newton-Verfahren:

(i) **Variante 1:** An Stelle des linearen Gleichungssystems

$$\nabla^2 f(x^{[i]}) d^{[i]} = -\nabla f(x^{[i]})$$

zur Bestimmung der Newton-Richtung wird das Gleichungssystem

$$H_i d^{[i]} = -\nabla f(x^{[i]}) \tag{3.18}$$

zur Bestimmung der Quasi-Newton-Richtung $d^{[i]}$ gelöst.

(ii) **Variante 2:** Anstatt die Newton-Richtung über

$$d^{[i]} = -\nabla^2 f(x^{[i]})^{-1} \cdot \nabla f(x^{[i]})$$

zu berechnen, wird die Quasi-Newton-Richtung durch

$$d^{[i]} = -B_i \cdot \nabla f(x^{[i]}), \tag{3.19}$$

bestimmt. Beachte, daß hierbei nur eine Matrix-Vektor-Multiplikation anfällt.

Im Idealfall gilt $H_i = B_i^{-1}$ und beide Varianten fallen formal zusammen.
Dies führt auf den folgenden konzeptionellen Algorithmus:

Algorithmus: Quasi-Newton-Verfahren

(i) Wähle einen Startvektor $x^{[0]} \in \mathbb{R}^n$ und eine symmetrische, positiv definite
Startmatrix H_0 (bzw. B_0; z.B. die Einheitsmatrix) und setze $i = 0$.

(ii) Falls ein Abbruchkriterium erfüllt ist, STOP.

(iii) Berechne die Suchrichtung $d^{[i]}$ gemäß (3.18) (bzw. (3.19)).

(iv) Bestimme eine Schrittweite $\alpha_i > 0$ durch eine Schrittweitenstrategie.

(v) Setze $x^{[i+1]} = x^{[i]} + \alpha_i d^{[i]}$, berechne den Update H_{i+1} (bzw. B_{i+1}),
setze $i := i + 1$, und gehe zu (ii).

3.10.1 Konstruktion von Update-Formeln

Wir betrachten zunächst wieder das nichtlineare Gleichungssystem

$$g(x) = 0 \quad \text{mit} \quad g(x) = \nabla f(x).$$

Es sei jedoch bemerkt, daß die folgenden Betrachtungen für allgemeine nichtlineare Gleichungssysteme gelten, die nicht notwendig durch Minimierungsprobleme motiviert sind.
Unter den Voraussetzungen von Satz 3.9.8 erhielten wir lokal superlineare bzw. quadratische Konvergenz der Folge $\{x^{[i]}\}$, die durch

$$\nabla^2 f(x^{[i]}) \left(x^{[i+1]} - x^{[i]}\right) = -\nabla f(x^{[i]}), \qquad i = 0, 1, \ldots$$

bzw. durch

$$x^{[i+1]} = x^{[i]} - \nabla^2 f(x^{[i]})^{-1} \nabla f(x^{[i]}), \qquad i = 0, 1, 2, \ldots.$$

definiert ist.

Wir ersetzen nun die Hessematrix wie oben angedeutet durch die Matrix H_i bzw. die Inverse durch B_i und erhalten die Iterationsvorschriften

$$H_i \left(x^{[i+1]} - x^{[i]} \right) = -\nabla f(x^{[i]}), \qquad i = 0, 1, 2, \ldots, \tag{3.20}$$

bzw.

$$x^{[i+1]} = x^{[i]} - B_i \nabla f(x^{[i]}), \qquad i = 0, 1, 2, \ldots. \tag{3.21}$$

Für die folgenden Betrachtungen gelte $H_i = B_i^{-1}$, so daß wir uns auf die erste Variante beschränken können.

Welche Eigenschaften müssen die Matrizen H_i (bzw. B_i) besitzen, damit dieses Verfahren ähnlich schnell konvergiert wie das Newtonverfahren (also mindestens superlinear)?

Der folgende Satz charakterisiert die superlineare Konvergenz (und gilt entsprechend für allgemeine nichtlineare Gleichungssysteme). Hierbei verwenden wir die Abkürzungen

$$d^{[i]} := x^{[i+1]} - x^{[i]}, \qquad y^{[i]} := \nabla f(x^{[i+1]}) - \nabla f(x^{[i]}).$$

Satz 3.10.1 (Dennis, Moré)

Es sei $D \subseteq \mathbb{R}^n$ konvex und offen und $f : D \to \mathbb{R}$ zweimal stetig differenzierbar. $\hat{x} \in D$ sei stationärer Punkt von f und $\nabla^2 f(\hat{x})$ sei invertierbar. Desweiteren erfülle $\nabla^2 f(\cdot)$ die Lipschitz-Bedingung

$$\|\nabla^2 f(x) - \nabla^2 f(y)\| \le L \cdot \|x - y\| \qquad \forall x, y \in D.$$

Die Folge $\{x^{[i]}\}$ sei durch (3.20) definiert und es gelte

$$\lim_{i \to \infty} x^{[i]} = \hat{x}, \qquad x^{[i]} \neq \hat{x}, \qquad x^{[i]} \in D \qquad \forall i.$$

Schließlich seien die Matrizen H_i für alle i invertierbar.
Dann sind äquivalent:

(a) Die Folge $\{x^{[i]}\}$ ist superlinear konvergent, d.h. es gilt

$$\lim_{i \to \infty} \frac{\|x^{[i+1]} - \hat{x}\|}{\|x^{[i]} - \hat{x}\|} = 0.$$

(b) Es gilt

$$\lim_{i \to \infty} \frac{\| \left(H_i - \nabla^2 f(\hat{x}) \right) d^{[i]} \|}{\|d^{[i]}\|} = 0. \tag{3.22}$$

(c) Es gilt

$$\lim_{i\to\infty} \frac{\|H_i d^{[i]} - y^{[i]}\|}{\|d^{[i]}\|} = 0. \tag{3.23}$$

Beweis: (vgl. Jarre und Stoer [JS04], S. 175)
Wir zeigen nur die Äquivalenz von (a) und (c). Wegen

$$y^{[i]} = \nabla f(x^{[i+1]}) - \nabla f(x^{[i]}) \approx \nabla^2 f(\hat{x}) d^{[i]}$$

ist damit auch (b) plausibel.

(i) $(a) \Rightarrow (c)$:
Gleichung (3.20) liefert $H_i d^{[i]} = -g(x^{[i]})$. Aus der Definition von $y^{[i]}$ folgt

$$g(x^{[i+1]}) = y^{[i]} + g(x^{[i]}) = y^{[i]} - H_i d^{[i]}.$$

Der Mittelwertsatz in Integralform liefert

$$g(x^{[i+1]}) = g(x^{[i+1]}) - g(\hat{x}) = \int_0^1 g'(\hat{x} + t(x^{[i+1]} - \hat{x}))\left(x^{[i+1]} - \hat{x}\right) dt = G_i\left(x^{[i+1]} - \hat{x}\right)$$

mit

$$G_i = \int_0^1 g'(\hat{x} + t(x^{[i+1]} - \hat{x}))dt.$$

Aus der Stetigkeit von $g'(\cdot)$ folgt mit $\lim x^{[i]} = \hat{x}$ sofort $G_i \to g'(\hat{x})$. Insbesondere ist $\|G_i\| \le c$ für eine von i unabhängige Konstante c. Es folgt

$$\|g(x^{[i+1]})\| \le c\|x^{[i+1]} - \hat{x}\|.$$

Wegen (a) gilt

$$c_i := \frac{\|x^{[i+1]} - \hat{x}\|}{\|x^{[i]} - \hat{x}\|} \to 0.$$

Damit gilt

$$\|d^{[i]}\| = \|x^{[i+1]} - \hat{x} + \hat{x} - x^{[i]}\| \ge \|x^{[i]} - \hat{x}\| - \|x^{[i+1]} - \hat{x}\| = (1 - c_i)\|x^{[i]} - \hat{x}\|.$$

Zusammen folgt

$$\frac{\|y^{[i]} - H_i d^{[i]}\|}{\|d^{[i]}\|} = \frac{\|g(x^{[i+1]})\|}{\|d^{[i]}\|} \le \frac{c\|x^{[i+1]} - \hat{x}\|}{(1 - c_i)\|x^{[i]} - \hat{x}\|} = \frac{c \cdot c_i}{1 - c_i} \to 0.$$

(ii) $(a) \Leftarrow (c)$:
(c) impliziert

$$s_i := \frac{\|g(x^{[i+1]})\|}{\|d^{[i]}\|} \to 0.$$

Die Inverse einer Matrix hängt stetig von den Komponenten der Matrix ab. Aus $\lim_{i\to\infty} G_i = g'(\hat{x})$ und der Invertierbarkeit von $g'(\hat{x})$ folgen $\lim_{i\to\infty} G_i^{-1} = g'(\hat{x})^{-1}$ und die Beschränktheit von G_i^{-1} gemäß $\|G_i^{-1}\| \le \tilde{c}$ (vgl. Hilfssatz 3.9.9). Dann gilt

$$\begin{aligned}\|x^{[i+1]} - \hat{x}\| &\le \|G_i^{-1}\| \cdot \|g(x^{[i+1]})\| \\ &\le \tilde{c}\|g(x^{[i+1]})\| \\ &= \tilde{c}s_i\|x^{[i+1]} - x^{[i]}\| \\ &\le \tilde{c}s_i\left(\|x^{[i+1]} - \hat{x}\| + \|\hat{x} - x^{[i]}\|\right)\end{aligned}$$

und somit

$$(1 - \tilde{c}s_i)\|x^{[i+1]} - \hat{x}\| \le \tilde{c}s_i\|x^{[i]} - \hat{x}\|.$$

Für große i ist $1 - \tilde{c}s_i > 0$ und es folgt die Behauptung aus

$$\frac{\|x^{[i+1]} - \hat{x}\|}{\|x^{[i]} - \hat{x}\|} \le \frac{\tilde{c}s_i}{1 - \tilde{c}s_i} \to 0.$$

□

Welche Konsequenz hat der Satz für die Wahl von H_i?
Bedingung (c) wäre sofort erfüllt, wenn $H_i d^{[i]} = y^{[i]} = \nabla f(x^{[i+1]}) - \nabla f(x^{[i]})$ gelten würde. Allerdings ist diese Bedingung i.a. nicht erfüllbar, da $d^{[i]}$ durch (3.20) gegeben ist.
Allerdings kann man die neue Matrix H_{i+1} so bestimmen, daß

$$H_{i+1}d^{[i]} = y^{[i]} \tag{3.24}$$

erfüllt ist. Diese Bedingung heißt **Quasi-Newton-Gleichung** oder **Sekantenbedingung**.

Bemerkung 3.10.2

Die Quasi-Newton-Gleichung läßt sich auch wie folgt motivieren (vgl. Alt [Alt02], S. 124). Sei $x^{[i+1]}$ berechnet. Wir betrachten in Analogie zu Beispiel 3.9.2 die quadratische Approximation von f in $x^{[i+1]}$:

$$\hat{f}(y) = f(x^{[i+1]}) + \nabla f(x^{[i+1]})^\top (y - x^{[i+1]}) + \frac{1}{2}(y - x^{[i+1]})^\top H_{i+1}(y - x^{[i+1]}).$$

Es gilt $\hat{f}(x^{[i+1]}) = f(x^{[i+1]})$ und $\nabla\hat{f}(x^{[i+1]}) = \nabla f(x^{[i+1]})$. Damit $\hat{f}$ zumindest in Richtung $d^{[i]}$ eine gute Approximation für f ist, wählen wir H_{i+1} so, daß auch $\nabla\hat{f}(x^{[i]}) = \nabla f(x^{[i]})$ gilt. Wegen

$$\nabla\hat{f}(x^{[i]}) = \nabla f(x^{[i+1]}) + H_{i+1}\left(x^{[i]} - x^{[i+1]}\right) = \nabla f(x^{[i+1]}) - H_{i+1}d^{[i]}$$

folgt dann wieder die Quasi-Newton-Gleichung (3.24).

Im folgenden werden wir folgende Abkürzungen verwenden:

$$x = x^{[i]},\ H = H_i,\ B = B_i,\ x^+ = x^{[i+1]},\ H_+ = H_{i+1},\ B_+ = B_{i+1},\ d = x^+ - x,\ y = y^{[i]}.$$

Es gilt nun eine Matrix H_+ zu erzeugen, die die Quasi-Newton-Gleichung (3.24) erfüllt (es gibt unendlich viele!).

Rang-1 Update von Broyden:

Ein erster Ansatz ergibt sich, wenn man fordert, daß sich H_+ und H nur entlang der Richtung $x^+ - x$ unterscheiden und für alle $y \perp (x^+ - x)$ identisch sind:

$$(H_+ - H)y = 0 \qquad \forall y \perp (x^+ - x).$$

Diese Forderung wird erfüllt durch die Matrix

$$H_+ = H + \frac{ud^\top}{d^\top d}, \qquad u \in \mathbb{R}^n.$$

Der Vektor $u \in \mathbb{R}^n$ wird durch die Quasi-Newton-Bedingung (3.24) festgelegt:

$$y = H_+ d = Hd + u \quad \Rightarrow \quad u = y - Hd.$$

Wir erhalten die Update-Formel

$$H_+ = H + \frac{(y - Hd)d^\top}{d^\top d}, \qquad d = x^+ - x, \quad y = \nabla f(x^+) - \nabla f(x). \tag{3.25}$$

Sie heißt **Broyden-Rang-1-Update**.

Einen interessanten Zusammenhang mit der Optimierung liefert

Satz 3.10.3

H_+ ist die eindeutig bestimmte Lösung des Minimierungsproblems

$$\min_{A \in \mathbb{R}^{n \times n}} \|A - H\|_F \qquad \text{unter} \qquad Ad = y.$$

($\|A\|_F = \sqrt{\sum_{i,j=1}^n |a_{ij}|^2}$ bezeichnet die Frobenius-Norm).

Weiterhin haben Broyden, Dennis und Moré gezeigt:

Satz 3.10.4

Es sei $D \subseteq \mathbb{R}^n$ konvex und offen und $f : D \to \mathbb{R}$ zweimal stetig differenzierbar. $\hat{x} \in D$ sei stationärer Punkt von f und $\nabla^2 f(\hat{x})$ sei invertierbar. Desweiteren erfülle $\nabla^2 f(\cdot)$ die Lipschitz-Bedingung

$$\|\nabla^2 f(x) - \nabla^2 f(y)\| \le L \cdot \|x - y\| \qquad \forall x, y \in D.$$

Dann gibt es ein $r > 0$, so daß das Verfahren (3.20) mit H_i aus (3.25) für alle $x^{[0]} \in U_r(\hat{x})$ und alle $H_0 \in \mathbb{R}^{n \times n}$ mit $\|H_0 - \nabla^2 f(\hat{x})\| \le r$ wohldefiniert ist und superlinear gegen $\hat{x}$ konvergiert.

Es sei bemerkt, daß i.a. **nicht** $H_i \to \nabla^2 f(\hat{x})$ gilt!

Die Rang-1-Update-Formel (3.25) bleibt allerdings primär auf die Lösung **nichtlinearer Gleichungssysteme** beschränkt, da sie für die Optimierung wichtige Eigenschaften nicht erfüllt: H_+ ist nicht symmetrisch und i.a. nicht positiv definit. Bei der Globalisierung des Quasi-Newton-Verfahrens sind daher ähnliche Schwierigkeiten zu erwarten, wie beim exakten Newton-Verfahren auch.

Bemerkung 3.10.5

Das lineare Gleichungssystem (3.20) mit H_+ aus (3.25) kann mit der **Sherman-Morrison-Formel**

$$(A+uv^\top)^{-1} = A^{-1} - \frac{A^{-1}uv^\top A^{-1}}{1+v^\top A^{-1}u} \tag{3.26}$$

gelöst werden. Die Matrix $A+uv^\top$ mit $u,v \in \mathbb{R}^n$ ist genau dann singulär, wenn der Nenner $1+v^\top A^{-1}u$ verschwindet.

Symmetrische Rang-1 Formel (SR1-Formel):

Ähnlich wie beim Rang-1 Update nach Broyden machen wir den Ansatz

$$H_+ = H + \gamma uu^\top.$$

Beachte, daß H_+ symmetrisch ist, falls H symmetrisch war. Der Vektor $u \in \mathbb{R}^n$ wird wieder durch die Quasi-Newton-Bedingung (3.24) festgelegt:

$$y = H_+ d = Hd + \gamma u\left(u^\top d\right) \quad \Leftarrow \quad u = y - Hd,\ \gamma = \frac{1}{u^\top d}.$$

Wir erhalten die Update-Formel

$$H_+ = H + \frac{(y-Hd)(y-Hd)^\top}{(y-Hd)^\top d}, \qquad d = x^+ - x, \quad y = \nabla f(x^+) - \nabla f(x). \tag{3.27}$$

Sie heißt **symmetrische Rang-1-Formel (SR1-Formel)**. Sie erhält zwar die Symmetrie, aber nicht die positive Definitheit. Zudem kann der Nenner mitunter verschwinden.

Fazit: Rang-1-Updates genügen nicht, um Symmetrie und positive Definitheit zu erhalten. Daher werden Rang-2-Update-Formeln benötigt.

BFGS-Formel nach Broyden, Fletcher, Goldfarb, Shanno:

Wir wählen den Rang-2-Ansatz

$$H_+ = H + \gamma uu^\top + \delta vv^\top, \qquad u,v \in \mathbb{R}^n.$$

Die Quasi-Newton-Bedingung (3.24) liefert:

$$y = H_+ d = Hd + \gamma u\left(u^\top d\right) + \delta v\left(v^\top d\right).$$

Mit anderen Worten: $y - Hd$ ist Linearkombination von u und v. Mit den Setzungen $u := y$ und $v := -Hd$ folgt

$$\gamma = \frac{1}{u^\top d} = \frac{1}{y^\top d}, \quad \delta = \frac{1}{v^\top d} = -\frac{1}{d^\top H d}$$

und damit die **BFGS-Formel**

$$H_+ = H + \frac{yy^\top}{y^\top d} - \frac{(Hd)(Hd)^\top}{d^\top H d}, \quad d = x^+ - x, \quad y = \nabla f(x^+) - \nabla f(x). \tag{3.28}$$

Es gilt

Hilfssatz 3.10.6

Ist H symmetrisch und positiv definit und gilt $d^\top y > 0$, so ist auch H_+ gemäß (3.28) symmetrisch und positiv definit.

Beweis: Übung □

Bemerkung 3.10.7

Im Hinblick auf eine Globalisierung des Quasi-Newton-Verfahrens mit BFGS-Update ist es wichtig, daß eine Schrittweitenregel gewählt wird, die die Bedingung $d^\top y > 0$ automatisch sicherstellt. Dies ist für die Wolfe-Powell-Schrittweitenregel der Fall. Andernfalls gibt es noch die Möglichkeit, auf die Suchrichtung $d = -\nabla f(x)$ umzuschalten, sobald $d^\top y > 0$ nicht erfüllt ist.

Alternativ kann eine bekannte Update-Formel mit Hilfe der Sherman-Morrison-Formel direkt invertiert werden. Für die BFGS-Formel (3.28) ergibt sich so die **BFGS-Update-Formel für die inverse Matrix**

$$B_+ = B + \left(1 + \frac{y^\top B y}{d^\top y}\right)\frac{dd^\top}{d^\top y} - \frac{d(By)^\top + (By)d^\top}{d^\top y} \tag{3.29}$$

Da B_+ durch Invertierung aus H_+ in (3.28) hervorgeht, erhält B_+ nach Hilfssatz 3.10.6 ebenfalls die positive Definitheit.

DFP-Formel nach Davidon, Fletcher, Powell:

Es ist ebenfalls möglich, für die Inverse Hessematrix B_+ eine Update-Formel analog zur BFGS-Formel herzuleiten. Die Quasi-Newton-Gleichung (3.24) bleibt dann sinngemäß erhalten, indem sie formal mit $H_{i+1}^{-1} = B_{i+1}$ durchmultipliziert wird:

$$d^{[i]} = B_{i+1} y^{[i]}. \tag{3.30}$$

Dies führt dann auf die **DFP-Formel**

$$B_+ = B + \frac{dd^\top}{d^\top y} - \frac{(By)(By)^\top}{y^\top B y}, \quad d = x^+ - x, \quad y = \nabla f(x^+) - \nabla f(x). \tag{3.31}$$

Analog zur BFGS-Formel läßt sich durch Invertierung der DFP-Formel (3.31) eine entsprechende Update-Formel für H_+ herleiten.

Allgemeine Update-Formeln:

Die oben erwähnten BFGS- und DFP-Formeln lassen sich auch aus einem allgemeineren Ansatz ableiten. Dieser Ansatz basiert auf einem zu Satz 3.10.3 analogen Satz:

Satz 3.10.8

Sei $M = M^\top \in \mathbb{R}^{n\times n}$ eine invertierbare Matrix, $y, d \in \mathbb{R}^n$, $d \neq 0$ und $c = M^{-2}d$. Es gelte $H = H^\top$. Dann wird das Minimum des Optimierungsproblems

$$\min_{A\in\mathbb{R}^{n\times n}} \|M(A-H)M\|_F \qquad \textit{unter} \qquad Ad = y,\ A = A^\top$$

angenommen durch

$$H_+ = H + \frac{(y-Hd)c^\top + c(y-Hd)^\top}{c^\top d} - \frac{(y-Hd)^\top d}{(c^\top d)^2} cc^\top.$$

Beweis: Jarre und Stoer [JS04], S. 178. □

Man sucht also unter allen symmetrischen Matrizen, die die Quasi-Newton-Bedingung erfüllen, diejenige die eine gewichtete Frobenius-Norm minimiert. Jede Wahl der Gewichtsmatrizen M führt auf eine eigene Update-Formel. Für $M = I$ folgt $c = d$ und man erhält eine weitere Update-Formel – die **PSB-Formel (Powell symmetric Broyden)**:

$$H_+ = H + \frac{(y-Hd)d^\top + d(y-Hd)^\top}{d^\top d} - \frac{(y-Hd)^\top d}{(d^\top d)^2} dd^\top. \tag{3.32}$$

Diese erhält allerdings nicht die positive Definitheit.

Es gilt jedoch ein zu Satz 3.10.4 analoger Satz, der die lokal superlineare Konvergenz des PSB-Verfahrens liefert, vgl. Satz 11.21 in Geiger und Kanzow [GK99].

Durch andere Wahl von M lassen sich auch die BFGS- und die DFP-Formeln herleiten, siehe hierzu Jarre und Stoer [JS04] und Geiger und Kanzow [GK99]. Es gilt zudem ein zu Satz 3.10.8 analoger Satz für die inverse Matrix B.

Beispiel 3.10.9 (Funktion von Rosenbrock)

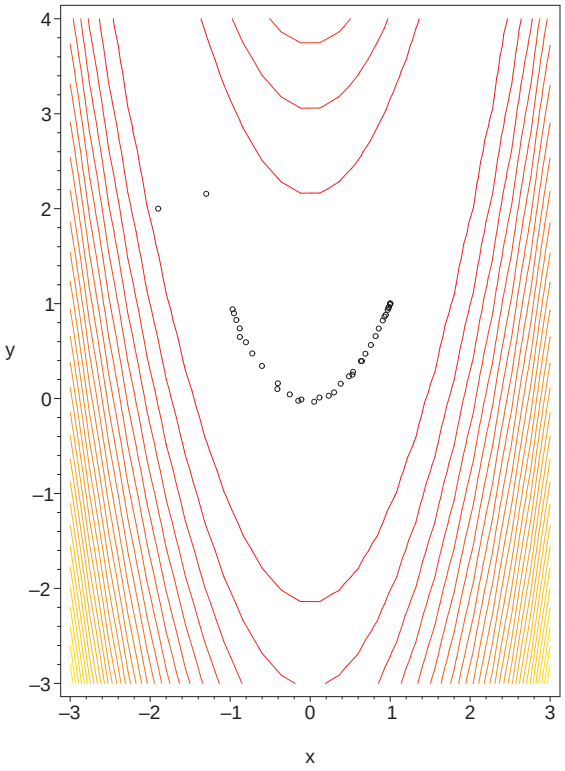
Wir betrachten wieder die Rosenbrockfunktion aus Beispiel 3.9.13 und minimieren diese mit dem Quasi-Newton-Verfahren mit BFGS-Update (3.28), Wolfe-Powell-Schrittweitenbestimmung und den Parametern $\gamma = 1.5$, $\sigma = 0.01$, $\varrho = 0.9$, $\tau_1 = \tau_2 = 0.5$.

Für den Startpunkt $x^{[0]} = (-1.9, 2)^\top$ liefert es folgende (gekürzte) Ausgabe:

ITER	X(1)	X(2)	||GRAD||	||DX||	P=1	P=2	ALPHA	VAL
0	-0.1900000E+01	0.2000000E+01	0.1270869E+04	0.0000000E+00	0.0000000E+00	0.0000000E+00	0.0000000E+00	0.2676200E+03
1	-0.1299707E+01	0.2157227E+01	0.2563943E+03	0.1270869E+04	0.8392488E+00	0.2735873E+00	0.4882812E-03	0.2718995E+02
2	-0.8795055E+00	0.6491019E+00	0.5365338E+02	0.2504912E+02	0.7426736E+00	0.2884778E+00	0.6250000E-01	0.5080774E+01

3	-0.9695494E+00	0.9407219E+00	0.3671907E+01	0.6104100E+00	0.1030576E+01	0.5390096E+00	0.5000000E+00	0.3879173E+01
4	-0.9537856E+00	0.8982665E+00	0.8582907E+01	0.4528750E-01	0.9928905E+00	0.5038925E+00	0.1000000E+01	0.3830367E+01
5	-0.9240251E+00	0.8281261E+00	0.1430104E+02	0.7619298E-01	0.9873516E+00	0.5046694E+00	0.1000000E+01	0.3767903E+01
7	-0.8025736E+00	0.5932189E+00	0.2239531E+02	0.1664492E+00	0.9728386E+00	0.5121564E+00	0.1000000E+01	0.3508408E+01
9	-0.6036296E+00	0.3439663E+00	0.9099631E+01	0.1783787E+00	0.9612907E+00	0.5333392E+00	0.1000000E+01	0.2613254E+01
11	-0.4033210E+00	0.1604669E+00	0.3192217E+01	0.6084060E-01	0.9774289E+00	0.5842240E+00	0.1000000E+01	0.1969794E+01
13	-0.1494509E+00	-0.2272514E-01	0.1030268E+02	0.2507613E+00	0.9743519E+00	0.6170407E+00	0.5000000E+00	0.1524284E+01
15	0.5099375E-01	-0.3406827E-01	0.7423354E+01	0.1634243E+00	0.9345593E+00	0.6222867E+00	0.1000000E+01	0.1035072E+01
17	0.2321091E+00	0.2867909E-01	0.5102759E+01	0.2341734E+00	0.9334192E+00	0.7036639E+00	0.5000000E+00	0.6531380E+00
19	0.3819810E+00	0.1564871E+00	0.3551146E+01	0.1236471E+00	0.8947684E+00	0.7656311E+00	0.1000000E+01	0.3931362E+00
21	0.4860346E+00	0.2338609E+00	0.7391879E+00	0.4562315E-01	0.1042350E+01	0.1177686E+01	0.1000000E+01	0.2647215E+00
23	0.6467575E+00	0.3949859E+00	0.7076391E+01	0.3159072E+00	0.8197120E+00	0.9590923E+00	0.5000000E+00	0.1791132E+00
25	0.6902197E+00	0.4728391E+00	0.8006143E+00	0.9574220E-01	0.8650130E+00	0.1223740E+01	0.1000000E+01	0.9723418E-01
27	0.8176609E+00	0.6577908E+00	0.3825734E+01	0.1105562E+00	0.7785550E+00	0.1563220E+01	0.1000000E+01	0.4486510E-01
29	0.9447914E+00	0.8810946E+00	0.4835292E+01	0.1679440E+00	0.4390532E+00	0.1470417E+01	0.1000000E+01	0.1635644E-01
31	0.9292466E+00	0.8623640E+00	0.3608330E+00	0.4601132E-01	0.7709086E+00	0.3840214E+01	0.1000000E+01	0.5134919E-02
33	0.9673010E+00	0.9346436E+00	0.3906330E+00	0.3492083E-01	0.1907252E+01	0.4977572E+02	0.1000000E+01	0.1174821E-02
35	0.9963214E+00	0.9921323E+00	0.2270727E+00	0.4105488E-01	0.1746233E+00	0.3510957E+01	0.1000000E+01	0.4098374E-04
37	0.1000010E+01	0.1000011E+01	0.4291218E-02	0.2836904E-02	0.5153904E-02	0.1825665E+01	0.1000000E+01	0.9231003E-08
38	0.9999960E+00	0.9999929E+00	0.3824911E-03	0.2256154E-04	0.5620856E+00	0.3863234E+05	0.1000000E+01	0.8635387E-10
39	0.1000000E+01	0.1000000E+01	0.3589352E-07	0.8137089E-05	0.5023790E-02	0.6142959E+03	0.1000000E+01	0.3375678E-15
40	0.1000000E+01	0.1000000E+01	0.1579589E-08	0.4090692E-07	0.4339409E-02	0.1056198E+06	0.1000000E+01	0.7589517E-20
41	0.1000000E+01	0.1000000E+01	0.8964938E-12	0.1782872E-09	0.1071371E-04	0.6009302E+05	0.1000000E+01	0.4017767E-27
42	0.1000000E+01	0.1000000E+01	0.0000000E+00	0.1901604E-14	0.0000000E+00	0.0000000E+00	0.1000000E+01	0.0000000E+00

Man sieht sehr schön die superlineare Konvergenz ab Iteration 35. Davor liegt nur lineare Konvergenz vor. Die folgende Abbildung zeigt die Lage der Iterationspunkte $x^{[i]}$.



Bemerkung 3.10.10

- *Es gibt eine Vielzahl weiterer Update-Formeln. Z.B. enthält die sogenannte Broyden-Klasse Konvexkombinationen der BFGS- und DFP-Formeln.*

- *In der Praxis haben sich vor allem die BFGS-Formeln als besonders effizient erwiesen. Insbesondere konnte Powell für die globale Variante des BFGS-Verfahrens mit Wolfe-Powell-Schrittweite globale Konvergenz und unter Zusatzforderungen sogar superlineare Konvergenz zeigen, vgl. auch Satz 11.33 in Geiger und Kanzow [GK99].*

Dabei läßt sich (3.29) auch schreiben als

$$B_+ = B + \frac{(d - By)d^\top + d(d - By)^\top}{d^\top y} - \frac{(d - By)^\top y dd^\top}{(d^\top y)^2}.$$

- *Das Arbeiten mit der Inversen Update-Formel für B_+ an Stelle von H_+ scheint auf den ersten Blick effizienter zu sein, da lediglich eine Matrix-Vektor-Multiplikation zur Berechnung von $x^{[i+1]}$ in (3.21) notwendig ist. In (3.20) muß hingegen ein lineares Gleichungssystem gelöst werden. Jedoch kann die Cholesky-Zerlegung von H_+ durch Aufdatierung der Cholesky-Zerlegung von H berechnet werden. Ist nämlich $H = LL^\top$, so folgt für den BFGS-Update*

$$\begin{aligned} H_+ &= JJ^\top, \\ J &= L + \frac{(y - Lw)w^\top}{w^\top w}, \\ w &= \sqrt{\frac{y^\top d}{d^\top H d}} L^\top d. \end{aligned}$$

Dabei ist $y^\top d > 0$ vorausgesetzt. Allerdings ist J i.a. keine linke untere Dreicksmatrix. Also berechnet man eine QR-Zerlegung von $J^\top$. Dann gilt

$$H_+ = JJ^\top = R^\top Q^\top QR = R^\top R.$$

Mit $L_+ := R^\top$ hat man die neue Cholesky-Zerlegung von H_+. Unter Ausnutzung der speziellen Struktur von $J^\top$ kann die QR-Zerlegung sehr kostengünstig berechnet werden.

Dies hat zusätzlich den Vorteil, daß man an der Cholesky-Zerlegung gewissermaßen den Grad der positiven Definitheit ablesen kann, während diese Kontrollmöglichkeit bei der Variante (3.21) nicht auftritt.

3.11 CG-Verfahren

Wir diskutieren das **konjugierte Gradientenverfahren (CG-Verfahren)**, welches insbesondere für sehr hochdimensionale Probleme geeignet ist. Es kommt im Wesentlichen mit Matrix-Vektor Multiplikationen aus und wird häufig auch zur Lösung von sehr großen, aber dünn besetzten linearen Gleichungssystemen verwendet.

3.11.1 Quadratische Funktionen

Das konjugierte Gradientenverfahren für quadratische Optimierungsprobleme der Form

$$f(x) = \frac{1}{2} x^\top Ax + b^\top x \to \min \tag{3.33}$$

mit einer symmetrischen und positiv definiten Matrix $A \in \mathbb{R}^{n\times n}$ und einem Vektor $b \in \mathbb{R}^n$ geht auf Hestenes und Stiefel [HS52] zurück.
Ziel ist es, einen Algorithmus zu konstruieren, der

- in endlich vielen Schritten terminiert und
- das Updaten der Hessematrix in jedem Schritt vermeidet.

Bemerkung 3.11.1

Der Zusammenhang mit linearen Gleichungssystemen wird deutlich, wenn man die notwendigen (und wegen der strengen Konvexität von f sogar hinreichenden) Bedingungen für ein Minimum $\hat{x}$ von f auswertet:

$$0 = \nabla f(\hat{x}) = A\hat{x} + b.$$

Offensichtlich ist die Lösung des Gleichungssystems $Ax = -b$ äquivalent zur Lösung des Minimierungsproblems (3.33), falls A symmetrisch und positiv definit ist. Das CG-Verfahren wird insbesondere angewendet, wenn A sehr groß, aber dünn besetzt ist.
Als Beispiel sei die partielle Differentialgleichung

$$\begin{aligned} -u_{xx}(x,y) - u_{yy}(x,y) &= f(x,y), & (x,y) &\in \Omega := (0,1)\times(0,1),\\ u(x,y) &= 0, & (x,y) &\in \partial\Omega \end{aligned}$$

genannt. Diskretisierung auf dem äquidistanten Gitter

$$G := \{(x_i, y_j) \mid x_i = ih,\ y_j = jh,\ 0 \le i,j \le N\}, \quad h = 1/N$$

und Approximation der zweiten partiellen Ableitungen durch

$$\begin{aligned} u_{xx}(x_i,y_j) &\approx \frac{u_{i+1,j} - 2u_{i,j} + u_{i-1,j}}{h^2}, \quad 1 \le i,j \le N-1,\\ u_{yy}(x_i,y_j) &\approx \frac{u_{i,j+1} - 2u_{i,j} + u_{i,j-1}}{h^2}, \quad 1 \le i,j \le N-1, \end{aligned}$$

mit $u_{i,j} \approx u(x_i,y_j)$ liefert das lineare Gleichungssystem $Au = b$ für den Vektor

$$u = (u_{1,1}, u_{1,2}, \dots, u_{1,N-1}, u_{2,1}, u_{2,2}, \dots, u_{2,N-1}, \dots, u_{N-1,1}, u_{N-1,2}, \dots, u_{N-1,N-1})^\top \in \mathbb{R}^{(N-1)^2},$$

mit

$$b = h^2(f_{1,1}, f_{1,2}, \dots, f_{1,N-1}, f_{2,1}, f_{2,2}, \dots, f_{2,N-1}, \dots, f_{N-1,1}, f_{N-1,2}, \dots, f_{N-1,N-1})^\top \in \mathbb{R}^{(N-1)^2},$$

$f_{i,j} := f(x_i,y_j)$,

$$A = \begin{pmatrix} M_1 & D_2 & & & \\ D_2 & M_2 & D_3 & & \\ & \ddots & \ddots & \ddots & \\ & & D_{N-2} & M_{N-2} & D_{N-1} \\ & & & D_{N-1} & M_{N-1} \end{pmatrix}, \quad M_i = \begin{pmatrix} 4 & -1 & & & \\ -1 & 4 & -1 & & \\ & \ddots & \ddots & \ddots & \\ & & -1 & 4 & -1 \\ & & & -1 & 4 \end{pmatrix} \in \mathbb{R}^{(N-1)^2}$$

und $D_i = -I \in \mathbb{R}^{(N-1)\times(N-1)}$. Die Matrix A ist sehr groß (Dimension wächst quadratisch mit N), aber sehr dünn besetzt (pro Zeile sind maximal 5 Einträge $\neq 0$). Zudem ist sie positiv definit und symmetrisch.

Das CG-Verfahren basiert auf der Konstruktion bzgl. des Skalarprodukts $\langle x, y\rangle_A := x^\top Ay$ orthogonaler Richtungen.

Definition 3.11.2

Sei $A \in \mathbb{R}^{n\times n}$ symmetrisch und positiv definit. Die Vektoren $s_1, \dots, s_m \in \mathbb{R}^n$, $s_i \neq 0$, $i = 1, \dots, m$ heißen **A-orthogonal** *oder* **A-konjugiert**, *falls gilt*

$$(s_i)^\top A s_j = 0 \quad \textit{für} \quad i \neq j.$$

Beachte, daß A-orthogonale Richtungen linear unabhängig sind, denn

$$\sum_{i=1}^m \alpha_i s_i = 0 \quad \Rightarrow \quad 0 = s_k^\top A^\top \left(\sum_{i=1}^m \alpha_i s_i \right) = \alpha_k \underbrace{s_k^\top A^\top s_k}_{>0} \quad \Rightarrow \quad \alpha_k = 0.$$

Insbesondere bilden n A-orthogonale Richtungen eine Basis des $\mathbb{R}^n$ und jeder Punkt des $\mathbb{R}^n$ kann als Linearkombination von (maximal) n A-orthogonalen Richtungen dargestellt werden. Insbesondere besitzt auch das Minimum $\hat{x}$ der Funktion f eine solche Darstellung. Der folgende Satz zeigt, wie die Koeffizienten dieser Linearkombination auszusehen haben.

Satz 3.11.3

Sei f gegeben durch (3.33) und $A \in \mathbb{R}^{n\times n}$ symmetrisch und positiv definit. Die Vektoren $d^{[0]}, \dots, d^{[n-1]}$ seien A-orthogonal und $x^{[0]} \in \mathbb{R}^n$ sei beliebig. Desweiteren gelte

$$x^{[i+1]} = x^{[i]} + \alpha_i d^{[i]}, \qquad i = 0, \dots, n-1$$

mit exakter Schrittweite

$$\alpha_i = \textit{argmin}_{\alpha\in\mathbb{R}} f(x^{[i]} + \alpha d^{[i]}).$$

Dann gilt

$$f(x^{[n]}) = \min_{x\in\mathbb{R}^n} f(x).$$

Beweis: (vgl. Jarre und Stoer [JS04])
Jeder Vektor $v \in \mathbb{R}^n$ läßt sich darstellen als

$$v = \sum_{j=0}^{n-1} \gamma_j d^{[j]} \quad \Rightarrow \quad \left(d^{[i]}\right)^\top Av = \gamma_i \underbrace{\left(d^{[i]}\right)^\top Ad^{[i]}}_{>0} \quad \Rightarrow \quad \gamma_i = \frac{\left(d^{[i]}\right)^\top Av}{\left(d^{[i]}\right)^\top Ad^{[i]}}.$$

Damit folgt

$$v = \sum_{j=0}^{n-1} \frac{\left(d^{[j]}\right)^\top Av}{\left(d^{[j]}\right)^\top Ad^{[j]}} d^{[j]}. \tag{3.34}$$

Im Minimum des i-ten Schrittes gilt

$$\varphi'(\alpha_i) = \nabla f(x^{[i]} + \alpha_i d^{[i]})^\top d^{[i]} = 0.$$

Es folgt

$$\begin{aligned}
0 &= \left(d^{[i]}\right)^\top \nabla f(x^{[i+1]}) \\
&= \left(d^{[i]}\right)^\top \left(Ax^{[i+1]} + b\right) \\
&= \left(d^{[i]}\right)^\top \left(A\left(x^{[0]} + \sum_{j=0}^{i-1} \alpha_j d^{[j]} + \alpha_i d^{[i]}\right) + b\right) \\
&= \left(d^{[i]}\right)^\top \left(Ax^{[0]} + b\right) + \alpha_i \underbrace{\left(d^{[i]}\right)^\top Ad^{[i]}}_{>0}.
\end{aligned}$$

Also ist

$$\alpha_i = -\frac{\left(d^{[i]}\right)^\top \left(Ax^{[0]} + b\right)}{\left(d^{[i]}\right)^\top Ad^{[i]}}.$$

Wegen $x^{[n]} = x^{[0]} + \sum_{i=0}^{n-1} \alpha_i d^{[i]}$ folgt

$$x^{[n]} = x^{[0]} - \sum_{i=0}^{n-1} \frac{\left(d^{[i]}\right)^\top \left(Ax^{[0]} + b\right)}{\left(d^{[i]}\right)^\top Ad^{[i]}} d^{[i]} = x^{[0]} - \sum_{i=0}^{n-1} \frac{\left(d^{[i]}\right)^\top A\overbrace{\left(x^{[0]} + A^{-1}b\right)}^{=:v}}{\left(d^{[i]}\right)^\top Ad^{[i]}} d^{[i]}.$$

Definition von $v := x^{[0]} + A^{-1}b$ zusammen mit der Darstellung von v in (3.34) liefert

$$x^{[n]} = x^{[0]} - \left(x^{[0]} + A^{-1}b\right) = -A^{-1}b = \operatorname{argmin}_{x \in \mathbb{R}^n} f(x).$$

□

Der Satz liefert mit anderen Worten die Darstellung

$$\hat{x} = x^{[0]} + \sum_{k=0}^{n-1} \alpha_k d^{[k]}$$

für ein Minimum $\hat{x}$ von f.
Es bleibt die Frage, wie A-orthogonale Richtungen berechnet werden können. Für eine beliebige Basis $\{v_1, \ldots, v_n\}$ des $\mathbb{R}^n$ kann dies beispielsweise durch das **Gram-Schmidtsche Orthogonalisierungsverfahren** bzgl. des Skalarprodukts $\langle \cdot, \cdot \rangle_A$ erfolgen:

$s_1 := v_1$
`for` $k = 2, \dots, n$

$$s_k := -v_k + \sum_{j=1}^{k-1} \frac{\langle s_j, v_k \rangle_A}{\langle s_j, s_j \rangle_A} s_j$$

`end`

Allerdings legt dieses Verfahren die Suchrichtungen von vornherein fest und liefert i.a. auch keine Abstiegsrichtungen.
Besser ist es, die Richtungen schrittweise zu erzeugen und darauf zu achten, daß Abstiegsrichtungen entstehen (was insbesondere im Hinblick auf nicht-quadratische Funktionen von Bedeutung ist). Zur Abkürzung sei ab jetzt

$$g^{[i+1]} := \nabla f(x^{[i+1]}) = Ax^{[i+1]} + b = A(x^{[i]} + \alpha_i d^{[i]}) + b. \tag{3.35}$$

Wir starten mit $d^{[0]} = -\nabla f(x^{[0]}) = -g^{[0]}$ und gehen davon aus, daß bereits A-orthogonale Richtungen $d^{[0]}, \dots, d^{[i]}$ vorliegen. Zusätzlich sei $g^{[j]} \neq 0$, $0 \leq j \leq i+1$ (ansonsten wären wir fertig).

(i) Zunächst bestimmen wir die Schrittweite α_i durch exakte eindimensionale Minimierung von f in $x^{[i]}$ in Richtung $d^{[i]}$. Wie im Beweis zu Satz 3.11.3 folgt mit (3.35) dann

$$(g^{[i+1]})^\top d^{[i]} = 0. \tag{3.36}$$

Wegen $g^{[i+1]} = g^{[i]} + \alpha_i A d^{[i]}$ folgt

$$g^{[i+1]} - g^{[i]} = \alpha_i A d^{[i]}. \tag{3.37}$$

Zusammen mit (3.36) folgt

$$\alpha_i = -\frac{(g^{[i]})^\top d^{[i]}}{(d^{[i]})^\top A d^{[i]}}. \tag{3.38}$$

Diese Beziehungen gelten für alle i.

(ii) Aus (3.37) folgt für $j < i$ die Beziehung

$$\left(g^{[i+1]} - g^{[i]}\right)^\top d^{[j]} = \alpha_i (d^{[i]})^\top A d^{[j]} = 0 \tag{3.39}$$

und damit auch

$$(g^{[i+1]})^\top d^{[j]} = (g^{[j+1]})^\top d^{[j]} + \sum_{k=j+1}^{i} \left(g^{[k+1]} - g^{[k]}\right)^\top d^{[j]} \overset{(3.36),(3.39)}{=} 0.$$

Es folgt

$$(g^{[i+1]})^\top d^{[j]} = 0 \qquad \forall j \leq i. \tag{3.40}$$

(iii) Gilt $g^{[i+1]} \neq 0$, so bedeutet (3.40), daß

$$g^{[i+1]} \perp \operatorname{span}(d^{[0]}, \dots, d^{[i]}). \tag{3.41}$$

Analog zur Gram-Schmidt-Orthogonalisierung machen wir den Ansatz

$$d^{[i+1]} = -g^{[i+1]} + \sum_{j=0}^{i} \frac{\langle d^{[j]}, g^{[i+1]} \rangle_A}{\langle d^{[j]}, d^{[j]} \rangle_A} d^{[j]}. \tag{3.42}$$

In der Tat folgt aus der A-Orthogonalität von $\{d^{[0]}, \dots, d^{[i]}\}$ die Beziehung $(d^{[i+1]})^\top A d^{[j]} = 0$ für $j = 0, \dots, i$. Somit ist auch $\{d^{[0]}, \dots, d^{[i+1]}\}$ A-orthogonal.

Zudem ist $d^{[i+1]}$ auch Abstiegsrichtung, denn Multiplikation von links mit $(g^{[i+1]})^\top$ liefert zusammen mit (3.40) die Beziehung

$$(g^{[i+1]})^\top d^{[i+1]} = -\|g^{[i+1]}\|^2 < 0. \tag{3.43}$$

Beachte, daß (3.38) $\alpha_{i+1} > 0$ liefert.

(iv) Wir zeigen, daß in (3.42) alle Summanden mit $j < i$ verschwinden. Wegen (3.42) ist

$$g^{[i+1]} \in \operatorname{span}(d^{[0]}, \dots, d^{[i+1]}) \quad \text{bzw.} \quad g^{[j]} \in \operatorname{span}(d^{[0]}, \dots, d^{[j]}),\ 0 \leq j \leq i+1.$$

Wegen (3.41) folgt

$$(g^{[i+1]})^\top g^{[j]} = 0 \qquad \forall j \leq i. \tag{3.44}$$

Für $j < i$ folgt dann

$$(g^{[i+1]})^\top A d^{[j]} \overset{(3.37)}{=} \frac{1}{\alpha_j} (g^{[i+1]})^\top \left(g^{[j+1]} - g^{[j]} \right) = 0.$$

(3.42) reduziert sich somit zu

$$d^{[i+1]} = -g^{[i+1]} + \frac{\langle d^{[i]}, g^{[i+1]} \rangle_A}{\langle d^{[i]}, d^{[i]} \rangle_A} d^{[i]}. \tag{3.45}$$

Schließlich folgt mit $A d^{[i]} = \frac{1}{\alpha_i} \left(g^{[i+1]} - g^{[i]} \right)$ gemäß (3.37):

$$\begin{aligned}
(g^{[i+1]})^\top A d^{[i]} &= \frac{1}{\alpha_i} \left(\|g^{[i+1]}\|^2 - (g^{[i+1]})^\top g^{[i]} \right) \\
&\overset{(3.44)}{=} \frac{1}{\alpha_i} \|g^{[i+1]}\|^2, \\
(d^{[i]})^\top A d^{[i]} &\overset{(3.38)}{=} -\frac{1}{\alpha_i} (g^{[i]})^\top d^{[i]} \\
&\overset{(3.43)}{=} \frac{1}{\alpha_i} \|g^{[i]}\|^2.
\end{aligned}$$

Einsetzen in (3.45) liefert

$$d^{[i+1]} = -g^{[i+1]} + \frac{\|g^{[i+1]}\|^2}{\|g^{[i]}\|^2} d^{[i]}. \tag{3.46}$$

Die obigen Betrachtungen zusammen mit Satz 3.11.3 und der Tatsache, daß maximal n A-orthogonale Richtungen existieren, zeigen:

Satz 3.11.4

Sei f gegeben durch (3.33) und $A \in \mathbb{R}^{n\times n}$ symmetrisch und positiv definit. Dann endet das obige Verfahren für einen beliebigen Startwert $x^{[0]} \in \mathbb{R}^n$ nach $m \le n$ Schritten mit $\nabla f(x^{[m]}) = 0$.
Für das kleinste $m \le n$ mit $\nabla f(x^{[m]}) = 0$ gelten für alle $l \le m$ die Orthogonalitätsrelationen:

$$\begin{aligned}
(d^{[i]})^\top A d^{[j]} &= 0, \qquad 0 \le j < i \le l,\\
(g^{[i]})^\top g^{[j]} &= 0, \qquad 0 \le j < i \le l, \ \textit{vgl. (3.44)}\\
(g^{[i]})^\top d^{[j]} &= 0, \qquad 0 \le j < i \le l, \ \textit{vgl. (3.40)}\\
(g^{[i]})^\top d^{[i]} &= -\|g^{[i]}\|^2, \qquad 0 \le i \le l, \textit{vgl. (3.43)}.
\end{aligned}$$

Die obigen Betrachtungen liefern ein konstruktives Verfahren zur Berechnung der A-orthogonalen Suchrichtungen. Zusammen mit Satz 3.11.3 folgt daraus die Endlichkeit des folgenden Algorithmus:

Algorithmus: CG-Verfahren

(i) Wähle einen Startvektor $x^{[0]} \in \mathbb{R}^n$, $\varepsilon \ge 0$, berechne $g^{[0]} = \nabla f(x^{[0]}) = Ax^{[0]} + b$ und setze $d^{[0]} = -g^{[0]}$, $i = 0$.

(ii) Falls $\|g^{[i]}\| \le \varepsilon$, STOP.

(iii) Berechne

$$\begin{aligned}
\alpha_i &= \frac{\|g^{[i]}\|^2}{(d^{[i]})^\top A d^{[i]}},\\
x^{[i+1]} &= x^{[i]} + \alpha_i d^{[i]},\\
g^{[i+1]} &= Ax^{[i+1]} + b = g^{[i]} + \alpha_i A d^{[i]},\\
\beta_i &= \frac{\|g^{[i+1]}\|^2}{\|g^{[i]}\|^2},\\
d^{[i+1]} &= -g^{[i+1]} + \beta_i d^{[i]}.
\end{aligned}$$

(iv) Setze $i := i + 1$, und gehe zu (ii).

Beispiel 3.11.5

Wir betrachten die partielle Differentialgleichung

$$\begin{aligned} -u_{xx}(x,y) - u_{yy}(x,y) &= f(x,y), \qquad (x,y) \in \Omega := (0,1) \times (0,1), \\ u(x,y) &= 0, \qquad (x,y) \in \partial\Omega \end{aligned}$$

mit

$$f(x,y) = 1000 \cdot \sin((x-0.5) \cdot (y-0.5)).$$

Die partielle Differentialgleichung wird wie in Bemerkung 3.11.1 diskretisiert und führt auf das lineare Gleichungssystem $Au = b$. Anschließend wird das äquivalente Problem

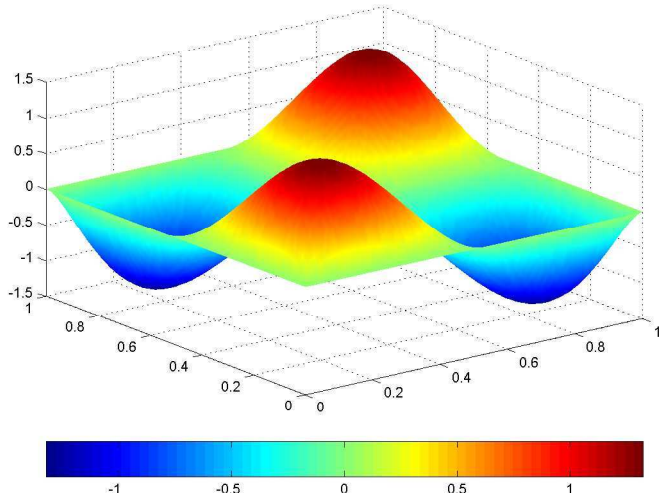
$$\frac{1}{2} u^\top A u - b^\top u \to \min$$

mit dem CG-Verfahren gelöst. Dabei werden nur die von Null verschiedenen Einträge von A mit einer entsprechenden Indizierung gespeichert, so daß Multiplikationen Ad sehr effizient durchgeführt werden können. Für $N = 100$ hat A die Dimension $99^2 = 9801$ und von den $9801^2 = 96059601$ Einträgen sind lediglich 48609 Einträge von Null verschieden.

Als Startwert wählen wir $u^{[0]} = 0$. Für $N = 100$ und $\varepsilon = 10^{-8}$ ergibt sich folgende Ausgabe:

```
  48609
ITER       ||GRAD||          ||DX||            ALPHA             BETA
------------------------------------------------------------------------
   0   0.8055956E+00   0.0000000E+00   0.0000000E+00   0.0000000E+00
   1   0.2271664E+01   0.6796640E+01   0.8287775E+01   0.7951592E+01
   3   0.2274125E+01   0.7564077E+01   0.7990598E+00   0.1092265E+01
   5   0.2100902E+01   0.7055373E+01   0.5045940E+00   0.8358480E+00
   7   0.1932038E+01   0.6517664E+01   0.4768494E+00   0.8424729E+00
   9   0.1767712E+01   0.5963891E+01   0.4706184E+00   0.8557311E+00
  11   0.1611152E+01   0.5421028E+01   0.4690171E+00   0.8656038E+00
  13   0.1463812E+01   0.4902665E+01   0.4686974E+00   0.8722568E+00
  15   0.1326138E+01   0.4414901E+01   0.4687291E+00   0.8765485E+00
  17   0.1198085E+01   0.3960077E+01   0.4688082E+00   0.8791410E+00
  19   0.1079372E+01   0.3538615E+01   0.4688279E+00   0.8804796E+00
  21   0.9696041E+00   0.3149930E+01   0.4687504E+00   0.8808559E+00
  23   0.8683385E+00   0.2792889E+01   0.4685635E+00   0.8804600E+00
  25   0.7751144E+00   0.2466078E+01   0.4682637E+00   0.8794144E+00
...
 135   0.3018280E-06   0.5503303E-06   0.4188414E+00   0.6769555E+00
 137   0.2309873E-06   0.4445599E-06   0.4462299E+00   0.7570287E+00
 139   0.1747984E-06   0.3426776E-06   0.4463246E+00   0.7220239E+00
 141   0.1093353E-06   0.1928130E-06   0.4093333E+00   0.6062993E+00
 143   0.6981008E-07   0.1201987E-06   0.4214496E+00   0.6967370E+00
 145   0.5219784E-07   0.9674677E-07   0.4486677E+00   0.7779052E+00
 147   0.3285968E-07   0.5685132E-07   0.3985365E+00   0.6319190E+00
 149   0.2182741E-07   0.3802979E-07   0.4169753E+00   0.7148059E+00
 151   0.1323888E-07   0.2172648E-07   0.3884909E+00   0.5765313E+00
 153   0.8588212E-08   0.1449120E-07   0.4170621E+00   0.7137157E+00
```

Die folgende Abbildung zeigt die numerische Lösung u der partiellen Differentialgleichung:



Bemerkung 3.11.6

- *Der wesentliche Aufwand des CG-Verfahrens besteht in der Berechnung des Matrix-Vektor-Produkts $Ad^{[i]}$. Ist A sehr groß, aber dünn besetzt, so müssen nur die nicht-verschwindenden Elemente von A mit einer entsprechenden Indizierung gespeichert werden.*

- *Es stellt sich heraus, daß die Suchrichtung $d^{[i]}$ – bis auf Skalierung – auch durch das Optimierungsproblem*

$$(g^{[i]})^\top d \to \min \quad \textit{unter} \quad d^\top d = 1,\ d^\top A d^{[j]} = 0,\ j = 1, \ldots, i-1$$

  ensteht, vgl. Hestenes [Hes80].

- *Kelley [Kel95], Th. 2.2.3, S. 15, zeigt, daß das CG-Verfahren das Minimum in höchstens k Schritten liefert, wobei k die Anzahl der verschiedenen Eigenwerte von A bezeichnet.*

- *Möchte man das Gleichungssystem $Ax+b=0$ lösen, wobei A nicht symmetrisch und positiv definit ist, so kann man stattdessen die quadratische Funktion $f(x) = \frac{1}{2}\|Ax+b\|^2$ minimieren. Notwendig erfüllt ein Minimum $\hat{x}$ die Bedingung $0 = \nabla f(\hat{x}) = A^\top Ax + A^\top b$ – die Normalengleichungen. Hat A vollen Rang, so ist $A^\top A$ positiv definit. Dieser Ansatz führt auf CGNR-Verfahren, vgl. Kelley [Kel95], S. 25.*

  Nachteil: Die Kondition von $A^\top A$ ist in der Regel deutlich schlechter als die von A.

3.11.2 Präkonditionierung

Obwohl das CG-Verfahren (zumindest theoretisch) nach maximal n Schritten terminiert, ist es in der Praxis aus Aufwandsgründen oft nicht möglich, n Iterationen tatsächlich durchzuführen (n kann sehr groß sein, etwa 1 Million).

Daher ist es sinnvoll, das CG-Verfahren als iteratives Verfahren zu betrachten. Es stellt sich dann die Frage nach der Konvergenzgeschwindigkeit. Es läßt sich folgende Abschätzung beweisen:

$$\|x^{[i]} - \hat{x}\|_A \leq 2 \left(\frac{\sqrt{\kappa(A)} - 1}{\sqrt{\kappa(A)} + 1} \right)^i \|x^{[0]} - \hat{x}\|_A, \tag{3.47}$$

wobei $\kappa(A) = \lambda_{max}/\lambda_{min}$ die Kondition der Matrix A bzgl. der Norm $\|\cdot\|_2$ bezeichnet.

