



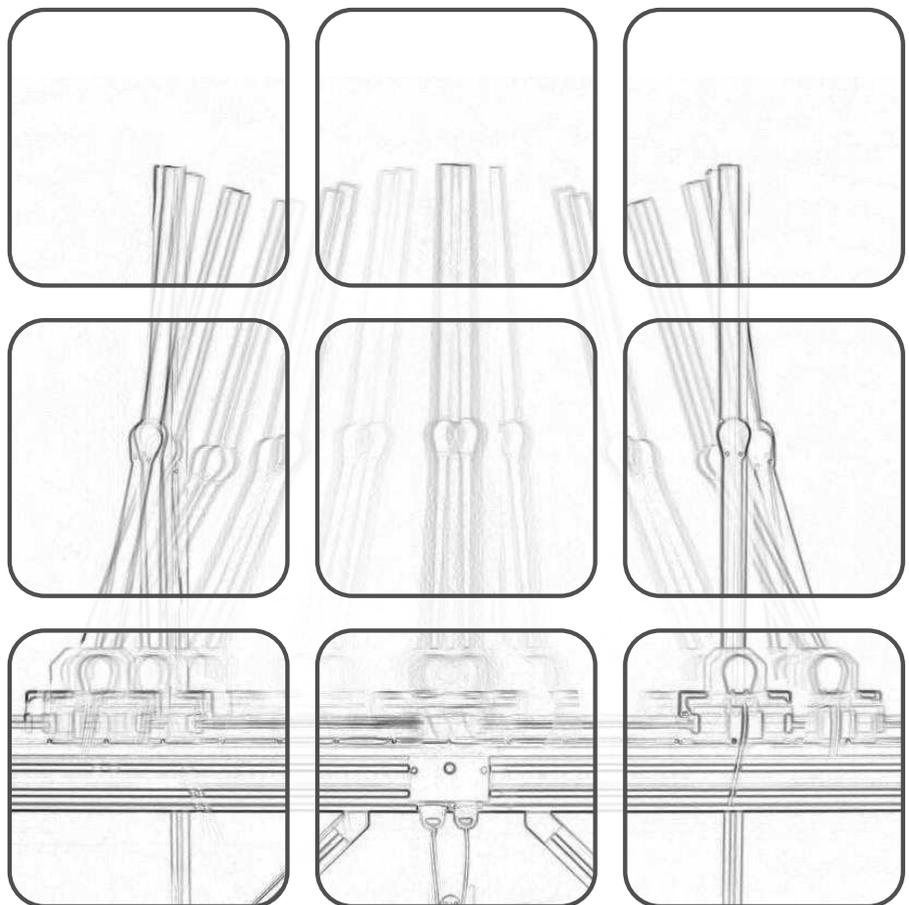
TECHNISCHE  
UNIVERSITÄT  
WIEN



Vorlesung und Übung  
WS 2017/2018

Andreas STEINBÖCK

# OPTIMIERUNG



## **Optimierung**

Vorlesung und Übung  
WS 2017/2018

Andreas STEINBÖCK

TU Wien  
Institut für Automatisierungs- und Regelungstechnik  
Gruppe für komplexe dynamische Systeme

Gußhausstraße 27–29  
1040 Wien  
Telefon: +43 1 58801 – 37615  
Internet: <http://www.acin.tuwien.ac.at>

© Institut für Automatisierungs- und Regelungstechnik, TU Wien

# Inhaltsverzeichnis

<b>1</b>	<b>Einleitung</b>	<b>2</b>
1.1	Statische Optimierungsprobleme . . . . .	2
1.1.1	Mathematische Formulierung . . . . .	2
1.1.2	Beispiele . . . . .	4
1.2	Dynamische Optimierungsprobleme . . . . .	7
1.2.1	Mathematische Formulierung . . . . .	7
1.2.2	Beispiele . . . . .	8
1.3	Mathematische Grundlagen . . . . .	12
1.3.1	Infimum, Supremum, Minimum und Maximum . . . . .	12
1.3.2	Existenz von Minima und Maxima . . . . .	13
1.3.3	Gradient und Hessematrix . . . . .	14
1.3.4	Berechnung von Ableitungen . . . . .	17
1.3.5	Konvexität . . . . .	21
1.4	Literatur . . . . .	25
<b>2</b>	<b>Statische Optimierung: Unbeschränkter Fall</b>	<b>27</b>
2.1	Optimalitätsbedingungen . . . . .	27
2.2	Rechnergestützte Minimierungsverfahren: Grundlagen . . . . .	30
2.3	Liniensuchverfahren . . . . .	32
2.3.1	Wahl der Schrittweite . . . . .	33
2.3.2	Wahl der Suchrichtung . . . . .	38
2.4	Methode der Vertrauensbereiche . . . . .	53
2.5	Direkte Suchverfahren . . . . .	55
2.6	Beispiel: Rosenbrock's „Bananenfunktion“ . . . . .	59
2.7	Literatur . . . . .	65
<b>3</b>	<b>Statische Optimierung mit Beschränkungen</b>	<b>66</b>
3.1	Optimalitätsbedingungen . . . . .	67
3.1.1	Optimalitätsbedingungen basierend auf zulässigen Richtungen . . . . .	67
3.1.2	Optimalitätsbedingungen mit Lagrange-Multiplikatoren . . . . .	70
3.2	Rechnergestützte Optimierungsverfahren . . . . .	83
3.2.1	Methode der aktiven Beschränkungen . . . . .	84
3.2.2	Gradienten-Projektionsmethode . . . . .	86
3.2.3	Reduzierte Gradientenmethode . . . . .	92
3.2.4	Sequentielle quadratische Programmierung (SQP) . . . . .	97
3.2.5	Methode der Straf- und Barrierefunktionen . . . . .	103
3.3	Beispiel: Rosenbrock's „Bananenfunktion“ . . . . .	106
3.4	Software-Übersicht . . . . .	109

---

3.5	Literatur . . . . .	111
<b>4</b>	<b>Dynamische Optimierung</b>	<b>112</b>
4.1	Grundlagen der Variationsrechnung . . . . .	112
4.1.1	Problemformulierung . . . . .	112
4.1.2	Optimalitätsbedingungen . . . . .	113
4.1.3	Stückweise stetig differenzierbare Extremale . . . . .	123
4.2	Entwurf von Optimalsteuerungen . . . . .	126
4.2.1	Problemformulierung . . . . .	126
4.2.2	Existenz einer Lösung . . . . .	127
4.2.3	Variationsformulierung . . . . .	129
4.2.4	Minimumsprinzip von Pontryagin . . . . .	145
4.2.5	Minimumsprinzip für eingangsaffine Systeme . . . . .	151
4.2.6	Der singuläre Fall . . . . .	158
4.3	Literatur . . . . .	164

# Vorwort

Wesentliche Teile dieses Skriptums wurden von Prof. Dr.-Ing. Knut GRAICHEN und Univ.-Prof. Dr. techn. Andreas KUGI verfasst. Ihnen gebührt aufrichtiger Dank dafür. Fragen sowie Korrektur- und Verbesserungsvorschläge zu diesem Skriptum können Sie jederzeit an Andreas STEINBÖCK richten.

# 1 Einleitung

Unter *Optimierung* versteht man gemeinhin die Suche nach einem im Sinne einer bestimmten Zielsetzung bestmöglichen Punkt (optimale Lösung) in einem Entscheidungsraum, wobei bei dieser Suche meist Nebenbedingungen zu berücksichtigen sind. Zur Systematisierung solcher Entscheidungsfindungsprozesse können mathematische Formulierungen und Lösungen von Optimierungsaufgaben (Optimierungsproblemen) verwendet werden. Das vorliegende Skriptum gibt einen Überblick über die mathematische Formulierung und Lösung von Optimierungsaufgaben.

Es wird grundsätzlich zwischen *statischen* und *dynamischen* Optimierungsproblemen unterschieden:

- *Statisches Optimierungsproblem*: Minimierung einer Funktion mit Optimierungsvariablen, die Elemente eines finit-dimensionalen Raumes (z. B. dem Euklidischen Raum) sind
- *Dynamisches Optimierungsproblem*: Minimierung eines Funktionals mit Optimierungsvariablen, die Elemente eines unendlich-dimensionalen Raumes sind (z. B. Zeitfunktionen)

In diesem Abschnitt soll anhand von Beispielen der prinzipielle Unterschied zwischen statischen und dynamischen Optimierungsaufgaben verdeutlicht werden.

## 1.1 Statische Optimierungsprobleme

Unter einem *statischen Optimierungsproblem* wird das Minimieren einer Funktion  $f(\mathbf{x})$  unter Berücksichtigung gewisser Nebenbedingungen verstanden, wobei die Optimierungsvariablen  $\mathbf{x}$  Elemente des Euklidischen Raumes  $\mathbb{R}^n$  sind.

### 1.1.1 Mathematische Formulierung

Die Standardformulierung eines statischen Optimierungsproblems lautet

$$\min_{\mathbf{x} \in \mathbb{R}^n} f(\mathbf{x}) \quad \text{Kostenfunktion} \quad (1.1a)$$

$$\text{u.B.v. } g_i(\mathbf{x}) = 0, \quad i = 1, \dots, p \quad \text{Gleichungsbeschränkungen} \quad (1.1b)$$

$$h_i(\mathbf{x}) \leq 0, \quad i = 1, \dots, q \quad \text{Ungleichungsbeschränkungen.} \quad (1.1c)$$

Ist ein Optimierungsproblem ohne die Gleichungs- und Ungleichungsbeschränkungen (1.1b) und (1.1c) gegeben, spricht man von einem *unbeschränkten Optimierungsproblem*. Im allgemeinen Fall, d. h. unter Berücksichtigung der Nebenbedingungen (1.1b)–(1.1c), handelt es sich um ein *beschränktes Optimierungsproblem*.

Die Menge  $\mathcal{X} \subset \mathbb{R}^n$ , die die Gleichungs- und Ungleichungsbeschränkungen (1.1b) und (1.1c) erfüllt,

$$\mathcal{X} = \{ \mathbf{x} \in \mathbb{R}^n \mid g_i(\mathbf{x}) = 0, i = 1, \dots, p, \quad h_i(\mathbf{x}) \leq 0, i = 1, \dots, q \} \quad (1.2)$$

wird als *zulässiges Gebiet* oder *zulässige Menge* (englisch: *admissible* oder *feasible region*) und jedes  $\mathbf{x} \in \mathcal{X}$  als *zulässiger Punkt* bezeichnet. Damit lässt sich das statische Optimierungsproblem (1.1) auch in der äquivalenten Form

$$\min_{\mathbf{x} \in \mathcal{X}} f(\mathbf{x}) \quad (1.3)$$

angeben. Im Falle von unbeschränkten Problemen gilt  $\mathcal{X} = \mathbb{R}^n$ .

Es ist direkt ersichtlich, dass  $\mathcal{X}$  nicht die leere Menge sein darf, da das Optimierungsproblem (1.3) ansonsten keine Lösung besitzt. Eine weitere notwendige Bedingung für  $\mathcal{X}$  kann aus den Gleichungsbeschränkungen (1.1b) abgeleitet werden, da sich durch die algebraischen Restriktionen  $g_i(\mathbf{x}) = 0$  die Anzahl der freien Optimierungsvariablen  $\mathbf{x} \in \mathbb{R}^n$  auf  $n - p$  reduziert. Somit darf die Anzahl  $p$  der Gleichungsbeschränkungen (1.1b) nicht größer als die Anzahl der Optimierungsvariablen  $\mathbf{x} \in \mathbb{R}^n$  sein, da die zulässige Menge  $\mathcal{X}$  ansonsten leer wäre.

In der Vergangenheit hat sich die Formulierung als Minimierungsproblem (1.1a) standardisiert. Analog dazu kann ein Maximierungsproblem ebenfalls als Minimierungsproblem gemäß (1.1a) geschrieben werden:

$$\max_{\mathbf{x} \in \mathcal{X}} f(\mathbf{x}) = \min_{\mathbf{x} \in \mathcal{X}} -f(\mathbf{x}).$$

Neben der Bezeichnung statische Optimierung werden häufig auch die Begriffe *Mathematische Programmierung* oder *Endlich-Dimensionale Optimierung* verwendet.

Der Begriff „Programmierung“ ist eher im Sinne von „Planung“ zu verstehen als im Sinne der Erstellung eines Computerprogramms. Er wurde schon Mitte der 1940er Jahre von George Dantzig, einem der Begründer der Linearen Optimierung, geprägt, bevor Computer zur Lösung linearer Optimierungsprobleme eingesetzt wurden.

Unterschieden werden bei statischen Optimierungsproblemen häufig folgende Klassen:

- *Lineare Programmierung*: Die Kostenfunktion und die Beschränkungen sind linear (genauer affin).
- *Quadratische Programmierung*: Die Kostenfunktion ist quadratisch, während die Beschränkungen linear (genauer affin) sind.
- *Nichtlineare Programmierung*: Die Kostenfunktion oder mindestens eine Beschränkung ist nichtlinear.
- *Konvexe Programmierung*: Konvexität ist ein mathematischer Begriff, der im Hinblick auf die Optimierung eine besondere Bedeutung spielt, denn er erlaubt es, eine Klasse von Optimierungsproblemen zu formulieren, für die die notwendigen Optimalitätsbedingungen erster Ordnung gleichzeitig hinreichende Bedingungen für ein globales Optimum sind.
- *Integer-Programmierung*: Alle Variablen sind diskret.
- *Mixed-Integer-Programmierung*: Kontinuierliche und diskrete Variablen treten auf.

### 1.1.2 Beispiele

Insbesondere die *lineare Programmierung* wird häufig bei ökonomischen Fragestellungen, wie Produktions-, Planungs- oder Investitionsproblemen, eingesetzt. Das folgende Beispiel ist ein stark vereinfachtes Beispiel einer Portfolio-Optimierung.

*Beispiel 1.1 (Portfolio-Optimierung).* Ein Anleger möchte 10.000 Euro gewinnbringend investieren und hat die Auswahl zwischen drei Aktienfonds mit unterschiedlicher Gewinnerwartung und Risikoeinstufung:

Fonds	erwarteter Gewinn/Jahr	Risikoeinstufung
A	10 %	4
B	7 %	2
C	4 %	1

Der Anleger möchte nach einem Jahr mindestens 600 Euro Gewinn erzielen. Andererseits möchte er sein Geld eher konservativ anlegen, d. h. er möchte mindestens 4.000 Euro in Fonds C investieren und das Risiko minimieren. Wie muss der Anleger die 10.000 Euro verteilen, damit diese Kriterien erfüllt werden?

Zunächst werden die Optimierungsvariablen  $x_1$ ,  $x_2$ ,  $x_3$  eingeführt, die den prozentualen Anteil der investierten 10.000 Euro an den jeweiligen Fonds A, B, C kennzeichnen. Dabei kann  $x_3$  durch die Beziehung

$$x_3 = 1 - x_1 - x_2$$

ersetzt werden. Der geforderte Mindestgewinn von 600 Euro lässt sich als die Beschränkung

$$10.000[0.1x_1 + 0.07x_2 + 0.04(1 - x_1 - x_2)] \geq 600 \quad \Rightarrow \quad 6x_1 + 3x_2 \geq 2 \quad (1.4)$$

ausdrücken. Die Mindestanlage von 4.000 Euro in Fonds C führt zu

$$10.000(1 - x_1 - x_2) \geq 4.000 \quad \Rightarrow \quad x_1 + x_2 \leq 0.6. \quad (1.5)$$

Des Weiteren müssen  $x_1 \geq 0$ ,  $x_2 \geq 0$  und  $x_3 \geq 0$  erfüllt sein. Das Ziel ist die Minimierung des Anlagerisikos, was sich durch die Funktion

$$f(\mathbf{x}) = 4x_1 + 2x_2 + (1 - x_1 - x_2) = 1 + 3x_1 + x_2 \quad (1.6)$$

ausdrücken lässt.

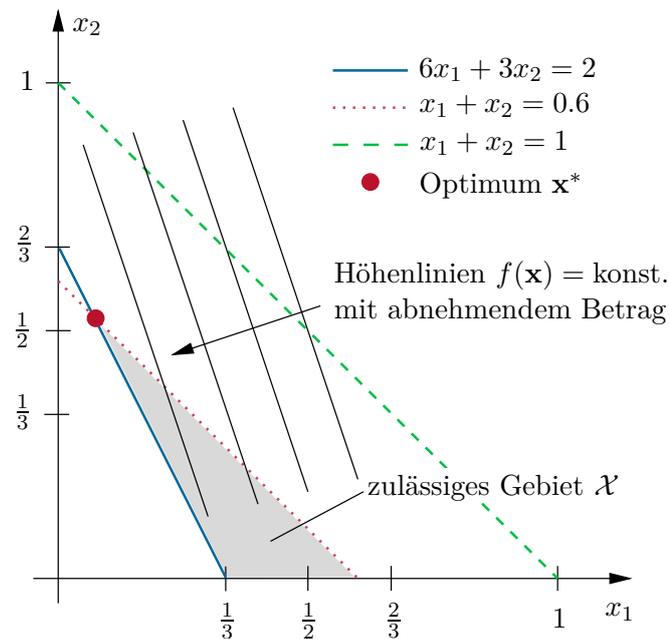


Abbildung 1.1: Veranschaulichung der Portfolio-Optimierung in Beispiel 1.1.

Somit kann das statische Optimierungsproblem in der Form

$$\min_{\mathbf{x} \in \mathbb{R}^2} f(\mathbf{x}) = 1 + 3x_1 + x_2 \quad (1.7a)$$

$$\text{u.B.v. } 6x_1 + 3x_2 \geq 2 \quad (1.7b)$$

$$x_1 + x_2 \leq 0.6 \quad (1.7c)$$

$$x_1 + x_2 \leq 1 \quad (1.7d)$$

$$x_1, x_2 \geq 0 \quad (1.7e)$$

geschrieben werden. Abbildung 1.1 stellt die einzelnen Beschränkungen sowie das zulässige Gebiet grafisch dar. Aus dem Verlauf der Höhenlinien  $f(\mathbf{x}) = \text{konst.}$  der Kostenfunktion (1.7a) ist direkt ersichtlich, dass der Punkt  $\mathbf{x}^*$  jene Ecke des zulässigen Gebiets  $\mathcal{X}$  mit dem niedrigsten Wert von  $f(\mathbf{x})$  ist. Somit ergibt sich für die optimale Verteilung der 10.000 Euro auf die einzelnen Fonds

$$x_1^* = \frac{1}{15}, \quad x_2^* = \frac{8}{15}, \quad x_3^* = \frac{6}{15}. \quad (1.8)$$

Das folgende (akademische) Beispiel der quadratischen Programmierung soll den Einfluss von Beschränkungen auf eine optimale Lösung verdeutlichen.

*Beispiel 1.2.* Betrachtet wird das (zunächst) unbeschränkte Problem

$$\min_{\mathbf{x} \in \mathbb{R}^2} f(\mathbf{x}) = (x_1 - 2)^2 + (x_2 - 1)^2. \quad (1.9)$$

Die Höhenlinien  $f(\mathbf{x}) = \text{konst.}$  der Funktion  $f(\mathbf{x})$  sind in Abbildung 1.2 in Abhängigkeit der beiden Optimierungsvariablen  $\mathbf{x} = [x_1 \ x_2]^T$  dargestellt. Es ist direkt ersichtlich, dass das Minimum  $f(\mathbf{x}^*) = 0$  an der Stelle  $\mathbf{x}^* = [2 \ 1]^T$  auftritt.

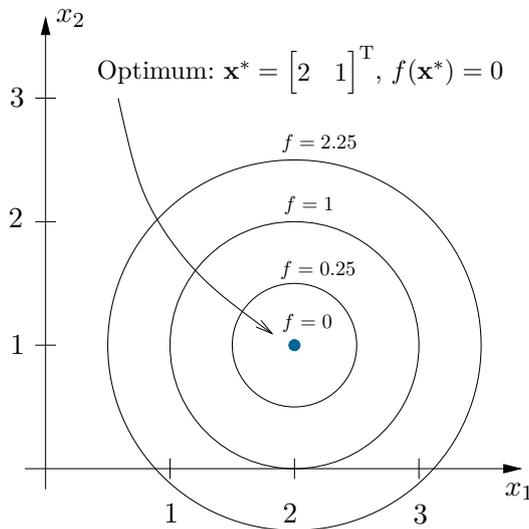


Abb. 1.2: Geometrische Darstellung des unbeschränkten Optimierungsproblems (1.9).

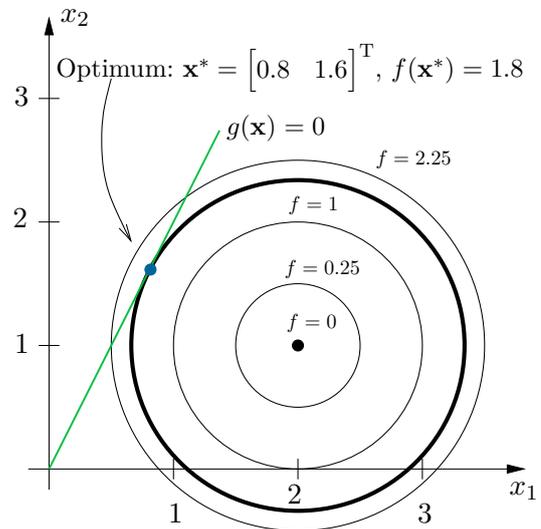


Abb. 1.3: Geometrische Darstellung des beschränkten Optimierungsproblems (1.9), (1.10).

Um den Einfluss verschiedener Beschränkungen zu untersuchen, wird zunächst eine zusätzliche Gleichungsbeschränkung der Form (1.1b) betrachtet

$$g(\mathbf{x}) = x_2 - 2x_1 = 0. \quad (1.10)$$

Die Gleichungsbeschränkung entspricht einer algebraischen Zwangsbedingung, wodurch lediglich noch eine Optimierungsvariable frei wählbar ist. Geometrisch interpretiert bedeutet dies, dass eine mögliche Lösung auf der Geraden liegen muss, die durch (1.10) definiert wird, siehe Abbildung 1.3. Die optimale Lösung liegt dabei auf dem tangentialen Berührungspunkt der Geraden  $g(\mathbf{x}) = 0$  mit der Höhenlinie  $f(\mathbf{x}) = 1.8$ .

Anstelle der Gleichungsbeschränkung (1.10) wird nun die Ungleichungsbeschränkung

$$h_1(\mathbf{x}) = x_1 + x_2 - 2 \leq 0 \quad (1.11)$$

betrachtet, wodurch sich die Menge der zulässigen Punkte  $\mathbf{x} = [x_1 \ x_2]^T$  auf das Gebiet links unterhalb der Geraden  $h_1(\mathbf{x}) = 0$  beschränkt (siehe Abbildung 1.4). Das Optimum  $f(\mathbf{x}^*) = 0.5$  an der Stelle  $\mathbf{x}^* = [1.5 \ 0.5]^T$  befindet sich an der Grenze des zulässigen Gebiets und liegt, wie im vorherigen Szenario, auf einer Höhenlinie, die die Gerade  $h_1(\mathbf{x}) = 0$  tangential berührt.

Zusätzlich zur ersten Ungleichungsbeschränkung (1.11) soll eine weitere Ungleichung der Form

$$h_2(\mathbf{x}) = x_1^2 - x_2 \leq 0 \quad (1.12)$$

betrachtet werden, durch die sich die Menge der zulässigen Punkte weiter verkleinert (Abbildung 1.5). Der optimale Punkt  $\mathbf{x}^* = [1 \ 1]^T$  mit dem Minimum  $f(\mathbf{x}^*) = 1$  liegt nun im Schnittpunkt der Kurven  $h_1(\mathbf{x}) = 0$  und  $h_2(\mathbf{x}) = 0$ , d. h. beide Beschränkungen (1.11) und (1.12) sind aktiv.

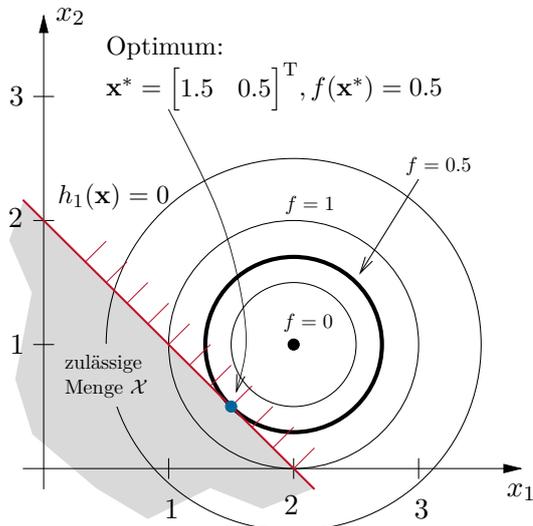


Abb. 1.4: Geometrische Darstellung des beschränkten Optimierungsproblems (1.9), (1.11).

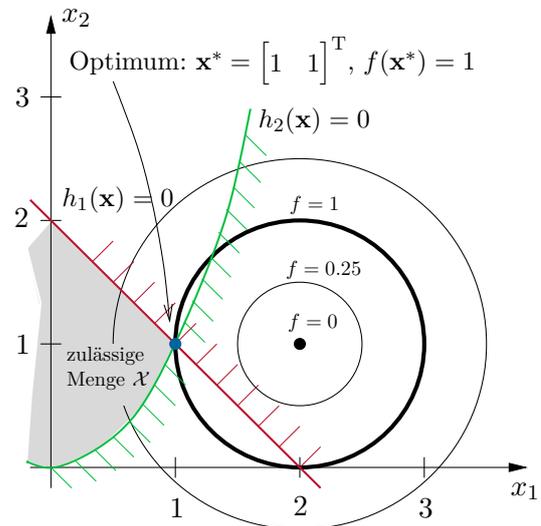


Abb. 1.5: Geometrische Darstellung des beschränkten Optimierungsproblems (1.9), (1.11), (1.12).

Das obige Beispiel 1.2 verdeutlicht den Einfluss von Gleichungs- und Ungleichungsbeschränkungen auf die Lösung (und Lösbarkeit) des Optimierungsproblems (1.1). Die systematische Untersuchung von statischen Optimierungsproblemen sowie die zugehörigen Verfahren zur numerischen Lösung werden in den folgenden Abschnitten behandelt.

## 1.2 Dynamische Optimierungsprobleme

Bei den Problemstellungen der statischen Optimierung im vorherigen Abschnitt 1.1 stellen die Optimierungsvariablen  $\mathbf{x}$  Elemente aus einem finit-dimensionalen Raum, meist dem Euklidischen Raum  $\mathbb{R}^n$ , dar. Bei der dynamischen Optimierung hingegen sind Funktionen einer unabhängigen Variablen zu bestimmen. Da es sich bei der unabhängigen Variablen meistens um die Zeit  $t$  handelt, wird in diesem Zusammenhang von *dynamischer Optimierung* gesprochen.

### 1.2.1 Mathematische Formulierung

Die generelle Struktur eines dynamischen Optimierungsproblems lautet

$$\min_{\mathbf{u}(\cdot)} J(\mathbf{u}) = \varphi(t_1, \mathbf{x}(t_1)) + \int_{t_0}^{t_1} l(t, \mathbf{x}(t), \mathbf{u}(t)) dt \quad \text{Kostenfunktional} \quad (1.13a)$$

$$\text{u.B.v. } \dot{\mathbf{x}} = \mathbf{f}(t, \mathbf{x}, \mathbf{u}), \quad \mathbf{x}(t_0) = \mathbf{x}_0 \quad \text{Systemdynamik \& AB} \quad (1.13b)$$

$$\boldsymbol{\psi}(t_1, \mathbf{x}(t_1)) = \mathbf{0} \quad \text{Endbedingungen (EB)} \quad (1.13c)$$

$$h_i(\mathbf{x}, \mathbf{u}) \leq 0, \quad i = 1, \dots, q \quad \text{Ungleichungsbeschr.} \quad (1.13d)$$

Dabei stellt  $\mathbf{u} \in \mathbb{R}^m$  die Eingangsgröße des nichtlinearen Systems (1.13b) mit dem Zustand  $\mathbf{x} \in \mathbb{R}^n$  dar. Zusätzlich zu den Anfangsbedingungen in (1.13b) sind häufig Endbedingungen der Form (1.13c) gegeben, um z. B. einen gewünschten Zustand  $\mathbf{x}_f$  zur Endzeit  $t_1$  zu erreichen (also  $\boldsymbol{\psi}(t_1, \mathbf{x}(t_1)) = \mathbf{x}(t_1) - \mathbf{x}_f$ ). In der Praxis treten häufig Ungleichungsbeschränkungen (1.13d) auf, die z. B. die Begrenzung einer Stellgröße oder Sicherheitsschranken eines Zustandes darstellen können.

Die Problemstellung der dynamischen Optimierung besteht nun darin, eine Eingangstrajektorie  $\mathbf{u}(t)$ ,  $t \in [t_0, t_1]$  derart zu finden, dass die Zustandstrajektorie  $\mathbf{x}(t)$ ,  $t \in [t_0, t_1]$  des dynamischen Systems (1.13b), die Endbedingungen (1.13c) und Beschränkungen (1.13d) erfüllt und gleichzeitig das Kostenfunktional (1.13a) minimiert wird. Abhängig davon, ob  $t_1$  vorgegeben oder unbekannt ist, spricht man von einer *festen* oder *freien Endzeit*  $t_1$ .

Neben der Bezeichnung dynamische Optimierung werden häufig auch die Begriffe *Unendlich-Dimensionale Optimierung*, *Optimalsteuerungsproblem* oder *Dynamische Programmierung* verwendet. Im Folgenden sind einige Beispiele angegeben, um die Problem- und Aufgabenstellung der dynamischen Optimierung zu erläutern.

### 1.2.2 Beispiele

*Beispiel 1.3 (Inverses Pendel).* Ein klassisches Problem in der Regelungstechnik ist das inverse Pendel, das an einem Wagen drehbar befestigt ist. Als Beispielproblem soll das seitliche Versetzen des Pendels betrachtet werden

$$\min_{u(\cdot)} \quad J(u) = \int_0^{t_1} 1 + c u^2 \, dt, \quad (1.14a)$$

$$\text{u.B.v.} \quad \begin{bmatrix} 1 & \varepsilon \cos \theta \\ \cos \theta & 1 \end{bmatrix} \begin{bmatrix} \ddot{x} \\ \ddot{\theta} \end{bmatrix} = \begin{bmatrix} \varepsilon \dot{\theta}^2 \sin \theta + u \\ -\sin \theta \end{bmatrix}, \quad \varepsilon = m/(M + m) \quad (1.14b)$$

$$\mathbf{x}(0) = [0 \ 0 \ \pi \ 0]^T, \quad \mathbf{x}(t_1) = [1 \ 0 \ \pi \ 0]^T, \quad (1.14c)$$

$$-1 \leq u \leq 1. \quad (1.14d)$$

Die vereinfachten Bewegungsgleichungen (1.14b) für die Zustände  $\mathbf{x} = [x \ \dot{x} \ \theta \ \dot{\theta}]^T$  sind normiert. Der Eingang  $u$  stellt die am Wagen angreifende Kraft dar und ist durch (1.14d) beschränkt. Die Masse des Pendels wird mit  $m$ , diejenige des Wagens mit  $M$  bezeichnet. Abbildung 1.6 zeigt exemplarisch das seitliche Versetzen des Pendels, um die Bewegung zu verdeutlichen.

Das Kostenfunktional (1.14a) und somit der Charakter des Optimierungsproblems hängt von dem Parameter  $c$  ab. Für  $c = 0$  ergibt sich die Aufgabe, die Endzeit  $t_1$  zu

minimieren

$$J(u) = \int_0^{t_1} 1 \, dt = t_1. \quad (1.15)$$

Für  $c > 0$  wird der Eingang  $u$  im Kostenfunktional und somit der Aspekt der Energieoptimalität mitberücksichtigt.

Abbildung 1.7 zeigt die optimalen Trajektorien für den Parameterwert  $\varepsilon = 0.5$  sowie für die Werte  $c = 0$ ,  $c = 0.25$  und  $c = 1$ . Für  $c = 0$  weist der Eingang  $u$  ein Bang-bang-Verhalten auf, während für  $c > 0$  die Steueramplituden kleiner werden und die benötigte Zeit  $t_1$  zunimmt.

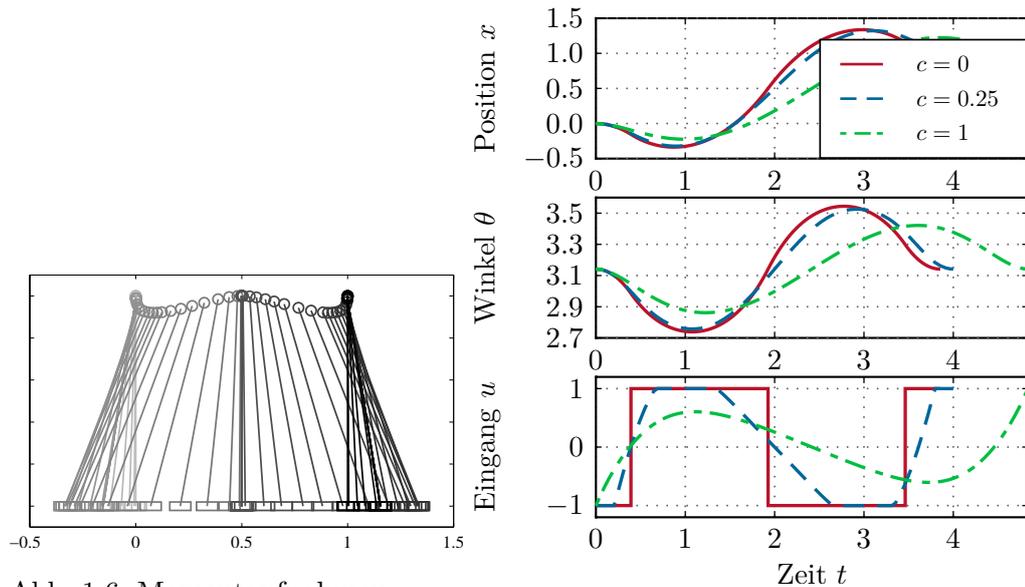


Abb. 1.6: Momentaufnahmen

beim Versetzen des Abb. 1.7: Optimale Trajektorien beim Ver-  
inversen Pendels. setzen des inversen Pendels.

Das inverse Pendel ist ein gutes Beispiel um zu verdeutlichen, dass nicht zu jedem Optimierungsproblem eine Lösung existiert, insbesondere wenn die Endzeit  $t_1$  nicht festgelegt ist. Wie aus Abbildung 1.7 ersichtlich, vergrößert sich die Endzeit  $t_1$  bei zunehmender Gewichtung von  $u^2$  im Vergleich zum zeitoptimalen Anteil in dem Kostenfunktional (1.14a). Wenn reine Energieoptimalität gefordert würde, d. h.

$$J(u) = \int_0^{t_1} u^2 \, dt, \quad (1.16)$$

hätte das Optimierungsproblem keine Lösung, da das Versetzen des Pendels dann unendlich langsam mit  $t_1 \rightarrow \infty$  ablaufen würde.

**Beispiel 1.4 (Goddard-Rakete [1.1, 1.2]).** Ein klassisches Optimierungsproblem aus der Raumfahrt ist die Maximierung der Flughöhe einer Rakete unter dem Einfluss von Luftreibung und Erdbeschleunigung. Dieses Problem wurde von dem amerikani-

schen Raketenpionier Robert H. Goddard im Jahr 1919 aufgestellt und kann in der normierten Form

$$\min_{u(\cdot)} -h(t_1) \quad (1.17a)$$

$$\text{u.B.v. } \dot{h} = v, \quad \dot{v} = \frac{u - D(h, v)}{m} - \frac{1}{h^2}, \quad \dot{m} = -\frac{u}{c}, \quad (1.17b)$$

$$h(0) = 1, \quad v(0) = 0, \quad m(0) = 1, \quad m(t_1) = 0.6, \quad (1.17c)$$

$$0 \leq u \leq 3.5 \quad (1.17d)$$

geschrieben werden.

Die Zustandsgrößen sind die Flughöhe  $h$ , die Geschwindigkeit  $v$  und die Masse  $m$  der Rakete. Die Luftreibung  $D(h, v)$  hängt über die Funktion

$$D(h, v) = D_0 v^2 e^{\beta(1-h)} \quad (1.18)$$

von den Zuständen  $h$  und  $v$  ab. Die Randbedingungen in (1.17c) umfassen die normierten Anfangsbedingungen sowie die Endbedingung für  $m(t_1)$ , die dem Leergewicht der Rakete ohne Treibstoff entspricht. Der Eingang des Systems ist der Schub  $u$ , der innerhalb der Beschränkungen (1.17d) liegen muss.

In Abbildung 1.8 sind die optimalen Trajektorien für die Goddard-Rakete dargestellt. Die verwendeten Parameterwerte lauten  $c = 0.5$ ,  $D_0 = 310$  und  $\beta = 500$ . Der Schub  $u(t)$  ist am Anfang maximal und weist dann einen parabelförmigen Verlauf auf, bevor der Treibstoff verbraucht ist. Dieses Verhalten wird durch den Luftwiderstand  $D(h, v)$  hervorgerufen, der mit zunehmender Höhe abnimmt. Es ist somit „optimaler“, im Falle eines hohen Luftwiderstandes nicht mit vollem Schub zu fliegen.

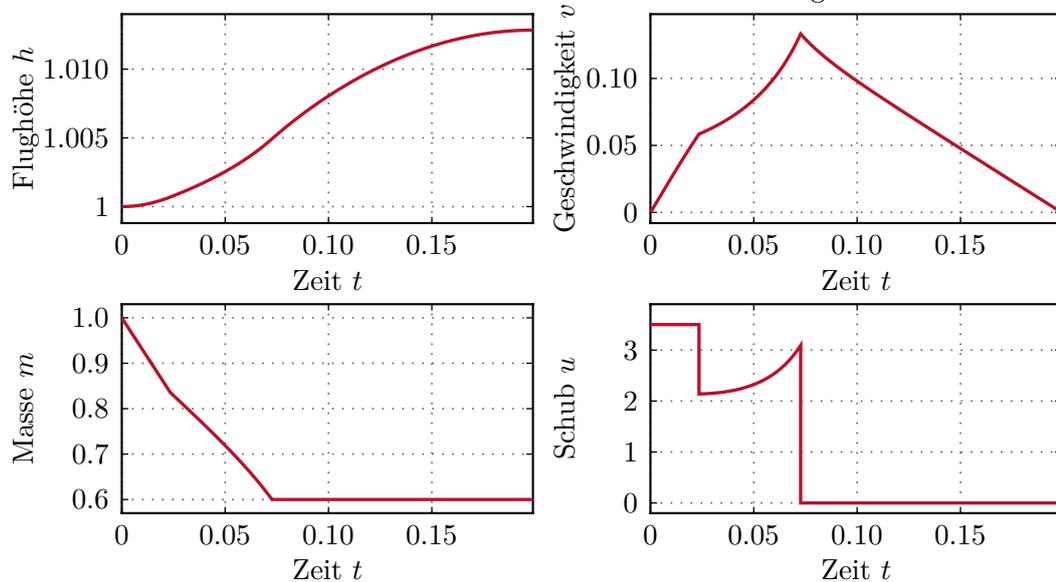


Abbildung 1.8: Trajektorien für die Goddard-Rakete in Beispiel 1.4.

*Beispiel 1.5 (Ökonomisches Modell [1.3, 1.4]).* Ein weiterer Anwendungszweig der dynamischen Optimierung sind wirtschaftliche Prozesse. Das folgende Beispiel beschreibt das Verhalten eines typischen Konsumenten, der Konsum, Freizeit und Bildung über die Lebensdauer maximieren will. Der Bildungsgrad  $B$  und das Kapital  $K$  eines durchschnittlichen Konsumenten lassen sich durch folgendes Modell beschreiben

$$\dot{B} = \overbrace{B^\varepsilon u_2 u_3}^{\text{Weiterbildung}} - \overbrace{\delta B}^{\text{Vergessen}}, \quad B(0) = B_0 \quad (1.19a)$$

$$\dot{K} = \underbrace{iK}_{\text{Verzinsung}} + \underbrace{B u_2 g(u_3)}_{\text{Einkommen}} - \underbrace{u_1}_{\text{Konsum}}, \quad K(0) = K_0. \quad (1.19b)$$

Die Eingangsgrößen sind der Konsum  $u_1$ , der Anteil der Arbeitszeit an der Gesamtzeit  $u_2$  sowie der Anteil der Fortbildungszeit an der Arbeitszeit  $u_3$ . Die Eingänge unterliegen den Beschränkungen

$$u_1 > 0, \quad 0 \leq u_2 \leq 1, \quad 0 \leq u_3 < 1. \quad (1.20)$$

Das Optimierungsziel des Konsumenten ist die Maximierung von Konsum, Freizeit und Bildung über die Lebensdauer von  $t_1 = 75$  Jahren, was in dem (zu minimierenden) Kostenfunktional

$$J(\mathbf{u}) = -K^\kappa(t_1) - \int_{t_0}^{t_1} U(t, u_1, u_2, B) e^{-\rho t} dt. \quad (1.21)$$

ausgedrückt ist. Der Integralanteil

$$U(t, u_1, u_2, B) = \alpha_0 u_1^\alpha + \beta_0 (1 - u_2)^\beta + \gamma_0 t B^\gamma \quad (1.22)$$

gewichtet dabei den Konsum  $u_1$ , die Freizeit  $1 - u_2$  und den Bildungsgrad  $B$ , während der Endwert  $-K^\kappa(t_1)$  in (1.21) zusätzlich das Vererbungskapital berücksichtigt.

Die optimalen Zeitverläufe des Bildungsgrades  $B(t)$  und des Kapitals  $K(t)$  sind in Abbildung 1.9 dargestellt. Die Funktion  $g(u_3)$  in (1.19b) ist durch die Parabel  $g(u_3) = 1 - (1 - a)u_3 - a u_3^2$  gegeben. Die verwendeten Parameterwerte lauten  $a = 0.3$ ,  $\alpha = -1$ ,  $\alpha_0 = -1$ ,  $\beta = -0.5$ ,  $\beta_0 = -1$ ,  $\gamma = 0.2$ ,  $\gamma_0 = 5$ ,  $\kappa = 0.2$ ,  $\rho = 0.01$ ,  $\varepsilon = 0.35$ ,  $\delta = 0.01$ ,  $i = 0.04$ ,  $B_0 = 1$  und  $K_0 = 30$ .

Die ersten 17 Jahre stellen die Lernphase dar (d. h.  $u_3 = 1$ ). Daraufhin folgt eine lange Arbeitsphase von 34 Jahren mit einem hohen Maß an Weiterbildung, bevor in den nächsten 10 Jahren (52.–61. Lebensjahr) eine reine Arbeitsphase mit zusätzlich reduzierter Arbeitszeit  $u_2$  stattfindet. Ab dem 62. Lebensjahr setzt der Ruhestand ein. Der Bildungsgrad  $B$  ist besonders hoch im Alter von 30–60 Jahren. Das Kapital  $K$  ist negativ während der ersten Lebenshälfte, was der Aufnahme eines Kredites entspricht. Im Laufe des Lebens wird dies aber durch das steigende Einkommen kompensiert.

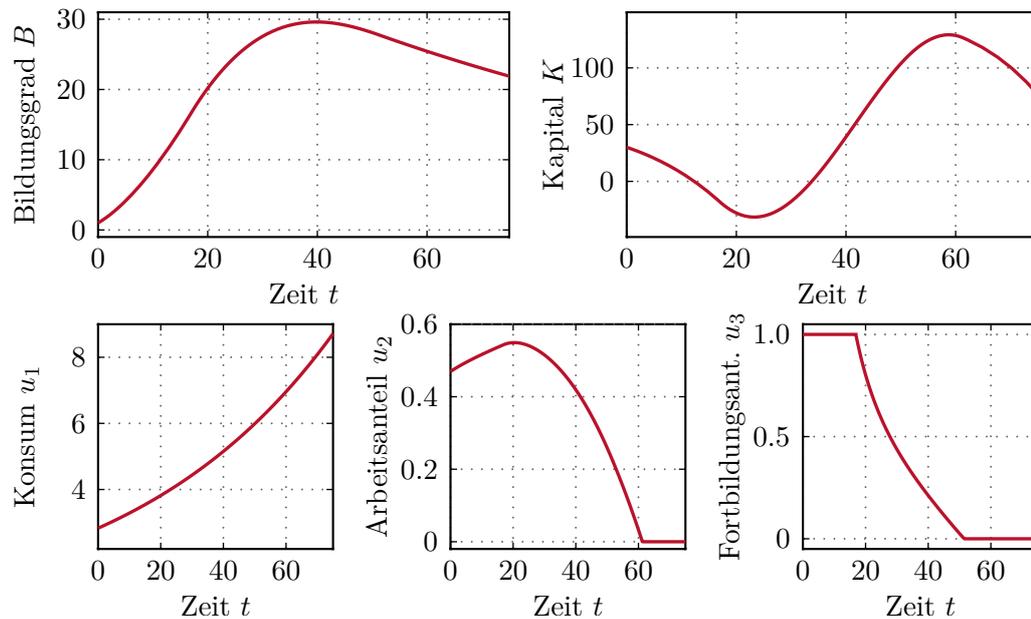


Abbildung 1.9: Optimale Trajektorien für das Konsumentenverhalten in Beispiel 1.5.

## 1.3 Mathematische Grundlagen

In diesem Abschnitt werden kurz einige mathematische Begriffe und Grundkonzepte erläutert, die das Verständnis der weiteren Kapitel erleichtern.

### 1.3.1 Infimum, Supremum, Minimum und Maximum

**Definition 1.1 (Infimum und Supremum).** Es sei  $\mathcal{Y} \subset \mathbb{R}$  eine nichtleere Menge. Das Infimum von  $\mathcal{Y}$ , kurz  $\inf \mathcal{Y}$  geschrieben, bezeichnet die größte untere Schranke von  $\mathcal{Y}$ , d. h. es existiert eine Zahl  $\alpha$  so, dass gilt

- (a)  $x \geq \alpha$  für alle  $x \in \mathcal{Y}$
- (b) für alle  $\bar{\alpha} > \alpha$  existiert ein  $x \in \mathcal{Y}$  so, dass  $x < \bar{\alpha}$ .

Das Supremum von  $\mathcal{Y}$ , kurz  $\sup \mathcal{Y}$  geschrieben, bezeichnet die kleinste obere Schranke von  $\mathcal{Y}$ , d. h. es existiert eine Zahl  $\alpha$  so, dass gilt

- (a)  $x \leq \alpha$  für alle  $x \in \mathcal{Y}$
- (b) für alle  $\bar{\alpha} < \alpha$  existiert ein  $x \in \mathcal{Y}$  so, dass  $x > \bar{\alpha}$ .

Existiert für eine nichtleere Menge  $\mathcal{Y}$  ein Infimum oder ein Supremum, so muss dieses nicht automatisch in  $\mathcal{Y}$  enthalten sein. Als Beispiel dazu betrachte man die Menge  $\mathcal{Y} = \{x \in \mathbb{R} \mid x > 0\} = (0, +\infty)$ . In diesem Fall gilt offensichtlich  $0 = \inf \mathcal{Y} \notin \mathcal{Y}$ .

Für die folgende Definition wird angenommen, dass  $\mathcal{X} \subset \mathbb{R}^n$  die zulässige Menge des betrachteten Optimierungsproblems gemäß (1.2) bezeichnet.

**Definition 1.2 (Globale und lokale Minima).** Die Funktion  $f(\mathbf{x})$  besitzt in  $\mathcal{X}$  an der Stelle  $\mathbf{x}^*$

- (a) ein *lokales Minimum*, falls ein  $\varepsilon > 0$  so existiert, dass gilt  $f(\mathbf{x}^*) \leq f(\mathbf{x})$  für alle  $\mathbf{x} \in \mathcal{U}_\varepsilon \cap \mathcal{X}$ , wobei  $\mathcal{U}_\varepsilon$  eine hinreichend kleine  $\varepsilon$ -Umgebung von  $\mathbf{x}^*$  bezeichnet,
- (b) ein *strikt lokales Minimum*, falls ein  $\varepsilon > 0$  so existiert, dass gilt  $f(\mathbf{x}^*) < f(\mathbf{x})$  für alle  $\mathbf{x} \in \mathcal{U}_\varepsilon \setminus \{\mathbf{x}^*\} \cap \mathcal{X}$
- (c) ein *globales (absolutes) Minimum*, falls  $f(\mathbf{x}^*) \leq f(\mathbf{x})$  für alle  $\mathbf{x} \in \mathcal{X}$ ,
- (d) ein *strikt (eindeutiges) globales Minimum*, falls  $f(\mathbf{x}^*) < f(\mathbf{x})$  für alle  $\mathbf{x} \in \mathcal{X} \setminus \{\mathbf{x}^*\}$ .

Abbildung 1.10 gibt eine grafische Darstellung der unterschiedlichen Arten von Minima. Definition 1.2 lässt sich direkt auf lokale und globale Maxima übertragen.

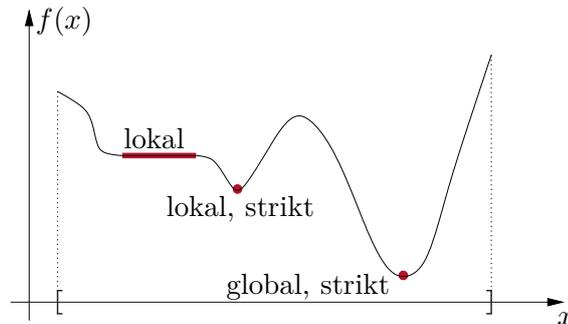


Abbildung 1.10: Verschiedene Minima einer Funktion  $f(x)$  mit  $x \in \mathbb{R}$ .

An dieser Stelle sei betont werden, dass ein Punkt  $\mathbf{x}^*$ , der die Funktion  $f(\mathbf{x})$  in der Menge  $\mathcal{X}$  minimiert bzw. maximiert, in  $\mathcal{X}$  enthalten sein muss. Ein Punkt  $\mathbf{x}$ , dessen Funktionswert  $f(\mathbf{x})$  gerade dem Infimum  $\inf\{f(\mathbf{x}) \mid \mathbf{x} \in \mathcal{X}\}$  bzw. dem Supremum  $\sup\{f(\mathbf{x}) \mid \mathbf{x} \in \mathcal{X}\}$  entspricht, muss jedoch nicht existieren, auch nicht außerhalb von  $\mathcal{X}$ .

Die Menge aller Minima wird oftmals auch in der Form

$$\mathcal{G} = \arg \min\{f(\mathbf{x}) \mid \mathbf{x} \in \mathcal{X}\} \quad (1.23)$$

angeschrieben, wobei die Menge sowohl leer sein kann als auch aus endlich oder unendlich vielen Punkten bestehen kann. Im Falle eines strikten globalen Minimums in  $\mathcal{X}$  versteht man unter dem Ausdruck  $\bar{\mathbf{x}} = \arg \min\{f(\mathbf{x}) \mid \mathbf{x} \in \mathcal{X}\}$  meist jene Funktion, die gerade den Punkt  $\bar{\mathbf{x}} \in \mathcal{X}$  zurückgibt, der die Funktion  $f(\mathbf{x})$  global minimiert.

### 1.3.2 Existenz von Minima und Maxima

Abbildung 1.11 zeigt drei Fälle, bei denen kein Minimum existiert. In Abbildung 1.11(a) ist das Infimum von  $f(x)$  in der Menge  $\mathcal{X} = (\alpha, \beta)$  durch  $f(\beta)$  gegeben. Da aber  $\mathcal{X}$

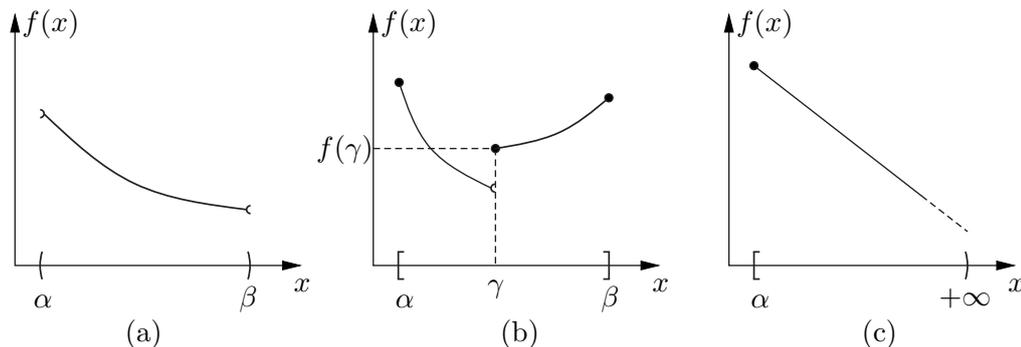


Abbildung 1.11: Nichtexistenz von Minima.

nicht abgeschlossen ist und somit  $\beta \notin \mathcal{X}$ , existiert in diesem Fall kein Minimum. In Abbildung 1.11(b) ist der linksseitige Grenzwert  $\lim_{x \rightarrow \gamma^-} f(x)$  das Infimum von  $f(x)$  in der Menge  $\mathcal{X} = [\alpha, \beta]$ . Auch in diesem Fall existiert auf Grund der Unstetigkeit von  $f(x)$  das Minimum nicht. Im letzten Fall, Abbildung 1.11(c), existiert das Minimum ebenfalls nicht, da  $f(x)$  in der unbeschränkten Menge  $\mathcal{X} = \{x \in \mathbb{R} \mid x \geq \alpha\}$  nach unten hin nicht beschränkt ist.

Der nachfolgende Satz gibt nun Bedingungen für die Existenz einer Lösung von Optimierungsproblemen an.

**Satz 1.1 (Weierstrass).** *Es sei  $\mathcal{X}$  eine nichtleere und kompakte (abgeschlossene und beschränkte) Menge und  $f : \mathcal{X} \rightarrow \mathbb{R}$  stetig auf  $\mathcal{X}$ . Dann ist die Menge aller Minima  $\mathcal{G} = \arg \min\{f(\mathbf{x}) \mid \mathbf{x} \in \mathcal{X}\}$  nichtleer und kompakt.*

Der Beweis dieses Satzes ist beispielsweise in [1.5, 1.6] zu finden. Es sei an dieser Stelle jedoch betont, dass Satz 1.1 nur eine hinreichende Bedingung für die Existenz einer optimalen Lösung angibt. Als Beispiel dazu betrachte man die Minimierungsaufgabe  $\min_{x \in (-1,1)} x^2$ , die zeigt, dass mit  $x = 0$  ein Minimum gegeben ist, obwohl die Menge  $\mathcal{X} = \{x \in \mathbb{R} \mid -1 < x < 1\}$  offen und damit nicht kompakt ist.

### 1.3.3 Gradient und Hessematrix

Die Berechnung von Ableitungen erster und zweiter Ordnung einer Kostenfunktion  $f(\mathbf{x})$  ist von fundamentaler Bedeutung in der Optimierung. Da im Falle von unstetigen Funktionen oder unstetigen Ableitungen Probleme auftreten können (sowohl numerischer als auch theoretischer Natur), wird oft angenommen, dass alle Funktionen eines Optimierungsproblems stetig und hinreichend oft differenzierbar sind. So nicht anders erwähnt, gilt dies auch für diese Vorlesung. Im Rahmen der Optimierungsalgorithmen spielen der Gradient und die Hessematrix eine bedeutende Rolle.

**Definition 1.3 (Gradient).** Es sei  $f : \mathcal{X} \rightarrow \mathbb{R}$  eine stetig differenzierbare Funktion, d. h.  $f \in C^1$ . Dann bezeichnet

$$(\nabla f)(\mathbf{x}) = \left( \frac{\partial f}{\partial \mathbf{x}} \right)^T = \begin{bmatrix} \frac{\partial f}{\partial x_1} \\ \vdots \\ \frac{\partial f}{\partial x_n} \end{bmatrix} \quad (1.24)$$

den Gradienten (also die 1. partielle Ableitung) von  $f(\mathbf{x})$  an der Stelle  $\mathbf{x} = [x_1 \ \dots \ x_n]^T$ .

**Definition 1.4 (Hessematrix).** Es sei  $f : \mathcal{X} \rightarrow \mathbb{R}$  eine zweifach stetig differenzierbare Funktion, d. h.  $f \in C^2$ . Dann bezeichnet

$$(\nabla^2 f)(\mathbf{x}) = \begin{bmatrix} \frac{\partial^2 f}{\partial x_1^2} & \cdots & \frac{\partial^2 f}{\partial x_1 \partial x_n} \\ \vdots & & \vdots \\ \frac{\partial^2 f}{\partial x_n \partial x_1} & \cdots & \frac{\partial^2 f}{\partial x_n^2} \end{bmatrix} \quad (1.25)$$

die Hessematrix (also die 2. partielle Ableitung) von  $f(\mathbf{x})$  an der Stelle  $\mathbf{x} = [x_1 \ \dots \ x_n]^T$ .

Im Falle von Funktionen  $f(x)$  mit nur einem skalaren Argument wird die  $\nabla$ -Notation häufig durch  $f'(x)$  und  $f''(x)$  ersetzt.

Aus der Stetigkeit der 2. partiellen Ableitungen folgt Kommutativität, d. h.

$$\frac{\partial^2 f}{\partial x_i \partial x_j} = \frac{\partial^2 f}{\partial x_j \partial x_i}.$$

Daraus ergibt sich  $(\nabla^2 f)(\mathbf{x}) = (\nabla^2 f)^T(\mathbf{x})$ , d. h. die Hessematrix ist symmetrisch. Folglich hat sie stets rein reelle Eigenwerte. In der Optimierung ist oft von Bedeutung, ob Hessematrizen positiv (semi-)definit sind. Diese Eigenschaft kann wie folgt untersucht werden.

**Satz 1.2 (Definitheit von Matrizen).** Die Definitheit einer symmetrischen Matrix  $\mathbf{A} \in \mathbb{R}^{n \times n}$  lässt sich durch folgende Bedingungen charakterisieren:

Matrix $\mathbf{A}$ ist	(a) für alle $\mathbf{p} \in \mathbb{R}^n$ mit $\mathbf{p} \neq \mathbf{0}$ gilt	(b) alle $n$ Eigen- werte $\lambda_i$ sind	(c) für alle $n$ Haupt- minoren $D_i$ gilt
positiv semi-definit:	$\mathbf{p}^T \mathbf{A} \mathbf{p} \geq 0$	$\geq 0$	-
positiv definit:	$\mathbf{p}^T \mathbf{A} \mathbf{p} > 0$	$> 0$	$D_i > 0$
negativ semi-definit:	$\mathbf{p}^T \mathbf{A} \mathbf{p} \leq 0$	$\leq 0$	-
negativ definit:	$\mathbf{p}^T \mathbf{A} \mathbf{p} < 0$	$< 0$	$(-1)^{i+1} D_i < 0$

Die Eigenwerte  $\lambda_i$ ,  $i = 1, \dots, n$  der Matrix  $\mathbf{A}$  sind die Lösungen der Gleichung

$$\det(\lambda \mathbf{E} - \mathbf{A}) = 0,$$

wobei  $\mathbf{E}$  die Einheitsmatrix der Dimension  $n$  darstellt. Die Hauptminoren  $D_i$  sind die Determinanten der linken oberen Untermatrizen von  $\mathbf{A}$ ,

$$D_1 = \det\left(\begin{bmatrix} a_{11} \end{bmatrix}\right), \quad D_2 = \det\left(\begin{bmatrix} a_{11} & a_{12} \\ a_{12} & a_{22} \end{bmatrix}\right), \quad \dots, \quad D_n = \det\left(\begin{bmatrix} a_{11} & \dots & a_{1n} \\ \vdots & \ddots & \vdots \\ a_{1n} & \dots & a_{nn} \end{bmatrix}\right),$$

wobei  $a_{ij}$  die Elemente der  $i$ -ten Zeile und  $j$ -ten Spalte der Matrix  $\mathbf{A}$  bezeichnen. Um die Definitheit einer symmetrischen Matrix  $\mathbf{A}$  zu bestimmen, muss lediglich eine der drei Bedingungen (a)–(c) ausgewertet werden, da jede für sich notwendig und hinreichend ist. Das Kriterium (c) wird auch *Sylvester-Kriterium* genannt und kann nicht für semi-definite Matrizen verwendet werden.

Bei der Abschätzung von Funktionen werden häufig der Gradient und die Hessematrix im Rahmen des *Mittelwertsatzes (Satz von Taylor)* verwendet.

**Satz 1.3 (Mittelwertsatz, Satz von Taylor).** Es sei  $f(\mathbf{x})$  eine stetig differenzierbare Funktion, d. h.  $f \in C^1$ , in einer Menge  $\mathcal{X}$ , die das Liniensegment  $[\mathbf{x}_1, \mathbf{x}_2]$  beinhaltet, dann existiert eine reelle Zahl  $\alpha$ ,  $0 \leq \alpha \leq 1$  so, dass gilt

$$f(\mathbf{x}_2) = f(\mathbf{x}_1) + (\mathbf{x}_2 - \mathbf{x}_1)^T (\nabla f)(\alpha \mathbf{x}_1 + (1 - \alpha) \mathbf{x}_2). \quad (1.26)$$

Ist die Funktion  $f(\mathbf{x})$  zweifach stetig differenzierbar, d. h.  $f \in C^2$ , dann existiert eine reelle Zahl  $\alpha$ ,  $0 \leq \alpha \leq 1$  so, dass die Beziehung

$$f(\mathbf{x}_2) = f(\mathbf{x}_1) + (\mathbf{x}_2 - \mathbf{x}_1)^T (\nabla f)(\mathbf{x}_1) + \frac{1}{2} (\mathbf{x}_2 - \mathbf{x}_1)^T (\nabla^2 f)(\alpha \mathbf{x}_1 + (1 - \alpha) \mathbf{x}_2) (\mathbf{x}_2 - \mathbf{x}_1) \quad (1.27)$$

gilt.

### 1.3.4 Berechnung von Ableitungen

Zur konkreten Berechnung von Ableitungen können verschiedene Verfahren verwendet werden.

#### Analytisches Differenzieren

Ist  $f(\mathbf{x})$  als analytischer Ausdruck gegeben, so können Ableitungen direkt mittels analytischer Differenzierung berechnet werden.

#### Algorithmisches Differenzieren

In der Optimierung steht die Kostenfunktion  $f(\mathbf{x})$  häufig nicht als geschlossener analytischer Ausdruck zur Verfügung sondern in Form von Funktionen (Algorithmen), die in einem Computerprogramm realisiert sind. In diesem Fall bietet das *Algorithmische Differenzieren*, gelegentlich auch *Automatisches Differenzieren* genannt, eine komfortable Möglichkeit Ableitungen zu berechnen. Das Verfahren nutzt die Regeln der analytischen Differentiation (Differentiationsregeln für elementare Funktionen, Kettenregel, Produktregel, Summenregel, Quotientenregel, Ableitungsregel für inverse Funktionen, etc.) um ein neues Computerprogramm zu erstellen, das die gewünschten Ableitungen von  $f(\mathbf{x})$  berechnet. Die Programmschritte werden dabei so organisiert, dass eine effiziente und zugleich möglichst genaue Berechnung der Ableitungen erreicht wird. Mehr Informationen über das Algorithmische Differenzieren finden sich z. B. unter <http://www.autodiff.org> oder in [1.7].

#### Ableitungsberechnung mit Differenzenquotienten

Eine näherungsweise Ableitungsberechnung ist auch numerisch durch Bildung von Differenzenquotienten möglich [1.8]. Für eine allgemeine, hinreichend oft differenzierbare Funktion  $f(\mathbf{x})$  sind in Tabelle 1.1 Beispiele für Differenzenquotienten gegeben. Hierbei ist  $h$  die Schrittweite und  $\mathbf{e}_i$  der Einheitsvektor mit dem Eintrag 1 an der Stelle  $i$ . Die Tabelle enthält auch die Ordnungen der Fehler, welche bei dieser näherungsweise Ableitungsberechnung entstehen können. Das sind *Abschneidefehler* und *Rundungsfehler*, wobei  $e_r$  der maximale relative Fehler zufolge von Rundungsoperationen bei Gleitkommaarithmetik ist.

Die Herleitung von Differenzenquotienten und die Berechnung der zugehörigen Abschneidefehler können mittels Taylorreihenentwicklung am Punkt  $\mathbf{x}$  bzw. dem Mittelwertsatz 1.3 erfolgen [1.9]. Der Abschneidefehler ergibt sich, weil die Taylorreihenentwicklung bei einer bestimmten Ordnung abgebrochen wird und eine finite Schrittweite  $h > 0$  verwendet werden muss. Der Abschneidefehler ist also der Methode geschuldet und entsteht selbst bei exakter Berechnung des Funktionswertes  $f(\mathbf{x})$ .

Der Rundungsfehler ist auf die praktisch nicht exakte numerische Berechnung von  $f(\mathbf{x})$ ,  $f(\mathbf{x} \pm h\mathbf{e}_i)$  und  $f(\mathbf{x} \pm h\mathbf{e}_i \pm h\mathbf{e}_j)$  bei Verwendung von Gleitkommaarithmetik zurückzuführen. Wie auch nachfolgendes Beispiel zeigt, kann der Rundungsfehler für  $h \rightarrow 0$  unbeschränkt anwachsen.

Ableitung, Richtung	Formel	Abschneide- fehler	Rundungs- fehler
1. Ableitung, vorwärts	$(\nabla f)(\mathbf{x}) \approx \left[ \frac{f(\mathbf{x} + h\mathbf{e}_i) - f(\mathbf{x})}{h} \right]_{i=1,\dots,n}$	$\mathcal{O}(h)$	$\mathcal{O}(e_r h^{-1})$
1. Ableitung, rückwärts	$(\nabla f)(\mathbf{x}) \approx \left[ \frac{f(\mathbf{x}) - f(\mathbf{x} - h\mathbf{e}_i)}{h} \right]_{i=1,\dots,n}$	$\mathcal{O}(h)$	$\mathcal{O}(e_r h^{-1})$
1. Ableitung, zentral	$(\nabla f)(\mathbf{x}) \approx \left[ \frac{f(\mathbf{x} + h\mathbf{e}_i) - f(\mathbf{x} - h\mathbf{e}_i)}{2h} \right]_{i=1,\dots,n}$	$\mathcal{O}(h^2)$	$\mathcal{O}(e_r h^{-1})$
2. Ableitung, vorwärts	$(\nabla^2 f)(\mathbf{x}) \approx [f(\mathbf{x} + h\mathbf{e}_i + h\mathbf{e}_j) - f(\mathbf{x} + h\mathbf{e}_i) - f(\mathbf{x} + h\mathbf{e}_j) + f(\mathbf{x})]_{i,j=1,\dots,n}/h^2$	$\mathcal{O}(h)$	$\mathcal{O}(e_r h^{-2})$
2. Ableitung, zentral	$(\nabla^2 f)(\mathbf{x}) \approx [f(\mathbf{x} + h\mathbf{e}_i + h\mathbf{e}_j) - f(\mathbf{x} + h\mathbf{e}_i - h\mathbf{e}_j) - f(\mathbf{x} - h\mathbf{e}_i + h\mathbf{e}_j) + f(\mathbf{x} - h\mathbf{e}_i - h\mathbf{e}_j)]_{i,j=1,\dots,n}/(4h^2)$	$\mathcal{O}(h^2)$	$\mathcal{O}(e_r h^{-2})$

Tabelle 1.1: Differenzenquotienten.

*Beispiel 1.6.* Beispielhaft soll nun die Herleitung und Fehlerberechnung für den Vorwärtsdifferenzenquotient zur Approximation des Gradienten (erste Zeile in Tabelle 1.1) durchgeführt werden. Eine Taylorreihenentwicklung von  $f(\mathbf{x})$  liefert unter Berücksichtigung des Mittelwertsatzes 1.3

$$f(\mathbf{x} + h\mathbf{e}_i) = f(\mathbf{x}) + h \frac{\partial f}{\partial x_i} \Big|_{\mathbf{x}} + \frac{h^2}{2} \frac{\partial^2 f}{\partial x_i^2} \Big|_{\mathbf{x} + \alpha h\mathbf{e}_i} \quad (1.28)$$

mit  $\alpha \in (0, 1)$ . Daraus folgt

$$\frac{\partial f}{\partial x_i} \Big|_{\mathbf{x}} = \frac{f(\mathbf{x} + h\mathbf{e}_i) - f(\mathbf{x})}{h} - \underbrace{\frac{h}{2} \frac{\partial^2 f}{\partial x_i^2} \Big|_{\mathbf{x} + \alpha h\mathbf{e}_i}}_{\text{Abschneidefehler, } \mathcal{O}(h)} \quad (1.29)$$

Geht man nun davon aus, dass  $e_r$  der maximale relative Fehler durch die in der Gleitkommaarithmetik notwendigen Rundungsoperationen bei der Auswertung der Funktionen  $f(\mathbf{x} + h\mathbf{e}_i)$  und  $f(\mathbf{x})$  ist, so ergibt sich im schlechtesten Fall der berechnete

Wert

$$\begin{aligned} & \frac{f(\mathbf{x} + h\mathbf{e}_i)(1 + e_r) - f(\mathbf{x})(1 - e_r)}{h} \\ &= \frac{f(\mathbf{x} + h\mathbf{e}_i) - f(\mathbf{x})}{h} + \underbrace{\frac{(f(\mathbf{x} + h\mathbf{e}_i) + f(\mathbf{x}))e_r}{h}}_{\text{Rundungsfehler, } O(e_r h^{-1})} \end{aligned} \quad (1.30)$$

als Approximation von  $\partial f / \partial x_i |_{\mathbf{x}}$ . Der Gesamtfehler folgt daher in der Form

$$\begin{aligned} & \underbrace{\frac{f(\mathbf{x} + h\mathbf{e}_i) - f(\mathbf{x})}{h}}_{\text{Berechneter Wert}} + \underbrace{\frac{(f(\mathbf{x} + h\mathbf{e}_i) + f(\mathbf{x}))e_r}{h}}_{\text{Rundungsfehler}} - \underbrace{\frac{\partial f}{\partial x_i} \Big|_{\mathbf{x}}}_{\text{Exakter Wert}} \\ &= \frac{h}{2} \frac{\partial^2 f}{\partial x_i^2} \Big|_{\mathbf{x} + \alpha h \mathbf{e}_i} + \frac{(f(\mathbf{x} + h\mathbf{e}_i) + f(\mathbf{x}))e_r}{h}. \end{aligned} \quad (1.31)$$

Abbildung 1.12 zeigt wie in diesem Fall Abschneide-, Rundungs- und Gesamtfehler von  $h$  abhängen. Es existiert also eine optimale Schrittweite  $h$ , um den Gesamtfehler zu minimieren.

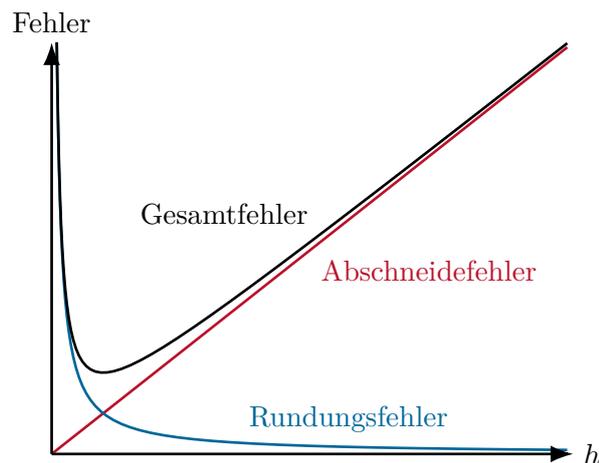


Abbildung 1.12: Fehler bei der Approximation des Gradienten durch den Vorwärtsdifferenzenquotienten in Abhängigkeit der Schrittweite  $h$ .

In Tabelle 1.1 und Beispiel 1.6 wurde nur der Rundungsfehler durch Gleitkommaarithmetik bei der Auswertung der Funktionen  $f$  berücksichtigt. Tatsächlich treten solche Rundungsfehler aber auch bei der Berechnung von Differenzenquotienten selbst auf, also z. B. bei der Auswertung von

$$\frac{f(\mathbf{x} + h\mathbf{e}_i) - f(\mathbf{x})}{h}. \quad (1.32)$$

Hierbei kann es durch die Differenzbildung im Zähler dieses Ausdrucks zu erheblichen Genauigkeitsverlusten durch *Auslöschungsfelder* kommen [1.10]. Beim Rechnen mit Gleitkommaarithmetik müssen alle Eingangsgrößen, Zwischenergebnisse der auftretenden Ele-

mentaroperationen und Endergebnisse auf die Zahl der Mantissenstellen gerundet werden. Als *Auslöschung* bezeichnet man das Annullieren führender Mantissenstellen bei der Subtraktion zweier (ähnlicher) Zahlen [1.11]. Geht man davon aus, dass mit  $M$  Mantissenstellen gerechnet wird und zwei zu subtrahierende Zahlen sich in  $m$  führenden Mantissenstellen gleichen, so bleiben im Ergebnis nur noch  $M - m$  gültige von Null verschiedene Mantissenstellen übrig, was direkt die erzielbare Genauigkeit einschränkt. Je ähnlicher sich daher zwei zu subtrahierende Zahlen sind, desto kleiner wird  $M - m$ , was bei der Berechnung von Differenzenquotienten im Falle  $h \rightarrow 0$  zur Unbrauchbarkeit des Ergebnisses führt.

*Beispiel 1.7.* Man betrachte die Rechnung

$$0.123\,456 - 0.123\,455 = 0.000\,001 . \quad (1.33)$$

Ist in den beiden Eingangsgrößen auch nur die letzte Nachkommastelle zufolge von früheren Rundungsfehlern unsicher, so kann das Ergebnis 0.000 001 völlig falsch sein, d. h. keine einzige gültige von Null verschiedene Nachkommastelle besitzen.

Neben den Fehlerordnungen, Rundungsfehlern und Auslöschungsfehlern spielt auch der jeweilige Rechenaufwand bei der Wahl eines Differenzenquotienten eine Rolle. Bei der näherungsweise Berechnung von  $(\nabla f)(\mathbf{x})$  erfordern einseitige Differenzenquotienten  $n + 1$  Auswertungen der Funktion  $f$  und der zentrale Differenzenquotient  $2n$  solche Auswertungen. D. h. bei der näherungsweise Berechnung von  $(\nabla f)(\mathbf{x})$  steigt der Berechnungsaufwand mit der Ordnung  $O(n)$ . Bei der näherungsweise Berechnung von  $(\nabla^2 f)(\mathbf{x})$  steigt der Berechnungsaufwand mit der Ordnung  $O(n^2)$ .

### Ableitungsberechnung mittels komplexer Funktionsauswertung

Im Englischen ist die nachfolgend beschriebene Methode zur näherungsweise Ableitungsberechnung als *complex step derivative* bekannt [1.12, 1.13]. Diese numerische Methode ist nur auf *holomorphe* Funktionen anwendbar. Holomorphe Funktionen werden auch als komplex differenzierbar bezeichnet.

**Definition 1.5 (Holomorphe Funktion).** Eine Funktion  $f : \mathcal{U} \rightarrow \mathbb{C}$  mit  $\mathcal{U} \subseteq \mathbb{C}$  nennt man *holomorph*, falls der Grenzwert

$$\lim_{h \rightarrow 0} \frac{f(z+h) - f(z)}{h} \quad \forall z \in \mathcal{U} \quad (1.34)$$

mit  $h \in \mathbb{C}$  existiert.

Eine Funktion  $f : \mathcal{U} \rightarrow \mathbb{C}$  in mehreren Variablen, d. h.  $\mathcal{U} \subseteq \mathbb{C}^n$ , ist genau dann holomorph, wenn sie holomorph ist bezüglich jeder einzelnen Variable bei festgehaltenen übrigen Variablen.

Die erste Ableitung einer holomorphen reellen Funktion  $f(\mathbf{x})$ , d. h.  $f : \mathbb{R}^n \rightarrow \mathbb{R}$ , kann näherungsweise mit der Formel

$$(\nabla f)(\mathbf{x}) \approx \left[ \frac{\operatorname{Im}(f(\mathbf{x} + h\mathbf{e}_i))}{h} \right]_{i=1, \dots, n} \quad (1.35)$$

berechnet werden, wobei  $h > 0$  eine kleine reelle Schrittweite darstellt. Zur Herleitung von (1.35), wird  $f(\mathbf{x})$  zunächst in eine Taylorreihe

$$f(\mathbf{x} + h\mathbf{I}e_i) = f(\mathbf{x}) + h\mathbf{I} \left. \frac{\partial f}{\partial x_i} \right|_{\mathbf{x}} - \frac{h^2}{2} \left. \frac{\partial^2 f}{\partial x_i^2} \right|_{\mathbf{x}} - \frac{h^3}{6} \mathbf{I} \left. \frac{\partial^3 f}{\partial x_i^3} \right|_{\mathbf{x} + \alpha h \mathbf{I}e_i} \quad (1.36)$$

mit  $\alpha \in (0, 1)$  entwickelt. Dividiert man den Imaginärteil von (1.36) durch  $h$ , so folgen daraus

$$\left. \frac{\partial f}{\partial x_i} \right|_{\mathbf{x}} = \frac{\operatorname{Im}(f(\mathbf{x} + h\mathbf{I}e_i))}{h} + \underbrace{\frac{h^2}{6} \left. \frac{\partial^3 f}{\partial x_i^3} \right|_{\mathbf{x} + \alpha h \mathbf{I}e_i}}_{\text{Abschneidefehler, } O(h^2)} \quad (1.37)$$

und somit direkt die Komponenten von (1.35). In ähnlicher Weise werden in [1.14] Formeln zur näherungsweisen Berechnung von zweiten Ableitungen mittels komplexer Funktionsauswertung hergeleitet.

$(\nabla f)(\mathbf{x})$  kann also mit nur  $n$  komplexen Auswertungen der Funktion  $f$  näherungsweise berechnet werden, wobei ein Abschneidefehler der Ordnung  $O(h^2)$  erzielt wird. Der Rundungsfehler besitzt die Ordnung  $O(e_r h^{-1})$ . Ein zentraler Vorteil dieser Methode gegenüber der Ableitungsberechnung mit Differenzenquotienten ist, dass bei der Berechnung der ersten Ableitung kein Auslöschungsfehler auftritt (keine Subtraktion ähnlicher Zahlen nötig), weshalb  $h$  sehr klein gewählt werden kann. Dies gilt im Allgemeinen nicht mehr für höhere Ableitungen. Der Nachteil dieser Methode liegt im geringfügig höheren numerischen Aufwand, da mit komplexen Zahlen gerechnet werden muss. In [1.13] wird der Zusammenhang zwischen dieser Methode und dem Algorithmischen Differenzieren diskutiert.

Aus dem Realteil von (1.36) folgt noch

$$f(\mathbf{x}) = \operatorname{Re}(f(\mathbf{x} + h\mathbf{I}e_i)) + O(h^2). \quad (1.38)$$

Wird also eine Ableitung  $\partial f / \partial x_i$  berechnet, so erhält man ohne zusätzliche Funktionsauswertungen auch den Wert  $f(\mathbf{x})$ .

### 1.3.5 Konvexität

Die Eigenschaft der Konvexität ist von großer Bedeutung in der Optimierung und erlaubt häufig eine einfache (numerische) Lösung des Optimierungsproblems. Der Begriff *konvex* kann sowohl auf Mengen als auch auf Funktionen angewandt werden.

#### 1.3.5.1 Konvexe Mengen

**Definition 1.6 (Konvexe Menge).** Eine Menge  $\mathcal{X} \subseteq \mathbb{R}^n$  nennt man *konvex*, falls für alle  $\mathbf{x}, \mathbf{y} \in \mathcal{X}$  und alle reellen Zahlen  $\alpha$  mit  $0 < \alpha < 1$  gilt

$$(\alpha \mathbf{x} + (1 - \alpha) \mathbf{y}) \in \mathcal{X}. \quad (1.39)$$

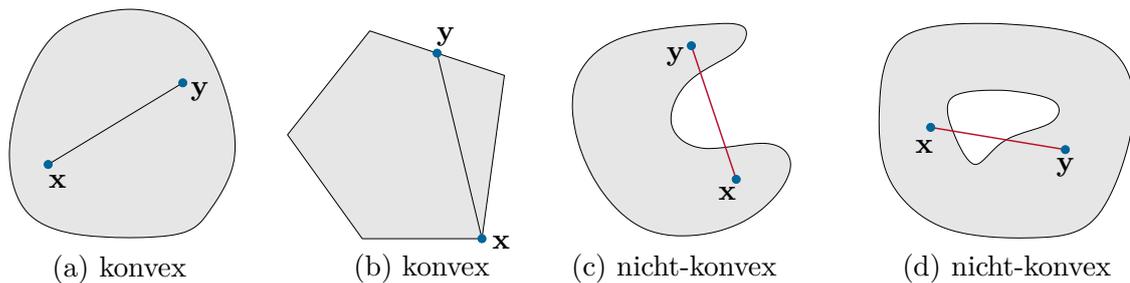


Abbildung 1.13: Beispiele von konvexen und nicht-konvexen Mengen.

Eine geometrische Interpretation dieser Definition ist, dass eine Menge  $\mathcal{X} \subseteq \mathbb{R}^n$  genau dann konvex ist, falls die Verbindungslineal zwischen zwei beliebigen Punkten  $\mathbf{x}, \mathbf{y} \in \mathcal{X}$  komplett in  $\mathcal{X}$  enthalten ist. Abbildung 1.13 zeigt einige Beispiele konvexer und nicht-konvexer Mengen.

Konvexe Mengen besitzen folgende Eigenschaften:

- (a) Die *Schnittmenge* von konvexen Mengen ist wiederum konvex.
- (b) Wenn  $\mathcal{X}$  eine konvexe Menge ist und  $\alpha$  eine feste reelle Zahl, dann ist die Menge

$$\{\alpha \mathbf{x} \mid \mathbf{x} \in \mathcal{X}\}$$

ebenfalls konvex.

- (c) Das Bild einer konvexen Menge unter einer linearen Transformation ist konvex.
- (d) Wenn  $\mathcal{X}$  und  $\mathcal{Y}$  konvexe Mengen sind, dann ist die Menge

$$\{\mathbf{x} + \mathbf{y} \mid \mathbf{x} \in \mathcal{X}, \mathbf{y} \in \mathcal{Y}\}$$

ebenfalls konvex.

Diese Eigenschaften sind u. a. bei der Charakterisierung der Konvexität der zulässigen Menge  $\mathcal{X}$  eines Optimierungsproblems von Bedeutung.

### 1.3.5.2 Konvexe Funktionen

**Definition 1.7 (Konvexe und konkave Funktionen).** Es sei  $\mathcal{X} \subseteq \mathbb{R}^n$  eine konvexe Menge. Man nennt die Funktion  $f : \mathcal{X} \rightarrow \mathbb{R}$  *konvex* auf  $\mathcal{X}$ , falls für alle  $\mathbf{x}, \mathbf{y} \in \mathcal{X}$  und alle reellen Zahlen  $\alpha$  mit  $0 \leq \alpha \leq 1$  gilt

$$f(\alpha \mathbf{x} + (1 - \alpha) \mathbf{y}) \leq \alpha f(\mathbf{x}) + (1 - \alpha) f(\mathbf{y}). \quad (1.40)$$

Die Funktion  $f$  nennt man *strikt konvex*, falls für alle  $\alpha$  mit  $0 < \alpha < 1$  und  $\mathbf{x} \neq \mathbf{y}$  gilt

$$f(\alpha \mathbf{x} + (1 - \alpha) \mathbf{y}) < \alpha f(\mathbf{x}) + (1 - \alpha) f(\mathbf{y}). \quad (1.41)$$

Man nennt die Funktion  $f$  (*strikt*) *konkav*, falls  $-f$  (strikt) konvex ist.

Die Definition 1.7 kann wie folgt geometrisch interpretiert werden: Eine Funktion  $f$  ist konvex (konkav), falls für alle  $\mathbf{x} \in \mathcal{X}$ ,  $\mathbf{y} \in \mathcal{X}$  und  $0 < \alpha < 1$  alle Funktionswerte  $f(\alpha \mathbf{x} + (1 - \alpha) \mathbf{y})$  unterhalb (oberhalb) oder auf der Verbindungslinie zwischen  $f(\mathbf{x})$  und  $f(\mathbf{y})$  liegen. Abbildung 1.14 zeigt einige Beispiele konvexer und konkaver Funktionen. Es ist direkt ersichtlich, dass affine Funktionen sowohl konkav als auch konvex sind.

