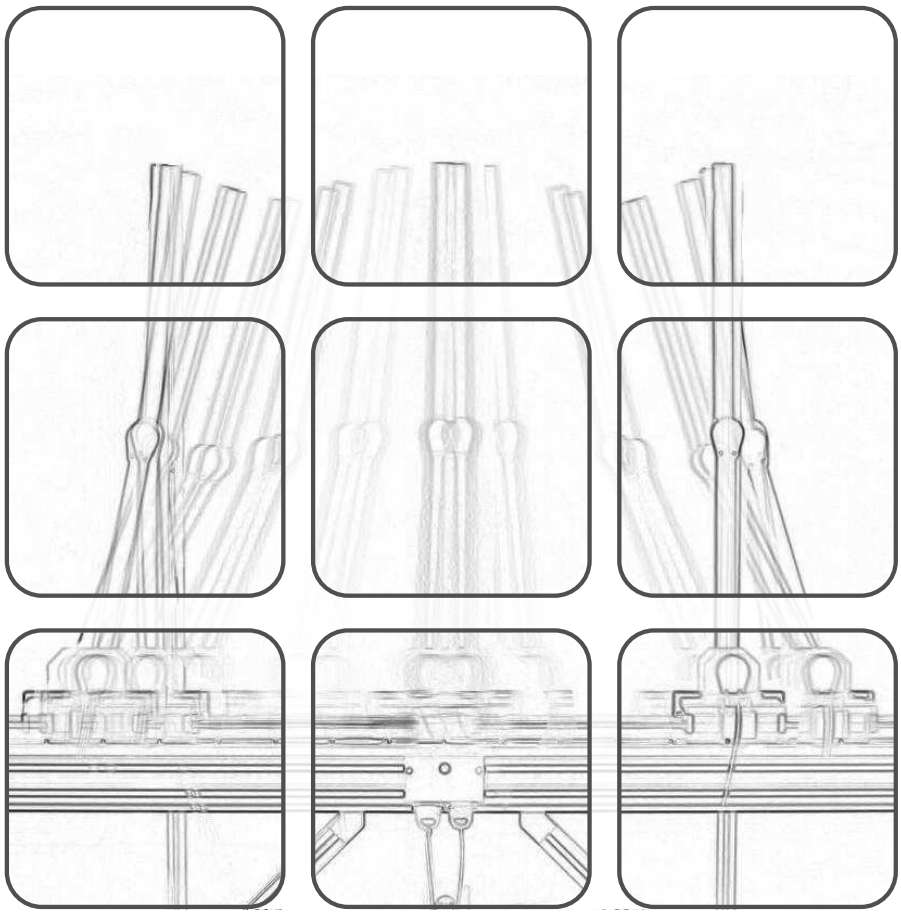
TECHNISCHE UNIVERSITÄT WIEN

AUTOMATION & CONTROL INSTITUTE
INSTITUT FÜR AUTOMATISIERUNGS- & REGELUNGSTECHNIK

Vorlesung und Übung
WS 2023/2024

Andreas STEINBÖCK

OPTIMIERUNG



Optimierung

Vorlesung und Übung
WS 2023/2024

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: https://www.acin.tuwien.ac.at

Inhaltsverzeichnis

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

$$\begin{aligned}
&\min_{\mathbf{x}\in\mathbb{R}^n} && f(\mathbf{x}) && && \text{Kostenfunktion} && (1.1a)\\
&\text{u.B.v.} && g_i(\mathbf{x}) = 0, && i = 1,\dots,p && \text{Gleichungsbeschränkungen} && (1.1b)\\
&&& h_i(\mathbf{x}) \le 0, && i = 1,\dots,q && \text{Ungleichungsbeschränkungen.} && (1.1c)
\end{aligned}$$

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) und (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,

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

wird als *zulässiges Gebiet* oder *zulässige Menge* (englisch: *admissible region* 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

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

angeben. Im Falle von unbeschränkten Problemen gilt $\mathcal{X} = \mathbb{R}^n$. $\mathcal{X}$ darf nicht die leere Menge sein, da sonst das Optimierungsproblem (1.3) keine Lösung besitzt. Eine weitere notwendige Bedingung für $\mathcal{X}$ kann aus den Gleichungsbeschränkungen (1.1b) abgeleitet werden, da sich durch die p 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 sein kann.

In der Literatur hat sich weitgehend die Formulierung als Minimierungsproblem (1.1) oder (1.3) durchgesetzt. Analog dazu kann ein Maximierungsproblem ebenfalls als Minimierungsproblem gemäß (1.3) geschrieben werden:

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

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.

Bei statischen Optimierungsproblemen werden häufig folgende Klassen unterschieden:

- *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 Rolle spielt. Ein Optimierungsproblem ist konvex, wenn die Kostenfunktion eine konvexe Funktion und das zulässige Gebiet eine konvexe Menge ist. Bei konvexen Optimierungsproblemen sind die notwendigen Optimalitätsbedingungen erster Ordnung gleichzeitig hinreichend für ein globales Optimum.
- *Integer-Programmierung*: Alle Optimierungsvariablen sind diskret.
- *Mixed-Integer-Programmierung*: Es treten kontinuierliche und diskrete Optimierungsvariablen auf.

1.1.2 Beispiele

Insbesondere die *lineare Programmierung* wird häufig bei ökonomischen Fragestellungen, wie Produktions-, Planungs- oder Investitionsproblemen, eingesetzt. Das folgende Beispiel zeigt eine einfache 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 gemäß der oben gegebenen Risikoeinstufung 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

$$x_3 = 1 - x_1 - x_2 \tag{1.4}$$

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

$$10.000(0.1x_1 + 0.07x_2 + 0.04(1 - x_1 - x_2)) \geq 600 \quad \Rightarrow \quad 6x_1 + 3x_2 \geq 2 \tag{1.5}$$

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

$$10.000(1 - x_1 - x_2) \geq 4.000 \quad \Rightarrow \quad x_1 + x_2 \leq 0.6. \tag{1.6}$$

Des Weiteren müssen $x_1 \geq 0$, $x_2 \geq 0$ und $x_3 \geq 0$ erfüllt sein. Mit diesen Beschränkungen und (1.6) ist auch sichergestellt, dass $x_1 \leq 1$, $x_2 \leq 1$ und $x_3 \leq 1$. Das Ziel ist die Minimierung des Anlagerisikos, was sich durch die Funktion

$$f(\mathbf{x}) = 4x_1 + 2x_2 + (1 - x_1 - x_2) = 1 + 3x_1 + x_2 \tag{1.7}$$

ausdrücken lässt.

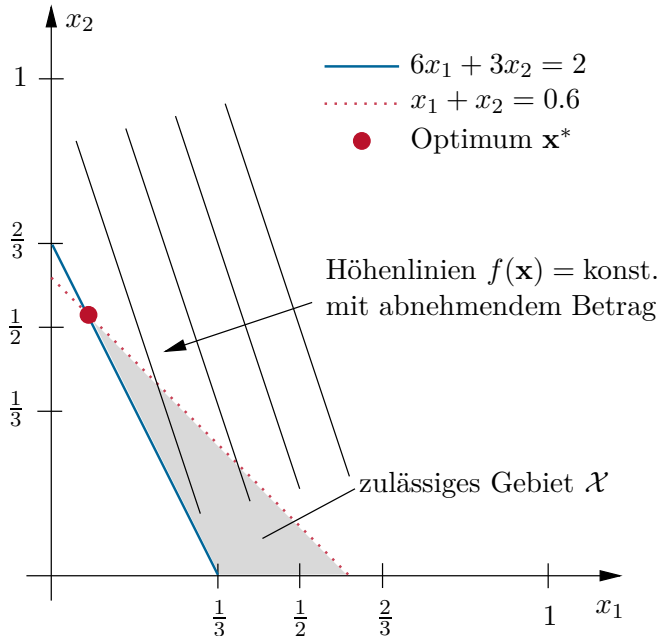


Abbildung 1.1: Grafische Lösung zur Portfolio-Optimierung.

Somit kann das statische Optimierungsproblem in der Form

$$\begin{aligned} \min_{\mathbf{x}\in\mathbb{R}^2} \quad & f(\mathbf{x}) = 1 + 3x_1 + x_2 && \text{(1.8a)} \\ \text{u.B.v.} \quad & 6x_1 + 3x_2 \geq 2 && \text{(1.8b)} \\ & x_1 + x_2 \leq 0.6 && \text{(1.8c)} \\ & x_1 \geq 0 && \text{(1.8d)} \\ & x_2 \geq 0 && \text{(1.8e)} \end{aligned}$$

geschrieben werden. Abbildung 1.1 stellt die einzelnen Beschränkungen sowie das zulässige Gebiet grafisch dar. Aus dem Verlauf der Höhenlinien $f(\mathbf{x}) =$ konst. der Kostenfunktion (1.8a) 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

$$x_1^* = \frac{1}{15}, \quad x_2^* = \frac{8}{15}, \quad x_3^* = \frac{6}{15}. \tag{1.9}$$

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

Beispiel 1.2. Betrachtet wird das (zunächst) unbeschränkte Problem

$$\min_{\mathbf{x}\in\mathbb{R}^2} \; f(\mathbf{x}) = (x_1 - 2)^2 + (x_2 - 1)^2 . \tag{1.10}$$

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

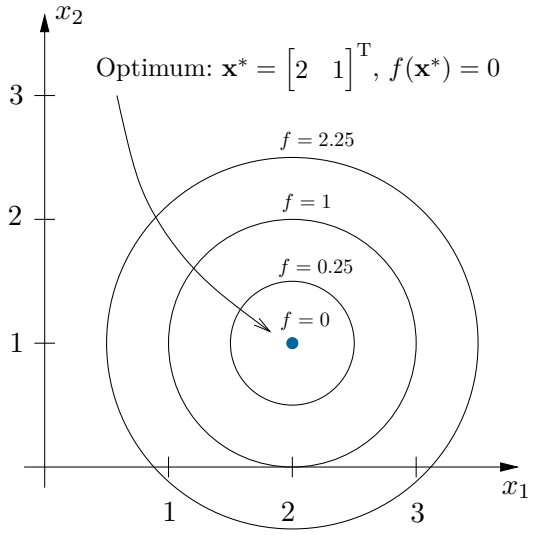


Abb. 1.2: Geometrische Darstellung des unbeschränkten Optimierungsproblems (1.10).

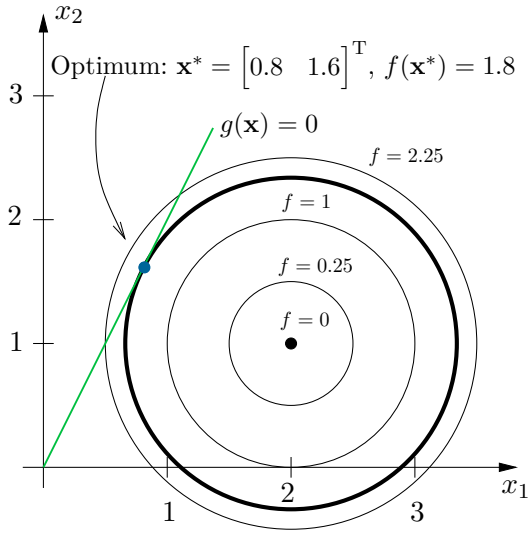


Abb. 1.3: Geometrische Darstellung des beschränkten Optimierungsproblems (1.10), (1.11).

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

$$g(\mathbf{x}) = x_2 - 2x_1 = 0\,. \tag{1.11}$$

Die Gleichungsbeschränkung entspricht einer algebraischen Zwangsbedingung, wodurch lediglich noch eine Optimierungsvariable frei wählbar ist. Geometrisch interpretiert bedeutet dies, dass eine mögliche Lösung auf jener Geraden liegen muss, die durch (1.11) definiert wird (siehe Abbildung 1.3). Die optimale Lösung liegt dabei auf dem tangentialen Berührpunkt der Geraden $g(\mathbf{x}) = 0$ mit der Höhenlinie $f(\mathbf{x}) = 1.8$.

Anstelle der Gleichungsbeschränkung (1.11) wird nun die Ungleichungsbeschränkung

$$h_1(\mathbf{x}) = x_1 + x_2 - 2 \leq 0 \tag{1.12}$$

betrachtet, wodurch sich die Menge der zulässigen Punkte $\mathbf{x} = [x_1 \quad x_2]^{\mathrm{T}}$ 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 $\mathbf{x}^* = [1.5 \quad 0.5]^{\mathrm{T}}$ 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 (1.12) soll eine weitere Ungleichung der Form

$$h_2(\mathbf{x}) = x_1^2 - x_2 \leq 0 \tag{1.13}$$

betrachtet werden, durch die sich die Menge der zulässigen Punkte weiter verkleinert (siehe Abbildung 1.5). Der optimale Punkt $\mathbf{x}^* = [1 \quad 1]^{\mathrm{T}}$ 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 (1.12) und (1.13) sind aktiv.

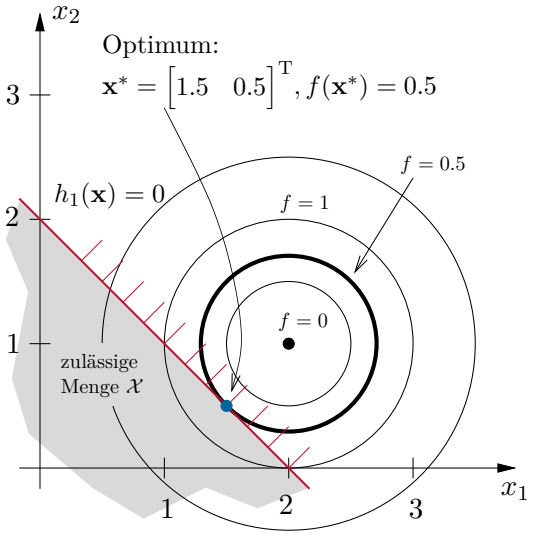


Abb. 1.4: Geometrische Darstellung des beschränkten Optimierungsproblems (1.10), (1.12).

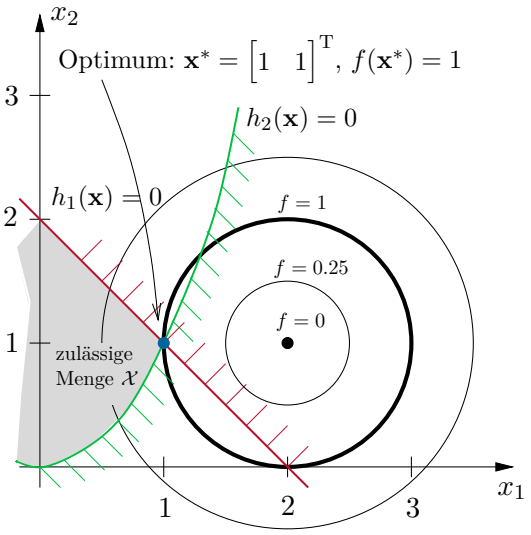


Abb. 1.5: Geometrische Darstellung des beschränkten Optimierungsproblems (1.10), (1.12), (1.13).

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

1.2 Dynamische Optimierungsprobleme

Bei den Problemstellungen der statischen Optimierung im vorangegangenen Abschnitt stellen die Optimierungsvariablen $\mathbf{x}$ Elemente aus einem finit-dimensionalen Raum, meist dem Euklidischen Raum $\mathbb{R}^n$, dar. Bei der dynamischen Optimierung hingegen wird in einem Raum von Funktionen einer unabhängigen Variablen nach einem Optimum gesucht. 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

$$\min_{\mathbf{u}(\cdot)} \quad J(\mathbf{u}) = \varphi(t_1, \mathbf{x}(t_1)) + \int_{t_0}^{t_1} l(t, \mathbf{x}(t), \mathbf{u}(t)) \, \mathrm{d}t \qquad \text{Kostenfunktional} \tag{1.14a}$$

$$\begin{aligned}
\text{u.B.v.} \quad & \dot{\mathbf{x}} = \mathbf{f}(t, \mathbf{x}, \mathbf{u}) && \text{Systemdynamik} && (1.14b)\\
& \mathbf{x}(t_0) = \mathbf{x}_0 && \text{Anfangsbedingungen} && (1.14c)\\
& \boldsymbol{\psi}(t_1, \mathbf{x}(t_1)) = \mathbf{0} && \text{Endbedingungen} && (1.14d)\\
& h_i(\mathbf{x}, \mathbf{u}) \leq 0, \quad i = 1, \ldots, q && \text{Ungleichungsbeschr.} && (1.14e)
\end{aligned}$$

Dabei stellt $\mathbf{u} \in \mathbb{R}^m$ die Eingangsgröße des nichtlinearen Systems (1.14b) mit dem Zustand $\mathbf{x} \in \mathbb{R}^n$ dar. Zusätzlich zu den Anfangsbedingungen (1.14c) sind häufig Endbedingungen der Form (1.14d) gegeben, um z. B. einen gewünschten Zustand $\mathbf{x}_\mathrm{f}$ zur Endzeit t_1 zu erreichen (also $\boldsymbol{\psi}(t_1, \mathbf{x}(t_1)) = \mathbf{x}(t_1) - \mathbf{x}_\mathrm{f}$). In der Praxis treten häufig Ungleichungsbeschränkungen (1.14e) auf, die z. B. die Begrenzung einer Stellgröße oder Sicherheitsschranken eines Zustandes darstellen können.

Die Aufgabe 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 (1.14b) mit den Anfangsbedingungen (1.14c) die Endbedingungen (1.14d) erfüllt, die Beschränkungen (1.14e) erfüllt werden und gleichzeitig das Kostenfunktional (1.14a) 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. Die folgenden Beispiele erläutern die Problem- und Aufgabenstellung der dynamischen Optimierung.

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

$$\min_{u(\cdot), t_1} \quad J(u) = \int_0^{t_1} 1 + c\,u^2 \;\mathrm{d}t\,, \tag{1.15a}$$

$$\text{u.B.v.} \quad \begin{bmatrix} 1 & \varepsilon\cos\theta \\ \cos\theta & 1 \end{bmatrix} \begin{bmatrix} \ddot{x} \\ \ddot{\theta} \end{bmatrix} = \begin{bmatrix} \varepsilon\dot{\theta}^2 \sin\theta + u \\ -\sin\theta \end{bmatrix}, \quad \varepsilon = m/(M+m) \tag{1.15b}$$

$$\mathbf{x}(0) = \begin{bmatrix} 0 & 0 & \pi & 0 \end{bmatrix}^\mathrm{T}, \quad \mathbf{x}(t_1) = \begin{bmatrix} 1 & 0 & \pi & 0 \end{bmatrix}^\mathrm{T}, \tag{1.15c}$$

$$-1 \leq u \leq 1\,. \tag{1.15d}$$

Die vereinfachten Bewegungsgleichungen (1.15b) für die Zustände $\mathbf{x} = [x \quad \dot{x} \quad \theta \quad \dot{\theta}]^\mathrm{T}$ sind normiert. Der Eingang u stellt die am Wagen angreifende Kraft dar und ist durch (1.15d) 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.

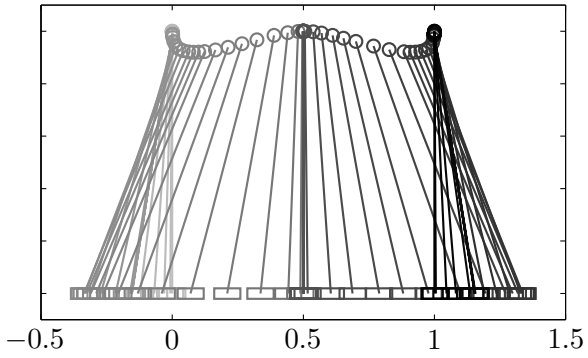


Abbildung 1.6: Momentaufnahmen beim Versetzens des inversen Pendels.

Das Kostenfunktional (1.15a) 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

$$J(u) = \int_0^{t_1} 1 \ \mathrm{d}t = t_1 \,. \tag{1.16}$$

Für $c > 0$ wird der Eingang u im Kostenfunktional (und somit der Stellgrößenaufwand) mitberücksichtigt.

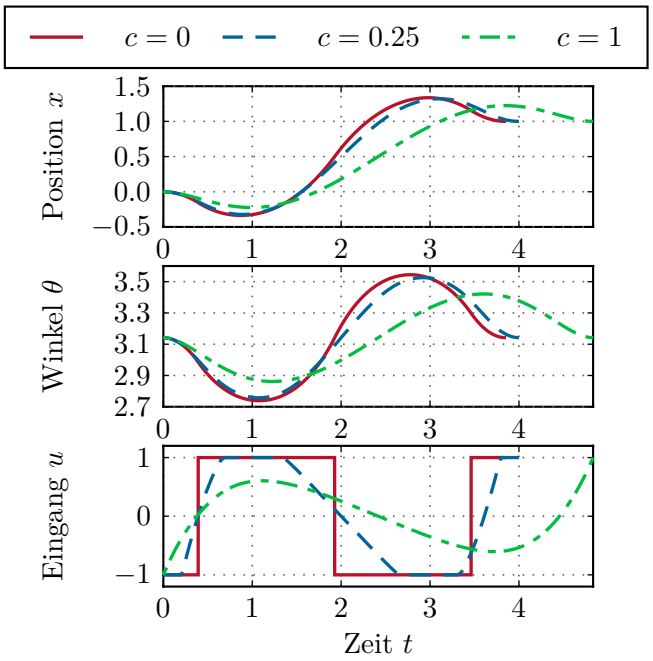


Abbildung 1.7: Optimale Trajektorien beim Versetzen des inversen Pendels.

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 sogenanntes Bang-Bang-Verhalten auf, während für $c > 0$ die Steueramplituden kleiner werden und die benötigte Zeit t_1 zunimmt.

Dieses Beispiel verdeutlicht, 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 (1.15a). Wenn nur die Stellgröße u gewichtet würde, d. h.

$$J(u) = \int_0^{t_1} u^2 \, \mathrm{d}t \,, \tag{1.17}$$

hätte das Optimierungsproblem keine Lösung, da das Versetzen des Pendels dann unendlich langsam, d. h. mit mit $t_1 \to \infty$, 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 Robert H. Goddard im Jahr 1919 aufgestellt und kann in der normierten Form

$$\min_{u(\cdot)} \quad -h(t_1) \tag{1.18a}$$

$$\text{u.B.v.} \quad \dot{h} = v, \quad \dot{v} = \frac{u - D(h,v)}{m} - \frac{1}{h^2} \,, \quad \dot{m} = -\frac{u}{c} \,, \tag{1.18b}$$

$$h(0) = 1, \quad v(0) = 0, \quad m(0) = 1, \quad m(t_1) = 0.6 \,, \tag{1.18c}$$

$$0 \le u \le 3.5 \tag{1.18d}$$

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

$$D(h,v) = D_0 \, v^2 \mathrm{e}^{[\beta(1-h)]} \tag{1.19}$$

von den Zuständen h und v ab. Die Randbedingungen in (1.18c) 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 (1.18d) 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 im Falle eines hohen Luftwiderstandes *optimaler* nicht permanent mit vollem Schub zu fliegen.

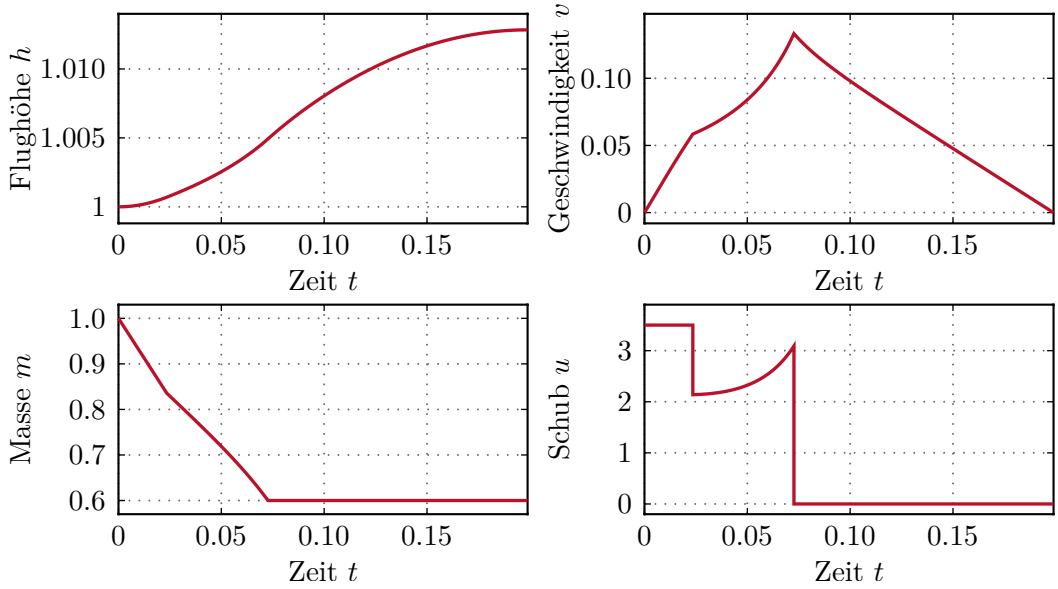


Abbildung 1.8: Trajektorien für die Goddard-Rakete.

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 optimieren will. Der Bildungsgrad B und das Kapital K eines durchschnittlichen Konsumenten lassen sich durch folgendes Modell beschreiben

$$\dot{B} = \overbrace{B^{\varepsilon} u_2 u_3}^{\text{Weiterbildung}} - \overbrace{\delta B}^{\text{Vergessen}}, \qquad B(0) = B_0 \tag{1.20a}$$

$$\dot{K} = \underbrace{iK}_{\text{Verzinsung}} + \underbrace{B u_2 g(u_3)}_{\text{Einkommen}} - \underbrace{u_1}_{\text{Konsum}}, \qquad K(0) = K_0. \tag{1.20b}$$

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

$$u_1 > 0, \quad 0 \le u_2 \le 1, \quad 0 \le u_3 < 1. \tag{1.21}$$

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

$$J(\mathbf{u}) = -K^{\kappa}(t_1) - \int_{t_0}^{t_1} U(t, u_1, u_2, B) \mathrm{e}^{-\rho t} \, \mathrm{d}t \,. \tag{1.22}$$

ausgedrückt ist. Die Nutzenfunktion

$$U(t, u_1, u_2, B) = \alpha_0 u_1^{\alpha} + \beta_0 (1 - u_2)^{\beta} + \gamma_0 t B^{\gamma} \tag{1.23}$$

gewichtet dabei den Konsum u_1, die Freizeit $1 - u_2$ und den Bildungsgrad B, während der Endwert $-K^{\kappa}(t_1)$ in (1.22) 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 (1.20b) ist durch die Parabel $g(u_3) = 1 - (1 - a)u_3 - au_3^2$ 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.

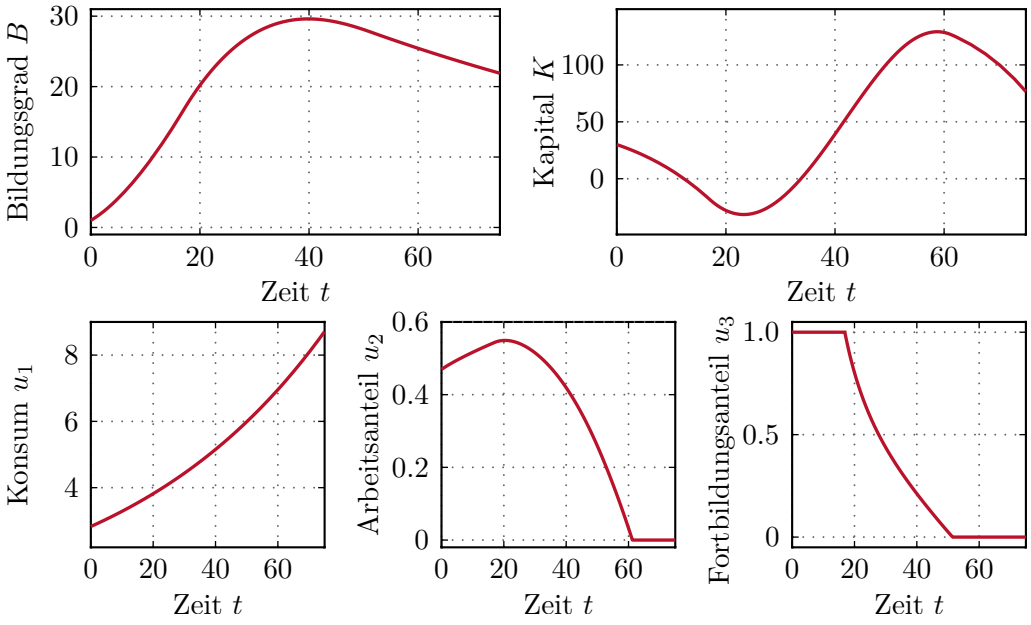


Abbildung 1.9: Optimale Trajektorien für das Konsumentenverhalten.

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 und Minimum

Definition 1.1 (Infimum)**.** 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 $\alpha = \inf \mathcal{Y}$ 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 $x < \bar{\alpha}$.

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

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 für eine Norm $||\cdot||$ ein $\varepsilon > 0$ so existiert, dass gilt $f(\mathbf{x}^*) \leq f(\mathbf{x})$ für alle $\mathbf{x} \in \{\mathbf{x} \in \mathcal{X} \,|\, ||\mathbf{x} - \mathbf{x}^*|| < \varepsilon\}$,

(b) ein *striktes lokales Minimum*, falls für eine Norm $||\cdot||$ ein $\varepsilon > 0$ so existiert, dass gilt $f(\mathbf{x}^*) < f(\mathbf{x})$ für alle $\mathbf{x} \in \{\mathbf{x} \in \mathcal{X} \,|\, ||\mathbf{x} - \mathbf{x}^*|| < \varepsilon,\ \mathbf{x} \neq \mathbf{x}^*\}$,

(c) ein *globales Minimum*, falls $f(\mathbf{x}^*) \leq f(\mathbf{x})$ für alle $\mathbf{x} \in \mathcal{X}$, und

(d) ein *striktes globales Minimum*, falls $f(\mathbf{x}^*) < f(\mathbf{x})$ für alle $\mathbf{x} \in \mathcal{X} \backslash \{\mathbf{x}^*\}$.

Abbildung 1.10 zeigt unterschiedliche Arten von Minima für eine Funktion in einer Optimierungsvariablen.

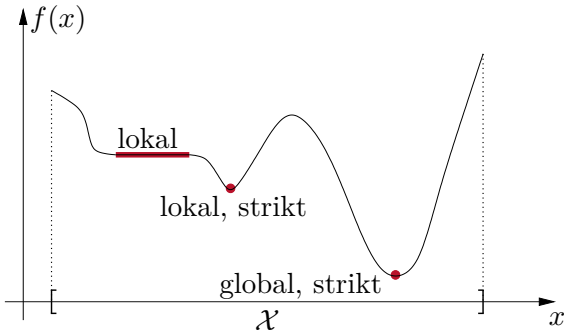


Abbildung 1.10: Verschiedene Minima einer Funktion $f(x)$ mit $x \in \mathcal{X} \subset \mathbb{R}$.

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

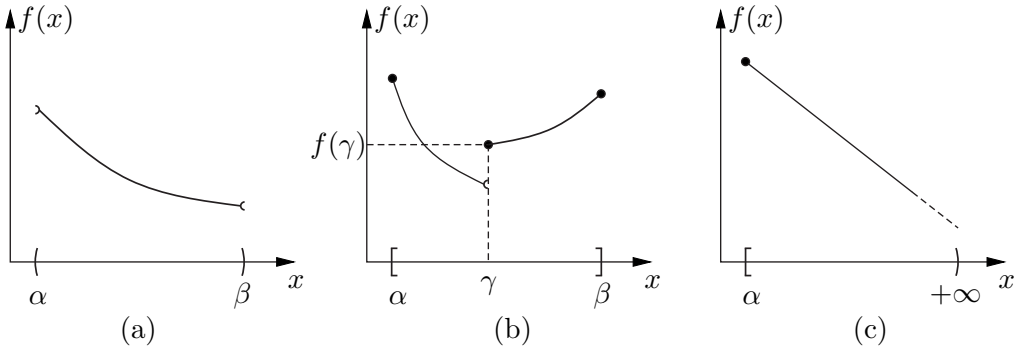


Abbildung 1.11: Nichtexistenz von Minima.

Die Menge aller Minima wird oftmals in der Form

$$\mathcal{G} = \arg\min\{f(\mathbf{x}) \,|\, \mathbf{x} \in \mathcal{X}\} \tag{1.24}$$

angeschrieben, wobei $\mathcal{G}$ 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 $\bar{\mathbf{x}} = \arg\min\{f(\mathbf{x}) \,|\, \mathbf{x} \in \mathcal{X}\}$ 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

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 $\mathcal{X}$ nicht abgeschlossen ist und somit $\beta \notin \mathcal{X}$, existiert in diesem Fall kein Minimum. In Abbildung 1.11(b) ist der linksseitige Grenzwert $\lim_{x\to\gamma^-} f(x)$ 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 $\mathcal{X} = \{x \in \mathbb{R} \,|\, x \geq \alpha\}$ nach unten hin nicht beschränkt ist.

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

Satz 1.1 (Satz von Weierstrass)**.** *Es sei $\mathcal{X}$ eine* nichtleere *und* kompakte (abgeschlossene und beschränkte) *Menge und $f : \mathcal{X} \to \mathbb{R}$* 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 betont, dass Satz 1.1 eine hinreichende jedoch nicht notwendige 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 $\mathcal{X} = (-1, 1)$ 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 $f : \mathcal{X} \to \mathbb{R}$ eine stetig differenzierbare Funktion, d. h. $f \in C^1$. Dann bezeichnet

$$(\nabla f)(\mathbf{x}) = \left(\frac{\partial f}{\partial \mathbf{x}}\right)^{\mathrm{T}} = \begin{bmatrix} \frac{\partial f}{\partial x_1} \\ \vdots \\ \frac{\partial f}{\partial x_n} \end{bmatrix} \tag{1.25}$$

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

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

$$\left(\nabla^2 f\right)(\mathbf{x}) = \begin{bmatrix} \frac{\partial^2 f}{\partial x_1^2} & \cdots & \frac{\partial^2 f}{\partial x_1 \partial x_n} \\ \vdots & & \vdots \\ \frac{\partial^2 f}{\partial x_n \partial x_1} & \cdots & \frac{\partial^2 f}{\partial x_n^2} \end{bmatrix} \tag{1.26}$$

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

Im Falle von Funktionen $f(x)$ mit nur einem skalaren Argument wird die ∇-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 $(\nabla^2 f)(\mathbf{x}) = (\nabla^2 f)^{\mathrm{T}}(\mathbf{x})$, 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 *Eigenwerte* λ_i *sind*	*(c) für alle* n *Hauptminoren* D_i *gilt*
positiv semi-definit:	$\mathbf{p}^{\mathrm{T}}\mathbf{A}\,\mathbf{p} \geq 0$	≥ 0	-
positiv definit:	$\mathbf{p}^{\mathrm{T}}\mathbf{A}\,\mathbf{p} > 0$	> 0	$D_i > 0$
negativ semi-definit:	$\mathbf{p}^{\mathrm{T}}\mathbf{A}\,\mathbf{p} \leq 0$	≤ 0	-
negativ definit:	$\mathbf{p}^{\mathrm{T}}\mathbf{A}\,\mathbf{p} < 0$	< 0	$(-1)^{i+1}D_i < 0$

Die Eigenwerte λ_i, $i = 1, \ldots, n$ 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 $\mathbf{A}$,

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

wobei a_{ij} das Element in der i-ten Zeile und j-ten Spalte von $\mathbf{A}$ bezeichnet, d. h. $\mathbf{A} = [a_{ij}]_{i,j=1,\ldots,n}$. Um die Definitheit einer symmetrischen Matrix $\mathbf{A}$ zu bestimmen, muss lediglich eine der drei Bedingungen (a)–(c) in Satz 1.2 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.

Die positive Definitheit einer symmetrischen Matrix $\mathbf{A}$ kann alternativ zu den Bedingungen (a)–(c) aus Satz 1.2 auch mit Hilfe der *Cholesky-Faktorisierung* überprüft werden. Gemäß dieser Methode gilt, dass eine symmetrische Matrix $\mathbf{A}$ genau dann positiv definit ist, wenn sie sich in der Form $\mathbf{A} = \mathbf{G}\mathbf{G}^{\mathrm{T}}$ faktorisieren lässt, wobei $\mathbf{G}$ eine untere Dreiecksmatrix mit positiven Diagonaleinträgen ist.

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* α*,* $0 \leq \alpha \leq 1$ *so, dass gilt*

$$f(\mathbf{x}_2) = f(\mathbf{x}_1) + (\mathbf{x}_2 - \mathbf{x}_1)^{\mathrm{T}}(\nabla f)(\mathbf{x}_1 + \alpha(\mathbf{x}_2 - \mathbf{x}_1)) \; . \tag{1.27}$$

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

$$f(\mathbf{x}_2) = f(\mathbf{x}_1) + (\mathbf{x}_2 - \mathbf{x}_1)^{\mathrm{T}}(\nabla f)(\mathbf{x}_1) + \frac{1}{2}(\mathbf{x}_2 - \mathbf{x}_1)^{\mathrm{T}}\Big(\nabla^2 f\Big)(\mathbf{x}_1 + \alpha(\mathbf{x}_2 - \mathbf{x}_1))(\mathbf{x}_2 - \mathbf{x}_1) \tag{1.28}$$

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, das eine effiziente und zugleich möglichst genaue Berechnung der Ableitungen erreicht wird. Mehr Informationen über das algorithmische Differenzieren ist z. B. unter `http://www.autodiff.org` oder in [1.7] zu finden.

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 näherungsweisen 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

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

Tabelle 1.1: Differenzenquotienten.

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 $h \to 0$ unbeschränkt anwachsen.

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

$$f(\mathbf{x}+h\mathbf{e}_i) = f(\mathbf{x}) + h\left.\frac{\partial f}{\partial x_i}\right|_{\mathbf{x}} + \left.\frac{h^2}{2}\frac{\partial^2 f}{\partial x_i^2}\right|_{\mathbf{x}+\alpha h\mathbf{e}_i} \tag{1.29}$$

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

$$\left.\frac{\partial f}{\partial x_i}\right|_{\mathbf{x}} = \frac{f(\mathbf{x}+h\mathbf{e}_i)-f(\mathbf{x})}{h} \underbrace{- \left.\frac{h}{2}\frac{\partial^2 f}{\partial x_i^2}\right|_{\mathbf{x}+\alpha h\mathbf{e}_i}}_{\text{Abschneidefehler, } \mathcal{O}(h)}. \tag{1.30}$$

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

$$\begin{aligned}&\frac{f(\mathbf{x}+h\mathbf{e}_i)(1+e_r)-f(\mathbf{x})(1-e_r)}{h}\\&\quad=\frac{f(\mathbf{x}+h\mathbf{e}_i)-f(\mathbf{x})}{h}+\underbrace{\frac{(f(\mathbf{x}+h\mathbf{e}_i)+f(\mathbf{x}))e_r}{h}}_{\text{Rundungsfehler, } \mathcal{O}(e_r h^{-1})}\end{aligned} \tag{1.31}$$

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

$$\begin{aligned}&\underbrace{\frac{f(\mathbf{x}+h\mathbf{e}_i)-f(\mathbf{x})}{h}+\frac{(f(\mathbf{x}+h\mathbf{e}_i)+f(\mathbf{x}))e_r}{h}}_{\text{Berechneter Wert}}-\underbrace{\left.\frac{\partial f}{\partial x_i}\right|_{\mathbf{x}}}_{\text{Exakter Wert}}\\&\quad=\frac{h}{2}\left.\frac{\partial^2 f}{\partial x_i^2}\right|_{\mathbf{x}+\alpha h\mathbf{e}_i}+\frac{(f(\mathbf{x}+h\mathbf{e}_i)+f(\mathbf{x}))e_r}{h}\,.\end{aligned} \tag{1.32}$$

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

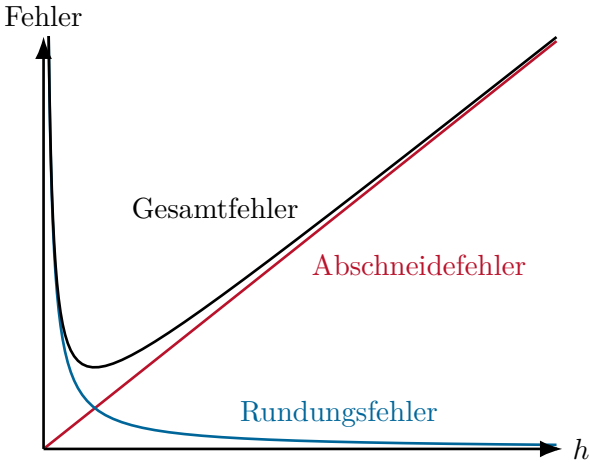


Abbildung 1.12: Absolutwerte der Fehler bei der Approximation des Gradienten durch den Vorwärtsdifferenzenquotienten in Abhängigkeit der Schrittweite h.

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

$$\frac{f(\mathbf{x}+h\mathbf{e}_i)-f(\mathbf{x})}{h}\,. \tag{1.33}$$

Bei der Differenzbildung im Zähler dieses Ausdrucks kann es zu erheblichen Genauigkeits-

verlusten durch *Auslöschungsfehler* kommen [1.10]. Beim Rechnen mit Gleitkommaarithmetik müssen alle Eingangsgrößen, Zwischenergebnisse der auftretenden Elementaroperationen und Endergebnisse auf die Zahl der Mantissenstellen gerundet werden. Als *Auslöschung* bezeichnet man das Annulieren 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 $h \to 0$ zur Unbrauchbarkeit des Ergebnisses führt.

Beispiel 1.7 (Auslöschungsfehler). Man betrachte die Rechnung

$$0.123\,456 - 0.123\,455 = 0.000\,001\ . \tag{1.34}$$

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 näherungsweisen 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 näherungsweisen Berechnung von $(\nabla f)(\mathbf{x})$ steigt der Berechnungsaufwand mit der Ordnung $\mathcal{O}(n)$. Bei der näherungsweisen Berechnung von $(\nabla^2 f)(\mathbf{x})$ steigt der Berechnungsaufwand mit der Ordnung $\mathcal{O}(n^2)$.

Ableitungsberechnung mittels komplexer Funktionsauswertung

Im Englischen ist die nachfolgend beschriebene Methode zur näherungsweisen 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 $f : \mathcal{X} \to \mathbb{C}$ mit $\mathcal{X} \subseteq \mathbb{C}$ nennt man *holomorph*, falls der Grenzwert

$$\lim_{h \to 0} \frac{f(z+h) - f(z)}{h} \qquad \forall\, z \in \mathcal{X} \tag{1.35}$$

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

Eine Funktion $f : \mathcal{X} \to \mathbb{C}$ in mehreren Variablen, d. h. $\mathcal{X} \subseteq \mathbb{C}^n$, 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. $f : \mathbb{R}^n \to \mathbb{R}$, kann

näherungsweise mit der Formel

$$(\nabla f)(\mathbf{x}) \approx \left[\frac{\mathrm{Im}(f(\mathbf{x}+h\mathrm{I}\mathbf{e}_i))}{h}\right]_{i=1,\ldots,n} \tag{1.36}$$

berechnet werden, wobei $h > 0$ eine kleine reelle Schrittweite darstellt. Zur Herleitung von (1.36), wird $f(\mathbf{x})$ zunächst in eine Taylorreihe

$$f(\mathbf{x}+h\mathrm{I}\mathbf{e}_i) = f(\mathbf{x}) + h\mathrm{I}\left.\frac{\partial f}{\partial x_i}\right|_{\mathbf{x}} - \frac{h^2}{2}\left.\frac{\partial^2 f}{\partial x_i^2}\right|_{\mathbf{x}} - \frac{h^3}{6}\mathrm{I}\left.\frac{\partial^3 f}{\partial x_i^3}\right|_{\mathbf{x}+\alpha h\mathrm{I}\mathbf{e}_i} \tag{1.37}$$

mit $\alpha \in (0,1)$ entwickelt. Dividiert man den Imaginärteil von (1.37) durch h, so folgen daraus

$$\left.\frac{\partial f}{\partial x_i}\right|_{\mathbf{x}} = \frac{\mathrm{Im}(f(\mathbf{x}+h\mathrm{I}\mathbf{e}_i))}{h} \underbrace{+\frac{h^2}{6}\left.\frac{\partial^3 f}{\partial x_i^3}\right|_{\mathbf{x}+\alpha h\mathrm{I}\mathbf{e}_i}}_{\text{Abschneidefehler, } \mathcal{O}(h^2)} \tag{1.38}$$

und somit direkt die Komponenten von (1.36). 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 $\mathcal{O}(h^2)$ erzielt wird. Der Rundungsfehler besitzt die Ordnung $\mathcal{O}(e_r h^{-1})$. 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 (1.37) folgt noch

$$f(\mathbf{x}) = \mathrm{Re}(f(\mathbf{x}+h\mathrm{I}\mathbf{e}_i)) + \mathcal{O}(h^2)\ . \tag{1.39}$$

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

1.3.5 Satz über implizite Funktionen

Satz 1.4 (Satz über implizite Funktionen)**.** *Es sei* $\mathbf{f} : \mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R}^m$ *eine stetig differenzierbare Funktion. Sind an einem Punkt* $(\bar{\mathbf{x}}, \bar{\mathbf{y}}) \in \mathbb{R}^n \times \mathbb{R}^m$

$$\mathbf{f}(\bar{\mathbf{x}}, \bar{\mathbf{y}}) = \mathbf{0} \tag{1.40}$$

und

$$\mathrm{rang}\left(\left.\frac{\partial \mathbf{f}(\bar{\mathbf{x}}, \mathbf{y})}{\partial \mathbf{y}}\right|_{\mathbf{y}=\bar{\mathbf{y}}}\right) = m \tag{1.41}$$

erfüllt, so existieren eine offene Umgebung $\mathcal{X} \in \mathbb{R}^n$ von $\bar{\mathbf{x}}$, eine offene Umgebung $\mathcal{Y} \in \mathbb{R}^m$ von $\bar{\mathbf{y}}$ und eine stetig differenzierbare Funktion $\mathbf{g} : \mathcal{X} \to \mathcal{Y}$ genau so, dass

$$\mathbf{f}(\mathbf{x}, \mathbf{y}) = \mathbf{0} \quad \Leftrightarrow \quad \mathbf{y} = \mathbf{g}(\mathbf{x}) \qquad \forall\, (\mathbf{x}, \mathbf{y}) \in \mathcal{X} \times \mathcal{Y} \tag{1.42}$$

gilt.

Folglich gilt auch $\mathbf{f}(\mathbf{x}, \mathbf{g}(\mathbf{x})) = \mathbf{0}\ \forall\, \mathbf{x} \in \mathcal{X}$. Für gegebenes $\mathbf{x}$ kann der eindeutige Wert $\mathbf{g}(\mathbf{x})$ (ohne explizite Kenntnis der Funktion $\mathbf{g}(\mathbf{x})$) durch numerisches Lösen der Gleichung

$$\mathbf{f}(\mathbf{x}, \mathbf{y}) = \mathbf{0} \tag{1.43}$$

nach $\mathbf{y}$ berechnet werden. Die Jacobi-Matrix $(\nabla \mathbf{g})(\mathbf{x}) \in \mathbb{R}^{n \times m}$ kann (ohne explizite Kenntnis der Funktion $\mathbf{g}(\mathbf{x})$) durch *implizites Ableiten* in der Form

$$(\nabla \mathbf{g})(\mathbf{x}) = -\frac{\partial \mathbf{f}(\mathbf{x}, \mathbf{g}(\mathbf{x}))}{\partial \mathbf{x}} \left(\frac{\partial \mathbf{f}(\mathbf{x}, \mathbf{y})}{\partial \mathbf{y}} \bigg|_{\mathbf{y}=\mathbf{g}(\mathbf{x})} \right)^{-1} \tag{1.44}$$

berechnet werden.

1.3.6 Konvexität

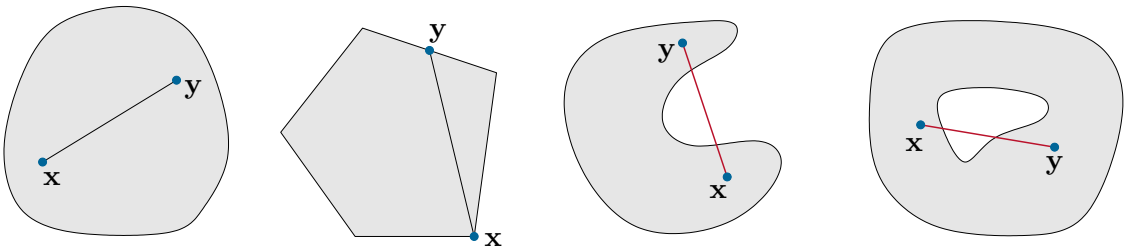
Die Eigenschaft der Konvexität ist von großer Bedeutung in der Optimierung und erlaubt häufig eine einfache (numerische) Lösung einer Optimierungsaufgabe. Der Begriff *konvex* kann sowohl auf Mengen als auch auf Funktionen angewandt werden.

1.3.6.1 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 α mit $0 < \alpha < 1$ gilt

$$(\alpha\, \mathbf{x} + (1 - \alpha)\, \mathbf{y}) \in \mathcal{X}\,. \tag{1.45}$$

Eine geometrische Interpretation dieser Definition ist, dass eine Menge $\mathcal{X} \subseteq \mathbb{R}^n$ genau dann konvex ist, falls die Verbindungslinie zwischen zwei beliebigen Punkten $\mathbf{x}, \mathbf{y} \in \mathcal{X}$ komplett in $\mathcal{X}$ enthalten ist. Abbildung 1.13 zeigt für den Raum $\mathbb{R}^2$ einige Beispiele konvexer und nicht-konvexer Mengen.



(a) konvex (b) konvex (c) nicht-konvex (d) nicht-konvex

Abbildung 1.13: Beispiele von konvexen und nicht-konvexen Mengen im $\mathbb{R}^2$.

Konvexe Mengen besitzen folgende Eigenschaften:

(a) Die *Schnittmenge* von konvexen Mengen ist wiederum konvex.

(b) Wenn $\mathcal{X} \subseteq \mathbb{R}^n$ eine konvexe Menge, $\mathbf{A} \in \mathbb{R}^{m \times n}$ eine feste Matrix und $\mathbf{b} \in \mathbb{R}^m$ ein fester Vektor ist, dann ist die Menge

$$\{\mathbf{A}\mathbf{x} + \mathbf{b} \,|\, \mathbf{x} \in \mathcal{X}\} \subseteq \mathbb{R}^m \tag{1.46}$$

ebenfalls konvex. D. h. das Bild einer konvexen Menge unter einer *affinen Transformation* ist konvex.

(c) Wenn $\mathcal{X}$ und $\mathcal{Y}$ konvexe Mengen sind, dann ist die Menge

$$\{\mathbf{x} + \mathbf{y} \,|\, \mathbf{x} \in \mathcal{X},\ \mathbf{y} \in \mathcal{Y}\} \tag{1.47}$$

ebenfalls konvex.

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

1.3.6.2 Konvexe Funktionen

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

$$f(\alpha\,\mathbf{x} + (1-\alpha)\,\mathbf{y}) \leq \alpha\, f(\mathbf{x}) + (1-\alpha)\, f(\mathbf{y}). \tag{1.48}$$

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

$$f(\alpha\,\mathbf{x} + (1-\alpha)\,\mathbf{y}) < \alpha\, f(\mathbf{x}) + (1-\alpha)\, f(\mathbf{y}). \tag{1.49}$$

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

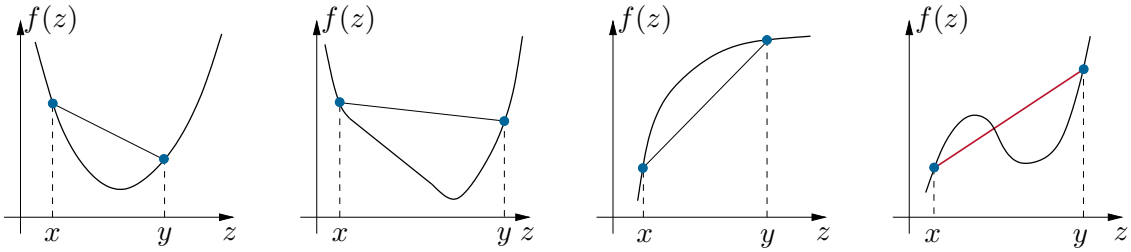
Die Definition 1.7 kann wie folgt geometrisch interpretiert werden: Eine Funktion f ist konvex (konkav), falls für alle $\mathbf{x} \in \mathcal{X}$, $\mathbf{y} \in \mathcal{X}$ und $0 < \alpha < 1$ alle Funktionswerte $f(\alpha\,\mathbf{x} + (1-\alpha)\,\mathbf{y})$ entlang der Achse f unterhalb (oberhalb)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

Konvexe Funktionen besitzen folgende Eigenschaften:

(a) Die Summenfunktion

$$f(\mathbf{x}) = \sum_{i=1}^{k} a_i\, f_i(\mathbf{x}) \tag{1.50}$$

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



(a) strikt konvex (b) konvex (nicht strikt) (c) konkav (d) weder konvex noch konkav

Abbildung 1.14: Beispiele von konvexen und konkaven Funktionen.

(b) Ist die Funktion $f(\mathbf{y})$ auf der konvexen Menge $\mathcal{Y} \subseteq \mathbb{R}^m$ konvex und existieren eine konvexe Menge $\mathcal{X} \subseteq \mathbb{R}^n$, eine feste Matrix $\mathbf{A} \in \mathbb{R}^{m \times n}$ und ein fester Vektor $\mathbf{b} \in \mathbb{R}^m$ so, dass

$$\{\mathbf{A}\mathbf{x} + \mathbf{b} \,|\, \mathbf{x} \in \mathcal{X}\} \subseteq \mathcal{Y} \tag{1.51}$$

gilt, dann ist die Funktion

$$\tilde{f}(\mathbf{x}) = f(\mathbf{A}\mathbf{x} + \mathbf{b}) \tag{1.52}$$

konvex auf $\mathcal{X}$.

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

$$\{\mathbf{x} \in \mathcal{X} \,|\, f(\mathbf{x}) \leq c\} \tag{1.53}$$

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

(d) *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

$$f(\mathbf{y}) \geq f(\mathbf{x}) + (\mathbf{y} - \mathbf{x})^{\mathrm{T}} (\nabla f)(\mathbf{x}) \tag{1.54}$$

erfüllt ist. Die geometrische Interpretation der Ungleichung (1.54) ist, dass an jedem Punkt $\mathbf{x}$ einer konvexen Funktion $f(\mathbf{x})$ eine sogenannte *stützende Hyperebene* (skalarer Fall: *stützende Tangente*) existieren muss, oberhalb oder auf der $f(\mathbf{x})$ verläuft. Dies ist in Abbildung 1.16 veranschaulicht.

(e) *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.

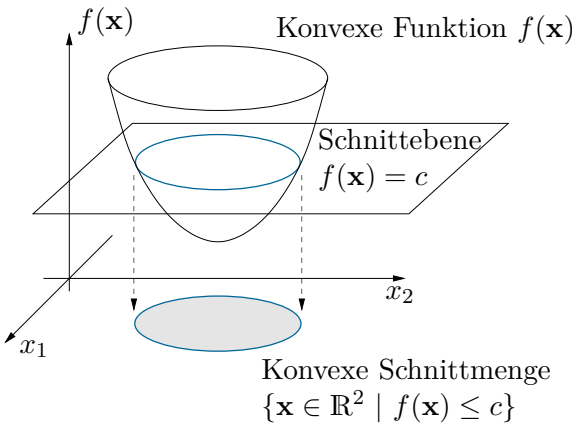


Abb. 1.15: Konvexe Menge, die durch den Schnitt einer konvexen Funktion $f(\mathbf{x})$ mit der Ebene $f(\mathbf{x}) = \text{konst.}$ entsteht.