Die Fehlerabschätzung zeigt, daß das Verfahren umso schneller konvergiert, je näher die Kondition $\kappa(A)$ bei Eins liegt. Umgekehrt ist die Konvergenz umso langsamer, je größer die Kondition von A ist.

Es ist daher erstrebenswert, eine möglichst gut konditionierte Matrix A zu haben.

Wir versuchen dies zu erreichen, indem wir eine Variablentransformation durchführen:

$$\begin{aligned} x &:= Sz, \quad S \in \mathbb{R}^{n \times n} \text{ regulär}, \\ \hat{f}(z) &:= f(Sz) = \frac{1}{2} z^\top S^\top A S z + (S^\top b)^\top z. \end{aligned}$$

Nun wenden wir das übliche CG-Verfahren auf $\hat{f}$ an und erhalten die Größen

$$\hat{x}^{[i]}, \quad \hat{g}^{[i]}, \quad \hat{d}^{[i]}, \quad \hat{\alpha}_i, \quad \hat{\beta}_i.$$

Rücksubstitution liefert

$$x^{[i]} = S\hat{x}^{[i]}, \quad S^\top g^{[i]} = \hat{g}^{[i]}, \quad d^{[i]} = S\hat{d}^{[i]}.$$

Unter Berücksichtigung der Rücksubstitution ergibt sich:

Algorithmus: Präkonditioniertes CG-Verfahren

(i) Wähle eine symmetrische und positiv definite Matrix $B \in \mathbb{R}^{n\times n}$, einen Startvektor $x^{[0]} \in \mathbb{R}^n$, $\varepsilon \geq 0$, berechne $g^{[0]} = Ax^{[0]} + b$ und setze $d^{[0]} = -Bg^{[0]}$, $i = 0$.

(ii) Falls $\|g^{[i]}\| \leq \varepsilon$, STOP.

(iii) Berechne

$$\begin{aligned}
\hat{\alpha}_i &= \frac{(g^{[i]})^\top B g^{[i]}}{(d^{[i]})^\top A d^{[i]}}, \\
x^{[i+1]} &= x^{[i]} + \hat{\alpha}_i d^{[i]}, \\
g^{[i+1]} &= g^{[i]} + \hat{\alpha}_i A d^{[i]}, \\
\hat{\beta}_i &= \frac{(g^{[i+1]})^\top B g^{[i+1]}}{(g^{[i]})^\top B g^{[i]}}, \\
d^{[i+1]} &= -Bg^{[i+1]} + \hat{\beta}_i d^{[i]}.
\end{aligned}$$

(iv) Setze $i := i + 1$, und gehe zu (ii).

Im Algorithmus tritt nur noch die symmetrische und positiv definite Matrix $B = SS^\top$ auf. B ist so zu wählen, daß die Matrix BA eine möglichst kleine Kondition besitzt. Wegen

$$S^{-1}BAS = S^{-1}SS^\top AS = S^\top AS$$

ist BA ähnlich zu $S^\top AS$ und folglich besitzen BA und $S^\top AS$ dieselben Eigenwerte und somit dieselbe Kondition.
Desweiteren sollte B so gewählt werden, daß die Multiplikationen $g \mapsto Bg$ kostengünstig ausgewertet werden können.
Mögliche Ansätze für B sind:

(a) $B = D^{-1}$, wobei D die Diagonale von A bezeichnet.

(b) Berechne die Cholesky-Zerlegung $A = L \cdot L^\top$, approximiere L durch Weglassen kleiner Elemente durch $\hat{L}$ und setze $B = \hat{L}^{-\top}\hat{L}^{-1}$. Dann gilt $BA = \hat{L}^{-\top}\hat{L}^{-1}LL^\top \approx I$. Dieses Verfahren ist brauchbar für dünn besetzte Matrizen.

3.11.3 CG-Verfahren für allgemeine Funktionen

Der CG-Algorithmus läßt sich formal auf beliebige stetig differenzierbare Zielfunktionen übertragen. Lediglich die exakte Liniensuche muß durch eine Schrittweitenstrategie ersetzt werden. Fletcher und Reeves nutzten dabei die strenge Wolfe-Powell-Bedingung und

formulierten

Algorithmus: Fletcher-Reeves-Verfahren

(i) Wähle einen Startvektor $x^{[0]} \in \mathbb{R}^n$, $\varepsilon \geq 0$, $\sigma \in (0, 1/2)$, $\rho \in (\sigma, 1/2)$, setze $d^{[0]} = -\nabla f(x^{[0]})$, $i = 0$.

(ii) Falls $\|\nabla f(x^{[i]})\| \leq \varepsilon$, STOP.

(iii) Berechne eine Schrittweite $\alpha_i > 0$, die den strengen Wolfe-Powell-Bedingungen (3.6)-(3.7) genügt.

(iv) Setze

$$\begin{aligned} x^{[i+1]} &= x^{[i]} + \alpha_i d^{[i]}, \\ \beta_i &= \frac{\|\nabla f(x^{[i+1]})\|^2}{\|\nabla f(x^{[i]})\|^2}, \\ d^{[i+1]} &= -\nabla f(x^{[i+1]}) + \beta_i d^{[i]}. \end{aligned}$$

(v) Setze $i := i + 1$, und gehe zu (ii).

In Geiger und Kanzow [GK99] wird folgender Konvergenzsatz bewiesen:

Satz 3.11.7

Sei $f : \mathbb{R}^n \to \mathbb{R}$ stetig differenzierbar und nach unten beschränkt. ∇f sei Lipschitz-stetig auf der Niveaumenge $lev(f, f(x^{[0]}))$ und es gelte $\varepsilon = 0$. Dann ist das Fletcher-Reeves-Verfahren wohldefiniert (es erzeugt Abstiegsrichtungen) und es gilt

$$\liminf_{i \to \infty} \|\nabla f(x^{[i]})\| = 0$$

für jede durch den Algorithmus erzeugte Folge $\{x^{[i]}\}$.

Beweis:

(a) Zur Wohldefiniertheit: Der Algorithmus ist wohldefiniert, falls es gelingt, eine strenge Wolfe-Powell-Schrittweite zu bestimmen. Gemäß Satz 3.7.3 existieren strenge Wolfe-Powell-Schrittweiten, wenn noch gezeigt wird, daß $\nabla f(x^{[i]})^\top d^{[i]} < 0$ gilt.

Durch vollständige Induktion zeigen wir für alle i:

$$-\sum_{j=0}^{\infty} \varrho^j \leq -\sum_{j=0}^{i} \varrho^j \leq \frac{\nabla f(x^{[i]})^\top d^{[i]}}{\|\nabla f(x^{[i]})\|^2} \leq -2 + \sum_{j=0}^{i} \varrho^i \leq -2 + \sum_{j=0}^{\infty} \varrho^i. \tag{3.48}$$

Wegen $\varrho \in (0, 1/2)$ steht dann auf der rechten Seite

$$-2 + \sum_{j=0}^{\infty} \varrho^i = -2 + \frac{1}{1-\varrho} = \frac{2\varrho - 1}{1-\varrho} < 0,$$

womit die Abstiegsbedingung erfüllt wäre.

Wegen $d^{[0]} = -\nabla f(x^{[0]})$ gelten die Ungleichungen für $i = 0$.

Die Ungleichungen seien nun erfüllt für $i \in \mathbb{N}$. Aus den strengen Wolfe-Powell-Bedingungen folgt

$$\varrho \nabla f(x^{[i]})^\top d^{[i]} \leq \nabla f(x^{[i+1]})^\top d^{[i]} \leq -\varrho \nabla f(x^{[i]})^\top d^{[i]}$$

und somit auch

$$-1 + \varrho \frac{\nabla f(x^{[i]})^\top d^{[i]}}{\|\nabla f(x^{[i]})\|^2} \leq -1 + \frac{\nabla f(x^{[i+1]})^\top d^{[i]}}{\|\nabla f(x^{[i]})\|^2} \leq -1 - \varrho \frac{\nabla f(x^{[i]})^\top d^{[i]}}{\|\nabla f(x^{[i]})\|^2}.$$

Der Algorithmus liefert die Suchrichtung

$$d^{[i+1]} = -\nabla f(x^{[i+1]}) + \beta_i d^{[i]}, \qquad \beta_i = \|\nabla f(x^{[i+1]})\|^2 / \|\nabla f(x^{[i]})\|^2.$$

Dies liefert

$$\frac{\nabla f(x^{[i+1]})^\top d^{[i+1]}}{\|\nabla f(x^{[i+1]})\|^2} = -1 + \frac{\nabla f(x^{[i+1]})^\top d^{[i]}}{\|\nabla f(x^{[i]})\|^2}.$$

Ersetzen des rechten Ausdrucks in vorangehender Ungleichung liefert zusammen mit der Induktionsvoraussetzung

$$\begin{aligned}
-\sum_{j=0}^{i+1} \varrho^j = -1 - \varrho \sum_{j=0}^{i} \varrho^j \quad &\leq \quad -1 + \varrho \frac{\nabla f(x^{[i]})^\top d^{[i]}}{\|\nabla f(x^{[i]})\|^2} \\
&\leq \quad -1 + \frac{\nabla f(x^{[i+1]})^\top d^{[i]}}{\|\nabla f(x^{[i]})\|^2} \\
&= \quad \frac{\nabla f(x^{[i+1]})^\top d^{[i+1]}}{\|\nabla f(x^{[i+1]})\|^2} \\
&\leq \quad -1 - \varrho \underbrace{\frac{\nabla f(x^{[i]})^\top d^{[i]}}{\|\nabla f(x^{[i]})\|^2}}_{\geq -\sum_{j=0}^{i} \varrho^j} \\
&\leq \quad -1 + \varrho \sum_{j=0}^{i} \varrho^j \\
&= \quad -2 + \sum_{j=0}^{i+1} \varrho^j.
\end{aligned}$$

Dies liefert die Behauptung.

(b) Sei $\{x^{[i]}\}$ eine durch das Fletcher-Reeves-Verfahren erzeugte Folge. Dann gelten nach (a) die Ungleichungen (3.48) und somit

$$-\frac{1}{1-\varrho} \leq -\sum_{j=0}^{i} \varrho^j \leq \frac{\nabla f(x^{[i]})^\top d^{[i]}}{\|\nabla f(x^{[i]})\|^2} \leq \frac{2\varrho-1}{1-\varrho} < 0 \qquad \forall i \in \mathbb{N}. \tag{3.49}$$

Wir führen den Beweis indirekt und nehmen an, daß

$$\liminf_{i\to\infty} \|\nabla f(x^{[i]})\| > 0.$$

Dann existiert $\varepsilon > 0$ mit $\|\nabla f(x^{[i]})\| \geq \varepsilon$ für alle $i \in \mathbb{N}$. Aus der Effizienz der strengen Wolfe-Powell-Regel (vgl. Satz 3.7.4) folgt die Existenz von $C > 0$ mit

$$\begin{aligned} f(x^{[i]}) - f(x^{[i+1]}) &\geq C\left(\frac{\nabla f(x^{[i]})^\top d^{[i]}}{\|d^{[i]}\|}\right)^2 \\ &\overset{(3.49)}{\geq} C\left(\frac{1-2\varrho}{1-\varrho}\right)^2 \frac{1}{\gamma_i} \end{aligned}$$

für alle $\in \mathbb{N}$ mit $\gamma_i = \|d^{[i]}\|^2/\|\nabla f(x^{[i]})\|^4$. Mit $d^{[i]}$ und β_i aus dem Fletcher-Reeves-Verfahren und den strengen Wolfe-Powell Bedingungen folgt für alle $i \in \mathbb{N}$:

$$\begin{aligned} \gamma_i &= \frac{(d^{[i]})^\top d^{[i]}}{\|\nabla f(x^{[i]})\|^4} \\ &= \frac{(-\nabla f(x^{[i]}) + \beta_{i-1} d^{[i-1]})^\top (-\nabla f(x^{[i]}) + \beta_{i-1} d^{[i-1]})}{\|\nabla f(x^{[i]})\|^4} \\ &= \frac{1}{\|\nabla f(x^{[i]})\|^2} - \frac{2}{\|\nabla f(x^{[i]})\|^2 \|\nabla f(x^{[i-1]})\|^2} \underbrace{\nabla f(x^{[i]})^\top d^{[i-1]}}_{\geq \varrho \nabla f(x^{[i-1]})^\top d^{[i-1]}} + \gamma_{i-1} \\ &\overset{WP}{\leq} \frac{1}{\|\nabla f(x^{[i]})\|^2} - \frac{2\varrho}{\|\nabla f(x^{[i]})\|^2} \cdot \frac{\nabla f(x^{[i-1]})^\top d^{[i-1]}}{\|\nabla f(x^{[i-1]})\|^2} + \gamma_{i-1} \\ &\overset{(3.49)}{\leq} \frac{1}{\|\nabla f(x^{[i]})\|^2} + \frac{2\varrho}{\|\nabla f(x^{[i]})\|^2} \cdot \frac{1}{1-\varrho} + \gamma_{i-1} \\ &= \frac{1+\varrho}{1-\varrho} \cdot \frac{1}{\|\nabla f(x^{[i]})\|^2} + \gamma_{i-1}. \end{aligned}$$

Induktiv ergibt sich

$$\begin{aligned} \gamma_i &\leq \frac{1+\varrho}{1-\varrho} \cdot \frac{1}{\|\nabla f(x^{[i]})\|^2} + \gamma_{i-1} \\ &\vdots \\ &\leq \frac{1+\varrho}{1-\varrho} \sum_{j=1}^{i} \frac{1}{\|\nabla f(x^{[j]})\|^2} + \underbrace{\gamma_0}_{=\|\nabla f(x^{[0]})\|^2/\|\nabla f(x^{[0]})\|^4} \\ &\leq \frac{1+\varrho}{1-\varrho} \sum_{j=0}^{i} \frac{1}{\|\nabla f(x^{[j]})\|^2} \\ &\leq \frac{1}{\varepsilon^2} \cdot \frac{1+\varrho}{1-\varrho}(i+1). \end{aligned}$$

Damit folgt weiter

$$f(x^{[i]}) - f(x^{[i+1]}) \geq C\varepsilon^2 \frac{(1-2\varrho)^2}{(1-\varrho)(1+\varrho)} \cdot \frac{1}{i+1}.$$

Damit folgt

$$f(x^{[0]}) - f(x^{[i+1]}) = \sum_{j=0}^{i} f(x^{[j]}) - f(x^{[j+1]}) \geq C\varepsilon^2 \frac{(1-2\varrho)^2}{(1-\varrho)(1+\varrho)} \cdot \sum_{j=0}^{i} \frac{1}{1+j} \to \infty.$$

Grenzübergang und Divergenz der harmonischen Reihe liefert

$$f(x^{[i]}) \to -\infty$$

im Widerspruch zur Beschränktheit von f. Dies zeigt $\liminf_{i\to\infty} \|\nabla f(x^{[i]})\| = 0$.

□

Beachte, daß der vorstehende Satz **nicht** besagt, daß **jeder** Häufungspunkt von $\{x^{[i]}\}$ ein stationärer Punkt von f ist.
Es gibt weitere Varianten des Verfahrens, die sich hauptsächlich durch die Berechnung von β_i unterscheiden:

(a) **Polak-Ribière-Verfahren:**

$$\beta_i = \frac{\left(\nabla f(x^{[i+1]}) - \nabla f(x^{[i]})\right)^\top \nabla f(x^{[i+1]})}{\|\nabla f(x^{[i]})\|^2}.$$

Dieses Verfahren ist dem Fletcher-Reeves-Verfahren in der Praxis häufig überlegen, jedoch besitzt es weniger schöne theoretische Konvergenzeigenschaften. In der Praxis werden Restarts durchgeführt, d.h. es wird $d^{[i]} = -\nabla f(x^{[i]})$ gesetzt, falls $d^{[i]}$ keine Abstiegsrichtung sein sollte oder falls die Winkelbedingung nicht erfüllt ist.

(b) **Hestenes-Stiefel-Verfahren:**

$$\beta_i = \frac{\left(\nabla f(x^{[i+1]}) - \nabla f(x^{[i]})\right)^\top \nabla f(x^{[i+1]})}{\left(\nabla f(x^{[i+1]}) - \nabla f(x^{[i]})\right)^\top d^{[i]}}.$$

Bei Verwendung der Curry-Schrittweitenregel stimmt dieses Verfahren mit dem Verfahren von Polak-Ribière überein.

(c) **Myers-Verfahren:**

$$\beta_i = -\frac{\|\nabla f(x^{[i+1]})\|^2}{\nabla f(x^{[i]})^\top d^{[i]}}.$$

Alle erwähnten Varianten stimmen für streng konvexe quadratische Zielfunktionen (3.33) überein.

3.12 Trust-Region-Verfahren

In $x \in \mathbb{R}^n$ betrachten wir wiederum die quadratische Approximation

$$\hat{f}(d) = f(x) + \nabla f(x)^\top d + \frac{1}{2} d^\top \nabla^2 f(x) d \tag{3.50}$$

der zweimal stetig differenzierbaren Funktion $f : \mathbb{R}^n \to \mathbb{R}$. Diese Approximation hatten wir in Beispiel 3.9.2 bereits genutzt, um das Newton-Verfahren zu motivieren. Allerdings hatten wir in Beispiel 3.9.2 noch verlangt, daß die Hessematrix $\nabla^2 f(x)$ von f positiv definit ist, um zu garantieren, daß $\hat{f}$ ein Minimum besitzt.
Ist die Hessematrix nicht positiv definit, so besitzt $\hat{f}$ u.U. kein Minimum. Dieses Problem hatten wir auch schon in Beispiel 3.9.1 für die lineare Funktion $f(x) + \nabla f(x)^\top d$. Dort führten wir einfach die Nebenbedingung $\|d\| \le 1$ ein, um ein wohldefiniertes Minimierungsproblem zu erhalten. Als Lösung ergab sich in Beispiel 3.9.1 die Richtung $d = -\nabla f(x)/\|\nabla f(x)\|$, also das Gradientenverfahren.
Es ist daher naheliegend, auch für d in (3.50) eine Nebenbedingung zu formulieren:

$$\|d\| \le \Delta, \qquad \Delta > 0.$$

Das Minimierungsproblem

$$\hat{f}(d) = f(x) + \nabla f(x)^\top d + \frac{1}{2} d^\top \nabla^2 f(x) d \to \min \quad \text{unter} \quad \|d\| \le \Delta$$

besitzt nach dem Satz 1.1.1 von Weierstrass stets eine Lösung ($\hat{f}$ ist stetig, $\|d\| \le \Delta$ ist kompakt) – insbesondere auch für eine nicht positiv definite Hessematrix.
Die Idee des **Trust-Region-Verfahrens** ist es nun, den Wert Δ iterativ anzupassen. Es erzeugt Folgen $\{x^{[i]}\}$ und $\{\Delta_i\}$ mit

$$x^{[i+1]} = x^{[i]} + d^{[i]}, \quad i = 0, 1, 2, \ldots,$$

wobei $d^{[i]}$ Lösung ist vom

Trust-Region-Teilproblem: Minimiere

$$\hat{f}_i(d) = f(x^{[i]}) + \nabla f(x^{[i]})^\top d + \frac{1}{2} d^\top \nabla^2 f(x^{[i]}) d \tag{3.51}$$

unter der Nebenbedingung

$$\|d\| \le \Delta_i, \qquad \Delta_i > 0. \tag{3.52}$$

Für kleines $\Delta_i > 0$ ist $\hat{f}_i(d)$ sicherlich eine gute Näherung für $f(x^{[i]} + d)$, da $\hat{f}_i(0) = f(x^{[i]})$ und $\nabla \hat{f}_i(0) = \nabla f(x^{[i]})$ gelten. Damit ist zu erwarten, daß die Lösung des Teilproblems eine gute Näherung für das Minimum von $f(x^{[i]} + d)$ auf dem durch (3.52) gegebenem

Bereich liefert. Man kann also der Approximation (3.51) auf diesem Bereich „vertrauen“– daher auch der Name **Vertrauensbereich (Trust Region)**.

Bemerkung 3.12.1

- *Durch die Anpassung von $\Delta_i > 0$ von Schritt zu Schritt erübrigt sich eine Liniensuche für die Funktion $\varphi(\alpha) = f(x^{[i]} + \alpha d^{[i]})$. I.a. sind die Lösungen $d^{[i]}$ des Trust-Region-Teilproblems keine Abstiegsrichtungen von f in $x^{[i]}$.*
- *Die euklidische Norm in (3.52) kann prinzipiell auch durch andere Normen, etwa die Maximumsnorm $\|\cdot\|_\infty$, ersetzt werden. Dies führt natürlich zu einem anderen Trust-Region-Verfahren.*

Anpassung des Vertrauensbereichs:

Es stellt sich die Frage, wie Δ_i gewählt werden soll.
Ein üblicher Zugang ist es, die **tatsächliche Abnahme** des Funktionswertes

$$a_i := f(x^{[i]}) - f(x^{[i+1]})$$

mit der **vorausgesagten Abnahme** der Approximation

$$p_i := \hat{f}_i(0) - \hat{f}_i(d^{[i]})$$

zu vergleichen. Der Quotient

$$q_i := \frac{a_i}{p_i} \tag{3.53}$$

dient als Maß für die Übereinstimmung von a_i und p_i. Er ist wohldefiniert, denn $p_i = 0$ impliziert

$$\hat{f}_i(d^{[i]}) = \hat{f}_i(0) = f(x^{[i]}) \le f(x^{[i]}) + \nabla f(x^{[i]})d + \frac{1}{2}d^\top \nabla^2 f(x^{[i]})d \qquad \forall \|d\| \le \Delta_i. \tag{3.54}$$

Damit ist (auch) $\hat{d} = 0$ Lösung des Trust-Region-Teilproblems und es gilt notwendig (da 0 innerer Punkt des Vertrauensbereichs ist)

$$0 = \nabla \hat{f}_i(\hat{d}) = \nabla f(x^{[i]}).$$

Also ist in $x^{[i]}$ die notwendige Bedingung erster Ordnung erfüllt und wir können das Verfahren beenden. Zusätzlich folgt aus Ungleichung (3.54) noch

$$d^\top \nabla^2 f(x^{[i]})d \ge 0 \qquad \forall d \in \mathbb{R}^n.$$

Dies ist gerade die notwendige Bedingung zweiter Ordnung für f. Wir haben so also auch das Abbruchkriterium $\hat{f}_i(d^{[i]}) = f(x^{[i]})$ entdeckt.
Je näher q_i bei Eins liegt, desto vertrauenswürdiger ist Δ_i. Zu gegebenen Zahlen $0 < \delta_1 < \delta_2 < 1$ überprüft man die folgenden Fälle:

(a) $q_i \in [\delta_1, \delta_2)$:
Die tatsächliche und die vorhergesagte Abnahme stimmen in etwa überein. Akzeptiere $x^{[i+1]} = x^{[i]} + d^{[i]}$ und verwende $\Delta_{i+1} \approx \Delta_i$.

(b) $q_i \geq \delta_2$:
Die tatsächliche Abnahme ist größer als erwartet. Akzeptiere $x^{[i+1]} = x^{[i]} + d^{[i]}$ und wähle $\Delta_{i+1} \geq \Delta_i$ in der Hoffnung, daß im nächsten Schritt ein noch größerer Abstieg erfolgt.

(c) $q_i < \delta_1$:
Die tatsächliche Abnahme ist deutlich kleiner als die vorhergesagte Abnahme. Die lokale Approximation ist nicht vertrauenswürdig: Setze $x^{[i+1]} := x^{[i]}$ (Nullschritt), verkleinere Δ_i und wiederhole den Schritt.

Im Detail erhalten wir den folgenden Algorithmus:

Algorithmus: Trust-Region-Newton-Verfahren

(i) Wähle einen Startvektor $x^{[0]} \in \mathbb{R}^n$ und Konstanten $0 < \delta_1 < \delta_2 < 1$, $0 < \sigma_1 < 1 < \sigma_2$, $\Delta_0 > 0$. Setze $i = 0$.

(ii) Berechne eine Lösung $d^{[i]}$ des Trust-Region-Teilproblems.

(iii) Falls $f(x^{[i]}) = \hat{f}_i(d^{[i]})$, STOP.

(iv) Berechne den Quotienten

$$q_i = \frac{f(x^{[i]}) - f(x^{[i]} + d^{[i]})}{\hat{f}_i(0) - \hat{f}_i(d^{[i]})}.$$

Falls $q_i \geq \delta_1$ (erfolgreicher Schritt):

1. Setze $x^{[i+1]} = x^{[i]} + d^{[i]}$.
2. Wähle

$$\Delta_{i+1} \in \begin{cases} [\sigma_1 \Delta_i, \Delta_i], & \text{falls } q_i \in [\delta_1, \delta_2), \\ [\Delta_i, \sigma_2 \Delta_i], & \text{falls } q_i \geq \delta_2 \end{cases}$$

Setze $i := i + 1$ und gehe zu (ii).

(v) Es gilt $q_i < \delta_1$ (Nullschritt):

1. Wähle $\Delta_{i+1} \in (0, \sigma_1 \Delta_i]$.
2. Setze $x^{[i+1]} = x^{[i]}$.

Setze $i := i + 1$ und gehe zu (ii).

3.12.1 Konvergenzanalyse

Wir folgen der Darstellung in Alt [Alt02]. Der folgende Hilfssatz gibt eine Abschätzung für den erwarteten Abstieg an.

Hilfssatz 3.12.2

Sei $d^{[i]} \in \mathbb{R}^n$ globale Lösung des Trust-Region-Teilproblems. Dann gilt

$$f(x^{[i]}) - \hat{f}_i(d^{[i]}) \geq \frac{1}{2}\|\nabla f(x^{[i]})\| \min\left\{\Delta_i, \frac{\|\nabla f(x^{[i]})\|}{\|\nabla^2 f(x^{[i]})\|}\right\}.$$

Beweis: $d = 0$ ist zulässig für das Trust-Region-Teilproblem und daher gilt

$$f(x^{[i]}) - \hat{f}_i(d^{[i]}) \geq f(x^{[i]}) - \hat{f}_i(0) = 0.$$

Würde hier Gleichheit gelten, so folgte aus den vorangegangenen Betrachtungen $\nabla f(x^{[i]}) = 0$ und die Behauptung wäre gezeigt. Sei nun $\nabla f(x^{[i]}) \neq 0$ und $d \in \mathbb{R}^n$ mit $\|d\| \leq \Delta_i$ beliebig. Dann gilt

$$\begin{aligned} f(x^{[i]}) - \hat{f}_i(d^{[i]}) &\geq f(x^{[i]}) - \hat{f}_i(d) \\ &= -\nabla f(x^{[i]})^\top d - \frac{1}{2} d^\top \nabla^2 f(x^{[i]}) d \\ &\geq -\nabla f(x^{[i]})^\top d - \frac{1}{2}\|\nabla^2 f(x^{[i]})\| \cdot \|d\|^2. \end{aligned}$$

Ist $\Delta_i\|\nabla^2 f(x^{[i]})\| \leq \|\nabla f(x^{[i]})\|$, so liefert das spezielle $d = -\Delta_i \nabla f(x^{[i]})/\|\nabla f(x^{[i]})\|$ die Beziehung

$$f(x^{[i]}) - \hat{f}_i(d^{[i]}) \geq \Delta_i\|\nabla f(x^{[i]})\| - \frac{1}{2}\Delta_i^2\|\nabla^2 f(x^{[i]})\| \geq \frac{1}{2}\Delta_i\|\nabla f(x^{[i]})\|.$$

Ist $\Delta_i\|\nabla^2 f(x^{[i]})\| > \|\nabla f(x^{[i]})\|$, so liefert das spezielle $d = -\nabla f(x^{[i]})/\|\nabla^2 f(x^{[i]})\|$ die Beziehung

$$f(x^{[i]}) - \hat{f}_i(d^{[i]}) \geq \frac{\|\nabla f(x^{[i]})\|^2}{\|\nabla^2 f(x^{[i]})\|} - \frac{1}{2}\frac{\|\nabla f(x^{[i]})\|^2}{\|\nabla^2 f(x^{[i]})\|} = \frac{1}{2}\frac{\|\nabla f(x^{[i]})\|^2}{\|\nabla^2 f(x^{[i]})\|}.$$

Beide Abschätzungen zusammen liefern die Behauptung. □

Hilfssatz 3.12.3

Es sei $f : \mathbb{R}^n \to \mathbb{R}$ zweimal stetig differenzierbar und $x^{[0]} \in \mathbb{R}^n$. Die Niveaumenge $lev(f, f(x^{[0]}))$ sei kompakt. Endet das Trust-Region-Newton-Verfahren nicht nach endlich vielen Schritten, dann gilt

$$\liminf_{i\to\infty} \|\nabla f(x^{[i]})\| = 0.$$

Beweis: Angenommen, die Behauptung ist falsch. Dann gibt es ein $\varepsilon > 0$ mit $\|\nabla f(x^{[i]})\| \geq \varepsilon$ für alle $i \in \mathbb{N}$. Wir zeigen, daß dann die folgenden Bedingungen gelten:

(a) $\sum_{i=0}^{\infty} \Delta_i < \infty$, was $\lim_{i\to\infty} \Delta_i = 0$ impliziert.

(b) $\lim_{i\to\infty} q_i = 1$, was $q_i \geq \delta_2$ und damit $\Delta_{i+1} \geq \Delta_i > 0$ für hinreichend großes i impliziert.

Beide Bedingungen zusammen sind widersprüchlich.
Zu (a): Gibt es nur endlich viele erfolgreiche Schritte, dann gilt

$$\Delta_{i+1} \leq \sigma_1 \Delta_i \leq \sigma_1^2 \Delta_{i-1} \leq \ldots \leq \sigma_1^{i+1-i_0} \Delta_{i_0}$$

für alle $i \geq i_0$. Wegen $\sigma_1 \in (0,1)$ gilt dann (a).
Gibt es unendlich viele erfolgreiche Schritte, so sei J die Indexmenge der erfolgreichen Schritte. Für $i \in J$ gilt $q_i \geq \delta_1$. Aus der Definition von q_i und Hilfssatz 3.12.2 folgt

$$f(x^{[i]}) - f(x^{[i+1]}) \geq \frac{\delta_1}{2} \|\nabla f(x^{[i]})\| \min\left\{\Delta_i, \frac{\|\nabla f(x^{[i]})\|}{\|\nabla^2 f(x^{[i]})\|}\right\}.$$

Wegen $\|\nabla f(x^{[i]})\| \geq \varepsilon$ folgt mit $\beta := \max_{x \in \mathrm{lev}(f, f(x^{[0]}))} \|\nabla^2 f(x)\|$ für $i \in J$

$$f(x^{[i]}) - f(x^{[i+1]}) \geq \frac{\delta_1 \varepsilon}{2} \min\left\{\Delta_i, \frac{\varepsilon}{\beta}\right\}. \tag{3.55}$$

Aus $f(x^{[i+1]}) \leq f(x^{[i]})$ und der Beschränktheit von $\{f(x^{[i]})\}$ (f ist stetig auf Kompaktum) folgt damit

$$\frac{\delta_1 \varepsilon}{2} \sum_{i\in J} \min\left\{\Delta_i, \frac{\varepsilon}{\beta}\right\} \leq \sum_{i\in J} \left(f(x^{[i]}) - f(x^{[i+1]})\right) \leq \sum_{i=0}^{\infty} \left(f(x^{[i]}) - f(x^{[i+1]})\right) < \infty.$$

Insbesondere kann $\min\left\{\Delta_i, \frac{\varepsilon}{\beta}\right\} = \frac{\varepsilon}{\beta}$ nur für endlich viele $i \in J$ gelten und es folgt $\sum_{i\in J} \Delta_i < \infty$.
Um $\sum_{i\notin J} \Delta_i < \infty$ zu zeigen, betrachten wir die (endlich vielen) Indizes, die zwischen zwei aufeinander folgenden erfolgreichen Schritten liegen. Sei $i \in J$ und $i+k \notin J$, $k = 1, \ldots, j$. Dann gilt $\Delta_{i+1} \leq \sigma_2 \Delta_i$, $i \in J$, und für $j > 1$ gilt

$$\Delta_{i+k} \leq \sigma_1 \Delta_{i+k-1} \leq \sigma_1^{k-1} \Delta_{i+1} \leq \sigma_2 \sigma_1^{k-1} \Delta_i, \qquad k = 2, \ldots, j.$$

Aufsummieren liefert

$$\sum_{k=1}^{j} \Delta_{i+k} \leq \frac{\sigma_2}{1-\sigma_1} \Delta_i$$

und damit

$$\sum_{i\notin J} \Delta_i \leq \frac{\sigma_2}{1-\sigma_1} \sum_{i\in J} \Delta_i < \infty.$$

Dies zeigt (a).

Zu (b): Es gilt

$$q_i - 1 = \frac{f(x^{[i]}) + \nabla f(x^{[i]})^\top d^{[i]} + \frac{1}{2}(d^{[i]})^\top \nabla^2 f(x^{[i]}) d^{[i]} - f(x^{[i]} + d^{[i]})}{f(x^{[i]}) - \hat{f}_i(d^{[i]})}.$$

Wegen $\Delta_i \to 0$, $\|\nabla f(x^{[i]})\| \geq \varepsilon$ und $\|\nabla f(x^{[i]})\|/\|\nabla^2 f(x^{[i]})\| \geq \varepsilon/\beta$ erhalten wir mit Hilfssatz 3.12.2 die untere Schranke $\varepsilon\Delta_i/2$ für den Nenner und damit

$$\begin{aligned} |q_i - 1| &\leq \frac{2}{\varepsilon\Delta_i} \left| f(x^{[i]}) + (\nabla f(x^{[i]}))^\top d^{[i]} + \frac{1}{2}(d^{[i]})^\top \nabla^2 f(x^{[i]}) d^{[i]} - f(x^{[i]} + d^{[i]}) \right| \\ &\overset{\|d^{[i]}\| \leq \Delta_i, Taylor}{\leq} \frac{1}{\varepsilon\|d^{[i]}\|} \left| (d^{[i]})^\top \left(\nabla^2 f(x^{[i]}) - \nabla^2 f(x^{[i]} + t_i d^{[i]}) \right) d^{[i]} \right| \end{aligned}$$

mit $t_i \in (0,1)$. Mit $\|d^{[i]}\| \leq \Delta_i \to 0$ folgt $q_i \to 1$. □

Der folgende Satz liefert eine globale Konvergenzaussage des Trust-Region-Newton-Verfahrens:

Satz 3.12.4 (Globaler Konvergenzsatz)

Es sei $f : \mathbb{R}^n \to \mathbb{R}$ zweimal stetig differenzierbar und $x^{[0]} \in \mathbb{R}^n$. Die Niveaumenge $lev(f, f(x^{[0]}))$ sei kompakt. Endet das Trust-Region-Newton-Verfahren nicht nach endlich vielen Schritten, dann gilt

$$\lim_{i\to\infty} \|\nabla f(x^{[i]})\| = 0.$$

Die Folge $\{x^{[i]}\}$ besitzt mindestens einen Häufungspunkt. Für jeden Häufungspunkt $\hat{x}$ gilt $\nabla f(\hat{x}) = 0$.

Beweis: Zunächst bemerken wir, daß ∇f Lipschitz-stetig auf der Niveaumenge ist, da diese kompakt und $\nabla^2 f$ stetig ist.
Angenommen, es gilt nicht $\nabla f(x^{[i]}) \to 0$. Dann gibt es $\varepsilon > 0$ und eine Teilfolge $\{x^{[i_k]}\}$ von $\{x^{[i]}\}$ mit

$$\|\nabla f(x^{[i_k]})\| \geq 2\varepsilon.$$

Nach Hilfssatz 3.12.3 gilt $\liminf_{i\to\infty} \|\nabla f(x^{[i]}\| = 0$ und somit gibt es unendlich viele Indizes i mit $\|\nabla f(x^{[i]})\| < \varepsilon$.
Für $k = 1, 2, \ldots$ muß es daher zu i_k ein $l_k > i_k$ geben mit

$$\|\nabla f(x^{[i]})\| \geq \varepsilon, \quad i_k \leq i < l_k, \quad \|\nabla f(x^{[l_k]})\| < \varepsilon.$$

Ist Schritt i mit $i_k \leq i < l_k$ erfolgreich, so folgt wegen $\|d^{[i]}\| = \|x^{[i+1]} - x^{[i]}\| \leq \Delta_i$ analog zu (3.55) die Ungleichung

$$f(x^{[i]}) - f(x^{[i+1]}) \geq \frac{\delta_1 \varepsilon}{2} \min\left\{ \|x^{[i+1]} - x^{[i]}\|, \frac{\varepsilon}{\beta} \right\} \geq 0. \tag{3.56}$$

Demnach fällt die Folge $\{f(x^{[i]})\}$ monoton und wegen der vorausgesetzten Kompaktheit der Niveaumenge ist f dort nach unten beschränkt. Damit konvergiert $\{f(x^{[i]})\}$. Damit folgt für hinreichend großes i zwangsläufig

$$\min\left\{\|x^{[i+1]}-x^{[i]}\|, \frac{\varepsilon}{\beta}\right\} = \|x^{[i+1]}-x^{[i]}\|, \qquad i_k \le i < l_k$$

und somit

$$f(x^{[i]}) - f(x^{[i+1]}) \ge \frac{\delta_1\varepsilon}{2}\|x^{[i+1]}-x^{[i]}\|, \qquad i_k \le i < l_k.$$

Diese Ungleichung gilt trivialerweise auch für Nullschritte, da dann $x^{[i+1]} = x^{[i]}$ gilt. Die Dreiecksungleichung liefert

$$\begin{aligned}
\|x^{[i_k]} - x^{[l_k]}\| &\le \sum_{i=i_k}^{l_k-1} \|x^{[i+1]} - x^{[i]}\| \\
&\le \frac{2}{\delta_1\varepsilon} \sum_{i=i_k}^{l_k-1} \left(f(x^{[i]}) - f(x^{[i+1]})\right) \\
&= \frac{2}{\delta_1\varepsilon} \left(f(x^{[i_k]}) - f(x^{[l_k]})\right).
\end{aligned}$$

Mit der Konvergenz der Funktionswerte konvergiert auch $\|x^{[i_k]}-x^{[l_k]}\| \to 0$. Die Lipschitz-Stetigkeit von ∇f liefert

$$\|\nabla f(x^{[i_k]}) - \nabla f(x^{[l_k]})\| \le L\|x^{[i_k]} - x^{[l_k]}\| \to 0.$$

Andererseits gilt aber

$$\|\nabla f(x^{[i_k]}) - \nabla f(x^{[l_k]})\| \ge \|\nabla f(x^{[i_k]})\| - \|\nabla f(x^{[l_k]})\| \ge 2\varepsilon - \varepsilon = \varepsilon.$$

Dies ist ein Widerspruch. Es gilt also $\nabla f(x^{[i]}) \to 0$. Wegen der vorausgesetzten Kompaktheit der Niveaumenge besitzt die Folge $\{x^{[i]}\}$ mindestens einen Häufungspunkt. Der Rest folgt aus der Stetigkeit von ∇f. □

Der folgende Satz behandelt die lokale Konvergenzgeschwindigkeit des Trust-Region-Verfahrens. Ein Beweis findet sich in Alt [Alt02], S. 152/153.

Satz 3.12.5 (Lokaler Konvergenzsatz)
Es sei $f : \mathbb{R}^n \to \mathbb{R}$ zweimal stetig differenzierbar und $x^{[0]} \in \mathbb{R}^n$. Das Trust-Region-Newton-Verfahren ende nicht nach endlich vielen Schritten und die Niveaumenge $\mathrm{lev}(f, f(x^{[0]}))$ sei kompakt. Dann gelten:

(a) Es gibt mindestens einen Häufungspunkt $\hat{x}$ von $\{x^{[i]}\}$ mit

$$\nabla f(\hat{x}) = 0, \qquad \nabla^2 f(\hat{x}) \text{ positiv semidefinit.}$$

(b) Ist $\hat{x}$ ein Häufungspunkt von $\{x^{[i]}\}$ mit positiv definiter Hessematrix $\nabla^2 f(\hat{x})$, dann konvergiert $\{x^{[i]}\}$ gegen $\hat{x}$ und das Trust-Region-Newton-Verfahren geht nach endlich vielen Schritten in das lokale Newtonverfahren über. Insbesondere vererben sich die Konvergenzeigenschaften (superlinear, quadratisch) des lokalen Newtonverfahrens.

Bemerkung 3.12.6

Der Idee der Quasi-Newton-Verfahren folgend, kann die exakte Hessematrix durch eine leichter zu berechnende Approximation $H \approx \nabla^2 f(x)$ ersetzt werden. Dabei können wir wieder auf die im Zusammenhang mit Quasi-Newton-Verfahren diskutierten Update-Formeln (BFGS, DFP, SR1, ...) zurück greifen. Die positive Definitheit ist im Zusammenhang von Trust-Region-Verfahren nicht wesentlich. Beachte auch, daß selbst die BFGS-Formel nicht notwendig positiv definite Matrizen liefert, da hier mangels einer Schrittweitenstrategie nicht garantiert werden kann, daß $y^\top d > 0$ gilt. Konvergenzbeweise finden sich in Jarre und Stoer [JS04] und Geiger und Kanzow [GK99]. Dabei wird u.a. verlangt, daß die Folge der Update-Matrizen beschränkt bleibt.

3.12.2 Lösung des Trust-Region-Teilproblems

Die wesentliche Schwierigkeit bei der Durchführung des Trust-Newton-Verfahrens besteht in der Lösung des Trust-Region-Teilproblems. Der folgende Satz charakterisiert die optimale Lösung des Problems.

Satz 3.12.7

$\hat{d}$ ist genau dann eine globale Lösung des quadratischen Teilproblems (3.51) und (3.52), wenn es ein $\lambda \in \mathbb{R}$ gibt mit

(a) $\lambda \geq 0$, $\|\hat{d}\| \leq \Delta_i$, $\lambda(\|\hat{d}\| - \Delta_i) = 0$;

(b) $(\nabla^2 f(x^{[i]}) + \lambda I)\hat{d} = -\nabla f(x^{[i]})$;

(c) $\nabla^2 f(x^{[i]}) + \lambda I$ *ist positiv semidefinit.*

Beweis:

$\Leftarrow$: Es gelten die Bedingungen (a)-(c) für $\hat{d}$. Sei $d \in \mathbb{R}^n$ mit $\|d\| \leq \Delta_i$ beliebig. Dann

folgt

$$\begin{aligned}
\hat{f}_i(d) - \hat{f}_i(\hat{d}) &= \nabla f(x^{[i]})^\top (d-\hat{d}) + \frac{1}{2} d^\top \nabla^2 f(x^{[i]}) d - \frac{1}{2} \hat{d}^\top \nabla^2 f(x^{[i]}) \hat{d} \\
&= \left(\nabla f(x^{[i]}) + \nabla^2 f(x^{[i]})\hat{d}\right)^\top (d-\hat{d}) + \frac{1}{2}(d-\hat{d})^\top \nabla^2 f(x^{[i]})(d-\hat{d}) \\
&\overset{(b)}{=} -\lambda (d-\hat{d})^\top \hat{d} + \frac{1}{2}(d-\hat{d})^\top \left(\nabla^2 f(x^{[i]}) + \lambda I\right)(d-\hat{d}) - \frac{1}{2}\lambda \|d-\hat{d}\|^2 \\
&\overset{(c)}{\geq} \frac{1}{2}\lambda(\|\hat{d}\|^2 - \|d\|^2) \\
&= \frac{1}{2}\lambda(\|\hat{d}\|^2 - \Delta_i^2) + \frac{1}{2}\lambda(\Delta_i^2 - \|d\|^2) \\
&\overset{(a)}{=} \frac{1}{2}\lambda(\Delta_i^2 - \|d\|^2) \\
&\geq 0.
\end{aligned}$$

$\Rightarrow$: Sei nun $\hat{d}$ ein globales Minimum. Ist $\|\hat{d}\| < \Delta_i$, so ist $\hat{d}$ auch ein globales Minimum von $\hat{f}_i$ und (a)-(c) folgen mit $\lambda = 0$ aus den notwendigen Bedingungen in Satz 3.1.1.

Ist $\|\hat{d}\| = \Delta_i$, so ist die Nebenbedingung

$$c(d) = \frac{1}{2}\left(d^\top d - \Delta_i^2\right) \leq 0$$

aktiv und es gilt $\nabla c(\hat{d}) = \hat{d} \neq 0$ (später heißt diese Bedingung „Linear Independence Constraint Qualification (LICQ)“). Wir verwenden nun die später im Abschnitt über notwendige Bedingungen für restringierte Optimierungsprobleme bewiesenen notwendigen Bedingungen erster und zweiter Ordnung. Diese besagen, daß es einen Multiplikator $\lambda \geq 0$ (wegen LICQ ist dieser sogar eindeutig) gibt mit

$$\begin{aligned}
\nabla \hat{f}_i(\hat{d}) + \lambda \nabla c(\hat{d}) &= 0, \\
\lambda(\|\hat{d}\| - \Delta_i) &= 0, \\
z^\top \nabla^2 (\hat{f}_i(\hat{d}) + \lambda I) z &\geq 0 \qquad \forall z \in \mathbb{R}^n \ : \ \nabla c(\hat{d})^\top z = 0.
\end{aligned}$$

Die erste Bedingung liefert gerade (b) und die zweite (a).

Zu zeigen bleibt, daß die dritte Bedingung auch für $\hat{d}^\top z = \nabla c(\hat{d})^\top z \neq 0$ gilt. Sei nun $z \in \mathbb{R}^n$ mit $\hat{d}^\top z \neq 0$. Setze

$$d := \hat{d} + \alpha z, \quad \alpha := -2\frac{z^\top \hat{d}}{\|z\|^2} \neq 0.$$

Wegen

$$d = \left(I - 2v \cdot v^\top\right)\hat{d}, \qquad v = \frac{z}{\|z\|}$$

entsteht d aus $\hat{d}$ durch Spiegelung an der zu z senkrechten Hyperebene (Householder-Transformation). Außerdem gilt $\|d\| = \|\hat{d}\| = \Delta_i$. Analog zum ersten Beweisteil zeigt man

$$\begin{aligned} 0 &\leq \hat{f}_i(d) - \hat{f}_i(\hat{d}) \\ &= -\lambda(d-\hat{d})^\top \hat{d} + \frac{1}{2}(d-\hat{d})^\top \left(\nabla^2 f(x^{[i]}) + \lambda I\right)(d-\hat{d}) - \frac{1}{2}\lambda\|d-\hat{d}\|^2 \\ &= \frac{1}{2}(d-\hat{d})^\top \left(\nabla^2 f(x^{[i]}) + \lambda I\right)(d-\hat{d}) - \frac{\lambda}{2}\underbrace{\left(\|\hat{d}\|^2 - \|d\|^2\right)}_{=0} \\ &= \frac{\alpha^2}{2} z^\top \left(\nabla^2 f(x^{[i]}) + \lambda I\right) z. \end{aligned}$$

Dies zeigt (c).

□

Es wird versucht, die Bedingungen (a)-(c) in Satz 3.12.7 algorithmisch zu erfüllen. Gelingt dies, so ist eine globale Lösung des Trust-Region-Teilproblems gefunden. Sei nun $H := \nabla^2 f(x^{[i]})$, $g := \nabla f(x^{[i]})$ und $\Delta := \Delta_i$. Die Hessematrix H ist symmetrisch (da f zweimal stetig differenzierbar ist) und besitzt daher eine Zerlegung

$$H = T^{-1}\Lambda T, \quad \Lambda = \operatorname{diag}(\lambda_1, \ldots, \lambda_n),$$

wobei Λ die Eigenwerte von H enthält.
Wir betrachten nun die Bedingungen (a)-(c) etwas genauer:

- Es bezeichne $d(\lambda)$ die Lösung (so sie existiert) des Gleichungssystems (b)

  $$(H + \lambda I)\, d = -g \quad \Leftrightarrow \quad T^{-1}(\Lambda + \lambda I)Td = -g \quad \Leftrightarrow \quad (\Lambda + \lambda I)Td = -Tg.$$

  Da Λ eine Diagonalmatrix ist, existiert $(\Lambda + \lambda I)^{-1}$ genau dann, wenn $\lambda \neq -\lambda_i$ für alle $i = 1, \ldots, n$ ist und es gilt

  $$d(\lambda) = -T^{-1}(\Lambda + \lambda I)^{-1}Tg = -(T^{-1}(\Lambda + \lambda I)T)^{-1}g = -(H + \lambda I)^{-1}g.$$

  In der Regel gilt $\|d(\lambda)\| \to \infty$ für $\lambda \to -\lambda_i$ mit $1 \leq i \leq n$.

- (a) impliziert:

  $$\begin{aligned} \|d\| < \Delta &\Rightarrow \lambda = 0, \\ \lambda > 0 &\Rightarrow \|d\| = \Delta. \end{aligned}$$

- Die Bedingung (c) ist erfüllt, wenn $\lambda \geq -\lambda_{min}$ gilt, wobei λ_{min} den kleinsten Eigenwert von H bezeichnet.

Wir versuchen nun λ so zu bestimmen, daß $\|d(\lambda)\| = \Delta$ gilt. Wegen der möglichen Unbeschränktheit von $\|d(\lambda)\|$ für $\lambda = -\lambda_i$ betrachten wir stattdessen die äquivalente **nichtlineare Gleichung**

$$\psi(\lambda) := \frac{1}{\Delta} - \frac{1}{w(\lambda)} = 0 \tag{3.57}$$

mit

$$w(\lambda) := \|d(\lambda)\|, \qquad w'(\lambda) = \frac{d(\lambda)^\top d'(\lambda)}{w(\lambda)}, \qquad \psi'(\lambda) = \frac{w'(\lambda)}{w(\lambda)^2}.$$

Anwendung des **Newtonverfahrens** auf (3.57) zur Bestimmung einer Nullstelle $\hat{\lambda}$ von ψ liefert für den Startwerte $\lambda^{[0]}$ die Iteration

$$\lambda^{[k+1]} = \lambda^{[k]} - \frac{\psi(\lambda^{[k]})}{\psi'(\lambda^{[k]})} = \lambda^{[k]} - \frac{w(\lambda^{[k]})^2}{w'(\lambda^{[k]})} \left(\frac{1}{\Delta} - \frac{1}{w(\lambda^{[k]})} \right), \qquad k = 0, 1, 2, \ldots. \tag{3.58}$$

Die Ableitung $d'(\lambda)$ kann wie folgt bestimmt werden. Es gilt $(H + \lambda I)d(\lambda) = -g$. Differentiation dieser Gleichung bzgl. λ liefert

$$d'(\lambda) = -(H + \lambda I)^{-1} d(\lambda).$$

Damit wird

$$w'(\lambda) = \frac{d(\lambda)^\top d'(\lambda)}{w(\lambda)} = -\frac{d(\lambda)^\top (H + \lambda I)^{-1} d(\lambda)}{w(\lambda)}.$$

Cholesky-Zerlegung $H + \lambda I = L \cdot L^\top$ liefert

$$d^\top (H + \lambda I)^{-1} d = d^\top L^{-\top} \underbrace{L^{-1} d}_{=q} = q^\top q = \sum_{i=1}^{n} q_i^2$$

mit q aus $Lq = d$.

Insgesamt erhält man aus den obigen Betrachtungen den folgenden Algorithmus zur Lösung des Trust-Region-Teilproblems nach Moré und Sorensen:

Algorithmus: Trust-Region-Hilfsproblem

(i) Gegeben sei $x^{[i]} \in \mathbb{R}^n$, $\Delta := \Delta_i > 0$, $H := \nabla^2 f(x^{[i]})$, $g := \nabla f(x^{[i]})$.

(ii) Berechne (falls möglich) die Cholesky-Zerlegung $H = L \cdot L^\top$.
Falls diese existiert, löse $L \cdot L^\top d = -g$.
Falls $\|d\| \leq \Delta$: STOP.

(iii) Wähle $\lambda > 0$.

(iv) Berechne (falls möglich) die Cholesky-Zerlegung $H + \lambda I = L \cdot L^\top$.
Falls diese nicht existiert, vergrößere λ und gehe zu (iv).

(v) Löse $L \cdot L^\top d = -g$ und setze $w := \|d\|$.
Löse $Lq = d$ und berechne

$$w' := -\frac{1}{w} \sum_{i=1}^{n} q_i^2, \qquad \lambda := \lambda - \frac{w^2}{w'} \left(\frac{1}{\Delta} - \frac{1}{w} \right).$$

Falls (3.57) genau genug erfüllt ist, STOP.
Gehe zu (iv).

Erläuterungen:

- In (ii) wird getestet, ob die Bedingungen (a)-(c) für $\lambda = 0$ bereits erfüllt sind.
- Ist (ii) nicht erfüllt, so muß für $\lambda > 0$ nach (a) zwangsläufig $\|d\| = \Delta$ gelten und das oben beschriebene Newtonverfahren kommt zum Einsatz.

Bemerkung 3.12.8

In der Regel genügt es, das Trust-Region-Teilproblem mit geringer Genauigkeit zu lösen. Zudem gibt es zahlreiche alternative Methoden zur Lösung des Trust-Region-Teilproblems, vgl. Geiger und Kanzow [GK99]:

- *Anstatt das beschränkte Teilproblem zu lösen, wird dieses äquivalent auf die unrestringierte Minimierung einer exakten Penaltyfunktion zurückgeführt. Suchrichtungen werden hierbei durch ein nichtglattes Newtonverfahren berechnet. Schrittweiten werden mit Hilfe der exakten Penaltyfunktion und dem Armijoverfahren bestimmt.*
- *Eine Modifikation des CG-Verfahrens zur approximativen Lösung des Trust-Region-Teilproblems führt auf* **inexakte Trust-Region-Verfahren**.
- *Teilraum-Trust-Region-Verfahren schränken den zulässigen Bereich des Teilproblems zusätzlich durch $d \in V_k \subseteq \mathbb{R}^n$ ein. Die Dimension von V_k ist dabei in der Regel klein,*

z.B. kann V_k aufgespannt sein durch die negative Gradientenrichtung und die Newtonrichtung. Dies erlaubt dann mitunter eine effiziente Lösung des Teilproblems.

In Alt [Alt02] ist eine Variante des Trust-Region-Verfahrens angegeben, bei dem nicht Δ gesteuert wird, sondern λ.

3.13 Gauss-Newton-Verfahren und nichtlineare Ausgleichsprobleme

Ein Experiment liefert die Meßpunkte (t_i, y_i), $i = 1, \ldots, m$. Der dem Experiment zu Grunde liegende Vorgang werde durch die Funktion $f(t, p)$ modelliert, die einen funktionalen Zusammenhang zwischen den Meßstellen t_i und den Meßwerten y_i herstellt. Allerdings hängt die Funktion auch noch vom unbekannten Parameter $p \in \mathbb{R}^{n_p}$ ab. In der Praxis sind die Meßwerte verrauscht bzw. fehlerbehaftet, so daß es in der Regel keinen Parameter p gibt, der die Meßpunkte exakt reproduziert. Daher wird versucht, die Meßpunkte so gut wie möglich zu approximieren, indem

$$\frac{1}{2} \sum_{i=1}^{m} \left(y_i - f(t_i, p) \right)^2$$

bezüglich p minimiert wird. Häufig sind zusätzlich noch Nebenbedingungen an den Parameter p gegeben. Ein erstes Ausgleichsproblem wurde in Beispiel 2.4.1 diskutiert. Die Funktion f muß allerdings nicht immer analytisch gegeben sein, wie das folgende Beispiel zeigt.

Beispiel 3.13.1

Wir betrachten die chemische Reaktion

$$C_1 + C_2 \quad \underset{k_-}{\overset{k_+}{\rightleftharpoons}} \quad C_3$$

Das **Massenwirkungsgesetz** *besagt, daß die Reaktionsgeschwindigkeit der Stoffe proportional zum Produkt der Konzentrationen der Reaktionspartner ist. Bezeichnen $c_i(t)$, $i = 1, 2, 3$ die Konzentrationen der Stoffe C_i, $i = 1, 2, 3$ zum Zeitpunkt t, so führt das Massenwirkungsgesetz auf das Differentialgleichungssystem*

$$\begin{aligned} \dot{c}_1(t) &= -k_+ \cdot c_1(t) \cdot c_2(t) + k_- \cdot c_3(t), \\ \dot{c}_2(t) &= -k_+ \cdot c_1(t) \cdot c_2(t) + k_- \cdot c_3(t), \\ \dot{c}_3(t) &= k_+ \cdot c_1(t) \cdot c_2(t) - k_- \cdot c_3(t) \end{aligned}$$

mit Konstanten k_+ und k_-. Zum Zeitpunkt $t = 0$ seien die Anfangskonzentrationen bekannt. Dies liefert Anfangsbedingungen $c_i(0) = c_{i,0}$, $i = 1, 2, 3$. Zusammen mit den Differentialgleichungen liegt also ein **Anfangswertproblem** *vor. Dieses besitzt (unter geeigneten Voraussetzungen) auf dem Zeitintervall $[0, T]$ eine Lösung, die allerdings noch von*

den Konstanten k_+ und k_- abhängt. Um diese Abhängigkeit anzudeuten, bezeichnen wir die Lösung mit $c_i(t; k_+, k_-)$ für $t \in [0,T]$. In der Regel sind die Konstanten k_+ und k_- nicht bekannt. Aber es ist häufig möglich, die Konzentrationen einiger (oder sogar aller) beteiligter Stoffe zu verschiedenen Zeitpunkten $t_j \in [0,T]$, $j = 1,\dots,N$ zu messen. Zur Vereinfachung nehmen wir an, daß nur die Konzentration c_3 gemessen werden kann. Dies liefert dann Meßwerte $y_j \approx c_3(t_j)$, $j = 1,\dots,N$. Natürlich sind die Meßwerte in der Praxis fehlerbehaftet. Die Aufgabe besteht nun darin, k_+ und k_- so zu bestimmen, daß das mathematische Modell (Anfangswertproblem) die Meßdaten möglichst gut widergibt. Dazu lösen wir das nichtlineare Ausgleichsproblem (Least-Squares-Problem)

$$\min_{k_+,k_-\in\mathbb{R}} \frac{1}{2}\sum_{j=1}^{N} \left(y_j - c_3(t_j; k_+, k_-)\right)^2 .$$

Zu beachten ist, daß ein Anfangswertproblem gelöst werden muß, um die Werte $c_3(t_j, k_+, k_-)$ zu bekommen. Dies ist in der Regel wiederum nur numerisch möglich.