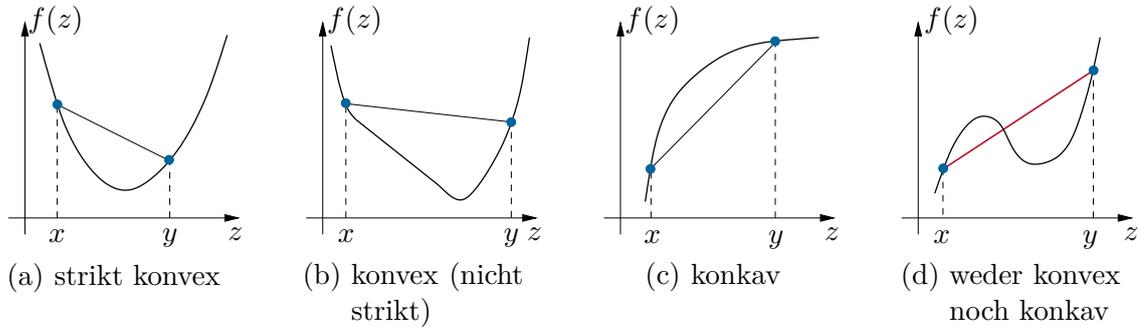


Abbildung 1.14: Beispiele von konvexen und konkaven Funktionen.

Konvexe Funktionen besitzen folgende Eigenschaften:

- (a) Die Summenfunktion

$$f(\mathbf{x}) = \sum_{i=1}^k a_i f_i(\mathbf{x}) \quad (1.42)$$

von auf der konvexen Menge  $\mathcal{X}$  konvexen Funktionen  $f_i(\mathbf{x})$ ,  $i = 1, \dots, k$  mit den reellen Koeffizienten  $a_i \geq 0$ ,  $i = 1, \dots, k$  ist auf  $\mathcal{X}$  ebenfalls konvex.

- (b) Ist die Funktion  $f(\mathbf{x})$  auf der konvexen Menge  $\mathcal{X}$  konvex, so ist die Menge

$$\{\mathbf{x} \in \mathcal{X} \mid f(\mathbf{x}) \leq c\} \quad (1.43)$$

für alle reellen Zahlen  $c \in \mathbb{R}$  ebenfalls konvex, siehe Abbildung 1.15.

- (c) Eine stetig differenzierbare Funktion  $f \in C^1$  ist genau dann konvex auf der konvexen Menge  $\mathcal{X}$ , wenn für alle  $\mathbf{x}, \mathbf{y} \in \mathcal{X}$  die Ungleichung

$$f(\mathbf{y}) \geq f(\mathbf{x}) + (\mathbf{y} - \mathbf{x})^T (\nabla f)(\mathbf{x}) \quad (1.44)$$

erfüllt ist. Die geometrische Interpretation der Ungleichung (1.44) ist, dass an jedem Punkt  $\mathbf{x}$  einer konvexen Funktion  $f(\mathbf{x})$  eine sogenannte *stützende Hyperebene* (skalärer Fall: *stützende Tangente*) existieren muss, oberhalb oder auf der  $f(\mathbf{x})$  verläuft. Dies ist in Abbildung 1.16 veranschaulicht.

- (d) Eine zweifach stetig differenzierbare Funktion  $f \in C^2$  ist genau dann konvex auf der konvexen Menge  $\mathcal{X}$ , wenn die Hessematrix  $(\nabla^2 f)(\mathbf{x})$  positiv semi-definit für alle  $\mathbf{x} \in \mathcal{X}$  ist. Falls die Hessematrix  $(\nabla^2 f)(\mathbf{x})$  positiv definit ist, so folgt daraus die strikte Konvexität der Funktion  $f(\mathbf{x})$ . Die Umkehrung dieser Aussage ist jedoch nicht gültig, wie man sich anhand der Funktion  $f(x) = x^4$  überzeugen kann. Diese Funktion ist strikt konvex, aber die zugehörige Hessematrix an der Stelle  $x = 0$  ist identisch Null.

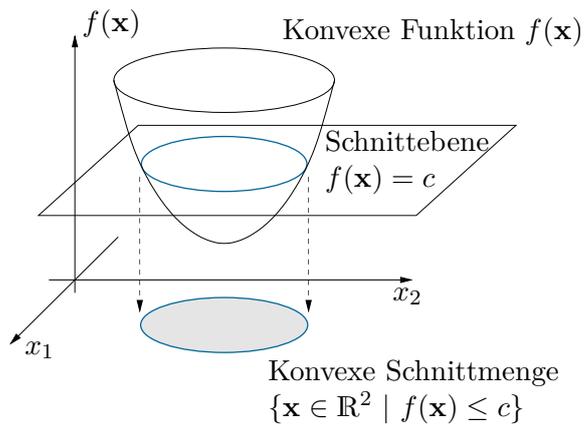


Abb. 1.15: Konvexe Menge, die durch den Schnitt einer konvexen Funktion  $f(\mathbf{x})$  mit der Ebene  $f(\mathbf{x}) = \text{konst.}$  entsteht.

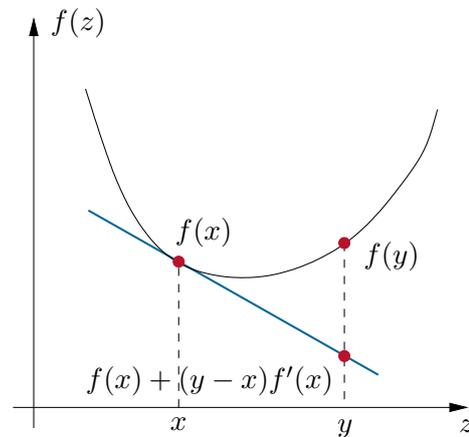


Abb. 1.16: Stützende Tangente einer konvexen Funktion  $f(z)$ .

**Aufgabe 1.1.** Beweisen Sie die Eigenschaften (a)–(d) von konvexen Funktionen. Nutzen Sie für den Beweis der Eigenschaft (d) den Mittelwertsatz, siehe Satz 1.3, im Speziellen (1.27).

**Aufgabe 1.2.** Zeigen Sie, dass die Funktion  $f(\mathbf{x}) = x_1^4 + x_1^2 - 2x_1x_2 + x_2^2$  über ihrem gesamten Definitionsbereich  $\mathbf{x} = [x_1 \ x_2]^T \in \mathbb{R}^2$  konvex ist.

## 1.4 Literatur

- [1.1] A. E. Bryson, Jr., *Dynamic Optimization*. Addison-Wesley, 1999.
- [1.2] R. H. Goddard, „A method of reaching extreme altitudes“, *Smithsonian Miscellaneous Collections*, Jg. 71, Nr. 2, 1919.
- [1.3] K. Pohmer, *Mikroökonomische Theorie der personellen Einkommens- und Vermögensverteilung*, Ser. Studies in Contemporary Economics. Springer, 1985, Bd. 16.
- [1.4] H. J. Oberle und R. Rosendahl, „Numerical computation of a singular-state subarc in an economic optimal control model“, *Optimal Control Applications and Methods*, Jg. 27, Nr. 4, S. 211–235, 2006.
- [1.5] M. Bazaraa, H. Sherali und C. Shetty, *Nonlinear Programming: Theory and Algorithms*, 3. Aufl. John Wiley & Sons, 2006.
- [1.6] I. Griva, S. Nash und A. Sofer, *Linear and Nonlinear Optimization*, 2. Aufl. Society for Industrial und Applied Mathematics, 2009.
- [1.7] A. Griewank und A. Walther, *Evaluating Derivatives: Principles and Techniques of Algorithmic Differentiation*, 2. Aufl., Ser. Other Titles in Applied Mathematics. Society for Industrial und Applied Mathematics, 2008.
- [1.8] D. Lynch, *Numerical Partial Differential Equations for Environmental Scientists and Engineers - A First Practical Course*. New York: Springer, 2005.
- [1.9] M. Hanke-Bourgeois, *Grundlagen der Numerischen Mathematik und des Wissenschaftlichen Rechnens*, 3. Aufl. Vieweg+Teubner, 2009.
- [1.10] P. E. Gill, W. Murray und M. H. Wright, *Practical Optimization*. Academic Press, 1981.
- [1.11] H.R. Schwarz und N. Köckler, *Numerische Mathematik*, 8. Aufl. Wiesbaden: Vieweg+Teubner, 2011.
- [1.12] J. Lyness und C. Moler, „Numerical differentiation of analytic functions“, *SIAM Journal on Numerical Analysis*, Jg. 4, Nr. 2, S. 202–210, 1967.
- [1.13] J. Martins, P. Sturdza und J. Alonso, „The complex-step derivative approximation“, *ACM Transactions on Mathematical Software*, Jg. 29, Nr. 3, S. 245–262, 2003.
- [1.14] R. Abreu, „Complex steps finite differences with applications to seismic problems“, Diss., University of Granada, Granada, Spanien, 2013.
- [1.15] J. Nocedal und S. J. Wright, *Numerical Optimization*, 2. Aufl., Ser. Springer Series in Operations Research and Financial Engineering. Springer, 2006.
- [1.16] S. Boyd und L. Vandenberghe, *Convex Optimization*. Cambridge University Press, 2004.
- [1.17] M. Papageorgiou, M. Leibold und M. Buss, *Optimierung: Statische, dynamische, stochastische Verfahren für die Anwendung*, 4. Aufl. Springer, 2015.
- [1.18] D. P. Bertsekas, *Nonlinear Programming*, 2. Aufl. Athena Scientific, 1999.

- 
- [1.19] D. G. Luenberger und Y. Ye, *Linear and Nonlinear Programming*, 3. Aufl., Ser. International Series in Operations Research & Management Science. Springer, 2008, Bd. 116.
- [1.20] B. C. Chachuat, „Nonlinear and Dynamic Optimization: From Theory to Practice“, abrufbar unter <http://infoscience.epfl.ch/record/111939>, Laboratoire d'Automatique, École Polytechnique Fédérale de Lausanne, 2007, (besucht am 28.09.2017).

## 2 Statische Optimierung: Unbeschränkter Fall

In diesem Kapitel werden unbeschränkte statische Optimierungsprobleme der Art

$$\min_{\mathbf{x} \in \mathbb{R}^n} f(\mathbf{x}) \quad (2.1)$$

betrachtet. Die Abschnitte 2.2 bis 2.6 behandeln numerische Verfahren zur Lösung solcher Optimierungsprobleme. Im nachfolgenden Abschnitt werden Optimalitätsbedingungen für das Problem (2.1) formuliert. Zur Definition der Begriffe lokaler und globaler Minima sei auf Abschnitt 1.3.1 und im Speziellen auf Definition 1.2 verwiesen.

### 2.1 Optimalitätsbedingungen

**Satz 2.1 (Notwendige Optimalitätsbedingung erster Ordnung).** *Es sei  $f \in C^1$  eine Funktion definiert auf  $\mathbb{R}^n$ . Wenn  $\mathbf{x}^*$  ein lokales Minimum von  $f$  auf  $\mathbb{R}^n$  ist, dann gilt*

$$(\nabla f)(\mathbf{x}^*) = \mathbf{0}. \quad (2.2)$$

*Beweis.* Man betrachte die Funktion  $g(\alpha) = f(\mathbf{x}^* + \alpha \mathbf{d})$  mit einer beliebigen Richtung  $\mathbf{d} \in \mathbb{R}^n$  und  $\alpha \geq 0$ . Für beliebige Werte  $\alpha \geq 0$  gilt  $\mathbf{x}^* + \alpha \mathbf{d} \in \mathbb{R}^n$ , so dass die Funktion  $g(\alpha)$  wohldefiniert ist. Diese Funktion muss am Punkt  $\alpha = 0$  ein lokales Minimum besitzen. Entwickelt man  $g(\alpha)$  um den Punkt  $\alpha = 0$  in eine Taylorreihe und bricht diese nach dem linearen Glied ab, so erhält man

$$g(\alpha) = g(0) + g'(0)\alpha + \mathcal{O}(\alpha^2) \quad (2.3)$$

mit  $g'(0) = \mathbf{d}^T(\nabla f)(\mathbf{x}^*)$ . Der Restterm  $\mathcal{O}(\alpha^2)$  klingt quadratisch nach Null ab, d. h. schneller als der lineare Term  $g'(0)\alpha$ . Wäre nun  $g'(0) < 0$ , dann würde für ein hinreichend kleines  $\alpha > 0$  gelten  $g(\alpha) - g(0) < 0$ , was im Widerspruch zu der Annahme steht, dass  $\alpha = 0$  bzw.  $\mathbf{x}^*$  ein Minimum ist. Daher muss gelten  $g'(0) = \mathbf{d}^T(\nabla f)(\mathbf{x}^*) \geq 0$ . Dies kann aber nur dann für beliebige  $\mathbf{d} \in \mathbb{R}^n$  erfüllt sein, wenn  $(\nabla f)(\mathbf{x}^*) = \mathbf{0}$  gilt.  $\square$

*Beispiel 2.1.* Man betrachte das Optimierungsproblem

$$\min_{\mathbf{x} \in \mathbb{R}^2} f(x_1, x_2) = x_1^2 - x_1x_2 + x_2^2 - 3x_2. \quad (2.4)$$

Berechnet man nun die notwendige Optimalitätsbedingung erster Ordnung gemäß (2.2)

$$2x_1 - x_2 = 0 \quad (2.5a)$$

$$-x_1 + 2x_2 = 3, \quad (2.5b)$$

dann erkennt man, dass  $\mathbf{x}^* = [1 \ 2]^T$  eine eindeutige Lösung von (2.5) ist. Man kann zeigen, dass  $\mathbf{x}^*$  in diesem Fall sogar ein globales Minimum ist.

Die Optimalitätsbedingung gemäß Satz 2.1 ist notwendig aber nicht hinreichend. Die Bedingung gibt lediglich an, dass es sich bei dem betreffenden Punkt um einen *Extremalpunkt* (auch als *stationärer Punkt* bezeichnet) handelt, und wird von einem Minimum, Maximum oder Sattelpunkt gleichermaßen erfüllt, siehe Abbildung 2.1.

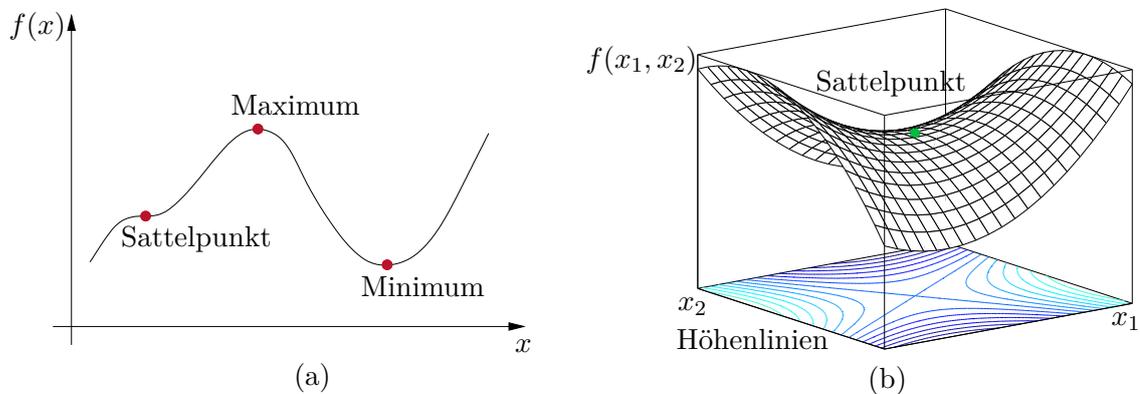


Abbildung 2.1: Beispiele von stationären Punkten im ein- und zweidimensionalen Fall.

**Satz 2.2 (Notwendige Optimalitätsbedingungen zweiter Ordnung).** *Es sei  $f \in C^2$  eine Funktion definiert auf  $\mathbb{R}^n$ . Wenn  $\mathbf{x}^*$  ein lokales Minimum von  $f$  auf  $\mathbb{R}^n$  ist, dann gelten die Bedingungen*

$$(a) \quad (\nabla f)(\mathbf{x}^*) = \mathbf{0} \quad (2.6a)$$

$$(b) \quad (\nabla^2 f)(\mathbf{x}^*) \text{ ist positiv semi-definit.} \quad (2.6b)$$

**Aufgabe 2.1.** Beweisen Sie Satz 2.2. **Hinweis:** Orientieren Sie sich dabei am Beweis von Satz 2.1.

Auch Satz 2.2 beschreibt lediglich notwendige Optimalitätsbedingungen, wie man sich einfach anhand der Funktion  $f(x) = x^3$  überzeugen kann. Diese Funktion besitzt an der Stelle  $x^* = 0$  einen Extrempunkt ( $f'(x^*) = 3(x^*)^2 = 0$ ) und obwohl die zweite Ableitung  $f''(x^*) = 6x^* = 0$  positiv semi-definit ist, ist  $x^* = 0$  kein Minimum. Die Funktion hat an der Stelle  $x^* = 0$  einen Sattelpunkt.

Es können nun folgende hinreichende Optimalitätsbedingungen angegeben werden.

**Satz 2.3** (Hinreichende Optimalitätsbedingungen zweiter Ordnung). *Es sei  $f \in C^2$  eine Funktion definiert auf  $\mathbb{R}^n$ . Wenn die Bedingungen*

$$(a) \quad (\nabla f)(\mathbf{x}^*) = \mathbf{0} \quad (2.7a)$$

$$(b) \quad (\nabla^2 f)(\mathbf{x}^*) \text{ ist positiv definit} \quad (2.7b)$$

*erfüllt sind, dann ist  $\mathbf{x}^*$  ein striktes lokales Minimum von  $f$  auf  $\mathbb{R}^n$ .*

**Aufgabe 2.2.** Beweisen Sie Satz 2.3.

**Beispiel 2.2.** Für das Optimierungsproblem

$$\min_{\mathbf{x} \in \mathbb{R}^2} f(\mathbf{x}) = x_1^2 + ax_2^2 - x_1x_2 \quad (2.8)$$

sollen die stationären Werte  $\mathbf{x}^*$  in Abhängigkeit des Parameters  $a \neq \frac{1}{4}$  charakterisiert werden. Der Gradient und die Hessematrix von  $f(\mathbf{x})$  ergeben sich zu

$$(\nabla f)(\mathbf{x}) = \begin{bmatrix} 2x_1 - x_2 \\ 2ax_2 - x_1 \end{bmatrix}, \quad (\nabla^2 f)(\mathbf{x}) = \begin{bmatrix} 2 & -1 \\ -1 & 2a \end{bmatrix}. \quad (2.9)$$

Aus  $(\nabla f)(\mathbf{x}^*) = \mathbf{0}$  folgt  $\mathbf{x}^* = [0 \ 0]^T$  als einziger stationärer Punkt. Die Definitheit der Hessematrix  $(\nabla^2 f)(\mathbf{x})$  an der Stelle  $\mathbf{x}^*$  lässt sich mit Hilfe der Hauptminoren (Sylvesterkriterium, siehe (c) in Satz 1.2) untersuchen

$$D_1 = 2, \quad D_2 = 4a - 1. \quad (2.10)$$

Somit ist  $(\nabla^2 f)(\mathbf{x}^*)$  positiv definit für  $a > \frac{1}{4}$  und  $\mathbf{x}^* = [0 \ 0]^T$  stellt ein striktes Minimum dar. Für  $a < \frac{1}{4}$  ist  $D_1 > 0$  und  $D_2 < 0$  und  $(\nabla^2 f)(\mathbf{x})$  somit *indefinit*. In diesem Fall ist  $\mathbf{x}^* = [0 \ 0]^T$  ein *Sattelpunkt*, ähnlich wie er in Abbildung 2.1(b) für  $a = -1$  dargestellt ist.

Wenn die Funktion  $f(\mathbf{x})$  konvex ist, dann ist die notwendige Optimalitätsbedingung erster Ordnung gemäß Satz 2.1 auch *hinreichend*. Um dies zu sehen, beachte man, dass mit beliebigem  $\mathbf{y} \in \mathbb{R}^n$  wegen der Konvexität von  $f(\mathbf{x})$  mit dem Minimum  $\mathbf{x}^*$  die Ungleichung

$$f(\mathbf{y}) \geq f(\mathbf{x}^*) + \underbrace{(\mathbf{y} - \mathbf{x}^*)^T (\nabla f)(\mathbf{x}^*)}_{=0} = f(\mathbf{x}^*) \quad (2.11)$$

gilt. Die Sätze 2.1 bis 2.3 liefern nur Aussagen zu lokalen Minima. Wenn die Funktion  $f(\mathbf{x})$  konvex oder strikt konvex ist, dann können nachfolgende Bedingungen für globale Minima angegeben werden.

**Satz 2.4** (Globale Minima einer konvexen Funktion). *Es sei  $f(\mathbf{x})$  eine konvexe Funktion auf  $\mathbb{R}^n$ . Die Menge aller Minima  $\mathcal{G} = \arg \min\{f(\mathbf{x}) \mid \mathbf{x} \in \mathbb{R}^n\}$  ist konvex. Jedes lokale Minimum  $\mathbf{x}^* \in \mathcal{G}$  von  $f$  ist auch ein globales Minimum. Ist  $f(\mathbf{x})$  strikt*

*konvex, so ist  $\mathbf{x}^*$  ein striktes globales Minimum.*

*Beweis.* Angenommen  $c$  beschreibt den minimalen Wert von  $f$ . Dann ist die Menge  $\mathcal{G} = \{\mathbf{x} \mid \mathbf{x} \in \mathbb{R}^n, f(\mathbf{x}) \leq c\}$  gemäß (1.43) konvex, womit der erste Teil des Satzes gezeigt ist.

Der zweite Teil des Satzes kann mittels Beweis durch Widerspruch gezeigt werden. Angenommen  $\mathbf{x}^*$  ist ein lokales Minimum aber kein globales Minimum von  $f$  auf  $\mathbb{R}^n$ . Dann existiert ein Punkt  $\mathbf{y} \in \mathbb{R}^n$ , der die Ungleichung  $f(\mathbf{y}) < f(\mathbf{x}^*)$  erfüllt. Gemäß der Definition 1.7 für konvexe Funktionen gilt dann für alle  $\alpha \in [0, 1]$

$$f(\alpha\mathbf{x}^* + (1-\alpha)\mathbf{y}) \leq \alpha f(\mathbf{x}^*) + (1-\alpha)f(\mathbf{y}) < \alpha f(\mathbf{x}^*) + (1-\alpha)f(\mathbf{x}^*) = f(\mathbf{x}^*). \quad (2.12)$$

Dies zeigt, dass mit  $\alpha \rightarrow 1$  Punkte  $\alpha\mathbf{x}^* + (1-\alpha)\mathbf{y}$  gefunden werden können, die beliebig nahe bei  $\mathbf{x}^*$  liegen, deren Funktionswerte  $f(\alpha\mathbf{x}^* + (1-\alpha)\mathbf{y})$  aber strikt kleiner sind als  $f(\mathbf{x}^*)$ . Dies verletzt die Definition eines lokalen Minimums an der Stelle  $\mathbf{x}^*$  und steht daher im Widerspruch zu Annahme. Folglich kann kein Punkt  $\mathbf{y} \in \mathbb{R}^n$  existieren, der die Ungleichung  $f(\mathbf{y}) < f(\mathbf{x}^*)$  erfüllt und  $\mathbf{x}^*$  ist ein globales Minimum.

In ähnlicher Weise kann der dritte Teil des Satzes gezeigt werden. Die Annahme, dass  $\mathbf{x}^*$  kein striktes globales Minimum ist, also ein Punkt  $\mathbf{y} \in \mathbb{R}^n$  existiert, der die Gleichung  $f(\mathbf{y}) = f(\mathbf{x}^*)$  erfüllt, führt für eine strikt konvexe Funktion  $f(\mathbf{x})$  auf einen Widerspruch.  $\square$

## 2.2 Rechnergestützte Minimierungsverfahren: Grundlagen

Da die Bestimmung eines (lokal) optimalen Punktes  $\mathbf{x}^*$  von (2.1) durch analytische Lösung der Stationaritätsbedingung  $(\nabla f)(\mathbf{x}^*) = \mathbf{0}$  von (2.7a) ( $n$  nichtlineare Gleichungen in  $\mathbf{x}^*$ ) nur in seltenen Fällen möglich ist, ist man im Allgemeinen auf *numerische Verfahren* zur Suche von  $\mathbf{x}^*$  angewiesen. Viele der in dieser Vorlesung besprochenen Algorithmen finden den exakten Punkt  $\mathbf{x}^*$  nicht in einer endlichen Anzahl von Rechenschritten, sondern generieren eine Folge  $\{\mathbf{x}_k\}$ , entlang welcher die zu optimierende Funktion  $f(\mathbf{x})$  abnimmt, d. h.

$$f(\mathbf{x}_{k+1}) < f(\mathbf{x}_k), \quad k = 0, 1, 2, \dots, \quad (2.13)$$

und die zumindest für  $k \rightarrow \infty$  gegen  $\mathbf{x}^*$  konvergieren soll, d. h.

$$\lim_{k \rightarrow \infty} \mathbf{x}_k = \mathbf{x}^*. \quad (2.14)$$

In der englischsprachigen Literatur werden solche Algorithmen auch als *iterative descent algorithms* bezeichnet. Neben der anhand von (2.14) zu beantwortenden Frage, ob ein Algorithmus prinzipiell gegen die richtige Lösung  $\mathbf{x}^*$  konvergiert, interessiert, wie rasch er dies tut. Es ist also das (globale) Konvergenzverhalten des Algorithmus zu analysieren. Zumeist wird diese Analyse basierend auf einer *Fehlerfunktion*  $e : \mathbb{R}^n \rightarrow \mathbb{R}$ , welche  $e(\mathbf{x}) \geq 0$  für alle  $\mathbf{x} \in \mathbb{R}^n$  und  $e(\mathbf{x}^*) = 0$  erfüllt, durchgeführt. Als Fehlerfunktion kann z. B. der Euklidische Abstand

$$e(\mathbf{x}) = \|\mathbf{x} - \mathbf{x}^*\| \geq 0 \quad (2.15a)$$

oder die Kostendifferenz

$$e(\mathbf{x}) = f(\mathbf{x}) - f(\mathbf{x}^*) \geq 0 \quad (2.15b)$$

verwendet werden [2.1]. Das Konvergenzverhalten eines Algorithmus kann nun anhand der zu  $\{\mathbf{x}_k\}$  gehörenden Folge  $\{e_k\}$  mit  $e_k = e(\mathbf{x}_k)$  analysiert werden. Zunächst sollen dazu die Begriffe *Konvergenzordnung* und *Konvergenzrate* einer Folge von Skalaren definiert werden.

**Definition 2.1 (Konvergenzordnung, Konvergenzrate).** Es sei  $\{e_k\}$  eine Folge von Skalaren, die gegen den Grenzwert 0 konvergiert. Die *Konvergenzordnung* der Folge  $\{e_k\}$  ist das Supremum der nichtnegativen Zahlen  $p$ , für die gilt

$$0 \leq \lim_{k \rightarrow \infty} \frac{|e_{k+1}|}{|e_k|^p} = \mu < \infty. \quad (2.16)$$

Als zugehörige *Konvergenzrate* bezeichnet man die Zahl  $\mu$ . Es werden folgende Fälle unterschieden:

- Im Fall  $p = 1$  und  $\mu \in (0, 1)$  spricht man von *linearer* Konvergenz.
- Im Fall  $p > 1$  mit  $\mu > 0$  oder  $p = 1$  mit  $\mu = 0$  spricht man von *superlinearer* Konvergenz.
- Im Fall  $p = 2$  mit  $\mu > 0$  spricht man von *quadratischer* Konvergenz.
- Im Fall  $p = 3$  mit  $\mu > 0$  spricht man von *kubischer* Konvergenz.

Im Wesentlichen beschreiben die Konvergenzordnung und die Konvergenzrate das Verhalten einer Folge für  $k \rightarrow \infty$ . Größere Werte der Konvergenzordnung  $p$  bedeuten, dass die Folge schneller konvergiert, da die Folgeelemente  $e_k$  (zumindest für sehr große Werte von  $k$ ) mit der  $p$ -ten Potenz abnehmen. Analoges gilt für kleinere Werte der Konvergenzrate  $\mu$ .

**Beispiel 2.3.** Die Folge  $\{a^k\}$  mit  $0 < a < 1$  konvergiert mit der Konvergenzordnung  $p = 1$  und der Konvergenzrate  $\mu = a$  gegen Null. Zunächst gilt, dass nur für  $p \leq 1$  die Bedingung

$$\lim_{k \rightarrow \infty} \frac{a^{k+1}}{a^{kp}} = \lim_{k \rightarrow \infty} a^{1+k(1-p)} < \infty \quad (2.17)$$

erfüllt ist. Mit  $p = 1$  folgt dann

$$\lim_{k \rightarrow \infty} \frac{a^{k+1}}{a^k} = a = \mu \quad (2.18)$$

für die Konvergenzrate  $\mu$ .

**Aufgabe 2.3.** Zeigen Sie, dass die Folge  $\{a^{2^k}\}$  mit  $0 < a < 1$  mit der Konvergenzordnung 2 und der Konvergenzrate 1 gegen 0 konvergiert.

**Beispiel 2.4.** Die Folge  $\{\frac{1}{k^k}\}$  hat eine lineare Konvergenzordnung, da nur für  $p \leq 1$  die Bedingung

$$\lim_{k \rightarrow \infty} \frac{k^{kp}}{(k+1)^{k+1}} < \infty \quad (2.19)$$

erfüllt ist. Mit  $p = 1$  ergibt sich dann

$$\lim_{k \rightarrow \infty} \frac{k^k}{(k+1)^{k+1}} = \lim_{k \rightarrow \infty} \frac{1}{k+1} \left( \frac{k}{k+1} \right)^k = \mu = 0. \quad (2.20)$$

Folglich konvergiert die Folge  $\{\frac{1}{k^k}\}$  superlinear gegen Null.

Abschließend stellt sich die Frage, ob das beobachtete Konvergenzverhalten eines Optimierungsalgorithmus von der gewählten Fehlerfunktion  $e(\mathbf{x})$  abhängt. Es lässt sich zeigen (vgl. [2.2]), dass die Konvergenzordnung eines Optimierungsalgorithmus von der Wahl der Fehlerfunktion  $e(\mathbf{x})$  weitgehend unabhängig ist. Dies gilt nicht für die Konvergenzrate.

Die bekanntesten numerischen Verfahren zur Lösung der unbeschränkten statischen Optimierungsaufgabe (2.1) sind die so genannten *Liniensuchverfahren* (englisch: *line search methods*). Der folgende Abschnitt gibt einen kurzen Überblick über die bekanntesten Liniensuchverfahren. Im Anschluss daran werden mit der *Methode der Vertrauensbereiche* und dem *direkten Suchverfahren* zwei alternative Lösungsmethoden für unbeschränkte statische Optimierungsaufgaben vorgestellt.

## 2.3 Liniensuchverfahren

---

```

Initialisierung:   $\mathbf{x}_0$       (Startlösung)
                    $k = 0$     (Startindex)

while   $\mathbf{x}_k$  ist nicht optimal
           Wähle geeignete Suchrichtung  $\mathbf{s}_k$ 
           Wähle optimale Schrittweite gemäß
            $\alpha_k = \arg \min_{\alpha > 0} f(\mathbf{x}_k + \alpha \mathbf{s}_k)$ 
            $\mathbf{x}_{k+1} \leftarrow \mathbf{x}_k + \alpha_k \mathbf{s}_k$ 
            $k \leftarrow k + 1$ 

end

```

---

Tabelle 2.1: Genereller Ablauf eines Liniensuchverfahrens.

Tabelle 2.1 zeigt die grundsätzliche algorithmische Struktur eines Liniensuchverfahrens. Zum Iterationsschritt  $k$  ermittelt man vorerst eine geeignete *Suchrichtung* bzw. *Abstiegsrichtung*  $\mathbf{s}_k$ . Sie soll so gewählt werden, dass, wenn man sich hinreichend wenig vom Punkt  $\mathbf{x}_k$  aus in diese Richtung bewegt, also

$$\mathbf{x}_{k+1} = \mathbf{x}_k + \alpha_k \mathbf{s}_k \quad (2.21)$$

mit einer geeigneten *Schrittweite*  $\alpha_k > 0$ , die Abstiegsbedingung (2.13), d. h.

$$f(\mathbf{x}_{k+1}) = f(\mathbf{x}_k + \alpha_k \mathbf{s}_k) < f(\mathbf{x}_k) \quad (2.22)$$

erfüllt ist. Nun wird die optimale Schrittweite  $\alpha_k > 0$  durch Lösung des *skalaren Optimierungsproblems*

$$\min_{\alpha_k > 0} g(\alpha_k) = f(\mathbf{x}_k + \alpha_k \mathbf{s}_k) \quad (2.23)$$

bestimmt. Die Iteration wird solange wiederholt, bis ein Abbruchkriterium erfüllt ist, z. B. bis eine gewählte Fehlerfunktion betragsmäßig kleiner als ein vorgegebener Schwellwert ist.

Abbildung 2.2 veranschaulicht das Prinzip der Liniensuche anhand von einem Iterationsschritt für eine (nicht konvexe) Kostenfunktion  $f(\mathbf{x})$  mit  $\mathbf{x} \in \mathbb{R}^2$  bei einer gegebenen Suchrichtung  $\mathbf{s}_k$ . In diesem Zusammenhang wird auch der Name *Liniensuchverfahren* verständlich, da sich bei gegebener Suchrichtung  $\mathbf{s}_k$  die Optimierungsaufgabe, d. h. die Wahl der Schrittweite  $\alpha_k$ , auf das Auffinden eines Minimums entlang einer Linie reduziert.

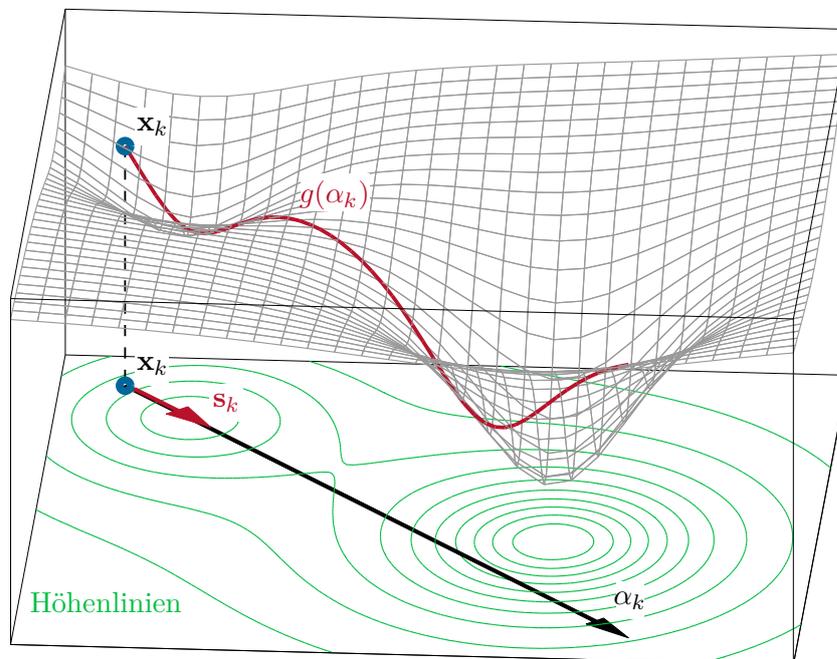


Abbildung 2.2: Veranschaulichung der Wahl der Schrittweite gemäß (2.23).

## 2.3.1 Wahl der Schrittweite

### 2.3.1.1 Intervallschachtelungsverfahren („Goldener Schnitt“)

Das *Intervallschachtelungsverfahren* generiert für das skalare Optimierungsproblem (2.23) eine konvergierende Folge von Intervallschachtelungen, um das Minimum von  $g(\alpha_k)$  einzugrenzen.

Zunächst muss ein Intervall  $[l_0, r_0]$  gefunden werden, in dem die Funktion  $g(\alpha_k)$  ein Minimum aufweist, siehe Abbildung 2.3. Dies kann z. B. dadurch erreicht werden, dass mit

einem hinreichend kleinen  $l_0$  gestartet und  $r_0$  (ausgehend von  $l_0$ ) sukzessive vergrößert wird, bis der Funktionswert  $g(r_0)$  anfängt zuzunehmen. Für das Folgende wird vorausgesetzt, dass die Funktion  $g(\alpha_k)$  stetig und *unimodal* im Intervall  $[l_0, r_0]$  ist, d. h. die Funktion  $g(\alpha_k)$  hat ein eindeutiges lokales Minimum im offenen Intervall  $(l_0, r_0)$ .

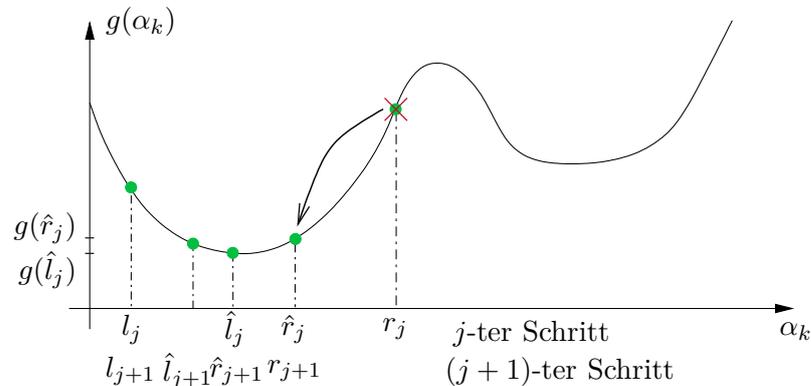


Abbildung 2.3: Veranschaulichung des Intervallschachtelungsverfahrens nach dem Prinzip des „Goldenen Schnittes“.

Zum Iterationsschritt  $j$  liege das Intervall  $[l_j, r_j]$  vor, das nach wie vor jenen Wert  $\alpha_k^*$  beinhaltet, der die Funktion  $g(\alpha_k)$  minimiert. Nun werden zwei neue Punkte  $\hat{l}_j$  und  $\hat{r}_j$ ,  $l_j < \hat{l}_j < \hat{r}_j < r_j$  in der Form

$$\hat{l}_j = l_j + (1 - a)(r_j - l_j) = al_j + (1 - a)r_j \quad (2.24a)$$

$$\hat{r}_j = l_j + a(r_j - l_j) = (1 - a)l_j + ar_j \quad (2.24b)$$

mit dem Parameter  $a \in (\frac{1}{2}, 1)$  berechnet. Es gilt nun folgender Satz:

**Satz 2.5 (Zur Intervallschachtelung).** *Es sei  $l_j < \hat{l}_j < \hat{r}_j < r_j$  und  $g(\alpha_k)$  eine stetige unimodale Funktion auf dem Intervall  $[l_j, r_j]$ . Bezeichnet man mit  $\alpha_k^*$  das Minimum auf  $(l_j, r_j)$ , dann gilt  $\alpha_k^* \in [l_j, \hat{r}_j]$ , wenn  $g(\hat{l}_j) \leq g(\hat{r}_j)$  bzw.  $\alpha_k^* \in [\hat{l}_j, r_j]$ , wenn  $g(\hat{l}_j) \geq g(\hat{r}_j)$ .*

*Beweis.* Man betrachte den Fall  $g(\hat{l}_j) \leq g(\hat{r}_j)$ . Angenommen,  $\alpha_k^* > \hat{r}_j$ , dann gilt  $\hat{l}_j < \alpha_k^*$ . Da  $g(\hat{l}_j) \leq g(\hat{r}_j)$  und  $g(\alpha_k^*) \leq g(\hat{r}_j)$  ist, muss ein Punkt  $\bar{\alpha}_k \in (\hat{l}_j, \alpha_k^*)$  so existieren, dass gilt  $g(\bar{\alpha}_k) = \max_{\alpha_k \in [\hat{l}_j, \alpha_k^*]} g(\alpha_k)$ , womit  $\bar{\alpha}_k$  ein lokales Maximum von  $g(\alpha_k)$  im Intervall  $[l_j, r_j]$  beschreibt. Die Existenz eines lokalen Maximums ist aber aufgrund der geforderten Unimodalität von  $g(\alpha_k)$  nicht möglich. Für  $g(\hat{l}_j) \geq g(\hat{r}_j)$  folgt der Beweis auf analoge Art und Weise.  $\square$

Gemäß Satz 2.5 wird zum nächsten Iterationsschritt  $j + 1$  für den Fall  $g(\hat{l}_j) \leq g(\hat{r}_j)$  der äußere Punkt  $r_j$  verworfen und das Intervall ergibt sich demnach zu  $[l_{j+1}, r_{j+1}]$  mit  $l_{j+1} = l_j$ ,  $r_{j+1} = \hat{r}_j$ , siehe Abbildung 2.3. Für  $g(\hat{l}_j) \geq g(\hat{r}_j)$  folgt das Intervall zum Iterationsschritt  $j + 1$  zu  $[l_{j+1}, r_{j+1}]$  mit  $l_{j+1} = \hat{l}_j$ ,  $r_{j+1} = r_j$ .

Für die weitere Betrachtung nehme man an, dass, wie in Abbildung 2.3 dargestellt,  $g(\hat{l}_j) \leq g(\hat{r}_j)$  ist. Die nachfolgenden Schritte lassen sich direkt auf den anderen Fall übertragen. Man führt zunächst eine weitere Iteration zur Berechnung der Zwischenpunkte gemäß (2.24) in der Form

$$\hat{l}_{j+1} = al_{j+1} + (1-a)r_{j+1} = (1-a+a^2)l_j + (1-a)ar_j \quad (2.25a)$$

$$\hat{r}_{j+1} = (1-a)l_{j+1} + ar_{j+1} = (1-a^2)l_j + a^2r_j \quad (2.25b)$$

durch. Aus einem Koeffizientenvergleich von (2.24a) und (2.25b) folgt, dass  $\hat{r}_{j+1} = \hat{l}_j$  genau dann erreicht wird, wenn  $a^2 = 1 - a$  gilt, d. h. wenn

$$a = \frac{\sqrt{5} - 1}{2} \approx 0.618. \quad (2.26)$$

Diese Wahl von  $a$  hat den Vorteil, dass je Iteration nur ein neuer Zwischenpunkt berechnet werden muss. Man beachte, dass die Berechnung jedes Zwischenpunktes mit einer Auswertung der Kostenfunktion  $g(\alpha_k)$  verbunden ist. Die Zahl  $1/a = 1 + a \approx 1.618$  ist bekannt als die Verhältniszahl des *Goldenen Schnittes*. Tabelle 2.2 fasst den Algorithmus nochmals zusammen.

Abschließend kann der optimale Wert  $\alpha_k^*$  entweder aus der *Mittelung* der letzten Intervallgrenzen  $\alpha_k^* = (l_j + r_j)/2$  oder aus einer *quadratischen Interpolation* zwischen den kleinsten drei der vier Funktionswerte gewonnen werden. Das Intervallschachtelungsverfahren ist ein *einfaches und robustes* Verfahren, das allerdings im Vergleich zu anderen Verfahren meist mehr Iterationen benötigt.

### 2.3.1.2 Quadratische Interpolation

Eine sehr effiziente Methode zur Lösung des skalaren Optimierungsproblems (2.23) besteht darin, durch drei Punkte eine quadratische Interpolationsfunktion zu legen. Dazu nehme man an, dass die voneinander paarweise verschiedenen Punkte  $\alpha_{k,1}$ ,  $\alpha_{k,2}$  und  $\alpha_{k,3}$  sowie deren Funktionswerte  $g_1 = g(\alpha_{k,1})$ ,  $g_2 = g(\alpha_{k,2})$  und  $g_3 = g(\alpha_{k,3})$  bekannt sind. Die quadratische Interpolationsfunktion  $q(\alpha_k)$  durch diese Punkte lautet dann

$$q(\alpha_k) = \sum_{i=1}^3 g_i \frac{\prod_{j \neq i} (\alpha_k - \alpha_{k,j})}{\prod_{j \neq i} (\alpha_{k,i} - \alpha_{k,j})} \quad (2.27)$$

und der optimale Wert  $\alpha_k^*$  errechnet sich zu

$$\alpha_k^* = \frac{1}{2} \frac{g_1(\alpha_{k,2}^2 - \alpha_{k,3}^2) + g_2(\alpha_{k,3}^2 - \alpha_{k,1}^2) + g_3(\alpha_{k,1}^2 - \alpha_{k,2}^2)}{g_1(\alpha_{k,2} - \alpha_{k,3}) + g_2(\alpha_{k,3} - \alpha_{k,1}) + g_3(\alpha_{k,1} - \alpha_{k,2})}. \quad (2.28)$$

Der so errechnete Wert  $\alpha_k^*$  sollte als optimale Schrittweite nur akzeptiert werden, wenn die Interpolationsfunktion  $q(\alpha_k)$  strikt konvex ist (nur dann ist  $\alpha_k^*$  tatsächlich ein Minimum von  $q(\alpha_k)$ ) und wenn die Bedingungen  $\alpha_k^* > 0$ ,  $g(\alpha_k^*) \leq g_1$ ,  $g(\alpha_k^*) \leq g_2$ ,  $g(\alpha_k^*) \leq g_3$  und  $g(\alpha_k^*) \leq g(0)$  gelten.

---

<b>Initialisierung:</b>	$l_0, r_0$	(Startintervall mit Minimum im Inneren)
	$j = 0$	(Startindex)
	$a = 0.618$	(Goldener Schnitt-Parameter)
	$\varepsilon_{lr}, \varepsilon_g$	(Schranken für Abbruch)
	$\hat{l}_0 \leftarrow al_0 + (1 - a)r_0$	(innere Punkte)
	$\hat{r}_0 \leftarrow (1 - a)l_0 + ar_0$	
<b>repeat</b>		
	<b>if</b> $g(\hat{l}_j) > g(\hat{r}_j)$	
	$l_{j+1} \leftarrow \hat{l}_j$	
	$r_{j+1} \leftarrow r_j$	
	$\hat{l}_{j+1} \leftarrow \hat{r}_j$	
	$\hat{r}_{j+1} \leftarrow (1 - a)l_{j+1} + ar_{j+1}$	
	<b>else</b> (d. h. $g(\hat{l}_j) \leq g(\hat{r}_j)$ )	
	$r_{j+1} \leftarrow \hat{r}_j$	
	$l_{j+1} \leftarrow l_j$	
	$\hat{r}_{j+1} \leftarrow \hat{l}_j$	
	$\hat{l}_{j+1} \leftarrow al_{j+1} + (1 - a)r_{j+1}$	
	<b>end if</b>	
	$j \leftarrow j + 1$	
<b>until</b>	$ r_j - l_j  \leq \varepsilon_{lr}$ <b>or</b> $ g(r_j) - g(l_j)  \leq \varepsilon_g$	

---

Tabelle 2.2: Intervallschachtelungsverfahren („Goldener Schnitt“).

**Aufgabe 2.4.** Zeigen Sie die Gültigkeit von (2.28).

### 2.3.1.3 Heuristische Wahl der Schrittweite

In der Praxis muss bei der Wahl der Schrittweite häufig ein Kompromiss zwischen numerischem Aufwand und erreichter Genauigkeit in Kauf genommen werden. Ungenauigkeiten treten z. B. auf, wenn ein iterativer Algorithmus zur Bestimmung der optimalen Schrittweite vorzeitig abgebrochen wird. Es gibt nun unterschiedliche *heuristische Abbruchkriterien*, die im Folgenden kurz erläutert werden. Den weiteren Betrachtungen liege das *skalare Optimierungsproblem*, siehe (2.23),

$$\min_{\alpha_k > 0} g(\alpha_k) = f(\mathbf{x}_k + \alpha_k \mathbf{s}_k) \quad (2.29)$$

zugrunde.

**Armijo-Bedingung:** Entwickelt man  $g(\alpha_k)$  um  $\alpha_k = 0$  in eine Taylorreihe und bricht

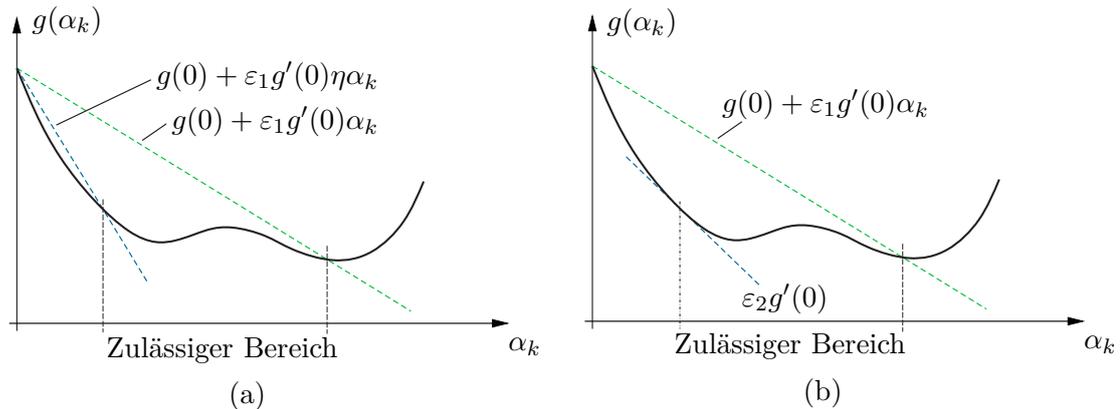


Abbildung 2.4: Veranschaulichung der Armijo- und Wolfe-Bedingung.

nach dem linearen Glied ab, so erhält man

$$g(\alpha_k) \approx g(0) + g'(0)\alpha_k . \quad (2.30)$$

Bei der *Armijo-Bedingung* wird nun die Schrittweite  $\alpha_k$  so gewählt, dass für ein festes  $\varepsilon_1$ ,  $0 < \varepsilon_1 < 1$ , die Ungleichung

$$g(\alpha_k) \leq g(0) + \varepsilon_1 g'(0)\alpha_k \quad (2.31)$$

erfüllt ist. Dies garantiert, dass die Schrittweite  $\alpha_k$  nach oben hin beschränkt ist. Um sicherzustellen, dass die Schrittweite nicht zu klein wird, führt man einen Parameter  $\eta > 1$  ein und fordert zusätzlich, dass die Schrittweite  $\alpha_k$  der Ungleichung

$$g(\alpha_k) > g(0) + \varepsilon_1 g'(0)\eta\alpha_k \quad (2.32)$$

genügt. Abbildung 2.4(a) gibt eine grafische Veranschaulichung dieses Sachverhaltes. In der Praxis geht man häufig so vor, dass man in einem ersten Schritt einen (weitgehend beliebigen) Startwert für  $\alpha_k$  wählt. Ist für dieses  $\alpha_k$  die Ungleichung (2.31) erfüllt, dann erhöht man  $\alpha_k$  sukzessive um den Faktor  $\eta$  solange, bis die Ungleichung (2.31) erstmals verletzt wird. Der vorletzte Wert von  $\alpha_k$  wird dann als geeignete Schrittweite gewählt. Umgekehrt, wenn der Startwert von  $\alpha_k$  die Ungleichung (2.31) nicht erfüllt, dann wird  $\alpha_k$  sukzessive durch den Faktor  $\eta$  dividiert, bis erstmals die Ungleichung (2.31) erfüllt ist. Typische Parameterwerte sind  $\varepsilon_1 = 0.1$  und  $\eta = 2$ . Man beachte jedoch, dass bei zu großem  $\varepsilon_1$  die Abstiegsbedingung zu restriktiv wird.

**Wolfe-Bedingung:** Wenn die Ableitungen der Kostenfunktion  $g(\alpha_k)$  sehr einfach berechnet werden können, eignet sich als Alternative zur Armijo-Bedingung die so genannte *Wolfe-Bedingung*. Dabei wird ein weiterer Parameter  $\varepsilon_2$  mit  $0 < \varepsilon_1 < \varepsilon_2 < 1$  eingeführt und von der Schrittweite  $\alpha_k$  wird gefordert, dass sie die Ungleichungen (2.31) und

$$g'(\alpha_k) \geq \varepsilon_2 g'(0) \quad (2.33)$$

erfüllt. Abbildung 2.4(b) gibt eine grafische Veranschaulichung dieses Sachverhaltes. Typische Werte für  $\varepsilon_2$  sind 0.9, wenn die Suchrichtung  $\mathbf{s}_k$  über die Newton-Methode oder die Quasi-Newton-Methode und 0.1, wenn  $\mathbf{s}_k$  über die konjugierte Gradientenmethode bestimmt wurde.

## 2.3.2 Wahl der Suchrichtung

### 2.3.2.1 Gradientenmethode

Bei der *Gradientenmethode*, sie wird auch *Methode des steilsten Abstiegs* (englisch: *steepest descent method*) genannt, wählt man als Suchrichtung  $\mathbf{s}_k$  in (2.21) den *negativen Gradienten* an der Stelle  $\mathbf{x}_k$ , d. h. die Abstiegsrichtung

$$\mathbf{s}_k = -(\nabla f)(\mathbf{x}_k) . \quad (2.34)$$

Entwickelt man  $g(\alpha_k) = f(\mathbf{x}_k + \alpha_k \mathbf{s}_k)$  um den Punkt  $\alpha_k = 0$  in eine Taylorreihe mit  $\mathbf{s}_k$  gemäß (2.34)

$$g(\alpha_k) = f(\mathbf{x}_k + \alpha_k \mathbf{s}_k) = f(\mathbf{x}_{k+1}) = f(\mathbf{x}_k) - \alpha_k \|(\nabla f)(\mathbf{x}_k)\|_2^2 + o(\alpha_k) , \quad (2.35)$$

wobei  $o(\alpha_k)$  den Restterm bezeichnet, der schneller nach Null abklingt als  $\alpha_k$ , dann erkennt man unmittelbar, dass für hinreichend kleines  $\alpha_k$  die Ungleichungsbedingung (2.13) für  $(\nabla f)(\mathbf{x}_k) \neq \mathbf{0}$  erfüllt ist.

Um die Konvergenzeigenschaften der Gradientenmethode näher zu untersuchen, betrachte man das quadratische Minimierungsproblem

$$\min_{\mathbf{x} \in \mathbb{R}^n} f(\mathbf{x}) = \frac{1}{2} \mathbf{x}^T \mathbf{Q} \mathbf{x} - \mathbf{x}^T \mathbf{b} \quad (2.36)$$

mit der positiv definiten Matrix  $\mathbf{Q} \in \mathbb{R}^{n \times n}$ . Da die Hessematrix  $(\nabla^2 f)(\mathbf{x}) = \mathbf{Q}$  von  $f(\mathbf{x})$  positiv definit ist, folgt aus der Eigenschaft (d) konvexer Funktionen von Abschnitt 1.3.5.2 die strikte Konvexität von  $f(\mathbf{x})$ . Auf Grund der Sätze 2.1 und 2.4 ergibt sich daher das strikte globale Minimum  $\mathbf{x}^*$  von  $f(\mathbf{x})$  aus der Beziehung

$$(\nabla f)(\mathbf{x}^*) = \mathbf{g}(\mathbf{x}^*) = \mathbf{Q} \mathbf{x}^* - \mathbf{b} = \mathbf{0} \quad (2.37)$$

in der Form

$$\mathbf{x}^* = \mathbf{Q}^{-1} \mathbf{b} . \quad (2.38)$$

Die Iterationsvorschrift gemäß Gradientenmethode lautet in diesem Fall, siehe (2.21)

$$\mathbf{x}_{k+1} = \mathbf{x}_k - \alpha_k \mathbf{g}_k \quad \text{mit} \quad \mathbf{g}_k = \mathbf{g}(\mathbf{x}_k) = \mathbf{Q} \mathbf{x}_k - \mathbf{b} . \quad (2.39)$$

Die optimale Schrittweite  $\alpha_k^*$  kann durch explizites Lösen des Optimierungsproblems gemäß (2.23)

$$\min_{\alpha_k > 0} f(\mathbf{x}_k + \alpha_k \mathbf{s}_k) = \frac{1}{2} (\mathbf{x}_k - \alpha_k \mathbf{g}_k)^T \mathbf{Q} (\mathbf{x}_k - \alpha_k \mathbf{g}_k) - (\mathbf{x}_k - \alpha_k \mathbf{g}_k)^T \mathbf{b} \quad (2.40)$$

in der Form

$$\alpha_k^* = \frac{\mathbf{g}_k^T \mathbf{g}_k}{\mathbf{g}_k^T \mathbf{Q} \mathbf{g}_k} \quad (2.41)$$

berechnet werden.

**Aufgabe 2.5.** Zeigen Sie die Gültigkeit von (2.41).

Zusammenfassend lässt sich damit die Gradientenmethode für die quadratische Kostenfunktion (2.36) wie folgt

$$\mathbf{x}_{k+1} = \mathbf{x}_k - \frac{\mathbf{g}_k^T \mathbf{g}_k}{\mathbf{g}_k^T \mathbf{Q} \mathbf{g}_k} \mathbf{g}_k, \quad \mathbf{g}_k = \mathbf{Q} \mathbf{x}_k - \mathbf{b} \quad (2.42)$$

anschreiben.

Für die weiteren Überlegungen ist es vorteilhaft, anstelle von  $f(\mathbf{x})$  die Kostenfunktion

$$F(\mathbf{x}) = f(\mathbf{x}) + \frac{1}{2}(\mathbf{x}^*)^T \mathbf{Q} \mathbf{x}^* = \frac{1}{2}(\mathbf{x} - \mathbf{x}^*)^T \mathbf{Q}(\mathbf{x} - \mathbf{x}^*) \quad (2.43)$$

zu betrachten. Da sich die beiden Kostenfunktionen  $f(\mathbf{x})$  und  $F(\mathbf{x})$  lediglich um eine Konstante unterscheiden, sind ihre Formen und Minima  $\mathbf{x}^*$  identisch.

**Lemma 2.1** (Zur Konvergenzrate des Gradientenverfahrens). *Mit der Iterationsvorschrift des Gradientenverfahrens (2.42) gilt für die Kostenfunktion  $F(\mathbf{x})$  die Beziehung*

$$F(\mathbf{x}_{k+1}) = \left( 1 - \frac{(\mathbf{g}_k^T \mathbf{g}_k)^2}{\mathbf{g}_k^T \mathbf{Q} \mathbf{g}_k \mathbf{g}_k^T \mathbf{Q}^{-1} \mathbf{g}_k} \right) F(\mathbf{x}_k). \quad (2.44)$$

*Beweis.* Aus (2.38) und (2.42) erhält man

$$\mathbf{g}_k = \mathbf{Q} \mathbf{x}_k - \mathbf{b} = \mathbf{Q}(\mathbf{x}_k - \mathbf{x}^*). \quad (2.45)$$

Folglich gilt

$$F(\mathbf{x}_k) = \frac{1}{2} \mathbf{g}_k^T \mathbf{Q}^{-1} \mathbf{g}_k. \quad (2.46)$$

Aus diesen Beziehungen und der Iterationsvorschrift (2.42) lässt sich nun direkt (2.44) berechnen.

$$\begin{aligned} F(\mathbf{x}_{k+1}) &= \frac{1}{2} \left( \mathbf{x}_k - \frac{\mathbf{g}_k^T \mathbf{g}_k}{\mathbf{g}_k^T \mathbf{Q} \mathbf{g}_k} \mathbf{g}_k - \mathbf{x}^* \right)^T \mathbf{Q} \left( \mathbf{x}_k - \frac{\mathbf{g}_k^T \mathbf{g}_k}{\mathbf{g}_k^T \mathbf{Q} \mathbf{g}_k} \mathbf{g}_k - \mathbf{x}^* \right) \\ &= \frac{1}{2} (\mathbf{x}_k - \mathbf{x}^*)^T \mathbf{Q} (\mathbf{x}_k - \mathbf{x}^*) - \frac{1}{2} \frac{(\mathbf{g}_k^T \mathbf{g}_k)^2}{\mathbf{g}_k^T \mathbf{Q} \mathbf{g}_k} \\ &= \left( 1 - \frac{(\mathbf{g}_k^T \mathbf{g}_k)^2}{\mathbf{g}_k^T \mathbf{Q} \mathbf{g}_k \mathbf{g}_k^T \mathbf{Q}^{-1} \mathbf{g}_k} \right) F(\mathbf{x}_k) \end{aligned} \quad (2.47)$$

□

Um nun die Konvergenzrate der Gradientenmethode für die quadratische Kostenfunktion abschätzen zu können, benötigt man noch folgendes Lemma.

**Lemma 2.2 (Ungleichung von Kantorovich).** Es sei  $\mathbf{Q} \in \mathbb{R}^{n \times n}$  eine symmetrische positiv definite Matrix. Für jeden Vektor  $\mathbf{x} \in \mathbb{R}^n$  gilt dann die Ungleichung

$$\frac{(\mathbf{x}^T \mathbf{x})^2}{(\mathbf{x}^T \mathbf{Q} \mathbf{x})(\mathbf{x}^T \mathbf{Q}^{-1} \mathbf{x})} \geq \frac{4\lambda_{\min} \lambda_{\max}}{(\lambda_{\min} + \lambda_{\max})^2}, \quad (2.48)$$

wobei  $\lambda_{\min}$  und  $\lambda_{\max}$  den kleinsten und größten (reellen und positiven) Eigenwert der Matrix  $\mathbf{Q}$  bezeichnen.

**Aufgabe 2.6.** Beweisen Sie Lemma 2.2. **Hinweis:** Dieser Beweis ist z. B. in [2.2] skizziert.

Damit lässt sich folgender Satz angeben.

**Satz 2.6 (Konvergenz der Gradientenmethode — Quadratische Kostenfunktion).** Für jeden Anfangswert  $\mathbf{x}_0 \in \mathbb{R}^n$  konvergiert die Iterationsvorschrift (2.42) der Gradientenmethode gegen das eindeutige globale Minimum  $\mathbf{x}^*$  der Kostenfunktion  $f(\mathbf{x})$  gemäß (2.36) bzw.  $F(\mathbf{x})$  gemäß (2.43) linear mit der Konvergenzrate

$$F(\mathbf{x}_{k+1}) \leq \left( \frac{\kappa - 1}{\kappa + 1} \right)^2 F(\mathbf{x}_k), \quad (2.49)$$

wobei  $\kappa = \lambda_{\max}/\lambda_{\min}$  die spektrale Konditionszahl der Matrix  $\mathbf{Q}$ , also das Verhältnis des größten zum kleinsten (reellen und positiven) Eigenwert  $\lambda_{\max}$  und  $\lambda_{\min}$  der Matrix  $\mathbf{Q}$ , bezeichnet.

*Beweis.* Aus den Lemmas 2.1 und 2.2 folgt unmittelbar

$$F(\mathbf{x}_{k+1}) \leq \left\{ 1 - \frac{4\lambda_{\min} \lambda_{\max}}{(\lambda_{\min} + \lambda_{\max})^2} \right\} F(\mathbf{x}_k) = \left( \frac{\lambda_{\max} - \lambda_{\min}}{\lambda_{\max} + \lambda_{\min}} \right)^2 F(\mathbf{x}_k). \quad (2.50)$$

□

Satz 2.6 lässt sich nun wie folgt interpretieren. Auf Grund der positiven Definitheit der Matrix  $\mathbf{Q}$  sind die Höhenlinien ( $f(\mathbf{x}) = \text{konst.}$ ) der Kostenfunktion (2.36)  $n$ -dimensionale Ellipsoide, deren Achsen mit den Richtungen der  $n$  paarweise orthogonalen Eigenvektoren der Matrix  $\mathbf{Q}$  zusammenfallen und deren Längen invers proportional zum jeweiligen (positiv reellen) Eigenwert sind. Der Gradient ( $\nabla f$ )( $\mathbf{x}_k$ ) steht orthogonal zur Höhenlinie durch den Punkt  $\mathbf{x}_k$ , siehe Abbildungen 2.5 und 2.6. Wenn die Eigenwerte von  $\mathbf{Q}$  in (2.36) alle in der gleichen Größenordnung liegen, weist die Gradientenmethode ein gutes Konvergenzverhalten auf, im Falle von  $\lambda_{\min} = \lambda_{\max}$  bzw.  $\kappa = 1$  konvergiert das Verfahren sogar in einem einzigen Schritt, siehe Abbildung 2.5. Bei schlecht konditionierten Problemen ( $\kappa$  sehr groß) konvergiert die Gradientenmethode sehr langsam, siehe Abbildung 2.6.

Die Gradientenmethode kann natürlich auch auf nichtquadratische Kostenfunktionen angewandt werden. Für diesen Fall beschreibt der nachfolgende Satz das Konvergenzverhalten der Gradientenmethode. Sein Beweis findet sich z. B. in [2.2].

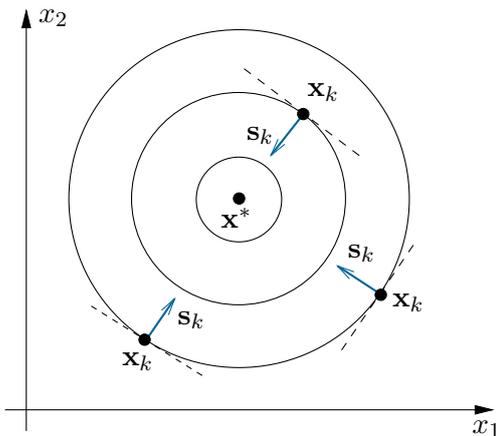


Abb. 2.5: Beispiel eines ideal konditionierten Problems für die Gradientenmethode.

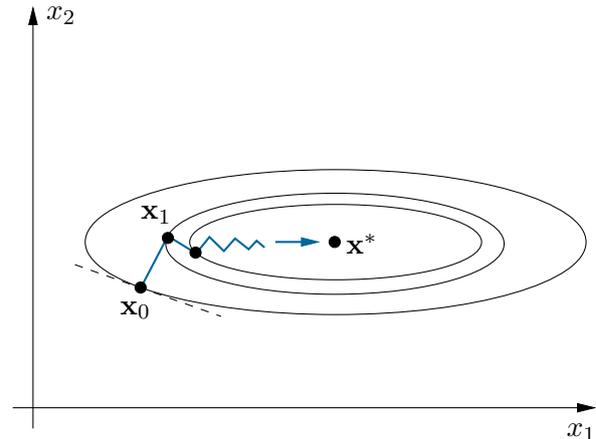


Abb. 2.6: Beispiel eines schlecht konditionierten Problems für die Gradientenmethode.

**Satz 2.7 (Konvergenz der Gradientenmethode — Allgemeine Kostenfunktion).** Gegeben sei die Kostenfunktion  $f \in C^2$  definiert im  $\mathbb{R}^n$  mit  $\mathbf{x}^*$  als lokales Minimum. Angenommen, die Hessematrix  $(\nabla^2 f)(\mathbf{x}^*)$  hat den kleinsten und größten Eigenwert  $\lambda_{\min} > 0$  und  $\lambda_{\max} > 0$  und die spektrale Konditionszahl  $\kappa = \lambda_{\max}/\lambda_{\min}$ . Wenn die Folge  $\{\mathbf{x}_k\}$  generiert durch die Gradientenmethode

$$\mathbf{x}_{k+1} = \mathbf{x}_k - \alpha_k (\nabla f)(\mathbf{x}_k) \quad (2.51)$$

für eine geeignete Schrittweite  $\alpha_k$  gegen das lokale Minimum  $\mathbf{x}^*$  konvergiert, dann konvergiert die Folge  $\{f(\mathbf{x}_k)\}$  linear gegen  $f(\mathbf{x}^*)$  mit einer Konvergenzrate von maximal  $\left(\frac{\kappa-1}{\kappa+1}\right)^2$ .

Schlecht konditionierte Problemstellungen bei der Gradientenmethode können mitunter durch eine geeignete *Skalierung* oder *Transformation* verbessert werden. Die Idee beruht darauf, dass die Aufgabe, ein Minimum der Kostenfunktion  $f(\mathbf{x})$  zu finden, äquivalent dazu ist, für die Funktion  $h(\mathbf{y}) = f(\mathbf{T}\mathbf{y})$  mit  $\mathbf{x} = \mathbf{T}\mathbf{y}$  und der regulären Matrix  $\mathbf{T} \in \mathbb{R}^{n \times n}$  ein Minimum zu suchen. Entwickelt man die Funktion  $h(\mathbf{y})$  um den optimalen Punkt  $\mathbf{y}^* = \mathbf{T}^{-1}\mathbf{x}^*$  in eine Taylorreihe

$$\begin{aligned} h(\mathbf{y}) &= h(\mathbf{y}^*) + (\nabla h)^T(\mathbf{y}^*)(\mathbf{y} - \mathbf{y}^*) + \frac{1}{2}(\mathbf{y} - \mathbf{y}^*)^T (\nabla^2 h)(\mathbf{y}^*)(\mathbf{y} - \mathbf{y}^*) + \dots \\ &= h(\mathbf{y}^*) + (\nabla f)^T(\mathbf{x}^*)\mathbf{T}(\mathbf{y} - \mathbf{y}^*) + \frac{1}{2}(\mathbf{y} - \mathbf{y}^*)^T \mathbf{T}^T (\nabla^2 f)(\mathbf{x}^*)\mathbf{T}(\mathbf{y} - \mathbf{y}^*) + \dots, \end{aligned} \quad (2.52)$$

so erkennt man, dass durch geeignete Wahl von  $\mathbf{T}$  die Verteilung der Eigenwerte der Hessematrix

$$(\nabla^2 h)(\mathbf{y}^*) = \mathbf{T}^T (\nabla^2 f)(\mathbf{x}^*)\mathbf{T} \quad (2.53)$$

gegenüber jener der Eigenwerte von  $(\nabla^2 f)(\mathbf{x}^*)$  verbessert werden kann. Aus (2.53) folgt, dass mit der idealen Wahl  $\mathbf{T} = (\nabla^2 f)^{-\frac{1}{2}}(\mathbf{x}^*)$  für die Hessematrix  $(\nabla^2 h)(\mathbf{y}^*) = \mathbf{E}$  mit der Einheitsmatrix  $\mathbf{E} \in \mathbb{R}^{n \times n}$  folgen würde und das Gradientenverfahren bei quadratischen Optimierungsproblemen nach einem Schritt konvergieren würde (vgl. Abbildung 2.5). Praktisch ist diese Vorgehensweise sehr ähnlich zur nachfolgend beschriebenen Newton-Methode. Sie hat aber den Nachteil, dass die Hessematrix explizit berechnet werden muss. Um diesen Rechenaufwand zu vermeiden, wird alternativ häufig eine Diagonalmatrix  $\mathbf{T}$  verwendet, deren Diagonaleinträge beispielsweise in der Form  $T_{ii} = ((\nabla^2 f)_{ii}(\mathbf{x}^*))^{-\frac{1}{2}}$ ,  $i = 1, \dots, n$  gewählt werden können. Wird eine Diagonalmatrix  $\mathbf{T}$  verwendet, so führt dies zu einer reinen Skalierung oder Normierung der Optimierungsvariablen.

Die Vor- und Nachteile der Gradientenmethode lassen sich wie folgt zusammenfassen:

- + einfaches Verfahren
- + Hessematrix  $(\nabla^2 f)(\mathbf{x}_k)$  (Berechnungsaufwand  $\mathcal{O}(n^2)$ ) nicht erforderlich, nur der Gradient  $(\nabla f)(\mathbf{x}_k)$  (Berechnungsaufwand  $\mathcal{O}(n)$ ) wird benötigt
- + Konvergenz auch für Startwerte, die weiter vom Minimum entfernt sind
- langsame Konvergenz bei schlecht konditionierten und schlecht skalierten Problemen
- lediglich lineare Konvergenzordnung

### 2.3.2.2 Konjugierte Gradientenmethode

Die konjugierte Gradientenmethode (englisch: *conjugate gradient method* oder kurz *C-G method*) versucht nun bei nur geringfügig erhöhtem Rechenaufwand, eine schnellere Konvergenz als die Gradientenmethode zu erreichen. Ursprünglich wurde diese Methode für hochdimensionale quadratische Probleme der Form (siehe auch (2.36))

$$\min_{\mathbf{x} \in \mathbb{R}^n} f(\mathbf{x}) = \frac{1}{2} \mathbf{x}^T \mathbf{Q} \mathbf{x} - \mathbf{x}^T \mathbf{b} \quad (2.54)$$

mit der positiv definiten Matrix  $\mathbf{Q} \in \mathbb{R}^{n \times n}$  entwickelt. Bevor nun die Methode genauer erläutert wird, sollen einige Grundlagen dazu erarbeitet werden.

**Definition 2.2 (Q-Orthogonalität).** Zwei Vektoren  $\mathbf{d}_1$  und  $\mathbf{d}_2$  heißen *konjugiert bezüglich einer positiv definiten Matrix Q* bzw. *Q-orthogonal*, wenn gilt  $\mathbf{d}_1^T \mathbf{Q} \mathbf{d}_2 = 0$ .

Für  $\mathbf{Q} = \mathbf{E}$  fällt der Begriff der Konjugiertheit mit dem klassischen Begriff der Orthogonalität zusammen. Eine Menge von Vektoren  $\mathbf{d}_0, \mathbf{d}_1, \dots, \mathbf{d}_r$  ist *Q-orthogonal*, wenn  $\mathbf{d}_i^T \mathbf{Q} \mathbf{d}_j = 0$  für alle  $i \neq j$ . Es gilt nun folgendes Lemma.

**Lemma 2.3 (Q-Orthogonalität positiv definiter Matrizen).** Wenn die Matrix  $\mathbf{Q}$  positiv definit ist und eine Menge von nichttrivialen Vektoren  $\mathbf{d}_0, \mathbf{d}_1, \dots, \mathbf{d}_r$  Q-orthogonal ist, dann sind die Vektoren  $\mathbf{d}_j$ ,  $j = 0, \dots, r$  linear unabhängig.

**Aufgabe 2.7.** Beweisen Sie das Lemma 2.3.

Für das Folgende sei angenommen, dass  $\mathbf{d}_0, \mathbf{d}_1, \dots, \mathbf{d}_{n-1}$  nichttriviale Q-orthogonale Vektoren der Matrix  $\mathbf{Q}$  der Kostenfunktion des Optimierungsproblems (2.54) sind. Nach

Lemma 2.3 sind die  $n$  Vektoren linear unabhängig und spannen daher den  $\mathbb{R}^n$  auf. Die optimale Lösung  $\mathbf{x}^*$  des Optimierungsproblems (2.54) lässt sich somit als Linearkombination der  $\mathbf{Q}$ -orthogonalen Vektoren in der Form

$$\mathbf{x}^* = \underbrace{\begin{bmatrix} \mathbf{d}_0 & \mathbf{d}_1 & \dots & \mathbf{d}_{n-1} \end{bmatrix}}_{=\mathbf{D}} \boldsymbol{\eta} \quad (2.55)$$

mit  $\boldsymbol{\eta} \in \mathbb{R}^n$  darstellen. Alternativ zur direkten Berechnung von  $\boldsymbol{\eta}$  aus (2.55) erhält man mit  $\mathbf{Q}\mathbf{x}^* = \mathbf{b}$  aus (2.37) durch Vormultiplikation von (2.55) mit  $\mathbf{D}^T\mathbf{Q}$

$$\mathbf{D}^T\mathbf{Q}\mathbf{x}^* = \mathbf{D}^T\mathbf{b} = \mathbf{D}^T\mathbf{Q}\mathbf{D}\boldsymbol{\eta}. \quad (2.56)$$

Der sich daraus ergebende Wert für  $\boldsymbol{\eta}$  wird wieder in (2.55) eingesetzt und man erhält für die optimale Lösung

$$\mathbf{x}^* = \mathbf{D}(\mathbf{D}^T\mathbf{Q}\mathbf{D})^{-1}\mathbf{D}^T\mathbf{b} = \sum_{i=0}^{n-1} \frac{\mathbf{d}_i^T\mathbf{b}}{\mathbf{d}_i^T\mathbf{Q}\mathbf{d}_i} \mathbf{d}_i, \quad (2.57)$$

wobei hier die  $\mathbf{Q}$ -Orthogonalität der Vektoren  $\mathbf{d}_i$  (Diagonalität der Matrix  $\mathbf{D}^T\mathbf{Q}\mathbf{D}$ ) ausgenutzt wurde. Diese Darstellung bildet auch die Grundlage für den nächsten Satz.

**Satz 2.8 (Konjugierte Richtung).** Die Vektoren  $\mathbf{d}_0, \mathbf{d}_1, \dots, \mathbf{d}_{n-1}$  seien nichttriviale  $\mathbf{Q}$ -orthogonale Vektoren der Matrix  $\mathbf{Q}$  der Kostenfunktion des Optimierungsproblems (2.54). Für jeden Anfangswert  $\mathbf{x}_0$  konvergiert die Folge

$$\mathbf{x}_{k+1} = \mathbf{x}_k + \alpha_k \mathbf{d}_k, \quad \alpha_k = -\frac{\mathbf{g}_k^T \mathbf{d}_k}{\mathbf{d}_k^T \mathbf{Q} \mathbf{d}_k}, \quad \mathbf{g}_k = \mathbf{Q}\mathbf{x}_k - \mathbf{b} \quad (2.58)$$

nach spätestens  $n$  Iterationsschritten gegen die eindeutige optimale Lösung  $\mathbf{x}^*$ , d. h.  $\mathbf{x}_n = \mathbf{x}^*$ .

*Beweis.* Nach Lemma 2.3 sind die Vektoren  $\mathbf{d}_0, \mathbf{d}_1, \dots, \mathbf{d}_{n-1}$  linear unabhängig und daher findet man geeignete Skalare  $\alpha_i$  so, dass gilt

$$\mathbf{x}^* - \mathbf{x}_0 = \alpha_0 \mathbf{d}_0 + \alpha_1 \mathbf{d}_1 + \dots + \alpha_{n-1} \mathbf{d}_{n-1}. \quad (2.59)$$

Multipliziert man (2.59) von links mit  $\mathbf{Q}$  und bildet man das Skalarprodukt mit  $\mathbf{d}_k$ , dann erhält man

$$\alpha_k = \frac{\mathbf{d}_k^T \mathbf{Q}(\mathbf{x}^* - \mathbf{x}_0)}{\mathbf{d}_k^T \mathbf{Q} \mathbf{d}_k}. \quad (2.60)$$

Aus (2.58) folgt durch rekursives Einsetzen

$$\mathbf{x}_k - \mathbf{x}_0 = \alpha_0 \mathbf{d}_0 + \alpha_1 \mathbf{d}_1 + \dots + \alpha_{k-1} \mathbf{d}_{k-1} \quad (2.61)$$

und auf Grund der  $\mathbf{Q}$ -Orthogonalität der Vektoren  $\mathbf{d}_i$  gilt

$$\mathbf{d}_k^T \mathbf{Q}(\mathbf{x}_k - \mathbf{x}_0) = 0. \quad (2.62)$$

Setzt man (2.62) in (2.60) ein, so erhält man unmittelbar das Ergebnis von (2.58)

$$\alpha_k = \frac{\mathbf{d}_k^T \mathbf{Q}(\mathbf{x}^* - \mathbf{x}_k)}{\mathbf{d}_k^T \mathbf{Q} \mathbf{d}_k} = -\frac{\mathbf{d}_k^T (\mathbf{Q} \mathbf{x}_k - \mathbf{b})}{\mathbf{d}_k^T \mathbf{Q} \mathbf{d}_k} = -\frac{\mathbf{d}_k^T \mathbf{g}_k}{\mathbf{d}_k^T \mathbf{Q} \mathbf{d}_k}. \quad (2.63)$$

□

Für eine geometrische Interpretation der konjugierten Gradientenmethode betrachte man die linearen Unterräume  $\mathcal{B}_k = \text{span}\{\mathbf{d}_0, \mathbf{d}_1, \dots, \mathbf{d}_{k-1}\}$ . Man kann nun mit Hilfe der vollständigen Induktion zeigen, dass der Gradient  $\mathbf{g}_k$  zum Iterationsschritt  $k$  orthogonal auf den Unterraum  $\mathcal{B}_k$  ist. Da  $\mathcal{B}_0 = \{\}$ , ist die Aussage trivialerweise für  $k = 0$  erfüllt. Angenommen es gilt  $\mathbf{g}_k \perp \mathcal{B}_k$ , dann soll im nächsten Schritt gezeigt werden, dass  $\mathbf{g}_{k+1} \perp \mathcal{B}_{k+1}$  ist. Aus (2.58) folgt

$$\mathbf{g}_{k+1} = \mathbf{Q} \mathbf{x}_{k+1} - \mathbf{b} = \mathbf{Q} \mathbf{x}_k - \mathbf{b} + \alpha_k \mathbf{Q} \mathbf{d}_k = \mathbf{g}_k + \alpha_k \mathbf{Q} \mathbf{d}_k \quad (2.64)$$

und damit gilt wegen der Definition von  $\alpha_k$  gemäß (2.58)

$$\mathbf{d}_k^T \mathbf{g}_{k+1} = \mathbf{d}_k^T \mathbf{g}_k + \alpha_k \mathbf{d}_k^T \mathbf{Q} \mathbf{d}_k = 0. \quad (2.65)$$

Basierend auf diesem Ergebnis und der  $\mathbf{Q}$ -Orthogonalität der Vektoren  $\mathbf{d}_i$  kann nun rekursiv für  $j = k - 1, j = k - 2, \dots, j = 0$

$$\mathbf{d}_j^T \mathbf{g}_{k+1} = \mathbf{d}_j^T \mathbf{g}_k + \alpha_k \mathbf{d}_j^T \mathbf{Q} \mathbf{d}_k = 0 \quad (2.66)$$

gezeigt werden, womit obige Aussage bewiesen ist. Die konjugierte Gradientenmethode kann nun so geometrisch interpretiert werden, dass  $\mathbf{x}_k$  die Kostenfunktion  $f(\mathbf{x})$  jeweils im affinen Unterraum  $\mathbf{x}_0 + \mathcal{B}_k$  minimiert.

Die Frage, die es nun noch zu klären gilt, ist, wie die  $\mathbf{Q}$ -orthogonalen Vektoren  $\mathbf{d}_0, \mathbf{d}_1, \dots, \mathbf{d}_{n-1}$  festgelegt werden. Diese werden bei der konjugierten Gradientenmethode sukzessive bestimmt, wie dies im folgenden Satz beschrieben ist. Seine Herleitung findet sich z. B. in [2.3].

**Satz 2.9 (Konjugierte Gradientenmethode).** *Für jeden Anfangswert  $\mathbf{x}_0$  konvergiert die Folge*

$$\mathbf{x}_{k+1} = \mathbf{x}_k + \alpha_k \mathbf{d}_k \quad (2.67a)$$

$$\alpha_k = -\frac{\mathbf{d}_k^T \mathbf{g}_k}{\mathbf{d}_k^T \mathbf{Q} \mathbf{d}_k} \quad (2.67b)$$

$$\mathbf{d}_{k+1} = -\mathbf{g}_{k+1} + \beta_k \mathbf{d}_k \quad (2.67c)$$

$$\beta_k = \frac{\mathbf{g}_{k+1}^T \mathbf{Q} \mathbf{d}_k}{\mathbf{d}_k^T \mathbf{Q} \mathbf{d}_k} \quad (2.67d)$$

mit  $\mathbf{g}_k = \mathbf{Q} \mathbf{x}_k - \mathbf{b}$  und  $\mathbf{d}_0 = -\mathbf{g}_0 = \mathbf{b} - \mathbf{Q} \mathbf{x}_0$  in höchstens  $n$  Iterationsschritten gegen die eindeutige optimale Lösung  $\mathbf{x}^*$  des Optimierungsproblems (2.54).

Es kann sehr einfach gezeigt werden, dass die iterativ bestimmten Vektoren  $\mathbf{d}_i$  die Eigenschaft der  $\mathbf{Q}$ -Orthogonalität aufweisen. Man erkennt, dass im ersten Iterationsschritt

mit  $\mathbf{d}_0 = -\mathbf{g}_0$  ein reiner Gradientenschritt durchgeführt wird und anschließend die neue Suchrichtung  $\mathbf{d}_{k+1}$  über eine Linearkombination des momentanen Gradienten  $\mathbf{g}_{k+1}$  und der vorigen Suchrichtung  $\mathbf{d}_k$  bestimmt wird.