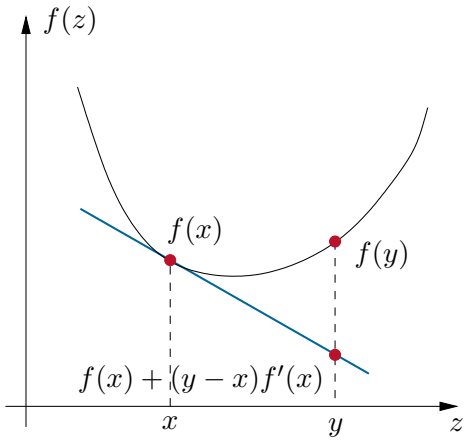


Abb. 1.16: Stützende Tangente einer konvexen Funktion $f(z)$.

Aufgabe 1.1. Beweisen Sie die Eigenschaften (a)–(e) von konvexen Funktionen. Nutzen Sie für den Beweis der Eigenschaft (e) den Mittelwertsatz, siehe Satz 1.3, im Speziellen (1.28).

Aufgabe 1.2. Zeigen Sie, dass die Funktion $f(\mathbf{x}) = x_1^4 + x_1^2 - 2x_1x_2 + x_2^2$ mit $\mathbf{x} = [x_1 \quad x_2]^{\mathrm{T}} \in \mathbb{R}^2$ über ihrem gesamten Definitionsbereich $\mathbb{R}^2$ 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* (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* (Other Titles in Applied Mathematics), 2. Aufl. 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* (Springer Series in Operations Research and Financial Engineering), 2. Aufl. 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* (International Series in Operations Research & Management Science), 3. Aufl. 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 30. 09. 2020).

2 Statische Optimierung: Unbeschränkter Fall

In diesem Kapitel werden unbeschränkte statische Optimierungsprobleme der Art

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

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.

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*

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

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

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

mit $g'(0) = \mathbf{d}^{\mathrm{T}}(\nabla f)(\mathbf{x}^*)$. 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 $g'(0) = \mathbf{d}^{\mathrm{T}}(\nabla f)(\mathbf{x}^*) \geq 0$. Dies kann aber nur dann für beliebige $\mathbf{d} \in \mathbb{R}^n$ erfüllt sein, wenn $(\nabla f)(\mathbf{x}^*) = \mathbf{0}$ gilt. □

Beispiel 2.1. Man betrachte das Optimierungsproblem

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

Aus den notwendigen Optimalitätsbedingungen erster Ordnung gemäß (2.2)

$$2x_1 - x_2 = 0 \tag{2.5a}$$
$$-x_1 + 2x_2 = 3\ , \tag{2.5b}$$

folgt die eindeutige Lösung $\mathbf{x}^* = [1 \quad 2]^{\mathrm{T}}$. Es kann mit Hilfe des nachfolgenden Satzes 2.3 gezeigt werden, dass $\mathbf{x}^*$ in diesem Fall sogar ein striktes 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.

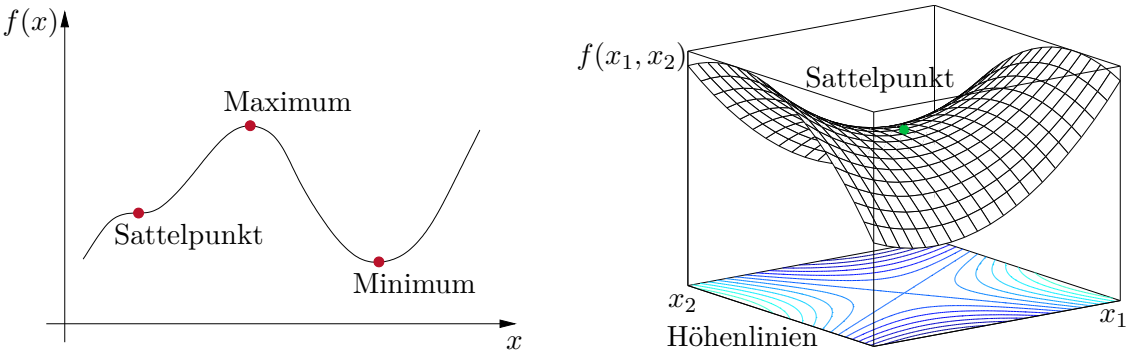


Abbildung 2.1: Beispiele von stationären Punkten im ein- und zweidimensionalen Fall.

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*

$$(a) \quad (\nabla f)(\mathbf{x}^*) = \mathbf{0} \tag{2.6a}$$
$$(b) \quad \left(\nabla^2 f\right)(\mathbf{x}^*) \text{ ist positiv semi-definit.} \tag{2.6b}$$

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 Extremalpunkt ($f'(x^*) = 3(x^*)^2 = 0$) 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*

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

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

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

$$\min_{\mathbf{x}\in\mathbb{R}^2} f(\mathbf{x}) = \; x_1^2 + a x_2^2 - x_1 x_2 \tag{2.8}$$

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

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

Aus $(\nabla f)(\mathbf{x}^*) = \mathbf{0}$ folgt $\mathbf{x}^* = [0 \quad 0]^{\mathrm{T}}$ 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

$$D_1 = 2, \quad D_2 = 4a - 1\,. \tag{2.10}$$

Somit ist $(\nabla^2 f)(\mathbf{x}^*)$ positiv definit für $a > \frac{1}{4}$ und $\mathbf{x}^* = [0 \quad 0]^{\mathrm{T}}$ 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 $\mathbf{x}^* = [0 \quad 0]^{\mathrm{T}}$ 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

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

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 $\mathcal{G} = \arg\min\{f(\mathbf{x}) \,|\, \mathbf{x} \in \mathbb{R}^n\}$ 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 $\mathcal{G} = \{\mathbf{x} \,|\, \mathbf{x} \in \mathbb{R}^n,\ f(\mathbf{x}) \leq c\}$ gemäß (1.53) 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 $\alpha \in [0, 1]$

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

Dies zeigt, dass mit $\alpha \to 1$ 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. □

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.

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

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

$$\lim_{k\to\infty} \mathbf{x}_k = \mathbf{x}^*. \tag{2.14}$$

Solche Algorithmen werden auch als *iterative Abstiegsverfahren* (englisch: iterative descent methods) bezeichnet und (2.13) als *Abstiegsbedingung*. 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* $e : \mathbb{R}^n \to \mathbb{R}$, 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 Abstand

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

im Sinne einer Norm $||\cdot||$ oder die Kostendifferenz

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

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 die kleinste obere Schranke (Supremum) der nichtnegativen Zahlen p, für die gilt

$$0 \leq \lim_{k\to\infty} \frac{|e_{k+1}|}{|e_k|^p} = \mu < \infty. \tag{2.16}$$

Als zugehörige *Konvergenzrate* bezeichnet man die Zahl μ. 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 $k \to \infty$. Größere Werte der Konvergenzordnung p bedeuten, dass die Folge schneller konvergiert, da die Folgenelemente e_k (zumindest für sehr große Werte von k) mit der p-ten Potenz abnehmen. Analoges gilt für kleinere Werte der Konvergenzrate μ.

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

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

erfüllt ist. Mit $p = 1$ folgt dann

$$\lim_{k\to\infty} \frac{a^{k+1}}{a^k} = a = \mu \tag{2.18}$$

für die Konvergenzrate μ.

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 die Konvergenzordnung 1, da nur für $p \leq 1$ die Bedingung

$$\lim_{k\to\infty} \frac{k^{kp}}{(k+1)^{k+1}} = \lim_{k\to\infty} \frac{k}{k+1}\left(\frac{k}{k+1}\right)^k k^{k(p-1)-1} = \lim_{k\to\infty} 1\frac{1}{\mathrm{e}} k^{k(p-1)-1} < \infty \quad (2.19)$$

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

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

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 gängige 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.

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

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

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

bestimmt. Die Iteration wird solange wiederholt, bis ein Abbruchkriterium erfüllt ist, z. B. bis eine gewählte Fehlerfunktion betraglich 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 α_k, auf das Auffinden eines Minimums entlang einer Linie reduziert.

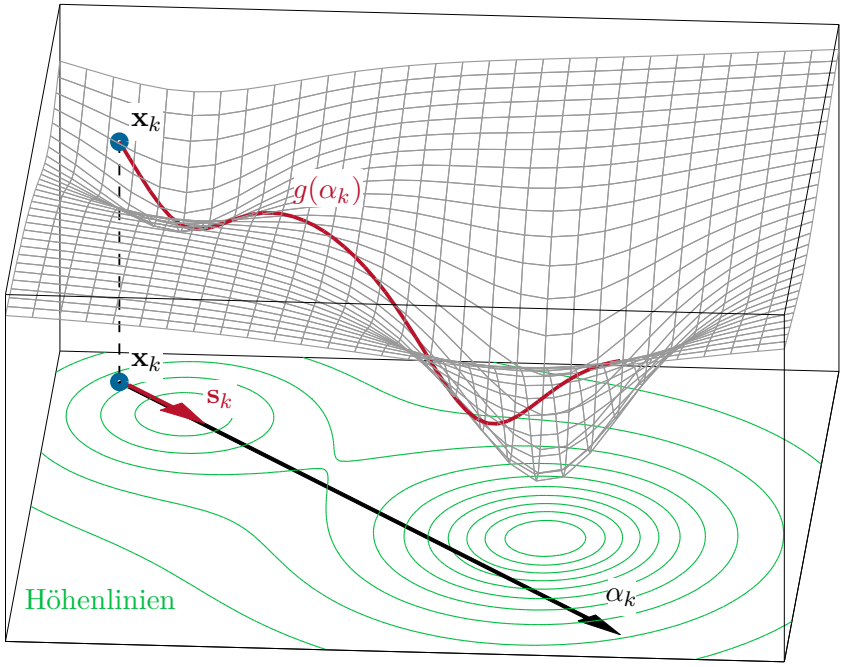


Abbildung 2.2: Veranschaulichung der Wahl der Schrittweite gemäß (2.23).

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 $g(\alpha_k)$ im Intervall $[l_0, r_0]$ stetig ist und genau ein eindeutiges Minimum besitzt.

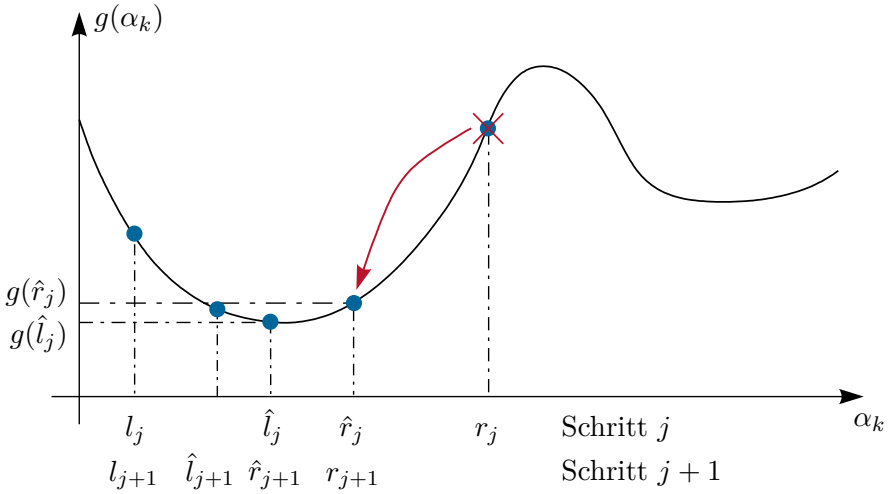


Abbildung 2.3: Veranschaulichung des Intervallschachtelungsverfahrens nach dem Prinzip des „Goldenen Schnittes“.

Zum Iterationsschritt j liege das Intervall $[l_j, r_j]$ vor, das nach wie vor jenen Wert α_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

$$\hat{l}_j = l_j + (1-a)(r_j - l_j) = a l_j + (1-a) r_j \tag{2.24a}$$
$$\hat{r}_j = l_j + a(r_j - l_j) \qquad = (1-a) l_j + a r_j \tag{2.24b}$$

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 Funktion mit genau einem eindeutigen Minimum auf dem Intervall $[l_j, r_j]$. Bezeichnet man dieses Minimum mit $\alpha_k^* \in (l_j, r_j)$, 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 Forderung genau eines eindeutigen Minimums von $g(\alpha_k)$ nicht möglich. Für $g(\hat{l}_j) \geq g(\hat{r}_j)$ folgt der Beweis auf analoge Art und Weise. □

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 mit $l_{j+1} = l_j$ und $r_{j+1} = \hat{r}_j = (1-a)l_j + ar_j$ eine weitere Iteration zur Berechnung der Zwischenpunkte gemäß (2.24) in der Form

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

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

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

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

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 α_k^* entweder aus der *Mittelung* der letzten Intervallgrenzen $\alpha_k^* = (l_j + r_j)/2$ oder aus einer *quadratischen Interpolation* (siehe nachfolgender Abschnitt) zwischen den kleinsten drei Funktionswerten 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. Dieser Ansatz wird gelegentlich auch als Newton-Methode bezeichnet. Es wird angenommen, 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. Gemäß Lagrangescher Interpolationsformel lautet die quadratische Interpolationsfunktion $q(\alpha_k)$ durch diese Punkte dann

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

und der optimale Wert α_k^* folgt in der Form

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

Der so errechnete Wert α_k^* sollte als optimale Schrittweite nur akzeptiert werden, wenn die Interpolationsfunktion $q(\alpha_k)$ strikt konvex ist (nur dann ist α_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.

Initialisierung: $l_0,\, r_0$ (Startintervall mit Minimum im Inneren)
$j = 0$ (Startindex)
$a = 0.618$ (Goldener Schnitt-Parameter)
ε_{lr} (Schranke für Abbruch)
$\hat{l}_0 \leftarrow a l_0 + (1-a) r_0$ (innere Punkte)
$\hat{r}_0 \leftarrow (1-a) l_0 + a r_0$

repeat

if $g(\hat{l}_j) > g(\hat{r}_j)$
$l_{j+1} \leftarrow \hat{l}_j$
$r_{j+1} \leftarrow r_j$
$\hat{l}_{j+1} \leftarrow \hat{r}_j$
$\hat{r}_{j+1} \leftarrow (1-a) l_{j+1} + a r_{j+1}$
else (d. h. $g(\hat{l}_j) \leq g(\hat{r}_j)$)
$r_{j+1} \leftarrow \hat{r}_j$
$l_{j+1} \leftarrow l_j$
$\hat{r}_{j+1} \leftarrow \hat{l}_j$
$\hat{l}_{j+1} \leftarrow a l_{j+1} + (1-a) r_{j+1}$
end if
$j \leftarrow j + 1$

until $|r_j - l_j| \leq \varepsilon_{lr}$
$\alpha_k^* = (l_j + r_j)/2$

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),

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

zugrunde.

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

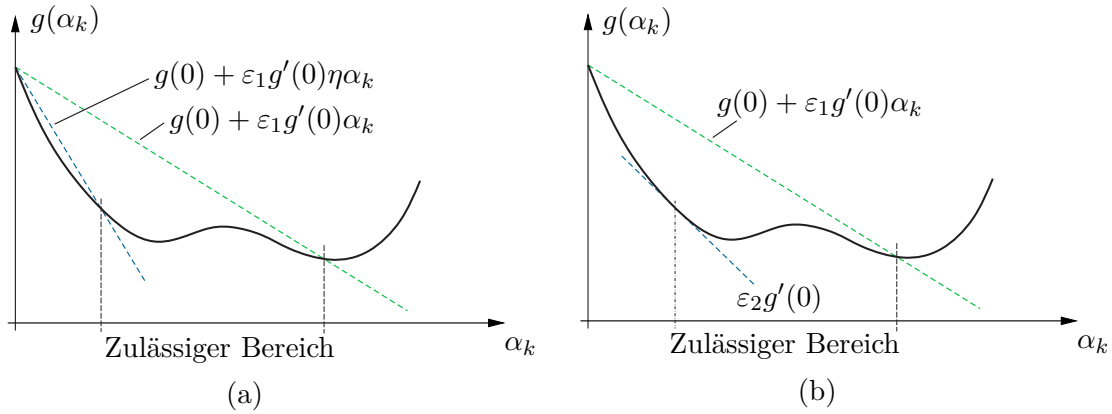


Abbildung 2.4: Veranschaulichung der Armijo-Goldstein-Bedingung (a) und Wolfe-Bedingung (b).

und bricht nach dem linearen Glied ab, so erhält man

$$g(\alpha_k) \approx g(0) + g'(0)\alpha_k \ . \tag{2.30}$$

Bei der *Armijo-Goldstein-Bedingung* wird nun die Schrittweite α_k so gewählt, dass für ein festes ε_1, $0 < \varepsilon_1 < 1$, die Ungleichung

$$g(\alpha_k) \leq g(0) + \varepsilon_1 g'(0)\alpha_k \tag{2.31}$$

erfüllt ist. Dies garantiert, dass die Schrittweite α_k nach oben hin beschränkt ist. Um sicherzustellen, dass die Schrittweite nicht zu klein wird, führt man bei der *Armijo-Goldstein-Bedingung* zusätzlich einen Parameter η mit $1/\varepsilon_1 > \eta > 1$ ein und fordert, dass die Schrittweite α_k der Ungleichung

$$g(\alpha_k) > g(0) + \varepsilon_1 g'(0)\eta\alpha_k \tag{2.32}$$

genügt. Abbildung 2.4(a) zeigt eine grafische Veranschaulichung der Armijo-Goldstein-Bedingung. In der Praxis geht man häufig so vor, dass man in einem ersten Schritt einen (weitgehend beliebigen) Startwert für α_k wählt. Ist für dieses α_k die Ungleichung (2.31) erfüllt, dann erhöht man α_k sukzessive um den Faktor η solange, bis die Ungleichung (2.31) erstmals verletzt wird. Der vorletzte Wert von α_k wird dann als geeignete Schrittweite gewählt. Umgekehrt, wenn der Startwert von α_k die Ungleichung (2.31) nicht erfüllt, dann wird α_k sukzessive durch den Faktor η 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 ε_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-Goldstein-Bedingung die so genannte *Wolfe-Bedingung*. Dabei wird ein weiterer Parameter ε_2 mit $0 < \varepsilon_1 < \varepsilon_2 < 1$ eingeführt und von der Schrittweite α_k wird gefordert, dass sie die Ungleichungen (2.31) und

$$g'(\alpha_k) \geq \varepsilon_2 g'(0) \tag{2.33}$$

erfüllt. Abbildung 2.4(b) gibt eine grafische Veranschaulichung dieses Sachverhaltes. Typische Werte für ε_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

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

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

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

wobei $\mathcal{O}(\alpha_k^2)$ den quadratischen Restterm bezeichnet, so zeigt sich, dass für hinreichend kleines α_k die Abstiegsbedingung (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

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

mit der symmetrischen 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 (e) konvexer Funktionen von Abschnitt 1.3.6.2 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

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

in der Form

$$\mathbf{x}^* = \mathbf{Q}^{-1} \mathbf{b} \ . \tag{2.38}$$

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

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

Die optimale Schrittweite α_k^* kann durch explizites Lösen des Optimierungsproblems gemäß (2.23)

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

in der Form

$$\alpha_k^* = \frac{\mathbf{g}_k^{\mathrm{T}} \mathbf{g}_k}{\mathbf{g}_k^{\mathrm{T}} \mathbf{Q} \mathbf{g}_k} \tag{2.41}$$

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

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

anschreiben.

Für weitere Überlegungen zum Konvergenzverhalten der Gradientenmethode ist es vorteilhaft, unter Verwendung von (2.38) anstelle von $f(\mathbf{x})$ die Fehlerfunktion

$$\begin{aligned} e(\mathbf{x}) &= ||\mathbf{x} - \mathbf{x}^*||_{\mathbf{Q}} = \sqrt{(\mathbf{x} - \mathbf{x}^*)^{\mathrm{T}}\mathbf{Q}(\mathbf{x} - \mathbf{x}^*)} \\ &= \sqrt{\mathbf{x}^{\mathrm{T}}\mathbf{Q}\mathbf{x} - 2\mathbf{x}^{\mathrm{T}}\mathbf{Q}\mathbf{x}^* + (\mathbf{x}^*)^{\mathrm{T}}\mathbf{Q}\mathbf{x}^*} = \sqrt{2f(\mathbf{x}) + (\mathbf{x}^*)^{\mathrm{T}}\mathbf{Q}\mathbf{x}^*} \end{aligned} \tag{2.43}$$

zu betrachten. Die Minima $\mathbf{x}^*$ der Funktionen $f(\mathbf{x})$ und $e(\mathbf{x})$ sind identisch. Die Vektornorm $||\cdot||_{\mathbf{Q}}$ mit einer symmetrischen positiv definiten Matrix $\mathbf{Q}$ wird auch als *Energienorm* bezeichnet.

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

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

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

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

Folglich gilt

$$e(\mathbf{x}_k) = \sqrt{\mathbf{g}_k^{\mathrm{T}}\mathbf{Q}^{-1}\mathbf{g}_k} \ . \tag{2.46}$$

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

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

□

Um nun die Konvergenzrate der Gradientenmethode bei Verwendung der Fehlerfunktion $e(\mathbf{x})$ abschätzen zu können, wird noch folgendes Lemma benötigt.

Lemma 2.2 (Ungleichung von Kantorovich). *Es sei $\mathbf{Q} \in \mathbb{R}^{n\times n}$ eine symmetrische positiv definite Matrix. Für jeden Vektor $\mathbf{x} \in \mathbb{R}^n$ mit $\mathbf{x} \neq \mathbf{0}$ gilt dann die Ungleichung*

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

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

Der Beweis von Lemma 2.2 ist z. B. in [2.2] skizziert.

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). Die Fehlerfunktion $e(\mathbf{x})$ gemäß (2.43) konvergiert dabei* linear *gegen* 0 mit der Konvergenzrate

$$e(\mathbf{x}_{k+1}) \leq \frac{\kappa-1}{\kappa+1} e(\mathbf{x}_k) \ , \tag{2.49}$$

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

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

□

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 für Beispiele mit $\mathbf{x} \in \mathbb{R}^2$. 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 (κ sehr groß) kann die Gradientenmethode sehr langsam konvergieren, siehe Abbildung 2.6. Das in Abbildung 2.6 gezeigte Verhalten wird im Englischen als *zigzagging* bezeichnet.

Die Gradientenmethode kann natürlich auch auf nichtquadratische Kostenfunktionen angewandt werden. Für diesen Fall beschreibt der nachfolgende Satz das Konvergenzverhalten

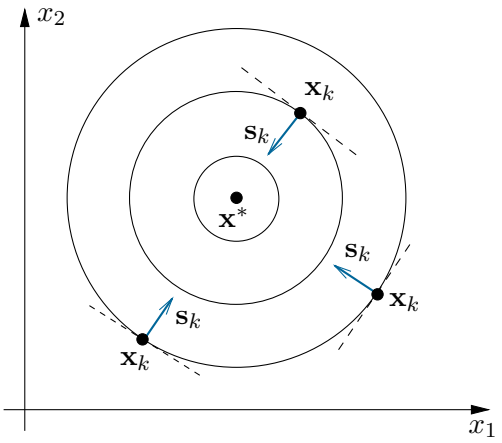


Abb. 2.5: Beispiel eines ideal konditionierten Problems für die Gradientenmethode.

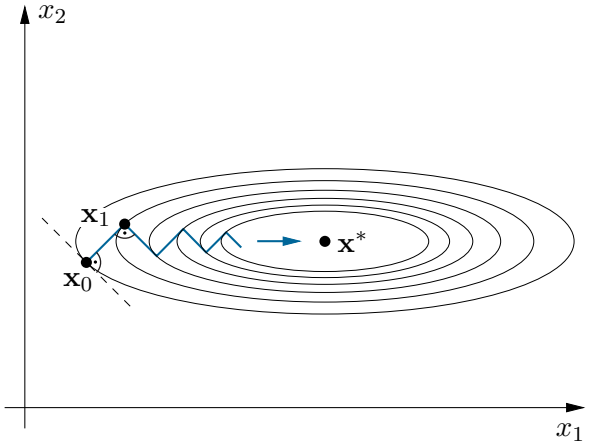


Abb. 2.6: Beispiel eines schlecht konditionierten Problems für die Gradientenmethode.

der Gradientenmethode. Sein Beweis findet sich z. B. in [2.2].

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*

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

für eine geeignete Schrittweite α_k gegen das lokale Minimum $\mathbf{x}^*$ konvergiert, dann konvergiert die Folge $\{e(\mathbf{x}_k)\}$ linear gegen 0 mit einer Konvergenzrate die im besten Fall $\frac{\kappa-1}{\kappa+1}$ beträgt.

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

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

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

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

gegenüber der Verteilung 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 später beschriebenen Newton-Methode. Sie hat aber den Nachteil, dass die gesamte Hessematrix explizit berechnet werden muss. Um diesen Rechenaufwand zu vermeiden, wird alternativ häufig eine Diagonalmatrix $\mathbf{T}$ verwendet, deren Diagonaleinträge beispielsweise in der Form $T_{ii} = \left((\nabla^2 f)_{ii}(\mathbf{x}^*)\right)^{-\frac{1}{2}}$, $i = 1, \ldots, n$ gewählt werden können. Wird eine Diagonalmatrix $\mathbf{T}$ verwendet, so führt dies zu einer reinen Skalierung (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 *CG method*) versucht nun bei nur geringfügig erhöhtem Rechen- und Speicheraufwand eine schnellere Konvergenz als die Gradientenmethode zu erreichen. Weitere Informationen zu der Methode finden sich z. B. in [2.2–2.5]. Ursprünglich wurde diese Methode zur Lösung hochdimensionaler quadratischer Probleme der Form (siehe auch (2.36))

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

sowie hochdimensionaler linearer Gleichungssysteme der Form

$$\mathbf{0} = \mathbf{Q}\mathbf{x} - \mathbf{b} \tag{2.55}$$

jeweils mit einer positiv definiten Matrix $\mathbf{Q} \in \mathbb{R}^{n\times n}$ und einem beliebigen Vektor $\mathbf{b} \in \mathbb{R}^n$ entwickelt. Da die Kostenfunktion (2.54) strikt konvex ist, entspricht (2.55) einem hinreichenden Optimalitätskriterium erster Ordnung für ein striktes globales Minimum. Soll ein lineares Gleichungssystem

$$\mathbf{0} = \mathbf{A}\mathbf{x} - \mathbf{c} \tag{2.56}$$

mit einer regulären aber nicht positiv definiten Matrix $\mathbf{A} \in \mathbb{R}^{n\times n}$ und einem beliebigen Vektor $\mathbf{c} \in \mathbb{R}^n$ gelöst werden, so kann es durch linksseitige Multiplikation mit $\mathbf{A}^{\mathrm{T}}$ auf die Form

$$\mathbf{0} = \mathbf{A}^{\mathrm{T}}\mathbf{A}\mathbf{x} - \mathbf{A}^{\mathrm{T}}\mathbf{c} \tag{2.57}$$

mit der positiv definiten Matrix $\mathbf{A}^{\mathrm{T}}\mathbf{A}$ gebracht werden und ebenfalls mit der konjugierten Gradientenmethode gelöst werden.

Bevor nun die konjugierte Gradientenmethode konkret erläutert wird, sollen einige Grundlagen dazu erarbeitet werden.

Definition 2.2 (**Q**-Orthogonalität)**.** Nichttriviale Vektoren $\mathbf{d}_j,\ j = 0, \dots, k$ heißen *konjugiert bezüglich einer positiv definiten Matrix* $\mathbf{Q}$ bzw. $\mathbf{Q}$*-orthogonal*, wenn gilt $\mathbf{d}_i^{\mathrm{T}}\mathbf{Q}\mathbf{d}_j = 0\ \forall\, i \neq j$. Die Vektoren $\mathbf{d}_j,\ j = 0, \dots, k$ sind dann linear unabhängig.

Für $\mathbf{Q} = \mathbf{E}$ fällt der Begriff der $\mathbf{Q}$-Orthogonalität folglich mit dem klassischen Begriff der Orthogonalität zusammen.

Aufgabe 2.6. Beweisen Sie, dass $\mathbf{Q}$-orthogonale Vektoren $\mathbf{d}_j,\ j = 0, \dots, k$ linear unabhängig sind.

Bei der konjugierten Gradientenmethode wird ausgehend von einer Startlösung $\mathbf{x}_0$ iterativ entlang von $\mathbf{Q}$-orthogonalen Suchrichtungen $\mathbf{d}_j$ die optimale Lösung $\mathbf{x}^*$ des Optimierungsproblems (2.54) gesucht. In jedem Iterationsschritt k gilt daher

$$\mathbf{x}_{k+1} = \mathbf{x}_k + \alpha_k \mathbf{d}_k \tag{2.58}$$

mit der optimalen Schrittweite α_k sowie

$$\mathbf{x}_{k+1} - \mathbf{x}_0 = \sum_{j=0}^{k} \alpha_j \mathbf{d}_j\ . \tag{2.59}$$

Für den Gradienten von (2.54) wird die Abkürzung

$$\mathbf{g}_k = (\nabla f)(\mathbf{x}_k) = \mathbf{Q}\mathbf{x}_k - \mathbf{b} \tag{2.60}$$

eingeführt. Es soll nun gezeigt werden, dass die optimalen Schrittweiten α_j mit geringem mathematischem Aufwand berechnet werden können. Aus (2.54) und (2.59) folgt unter Ausnützung der $\mathbf{Q}$-Orthogonalität der Vektoren $\mathbf{d}_j$

$$\begin{aligned} f(\mathbf{x}_k) &= \frac{1}{2}\left(\mathbf{x}_0 + \sum_{j=0}^{k-1} \alpha_j \mathbf{d}_j\right)^{\mathrm{T}} \mathbf{Q}\left(\mathbf{x}_0 + \sum_{j=0}^{k-1} \alpha_j \mathbf{d}_j\right) - \left(\mathbf{x}_0 + \sum_{j=0}^{k-1} \alpha_j \mathbf{d}_j\right)^{\mathrm{T}} \mathbf{b} \\ &= \frac{1}{2}\mathbf{x}_0^{\mathrm{T}}\mathbf{Q}\mathbf{x}_0 - \mathbf{x}_0^{\mathrm{T}}\mathbf{b} + \sum_{j=0}^{k-1}\left(\frac{1}{2}\alpha_j^2 \mathbf{d}_j^{\mathrm{T}}\mathbf{Q}\mathbf{d}_j + \alpha_j \mathbf{g}_0^{\mathrm{T}}\mathbf{d}_j\right)\ . \end{aligned} \tag{2.61}$$

Das Problem zerfällt also in unabhängige skalare Optimierungsprobleme für die Schrittweiten α_j. Deren Lösung lautet

$$\alpha_j = -\frac{\mathbf{g}_0^{\mathrm{T}}\mathbf{d}_j}{\mathbf{d}_j^{\mathrm{T}}\mathbf{Q}\mathbf{d}_j} = -\frac{\mathbf{g}_j^{\mathrm{T}}\mathbf{d}_j}{\mathbf{d}_j^{\mathrm{T}}\mathbf{Q}\mathbf{d}_j}\ , \tag{2.62}$$

wobei hier die Identität

$$\mathbf{g}_j^{\mathrm{T}}\mathbf{d}_j = \left(\mathbf{Q}\left(\mathbf{x}_0 + \sum_{i=0}^{j-1} \alpha_i \mathbf{d}_i\right) - \mathbf{b}\right)^{\mathrm{T}} \mathbf{d}_j = \mathbf{g}_0^{\mathrm{T}}\mathbf{d}_j \tag{2.63}$$

verwendet wurde. Wird also in jedem Iterationsschritt (2.58) die Schrittweite α_k gemäß (2.62) gewählt, so minimiert $\mathbf{x}_{k+1}$ die Kostenfunktion $f(\mathbf{x}_{k+1})$ in dem von den Vektoren

$\mathbf{d}_j,\ j = 0, \ldots, k$ aufgespannten Unterraum des $\mathbb{R}^n$. Nachdem die Vektoren $\mathbf{d}_j,\ j = 0, \ldots, k$ linear unabhängig sind (siehe Definition 2.2), ergibt sich daraus nach spätestens $k+1 = n$ Iterationsschritten die optimale Lösung $\mathbf{x}^* = \mathbf{x}_{k+1}$ des Optimierungsproblems (2.54). Wie die noch unbekannten Vektoren $\mathbf{d}_j$ zu wählen sind, geht aus dem nachfolgend beschriebenen Algorithmus der konjugierten Gradientenmethode hervor.

Satz 2.8 (Konjugierte Gradientenmethode)**.** *Für jeden Anfangswert $\mathbf{x}_0$ konvergiert die Folge*

$$\alpha_k = \frac{\mathbf{g}_k^{\mathrm{T}} \mathbf{g}_k}{\mathbf{d}_k^{\mathrm{T}} \mathbf{Q} \mathbf{d}_k} \tag{2.64a}$$

$$\mathbf{x}_{k+1} = \mathbf{x}_k + \alpha_k \mathbf{d}_k \tag{2.64b}$$

$$\beta_k = \frac{\mathbf{g}_{k+1}^{\mathrm{T}} \mathbf{g}_{k+1}}{\mathbf{g}_k^{\mathrm{T}} \mathbf{g}_k} \tag{2.64c}$$

$$\mathbf{d}_{k+1} = -\mathbf{g}_{k+1} + \beta_k \mathbf{d}_k \tag{2.64d}$$

mit $\mathbf{g}_k = \mathbf{Q}\mathbf{x}_k - \mathbf{b}$ und dem Anfangswert

$$\mathbf{d}_0 = -\mathbf{g}_0 = \mathbf{b} - \mathbf{Q}\mathbf{x}_0 \tag{2.65}$$

in höchstens n Iterationsschritten gegen die eindeutige optimale Lösung $\mathbf{x}^*$ des Optimierungsproblems (2.54). Die so konstruierte Folge von Suchrichtungen und Gradienten weist die Eigenschaften

$$\mathbf{g}_i^{\mathrm{T}} \mathbf{g}_j = 0 \qquad \forall\, i \neq j \tag{2.66a}$$

$$\mathbf{g}_i^{\mathrm{T}} \mathbf{d}_j = 0 \qquad \forall\, i > j \tag{2.66b}$$

$$\mathbf{d}_i^{\mathrm{T}} \mathbf{Q} \mathbf{d}_j = 0 \qquad \forall\, i \neq j \tag{2.66c}$$

auf. D. h. die Gradienten sind zueinander orthogonal, die Gradienten sind orthogonal auf alle bisherigen Suchrichtungen und die Suchrichtungen sind $\mathbf{Q}$-orthogonal.

Beweis. Entsprechend der Iterationsvorschrift (2.64b) gilt der Zusammenhang

$$\mathbf{g}_{k+1} = \mathbf{Q}\mathbf{x}_{k+1} - \mathbf{b} = \mathbf{Q}\mathbf{x}_k - \mathbf{b} + \alpha_k \mathbf{Q}\mathbf{d}_k = \mathbf{g}_k + \alpha_k \mathbf{Q}\mathbf{d}_k\ . \tag{2.67}$$

Nachfolgend werden ohne explizite Erwähnung laufend die Beziehungen (2.64) und (2.65) verwendet. Zunächst sollen die Eigenschaften (2.66) mittels vollständiger Induktion gezeigt werden. Zum Induktionsbeginn gilt unter Zuhilfenahme von (2.67)

$$\mathbf{g}_0^{\mathrm{T}} \mathbf{g}_1 = \mathbf{g}_0^{\mathrm{T}} (\mathbf{g}_0 + \alpha_0 \mathbf{Q}\mathbf{d}_0) = \mathbf{g}_0^{\mathrm{T}} \mathbf{g}_0 - \mathbf{d}_0^{\mathrm{T}} \alpha_0 \mathbf{Q}\mathbf{d}_0 = 0 \tag{2.68a}$$

für (2.66a) und daher

$$\mathbf{d}_0^{\mathrm{T}} \mathbf{g}_1 = 0 \tag{2.68b}$$

für (2.66b). Aus (2.67) folgt $\mathbf{Q}\mathbf{d}_0 = (\mathbf{g}_1 - \mathbf{g}_0)/\alpha_0$. Mit dieser Beziehung und (2.68a)

ergibt sich die Gültigkeit von (2.66c) zum Induktionsbeginn in der Form

$$\mathbf{d}_0^{\mathrm{T}}\mathbf{Q}\mathbf{d}_1 = \mathbf{d}_0^{\mathrm{T}}\mathbf{Q}(-\mathbf{g}_1 + \beta_0\mathbf{d}_0) = \frac{1}{\alpha_0}(\mathbf{g}_1 - \mathbf{g}_0)^{\mathrm{T}}\left(-\mathbf{g}_1 - \frac{\mathbf{g}_1^{\mathrm{T}}\mathbf{g}_1}{\mathbf{g}_0^{\mathrm{T}}\mathbf{g}_0}\mathbf{g}_0\right) = 0\ . \tag{2.68c}$$

Basierend auf der Induktionsannahme, dass (2.66) bei festem Index i für alle $j < i$ gilt, soll nun gezeigt werden, dass (2.66) auch für alle $j < i+1$ gilt. Dabei wird (2.67) mehrfach verwendet. Aus

$$\mathbf{g}_{i+1}^{\mathrm{T}}\mathbf{g}_j = (\mathbf{g}_i + \alpha_i\mathbf{Q}\mathbf{d}_i)^{\mathrm{T}}\mathbf{g}_j = \mathbf{g}_i^{\mathrm{T}}\mathbf{g}_j + \alpha_i\mathbf{d}_i^{\mathrm{T}}\mathbf{Q}(-\mathbf{d}_j + \beta_{j-1}\mathbf{d}_{j-1}) = 0 \tag{2.69a}$$

folgt im ersten Schritt die Gültigkeit von (2.66a) für alle $j < i+1$. Unter Verwendung von (2.69a) folgt aus

$$\begin{aligned}\mathbf{g}_{i+1}^{\mathrm{T}}\mathbf{d}_j &= \mathbf{g}_{i+1}^{\mathrm{T}}(-\mathbf{g}_j + \beta_{j-1}\mathbf{d}_{j-1})\\ &= \mathbf{g}_{i+1}^{\mathrm{T}}\beta_{j-1}\mathbf{d}_{j-1} = (\mathbf{g}_i + \alpha_i\mathbf{Q}\mathbf{d}_i)^{\mathrm{T}}\beta_{j-1}\mathbf{d}_{j-1} = 0\end{aligned} \tag{2.69b}$$

im zweiten Schritt die Gültigkeit von (2.66b) für alle $j < i+1$. Wieder unter Verwendung von (2.69a) folgt aus

$$\begin{aligned}\mathbf{d}_{i+1}^{\mathrm{T}}\mathbf{Q}\mathbf{d}_j &= (-\mathbf{g}_{i+1} + \beta_i\mathbf{d}_i)^{\mathrm{T}}\mathbf{Q}\mathbf{d}_j = -\mathbf{g}_{i+1}^{\mathrm{T}}\frac{1}{\alpha_j}(\mathbf{g}_{j+1} - \mathbf{g}_j) + \beta_i\mathbf{d}_i^{\mathrm{T}}\mathbf{Q}\mathbf{d}_j\\ &= -\mathbf{g}_{i+1}^{\mathrm{T}}\mathbf{g}_{j+1}\frac{\mathbf{d}_j^{\mathrm{T}}\mathbf{Q}\mathbf{d}_j}{\mathbf{g}_j^{\mathrm{T}}\mathbf{g}_j} + \beta_i\mathbf{d}_i^{\mathrm{T}}\mathbf{Q}\mathbf{d}_j = 0\end{aligned} \tag{2.69c}$$

im dritten Schritt die Gültigkeit von (2.66c) für alle $j < i+1$. Damit ist (2.66) gezeigt.

Wegen

$$\alpha_k = -\frac{\mathbf{g}_k^{\mathrm{T}}\mathbf{d}_k}{\mathbf{d}_k^{\mathrm{T}}\mathbf{Q}\mathbf{d}_k} = \frac{\mathbf{g}_k^{\mathrm{T}}(\mathbf{g}_k - \beta_{k-1}\mathbf{d}_{k-1})}{\mathbf{d}_k^{\mathrm{T}}\mathbf{Q}\mathbf{d}_k} = \frac{\mathbf{g}_k^{\mathrm{T}}\mathbf{g}_k}{\mathbf{d}_k^{\mathrm{T}}\mathbf{Q}\mathbf{d}_k} \tag{2.70}$$

liefert (2.64a) die optimale Schrittweite gemäß (2.62). Dass die konjugierte Gradientenmethode nach höchstens n Iterationsschritten zur optimalen Lösung $\mathbf{x}^*$ konvergiert, wenn die Schrittweiten α_k optimal gewählt werden und die Vektoren $\mathbf{d}_k,\ k = 0, \ldots, n-1$ $\mathbf{Q}$-orthogonal sind, folgt bereits aus den Überlegungen, die dem Satz 2.8 vorangegangen sind. □

Aufgabe 2.7. Zeigen Sie, dass die Suchrichtungen $\mathbf{d}_k$ der konjugierten Gradientenmethode nach Satz 2.8 Abstiegsrichtungen sind.

Anders als es die Bezeichnung *konjugierte Gradientenmethode* erwarten lässt, sind nicht die lokalen Gradienten $\mathbf{g}_k$ der Kostenfunktion $\mathbf{Q}$-orthogonal im Sinne der Definition 2.2, sondern die Suchrichtungen $\mathbf{d}_k$. Gemäß (2.64d) bildet der Vektor $\mathbf{d}_{k+1}$ eine Linearkombination der Richtung des steilsten Abstieges (negativer Gradient $-\mathbf{g}_{k+1}$) und der vorangegangenen Suchrichtung $\mathbf{d}_k$. Der Faktor β_k wird dabei gemäß (2.64c) genau so gewählt, dass die $\mathbf{Q}$-Orthogonalität zwischen $\mathbf{d}_{k+1}$ und allen vorangegangenen Suchrichtungen sichergestellt ist. Die Gleichung (2.64c) wird auch *Formel von Fletcher-Reeves*

genannt. Alternativ dazu kann die sogenannte *Formel von Hestenes-Stiefel*

$$\beta_k = \frac{(\mathbf{g}_{k+1} - \mathbf{g}_k)^{\mathrm{T}} \mathbf{g}_{k+1}}{(\mathbf{g}_{k+1} - \mathbf{g}_k)^{\mathrm{T}} \mathbf{d}_k} \tag{2.71}$$

verwendet werden.

Der in Satz 2.8 angegebene Algorithmus der konjugierten Gradientenmethode verursacht nur einen geringen numerischen Aufwand. Konkret ist in jeder Iteration nur einmal das Matrix-Vektor-Produkt $\mathbf{Q}\mathbf{d}_k$ auszurechnen. Der Aufwand dieser Operation reduziert sich deutlich, wenn $\mathbf{Q}$ dünn besetzt ist. Alle übrigen Rechenschritte der konjugierten Gradientenmethode sind elementare Vektoroperationen (inneres Produkt zweier Vektoren, Skalierung eines Vektors und Summe von Vektoren). Außerdem müssen mit Ausnahme der Matrix $\mathbf{Q}$ ausschließlich Vektoren und Skalare gespeichert werden. Alternativ zu (2.64a) kann die Schrittweite α_k auch mit einem der in Abschnitt 2.3.1 beschriebenen Verfahren bestimmt werden, sofern dieses tatsächlich die optimale Schrittweite liefert. Die explizite Berechnung des Gradienten $\mathbf{g}_k$ gemäß (2.60) kann durch eine iterative Berechnung gemäß (2.67) ersetzt werden. Abwandlungen des in Satz 2.8 angegebenen Algorithmus finden sich z. B. in [2.3, 2.5, 2.6].

Der nachfolgende Satz (siehe [2.4, 2.7]) gibt eine Aussage über das Konvergenzverhalten der konjugierten Gradientenmethode.

Satz 2.9 (Konvergenz der konjugierten Gradientenmethode — Quadratische Kostenfunktion)**.** *Für jeden Anfangswert $\mathbf{x}_0 \in \mathbb{R}^n$ konvergiert die konjugierte Gradientenmethode gemäß Satz 2.8 gegen das eindeutige globale Minimum $\mathbf{x}^*$ der Kostenfunktion $f(\mathbf{x})$ aus (2.54). Die Fehlerfunktion $e(\mathbf{x})$ gemäß (2.43) konvergiert dabei* linear *gegen* 0 mit der Konvergenzrate

$$e(\mathbf{x}_{k+1}) \leq \frac{\sqrt{\kappa} - 1}{\sqrt{\kappa} + 1} e(\mathbf{x}_k) \ , \tag{2.72}$$

wobei $\kappa = \lambda_{\max}/\lambda_{\min}$ wieder die spektrale Konditionszahl *der Matrix $\mathbf{Q}$ bezeichnet.*

Die konjugierte Gradientenmethode ist ein sogenanntes *Krylov-Unterraum-Verfahren*. Um dies zu sehen, wird zunächst der Begriff des Krylov-Unterraums definiert.

Definition 2.3 (Krylov-Unterraum)**.** Für eine quadratische Matrix $\mathbf{Q} \in \mathbb{R}^{n \times n}$ und einen Vektor $\mathbf{g} \in \mathbb{R}^n$ wird

$$\mathcal{K}_m(\mathbf{Q}, \mathbf{g}) = \mathrm{span}\{\mathbf{g}, \mathbf{Q}\mathbf{g}, \ldots, \mathbf{Q}^{m-1}\mathbf{g}\} \tag{2.73}$$

mit $m \leq n$ als Krylov-Unterraum der Ordnung m bezeichnet.

Mit einem Krylov-Unterraum-Verfahren wird das lineare Gleichungssystem $\mathbf{0} = \mathbf{Q}\mathbf{x} - \mathbf{b}$ ausgehend von einer Startlösung $\mathbf{x}_0$ iterativ so gelöst, dass $\mathbf{x}_k - \mathbf{x}_0$ im Krylov-Unterraum $\mathcal{K}_k(\mathbf{Q}, \mathbf{g}_0)$ mit $\mathbf{g}_0 = \mathbf{Q}\mathbf{x}_0 - \mathbf{b}$ liegt und eine Norm des Residuums $\mathbf{g}_k = \mathbf{Q}\mathbf{x}_k - \mathbf{b}$ minimiert. In jedem Iterationsschritt wird die Dimension des Krylov-Unterraums um eins erhöht.

Satz 2.10 (Konjugierte Gradientenmethode als Krylov-Unterraum-Verfahren)**.** *Die konjugierte Gradientenmethode gemäß Satz 2.8 ist ein Krylov-Unterraum-Verfahren bei dem ausgehend von einer Startlösung* $\mathbf{x}_0$ *in jedem Iterationsschritt* $k > 0$ *das Optimierungsproblem*

$$\min_{\mathbf{x}_k \in \mathbf{x}_0 + \mathcal{K}_k(\mathbf{Q}, \mathbf{g}_0)} \frac{1}{2}\mathbf{x}_k^{\mathrm{T}}\mathbf{Q}\mathbf{x}_k - \mathbf{x}_k^{\mathrm{T}}\mathbf{b} \tag{2.74}$$

mit einer symmetrisch positiv definiten Matrix $\mathbf{Q}$*, dem Krylov-Unterraum*

$$\begin{aligned}\mathcal{K}_k(\mathbf{Q}, \mathbf{g}_0) &= \operatorname{span}\{\mathbf{g}_0, \mathbf{Q}\mathbf{g}_0, \ldots, \mathbf{Q}^{k-1}\mathbf{g}_0\} \\ &= \operatorname{span}\{\mathbf{g}_0, \mathbf{g}_1, \ldots, \mathbf{g}_{k-1}\} = \operatorname{span}\{\mathbf{d}_0, \mathbf{d}_1, \ldots, \mathbf{d}_{k-1}\}\end{aligned} \tag{2.75}$$

und $\mathbf{g}_0 = \mathbf{Q}\mathbf{x}_0 - \mathbf{b}$ *gelöst wird.*

Aufgabe 2.8. Beweisen Sie (2.75). Dies ist besonders einfach mit Hilfe der vollständigen Induktion. Der Rest des Beweises von Satz 2.10 folgt aus den Überlegungen, die dem Satz 2.8 vorangegangenen sind.

Eine gelegentlich verwendete Erweiterung des Algorithmus nach Satz 2.8 ist die sogenannte *partielle konjugierte Gradientenmethode*. Dabei werden lediglich $m + 1 < n$ Iterationsschritte ausgeführt, ehe das Verfahren an dem so erhaltenen Punkt $\mathbf{x}_{m+1}$ neu initialisiert wird, d. h. es wird mit den Anfangswerten $\mathbf{x}_0 \leftarrow \mathbf{x}_{m+1}$ und $\mathbf{d}_0 = \mathbf{b} - \mathbf{Q}\mathbf{x}_0$ neu gestartet. Für das Konvergenzverhalten dieser Methode gilt folgender Satz, welcher z. B. in [2.2] bewiesen wird.

Satz 2.11 (Partielle konjugierte Gradientenmethode)**.** *Gegeben ist das Optimierungsproblem (2.54) mit der quadratischen Kostenfunktion* $f(\mathbf{x})$*. 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 nach jeweils* $m + 1$ *Schritten neu gestartet wird, für die Fehlerfunktion* $e(\mathbf{x})$ *gemäß (2.43) das Konvergenzverhalten*

$$e(\mathbf{x}_{k+1}) \leq \frac{r - l}{r + l} e(\mathbf{x}_k) \ . \tag{2.76}$$

Im Sinne dieses Satzes wird der Punkt $\mathbf{x}_{k+1}$ durch $m + 1$ Zwischeniterationen nach Satz 2.8 mit dem Anfangswert $\mathbf{x}_k$ erreicht. Für $m = 0$ (reine Gradientenmethode) liefert Satz 2.11 die gleichen Aussagen wie Satz 2.6.

Soll die konjugierte Gradientenmethode für *nichtquadratische Kostenfunktionen* $f(\mathbf{x})$ verwendet werden, so müssen in Satz 2.8 lediglich die Substitutionen

$$\mathbf{g}_k \leftrightarrow (\nabla f)(\mathbf{x}_k) \ , \qquad \mathbf{Q} \leftrightarrow \left(\nabla^2 f\right)(\mathbf{x}_k) \tag{2.77}$$

vorgenommen werden. In diesem Fall wird der Algorithmus im Allgemeinen nicht mehr nach spätestens n Schritten terminieren. Für nichtquadratische Kostenfunktionen $f(\mathbf{x})$ hat sich auch die Anwendung der partiellen konjugierten Gradientenmethode als günstig erwiesen, da die Eigenschaft der $\mathbf{Q}$-Orthogonalität der Suchrichtungen mit fortschreitenden Iterationen mehr und mehr verletzt sein kann. Um die aufwändige Berechnung der

Hessematrix $(\nabla^2 f)(\mathbf{x}_k)$ zu vermeiden, kann die Schrittweite α_k wieder alternativ mit einem der in Abschnitt 2.3.1 beschriebenen Verfahren festgelegt werden.

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

$$f(\mathbf{x}) \approx f(\mathbf{x}_k) + \mathbf{s}_k^{\mathrm{T}}(\nabla f)(\mathbf{x}_k) + \frac{1}{2}\mathbf{s}_k^{\mathrm{T}}\Big(\nabla^2 f\Big)(\mathbf{x}_k)\mathbf{s}_k \ , \tag{2.78}$$

siehe Abbildung 2.7. Die so genannte *Newton-Richtung* $\mathbf{s}_k$ ergibt sich unmittelbar durch Minimierung der rechten Seite von (2.78) bezüglich $\mathbf{s}_k$ in der Form

$$\mathbf{s}_k = -\Big(\nabla^2 f\Big)^{-1}(\mathbf{x}_k)(\nabla f)(\mathbf{x}_k) \ . \tag{2.79}$$

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 (2.79) 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, 2.8] zu finden.

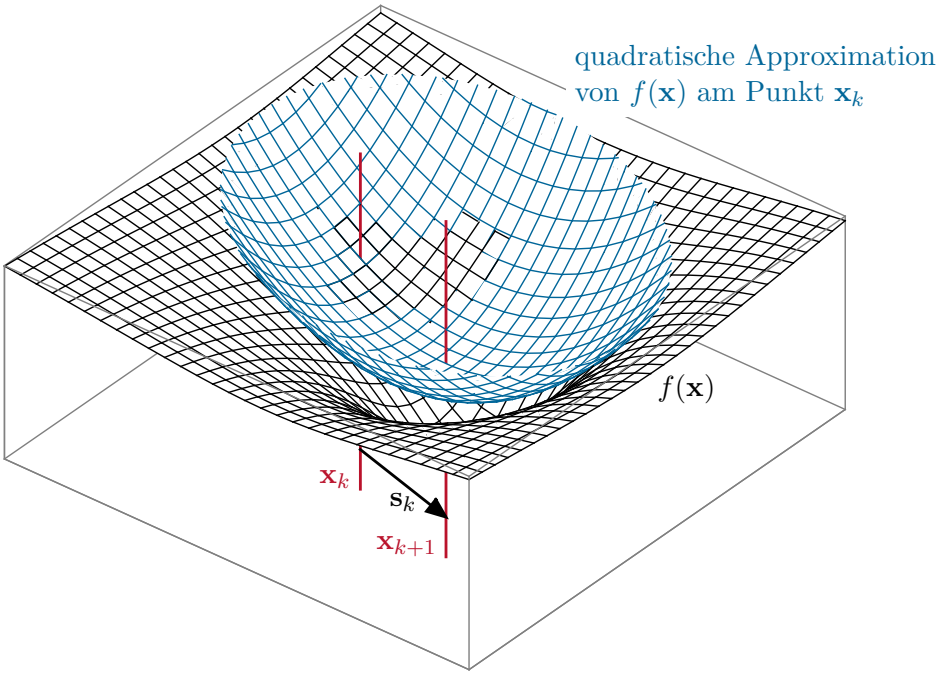


Abbildung 2.7: Lokale quadratische Approximation der Kostenfunktion im Zuge der Newton-Iteration (bei Schrittweite $\alpha_k = 1$).

Satz 2.12 (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

$$\mathbf{x}_{k+1} = \mathbf{x}_k - \Big(\nabla^2 f\Big)^{-1}(\mathbf{x}_k)(\nabla f)(\mathbf{x}_k) \tag{2.80}$$

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

Aufgabe 2.9. Zeigen Sie, dass ein Newton-Iterationsschritt gemäß (2.80) exakt einem Iterationsschritt der Newton-Raphson-Methode zur Lösung der Gleichung

$$(\nabla f)(\mathbf{x}^*) = \mathbf{0}\ , \tag{2.81}$$

also der notwendigen Optimalitätsbedingung erster Ordnung, entspricht.

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

$$\mathbf{x}_{k+1} = \mathbf{x}_k - \alpha_k\Big(\nabla^2 f\Big)^{-1}(\mathbf{x}_k)(\nabla f)(\mathbf{x}_k) \tag{2.82}$$

lautet und die Abstiegsbedingung (2.22) erfüllt ist. Das Verfahren wird dann häufig *gedämpfte Newton-Methode* genannt und α_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 eine iterative Schrittweitensuche mit dem Wert $\alpha_k = 1$ beginnt. Strategien zur Berechnung der Schrittweite α_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 (2.78) 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 (2.82) die Iterationsvorschrift

$$\mathbf{x}_{k+1} = \mathbf{x}_k - \alpha_k \mathbf{M}_k (\nabla f)(\mathbf{x}_k) \ , \quad \mathbf{M}_k = \Big(\Big(\nabla^2 f \Big)(\mathbf{x}_k) + \varepsilon_k \mathbf{E} \Big)^{-1} \tag{2.83}$$

mit einem geeigneten positiven Parameter ε_k verwendet werden. Die Iterationsvorschrift (2.83) geht für $\varepsilon_k = 0$ in die Newton-Methode gemäß (2.80) und für sehr große ε_k in die Gradientenmethode gemäß (2.51) über. Eine geeignete Wahl von ε_k erweist sich jedoch als nicht sehr einfach. Typischerweise wird ε_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.