Das resultierende Optimierungsproblem ist ein spezielles nichtlineares Ausgleichsproblem. Ein allgemeines (unbeschränktes) Ausgleichsproblem lautet:

Nichtlineares Ausgleichsproblem (Least-Squares-Problem):
Gegeben sei die Funktion $F : \mathbb{R}^n \to \mathbb{R}^m$. Finde $x \in \mathbb{R}^n$, so daß

$$f(x) := \frac{1}{2}\|F(x)\|^2 = \frac{1}{2}F(x)^\top F(x) = \frac{1}{2}\sum_{i=1}^{m} F_i(x)^2$$

minimal wird.

In der Regel ist m sehr viel größer als n (es liegen also mehr Meßwerte als Parameter vor). Im folgenden sei F hinreichend oft differenzierbar.

3.13.1 Das lokale Gauss-Newton-Verfahren

In einem lokalen Minimum $\hat{x}$ von f gilt notwendig

$$0 = \nabla f(\hat{x}) = F'(\hat{x})^\top F(\hat{x}).$$

Diese **nichtlineare Gleichung** könnten wir wieder mit dem **(lokalen) Newtonverfahren**

$$x^{[i+1]} = x^{[i]} - \nabla^2 f(x^{[i]})^{-1}\nabla f(x^{[i]}), \qquad i = 0,1,2,\dots$$

lösen. Allerdings wird in der Iterationsvorschrift die Hessematrix

$$\nabla^2 f(x) = F'(x)^\top F'(x) + \sum_{i=1}^{m} F_i(x)\nabla^2 F_i(x)$$

benötigt. Diese ist in der Regel ($m \gg n!$) jedoch sehr aufwendig zu berechnen.
Idee: Wir vernachlässigen die Summanden $F_i(x)\nabla^2 F_i(x)$ in der Hessematrix mit der Begründung, daß man für ein „vernünftiges“ Ausgleichsproblem $F(\hat{x}) \approx 0$ erwarten kann. Wir erhalten den

Algorithmus: Lokales Gauss-Newton-Verfahren

(i) Wähle einen Startvektor $x^{[0]} \in \mathbb{R}^n$ und setze $i = 0$.

(ii) Falls ein Abbruchkriterium erfüllt ist, STOP.

(iii) Berechne (falls möglich) die Suchrichtung

$$d^{[i]} = -\left(F'(x^{[i]})^\top F'(x^{[i]})\right)^{-1} F'(x^{[i]})^\top F(x^{[i]}).$$

Setze $x^{[i+1]} = x^{[i]} + d^{[i]}$, $i := i+1$ und gehe zu (ii).

Bemerkung 3.13.2

Das Gauss-Newton-Verfahren kann auch wie folgt motiviert werden: Wir ersetzen die nichtlineare Funktion F in f durch ihre **Linearisierung in** $x^{[i]}$*:*

$$\hat{F}_i(x) := F(x^{[i]}) + F'(x^{[i]})(x - x^{[i]}).$$

Im Punkt $x^{[i]}$ wird nun das lineare Ausgleichsproblem

$$\hat{f}(x) = \frac{1}{2}\|\hat{F}_i(x)\|^2 = \frac{1}{2}\|F(x^{[i]}) + F'(x^{[i]})(x - x^{[i]})\|^2 \to \min.$$

für x gelöst. Die Lösung ist dann der neue Iterationspunkt $x^{[i+1]}$. Dieser erfüllt notwendig die **Gauss'sche Normalengleichung**

$$F'(x^{[i]})^\top F'(x^{[i]})(x^{[i+1]} - x^{[i]}) = -F'(x^{[i]})^\top F(x^{[i]}),$$

welche gerade die Bedingung $\nabla \hat{f}(x^{[i+1]}) = 0$ darstellt. Umformung liefert das lokale Gauss-Newton-Verfahren. Numerisch wird $x^{[i+1]}$ aus Stabilitätsgründen **nicht** *über die Normalengleichungen, sondern über das lineare Ausgleichsproblem bestimmt, vgl. Abschnitt 3.13.3.*

Zunächst diskutieren wir die lokale Konvergenz des Gauss-Newton-Verfahrens im Fall $F(\hat{x}) = 0$, vgl. Kosmol [Kos93], S. 81.

Satz 3.13.3

Es sei $D \subseteq \mathbb{R}^n$ offen und $F : D \to \mathbb{R}^m$ sei stetig differenzierbar. Sei $\hat{x} \in D$ mit $F(\hat{x}) = 0$.

$F'(\hat{x})^\top F'(\hat{x})$ sei invertierbar. Dann existiert ein $r > 0$, so daß das lokale Gauss-Newton-Verfahren für jeden Startwert $x^{[0]} \in U_r(\hat{x})$ wohldefiniert ist und $\{x^{[i]}\}$ konvergiert superlinear gegen $\hat{x}$.
Ist F' zusätzlich Lipschitz-stetig in D, so konvergiert $\{x^{[i]}\}$ sogar quadratisch.

Beweis: Wir interpretieren das Gauss-Newton-Verfahren als Fixpunktiteration für die Fixpunktfunktion

$$G(x) = x - \left(F'(x)^\top F'(x)\right)^{-1} F'(x)^\top F(x).$$

Beachte, daß $\hat{x}$ wegen $F(\hat{x}) = 0$ und der Invertierbarkeit von $F'(\hat{x})^\top F'(\hat{x})$ ein Fixpunkt von G ist. Definiere $H(x) := F'(x)^\top F'(x)$. Da $H(\hat{x})$ invertierbar ist, gibt es nach Hilfssatz 3.9.9 ein $r_1 > 0$, so daß $H(x)$ invertierbar ist mit $\|H(x)^{-1}\| \le C_1$ für alle $x \in U_{r_1}(\hat{x})$. Definiere

$$\mu := \frac{1}{\|H(\hat{x})^{-1}\|}.$$

Weiter gilt für alle $x \in U_{r_1}(\hat{x})$:

$$\begin{aligned}
\|G(x) - \hat{x}\| &= \|H(\hat{x})^{-1} H(\hat{x})(G(x) - \hat{x})\| \\
&\le \|H(\hat{x})^{-1}\| \cdot \|H(\hat{x})(G(x) - \hat{x})\| \\
&= \frac{1}{\mu} \|H(\hat{x})(G(x) - \hat{x})\| \\
&\le \frac{1}{\mu} \left(\|H(\hat{x}) - H(x)\| \cdot \|G(x) - \hat{x}\| + \|H(x)(G(x) - \hat{x})\| \right).
\end{aligned}$$

Folglich gilt für alle $x \in U_{r_1}(\hat{x})$:

$$\mu \|G(x) - \hat{x}\| \le \|H(\hat{x}) - H(x)\| \cdot \|G(x) - \hat{x}\| + \|H(x)(G(x) - \hat{x})\|. \tag{3.59}$$

Da H stetig ist, gibt es für jedes $\varepsilon > 0$ ein $r_2 > 0$ mit

$$\|H(x) - H(\hat{x})\| \le \varepsilon \qquad \forall x \in U_{r_2}(\hat{x}).$$

Ebenso gilt $\|F'(x)^\top\| \le M$ für alle $x \in U_{r_1}(\hat{x})$ und für ε gibt es ein $r_3 > 0$ mit

$$\|F'(x) - F'(\hat{x})\| \le \varepsilon/2 \qquad \forall x \in U_{r_3}(\hat{x}).$$

Damit folgt

$$\|F'(x) - F'(\hat{x} + t(x - \hat{x}))\| \le \|F'(x) - F'(\hat{x})\| + \|F'(\hat{x}) - F'(\hat{x} + t(x - \hat{x}))\| \le \varepsilon$$

für alle $x \in U_{r_3}(\hat{x})$ und alle $0 \le t \le 1$. Sei nun $r := \min\{r_1, r_2, r_3\}$. Mit $F(\hat{x}) = 0$ folgt

dann

$$\begin{aligned}
\|H(x)(G(x)-\hat{x})\| &= \|H(x)(G(x)-\hat{x}) - H(x)(G(x)-x) - F'(x)^\top F(x) + F'(x)^\top F(\hat{x})\| \\
&= \|F'(x)^\top F'(x)(x-\hat{x}) - F'(x)^\top F(x) + F'(x)^\top F(\hat{x})\| \\
&\le \|F'(x)^\top\| \cdot \|F'(x)(x-\hat{x}) - F(x) + F(\hat{x})\| \\
&\le M \cdot \|F'(x)(x-\hat{x}) - F(x) + F(\hat{x})\| \\
&= M \cdot \left\| \int_0^1 \left(F'(x) - F'(\hat{x}+t(x-\hat{x}))\right)(x-\hat{x}) dt \right\| \\
&\le M \cdot \int_0^1 \underbrace{\|F'(x) - F'(\hat{x}+t(x-\hat{x}))\|}_{\le \varepsilon} dt \cdot \|x-\hat{x}\| \\
&\le M\varepsilon \|x-\hat{x}\| \qquad (3.60)
\end{aligned}$$

für alle $x \in U_r(\hat{x})$.
Zusammen mit (3.59) folgt dann

$$\mu\|G(x)-\hat{x}\| \le \varepsilon\|G(x)-\hat{x}\| + M\varepsilon\|x-\hat{x}\| \qquad \forall x \in U_r(\hat{x}).$$

Umstellen liefert

$$\|G(x)-\hat{x}\| \le \frac{M\varepsilon}{\mu-\varepsilon}\|x-\hat{x}\| \qquad \forall x \in U_r(\hat{x}). \qquad (3.61)$$

Wählt man $\varepsilon > 0$ so, daß $\frac{M\varepsilon}{\mu-\varepsilon} < 1$ gilt, so folgt die Konvergenz aus Hilfssatz 3.9.5.

Superlineare Konvergenz:

Die superlineare Konvergenz folgt aus (3.61), denn für $x^{[i]} \to \hat{x}$ gilt für hinreichend großes i

$$\frac{\|x^{[i+1]}-\hat{x}\|}{\|x^{[i]}-\hat{x}\|} = \frac{\|G(x^{[i]})-\hat{x}\|}{\|x^{[i]}-\hat{x}\|} \le \frac{M\varepsilon}{\mu-\varepsilon},$$

wobei $\varepsilon > 0$ beliebig ist. Für $\varepsilon \to 0$ zeigt dies die superlineare Konvergenz.

Quadratische Konvergenz:

Ist F' Lipschitz-stetig mit Konstante L, so gilt in (3.60) sogar

$$\begin{aligned}
\|H(x)(G(x)-\hat{x})\| &\le M \cdot \int_0^1 \underbrace{\|F'(x) - F'(\hat{x}+t(x-\hat{x}))\|}_{\le L(1-t)\|x-\hat{x}\|} dt \cdot \|x-\hat{x}\| \\
&\le M\frac{L}{2}\|x-\hat{x}\|^2.
\end{aligned}$$

Zusammen mit (3.59) folgt dann nach Umformung

$$\|G(x)-\hat{x}\| \le \frac{M \cdot \frac{L}{2}}{\mu-\varepsilon}\|x-\hat{x}\|^2 \qquad \forall x \in U_r(\hat{x}).$$

Also die quadratische Konvergenz. □

In der Praxis kann man nicht erwarten, daß $f(\hat{x}) = 0$ bzw. $F(\hat{x}) = 0$ gilt. Ist jedoch $\|F(\hat{x})\|$ hinreichend klein, so kann man zumindest lokal lineare Konvergenz nachweisen:

Satz 3.13.4

Es sei $D \subseteq \mathbb{R}^n$ offen und $F : D \to \mathbb{R}^m$ sei stetig differenzierbar. F' sei Lipschitz-stetig in D mit Konstante L. Sei $\hat{x} \in D$ stationärer Punkt mit $\nabla f(\hat{x}) = F'(\hat{x})^\top F(\hat{x}) = 0$. Desweiteren sei $Rang(F'(\hat{x})) = n$ und es gelte

$$\mu = \frac{1}{\| \left(F'(\hat{x})^\top F'(\hat{x})\right)^{-1} \|} > 1, \qquad L\|F(\hat{x})\| \le \|(F'(\hat{x})^\top F'(\hat{x}))^{-1}\| = \frac{1}{\mu}. \tag{3.62}$$

Dann existiert ein $r > 0$, so daß das lokale Gauss-Newton-Verfahren für jeden Startwert $x^{[0]} \in U_r(\hat{x})$ wohldefiniert ist und $\{x^{[i]}\}$ konvergiert linear gegen $\hat{x}$.

Beweis: In (3.61) fügen wir anstatt $F'(x)^\top F(\hat{x})$ den Term $F'(\hat{x})^\top F(\hat{x})$ ein und erhalten analog

$$\begin{aligned}
\|H(x)(G(x) - \hat{x})\| &= \|H(x)(G(x) - \hat{x}) - H(x)(G(x) - x) - F'(x)^\top F(x) + F'(\hat{x})^\top F(\hat{x})\| \\
&= \|F'(x)^\top F'(x)(x - \hat{x}) - F'(x)^\top F(x) + F'(x)^\top F(\hat{x}) - F'(x)^\top F(\hat{x}) \\
&\qquad + F'(\hat{x})^\top F(\hat{x})\| \\
&\le \underbrace{\|F'(x)^\top\|}_{\le M} \cdot \underbrace{\|F'(x)(x - \hat{x}) - F(x) + F(\hat{x})\|}_{\le \varepsilon\|x - \hat{x}\|} \\
&\quad + \underbrace{\|F'(\hat{x})^\top - F'(x)^\top\|}_{\le L\|x - \hat{x}\|} \cdot \underbrace{\|F(\hat{x})\|}_{\le 1/(L\mu)} \\
&\le \left(M\varepsilon + \frac{1}{\mu}\right) \|x - \hat{x}\|.
\end{aligned}$$

Zusammen mit (3.59) folgt dann nach Umformung

$$\|G(x) - \hat{x}\| \le \frac{M\varepsilon + \frac{1}{\mu}}{\mu - \varepsilon}\|x - \hat{x}\| \qquad \forall x \in U_r(\hat{x}).$$

Beachte hierbei, daß die Bedingung $\text{Rang}(F'(\hat{x})) = n$ die Invertierbarkeit von $F'(\hat{x})^\top F'(\hat{x})$ sichert.

Die Konstante wird kleiner als eins, wenn

$$\frac{M\varepsilon + \frac{1}{\mu}}{\mu - \varepsilon} < 1 \qquad \Leftrightarrow \qquad M\varepsilon + \frac{1}{\mu} < \mu - \varepsilon \qquad \Leftrightarrow \qquad \varepsilon(M + 1) < \mu - \frac{1}{\mu} = \frac{\mu^2 - 1}{\mu}.$$

Unter der Voraussetzung $\mu > 1$ gibt es solch ein $\varepsilon > 0$. □

Bemerkung 3.13.5

- *Jarre und Stoer [JS04] zeigen auf S. 188, daß das lokale Gauss-Newton-Verfahren unter geeigneten Voraussetzungen linear gegen einen stationären Punkt konvergiert, falls die Eigenwerte der Matrix*

$$R := \left(F'(\hat{x})^\top F'(\hat{x})\right)^{-1/2} \left(\sum_{i=1}^{m} F_i(\hat{x}) \nabla^2 F_i(\hat{x})\right) \left(F'(\hat{x})^\top F'(\hat{x})\right)^{-1/2}$$

die Bedingungen

$$-1 < \lambda_{min}(R) \leq \lambda_{max}(R) < 1$$

erfüllen.

- *Es gibt auch Quasi-Newton Ansätze für Ausgleichsprobleme. Hierbei wird der aufwendig zu berechnende Anteil $\sum_{j=1}^{m} F_j(x^{[i]}) \nabla^2 F_j(x^{[i]})$ in der Hessematrix von f durch eine Update-Matrix H_i ersetzt. Hierbei erweist sich die PSB-Update-Formel als geeignet, vgl. Jarre und Stoer [JS04], S. 190.*

3.13.2 Globalisierung des Gauss-Newton-Verfahrens

Es liegt nahe, das lokale Gauss-Newton-Verfahren durch Verwendung einer Schrittweitenstrategie zu globalisieren.

Algorithmus: Globales Gauss-Newton-Verfahren

(i) Wähle einen Startvektor $x^{[0]} \in \mathbb{R}^n$ und setze $i = 0$.

(ii) Falls ein Abbruchkriterium erfüllt ist, STOP.

(iii) Berechne (falls möglich) die Suchrichtung

$$d^{[i]} = -\left(F'(x^{[i]})^\top F'(x^{[i]})\right)^{-1} F'(x^{[i]})^\top F(x^{[i]}).$$

(iv) Berechne eine Schrittweite $\alpha_i > 0$ mit einer effizienten Schrittweitenstrategie.

(v) Setze $x^{[i+1]} = x^{[i]} + \alpha_i d^{[i]}$, $i := i + 1$ und gehe zu (ii).

Durch Zurückführung auf ein allgemeines Abstiegsverfahren beweist Kosmol [Kos93] auf S. 135 den folgenden Satz:

Satz 3.13.6

Es sei $D \subseteq \mathbb{R}^n$ offen und $F : D \to \mathbb{R}^m$ sei stetig differenzierbar. Es sei $x^{[0]} \in D$ und

Rang$(F'(x)) = n$ *für alle* $x \in lev(f, f(x^{[0]}))$. *Desweiteren sei* $F'(x)^\top F'(x)$ *gleichmäßig positiv definit auf* $lev(f, f(x^{[0]}))$, *d.h. es gebe Konstanten* $\mu_1, \mu_2 > 0$ *mit*

$$\mu_1 \|d\|^2 \leq d^\top F'(x)^\top F'(x) d \leq \mu_2 \|d\|^2 \qquad \forall x \in lev(f, f(x^{[0]})),\ d \in \mathbb{R}^n.$$

Dann ist das Gauss-Newton-Verfahren mit Startwert $x^{[0]}$ *durchführbar und es gilt*

$$\liminf_{i \to \infty} \|\nabla f(x^{[i]})\| = 0.$$

Darüber hinaus kann unter der Voraussetzung $F(\hat{x}) = 0$ noch gezeigt werden, daß das globale Verfahren in das lokale Verfahren übergeht und somit quadratische Konvergenz erzielt werden kann, vgl. Kosmol [Kos93], S. 135.

3.13.3 Lösung des linearen Ausgleichsproblems

Wir gehen noch kurz auf die Lösung des linearen Ausgleichsproblems ein und betrachten:

Lineares Ausgleichsproblem:
Gegeben seien $A \in \mathbb{R}^{m \times n}$ und $b \in \mathbb{R}^m$. Finde $x \in \mathbb{R}^n$, so daß

$$f(x) := \frac{1}{2}\|Ax + b\|^2 = \frac{1}{2}(Ax + b)^\top (Ax + b)$$

minimal wird.

Wir setzen im folgenden **nicht** voraus, daß A vollen Rang hat. Der Rang von A sei r. Dann liefert die QR-Zerlegung von A mit Spaltenpivoting die Zerlegung

$$AP = Q \begin{pmatrix} R \\ \Theta \end{pmatrix},$$

wobei $P \in \mathbb{R}^{n \times n}$ Permutationsmatrix und $Q \in \mathbb{R}^{m \times m}$ orthogonal ist. Desweiteren gilt $R = (R_1 \mid R_2) \in \mathbb{R}^{r \times r} \times \mathbb{R}^{r \times (n-r)}$, wobei $R_1 \in \mathbb{R}^{r \times r}$ invertierbar ist.
Da sich die euklidische Norm unter orthogonalen Transformationen nicht ändert, erhalten wir für die Zielfunktion

$$\begin{aligned} \frac{1}{2}\|Ax + b\|^2 &= \frac{1}{2}\|Q^\top AP \underbrace{P^\top x}_{=z} + \underbrace{Q^\top b}_{=(c_1, c_2)^\top}\|^2 \\ &= \frac{1}{2}\left\| \begin{pmatrix} R \\ 0 \end{pmatrix} z + \begin{pmatrix} c_1 \\ c_2 \end{pmatrix} \right\|^2 \\ &= \frac{1}{2}\|Rz + c_1\|^2 + \frac{1}{2}\|c_2\|^2. \end{aligned}$$

Die Lösung ist festgelegt durch

$$Rz = -c_1 \quad \overset{z=(z_1,z_2)^\top}{\Leftrightarrow} \quad R_1 z_1 + R_2 z_2 = -c_1.$$

Da R_1 regulär ist, ist diese Gleichung für $z_1 = -R_1^{-1} c_1$ und $z_2 = 0$ erfüllt. Eine Lösung des linearen Ausgleichsproblems ist also

$$x = P \begin{pmatrix} -R_1^{-1} c_1 \\ 0 \end{pmatrix}.$$

Bemerkung 3.13.7

Hat A vollen Rang, so ist die Matrix R invertierbar.
Ein alternativer Weg zur Lösung des linearen Ausgleichsproblems führt zur **Gauss'schen Normalengleichung**

$$A^\top A x = -A^\top b,$$

welche gerade die notwendige Bedingung $\nabla f(x) = 0$ darstellt. Die Matrix $A^\top A$ ist positiv definit, falls A vollen Rang hat. Allerdings ist diese Variante aus numerischer Sicht sehr problematisch, da die Matrix $A^\top A$ in der Regel eine deutlich schlechtere Kondition als A (welche häufig schon sehr schlecht ist) aufweist.

Kapitel 4

Konvexe Optimierung

Wir diskutieren konvexe Optimierungsprobleme. Es sei $S \subseteq \mathbb{R}^n$ eine konvexe Menge. $\mathcal{I} = \{1, \ldots, m\}$ und $\mathcal{J} = \{1, \ldots, p\}$ seien endliche Indexmengen mit $|\mathcal{I}| = m$ und $|\mathcal{J}| = p$. Die Funktionen $f : S \to \mathbb{R}$ und $g_i : S \to \mathbb{R}$, $i \in \mathcal{I}$ seien konvex. Darüber hinaus seien affin-lineare Funktionen

$$h_j(x) = a_j^\top x + b_j, \qquad a_j \in \mathbb{R}^n,\ b_j \in \mathbb{R},\ j \in \mathcal{J}$$

gegeben (diese sind ebenfalls konvex). Zur Abkürzung setze $g = (g_i \mid i \in \mathcal{I})$ und $h = (h_j \mid j \in \mathcal{J})$.

Konvexes Optimierungsproblem:
Minimiere $f(x)$ unter den Nebenbedingungen $x \in S$ und

$$\begin{aligned} g_i(x) &\leq 0, \qquad i \in \mathcal{I}, \\ h_j(x) &= 0, \qquad j \in \mathcal{J}. \end{aligned}$$

Natürlich setzen wir voraus, daß der **zulässige Bereich** Σ nichtleer ist:

$$\Sigma := \{x \in S \mid g_i(x) \leq 0,\ i \in \mathcal{I},\ h_j(x) = 0,\ j \in \mathcal{J}\} \neq \emptyset.$$

Beachte, daß Σ konvex ist, da S, g und h konvex sind (Beweis?). Nach Satz 3.3.4 wissen wir bereits, daß die Menge der globalen Minima konvex ist und daß jedes lokale Minimum des konvexen Minimierungsproblems zugleich globales Minimum ist.
Die Menge S enthält häufig Vorzeichenbeschränkungen:

$$S = \{x \in \mathbb{R}^n \mid x \geq 0\}.$$

Sie kann aber auch kompliziertere Mengen beschreiben, die nicht durch endlich viele (konvexe) Ungleichungen $g(x) \leq 0$ und (affine) Gleichungen $h(x) = 0$ beschrieben werden können. Beachte, daß die Beschränkung $x \geq 0$ formal auch als $g(x) := -x \leq 0$ geschrieben werden kann. Insofern ist die Formulierung des konvexen Optimierungsproblems häufig nicht eindeutig.

Der einfachste Vertreter der Klasse der konvexen Optimierungsprobleme ist das lineare Optimierungsproblem, bei dem sämtliche Funktionen affin-linear sind und $S = \mathbb{R}^n$ gilt. Es sei bemerkt, daß eine konvexe Funktion $f : D \subseteq \mathbb{R}^n \to \mathbb{R}$, D konvex, schöne Eigenschaften besitzt:

- f ist **lokal Lipschitz-stetig im Inneren ihres Definitionsbereichs** D (und damit auch **stetig**), d.h. zu jedem $x \in \operatorname{int}(D)$ gibt es $r > 0$ und $L = L(x) > 0$ mit

  $$|f(y_1) - f(y_2)| \le L\|y_1 - y_2\| \qquad \forall y_1, y_2 \in U_r(x),$$

  vgl. Geiger und Kanzow [GK02], S. 327.

- Sei D offen. Dann ist f **richtungsdifferenzierbar** für alle $x \in D$ und alle Richtungen $d \in \mathbb{R}^n$, d.h. die Richtungsableitung

  $$f'(x;d) := \lim_{h \downarrow 0} \frac{f(x+hd) - f(x)}{h}$$

  existiert. Genauer gilt sogar

  $$f'(x;d) = \inf_{h>0} \frac{f(x+hd) - f(x)}{h},$$

  vgl. Geiger und Kanzow [GK02], S. 329.

4.1 Trennungssätze

Zur Herleitung notwendiger Optimalitätsbedingungen benötigen wir Resultate über die Trennbarkeit von konvexen Mengen durch Hyperebenen.

Definition 4.1.1 **(Hyperebene)**
Sei $a \in \mathbb{R}^n$, $a \neq 0$ und $\gamma \in \mathbb{R}$. Die Menge

$$H = \{x \in \mathbb{R}^n \mid a^\top x = \gamma\}$$

heißt **Hyperebene**. *Die Menge*

$$H_+ = \{x \in \mathbb{R}^n \mid a^\top x \ge \gamma\}$$

heißt **positiver Halbraum von** H *und*

$$H_- = \{x \in \mathbb{R}^n \mid a^\top x \le \gamma\}$$

heißt **negativer Halbraum von** H.

Anschaulich trennt eine Hyperebene den $\mathbb{R}^n$ in zwei Hälften, nämlich in den positiven und in den negativen Halbraum.

Definition 4.1.2 **(Trennung von Mengen)**
Es seien $A, B \subseteq \mathbb{R}^n$ nichtleere Mengen. Die Hyperebene $H = \{x \in \mathbb{R}^n \mid a^\top x = \gamma\}$ **trennt** *A und B, wenn*

$$\begin{aligned} a^\top x &\leq \gamma, \qquad \forall x \in A, \\ a^\top x &\geq \gamma, \qquad \forall x \in B. \end{aligned}$$

Die Hyperebene trennt A und B **strikt***, wenn*

$$\begin{aligned} a^\top x &< \gamma, \qquad \forall x \in A, \\ a^\top x &> \gamma, \qquad \forall x \in B. \end{aligned}$$

Die Hyperebene trennt A und B **eigentlich***, wenn zusätzlich nicht beide Mengen ganz in H enthalten sind.*

Abbildung 4.1 veranschaulicht die Begriffe.

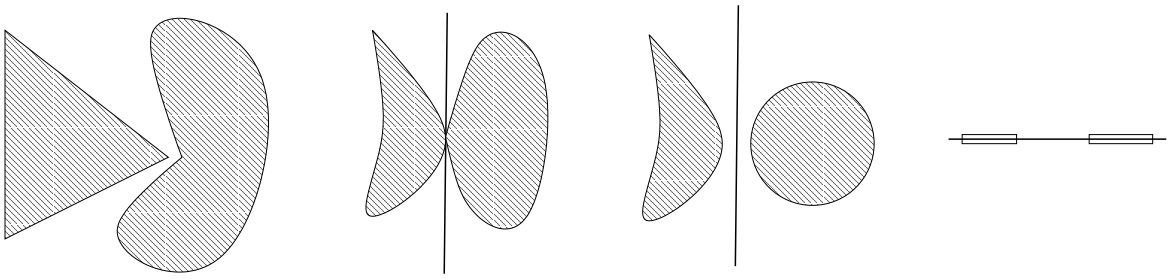


Abbildung 4.1: Trennung von Mengen durch Hyperebenen: Von links nach rechts sind folgende Fälle dargestellt: keine Trennung möglich, keine strikte Trennung möglich, strikte Trennung, keine eigentliche Trennung.

Natürlich stellt sich die Frage, welche Mengen durch Hyperebenen getrennt werden können. Die folgenden Sätze finden sich u.a. in Rockafellar [Roc70], Jarre und Stoer [JS04], Geiger und Kanzow [GK02], Mangasarian [Man94] und Bazaraa et al. [BSS93].
Der folgende Satz behandelt die Trennbarkeit eines Punktes und einer konvexen Menge.

Satz 4.1.3
Sei $A \subseteq \mathbb{R}^n$ konvex und $0 \notin A$. Dann existiert eine Hyperebene, die $\{0\}$ von A trennt. Ist A zusätzlich abgeschlossen, so kann $\{0\}$ strikt von A getrennt werden.

Mit Hilfe dieses Satzes läßt sich die Trennbarkeit zweier disjunkter konvexer Mengen A, B beweisen:

Satz 4.1.4
Seien $A, B \subseteq \mathbb{R}^n$ nichtleere konvexe Mengen mit $A \cap B = \emptyset$. Dann gibt es eine Hyperebene,

die A und B trennt.

Eine Verschärfung des Satzes lautet:

Satz 4.1.5

Seien $A, B \subseteq \mathbb{R}^n$ nichtleere konvexe Mengen mit $A \cap B = \emptyset$. A sei abgeschlossen und B sei kompakt. Dann gibt es eine Hyperebene, die A und B strikt trennt.

Abschließend beschäftigen wir uns mit der Charakterisierung eigentlich trennbarer Mengen und benötigen hierfür einige Begriffe.

Definition 4.1.6 **(affine Menge, affine Hülle, relativ Inneres)**

Sei $A \subseteq \mathbb{R}^n$ eine Menge.

(a) A heißt **affine Menge***, wenn*

$$(1-\lambda)x + \lambda y \in A \qquad \forall x, y \in A,\ \lambda \in \mathbb{R}.$$

(b) Die Menge

$$\begin{aligned} \mathit{aff}(A) &:= \bigcap \{M \mid M \text{ ist affine Menge und } A \subseteq M\} \\ &= \left\{ \sum_{i=1}^m \lambda_i x_i \mid m \in \mathbb{N},\ \sum_{i=1}^m \lambda_i = 1,\ x_i \in A,\ i = 1, \ldots, m \right\} \end{aligned}$$

heißt **affine Hülle von** *A.*

(c) Die Menge

$$\mathit{relint}(A) := \{x \in \mathit{aff}(A) \mid \exists \varepsilon > 0 : U_\varepsilon(x) \cap \mathit{aff}(A) \subseteq A\}$$

heißt **relativ Inneres von** *A. Die Menge $\mathit{cl}(A) \setminus \mathit{relint}(A)$ heißt* **relativer Rand von** *A.*

Beachte, daß jede affine Menge A in $\mathbb{R}^n$ dargestellt werden kann durch $A = \{x \in \mathbb{R}^n \mid Bx = b\}$ mit einer Matrix $B \in \mathbb{R}^{m \times n}$ und einem Vektor $b \in \mathbb{R}^m$. Das relativ Innere einer konvexen Menge A kann auch wie folgt charakterisiert werden:

$$x \in \operatorname{relint}(A) \qquad \Leftrightarrow \qquad \forall y \in \operatorname{aff}(A) \exists \varepsilon > 0 \ :\ x \pm \varepsilon(y - x) \in A.$$

Es gilt

Satz 4.1.7

Sei $A \subseteq \mathbb{R}^n$ nichtleer und konvex. Dann ist $\mathit{relint}(A) \neq \emptyset$.

Abschließend zitieren wir das Hauptresultat dieses Abschnitts:

Satz 4.1.8

Seien $A, B \subseteq \mathbb{R}^n$ nichtleer und konvex. Genau dann existiert eine Hyperebene, die A und B eigentlich trennt, wenn

$$\mathit{relint}(A) \cap \mathit{relint}(B) = \emptyset.$$

4.2 Optimalitätsbedingungen

Wir wollen notwendige Bedingungen für das konvexe Optimierungsproblem herleiten und setzen voraus, daß $\hat{x}$ optimal ist. Betrachte die Mengen

$$\begin{aligned}
A &= \{(r,y,z) \in \mathbb{R}^{1+m+p} \mid r < f(\hat{x}),\ y_i \le 0,\ i \in \mathcal{I},\ z_j = 0,\ j \in \mathcal{J}\},\\
B &= \{(r,y,z) \in \mathbb{R}^{1+m+p} \mid \exists x \in S\ :\ r \ge f(x),\ y_i \ge g_i(x),\ i \in \mathcal{I},\ z_j = h_j(x),\ j \in \mathcal{J}\},\\
D &= \{(f(x),g(x),h(x)) \in \mathbb{R}^{1+m+p} \mid x \in S\},
\end{aligned}$$

vgl. Abbildung 4.2.

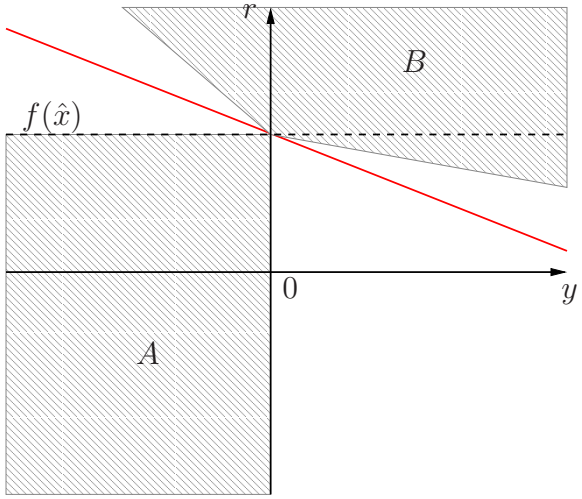


Abbildung 4.2: Trennbarkeit der Mengen A und B am Beispiel einer Ungleichungsnebenbedingung $g(x) \le 0$ mit $g(\hat{x}) = 0$ und ohne Gleichungsrestriktionen.

Offensichtlich ist A konvex und $D \subseteq B$. B ist auch konvex, denn für $(r,y,z), (\tilde{r},\tilde{y},\tilde{z}) \in B$ gilt für $x, \tilde{x} \in S$, $0 \le \lambda \le 1$:

$$\begin{aligned}
f(\lambda x + (1-\lambda)\tilde{x}) &\le \lambda f(x) + (1-\lambda) f(\tilde{x}) \le \lambda r + (1-\lambda)\tilde{r},\\
g(\lambda x + (1-\lambda)\tilde{x}) &\le \lambda g(x) + (1-\lambda) g(\tilde{x}) \le \lambda y + (1-\lambda)\tilde{y},\\
h(\lambda x + (1-\lambda)\tilde{x}) &= \lambda h(x) + (1-\lambda) h(\tilde{x}) = \lambda z + (1-\lambda)\tilde{z},\\
\lambda x + (1-\lambda)\tilde{x} &\in S.
\end{aligned}$$

Annahme: Es gibt $(r, y, z) \in A \cap B$. Dann gibt es $x \in S$ mit

$$\begin{aligned} g_i(x) &\leq 0, \qquad i \in \mathcal{I}, \\ h_j(x) &= 0, \qquad j \in \mathcal{J}, \\ f(x) &\leq r < f(\hat{x}) \end{aligned}$$

im Widerspruch zur Minimalität von $\hat{x}$ (beachte, daß x zulässig ist!).
Also gilt

$$A \cap B = \emptyset.$$

Gemäß Trennungssatz 4.1.4 können die Mengen A und B durch eine Hyperebene getrennt werden (gemäß Satz 4.1.8 ist diese Trennung sogar eigentlich). Wegen $D \subseteq B$ können auch A und D getrennt werden. Es exisieren also nichtverschwindende Multiplikatoren $(l_0, \lambda, \mu) \in \mathbb{R}^{1+m+p}$, $(l_0, \lambda, \mu) \neq 0$ und $\gamma \in \mathbb{R}$ mit

$$l_0 r + \lambda^\top y + \mu^\top z \leq \gamma \leq l_0 f(x) + \lambda^\top g(x) + \mu^\top h(x) \tag{4.1}$$

für alle $(r, y, z) \in A$, $x \in S$. Aus Stetigkeitsgründen bleiben diese Ungleichungen für $(r, y, z) \in \operatorname{cl}(A)$ erhalten. Wegen $(f(\hat{x}), g(\hat{x}), h(\hat{x})) \in \operatorname{cl}(A) \cap D$ folgt

$$\gamma = l_0 f(\hat{x}) + \lambda^\top g(\hat{x}) + \mu^\top h(\hat{x})$$

und insbesondere

$$l_0 f(\hat{x}) + \lambda^\top g(\hat{x}) + \mu^\top h(\hat{x}) \leq l_0 f(x) + \lambda^\top g(x) + \mu^\top h(x) \tag{4.2}$$

für alle $x \in S$.

(i) **Behauptung:** Es gilt $l_0 \geq 0$.

Beweis: Wähle in (4.1) $\hat{x} \in S$ und $r < f(\hat{x})$. Dann ist $(r, g(\hat{x}), h(\hat{x})) \in A$, denn $g(\hat{x}) \leq 0$, $h(\hat{x}) = 0$. (4.1) reduziert sich auf

$$l_0 r \leq l_0 f(\hat{x}) \qquad \forall r < f(\hat{x}).$$

Wäre $l_0 < 0$, so ist diese Bedingung wegen $l_0 r \to \infty$ für $r \to -\infty$ nicht erfüllt.

(ii) **Behauptung:** Es gilt $\lambda_i \geq 0$ für alle $i \in \mathcal{I}$.

Beweis: Wähle in (4.1) $\hat{x} \in S$, $r = f(\hat{x})$, $z = h(\hat{x}) = 0$, $i_0 \in \mathcal{I}$ beliebig und

$$y_i = \begin{cases} g_{i_0}(\hat{x}) - 1, & \text{für } i = i_0, \\ g_i(\hat{x}), & \text{für } i \neq i_0, \end{cases} \qquad i \in \mathcal{I}.$$

Dann ist $(r, y, z) \in \operatorname{cl}(A)$ und (4.1) reduziert sich auf $-\lambda_{i_0} \leq 0$ bzw. $\lambda_{i_0} \geq 0$. Da i_0 beliebig war, folgt die Behauptung.

(iii) **Behauptung:** Es gilt $\lambda_i g_i(\hat{x}) = 0$ für alle $i \in \mathcal{I}$.

Beweis: Wähle in (4.1) $\hat{x} \in S$, $r = f(\hat{x})$, $y = 0$, $z = h(\hat{x}) = 0$. Dann ist $(r, y, z) \in \mathrm{cl}(A)$ und (4.1) reduziert sich auf

$$0 \leq \lambda^\top g(\hat{x}) = \sum_{i \in \mathcal{I}} \underbrace{\lambda_i}_{\geq 0} \underbrace{g_i(\hat{x})}_{\leq 0}.$$

Also muß $\lambda^\top g(\hat{x}) = 0$ gelten, was gleichbedeutend ist mit $\lambda_i g_i(\hat{x}) = 0$ für alle i.

Unter Verwendung der **Lagrangefunktion**

$$L(x, l_0, \lambda, \mu) := l_0 f(x) + \lambda^\top g(x) + \mu^\top h(x) \tag{4.3}$$

in (4.2) haben wir die folgenden notwendigen Bedingungen bewiesen:

Satz 4.2.1 (Notwendige Bedingungen nach Fritz John)
Sei $\hat{x}$ optimal für das konvexe Optimierungsproblem. Dann existieren Multiplikatoren $(l_0, \lambda, \mu) \in \mathbb{R}^{1+m+p}$, $(l_0, \lambda, \mu) \neq 0$, so daß die folgenden Bedingungen erfüllt sind:

(a) **Vorzeichenbedingungen:**

$$l_0 \geq 0, \qquad \lambda_i \geq 0, \qquad i = 1, \ldots, m. \tag{4.4}$$

(b) **Minimalität der Lagrangefunktion:**

$$L(\hat{x}, l_0, \lambda, \mu) \leq L(x, l_0, \lambda, \mu) \qquad \forall x \in S. \tag{4.5}$$

(c) **Komplementaritätsbedingungen:**

$$\lambda_i g_i(\hat{x}) = 0, \qquad i = 1, \ldots, m. \tag{4.6}$$

(d) **Zulässigkeit:**

$$\hat{x} \in \Sigma. \tag{4.7}$$

Bemerkung 4.2.2

- *Jeder Vektor $(x, l_0, \lambda, \mu) \in \mathbb{R}^{n+1+m+p}$ mit $(l_0, \lambda, \mu) \neq 0$, der die Bedingungen (4.4)-(4.7) erfüllt, heißt* **Fritz-John-Punkt**. *Die Multiplikatoren l_0, λ, μ heißen auch* **Lagrange-Multiplikatoren**. *Die Bedingungen (4.4)-(4.7) heißen* **Fritz-John-Bedingungen**.
- *Die wesentliche Aussage des Satzes ist, daß es einen von Null verschiedenen Vektor (l_0, λ, μ) gibt. Beachte, daß $(l_0, \lambda, \mu) = 0$ trivialerweise die Fritz-John-Bedingungen erfüllt.*

- *Die Komplementaritätsbedingungen sind äquivalent mit*

$$g_i(\hat{x}) < 0 \qquad \Rightarrow \qquad \lambda_i = 0$$

für $i \in \mathcal{I}$.

Im Fall $l_0 = 0$ tritt die Zielfunktion f in den Fritz-John-Bedingungen nicht auf. Wir werden später noch sehen, daß dies in gewisser Weise eine entartete Situation darstellt. Gilt $l_0 > 0$, so können wir o.B.d.A. $l_0 = 1$ wählen, da die Multiplikatoren linear auftreten. In diesem Fall spricht man von den **Karush-Kuhn-Tucker-Bedingungen (KKT-Bedingungen)**. Diese sind sogar hinreichend für Optimalität:

Satz 4.2.3 (Hinreichende Bedingung)
Sei $\hat{x}$ zulässig für das konvexe Optimierungsproblem, d.h. $\hat{x} \in \Sigma$. Sind die Fritz-John-Bedingungen (4.4)-(4.7) mit $l_0 = 1$ erfüllt, so ist $\hat{x}$ optimal.

Beweis: Gemäß (4.5) und (4.6) gilt

$$f(\hat{x}) \le f(x) + \lambda^\top g(x) + \mu^\top h(x) \qquad \forall x \in S.$$

Für alle zulässigen $x \in \Sigma$ gilt mit (4.4)

$$\underbrace{\lambda^\top}_{\ge 0} \underbrace{g(x)}_{\le 0} \le 0, \qquad h(x) = 0$$

und somit

$$f(\hat{x}) \le f(x) \qquad \forall x \in \Sigma.$$

□

Der vorangehende Satz zeigt, daß der Fall $l_0 = 1$ von besonderer Bedeutung ist. Es stellt sich nun die Frage, ob es Bedingungen gibt, unter denen stets $l_0 = 1$ gewählt werden kann. Bei der nun folgenden Herleitung einer solchen Bedingung – der **Slater-Bedingung** – beschränken wir uns zur Vereinfachung auf den Fall $p = 0$, d.h. es mögen **keine Gleichungsrestriktionen** auftreten. Wir nehmen nun an, daß die Fritz-John-Bedingungen für diesen Spezialfall mit $l_0 = 0$ erfüllt seien. Es gibt also einen Multiplikator $0 \neq \lambda \in \mathbb{R}^m$ mit $\lambda \ge 0$ und

$$0 = \lambda^\top g(\hat{x}) \le \lambda^\top g(x) \qquad \forall x \in S.$$

Diese Bedingung hatten wir erhalten durch Trennung der Mengen A und D. Sie ist **nicht** erfüllt, wenn es ein $\tilde{x} \in S$ gibt mit $g(\tilde{x}) < 0$, da dann wegen $\lambda \ge 0$ und $\lambda \neq 0$ sofort $\lambda^\top g(\tilde{x}) < 0$ folgt. Geometrisch bedeutet dies, daß die Projektionen

$$\mathrm{proj}_{\mathbb{R}^m}(A) = \{y \in \mathbb{R}^m \mid y \le 0\}, \qquad \mathrm{proj}_{\mathbb{R}^m}(D) = \{g(x) \mid x \in S\}$$

nicht trennbar sind. Wegen $g(\tilde{x}) < 0$ ist $y = g(\tilde{x}) < 0$ nämlich ein innerer Punkt von $\text{proj}_{\mathbb{R}^m}(A)$. Mit anderen Worten: Gibt es ein $\tilde{x} \in S$ mit $g(\tilde{x}) < 0$, so können die Fritz-John-Bedingungen nicht mit $l_0 = 0$ gelten. Andererseits ist die Gültigkeit der Fritz-John-Bedingungen bewiesen. Folglich muß $l_0 > 0$ und somit o.B.d.A. $l_0 = 1$ gelten. Dies beweist

Satz 4.2.4 **(Karush-Kuhn-Tucker-Bedingungen (KKT-Bedingungen))**
Im konvexen Optimierungsproblem sei $p = 0$ (es treten keine Gleichungsrestriktionen auf!) und $\hat{x}$ sei optimal. Desweiteren sei die **Slater-Bedingung** *erfüllt:*

$$\exists \tilde{x} \in S \ : \ g(\tilde{x}) < 0.$$

Dann gelten die **KKT-Bedingungen:** *Es gibt Multiplikatoren $\lambda \in \mathbb{R}^m$, so daß die folgenden Bedingungen erfüllt sind:*

(a) **Vorzeichenbedingungen:**

$$\lambda_i \geq 0, \qquad i = 1, \ldots, m. \tag{4.8}$$

(b) **Minimalität der Lagrangefunktion mit $l_0 = 1$:**

$$f(\hat{x}) + \lambda^\top g(\hat{x}) \leq f(x) + \lambda^\top g(x) \qquad \forall x \in S. \tag{4.9}$$

(c) **Komplementaritätsbedingungen:**

$$\lambda_i g_i(\hat{x}) = 0, \qquad i = 1, \ldots, m. \tag{4.10}$$

(d) **Zulässigkeit:**

$$\hat{x} \in \Sigma. \tag{4.11}$$

Beachte, daß die KKT-Bedingungen genau die Fritz-John-Bedingungen mit $l_0 = 1$ sind. Die KKT-Bedingungen stellen den nicht-entarteten Fall dar, da hier die Zielfunktion f explizit in den Bedingungen auftaucht.

Beispiel 4.2.5
Wir betrachten das konvexe Optimierungsproblem

$$\begin{array}{ll} \textit{Minimiere} & (x_1 - 2)^2 + (x_2 - 3)^2 \\ \textit{unter} & (x_1, x_2)^\top \in S := \{(x_1, x_2)^\top \in \mathbb{R}^2 \mid x_2 + \frac{1}{2}x_1 - \frac{1}{2} = 0\} \\ & g_1(x_1, x_2) := x_2 + 2{x_1}^2 - 2 \leq 0, \\ & g_2(x_1, x_2) := {x_1}^2 - x_2 - 1 \leq 0. \end{array}$$

Wir versuchen, einen KKT-Punkt zu finden. Einen solchen gibt es, da die Slater-Bedingung z.B. für $x_1 = 0, x_2 = 1/2$ erfüllt ist. Die Lagrangefunktion für $l_0 = 1$ lautet

$$L(x_1, x_2, \lambda_1, \lambda_2) = (x_1 - 2)^2 + (x_2 - 3)^2 + \lambda_1(x_2 + 2{x_1}^2 - 2) + \lambda_2({x_1}^2 - x_2 - 1).$$

Die KKT-Bedingungen lauten wie folgt: Sei $\hat{x} = (\hat{x}_1, \hat{x}_2)^\top$ *optimal für das Optimierungsproblem. Dann gibt es Multiplikatoren* $\lambda = (\lambda_1, \lambda_2)^\top$ *mit*

$$\begin{aligned} \lambda_1, \lambda_2 &\geq 0, \\ \lambda_i g_i(\hat{x}_1, \hat{x}_2) &= 0, \quad i = 1, 2, \\ L(\hat{x}, \lambda_1, \lambda_2) &\leq L(x, \lambda_1, \lambda_2), \qquad \forall (x_1, x_2)^\top \in S. \end{aligned}$$

Jeder Punkt $(x_1, x_2)^\top \in S$ *erfüllt*

$$x_1 = 1 - 2x_2.$$

Einsetzen in die Lagrangefunktion liefert

$$\tilde{L}(x_2, \lambda_1, \lambda_2) = 5x_2^2 - 2x_2 + 10 + \lambda_1 x_2(8x_2 - 7) + \lambda_2 x_2(4x_2 - 5).$$

Das Optimum minimiert die Lagrangefunktion auf S*, was äquivalent ist mit der Minimierung von* $\tilde{L}(x_2, \lambda_1, \lambda_2)$ *auf* $x_2 \in \mathbb{R}$*. Notwendig gilt für ein Minimum von* $\tilde{L}$

$$0 = L'_{x_2}(\hat{x}_2, \lambda_1, \lambda_2) = 10\hat{x}_2 - 2 + \lambda_1(16\hat{x}_2 - 7) + \lambda_2(8\hat{x}_2 - 5). \tag{4.12}$$

Annahme: $g_1(\hat{x}_1, \hat{x}_2) < 0$ *und* $g_2(\hat{x}_1, \hat{x}_2) < 0$*, d.h. im Optimalpunkt sind beide Ungleichungsrestriktionen nicht aktiv. Aus den Komplementaritätsbedingungen folgt dann* $\lambda_1 = \lambda_2 = 0$ *und damit*

$$0 = 10\hat{x}_2 - 2 \qquad \Rightarrow \qquad \hat{x}_2 = \frac{1}{5} \quad \Rightarrow \quad \hat{x}_1 = 1 - 2\hat{x}_2 = \frac{3}{5}.$$

Wir haben also unter den Annahmen $g_1(\hat{x}_1, \hat{x}_2) < 0$ *und* $g_2(\hat{x}_1, \hat{x}_2) < 0$ *den Kandidaten* $(\hat{x}_1, \hat{x}_2)^\top = (3/5, 1/5)^\top$ *erhalten. Tatsächlich gilt* $g_1(\hat{x}_1, \hat{x}_2) = -\frac{27}{25} < 0$ *und* $g_2(\hat{x}_1, \hat{x}_2) = -\frac{21}{25} < 0$ *und unsere Annahme war (ganz zufällig) richtig.*
Damit ist

$$\hat{x} = \left(\frac{3}{5}, \frac{1}{5}\right)^\top, \hat{\lambda} = (0, 0)^\top, f(\hat{x}) = \frac{49}{5}$$

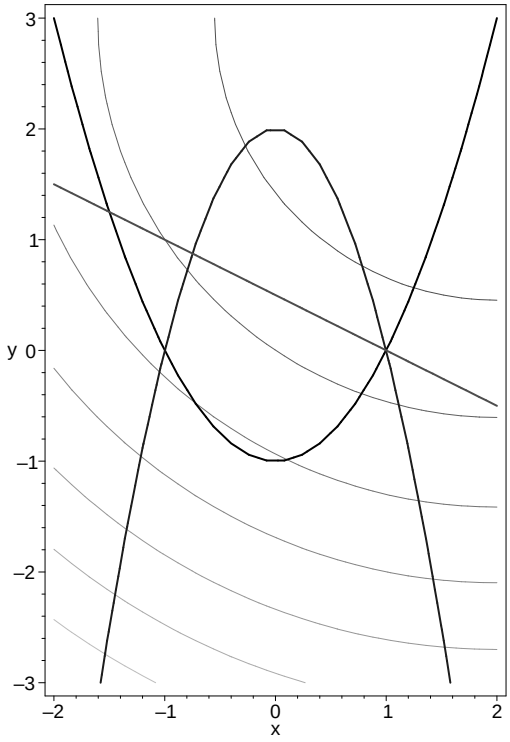
KKT-Punkt und wegen Satz 4.2.3 auch optimal.

Bemerkung: *Natürlich haben wir uns die Sache etwas einfach gemacht. Im allgemeinen ist es notwendig, sämtliche Kombinationen von aktiven und nichtaktiven Beschränkungen* g_1 *und* g_2 *zu betrachten.*
So führt die Annahme, daß g_1 *und* g_2 *aktiv sind, auf den einzelnen Punkt* $\hat{x} = (1, 0)^\top \in S$*, der die beiden Beschränkungen erfüllt. Diejenigen Multiplikatoren, die (4.12) erfüllen lauten*

$$\lambda_1 = -\frac{2}{7} - \frac{5}{7}\lambda_2.$$

Wegen $\lambda_2 \geq 0$ *ist* $\lambda_1 < 0$ *und die Vorzeichenbedingungen sind nie erfüllt. Damit gibt es keinen KKT-Punkt, falls* g_1 *und* g_2 *gleichzeitig aktiv sind.*

Die beiden verbleibenden Kombinationen g_1 aktiv und g_2 nicht aktiv (hier ist $\hat{x}_1 = -3/4, \hat{x}_2 = 7/8$ zulässig, aber $\lambda_1 = -27/28 - 2/7\lambda_2 < 0$ für $\lambda_2 \geq 0$) bzw. g_2 aktiv und g_1 nicht aktiv (hier gibt es keinen zulässigen Punkt) liefern ebenfalls keine KKT-Punkte.



Bemerkung 4.2.6

- *Die hinreichende Bedingung in Satz 4.2.3 gilt genauso auch für nichtkonvexe Probleme!*

- *Die Minimalitätsbedingungen (4.5) und (4.9) besagen gerade, daß $\hat{x}$ die Lagrangefunktion über S minimiert. Sind die Funktionen f und g_i, $i \in \mathcal{I}$ differenzierbar und gilt $S = \mathbb{R}^n$, so gilt notwendig*

$$\nabla_x L(\hat{x}, l_0, \lambda, \mu) = 0.$$

- *Es gibt eine abgeschwächte Form der Slater-Bedingung für konvexe Optimierungsprobleme mit Gleichungs- und Ungleichungsrestriktionen, die die Gültigkeit der Fritz-John-Bedingungen mit $l_0 = 1$ (KKT-Bedingungen) impliziert. Dabei werden noch lineare und nichtlineare Ungleichungsrestriktionen unterschieden gemäß $\mathcal{I} = \mathcal{I}_1 \cup \mathcal{I}_2$, wobei $\mathcal{I}_1$ die nichtlinearen und $\mathcal{I}_2$ die linearen Ungleichungsrestriktionen indizieren.*

Die **abgeschwächte Slater-Bedingung** *lautet:*

$$\begin{aligned} \exists \tilde{x} \in \mathit{relint}(S) \quad : \quad & g_i(\tilde{x}) < 0, \quad i \in \mathcal{I}_1, \\ & g_i(\tilde{x}) \le 0, \quad i \in \mathcal{I}_2, \\ & h_j(\tilde{x}) = 0, \quad j \in \mathcal{J}, \end{aligned}$$

vgl. Blum und Oettli [BO75], S. 76. Ist die abgeschwächte Slater-Bedingung erfüllt, so gelten die Fritz-John-Bedingungen in Satz 4.2.1 mit $l_0 = 1$ und man spricht dann analog zu Satz 4.2.4 wieder von den KKT-Bedingungen.

4.3 Schnittebenenverfahren

Wir betrachten wieder das konvexe Optimierungsproblem ohne Gleichungsrestriktionen, d.h. $p = 0$.

Das Schnittebenenverfahren nach Kelly basiert auf einer sukzessiven äußeren Approximation der zulässigen Menge unter Verwendung von Linearisierungen der Ungleichungen. Nichtzulässige Punkte werden durch Hinzunahme weiterer Beschränkungen abgeschnitten, vgl. Abbildung 4.3.

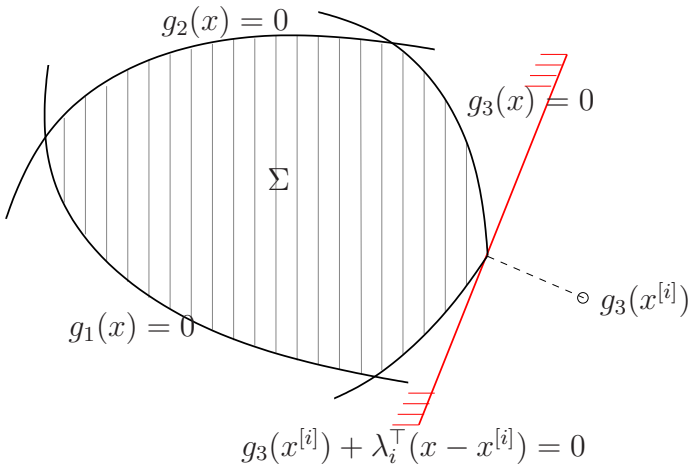


Abbildung 4.3: Motivation des Schnittebenenverfahrens nach Kelly für konvexe Ungleichungsbeschränkungen.

Algorithmus: Schnittebenenverfahren nach Kelly

(i) Gegeben sei eine konvexe und kompakte Menge $S_0 \neq \emptyset$ mit $\Sigma \subseteq S_0 \subseteq S$.
Setze $i = 0$.

(ii) Berechne $x^{[i]} \in S_i$ mit $f(x^{[i]}) = \min\{f(x) \mid x \in S_i\}$.

(iii) Falls $x^{[i]} \in \Sigma$ gilt, STOP ($x^{[i]}$ ist optimal).
Falls weitere Abbruchbedingungen erfüllt sind, STOP.

(iv) Bestimme einen Index $k_i \in \{1, \ldots, m\}$ mit

$$g_{k_i}(x^{[i]}) = \max_{j=1,\ldots,m} g_j(x^{[i]})$$

und setze

$$S_{i+1} = S_i \cap \{x \in \mathbb{R}^n \mid g_{k_i}(x^{[i]}) + \lambda_i^\top (x - x^{[i]}) \leq 0\},.$$

wobei λ_i ein Subgradient von g_{i_k} in $x^{[i]}$.

(v) Falls $S_{i+1} = \emptyset$, STOP mit Fehlermeldung.
Andernfalls setze $i = i + 1$ und gehe zu (ii).