Für viele praktische Fragestellungen zeigt die sogenannte *partielle konjugierte Gradientenmethode* große Vorteile. Dabei wird die konjugierte Gradientenmethode von Satz 2.9 lediglich für  $m + 1 < n$  Iterationsschritte ausgeführt ehe sie mit dem so erhaltenen Punkt als Startlösung neu gestartet wird. Bei jedem Aufruf des Verfahrens werden  $m + 1$  Iterationen durchgeführt. In diesem Zusammenhang kann folgender Satz angegeben werden, welcher z. B. in [2.2] bewiesen wird.

**Satz 2.10 (Partielle konjugierte Gradientenmethode).** *Gegeben ist das Optimierungsproblem (2.54) mit der Kostenfunktion  $f(\mathbf{x})$  oder äquivalent dazu mit der Kostenfunktion  $F(\mathbf{x})$  gemäß (2.43). Wenn nun die positiv definite Matrix  $\mathbf{Q}$   $n - m$  Eigenwerte im Intervall  $[l, r]$  ( $l > 0$ ) und  $m$  Eigenwerte größer als  $r$  besitzt, dann zeigt die partielle konjugierte Gradientenmethode, welche alle  $m + 1$  Schritte neu gestartet wird, das Konvergenzverhalten*

$$F(\mathbf{x}_{k+1}) \leq \left( \frac{r-l}{r+l} \right)^2 F(\mathbf{x}_k). \quad (2.68)$$

Man beachte, dass der Punkt  $\mathbf{x}_{k+1}$  durch  $(m + 1)$ -fache Zwischeniteration nach Satz 2.9 mit dem Anfangswert  $\mathbf{x}_k$  entsteht. Satz 2.10 zeigt, dass durch Anwendung der partiellen konjugierten Gradientenmethode das schlechte Konvergenzverhalten der Gradientenmethode bei schlecht konditionierten Systemen (vergleiche dazu Satz 2.7) umgangen werden kann.

Für *nichtquadratische Kostenfunktionen*  $f(\mathbf{x})$  müssen in Satz 2.9 lediglich die Substitutionen

$$\mathbf{g}_k \leftrightarrow (\nabla f)(\mathbf{x}_k) \quad \text{und} \quad \mathbf{Q} \leftrightarrow (\nabla^2 f)(\mathbf{x}_k) \quad (2.69)$$

vorgenommen werden. An dieser Stelle ist jedoch zu erwähnen, dass der Algorithmus im Allgemeinen nicht wie im quadratischen Fall in  $n$  Schritten terminieren wird. Um die aufwändige Berechnung der Hessematrix  $(\nabla^2 f)(\mathbf{x}_k)$  zu vermeiden, kann die Bestimmung von  $\alpha_k$  in (2.67b) nach Satz 2.9 über ein Verfahren aus Abschnitt 2.3.1 erfolgen und  $\beta_k$  in (2.67d) wird beispielsweise durch die sogenannte *Formel von Fletcher-Reeves*

$$\beta_k = \frac{\mathbf{g}_{k+1}^T \mathbf{g}_{k+1}}{\mathbf{g}_k^T \mathbf{g}_k} \quad (2.70)$$

ersetzt.

Die Vor- und Nachteile der Konjugierten Gradientenmethode können wie folgt zusammengefasst werden:

- + einfaches Verfahren, geringer Rechenaufwand und Speicherbedarf, geeignet für große Optimierungsprobleme
- + nur der Gradient  $(\nabla f)(\mathbf{x}_k)$  (Berechnungsaufwand  $\mathcal{O}(n)$ ) wird benötigt
- + konvergiert bei quadratischen Optimierungsproblemen nach spätestens  $n$  Iterationsschritten

- Konvergenzverhalten variiert je nach Problemstellung, Konvergenzverhalten besser als jenes der Gradientenmethode

### 2.3.2.3 Newton-Methode

Die Idee der Newton-Methode besteht darin, die allgemeine Kostenfunktion  $f(\mathbf{x})$  lokal durch eine quadratische Funktion zu approximieren und diese zu minimieren. Entwickelt man  $f(\mathbf{x}) = f(\mathbf{x}_k + \mathbf{s}_k)$  um den Iterationspunkt  $\mathbf{x}_k$  in eine Taylorreihe und bricht diese nach dem quadratischen Term ab, so erhält man

$$f(\mathbf{x}) \approx f(\mathbf{x}_k) + \mathbf{s}_k^T (\nabla f)(\mathbf{x}_k) + \frac{1}{2} \mathbf{s}_k^T (\nabla^2 f)(\mathbf{x}_k) \mathbf{s}_k, \quad (2.71)$$

siehe Abbildung 2.7. Die so genannte *Newton-Richtung*  $\mathbf{s}_k$  ergibt sich unmittelbar durch

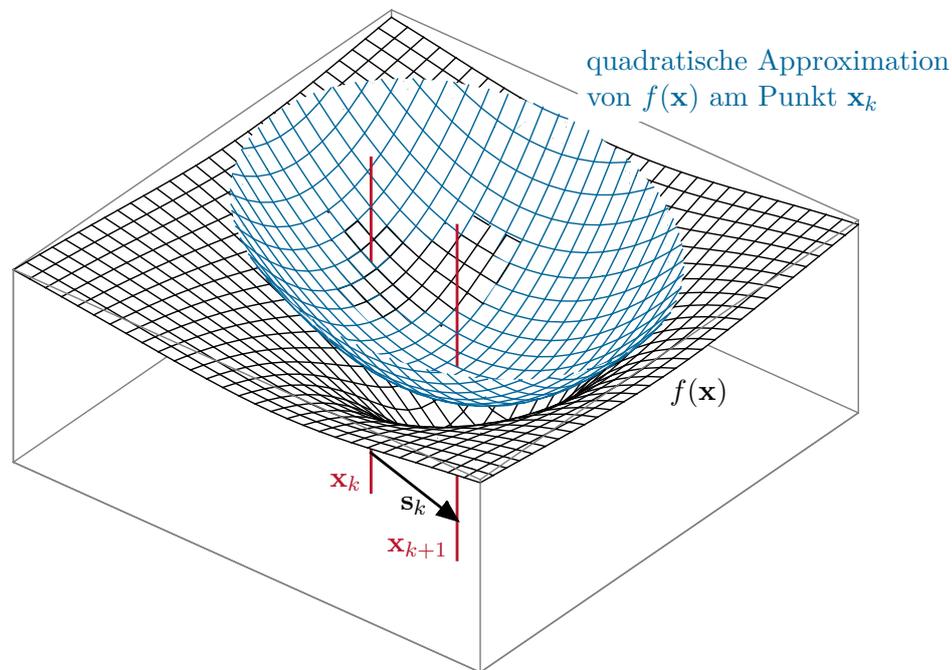


Abbildung 2.7: Lokale quadratische Approximation der Kostenfunktion im Zuge der Newton-Iteration (bei Schrittweite  $\alpha_k = 1$ ).

Minimierung der rechten Seite von (2.71) bezüglich  $\mathbf{s}_k$  in der Form

$$\mathbf{s}_k = -(\nabla^2 f)^{-1}(\mathbf{x}_k) (\nabla f)(\mathbf{x}_k). \quad (2.72)$$

Falls die Hessematrix  $(\nabla^2 f)(\mathbf{x}^*)$  am Minimum positiv definit ist, existiert in einer Umgebung um das Minimum die Inverse  $(\nabla^2 f)^{-1}(\mathbf{x}_k)$  und die Methode ist wohldefiniert. Man beachte, dass die Berechnung von  $\mathbf{s}_k$  gemäß (2.72) keine tatsächliche Inversion von  $(\nabla^2 f)(\mathbf{x}_k)$  erfordert. Der nachfolgende Satz gibt die Konvergenzordnung der Newton-Methode an. Sein Beweis ist z. B. in [2.2–2.4] zu finden.

**Satz 2.11 (Konvergenzordnung der Newton-Methode).** Gegeben sei die Kostenfunktion  $f \in C^3$  definiert im  $\mathbb{R}^n$  mit dem lokalen Minimum  $\mathbf{x}^*$ . Wenn die Hessematrix  $(\nabla^2 f)(\mathbf{x}^*)$  positiv definit ist und der Anfangswert  $\mathbf{x}_0$  in einer hinreichend nahen Umgebung des Minimums liegt, dann konvergiert die Newton-Iteration

$$\mathbf{x}_{k+1} = \mathbf{x}_k - (\nabla^2 f)^{-1}(\mathbf{x}_k)(\nabla f)(\mathbf{x}_k) \quad (2.73)$$

mit der Konvergenzordnung 2 gegen das Minimum  $\mathbf{x}^*$ .

Für die praktische Anwendung der Methode führt man noch eine geeignete Schrittweite  $\alpha_k$  ein, so dass die Iterationsvorschrift

$$\mathbf{x}_{k+1} = \mathbf{x}_k - \alpha_k (\nabla^2 f)^{-1}(\mathbf{x}_k)(\nabla f)(\mathbf{x}_k) \quad (2.74)$$

lautet und die Abstiegsbedingung (2.22) erfüllt ist. Das Verfahren wird dann häufig *gedämpfte Newton-Methode* genannt und  $\alpha_k$  wird als *Dämpfungsparameter* bezeichnet. Gelegentlich wird die Einschränkung  $\alpha_k \leq 1$  verwendet. Es ist zu erwarten, dass in der Nähe des Minimums die optimale Schrittweite  $\alpha_k \approx 1$  ist, weshalb man typischerweise die Iteration mit dem Wert  $\alpha_k = 1$  beginnt. Strategien zur Berechnung der Schrittweite  $\alpha_k$  wurden bereits im Abschnitt 2.3.1 erläutert.

Ein Problem, das in diesem Zusammenhang gelegentlich auftritt, besteht im Verlust der positiven Definitheit von  $(\nabla^2 f)(\mathbf{x}_k)$ , wenn  $\mathbf{x}_k$  zu weit vom Minimum entfernt ist. Dann besitzt die rechte Seite von (2.71) kein oder kein eindeutiges Minimum und die Invertierbarkeit von  $(\nabla^2 f)(\mathbf{x}_k)$  kann verloren gehen. Ist  $(\nabla^2 f)(\mathbf{x}_k)$  nicht positiv definit, so kann statt (2.74) die Iterationsvorschrift

$$\mathbf{x}_{k+1} = \mathbf{x}_k - \alpha_k \mathbf{M}_k (\nabla f)(\mathbf{x}_k), \quad \mathbf{M}_k = \left( (\nabla^2 f)(\mathbf{x}_k) + \varepsilon_k \mathbf{E} \right)^{-1} \quad (2.75)$$

mit einem geeigneten positiven Parameter  $\varepsilon_k$  verwendet werden. Die Iterationsvorschrift (2.75) geht für  $\varepsilon_k = 0$  in die Newton-Methode gemäß (2.73) und für sehr große  $\varepsilon_k$  in die Gradientenmethode gemäß (2.51) über. Eine geeignete Wahl von  $\varepsilon_k$  erweist sich jedoch als nicht sehr einfach. Typischerweise wird  $\varepsilon_k$  beginnend bei einem Startwert  $\varepsilon_k > 0$  sukzessive erhöht, bis die Matrix  $(\nabla^2 f)(\mathbf{x}_k) + \varepsilon_k \mathbf{E}$  positiv definit ist. Dies kann sowohl basierend auf den Eigenwerten als auch beispielsweise über die *Cholesky-Faktorisierung* überprüft werden. Für die Cholesky-Faktorisierung gilt nämlich, dass eine Matrix  $\mathbf{M}$  genau dann positiv definit ist, wenn sich die Matrix in der Form  $\mathbf{M} = \mathbf{G}\mathbf{G}^T$  faktorisieren lässt, wobei  $\mathbf{G}$  eine untere Dreiecksmatrix mit positiven Diagonaleinträgen ist.

**Aufgabe 2.8.** Zeigen Sie, dass die Newton-Methode für quadratische Kostenfunktionen

$$\min_{\mathbf{x} \in \mathbb{R}^n} \frac{1}{2} \mathbf{x}^T \mathbf{Q} \mathbf{x} - \mathbf{b}^T \mathbf{x} \quad (2.76)$$

mit der positiv definiten Matrix  $\mathbf{Q}$  unabhängig vom Startpunkt  $\mathbf{x}_0$  innerhalb von nur einem Iterationsschritt konvergiert.

Die Vor- und Nachteile der Newton-Methode können wie folgt zusammengefasst werden:

- + Konvergenzordnung von 2, wenn die Hessematrix  $(\nabla^2 f)(\mathbf{x}_k)$  positiv definit ist, was zumindest in der Nähe des Minimums  $\mathbf{x}^*$  der Fall ist
- + Konvergenzverhalten besser als jenes der konjugierten Gradientenmethode
- außerhalb einer hinreichend kleinen Umgebung um das Minimum ist  $(\nabla^2 f)(\mathbf{x}_k)$  im Allgemeinen nicht positiv definit
- Berechnungsaufwand  $\mathcal{O}(n^2)$  für die benötigte Hessematrix  $(\nabla^2 f)(\mathbf{x}_k)$ , Berechnungsaufwand  $\mathcal{O}(n^3)$  für die Suchrichtung  $\mathbf{s}_k$

### 2.3.2.4 Quasi-Newton-Methode

Einer der Hauptnachteile der Newton-Methode liegt in der aufwändigen Berechnung der Hessematrix  $(\nabla^2 f)(\mathbf{x}_k)$ . Aus diesem Grund versucht man bei der Quasi-Newton-Methode die *inverse* Hessematrix iterativ zu bestimmen. Für das Weitere sei angenommen, dass die Kostenfunktion  $f \in C^2$  ist und für die Punkte  $\mathbf{x}_{k+1}$  und  $\mathbf{x}_k$  gilt  $\mathbf{g}_{k+1} = (\nabla f)(\mathbf{x}_{k+1})$  und  $\mathbf{g}_k = (\nabla f)(\mathbf{x}_k)$ . Aus einer Taylorreihenentwicklung folgt die Näherung

$$\mathbf{g}_{k+1} - \mathbf{g}_k \approx (\nabla^2 f)(\mathbf{x}_k) \mathbf{p}_k \quad (2.77)$$

mit  $\mathbf{p}_k = \mathbf{x}_{k+1} - \mathbf{x}_k$ . Nimmt man nun an, dass die Hessematrix  $(\nabla^2 f)(\mathbf{x}_k) = \mathbf{K}$  konstant ist, dann gilt

$$\mathbf{q}_k = \mathbf{g}_{k+1} - \mathbf{g}_k = \mathbf{K} \mathbf{p}_k . \quad (2.78)$$

Wenn  $n$  linear unabhängige Vektoren  $\mathbf{p}_0, \mathbf{p}_1, \dots, \mathbf{p}_{n-1}$  mit den zugehörigen  $\mathbf{q}_j$ ,  $j = 0, \dots, n-1$  zur Verfügung stehen, dann lässt sich die Hessematrix in der Form

$$\mathbf{K} = \begin{bmatrix} \mathbf{q}_0 & \mathbf{q}_1 & \dots & \mathbf{q}_{n-1} \end{bmatrix} \begin{bmatrix} \mathbf{p}_0 & \mathbf{p}_1 & \dots & \mathbf{p}_{n-1} \end{bmatrix}^{-1} \quad (2.79)$$

berechnen. Das Ziel ist es nun, unter der Annahme einer konstanten Hessematrix  $\mathbf{K}$  in  $n$  Iterationsschritten die inverse Hessematrix  $\mathbf{K}^{-1}$  iterativ in der Form

$$\mathbf{H}_{k+1} \mathbf{q}_j = \mathbf{p}_j, \quad j = 0, \dots, k \quad (2.80)$$

zu konstruieren, dass  $\mathbf{H}_n = \mathbf{K}^{-1}$  gilt. Diese iterative Konstruktion kann auf unterschiedliche Art und Weise erfolgen. Eine mögliche Variante wird im Folgenden beschrieben. Da die Hessematrix und ihre Inverse symmetrisch sind, ist es naheliegend, auch eine symmetrische Matrix für die Rekursion

$$\mathbf{H}_{k+1} = \mathbf{H}_k + \gamma_k \mathbf{z}_k \mathbf{z}_k^T \quad (2.81)$$

anzusetzen. Das dyadische Produkt  $\mathbf{z}_k \mathbf{z}_k^T$  erhält die Symmetrie und hat höchstens den Rang 1, weshalb diese Korrektur auch als *Rang 1 Korrektur* bezeichnet wird. Setzt man (2.81) in (2.80) ein, so erhält man für  $j = k$

$$\mathbf{p}_k = \mathbf{H}_{k+1} \mathbf{q}_k = \mathbf{H}_k \mathbf{q}_k + \gamma_k \mathbf{z}_k \mathbf{z}_k^T \mathbf{q}_k . \quad (2.82)$$

Mit

$$(\mathbf{p}_k - \mathbf{H}_k \mathbf{q}_k)(\mathbf{p}_k - \mathbf{H}_k \mathbf{q}_k)^T = \gamma_k^2 \mathbf{z}_k \mathbf{z}_k^T \mathbf{q}_k \left( \mathbf{z}_k \mathbf{z}_k^T \mathbf{q}_k \right)^T = \gamma_k^2 \mathbf{z}_k \underbrace{\mathbf{z}_k^T \mathbf{q}_k \mathbf{q}_k^T \mathbf{z}_k}_{(\mathbf{z}_k^T \mathbf{q}_k)^2} \mathbf{z}_k^T \quad (2.83)$$

lässt sich (2.81) in der Form

$$\mathbf{H}_{k+1} = \mathbf{H}_k + \frac{(\mathbf{p}_k - \mathbf{H}_k \mathbf{q}_k)(\mathbf{p}_k - \mathbf{H}_k \mathbf{q}_k)^\top}{\gamma_k (\mathbf{z}_k^\top \mathbf{q}_k)^2} \quad (2.84)$$

anschreiben. Bildet man von (2.82) das Skalarprodukt mit  $\mathbf{q}_k$

$$\mathbf{q}_k^\top \mathbf{p}_k = \mathbf{q}_k^\top \mathbf{H}_k \mathbf{q}_k + \gamma_k (\mathbf{z}_k^\top \mathbf{q}_k)^2, \quad (2.85)$$

dann kann (2.84) wie folgt

$$\mathbf{H}_{k+1} = \mathbf{H}_k + \frac{(\mathbf{p}_k - \mathbf{H}_k \mathbf{q}_k)(\mathbf{p}_k - \mathbf{H}_k \mathbf{q}_k)^\top}{\mathbf{q}_k^\top (\mathbf{p}_k - \mathbf{H}_k \mathbf{q}_k)} \quad (2.86)$$

geschrieben werden. Damit lässt sich folgender Satz formulieren.

**Satz 2.12 (Quasi-Newton-Methode — Rang 1 Korrektur).** *Angenommen  $\mathbf{K}$  ist eine konstante symmetrische Matrix und  $\mathbf{p}_0, \mathbf{p}_1, \dots, \mathbf{p}_k$  sind linear unabhängige Vektoren. Mit  $\mathbf{q}_j = \mathbf{K} \mathbf{p}_j$ ,  $j = 0, \dots, k$  gilt für jede symmetrische Startmatrix  $\mathbf{H}_0$  und die Iterationsvorschrift*

$$\mathbf{H}_{j+1} = \mathbf{H}_j + \frac{(\mathbf{p}_j - \mathbf{H}_j \mathbf{q}_j)(\mathbf{p}_j - \mathbf{H}_j \mathbf{q}_j)^\top}{\mathbf{q}_j^\top (\mathbf{p}_j - \mathbf{H}_j \mathbf{q}_j)} \quad (2.87)$$

die Beziehung

$$\mathbf{p}_j = \mathbf{H}_{k+1} \mathbf{q}_j, \quad j = 0, \dots, k. \quad (2.88)$$

**Aufgabe 2.9.** Beweisen Sie Satz 2.12 mit vollständiger Induktion. **Hinweis:** Dieser Beweis ist z. B. in [2.3] skizziert.

Ein zentraler Nachteil der Rang 1 Korrektur ist, dass die positive Definitheit von  $\mathbf{H}_{k+1}$  nur gesichert ist, wenn  $\mathbf{q}_k^\top (\mathbf{p}_k - \mathbf{H}_k \mathbf{q}_k) > 0$  gilt. Aus diesem Grund wurden weitere iterative Korrekturformeln für  $\mathbf{H}_k$  entwickelt. Beispiele dafür sind die Korrekturformel nach Davidon-Fletcher-Powell (DFP)

$$\mathbf{H}_{k+1} = \mathbf{H}_k + \frac{\mathbf{p}_k \mathbf{p}_k^\top}{\mathbf{q}_k^\top \mathbf{p}_k} - \frac{\mathbf{H}_k \mathbf{q}_k \mathbf{q}_k^\top \mathbf{H}_k}{\mathbf{q}_k^\top \mathbf{H}_k \mathbf{q}_k} \quad (2.89)$$

und die etwas häufiger verwendete Korrekturformel nach Broyden-Fletcher-Goldfarb-Shanno (BFGS)

$$\mathbf{H}_{k+1} = \left( \mathbf{E} - \frac{\mathbf{p}_k \mathbf{q}_k^\top}{\mathbf{q}_k^\top \mathbf{p}_k} \right) \mathbf{H}_k \left( \mathbf{E} - \frac{\mathbf{q}_k \mathbf{p}_k^\top}{\mathbf{q}_k^\top \mathbf{p}_k} \right) + \frac{\mathbf{p}_k \mathbf{p}_k^\top}{\mathbf{q}_k^\top \mathbf{p}_k}. \quad (2.90)$$

Beide werden als *Rang 2 Korrekturformeln* bezeichnet, da die aktuelle Approximation der inversen Hessematrix jeweils durch eine Matrix mit Rang 2 korrigiert wird. Linearkombinationen der obigen beiden Formeln in der Art  $\mathbf{H}_{k+1} = \phi \mathbf{H}_{k+1}^{\text{DFP}} + (1 - \phi) \mathbf{H}_{k+1}^{\text{BFGS}}$  mit  $\phi \in (0, 1)$  können ebenfalls verwendet werden. Alle so erhaltenen Rang 2 Korrekturformeln

bilden die sogenannte *Broyden Familie*. Diese Korrekturformeln erhalten natürlich die Symmetrie von  $\mathbf{H}_k$ . Ferner lässt sich zeigen (siehe [2.1, 2.4, 2.5]), dass sie die positive Definitheit von  $\mathbf{H}_k$  erhalten, wenn

$$\mathbf{q}_k^T \mathbf{p}_k > 0 \quad (2.91)$$

erfüllt ist.

Basierend auf der aktuellen Schätzung  $\mathbf{H}_k$  der inversen Hessematrix wird bei der Quasi-Newton-Methode die Suchrichtung in der Form

$$\mathbf{s}_k = -\mathbf{H}_k(\nabla f)(\mathbf{x}_k) \quad (2.92)$$

gewählt. Man beachte, dass hierfür, anders als bei der Newton-Methode (siehe (2.72)), lediglich die Kenntnis des Gradienten  $(\nabla f)(\mathbf{x}_k)$  und eine Matrixmultiplikation von Nöten sind. Der Algorithmus der Quasi-Newton-Methode ist unter Verwendung der BFGS-Korrekturformel in Tabelle 2.3 zusammengefasst.

---

<b>Initialisierung:</b>	$\mathbf{H}_0$	(Startwert, positiv definite Matrix)
	$k = 0$	(Startindex)
	$\mathbf{x}_0$	(Startlösung)
	$\mathbf{g}_0 = (\nabla f)(\mathbf{x}_0)$	(Gradient an der Stelle $\mathbf{x}_0$ )
	$\varepsilon_f, \varepsilon_x$	(Schwellwerte für Abbruchkriterien)

**repeat**

Schritt 1: Berechne die Suchrichtung  $\mathbf{s}_k = -\mathbf{H}_k \mathbf{g}_k$

Schritt 2: Löse die Minimierungsaufgabe  $\min_{\alpha_k \geq 0} f(\mathbf{x}_k + \alpha_k \mathbf{s}_k)$

Schritt 3:  $\mathbf{x}_{k+1} = \mathbf{x}_k + \alpha_k \mathbf{s}_k$

$$\mathbf{p}_k = \mathbf{x}_{k+1} - \mathbf{x}_k = \alpha_k \mathbf{s}_k$$

$$\mathbf{g}_{k+1} = (\nabla f)(\mathbf{x}_{k+1})$$

$$\mathbf{q}_k = \mathbf{g}_{k+1} - \mathbf{g}_k$$

Schritt 4: BFGS-Korrekturformel

$$\mathbf{H}_{k+1} = \left( \mathbf{E} - \frac{\mathbf{p}_k \mathbf{q}_k^T}{\mathbf{q}_k^T \mathbf{p}_k} \right) \mathbf{H}_k \left( \mathbf{E} - \frac{\mathbf{q}_k \mathbf{p}_k^T}{\mathbf{q}_k^T \mathbf{p}_k} \right) + \frac{\mathbf{p}_k \mathbf{p}_k^T}{\mathbf{q}_k^T \mathbf{p}_k}$$

**until**  $\|\mathbf{x}_{k+1} - \mathbf{x}_k\| \leq \varepsilon_x$  **or**  $|f(\mathbf{x}_{k+1}) - f(\mathbf{x}_k)| \leq \varepsilon_f$

---

Tabelle 2.3: Quasi-Newton-Methode mit der BFGS-Korrekturformel.

Die Erfüllung der Bedingung (2.91) lässt sich durch eine geeignete Wahl der Schrittweite  $\alpha_k$  in  $\mathbf{x}_{k+1} = \mathbf{x}_k + \alpha_k \mathbf{s}_k$  sicherstellen. Genügt die Wahl der Schrittweite beispielsweise der Wolfe-Bedingung gemäß Abschnitt 2.3.1.3, so ist (2.91) automatisch erfüllt. Um dies zu sehen, beachte man, dass aus (2.33)

$$g'(\alpha_k) = \mathbf{g}_{k+1}^T \mathbf{s}_k \geq \varepsilon_2 g'(0) = \varepsilon_2 \mathbf{g}_k^T \mathbf{s}_k \quad (2.93)$$

mit  $0 < \varepsilon_2 < 1$  folgt. Daraus erhält man

$$(\mathbf{g}_{k+1} - \mathbf{g}_k)^T \mathbf{s}_k = \mathbf{q}_k^T \mathbf{s}_k \geq \underbrace{(\varepsilon_2 - 1) \mathbf{g}_k^T \mathbf{s}_k}_{> 0}, \quad (2.94)$$

wobei die rechte Seite dieser Ungleichung (abseits des optimalen Punktes  $\mathbf{x}^*$ ) strikt positiv sein muss, da  $\mathbf{s}_k$  eine Abstiegsrichtung ist. Mit  $\mathbf{p}_k = \mathbf{x}_{k+1} - \mathbf{x}_k = \alpha_k \mathbf{s}_k$  folgt schließlich aus (2.94), dass

$$\mathbf{q}_k^T \mathbf{p}_k = \alpha_k \mathbf{q}_k^T \mathbf{s}_k \geq \alpha_k (\varepsilon_2 - 1) \mathbf{g}_k^T \mathbf{s}_k > 0 \quad (2.95)$$

gilt und (2.91) somit erfüllt ist.

Für unbeschränkte nichtlineare Optimierungsprobleme mit konvexer Kostenfunktion  $f(\mathbf{x})$  konvergiert die Quasi-Newton-Methode mit superlinearer Konvergenzordnung. Für das unbeschränkte quadratische Optimierungsproblem (2.54) konvergiert die Quasi-Newton-Methode nach spätestens  $n$  Iterationsschritten. Konvergiert die Methode in diesem Fall genau nach  $n$  Iterationsschritten, so kann gezeigt werden (siehe [2.1]), dass  $\mathbf{H}_n = (\nabla^2 f)^{-1}(\mathbf{x}^*) = \mathbf{Q}^{-1}$ , d. h. der Algorithmus liefert die exakte inverse Hessematrix.

Die Vor- und Nachteile der Quasi-Newton-Methode können wie folgt zusammengefasst werden:

- + einfaches Verfahren mit moderatem Rechenaufwand
- + nur der Gradient  $(\nabla f)(\mathbf{x}_k)$  (Berechnungsaufwand  $\mathcal{O}(n)$ ) wird benötigt
- + konvergiert bei quadratischen Optimierungsproblemen nach spätestens  $n$  Iterationsschritten
- + generell superlineares Konvergenzverhalten
- Matrix  $\mathbf{H}_k$  muss gespeichert werden (Speicherplatzbedarf  $\mathcal{O}(n^2)$ )

### 2.3.2.5 Gauss-Newton-Methode

Bei der Gauss-Newton-Methode wird die aufwändige Berechnung der Hessematrix  $(\nabla^2 f)(\mathbf{x}_k)$  durch eine weniger rechenintensive näherungsweise Berechnung ersetzt. Diese Optimierungsmethode ist nur anwendbar, wenn die Kostenfunktion  $f(\mathbf{x})$  die Struktur

$$f(\mathbf{x}) = \frac{1}{2} \|\mathbf{r}(\mathbf{x})\|_2^2 = \frac{1}{2} \sum_{i=1}^m r_i^2(\mathbf{x}) \quad (2.96)$$

mit einer beliebigen Funktion  $\mathbf{r} : \mathbb{R}^n \rightarrow \mathbb{R}^m$  besitzt. Die Komponenten  $r_i(\mathbf{x})$  von  $\mathbf{r}(\mathbf{x})$  sollen  $r_i \in C^2$  erfüllen. Die exakte Hessematrix von  $f(\mathbf{x})$  lautet in diesem Fall

$$\left(\nabla^2 f\right)(\mathbf{x}) = (\nabla \mathbf{r})(\mathbf{x})(\nabla \mathbf{r})^T(\mathbf{x}) + \sum_{i=1}^m r_i(\mathbf{x}) \left(\nabla^2 r_i\right)(\mathbf{x}), \quad (2.97)$$

wobei die Spalten der Jacobi-Matrix  $(\nabla \mathbf{r})(\mathbf{x}) \in \mathbb{R}^{n \times m}$  die Gradienten  $(\nabla r_i)(\mathbf{x})$  enthalten. Bei der Gauss-Newton-Methode wird der zweite Summand in (2.97) vernachlässigt, so dass sich die Näherung

$$\left(\nabla^2 f\right)(\mathbf{x}) \approx (\nabla \mathbf{r})(\mathbf{x})(\nabla \mathbf{r})^T(\mathbf{x}) \quad (2.98)$$

ergibt. Der damit verbundene Näherungsfehler ist also klein, wenn  $r_i(\mathbf{x})$  oder  $(\nabla^2 r_i)(\mathbf{x}) \forall i = 1, \dots, m$  betragslich kleine Werte annimmt. Ist z. B. bekannt, dass  $f(\mathbf{x}^*) = 0$ , so folgt daraus  $r_i(\mathbf{x}^*) = 0 \forall i = 1, \dots, m$ . Der Fall  $(\nabla^2 r_i)(\mathbf{x}) = \mathbf{0} \forall i = 1, \dots, m$  tritt ein, wenn  $\mathbf{r}(\mathbf{x})$  affin in  $\mathbf{x}$  ist (vgl. Aufgabe 2.10).

Unter Verwendung von des Gradienten  $(\nabla f)(\mathbf{x}) = (\nabla \mathbf{r})(\mathbf{x})\mathbf{r}(\mathbf{x})$  und der Approximation (2.98) ergibt sich der Gauss-Newton-Schritt

$$\begin{aligned} \mathbf{s}_k &= -((\nabla \mathbf{r})(\mathbf{x}_k)(\nabla \mathbf{r})^T(\mathbf{x}_k))^{-1}(\nabla f)(\mathbf{x}_k) \\ &= -((\nabla \mathbf{r})(\mathbf{x}_k)(\nabla \mathbf{r})^T(\mathbf{x}_k))^{-1}(\nabla \mathbf{r})(\mathbf{x}_k)\mathbf{r}(\mathbf{x}_k) \end{aligned} \quad (2.99)$$

Die Iterationsvorschrift der Gauss-Newton-Methode lautet damit

$$\mathbf{x}_{k+1} = \mathbf{x}_k - ((\nabla \mathbf{r})(\mathbf{x}_k)(\nabla \mathbf{r})^T(\mathbf{x}_k))^{-1}(\nabla \mathbf{r})(\mathbf{x}_k)\mathbf{r}(\mathbf{x}_k) . \quad (2.100)$$

Für die praktische Anwendung der Methode führt man noch eine geeignete Schrittweite  $\alpha_k$  ein, so dass die Iterationsvorschrift

$$\mathbf{x}_{k+1} = \mathbf{x}_k - \alpha_k((\nabla \mathbf{r})(\mathbf{x}_k)(\nabla \mathbf{r})^T(\mathbf{x}_k))^{-1}(\nabla \mathbf{r})(\mathbf{x}_k)\mathbf{r}(\mathbf{x}_k) \quad (2.101)$$

lautet und die Abstiegsbedingung (2.22) erfüllt ist. Das Verfahren wird dann häufig *gedämpfte Gauss-Newton-Methode* genannt und  $\alpha_k$  wird als *Dämpfungsparameter* bezeichnet. Gelegentlich wird die Einschränkung  $\alpha_k \leq 1$  verwendet.

Man beachte, dass die Berechnung von  $\mathbf{s}_k$  gemäß (2.99) keine tatsächliche Inversion von  $(\nabla \mathbf{r})(\mathbf{x}_k)(\nabla \mathbf{r})^T(\mathbf{x}_k)$  erfordert. Die Wahl der Suchrichtung gemäß (2.99) ist nur sinnvoll, wenn die  $n \times n$  Matrix  $(\nabla \mathbf{r})(\mathbf{x})(\nabla \mathbf{r})^T(\mathbf{x})$  positiv definit ist. Ist sie singulär (oder zumindest schlecht konditioniert), kann ähnlich wie bei der Newton-Methode die alternative Iterationsvorschrift

$$\mathbf{x}_{k+1} = \mathbf{x}_k - \alpha_k \mathbf{M}_k (\nabla f)(\mathbf{x}_k) , \quad \mathbf{M}_k = \left( (\nabla \mathbf{r})(\mathbf{x}_k)(\nabla \mathbf{r})^T(\mathbf{x}_k) + \varepsilon_k \mathbf{E} \right)^{-1} \quad (2.102)$$

mit einem geeigneten positiven Parameter  $\varepsilon_k$  verwendet werden. Das Verfahren wird dann als *Levenberg-Marquardt-Methode* bezeichnet. Die Iterationsvorschrift (2.102) geht für  $\varepsilon_k = 0$  in die Gauss-Newton-Methode gemäß (2.101) und für sehr große  $\varepsilon_k$  in die Gradientenmethode gemäß (2.51) über.

Der nachfolgende Satz liefert eine Aussage über die Konvergenzordnung der Gauss-Newton-Methode. Sein Beweis ist z. B. in [2.6] zu finden.

**Satz 2.13 (Konvergenzordnung der Gauss-Newton-Methode).** *Gegeben sei die Kostenfunktion  $f(\mathbf{x}) = \frac{1}{2} \|\mathbf{r}(\mathbf{x})\|_2^2$  mit  $\mathbf{r} \in C^2$  definiert im  $\mathbb{R}^n$  und dem lokalen Minimum  $\mathbf{x}^*$ . Wenn die Jacobi-Matrix  $(\nabla \mathbf{r})(\mathbf{x}^*)$  zeilenregulär ist,  $\sum_{i=1}^m r_i(\mathbf{x}^*)(\nabla^2 r_i)(\mathbf{x}^*) = \mathbf{0}$  gilt und der Anfangswert  $\mathbf{x}_0$  in einer hinreichend nahen Umgebung des Minimums liegt, dann konvergiert die Gauss-Newton-Iteration*

$$\mathbf{x}_{k+1} = \mathbf{x}_k - ((\nabla \mathbf{r})(\mathbf{x}_k)(\nabla \mathbf{r})^T(\mathbf{x}_k))^{-1}(\nabla \mathbf{r})(\mathbf{x}_k)\mathbf{r}(\mathbf{x}_k) \quad (2.103)$$

*mit der Konvergenzordnung 2 gegen das Minimum  $\mathbf{x}^*$ .*

Ist der Term  $\sum_{i=1}^m r_i(\mathbf{x}^*)(\nabla^2 r_i)(\mathbf{x}^*) \neq \mathbf{0}$  aber klein gegenüber  $(\nabla \mathbf{r})(\mathbf{x}_k)(\nabla \mathbf{r})^T(\mathbf{x}_k)$ , so kann zumindest lineares Konvergenzverhalten nachgewiesen werden. Wenn jedoch  $\sum_{i=1}^m r_i(\mathbf{x}^*)(\nabla^2 r_i)(\mathbf{x}^*)$  betragslich zu große Werte annimmt, kann es sein, dass die Gauss-Newton-Methode nicht konvergiert [2.6].

**Aufgabe 2.10.** Zeigen Sie, dass die Gauss-Newton-Methode für die Optimierungsaufgabe

$$\min_{\mathbf{x} \in \mathbb{R}^n} \frac{1}{2} \|\mathbf{r}(\mathbf{x})\|_2^2 \quad \text{mit} \quad \mathbf{r}(\mathbf{x}) = \mathbf{A}\mathbf{x} + \mathbf{b} \quad (2.104)$$

und einer spaltenregulären Matrix  $\mathbf{A}$  (dies impliziert  $m \geq n$ ) unabhängig vom Startpunkt  $\mathbf{x}_0$  innerhalb von nur einem Iterationsschritt zum optimalen Punkt  $\mathbf{x}^* = -(\mathbf{A}^T \mathbf{A})^{-1} \mathbf{A}^T \mathbf{b}$  konvergiert.

Kostenfunktionen wie in der Optimierungsaufgabe (2.104) treten z. B. bei der Methode der kleinsten Fehlerquadrate mit parametrisch linearem Modell (lineare Regressionsanalyse) [2.7] auf. Kostenfunktionen der Art (2.96) treten z. B. bei der Methode der kleinsten Fehlerquadrate mit parametrisch nichtlinearem Modell (nichtlineare Regressionsanalyse) [2.8] auf.

**Aufgabe 2.11.** Zeigen Sie, dass die iterative Lösung der nichtlinearen Gleichung

$$\mathbf{r}(\mathbf{x}) = \mathbf{0} \quad (2.105)$$

mit  $\mathbf{r} : \mathbb{R}^n \rightarrow \mathbb{R}^n$  gemäß dem Newton-Raphson-Verfahren genau dem Gauss-Newton-Verfahren angewandt auf die Optimierungsaufgabe

$$\min_{\mathbf{x} \in \mathbb{R}^n} \frac{1}{2} \|\mathbf{r}(\mathbf{x})\|_2^2 \quad (2.106)$$

entspricht.

Die Vor- und Nachteile der Gauss-Newton-Methode können wie folgt zusammengefasst werden:

- + keine zweiten Ableitungen nötig, nur der Gradient  $(\nabla \mathbf{r})(\mathbf{x}_k)$  wird benötigt
- + konvergiert im ersten Iterationsschritt, wenn  $\mathbf{r}(\mathbf{x})$  affin in  $\mathbf{x}$  ist
- + unter bestimmten Voraussetzungen kann quadratisches Konvergenzverhalten erreicht werden
- Konvergenz hängt vom jeweiligen Problem ab und ist nicht garantiert
- Berechnungsaufwand  $\mathcal{O}(mn)$  für die Jacobi-Matrix  $(\nabla \mathbf{r})(\mathbf{x}_k)$ , Berechnungsaufwand  $\mathcal{O}(n^3)$  für die Suchrichtung  $\mathbf{s}_k$

## 2.4 Methode der Vertrauensbereiche

Bei den Liniensuchverfahren wird eine geeignete Abstiegsrichtung (Suchrichtung)  $\mathbf{s}_k$  (beispielsweise der *negative Gradient* an der Stelle  $\mathbf{x}_k$  gemäß (2.34) bei der Gradientenmethode

oder die *Newton-Richtung* gemäß (2.72) bei der Newton-Methode) gewählt und anschließend wird über das skalare Optimierungsproblem (2.23) die (optimale) Schrittweite  $\alpha_k > 0$  in diese Abstiegsrichtung bestimmt. Bei der Methode der Vertrauensbereiche (englisch: *trust region method*) wird die zu minimierende Kostenfunktion  $f(\mathbf{x})$  in der Umgebung von  $\mathbf{x}_k$  durch eine quadratische Ansatzfunktion  $m_k$  in der Form

$$f(\mathbf{x}_k + \mathbf{s}_k) \approx m_k(\mathbf{s}_k) = f(\mathbf{x}_k) + \mathbf{s}_k^T (\nabla f)(\mathbf{x}_k) + \frac{1}{2} \mathbf{s}_k^T \mathbf{B}_k \mathbf{s}_k \quad (2.107)$$

mit einer geeigneten symmetrischen Matrix  $\mathbf{B}_k$  approximiert. Der Approximationsfehler der quadratischen Ansatzfunktion ist in der Größenordnung von  $\|\mathbf{s}_k\|^2$  und wenn  $\mathbf{B}_k$  mit der Hessematrix  $(\nabla^2 f)(\mathbf{x}_k)$  übereinstimmt sogar von  $\|\mathbf{s}_k\|^3$ . Der *Vertrauensbereich* wird durch den Parameter  $\Delta_k$  charakterisiert und beschreibt jene Umgebung um den Punkt  $\mathbf{x}_k$ , in der die Kostenfunktion  $f(\mathbf{x}_k + \mathbf{s}_k)$  hinreichend genau durch die quadratische Ansatzfunktion  $m_k(\mathbf{s}_k)$  beschrieben wird. Dabei wird in jedem Iterationsschritt das Optimierungsproblem

$$\begin{aligned} \min_{\mathbf{s}_k \in \mathbb{R}^n} \quad & m_k(\mathbf{s}_k) = f(\mathbf{x}_k) + \mathbf{s}_k^T (\nabla f)(\mathbf{x}_k) + \frac{1}{2} \mathbf{s}_k^T \mathbf{B}_k \mathbf{s}_k \\ \text{u.B.v.} \quad & \|\mathbf{s}_k\| \leq \Delta_k \end{aligned} \quad (2.108)$$

für ein geeignetes  $\Delta_k > 0$  gelöst. Man beachte, dass (im Gegensatz zu den meisten Liniensuchverfahren) die Abstiegsrichtung und die Schrittweite *gleichzeitig* bestimmt werden.

Ein wesentlicher Entwurfsfreiheitsgrad dieser Methode liegt nun in der Wahl von  $\Delta_k$ . Dazu wird in jedem Iterationsschritt *die Übereinstimmung der quadratischen Ansatzfunktion  $m_k$  mit der Kostenfunktion  $f$*  überprüft, indem das Verhältnis

$$\rho_k(\mathbf{s}_k) = \frac{f(\mathbf{x}_k) - f(\mathbf{x}_k + \mathbf{s}_k)}{m_k(\mathbf{0}) - m_k(\mathbf{s}_k)} \quad (2.109)$$

berechnet wird. Der Zählerterm in (2.109) beschreibt die *tatsächliche Reduktion* der Kostenfunktion während der Nennerterm die *prädiizierte Reduktion* wiedergibt. Der Nennerterm von  $\rho_k(\mathbf{s}_k)$  ist stets größer gleich Null, da  $\mathbf{s}_k$  die Funktion  $m_k$  gemäß (2.108) innerhalb des Vertrauensbereiches minimiert und der Punkt  $\mathbf{s}_k = \mathbf{0}$  im Vertrauensbereich liegt. Ist nun  $\rho_k(\mathbf{s}_k) < 0$ , so bedeutet dies, dass der Wert der Kostenfunktion am nächsten Iterationspunkt  $f(\mathbf{x}_k + \mathbf{s}_k)$  größer als am vorigen Iterationspunkt  $f(\mathbf{x}_k)$  ist, weshalb dieser Iterationsschritt verworfen und der Vertrauensbereich verkleinert werden muss. Andererseits kann bei  $\rho_k(\mathbf{s}_k) \approx 1$  der Vertrauensbereich vergrößert werden, da die Kostenfunktion  $f(\mathbf{x}_k)$  in diesem Fall gut von der Ansatzfunktion beschrieben wird. Für den Fall, dass  $\rho_k(\mathbf{s}_k)$  positiv und deutlich kleiner als 1 ist, wird der Vertrauensbereich im nächsten Schritt verkleinert.

Der Algorithmus der Methode der Vertrauensbereiche ist in Tabelle 2.4 aufgelistet. Man beachte, dass hier  $\bar{\Delta}$  die obere Grenze des zulässigen Vertrauensbereiches beschreibt und dass eine Vergrößerung des Vertrauensbereiches im nächsten Iterationsschritt nur dann stattfindet, wenn  $\mathbf{s}_k$  durch die Grenze des Vertrauensbereiches beschränkt wurde (Bedingung  $\|\mathbf{s}_k\| = \Delta_k$ ).

Vorschläge für die in Tabelle 2.4 nicht dargestellte Iterationsvorschrift der Matrix  $\mathbf{B}_k$  werden z. B. in [2.1, 2.3, 2.4] beschrieben. Eine Möglichkeit ist die Berechnung von  $\mathbf{B}_k$  in

---

<b>Initialisierung:</b>	$\bar{\Delta}, \Delta_0 \in (0, \bar{\Delta})$	(Vertrauensbereich: Grenz- & Startwert)
	$\eta \in [0, \frac{1}{4})$	(Parameter)
	$k \leftarrow 0$	(Iterationsindex)
	$\varepsilon_x, \varepsilon_f$	(Abbruchkriterien)
<b>repeat</b>		
	$m_k(\mathbf{s}_k)$ nach (2.107)	(Modell)
	$\mathbf{s}_k$ Lösung von (2.108)	(evtl. approximativ gelöst)
	$\rho_k$ nach (2.109)	(Modellgüte)
	<b>if</b> $\rho_k < \frac{1}{4}$	
	$\Delta_{k+1} \leftarrow \frac{1}{4}\Delta_k$	(Reduktion)
	<b>else if</b> $\rho_k > \frac{3}{4}$ <b>and</b> $\ \mathbf{s}_k\  = \Delta_k$	
	$\Delta_{k+1} \leftarrow \min\{2\Delta_k, \bar{\Delta}\}$	(Vergrößerung)
	<b>else</b>	
	$\Delta_{k+1} \leftarrow \Delta_k$	
	<b>end if</b>	
	<b>if</b> $\rho_k > \eta$	
	$\mathbf{x}_{k+1} \leftarrow \mathbf{x}_k + \mathbf{s}_k$	(nächster Schritt)
	$\mathbf{B}_{k+1} \leftarrow \mathbf{B}_k + \dots$	(Aktualisierung der Ansatzfunktion)
	<b>else</b>	
	$\mathbf{x}_{k+1} \leftarrow \mathbf{x}_k$	(Schritt mit $\Delta_{k+1} < \Delta_k$ wiederholen)
	<b>end if</b>	
	$k \leftarrow k + 1$	
<b>until</b> $\ \mathbf{x}_k - \mathbf{x}_{k-1}\  \leq \varepsilon_x$ <b>or</b> $ f(\mathbf{x}_k) - f(\mathbf{x}_{k-1})  \leq \varepsilon_f$		

---

Tabelle 2.4: Methode der Vertrauensbereiche.

der Form (2.98), d. h.  $\mathbf{B}_k = (\nabla \mathbf{r})(\mathbf{x}_k)(\nabla \mathbf{r})^T(\mathbf{x}_k)$  gemäß der Gauss-Newton-Methode. Es kann gezeigt werden, dass dann die Methode der Vertrauensbereiche mit der Levenberg-Marquardt-Methode übereinstimmt [2.6].

## 2.5 Direkte Suchverfahren

Die bisher betrachteten sogenannten *ableitungsbehafteten* Lösungsverfahren verwenden den Gradienten  $(\nabla f)(\mathbf{x}_k)$  (und mitunter die Hessematrix  $(\nabla^2 f)(\mathbf{x}_k)$ ), um mittels einer geeigneten Iterationsvorschrift einen neuen Punkt  $\mathbf{x}_{k+1}$  zu bestimmen. Es soll dabei eine hinreichend gute Reduktion der Kostenfunktion  $f(\mathbf{x}_{k+1}) < f(\mathbf{x}_k)$  erreicht werden.

Allerdings sind in manchen praktischen Fällen die dazu erforderlichen Ableitungen nicht

verfügbar oder mit vertretbarem Aufwand berechenbar, da das betrachtete Problem *zu komplex* oder *nicht stetig differenzierbar* ist. Abhilfe verschaffen in diesem Fall sogenannte *direkte* oder *ableitungsfreie Suchverfahren*, die mit Hilfe von Stichproben eine Reihe von Funktionswerten berechnen, um daraus einen neuen Iterationspunkt  $\mathbf{x}_{k+1}$  zu bestimmen.

Ein bekanntes und gleichzeitig einfaches Verfahren in der nichtlinearen Optimierung ist das *Simplex-Verfahren* nach *Nelder* und *Mead*. Dieses Verfahren unterscheidet sich grundsätzlich vom Simplex-Algorithmus in der *Linearen Programmierung* und sollte nicht mit ihm verwechselt werden.

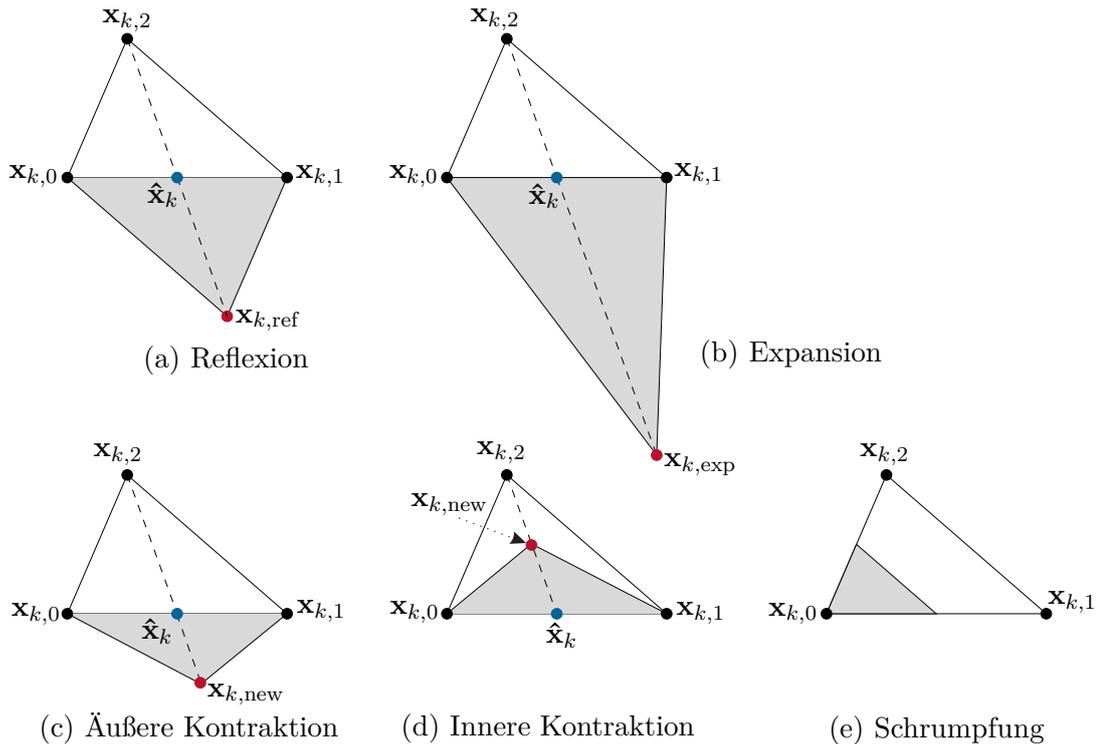


Abbildung 2.8: Operationen des Simplex-Verfahrens nach Nelder und Mead für  $\beta = 1$ ,  $\gamma = 1$  und  $\theta = 1/2$ .

Der Algorithmus basiert im Wesentlichen auf der Iteration eines sogenannten *Simplex* im  $n$ -dimensionalen Parameterraum. Unter einem Simplex versteht man in diesem Zusammenhang jene konvexe Hülle, die von  $n + 1$  Punkten  $\mathbf{x}_{k,i}$ ,  $i = 0, \dots, n$  zum Iterationsschritt  $k$  im  $n$ -dimensionalen Suchraum aufgespannt wird (für  $n = 1$  ist dies eine Linie, für  $n = 2$  ein Dreieck, etc.). Man bezeichne nun im Weiteren mit  $\mathbf{x}_{k,\min}$  und  $\mathbf{x}_{k,\max}$  jene Punkte  $\mathbf{x}_{k,i}$ ,  $i = 0, \dots, n$ , die die Kostenfunktion  $f$  möglichst klein bzw. groß machen, d. h. es gilt

$$\begin{aligned} f(\mathbf{x}_{k,\min}) &= \min_{i=0,\dots,n} f(\mathbf{x}_{k,i}) \\ f(\mathbf{x}_{k,\max}) &= \max_{i=0,\dots,n} f(\mathbf{x}_{k,i}) . \end{aligned} \quad (2.110)$$

Der *Schwerpunkt* oder Mittelpunkt des Simplex  $\hat{\mathbf{x}}_k$  gebildet durch alle Punkte außer  $\mathbf{x}_{k,\max}$

errechnet sich zu

$$\hat{\mathbf{x}}_k = \frac{1}{n} \left( \sum_{i=0}^n \mathbf{x}_{k,i} - \mathbf{x}_{k,\max} \right). \quad (2.111)$$

Der Algorithmus beruht nun auf der Idee, den Punkt  $\mathbf{x}_{k,\max}$  im Simplex durch einen anderen Punkt mit einem niedrigeren Kostenfunktionswert zu ersetzen. Eine wichtige Operationen dabei ist die Berechnung des *Reflexionspunktes*

$$\mathbf{x}_{k,\text{ref}} = \hat{\mathbf{x}}_k + (\hat{\mathbf{x}}_k - \mathbf{x}_{k,\max}), \quad (2.112)$$

der auf einer Geraden durch die Punkte  $\mathbf{x}_{k,\max}$  und  $\hat{\mathbf{x}}_k$  liegt und symmetrisch bezüglich  $\hat{\mathbf{x}}_k$  zu  $\mathbf{x}_{k,\max}$  ist, siehe Abbildung 2.8(a). Abhängig von  $f(\mathbf{x}_{k,\text{ref}})$  im Vergleich zu den Funktionswerten der anderen Punkte des Simplex außer  $\mathbf{x}_{k,\max}$  wird ein neuer Punkt  $\mathbf{x}_{k,\text{new}}$  konstruiert, der im nächsten Iterationsschritt den Punkt  $\mathbf{x}_{k,\max}$  ersetzt. Der Algorithmus ist in seiner Grundfunktion in Tabelle 2.5 aufgelistet und die unterschiedlichen Operationen sind grafisch in Abbildung 2.8 dargestellt. Man beachte, dass die Schrumpfung im Algorithmus von Tabelle 2.5 stets bezüglich des minimierenden Punktes  $\mathbf{x}_{k,\min}$  ausgeführt wird, d. h. in Abbildung 2.8(e) gilt  $\mathbf{x}_{k,1} = \mathbf{x}_{k,\min}$ . Während der Iteration wandert der Simplex in Richtung des Optimums und zieht sich sukzessive zusammen. Allerdings ist die Konvergenz im Allgemeinen nicht garantiert und es kann vorkommen, dass das Simplex-Verfahren zu einem *nicht-optimalen Punkt* konvergiert. In der Praxis führt das Simplex-Verfahren dennoch häufig zu guten Ergebnissen und akzeptablem Konvergenzverhalten.

---

<b>Initialisierung:</b>	$\mathbf{x}_{0,i}, i = 0, \dots, n$ (Startsimplex)	
	$k \leftarrow 0$ (Iterationsindex)	
	$\beta > 0$ (Reflexionskoeffizient, typisch $\beta = 1$ )	
	$\gamma > 0$ (Expansionskoeffizient, typisch $\gamma = 1$ )	
	$\theta \in (0, 1)$ (Kontraktionskoeffizient, typisch $\theta = 1/2$ )	
	$\varepsilon_x, \varepsilon_f$ (Abbruchkriterien)	
<b>repeat</b>		
	$\mathbf{x}_{k,\min}, \mathbf{x}_{k,\max}$ gemäß (2.110)	(Punkte mit min. und max. Kostenfunktionswert)
	$\hat{\mathbf{x}}_k$ gemäß (2.111)	(Schwerpunkt)
	$\mathbf{x}_{k,\text{ref}} = \hat{\mathbf{x}}_k + \beta(\hat{\mathbf{x}}_k - \mathbf{x}_{k,\max})$	(Reflexionsschritt, Abb. 2.8(a))
	<b>if</b> $f(\mathbf{x}_{k,\text{ref}}) < f(\mathbf{x}_{k,\min})$	
	$\mathbf{x}_{k,\text{exp}} = \mathbf{x}_{k,\text{ref}} + \gamma(\mathbf{x}_{k,\text{ref}} - \hat{\mathbf{x}}_k)$	(Expansionsschritt, Abb. 2.8(b))
	<b>if</b> $f(\mathbf{x}_{k,\text{exp}}) < f(\mathbf{x}_{k,\text{ref}})$	
	$\mathbf{x}_{k,\text{new}} = \mathbf{x}_{k,\text{exp}}$	
	<b>else</b>	
	$\mathbf{x}_{k,\text{new}} = \mathbf{x}_{k,\text{ref}}$	(Reflexionspunkt beibehalten)
	<b>end if</b>	
	<b>else if</b> $f(\mathbf{x}_{k,\text{ref}}) > \max_{\mathbf{x}_{k,i} \neq \mathbf{x}_{k,\max}, i=0,\dots,n} f(\mathbf{x}_{k,i})$	
	<b>if</b> $f(\mathbf{x}_{k,\max}) \leq f(\mathbf{x}_{k,\text{ref}})$	
	$\mathbf{x}_{k,\text{new}} = \theta \mathbf{x}_{k,\max} + (1 - \theta) \hat{\mathbf{x}}_k$	(Innere Kontraktion, Abb. 2.8(d))
	<b>else</b>	
	$\mathbf{x}_{k,\text{new}} = \theta \mathbf{x}_{k,\text{ref}} + (1 - \theta) \hat{\mathbf{x}}_k$	(Äußere Kontraktion, Abb. 2.8(c))
	<b>end if</b>	
	<b>else</b>	
	$\mathbf{x}_{k,\text{new}} = \mathbf{x}_{k,\text{ref}}$	(Reflexionspunkt beibehalten)
	<b>end if</b>	
	<b>if</b> $f(\mathbf{x}_{k,\text{new}}) \geq f(\mathbf{x}_{k,\max})$	(ev. bei nichtkonv. Kostenfkt.)
	$\mathbf{x}_{k+1,i} \leftarrow \frac{1}{2}(\mathbf{x}_{k,i} + \mathbf{x}_{k,\min}), i = 0, \dots, n$	(Schrumpfung, Abb. 2.8(e))
	<b>else</b>	
	$\mathbf{x}_{k,\max} \leftarrow \mathbf{x}_{k,\text{new}}$	
	$\mathbf{x}_{k+1,i} \leftarrow \mathbf{x}_{k,i}, i = 0, \dots, n$	
	<b>end if</b>	
	$k \leftarrow k + 1$	
<b>until</b>	$\ \mathbf{x}_k - \mathbf{x}_{k-1}\  \leq \varepsilon_x$ <b>or</b> $ f(\mathbf{x}_k) - f(\mathbf{x}_{k-1})  \leq \varepsilon_f$	

---

Tabelle 2.5: Simplex-Verfahren nach Nelder und Mead.

## 2.6 Beispiel: Rosenbrock's „Bananenfunktion“

Ein bekanntes Beispiel in der Optimierung ist das *Rosenbrock*-Problem

$$\min_{\mathbf{x} \in \mathbb{R}^2} f(\mathbf{x}) \quad \text{mit} \quad f(\mathbf{x}) = 100(x_2 - x_1^2)^2 + (x_1 - 1)^2. \quad (2.113)$$

Abbildung 2.9 zeigt das Profil und die Höhenlinien der Funktion, die auch als *Bananenfunktion* bezeichnet wird. Das Rosenbrock-Problem soll als Beispiel verwendet werden, um die *Konvergenzeigenschaften der behandelten Verfahren* numerisch mit Hilfe von MATLAB zu untersuchen.

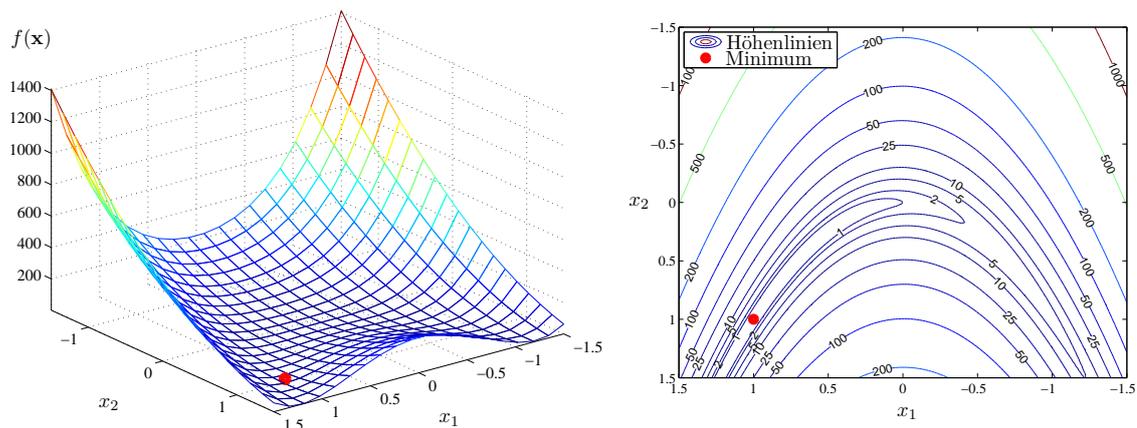


Abbildung 2.9: Profil und Höhenlinien von Rosenbrock's Bananenfunktion.

**Aufgabe 2.12.** Verifizieren Sie, dass der Punkt  $\mathbf{x}^* = [1 \ 1]^T$  ein Minimum darstellt. Ist das Minimum  $\mathbf{x}^*$  global und eindeutig? Sind die Funktion  $f(\mathbf{x})$  und das Optimierungsproblem (2.113) konvex?

Zur Lösung von unbeschränkten Optimierungsproblemen stellt die *Optimization Toolbox* von MATLAB die folgenden Funktionen zur Verfügung

- **fminunc:** Liniensuche: Gradientenverfahren, Quasi-Newton-Verfahren  
Methode der Vertrauensbereiche: Newton-Verfahren
- **fminsearch:** Simplex-Verfahren nach Nelder-Mead.

Eine empfehlenswerte Alternative ist die frei zugängliche MATLAB-Funktion **minFunc** [2.9], die eine große Auswahl an Liniensuchverfahren bietet. Tabelle 2.6 zeigt einige Vergleichsdaten für die numerische Lösung des Rosenbrock-Problems (ausgehend vom Startwert  $\mathbf{x}_0 = [-1 \ -1]^T$ ), die mit Hilfe von **fminunc**, **fminsearch** und **minFunc** berechnet wurden.

Abbildung 2.11 stellt zusätzlich die Iterationsverläufe für die Verfahren dar, die unter **fminunc** und **fminsearch** implementiert sind. In der Code-Auflistung 2.1 am Ende dieses Abschnitts ist der MATLAB-Code für das Rosenbrock-Problem (2.113) angegeben, um zu verdeutlichen, wie die einzelnen Optimierungsverfahren mit **fminunc** und **fminsearch** angesprochen werden können.

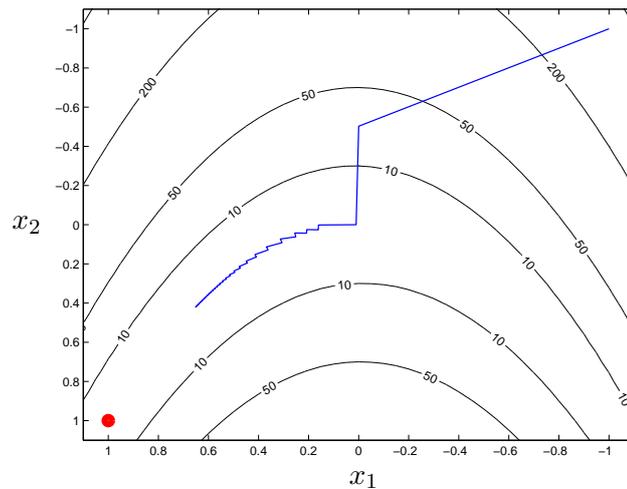


Abbildung 2.10: Darstellung der Iterationen des Gradientenverfahrens.

Verfahren	Funktion	Iter.	$f(\mathbf{x}^*)$	$\ (\nabla f)(\mathbf{x}^*)\ $	Funktionsaufrufe		
					$f(\mathbf{x})$	$(\nabla f)(\mathbf{x})$	$(\nabla^2 f)(\mathbf{x})$
LS: Gradientenverfahren	fminunc	57	0.1232	1.1978	200	200	–
LS: Konj. Gradientenm.	minFunc	28	$6.9 \cdot 10^{-18}$	$9.6 \cdot 10^{-8}$	68	68	–
LS: Newton-Verfahren	minFunc	20	$3.8 \cdot 10^{-16}$	$7.3 \cdot 10^{-7}$	32	32	26
LS: Quasi-Newton (BFGS)	fminunc	23	$5.4 \cdot 10^{-12}$	$9.2 \cdot 10^{-6}$	29	29	–
LS: Gauss-Newton-Verf.	minFunc	4	$2.8 \cdot 10^{-29}$	$1.1 \cdot 10^{-14}$	9	9	–
VB: Newton-Verfahren	fminunc	25	$2.2 \cdot 10^{-18}$	$2.1 \cdot 10^{-8}$	26	26	26
Nelder-Mead Simplex-Verf.	fminsearch	67	$5.3 \cdot 10^{-10}$	–	125	–	–

Tabelle 2.6: Vergleich der numerischen Verfahren für das Rosenbrock-Problem (LS=Liniensuche, VB=Methode der Vertrauensbereiche).

Beim *Gradientenverfahren* fällt die *langsame Konvergenz* auf, weil auch nach dem Erreichen der maximalen Anzahl an Funktionsauswertungen von 200 das Minimum noch immer nicht erreicht ist. In Abbildung 2.10 ist der Iterationsverlauf des Gradientenverfahrens über den Höhenlinien der Rosenbrock-Funktion (2.113) dargestellt. Erkennbar ist, dass sich das Gradientenverfahren an der maximalen Abstiegsrichtung orientiert, die orthogonal zur jeweiligen Höhenlinie verläuft. In Richtung des Minimums werden die Iterationsschritte immer kleiner. Die niedrige Konvergenzgeschwindigkeit wurde bereits in Abbildung 2.6 veranschaulicht und soll in der folgenden Aufgabe näher untersucht werden.

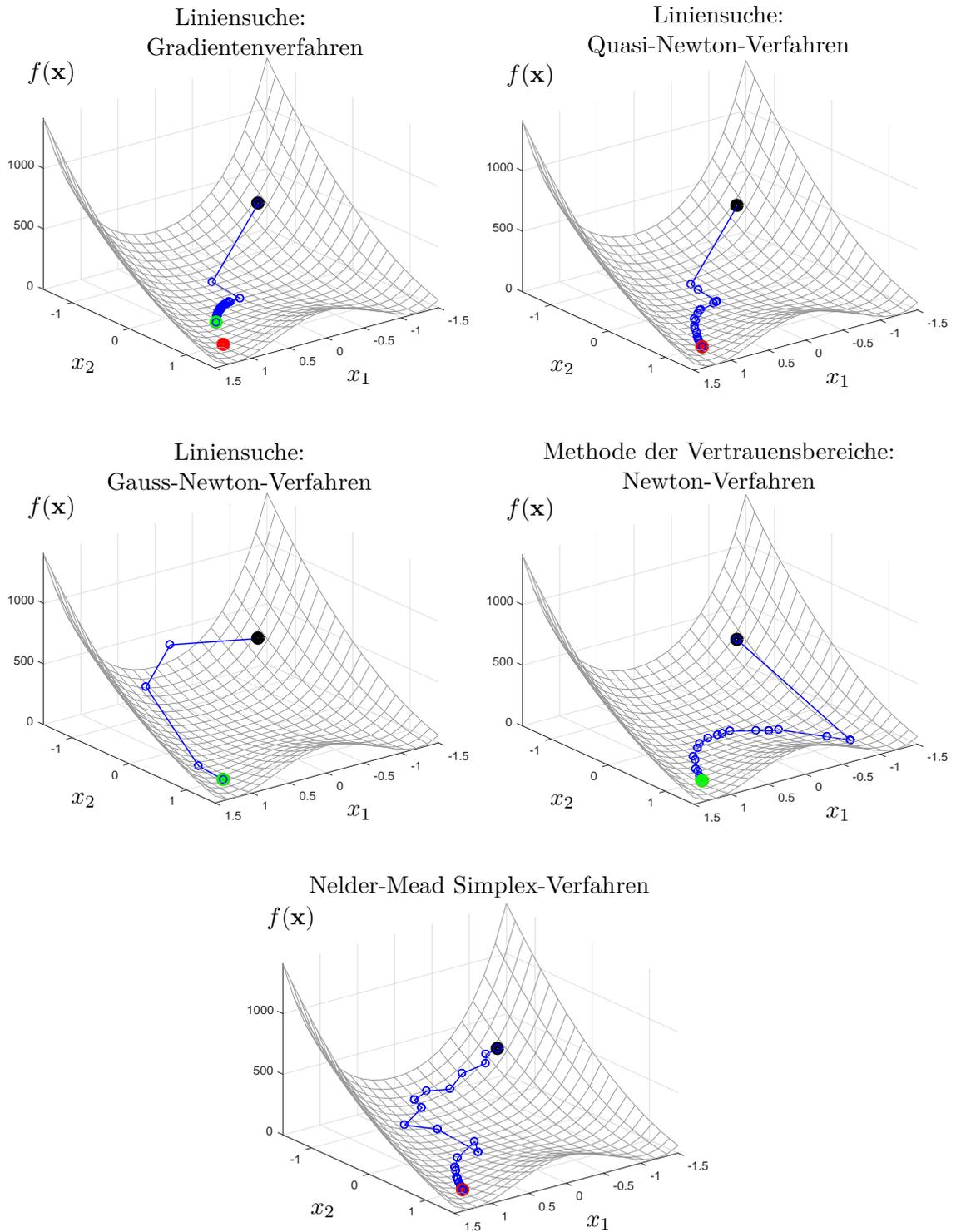


Abbildung 2.11: Rosenbrock's Bananenfunktion: Vergleich der numerischen Verfahren mit fminunc, fminsearch und minFunc.

**Aufgabe 2.13.** Berechnen Sie für das Minimum  $\mathbf{x}^* = [1 \ 1]^T$  des Rosenbrock-Problems (2.113) die Konvergenzrate des Gradientenverfahrens gemäß Satz 2.7.

Das Konvergenzverhalten des *Quasi-Newton-Verfahrens* in Abbildung 2.11 ist wesentlich besser als beim Gradientenverfahren. Das *Newton-Verfahren (Methode der Vertrauensbereiche)* in Abbildung 2.11 startet zunächst in die „falsche“ Richtung, was durch die *quadratische Approximation* (2.107) am Startpunkt  $\mathbf{x}_0 = [-1 \ -1]^T$  zu erklären ist, deren Minimum in der Nähe von  $\mathbf{x} \approx [-1 \ 1]^T$  liegt. Allerdings sind die einzelnen Schritte entlang des Tales der Rosenbrock-Funktion deutlich größer, da das Newton-Verfahren die *Hessematrix explizit verwendet* und nicht auf eine Approximation angewiesen ist wie im Fall des Quasi-Newton-Verfahrens. Das *Gauss-Newton-Verfahren* weist für die Rosenbrock-Funktion ein sehr gutes Konvergenzverhalten auf. Trotz der Approximation der Hessematrix kann auch im Tal der Kostenfunktion mit großer Schrittweite das Minimum gefunden werden. Dieses besonders gute Verhalten ist aber auf die Form der Kostenfunktion zurückzuführen, welche Terme enthält, die affin in  $\mathbf{x}$  sind (vgl. dazu Aufgabe 2.10).

Zusätzlich sind in Tabelle 2.6 und Abbildung 2.11 die Ergebnisse für das *Simplex-Verfahren von Nelder-Mead* angegeben, die mit der MATLAB-Funktion `fminsearch` erzielt wurden. Allerdings bietet die Grafikausgabe unter `fminsearch` nicht die Möglichkeit, die einzelnen Simplexe darzustellen. In der nächsten Aufgabe soll das Simplex-Verfahren deshalb näher untersucht werden, um einen Eindruck von den Simplex-Operationen und der Robustheit des Verfahrens zu erhalten.

**Aufgabe 2.14.** Schreiben Sie eine MATLAB-Funktion, die das Rosenbrock-Problem (2.113) mit Hilfe des Simplex-Verfahrens nach Nelder-Mead numerisch löst (siehe den Algorithmus von Tabelle 2.5). Konstruieren Sie den ersten Simplex mit den Eckpunkten

$$\mathbf{x}_{0,1} = \begin{bmatrix} -1 \\ -1 \end{bmatrix}, \quad \mathbf{x}_{0,2} = \mathbf{x}_{0,1} + s \begin{bmatrix} 1 \\ 0 \end{bmatrix}, \quad \mathbf{x}_{0,3} = \mathbf{x}_{0,1} + s \begin{bmatrix} 0 \\ 1 \end{bmatrix} \quad (2.114)$$

in Abhängigkeit der Seitenlänge  $s = 1$ . Verwenden Sie für die Abbruchbedingung in Tabelle 2.5 die Schranke  $\varepsilon_f = 10^{-9}$  und vergleichen Sie Ihre Ergebnisse mit jenen von `fminsearch` in Tabelle 2.6. Stellen Sie die Simplexe aus den einzelnen Iterationen grafisch dar. Untersuchen Sie die Robustheit und das Konvergenzverhalten des Simplex-Verfahrens für verschiedene Seitenlängen  $s$  des Startsimplex und unterschiedliche Startpunkte  $\mathbf{x}_{0,1}$ .

**Aufgabe 2.15.** Schreiben Sie eine MATLAB-Funktion, die das Rosenbrock-Problem (2.113) mit Hilfe des Newton-Verfahrens (Liniensuche) löst, siehe Abschnitt 2.3.2.3. Verwenden Sie die heuristischen Bedingungen von Abschnitt 2.3.1.3 für die Schrittweitenbestimmung von  $\alpha_k$ . Vergleichen Sie die Konvergenzresultate mit den Werten in Tabelle 2.6. Stellen Sie die einzelnen Iterationen grafisch dar und variieren Sie die Startpunkte  $\mathbf{x}_0$ .

Listing 2.1: MATLAB-Code für das Rosenbrock-Problem (fminunc, fminsearch, minFunc).

```

function Xopt = rosenbrock_problem(Xinit,methodQ)
% -----
% Xinit: Startpunkt
% methodQ: 1 - fminunc: Liniensuche: Gradientenverfahren
%           2 - fminunc: Liniensuche: Quasi-Newton
%           3 - fminunc: Methode der Vertrauensbereiche: Newton-Verfahren
%           4 - fminsearch: Nelder-Mead Simplex-Verfahren
%           5 - minFunc: CG-Verfahren
%           6 - minFunc: Newton-Verfahren
%           7 - minFunc: Gauss-Newton-Verfahren
global old
old = [Xinit; rosenbrock(Xinit)];

% Optionen für fminunc:
opt_fminu = optimoptions('fminunc','Display','iter','PlotFcns',@plot_iterates);
% Optionen für fminsearch:
opt_fmins = optimset('Display','iter','PlotFcns',@plot_iterates);
% Optionen für minFunc:
opt=struct('display','iter','outputFcn',@plot_iterates);

% Generelles: GradObj gibt an, dass die Funktion zur Berechnung der
% Kostenfunktion auch den Wert des Gradienten retourniert,
% ist auch Hessian auf on, muss zusätzlich die Hessematrix zurückgegeben
% werden
switch methodQ
case 1 % fminunc: Liniensuche: Gradientenverfahren
    opt_fminu = optimoptions(opt_fminu,'Algorithm','quasi-newton','HessUpdate','steepdesc',...
        'GradObj','on');
    [Xopt,fopt,exitflag,output] = fminunc(@rosenbrock,Xinit,opt_fminu);
case 2 % fminunc: Liniensuche: Quasi-Newton
    opt_fminu = optimoptions(opt_fminu,'Algorithm','quasi-newton','HessUpdate','bfgs',...
        'GradObj','on');
    [Xopt,fopt,exitflag,output] = fminunc(@rosenbrock,Xinit,opt_fminu);
case 3 % fminunc: Methode der Vertrauensbereiche: Newton-Verfahren
    opt_fminu = optimoptions(opt_fminu,'Algorithm','trust-region','Hessian','on','GradObj','on');
    [Xopt,fopt,exitflag,output] = fminunc(@rosenbrock,Xinit,opt_fminu);
case 4 % fminsearch: Nelder-Mead Simplex-Verfahren
    [Xopt,fopt,exitflag,output] = fminsearch(@rosenbrock,Xinit,opt_fmins);
case 5 % minFunc: CG-Verfahren
    figure
    opt.Method = 'cg';
    [Xopt,fopt,exitflag,output] = minFunc(@rosenbrock,Xinit,opt);
case 6 % minFunc: Newton-Verfahren
    figure
    opt.Method = 'newton';
    [Xopt,fopt,exitflag,output] = minFunc(@rosenbrock,Xinit,opt);
case 7 % minFunc: Gauss-Newton-Verfahren
    figure
    opt.Method = 'newton';
    [Xopt,fopt,exitflag,output] = minFunc(@rosenbrock_gauss_newton,Xinit,opt);
end

function [f, grad, H] = rosenbrock(x)
% -----
grad = {}; H = {};
f = 100*(x(2)-x(1)^2)^2 + (x(1)-1)^2; % Rosenbrock-Funktion
if nargin>1 % falls Gradient angefordert wird
    grad = [ -400*(x(2)-x(1)^2)*x(1)+2*(x(1)-1);
            200*(x(2)-x(1)^2) ];
end
if nargin>2 % falls Hessematrix angefordert wird
    H = [ -400*(x(2)-3*x(1)^2)+2, -400*x(1);
         200*(x(2)-x(1)^2), 0 ];
end

```

```

        -400*x(1),          200 ];
end

function [f, grad, H]=rosenbrock_gauss_newton(x)
% -----
% Rosenbrock-Funktion wird umformuliert als Norm einer
% vektorwertigen Funktion r(x):
% f(x)=100*(x2-x1^2)^2+(x1-1)^2 = 1/2*norm(r(x))^2
% mit
% r1(x)=sqrt(200)*(x2-x1^2)
% r2(x)=sqrt(2)*(x1-1)
r=[sqrt(200)*(x(2)-x(1)^2);
  sqrt(2)*(x(1)-1)];
f=1/2*norm(r)^2;
if nargout>1
    grad_r=[-sqrt(200)*2*x(1),sqrt(2);
            sqrt(200),0];
    grad=grad_r*r;
end
if nargout>2
    H=grad_r*grad_r';
end

function stop = plot_iterates(x,info,state)
% -----
global old
f = rosenbrock(x);
switch state
    case 'init' % Grafische Ausgabe:
                  % Initialisierung
        plot_surface(x,f);
    case 'iter' % Iterationen
        plot3([old(1),x(1)], [old(2),x(2)], [old(3),f], 'b-o', 'LineWidth',1);
    case 'done' % nach letzter Iteration
        plot3(x(1),x(2),f, 'go', 'LineWidth',5);
end
stop = false; % kein Abbruchkriterium
old = [x;f];

function plot_surface(x,f)
% -----
[X1,X2] = meshgrid(-1.5:0.15:1.5); % 3D-Profil von
F = 100*(X2-X1.^2).^2 + (X1-1).^2; % Rosenbrock-Funktion
h = surf(X1,X2,F, 'EdgeColor',0.6*[1,1,1], 'FaceColor', 'none');
hold on; axis tight;
plot3(x(1),x(2),f, 'ko', 'LineWidth',5); % Startpunkt
plot3(1,1,0, 'ro', 'LineWidth',5); % optimale Lösung
xlabel('x_1'); ylabel('x_2'); zlabel('f')
set(gcf, 'ToolBar', 'figure'); % Aktivieren der Menüleiste (Zoom, etc.)
set(gca, 'Xdir', 'reverse', 'Ydir', 'reverse');
set(gca, 'clipping', 'off');

```

## 2.7 Literatur

- [2.1] D. P. Bertsekas, *Nonlinear Programming*, 2. Aufl. Athena Scientific, 1999.
- [2.2] D. G. Luenberger und Y. Ye, *Linear and Nonlinear Programming*, 3. Aufl., Ser. International Series in Operations Research & Management Science. Springer, 2008, Bd. 116.
- [2.3] J. Nocedal und S. J. Wright, *Numerical Optimization*, 2. Aufl., Ser. Springer Series in Operations Research and Financial Engineering. Springer, 2006.
- [2.4] I. Griva, S. Nash und A. Sofer, *Linear and Nonlinear Optimization*, 2. Aufl. Society for Industrial und Applied Mathematics, 2009.
- [2.5] R. Fletcher, *Practical methods of optimization*, 2. Aufl. John Wiley & Sons, 1987.
- [2.6] J. Dennis und R. Schnabel, *Numerical Methods for Unconstrained Optimization and Nonlinear Equations*, Ser. Classics in Applied Mathematics. Society for Industrial und Applied Mathematics, 1996.
- [2.7] W. Kemmetmüller, *Skriptum zur VO Regelungssysteme 1 (WS 2016/2017)*, Institut für Automatisierungs- und Regelungstechnik, TU Wien, 2016. Adresse: <http://www.acin.tuwien.ac.at/lehre/master/regelungssysteme-1/>.
- [2.8] M. Papageorgiou, M. Leibold und M. Buss, *Optimierung: Statische, dynamische, stochastische Verfahren für die Anwendung*, 4. Aufl. Springer, 2015.
- [2.9] M. Schmidt, „minFunc“, abrufbar unter <http://www.cs.ubc.ca/~schmidt/Software/minFunc.html>, University of British Columbia, Vancouver, 2005, (besucht am 28.09.2016).
- [2.10] S. Boyd und L. Vandenberghe, *Convex Optimization*. Cambridge University Press, 2004.
- [2.11] C. T. Kelley, *Iterative Methods for Optimization*. Society for Industrial und Applied Mathematics, 1999.
- [2.12] B. C. Chachuat, „Nonlinear and Dynamic Optimization: From Theory to Practice“, abrufbar unter <http://infoscience.epfl.ch/record/111939>, Laboratoire d'Automatique, École Polytechnique Fédérale de Lausanne, 2007, (besucht am 28.09.2017).
- [2.13] H. T. Jongen, K. Meer und E. Triesch, *Optimization Theory*. Kluwer Academic Publishers, 2004.