Aufgabe 2.10. Zeigen Sie, dass die Newton-Methode für quadratische Kostenfunktionen

$$\min_{\mathbf{x} \in \mathbb{R}^n} \ \frac{1}{2} \mathbf{x}^{\mathrm{T}} \mathbf{Q} \mathbf{x} - \mathbf{b}^{\mathrm{T}} \mathbf{x} \tag{2.84}$$

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 $\mathbf{x}^*$ zumeist der Fall ist
+ Konvergenzverhalten besser als jenes der konjugierten Gradientenmethode
– nicht geeignet in Gebieten in denen $(\nabla^2 f)(\mathbf{x}_k)$ nicht positiv definit ist
– 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

$$\mathbf{g}_{k+1} - \mathbf{g}_k \approx \Big(\nabla^2 f \Big)(\mathbf{x}_k) \mathbf{p}_k \tag{2.85}$$

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

$$\mathbf{q}_k = \mathbf{g}_{k+1} - \mathbf{g}_k = \mathbf{K} \mathbf{p}_k \ . \tag{2.86}$$

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

$$\mathbf{K} = \begin{bmatrix} \mathbf{q}_0 & \mathbf{q}_1 & \ldots & \mathbf{q}_{n-1} \end{bmatrix} \begin{bmatrix} \mathbf{p}_0 & \mathbf{p}_1 & \ldots & \mathbf{p}_{n-1} \end{bmatrix}^{-1} \tag{2.87}$$

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

$$\mathbf{H}_{k+1}\mathbf{q}_j = \mathbf{p}_j, \quad j = 0, \ldots, k \tag{2.88}$$

zu konstruieren, woraus sich $\mathbf{H}_n = \mathbf{K}^{-1}$ ergibt. 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

$$\mathbf{H}_{k+1} = \mathbf{H}_k + \mathbf{z}_k \mathbf{z}_k^{\mathrm{T}} \tag{2.89}$$

anzusetzen. Das dyadische Produkt $\mathbf{z}_k \mathbf{z}_k^{\mathrm{T}}$ erhält die Symmetrie und hat höchstens den Rang 1, weshalb diese Korrektur auch als *Rang 1 Korrektur* bezeichnet wird. Setzt man (2.89) in (2.88) ein, so erhält man für $j = k$

$$\mathbf{p}_k = \mathbf{H}_{k+1}\mathbf{q}_k = \mathbf{H}_k\mathbf{q}_k + \mathbf{z}_k \mathbf{z}_k^{\mathrm{T}}\mathbf{q}_k \tag{2.90}$$

und nach Skalarmultiplikation mit $\mathbf{q}_k$

$$\mathbf{q}_k^{\mathrm{T}}\mathbf{p}_k = \mathbf{q}_k^{\mathrm{T}}\mathbf{H}_k\mathbf{q}_k + (\mathbf{z}_k^{\mathrm{T}}\mathbf{q}_k)^2 \ . \tag{2.91}$$

Aus (2.90) und (2.91) folgt

$$\begin{aligned} (\mathbf{p}_k - \mathbf{H}_k\mathbf{q}_k)(\mathbf{p}_k - \mathbf{H}_k\mathbf{q}_k)^{\mathrm{T}} &= \mathbf{z}_k \mathbf{z}_k^{\mathrm{T}}\mathbf{q}_k(\mathbf{z}_k \mathbf{z}_k^{\mathrm{T}}\mathbf{q}_k)^{\mathrm{T}} = \mathbf{z}_k \mathbf{z}_k^{\mathrm{T}}\mathbf{q}_k\mathbf{q}_k^{\mathrm{T}}\mathbf{z}_k \mathbf{z}_k^{\mathrm{T}} \\ &= \mathbf{z}_k \mathbf{z}_k^{\mathrm{T}}(\mathbf{z}_k^{\mathrm{T}}\mathbf{q}_k)^2 = \mathbf{z}_k \mathbf{z}_k^{\mathrm{T}}\mathbf{q}_k^{\mathrm{T}}(\mathbf{p}_k - \mathbf{H}_k\mathbf{q}_k) \ . \end{aligned} \tag{2.92}$$

Wird der daraus resultierende Ausdruck für $\mathbf{z}_k \mathbf{z}_k^{\mathrm{T}}$ in (2.89) eingesetzt, so ergibt sich

$$\mathbf{H}_{k+1} = \mathbf{H}_k + \frac{(\mathbf{p}_k - \mathbf{H}_k\mathbf{q}_k)(\mathbf{p}_k - \mathbf{H}_k\mathbf{q}_k)^{\mathrm{T}}}{\mathbf{q}_k^{\mathrm{T}}(\mathbf{p}_k - \mathbf{H}_k\mathbf{q}_k)} \ . \tag{2.93}$$

Damit lässt sich folgender Satz formulieren.

Satz 2.13 (Quasi-Newton-Methode — Rang 1 Korrektur). *Angenommen $\mathbf{K}$ ist eine konstante symmetrische Matrix und $\mathbf{p}_0, \mathbf{p}_1, \ldots, \mathbf{p}_k$ sind linear unabhängige Vektoren. Mit $\mathbf{q}_j = \mathbf{K}\mathbf{p}_j$, $j = 0, \ldots, k$ gilt für jede symmetrische Startmatrix $\mathbf{H}_0$ und die Iterationsvorschrift*

$$\mathbf{H}_{j+1} = \mathbf{H}_j + \frac{(\mathbf{p}_j - \mathbf{H}_j\mathbf{q}_j)(\mathbf{p}_j - \mathbf{H}_j\mathbf{q}_j)^{\mathrm{T}}}{\mathbf{q}_j^{\mathrm{T}}(\mathbf{p}_j - \mathbf{H}_j\mathbf{q}_j)} \tag{2.94}$$

die Beziehung

$$\mathbf{p}_j = \mathbf{H}_{k+1}\mathbf{q}_j, \quad j = 0, \ldots, k \ . \tag{2.95}$$

Aufgabe 2.11. Beweisen Sie Satz 2.13 mit vollständiger Induktion. **Hinweis:** Dieser Beweis ist z. B. in [2.8] skizziert.

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

$$\mathbf{H}_{k+1} = \mathbf{H}_k + \frac{\mathbf{p}_k\mathbf{p}_k^{\mathrm{T}}}{\mathbf{q}_k^{\mathrm{T}}\mathbf{p}_k} - \frac{\mathbf{H}_k\mathbf{q}_k\mathbf{q}_k^{\mathrm{T}}\mathbf{H}_k}{\mathbf{q}_k^{\mathrm{T}}\mathbf{H}_k\mathbf{q}_k} \tag{2.96}$$

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

$$\mathbf{H}_{k+1} = \left(\mathbf{E} - \frac{\mathbf{p}_k\mathbf{q}_k^{\mathrm{T}}}{\mathbf{q}_k^{\mathrm{T}}\mathbf{p}_k}\right)\mathbf{H}_k\left(\mathbf{E} - \frac{\mathbf{q}_k\mathbf{p}_k^{\mathrm{T}}}{\mathbf{q}_k^{\mathrm{T}}\mathbf{p}_k}\right) + \frac{\mathbf{p}_k\mathbf{p}_k^{\mathrm{T}}}{\mathbf{q}_k^{\mathrm{T}}\mathbf{p}_k} \ . \tag{2.97}$$

Beide werden als *Rang* 2 *Korrekturformeln* bezeichnet, da die aktuelle Approximation der inversen Hessematrix jeweils durch Addition einer Matrix mit Rang 2 korrigiert wird. Linearkombinationen der obigen beiden Formeln in der Art $\mathbf{H}_{k+1} = \phi\mathbf{H}_{k+1}^{\mathrm{DFP}} + (1-\phi)\mathbf{H}_{k+1}^{\mathrm{BFGS}}$ 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, 2.3, 2.4]), dass sie die positive Definitheit von $\mathbf{H}_k$ erhalten, wenn

$$\mathbf{q}_k^{\mathrm{T}}\mathbf{p}_k > 0 \tag{2.98}$$

erfüllt ist.

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

$$\mathbf{s}_k = -\mathbf{H}_k(\nabla f)(\mathbf{x}_k) \tag{2.99}$$

gewählt. Man beachte, dass hierfür, anders als bei der Newton-Methode (siehe (2.79)), 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.

Die Erfüllung der Bedingung (2.98) lässt sich durch eine geeignete Wahl der Schrittweite $\alpha_k > 0$ in $\mathbf{x}_{k+1} = \mathbf{x}_k + \alpha_k\mathbf{s}_k$ bzw. $\mathbf{p}_k = \mathbf{x}_{k+1} - \mathbf{x}_k = \alpha_k\mathbf{s}_k$ sicherstellen. Genügt die Wahl der Schrittweite α_k beispielsweise der Wolfe-Bedingung gemäß Abschnitt 2.3.1.3, so ist (2.98) automatisch erfüllt. Um dies zu sehen, beachte man, dass aus (2.33)

$$g'(\alpha_k) = \mathbf{g}_{k+1}^{\mathrm{T}}\mathbf{s}_k \geq \varepsilon_2 g'(0) = \varepsilon_2\mathbf{g}_k^{\mathrm{T}}\mathbf{s}_k \tag{2.100}$$

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

$$\mathbf{q}_k^{\mathrm{T}}\mathbf{p}_k = \alpha_k\mathbf{q}_k^{\mathrm{T}}\mathbf{s}_k = \alpha_k(\mathbf{g}_{k+1} - \mathbf{g}_k)^{\mathrm{T}}\mathbf{s}_k \geq \underbrace{\alpha_k(\varepsilon_2 - 1)\mathbf{g}_k^{\mathrm{T}}\mathbf{s}_k}_{>0} \ , \tag{2.101}$$

Initialisierung: $\mathbf{H}_0$ (Startwert, positiv definite Matrix)
$k = 0$ (Startindex)
$\mathbf{x}_0$ (Startlösung)
$\mathbf{g}_0 = (\nabla f)(\mathbf{x}_0)$ (Gradient an der Stelle $\mathbf{x}_0$)
ε_x (Schwellwert für Abbruchkriterium)

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

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

$k \leftarrow k + 1$

until $\|\mathbf{x}_{k+1} - \mathbf{x}_k\| \leq \varepsilon_x$

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

wobei die rechte Seite dieser Ungleichung (abseits des optimalen Punktes $\mathbf{x}^*$) strikt positiv sein muss, da $\mathbf{s}_k$ eine Abstiegsrichtung ist. Somit ist (2.98) erfüllt.

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
+ meist 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

$$f(\mathbf{x}) = \frac{1}{2}\|\mathbf{r}(\mathbf{x})\|_2^2 = \frac{1}{2}\sum_{i=1}^{m} r_i^2(\mathbf{x}) \tag{2.102}$$

mit einer beliebigen Funktion $\mathbf{r} : \mathbb{R}^n \to \mathbb{R}^m$ 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

$$\Big(\nabla^2 f\Big)(\mathbf{x}) = (\nabla \mathbf{r})(\mathbf{x})(\nabla \mathbf{r})^{\mathrm{T}}(\mathbf{x}) + \sum_{i=1}^{m} r_i(\mathbf{x})\Big(\nabla^2 r_i\Big)(\mathbf{x}) \ , \tag{2.103}$$

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.

Aufgabe 2.12. Rechnen Sie (2.103) ausgehend von (2.102) nach.

Bei der Gauss-Newton-Methode wird der zweite Summand in (2.103) vernachlässigt, so dass sich die Näherung

$$\Big(\nabla^2 f\Big)(\mathbf{x}) \approx (\nabla \mathbf{r})(\mathbf{x})(\nabla \mathbf{r})^{\mathrm{T}}(\mathbf{x}) \tag{2.104}$$

ergibt. Der damit verbundene Näherungsfehler ist also klein, wenn $r_i(\mathbf{x})$ oder $(\nabla^2 r_i)(\mathbf{x})$ $\forall\, i = 1, \ldots, m$ betraglich kleine Werte annimmt. Ist z. B. bekannt, dass $f(\mathbf{x}^*) = 0$, so folgt daraus $r_i(\mathbf{x}^*) = 0$ $\forall\, i = 1, \ldots, m$. Der Fall $(\nabla^2 r_i)(\mathbf{x}) = \mathbf{0}$ $\forall\, i = 1, \ldots, m$ tritt ein, wenn $\mathbf{r}(\mathbf{x})$ affin in $\mathbf{x}$ ist (vgl. Aufgabe 2.13).

Unter Verwendung des Gradienten $(\nabla f)(\mathbf{x}) = (\nabla \mathbf{r})(\mathbf{x})\mathbf{r}(\mathbf{x})$ und der Approximation (2.104) ergibt sich der Gauss-Newton-Schritt

$$\begin{aligned} \mathbf{s}_k &= -((\nabla \mathbf{r})(\mathbf{x}_k)(\nabla \mathbf{r})^{\mathrm{T}}(\mathbf{x}_k))^{-1}(\nabla f)(\mathbf{x}_k) \\ &= -((\nabla \mathbf{r})(\mathbf{x}_k)(\nabla \mathbf{r})^{\mathrm{T}}(\mathbf{x}_k))^{-1}(\nabla \mathbf{r})(\mathbf{x}_k)\mathbf{r}(\mathbf{x}_k) \ . \end{aligned} \tag{2.105}$$

Die Iterationsvorschrift der Gauss-Newton-Methode lautet damit

$$\mathbf{x}_{k+1} = \mathbf{x}_k - ((\nabla \mathbf{r})(\mathbf{x}_k)(\nabla \mathbf{r})^{\mathrm{T}}(\mathbf{x}_k))^{-1}(\nabla \mathbf{r})(\mathbf{x}_k)\mathbf{r}(\mathbf{x}_k) \ . \tag{2.106}$$

Mit der Gauss-Newton-Methode wird also im Iterationsschritt k die lokale quadratische Approximation

$$f(\mathbf{x}_k) + (\nabla f)^{\mathrm{T}}(\mathbf{x}_k)(\mathbf{x} - \mathbf{x}_k) + \frac{1}{2}(\mathbf{x} - \mathbf{x}_k)^{\mathrm{T}}(\nabla \mathbf{r})(\mathbf{x}_k)(\nabla \mathbf{r})^{\mathrm{T}}(\mathbf{x}_k)(\mathbf{x} - \mathbf{x}_k) \tag{2.107}$$

von $f(\mathbf{x})$ am Punkt $\mathbf{x}_k$ bezüglich $\mathbf{x}$ exakt minimiert. Für die praktische Anwendung der Methode führt man in (2.106) noch eine geeignete Schrittweite α_k ein, so dass die Iterationsvorschrift

$$\mathbf{x}_{k+1} = \mathbf{x}_k - \alpha_k((\nabla \mathbf{r})(\mathbf{x}_k)(\nabla \mathbf{r})^{\mathrm{T}}(\mathbf{x}_k))^{-1}(\nabla \mathbf{r})(\mathbf{x}_k)\mathbf{r}(\mathbf{x}_k) \tag{2.108}$$

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

Man beachte, dass die Berechnung von $\mathbf{s}_k$ gemäß (2.105) keine tatsächliche Inversion von $(\nabla\mathbf{r})(\mathbf{x}_k)(\nabla\mathbf{r})^{\mathrm{T}}(\mathbf{x}_k)$ erfordert. Die Wahl der Suchrichtung gemäß (2.105) ist nur sinnvoll, wenn die $n \times n$ Matrix $(\nabla\mathbf{r})(\mathbf{x})(\nabla\mathbf{r})^{\mathrm{T}}(\mathbf{x})$ positiv definit ist. Ist sie nicht positiv definit (oder schlecht konditioniert), kann ähnlich wie bei der Newton-Methode die alternative Iterationsvorschrift

$$\mathbf{x}_{k+1} = \mathbf{x}_k - \alpha_k \mathbf{M}_k (\nabla f)(\mathbf{x}_k) \;, \quad \mathbf{M}_k = \Big((\nabla\mathbf{r})(\mathbf{x}_k)(\nabla\mathbf{r})^{\mathrm{T}}(\mathbf{x}_k) + \varepsilon_k \mathbf{E} \Big)^{-1} \tag{2.109}$$

mit einem geeigneten positiven Parameter $\varepsilon_k \geq 0$ verwendet werden. Das Verfahren wird dann (für den Fall $\alpha_k = 1$) als *Levenberg-Marquardt-Methode* bezeichnet. Die Iterationsvorschrift (2.109) geht für $\varepsilon_k \to 0$ in die Gauss-Newton-Methode gemäß (2.108) und für sehr große ε_k in die Gradientenmethode gemäß (2.51) über. Die Levenberg-Marquardt-Methode kombiniert daher das Verhalten der Gauss-Newton-Methode mit jenem der Gradientenmethode. Häufig wird dabei der Parameter ε_k ausgehend von einem hohen Anfangswert heuristisch adaptiert. Dazu wird die tatsächliche Kostenfunktionsänderung $f(\mathbf{x}_{k+1}) - f(\mathbf{x}_k)$ betrachtet. Ist sie positiv (kein Abstieg), so wird der Schritt verworfen und ε_k vergrößert (Verhalten wird der Gradientenmethode ähnlicher). Ist sie negativ, so wird der Schritt akzeptiert. Stimmt sie überdies gut mit der gemäß (2.107) prädizierten (negativen) Kostenfunktionsänderung $(\nabla f)^{\mathrm{T}}(\mathbf{x}_k)(\mathbf{x}_{k+1} - \mathbf{x}_k) + \frac{1}{2}(\mathbf{x}_{k+1} - \mathbf{x}_k)^{\mathrm{T}}(\nabla\mathbf{r})(\mathbf{x}_k)(\nabla\mathbf{r})^{\mathrm{T}}(\mathbf{x}_k)(\mathbf{x}_{k+1} - \mathbf{x}_k)$ überein, wird ε_k verkleinert (Verhalten wird der Gauss-Newton-Methode ähnlicher). Eine detaillierter dargestellte, ähnliche Heuristik findet sich in Abschnitt 2.4.

Der nachfolgende Satz liefert eine Aussage über die Konvergenzordnung der Gauss-Newton-Methode. Sein Beweis ist z. B. in [2.9] zu finden.

Satz 2.14 (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

$$\mathbf{x}_{k+1} = \mathbf{x}_k - \big((\nabla\mathbf{r})(\mathbf{x}_k)(\nabla\mathbf{r})^{\mathrm{T}}(\mathbf{x}_k) \big)^{-1} (\nabla\mathbf{r})(\mathbf{x}_k)\mathbf{r}(\mathbf{x}_k) \tag{2.110}$$

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 $(\nabla\mathbf{r})(\mathbf{x}_k)(\nabla\mathbf{r})^{\mathrm{T}}(\mathbf{x}_k)$, so kann zumindest lineares Konvergenzverhalten nachgewiesen werden. Wenn jedoch $\sum_{i=1}^m r_i(\mathbf{x}^*)(\nabla^2 r_i)(\mathbf{x}^*)$ betraglich zu große Werte annimmt, kann es sein, dass die Gauss-Newton-Methode nicht konvergiert [2.9].

Aufgabe 2.13. Zeigen Sie, dass die Gauss-Newton-Methode für die Optimierungsaufgabe

$$\min_{\mathbf{x}\in\mathbb{R}^n} \; \frac{1}{2}\|\mathbf{r}(\mathbf{x})\|_2^2 \qquad \text{mit} \qquad \mathbf{r}(\mathbf{x}) = \mathbf{A}\mathbf{x} + \mathbf{b} \tag{2.111}$$

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 $\mathbf{x}^* = -(\mathbf{A}^\mathrm{T}\mathbf{A})^{-1}\mathbf{A}^\mathrm{T}\mathbf{b}$ konvergiert.

Kostenfunktionen wie in der Optimierungsaufgabe (2.111) treten z. B. bei der Methode der kleinsten Fehlerquadrate mit parametrisch linearem Modell (lineare Regressionsanalyse) auf. Kostenfunktionen der Art (2.102) treten z. B. bei der Methode der kleinsten Fehlerquadrate mit parametrisch nichtlinearem Modell (nichtlineare Regressionsanalyse) [2.10] auf.

Aufgabe 2.14. Zeigen Sie, dass die iterative Lösung der nichtlinearen Gleichung

$$\mathbf{r}(\mathbf{x}) = \mathbf{0} \tag{2.112}$$

mit $\mathbf{r} : \mathbb{R}^n \to \mathbb{R}^n$ und $\mathbf{r} \in C^1$ gemäß dem Newton-Raphson-Verfahren genau der Gauss-Newton-Methode angewandt auf die Optimierungsaufgabe

$$\min_{\mathbf{x}\in\mathbb{R}^n} \ \frac{1}{2}\|\mathbf{r}(\mathbf{x})\|_2^2 \tag{2.113}$$

entspricht.

Lösung von Aufgabe 2.14. Damit die Gleichung (2.112) eine eindeutige Lösung $\mathbf{x}^*$ besitzt, muss $(\nabla\mathbf{r})^\mathrm{T}(\mathbf{x}^*)$ regulär sein. Dies muss wegen $\mathbf{r} \in C^1$ auch für $(\nabla\mathbf{r})^\mathrm{T}(\mathbf{x}_k)$ in einer Umgebung von $\mathbf{x}^*$ gelten. Die Iterationsvorschrift beim Newton-Raphson-Verfahren lautet

$$\mathbf{x}_{k+1} = \mathbf{x}_k - ((\nabla\mathbf{r})^\mathrm{T}(\mathbf{x}_k))^{-1}\mathbf{r}(\mathbf{x}_k) \ . \tag{2.114}$$

Die Iterationsvorschrift (2.106) der Gauss-Newton-Methode vereinfacht sich bei regulärem $(\nabla\mathbf{r})^\mathrm{T}(\mathbf{x}_k)$ zu

$$\begin{aligned} \mathbf{x}_{k+1} &= \mathbf{x}_k - ((\nabla\mathbf{r})(\mathbf{x}_k)(\nabla\mathbf{r})^\mathrm{T}(\mathbf{x}_k))^{-1}(\nabla\mathbf{r})(\mathbf{x}_k)\mathbf{r}(\mathbf{x}_k) \\ &= \mathbf{x}_k - ((\nabla\mathbf{r})^\mathrm{T}(\mathbf{x}_k))^{-1}\mathbf{r}(\mathbf{x}_k) \ . \end{aligned} \tag{2.115}$$

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

+ keine zweiten Ableitungen nötig, nur die Jacobi-Matrix $(\nabla\mathbf{r})(\mathbf{x}_k)$ 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 (2.79) 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

$$f(\mathbf{x}_k + \mathbf{s}_k) \approx m_k(\mathbf{s}_k) = f(\mathbf{x}_k) + \mathbf{s}_k^{\mathrm{T}}(\nabla f)(\mathbf{x}_k) + \frac{1}{2}\mathbf{s}_k^{\mathrm{T}}\mathbf{B}_k\mathbf{s}_k \tag{2.116}$$

mit einer geeigneten symmetrischen Matrix $\mathbf{B}_k$ approximiert. Grundsätzlich sollte $\mathbf{B}_k$ eine mit geringem Aufwand berechenbare Näherung der Hessematrix $(\nabla^2 f)(\mathbf{x}_k)$ sein. Der mit $m_k(\mathbf{s}_k)$ verbundene Approximationsfehler 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 $\|\mathbf{s}_k\|_2 \leq \Delta_k$ mit einem skalaren Parameter Δ_k charakterisiert und definiert eine 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 *beschränkte* Optimierungsproblem

$$\min_{\mathbf{s}_k \in \mathbb{R}^n} \quad m_k(\mathbf{s}_k) = f(\mathbf{x}_k) + \mathbf{s}_k^{\mathrm{T}}(\nabla f)(\mathbf{x}_k) + \frac{1}{2}\mathbf{s}_k^{\mathrm{T}}\mathbf{B}_k\mathbf{s}_k \tag{2.117a}$$

$$\text{u.B.v.} \quad \|\mathbf{s}_k\|_2 \leq \Delta_k \tag{2.117b}$$

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 Δ_k. Dazu wird in jedem Iterationsschritt *die Übereinstimmung der quadratischen Ansatzfunktion m_k mit der Kostenfunktion f* überprüft, indem das Verhältnis

$$\rho_k(\mathbf{s}_k) = \frac{f(\mathbf{x}_k) - f(\mathbf{x}_k + \mathbf{s}_k)}{m_k(\mathbf{0}) - m_k(\mathbf{s}_k)} \tag{2.118}$$

berechnet wird. Der Zählerterm in (2.118) beschreibt die *tatsächliche Reduktion* der Kostenfunktion, während der Nennerterm die *prädizierte Reduktion* wiedergibt. Dieser Nennerterm ist stets größer gleich Null, da $\mathbf{s}_k$ die Funktion m_k gemäß (2.117) innerhalb des Vertrauensbereiches minimiert. 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.

Ein Algorithmus zur Methode der Vertrauensbereiche ist in Tabelle 2.4 aufgelistet. Man beachte, dass hier $\bar{\Delta}$ eine obere Schranke für Δ_k darstellt und dass eine Vergrößerung des

Initialisierung: $\bar{\Delta}, \Delta_0 \in (0, \bar{\Delta})$ (Vertrauensbereich: Grenz- & Startwert)
$\eta \in [0, \frac{1}{4})$ (Parameter)
$k \leftarrow 0$ (Startindex)
$\mathbf{x}_0$ (Startlösung)
ε_x (Schwellwert für Abbruchkriterium)

repeat

$m_k(\mathbf{s}_k)$ nach (2.116) (Modell)
$\mathbf{s}_k$ Lösung von (2.117) (evtl. approximativ gelöst)
ρ_k nach (2.118) (Modellgüte)

if $\rho_k < \frac{1}{4}$
$\Delta_{k+1} \leftarrow \frac{1}{4}\Delta_k$ (Reduktion)
else if $\rho_k > \frac{3}{4}$ **and** $||\mathbf{s}_k||_2 = \Delta_k$
$\Delta_{k+1} \leftarrow \min\{2\Delta_k, \bar{\Delta}\}$ (Vergrößerung)
else
$\Delta_{k+1} \leftarrow \Delta_k$
end if

if $\rho_k > \eta$
$\mathbf{x}_{k+1} \leftarrow \mathbf{x}_k + \mathbf{s}_k$ (nächster Schritt)
$\mathbf{B}_{k+1} \leftarrow \mathbf{B}_k + \ldots$ (Approximation der Hessematrix)
else
$\mathbf{x}_{k+1} \leftarrow \mathbf{x}_k$ (Schritt mit $\Delta_{k+1} < \Delta_k$ wiederholen)
end if

$k \leftarrow k + 1$

until $\|\mathbf{x}_k - \mathbf{x}_{k-1}\|_2 \leq \varepsilon_x$

Tabelle 2.4: Methode der Vertrauensbereiche.

Vertrauensbereiches im nächsten Iterationsschritt nur dann stattfindet, wenn $\mathbf{s}_k$ durch die Grenze des Vertrauensbereiches beschränkt wurde (Bedingung $\|\mathbf{s}_k\|_2 = \Delta_k$).

Vorschläge für die in Tabelle 2.4 nicht dargestellte Iterationsvorschrift der Matrix $\mathbf{B}_k$ werden z. B. in [2.1, 2.4, 2.8] beschrieben. Wenn sich die Kostenfunktion $f(\mathbf{x}_k)$ in der Form (2.102), d. h. $f(\mathbf{x}) = \frac{1}{2}\|\mathbf{r}(\mathbf{x})\|_2^2$ darstellen lässt, kann die Berechnung von $\mathbf{B}_k$ wie bei der Gauss-Newton-Methode in der Form (2.104) erfolgen, d. h. $\mathbf{B}_k = (\nabla\mathbf{r})(\mathbf{x}_k)(\nabla\mathbf{r})^{\mathrm{T}}(\mathbf{x}_k)$. Es wird nun gezeigt, dass bei dieser Wahl die Methode der Vertrauensbereiche exakt mit der Levenberg-Marquardt-Methode übereinstimmt [2.9]. Dazu wird die Ungleichungsbe-

schränkung in (2.117) äquivalent umgeschrieben, womit sich das Optimierungsproblem

$$\min_{\mathbf{s}_k \in \mathbb{R}^n} \quad m_k(\mathbf{s}_k) = f(\mathbf{x}_k) + \mathbf{s}_k^{\mathrm{T}}(\nabla f)(\mathbf{x}_k) + \frac{1}{2}\mathbf{s}_k^{\mathrm{T}}(\nabla \mathbf{r})(\mathbf{x}_k)(\nabla \mathbf{r})^{\mathrm{T}}(\mathbf{x}_k)\mathbf{s}_k \tag{2.119a}$$

$$\text{u.B.v.} \quad h(\mathbf{s}_k) = \frac{1}{2}\left(\mathbf{s}_k^{\mathrm{T}}\mathbf{s}_k - \Delta_k^2\right) \leq 0 \tag{2.119b}$$

ergibt. Basierend auf Satz 3.10 lauten notwendige Optimalitätsbedingungen für das *beschränkte* Optimierungsproblem (2.119)

$$(\nabla f)(\mathbf{x}_k) + (\nabla \mathbf{r})(\mathbf{x}_k)(\nabla \mathbf{r})^{\mathrm{T}}(\mathbf{x}_k)\mathbf{s}_k^* + \varepsilon_k^*\mathbf{s}_k^* = \mathbf{0} \tag{2.120a}$$

$$\varepsilon_k^* \geq 0 \tag{2.120b}$$

$$\varepsilon_k^* h(\mathbf{s}_k^*) = 0\ . \tag{2.120c}$$

mit einem noch zu bestimmenden Multiplikator ε_k^*. Aus (2.119b) und (2.120) lassen sich $\mathbf{s}_k^*$ und ε_k^* eindeutig berechnen. Für $\mathbf{s}_k^*$ folgt

$$\mathbf{s}_k^* = -\left((\nabla \mathbf{r})(\mathbf{x}_k)(\nabla \mathbf{r})^{\mathrm{T}}(\mathbf{x}_k) + \varepsilon_k^*\mathbf{E}\right)^{-1}(\nabla f)(\mathbf{x}_k)\ , \tag{2.121}$$

was genau der Levenberg-Marquardt-Methode (vgl. (2.109) mit $\alpha_k = 1$) entspricht. Bei der Methode der Vertrauensbereiche wird der Parameter Δ_k gewählt. Bei der Levenberg-Marquardt-Methode wird der Parameter ε_k gewählt. Diese Parameter lassen sich über (2.119b) und (2.120) ineinander umrechnen und es gilt $\varepsilon_k = 0$ für $\Delta_k \to \infty$ sowie $\varepsilon_k \to \infty$ für $\Delta_k \to 0$.

Mit der in Tabelle 2.5 zusammengefassten sogenannten *Dogleg-Methode* wird eine näherungsweise Lösung des Optimierungsproblems (2.117) berechnet. Sie kombiniert dabei das Verhalten der Newton-Methode mit jenem der Gradientenmethode. Wenn mit einem unbeschränkten Gradientenschritt der Vertrauensbereich verlassen würde, wird $\mathbf{s}_k$ durch Einschränkung des Gradientenschrittes auf den Vertrauensbereich gewählt. Wenn mit einem unbeschränkten Newtonschritt der Vertrauensbereich nicht verlassen wird, wird $\mathbf{s}_k$ entsprechend diesem Schritt gewählt. Andernfalls wird zwischen dem unbeschränkten Gradientenschritt und dem unbeschränkten Newtonschritt genau so linear interpoliert, dass $\mathbf{x}_{k+1} = \mathbf{x}_k + \mathbf{s}_k$ am Rand des Vertrauensbereiches zu liegen kommt.

Für ein Beispiel mit $n = 2$ zeigt Abbildung 2.8 die Niveaulinien einer lokalen quadratischen Ansatzfunktion $m_k(\mathbf{s}_k)$ in einer Umgebung von $\mathbf{x}_k$. Die grüne Linie beinhaltet die Näherungslösungen $\mathbf{x}_{k+1} = \mathbf{x}_k + \mathbf{s}_k$ der Dogleg Methode für verschiedene Werte $\Delta_k \in (0, \infty)$. Zum Vergleich zeigt die rote Linie die exakten Lösungen des Optimierungsproblems (2.117) für verschiedene Werte $\Delta_k \in (0, \infty)$. Im Fall $\mathbf{B}_k = (\nabla \mathbf{r})(\mathbf{x}_k)(\nabla \mathbf{r})^{\mathrm{T}}(\mathbf{x}_k)$ stimmt die rote Linie natürlich auch mit den Lösungen der Levenberg-Marquardt-Methode überein.

Bemerkung 2.1. Die Lösungen einer nichtlinearen Gleichung

$$\mathbf{r}(\mathbf{x}) = \mathbf{0} \tag{2.122}$$

$m_k(\mathbf{s}_k)$ nach (2.116) (Modell)

$\mathbf{s}_{G,k} = -\frac{(\nabla f)(\mathbf{x}_k)^{\mathrm{T}}(\nabla f)(\mathbf{x}_k)}{(\nabla f)(\mathbf{x}_k)^{\mathrm{T}}\mathbf{B}_k(\nabla f)(\mathbf{x}_k)}(\nabla f)(\mathbf{x}_k)$ (unbeschränkter Gradientenschritt)

if $\|\mathbf{s}_{G,k}\|_2 \geq \Delta_k$

  $\mathbf{s}_k = \Delta_k \frac{\mathbf{s}_{G,k}}{\|\mathbf{s}_{G,k}\|_2}$

else

  $\mathbf{s}_{N,k} = -\mathbf{B}_k^{-1}(\nabla f)(\mathbf{x}_k)$ (unbeschränkter Newtonschritt)

  if $\|\mathbf{s}_{N,k}\|_2 \leq \Delta_k$

    $\mathbf{s}_k = \mathbf{s}_{N,k}$

  else

    $\mathbf{s}_k = \beta_k \mathbf{s}_{G,k} + (1-\beta_k)\mathbf{s}_{N,k}$ (Interpolation)

    mit $\beta_k \in (0,1)$ so, dass $\|\mathbf{s}_k\|_2 = \Delta_k$

  end if

end if

Tabelle 2.5: Näherungsweise Lösung von (2.117) mittels Dogleg-Methode.

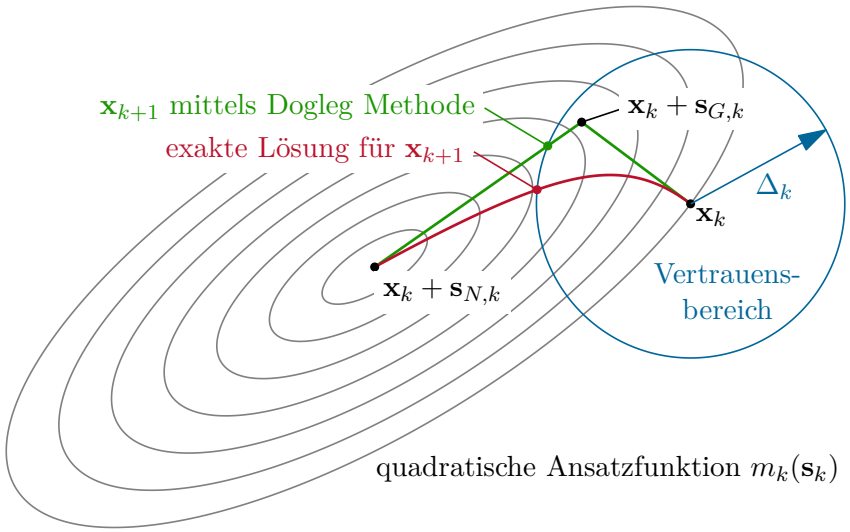


Abbildung 2.8: Lösung des Optimierungsproblems (2.117) im Rahmen einer Iteration der Methode der Vertrauensbereiche, Vergleich zwischen exakter Lösung (rot) und Dogleg Methode (grün).

mit $\mathbf{r} : \mathbb{R}^n \to \mathbb{R}^n$ können durch Lösen der Optimierungsaufgabe

$$\min_{\mathbf{x}\in\mathbb{R}^n} \frac{1}{2}\|\mathbf{r}(\mathbf{x})\|_2^2 \tag{2.123}$$

berechnet werden (siehe dazu auch Aufgabe 2.14). Diese Möglichkeit zur Lösung nichtlinearer Gleichungen wird von der MATLAB-Funktion fsolve aus der Optimization Toolbox genutzt. Die Funktion fsolve verwendet standardmäßig die Methode der Vertrauensbereiche, wobei zwischen der Basisvariante, der Dogleg-Methode und der Levenberg-Marquardt-Methode ausgewählt werden kann (siehe Dokumentation der Funktion in MATLAB).

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 hohe 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 nicht 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.

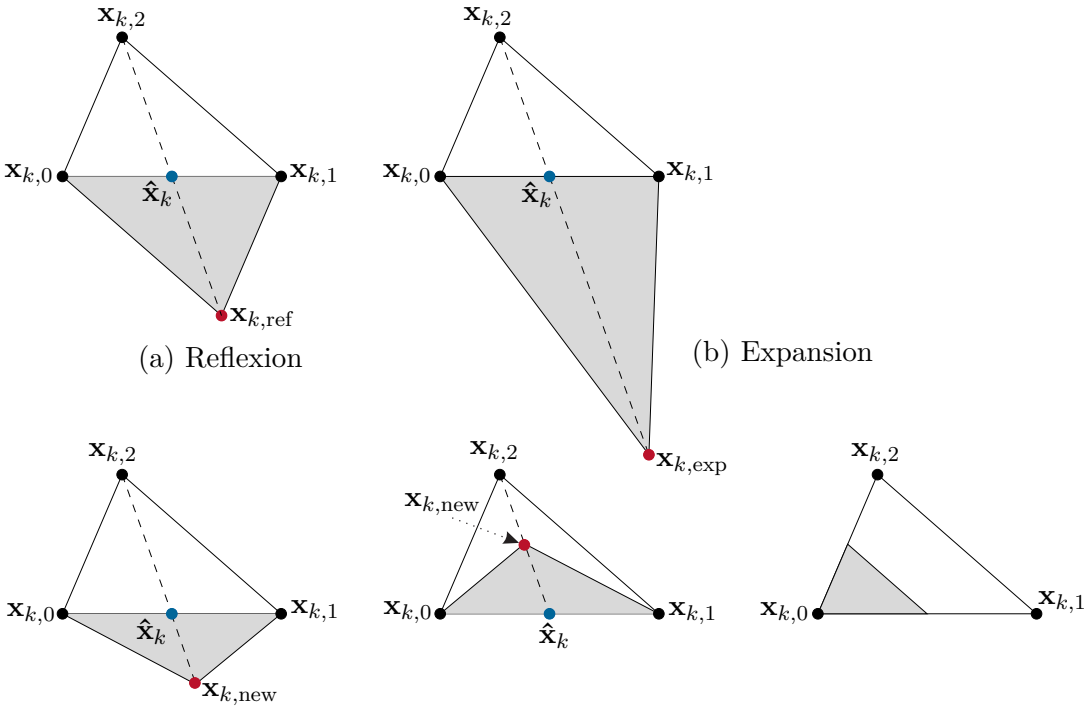
Der Algorithmus basiert auf der Iteration eines sogenannten *Simplex* im n-dimensionalen Raum der Optimierungsvariablen. Unter einem Simplex versteht man in diesem Zusammenhang jene konvexe Hülle, die von $n+1$ Punkten $\mathbf{x}_{k,i},\ i = 0, \ldots, n$ zum Iterationsschritt k im n-dimensionalen Suchraum aufgespannt wird (für $n = 1$ ist dies eine Linie, für $n = 2$ ein Dreieck, etc.). Im Weiteren werden mit $\mathbf{x}_{k,\min}$ und $\mathbf{x}_{k,\max}$ jene Eckpunkte des Simplex bezeichnet welche den kleinsten und größten Kostenfunktionswert f aufweisen, d. h. es gilt

$$\begin{aligned} f(\mathbf{x}_{k,\min}) &= \min_{i=0,\ldots,n} f(\mathbf{x}_{k,i}) \\ f(\mathbf{x}_{k,\max}) &= \max_{i=0,\ldots,n} f(\mathbf{x}_{k,i}) \ . \end{aligned} \tag{2.124}$$

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

$$\hat{\mathbf{x}}_k = \frac{1}{n}\left(\sum_{i=0}^{n} \mathbf{x}_{k,i} - \mathbf{x}_{k,\max}\right) . \tag{2.125}$$

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



(a) Reflexion (b) Expansion

(c) Äußere Kontraktion (d) Innere Kontraktion (e) Schrumpfung

Abbildung 2.9: Operationen des Simplex-Verfahrens nach Nelder und Mead für $\beta = 1$, $\gamma = 1$ und $\theta = 1/2$.

Operation dabei ist die Berechnung des *Reflexionspunktes*

$$\mathbf{x}_{k,\text{ref}} = \hat{\mathbf{x}}_k + (\hat{\mathbf{x}}_k - \mathbf{x}_{k,\text{max}}) \ , \tag{2.126}$$

der auf einer Geraden durch die Punkte $\mathbf{x}_{k,\text{max}}$ und $\hat{\mathbf{x}}_k$ liegt und symmetrisch bezüglich $\hat{\mathbf{x}}_k$ zu $\mathbf{x}_{k,\text{max}}$ ist, siehe Abbildung 2.9(a). Abhängig von $f(\mathbf{x}_{k,\text{ref}})$ im Vergleich zu den Funktionswerten der anderen Punkte des Simplex wird ein neuer Punkt $\mathbf{x}_{k,\text{new}}$ konstruiert, der im nächsten Iterationsschritt den Punkt $\mathbf{x}_{k,\text{max}}$ ersetzt. Der Algorithmus ist in seiner Grundfunktion in Tabelle 2.6 aufgelistet und die unterschiedlichen Operationen sind grafisch in Abbildung 2.9 dargestellt. Man beachte, dass die Schrumpfung im Algorithmus von Tabelle 2.6 stets bezüglich des Eckpunktes $\mathbf{x}_{k,\text{min}}$ mit dem kleinsten Kostenfunktionswert ausgeführt wird, d. h. in Abbildung 2.9(e) gilt $\mathbf{x}_{k,0} = \mathbf{x}_{k,\text{min}}$. 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.

Initialisierung: $\mathbf{x}_{0,i},\ i=0,\dots,n$ (Startsimplex)
$k \leftarrow 0$ (Iterationsindex)
$\beta > 0$ (Reflexionskoeffizient, typisch $\beta = 1$)
$\gamma > 0$ (Expansionskoeffizient, typisch $\gamma = 1$)
$\theta \in (0,1)$ (Kontraktionskoeffizient, typisch $\theta = 1/2$)
ε_x (Schwellwert für Abbruchkriterium)

repeat

$\mathbf{x}_{k,\min}$, $\mathbf{x}_{k,\max}$ gemäß (2.124) (Punkte mit min. und max. Kostenfunktionswert)

$\hat{\mathbf{x}}_k$ gemäß (2.125) (Schwerpunkt)

$\mathbf{x}_{k,\text{ref}} = \hat{\mathbf{x}}_k + \beta(\hat{\mathbf{x}}_k - \mathbf{x}_{k,\max})$ (Reflexionsschritt, Abb. 2.9(a))

if $f(\mathbf{x}_{k,\text{ref}}) < f(\mathbf{x}_{k,\min})$

$\mathbf{x}_{k,\text{exp}} = \mathbf{x}_{k,\text{ref}} + \gamma(\mathbf{x}_{k,\text{ref}} - \hat{\mathbf{x}}_k)$ (Expansionsschritt, Abb. 2.9(b))

if $f(\mathbf{x}_{k,\text{exp}}) < f(\mathbf{x}_{k,\text{ref}})$

$\mathbf{x}_{k,\text{new}} = \mathbf{x}_{k,\text{exp}}$

else

$\mathbf{x}_{k,\text{new}} = \mathbf{x}_{k,\text{ref}}$ (Reflexionspunkt beibehalten)

end if

else if $f(\mathbf{x}_{k,\text{ref}}) > \max\limits_{\mathbf{x}_{k,i} \neq \mathbf{x}_{k,\max},\ i=0,\dots,n} f(\mathbf{x}_{k,i})$

if $f(\mathbf{x}_{k,\max}) \leq f(\mathbf{x}_{k,\text{ref}})$

$\mathbf{x}_{k,\text{new}} = \theta\mathbf{x}_{k,\max} + (1-\theta)\hat{\mathbf{x}}_k$ (Innere Kontraktion, Abb. 2.9(d))

else

$\mathbf{x}_{k,\text{new}} = \theta\mathbf{x}_{k,\text{ref}} + (1-\theta)\hat{\mathbf{x}}_k$ (Äußere Kontraktion, Abb. 2.9(c))

end if

else

$\mathbf{x}_{k,\text{new}} = \mathbf{x}_{k,\text{ref}}$ (Reflexionspunkt beibehalten)

end if

if $f(\mathbf{x}_{k,\text{new}}) \geq f(\mathbf{x}_{k,\max})$ (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. 2.9(e))

else

$\mathbf{x}_{k,\max} \leftarrow \mathbf{x}_{k,\text{new}}$

$\mathbf{x}_{k+1,i} \leftarrow \mathbf{x}_{k,i},\ i=0,\dots,n$

end if

$k \leftarrow k+1$

until $\max_{i=0,\dots,n} \|\mathbf{x}_{k,i} - \mathbf{x}_{k,\min}\|_\infty \leq \varepsilon_x$

Tabelle 2.6: Simplex-Verfahren nach Nelder und Mead.

2.6 Beispiel: Rosenbrock's „Bananenfunktion“

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

$$\min_{\mathbf{x}\in\mathbb{R}^2} f(\mathbf{x}) \quad \text{mit} \quad f(\mathbf{x}) = 100(x_2 - x_1^2)^2 + (x_1 - 1)^2 \,. \tag{2.127}$$

Abbildung 2.10 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.

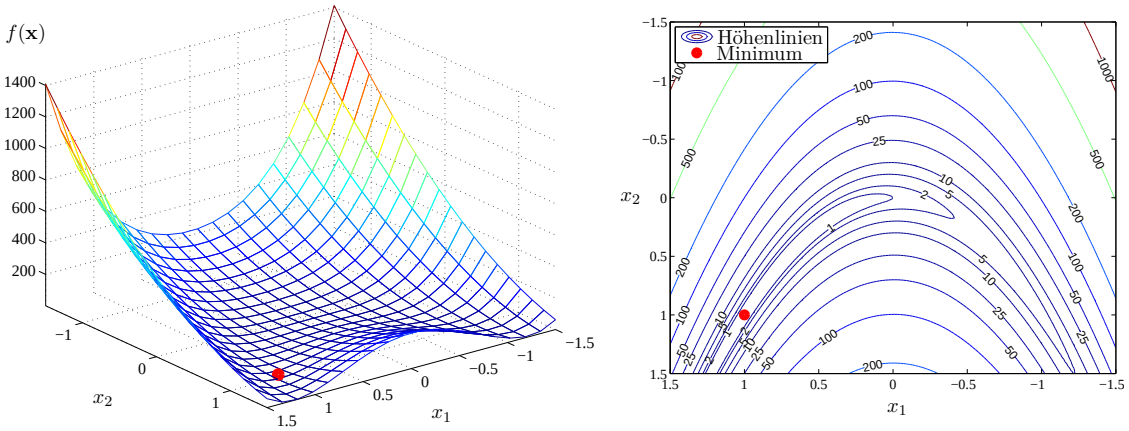


Abbildung 2.10: Profil und Höhenlinien von Rosenbrock's Bananenfunktion.

> *Aufgabe* 2.15. Verifizieren Sie, dass der Punkt $\mathbf{x}^* = [1 \quad 1]^{\mathrm{T}}$ ein Minimum darstellt. Ist das Minimum $\mathbf{x}^*$ global und eindeutig? Sind die Funktion $f(\mathbf{x})$ und das Optimierungsproblem (2.127) konvex?

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

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

Eine empfehlenswerte Alternative ist die frei zugängliche MATLAB-Funktion minFunc [2.11], die eine große Auswahl an Liniensuchverfahren bietet. Tabelle 2.7 zeigt einige Vergleichsdaten für die numerische Lösung des Rosenbrock-Problems (ausgehend vom Startwert $\mathbf{x}_0 = [-1 \quad -1]^{\mathrm{T}}$), die mit Hilfe von fminunc, fminsearch und minFunc berechnet wurden.

Abbildung 2.12 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 (2.127) angegeben, um zu verdeutlichen, wie die einzelnen Optimierungsverfahren mit fminunc und fminsearch aufgerufen werden können.

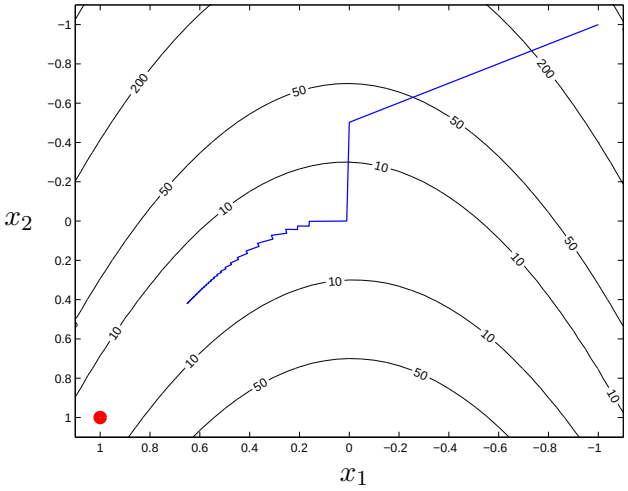


Abbildung 2.11: Darstellung der Iterationen des Gradientenverfahrens.

Verfahren	Funktion	Iter.	$f(\mathbf{x}^*)$	$\|(\nabla f)(\mathbf{x}^*)\|$	Funktionsaufrufe $f(\mathbf{x})$	$(\nabla f)(\mathbf{x})$	$\left(\nabla^2 f\right)(\mathbf{x})$
LS: Gradientenmethode	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-Methode	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-Meth.	minFunc	4	$2.8 \cdot 10^{-29}$	$1.1 \cdot 10^{-14}$	9	9	–
VB: Newton-Methode	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 2.7: 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 2.11 ist der Iterationsverlauf des Gradientenverfahrens über den Höhenlinien der Rosenbrock-Funktion (2.127) dargestellt. Das Gradientenverfahren verwendet die Richtung des steilsten Abstieges, welche 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.

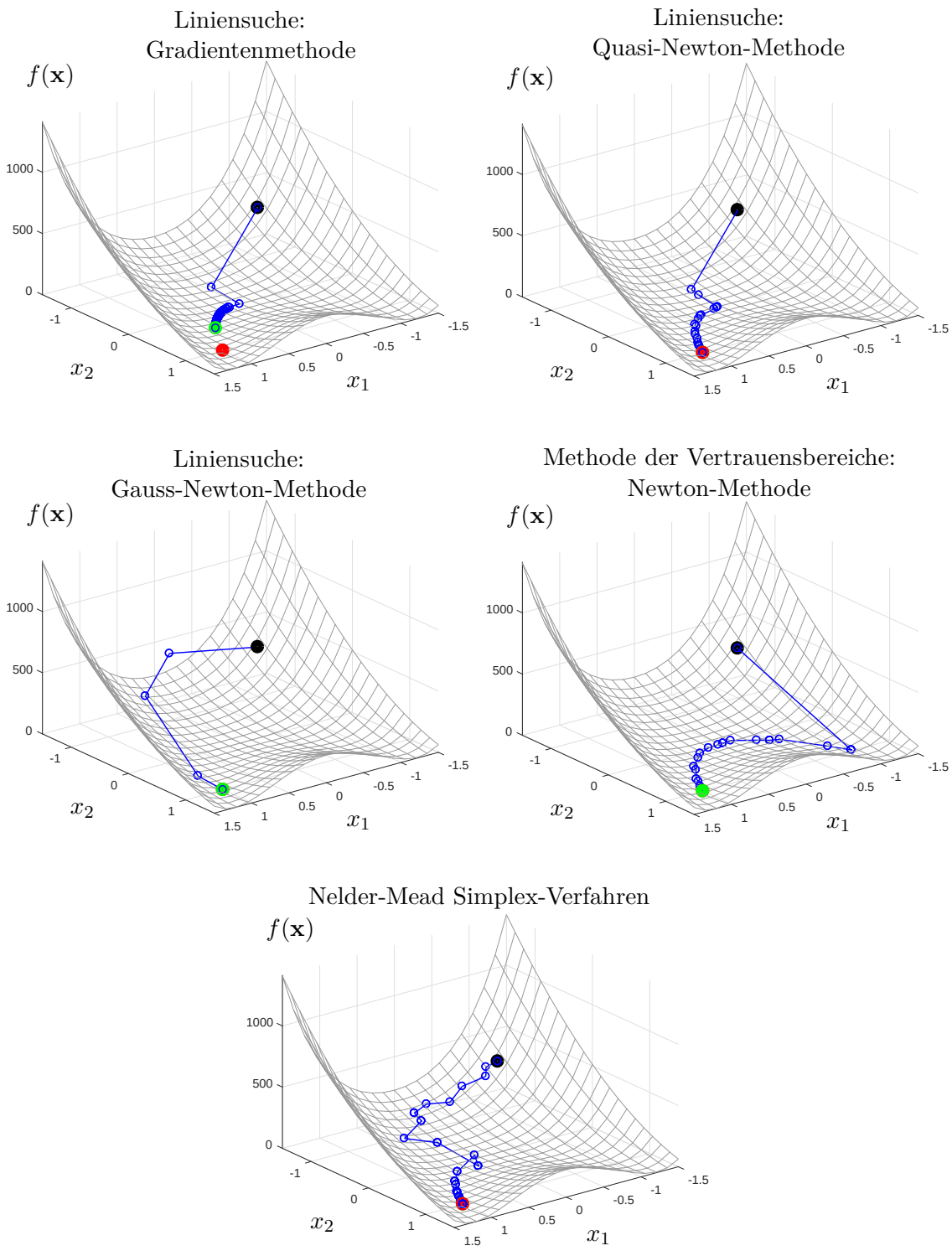


Abbildung 2.12: Rosenbrock's Bananenfunktion: Vergleich der numerischen Verfahren mit fminunc, fminsearch und minFunc.

Aufgabe 2.16. Berechnen Sie für das Minimum $\mathbf{x}^* = [1 \quad 1]^{\mathrm{T}}$ des Rosenbrock-Problems (2.127) die Konvergenzrate des Gradientenverfahrens gemäß Satz 2.7.

Das Konvergenzverhalten der *Quasi-Newton-Methode* in Abbildung 2.12 ist wesentlich besser als beim Gradientenverfahren. Die *Newton-Methode (Methode der Vertrauensbereiche)* in Abbildung 2.12 startet zunächst in die „falsche“ Richtung, was durch die *quadratische Approximation* (2.116) am Startpunkt $\mathbf{x}_0 = [-1 \quad -1]^{\mathrm{T}}$ zu erklären ist, deren Minimum in der Nähe von $\mathbf{x} \approx [-1 \quad 1]^{\mathrm{T}}$ liegt. Allerdings sind die einzelnen Schritte entlang des Tales der Rosenbrock-Funktion deutlich größer, da die Newton-Methode die *Hessematrix explizit verwendet* und nicht auf eine Approximation angewiesen ist wie im Fall der Quasi-Newton-Methode. Die *Gauss-Newton-Methode* 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 dominante Terme enthält, die affin in $\mathbf{x}$ sind (vgl. dazu Aufgabe 2.13).

Zusätzlich sind in Tabelle 2.7 und Abbildung 2.12 die Ergebnisse für das *Simplex-Verfahren von Nelder-Mead*, welches in der MATLAB-Funktion fminsearch implementiert ist, angegeben. 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 2.17. Schreiben Sie eine MATLAB-Funktion, die das Rosenbrock-Problem (2.127) mit Hilfe des Simplex-Verfahrens nach Nelder-Mead numerisch löst (siehe den Algorithmus von Tabelle 2.6). Konstruieren Sie den ersten Simplex mit den Eckpunkten

$$\mathbf{x}_{0,1} = \begin{bmatrix} -1 \\ -1 \end{bmatrix}, \quad \mathbf{x}_{0,2} = \mathbf{x}_{0,1} + s \begin{bmatrix} 1 \\ 0 \end{bmatrix}, \quad \mathbf{x}_{0,3} = \mathbf{x}_{0,1} + s \begin{bmatrix} 0 \\ 1 \end{bmatrix} \tag{2.128}$$

in Abhängigkeit der Seitenlänge $s = 1$. Verwenden Sie für die Abbruchbedingung in Tabelle 2.6 die Schranke $\varepsilon_f = 10^{-9}$ und vergleichen Sie Ihre Ergebnisse mit jenen von fminsearch in Tabelle 2.7. 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 2.18. Schreiben Sie eine MATLAB-Funktion, die das Rosenbrock-Problem (2.127) mit Hilfe der Newton-Methode (Liniensuche) löst, siehe Abschnitt 2.3.2.3. Verwenden Sie die heuristischen Bedingungen von Abschnitt 2.3.1.3 für die Schrittweitenbestimmung von α_k. Vergleichen Sie die Konvergenzergebnisse mit den Werten in Tabelle 2.7. Stellen Sie die einzelnen Iterationen grafisch dar und variieren Sie die Startpunkte $\mathbf{x}_0$.