Erläuterung: $\lambda \in \mathbb{R}^n$ heißt **Subgradient** der konvexen Funktion $g : \mathbb{R}^n \to \mathbb{R}$ in $x \in \mathbb{R}^n$, wenn

$$g(y) \geq g(x) + \lambda^\top (y - x) \qquad \forall y \in \mathbb{R}^n.$$

Die Menge $\partial g(x) = \{\lambda \in \mathbb{R}^n \mid \lambda \text{ ist Subgradient von } g \text{ in } x\}$ aller Subgradienten heißt **Subdifferential** von g in x. Das Subdifferential verallgemeinert den Begriff des Gradienten auf nichtdifferenzierbare (konvexe) Funktionen. Für differenzierbare Funktionen besteht das Subdifferential nur aus dem Gradienten von g.

Satz 4.3.1

Falls $S_i = \emptyset$ für einen Index $i \in \mathbb{N}$, dann ist $\Sigma = \emptyset$. Falls $x^{[i]} \in \Sigma$ für einen Index $i \in \mathbb{N} \cup \{0\}$, dann ist $x^{[i]}$ optimal. In allen anderen Fällen generiert der Algorithmus eine unendliche Folge $\{x^{[i]}\}$ mit mindestens einem Häufungspunkt. Jeder Häufungspunkt $\hat{x}$ dieser Folge ist eine optimale Lösung und die Folge $\{f(x^{[i]})\}$ konvergiert monoton steigend gegen $f(\hat{x})$.

Beweis: Wir zeigen induktiv, daß $\Sigma \subseteq S_{i+1} \subsetneq S_i$ und $f(x^{[i]}) \leq f(x^{[i+1]}) \leq w$ für alle i, wobei w den optimalen Funktionswert bezeichnet.

Für S_0 wurde $\Sigma \subseteq S_0$ vorausgesetzt. Sei nun $\Sigma \subseteq S_i$ für ein $i \geq 0$.

Da g_{k_i} konvex und λ_i Subgradient ist, folgt

$$g_{k_i}(x^{[i]}) + \lambda_i^\top (x - x^{[i]}) \le g_{k_i}(x) \le 0 \qquad \forall x \in \Sigma, \tag{4.13}$$

also $\Sigma \subseteq S_{i+1}$.
Die zusätzliche Beschränkung (4.13) schneidet tatsächlich etwas von S_i ab, solange $x^{[i]}$ noch nicht optimal ist, denn für $x^{[i]}$ gilt

$$g_{k_i}(x^{[i]}) + \lambda_i^\top (x - x^{[i]})\big|_{x=x^{[i]}} = g_{k_i}(x^{[i]}).$$

Da $x^{[i]} \notin \Sigma$ und $x^{[i]} \in S_i \subseteq S_0 \subseteq S$ muß es einen Index j mit $g_j(x^{[i]}) > 0$ geben. Mit der Definition von k_i, d.h.

$$g_{k_i}(x^{[i]}) = \max_{j=1,\dots,m} g_j(x^{[i]}),$$

folgt $g_{k_i}(x^{[i]}) > 0$. Also ist (4.13) nicht erfüllt für $x = x^{[i]}$ und $x^{[i]} \notin S_{i+1}$. Wir haben also $S_{i+1} \subsetneq S_i$ gezeigt.
Die Monotonie der Werte $f(x^{[i]})$ folgt direkt aus der Inklusion $\Sigma \subseteq S_{i+1} \subsetneq S_i$ und wir erhalten

$$f(x^{[0]}) \le f(x^{[1]}) \le \dots \le w.$$

Angenommen, der Algorithmus endet nicht. Dann generiert er eine unendliche Folge $\{x^{[i]}\} \subseteq S_0$. Da S_0 kompakt ist, gibt es mindestens einen Häufungspunkt $\hat{x}$ der Folge mit $\hat{x} \in S_0 \subseteq S$ und $f(\hat{x}) \le w$. Die letzte Beziehung folgt aus der Stetigkeit von f auf S (f ist konvex und daher lokal Lipschitz-stetig im Inneren des Definitionsbereichs). Für alle $i \in \mathbb{N} \cup \{0\}$ und $j = 1, \dots, m$ gilt

$$g_j(x^{[i]}) + \lambda_i^\top (x^{[k]} - x^{[i]}) \le g_{k_i}(x^{[i]}) + \lambda_i^\top (x^{[k]} - x^{[i]}) \le 0 \quad \forall k > i.$$

Daraus folgt

$$g_j(x^{[i]}) + \lambda_i^\top (\hat{x} - x^{[i]}) \le 0,$$

und weiter

$$g_j(x^{[i]}) \le c \cdot \|\hat{x} - x^{[i]}\|_2$$

für eine Konstante c, da $\{\lambda_0, \lambda_1, \dots\}$ beschränkt ist (das Subdifferential einer konvexen Funktion ist kompakt). Zusammenfassend erhalten wir $g_j(\hat{x}) \le 0$ für $j = 1, \dots, m$. Daher ist $\hat{x}$ zulässig und optimal wegen $f(\hat{x}) \le w$. □

Für eine praktische Implementierung des Schnittebenenverfahrens sei

$$S := \{x = (x_1, \dots, x_n)^\top \in \mathbb{R}^n \mid l_i \le x_i \le u_i,\ i = 1, \dots, n\},$$

mit gegebenen Schranken $l_i \le u_i$ für $i = 1, \dots, n$. Zusätzlich sei die Zielfunktion f linear, d.h. $f(x) = c^\top x$ mit $c \in \mathbb{R}^n$. Die letzte Annahme ist keine wesentliche Einschränkung, da das konvexe Optimierungsproblem äquivalent ist zu

$$(P') \qquad \begin{array}{ll} \text{Minimiere} & \alpha \\ \text{unter} & (x,\alpha)^\top \in \hat{\Sigma}, \end{array}$$

mit

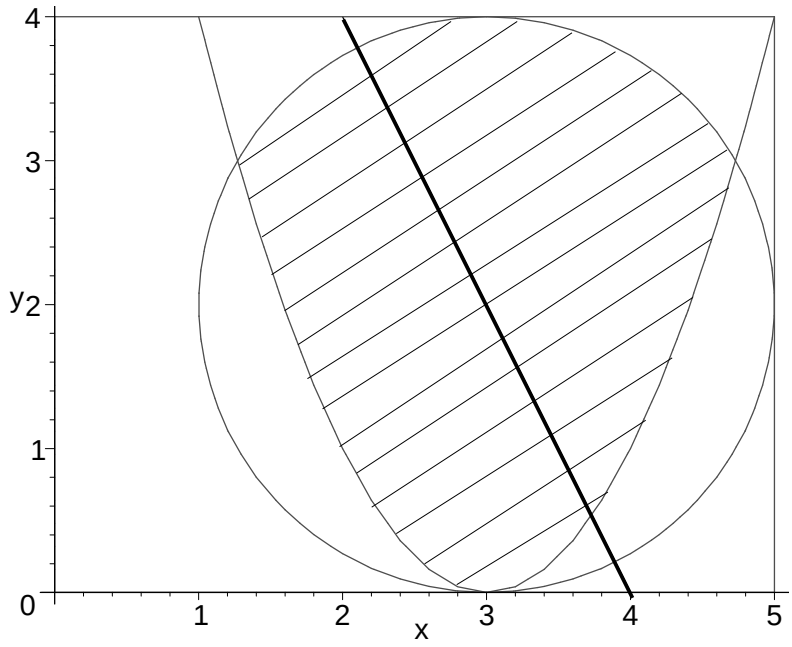
$$\hat{\Sigma} := \{(x,\alpha) \in S \times \mathbb{R} \mid g_i(x) \le 0,\ i = 1,\dots,m,\ f(x) \le \alpha\}.$$

Beachte, daß die Zielfunktion in (P') linear ist.
Außerdem nehmen wir an, daß ein Algorithmus zur Berechnung eines Subgradienten von g_i zur Verfügung steht. Ist g_i differenzierbar, so ist der Subgradient gerade der gewöhnliche Gradient.
Unter der Voraussetzung, daß die Startmenge S_0 durch endlich viele lineare Gleichungen und Ungleichungen beschrieben wird, sind sämtliche Teilprobleme des Schnittebenenverfahrens linear und wir können z.B. ein Simplexverfahren zur Lösung verwenden.

Beispiel 4.3.2

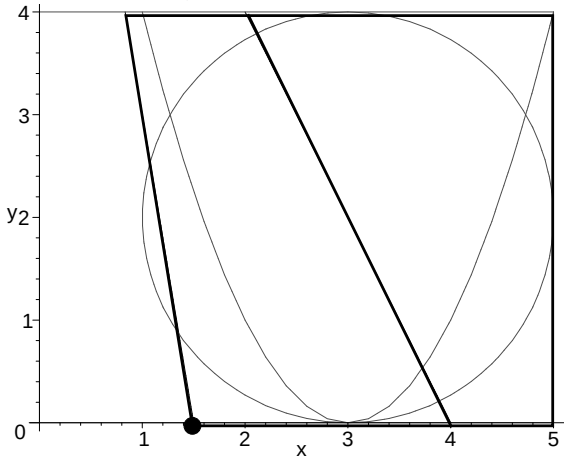
$$\begin{array}{ll} \textit{Minimiere} & 2x_1 + x_2 \\ \textit{unter} & 0 \le x_1 \le 5,\ 0 \le x_2 \le 4, \\ & (x_1-3)^2 + (x_2-2)^2 - 4 \le 0, \\ & (x_1-3)^2 - x_2 \le 0. \end{array}$$

Der zulässige Bereich und die Höhenlinien der Zielfunktion sind unten abgebildet:

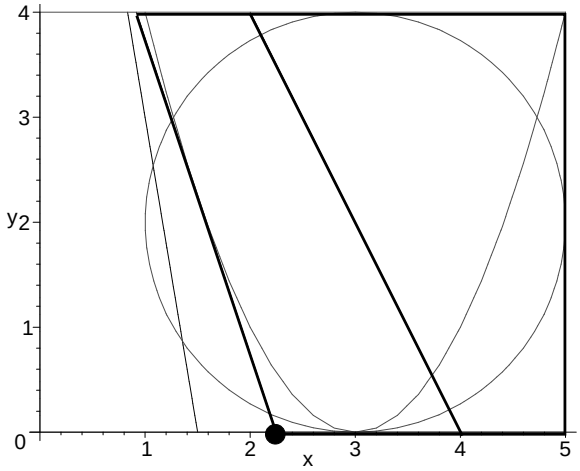


Die folgenden Bilder zeigen die ersten drei Schnitte:

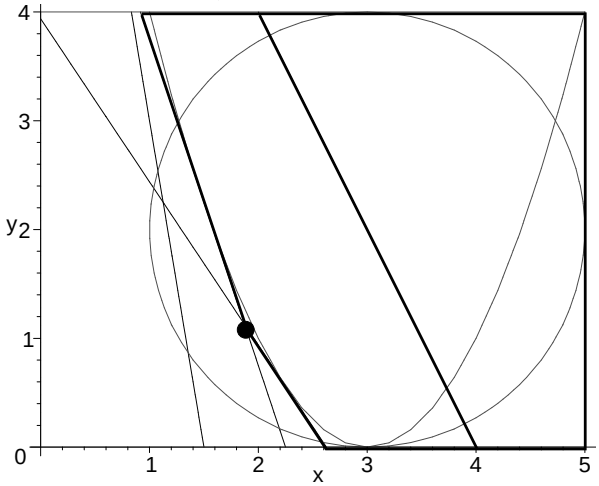
Schnitt 1: $-6x_1 - x_2 \leq -9$
Optimalwert: $x_1 = 1.5$, $x_2 = 0$



Schnitt 2: $-3x_1 - x_2 \leq -6.75$
Optimalwert: $x_1 = 2.25$, $x_2 = 0$



Schnitt 3: $-1.5x_1 - x_2 \leq -3.9375$
Optimalwert: $x_1 = 1.875$, $x_2 = 1.125$



Lösung:

```
ITER   KI           GKI                        F
-------------------------------------------------------------
   1    2   0.9000000000000000E+01  -0.3000000000000000E+01
   2    2   0.2250000000000000E+01  -0.4500000000000000E+01
   3    2   0.5625000000000000E+00  -0.4875000000000000E+01
   4    2   0.1406250000000000E+00  -0.4968750000000000E+01
   5    2   0.3515625000000000E-01  -0.4992187500000000E+01
   6    2   0.8789062500000000E-02  -0.4998046875000000E+01
   7    2   0.2197265625000000E-02  -0.4999511718750000E+01
   8    2   0.5493164062500000E-03  -0.4999877929687500E+01
   9    2   0.1373291015625000E-03  -0.4999969482421875E+01
  10    2   0.3433227539062500E-04  -0.4999992370605469E+01
  11    2   0.8583068847656250E-05  -0.4999998092651367E+01
  12    2   0.2145767211914062E-05  -0.4999999523162842E+01
  13    2   0.5364418029785156E-06  -0.4999999880790710E+01
  14    2   0.1341104507446289E-06  -0.4999999970197678E+01
  15    2   0.3352761268615723E-07  -0.4999999992549419E+01
  16    2   0.8381903171539307E-08  -0.4999999998137355E+01
  17    2   0.2095475792884827E-08  -0.4999999999534339E+01
  18    2   0.5238689482212067E-09  -0.4999999999883585E+01
  19    2   0.1309672370553017E-09  -0.4999999999970896E+01
  20    2   0.3274180926382542E-10  -0.4999999999992724E+01
  21    2   0.8185452315956354E-11  -0.4999999999998181E+01
  22    2   0.2046363078989089E-11  -0.4999999999999545E+01
  23    2   0.5115907697472721E-12  -0.4999999999999886E+01
  24    2   0.1278976924368180E-12  -0.4999999999999972E+01
  25    2   0.3197442310920451E-13  -0.4999999999999993E+01
  26    2   0.7993605777301127E-14  -0.4999999999999998E+01
  27    2   0.1998401444325282E-14  -0.5000000000000000E+01
  28    2   0.5427516075462435E-15  -0.5000000000000000E+01
  29    2  -0.5305543331057816E-15  -0.5000000000000000E+01

OPTIMAL SOLUTION FOUND (FEASIBILITY TOLERANCE   1.E-16):
X=         0.2000000004954636E+01        0.9999999900907280E+00
```

Kapitel 5

Restringierte Optimierung

In den folgenden Abschnitten werden wir zunächst notwendige und hinreichende Bedingungen für restringierte Optimierungsprobleme herleiten. Darauf aufbauend wird die Abhängigkeit der Lösung von Parametern untersucht und das duale Problem betrachtet. Schließlich untersuchen wir Lösungsverfahren, darunter die sequentielle quadratische Programmierung (SQP). Wir greifen dabei auf die Bücher [BS79], [GMW81], [FM90], [Spe93], [Man94], [GK99], [Alt02], [GK02] zurück.
Wir interessieren uns für das allgemeine Optimierungsproblem

$$\text{Minimiere } f(x) \qquad \text{unter} \qquad x \in \Sigma \tag{5.1}$$

mit $\Sigma \subseteq \mathbb{R}^n$, $\Sigma \neq \emptyset$.
Da die Aussagen für allgemeine zulässige Bereiche Σ sehr limitiert sind, werden wir in der Regel Optimierungsprobleme mit speziell strukturiertem zulässigen Bereich Σ betrachten. Dazu seien stetig differenzierbare Funktionen

$$\begin{aligned} f &: \mathbb{R}^n \to \mathbb{R}, \\ g = (g_1, \ldots, g_m)^\top &: \mathbb{R}^n \to \mathbb{R}^m, \\ h = (h_1, \ldots, h_p)^\top &: \mathbb{R}^n \to \mathbb{R}^p \end{aligned}$$

und eine abgeschlossene und konvexe Menge $S \subseteq \mathbb{R}^n$ mit $\operatorname{int}(S) \neq \emptyset$ gegeben. Mit

$$\Sigma := \{x \in S \mid g_i(x) \leq 0,\ i = 1, \ldots, m,\ h_j(x) = 0,\ j = 1, \ldots, p\} \tag{5.2}$$

erhalten wir das

(Standard-)Optimierungsproblem: Finde $x \in \mathbb{R}^n$, so daß $f(x)$ minimal wird unter den Nebenbedingungen

$$\begin{aligned} g_i(x) &\leq 0, \quad i = 1, \ldots, m, \\ h_j(x) &= 0, \quad j = 1, \ldots, p, \\ x &\in S. \end{aligned}$$

Beachte, daß Σ in (5.2) abgeschlossen ist, da g_i und h_j stetig sind und S abgeschlossen ist. Wir benötigen einige Begriffe und Definitionen.

Definition 5.0.3

Sei $x \in \Sigma$. Die Beschränkung $g_i(x) \leq 0$ heißt **aktiv in** *x, wenn $g_i(x) = 0$ gilt. Sie heißt* **inaktiv in** *$x \in \mathbb{R}^n$, wenn $g_i(x) < 0$ gilt. Die Menge*

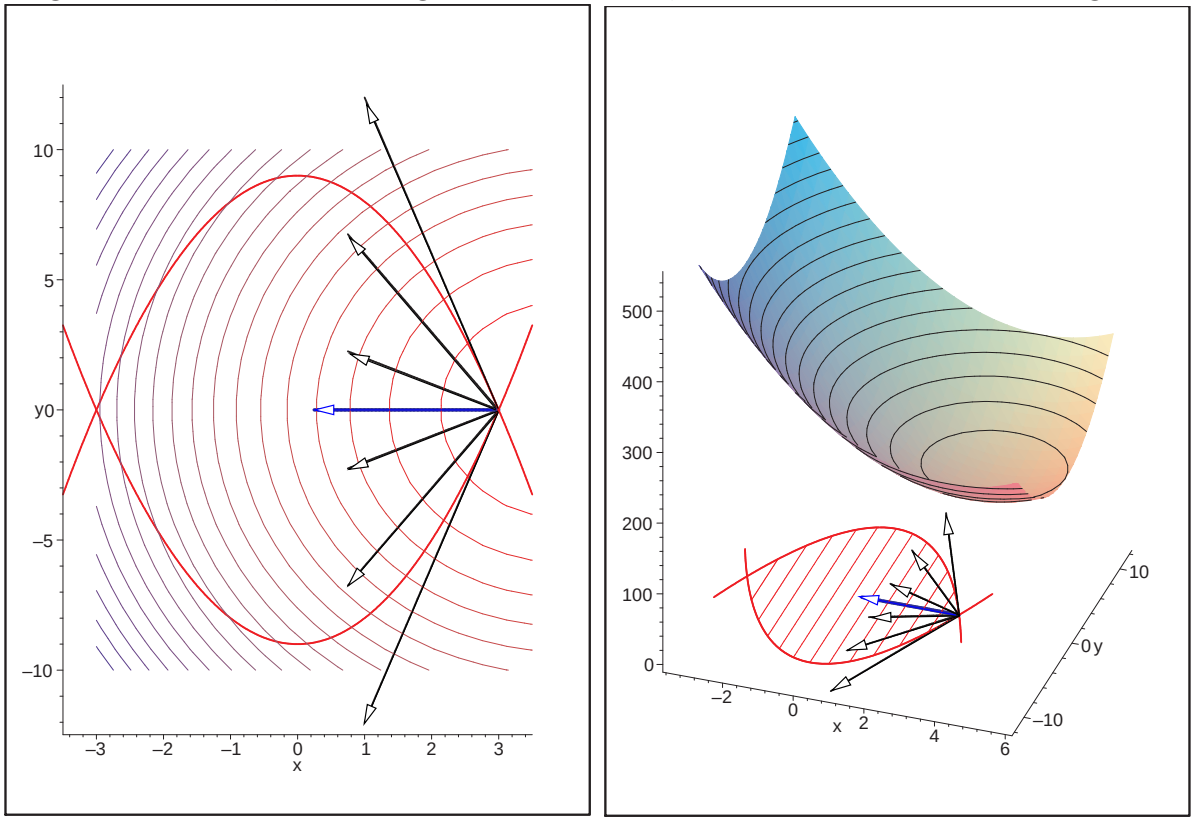
$$A(x) := \{i \mid g_i(x) = 0,\ 1 \leq i \leq m\}$$

heißt **Indexmenge der (in x) aktiven Ungleichungsrestriktionen**.

5.1 Geometrie und Tangentialkegel

Wir leiten geometrisch motivierte notwendige und hinreichende Bedingungen für das allgemeine Optimierungsproblem (5.1) her.

Sei $\hat{x}$ ein lokales Minimum von (5.1). Die beiden Abbildungen zeigen anschaulich, daß die Zielfunktion f entlang aller Richtungen, die in den zulässigen Bereich zeigen oder tangential verlaufen, notwendig nicht fallen kann, da sonst kein Minimum vorläge.



Wir formalisieren den Begriff tangential.

Definition 5.1.1 **(Tangentialkegel)**

Sei $x \in \Sigma$. Die Menge

$$T(\Sigma, x) = \left\{ d \in \mathbb{R}^n \;\middle|\; \begin{array}{l} \text{es gibt eine Folge } \{\alpha_k\}_{k\in\mathbb{N}}, \alpha_k \downarrow 0 \text{ und eine} \\ \text{Folge } \{x_k\}_{k\in\mathbb{N}}, x_k \in \Sigma \text{ mit } \lim_{k\to\infty} x_k = x, \\ \text{so daß } \lim_{k\to\infty} \frac{x_k - x}{\alpha_k} = d \text{ gilt.} \end{array} \right\}$$

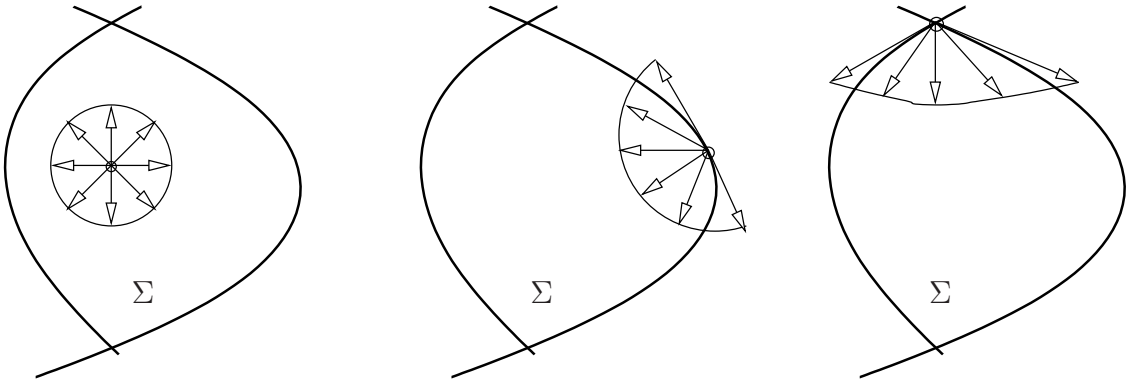


Abbildung 5.1: Tangentialkegel in verschiedenen Punkten des zulässigen Bereichs Σ.

heißt **Tangentialkegel an Σ in x.** *Jedes Element $d \in T(\Sigma, x)$ heißt* **tangentiale Richtung**.

Anschaulich bedeutet das: Gibt es in Σ eine Punktfolge x_k, die x beliebig nahe kommt, dann ist die Richtung

$$d = \lim_{k\to\infty} \frac{x_k - x}{\|x_k - x\|}$$

und alle positiven Vielfachen im Tangentialkegel enthalten. Falls der Grenzwert nicht existiert, liegen alle konvergenten Teilfolgen im Tangentialkegel, vgl. Abbildung 5.1.
Die Menge $T(\Sigma, x)$ ist – wie man leicht nachprüft – ein sogenannter Kegel (mit Spitze 0) und $x + T(\Sigma, x)$ stellt somit eine kegelförmige Approximation des zulässigen Bereichs dar.

Definition 5.1.2

Eine Menge $K \subseteq \mathbb{R}^n$ heißt **Kegel (mit Spitze 0)**, *wenn mit $d \in K$ auch $\alpha d \in K$ für alle $\alpha \geq 0$ gilt.*

Satz 5.1.3

Der Tangentialkegel $T(\Sigma, x)$ ist ein nichtleerer und abgeschlossener Kegel.

Beweis: Wegen $0 \in T(\Sigma, x)$ ist der Tangentialkegel nichtleer und mit d liegt auch jedes Vielfache αd, $\alpha > 0$ im Tangentialkegel.
Zu zeigen bleibt die Abgeschlossenheit. Sei $\{d_i\} \subseteq T(\Sigma, x)$ eine Folge mit $d_i \to d$.
Zu jedem $i \in \mathbb{N}$ gibt es dann eine Folge $\{\alpha_{i,k}\}_{k\in\mathbb{N}}$, $\alpha_{i,k} \downarrow 0$ und eine Folge $\{x_{i,k}\}_{k\in\mathbb{N}}$, $x_{i,k} \in \Sigma$ mit $\lim_{k\to\infty} x_{i,k} = x$, so daß

$$\lim_{k\to\infty} \frac{x_{i,k} - x}{\alpha_{i,k}} = d_i.$$

Zu jedem $i \in \mathbb{N}$ gibt es daher einen Index $k(i)$ mit

$$\|x_{i,k(i)} - x\| \leq \frac{1}{i}, \qquad \alpha_{i,k(i)} \leq \frac{1}{i}, \qquad \|\frac{x_{i,k(i)} - x}{\alpha_{i,k(i)}} - d_i\| \leq \frac{1}{i}.$$

Wegen

$$\|\frac{x_{i,k(i)} - x}{\alpha_{i,k(i)}} - d\| \leq \|\frac{x_{i,k(i)} - x}{\alpha_{i,k(i)}} - d_i\| + \|d_i - d\| \leq \frac{1}{i} + \|d_i - d\| \to 0$$

ergeben sich für $i \to \infty$ also Folgen $\{x_{i,k(i)}\}_{i\in\mathbb{N}}$, $\{\alpha_{i,k(i)}\}_{i\in\mathbb{N}}$, so daß $d \in T(\Sigma, x)$. □

Die folgende notwendige Bedingung erster Ordnung verwendet den Tangentialkegel und besagt anschaulich, daß es entlang der tangentialen Richtungen nicht bergab gehen darf.

Satz 5.1.4

Sei $\hat{x}$ lokales Minimum von (5.1) und $f : \mathbb{R}^n \to \mathbb{R}$ sei differenzierbar in $\hat{x}$. Dann gilt

$$\nabla f(\hat{x})^\top d \geq 0 \qquad \forall d \in T(\Sigma, \hat{x}).$$

Beweis: Sei $d \in T(\Sigma, \hat{x})$. Dann existieren Folgen $\alpha_k \downarrow 0$ und $x_k \in \Sigma$, $x_k \to \hat{x}$ mit $(x_k - \hat{x})/\alpha_k \to d$. Da $\hat{x}$ lokales Minimum und f differenzierbar in $\hat{x}$ ist, folgt

$$0 \leq f(x_k) - f(\hat{x}) = \nabla f(\hat{x})^\top (x_k - \hat{x}) + o(\|x_k - \hat{x}\|),$$

d.h. es gibt eine Folge $\xi_k \to 0$ mit

$$0 \leq \nabla f(\hat{x})^\top (x_k - \hat{x}) + \|x_k - \hat{x}\| \cdot \xi_k.$$

Division durch $\alpha_k > 0$ liefert

$$0 \leq \nabla f(\hat{x})^\top \underbrace{\left(\frac{x_k - \hat{x}}{\alpha_k}\right)}_{\to d} + \underbrace{\left\|\frac{x_k - \hat{x}}{\alpha_k}\right\|}_{\to \|d\|} \cdot \xi_k.$$

Da $d \in T(\Sigma, \hat{x})$ beliebig gewählt war, folgt die Behauptung. □

Definition 5.1.5 **(Normalkegel)**

Der **(negative) Normalkegel an Σ in $x \in \Sigma$** *ist definiert als*

$$N(\Sigma, x) := \{z \in \mathbb{R}^n \mid d^\top z \leq 0 \quad \forall d \in T(\Sigma, x)\}.$$

Die Elemente von $N(\Sigma, x)$ heißen **normale Richtungen an Σ in** *x, vgl. Abbildung 5.2.*

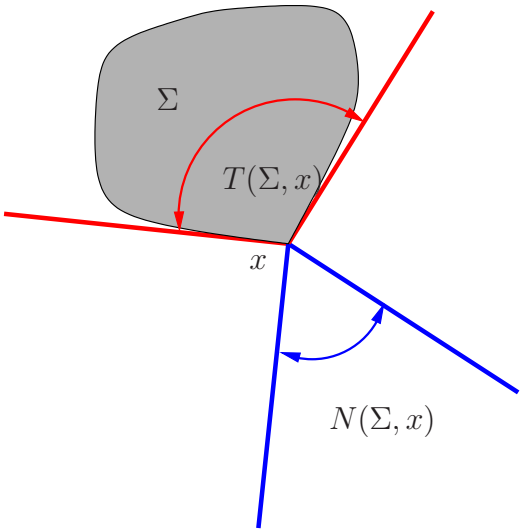


Abbildung 5.2: Tangentialkegel (rot) und Normalkegel (blau) an die Menge Σ

Unter Verwendung des Normalkegels läßt sich Satz 5.1.4 auch wie folgt formulieren:

Satz 5.1.6

Sei $\hat{x}$ lokales Minimum von (5.1) und $f : \mathbb{R}^n \to \mathbb{R}$ sei differenzierbar in $\hat{x}$. Dann gilt

$$-\nabla f(\hat{x}) \in N(\Sigma, \hat{x}).$$

Abschließend formulieren wir noch eine hinreichende Bedingung mit Hilfe des Tangentialkegels. Sie besagt anschaulich, daß ein zulässiger Punkt optimal ist, wenn es für sämtliche tangentialen Richtungen streng „bergauf“ geht.

Satz 5.1.7 (Hinreichende Bedingung erster Ordnung)

Sei $\hat{x} \in \Sigma$ und f differenzierbar in $\hat{x}$. Es gelte

$$\nabla f(\hat{x})^\top d > 0 \qquad \forall d \in T(\Sigma, \hat{x}) \setminus \{0\}.$$

Dann existiert eine Umgebung U von $\hat{x}$ und ein $\alpha > 0$ mit

$$f(x) \geq f(\hat{x}) + \alpha \|x - \hat{x}\| \qquad \forall x \in \Sigma \cap U$$

(insbesondere ist $\hat{x}$ also lokales Minimum).

Beweis: Angenommen, die Aussage ist falsch. Dann gibt es zu jeder Kugelumgebung

um $\hat{x}$ mit Radius $1/i$ einen Punkt $x_i \in \Sigma$ mit

$$f(x_i) - f(\hat{x}) < \frac{1}{i}\|x_i - \hat{x}\|, \quad \|x_i - \hat{x}\| < \frac{1}{i}, \quad \forall i \in \mathbb{N}. \tag{5.3}$$

Da die Einheitskugel bzgl. $\|\cdot\|$ in $\mathbb{R}^n$ kompakt ist, existiert eine konvergente Teilfolge von $\{x_i\}$ mit

$$\lim_{k\to\infty} \frac{x_{i_k} - \hat{x}}{\|x_{i_k} - \hat{x}\|} = \hat{d}, \quad \lim_{k\to\infty} \|x_{i_k} - \hat{x}\| = 0,$$

d.h. $\hat{d} \in T(\Sigma, \hat{x}) \setminus \{0\}$. Grenzübergang in (5.3) liefert den Widerspruch

$$\nabla f(\hat{x})^\top \hat{d} = \lim_{k\to\infty} \frac{f(x_{i_k}) - f(\hat{x})}{\|x_{i_k} - \hat{x}\|} \leq 0.$$

□

Bemerkung 5.1.8

- *Gilt $\hat{x} \in int(\Sigma)$ oder $\Sigma = \mathbb{R}^n$ (unrestringierter Fall), so folgt $T(\Sigma, \hat{x}) = \mathbb{R}^n$ und die notwendige Bedingung in Satz 5.1.4 impliziert $\nabla f(\hat{x}) = 0$. Beachte insbesondere, daß die hinreichende Bedingung in Satz 5.1.7 in diesem Fall für differenzierbare Funktionen f nicht erfüllbar ist!*

- *Der Tangentialkegel enthält die sogenannten* **zulässigen Richtungen***:*

  $$Z(\Sigma, x) = \{d \in \mathbb{R}^n \mid \exists \bar{\alpha} > 0 \; \forall \alpha \in (0, \bar{\alpha}] : \; x + \alpha d \in \Sigma\}$$

  Eine geometrisch einleuchtende und leicht zu beweisende notwendige Bedingung für ein lokales Minimum $\hat{x}$ ist:

  $$Z(\Sigma, \hat{x}) \cap \{d \in \mathbb{R}^n \mid \nabla f(\hat{x})^\top d < 0\} = \emptyset,$$

  d.h. in einem lokalen Minimum darf es keine zulässige Richtung geben, die zugleich Abstiegsrichtung ist, vgl. Definition 3.5.1, Hilfssatz 3.5.2. .

5.2 Notwendige Bedingungen für Standard-Optimierungsprobleme

Die notwendigen Bedingungen in Abschnitt 5.1 sind in der Praxis gar nicht oder nur sehr schwer nachprüfbar und verwertbar, da der Tangentialkegel schwer zu bestimmen ist.
In diesem Abschnitt werden wir notwendige Bedingungen für Standard-Optimierungsprobleme herleiten. Wir starten mit notwendigen Bedingungen erster Ordnung vom Fritz-John-Typ. Die Bezeichnung „erster Ordnung“ ist darin begründet, daß nur erste Ableitungen vorkommen. Unter Verwendung der Lagrangefunktion

$$L(x, l_0, \lambda, \mu) = l_0 f(x) + \lambda^\top g(x) + \mu^\top h(x) = l_0 f(x) + \sum_{i=1}^{m} \lambda_i g_i(x) + \sum_{j=1}^{p} \mu_j h_j(x)$$

gilt

Satz 5.2.1 (Notwendige Bedingungen erster Ordnung, Fritz-John Bedingungen)
Sei $\hat{x}$ ein lokales Minimum des Standard-Optimierungsproblems. Die Menge S sei abgeschlossen und konvex mit $\text{int}(S) \neq \emptyset$. Die Funktionen f, g_i, $i = 1, \ldots, m$ und h_j, $j = 1, \ldots, p$ seien stetig differenzierbar. Dann existieren Multiplikatoren $l_0 \in \mathbb{R}$, $\lambda = (\lambda_1, \ldots, \lambda_m)^\top \in \mathbb{R}^m$ und $\mu = (\mu_1, \ldots, \mu_p)^\top \in \mathbb{R}^p$ mit $(l_0, \lambda, \mu) \neq 0$, so daß die folgenden Bedingungen gelten:

(a) **Vorzeichenbedingungen:**

$$l_0 \geq 0, \qquad \lambda_i \geq 0, \qquad i = 1, \ldots, m. \tag{5.4}$$

(b) **Optimalitätsbedingung:**

$$L'_x(\hat{x}, l_0, \lambda, \mu)(x - \hat{x}) \geq 0 \qquad \forall x \in S. \tag{5.5}$$

bzw.

$$\left(l_0 f'(\hat{x}) + \sum_{i=1}^{m} \lambda_i g'_i(\hat{x}) + \sum_{j=1}^{p} \mu_j h'_j(\hat{x}) \right) (x - \hat{x}) \geq 0 \qquad \forall x \in S. \tag{5.6}$$

(c) **Komplementaritätsbedingungen:**

$$\lambda_i g_i(\hat{x}) = 0, \qquad i = 1, \ldots, m. \tag{5.7}$$

(d) **Zulässigkeit:**

$$\hat{x} \in \Sigma. \tag{5.8}$$

Beweis: Betrachte das linearisierte Problem

Minimiere	$f(\hat{x}) + f'(\hat{x})(x - \hat{x})$
bzgl.	$x \in S$
unter	$g_i(\hat{x}) + g'_i(\hat{x})(x - \hat{x}) \leq 0, \quad i = 1, \ldots, m,$
	$h_j(\hat{x}) + h'_j(\hat{x})(x - \hat{x}) = 0, \quad j = 1, \ldots, p.$

Für den ersten Teil des Beweises nehmen wir zunächst an, daß die Abbildung $h'(\hat{x})(\cdot) : \mathbb{R}^n \to \mathbb{R}^p$ surjektiv ist (die Matrix $h'(\hat{x})$ besitze also vollen Rang bzw. die Gradienten $\nabla h_j(\hat{x})$ sind linear unabhängig) und daß es einen zulässigen Punkt $x \in \text{int}(S)$ für das linearisierte Problem gibt mit

$$g_i(\hat{x}) + g'_i(\hat{x})(x - \hat{x}) < 0, \quad i = 1, \ldots, m, \tag{5.9}$$

$$h_j(\hat{x}) + h'_j(\hat{x})(x - \hat{x}) = 0, \quad j = 1, \ldots, p, \tag{5.10}$$

$$f(\hat{x}) + f'(\hat{x})(x - \hat{x}) < f(\hat{x}). \tag{5.11}$$

(i) Da $h'(\hat{x})$ surjektiv ist, gibt es $y_j \in \mathbb{R}^n$, $j = 1, \dots, p$ mit

$$h'(\hat{x})y_j = e_j, \quad j = 1, \dots, p, \tag{5.12}$$

wobei $e_j \in \mathbb{R}^p$ den j-ten Einheitsvektor bezeichnet. Betrachte das nichtlineare Gleichungssystem

$$H(t, r) := h\left(\hat{x} + t(x - \hat{x}) + \sum_{j=1}^{p} r_j y_j\right) = 0$$

für die Variablen $t \in \mathbb{R}$ und $r = (r_1, \dots, r_p)^\top \in \mathbb{R}^p$. Die Jacobimatrix von H bzgl. r in $(t, r) = (0, 0)$ ist invertierbar wegen

$$\frac{\partial H}{\partial r}(0, 0) = h'(\hat{x}) \cdot (y_1 | \cdots | y_p) = I.$$

Desweiteren gilt $H(0, 0) = h(\hat{x}) = 0$. Daher sind die Voraussetzungen des Satzes über implizite Funktionen erfüllt und Anwendung des Satzes liefert die Existenz von $\varepsilon_1 > 0$ und $r : (-\varepsilon_1, \varepsilon_1) \to \mathbb{R}^p$, $t \mapsto r(t)$ mit

$$G(t) := H(t, r(t)) = h\left(\hat{x} + t(x - \hat{x}) + \sum_{j=1}^{p} r_j(t) y_j\right) = 0, \quad \forall t \in (-\varepsilon_1, \varepsilon_1)$$

und $r(0) = 0$. Zusätzlich gilt

$$0 = \frac{dG}{dt}(0) = h'(\hat{x}) \cdot \left(x - \hat{x} + \sum_{j=1}^{p} y_j \frac{dr_j}{dt}(0)\right) \overset{(5.10),(5.12)}{=} \frac{dr}{dt}(0).$$

Dies zeigt, daß die Kurve

$$x(t) = \hat{x} + t(x - \hat{x}) + \sum_{j=1}^{p} r_j(t) y_j$$

für $t \in (-\varepsilon_1, \varepsilon_1)$ zulässig bleibt für die nichtlinearen Gleichungsrestriktionen. Zudem gilt

$$x(0) = \hat{x}, \quad x'(0) = x - \hat{x}.$$

(ii) Nun betrachten wir die in $\hat{x}$ inaktiven Ungleichungsrestriktionen $g_i(\hat{x})$, $i \notin A(\hat{x})$. Wegen der Stetigkeit von g_i gilt $g_i(x(t)) < 0$ für die Kurve $x(t)$ aus (i) mit hinreichend kleinem t.

Für die in $\hat{x}$ aktiven Ungleichungsrestriktionen $g_i(\hat{x}) = 0$, $i \in A(\hat{x})$ gilt $g_i(x(0)) = 0$, $i \in A(\hat{x})$ und

$$\frac{d}{dt} g_i(x(t))\Big|_{t=0} = g_i'(\hat{x}) \cdot \frac{dx}{dt}(0) = g_i'(\hat{x})(x - \hat{x}) \overset{(5.9)}{<} 0.$$

Da g_i stetig differenzierbar ist, gilt $g_i(x(t)) < 0$, $i \in A(\hat{x})$ für hinreichend kleines $t > 0$.

Insgesamt haben wir gezeigt, daß die Kurve $x(t)$ für hireichend kleines $t > 0$ auch zulässig ist für die nichtlinearen Ungleichungsrestriktionen.

(iii) Die Zielfunktion f wird analog zu (ii) behandelt. Es gilt

$$\frac{d}{dt} f(x(t))\Big|_{t=0} = f'(\hat{x}) \cdot \frac{dx}{dt}(0) = f'(\hat{x})(x - \hat{x}) \overset{(5.11)}{<} 0.$$

Also ist $x - \hat{x}$ eine Abstiegsrichtung von f in $\hat{x}$. Da f stetig differenzierbar ist, gilt $f(x(t)) < f(\hat{x})$ für hinreichend kleines $t > 0$ (Hieraus ergibt sich später ein Widerspruch zur lokalen Minimalität von $\hat{x}$).

(iv) Der Punkt x im linearisierten Problem mit (5.9)-(5.11) ist ein innerer Punkt von S, vgl. Abbildung 5.3.

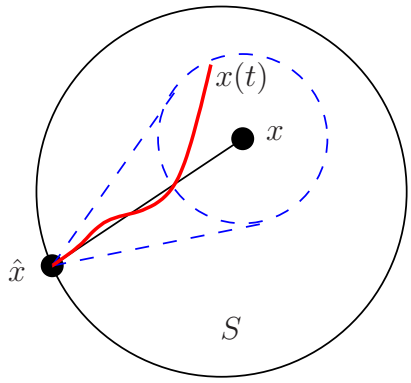


Abbildung 5.3: Fritz-John Bedingungen unter der Beschränkung $x \in S$. Die Annahme $\operatorname{int}(S) \neq \emptyset$ ist wesentlich. Für hinreichend kleines $t > 0$ bleibt die Kurve $x(t)$ zulässig.

Dann gibt es eine Umgebung $U_\delta(x)$, so daß $\hat{x} + t(z - \hat{x}) \in S$ für alle $0 \le t \le 1$ und alle $z \in U_\delta(x)$. Wegen

$$0 = r'(0) = \lim_{t \to 0} \frac{r(t) - r(0)}{t} = \lim_{t \to 0} \frac{r(t)}{t}$$

existiert $\varepsilon_2 > 0$ mit

$$\left\| \sum_{j=1}^{p} \frac{r_j(t)}{t} y_j \right\| < \delta$$

für alle $0 < t < \varepsilon_2$. Daher gilt

$$x(t) = \hat{x} + t(x - \hat{x}) + \sum_{j=1}^{p} r_j(t) y_j = \hat{x} + t\Big(\underbrace{x + \sum_{j=1}^{p} \frac{r_j(t)}{t} y_j}_{\in U_\delta(x)} - \hat{x} \Big) \in S.$$

Die Punkte (i)-(iv) zeigen das Folgende: Ist $\hat{x}$ ein lokales Minimum des Standard-Optimierungsproblems und $h'(\hat{x})$ surjektiv, dann existiert kein $x \in \operatorname{int}(S)$ mit (5.9)-(5.11), denn gäbe es ein solches x, so existierte nach (i)-(iv) eine für das Standard-Optimierungsproblem zulässige Kurve $x(t)$ mit $f(x(t)) < f(\hat{x})$ für hinreichend kleines $t > 0$.
Betrachte die nichtleeren konvexen Mengen

$$A := \left\{ \begin{pmatrix} z_f \\ z_g \\ z_h \end{pmatrix} \mid z_f \leq f(\hat{x}),\ z_g \leq 0,\ z_h = 0 \right\},$$

$$B := \left\{ \begin{pmatrix} f(\hat{x}) + f'(\hat{x})(x-\hat{x}) \\ g(\hat{x}) + g'(\hat{x})(x-\hat{x}) \\ h(\hat{x}) + h'(\hat{x})(x-\hat{x}) \end{pmatrix} \;\middle|\; x \in S \right\}.$$

Ist $h'(\hat{x})$ nicht surjektiv, so können diese Mengen trivialerweise durch eine Hyperebene getrennt werden, da beide Mengen in einer Hyperebene enthalten sind.
Ist $h'(\hat{x})$ surjektiv, so zeigen die Betrachtungen in (i)-(iv), daß $\operatorname{relint}(A) \cap \operatorname{relint}(B) = \emptyset$. Trennungssatz 4.1.8 liefert die Existenz einer Hyperebene, die A und B trennt. Es existieren also nichttriviale Multiplikatoren $l_0 \in \mathbb{R}, \lambda \in \mathbb{R}^m, \mu \in \mathbb{R}^p$ mit

$$\begin{aligned} l_0 z_f + \lambda^\top z_g + \mu^\top z_h \quad \leq \quad & l_0 \left(f(\hat{x}) + f'(\hat{x})(x-\hat{x})\right) \\ & + \lambda^\top \left(g(\hat{x}) + g'(\hat{x})(x-\hat{x})\right) \\ & + \mu^\top \left(h(\hat{x}) + h'(\hat{x})(x-\hat{x})\right) \end{aligned}$$

für alle $x \in S$, $z_f \leq f(\hat{x})$, $z_g \leq 0$ und $z_h = 0$. Äquivalent gilt

$$0 \leq l_0(f(\hat{x}) - z_f) + \lambda^\top (g(\hat{x}) - z_g) + \left(l_0 f'(\hat{x}) + \lambda^\top g'(\hat{x}) + \mu^\top h'(\hat{x})\right)(x - \hat{x})$$

für alle $x \in S$, $z_f \leq f(\hat{x})$ und $z_g \leq 0$. Die Wahl $z_g = g(\hat{x}) \leq 0$ und $z_f = f(\hat{x})$ liefert

$$\left(l_0 f'(\hat{x}) + \lambda^\top g'(\hat{x}) + \mu^\top h'(\hat{x})\right)(x - \hat{x}) \geq 0$$

für alle $x \in S$. Die Wahl $x = \hat{x} \in S$ liefert

$$0 \leq l_0(\underbrace{f(\hat{x}) - z_f}_{\geq 0}) + \lambda^\top(\underbrace{g(\hat{x})}_{\leq 0} - \underbrace{z_g}_{\leq 0})$$

für alle $z_f \leq f(\hat{x})$ und $z_g \leq 0$. Die letztere Ungleichung impliziert

$$l_0 \geq 0, \quad \lambda \geq 0, \quad \lambda^\top g(\hat{x}) = 0.$$

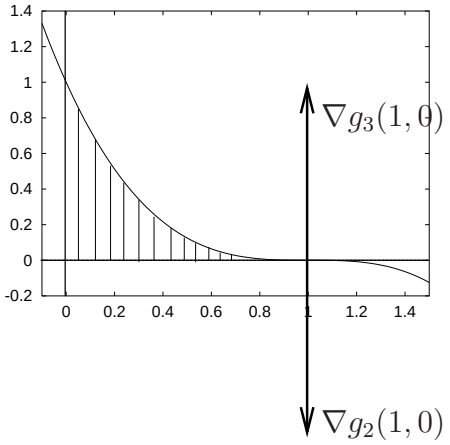
Dies vervollständigt den Beweis. □

Beispiel 5.2.2

Betrachte das Optimierungsproblem

$$\begin{array}{ll} \textit{Minimiere} & f(x_1,x_2) = -x_1 \\ \textit{bzgl.} & (x_1,x_2)^\top \in S := \mathbb{R}^2 \\ \textit{unter} & g_1(x_1,x_2) = -x_1 \le 0, \\ & g_2(x_1,x_2) = -x_2 \le 0, \\ & g_3(x_1,x_2) = -(1-x_1)^3 + x_2 \le 0. \end{array}$$

Das Optimum wird in $\hat{x}_1 = 1, \hat{x}_2 = 0$ *angenommen, siehe Abbildung.*



Wegen $S = \mathbb{R}^2$ *reduziert sich die Optimalitätsbedingung (5.5) bzw. (5.6) für* $\hat{x} = (\hat{x}_1, \hat{x}_2)^\top$ *und* $\lambda = (\lambda_1, \lambda_2, \lambda_3)^\top \ge 0$ *unter Ausnutzung der Komplementaritätsbedingung auf*

$$\begin{aligned} 0 &= l_0 \nabla f(\hat{x}_1,\hat{x}_2) + \underbrace{\lambda_1}_{=0,\ da\ g_1(1,0)=-1<0} \nabla g_1(\hat{x}_1,\hat{x}_2) + \underbrace{\lambda_2}_{\ge 0} \nabla g_2(\hat{x}_1,\hat{x}_2) + \underbrace{\lambda_3}_{\ge 0} \nabla g_3(\hat{x}_1,\hat{x}_2) \\ &= l_0 \begin{pmatrix} -1 \\ 0 \end{pmatrix} + \lambda_2 \begin{pmatrix} 0 \\ -1 \end{pmatrix} + \lambda_3 \begin{pmatrix} 0 \\ 1 \end{pmatrix}. \end{aligned}$$

Offensichtlich ist diese Bedingung nur für $l_0 = 0$ *und* $\lambda_2 = \lambda_3$ *erfüllt.*

Wie im Abschnitt über konvexe Optimierung heißt jeder Vektor $(x, l_0, \lambda, \mu) \in \mathbb{R}^{n+1+m+p}$ mit $(l_0, \lambda, \mu) \neq 0$, der die Fritz-John-Bedingungen (5.4)-(5.8) erfüllt, **Fritz-John Punkt** des Standard-Optimierungsproblems. Die Multiplikatoren l_0, λ und μ heißen **Lagrange-Multiplikatoren**. Die wesentliche Aussage des Satzes ist wiederum, daß es einen **nichttrivialen** Vektor $(l_0, \lambda, \mu) \neq 0$ gibt. Für $(l_0, \lambda, \mu) = 0$ sind die Fritz-John-Bedingungen trivial. Wie im konvexen Fall, kann auch hier der Fall $l_0 = 0$ auftreten, vgl. Beispiel 5.2.2. In diesem Fall tritt die Zielfunktion f nicht in den Fritz-John-Bedingungen auf. Im folgenden sind wir an Bedingungen interessiert, die $l_0 \neq 0$ liefern. Da alle Multiplikatoren linear auftreten, kann dann o.B.d.A. $l_0 = 1$ gewählt werden. Solche Bedingungen hei-

ßen **Regularitätsbedingungen** oder **Constraint Qualifications**. Sind die Fritz-John-Bedingungen mit $l_0 = 1$ erfüllt, so heißen sie **Karush-Kuhn-Tucker (KKT) Bedingungen**. Im konvexen Fall hatten wir bereits eine Regularitätsbedingung kennen gelernt: die Slater-Bedingung. Wir werden nun einige gebräuchliche Regularitätsbedingungen für nichtkonvexe Probleme vorstellen:

- Regularitätsbedingung von Robinson
- Regularitätsbedingung von Mangasarian-Fromowitz
- Regularitätsbedingung der linearen Unabhängigkeit (Linear Independence Constraint Qualification, LICQ)

Darüber hinaus gibt es zahlreiche weitere Bedingungen (Abadie, Kuhn-Tucker, Guignard, Arrow-Hurrwitz,...).
Wir starten mit der Regularitätsbedingung von Robinson, die Robinson [Rob76, Rob82] ursprünglich im Zusammenhang mit Stabilitätsuntersuchungen für verallgemeinerte Ungleichungen formulierte.

Definition 5.2.3 **(Regularitätsbedingung von Robinson)**
Die **Regularitätsbedingung von Robinson** *ist in $\hat{x}$ erfüllt, wenn*

$$(0,0)^\top \in int(M),$$

wobei

$$M := \left\{ \begin{pmatrix} g(\hat{x}) + g'(\hat{x})(x-\hat{x}) - k \\ h'(\hat{x})(x-\hat{x}) \end{pmatrix} \;\middle|\; x \in S,\ k \leq 0 \right\}.$$

Satz 5.2.4 **(Karush-Kuhn-Tucker (KKT) Bedingungen I)**
Es seien die Voraussetzungen von Satz 5.2.1 erfüllt. Zusätzlich sei die Regularitätsbedingung von Robinson in $\hat{x}$ erfüllt. Dann gelten die Aussagen von Satz 5.2.1 mit $l_0 = 1$.

Beweis: Angenommen, es gilt $l_0 = 0$ in Satz 5.2.1. Dann werden die Projektionen

$$\begin{aligned} \text{proj}_{\mathbb{R}^{m+p}}(A) &:= \left\{ \begin{pmatrix} z_g \\ z_h \end{pmatrix} \mid z_g \leq 0,\ z_h = 0 \right\}, \\ \text{proj}_{\mathbb{R}^{m+p}}(B) &:= \left\{ \begin{pmatrix} g(\hat{x}) + g'(\hat{x})(x-\hat{x}) \\ h(\hat{x}) + h'(\hat{x})(x-\hat{x}) \end{pmatrix} \;\middle|\; x \in S \right\} \end{aligned}$$

von A und B im Beweis von Satz 5.2.1 durch eine Hyperebene getrennt, d.h. es gibt $(0,0)^\top \neq (\lambda,\mu)^\top \in \mathbb{R}^{m+p}$ mit

$$\lambda^\top k \leq \lambda^\top (g(\hat{x}) + g'(\hat{x})(x-\hat{x})) + \mu^\top (h(\hat{x}) + h'(\hat{x})(x-\hat{x}))$$

für alle $x \in S$ und alle $k \leq 0$. Folglich kann $(0,0)^\top$ von $\text{proj}_{\mathbb{R}^{m+p}}(B) - \text{proj}_{\mathbb{R}^{m+p}}(A)$ getrennt werden, da

$$0 \leq \lambda^\top (g(\hat{x}) + g'(\hat{x})(x - \hat{x}) - k) + \mu^\top h'(\hat{x})(x - \hat{x})$$

für alle $x \in S$ und alle $k \leq 0$ gilt. Beachte, daß $M = \text{proj}_{\mathbb{R}^{m+p}}(B) - \text{proj}_{\mathbb{R}^{m+p}}(A)$ gilt und daß nach Voraussetzung $(0,0)^\top \in \text{int}(M)$ ist. Also ist es unmöglich, $(0,0)^\top$ von M zu trennen. Dies liefert einen Widerspruch und es kann nicht $l_0 = 0$ gelten. □

Eine weitere Regularitätsbedingung wurde von Mangasarian und Fromowitz [MF67] formuliert. Tatsächlich ist sie sogar äquivalent mit der Regularitätsbedingung von Robinson.

Definition 5.2.5 **(Regularitätsbedingung von Mangasarian-Fromowitz)**
Die **Regularitätsbedingung von Mangasarian-Fromowitz** *ist in $\hat{x}$ erfüllt, wenn die folgenden Bedingungen gelten:*

(a) Die Gradienten $\nabla h_j(\hat{x})$, $j = 1, \ldots, p$ sind linear unabhängig.

(b) Es gibt einen Vektor $\hat{d} \in int(S - \{\hat{x}\})$ mit

$$\nabla g_i(\hat{x})^\top \hat{d} < 0 \text{ für } i \in A(\hat{x}) \text{ und } \nabla h_j(\hat{x})^\top \hat{d} = 0 \text{ für } j = 1, \ldots, p.$$

Satz 5.2.6 **(Karush-Kuhn-Tucker (KKT) Bedingungen II)**
Es seien die Voraussetzungen von Satz 5.2.1 erfüllt. Zusätzlich sei die Regularitätsbedingung von Mangasarian-Fromowitz in $\hat{x}$ erfüllt. Dann gelten die Aussagen von Satz 5.2.1 mit $l_0 = 1$.

Beweis: Angenommen, es gilt $l_0 = 0$ in Satz 5.2.1, d.h. es gibt Multiplikatoren $\lambda_i \geq 0$, $i = 1, \ldots, m$, μ_j, $j = 1, \ldots, p$ nicht alle Null mit $\lambda_i g_i(\hat{x}) = 0$ für alle $i = 1, \ldots, m$ und

$$L'_x(\hat{x}, 0, \lambda, \mu)d = \left(\sum_{i=1}^{m} \lambda_i g'_i(\hat{x}) + \sum_{j=1}^{p} \mu_j h'_j(\hat{x}) \right) d \geq 0 \quad \forall d \in S - \{\hat{x}\}. \tag{5.13}$$

Sei $\hat{d}$ der Vektor aus der Regularitätsbedingung von Mangasarian-Fromowitz. Dann folgt

$$\sum_{i=1}^{m} \lambda_i \nabla g_i(\hat{x})^\top \hat{d} = \sum_{i \in A(\hat{x})} \lambda_i \nabla g_i(\hat{x})^\top \hat{d} \;\geq\; 0.$$

Wegen $\lambda_i \geq 0$ und $\nabla g_i(\hat{x})^\top \hat{d} < 0$ für $i \in A(\hat{x})$ kann diese Ungleichung nur für $\lambda_i = 0$, $i \in A(\hat{x})$ gelten. Wegen $\lambda_i = 0$ für $i \notin A(\hat{x})$ gilt dann $\lambda_i = 0$ für alle $i = 1, \ldots, m$.
Also reduziert sich Ungleichung (5.13) auf

$$0 \leq \sum_{j=1}^{p} \mu_j \nabla h_j(\hat{x})^\top d = \sum_{j=1}^{p} \mu_j (\nabla h_j(\hat{x})^\top d - \nabla h_j(\hat{x})^\top \hat{d}) = \sum_{j=1}^{p} \mu_j \nabla h_j(\hat{x})^\top (d - \hat{d})$$

für alle $d \in S - \{\hat{x}\}$. Da $\hat{d} \in \text{int}(S - \{\hat{x}\})$ ist und $\nabla h_j(\hat{x})$, $j = 1, \ldots, p$ linear unabhängig sind ($h'(\hat{x})$ ist also surjektiv), kann diese Ungleichung nur für $\mu_j = 0$, $j = 1, \ldots, p$ gelten. Insgesamt haben wir also $(l_0, \lambda, \mu) = 0$ gezeigt, was der Aussage widerspricht, daß nicht alle Multiplikatoren gleich Null sind. □

Die folgende Regularitätsbedingung der linearen Unabhängigkeit ist im Hinblick auf die Formulierung hinreichender Bedingungen und die Sensitivitätsanalyse sehr wichtig. Aus ihr folgt die Regularitätsbedingung von Mangasarian-Fromowitz. Sie stellt sogar sicher, daß die Lagrange-Multiplikatoren eindeutig sind, was bei den vorangehenden Bedingungen nicht der Fall ist.