### 3 Statische Optimierung mit Beschränkungen

Den nachfolgenden Betrachtungen liegt das statische Optimierungsproblem mit Gleichungs- und Ungleichungsbeschränkungen gemäß (1.1) in der Form

$$\min_{\mathbf{x} \in \mathbb{R}^n} f(\mathbf{x}) \quad \text{Kostenfunktion} \quad (3.1a)$$

$$\text{u.B.v. } g_i(\mathbf{x}) = 0, \quad i = 1, \dots, p \quad \text{Gleichungsbeschränkungen} \quad (3.1b)$$

$$h_i(\mathbf{x}) \leq 0, \quad i = 1, \dots, q \quad \text{Ungleichungsbeschränkungen} \quad (3.1c)$$

mit  $p \leq n$ , der stetigen Funktion  $f(\mathbf{x})$  und den stetig differenzierbaren Funktionen  $g_i(\mathbf{x})$ ,  $i = 1, \dots, p$  und  $h_i(\mathbf{x})$ ,  $i = 1, \dots, q$  zu Grunde. Fasst man alle Gleichungs- und Ungleichungsbeschränkungen in Vektoren der Form  $\mathbf{g}(\mathbf{x}) = [g_1(\mathbf{x}) \ \dots \ g_p(\mathbf{x})]^T$  und  $\mathbf{h}(\mathbf{x}) = [h_1(\mathbf{x}) \ \dots \ h_q(\mathbf{x})]^T$  zusammen, so kann das Optimierungsproblem (3.1) in der Form

$$\min_{\mathbf{x} \in \mathbb{R}^n} f(\mathbf{x}) \quad (3.2a)$$

$$\text{u.B.v. } \mathbf{g}(\mathbf{x}) = \mathbf{0} \quad (3.2b)$$

$$\mathbf{h}(\mathbf{x}) \leq \mathbf{0} \quad (3.2c)$$

dargestellt werden. Noch kompakter lässt sich dies äquivalent in der Form (1.3)

$$\min_{\mathbf{x} \in \mathcal{X}} f(\mathbf{x}) \quad (3.3a)$$

mit der zulässigen Menge

$$\mathcal{X} = \{ \mathbf{x} \in \mathbb{R}^n \mid \mathbf{g}(\mathbf{x}) = \mathbf{0}, \mathbf{h}(\mathbf{x}) \leq \mathbf{0} \} \quad (3.3b)$$

anschreiben. Jedes  $\mathbf{x} \in \mathcal{X}$  wird als zulässiger Punkt bezeichnet.

Wie im Verlauf dieses Kapitels noch klarer werden sollte, ist die Berücksichtigung von allgemeinen nichtlinearen Ungleichungsbeschränkungen (3.2c) zumeist schwieriger als die Berücksichtigung von Gleichungsbeschränkungen (3.2b). Eine Möglichkeit, (3.2) äquivalent ohne nichtlineare Ungleichungsbeschränkungen zu formulieren, ist die Verwendung sogenannter *Schlupfvariablen* (englisch: *slack variables*). Dies führt auf die Formulierung

$$\min_{\substack{\mathbf{x} \in \mathbb{R}^n, \\ \mathbf{x}_s \in \mathbb{R}^q}} f(\mathbf{x}) \quad (3.4a)$$

$$\text{u.B.v. } \mathbf{g}(\mathbf{x}) = \mathbf{0} \quad (3.4b)$$

$$\mathbf{h}(\mathbf{x}) + \mathbf{x}_s = \mathbf{0} \quad (3.4c)$$

$$\mathbf{x}_s \geq \mathbf{0}. \quad (3.4d)$$

Hierbei wurde (3.2c) durch die zusätzlichen Gleichungsbeschränkungen (3.4c) und die (wesentlich einfacheren) Ungleichungsbeschränkungen (3.4d) ersetzt. Die Schlupfvariablen  $\mathbf{x}_s$  stellen zusätzliche Optimierungsvariablen dar, d. h. die Dimension des Optimierungsproblems erhöht sich um  $q$ .

## 3.1 Optimalitätsbedingungen

### 3.1.1 Optimalitätsbedingungen basierend auf zulässigen Richtungen

Um Bedingungen für ein lokales Minimum  $\mathbf{x}^*$  der Optimierungsaufgabe (3.3) zu formulieren, wird zunächst der Begriff einer *zulässigen Richtung* definiert.

**Definition 3.1 (Zulässige Richtung).** Der Vektor  $\mathbf{d} \in \mathbb{R}^n$  wird als zulässige Richtung am Punkt  $\mathbf{x} \in \mathcal{X}$  bezeichnet, wenn ein  $\bar{\alpha} > 0$  so existiert, dass  $\mathbf{x} + \alpha \mathbf{d} \in \mathcal{X}$  für alle  $\alpha \in [0, \bar{\alpha}]$  gilt.

Eine zulässige Richtung muss nicht an jedem zulässigen Punkt  $\mathbf{x} \in \mathcal{X}$  existieren. Als Beispiel dafür betrachte man die in Abbildung 3.1a gezeigte zulässige Menge  $\mathcal{X} \in \mathbb{R}^2$ , welche z. B. durch eine nichtlineare Gleichungsnebenbedingung definiert werden kann. In diesem Fall existiert an keinem Punkt  $\mathbf{x} \in \mathcal{X}$  eine zulässige Richtung im Sinne der Definition 3.1.

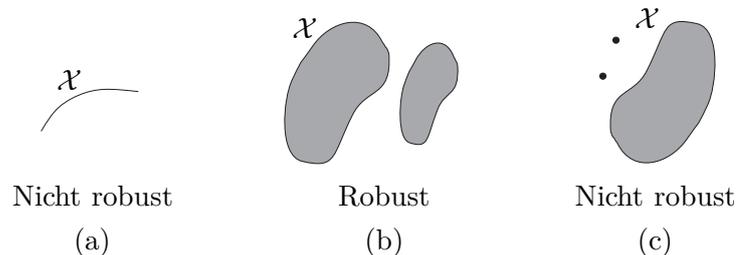


Abbildung 3.1: Robuste und nicht robuste Mengen im  $\mathbb{R}^2$ .

Wenn  $\mathcal{X}$  eine *robuste Menge* ist, dann ist gesichert, dass an jedem zulässigen Punkt  $\mathbf{x} \in \mathcal{X}$  eine zulässige Richtung existiert. Eine Menge  $\mathcal{X}$  ist genau dann *robust*, wenn jeder Punkt am Rand von  $\mathcal{X}$  über das (nichtleere) Innere der Menge  $\mathcal{X}$  erreicht werden kann. Abbildung 3.1 zeigt Beispiele für robuste und nicht robuste Mengen im  $\mathbb{R}^2$ . Eine robuste Menge kann offen oder abgeschlossen sein.

Methodisch kann bei der numerischen Suche von optimalen Lösungen auf einer robusten Menge  $\mathcal{X}$  ähnlich vorgegangen werden, wie bei den in Abschnitt 2.3 beschriebenen Liniensuchverfahren für unbeschränkte statische Optimierungsprobleme. Dabei wird iterativ entlang von Abstiegsrichtungen  $\mathbf{d}$ , die gleichzeitig zulässigen Richtungen sind, gesucht und durch Eingrenzung der Schrittweite  $\alpha$  darauf geachtet, dass das zulässige Gebiet  $\mathcal{X}$  nicht verlassen wird, d. h. dass die Ungleichungsbeschränkungen (3.2c) stets eingehalten werden. Dies ist besonders einfach, wenn die Formulierung (3.4) mit Schlupfvariablen verwendet wird, denn dann müssen nur die einfachen Ungleichungen (3.4d) erfüllt werden.

**Satz 3.1** (Notwendige Optimalitätsbedingungen erster Ordnung). *Es sei  $\mathcal{X} \subseteq \mathbb{R}^n$  die zulässige Menge des Optimierungsproblems (3.3) und  $f \in C^1$  eine Funktion definiert auf  $\mathcal{X}$ . Wenn  $\mathbf{x}^*$  ein lokales Minimum von  $f$  auf  $\mathcal{X}$  ist, dann gilt für jede zulässige Richtung  $\mathbf{d}$  am Punkt  $\mathbf{x}^*$  die Ungleichungsbedingung*

$$\mathbf{d}^T(\nabla f)(\mathbf{x}^*) \geq 0. \quad (3.5)$$

Liegt  $\mathbf{x}^*$  im Inneren von  $\mathcal{X}$ , dann gilt zusätzlich

$$(\nabla f)(\mathbf{x}^*) = \mathbf{0}. \quad (3.6)$$

*Beweis.* Wenn  $\mathbf{d}$  eine zulässige Richtung am Punkt  $\mathbf{x}^*$  ist, dann existiert ein  $\bar{\alpha} > 0$  so, dass  $\mathbf{x}^* + \alpha\mathbf{d} \in \mathcal{X}$  für alle  $\alpha \in [0, \bar{\alpha}]$ . Nun definiert man für  $0 \leq \alpha \leq \bar{\alpha}$  die Funktion  $g(\alpha) = f(\mathbf{x}^* + \alpha\mathbf{d})$ . Sie muss am Punkt  $\alpha = 0$  ein lokales Minimum besitzen. Entwickelt man  $g(\alpha)$  um den Punkt  $\alpha = 0$  in eine Taylorreihe und bricht diese nach dem linearen Glied ab, so erhält man

$$g(\alpha) = g(0) + g'(0)\alpha + \mathcal{O}(\alpha^2) \quad (3.7)$$

mit  $g'(0) = \mathbf{d}^T(\nabla f)(\mathbf{x}^*)$ . Wäre nun  $g'(0) < 0$ , dann würde für ein hinreichend kleines  $\alpha > 0$  gelten  $g(\alpha) - g(0) < 0$ , was im Widerspruch zu der Annahme steht, dass  $\alpha = 0$  bzw.  $\mathbf{x}^*$  ein Minimum ist. Daher muss gelten  $g'(0) = \mathbf{d}^T(\nabla f)(\mathbf{x}^*) \geq 0$ .

Wenn  $\mathbf{x}^*$  im Inneren von  $\mathcal{X}$  liegt, dann ist jede Richtung  $\mathbf{d} \in \mathbb{R}^n$  am Punkt  $\mathbf{x}^*$  zulässig. Damit aber  $g'(0) = \mathbf{d}^T(\nabla f)(\mathbf{x}^*) \geq 0$  für beliebige  $\mathbf{d} \in \mathbb{R}^n$  erfüllt ist, muss  $(\nabla f)(\mathbf{x}^*) = \mathbf{0}$  gelten.  $\square$

Natürlich impliziert Satz 3.1 die notwendige Optimalitätsbedingung für unbeschränkte statische Optimierungsprobleme gemäß Satz 2.1. Dann gilt  $\mathcal{X} = \mathbb{R}^n$  und  $\mathbf{x}^*$  liegt immer im Inneren von  $\mathcal{X}$ .

*Beispiel 3.1.* Man betrachte die Optimierungsaufgabe

$$\min_{\mathbf{x} \in \mathcal{X}} f(x_1, x_2) = x_1^2 - 2x_1 + x_2 + x_1x_2 \quad (3.8)$$

mit der zulässigen Menge

$$\mathcal{X} = \{ \mathbf{x} \in \mathbb{R}^2 \mid x_1 \geq 0, x_2 \geq 0 \}. \quad (3.9)$$

Das Problem hat an der Stelle  $\mathbf{x}^* = [1 \ 0]^T$  ein globales Minimum. Wertet man den Gradienten an der Stelle  $\mathbf{x}^*$  aus, so erhält man

$$\frac{\partial}{\partial x_1} f(\mathbf{x}^*) = 2x_1^* - 2 + x_2^* = 0 \quad (3.10a)$$

$$\frac{\partial}{\partial x_2} f(\mathbf{x}^*) = 1 + x_1^* = 2. \quad (3.10b)$$

In diesem Fall liegt  $\mathbf{x}^*$  am Rand von  $\mathcal{X}$  und der Gradient  $(\nabla f)(\mathbf{x}^*)$  verschwindet nicht. Es ist aber die notwendige Bedingung (3.5) für alle zulässigen Richtungen  $\mathbf{d}$  erfüllt. Die zweite Komponente von  $\mathbf{d}$  muss wegen der Definition von  $\mathcal{X}$  gemäß (3.9) größer gleich Null sein.

**Satz 3.2 (Notwendige Optimalitätsbedingungen zweiter Ordnung).** *Es sei  $\mathcal{X} \subseteq \mathbb{R}^n$  die zulässige Menge des Optimierungsproblems (3.3) und  $f \in C^2$  eine Funktion definiert auf  $\mathcal{X}$ . Wenn  $\mathbf{x}^*$  ein lokales Minimum von  $f$  auf  $\mathcal{X}$  ist, dann gelten für jede zulässige Richtung  $\mathbf{d}$  am Punkt  $\mathbf{x}^*$  die Bedingungen*

$$(a) \quad \mathbf{d}^T (\nabla f)(\mathbf{x}^*) \geq 0 \quad (3.11a)$$

$$(b) \quad \text{wenn } \mathbf{d}^T (\nabla f)(\mathbf{x}^*) = 0, \text{ dann } \mathbf{d}^T (\nabla^2 f)(\mathbf{x}^*) \mathbf{d} \geq 0. \quad (3.11b)$$

Liegt  $\mathbf{x}^*$  im Inneren von  $\mathcal{X}$ , dann gelten zusätzlich die Bedingungen

$$(a) \quad (\nabla f)(\mathbf{x}^*) = \mathbf{0} \quad (3.12a)$$

$$(b) \quad (\nabla^2 f)(\mathbf{x}^*) \text{ ist positiv semi-definit.} \quad (3.12b)$$

**Aufgabe 3.1.** Beweisen Sie Satz 3.2. **Hinweis:** Orientieren Sie sich dabei am Beweis von Satz 3.1.

**Aufgabe 3.2.** Es wird die Optimierungsaufgabe

$$\min_{\mathbf{x} \in \mathcal{X}} f(x_1, x_2) = x_1^3 - x_1^2 x_2 + 2x_2^2 \quad (3.13)$$

mit der zulässigen Menge

$$\mathcal{X} = \{ \mathbf{x} \in \mathbb{R}^2 \mid x_1 \geq 0, x_2 \geq 0 \} \quad (3.14)$$

betrachtet. Zeigen Sie, dass der Punkt  $\mathbf{x}^* = [6 \ 9]^T$  zwar die Optimalitätsbedingung erster Ordnung erfüllt, aber trotzdem kein lokales Minimum beschreibt.

Wenn die Funktion  $f(\mathbf{x})$  und die zulässige Menge  $\mathcal{X}$  konvex sind, dann sind die notwendigen Optimalitätsbedingungen erster Ordnung gemäß Satz 3.1 auch *hinreichend*. Um dies zu sehen, beachte man, dass mit beliebigem  $\mathbf{y} \in \mathcal{X}$ ,  $\mathbf{y} - \mathbf{x}^*$  eine zulässige Richtung am Punkt  $\mathbf{x}^*$  darstellt und wegen der Konvexität von  $f(\mathbf{x})$  die Ungleichung

$$f(\mathbf{y}) \geq f(\mathbf{x}^*) + \underbrace{(\mathbf{y} - \mathbf{x}^*)^T (\nabla f)(\mathbf{x}^*)}_{\geq 0} \geq f(\mathbf{x}^*) \quad (3.15)$$

gilt. Die Sätze 3.1 und 3.2 liefern nur Aussagen zu lokalen Minima. Wenn die Funktion  $f(\mathbf{x})$  konvex oder strikt konvex ist, dann können nachfolgende Bedingungen für globale Minima angegeben werden.

**Satz 3.3** (Globale Minima einer konvexen Funktion). *Es sei  $f(\mathbf{x})$  eine konvexe Funktion auf einer konvexen Menge  $\mathcal{X}$ . Die Menge aller Minima  $\mathcal{G} = \arg \min\{f(\mathbf{x}) \mid \mathbf{x} \in \mathcal{X}\}$  ist konvex. Jedes lokale Minimum  $\mathbf{x}^* \in \mathcal{G}$  von  $f$  ist auch ein globales Minimum. Ist  $f(\mathbf{x})$  strikt konvex, so ist  $\mathbf{x}^*$  ein striktes globales Minimum.*

Der Beweis dieses Satzes erfolgt analog zum Beweis von Satz 2.4

### 3.1.2 Optimalitätsbedingungen mit Lagrange-Multiplikatoren

In diesem Abschnitt werden Optimalitätsbedingungen mit Hilfe von Lagrange-Multiplikatoren formuliert. Sie setzen den Gradienten  $(\nabla f)(\mathbf{x})$  der Kostenfunktion in Beziehung zu den Jacobi-Matrizen  $(\nabla \mathbf{g})(\mathbf{x})$  und  $(\nabla \mathbf{h})(\mathbf{x})$  der Beschränkungen. An optimalen Punkten  $\mathbf{x}^*$  muss  $(\nabla f)(\mathbf{x}^*)$  im Bild dieser Jacobi-Matrizen liegen.

Man bezeichnet eine Ungleichungsbeschränkung  $h_i(\mathbf{x}) \leq 0$  als *aktiv* an einem zulässigen Punkt  $\mathbf{x}$ , wenn  $h_i(\mathbf{x}) = 0$  und als *inaktiv*, falls  $h_i(\mathbf{x}) < 0$ . Eine Gleichungsbeschränkung  $g_i(\mathbf{x}) = 0$  ist demnach aktiv an jedem zulässigen Punkt  $\mathbf{x}$ . Inaktive Ungleichungsbeschränkungen an einem zulässigen Punkt  $\mathbf{x}$  haben keinen Einfluss auf die Lösung der Optimierungsaufgabe in einer hinreichend kleinen Umgebung von  $\mathbf{x}$ . Würde man also die Menge der (am optimalen Punkt) aktiven Ungleichungsbeschränkungen kennen, so könnte man die inaktiven Ungleichungsbeschränkungen vernachlässigen und die aktiven Ungleichungsbeschränkungen durch Gleichungsbeschränkungen ersetzen. Deshalb soll in einem ersten Schritt das Optimierungsproblem (3.2) mit reinen Gleichungsbeschränkungen  $\mathbf{g}(\mathbf{x}) = \mathbf{0}$  betrachtet werden.

#### 3.1.2.1 Reine Gleichungsbeschränkungen

Treten ausschließlich Gleichungsbeschränkungen auf, so reduziert sich das Optimierungsproblem (3.2) auf die Form

$$\min_{\mathbf{x} \in \mathbb{R}^n} f(\mathbf{x}) \quad (3.16a)$$

$$\text{u.B.v. } \mathbf{g}(\mathbf{x}) = \mathbf{0} . \quad (3.16b)$$

Für die *zulässige Menge* gilt dann

$$\mathcal{X} = \{\mathbf{x} \in \mathbb{R}^n \mid g_i(\mathbf{x}) = 0, i = 1, \dots, p\} . \quad (3.17)$$

**Definition 3.2** (Regulärer Punkt bei Gleichungsbeschränkungen, LICQ). Ein zulässiger Punkt  $\mathbf{x} \in \mathcal{X}$  der Optimierungsaufgabe (3.2) mit ausschließlich Gleichungsbeschränkungen ( $q = 0$ ) ist *regulär*, wenn die Gradientenvektoren  $(\nabla g_i)(\mathbf{x})$ ,  $i = 1, \dots, p$  linear unabhängig sind. D. h. die Bedingung

$$\text{rang}((\nabla \mathbf{g})(\mathbf{x})) = \text{rang}\left(\begin{bmatrix} (\nabla g_1)(\mathbf{x}) & (\nabla g_2)(\mathbf{x}) & \dots & (\nabla g_p)(\mathbf{x}) \end{bmatrix}\right) = p, \quad (3.18)$$

welche im Englischen auch als *linear independence constraint qualification (LICQ)* bekannt ist, muss an diesem Punkt erfüllt sein.

Folglich sind an einem regulären Punkt die Gleichungsbeschränkungen  $\mathbf{g}(\mathbf{x})$  *funktional unabhängig*. Die Menge der regulären Punkte ist eine Untermenge von  $\mathcal{X}$ .

Man beachte, dass die Regularität eines Punktes gemäß Definition 3.2 direkt von der Formulierung der Gleichungsbeschränkungen abhängt. Als Beispiel dazu betrachte man die zunächst äquivalent erscheinenden Gleichungsbeschränkungen  $g(x_1, x_2) = x_1 = 0$  und  $g(x_1, x_2) = x_1^2 = 0$  im  $\mathbb{R}^2$ . Beide Beschränkungen definieren die gleiche zulässige Menge  $\mathcal{X}$ , nämlich die gesamte  $x_2$ -Achse. Für  $g(x_1, x_2) = x_1$  ist jeder Punkt von  $\mathcal{X}$  ein regulärer Punkt gemäß der Definition 3.2. Für  $g(x_1, x_2) = x_1^2$  jedoch ist kein Punkt von  $\mathcal{X}$  regulär.

Die zulässige Menge  $\mathcal{X} = \{\mathbf{x} \in \mathbb{R}^n \mid g_i(\mathbf{x}) = 0, i = 1, \dots, p\}$  mit den stetig differenzierbaren Funktionen  $g_i(\mathbf{x}), i = 1, \dots, p$  beschreibt eine  $(n - p)$ -dimensionale  $C^1$ -Mannigfaltigkeit (siehe dazu Anhang A des Skriptums Regelungssysteme 2). Der zugehörige  $(n - p)$ -dimensionale Tangentialraum  $\mathcal{T}_{\mathbf{x}}\mathcal{X}$  an einem regulären Punkt  $\mathbf{x}$  wird durch  $n - p$  linear unabhängige Vektoren aufgespannt.  $\mathcal{T}_{\mathbf{x}}\mathcal{X}$  lässt sich nun als Annulator des  $p$ -dimensionalen Kotangentialraumes, welcher durch die exakten Differentiale  $dg_i : \mathcal{T}_{\mathbf{x}}\mathcal{X} \rightarrow \mathbb{R}, i = 1, \dots, p$  gebildet wird, definieren. Das heißt, es gilt

$$\mathcal{T}_{\mathbf{x}}\mathcal{X} = \left\{ \mathbf{d} \mid dg_i(\mathbf{d}) = \mathbf{L}_{\mathbf{d}}g_i(\mathbf{x}) = \underbrace{\left( \frac{\partial}{\partial \mathbf{x}} g_i(\mathbf{x}) \right)}_{(\nabla g_i)^T(\mathbf{x})} \mathbf{d} = 0, i = 1, \dots, p \right\}. \quad (3.19)$$

Man beachte, dass die Vektoren  $\mathbf{d}$  in (3.19) im allgemeinen Fall, d. h. bei nichtlinearen Gleichungsnebenbedingungen, keine zulässigen Richtungen im Sinne der Definition 3.1 sind. Als Vorbereitung für die Formulierung notwendiger Optimalitätsbedingungen des Optimierungsproblems (3.2) mit reinen Gleichungsnebenbedingungen sei folgendes Lemma angegeben.

**Lemma 3.1** (Zu den Optimalitätsbedingungen mit Gleichungsbeschränkungen). *Es sei  $\mathbf{x}^* \in \mathcal{X}$  mit  $\mathcal{X} = \{\mathbf{x} \in \mathbb{R}^n \mid g_i(\mathbf{x}) = 0, i = 1, \dots, p\}$  ein regulärer Punkt und ein lokaler Extrempunkt von  $f(\mathbf{x})$  unter den Gleichungsnebenbedingungen  $\mathbf{g}(\mathbf{x}) = \mathbf{0}$  mit  $f, g_1, \dots, g_p \in C^1$ . Für alle  $\mathbf{d}$ , die die Bedingung*

$$(\nabla \mathbf{g})^T(\mathbf{x}^*)\mathbf{d} = \mathbf{0} \quad (3.20)$$

*erfüllen, muss auch gelten*

$$(\nabla f)^T(\mathbf{x}^*)\mathbf{d} = 0. \quad (3.21)$$

*Beweisskizze:* Da  $\mathbf{x}^*$  ein regulärer Punkt von  $\mathcal{X}$  ist, liegt jedes  $\mathbf{d}$ , das (3.20) erfüllt, gemäß (3.19) im Tangentialraum  $\mathcal{T}_{\mathbf{x}^*}\mathcal{X}$ . Mit  $\mathbf{x}(\alpha), \alpha \in (-\bar{\alpha}, \bar{\alpha}), \bar{\alpha} > 0$  bezeichne man im Weiteren eine stetig differenzierbare Kurve parametrisiert in  $\alpha$  durch den Punkt  $\mathbf{x}^*$  mit dem Tangentialvektor  $\mathbf{d}$ , so dass gilt  $\mathbf{x}(0) = \mathbf{x}^*$  und  $\left( \frac{d}{d\alpha} \mathbf{x} \right)(0) = \mathbf{d}$ . Da nun  $\mathbf{x}^*$

ein lokaler Extrempunkt ist, muss die Beziehung

$$\left. \frac{d}{d\alpha} f(\mathbf{x}(\alpha)) \right|_{\alpha=0} = \underbrace{\left( \frac{\partial}{\partial \mathbf{x}} f \right) (\mathbf{x}(0))}_{(\nabla f)^T(\mathbf{x}^*)} \underbrace{\left( \frac{d}{d\alpha} \mathbf{x} \right) (0)}_{\mathbf{d}} = 0 \quad (3.22)$$

gelten.  $\square$

Lemma 3.1, speziell (3.21), besagt also, dass  $(\nabla f)(\mathbf{x}^*)$  orthogonal auf den Tangentialraum  $\mathcal{T}_{\mathbf{x}^*} \mathcal{X}$  steht. Dies bedeutet für einen regulären Punkt  $\mathbf{x}^*$ , dass  $(\nabla f)(\mathbf{x}^*)$  sich als Linearkombination von  $(\nabla g_i)(\mathbf{x}^*)$ ,  $i = 1, \dots, p$  darstellen lassen muss, d. h. im Bild von  $(\nabla \mathbf{g})(\mathbf{x}^*)$  liegt. Dies motiviert die Einführung des so genannten *Lagrange-Multiplikators*  $\boldsymbol{\lambda}$  im folgenden Satz.

**Satz 3.4 (Notwendige Optimalitätsbedingungen erster Ordnung).** *Es sei  $\mathbf{x}^* \in \mathcal{X}$  mit  $\mathcal{X} = \{\mathbf{x} \in \mathbb{R}^n \mid g_i(\mathbf{x}) = 0, i = 1, \dots, p\}$  ein regulärer Punkt und ein lokaler Extrempunkt von  $f(\mathbf{x})$  unter den Gleichungsnebenbedingungen  $\mathbf{g}(\mathbf{x}) = \mathbf{0}$  mit  $f, g_1, \dots, g_p \in C^1$ . Dann existiert ein eindeutiges  $\boldsymbol{\lambda}^* \in \mathbb{R}^p$  so, dass gilt*

$$(\nabla f)(\mathbf{x}^*) + (\nabla \mathbf{g})(\mathbf{x}^*) \boldsymbol{\lambda}^* = \mathbf{0} . \quad (3.23)$$

Die notwendige Optimalitätsbedingung (3.23) und die Gleichungsnebenbedingungen  $\mathbf{g}(\mathbf{x}^*) = \mathbf{0}$  bilden ein System von  $n + p$  Gleichungen in den  $n + p$  Unbekannten  $(\mathbf{x}^*, \boldsymbol{\lambda}^*)$ . Man kann nun die *Lagrangefunktion*

$$L(\mathbf{x}, \boldsymbol{\lambda}) = f(\mathbf{x}) + \boldsymbol{\lambda}^T \mathbf{g}(\mathbf{x}) \quad (3.24)$$

einführen und die notwendigen Optimalitätsbedingungen von Satz 3.4 in der Form

$$\left( \frac{\partial}{\partial \mathbf{x}} L \right)^T (\mathbf{x}^*, \boldsymbol{\lambda}^*) = (\nabla f)(\mathbf{x}^*) + (\nabla \mathbf{g})(\mathbf{x}^*) \boldsymbol{\lambda}^* = \mathbf{0} \quad (3.25a)$$

$$\left( \frac{\partial}{\partial \boldsymbol{\lambda}} L \right)^T (\mathbf{x}^*, \boldsymbol{\lambda}^*) = \mathbf{g}(\mathbf{x}^*) = \mathbf{0} \quad (3.25b)$$

schreiben.

**Beispiel 3.2.** Man betrachte das Optimierungsproblem (3.1) mit einer Gleichungsbeschränkung in der Form

$$\min_{\mathbf{x} \in \mathbb{R}^2} f(\mathbf{x}) = (x_1 - 2)^2 + (x_2 - 1)^2 \quad (3.26a)$$

$$\text{u.B.v. } g(\mathbf{x}) = x_2 - 2x_1 = 0 . \quad (3.26b)$$

Abbildung 3.2 stellt die Gerade  $g(\mathbf{x}) = 0$  und die Höhenlinien von  $f(\mathbf{x})$  dar.

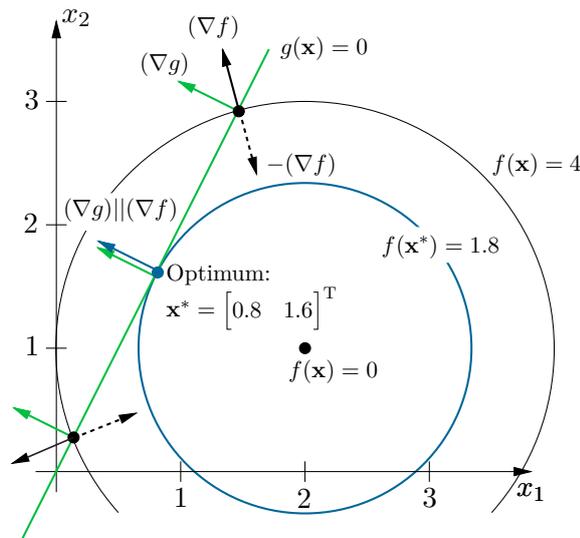


Abbildung 3.2: Veranschaulichung von Beispiel 3.2 mit einer Gleichungsbeschränkung.

Da optimale Punkte  $\mathbf{x}^*$  auf der Geraden  $g(\mathbf{x}) = 0$  liegen müssen, existieren z. B. für die Höhenlinie  $f(\mathbf{x}) = 4$  zwei Schnittpunkte. Es ist direkt ersichtlich, dass für Höhenlinien mit  $f(\mathbf{x}) < 4$  die Schnittpunkte dichter zusammenwandern und schließlich zum Minimum

$$\mathbf{x}^* = [0.8 \quad 1.6]^T, \quad f(\mathbf{x}^*) = 1.8 \quad (3.27)$$

führen. Die Gradienten der Funktionen  $f(\mathbf{x})$  und  $g(\mathbf{x})$  lauten

$$(\nabla f)(\mathbf{x}) = \begin{bmatrix} 2(x_1 - 2) \\ 2(x_2 - 1) \end{bmatrix}, \quad (\nabla g)(\mathbf{x}) = \begin{bmatrix} -2 \\ 1 \end{bmatrix} \quad (3.28)$$

und damit errechnen sich die notwendigen Optimalitätsbedingungen (3.25) mit der Lagrangefunktion  $L(x_1, x_2, \lambda) = f(\mathbf{x}) + \lambda g(\mathbf{x}) = (x_1 - 2)^2 + (x_2 - 1)^2 + \lambda(x_2 - 2x_1)$  zu

$$\frac{\partial}{\partial x_1} L(x_1, x_2, \lambda) = 2(x_1 - 2) - 2\lambda = 0 \quad (3.29a)$$

$$\frac{\partial}{\partial x_2} L(x_1, x_2, \lambda) = 2(x_2 - 1) + \lambda = 0 \quad (3.29b)$$

$$\frac{\partial}{\partial \lambda} L(x_1, x_2, \lambda) = x_2 - 2x_1 = 0. \quad (3.29c)$$

Die Lösung dieses Gleichungssystems liefert den optimalen Punkt  $x_1^* = 0.8$ ,  $x_2^* = 1.6$  und  $\lambda^* = -1.2$ .

**Aufgabe 3.3.** Zeigen Sie, dass unter allen möglichen Quadern mit der gegebenen Oberfläche  $A$  der Würfel mit der Seitenlänge  $\sqrt{A/6}$  das größte Volumen besitzt.

**Aufgabe 3.4.** Gegeben ist ein nichtlineares zeitvariantes Abtastsystem der Form

$$x_{k+1} = \varphi_k(x_k, u_k), \quad x_0 = x(0) \quad (3.30)$$

mit dem Zustand  $x$  und dem Eingang  $u$ . Gesucht sind die Steuerfolge  $(u_0, u_1, \dots, u_N)$  und die zugehörigen Zustände  $(x_0, x_1, \dots, x_N)$  so, dass die Kostenfunktion

$$J = \sum_{k=0}^N \psi_k(x_k, u_k) \quad (3.31)$$

minimiert wird und die Endbedingungen  $g(x_{N+1}) = 0$  erfüllt ist. Nehmen Sie dabei an, dass die partiellen Ableitungen erster Ordnung aller auftretenden Funktionen stetig sind und die LICQ Bedingung gemäß Definition 3.2 erfüllt ist. Zeigen Sie, dass mit der optimalen Lösung die Gleichungen

$$\lambda_{k-1} = \lambda_k \left( \frac{\partial}{\partial x} \varphi_k \right) (x_k, u_k) + \left( \frac{\partial}{\partial x} \psi_k \right) (x_k, u_k), \quad k = 1, \dots, N \quad (3.32a)$$

$$\lambda_N = \mu \left( \frac{\partial}{\partial x} g \right) (x_{N+1}) \quad (3.32b)$$

$$0 = \lambda_k \left( \frac{\partial}{\partial u} \varphi_k \right) (x_k, u_k) + \left( \frac{\partial}{\partial u} \psi_k \right) (x_k, u_k), \quad k = 0, \dots, N \quad (3.32c)$$

mit einem geeigneten Wert  $\mu$  verbunden sind.

**Satz 3.5 (Notwendige Optimalitätsbedingungen zweiter Ordnung).** *Es sei  $\mathbf{x}^* \in \mathcal{X}$  mit  $\mathcal{X} = \{\mathbf{x} \in \mathbb{R}^n \mid g_i(\mathbf{x}) = 0, i = 1, \dots, p\}$  ein regulärer Punkt und ein lokales Minimum von  $f(\mathbf{x})$  unter den Gleichungsnebenbedingungen  $\mathbf{g}(\mathbf{x}) = \mathbf{0}$  mit  $f, g_1, \dots, g_p \in C^2$ . Dann existiert ein eindeutiges  $\boldsymbol{\lambda}^* \in \mathbb{R}^p$  so, dass gilt*

$$\left( \frac{\partial}{\partial \mathbf{x}} L \right)^T (\mathbf{x}^*, \boldsymbol{\lambda}^*) = (\nabla f)(\mathbf{x}^*) + (\nabla \mathbf{g})(\mathbf{x}^*) \boldsymbol{\lambda}^* = \mathbf{0} \quad (3.33)$$

und

$$\mathbf{d}^T (\nabla^2 L)(\mathbf{x}^*, \boldsymbol{\lambda}^*) \mathbf{d} = \mathbf{d}^T \left( (\nabla^2 f)(\mathbf{x}^*) + \sum_{i=1}^p \lambda_i^* (\nabla^2 g_i)(\mathbf{x}^*) \right) \mathbf{d} \geq 0 \quad (3.34)$$

für alle  $\mathbf{d} \in \mathcal{T}_{\mathbf{x}^*} \mathcal{X}$  mit der Lagrangefunktion  $L$  gemäß (3.24).

Die Sätze 3.4 und 3.5 geben notwendige Bedingungen an, die ein lokales Minimum des beschränkten Optimierungsproblems erfüllen muss. Der nächste Satz formuliert nun hinreichende Bedingungen für ein striktes lokales Minimum des Optimierungsproblems (3.1) mit reinen Gleichungsnebenbedingungen.

**Satz 3.6 (Hinreichende Optimalitätsbedingungen zweiter Ordnung).** *Gesucht ist das Minimum der Kostenfunktion  $f(\mathbf{x})$  unter den Gleichungsnebenbedingungen  $g_i(\mathbf{x}) =$*

$0, i = 1, \dots, p$  mit  $f, g_1, \dots, g_p \in C^2$ . Wenn  $(\mathbf{x}^*, \boldsymbol{\lambda}^*)$  so existieren, dass

$$\left( \frac{\partial}{\partial \mathbf{x}} L \right)^T (\mathbf{x}^*, \boldsymbol{\lambda}^*) = \mathbf{0} \quad (3.35)$$

$g_i(\mathbf{x}^*) = 0, i = 1, \dots, p$  und

$$\mathbf{d}^T (\nabla^2 L) (\mathbf{x}^*, \boldsymbol{\lambda}^*) \mathbf{d} > 0 \quad \forall \mathbf{d} \in \mathcal{T}_{\mathbf{x}^*} \mathcal{X}, \mathbf{d} \neq \mathbf{0} \quad (3.36)$$

mit der Lagrangefunktion  $L$  gemäß (3.24), dann ist  $\mathbf{x}^*$  ein striktes lokales Minimum.

Man erkennt aus Satz 3.6, dass die Matrix  $\mathbf{L} = (\nabla^2 L)(\mathbf{x}^*, \boldsymbol{\lambda}^*)$  eine ähnliche Rolle wie die Hessematrix  $(\nabla^2 f)(\mathbf{x}^*)$  der Kostenfunktion  $f(\mathbf{x})$  im unbeschränkten Fall spielt (siehe die Sätze 2.2 und 2.3). In der Tat wird das Konvergenzverhalten des beschränkten Optimierungsproblems durch die Eigenwerte der Matrix  $\mathbf{L}$  eingeschränkt auf den Tangentialraum  $\mathcal{T}_{\mathbf{x}^*} \mathcal{X}$  bestimmt. Um nun die Erfüllung von (3.36) zu überprüfen, kann die Matrix  $\mathbf{L}$  auf den Tangentialraum  $\mathcal{T}_{\mathbf{x}^*} \mathcal{X}$  projiziert und dort auf positive Definitheit getestet werden. Dazu verwendet man die Transformationsmatrix  $\mathbf{T} \in \mathbb{R}^{n \times (n-p)}$ , deren Spaltenvektoren eine orthonormale Basis von  $\mathcal{T}_{\mathbf{x}^*} \mathcal{X}$  bilden, d. h.  $\mathbf{T}^T \mathbf{T} = \mathbf{E}$ . Es lässt sich nun für jedes  $\mathbf{d} \in \mathcal{T}_{\mathbf{x}^*} \mathcal{X}$  stets ein  $\mathbf{z} \in \mathbb{R}^{n-p}$  so finden, dass gilt  $\mathbf{d} = \mathbf{Tz}$ . Setzt man  $\mathbf{d} = \mathbf{Tz}$  in (3.36) ein, so erhält man

$$\mathbf{d}^T (\nabla^2 L) (\mathbf{x}^*, \boldsymbol{\lambda}^*) \mathbf{d} = \mathbf{d}^T \mathbf{L} \mathbf{d} = \mathbf{z}^T \mathbf{T}^T \mathbf{L} \mathbf{T} \mathbf{z} > 0 \quad \forall \mathbf{z} \in \mathbb{R}^{n-p}, \mathbf{z} \neq \mathbf{0}. \quad (3.37)$$

Die Projektion der symmetrischen Matrix  $\mathbf{L}$  in den Tangentialraum  $\mathcal{T}_{\mathbf{x}^*} \mathcal{X}$  ergibt sich also in der Form

$$\mathbf{L}_S = \mathbf{T}^T \mathbf{L} \mathbf{T}, \quad (3.38)$$

und die Überprüfung der Erfüllung von (3.36) reduziert sich auf die Prüfung der positiven Definitheit von  $\mathbf{L}_S$ . Es sei nun  $\lambda$  ein Eigenwert von  $\mathbf{L}_S$  und  $\mathbf{v}$  der zugehörige Eigenvektor (zugleich Links- und Rechtseigenvektor). Folglich gilt

$$0 = \mathbf{v}^T (\lambda \mathbf{E} - \mathbf{L}_S) \mathbf{v} = \mathbf{v}^T (\lambda \mathbf{T}^T \mathbf{T} - \mathbf{T}^T \mathbf{L} \mathbf{T}) \mathbf{v} = \mathbf{v}^T \mathbf{T}^T (\lambda \mathbf{E} - \mathbf{L}) \mathbf{T} \mathbf{v}. \quad (3.39)$$

Aus diesem Ergebnis und der Spaltenregularität von  $\mathbf{T}$  folgt, dass die Singularität von  $\lambda \mathbf{E} - \mathbf{L}_S$  die Singularität von  $\lambda \mathbf{E} - \mathbf{L}$  impliziert. Damit ist gezeigt, dass jeder Eigenwert von  $\mathbf{L}_S$  auch ein Eigenwert von  $\mathbf{L}$  ist.

**Beispiel 3.3.** Man betrachte das Optimierungsproblem (3.1) mit einer Gleichungsbeschränkung in der Form

$$\min_{\mathbf{x} \in \mathbb{R}^3} f(\mathbf{x}) = x_1 + x_2^2 + x_2 x_3 + 2x_3^2 \quad (3.40a)$$

$$\text{u.B.v. } g(\mathbf{x}) = x_1^2 + x_2^2 + x_3^2 - 1 = 0. \quad (3.40b)$$

Die notwendigen Optimalitätsbedingungen erster Ordnung nach Satz 3.4 lauten

$$\frac{\partial}{\partial x_1} L(\mathbf{x}^*, \boldsymbol{\lambda}^*) = 1 + 2\lambda^* x_1^* = 0 \quad (3.41a)$$

$$\frac{\partial}{\partial x_2} L(\mathbf{x}^*, \lambda^*) = 2x_2^* + x_3^* + 2\lambda^* x_2^* = 0 \quad (3.41b)$$

$$\frac{\partial}{\partial x_3} L(\mathbf{x}^*, \lambda^*) = x_2^* + 4x_3^* + 2\lambda^* x_3^* = 0 \quad (3.41c)$$

$$\frac{\partial}{\partial \lambda} L(\mathbf{x}^*, \lambda^*) = (x_1^*)^2 + (x_2^*)^2 + (x_3^*)^2 - 1 = 0. \quad (3.41d)$$

Mit dem regulären Punkt  $\mathbf{x}^* = [1 \ 0 \ 0]^T$  und dem Lagrange-Multiplikator  $\lambda^* = -1/2$  ist eine Lösung von (3.41) gegeben. Zur Prüfung der notwendigen Optimalitätsbedingung zweiter Ordnung gemäß Satz 3.5 wird die Matrix

$$\mathbf{L} = (\nabla^2 L)(\mathbf{x}^*, \lambda^*) = \begin{bmatrix} -1 & 0 & 0 \\ 0 & 1 & 1 \\ 0 & 1 & 3 \end{bmatrix} \quad (3.42)$$

benötigt.

Um den Tangentialraum  $\mathcal{T}_{\mathbf{x}^*} \mathcal{X}$  der Mannigfaltigkeit  $\mathcal{X} = \{\mathbf{x} \in \mathbb{R}^3 \mid x_1^2 + x_2^2 + x_3^2 = 1\}$  gemäß (3.19) zu berechnen, bestimme man vorerst den Ausdruck

$$\left( \frac{\partial}{\partial \mathbf{x}} g \right) (\mathbf{x}^*) = \left[ 2x_1 \ 2x_2 \ 2x_3 \right] \Big|_{\mathbf{x}=\mathbf{x}^*} = \left[ 2 \ 0 \ 0 \right]. \quad (3.43)$$

Aus (3.43) und der Definition (3.19) für den Tangentialraum am Punkt  $\mathbf{x}^*$  folgt, dass die erste Komponente aller Vektoren  $\mathbf{d} \in \mathcal{T}_{\mathbf{x}^*} \mathcal{X}$  identisch Null sein muss. Folglich ist für alle  $\mathbf{d} \in \mathcal{T}_{\mathbf{x}^*} \mathcal{X}$  die Beziehung (3.36) von Satz 3.6 erfüllt, denn die Submatrix  $\mathbf{L}_{[2..3,2..3]}$  ist positiv definit. Alternativ erhält man dieses Ergebnis durch Projektion der Matrix  $\mathbf{L}$ . Wählt man dazu zwei orthogonale Basisvektoren des Tangentialraumes  $\mathcal{T}_{\mathbf{x}^*} \mathcal{X}$  und fasst man diese als Spaltenvektoren in der Matrix  $\mathbf{T}$  zusammen, z. B.

$$\mathbf{T} = \begin{bmatrix} 0 & 0 \\ 1 & 0 \\ 0 & 1 \end{bmatrix}, \quad (3.44)$$

dann kann die Matrix  $\mathbf{L}$  von (3.42) gemäß (3.38) wie folgt

$$\mathbf{L}_S = \mathbf{T}^T \mathbf{L} \mathbf{T} = \begin{bmatrix} 1 & 1 \\ 1 & 3 \end{bmatrix} \quad (3.45)$$

in den Tangentialraum projiziert werden. Da die Matrix  $\mathbf{L}_S$  positiv definit ist, folgt aus Satz 3.6, dass der Punkt  $\mathbf{x}^* = [1 \ 0 \ 0]^T$  ein striktes lokales Minimum des beschränkten Optimierungsproblems (3.40) ist.

Angenommen der Punkt  $\mathbf{x}^*$  ist eine Lösung des beschränkten Optimierungsproblems (3.16) mit dem zugehörigen Lagrange-Multiplikator  $\lambda^*$ . Dann lässt sich  $\lambda^*$  wie folgt interpretieren.

**Satz 3.7** (Sensitivitätstheorem des Lagrange-Multiplikators bei Gleichungsbeschränkungen). Für  $f, g_1, \dots, g_p \in C^2$  betrachte man folgende Familie beschränkter Optimierungsprobleme

$$\min_{\mathbf{x} \in \mathbb{R}^n} f(\mathbf{x}) \quad (3.46a)$$

$$\text{u.B.v. } \mathbf{g}(\mathbf{x}) = \mathbf{c}_g \quad (3.46b)$$

mit  $\mathbf{c}_g \in \mathbb{R}^p$ . Angenommen, für  $\mathbf{c}_g = \mathbf{0}$  sei  $\mathbf{x}^*$  ein regulärer Punkt und erfülle gemeinsam mit dem Lagrange-Multiplikator  $\boldsymbol{\lambda}^*$  die hinreichenden Optimalitätsbedingungen zweiter Ordnung von Satz 3.6 für ein striktes lokales Minimum. Dann existiert für jedes  $\mathbf{c}_g \in \mathbb{R}^p$  in einer Umgebung von  $\mathbf{0}$  ein lokales Minimum des Optimierungsproblems (3.46) an einer Stelle  $(\mathbf{x}(\mathbf{c}_g), \boldsymbol{\lambda}(\mathbf{c}_g))$ , welche stetig von  $\mathbf{c}_g$  abhängt mit  $\mathbf{x}(\mathbf{0}) = \mathbf{x}^*$  und  $\boldsymbol{\lambda}(\mathbf{0}) = \boldsymbol{\lambda}^*$ . Ferner gilt die Beziehung

$$\left. \frac{d}{d\mathbf{c}_g} f(\mathbf{x}(\mathbf{c}_g)) \right|_{\mathbf{c}_g=\mathbf{0}} = -(\boldsymbol{\lambda}^*)^T. \quad (3.47)$$

*Beweisskizze:* Die notwendigen Optimalitätsbedingungen erster Ordnung für das Optimierungsproblem (3.46) lauten

$$(\nabla f)(\mathbf{x}) + (\nabla \mathbf{g})(\mathbf{x})\boldsymbol{\lambda} = \mathbf{0} \quad (3.48a)$$

$$\mathbf{g}(\mathbf{x}) = \mathbf{c}_g. \quad (3.48b)$$

Berechnet man die Jacobi-Matrix von (3.48) an der Stelle  $(\mathbf{x}^*, \boldsymbol{\lambda}^*)$ , also für  $\mathbf{c}_g = \mathbf{0}$ , dann erhält man (siehe auch (3.34))

$$\begin{bmatrix} (\nabla^2 L)(\mathbf{x}^*, \boldsymbol{\lambda}^*) & (\nabla \mathbf{g})(\mathbf{x}^*) \\ (\nabla \mathbf{g})^T(\mathbf{x}^*) & \mathbf{0} \end{bmatrix}. \quad (3.49)$$

Da nach den Voraussetzungen von Satz 3.6 die Matrix  $(\nabla^2 L)(\mathbf{x}^*, \boldsymbol{\lambda}^*)$  positiv definit auf  $\mathcal{T}_{\mathbf{x}^*} \mathcal{X}$  sein muss (striktes lokales Minimum an der Stelle  $\mathbf{x}^*$ ) und die Matrix  $(\nabla \mathbf{g})(\mathbf{x}^*)$  regulär ist ( $\mathbf{x}^*$  ist ein regulärer Punkt), ist die Jacobi-Matrix (3.49) regulär auf  $\mathcal{T}_{\mathbf{x}^*} \mathcal{X}$ .

**Aufgabe 3.5.** Beweisen Sie diese Behauptung.

Mit Hilfe des Satzes über implizite Funktionen kann daraus geschlossen werden, dass  $\mathbf{x}(\mathbf{c}_g)$  und  $\boldsymbol{\lambda}(\mathbf{c}_g)$  stetig differenzierbare Funktionen in  $\mathbf{c}_g$  sind.

Die Ableitung von (3.48b) nach  $\mathbf{c}_g$  am Punkt  $\mathbf{c}_g = \mathbf{0}$  liefert

$$\left. \frac{d\mathbf{g}(\mathbf{x}(\mathbf{c}_g))}{d\mathbf{c}_g} \right|_{\mathbf{c}_g=\mathbf{0}} = (\nabla \mathbf{g})^T(\mathbf{x}^*) \left. \frac{d\mathbf{x}}{d\mathbf{c}_g} \right|_{\mathbf{c}_g=\mathbf{0}} = \mathbf{E}. \quad (3.50)$$

Wird die Transponierte von (3.48a) rechtsseitig mit  $\frac{d\mathbf{x}}{d\mathbf{c}_g}$  multipliziert, so ergibt sich

am Punkt  $\mathbf{c}_g = \mathbf{0}$

$$(\nabla f)^T(\mathbf{x}^*) \frac{d\mathbf{x}}{d\mathbf{c}_g} \Big|_{\mathbf{c}_g=\mathbf{0}} + \underbrace{(\boldsymbol{\lambda}^*)^T (\nabla \mathbf{g})^T(\mathbf{x}^*) \frac{d\mathbf{x}}{d\mathbf{c}_g} \Big|_{\mathbf{c}_g=\mathbf{0}}}_{\mathbf{E}} = \mathbf{0}. \quad (3.51)$$

Gemeinsam mit (3.50) folgt daraus

$$\frac{d}{d\mathbf{c}_g} f(\mathbf{x}(\mathbf{c}_g)) \Big|_{\mathbf{c}_g=\mathbf{0}} + (\boldsymbol{\lambda}^*)^T = \mathbf{0} \quad (3.52)$$

und damit (3.47). □

### 3.1.2.2 Gleichungs- und Ungleichungsbeschränkungen

Ausgangspunkt der weiteren Betrachtungen ist das Optimierungsproblem mit Gleichungs- und Ungleichungsbeschränkungen (3.1) bzw. (3.2). In diesem Fall wird die Menge der Indizes aller am aktuellen Punkt  $\mathbf{x}$  aktiven Ungleichungsbeschränkungen in der Form

$$J = J(\mathbf{x}) = \{j \in \mathbb{N} \mid 1 \leq j \leq q, h_j(\mathbf{x}) = 0\} \quad (3.53)$$

definiert. Wieder soll

$$\mathcal{X} = \{\mathbf{x} \in \mathbb{R}^n \mid g_i(\mathbf{x}) = 0, i = 1, \dots, p, h_j(\mathbf{x}) \leq 0, j = 1, \dots, q\} \quad (3.54)$$

die zulässige Menge genannt werden. Ferner beschreibt

$$\bar{\mathcal{X}} = \bar{\mathcal{X}}(\mathbf{x}) = \{\bar{\mathbf{x}} \in \mathcal{X} \mid h_j(\bar{\mathbf{x}}) = 0, j \in J(\mathbf{x})\} \quad (3.55)$$

die durch die Gleichungs- und aktiven Ungleichungsbeschränkungen definierte Mannigfaltigkeit. Man beachte, dass  $J$  und  $\bar{\mathcal{X}}$  vom aktuell betrachteten Punkt  $\mathbf{x}$  abhängen.

**Definition 3.3** (Regulärer Punkt bei Gleichungs- und Ungleichungsbeschränkungen, LICQ). Ein zulässiger Punkt  $\mathbf{x} \in \mathcal{X}$  der Optimierungsaufgabe (3.2) mit Gleichungs- und Ungleichungsbeschränkungen ist *regulär*, wenn die Gradientenvektoren  $(\nabla g_i)(\mathbf{x})$ ,  $i = 1, \dots, p$  und  $(\nabla h_j)(\mathbf{x})$ ,  $j \in J$  mit  $J$  gemäß (3.53) linear unabhängig sind. D. h. die Bedingung

$$\text{rang} \left( \left[ [(\nabla g_i)(\mathbf{x})]_{i=1, \dots, p} \quad [(\nabla h_j)(\mathbf{x})]_{j \in J} \right] \right) = p + |J|, \quad (3.56)$$

muss erfüllt sein, welche im Englischen auch als *linear independence constraint qualification* (LICQ) bekannt ist.

**Satz 3.8** (Karush-Kuhn-Tucker (KKT) notwendige Optimalitätsbedingungen erster Ordnung). Angenommen,  $\mathbf{x}^*$  sei ein lokales Minimum des Optimierungsproblems (3.1)

bzw. (3.2)

$$\min_{\mathbf{x} \in \mathbb{R}^n} f(\mathbf{x}) \quad (3.57a)$$

$$\text{u.B.v. } g_i(\mathbf{x}) = 0, \quad i = 1, \dots, p \quad (3.57b)$$

$$h_i(\mathbf{x}) \leq 0, \quad i = 1, \dots, q \quad (3.57c)$$

mit  $f, g_1, \dots, g_p, h_1, \dots, h_q \in C^1$ . Im Weiteren sei  $\mathbf{x}^*$  ein regulärer Punkt der Beschränkungen (3.57b) und (3.57c). Dann existieren eindeutige Lagrange-Multiplikatoren  $((\boldsymbol{\lambda}^*)^T, (\boldsymbol{\mu}^*)^T)$  mit  $\boldsymbol{\lambda}^* \in \mathbb{R}^p$  und  $\boldsymbol{\mu}^* \in \mathbb{R}^q$  so, dass die Bedingungen

$$(\nabla f)(\mathbf{x}^*) + (\nabla \mathbf{g})(\mathbf{x}^*)\boldsymbol{\lambda}^* + (\nabla \mathbf{h})(\mathbf{x}^*)\boldsymbol{\mu}^* = \mathbf{0} \quad (3.58a)$$

$$\boldsymbol{\mu}^* \geq \mathbf{0} \quad (3.58b)$$

$$\mathbf{h}^T(\mathbf{x}^*)\boldsymbol{\mu}^* = 0 \quad (3.58c)$$

mit den Jacobi-Matrizen  $(\nabla \mathbf{g})(\mathbf{x}^*) = [(\nabla g_1)(\mathbf{x}^*) \quad \dots \quad (\nabla g_p)(\mathbf{x}^*)]$  und  $(\nabla \mathbf{h})(\mathbf{x}^*) = [(\nabla h_1)(\mathbf{x}^*) \quad \dots \quad (\nabla h_q)(\mathbf{x}^*)]$  erfüllt sind.

*Beweisskizze:* Da  $\boldsymbol{\mu}^* \geq \mathbf{0}$  und  $\mathbf{h}(\mathbf{x}^*) \leq \mathbf{0}$  folgt aus (3.58c), dass eine Komponente  $\mu_j^*$  von  $\boldsymbol{\mu}^*$  nur dann von Null verschieden sein kann, wenn die zugehörige Ungleichungsbedingung aktiv ist, d. h.  $h_j(\mathbf{x}^*) = 0$  gilt. Diese so genannte *complementary slackness condition* hat zur Folge, dass  $h_j(\mathbf{x}^*) < 0$  stets  $\mu_j^* = 0$  und  $\mu_j^* > 0$  stets  $h_j(\mathbf{x}^*) = 0$  impliziert.

Da  $\mathbf{x}^*$  ein lokales Minimum des beschränkten Optimierungsproblems (3.57) beschreibt, ist es auch ein lokales Minimum für jenes Optimierungsproblem, bei dem alle aktiven Ungleichungsbeschränkungen durch Gleichungsbeschränkungen ersetzt werden. Dann gibt (3.58a) mit  $\mu_j^* = 0$  falls  $h_j(\mathbf{x}^*) < 0$  exakt die Bedingung (3.23) des bei gleichungsbeschränkten Problemen anwendbaren Satzes 3.4 wieder.

Um zu zeigen, dass  $\boldsymbol{\mu}^* \geq \mathbf{0}$  gelten muss, schreibt man (3.57c) in

$$\mathbf{h}(\mathbf{x}) \leq \mathbf{c}_h \leq \mathbf{0}$$

mit  $\mathbf{c}_h \in \mathbb{R}^q$  um, wobei am optimalen Punkt  $\mathbf{c}_h = \mathbf{0}$  gelten soll. Ähnlich zum Beweis von Satz 3.7 (siehe auch Satz 3.11) folgt, dass  $\mathbf{x}(\mathbf{c}_h)$ ,  $\boldsymbol{\lambda}(\mathbf{c}_h)$  und  $\boldsymbol{\mu}(\mathbf{c}_h)$  stetig differenzierbare Funktionen von  $\mathbf{c}_h$  sind und

$$(\boldsymbol{\mu}^*)^T = - \left. \frac{d}{d\mathbf{c}_h} f(\mathbf{x}(\mathbf{c}_h)) \right|_{\mathbf{c}_h = \mathbf{0}}$$

gilt. Hierbei gilt  $\mu_j^* = 0 \quad \forall j \in \{1, \dots, q\} \setminus J$ , d. h. für alle inaktiven Ungleichungsbeschränkungen  $h_j(\mathbf{x}^*) < 0$ . Die Entwicklung von  $f$  in eine Taylorreihe um den

optimalen Punkt liefert

$$f(\mathbf{x}(\mathbf{c}_h)) = f(\mathbf{x}(\mathbf{0})) + \left. \frac{d}{d\mathbf{c}_h} f(\mathbf{x}(\mathbf{c}_h)) \right|_{\mathbf{c}_h=\mathbf{0}} \mathbf{c}_h + \mathcal{O}(\mathbf{c}_h^2) = f(\mathbf{x}(\mathbf{0})) - \boldsymbol{\mu}^* \mathbf{c}_h + \mathcal{O}(\mathbf{c}_h^2).$$

Aus diesem Ergebnis folgt  $\boldsymbol{\mu}^* \geq \mathbf{0}$ , da  $\mathbf{c}_h \leq \mathbf{0}$  und da wegen der Optimalität von  $\mathbf{x}(\mathbf{0})$  die Ungleichung  $f(\mathbf{x}(\mathbf{0})) \leq f(\mathbf{x}(\mathbf{c}_h))$  zumindest für  $\mathbf{c}_h \rightarrow \mathbf{0}^-$  erfüllt sein muss.  $\square$

Man beachte, dass *nicht* jedes lokale Minimum die KKT-Bedingungen (3.58) erfüllt. Dies ist nur der Fall, wenn die Beschränkungen (3.57b) und (3.57c) gewisse Voraussetzungen erfüllen, die im Englischen auch als *constraint qualification (CQ)* bezeichnet werden, vgl. [3.1–3.3]. Diese Voraussetzungen sind jedenfalls erfüllt, wenn  $\mathbf{x}^*$  ein regulärer Punkt gemäß Definition 3.3 ist, d. h. wenn die LICQ Bedingung erfüllt ist. Diese garantiert, dass die Lagrange-Multiplikatoren  $((\boldsymbol{\lambda}^*)^T, (\boldsymbol{\mu}^*)^T)$  *eindeutig* sind. Es existieren auch andere CQ Bedingungen, die diese Eindeutigkeit nicht garantieren.

Ein Problem bei der Berechnung von  $\mathbf{x}^*$  gemäß Satz 3.8 ist, dass man *a priori* nicht weiß, welche Ungleichungsbeschränkungen aktiv sind. Eigentlich müssten sämtliche Kombinationen aktiver und inaktiver Ungleichungsbeschränkungen überprüft werden, um mögliche (lokale) Minima zu finden. Das nachfolgende Beispiel zeigt dies für eine einfache Optimierungsaufgabe.

*Beispiel 3.4.* Man betrachte das Optimierungsproblem (3.1) mit zwei Ungleichungsbeschränkungen in der Form

$$\min_{\mathbf{x} \in \mathbb{R}^3} f(\mathbf{x}) = \frac{1}{2}(x_1^2 + x_2^2 + x_3^2) \quad (3.59a)$$

$$\text{u.B.v. } h_1(\mathbf{x}) = x_1 + x_2 + x_3 + 3 \leq 0 \quad (3.59b)$$

$$h_2(\mathbf{x}) = x_1 \leq 0. \quad (3.59c)$$

Mit  $(\nabla h_1)(\mathbf{x}) = [1 \ 1 \ 1]^T$  und  $(\nabla h_2)(\mathbf{x}) = [1 \ 0 \ 0]^T$  erkennt man, dass jeder zulässige Punkt ein regulärer Punkt ist. Die KKT-Bedingungen (3.58) lauten in diesem Fall

$$\underbrace{\begin{bmatrix} x_1^* \\ x_2^* \\ x_3^* \end{bmatrix}}_{(\nabla f)(\mathbf{x}^*)} + \underbrace{\begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix}}_{(\nabla h_1)(\mathbf{x}^*)} \mu_1^* + \underbrace{\begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix}}_{(\nabla h_2)(\mathbf{x}^*)} \mu_2^* = \mathbf{0} \quad (3.60a)$$

$$\mu_1^* \geq 0 \quad (3.60b)$$

$$\mu_2^* \geq 0 \quad (3.60c)$$

$$\mu_1^*(x_1^* + x_2^* + x_3^* + 3) + \mu_2^* x_1^* = 0 \quad (3.60d)$$

$$x_1^* + x_2^* + x_3^* + 3 \leq 0 \quad (3.60e)$$

$$x_1^* \leq 0. \quad (3.60f)$$

Nun können vier Fälle unterschieden werden.

- a) Beide Ungleichungsbeschränkungen sind inaktiv, d. h.  $h_1(\mathbf{x}^*) = x_1^* + x_2^* + x_3^* + 3 < 0$  und  $h_2(\mathbf{x}^*) = x_1^* < 0$ . Damit folgt  $\mu_1^* = \mu_2^* = 0$  und  $x_1^* = x_2^* = x_3^* = 0$  wäre die einzige Lösung, die aber nicht zulässig ist, da sie die Ungleichungsbedingung  $h_1(\mathbf{x}^*)$  verletzt.
- b) Die Ungleichungsbeschränkung  $h_1(\mathbf{x}^*)$  ist inaktiv und  $h_2(\mathbf{x}^*)$  ist aktiv, d. h.  $h_1(\mathbf{x}^*) = x_1^* + x_2^* + x_3^* + 3 < 0$ ,  $h_2(\mathbf{x}^*) = x_1^* = 0$ ,  $\mu_1^* = 0$  und  $\mu_2^* > 0$ . Die Lösung  $x_2^* = x_3^* = 0$  ist wiederum kein zulässiger Punkt.
- c) Die Ungleichungsbeschränkung  $h_1(\mathbf{x}^*)$  ist aktiv und  $h_2(\mathbf{x}^*)$  ist inaktiv, d. h.  $h_1(\mathbf{x}^*) = x_1^* + x_2^* + x_3^* + 3 = 0$ ,  $h_2(\mathbf{x}^*) = x_1^* < 0$ ,  $\mu_1^* > 0$  und  $\mu_2^* = 0$ . Die Lösung  $x_1^* = x_2^* = x_3^* = -1$  und  $\mu_1^* = 1$ ,  $\mu_2^* = 0$  ist damit ein zulässiger Kandidat für ein (lokales) Minimum.
- d) Beide Ungleichungsbeschränkungen sind aktiv, d. h.  $h_1(\mathbf{x}^*) = x_1^* + x_2^* + x_3^* + 3 = 0$ ,  $h_2(\mathbf{x}^*) = x_1^* = 0$ ,  $\mu_1^* > 0$  und  $\mu_2^* > 0$ . Da wegen  $x_1^* = 0$  die Beziehung  $\mu_1^* = -\mu_2^*$  gelten muss, widerspricht dies der Forderung  $\mu_1^* > 0$  und  $\mu_2^* > 0$ .

Ob es sich nun beim Punkt  $x_1^* = x_2^* = x_3^* = -1$  tatsächlich um ein (lokales) Minimum handelt, kann basierend auf den nächsten Sätzen geklärt werden.

**Satz 3.9** (KKT notwendige Optimalitätsbedingungen zweiter Ordnung). *Angenommen,  $\mathbf{x}^*$  sei ein lokales Minimum des Optimierungsproblems (3.1) bzw. (3.2)*

$$\min_{\mathbf{x} \in \mathbb{R}^n} f(\mathbf{x}) \quad (3.61a)$$

$$\text{u.B.v. } g_i(\mathbf{x}) = 0, \quad i = 1, \dots, p \quad (3.61b)$$

$$h_i(\mathbf{x}) \leq 0, \quad i = 1, \dots, q \quad (3.61c)$$

mit  $f, g_1, \dots, g_p, h_1, \dots, h_q \in C^2$ . Im Weiteren sei  $\mathbf{x}^*$  ein regulärer Punkt der Beschränkungen (3.61b) und (3.61c). Dann existieren eindeutige Lagrange-Multiplikatoren  $((\boldsymbol{\lambda}^*)^T, (\boldsymbol{\mu}^*)^T)$  mit  $\boldsymbol{\lambda}^* \in \mathbb{R}^p$  und  $\boldsymbol{\mu}^* \in \mathbb{R}^q$  so, dass die Bedingungen

$$(\nabla f)(\mathbf{x}^*) + (\nabla \mathbf{g})(\mathbf{x}^*)\boldsymbol{\lambda}^* + (\nabla \mathbf{h})(\mathbf{x}^*)\boldsymbol{\mu}^* = \mathbf{0} \quad (3.62a)$$

$$\boldsymbol{\mu}^* \geq \mathbf{0} \quad (3.62b)$$

$$\mathbf{h}^T(\mathbf{x}^*)\boldsymbol{\mu}^* = 0 \quad (3.62c)$$

mit den Jacobi-Matrizen  $(\nabla \mathbf{g})(\mathbf{x}^*) = [(\nabla g_1)(\mathbf{x}^*) \ \dots \ (\nabla g_p)(\mathbf{x}^*)]$  und  $(\nabla \mathbf{h})(\mathbf{x}^*) = [(\nabla h_1)(\mathbf{x}^*) \ \dots \ (\nabla h_q)(\mathbf{x}^*)]$  erfüllt sind und

$$\mathbf{d}^T \underbrace{\left( (\nabla^2 f)(\mathbf{x}^*) + \sum_{i=1}^p \lambda_i^* (\nabla^2 g_i)(\mathbf{x}^*) + \sum_{j=1}^q \mu_j^* (\nabla^2 h_j)(\mathbf{x}^*) \right)}_{(\nabla^2 L)(\mathbf{x}^*, \boldsymbol{\lambda}^*, \boldsymbol{\mu}^*)} \mathbf{d} \geq 0 \quad (3.63)$$

für alle  $\mathbf{d} \in \mathcal{T}_{\mathbf{x}^*} \bar{\mathcal{X}}$  mit  $\bar{\mathcal{X}} = \{\mathbf{x} \in \mathbb{R}^n \mid g_i(\mathbf{x}) = 0, i = 1, \dots, p, h_j(\mathbf{x}) = 0, j \in J\}$  und der Lagrangefunktion  $L(\mathbf{x}^*, \boldsymbol{\lambda}^*, \boldsymbol{\mu}^*) = f(\mathbf{x}^*) + (\boldsymbol{\lambda}^*)^T \mathbf{g}(\mathbf{x}^*) + (\boldsymbol{\mu}^*)^T \mathbf{h}(\mathbf{x}^*)$  gilt. Mit  $J$  wird dabei wieder die Menge der Indizes der am Punkt  $\mathbf{x}^*$  aktiven Ungleichungsbeschränkungen bezeichnet, d. h. es gilt  $h_j(\mathbf{x}^*) = 0, j \in J$ .

**Satz 3.10** (KKT hinreichende Optimalitätsbedingungen zweiter Ordnung). Gesucht ist das (lokale) Minimum des Optimierungsproblems (3.1) bzw. (3.2)

$$\min_{\mathbf{x} \in \mathbb{R}^n} f(\mathbf{x}) \quad (3.64a)$$

$$\text{u.B.v. } g_i(\mathbf{x}) = 0, \quad i = 1, \dots, p \quad (3.64b)$$

$$h_i(\mathbf{x}) \leq 0, \quad i = 1, \dots, q \quad (3.64c)$$

mit  $f, g_1, \dots, g_p, h_1, \dots, h_q \in C^2$ . Wenn für einen regulären Punkt  $\mathbf{x}^*$  der Beschränkungen (3.64b) und (3.64c), Größen  $\mathbf{x}^* \in \mathbb{R}^n$ ,  $\boldsymbol{\lambda}^* \in \mathbb{R}^p$  und  $\boldsymbol{\mu}^* \in \mathbb{R}^q$  so existieren, dass gilt

$$(\nabla f)(\mathbf{x}^*) + (\nabla \mathbf{g})(\mathbf{x}^*) \boldsymbol{\lambda}^* + (\nabla \mathbf{h})(\mathbf{x}^*) \boldsymbol{\mu}^* = \mathbf{0} \quad (3.65a)$$

$$\boldsymbol{\mu}^* \geq \mathbf{0} \quad (3.65b)$$

$$\mathbf{h}^T(\mathbf{x}^*) \boldsymbol{\mu}^* = 0 \quad (3.65c)$$

mit den Jacobi-Matrizen  $(\nabla \mathbf{g})(\mathbf{x}^*) = [(\nabla g_1)(\mathbf{x}^*) \ \dots \ (\nabla g_p)(\mathbf{x}^*)]$  und  $(\nabla \mathbf{h})(\mathbf{x}^*) = [(\nabla h_1)(\mathbf{x}^*) \ \dots \ (\nabla h_q)(\mathbf{x}^*)]$  und

$$\mathbf{d}^T \underbrace{\left( (\nabla^2 f)(\mathbf{x}^*) + \sum_{i=1}^p \lambda_i^* (\nabla^2 g_i)(\mathbf{x}^*) + \sum_{j=1}^q \mu_j^* (\nabla^2 h_j)(\mathbf{x}^*) \right)}_{(\nabla^2 L)(\mathbf{x}^*, \boldsymbol{\lambda}^*, \boldsymbol{\mu}^*)} \mathbf{d} > 0 \quad (3.66)$$

für alle  $\mathbf{d} \in \mathcal{T}_{\mathbf{x}^*} \bar{\mathcal{X}}, \mathbf{d} \neq \mathbf{0}$  und  $\bar{\mathcal{X}} = \{\mathbf{x} \in \mathbb{R}^n \mid g_i(\mathbf{x}) = 0, i = 1, \dots, p, h_j(\mathbf{x}) = 0, j \in J\}$  sowie der Lagrangefunktion  $L(\mathbf{x}^*, \boldsymbol{\lambda}^*, \boldsymbol{\mu}^*) = f(\mathbf{x}^*) + (\boldsymbol{\lambda}^*)^T \mathbf{g}(\mathbf{x}^*) + (\boldsymbol{\mu}^*)^T \mathbf{h}(\mathbf{x}^*)$ , dann ist  $\mathbf{x}^*$  ein striktes (lokales) Minimum. Mit  $J$  wird dabei wieder die Menge der Indizes der am Punkt  $\mathbf{x}^*$  aktiven Ungleichungsbeschränkungen bezeichnet, d. h. es gilt  $h_j(\mathbf{x}^*) = 0, j \in J$ .

**Aufgabe 3.6.** Zeigen Sie, dass der Punkt  $x_1^* = x_2^* = x_3^* = -1$  von Beispiel 3.4 ein globales Minimum ist.

Angenommen der Punkt  $\mathbf{x}^*$  ist eine Lösung des beschränkten Optimierungsproblems (3.2). Dann können die Lagrange-Multiplikatoren  $\boldsymbol{\lambda}^*$  und  $\boldsymbol{\mu}^*$  wieder als Sensitivitäten des Kostenfunktionswerts am optimalen Punkt bezüglich einer Änderung der Beschränkungen interpretieren werden.

**Satz 3.11 (Sensitivitätstheorem der Lagrange-Multiplikatoren bei Gleichungs- und Ungleichungsbeschränkungen).** Für  $f, g_1, \dots, g_p, h_1, \dots, h_q \in C^2$  betrachte man folgende Familie beschränkter Optimierungsprobleme

$$\min_{\mathbf{x} \in \mathbb{R}^n} f(\mathbf{x}) \quad (3.67a)$$

$$\text{u.B.v. } \mathbf{g}(\mathbf{x}) = \mathbf{c}_g \quad (3.67b)$$

$$\mathbf{h}(\mathbf{x}) \leq \mathbf{c}_h \quad (3.67c)$$

mit  $\mathbf{c}_g \in \mathbb{R}^p$  und  $\mathbf{c}_h \in \mathbb{R}^q$ . Angenommen, für  $\mathbf{c}_g = \mathbf{0}$  und  $\mathbf{c}_h = \mathbf{0}$  sei  $\mathbf{x}^*$  ein regulärer Punkt und erfülle gemeinsam mit den Lagrange-Multiplikatoren  $\boldsymbol{\lambda}^*$  und  $\boldsymbol{\mu}^*$  die hinreichenden Optimalitätsbedingungen zweiter Ordnung von Satz 3.10 für ein striktes lokales Minimum. Dann existiert für jedes Paar  $(\mathbf{c}_g, \mathbf{c}_h) \in \mathbb{R}^{p+q}$  in einer Umgebung von  $(\mathbf{0}, \mathbf{0})$ , in der  $J$  konstant ist, ein lokales Minimum des Optimierungsproblems (3.67) an einer Stelle  $(\mathbf{x}(\mathbf{c}_g, \mathbf{c}_h), \boldsymbol{\lambda}(\mathbf{c}_g, \mathbf{c}_h), \boldsymbol{\mu}(\mathbf{c}_g, \mathbf{c}_h))$ , welche stetig von  $(\mathbf{c}_g, \mathbf{c}_h)$  abhängt mit  $\mathbf{x}(\mathbf{0}, \mathbf{0}) = \mathbf{x}^*$ ,  $\boldsymbol{\lambda}(\mathbf{0}, \mathbf{0}) = \boldsymbol{\lambda}^*$  und  $\boldsymbol{\mu}(\mathbf{0}, \mathbf{0}) = \boldsymbol{\mu}^*$ . Ferner gelten die Beziehungen

$$\left. \frac{d}{d\mathbf{c}_g} f(\mathbf{x}(\mathbf{c}_g, \mathbf{c}_h)) \right|_{\mathbf{c}_g=\mathbf{0}, \mathbf{c}_h=\mathbf{0}} = -(\boldsymbol{\lambda}^*)^T \quad (3.68a)$$

$$\left. \frac{d}{d\mathbf{c}_h} f(\mathbf{x}(\mathbf{c}_g, \mathbf{c}_h)) \right|_{\mathbf{c}_g=\mathbf{0}, \mathbf{c}_h=\mathbf{0}} = -(\boldsymbol{\mu}^*)^T. \quad (3.68b)$$

**Aufgabe 3.7.** Beweisen Sie Satz 3.11. **Hinweis:** Orientieren Sie sich dabei am Beweis von Satz 3.7.

Man beachte, dass Satz 3.11 nur in einer solchen Umgebung von  $(\mathbf{c}_g, \mathbf{c}_h) = (\mathbf{0}, \mathbf{0})$  gilt, in der  $J$ , die Menge der Indizes der aktiven Ungleichungsbeschränkungen, konstant ist. Wie es sein muss, folgt aus Satz 3.11  $\mu_i^* = 0 \forall i \in \{1, \dots, q\} \setminus J$ .

## 3.2 Rechnergestützte Optimierungsverfahren

Als Ausgangspunkt betrachte man wiederum das beschränkte Optimierungsproblem (3.2)

$$\min_{\mathbf{x} \in \mathbb{R}^n} f(\mathbf{x}) \quad (3.69a)$$

$$\text{u.B.v. } \mathbf{g}(\mathbf{x}) = \mathbf{0} \quad (3.69b)$$

$$\mathbf{h}(\mathbf{x}) \leq \mathbf{0} \quad (3.69c)$$

mit  $p$  Gleichungsbeschränkungen  $g_1(\mathbf{x}), \dots, g_p(\mathbf{x})$ ,  $q$  Ungleichungsbeschränkungen  $h_1(\mathbf{x}), \dots, h_q(\mathbf{x})$  und  $n$  Optimierungsvariablen  $x_1, \dots, x_n$ . Da die Bestimmung eines (lokal) optimalen Punktes  $\mathbf{x}^*$  von (3.69) durch analytische Lösung von Optimalitätsbedingungen (nichtlineare Gleichungen sowie Ungleichungen in  $\mathbf{x}^*$ ,  $\boldsymbol{\lambda}^*$  und  $\boldsymbol{\mu}^*$ ) in vielen Fällen nicht möglich ist, ist man im Allgemeinen auf *numerische Verfahren* zur Suche von  $\mathbf{x}^*$  angewiesen. Einen Überblick über einige dieser Verfahren gibt der aktuelle Abschnitt.

Zur Lösung des Problems (3.69) können im Rahmen der *Methode der aktiven Beschränkungen* aktive und inaktive Ungleichungsbeschränkungen unterschiedlich behandelt werden. Mit der *Gradienten-Projektionsmethode* und der *reduzierten Gradientenmethode* wird während der iterativen Lösungssuche eine Fortbewegung im zulässigen Gebiet sichergestellt. Alternativ kann das Problem auch durch *sequentielle quadratische Programmierung* gelöst werden, wobei hier die Optimierungsaufgabe durch eine Folge von quadratischen Programmen approximiert wird. Es besteht ferner die Möglichkeit, das beschränkte Optimierungsproblem mit Hilfe von so genannten *Straf- bzw. Barrierefunktionen* in ein unbeschränktes Optimierungsproblem zu transformieren.

### 3.2.1 Methode der aktiven Beschränkungen

Ohne Einschränkung der Allgemeinheit sei für die folgenden Betrachtungen angenommen, dass keine Gleichungsbeschränkungen vorhanden sind. Aus Satz 3.8 weiß man, dass die notwendigen Optimalitätsbedingungen für ein lokales Minimum durch

$$(\nabla f)(\mathbf{x}^*) + (\nabla \mathbf{h})(\mathbf{x}^*)\boldsymbol{\mu}^* = \mathbf{0} \quad (3.70a)$$

$$h_i(\mathbf{x}^*) = 0, \quad i \in J \quad (3.70b)$$

$$h_i(\mathbf{x}^*) < 0, \quad i \in \{1, \dots, q\} \setminus J \quad (3.70c)$$

$$\mu_i^* \geq 0, \quad i \in J \quad (3.70d)$$

$$\mu_i^* = 0, \quad i \in \{1, \dots, q\} \setminus J \quad (3.70e)$$

gegeben sind. Mit  $J$  wird dabei wieder die Menge der Indizes der am Punkt  $\mathbf{x}^*$  aktiven Ungleichungsbeschränkungen bezeichnet. Wenn man nun das beschränkte Optimierungsproblem für eine angenommene Menge der aktiven Ungleichungsbeschränkungen löst und diese Lösung die nichtaktiven Ungleichungsbeschränkungen erfüllt sowie ausschließlich nichtnegative Lagrange-Multiplikatoren  $\mu_i^*$  beinhaltet, dann kann die Lösung als Kandidat für das Minimum angesehen werden.

Die Idee der Methode der aktiven Beschränkungen (englisch: *active set method*) beruht darauf, in jedem Iterationsschritt eine *Arbeitsmenge* festzulegen, welche die am aktuellen Iterationspunkt  $\mathbf{x}_k$  aktiven Ungleichungsbeschränkungen beinhaltet. Sind Gleichungsbeschränkungen vorhanden, können diese auf analoge Art und Weise in der Arbeitsmenge berücksichtigt werden. Der aktuelle Iterationspunkt  $\mathbf{x}_k$  ist daher zulässig im Hinblick auf die Arbeitsmenge. Um zum nächsten Iterationspunkt  $\mathbf{x}_{k+1}$  zu gelangen, erfolgt eine Bewegung entlang der durch die Arbeitsmenge definierten Mannigfaltigkeit. Ziel ist es, so zu einem hinsichtlich der Kostenfunktion verbesserten Punkt zu gelangen. Die verschiedenen Verfahren unterscheiden sich dadurch, wie die Bewegung entlang der Mannigfaltigkeit, die durch die Arbeitsmenge definiert ist, erfolgt, wobei dies auch wesentlich das Konvergenzverhalten des Verfahrens bestimmt.

Angenommen,  $W$  bezeichne die Arbeitsmenge, also die Indexmenge der aktiven Ungleichungsbeschränkungen zu einem Iterationsschritt. Dann besteht die Aufgabe in diesem Iterationsschritt darin, für das Optimierungsproblem

$$\min_{\mathbf{x} \in \mathbb{R}^n} f(\mathbf{x}) \quad \text{Kostenfunktion} \quad (3.71a)$$

$$\text{u.B.v. } h_i(\mathbf{x}) = 0, \quad i \in W \quad \text{angenommene aktive Ungleichungsbeschr.} \quad (3.71b)$$

eine Lösung  $\mathbf{x}_W^*$  zu finden, für die gilt  $h_i(\mathbf{x}_W^*) < 0$  für alle  $i \in \{1, \dots, q\} \setminus W$ . Dazu löse man das Gleichungssystem

$$(\nabla f)(\mathbf{x}_W^*) + \sum_{i \in W} \mu_{i,W}^* (\nabla h_i)(\mathbf{x}_W^*) = \mathbf{0} \quad (3.72a)$$

$$h_i(\mathbf{x}_W^*) = 0, \quad i \in W \quad (3.72b)$$

nach  $\mathbf{x}_W^*$  und  $\mu_{i,W}^*$ ,  $i \in W$ . Wenn nun  $\mu_{i,W}^* \geq 0$  für alle  $i \in W$ , dann ist  $\mathbf{x}_W^*$  eine mögliche lokale Lösung des beschränkten Optimierungsproblems. Existiert hingegen ein  $k \in W$ , für das gilt  $\mu_{k,W}^* < 0$ , dann kann die Kostenfunktion weiter reduziert werden, indem die Beschränkung  $h_k(\mathbf{x})$  inaktiv gesetzt wird, d. h. der Index  $k$  wird aus der Indexmenge  $W$  entfernt. Dies folgt unmittelbar aus dem Beweis von Satz 3.8, denn wenn statt der aktiven Ungleichungsbeschränkung  $h_k(\mathbf{x}) = 0$  die Gleichungsbeschränkung  $h_k(\mathbf{x}) = c_h$  mit einem kleinen Wert  $c_h < 0$  verwendet wird, d. h.  $\mathbf{x}$  wird vom Rand in das Innere des Gebietes der Ungleichungsbeschränkung  $h_k(\mathbf{x}) < 0$  bewegt, dann ändert sich für  $\mu_{k,W}^* < 0$  und  $\mathbf{x}(0) = \mathbf{x}_W^*$  die Kostenfunktion in der Form

$$f(\mathbf{x}(c_h)) \approx f(\mathbf{x}_W^*) + \underbrace{\frac{d}{dc_h} f(\mathbf{x}(c_h)) \Big|_{c_h=0}}_{-\mu_{k,W}^* > 0} \underbrace{c_h}_{< 0} < f(\mathbf{x}_W^*) . \quad (3.73)$$

Dies zeigt also, dass durch eine Bewegung in das Innere des Gebietes der Ungleichungsbeschränkung  $h_k(\mathbf{x}) < 0$  die Kostenfunktion weiter minimiert werden kann. Abbildung 3.3 veranschaulicht diesen Sachverhalt grafisch für die Ungleichungsbeschränkung  $h_1(\mathbf{x}) \leq 0$ .

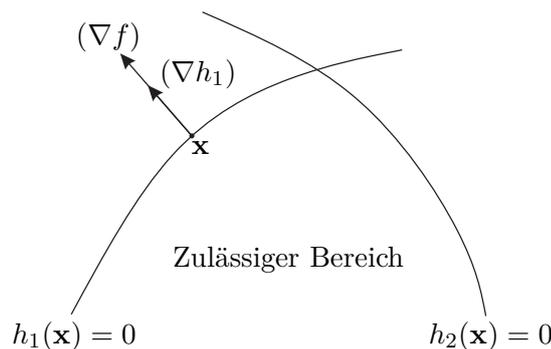


Abbildung 3.3: Zur Deaktivierung von Ungleichungsbeschränkungen.