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

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

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

  % Generelles: 'SpecifyObjectiveGradient' gibt an, dass die Funktion zur Berechnung der
  % Kostenfunktion auch den Wert des Gradienten retourniert. Ist auch 'HessianFcn' auf
  % 'objective' gesetzt, muss zusätzlich die Hessematrix zurückgegeben werden.
  switch methodQ
    case 1     % fminunc: Liniensuche: Gradientenmethode
      opt_fminu = optimoptions(opt_fminu,'Algorithm','quasi-newton','HessUpdate','steepdesc',...
          'SpecifyObjectiveGradient',true);
      [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',...
          'SpecifyObjectiveGradient',true);
      [Xopt,fopt,exitflag,output] = fminunc(@rosenbrock,Xinit,opt_fminu);
    case 3     % fminunc: Methode der Vertrauensbereiche: Newton-Methode
      opt_fminu = optimoptions(opt_fminu,'Algorithm','trust-region','HessianFcn','objective',...
          'SpecifyObjectiveGradient',true);
      [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-Methode
      figure
      opt.Method = 'cg';
      [Xopt,fopt,exitflag,output] = minFunc(@rosenbrock,Xinit,opt);
    case 6     % minFunc: Newton-Methode
      figure
      opt.Method = 'newton';
      [Xopt,fopt,exitflag,output] = minFunc(@rosenbrock,Xinit,opt);
    case 7    % minFunc: Gauss-Newton-Methode
      figure
      opt.Method = 'newton';
      [Xopt,fopt,exitflag,output] = minFunc(@rosenbrock_gauss_newton,Xinit,opt);
  end
end

function [f, grad, H] = rosenbrock(x)
  grad = [];
  H = [];
  f = 100*(x(2)-x(1)^2)^2 + (x(1)-1)^2;                    % Rosenbrock-Funktion
  if nargout>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 nargout>2                                              % falls Hessematrix angefordert wird
```

```
    H = [ -400*(x(2)-3*x(1)^2)+2, -400*x(1);
          -400*x(1),              200 ];
  end
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
end

function stop = plot_iterates(x,info,state)
  global old
  f = rosenbrock(x);
  switch state                                          % Grafische Ausgabe:
    case 'init'                                         % 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];
end