Definition 5.2.7 **(Regularitätsbedingung der linearen Unabhängigkeit (LICQ))**
Die **Regularitätsbedingung der linearen Unabhängigkeit** *oder* **Linear Independence Constraint Qualification (LICQ)** *ist in $\hat{x}$ erfüllt, wenn die folgenden Bedingungen gelten:*

(a) $\hat{x} \in int(S)$;

(b) Die Gradienten $\nabla g_i(\hat{x})$, $i \in A(\hat{x})$ und $\nabla h_j(\hat{x})$, $j = 1, \ldots, p$ sind linear unabhängig.

Satz 5.2.8 **(Karush-Kuhn-Tucker (KKT) Bedingungen III)**
Es seien die Voraussetzungen von Satz 5.2.1 erfüllt. Zusätzlich sei die Regularitätsbedingung der linearen Unabhängigkeit in $\hat{x}$ erfüllt. Dann gelten die Aussagen von Satz 5.2.1 mit $l_0 = 1$. Insbesondere gilt

$$\nabla_x L(\hat{x}, l_0 = 1, \lambda, \mu) = \nabla f(\hat{x}) + \sum_{i=1}^{m} \lambda_i \nabla g_i(\hat{x}) + \sum_{j=1}^{p} \mu_j \nabla h_j(\hat{x}) = 0.$$

Desweiteren sind die Multiplikatoren λ und μ eindeutig.

Beweis: Angenommen, es gilt $l_0 = 0$ in Satz 5.2.1. Da $\hat{x} \in \text{int}(S)$ gilt, muß Ungleichung (5.5) für alle $x \in \mathbb{R}^n$ gelten und es folgt

$$\sum_{i=1}^{m} \lambda_i \nabla g_i(\hat{x}) + \sum_{j=1}^{p} \mu_j \nabla h_j(\hat{x}) = \sum_{i \in A(\hat{x})} \lambda_i \nabla g_i(\hat{x}) + \sum_{j=1}^{p} \mu_j \nabla h_j(\hat{x}) = 0.$$

Die lineare Unabhängigkeit der Gradienten liefert $\lambda_i = \mu_j = 0$ für alle $i = 1, \ldots, m$, $j = 1, \ldots, p$. Dies widerspricht wieder der Aussage, daß nicht alle Multiplikatoren gleich Null sind.
Die Eindeutigkeit der Multiplikatoren ergibt sich wie folgt. Es seien λ_i, $i = 1, \ldots, m$, μ_j, $j = 1, \ldots, p$ und $\tilde{\lambda}_i$, $i = 1, \ldots, m$, $\tilde{\mu}_j$, $j = 1, \ldots, p$ Multiplikatoren, die die KKT-

Bedingungen erfüllen. Dann folgt aus $\hat{x} \in \operatorname{int}(S)$ wieder

$$0 = \nabla f(\hat{x}) + \sum_{i=1}^{m} \lambda_i \nabla g_i(\hat{x}) + \sum_{j=1}^{p} \mu_j \nabla h_j(\hat{x}),$$
$$0 = \nabla f(\hat{x}) + \sum_{i=1}^{m} \tilde{\lambda}_i \nabla g_i(\hat{x}) + \sum_{j=1}^{p} \tilde{\mu}_j \nabla h_j(\hat{x}).$$

Subtraktion der Gleichungen liefert

$$0 = \sum_{i=1}^{m} (\lambda_i - \tilde{\lambda}_i) \nabla g_i(\hat{x}) + \sum_{j=1}^{p} (\mu_j - \tilde{\mu}_j) \nabla h_j(\hat{x}).$$

Für inaktive Ungleichungen gilt $\lambda_i = \tilde{\lambda}_i = 0$, $i \notin A(\hat{x})$. Da die Gradienten der aktiven Beschränkungen linear unabhängig sind, folgt $0 = \lambda_i - \tilde{\lambda}_i$, $i \in A(\hat{x})$ und $0 = \mu_j - \tilde{\mu}_j$, $j = 1, \ldots, p$. Also sind die Lagrange-Multiplikatoren eindeutig. □

Bemerkung 5.2.9

Es gibt zahlreiche andere Regularitätsbedingungen. Eine der schwächsten Bedingungen (in dem Sinne, daß sie aus den bisher diskutierten Regularitätsbedingungen folgt) ist die **Regularitätsbedingung von Abadie**, *welche im Fall $S = \mathbb{R}^n$ fordert, daß*

$$T_{lin}(\hat{x}) = T(\Sigma, \hat{x})$$

gilt, wobei

$$T_{lin}(x) \;=\; \{d \in \mathbb{R}^n \mid \nabla g_i(x)^\top d \le 0,\ i \in A(x),\ \nabla h_j(x)^\top d = 0,\ j = 1, \ldots, p\}$$

den **linearisierenden Kegel in** $\hat{x}$ *bezeichnet. (Im allgemeinen gilt nur $T(\Sigma, \hat{x}) \subseteq T_{lin}(\hat{x})$, wie das Beispiel $n = 2, m = 3, p = 0$, $g(x_1, x_2) = (-(1 - x_1)^3 + x_2, -x_1, -x_2)^\top$ verdeutlicht, denn für $x = (1, 0)^\top$ gilt $T(\Sigma, x) = \{(\alpha, 0)^\top \mid \alpha \le 0\}$ und $T_{lin}(x) = \{(\alpha, 0)^\top \mid \alpha \in \mathbb{R}\}$.)*

Es sei erwähnt, daß der Tangentialkegel $T(\Sigma, x)$ unabhängig von der Darstellung des zulässigen Bereichs Σ durch Ungleichungen und Gleichungen ist, während der linearisierende Kegel $T_{lin}(x)$ von den Funktionen g_i und h_j, die Σ beschreiben, abhängt (also: zusätzliche Restriktionen, die Σ nicht ändern, können $T_{lin}(x)$ ändern).

Die Bedingung von Abadie besitzt zudem den Nachteil, daß sie nur sehr schwer überprüfbar ist. Außerdem garantiert sie im Gegensatz zu den zuvor behandelten Bedingungen nicht die lokale Stabilität des zulässigen Bereichs unter Störungen.

Im folgenden beschränken wir uns stets auf den wichtigen Spezialfall $S = \mathbb{R}^n$, der insbesondere für die Konstruktion numerischer Verfahren, wie z.B. SQP Verfahren, von Interesse

ist. Die Optimalitätsbedingung (5.5) bzw. (5.6) muß dann für alle $x \in \mathbb{R}^n$ gelten, was gleichbedeutend ist mit

$$l_0 \nabla f(\hat{x}) + \sum_{i=1}^{m} \lambda_i \nabla g_i(\hat{x}) + \sum_{j=1}^{p} \mu_j \nabla h_j(\hat{x}) = 0.$$

Die diskutierten Regularitätsbedingungen sichern zudem $l_0 = 1$.
In dieser Darstellung kann man auch die **geometrische Bedeutung der KKT-Bedingungen** ablesen:

$$-\nabla f(\hat{x}) = \sum_{i=1}^{m} \lambda_i \nabla g_i(\hat{x}) + \sum_{j=1}^{p} \mu_j \nabla h_j(\hat{x}).$$

Der negative Gradient von f kann in einem Optimum also als **Linearkombination der Gradienten der aktiven Nebenbedingungen** dargestellt werden. Zu beachten sind hierbei aber die Vorzeichenbedingungen für λ! Es ist also keine beliebige Linearkombination. Beispiel 5.2.2 zeigt zudem, daß eine solche Darstellung nicht immer möglich ist (Fritz-John Fall mit $l_0 = 0$!).
Spezialfälle:

- Für $m = p = 0$ ergibt sich $\nabla f(\hat{x}) = 0$ ($\rightarrow$ unrestringierte Optimierung).
- Für $m = 0, p > 0$ (keine Ungleichungsrestriktionen) ergibt sich die Multiplikatorenregel von Lagrange

  $$\nabla f(\hat{x}) + \sum_{j=1}^{p} \mu_j \nabla h_j(\hat{x}) = 0,$$

  welche in der Analysis-Vorlesung unter der LICQ-Bedingung bewiesen wird.

Abschließend formulieren wir notwendige Bedingungen zweiter Ordnung, wobei wir voraussetzen, daß $(\hat{x}, \hat{\lambda}, \hat{\mu})$ ein KKT-Punkt des Standard-Optimierungsproblems mit $S = \mathbb{R}^n$ ist. Wir benötigen den sogenannten **kritischen Kegel**

$$\begin{aligned} T_K(\hat{x}) \;:=\; \{d \in \mathbb{R}^n \mid\; & \nabla g_i(\hat{x})^\top d \le 0,\ i \in A(\hat{x}), \hat{\lambda}_i = 0, \\ & \nabla g_i(\hat{x})^\top d = 0,\ i \in A(\hat{x}), \hat{\lambda}_i > 0, \\ & \nabla h_j(\hat{x})^\top d = 0,\ j = 1, \ldots, p\}. \end{aligned}$$

Satz 5.2.10 **(Notwendige Bedingungen zweiter Ordnung)**
Es seien f, g_i, $i = 1, \ldots, m$, und h_j, $j = 1, \ldots, p$ zweimal stetig differenzierbar und $S = \mathbb{R}^n$. Desweiteren seien $(\hat{x}, \hat{\lambda}, \hat{\mu})$ ein KKT-Punkt, $\hat{x}$ ein lokales Minimum des Standard-Optimierungsproblems und die Gradienten $\nabla g_i(\hat{x})$, $i \in A(\hat{x})$ und $\nabla h_j(\hat{x})$, $j = 1, \ldots, p$ seien linear unabhängig (LICQ). Dann gilt

$$d^\top \nabla^2_{xx} L(\hat{x}, l_0 = 1, \hat{\lambda}, \hat{\mu}) d \ge 0 \qquad \forall d \in T_K(\hat{x})$$

(Die Hessematrix der Lagrange-Funktion ist positiv semidefinit auf dem kritischen Kegel).

Beweis: Es sei $d \in T_K(\hat{x})$, $d \neq 0$. Definiere

$$\begin{aligned} A_0(\hat{x}) &:= \{i \in A(\hat{x}) \mid \hat{\lambda}_i = 0\}, \\ A_>(\hat{x}) &:= \{i \in A(\hat{x}) \mid \hat{\lambda}_i > 0\}, \\ A_0^<(\hat{x}) &:= \{i \in A_0(\hat{x}) \mid g_i'(\hat{x})(d) < 0\}, \\ A_0^=(\hat{x}) &:= \{i \in A_0(\hat{x}) \mid g_i'(\hat{x})(d) = 0\}. \end{aligned}$$

Da die Vektoren $\nabla g_i(\hat{x})$, $i \in A_>(\hat{x}) \cup A_0^=(\hat{x})$ und $\nabla h_j(\hat{x})$, $j = 1, \ldots, p$ linear unabhängig sind, kann man analog zum Beweis der Fritz-John-Bedingungen zeigen, daß es ein $\varepsilon > 0$ und eine zweimal stetig differenzierbare Kurve $x : (-\varepsilon, \varepsilon) \to \mathbb{R}^n$ gibt mit $x(0) = \hat{x}$, $x'(0) = d$, $g_i(x(t)) = 0$, $i \in A_>(\hat{x}) \cup A_0^=(\hat{x})$, $h_j(x(t)) = 0$, $j = 1, \ldots, p$ für alle $t \in (-\varepsilon, \varepsilon)$ und $x(t) \in \Sigma$ für alle $t \in [0, \varepsilon)$. Sei

$$\varphi(t) := L(x(t), l_0 = 1, \hat{\lambda}, \hat{\mu}), \qquad t \in (-\varepsilon, \varepsilon).$$

φ ist zweimal stetig differenzierbar mit

$$\varphi'(t) = \nabla_x L(x(t), l_0 = 1, \hat{\lambda}, \hat{\mu})^\top x'(t)$$

und

$$\varphi''(t) = \nabla_x L(x(t), l_0 = 1, \hat{\lambda}, \hat{\mu})^\top x''(t) + x'(t)^\top \nabla_{xx}^2 L(x(t), l_0 = 1, \hat{\lambda}, \hat{\mu}) x'(t).$$

Da $(\hat{x}, \hat{\lambda}, \hat{\mu})$ KKT-Punkt ist, folgt

$$\varphi'(0) = \nabla_x L(\hat{x}, l_0 = 1, \hat{\lambda}, \hat{\mu})^\top d = 0$$

und

$$\varphi''(0) = d^\top \nabla_{xx}^2 L(\hat{x}, l_0 = 1, \hat{\lambda}, \hat{\mu}) d.$$

Angenommen, es gilt $\varphi''(0) < 0$. Dann folgt aus der Stetigkeit von φ'', daß $\varphi''(t) < 0$ für alle hinreichend kleinen $t \in (-\varepsilon, \varepsilon)$ gilt. Taylorentwicklung von φ in $t = 0$ liefert

$$\varphi(t) = \varphi(0) + t\varphi'(0) + \frac{t^2}{2}\varphi''(\xi_t)$$

für alle $t \in (-\varepsilon, \varepsilon)$ mit ξ_t zwischen 0 und t. Aus $\varphi'(0) = 0$ und $\varphi''(\xi_t) < 0$ für alle hinreichend kleinen $t \in (-\varepsilon, \varepsilon)$ erhalten wir $\varphi(t) < \varphi(0)$. Wegen

$$\varphi(0) = L(\hat{x}, l_0 = 1, \hat{\lambda}, \hat{\mu}) = f(\hat{x})$$

und

$$\varphi(t) = L(x(t), l_0 = 1, \hat{\lambda}, \hat{\mu}) = f(x(t))$$

(um dies zu sehen werden die Eigenschaften von $x(t)$ und die Tatsache, daß $(\hat{x}, \hat{\lambda}, \hat{\mu})$ KKT-Punkt ist, ausgenutzt), folgt $f(x(t)) < f(\hat{x})$ für alle hinreichend kleinen $t \in [0, \varepsilon)$. Wegen $x(t) \in \Sigma$ widerspricht dies der Minimalität von $\hat{x}$. □

5.3 Hinreichende Bedingungen

Um entscheiden zu können, ob ein Punkt, der die notwendigen Bedingungen erfüllt, tatsächlich optimal ist, werden hinreichende Bedingungen benötigt. Eine hinreichende Bedingung zweiter Ordnung ist durch den folgenden Satz gegeben.

Satz 5.3.1 (Hinreichende Bedingung zweiter Ordnung)
Seien f, g_i, $i = 1, \dots, m$ und h_j, $j = 1, \dots, p$ zweimal stetig differenzierbar und $S = \mathbb{R}^n$. Weiter sei $(\hat{x}, \hat{\lambda}, \hat{\mu})$ ein KKT-Punkt des Standard-Optimierungsproblems mit

$$d^\top \nabla^2_{xx} L(\hat{x}, l_0 = 1, \hat{\lambda}, \hat{\mu}) d > 0 \qquad \forall d \in T_K(\hat{x}),\ d \neq 0. \tag{5.14}$$

Dann existiert eine Umgebung U von $\hat{x}$ und ein $\alpha > 0$ mit

$$f(x) \geq f(\hat{x}) + \alpha \|x - \hat{x}\|^2 \qquad \forall x \in \Sigma \cap U$$

(insbesondere ist $\hat{x}$ also lokales Minimum und f wächst lokal mindestens quadratisch).

Beweis:

(a) Sei $d \in T(\Sigma, \hat{x})$, $d \neq 0$. Dann existieren Folgen $x_k \in \Sigma$, $x_k \to \hat{x}$ und $\alpha_k \downarrow 0$ mit

$$\lim_{k\to\infty} \frac{x_k - \hat{x}}{\alpha_k} = d.$$

Für $i \in A(\hat{x})$ folgt mit Hilfe des Mittelwertsatzes

$$0 \geq \frac{g_i(x_k) - g_i(\hat{x})}{\alpha_k} = \nabla g_i(\xi_k)^\top \left(\frac{x_k - \hat{x}}{\alpha_k} \right) \overset{k\to\infty}{\longrightarrow} \nabla g_i(\hat{x})^\top d,$$

wobei ξ_k zwischen $\hat{x}$ und x_k liegt. Analog zeigt man $\nabla h_j(\hat{x})^\top d = 0$ für $j = 1, \dots, p$. Da $(\hat{x}, \hat{\lambda}, \hat{\mu})$ ein KKT-Punkt ist, gilt

$$\nabla f(\hat{x})^\top d = -\sum_{i=1}^{m} \underbrace{\hat{\lambda}_i}_{\geq 0} \underbrace{\nabla g_i(\hat{x})^\top d}_{\leq 0} - \sum_{j=1}^{p} \hat{\mu}_j \underbrace{\nabla h_j(\hat{x})^\top d}_{=0} \geq 0.$$

Also erfüllt $\hat{x}$ die notwendige Bedingung $\nabla f(\hat{x})^\top d \geq 0$ für alle $d \in T(\Sigma, \hat{x})$.

(b) Angenommen, die Aussage des Satzes ist falsch. Dann gibt es zu jeder Kugelumgebung um $\hat{x}$ mit Radius $1/i$ einen Punkt $x_i \in \Sigma$ mit

$$f(x_i) - f(\hat{x}) < \frac{1}{i} \|x_i - \hat{x}\|^2 , \quad \|x_i - \hat{x}\| \leq \frac{1}{i} \qquad \forall i \in \mathbb{N}. \tag{5.15}$$

Da die Einheitskugel bzgl. $\|\cdot\|$ in $\mathbb{R}^n$ kompakt ist, existiert eine konvergente Teilfolge $\{x_{i_k}\}$ von $\{x_i\}$ mit

$$\lim_{k\to\infty} \frac{x_{i_k} - \hat{x}}{\|x_{i_k} - \hat{x}\|} = \hat{d}, \quad \lim_{k\to\infty} \|x_{i_k} - \hat{x}\| = 0.$$

Also gilt $\hat{d} \in T(\Sigma, \hat{x}) \setminus \{0\}$. Grenzübergang in (5.15) liefert

$$\nabla f(\hat{x})^\top \hat{d} = \lim_{k\to\infty} \frac{f(x_{i_k}) - f(\hat{x})}{\|x_{i_k} - \hat{x}\|} \leq 0.$$

Zusammen mit (a) folgt

$$\nabla f(\hat{x})^\top \hat{d} = 0.$$

(c) Da $\hat{x}$ ein KKT-Punkt ist, folgt

$$\nabla f(\hat{x})^\top \hat{d} = -\sum_{i\in A(\hat{x})} \underbrace{\hat{\lambda}_i}_{\geq 0} \underbrace{\nabla g_i(\hat{x})^\top \hat{d}}_{\leq 0} - \sum_{j=1}^{p} \hat{\mu}_j \underbrace{\nabla h_j(\hat{x})^\top \hat{d}}_{=0} = 0.$$

Also ist $\nabla g_i(\hat{x})^\top \hat{d} = 0$, falls $\hat{\lambda}_i > 0$. Dies zeigt, daß $\hat{d} \in T_K(\hat{x})$.

Gemäß (5.15) gilt

$$\lim_{k\to\infty} \frac{f(x_{i_k}) - f(\hat{x})}{\|x_{i_k} - \hat{x}\|^2} \leq \lim_{k\to\infty} \frac{1}{i_k} = 0 \tag{5.16}$$

für die Richtung $\hat{d}$. Weiter gilt

$$\begin{aligned}
L(x_{i_k}, l_0 = 1, \hat{\lambda}, \hat{\mu}) &= f(x_{i_k}) + \sum_{i=1}^{m} \underbrace{\hat{\lambda}_i}_{\geq 0} \underbrace{g_i(x_{i_k})}_{\leq 0} + \sum_{j=1}^{p} \hat{\mu}_j \underbrace{h_j(x_{i_k})}_{=0} \leq f(x_{i_k}) \ , \\
L(\hat{x}, l_0 = 1, \hat{\lambda}, \hat{\mu}) &= f(\hat{x}) + \sum_{i=1}^{m} \underbrace{\hat{\lambda}_i g_i(\hat{x})}_{=0} + \sum_{j=1}^{p} \hat{\mu}_j \underbrace{h_j(\hat{x})}_{=0} = f(\hat{x}) \ , \\
\nabla_x L(\hat{x}, l_0 = 1, \hat{\lambda}, \hat{\mu}) &= \nabla f(\hat{x}) + \sum_{i=1}^{m} \hat{\lambda}_i \nabla g_i(\hat{x}) + \sum_{j=1}^{p} \hat{\mu}_j \nabla h_j(\hat{x}) = 0.
\end{aligned}$$

Taylorentwicklung von L bzgl. x in $\hat{x}$ liefert

$$\begin{aligned}
f(x_{i_k}) \geq L(x_{i_k}, l_0 = 1, \hat{\lambda}, \hat{\mu}) &= L(\hat{x}, l_0 = 1, \hat{\lambda}, \hat{\mu}) + \nabla_x L(\hat{x}, l_0 = 1, \hat{\lambda}, \hat{\mu})^\top (x_{i_k} - \hat{x}) \\
&\quad + \frac{1}{2}(x_{i_k} - \hat{x})^\top \nabla^2_{xx} L(\xi_k, l_0 = 1, \hat{\lambda}, \hat{\mu})(x_{i_k} - \hat{x}) \\
&= f(\hat{x}) + \frac{1}{2}(x_{i_k} - \hat{x})^\top \nabla^2_{xx} L(\xi_k, l_0 = 1, \hat{\lambda}, \hat{\mu})(x_{i_k} - \hat{x}),
\end{aligned}$$

wobei ξ_k zwischen $\hat{x}$ und x_{i_k} liegt. Division durch $\|x_{i_k} - \hat{x}\|^2$ und Grenzübergang liefert zusammen mit (5.16) die Beziehung

$$0 \geq \frac{1}{2}\hat{d}^\top \nabla^2_{xx} L(\hat{x}, l_0 = 1, \hat{\lambda}, \hat{\mu})\hat{d}.$$

Dies widerspricht der Voraussetzung, daß $d^\top \nabla^2_{xx} L(\hat{x}, \hat{\lambda}, \hat{\mu}) d > 0$ für alle $d \in T_K(\hat{x})$ gilt.

□

Der Begriff „kritischer Kegel“ läßt sich wie folgt begründen. Sei $d \in T_K(\hat{x})$. Aus den KKT-Bedingungen folgt dann

$$0 = \nabla f(\hat{x})^\top d + \sum_{i \in A(\hat{x})} \lambda_i \nabla g_i(\hat{x})^\top d + \sum_{j=1}^{p} \mu_j \nabla h_j(\hat{x})^\top d \overset{d \in T_K(\hat{x})}{=} \nabla f(\hat{x})^\top d.$$

D.h. für die kritischen Richtungen $d \in T_K(\hat{x})$ verschwindet die Richtungsableitung $\nabla f(\hat{x})^\top d$. In diesem „kritischen Fall“ kann man nichts über das lokale Verhalten von f aussagen und man benötigt Zusatzforderungen an die Krümmung (2. Ableitung!), um die lokale Optimalität sicherzustellen. Wäre $\nabla f(\hat{x})^\top d \neq 0$, so folgte, daß d entweder Auf- oder Abstiegsrichtung von f in $\hat{x}$ ist.
Generell sind bei diesen Betrachtungen Richtungen mit $\nabla g_i(\hat{x})^\top d > 0$ für ein $i \in A(\hat{x})$ oder $\nabla h_j(\hat{x})^\top d \neq 0$ für ein $1 \leq j \leq p$ uninteressant, da der zulässige Bereich entlang dieser Richtungen sofort verlassen wird. Daher ist es auch unerheblich, was die Zielfunktion entlang solcher Richtungen macht. Es sind also nur Richtungen aus dem linearisierenden Kegel $T_{lin}(\hat{x})$ von Interesse. Für die Richtungen $d \in T_{lin}(\hat{x}) \setminus T_K(\hat{x})$ gilt

$$\begin{aligned}
\nabla g_i(\hat{x})^\top d &\leq 0, \qquad i \in A(\hat{x}), \lambda_i = 0,\\
\nabla g_i(\hat{x})^\top d &< 0, \qquad i \in A(\hat{x}), \lambda_i > 0,\\
\nabla h_j(\hat{x})^\top d &= 0, \qquad j = 1, \ldots, p.
\end{aligned}$$

Aus den KKT-Bedingungen folgt hierfür

$$0 = \nabla f(\hat{x})^\top d + \sum_{i \in A(\hat{x}), \lambda_i > 0} \underbrace{\lambda_i}_{>0} \underbrace{\nabla g_i(\hat{x})^\top d}_{<0} \qquad \Rightarrow \qquad \nabla f(\hat{x})^\top d > 0,$$

d.h. für diese Richtungen wächst f lokal. Man benötigt hierfür also keine Extraforderungen.

5.4 Sensitivität und parametrische Optimierung

Häufig hängen Optimierungsprobleme von Parametern ab, die selbst **keine** Optimierungsvariablen sind. Es stellt sich dann die Frage, wie die Lösung des Optimierungsproblems von den Parametern abhängt bzw. wie sensitiv diese Abhängigkeit ist. Wir werden sehen, daß die Lösung unter geeigneten Voraussetzungen (u.a. hinreichende Bedingung zweiter Ordnung, LICQ) sogar stetig differenzierbar von den Parametern abhängt. Dieses Resultat wird auch bei den später zu diskutierenden SQP Verfahren von Bedeutung sein.
Wir untersuchen parametrische Optimierungsprobleme der folgenden Form.

Parametrisches Optimierungsproblem $(NLP(w))$**:**
Für einen gegebenen Parameter $w \in \mathbb{R}^q$ finde $x \in \mathbb{R}^n$, so daß $f(x,w)$ minimal wird unter den Nebenbedingungen

$$\begin{aligned} g_i(x,w) &\leq 0, \quad i = 1,\dots,m, \\ h_j(x,w) &= 0, \quad j = 1,\dots,p. \end{aligned}$$

Sei $\hat{w}$ ein fester Parameter, genannt **Nominalparameter**. Wir interessieren uns für das Verhalten der optimalen Lösung $\hat{x}(w)$ als Funktion von w in einer Umgebung des Nominalparameters $\hat{w}$.
Der zulässige Bereich des parametrischen Optimierungsproblems hängt von w ab und lautet

$$\Sigma(w) := \{x \in \mathbb{R}^n \mid g_i(x,w) \leq 0,\ i = 1,\dots,m,\ h_j(x,w) = 0,\ j = 1,\dots,p\}.$$

Ebenso ist auch die Indexmenge der aktiven Ungleichungsrestriktionen von w abhängig:

$$A(x,w) := \{i \mid g_i(x,w) = 0,\ 1 \leq i \leq m\}.$$

Die Lagrangefunktion lautet entsprechend

$$L(x, l_0, \lambda, \mu, w) = l_0 f(x,w) + \lambda^\top g(x,w) + \mu^\top h(x,w).$$

Definition 5.4.1 **(Streng reguläre Lösung)**
Ein lokales Minimum $\hat{x}$ von $(NLP(w))$ heißt **streng regulär** *wenn die folgenden Bedingungen gelten:*

- *$\hat{x}$ ist zulässig, d.h. $\hat{x} \in \Sigma(w)$.*
- *In $\hat{x}$ ist die Regularitätsbedingung der linearen Unabhängigkeit (LICQ) erfüllt, d.h. die Gradienten $\nabla_x g_i(\hat{x},w)$, $i \in A(\hat{x},w)$, $\nabla_x h_j(\hat{x},w)$, $j = 1,\dots,p$ sind linear unabhängig.*
- *Die KKT-Bedingungen gelten in $(\hat{x}, \hat{\lambda}, \hat{\mu})$ (also insbesondere $l_0 = 1$ in der Lagrangefunktion).*
- *Die* **strikte Komplementaritätsbedingung**

  $$\hat{\lambda}_i - g_i(\hat{x},w) > 0 \qquad \forall i = 1,\dots,m$$

  ist erfüllt.
- *Die hinreichende Bedingung zweiter Ordnung (5.14) ist erfüllt.*

Beachte: Die strikte Komplementaritätsbedingung schließt den Fall aus, daß $\hat{\lambda}_i = 0$ und $g_i(\hat{x}, w) = 0$ **gleichzeitig** gelten. Wegen $\lambda_i g_i(\hat{x}, w) = 0$ gilt also entweder $\lambda_i = 0, g_i(\hat{x}, w) < 0$ oder $\lambda_i > 0, g_i(\hat{x}, w) = 0$.
Das folgende sehr wichtige Resultat findet sich in den Büchern [Fia83], [FM90], [Spe93].

Satz 5.4.2 **(Sensitivitätssatz)**
Seien $f, g_1, \ldots, g_m, h_1, \ldots, h_p : \mathbb{R}^n \times \mathbb{R}^q \to \mathbb{R}$ zweimal stetig differenzierbar und $\hat{w}$ Nominalparameter. Sei $\hat{x}$ ein streng reguläres lokales Minimum von $(NLP(\hat{w}))$, wobei $\hat{\lambda}$, $\hat{\mu}$ die entsprechenden Lagrange-Multiplikatoren bezeichnen. Dann existieren Umgebungen $V_\varepsilon(\hat{w})$ und $U_\delta(\hat{x}, \hat{\lambda}, \hat{\mu})$, so daß $(NLP(w))$ ein eindeutiges, streng reguläres lokales Minimum

$$(x(w), \lambda(w), \mu(w)) \in U_\delta(\hat{x}, \hat{\lambda}, \hat{\mu})$$

für alle $w \in V_\varepsilon(\hat{w})$ besitzt. Die Indexmenge der aktiven Ungleichungen ändert sich in dieser Umgebung nicht, d.h. es gilt $A(\hat{x}, \hat{w}) = A(x(w), w)$ für alle $w \in V_\varepsilon(\hat{w})$.
Zusätzlich ist $(x(w), \lambda(w), \mu(w))$ stetig differenzierbar bzgl. w mit

$$\begin{pmatrix} \frac{dx}{dw}(\hat{w}) \\ \frac{d\lambda}{dw}(\hat{w}) \\ \frac{d\mu}{dw}(\hat{w}) \end{pmatrix} = - \begin{pmatrix} \nabla^2_{xx} L & (g'_x)^\top & (h'_x)^\top \\ \hat{\Lambda} \cdot g'_x & \hat{\Gamma} & 0 \\ h'_x & 0 & 0 \end{pmatrix}^{-1} \cdot \begin{pmatrix} \nabla^2_{xw} L \\ \hat{\Lambda} \cdot g'_w \\ h'_w \end{pmatrix}, \tag{5.17}$$

wobei $\hat{\Lambda} = \operatorname{diag}(\hat{\lambda}_1, \ldots, \hat{\lambda}_m)$, $\hat{\Gamma} = \operatorname{diag}(g_1, \ldots, g_m)$ ist. Alle Funktionen und Ableitungen sind in $(\hat{x}, l_0 = 1, \hat{\lambda}, \hat{\mu}, \hat{w})$ ausgewertet.

Beweis: Der Beweis basiert auf dem Satz über implizite Funktionen. Betrachte die nichtlineare Gleichung (KKT-Bedingungen!)

$$F(x, \lambda, \mu, w) := \begin{pmatrix} \nabla_x L(x, l_0 = 1, \lambda, \mu, w) \\ \Lambda \cdot g(x, w) \\ h(x, w) \end{pmatrix} = 0, \tag{5.18}$$

wobei $\Lambda := \operatorname{diag}(\lambda_1, \ldots, \lambda_m)$ ist. F ist stetig differenzierbar und es gilt

$$F(\hat{x}, \hat{\lambda}, \hat{\mu}, \hat{w}) = 0.$$

Um den Satz über implizite Funktionen anwenden zu können, müssen wir die Invertierbarkeit der Jacobimatrix

$$F'_{(x,\lambda,\mu)}(\hat{x}, \hat{\lambda}, \hat{\mu}, \hat{w}) = \begin{pmatrix} \nabla^2_{xx} L(\hat{x}, l_0 = 1, \hat{\lambda}, \hat{\mu}, \hat{w}) & (g'_x(\hat{x}, \hat{w}))^\top & (h'_x(\hat{x}, \hat{w}))^\top \\ \hat{\Lambda} \cdot g'_x(\hat{x}, \hat{w}) & \hat{\Gamma} & 0 \\ h'_x(\hat{x}, \hat{w}) & 0 & 0 \end{pmatrix}$$

zeigen ($\rightarrow$ Übungsaufgabe).
Gemäß des Satzes über implizite Funktionen existieren Umgebungen $V_\varepsilon(\hat{w})$ und $U_\delta(\hat{x},\hat{\lambda},\hat{\mu})$ und eindeutig definierte Funktionen

$$(x(\cdot),\lambda(\cdot),\mu(\cdot)) : V_\varepsilon(\hat{w}) \to U_\delta(\hat{x},\hat{\lambda},\hat{\mu})$$

mit

$$F(x(w),\lambda(w),\mu(w),w) = 0 \tag{5.19}$$

für alle $w \in V_\varepsilon(\hat{w})$. Darüber hinaus sind dises Funktionen stetig differenzierbar und (5.17) entsteht durch Differentiation der Identität (5.19) nach w.
Es bleibt zu zeigen, daß $x(w)$ tatsächlich ein streng reguläres lokales Minimum von $(NLP(w))$ ist. O.B.d.A. sei $A(\hat{x},\hat{w}) = \{\ell+1,\dots,m\}$. Die Stetigkeit von $x(w),\lambda(w)$ und g zusammen mit $\lambda_i(\hat{w}) = \hat{\lambda}_i > 0$, $i = \ell+1,\dots,m$ und $g_i(x(\hat{w}),\hat{w}) = g_i(\hat{x},\hat{w}) < 0$, $i = 1,\dots,\ell$ garantiert $\lambda_i(w) > 0$, $i = \ell+1,\dots,m$ und $g_i(x(w),w) < 0$, $i = 1,\dots,\ell$ für w hinreichend nahe bei $\hat{w}$.
Aus (5.19) folgt $g_i(x(w),w) = 0$, $i = \ell+1,\dots,m$ und $h_j(x(w),w) = 0$, $j = 1,\dots,p$. Also: $x(w) \in \Sigma(w)$ und die KKT-Bedingungen sind erfüllt. Insbesondere bleibt auch die Indexmenge $A(x(w),w) = A(\hat{x},\hat{w})$ unverändert in einer Umgebung von $\hat{w}$.
Schließlich müssen wir noch zeigen, daß $\nabla^2_{xx}L(x(w),l_0 = 1,\lambda(w),\mu(w),w)$ für w hinreichend nahe bei $\hat{w}$ positiv definit bleibt auf dem kritischen Kegel $T_K(x(w))$. Beachte, daß der kritische Kegel $T_K(x(w))$ mit w variiert. Bisher wissen wir nur, daß $\nabla^2_{xx}L(\hat{w}) := \nabla^2_{xx}L(x(\hat{w}),l_0 = 1,\lambda(\hat{w}),\mu(\hat{w}),\hat{w})$ positiv definit auf $T_K(x(\hat{w}))$ ist, was äquivalent ist zur Existenz von $\alpha > 0$ mit $d^\top \nabla^2_{xx}L(\hat{w})d \geq \alpha\|d\|^2$ für alle $d \in T_K(x(\hat{w}))$. Auf Grund der strikten Komplementarität in einer Umgebung von $\hat{w}$ gilt

$$T_K(x(w)) = \left\{ d \in \mathbb{R}^{n_x} \;\middle|\; \begin{array}{rcl} g_i'(x(w),w)(d) &=& 0,\ i \in A(\hat{x},\hat{w}), \\ h_j'(x(w),w)(d) &=& 0,\ j = 1,\dots,m \end{array} \right\}$$

nahe bei $\hat{w}$. Angenommen, für jedes $i \in \mathbb{N}$ gibt es ein $w^i \in \mathbb{R}^q$ mit $\|w^i - \hat{w}\| \leq \frac{1}{i}$, so daß es für alle $j \in \mathbb{N}$ ein $d^{i_j} \in T_K(x(w^i))$, $d^{i_j} \neq 0$ gibt mit

$$(d^{i_j})^\top \nabla^2_{xx}L(w^{i_j})d^{i_j} < \frac{1}{j}\|d^{i_j}\|^2.$$

Da die Einheitskugel bzgl. $\|\cdot\|$ kompakt ist in $\mathbb{R}^n$, gibt es eine konvergente Teilfolge $\{w^{i_{j_k}}\}$ mit $\lim_{k\to\infty} w^{i_{j_k}} = \hat{w}$ und

$$\lim_{k\to\infty} \frac{d^{i_{j_k}}}{\|d^{i_{j_k}}\|} = \hat{d}, \quad \|\hat{d}\| = 1, \quad \hat{d} \in T_K(x(\hat{w}))$$

und

$$\hat{d}^\top \nabla^2_{xx}L(\hat{w})\hat{d} \leq 0.$$

Dies widerspricht der positiven Definitheit von $\nabla^2_{xx}L(\hat{w})$. □

Bemerkung 5.4.3

- *Gemäß (5.17) können die* **Sensitivitätsableitungen** $\frac{dx}{dw}(\hat{w})$, $\frac{d\lambda}{dw}(\hat{w})$ *und* $\frac{d\mu}{dw}(\hat{w})$ *durch Lösen eines* **linearen Gleichungssystems** *numerisch berechnet werden. Allerdings werden hierzu die zweiten Ableitungen* $\nabla^2_{xx}L$ *und* $\nabla^2_{xw}L$ *benötigt, die häufig nur numerisch berechnet werden können.*

- *Die Größenordnung der Sensitivitätsableitungen erlaubt eine quantitative Aussage darüber, wie stark die Lösung von den Komponenten des Parametervektors* $w \in \mathbb{R}^q$ *abhängt.*

- *Die Sensitivitätsableitungen können auch zur Berechnung von optimalen Lösungen in Echtzeit verwendet werden, vgl. etwa Büskens und Maurer [BM01] und Büskens und Gerdts [BG01].*

- *Leider macht der Sensitivitätssatz keine Aussagen darüber, wie groß die Umgebung* $V_\varepsilon(\hat{w})$ *ist.*

  Problem: Für zu große Abweichungen w *vom Nominalwert* $\hat{w}$ *wird sich im allgemeinen die Indexmenge* $A(\hat{x}, w)$ *ändern. Für diese Situation liefert der Satz keine Aussage.*

- *Die hinreichende Bedingung zweiter Ordnung (5.14) bei gleichzeitiger Gültigkeit der strikten Komplementaritätsbedingung besagt, daß die Hessematrix der Lagrangefunktion positiv definit auf dem Kern der aktiven Nebenbedingungen ist:*

$$d^\top \nabla^2_{xx} L(\hat{x}, l_0 = 1, \hat{\lambda}, \hat{\mu}, \hat{w}) d > 0 \qquad \forall 0 \neq d \in \mathbb{R}^n \ : \ Ad = 0,$$

  wobei

$$A := \begin{pmatrix} \nabla g_i(\hat{x}, \hat{w})^\top, \ i \in A(\hat{x}, \hat{w}) \\ \nabla h_j(\hat{x}, \hat{w})^\top, \ j = 1, \ldots, p \end{pmatrix} \in \mathbb{R}^{s \times n}, \ s = |A(\hat{x}, \hat{w})| + p.$$

  Sie kann numerisch mit Hilfe einer QR-Zerlegung von $A^\top \in \mathbb{R}^{n \times s}$ *überprüft werden:*

$$A^\top = Q \begin{pmatrix} R \\ 0 \end{pmatrix}, \qquad Q = (Y, Z) \in \mathbb{R}^{n \times n}, \ Y \in \mathbb{R}^{n \times s}, \ Z \in \mathbb{R}^{n \times (n-s)}, \ R \in \mathbb{R}^{s \times s}.$$

  R ist regulär, da die Spalten von $A^\top$ linear unabhängig sind, Q ist orthogonal, Y ist eine Orthonormalbasis des Bildes von $A^\top$, Z ist eine Orthonormalbasis von $Bild(A^\top)^\perp = ker(A)$.

  Jedes $d \in ker(A) = \{d \in \mathbb{R}^n \mid Ad = 0\}$ kann also eindeutig dargestellt werden als $d = Zd_Z$ mit einem $d_Z \in \mathbb{R}^{n-s}$.

Damit ist die Aufgabe, die positive Definitheit von $\nabla^2_{xx} L$ auf $\ker(A)$ zu überprüfen, äquivalent zur Überprüfung der **reduzierten Hessematrix**

$$Z^\top \nabla^2_{xx} L(\hat{x}, l_0 = 1, \hat{\lambda}, \hat{\mu}, \hat{w}) Z \in \mathbb{R}^{(n-s)\times(n-s)}$$

auf positive Definitheit.

5.5 Dualität

Wir betrachten das **primale Optimierungsproblem (Primalproblem)**

$$\begin{array}{ll} \text{Minimiere} & f(x) \\ \text{unter} & x \in S, \\ & g_i(x) \le 0, \quad i = 1, \dots, m, \\ & h_j(x) = 0, \quad j = 1, \dots, p. \end{array}$$

Darin sei $S \subseteq \mathbb{R}^n$ eine nichtleere Menge. Wir wollen das sogenannte duale Problem herleiten bzw. motivieren. Dazu betten wir das primale Optimierungsproblem in eine Schar von **gestörten primalen Optimierungsproblemen** ein, indem wir die Ungleichungs- und Gleichungsnebenbedingungen stören:

$$\begin{array}{ll} \text{Minimiere} & f(x) \\ \text{unter} & x \in S, \\ & g_i(x) \le y_i, \quad i = 1, \dots, m, \\ & h_j(x) = z_j, \quad j = 1, \dots, p. \end{array}$$

Hierbei seien $y = (y_1, \dots, y_m)^\top \in \mathbb{R}^m$ und $z = (z_1, \dots, z_p)^\top \in \mathbb{R}^p$ beliebige Störungsvektoren. Das Ausgangsproblem ergibt sich für $y = 0$ und $z = 0$.

Definition 5.5.1 **(Minimalwertfunktion)**

Die Funktion

$$\Phi : \mathbb{R}^{m+p} \to \bar{\mathbb{R}} := \mathbb{R} \cup \{\infty, -\infty\}$$

mit

$$\Phi(y, z) := \inf\{f(x) \mid g_i(x) \le y_i,\ i = 1, \dots, m,\ h_j(x) = z_j,\ j = 1, \dots, p,\ x \in S\}$$

heißt **Minimalwertfunktion** *für das gestörte Problem.*

Das duale Problem ist motiviert durch die Aufgabe, den Graphen der Minimalwertfunk-

tion Φ von unten durch eine Hyperebene

$$\lambda^\top y + \mu^\top z + r = \gamma \tag{5.20}$$

mit Normalenvektor $(\lambda, \mu, 1)^\top \in \mathbb{R}^{m+p+1}$ und den Variablen $(y, z, r)^\top \in \mathbb{R}^{m+p+1}$ abzustützen, vgl. Abbildung 5.4. Der Schnittpunkt dieser Hyperebene mit der r-Achse in $y = 0, z = 0$ ist $(0, 0, \gamma)^\top$.

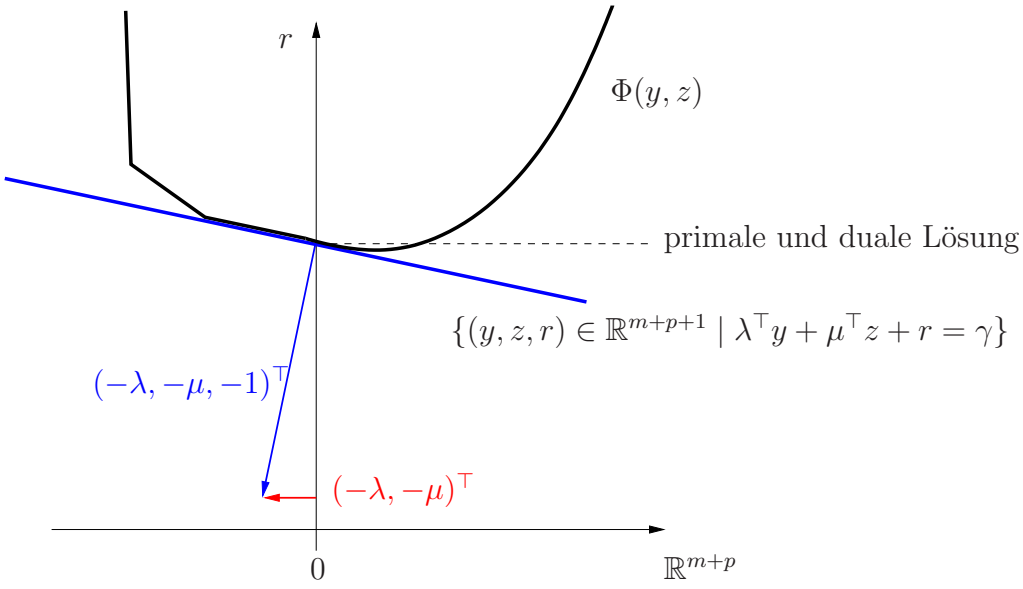


Abbildung 5.4: Grafische Interpretation des Dualproblems: Abstützung des Graphen der Minimalwertfunktion durch eine Hyperbene mit Normalenvektor $(\lambda, \mu, 1)^\top$ bzw. $(-\lambda, -\mu, -1)^\top$.

Das duale Problem läßt sich dann wie folgt formulieren:

Maximiere	γ	
bzgl.	$\lambda \in \mathbb{R}^m, \mu \in \mathbb{R}^p$	
unter	$r \leq \Phi(y, z)$	$\forall y \in \mathbb{R}^m, z \in \mathbb{R}^p.$

wobei r gemäß (5.20) durch $r = \gamma - \lambda^\top y - \mu^\top z$ gegeben ist. Damit lautet das Dualprogramm

Maximiere	γ	
bzgl.	$\lambda \in \mathbb{R}^m, \mu \in \mathbb{R}^p$	
unter	$\gamma - \lambda^\top y - \mu^\top z \leq \Phi(y, z)$	$\forall y \in \mathbb{R}^m, z \in \mathbb{R}^p.$

Mit der Definition von Φ ist dieses Problem wiederum äquivalent zu

Maximiere	γ	
bzgl.	$\lambda \in \mathbb{R}^m, \mu \in \mathbb{R}^p$	
unter	$\gamma \leq f(x) + \lambda^\top y + \mu^\top z$	$\forall y \in \mathbb{R}^m, z \in \mathbb{R}^p,\ x \in S,\ g(x) \leq y,\ h(x) = z.$

Wegen $h(x) = z$ ist dieses Problem wiederum äquivalent mit dem Problem

$$\begin{array}{ll} \text{Maximiere} & \gamma \\ \text{bzgl.} & \lambda \in \mathbb{R}^m, \mu \in \mathbb{R}^p \\ \text{unter} & \gamma \leq f(x) + \lambda^\top y + \mu^\top h(x) \qquad \forall y \in \mathbb{R}^m,\ x \in S,\ g(x) \leq y. \end{array}$$

Nun unterscheiden wir zwei Fälle. Zunächst untersuchen wir die Nebenbedingung

$$\gamma \leq f(x) + \lambda^\top y + \mu^\top h(x) \qquad \forall y \in \mathbb{R}^m,\ x \in S,\ g(x) \leq y \tag{5.21}$$

und nehmen an, daß es eine Komponente $\lambda_i < 0$ gibt. Damit gilt $\lambda_i y_i \to -\infty$ für $y_i \to \infty$. Für $y_i \to \infty$ ist dann auch $g_i(x) \leq y_i$ für $x \in S$ erfüllt. Es folgt, daß (5.21) nur für $\gamma = -\infty$ erfüllt ist. Da γ maximiert werden soll, ist dieser Fall uninteressant. Wir können also die Bedingung $\lambda \in \mathbb{R}^m$ durch $\lambda \geq 0$ ersetzen. In diesem Fall gilt wegen $g(x) \leq y$ stets $\lambda^\top g(x) \leq \lambda^\top y$ und die Bedingung (5.21) kann durch

$$\gamma \leq f(x) + \lambda^\top g(x) + \mu^\top h(x) \qquad \forall x \in S$$

ersetzt werden. Also ist das duale Problem äquivalent mit

$$\begin{array}{ll} \text{Maximiere} & \gamma \\ \text{bzgl.} & \lambda \geq 0, \mu \in \mathbb{R}^p \\ \text{unter} & \gamma \leq f(x) + \lambda^\top g(x) + \mu^\top h(x) \qquad \forall x \in S. \end{array}$$

Mit der Lagrangefunktion

$$L(x, l_0 = 1, \lambda, \mu) = f(x) + \lambda^\top g(x) + \mu^\top h(x)$$

und der **dualen Zielfunktion**

$$\psi(\lambda, \mu) := \inf_{x \in S} L(x, l_0 = 1, \lambda, \mu)$$

lautet das

Dualproblem:

$$\max_{\lambda \geq 0, \mu \in \mathbb{R}^p} \psi(\lambda, \mu) = \max_{\lambda \geq 0, \mu \in \mathbb{R}^p} \inf_{x \in S} L(x, l_0 = 1, \lambda, \mu).$$

Beispiel 5.5.2

- *(vgl. [GK02], S. 316)*

$$\begin{array}{ll} \textit{Minimiere} & f(x_1, x_2) = x_1^2 - x_2^2 \\ \textit{unter} & (x_1, x_2) \in \mathbb{R}^2 \\ & g(x_1, x_2) = x_1^2 + x_2^2 - 1 \leq 0. \end{array}$$

Die Lösung dieses Primalproblems ist $f(0,\pm 1) = -1$. *Die duale Zielfunktion ist*

$$\psi(\lambda) = \inf_{(x_1,x_2)\in\mathbb{R}^2} \{x_1^2 - x_2^2 + \lambda(x_1^2 + x_2^2 - 1)\} = \begin{cases} -\infty, & \text{falls } 0 \le \lambda < 1, \\ -\lambda, & \text{falls } \lambda \ge 1. \end{cases}$$

Die optimale Lösung des Dualproblems $\max_{\lambda\ge 0} \psi(\lambda)$ *ist gegeben durch* $\lambda = 1$ *mit Wert* -1. *Hier besitzen Primal- und Dualproblem den gleichen Zielfunktionswert.*

- *Gegeben sei das primale lineare Optimierungsproblem*

$$\min c^\top x \qquad \text{unter} \qquad Ax = b,\ x \in S = \{x \in \mathbb{R}^n \mid x \ge 0\}.$$

Die duale Zielfunktion lautet

$$\begin{aligned} \psi(\lambda) &= \inf_{x\ge 0} \left(c^\top x + \lambda^\top (b - Ax)\right) \\ &= \inf_{x\ge 0} \left(c^\top - \lambda^\top A\right) x + \lambda^\top b \\ &= \begin{cases} \lambda^\top b, & \text{falls } c^\top - \lambda^\top A \ge 0, \\ -\infty, & \text{sonst.} \end{cases} \end{aligned}$$

Damit lautet das duale Problem

$$\max b^\top \lambda \qquad \text{unter} \qquad A^\top \lambda \le c.$$

Im folgenden werden das primale und das duale Problem zueinander in Beziehung gesetzt.

Satz 5.5.3 (Schwacher Dualitätssatz)
Es sei x *ein primal zulässiger Punkt*[1] *und* (λ, μ) *sei dual zulässiger Punkt*[2]. *Dann gilt für den Minimalwert* $w(P)$ *des Primalproblems und den Maximalwert* $w(D)$ *des Dualproblems die Einschließung*

$$\psi(\lambda,\mu) \le w(D) \le w(P) \le f(x).$$

Beweis: Für alle primal und dual zulässigen Punkte x und λ, μ gilt

$$\psi(\lambda,\mu) = \inf_{x\in S} L(x, l_0 = 1, \lambda, \mu) \le f(x) + \underbrace{\lambda^\top}_{\ge 0} \underbrace{g(x)}_{\le 0} + \mu^\top \underbrace{h(x)}_{=0} \le f(x).$$

Die linke Seite hängt nicht von x ab, während die rechte Seite nicht von λ, μ abhängt. Übergang zum Supremum bzgl. $\lambda \ge 0, \mu$ auf der linken Seite und zum Infimum bzgl. $x \in S, g(x) \le 0, h(x) = 0$ auf der rechten Seite liefert die Behauptung. □

Stimmen der primale und der duale Zielfunktionswert überein, so sind beide Probleme bereits optimal gelöst, denn es gilt:

[1] Ein primal zulässiger Punkt x erfüllt $x \in S$, $g(x) \le 0$ und $h(x) = 0$
[2] (λ, μ) heißt dual zulässig, falls $\lambda \ge 0$ gilt.

Satz 5.5.4 **(Hinreichendes Optimalitätskriterium)**
Es gelte $\psi(\hat\lambda,\hat\mu) = f(\hat x)$, wobei $\hat x$ primal zulässig und $(\hat\lambda,\hat\mu)$ dual zulässig seien. Dann ist $\hat x$ optimal für das Primalproblem und $(\hat\lambda,\hat\mu)$ ist optimal für das Dualproblem.
Außerdem gilt die **Komplementaritätsbedingung**

$$\hat\lambda_i = 0, \ \textit{falls}\ g_i(\hat x) < 0 \qquad (i = 1,\dots,m).$$

Beweis: Die erste Aussage folgt aus dem schwachen Dualitätssatz. Komplementaritätsbedingung: Angenommen es gilt $g_i(\hat x) < 0$ und $\hat\lambda_i > 0$ für mindestens ein $1 \le i \le m$. Dann folgt

$$\psi(\hat\lambda,\hat\mu) = \inf_{x\in S} L(x, l_0 = 1, \hat\lambda, \hat\mu) \le f(\hat x) + \underbrace{\hat\lambda^\top g(\hat x)}_{<0} + \underbrace{\hat\mu^\top h(\hat x)}_{=0} < f(\hat x)$$

im Widerspruch zur Voraussetzung $\psi(\hat\lambda,\hat\mu) = f(\hat x)$. □

Das folgende Beispiel zeigt, daß tatsächlich der Fall $w(D) < w(P)$ eintreten kann. In diesem Fall spricht man von einer **Dualitätslücke**, vgl. Abbildung 5.5.

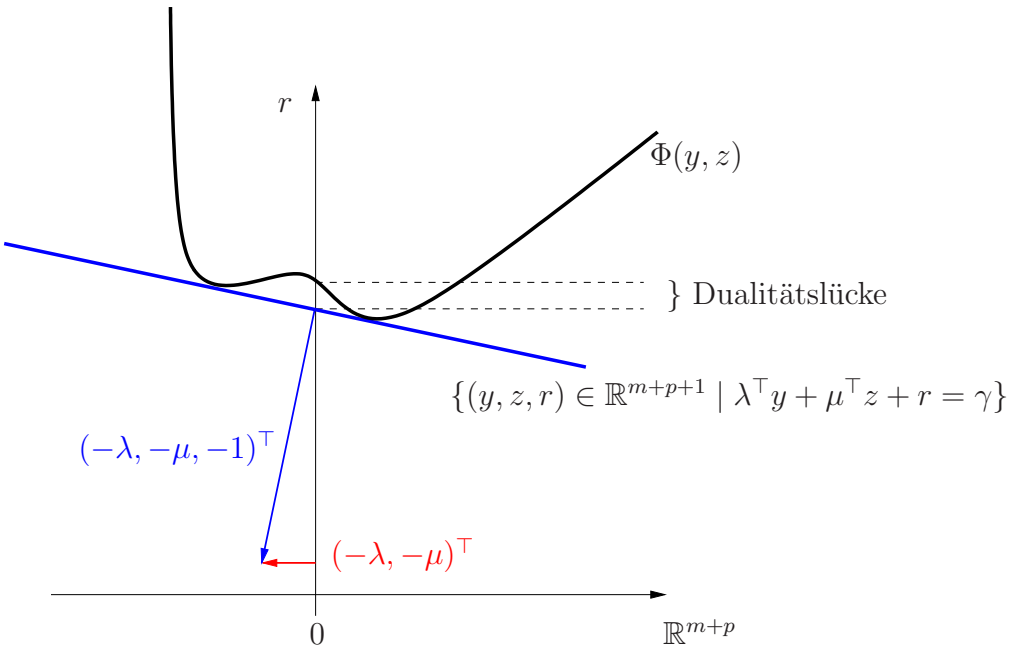


Abbildung 5.5: Grafische Interpretation des Dualproblems: Abstützung des Graphen der Minimalwertfunktion durch eine Hyperbene mit Normalenvektor $(\lambda,\mu,1)^\top$ bzw. $(-\lambda,-\mu,-1)^\top$ und Auftreten einer Dualitätslücke.

Beispiel 5.5.5 **(Dualitätslücke (vgl. [BS79], S. 181))**

$$\begin{array}{ll} \textit{Minimiere} & f(x_1,x_2) = -2x_1 + x_2 \\ \textit{unter} & (x_1,x_2) \in S = \{(0,0),(0,4),(4,4),(4,0),(1,2),(2,1)\}, \\ & 0 = h(x_1,x_2) = x_1 + x_2 - 3. \end{array}$$

Die Lösung dieses Primalproblems ist $(2,1)$ *mit* $f(2,1) = -3$*. Die duale Zielfunktion ist*

$$\psi(\mu) = \min\{-2x_1+3+\mu(x_1+x_2-3) \mid (x_1,x_2) \in S\} = \begin{cases} -4+5\mu, & \textit{falls } \mu \le -1, \\ -8+\mu, & \textit{falls } -1 \le \mu \le 2, \\ -3\mu, & \textit{falls } \mu \ge 2. \end{cases}$$

Die optimale Lösung des Dualproblems $\max_{\mu\in\mathbb{R}} \psi(\mu)$ *ist gegeben durch* $\mu = 2$ *mit Wert* -6*. Wegen* $-6 < -3$ *besteht eine Dualitätslücke.*

Es stellt sich die Frage, unter welchen Bedingungen **keine** Dualitätslücke auftritt. Der folgende Satz liefert hinreichende Bedingungen hierfür.