Natürlich kann es umgekehrt passieren, dass durch die iterative Lösung des beschränkten Optimierungsproblems eine in der Arbeitsmenge als inaktiv erachtete Ungleichungsbeschränkung verletzt wird, d. h. es existiert ein  $k \in \{1, \dots, q\} \setminus W$  so, dass  $h_k(\mathbf{x}_W^*) > 0$ . In

diesem Fall muss die Indexmenge  $W$  um den Index  $k$  dieser Ungleichungsbeschränkung erweitert werden.

Die Methode der aktiven Beschränkungen beruht im Wesentlichen auf den obigen Überlegungen, wobei die Konvergenz durch den nachfolgenden Satz sichergestellt werden kann.

**Satz 3.12 (Konvergenz der Methode der aktiven Beschränkungen).** *Angenommen, für aufeinanderfolgende Arbeitsmengen  $W$  ist das Optimierungsproblem (3.71) wohldefiniert und hat eine eindeutige nichtdegenerierte Lösung (d. h. für alle  $i \in W$ ,  $\mu_{i,W}^* \neq 0$ ), dann konvergiert die Methode der aktiven Beschränkungen gegen die Lösung des zugrundeliegenden beschränkten Optimierungsproblems.*

Eine Schwierigkeit in der praktischen Anwendung ist, dass die Iterationslösungen exakte Lösungen des unterlagerten Minimierungsproblems sein müssen, da ansonsten Vorzeichen der Lagrange-Multiplikatoren falsch sein können. Ferner muss verhindert werden, dass zwischen gleichen Arbeitsmengen wiederkehrend hin- und hergesprungen wird. Dazu gibt es eine Vielzahl unterschiedlicher Strategien, die sich mehr oder weniger von dem vorgestellten Basisalgorithmus unterscheiden.

### 3.2.2 Gradienten-Projektionsmethode

Die Grundidee dieser Methode ist es, die Lösung iterativ entlang jener Mannigfaltigkeit zu suchen, die durch die jeweilige Arbeitsmenge definiert ist. Als Suchrichtung wird der in den Tangentialraum der Mannigfaltigkeit projizierte negative Gradient der Kostenfunktion verwendet.

#### 3.2.2.1 Lineare Beschränkungen

Es wird zunächst das Optimierungsproblem

$$\min_{\mathbf{x} \in \mathbb{R}^n} f(\mathbf{x}) \quad (3.74a)$$

$$\text{u.B.v. } g_i(\mathbf{x}) = \mathbf{a}_{g,i}^T \mathbf{x} - b_{g,i} = 0, \quad i = 1, \dots, p \quad (3.74b)$$

$$h_i(\mathbf{x}) = \mathbf{a}_{h,i}^T \mathbf{x} - b_{h,i} \leq 0, \quad i = 1, \dots, q \quad (3.74c)$$

mit linearen Gleichungs- und Ungleichungsbeschränkungen betrachtet. Zu einem Iterationsschritt  $k$  sei  $\mathbf{x}_k$  der aktuell gefundene Punkt. Es wird angenommen, dass an diesem Punkt in Summe  $\bar{p} = p + |W| < n$  Gleichungs- und Ungleichungsbeschränkungen aktiv sind, wobei  $W$  die aktuelle Arbeitsmenge ist. Die Gleichungs- und aktiven Ungleichungsbeschränkungen werden im Vektor

$$\bar{\mathbf{g}}(\mathbf{x}) = \begin{bmatrix} [g_i(\mathbf{x})]_{i=1,\dots,p} \\ [h_i(\mathbf{x})]_{i \in W} \end{bmatrix}, \quad (3.75)$$

zusammengefasst, dessen (konstante) Jacobi-Matrix mit

$$\mathbf{A} = (\nabla \bar{\mathbf{g}})(\mathbf{x}) = \begin{bmatrix} [\mathbf{a}_{g,i}]_{i=1,\dots,p} & [\mathbf{a}_{h,i}]_{i \in W} \end{bmatrix} \in \mathbb{R}^{n \times \bar{p}} \quad (3.76)$$

abgekürzt wird. Folglich wird der Tangentialraum der durch die aktuell aktiven Beschränkungen definierten Mannigfaltigkeit  $\bar{\mathcal{X}} = \{\mathbf{x} \in \mathcal{X} \mid \bar{\mathbf{g}}(\mathbf{x}) = \mathbf{0}\}$  von den Vektoren des Nullraums  $\text{Kern}(\mathbf{A}^\top)$  aufgespannt, d. h.

$$\mathcal{T}_{\mathbf{x}_k} \bar{\mathcal{X}} = \mathcal{T} \bar{\mathcal{X}} = \{\mathbf{d} \in \mathbb{R}^n \mid \mathbf{A}^\top \mathbf{d} = \mathbf{0}\} \quad (3.77)$$

(vgl. auch (3.19)). Da die Erfüllung der LICQ Bedingung vorausgesetzt wird, ist die Matrix  $\mathbf{A}$  spaltenregulär, d. h.  $\text{rang}(\mathbf{A}) = \bar{p}$ , und es gilt  $\dim(\text{Kern}(\mathbf{A}^\top)) = n - \bar{p}$ . Der Gradient  $(\nabla f)(\mathbf{x}_k)$  im Iterationsschritt  $k$  steht nun im Allgemeinen nicht orthogonal auf  $\mathcal{T} \bar{\mathcal{X}}$ . Aufgrund von  $\mathbb{R}^n = \text{Kern}(\mathbf{A}^\top) \oplus \text{Bild}(\mathbf{A})$  mit  $\text{Bild}(\mathbf{A})$  als dem Bild von  $\mathbf{A}$  lässt sich der negative Gradient  $-(\nabla f)(\mathbf{x}_k)$  immer in der Form

$$-(\nabla f)(\mathbf{x}_k) = \mathbf{d}_k + \mathbf{A} \boldsymbol{\sigma}_k \quad (3.78)$$

für geeignete  $\mathbf{d}_k \in \mathcal{T} \bar{\mathcal{X}}$  und  $\boldsymbol{\sigma}_k \in \mathbb{R}^{\bar{p}}$  anschreiben. Wird (3.78) linksseitig mit  $\mathbf{A}^\top$  multipliziert, so ergibt sich aufgrund der Spaltenregularität von  $\mathbf{A}$  die Beziehung

$$\boldsymbol{\sigma}_k = -(\mathbf{A}^\top \mathbf{A})^{-1} \mathbf{A}^\top (\nabla f)(\mathbf{x}_k) . \quad (3.79)$$

Einsetzen von (3.79) in (3.78) führt auf den projizierten Gradient  $\mathbf{d}_k$  in der Form

$$\mathbf{d}_k = -\mathbf{P}_k (\nabla f)(\mathbf{x}_k) \quad \text{mit} \quad \mathbf{P}_k = \mathbf{E} - \mathbf{A} (\mathbf{A}^\top \mathbf{A})^{-1} \mathbf{A}^\top . \quad (3.80)$$

$\mathbf{P}_k$  wird dabei als *Projektionsmatrix* bezeichnet, weil sie den negativen Gradienten  $-(\nabla f)(\mathbf{x}_k)$  in den Tangentialraum  $\mathcal{T} \bar{\mathcal{X}}$  projiziert.

**Aufgabe 3.8.** Zeigen Sie, dass eine Projektionsmatrix  $\mathbf{P}$  die Eigenschaften  $\mathbf{P}^\top = \mathbf{P}$  sowie  $\mathbf{P}^2 = \mathbf{P}$  erfüllt.

**Aufgabe 3.9.** Zeigen Sie, dass man den projizierten Gradienten  $\mathbf{d}_k$  auch als Lösung des beschränkten Optimierungsproblems

$$\min_{\mathbf{d}_k \in \mathbb{R}^n} \|(\nabla f)(\mathbf{x}_k) + \mathbf{d}_k\|_2^2 \quad (3.81a)$$

$$\text{u.B.v.} \quad \mathbf{A}^\top \mathbf{d}_k = \mathbf{0} \quad (3.81b)$$

erhalten kann.

Aus (3.77) und (3.78) folgt, dass  $(\nabla f)(\mathbf{x}_k) + \mathbf{d}_k = -\mathbf{A} \boldsymbol{\sigma}_k$  orthogonal auf  $\mathcal{T} \bar{\mathcal{X}}$  und damit  $\mathbf{d}_k$  steht. Wenn  $\mathbf{d}_k \neq \mathbf{0}$ , dann ist  $\mathbf{d}_k$  eine *zulässige Abstiegsrichtung* des beschränkten Optimierungsproblems am Punkt  $\mathbf{x}_k$ , denn es gilt

$$(\nabla f)^\top(\mathbf{x}_k) \mathbf{d}_k = \left( (\nabla f)^\top(\mathbf{x}_k) + \mathbf{d}_k^\top - \mathbf{d}_k^\top \right) \mathbf{d}_k = -\|\mathbf{d}_k\|_2^2 < 0 . \quad (3.82)$$

Da nun eine zulässige Abstiegsrichtung  $\mathbf{d}_k$  festliegt, muss lediglich die Schrittweite  $\alpha_k$  zum neuen Iterationspunkt  $\mathbf{x}_{k+1} = \mathbf{x}_k + \alpha_k \mathbf{d}_k$  beispielsweise mit dem skalaren Optimierungsproblem (2.23) bestimmt werden. Es ist hierbei zu beachten, dass die Schrittweite  $\alpha_k$  durch

die am Punkt  $\mathbf{x}_k$  inaktiven Ungleichungsnebenbedingungen nach oben hin beschränkt sein kann. Wenn  $\mathbf{d}_k = \mathbf{0}$  gilt, dann folgt aus (3.78)

$$(\nabla f)(\mathbf{x}_k) + \mathbf{A}\boldsymbol{\sigma}_k = \mathbf{0} . \quad (3.83)$$

Dies entspricht der KKT-Bedingung (3.58a) von Satz 3.8 für das Optimierungsproblem (3.74), wobei  $\boldsymbol{\sigma}_k$  mit den Lagrange-Multiplikatoren der aktiven Beschränkungen übereinstimmt. Wenn keine inaktiven Ungleichungsbeschränkungen verletzt werden und alle Einträge  $\sigma_{k,i}$ ,  $i = p+1, \dots, \bar{p}$  nichtnegativ sind, dann erfüllt der Punkt  $\mathbf{x}_k$  die notwendigen KKT-Bedingungen für ein (lokales) Minimum. Wenn jedoch ein  $j \in \{p+1, \dots, \bar{p}\}$  so existiert, dass  $\sigma_{k,j} < 0$  gilt, dann kann die Kostenfunktion weiter verkleinert werden, indem die Ungleichungsbeschränkung  $h_j(\mathbf{x}) = \mathbf{a}_{h,j}^T \mathbf{x} - b_{h,j} \leq 0$  inaktiv gesetzt wird. Der Algorithmus der Gradienten-Projektionsmethode ist in Tabelle 3.1 zusammengefasst.

### 3.2.2.2 Nichtlineare Beschränkungen

Es wird nun nun anstelle des Optimierungsproblems (3.74) mit linearen Beschränkungen das Optimierungsproblem

$$\min_{\mathbf{x} \in \mathbb{R}^n} f(\mathbf{x}) \quad (3.84a)$$

$$\text{u.B.v. } g_i(\mathbf{x}) = 0, \quad i = 1, \dots, p \quad (3.84b)$$

$$h_i(\mathbf{x}) \leq 0, \quad i = 1, \dots, q \quad (3.84c)$$

mit nichtlinearen Beschränkungen betrachtet. Die Gleichungs- und aktiven Ungleichungsbeschränkungen werden wieder im Vektor

$$\bar{\mathbf{g}}(\mathbf{x}) = \begin{bmatrix} [g_i(\mathbf{x})]_{i=1, \dots, p} \\ [h_i(\mathbf{x})]_{i \in W} \end{bmatrix} \quad (3.85)$$

mit der Dimension  $\bar{p} = p + |W|$  zusammengefasst. Diese aktuell aktiven Beschränkungen definieren die Mannigfaltigkeit  $\bar{\mathcal{X}} = \{\mathbf{x} \in \mathcal{X} \mid \bar{\mathbf{g}}(\mathbf{x}) = \mathbf{0}\}$ .

Im Falle nichtlinearer Beschränkungen ergeben sich Schwierigkeiten, da Elemente des Tangentialraums  $\mathcal{T}_{\mathbf{x}}\bar{\mathcal{X}}$  nicht unbedingt auch zulässige Richtungen sind. Nachfolgend wird kurz beschrieben wie mit dieser Situation umzugehen ist.

Zunächst wird der negative Gradient  $-(\nabla f)(\mathbf{x}_k)$  an einem Punkt  $\mathbf{x}_k$  mit Hilfe der Projektionsmatrix (vergleiche (3.80))

$$\mathbf{P}_k = \mathbf{E} - (\nabla \bar{\mathbf{g}})(\mathbf{x}_k) \left( (\nabla \bar{\mathbf{g}})^T(\mathbf{x}_k) (\nabla \bar{\mathbf{g}})(\mathbf{x}_k) \right)^{-1} (\nabla \bar{\mathbf{g}})^T(\mathbf{x}_k) \quad (3.86)$$

mit  $(\nabla \bar{\mathbf{g}})(\mathbf{x}) = \begin{bmatrix} [(\nabla g_i)(\mathbf{x})]_{i=1, \dots, p} & [(\nabla h_i)(\mathbf{x})]_{i \in W} \end{bmatrix}$  in den Tangentialraum  $\mathcal{T}_{\mathbf{x}_k}\bar{\mathcal{X}}$  projiziert.

Abbildung 3.4 veranschaulicht diese Projektion grafisch. Es ist unmittelbar ersichtlich, dass im Gegensatz zum Problem mit linearen Beschränkungen der Punkt

$$\bar{\mathbf{x}}_{k+1} = \mathbf{x}_k + \alpha_k \mathbf{d}_k \quad \text{mit} \quad \mathbf{d}_k = -\mathbf{P}_k (\nabla f)(\mathbf{x}_k) \quad (3.87)$$

auch für hinreichend kleines  $\alpha_k$  im Allgemeinen nicht mehr auf der Mannigfaltigkeit  $\bar{\mathcal{X}}$  liegt. Es ist deshalb eine weitere Bewegung vom Punkt  $\bar{\mathbf{x}}_{k+1}$  orthogonal zu  $\mathbf{d}_k$  nötig, um

---

<b>Initialisierung:</b>	$\mathbf{x}_0$ (Zulässiger Startpunkt)
	$k = 0$ (Startindex)
	stop = 0 (Abbruch-Flag)
<b>repeat</b>	
Schritt 1:	Suche für den Punkt $\mathbf{x}_k$ die Menge der aktiven Beschränkungen (Mannigfaltigkeit $\bar{\mathcal{X}}$ ) mit der zugehörigen Arbeitsmenge $W$ .
Schritt 2:	Projiziere den negativen Gradienten $-(\nabla f)(\mathbf{x}_k)$ in der Form $\mathbf{d}_k = -\mathbf{P}_k(\nabla f)(\mathbf{x}_k)$ mit Hilfe der Projektionsmatrix $\mathbf{P}_k$ gemäß (3.80) in den Tangentialraum $\mathcal{T}\bar{\mathcal{X}}$ .
Schritt 3:	
<b>if</b> $\mathbf{d}_k \neq \mathbf{0}$	
Berechne	$\alpha_{k,1} = \max\{\alpha_k \mid \mathbf{x}_k + \alpha_k \mathbf{d}_k \in \mathcal{X}\}$ $\alpha_{k,2} = \arg \min_{0 < \alpha_k < \alpha_{k,1}} f(\mathbf{x}_k + \alpha_k \mathbf{d}_k)$
	und setze $\mathbf{x}_{k+1} = \mathbf{x}_k + \alpha_{k,2} \mathbf{d}_k$ und $k \leftarrow k + 1$ .
<b>else</b> (d. h. $\mathbf{d}_k = \mathbf{0}$ )	
Berechne $\boldsymbol{\sigma}_k = -(\mathbf{A}^T \mathbf{A})^{-1} \mathbf{A}^T (\nabla f)(\mathbf{x}_k)$ (siehe (3.79))	
1.	Wenn $\sigma_{k,j} \geq 0$ für alle $j = p + 1, \dots, \bar{p}$ gilt, dann erfüllt $\mathbf{x}_k$ die KKT-Bedingungen, setze stop=1.
2.	Wenn $\sigma_{k,j} \geq 0$ nicht für alle $j = p + 1, \dots, \bar{p}$ gilt, dann streiche jene Ungleichungsbeschränkung, die zur negativsten Komponente $\sigma_{k,j}$ , $j = p + 1, \dots, \bar{p}$ gehört, passe die Indexmenge $W$ und die Matrix $\mathbf{A}$ entsprechend an und gehe zu Schritt 2 in der nächsten Iteration.
<b>end</b>	
<b>until</b> stop == 1	

---

Tabelle 3.1: Gradienten-Projektionsmethode.

auf die Mannigfaltigkeit  $\bar{\mathcal{X}}$  zu gelangen. Die Idee dabei ist, ein  $\boldsymbol{\eta}_k \in \mathbb{R}^{\bar{p}}$  so zu suchen, dass gilt

$$\bar{\mathbf{g}}(\mathbf{x}_{k+1}) = \mathbf{0} \quad \text{mit} \quad \mathbf{x}_{k+1} = \bar{\mathbf{x}}_{k+1} + (\nabla \bar{\mathbf{g}})(\mathbf{x}_k) \boldsymbol{\eta}_k. \quad (3.88)$$

Iterativ kann  $\boldsymbol{\eta}_k$  im Sinne einer Approximation erster Ordnung wie folgt berechnet werden. Unter Verwendung von (3.88) wird der Ausdruck  $\bar{\mathbf{g}}(\mathbf{x}_{k+1})$  bezüglich  $\mathbf{x}_{k+1}$  am Punkt  $\mathbf{x}_k$  in

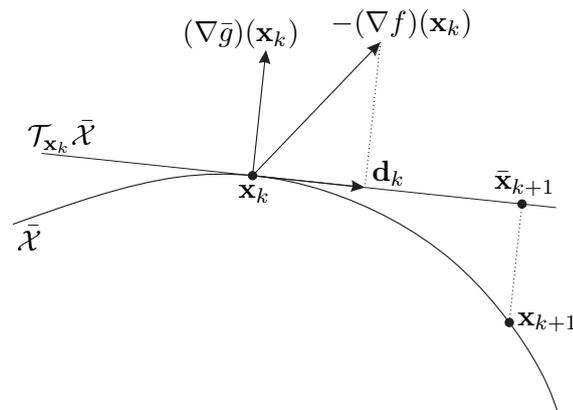


Abbildung 3.4: Gradienten-Projektionsmethode.

eine Reihe entwickelt, welche nach dem linearen Glied abgebrochen wird, d. h.

$$\begin{aligned} \mathbf{0} &= \bar{\mathbf{g}}(\mathbf{x}_{k+1}) \approx \bar{\mathbf{g}}(\mathbf{x}_k) + (\nabla \bar{\mathbf{g}})^\top(\mathbf{x}_k)(\mathbf{x}_{k+1} - \mathbf{x}_k) \\ &= \bar{\mathbf{g}}(\mathbf{x}_k) + (\nabla \bar{\mathbf{g}})^\top(\mathbf{x}_k)(\mathbf{x}_{k+1} - \bar{\mathbf{x}}_{k+1} + \bar{\mathbf{x}}_{k+1} - \mathbf{x}_k) \\ &\approx \bar{\mathbf{g}}(\bar{\mathbf{x}}_{k+1}) + (\nabla \bar{\mathbf{g}})^\top(\mathbf{x}_k)(\nabla \bar{\mathbf{g}})(\mathbf{x}_k)\boldsymbol{\eta}_k. \end{aligned} \quad (3.89)$$

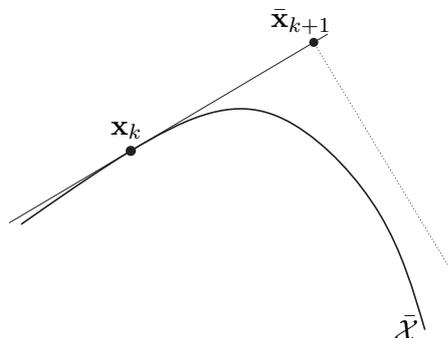
Daraus folgen

$$\boldsymbol{\eta}_k = -\left[(\nabla \bar{\mathbf{g}})^\top(\mathbf{x}_k)(\nabla \bar{\mathbf{g}})(\mathbf{x}_k)\right]^{-1} \bar{\mathbf{g}}(\bar{\mathbf{x}}_{k+1}) \quad (3.90a)$$

$$\mathbf{x}_{k+1} = \bar{\mathbf{x}}_{k+1} - (\nabla \bar{\mathbf{g}})(\mathbf{x}_k) \left[(\nabla \bar{\mathbf{g}})^\top(\mathbf{x}_k)(\nabla \bar{\mathbf{g}})(\mathbf{x}_k)\right]^{-1} \bar{\mathbf{g}}(\bar{\mathbf{x}}_{k+1}). \quad (3.90b)$$

Diese Ausdrücke beinhalten die gleichen Matrizen wie die Projektionsmatrix  $\mathbf{P}_k$  gemäß (3.86). Gleichung (3.90b) kann iterativ ausgeführt werden, um einen zulässigen Punkt  $\mathbf{x}_{k+1}$  zu finden.

Man beachte, dass ein Vektor  $\boldsymbol{\eta}_k$  gemäß (3.88) nicht immer existieren muss, siehe beispielsweise Abbildung 3.5. Tritt dieses Problem auf, muss die Schrittweite  $\alpha_k$  in (3.87) reduziert werden.

Abbildung 3.5: Nichtexistenz von  $\boldsymbol{\eta}_k$ , um auf  $\bar{\mathcal{X}}$  zu kommen.

Ein weiteres Problem, das bei der Gradienten-Projektionsmethode mit nichtlinearen Beschränkungen auftreten kann, besteht darin, dass Ungleichungsbeschränkungen, welche am Punkt  $\mathbf{x}_k$  inaktiv waren, bei einer Bewegung in Richtung des projizierten negativen Gradienten verletzt werden können, siehe Abbildung 3.6. Typischerweise werden in diesem Zusammenhang Verfahren zur Interpolation zwischen den Punkten  $\mathbf{x}_k$  und  $\bar{\mathbf{x}}_{k+1}$  eingesetzt, um zu einem neuen Punkt  $\bar{\bar{\mathbf{x}}}_{k+1}$  bzw. nach der Projektion  $\mathbf{x}_{k+1}$  zu gelangen, der die ursprünglich inaktiven Ungleichungsbeschränkungen nicht verletzt.

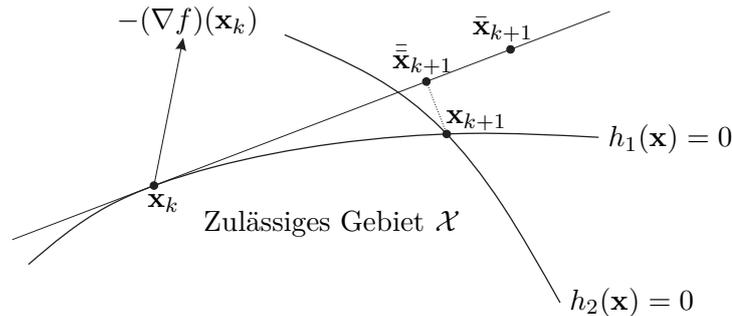


Abbildung 3.6: Interpolation zur Einhaltung der ursprünglich inaktiven Ungleichungsbeschränkungen.

**Aufgabe 3.10.** Lösen Sie das beschränkte Optimierungsproblem (3.40) aus Beispiel 3.3 numerisch mit Hilfe der Gradienten-Projektionsmethode. Erstellen Sie dafür ein Computerprogramm, wobei Sie für die Bestimmung der Schrittweite  $\alpha_k$  wahlweise selbst einen Algorithmus (z. B. aus Abschnitt 2.3.1) programmieren oder vorgefertigte Funktionen einsetzen können. Verwenden Sie als Startpunkt  $\mathbf{x}_0 = [1/\sqrt{2} \ 1/\sqrt{2} \ 0]^T$ . Zur Projektion des Punktes  $\bar{\mathbf{x}}_{k+1}$  zurück auf die Mannigfaltigkeit  $\bar{\mathcal{X}}$ , also zum Finden eines zulässigen Punktes  $\mathbf{x}_{k+1}$  können Sie (3.90b) iterativ anwenden.

**Lösung von Aufgabe 3.10.** Am Startpunkt beträgt der Kostenfunktionswert  $f(\mathbf{x}_0) = 1.207107$ . Im Verlauf der ersten Iteration ergeben sich folgende Zwischenergebnisse:

$$\mathbf{P}_0 = \frac{1}{2} \begin{bmatrix} 1 & -1 & 0 \\ -1 & 1 & 0 \\ 0 & 0 & 2 \end{bmatrix}, \quad \mathbf{d}_0 = \begin{bmatrix} \sqrt{2} - 1/2 \\ 1/2 - \sqrt{2} \\ -1/\sqrt{2} \end{bmatrix}, \quad \alpha_0 = 0.246289,$$

$$\bar{\mathbf{x}}_1 = \begin{bmatrix} 0.758115 \\ 0.656099 \\ -0.174153 \end{bmatrix}, \quad \mathbf{x}_1 = \begin{bmatrix} 0.745439 \\ 0.643422 \\ -0.174153 \end{bmatrix}, \quad f(\mathbf{x}_1) = 1.108035$$

Nach zehn Iterationen gelangt der Algorithmus zum Punkt

$$\mathbf{x}_{10} = \begin{bmatrix} 0.997\,635 \\ 0.065\,201 \\ -0.021\,753 \end{bmatrix}$$

mit dem Kostenfunktionswert  $f(\mathbf{x}_{10}) = 1.001\,414$ . Aus Beispiel 3.3 ist bekannt, dass die exakte Lösung  $\mathbf{x}^* = [1 \ 0 \ 0]^T$  mit  $f(\mathbf{x}^*) = 1$  lautet.

Probleme, die sich bei der Gradienten-Projektionsmethode im Zusammenhang mit nichtlinearen Beschränkungen ergeben und eine weitere Bewegung vom Tangentialraum  $\mathcal{T}_{\mathbf{x}_k} \bar{\mathcal{X}}$  zurück auf die Mannigfaltigkeit  $\bar{\mathcal{X}}$  erfordern (siehe Abbildung 3.4), können vermieden werden, wenn bereits bei der Gradientenberechnung eine Bewegung entlang der Mannigfaltigkeit  $\bar{\mathcal{X}}$  erzwungen wird. Die nachfolgend beschriebene Methode tut dies.

### 3.2.3 Reduzierte Gradientenmethode

Die Idee der *reduzierten Gradientenmethode* ist, dass grundsätzlich nur Bewegungen auf der Mannigfaltigkeit, die durch die aktuelle Arbeitsmenge definiert ist, erlaubt werden. Bei der Gradientenberechnung wird dies im Sinne der Kettenregel berücksichtigt. Gelegentlich wird die Methode auch als *generalisierte reduzierte Gradientenmethode* bezeichnet.

Man betrachte wieder das Optimierungsproblem

$$\min_{\mathbf{x} \in \mathbb{R}^n} f(\mathbf{x}) \quad (3.91a)$$

$$\text{u.B.v. } g_i(\mathbf{x}) = 0, \quad i = 1, \dots, p \quad (3.91b)$$

$$h_i(\mathbf{x}) \leq 0, \quad i = 1, \dots, q \quad (3.91c)$$

mit nichtlinearen Beschränkungen. Analog zu Abschnitt 3.2.2.2 sei  $W$  die aktuelle Arbeitsmenge und der Vektor  $\bar{\mathbf{g}}(\mathbf{x})$  mit der Dimension  $\bar{p} < n$  fasse die Gleichungs- und die im aktuellen Iterationsschritt aktiven Ungleichungsbeschränkungen zusammen, welche die Mannigfaltigkeit  $\bar{\mathcal{X}}$  definieren.

Die Optimierungsvariablen werden nun im aktuellen Iterationsschritt so in  $n - \bar{p}$  unabhängige Variablen  $\mathbf{x}_I$  und  $\bar{p}$  abhängige Variablen  $\mathbf{x}_D$  partitioniert, dass

$$\mathbf{x} = \begin{bmatrix} \mathbf{x}_I \\ \mathbf{x}_D \end{bmatrix} \quad (3.92)$$

gilt und die Matrix

$$\mathbf{A}_D(\mathbf{x}) = \left( \frac{\partial \bar{\mathbf{g}}(\mathbf{x})}{\partial \mathbf{x}_D} \right) = \begin{bmatrix} \frac{\partial \bar{g}_1}{\partial x_{D,1}} & \dots & \frac{\partial \bar{g}_1}{\partial x_{D,\bar{p}}} \\ \vdots & & \vdots \\ \frac{\partial \bar{g}_{\bar{p}}}{\partial x_{D,1}} & \dots & \frac{\partial \bar{g}_{\bar{p}}}{\partial x_{D,\bar{p}}} \end{bmatrix} \quad (3.93)$$

regulär ist. Die in (3.92) gewählte Reihenfolge der Variablen stellt keine Einschränkung dar, da sie durch Umsortieren immer hergestellt werden kann. Im Verlauf der Iterationen

kann sich die Partitionierung (3.92) ändern. Gemäß dem Satz über implizite Funktionen ist bei gegebenem  $\mathbf{x}_I$  mit der geforderten Regularität der Matrix  $\mathbf{A}_D$  (zumindest lokal) sichergestellt, dass  $\mathbf{x}_D$  aus  $\bar{\mathbf{g}}(\mathbf{x}) = \mathbf{0}$  eindeutig ausgerechnet werden kann und stetig differenzierbar von  $\mathbf{x}_I$  abhängt. Formal existiert damit eine (zumindest lokal definierte) stetig differenzierbare Funktion  $\bar{\mathbf{g}}_D^{-1}(\mathbf{x}_I)$  so, dass

$$\bar{\mathbf{g}}(\mathbf{x}) = \mathbf{0} \quad \Leftrightarrow \quad \mathbf{x}_D = \bar{\mathbf{g}}_D^{-1}(\mathbf{x}_I). \quad (3.94)$$

Damit lässt sich das *reduzierte unbeschränkte Optimierungsproblem*

$$\min_{\mathbf{x}_I \in \mathbb{R}^{n-\bar{p}}} f\left(\begin{bmatrix} \mathbf{x}_I \\ \bar{\mathbf{g}}_D^{-1}(\mathbf{x}_I) \end{bmatrix}\right) \quad (3.95)$$

formulieren, welches nur  $n - \bar{p}$  Optimierungsvariablen besitzt und in der aktuellen Iteration lokal äquivalent zum ursprünglichen (höherdimensionaleren) Optimierungsproblem (3.91) ist. Die Lösungssuche wird damit automatisch auf die Mannigfaltigkeit  $\mathcal{X}$  eingeschränkt. Aus der Lösung  $\mathbf{x}_I^*$  des unbeschränkten Optimierungsproblems (3.95) folgt schließlich noch der optimale Wert  $\mathbf{x}_D^* = \bar{\mathbf{g}}_D^{-1}(\mathbf{x}_I^*)$  für die abhängigen Variablen.

*Bemerkung 3.1.* Gerade bei regelungstechnischen Optimierungsaufgaben ist die in (3.92) vorgenommene Einteilung der Optimierungsvariablen in unabhängige und abhängige Variablen häufig sehr einfach möglich. Die zu optimierenden *Stellgrößen* des Systems stellen unabhängige Variablen dar, während die *Zustandsgrößen* abhängige Variablen darstellen, deren Werte durch Zustandsgleichungen eindeutig definiert und im Allgemeinen einfach berechenbar sind (siehe dazu auch Aufgabe 3.4). Da die meisten dynamischen Systeme viele Zustandsgrößen aber nur wenige Stellgrößen besitzen, ist damit die Dimension des Optimierungsproblems (3.95) erheblich kleiner als jene von (3.91).

Die Lösung  $\mathbf{x}_I^*$  des unbeschränkten Problems (3.95) kann z. B. mit den in Kapitel 2 vorgestellten Methoden gefunden werden. Die dabei häufig benötigte (totale) Ableitung der Kostenfunktion (3.95) bezüglich  $\mathbf{x}_I$  (Gradient) kann wie folgt berechnet werden. Zunächst bilde man das totale Differenzial von  $\bar{\mathbf{g}}(\mathbf{x}) = \mathbf{0}$

$$\begin{aligned} d\bar{\mathbf{g}}(\mathbf{x}) &= \underbrace{\frac{\partial \bar{\mathbf{g}}(\mathbf{x})}{\partial \mathbf{x}_I}}_{= \mathbf{A}_I(\mathbf{x})} d\mathbf{x}_I + \mathbf{A}_D(\mathbf{x}) d\mathbf{x}_D = \mathbf{0} \end{aligned} \quad (3.96)$$

aus dem

$$\frac{d\mathbf{x}_D}{d\mathbf{x}_I} = \frac{d\bar{\mathbf{g}}_D^{-1}(\mathbf{x}_I)}{d\mathbf{x}_I} = -\mathbf{A}_D^{-1} \mathbf{A}_I \quad (3.97)$$

folgt. Daraus ergibt sich die gesuchte totale Ableitung

$$\begin{aligned} \left(\frac{df(\mathbf{x})}{d\mathbf{x}_I}\right)^T &= \left(\frac{\partial f(\mathbf{x})}{\partial \mathbf{x}_I} - \frac{\partial f(\mathbf{x})}{\partial \mathbf{x}_D} \mathbf{A}_D^{-1} \mathbf{A}_I\right)^T = \underbrace{\left[\mathbf{E} \quad -(\mathbf{A}_D^{-1} \mathbf{A}_I)^T\right]}_{= \bar{\mathbf{P}}(\mathbf{x})} (\nabla f)(\mathbf{x}), \end{aligned} \quad (3.98)$$

welche auch als *reduzierter Gradient* bezeichnet wird. In analoger Weise folgt die (totale) zweite Ableitung der Kostenfunktion (3.95) bezüglich  $\mathbf{x}_I$  (Hessematrix) in der Form

$$\frac{d^2 f(\mathbf{x})}{d\mathbf{x}_I^T d\mathbf{x}_I} = \begin{bmatrix} \frac{d^2 f}{dx_{I,1}^2} & \cdots & \frac{d^2 f}{dx_{I,1} dx_{I,n-\bar{p}}} \\ \vdots & & \vdots \\ \frac{d^2 f}{dx_{I,n-\bar{p}} dx_{I,1}} & \cdots & \frac{d^2 f}{dx_{I,n-\bar{p}}^2} \end{bmatrix} = \bar{\mathbf{P}}(\mathbf{x}) (\nabla^2 f)(\mathbf{x}) \bar{\mathbf{P}}^T(\mathbf{x}), \quad (3.99)$$

welche auch als *reduzierte Hessematrix* bezeichnet wird. Bei der Berechnung dieser Ableitungen ist es also nicht notwendig, die Funktion  $\bar{\mathbf{g}}_D^{-1}(\mathbf{x}_I)$  explizit zu kennen.

Man beachte, dass mit dem reduzierten unbeschränkten Optimierungsproblem (3.95) noch nicht sichergestellt ist, dass die in der aktuellen Iteration gefundene Lösung  $\mathbf{x}_I^*$ ,  $\mathbf{x}_D^*$  die KKT-Bedingungen gemäß Satz 3.8 für das ursprüngliche beschränkte Optimierungsproblem (3.91) erfüllt. Zudem ist es häufig nicht möglich, einen analytischen Ausdruck für die Funktion  $\bar{\mathbf{g}}_D^{-1}(\mathbf{x}_I)$ , welche auch in (3.95) auftritt, zu finden. Praktisch wird daher meist nicht das reduzierte Optimierungsproblem (3.95) selbst gelöst, sondern nur die zugehörige Ableitung (3.98) (gegebenenfalls auch (3.99)) berechnet, um im Zuge der iterativen Lösungssuche entlang der aktuellen Mannigfaltigkeit  $\bar{\mathcal{X}}$  eine neue Suchrichtung  $\mathbf{s}_{I,k}$  für die unabhängigen Variablen  $\mathbf{x}_I$  zu bestimmen. Im einfachsten Fall der Gradientenmethode gilt  $\mathbf{s}_{I,k} = -(\mathrm{d}f(\mathbf{x}_k)/\mathrm{d}\mathbf{x}_{I,k})^T$ . Es können aber auch die weiteren in Abschnitt 2.3.2 besprochenen Methoden zur Wahl einer Suchrichtung  $\mathbf{s}_{I,k}$  verwendet werden.

Tabelle 3.2 fasst den zugehörigen Algorithmus zusammen. Im Zuge dieses Algorithmus muss die Funktion  $\bar{\mathbf{g}}_D^{-1}(\mathbf{x}_I)$  weder bekannt sein noch ausgewertet werden. Für einen bestimmten Wert  $\mathbf{x}_I$  kann  $\mathbf{x}_D$  stets als Lösung der Gleichung  $\bar{\mathbf{g}}(\mathbf{x}) = \mathbf{0}$  berechnet werden. Folglich ist  $\mathbf{x}_k$  immer ein zulässiger Punkt im Sinne der Gleichungs- und aktiven Ungleichungsbeschränkungen, was bei einem vorzeitigen Abbruch der Iteration von Vorteil sein kann.

Wie nachfolgendes Lemma zeigt, kann der reduzierte Gradient  $(\mathrm{d}f(\mathbf{x})/\mathrm{d}\mathbf{x}_I)^T$  alternativ zu (3.98) auch mit Hilfe der Lagrangefunktion berechnet werden.

**Lemma 3.2** (Berechnung des reduzierten Gradienten mit Hilfe der Lagrangefunktion).

Es sei  $\mathbf{x} \in \mathcal{X}$  ein regulärer Punkt des Optimierungsproblems (3.91). Für eine gegebene Partitionierung (3.92) kann der reduzierte Gradient  $(\mathrm{d}f(\mathbf{x})/\mathrm{d}\mathbf{x}_I)^T$  mit Hilfe der Lagrangefunktion

$$L(\mathbf{x}, \bar{\boldsymbol{\lambda}}) = f(\mathbf{x}) + \bar{\boldsymbol{\lambda}}^T \bar{\mathbf{g}}(\mathbf{x}) \quad (3.100)$$

in der Form

$$\left(\frac{\mathrm{d}f(\mathbf{x})}{\mathrm{d}\mathbf{x}_I}\right)^T = \left(\frac{\partial}{\partial \mathbf{x}_I} L\right)^T(\mathbf{x}, \bar{\boldsymbol{\lambda}}) = \left(\frac{\partial f(\mathbf{x})}{\partial \mathbf{x}_I}\right)^T + \left(\frac{\partial \bar{\mathbf{g}}(\mathbf{x})}{\partial \mathbf{x}_I}\right)^T \bar{\boldsymbol{\lambda}} \quad (3.101)$$

berechnet werden, wobei  $\mathbf{x}_D$  und  $\bar{\boldsymbol{\lambda}}$  aus den Bedingungen

$$\left(\frac{\partial}{\partial \boldsymbol{\lambda}} L\right)^T(\mathbf{x}, \bar{\boldsymbol{\lambda}}) = \bar{\mathbf{g}}(\mathbf{x}) = \mathbf{0} \quad (3.102a)$$

---

**Initialisierung:**  $\mathbf{x}_0$  (Zulässiger Startpunkt)  
 $k = 0$  (Startindex)  
 $\text{stop} = 0$  (Abbruch-Flag)

**repeat**

Schritt 1: Suche für den Punkt  $\mathbf{x}_k$  die Menge der aktiven Beschränkungen (Mannigfaltigkeit  $\bar{\mathcal{X}}$ ) mit der zugehörigen Arbeitsmenge  $W$ .

Schritt 2: Partitioniere  $\mathbf{x}$  gemäß (3.92) in unabhängige Variablen  $\mathbf{x}_I$  und abhängige Variablen  $\mathbf{x}_D$  so, dass  $\mathbf{A}_D(\mathbf{x}_k)$  gemäß (3.93) regulär ist.

Schritt 3: Berechne am Punkt  $\mathbf{x}_k$  den reduzierten Gradienten  $\mathbf{d}_{I,k} = (df(\mathbf{x}_k)/d\mathbf{x}_I)^T$  gemäß (3.98) und gegebenenfalls die reduzierte Hessematrix gemäß (3.99).

Schritt 4:

**if**  $\mathbf{d}_{I,k} \neq \mathbf{0}$

Wähle basierend auf  $\mathbf{d}_{I,k}$  (und der reduzierten Hessematrix) eine geeignete Suchrichtung  $\mathbf{s}_{I,k}$  im Raum der unabhängigen Variablen  $\mathbf{x}_I$ .  
 Berechne

$$\alpha_{k,1} = \max_{\alpha_k} \left[ \begin{array}{l} \alpha_k \\ \text{u.B.v.} \left[ \begin{array}{l} \mathbf{x}_{I,k} + \alpha_k \mathbf{s}_{I,k} \\ \bar{\mathbf{g}}_D^{-1}(\mathbf{x}_{I,k} + \alpha_k \mathbf{s}_{I,k}) \end{array} \right] \in \mathcal{X} \end{array} \right]$$

$$\alpha_{k,2} = \arg \min_{0 < \alpha_k < \alpha_{k,1}} f \left( \left[ \begin{array}{l} \mathbf{x}_{I,k} + \alpha_k \mathbf{s}_{I,k} \\ \bar{\mathbf{g}}_D^{-1}(\mathbf{x}_{I,k} + \alpha_k \mathbf{s}_{I,k}) \end{array} \right] \right),$$

setze  $\mathbf{x}_{I,k+1} = \mathbf{x}_{I,k} + \alpha_{k,2} \mathbf{s}_{I,k}$ , berechne  $\mathbf{x}_{D,k+1}$  aus  $\bar{\mathbf{g}}(\mathbf{x}_{k+1}) = \mathbf{0}$  und setze  $k \leftarrow k + 1$ .

**else** (d. h.  $\mathbf{d}_{I,k} = \mathbf{0}$ )

Prüfe ob KKT-Bedingungen (3.57) erfüllt sind.

1. Wenn Punkt  $\mathbf{x}_k$  die KKT-Bedingungen erfüllt, setze  $\text{stop}=1$ .
2. Andernfalls passe die Arbeitsmenge  $W$  durch Hinzunahme von verletzten Ungleichungsbeschränkungen ( $h_i(\mathbf{x}_k) > 0$ ) oder durch Streichung der Ungleichungsbeschränkung mit dem kleinsten negativen Lagrange-Multiplikator ( $\mu_i < 0$ ) an und gehe zu Schritt 2 in der nächsten Iteration.

**end**

**until**  $\text{stop} == 1$

---

Tabelle 3.2: Reduzierte Gradientenmethode.

$$\left(\frac{\partial}{\partial \mathbf{x}_D} L\right)^T(\mathbf{x}, \bar{\boldsymbol{\lambda}}) = \left(\frac{\partial f(\mathbf{x})}{\partial \mathbf{x}_D}\right)^T + \left(\frac{\partial \bar{\mathbf{g}}(\mathbf{x})}{\partial \mathbf{x}_D}\right)^T \bar{\boldsymbol{\lambda}} = \mathbf{0} \quad (3.102b)$$

folgen.

**Aufgabe 3.11.** Beweisen Sie Lemma 3.2.

Gegenüber (3.98) hat die Berechnung des reduzierten Gradienten gemäß Lemma 3.2 den Vorteil, dass die Matrix  $\mathbf{A}_D(\mathbf{x})$  nicht invertiert werden muss. Zur Berechnung von  $\bar{\boldsymbol{\lambda}}$  muss nur die lineare Gleichung (3.102b) gelöst werden. Aus Lemma 3.2 folgt direkt, dass die Stationaritätsbedingung  $(df(\mathbf{x})/d\mathbf{x}_I)^T = \mathbf{0}$  genau auf die notwendige Optimalitätsbedingung erster Ordnung für Optimierungsprobleme mit reinen Gleichungsbeschränkungen (siehe Satz 3.4 sowie (3.25)) führt.

In [3.4–3.6] werden Varianten der reduzierten Gradientenmethode beschrieben, die als Grundlage die Problemformulierung (3.4) mit Schlupfvariablen  $\mathbf{x}_s$  verwenden. Dabei werden die Optimierungsvariablen in unabhängige, abhängige und fixierte Variablen partitioniert. Als fixierte Variablen gelten Schlupfvariablen mit dem Wert  $x_{s,i} = 0$ , d. h. jene, die zu einer aktuell aktiven Ungleichungsbeschränkung gehören. Fixierte Variablen haben keinen direkten Einfluss auf den reduzierten Gradienten.

**Aufgabe 3.12.** Lösen Sie das beschränkte Optimierungsproblem (3.40) aus Beispiel 3.3 numerisch mit Hilfe der reduzierten Gradientenmethode. Erstellen Sie dafür ein Computerprogramm, wobei Sie für die Bestimmung der Schrittweite  $\alpha_k$  wahlweise selbst einen Algorithmus (z. B. aus Abschnitt 2.3.1) programmieren oder vorgefertigte Funktionen einsetzen können. Verwenden Sie als Startpunkt  $\mathbf{x}_0 = [1/\sqrt{2} \ 1/\sqrt{2} \ 0]^T$ . Wählen Sie die Partitionierung  $\mathbf{x}_I = [x_2 \ x_3]^T$  und  $x_D = x_1$ . Warum ist diese Aufteilung in Anbetracht des aus Beispiel 3.3 bekannten optimalen Punktes  $\mathbf{x}^* = [1 \ 0 \ 0]^T$  sinnvoll? Vergleichen Sie die Ergebnisse mit jenen von Aufgabe 3.10.

**Lösung von Aufgabe 3.12.** Die Partitionierung  $\mathbf{x}_I = [x_2 \ x_3]^T$  und  $x_D = x_1$  ist die einzig mögliche, die am optimalen Punkt  $\mathbf{x}^*$  die Regularität von  $A_D(\mathbf{x}^*) = \partial g(\mathbf{x}^*)/\partial x_D$  sicherstellt. Am Startpunkt beträgt der Kostenfunktionswert  $f(\mathbf{x}_0) = 1.207107$ . Im Verlauf der ersten Iteration ergeben sich bei Verwendung der Suchrichtung  $\mathbf{s}_{I,k} = -(df(\mathbf{x}_k)/d\mathbf{x}_{I,k})^T$  (einfache Gradientenmethode) folgende Zwischenergebnisse:

$$\mathbf{A}_I(\mathbf{x}_0) = \begin{bmatrix} \sqrt{2} & 0 \end{bmatrix}, \quad A_D(\mathbf{x}_0) = \sqrt{2}, \quad \mathbf{d}_0 = \begin{bmatrix} 1 - \sqrt{2} \\ -1/\sqrt{2} \end{bmatrix},$$

$$\alpha_0 = 0.343906, \quad \mathbf{x}_1 = \begin{bmatrix} 0.788687 \\ 0.564656 \\ -0.243178 \end{bmatrix}, \quad f(\mathbf{x}_1) = 1.088483$$

Nach zehn Iterationen gelangt der Algorithmus zum Punkt

$$\mathbf{x}_{10} = \begin{bmatrix} 1.000\,000 \\ 0.000\,002 \\ 0.000\,001 \end{bmatrix}$$

mit dem Kostenfunktionswert  $f(\mathbf{x}_{10}) = 1.000\,000$ . Ein Vergleich mit Aufgabe 3.10 zeigt, dass für dieses Optimierungsproblem und den hier verwendeten Startpunkt die reduzierte Gradientenmethode deutlich schneller konvergiert als die Gradientenprojektionsmethode.

### 3.2.4 Sequentielle quadratische Programmierung (SQP)

#### 3.2.4.1 Lokales SQP-Verfahren

Für die Motivation des SQP-Verfahrens betrachte man das beschränkte Optimierungsproblem

$$\min_{\mathbf{x} \in \mathbb{R}^n} f(\mathbf{x}) \quad (3.103a)$$

$$\text{u.B.v. } \mathbf{g}(\mathbf{x}) = \mathbf{0} \quad (3.103b)$$

mit  $f \in C^2$  und  $p < n$  Gleichungsbeschränkungen  $g_1(\mathbf{x}), \dots, g_p(\mathbf{x}) \in C^2$ . Nach Satz 3.8 lauten die notwendigen Optimalitätsbedingungen (KKT-Bedingungen) erster Ordnung für einen optimalen Punkt  $(\mathbf{x}^*, \boldsymbol{\lambda}^*)$  mit der Lagrangefunktion  $L(\mathbf{x}, \boldsymbol{\lambda}) = f(\mathbf{x}) + \boldsymbol{\lambda}^T \mathbf{g}(\mathbf{x})$

$$\begin{bmatrix} \left(\frac{\partial}{\partial \mathbf{x}} L\right)^T(\mathbf{x}^*, \boldsymbol{\lambda}^*) \\ \left(\frac{\partial}{\partial \boldsymbol{\lambda}} L\right)^T(\mathbf{x}^*, \boldsymbol{\lambda}^*) \end{bmatrix} = \begin{bmatrix} (\nabla f)(\mathbf{x}^*) + (\nabla \mathbf{g})(\mathbf{x}^*) \boldsymbol{\lambda}^* \\ \mathbf{g}(\mathbf{x}^*) \end{bmatrix} = \mathbf{0}, \quad (3.104)$$

wobei gilt  $(\nabla \mathbf{g})(\mathbf{x}^*) = [(\nabla g_1)(\mathbf{x}^*) \ \dots \ (\nabla g_p)(\mathbf{x}^*)]$ . Eine Möglichkeit, die  $n + p$  Gleichungen (3.104) in den  $n + p$  Unbekannten  $(\mathbf{x}^*, \boldsymbol{\lambda}^*)$  rekursiv numerisch zu lösen, ist das Newton-Raphson Verfahren mit der Iterationsvorschrift

$$\begin{bmatrix} \mathbf{x}_{k+1} \\ \boldsymbol{\lambda}_{k+1} \end{bmatrix} = \begin{bmatrix} \mathbf{x}_k \\ \boldsymbol{\lambda}_k \end{bmatrix} + \begin{bmatrix} \mathbf{p}_{\mathbf{x},k} \\ \mathbf{p}_{\boldsymbol{\lambda},k} \end{bmatrix} \quad (3.105a)$$

$$\underbrace{\begin{bmatrix} \mathbf{L}(\mathbf{x}_k, \boldsymbol{\lambda}_k) & (\nabla \mathbf{g})(\mathbf{x}_k) \\ (\nabla \mathbf{g})^T(\mathbf{x}_k) & \mathbf{0} \end{bmatrix}}_{\mathbf{M}_k} \begin{bmatrix} \mathbf{p}_{\mathbf{x},k} \\ \mathbf{p}_{\boldsymbol{\lambda},k} \end{bmatrix} = - \begin{bmatrix} (\nabla f)(\mathbf{x}_k) + (\nabla \mathbf{g})(\mathbf{x}_k) \boldsymbol{\lambda}_k \\ \mathbf{g}(\mathbf{x}_k) \end{bmatrix} \quad (3.105b)$$

und der Hessematrix  $\mathbf{L}(\mathbf{x}_k, \boldsymbol{\lambda}_k) = \left(\frac{\partial^2}{\partial \mathbf{x}^2} L\right)(\mathbf{x}_k, \boldsymbol{\lambda}_k)$ . Die Matrix  $\mathbf{M}_k$  in (3.105b) hat vollen Rang, d. h. sie kann invertiert werden, wenn die Gleichungsbeschränkungen der LICQ Bedingung genügen ( $(\nabla \mathbf{g})(\mathbf{x}_k)$  ist spaltenregulär) und für alle  $\mathbf{d} \neq \mathbf{0}$  mit der Eigenschaft  $(\nabla \mathbf{g})^T \mathbf{d} = \mathbf{0}$  die Bedingung  $\mathbf{d}^T \mathbf{L}(\mathbf{x}_k, \boldsymbol{\lambda}_k) \mathbf{d} > 0$  erfüllt ist. Wird letztere Bedingung am Punkt  $(\mathbf{x}^*, \boldsymbol{\lambda}^*)$  eingehalten, so gilt dies wegen der getroffenen Differenzierbarkeitsannahmen

auch für Punkte  $(\mathbf{x}_k, \boldsymbol{\lambda}_k)$  in einer hinreichend kleinen Umgebung um  $(\mathbf{x}^*, \boldsymbol{\lambda}^*)$  und gemäß Satz 3.6 ist  $(\mathbf{x}^*, \boldsymbol{\lambda}^*)$  dann ein striktes lokales Minimum. Wenn in einem Iterationsschritt  $\mathbf{p}_{\mathbf{x},k} = \mathbf{0}$  gilt, so ist aus (3.105b) ersichtlich, dass damit auch ein Punkt  $\mathbf{x}^* = \mathbf{x}_{k+1} = \mathbf{x}_k$ ,  $\boldsymbol{\lambda}^* = \boldsymbol{\lambda}_{k+1}$  gefunden wurde, der die KKT-Bedingungen (3.104) des ursprünglichen Optimierungsproblems (3.103) erfüllt.

Ein Schritt  $k$  der Iterationsvorschrift (3.105) kann nun äquivalent auch als *Lösung des quadratischen Programms*

$$\min_{\tilde{\mathbf{p}} \in \mathbb{R}^n} f(\mathbf{x}_k) + (\nabla f)^T(\mathbf{x}_k)\tilde{\mathbf{p}} + \frac{1}{2}\tilde{\mathbf{p}}^T \mathbf{L}(\mathbf{x}_k, \boldsymbol{\lambda}_k)\tilde{\mathbf{p}} \quad (3.106a)$$

$$\text{u.B.v. } (\nabla \mathbf{g})^T(\mathbf{x}_k)\tilde{\mathbf{p}} + \mathbf{g}(\mathbf{x}_k) = \mathbf{0} \quad (3.106b)$$

aufgefasst werden. Die KKT-Bedingungen für (3.106) lauten

$$\begin{bmatrix} \mathbf{L}(\mathbf{x}_k, \boldsymbol{\lambda}_k) & (\nabla \mathbf{g})(\mathbf{x}_k) \\ (\nabla \mathbf{g})^T(\mathbf{x}_k) & \mathbf{0} \end{bmatrix} \begin{bmatrix} \tilde{\mathbf{p}}^* \\ \tilde{\boldsymbol{\lambda}}^* \end{bmatrix} = - \begin{bmatrix} (\nabla f)(\mathbf{x}_k) \\ \mathbf{g}(\mathbf{x}_k) \end{bmatrix} \quad (3.107)$$

mit dem Lagrange-Multiplikator  $\tilde{\boldsymbol{\lambda}}$ . Um diese Äquivalenz zu sehen, wird zunächst  $\mathbf{p}_{\boldsymbol{\lambda},k} = \boldsymbol{\lambda}_{k+1} - \boldsymbol{\lambda}_k$  in (3.105b) eingesetzt, woraus sich

$$\begin{bmatrix} \mathbf{L}(\mathbf{x}_k, \boldsymbol{\lambda}_k) & (\nabla \mathbf{g})(\mathbf{x}_k) \\ (\nabla \mathbf{g})^T(\mathbf{x}_k) & \mathbf{0} \end{bmatrix} \begin{bmatrix} \mathbf{p}_{\mathbf{x},k} \\ \boldsymbol{\lambda}_{k+1} \end{bmatrix} = - \begin{bmatrix} (\nabla f)(\mathbf{x}_k) \\ \mathbf{g}(\mathbf{x}_k) \end{bmatrix} \quad (3.108)$$

ergibt. Ein Vergleich von (3.107) mit (3.108) bestätigt nun, dass statt der Lösung des Gleichungssystems (3.108) auch das Minimum des quadratischen Programms (3.106) berechnet werden kann. Die eigentliche Iterationsvorschrift, welche äquivalent zu (3.105) ist, lautet damit  $\boldsymbol{\lambda}_{k+1} = \tilde{\boldsymbol{\lambda}}^*$  und  $\mathbf{x}_{k+1} = \mathbf{x}_k + \tilde{\mathbf{p}}^*$  (mit  $\mathbf{p}_{\mathbf{x},k} = \tilde{\mathbf{p}}^*$ ). Wenn nun in einem Iterationsschritt  $k$  für (3.106) die Lösung  $\tilde{\mathbf{p}}^* = \mathbf{0}$  gefunden wird, so ist aus (3.107) ersichtlich, dass damit auch ein Punkt  $\mathbf{x}^* = \mathbf{x}_k$ ,  $\boldsymbol{\lambda}^* = \tilde{\boldsymbol{\lambda}}^*$  gefunden wurde, der die KKT-Bedingungen (3.104) des ursprünglichen Optimierungsproblems (3.103) erfüllt. Da wiederkehrend quadratische Programme gelöst werden, bezeichnet man dieses Verfahren als *sequentielle quadratische Programmierung*.

Die vorangegangenen Überlegungen motivieren die Erweiterung der SQP-Methode auf allgemeine nichtlineare Optimierungsprobleme der Form

$$\min_{\mathbf{x} \in \mathbb{R}^n} f(\mathbf{x}) \quad (3.109a)$$

$$\text{u.B.v. } \mathbf{g}(\mathbf{x}) = \mathbf{0} \quad (3.109b)$$

$$\mathbf{h}(\mathbf{x}) \leq \mathbf{0} \quad (3.109c)$$

mit  $f \in C^2$ ,  $p < n$  Gleichungsbeschränkungen  $g_1(\mathbf{x}), \dots, g_p(\mathbf{x}) \in C^2$  und  $q$  Ungleichungsbeschränkungen  $h_1(\mathbf{x}), \dots, h_q(\mathbf{x}) \in C^2$ . Die zugehörige Lagrangefunktion lautet  $L(\mathbf{x}, \boldsymbol{\lambda}, \boldsymbol{\mu}) = f(\mathbf{x}) + \boldsymbol{\lambda}^T \mathbf{g}(\mathbf{x}) + \boldsymbol{\mu}^T \mathbf{h}(\mathbf{x})$ . Das Optimierungsproblem (3.109) wird in jedem Iterationsschritt durch das *quadratische Programm*

$$\min_{\tilde{\mathbf{p}} \in \mathbb{R}^n} f(\mathbf{x}_k) + (\nabla f)^T(\mathbf{x}_k)\tilde{\mathbf{p}} + \frac{1}{2}\tilde{\mathbf{p}}^T \mathbf{L}(\mathbf{x}_k, \boldsymbol{\lambda}_k, \boldsymbol{\mu}_k)\tilde{\mathbf{p}} \quad (3.110a)$$

$$\text{u.B.v. } (\nabla \mathbf{g})^\top(\mathbf{x}_k)\tilde{\mathbf{p}} + \mathbf{g}(\mathbf{x}_k) = \mathbf{0} \quad (3.110b)$$

$$(\nabla \mathbf{h})^\top(\mathbf{x}_k)\tilde{\mathbf{p}} + \mathbf{h}(\mathbf{x}_k) \leq \mathbf{0} \quad (3.110c)$$

mit  $\mathbf{L}(\mathbf{x}_k, \boldsymbol{\lambda}_k, \boldsymbol{\mu}_k) = \left(\frac{\partial^2}{\partial \mathbf{x}^2} L\right)(\mathbf{x}_k, \boldsymbol{\lambda}_k, \boldsymbol{\mu}_k)$  approximiert. Die KKT-Bedingungen für (3.110) lauten (siehe Satz 3.8)

$$(\nabla f)(\mathbf{x}_k) + \mathbf{L}(\mathbf{x}_k, \boldsymbol{\lambda}_k, \boldsymbol{\mu}_k)\tilde{\mathbf{p}}^* + (\nabla \mathbf{g})(\mathbf{x}_k)\tilde{\boldsymbol{\lambda}}^* + (\nabla \mathbf{h})(\mathbf{x}_k)\tilde{\boldsymbol{\mu}}^* = \mathbf{0} \quad (3.111a)$$

$$\tilde{\boldsymbol{\mu}}^* \geq \mathbf{0} \quad (3.111b)$$

$$\left((\nabla \mathbf{h})^\top(\mathbf{x}_k)\tilde{\mathbf{p}}^* + \mathbf{h}(\mathbf{x}_k)\right)^\top \tilde{\boldsymbol{\mu}}^* = 0 \quad (3.111c)$$

$$(\nabla \mathbf{g})^\top(\mathbf{x}_k)\tilde{\mathbf{p}}^* + \mathbf{g}(\mathbf{x}_k) = \mathbf{0} \quad (3.111d)$$

$$(\nabla \mathbf{h})^\top(\mathbf{x}_k)\tilde{\mathbf{p}}^* + \mathbf{h}(\mathbf{x}_k) \leq \mathbf{0} \quad (3.111e)$$

mit den Lagrange-Multiplikatoren  $\tilde{\boldsymbol{\lambda}}$  und  $\tilde{\boldsymbol{\mu}}$ . Die Iterationsvorschrift des SQP-Verfahrens lautet analog zum gleichungsbeschränkten Fall  $\mathbf{x}_{k+1} = \mathbf{x}_k + \tilde{\mathbf{p}}^*$ ,  $\boldsymbol{\lambda}_{k+1} = \tilde{\boldsymbol{\lambda}}^*$  und  $\boldsymbol{\mu}_{k+1} = \tilde{\boldsymbol{\mu}}^*$ . Ergibt sich in einem Iterationsschritt  $k$  für (3.110) die Lösung  $\tilde{\mathbf{p}}^* = \mathbf{0}$ , so folgt aus (3.111), dass damit ein Punkt  $\mathbf{x}^* = \mathbf{x}_k$ ,  $\boldsymbol{\lambda}^* = \tilde{\boldsymbol{\lambda}}^*$ ,  $\boldsymbol{\mu}^* = \tilde{\boldsymbol{\mu}}^*$  gefunden wurde, der die KKT-Bedingungen des ursprünglichen Optimierungsproblems (3.109) erfüllt.

Unter bestimmten Voraussetzungen kann eine quadratische Konvergenzordnung (vergleiche Satz 2.11) des SQP-Verfahrens gegen  $(\mathbf{x}^*, \boldsymbol{\lambda}^*, \boldsymbol{\mu}^*)$  gezeigt werden. Allerdings gilt diese Aussage im Allgemeinen nur für Startwerte  $(\mathbf{x}_0, \boldsymbol{\lambda}_0, \boldsymbol{\mu}_0)$  in einer hinreichend kleinen Umgebung um  $(\mathbf{x}^*, \boldsymbol{\lambda}^*, \boldsymbol{\mu}^*)$ . Man spricht deshalb auch vom *lokalen SQP-Verfahren*, welches in Tabelle 3.3 zusammengefasst ist.

---

<b>Initialisierung:</b>	$\mathbf{x}_0$	(Zulässiger Startpunkt)
	$\boldsymbol{\lambda}_0, \boldsymbol{\mu}_0$	(Startwerte der Lagrange-Multiplikatoren)
	$k = 0$	(Startindex)
	$\varepsilon$	(Abbruchkriterium)

**repeat**

Schritt 1: Berechne  $f(\mathbf{x}_k)$ ,  $(\nabla f)(\mathbf{x}_k)$ ,  $\mathbf{L}(\mathbf{x}_k, \boldsymbol{\lambda}_k, \boldsymbol{\mu}_k)$ ,  $\mathbf{g}(\mathbf{x}_k)$ ,  $(\nabla \mathbf{g})(\mathbf{x}_k)$ ,  $\mathbf{h}(\mathbf{x}_k)$ ,  $(\nabla \mathbf{h})(\mathbf{x}_k)$ .

Schritt 2: Berechne  $\tilde{\mathbf{p}}^*$ ,  $\tilde{\boldsymbol{\lambda}}^*$ ,  $\tilde{\boldsymbol{\mu}}^*$  durch Lösen des Optimierungsproblems (3.110).

Schritt 3: Setze  $\mathbf{x}_{k+1} \leftarrow \mathbf{x}_k + \tilde{\mathbf{p}}^*$ ,  $\boldsymbol{\lambda}_{k+1} \leftarrow \tilde{\boldsymbol{\lambda}}^*$ ,  $\boldsymbol{\mu}_{k+1} \leftarrow \tilde{\boldsymbol{\mu}}^*$ ,  $k \leftarrow k + 1$ .

**until**  $\|\tilde{\mathbf{p}}^*\| \leq \varepsilon$

---

Tabelle 3.3: Lokales SQP-Verfahren.

Das quadratische Programm (3.110) für einen Iterationsschritt  $k$  kann beispielsweise über die Methode der aktiven Beschränkungen (siehe Abschnitt 3.2.1) gelöst werden. Zur Veranschaulichung eines solchen Iterationsschritts betrachte man das nachfolgende Beispiel.

*Beispiel 3.5.* Gegeben ist das quadratische Optimierungsproblem

$$\min_{\tilde{\mathbf{p}} \in \mathbb{R}^2} \frac{1}{2} \tilde{\mathbf{p}}^T \mathbf{H} \tilde{\mathbf{p}} + \mathbf{c}^T \tilde{\mathbf{p}} \quad (3.112a)$$

$$\text{u.B.v. } \mathbf{a}_1^T \tilde{\mathbf{p}} - b_1 = -\tilde{p}_1 + 2\tilde{p}_2 - 2 \leq 0 \quad \text{Ungleichungsbeschr. U1} \quad (3.112b)$$

$$\mathbf{a}_2^T \tilde{\mathbf{p}} - b_2 = \tilde{p}_1 + 2\tilde{p}_2 - 6 \leq 0 \quad \text{Ungleichungsbeschr. U2} \quad (3.112c)$$

$$\mathbf{a}_3^T \tilde{\mathbf{p}} - b_3 = \tilde{p}_1 - 2\tilde{p}_2 - 2 \leq 0 \quad \text{Ungleichungsbeschr. U3} \quad (3.112d)$$

$$\mathbf{a}_4^T \tilde{\mathbf{p}} - b_4 = -\tilde{p}_1 \leq 0 \quad \text{Ungleichungsbeschr. U4} \quad (3.112e)$$

$$\mathbf{a}_5^T \tilde{\mathbf{p}} - b_5 = -\tilde{p}_2 \leq 0 \quad \text{Ungleichungsbeschr. U5} \quad (3.112f)$$

mit  $\mathbf{H} = \begin{bmatrix} 2 & 0 \\ 0 & 2 \end{bmatrix}$  und  $\mathbf{c} = [-2 \quad -5]^T$ . Als Startpunkt für die Methode der aktiven

Beschränkungen wählt man den zulässigen Punkt  $\tilde{\mathbf{p}}_0 = [2 \quad 0]^T$ , an dem die Ungleichungsbeschränkungen U3 und U5 aktiv sind. Die Arbeitsmenge  $W_0$  der aktiven Ungleichungsbeschränkungen lautet damit  $W_0 = \{3, 5\}$ . Der nächste Iterationspunkt  $\tilde{\mathbf{p}}_{j+1}$  wird in der Form  $\tilde{\mathbf{p}}_{j+1} = \tilde{\mathbf{p}}_j + \tilde{\mathbf{s}}_j^*$  angesetzt. Der optimale Schritt  $\tilde{\mathbf{s}}_j^*$  folgt aus

$$\min_{\tilde{\mathbf{s}}_j \in \mathbb{R}^2} \frac{1}{2} (\tilde{\mathbf{p}}_j + \tilde{\mathbf{s}}_j)^T \mathbf{H} (\tilde{\mathbf{p}}_j + \tilde{\mathbf{s}}_j) + \mathbf{c}^T (\tilde{\mathbf{p}}_j + \tilde{\mathbf{s}}_j) \quad (3.113a)$$

$$\text{u.B.v. } \mathbf{a}_w^T (\tilde{\mathbf{p}}_j + \tilde{\mathbf{s}}_j) - b_w = \mathbf{a}_w^T \tilde{\mathbf{s}}_j = 0, \quad \forall w \in W_j. \quad (3.113b)$$

Die KKT-Bedingungen für (3.113) lauten mit der Matrix  $\mathbf{A}_j$ , deren Spalten durch  $\mathbf{a}_w$ ,  $w \in W_j$ , gegeben sind, und den Lagrange-Multiplikatoren  $\tilde{\boldsymbol{\nu}}_j$

$$\mathbf{H} \tilde{\mathbf{s}}_j^* + \mathbf{A}_j \tilde{\boldsymbol{\nu}}_j^* = -\mathbf{H} \tilde{\mathbf{p}}_j - \mathbf{c} \quad (3.114a)$$

$$\mathbf{a}_w^T \tilde{\mathbf{s}}_j^* = 0, \quad \forall w \in W_j. \quad (3.114b)$$

Für  $j = 0$  erhält man als Lösung von (3.114)

$$\begin{bmatrix} 2 & 0 & 1 & 0 \\ 0 & 2 & -2 & -1 \\ 1 & -2 & 0 & 0 \\ 0 & -1 & 0 & 0 \end{bmatrix} \begin{bmatrix} \tilde{\mathbf{s}}_0^* \\ \tilde{\boldsymbol{\nu}}_0^* \end{bmatrix} = \begin{bmatrix} -2 \\ 5 \\ 0 \\ 0 \end{bmatrix} \quad (3.115)$$

die Größen  $\tilde{\mathbf{s}}_0^* = \mathbf{0}$  und  $\tilde{\boldsymbol{\nu}}_0^* = [-2 \quad -1]^T$  und somit  $\tilde{\mathbf{p}}_1 = \tilde{\mathbf{p}}_0 = [2 \quad 0]^T$ . Nun wird die Ungleichung U3 (Lagrange-Multiplikator mit negativstem Wert) inaktiv gesetzt ( $W_1 = \{5\}$ ) und (3.114) erneut für  $j = 1$  gelöst, d. h. aus

$$\begin{bmatrix} 2 & 0 & 0 \\ 0 & 2 & -1 \\ 0 & -1 & 0 \end{bmatrix} \begin{bmatrix} \tilde{\mathbf{s}}_1^* \\ \tilde{\boldsymbol{\nu}}_1^* \end{bmatrix} = \begin{bmatrix} -2 \\ 5 \\ 0 \end{bmatrix} \quad (3.116)$$

folgen  $\tilde{\mathbf{s}}_1^* = [-1 \ 0]^T$  und  $\tilde{\nu}_1^* = -5$ . Damit kann der neue Iterationspunkt  $\tilde{\mathbf{p}}_2$  zu  $\tilde{\mathbf{p}}_2 = \tilde{\mathbf{p}}_1 + \tilde{\mathbf{s}}_1^*$  berechnet werden, vorausgesetzt es werden keine inaktiven Ungleichungsbeschränkungen verletzt. Um diesen Fall zu berücksichtigen, wird eine Schrittweite  $\beta_j > 0$  verwendet, so dass  $\tilde{\mathbf{p}}_{j+1} = \tilde{\mathbf{p}}_j + \beta_j \tilde{\mathbf{s}}_j^*$  gilt. Die Wahl der Schrittweite  $\beta_j$  erfolgt nun auf Basis folgender Überlegungen. Wenn für alle inaktiven (affinen) Ungleichungsbeschränkungen gilt  $\mathbf{a}_i^T \tilde{\mathbf{s}}_j^* \leq 0$ , dann kann  $\beta_j > 0$  beliebig gewählt werden ohne eine Ungleichung  $\mathbf{a}_i^T (\tilde{\mathbf{p}}_j + \beta_j \tilde{\mathbf{s}}_j^*) \leq b_i$ ,  $i \notin W_j$  zu verletzen. Falls hingegen  $\mathbf{a}_i^T \tilde{\mathbf{s}}_j^* > 0$ , dann ist die Schrittweite durch  $\beta_j \leq \frac{b_i - \mathbf{a}_i^T \tilde{\mathbf{p}}_j}{\mathbf{a}_i^T \tilde{\mathbf{s}}_j^*}$  begrenzt. Da das Minimum der quadratischen Funktion für  $\beta_j = 1$  erreicht wird, folgt die Wahl der Schrittweite zu

$$\beta_j = \min \left\{ 1, \min_{i \notin W_j, \mathbf{a}_i^T \tilde{\mathbf{s}}_j^* > 0} \frac{b_i - \mathbf{a}_i^T \tilde{\mathbf{p}}_j}{\mathbf{a}_i^T \tilde{\mathbf{s}}_j^*} \right\}. \quad (3.117)$$

Im vorliegenden Fall gilt  $\beta_1 = \min \left\{ 1, \underbrace{4}_{U1}, \underbrace{2}_{U4} \right\} = 1$  und damit  $\tilde{\mathbf{p}}_2 = [1 \ 0]^T$ .

Da die optimale Schrittweite  $\beta_1 = 1$  möglich ist, muss (3.114) nicht erneut gelöst werden. Es kann direkt die Ungleichung U5 (Lagrange-Multiplikator ist negativ) inaktiv gesetzt ( $W_2 = \{ \}$ ) und das unbeschränkte Optimierungsproblem zu  $\tilde{\mathbf{s}}_2^* = [0 \ 2.5]^T$  gelöst werden. Die maximale Schrittweite gemäß (3.117) errechnet sich zu  $\beta_2 = \min \left\{ 1, \underbrace{0.6}_{U1}, \underbrace{1}_{U2} \right\} = 0.6$  und damit folgt  $\tilde{\mathbf{p}}_3 = [1 \ 1.5]^T$ . Man erkennt

nun, dass die Ungleichung U1 aktiv ist (der Wert  $\beta_2 = 0.6$  wurde gerade durch die Ungleichungsbeschränkung U1 bestimmt), weshalb die Arbeitsmenge der aktiven Beschränkungen zu  $W_3 = \{1\}$  gesetzt wird. Aus (3.114) folgt mit

$$\begin{bmatrix} 2 & 0 & -1 \\ 0 & 2 & 2 \\ -1 & 2 & 0 \end{bmatrix} \begin{bmatrix} \tilde{\mathbf{s}}_3^* \\ \tilde{\nu}_3^* \end{bmatrix} = \begin{bmatrix} 0 \\ 2 \\ 0 \end{bmatrix} \quad (3.118)$$

die Lösung  $\tilde{\mathbf{s}}_3^* = [0.4 \ 0.2]^T$  und  $\tilde{\nu}_3^* = 0.8$ . Die Schrittweite  $\beta_3$  folgt aus der Beziehung (3.117) zu  $\beta_3 = \min \left\{ 1, \underbrace{2.5}_{U2} \right\} = 1$ . Daher und auf Grund des positiven Lagrange-

Multiplikators  $\tilde{\nu}_3^* = 0.8$  stellt  $\tilde{\mathbf{p}}^* = \tilde{\mathbf{p}}_3 + \beta_3 \tilde{\mathbf{s}}_3^* = [1.4 \ 1.7]^T$  die optimale Lösung von (3.112) dar. Der Lagrange-Multiplikator für das quadratische Programm (3.112) lautet  $\tilde{\boldsymbol{\mu}}^* = [\tilde{\nu}_3^* \ 0 \ 0 \ 0 \ 0]^T$ .

Für die Formulierung des quadratischen Programms (3.110) wird in jedem Iterationsschritt die Hessematrix der Lagrangefunktion  $L(\mathbf{x}_k, \boldsymbol{\lambda}_k, \boldsymbol{\mu}_k)$  benötigt. Damit können folgende Probleme verbunden sein. Die exakte Hessematrix ist in vielen Anwendungen nicht bekannt und ihre näherungsweise numerische Berechnung mit finiten Differenzen (vgl. Abschnitt 1.3.4) kann aufwändig und ungenau sein. Ferner kann die Hessematrix indefinit sein, was insbesondere dann vorkommt, wenn das Verfahren nicht in einer hinreichend kleinen Umgebung um  $(\mathbf{x}^*, \boldsymbol{\lambda}^*, \boldsymbol{\mu}^*)$  gestartet wird. Eine indefinite Hessematrix erschwert

die Lösung des quadratischen Programms. Aus diesen Gründen ersetzt man in der Praxis die Hessematrix  $\mathbf{L}(\mathbf{x}_k, \boldsymbol{\lambda}_k, \boldsymbol{\mu}_k)$  beim SQP-Verfahren häufig durch eine geeignete *positiv definite Approximation*  $\mathbf{H}_k$ . Für die Berechnung von  $\mathbf{H}_k$  in jedem Iterationsschritt  $k$  kann in Analogie zur Quasi-Newton-Methode die *modifizierte BFGS Methode* (siehe z. B. [3.7])

$$\mathbf{H}_{k+1} = \mathbf{H}_k + \frac{\mathbf{q}_k \mathbf{q}_k^T}{\mathbf{q}_k^T \mathbf{d}_k} - \frac{\mathbf{H}_k \mathbf{d}_k \mathbf{d}_k^T \mathbf{H}_k}{\mathbf{d}_k^T \mathbf{H}_k \mathbf{d}_k} \quad (3.119a)$$

mit

$$\mathbf{d}_k = \mathbf{x}_{k+1} - \mathbf{x}_k \quad (3.119b)$$

$$\mathbf{y}_k = \left( \frac{\partial}{\partial \mathbf{x}} L \right)^T (\mathbf{x}_{k+1}, \boldsymbol{\lambda}_k, \boldsymbol{\mu}_k) - \left( \frac{\partial}{\partial \mathbf{x}} L \right)^T (\mathbf{x}_k, \boldsymbol{\lambda}_k, \boldsymbol{\mu}_k) \quad (3.119c)$$

$$\theta_k = \begin{cases} 1, & \text{wenn } \mathbf{d}_k^T \mathbf{y}_k \geq 0.2 \mathbf{d}_k^T \mathbf{H}_k \mathbf{d}_k \\ \frac{0.8 \mathbf{d}_k^T \mathbf{H}_k \mathbf{d}_k}{\mathbf{d}_k^T \mathbf{H}_k \mathbf{d}_k - \mathbf{d}_k^T \mathbf{y}_k} & \text{sonst} \end{cases} \quad (3.119d)$$

$$\mathbf{q}_k = \theta_k \mathbf{y}_k + (1 - \theta_k) \mathbf{H}_k \mathbf{d}_k \quad (3.119e)$$

verwendet werden. Sie wird auch *gedämpfte BFGS Methode* genannt. Man beachte, dass hier direkt die Hessematrix und nicht deren Inverse wie in Abschnitt 2.3.2.4 approximiert wird. Unter Verwendung von (3.119) ist garantiert, dass  $\mathbf{H}_{k+1}$  symmetrisch und positiv definit ist, wenn  $\mathbf{H}_k$  symmetrisch und positiv definit war. Damit kann das lokale SQP-Verfahren gemäß Tabelle 3.3 dahingehend modifiziert werden, dass ausgehend von einer symmetrischen, positiv definiten Matrix  $\mathbf{H}_0$  für  $k > 0$  in Schritt 1 des Verfahrens  $\mathbf{H}_k$  gemäß (3.119) statt  $\mathbf{L}(\mathbf{x}_k, \boldsymbol{\lambda}_k, \boldsymbol{\mu}_k)$  berechnet wird. Dadurch verschlechtert sich zwar das Konvergenzverhalten des Verfahrens, es kann aber zumindest noch superlineare Konvergenz in einer hinreichend kleinen Umgebung um  $(\mathbf{x}^*, \boldsymbol{\lambda}^*, \boldsymbol{\mu}^*)$  nachgewiesen werden.

### 3.2.4.2 Globalisierung des SQP-Verfahrens

Im Allgemeinen konvergiert der SQP-Algorithmus gemäß Tabelle 3.3 nur für Startwerte  $(\mathbf{x}_0, \boldsymbol{\lambda}_0, \boldsymbol{\mu}_0)$ , die in einer hinreichend kleinen Umgebung um  $(\mathbf{x}^*, \boldsymbol{\lambda}^*, \boldsymbol{\mu}^*)$  liegen. Um zu erreichen, dass das SQP-Verfahren (idealerweise) für beliebige Startwerte konvergiert, führt man eine Globalisierung des Verfahrens durch. In Analogie zur Newton-Methode (siehe Abschnitt 2.3.2.3) wird dies durch die Einführung einer Schrittweite  $\alpha_k > 0$  erzielt. In Schritt 3 des Algorithmus berechnet man  $\mathbf{x}_{k+1}$  daher zu

$$\mathbf{x}_{k+1} = \mathbf{x}_k + \alpha_k \tilde{\mathbf{p}}^* \quad (3.120)$$

Die Schrittweite  $\alpha_k$  folgt aus einem geeignet formulierten Liniensuchproblem. Die Kostenfunktion dieses Liniensuchproblems muss eine Bewertung ermöglichen, ob  $\mathbf{x}_{k+1}$  *besser* ist als  $\mathbf{x}_k$ . Während diese Bewertung im unbeschränkten Fall direkt anhand der Kostenfunktionswerte an den Punkten  $\mathbf{x}_k$  und  $\mathbf{x}_{k+1}$  möglich ist, trifft dies für ein beschränktes Optimierungsproblem im Allgemeinen nicht mehr zu. Insbesondere gilt beim SQP-Verfahren im Allgemeinen nicht mehr strikt  $\mathbf{x}_k \in \mathcal{X}$ . Damit kann  $\mathbf{x}_{k+1}$  zwar den Kostenfunktionswert verbessern aber möglicherweise zu einer stärkeren Verletzung von Beschränkungen führen.

Um für die Wahl von  $\alpha_k$  einen guten Kompromiss zu finden, verwendet man daher eine so genannte *Bewertungsfunktion* (englisch: *merit function*). Eine gängige Wahl dafür ist die Funktion

$$l(\mathbf{x}, \eta) = f(\mathbf{x}) + \eta \left( \sum_{i=1}^p |g_i(\mathbf{x})| + \sum_{i=1}^q \max\{0, h_i(\mathbf{x})\} \right) \quad (3.121)$$

mit einem Gewichtungsfaktor  $\eta > 0$ . Damit folgt die optimale Schrittweite  $\alpha_k$  aus dem Liniensuchproblem

$$\alpha_k = \arg \min_{\alpha} l(\mathbf{x}_k + \alpha \tilde{\mathbf{p}}^*, \eta). \quad (3.122)$$

In der Praxis wird  $\alpha_k$  so gewählt, dass mit  $l(\mathbf{x}_k + \alpha_k \tilde{\mathbf{p}}^*, \eta)$  eine hinreichende Verbesserung gegenüber  $l(\mathbf{x}_k, \eta)$  erreicht wird. Dies kann beispielsweise mit einem der Verfahren zur Schrittweitenwahl aus Abschnitt 2.3.1 erfolgen. Üblicherweise ist eine Anpassung von  $\eta$  in jedem Iterationsschritt des SQP-Verfahrens erforderlich (vgl. [3.4]). In seltenen Fällen besitzt die Bewertungsfunktion gemäß (3.121) die gute Eigenschaft, dass ein lokales Minimum  $\mathbf{x}^*$  von (3.109) auch ein lokales Minimum von  $l(\mathbf{x}, \eta)$  ist. Man spricht dann auch von einer *exakten Bewertungsfunktion*.

### 3.2.5 Methode der Straf- und Barrierefunktionen

Mit Hilfe von Straf- und Barrierefunktionen lassen sich beschränkte in unbeschränkte Optimierungsprobleme überführen, welche dann z. B. mit den in Abschnitt 2 beschriebenen Methoden gelöst werden können.