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');
end
```

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* (International Series in Operations Research & Management Science), 3. Aufl. Springer, 2008, Bd. 116.

[2.3] R. Fletcher, *Practical methods of optimization*, 2. Aufl. John Wiley & Sons, 1987.

[2.4] I. Griva, S. Nash und A. Sofer, *Linear and Nonlinear Optimization*, 2. Aufl. Society for Industrial und Applied Mathematics, 2009.

[2.5] E. Chong und S. Żak, *An Introduction to Optimization* (Wiley Series in Discrete Mathematics and Optimization), 4. Aufl. Wiley, 2013.

[2.6] M. Hestenes und E. Stiefel, „Methods of conjugate gradients for solving linear systems," *Journal of Research of the National Bureau of Standards*, Jg. 49, Nr. 6, S. 409–436, 1952.

[2.7] J. Shewchuk, *An introduction to the conjugate gradient method without the agonizing pain*, Technical Report CMU-CS-94-125, Pittsburgh, PA: School of Computer Science, Carnegie Mellon University, 1994.

[2.8] J. Nocedal und S. J. Wright, *Numerical Optimization* (Springer Series in Operations Research and Financial Engineering), 2. Aufl. Springer, 2006.

[2.9] J. Dennis und R. Schnabel, *Numerical Methods for Unconstrained Optimization and Nonlinear Equations* (Classics in Applied Mathematics). Society for Industrial und Applied Mathematics, 1996.

[2.10] M. Papageorgiou, M. Leibold und M. Buss, *Optimierung: Statische, dynamische, stochastische Verfahren für die Anwendung*, 4. Aufl. Springer, 2015.

[2.11] M. Schmidt, „minFunc," abrufbar unter http://www.cs.ubc.ca/~schmidtm/Software/minFunc.html, University of British Columbia, Vancouver, 2005. (besucht am 28. 09. 2016).

[2.12] S. Boyd und L. Vandenberghe, *Convex Optimization*. Cambridge University Press, 2004.

[2.13] C. T. Kelley, *Iterative Methods for Optimization*. Society for Industrial und Applied Mathematics, 1999.

[2.14] 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 30. 09. 2020).

[2.15] 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

$$\begin{aligned} \min_{\mathbf{x}\in\mathbb{R}^n} \quad & f(\mathbf{x}) && \text{Kostenfunktion} && (3.1a)\\ \text{u.B.v.} \quad & g_i(\mathbf{x}) = 0, \quad i = 1,\dots,p && \text{Gleichungsbeschränkungen} && (3.1b)\\ & h_i(\mathbf{x}) \le 0, \quad i = 1,\dots,q && \text{Ungleichungsbeschränkungen} && (3.1c) \end{aligned}$$

mit $p \le n$, 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 $\mathbf{g}(\mathbf{x}) = [g_1(\mathbf{x}) \quad \dots \quad g_p(\mathbf{x})]^{\mathrm{T}}$ und $\mathbf{h}(\mathbf{x}) = [h_1(\mathbf{x}) \quad \dots \quad h_q(\mathbf{x})]^{\mathrm{T}}$ zusammen, so kann das Optimierungsproblem (3.1) in der Form

$$\begin{aligned} \min_{\mathbf{x}\in\mathbb{R}^n} \quad & f(\mathbf{x}) && (3.2a)\\ \text{u.B.v.} \quad & \mathbf{g}(\mathbf{x}) = \mathbf{0} && (3.2b)\\ & \mathbf{h}(\mathbf{x}) \le \mathbf{0} && (3.2c) \end{aligned}$$

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

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

mit der zulässigen Menge

$$\mathcal{X} = \{\, \mathbf{x} \in \mathbb{R}^n \,|\, \mathbf{g}(\mathbf{x}) = \mathbf{0},\ \mathbf{h}(\mathbf{x}) \le \mathbf{0} \,\} \tag{3.3b}$$

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

Die Berücksichtigung von allgemeinen nichtlinearen Ungleichungsbeschränkungen (3.2c) ist 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

$$\begin{aligned} \min_{\substack{\mathbf{x}\in\mathbb{R}^n,\\ \mathbf{x}_s\in\mathbb{R}^q}} \quad & f(\mathbf{x}) && (3.4a)\\ \text{u.B.v.} \quad & \mathbf{g}(\mathbf{x}) = \mathbf{0} && (3.4b)\\ & \mathbf{h}(\mathbf{x}) + \mathbf{x}_s = \mathbf{0} && (3.4c) \end{aligned}$$

$$\mathbf{x}_s \geq \mathbf{0} \ . \tag{3.4d}$$

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.

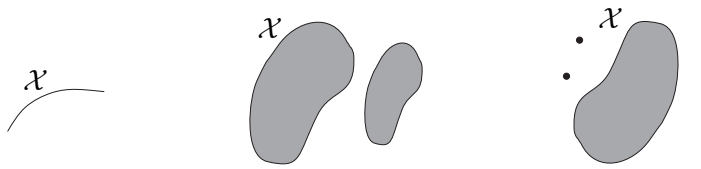
3.1 Optimalitätsbedingungen

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.



Nicht robust (a) Robust (b) Nicht robust (c)

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 zulässige Richtungen sind, gesucht und durch Eingrenzung der Schrittweite α 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*

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

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

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

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

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

mit $g'(0) = \mathbf{d}^{\mathrm{T}}(\nabla f)(\mathbf{x}^*)$. 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 $g'(0) = \mathbf{d}^{\mathrm{T}}(\nabla f)(\mathbf{x}^*) \geq 0$.

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 $g'(0) = \mathbf{d}^{\mathrm{T}}(\nabla f)(\mathbf{x}^*) \geq 0$ für beliebige $\mathbf{d} \in \mathbb{R}^n$ erfüllt ist, muss $(\nabla f)(\mathbf{x}^*) = \mathbf{0}$ gelten. □

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 $\mathcal{X}$.

Satz 3.2 (Hinreichende 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 $\mathcal{X}$. Gilt für jede zulässige Richtung $\mathbf{d} \neq \mathbf{0}$ an einem Punkt $\mathbf{x}^* \in \mathcal{X}$ die Ungleichungsbedingung*

$$\mathbf{d}^{\mathrm{T}}(\nabla f)(\mathbf{x}^*) > 0\ , \tag{3.8}$$

so ist $\mathbf{x}^*$ ein striktes lokales Minimum von f auf $\mathcal{X}$. Der Punkt $\mathbf{x}^*$ muss folglich am Rand von $\mathcal{X}$ liegen.

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

Für allgemeine Kostenfunktionen $f \in C^1$ und optimale Punkte $\mathbf{x}^*$ im Inneren von $\mathcal{X}$ existiert (ohne weitere Voraussetzungen, siehe z. B. Satz 3.5) keine hinreichende Optimalitätsbedingung erster Ordnung.

Beispiel 3.1. Es wird die Optimierungsaufgabe

$$\min_{\mathbf{x}\in\mathcal{X}} \; f(x_1, x_2) = x_1^2 - 2x_1 + x_2 + x_1 x_2 \tag{3.9}$$

mit der zulässigen Menge

$$\mathcal{X} = \Big\{ \mathbf{x} \in \mathbb{R}^2 \,|\, x_1 \geq 0,\; x_2 \geq 0 \Big\} \tag{3.10}$$

betrachtet. Es soll untersucht werden, ob der Punkt $\mathbf{x}^* = [1 \quad 0]^{\mathrm{T}}$ ein Minimum ist. Eine Auswertung des Gradienten an der Stelle $\mathbf{x}^*$ ergibt

$$(\nabla f)(\mathbf{x}^*) = \begin{bmatrix} 2x_1^* - 2 + x_2^* \\ 1 + x_1^* \end{bmatrix} = \begin{bmatrix} 0 \\ 2 \end{bmatrix} . \tag{3.11}$$

In diesem Fall liegt $\mathbf{x}^*$ am Rand von $\mathcal{X}$ und der Gradient $(\nabla f)(\mathbf{x}^*)$ verschwindet nicht. Wegen der Definition von $\mathcal{X}$ gemäß (3.10) muss für alle zulässigen Richtungen $\mathbf{d}$ am Punkt $\mathbf{x}^*$ gelten, dass die zweite Komponente von $\mathbf{d}$ größer gleich Null ist. Folglich ist die notwendige Bedingung (3.5) für alle zulässigen Richtungen $\mathbf{d}$ erfüllt. Die hinreichende Bedingung (3.8) ist jedoch nicht für alle zulässigen Richtungen $\mathbf{d} \neq \mathbf{0}$ erfüllt. Gemäß den Optimalitätsbedingungen erster Ordnung ist daher weder ausgeschlossen noch gesichert, dass $\mathbf{x}^*$ ein Minimum darstellt.

Satz 3.3 (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*

$$(a) \quad \mathbf{d}^{\mathrm{T}}(\nabla f)(\mathbf{x}^*) \geq 0 \tag{3.12a}$$

$$(b) \quad \text{wenn } \mathbf{d}^{\mathrm{T}}(\nabla f)(\mathbf{x}^*) = 0, \text{ dann } \mathbf{d}^{\mathrm{T}}\Big(\nabla^2 f\Big)(\mathbf{x}^*)\mathbf{d} \geq 0 \; . \tag{3.12b}$$

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

$$(a) \quad (\nabla f)(\mathbf{x}^*) = \mathbf{0} \tag{3.13a}$$

$$(b) \quad \Big(\nabla^2 f\Big)(\mathbf{x}^*) \text{ ist positiv semi-definit.} \tag{3.13b}$$

Aufgabe 3.2. Beweisen Sie Satz 3.3. **Hinweis:** Orientieren Sie sich dabei am Beweis von Satz 3.1.

Aufgabe 3.3. Es wird die Optimierungsaufgabe

$$\min_{\mathbf{x}\in\mathcal{X}} \; f(x_1, x_2) = x_1^3 - x_1^2 x_2 + 2x_2^2 \tag{3.14}$$

mit der zulässigen Menge

$$\mathcal{X} = \left\{ \mathbf{x} \in \mathbb{R}^2 \,|\, x_1 \geq 0,\ x_2 \geq 0 \right\} \tag{3.15}$$

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

Satz 3.4 (Hinreichende 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 $\mathcal{X}$. Erfüllt ein Punkt $\mathbf{x}^* \in \mathcal{X}$ für jede dort zulässige Richtung $\mathbf{d} \neq \mathbf{0}$ die Bedingungen*

$$\text{(a)} \quad \mathbf{d}^{\mathrm{T}}(\nabla f)(\mathbf{x}^*) \geq 0 \tag{3.16a}$$

$$\text{(b)} \quad \text{wenn } \mathbf{d}^{\mathrm{T}}(\nabla f)(\mathbf{x}^*) = 0, \text{ dann } \mathbf{d}^{\mathrm{T}}\left(\nabla^2 f\right)(\mathbf{x}^*)\mathbf{d} > 0\ , \tag{3.16b}$$

so ist $\mathbf{x}^*$ ein striktes lokales Minimum von f auf $\mathcal{X}$. Liegt $\mathbf{x}^*$ im Inneren von $\mathcal{X}$, dann reicht daher die Erfüllung der Bedingungen

$$\text{(a)} \quad (\nabla f)(\mathbf{x}^*) = \mathbf{0} \tag{3.17a}$$

$$\text{(b)} \quad \left(\nabla^2 f\right)(\mathbf{x}^*) \text{ ist positiv definit.} \tag{3.17b}$$

Der Beweis dieses Satzes erfolgt analog zu den Beweisen der vorangegangenen Sätze.

Beispiel 3.2 (Fortsetzung von Beispiel 3.1). Es wird für das in Beispiel 3.1 gegebene Optimierungsproblem die Erfüllung der Optimalitätsbedingungen zweiter Ordnung am Punkt $\mathbf{x}^* = [1 \quad 0]^{\mathrm{T}}$ untersucht. Eine Auswertung der Hessematrix an der Stelle $\mathbf{x}^*$ ergibt

$$\left(\nabla^2 f\right)(\mathbf{x}^*) = \begin{bmatrix} 2 & 1 \\ 1 & 0 \end{bmatrix}\ . \tag{3.18}$$

Für alle zulässigen Richtungen $\mathbf{d}$ am Punkt $\mathbf{x}^*$ muss die zweite Komponente von $\mathbf{d}$ größer gleich Null sein. Die Erfüllung der Bedingung (3.12a) wurde bereits in Beispiel 3.1 gezeigt. Aus dem Gradienten $(\nabla f)(\mathbf{x}^*) = [0 \quad 2]^{\mathrm{T}}$ folgt, dass $\mathbf{d}^{\mathrm{T}}(\nabla f)(\mathbf{x}^*) = 0$ für all jene zulässigen Richtungen $\mathbf{d}$ gilt deren zweite Komponente gleich Null ist. Für solche zulässigen Richtungen $\mathbf{d}$ ergibt sich

$$\mathbf{d}^{\mathrm{T}}\left(\nabla^2 f\right)(\mathbf{x}^*)\mathbf{d} \geq 0\ . \tag{3.19}$$

Die Bedingung 3.12b ist daher erfüllt und der Punkt $\mathbf{x}^*$ genügt den notwendigen Optimalitätsbedingungen zweiter Ordnung. Da für alle zulässigen Richtungen $\mathbf{d} \neq \mathbf{0}$ welche $\mathbf{d}^{\mathrm{T}}(\nabla f)(\mathbf{x}^*) = 0$ erfüllen (erste Komponenten ungleich Null, zweite Komponente gleich Null)

$$\mathbf{d}^{\mathrm{T}}\left(\nabla^2 f\right)(\mathbf{x}^*)\mathbf{d} > 0 \tag{3.20}$$

gilt, genügt der Punkt $\mathbf{x}^*$ auch den hinreichenden Optimalitätsbedingungen zweiter Ordnung, d. h. $\mathbf{x}^*$ ist ein striktes lokales Minimum.

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* für ein globales Minimum. Um dies zu sehen, beachte man, dass mit beliebigem $\mathbf{x} \in \mathcal{X}$, $\mathbf{x} - \mathbf{x}^*$ eine zulässige Richtung am Punkt $\mathbf{x}^*$ darstellt und wegen der Konvexität von $f(\mathbf{x})$ die Ungleichung

$$f(\mathbf{x}) \geq f(\mathbf{x}^*) + \underbrace{(\mathbf{x} - \mathbf{x}^*)^{\mathrm{T}}(\nabla f)(\mathbf{x}^*)}_{\geq 0} \geq f(\mathbf{x}^*) \tag{3.21}$$

gilt. Die Sätze 3.1 bis 3.4 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 3.5 (Globale Minima einer konvexen Funktion)**.** *Es sei $f(\mathbf{x})$ eine konvexe Funktion auf einer konvexen Menge $\mathcal{X}$. Die Menge aller Minima $\mathcal{G} = \arg\min\{f(\mathbf{x}) \,|\, \mathbf{x} \in \mathcal{X}\}$ 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 schon vorab 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

$$\min_{\mathbf{x} \in \mathbb{R}^n} \quad f(\mathbf{x}) \tag{3.22a}$$

$$\text{u.B.v.} \quad \mathbf{g}(\mathbf{x}) = \mathbf{0}\ . \tag{3.22b}$$

Für die *zulässige Menge* gilt dann

$$\mathcal{X} = \{\mathbf{x} \in \mathbb{R}^n \,|\, g_i(\mathbf{x}) = 0,\ i = 1, \dots, p\} \ . \tag{3.23}$$

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

$$\operatorname{rang}((\nabla \mathbf{g})(\mathbf{x})) = \operatorname{rang}\Big(\Big[(\nabla g_1)(\mathbf{x}) \quad (\nabla g_2)(\mathbf{x}) \quad \dots \quad (\nabla g_p)(\mathbf{x})\Big]\Big) = p, \tag{3.24}$$

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 $\mathcal{X}$.

Es ist zu beachten, 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 $g(\mathbf{x}) = x_1 = 0$ und $g(\mathbf{x}) = x_1^2 = 0$ im $\mathbb{R}^n$. Beide Beschränkungen definieren die gleiche zulässige Menge $\mathcal{X}$, nämlich in der Form $\mathcal{X} = \{\mathbf{x} \in \mathbb{R}^n \,|\, x_1 = 0\}$. Für $g(\mathbf{x}) = x_1$ ist jeder Punkt von $\mathcal{X}$ ein regulärer Punkt gemäß der Definition 3.2. Für $g(\mathbf{x}) = x_1^2$ jedoch ist kein Punkt von $\mathcal{X}$ regulär.

Die zulässige Menge $\mathcal{X} = \{\mathbf{x} \in \mathbb{R}^n \,|\, g_i(\mathbf{x}) = 0,\ i = 1, \dots, p\}$ 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 [3.1]). 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 Annullator des p-dimensionalen Kotangentialraumes, welcher durch die exakten Differentiale $\mathrm{d}g_i : \mathcal{T}_{\mathbf{x}}\mathcal{X} \to \mathbb{R}$, $i = 1, \dots, p$ gebildet wird, definieren. Das heißt, es gilt

$$\mathcal{T}_{\mathbf{x}}\mathcal{X} = \left\{ \mathbf{d} \in \mathbb{R}^n \;\middle|\; \mathrm{d}g_i(\mathbf{d}) = \mathrm{L}_{\mathbf{d}} g_i(\mathbf{x}) = \underbrace{\left(\frac{\partial}{\partial \mathbf{x}} g_i(\mathbf{x})\right)}_{(\nabla g_i)^{\mathrm{T}}(\mathbf{x})} \mathbf{d} = 0,\ i = 1, \dots, p \right\} . \tag{3.25}$$

Man beachte, dass die Vektoren $\mathbf{d}$ in (3.25) im allgemeinen Fall, d. h. bei nichtlinearen Gleichungsnebenbedingungen, keine zulässigen Richtungen im Sinne der Definition 3.1 sind. Im speziellen Fall $p = n$ folgt für jeden regulären Punkt $\mathbf{x} \in \mathcal{X}$ der Tangentialraum $\mathcal{T}_{\mathbf{x}}\mathcal{X} = \{\mathbf{0}\}$. 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* $\mathcal{X} = \{\mathbf{x} \in \mathbb{R}^n \,|\, g_i(\mathbf{x}) = 0,\ i = 1, \dots, p\}$ *ein regulärer Punkt und ein lokaler Extremalpunkt 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

$$(\nabla \mathbf{g})^{\mathrm{T}}(\mathbf{x}^*)\mathbf{d} = \mathbf{0} \tag{3.26}$$

erfüllen, muss auch gelten

$$(\nabla f)^{\mathrm{T}}(\mathbf{x}^*)\mathbf{d} = 0 \ . \tag{3.27}$$

Beweisskizze: Da $\mathbf{x}^* \in \mathcal{X}$ ein regulärer Punkt ist, liegt jedes $\mathbf{d}$ das (3.26) erfüllt gemäß (3.25) 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 parametriert in α durch den Punkt $\mathbf{x}^*$ mit dem Tangentialvektor $\mathbf{d}$, so dass gilt $\mathbf{x}(0) = \mathbf{x}^*$ und $\left(\frac{\mathrm{d}}{\mathrm{d}\alpha}\mathbf{x}\right)(0) = \mathbf{d}$. Da nun $\mathbf{x}^*$ ein lokaler Extremalpunkt ist, muss die Beziehung

$$\frac{\mathrm{d}}{\mathrm{d}\alpha} f(\mathbf{x}(\alpha))\Big|_{\alpha=0} = \underbrace{\left(\frac{\partial}{\partial \mathbf{x}} f\right)(\mathbf{x}(0))}_{(\nabla f)^{\mathrm{T}}(\mathbf{x}^*)} \underbrace{\left(\frac{\mathrm{d}}{\mathrm{d}\alpha}\mathbf{x}\right)(0)}_{\mathbf{d}} = 0 \tag{3.28}$$

gelten. □

Lemma 3.1, speziell (3.27), 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 eindeutig 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 3.6 (Notwendige Optimalitätsbedingungen erster Ordnung)**.** *Es sei $\mathbf{x}^* \in \mathcal{X}$ mit $\mathcal{X} = \{\mathbf{x} \in \mathbb{R}^n \,|\, g_i(\mathbf{x}) = 0,\ i = 1, \dots, p\}$ ein regulärer Punkt und ein lokaler Extremalpunkt 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*

$$(\nabla f)(\mathbf{x}^*) + (\nabla \mathbf{g})(\mathbf{x}^*)\boldsymbol{\lambda}^* = \mathbf{0} \ . \tag{3.29}$$

Die notwendige Optimalitätsbedingung (3.29) 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*

$$L(\mathbf{x}, \boldsymbol{\lambda}) = f(\mathbf{x}) + \boldsymbol{\lambda}^{\mathrm{T}}\mathbf{g}(\mathbf{x}) \tag{3.30}$$

einführen und die notwendigen Optimalitätsbedingungen von Satz 3.6 in der Form

$$\left(\frac{\partial}{\partial \mathbf{x}} L\right)^{\mathrm{T}}(\mathbf{x}^*, \boldsymbol{\lambda}^*) = (\nabla f)(\mathbf{x}^*) + (\nabla \mathbf{g})(\mathbf{x}^*)\boldsymbol{\lambda}^* = \mathbf{0} \tag{3.31a}$$

$$\left(\frac{\partial}{\partial \boldsymbol{\lambda}} L\right)^{\mathrm{T}}(\mathbf{x}^*, \boldsymbol{\lambda}^*) = \mathbf{g}(\mathbf{x}^*) = \mathbf{0} \tag{3.31b}$$

schreiben.

Wenn die Funktion $f(\mathbf{x})$ und die zulässige Menge $\mathcal{X}$ konvex sind, dann sind die notwendigen Optimalitätsbedingungen erster Ordnung gemäß Satz 3.6 auch *hinreichend* für ein

globales Minimum. Um dies zu sehen, beachte man zunächst, dass für konvexes $\mathcal{X}$ die Gleichungsbeschränkungen $\mathbf{g}(\mathbf{x}) = \mathbf{0}$ affin in $\mathbf{x}$ sein müssen, d. h. sie haben die Struktur

$$\mathbf{g}(\mathbf{x}) = \mathbf{A}^{\mathrm{T}}\mathbf{x} + \mathbf{b} = \mathbf{0} \tag{3.32}$$

und es gilt

$$(\mathbf{x} - \mathbf{x}^*)^{\mathrm{T}}(\nabla \mathbf{g})(\mathbf{x}^*) = (\mathbf{x} - \mathbf{x}^*)^{\mathrm{T}}\mathbf{A} = \mathbf{0} \qquad \forall\, \mathbf{x} \in \mathcal{X},\ \mathbf{x}^* \in \mathcal{X}\ . \tag{3.33}$$

Aus der Konvexität von $f(\mathbf{x})$ auf $\mathcal{X}$ folgt mit (3.29) und (3.33) die Ungleichung

$$\begin{aligned} f(\mathbf{x}) &\geq f(\mathbf{x}^*) + (\mathbf{x} - \mathbf{x}^*)^{\mathrm{T}}(\nabla f)(\mathbf{x}^*) \\ &= f(\mathbf{x}^*) - \underbrace{(\mathbf{x} - \mathbf{x}^*)^{\mathrm{T}}(\nabla \mathbf{g})(\mathbf{x}^*)}_{=\mathbf{0}}\boldsymbol{\lambda}^* = f(\mathbf{x}^*) \qquad \forall\, \mathbf{x} \in \mathcal{X} \end{aligned} \tag{3.34}$$

und damit die globale Optimalität von $\mathbf{x}^*$.

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

$$\min_{\mathbf{x}\in\mathbb{R}^2} \quad f(\mathbf{x}) = (x_1 - 2)^2 + (x_2 - 1)^2 \tag{3.35a}$$

$$\text{u.B.v.} \quad g(\mathbf{x}) = x_2 - 2x_1 = 0 \tag{3.35b}$$

(siehe auch Beispiel 1.2). Abbildung 3.2 stellt die Gerade $g(\mathbf{x}) = 0$ und die Höhenlinien von $f(\mathbf{x})$ dar.

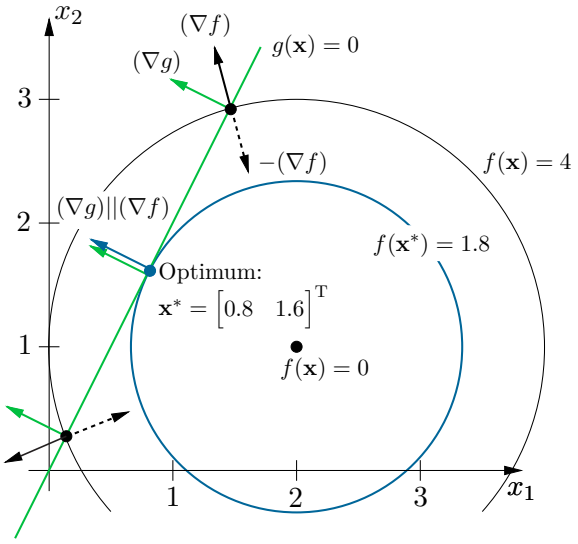


Abbildung 3.2: Veranschaulichung von Beispiel 3.3 mit einer Gleichungsbeschränkung.

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

$$\mathbf{x}^* = \begin{bmatrix} 0.8 & 1.6 \end{bmatrix}^{\mathrm{T}}, \quad f(\mathbf{x}^*) = 1.8 \tag{3.36}$$

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

$$(\nabla f)(\mathbf{x}) = \begin{bmatrix} 2(x_1 - 2) \\ 2(x_2 - 1) \end{bmatrix}, \quad (\nabla g)(\mathbf{x}) = \begin{bmatrix} -2 \\ 1 \end{bmatrix} \tag{3.37}$$

und damit errechnen sich die notwendigen Optimalitätsbedingungen (3.31) 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

$$\frac{\partial}{\partial x_1} L(x_1, x_2, \lambda) = 2(x_1 - 2) - 2\lambda = 0 \tag{3.38a}$$

$$\frac{\partial}{\partial x_2} L(x_1, x_2, \lambda) = 2(x_2 - 1) + \lambda = 0 \tag{3.38b}$$

$$\frac{\partial}{\partial \lambda} L(x_1, x_2, \lambda) = x_2 - 2x_1 = 0 \ . \tag{3.38c}$$

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

Aufgabe 3.4. 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 3.5. Gegeben ist ein nichtlineares zeitvariantes Abtastsystem der Form

$$x_{k+1} = \varphi_k(x_k, u_k), \quad x_0 = x(0) \tag{3.39}$$

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

$$J = \sum_{k=0}^{N} \psi_k(x_k, u_k) \tag{3.40}$$

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

$$\lambda_{k-1} = \lambda_k \left(\frac{\partial}{\partial x} \varphi_k \right)(x_k, u_k) + \left(\frac{\partial}{\partial x} \psi_k \right)(x_k, u_k), \quad k = 1, \ldots, N \tag{3.41a}$$

$$\lambda_N = \mu \left(\frac{\partial}{\partial x} g \right)(x_{N+1}) \tag{3.41b}$$

$$0 = \lambda_k \left(\frac{\partial}{\partial u} \varphi_k \right)(x_k, u_k) + \left(\frac{\partial}{\partial u} \psi_k \right)(x_k, u_k), \quad k = 0, \ldots, N \tag{3.41c}$$

mit einem geeigneten Wert μ verbunden sind.

Satz 3.7 (Notwendige Optimalitätsbedingungen zweiter Ordnung). *Es sei $\mathbf{x}^* \in \mathcal{X}$ mit $\mathcal{X} = \{\mathbf{x} \in \mathbb{R}^n \,|\, g_i(\mathbf{x}) = 0,\ i = 1, \ldots, p\}$ ein regulärer Punkt und ein lokales Minimum von $f(\mathbf{x})$ unter den Gleichungsnebenbedingungen $\mathbf{g}(\mathbf{x}) = \mathbf{0}$ mit $f, g_1, \ldots, g_p \in C^2$. Dann existiert ein eindeutiges $\boldsymbol{\lambda}^* \in \mathbb{R}^p$ so, dass gilt*

$$\left(\frac{\partial}{\partial \mathbf{x}} L \right)^{\mathrm{T}} (\mathbf{x}^*, \boldsymbol{\lambda}^*) = (\nabla f)(\mathbf{x}^*) + (\nabla \mathbf{g})(\mathbf{x}^*) \boldsymbol{\lambda}^* = \mathbf{0} \tag{3.42}$$

und

$$\mathbf{d}^{\mathrm{T}} \left(\nabla^2 L \right)(\mathbf{x}^*, \boldsymbol{\lambda}^*) \mathbf{d} = \mathbf{d}^{\mathrm{T}} \left(\left(\nabla^2 f \right)(\mathbf{x}^*) + \sum_{i=1}^{p} \lambda_i^* \left(\nabla^2 g_i \right)(\mathbf{x}^*) \right) \mathbf{d} \geq 0 \tag{3.43}$$

für alle $\mathbf{d} \in \mathcal{T}_{\mathbf{x}^*} \mathcal{X}$ mit der Lagrangefunktion L gemäß (3.30).

Die Sätze 3.6 und 3.7 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 3.8 (Hinreichende Optimalitätsbedingungen zweiter Ordnung). *Gesucht ist das Minimum der Kostenfunktion $f(\mathbf{x})$ unter den Gleichungsnebenbedingungen $g_i(\mathbf{x}) = 0,\ i = 1, \ldots, p$ mit $f, g_1, \ldots, g_p \in C^2$. Wenn $(\mathbf{x}^*, \boldsymbol{\lambda}^*)$ so existieren, dass*

$$\left(\frac{\partial}{\partial \mathbf{x}} L \right)^{\mathrm{T}} (\mathbf{x}^*, \boldsymbol{\lambda}^*) = \mathbf{0} \tag{3.44}$$

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

$$\mathbf{d}^{\mathrm{T}} \left(\nabla^2 L \right)(\mathbf{x}^*, \boldsymbol{\lambda}^*) \mathbf{d} > 0 \qquad \forall\, \mathbf{d} \in \mathcal{T}_{\mathbf{x}^*} \mathcal{X},\ \mathbf{d} \neq \mathbf{0} \tag{3.45}$$

mit der Lagrangefunktion L gemäß (3.30), dann ist $\mathbf{x}^*$ ein striktes lokales Minimum.

Im speziellen Fall $p = n$ ist folglich die Optimalitätsbedingung erster Ordnung gemäß Satz 3.6 auch hinreichend für ein striktes lokales Minimum. Es gilt dann nämlich $\mathcal{T}_{\mathbf{x}} \mathcal{X} = \{\mathbf{0}\}$, so dass die Bedingungen (3.43) und (3.45) unabhängig von $(\nabla^2 L)(\mathbf{x}^*, \boldsymbol{\lambda}^*)$ erfüllt werden.

Für den üblichen Fall $p < n$ zeigt sich anhand von Satz 3.8, 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). Dieser Zusammenhang erklärt auch, warum das *Konvergenzverhalten von iterativen Lösungsverfahren des beschränkten Optimierungsproblems durch die Eigenwerte der Matrix $\mathbf{L}$ eingeschränkt auf den Tangentialraum $\mathcal{T}_{\mathbf{x}^*} \mathcal{X}$ beeinflusst* werden. Um nun die Erfüllung von (3.45) zu überprüfen, kann

die Matrix $\mathbf{L}$ in 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. $\mathbf{T}^{\mathrm{T}}\mathbf{T} = \mathbf{E}$. 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 $\mathbf{d} = \mathbf{T}\mathbf{z}$. Setzt man $\mathbf{d} = \mathbf{T}\mathbf{z}$ in (3.45) ein, so erhält man

$$\mathbf{d}^{\mathrm{T}}\Big(\nabla^2 L\Big)(\mathbf{x}^*, \boldsymbol{\lambda}^*)\mathbf{d} = \mathbf{d}^{\mathrm{T}}\mathbf{L}\mathbf{d} = \mathbf{z}^{\mathrm{T}}\mathbf{T}^{\mathrm{T}}\mathbf{L}\mathbf{T}\mathbf{z} > 0 \qquad \forall\, \mathbf{z} \in \mathbb{R}^{n-p},\ \mathbf{z} \neq \mathbf{0}. \tag{3.46}$$

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

$$\mathbf{L}_{\mathcal{X}} = \mathbf{T}^{\mathrm{T}}\mathbf{L}\mathbf{T}, \tag{3.47}$$

und die Überprüfung der Erfüllung von (3.45) reduziert sich auf die Prüfung der positiven Definitheit von $\mathbf{L}_{\mathcal{X}}$. Es sei nun λ ein Eigenwert von $\mathbf{L}_{\mathcal{X}}$ und $\mathbf{v}$ der zugehörige Eigenvektor (zugleich Links- und Rechtseigenvektor). Folglich gilt

$$0 = \mathbf{v}^{\mathrm{T}}(\lambda\mathbf{E} - \mathbf{L}_{\mathcal{X}})\mathbf{v} = \mathbf{v}^{\mathrm{T}}(\lambda\mathbf{T}^{\mathrm{T}}\mathbf{T} - \mathbf{T}^{\mathrm{T}}\mathbf{L}\mathbf{T})\mathbf{v} = \mathbf{v}^{\mathrm{T}}\mathbf{T}^{\mathrm{T}}(\lambda\mathbf{E} - \mathbf{L})\mathbf{T}\mathbf{v}. \tag{3.48}$$

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

Beispiel 3.4. Man betrachte das Optimierungsproblem (3.1) mit einer Gleichungsbeschränkung in der Form

$$\min_{\mathbf{x}\in\mathbb{R}^3} \quad f(\mathbf{x}) = x_1 + x_2^2 + x_2x_3 + 2x_3^2 \tag{3.49a}$$

$$\text{u.B.v.} \quad g(\mathbf{x}) = x_1^2 + x_2^2 + x_3^2 - 1 = 0\,. \tag{3.49b}$$

Die notwendigen Optimalitätsbedingungen erster Ordnung nach Satz 3.6 lauten

$$\frac{\partial}{\partial x_1}L(\mathbf{x}^*, \lambda^*) = 1 + 2\lambda^* x_1^* = 0 \tag{3.50a}$$

$$\frac{\partial}{\partial x_2}L(\mathbf{x}^*, \lambda^*) = 2x_2^* + x_3^* + 2\lambda^* x_2^* = 0 \tag{3.50b}$$

$$\frac{\partial}{\partial x_3}L(\mathbf{x}^*, \lambda^*) = x_2^* + 4x_3^* + 2\lambda^* x_3^* = 0 \tag{3.50c}$$

$$\frac{\partial}{\partial \lambda}L(\mathbf{x}^*, \lambda^*) = (x_1^*)^2 + (x_2^*)^2 + (x_3^*)^2 - 1 = 0\ . \tag{3.50d}$$

Mit dem regulären Punkt $\mathbf{x}^* = [1 \quad 0 \quad 0]^{\mathrm{T}}$ und dem Lagrange-Multiplikator $\lambda^* = -1/2$ ist eine Lösung von (3.50) gegeben. Zur Prüfung der Optimalitätsbedingung zweiter Ordnung gemäß Satz 3.8 wird die Matrix

$$\mathbf{L} = \Big(\nabla^2 L\Big)(\mathbf{x}^*, \boldsymbol{\lambda}^*) = \begin{bmatrix} -1 & 0 & 0 \\ 0 & 1 & 1 \\ 0 & 1 & 3 \end{bmatrix} \tag{3.51}$$

benötigt.

Um den Tangentialraum $\mathcal{T}_{\mathbf{x}^*}\mathcal{X}$ der Mannigfaltigkeit $\mathcal{X} = \{\mathbf{x} \in \mathbb{R}^3 \,|\, x_1^2 + x_2^2 + x_3^2 = 1\}$ gemäß (3.25) zu berechnen, bestimme man vorerst den Ausdruck

$$\left(\frac{\partial}{\partial \mathbf{x}} g\right)(\mathbf{x}^*) = \begin{bmatrix} 2x_1 & 2x_2 & 2x_3 \end{bmatrix}\Big|_{\mathbf{x}=\mathbf{x}^*} = \begin{bmatrix} 2 & 0 & 0 \end{bmatrix}. \tag{3.52}$$

Aus (3.52) und der Definition (3.25) 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 (3.45) von Satz 3.8 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 orthonormale Basisvektoren des Tangentialraumes $\mathcal{T}_{\mathbf{x}^*}\mathcal{X}$ und fasst man diese als Spaltenvektoren in der Matrix $\mathbf{T}$ zusammen, z. B.

$$\mathbf{T} = \begin{bmatrix} 0 & 0 \\ 1 & 0 \\ 0 & 1 \end{bmatrix}, \tag{3.53}$$

dann kann die Matrix $\mathbf{L}$ von (3.51) gemäß (3.47) wie folgt

$$\mathbf{L}_{\mathcal{X}} = \mathbf{T}^{\mathrm{T}}\mathbf{L}\mathbf{T} = \begin{bmatrix} 1 & 1 \\ 1 & 3 \end{bmatrix} \tag{3.54}$$

in den Tangentialraum projiziert werden. Da die Matrix $\mathbf{L}_{\mathcal{X}}$ positiv definit ist, folgt aus Satz 3.8, dass der Punkt $\mathbf{x}^* = [1 \quad 0 \quad 0]^{\mathrm{T}}$ ein striktes lokales Minimum des beschränkten Optimierungsproblems (3.49) ist.

Angenommen der Punkt $\mathbf{x}^*$ ist eine Lösung des beschränkten Optimierungsproblems (3.22) mit dem zugehörigen Lagrange-Multiplikator $\boldsymbol{\lambda}^*$. Dann lässt sich $\boldsymbol{\lambda}^*$ wie folgt interpretieren.

Satz 3.9 (Sensitivitätstheorem des Lagrange-Multiplikators bei Gleichungsbeschränkungen)**.** *Für $f, g_1, \ldots, g_p \in C^2$ betrachte man folgende Familie beschränkter Optimierungsprobleme*

$$\min_{\mathbf{x}\in\mathbb{R}^n} \quad f(\mathbf{x}) \tag{3.55a}$$

$$\text{u.B.v.} \quad \mathbf{g}(\mathbf{x}) = \mathbf{c}_g \tag{3.55b}$$

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 3.8 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 (3.55) 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*

$$\frac{\mathrm{d}}{\mathrm{d}\mathbf{c}_g} f(\mathbf{x}(\mathbf{c}_g))\bigg|_{\mathbf{c}_g=\mathbf{0}} = -(\boldsymbol{\lambda}^*)^{\mathrm{T}} \ . \tag{3.56}$$

Beweisskizze: Die notwendigen Optimalitätsbedingungen erster Ordnung für das Optimierungsproblem (3.55) lauten

$$(\nabla f)(\mathbf{x}) + (\nabla \mathbf{g})(\mathbf{x})\boldsymbol{\lambda} = \mathbf{0} \tag{3.57a}$$

$$\mathbf{g}(\mathbf{x}) = \mathbf{c}_g \ . \tag{3.57b}$$

Berechnet man die Jacobi-Matrix von (3.57) an der Stelle $(\mathbf{x}^*, \boldsymbol{\lambda}^*)$, also für $\mathbf{c}_g = \mathbf{0}$, dann erhält man (siehe auch (3.43))

$$\begin{bmatrix} (\nabla^2 L)(\mathbf{x}^*, \boldsymbol{\lambda}^*) & (\nabla \mathbf{g})(\mathbf{x}^*) \\ (\nabla \mathbf{g})^{\mathrm{T}}(\mathbf{x}^*) & \mathbf{0} \end{bmatrix} . \tag{3.58}$$

Da nach den Voraussetzungen von Satz 3.8 die Matrix $(\nabla^2 L)(\mathbf{x}^*, \boldsymbol{\lambda}^*)$ die Ungleichung $\mathbf{d}^{\mathrm{T}}(\nabla^2 L)(\mathbf{x}^*, \boldsymbol{\lambda}^*)\mathbf{d} > 0 \ \forall \mathbf{d} \in \mathcal{T}_{\mathbf{x}^*}\mathcal{X}, \ \mathbf{d} \neq \mathbf{0}$ erfüllt (striktes lokales Minimum an der Stelle $\mathbf{x}^*$) und die Matrix $(\nabla \mathbf{g})(\mathbf{x}^*)$ spaltenregulär ist ($\mathbf{x}^*$ ist ein regulärer Punkt), ist die Jacobi-Matrix (3.58) regulär.

> *Aufgabe* 3.6. Beweisen Sie diese Behauptung.

Mit Hilfe des Satzes über implizite Funktionen (Satz 1.4) kann daraus geschlossen werden, dass $\mathbf{x}(\mathbf{c}_g)$ und $\boldsymbol{\lambda}(\mathbf{c}_g)$ in einer Umgebung von $\mathbf{c}_g = \mathbf{0}$ stetig differenzierbare Funktionen in $\mathbf{c}_g$ sind.

Die Ableitung von (3.57b) nach $\mathbf{c}_g$ am Punkt $\mathbf{c}_g = \mathbf{0}$ liefert

$$\frac{\mathrm{d}\mathbf{g}(\mathbf{x}(\mathbf{c}_g))}{\mathrm{d}\mathbf{c}_g}\bigg|_{\mathbf{c}_g=\mathbf{0}} = (\nabla \mathbf{g})^{\mathrm{T}}(\mathbf{x}^*) \frac{\mathrm{d}\mathbf{x}}{\mathrm{d}\mathbf{c}_g}\bigg|_{\mathbf{c}_g=\mathbf{0}} = \mathbf{E}. \tag{3.59}$$

Wird die Transponierte von (3.57a) rechtsseitig mit $\frac{\mathrm{d}\mathbf{x}}{\mathrm{d}\mathbf{c}_g}$ multipliziert, so ergibt sich am Punkt $\mathbf{c}_g = \mathbf{0}$

$$(\nabla f)^{\mathrm{T}}(\mathbf{x}^*) \frac{\mathrm{d}\mathbf{x}}{\mathrm{d}\mathbf{c}_g}\bigg|_{\mathbf{c}_g=\mathbf{0}} + (\boldsymbol{\lambda}^*)^{\mathrm{T}} \underbrace{(\nabla \mathbf{g})^{\mathrm{T}}(\mathbf{x}^*) \frac{\mathrm{d}\mathbf{x}}{\mathrm{d}\mathbf{c}_g}\bigg|_{\mathbf{c}_g=\mathbf{0}}}_{\mathbf{E}} = \mathbf{0}. \tag{3.60}$$

Gemeinsam mit (3.59) folgt daraus

$$\frac{\mathrm{d}}{\mathrm{d}\mathbf{c}_g} f(\mathbf{x}(\mathbf{c}_g))\bigg|_{\mathbf{c}_g=\mathbf{0}} + (\boldsymbol{\lambda}^*)^{\mathrm{T}} = \mathbf{0} \tag{3.61}$$

und damit (3.56). □

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

$$J = J(\mathbf{x}) = \{j \in \mathbb{N} \,|\, 1 \leq j \leq q,\ h_j(\mathbf{x}) = 0\} \tag{3.62}$$

definiert. Wieder soll

$$\mathcal{X} = \{\mathbf{x} \in \mathbb{R}^n \,|\, g_i(\mathbf{x}) = 0,\ i = 1, \ldots, p,\ h_j(\mathbf{x}) \leq 0,\ j = 1, \ldots, q\} \tag{3.63}$$

die *zulässige Menge* genannt werden. Ferner beschreibt

$$\bar{\mathcal{X}} = \bar{\mathcal{X}}(\mathbf{x}) = \{\bar{\mathbf{x}} \in \mathcal{X} \,|\, h_j(\bar{\mathbf{x}}) = 0,\ j \in J(\mathbf{x})\} \tag{3.64}$$

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

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 $(\nabla g_i)(\mathbf{x})$, $i = 1, \ldots, p$ und $(\nabla h_j)(\mathbf{x})$, $j \in J$ mit J gemäß (3.62) linear unabhängig sind. D. h. die Bedingung

$$\operatorname{rang}\Big(\Big[[(\nabla g_i)(\mathbf{x})]_{i=1,\ldots,p} \quad [(\nabla h_j)(\mathbf{x})]_{j\in J}\Big]\Big) = p + |J|, \tag{3.65}$$

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

Satz 3.10 (Karush-Kuhn-Tucker (KKT) notwendige Optimalitätsbedingungen erster Ordnung)**.** *Angenommen, $\mathbf{x}^*$ sei ein lokales Minimum des Optimierungsproblems (3.1) bzw. (3.2)*

$$\min_{\mathbf{x}\in\mathbb{R}^n} \quad f(\mathbf{x}) \tag{3.66a}$$

$$\text{u.B.v.} \quad g_i(\mathbf{x}) = 0, \quad i = 1, \ldots, p \tag{3.66b}$$

$$h_i(\mathbf{x}) \leq 0, \quad i = 1, \ldots, q \tag{3.66c}$$

mit $f, g_1, \ldots, g_p, h_1, \ldots, h_q \in C^1$. Im Weiteren sei $\mathbf{x}^*$ ein regulärer Punkt der Beschränkungen (3.66b) und (3.66c). Dann existieren eindeutige Lagrange-Multiplikatoren $((\boldsymbol{\lambda}^*)^{\mathrm{T}}, (\boldsymbol{\mu}^*)^{\mathrm{T}})$ mit $\boldsymbol{\lambda}^* \in \mathbb{R}^p$ und $\boldsymbol{\mu}^* \in \mathbb{R}^q$ so, dass die Bedingungen

$$(\nabla f)(\mathbf{x}^*) + (\nabla \mathbf{g})(\mathbf{x}^*)\boldsymbol{\lambda}^* + (\nabla \mathbf{h})(\mathbf{x}^*)\boldsymbol{\mu}^* = \mathbf{0} \tag{3.67a}$$

$$\boldsymbol{\mu}^* \geq \mathbf{0} \tag{3.67b}$$

$$\mathbf{h}^{\mathrm{T}}(\mathbf{x}^*)\boldsymbol{\mu}^* = 0 \tag{3.67c}$$

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

Beweisskizze: Aus (3.66c), (3.67b) und (3.67c) folgt $h_i(\mathbf{x}^*)\mu_i^* = 0 \;\forall\, i = 1, \dots, q$. Außerdem folgt, dass eine Komponente μ_i^* nur dann von Null verschieden sein kann, wenn die zugehörige Ungleichungsbedingung aktiv ist, d. h. $h_i(\mathbf{x}^*) = 0$ gilt. Umgekehrt kann eine Ungleichungsbeschränkung $h_i(\mathbf{x}^*) \leq 0$ nur inaktiv sein, wenn der zugehörige Lagrange-Multiplikator $\mu_i^* = 0$ erfüllt. Diese so genannte *complementary slackness condition* besagt also, dass $h_i(\mathbf{x}^*) < 0$ stets $\mu_i^* = 0$ und $\mu_i^* > 0$ stets $h_i(\mathbf{x}^*) = 0$ impliziert.

Da $\mathbf{x}^*$ ein lokales Minimum des beschränkten Optimierungsproblems (3.66) beschreibt, ist es auch ein lokales Minimum für jenes Optimierungsproblem, bei dem alle aktiven Ungleichungsbeschränkungen durch Gleichungsbeschränkungen ersetzt werden. Dann gibt (3.67a) mit $\mu_i^* = 0$ falls $h_i(\mathbf{x}^*) < 0$ exakt die Bedingung (3.29) des bei gleichungsbeschränkten Problemen anwendbaren Satzes 3.6 wieder.

Um zu zeigen, dass $\boldsymbol{\mu}^* \geq \mathbf{0}$ gelten muss, schreibt man (3.66c) in

$$\mathbf{h}(\mathbf{x}) \leq \mathbf{c}_h \leq \mathbf{0} \tag{3.68}$$

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 3.9 (siehe auch Satz 3.13) 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

$$\frac{\mathrm{d}}{\mathrm{d}\mathbf{c}_h} f(\mathbf{x}(\mathbf{c}_h))\bigg|_{\mathbf{c}_h=\mathbf{0}} = -(\boldsymbol{\mu}^*)^{\mathrm{T}} \tag{3.69}$$

gilt. Hierbei gilt $\mu_i^* = 0 \;\forall\, i \in \{1, \dots, q\} \backslash J$, d. h. für alle inaktiven Ungleichungsbeschränkungen $h_i(\mathbf{x}^*) < 0$. Die Entwicklung von f in eine Taylorreihe um den optimalen Punkt liefert

$$\begin{aligned} f(\mathbf{x}(\mathbf{c}_h)) &= f(\mathbf{x}(\mathbf{0})) + \frac{\mathrm{d}}{\mathrm{d}\mathbf{c}_h} f(\mathbf{x}(\mathbf{c}_h))\bigg|_{\mathbf{c}_h=\mathbf{0}} \mathbf{c}_h + \mathcal{O}(\mathbf{c}_h^2) \\ &= f(\mathbf{x}(\mathbf{0})) - (\boldsymbol{\mu}^*)^{\mathrm{T}}\mathbf{c}_h + \mathcal{O}(\mathbf{c}_h^2)\ . \end{aligned} \tag{3.70}$$

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 $\mathbf{c}_h \to \mathbf{0}^-$ erfüllt sein muss. □

Man beachte, dass *nicht* jedes lokale Minimum die KKT-Bedingungen (3.67) erfüllt. Dies ist nur der Fall, wenn die Beschränkungen (3.66b) und (3.66c) gewisse Voraussetzungen erfüllen, die im Englischen auch als *constraint qualification (CQ)* bezeichnet werden, vgl. [3.2–3.4]. 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 $((\boldsymbol{\lambda}^*)^{\mathrm{T}}, (\boldsymbol{\mu}^*)^{\mathrm{T}})$ *eindeutig* sind. Es existieren auch andere CQ Bedingungen, die diese Eindeutigkeit nicht garantieren.

Aufgabe 3.7. Rechnen Sie mithilfe von Satz 3.10 die Beziehungen (2.120) nach. Dabei stellt ε_k einen Lagrange-Multiplikator dar.

Ein Problem bei der Berechnung von $\mathbf{x}^*$ gemäß Satz 3.10 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 3.5. Es wird das Optimierungsproblem (3.1) mit $q = 2$ Ungleichungsbeschränkungen in der Form

$$\min_{\mathbf{x}\in\mathbb{R}^3} \quad f(\mathbf{x}) = \frac{1}{2}(x_1^2 + x_2^2 + x_3^2) \tag{3.71a}$$

$$\text{u.B.v.} \quad h_1(\mathbf{x}) = x_1 + x_2 + x_3 + 3 \leq 0 \tag{3.71b}$$

$$h_2(\mathbf{x}) = x_1 \leq 0 \tag{3.71c}$$

betrachtet. Es handelt sich hierbei um ein konvexes Optimierungsproblem. Wegen $(\nabla h_1)(\mathbf{x}) = [1 \quad 1 \quad 1]^{\mathrm{T}}$ und $(\nabla h_2)(\mathbf{x}) = [1 \quad 0 \quad 0]^{\mathrm{T}}$ ist jeder zulässige Punkt ein regulärer Punkt. Die KKT-Bedingungen (3.67) lauten in diesem Fall

$$\underbrace{\begin{bmatrix} x_1^* \\ x_2^* \\ x_3^* \end{bmatrix}}_{(\nabla f)(\mathbf{x}^*)} + \underbrace{\begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix}}_{(\nabla h_1)(\mathbf{x}^*)} \mu_1^* + \underbrace{\begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix}}_{(\nabla h_2)(\mathbf{x}^*)} \mu_2^* = \mathbf{0} \tag{3.72a}$$

$$\mu_1^* \geq 0 \tag{3.72b}$$

$$\mu_2^* \geq 0 \tag{3.72c}$$

$$\mu_1^*(x_1^* + x_2^* + x_3^* + 3) + \mu_2^* x_1^* = 0 \tag{3.72d}$$

$$x_1^* + x_2^* + x_3^* + 3 \leq 0 \tag{3.72e}$$

$$x_1^* \leq 0. \tag{3.72f}$$

Nun können $2^q = 4$ 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 gemäß (3.72a) 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 $\mu_2^* \geq 0$. Die Lösung $x_2^* = x_3^* = 0$ von (3.72a) ist wiederum kein zulässiger Punkt, da $h_1(\mathbf{x}^*) \leq 0$ nicht erfüllt wird.

c) Die Ungleichungsbeschränkung $h_1(\mathbf{x}^*)$ ist aktiv und $h_2(\mathbf{x}^*)$ ist inaktiv, d. h. $h_1(\mathbf{x}^*) = x_1^* + x_2^* + x_3^* + 3 = 0$, $h_2(\mathbf{x}^*) = x_1^* < 0$, $\mu_1^* \geq 0$ 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. $h_1(\mathbf{x}^*) = x_1^* + x_2^* + x_3^* + 3 = 0$, $h_2(\mathbf{x}^*) = x_1^* = 0$, $\mu_1^* \geq 0$ und $\mu_2^* \geq 0$. Aus der ersten Zeile von (3.72a) folgen dementsprechend $\mu_1^* = 0$ und $\mu_2^* = 0$. Dies würde wieder auf die Lösung $x_2^* = x_3^* = 0$ von (3.72a) führen, welche keinen zulässigen Punkt darstellt.

Ob es sich beim Punkt $x_1^* = x_2^* = x_3^* = -1$ tatsächlich um ein Minimum handelt, kann basierend auf den nachfolgenden Ausführungen zu konvexen Optimierungsproblemen einfach geklärt werden.

Wenn die Funktion $f(\mathbf{x})$ und die zulässige Menge $\mathcal{X}$ konvex sind, dann sind die notwendigen Optimalitätsbedingungen erster Ordnung gemäß Satz 3.10 auch *hinreichend* für ein globales Minimum. Um dies zu sehen, beachte man zunächst wieder, dass die Gleichungsbeschränkungen $\mathbf{g}(\mathbf{x}) = \mathbf{0}$ für konvexes $\mathcal{X}$ affin in $\mathbf{x}$ sein müssen, d. h. es gilt gemäß (3.33)

$$(\mathbf{x} - \mathbf{x}^*)^{\mathrm{T}} (\nabla \mathbf{g})(\mathbf{x}^*) = \mathbf{0} \qquad \forall\, \mathbf{x} \in \mathcal{X},\ \mathbf{x}^* \in \mathcal{X}\ . \tag{3.73}$$

Außerdem müssen für die Ungleichungsbeschränkungen $\mathbf{h}(\mathbf{x}) \leq \mathbf{0}$

$$\mu_j^* \geq 0 \qquad \forall\, j \in J(\mathbf{x}^*) \tag{3.74a}$$

$$\mu_j^* = 0 \qquad \forall\, j \in \{1, \ldots, q\} \backslash J(\mathbf{x}^*) \tag{3.74b}$$

sowie für konvexes $\mathcal{X}$

$$(\mathbf{x} - \mathbf{x}^*)^{\mathrm{T}} (\nabla h_j)(\mathbf{x}^*) \leq 0 \qquad \forall\, j \in J(\mathbf{x}^*),\ \mathbf{x} \in \mathcal{X},\ \mathbf{x}^* \in \mathcal{X} \tag{3.74c}$$

gelten. Aus der Konvexität von $f(\mathbf{x})$ auf $\mathcal{X}$ folgt mit (3.67a), (3.73) und (3.74) die Ungleichung