Satz 5.5.6 **(Starker Dualitätssatz)**

$S \subseteq \mathbb{R}^n$ sei nichtleer und konvex. Die Funktionen f und g_i, $i = 1,\dots,m$ seien konvex. Die Funktionen h_j, $j = 1,\dots,p$ seien affin linear[3]*. Es seien $w(P)$ bzw. $w(D)$ die Zielfunktionswerte des Primal- bzw. Dualproblems. Es sei $w(P)$ endlich und es gebe ein $y \in relint(S)$ mit*

$$\begin{aligned} g_i(y) &< 0, \qquad i = 1,\dots,m, \\ h_j(y) &= 0, \qquad j = 1,\dots,p \end{aligned}$$

*(***Slater Bedingung***). Dann ist das Dualproblem lösbar und es gilt $w(P) = w(D)$.*

Beweis: (siehe [GK02], S. 323)

Seien $h_j(x) = a_j^\top x - \gamma_j$, $j = 1,\dots,p$. Zunächst setzen wir voraus, daß die Vektoren a_j, $j = 1,\dots,p$ linear unabhängig sind und daß das Innere von S nichtleer ist. Definiere

$$Q := \{(y,z,r) \in \mathbb{R}^m \times \mathbb{R}^p \times \mathbb{R} \mid \exists x \in S : g(x) \le y,\ h(x) = z,\ f(x) \le r\}.$$

Es läßt sich zeigen, daß Q konvex und nichtleer ist. Der Punkt $(0,0,w(P))$ ist kein innerer Punkt von Q, denn andernfalls gäbe es ein $\delta > 0$ mit $(0,0,w(P)-\delta) \in Q$ im Widerspruch zur Minimalität von $w(P)$. Nach dem Trennungssatz 4.1.3 gibt es eine trennende Hyperebene, die $(0,0,w(P))$ enthält und Q in einem ihrer abgeschlossenen Halbräume enthält. Mit anderen Worten: Es gibt einen von Null verschiedenen Vektor $(0,0,0) \neq (\hat\lambda,\hat\mu,\hat\gamma) \in \mathbb{R}^m \times \mathbb{R}^p \times \mathbb{R}$ mit

$$\hat\gamma w(P) \le \hat\lambda^\top y + \hat\mu^\top z + \hat\gamma r \qquad \forall (y,z,r) \in Q. \tag{5.22}$$

[3] $h_j(x) = a_j^\top x - \gamma_j$ mit $a_j \in \mathbb{R}^n$ und $\gamma_j \in \mathbb{R}$

Da mit $(y,z,r) \in Q$ auch $(y,z,r+\delta) \in Q$ für alle $\delta \geq 0$ gilt, folgt, daß $\hat{\gamma} \geq 0$ gelten muß, da $w(P)$ endlich ist. Analog folgt auch $\hat{\lambda} \geq 0$.

Zeige: $\hat{\gamma} > 0$.

Annahme: $\hat{\gamma} = 0$. Dann folgt aus (5.22)

$$0 \leq \hat{\lambda}^\top y + \hat{\mu}^\top z \qquad \forall (y,z,r) \in Q$$

bzw.

$$0 \leq \hat{\lambda}^\top g(x) + \hat{\mu}^\top h(x) \qquad \forall x \in S. \tag{5.23}$$

Für das y aus der Slater-Bedingung gilt speziell $g(y) < 0$ und $h(y) = 0$. Damit folgt sofort $\hat{\lambda} = 0$. (5.23) läßt sich damit neu formulieren:

$$0 \leq \hat{\mu}^\top (h(x) - h(y)) = \left(\sum_{j=1}^{p} \hat{\mu}_j a_j \right)^\top (x-y) \qquad \forall x \in S. \tag{5.24}$$

Wir hatten vorausgesetzt, daß das Innere von S nichtleer ist, also sind das Innere von S und $relint(S)$ identisch. Wegen $y \in relint(S)$ ist y ein innerer Punkt von S. Damit gilt $y \pm \varepsilon_k e_k \in S$ für hinreichend kleines $\varepsilon_k > 0$ für alle Einheitsvektoren e_k, $k = 1, \ldots, n$. Aus (5.24) folgt für diese Punkte

$$0 \leq \pm \varepsilon_k \left(\sum_{j=1}^{p} \hat{\mu}_j a_j \right)_k, \qquad k = 1, \ldots, n.$$

Daraus folgt

$$0 = \sum_{j=1}^{p} \hat{\mu}_j a_j$$

und mit der vorausgesetzten linearen Unabhängigkeit der a_j folgt $\hat{\mu} = 0$. Wir haben also insgesamt $(\hat{\lambda}, \hat{\mu}, \hat{\gamma}) = (0,0,0)$, im Widerspruch zu $(\hat{\lambda}, \hat{\mu}, \hat{\gamma}) \neq (0,0,0)$.

Also gilt $\hat{\gamma} > 0$ und o.B.d.A. können wir $\hat{\gamma} = 1$ wählen. Mit $y = g(x), z = h(x), r = f(x)$ folgt aus (5.22)

$$w(P) \leq f(x) + \hat{\lambda}^\top g(x) + \hat{\mu}^\top h(x) \qquad \forall x \in S$$

bzw.

$$w(P) \leq \inf_{x \in S} L(x, l_0 = 1, \hat{\lambda}, \hat{\mu}) = \psi(\hat{\lambda}, \hat{\mu}).$$

Andererseits gilt

$$\psi(\hat{\lambda}, \hat{\mu}) \leq \sup_{\lambda \geq 0, \mu \in \mathbb{R}^p} \psi(\lambda, \mu) = w(D).$$

Beides zusammen ergibt $w(P) \leq w(D)$ und nach dem schwachen Dualitätssatz muß Gleichheit gelten, also $w(P) = w(D)$. Damit ist die Behauptung unter den obigen Zusatzvoraussetzungen gezeigt.

Seien nun die Vektoren a_j, $j = 1, \ldots, p$ linear abhängig und S habe ein nichtleeres Inneres. Aufgrund der Slater-Bedingung hat das Gleichungssystem $a_j^\top x - \gamma_j = 0$, $j = 1, \ldots, p$ zumindest die Lösung y und einige Gleichungen sind linear abhängig. Eliminieren wir die linear abhängigen Gleichungen aus dem Optimierungsproblem, so befinden wir uns wieder im zuerst diskutierten Fall der linear unabhängigen Restriktionen und der obige Beweis funktioniert wie zuvor. Schließlich ergänzen wir die weggelassenen linear abhängigen Gleichungsrestriktionen mit Multiplikator $\mu_j = 0$ in der Lagrangefunktion.

Es bleibt der Fall zu untersuchen, daß das Innere von S leer ist. Dann ist die affine Hülle $aff(S)$ ein affiner Unterraum des $\mathbb{R}^n$ mit Dimension $k < n$ und kann als

$$aff(S) = \{x \in \mathbb{R}^n \mid x = Cu + d,\ u \in \mathbb{R}^k\}$$

mit einer Matrix $C \in \mathbb{R}^{n \times k}$ vom Rang k und einem Vektor $d \in \mathbb{R}^n$ dargestellt werden. Betrachte nun das transformierte Problem $(\bar{P})$

$$\min_u \bar{f}(u) \text{ unter } u \in U,\ \bar{g}(u) \leq 0,\ \bar{h}(u) = 0,$$

wobei

$$\begin{aligned} \bar{f}(u) &:= f(Cu + d), \\ \bar{g}(u) &:= g(Cu + d), \\ \bar{h}(u) &:= h(Cu + d), \\ U &:= \{u \in \mathbb{R}^k \mid Cu + d \in S\}. \end{aligned}$$

Es zeigt sich, daß dieses transformierte Problem die Voraussetzungen des Satzes erfüllt, wobei der durch $y = C\hat{u} + d$ gegebene Vektor $\hat{u} \in \mathbb{R}^k$ die Rolle des Slaterpunktes übernimmt. Da $y \in relint(S)$ gilt, ist $\hat{u}$ im Inneren von U. Das zum transformierten Problem gehörende duale Problem $(\bar{D})$ ist also lösbar und es gilt $w(\bar{P}) = w(\bar{D})$. Ausserdem hat sich der Zielfunktionswert durch die Transformation nicht verändert und es gilt $w(P) = w(\bar{P})$. Für die duale Zielfunktion gilt

$$\begin{aligned} \bar{\psi}(\lambda, \mu) &= \inf_{u \in U} \big(\bar{f}(u) + \lambda^\top \bar{g}(u) + \mu^\top \bar{h}(u)\big) \\ &= \inf_{Cu+d \in S} \big(f(Cu + d) + \lambda^\top g(Cu + d) + \mu^\top h(Cu + d)\big) \\ &= \inf_{x \in S} \big(f(x) + \lambda^\top g(x) + \mu^\top h(x)\big) \\ &= \psi(\lambda, \mu). \end{aligned}$$

Die beiden dualen Zielfunktionen stimmen also miteinander überein. Damit ist alles gezeigt. □

Bemerkung 5.5.7

Der starke Dualitätssatz kann noch weiter abgeschwächt werden, da auf die Gültigkeit der Slaterbedingung für affin-lineare Ungleichungsrestriktionen verzichtet werden kann.

Der Vorteil des Dualproblems besteht darin, daß die duale Zielfunktion $\psi(\lambda, \mu)$ auf ihrer **Domäne (wesentlicher Definitionsbereich)**

$$dom(\psi) := \{(\lambda, \mu)^\top \in \mathbb{R}^{m+p} \mid \lambda \geq 0,\ \psi(\lambda, \mu) > -\infty\}$$

konkav ist, d.h. $-\psi$ ist konvex, siehe [GK02], S. 320. Damit ist das zum Dualprogramm äquivalente Problem $\min_{\lambda \geq 0, \mu \in \mathbb{R}^p} -\psi(\lambda, \mu)$ ein konvexes Problem und kann u.U. leichter gelöst werden als das zugehörige Primalproblem (insbesondere falls die duale Zielfunktion ψ leicht berechenbar ist). Insbesondere ist jede Lösung des dualen Problems bereits ein globales Maximum der dualen Zielfunktion. Kann man zudem noch das Auftreten einer Dualitätslücke ausschliessen, so besteht ein möglicher Ansatz zur Lösung des Primalproblems darin, das duale Problem zu lösen. Gemäß Satz 5.5.4 hätte man dann auch das Primalproblem gelöst. Kann das Auftreten einer Dualitätslücke nicht ausgeschlossen werden, so liefert das Dualproblem gemäß Satz 5.5.3 zumindest eine untere Schranke des optimalen primalen Zielfunktionswerts (dies kann z.B. bei Branch&Bound Verfahren ausgenutzt werden).

5.6 Quadratische Optimierung

Wir widmen uns nun einem Verfahren zur Lösung des wohl einfachsten nichtlinearen Optimierungsproblems – dem quadratischen Optimierungsproblem. Zunächst beschränken wir uns auf den Fall mit Gleichungsrestriktionen.

Quadratisches Optierungsproblem mit Gleichungsbeschränkungen:

Für eine symmetrische Matrix $W \in \mathbb{R}^{n \times n}$, Vektoren $c \in \mathbb{R}^n$, $b_j \in \mathbb{R}^n$, $j = 1, \ldots, p$ und Zahlen $v_j \in \mathbb{R}$, $j = 1, \ldots, p$, minimiere

$$f(x) := \frac{1}{2} x^\top W x + c^\top x$$

unter den Nebenbedingungen

$$h_j(x) := b_j^\top x - v_j \quad = \quad 0, \qquad j = 1, \ldots, p.$$

Mit

$$B := \begin{pmatrix} b_1^\top \\ \vdots \\ b_p^\top \end{pmatrix}, \qquad v = \begin{pmatrix} v_1 \\ \vdots \\ v_p \end{pmatrix}$$

lautet das Problem in Matrixschreibweise

$$\text{Minimiere} \quad \frac{1}{2}x^\top W x + c^\top x \quad \text{unter} \quad Bx = v.$$

Da die Regularitätsbedingung von Abadie für lineare Nebenbedingungen stets erfüllt ist, genügt es, die KKT-Bedingungen zu untersuchen. Sei also $\hat{x}$ lokales Minimum des quadratischen Optimierungsproblems. Dann gibt es Lagrange-Multiplikatoren $\mu \in \mathbb{R}^p$ mit

$$\begin{aligned} W\hat{x} + c + \sum_{j=1}^{p} \mu_j b_j &= 0, \\ b_j^\top \hat{x} - v_j &= 0, \qquad j = 1, \ldots, p \end{aligned}$$

bzw. in Matrixschreibweise

$$\begin{pmatrix} W & B^\top \\ B & 0 \end{pmatrix} \begin{pmatrix} \hat{x} \\ \mu \end{pmatrix} = \begin{pmatrix} -c \\ v \end{pmatrix}. \tag{5.25}$$

Beachte, daß hier wieder die sogenannte KKT-Matrix auftritt. Ist W positiv definit auf dem Kern von B und $\text{Rang}(B) = p$, so ist die Matrix invertierbar ($\rightarrow$ Übung). Ist W positiv semidefinit, so ist f konvex und jede Lösung des Gleichungssystems ist zugleich globale Lösung des quadratischen Optimierungsproblems.

Im Hinblick auf ein später zu diskutierendes iteratives Verfahren formen wir (5.25) um und setzen dazu $\hat{x} = x^{[k]} + d$, wobei $x^{[k]}$ ein beliebiger zulässiger Punkt mit $h(x^{[k]}) = 0$ sei.

Dann ist (5.25) äquivalent mit

$$\begin{pmatrix} W & B^\top \\ B & 0 \end{pmatrix} \begin{pmatrix} x^{[k]} + d \\ \mu \end{pmatrix} = \begin{pmatrix} -c \\ v \end{pmatrix}.$$

Äquivalent gilt

$$\begin{pmatrix} W & B^\top \\ B & 0 \end{pmatrix} \begin{pmatrix} d \\ \mu \end{pmatrix} = \begin{pmatrix} -c \\ v \end{pmatrix} - \begin{pmatrix} W & B^\top \\ B & 0 \end{pmatrix} \begin{pmatrix} x^{[k]} \\ 0 \end{pmatrix} = \begin{pmatrix} -c - Wx^{[k]} \\ v - Bx^{[k]} \end{pmatrix} = \begin{pmatrix} -\nabla f(x^{[k]}) \\ 0 \end{pmatrix}.$$

Diese Betrachtungen zeigen

Satz 5.6.1

Ist $x^{[k]} \in \mathbb{R}^n$ zulässig, so erfüllen $x = x^{[k]} + d$ und $\mu \in \mathbb{R}^p$ die KKT-Bedingungen des gleichungsbeschränkten quadratischen Optimierungsproblems, wenn (d, μ) das lineare Gleichungssystem

$$\begin{pmatrix} W & B^\top \\ B & 0 \end{pmatrix} \begin{pmatrix} d \\ \mu \end{pmatrix} = \begin{pmatrix} -\nabla f(x^{[k]}) \\ 0 \end{pmatrix} \tag{5.26}$$

löst.

Wir lassen nun auch Ungleichungen zu und betrachten allgemeine quadratische Optimierungsprobleme

Quadratisches Optierungsproblem:

Für eine symmetrische Matrix $W \in \mathbb{R}^{n\times n}$, Vektoren $c \in \mathbb{R}^n$, $a_i \in \mathbb{R}^n$, $i = 1, \dots, m$, $b_j \in \mathbb{R}^n$, $j = 1, \dots, p$ und Zahlen u_i, $i = 1, \dots, m$, $v_j \in \mathbb{R}$, $j = 1, \dots, p$, minimiere

$$f(x) := \frac{1}{2} x^\top W x + c^\top x$$

unter den Nebenbedingungen

$$\begin{aligned} g_i(x) := a_i^\top x - u_i &\leq 0, \qquad i \in \mathcal{I} := \{1, \dots, m\}, \\ h_j(x) := b_j^\top x - v_j &= 0, \qquad j \in \mathcal{J} := \{1, \dots, p\}. \end{aligned}$$

Mit

$$A := \begin{pmatrix} a_1^\top \\ \vdots \\ a_m^\top \end{pmatrix}, \qquad u = \begin{pmatrix} u_1 \\ \vdots \\ u_m \end{pmatrix}$$

lautet das Problem in Matrixschreibweise

$$\text{Minimiere} \quad \frac{1}{2} x^\top W x + c^\top x \quad \text{unter} \quad Ax \leq u,\ Bx = v.$$

Auswertung der KKT-Bedingungen in einem lokalen Minimum $\hat{x}$ liefert

$$\begin{aligned} \lambda_i &\geq 0, \qquad i = 1, \dots, m, \\ \lambda_i g_i(\hat{x}) &= 0, \qquad i = 1, \dots, m, \\ W\hat{x} + c + \sum_{i=1}^m \lambda_i a_i + \sum_{j=1}^p \mu_j b_j &= 0, \\ a_i^\top \hat{x} - u_i &\leq 0, \qquad i = 1, \dots, m, \\ b_j^\top \hat{x} - v_j &= 0, \qquad j = 1, \dots, p. \end{aligned}$$

Dieses System von Gleichungen und Ungleichungen läßt sich nicht so einfach lösen wie im gleichungsbeschränkten Fall. Das Problem ist darin begründet, daß die Indexmenge $A(\hat{x})$

der aktiven Ungleichungsbeschränkungen unbekannt ist. Wäre sie bekannt, so könnten die inaktiven Ungleichungsbeschränkungen weggelassen werden, da sie keinen Einfluß auf das Optimum haben und man erhielte das äquivalente Problem

$$\begin{array}{lllll} \text{Minimiere} & \frac{1}{2}x^\top Wx + c^\top x & & & \\ \text{unter} & g_i(x) = a_i^\top x - u_i & = & 0, & i \in A(\hat{x}), \\ & h_j(x) = b_j^\top x - v_j & = & 0, & j \in \mathcal{J}. \end{array} \tag{5.27}$$

Dieses ist ein quadratisches Optimierungsproblem mit Gleichungsbeschränkungen, dessen Lösung durch Satz 5.6.1 charakterisiert ist.
Die Idee der **Strategie der aktiven Mengen** zur Lösung des allgemeinen quadratischen Optimierungsproblems besteht nun darin, die unbekannte Indexmenge $A(\hat{x})$ in (5.27) durch eine Schätzung $\mathcal{I}_s \subseteq \mathcal{I}$ zu ersetzen und diese iterativ anzupassen.
Sei $x^{[k]}$ der aktuelle Iterationspunkt, $x^{[k]}$ sei zulässig und $\mathcal{I}_s^k \subseteq \mathcal{I}$ sei die aktuelle Schätzung der aktiven Menge. Löse dann das Hilfsproblem

Hilfsproblem:
Minimiere

$$f(x^{[k]} + d)$$

bzgl. $d \in \mathbb{R}^n$ unter den Nebenbedingungen

$$\begin{aligned} a_i^\top d &= 0, \qquad i \in \mathcal{I}_s^k, \\ b_j^\top d &= 0, \qquad j \in \mathcal{J}. \end{aligned}$$

Die Strategie zur Anpassung der Indexmenge $\mathcal{I}_s^k$ hängt nun ab von der Lösung d und den zugehörigen Lagrange-Multiplikatoren λ_i, $i \in \mathcal{I}_s^k$ und μ_j, $j = 1, \ldots, m$ des Hilfsproblems. Folgende Fälle können eintreten:

(a) Besitzt das Hilfsproblem die Lösung $d = 0$, so liefern die KKT-Bedingungen für das Hilfssproblem

$$\nabla f(x^{[k]}) + \sum_{i \in \mathcal{I}_s^k} \lambda_i a_i + \sum_{j=1}^{m} \mu_j b_j = 0.$$

(i) Sind alle $\lambda_i \geq 0$, $i \in \mathcal{I}_s^k$, so wird $x^{[k]}$ als Lösung akzeptiert, da die KKT-Bedingungen für das Ausgangsproblem erfüllt sind, wenn man noch $\lambda_i = 0$, $i \in \mathcal{I} \setminus \mathcal{I}_s^k$ setzt.

(ii) **Deaktivierungsschritt:**
Gibt es einen Index $i \in \mathcal{I}_s^k$ mit $\lambda_i < 0$, so erfüllt $x^{[k]}$ die KKT-Bedingungen des Ausgangsproblems nicht, ist also nicht optimal. Andererseits ist $x^{[k]}$ Minimum

von f unter den Nebenbedingungen $\mathcal{I}_s^k \cup \mathcal{J}$. Daher muß der zulässige Bereich vergrößert werden, d.h. die Indexmenge $\mathcal{I}_s^k$ wird verkleinert.
Bestimme dazu denjenigen Index $q \in \mathcal{I}_s^k$ mit

$$\lambda_q = \min_{i \in \mathcal{I}_s^k} \lambda_i < 0$$

und setze

$$\mathcal{I}_s^{k+1} := \mathcal{I}_s^k \setminus \{q\}.$$

(b) Besitzt das Hilfsproblem eine Lösung $d \neq 0$, so ist d eine Abstiegsrichtung von f im Punkt $x^{[k]}$, die die Nebenbedingungen $\mathcal{I}_s^k \cup \mathcal{J}$ erfüllt, d.h. es gilt

$$a_i^\top d = 0, \quad i \in \mathcal{I}_s^k, \qquad b_j^\top d = 0, \quad j \in \mathcal{J}. \tag{5.28}$$

(i) Ist $x^{[k]} + d$ zulässig für das Ausgangsproblem, d.h. gilt

$$a_i^\top (x^{[k]} + d) \leq u_i, \qquad i \in \mathcal{I} \setminus \mathcal{I}_s^k,$$

so setze

$$x^{[k+1]} := x^{[k]} + d, \qquad \mathcal{I}_s^{k+1} := \mathcal{I}_s^k.$$

(ii) **Aktivierungsschritt:**
Ist $x^{[k]} + d$ unzulässig für das Ausgangsproblem, so bestimme eine möglichst große Schrittweite $\alpha_k \geq 0$, so daß $x^{[k]} + \alpha_k d$ zulässig bleibt:

$$a_i^\top (x^{[k]} + \alpha_k d) \leq u_i \qquad \forall i \in \mathcal{I} \setminus \mathcal{I}_s^k$$

bzw.

$$\alpha_k \left(a_i^\top d\right) \leq u_i - a_i^\top x^{[k]} \qquad \forall i \in \mathcal{I} \setminus \mathcal{I}_s^k \tag{5.29}$$

Beachte, daß $x^{[k]} + \alpha d$ für alle α zulässig bleibt für die Nebenbedingungen $\mathcal{I}_s^k \cup \mathcal{J}$, da $x^{[k]}$ zulässig ist und (5.28) gilt.
Da $x^{[k]}$ zulässig ist und $x^{[k]} + \alpha d$ mit $\alpha = 1$ unzulässig ist, muß es in (5.29) einen Index $i \in \mathcal{I} \setminus \mathcal{I}_s^k$ mit $a_i^\top d > 0$ geben. Bestimme also

$$\alpha_k := \min \left\{ \frac{u_i - a_i^\top x^{[k]}}{a_i^\top d} \;\middle|\; i \in \mathcal{I} \setminus \mathcal{I}_s^k,\ a_i^\top d > 0 \right\}.$$

Sei $r \in \mathcal{I} \setminus \mathcal{I}_s^k$ ein (nicht notwendig eindeutiger) Index, für den dieses Minimum angenommen wird. Der Fall $\alpha = 0$ kann auftreten, wenn mehrere Nebenbedingungen gleichzeitig aktiv werden (Entartung).
Setze

$$x^{[k+1]} := x^{[k]} + \alpha_k d, \qquad \mathcal{I}_s^{k+1} := \mathcal{I}_s^k \cup \{r\}.$$

Zusammenfassend erhalten wir den folgenden Algorithmus.

Algorithmus: Strategie der aktiven Menge

(i) Sei $x^{[0]}$ zulässig für das quadratische Optimierungsproblem. Setze $k := 0$ und

$$\mathcal{I}_s^0 := \{i \in \mathcal{I} \mid a_i^\top x^{[0]} = u_i\}.$$

(ii) Bestimme eine Lösung $(d, \lambda_{\mathcal{I}_s^k}, \mu)$ des Hilfsproblems durch Lösen des Gleichungssystems

$$\begin{pmatrix} W & A_{\mathcal{I}_s^k}^\top & B^\top \\ A_{\mathcal{I}_s^k} & 0 & 0 \\ B & 0 & 0 \end{pmatrix} \begin{pmatrix} d \\ \lambda_{\mathcal{I}_s^k} \\ \mu \end{pmatrix} = \begin{pmatrix} -\nabla f(x^{[k]}) \\ 0 \\ 0 \end{pmatrix}$$

(vgl. (5.26)). Hierin sind

$$A_{\mathcal{I}_s^k} := \left(a_i^\top\right)_{i \in \mathcal{I}_s^k}, \qquad \lambda_{\mathcal{I}_s^k} := (\lambda_i)_{i \in \mathcal{I}_s^k}.$$

(iii) Ist $d = 0$ und $\lambda_{\mathcal{I}_s^k} \geq 0$, so setze $\lambda_i = 0$ für $i \in \mathcal{I} \setminus \mathcal{I}_s^k$ und STOP.

(iv) Ist $d = 0$ und $\lambda_q := \min\{\lambda_i \mid i \in \mathcal{I}_s^k\}$, so setze

$$\mathcal{I}_s^{k+1} := \mathcal{I}_s^k \setminus \{q\}.$$

Setze $k := k + 1$ und gehe zu (ii).

(v) Gilt $a_i^\top (x^{[k]} + d) \leq u_i$, $i \in \mathcal{I} \setminus \mathcal{I}_s^k$, so setze

$$x^{[k+1]} := x^{[k]} + d, \qquad \mathcal{I}_s^{k+1} := \mathcal{I}_s^k.$$

Setze $k := k + 1$ und gehe zu (ii).

(vi) Bestimme $r \in \mathcal{I} \setminus \mathcal{I}_s^k$ mit

$$\alpha_k := \frac{u_r - a_r^\top x^{[k]}}{a_r^\top d} = \min \left\{ \frac{u_i - a_i^\top x^{[k]}}{a_i^\top d} \;\middle|\; i \in \mathcal{I} \setminus \mathcal{I}_s^k,\ a_i^\top d > 0 \right\}$$

und setze

$$x^{[k+1]} := x^{[k]} + \alpha_k d, \qquad \mathcal{I}_s^{k+1} := \mathcal{I}_s^k \cup \{r\}.$$

Setze $k := k + 1$ und gehe zu (ii).

Bemerkung 5.6.2

- *Ein zulässiger Startpunkt $x^{[0]}$ für das Active-Set-Verfahren kann analog zu Phase I*

des Simplexverfahrens berechnet werden.

- *Ist W positiv definit und sind die Vektoren a_i, $i \in \mathcal{I}_s^0$ und b_j, $j \in \mathcal{J}$ linear unabhängig, so ist der Algorithmus wohldefiniert und endet nach endlich vielen Schritten mit der eindeutig bestimmten Lösung.*

- *Es gibt alternative Verfahren zur Lösung quadratischer Optimierungsprobleme:*
  - *Das Verfahren von Goldfarb-Idnani [GI83] für streng konvexe quadratische Optimierungsprobleme ist eine Active-Set-Strategie für das duale Problem. Es hat den Vorteil, daß kein zulässiger Startpunkt berechnet werden muß.*
  - *Der simplexartige Algorithmus von Lemke (vgl. Cottle et al. [CPS92]) zur Lösung linearer Komplementaritätsprobleme kann ebenfalls zur Lösung quadratischer Optimierungsprobleme verwendet werden. Zu beachten ist hierbei, daß die KKT-Bedingungen des quadratischen Optimierungsproblems ein lineares Komplementaritätsproblem darstellen.*
  - *Innere-Punkt-Verfahren (vgl. Vanderbei [Van01]) stellen eine weitere alternative zur Lösung quadratischer Optimierungsprobleme dar.*
  - *Gill et al. [GM78, GMSW91] beschreiben ein Verfahren, welches auch nichtkonvexe quadratische Probleme lösen kann.*

Beispiel 5.6.3

Das folgende Beispiel bezieht sich auf die etwas allgemeinere Aufgabenstellung

$$\min \frac{1}{2} x^\top W x + c^\top x \quad \textit{unter} \quad l \leq \begin{pmatrix} x \\ Ax \end{pmatrix} \leq u,$$

$W \in \mathbb{R}^{n\times n}$, $c, x \in \mathbb{R}^n$, $A \in \mathbb{R}^{m\times n}$, $l, u \in \mathbb{R}^{n+m}$.
Spezielle Daten:

$$W = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}, \quad c = \begin{pmatrix} 2 \\ 1 \end{pmatrix}, \quad A = \begin{pmatrix} 1 & 1 \\ -1 & 1 \\ 1 & 0 \end{pmatrix}, \quad l = \begin{pmatrix} -\infty \\ -1 \\ 0 \\ -\infty \\ -\infty \end{pmatrix}, u = \begin{pmatrix} 5 \\ 2 \\ 5 \\ 2 \\ 5 \end{pmatrix}.$$

Startschätzung: $x^{[0]} = (5, 0)^\top$.
Resultat des Algorithmus:

```
------------------------------------------------------
INITIAL X IS FEASIBLE!
STARTING OPTIMIZATION PHASE ...
------------------------------------------------------
------------------------------------------------------
                   ITERATION  0
------------------------------------------------------
OBJ =        0.2250000000000000E+02
KKT =        0.7000000000000000E+01
CON =        0.0000000000000000E+00
```

```
VARIABLE     LB                              VALUE                           UB                          STATUS      LAMBDA
 1          -0.1000000000000000E+21          0.5000000000000000E+01          0.5000000000000000E+01      UB          0.0000000000000000E+00
 2          -0.1000000000000000E+01          0.0000000000000000E+00          0.2000000000000000E+01      IA          0.0000000000000000E+00
CONSTRAINT   LB                              VALUE                           UB                          STATUS      LAMBDA
 3           0.0000000000000000E+00          0.5000000000000000E+01          0.5000000000000000E+01      UB          0.0000000000000000E+00
 4          -0.1000000000000000E+21          0.5000000000000000E+01          0.2000000000000000E+01      IA          0.0000000000000000E+00
 5          -0.1000000000000000E+21          0.5000000000000000E+01          0.5000000000000000E+01      UB          0.0000000000000000E+00
DEACTIVATING LINEAR DEPENDENT CONSTRAINT  5
DEACTIVATING CONSTRAINT  1 WITH MULTIPLIER VALUE LAMBDA =  -6.
-----------------------------------------------------
                  ITERATION  1
-----------------------------------------------------
OBJ =        0.2250000000000000E+02
KKT =        0.6000000000000000E+01
CON =        0.0000000000000000E+00
VARIABLE     LB                              VALUE                           UB                          STATUS      LAMBDA
 1          -0.1000000000000000E+21          0.5000000000000000E+01          0.5000000000000000E+01      IA          0.0000000000000000E+00
 2          -0.1000000000000000E+01          0.0000000000000000E+00          0.2000000000000000E+01      IA          0.0000000000000000E+00
CONSTRAINT   LB                              VALUE                           UB                          STATUS      LAMBDA
 3           0.0000000000000000E+00          0.5000000000000000E+01          0.5000000000000000E+01      UB         -0.1000000000000000E+01
 4          -0.1000000000000000E+21         -0.5000000000000000E+01          0.2000000000000000E+01      IA          0.0000000000000000E+00
 5          -0.1000000000000000E+21          0.5000000000000000E+01          0.5000000000000000E+01      IA          0.0000000000000000E+00
ACTIVATING CONSTRAINT  2 OF TYPE UB
-----------------------------------------------------
                  ITERATION  2
-----------------------------------------------------
OBJ =        0.1450000000000000E+02
KKT =        0.1000000000000000E+01
CON =        0.0000000000000000E+00
VARIABLE     LB                              VALUE                           UB                          STATUS      LAMBDA
 1          -0.1000000000000000E+21          0.3000000000000000E+01          0.5000000000000000E+01      IA          0.0000000000000000E+00
 2          -0.1000000000000000E+01          0.2000000000000000E+01          0.2000000000000000E+01      UB          0.0000000000000000E+00
CONSTRAINT   LB                              VALUE                           UB                          STATUS      LAMBDA
 3           0.0000000000000000E+00          0.5000000000000000E+01          0.5000000000000000E+01      UB         -0.4000000000000000E+01
 4          -0.1000000000000000E+21         -0.1000000000000000E+01          0.2000000000000000E+01      IA          0.0000000000000000E+00
 5          -0.1000000000000000E+21          0.3000000000000000E+01          0.5000000000000000E+01      IA          0.0000000000000000E+00
DEACTIVATING CONSTRAINT  3 WITH MULTIPLIER VALUE LAMBDA =  -5.
-----------------------------------------------------
                  ITERATION  3
-----------------------------------------------------
OBJ =        0.1450000000000000E+02
KKT =        0.5000000000000000E+01
CON =        0.0000000000000000E+00
VARIABLE     LB                              VALUE                           UB                          STATUS      LAMBDA
 1          -0.1000000000000000E+21          0.3000000000000000E+01          0.5000000000000000E+01      IA          0.0000000000000000E+00
 2          -0.1000000000000000E+01          0.2000000000000000E+01          0.2000000000000000E+01      UB          0.2000000000000000E+01
CONSTRAINT   LB                              VALUE                           UB                          STATUS      LAMBDA
 3           0.0000000000000000E+00          0.5000000000000000E+01          0.5000000000000000E+01      IA          0.0000000000000000E+00
 4          -0.1000000000000000E+21         -0.1000000000000000E+01          0.2000000000000000E+01      IA          0.0000000000000000E+00
 5          -0.1000000000000000E+21          0.3000000000000000E+01          0.5000000000000000E+01      IA          0.0000000000000000E+00
ACTIVATING CONSTRAINT  4 OF TYPE UB
-----------------------------------------------------
                  ITERATION  4
-----------------------------------------------------
OBJ =        0.4000000000000000E+01
KKT =        0.2000000000000000E+01
CON =        0.0000000000000000E+00
VARIABLE     LB                              VALUE                           UB                          STATUS      LAMBDA
 1          -0.1000000000000000E+21          0.0000000000000000E+00          0.5000000000000000E+01      IA          0.0000000000000000E+00
 2          -0.1000000000000000E+01          0.2000000000000000E+01          0.2000000000000000E+01      UB         -0.3000000000000000E+01
CONSTRAINT   LB                              VALUE                           UB                          STATUS      LAMBDA
 3           0.0000000000000000E+00          0.2000000000000000E+01          0.5000000000000000E+01      IA          0.0000000000000000E+00
 4          -0.1000000000000000E+21          0.2000000000000000E+01          0.2000000000000000E+01      UB          0.0000000000000000E+00
 5          -0.1000000000000000E+21          0.0000000000000000E+00          0.5000000000000000E+01      IA          0.0000000000000000E+00
DEACTIVATING CONSTRAINT  2 WITH MULTIPLIER VALUE LAMBDA =  -5.
-----------------------------------------------------
                  ITERATION  5
-----------------------------------------------------
OBJ =        0.4000000000000000E+01
KKT =        0.5000000000000000E+01
CON =        0.0000000000000000E+00
VARIABLE     LB                              VALUE                           UB                          STATUS      LAMBDA
 1          -0.1000000000000000E+21          0.0000000000000000E+00          0.5000000000000000E+01      IA          0.0000000000000000E+00
 2          -0.1000000000000000E+01          0.2000000000000000E+01          0.2000000000000000E+01      IA          0.0000000000000000E+00
CONSTRAINT   LB                              VALUE                           UB                          STATUS      LAMBDA
 3           0.0000000000000000E+00          0.2000000000000000E+01          0.5000000000000000E+01      IA          0.0000000000000000E+00
 4          -0.1000000000000000E+21          0.2000000000000000E+01          0.2000000000000000E+01      UB          0.2000000000000000E+01
 5          -0.1000000000000000E+21          0.0000000000000000E+00          0.5000000000000000E+01      IA          0.0000000000000000E+00
ACTIVATING CONSTRAINT  3 OF TYPE LB
-----------------------------------------------------
                  ITERATION  6
-----------------------------------------------------
OBJ =        0.6661338147750939E-15
KKT =        0.1500000000000000E+01
CON =        0.0000000000000000E+00
VARIABLE     LB                              VALUE                           UB                          STATUS      LAMBDA
 1          -0.1000000000000000E+21         -0.9999999999999998E+00          0.5000000000000000E+01      IA          0.0000000000000000E+00
 2          -0.1000000000000000E+01          0.1000000000000000E+01          0.2000000000000000E+01      IA          0.0000000000000000E+00
CONSTRAINT   LB                              VALUE                           UB                          STATUS      LAMBDA
 3           0.0000000000000000E+00          0.0000000000000000E+00          0.5000000000000000E+01      LB          0.0000000000000000E+00
 4          -0.1000000000000000E+21          0.2000000000000000E+01          0.2000000000000000E+01      UB         -0.4999999999999998E+00
 5          -0.1000000000000000E+21         -0.9999999999999998E+00          0.5000000000000000E+01      IA          0.0000000000000000E+00
DEACTIVATING CONSTRAINT  4 WITH MULTIPLIER VALUE LAMBDA =  -0.5
-----------------------------------------------------
                  ITERATION  7
-----------------------------------------------------
OBJ =        0.6661338147750939E-15
KKT =        0.5000000000000000E+00
CON =        0.0000000000000000E+00
VARIABLE     LB                              VALUE                           UB                          STATUS      LAMBDA
 1          -0.1000000000000000E+21         -0.9999999999999998E+00          0.5000000000000000E+01      IA          0.0000000000000000E+00
 2          -0.1000000000000000E+01          0.1000000000000000E+01          0.2000000000000000E+01      IA          0.0000000000000000E+00
CONSTRAINT   LB                              VALUE                           UB                          STATUS      LAMBDA
 3           0.0000000000000000E+00          0.0000000000000000E+00          0.5000000000000000E+01      LB         -0.1500000000000000E+01
 4          -0.1000000000000000E+21          0.2000000000000000E+01          0.2000000000000000E+01      IA          0.0000000000000000E+00
 5          -0.1000000000000000E+21         -0.9999999999999998E+00          0.5000000000000000E+01      IA          0.0000000000000000E+00
-----------------------------------------------------
                  ITERATION  8
-----------------------------------------------------
OBJ =       -0.2499999999999994E+00
KKT =        0.2220446049250313E-15
CON =        0.1110223024625157E-15
VARIABLE     LB                              VALUE                           UB                          STATUS      LAMBDA
 1          -0.1000000000000000E+21         -0.4999999999999999E+00          0.5000000000000000E+01      IA          0.0000000000000000E+00
 2          -0.1000000000000000E+01          0.5000000000000002E+00          0.2000000000000000E+01      IA          0.0000000000000000E+00
CONSTRAINT   LB                              VALUE                           UB                          STATUS      LAMBDA
```

```
 3          0.0000000000000000E+00       -0.1110223024625157E-15       0.5000000000000000E+01    LB      -0.1500000000000000E+01
 4         -0.1000000000000000E+21        0.1000000000000000E+01       0.2000000000000000E+01    IA       0.0000000000000000E+00
 5         -0.1000000000000000E+21       -0.4999999999999999E+00       0.5000000000000000E+01    IA       0.0000000000000000E+00

OPTIMAL SOLUTION FOUND AT ITERATION  8
```

5.6.1 Effiziente Lösung der Gleichungssysteme

Wir diskutieren, wie lineare Gleichungssysteme der Form

$$\begin{pmatrix} W & A^\top \\ A & 0 \end{pmatrix} \begin{pmatrix} x \\ \mu \end{pmatrix} = \begin{pmatrix} d \\ f \end{pmatrix}$$

mit $W \in \mathbb{R}^{n\times n}$, $A \in \mathbb{R}^{m\times n}$, $d \in \mathbb{R}^n$ und $f \in \mathbb{R}^m$ effizient gelöst werden können. Die Matrix W sei positiv definit auf dem Kern von A und $\mathrm{Rang}(A) = m$, $m \le n$. Unter diesen Voraussetzungen ist die Matrix invertierbar und das Gleichungssystem ist eindeutig lösbar.

(a) **QR-Methode:**

QR-Zerlegung von $A^\top$:

$$A^\top = Q \begin{pmatrix} R \\ 0 \end{pmatrix}, \qquad Q = (Y, Z) \in \mathbb{R}^{n\times n},\ Y \in \mathbb{R}^{n\times m},\ Z \in \mathbb{R}^{n\times(n-m)},\ R \in \mathbb{R}^{m\times m}.$$

R ist reguläre rechte obere Dreiecksmatrix, Q ist orthogonal, Y ist eine Orthonormalbasis des Bildes von $A^\top$, Z ist eine Orthonormalbasis von $Bild(A^\top)^\perp = \ker(A)$. Jedes $x \in \mathbb{R}^n$ kann als $x = Yx_Y + Zx_z$ mit eindeutig bestimmten Vektoren $x_Y \in \mathbb{R}^m$ und $x_z \in \mathbb{R}^{n-m}$ dargestellt werden. Zudem gilt $\ker(A) = \{Zx_Z \mid x_Z \in \mathbb{R}^{n-m}\}$.

Transformation des Gleichungssystems:

$$\begin{aligned} Wx + A^\top\mu = d \quad &\Leftrightarrow \quad Q^\top WQQ^\top x + Q^\top A^\top \mu = Q^\top d \\ Ax = f \quad &\Leftrightarrow \quad AQQ^\top x = f \end{aligned}$$

In Matrixschreibweise:

$$\begin{pmatrix} Y^\top WY & Y^\top WZ & R \\ Z^\top WY & Z^\top WZ & 0 \\ R^\top & 0 & 0 \end{pmatrix} \begin{pmatrix} x_Y \\ x_Z \\ \mu \end{pmatrix} = \begin{pmatrix} Y^\top d \\ Z^\top d \\ f \end{pmatrix}$$

mit $x_Y := Y^\top x$ und $x_Z := Z^\top x$. Dieses Gleichungssystem hat Dreiecksgestalt und kann blockweise gelöst werden:

(i) Löse $R^\top x_Y = f$ durch Rückwärtssubstitution.
(ii) Löse $Z^\top WZx_Z = Z^\top d - Z^\top WYx_Y$ mittels Cholesky-Zerlegung von $Z^\top WZ$.
(iii) Löse $R\mu = Y^\top d - Y^\top WYx_Y - Y^\top WZx_Z$ durch Rückwärtssubstitution.

Spezialfall: $f = 0$:

(i) $x_Y = 0$.
(ii) Löse $Z^\top W Z x_Z = Z^\top d$ mittels Cholesky-Zerlegung von $Z^\top W Z$.
(iii) Löse $R\mu = Y^\top d - Y^\top W Z x_Z$ durch Rückwärtssubstitution.

(b) **Eliminationsmethode:**

Wir nehmen an, daß A partitioniert ist gemäß

$$A = (A_1 \mid A_2), \qquad A_1 \in \mathbb{R}^{m\times m},\ A_2 \in \mathbb{R}^{m\times(n-m)},$$

wobei A_1 invertierbar sei (numerisch kann dieses mit Hilfe der Gauss-Elimination mit Zeilen- und Spaltenpivoting realisiert werden).

Entsprechend seien x und W partitioniert: $x = (x_1, x_2)^\top \in \mathbb{R}^{m+(n-m)}$, $W = (W_1 \mid W_2)$, $W_1 \in \mathbb{R}^{n\times m}$, $W_2 \in \mathbb{R}^{n\times(n-m)}$.

Transformation des Gleichungssystems:

$$Ax = f \quad \Leftrightarrow \quad A_1x_1 + A_2x_2 = f \quad \Leftrightarrow \quad x_1 = A_1^{-1}\left(f - A_2x_2\right).$$

Weiter:

$$W_1x_1 + W_2x_2 + A^\top\mu = d \quad \Leftrightarrow \quad \underbrace{\left(W_2 - W_1A_1^{-1}A_2\right)}_{=:\begin{pmatrix} Z_1 \\ Z_2 \end{pmatrix}} x_2 + A^\top\mu = \underbrace{d - W_1A_1^{-1}f}_{=:\begin{pmatrix} \tilde d_1 \\ \tilde d_2 \end{pmatrix}}.$$

Die erste „Zeile“ liefert

$$\mu = A_1^{-\top}\left(\tilde d_1 - Z_1x_2\right).$$

Einsetzen in die zweite Gleichung liefert

$$\left(Z_2 - (A_1^{-1}A_2)^\top Z_1\right) x_2 = \tilde d_2 - (A_1^{-1}A_2)^\top \tilde d_1.$$

Zusammenfassend entsteht der folgende Algorithmus zur Lösung des Gleichungssystems:

(i) Berechne M als Lösung von $A_1 \cdot M = A_2$ und r als Lösung von $A_1 \cdot r = f$.
(ii) Berechne $Z := \begin{pmatrix} Z_1 \\ Z_2 \end{pmatrix} = W_2 - W_1 \cdot M$, $\tilde d := \begin{pmatrix} \tilde d_1 \\ \tilde d_2 \end{pmatrix} = d - W_1 r$.
(iii) Löse $\left(Z_2 - M^\top Z_1\right) x_2 = \tilde d_2 - M^\top \tilde d_1$.
(iv) Löse $A_1^\top \mu = \tilde d_1 - Z_1 x_2$.
(v) Berechne $x_1 = r - Mx_2$.

5.7 Lagrange-Newton-Verfahren

In diesem Abschnitt beschränken wir uns auf Standard-Optimierungsprobleme ohne Ungleichungsrestriktionen ($m = 0$) und $S = \mathbb{R}^n$ und betrachten das

Gleichungsrestringiertes Optimierungsproblem:
Minimiere $f(x)$ bezüglich $x \in \mathbb{R}^n$ unter den Nebenbedingungen

$$h_j(x) = 0, \quad j = 1, \dots, p.$$

Es sei $\hat{x}$ ein lokales Minimum und die Gradienten $\nabla h_j(\hat{x})$, $j = 1, \dots, p$ seien linear unabhängig, d.h. es gelte die **Regularitätsbedingung der linearen Unabhängigkeit (LICQ)**. Dann gelten die KKT-Bedingungen: Es existieren Multiplikatoren $\hat{\mu} = (\hat{\mu}_1, \dots, \hat{\mu}_p)^\top \in \mathbb{R}^p$ mit

$$\begin{aligned} \nabla_x L(\hat{x}, l_0 = 1, \hat{\mu}) = \nabla f(\hat{x}) + \sum_{j=1}^{p} \hat{\mu}_j \nabla h_j(\hat{x}) &= 0, \\ h_j(\hat{x}) &= 0, \quad j = 1, \dots, p, \end{aligned}$$

wobei L wieder die Lagrange-Funktion bezeichnet. Dies ist ein **nichtlineares Gleichungssystem** für $\hat{x}$ und $\hat{\mu}$. Mit $F : \mathbb{R}^n \times \mathbb{R}^p \to \mathbb{R}^{n+p}$ und

$$F(x, \mu) := \begin{pmatrix} \nabla_x L(x, l_0 = 1, \mu) \\ h(x) \end{pmatrix}, \quad h(x) = \begin{pmatrix} h_1(x) \\ \vdots \\ h_p(x) \end{pmatrix}$$

lautet es

$$F(\hat{x}, \hat{\mu}) = 0. \tag{5.30}$$

Das **Lagrange-Newton-Verfahren** basiert auf der Anwendung des Newton-Verfahrens zur numerischen Lösung der notwendigen Bedingungen (5.30). Dies führt auf den folgenden Algorithmus:

Algorithmus: Lagrange-Newton-Verfahren

(i) Wähle Startschätzungen $x^{[0]} \in \mathbb{R}^n$ und $\mu^{[0]} \in \mathbb{R}^p$, $\varepsilon > 0$ und setze $k = 0$.

(ii) Falls $\|F(x^{[k]}, \mu^{[k]})\| \leq \varepsilon$, STOP.

(iii) Löse das lineare Gleichungssystem

$$\begin{pmatrix} \nabla^2_{xx} L(x^{[k]}, l_0 = 1, \mu^{[k]}) & h'(x^{[k]})^\top \\ h'(x^{[k]}) & 0 \end{pmatrix} \begin{pmatrix} d \\ v \end{pmatrix} = - \begin{pmatrix} \nabla_x L(x^{[k]}, l_0 = 1, \mu^{[k]}) \\ h(x^{[k]}) \end{pmatrix} \tag{5.31}$$

und setze

$$x^{[k+1]} = x^{[k]} + d, \qquad \mu^{[k+1]} = \mu^{[k]} + v. \tag{5.32}$$

(iv) Setze $k := k + 1$ und gehe zu (ii).

Ausnutzung und Anpassung der bekannten Konvergenzresultate für das Newtonverfahren auf dieses spezielle System liefert den folgenden lokalen Konvergenzsatz.

Satz 5.7.1 (Lokal quadratische Konvergenz)

(i) Sei $(\hat{x}, \hat{\mu})$ ein KKT-Punkt.

(ii) Die Funktionen f, h_j, $j = 1, \ldots, p$ seien zweimal stetig differenzierbar.

(iii) Die Matrix

$$\begin{pmatrix} \nabla^2_{xx} L(\hat{x}, l_0 = 1, \hat{\mu}) & h'(\hat{x})^\top \\ h'(\hat{x}) & 0 \end{pmatrix} \tag{5.33}$$

sei invertierbar.

Dann existiert ein $r > 0$, so daß das Lagrange-Newton-Verfahren für alle Startwerte $(x^{[0]}, \mu^{[0]}) \in U_r(\hat{x}, \hat{\mu})$ wohldefiniert ist und die Folge $\{(x^{[i]}, \mu^{[i]})\}$ konvergiert superlinear gegen $(\hat{x}, \hat{\mu})$.
Sind die zweiten Ableitungen f'' und h_j'', $j = 1, \ldots, p$ sogar Lipschitz-stetig, so ist die Konvergenz quadratisch.

Beweis: Der Beweis lautet beinahe wörtlich wie der Beweis zu Satz 3.9.8 im Abschnitt über das Newtonverfaren für unrestringierte Probleme. □

Beispiel 5.7.2

$$\begin{aligned} &\textit{Minimiere} && 2x_1^4 + x_2^4 + 4x_1^2 - x_1 x_2 + 6x_2^2 \\ &\textit{unter} && 2x_1 - x_2 = -4, \ x_1, x_2 \in \mathbb{R}. \end{aligned}$$

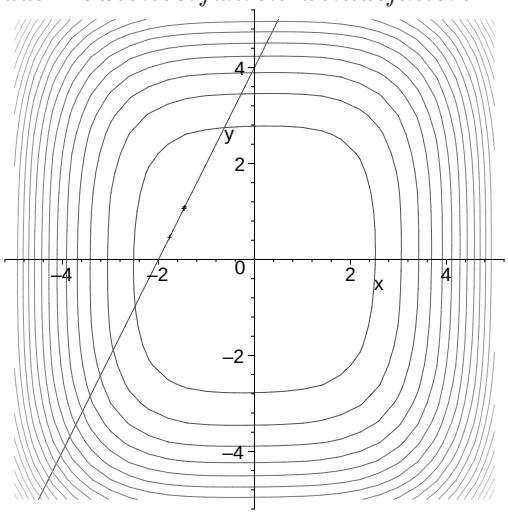
Der Gradient von f ist

$$\nabla f(x) = \begin{pmatrix} 8x_1^3 + 8x_1 - x_2 \\ 4x_2^3 - x_1 + 12x_2 \end{pmatrix}.$$

Die Hessematrix von f lautet

$$\nabla^2 f(x) = \begin{pmatrix} 24x_1^2 + 8 & -1 \\ -1 & 12x_2^2 + 12 \end{pmatrix}.$$

$\nabla^2 f$ ist symmetrisch und diagonaldominant und damit positiv definit. Der Rang von A ist 1. Damit ist das Newtonverfahren wohldefiniert.



Ausgabe des Lagrange-Newton-Verfahrens:

```
ITERATION  0
  ZIELFUNKTION  =         0.0000000000000000E+00
  BESCHRAENKUNG =         0.4000000000000000E+01
  NORM KKT      =         0.0000000000000000E+00
  X=              0.0000000000000000E+00        0.0000000000000000E+00
  LAMBDA=         0.0000000000000000E+00

ITERATION  1
  ZIELFUNKTION  =         0.3425678372606001E+02
  BESCHRAENKUNG =         0.0000000000000000E+00
  NORM KKT      =         0.4430579634007108E+02
  X=             -0.1769230769230769E+01        0.4615384615384616E+00
  LAMBDA=         0.7307692307692307E+01

ITERATION  2
  ZIELFUNKTION  =         0.2756373273045131E+02
  BESCHRAENKUNG =         0.0000000000000000E+00
  NORM KKT      =         0.5155005942054521E+01
  X=             -0.1452391998833495E+01        0.1095216002333011E+01
  LAMBDA=         0.1660806234685793E+02

ITERATION  3
  ZIELFUNKTION  =         0.2754452288430214E+02
  BESCHRAENKUNG =         0.0000000000000000E+00
  NORM KKT      =         0.1497798610217076E-01
  X=             -0.1467843643905216E+01        0.1064312712189567E+01
  LAMBDA=         0.1904961295898514E+02

ITERATION  4
  ZIELFUNKTION  =         0.2754452202163215E+02
  BESCHRAENKUNG =         0.0000000000000000E+00
  NORM KKT      =         0.6772437517980240E-06
  X=             -0.1467948113419141E+01        0.1064103773161718E+01
  LAMBDA=         0.1905680332151563E+02

ITERATION  5
  ZIELFUNKTION  =         0.2754452202163215E+02
  BESCHRAENKUNG =         0.0000000000000000E+00
```

```
NORM KKT      =         0.3552713678800501E-14
X=             -0.1467948118040920E+01        0.1064103763918160E+01
LAMBDA=         0.1905680364713604E+02
```

Bemerkung 5.7.3

- *In (5.33) tritt wieder die KKT-Matrix auf (vgl. auch die Abschnitte über parametrische Optimierung und quadratische Optimierung). Die KKT-Matrix ist beispielsweise invertierbar, wenn die folgenden Bedingungen erfüllt sind (→ Übung):*

  (i) Die Gradienten $\nabla h_j(\hat{x})$, $j = 1, \dots, p$ sind linear unabhängig (LICQ).

  (ii) Es gilt

  $$d^\top \nabla^2_{xx} L(\hat{x}, l_0 = 1, \hat{\mu}) d > 0 \qquad \forall d \neq 0 \ : \ h'(\hat{x}) d = 0.$$

  Dies ist gerade die hinreichende Bedingung zweiter Ordnung, vgl. (5.14).

- *Treten auch Ungleichungen $g_i(x) \leq 0$, $i = 1, \dots, m$ auf, so lauten die KKT-Bedingungen (unter einer geeigneten Regularitätsbedingung)*

$$\begin{aligned} \nabla_x L(\hat{x}, l_0 = 1, \hat{\lambda}, \hat{\mu}) &= 0, \\ h(\hat{x}) &= 0, \\ g(\hat{x}) &\leq 0, \\ \hat{\lambda}_i g_i(\hat{x}) &= 0, \qquad i = 1, \dots, m, \\ \lambda_i &\geq 0, \qquad i = 1, \dots, m. \end{aligned}$$

Unter Verwendung einer sogenannten **NCP-Funktion** *φ mit der Eigenschaft*

$$\varphi(a, b) = 0 \qquad \Leftrightarrow \qquad a \geq 0, \ b \geq 0, \ ab = 0$$

lauten die KKT-Bedingungen äquivalent

$$\begin{aligned} \nabla_x L(\hat{x}, l_0 = 1, \hat{\lambda}, \hat{\mu}) &= 0, \\ h(\hat{x}) &= 0, \\ \varphi(-g_i(\hat{x}), \hat{\lambda}_i) &= 0, \qquad i = 1, \dots, m. \end{aligned}$$

Dies ist wieder ein nichtlineares Gleichungssystem der Form $F(x, \lambda, \mu) = 0$, allerdings ist F häufig nicht mehr differenzierbar. Es kommen **nichtdifferenzierbare Newtonverfahren** *zum Einsatz. Beispiele für NCP-Funktionen sind*

$$\varphi(a, b) = \min\{a, b\}, \qquad \varphi(a, b) = \sqrt{a^2 + b^2} - a - b.$$

Für Details sei auf Geiger und Kanzow [GK02] verwiesen.

5.8 Sequentielle quadratische Programmierung

Die sequentielle quadratische Programmierung (SQP) wird, z.B., in [Han77], [Pow78], [GMW81], [Sto85], [Sch81], [Sch83], [Alt02], [GK02] behandelt. Es existieren diverse Implementierungen, z.B. [Sch85], [Kra88], [GMSW98], [GMS02]. Spezielle Anpassungen auf diskretisierte Optimalsteuerungsprobleme werden in [GMS94], [Sch96], [Ste95], [BH99] besprochen.

Wir starten zunächst mit einer Beobachtung. Das lineare Gleichungssystem (5.31) in (iii) des Lagrange-Newton-Verfahrens entsteht auch auf andere Art. Wir erinnern uns an das Newton-Verfahren für unrestringierte Optimierungsprobleme. Dort hatten wir das Newtonverfahren auf zwei Arten motiviert: 1. Anwendung des Newtonverfahrens auf die notwendigen Bedingung $\nabla f = 0$ (indirekter Ansatz); 2. lokale Approximation der Zielfunktion durch eine quadratische Funktion (direkter Ansatz). Beide Ansätze lieferten das gleiche Verfaren.