#### 3.2.5.1 Straffunktionen

Die grundlegende Idee der Methode der Straffunktionen besteht darin, das *beschränkte Optimierungsproblem*

$$\min_{\mathbf{x} \in \mathcal{X}} f(\mathbf{x}) \quad (3.123)$$

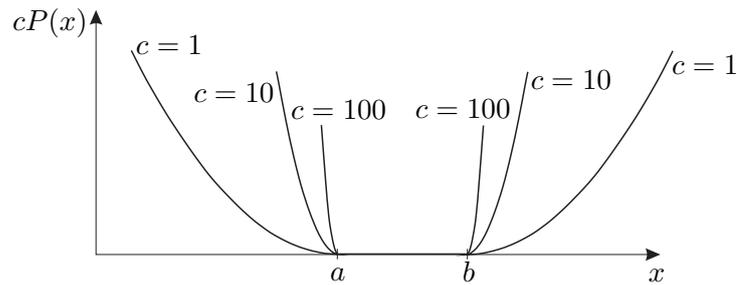
mit der zulässigen Menge  $\mathcal{X}$  in ein *unbeschränktes Optimierungsproblem* der Form

$$\min_{\mathbf{x} \in \mathbb{R}^n} f(\mathbf{x}) + cP(\mathbf{x}) \quad (3.124)$$

mit einem positiven Gewichtsparameter  $c$  und der stetigen Funktion  $P(\mathbf{x})$  überzuführen. Die Funktion  $P(\mathbf{x})$  wird als *Straffunktion* bezeichnet und besitzt die Eigenschaft, dass  $P(\mathbf{x}) \geq 0$  für alle  $\mathbf{x} \in \mathbb{R}^n$  und  $P(\mathbf{x}) = 0$  genau dann, wenn  $\mathbf{x} \in \mathcal{X}$ . Die zulässige Menge  $\mathcal{X}$  ist typischerweise implizit über Gleichungs- und Ungleichungsbeschränkungen definiert. Für  $\mathcal{X} = \{\mathbf{x} \in \mathbb{R}^n \mid g_i(\mathbf{x}) = 0, i = 1, \dots, p, h_j(\mathbf{x}) \leq 0, j = 1, \dots, q\}$  kann als Straffunktion beispielsweise

$$P(\mathbf{x}) = \frac{1}{2} \sum_{i=1}^p (g_i(\mathbf{x}))^2 + \frac{1}{2} \sum_{i=1}^q (\max\{0, h_i(\mathbf{x})\})^2 \quad (3.125)$$

verwendet werden. Abbildung 3.7 zeigt für einen eindimensionalen Fall den Verlauf der Straffunktionen  $cP(x)$  für unterschiedliche Werte von  $c > 0$  und zwei Ungleichungsbeschränkungen mit  $h_1(x) = x - b$  und  $h_2(x) = a - x$ .

Abbildung 3.7: Verlauf der Straffunktionen  $cP(x)$ .

Für größere Werte von  $c$  ist zu erwarten, dass die Lösung des unbeschränkten Optimierungsproblems (3.124) zumindest in der Nähe des zulässigen Gebiets  $\mathcal{X}$  zu liegen kommt und für  $c \rightarrow \infty$  wird die Lösung von (3.124) gegen jene von (3.123) konvergieren. Dabei wird so vorgegangen, dass für eine gegen Unendlich strebende Folge  $\{c_l\}$ ,  $l = 1, 2, \dots$  mit  $c_1 > 0$  und  $c_{l+1} > c_l$  das unbeschränkte Optimierungsproblem (3.124) wiederkehrend mit  $c = c_l$  gelöst wird und man jeweils einen optimalen Punkt  $\mathbf{x}_l^*$  erhält. Für die Lösung des Optimierungsproblems (3.124) mittels numerischer Verfahren bietet es sich an, den Punkt  $\mathbf{x}_l^*$  als Startpunkt für die Optimierungsaufgabe mit  $c_{l+1}$  zu verwenden. Bei der Methode der Straffunktionen gelten folgende Hilfssätze.

**Lemma 3.3 (Ungleichungen bei der Methode der Straffunktionen).** Für  $c_{l+1} > c_l > 0$  und den zugehörigen Lösungen  $\mathbf{x}_l^*$  und  $\mathbf{x}_{l+1}^*$  des unbeschränkten Optimierungsproblems (3.124) gelten folgende Ungleichungen

$$f(\mathbf{x}_l^*) + c_l P(\mathbf{x}_l^*) \leq f(\mathbf{x}_{l+1}^*) + c_{l+1} P(\mathbf{x}_{l+1}^*) \quad (3.126a)$$

$$P(\mathbf{x}_l^*) \geq P(\mathbf{x}_{l+1}^*) \quad (3.126b)$$

$$f(\mathbf{x}_l^*) \leq f(\mathbf{x}_{l+1}^*) . \quad (3.126c)$$

*Beweis.* Aufgrund von  $c_{l+1} > c_l$  und der Definitionen von  $\mathbf{x}_l^*$  und  $\mathbf{x}_{l+1}^*$  gilt unmittelbar

$$f(\mathbf{x}_{l+1}^*) + c_{l+1} P(\mathbf{x}_{l+1}^*) \geq f(\mathbf{x}_{l+1}^*) + c_l P(\mathbf{x}_{l+1}^*) \geq f(\mathbf{x}_l^*) + c_l P(\mathbf{x}_l^*), \quad (3.127)$$

womit (3.126a) gezeigt ist. Aus

$$-f(\mathbf{x}_{l+1}^*) - c_l P(\mathbf{x}_{l+1}^*) \leq -f(\mathbf{x}_l^*) - c_l P(\mathbf{x}_l^*) \quad (3.128a)$$

$$f(\mathbf{x}_{l+1}^*) + c_{l+1} P(\mathbf{x}_{l+1}^*) \leq f(\mathbf{x}_l^*) + c_{l+1} P(\mathbf{x}_l^*) \quad (3.128b)$$

folgt

$$(c_{l+1} - c_l) P(\mathbf{x}_{l+1}^*) \leq (c_{l+1} - c_l) P(\mathbf{x}_l^*) \quad (3.129)$$

und daher (3.126b). Aus (3.127) erhält man

$$f(\mathbf{x}_{l+1}^*) + c_l \underbrace{(P(\mathbf{x}_{l+1}^*) - P(\mathbf{x}_l^*))}_{\leq 0} \geq f(\mathbf{x}_l^*), \quad (3.130)$$

woraus sich schließlich (3.126c) ergibt.  $\square$

**Lemma 3.4 (Methode der Straffunktionen).** Wenn  $\mathbf{x}^*$  die Lösung des beschränkten Optimierungsproblems (3.123) ist, dann gilt für jedes  $l$  der Folge  $\{c_l\}$

$$f(\mathbf{x}^*) \geq f(\mathbf{x}_l^*) + c_l P(\mathbf{x}_l^*) \geq f(\mathbf{x}_l^*) . \quad (3.131)$$

**Aufgabe 3.13.** Beweisen Sie Lemma 3.4.

**Satz 3.13 (Konvergenz der Methode der Straffunktionen).** Angenommen,  $\{\mathbf{x}_l^*\}$  sei eine Folge von Punkten, die durch die Lösung des unbeschränkten Optimierungsproblems (3.124) für eine gegen Unendlich strebende Folge  $\{c_l\}$ ,  $l = 1, 2, \dots$  mit  $c_1 > 0$  und  $c_{l+1} > c_l$  erhalten wurde. Dann ist jeder Grenzwert der Folge  $\{\mathbf{x}_l^*\}$  eine Lösung des beschränkten Optimierungsproblems (3.123).

### 3.2.5.2 Barrierefunktionen

Barriermethoden sind auf das beschränkte Optimierungsproblem (3.123) dann anwendbar, wenn die zulässige Menge  $\mathcal{X}$  eine *robuste Menge* ist (siehe Abbildung 3.1). Folglich können Gleichungsbeschränkungen  $\mathbf{g}(\mathbf{x}) = \mathbf{0}$  nicht durch Barrierefunktionen ersetzt werden. Es sei nun  $\mathcal{X}$  das (nichtleere) Innere von  $\check{\mathcal{X}}$ .

Eine *Barrierefunktion*  $B(\mathbf{x})$  ist auf  $\check{\mathcal{X}}$  definiert und hat die Eigenschaften, dass sie stetig ist, dass sie für alle  $\mathbf{x} \in \check{\mathcal{X}}$  nichtnegativ ist und dass  $B(\mathbf{x}) \rightarrow \infty$  gilt wenn sich  $\mathbf{x}$  dem Rand von  $\mathcal{X}$  nähert.

Angenommen,  $\mathcal{X} = \{\mathbf{x} \in \mathbb{R}^n \mid h_i(\mathbf{x}) \leq 0, i = 1, \dots, q\}$  sei eine robuste Menge mit dem Inneren  $\check{\mathcal{X}} = \{\mathbf{x} \in \mathbb{R}^n \mid h_i(\mathbf{x}) < 0, i = 1, \dots, q\}$ , dann kann als Barrierefunktion beispielsweise

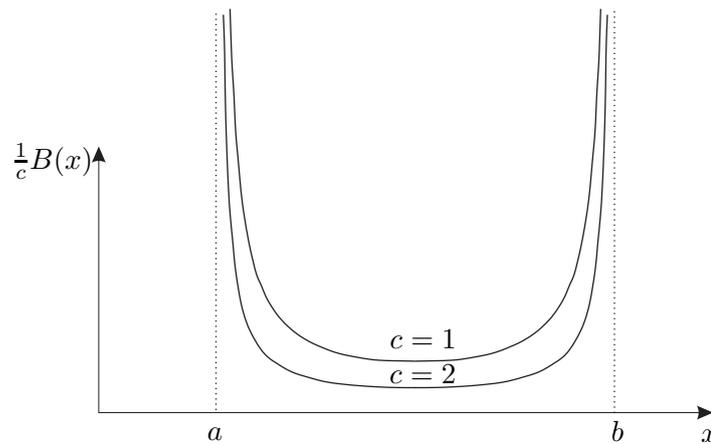
$$B(\mathbf{x}) = - \sum_{i=1}^q \frac{1}{h_i(\mathbf{x})} \quad (3.132)$$

verwendet werden. Abbildung 3.8 zeigt für den eindimensionalen Fall den Verlauf der Barrierefunktion  $\frac{1}{c}B(x)$  für unterschiedliche Werte des Gewichtungsparameters  $c > 0$  und zwei Ungleichungsbeschränkungen mit  $h_1(x) = x - b$  und  $h_2(x) = a - x$ . Eine alternative Möglichkeit, eine Barrierefunktion zu konstruieren, bietet beispielsweise die Funktion

$$B(\mathbf{x}) = - \sum_{i=1}^q \log(-h_i(\mathbf{x})) . \quad (3.133)$$

Die Vorgehensweise ist nun ähnlich zur Methode der Straffunktionen. Es wird für eine gegen Unendlich strebende Folge  $\{c_l\}$ ,  $l = 1, 2, \dots$  mit  $c_1 > 0$  und  $c_{l+1} > c_l$  das unbeschränkte Optimierungsproblem

$$\min_{\mathbf{x} \in \check{\mathcal{X}}} f(\mathbf{x}) + \frac{1}{c_l} B(\mathbf{x}) \quad (3.134)$$

Abbildung 3.8: Verlauf der Barrierefunktionen  $\frac{1}{c}B(x)$ .

gelöst und mit  $\mathbf{x}_l^*$  der jeweilige optimale Punkt bezeichnet. Aufgrund des Definitionsbereiches der Barrierefunktion muss  $\mathbf{x}$  stets im (nicht leeren) Inneren  $\check{\mathcal{X}}$  des zulässigen Gebiets  $\mathcal{X}$  liegen. Auch hier bietet es sich zur numerischen Lösung an,  $\mathbf{x}_l^*$  als Startpunkt für die Optimierung mit  $c_{l+1} > c_l$  zu verwenden. Es gilt nun folgender Satz.

**Satz 3.14 (Konvergenz der Methode der Barrierefunktionen).** *Angenommen,  $\{\mathbf{x}_l^*\}$  sei eine Folge von Punkten, die durch die Lösung des (unbeschränkten) Optimierungsproblems (3.134) für eine gegen Unendlich strebende Folge  $\{c_l\}$ ,  $l = 1, 2, \dots$  mit  $c_1 > 0$  und  $c_{l+1} > c_l$  erhalten wurde. Dann ist jeder Grenzwert der Folge  $\{\mathbf{x}_l^*\}$  eine Lösung des beschränkten Optimierungsproblems (3.123).*

### 3.3 Beispiel: Rosenbrock's „Bananenfunktion“

Es wird das beschränkte Optimierungsproblem (vgl. (2.113))

$$\min_{\mathbf{x} \in \mathbb{R}^2} f(\mathbf{x}) = 100(x_2 - x_1^2)^2 + (x_1 - 1)^2 \quad (3.135a)$$

$$\text{u.B.v. } x_1^2 + x_2^2 \geq 0.5^2 \quad (3.135b)$$

betrachtet. Die Kostenfunktion („Bananenfunktion“) ist gemeinsam mit dem Rand des zulässigen Gebiets  $\mathcal{X}$  in Abbildung 3.9 dargestellt.

Zur Lösung von beschränkten Optimierungsproblemen bietet sich in MATLAB der Befehl `fmincon` an. In diesem Befehl sind die folgenden vier Algorithmen implementiert:

1. `interior-point`: Verwendet logarithmische Barrierefunktionen.
2. `active-set`: Verwendet die sequentielle quadratische Programmierung mit unterlagerter Lösung des quadratischen Programms nach der Methode der aktiven Beschränkungen.
3. `sqp`: Ähnlich `active-set`, unterscheidet sich aber in den unterlagerten Programmroutinen sowie den Eigenschaften der Iteration zum Minimum.

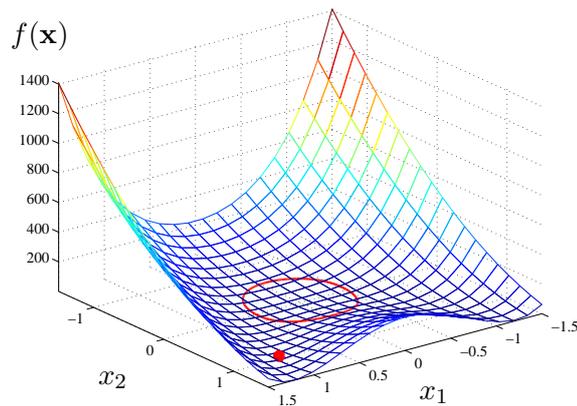


Abbildung 3.9: Profil der Rosenbrock Bananenfunktion und Rand des zulässigen Gebiets.

4. **trust region reflective**: Methode der Vertrauensbereiche, erweitert auf Optimierungsprobleme mit entweder Beschränkungen der Form  $\mathbf{Ax} = \mathbf{b}$  oder Beschränkungen der Form  $\mathbf{l} \leq \mathbf{x} \leq \mathbf{u}$ , wobei  $\mathbf{l}$  bzw.  $\mathbf{u}$  untere bzw. obere Schranken von  $\mathbf{x}$  bezeichnen.

Die Lösung des Optimierungsproblems (3.135) ist damit nur mit den Methoden **active-set**, **interior-point** und **sqp** möglich, weil **trust region reflective** keine nichtlinearen Beschränkungen zulässt. Die Ergebnisse der drei angesprochenen Algorithmen sind in Abbildung 3.10 dargestellt. Die Algorithmen **active-set** und **sqp** finden das globale Minimum, **interior-point** konvergiert zu einem anderen lokalen Minimum. Die gewählten Einstellungen der einzelnen Algorithmen sind in der MATLAB-Implementierung in Code-Auflistung 3.1 ersichtlich.

Listing 3.1: MATLAB-Code für die beschränkte Optimierung der Rosenbrock'schen Bananenfunktion.

```
function [Xopt,fopt,exitflag,output] = rosenbrock_problem_constrained(Xinit,testCase)
% -----
% Xinit: Startpunkt
% testCase: 1 - Active-Set
%           2 - SQP
%           3 - Interior Point

old = [Xinit; rosenbrock(Xinit)];
opt = optimoptions('fmincon','Display','iter','PlotFcns',@plot_iterates); % Optionen für die Ausgabe

switch testCase
case 1, % Active-Set mit SQP
    opt = optimoptions(opt,'Algorithm','active-set','GradObj','on','MaxFunEvals',1000,'TolFun',1e-12);
    [Xopt,fopt,exitflag,output] = fmincon(@rosenbrock,Xinit,[],[],[],[],[],[],@nonlconstr1,opt);
case 2, % SQP
    opt = optimoptions(opt,'Algorithm','sqp','GradObj','on','MaxFunEvals',2000,'TolX',1e-18);
    [Xopt,fopt,exitflag,output] = fmincon(@rosenbrock,Xinit,[],[],[],[],[],[],@nonlconstr1,opt);
case 3, % Interior-Point
    opt = optimoptions(opt,'Algorithm','interior-point','GradObj','on','TolFun',1e-12);
    [Xopt,fopt,exitflag,output] = fmincon(@rosenbrock,Xinit,[],[],[],[],[],[],@nonlconstr1,opt);
end

function [c,ceq] = nonlconstr1(x) % Nichtlineare Beschränkungsfunktion
```

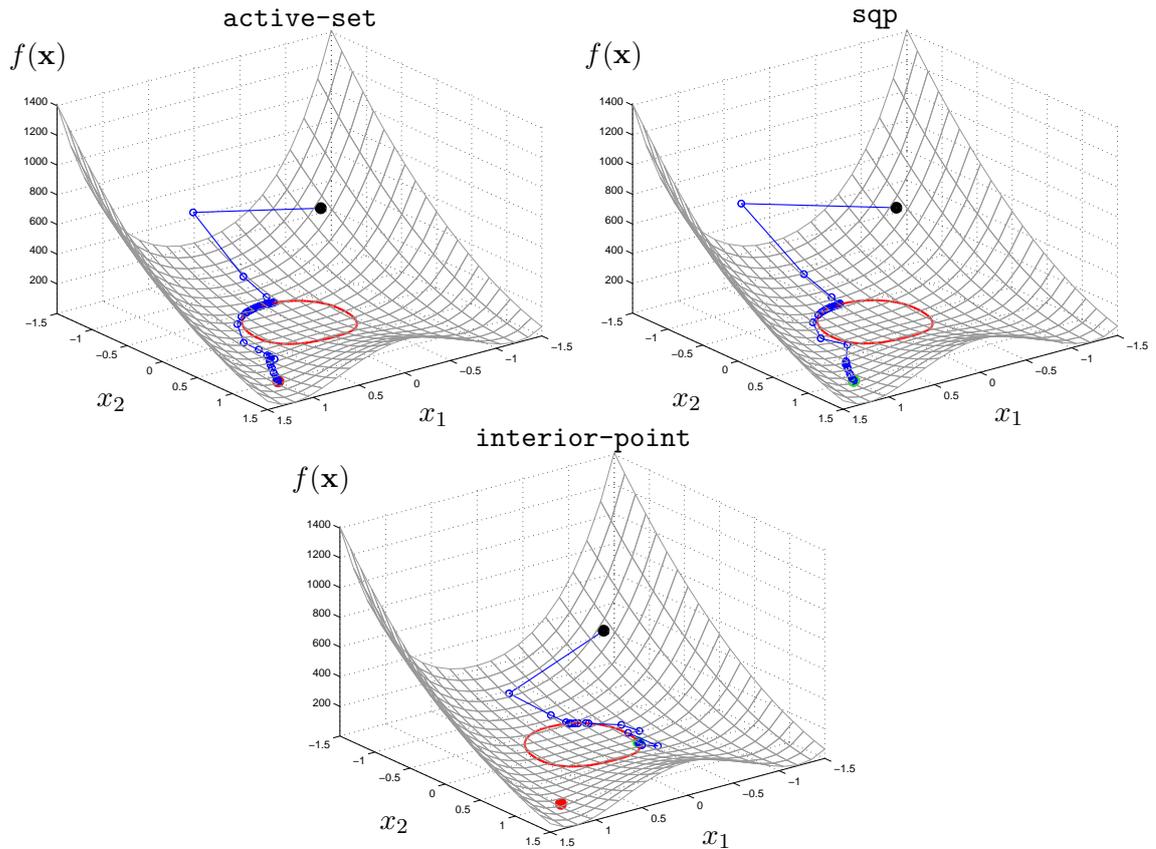


Abbildung 3.10: Rosenbrock Bananenfunktion: Vergleich der numerischen Verfahren aus fmincon.

```

c = 0.5^2 - (x(1))^2 - (x(2))^2; ceq = [];

function stop = plot_iterates(x,info,state)
global old
f = rosenbrock(x);
switch state
case 'init',
    plot_surface(x,f);
    plot_constraints;
case 'iter',
    plot3([old(1),x(1)],[old(2),x(2)],[old(3),f],'b-o','LineWidth',1);
case 'done',
    plot3(x(1),x(2),f,'go','LineWidth',5);
end
stop = false;
old = [x,f];

function plot_constraints % Zeichnen der eingestellten Beschränkung
[X1,X2] = meshgrid(-1.5:0.15:1.5);
x1_values = [-1.5:0.15:1.5]; x2_values = [-1.5:0.15:1.5]; r = 0.5;
x1_plot_values = [-r:0.01:r]; x2_plot_values1 = sqrt(r^2-(x1_plot_values).^2);
x2_plot_values2 = -sqrt(r^2-(x1_plot_values).^2);
z_values1 = 100*(x2_plot_values1-x1_plot_values.^2).^2 + (x1_plot_values-1).^2;
z_values2 = 100*(x2_plot_values2-x1_plot_values.^2).^2 + (x1_plot_values-1).^2;

```

```

line(x1_plot_values,x2_plot_values1,z_values1,'LineWidth',2,'color','r');
line(x1_plot_values,x2_plot_values2,z_values2,'LineWidth',2,'color','r');

function plot_surface(x,f) % Zeichnen der Rosenbrock-Funktion mit Startpunkt und optimalem Punkt
[X1,X2] = meshgrid(-1.5:0.15:1.5);          % 3D-Profil von
F = 100*(X2-X1.^2).^2 + (X1-1).^2;          % Rosenbrock-Funktion
h = surf(X1,X2,F,'EdgeColor',0.6*[1,1,1],'FaceColor','none');
hold on; axis tight;
plot3(x(1),x(2),f,'ko','LineWidth',5);      % Startpunkt
plot3(1,1,0,'ro','LineWidth',5);          % optimale Lösung
xlabel('x_1'); ylabel('x_2'); zlabel('f')
set(gcf,'ToolBar','figure');              % Aktivieren der Menüleiste (Zoom, etc.)
set(gca,'Xdir','reverse','Ydir','reverse');

function [f, grad, H] = rosenbrock(x)
grad = {}; H = {};
f = 100*(x(2)-x(1)^2)^2 + (x(1)-1)^2;      % Rosenbrock-Funktion
if nargin>1,                               % falls Gradient angefordert wird
    grad = [ -400*(x(2)-x(1)^2)*x(1)+2*(x(1)-1); 200*(x(2)-x(1)^2) ];
end
if nargin>2,                               % falls Hessematrix angefordert wird
    H = [ -400*(x(2)-3*x(1)^2)+2, -400*x(1); -400*x(1), 200 ];
end

```

## 3.4 Software-Übersicht

Im Folgenden ist eine Auswahl an Software zur Lösung von statischen Optimierungsproblemen zusammengestellt.

### Lineare Optimierung

- linprog: MATLAB Optimization Toolbox (kostenpflichtig)
- CPLEX (kostenpflichtig)  
<http://www.ilog.com/products/cplex>
- GLPK: „GNU Linear Programming Kit“ (frei zugänglich)  
<http://www.gnu.org/software/glpk>
- lp\_solve: Mixed-Integer Lineare Optimierung (frei zugänglich)  
<http://lpsolve.sourceforge.net>

### Quadratische Optimierung

- quadprog: MATLAB Optimization Toolbox (kostenpflichtig)
- CPLEX (kostenpflichtig)  
<http://www.ilog.com/products/cplex>
- OOQP (frei zugänglich)  
<http://pages.cs.wisc.edu/~swright/ooqp>
- qpOASES (frei zugänglich)  
<https://projects.coin-or.org/qpOASES>

- CVX (frei zugänglich)  
<http://cvxr.com/cvx/>
- LOQO (kostenpflichtig)  
<http://www.princeton.edu/~rvdb>

### Nichtlineare Optimierung

- fmincon: MATLAB Optimization Toolbox (kostenpflichtig)
- LOQO (kostenpflichtig)  
<http://www.princeton.edu/~rvdb>
- MINOS (kostenpflichtig)  
[http://www.sbsi-sol-optimize.com/asp/sol\\_product\\_minos.htm](http://www.sbsi-sol-optimize.com/asp/sol_product_minos.htm)
- SNOPT (kostenpflichtig, aber Studentenversion frei zugänglich)  
[http://www.sbsi-sol-optimize.com/asp/sol\\_product\\_snopt.htm](http://www.sbsi-sol-optimize.com/asp/sol_product_snopt.htm)
- DONLP2 (frei zugänglich)  
<ftp://ftp.mathematik.tu-darmstadt.de/pub/department/software/opti>
- IPOPT (frei zugänglich)  
<https://projects.coin-or.org/Ipopt>
- NLOPT (frei zugänglich)  
<http://ab-initio.mit.edu/wiki/index.php/NLopt>
- YALMIP (frei zugänglich)  
<https://yalmip.github.io/>

### Modellierungssprachen

Viele der oben angegebenen Optimierer unterstützen eine Anbindung an MATLAB (z. B. `lp_solve`, SNOPT, DONLP2, IPOPT) oder an eine der Modellierungssprachen AMPL (z. B. LOQO, GLPK, IPOPT) oder GAMS (z. B. MINOS). Diese Sprachen bieten eine symbolorientierte Syntax zum Formulieren von Optimierungsproblemen:

- AMPL: “A Mathematical Programming Language”  
<http://www.ampl.com>
- GAMS: “General Algebraic Modeling System”  
<http://www.gams.com>
- OPL: “Optimization Programming Language”  
<https://www-01.ibm.com/software/commerce/optimization/modeling/>

### 3.5 Literatur

- [3.1] I. Griva, S. Nash und A. Sofer, *Linear and Nonlinear Optimization*, 2. Aufl. Society for Industrial und Applied Mathematics, 2009.
- [3.2] D. P. Bertsekas, *Nonlinear Programming*, 2. Aufl. Athena Scientific, 1999.
- [3.3] M. Bazararaa, H. Serali und C. Shetty, *Nonlinear Programming: Theory and Algorithms*, 3. Aufl. John Wiley & Sons, 2006.
- [3.4] L. Biegler, *Nonlinear Programming: Concepts, Algorithms, and Applications to Chemical Processes*. Society for Industrial und Applied Mathematics, 2010.
- [3.5] D. G. Luenberger und Y. Ye, *Linear and Nonlinear Programming*, 3. Aufl., Ser. International Series in Operations Research & Management Science. Springer, 2008, Bd. 116.
- [3.6] S. S. Rao, *Engineering Optimization, Theory and Practice*, 4. Aufl. John Wiley & Sons, 2009.
- [3.7] J. Nocedal und S. J. Wright, *Numerical Optimization*, 2. Aufl., Ser. Springer Series in Operations Research and Financial Engineering. Springer, 2006.
- [3.8] S. Boyd und L. Vandenberghe, *Convex Optimization*. Cambridge University Press, 2004.
- [3.9] M. Papageorgiou, M. Leibold und M. Buss, *Optimierung: Statische, dynamische, stochastische Verfahren für die Anwendung*, 4. Aufl. Springer, 2015.
- [3.10] C. T. Kelley, *Iterative Methods for Optimization*. Society for Industrial und Applied Mathematics, 1999.
- [3.11] B. C. Chachuat, „Nonlinear and Dynamic Optimization: From Theory to Practice“, abrufbar unter <http://infoscience.epfl.ch/record/111939>, Laboratoire d'Automatique, École Polytechnique Fédérale de Lausanne, 2007, (besucht am 28.09.2017).
- [3.12] P. E. Gill, W. Murray und M. H. Wright, *Practical Optimization*. Academic Press, 1981.
- [3.13] S.-P. Han, „A Globally Convergent Method for Nonlinear Programming“, *Journal of Optimization Theory and Applications*, Jg. 22, Nr. 3, S. 297–309, 1977.
- [3.14] M. J. D. Powell, „A Fast Algorithm for Nonlinearly Constrained Optimization Calculations“, in *Numerical Analysis*, Ser. Lecture Notes in Mathematics, G. A. Watson, Hrsg., Bd. 630, Springer, 1978, S. 144–157.
- [3.15] K. Schittkowski, „On the Convergence of a Sequential Quadratic Programming Method with an Augmented Lagrangian Line Search Function“, *Mathematische Operationsforschung und Statistik. Series Optimization*, Jg. 14, Nr. 2, S. 197–216, 1983.

# 4 Dynamische Optimierung

## 4.1 Grundlagen der Variationsrechnung

### 4.1.1 Problemformulierung

Im Gegensatz zu den bisher betrachteten statischen Optimierungsproblemen, bei denen die Optimierungsvariablen  $\mathbf{x}$  in einem *finit-dimensionalen Euklidischen Vektorraum*  $\mathbb{R}^n$  definiert sind, wird bei dynamischen Optimierungsaufgaben nach dem Minimum (Maximum) eines *Kostenfunctionals*  $J : \mathcal{V} \rightarrow \mathbb{R}$  bezüglich einer (reellen vektorwertigen) Funktion  $\mathbf{x}(t)$  aus einem geeigneten *Funktionsraum*  $\mathcal{V}$  gesucht. In vielen Fällen entspricht die *unabhängige Variable*  $t$  der Zeit. Die totale Ableitung nach  $t$  wird mit  $(\dot{\cdot}) = d(\cdot)/dt$  abgekürzt. Typischerweise hat das Kostenfunktional die Form (*Lagrange Problem der Variationsrechnung*)

$$J(\mathbf{x}) = \int_{t_0}^{t_1} l(t, \mathbf{x}(t), \dot{\mathbf{x}}(t)) dt \quad (4.1)$$

oder (*Bolza Problem der Variationsrechnung*)

$$J(\mathbf{x}) = \varphi(t_0, \mathbf{x}(t_0), t_1, \mathbf{x}(t_1)) + \int_{t_0}^{t_1} l(t, \mathbf{x}(t), \dot{\mathbf{x}}(t)) dt . \quad (4.2)$$

Dabei wird  $\mathbf{x}(t) = [x_1(t) \ \dots \ x_n(t)]^T : [t_0, t_1] \rightarrow \mathbb{R}^n$  häufig als *Trajektorie* bezeichnet. Die reellwertige Funktion  $l(t, \mathbf{x}(t), \dot{\mathbf{x}}(t)) : \mathbb{R} \times \mathbb{R}^n \times \mathbb{R}^n \rightarrow \mathbb{R}$  nennt man *Lagrangesche Dichte* und  $\varphi(t_0, \mathbf{x}(t_0), t_1, \mathbf{x}(t_1)) : \mathbb{R} \times \mathbb{R}^n \times \mathbb{R} \times \mathbb{R}^n \rightarrow \mathbb{R}$  beschreibt die *Rand- oder Endkostenfunktion* (englisch: *boundary costs* oder *terminal costs*). Die Lagrangesche Dichte  $l$  sollte nicht mit der in Abschnitt 3.1.2.1 eingeführten Lagrangefunktion  $L$  verwechselt werden.

Man nennt eine Trajektorie  $\mathbf{x}(t)$  *zulässig*, wenn im Intervall  $[t_0, t_1]$  sämtliche Beschränkungen eingehalten werden. Die Menge aller zulässigen Trajektorien wird im Weiteren mit  $\mathcal{X}$  bezeichnet. Die einfachste Form solcher Beschränkungen ist, dass beide Endpunkte fixiert sind, d. h.  $\mathbf{x}(t_0) = \mathbf{x}_0$  und  $\mathbf{x}(t_1) = \mathbf{x}_1$  bzw.  $\mathcal{X} = \{\mathbf{x}(t) \in \mathcal{V} \mid \mathbf{x}(t_0) = \mathbf{x}_0, \mathbf{x}(t_1) = \mathbf{x}_1\}$ . Eine weitere Möglichkeit besteht darin, dass die Trajektorie zwar an einem festen Punkt  $(t_0, \mathbf{x}_0)$  startet aber zu einem *freien* Zeitpunkt  $t_1 \in [t_0, T]$  auf einer vorgegebenen Kurve  $\Gamma$  definiert durch  $\mathbf{x} = \mathbf{g}(t)$  mit  $t_0 \leq t \leq T$  zu liegen kommen muss. In diesem Fall ist die freie Endzeit  $t_1$  eine zu optimierende Größe und die zulässige Menge lautet  $\mathcal{X} = \{\mathbf{x}(t) \in \mathcal{V} \mid \mathbf{x}(t_0) = \mathbf{x}_0, \mathbf{x}(t_1) = \mathbf{g}(t_1)\}$ . Andere mögliche Beschränkungen sind so genannte *Pfadbeschränkungen* (englisch: *path constraints*) oder *Beschränkungen der Lagrangeschen Form*

$$\psi(t, \mathbf{x}(t), \dot{\mathbf{x}}(t)) = 0 \quad \text{bzw.} \quad \psi(t, \mathbf{x}(t), \dot{\mathbf{x}}(t)) \leq 0, \quad t \in I \subseteq [t_0, t_1] \quad (4.3)$$

und *isoperimetrische Beschränkungen* der Form

$$\int_{t_0}^{t_1} \psi(t, \mathbf{x}(t), \dot{\mathbf{x}}(t)) dt = 0. \quad (4.4)$$

Die bei der Variationsrechnung typischerweise betrachteten Funktionenräume sind die im Intervall  $[t_0, t_1]$  *stetig differenzierbaren Funktionen*  $(C^1[t_0, t_1])^n$  und die *stückweise stetig differenzierbaren Funktionen*, im Weiteren als  $(\hat{C}^1[t_0, t_1])^n$  bezeichnet. Elemente des Funktionenraumes  $(\hat{C}^1[t_0, t_1])^n$  werden dabei im Folgenden auch als global stetig angenommen. Die Definition des *globalen Minimums*  $\mathbf{x}^*$  eines Kostenfunktional  $J(\mathbf{x})$  lässt sich ohne Angabe einer Norm direkt in der Form

$$J(\mathbf{x}^*) \leq J(\mathbf{x}), \quad \forall \mathbf{x} \in \mathcal{X} \quad (4.5)$$

angeben. Die Beschreibung des *lokalen* Verhaltens in der Umgebung des Minimums  $\mathbf{x}^*$  hingegen verlangt die Definition einer Norm. Eine zulässige Lösung  $\mathbf{x}^*$  ist ein *lokales Minimum* in  $\mathcal{X}$  bezüglich der Norm  $\|\cdot\|$ , wenn gilt

$$\exists \gamma > 0 \text{ so, dass gilt } J(\mathbf{x}^*) \leq J(\mathbf{x}), \quad \forall \mathbf{x} \in \mathcal{X} \cap B_\gamma(\mathbf{x}^*) \quad (4.6)$$

mit  $B_\gamma(\mathbf{x}^*) = \{\mathbf{x} \in \mathcal{V} \mid \|\mathbf{x} - \mathbf{x}^*\| < \gamma\}$ . Da in infinit-dimensionalen Vektorräumen Normen grundsätzlich nicht äquivalent sind, kann  $\mathbf{x}^*$  zwar bezüglich einer Norm ein lokales Minimum sein, aber bezüglich einer anderen Norm nicht. Im Funktionenraum  $(C^1[t_0, t_1])^n$  werden häufig die Normen

$$\|\mathbf{x}(t)\|_\infty := \max_{t_0 \leq t \leq t_1} \|\mathbf{x}(t)\| \quad \text{und} \quad \|\mathbf{x}(t)\|_{1,\infty} := \max_{t_0 \leq t \leq t_1} \|\mathbf{x}(t)\| + \max_{t_0 \leq t \leq t_1} \|\dot{\mathbf{x}}(t)\| \quad (4.7)$$

verwendet, wobei  $\|\mathbf{x}(t)\|$  eine Norm im finit-dimensionalen Vektorraum  $\mathbb{R}^n$  beschreibt.

### 4.1.2 Optimalitätsbedingungen

Zur Herleitung notwendiger Optimalitätsbedingungen benötigt man den Begriff der *Variation eines Funktional*.

**Definition 4.1** (*Variation eines Funktional, Gâteaux Ableitung*). Die *erste Variation des Funktional*  $J(\mathbf{x})$  am Punkt  $\mathbf{x} \in \mathcal{V}$  in Richtung  $\boldsymbol{\xi} \in \mathcal{V}$ , auch als *Gâteaux Ableitung* von  $J(\mathbf{x})$  bezüglich  $\boldsymbol{\xi}$  am Punkt  $\mathbf{x}$  bezeichnet, ist in der Form

$$\delta J(\mathbf{x}; \boldsymbol{\xi}) := \lim_{\eta \rightarrow 0} \frac{J(\mathbf{x} + \eta \boldsymbol{\xi}) - J(\mathbf{x})}{\eta} = \left. \frac{d}{d\eta} J(\mathbf{x} + \eta \boldsymbol{\xi}) \right|_{\eta=0} \quad (4.8)$$

definiert. Falls  $\delta J(\mathbf{x}; \boldsymbol{\xi})$  für alle  $\boldsymbol{\xi} \in \mathcal{V}$  definiert ist, dann nennt man  $J(\mathbf{x})$  *Gâteaux differenzierbar* am Punkt  $\mathbf{x}$ .

Für die Existenz der Gâteaux Ableitung muss nicht nur das Funktional  $J(\mathbf{x})$  definiert sein, sondern auch die Ableitung von  $J(\mathbf{x} + \eta \boldsymbol{\xi})$  bezüglich  $\eta$  an der Stelle  $\eta = 0$  existieren.

**Beispiel 4.1.** Die Gâteaux Ableitung des Funktionals  $J(x) = \int_{t_0}^{t_1} x^2(t) dt$ ,  $x \in C^1[t_0, t_1]$  lautet

$$\begin{aligned} \delta J(x; \xi) &= \lim_{\eta \rightarrow 0} \frac{\int_{t_0}^{t_1} (x(t) + \eta \xi(t))^2 dt - \int_{t_0}^{t_1} x^2(t) dt}{\eta} \\ &= \lim_{\eta \rightarrow 0} \left( \int_{t_0}^{t_1} 2x(t)\xi(t) dt + \eta \int_{t_0}^{t_1} \xi^2(t) dt \right) = 2 \int_{t_0}^{t_1} x(t)\xi(t) dt \end{aligned} \quad (4.9)$$

für alle  $\xi \in C^1[t_0, t_1]$ , weshalb  $J(x)$  an jedem Punkt  $x \in C^1[t_0, t_1]$  Gâteaux differenzierbar ist.

**Beispiel 4.2.** Man betrachte das Funktional  $J(x) = \int_0^1 |x(t)| dt$ ,  $x \in C^1[0, 1]$ , welches für jedes  $x \in C^1[0, 1]$  im endlichen Intervall  $[0, 1]$  einen finiten Wert liefert. Für  $x_0(t) = 0$  und  $\xi_0(t) = t$  lautet die Gâteaux Ableitung (4.8)

$$\delta J(x_0; \xi_0) = \lim_{\eta \rightarrow 0} \frac{1}{\eta} \left( \int_0^1 |x_0 + \eta \xi_0| dt - \int_0^1 |x_0| dt \right) = \quad (4.10a)$$

$$= \lim_{\eta \rightarrow 0} \operatorname{sgn}(\eta) \int_0^1 |t| dt = \begin{cases} \frac{1}{2}, & \eta \rightarrow +0 \\ -\frac{1}{2}, & \eta \rightarrow -0 \end{cases} \quad (4.10b)$$

Dabei erkennt man, dass in Richtung  $\xi_0 = t$  an der Stelle  $x_0 = 0$  die Gâteaux Ableitung nicht existiert.

Die Gâteaux Ableitung ist eine *lineare Operation*, weshalb gilt

$$\delta(J_1 + J_2)(\mathbf{x}; \boldsymbol{\xi}) = \delta J_1(\mathbf{x}; \boldsymbol{\xi}) + \delta J_2(\mathbf{x}; \boldsymbol{\xi}) \quad (4.11)$$

und für jedes reelle  $\alpha$  gilt die Beziehung

$$\delta J(\mathbf{x}; \alpha \boldsymbol{\xi}) = \alpha \delta J(\mathbf{x}; \boldsymbol{\xi}) \quad (4.12)$$

Basierend auf der Gâteaux Ableitung lässt sich nun der Begriff der *zulässigen Richtung* eines Funktionals definieren.

**Definition 4.2 (Zulässige Richtung).**  $J : \mathcal{X} \rightarrow \mathbb{R}$  sei ein Funktional welches in einer Teilmenge  $\mathcal{X}$  eines normierten linearen Vektorraums  $(\mathcal{V}, \|\cdot\|)$  definiert ist. An einem (zulässigen) Punkt  $\mathbf{x}$  im Inneren von  $\mathcal{X}$  bezeichnet man  $\boldsymbol{\xi} \in \mathcal{V}$  mit  $\boldsymbol{\xi} \neq \mathbf{0}$  als *zulässige Richtung*, wenn

- (a)  $\delta J(\mathbf{x}; \boldsymbol{\xi})$  existiert und
- (b) ein (hinreichend kleines)  $\varepsilon > 0$  existiert, so dass  $\mathbf{x} + \eta \boldsymbol{\xi} \in \mathcal{X}$  für alle  $\eta \in (-\varepsilon, \varepsilon)$  gilt.

Die Bedingung (b) verlangt natürlich, dass  $\mathbf{x}$  im Inneren von  $\mathcal{X}$  liegt. Abbildung 4.1 zeigt ein Beispiel für eine zulässige Richtung  $\xi(t)$  im Fall  $\mathcal{X} = \{x(t) \in C[t_0, t_1] \mid x(t_0) = x_0, x(t_1) = x_1\}$ . Eine zulässige Richtung  $\boldsymbol{\xi}$  am Punkt  $\bar{\mathbf{x}}$  für die gilt  $\delta J(\bar{\mathbf{x}}; \boldsymbol{\xi}) < 0$  wird

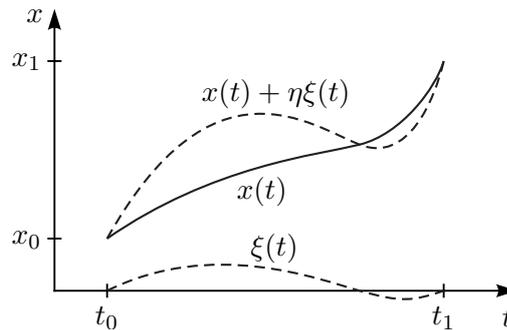


Abbildung 4.1: Zulässige Richtung im Fall  $\mathcal{X} = \{x(t) \in C[t_0, t_1] \mid x(t_0) = x_0, x(t_1) = x_1\}$ .

*Abstiegsrichtung* des Funktionals  $J$  am Punkt  $\bar{\mathbf{x}}$  bezeichnet. Dies stellt eine Generalisierung der Abstiegsrichtung  $\mathbf{d}$  der Kostenfunktion  $f(\mathbf{x})$  im finit-dimensionalen Fall mit  $\mathbf{d}^T(\nabla f)(\bar{\mathbf{x}}) < 0$  am Punkt  $\bar{\mathbf{x}}$  gemäß Satz 2.1 dar. Es gilt nun folgendes Lemma.

**Lemma 4.1 (Ausschluss eines Minimums).** *Wenn  $J$  ein Funktional in einem normierten linearen Vektorraum  $(\mathcal{V}, \|\cdot\|)$  beschreibt und an einem Punkt  $\bar{\mathbf{x}} \in \mathcal{X}$  eine zulässige Richtung  $\boldsymbol{\xi} \in \mathcal{V}$  so existiert, dass gilt  $\delta J(\bar{\mathbf{x}}; \boldsymbol{\xi}) < 0$ , dann kann  $\bar{\mathbf{x}}$  kein lokales Minimum sein.*

*Beweisskizze:* Gemäß Definition 4.1 gilt

$$\delta J(\bar{\mathbf{x}}; \boldsymbol{\xi}) = \lim_{\eta \rightarrow 0} \frac{J(\bar{\mathbf{x}} + \eta \boldsymbol{\xi}) - J(\bar{\mathbf{x}})}{\eta} < 0 \quad (4.13)$$

und es existiert ein  $\gamma > 0$  so, dass

$$J(\bar{\mathbf{x}} + \eta \boldsymbol{\xi}) < J(\bar{\mathbf{x}}), \quad \forall \eta \in (0, \gamma). \quad (4.14)$$

Da nun  $\boldsymbol{\xi}$  eine zulässige Richtung gemäß Definition 4.2 ist, kann das Funktional  $J$  am Punkt  $\bar{\mathbf{x}}$  in Richtung  $\eta \boldsymbol{\xi}$  mit beliebigem  $\eta \in (0, \gamma)$  weiter verkleinert werden. Da unabhängig von der verwendeten Norm  $\|\bar{\mathbf{x}} + \eta \boldsymbol{\xi} - \bar{\mathbf{x}}\| = \|\eta \boldsymbol{\xi}\| \rightarrow 0$  für  $\eta \rightarrow 0$  gilt, findet man stets einen hinreichend kleinen Wert  $\eta \in (0, \gamma)$ , so dass  $\bar{\mathbf{x}} + \eta \boldsymbol{\xi}$  im Sinne der Norm  $\|\cdot\|$  in der Umgebung von  $\bar{\mathbf{x}}$  liegt. Folglich kann  $\bar{\mathbf{x}}$  kein lokales Minimum sein.  $\square$

Die notwendigen Bedingungen erster Ordnung für ein lokales Minimum eines Funktionals lassen sich nun wie folgt formulieren [4.1].

**Satz 4.1 (Notwendige Bedingungen erster Ordnung).** *Angenommen  $\mathbf{x}^* \in \mathcal{X}$  ist ein (lokales) Minimum des Funktionals  $J$ , welches in einer Teilmenge  $\mathcal{X}$  eines normierten linearen Vektorraums  $(\mathcal{V}, \|\cdot\|)$  definiert ist. Dann gilt*

$$\delta J(\mathbf{x}^*; \boldsymbol{\xi}) = 0 \quad (4.15)$$

für alle zulässigen Richtungen  $\xi$  gemäß Definition 4.2 an der Stelle  $\mathbf{x}^*$ .

Im nächsten Schritt betrachte man das Lagrange Problem der Variationsrechnung gemäß (4.1) mit festem Anfangs- und Endpunkt.

**Satz 4.2 (Euler-Lagrange Gleichungen).** Gegeben sei das Funktional

$$J(\mathbf{x}) = \int_{t_0}^{t_1} l(t, \mathbf{x}(t), \dot{\mathbf{x}}(t)) dt \quad (4.16)$$

mit der zulässigen Menge  $\mathcal{X} = \{ \mathbf{x}(t) \in (C^1[t_0, t_1])^n \mid \mathbf{x}(t_0) = \mathbf{x}_0, \mathbf{x}(t_1) = \mathbf{x}_1 \}$  und der stetig differenzierbaren Lagrangeschen Dichte  $l : \mathbb{R} \times \mathbb{R}^n \times \mathbb{R}^n \rightarrow \mathbb{R}$ . Wenn  $\mathbf{x}^*(t)$  ein (lokales) Minimum von  $J(\mathbf{x})$  auf  $\mathcal{X}$  bezeichnet, dann erfüllt  $\mathbf{x}^*(t)$  die Euler-Lagrange Gleichungen

$$\frac{d}{dt} \left( \frac{\partial}{\partial \dot{x}_i} l \right) (t, \mathbf{x}^*(t), \dot{\mathbf{x}}^*(t)) - \left( \frac{\partial}{\partial x_i} l \right) (t, \mathbf{x}^*(t), \dot{\mathbf{x}}^*(t)) = 0 \quad (4.17)$$

für alle  $t \in [t_0, t_1]$  und  $i = 1, \dots, n$ .

*Beweis.* Da  $\mathbf{x}^*$  ein Minimum ist, muss wegen Satz 4.1 gelten

$$\begin{aligned} \delta J(\mathbf{x}^*; \xi) &= \frac{d}{d\eta} J(\mathbf{x}^* + \eta \xi) \Big|_{\eta=0} = \int_{t_0}^{t_1} \frac{d}{d\eta} l(t, \mathbf{x}^*(t) + \eta \xi(t), \dot{\mathbf{x}}^*(t) + \eta \dot{\xi}(t)) dt \Big|_{\eta=0} \\ &= \int_{t_0}^{t_1} \left[ \left( \frac{\partial}{\partial \mathbf{x}} l \right) (t, \mathbf{x}^*(t), \dot{\mathbf{x}}^*(t)) \xi + \left( \frac{\partial}{\partial \dot{\mathbf{x}}} l \right) (t, \mathbf{x}^*(t), \dot{\mathbf{x}}^*(t)) \dot{\xi} \right] dt = 0. \end{aligned} \quad (4.18)$$

Wegen der stetigen Differenzierbarkeit der Lagrangeschen Dichte  $l$  und da  $\xi \in (C^1[t_0, t_1])^n$  ist der Integrand von (4.18) im Intervall  $[t_0, t_1]$  stetig und daher ist das Funktional  $J(\mathbf{x})$  an allen Punkten  $\mathbf{x} \in (C^1[t_0, t_1])^n$  Gâteaux differenzierbar. Eine nach Definition 4.2 zulässige Richtung  $\xi$  muss die Bedingungen  $\xi(t_0) = \mathbf{0}$  und  $\xi(t_1) = \mathbf{0}$  erfüllen. Führt man für den zweiten Summanden in der zweiten Zeile von (4.18) eine partielle Integration durch, so erhält man

$$\int_{t_0}^{t_1} \left( \frac{\partial}{\partial \mathbf{x}} l \right) \xi + \left( \frac{\partial}{\partial \dot{\mathbf{x}}} l \right) \dot{\xi} dt = \int_{t_0}^{t_1} \left[ \left( \frac{\partial}{\partial \mathbf{x}} l \right) - \frac{d}{dt} \left( \frac{\partial}{\partial \dot{\mathbf{x}}} l \right) \right] \xi dt + \underbrace{\left[ \left( \frac{\partial}{\partial \dot{\mathbf{x}}} l \right) \xi \right]_{t_0}^{t_1}}_{=0} = 0. \quad (4.19)$$

Wählt man nun für ein festes  $i = 1, \dots, n$  die Richtung  $\xi = [\xi_1 \ \dots \ \xi_n]^T \in (C^1[t_0, t_1])^n$  so, dass gilt  $\xi_j = 0$  für  $\forall j$  mit  $j \neq i$  und  $\xi_i(t_0) = \xi_i(t_1) = 0$ , dann ergibt sich

$$\int_{t_0}^{t_1} \left[ \left( \frac{\partial}{\partial x_i} l \right) - \frac{d}{dt} \left( \frac{\partial}{\partial \dot{x}_i} l \right) \right] \xi_i dt = 0. \quad (4.20)$$

Gemäß dem nachfolgend angeführten *Fundamentallemma der Variationsrechnung* folgt aus (4.20) unmittelbar das Ergebnis (4.17).  $\square$

**Lemma 4.2 (Fundamentallemma der Variationsrechnung).** *Angenommen  $g(t)$  ist eine stückweise stetige Funktion auf dem Intervall  $[t_0, t_1]$  und es gilt*

$$\int_{t_0}^{t_1} g(t) \xi_i(t) dt = 0 \quad (4.21)$$

*für alle stückweise stetigen Funktionen  $\xi_i(t)$  im Intervall  $[t_0, t_1]$ , dann folgt fast überall (abgesehen von einer abzählbaren Menge von Punkten)  $g(t) = 0$ ,  $t \in [t_0, t_1]$ .*

Eine Funktion  $\bar{\mathbf{x}}(t)$ , die die Euler-Lagrange Gleichungen (4.17) erfüllt, wird auch als *stationäre Funktion der Lagrangeschen Dichte  $l$*  bezeichnet. In manchen Literaturstellen werden diese Funktionen auch als *extremale Funktionen* oder nur *Extremale* bezeichnet, obwohl es sein kann, dass sie weder ein Minimum noch ein Maximum des Kostenfunktional beschreiben.

Mit Satz 4.2 ist es also gelungen, die Minimierung des Funktional (4.1) in ein *Zweipunkt-randwertproblem* mit den Euler-Lagrange Gleichungen umzuformulieren. Das erhaltene Randwertproblem kann meist mit gängigen numerischen Methoden [4.2–4.4], wie z. B. Finite-Differenzenverfahren, Einfach-Schießverfahren, Mehrfach-Schießverfahren und Kollokationsverfahren, gelöst werden. Die Lösung der Euler-Lagrange Gleichungen (4.17) kann für Spezialfälle auch mit Hilfe so genannter *erster Integrale* formuliert werden:

- (a) Die Lagrangesche Dichte hängt nicht von der unabhängigen Variablen  $t$  ab, d. h.  $l = l(\mathbf{x}, \dot{\mathbf{x}})$ . Mit der *Hamiltonfunktion*

$$H = \left( \frac{\partial}{\partial \dot{\mathbf{x}}} l \right) \dot{\mathbf{x}} - l(\mathbf{x}, \dot{\mathbf{x}}) \quad (4.22)$$

folgt aus den Euler-Lagrange Gleichungen (4.17), dass

$$\begin{aligned} \frac{d}{dt} H &= \frac{d}{dt} \left( \frac{\partial}{\partial \dot{\mathbf{x}}} l \right) \dot{\mathbf{x}} + \left( \frac{\partial}{\partial \dot{\mathbf{x}}} l \right) \ddot{\mathbf{x}} - \left( \frac{\partial}{\partial \mathbf{x}} l \right) \dot{\mathbf{x}} - \left( \frac{\partial}{\partial \dot{\mathbf{x}}} l \right) \ddot{\mathbf{x}} \\ &= \left( \frac{d}{dt} \left( \frac{\partial}{\partial \dot{\mathbf{x}}} l \right) - \left( \frac{\partial}{\partial \mathbf{x}} l \right) \right) \dot{\mathbf{x}} = 0. \end{aligned} \quad (4.23)$$

D. h. die Hamiltonfunktion  $H$  ist entlang von stationären Funktionen konstant und bildet damit eine *Invariante* des Systems.

- (b) Die Lagrangesche Dichte hängt nicht von  $\mathbf{x}$  ab, d. h.  $l = l(t, \dot{\mathbf{x}})$ . Dann folgt aus den Euler-Lagrange Gleichungen (4.17), dass  $\frac{\partial}{\partial \dot{x}_i} l$ ,  $i = 1, \dots, n$  *Invarianten* des Systems sind, denn es gilt

$$\frac{d}{dt} \left( \frac{\partial}{\partial \dot{x}_i} l \right) = 0. \quad (4.24)$$

**Aufgabe 4.1.** Nehmen Sie an, dass die Lagrangesche Dichte  $l(\mathbf{x}, \dot{\mathbf{x}})$  die Lagrangefunktion eines Starrkörpersystems ist (siehe Skriptum Fachvertiefung: Automatisierungs- und Regelungstechnik oder Regelungssysteme 2) und  $\mathbf{x}$  bzw.  $\dot{\mathbf{x}}$  die generalisierten Lagekoordinaten und deren Geschwindigkeiten bezeichnen. Geben Sie eine physikalische Interpretation der Hamiltonfunktion  $H$  von (4.22) und der darin auftretenden Größen  $\frac{\partial}{\partial \dot{x}_i} l$ ,  $i = 1, \dots, n$  an.

**Bemerkung 4.1.** Konservative Starrkörpersysteme erfüllen die Euler-Lagrange Gleichungen (4.17). Dies gilt im Allgemeinen nicht für nicht-konservative Starrkörpersysteme. Für sie lauten die Euler-Lagrange Gleichungen

$$\frac{d}{dt} \left( \frac{\partial}{\partial \dot{x}_i} l \right) - \left( \frac{\partial}{\partial x_i} l \right) = \tau_i \quad (4.25)$$

für alle  $t \in [t_0, t_1]$  und  $i = 1, \dots, n$  mit den externen generalisierten Kräften  $\tau_j$  (siehe Skriptum Fachvertiefung: Automatisierungs- und Regelungstechnik oder Regelungssysteme 2).

**Bemerkung 4.2.** Der Begriff *Lagrangefunktion* hat in der Mechanik eine andere Bedeutung als in der Optimierung.

**Beispiel 4.3 (Elastischer Zugstab belastet durch Eigengewicht).** Ein gerader, linear elastischer Stab habe die Zugsteifigkeit  $k$  und im unbelasteten Zustand die Masse pro Längeneinheit  $\bar{m}$  und die Länge  $x_1$ . Der Stab wird am Punkt  $x = x_0 = 0$  senkrecht befestigt und durch sein Eigengewicht (Erdbeschleunigung  $g$ ) belastet. Es soll das Verschiebungsfeld  $y(x)$  zufolge der Eigengewichtsbelastung berechnet werden. Die Längskoordinate  $x$  sei materialfest, d. h. sie wird im unbelasteten Zustand gemessen.

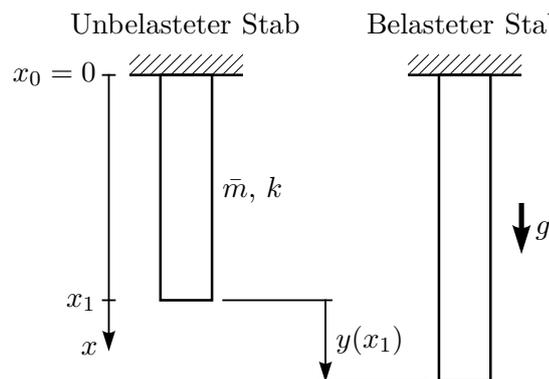


Abbildung 4.2: Elastischer Zugstab belastet durch Eigengewicht.

Zur Lösung dieser Aufgabe kann das *Hamiltonsche Prinzip* der Mechanik [4.5, 4.6] verwendet werden. Angewandt auf den Sonderfall der hier rein statischen Beanspruchung besagt es, dass die potentielle Energie des Stabes im statischen Gleichgewicht

extremal sein muss [4.7]. Für ein stabiles statisches Gleichgewicht muss sie minimal sein.

Die bis auf einen konstanten Term definierte potentielle Energie

$$J(y) = \int_0^{x_1} \frac{k(y'(x))^2}{2} - \bar{m}gy(x) \, dx \quad (4.26)$$

des Stabes setzt sich aus der Dehnungsenergie mit der Längsdehnung  $y'(x)$  und der potentiellen Höhenenergie zusammen. Am Befestigungspunkt  $x = x_0 = 0$  des Stabes darf keine Verschiebung auftreten und es gilt

$$y(0) = 0 . \quad (4.27a)$$

Da am freien Ende  $x = x_1$  des Stabes die Zugkraft 0 beträgt, muss dort auch die Dehnung verschwinden, d. h.

$$y'(x_1) = 0 . \quad (4.27b)$$

Die Minimierung des Funktional (4.26) unter Berücksichtigung der Randbedingungen (4.27) kann mit Hilfe der Variationsrechnung erfolgen. Da in der Lagrangeschen Dichte  $l(y, y') = k(y')^2/2 - \bar{m}gy$  die unabhängige Variable  $x$  nicht explizit auftritt, muss gemäß (4.23) die Hamiltonfunktion eine Invariante des Systems sein, d. h.

$$H = \left( \frac{\partial}{\partial y'} l \right) (y, y') y' - l(y, y') = \frac{k(y')^2}{2} + \bar{m}gy = c_1 = \text{konst.} \quad (4.28)$$

Die Integration dieser Differentialgleichung liefert

$$\left[ -\sqrt{2k} \frac{\sqrt{c_1 - \bar{m}gy}}{\bar{m}g} \right]_{y(0)}^{y(x)} = x . \quad (4.29)$$

Die Werte  $c_1$  und  $y(0)$  folgen schließlich aus den Randbedingungen (4.27) und für die Lösung ergibt sich

$$y(x) = \frac{\bar{m}g}{k} \left( x_1 x - \frac{x^2}{2} \right) . \quad (4.30)$$

Alternativ kann diese Aufgabe natürlich auch direkt mit Satz 4.2 gelöst werden. Aus der Euler-Lagrange Gleichung (4.17) folgt

$$\frac{d}{dx} \left( \frac{\partial}{\partial y'} l \right) (y, y') - \frac{\partial}{\partial y} l(y, y') = ky'' + \bar{m}g = 0 . \quad (4.31)$$

Die Integration dieser Differentialgleichung liefert bei Berücksichtigung der Randbedingungen (4.27) ebenfalls die Lösung (4.30).

Analog zum finit-dimensionalen Fall, siehe Satz 2.2, können auch für die Minimierung von Funktionalen notwendige Bedingungen zweiter Ordnung formuliert werden.

**Satz 4.3** (Notwendige Bedingungen zweiter Ordnung - Legendre Bedingung). *Angenommen  $\mathbf{x}^* \in \mathcal{X}$  ist ein lokales Minimum des Funktionals*

$$J(\mathbf{x}) = \int_{t_0}^{t_1} l(t, \mathbf{x}(t), \dot{\mathbf{x}}(t)) dt \quad (4.32)$$

mit der zulässigen Menge  $\mathcal{X} = \{\mathbf{x}(t) \in (C^1[t_0, t_1])^n \mid \mathbf{x}(t_0) = \mathbf{x}_0, \mathbf{x}(t_1) = \mathbf{x}_1\}$  und der zweifach stetig differenzierbaren Lagrangeschen Dichte  $l : \mathbb{R} \times \mathbb{R}^n \times \mathbb{R}^n \rightarrow \mathbb{R}$ , dann erfüllt  $\mathbf{x}^*$  die Euler-Lagrange Gleichungen (4.17) und die so genannte Legendre Bedingung

$$\mathbf{d}^T \left( \frac{\partial^2 l}{\partial \dot{\mathbf{x}}^2} \right) (t, \mathbf{x}^*(t), \dot{\mathbf{x}}^*(t)) \mathbf{d} \geq 0 \quad \forall \mathbf{d} \in \mathbb{R}^n, t \in [t_0, t_1]. \quad (4.33)$$

**Satz 4.4** (Hinreichende Bedingungen zweiter Ordnung - Konvexitätsbedingung). *Gegeben sei das Funktional*

$$J(\mathbf{x}) = \int_{t_0}^{t_1} l(t, \mathbf{x}(t), \dot{\mathbf{x}}(t)) dt \quad (4.34)$$

mit der zulässigen Menge  $\mathcal{X} = \{\mathbf{x}(t) \in (C^1[t_0, t_1])^n \mid \mathbf{x}(t_0) = \mathbf{x}_0, \mathbf{x}(t_1) = \mathbf{x}_1\}$  und der zweifach stetig differenzierbaren Lagrangeschen Dichte  $l : \mathbb{R} \times \mathbb{R}^n \times \mathbb{R}^n \rightarrow \mathbb{R}$ . Erfüllt eine Funktion  $\mathbf{x}^* \in \mathcal{X}$  die Euler-Lagrange Gleichungen (4.17) und die sogenannte Konvexitätsbedingung

$$\mathbf{d}^T \begin{bmatrix} \left( \frac{\partial^2 l}{\partial \mathbf{x}^2} \right) (t, \mathbf{x}^*(t), \dot{\mathbf{x}}^*(t)) & \left( \frac{\partial^2 l}{\partial \mathbf{x} \partial \dot{\mathbf{x}}} \right) (t, \mathbf{x}^*(t), \dot{\mathbf{x}}^*(t)) \\ \left( \frac{\partial^2 l}{\partial \mathbf{x} \partial \dot{\mathbf{x}}} \right) (t, \mathbf{x}^*(t), \dot{\mathbf{x}}^*(t)) & \left( \frac{\partial^2 l}{\partial \dot{\mathbf{x}}^2} \right) (t, \mathbf{x}^*(t), \dot{\mathbf{x}}^*(t)) \end{bmatrix} \mathbf{d} \geq 0 \quad \forall \mathbf{d} \in \mathbb{R}^{2n}, t \in [t_0, t_1], \quad (4.35)$$

d. h.  $l(t, \mathbf{x}(t), \dot{\mathbf{x}}(t))$  ist lokal konvex in  $\mathbf{x}(t)$  und  $\dot{\mathbf{x}}(t)$ , dann ist  $\mathbf{x}^*(t)$  ein lokales Minimum des Funktionals  $J$ . Ist  $l(t, \mathbf{x}(t), \dot{\mathbf{x}}(t))$  sogar strikt lokal konvex in  $\mathbf{x}(t)$  und  $\dot{\mathbf{x}}(t)$  (Gleichung (4.35) ist dann nur für  $\mathbf{d} = \mathbf{0}$  mit Gleichheit erfüllt), dann ist  $\mathbf{x}^*(t)$  ein striktes lokales Minimum.

Der Beweis zu Satz 4.4 findet sich z. B. in [4.1, 4.8].

Satz 4.2 behandelt das Lagrange Problem der Variationsrechnung (4.1). Im nächsten Schritt soll das Bolza Problem der Variationsrechnung (4.2) mit freier Endzeit näher untersucht werden.

**Satz 4.5** (Euler-Lagrange Gleichungen für freie Endzeit). *Gegeben sei das Funktional*

$$J(t_1, \mathbf{x}) = \varphi(t_1, \mathbf{x}(t_1)) + \int_{t_0}^{t_1} l(t, \mathbf{x}(t), \dot{\mathbf{x}}(t)) dt \quad (4.36)$$

mit der zulässigen Menge  $\mathcal{X} = \{(t_1, \mathbf{x}(t)) \in (t_0, T) \times (C^1[t_0, T])^n \mid \mathbf{x}(t_0) = \mathbf{x}_0\}$ , der hinreichend großen Zeit  $T \gg t_1$ , der stetig differenzierbaren Lagrangeschen Dichte  $l : \mathbb{R} \times \mathbb{R}^n \times \mathbb{R}^n \rightarrow \mathbb{R}$  und der stetig differenzierbaren Endkostenfunktion  $\varphi : \mathbb{R} \times \mathbb{R}^n \rightarrow \mathbb{R}$ .

Wenn  $(t_1^*, \mathbf{x}^*(t))$  ein (lokales) Minimum von  $J(\mathbf{x})$  auf  $\mathcal{X}$  bezeichnet, dann erfüllt  $\mathbf{x}^*(t)$  die Euler-Lagrange Gleichungen (4.17) im Intervall  $[t_0, t_1^*]$  und es gelten die Anfangsbedingung  $\mathbf{x}^*(t_0) = \mathbf{x}_0$  sowie die Transversalitätsbedingungen

$$\left[ \frac{\partial}{\partial \dot{\mathbf{x}}} l + \frac{\partial}{\partial \mathbf{x}} \varphi \right]_{t=t_1^*, \mathbf{x}=\mathbf{x}^*} = \mathbf{0}^T \quad (4.37a)$$

$$\left[ l - \left( \frac{\partial}{\partial \dot{\mathbf{x}}} l \right) \dot{\mathbf{x}} + \frac{\partial}{\partial t} \varphi \right]_{t=t_1^*, \mathbf{x}=\mathbf{x}^*} = 0 . \quad (4.37b)$$

*Beweis.* Wenn man die Endzeit  $t_1^*$  fixiert, dann folgt aus Satz 4.2 unmittelbar, dass die optimale Lösung  $\mathbf{x}^*(t)$  im Intervall  $[t_0, t_1^*]$  die Euler-Lagrange Gleichungen (4.17) erfüllt. Um die optimale Endzeit  $t_1^*$  zu berechnen, nimmt man an, dass  $\mathbf{x}(t)$  in einem hinreichend großen Intervall  $[t_0, T]$ ,  $T \gg t_1^*$  definiert ist und betrachtet den linearen Funktionenraum  $\mathbb{R} \times (C^1[t_0, T])^n$ . Die Gâteaux Ableitung gemäß Definition 4.1 wird dann in der Form

$$\begin{aligned} \delta J(t_1, \mathbf{x}; \xi_{t_1}, \xi_x) &:= \lim_{\eta \rightarrow 0} \frac{J(t_1 + \eta \xi_{t_1}, \mathbf{x} + \eta \xi_x) - J(t_1, \mathbf{x})}{\eta} \\ &= \left. \frac{d}{d\eta} J(t_1 + \eta \xi_{t_1}, \mathbf{x} + \eta \xi_x) \right|_{\eta=0} \end{aligned} \quad (4.38)$$

erweitert und die notwendige Bedingung für ein Minimum (4.15) von Satz 4.1 ausgewertet. Wendet man nun (4.38) für beliebiges  $\eta$  auf (4.36) an, so erhält man

$$\begin{aligned} &\frac{d}{d\eta} J(t_1^* + \eta \xi_{t_1}, \mathbf{x}^* + \eta \xi_x) = \\ &= \frac{d}{d\eta} \varphi(t_1^* + \eta \xi_{t_1}, \mathbf{x}^*(t_1^* + \eta \xi_{t_1}) + \eta \xi_x(t_1^* + \eta \xi_{t_1})) \\ &\quad + \frac{d}{d\eta} \int_{t_0}^{t_1^* + \eta \xi_{t_1}} l(t, \mathbf{x}^* + \eta \xi_x, \dot{\mathbf{x}}^* + \eta \dot{\xi}_x) dt \\ &= \left[ \frac{\partial}{\partial t} \varphi \xi_{t_1} + \frac{\partial}{\partial \mathbf{x}} \varphi \left[ \underbrace{\frac{\partial}{\partial t} \mathbf{x}}_{\dot{\mathbf{x}}} + \eta \underbrace{\frac{\partial}{\partial t} \xi_x}_{\dot{\xi}_x} \right] \xi_{t_1} + \frac{\partial}{\partial \mathbf{x}} \varphi \xi_x \right]_{t=t_1^* + \eta \xi_{t_1}, \mathbf{x}=\mathbf{x}^* + \eta \xi_x} \\ &\quad + \xi_{t_1} \left[ l(t, \mathbf{x}^* + \eta \xi_x, \dot{\mathbf{x}}^* + \eta \dot{\xi}_x) \right]_{t=t_1^* + \eta \xi_{t_1}} \\ &\quad + \int_{t_0}^{t_1^* + \eta \xi_{t_1}} \frac{d}{d\eta} l(t, \mathbf{x}^* + \eta \xi_x, \dot{\mathbf{x}}^* + \eta \dot{\xi}_x) dt \\ &= \left[ \frac{\partial}{\partial t} \varphi \xi_{t_1} + \frac{\partial}{\partial \mathbf{x}} \varphi (\dot{\mathbf{x}} + \eta \dot{\xi}_x) \xi_{t_1} + \frac{\partial}{\partial \mathbf{x}} \varphi \xi_x \right]_{t=t_1^* + \eta \xi_{t_1}, \mathbf{x}=\mathbf{x}^* + \eta \xi_x} \\ &\quad + \xi_{t_1} \left[ l(t, \mathbf{x}^* + \eta \xi_x, \dot{\mathbf{x}}^* + \eta \dot{\xi}_x) \right]_{t=t_1^* + \eta \xi_{t_1}} \end{aligned} \quad (4.39)$$

$$\begin{aligned}
& + \left[ \frac{\partial}{\partial \dot{\mathbf{x}}} l(t, \mathbf{x}^* + \eta \boldsymbol{\xi}_x, \dot{\mathbf{x}}^* + \eta \dot{\boldsymbol{\xi}}_x) \boldsymbol{\xi}_x \right]_{t_0}^{t_1^* + \eta \xi_{t_1}} \\
& + \int_{t_0}^{t_1^* + \eta \xi_{t_1}} \left( \frac{\partial}{\partial \mathbf{x}} l(t, \mathbf{x}^* + \eta \boldsymbol{\xi}_x, \dot{\mathbf{x}}^* + \eta \dot{\boldsymbol{\xi}}_x) \right. \\
& \left. - \frac{d}{dt} \left( \frac{\partial}{\partial \dot{\mathbf{x}}} l \right) (t, \mathbf{x}^* + \eta \boldsymbol{\xi}_x, \dot{\mathbf{x}}^* + \eta \dot{\boldsymbol{\xi}}_x) \right) \boldsymbol{\xi}_x dt .
\end{aligned}$$

Wertet man (4.39) für  $\eta = 0$  aus, so lautet die notwendige Optimalitätsbedingung

$$\begin{aligned}
\delta J(t_1^*, \mathbf{x}^*; \xi_{t_1}, \boldsymbol{\xi}_x) &= \frac{d}{d\eta} J(t_1^* + \eta \xi_{t_1}, \mathbf{x}^* + \eta \boldsymbol{\xi}_x) \Big|_{\eta=0} \\
&= \left[ \frac{\partial}{\partial t} \varphi \xi_{t_1} + \frac{\partial}{\partial \mathbf{x}} \varphi (\dot{\mathbf{x}} \xi_{t_1} + \boldsymbol{\xi}_x) + \xi_{t_1} l \right]_{t=t_1^*, \mathbf{x}=\mathbf{x}^*} + \left[ \frac{\partial}{\partial \dot{\mathbf{x}}} l(t, \mathbf{x}^*, \dot{\mathbf{x}}^*) \boldsymbol{\xi}_x \right]_{t_0}^{t_1^*} \\
&+ \int_{t_0}^{t_1^*} \left[ \frac{\partial}{\partial \mathbf{x}} l(t, \mathbf{x}^*, \dot{\mathbf{x}}^*) - \frac{d}{dt} \left( \frac{\partial}{\partial \dot{\mathbf{x}}} l \right) (t, \mathbf{x}^*, \dot{\mathbf{x}}^*) \right] \boldsymbol{\xi}_x dt = 0 .
\end{aligned} \tag{4.40}$$

Da der Anfangswert mit  $\mathbf{x}(t_0) = \mathbf{x}_0$  festgelegt ist, muss für eine zulässige Richtung  $\boldsymbol{\xi}_x$  die Bedingung  $\boldsymbol{\xi}_x(t_0) = \mathbf{0}$  gelten. Im Weiteren erfüllt die optimale Lösung  $\mathbf{x}^*(t)$  im Intervall  $[t_0, t_1^*]$  die Euler-Lagrange Gleichungen (4.17), weshalb sich (4.40) zu

$$\begin{aligned}
& \left[ \frac{\partial}{\partial t} \varphi \xi_{t_1} + \frac{\partial}{\partial \mathbf{x}} \varphi (\dot{\mathbf{x}} \xi_{t_1} + \boldsymbol{\xi}_x) + \xi_{t_1} l \right]_{t=t_1^*, \mathbf{x}=\mathbf{x}^*} \\
& + \left[ \frac{\partial}{\partial \dot{\mathbf{x}}} l(t, \mathbf{x}, \dot{\mathbf{x}}) (\boldsymbol{\xi}_x + \dot{\mathbf{x}} \xi_{t_1} - \dot{\mathbf{x}} \xi_{t_1}) \right]_{t=t_1^*, \mathbf{x}=\mathbf{x}^*} \\
&= \left[ l - \frac{\partial}{\partial \dot{\mathbf{x}}} l \dot{\mathbf{x}} + \frac{\partial}{\partial t} \varphi \right]_{t=t_1^*, \mathbf{x}=\mathbf{x}^*} \xi_{t_1} \\
& + \left[ \frac{\partial}{\partial \mathbf{x}} \varphi + \frac{\partial}{\partial \dot{\mathbf{x}}} l \right]_{t=t_1^*, \mathbf{x}=\mathbf{x}^*} (\dot{\mathbf{x}}^*(t_1^*) \xi_{t_1} + \boldsymbol{\xi}_x(t_1^*)) = 0
\end{aligned} \tag{4.41}$$

vereinfacht. Wenn die Endzeit  $t_1$  und der Endwert  $\mathbf{x}(t_1)$  frei sind, dann sind  $\xi_{t_1}$  und  $\boldsymbol{\xi}_x(t_1^*)$  unabhängig voneinander frei wählbar, weshalb (4.41) nur dann Null ist, wenn die *Transversalitätsbedingungen* (4.37) erfüllt sind.  $\square$

Das Ergebnis von Satz 4.5 lässt sich nun wie folgt verallgemeinern.

- (a) Wenn die *Endzeit*  $t_1$  *fest ist*, dann gilt  $\xi_{t_1} = 0$ , womit automatisch der erste Term von (4.41) verschwindet. Es liegt somit keine Transversalitätsbedingung (4.37b) vor.
  - (i) Wenn für eine Komponente  $x_k(t)$  von  $\mathbf{x}(t)$  gilt, dass *deren Endwert*  $x_k(t_1) = \bar{x}_k$  mit  $\bar{x}_k = \text{const.}$  *fest ist*, so folgt daraus  $\xi_{x,k}(t_1) = 0$ , womit automatisch der zugehörige Eintrag im zweiten Term von (4.41) verschwindet. Damit liegt für diese Komponente keine Transversalitätsbedingung vor. Dieser Fall entspricht dem Ergebnis von Satz 4.2.

- (ii) Wenn für eine Komponente  $x_k(t)$  von  $\mathbf{x}(t)$  gilt, dass deren Endwert  $x_k(t_1)$  frei ist, dann lautet die *Transversalitätsbedingung* gemäß (4.41) für diese Komponente

$$\left[ \frac{\partial}{\partial \dot{x}_k} l + \frac{\partial}{\partial x_k} \varphi \right]_{t=t_1^*, \mathbf{x}=\mathbf{x}^*} = 0. \quad (4.42)$$

- (b) Wenn die Endzeit  $t_1$  frei ist, dann muss die Transversalitätsbedingung (4.37b)

$$\left[ l - \left( \frac{\partial}{\partial \dot{\mathbf{x}}} l \right) \dot{\mathbf{x}} + \frac{\partial}{\partial t} \varphi \right]_{t=t_1^*, \mathbf{x}=\mathbf{x}^*} = 0 \quad (4.43)$$

gelten.

- (i) Wenn für eine Komponente  $x_k(t)$  von  $\mathbf{x}(t)$  gilt, dass deren Endwert  $x_k(t_1^*) = \bar{x}_k$  mit  $\bar{x}_k = \text{const.}$  fest ist, dann muss für diese Komponente eine zulässige Richtung  $(\xi_{t_1}, \xi_{x,k})$  die Bedingung

$$\bar{x}_k = x_k^*(t_1^* + \eta \xi_{t_1}) + \eta \xi_{x,k}(t_1^* + \eta \xi_{t_1}) \quad (4.44)$$

bzw.

$$0 = \frac{d}{d\eta} \bar{x}_k \Big|_{\eta=0} = \xi_{x,k}(t_1^*) + \xi_{t_1} \dot{x}_k^*(t_1^*) \quad (4.45)$$

erfüllen. Damit verschwindet der zugehörige Eintrag im zweiten Term von (4.41) und es liegt keine weitere Transversalitätsbedingung für diese Komponente vor.

- (ii) Wenn für eine Komponente  $x_k(t)$  von  $\mathbf{x}(t)$  gilt, dass deren Endwert  $x_k(t_1^*)$  frei ist, dann lautet, analog zum Fall (a)(ii), die *Transversalitätsbedingung* für diese Komponente

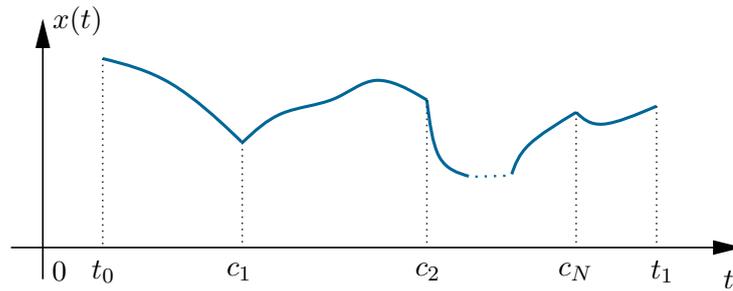
$$\left[ \frac{\partial}{\partial \dot{x}_k} l + \frac{\partial}{\partial x_k} \varphi \right]_{t=t_1^*, \mathbf{x}=\mathbf{x}^*} = 0. \quad (4.46)$$

### 4.1.3 Stückweise stetig differenzierbare Extremale

Bei den bisherigen Betrachtungen, siehe im Speziellen die Sätze 4.2 bis 4.5, wurde stets angenommen, dass  $\mathbf{x}(t)$  im Funktionenraum der im Intervall  $[t_0, t_1]$  (vektorwertigen) stetig differenzierbaren Funktionen  $(C^1[t_0, T])^n$  definiert ist. Im Weiteren soll dies auf den Funktionenraum der stückweise stetig differenzierbaren Funktionen  $(\hat{C}^1[t_0, T])^n$  erweitert werden, wobei zusätzlich die globale Stetigkeit vorausgesetzt wird. Man nennt nun eine reellwertige Funktion  $x(t) \in \hat{C}^1[t_0, t_1]$  *stückweise stetig differenzierbar*, wenn sie stetig ist und eine *Partitionierung*  $t_0 = c_0 < c_1 < \dots < c_{N+1} = t_1$  mit  $N < \infty$  so existiert, dass die Funktion  $x(t)$  in allen Intervallen  $(c_k, c_{k+1})$ ,  $k = 0, \dots, N$  stetig differenzierbar ist, siehe Abbildung 4.3. Die inneren Punkte  $c_1, \dots, c_N$  werden als *Eckpunkte von  $x(t)$*  bezeichnet. Für stückweise stetig differenzierbare Funktionen  $\hat{x}(t) \in \hat{C}^1[t_0, t_1]$  lauten die Normen gemäß (4.7)

$$\|\hat{\mathbf{x}}(t)\|_\infty := \max_{t_0 \leq t \leq t_1} \|\hat{\mathbf{x}}(t)\| \quad \text{und} \quad \|\hat{\mathbf{x}}(t)\|_{1,\infty} := \max_{t_0 \leq t \leq t_1} \|\hat{\mathbf{x}}(t)\| + \sup_{t \in \bigcup_{k=0}^N (c_k, c_{k+1})} \left\| \frac{d}{dt} \hat{\mathbf{x}}(t) \right\|. \quad (4.47)$$

Es gilt nun folgender Satz, welcher z. B. in [4.9] bewiesen wird.

Abbildung 4.3: Beispiel einer Funktion  $x(t) \in \hat{C}^1[t_0, t_1]$ .

**Satz 4.6** (Stückweise stetig vs. stetig differenzierbare Extremale). *Angenommen  $\mathbf{x}^* \in \mathcal{X}$  ist ein (lokales) Minimum des Funktional*

$$J(\mathbf{x}) = \int_{t_0}^{t_1} l(t, \mathbf{x}(t), \dot{\mathbf{x}}(t)) dt \quad (4.48)$$

mit der zulässigen Menge  $\mathcal{X} = \{\mathbf{x}(t) \in (C^1[t_0, t_1])^n \mid \mathbf{x}(t_0) = \mathbf{x}_0, \mathbf{x}(t_1) = \mathbf{x}_1\}$  und der stetig differenzierbaren Lagrangeschen Dichte  $l : \mathbb{R} \times \mathbb{R}^n \times \mathbb{R}^n \rightarrow \mathbb{R}$ , dann ist  $\mathbf{x}^* \in \hat{\mathcal{X}}$  auch ein (lokales) Minimum des Funktional (4.48) in der zulässigen Menge  $\hat{\mathcal{X}} = \{\mathbf{x}(t) \in (\hat{C}^1[t_0, t_1])^n \mid \mathbf{x}(t_0) = \mathbf{x}_0, \mathbf{x}(t_1) = \mathbf{x}_1\}$ . Handelt es sich um ein lokales Minimum, so ist der Begriff lokal bezüglich der gleichen Norm  $\|\cdot\|_\infty$  bzw.  $\|\cdot\|_{1,\infty}$  zu verstehen.

Man kann nun zeigen, dass eine extremale Lösung  $\hat{\mathbf{x}}^*(t) \in (\hat{C}^1[t_0, t_1])^n$  im gesamten Intervall  $[t_0, t_1]$  außer an den Eckpunkten  $c_1, \dots, c_N$  die Euler-Lagrange Gleichungen (4.17) und die Legendre-Bedingung (4.33) erfüllt. Die Transversalitätsbedingungen (4.42), (4.43) und (4.46) bleiben im Falle stückweise stetig differenzierbarer Extremale *unverändert*. Die Unstetigkeiten von  $\frac{d}{dt}\hat{\mathbf{x}}^*(t)$  an den Eckpunkten  $t = c_k$ ,  $k = 1, \dots, N$  unterliegen nun folgenden Einschränkungen:

**Satz 4.7** (Weierstrass-Erdmann Bedingungen). *Angenommen  $\hat{\mathbf{x}}^* \in \hat{\mathcal{X}}$  ist ein (lokales) Minimum des Funktional*

$$J(\hat{\mathbf{x}}) = \int_{t_0}^{t_1} l(t, \hat{\mathbf{x}}(t), \dot{\hat{\mathbf{x}}}(t)) dt \quad (4.49)$$

mit der zulässigen Menge  $\hat{\mathcal{X}} = \{\hat{\mathbf{x}}(t) \in (\hat{C}^1[t_0, t_1])^n \mid \hat{\mathbf{x}}(t_0) = \hat{\mathbf{x}}_0, \hat{\mathbf{x}}(t_1) = \hat{\mathbf{x}}_1\}$ , wobei die Lagrangesche Dichte  $l$  sowie die partiellen Ableitungen  $\frac{\partial}{\partial \hat{x}_i} l$  und  $\frac{\partial}{\partial \dot{\hat{x}}_i} l$  im Gebiet  $[t_0, t_1] \times \mathbb{R}^n \times \mathbb{R}^n$  stetig bezüglich ihrer Argumente  $t$ ,  $\hat{\mathbf{x}}(t)$  und  $\dot{\hat{\mathbf{x}}}(t)$  sind. Dann gilt für jeden Eckpunkt  $c \in (t_0, t_1)$  von  $\hat{\mathbf{x}}^*(t)$ , dass die Bedingungen

$$\left( \frac{\partial}{\partial \hat{\mathbf{x}}} l \right) (c, \hat{\mathbf{x}}^*(c), \dot{\hat{\mathbf{x}}}^*(c^-)) = \left( \frac{\partial}{\partial \hat{\mathbf{x}}} l \right) (c, \hat{\mathbf{x}}^*(c), \dot{\hat{\mathbf{x}}}^*(c^+)) \quad (4.50a)$$

$$H(c, \hat{\mathbf{x}}^*(c), \dot{\hat{\mathbf{x}}}^*(c^-)) = H(c, \hat{\mathbf{x}}^*(c), \dot{\hat{\mathbf{x}}}^*(c^+)) \quad (4.50b)$$

mit der Hamiltonfunktion

$$H(t, \hat{\mathbf{x}}(t), \dot{\hat{\mathbf{x}}}(t)) = \left( \frac{\partial}{\partial \dot{\hat{\mathbf{x}}}} l \right) (t, \hat{\mathbf{x}}(t), \dot{\hat{\mathbf{x}}}(t)) \dot{\hat{\mathbf{x}}}(t) - l(t, \hat{\mathbf{x}}(t), \dot{\hat{\mathbf{x}}}(t)) \quad (4.51)$$

erfüllt sind, wobei  $\dot{\hat{\mathbf{x}}}^*(c^-)$  und  $\dot{\hat{\mathbf{x}}}^*(c^+)$  den links- bzw. rechtsseitigen Grenzwert von  $\dot{\hat{\mathbf{x}}}^*(t)$  an der Stelle  $t = c$  bezeichnen.