$$\begin{aligned} f(\mathbf{x}) &\geq f(\mathbf{x}^*) + (\mathbf{x} - \mathbf{x}^*)^{\mathrm{T}} (\nabla f)(\mathbf{x}^*) \\ &= f(\mathbf{x}^*) - \underbrace{(\mathbf{x} - \mathbf{x}^*)^{\mathrm{T}} (\nabla \mathbf{g})(\mathbf{x}^*)}_{= \mathbf{0}} \boldsymbol{\lambda}^* - \underbrace{(\mathbf{x} - \mathbf{x}^*)^{\mathrm{T}} (\nabla \mathbf{h})(\mathbf{x}^*) \boldsymbol{\mu}^*}_{\leq 0} \\ &\geq f(\mathbf{x}^*) \qquad \forall\, \mathbf{x} \in \mathcal{X} \end{aligned} \tag{3.75}$$

und damit die globale Optimalität von $\mathbf{x}^*$.

Satz 3.11 (KKT notwendige Optimalitätsbedingungen zweiter Ordnung)**.** *Angenommen, $\mathbf{x}^*$ sei ein lokales Minimum des Optimierungsproblems (3.1) bzw. (3.2)*

$$\min_{\mathbf{x} \in \mathbb{R}^n} \quad f(\mathbf{x}) \tag{3.76a}$$

$$\text{u.B.v.} \quad g_i(\mathbf{x}) = 0, \quad i = 1, \ldots, p \tag{3.76b}$$

$$h_i(\mathbf{x}) \leq 0, \quad i = 1, \ldots, q \tag{3.76c}$$

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 (3.76b) und (3.76c). Dann existieren eindeutige Lagrange-Multiplikatoren $((\boldsymbol{\lambda}^*)^\mathrm{T}, (\boldsymbol{\mu}^*)^\mathrm{T})$ mit $\boldsymbol{\lambda}^* \in \mathbb{R}^p$ und $\boldsymbol{\mu}^* \in \mathbb{R}^q$ so, dass die Bedingungen

$$(\nabla f)(\mathbf{x}^*) + (\nabla \mathbf{g})(\mathbf{x}^*)\boldsymbol{\lambda}^* + (\nabla \mathbf{h})(\mathbf{x}^*)\boldsymbol{\mu}^* = \mathbf{0} \tag{3.77a}$$

$$\boldsymbol{\mu}^* \geq \mathbf{0} \tag{3.77b}$$

$$\mathbf{h}^\mathrm{T}(\mathbf{x}^*)\boldsymbol{\mu}^* = 0 \tag{3.77c}$$

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

$$\mathbf{d}^\mathrm{T} \underbrace{\left(\left(\nabla^2 f\right)(\mathbf{x}^*) + \sum_{i=1}^{p} \lambda_i^* \left(\nabla^2 g_i\right)(\mathbf{x}^*) + \sum_{j=1}^{q} \mu_j^* \left(\nabla^2 h_j\right)(\mathbf{x}^*) \right)}_{(\nabla^2 L)(\mathbf{x}^*, \boldsymbol{\lambda}^*, \boldsymbol{\mu}^*)} \mathbf{d} \geq 0 \qquad \forall\, \mathbf{d} \in \mathcal{T}_{\mathbf{x}^*}\bar{\mathcal{X}} \tag{3.78}$$

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 $L(\mathbf{x}^*, \boldsymbol{\lambda}^*, \boldsymbol{\mu}^*) = f(\mathbf{x}^*) + (\boldsymbol{\lambda}^*)^\mathrm{T}\mathbf{g}(\mathbf{x}^*) + (\boldsymbol{\mu}^*)^\mathrm{T}\mathbf{h}(\mathbf{x}^*)$ 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 3.12 (KKT hinreichende Optimalitätsbedingungen zweiter Ordnung)**.** *Gesucht ist das (lokale) Minimum des Optimierungsproblems (3.1) bzw. (3.2)*

$$\min_{\mathbf{x} \in \mathbb{R}^n} \quad f(\mathbf{x}) \tag{3.79a}$$

$$\text{u.B.v.} \quad g_i(\mathbf{x}) = 0, \quad i = 1, \dots, p \tag{3.79b}$$

$$h_i(\mathbf{x}) \leq 0, \quad i = 1, \dots, q \tag{3.79c}$$

mit $f, g_1, \dots, g_p, h_1, \dots, h_q \in C^2$. Wenn für einen regulären Punkt $\mathbf{x}^*$ der Beschränkungen (3.79b) und (3.79c), Größen $\mathbf{x}^* \in \mathbb{R}^n$, $\boldsymbol{\lambda}^* \in \mathbb{R}^p$ und $\boldsymbol{\mu}^* \in \mathbb{R}^q$ so existieren, dass gilt

$$(\nabla f)(\mathbf{x}^*) + (\nabla \mathbf{g})(\mathbf{x}^*)\boldsymbol{\lambda}^* + (\nabla \mathbf{h})(\mathbf{x}^*)\boldsymbol{\mu}^* = \mathbf{0} \tag{3.80a}$$

$$\boldsymbol{\mu}^* \geq \mathbf{0} \tag{3.80b}$$

$$\mathbf{h}^\mathrm{T}(\mathbf{x}^*)\boldsymbol{\mu}^* = 0 \tag{3.80c}$$

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

$$\mathbf{d}^{\mathrm{T}} \underbrace{\left(\Big(\nabla^2 f\Big)(\mathbf{x}^*) + \sum_{i=1}^{p} \lambda_i^* \Big(\nabla^2 g_i\Big)(\mathbf{x}^*) + \sum_{j=1}^{q} \mu_j^* \Big(\nabla^2 h_j\Big)(\mathbf{x}^*) \right)}_{(\nabla^2 L)(\mathbf{x}^*, \boldsymbol{\lambda}^*, \boldsymbol{\mu}^*)} \mathbf{d} > 0 \qquad \forall\, \mathbf{d} \in \mathcal{T}_{\mathbf{x}^*} \bar{\mathcal{X}},\ \mathbf{d} \neq \mathbf{0}\ , \tag{3.81}$$

$\bar{\mathcal{X}} = \{\mathbf{x} \in \mathbb{R}^n \,|\, g_i(\mathbf{x}) = 0, i = 1, \dots, p,\ h_j(\mathbf{x}) = 0,\ j \in J\}$ *sowie der Lagrangefunktion* $L(\mathbf{x}^*, \boldsymbol{\lambda}^*, \boldsymbol{\mu}^*) = f(\mathbf{x}^*) + (\boldsymbol{\lambda}^*)^{\mathrm{T}} \mathbf{g}(\mathbf{x}^*) + (\boldsymbol{\mu}^*)^{\mathrm{T}} \mathbf{h}(\mathbf{x}^*)$*, 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* $h_j(\mathbf{x}^*) = 0,\ j \in J$*.*

Im speziellen Fall $p + |J| = n$ ist folglich die Optimalitätsbedingung erster Ordnung gemäß Satz 3.10 auch hinreichend für ein striktes lokales Minimum. Es gilt dann nämlich $\mathcal{T}_{\mathbf{x}^*} \bar{\mathcal{X}} = \{\mathbf{0}\}$, so dass die Bedingungen (3.78) und (3.81) unabhängig von $(\nabla^2 L)(\mathbf{x}^*, \boldsymbol{\lambda}^*)$ erfüllt werden.

Um in allen anderen Fällen die Erfüllung von (3.78) und (3.81) zu überprüfen, kann analog zu den Bedingungen (3.43) und (3.45) vorgegangen werden, siehe dazu (3.46) bis (3.48). D. h. es wird die positive (Semi-) Definitheit der Projektion von $(\nabla^2 L)(\mathbf{x}^*, \boldsymbol{\lambda}^*, \boldsymbol{\mu}^*)$ in den Tangentialraum $\mathcal{T}_{\mathbf{x}^*} \bar{\mathcal{X}}$ untersucht.

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 3.13 (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*

$$\min_{\mathbf{x} \in \mathbb{R}^n} \quad f(\mathbf{x}) \tag{3.82a}$$

$$\text{u.B.v.} \quad \mathbf{g}(\mathbf{x}) = \mathbf{c}_g \tag{3.82b}$$

$$\mathbf{h}(\mathbf{x}) \leq \mathbf{c}_h \tag{3.82c}$$

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 3.12 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* (3.82) *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*

$$\frac{\mathrm{d}}{\mathrm{d}\mathbf{c}_g} f(\mathbf{x}(\mathbf{c}_g,\mathbf{c}_h))\bigg|_{\mathbf{c}_g=\mathbf{0},\mathbf{c}_h=\mathbf{0}} = -(\boldsymbol{\lambda}^*)^{\mathrm{T}} \tag{3.83a}$$

$$\frac{\mathrm{d}}{\mathrm{d}\mathbf{c}_h} f(\mathbf{x}(\mathbf{c}_g,\mathbf{c}_h))\bigg|_{\mathbf{c}_g=\mathbf{0},\mathbf{c}_h=\mathbf{0}} = -(\boldsymbol{\mu}^*)^{\mathrm{T}} \ . \tag{3.83b}$$

Aufgabe 3.8. Beweisen Sie Satz 3.13. **Hinweis:** Orientieren Sie sich dabei am Beweis von Satz 3.9.

Man beachte, dass Satz 3.13 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 3.13 $\mu_i^* = 0 \ \forall\, i \in \{1,\dots,q\}\backslash J$.

Beispiel 3.6. Ein Unternehmen produziert täglich x Einheiten eines Produkts mit einem Verkaufspreis P in €/Stück. Die variablen Produktionskosten betragen K_v in €/Stück. Für die Produktion besitzt das Unternehmen m gleiche Maschinen, wobei jede Maschine täglich x^+ Produkte fertigen kann.

a) Wieviele Produkte soll das Unternehmen mit den vorhandenen m Maschinen täglich fertigen, um seinen Deckungsbeitrag zu maximieren?

b) Dem Unternehmen werden einige weitere Maschinen des gleichen Types zur Miete angeboten. Bei welchem Tagesmietsatz M in €/Maschine sollte das Angebot angenommen werden?

Es ist zunächst die Minimierungsaufgabe

$$\min_{x\in\mathbb{R}} \quad f(x) = -(P-K_v)x \tag{3.84a}$$

$$\text{u.B.v.} \quad h_1(x) = -x \le 0 \tag{3.84b}$$

$$h_2(x) = \frac{x}{x^+} - m \le 0 \tag{3.84c}$$

zu lösen. Es handelt sich hierbei um ein lineares Programm. Die KKT-Bedingungen für (3.84) lauten

$$\underbrace{-(P-K_v)}_{(\nabla f)(x^*)} + \underbrace{(-1)}_{(\nabla h_1)(x^*)}\mu_1^* + \underbrace{\frac{1}{x^+}}_{(\nabla h_2)(x^*)}\mu_2^* = 0 \tag{3.85a}$$

$$\mu_1^* \ge 0 \tag{3.85b}$$

$$\mu_2^* \ge 0 \tag{3.85c}$$

$$-x^*\mu_1^* + \left(\frac{x^*}{x^+} - m\right)\mu_2^* = 0 \tag{3.85d}$$

$$-x^* \le 0 \tag{3.85e}$$

$$\frac{x^*}{x^+} - m \leq 0 \tag{3.85f}$$

mit den Lagrange-Multiplikatoren μ_1^* und μ_2^*. Für $m > 0$ kann nur eine der beiden Ungleichungsnebenbedingungen (3.84b) und (3.84c) aktiv sein. Es müssen daher nur zwei Fälle unterschieden werden:

Im Fall $h_1(x^*) = -x^* = 0$ und $h_2(x^*) = -m < 0$ folgt aus (3.85d) $\mu_2^* = 0$ und aus (3.85a)

$$\mu_1^* = -\frac{(\nabla f)(x^*)}{(\nabla h_1)(x^*)} = -(P - K_v) \ . \tag{3.86}$$

Die Bedingung $\mu_1^* \geq 0$ (siehe (3.85b)) wird nur für $K_v \geq P$ erfüllt. Wenn also die variablen Produktionskosten K_v den Verkaufspreis P übersteigen, sollte nichts produziert werden und das Angebot zur Maschinenmiete sollte ausgeschlagen werden.

Im Fall $h_1(x^*) = -x^* < 0$ und $h_2(x^*) = x^*/x^+ - m = 0$ folgt aus (3.85d) $\mu_1^* = 0$ und aus (3.85a)

$$\mu_2^* = -\frac{(\nabla f)(x^*)}{(\nabla h_2)(x^*)} = (P - K_v)x^+ \ . \tag{3.87}$$

Die Bedingung $\mu_2^* \geq 0$ (siehe (3.85c)) wird nur für $P \geq K_v$ erfüllt. Wenn also der Verkaufspreis P die variablen Produktionskosten K_v übersteigt, sollten mit den verfügbaren m Maschinen täglich $x^* = x^+ m$ Produkte erzeugt werden. In diesem Fall sollte das Angebot zur Maschinenmiete nur dann angenommen (und zur Produktionssteigerung genützt) werden, wenn der Mietpreis M in €/Maschine den zusätzlich (je Maschine) erwirtschafteten Deckungsbeitrag $(P - K_v)x^+$ nicht übersteigt. Der Grenzpreis für die Maschinentagesmiete entspricht folglich dem Lagrange-Multiplikator μ_2^* in €/Maschine. Dieser Grenzpreis wird häufig auch als Schattenpreis der Engpasskapazität bezeichnet.

Abschließend sollen kurz die Einheiten einiger verwendeter Größen besprochen werden. Die Kostenfunktion $f(x)$ und auch die Lagrangefunktion haben die Einheit €. Die Ungleichungsnebenbedingung $h_1(x)$ ist in der Einheit Produkt formuliert und μ_1 besitzt die Einheit €/Produkt. Die Ungleichungsnebenbedingung $h_2(x)$ ist in der Einheit Maschine formuliert und μ_2 besitzt die Einheit €/Maschine.

3.2 Rechnergestützte Optimierungsverfahren

Als Ausgangspunkt betrachte man wiederum das beschränkte Optimierungsproblem (3.2)

$$\min_{\mathbf{x} \in \mathbb{R}^n} \quad f(\mathbf{x}) \tag{3.88a}$$

$$\text{u.B.v.} \quad \mathbf{g}(\mathbf{x}) = \mathbf{0} \tag{3.88b}$$

$$\mathbf{h}(\mathbf{x}) \leq \mathbf{0} \tag{3.88c}$$

mit p Gleichungsbeschränkungen $g_1(\mathbf{x}), \ldots, g_p(\mathbf{x})$, q Ungleichungsbeschränkungen $h_1(\mathbf{x})$, $\ldots, h_q(\mathbf{x})$ und n Optimierungsvariablen $x_1, \ldots, x_n$. Da die Bestimmung eines (lokal) optimalen Punktes $\mathbf{x}^*$ von (3.88) 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 (3.88) 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 3.10 weiß man, dass die notwendigen Optimalitätsbedingungen für ein lokales Minimum durch

$$(\nabla f)(\mathbf{x}^*) + (\nabla \mathbf{h})(\mathbf{x}^*)\boldsymbol{\mu}^* = \mathbf{0} \tag{3.89a}$$
$$h_i(\mathbf{x}^*) = 0, \qquad i \in J \tag{3.89b}$$
$$h_i(\mathbf{x}^*) < 0, \qquad i \in \{1, \dots, q\} \backslash J \tag{3.89c}$$
$$\mu_i^* \geq 0, \qquad i \in J \tag{3.89d}$$
$$\mu_i^* = 0, \qquad i \in \{1, \dots, q\} \backslash J \tag{3.89e}$$

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 beinhaltet, dann kann die Lösung als Kandidat für das Minimum von (3.89) angesehen werden.

Die Idee der Methode der aktiven Beschränkungen (englisch: *active set method*) beruht darauf, in jedem Iterationsschritt k eine *Arbeitsmenge* W_k festzulegen, welche die Indizes der am aktuellen Iterationspunkt $\mathbf{x}_k$ als aktiv betrachteten 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 W_k. Um zum nächsten Iterationspunkt $\mathbf{x}_{k+1}$ zu gelangen, erfolgt eine Bewegung entlang der durch die Arbeitsmenge W_k definierten Mannigfaltigkeit. Ziel ist es, so zu einem hinsichtlich der Kostenfunktion verbesserten Punkt zu gelangen. Die Bewegung auf der Mannigfaltigkeit kann z. B. mit der *Gradienten-Projektionsmethode* (siehe Abschnitt 3.2.2), mit der *reduzierten Gradientenmethode* (siehe Abschnitt 3.2.3) oder bei quadratischen Programmen durch analytische Lösung (siehe Beispiel 3.7 am Ende dieses Abschnitts) erfolgen. Wie diese Bewegung gewählt wird, beeinflusst auch wesentlich das Konvergenzverhalten des Lösungsalgorithmus. Die Methode der aktiven Beschränkungen ist also kein eigenständiges Lösungsverfahren

für beschränkte Optimierungsprobleme sondern lediglich eine Strategie für den Umgang mit Ungleichungsbeschränkungen im Zuge einer iterativen Lösungssuche.

Angenommen, W_k bezeichne die Arbeitsmenge, also die Indexmenge der als aktiv betrachteten Ungleichungsbeschränkungen am aktuellen Punkt $\mathbf{x}_k$. Dann besteht die Aufgabe im aktuellen Iterationsschritt k darin, eine Lösung $\mathbf{x}_{k+1}$ des Optimierungsproblems

$$\mathbf{x}_{k+1} = \arg\min_{\mathbf{x}\in\mathbb{R}^n} f(\mathbf{x}) \qquad \text{Kostenfunktion} \tag{3.90a}$$

$$\text{u.B.v.} \quad h_i(\mathbf{x}) = 0, \quad i \in W_k \qquad \text{aktive Ungleichungsbeschränkungen} \tag{3.90b}$$

zu finden. Dazu kann das Gleichungssystem

$$(\nabla f)(\mathbf{x}_{k+1}) + \sum_{i\in W_k} \lambda^*_{k,i} (\nabla h_i)(\mathbf{x}_{k+1}) = \mathbf{0} \tag{3.91a}$$

$$h_i(\mathbf{x}_{k+1}) = 0, \quad i \in W_k \tag{3.91b}$$

nach $\mathbf{x}_{k+1}$ und $\lambda^*_{k,i}$, $i \in W_k$ gelöst werden. Wenn nun $\lambda^*_{k,i} \geq 0$ für alle $i \in W_k$ und $h_i(\mathbf{x}_{k+1}) < 0$ für alle $i \in \{1,\dots,q\}\backslash W_k$, dann ist $\mathbf{x}_{k+1}$ eine mögliche lokale Lösung des beschränkten Optimierungsproblems (3.89).

Existiert hingegegen ein $j \in W_k$, für das gilt $\lambda^*_{k,j} < 0$, dann kann die Kostenfunktion weiter reduziert werden, indem die Beschränkung $h_j(\mathbf{x})$ inaktiv gesetzt wird, d. h. der Index j wird aus der neuen Indexmenge W_{k+1} entfernt. Dies folgt unmittelbar aus dem Beweis von Satz 3.10. Wird statt der aktiven Ungleichungsbeschränkung $h_j(\mathbf{x}) = 0$ die Gleichungsbeschränkung $h_j(\mathbf{x}) = c_h$ mit einem kleinen Wert $c_h < 0$ verwendet, d. h. $\mathbf{x}$ wird vom Rand in das Innere des Gebietes der Ungleichungsbeschränkung $h_j(\mathbf{x}) \leq 0$ bewegt, dann ändert sich für $\lambda^*_{k,j} < 0$ und $\mathbf{x}(0) = \mathbf{x}_{k+1}$ die Kostenfunktion in der Form

$$f(\mathbf{x}(c_h)) \approx f(\mathbf{x}_{k+1}) + \underbrace{\frac{\mathrm{d}}{\mathrm{d}c_h} f(\mathbf{x}(c_h))\Big|_{c_h=0}}_{-\lambda^*_{k,j}>0} \underbrace{c_h}_{<0} < f(\mathbf{x}_{k+1}) \; . \tag{3.92}$$

Dies zeigt also, dass durch eine Bewegung in das Innere des Gebietes der Ungleichungsbeschränkung $h_j(\mathbf{x}) < 0$ die Kostenfunktion weiter minimiert werden kann. Abbildung 3.3 veranschaulicht diesen Sachverhalt für die Ungleichungsbeschränkung $h_1(\mathbf{x}) \leq 0$.

Natürlich kann es umgekehrt passieren, dass durch die iterative Lösung des beschränkten Optimierungsproblems eine in der Arbeitsmenge W_k als inaktiv erachtete Ungleichungsbeschränkung vom neuen Punkt $\mathbf{x}_{k+1}$ verletzt wird, d. h. es existiert ein $j \in \{1,\dots,q\}\backslash W_k$ so, dass $h_j(\mathbf{x}_{k+1}) > 0$. In diesem Fall muss die neue Indexmenge W_{k+1} um den Index j dieser Ungleichungsbeschränkung erweitert werden. Alternativ ist es möglich, in jeder Iteration nur solche Bewegungen zuzulassen die zu zulässigen Punkten $\mathbf{x}_{k+1}$ führen, d. h. es gilt stets $\mathbf{x}_{k+1} \in \mathcal{X}$.

Der nachfolgende Satz liefert eine Aussage zur Konvergenz der Methode der aktiven Beschränkungen. Sein Beweis ist in [3.5] zu finden.

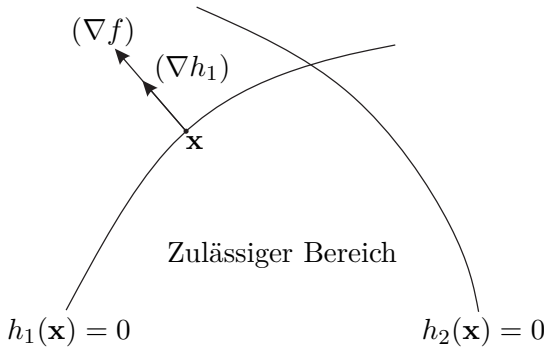


Abbildung 3.3: Zur Deaktivierung von Ungleichungsbeschränkungen.

Satz 3.14 (Konvergenz der Methode der aktiven Beschränkungen)**.** *Angenommen für aufeinanderfolgende Arbeitsmengen W_k ist das Optimierungsproblem (3.90) wohldefiniert und es besitzt eine eindeutige Lösung, die $\lambda^*_{k,i} \neq 0 \; \forall \, i \in W_k$ erfüllt. Wird diese eindeutige Lösung in jeder Iteration k exakt berechnet, so konvergiert die Methode der aktiven Beschränkungen gegen die Lösung des zugrundeliegenden beschränkten Optimierungsproblems.*

Der Fall, dass für eine Ungleichungsbeschränkung gleichzeitig $h_i(\mathbf{x}_{k+1}) = 0$ und $\lambda^*_{k,i} = 0$ gelten, wird als *degenerierte Lösung* bezeichnet. In diesem Fall ist die Ungleichungsbeschränkung zwar aktiv, im Sinne von Satz 3.13 ist aber der optimale Kostenfunktionswert dennoch insensitiv gegenüber einer Änderung dieser Beschränkung.

Eine Schwierigkeit bei der praktischen Anwendung von Satz 3.14 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. Wie in [3.5] beschrieben, gibt es dazu Erweiterungen des hier vorgestellten Basisalgorithmus.

Beispiel 3.7. Gegeben ist das quadratische Optimierungsproblem

$$\min_{\mathbf{x}\in\mathbb{R}^2} \quad \frac{1}{2}\mathbf{x}^{\mathrm{T}}\mathbf{H}\mathbf{x} + \mathbf{c}^{\mathrm{T}}\mathbf{x} \tag{3.93a}$$

$$\text{u.B.v.} \quad \mathbf{a}_1^{\mathrm{T}}\mathbf{x} - b_1 = -x_1 + 2x_2 - 2 \leq 0 \qquad \text{Ungleichungsbeschr. U1} \tag{3.93b}$$

$$\mathbf{a}_2^{\mathrm{T}}\mathbf{x} - b_2 = x_1 + 2x_2 - 6 \leq 0 \qquad \text{Ungleichungsbeschr. U2} \tag{3.93c}$$

$$\mathbf{a}_3^{\mathrm{T}}\mathbf{x} - b_3 = x_1 - 2x_2 - 2 \leq 0 \qquad \text{Ungleichungsbeschr. U3} \tag{3.93d}$$

$$\mathbf{a}_4^{\mathrm{T}}\mathbf{x} - b_4 = -x_1 \leq 0 \qquad \text{Ungleichungsbeschr. U4} \tag{3.93e}$$

$$\mathbf{a}_5^{\mathrm{T}}\mathbf{x} - b_5 = -x_2 \leq 0 \qquad \text{Ungleichungsbeschr. U5} \tag{3.93f}$$

mit $\mathbf{H} = \begin{bmatrix} 2 & 0 \\ 0 & 2 \end{bmatrix}$ und $\mathbf{c} = [-2 \quad -5]^{\mathrm{T}}$.

Zur Anwendung der Methode der aktiven Beschränkungen wird nun der jeweils

nächste Iterationspunkt $\mathbf{x}_{k+1}$ in der Form $\mathbf{x}_{k+1} = \mathbf{x}_k + \mathbf{s}_k^*$ angesetzt. Der optimale Schritt $\mathbf{s}_k^*$ folgt aus dem unterlagerten quadratischen Optimierungsproblem

$$\min_{\mathbf{s}_k \in \mathbb{R}^2} \quad \frac{1}{2}(\mathbf{x}_k + \mathbf{s}_k)^{\mathrm{T}}\mathbf{H}(\mathbf{x}_k + \mathbf{s}_k) + \mathbf{c}^{\mathrm{T}}(\mathbf{x}_k + \mathbf{s}_k) \tag{3.94a}$$

$$\text{u.B.v.} \quad \mathbf{a}_i^{\mathrm{T}}(\mathbf{x}_k + \mathbf{s}_k) - b_i = \mathbf{a}_i^{\mathrm{T}}\mathbf{s}_k = 0, \ \forall\, i \in W_k \ . \tag{3.94b}$$

Die KKT-Bedingungen für (3.94) lauten mit der Matrix $\mathbf{A}_k$, deren Spalten durch $\mathbf{a}_i,\ i \in W_k$ gegeben sind, und den Lagrange-Multiplikatoren $\boldsymbol{\lambda}_k$

$$\mathbf{H}\mathbf{s}_k^* + \mathbf{A}_k\boldsymbol{\lambda}_k^* = -\mathbf{H}\mathbf{x}_k - \mathbf{c} \tag{3.95a}$$

$$\mathbf{A}_k^{\mathrm{T}}\mathbf{s}_k^* = \mathbf{0} \ . \tag{3.95b}$$

Als Startpunkt für die Methode der aktiven Beschränkungen wird der zulässige Punkt $\mathbf{x}_0 = [2 \quad 0]^{\mathrm{T}}$ gewählt, an dem die Ungleichungsbeschränkungen U3 und U5 aktiv sind. Die Arbeitsmenge W_0 der in der ersten Iteration aktiven Ungleichungsbeschränkungen lautet damit $W_0 = \{3, 5\}$. Für $k = 0$ erhält man als Lösung von (3.95)

$$\begin{bmatrix} 2 & 0 & 1 & 0 \\ 0 & 2 & -2 & -1 \\ 1 & -2 & 0 & 0 \\ 0 & -1 & 0 & 0 \end{bmatrix} \begin{bmatrix} \mathbf{s}_0^* \\ \boldsymbol{\lambda}_0^* \end{bmatrix} = \begin{bmatrix} -2 \\ 5 \\ 0 \\ 0 \end{bmatrix} \tag{3.96}$$

die Größen $\mathbf{s}_0^* = \mathbf{0}$ und $\boldsymbol{\lambda}_0^* = [-2 \quad -1]^{\mathrm{T}}$ und somit $\mathbf{x}_1 = \mathbf{x}_0 = [2 \quad 0]^{\mathrm{T}}$. Nun wird die Ungleichung U3 (Lagrange-Multiplikator mit negativstem Wert) inaktiv gesetzt ($W_1 = \{5\}$) und (3.95) erneut für $k = 1$ gelöst, d. h. aus

$$\begin{bmatrix} 2 & 0 & 0 \\ 0 & 2 & -1 \\ 0 & -1 & 0 \end{bmatrix} \begin{bmatrix} \mathbf{s}_1^* \\ \lambda_1^* \end{bmatrix} = \begin{bmatrix} -2 \\ 5 \\ 0 \end{bmatrix} \tag{3.97}$$

folgen $\mathbf{s}_1^* = [-1 \quad 0]^{\mathrm{T}}$ und $\lambda_1^* = -5$. Damit kann der neue Iterationspunkt $\mathbf{x}_2 = \mathbf{x}_1 + \mathbf{s}_1^*$ berechnet werden, vorausgesetzt es werden keine inaktiven Ungleichungsbeschränkungen verletzt. Um diesen Fall zu berücksichtigen, wird eine Schrittweite $\alpha_k > 0$ verwendet und so nach oben beschränkt, dass $\mathbf{x}_{k+1} = \mathbf{x}_k + \alpha_k\mathbf{s}_k^* \in \mathcal{X}$ gilt. Die Wahl der Schrittweite α_k erfolgt nun auf Basis folgender Überlegungen. Wenn für alle inaktiven (affinen) Ungleichungsbeschränkungen gilt $\mathbf{a}_j^{\mathrm{T}}\mathbf{s}_k^* \leq 0$, dann kann $\alpha_k > 0$ beliebig gewählt werden ohne eine Ungleichung $\mathbf{a}_j^{\mathrm{T}}(\mathbf{x}_k + \alpha_k\mathbf{s}_k^*) \leq b_j,\ j \notin W_k$ zu verletzen. Falls hingegen $\mathbf{a}_j^{\mathrm{T}}\mathbf{s}_k^* > 0$, dann ist die Schrittweite durch $\alpha_k \leq \frac{b_j - \mathbf{a}_j^{\mathrm{T}}\mathbf{x}_k}{\mathbf{a}_j^{\mathrm{T}}\mathbf{s}_k^*}$ begrenzt. Da das Minimum der in α_k quadratischen Funktion $\frac{1}{2}(\mathbf{x}_k + \alpha_k\mathbf{s}_k^*)^{\mathrm{T}}\mathbf{H}(\mathbf{x}_k + \alpha_k\mathbf{s}_k^*) + \mathbf{c}^{\mathrm{T}}(\mathbf{x}_k + \alpha_k\mathbf{s}_k^*)$

für $\alpha_k = 1$ erreicht wird, folgt die Wahl der Schrittweite zu

$$\alpha_k = \min\left\{1, \min_{j \notin W_k, \mathbf{a}_j^{\mathrm{T}} \mathbf{s}_k^* > 0} \frac{b_j - \mathbf{a}_j^{\mathrm{T}} \mathbf{x}_k}{\mathbf{a}_j^{\mathrm{T}} \mathbf{s}_k^*}\right\}. \tag{3.98}$$

Im vorliegenden Fall gilt $\alpha_1 = \min\{1, \underbrace{4}_{\text{U1}}, \underbrace{2}_{\text{U4}}\} = 1$ und damit $\mathbf{x}_2 = [1 \quad 0]^{\mathrm{T}}$. Da die optimale Schrittweite $\alpha_1 = 1$ möglich ist, muss (3.95) nicht erneut gelöst werden. Es kann direkt die Ungleichung U5 (Lagrange-Multiplikator ist negativ) inaktiv gesetzt ($W_2 = \{\ \}$) und das unbeschränkte Optimierungsproblem zu $\mathbf{s}_2^* = [0 \quad 2.5]^{\mathrm{T}}$ gelöst werden. Die maximale Schrittweite gemäß (3.98) errechnet sich zu $\alpha_2 = \min\{1, \underbrace{0.6}_{\text{U1}}, \underbrace{1}_{\text{U2}}\} = 0.6$ und damit folgt $\mathbf{x}_3 = [1 \quad 1.5]^{\mathrm{T}}$. An diesem Punkt ist die Ungleichung U1 aktiv (der Wert $\alpha_2 = 0.6$ wurde durch die Ungleichungsbeschränkung U1 determiniert), weshalb die Arbeitsmenge der aktiven Beschränkungen zu $W_3 = \{1\}$ gesetzt wird. Aus (3.95) folgt mit

$$\begin{bmatrix} 2 & 0 & -1 \\ 0 & 2 & 2 \\ -1 & 2 & 0 \end{bmatrix} \begin{bmatrix} \mathbf{s}_3^* \\ \lambda_3^* \end{bmatrix} = \begin{bmatrix} 0 \\ 2 \\ 0 \end{bmatrix} \tag{3.99}$$

die Lösung $\mathbf{s}_3^* = [0.4 \quad 0.2]^{\mathrm{T}}$ und $\lambda_3^* = 0.8$. Die Schrittweite α_3 folgt aus der Beziehung (3.98) zu $\alpha_3 = \min\{1, \underbrace{2.5}_{\text{U2}}\} = 1$. Daher und auf Grund des positiven Lagrange-Multiplikators $\lambda_3^* = 0.8$ stellt $\mathbf{x}^* = \mathbf{x}_3 + \alpha_3 \mathbf{s}_3^* = [1.4 \quad 1.7]^{\mathrm{T}}$ die optimale Lösung von (3.93) dar. Der Lagrange-Multiplikator für das quadratische Programm (3.93) lautet $\boldsymbol{\mu}^* = [\lambda_3^* \quad 0 \quad 0 \quad 0 \quad 0]^{\mathrm{T}}$.

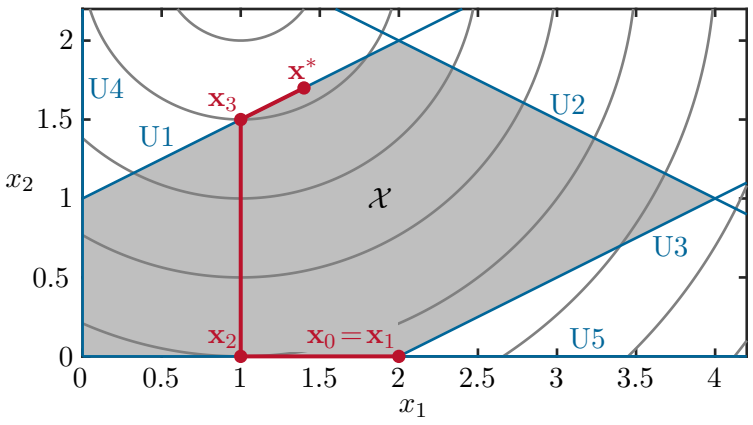


Abbildung 3.4: Anwendung der Methode der aktiven Beschränkungen auf das beschränkte quadratische Optimierungsproblem (3.93).

Abbildung 3.4 zeigt in Grau Höhenlinien der Kostenfunktion (3.93a), in Blau Linien entlang denen eine der Ungleichungsbeschränkungen (3.93b)–(3.93f) aktiv ist, in Grau die zulässige Menge $\mathcal{X}$ und in Rot den Linienzug dem die Methode der aktiven Beschränkungen bei der Lösungssuche folgt.

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

$$\min_{\mathbf{x}\in\mathbb{R}^n} \quad f(\mathbf{x}) \tag{3.100a}$$

$$\text{u.B.v.} \quad g_i(\mathbf{x}) = \mathbf{a}_{g,i}^{\mathrm{T}}\mathbf{x} - b_{g,i} = 0, \qquad i = 1,\dots,p \tag{3.100b}$$

$$h_i(\mathbf{x}) = \mathbf{a}_{h,i}^{\mathrm{T}}\mathbf{x} - b_{h,i} \leq 0, \qquad i = 1,\dots,q \tag{3.100c}$$

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

$$\bar{\mathbf{g}}(\mathbf{x}) = \begin{bmatrix} [g_i(\mathbf{x})]_{i=1,\dots,p} \\ [h_i(\mathbf{x})]_{i\in W} \end{bmatrix}, \tag{3.101}$$

zusammengefasst, dessen (konstante) Jacobi-Matrix mit

$$\mathbf{A} = (\nabla\bar{\mathbf{g}})(\mathbf{x}) = \begin{bmatrix} [\mathbf{a}_{g,i}]_{i=1,\dots,p} & [\mathbf{a}_{h,i}]_{i\in W} \end{bmatrix} \in \mathbb{R}^{n\times\bar{p}} \tag{3.102}$$

abgekürzt wird. Folglich wird der Tangentialraum der durch die aktuell aktiven Beschränkungen definierten Mannigfaltigkeit $\bar{\mathcal{X}} = \{\mathbf{x} \in \mathcal{X} \,|\, \bar{\mathbf{g}}(\mathbf{x}) = \mathbf{0}\}$ von den Vektoren des Nullraums $\mathrm{Kern}(\mathbf{A}^{\mathrm{T}})$ aufgespannt, d. h.

$$\mathcal{T}_{\mathbf{x}_k}\bar{\mathcal{X}} = \mathcal{T}\bar{\mathcal{X}} = \{\mathbf{d} \in \mathbb{R}^n \,|\, \mathbf{A}^{\mathrm{T}}\mathbf{d} = \mathbf{0}\} \tag{3.103}$$

(vgl. auch (3.25)). Da die Erfüllung der LICQ Bedingung vorausgesetzt wird, ist die Matrix $\mathbf{A}$ spaltenregulär, d. h. $\mathrm{rang}(\mathbf{A}) = \bar{p}$, und es gilt $\dim(\mathrm{Kern}(\mathbf{A}^{\mathrm{T}})) = n - \bar{p}$. Der Gradient $(\nabla f)(\mathbf{x}_k)$ im Iterationsschritt k steht nun im Allgemeinen nicht orthogonal auf $\mathcal{T}\bar{\mathcal{X}}$. Aufgrund von $\mathbb{R}^n = \mathrm{Kern}(\mathbf{A}^{\mathrm{T}}) \oplus \mathrm{Bild}(\mathbf{A})$ mit $\mathrm{Bild}(\mathbf{A})$ als dem Bild von $\mathbf{A}$ lässt sich der negative Gradient $-(\nabla f)(\mathbf{x}_k)$ immer in der Form

$$-(\nabla f)(\mathbf{x}_k) = \mathbf{d}_k + \mathbf{A}\boldsymbol{\sigma}_k \tag{3.104}$$

für geeignete $\mathbf{d}_k \in \mathcal{T}\bar{\mathcal{X}}$ und $\boldsymbol{\sigma}_k \in \mathbb{R}^{\bar{p}}$ anschreiben. Wird (3.104) linksseitig mit $\mathbf{A}^{\mathrm{T}}$ multipliziert, so ergibt sich aufgrund der Spaltenregularität von $\mathbf{A}$ die Beziehung

$$\boldsymbol{\sigma}_k = -\Big(\mathbf{A}^{\mathrm{T}}\mathbf{A}\Big)^{-1}\mathbf{A}^{\mathrm{T}}(\nabla f)(\mathbf{x}_k) \ . \tag{3.105}$$

Einsetzen von (3.105) in (3.104) führt auf den projizierten Gradient $\mathbf{d}_k$ in der Form

$$\mathbf{d}_k = -\mathbf{P}_k(\nabla f)(\mathbf{x}_k) \quad \text{mit} \quad \mathbf{P}_k = \mathbf{E} - \mathbf{A}\Big(\mathbf{A}^{\mathrm{T}}\mathbf{A}\Big)^{-1}\mathbf{A}^{\mathrm{T}} \ . \tag{3.106}$$

$\mathbf{P}_k$ 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 3.9. Zeigen Sie, dass eine Projektionsmatrix $\mathbf{P}$ die Eigenschaften $\mathbf{P}^{\mathrm{T}} = \mathbf{P}$ sowie $\mathbf{P}^2 = \mathbf{P}$ erfüllt.

Aufgabe 3.10. Zeigen Sie, dass der projizierte Gradient $\mathbf{d}_k$ auch durch Lösen des beschränkten Optimierungsproblems

$$\min_{\mathbf{d}_k \in \mathbb{R}^n} \quad \|(\nabla f)(\mathbf{x}_k) + \mathbf{d}_k\|_2^2 \tag{3.107a}$$

$$\text{u.B.v.} \quad \mathbf{A}^{\mathrm{T}}\mathbf{d}_k = \mathbf{0} \tag{3.107b}$$

berechnet werden kann.

Aufgabe 3.11. Zeigen Sie, dass $\boldsymbol{\sigma}_k$ gemäß (3.105) auch durch Lösen des unbeschränkten Optimierungsproblems

$$\min_{\boldsymbol{\sigma}_k \in \mathbb{R}^{\bar{p}}} \quad \|(\nabla f)(\mathbf{x}_k) + \mathbf{A}\boldsymbol{\sigma}_k\|_2^2 \tag{3.108}$$

berechnet werden kann.

Wenn $\mathbf{d}_k \neq \mathbf{0}$, dann gilt wegen (3.104) und $\mathbf{A}^{\mathrm{T}}\mathbf{d}_k = \mathbf{0}$

$$\mathbf{d}_k^{\mathrm{T}}(\nabla f)(\mathbf{x}_k) = \mathbf{d}_k^{\mathrm{T}}(-\mathbf{d}_k - \mathbf{A}\boldsymbol{\sigma}_k) = -\mathbf{d}_k^{\mathrm{T}}\mathbf{d}_k = -\|\mathbf{d}_k\|_2^2 < 0 \ . \tag{3.109}$$

Folglich ist mit $\mathbf{d}_k$ eine *zulässige Abstiegsrichtung* des beschränkten Optimierungsproblems am Punkt $\mathbf{x}_k$ gefunden. In weiterer Folge muss lediglich die Schrittweite α_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 α_k durch die am Punkt $\mathbf{x}_k$ inaktiven Ungleichungsnebenbedingungen nach oben hin beschränkt sein kann. Wenn jedoch $\mathbf{d}_k = \mathbf{0}$ gilt, dann folgt aus (3.104)

$$(\nabla f)(\mathbf{x}_k) + \mathbf{A}\boldsymbol{\sigma}_k = \mathbf{0} \ . \tag{3.110}$$

Dies entspricht der KKT-Bedingung (3.67a) von Satz 3.10 für das Optimierungsproblem (3.100), wobei $\boldsymbol{\sigma}_k$ mit den Lagrange-Multiplikatoren der aktiven Beschränkungen übereinstimmt. Wenn keine inaktiven Ungleichungsbeschränkungen verletzt werden und alle Einträge $\sigma_{k,i},\ i = p+1, \ldots, \bar{p}$ nichtnegativ sind, dann erfüllt der Punkt $\mathbf{x}_k$ die notwendigen

KKT-Bedingungen für ein (lokales) Minimum. Wenn jedoch ein $j \in \{p+1, \ldots, \bar{p}\}$ so existiert, dass $\sigma_{k,j} < 0$ gilt, dann kann die Kostenfunktion weiter verkleinert werden, indem die Ungleichungsbeschränkung $h_j(\mathbf{x}) = \mathbf{a}_{h,j}^{\mathrm{T}}\mathbf{x} - b_{h,j} \leq 0$ inaktiv gesetzt wird. Der Algorithmus der Gradienten-Projektionsmethode ist in Tabelle 3.1 zusammengefasst.

Initialisierung: $\mathbf{x}_0$ (Zulässiger Startpunkt)
$k = 0$ (Startindex)
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: 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 (3.106) in den Tangentialraum $\mathcal{T}\bar{\mathcal{X}}$.

Schritt 3:

if $\mathbf{d}_k \neq \mathbf{0}$

Berechne

$$\begin{aligned} \alpha_{k,1} &= \max\{\alpha_k \,|\, \mathbf{x}_k + \alpha_k \mathbf{d}_k \in \mathcal{X}\} \\ \alpha_{k,2} &= \arg \min_{0<\alpha_k<\alpha_{k,1}} f(\mathbf{x}_k + \alpha_k \mathbf{d}_k) \end{aligned}$$

und setze $\mathbf{x}_{k+1} = \mathbf{x}_k + \alpha_{k,2}\mathbf{d}_k$ und $k \leftarrow k+1$.

else (d. h. $\mathbf{d}_k = \mathbf{0}$)

Berechne $\boldsymbol{\sigma}_k = -\left(\mathbf{A}^{\mathrm{T}}\mathbf{A}\right)^{-1}\mathbf{A}^{\mathrm{T}}(\nabla f)(\mathbf{x}_k)$ (siehe (3.105))

1. Wenn $\sigma_{k,j} \geq 0$ für alle $j = p+1, \ldots, \bar{p}$ gilt, dann erfüllt $\mathbf{x}_k$ die KKT-Bedingungen, setze stop=1.
2. Wenn $\sigma_{k,j} \geq 0$ nicht für alle $j = p+1, \ldots, \bar{p}$ gilt, dann streiche jene Ungleichungsbeschränkung, die zur negativsten Komponente $\sigma_{k,j},\ j = p+1, \ldots, \bar{p}$ gehört, passe die Indexmenge W und die Matrix $\mathbf{A}$ entsprechend an und gehe zu Schritt 2 in der nächsten Iteration.

end

until stop $== 1$

Tabelle 3.1: Gradienten-Projektionsmethode.

3.2.2.2 Nichtlineare Beschränkungen

Es wird nun anstelle des Optimierungsproblems (3.100) mit linearen Beschränkungen das Optimierungsproblem

$$\begin{aligned} \min_{\mathbf{x}\in\mathbb{R}^n} \quad & f(\mathbf{x}) && && \text{(3.111a)} \\ \text{u.B.v.} \quad & g_i(\mathbf{x}) = 0, && i = 1,\dots,p && \text{(3.111b)} \\ & h_i(\mathbf{x}) \le 0, && i = 1,\dots,q && \text{(3.111c)} \end{aligned}$$

mit nichtlinearen Beschränkungen betrachtet. Die Gleichungs- und aktiven Ungleichungsbeschränkungen werden wieder im Vektor

$$\bar{\mathbf{g}}(\mathbf{x}) = \begin{bmatrix} [g_i(\mathbf{x})]_{i=1,\dots,p} \\ [h_i(\mathbf{x})]_{i\in W} \end{bmatrix} \tag{3.112}$$

mit der Dimension $\bar{p} = p + |W|$ zusammengefasst. Diese aktuell aktiven Beschränkungen definieren die Mannigfaltigkeit $\bar{\mathcal{X}} = \{\mathbf{x} \in \mathcal{X} \,|\, \bar{\mathbf{g}}(\mathbf{x}) = \mathbf{0}\}$.

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 (3.106))

$$\mathbf{P}_k = \mathbf{E} - (\nabla\bar{\mathbf{g}})(\mathbf{x}_k)\Big((\nabla\bar{\mathbf{g}})^{\mathrm{T}}(\mathbf{x}_k)(\nabla\bar{\mathbf{g}})(\mathbf{x}_k)\Big)^{-1}(\nabla\bar{\mathbf{g}})^{\mathrm{T}}(\mathbf{x}_k) \tag{3.113}$$

mit $(\nabla\bar{\mathbf{g}})(\mathbf{x}) = \Big[[(\nabla g_i)(\mathbf{x})]_{i=1,\dots,p} \quad [(\nabla h_i)(\mathbf{x})]_{i\in W}\Big]$ in den Tangentialraum $\mathcal{T}_{\mathbf{x}_k}\bar{\mathcal{X}}$ projiziert.

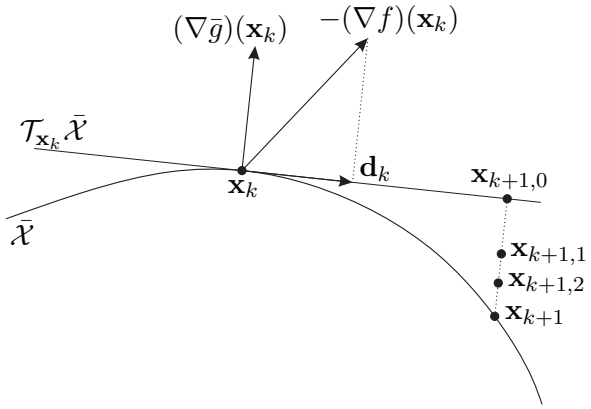


Abbildung 3.5: Gradienten-Projektionsmethode.

Abbildung 3.5 veranschaulicht diese Projektion grafisch. Es ist unmittelbar ersichtlich, dass (im Gegensatz zum Problem mit linearen Beschränkungen) der Punkt

$$\mathbf{x}_{k+1,0} = \mathbf{x}_k + \alpha_k\mathbf{d}_k \quad \text{mit} \quad \mathbf{d}_k = -\mathbf{P}_k(\nabla f)(\mathbf{x}_k) \tag{3.114}$$

auch für hinreichend kleines α_k im Allgemeinen nicht mehr auf der Mannigfaltigkeit $\bar{\mathcal{X}}$ liegt. Es kann deshalb eine weitere Bewegung vom Punkt $\mathbf{x}_{k+1,0}$ orthogonal zu $\mathbf{d}_k$ nötig sein, um wieder auf die Mannigfaltigkeit $\bar{\mathcal{X}}$ und somit in den zulässigen Bereich $\mathcal{X}$ zu gelangen. Die Idee dabei ist, ein $\boldsymbol{\eta}_k \in \mathbb{R}^{\bar{p}}$ so zu berechnen, dass

$$\bar{\mathbf{g}}(\mathbf{x}_{k+1}) = \mathbf{0} \quad \text{mit} \quad \mathbf{x}_{k+1} = \mathbf{x}_{k+1,0} + (\nabla\bar{\mathbf{g}})(\mathbf{x}_k)\boldsymbol{\eta}_k \tag{3.115}$$

gilt. Diese Berechnung kann iterativ erfolgen. Dabei wird $\boldsymbol{\eta}_k$ aus den Teilstücken $\Delta\boldsymbol{\eta}_{k,l}$ mit $l = 0, 1, \ldots$ zusammengesetzt, d. h. $\boldsymbol{\eta}_k = \sum_l \Delta\boldsymbol{\eta}_{k,l}$. Es wird also beginnend am Punkt $\mathbf{x}_{k+1,0}$ mit der Iterationsvorschrift

$$\mathbf{x}_{k+1,l+1} = \mathbf{x}_{k+1,l} + (\nabla\bar{\mathbf{g}})(\mathbf{x}_k)\Delta\boldsymbol{\eta}_{k,l} \tag{3.116}$$

eine Punktfolge $\{\mathbf{x}_{k+1,l}\}$ konstruiert, die zu einem zulässigen Punkt $\mathbf{x}_{k+1} \in \bar{\mathcal{X}}$ konvergiert. Ausgehend von einem Punkt $\mathbf{x}_{k+1,l}$ mit $\bar{\mathbf{g}}(\mathbf{x}_{k+1,l}) \neq \mathbf{0}$ soll daher ein verbesserter Punkt $\mathbf{x}_{k+1,l+1}$ mit $\bar{\mathbf{g}}(\mathbf{x}_{k+1,l+1}) = \mathbf{0}$ gefunden werden. Unter Ausnützung der Orthogonalitätsbeziehung $(\nabla\bar{\mathbf{g}})^{\mathrm{T}}(\mathbf{x}_k)(\mathbf{x}_{k+1,0} - \mathbf{x}_k) = \mathbf{0}$ liefert eine Taylorreihenentwicklung von $\bar{\mathbf{g}}(\mathbf{x})$ am Punkt $\mathbf{x}_k$ bis zum linearen Glied

$$\begin{aligned}
\bar{\mathbf{g}}(\mathbf{x}) &\approx \bar{\mathbf{g}}(\mathbf{x}_k) + (\nabla\bar{\mathbf{g}})^{\mathrm{T}}(\mathbf{x}_k)(\mathbf{x} - \mathbf{x}_k) \\
&= \bar{\mathbf{g}}(\mathbf{x}_k) + (\nabla\bar{\mathbf{g}})^{\mathrm{T}}(\mathbf{x}_k)(\mathbf{x} - \mathbf{x}_{k+1,0} + \mathbf{x}_{k+1,0} - \mathbf{x}_k) \\
&= \bar{\mathbf{g}}(\mathbf{x}_k) + (\nabla\bar{\mathbf{g}})^{\mathrm{T}}(\mathbf{x}_k)(\mathbf{x} - \mathbf{x}_{k+1,0}) \ .
\end{aligned} \tag{3.117}$$

Im Speziellen gilt unter Verwendung von (3.116)

$$\begin{aligned}
\mathbf{0} = \bar{\mathbf{g}}(\mathbf{x}_{k+1,l+1}) &\approx \bar{\mathbf{g}}(\mathbf{x}_k) + (\nabla\bar{\mathbf{g}})^{\mathrm{T}}(\mathbf{x}_k)(\mathbf{x}_{k+1,l+1} - \mathbf{x}_{k+1,0}) \\
&= \bar{\mathbf{g}}(\mathbf{x}_k) + (\nabla\bar{\mathbf{g}})^{\mathrm{T}}(\mathbf{x}_k)(\mathbf{x}_{k+1,l} + (\nabla\bar{\mathbf{g}})(\mathbf{x}_k)\Delta\boldsymbol{\eta}_{k,l} - \mathbf{x}_{k+1,0})
\end{aligned} \tag{3.118a}$$

$$\bar{\mathbf{g}}(\mathbf{x}_{k+1,l}) \approx \bar{\mathbf{g}}(\mathbf{x}_k) + (\nabla\bar{\mathbf{g}})^{\mathrm{T}}(\mathbf{x}_k)(\mathbf{x}_{k+1,l} - \mathbf{x}_{k+1,0}) \ . \tag{3.118b}$$

Die Differenz von (3.118a) und (3.118b) lautet

$$-\bar{\mathbf{g}}(\mathbf{x}_{k+1,l}) = (\nabla\bar{\mathbf{g}})^{\mathrm{T}}(\mathbf{x}_k)(\nabla\bar{\mathbf{g}})(\mathbf{x}_k)\Delta\boldsymbol{\eta}_{k,l} \ . \tag{3.119}$$

Daraus folgt

$$\Delta\boldsymbol{\eta}_{k,l} = -\Big[(\nabla\bar{\mathbf{g}})^{\mathrm{T}}(\mathbf{x}_k)(\nabla\bar{\mathbf{g}})(\mathbf{x}_k)\Big]^{-1}\bar{\mathbf{g}}(\mathbf{x}_{k+1,l}) \tag{3.120}$$

und nach Einsetzen in (3.116)

$$\mathbf{x}_{k+1,l+1} = \mathbf{x}_{k+1,l} - (\nabla\bar{\mathbf{g}})(\mathbf{x}_k)\Big[(\nabla\bar{\mathbf{g}})^{\mathrm{T}}(\mathbf{x}_k)(\nabla\bar{\mathbf{g}})(\mathbf{x}_k)\Big]^{-1}\bar{\mathbf{g}}(\mathbf{x}_{k+1,l}) \ . \tag{3.121}$$

Diese Gleichung beinhaltet ähnliche Ausdrücke wie die Projektionsmatrix $\mathbf{P}_k$ gemäß (3.113). Gleichung (3.121) wird iterativ ausgeführt, um zu einem zulässigen Punkt $\mathbf{x}_{k+1}$ zu kommen.

Es ist zu beachten, dass ein Vektor $\boldsymbol{\eta}_k$ gemäß (3.115) nicht immer existieren muss, siehe beispielsweise Abbildung 3.6. Tritt dieses Problem auf, muss die Schrittweite α_k in (3.114) reduziert werden.

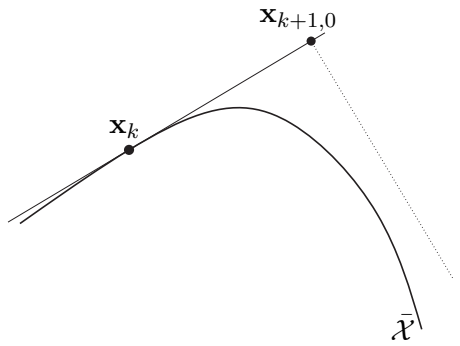


Abbildung 3.6: Nichtexistenz von $\boldsymbol{\eta}_k$, um auf $\bar{\mathcal{X}}$ zu kommen.

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 3.7. Typischerweise werden in diesem Zusammenhang Verfahren zur Interpolation zwischen den Punkten $\mathbf{x}_k$ und $\mathbf{x}_{k+1,0}$ eingesetzt, um zu einem neuen Punkt $\mathbf{x}_{k+1,1}$ zu kommen. Es wird also für die Berechnung von $\mathbf{x}_{k+1,1}$ nicht die Iterationsvorschrift (3.121) verwendet. Ausgehend von $\mathbf{x}_{k+1,1}$ wird wieder mittels (3.121) eine Punktfolge konstruiert um zu einem neuen Punkt $\mathbf{x}_{k+1} \in \bar{\mathcal{X}}$ zu gelangen, der die ursprünglich inaktiven Ungleichungsbeschränkungen nicht verletzt.

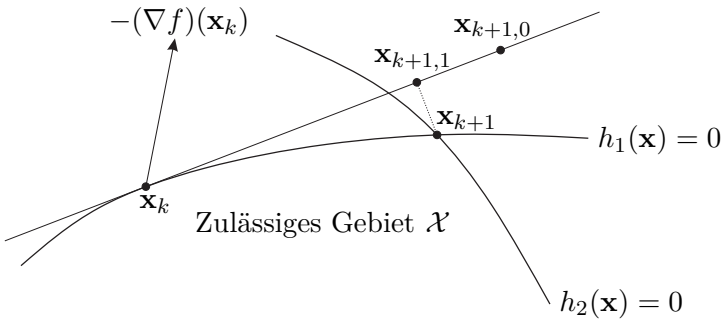


Abbildung 3.7: Interpolation zur Einhaltung der ursprünglich inaktiven Ungleichungsbeschränkungen.

Aufgabe 3.12. Lösen Sie das beschränkte Optimierungsproblem (3.49) aus Beispiel 3.4 numerisch mit Hilfe der Gradienten-Projektionsmethode. Erstellen Sie dafür ein Computerprogramm, wobei Sie für die Bestimmung der Schrittweite α_k wahlweise selbst einen Algorithmus (z. B. aus Abschnitt 2.3.1) programmieren oder vorgefertigte Funktionen einsetzen können. Verwenden Sie als Startpunkt $\mathbf{x}_0 = [1/\sqrt{2} \quad 1/\sqrt{2} \quad 0]^{\mathrm{T}}$. Zur Projektion des Punktes $\mathbf{x}_{k+1,0}$ zurück auf die Mannigfaltigkeit $\bar{\mathcal{X}}$, also zum Finden eines zulässigen Punktes $\mathbf{x}_{k+1}$ können Sie (3.121) iterativ anwenden.

Lösung von Aufgabe 3.12. Am Startpunkt beträgt der Kostenfunktionswert $f(\mathbf{x}_0) = 1.207\,107$. Im Verlauf der ersten Iteration ergeben sich folgende Zwischenergebnisse:

$$\mathbf{P}_0 = \frac{1}{2}\begin{bmatrix} 1 & -1 & 0 \\ -1 & 1 & 0 \\ 0 & 0 & 2 \end{bmatrix}, \qquad \mathbf{d}_0 = \frac{1}{2}\begin{bmatrix} \sqrt{2}-1 \\ 1-\sqrt{2} \\ -\sqrt{2} \end{bmatrix}, \qquad \alpha_0 = 0.246\,289\,,$$

$$\mathbf{x}_{1,0} = \begin{bmatrix} 0.758\,115 \\ 0.656\,099 \\ -0.174\,153 \end{bmatrix}, \qquad \mathbf{x}_1 = \begin{bmatrix} 0.745\,439 \\ 0.643\,422 \\ -0.174\,153 \end{bmatrix}, \qquad f(\mathbf{x}_1) = 1.108\,035$$

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 3.4 ist bekannt, dass die exakte Lösung $\mathbf{x}^* = [1 \quad 0 \quad 0]^{\mathrm{T}}$ mit $f(\mathbf{x}^*) = 1$ lautet.

Probleme, die sich bei der Gradienten-Projektionsmethode im Zusammenhang mit nichtlinearen Beschränkungen ergeben und möglicherweise eine weitere Bewegung vom Tangentialraum $\mathcal{T}_{\mathbf{x}_k}\bar{\mathcal{X}}$ zurück auf die Mannigfaltigkeit $\bar{\mathcal{X}}$ erfordern (siehe Abbildung 3.5), 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* [3.2, 3.5–3.7] 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

$$\min_{\mathbf{x}\in\mathbb{R}^n} \quad f(\mathbf{x}) \tag{3.122a}$$

$$\text{u.B.v.} \quad g_i(\mathbf{x}) = 0, \qquad i = 1,\dots,p \tag{3.122b}$$

$$h_i(\mathbf{x}) \le 0, \qquad i = 1,\dots,q \tag{3.122c}$$

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

$$\mathbf{x} = \begin{bmatrix} \mathbf{x}_I \\ \mathbf{x}_D \end{bmatrix} \tag{3.123}$$