Wir betrachten nun wieder das gleichungsrestringierte Optimierungsproblem und approximieren es lokal im Punkt $(x^{[k]}, \mu^{[k]})$ durch das **quadratische Optimierungsproblem**

$$\begin{aligned} \min_{d\in\mathbb{R}^n} \quad & \frac{1}{2} d^\top \nabla^2_{xx} L(x^{[k]}, l_0 = 1, \mu^{[k]}) d + \nabla f(x^{[k]})^\top d \\ \text{unter} \quad & h(x^{[k]}) + h'(x^{[k]}) d \;=\; 0. \end{aligned}$$

Die Lagrange-Funktion für das quadratische Problem ist gegeben durch (wir lassen l_0 weg, da für quadratische Probleme stets $l_0 = 1$ gilt, Abadie!)

$$L_{QP}(d, \eta) := \frac{1}{2} d^\top \nabla^2_{xx} L(x^{[k]}, l_0 = 1, \mu^{[k]}) d + \nabla f(x^{[k]})^\top d + \eta^\top \left(h(x^{[k]}) + h'(x^{[k]}) d \right).$$

Auswertung der notwendigen Bedingungen erster Ordnung führt auf

$$\begin{aligned} \nabla^2_{xx} L(x^{[k]}, l_0 = 1, \mu^{[k]}) d + \nabla f(x^{[k]}) + h'(x^{[k]})^\top \eta &= 0, \\ h(x^{[k]}) + h'(x^{[k]}) d &= 0, \end{aligned}$$

bzw.

$$\begin{pmatrix} \nabla^2_{xx} L(x^{[k]}, l_0 = 1, \mu^{[k]}) & h'(x^{[k]})^\top \\ h'(x^{[k]}) & 0 \end{pmatrix} \cdot \begin{pmatrix} d \\ \eta \end{pmatrix} = - \begin{pmatrix} \nabla f(x^{[k]}) \\ h(x^{[k]}) \end{pmatrix}. \tag{5.34}$$

Subtraktion von $h'(x^{[k]})^\top \mu^{[k]}$ auf beiden Seiten der ersten Gleichung in (5.34) liefert das lineare Gleichungssystem

$$\begin{pmatrix} \nabla^2_{xx} L(x^{[k]}, l_0 = 1, \mu^{[k]}) & h'(x^{[k]})^\top \\ h'(x^{[k]}) & 0 \end{pmatrix} \cdot \begin{pmatrix} d \\ \eta - \mu^{[k]} \end{pmatrix} = - \begin{pmatrix} \nabla_x L(x^{[k]}, l_0 = 1, \mu^{[k]}) \\ h(x^{[k]}) \end{pmatrix}. \tag{5.35}$$

Ein Vergeich von (5.35) mit (5.31) zeigt, daß diese zwei Gleichungssysteme identisch sind, wenn man noch $v := \eta - \mu^{(k)}$ definiert. Aus (5.32) folgt, daß die neuen Iterierten durch

$$x^{[k+1]} = x^{[k]} + d, \qquad \mu^{[k+1]} = \mu^{[k]} + v = \eta.$$

gegeben sind.

Zusammenfassung:
Für gleichungsrestringierte Optimierungsprobleme ist das Lagrange-Newton-Verfahren identisch mit dem oben hergeleiteten sukzessiven quadratischen Optimierungsverfahren, wenn der Multiplikator η des quadratischen Hilfsproblems als neue Approximation für den Multiplikator μ des Ausgangsproblems verwendet wird.

Diese Beobachtung motiviert die folgende Erweiterung des quadratischen Hilfsproblems für Standard-Optimierungsprobleme (mit $S = \mathbb{R}^n$) mit **Gleichungs- und Ungleichungsrestriktionen**:

QP Problem $(QP(x^{[k]}, \lambda^{[k]}, \mu^{[k]}))$**:**

$$\begin{aligned} \min_{d\in\mathbb{R}^n} \quad & \frac{1}{2} d^\top \nabla^2_{xx} L(x^{[k]}, l_0 = 1, \lambda^{[k]}, \mu^{[k]}) d + \nabla f(x^{[k]})^\top d \\ \text{unter} \quad & g_i(x^{[k]}) + \nabla g_i(x^{[k]})^\top d \;\leq\; 0, \quad i = 1, \ldots, m, \\ & h_j(x^{[k]}) + \nabla h_j(x^{[k]})^\top d \;=\; 0, \quad j = 1, \ldots, p. \end{aligned}$$

Sukzessive quadratische Approximation liefert das lokale SQP Verfahren:

Algorithmus: Lokales SQP Verfahren

(i) Wähle Startwerte $(x^{[0]}, \lambda^{[0]}, \mu^{[0]}) \in \mathbb{R}^n \times \mathbb{R}^m \times \mathbb{R}^p$ und setze $k = 0$.

(ii) Falls $(x^{[k]}, \lambda^{[k]}, \mu^{[k]})$ ein KKT-Punkt des Standard-Optimierungsproblems ist, STOP.

(iii) Berechne einen KKT-Punkt $(d^{[k]}, \lambda^{[k+1]}, \mu^{[k+1]}) \in \mathbb{R}^n \times \mathbb{R}^m \times \mathbb{R}^p$ des quadratischen Optimierungsproblems $(QP(x^{[k]}, \lambda^{[k]}, \mu^{[k]}))$.

(iv) Setze $x^{[k+1]} = x^{[k]} + d^{[k]}$, $k := k + 1$ und gehe zu (ii).

Bemerkung 5.8.1

- *Es ist nicht notwendig, die Indexmenge $A(\hat{x})$ der aktiven Ungleichungsnebenbedingungen im Voraus zu kennen.*
- *Die Iterierten $x^{[k]}$ sind in der Regel nicht zulässig, d.h. es gilt i.a. $x^{[k]} \notin \Sigma$.*

Die lokale Konvergenz des SQP Verfahrens wird im folgenden Satz formuliert.

Satz 5.8.2 (Lokale Konvergenz des SQP-Verfarens)

(i) Sei $\hat{x}$ lokales Minimum des Standard-Optimierungsproblems mit $S = \mathbb{R}^n$.

(ii) Die Funktionen f, g_i, $i = 1, \dots, m$ und h_j, $j = 1, \dots, p$ seien zweimal stetig differenzierbar mit Lipschitz-stetigen zweiten Ableitungen f'', g_i'', $i = 1, \dots, m$ und h_j'', $j = 1, \dots, p$.

(iii) Die Gradienten $\nabla g_i(\hat{x})$, $i \in A(\hat{x})$ und $\nabla h_j(\hat{x})$, $j = 1, \dots, p$ seien linear unabhängig (LICQ).

(iv) Die strikte Komplementaritätsbedingung $\hat{\lambda}_i - g_i(\hat{x}) > 0$ für alle $i \in A(\hat{x})$ sei erfüllt.

(v) Es gelte die hinreichende Bedingung zweiter Ordnung:

$$d^\top \nabla^2_{xx} L(\hat{x}, l_0 = 1, \hat{\lambda}, \hat{\mu}) d > 0$$

für alle $0 \neq d \in \mathbb{R}^n$ mit

$$\nabla g_i(\hat{x})^\top d = 0, \quad i \in A(\hat{x}), \quad \nabla h_j(\hat{x})^\top d = 0, \quad j = 1, \dots, p.$$

Dann existiert ein $r > 0$, so daß für beliebige Startwerte

$$(x^{[0]}, \lambda^{[0]}, \mu^{[0]}) \in U_r(\hat{x}, \hat{\lambda}, \hat{\mu})$$

alle QP-Probleme $(QP(x^{[k]}, \lambda^{[k]}, \mu^{[k]}))$ eine lokal eindeutige Lösung $d^{[k]}$ mit eindeutigen Multiplikatoren $\lambda^{[k+1]}$ und $\mu^{[k+1]}$ besitzen.
Desweiteren konvergiert die Folge $\{(x^{[k]}, \lambda^{[k]}, \mu^{[k]})\}$ quadratisch gegen $(\hat{x}, \hat{\lambda}, \hat{\mu})$.

Beweis: Der Beweis verwendet den Sensitivitätssatz 5.4.2 für parametrische Optimierungsprobleme, um zu zeigen, daß die Indexmenge der aktiven Beschränkungen in einer Umgebung der Lösung konstant bleibt. Dann kann gezeigt werden, daß das SQP-Verfahren lokal mit dem Lagrange-Newton-Verfahren übereinstimmt, welches quadratisch konvergiert.

1. Wir betrachten $(QP(\hat{x},\hat{\lambda},\hat{\mu}))$ als ungestörtes quadratisches Optimierungsproblem mit Nominalparameter $\hat{w} = (\hat{x},\hat{\lambda},\hat{\mu})$. Desweiteren bemerken wir, daß die KKT-Bedingungen für $(QP(\hat{w}))$ und das Standard-Optimierungsproblem im Fall $\hat{d} = 0$ übereinstimmen. Also ist $(0,\hat{\lambda},\hat{\mu})$ ein KKT-Punkt von $(QP(\hat{w}))$. Die Voraussetzungen (iii)-(v) garantieren, daß $\hat{d} = 0$ ein streng reguläres lokales Minimum von $(QP(\hat{w}))$ ist.

   Daher können wir den Sensitivitätssatz 5.4.2 anwenden: Es existiert eine Umgebung $V_\varepsilon(\hat{w})$ von $\hat{w}$, so daß $(QP(w))$ ein eindeutiges streng reguläres lokales Minimum $(d(w),\lambda(w),\mu(w))$ für jedes $w \in V_\varepsilon(\hat{w})$ besitzt. Dabei ist $(d(w),\lambda(w),\mu(w))$ stetig differenzierbar bezüglich w.

   Zusätzlich bleibt die Indexmenge der aktiven Nebenbedingungen in der Umgebung konstant: $A(\hat{x}) = A_{QP}(\hat{d}=0,\hat{w}) = A_{QP}(d(w),w)$. $A_{QP}(d,w)$ bezeichnet dabei die Indexmenge der aktiven Ungleichungen von $(QP(w))$ in d.

2. Wegen der Stetigkeit der Nebenbedingungen können wir die in $\hat{x}$ inaktiven Nebenbedingungen des Standard-Optimierungsproblems lokal weglassen und erhalten das lokal äquivalente Problem

$$\min f(x) \quad \text{unter} \quad g_i(x) = 0,\ i \in A(\hat{x}),\ h_j(x) = 0,\ j = 1,\dots,p.$$

   Darauf können wir das Lagrange-Newton-Verfahren anwenden und unter den Voraussetzungen (i)-(v) liefert Satz 5.7.1 die lokal quadratische Konvergenz

$$(x^{[k]},\lambda^{[k]}_{A(\hat{x})},\mu^{[k]}) \to (\hat{x},\hat{\lambda}_{A(\hat{x})},\hat{\mu}).$$

   Beachte, daß die Multiplikatoren unter der Voraussetzung (iii) eindeutig sind. Wir ergänzen $\lambda_i^{[k]} = 0$ für $i \notin A(\hat{x})$ und erhalten

$$w^{[k]} := (x^{[k]},\lambda^{[k]},\mu^{[k]}) \to \hat{w} = (\hat{x},\hat{\lambda},\hat{\mu}).$$

3. Es bezeichne δ den Konvergenzradius des Lagrange-Newton-Verfahrens. Sei $r := \min\{\varepsilon,\delta\}$. Für $w^{[0]} \in U_r(\hat{w})$ bleiben alle folgenden Iterierten $w^{[k]}$ des Lagrange-Newton-Verfahrens in dieser Umgebung. Weiter erfüllen $(d^{[k]},\lambda^{[k+1]},\mu^{[k+1]})$ mit $d^{[k]} = x^{[k+1]} - x^{[k]}$ die notwendigen Bedingungen von $(QP(w^{[k]}))$, vgl. (5.34) und (5.35). Gemäß 1. ist die Lösung $(d(w^{[k]}),\lambda(w^{[k]}),\mu(w^{[k]}))$ von $(QP(w^{[k]}))$ eindeutig. Also fallen die SQP-Iteration und die Lagrange-Newton-Iteration zusammen.

□

Bemerkung 5.8.3 (Approximation der Hessematrix)

Die Verwendung der exakten Hessematrix $\nabla^2_{xx}L$ der Lagrange-Funktion im QP-Problem hat zwei Nachteile:

- *In vielen Anwendungen ist die Hessematrix nicht explizit bekannt. Die numerische Approximation durch finite Differenzen ist sehr aufwendig und ungenau.*
- *Die Hessematrix kann indefinit sein. Dies erschwert die Lösung der QP-Hilfsprobleme erheblich. Es ist daher wünschenswert, die Hessematrix durch eine positiv definite Matrix zu ersetzen (→ Quasi-Newton-Verfahren).*

In der Praxis wird die Hessematrix der Lagrange-Funktion in Iteration k durch eine geeignete Matrix H_k ersetzt. Powell [Pow78] schlug vor, die modifizierte BFGS-Update-Formel

$$H_{k+1} = H_k + \frac{q^{[k]}(q^{[k]})^\top}{(q^{[k]})^\top d^{[k]}} - \frac{H_k d^{[k]}(d^{[k]})^\top H_k}{(d^{[k]})^\top H_k d^{[k]}}, \tag{5.36}$$

mit

$$\begin{aligned}
d^{[k]} &= x^{[k+1]} - x^{[k]},\\
q^{[k]} &= \theta_k y^{[k]} + (1-\theta_k) H_k d^{[k]},\\
y^{[k]} &= \nabla_x L(x^{[k+1]}, \lambda^{[k]}, \mu^{[k]}) - \nabla_x L(x^{[k]}, \lambda^{[k]}, \mu^{[k]}),\\
\theta_k &= \begin{cases} 1, & \text{falls } (d^{[k]})^\top y^{[k]} \geq 0.2 (d^{[k]})^\top H_k d^{[k]},\\ \frac{0.8 (d^{[k]})^\top H_k d^{[k]}}{(d^{[k]})^\top H_k d^{[k]} - (d^{[k]})^\top y^{[k]}}, & \text{sonst} \end{cases}
\end{aligned}$$

zu verwenden. Diese Update-Formel garantiert, daß H_{k+1} symmetrisch und positiv definit bleibt, wenn H_k symmetrisch und positiv definit war. Für $\theta_k = 1$ entsteht die BFGS-Formel, welche schon bei den Quasi-Newton-Verfahren verwendet wurde (allerdings mußte dort durch Wahl einer geeigneten Schrittweitenstrategie noch die Bedingung $(d^{[k]})^\top y^{[k]} > 0$ garantiert werden).
Wird die modifizierte BFGS-Update-Formel verwendet, kann immerhin noch superlineare Konvergenz nachgewiesen werden.

5.8.1 Globalisierung des SQP-Verfahrens

Das Konvergenzresultat zeigt, daß das SQP-Verfahren für alle Startwerte, die in einer Umgebung eines lokalen Minimums liegen, konvergent ist. In der Praxis ist diese Umgebung jedoch unbekannt und kann sehr klein sein. Daher ist es notwendig, das SQP-Verfahren zu globalisieren, so daß es (unter geeigneten Bedingungen) für beliebige Startwerte konvergiert. Wie im unrestringierten Fall wird dies durch Einführung einer Schrittweite $\alpha_k > 0$ erreicht. Die neue Iterierte ist gegeben durch

$$x^{[k+1]} = x^{[k]} + \alpha_k d^{[k]},$$

wobei $d^{[k]}$ wie zuvor ein quadratisches Hilfsproblem löst. Zur Bestimmung der Schrittweite α_k wird wieder eine eindimensionale **Liniensuche** in Richtung $d^{[k]}$ durchgeführt.

Problem: Wann ist $x^{[k+1]}$ „besser“ als $x^{[k]}$?

Im unrestringierten Fall konnte diese Frage leicht durch einen Vergleich der Zielfunktionswerte beantwortet werden: $x^{[k+1]}$ ist besser als $x^{[k]}$, wenn $f(x^{[k+1]}) < f(x^{[k]})$ gilt.
Im restringierten Fall ist dies nicht mehr so einfach, da die Iterierten $x^{[k]}$ des SQP-Verfahrens i.a. unzulässig sind. Eine Verbesserung kann also sowohl an Hand der Zielfunktionswerte als auch an Hand der Verletzungen der Nebenbedingungen gemessen werden. Dies führt auf sogenannte **Bewertungsfunktionen** (**penalty function, merit function**). An Hand der Werte der Bewertungsfunktion wird es möglich zu entscheiden, ob die neue Iterierte $x^{[k+1]}$ in gewissem Sinne „besser“ ist als die alte Iterierte $x^{[k]}$. Dabei ist die neue Iterierte besser als die alte, falls entweder ein hinreichender Abstieg in der Zielfunktion f oder eine weniger starke Verletzung der Nebenbedingungen erreicht wird, wobei sich das jeweils andere Kriterium nicht substantiell verschlechtern darf.
Eine allgemeine Klasse von Bewertungsfunktionen wird durch

$$P_r(x;\eta) := f(x) + \eta \cdot r(x) \tag{5.37}$$

definiert, wobei $\eta > 0$ einen Gewichtungsparameter und $r : \mathbb{R}^n \to [0,\infty)$ eine stetige Funktion mit der Eigenschaft

$$r(x) = 0 \qquad \Leftrightarrow \qquad x \in \Sigma$$

bezeichnen.

Beispiel 5.8.4 (Bewertungsfunktion)
Eine typische Bewertungsfunktion für das Standard-Optimierungsproblem (mit $S = \mathbb{R}^n$), die auf der 1-Norm basiert, ist die l_1-Bewertungsfunktion:

$$l_1(x;\eta) := f(x) + \eta \left(\sum_{i=1}^{m} \max\{0, g_i(x)\} + \sum_{j=1}^{p} |h_j(x)| \right), \qquad \eta > 0.$$

Beachte, daß unzulässige Punkte $x \notin \Sigma$ durch die Terme

$$\sum_{i=1}^{m} \max\{0, g_i(x)\} + \sum_{j=1}^{p} |h_j(x)| > 0$$

bestraft werden.
Allgemeinere Bewertungsfunktionen basieren auf der q-Norm:

$$l_q(x;\eta) = f(x) + \eta \left(\sum_{i=1}^{m} (\max\{0, g_i(x)\})^q + \sum_{j=1}^{p} |h_j(x)|^q \right)^{1/q}, \ 1 \le q < \infty,$$

und

$$l_\infty(x;\eta) = f(x) + \eta \max\{0, g_1(x), \ldots, g_m(x), |h_1(x)|, \ldots, |h_p(x)|\}.$$

Von besonderem Interesse sind die sogenannten exakten Bewertungsfunktionen, da für diese Bewertungsfunktionen lokale Minima des restringierten Ausgangsproblems auch lokale Minima der unrestringierten Bewertungsfunktion sind. Für diese Bewertungsfunktionen besteht ein Ansatz zur Lösung des restringierten Problems darin, stattdessen die unrestringierte Bewertungsfunktion zu minieren.

Definition 5.8.5 **(Exakte Bewertungsfunktion)**
Die Bewertungsfunktion $P_r(x;\eta)$ in (5.37) heißt **exakt in einem lokalen Minimum** $\hat{x}$ **des Standard-Optimierungsproblems (mit** $S = \mathbb{R}^n$**)**, *falls es einen* **endlichen** *(!) Parameter $\hat{\eta} > 0$ gibt, so daß $\hat{x}$ ein lokales Minimum von $P_r(\cdot;\eta)$ für alle $\eta \geq \hat{\eta}$ ist.*

Beispiel 5.8.6
Abbildung 5.6 zeigt die l_1-Bewertungsfunktion für verschiedene Werte von η für das Optimierungsproblem mit den Daten

$$\begin{aligned} f(x,y) &= (x-2)^2 + (y-3)^2, \\ h(x,y) &= y + \frac{x}{2} - \frac{1}{2}, \\ g_1(x,y) &= y + 2x^2 - 2, \\ g_2(x,y) &= x^2 - y - 1. \end{aligned}$$

Die optimale Lösung ist gegeben durch $\hat{x} = (3/5, 1/5)^\top$, $\hat{\lambda} = (0,0)^\top$ und $\hat{\mu} = 28/5$. Die Restriktionen g_1 und g_2 sind nicht aktiv in $\hat{x}$.

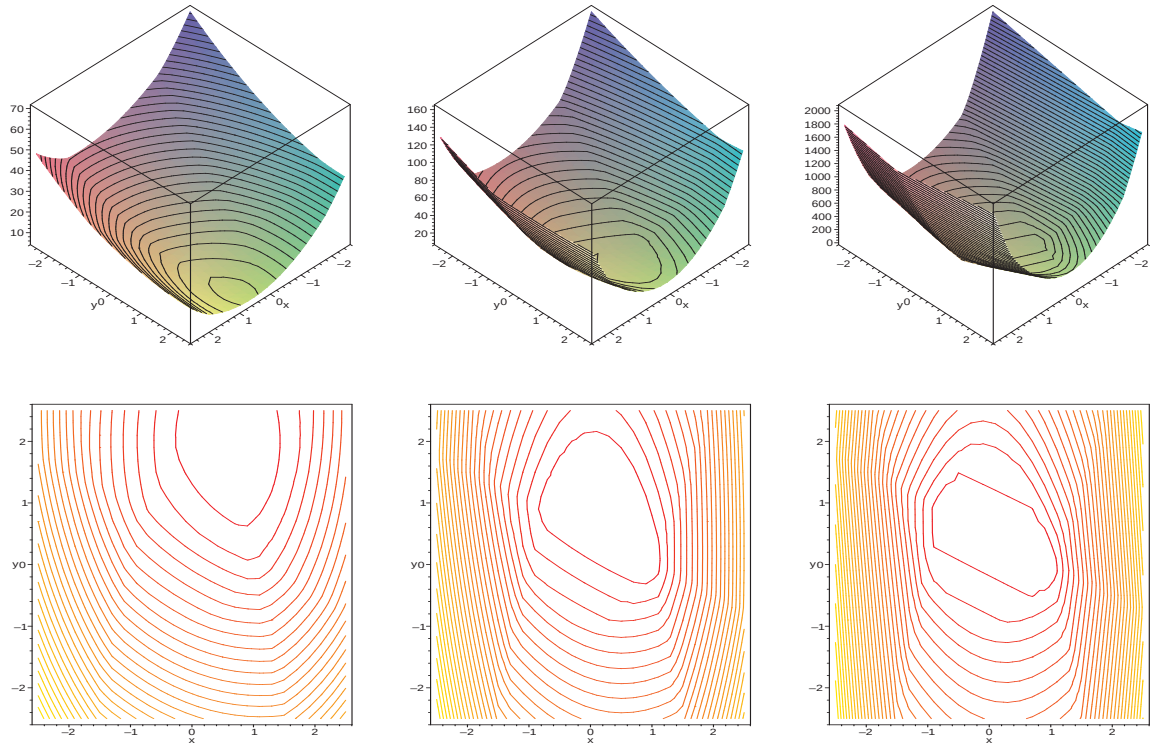


Abbildung 5.6: 3D-Darstellung (oben) und Höhenlinien (unten) der l_1-Bewertungsfunktion für $\eta = 1$ (links), $\eta = 28/5$ (Mitte) und $\eta = 100$ (rechts).

Dummerweise kann gezeigt werden, daß die Bewertungsfunktion $P_r(x;\eta)$ aus (5.37) in einem lokalen Minimum $\hat{x}$ **nicht differenzierbar** ist, falls sie exakt ist und $\nabla f(\hat{x}) \neq 0$ gilt. Beachte, daß die Bedingung $\nabla f(\hat{x}) \neq 0$ den Normalfall in der restringierten Optimierung darstellt.
Der folgende Satz sagt aus, daß die Bewertungsfunktionen l_q für $1 \leq q \leq \infty$ exakt sind, wenn eine Regularitätsbedingung gilt.

Satz 5.8.7

Sei $\hat{x} \in \Sigma$ ein isoliertes lokales Minimum des Standard-Optimierungsproblems mit $S = \mathbb{R}^n$, welches die Regularitätsbedingung von Mangasarian-Fromowitz erfüllt. Dann ist l_q exakt für $1 \leq q \leq \infty$.

Beweis: Geiger and Kanzow [GK02], S. 225 □

Satz 5.8.7 legt es nahe, das restringierte Standard-Optimierungsproblem durch das unrestringierte Minimierungsproblem

$$\min_{x \in \mathbb{R}^n} l_q(x;\eta)$$

mit hinreichend großem $\eta > 0$ zu ersetzen. Diese Idee wird im SQP-Verfahren ausge-

nutzt, um eine Schrittweite α mittels eindimensionaler Liniensuche für $l_q(x^{[k]} + \alpha d^{[k]}; \eta)$ durchzuführen. Im folgenden beschränken wir uns auf die l_1-Bewertungsfunktion. Wie oben erwähnt, ist die exakte l_1-Bewertungsfunktion nicht differenzierbar. Allerdings ist sie immerhin noch richtungsdifferenzierbar, d.h. der Grenzwert

$$l_1'(x; d; \eta) := \lim_{h \downarrow 0} \frac{l_1(x + hd; \eta) - l_1(x; \eta)}{h}.$$

existiert für alle $x \in \mathbb{R}^n$, $d \in \mathbb{R}^n$ (vgl. Geiger und Kanzow [GK02], S. 252):

Hilfssatz 5.8.8

Seien f, g_i, $i = 1, \ldots, m$, h_j, $j = 1, \ldots, p$ stetig differenzierbar. Die Richtungsableitung der exakten l_1-Penaltyfunktion in $x \in \mathbb{R}^n$ in Richtung $d \in \mathbb{R}^n$ ist gegeben durch

$$\begin{aligned} l_1'(x; d; \eta) \;=\; & \nabla f(x)^\top d + \eta \sum_{i \,:\, g_i(x) > 0} \nabla g_i(x)^\top d + \eta \sum_{i \,:\, g_i(x) = 0} \max\{0, \nabla g_i(x)^\top d\} \\ & + \eta \sum_{j \,:\, h_j(x) > 0} \nabla h_j(x)^\top d - \eta \sum_{j \,:\, h_j(x) < 0} \nabla h_j(x)^\top d \\ & + \eta \sum_{j \,:\, h_j(x) = 0} |\nabla h_j(x)^\top d| \end{aligned}$$

Das SQP-Verfahren sei bis zur Iteration k fortgeschritten und $d^{[k]}$ sei die optimale Lösung des QP-Hilfsproblems. Falls nun

$$l_1'(x^{[k]}; d^{[k]}; \eta) < 0$$

gilt und $d^{[k]}$ somit eine Abstiegsrichtung von l_1 in $x^{[k]}$ ist, kann z.B. mit dem Armijo-Verfahren eine eindimensionale Liniensuche für die Funktion

$$\varphi(\alpha) := l_1(x^{[k]} + \alpha d^{[k]}; \eta), \qquad \alpha \geq 0$$

durchgeführt werden. Tatsächlich ist die Lösung $d^{[k]}$ des QP-Problems unter bestimmten Bedingungen eine Abstiegsrichtung von l_1 in $x^{[k]}$.

Satz 5.8.9

Sei $(d^{[k]}, \lambda^{[k+1]}, \mu^{[k+1]})$ mit $d^{[k]} \neq 0$ ein KKT-Punkt des quadratischen Hilfsproblems $(QP(x^{[k]}, \lambda^{[k]}, \mu^{[k]}))$, wobei die Hessematrix $\nabla^2_{xx} L$ durch eine symmetrische und positiv definite Matrix $H_k \in \mathbb{R}^{n \times n}$ ersetzt sei. Desweiteren gelte

$$\eta \geq \max\{\lambda_1^{[k+1]}, \ldots, \lambda_m^{[k+1]}, |\mu_1^{[k+1]}|, \ldots, |\mu_p^{[k+1]}|\}. \tag{5.38}$$

Dann gilt

$$l_1'(x^{[k]}; d^{[k]}; \eta) \leq -(d^{[k]})^\top H_k d^{[k]} < 0,$$

d.h. $d^{[k]}$ ist Abstiegsrichtung von l_1 in $x^{[k]}$.

Beweis: Zur Vereinfachung seien $d := d^{[k]} \neq 0$, $x = x^{[k]}$, $\lambda := \lambda^{[k+1]}$ und $\mu := \mu^{[k+1]}$. Aus den KKT-Bedingungen des quadratischen Teilproblems erhalten wir

$$\nabla f(x)^\top d = -d^\top H_k d - \sum_{i=1}^m \lambda_i \nabla g_i(x)^\top d - \sum_{j=1}^p \mu_j \nabla h_j(x)^\top d.$$

Die Komplementaritätsbedingungen liefern

$$\lambda_i \nabla g_i(x)^\top d = -\lambda_i g_i(x).$$

Aus den Nebenbedingungen des QP-Problems folgen

$$\begin{aligned} \nabla g_i(x)^\top d &\leq -g_i(x), \\ \mu_j \nabla h_j(x)^\top d &= -\mu_j h_j(x). \end{aligned}$$

Einsetzen dieser Beziehungen in $l_1'(x;d;\eta)$ liefert

$$\begin{aligned} l_1'(x;d;\eta) &= -d^\top H_k d - \sum_{i=1}^m \lambda_i \nabla g_i(x)^\top d - \sum_{j=1}^p \mu_j \nabla h_j(x)^\top d \\ &\quad + \eta \sum_{i:g_i(x)>0} \nabla g_i(x)^\top d + \eta \sum_{i:g_i(x)=0} \max\{0, \nabla g_i(x)^\top d\} \\ &\quad + \eta \sum_{j:h_j(x)>0} \nabla h_j(x)^\top d - \eta \sum_{j:h_j(x)<0} \nabla h_j(x)^\top d + \eta \sum_{j:h_j(x)=0} |\nabla h_j(x)^\top d| \\ &= -d^\top H_k d + \sum_{i=1}^m \lambda_i g_i(x) + \sum_{j=1}^p \mu_j h_j(x) \\ &\quad + \eta \sum_{i:g_i(x)>0} \nabla g_i(x)^\top d + \eta \sum_{i:g_i(x)=0} \max\{0, \nabla g_i(x)^\top d\} \\ &\quad + \eta \sum_{j:h_j(x)>0} \nabla h_j(x)^\top d - \eta \sum_{j:h_j(x)<0} \nabla h_j(x)^\top d + \eta \sum_{j:h_j(x)=0} |\nabla h_j(x)^\top d| \\ &= -d^\top H_k d + \sum_{i:g_i(x)<0} \underbrace{\lambda_i}_{\geq 0} \underbrace{g_i(x)}_{<0} \\ &\quad + \sum_{i:g_i(x)>0} \big(\eta \nabla g_i(x)^\top d + \lambda_i g_i(x)\big) + \eta \sum_{i:g_i(x)=0} \max\{0, \underbrace{g_i(x) + \nabla g_i(x)^\top d}_{\leq 0}\} \\ &\quad + \sum_{j:h_j(x)>0} \big(\eta \nabla h_j(x)^\top d + \mu_j h_j(x)\big) - \sum_{j:h_j(x)<0} \big(\eta \nabla h_j(x)^\top d - \mu_j h_j(x)\big) \\ &\quad + \eta \sum_{j:h_j(x)=0} |\underbrace{h_j(x) + \nabla h_j(x)^\top d}_{=0}| \\ &\leq -d^\top H_k d + \sum_{i:g_i(x)>0} \big(\eta \nabla g_i(x)^\top d + \lambda_i g_i(x)\big) \\ &\quad + \sum_{j:h_j(x)>0} \big(\eta \nabla h_j(x)^\top d + \mu_j h_j(x)\big) - \sum_{j:h_j(x)<0} \big(\eta \nabla h_j(x)^\top d - \mu_j h_j(x)\big). \end{aligned}$$

Beachte

$$\begin{aligned} g_i(x) + \nabla g_i(x)^\top d \le 0 \quad &\overset{\eta \ge 0}{\Rightarrow} \quad \eta \nabla g_i(x)^\top d \le -\eta g_i(x), \\ h_j(x) + \nabla h_j(x)^\top d = 0 \quad &\overset{\eta \ge 0}{\Rightarrow} \quad \eta \nabla h_j(x)^\top d = -\eta h_j(x). \end{aligned}$$

Ausnutzen dieser Beziehungen liefert

$$\begin{aligned} l_1'(x; d; \eta) &\le -d^\top H_k d + \sum_{i: g_i(x) > 0} \underbrace{(\lambda_i - \eta)}_{\le 0} \underbrace{g_i(x)}_{>0} \\ &\quad + \sum_{j: h_j(x) > 0} \underbrace{(\mu_j - \eta)}_{\le 0} \underbrace{h_j(x)}_{>0} - \sum_{j: h_j(x) < 0} \underbrace{(-\mu_j - \eta)}_{\le 0} \underbrace{h_j(x)}_{<0} \\ &\le -d^\top H_k d. \end{aligned}$$

□

Satz 5.8.9 zeigt, wie η gewählt werden muß, damit $d^{[k]} \neq 0$ eine Abstiegsrichtung ist (im Fall $d^{[k]} = 0$ sind bereits die KKT-Bedingungen des Ausgangsproblems erfüllt). Man wird η iterativ anpassen müssen, damit (5.38) erfüllt ist, etwa in der Form:

$$\eta_{k+1} := \max\{\eta_k, \max\{\lambda_1^{[k+1]}, \dots, \lambda_m^{[k+1]}, |\mu_1^{[k+1]}|, \dots, |\mu_p^{[k+1]}|\} + \varepsilon\}, \tag{5.39}$$

wobei $\varepsilon \ge 0$ ist.

Insgesamt erhalten wir das globale SQP-Verfahren:

Algorithmus: Globalisiertes SQP-Verfahren

(i) Wähle Startwerte $(x^{[0]}, \lambda^{[0]}, \mu^{[0]}) \in \mathbb{R}^n \times \mathbb{R}^m \times \mathbb{R}^p$, $H_0 \in \mathbb{R}^{n\times n}$ symmetrisch und positiv definit, $\beta \in (0,1)$, $\sigma \in (0,1)$ und setze $k = 0$.

(ii) Falls $(x^{[k]}, \lambda^{[k]}, \mu^{[k]})$ ein KKT-Punkt des Standard-Optimierungsproblems ist, STOP.

(iii) **QP-Hilfsproblem:**
Berechne einen KKT-Punkt $(d^{[k]}, \lambda^{[k+1]}, \mu^{[k+1]}) \in \mathbb{R}^n \times \mathbb{R}^m \times \mathbb{R}^p$ des quadratischen Hilfsproblems $(QP(x^{[k]}, \lambda^{[k]}, \mu^{[k]}))$, wobei die Hessematrix L''_{xx} durch die modifizierte BFGS-Update-Matrix H_k ersetzt ist.

(iv) Passe η gemäß (5.39) an.

(v) **Armijo-Regel:**
Bestimme eine Schrittweite $\alpha_k = \max\{\beta^j \mid j = 0, 1, 2, \ldots\}$ mit

$$l_1(x^{[k]} + \alpha_k d^{[k]}; \eta) \le l_1(x^{[k]}; \eta) + \sigma\alpha_k l_1'(x^{[k]}; d^{[k]}; \eta).$$

(vi) **Modifizierter BFGS-Update:**
Berechne H_{k+1} gemäß der Update-Formel (5.36).

(vii) Setze $x^{[k+1]} := x^{[k]} + \alpha_k d^{[k]}$, $k := k + 1$ und gehe zu (ii).

Beispiel 5.8.10

Minimiere

$$f(x) = -x_2x_6 + x_1x_7 - x_3x_7 - x_5x_8 + x_4x_9 + x_3x_8$$

bzgl. $x \in \mathbb{R}^9$ *unter den Nebenbedingungen*

$$x_1 \ge 0,\ -1 \le x_3 \le 1,\ x_5 \ge 0,\ x_6 \ge 0,\ x_7 \ge 0,\ x_8 \le 0,\ x_9 \le 0$$

und

$$x_2 - x_1 \ge 0, \quad x_3 - x_2 \ge 0, \quad x_3 - x_4 \ge 0, \quad x_4 - x_5 \ge 0$$

und

$$\begin{aligned}
x_1^2 + x_6^2 &\leq 1,\\
(x_2 - x_1)^2 + (x_7 - x_6)^2 &\leq 1,\\
(x_3 - x_1)^2 + x_6^2 &\leq 1,\\
(x_1 - x_4)^2 + (x_6 - x_8)^2 &\leq 1,\\
(x_1 - x_5)^2 + (x_6 - x_9)^2 &\leq 1,\\
x_2^2 + x_7^2 &\leq 1,\\
(x_3 - x_2)^2 + x_7^2 &\leq 1,\\
(x_4 - x_2)^2 + (x_8 - x_7)^2 &\leq 1,\\
(x_2 - x_5)^2 + (x_7 - x_9)^2 &\leq 1,\\
(x_4 - x_3)^2 + x_8^2 &\leq 1,\\
(x_5 - x_3)^2 + x_9^2 &\leq 1,\\
x_4^2 + x_8^2 &\leq 1,\\
(x_4 - x_5)^2 + (x_9 - x_8)^2 &\leq 1,\\
x_5^2 + x_9^2 &\leq 1.
\end{aligned}$$

Startschätzung:

$$x^{[0]} = (0.1, 0.125, 2/3, 0.142857, 1/9, 0.2, 0.25, -0.2, -0.25)^\top.$$

Die Startschätzung der Multiplikatoren ist Null.

Ausgabe des SQP-Verfahrens:

```
---------------------------------------- SQP VERSION 1.1 (C) Matthias Gerdts, University of Bayreuth, 2004 ----------------------------------------

      NUMBER OF VARIABLES              :          9
      NUMBER OF CONSTRAINTS            :         18
      METHOD                           :  SEQUENTIAL QUADRATIC PROGRAMMING (SQP)
      MERIT FUNCTION                   :  L1-PENALTY FUNCTION
      MULTIPLIER UPDATE RULE           :  POWELL
      OPTIMALITY TOLERANCE             :  0.149E-07
      FEASIBILITY TOLERANCE            :  0.100E-11
      LINE SEARCH PARAMETER            :  SIGMA=  0.100E+00 BETA=  0.900E+00
      MAXIMUM NUMBER OF ITERATIONS     :      10000
      INFINITY                         :  0.100E+21
      ROUNDOFF TOLERANCE               :  0.300E-12
      REAL WORK SPACE PROVIDED         :       5338  NEEDED :       5338
      INTEGER WORK SPACE PROVIDED      :         82  NEEDED :         82

------------------------------------------------------------------------------------------------------------------------------------------------
 ITER    QPIT     ALPHA                OBJ                NB          KKT         PEN         |D|         DELTA       RDELTA        F/G
------------------------------------------------------------------------------------------------------------------------------------------------
   0       0    0.0000E+00    -0.3134920277777778E+00  0.0000E+00  0.5667E+00  0.4243E+01  0.0000E+00  0.0000E+00  0.1000E+01     1/   1    i
   1       7    0.1000E+01    -0.1593788677043482E+01  0.7834E+00  0.1217E+01  0.2138E+01  0.8139E+00  0.0000E+00  0.1000E+01     2/   2    i
   2      10    0.1000E+01    -0.1520709520745054E+01  0.3302E+00  0.2969E+00  0.1236E+01  0.4374E+00  0.3790E+00  0.1000E+01     3/   3    i
   3       3    0.1000E+01    -0.1388511910024717E+01  0.7326E-01  0.7589E-01  0.7931E+00  0.1180E+00  0.1801E+00  0.1000E+01     4/   4    i
   4       3    0.1000E+01    -0.1352359503836702E+01  0.4336E-02  0.9657E-02  0.6253E+00  0.4386E-01  0.3948E-01  0.1000E+01     5/   5    i
   5       3    0.1000E+01    -0.1349928698009139E+01  0.2504E-04  0.4006E-02  0.5547E+00  0.7860E-02  0.2493E-02  0.1000E+01     6/   6    i
   6       3    0.1000E+01    -0.1350018091962160E+01  0.7467E-04  0.3736E-02  0.5284E+00  0.1358E-01  0.1478E-04  0.1000E+01     7/   7    i
   7       3    0.1000E+01    -0.1349983850538642E+01  0.2758E-04  0.1499E-02  0.5174E+00  0.6417E-02  0.6179E-04  0.1000E+01     8/   8    i
   8       3    0.1000E+01    -0.1349963012462895E+01  0.2394E-06  0.2028E-03  0.5128E+00  0.4120E-03  0.2106E-04  0.1000E+01     9/   9    i
   9       3    0.1000E+01    -0.1349962886083232E+01  0.8269E-09  0.1035E-04  0.5106E+00  0.4345E-04  0.1275E-06  0.1000E+01    10/  10    i
  10       3    0.1000E+01    -0.1349962885862414E+01  0.3006E-11  0.3058E-06  0.5096E+00  0.2714E-05  0.2240E-09  0.1000E+01    11/  11    i
  11       3    0.1000E+01    -0.1349962885860211E+01  0.2665E-14  0.4422E-08  0.5091E+00  0.5736E-07  0.2205E-11  0.1000E+01    12/  12    i
------------------------------------------------------------------------------------------------------------------------------------------------
KKT CONDITIONS SATISFIED (IER= 0)!
SOLUTION:

OBJ =       -0.1349962885860211E+01
KKT =        0.4421887098724018E-08
CON =        0.266453525910037 6E-14
X   =
             0.6094665336054564E-01
             0.5976493035302869E+00
             0.1000000000000000E+01
             0.5976493034306842E+00
             0.6094665324738306E-01
```

```
            0.3437714533890817E+00
            0.5000000000868919E+00
           -0.4999999999131094E+00
           -0.3437714530799649E+00
CONSTRAINT    LB                      VALUE                   UB                      STATUS      LAMBDA
    1          0.0000000000000000E+00  0.6094665336054564E-01  0.1000000000000000E+21  IA          0.0000000000000000E+00
    2         -0.1000000000000000E+21  0.5976493035302869E+00  0.1000000000000000E+21  IA          0.0000000000000000E+00
    3         -0.1000000000000000E+01  0.1000000000000000E+01  0.1000000000000000E+01  UB          0.6875429052209211E+00
    4         -0.1000000000000000E+21  0.5976493034306842E+00  0.1000000000000000E+21  IA          0.0000000000000000E+00
    5          0.0000000000000000E+00  0.6094665324738306E-01  0.1000000000000000E+21  IA          0.0000000000000000E+00
    6          0.0000000000000000E+00  0.3437714533890817E+00  0.1000000000000000E+21  IA          0.0000000000000000E+00
    7          0.0000000000000000E+00  0.5000000000868919E+00  0.1000000000000000E+21  IA          0.0000000000000000E+00
    8         -0.1000000000000000E+21  -0.4999999999131094E+00 0.0000000000000000E+00  IA          0.0000000000000000E+00
    9         -0.1000000000000000E+21  -0.3437714530799649E+00 0.0000000000000000E+00  IA          0.0000000000000000E+00
CONSTRAINT    LB                      VALUE                   UB                      STATUS      LAMBDA
   10          0.0000000000000000E+00  0.5367026501697413E+00  0.1000000000000000E+21  IA          0.0000000000000000E+00
   11          0.0000000000000000E+00  0.4023506964697131E+00  0.1000000000000000E+21  IA          0.0000000000000000E+00
   12          0.0000000000000000E+00  0.4023506965693158E+00  0.1000000000000000E+21  IA          0.0000000000000000E+00
   13          0.0000000000000000E+00  0.5367026501833012E+00  0.1000000000000000E+21  IA          0.0000000000000000E+00
   14         -0.1000000000000000E+21  0.1218933067210921E+00  0.1000000000000000E+01  IA          0.0000000000000000E+00
   15         -0.1000000000000000E+21  0.3124570935025335E+00  0.1000000000000000E+01  IA          0.0000000000000000E+00
   16         -0.1000000000000000E+21  0.1000000000000001E+01  0.1000000000000000E+01  UB          0.8318406155170190E-01
   17         -0.1000000000000000E+21  0.1000000000000001E+01  0.1000000000000000E+01  UB          0.3202624887781933E+00
   18         -0.1000000000000000E+21  0.4727152482359042E+00  0.1000000000000000E+01  IA          0.0000000000000000E+00
   19         -0.1000000000000000E+21  0.6071846900971289E+00  0.1000000000000000E+01  IA          0.0000000000000000E+00
   20         -0.1000000000000000E+21  0.4118860830365552E+00  0.1000000000000000E+01  IA          0.0000000000000000E+00
   21         -0.1000000000000000E+21  0.1000000000000003E+01  0.1000000000000000E+01  UB          0.1992983286519636E+00
   22         -0.1000000000000000E+21  0.1000000000000002E+01  0.1000000000000000E+01  UB          0.3202624872860010E+00
   23         -0.1000000000000000E+21  0.4118860829429231E+00  0.1000000000000000E+01  IA          0.0000000000000000E+00
   24         -0.1000000000000000E+21  0.1000000000000001E+01  0.1000000000000000E+01  UB          0.8318406609821495E-01
   25         -0.1000000000000000E+21  0.6071846898042915E+00  0.1000000000000000E+01  IA          0.0000000000000000E+00
   26         -0.1000000000000000E+21  0.3124570935593750E+00  0.1000000000000000E+01  IA          0.0000000000000000E+00
   27         -0.1000000000000000E+21  0.1218933064947673E+00  0.1000000000000000E+01  IA          0.0000000000000000E+00
```

Bemerkung 5.8.11

- *Für einen Konvergenzbeweis sei auf Han [Han77] verwiesen. Im günstigsten Fall geht dabei das globale SQP-Verfahren nach endlich vielen Schritten in das lokale über, welches (unter entsprechenden Voraussetzungen) mindestens superlinear konvergiert. Dies bedeutet, daß dann die Schrittweite $\alpha_k = 1$ bei der Armijo-Regel akzeptiert wird.*

  Es gibt jedoch Beispiele bei denen die Schrittweite $\alpha_k = 1$ nicht akzeptiert wird und somit die superlineare Konvergenz verhindert wird. Dieser Effekt geht auf Maratos zurück und heißt daher **Maratos-Effekt**. *Strategien zur Vermeidung des Effekts sind in Geiger und Kanzow [GK02] ab S. 258 beschrieben.*

- *Es gibt auch differenzierbare exakte Bewertungsfunktionen, diese sind allerdings nicht von der Gestalt in (5.37). Eine häufig benutzte differenzierbare exakte Bewertungsfunktion für das Standard-Optimierungsproblem (mit $S = \mathbb{R}^n$) ist die* **erweiterte Lagrange-Funktion**

$$\begin{aligned} L_a(x,\lambda,\mu;\eta) &= f(x) + \mu^\top h(x) + \frac{\eta}{2}\|h(x)\|^2 \\ &\quad + \frac{1}{2\eta}\sum_{i=1}^{m}\left(\left(\max\{0,\lambda_i+\eta g_i(x)\}\right)^2 - \lambda_i^2\right) \\ &= f(x) + \sum_{j=1}^{p}\left(\mu_j h_j(x) + \frac{\eta}{2}h_j(x)^2\right) \\ &\quad + \sum_{i=1}^{m}\begin{cases} \lambda_i g_i(x) + \frac{\eta}{2}g_i(x)^2, & \text{falls } \lambda_i + \eta g_i(x) \geq 0, \\ -\frac{\lambda_i^2}{2\eta}, & \text{sonst.} \end{cases} \end{aligned} \tag{5.40}$$

  Ein SQP-Verfahren unter Verwendung der erweiterten Lagrange-Funktion wird in

Schittkowsi [Sch81, Sch83] diskutiert. Hierbei wird eine Liniensuche für die Funktion

$$\varphi(\alpha) := L_a\left(\begin{pmatrix} x^{[k]} \\ \lambda^{[k]} \\ \mu^{[k]} \end{pmatrix} + \alpha_k \begin{pmatrix} d^{[k]} \\ \lambda_{QP} - \lambda^{[k]} \\ \mu_{QP} - \mu^{[k]} \end{pmatrix}; \eta\right)$$

durchgeführt, wobei λ_{QP} und μ_{QP} die Multiplikatoren des QP-Problems bezeichnen. Die neuen Iterierten lauten

$$\begin{pmatrix} x^{[k+1]} \\ \lambda^{[k+1]} \\ \mu^{[k+1]} \end{pmatrix} = \begin{pmatrix} x^{[k]} \\ \lambda^{[k]} \\ \mu^{[k]} \end{pmatrix} + \alpha_k \begin{pmatrix} d^{[k]} \\ \lambda_{QP} - \lambda^{[k]} \\ \mu_{QP} - \mu^{[k]} \end{pmatrix}.$$

- *In praktischen Anwendungen wird anstatt eines einzelnen Gewichtungsparameters η jeder Summand der Strafterme in der Bewertungsfunktion individuell gewichtet, etwa durch η_i, $i = 1, \dots, m$ und $\hat{\eta}_j$, $j = 1, \dots, p$. Powell [Pow78] schlug folgende Update-Formel vor:*

$$\begin{aligned} \eta_i^{[k+1]} &:= \max\{|\lambda_i^{[k+1]}|, \frac{1}{2}(\eta_i^{[k]} + |\lambda_i^{[k+1]}|)\}, \qquad i = 1, \dots, m, \\ \hat{\eta}_j^{[k+1]} &:= \max\{|\mu_j^{[k+1]}|, \frac{1}{2}(\hat{\eta}_j^{[k]} + |\mu_j^{[k+1]}|)\}, \qquad j = 1, \dots, p. \end{aligned}$$

  Diese Formel hat sich in der Praxis bewährt.

- *Anstatt eine eindimensionale Liniensuche für eine Bewertungsfunktion durchzuführen, kann auch der Trust-Region-Ansatz mit dem SQP-Ansatz gekoppelt werden. Dies führt zu* **Trust-Region-SQP-Verfahren**.

- *Eine aktuelle Entwicklung sind die sogenannten* **Filter-SQP-Verfahren**, *siehe Fletcher und Leyffer [FL02] und Fletcher et al. [FLT02]. Diese ersetzen die Liniensuche für eine Bewertungsfunktion durch ein Verfahren, welches gute Suchrichtungen nach einem geeigneten Kriterien aus einer Menge von möglichen Suchrichtungen „herausfiltert“.*

5.8.2 Inkonsistentes QP Problem

Bisher haben wir stets vorausgesetzt, daß das QP Problem eine Lösung besitzt. Dies muß aber nicht der Fall sein, wie das folgende Beispiel zeigt.

Beispiel 5.8.12

Betrachte die Nebenbedingung

$$g(x) = 1 - x^2 \leq 0$$

und $x^{[0]} = 0$. Im QP Problem erhalten wir die Nebenbedingung

$$g(x^{[0]}) + g'(x^{[0]}) \cdot d = 1 \leq 0.$$

Offensichtlich ist diese niemals erfüllt.

Powell [Pow78] schlug vor, die Nebenbedingungen des QP Problems zu relaxieren, so daß das relaxierte QP Problem zulässig ist. Das ursprüngliche QP Problem wird ersetzt durch das relaxierte QP Problem.

Relaxiertes QP Problem $(RQP(x^{[k]}, \lambda^{[k]}, \mu^{[k]}))$**:**

$$\begin{aligned} \min_{d\in\mathbb{R}^n,\, \delta\in[0,1]} \quad & \frac{1}{2} d^\top H_k d + \nabla f(x^{[k]})^\top d + \frac{\eta}{2}\delta^2 \\ \text{unter} \quad & g_i(x^{[k]})(1-\sigma_i\delta) + \nabla g_i(x^{[k]})^\top d \;\leq\; 0, \quad i = 1,\ldots,m, \\ & h_j(x^{[k]})(1-\delta) + \nabla h_j(x^{[k]})^\top d \;=\; 0, \quad j = 1,\ldots,p. \end{aligned}$$

Hierin ist

$$\sigma_i = \begin{cases} 0, & \text{falls } g_i(x^{[k]}) < 0, \\ 1, & \text{sonst,} \end{cases} \qquad i = 1,\ldots,m.$$

Der Punkt $d = 0$ und $\delta = 1$ ist stets zulässig für $(RQP(x^{[k]}, \lambda^{[k]}, \mu^{[k]}))$. Erfüllt die optimale Lösung (d, δ) von $(RQP(x^{[k]}, \lambda^{[k]}, \mu^{[k]}))$ die Beziehung $\delta = 0$, dann ist d auch optimal für das ursprüngliche QP Problem $(QP(x^{[k]}, \lambda^{[k]}, \mu^{[k]}))$. Um tatsächlich $\delta = 0$ zu erreichen, muß der Gewichtungsparameter η, der auch in der Bewertungsfunktion auftritt, hinreichend groß sein.

5.9 Penalty-Verfahren

Wir betrachten das gleichungsrestringierte Problem

$$\min f(x) \quad \text{unter} \quad x \in \Sigma$$

mit

$$\Sigma = \{x \in \mathbb{R}^n \mid h_j(x) = 0,\ j = 1,\ldots,p\}.$$

Darin seien die Funktionen $f : \mathbb{R}^n \to \mathbb{R}$ und $h_j : \mathbb{R}^n \to \mathbb{R}$, $j = 1,\ldots,p$ stetig.

Die Idee des Penalty-Verfahrens besteht darin, die Lösung $\hat{x}$ des Ausgangsproblems iterativ durch die Lösungen von unrestringierten Hilfsproblemen zu approximieren. Diese

Hilfsprobleme bestehen in der Minimierung der **Penalty-Funktion**

$$P(x;\eta) = f(x) + \frac{\eta}{2}\sum_{j=1}^{p}(h_j(x))^2$$

für geeignete Werte von $\eta > 0$. Durch die Ankopplung der Nebenbedingungen wird ein Verlassen des zulässigen Bereichs Σ „bestraft“. Die Konstante η stellt einen Gewichtungsfaktor dar, mit dessen Hilfe die „Stärke der Bestrafung“ gesteuert werden kann. Das Penalty-Verfahren ist durch folgenden Algorithmus gegeben.

Algorithmus: Penalty-Verfahren

(i) Wähle $\eta_0 > 0$ und setze $k = 0$.

(ii) Bestimme $x^{[k]}$ als Lösung von

$$\min_x P(x;\eta_k)$$

(iii) Ist $h(x^{[k]}) = 0$, STOP.

(iv) Bestimme $\eta_{k+1} > \eta_k$, setze $k := k+1$ und gehe zu (ii).

Da P i.a. nicht differenzierbar ist, werden für das Hilfsproblem in Schritt (ii) Verfahren der unrestringierten, nichtdifferenzierbaren Optimierung benötigt.
Es stellt sich natürlich die Frage, ob das Verfahren tatsächlich gegen eine Lösung des Ausgangsproblems konvergiert.

Satz 5.9.1

Seien f und h_j, $j = 1,\dots,p$ stetig und $\{\eta_k\}$ streng monoton wachsend mit $\eta_k \to \infty$. Die zulässige Menge $\Sigma := \{x \in \mathbb{R}^n \mid h_j(x) = 0, j = 1,\dots,p\}$ sei nichtleer, und $\{x^{[k]}\}$ sei eine durch den Algorithmus erzeugte Folge (Existenz vorausgesetzt). Dann gelten die folgenden Aussagen:

(a) $\{P(x^{[k]};\eta_k)\}$ ist monoton wachsend.

(b) $\{\|h(x^{[k]})\|\}$ ist monoton fallend.

(c) $\{f(x^{[k]})\}$ ist monoton wachsend.

(d) Es ist $\lim_{k\to\infty} h(x^{[k]}) = 0$.

(e) Jeder Häufungspunkt der Folge $\{x^{[k]}\}$ ist eine Lösung des Ausgangsproblems.

Beweis: (vgl. [GK02], S. 208)

(a) Aus $\eta_{k+1} > \eta_k$ und der Definition von $x^{[k]}$ folgt

$$P(x^{[k]};\eta_k) \le P(x^{[k+1]};\eta_k) \le P(x^{[k+1]};\eta_{k+1}).$$

(b) Es gilt

$$P(x^{[k]};\eta_k) + P(x^{[k+1]};\eta_{k+1}) \le P(x^{[k+1]};\eta_k) + P(x^{[k]};\eta_{k+1}).$$

Mit der Definition von P folgt

$$\eta_k\|h(x^{[k]})\|^2 + \eta_{k+1}\|h(x^{[k+1]})\|^2 \le \eta_k\|h(x^{[k+1]})\|^2 + \eta_{k+1}\|h(x^{[k]})\|^2$$

bzw.

$$(\eta_k - \eta_{k+1})\left(\|h(x^{[k]})\|^2 - \|h(x^{[k+1]})\|^2\right) \le 0.$$

Wegen $\eta_k < \eta_{k+1}$ folgt $\|h(x^{[k]})\|^2 \ge \|h(x^{[k+1]})\|^2$ für alle k.

(c) folgt aus $P(x^{[k]};\eta_k) \le P(x^{[k+1]};\eta_k)$ und Teil (b).

(d) Wegen $\Sigma \neq \emptyset$ folgt

$$f(x^{[k]}) \le P(x^{[k]};\eta_k) \le \inf_{x\in\Sigma} P(x;\eta_k) = \inf_{x\in\Sigma} f(x) =: \hat{f} < \infty.$$

Wegen $\eta_k \to \infty$ und $f(x^{[k]}) \ge f(x^{[0]})$ nach (c) folgt $\lim_{k\to\infty}\|h(x^{[k]})\| = 0$.

(e) Sei $\hat{x}$ Häufungspunkt der Folge $\{x^{[k]}\}$ und $\{x^{[k_j]}\}$, $j = 1, 2, \ldots$ eine gegen $\hat{x}$ konvergente Teilfolge. Nach (d) gilt $h(\hat{x}) = 0$, d.h. $\hat{x}$ ist zulässig. Ausserdem gilt

$$f(\hat{x}) = \lim_{j\to\infty} f(x^{[k_j]}) \le \lim_{j\to\infty} P(x^{[k_j]};\eta_{k_j}) \le \inf_{x\in\Sigma} f(x) =: \hat{f}.$$

Daraus folgt (e). □

Bemerkung 5.9.2

Da nur die Stetigkeit der auftretenden Funktionen benötigt wird, ist das Verfahren auch auf Problemstellungen mit Ungleichungsnebenbedingungen

$$g_i(x) \le 0, \quad i = 1, \ldots, m$$

anwendbar. Denn diese Nebenbedingungen können äquivalent als stetige Nebenbedingungen

$$\max\{0, g_i(x)\} = 0, \quad i = 1, \ldots, m$$

geschrieben werden. Die Penaltyfunktion lautet dann

$$P(x;\eta) = f(x) + \frac{\eta}{2}\sum_{j=1}^{p}(h_j(x))^2 + \frac{\eta}{2}\sum_{i=1}^{m}\left(\max\{0, g_i(x)\}\right)^2.$$

Ein wesentlicher Nachteil des Penalty-Verfahrens ist die Tatsache, daß die Gewichtungsfaktoren η_k gegen ∞ streben müssen, um Konvergenz zu erhalten. Dies führt dazu, daß die Teilprobleme in (ii) des Algorithmus für grosses η_k sehr schlecht konditioniert sind[4] und numerisch nur sehr schwer zu lösen sind.