Die Weierstrass-Erdmann Bedingungen besagen also, dass an den Eckpunkten einer (lokal) extremalen Trajektorie  $\hat{\mathbf{x}}^*(t) \in (\hat{C}^1[t_0, t_1])^n$  nur jene Unstetigkeiten von  $\dot{\hat{\mathbf{x}}}^*$  erlaubt sind, die die zeitliche Stetigkeit von  $\frac{\partial}{\partial \dot{\hat{\mathbf{x}}}} l$  und die zeitliche Stetigkeit der Hamiltonfunktion  $H$  erhalten. Die Weierstrass-Erdmann Bedingungen sind *notwendige* Optimalitätsbedingungen. Ihre Herleitung findet sich z. B. in [4.8].

*Beispiel 4.4.* Gesucht ist ein (lokales) Minimum  $x^* \in \mathcal{X}$  des Funktionals

$$J(x) = \int_{-1}^1 x^2(t)(1 - \dot{x}(t))^2 dt \quad (4.52)$$

in der zulässigen Menge  $\mathcal{X} = \{x(t) \in C^1[-1, 1] \mid x(-1) = 0, x(1) = 1\}$ . Da die Lagrange Dichte nicht explizit von der Zeit  $t$  abhängt, ist die Hamiltonfunktion

$$H = \left( \frac{\partial}{\partial \dot{x}} l \right) \dot{x} - l = -2x^2(1 - \dot{x})\dot{x} - x^2(1 - \dot{x})^2 = x^2(\dot{x}^2 - 1) = -k_1 \quad (4.53)$$

für alle Zeiten  $t \in [-1, 1]$  konstant mit der Konstanten  $k_1$  und damit eine Invariante des Systems, siehe auch (4.23). Ersetzt man  $x^2(t) = z(t)$  und  $2x(t)\dot{x}(t) = \dot{z}(t)$  in (4.53), so ergibt sich

$$z(t) - \frac{1}{4}\dot{z}^2(t) = k_1. \quad (4.54)$$

Die Lösung von (4.54) lautet

$$z(t) = (t + k_2)^2 + k_1 \quad (4.55)$$

mit der Konstanten  $k_2$ . Mit  $x(-1) = 0$  und  $x(1) = 1$  sowie  $z(t) = x^2(t)$  folgen die Konstanten  $k_1$  und  $k_2$  zu  $k_1 = -\left(\frac{3}{4}\right)^2$  und  $k_2 = \frac{1}{4}$  und die mögliche stationäre Lösung  $\bar{x}(t)$  des Kostenfunktionals (4.52) lautet

$$\bar{x}(t) = \pm \sqrt{\left(t + \frac{1}{4}\right)^2 - \left(\frac{3}{4}\right)^2}. \quad (4.56)$$

Die Wurzel liefert nur für  $t \geq \frac{1}{2}$  und  $t \leq -1$  ein reellwertiges Ergebnis, weshalb  $\bar{x}(t)$  keine stationäre Lösung von (4.52) in der zulässigen Menge  $\mathcal{X} = \{x(t) \in C^1[-1, 1] \mid x(-1) = 0, x(1) = 1\}$  darstellt.

Im nächsten Schritt soll das Kostenfunktional (4.52) in der zulässigen Menge  $\hat{\mathcal{X}} = \{\hat{x}(t) \in \hat{C}^1[-1, 1] \mid \hat{x}(-1) = 0, \hat{x}(1) = 1\}$  minimiert werden. Die Weierstrass-Erdmann Bedingung (4.50a) besagt nun, dass an einem Eckpunkt  $c \in (-1, 1)$  gilt

$$-2\hat{x}^2(c)[1 - \dot{\hat{x}}(c^-)] = -2\hat{x}^2(c)[1 - \dot{\hat{x}}(c^+)] \quad (4.57)$$

und folglich

$$\hat{x}^2(c)[\dot{\hat{x}}(c^+) - \dot{\hat{x}}(c^-)] = 0. \quad (4.58)$$

Da an einem Eckpunkt  $t = c$  gilt  $\dot{\hat{x}}(c^+) \neq \dot{\hat{x}}(c^-)$ , muss zur Erfüllung von (4.58) die Bedingung  $\hat{x}(c) = 0$  eingehalten werden. D. h. Eine Unstetigkeit in  $\hat{x}(t)$  kann nur an Stellen auftreten, an denen der Wert von  $\hat{x}(t)$  selbst identisch Null ist. Den minimalen Wert des Kostenfunktionals (4.52), nämlich den Wert Null, erhält man, wenn  $\hat{x}(t) = 0$  oder  $\hat{x}(t) = 1$  für alle  $t$  in  $[-1, 1]$  gilt. Außerdem sind die Randbedingungen  $\hat{x}(-1) = 0$  und  $\hat{x}(1) = 1$  zu erfüllen. Aus diesen Überlegungen folgt, dass für die optimale Lösung  $\hat{x}(t) = 0 \forall t \in [-1, c]$  und  $\hat{x}(t) = 1 \forall t \in (c, 1]$  gelten muss. Daraus ergibt sich der Umschaltzeitpunkt  $c = 0$  und die optimale Lösung

$$\hat{x}^*(t) = \begin{cases} 0 & \text{für } -1 \leq t \leq 0 \\ t & \text{für } 0 < t \leq 1. \end{cases} \quad (4.59)$$

Diese Lösung ist das eindeutige globale Minimum des Kostenfunktionals (4.52) in der zulässigen Menge  $\hat{\mathcal{X}} = \{\hat{x}(t) \in \hat{C}^1[-1, 1] \mid \hat{x}(-1) = 0, \hat{x}(1) = 1\}$ .

## 4.2 Entwurf von Optimalsteuerungen

### 4.2.1 Problemformulierung

Eine typische Optimalsteuerungsaufgabe besteht darin, für ein dynamisches System beschrieben durch die Differenzialgleichungen

$$\dot{\mathbf{x}} = \mathbf{f}(t, \mathbf{x}(t), \mathbf{u}(t)) \quad (4.60a)$$

mit der Zeit  $t \in \mathbb{R}$ , dem Zustand  $\mathbf{x} \in \mathbb{R}^n$ , dem Anfangszustand

$$\mathbf{x}(t_0) = \mathbf{x}_0 \quad (4.60b)$$

und dem Stelleingang  $\mathbf{u} \in \mathbb{R}^m$  eine geeignete Steuertrajektorie  $\mathbf{u}(t), t \in [t_0, t_1]$  so zu finden, dass ein Kostenfunktional der Form (siehe auch (4.2))

$$J(\mathbf{u}) = \varphi(t_0, \mathbf{x}(t_0), t_1, \mathbf{x}(t_1)) + \int_{t_0}^{t_1} l(t, \mathbf{x}(t), \mathbf{u}(t)) dt \quad (4.61)$$

bezüglich  $\mathbf{u}(t)$  minimiert wird und dabei allfällige Beschränkungen für  $\mathbf{x}(t)$  und  $\mathbf{u}(t)$  eingehalten werden. Die Abhängigkeit der Funktion  $\varphi$  von  $t_0$  und  $\mathbf{x}(t_0)$  ist nur dann relevant, wenn die Anfangsbedingung (4.60b) nicht vorhanden ist. Beim Kostenfunktional

$J(\mathbf{u})$  unterscheidet man im Allgemeinen zwischen der *Bolza-Form* (4.61), der *Lagrange-Form*

$$J(\mathbf{u}) = \int_{t_0}^{t_1} l(t, \mathbf{x}(t), \mathbf{u}(t)) dt \quad (4.62)$$

und der *Mayer-Form*

$$J(\mathbf{u}) = \varphi(t_0, \mathbf{x}(t_0), t_1, \mathbf{x}(t_1)) . \quad (4.63)$$

**Aufgabe 4.2.** Zeigen Sie, dass die Lagrange-Form in die Mayer-Form übergeführt werden kann, indem man einen zusätzlichen Zustand

$$\dot{x}_{n+1} = l(t, \mathbf{x}, \mathbf{u}), \quad x_{n+1}(t_0) = 0 \quad (4.64)$$

einführt und das Kostenfunktional in der Form  $J(\mathbf{u}) = x_{n+1}(t_1)$  anschreibt.

Zeigen Sie, dass die Mayer-Form in die Lagrange-Form übergeführt werden kann, indem man einen zusätzlichen Zustand

$$\dot{x}_{n+1} = 0, \quad x_{n+1}(t_0) = \frac{1}{t_1 - t_0} \varphi(t_0, \mathbf{x}(t_0), t_1, \mathbf{x}(t_1)) \quad (4.65)$$

einführt und das Kostenfunktional in der Form  $J(\mathbf{u}) = \int_{t_0}^{t_1} x_{n+1}(t) dt$  anschreibt.

Zeigen Sie, wie man eine Bolza-Form in die Mayer- oder Lagrange-Form überführt.

Bei den möglichen Beschränkungen unterscheidet man zwischen *Punktbeschränkungen*, beispielsweise *Endpunktbeschränkungen* der Form

$$\psi(t_1, \mathbf{x}(t_1)) \leq 0, \quad (4.66)$$

*Pfadbeschränkungen*

$$\psi(t, \mathbf{x}(t), \mathbf{u}(t)) \leq 0, \quad \forall t \in I \subset [t_0, t_1], \quad (4.67)$$

und *isoperimetrischen Beschränkungen*

$$\int_{t_0}^{t_1} \psi(t, \mathbf{x}(t), \mathbf{u}(t)) dt \leq 0 . \quad (4.68)$$

Häufig sind Pfadbeschränkungen (4.67) schwieriger zu berücksichtigen, wenn sie nicht von der Stellgröße abhängen.

### 4.2.2 Existenz einer Lösung

Im Skriptum Regelungssysteme 2 wurden hinreichende Bedingungen für die lokale Eindeutigkeit und Existenz der Lösung eines *Anfangswertproblems* angegeben. Sie besagen, wenn  $\mathbf{g}(t, \mathbf{x})$  stückweise stetig in  $t$  ist und der Lipschitz-Bedingung

$$\|\mathbf{g}(t, \mathbf{x}) - \mathbf{g}(t, \mathbf{y})\| \leq L \|\mathbf{x} - \mathbf{y}\|, \quad 0 < L < \infty \quad (4.69)$$

für alle  $\mathbf{x}, \mathbf{y} \in B_\gamma = \{\mathbf{x} \in \mathbb{R}^n \mid \|\mathbf{x} - \mathbf{x}_0\| \leq \gamma\}$  und alle  $t \in [t_0, t_0 + \tau]$  genügt, dann existiert ein  $\delta \in (0, \tau]$  so, dass das Anfangswertproblem

$$\dot{\mathbf{x}} = \mathbf{g}(t, \mathbf{x}), \quad \mathbf{x}(t_0) = \mathbf{x}_0 \quad (4.70)$$

für  $t \in [t_0, t_0 + \delta]$  genau eine Lösung besitzt. Da hier nur stückweise Stetigkeit von  $\mathbf{g}(t, \mathbf{x})$  in  $t$  gefordert wird, sind entsprechend (4.60) mit  $\mathbf{g}(t, \mathbf{x}) := \mathbf{f}(t, \mathbf{x}(t), \mathbf{u}(t))$  für die Stellgrößen  $\mathbf{u}(t)$  auch *stückweise stetige Funktionen* zugelassen, d. h.  $\mathbf{u}(t) \in (\hat{C}[t_0, t_1])^m$ . Man nennt eine reellwertige Funktion  $u(t) \in \hat{C}[t_0, t_1]$  *stückweise stetig*, wenn eine *Partitionierung*  $t_0 = c_0 < c_1 < \dots < c_{N+1} = t_1$  mit  $N < \infty$  so existiert, dass die Funktion  $u(t)$  in allen Intervallen  $(c_k, c_{k+1})$ ,  $k = 0, \dots, N$  stetig ist. Für stückweise stetige Stellgrößen  $\mathbf{u}(t)$  sind die zugehörigen Zustandsgrößen von (4.60) stückweise stetig differenzierbar, d. h.  $\mathbf{x}(t) \in (\hat{C}^1[t_0, t_1])^n$ , wobei die Eckpunkte mit den Unstetigkeitsstellen der Stellgrößen übereinstimmen. Wie im Skriptum Regelungssysteme 2 beschrieben, ist die die Stetigkeit von  $\mathbf{g}(t, \mathbf{x})$  und  $\left(\frac{\partial \mathbf{g}}{\partial \mathbf{x}}\right)(t, \mathbf{x})$  bezüglich  $\mathbf{x}$  auf der Menge  $[t_0, t_0 + \tau] \times B_\gamma$  *hinreichend* dafür, dass  $\mathbf{g}(t, \mathbf{x})$  die Lipschitz-Bedingung (4.69) lokal erfüllt.

Bei den meisten praktischen Anwendungen unterliegen die Stellgrößen gewissen Beschränkungen, d. h.  $\mathbf{u}(t) \in U \subset \mathbb{R}^m$ . Häufig auftretende Beschränkungen sind z. B. sogenannte *box constraints*

$$u_i^- \leq u_i \leq u_i^+, \quad i = 1, \dots, m. \quad (4.71)$$

Eine stückweise stetige Stellgröße  $\mathbf{u}(t)$  im Intervall  $t_0 \leq t \leq t_1$  mit  $\mathbf{u}(t) \in U$  für alle  $t \in [t_0, t_1]$  bezeichnet man im Weiteren als *zulässige Stellgröße*. Für das Folgende sei angenommen, dass  $\bar{\mathbf{x}}(t; \mathbf{x}_0, \mathbf{u}(t))$  die Lösung von (4.60) zum Zeitpunkt  $t$  für den Anfangswert  $\mathbf{x}(t_0) = \mathbf{x}_0$  und die Stellgröße  $\mathbf{u}(t)$ ,  $t_0 \leq t \leq t_1$  bezeichnet. Dann nennt man eine zulässige Stellgröße *realisierbar*, wenn  $\bar{\mathbf{x}}(t; \mathbf{x}_0, \mathbf{u}(t))$  im gesamten Intervall  $t_0 \leq t \leq t_1$  definiert ist und sämtliche Beschränkungen einhält.

Das Problem bei der Existenz einer Lösung des *Optimalsteuerungsproblems* besteht häufig darin, dass die Menge der realisierbaren Lösungen *nicht kompakt* ist. Es könnte beispielsweise passieren, dass die Lösung von (4.60) innerhalb des Optimierungsintervalls  $[t_0, t_1]$  nach Unendlich strebt und in Folge das Kostenfunktional unendlich wird. Dieses Phänomen ist auch unter dem Namen *finite escape time* bekannt. Um derartige Fälle auszuschließen, fordert man oft, dass die Lösungen des dynamischen Systems (4.60) beschränkt sind, also dass gilt

$$\|\bar{\mathbf{x}}(t; \mathbf{x}_0, \mathbf{u}(t))\| \leq \alpha, \quad \forall t \in [t_0, t_1] \quad (4.72)$$

für ein finites  $\alpha > 0$ . Man beachte, dass bezüglich  $\mathbf{x}$  affine Systeme der Form

$$\dot{\mathbf{x}} = \mathbf{A}(t, \mathbf{u})\mathbf{x} + \mathbf{b}(t, \mathbf{u}), \quad \mathbf{x}(t_0) = \mathbf{x}_0 \quad (4.73)$$

diese Eigenschaft stets erfüllen und in endlicher Zeit nicht nach Unendlich streben können.

Wenn das Optimierungsintervall selbst unendlich ist, dann ist die Menge der realisierbaren Stellgrößen unbeschränkt und nicht kompakt, weshalb stets ein kompaktes (finites) Zeitintervall  $[t_0, T]$  mit hinreichend großem  $T > t_1$  gewählt werden sollte. Dies wird anhand des nachfolgenden Beispiels näher erläutert.

**Beispiel 4.5.** Gegeben ist eine Punktmasse  $m$ , die über eine Kraft  $u(t)$  mit  $0 \leq u(t) \leq 1$  für alle  $t$  im Optimierungsintervall  $[t_0, t_1]$  beschleunigt wird. Die Aufgabe besteht nun darin, die Stellgröße  $u(t)$  so zu bestimmen, dass ausgehend von der Anfangsposition  $x(t_0) = x_0$  die Masse nach der Zeit  $t = t_1$  die Position  $x(t_1) = x_1$  erreicht und dabei

das Kostenfunktional

$$J(u) = \int_{t_0}^{t_1} u^2(t) dt \quad (4.74)$$

minimiert. Man erkennt unmittelbar, dass für  $x_1 > x_0$  die Stellgröße  $u(t) \equiv 0$  nicht realisierbar ist und somit für jede realisierbare Stellgröße  $J(u) > 0$  gelten muss. Betrachtet man nun die Folge der realisierbaren konstanten Stellgrößen  $u_k(t) = \frac{1}{k}$ ,  $k \geq 1$  für alle  $t \geq t_0$ , so erhält man als Lösung des Differentialgleichungssystems

$$\dot{x}_k = v_k, \quad x_k(t_0) = x_0 \quad (4.75a)$$

$$m\dot{v}_k = u_k, \quad v_k(t_0) = 0 \quad (4.75b)$$

das Ergebnis

$$x_k(t) = x_0 + \frac{1}{2mk}(t - t_0)^2 \quad (4.76a)$$

$$v_k(t) = \frac{1}{mk}(t - t_0) . \quad (4.76b)$$

Die Zeit  $t_1$ , nach der der Zustand  $x_k(t_1) = x_1$  erreicht wird, errechnet sich direkt aus (4.76a) zu

$$t_{1,k} = t_0 + \sqrt{2mk\sqrt{x_1 - x_0}} . \quad (4.77)$$

Damit erhält man für  $u_k(t) = \frac{1}{k}$  im Optimierungsintervall  $[t_0, t_{1,k}]$  den Wert des Kostenfunktionals zu

$$J(u_k) = \int_{t_0}^{t_{1,k}} \frac{1}{k^2} dt = \sqrt{\frac{2m}{k^3}} \sqrt{x_1 - x_0} . \quad (4.78)$$

Es gilt nun  $J(u_k) \rightarrow 0$  für  $k \rightarrow \infty$  und damit  $t_{1,k} \rightarrow \infty$ . Man erkennt also, dass gilt  $\inf J(u_k) = 0$ , d. h. das Problem hat *kein Minimum*.

Damit die Existenz einer Lösung des *Optimalsteuerungsproblems* auch tatsächlich gewährleistet ist, müssen weitere Einschränkungen der zulässigen Steuerungen vorgenommen werden. Zwei Möglichkeiten sollen im Folgenden kurz aufgezeigt werden. Einerseits besteht die Möglichkeit, zu fordern, dass die Stellgröße der *zusätzlichen Lipschitz-Bedingung*

$$\|\mathbf{u}(t) - \mathbf{u}(s)\| \leq L_u |t - s|, \quad 0 < L_u < \infty, \quad \forall s, t \in [t_0, t_1] \quad (4.79)$$

genügt und andererseits kann die Klasse der zulässigen Stellgrößen auf die *stückweise konstanten Stellgrößen* mit einer finiten Anzahl an Unstetigkeitsstellen eingeschränkt werden.

### 4.2.3 Variationsformulierung

Im Folgenden werden die notwendigen Optimalitätsbedingungen erster Ordnung für ein Optimalsteuerungsproblem mit fester Endzeit und freiem Endwert formuliert.

**Satz 4.8** (Optimalsteuerungsproblem mit fester Endzeit und freiem Endwert). *Gesucht ist die Stellgröße  $\mathbf{u} \in (C[t_0, t_1])^m$  so, dass das Kostenfunktional (Bolza-Form)*

$$J(\mathbf{u}) = \varphi(\mathbf{x}(t_1)) + \int_{t_0}^{t_1} l(t, \mathbf{x}(t), \mathbf{u}(t)) dt \quad (4.80)$$

*unter der Gleichungsbeschränkung (dynamisches System)*

$$\dot{\mathbf{x}} = \mathbf{f}(t, \mathbf{x}(t), \mathbf{u}(t)), \quad \mathbf{x}(t_0) = \mathbf{x}_0 \quad (4.81)$$

*mit fester Anfangszeit  $t_0$  und fester Endzeit  $t_1 > t_0$  minimiert wird. Dabei wird angenommen, dass  $l$  und  $\mathbf{f}$  stetig in  $t$  und stetig differenzierbar bezüglich  $\mathbf{x}$  und  $\mathbf{u}$  für alle  $(t, \mathbf{x}, \mathbf{u}) \in [t_0, t_1] \times \mathbb{R}^n \times \mathbb{R}^m$  sind und die Funktion  $\varphi$  stetig und stetig differenzierbar bezüglich  $\mathbf{x}_1$  für alle  $\mathbf{x}_1 \in \mathbb{R}^n$  ist. Wenn  $\mathbf{u}^*(t) \in (C[t_0, t_1])^m$  die optimale Lösung des Optimierungsproblems bezeichnet und  $\mathbf{x}^*(t) \in (C^1[t_0, t_1])^n$  die zugehörige Lösung des Anfangswertproblems (4.81) ist, dann existiert ein  $\boldsymbol{\lambda}^*(t) \in (C^1[t_0, t_1])^n$  so, dass gilt*

$$\dot{\mathbf{x}}^* = \mathbf{f}(t, \mathbf{x}^*(t), \mathbf{u}^*(t)), \quad \mathbf{x}^*(t_0) = \mathbf{x}_0 \quad (4.82a)$$

$$\dot{\boldsymbol{\lambda}}^* = - \left( \frac{\partial}{\partial \mathbf{x}} l \right)^T (t, \mathbf{x}^*(t), \mathbf{u}^*(t)) - \left( \frac{\partial}{\partial \mathbf{x}} \mathbf{f} \right)^T (t, \mathbf{x}^*(t), \mathbf{u}^*(t)) \boldsymbol{\lambda}^*(t) \quad (4.82b)$$

$$\boldsymbol{\lambda}^*(t_1) = \left( \frac{d}{d\mathbf{x}_1} \varphi \right)^T (\mathbf{x}^*(t_1))$$

$$\mathbf{0} = \left( \frac{\partial}{\partial \mathbf{u}} l \right)^T (t, \mathbf{x}^*(t), \mathbf{u}^*(t)) + \left( \frac{\partial}{\partial \mathbf{u}} \mathbf{f} \right)^T (t, \mathbf{x}^*(t), \mathbf{u}^*(t)) \boldsymbol{\lambda}^*(t) \quad (4.82c)$$

*für  $t_0 \leq t \leq t_1$ . Die Gleichungen (4.82) werden als Euler-Lagrange Gleichungen des Optimalsteuerungsproblems und  $\boldsymbol{\lambda}^*(t)$  als adjungierter Zustand oder Kozustand bezeichnet.*

*Beweis.* Für den Beweis dieses Satzes wird im ersten Schritt die Gâteaux Ableitung  $\delta J(\mathbf{u}; \boldsymbol{\xi}_u)$  mit  $\mathbf{u}(t), \boldsymbol{\xi}_u(t) \in (C[t_0, t_1])^m$  allgemein berechnet. Im zweiten Schritt wird diese Gâteaux Ableitung in die Optimalitätsbedingung gemäß Satz 4.1 eingesetzt.

Es sei  $\mathbf{x}(t)$  die Lösungstrajektorie von (4.81) für einen gegebenen Eingang  $\mathbf{u}(t)$ . Die Gâteaux Ableitung von  $\mathbf{x}(t)$  bezüglich  $\boldsymbol{\xi}_u$  am Punkt  $\mathbf{u}$  soll mit  $\delta \mathbf{x}(\mathbf{u}; \boldsymbol{\xi}_u)|_t \in \mathbb{R}^n$  bezeichnet werden. Wegen der Anfangsbedingung  $\mathbf{x}(t_0) = \mathbf{x}_0$  gilt

$$\delta \mathbf{x}(\mathbf{u}; \boldsymbol{\xi}_u)|_{t_0} = \mathbf{0} . \quad (4.83a)$$

Die Gâteaux Ableitung der Differentialgleichung  $\dot{\mathbf{x}} = \mathbf{f}(t, \mathbf{x}(t), \mathbf{u}(t))$  lautet

$$\delta \dot{\mathbf{x}}(\mathbf{u}; \boldsymbol{\xi}_u)|_t = \frac{\partial}{\partial \mathbf{x}} \mathbf{f}(t, \mathbf{x}(t), \mathbf{u}(t)) \delta \mathbf{x}(\mathbf{u}; \boldsymbol{\xi}_u)|_t + \frac{\partial}{\partial \mathbf{u}} \mathbf{f}(t, \mathbf{x}(t), \mathbf{u}(t)) \boldsymbol{\xi}_u(t) . \quad (4.83b)$$

Um aus dem linearen (zeitvarianten) Anfangswertproblem (4.83) die Gâteaux Ableitung  $\delta \mathbf{x}(\mathbf{u}; \boldsymbol{\xi}_u)|_\tau$  zum Zeitpunkt  $\tau \in [t_0, t_1]$  auszurechnen, betrachte man die zur

Dynamikmatrix  $\frac{\partial}{\partial \mathbf{x}} \mathbf{f}(t, \mathbf{x}(t), \mathbf{u}(t))$  gehörige Transitionsmatrix

$$\Phi(\tau, t) \quad (4.84)$$

für das Zeitintervall  $[t, \tau]$ . Mit dieser Transitionsmatrix folgt die Lösung von (4.83) in der Form

$$\delta \mathbf{x}(\mathbf{u}; \boldsymbol{\xi}_u)|_\tau = \int_{t_0}^{\tau} \Phi(\tau, t) \frac{\partial}{\partial \mathbf{u}} \mathbf{f}(t, \mathbf{x}(t), \mathbf{u}(t)) \boldsymbol{\xi}_u(t) dt . \quad (4.85)$$

Für die gesuchte Gâteaux Ableitung  $\delta J(\mathbf{u}; \boldsymbol{\xi}_u)$  an einem allgemeinen Punkt  $\mathbf{u} \in (C[t_0, t_1])^m$  bezüglich  $\boldsymbol{\xi}_u$  ergibt sich daher unter Verwendung der Kettenregel

$$\begin{aligned} & \delta J(\mathbf{u}; \boldsymbol{\xi}_u) \\ &= \frac{d}{d\mathbf{x}_1} \varphi(\mathbf{x}(t_1)) \delta \mathbf{x}(\mathbf{u}; \boldsymbol{\xi}_u)|_{t_1} \\ & \quad + \int_{t_0}^{t_1} \frac{\partial}{\partial \mathbf{x}} l(\tau, \mathbf{x}(\tau), \mathbf{u}(\tau)) \delta \mathbf{x}(\mathbf{u}; \boldsymbol{\xi}_u)|_\tau + \frac{\partial}{\partial \mathbf{u}} l(\tau, \mathbf{x}(\tau), \mathbf{u}(\tau)) \boldsymbol{\xi}_u(\tau) d\tau \\ &= \frac{d}{d\mathbf{x}_1} \varphi(\mathbf{x}(t_1)) \int_{t_0}^{t_1} \Phi(t_1, t) \frac{\partial}{\partial \mathbf{u}} \mathbf{f}(t, \mathbf{x}(t), \mathbf{u}(t)) \boldsymbol{\xi}_u(t) dt \\ & \quad + \int_{t_0}^{t_1} \left( \frac{\partial}{\partial \mathbf{x}} l(\tau, \mathbf{x}(\tau), \mathbf{u}(\tau)) \int_{t_0}^{\tau} \Phi(\tau, t) \frac{\partial}{\partial \mathbf{u}} \mathbf{f}(t, \mathbf{x}(t), \mathbf{u}(t)) \boldsymbol{\xi}_u(t) dt \right. \\ & \quad \quad \left. + \frac{\partial}{\partial \mathbf{u}} l(\tau, \mathbf{x}(\tau), \mathbf{u}(\tau)) \boldsymbol{\xi}_u(\tau) \right) d\tau \quad (4.86) \\ &= \int_{t_0}^{t_1} \frac{d}{d\mathbf{x}_1} \varphi(\mathbf{x}(t_1)) \Phi(t_1, t) \frac{\partial}{\partial \mathbf{u}} \mathbf{f}(t, \mathbf{x}(t), \mathbf{u}(t)) \boldsymbol{\xi}_u(t) dt \\ & \quad + \int_{t_0}^{t_1} \int_t^{t_1} \frac{\partial}{\partial \mathbf{x}} l(\tau, \mathbf{x}(\tau), \mathbf{u}(\tau)) \Phi(\tau, t) \frac{\partial}{\partial \mathbf{u}} \mathbf{f}(t, \mathbf{x}(t), \mathbf{u}(t)) \boldsymbol{\xi}_u(t) d\tau dt \\ & \quad + \int_{t_0}^{t_1} \frac{\partial}{\partial \mathbf{u}} l(t, \mathbf{x}(t), \mathbf{u}(t)) \boldsymbol{\xi}_u(t) dt \\ &= \int_{t_0}^{t_1} \left[ \frac{d}{d\mathbf{x}_1} \varphi(\mathbf{x}(t_1)) \Phi(t_1, t) \frac{\partial}{\partial \mathbf{u}} \mathbf{f}(t, \mathbf{x}(t), \mathbf{u}(t)) + \frac{\partial}{\partial \mathbf{u}} l(t, \mathbf{x}(t), \mathbf{u}(t)) \right. \\ & \quad \left. + \int_t^{t_1} \frac{\partial}{\partial \mathbf{x}} l(\tau, \mathbf{x}(\tau), \mathbf{u}(\tau)) \Phi(\tau, t) d\tau \frac{\partial}{\partial \mathbf{u}} \mathbf{f}(t, \mathbf{x}(t), \mathbf{u}(t)) \right] \boldsymbol{\xi}_u(t) dt . \end{aligned}$$

Aufgrund von Satz 4.1 muss am optimalen Punkt  $\mathbf{u}^* \in (C[t_0, t_1])^m$  die Bedingung  $\delta J(\mathbf{u}^*; \boldsymbol{\xi}_u) = 0$  für alle zulässigen Richtungen  $\boldsymbol{\xi}_u \in (C[t_0, t_1])^m$  erfüllt sein. Gemäß dem Fundamentallemma der Variationsrechnung Lemma 4.2 folgt daher aus (4.86)

$$\begin{aligned} & \frac{d}{d\mathbf{x}_1} \varphi(\mathbf{x}^*(t_1)) \Phi^*(t_1, t) \frac{\partial}{\partial \mathbf{u}} \mathbf{f}(t, \mathbf{x}^*(t), \mathbf{u}^*(t)) + \frac{\partial}{\partial \mathbf{u}} l(t, \mathbf{x}^*(t), \mathbf{u}^*(t)) \\ & + \int_t^{t_1} \frac{\partial}{\partial \mathbf{x}} l(\tau, \mathbf{x}^*(\tau), \mathbf{u}^*(\tau)) \Phi^*(\tau, t) d\tau \frac{\partial}{\partial \mathbf{u}} \mathbf{f}(t, \mathbf{x}^*(t), \mathbf{u}^*(t)) = \mathbf{0} \quad (4.87) \end{aligned}$$

für alle  $t \in [t_0, t_1]$ . Die Transitionsmatrix (4.84) ist auch für das lineare Anfangswertproblem (4.82b) anwendbar. In diesem Fall lautet die zur Dynamikmatrix  $-\left(\frac{\partial}{\partial \mathbf{x}} \mathbf{f}\right)^T(t, \mathbf{x}^*(t), \mathbf{u}^*(t))$  und zum Zeitintervall  $[\tau, t]$  gehörige Transitionsmatrix  $\Phi^{*T}(\tau, t)$ . Damit ergibt sich die Lösung von (4.82b) in der Form

$$\begin{aligned} \boldsymbol{\lambda}^*(t) &= \Phi^{*T}(t_1, t) \left( \frac{d}{d\mathbf{x}_1} \varphi \right)^T(\mathbf{x}^*(t_1)) - \int_{t_1}^t \Phi^{*T}(\tau, t) \left( \frac{\partial}{\partial \mathbf{x}} l \right)^T(\tau, \mathbf{x}^*(\tau), \mathbf{u}^*(\tau)) d\tau \\ &= \Phi^{*T}(t_1, t) \left( \frac{d}{d\mathbf{x}_1} \varphi \right)^T(\mathbf{x}^*(t_1)) + \int_t^{t_1} \Phi^{*T}(\tau, t) \left( \frac{\partial}{\partial \mathbf{x}} l \right)^T(\tau, \mathbf{x}^*(\tau), \mathbf{u}^*(\tau)) d\tau . \end{aligned} \quad (4.88)$$

Einsetzen dieser Lösung in (4.82c) zeigt die Äquivalenz zwischen den Bedingungen (4.82c) und (4.87).  $\square$

Aus (4.82) folgt, dass sich die notwendigen Optimalitätsbedingungen für das Optimalsteuerungsproblem (4.80) und (4.81) aus  $2n$  Differentialgleichungen in  $\mathbf{x}^*$  und  $\boldsymbol{\lambda}^*$  und  $m$  algebraischen Gleichungen zusammensetzen. Da für die Differentialgleichung in  $\mathbf{x}^*$  der Wert zum Anfangszeitpunkt  $t = t_0$  und für die Differentialgleichung in  $\boldsymbol{\lambda}^*$  der Wert zum Endzeitpunkt  $t = t_1$  gegeben ist, handelt es sich um ein *Zweipunkttrandwertproblem*. Analog zum Lagrange-Multiplikator in Satz 3.7 lässt sich der adjungierte Zustand  $\boldsymbol{\lambda}(t)$  gemäß (4.88) in der Form interpretieren, dass  $\boldsymbol{\lambda}(t)$  (zu einem bestimmten Zeitpunkt  $t$ ) der *Sensitivität des Kostenfunktional* (4.80) bezüglich einer (sprungförmigen) Änderung des Zustandes  $\mathbf{x}(t)$  (zum selben Zeitpunkt  $t$ ) entspricht.

Um die Gâteaux Ableitung  $\delta J(\mathbf{u}; \boldsymbol{\xi}_u)$  und die Optimalitätsbedingungen von Satz 4.8 alternativ mit Hilfe eines Lagrangefunktional zu formulieren, wird zunächst in Erweiterung zur Definition 4.1 der Begriff der *partiellen Gâteaux Ableitung* definiert.

**Definition 4.3 (Partielle Gâteaux Ableitung).** Die *partielle Gâteaux Ableitung des Funktional*  $J(\mathbf{x})$  am Punkt  $\mathbf{x} \in \mathcal{V}$  bezüglich  $x_i$  mit  $i \in \{1, \dots, n\}$  in Richtung  $\boldsymbol{\xi} = [0 \ \dots \ 0 \ \xi_i \ 0 \ \dots \ 0]^T \in \mathcal{V}$  (d. h.  $\xi_j = 0 \ \forall j \neq i$ ) ist in der Form

$$\delta_{x_i} J(\mathbf{x}; \boldsymbol{\xi}) := \lim_{\eta \rightarrow 0} \frac{J(\mathbf{x} + \eta \boldsymbol{\xi}) - J(\mathbf{x})}{\eta} = \left. \frac{d}{d\eta} J(\mathbf{x} + \eta \boldsymbol{\xi}) \right|_{\eta=0} \quad (4.89)$$

definiert.

Wird nun für das Kostenfunktional (4.80) mit den Gleichungsbeschränkungen (4.81) das *Lagrangefunktional*

$$\begin{aligned} L(\mathbf{x}, \mathbf{u}, \boldsymbol{\lambda}) &= J(\mathbf{u}) + \int_{t_0}^{t_1} \boldsymbol{\lambda}^T(t) (\mathbf{f}(t, \mathbf{x}(t), \mathbf{u}(t)) - \dot{\mathbf{x}}(t)) dt \\ &= \varphi(\mathbf{x}(t_1)) + \int_{t_0}^{t_1} l(t, \mathbf{x}(t), \mathbf{u}(t)) + \boldsymbol{\lambda}^T(t) (\mathbf{f}(t, \mathbf{x}(t), \mathbf{u}(t)) - \dot{\mathbf{x}}(t)) dt \end{aligned} \quad (4.90)$$

eingeführt (vgl. dazu die Lagrangefunktion (3.24) für ein statisches Optimierungsproblem mit Gleichungsbeschränkungen), so ergibt sich die Gâteaux Ableitung  $\delta J(\mathbf{u}; \boldsymbol{\xi}_u)$  an einem

allgemeinen Punkt  $\mathbf{u} \in (C[t_0, t_1])^m$  bezüglich  $\boldsymbol{\xi}_u(t) \in (C[t_0, t_1])^m$  in der Form

$$\begin{aligned} \delta J(\mathbf{u}; \boldsymbol{\xi}_u) &= \delta_{\mathbf{u}} L(\mathbf{x}, \mathbf{u}, \boldsymbol{\lambda}; \boldsymbol{\xi}_u) \\ &= \int_{t_0}^{t_1} \left[ \frac{\partial}{\partial \mathbf{u}} l(t, \mathbf{x}(t), \mathbf{u}(t)) + \boldsymbol{\lambda}^T(t) \frac{\partial}{\partial \mathbf{u}} \mathbf{f}(t, \mathbf{x}(t), \mathbf{u}(t)) \right] \boldsymbol{\xi}_u(t) dt, \end{aligned} \quad (4.91a)$$

wobei die Bedingungen

$$\begin{aligned} \mathbf{0} &= \delta_{\mathbf{x}} L(\mathbf{x}, \mathbf{u}, \boldsymbol{\lambda}; \boldsymbol{\xi}_x) \\ &= \frac{d}{d\mathbf{x}_1} \varphi(\mathbf{x}(t_1)) \boldsymbol{\xi}_x(t_1) \\ &\quad + \int_{t_0}^{t_1} \frac{\partial}{\partial \mathbf{x}} l(t, \mathbf{x}(t), \mathbf{u}(t)) \boldsymbol{\xi}_x(t) + \boldsymbol{\lambda}^T(t) \left( \frac{\partial}{\partial \mathbf{x}} \mathbf{f}(t, \mathbf{x}(t), \mathbf{u}(t)) \boldsymbol{\xi}_x(t) - \dot{\boldsymbol{\xi}}_x(t) \right) dt \end{aligned} \quad (4.91b)$$

$$\begin{aligned} &= \left[ \frac{d}{d\mathbf{x}_1} \varphi(\mathbf{x}(t_1)) - \boldsymbol{\lambda}^T(t_1) \right] \boldsymbol{\xi}_x(t_1) + \boldsymbol{\lambda}^T(t_0) \boldsymbol{\xi}_x(t_0) \\ &\quad + \int_{t_0}^{t_1} \left[ \frac{\partial}{\partial \mathbf{x}} l(t, \mathbf{x}(t), \mathbf{u}(t)) + \boldsymbol{\lambda}^T(t) \frac{\partial}{\partial \mathbf{x}} \mathbf{f}(t, \mathbf{x}(t), \mathbf{u}(t)) + \dot{\boldsymbol{\lambda}}^T(t) \right] \boldsymbol{\xi}_x(t) dt \\ \mathbf{0} &= \delta_{\boldsymbol{\lambda}} L(\mathbf{x}, \mathbf{u}, \boldsymbol{\lambda}; \boldsymbol{\xi}_\lambda) = \int_{t_0}^{t_1} \boldsymbol{\xi}_\lambda^T(t) [\mathbf{f}(t, \mathbf{x}(t), \mathbf{u}(t)) - \dot{\mathbf{x}}(t)] dt \end{aligned} \quad (4.91c)$$

für alle zulässigen Richtungen  $\boldsymbol{\xi}_x(t) \in (C[t_0, t_1])^n$  mit  $\boldsymbol{\xi}_x(t_0) = \mathbf{0}$  zufolge der Anfangsbedingung  $\mathbf{x}(t_0) = \mathbf{x}_0$  und für alle zulässigen Richtungen  $\boldsymbol{\xi}_\lambda(t) \in (C[t_0, t_1])^n$  einzuhalten sind. Gemäß Fundamentallemma der Variationsrechnung Lemma 4.2 folgt aus (4.91b)

$$\begin{aligned} \dot{\boldsymbol{\lambda}} &= - \left( \frac{\partial}{\partial \mathbf{x}} l \right)^T (t, \mathbf{x}(t), \mathbf{u}(t)) - \left( \frac{\partial}{\partial \mathbf{x}} \mathbf{f} \right)^T (t, \mathbf{x}(t), \mathbf{u}(t)) \boldsymbol{\lambda}(t) \\ \boldsymbol{\lambda}(t_1) &= \left( \frac{d}{d\mathbf{x}_1} \varphi \right)^T (\mathbf{x}(t_1)) \end{aligned} \quad (4.92)$$

und aus (4.91c) die Differentialgleichung (4.81). Um zu sehen, dass (4.91) tatsächlich genau die Gâteaux Ableitung  $\delta J(\mathbf{u}; \boldsymbol{\xi}_u)$  gemäß (4.86) liefert, kann die Lösung  $\boldsymbol{\lambda}(t)$  von (4.92) analog zu (4.88) mit Hilfe der zur Dynamikmatrix  $-\left(\frac{\partial}{\partial \mathbf{x}} \mathbf{f}\right)^T(t, \mathbf{x}(t), \mathbf{u}(t))$  und zum Zeitintervall  $[\tau, t]$  gehörigen Transitionsmatrix  $\Phi^T(\tau, t)$  formuliert und in (4.91a) eingesetzt werden. Ein Vergleich zwischen der hier beschriebenen Berechnung der Gâteaux Ableitung  $\delta J(\mathbf{u}; \boldsymbol{\xi}_u)$  anhand des Lagrangefunktional  $L(\mathbf{x}, \mathbf{u}, \boldsymbol{\lambda})$  mit der Berechnung des reduzierten Gradienten mit Hilfe der Lagrangefunktion in der beschränkten statischen Optimierung gemäß Lemma 3.2 zeigt die strukturelle Ähnlichkeit dieser beiden Methoden. Da am optimalen Punkt  $\mathbf{u}^* \in (C[t_0, t_1])^m$  gemäß Satz 4.1 die Bedingung  $\delta J(\mathbf{u}^*; \boldsymbol{\xi}_u) = 0$  für alle zulässigen Richtungen  $\boldsymbol{\xi}_u(t) \in (C[t_0, t_1])^m$  erfüllt sein muss, folgen aus (4.91a), (4.92) und (4.81) unter Verwendung von Lemma 4.2 genau die Optimalitätsbedingungen (4.82) von Satz 4.8 (Euler-Lagrange Gleichungen).

Alternativ lassen sich diese mit Hilfe der *Hamiltonfunktion*

$$H(t, \mathbf{x}, \mathbf{u}, \boldsymbol{\lambda}) = l(t, \mathbf{x}, \mathbf{u}) + \boldsymbol{\lambda}^T(t) \mathbf{f}(t, \mathbf{x}, \mathbf{u}) \quad (4.93)$$

auch in der Form

$$\dot{\mathbf{x}}^* = \left( \frac{\partial}{\partial \boldsymbol{\lambda}} H \right)^T (t, \mathbf{x}^*(t), \mathbf{u}^*(t), \boldsymbol{\lambda}^*(t)), \quad \mathbf{x}^*(t_0) = \mathbf{x}_0 \quad (4.94a)$$

$$\dot{\boldsymbol{\lambda}}^* = - \left( \frac{\partial}{\partial \mathbf{x}} H \right)^T (t, \mathbf{x}^*(t), \mathbf{u}^*(t), \boldsymbol{\lambda}^*(t)), \quad \boldsymbol{\lambda}^*(t_1) = \left( \frac{\partial}{\partial \mathbf{x}_1} \varphi \right)^T (\mathbf{x}^*(t_1)) \quad (4.94b)$$

$$\mathbf{0} = \left( \frac{\partial}{\partial \mathbf{u}} H \right)^T (t, \mathbf{x}^*(t), \mathbf{u}^*(t), \boldsymbol{\lambda}^*(t)) \quad (4.94c)$$

für  $t_0 \leq t \leq t_1$  anschreiben. Man beachte, dass sich die hier in der dynamischen Optimierung verwendete Hamiltonfunktion  $H$  im Vorzeichen von jener der Variationsrechnung (siehe (4.22)) unterscheidet. Die Bedingung (4.94c) zeigt, dass  $\mathbf{u}^*$  ein *stationärer Punkt* der Hamiltonfunktion  $H$  sein muss. Die Ableitung der Hamiltonfunktion entlang der optimalen Lösung  $(\mathbf{x}^*(t), \mathbf{u}^*(t), \boldsymbol{\lambda}^*(t))$  nach der Zeit liefert

$$\begin{aligned} \frac{d}{dt} H &= \frac{\partial}{\partial t} H + \left( \frac{\partial}{\partial \mathbf{x}} H \right) \dot{\mathbf{x}}^* + \underbrace{\left( \frac{\partial}{\partial \mathbf{u}} H \right) \dot{\mathbf{u}}^*}_{=0} + \left( \frac{\partial}{\partial \boldsymbol{\lambda}} H \right) \dot{\boldsymbol{\lambda}}^* \\ &= \frac{\partial}{\partial t} H - (\dot{\boldsymbol{\lambda}}^*)^T \mathbf{f} + (\dot{\mathbf{x}}^*)^T \dot{\boldsymbol{\lambda}}^* = \frac{\partial}{\partial t} H . \end{aligned} \quad (4.95)$$

Wenn daher weder  $\mathbf{f}$  noch  $l$  explizit von der Zeit  $t$  abhängen, ist die Hamiltonfunktion  $H$  eine *Invariante* des Zweipunkttrandwertproblems (4.94).

Im Weiteren muss ähnlich zur Legendre Bedingung gemäß Satz 4.3 für ein Minimum des Kostenfunktional  $J(\mathbf{u})$  die *notwendige Bedingung zweiter Ordnung*

$$\mathbf{d}^T \frac{\partial^2}{\partial \mathbf{u}^2} H(t, \mathbf{x}^*(t), \mathbf{u}^*(t), \boldsymbol{\lambda}^*(t)) \mathbf{d} \geq 0 \quad \forall \mathbf{d} \in \mathbb{R}^m, t \in [t_0, t_1] . \quad (4.96)$$

erfüllt sein. Sie wird auch *Legendre-Clebsch Bedingung* genannt.

In Satz 4.8 wurde angenommen, dass die optimale Stellgröße  $\mathbf{u}^*$  stetig ist, d. h.  $\mathbf{u}^*(t) \in (C[t_0, t_1])^m$ . Für manche Beispiele findet man keine Lösung der Euler-Lagrange Gleichungen (4.82) in der Klasse der stetigen Stellgrößen. Aus diesem Grund sucht man Extremale in der erweiterten Klasse der stückweise stetigen Stellgrößen  $(\hat{C}[t_0, t_1])^m$ . Wie bereits im Abschnitt 4.2.2 diskutiert, sind für stückweise stetige Stellgrößen  $\mathbf{u}(t)$  die zugehörigen Zustandsgrößen  $\mathbf{x}(t)$  von (4.81) stückweise stetig differenzierbar, d. h.  $\mathbf{x}(t) \in (\hat{C}^1[t_0, t_1])^n$ , wobei die Eckpunkte mit den Unstetigkeitsstellen der Stellgrößen übereinstimmen. Bezeichnet man mit  $\hat{\mathbf{u}}^*(t) \in (\hat{C}[t_0, t_1])^m$  die optimale Stellgröße und mit  $\hat{\mathbf{x}}^*(t)$  und  $\hat{\boldsymbol{\lambda}}^*(t)$  den zugehörigen Zustand und den adjungierten Zustand des Optimalsteuerungsproblems (4.80), (4.81), dann gelten für jeden Eckpunkt  $c \in (t_0, t_1)$  die Bedingungen

$$\hat{\mathbf{x}}^*(c^-) = \hat{\mathbf{x}}^*(c^+) \quad (4.97a)$$

$$\hat{\boldsymbol{\lambda}}^*(c^-) = \hat{\boldsymbol{\lambda}}^*(c^+) \quad (4.97b)$$

$$H(c^-, \hat{\mathbf{x}}^*(c), \hat{\mathbf{u}}^*(c^-), \hat{\boldsymbol{\lambda}}^*(c)) = H(c^+, \hat{\mathbf{x}}^*(c), \hat{\mathbf{u}}^*(c^+), \hat{\boldsymbol{\lambda}}^*(c)) , \quad (4.97c)$$

wobei die Argumente  $c^-$  bzw.  $c^+$  jeweils den links- bzw. rechtsseitigen Grenzwert liefern sollen. Man beachte, dass (4.97b) und (4.97c) den Weierstrass-Erdmann Bedingungen gemäß Satz 4.7 entsprechen.

Im Folgenden werden die notwendigen Bedingungen erster Ordnung für ein Optimalsteuerungsproblem mit freier Endzeit und allgemeinen Endbeschränkungen formuliert.

**Satz 4.9 (Optimalsteuerungsproblem mit freier Endzeit und Endbeschränkungen).** Gesucht ist die Stellgröße  $\mathbf{u} \in (C[t_0, t_1])^m$  so, dass das Kostenfunktional (Bolza-Form)

$$J(\mathbf{u}, t_1) = \varphi(t_1, \mathbf{x}(t_1)) + \int_{t_0}^{t_1} l(t, \mathbf{x}(t), \mathbf{u}(t)) dt \quad (4.98)$$

unter den Gleichungsbeschränkungen

$$\dot{\mathbf{x}} - \mathbf{f}(t, \mathbf{x}(t), \mathbf{u}(t)) = \mathbf{0}, \quad \mathbf{x}(t_0) = \mathbf{x}_0 \quad (4.99a)$$

$$G_k(\mathbf{u}, t_1) = \psi_k(t_1, \mathbf{x}(t_1)) = 0, \quad k = 1, \dots, p \quad (4.99b)$$

mit fester Anfangszeit  $t_0$  und freier Endzeit  $t_1 \ll T$  minimiert wird. Dabei wird angenommen, dass  $l$  und  $\mathbf{f}$  stetig in  $t$ ,  $\mathbf{x}$  und  $\mathbf{u}$  und stetig differenzierbar bezüglich  $\mathbf{x}$  und  $\mathbf{u}$  für alle  $(t, \mathbf{x}, \mathbf{u}) \in [t_0, T] \times \mathbb{R}^n \times \mathbb{R}^m$  sind und die Funktionen  $\varphi$  und  $\psi_k$ ,  $k = 1, \dots, p$  stetig und stetig differenzierbar bezüglich  $t_1$  und  $\mathbf{x}_1$  für alle  $(t_1, \mathbf{x}_1) \in [t_0, T] \times \mathbb{R}^n$  sind. Weiters sei  $(\mathbf{u}^*, t_1^*) \in (C[t_0, T])^m \times [t_0, T]$  die optimale Lösung des Optimierungsproblems und  $\mathbf{x}^* \in (C^1[t_0, T])^n$  die zugehörige Lösung des Anfangswertproblems (4.99a). Darüber hinaus wird angenommen, dass für  $p$  unabhängige zulässige Richtungen  $(\xi_{x,k}, \xi_{t_1,k}) \in (C[t_0, T])^m \times [t_0, T]$ ,  $k = 1, \dots, p$  die Regularitätsbedingung

$$\det \left( \begin{bmatrix} \delta G_1(\mathbf{u}^*, t_1^*; \xi_{x,1}, \xi_{t_1,1}) & \cdots & \delta G_1(\mathbf{u}^*, t_1^*; \xi_{x,p}, \xi_{t_1,p}) \\ \vdots & \ddots & \vdots \\ \delta G_p(\mathbf{u}^*, t_1^*; \xi_{x,1}, \xi_{t_1,1}) & \cdots & \delta G_p(\mathbf{u}^*, t_1^*; \xi_{x,p}, \xi_{t_1,p}) \end{bmatrix} \right) \neq 0 \quad (4.100)$$

gilt. Dann existieren ein  $\lambda^* \in (C^1[t_0, t_1^*])^n$  und ein  $\mu^* \in \mathbb{R}^p$  so, dass die Beziehungen

$$\dot{\mathbf{x}}^* = \left( \frac{\partial}{\partial \lambda} H \right)^\top (t, \mathbf{x}^*(t), \mathbf{u}^*(t), \lambda^*(t)), \quad \mathbf{x}^*(t_0) = \mathbf{x}_0 \quad (4.101a)$$

$$\dot{\lambda}^* = - \left( \frac{\partial}{\partial \mathbf{x}} H \right)^\top (t, \mathbf{x}^*(t), \mathbf{u}^*(t), \lambda^*(t)) \quad (4.101b)$$

$$\lambda^*(t_1^*) = \left( \frac{\partial}{\partial \mathbf{x}_1} \Phi \right)^\top (t_1^*, \mathbf{x}^*(t_1^*), \mu^*)$$

$$\mathbf{0} = \left( \frac{\partial}{\partial \mathbf{u}} H \right)^\top (t, \mathbf{x}^*(t), \mathbf{u}^*(t), \lambda^*(t)) \quad (4.101c)$$

für  $t_0 \leq t \leq t_1$  mit den Transversalitätsbedingungen

$$\psi(t_1^*, \mathbf{x}^*(t_1^*)) = \mathbf{0} \quad (4.102a)$$

$$\left(\frac{\partial}{\partial t_1}\Phi\right)(t_1^*, \mathbf{x}^*(t_1^*), \boldsymbol{\mu}^*) + H(t_1^*, \mathbf{x}^*(t_1^*), \mathbf{u}^*(t_1^*), \boldsymbol{\lambda}^*(t_1^*)) = 0, \quad (4.102b)$$

und der Hamiltonfunktion  $H = l + \boldsymbol{\lambda}^T \mathbf{f}$  sowie  $\Phi = \varphi + \boldsymbol{\mu}^T \boldsymbol{\psi}$  mit  $\boldsymbol{\psi}^T = [\psi_1 \ \psi_2 \ \dots \ \psi_p]$  erfüllt sind.

*Beweisskizze:* Für das Kostenfunktional (4.98) mit den Gleichungsbeschränkungen (4.99) wird das *Lagrangefunktional*

$$\begin{aligned} L(\mathbf{x}, \mathbf{u}, t_1, \boldsymbol{\lambda}, \boldsymbol{\mu}) &= J(\mathbf{u}, t_1) + \boldsymbol{\mu}^T \boldsymbol{\psi}(t_1, \mathbf{x}(t_1)) + \int_{t_0}^{t_1} \boldsymbol{\lambda}^T(t) (\mathbf{f}(t, \mathbf{x}(t), \mathbf{u}(t)) - \dot{\mathbf{x}}(t)) dt \\ &= \varphi(t_1, \mathbf{x}(t_1)) + \boldsymbol{\mu}^T \boldsymbol{\psi}(t_1, \mathbf{x}(t_1)) \\ &\quad + \int_{t_0}^{t_1} l(t, \mathbf{x}(t), \mathbf{u}(t)) + \boldsymbol{\lambda}^T(t) (\mathbf{f}(t, \mathbf{x}(t), \mathbf{u}(t)) - \dot{\mathbf{x}}(t)) dt \end{aligned} \quad (4.103)$$

formuliert. Die Gâteaux Ableitung  $\delta J(\mathbf{u}, t_1; \boldsymbol{\xi}_u, \xi_{t_1})$  an einem allgemeinen Punkt  $(\mathbf{u}, t_1) \in (C[t_0, T])^m \times [t_0, T]$  bezüglich  $(\boldsymbol{\xi}_u(t), \xi_{t_1}) \in (C[t_0, T])^m \times [t_0, T]$  lautet

$$\delta J(\mathbf{u}, t_1; \boldsymbol{\xi}_u, \xi_{t_1}) = \delta_{\mathbf{u}} L(\mathbf{x}, \mathbf{u}, t_1, \boldsymbol{\lambda}, \boldsymbol{\mu}; \boldsymbol{\xi}_u) + \delta_{t_1} L(\mathbf{x}, \mathbf{u}, t_1, \boldsymbol{\lambda}, \boldsymbol{\mu}; \xi_{t_1}) \quad (4.104a)$$

mit

$$\begin{aligned} \delta_{\mathbf{u}} L(\mathbf{x}, \mathbf{u}, t_1, \boldsymbol{\lambda}, \boldsymbol{\mu}; \boldsymbol{\xi}_u) &= \int_{t_0}^{t_1} \left[ \frac{\partial}{\partial \mathbf{u}} l(t, \mathbf{x}(t), \mathbf{u}(t)) + \boldsymbol{\lambda}^T(t) \frac{\partial}{\partial \mathbf{u}} \mathbf{f}(t, \mathbf{x}(t), \mathbf{u}(t)) \right] \boldsymbol{\xi}_u(t) dt \end{aligned} \quad (4.104b)$$

$$\begin{aligned} \delta_{t_1} L(\mathbf{x}, \mathbf{u}, t_1, \boldsymbol{\lambda}, \boldsymbol{\mu}; \xi_{t_1}) &= \left[ \frac{\partial}{\partial t_1} \varphi(t_1, \mathbf{x}(t_1)) + \frac{\partial}{\partial \mathbf{x}_1} \varphi(t_1, \mathbf{x}(t_1)) \dot{\mathbf{x}}(t_1) \right. \\ &\quad \left. + \boldsymbol{\mu}^T \left( \frac{\partial}{\partial t_1} \boldsymbol{\psi}(t_1, \mathbf{x}(t_1)) + \frac{\partial}{\partial \mathbf{x}_1} \boldsymbol{\psi}(t_1, \mathbf{x}(t_1)) \dot{\mathbf{x}}(t_1) \right) \right. \\ &\quad \left. + l(t_1, \mathbf{x}(t_1), \mathbf{u}(t_1)) + \boldsymbol{\lambda}^T(t_1) (\mathbf{f}(t_1, \mathbf{x}(t_1), \mathbf{u}(t_1)) - \dot{\mathbf{x}}(t_1)) \right] \xi_{t_1}, \end{aligned} \quad (4.104c)$$

wobei die Bedingungen

$$\begin{aligned}
\mathbf{0} &= \delta_{\mathbf{x}} L(\mathbf{x}, \mathbf{u}, t_1, \boldsymbol{\lambda}, \boldsymbol{\mu}; \boldsymbol{\xi}_x) \\
&= \frac{\partial}{\partial \mathbf{x}_1} \varphi(t_1, \mathbf{x}(t_1)) \boldsymbol{\xi}_x(t_1) + \boldsymbol{\mu}^T \frac{\partial}{\partial \mathbf{x}_1} \psi(t_1, \mathbf{x}(t_1)) \boldsymbol{\xi}_x(t_1) \\
&\quad + \int_{t_0}^{t_1} \frac{\partial}{\partial \mathbf{x}} l(t, \mathbf{x}(t), \mathbf{u}(t)) \boldsymbol{\xi}_x(t) + \boldsymbol{\lambda}^T(t) \left( \frac{\partial}{\partial \mathbf{x}} \mathbf{f}(t, \mathbf{x}(t), \mathbf{u}(t)) \boldsymbol{\xi}_x(t) - \dot{\boldsymbol{\xi}}_x(t) \right) dt \\
&= \left[ \frac{\partial}{\partial \mathbf{x}_1} \varphi(\mathbf{x}(t_1)) + \boldsymbol{\mu}^T \frac{\partial}{\partial \mathbf{x}_1} \psi(t_1, \mathbf{x}(t_1)) - \boldsymbol{\lambda}^T(t_1) \right] \boldsymbol{\xi}_x(t_1) + \boldsymbol{\lambda}^T(t_0) \boldsymbol{\xi}_x(t_0) \\
&\quad + \int_{t_0}^{t_1} \left[ \frac{\partial}{\partial \mathbf{x}} l(t, \mathbf{x}(t), \mathbf{u}(t)) + \boldsymbol{\lambda}^T(t) \frac{\partial}{\partial \mathbf{x}} \mathbf{f}(t, \mathbf{x}(t), \mathbf{u}(t)) + \dot{\boldsymbol{\lambda}}^T(t) \right] \boldsymbol{\xi}_x(t) dt
\end{aligned} \tag{4.104d}$$

$$\mathbf{0} = \delta_{\lambda} L(\mathbf{x}, \mathbf{u}, \boldsymbol{\lambda}, \boldsymbol{\mu}; \boldsymbol{\xi}_{\lambda}) = \int_{t_0}^{t_1} \boldsymbol{\xi}_{\lambda}^T(t) [\mathbf{f}(t, \mathbf{x}(t), \mathbf{u}(t)) - \dot{\mathbf{x}}(t)] dt \tag{4.104e}$$

$$\mathbf{0} = \left( \frac{\partial}{\partial \boldsymbol{\mu}} L \right)^T (\mathbf{x}, \mathbf{u}, t_1, \boldsymbol{\lambda}, \boldsymbol{\mu}) = \psi(t_1, \mathbf{x}(t_1)) \tag{4.104f}$$

für alle zulässigen Richtungen  $\boldsymbol{\xi}_x(t) \in (C[t_0, T])^n$  mit  $\boldsymbol{\xi}_x(t_0) = \mathbf{0}$  zufolge der Anfangsbedingung  $\mathbf{x}(t_0) = \mathbf{x}_0$  und für alle zulässigen Richtungen  $\boldsymbol{\xi}_{\lambda}(t) \in (C[t_0, T])^n$  einzuhalten sind. Gemäß Fundamentallema der Variationsrechnung Lemma 4.2 folgt aus (4.104d)

$$\dot{\boldsymbol{\lambda}} = - \left( \frac{\partial}{\partial \mathbf{x}} l \right)^T (t, \mathbf{x}(t), \mathbf{u}(t)) - \left( \frac{\partial}{\partial \mathbf{x}} \mathbf{f} \right)^T (t, \mathbf{x}(t), \mathbf{u}(t)) \boldsymbol{\lambda}(t) \tag{4.105a}$$

$$\boldsymbol{\lambda}(t_1) = \left( \frac{d}{d\mathbf{x}_1} \varphi \right)^T (\mathbf{x}(t_1)) + \boldsymbol{\mu}^T \frac{\partial}{\partial \mathbf{x}_1} \psi(t_1, \mathbf{x}(t_1)) \tag{4.105b}$$

und aus (4.104e) die Differenzialgleichung (4.99a). Wegen (4.105b) vereinfacht sich (4.104c) zu

$$\begin{aligned}
&\delta_{t_1} L(\mathbf{x}, \mathbf{u}, t_1, \boldsymbol{\lambda}, \boldsymbol{\mu}; \xi_{t_1}) \\
&= \left[ \frac{\partial}{\partial t_1} \varphi(t_1, \mathbf{x}(t_1)) + \boldsymbol{\mu}^T \frac{\partial}{\partial t_1} \psi(t_1, \mathbf{x}(t_1)) \right. \\
&\quad \left. + l(t_1, \mathbf{x}(t_1), \mathbf{u}(t_1)) + \boldsymbol{\lambda}^T(t_1) \mathbf{f}(t_1, \mathbf{x}(t_1), \mathbf{u}(t_1)) \right] \xi_{t_1} .
\end{aligned} \tag{4.106}$$

Da am optimalen Punkt  $(\mathbf{u}^*, t_1^*) \in (C[t_0, T])^m \times [t_0, T]$  gemäß Satz 4.1 die Bedingung  $\delta J(\mathbf{u}, t_1; \boldsymbol{\xi}_u, \xi_{t_1}) = 0$  für alle zulässigen Richtungen  $(\boldsymbol{\xi}_u, \xi_{t_1}) \in (C[t_0, T])^m \times [t_0, T]$  erfüllt sein muss, folgen aus (4.104a), (4.104b), (4.104f), (4.105), (4.106) und (4.99a) unter Verwendung von Lemma 4.2 genau die Optimalitätsbedingungen (4.101) und (4.102) von Satz 4.9.

Für den Beweis der Notwendigkeit der Regularitätsbedingung (4.100) wird auf [4.1] verwiesen. Diese Regularitätsbedingung sichert die Existenz einer Lösung  $\mathbf{u}(t)$ , so

dass der zugehörige Endzustand  $\mathbf{x}(t_1)$  die Gleichungsbeschränkung (4.99b) erfüllt.  $\square$

**Aufgabe 4.3.** Beweisen Sie Satz 4.9 ohne Verwendung des Lagrangefunktionals (4.103).

**Hinweis:** Orientieren Sie sich dabei am Beweis von Satz 4.8.

Zur Berechnung der  $m + 2n + p + 1$  unbekanntenen Größen  $(\mathbf{u}^*(t), \mathbf{x}^*(t), \boldsymbol{\lambda}^*(t), \boldsymbol{\mu}^*, t_1^*)$  stehen mit Satz 4.9  $m + 2n + p + 1$  Bedingungen zur Verfügung. Das sind  $m$  algebraische Gleichungen (4.101c),  $p + 1$  algebraische Gleichungen in Form der Transversalitätsbedingungen (4.102) und  $2n$  Differentialgleichungen (4.101a) und (4.101b) für  $\mathbf{x}^*$  und  $\boldsymbol{\lambda}^*$ . Zu diesen Differentialgleichungen gehören die in (4.101a) und (4.101b) angegebenen  $2n$  Randbedingungen. Aus den genannten Gleichungen lassen sich eindeutig die unbekanntenen Größen  $(\mathbf{u}^*(t), \mathbf{x}^*(t), \boldsymbol{\lambda}^*(t), \boldsymbol{\mu}^*, t_1^*)$  bestimmen.

Für eine Zusammenfassung der Ergebnisse von Satz 4.9 werden in weiterer Folge nur sogenannte partielle Endbedingungen der Form

$$\psi_j = x_k(t_1) - \bar{x}_k, \quad j = 1, \dots, p \quad (4.107)$$

mit  $\bar{x}_k = \text{konst.}$  als fixem Endwert der Komponente  $x_k(t)$  von  $\mathbf{x}(t)$  betrachtet. Für diesen Spezialfall kann die Endbedingung für  $\boldsymbol{\lambda}^*(t_1^*)$  in (4.101b) ersetzt werden und unter Berücksichtigung von (4.104d) und (4.106) gilt Folgendes:

- (a) Wenn die *Endzeit*  $t_1$  fest ist und damit  $\xi_{t_1} = 0$  gilt, wird (4.106) automatisch erfüllt. Es liegt somit keine Transversalitätsbedingung gemäß (4.102b) vor.
- (i) Wenn für eine Komponente  $x_k(t)$  von  $\mathbf{x}(t)$  gilt, dass deren *Endwert fest ist*, so folgt daraus  $\xi_{x,k}(t_1) = 0$ , womit automatisch der zugehörige Summand in der vorletzten Zeile von (4.104d) verschwindet. Damit liegt für diese Komponente keine Endbedingung für den zugehörigen adjungierten Zustand  $\lambda_k^*(t_1)$  in (4.105b) vor.
- (ii) Wenn für eine Komponente  $x_k(t)$  von  $\mathbf{x}(t)$  gilt, dass deren *Endwert frei ist*, so folgt daraus die Zulässigkeit von  $\xi_{x,k}(t_1) \neq 0$ . Aus dem zugehörigen Summanden in der vorletzten Zeile von (4.104d) ergibt sich daher, dass die Komponente des zugehörigen adjungierten Zustands  $\lambda_k^*(t_1)$  die Endbedingung

$$\lambda_k^*(t_1) = \frac{\partial}{\partial x_{1,k}} \varphi(t_1, \mathbf{x}^*(t_1)) \quad (4.108)$$

erfüllen muss.

- (b) Wenn die *Endzeit frei ist*, muss die Transversalitätsbedingung

$$\frac{\partial}{\partial t_1} \Phi(t_1^*, \mathbf{x}^*(t_1^*), \boldsymbol{\mu}^*) + H(t_1^*, \mathbf{x}^*(t_1^*), \mathbf{u}^*(t_1^*), \boldsymbol{\lambda}^*(t_1^*)) = 0, \quad H = l + \boldsymbol{\lambda}^T \mathbf{f} \quad (4.109)$$

gelten. In Abhängigkeit davon, ob der Endwert einer Komponente  $x_k(t)$  von  $\mathbf{x}(t)$  fest oder frei ist, können die Unterpunkte (i) und (ii) vom Fall (a) auch direkt hier angewandt werden.

Wenn in Satz 4.9 die Gleichungsbeschränkungen (4.99b) durch *Ungleichungsbeschränkungen* der Form

$$\psi_k(t_1, \mathbf{x}(t_1)) \leq 0, \quad k = 1, \dots, p \quad (4.110)$$

ersetzt werden, so ändert sich lediglich (4.102a) zu

$$\psi_k(t_1^*, \mathbf{x}^*(t_1^*)) \leq 0 \quad (4.111a)$$

$$\boldsymbol{\mu}^* \geq \mathbf{0} \quad (4.111b)$$

$$(\boldsymbol{\mu}^*)^\top \boldsymbol{\psi}(t_1^*, \mathbf{x}^*(t_1^*)) = 0, \quad (4.111c)$$

wobei (4.111c) auch als *complementary slackness condition* bezeichnet wird.

**Aufgabe 4.4.** Gesucht ist eine Lösung des Optimierungsproblems

$$\min_{u(\cdot)} \int_0^1 \frac{1}{2} u^2 + \frac{a}{2} x^2 dt, \quad a > 0 \quad (4.112a)$$

$$\text{u.B.v. } \dot{x} = u, \quad x(0) = 1, \quad x(1) = 0. \quad (4.112b)$$

Zeigen Sie, dass die Lösung durch

$$x^*(t) = \frac{1}{1 - e^{2\sqrt{a}}} (e^{\sqrt{a}t} - e^{\sqrt{a}(2-t)}), \quad u^*(t) = \frac{\sqrt{a}}{1 - e^{2\sqrt{a}}} (e^{\sqrt{a}t} + e^{\sqrt{a}(2-t)}) \quad (4.113)$$

gegeben ist und interpretieren Sie die Ergebnisse, die in Abbildung 4.4 für verschiedene Parameterwerte  $a$  dargestellt sind.

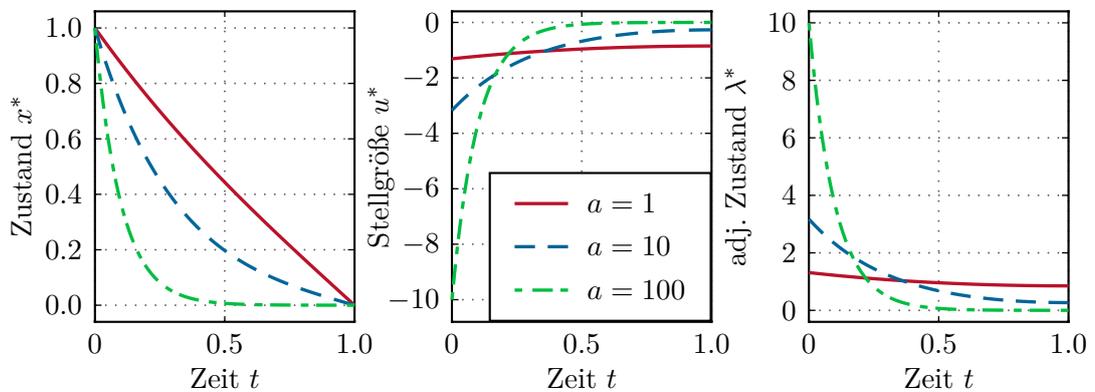


Abbildung 4.4: Optimale Trajektorien in Aufgabe 4.4.

**Aufgabe 4.5.** Gesucht ist eine Lösung des Optimierungsproblems

$$\min_{u(\cdot)} \frac{a}{2} x_2^2(1) + \int_0^1 \frac{1}{2} u^2 dt, \quad a \geq 0 \quad (4.114a)$$

$$\text{u.B.v. } \dot{x}_1 = x_2, \quad x_1(0) = 1, \quad x_1(1) = 0 \quad (4.114b)$$

$$\dot{x}_2 = u, \quad x_2(0) = 0. \quad (4.114c)$$

Zeigen Sie, dass sich für den (freien) Endzustand  $x_2^*(1) = -6/(4+a)$  in Abhängigkeit des Parameters  $a \geq 0$  ergibt und dass die optimale Lösung durch

$$x_1^*(t) = \frac{2(1+a)}{4+a} t^3 - \frac{3(2+a)}{a+4} t^2 + 1, \quad u^*(t) = \frac{12(1+a)}{4+a} t - \frac{6(2+a)}{a+4} \quad (4.115)$$

gegeben ist. Interpretieren Sie die Ergebnisse in Abbildung 4.5.

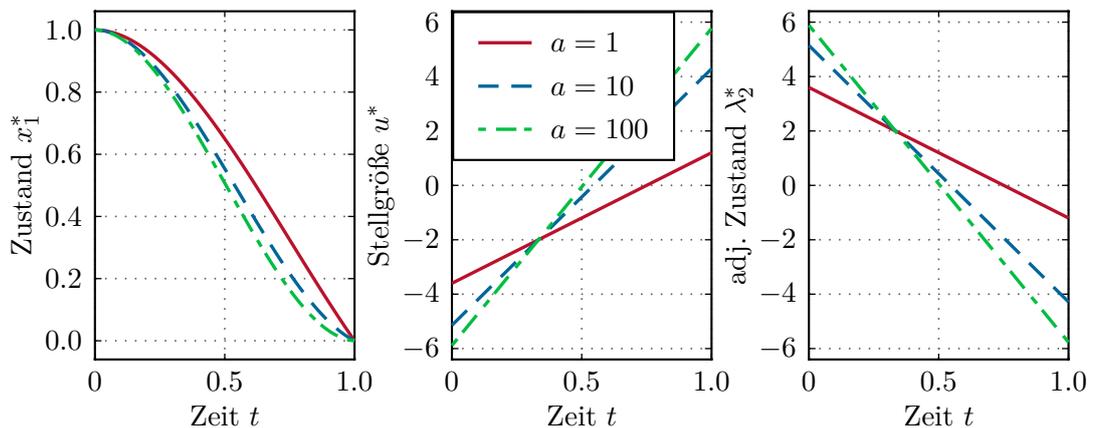


Abbildung 4.5: Optimale Trajektorien in Aufgabe 4.5.

**Beispiel 4.6.** Betrachtet wird eine Punktmasse mit der Masse  $m$  in der  $(x, y)$ -Ebene, auf die eine konstante Kraft  $F = ma$  wirkt. Die Stellgröße  $u$  des Problems ist der Winkel zwischen der Schubrichtung und der  $x$ -Achse, siehe Abbildung 4.6. Ziel ist es, die Punktmasse in *minimaler Zeit*  $[t_0 = 0, t_1^*]$  zu einem *fest vorgegebenen Zielpunkt*  $(\bar{x}_1, \bar{y}_1)$  zu steuern. Unter der Annahme, dass außer dem Schub keine weiteren Kräfte auftreten, kann das Optimalsteuerungsproblem wie folgt formuliert werden

$$\min_{u(\cdot)} t_1 \quad (4.116a)$$

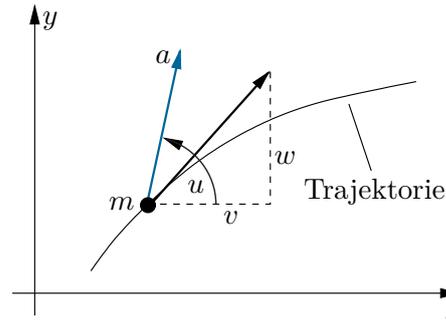
$$\text{u.B.v. } \dot{x} = v, \quad x(0) = x_0, \quad x(t_1) = \bar{x}_1 \quad (4.116b)$$

$$\dot{v} = a \cos(u), \quad v(0) = v_0 \quad (4.116c)$$

$$\dot{y} = w, \quad y(0) = y_0, \quad y(t_1) = \bar{y}_1 \quad (4.116d)$$

$$\dot{w} = a \sin(u), \quad w(0) = w_0. \quad (4.116e)$$

Man beachte, dass der Endzustand nur für die Position  $(x, y)$  aber nicht für die Geschwindigkeiten  $(v, w)$  vorgegeben ist.

Abbildung 4.6: Bewegung einer Punktmasse der Masse  $m$  in der  $(x, y)$ -Ebene.

Die beiden fest vorgegebenen Endwerte für  $x$  und  $y$  können als Gleichungsbeschränkungen gemäß (4.99b) in der Form

$$\psi_1(t_1, \mathbf{x}(t_1)) = x(t_1) - \bar{x}_1, \quad \psi_2(t_1, \mathbf{x}(t_1)) = y(t_1) - \bar{y}_1 \quad (4.117)$$

formuliert werden. Die Hamiltonfunktion  $H$  und die Funktion  $\Phi$  gemäß Satz 4.9 lauten dann für das vorliegende Optimierungsproblem

$$H(\mathbf{x}, u, \boldsymbol{\lambda}) = \lambda_x v + \lambda_v a \cos(u) + \lambda_y w + \lambda_w a \sin(u) \quad (4.118a)$$

$$\Phi(t_1, \mathbf{x}(t_1), \boldsymbol{\mu}) = \varphi + \mu_x \psi_1 + \mu_y \psi_2 = t_1 + \mu_x (x(t_1) - \bar{x}_1) + \mu_y (y(t_1) - \bar{y}_1) \quad (4.118b)$$

mit  $\mathbf{x} = [x \ v \ y \ w]^T$ , dem adjungierten Zustand  $\boldsymbol{\lambda} = [\lambda_x \ \lambda_v \ \lambda_y \ \lambda_w]^T$  und dem konstanten Lagrange-Multiplikator  $\boldsymbol{\mu} = [\mu_x \ \mu_y]^T$ . Die Randbedingungen für den adjungierten Zustand errechnen sich gemäß (4.101b) zu

$$\lambda_x^*(t_1^*) = \left( \frac{\partial}{\partial x_1} \Phi \right) (t_1^*, \mathbf{x}^*(t_1^*), \boldsymbol{\mu}^*) = \mu_x^*, \quad \lambda_v^*(t_1^*) = \left( \frac{\partial}{\partial v_1} \Phi \right) (t_1^*, \mathbf{x}^*(t_1^*), \boldsymbol{\mu}^*) = 0 \quad (4.119a)$$

$$\lambda_y^*(t_1^*) = \left( \frac{\partial}{\partial y_1} \Phi \right) (t_1^*, \mathbf{x}^*(t_1^*), \boldsymbol{\mu}^*) = \mu_y^*, \quad \lambda_w^*(t_1^*) = \left( \frac{\partial}{\partial w_1} \Phi \right) (t_1^*, \mathbf{x}^*(t_1^*), \boldsymbol{\mu}^*) = 0. \quad (4.119b)$$

Damit lautet das adjungierte System

$$\dot{\lambda}_x^* = - \left( \frac{\partial H}{\partial x} \right) (\mathbf{x}^*, \mathbf{u}^*, \boldsymbol{\lambda}^*) = 0, \quad \lambda_x^*(t_1^*) = \mu_x^* \quad (4.120a)$$

$$\dot{\lambda}_v^* = - \left( \frac{\partial H}{\partial v} \right) (\mathbf{x}^*, \mathbf{u}^*, \boldsymbol{\lambda}^*) = -\lambda_x^*, \quad \lambda_v^*(t_1^*) = 0 \quad (4.120b)$$

$$\dot{\lambda}_y^* = - \left( \frac{\partial H}{\partial y} \right) (\mathbf{x}^*, \mathbf{u}^*, \boldsymbol{\lambda}^*) = 0, \quad \lambda_y^*(t_1^*) = \mu_y^* \quad (4.120c)$$

$$\dot{\lambda}_w^* = - \left( \frac{\partial H}{\partial w} \right) (\mathbf{x}^*, \mathbf{u}^*, \boldsymbol{\lambda}^*) = -\lambda_y^*, \quad \lambda_w^*(t_1^*) = 0, \quad (4.120d)$$

woraus direkt

$$\lambda_x^* = \mu_x^*, \quad \lambda_v^* = \mu_x^*(t_1^* - t), \quad \lambda_y^* = \mu_y^*, \quad \lambda_w^* = \mu_y^*(t_1^* - t) \quad (4.121)$$

folgt. Des Weiteren muss die Hamiltonfunktion  $H$  gemäß (4.101) extremal sein, weshalb die Bedingung

$$\left(\frac{\partial H}{\partial u}\right)(\mathbf{x}^*, \mathbf{u}^*, \boldsymbol{\lambda}^*) = -\lambda_v^* a \sin(u^*) + \lambda_w^* a \cos(u^*) = 0, \quad (4.122)$$

erfüllt sein muss und  $u^*$  sich in der Form

$$\begin{aligned} \tan(u^*) &= \frac{\lambda_w^*}{\lambda_v^*} \stackrel{(4.121)}{=} \frac{\mu_y^*(t_1^* - t)}{\mu_x^*(t_1^* - t)} = \frac{\mu_y^*}{\mu_x^*} = \text{konst.} \\ \Rightarrow u^* &= \arctan\left(\frac{\mu_y^*}{\mu_x^*}\right) \quad \text{mit} \quad -\frac{\pi}{2} < u^* < \frac{\pi}{2} \end{aligned} \quad (4.123)$$

berechnen lässt. Die optimale Steuerung  $u^*$  ist also auf dem gesamten Zeitintervall  $[t_0, t_1^*]$  konstant und die zugehörigen optimalen Zustandstrajektorien  $\mathbf{x}^*(t)$  können durch Lösen der Differentialgleichungen (4.116) und Einsetzen der Anfangsbedingungen in der Form

$$x^*(t) = g_x(\mu_x^*, \mu_y^*, t) = x_0 + v_0 t + \frac{1}{2} a \cos(u^*) t^2, \quad \cos(u^*) = \frac{1}{\sqrt{1 + (\mu_y^*/\mu_x^*)^2}} \quad (4.124a)$$

$$v^*(t) = g_v(\mu_x^*, \mu_y^*, t) = v_0 + a \cos(u^*) t \quad (4.124b)$$

$$y^*(t) = g_y(\mu_x^*, \mu_y^*, t) = y_0 + w_0 t + \frac{1}{2} a \sin(u^*) t^2, \quad \sin(u^*) = \frac{\mu_y^*}{\mu_x^* \sqrt{1 + (\mu_y^*/\mu_x^*)^2}} \quad (4.124c)$$

$$w^*(t) = g_w(\mu_x^*, \mu_y^*, t) = w_0 + a \sin(u^*) t. \quad (4.124d)$$

bestimmt werden. Dabei wurden die trigonometrischen Beziehungen

$$\sin(\arctan(b)) = \frac{b}{\sqrt{1 + b^2}}, \quad \cos(\arctan(b)) = \frac{1}{\sqrt{1 + b^2}} \quad (4.125)$$

verwendet. Da die Endzeit  $t_1$  frei ist, muss zusätzlich die Transversalitätsbedingung (4.102b) gelten, wobei sich die Hamiltonfunktion (4.118a) aufgrund der Endbedingungen  $\lambda_v^*(t_1^*) = \lambda_w^*(t_1^*) = 0$  entsprechend vereinfacht

$$0 = \left(\frac{\partial}{\partial t_1} \Phi\right)(t_1^*, \mathbf{x}^*(t_1^*), \boldsymbol{\mu}^*) + H(\mathbf{x}^*(t_1^*), \mathbf{u}^*(t_1^*), \boldsymbol{\lambda}^*(t_1^*)) \quad (4.126a)$$

$$= 1 + \mu_x^* g_v(\mu_x^*, \mu_y^*, t_1^*) + \mu_y^* g_w(\mu_x^*, \mu_y^*, t_1^*). \quad (4.126b)$$

Mit Hilfe der zwei Gleichungsbeschränkungen (4.117) und der Transversalitätsbedingung (4.126) lässt sich ein Gleichungssystem für die verbleibenden drei Unbekannten

$\mu_x^*$ ,  $\mu_y^*$  und  $t_1^*$  in der Form

$$\begin{bmatrix} g_x(\mu_x^*, \mu_y^*, t_1^*) \\ g_y(\mu_x^*, \mu_y^*, t_1^*) \\ \mu_x^* g_v(\mu_x^*, \mu_y^*, t_1^*) + \mu_y^* g_w(\mu_x^*, \mu_y^*, t_1^*) \end{bmatrix} - \begin{bmatrix} \bar{x}_1 \\ \bar{y}_1 \\ -1 \end{bmatrix} = \mathbf{0} \quad (4.127)$$

formulieren, welches auf *numerischem Wege* gelöst werden kann. Eine geeignete MATLAB-Funktion zur Lösung von nichtlinearen Gleichungen ist mit dem Befehl *fsolve* aus der Optimization Toolbox gegeben. Die Funktion *fsolve* verwendet standardmäßig die Methode der Vertrauensbereiche (siehe Abschnitt 2.4), um ein Gleichungssystem in Residuenform  $\mathbf{F}(\mathbf{x}) = \mathbf{0}$  als Minimierungsproblem in  $\mathbf{x}$  zu lösen. Als Beispiel ist in der Code-Auflistung 4.1 der MATLAB-Code dargestellt, wie *fsolve* zur Lösung von (4.127) verwendet werden kann. Der gewünschte Endpunkt  $(\bar{x}_1, \bar{y}_1)$  wird beim Aufruf der Funktion *punktmasse(x1,y1)* übergeben, wobei angenommen wird, dass die Punktmasse am Punkt  $(x_0, y_0) = (0, 0)$  mit der Geschwindigkeit  $(v_0, w_0) = (0, 1)$  in vertikale Richtung startet. Abbildung 4.7 stellt die optimalen Bahnen  $x^*(t)$ ,  $y^*(t)$  der Punktmasse in der  $(x, y)$ -Ebene für verschiedene Endpunkte  $(\bar{x}_1, \bar{y}_1)$  dar. Die Pfeile zeigen die (jeweils konstante) Richtung  $u^* = \arctan(\mu_y^*/\mu_x^*)$  der angreifenden Kraft *ma* an.