gilt und die Matrix

$$\mathbf{A}_D(\mathbf{x}) = \left(\frac{\partial \bar{\mathbf{g}}(\mathbf{x})}{\partial \mathbf{x}_D} \right) = \begin{bmatrix} \frac{\partial \bar{g}_1}{\partial x_{D,1}} & \cdots & \frac{\partial \bar{g}_1}{\partial x_{D,\bar{p}}} \\ \vdots & & \vdots \\ \frac{\partial \bar{g}_{\bar{p}}}{\partial x_{D,1}} & \cdots & \frac{\partial \bar{g}_{\bar{p}}}{\partial x_{D,\bar{p}}} \end{bmatrix} \tag{3.124}$$

regulär ist. Die in (3.123) gewählte Reihenfolge der Variablen stellt keine Einschränkung dar, da sie durch Umsortieren immer hergestellt werden kann. Im Verlauf der Iterationen kann sich die Partitionierung (3.123) ändern. Gemäß dem Satz über implizite Funktionen (Satz 1.4) 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

$$\bar{\mathbf{g}}(\mathbf{x}) = \mathbf{0} \quad \Leftrightarrow \quad \mathbf{x}_D = \bar{\mathbf{g}}_D^{-1}(\mathbf{x}_I) \tag{3.125}$$

gilt. Damit lässt sich das *reduzierte Optimierungsproblem*

$$\min_{\mathbf{x}_I \in \mathbb{R}^{n-\bar{p}}} \quad f\left(\begin{bmatrix} \mathbf{x}_I \\ \bar{\mathbf{g}}_D^{-1}(\mathbf{x}_I) \end{bmatrix} \right) \tag{3.126}$$

formulieren, welches nur $n - \bar{p}$ Optimierungsvariablen besitzt und in der aktuellen Iteration lokal äquivalent zum ursprünglichen (höherdimensionaleren) Optimierungsproblem (3.122) ist. Die Lösungssuche wird damit automatisch auf die Mannigfaltigkeit $\bar{\mathcal{X}}$ eingeschränkt. Aus der Lösung $\mathbf{x}_I^*$ des Optimierungsproblems (3.126) folgt schließlich noch der optimale Wert $\mathbf{x}_D^* = \bar{\mathbf{g}}_D^{-1}(\mathbf{x}_I^*)$ für die abhängigen Variablen.

In günstigen Fällen ist das reduzierte Optimierungsproblem (3.126) *unbeschränkt*, d. h. beliebige Werte $\mathbf{x}_I \in \mathbb{R}^{n-\bar{p}}$ sind zulässig. In ungünstigen Fällen jedoch besitzt die Funktion $\bar{\mathbf{g}}_D^{-1}(\mathbf{x}_I)$ nur für ein eingeschränktes Definitionsgebiet von $\mathbf{x}_I$ eine reelle Lösung, d. h. das reduzierte Optimierungsproblem (3.126) ist *beschränkt*. Die Beschränkungen für $\mathbf{x}_I$ sind bei der Lösung von (3.126) zu berücksichtigen. Ob das reduzierte Optimierungsproblem (3.126) unbeschränkt oder beschränkt ist, hängt also vom Definitionsgebiet der Funktion $\bar{\mathbf{g}}_D^{-1}(\mathbf{x}_I)$ und damit auch von der Wahl der Partitionierung in (3.123) ab. Es ist daher zu empfehlen, diese Partitionierung so zu wählen, dass die unabhängigen Variablen $\mathbf{x}_I$ beliebige Werte annehmen dürfen (siehe dazu das Beispiel 3.8 am Ende dieses Abschnittes).

> *Bemerkung* 3.1. Gerade bei regelungstechnischen Optimierungsaufgaben ist die in (3.123) 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 3.5). Da die meisten dynamischen Systeme viele Zustandsgrößen aber nur wenige Stellgrößen besitzen, ist damit die Dimension des Optimierungsproblems (3.126) erheblich kleiner als jene von (3.122).

Wenn das reduzierte Problem (3.126) unbeschränkt ist, kann seine Lösung $\mathbf{x}_I^*$ z. B. mit den in Kapitel 2 vorgestellten Methoden berechnet werden. Andernfalls kann $\mathbf{x}_I^*$ z. B. mit den in aktuellen Kapitel vorgestellten Methoden berechnet werden. Generell wird bei der Lösung von (3.126) häufig die (totale) Ableitung der Kostenfunktion (3.126) bezüglich $\mathbf{x}_I$ (Gradient) benötigt. Diese Ableitung kann wie folgt berechnet werden. Zunächst bilde man das totale Differenzial von $\bar{\mathbf{g}}(\mathbf{x}) = \mathbf{0}$

$$\mathrm{d}\bar{\mathbf{g}}(\mathbf{x}) = \underbrace{\frac{\partial \bar{\mathbf{g}}(\mathbf{x})}{\partial \mathbf{x}_I}}_{= \mathbf{A}_I(\mathbf{x})} \mathrm{d}\mathbf{x}_I + \mathbf{A}_D(\mathbf{x})\,\mathrm{d}\mathbf{x}_D = \mathbf{0} \tag{3.127}$$

aus dem

$$\frac{\mathrm{d}\mathbf{x}_D}{\mathrm{d}\mathbf{x}_I} = \frac{\mathrm{d}\bar{\mathbf{g}}_D^{-1}(\mathbf{x}_I)}{\mathrm{d}\mathbf{x}_I} = -\mathbf{A}_D^{-1}\mathbf{A}_I \tag{3.128}$$

folgt. Daraus ergibt sich die gesuchte totale Ableitung

$$\left(\frac{\mathrm{d}f(\mathbf{x})}{\mathrm{d}\mathbf{x}_I}\right)^{\mathrm{T}} = \left(\frac{\partial f(\mathbf{x})}{\partial \mathbf{x}_I} - \frac{\partial f(\mathbf{x})}{\partial \mathbf{x}_D}\mathbf{A}_D^{-1}\mathbf{A}_I\right)^{\mathrm{T}} = \underbrace{\left[\begin{array}{c|c} \mathbf{E} & -(\mathbf{A}_D^{-1}\mathbf{A}_I)^{\mathrm{T}} \end{array}\right]}_{= \bar{\mathbf{P}}(\mathbf{x})} (\nabla f)(\mathbf{x}) \;, \tag{3.129}$$

welche auch als *reduzierter Gradient* bezeichnet wird. In analoger Weise folgt die (totale) zweite Ableitung der Kostenfunktion (3.126) bezüglich $\mathbf{x}_I$ (Hessematrix) in der Form

$$\frac{\mathrm{d}^2 f(\mathbf{x})}{\mathrm{d}\mathbf{x}_I^{\mathrm{T}}\,\mathrm{d}\mathbf{x}_I} = \begin{bmatrix} \frac{\mathrm{d}^2 f}{\mathrm{d}x_{I,1}^2} & \cdots & \frac{\mathrm{d}^2 f}{\mathrm{d}x_{I,1}\,\mathrm{d}x_{I,n-\bar{p}}} \\ \vdots & & \vdots \\ \frac{\mathrm{d}^2 f}{\mathrm{d}x_{I,n-\bar{p}}\,\mathrm{d}x_{I,1}} & \cdots & \frac{\mathrm{d}^2 f}{\mathrm{d}x_{I,n-\bar{p}}^2} \end{bmatrix} = \bar{\mathbf{P}}(\mathbf{x})\Big(\nabla^2 f\Big)(\mathbf{x})\bar{\mathbf{P}}^{\mathrm{T}}(\mathbf{x}) \;, \tag{3.130}$$

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. Im Fall $n - \bar{p} < \bar{p}$ kann es den Aufwand zur Berechnung von $\bar{\mathbf{P}}(\mathbf{x})$ gemäß (3.129) reduzieren, wenn statt dem Ausdruck $-\mathbf{A}_D^{-1}\mathbf{A}_I$, welcher eine Inversion der Matrix $\mathbf{A}_D$ erfordert, die Matrix $\mathrm{d}\mathbf{x}_D/\,\mathrm{d}\mathbf{x}_I$, die sich aus der Lösung des linearen Gleichungssystems

$$\mathbf{A}_D \frac{\mathrm{d}\mathbf{x}_D}{\mathrm{d}\mathbf{x}_I} = -\mathbf{A}_I \tag{3.131}$$

(vgl. (3.128)) ergibt, in (3.129) eingesetzt wird.

Es ist zu beachten, dass mit dem reduzierten Optimierungsproblem (3.126) noch nicht sichergestellt ist, dass die in der aktuellen Iteration gefundene Lösung $\mathbf{x}_I^*$, $\mathbf{x}_D^*$ die KKT-Bedingungen gemäß Satz 3.10 für das ursprüngliche beschränkte Optimierungsproblem

(3.122) erfüllt. Zudem ist es häufig nicht möglich, einen analytischen Ausdruck für die Funktion $\bar{\mathbf{g}}_D^{-1}(\mathbf{x}_I)$, welche auch in (3.126) auftritt, zu finden. Praktisch wird daher meist nicht das reduzierte Optimierungsproblem (3.126) selbst gelöst, sondern nur die zugehörige Ableitung (3.129) (gegebenenfalls auch (3.130)) berechnet, um im Zuge der iterativen Lösungssuche entlang der aktuellen Mannigfaltigkeit $\bar{\mathcal{X}}$ und unter Berücksichtigung allfälliger Beschränkungen für die unabhängigen Variablen $\mathbf{x}_I$ eine neue Suchrichtung $\mathbf{s}_{I,k}$ für $\mathbf{x}_I$ zu bestimmen. Im Falle eines unbeschränkten reduzierten Optimierungsproblems (3.126) gilt z. B. bei Verwendung der Gradientenmethode $\mathbf{s}_{I,k} = -(\mathrm{d}f(\mathbf{x}_k)/\,\mathrm{d}\mathbf{x}_{I,k})^{\mathrm{T}}$. Es können im unbeschränkten Fall aber auch die weiteren in Abschnitt 2.3.2 besprochenen Methoden zu Wahl einer Suchrichtung $\mathbf{s}_{I,k}$ verwendet werden.

Tabelle 3.2 fasst den zugehörigen Algorithmus im Falle eines unbeschränkten reduzierten Optimierungsproblem (3.126) zusammen. Sind jedoch Beschränkungen für $\mathbf{x}_I$ vorhanden, muss Schritt 4 des Algorithmus entsprechend angepasst werden. Im Zuge der reduzierten Gradientenmethode muss die Funktion $\bar{\mathbf{g}}_D^{-1}(\mathbf{x}_I)$ weder explizit 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 iterativen Lösungssuche von Vorteil sein kann.

Wie nachfolgendes Lemma zeigt, kann der reduzierte Gradient $(\mathrm{d}f(\mathbf{x})/\,\mathrm{d}\mathbf{x}_I)^{\mathrm{T}}$ alternativ zu (3.129) 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 (3.122). Für eine gegebene Partitionierung (3.123) kann der reduzierte Gradient $(\mathrm{d}f(\mathbf{x})/\,\mathrm{d}\mathbf{x}_I)^{\mathrm{T}}$ mit Hilfe der Lagrangefunktion*

$$L(\mathbf{x}, \bar{\boldsymbol{\lambda}}) = f(\mathbf{x}) + \bar{\boldsymbol{\lambda}}^{\mathrm{T}} \bar{\mathbf{g}}(\mathbf{x}) \tag{3.132}$$

in der Form

$$\left(\frac{\mathrm{d}f(\mathbf{x})}{\mathrm{d}\mathbf{x}_I}\right)^{\mathrm{T}} = \left(\frac{\partial}{\partial \mathbf{x}_I} L\right)^{\mathrm{T}} (\mathbf{x}, \bar{\boldsymbol{\lambda}}) = \left(\frac{\partial f(\mathbf{x})}{\partial \mathbf{x}_I}\right)^{\mathrm{T}} + \left(\frac{\partial \bar{\mathbf{g}}(\mathbf{x})}{\partial \mathbf{x}_I}\right)^{\mathrm{T}} \bar{\boldsymbol{\lambda}} \tag{3.133}$$

berechnet werden, wobei $\mathbf{x}_D$ und $\bar{\boldsymbol{\lambda}}$ aus den Bedingungen

$$\left(\frac{\partial}{\partial \bar{\boldsymbol{\lambda}}} L\right)^{\mathrm{T}} (\mathbf{x}, \bar{\boldsymbol{\lambda}}) = \bar{\mathbf{g}}(\mathbf{x}) = \mathbf{0} \tag{3.134a}$$

$$\left(\frac{\partial}{\partial \mathbf{x}_D} L\right)^{\mathrm{T}} (\mathbf{x}, \bar{\boldsymbol{\lambda}}) = \left(\frac{\partial f(\mathbf{x})}{\partial \mathbf{x}_D}\right)^{\mathrm{T}} + \left(\frac{\partial \bar{\mathbf{g}}(\mathbf{x})}{\partial \mathbf{x}_D}\right)^{\mathrm{T}} \bar{\boldsymbol{\lambda}} = \mathbf{0} \tag{3.134b}$$

folgen.

Aufgabe 3.13. Beweisen Sie Lemma 3.2.

Gegenüber (3.129) 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 (3.134b) gelöst werden. Aus Lemma 3.2 folgt direkt, dass

Initialisierung: $\mathbf{x}_0$ (Zulässiger Startpunkt)
$k = 0$ (Startindex)
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 (3.123) in unabhängige Variablen $\mathbf{x}_I$ und abhängige Variablen $\mathbf{x}_D$ so, dass $\mathbf{A}_D(\mathbf{x}_k)$ gemäß (3.124) regulär ist.

Schritt 3: Berechne am Punkt $\mathbf{x}_k$ den reduzierten Gradienten $\mathbf{d}_{I,k} = (\mathrm{d}f(\mathbf{x}_k)/\,\mathrm{d}\mathbf{x}_I)^{\mathrm{T}}$ gemäß (3.129) und gegebenenfalls die reduzierte Hessematrix gemäß (3.130).

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

$$\alpha_{k,1} = \max \alpha_k \quad \text{u.B.v.} \quad \begin{bmatrix} \mathbf{x}_{I,k} + \alpha_k \mathbf{s}_{I,k} \\ \bar{\mathbf{g}}_D^{-1}(\mathbf{x}_{I,k} + \alpha_k \mathbf{s}_{I,k}) \end{bmatrix} \in \mathcal{X}$$

$$\alpha_{k,2} = \arg \min_{0<\alpha_k<\alpha_{k,1}} f\left(\begin{bmatrix} \mathbf{x}_{I,k} + \alpha_k \mathbf{s}_{I,k} \\ \bar{\mathbf{g}}_D^{-1}(\mathbf{x}_{I,k} + \alpha_k \mathbf{s}_{I,k}) \end{bmatrix} \right),$$

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 (3.66) erfüllt sind.

1. Wenn Punkt $\mathbf{x}_k$ die KKT-Bedingungen erfüllt, setze 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 stop $== 1$

Tabelle 3.2: Reduzierte Gradientenmethode für ein unbeschränktes reduziertes Optimierungsproblem (3.126).

die Stationaritätsbedingung $(\mathrm{d}f(\mathbf{x})/\,\mathrm{d}\mathbf{x}_I)^{\mathrm{T}} = \mathbf{0}$ genau auf die notwendige Optimalitätsbedingung erster Ordnung für Optimierungsprobleme mit reinen Gleichungsbeschränkungen (siehe Satz 3.6 sowie (3.31)) führt.

In [3.5–3.7] 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.

Beispiel 3.8. Man betrachte das Optimierungsproblem

$$\min_{\mathbf{x}\in\mathbb{R}^2} \quad f(\mathbf{x}) = 1 - x_1^2 - x_2^2 \tag{3.135a}$$

$$\text{u.B.v.} \quad g(\mathbf{x}) = x_1^2 + x_2 - 1 = 0\ , \tag{3.135b}$$

welches nun in ein reduziertes Optimierungsproblem der Form (3.126) umgeschrieben werden soll.

Wird x_1 als unabhängige Variable gewählt, d. h. $x_I = x_1$ und $x_D = x_2$, so ergibt sich das *unbeschränkte* reduzierte Optimierungsproblem

$$\min_{x_1\in\mathbb{R}} \quad x_1^2 - x_1^4\ . \tag{3.136}$$

Es besitzt nur an der Stelle $x_1^* = 0$ ein lokales Minimum. Für den zugehörigen optimalen Wert der abhängigen Variable x_2 gilt $x_2^* = 1$. An diesem Punkt gilt für den reduzierten Gradienten gemäß (3.129) $\left(\frac{\mathrm{d}f(\mathbf{x})}{\mathrm{d}\mathbf{x}_I}\right)^{\mathrm{T}} = 0$. Mithilfe der hinreichenden KKT-Bedingungen gemäß den Satz 3.12 lässt sich leicht zeigen, dass der gefundene Punkt $\mathbf{x}^* = [0 \quad 1]^{\mathrm{T}}$ ein lokales Minimum des ursprünglichen beschränkten Optimierungsproblems (3.135) darstellt. Es ist dies das einzige lokale Minimum. Das Problem besitzt kein globales Minimum.

> *Aufgabe* 3.14. Zeigen Sie dass der gefundene Punkt $\mathbf{x}^* = [0 \quad 1]^{\mathrm{T}}$ das einzige lokale Minimum des ursprünglichen beschränkten Optimierungsproblems (3.135) darstellt. Begründen Sie warum das Problem kein globales Minimum besitzt.

Wird alternativ x_2 als unabhängige Variable gewählt, d. h. $x_I = x_2$ und $x_D = x_1$, so ergibt sich das *beschränkte* reduzierte Optimierungsproblem

$$\min_{x_2\in\mathbb{R}} \quad x_2 - x_2^2 \tag{3.137a}$$

$$\text{u.B.v.} \quad x_2 \leq 1\ . \tag{3.137b}$$

Die Beschränkung (3.137b) wird benötigt, damit $g(\mathbf{x})$ nach x_1 aufgelöst werden kann. Das reduzierte Optimierungsproblem (3.137) besitzt nur an der Stelle $x_2^* = 1$ ein lokales Minimum. Für den zugehörigen optimalen Wert der abhängigen Variable x_1 gilt $x_1^* = 0$. An diesem Punkt kann der reduzierte Gradient nicht berechnet werden, da

$A_D = 0$ gilt, d. h. A_D ist nicht regulär. Auch in diesem Fall wird also der Punkt $\mathbf{x}^* = [0 \quad 1]^\mathrm{T}$ als lokales Minimum des ursprünglichen beschränkten Optimierungsproblems (3.135) gefunden. Dieser Fall zeigt aber, dass die Berücksichtigung der Beschränkung (3.137b) notwendig ist, um das Optimum zu finden.

Die Wahl $x_I = x_1$ und $x_D = x_2$ erwies sich hier als günstig, weil sie zu einem unbeschränkten reduzierten Optimierungsproblem führt. Die alternative Wahl $x_I = x_2$ und $x_D = x_1$ ist ungünstig, da sie zu einem beschränkten reduzierten Optimierungsproblem führt.

Aufgabe 3.15. Lösen Sie das beschränkte Optimierungsproblem (3.49) aus Beispiel 3.4 numerisch mit Hilfe der reduzierten Gradientenmethode. Erstellen Sie dafür ein Computerprogramm, wobei Sie für die Bestimmung der Schrittweite α_k wahlweise selbst einen Algorithmus (z. B. aus Abschnitt 2.3.1) programmieren oder vorgefertigte Funktionen einsetzen können. Verwenden Sie als Startpunkt $\mathbf{x}_0 = [1/\sqrt{2} \quad 1/\sqrt{2} \quad 0]^\mathrm{T}$. Wählen Sie die Partitionierung $\mathbf{x}_I = [x_2 \quad x_3]^\mathrm{T}$ und $x_D = x_1$. Warum ist diese Aufteilung in Anbetracht des aus Beispiel 3.4 bekannten optimalen Punktes $\mathbf{x}^* = [1 \quad 0 \quad 0]^\mathrm{T}$ sinnvoll? Vergleichen Sie die Ergebnisse mit jenen von Aufgabe 3.12.

Lösung von Aufgabe 3.15. Die Partitionierung $\mathbf{x}_I = [x_2 \quad x_3]^\mathrm{T}$ 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 $f(\mathbf{x}_0) = 1.207\,107$. Im Verlauf der ersten Iteration ergeben sich bei Verwendung der Suchrichtung $\mathbf{s}_{I,k} = -(\mathrm{d}f(\mathbf{x}_k)/\,\mathrm{d}\mathbf{x}_{I,k})^\mathrm{T}$ (einfache Gradientenmethode) folgende Zwischenergebnisse:

$$\mathbf{A}_I(\mathbf{x}_0) = \begin{bmatrix} \sqrt{2} & 0 \end{bmatrix}, \qquad A_D(\mathbf{x}_0) = \sqrt{2}, \qquad \mathbf{d}_0 = \begin{bmatrix} 1-\sqrt{2} \\ -1/\sqrt{2} \end{bmatrix},$$

$$\alpha_0 = 0.343\,906, \qquad \mathbf{x}_1 = \begin{bmatrix} 0.788\,687 \\ 0.564\,656 \\ -0.243\,178 \end{bmatrix}, \qquad f(\mathbf{x}_1) = 1.088\,483$$

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 3.12 zeigt, dass für dieses Optimierungsproblem und den hier verwendeten Startpunkt die reduzierte Gradientenmethode deutlich schneller konvergiert als die Gradienten-Projektionsmethode.

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

$$\min_{\mathbf{x}\in\mathbb{R}^n} \quad f(\mathbf{x}) \tag{3.138a}$$

$$\text{u.B.v.} \quad \mathbf{g}(\mathbf{x}) = \mathbf{0} \tag{3.138b}$$

mit $f \in C^2$ und $p < n$ Gleichungsbeschränkungen $g_1(\mathbf{x}), \dots, g_p(\mathbf{x}) \in C^2$. Nach Satz 3.6 lauten die notwendigen Optimalitätsbedingungen (KKT-Bedingungen) erster Ordnung für einen optimalen Punkt $(\mathbf{x}^*, \boldsymbol{\lambda}^*)$ mit der Lagrangefunktion $L(\mathbf{x}, \boldsymbol{\lambda}) = f(\mathbf{x}) + \boldsymbol{\lambda}^{\mathrm{T}}\mathbf{g}(\mathbf{x})$

$$\begin{bmatrix} \left(\frac{\partial}{\partial \mathbf{x}} L\right)^{\mathrm{T}}(\mathbf{x}^*, \boldsymbol{\lambda}^*) \\ \left(\frac{\partial}{\partial \boldsymbol{\lambda}} L\right)^{\mathrm{T}}(\mathbf{x}^*, \boldsymbol{\lambda}^*) \end{bmatrix} = \begin{bmatrix} (\nabla f)(\mathbf{x}^*) + (\nabla \mathbf{g})(\mathbf{x}^*)\boldsymbol{\lambda}^* \\ \mathbf{g}(\mathbf{x}^*) \end{bmatrix} = \mathbf{0} \ , \tag{3.139}$$

wobei gilt $(\nabla \mathbf{g})(\mathbf{x}^*) = \begin{bmatrix} (\nabla g_1)(\mathbf{x}^*) & \dots & (\nabla g_p)(\mathbf{x}^*) \end{bmatrix}$. Eine Möglichkeit, die $n + p$ Gleichungen (3.139) in den $n + p$ Unbekannten $(\mathbf{x}^*, \boldsymbol{\lambda}^*)$ rekursiv numerisch zu lösen, ist das Newton-Raphson Verfahren mit der Iterationsvorschrift

$$\begin{bmatrix} \mathbf{x}_{k+1} \\ \boldsymbol{\lambda}_{k+1} \end{bmatrix} = \begin{bmatrix} \mathbf{x}_k \\ \boldsymbol{\lambda}_k \end{bmatrix} + \begin{bmatrix} \mathbf{p}_{\mathbf{x},k} \\ \mathbf{p}_{\boldsymbol{\lambda},k} \end{bmatrix} \tag{3.140a}$$

$$\underbrace{\begin{bmatrix} \mathbf{L}(\mathbf{x}_k, \boldsymbol{\lambda}_k) & (\nabla \mathbf{g})(\mathbf{x}_k) \\ (\nabla \mathbf{g})^{\mathrm{T}}(\mathbf{x}_k) & \mathbf{0} \end{bmatrix}}_{\mathbf{M}_k} \begin{bmatrix} \mathbf{p}_{\mathbf{x},k} \\ \mathbf{p}_{\boldsymbol{\lambda},k} \end{bmatrix} = -\begin{bmatrix} (\nabla f)(\mathbf{x}_k) + (\nabla \mathbf{g})(\mathbf{x}_k)\boldsymbol{\lambda}_k \\ \mathbf{g}(\mathbf{x}_k) \end{bmatrix} \tag{3.140b}$$

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 (3.140b) 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 $(\nabla \mathbf{g})^{\mathrm{T}}\mathbf{d} = \mathbf{0}$ die Bedingung $\mathbf{d}^{\mathrm{T}}\mathbf{L}(\mathbf{x}_k, \boldsymbol{\lambda}_k)\mathbf{d} > 0$ 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 3.8 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 (3.140b) 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 (3.139) des ursprünglichen Optimierungsproblems (3.138) erfüllt.

Ein Schritt k der Iterationsvorschrift (3.140) kann nun äquivalent auch als *Lösung des quadratischen Programms*

$$\min_{\tilde{\mathbf{p}}\in\mathbb{R}^n} \quad f(\mathbf{x}_k) + (\nabla f)^{\mathrm{T}}(\mathbf{x}_k)\tilde{\mathbf{p}} + \frac{1}{2}\tilde{\mathbf{p}}^{\mathrm{T}}\mathbf{L}(\mathbf{x}_k, \boldsymbol{\lambda}_k)\tilde{\mathbf{p}} \tag{3.141a}$$

$$\text{u.B.v.} \quad (\nabla \mathbf{g})^{\mathrm{T}}(\mathbf{x}_k)\tilde{\mathbf{p}} + \mathbf{g}(\mathbf{x}_k) = \mathbf{0} \tag{3.141b}$$

aufgefasst werden. Die Kostenfunktion (3.141a) folgt direkt aus der nach dem quadratischen Glied abgebrochenen Taylorreihenentwicklung $L(\mathbf{x}_k, \boldsymbol{\lambda}_k) + \left(\frac{\partial}{\partial \mathbf{x}} L\right)(\mathbf{x}_k, \boldsymbol{\lambda}_k)\tilde{\mathbf{p}} + \frac{1}{2}\tilde{\mathbf{p}}^{\mathrm{T}}\mathbf{L}(\mathbf{x}_k, \boldsymbol{\lambda}_k)\tilde{\mathbf{p}}$ unter Berücksichtigung von (3.141b). Die KKT-Bedingungen für (3.141) lauten

$$\begin{bmatrix} \mathbf{L}(\mathbf{x}_k, \boldsymbol{\lambda}_k) & (\nabla \mathbf{g})(\mathbf{x}_k) \\ (\nabla \mathbf{g})^{\mathrm{T}}(\mathbf{x}_k) & \mathbf{0} \end{bmatrix} \begin{bmatrix} \tilde{\mathbf{p}}^* \\ \tilde{\boldsymbol{\lambda}}^* \end{bmatrix} = -\begin{bmatrix} (\nabla f)(\mathbf{x}_k) \\ \mathbf{g}(\mathbf{x}_k) \end{bmatrix} \tag{3.142}$$

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 (3.140b) eingesetzt, woraus sich

$$\begin{bmatrix} \mathbf{L}(\mathbf{x}_k, \boldsymbol{\lambda}_k) & (\nabla \mathbf{g})(\mathbf{x}_k) \\ (\nabla \mathbf{g})^{\mathrm{T}}(\mathbf{x}_k) & \mathbf{0} \end{bmatrix} \begin{bmatrix} \mathbf{p}_{\mathbf{x},k} \\ \boldsymbol{\lambda}_{k+1} \end{bmatrix} = -\begin{bmatrix} (\nabla f)(\mathbf{x}_k) \\ \mathbf{g}(\mathbf{x}_k) \end{bmatrix} \tag{3.143}$$

ergibt. Ein Vergleich von (3.142) mit (3.143) bestätigt nun, dass statt der Lösung des Gleichungssystems (3.143) auch das Minimum des quadratischen Programms (3.141) berechnet werden kann. Die eigentliche Iterationsvorschrift, welche äquivalent zu (3.140) 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 (3.141) die Lösung $\tilde{\mathbf{p}}^* = \mathbf{0}$ gefunden wird, so ist aus (3.142) ersichtlich, dass damit auch ein Punkt $\mathbf{x}^* = \mathbf{x}_k$, $\boldsymbol{\lambda}^* = \tilde{\boldsymbol{\lambda}}^*$ gefunden wurde, der die KKT-Bedingungen (3.139) des ursprünglichen Optimierungsproblems (3.138) 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

$$\min_{\mathbf{x} \in \mathbb{R}^n} \quad f(\mathbf{x}) \tag{3.144a}$$

$$\text{u.B.v.} \quad \mathbf{g}(\mathbf{x}) = \mathbf{0} \tag{3.144b}$$

$$\mathbf{h}(\mathbf{x}) \leq \mathbf{0} \tag{3.144c}$$

mit $f \in C^2$, $p < n$ Gleichungsbeschränkungen $g_1(\mathbf{x}), \ldots, g_p(\mathbf{x}) \in C^2$ und q Ungleichungsbeschränkungen $h_1(\mathbf{x}), \ldots, h_q(\mathbf{x}) \in C^2$. Die zugehörige Lagrangefunktion lautet $L(\mathbf{x}, \boldsymbol{\lambda}, \boldsymbol{\mu}) = f(\mathbf{x}) + \boldsymbol{\lambda}^{\mathrm{T}}\mathbf{g}(\mathbf{x}) + \boldsymbol{\mu}^{\mathrm{T}}\mathbf{h}(\mathbf{x})$. Das Optimierungsproblem (3.144) wird in jedem Iterationsschritt durch das *quadratische Programm*

$$\min_{\tilde{\mathbf{p}} \in \mathbb{R}^n} \quad f(\mathbf{x}_k) + (\nabla f)^{\mathrm{T}}(\mathbf{x}_k)\tilde{\mathbf{p}} + \frac{1}{2}\tilde{\mathbf{p}}^{\mathrm{T}}\mathbf{L}(\mathbf{x}_k, \boldsymbol{\lambda}_k, \boldsymbol{\mu}_k)\tilde{\mathbf{p}} \tag{3.145a}$$

$$\text{u.B.v.} \quad (\nabla \mathbf{g})^{\mathrm{T}}(\mathbf{x}_k)\tilde{\mathbf{p}} + \mathbf{g}(\mathbf{x}_k) = \mathbf{0} \tag{3.145b}$$

$$(\nabla \mathbf{h})^{\mathrm{T}}(\mathbf{x}_k)\tilde{\mathbf{p}} + \mathbf{h}(\mathbf{x}_k) \leq \mathbf{0} \tag{3.145c}$$

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 (3.145) lauten (siehe Satz 3.10)

$$(\nabla f)(\mathbf{x}_k) + \mathbf{L}(\mathbf{x}_k, \boldsymbol{\lambda}_k, \boldsymbol{\mu}_k)\tilde{\mathbf{p}}^* + (\nabla \mathbf{g})(\mathbf{x}_k)\tilde{\boldsymbol{\lambda}}^* + (\nabla \mathbf{h})(\mathbf{x}_k)\tilde{\boldsymbol{\mu}}^* = \mathbf{0} \tag{3.146a}$$

$$\tilde{\boldsymbol{\mu}}^* \geq \mathbf{0} \tag{3.146b}$$

$$\Big((\nabla \mathbf{h})^{\mathrm{T}}(\mathbf{x}_k)\tilde{\mathbf{p}}^* + \mathbf{h}(\mathbf{x}_k)\Big)^{\mathrm{T}} \tilde{\boldsymbol{\mu}}^* = 0 \tag{3.146c}$$

$$(\nabla \mathbf{g})^{\mathrm{T}}(\mathbf{x}_k)\tilde{\mathbf{p}}^* + \mathbf{g}(\mathbf{x}_k) = \mathbf{0} \tag{3.146d}$$

$$(\nabla \mathbf{h})^{\mathrm{T}}(\mathbf{x}_k)\tilde{\mathbf{p}}^* + \mathbf{h}(\mathbf{x}_k) \leq \mathbf{0} \tag{3.146e}$$

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 (3.145) die Lösung $\tilde{\mathbf{p}}^* = \mathbf{0}$, so folgt aus (3.146), 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 (3.144) erfüllt.

Unter bestimmten Voraussetzungen kann eine quadratische Konvergenzordnung (vergleiche Satz 2.12) 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.

Initialisierung:	$\mathbf{x}_0$	(Zulässiger Startpunkt)
	$\boldsymbol{\lambda}_0$, $\boldsymbol{\mu}_0$	(Startwerte der Lagrange-Multiplikatoren)
	$k = 0$	(Startindex)
	ε	(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 (3.145).		
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 (3.145) für einen Iterationsschritt k kann beispielsweise über die Methode der aktiven Beschränkungen (siehe Abschnitt 3.2.1 sowie Beispiel 3.7) gelöst werden. Für die Formulierung des quadratischen Programms (3.145) wird in jedem Iterationsschritt die Hessematrix $\mathbf{L}(\mathbf{x}_k, \boldsymbol{\lambda}_k, \boldsymbol{\mu}_k)$ 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 es sein, dass die Bedingung $\mathbf{d}^{\mathrm{T}}\mathbf{L}(\mathbf{x}_k, \boldsymbol{\lambda}_k, \boldsymbol{\mu}_k)\mathbf{d} > 0$ für alle $\mathbf{d} \neq \mathbf{0}$ aus dem durch (3.145b) und (3.145c) definierten Tangentialraum nicht erfüllt ist, was insbesondere dann möglich ist, wenn das Verfahren nicht in einer hinreichend kleinen Umgebung um $(\mathbf{x}^*, \boldsymbol{\lambda}^*, \boldsymbol{\mu}^*)$ gestartet wird. Sollte die Hessematrix indefinit sein, so erschwert dies 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 gelegentlich durch eine geeignete *positiv definite Appro-*

ximation $\mathbf{H}_k$. 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. [3.8])

$$\mathbf{H}_{k+1} = \mathbf{H}_k + \frac{\mathbf{q}_k \mathbf{q}_k^{\mathrm{T}}}{\mathbf{q}_k^{\mathrm{T}} \mathbf{d}_k} - \frac{\mathbf{H}_k \mathbf{d}_k \mathbf{d}_k^{\mathrm{T}} \mathbf{H}_k}{\mathbf{d}_k^{\mathrm{T}} \mathbf{H}_k \mathbf{d}_k} \tag{3.147a}$$

mit

$$\mathbf{d}_k = \mathbf{x}_{k+1} - \mathbf{x}_k \tag{3.147b}$$

$$\mathbf{y}_k = \left(\frac{\partial}{\partial \mathbf{x}} L\right)^{\mathrm{T}} (\mathbf{x}_{k+1}, \boldsymbol{\lambda}_k, \boldsymbol{\mu}_k) - \left(\frac{\partial}{\partial \mathbf{x}} L\right)^{\mathrm{T}} (\mathbf{x}_k, \boldsymbol{\lambda}_k, \boldsymbol{\mu}_k) \tag{3.147c}$$

$$\theta_k = \begin{cases} 1, & \text{wenn } \mathbf{d}_k^{\mathrm{T}} \mathbf{y}_k \geq 0.2 \mathbf{d}_k^{\mathrm{T}} \mathbf{H}_k \mathbf{d}_k \\ \frac{0.8 \mathbf{d}_k^{\mathrm{T}} \mathbf{H}_k \mathbf{d}_k}{\mathbf{d}_k^{\mathrm{T}} \mathbf{H}_k \mathbf{d}_k - \mathbf{d}_k^{\mathrm{T}} \mathbf{y}_k} & \text{sonst} \end{cases} \tag{3.147d}$$

$$\mathbf{q}_k = \theta_k \mathbf{y}_k + (1 - \theta_k) \mathbf{H}_k \mathbf{d}_k \tag{3.147e}$$

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 (3.147) 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 (3.147) 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

Die im Zuge der iterativen Lösungssuche des SQP-Verfahrens auftretenden Punkte $\mathbf{x}_k$ erfüllen im Allgemeinen nicht strikt die Bedingung $\mathbf{x}_k \in \mathcal{X}$, d. h. sie können Beschränkungen verletzen. Ferner konvergiert der SQP-Algorithmus gemäß Tabelle 3.3 im Allgemeinen 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 Dämpfung bei den Newton-Methoden (siehe Abschnitte 2.3.2.3 bis 2.3.2.5) wird dies durch die Einführung einer Schrittweite $\alpha_k > 0$ erzielt. In Schritt 3 des Algorithmus berechnet man $\mathbf{x}_{k+1}$ daher in der Form

$$\mathbf{x}_{k+1} = \mathbf{x}_k + \alpha_k \tilde{\mathbf{p}}^*. \tag{3.148}$$

Die Schrittweite α_k folgt aus einem geeignet formulierten Liniensuchproblem. Die Kostenfunktion dieses Liniensuchproblems muss eine Bewertung ermöglichen, ob $\mathbf{x}_{k+1}$ *besser* ist als $\mathbf{x}_k$. Eine solche Bewertung kann für unbeschränkte Optimierungsprobleme direkt anhand der Kostenfunktionswerte an den Punkten $\mathbf{x}_k$ und $\mathbf{x}_{k+1}$ erfolgen. Dies gilt im Allgemeinen aber nicht für beschränkte Optimierungsprobleme. Ein Punkt $\mathbf{x}_{k+1}$ ist klarerweise besser als $\mathbf{x}_k$, wenn er sowohl die Kostenfunktion als auch die Verletzung von Beschränkungen reduziert. In vielen Fällen aber sind die Reduktion der Kostenfunktion

und die genauere Einhaltung der Beschränkungen gegensätzliche Ziele, so dass $\mathbf{x}_{k+1}$ gegenüber $\mathbf{x}_k$ zwar den Kostenfunktionswert verbessert jedoch die Beschränkungen stärker verletzt oder umgekehrt (siehe [3.2]). Um in dieser Hinsicht bei der Wahl von α_k einen guten Kompromiss zu finden, kann eine so genannte *Bewertungsfunktion* (englisch: *merit function*) verwendet werden. Eine gängige Wahl dafür ist die Funktion

$$l(\mathbf{x}, \eta) = f(\mathbf{x}) + \eta \left(\sum_{i=1}^{p} |g_i(\mathbf{x})| + \sum_{i=1}^{q} \max\{0, h_i(\mathbf{x})\} \right) . \tag{3.149}$$

Mit der Wahl des Faktors $\eta > 0$ wird die Gewichtung der Verletzung von Beschränkungen gegenüber der Kostenfunktion eingestellt. Die optimale Schrittweite α_k folgt nun aus dem Liniensuchproblem

$$\alpha_k = \arg\min_{\alpha} \; l(\mathbf{x}_k + \alpha \tilde{\mathbf{p}}^*, \eta) \; . \tag{3.150}$$

In der Praxis wird α_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 Verfahren zur Schrittweitenwahl aus Abschnitt 2.3.1 erfolgen. Häufig ist eine Anpassung von η in jedem Iterationsschritt des SQP-Verfahrens erforderlich (vgl. [3.6]). Nur in seltenen Fällen besitzen Bewertungsfunktionen die wünschenswerte Eigenschaft, dass ein lokales Minimum $\mathbf{x}^*$ von (3.144) 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*

$$\min_{\mathbf{x} \in \mathcal{X}} f(\mathbf{x}) \tag{3.151}$$

mit der zulässigen Menge $\mathcal{X}$ in ein *unbeschränktes Optimierungsproblem* der Form

$$\min_{\mathbf{x} \in \mathbb{R}^n} f(\mathbf{x}) + cP(\mathbf{x}) \tag{3.152}$$

mit einem positiven Gewichtungsparameter c und der stetigen Funktion $P(\mathbf{x})$ überzuführen. Die Funktion $P(\mathbf{x})$ wird als *Straffunktion* bezeichnet und besitzt die Eigenschaft, dass $P(\mathbf{x}) > 0$ für alle $\mathbf{x} \in \mathbb{R}^n \backslash \mathcal{X}$ und $P(\mathbf{x}) = 0$ für alle $\mathbf{x} \in \mathcal{X}$. Die zulässige Menge $\mathcal{X}$ ist typischerweise implizit über Gleichungs- und Ungleichungsbeschränkungen definiert. Für $\mathcal{X} = \{\mathbf{x} \in \mathbb{R}^n \,|\, g_i(\mathbf{x}) = 0, \; i = 1, \ldots, p, \; h_i(\mathbf{x}) \leq 0, \; i = 1, \ldots, q\}$ kann daher als Straffunktion

$$P(\mathbf{x}) = \frac{1}{2} \sum_{i=1}^{p} (g_i(\mathbf{x}))^2 + \frac{1}{2} \sum_{i=1}^{q} (\max\{0, h_i(\mathbf{x})\})^2 \tag{3.153}$$

verwendet werden. Je nach Wertebereich und physikalischer Bedeutung der Funktionen $g_i(\mathbf{x})$ und $h_i(\mathbf{x})$ kann es für das numerische Lösungsverhalten der Methode sinnvoll sein, die relative Bedeutung der einzelnen Summanden in $P(\mathbf{x})$ durch zusätzliche Gewichtungsfaktoren oder Normierungen zu beeinflussen (vgl. [3.7]). Abbildung 3.8 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 $h_2(x) = a - x$.

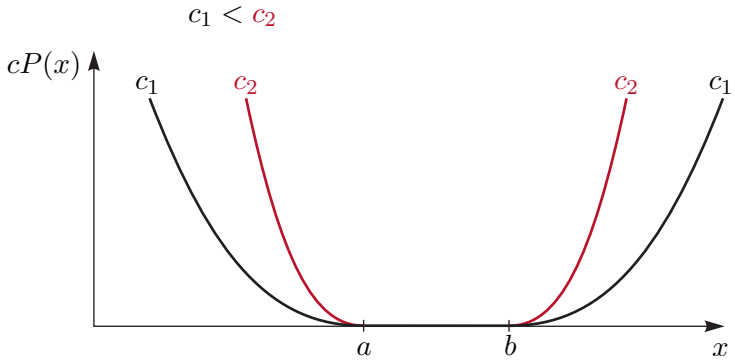


Abbildung 3.8: Straffunktionen $cP(x)$ für verschiedene Werte von c.

Für größere Werte von c ist zu erwarten, dass die Lösung des unbeschränkten Optimierungsproblems (3.152) zumindest in der Nähe des zulässigen Gebiets $\mathcal{X}$ zu liegen kommt und für $c \to \infty$ wird die Lösung von (3.152) gegen jene von (3.151) konvergieren. Dabei wird so vorgegangen, dass für eine gegen Unendlich strebende Folge $\{c_l\}$, $l = 1, 2, \ldots$ mit $c_1 > 0$ und $c_{l+1} > c_l$ das unbeschränkte Optimierungsproblem (3.152) wiederkehrend mit $c = c_l$ gelöst wird und man jeweils einen optimalen Punkt $\mathbf{x}_l^*$ erhält. Für die Lösung des Optimierungsproblems (3.152) 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 die zugehörigen Lösungen $\mathbf{x}_l^*$ und $\mathbf{x}_{l+1}^*$ des unbeschränkten Optimierungsproblems (3.152) gelten folgende Ungleichungen*

$$f(\mathbf{x}_l^*) + c_l P(\mathbf{x}_l^*) \leq f(\mathbf{x}_{l+1}^*) + c_{l+1} P(\mathbf{x}_{l+1}^*) \tag{3.154a}$$

$$P(\mathbf{x}_l^*) \geq P(\mathbf{x}_{l+1}^*) \tag{3.154b}$$

$$f(\mathbf{x}_l^*) \leq f(\mathbf{x}_{l+1}^*) \ . \tag{3.154c}$$

Beweis. Aufgrund von $c_{l+1} > c_l$ und der Definitionen von $\mathbf{x}_l^*$ und $\mathbf{x}_{l+1}^*$ gilt unmittelbar

$$f(\mathbf{x}_{l+1}^*) + c_{l+1} P(\mathbf{x}_{l+1}^*) \geq f(\mathbf{x}_{l+1}^*) + c_l P(\mathbf{x}_{l+1}^*) \geq f(\mathbf{x}_l^*) + c_l P(\mathbf{x}_l^*), \tag{3.155}$$

womit (3.154a) gezeigt ist. Aus

$$-f(\mathbf{x}_{l+1}^*) - c_l P(\mathbf{x}_{l+1}^*) \leq -f(\mathbf{x}_l^*) - c_l P(\mathbf{x}_l^*) \tag{3.156a}$$

$$f(\mathbf{x}_{l+1}^*) + c_{l+1}P(\mathbf{x}_{l+1}^*) \leq f(\mathbf{x}_l^*) + c_{l+1}P(\mathbf{x}_l^*) \tag{3.156b}$$

folgt

$$(c_{l+1} - c_l)P(\mathbf{x}_{l+1}^*) \leq (c_{l+1} - c_l)P(\mathbf{x}_l^*) \tag{3.157}$$

und mit $c_{l+1} > c_l$ daher (3.154b). Aus (3.155) erhält man

$$f(\mathbf{x}_{l+1}^*) + c_l \underbrace{\left(P(\mathbf{x}_{l+1}^*) - P(\mathbf{x}_l^*)\right)}_{\leq 0} \geq f(\mathbf{x}_l^*), \tag{3.158}$$

woraus sich schließlich (3.154c) ergibt. □

Lemma 3.4 (Methode der Straffunktionen). *Wenn $\mathbf{x}^*$ die Lösung des beschränkten Optimierungsproblems (3.151) ist, dann gilt für jede Iteration l*

$$f(\mathbf{x}^*) \geq f(\mathbf{x}_l^*) + c_l P(\mathbf{x}_l^*) \geq f(\mathbf{x}_l^*) \ . \tag{3.159}$$

Aufgabe 3.16. Beweisen Sie Lemma 3.4.

Satz 3.15 (Konvergenz der Methode der Straffunktionen). *Angenommen, $\{\mathbf{x}_l^*\}$ sei eine Folge von Punkten, die durch die Lösung des unbeschränkten Optimierungsproblems (3.152) für eine gegen Unendlich strebende Folge $\{c_l\}$, $l = 1, 2, \ldots$ 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 (3.151).*

3.2.5.2 Barrierefunktionen

Die Methode der Barrierefunktionen ist auf das beschränkte Optimierungsproblem (3.151) 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 $\check{\mathcal{X}}$ das (nichtleere) Innere von $\mathcal{X}$. Eine *Barrierefunktion* $B(\mathbf{x})$ ist auf $\check{\mathcal{X}}$ definiert und ist auf diesem Gebiet stetig. Außerdem gilt $B(\mathbf{x}) \to \infty$, wenn sich $\mathbf{x}$ dem Rand von $\mathcal{X}$ nähert.

Angenommen, $\mathcal{X} = \{\mathbf{x} \in \mathbb{R}^n \,|\, h_i(\mathbf{x}) \leq 0,\ i = 1, \ldots, q\}$ sei eine robuste Menge mit dem Inneren $\check{\mathcal{X}} = \{\mathbf{x} \in \mathbb{R}^n \,|\, h_i(\mathbf{x}) < 0,\ i = 1, \ldots, q\}$, dann kann als Barrierefunktion

$$B(\mathbf{x}) = -\sum_{i=1}^{q} \frac{1}{h_i(\mathbf{x})} \tag{3.160}$$

verwendet werden. Abbildung 3.9 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 die Funktion

$$B(\mathbf{x}) = -\sum_{i=1}^{q} \log(-h_i(\mathbf{x})) \ . \tag{3.161}$$

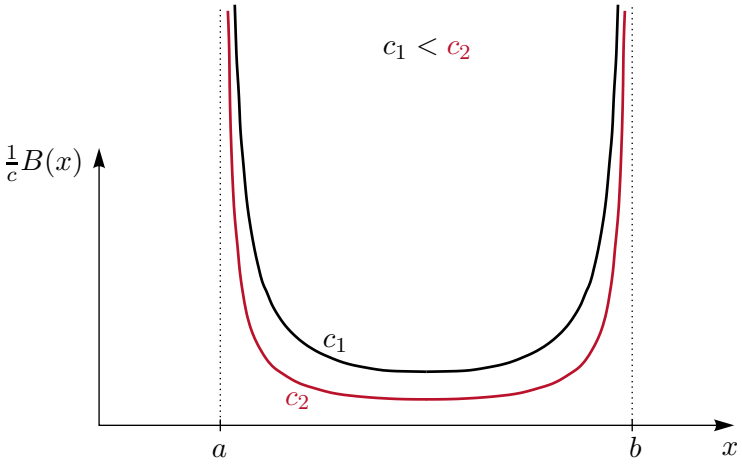


Abbildung 3.9: Barrierefunktionen $\frac{1}{c}B(x)$ für verschiedene Werte von c.

Je nach Wertebereich und physikalischer Bedeutung der Funktionen $h_i(\mathbf{x})$ kann es für das numerische Lösungsverhalten der Methode sinnvoll sein, die relative Bedeutung der einzelnen Summanden in $B(\mathbf{x})$ durch zusätzliche Gewichtungsfaktoren oder Normierungen zu beeinflussen (vgl. [3.7]).

Die Vorgehensweise bei der Methode der Barrierefunktionen ist nun ähnlich zur Methode der Straffunktionen. Es wird für eine gegen Unendlich strebende Folge $\{c_l\},\ l = 1, 2, \ldots$ mit $c_1 > 0$ und $c_{l+1} > c_l$ das Optimierungsproblem

$$\min_{\mathbf{x}\in\check{\mathcal{X}}} f(\mathbf{x}) + \frac{1}{c_l}B(\mathbf{x}) \tag{3.162}$$

gelöst und mit $\mathbf{x}_l^*$ der jeweilige optimale Punkt bezeichnet. Es handelt sich bei (3.162) noch immer um ein beschränktes Optimierungsproblem, dessen Beschränkungen sogar restriktiver sein können als jene des ursprünglichen Problems (3.151). Dennoch kann die Lösung von (3.162) mit Methoden der unbeschränkten statischen Optimierung, wie sie z. B. in Abschnitt 2 vorgestellt wurden, erfolgen, da der Kostenfunktionswert nahe dem Rand von $\mathcal{X}$ gegen Unendlich strebt. In diesem Zusammenhang ist bei der Verwendung von Methoden der unbeschränkten statischen Optimierung besonders darauf zu achten, dass die Kostenfunktion $f(\mathbf{x}) + B(\mathbf{x})/c_l$ nur im (nicht leeren) Inneren $\check{\mathcal{X}}$ des zulässigen Gebiets $\mathcal{X}$ definiert ist, d. h. nur dort ausgewertet werden darf. Folglich ist bei der Methode der Barrierefunktionen jeder Punkt $\mathbf{x}_l^*$ ein zulässiger Punkt, d. h. es gilt $\mathbf{x}_l^* \in \mathcal{X}$. Bei der numerischen Lösung von (3.162) bietet es sich wieder an, $\mathbf{x}_l^*$ als Startpunkt für die Optimierung mit $c_{l+1} > c_l$ zu verwenden. Bei der Methode der Barrierefunktionen gelten folgende Hilfssätze.

Lemma 3.5 (Ungleichungen bei der Methode der Barrierefunktionen)**.** *Für $c_{l+1} > c_l > 0$ und die zugehörigen Lösungen $\mathbf{x}_l^*$ und $\mathbf{x}_{l+1}^*$ des unbeschränkten Optimierungs-*

problems (3.162) *gelten folgende Ungleichungen*

$$f(\mathbf{x}_l^*) + \frac{1}{c_l} B(\mathbf{x}_l^*) \geq f(\mathbf{x}_{l+1}^*) + \frac{1}{c_{l+1}} B(\mathbf{x}_{l+1}^*) \tag{3.163a}$$

$$B(\mathbf{x}_l^*) \leq B(\mathbf{x}_{l+1}^*) \tag{3.163b}$$

$$f(\mathbf{x}_l^*) \geq f(\mathbf{x}_{l+1}^*) \ . \tag{3.163c}$$

Aufgabe 3.17. Beweisen Sie Lemma 3.5.

Lemma 3.6 (Methode der Barrierefunktionen)**.** *Wenn* $\mathbf{x}^*$ *die Lösung des Optimierungsproblems* (3.151)*ist, dann gilt für jede Iteration* l

$$f(\mathbf{x}^*) \leq f(\mathbf{x}_l^*) \leq f(\mathbf{x}_l^*) + \frac{1}{c_l} B(\mathbf{x}_l^*) \ . \tag{3.164}$$

Aufgabe 3.18. Beweisen Sie Lemma 3.6.

Satz 3.16 (Konvergenz der Methode der Barrierefunktionen)**.** *Angenommen,* $\{\mathbf{x}_l^*\}$ *sei eine Folge von Punkten, die durch die Lösung des Optimierungsproblems* (3.162) *für eine gegen Unendlich strebende Folge* $\{c_l\}$, $l = 1, 2, \ldots$ *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* (3.151)*.*

3.3 Beispiel: Rosenbrock's „Bananenfunktion“

Es wird das beschränkte Optimierungsproblem (vgl. (2.127))

$$\min_{\mathbf{x} \in \mathbb{R}^2} \quad f(\mathbf{x}) = 100(x_2 - x_1^2)^2 + (x_1 - 1)^2 \tag{3.165a}$$

$$\text{u.B.v.} \quad x_1^2 + x_2^2 \geq 0.5^2 \tag{3.165b}$$

betrachtet. Die Kostenfunktion („Bananenfunktion“) ist gemeinsam mit dem Rand des zulässigen Gebiets $\mathcal{X}$ in Abbildung 3.10 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.

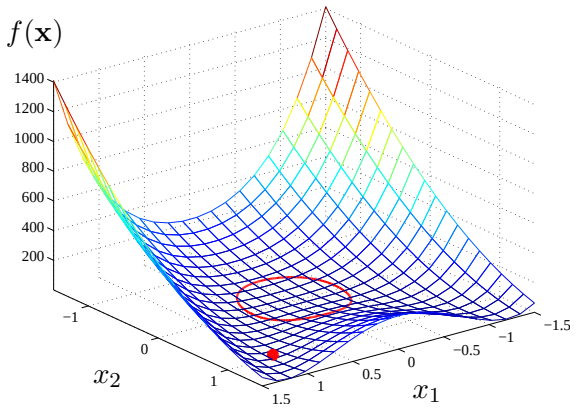


Abbildung 3.10: Profil der Rosenbrock Bananenfunktion und Rand des zulässigen Gebiets.

4. `trust region reflective`: Methode der Vertrauensbereiche, erweitert auf Optimierungsprobleme mit Beschränkungen der Form $\mathbf{Ax} = \mathbf{b}$ *oder* alternativ 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 (3.165) 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 3.11 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.

Code-Auflistung 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
  global old
  old = [Xinit; rosenbrock(Xinit)];
  % Optionen für die Ausgabe
  opt = optimoptions('fmincon','Display','iter','PlotFcns',@plot_iterates);

  switch testCase
    case 1     % Active-Set mit SQP
      opt = optimoptions(opt,'Algorithm','active-set','SpecifyObjectiveGradient',true);
      opt = optimoptions(opt,'MaxFunctionEvaluations',1000);
      [Xopt,fopt,exitflag,output] = fmincon(@rosenbrock,Xinit,[],[],[],[],[],[],@nonlconstr1,opt);
    case 2     % SQP
      opt = optimoptions(opt,'Algorithm','sqp','SpecifyObjectiveGradient',true);
      opt = optimoptions(opt,'MaxFunctionEvaluations',2000);
      [Xopt,fopt,exitflag,output] = fmincon(@rosenbrock,Xinit,[],[],[],[],[],[],@nonlconstr1,opt);
    case 3     % Interior-Point
      opt = optimoptions(opt,'Algorithm','interior-point','SpecifyObjectiveGradient',true);
      [Xopt,fopt,exitflag,output] = fmincon(@rosenbrock,Xinit,[],[],[],[],[],[],@nonlconstr1,opt);
  end
```