5.9.1 Schätzung der Lagrange-Multiplikatoren

Wir untersuchen, wie aus der Folge $\{x^{[k]}\}$ eine Folge $\{\mu^{[k]}\}$ von Näherungen der Lagrange-Multiplikatoren konstruiert werden kann, so daß $x^{[k]}$ und $\mu^{[k]}$ gegen einen KKT-Punkt $\hat{x}$ und $\hat{\mu}$ des Ausgangsproblems konvergieren.
Hierzu benötigen wir die stetige Differenzierbarkeit der Funktionen f und h_j, $j = 1, \dots, p$.
Ein KKT-Punkt $(\hat{x}, \hat{\mu})$ des Ausgangsproblems erfüllt

$$0 = \nabla f(\hat{x}) + \sum_{j=1}^{p} \hat{\mu}_j \nabla h_j(\hat{x}).$$

Da $x^{[k]}$ Minimalstelle der Penaltyfunktion mit Gewichtungsparameter η_k ist, gilt notwendig

$$0 = \nabla_x P(x^{[k]}; \eta_k) = \nabla f(x^{[k]}) + \eta_k \sum_{j=1}^{p} h_j(x^{[k]}) \nabla h_j(x^{[k]}).$$

Vergleicht man die beiden Ausdrücke, so liegt es nahe,

$$\mu_j^{[k]} = \eta_k h_j(x^{[k]}) \tag{5.41}$$

als Approximation der Lagrange-Multiplikatoren $\hat{\mu}_j$ zu verwenden.
Es gilt:

Satz 5.9.3

Seien f und h_j, $j = 1, \dots, p$ stetig differenzierbar und $\{x^{[k]}\}$ eine durch das Penalty-Verfahren erzeugte Folge mit $x^{[k]} \to \hat{x}$. Die Gradienten $\nabla h_j(\hat{x})$, $j = 1, \dots, p$ seien linear unabhängig, und $\{\mu^{[k]}\}$ sei durch (5.41) gegeben. Dann gelten:

(a) Die Folge $\{\mu^{[k]}\}$ konvergiert gegen einen Vektor $\hat{\mu}$.

(b) $(\hat{x}, \hat{\mu})$ ist ein KKT-Punkt des Ausgangsproblems.

Beweis: (vgl. [GK02], S. 211)

[4] Einige Eigenwerte der Hessematrix $\nabla^2_{xx} P(x^{[k]}; \eta_k)$ streben gegen ∞ und somit strebt die Spektral-Konditionszahl der Hessematrix gegen unendlich.

(a) Es seien $J_k := \frac{\partial h(x^{[k]})}{\partial x}$ und $\hat{J} := \frac{\partial h(\hat{x})}{\partial x}$ die Jacobimatrizen von $h = (h_1, \ldots, h_p)$ in $x^{[k]}$ bzw. $\hat{x}$. Aus der Stetigkeit der Jacobimatrizen folgt $J_k \to \hat{J}$. Da die Gradienten $\nabla h_j(\hat{x})$ linear unabhängig sind, ist die Matrix $J_k J_k^\top$ regulär und es gilt $\left(J_k J_k^\top\right)^{-1} \to \left(\hat{J}\hat{J}^\top\right)^{-1}$. Da $x^{[k]}$ ein Minimum von $P(x;\eta_k)$ ist, gilt

$$\begin{aligned} 0 &= \nabla_x P(x^{[k]};\eta_k) \\ &= \nabla f(x^{[k]}) + \eta_k \sum_{j=1}^{p} h_j(x^{[k]}) \nabla h_j(x^{[k]}) \\ &= \nabla f(x^{[k]}) + \sum_{j=1}^{p} \mu_j^{[k]} \nabla h_j(x^{[k]}) \\ &= \nabla f(x^{[k]}) + J_k^\top \mu^{[k]}. \end{aligned}$$

Multiplikation von links mit J_k liefert $J_k J_k^\top \mu^{[k]} = -J_k \nabla f(x^{[k]})$ bzw.

$$\mu^{[k]} = -\left(J_k J_k^\top\right)^{-1} J_k \nabla f(x^{[k]}) \to -\left(\hat{J}\hat{J}^\top\right)^{-1} \hat{J} \nabla f(\hat{x}) =: \hat{\mu}.$$

(b) Folgt aus

$$0 = \nabla_x P(x^{[k]};\eta_k) = \nabla f(x^{[k]}) + \eta_k \sum_{j=1}^{p} h_j(x^{[k]}) \nabla h_j(x^{[k]})$$

und $\mu_j^{[k]} = \eta_k h_j(x^{[k]}) \to \hat{\mu}$.

□

5.10 Multiplier-Penalty-Verfahren

Multiplier-Penalty-Verfahren ähneln den Penalty-Verfahren. Allerdings arbeiten sie mit einer exakten und differenzierbaren Penalty-Funktion – der erweiterten Lagrangefunktion. Wir betrachten wieder das gleichungsrestringierte Problem

$$\min_{x \in \mathbb{R}^n} f(x) \quad \text{unter} \quad h(x) = 0. \tag{5.42}$$

Darin seien die Funktionen $f : \mathbb{R}^n \to \mathbb{R}$ und $h = (h_1, \ldots, h_p)^\top : \mathbb{R}^n \to \mathbb{R}^p$ zweimal stetig differenzierbar.

Sei $\hat{x}$ lokales Minimum des Problems. Dann ist $\hat{x}$ für $\eta > 0$ auch ein lokales Minimum von

$$\min f(x) + \frac{\eta}{2}\|h(x)\|^2 \quad \text{unter} \quad h(x) = 0.$$

Die Lagrangefunktion für dieses Problem lautet

$$L_a(x, \mu; \eta) := f(x) + \frac{\eta}{2}\|h(x)\|^2 + \mu^\top h(x)$$

und heißt **erweiterte Lagrangefunktion** (**augmented Lagrangian**) oder **Multiplier-Penalty-Funktion**.
Es zeigt sich, daß L_a exakt ist.

Hilfssatz 5.10.1

Sei $(\hat{x}, \hat{\mu})$ KKT-Punkt von (5.42). Desweiteren sei die hinreichende Bedingung zweiter Ordnung (5.14) erfüllt. Dann existiert ein endliches $\bar{\eta} > 0$, so daß $\hat{x}$ für jedes $\eta \geq \bar{\eta}$ ein striktes lokales Minimum von $L_a(\cdot, \hat{\mu}; \eta)$ ist.

Beweis: siehe Geiger und Kanzow [GK02], S. 229 □

Auf Grund dieses Hilfssatzes kann man versuchen, das Ausgangsproblem (5.42) indirekt zu lösen, indem die erweiterte Lagrangefunktion minimiert wird:

$$\min_{x \in \mathbb{R}^n} L_a(x, \hat{\mu}; \eta).$$

Der Penalty-Parameter η muß jetzt, anders als bei den Penalty-Verfahren, nicht mehr gegen ∞ streben, da L_a exakt ist. Darüber hinaus ist L_a differenzierbar, so daß bekannte Verfahren aus der unrestringierten Optimierung eingesetzt werden können.

Problem: Der optimale Lagrangemultiplikator $\hat{\mu}$ ist unbekannt.

Wir versuchen nun, $\hat{\mu}$ geeignet zu approximieren. Seien η hinreichend groß und $x^{[k+1]}$ stationärer Punkt des Problems

$$\min_{x \in \mathbb{R}^n} L_a(x, \mu^{[k]}; \eta).$$

Dann gilt notwendig

$$0 = \nabla_x L_a(x^{[k+1]}, \mu^{[k]}; \eta) = \nabla f(x^{[k+1]}) + \sum_{j=1}^{p} \left(\mu_j^{[k]} + \eta h_j(x^{[k+1]}) \right) \nabla h_j(x^{[k+1]}).$$

Andererseits gilt in einem KKT-Punkt $(\hat{x}, \hat{\mu})$ von (5.42) notwendig

$$0 = \nabla_x L(\hat{x}, \hat{\mu}) = \nabla f(\hat{x}) + \sum_{j=1}^{p} \hat{\mu}_j \nabla h_j(\hat{x}).$$

Ein Vergleich beider Ausdrücke liefert die naheliegende Aufdatierungsvorschrift

$$\mu^{[k+1]} := \mu^{[k]} + \eta h(x^{[k+1]}).$$

Insgesamt entsteht das Multiplier-Penalty-Verfahren:

Algorithmus: Multiplier-Penalty-Verfahren

(i) Wähle $x^{[0]} \in \mathbb{R}^n$, $\mu^{[0]} \in \mathbb{R}^p$, $\eta_0 > 0$, $\sigma \in (0,1)$ und setze $k = 0$.

(ii) Ist $(x^{[k]}, \mu^{[k]})$ KKT-Punkt von (5.42), STOP.

(iii) Bestimme $x^{[k+1]}$ als Lösung von

$$\min_{x \in \mathbb{R}^n} L_a(x, \mu^{[k]}; \eta_k).$$

(iv) Setze $\mu^{[k+1]} := \mu^{[k]} + \eta_k h(x^{[k+1]})$.

(v) Ist $\|h(x^{[k+1]})\| \geq \sigma \|h(x^{[k]})\|$, so setze $\eta_{k+1} := 10\eta_k$. Andernfalls setze $\eta_{k+1} := \eta_k$.

(vi) Setze $k := k + 1$ und gehe zu (ii).

5.10.1 Anwendung auf Ungleichungen

Das Standard-Optimierungsproblem (mit $S = \mathbb{R}^n$)

$$\min_{x \in \mathbb{R}^n} f(x) \quad \text{unter} \quad g(x) \leq 0,\ h(x) = 0 \tag{5.43}$$

ist durch Einführung von Schlupfvariablen $s = (s_1, \ldots, s_m)^\top \in \mathbb{R}^m$ äquivalent mit dem gleichungsrestringierten Problem

$$\begin{array}{lrcl} \min\limits_{x,s} & f(x) & & \\ \text{unter} & g_i(x) + s_i^2 & = & 0, \quad i = 1, \ldots, m, \\ & h_j(x) & = & 0, \quad j = 1, \ldots, p. \end{array}$$

Die erweiterte Lagrange-Funktion hierfür lautet

$$\bar{L}_a(x, s, \lambda, \mu; \eta) = f(x) + \frac{\eta}{2}\|h(x)\|^2 + \mu^\top h(x) + \sum_{i=1}^{m} \left(\lambda_i \left(g_i(x) + s_i^2 \right) + \frac{\eta}{2} \left(g_i(x) + s_i^2 \right)^2 \right).$$

Für festes x kann die Minimierung bzgl. s explizit ausgeführt werden und man erhält

$$\hat{s}_i = \left(\max \left\{ 0, -\left(\frac{\lambda_i}{\eta} + g_i(x) \right) \right\} \right)^{1/2}, \quad i = 1, \ldots, m.$$

Einsetzen in die erweiterte Lagrange-Funktion liefert

$$\begin{aligned}\bar{L}_a(x,\lambda,\mu;\eta) &= f(x)+\mu^\top h(x)+\frac{\eta}{2}\|h(x)\|^2 \\ &\quad +\frac{1}{2\eta}\sum_{i=1}^{m}\left(\left(\max\{0,\lambda_i+\eta g_i(x)\}\right)^2-\lambda_i^2\right) \\ &= f(x)+\sum_{j=1}^{p}\left(\mu_j h_j(x)+\frac{\eta}{2}h_j(x)^2\right) \\ &\quad +\sum_{i=1}^{m}\begin{cases}\lambda_i g_i(x)+\frac{\eta}{2}g_i(x)^2, & \text{falls } \lambda_i+\eta g_i(x)\geq 0,\\ -\frac{\lambda_i^2}{2\eta}, & \text{sonst.}\end{cases}\end{aligned}$$

Beachte, daß diese Funktion nur noch stetig differenzierbar ist.
Für die Multiplikatoren ergeben sich die Aufdatierungsformeln

$$\begin{aligned}\mu^{[k+1]} &:= \mu^{[k]}+\eta h(x^{[k+1]}),\\ \lambda_i^{[k+1]} &:= \max\left\{0,\lambda_i^{[k]}+\eta g_i(x^{[k+1]})\right\}, \qquad i=1,\ldots,m.\end{aligned}$$

5.11 Innere-Punkt-Verfahren

Innere-Punkt-Verfahren basieren auf der Konstruktion von Näherungslösungen, die sich strikt im Inneren des zulässigen Bereichs Σ befinden und sich aus dem Inneren dem Optimum (welches in der Regel am Rand liegt) nähern. Erreicht wird dies durch Ankopplung von Straftermen, die zulässige Punkte auf dem Rand bestrafen (anders als bei Penalty-Verfahren, wo nur unzulässige Punkte bestraft werden!), vgl. Abbildung 5.7.

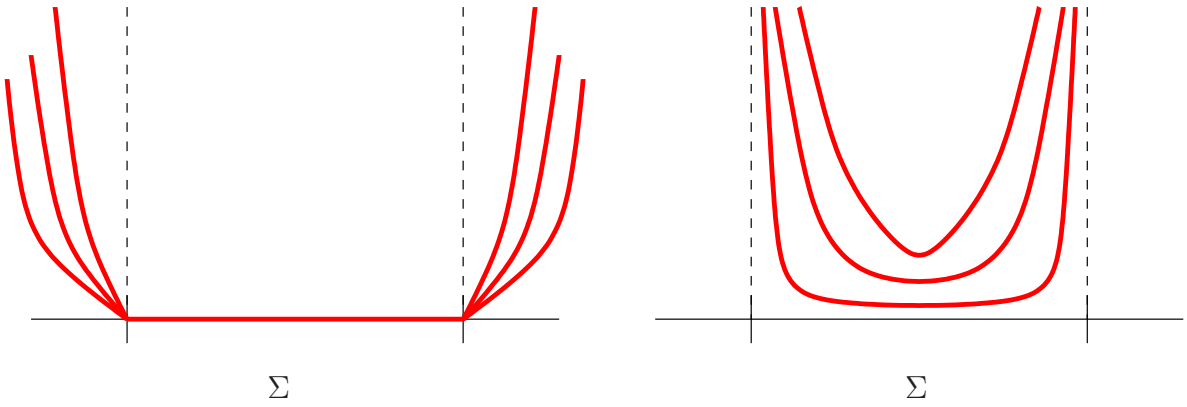


Abbildung 5.7: Qualitativer Unterschied zwischen Penalty-Verfahren (links) und Barriere-Verfahren (rechts). Dargestellt sind die Werte der Penalty-Funktion bzw. der Barriere-Funktion innerhalb und außerhalb des zulässigen Bereichs Σ.

Derartige Ansätze wurden bereits früh von Fiacco und McCormick [FM90] diskutiert, wurden zwischenzeitlich nicht sonderlich beachtet und haben mittlerweile eine Renaissance erfahren.

5.11.1 Lineare Optimierungsprobleme

Wir starten mit dem primalen linearen Optimierungsproblem

$$\text{Minimiere} \quad c^\top x \quad \text{unter} \quad Ax = b,\ x \geq 0 \tag{5.44}$$

mit $c, x \in \mathbb{R}^n$, $A \in \mathbb{R}^{m\times n}$, $b \in \mathbb{R}^m$. Mit der Lagrangefunktion

$$L(x, \lambda, \mu) = c^\top x + \lambda^\top(-x) + \mu^\top(b - Ax)$$

lauten die KKT-Bedingungen

$$\begin{aligned} A^\top\mu + \lambda &= c, && (5.45)\\ Ax &= b, && (5.46)\\ x &\geq 0, && (5.47)\\ \lambda &\geq 0, && (5.48)\\ \lambda_i x_i &= 0, \quad i = 1, \ldots, n. && (5.49) \end{aligned}$$

Nun eliminieren wir die Ungleichungen $x \geq 0$ im primalen Problem, indem wir diese durch Strafterme an die Zielfunktion koppeln und erhalten für $\eta > 0$ das sogenannte **(logarithmische) Barriere-Problem**

$$\text{Minimiere} \quad c^\top x - \eta \sum_{i=1}^{n} \log(x_i) \qquad \text{unter} \quad Ax = b. \tag{5.50}$$

Beachte, daß $\log(x_i) \to -\infty$ für $x_i \downarrow 0$. Das Barriere-Problem (5.50) ist nichtlinear. Die KKT-Bedingungen lauten

$$\begin{aligned} c_i - \frac{\eta}{x_i} - (A^\top\mu)_i &= 0, \qquad i = 1, \ldots, n,\\ Ax &= b. \end{aligned}$$

Mit den Definitionen $\lambda_i := \frac{\eta}{x_i}$, $i = 1, \ldots, n$ lauten sie

$$\begin{aligned} A^\top\mu + \lambda &= c, && (5.51)\\ Ax &= b, && (5.52)\\ \lambda_i x_i &= \eta, \quad i = 1, \ldots, n. && (5.53) \end{aligned}$$

Ein Vergleich mit (5.45)-(5.49) zeigt, daß die KKT-Bedingungen (5.51)- (5.53) für das Barriere-Problem als **Störung** der KKT-Bedingungen (5.45)-(5.49) interpretiert werden

können, wenn noch $x > 0$ und $\lambda > 0$ gilt. Die Störung tritt explizit durch den Gewichtungsparameter $\eta > 0$ in der Komplementaritätsbedingung (5.49) bzw. (5.53) auf.
Besitzt das nichtlineare Gleichungssystem (5.51)-(5.53) für jedes $\eta > 0$ eine Lösung

$$(x(\eta), \lambda(\eta), \mu(\eta)),$$

so besteht die Hoffnung, daß diese für $\eta \downarrow 0$ gegen eine Lösung des Ausgangsproblems (5.44) konvergiert. Die Menge

$$\{(x(\eta), \lambda(\eta), \mu(\eta)) \mid \eta > 0\}$$

heißt **zentraler Pfad**.
Da die KKT-Bedingungen für das konvexe Barriere-Problem notwendig (Abadie) und hinreichend sind und mit den zentralen Pfadbedingungen (5.51)-(5.52) übereinstimmen, gilt folgender Hilfssatz.

Hilfssatz 5.11.1

Sei $\eta > 0$. Das Barriere-Problem besitzt genau dann eine Lösung $x > 0$, wenn die zentralen Pfadbedingungen (5.51)-(5.52) eine Lösung (x, μ, λ) mit $x > 0$, $\lambda > 0$ besitzen.

Es stellt sich die Frage, wann die zentralen Pfadbedingungen eine Lösung $x > 0$, $\lambda > 0$ besitzen. Offenbar besitzen sie keine Lösung, wenn die Menge

$$\mathcal{F}^\circ := \{(x, \mu, \lambda) \mid A^\top \mu + \lambda = c,\ Ax = b,\ x > 0,\ \lambda > 0\}$$

leer ist. Andernfalls besitzen sie eine Lösung:

Satz 5.11.2

Sei $\mathcal{F}^\circ \neq \emptyset$. Dann besitzt das Barriere-Problem für jedes $\eta > 0$ eine Lösung $x > 0$ (wegen Hilfssatz 5.11.1 besitzen dann auch die zentralen Pfadbedingungen eine Lösung).

Beweis: Seien $\eta > 0$ und $(\hat{x}, \hat{\mu}, \hat{\lambda}) \in \mathcal{F}^\circ$. Dann gilt $A^\top \hat{\mu} + \hat{\lambda} = c, A\hat{x} = b, \hat{x} > 0, \hat{\lambda} > 0$.
Betrachte

$$\phi_\eta(x) = c^\top x - \eta \sum_{i=1}^{n} \log(x_i).$$

Wir zeigen, daß die Menge

$$L_\eta := \{x \in \mathbb{R}^n \mid Ax = b,\ x \geq 0,\ \phi_\eta(x) \leq \phi_\eta(\hat{x})\}$$

kompakt ist. Da sie offenbar abgeschlossen ist, müssen wir nur noch die Beschränktheit

zeigen. Sei nun $x \in L_\eta$. Dann gilt

$$\begin{aligned}
\phi_\eta(x) &= c^\top x - \eta \sum_{i=1}^n \log(x_i) \\
&= c^\top x - \hat{\mu}^\top (Ax - b) - \eta \sum_{i=1}^n \log(x_i) \\
&= c^\top x - x^\top A^\top \hat{\mu} + b^\top \hat{\mu} - \eta \sum_{i=1}^n \log(x_i) \\
&= c^\top x - x^\top (c - \hat{\lambda}) + b^\top \hat{\mu} - \eta \sum_{i=1}^n \log(x_i) \\
&= x^\top \hat{\lambda} + b^\top \hat{\mu} - \eta \sum_{i=1}^n \log(x_i).
\end{aligned}$$

Damit ist $\phi_\eta(x) \leq \phi_\eta(\hat{x})$ äquivalent mit

$$x^\top \hat{\lambda} + b^\top \hat{\mu} - \eta \sum_{i=1}^n \log(x_i) \leq \phi_\eta(\hat{x})$$

bzw.

$$\sum_{i=1}^n \Big(\hat{\lambda}_i x_i - \eta \log(x_i) \Big) \leq \phi_\eta(\hat{x}) - b^\top \hat{\mu} =: \kappa.$$

Die Funktion $x_i \mapsto p(x_i) := \hat{\lambda}_i x_i - \eta \log(x_i)$ hat die Eigenschaften $p(x_i) \to \infty$ für $x_i \to 0$ oder $x_i \to \infty$. Damit ist die Menge $\{x \in \mathbb{R}^n \mid \phi_\eta(x) \leq \phi_\eta(\hat{x})\}$ beschränkt und es folgt automatisch $x > 0$. Also ist auch die Menge L_η beschränkt und somit kompakt. Folglich nimmt die stetige Funktion ϕ_η auf der kompakten Menge L_η ihr Minimum an.

□

Zur numerischen Lösung der nichtlinearen Gleichungen (5.51)-(5.53) wird das Newton-Verfahren auf die Funktion

$$F(x, \mu, \lambda; \eta) := \begin{pmatrix} A^\top \mu + \lambda - c \\ Ax - b \\ X\Lambda e - \eta e \end{pmatrix}$$

angewendet. Hierbei benutzen wir die Notationen

$$X = diag(x_1, \ldots, x_n), \quad \Lambda = diag(\lambda_1, \ldots, \lambda_n), \quad e = (1, \ldots, 1)^\top.$$

Die Jacobimatrix von F lautet

$$F'_{(x,\mu,\lambda)})(x, \mu, \lambda; \eta) = \begin{pmatrix} 0 & A^\top & I \\ A & 0 & 0 \\ \Lambda & 0 & X \end{pmatrix}$$

Es gilt

Hilfssatz 5.11.3

Sei $(x,\mu,\lambda) \in \mathbb{R}^n \times \mathbb{R}^m \times \mathbb{R}^n$ ein Vektor mit $x > 0$ und $\lambda > 0$ und es gelte $Rang(A) = m$. Dann ist die Jacobimatrix $F'_{(x,\mu,\lambda)})(x,\mu,\lambda;\eta)$ für jedes $\eta > 0$ regulär.

Beweis: Der Beweis verläuft analog zum Beweis zur Regularität der KKT-Matrix ($\rightarrow$ Übung). □

Sei $w^{[k]} := (x^{[k]}, \mu^{[k]}, \lambda^{[k]})$ ein gegebener Iterationspunkt. Die Newtonkorrektur Δw ist durch das lineare Gleichungssystem

$$F'_w(w^{[k]})\Delta w = -F(w^{[k]})$$

bzw. durch

$$\begin{pmatrix} 0 & A^\top & I \\ A & 0 & 0 \\ \Lambda^{[k]} & 0 & X^{[k]} \end{pmatrix} \begin{pmatrix} \Delta x \\ \Delta\mu \\ \Delta\lambda \end{pmatrix} = - \begin{pmatrix} A^\top\mu^{[k]} + \lambda^{[k]} - c \\ Ax^{[k]} - b \\ X^{[k]}\Lambda^{[k]}e - \eta_k e \end{pmatrix} \tag{5.54}$$

gegeben, wobei $X^{[k]} = diag(x_1^{[k]}, \ldots, x_n^{[k]})$ und $\Lambda^{[k]} = diag(\lambda_1^{[k]}, \ldots, \Lambda_n^{[k]})$ sind. Das gedämpfte Newtonverfahren liefert die neue Iterierte

$$w^{[k+1]} := w^{[k]} + t_k\Delta w$$

mit einer Schrittweite $t_k > 0$.

Wir betrachten noch den Spezialfall eines **zulässigen Verfahrens**, welches dadurch ausgezeichnet ist, daß es die Bedingungen

$$\begin{aligned} A^\top\mu^{[k]} + \lambda^{[k]} - c &= 0, \\ Ax^{[k]} - b &= 0 \end{aligned}$$

erfüllt. Aus den ersten beiden Gleichungen in (5.54) folgen dann

$$\begin{aligned} A^\top\Delta\mu + \Delta\lambda &= 0, \\ A\Delta x &= 0. \end{aligned}$$

Für die neue Iterierte ergibt sich damit

$$\begin{aligned} A^\top\mu^{[k+1]} + \lambda^{[k+1]} - c &= A^\top\left(\mu^{[k]} + t_k\Delta\mu\right) + \lambda^{[k]} + t_k\Delta\lambda - c \\ &= A^\top\mu^{[k]} + \lambda^{[k]} - c \\ &= 0, \\ Ax^{[k+1]} - b &= A\left(x^{[k]} + t_k\Delta x\right) - b \\ &= Ax^{[k]} - b \\ &= 0. \end{aligned}$$

Mit anderen Worten: Erfüllt $(x^{[0]}, \mu^{[0]}, \lambda^{[0]})$ die Bedingungen (5.51)-(5.52), so erfüllen alle weiteren Iterierten diese Bedingungen ebenfalls.
Zusammenfassend erhalten wir das folgende (konzeptionelle) Innere-Punkt-Verfahren:

Algorithmus: Innere-Punkt-Verfahren

(i) Wähle $w^{[0]} := (x^{[0]}, \mu^{[0]}, \lambda^{[0]})$ mit

$$Ax^{[0]} = b, \qquad A^\top \mu^{[0]} + \lambda^{[0]} = c, \qquad x^{[0]} > 0, \qquad \lambda^{[0]} > 0,$$

$\varepsilon > 0$ und setze $k = 0$.

(ii) Ist $\zeta_k := \frac{(x^{[k]})^\top \lambda^{[k]}}{n} \leq \varepsilon$, STOP.

(iii) Wähle $\sigma_k \in [0,1]$ und berechne $\Delta w = (\Delta x, \Delta\mu, \Delta\lambda)$ als Lösung von

$$\begin{pmatrix} 0 & A^\top & I \\ A & 0 & 0 \\ \Lambda^{[k]} & 0 & X^{[k]} \end{pmatrix} \begin{pmatrix} \Delta x \\ \Delta\mu \\ \Delta\lambda \end{pmatrix} = - \begin{pmatrix} 0 \\ 0 \\ X^{[k]}\Lambda^{[k]} e - \sigma_k \zeta_k e \end{pmatrix}. \tag{5.55}$$

(iv) Setze $w^{[k+1]} := w^{[k]} + t_k \Delta w$, wobei $t_k > 0$ so gewählt ist, daß $x^{[k+1]} > 0$ und $\lambda^{[k+1]} > 0$ gelten.

(v) Setze $k := k + 1$ und gehe zu (ii).

Erläuterungen:

- Die Iterierten $x^{[k]}$ sind primal zulässig, da per Konstruktion stets

  $$Ax^{[k]} = b, \qquad x^{[k]} > 0$$

  gilt. Die Iterierten $\mu^{[k]}$ sind **dual zulässig**, da stets

  $$A^\top \mu^{[k]} + \lambda^{[k]} = c, \qquad \lambda^{[k]} > 0$$

  gilt. Beachte, daß das duale Problem zu (5.44) gegeben ist durch

  $$\max \quad b^\top \mu \quad \text{unter} \quad A^\top \mu \leq c,$$

  vgl. Beispiel 5.5.2. Nach Einführung von Schlupfvariablen $\lambda \geq 0$ ist dieses äquivalent mit

  $$\max \quad b^\top \mu \quad \text{unter} \quad A^\top \mu + \lambda = c,\ \lambda \geq 0.$$

- Hilfssatz 5.11.3 garantiert die Durchführbarkeit des Algorithmus, falls noch $\operatorname{Rang}(A) = m$ gefordert wird.

- Die einzige Bedingung in den KKT-Bedingungen (5.45)-(5.49), die von $(x^{[k]}, \mu^{[k]}, \lambda^{[k]})$ i.a. nicht erfüllt wird, ist die Komplementaritätsbedingung $(\lambda^{[k]})^\top x^{[k]} = 0$. Wir müssen also dafür sorgen, daß die Größe $(\lambda^{[k]})^\top x^{[k]}$ bzw.

$$\zeta_k := \frac{(x^{[k]})^\top \lambda^{[k]}}{n}$$

klein wird. Dies motiviert das Abbruchkriterium in Schritt (ii).

- Der Algorithmus enthält noch zwei Freiheitsgrade, die nicht näher spezifiziert sind: die **Schrittweite** $t_k > 0$ und den **Centering-Parameter** $\sigma_k > 0$ (beachte, daß der Term $\sigma_k \zeta_k$ die Rolle des Penalty-Parameters η übernimmt). Je nach Wahl dieser Parameter ergeben sich verschiedene Verfahren. Für Details verweisen wir auf Geiger und Kanzow [GK02], Jarre und Stoer [JS04], Wright [Wri97] und Vanderbei [Van01].

Satz 5.11.4 **(Konvergenzsatz)**
Seien $\varepsilon \in (0,1)$ beliebig und $\{(x^{[k]}, \mu^{[k]}, \lambda^{[k]})\}$ eine durch den Algorithmus erzeugte Folge. Es gelte

$$\zeta_{k+1} \leq \left(1 - \frac{\delta}{n^s}\right) \zeta_k, \qquad k = 0, 1, 2, \ldots \tag{5.56}$$

mit gewissen Parametern $\delta > 0$ und $s > 0$. Der Startvektor $(x^{[0]}, \mu^{[0]}, \lambda^{[0]})$ erfülle die Bedingung

$$\zeta_0 \leq \frac{1}{\varepsilon^\kappa}, \qquad \kappa > 0.$$

Dann existiert ein Index $K \in \mathbb{N}$ mit $K = \mathcal{O}(n^s |\log(\varepsilon)|)$ und $\zeta_k \leq \varepsilon$ für alle $k \geq K$.

Beweis: Anwendung des Logarithmus als monotone Funktion auf (5.56) liefert

$$\log(\zeta_{k+1}) \leq \log\left(1 - \frac{\delta}{n^s}\right) + \log(\zeta_k), \qquad k = 0, 1, 2, \ldots$$

Rekursive Anwendung führt auf

$$\log(\zeta_k) \leq k \log\left(1 - \frac{\delta}{n^s}\right) + \log(\zeta_0) \leq k \log\left(1 - \frac{\delta}{n^s}\right) + \kappa \log\left(\frac{1}{\varepsilon}\right).$$

Wegen $\log(1+\beta) \leq \beta$ für $\beta > -1$ gilt

$$\log(\zeta_k) \leq k\left(-\frac{\delta}{n^s}\right) + \kappa \log\left(\frac{1}{\varepsilon}\right).$$

Das Abbruchkriterium $\zeta_k \leq \varepsilon$ ist erfüllt, wenn gilt

$$k\left(-\frac{\delta}{n^s}\right) + \kappa \log\left(\frac{1}{\varepsilon}\right) \leq \log(\varepsilon).$$

Umformung liefert die Bedingung

$$k \geq (1+\kappa)\frac{n^s}{\delta}|\log(\varepsilon)| = \mathcal{O}(n^s|\log(\varepsilon)|).$$

□

Daß die Bedingung (5.56) erfüllbar ist, zeigt

Hilfssatz 5.11.5

Sei $(\Delta x, \Delta\mu, \Delta\lambda)$ *Lösung von (5.55) und*

$$\begin{aligned}(x^{[k]}(t), \mu^{[k]}(t), \lambda^{[k]}(t)) &= (x^{[k]}, \mu^{[k]}, \lambda^{[k]}) + t(\Delta x, \Delta\mu, \Delta\lambda),\\ \zeta_k(t) &= \frac{(x^{[k]}(t))^\top \lambda^{[k]}(t)}{n}.\end{aligned}$$

Dann gelten:

$$\begin{aligned}\Delta x^\top \Delta\lambda &= 0,\\ \zeta_k(t) &= (1-t(1-\sigma_k))\,\zeta_k.\end{aligned}$$

Beweis: Da $(\Delta x, \Delta\mu, \Delta\lambda)$ Gleichung (5.55) löst, gelten

$$\begin{aligned}A^\top\Delta\mu + \Delta\lambda &= 0,\\ A\Delta x &= 0,\\ \Lambda^{[k]}\Delta x + X^{[k]}\Delta\lambda &= -X^{[k]}\Lambda^{[k]}e + \sigma_k\zeta_k e.\end{aligned}$$

Multiplikation der ersten Gleichung von links mit $\Delta x^\top$ liefert $\Delta x^\top A^\top\Delta\mu + \Delta x^\top\Delta\lambda = 0$. Ausnutzen der zweiten Gleichung zeigt $\Delta x^\top\Delta\lambda = 0$.
Multiplikation der dritten Gleichung mit $e^\top$ von links liefert

$$(\lambda^{[k]})^\top\Delta x + (x^{[k]})^\top\Delta\lambda = -(x^{[k]})^\top\lambda^{[k]} + \sigma_k\zeta_k n = -(1-\sigma_k)(x^{[k]})^\top\lambda^{[k]}.$$

Zusammen mit $\Delta x^\top\Delta\lambda = 0$ folgt

$$\begin{aligned}n\zeta_k(t) &= (x^{[k]})^\top\lambda^{[k]} + t\left((\lambda^{[k]})^\top\Delta x + (x^{[k]})^\top\Delta\lambda\right) + t^2\Delta x^\top\Delta\lambda\\ &= (x^{[k]})^\top\lambda^{[k]}\,(1-t(1-\sigma_k))\\ &= n\zeta_k\,(1-t(1-\sigma_k))\,.\end{aligned}$$

□

Konkrete Realisierungen des Algorithmus versuchen, die Schrittweite t_k so zu wählen, daß die Iterierten $(x^{[k]}, \mu^{[k]}, \lambda^{[k]})$ in der Nähe des zentralen Pfades verbleiben und heißen **Pfadverfolgungsverfahren**. Dabei gibt es zulässige und unzulässige Varianten.

In der Regel werden die Umgebungen

$$N_2(\theta) := \{(x, \mu, \lambda) \in \mathcal{F}^\circ \mid \|X\Lambda e - \zeta e\|_2 \leq \theta\zeta\}$$

und

$$N_{-\infty}(\gamma) := \{(x, \mu, \lambda) \in \mathcal{F}^\circ \mid x_i\lambda_i \geq \gamma\zeta,\ i = 1, \ldots, n\}$$

mit $\zeta = x^\top\lambda/n$ und

$$\mathcal{F}^\circ := \{(x, \mu, \lambda) \mid A^\top\mu + \lambda = c,\ Ax = b,\ x > 0,\ \lambda > 0\}$$

verwendet.

Es zeigt sich, daß die Umgebung $N_2(\theta)$ in der Regel kleiner ist als $N_{-\infty}(\gamma)$, so daß Verfahren, die auf der Umgebung $N_2(\theta)$ basieren, in der Praxis langsamer konvergieren und kleinere Schrittweiten zulassen als diejenigen Verfahren, die auf der Umgebung $N_{-\infty}(\gamma)$ basieren. Die entsprechenden Verfahren heißen daher auch **short-step-Verfahren** bzw. **long-step-Verfahren**. Aus theoretischer Sicht besitzen die short-step-Verfahren häufig jedoch bessere Konvergenzeigenschaften.

5.11.2 Nichtlineare Optimierungsprobleme

Wir wollen hier nur die grundlegenden Ideen darstellen, da es sich um ein sehr aktives Forschungsgebiet handelt. Das folgende Verfahren stammt von Byrd et al. [BHN99].

Wir betrachten wieder das Standard-Optimierungsproblem mit $S = \mathbb{R}^n$

Finde $x \in \mathbb{R}^n$, so daß $f(x)$ minimal wird unter den Nebenbedingungen

$$\begin{aligned} g_i(x) &\leq 0, \quad i = 1, \ldots, m, \\ h_j(x) &= 0, \quad j = 1, \ldots, p. \end{aligned}$$

Ähnlich wie im linearen Fall eliminieren wir die Ungleichungsrestriktionen, indem wir Schlupfvariablen $s > 0$ einführen und diese über Logarithmus-Terme mit Gewichtungsfaktor $\eta > 0$ an die Zielfunktion koppeln. Wir erhalten das

Barriere-Problem: Finde $x \in \mathbb{R}^n$ und $s = (s_1, \ldots, s_m)^\top \in \mathbb{R}^m$, so daß

$$f(x) - \eta \sum_{i=1}^{m} \log(s_i)$$

minimal wird unter den Nebenbedingungen

$$\begin{aligned} g_i(x) + s_i &= 0, \qquad i = 1, \ldots, m, \\ h_j(x) &= 0, \qquad j = 1, \ldots, p. \end{aligned}$$

Die Lagrangefunktion des Barriere-Problems (mit $l_0 = 1$) lautet

$$L(x, s, \lambda, \mu) = f(x) - \eta \sum_{i=1}^{m} \log(s_i) + \sum_{i=1}^{m} \lambda_i (g_i(x) + s_i) + \sum_{j=1}^{p} \mu_j h_j(x).$$

Mit $S = diag(s_1, \ldots, s_m)$ lauten die KKT-Bedingungen des Barriere-Problems wie folgt:

$$0 = \nabla_x L(x, s, \lambda, \mu) = \nabla f(x) + \sum_{i=1}^{m} \lambda_i \nabla g_i(x) + \sum_{j=1}^{p} \mu_j \nabla h_j(x), \tag{5.57}$$

$$0 = \nabla_s L(x, s, \lambda, \mu) = -\eta S^{-1} e + \lambda, \tag{5.58}$$

$$0 = g_i(x) + s_i, \qquad i = 1, \ldots, m, \tag{5.59}$$

$$0 = h_j(x), \qquad j = 1, \ldots, p. \tag{5.60}$$

Gleichung (5.58) lautet nach Umformung

$$\eta e = S\lambda \qquad \Leftrightarrow \qquad \eta = s_i \lambda_i, \qquad i = 1, \ldots, m.$$

Zusammen mit $s_i = -g_i(x) < 0$, $i = 1, \ldots, m$ aus (5.59) folgt die mit $-\eta$ **gestörte Komplementaritätsbedingung**

$$-\eta = \lambda_i g_i(x), \qquad i = 1, \ldots, m. \tag{5.61}$$

Beim Innere-Punkte-Verfahren sorgt man nun wie im linearen Fall dafür, daß $s_i > 0$ gilt, was wegen (5.59) gleichbedeutend ist mit $g_i(x) < 0$. Wegen (5.61) und $-\eta < 0$ gilt dann zwangsläufig $\lambda_i > 0$ für $i = 1, \ldots, m$. Die von η abhängige Lösung $(x(\eta), s(\eta), \lambda(\eta), \mu(\eta))$ bestimmt wieder den zentralen Pfad (Existenz der Lösung vorausgesetzt).
Wie im linearen Fall wird das nichtlineare Gleichungssystem (5.57)-(5.60) in den Variablen x, s, λ, μ mit dem Newtonverfahren gelöst. Diese Vorgehensweise entspricht dem Lagrange-Newton-Verfahren. Da nur Gleichungsrestriktionen im Barriere-Problem auftreten, ist das Lagrange-Newton-Verfahren identisch mit dem (lokalen) SQP-Verfahren. Die Newton-Richtungen sind somit durch Lösen von quadratischen Hilfsproblemen gegeben.

Byrd et al. [BHN99] verwenden darüber hinaus ein Trust-Region-SQP-Verfahren, bei dem zur Berechnung einer Suchrichtung $d = (d_x, d_s)^\top$ das folgende quadratische Hilfsproblem gelöst wird:

Quadratisches Hilfsproblem: Finde $(d_x, d_s) \in \mathbb{R}^n \times \mathbb{R}^m$, so daß

$$\frac{1}{2} d_x^\top \nabla_{xx}^2 L(x^{[k]}, s^{[k]}, \lambda^{[k]}, \mu^{[k]}) d_x + \frac{1}{2} d_s^\top \nabla_{ss}^2 L(x^{[k]}, s^{[k]}, \lambda^{[k]}, \mu^{[k]}) d_s + \nabla f(x^{[k]})^\top d_x - \eta e^\top S_k^{-1} d_s$$

minimal wird unter den Nebenbedingungen

$$\begin{aligned} g_i(x^{[k]}) + \nabla g_i(x^{[k]})^\top d_x + s^{[k]} + d_s &= 0, \qquad i = 1, \ldots, m, \\ h_j(x^{[k]}) + \nabla h_j(x^{[k]})^\top d_x &= 0, \qquad j = 1, \ldots, p, \\ (d_x, d_s) &\in T_k. \end{aligned}$$

Darin ist $S_k = diag(s_1^{[k]}, \ldots, s_m^{[k]})$ und T_k bezeichnet den Vertrauensbereich, der in jeder Iteration geeignet angepaßt wird. Ob der Vertrauensbereich angepaßt werden muß, wird mit Hilfe der Bewertungsfunktion

$$l_2(x, s; \nu) = f(x) - \eta \sum_{i=1}^{m} \log(s_i) + \nu \left(\sum_{i=1}^{m} (g_i(x) + s_i)^2 + \sum_{j=1}^{p} h_j(x)^2 \right)^{1/2}$$

entschieden. Falls (d_x, d_s) einen hinreichenden Abstieg in der Bewertungsfunktion erzeugt, werden $x^{[k+1]} = x^{[k]} + d_x$ und $s^{[k+1]} = s^{[k]} + d_s$ als neue Iterierte akzeptiert, andernfalls muß der Vertrauensbereich T_k verkleinert werden.

Wir begnügen uns hier mit diesem Konzept eines Algorithmus und verweisen für die zahlreichen Details des Algorithmus auf Byrd et al. [BHN99]. Unter anderem muß geklärt werden, wie der Parameter η gegen Null streben soll und es muß dafür gesorgt werden, daß $s_i^{[k]} > 0$ für $i = 1, \ldots, m$ erfüllt ist.

Bemerkung 5.11.6

Interpretiert man λ wie im linearen Fall als duale Variable des Dualproblems

$$\max_{\lambda \geq 0,\, \mu} \psi(\lambda, \mu), \qquad \psi(\lambda, \mu) = \min_{x,s} L(x, s, \lambda, \mu),$$

so ist λ wegen $\lambda > 0$ dual strikt zulässig.

Beispiel 5.11.7

Wir möchten die Funktion

$$f(x_1, x_2) = x_1 + x_2$$

minimieren unter den Nebenbedingungen

$$\begin{aligned} g_1(x_1,x_2) = x_1^2 - x_2 &\leq 0, \\ g_2(x_1,x_2) = -x_1 &\leq 0. \end{aligned}$$

Das zugehörige Barriere-Problem lautet

$$\begin{array}{ll} \textit{Minimiere} & x_1 + x_2 - \eta(\log(s_1) + \log(s_2)) \\ \textit{unter} & x_1^2 - x_2 + s_1 = 0, \\ & -x_1 + s_2 = 0. \end{array}$$

Durch Elimination der Schlupfvariablen ist es äquivalent mit der unrestringierten Minimierung der Funktion

$$\phi(x_1,x_2) := x_1 + x_2 - \eta(\log(-x_1^2 + x_2) + \log(x_1)).$$

Ein lokales Minimum von ϕ erfüllt notwendig die Bedingungen

$$\begin{aligned} 0 &= 1 - \eta\frac{-2x_1}{-x_1^2+x_2} - \eta\frac{1}{x_1}, \\ 0 &= 1 - \eta\frac{1}{-x_1^2+x_2}. \end{aligned}$$

Die zweite Gleichung impliziert $1 = \eta\frac{1}{-x_1^2+x_2}$ und damit lautet die erste Gleichung nach einigen Umformungen

$$0 = x_1^2 + \frac{1}{2}x_1 - \frac{\eta}{2}.$$

Diese quadratische Gleichung besitzt die Lösungen

$$x_1 = -\frac{1}{4} \pm \sqrt{\frac{1}{16} + \frac{\eta}{2}}.$$

Wegen $\eta > 0$ erfüllt nur die Lösung mit „+“ die Bedingung $x_1 > 0$. Somit folgt

$$x_1(\eta) = -\frac{1}{4} + \sqrt{\frac{1}{16} + \frac{\eta}{2}}.$$

Schließlich folgt hiermit die Beziehung

$$x_2(\eta) = \eta + x_1(\eta)^2 = \frac{3}{2}\eta + \frac{1}{8} - \frac{1}{2}\sqrt{\frac{1}{16} + \frac{\eta}{2}}.$$

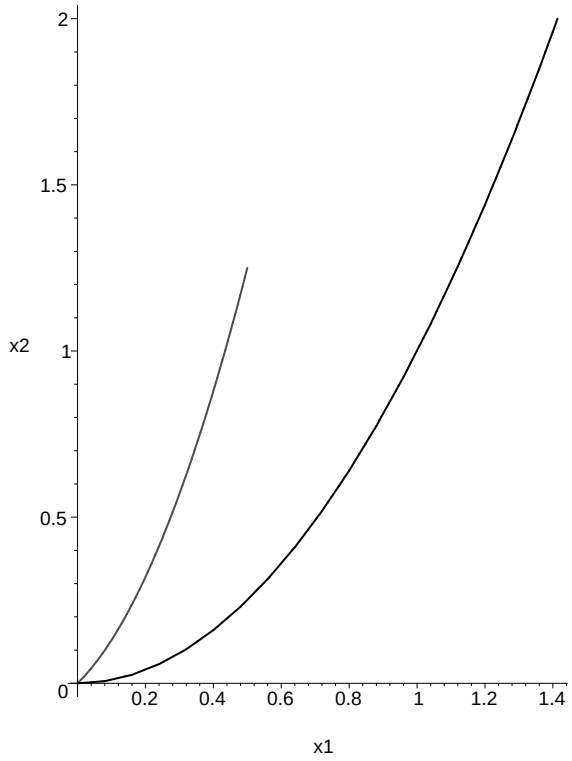
Es gilt

$$\begin{aligned} g_1(x_1(\eta),x_2(\eta)) = -\eta &< 0, \\ g_2(x_1(\eta),x_2(\eta)) = -x_2(\eta) &< 0. \end{aligned}$$

Damit sind $(x_1(\eta), x_2(\eta))$ *strikt zulässig. Da das Barriere-Problem konvex ist, ist* $(x_1(\eta), x_2(\eta))$ *ein lokales Minimum. Grenzübergang* $\eta \to 0$ *liefert*

$$x_1(\eta) \to 0, \qquad x_2(\eta) \to 0.$$

Der Punkt $(0,0)$ *ist gerade das Minimum des Ausgangsproblems.*
Die Abbildung zeigt den zentralen Pfad $\{(x_1(\eta), x_2(\eta)) \mid \eta > 0\}$.



Kapitel 6

Ausblick

Was wir nicht diskutiert haben:

- Nichtdifferenzierbare Optimierungsprobleme (Strukturoptimierung, Eigenwertoptimierung)
- Optimierungsprobleme in unendlichdimensionalen Vektorräumen (z.B. optimale Steuerprozesse)
- Kombinatorische und gemischt-ganzzahlige Optimierung (z.B. An- und Ausschaltvorgänge, Gangschaltung, ganzzahlige Ressoucen)
- Vektor-Optimierung (mehre Optimierungskriterien, die sich gegenseitig widersprechen)
- Globale Optimierung (Berechnung globaler Minima)
- Stochastische Optimierung (Genetische und evolutionäre Algorithmen)

Software im Netz:

- SCILAB: A Free Scientific Software Package; http://www.scilab.org
- GNU OCTAVE: A high-level language, primarily intended for numerical computations; http://www.octave.org/octave.html
- GAMS: Guide to Available Mathematical Software; http://gams.nist.gov/
- NETLIB: collection of mathematical software, papers, and databases; http://www.netlib.org/
- Decision Tree for Optimization Software; http://plato.la.asu.edu/guide.html
- NEOS GUIDE: www-fp.mcs.anl.gov/otc/Guide
- COPS: Large-Scale Optimization Problems; http://www-unix.mcs.anl.gov/∼more/cops/
- GALIB: set of C++ genetic algorithm objects; http://lancet.mit.edu/ga/
- Testprobleme: http://www.princeton.edu/∼rvdb/ampl/nlmodels/

- Spezielle SQP Verfahren:
  - NPSOL (NAG-Routine E04UCF), voll besetzte Probleme, P. Gill et al., University of California, San Diego
  - SNOPT (NAG-Routine), dünn besetzte Probleme, P. Gill et al., University of California, San Diego
  - NLPQP, voll besetzte Probleme, K. Schittkowsi, Universität Bayreuth

Literaturverzeichnis

[Alt02] Alt, W. *Nichtlineare Optimierung: Eine Einführung in Theorie, Verfahren und Anwendungen*. Vieweg, Braunschweig/Wiesbaden, 2002.

[BG01] Büskens, C. and Gerdts, M. *Real-time Optimization of DAE Systems*. In *Online Optimization of Large Scale Systems* (M. Grötschel, S. O. Krumke and J. Rambau, editors), pp. 117–128. Springer, 2001.

[BH75] Bryson, A. E. and Ho, Y.-C. *Applied Optimal Control*. Hemisphere Publishing Corporation, Washington, 1975.

[BH99] Betts, J. T. and Huffman, W. P. *Exploiting Sparsity in the Direct Transcription Method for Optimal Control*. Computational Optimization and Applications, 14 (2); 179–201, 1999.

[BHN99] Byrd, R. H., Hribar, M. E. and Nocedal, J. *An interior point algorithm for large-scale nonlinear programming*. SIAM Journal on Optimization, 9 (4); 877–900, 1999.

[BM68] Bracken, J. and McCormick, G. P. *Selected Applications of Nonlinear Programming*. John Wiley & Sons, New York-London-Sydney-Toronto, 1968.

[BM01] Büskens, C. and Maurer, H. *Sensitivity Analysis and Real-Time Optimization of Parametric Nonlinear Programming Problems*. In *Online Optimization of Large Scale Systems* (M. Grötschel, S. O. Krumke and J. Rambau, editors), pp. 3–16. Springer, 2001.

[BO75] Blum, E. and Oettli, W. *Mathematische Optimierung*. volume 20 of *Ökonometrie und Unternehmensforschung*. Springer-Verlag Berlin Heidelberg New York, Berlin, 1975.

[BS79] Bazaraa, M. S. and Shetty, C. M. *Nonlinear Programming: Theory and Algorithms*. John Wiley & Sons, 1979.

[BSS93] Bazaraa, M. S., Sherali, H. D. and Shetty, C. M. *Nonlinear Programming: Theory and Algorithms*. John Wiley & Sons, 2nd edition, 1993.

[CPS92] Cottle, R. W., Pang, J. S. and Stone, R. E. *The linear complementarity problem*. Academic Press, Boston, 1992.

[Fia83] Fiacco, A. V. *Introduction to Sensitivity and Stability Analysis in Nonlinear Programming*, volume 165 of *Mathematics in Science and Engineering*. Academic Press, New York, 1983.

[FL02] Fletcher, R. and Leyffer, S. *Nonlinear programming without a penalty function*. Mathematical Programming, 91A (2); 239–269, 2002.

[Fle03] Fletcher, R. *Practical Methods of Optimization*. John Wiley & Sons, Chichester–New York–Brisbane–Toronto–Singapore, 2nd edition, 2003.

[FLT02] Fletcher, R., Leyffer, S. and Toint, P. *On the global convergence of a filter-SQP algorithm*. SIAM Journal on Optimization, 13 (1); 44–59, 2002.

[FM90] Fiacco, A. V. and McCormick, G. P. *Nonlinear Programming: Sequential Unconstrained Minimization Techniques*, volume 4 of *Classics In Applied Mathematics*. SIAM, Philadelphia, 1990.

[GI83] Goldfarb, D. and Idnani, A. *A numerically stable dual method for solving strictly convex quadratic programs*. Mathematical Programming, 27; 1–33, 1983.

[GK99] Geiger, C. and Kanzow, C. *Numerische Verfahren zur Lösung unrestringierter Optimierungsaufgaben*. Springer, Berlin-Heidelberg-New York, 1999.

[GK02] Geiger, C. and Kanzow, C. *Theorie und Numerik restringierter Optimierungsaufgaben*. Springer, Berlin-Heidelberg-New York, 2002.

[GM78] Gill, P. E. and Murray, W. *Numerically stable methods for quadratic programming*. Mathematical Programming, 14; 349–372, 1978.

[GMS94] Gill, P. E., Murray, W. and Saunders, M. A. *Large-scale SQP Methods and their Application in Trajectory Optimization*, volume 115 of *International Series of Numerical Mathematics*, pp. 29–42. Birkhäuser, Basel, 1994.

[GMS02] Gill, P. E., Murray, W. and Saunders, M. A. *SNOPT: An SQP algorithm for large-scale constrained optimization*. SIAM Journal on Optimization, 12 (4); 979–1006, 2002.

[GMSW91] Gill, P. E., Murray, W., Saunders, M. A. and Wright, M. H. *Inertia-controlling methods for general quadratic programming*. SIAM Review, 33 (1); 1–36, 1991.

[GMSW98] Gill, P. E., Murray, W., Saunders, M. A. and Wright, M. H. *User's guide for NPSOL 5.0: A FORTRAN package for nonlinear programming*. Technical

Report NA 98-2, Department of Mathematics, University of California, San Diego,California, 1998.

[GMW81] Gill, P. E., Murray, W. and Wright, M. H. *Practical Optimization*. Academic Press, London, 1981.

[Han77] Han, S. P. *A Globally Convergent Method for Nonlinear Programming*. Journal of Optimization Theory and Applications, 22 (3); 297–309, 1977.

[Hes80] Hestenes, M. R. *Conjugate direction methods in optimization*, volume 12 of *Applications of Mathematics*. Springer, New York – Heidelberg – Berlin, 1980.

[HS52] Hestenes, M. R. and Stiefel, E. *Methods of conjugate gradients for solving linear systems*. J. Res. Natl. Bur. Stand., 49; 409–436, 1952.

[JS04] Jarre, F. and Stoer, J. *Optimierung*. Springer, Berlin-Heidelberg-New York, 2004.

[Kel95] Kelley, C. T. *Iterative Methods for Solving Linear and Nonlinear Equations*, volume 16 of *Frontiers in Applied Mathematics*. SIAM, Philadelphia, 1995.

[Kos93] Kosmol, P. *Methoden zur numerischen Behandlung nichtlinearer Gleichungen und Optimierungsaufgaben*. Teubner, Stuttgart, 2nd edition, 1993.

[Kra88] Kraft, D. *A Software Package for Sequential Quadratic Programming*. DFVLR-FB-88-28, Oberpfaffenhofen, 1988.

[LRWW98] Lagarius, J. C., Reeds, J. A., Wright, M. H. and Wright, P. E. *Convergence Properties of the Nelder-Mead Simplex Method in Low Dimensions*. SIAM Journal on Optimization, 9; 112–147, 1998.

[Man94] Mangasarian, O. L. *Nonlinear Programming*, volume 10 of *Classics In Applied Mathematics*. SIAM, Philadelphia, 1994.

[McK98] McKinnon, K. I. M. *Convergence of the Nelder-Mead Simplex Method to a Nonstationary Point*. SIAM Journal on Optimization, 9; 148–158, 1998.

[MF67] Mangasarian, O. L. and Fromowitz, S. *The Fritz John Necessary Optimality Conditions in the Presence of Equality and Inequality Constraints*. Journal of Mathematical Analysis and Applications, 17; 37–47, 1967.

[NM65] Nelder, J. A. and Mead, R. *A Simplex Method for Function Minimization*. Computer Journal, 7; 308–313, 1965.

[PCB02] Price, C. J., Coope, I. D. and Byatt, D. *A Convergent Variant of the Nelder-Mead Algorithm*. Journal of Optimization Theory and Applications, 113 (1); 5–19, 2002.

[Pow78] Powell, M. J. D. *A fast algorithm for nonlinearily constrained optimization calculation*. In *Numerical Analysis* (G. Watson, editor), volume 630 of *Lecture Notes in Mathematics*. Springer, Berlin-Heidelberg-New York, 1978.

[Rob76] Robinson, S. M. *Stability Theory for Systems of Inequalities, Part II: Differentiable Nonlinear Systems*. SIAM Journal on Numerical Analysis, 13 (4); 487–513, 1976.

[Rob82] Robinson, S. M. *Generalized Equations and their Solutions, Part II: Applications to Nonlinear Programming*. Mathematical Programming Study, 19; 200–221, 1982.

[Roc70] Rockafellar, R. T. *Convex Analysis*. Princeton University Press, New Jersey, 1970.

[Sch81] Schittkowski, K. *The Nonlinear Programming Method of Wilson, Han, and Powell with an Augmented Lagrangean Type Line Search Function. Part 1: Convergence Analysis, Part 2: An Efficient Implementation with Linear Least Squares Subproblems*. Numerische Mathematik, 383; 83–114, 115–127, 1981.

[Sch83] Schittkowski, K. *On the Convergence of a Sequential Quadratic Programming Method with an Augmented Lagrangean Line Search Function*. Mathematische Operationsforschung und Statistik, Series Optimization, 14 (2); 197–216, 1983.

[Sch85] Schittkowski, K. *NLPQL: A Fortran subroutine for solving constrained nonlinear programming problems*. Annals of Operations Research, 5; 484–500, 1985.

[Sch96] Schulz, V. H. *Reduced SQP Methods for Large-Scale Optimal Control Problems in DAE with Application to Path Planning Problems for Satellite Mounted Robots*. Ph.D. thesis, Interdisziplinäres Zentrum für Wissenschaftliches Rechnen, Universität Heidelberg, 1996.

[Spe93] Spellucci, P. *Numerische Verfahren der nichtlinearen Optimierung*. Birkhäuser, Basel, 1993.

[Ste95] Steinbach, M. C. *Fast Recursive SQP Methods for Large-Scale Optimal Control Problems*. Ph.D. thesis, Interdisziplinäres Zentrum für Wissenschaftliches Rechnen, Universität Heidelberg, 1995.

[Sto85] Stoer, J. *Principles of sequential quadratic programming methods for solving nonlinear programs*. In *Computational Mathematical Programming* (K. Schittkowski, editor), volume F15 of *NATO ASI Series*, pp. 165–207. Springer, Berlin-Heidelberg-New York, 1985.

[Van01] Vanderbei, R. J. *Linear programming. Foundations and extensions*. volume 37 of *International Series in Operations Research & Management Science*. Kluwer Academic Publishers, Dordrecht, 2001.

[Wri97] Wright, S. E. *Primal-Dual Interior-Point Methods*. SIAM, Philadelphia, 1997.