Listing 4.1: MATLAB-Code für das Punktmasse-Problem unter Verwendung von *fsolve*.

```
function [t,x,y,p] = punktmasse(x1,y1)
% -----
% (x1,y1): gewünschter Endpunkt
% (t,x,y): Trajektorien der Punktmasse
% p:      Parameterstruktur

p.a = 1; % Parameter
p.x0=0; p.v0=0; p.y0=0; p.w0=1; % Anfangsbedingungen
p.x1=x1; p.y1=y1; % Endbedingungen (Übergabe aus Funktionsaufruf)

opt = optimoptions('fsolve','Display','iter'); % Optionen
X0 = [-1,0,1]; % Startwert
Xopt = fsolve(@eqns,X0,opt,p); % Numerische Lösung mit fsolve
p.mux=Xopt(1); p.muy=Xopt(2); p.t1=Xopt(3); % Lösung

t = linspace(0,p.t1,100); % Trajektorien
x = xfct(p.mux,p.muy,t,p);
y = yfct(p.mux,p.muy,t,p);
% -----
function res = eqns(X,p) % Gleichungen in Residuenform
mux=X(1); muy=X(2); t1=X(3);
res = [ xfct(mux,muy,t1,p) - p.x1;
        yfct(mux,muy,t1,p) - p.y1;
        mux*vfct(mux,muy,t1,p) + muy*wfct(mux,muy,t1,p) + 1 ];
% -----
function x = xfct(mux,muy,t,p) % Funktionen für x und v
cosu = 1/sqrt(1+(muy/mux)^2);
x = p.x0 + p.v0*t + p.a/2*cosu*t.^2; % 't.^2' steht für komponentenweise Auswertung
function v = vfct(mux,muy,t,p)
cosu = 1/sqrt(1+(muy/mux)^2);
v = p.v0 + p.a*cosu*t;
% -----
```

```

function y = yfct(mux,muy,t,p)           % Funktionen für y und w
sinu = muy/(mux*sqrt(1+(muy/mux)^2));
y = p.y0 + p.w0*t + p.a/2*sinu*t.^2;   % 't.^2' steht für komponentenweise Auswertung
function w = wfct(mux,muy,t,p)
sinu = muy/(mux*sqrt(1+(muy/mux)^2));
w = p.w0 + p.a*sinu*t;

```

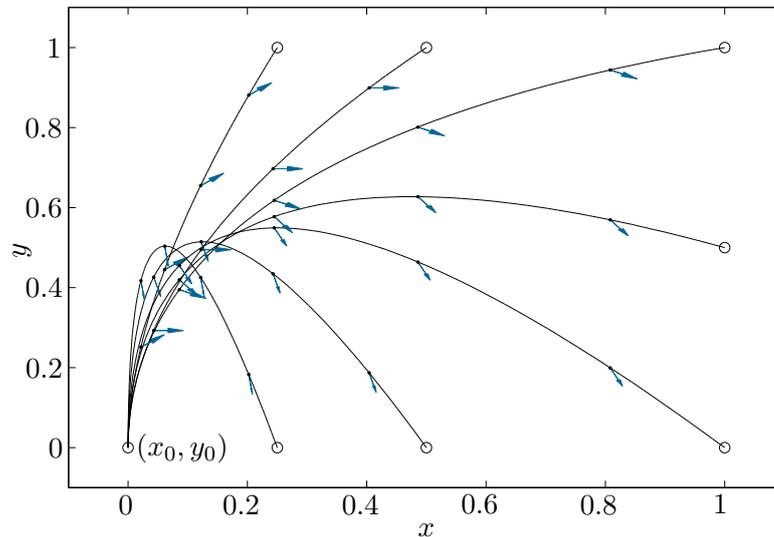


Abbildung 4.7: Zeitoptimale Steuerung einer Punktmasse zu verschiedenen Endpunkten.

**Aufgabe 4.6.** Gegeben ist ein lineares zeitvariantes Mehrgrößensystem der Form

$$\dot{\mathbf{x}} = \mathbf{A}(t)\mathbf{x} + \mathbf{B}(t)\mathbf{u}, \quad \mathbf{x}(t_0) = \mathbf{x}_0 \quad (4.128)$$

mit dem Zustand  $\mathbf{x} \in \mathbb{R}^n$  und dem Stelleingang  $\mathbf{u} \in \mathbb{R}^m$ . Zeigen Sie, dass das zeitvariante Zustandsregelgesetz

$$\mathbf{u}^*(t) = -\mathbf{R}^{-1}(t)\mathbf{B}^T(t)\mathbf{S}(t)\mathbf{x}^*(t) \quad (4.129)$$

mit  $\mathbf{S}(t)$  als Lösung der *Matrix-Riccati-Differentialgleichung*

$$\dot{\mathbf{S}} = -\mathbf{S}\mathbf{A} - \mathbf{A}^T\mathbf{S} + \mathbf{S}\mathbf{B}\mathbf{R}^{-1}\mathbf{B}^T\mathbf{S} - \mathbf{Q}, \quad \mathbf{S}(t_1) = \mathbf{S}_1, \quad t_0 \leq t \leq t_1 \quad (4.130)$$

das Kostenfunktional

$$J(\mathbf{u}) = \frac{1}{2} \int_{t_0}^{t_1} \mathbf{x}^T(t)\mathbf{Q}(t)\mathbf{x}(t) + \mathbf{u}^T(t)\mathbf{R}(t)\mathbf{u}(t) dt + \frac{1}{2} \mathbf{x}^T(t_1)\mathbf{S}_1\mathbf{x}(t_1) \quad (4.131)$$

mit der für alle Zeiten  $t_0 \leq t \leq t_1$  positiv definiten Matrix  $\mathbf{R}(t)$ , der für alle Zeiten  $t_0 \leq t \leq t_1$  positiv semidefiniten Matrix  $\mathbf{Q}(t)$  und der positiv semidefiniten Matrix

$S_1$  minimiert. Dieses Problem ist auch unter dem Namen *LQR (Linear Quadratic Regulator) Problem* bekannt.

#### 4.2.4 Minimumsprinzip von Pontryagin

Für das Weitere betrachte man im ersten Schritt die Minimierung des Kostenfunktional

$$J(\mathbf{u}) = \int_{t_0}^{t_1} l(\mathbf{x}(t), \mathbf{u}(t)) dt \quad (4.132)$$

mit der freien Endzeit  $t_1$  und einem festen Endzustand  $\mathbf{x}(t_1) = \mathbf{x}_1$  unter der Gleichungsbeschränkung des dynamischen Systems

$$\dot{\mathbf{x}} = \mathbf{f}(\mathbf{x}(t), \mathbf{u}(t)), \quad \mathbf{x}(t_0) = \mathbf{x}_0 \quad (4.133)$$

für die zulässigen Stellgrößen

$$\mathbf{u} \in (\hat{C}_U[t_0, T])^m := \left\{ \mathbf{u} \in (\hat{C}[t_0, T])^m \mid \mathbf{u}(t) \in U, \forall t_0 \leq t \leq T \right\} \quad (4.134)$$

mit einer hinreichend großen Zeit  $T \gg t_1$  und der nichtleeren Menge der Stellgrößenbeschränkungen  $U$ .

Man kann nun das Optimierungsproblem bestehend aus (4.132) und (4.133) durch Erweiterung des Zustandsvektors in der Form  $\bar{\mathbf{x}}^T = [\mathbf{x}^T \quad x_{n+1}]$  mit

$$x_{n+1}(t) := \int_{t_0}^t l(\mathbf{x}(\tau), \mathbf{u}(\tau)) d\tau \quad (4.135)$$

wie folgt umformulieren: Gesucht wird eine zulässige Stellgröße  $\mathbf{u}(t) \in (\hat{C}_U[t_0, T])^m$  und eine Endzeit  $t_1$  so, dass die Lösung des erweiterten Systems

$$\underbrace{\begin{bmatrix} \dot{\mathbf{x}} \\ \dot{x}_{n+1} \end{bmatrix}}_{\dot{\bar{\mathbf{x}}}} = \underbrace{\begin{bmatrix} \mathbf{f}(\mathbf{x}, \mathbf{u}) \\ l(\mathbf{x}, \mathbf{u}) \end{bmatrix}}_{\bar{\mathbf{f}}(\mathbf{x}, \mathbf{u})}, \quad \bar{\mathbf{x}}(t_0) = \begin{bmatrix} \mathbf{x}_0 \\ 0 \end{bmatrix} \quad (4.136)$$

beim Punkt  $\bar{\mathbf{x}}^T(t_1) = [\mathbf{x}_1^T \quad x_{n+1}(t_1)]$  terminiert und dabei  $x_{n+1}(t_1)$  möglichst klein gemacht wird. Abbildung 4.8 veranschaulicht diesen Sachverhalt.

Die Linie durch den Punkt  $(\mathbf{x}_1, 0)$  parallel zur  $x_{n+1}$ -Achse beschreibt alle Punkte einer Familie von Trajektorien  $\bar{\mathbf{x}}(t)$  des erweiterten Systems (4.136), die die Bedingung  $\mathbf{x}(t_1) = \mathbf{x}_1$  erfüllen und unterschiedliche Werte des Kostenfunktional  $x_{n+1}(t_1)$  ergeben. Keine andere Trajektorie kann diese vertikale Linie an einem kleineren Wert für  $x_{n+1}(t_1)$  schneiden als  $x_{n+1}^*(t_1^*)$ , der mit der optimalen Stellgröße  $\mathbf{u}^*(t)$  erreicht wird. Diese geometrischen Überlegungen sind auch der Ausgangspunkt für die Herleitung des Minimumsprinzips von Pontryagin. Diese Herleitung wird z. B. in [4.10, 4.11] gezeigt.

An dieser Stelle wird jedoch auf einen Beweis verzichtet und auf die am Ende des Kapitels angegebene Literatur verwiesen.

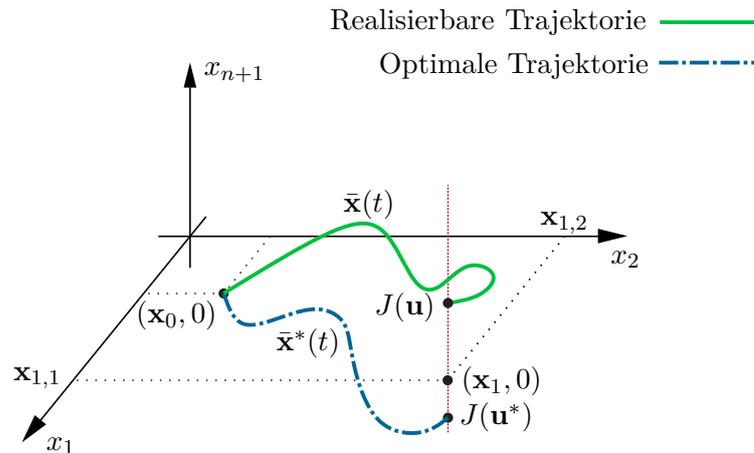


Abbildung 4.8: Zum Minimumsprinzip von Pontryagin.

**Satz 4.10** (Minimumsprinzip von Pontryagin, vorgeschriebener Endzustand). Gesucht ist die Stellgröße  $\mathbf{u} \in (\hat{C}_U[t_0, t_1])^m$  so, dass das Kostenfunktional (Lagrange-Form)

$$J(\mathbf{u}) = \int_{t_0}^{t_1} l(\mathbf{x}(t), \mathbf{u}(t)) dt \quad (4.137)$$

unter den Gleichungsbeschränkungen (dynamisches System)

$$\dot{\mathbf{x}} - \mathbf{f}(\mathbf{x}(t), \mathbf{u}(t)) = \mathbf{0}, \quad \mathbf{x}(t_0) = \mathbf{x}_0, \quad \mathbf{x}(t_1) = \mathbf{x}_1 \quad (4.138)$$

mit fester Anfangszeit  $t_0$  und freier Endzeit  $t_1 \ll T$  minimiert wird. Dabei wird angenommen, dass  $l$  und  $\mathbf{f}$  stetig in  $\mathbf{x}$  und  $\mathbf{u}$  und stetig differenzierbar bezüglich  $\mathbf{x}$  für alle  $(\mathbf{x}, \mathbf{u}) \in \mathbb{R}^n \times \mathbb{R}^m$  sind. Weiters sei  $(\mathbf{u}^*, t_1^*) \in (\hat{C}_U[t_0, T])^m \times [t_0, T]$  die optimale Lösung des Optimierungsproblems und  $\mathbf{x}^*$  die zugehörige Lösung von (4.138). Dann existiert ein  $\bar{\boldsymbol{\lambda}}^* \in (\hat{C}^1[t_0, t_1^*])^{n+1}$ ,  $\bar{\boldsymbol{\lambda}}^* = [\bar{\lambda}_1^* \dots \bar{\lambda}_{n+1}^*]^T \neq \mathbf{0}$  so, dass die Beziehung

$$\dot{\bar{\boldsymbol{\lambda}}^*} = - \left( \frac{\partial}{\partial \bar{\mathbf{x}}} H \right)^T (\mathbf{x}^*(t), \mathbf{u}^*(t), \bar{\boldsymbol{\lambda}}^*(t)) \quad (4.139)$$

für  $t_0 \leq t \leq t_1$  mit  $H(\mathbf{x}^*(t), \mathbf{u}^*(t), \bar{\boldsymbol{\lambda}}^*(t)) = \bar{\boldsymbol{\lambda}}^{*T} \bar{\mathbf{f}}(\mathbf{x}^*(t), \mathbf{u}^*(t))$  erfüllt ist, wobei  $\bar{\mathbf{f}}$  gemäß (4.136) definiert ist, und folgende Eigenschaften gelten:

- (a) Die optimale Lösung  $\mathbf{u}^*(t)$  minimiert die Funktion  $H(\mathbf{x}^*(t), \mathbf{u}(t), \bar{\boldsymbol{\lambda}}^*(t))$  für alle Zeiten  $t_0 \leq t \leq t_1^*$  in der Menge der Stellgrößenbeschränkungen  $U$ , d. h.

$$H(\mathbf{x}^*(t), \mathbf{v}, \bar{\boldsymbol{\lambda}}^*(t)) \geq H(\mathbf{x}^*(t), \mathbf{u}^*(t), \bar{\boldsymbol{\lambda}}^*(t)), \quad \forall \mathbf{v} \in U. \quad (4.140)$$

(b) Es gilt für alle Zeiten  $t_0 \leq t \leq t_1^*$

$$\bar{\lambda}_{n+1}^*(t) = \text{konst.} \geq 0 \quad (4.141a)$$

$$H(\mathbf{x}^*(t), \mathbf{u}^*(t), \bar{\boldsymbol{\lambda}}^*(t)) = \text{konst.} \quad (4.141b)$$

(c) Es gilt die folgende Transversalitätsbedingung (für  $t_1$  frei)

$$H(\mathbf{x}^*(t_1^*), \mathbf{u}^*(t_1^*), \bar{\boldsymbol{\lambda}}^*(t_1^*)) = 0. \quad (4.142)$$

Zur Berechnung der  $m + 2n + 3$  unbekanntenen Größen  $(\mathbf{u}^*(t), \bar{\mathbf{x}}^*(t), \bar{\boldsymbol{\lambda}}^*(t), t_1^*)$  stehen  $m + 2n + 3$  Bedingungen zur Verfügung. Das sind  $m$  algebraische Bedingungen aus der Forderung, dass gemäß (4.140) die Hamiltonfunktion  $H$  zu jedem Zeitpunkt  $t_0 \leq t \leq t_1^*$  am Punkt  $\mathbf{u}^*(t)$  in der Menge  $U$  ein Minimum aufweisen muss, eine algebraische Gleichung in Form der Transversalitätsbedingung (4.142) und  $2n + 2$  Differentialgleichungen für den erweiterten Zustand  $\bar{\mathbf{x}}^*$  gemäß (4.136) und den erweiterten adjungierten Zustand  $\bar{\boldsymbol{\lambda}}^*$  gemäß (4.139). Zu diesen Differentialgleichungen gehören  $n + 1$  Anfangsbedingungen  $\bar{\mathbf{x}}^*(t_0) = [\mathbf{x}_0^T \quad 0]^T$ ,  $n$  Endbedingungen  $\mathbf{x}^*(t_1^*) = \mathbf{x}_1$  sowie die Bedingung  $\bar{\lambda}_{n+1}^*(t_1^*) \geq 0$ . Man beachte, dass daraus noch nicht eindeutig die unbekanntenen Größen  $(\mathbf{u}^*(t), \bar{\mathbf{x}}^*(t), \bar{\boldsymbol{\lambda}}^*(t), t_1^*)$  bestimmt werden können. Es sind nun zwei Fälle zu unterscheiden:

(i) Für  $\bar{\lambda}_{n+1}^* = 0$  liegt ein *abnormaler Fall* vor, da dann wegen

$$H(\mathbf{x}^*, \mathbf{u}^*, \bar{\boldsymbol{\lambda}}^*) = \bar{\boldsymbol{\lambda}}^{*T} \bar{\mathbf{f}}(\mathbf{x}^*, \mathbf{u}^*) = \boldsymbol{\lambda}^{*T} \mathbf{f}(\mathbf{x}^*, \mathbf{u}^*) + \underbrace{\bar{\lambda}_{n+1}^*}_{=0} l(\mathbf{x}^*, \mathbf{u}^*) \quad (4.143)$$

die Hamiltonfunktion  $H$  und folglich auch die Optimalitätsbedingungen gemäß Satz 4.10 unabhängig von der Lagrangeschen Dichte  $l$  sind. In diesem Fall ist die Optimierungsaufgabe *nicht sinnvoll gestellt*.

(ii) Für  $\bar{\lambda}_{n+1}^* > 0$  liegt ein *normaler Fall* vor und  $\boldsymbol{\lambda}^*$  ist bis auf einen multiplikativen Faktor durch die genannten Gleichungen definiert. In der Praxis wählt man meist den Wert  $\bar{\lambda}_{n+1}^* = 1$  und erhält damit genau die Hamiltonfunktion wie sie bereits in Abschnitt 4.2.3 verwendet wurde, siehe (4.93).

Die notwendigen Bedingungen dafür, dass die Hamiltonfunktion  $H(\mathbf{x}, \mathbf{u}, \bar{\boldsymbol{\lambda}})$  für  $\bar{\lambda}_{n+1}^* = 1$  (normaler Fall) bezüglich  $\mathbf{u}$  minimal ist, wie in (4.140) gefordert, entsprechen den notwendigen Bedingungen erster und zweiter Ordnung (4.94c) und (4.96), d. h.

$$\left( \frac{\partial}{\partial \mathbf{u}} H \right) (\mathbf{x}^*, \mathbf{u}^*, \boldsymbol{\lambda}^*) = \mathbf{0} \quad (4.144a)$$

$$\mathbf{d}^T \left( \frac{\partial^2}{\partial \mathbf{u}^2} H \right) (\mathbf{x}^*, \mathbf{u}^*, \boldsymbol{\lambda}^*) \mathbf{d} \geq 0 \quad \forall \mathbf{d} \in \mathbb{R}^m, t \in [t_0, t_1^*], \quad (4.144b)$$

die im Rahmen der Variationsrechnung hergeleitet wurden. Trotz dieser Analogie ist das Minimumsprinzip von Pontryagin nach Satz 4.10 allgemeiner als die Ergebnisse der Variationsrechnung, da die Bedingung  $\frac{\partial}{\partial \mathbf{u}} H = \mathbf{0}$  im Allgemeinen nicht mehr gültig ist,

wenn das Minimum von  $H$  am Rand der Menge  $U$  der Stellgrößenbeschränkungen liegt. Im Weiteren fordert man beim Minimumsprinzip von Pontryagin lediglich die Stetigkeit von  $l$  und  $\mathbf{f}$  bezüglich  $\mathbf{u}$ , wohingegen bei der Herleitung der Euler-Lagrange Gleichungen die stetige Differenzierbarkeit bezüglich  $\mathbf{u}$  gefordert wurde, siehe Satz 4.8.

*Beispiel 4.7 (Abnormaler Fall).* Gegeben ist das Optimalsteuerungsproblem

$$\min_{u(\cdot)} \int_0^1 l(x, u) dt \quad (4.145a)$$

$$\text{u.B.v. } \dot{x} = u, \quad x(0) = 0, \quad x(1) = 1 \quad (4.145b)$$

$$u(t) \in [0, 1]. \quad (4.145c)$$

Es existiert nur eine realisierbare Steuerung  $u^*(t) = 1$ , die den Zustand  $x^*(t) = t$  von  $x^*(0) = 0$  nach  $x^*(1) = 1$  überführt. Somit ist die optimale Lösung unabhängig von der Wahl der Kostenfunktion  $l(x, u)$  und es liegt ein abnormaler Fall vor.

*Beispiel 4.8.* Gesucht ist das Minimum des Kostenfunktional

$$J(u) = \frac{1}{2} \int_0^1 u^2(t) dt \quad (4.146)$$

für das dynamische System

$$\dot{x} = -x + u, \quad x(0) = 1, \quad x(1) = 0 \quad (4.147)$$

unter Berücksichtigung der Stellgrößenbeschränkung  $-0.6 \leq u(t) \leq 0$  für alle  $0 \leq t \leq 1$ . Die Hamiltonfunktion  $H$  von Satz 4.10 für dieses Beispiel lautet

$$H(x, u, \bar{\lambda}) = \bar{\lambda}_1(-x + u) + \bar{\lambda}_2 \frac{1}{2} u^2 \quad (4.148)$$

und die adjungierten Zustände  $\bar{\lambda}$  erfüllen gemäß (4.139) und (4.141a) die Gleichungen

$$\frac{d}{dt} \bar{\lambda}_1^* = - \left( \frac{\partial}{\partial x} H \right) (x^*, u^*, \bar{\lambda}^*) = \bar{\lambda}_1^* \quad (4.149a)$$

$$\frac{d}{dt} \bar{\lambda}_2^* = 0. \quad (4.149b)$$

Daraus folgt die Lösung

$$\bar{\lambda}_1^*(t) = C_1 e^t \quad \text{und} \quad \bar{\lambda}_2^*(t) = C_2 \quad (4.150)$$

für geeignete Konstanten  $C_1$  und  $C_2 \geq 0$ . Im Weiteren setzt man  $C_2 = 1$ , da ein *normaler Fall* vorliegt. Die optimale Lösung  $u^*$  mit  $-0.6 \leq u^* \leq 0$  muss gemäß

(4.140) der Ungleichung

$$H(x^*(t), v, \bar{\lambda}^*(t)) \geq H(x^*(t), u^*(t), \bar{\lambda}^*(t)), \quad \forall v \in [-0.6, 0], \quad \forall t \in [0, 1] \quad (4.151)$$

genügen. Aus

$$\left(\frac{\partial}{\partial u} H\right)(x^*, u^*, \bar{\lambda}^*) = \bar{\lambda}_1^* + \bar{\lambda}_2^* u^* = \bar{\lambda}_1^* + u^* \quad (4.152)$$

folgt für die optimale Stellgröße unter Berücksichtigung der Stellgrößenbeschränkung

$$u^*(t) = \begin{cases} 0 & \text{für } \bar{\lambda}_1^* \leq 0 \\ -\bar{\lambda}_1^* = -C_1 e^t & \text{für } 0 < \bar{\lambda}_1^* < 0.6 \\ -0.6 & \text{für } \bar{\lambda}_1^* \geq 0.6 \end{cases} \quad (4.153)$$

Hieraus folgt, dass  $C_1 > 0$  gelten muss, denn für  $C_1 \leq 0$  ist  $\bar{\lambda}_1^* \leq 0$  und damit  $u^*(t) = 0$  für  $0 \leq t \leq 1$ , woraus aber wegen  $x^*(1) = e^{-1} \neq 0$  keine zulässige Lösung resultiert. Mit  $C_1 > 0$  bzw.  $\bar{\lambda}_1^* > 0$  muss deshalb die optimale Stellgröße  $u^*(t)$  zwischen  $-C_1 e^t$  und  $-0.6$  umschalten. Da  $\bar{\lambda}_1^*(t) = C_1 e^t$  streng monoton steigend in  $t$  ist, setzt man eine stückweise stetige Steuerung

$$u^*(t) = \begin{cases} -C_1 e^t & \text{falls } t \in [0, c^*] \\ -0.6 & \text{falls } t \in (c^*, 1] \end{cases} \quad (4.154)$$

mit einem Umschaltzeitpunkt (Eckpunkt)  $t = c^*$  an. Für das Zeitintervall  $[0, c^*]$  errechnet sich die Lösung von

$$\dot{x}_{(1)}^*(t) = -x_{(1)}^*(t) + u^*(t), \quad x_{(1)}^*(0) = 1 \quad (4.155)$$

zu

$$x_{(1)}^*(t) = -\frac{1}{2}C_1 e^t + e^{-t} \left(\frac{1}{2}C_1 + 1\right) \quad (4.156)$$

und für das Zeitintervall  $[c^*, 1]$  folgt aus

$$\dot{x}_{(2)}^*(t) = -x_{(2)}^*(t) + u^*(t), \quad x_{(2)}^*(1) = 0 \quad (4.157)$$

die Lösung zu

$$x_{(2)}^*(t) = 0.6(e^{1-t} - 1) \quad (4.158)$$

Nach (4.141b) muss die Hamiltonfunktion  $H(x^*(t), u^*(t), \bar{\lambda}^*(t))$  im gesamten Zeitintervall konstant sein. Daraus folgt

$$\begin{aligned} \bar{\lambda}_1^*(\tau_1) \left(-x_{(1)}^*(\tau_1) + u^*(\tau_1)\right) + \frac{1}{2}(u^*(\tau_1))^2 = \\ \bar{\lambda}_1^*(\tau_2) \left(-x_{(2)}^*(\tau_2) + u^*(\tau_2)\right) + \frac{1}{2}(u^*(\tau_2))^2 \quad \forall \tau_1 \in [0, c^*], \tau_2 \in (c^*, 1] \end{aligned} \quad (4.159a)$$

und nach Einsetzen

$$-C_1 \left( 1 + \frac{1}{2} C_1 \right) = -C_1 0.6e + \frac{1}{2} 0.6^2 . \quad (4.159b)$$

Hieraus ergeben sich die zwei möglichen Lösungen  $C_{1,1} = 0.436$  und  $C_{1,2} = 0.826$ . Der Zeitpunkt der Umschaltung (Eckpunkt)  $t = c^*$  folgt aus der Stetigkeitsbedingung der Zustandsgröße

$$x_{(1)}^*(c^*) = x_{(2)}^*(c^*) \quad (4.160)$$

zu  $c_1^* = 0.32$  für  $C_{1,1}$  und  $c_1^* = -0.32$  für  $C_{1,2}$ . Die für das betrachtete Zeitintervall  $0 \leq t \leq 1$  relevante Lösung lautet daher  $C_1 = C_{1,1} = 0.436$ .

Im nächsten Schritt wird gezeigt, wie sich das Minimumsprinzip von Pontryagin nach Satz 4.10 ändert, wenn die Endbedingung  $\mathbf{x}(t_1) = \mathbf{x}_1$  durch eine *Endbeschränkung* der Form  $\mathbf{x}(t_1) \in \mathcal{X}_1$  mit einer glatten Mannigfaltigkeit  $\mathcal{X}_1$  (siehe auch Abschnitt 3.1.2.1) der Dimension  $n - p$  ersetzt wird. Diese  $(n - p)$ -dimensionale Mannigfaltigkeit wird durch  $p$  Gleichungen der Form

$$\psi_k(\mathbf{x}) = 0, \quad k = 1, \dots, p \quad (4.161)$$

beschrieben. Der Tangentialraum  $\mathcal{T}_{\check{\mathbf{x}}}\mathcal{X}_1$  an der Stelle  $\mathbf{x} = \check{\mathbf{x}}$  ist dann in der Form

$$\mathcal{T}_{\check{\mathbf{x}}}\mathcal{X}_1 = \left\{ \mathbf{d} \mid \left( \frac{\partial}{\partial \mathbf{x}} \psi_k(\check{\mathbf{x}}) \right) \mathbf{d} = 0, k = 1, \dots, p \right\} \quad (4.162)$$

definiert (siehe auch Abschnitt 3.1.2.1).

**Satz 4.11 (Minimumsprinzip von Pontryagin, beschränkter Endzustand).** *Gesucht ist die Stellgröße  $\mathbf{u} \in (\hat{C}_U[t_0, t_1])^m$  so, dass das Kostenfunktional (Lagrange-Form)*

$$J(\mathbf{u}) = \int_{t_0}^{t_1} l(\mathbf{x}(t), \mathbf{u}(t)) dt \quad (4.163)$$

*unter den Gleichungsbeschränkungen (dynamisches System)*

$$\dot{\mathbf{x}} - \mathbf{f}(\mathbf{x}(t), \mathbf{u}(t)) = \mathbf{0}, \quad \mathbf{x}(t_0) = \mathbf{x}_0, \quad \mathbf{x}(t_1) \in \mathcal{X}_1 \quad (4.164)$$

*mit fester Anfangszeit  $t_0$ , freier Endzeit  $t_1 \ll T$  und der glatten  $(n - p)$ -dimensionalen Mannigfaltigkeit  $\mathcal{X}_1$  minimiert wird. Dabei wird angenommen, dass  $l$  und  $\mathbf{f}$  stetig in  $\mathbf{x}$  und  $\mathbf{u}$  und stetig differenzierbar bezüglich  $\mathbf{x}$  für alle  $(\mathbf{x}, \mathbf{u}) \in \mathbb{R}^n \times \mathbb{R}^m$  sind. Weiters sei  $(\mathbf{u}^*, t_1^*) \in (\hat{C}_U[t_0, T])^m \times [t_0, T]$  die optimale Lösung des Optimierungsproblems und  $\mathbf{x}^*$  die zugehörige Lösung von (4.164). Dann existiert ein  $\bar{\boldsymbol{\lambda}}^* \in (\hat{C}^1[t_0, t_1^*])^{n+1}$ ,  $\bar{\boldsymbol{\lambda}}^* = [(\boldsymbol{\lambda}^*)^\top \quad \bar{\lambda}_{n+1}^*]^\top = [\bar{\lambda}_1^* \quad \dots \quad \bar{\lambda}_{n+1}^*]^\top \neq \mathbf{0}$  so, dass die Beziehungen (4.139)–(4.142) von Satz 4.10 erfüllt sind und  $\boldsymbol{\lambda}^*(t_1^*) = [\lambda_1^*(t_1^*) \quad \dots \quad \lambda_n^*(t_1^*)]^\top$  orthogonal zum Tangentialraum  $\mathcal{T}_{\mathbf{x}^*(t_1^*)}\mathcal{X}_1$  ist, d. h. es gelten die Transversalitätsbedingungen*

$$(\boldsymbol{\lambda}^*)^\top(t_1^*) \mathbf{d} = 0, \quad \forall \mathbf{d} \in \mathcal{T}_{\mathbf{x}^*(t_1^*)}\mathcal{X}_1 . \quad (4.165)$$

Nach Satz 4.11 und (4.162) muss  $\lambda^*(t_1^*)$  sich also als Linearkombination der Gradienten  $\left(\frac{\partial}{\partial \mathbf{x}}\psi_k\right)(\mathbf{x}_1^*)$  mit  $k = 1, \dots, p$  darstellen lassen, d. h. in der Form

$$\lambda^*(t_1^*) = \sum_{k=1}^p \mu_k \left(\frac{\partial}{\partial \mathbf{x}}\psi_k\right)^T(\mathbf{x}_1^*), \quad \mathbf{x}_1^* = \mathbf{x}^*(t_1^*) \quad (4.166)$$

mit dem Lagrange-Multiplikator  $\boldsymbol{\mu} = [\mu_1 \ \dots \ \mu_p]^T \in \mathbb{R}^p$ . Die Bedingung

$$\text{rang}\left(\left(\frac{\partial}{\partial \mathbf{x}}\boldsymbol{\psi}\right)(\mathbf{x}_1^*)\right) = p, \quad \boldsymbol{\psi} = [\psi_1 \ \dots \ \psi_p]^T \quad (4.167)$$

entspricht der LICQ (linear independence constraint qualification) Bedingung der statischen Optimierung mit Gleichungsbeschränkungen, siehe auch Definition 3.2.

Für *zeitvariante Systeme*

$$\dot{\mathbf{x}} = \mathbf{f}(t, \mathbf{x}(t), \mathbf{u}(t)), \quad \mathbf{x}(t_0) = \mathbf{x}_0 \quad (4.168)$$

führt man eine weitere Zustandsgröße der Form  $x_{n+1} = t$  ein und entwirft für das erweiterte System

$$\frac{d}{dt} \underbrace{\begin{bmatrix} \mathbf{x} \\ x_{n+1} \end{bmatrix}}_{\mathbf{x}_e} = \underbrace{\begin{bmatrix} \mathbf{f}(x_{n+1}, \mathbf{x}, \mathbf{u}) \\ 1 \end{bmatrix}}_{\mathbf{f}_e(\mathbf{x}_e, \mathbf{u})} \quad (4.169)$$

die optimale Steuerung gemäß Satz 4.10 oder Satz 4.11. Dabei wird vorausgesetzt, dass  $\mathbf{f}$  und  $l$  stetig differenzierbar in  $t$  sind.

### 4.2.5 Minimumsprinzip für eingangsaffine Systeme

Den weiteren Betrachtungen liege das *eingangsaffine System*

$$\dot{\mathbf{x}} = \mathbf{f}(\mathbf{x}, \mathbf{u}) = \mathbf{f}_0(\mathbf{x}) + \sum_{i=1}^m \mathbf{f}_i(\mathbf{x})u_i \quad (4.170)$$

mit den Stellgrößenbeschränkungen der Form

$$\mathbf{u} \in U = [\mathbf{u}^-, \mathbf{u}^+] \quad \text{bzw.} \quad u_i \in [u_i^-, u_i^+], \quad i = 1, \dots, m \quad (4.171)$$

(box constraints) zugrunde.

#### 4.2.5.1 Kostenfunktional mit verbrauchsoptimalem Anteil

In der Literatur findet man im Zusammenhang mit dem Entwurf von *verbrauchsoptimalen Steuerungen* häufig Kostenfunktionale der Form

$$J(\mathbf{u}) = \int_{t_0}^{t_1} l_0(\mathbf{x}) + \sum_{i=1}^m r_i |u_i| \, dt, \quad r_i > 0. \quad (4.172)$$

Die zu (4.170) und (4.172) passende Hamiltonfunktion lautet (mit  $\bar{\lambda}_{n+1} = 1$ )

$$H(\mathbf{x}, \mathbf{u}, \boldsymbol{\lambda}) = l_0(\mathbf{x}) + \sum_{i=1}^m r_i |u_i| + \boldsymbol{\lambda}^T \left( \mathbf{f}_0(\mathbf{x}) + \sum_{i=1}^m \mathbf{f}_i(\mathbf{x}) u_i \right). \quad (4.173)$$

Da der Anteil  $l_0(\mathbf{x}) + \boldsymbol{\lambda}^T \mathbf{f}_0(\mathbf{x})$  in  $H$  unabhängig von  $\mathbf{u}$  ist, kann er im Minimierungsproblem (4.140) vernachlässigt werden. Die Minimierung von  $H$

$$\min_{u_i \in [u_i^-, u_i^+]} H_i(u_i) = r_i |u_i| + q_i(\mathbf{x}, \boldsymbol{\lambda}) u_i, \quad q_i(\mathbf{x}, \boldsymbol{\lambda}) = \boldsymbol{\lambda}^T \mathbf{f}_i(\mathbf{x}) \quad (4.174)$$

kann nun für jedes  $u_i$ ,  $i = 1, \dots, m$ , separat durchgeführt werden. Der Term  $q_i(\mathbf{x}, \boldsymbol{\lambda})$  spielt eine wichtige Rolle bei der Lösung dieses Problems. Im aktuellen Abschnitt wird davon ausgegangen, dass  $u_i^- < 0 < u_i^+$  gilt. Sind diese Bedingungen nicht erfüllt, lassen sich analog zu den nachfolgenden Ausführungen sehr einfach Vorschriften zur Bestimmung der optimalen Stellgröße ableiten. Abbildung 4.9 illustriert die unterschiedlichen Fälle a)–d) mit denen die optimale Stellgröße

$$u_i^* = \begin{cases} u_i^- & \text{falls } q_i(\mathbf{x}, \boldsymbol{\lambda}) > r_i \\ 0 & \text{falls } q_i(\mathbf{x}, \boldsymbol{\lambda}) \in (-r_i, r_i), \\ u_i^+ & \text{falls } q_i(\mathbf{x}, \boldsymbol{\lambda}) < -r_i \end{cases} \quad \forall i = 1, \dots, m \quad (4.175)$$

komponentenweise bestimmt wird. Ein kritischer Fall liegt vor, falls auf einem Subintervall  $I_s \subset [t_0, t_1]$  die Bedingung  $q_i(\mathbf{x}(t), \boldsymbol{\lambda}(t)) = \pm r_i$  identisch erfüllt ist. Die optimale Stellgröße  $u_i^*$  ist dann nicht mehr eindeutig aus der Minimierungsbedingung (4.174) bestimmbar. Dieser Fall wird als *singulär* bezeichnet und im Abschnitt 4.2.6 näher erläutert.

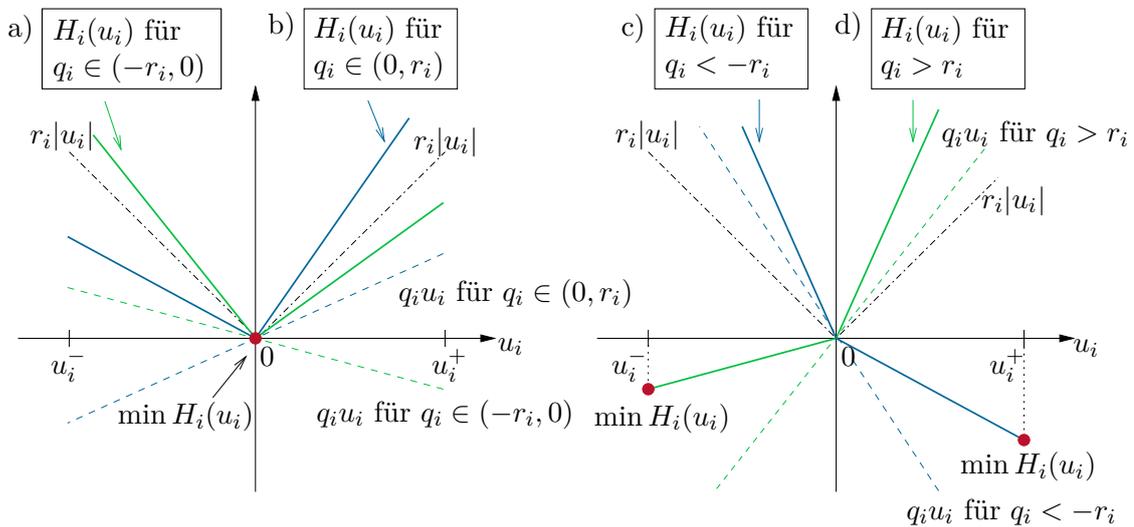


Abbildung 4.9: Verbrauchsoptimaler Fall, grafische Veranschaulichung von (4.174).

### 4.2.5.2 Kostenfunktional mit energieoptimalem Anteil

Unter dem Begriff *energieoptimale Steuerung* wird häufig die Minimierung eines Kostenfunktionals der Form

$$J(\mathbf{u}) = \int_{t_0}^{t_1} l_0(\mathbf{x}) + \frac{1}{2} \sum_{i=1}^m r_i u_i^2 dt, \quad r_i > 0 \quad (4.176)$$

verstanden. Analog zum vorherigen Fall kann die Minimierung der Hamiltonfunktion  $H$

$$\min_{u_i \in [u_i^-, u_i^+]} H_i(u_i) = \frac{1}{2} r_i u_i^2 + q_i(\mathbf{x}, \boldsymbol{\lambda}) u_i, \quad q_i(\mathbf{x}, \boldsymbol{\lambda}) = \boldsymbol{\lambda}^T \mathbf{f}_i(\mathbf{x}) \quad (4.177)$$

wieder für jedes  $u_i$ ,  $i = 1, \dots, m$ , separat erfolgen. Durch den quadratischen Term mit  $r_i > 0$  hätte die Funktion  $H_i(u_i)$  im unbeschränkten Fall stets ein *Minimum* an der Stelle

$$u_i^0 = -\frac{1}{r_i} q_i(\mathbf{x}, \boldsymbol{\lambda}). \quad (4.178)$$

Falls  $u_i^0$  innerhalb des zulässigen Intervalls  $[u_i^-, u_i^+]$  liegt, ist die optimale Lösung von (4.170), (4.171) und (4.176) durch  $u_i^* = u_i^0$  gegeben. Falls  $u_i^0$  außerhalb von  $[u_i^-, u_i^+]$  liegt, so befindet sich das Minimum von  $H_i(u_i)$  an der Schranke  $u_i^-$  oder  $u_i^+$ , da  $H_i(u_i)$  für  $u_i^0 < u_i^-$  (bzw.  $u_i^0 > u_i^+$ ) im Intervall  $[u_i^-, u_i^+]$  *streng monoton steigend (fallend)* ist, siehe Abbildung 4.10. Somit ist die optimale Stellgröße  $\mathbf{u}^*(t)$  komponentenweise wie folgt definiert

$$u_i^* = \begin{cases} u_i^- & \text{falls } u_i^0 \leq u_i^- \\ u_i^0 & \text{falls } u_i^0 \in (u_i^-, u_i^+) \\ u_i^+ & \text{falls } u_i^0 \geq u_i^+ \end{cases}, \quad i = 1, \dots, m. \quad (4.179)$$

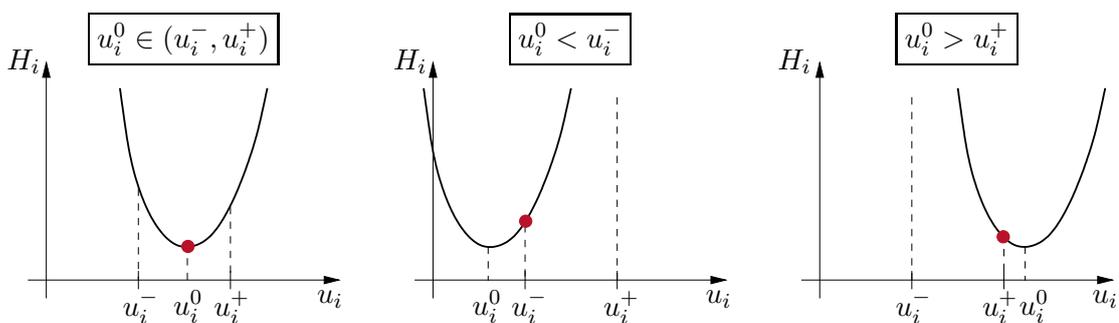


Abbildung 4.10: Energieoptimaler Fall, grafische Veranschaulichung von (4.177).

### 4.2.5.3 Zeitoptimales Kostenfunktional

Für zeitoptimale Probleme lautet das Kostenfunktional

$$J(\mathbf{u}) = \int_{t_0}^{t_1} 1 dt = t_1 - t_0 \quad (4.180)$$

und die Hamiltonfunktion lässt sich in der Form (mit  $\bar{\lambda}_{n+1} = 1$ )

$$H(\mathbf{x}, \mathbf{u}, \boldsymbol{\lambda}) = 1 + \boldsymbol{\lambda}^T \left( \mathbf{f}_0(\mathbf{x}) + \sum_{i=1}^m \mathbf{f}_i(\mathbf{x}) u_i \right) \quad (4.181)$$

anschreiben. Minimiert man die Hamiltonfunktion  $H$  wieder für jedes  $u_i$ ,  $i = 1, \dots, m$  separat

$$\min_{u_i \in [u_i^-, u_i^+]} H_i(u_i) = q_i(\mathbf{x}, \boldsymbol{\lambda}) u_i, \quad q_i(\mathbf{x}, \boldsymbol{\lambda}) = \boldsymbol{\lambda}^T \mathbf{f}_i(\mathbf{x}), \quad (4.182)$$

so erhält man die optimale Stellgröße  $\mathbf{u}^*$  direkt in Abhängigkeit des Vorzeichens von  $q_i(\mathbf{x}, \boldsymbol{\lambda})$ ,  $i = 1, \dots, m$ , in der Form

$$u_i^* = \begin{cases} u_i^- & \text{falls } q_i(\mathbf{x}, \boldsymbol{\lambda}) > 0 \\ u_i^+ & \text{falls } q_i(\mathbf{x}, \boldsymbol{\lambda}) < 0 \end{cases}, \quad i = 1, \dots, m. \quad (4.183)$$

Diese Steuerung wird in der Literatur häufig als *Bang-Bang-Steuerung* bezeichnet, da lediglich zwischen den Maximal- und Minimalwerten des Stellgrößenbereiches hin- und hergeschaltet wird. Ein singulärer Fall liegt vor, falls  $q_i(\mathbf{x}(t), \boldsymbol{\lambda}(t))$  auf einem Subintervall  $I_s \subset [t_0, t_1]$  identisch Null ist. Die Hamiltonfunktion  $H$  ist dann *unabhängig von*  $u_i$ , sodass  $H$  für jeden beliebigen Wert von  $u_i$  trivialerweise minimal ist. Die Minimumsforderung (4.182) ist damit zwar erfüllt, liefert aber keine Informationen über die Wahl von  $u_i$ .

In der Praxis wird der singuläre Fall oft durch einen zusätzlichen *Regularisierungsterm*

$$J(\mathbf{u}) = \int_{t_0}^{t_1} 1 + \frac{1}{2} \sum_{i=1}^m r_i u_i^2 dt, \quad r_i > 0 \quad (4.184)$$

vermieden, wobei  $r_i$  hinreichend klein gewählt wird, um annähernd Zeitoptimalität zu erzielen. Der Regularisierungsterm entspricht natürlich einem energieoptimalen Anteil, so dass (4.184) die Form (4.176) besitzt und die optimale Stellgröße  $\mathbf{u}^*$  gemäß (4.179) berechnet werden kann.

**Beispiel 4.9 (Doppelintegrator).** Zur Veranschaulichung des Minimumsprinzips von Pontryagin wird die *zeitminimale* Überführung eines Doppelintegrators mit beschränktem Eingang in den Ursprung  $\mathbf{x}(t_1) = \mathbf{0}$  betrachtet. Das Optimalsteuerungsproblem mit dem Zustand  $\mathbf{x} = [x_1 \ x_2]^T$  kann wie folgt formuliert werden

$$\min_{u(\cdot)} \int_{t_0=0}^{t_1} dt = t_1 \quad (4.185a)$$

$$\text{u.B.v. } \dot{x}_1 = x_2, \quad \dot{x}_2 = u \quad (4.185b)$$

$$\mathbf{x}(0) = \mathbf{x}_0, \quad \mathbf{x}(t_1) = \mathbf{0} \quad (4.185c)$$

$$|u(t)| \leq 1 \quad \forall t \in [0, t_1]. \quad (4.185d)$$

Mit der Hamiltonfunktion ( $\bar{\lambda}_{n+1} = 1$ )  $H(\mathbf{x}, u, \boldsymbol{\lambda}) = 1 + \lambda_1 x_2 + \lambda_2 u$  ergeben sich die adjungierten Zustände  $\boldsymbol{\lambda}^* = [\lambda_1^* \quad \lambda_2^*]^T$  aus (4.139) zu

$$\dot{\lambda}_1^* = 0 \quad \Rightarrow \quad \lambda_1^*(t) = C_1 \quad (4.186a)$$

$$\dot{\lambda}_2^* = -\lambda_1^* \quad \Rightarrow \quad \lambda_2^*(t) = -C_1 t + C_2, \quad (4.186b)$$

wobei  $C_1$  und  $C_2$  Integrationskonstanten darstellen. Die Minimierungsbedingung (4.140) für die Hamiltonfunktion führt auf die optimale Stellgröße

$$u^*(t) = \begin{cases} +1 & \text{falls } \lambda_2^* < 0 \\ -1 & \text{falls } \lambda_2^* > 0. \end{cases} \quad (4.187)$$

Der *singuläre Fall*, d. h.  $\lambda_2^*(t) = 0$  auf einem nicht verschwindenden Subintervall  $t \in I_s \subseteq [t_0, t_1]$ , kann hier nicht auftreten, da dann aufgrund von (4.186)  $\lambda_2^*(t) = 0$  und  $\lambda_1^*(t) = 0$  auf dem Gesamtintervall  $[0, t_1]$  gelten müsste. Dies widerspricht aber der Transversalitätsbedingung (4.142) für die *freie Endzeit*  $t_1$

$$H(\mathbf{x}^*, u^*, \boldsymbol{\lambda}^*)|_{t=t_1^*} = 1 + \lambda_1^*(t_1^*) x_2^*(t_1^*) + \lambda_2^*(t_1^*) u^*(t_1^*) = 0. \quad (4.188)$$

Der Fall  $\lambda_2^*(t) = 0$  kann also nur zu diskreten Zeitpunkten auftreten. Zu diesen Zeitpunkten wird  $u^*(t)$  zwischen  $-1$  und  $+1$  umgeschaltet. Da  $\lambda_2^*(t) = -C_1 t + C_2$ , gibt es *maximal einen Umschaltzeitpunkt*  $t = t_s$  im Zeitintervall  $[0, t_1^*]$ , an dem  $u^*(t)$  zwischen  $-1$  und  $+1$  wechselt. Somit existieren lediglich *vier mögliche Schaltsequenzen*  $\{+1\}$ ,  $\{-1\}$ ,  $\{+1, -1\}$ ,  $\{-1, +1\}$ , die für eine optimale Lösung in Frage kommen. Da  $u$  stets auf einem Zeitintervall konstant ist, stellen die Trajektorien in der  $(x_1, x_2)$ -Ebene *Parabeln* dar.

**Aufgabe 4.7.** Zeigen Sie, dass die Lösung von (4.185b) für  $u(t) = \pm 1 = \text{konst.}$  die folgende Parabelgleichung erfüllt

$$x_1 = \frac{x_2^2}{2u} + c, \quad u = \pm 1. \quad (4.189)$$

Das Optimierungsziel ist das *schnellstmögliche Erreichen des Ursprungs*  $\mathbf{x}(t_1) = \mathbf{0}$  ausgehend von einem beliebigen Anfangspunkt  $\mathbf{x}(0) = \mathbf{x}_0$ . Der Ursprung ist aber nur über die *Schaltkurve*  $x_1 = x_2^2/2$  für  $u = 1$  bzw.  $x_1 = -x_2^2/2$  für  $u = -1$  erreichbar, siehe Abbildung 4.11. Diese beiden Fälle können mit der Funktion

$$S(x_2) = -\frac{1}{2} x_2 |x_2| \quad (4.190)$$

unterschieden werden. Da lediglich die Schaltsequenzen  $\{+1\}$ ,  $\{-1\}$ ,  $\{+1, -1\}$ ,  $\{-1, +1\}$  für  $u$  in Frage kommen, gibt es nur eine Möglichkeit, um den Systemzustand mit  $\mathbf{x}(0) = \mathbf{x}_0 = [x_{1,0} \quad x_{2,0}]^T$  schnellstmöglich zum Ursprung  $\mathbf{x}(t_1) = \mathbf{0}$  zu bringen:

- Falls  $x_{1,0} = S(x_{2,0})$  gilt, ist keine Umschaltung notwendig, und  $\mathbf{x}(t_1) = \mathbf{0}$  wird direkt über die Schaltkurve mit  $u(t) = 1$  für  $x_{1,0} > 0$  oder  $u(t) = -1$  für  $x_{1,0} < 0$  erreicht, siehe Abbildung 4.11.
- Falls  $\mathbf{x}_0$  nicht auf der Schaltkurve liegt, d. h.  $x_{1,0} < S(x_{2,0})$  oder  $x_{1,0} > S(x_{2,0})$ , ist genau eine Umschaltung notwendig, um zunächst die Schaltkurve zu erreichen und anschließend entlang dieser Kurve zum Ursprung zu laufen, siehe Abbildung 4.12.

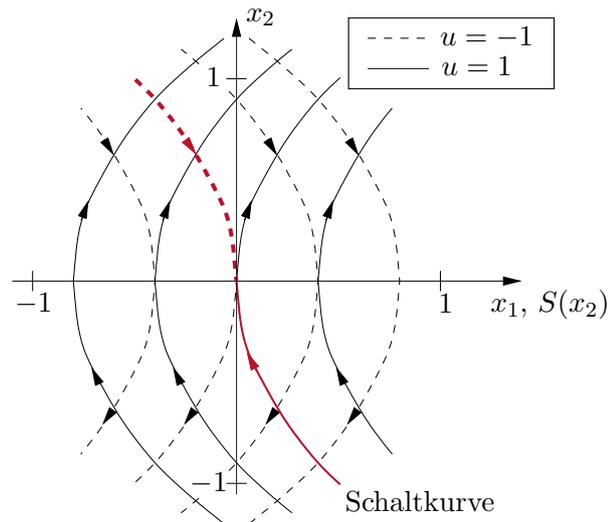


Abbildung 4.11: Mögliche Trajektorien des Doppelintegrators in der  $(x_1, x_2)$ -Ebene.

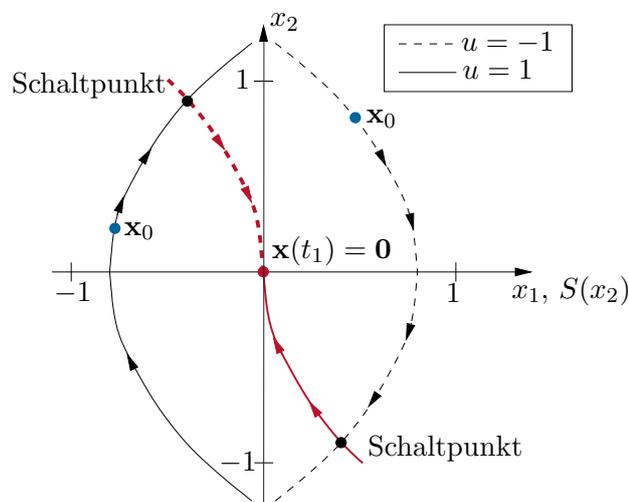


Abbildung 4.12: Optimale Umschaltung für verschiedene Anfangspunkte  $\mathbf{x}_0$ .

Das optimale Stellgesetz lautet also

$$u^*(t) = \begin{cases} +1 & \text{falls } x_1 < S(x_2) \\ +1 & \text{falls } x_1 = S(x_2) \text{ und } x_1 > 0 \\ -1 & \text{falls } x_1 > S(x_2) \\ -1 & \text{falls } x_1 = S(x_2) \text{ und } x_1 < 0. \end{cases} \quad (4.191)$$

Daraus können auch der Schaltzeitpunkt  $t_s$  und die minimale Endzeit  $t_1^*$  berechnet werden.

**Aufgabe 4.8.** Verifizieren Sie, dass der optimale Umschaltzeitpunkt  $t_s$  und die minimale Endzeit  $t_1^*$  wie folgt definiert sind

$$t_s = \begin{cases} x_{2,0} + \sqrt{\frac{1}{2}x_{2,0}^2 + x_{1,0}} & \text{falls } x_{1,0} > S(x_{2,0}) \\ -x_{2,0} + \sqrt{\frac{1}{2}x_{2,0}^2 - x_{1,0}} & \text{falls } x_{1,0} < S(x_{2,0}) \end{cases} \quad (4.192)$$

$$t_1^* = \begin{cases} x_{2,0} + \sqrt{2x_{2,0}^2 + 4x_{1,0}} & \text{falls } x_{1,0} > S(x_{2,0}) \\ -x_{2,0} + \sqrt{2x_{2,0}^2 - 4x_{1,0}} & \text{falls } x_{1,0} < S(x_{2,0}) \\ |x_{2,0}| & \text{falls } x_{1,0} = S(x_{2,0}). \end{cases} \quad (4.193)$$

Abbildung 4.13 zeigt die zeitoptimalen Trajektorien für den Doppelintegrator für verschiedene Anfangswerte  $\mathbf{x}_0 = [x_{1,0} \ x_{2,0}]^T$ .

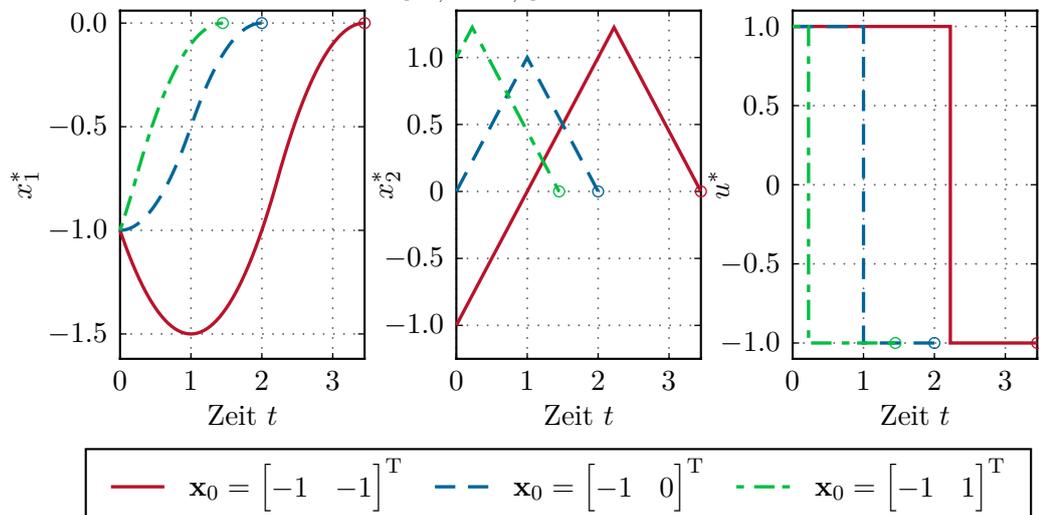


Abbildung 4.13: Zeitoptimale Trajektorien des Doppelintegrators für verschiedene Anfangswerte  $\mathbf{x}_0$ .

**Aufgabe 4.9.** Implementieren Sie das System (4.185b) mit dem optimalen Stellgesetz (4.191) in MATLAB/SIMULINK und verifizieren Sie die Ergebnisse in Abbildung 4.13 für verschiedene Anfangswerte  $\mathbf{x}_0$ . Verwenden Sie  $t_1^*$  gemäß (4.193) als Zeithorizont für die Simulation.

### 4.2.6 Der singuläre Fall

Wenn auf einem endlichen Subintervall  $I_s \subseteq [t_0, t_1]$  die optimale Stellgröße  $\mathbf{u}^*$  nicht aus der Minimierungsbedingung (4.140) bestimmt werden kann, so liegt ein singulärer Fall vor. Zur Verdeutlichung dieser Problematik soll im Weiteren das Optimalsteuerungsproblem

$$\min_{u(\cdot)} J(u) = \int_{t_0}^{t_1} l_0(\mathbf{x}) + l_1(\mathbf{x})u \, dt \quad (4.194a)$$

$$\text{u.B.v. } \dot{\mathbf{x}} = \mathbf{f}_0(\mathbf{x}) + \mathbf{f}_1(\mathbf{x})u, \quad \mathbf{x}(t_0) = \mathbf{x}_0 \quad (4.194b)$$

$$u \in \hat{C}_U[t_0, t_1] \quad (4.194c)$$

mit skalarer und affin auftretender Stellgröße  $u(t)$  betrachtet werden. Die grundsätzliche Vorgehensweise ist aber auch auf allgemeinere Optimalsteuerungsprobleme anwendbar. Die Hamiltonfunktion

$$H(\mathbf{x}, u, \boldsymbol{\lambda}) = l_0(\mathbf{x}) + \boldsymbol{\lambda}^T \mathbf{f}_0(\mathbf{x}) + (l_1(\mathbf{x}) + \boldsymbol{\lambda}^T \mathbf{f}_1(\mathbf{x}))u \quad (4.195)$$

ist affin in  $u$ . Die Funktion

$$\zeta(t) = \left( \frac{\partial}{\partial u} H \right) (\mathbf{x}, u, \boldsymbol{\lambda}) = l_1(\mathbf{x}) + \boldsymbol{\lambda}^T \mathbf{f}_1(\mathbf{x}) \quad (4.196)$$

wird als *Schaltfunktion* bezeichnet und für sie gilt entlang der optimalen Lösung  $\zeta^*(t) = \zeta(t)|_{\mathbf{x}=\mathbf{x}^*(t), \boldsymbol{\lambda}=\boldsymbol{\lambda}^*(t)}$ . Wenn  $\zeta^*(t) = 0$  auf einem endlichen Zeitintervall  $I_s \subseteq [t_0, t_1]$  gilt, so liefert die Minimierungsbedingung (4.140) keine Aussage für die optimale Stellgröße  $u^*(t) \forall t \in I_s$ . Man spricht in diesem Fall von einem *singulären Pfad* (englisch: *singular arc*) und es gilt folglich auch

$$\frac{d^k}{dt^k}(\zeta^*(t)) = 0 \quad \forall k \in \mathbb{N}, t \in I_s. \quad (4.197)$$

Entlang des singulären Pfades ist die Minimierungsbedingung (4.140) für alle zulässigen Stellgrößen erfüllt, d. h. es gilt nicht nur  $\frac{\partial H}{\partial u} = 0$ , sondern auch  $\frac{\partial^2 H}{\partial u^2} = 0$ . Um dennoch eine optimale Stellgröße  $u^*(t)$  ermitteln zu können, sucht man die *kleinste positive natürliche Zahl  $\bar{k}$*  so, dass gilt

$$\frac{\partial}{\partial u} \frac{d^{\bar{k}}}{dt^{\bar{k}}}(\zeta^*(t)) \neq 0. \quad (4.198)$$

Man kann zeigen, dass  $\bar{k}$  eine *gerade Zahl* sein muss und nennt  $p = \bar{k}/2$  die *Ordnung des singulären Pfades*. Entlang eines singulären Pfades müssen die Zustandsgrößen  $\mathbf{x}^*(t)$  und die adjungierten Zustände  $\boldsymbol{\lambda}^*(t)$  auf einer Mannigfaltigkeit definiert durch die Gleichungen

$$\frac{d^k}{dt^k}(\zeta^*(t)) = 0 \quad \forall k = 0, \dots, 2p - 1 \quad (4.199)$$

zu liegen kommen. Ähnlich der Legendre-Clebsch Bedingung (4.96) muss entlang eines singulären Pfades für alle Zeiten  $t \in I_s$  die sogenannte *generalisierte Legendre-Clebsch Bedingung*

$$(-1)^p \frac{\partial}{\partial u} \frac{d^{2p}}{dt^{2p}} (\zeta^*(t)) \geq 0 \quad (4.200)$$

erfüllt sein, vgl. [4.12].

*Beispiel 4.10.* Gesucht ist das Minimum des Kostenfunktional

$$J(u) = \frac{1}{2} \int_0^2 x_1^2(t) dt \quad (4.201)$$

für das dynamische System

$$\dot{x}_1 = x_2 + u \quad x_1(0) = 1 \quad x_1(2) = 0 \quad (4.202a)$$

$$\dot{x}_2 = -u \quad x_2(0) = 1 \quad x_2(2) = 0 \quad (4.202b)$$

unter Berücksichtigung der Stellgrößenbeschränkung  $-10 \leq u(t) \leq 10$  für alle  $t \in [0, 2]$ . Die Hamiltonfunktion  $H$  für dieses Beispiel lautet (mit  $\bar{\lambda}_{n+1} = \bar{\lambda}_3 = 1$ )

$$H(\mathbf{x}, u, \boldsymbol{\lambda}) = \lambda_1(x_2 + u) - \lambda_2 u + \frac{1}{2} x_1^2 \quad (4.203)$$

und die adjungierten Zustände  $\boldsymbol{\lambda}$  erfüllen gemäß (4.139) die Gleichungen

$$\frac{d}{dt} \lambda_1^* = - \left( \frac{\partial}{\partial x_1} H \right) (\mathbf{x}^*, u^*, \boldsymbol{\lambda}^*) = -x_1^* \quad (4.204a)$$

$$\frac{d}{dt} \lambda_2^* = - \left( \frac{\partial}{\partial x_2} H \right) (\mathbf{x}^*, u^*, \boldsymbol{\lambda}^*) = -\lambda_1^* . \quad (4.204b)$$

Die optimale Lösung  $u^*$  mit  $-10 \leq u^* \leq 10$  muss der Ungleichung (4.140)

$$H(\mathbf{x}^*(t), v, \boldsymbol{\lambda}^*(t)) \geq H(\mathbf{x}^*(t), u^*(t), \boldsymbol{\lambda}^*(t)), \quad \forall v \in [-10, 10] \quad (4.205)$$

genügen. Damit folgt zunächst für die optimale Stellgröße unter Berücksichtigung der Stellgrößenbeschränkung

$$u^*(t) = \begin{cases} 10 & \text{für } \lambda_1^* < \lambda_2^* \\ -10 & \text{für } \lambda_1^* > \lambda_2^* . \end{cases} \quad (4.206)$$

Für  $\lambda_1^* = \lambda_2^*$  tritt ein *singulärer Pfad* auf. Eine Auswertung von (4.198)

$$\left( \frac{\partial}{\partial u} H \right) (\mathbf{x}^*, u, \boldsymbol{\lambda}^*) = \lambda_1^* - \lambda_2^* = 0 \quad (4.207a)$$

$$\left( \frac{d}{dt} \frac{\partial}{\partial u} H \right) (\mathbf{x}^*, u, \boldsymbol{\lambda}^*) = \frac{d}{dt} \lambda_1^* - \frac{d}{dt} \lambda_2^* = -x_1^* + \lambda_1^* = 0 \quad (4.207b)$$

$$\left(\frac{d^2}{dt^2} \frac{\partial}{\partial u} H\right)(\mathbf{x}^*, u, \boldsymbol{\lambda}^*) = -x_2^* - u^* - x_1^* = 0 \quad (4.207c)$$

$$\left(\frac{\partial}{\partial u} \frac{d^2}{dt^2} \frac{\partial}{\partial u} H\right)(\mathbf{x}^*, u, \boldsymbol{\lambda}^*) = -1 \neq 0 \quad (4.207d)$$

liefert die Ordnung  $p = 1$  des singulären Pfades. Aus (4.207c) folgt

$$u^* = -x_2^* - x_1^* \quad (4.208)$$

und aus (4.207d) folgt

$$(-1)^1 \left(\frac{\partial}{\partial u} \frac{d^2}{dt^2} \frac{\partial}{\partial u} H\right)(\mathbf{x}^*, u, \boldsymbol{\lambda}^*) = 1 > 0. \quad (4.209)$$

Dies zeigt, dass die generalisierte Legendre-Clebsch Bedingung (4.200) hier erfüllt ist. Gemäß (4.141b) muss die Hamiltonfunktion  $H(\mathbf{x}^*(t), u^*(t), \boldsymbol{\lambda}^*(t))$  im gesamten Zeitintervall konstant sein, d. h.

$$\lambda_1^* x_2^* + \frac{1}{2}(x_1^*)^2 + (\lambda_1^* - \lambda_2^*)u^* = C = \text{konst.} \quad (4.210)$$

Entlang des singulären Pfades müssen  $\mathbf{x}^*(t)$  und  $\boldsymbol{\lambda}^*(t)$  auf der durch (4.207a) und (4.207b) definierten Mannigfaltigkeit  $\lambda_1^* = \lambda_2^* = x_1^*$  liegen (siehe auch (4.199)), weshalb sich für diesen Fall (4.210) zu

$$x_1^* x_2^* + \frac{1}{2}(x_1^*)^2 = C \quad (4.211)$$

vereinfacht.

**Aufgabe 4.10.** Plausibilisieren Sie, dass die optimale Stellgröße durch

$$u^*(t) = \begin{cases} 10 & \text{für } 0 \leq t \leq 0.299 \\ -x_2^* - x_1^* & \text{für } 0.299 < t < 1.927 \\ -10 & \text{für } 1.927 \leq t \leq 2 \end{cases} \quad (4.212)$$

gegeben ist.

**Aufgabe 4.11.** Im unebenen Gelände soll in direkter Linie eine Straße vom Ort  $y = y_0$  zum Ort  $y = y_1$  gebaut werden. Wie im linken Teil der Abbildung 4.14 skizziert, ist das Geländeprofil entlang der Trasse mit  $g(y)$  gegeben. Das Profil der zu errichtenden Straße wird mit  $h(y)$  bezeichnet. Die maximal zulässige Steigung der Straße sei  $s$ , d. h.  $|dh/dy| \leq s$ . Wird das Gelände für den Straßenbau aufgeschüttet, so ergeben sich Kosten in Höhe von  $k^+ > 0$  je aufgeschütteter Höheneinheit je Längeneinheit  $y$ . Wird das Gelände für den Straßenbau abgegraben, so ergeben sich Kosten in Höhe von  $k^- > 0$  je abgegrabener Höheneinheit je Längeneinheit  $y$ . Brücken oder Tunnels

sollen nicht gebaut werden.

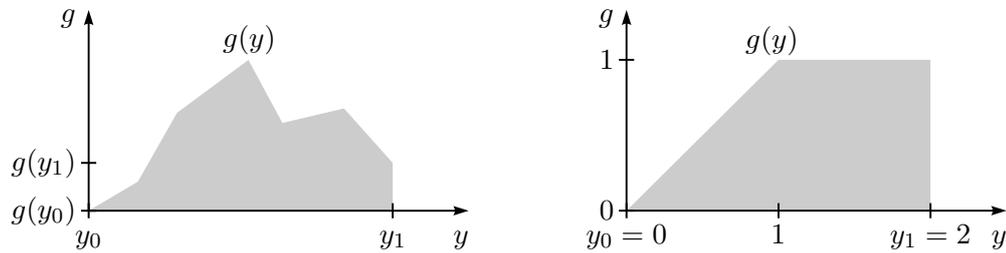


Abbildung 4.14: Geländeprofile.

Entlang von welchem optimalem Profil  $h^*(y)$  muss die Straße gebaut werden, um die Baukosten zu minimieren. Geben Sie zunächst allgemein die Optimalitätsbedingungen für diese Optimierungsaufgabe an. Berechnen Sie dann das optimale Profil  $h^*(y)$  für das im rechten Teil der Abbildung 4.14 skizzierte Geländeprofil  $g(y)$  und die Parameterwerte  $s = 1/2$  und  $k^+ = k^- = 1$ .

*Lösung von Aufgabe 4.11.* Unter Verwendung der Sprungfunktion

$$\sigma(t) = \begin{cases} 1 & \text{falls } t > 0 \\ 0 & \text{falls } t \leq 0 \end{cases} \quad (4.213)$$

ergeben sich die Gesamtkosten für den Straßenbau in der Form

$$\int_{y_0}^{y_1} (h - g) (k^+ \sigma(h - g) - k^- \sigma(g - h)) dy . \quad (4.214)$$

Dementsprechend muss die Optimierungsaufgabe

$$\min_{u(\cdot)} \int_{y_0}^{y_1} (h - g) (k^+ \sigma(h - g) - k^- \sigma(g - h)) dy \quad (4.215a)$$

$$\text{u.B.v. } \frac{dh}{dy} = u \quad (4.215b)$$

$$-s \leq u \leq s \quad (4.215c)$$

mit  $u(y) \in \hat{C}[y_0, y_1]$  gelöst werden. Dazu kann das Minimumsprinzip von Pontryagin herangezogen werden. Da  $g$  von  $y$  abhängt, wird der erweiterte Zustand  $\mathbf{x} = [h \quad y]^T$  und das erweiterte dynamische System

$$\frac{d\mathbf{x}}{dy} = \begin{bmatrix} u \\ 1 \end{bmatrix} \quad (4.216)$$

mit dem Anfangszustand  $x_2(y_0) = y_0$  verwendet. Mit  $\bar{\lambda}_{n+1} = 1$  (normaler Fall) folgt die Hamiltonfunktion

$$H(\mathbf{x}, u, \boldsymbol{\lambda}) = (h - g)(k^+ \sigma(h - g) - k^- \sigma(g - h)) + \boldsymbol{\lambda}^T \begin{bmatrix} u \\ 1 \end{bmatrix}. \quad (4.217)$$

Die optimale Lösung  $u^*(y)$ , die zugehörige Zustandstrajektorie  $\mathbf{x}^*(y)$  und die zugehörige Kozustandstrajektorie  $\boldsymbol{\lambda}^*(y)$  müssen (4.215c), (4.216), die Bedingung

$$H(\mathbf{x}^*(y), v, \boldsymbol{\lambda}^*(y)) \geq H(\mathbf{x}^*(y), u^*(y), \boldsymbol{\lambda}^*(y)) \quad \forall v \in [-s, s] \quad (4.218)$$

und die Differenzialgleichung

$$\frac{d\boldsymbol{\lambda}^*}{dy} = - \left( \frac{\partial}{\partial \mathbf{x}} H \right)^T (\mathbf{x}^*, u^*, \boldsymbol{\lambda}^*) = \begin{bmatrix} -1 \\ \frac{dg}{dy} \end{bmatrix} (k^+ \sigma(h - g) - k^- \sigma(g - h)) \quad (4.219)$$

erfüllen. Als Randbedingungen treten zusätzlich die Transversalitätsbedingungen

$$\lambda_1^*(y_0) = 0, \quad \lambda_1^*(y_1) = 0 \quad (4.220)$$

auf, da weder  $h(y_0)$  noch  $h(y_1)$  beschränkt sind (vgl. (4.165)).

**Bemerkung 4.3.** Würden  $h(y_0) = g(y_0)$  und  $h(y_1) = g(y_1)$  als weitere Bedingungen in (4.215) auftreten, so könnten keine Randbedingungen für  $\lambda_1^*(y_0)$  und  $\lambda_1^*(y_1)$  formuliert werden.

Für die Schaltfunktion ergibt sich gemäß (4.196)

$$\zeta(y) = \left( \frac{\partial}{\partial u} H \right) (\mathbf{x}, u, \boldsymbol{\lambda}) = \lambda_1. \quad (4.221)$$

Folglich gilt  $u^* = -s$  für  $\lambda_1^* > 0$  und  $u^* = s$  für  $\lambda_1^* < 0$ . Im Fall  $\lambda_1^* = 0$  tritt ein *singulärer Pfad* auf. Eine Auswertung von (4.197) liefert für den singulären Pfad

$$\frac{d\zeta^*(y)}{dy} = \frac{d\lambda_1^*}{dy} = k^- \sigma(g - h^*) - k^+ \sigma(h^* - g) = 0, \quad (4.222)$$

d. h. am singulären Pfad muss  $h^*(y) = g(y)$  und folglich  $u^* = \frac{dg}{dy}$  gelten. Die optimale Lösung  $u^*(y)$  kann daher in der Form

$$u^* = \begin{cases} -s & \text{falls } \lambda_1^* > 0 \\ s & \text{falls } \lambda_1^* < 0 \\ \frac{dg}{dy} & \text{falls } \lambda_1^* = 0 \end{cases} \quad (4.223)$$

zusammengefasst werden. Das optimale Profil  $h^*(y)$  der Straße kann schließlich durch einfache Integration von (4.215b) berechnet werden.

Für das im rechten Teil der Abbildung 4.14 angegebene Geländeprofil  $g(y)$  und die Parameterwerte  $s = 1/2$  und  $k^+ = k^- = 1$  wird  $\{s, 0\}$  als mögliche Schaltsequenz für die optimale Eingangstrajektorie  $u^*(y)$  verwendet. Diese Schaltsequenz ist naheliegend, da das gegebene Geländeprofil  $g(y)$  zunächst steiler ansteigt als es für die Straße zulässig ist und danach mit Steigung 0 fortsetzt, welche auch für den Straßenbau geeignet ist. Aus dieser Wahl der Schaltsequenz folgen

$$u^*(y) = \begin{cases} \frac{1}{2} & \text{falls } y \in [0, b) \\ 0 & \text{falls } y \in [b, 2] \end{cases}, \quad h^*(y) = \begin{cases} h_0 + \frac{y}{2} & \text{falls } y \in [0, b) \\ 1 & \text{falls } y \in [b, 2] \end{cases} \quad (4.224)$$

mit den noch zu bestimmenden Konstanten  $h_0$  und  $b$  (siehe Abbildung 4.15). Integration der ersten Zeile von (4.219) liefert unter Berücksichtigung von (4.220)

$$\lambda_1^*(y) = \begin{cases} -y & \text{falls } y \in [0, a) \\ y - 2a & \text{falls } y \in [a, b) \\ 0 & \text{falls } y \in [b, 2] \end{cases} \quad (4.225)$$

mit der noch zu bestimmenden Konstante  $a$ . Damit  $\lambda_1^*(y)$  am Punkt  $y = b$  stetig ist, muss

$$b - 2a = 0 \quad (4.226a)$$

gelten. Weiters müssen die Bedingungen

$$h^*(a) = h_0 + \frac{a}{2} = g(a) = a, \quad h^*(b) = h_0 + \frac{b}{2} = g(b) = 1 \quad (4.226b)$$

erfüllt sein (siehe Abbildung 4.15). Aus (4.226) folgen unmittelbar

$$h_0 = \frac{1}{3}, \quad a = \frac{2}{3}, \quad b = \frac{4}{3} \quad (4.227)$$

und damit das in Abbildung 4.15 rot dargestellte optimale Höhenprofil  $h^*(y)$  der Straße. Dieses Höhenprofil genügt den Optimalitätsbedingungen gemäß Satz 4.10 und der Bereich  $[b, y_1]$  stellt einen singulären Pfad dar, da in diesem Bereich (4.197) erfüllt ist. Die Schaltsequenz  $\{s, 0\}$  ist also optimal.

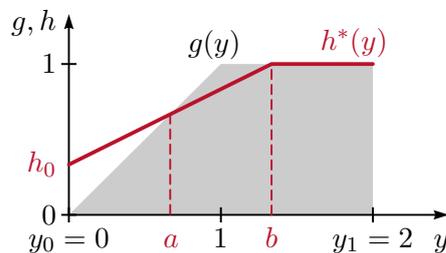


Abbildung 4.15: Optimales Höhenprofil der Straße.

## 4.3 Literatur

- [4.1] B. C. Chachuat, „Nonlinear and Dynamic Optimization: From Theory to Practice“, abrufbar unter <http://infoscience.epfl.ch/record/111939>, Laboratoire d'Automatique, École Polytechnique Fédérale de Lausanne, 2007, (besucht am 28.09.2017).
- [4.2] H.R. Schwarz und N. Köckler, *Numerische Mathematik*, 8. Aufl. Wiesbaden: Vieweg+Teubner, 2011.
- [4.3] M. Hermann, *Numerik gewöhnlicher Differentialgleichungen: Anfangs- und Randwertprobleme*. München: Oldenbourg, 2004.
- [4.4] J. Stoer und R. Bulirsch, *Introduction to Numerical Analysis*, 3. Aufl., Ser. Texts in Applied Mathematics 12. New York, Berlin: Springer, 2002.
- [4.5] C. Lanczos, *The Variational Principles of Mechanics*, 4. Aufl. New York: Dover, 1970.
- [4.6] L. Meirovitch, *Methods of Analytical Dynamics*, Ser. Advanced Engineering Series. New York: McGraw-Hill, 1970.
- [4.7] H. A. Mang und G. Hofstetter, *Festigkeitslehre*, 4. Aufl. Wien, New York: Springer, 2013.
- [4.8] M. I. Kamien und N. L. Schwartz, *Dynamic Optimization: The Calculus of Variations and Optimal Control in Economics and Management*, 2. Aufl. Amsterdam: Elsevier, 1991.
- [4.9] J. Troutman, *Variational Calculus and Optimal Control: Optimization with Elementary Convexity*, 2. Aufl., Ser. Undergraduate Texts in Mathematics. New York: Springer, 1996.
- [4.10] J. Macki und A. Strauss, *Introduction to Optimal Control Theory*, Ser. Undergraduate Texts in Mathematics. New York: Springer, 1982.
- [4.11] L. Pontryagin, V. Boltyanskii, R. Gamkrelidze und E. Mishchenko, *The Mathematical Theory of Optimal Processes*. Pergamon Press, 1964.
- [4.12] A. E. Bryson, Jr. und Y.-C. Ho, *Applied Optimal Control: Optimization, Estimation, and Control*. John Wiley & Sons, 1975.
- [4.13] R. F. Hartl, S. P. Sethi und R. G. Vickson, „A Survey of the Maximum Principles for Optimal Control Problems with State Constraints“, *SIAM Review*, Jg. 37, Nr. 2, S. 181–218, 1995.
- [4.14] M. Papageorgiou, M. Leibold und M. Buss, *Optimierung: Statische, dynamische, stochastische Verfahren für die Anwendung*, 4. Aufl. Springer, 2015.
- [4.15] M. Athans und P. L. Falb, *Optimal Control: An Introduction to the Theory and Its Applications*. New York: McGraw-Hill, 1966.
- [4.16] B. van Brunt, *The Calculus of Variations*, Ser. Universitext. Springer, 2004.
- [4.17] O. Föllinger, *Optimale Regelung und Steuerung*, Ser. Methoden der Regelungs- und Automatisierungstechnik. R. Oldenbourg Verlag, 1994.

- 
- [4.18] D. S. Naidu, *Optimal Control Systems*, Ser. Electrical Engineering Series. CRC Press, 2003.
- [4.19] D. E. Kirk, *Optimal Control Theory: An Introduction*. Dover Publications, 2004.