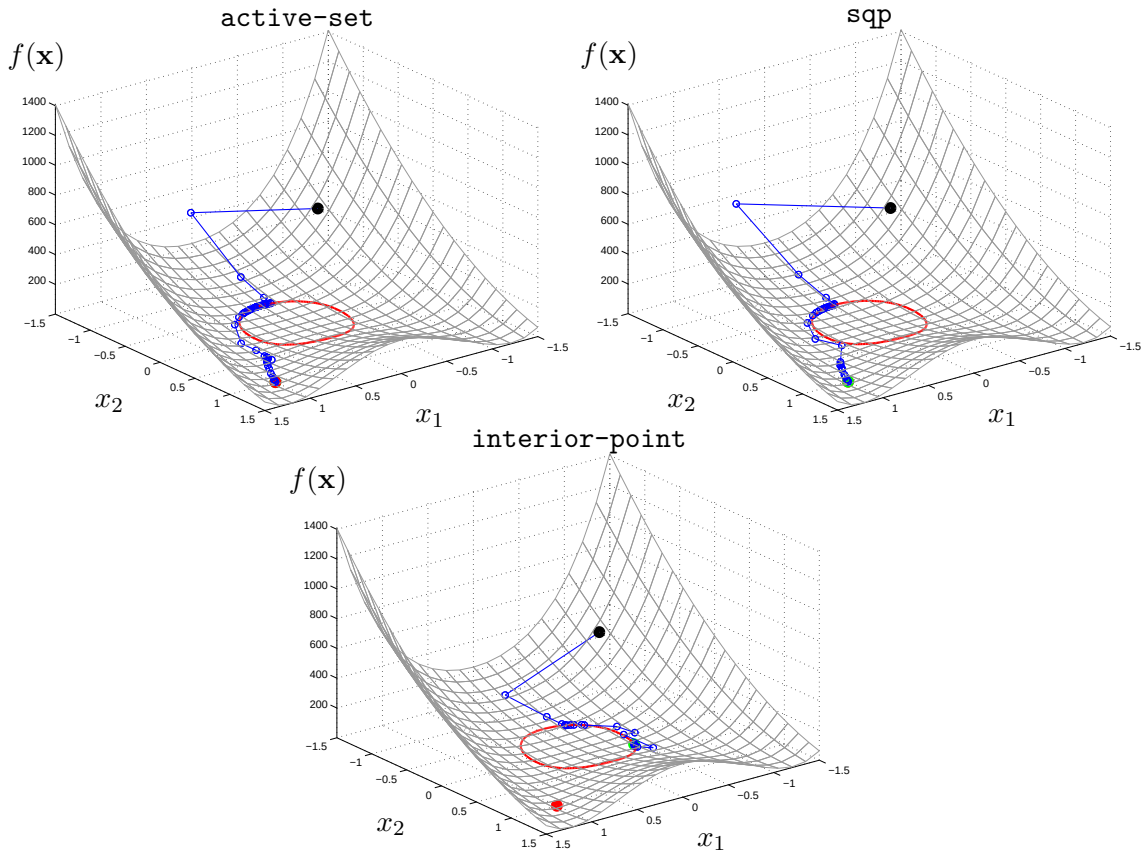


Abbildung 3.11: Rosenbrock Bananenfunktion: Vergleich der numerischen Verfahren aus `fmincon`.

```
end

function [c,ceq] = nonlconstr1(x) % Nichtlineare Beschränkungsfunktion
  r = 0.5;
  c = r^2 - (x(1))^2 - (x(2))^2; ceq = [];
end

function stop = plot_iterates(x,info,state)
  global old
  f = rosenbrock(x);
  switch state                                          % Grafische Ausgabe:
    case 'init'                                         % Initialisierung
      plot_surface(x,f);
      plot_constraints;
    case 'iter'                                         % Iterationen
      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;                                         % kein Abbruchkriterium
  old = [x;f];
end

function plot_constraints % Zeichnen der eingestellten Beschränkung
```

```
  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');
end

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
  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');
end

function [f, grad, H] = rosenbrock(x)
  grad = [];
  H = [];
  f = 100*(x(2)-x(1)^2)^2 + (x(1)-1)^2;                     % Rosenbrock-Funktion
  if nargout>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 nargout>2                                              % falls Hessematrix angefordert wird
    H = [ -400*(x(2)-3*x(1)^2)+2, -400*x(1); -400*x(1), 200 ];
  end
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)
- linprog: SciPy Bibliothek Optimization (frei zugänglich)
- 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`
- OSQP (frei zugänglich)
  `https://osqp.org`
- 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)
- minimize: SciPy Bibliothek Optimization (frei zugänglich)
- LOQO (kostenpflichtig)
  `http://www.princeton.edu/~rvdb`
- MINOS (kostenpflichtig)
  `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`
- 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] A. Kugi, *Skriptum zur VO Nichtlineare dynamische Systeme und Regelung (SS 2024)*, Institut für Automatisierungs- und Regelungstechnik, TU Wien, 2024. Adresse: https://www.acin.tuwien.ac.at/master/nichtlineare-dynamische-systeme-und-regelung/.

[3.2] I. Griva, S. Nash und A. Sofer, *Linear and Nonlinear Optimization*, 2. Aufl. Society for Industrial und Applied Mathematics, 2009.

[3.3] D. P. Bertsekas, *Nonlinear Programming*, 2. Aufl. Athena Scientific, 1999.

[3.4] M. Bazaraa, H. Sherali und C. Shetty, *Nonlinear Programming: Theory and Algorithms*, 3. Aufl. John Wiley & Sons, 2006.

[3.5] D. G. Luenberger und Y. Ye, *Linear and Nonlinear Programming* (International Series in Operations Research & Management Science), 3. Aufl. Springer, 2008, Bd. 116.

[3.6] L. Biegler, *Nonlinear Programming: Concepts, Algorithms, and Applications to Chemical Processes*. Society for Industrial und Applied Mathematics, 2010.

[3.7] S. S. Rao, *Engineering Optimization, Theory and Practice*, 4. Aufl. John Wiley & Sons, 2009.

[3.8] J. Nocedal und S. J. Wright, *Numerical Optimization* (Springer Series in Operations Research and Financial Engineering), 2. Aufl. Springer, 2006.

[3.9] S. Boyd und L. Vandenberghe, *Convex Optimization*. Cambridge University Press, 2004.

[3.10] M. Papageorgiou, M. Leibold und M. Buss, *Optimierung: Statische, dynamische, stochastische Verfahren für die Anwendung*, 4. Aufl. Springer, 2015.

[3.11] C. T. Kelley, *Iterative Methods for Optimization*. Society for Industrial und Applied Mathematics, 1999.

[3.12] 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 30. 09. 2020).

[3.13] P. E. Gill, W. Murray und M. H. Wright, *Practical Optimization*. Academic Press, 1981.

[3.14] S.-P. Han, „A Globally Convergent Method for Nonlinear Programming," *Journal of Optimization Theory and Applications*, Jg. 22, Nr. 3, S. 297–309, 1977.

[3.15] 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.

[3.16] 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 *Kostenfunktionals* $J : \mathcal{V} \to \mathbb{R}$ bezüglich einer (reellen vektorwertigen) Funktion $\mathbf{x}(t)$ aus einem geeigneten *Funktionenraum* $\mathcal{V}$ gesucht. In vielen Fällen entspricht die *unabhängige Variable* t der Zeit. Die totale Ableitung nach t wird mit $\dot{(\,\cdot\,)} = \mathrm{d}(\,\cdot\,)/\mathrm{d}t$ abgekürzt. Beispielhaft kann eine dynamische Optimierungsaufgabe in der Form

$$\min_{\mathbf{x}(\cdot)\in\mathcal{V}} \quad J(\mathbf{x}) = \int_{t_0}^{t_1} l(t, \mathbf{x}(t), \dot{\mathbf{x}}(t))\,\mathrm{d}t \qquad \text{Kostenfunktional} \tag{4.1a}$$

$$\text{u.B.v.} \quad \mathbf{x}(t_0) = \mathbf{x}_0 \qquad \text{Beschränkungen} \tag{4.1b}$$

$$\boldsymbol{\psi}_1(t_1, \mathbf{x}(t_1)) = \mathbf{0} \tag{4.1c}$$

$$\boldsymbol{\psi}_2(t, \mathbf{x}(t), \dot{\mathbf{x}}(t)) \leq \mathbf{0} \quad \forall\, t \in I \subseteq [t_0, t_1] \tag{4.1d}$$

angeschrieben werden. Typischerweise hat das Kostenfunktional die Form (*Lagrange Problem der Variationsrechnung*)

$$J(\mathbf{x}) = \int_{t_0}^{t_1} l(t, \mathbf{x}(t), \dot{\mathbf{x}}(t))\,\mathrm{d}t\ , \tag{4.2}$$

(*Bolza Problem der Variationsrechnung*)

$$J(\mathbf{x}) = \varphi(t_0, \mathbf{x}(t_0), t_1, \mathbf{x}(t_1)) + \int_{t_0}^{t_1} l(t, \mathbf{x}(t), \dot{\mathbf{x}}(t))\,\mathrm{d}t \tag{4.3}$$

oder (*Mayer Problem der Variationsrechnung*)

$$J(\mathbf{x}) = \varphi(t_0, \mathbf{x}(t_0), t_1, \mathbf{x}(t_1))\ . \tag{4.4}$$

Dabei wird $\mathbf{x}(t) = [x_1(t) \quad \dots \quad x_n(t)]^{\mathrm{T}} : [t_0, t_1] \to \mathbb{R}^n$ häufig als *Trajektorie* bezeichnet. Die reellwertige Funktion $l(t, \mathbf{x}(t), \dot{\mathbf{x}}(t)) : \mathbb{R} \times \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}$ nennt man *Lagrangesche Dichte* und $\varphi(t_0, \mathbf{x}(t_0), t_1, \mathbf{x}(t_1)) : \mathbb{R} \times \mathbb{R}^n \times \mathbb{R} \times \mathbb{R}^n \to \mathbb{R}$ 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. Es wird zwischen den folgenden Arten von Beschränkungen unterschieden:

- **Punktbeschränkungen:** Die einfachste Form von Punktbeschränkungen ist, dass beide Endpunkte fixiert sind, d. h. $\mathbf{x}(t_0) = \mathbf{x}_0$ und $\mathbf{x}(t_1) = \mathbf{x}_1$. Folglich gilt dann $\mathcal{X} = \{\mathbf{x}(t) \in \mathcal{V} \,|\, \mathbf{x}(t_0) = \mathbf{x}_0,\, \mathbf{x}(t_1) = \mathbf{x}_1\}$. Eine weitere Möglichkeit für Punktbeschränkungen ist, 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 einem vorgegebenen Gebiet zu liegen kommen muss, welches implizit durch Gleichungen der Art

$$\boldsymbol{\psi}(t_1, \mathbf{x}(t_1)) = \mathbf{0} \quad \text{bzw.} \quad \boldsymbol{\psi}(t_1, \mathbf{x}(t_1)) \leq \mathbf{0} \tag{4.5}$$

  definiert ist. In diesem Fall ist die freie Endzeit t_1 eine zu optimierende Größe und die zulässige Menge für $\mathbf{x}(t)$ lautet $\mathcal{X} = \{\mathbf{x}(t) \in \mathcal{V} \,|\, \mathbf{x}(t_0) = \mathbf{x}_0,\ \boldsymbol{\psi}(t_1, \mathbf{x}(t_1)) = \mathbf{0}\}$ bzw. $\mathcal{X} = \{\mathbf{x}(t) \in \mathcal{V} \,|\, \mathbf{x}(t_0) = \mathbf{x}_0,\ \boldsymbol{\psi}(t_1, \mathbf{x}(t_1)) \leq \mathbf{0}\}$.

- **Pfadbeschränkungen:** Pfadbeschränkungen (englisch: *path constraints*) können in der Form

$$\psi(t, \mathbf{x}(t), \dot{\mathbf{x}}(t)) = 0 \quad \text{bzw.} \quad \psi(t, \mathbf{x}(t), \dot{\mathbf{x}}(t)) \leq 0, \quad \forall\, t \in I \subseteq [t_0, t_1] \tag{4.6}$$

  mit einem Intervall I, dessen Länge größer Null ist, formuliert werden.

- **Isoperimetrische Beschränkungen:** Als isoperimetrische Beschränkungen werden Beschränkungen der Form

$$\int_{t_0}^{t_1} \psi(t, \mathbf{x}(t), \dot{\mathbf{x}}(t))\, \mathrm{d}t = 0 \quad \text{bzw.} \quad \int_{t_0}^{t_1} \psi(t, \mathbf{x}(t), \dot{\mathbf{x}}(t))\, \mathrm{d}t \leq 0 \tag{4.7}$$

  verstanden.

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*, welche im Weiteren als $(\hat{C}^1[t_0, t_1])^n$ bezeichnet werden. Elemente des Funktionenraumes $(\hat{C}^1[t_0, t_1])^n$ werden auch als global stetig angenommen. Die Definition eines *globalen Minimums* $\mathbf{x}^*$ eines Kostenfunktionals $J(\mathbf{x})$ lässt sich einfach direkt (ohne Verwendung einer Norm) in der Form

$$J(\mathbf{x}^*) \leq J(\mathbf{x}), \quad \forall \mathbf{x} \in \mathcal{X} \tag{4.8}$$

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

$$\exists \gamma > 0 \text{ so, dass gilt } J(\mathbf{x}^*) \leq J(\mathbf{x}), \quad \forall \mathbf{x} \in \{\mathbf{x} \in \mathcal{X} \,|\, \|\mathbf{x} - \mathbf{x}^*\| < \gamma\}\ . \tag{4.9}$$

Da in infinit-dimensionalen Vektorräumen Normen grundsätzlich nicht äquivalent sind, kann $\mathbf{x}^*$ zwar bezüglich einer Norm ein lokales Minimum sein, aber möglicherweise nicht bezüglich einer anderen Norm. Im Funktionenraum $(C^1[t_0, t_1])^n$ werden häufig die Normen

$$\|\mathbf{x}(t)\|_\infty := \max_{t_0 \leq t \leq t_1} \|\mathbf{x}(t)\| \quad \text{und} \quad \|\mathbf{x}(t)\|_{1,\infty} := \max_{t_0 \leq t \leq t_1} \|\mathbf{x}(t)\| + \max_{t_0 \leq t \leq t_1} \|\dot{\mathbf{x}}(t)\| \tag{4.10}$$

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 Funktionals*.

Definition 4.1 (Variation eines Funktionals, Gâteaux Ableitung)**.** Die *erste Variation des Funktionals* $J(\mathbf{x})$ am Punkt $\mathbf{x} \in \mathcal{V}$ in Richtung $\boldsymbol{\xi} \in \mathcal{V}$, auch als *Gâteaux Ableitung* von $J(\mathbf{x})$ am Punkt $\mathbf{x}$ in Richtung $\boldsymbol{\xi}$ bezeichnet, ist in der Form

$$\delta J(\mathbf{x};\boldsymbol{\xi}) := \lim_{\eta\to 0} \frac{J(\mathbf{x}+\eta\boldsymbol{\xi}) - J(\mathbf{x})}{\eta} = \frac{\mathrm{d}}{\mathrm{d}\eta} J(\mathbf{x}+\eta\boldsymbol{\xi})\Big|_{\eta=0} \tag{4.11}$$

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 also die Ableitung von $J(\mathbf{x}+\eta\boldsymbol{\xi})$ bezüglich η an der Stelle $\eta = 0$ existieren.

Beispiel 4.1. Die Gâteaux Ableitung des Funktionals $J(x) = \int_{t_0}^{t_1} x^2(t)\,\mathrm{d}t$, $x \in C^1[t_0,t_1]$ lautet

$$\begin{aligned}\delta J(x;\xi) &= \lim_{\eta\to 0} \frac{1}{\eta}\left(\int_{t_0}^{t_1} (x(t)+\eta\xi(t))^2\,\mathrm{d}t - \int_{t_0}^{t_1} x^2(t)\,\mathrm{d}t\right)\\ &= \lim_{\eta\to 0}\left(\int_{t_0}^{t_1} 2x(t)\xi(t)\,\mathrm{d}t + \eta\int_{t_0}^{t_1}\xi^2(t)\,\mathrm{d}t\right) = 2\int_{t_0}^{t_1} x(t)\xi(t)\,\mathrm{d}t\end{aligned} \tag{4.12}$$

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 $J(x) = \int_0^1 |x(t)|\,\mathrm{d}t$, $x \in C^1[0,1]$. Für $x_0(t) = 0$ und $\xi_0(t) = t$ lautet dessen Gâteaux Ableitung

$$\delta J(x_0;\xi_0) = \lim_{\eta\to 0}\frac{1}{\eta}\left(\int_0^1 |x_0+\eta\xi_0|\,\mathrm{d}t - \int_0^1 |x_0|\,\mathrm{d}t\right) = \tag{4.13a}$$

$$= \lim_{\eta\to 0}\operatorname{sgn}(\eta)\int_0^1 |t|\,\mathrm{d}t = \begin{cases}\frac{1}{2}, & \eta\to+0\\ -\frac{1}{2}, & \eta\to-0\end{cases}. \tag{4.13b}$$

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

$$\delta(J_1+J_2)(\mathbf{x};\boldsymbol{\xi}) = \delta J_1(\mathbf{x};\boldsymbol{\xi}) + \delta J_2(\mathbf{x};\boldsymbol{\xi}) \tag{4.14}$$

und für jedes reelle α gilt die Beziehung

$$\delta J(\mathbf{x};\alpha\boldsymbol{\xi}) = \alpha\delta J(\mathbf{x};\boldsymbol{\xi})\ . \tag{4.15}$$

Basierend auf der Gâteaux Ableitung lässt sich nun der Begriff der *zulässigen Richtung* eines Funktionals definieren.

Definition 4.2 (Zulässige Richtung). $J : \mathcal{X} \to \mathbb{R}$ 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.

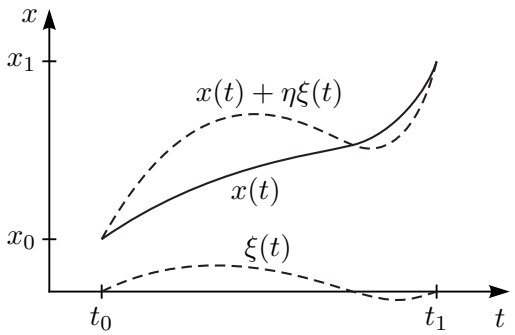


Abbildung 4.1: Zulässige Richtung im Fall $\mathcal{X} = \{x(t) \in C[t_0, t_1] \,|\, x(t_0) = x_0,\ x(t_1) = x_1\}$.

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 $\mathcal{X} = \{x(t) \in C[t_0, t_1] \,|\, x(t_0) = x_0,\ x(t_1) = x_1\}$. Eine zulässige Richtung $\boldsymbol{\xi}$ an einem Punkt $\mathbf{x}$ für die gilt $\delta J(\mathbf{x}; \boldsymbol{\xi}) < 0$ wird *Abstiegsrichtung* des Funktionals J am Punkt $\mathbf{x}$ bezeichnet. Dies stellt eine Generalisierung der Abstiegsrichtung $\mathbf{d}$ der Kostenfunktion $f(\mathbf{x})$ im finit-dimensionalen Fall mit $\mathbf{d}^{\mathrm{T}}(\nabla f)(\mathbf{x}) < 0$ am Punkt $\mathbf{x}$ 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 $\mathbf{x} \in \mathcal{X}$ eine zulässige Richtung $\boldsymbol{\xi} \in \mathcal{V}$ so existiert, dass gilt $\delta J(\mathbf{x}; \boldsymbol{\xi}) < 0$, dann kann $\mathbf{x}$ kein lokales Minimum sein.*

Beweisskizze: Gemäß Definition 4.1 gilt

$$\delta J(\mathbf{x}; \boldsymbol{\xi}) = \lim_{\eta \to 0} \frac{J(\mathbf{x} + \eta\boldsymbol{\xi}) - J(\mathbf{x})}{\eta} < 0 \tag{4.16}$$

und es existiert ein $\gamma > 0$ so, dass

$$J(\mathbf{x} + \eta\boldsymbol{\xi}) < J(\mathbf{x}), \quad \forall \eta \in (0, \gamma)\ . \tag{4.17}$$

Da nun $\boldsymbol{\xi}$ eine zulässige Richtung gemäß Definition 4.2 ist, kann das Funktional J am Punkt $\mathbf{x}$ in Richtung $\eta\boldsymbol{\xi}$ mit beliebigem $\eta \in (0, \gamma)$ weiter verkleinert werden. Da unabhängig von der verwendeten Norm $\|\mathbf{x} + \eta\boldsymbol{\xi} - \mathbf{x}\| = \|\eta\boldsymbol{\xi}\| \to 0$ für $\eta \to 0$ gilt,

findet man stets einen hinreichend kleinen Wert $\eta \in (0, \gamma)$, so dass $\mathbf{x} + \eta\boldsymbol{\xi}$ im Sinne der Norm $\|\cdot\|$ in der Umgebung von $\mathbf{x}$ liegt. Folglich kann $\mathbf{x}$ kein lokales Minimum sein. □

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*

$$\delta J(\mathbf{x}^*; \boldsymbol{\xi}) = 0 \tag{4.18}$$

für alle zulässigen Richtungen $\boldsymbol{\xi}$ gemäß Definition 4.2 an der Stelle $\mathbf{x}^*$.

Im nächsten Schritt betrachte man das Lagrange Problem der Variationsrechnung gemäß (4.2) mit festem Anfangs- und Endpunkt.

Satz 4.2 (Euler-Lagrange Gleichungen). *Gegeben sei das Funktional*

$$J(\mathbf{x}) = \int_{t_0}^{t_1} l(t, \mathbf{x}(t), \dot{\mathbf{x}}(t))\,\mathrm{d}t \tag{4.19}$$

mit der zulässigen Menge $\mathcal{X} = \left\{\mathbf{x}(t) \in (C^1[t_0, t_1])^n \,|\, \mathbf{x}(t_0) = \mathbf{x}_0,\ \mathbf{x}(t_1) = \mathbf{x}_1\right\}$ und der stetig differenzierbaren Lagrangeschen Dichte $l : \mathbb{R} \times \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}$. Wenn $\mathbf{x}^*(t)$ ein (lokales) Minimum von $J(\mathbf{x})$ auf $\mathcal{X}$ bezeichnet, dann erfüllt $\mathbf{x}^*(t)$ die Euler-Lagrange Gleichungen

$$\frac{\mathrm{d}}{\mathrm{d}t}\left(\frac{\partial}{\partial \dot{\mathbf{x}}} l\right)^{\mathrm{T}}(t, \mathbf{x}^*(t), \dot{\mathbf{x}}^*(t)) - \left(\frac{\partial}{\partial \mathbf{x}} l\right)^{\mathrm{T}}(t, \mathbf{x}^*(t), \dot{\mathbf{x}}^*(t)) = \mathbf{0} \tag{4.20}$$

für alle $t \in [t_0, t_1]$.

Beweis. Da $\mathbf{x}^*$ ein Minimum ist, muss wegen Satz 4.1 gelten

$$\begin{aligned}\delta J(\mathbf{x}^*; \boldsymbol{\xi}) &= \frac{\mathrm{d}}{\mathrm{d}\eta} J(\mathbf{x}^* + \eta\boldsymbol{\xi})\Big|_{\eta=0} = \int_{t_0}^{t_1} \frac{\mathrm{d}}{\mathrm{d}\eta} l\Big(t, \mathbf{x}^*(t) + \eta\boldsymbol{\xi}(t), \dot{\mathbf{x}}^*(t) + \eta\dot{\boldsymbol{\xi}}(t)\Big)\,\mathrm{d}t\Big|_{\eta=0} \\ &= \int_{t_0}^{t_1} \left[\left(\frac{\partial}{\partial \mathbf{x}} l\right)(t, \mathbf{x}^*(t), \dot{\mathbf{x}}^*(t))\boldsymbol{\xi} + \left(\frac{\partial}{\partial \dot{\mathbf{x}}} l\right)(t, \mathbf{x}^*(t), \dot{\mathbf{x}}^*(t))\dot{\boldsymbol{\xi}}\right] \mathrm{d}t = 0\ .\end{aligned} \tag{4.21}$$

Wegen der stetigen Differenzierbarkeit der Lagrangeschen Dichte l und da $\boldsymbol{\xi} \in (C^1[t_0, t_1])^n$ ist der Integrand von (4.21) 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 Defintion 4.2 zulässige Richtung $\boldsymbol{\xi}$ muss die Bedingungen $\boldsymbol{\xi}(t_0) = \mathbf{0}$ und $\boldsymbol{\xi}(t_1) = \mathbf{0}$ erfüllen. Führt man für den zweiten Summanden in der zweiten Zeile von

(4.21) eine partielle Integration durch, so erhält man

$$\int_{t_0}^{t_1} \left(\frac{\partial}{\partial \mathbf{x}} l\right)\boldsymbol{\xi} + \left(\frac{\partial}{\partial \dot{\mathbf{x}}} l\right)\dot{\boldsymbol{\xi}}\,\mathrm{d}t = \int_{t_0}^{t_1} \left[\left(\frac{\partial}{\partial \mathbf{x}} l\right) - \frac{\mathrm{d}}{\mathrm{d}t}\left(\frac{\partial}{\partial \dot{\mathbf{x}}} l\right)\right]\boldsymbol{\xi}\,\mathrm{d}t + \underbrace{\left[\left(\frac{\partial}{\partial \dot{\mathbf{x}}} l\right)\boldsymbol{\xi}\right]_{t_0}^{t_1}}_{=0} = 0\ . \tag{4.22}$$

Wählt man nun nacheinander für festes $i = 1, \ldots, n$ eine Richtung $\boldsymbol{\xi} = [\xi_1 \quad \ldots \quad \xi_n]^{\mathrm{T}} \in \left(C^1[t_0, t_1]\right)^n$ 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 jeweils

$$\int_{t_0}^{t_1} \left[\left(\frac{\partial}{\partial x_i} l\right) - \frac{\mathrm{d}}{\mathrm{d}t}\left(\frac{\partial}{\partial \dot{x}_i} l\right)\right]\xi_i\,\mathrm{d}t = 0\ . \tag{4.23}$$

Gemäß dem nachfolgend angeführten *Fundamentallemma der Variationsrechnung* folgt aus (4.23) mit $i = 1, \ldots, n$ unmittelbar das Ergebnis (4.20). □

Lemma 4.2 (Fundamentallemma der Variationsrechnung)**.** *Angenommen $g(t)$ ist eine stückweise stetige Funktion auf dem Intervall $[t_0, t_1]$ und es gilt*

$$\int_{t_0}^{t_1} g(t)\xi_i(t)\,\mathrm{d}t = 0 \tag{4.24}$$

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 $\mathbf{x}(t)$, die die Euler-Lagrange Gleichungen (4.20) 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 Kostenfunktionals beschreiben. Satz 4.2 stellt also nur eine notwendige Bedingung für eine optimale Lösung $\mathbf{x}^*(t)$ dar, wie auch das nachfolgende Beispiel zeigt.

Beispiel 4.3. Das Funktional $J(x) = \int_0^1 x(t)\dot{x}^2(t)\,\mathrm{d}t$, $x \in C^1[0, 1]$ mit den Randbedingungen $x(0) = x(1) = 0$ soll minimiert werden. Für diesen Fall ergibt sich die Euler-Lagrange Gleichung (4.20) in der Form

$$\frac{\mathrm{d}}{\mathrm{d}t}(2x(t)\dot{x}(t)) - \dot{x}^2(t) = \frac{\mathrm{d}^2}{\mathrm{d}t^2}\left(x^2(t)\right) - \dot{x}^2(t) = 0\ . \tag{4.25}$$

Diese Differentialgleichung besitzt für die Randbedingungen $x(0) = x(1) = 0$ die Lösung $x^*(t) = 0$. Folglich ist $x^*(t) = 0$ eine extremale Lösung und der zugehörige Wert des Kostenfunktionals lautet $J(x^*) = 0$. Dies ist aber kein Minimum, denn die ebenfalls zulässige Trajektorie $\breve{x}(t) = -\varepsilon t(1 - t)$ mit $\varepsilon > 0$ liefert für das Kostenfunktional $J(\breve{x}) = -\varepsilon^2/30 < 0$. Für $\varepsilon \to \infty$ gilt $J(\breve{x}) \to -\infty$, weshalb dieses Optimierungsproblem kein Optimum besitzt. Die Trajektorie $x^*(t) = 0$ stellt auch keine lokal optimale Lösung dar, denn mit $\varepsilon \to 0^+$ kommt $\breve{x}(t)$ der extremalen Lösung $x^*(t)$ im Sinne jeder Norm beliebig nahe.

Mit Satz 4.2 ist es also gelungen, die Minimierung des Funktionals (4.2) in ein *Zweipunktrandwertproblem* 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 (4.20) kann für Spezialfälle auch mit Hilfe so genannter *erster Integrale* formuliert werden:

(a) Die Lagrangesche Dichte hängt nicht von der unabhängigen Variable t ab, d. h. $l = l(\mathbf{x}, \dot{\mathbf{x}})$. Mit der *Hamiltonfunktion*

$$H = \left(\frac{\partial}{\partial \dot{\mathbf{x}}} l\right)\dot{\mathbf{x}} - l(\mathbf{x}, \dot{\mathbf{x}}) \tag{4.26}$$

folgt aus den Euler-Lagrange Gleichungen (4.20), dass

$$\begin{aligned} \frac{\mathrm{d}}{\mathrm{d}t} H &= \frac{\mathrm{d}}{\mathrm{d}t}\left(\frac{\partial}{\partial \dot{\mathbf{x}}} l\right)\dot{\mathbf{x}} + \left(\frac{\partial}{\partial \dot{\mathbf{x}}} l\right)\ddot{\mathbf{x}} - \left(\frac{\partial}{\partial \mathbf{x}} l\right)\dot{\mathbf{x}} - \left(\frac{\partial}{\partial \dot{\mathbf{x}}} l\right)\ddot{\mathbf{x}} \\ &= \left(\frac{\mathrm{d}}{\mathrm{d}t}\left(\frac{\partial}{\partial \dot{\mathbf{x}}} l\right) - \left(\frac{\partial}{\partial \mathbf{x}} l\right)\right)\dot{\mathbf{x}} = 0 \ . \end{aligned} \tag{4.27}$$

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 (4.20), dass $\frac{\partial}{\partial \dot{x}_i} l$, $i = 1, \ldots, n$ *Invarianten* des Systems sind, denn es gilt

$$\frac{\mathrm{d}}{\mathrm{d}t}\left(\frac{\partial}{\partial \dot{x}_i} l\right) = 0 \ . \tag{4.28}$$

Aufgabe 4.1. Nehmen Sie an, dass die Lagrangesche Dichte $l(\mathbf{x}, \dot{\mathbf{x}})$ die Lagrangefunktion eines Starrkörpersystems ist (siehe Skriptum [4.5]) und $\mathbf{x}$ bzw. $\dot{\mathbf{x}}$ die generalisierten Lagekoordinaten und deren Geschwindigkeiten bezeichnen. Geben Sie eine physikalische Interpretation der Hamiltonfunktion H von (4.26) und der darin auftretenden Größen $\frac{\partial}{\partial \dot{x}_i} l$, $i = 1, \ldots, n$ an.

Bemerkung 4.1. Konservative Starrkörpersysteme erfüllen die Euler-Lagrange Gleichungen (4.20). Dies gilt im Allgemeinen nicht für nicht-konservative Starrkörpersysteme. Für sie lauten die Euler-Lagrange Gleichungen

$$\frac{\mathrm{d}}{\mathrm{d}t}\left(\frac{\partial}{\partial \dot{x}_i} l\right) - \left(\frac{\partial}{\partial x_i} l\right) = \tau_i \tag{4.29}$$

für alle $t \in [t_0, t_1]$ und $i = 1, \ldots, n$ mit den externen generalisierten Kräften τ_j.

Bemerkung 4.2. Der Begriff *Lagrangefunktion* hat in der Mechanik eine andere Bedeutung als in der Optimierung.

Beispiel 4.4 (Elastischer Zugstab belastet durch Eigengewicht). Ein gerader, linear elastischer Stab hat die Zugsteifigkeit k und im unbelasteten Zustand die Masse pro Längeneinheit $\bar{m}$ sowie 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.

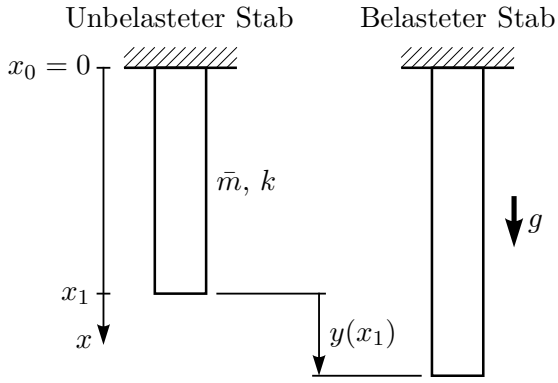


Abbildung 4.2: Elastischer Zugstab belastet durch Eigengewicht.

Zur Lösung dieser Aufgabe kann das *Hamiltonsche Prinzip* der Mechanik [4.6, 4.7] 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 [4.8]. Für ein stabiles statisches Gleichgewicht muss sie minimal sein.

Die bis auf einen konstanten Term definierte potentielle Energie

$$J(y) = \int_0^{x_1} \frac{k(y'(x))^2}{2} - \bar{m}gy(x) \; \mathrm{d}x \tag{4.30}$$

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

$$y(0) = 0 \; . \tag{4.31a}$$

Da am freien Ende $x = x_1$ des Stabes die Zugkraft 0 beträgt, muss dort auch die Dehnung verschwinden, d. h.

$$y'(x_1) = 0 \; . \tag{4.31b}$$

Die Minimierung des Funktionals (4.30) unter Berücksichtigung der Randbedingungen (4.31) 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 (4.27) die Hamiltonfunktion eine Invariante des Systems sein, d. h.

$$H = \left(\frac{\partial}{\partial y'} l\right)(y, y')y' - l(y, y') = \frac{k(y')^2}{2} + \bar{m}gy = c_1 = \text{konst.} \tag{4.32}$$

Die Integration dieser Differentialgleichung liefert

$$\left[-\sqrt{2k}\frac{\sqrt{c_1 - \bar{m}gy}}{\bar{m}g}\right]_{y(0)}^{y(x)} = x \ . \tag{4.33}$$

Die Werte c_1 und $y(0)$ folgen schließlich aus den Randbedingungen (4.31) und für die Lösung ergibt sich

$$y(x) = \frac{\bar{m}g}{k}\left(x_1 x - \frac{x^2}{2}\right) \ . \tag{4.34}$$

Alternativ kann diese Aufgabe natürlich auch direkt mit Satz 4.2 gelöst werden. Aus der Euler-Lagrange Gleichung (4.20) folgt

$$\frac{\mathrm{d}}{\mathrm{d}x}\left(\frac{\partial}{\partial y'} l\right)(y, y') - \frac{\partial}{\partial y} l(y, y') = ky'' + \bar{m}g = 0 \ . \tag{4.35}$$

Die Integration dieser Differentialgleichung liefert bei Berücksichtigung der Randbedingungen (4.31) ebenfalls die Lösung (4.34).

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*

$$J(\mathbf{x}) = \int_{t_0}^{t_1} l(t, \mathbf{x}(t), \dot{\mathbf{x}}(t))\,\mathrm{d}t \tag{4.36}$$

mit der zulässigen Menge $\mathcal{X} = \{\mathbf{x}(t) \in \left(C^1[t_0, t_1]\right)^n \,|\, \mathbf{x}(t_0) = \mathbf{x}_0, \mathbf{x}(t_1) = \mathbf{x}_1\}$ und der zweifach stetig differenzierbaren Lagrangeschen Dichte $l : \mathbb{R} \times \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}$, dann erfüllt $\mathbf{x}^*$ die Euler-Lagrange Gleichungen (4.20) und die so genannte Legendre Bedingung

$$\mathbf{d}^{\mathrm{T}}\left(\frac{\partial^2 l}{\partial \dot{\mathbf{x}}^2}\right)(t, \mathbf{x}^*(t), \dot{\mathbf{x}}^*(t))\mathbf{d} \geq 0 \qquad \forall\, \mathbf{d} \in \mathbb{R}^n,\ t \in [t_0, t_1] \ . \tag{4.37}$$

Satz 4.4 (Hinreichende Bedingungen zweiter Ordnung - Konvexitätsbedingung)**.** *Gegeben sei das Funktional*

$$J(\mathbf{x}) = \int_{t_0}^{t_1} l(t, \mathbf{x}(t), \dot{\mathbf{x}}(t))\,\mathrm{d}t \tag{4.38}$$

mit der zulässigen Menge $\mathcal{X} = \{\mathbf{x}(t) \in (C^1[t_0, t_1])^n \,|\, \mathbf{x}(t_0) = \mathbf{x}_0,\ \mathbf{x}(t_1) = \mathbf{x}_1\}$ *und der zweifach stetig differenzierbaren Lagrangeschen Dichte* $l : \mathbb{R} \times \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}$. *Erfüllt eine Funktion* $\mathbf{x}^* \in \mathcal{X}$ *die Euler-Lagrange Gleichungen* (4.20) *und die sogenannte* Konvexitätsbedingung

$$\mathbf{d}^{\mathrm{T}} \begin{bmatrix} \left(\frac{\partial^2 l}{\partial \mathbf{x}^2}\right)(t, \mathbf{x}^*(t), \dot{\mathbf{x}}^*(t)) & \left(\frac{\partial^2 l}{\partial \mathbf{x} \partial \dot{\mathbf{x}}}\right)(t, \mathbf{x}^*(t), \dot{\mathbf{x}}^*(t)) \\ \left(\frac{\partial^2 l}{\partial \dot{\mathbf{x}} \partial \mathbf{x}}\right)(t, \mathbf{x}^*(t), \dot{\mathbf{x}}^*(t)) & \left(\frac{\partial^2 l}{\partial \dot{\mathbf{x}}^2}\right)(t, \mathbf{x}^*(t), \dot{\mathbf{x}}^*(t)) \end{bmatrix} \mathbf{d} \geq 0 \quad \forall\, \mathbf{d} \in \mathbb{R}^{2n},\ t \in [t_0, t_1]\ , \tag{4.39}$$

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* (4.39) *ist dann nur für* $\mathbf{d} = \mathbf{0}$ *mit Gleichheit erfüllt), so ist* $\mathbf{x}^*(t)$ *ein striktes lokales Minimum.*

Der Beweis zu Satz 4.4 findet sich z. B. in [4.1, 4.9].

Aufgabe 4.2. Prüfen Sie, ob die in Beispiel 4.4 gefundene Lösung die Sätze 4.3 und 4.4 erfüllt.

Satz 4.2 behandelt das Lagrange Problem der Variationsrechnung (4.2). Im nächsten Schritt soll das *Bolza Problem der Variationsrechnung* (4.3) *mit freier Endzeit* näher untersucht werden.

Satz 4.5 (Euler-Lagrange Gleichungen für freie Endzeit)**.** *Gegeben sei das Funktional*

$$J(t_1, \mathbf{x}) = \varphi(t_1, \mathbf{x}(t_1)) + \int_{t_0}^{t_1} l(t, \mathbf{x}(t), \dot{\mathbf{x}}(t))\, \mathrm{d}t \tag{4.40}$$

mit der zulässigen Menge $\mathcal{X} = \left\{(t_1, \mathbf{x}(t)) \in (t_0, T) \times (C^1[t_0, T])^n \,|\, \mathbf{x}(t_0) = \mathbf{x}_0\right\}$, *der hinreichend großen Zeit* $T \gg t_1$, *der stetig differenzierbaren Lagrangeschen Dichte* $l : \mathbb{R} \times \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}$ *und der stetig differenzierbaren Endkostenfunktion* $\varphi : \mathbb{R} \times \mathbb{R}^n \to \mathbb{R}$. *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 (4.20) *im Intervall* $[t_0, t_1^*]$ *und es gelten die Anfangsbedingung* $\mathbf{x}^*(t_0) = \mathbf{x}_0$ *sowie die* Transversalitätsbedingungen

$$\left[\frac{\partial}{\partial \dot{\mathbf{x}}} l + \frac{\partial}{\partial \mathbf{x}} \varphi\right]_{t = t_1^*,\ \mathbf{x} = \mathbf{x}^*} = \mathbf{0}^{\mathrm{T}} \tag{4.41a}$$

$$\left[l - \left(\frac{\partial}{\partial \dot{\mathbf{x}}} l\right)\dot{\mathbf{x}} + \frac{\partial}{\partial t} \varphi\right]_{t = t_1^*,\ \mathbf{x} = \mathbf{x}^*} = 0\ . \tag{4.41b}$$

Beweis. Es wird angenommen, dass $\mathbf{x}(t)$ in einem hinreichend großen Intervall $[t_0, T]$, $T \gg t_1^*$ definiert ist, und es wird der lineare Funktionenraum $\mathbb{R} \times (C^1[t_0, T])^n$ betrachtet. Die Gâteaux Ableitung gemäß Definition 4.1 wird für das Funktional

$J(t_1, \mathbf{x})$ in der Form

$$\begin{aligned}\delta J(t_1, \mathbf{x}; \xi_{t_1}, \boldsymbol{\xi}_x) &:= \lim_{\eta \to 0} \frac{J(t_1 + \eta\xi_{t_1}, \mathbf{x} + \eta\boldsymbol{\xi}_x) - J(t_1, \mathbf{x})}{\eta} \\ &= \frac{\mathrm{d}}{\mathrm{d}\eta} J(t_1 + \eta\xi_{t_1}, \mathbf{x} + \eta\boldsymbol{\xi}_x)\Big|_{\eta=0}\end{aligned} \tag{4.42}$$

angeschrieben und später in die notwendige Bedingung für ein Minimum gemäß Satz 4.1 eingesetzt. Wendet man (4.42) für zunächst beliebiges η auf (4.40) an, so erhält man

$$\begin{aligned}&\frac{\mathrm{d}}{\mathrm{d}\eta} J(t_1^* + \eta\xi_{t_1}, \mathbf{x}^* + \eta\boldsymbol{\xi}_x) = \\ &= \left(\frac{\mathrm{d}}{\mathrm{d}\eta}\varphi\right)(t_1^* + \eta\xi_{t_1}, \mathbf{x}^*(t_1^* + \eta\xi_{t_1}) + \eta\boldsymbol{\xi}_x(t_1^* + \eta\xi_{t_1})) \\ &\quad + \frac{\mathrm{d}}{\mathrm{d}\eta} \int_{t_0}^{t_1^* + \eta\xi_{t_1}} l\Big(t, \mathbf{x}^* + \eta\boldsymbol{\xi}_x, \dot{\mathbf{x}}^* + \eta\dot{\boldsymbol{\xi}}_x\Big)\,\mathrm{d}t \\ &= \left[\left(\frac{\partial}{\partial t}\varphi\right)\xi_{t_1} + \left(\frac{\partial}{\partial \mathbf{x}}\varphi\right)\boldsymbol{\xi}_x + \left(\frac{\partial}{\partial \mathbf{x}}\varphi\right)\Big(\underbrace{\frac{\partial}{\partial t}\mathbf{x}}_{\dot{\mathbf{x}}} + \eta\,\underbrace{\frac{\partial}{\partial t}\boldsymbol{\xi}_x}_{\dot{\boldsymbol{\xi}}_x}\Big)\xi_{t_1}\right]_{t = t_1^* + \eta\xi_{t_1},\ \mathbf{x} = \mathbf{x}^* + \eta\boldsymbol{\xi}_x} \\ &\quad + \xi_{t_1}\Big[l\Big(t, \mathbf{x}^* + \eta\boldsymbol{\xi}_x, \dot{\mathbf{x}}^* + \eta\dot{\boldsymbol{\xi}}_x\Big)\Big]_{t = t_1^* + \eta\xi_{t_1}} \\ &\quad + \int_{t_0}^{t_1^* + \eta\xi_{t_1}} \left(\frac{\mathrm{d}}{\mathrm{d}\eta} l\right)\Big(t, \mathbf{x}^* + \eta\boldsymbol{\xi}_x, \dot{\mathbf{x}}^* + \eta\dot{\boldsymbol{\xi}}_x\Big)\,\mathrm{d}t \\ &= \left[\left(\frac{\partial}{\partial t}\varphi\right)\xi_{t_1} + \left(\frac{\partial}{\partial \mathbf{x}}\varphi\right)\boldsymbol{\xi}_x + \left(\frac{\partial}{\partial \mathbf{x}}\varphi\right)\Big(\dot{\mathbf{x}} + \eta\dot{\boldsymbol{\xi}}_x\Big)\xi_{t_1}\right]_{t = t_1^* + \eta\xi_{t_1},\ \mathbf{x} = \mathbf{x}^* + \eta\boldsymbol{\xi}_x} \\ &\quad + \xi_{t_1}\Big[l\Big(t, \mathbf{x}^* + \eta\boldsymbol{\xi}_x, \dot{\mathbf{x}}^* + \eta\dot{\boldsymbol{\xi}}_x\Big)\Big]_{t = t_1^* + \eta\xi_{t_1}} \\ &\quad + \int_{t_0}^{t_1^* + \eta\xi_{t_1}} \left(\left(\frac{\partial}{\partial \mathbf{x}} l\right)\Big(t, \mathbf{x}^* + \eta\boldsymbol{\xi}_x, \dot{\mathbf{x}}^* + \eta\dot{\boldsymbol{\xi}}_x\Big)\right. \\ &\quad \left. - \frac{\mathrm{d}}{\mathrm{d}t}\left(\frac{\partial}{\partial \dot{\mathbf{x}}} l\right)\Big(t, \mathbf{x}^* + \eta\boldsymbol{\xi}_x, \dot{\mathbf{x}}^* + \eta\dot{\boldsymbol{\xi}}_x\Big)\right)\boldsymbol{\xi}_x\,\mathrm{d}t \\ &\quad + \left[\left(\frac{\partial}{\partial \dot{\mathbf{x}}} l\right)\Big(t, \mathbf{x}^* + \eta\boldsymbol{\xi}_x, \dot{\mathbf{x}}^* + \eta\dot{\boldsymbol{\xi}}_x\Big)\boldsymbol{\xi}_x\right]_{t_0}^{t_1^* + \eta\xi_{t_1}}.\end{aligned} \tag{4.43}$$

Wertet man (4.43) für $\eta = 0$ aus, so lautet die notwendige Optimalitätsbedingung

$$\begin{aligned}\delta J(t_1^*, \mathbf{x}^*; \xi_{t_1}, \boldsymbol{\xi}_x) &= \frac{\mathrm{d}}{\mathrm{d}\eta} J(t_1^* + \eta\xi_{t_1}, \mathbf{x}^* + \eta\boldsymbol{\xi}_x)\Big|_{\eta=0} \\ &= \left[\left(\frac{\partial}{\partial t}\varphi\right)\xi_{t_1} + \left(\frac{\partial}{\partial \mathbf{x}}\varphi\right)(\dot{\mathbf{x}}\xi_{t_1} + \boldsymbol{\xi}_x) + \xi_{t_1} l\right]_{t = t_1^*, \mathbf{x} = \mathbf{x}^*}\end{aligned}$$

$$
\begin{aligned}
&+\int_{t_0}^{t_1^*}\left[\left(\frac{\partial}{\partial \mathbf{x}}l\right)(t,\mathbf{x}^*,\dot{\mathbf{x}}^*)-\frac{\mathrm{d}}{\mathrm{d}t}\left(\frac{\partial}{\partial \dot{\mathbf{x}}}l\right)(t,\mathbf{x}^*,\dot{\mathbf{x}}^*)\right]\boldsymbol{\xi}_x\,\mathrm{d}t\\
&+\left[\left(\frac{\partial}{\partial \dot{\mathbf{x}}}l\right)(t,\mathbf{x}^*,\dot{\mathbf{x}}^*)\boldsymbol{\xi}_x\right]_{t_0}^{t_1^*}\\
=&\left[\left(\frac{\partial}{\partial t}\varphi\right)\xi_{t_1}+\left(\frac{\partial}{\partial \mathbf{x}}\varphi\right)(\dot{\mathbf{x}}\xi_{t_1}+\boldsymbol{\xi}_x)+\xi_{t_1}l\right]_{t=t_1^*,\mathbf{x}=\mathbf{x}^*}\\
&+\left[\left(\frac{\partial}{\partial \dot{\mathbf{x}}}l\right)(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[\left(\frac{\partial}{\partial \dot{\mathbf{x}}}l\right)(t,\mathbf{x},\dot{\mathbf{x}})\boldsymbol{\xi}_x\right]_{t=t_0,\mathbf{x}=\mathbf{x}^*}\\
&+\int_{t_0}^{t_1^*}\left[\left(\frac{\partial}{\partial \mathbf{x}}l\right)(t,\mathbf{x}^*,\dot{\mathbf{x}}^*)-\frac{\mathrm{d}}{\mathrm{d}t}\left(\frac{\partial}{\partial \dot{\mathbf{x}}}l\right)(t,\mathbf{x}^*,\dot{\mathbf{x}}^*)\right]\boldsymbol{\xi}_x\,\mathrm{d}t\\
=&\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}^*}\xi_{t_1}\\
&+\left[\frac{\partial}{\partial \dot{\mathbf{x}}}l+\frac{\partial}{\partial \mathbf{x}}\varphi\right]_{t=t_1^*,\ \mathbf{x}=\mathbf{x}^*}(\dot{\mathbf{x}}^*(t_1^*)\xi_{t_1}+\boldsymbol{\xi}_x(t_1^*))-\left[\frac{\partial}{\partial \dot{\mathbf{x}}}l\right]_{t=t_0,\mathbf{x}=\mathbf{x}^*}\underbrace{\boldsymbol{\xi}_x(t_0)}_{=\mathbf{0}}\\
&+\int_{t_0}^{t_1^*}\left[\left(\frac{\partial}{\partial \mathbf{x}}l\right)(t,\mathbf{x}^*,\dot{\mathbf{x}}^*)-\frac{\mathrm{d}}{\mathrm{d}t}\left(\frac{\partial}{\partial \dot{\mathbf{x}}}l\right)(t,\mathbf{x}^*,\dot{\mathbf{x}}^*)\right]\boldsymbol{\xi}_x\,\mathrm{d}t=0
\end{aligned}
\tag{4.44}
$$

für beliebige zulässige Richtungen ξ_{t_1} und $\boldsymbol{\xi}_x$. Gemäß Fundamentallemma der Variationsrechnung Lemma 4.2 folgen aus (4.44) daher zunächst die Euler-Lagrange Gleichungen (4.20). 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. Da die Endzeit t_1 und der Endwert $\mathbf{x}(t_1)$ frei sind, können ξ_{t_1} und $\boldsymbol{\xi}_x(t_1^*)$ unabhängig voneinander frei gewählt werden. Folglich ist (4.44) nur dann Null, wenn die *Transversalitätsbedingungen* (4.41) erfüllt sind. □

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 die drittletzte Zeile von (4.44) verschwindet. Es liegt somit keine Transversalitätsbedingung (4.41b) 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 $\bar{x}_k=$ konst. fest ist*, so folgt daraus $\xi_{x,k}(t_1)=0$, womit automatisch der zugehörige Eintrag in der zweitletzten Zeile von (4.44) 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 (4.44) für diese Komponente

$$
\left[\frac{\partial}{\partial \dot{x}_k}l+\frac{\partial}{\partial x_k}\varphi\right]_{t=t_1^*,\ \mathbf{x}=\mathbf{x}^*}=0\ .
\tag{4.45}
$$

(b) Wenn *die Endzeit* t_1 *frei ist*, dann muss die Transversalitätsbedingung (4.41b)

$$\left[l - \left(\frac{\partial}{\partial \dot{\mathbf{x}}} l\right)\dot{\mathbf{x}} + \frac{\partial}{\partial t}\varphi\right]_{t=t_1^*,\ \mathbf{x}=\mathbf{x}^*} = 0 \tag{4.46}$$

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* $\bar{x}_k$ = konst. *fest ist*, dann muss für diese Komponente eine zulässige Richtung $(\xi_{t_1}, \xi_{x,k})$ die Bedingung

$$\bar{x}_k = x_k^*(t_1^* + \eta\xi_{t_1}) + \eta\xi_{x,k}(t_1^* + \eta\xi_{t_1}) \tag{4.47}$$

bzw.

$$0 = \frac{\mathrm{d}}{\mathrm{d}\eta}\bar{x}_k\bigg|_{\eta=0} = \xi_{x,k}(t_1^*) + \xi_{t_1}\dot{x}_k^*(t_1^*) \tag{4.48}$$

erfüllen. Damit verschwindet der zugehörige Eintrag in der zweitletzten Zeile von (4.44) 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

$$\left[\frac{\partial}{\partial \dot{x}_k} l + \frac{\partial}{\partial x_k}\varphi\right]_{t=t_1^*,\ \mathbf{x}=\mathbf{x}^*} = 0\ . \tag{4.49}$$

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 $[t_0, t_1]$ stetig differenzierbaren (vektorwertigen) Funktionen $\left(C^1[t_0, T]\right)^n$ 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* $t_0 = c_0 < c_1 < \ldots < c_{N+1} = t_1$ mit $N < \infty$ so existiert, dass die Funktion $x(t)$ in allen Intervallen (c_k, c_{k+1}), $k = 0, \ldots, N$ stetig differenzierbar ist, siehe Abbildung 4.3. Die inneren Punkte $c_1, \ldots, c_N$ 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 (4.10)

$$\|\hat{\mathbf{x}}(t)\|_\infty := \max_{t_0 \le t \le t_1} \|\hat{\mathbf{x}}(t)\| \quad \text{und} \quad \|\hat{\mathbf{x}}(t)\|_{1,\infty} := \max_{t_0 \le t \le t_1} \|\hat{\mathbf{x}}(t)\| + \sup_{t \in \cup_{k=0}^{N}(c_k, c_{k+1})} \left\|\frac{\mathrm{d}}{\mathrm{d}t}\hat{\mathbf{x}}(t)\right\| . \tag{4.50}$$

Es gilt nun folgender Satz, welcher z. B. in [4.10] bewiesen wird.

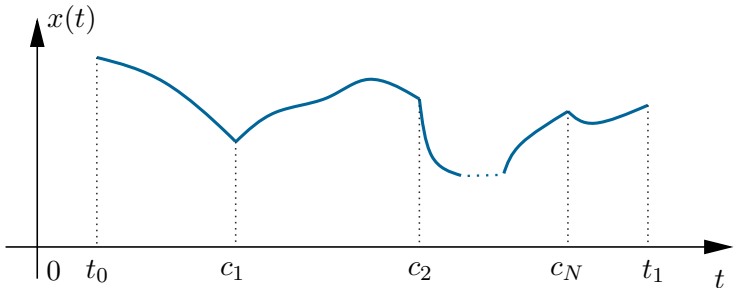


Abbildung 4.3: Beispiel einer Funktion $x(t) \in \hat{C}^1[t_0, t_1]$.

Satz 4.6 (Stückweise stetig vs. stetig differenzierbare Extremale)**.** *Angenommen* $\mathbf{x}^* \in \mathcal{X}$ *ist ein (lokales) Minimum des Funktionals*

$$J(\mathbf{x}) = \int_{t_0}^{t_1} l(t, \mathbf{x}(t), \dot{\mathbf{x}}(t))\,\mathrm{d}t \tag{4.51}$$

mit der zulässigen Menge $\mathcal{X} = \{\mathbf{x}(t) \in (C^1[t_0, t_1])^n \,|\, \mathbf{x}(t_0) = \mathbf{x}_0, \mathbf{x}(t_1) = \mathbf{x}_1\}$ *und der stetig differenzierbaren Lagrangeschen Dichte* $l : \mathbb{R} \times \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}$*, dann ist* $\mathbf{x}^* \in \hat{\mathcal{X}}$ *auch ein (lokales) Minimum des Funktionals* (4.51) *in der zulässigen Menge* $\hat{\mathcal{X}} = \{\mathbf{x}(t) \in (\hat{C}^1[t_0, t_1])^n \,|\, \mathbf{x}(t_0) = \mathbf{x}_0, \mathbf{x}(t_1) = \mathbf{x}_1\}$*. Handelt es sich um ein lokales Minimum, so ist der Begriff* lokal *bezüglich der gleichen Norm* $\|\cdot\|_\infty$ *bzw.* $\|\cdot\|_{1,\infty}$ *zu verstehen.*

Beweisskizze: Zu jedem $\hat{\mathbf{x}} \in \hat{\mathcal{X}}$ und $\varepsilon > 0$ existiert ein $\tilde{\mathbf{x}} \in \mathcal{X}$ so, dass

$$|J(\tilde{\mathbf{x}}) - J(\hat{\mathbf{x}})| < \varepsilon\ . \tag{4.52}$$

Diese plausible Aussage wird z. B. in [4.10] streng bewiesen.

Stellt $\mathbf{x}^* \in \mathcal{X}$ ein *globales* Minimum des Funktionals (4.51) dar, so muss

$$J(\hat{\mathbf{x}}) \geq J(\mathbf{x}^*)\ , \qquad \forall\, \hat{\mathbf{x}} \in \hat{\mathcal{X}} \tag{4.53}$$

gezeigt werden. Es gilt nun für beliebige $\hat{\mathbf{x}} \in \hat{\mathcal{X}}$ und $\varepsilon > 0$ mit $\tilde{\mathbf{x}} \in \mathcal{X}$, welches (4.52) erfüllt,

$$J(\hat{\mathbf{x}}) = J(\tilde{\mathbf{x}}) - (J(\tilde{\mathbf{x}}) - J(\hat{\mathbf{x}})) \geq J(\tilde{\mathbf{x}}) - |J(\tilde{\mathbf{x}}) - J(\hat{\mathbf{x}})| \geq J(\tilde{\mathbf{x}}) - \varepsilon \geq J(\mathbf{x}^*) - \varepsilon\ , \tag{4.54}$$

wobei hier die Optimalität von $\mathbf{x}^*$ im Funktionenraum $\mathcal{X}$ ausgenützt wurde. Im Grenzübergang $\varepsilon \to 0^+$ folgt aus (4.54) der zu zeigende Zusammenhang (4.53).

Stellt $\mathbf{x}^* \in \mathcal{X}$ nur ein *lokales* Minimum (vgl. (4.9)) des Funktionals (4.51) dar, so erfolgt der Beweis völlig analog, wobei $\hat{\mathbf{x}}$ und $\tilde{\mathbf{x}}$ auf das Gebiet $\{\mathbf{x} \in \hat{\mathcal{X}} \,|\, \|\mathbf{x} - \mathbf{x}^*\| < \gamma\}$ einzuschränken sind. □

Die Aussage von Satz 4.6 kann wie folgt genützt werden. Wenn ein lokales Minimum $\mathbf{x}^* \in \mathcal{X}$ entsprechend (4.9) gefunden wurde, so kann auf die Suche nach einem anderen

Minimum in der Umgebung $\{\mathbf{x} \in \hat{\mathcal{X}} \,|\, \|\mathbf{x} - \mathbf{x}^*\| < \gamma\}$ verzichtet werden. Wenn ein globales Minimum $\mathbf{x}^* \in \mathcal{X}$ gefunden wurde, so kann auf die Suche nach einem anderen globalen Minimum im Funktionenraum $\hat{\mathcal{X}}$ verzichtet werden.

Man kann nun zeigen, dass eine extremale Lösung $\hat{\mathbf{x}}^*(t) \in (\hat{C}[t_0, t_1])^n$ im gesamten Intervall $[t_0, t_1]$ außer an den Eckpunkten $c_1, \ldots, c_N$ die Euler-Lagrange Gleichungen (4.20) und die Legendre-Bedingung (4.37) erfüllt. Die Transversalitätsbedingungen (4.45), (4.46) und (4.49) bleiben im Falle stückweise stetig differenzierbarer Extremale *unverändert*. Die Unstetigkeiten von $\frac{\mathrm{d}}{\mathrm{d}t}\hat{\mathbf{x}}^*(t)$ an den Eckpunkten $t = c_k,\ k = 1, \ldots, N$ unterliegen nun folgenden Einschränkungen:

Satz 4.7 (Weierstrass-Erdmann Bedingungen)**.** *Angenommen $\hat{\mathbf{x}}^* \in \hat{\mathcal{X}}$ ist ein (lokales) Minimum des Funktionals*

$$J(\hat{\mathbf{x}}) = \int_{t_0}^{t_1} l\Big(t, \hat{\mathbf{x}}(t), \dot{\hat{\mathbf{x}}}(t)\Big)\,\mathrm{d}t \tag{4.55}$$

mit der zulässigen Menge $\hat{\mathcal{X}} = \{\hat{\mathbf{x}}(t) \in (\hat{C}^1[t_0, t_1])^n \,|\, \hat{\mathbf{x}}(t_0) = \hat{\mathbf{x}}_0, \hat{\mathbf{x}}(t_1) = \hat{\mathbf{x}}_1\}$, 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

$$\left(\frac{\partial}{\partial \dot{\hat{\mathbf{x}}}} l\right)\Big(c, \hat{\mathbf{x}}^*(c), \dot{\hat{\mathbf{x}}}^*(c^-)\Big) = \left(\frac{\partial}{\partial \dot{\hat{\mathbf{x}}}} l\right)\Big(c, \hat{\mathbf{x}}^*(c), \dot{\hat{\mathbf{x}}}^*(c^+)\Big) \tag{4.56a}$$

$$H\Big(c, \hat{\mathbf{x}}^*(c), \dot{\hat{\mathbf{x}}}^*(c^-)\Big) = H\Big(c, \hat{\mathbf{x}}^*(c), \dot{\hat{\mathbf{x}}}^*(c^+)\Big) \tag{4.56b}$$

mit der Hamiltonfunktion

$$H\Big(t, \hat{\mathbf{x}}(t), \dot{\hat{\mathbf{x}}}(t)\Big) = \left(\frac{\partial}{\partial \dot{\hat{\mathbf{x}}}} l\right)\Big(t, \hat{\mathbf{x}}(t), \dot{\hat{\mathbf{x}}}(t)\Big)\dot{\hat{\mathbf{x}}}(t) - l\Big(t, \hat{\mathbf{x}}(t), \dot{\hat{\mathbf{x}}}(t)\Big) \tag{4.57}$$

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 [4.9].

Beispiel 4.5. Gesucht ist ein (lokales) Minimum $x^* \in \mathcal{X}$ des Funktionals

$$J(x) = \int_{-1}^{1} x^2(t)(1 - \dot{x}(t))^2\,\mathrm{d}t \tag{4.58}$$

in der zulässigen Menge $\mathcal{X} = \{x(t) \in C^1[-1, 1] \,|\, x(-1) = 0, x(1) = 1\}$. Da die Lagrangesche Dichte nicht explizit von der Zeit t abhängt und wegen der Randbedingung

$x(-1) = 0$, folgt für die Hamiltonfunktion

$$H = \left(\frac{\partial}{\partial \dot{x}} l\right)\dot{x} - l = -2x^2(1-\dot{x})\dot{x} - x^2(1-\dot{x})^2 = x^2(\dot{x}^2 - 1) = 0 \tag{4.59}$$

für alle Zeiten $t \in [-1, 1]$. Die Hamiltonfunktion ist also konstant und damit eine Invariante des Systems, siehe auch (4.27). Mit den Substitutionen $x^2(t) = z(t)$ und $2x(t)\dot{x}(t) = \dot{z}(t)$ kann (4.59) in die Form

$$z(t) - \frac{1}{4}\dot{z}^2(t) = 0 \tag{4.60}$$

umgeschrieben werden. Die Lösung von (4.60) lautet

$$z(t) = (t + k)^2 \tag{4.61}$$

mit der Konstanten k. Mit der Randbedinung $x(1) = 1$ und daher $z(1) = 1$ folgt für die Konstante k der Wert $k = 0$ und die mögliche stationäre Lösung $\bar{x}(t)$ des Kostenfunktionals (4.58) lautet

$$\bar{x}(t) = \pm t \ . \tag{4.62}$$

Da $\bar{x}(t)$ die Randbedingung $x(-1) = 0$ nicht erfüllt, ist $\bar{x}(t)$ keine stationäre Lösung von (4.58) in der zulässigen Menge $\mathcal{X} = \{x(t) \in C^1[-1, 1] \,|\, x(-1) = 0,\ x(1) = 1\}$.

Im nächsten Schritt soll daher das Kostenfunktional (4.58) in der zulässigen Menge $\hat{\mathcal{X}} = \{\hat{x}(t) \in \hat{C}^1[-1, 1] \,|\, \hat{x}(-1) = 0,\ \hat{x}(1) = 1\}$ minimiert werden. Die Weierstrass-Erdmann Bedingung (4.56a) besagt, dass an einem Eckpunkt $c \in (-1, 1)$ gilt

$$-2\hat{x}^2(c)\Big(1 - \dot{\hat{x}}(c^-)\Big) = -2\hat{x}^2(c)\Big(1 - \dot{\hat{x}}(c^+)\Big) \tag{4.63}$$

und folglich

$$\hat{x}^2(c)\Big(\dot{\hat{x}}(c^+) - \dot{\hat{x}}(c^-)\Big) = 0 \ . \tag{4.64}$$

Da an einem Eckpunkt $t = c$ gilt $\dot{\hat{x}}(c^+) \neq \dot{\hat{x}}(c^-)$, muss zur Erfüllung von (4.64) die Bedingung $\hat{x}(c) = 0$ eingehalten werden. D. h. eine Unstetigkeit in $\dot{\hat{x}}(t)$ kann nur an Stellen auftreten, an denen der Wert von $\hat{x}(t)$ selbst identisch Null ist. Aus (4.59) folgt, dass zu jedem Zeitpunkt $t \in [-1, 1]$ $x(t) = 0$ oder $\dot{x}(t) = \pm 1$ gelten muss. Außerdem sind die Randbedingungen $\hat{x}(-1) = 0$ und $\hat{x}(1) = 1$ zu erfüllen. Der minimal mögliche Wert des Kostenfunktionals (4.58) ist 0. Er wird erreicht, wenn $\hat{x}(t) = 0$ oder $\dot{\hat{x}}(t) = 1$ für alle t in $[-1, 1]$ gilt. Aus diesen Überlegungen folgt, dass für die optimale Lösung $\hat{x}(t) = 0\ \forall t \in [-1, c]$ und $\dot{\hat{x}}(t) = 1$ sowie $x(t) = t\ \forall t \in (c, 1]$ gelten müssen. Aus der Stetigkeitsbedingung $\hat{x}(c) = 0 = c$ folgen der Umschaltzeitpunkt $c = 0$ und die optimale Lösung

$$\hat{x}^*(t) = \begin{cases} 0 & \text{für } -1 \leq t \leq 0 \\ t & \text{für } 0 < t \leq 1 \ . \end{cases} \tag{4.65}$$

Diese Lösung ist das eindeutige globale Minimum des Kostenfunktionals (4.58) in der zulässigen Menge $\hat{\mathcal{X}} = \{\hat{x}(t) \in \hat{C}^1[-1, 1] \,|\, \hat{x}(-1) = 0,\ \hat{x}(1) = 1\}$.

4.2 Entwurf von Optimalsteuerungen

4.2.1 Problemformulierung

Eine typische Optimalsteuerungsaufgabe besteht darin, für ein dynamisches System beschrieben durch die Differenzialgleichungen

$$\dot{\mathbf{x}} = \mathbf{f}(t, \mathbf{x}(t), \mathbf{u}(t)) \tag{4.66}$$

mit der Zeit $t \in \mathbb{R}$, dem *Zustand* $\mathbf{x} \in \mathbb{R}^n$, dem *Anfangszustand* $\mathbf{x}(t_0) = \mathbf{x}_0$ 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* $J(\mathbf{u})$ minimiert wird und dabei allfällige Beschränkungen für $\mathbf{x}(t)$ und $\mathbf{u}(t)$ eingehalten werden. Beispielhaft kann eine Optimalsteuerungsaufgabe in der Form

$$\begin{aligned}
\min_{\mathbf{u}(\cdot)} \quad & J(\mathbf{u}) = \int_{t_0}^{t_1} l(t, \mathbf{x}(t), \mathbf{u}(t))\,\mathrm{d}t && \text{Kostenfunktional} && \text{(4.67a)}\\
\text{u.B.v.} \quad & \dot{\mathbf{x}} = \mathbf{f}(t, \mathbf{x}(t), \mathbf{u}(t)) && \text{Beschränkungen} && \text{(4.67b)}\\
& \mathbf{x}(t_0) = \mathbf{x}_0 && && \text{(4.67c)}\\
& \boldsymbol{\psi}_1(t_1, \mathbf{x}(t_1)) = \mathbf{0} && && \text{(4.67d)}\\
& \boldsymbol{\psi}_2(t, \mathbf{x}(t), \mathbf{u}(t)) \leq \mathbf{0} \quad \forall\, t \in I \subseteq [t_0, t_1] && && \text{(4.67e)}
\end{aligned}$$

angeschrieben werden. Die sogenannte *Bolza-Form* des Kostenfunktionals (siehe auch (4.3)) lautet

$$J(\mathbf{u}) = \varphi(t_0, \mathbf{x}(t_0), t_1, \mathbf{x}(t_1)) + \int_{t_0}^{t_1} l(t, \mathbf{x}(t), \mathbf{u}(t))\,\mathrm{d}t\ . \tag{4.68}$$

Für die *Lagrange-Form* gilt

$$J(\mathbf{u}) = \int_{t_0}^{t_1} l(t, \mathbf{x}(t), \mathbf{u}(t))\,\mathrm{d}t \tag{4.69}$$

und für die *Mayer-Form*

$$J(\mathbf{u}) = \varphi(t_0, \mathbf{x}(t_0), t_1, \mathbf{x}(t_1))\ . \tag{4.70}$$

Die Abhängigkeit der Funktion φ von t_0 und $\mathbf{x}(t_0)$ ist nur dann relevant, wenn die Anfangsbedingung (4.67c) nicht vorhanden ist.

Aufgabe 4.3. Zeigen Sie, dass die Lagrange-Form in die Mayer-Form übergeführt werden kann, indem man einen zusätzlichen Zustand

$$\dot{x}_{n+1} = l(t, \mathbf{x}, \mathbf{u}), \quad x_{n+1}(t_0) = 0 \tag{4.71}$$

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

$$\dot{x}_{n+1} = 0, \quad x_{n+1}(t_0) = \frac{1}{t_1 - t_0}\varphi(t_0, \mathbf{x}(t_0), t_1, \mathbf{x}(t_1)) \tag{4.72}$$

einführt und das Kostenfunktional in der Form $J(\mathbf{u}) = \int_{t_0}^{t_1} x_{n+1}(t)\,\mathrm{d}t$ anschreibt.

Zeigen Sie, wie man eine Bolza-Form in die Mayer- oder Lagrange-Form überführt.

Bei den möglichen Beschränkungen unterscheidet man wieder zwischen *Punktbeschränkungen*, beispielsweise Endpunktbeschränkungen der Form

$$\boldsymbol{\psi}(t_1, \mathbf{x}(t_1)) = \mathbf{0} \quad \text{bzw.} \quad \boldsymbol{\psi}(t_1, \mathbf{x}(t_1)) \leq \mathbf{0}\ , \tag{4.73}$$

Pfadbeschränkungen

$$\psi(t, \mathbf{x}(t), \mathbf{u}(t)) = 0 \quad \text{bzw.} \quad \psi(t, \mathbf{x}(t), \mathbf{u}(t)) \leq 0, \quad \forall\, t \in I \subseteq [t_0, t_1] \tag{4.74}$$

und *isoperimetrischen Beschränkungen*

$$\int_{t_0}^{t_1} \psi(t, \mathbf{x}(t), \mathbf{u}(t))\,\mathrm{d}t = 0 \quad \text{bzw.} \quad \int_{t_0}^{t_1} \psi(t, \mathbf{x}(t), \mathbf{u}(t))\,\mathrm{d}t \leq 0\ . \tag{4.75}$$

Häufig sind Pfadbeschränkungen (4.74) schwieriger zu berücksichtigen, wenn sie nicht von der Stellgröße abhängen. Isoperimetrische Beschränkungen können durch Einführung eines zusätzlichen Zustandes $x_{n+1}(t)$ mit $\dot{x}_{n+1}(t) = \psi(t, \mathbf{x}(t), \mathbf{u}(t))$ und $x_{n+1}(t_0) = 0$ durch Endpunktbeschränkungen ersetzt werden.

4.2.2 Existenz und Eindeutigkeit einer Lösung

4.2.2.1 Existenz und Eindeutigkeit einer Lösung eines Anfangswertproblems

Im Skriptum [4.11] werden hinreichende Bedingungen für die lokale Existenz und Eindeutigkeit der Lösung eines *Anfangswertproblems* angegeben. Sie besagen, wenn eine Funktion $\mathbf{g}(t, \mathbf{x})$ stückweise stetig in t ist und der Lipschitz-Bedingung

$$\|\mathbf{g}(t, \mathbf{x}) - \mathbf{g}(t, \mathbf{y})\| \leq L\|\mathbf{x} - \mathbf{y}\|, \quad 0 < L < \infty \tag{4.76}$$

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

$$\dot{\mathbf{x}} = \mathbf{g}(t, \mathbf{x}), \quad \mathbf{x}(t_0) = \mathbf{x}_0 \tag{4.77}$$

für $t \in [t_0, t_0 + \delta]$ *genau eine Lösung* besitzt. Wie im Skriptum [4.11] beschrieben, ist 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 (4.76) lokal erfüllt.

Da für die obige Existenzaussage nur stückweise Stetigkeit von $\mathbf{g}(t, \mathbf{x})$ in t erforderlich ist, sind mit $\mathbf{g}(t, \mathbf{x}) := \mathbf{f}(t, \mathbf{x}(t), \mathbf{u}(t))$ entsprechend (4.66) 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* $t_0 = c_0 < c_1 < \ldots < c_{N+1} = t_1$ mit $N < \infty$ so existiert, dass die Funktion $u(t)$ in allen Intervallen (c_k, c_{k+1}), $k = 0, \ldots, N$ stetig ist. Für stückweise stetige Stellgrößen $\mathbf{u}(t)$ sind die zugehörigen Zustandsgrößen von (4.66) 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.

4.2.2.2 Existenz und Eindeutigkeit einer Lösung eines Optimalsteuerungsproblems

Die Frage, ob für ein Optimalsteuerungsproblem eine optimale Lösung existiert und sogar eindeutig ist, ist wesentlich schwieriger zu beantworten als die Frage nach der Existenz und Eindeutigkeit einer Lösung eines Anfangswertproblems.

Wie in Beispiel 4.3 gezeigt wurde, folgt aus der Existenz einer extremalen Lösung, welche z. B. durch Lösung von Euler-Lagrange Gleichungen (vgl. Satz 4.2) nachgewiesen werden kann, nicht automatisch die Existenz einer (lokal) optimalen Lösung. Damit aber eine optimale Lösung existieren kann, muss zumindest eine extremale Lösung existieren.

Existiert eine optimale Lösung und sind die extremalen Lösungen eindeutig, so ist auch die optimale Lösung eindeutig. Umgekehrt jedoch folgt aus der Eindeutigkeit einer optimalen Lösung nicht die Eindeutigkeit von extremalen Lösungen. Weitere Erläuterungen zu diesen Aussagen finden sich in [4.12].

In vielen Optimalsteuerungsproblemen unterliegen die Stellgrößen $\mathbf{u}(t)$ gewissen Beschränkungen, d. h. $\mathbf{u}(t) \in U \subset \mathbb{R}^m$. Eine stückweise stetige Stellgröße $\mathbf{u}(t)$ im Intervall $t_0 \leq t \leq t_1$ mit $\mathbf{u}(t) \in U$ für alle $t \in [t_0, t_1]$ wird im Weiteren als *zulässige Stellgröße* bezeichnet.

Eine zulässige Stellgröße wird als *realisierbar* bezeichnet, wenn die zugehörige Zustandstrajektorie $\mathbf{x}(t)$ von (4.66) im gesamten Intervall $t_0 \leq t \leq t_1$ eindeutig definiert ist (vgl. Abschnitt 4.2.2.1) und sämtliche Beschränkungen erfüllt. Wie die nachfolgenden Beispiele verdeutlichen, impliziert die Existenz einer realisierbaren Stellgröße nicht die Existenz einer optimalen Lösung eines Optimalsteuerungsproblems. Umgekehrt jedoch ist die Existenz einer realisierbaren Stellgröße natürlich eine Voraussetzung für die Existenz einer optimalen Lösung.

Beispiel 4.6. Es soll für das dynamische System $\dot{x} = u$ die Stellgröße $u(t) \geq 0$ so gewählt werden, dass der Zustand in minimaler Zeit vom Anfangszustand $x(t_0) = 0$ in den Endzustand $x(t_1) = 1$ (Endzeit t_1 ist frei) übergeführt wird, d. h. das Kostenfunktional

$$J(u) = t_1 - t_0 \tag{4.78}$$

ist zu minimieren. Es existiert keine realisierbare Stellgröße so, dass $J(u) = t_1 - t_0 = 0$ gilt. Die konstante Stellgröße

$$u(t) = \frac{1}{t_1 - t_0} \tag{4.79}$$

mit $t_1 - t_0 > 0$ ist realisierbar. Es existiert jedoch keine optimale Lösung, da zu jedem Wert $1/(t_1 - t_0)$ eine noch größere finite Stellgröße $u(t)$ gefunden werden kann, die $J(u) = t_1 - t_0 > 0$ weiter verkleinert.

Beispiel 4.7. Es sollen für das dynamische System $\dot{x} = u$ die Stellgröße $u(t) \in [0, 1]$ und die freie Endzeit $t_1 \geq t_0$ so gewählt werden, dass der Zustand vom Anfangszustand $x(t_0) = 0$ in den Endzustand $x(t_1) = 1$ übergeführt wird und das Kostenfunktional

$$J(u) = \int_{t_0}^{t_1} u^2(t)\,\mathrm{d}t \tag{4.80}$$

minimiert wird. Die konstante Stellgröße $u(t) = 0$ ist nicht realisierbar, d. h. für jede realisierbare Stellgröße muss $J(u) > 0$ gelten. Die konstante Stellgröße

$$u(t) = \frac{1}{t_1 - t_0} \tag{4.81}$$

mit $t_1 - t_0 \geq 1$ ist realisierbar und liefert für das Kostenfunktional den Wert

$$J(u) = \int_{t_0}^{t_1} \left(\frac{1}{t_1 - t_0} \right)^2 \mathrm{d}t = \frac{1}{t_1 - t_0} \ . \tag{4.82}$$

Es existiert jedoch keine optimale Lösung, da zu jedem Wert $u(t)$ eine noch kleinere von Null verschiedene Stellgröße gefunden werden kann, die $J(u) = 1/(t_1 - t_0) > 0$ mit $t_1 < \infty$ weiter verkleinert.

In beiden Beispielen ist die Menge der realisierbaren Lösungen *unbeschränkt* und daher *nicht kompakt*. Folglich kann der Satz von Weierstrass Satz 1.1 nicht angewandt werden. In Beispiel 4.6 ist die Stellgröße $u(t)$ selbst unbeschränkt. In Beispiel 4.7 kann die zu optimierende Größe t_1 beliebig hohe Werte annehmen, weshalb auch hier die Menge der realisierbaren Lösungen unbeschränkt (nicht kompakt) ist. Um letzteres Problem zu verhindern, kann bei Optimalsteuerungsproblemen mit freier Endzeit t_1 die Einschränkung $t_0 \leq t_1 \leq T$ mit einem hinreichend großen, festen Wert T verwendet werden.

In den beiden obigen Beispielen ist unmittelbar einsichtig, warum die Menge der realisierbaren Lösungen nicht kompakt ist. Schwieriger kann das Feststellen der Nichtkompaktheit der Menge der realisierbaren Stellgrößen sein, wenn sie mit einer Zustandstrajektorie von (4.66), die eine Beschränkung verletzt oder nicht definiert ist, im Zusammenhang steht. Exemplarisch dafür ist das nachfolgende Beispiel. Hier ist die Menge der realisierbaren Stellgrößen *nicht abgeschlossen* und daher *nicht kompakt*, da am Rand dieser Menge die zugehörige Zustandstrajektorie nicht mehr definiert ist. In diesem Beispiel existiert keine optimale Stellgröße.

Beispiel 4.8. Es soll für das dynamische System $\dot{x} = (1+x)^2 u$ mit dem Anfangszustand $x(0) = 0$ die Stellgröße $u(t) \in [0, 1]$ so gewählt werden, dass das Kostenfunktional

$$J(u) = \int_0^1 \frac{1}{1 + x(t)} \, \mathrm{d}t \tag{4.83}$$

minimiert wird. Der Endzustand $x(1)$ ist frei.

Da hier $x(t) \geq 0$ gilt, ist der Integrand in (4.83) stets nichtnegativ und finit. Um $J(u)$ zu minimieren, ist die Stellgröße $u(t)$ so zu wählen, dass $x(t)$ schnellstmöglich wächst. Die extremale Stellgröße $u(t) = 1$ würde das schnellste Wachstum von $x(t)$ hervorrufen, ist aber nicht realisierbar, da die zugehörige Lösung $x(t) = t/(1 - t)$ für $t = 1$ nicht definiert ist. In diesem Fall würde sich der Wert $J(u) = 1/2$ ergeben. Alternativ kann die konstante realisierbare Stellgröße $u(t) = \alpha$ mit $0 \leq \alpha < 1$ gewählt werden, welche $x(t) = \alpha t/(1 - \alpha t)$ und $J(u) = 1 - \alpha/2 > 1/2$ liefert. Es existiert in diesem Fall also keine optimale Lösung, da zu jeder realisierbaren Stellgröße $u(t)$ eine noch größere realisierbare Stellgröße gefunden werden kann, die $J(u) > 1/2$ weiter

verkleinert.

In diesem Beispiel strebt die optimale Zustandstrajektorie im betrachteten endlichen Zeitintervall $[t_0, t_1]$ gegen Unendlich. Im Englischen wird dieses Verhalten als *finite escape time* bezeichnet. Um derartige Fälle zu vermeiden, kann eine zusätzliche Pfadbeschränkung der Art

$$\|\mathbf{x}(t)\| \leq \alpha, \quad \forall\, t \in [t_0, t_1] \tag{4.84}$$

mit einem endlichen Wert $\alpha > 0$ verwendet werden [4.13]. Diese Beschränkung wird z. B. von dynamischen Systemen erfüllt, die einer der beiden Ungleichungen

$$|\mathbf{f}(t, \mathbf{x}, \mathbf{u}(t))| \leq \beta \|\mathbf{x}\|_1 + \gamma \tag{4.85a}$$

$$|\mathbf{x}^{\mathrm{T}} \mathbf{f}(t, \mathbf{x}, \mathbf{u}(t))| \leq \beta \|\mathbf{x}\|_2^2 + \gamma \tag{4.85b}$$

mit nichtnegativen Konstanten β und γ für alle $t \in [t_0, t_1]$, alle zulässigen $\mathbf{u}(t)$ und alle $\mathbf{x} \in \mathbb{R}^n$ genügen. Dies gilt z. B. für Systeme, die in $\mathbf{x}$ affin sind, d. h. die Struktur $\dot{\mathbf{x}} = \mathbf{A}(t, \mathbf{u})\mathbf{x} + \mathbf{b}(t, \mathbf{u})$ aufweisen.

Bislang wurde anhand von Negativbeispielen verdeutlicht, dass die Frage nach der Existenz einer Lösung eines Optimalsteuerungsproblems keineswegs einfach zu beantworten ist. Nachfolgend sollen Möglichkeiten skizziert werden, um diese Frage positiv zu beantworten.

Eine Methode zum Nachweis der Existenz einer optimalen Lösung eines Optimalsteuerungsproblems ist die folgende: Findet man eine untere Schranke $\underline{J} > -\infty$ so, dass $J(\mathbf{u}) \geq \underline{J}$ für alle realisierbaren Stellgrößen $\mathbf{u}$ gelten muss, so ist jede realisierbare Stellgröße $\mathbf{u}^*$, welche $J(\mathbf{u}^*) = \underline{J}$ liefert, optimal. Diese Methode wird im nachfolgenden Beispiel verwendet.

Beispiel 4.9. Es soll für das dynamische System $\dot{x} = 1 - (u_1^2 + u_2^2)$ mit dem Anfangszustand $x(0) = 0$ die Stellgröße $\mathbf{u}(t) \in \mathbb{R}^2$ so gewählt werden, dass das Kostenfunktional

$$J(u) = \int_0^1 x^2(t)\, \mathrm{d}t \tag{4.86}$$

minimiert wird. Der Endzustand $x(1)$ ist frei.

Der Wert $\underline{J} = 0$ ist eine unter Schranke für $J(\mathbf{u})$, da der Integrand von (4.86) stets nichtnegativ ist. Jede realisierbare Stellgröße $\mathbf{u}^*(t)$, die $\|\mathbf{u}^*(t)\|_2^2 = 1\ \forall\, t \in [0, 1]$ erfüllt, also z. B.

$$\mathbf{u}(t) = \begin{bmatrix} \sin(\omega t + \phi) \\ \cos(\omega t + \phi) \end{bmatrix} \tag{4.87}$$

mit beliebigen Werten ω und ϕ, ist optimal, da sie $x^*(t) = 0$ und $J(\mathbf{u}^*) = 0$ liefert.

Dieses Beispiel zeigt, dass die Existenz einer optimalen Lösung (Stellgröße) nicht deren Eindeutigkeit impliziert. Auch die hier gegebene Eindeutigkeit der optimalen Zustandstrajektorie $x^*(t) = 0$ ändert daran nichts.

Um die Existenz einer Lösung eines Optimalsteuerungsproblems sicherzustellen, können zwei im nachfolgenden Satz beschriebene Wege beschritten werden: Die zulässigen Stellgrößen können auf spezielle Mengen eingeschränkt werden oder von $\mathbf{f}(t, \mathbf{x}(t), \mathbf{u}(t))$ und

$l(t, \mathbf{x}(t), \mathbf{u}(t))$ werden gewisse Konvexitätseigenschaften gefordert. Beweise zum nachfolgenden Satz finden sich z. B. in [4.13, 4.14].

Satz 4.8 (Existenz einer optimalen Lösung)**.** *Werden die nachfolgenden Bedingungen erfüllt, so besitzt ein Optimalsteuerungsproblem mit dem Kostenfunktional (4.68) eine optimale Lösung.*

- *U ist eine kompakte Menge.*
- *Die Zustandstrajektorien von (4.66) genügen der Bedingung (4.84), d. h. sie sind beschränkt.*
- *Der Endzustand $\mathbf{x}(t_1)$ wird basierend auf (4.73) auf eine abgeschlossene Menge beschränkt.*
- *Die Funktionen $\mathbf{f}(t, \mathbf{x}(t), \mathbf{u}(t))$, $l(t, \mathbf{x}(t), \mathbf{u}(t))$ und $\varphi(t_0, \mathbf{x}(t_0), t_1, \mathbf{x}(t_1))$ sind stetig.*
- *Die Menge der realisierbaren Stellgrößen ist nichtleer.*
- *Es wird zumindest eine der folgenden Bedingungen erfüllt.*
  - *Es werden nur Stellgrößen $\mathbf{u}(t) \in U$ zugelassen, die die Lipschitz-Bedingung*

$$\|\mathbf{u}(t) - \mathbf{u}(s)\| \leq L_u |t - s|, \quad 0 < L_u < \infty, \quad \forall s, t \in [t_0, t_1] \tag{4.88}$$

    erfüllen.
  - *Es werden nur Stellgrößen $\mathbf{u}(t) \in U$ zugelassen, die stückweise konstant sind mit einer finiten Anzahl an Unstetigkeitsstellen.*
  - *Die Menge $\{[\mathbf{f}^{\mathrm{T}}(t, \mathbf{x}(t), \mathbf{v}) \quad l(t, \mathbf{x}(t), \mathbf{v})]^{\mathrm{T}} \,|\, \mathbf{v} \in U\} \in \mathbb{R}^{n+1}$ ist für feste Werte t und $\mathbf{x}(t)$ konvex.*

Weiterführende Aussagen zur Existenz und Eindeutigkeit von Lösungen eines Optimalsteuerungsproblems finden sich z. B. in [4.12–4.16].

4.2.3 Variationsformulierung

Im Folgenden werden die notwendigen Optimalitätsbedingungen erster Ordnung für ein Optimalsteuerungsproblem mit fester Endzeit und freiem Endwert formuliert.

Satz 4.9 (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)*

$$J(\mathbf{u}) = \varphi(\mathbf{x}(t_1)) + \int_{t_0}^{t_1} l(t, \mathbf{x}(t), \mathbf{u}(t)) \, \mathrm{d}t \tag{4.89}$$

unter der Gleichungsbeschränkung (dynamisches System)

$$\dot{\mathbf{x}} = \mathbf{f}(t, \mathbf{x}(t), \mathbf{u}(t)), \quad \mathbf{x}(t_0) = \mathbf{x}_0 \tag{4.90}$$

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 $\varphi(\mathbf{x}_1)$ 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 $\mathbf{x}^*(t) \in \left(C^1[t_0, t_1]\right)^n$ die zugehörige Lösung des Anfangswertproblems (4.90) ist, dann existiert ein $\boldsymbol{\lambda}^*(t) \in \left(C^1[t_0, t_1]\right)^n$ so, dass gilt

$$\dot{\mathbf{x}}^* = \mathbf{f}(t, \mathbf{x}^*(t), \mathbf{u}^*(t)), \quad \mathbf{x}^*(t_0) = \mathbf{x}_0 \tag{4.91a}$$

$$\begin{aligned} \dot{\boldsymbol{\lambda}}^* &= -\left(\frac{\partial}{\partial \mathbf{x}} l\right)^{\mathrm{T}}(t, \mathbf{x}^*(t), \mathbf{u}^*(t)) - \left(\frac{\partial}{\partial \mathbf{x}} \mathbf{f}\right)^{\mathrm{T}}(t, \mathbf{x}^*(t), \mathbf{u}^*(t))\boldsymbol{\lambda}^*(t) \\ \boldsymbol{\lambda}^*(t_1) &= \left(\frac{\mathrm{d}}{\mathrm{d}\mathbf{x}_1}\varphi\right)^{\mathrm{T}}(\mathbf{x}^*(t_1)) \end{aligned} \tag{4.91b}$$

$$\mathbf{0} = \left(\frac{\partial}{\partial \mathbf{u}} l\right)^{\mathrm{T}}(t, \mathbf{x}^*(t), \mathbf{u}^*(t)) + \left(\frac{\partial}{\partial \mathbf{u}} \mathbf{f}\right)^{\mathrm{T}}(t, \mathbf{x}^*(t), \mathbf{u}^*(t))\boldsymbol{\lambda}^*(t) \tag{4.91c}$$

für $t_0 \leq t \leq t_1$. Die Gleichungen (4.91) 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 (4.90) 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

$$\delta\mathbf{x}(\mathbf{u}; \boldsymbol{\xi}_u)|_{t_0} = \mathbf{0} \ . \tag{4.92a}$$

Die Gâteaux Ableitung der Differenzialgleichung $\dot{\mathbf{x}} = \mathbf{f}(t, \mathbf{x}(t), \mathbf{u}(t))$ lautet

$$\delta\dot{\mathbf{x}}(\mathbf{u}; \boldsymbol{\xi}_u)|_t = \frac{\partial}{\partial \mathbf{x}}\mathbf{f}(t, \mathbf{x}(t), \mathbf{u}(t))\delta\mathbf{x}(\mathbf{u}; \boldsymbol{\xi}_u)|_t + \frac{\partial}{\partial \mathbf{u}}\mathbf{f}(t, \mathbf{x}(t), \mathbf{u}(t))\boldsymbol{\xi}_u(t) \ . \tag{4.92b}$$

Um aus dem linearen (zeitvarianten) Anfangswertproblem (4.92) 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

$$\boldsymbol{\Phi}(\tau, t) \tag{4.93}$$

für das Zeitintervall $[t, \tau]$. Mit dieser Transitionsmatrix folgt die Lösung von (4.92) in der Form

$$\delta\mathbf{x}(\mathbf{u}; \boldsymbol{\xi}_u)|_\tau = \int_{t_0}^{\tau} \boldsymbol{\Phi}(\tau, t)\frac{\partial}{\partial \mathbf{u}}\mathbf{f}(t, \mathbf{x}(t), \mathbf{u}(t))\boldsymbol{\xi}_u(t)\,\mathrm{d}t \ . \tag{4.94}$$

Für die gesuchte Gâteaux Ableitung $\delta J(\mathbf{u};\boldsymbol{\xi}_u)$ an einem allgemeinen Punkt $\mathbf{u} \in (C[t_0,t_1])^m$ in Richtung $\boldsymbol{\xi}_u$ ergibt sich daher unter Verwendung der Kettenregel

$$\begin{aligned}
&\delta J(\mathbf{u};\boldsymbol{\xi}_u) \\
&\quad= \frac{\mathrm{d}}{\mathrm{d}\mathbf{x}_1}\varphi(\mathbf{x}(t_1))\delta\mathbf{x}(\mathbf{u};\boldsymbol{\xi}_u)|_{t_1} \\
&\qquad+ \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)\,\mathrm{d}\tau \\
&\quad= \frac{\mathrm{d}}{\mathrm{d}\mathbf{x}_1}\varphi(\mathbf{x}(t_1)) \int_{t_0}^{t_1} \boldsymbol{\Phi}(t_1,t)\frac{\partial}{\partial\mathbf{u}}\mathbf{f}(t,\mathbf{x}(t),\mathbf{u}(t))\boldsymbol{\xi}_u(t)\,\mathrm{d}t \\
&\qquad+ \int_{t_0}^{t_1} \Bigg(\frac{\partial}{\partial\mathbf{x}} l(\tau,\mathbf{x}(\tau),\mathbf{u}(\tau)) \int_{t_0}^{\tau} \boldsymbol{\Phi}(\tau,t)\frac{\partial}{\partial\mathbf{u}}\mathbf{f}(t,\mathbf{x}(t),\mathbf{u}(t))\boldsymbol{\xi}_u(t)\,\mathrm{d}t \\
&\qquad\qquad+ \frac{\partial}{\partial\mathbf{u}} l(\tau,\mathbf{x}(\tau),\mathbf{u}(\tau))\boldsymbol{\xi}_u(\tau)\Bigg)\,\mathrm{d}\tau \qquad (4.95) \\
&\quad= \int_{t_0}^{t_1} \frac{\mathrm{d}}{\mathrm{d}\mathbf{x}_1}\varphi(\mathbf{x}(t_1))\boldsymbol{\Phi}(t_1,t)\frac{\partial}{\partial\mathbf{u}}\mathbf{f}(t,\mathbf{x}(t),\mathbf{u}(t))\boldsymbol{\xi}_u(t)\,\mathrm{d}t \\
&\qquad+ \int_{t_0}^{t_1}\int_{t}^{t_1} \frac{\partial}{\partial\mathbf{x}} l(\tau,\mathbf{x}(\tau),\mathbf{u}(\tau))\boldsymbol{\Phi}(\tau,t)\frac{\partial}{\partial\mathbf{u}}\mathbf{f}(t,\mathbf{x}(t),\mathbf{u}(t))\boldsymbol{\xi}_u(t)\,\mathrm{d}\tau\,\mathrm{d}t \\
&\qquad+ \int_{t_0}^{t_1} \frac{\partial}{\partial\mathbf{u}} l(t,\mathbf{x}(t),\mathbf{u}(t))\boldsymbol{\xi}_u(t)\,\mathrm{d}t \\
&\quad= \int_{t_0}^{t_1} \Bigg[\frac{\mathrm{d}}{\mathrm{d}\mathbf{x}_1}\varphi(\mathbf{x}(t_1))\boldsymbol{\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)) \\
&\qquad\qquad+ \int_{t}^{t_1} \frac{\partial}{\partial\mathbf{x}} l(\tau,\mathbf{x}(\tau),\mathbf{u}(\tau))\boldsymbol{\Phi}(\tau,t)\,\mathrm{d}\tau\frac{\partial}{\partial\mathbf{u}}\mathbf{f}(t,\mathbf{x}(t),\mathbf{u}(t))\Bigg]\boldsymbol{\xi}_u(t)\,\mathrm{d}t\ .
\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 (4.95)

$$\begin{aligned}
&\frac{\mathrm{d}}{\mathrm{d}\mathbf{x}_1}\varphi(\mathbf{x}^*(t_1))\boldsymbol{\Phi}^*(t_1,t)\frac{\partial}{\partial\mathbf{u}}\mathbf{f}(t,\mathbf{x}^*(t),\mathbf{u}^*(t)) + \frac{\partial}{\partial\mathbf{u}} l(t,\mathbf{x}^*(t),\mathbf{u}^*(t)) \\
&+ \int_{t}^{t_1} \frac{\partial}{\partial\mathbf{x}} l(\tau,\mathbf{x}^*(\tau),\mathbf{u}^*(\tau))\boldsymbol{\Phi}^*(\tau,t)\,\mathrm{d}\tau\frac{\partial}{\partial\mathbf{u}}\mathbf{f}(t,\mathbf{x}^*(t),\mathbf{u}^*(t)) = \mathbf{0}
\end{aligned} \qquad (4.96)$$

für alle $t \in [t_0,t_1]$. Die Transitionsmatrix (4.93) ist auch für das lineare Anfangswertproblem (4.91b) anwendbar. In diesem Fall lautet die zur Dynamikmatrix $-\left(\frac{\partial}{\partial\mathbf{x}}\mathbf{f}\right)^{\mathrm{T}}(t,\mathbf{x}^*(t),\mathbf{u}^*(t))$ und zum Zeitintervall $[\tau,t]$ gehörige Transitionsmatrix $\boldsymbol{\Phi}^{*\mathrm{T}}(\tau,t)$. Damit ergibt sich die Lösung von (4.91b) in der Form

$$\begin{aligned}\boldsymbol{\lambda}^*(t) &= \boldsymbol{\Phi}^{*\mathrm{T}}(t_1,t)\left(\frac{\mathrm{d}}{\mathrm{d}\mathbf{x}_1}\varphi\right)^{\mathrm{T}}(\mathbf{x}^*(t_1)) - \int_{t_1}^{t}\boldsymbol{\Phi}^{*\mathrm{T}}(\tau,t)\left(\frac{\partial}{\partial\mathbf{x}}l\right)^{\mathrm{T}}(\tau,\mathbf{x}^*(\tau),\mathbf{u}^*(\tau))\,\mathrm{d}\tau \\ &= \boldsymbol{\Phi}^{*\mathrm{T}}(t_1,t)\left(\frac{\mathrm{d}}{\mathrm{d}\mathbf{x}_1}\varphi\right)^{\mathrm{T}}(\mathbf{x}^*(t_1)) + \int_{t}^{t_1}\boldsymbol{\Phi}^{*\mathrm{T}}(\tau,t)\left(\frac{\partial}{\partial\mathbf{x}}l\right)^{\mathrm{T}}(\tau,\mathbf{x}^*(\tau),\mathbf{u}^*(\tau))\,\mathrm{d}\tau\ .\end{aligned} \tag{4.97}$$

Einsetzen dieser Lösung in (4.91c) zeigt die Äquivalenz zwischen den Bedingungen (4.91c) und (4.96). □

Aus (4.91) folgt, dass sich die notwendigen Optimalitätsbedingungen für das Optimalsteuerungsproblem (4.89) und (4.90) aus $2n$ *Differentialgleichungen* in $\mathbf{x}^*$ und $\boldsymbol{\lambda}^*$ und m *algebraischen Gleichungen* zusammensetzen. Da für die Differentialgleichung von $\mathbf{x}^*$ der Wert zum Anfangszeitpunkt $t = t_0$ und für die Differentialgleichung von $\boldsymbol{\lambda}^*$ der Wert zum Endzeitpunkt $t = t_1$ gegeben ist, handelt es sich um ein *Zweipunktrandwertproblem*. Analog zum Lagrange-Multiplikator in Satz 3.9 lässt sich der adjungierte Zustand $\boldsymbol{\lambda}(t)$ gemäß (4.97) in der Form interpretieren, dass $\boldsymbol{\lambda}(t)$ (zu einem bestimmten Zeitpunkt t) der *Sensitivität des Kostenfunktionals* (4.89) 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 4.9 alternativ mit Hilfe eines Lagrangefunktionals 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 Funktionals $J(\mathbf{x}_1,\dots,\mathbf{x}_n)$ am Punkt $[\mathbf{x}_1^{\mathrm{T}},\dots,\mathbf{x}_n^{\mathrm{T}}]^{\mathrm{T}} \in \mathcal{V}$ bezüglich $\mathbf{x}_i$ mit $i \in \{1,\dots,n\}$ in Richtung $\boldsymbol{\xi}$ mit $\dim(\boldsymbol{\xi}) = \dim(\mathbf{x}_i)$ und $[\mathbf{x}_1^{\mathrm{T}},\dots,\mathbf{x}_i^{\mathrm{T}}+\boldsymbol{\xi}^{\mathrm{T}},\dots,\mathbf{x}_n^{\mathrm{T}}]^{\mathrm{T}} \in \mathcal{V}$ ist in der Form

$$\begin{aligned}\delta_{\mathbf{x}_i}J(\mathbf{x}_1,\dots,\mathbf{x}_n;\boldsymbol{\xi}) :&= \lim_{\eta\to 0}\frac{J(\mathbf{x}_1,\dots,\mathbf{x}_i+\eta\boldsymbol{\xi},\dots,\mathbf{x}_n)-J(\mathbf{x}_1,\dots,\mathbf{x}_n)}{\eta} \\ &= \frac{\mathrm{d}}{\mathrm{d}\eta}J(\mathbf{x}_1,\dots,\mathbf{x}_i+\eta\boldsymbol{\xi},\dots,\mathbf{x}_n)\Big|_{\eta=0}\end{aligned} \tag{4.98}$$

definiert.

Wird nun für das Kostenfunktional (4.89) mit den Gleichungsbeschränkungen (4.90) das *Lagrangefunktional*

$$\begin{aligned}L(\mathbf{x},\mathbf{u},\boldsymbol{\lambda}) &= J(\mathbf{u}) + \int_{t_0}^{t_1}\boldsymbol{\lambda}^{\mathrm{T}}(t)(\mathbf{f}(t,\mathbf{x}(t),\mathbf{u}(t)) - \dot{\mathbf{x}}(t))\,\mathrm{d}t \\ &= \varphi(\mathbf{x}(t_1)) + \int_{t_0}^{t_1} l(t,\mathbf{x}(t),\mathbf{u}(t)) + \boldsymbol{\lambda}^{\mathrm{T}}(t)(\mathbf{f}(t,\mathbf{x}(t),\mathbf{u}(t)) - \dot{\mathbf{x}}(t))\,\mathrm{d}t\end{aligned} \tag{4.99}$$

eingeführt (vgl. dazu die Lagrangefunktion (3.30) für ein statisches Optimierungsproblem mit Gleichungsbeschränkungen), so ergibt sich die Gâteaux Ableitung $\delta J(\mathbf{u};\boldsymbol{\xi}_u)$ an einem

allgemeinen Punkt $\mathbf{u} \in (C[t_0, t_1])^m$ in Richtung $\boldsymbol{\xi}_u(t) \in (C[t_0, t_1])^m$ in der Form

$$\begin{aligned} \delta J(\mathbf{u}; \boldsymbol{\xi}_u) &= \delta_{\mathbf{u}} L(\mathbf{x}, \mathbf{u}, \boldsymbol{\lambda}; \boldsymbol{\xi}_u) \\ &= \int_{t_0}^{t_1} \left[\frac{\partial}{\partial \mathbf{u}} l(t, \mathbf{x}(t), \mathbf{u}(t)) + \boldsymbol{\lambda}^{\mathrm{T}}(t) \frac{\partial}{\partial \mathbf{u}} \mathbf{f}(t, \mathbf{x}(t), \mathbf{u}(t)) \right] \boldsymbol{\xi}_u(t) \, \mathrm{d}t \ , \end{aligned} \tag{4.100a}$$

wobei die Bedingungen

$$\begin{aligned} \mathbf{0} &= \delta_{\mathbf{x}} L(\mathbf{x}, \mathbf{u}, \boldsymbol{\lambda}; \boldsymbol{\xi}_x) \\ &= \frac{\mathrm{d}}{\mathrm{d}\mathbf{x}_1} \varphi(\mathbf{x}(t_1)) \boldsymbol{\xi}_x(t_1) \\ &\quad + \int_{t_0}^{t_1} \frac{\partial}{\partial \mathbf{x}} l(t, \mathbf{x}(t), \mathbf{u}(t)) \boldsymbol{\xi}_x(t) + \boldsymbol{\lambda}^{\mathrm{T}}(t) \left(\frac{\partial}{\partial \mathbf{x}} \mathbf{f}(t, \mathbf{x}(t), \mathbf{u}(t)) \boldsymbol{\xi}_x(t) - \dot{\boldsymbol{\xi}}_x(t) \right) \mathrm{d}t \\ &= \left[\frac{\mathrm{d}}{\mathrm{d}\mathbf{x}_1} \varphi(\mathbf{x}(t_1)) - \boldsymbol{\lambda}^{\mathrm{T}}(t_1) \right] \boldsymbol{\xi}_x(t_1) + \boldsymbol{\lambda}^{\mathrm{T}}(t_0) \boldsymbol{\xi}_x(t_0) \\ &\quad + \int_{t_0}^{t_1} \left[\frac{\partial}{\partial \mathbf{x}} l(t, \mathbf{x}(t), \mathbf{u}(t)) + \boldsymbol{\lambda}^{\mathrm{T}}(t) \frac{\partial}{\partial \mathbf{x}} \mathbf{f}(t, \mathbf{x}(t), \mathbf{u}(t)) + \dot{\boldsymbol{\lambda}}^{\mathrm{T}}(t) \right] \boldsymbol{\xi}_x(t) \, \mathrm{d}t \end{aligned} \tag{4.100b}$$

$$\mathbf{0} = \delta_{\boldsymbol{\lambda}} L(\mathbf{x}, \mathbf{u}, \boldsymbol{\lambda}; \boldsymbol{\xi}_\lambda) = \int_{t_0}^{t_1} \boldsymbol{\xi}_\lambda^{\mathrm{T}}(t) [\mathbf{f}(t, \mathbf{x}(t), \mathbf{u}(t)) - \dot{\mathbf{x}}(t)] \, \mathrm{d}t \tag{4.100c}$$

für alle zulässigen Richtungen $\boldsymbol{\xi}_x(t) \in (C^1[t_0, t_1])^n$ mit $\boldsymbol{\xi}_x(t_0) = \mathbf{0}$ zufolge der Anfangsbedingung $\mathbf{x}(t_0) = \mathbf{x}_0$ und für alle zulässigen Richtungen $\boldsymbol{\xi}_\lambda(t) \in (C^1[t_0, t_1])^n$ einzuhalten sind. Gemäß Fundamentallemma der Variationsrechnung Lemma 4.2 folgt aus (4.100b)

$$\begin{aligned} \dot{\boldsymbol{\lambda}} &= -\left(\frac{\partial}{\partial \mathbf{x}} l \right)^{\mathrm{T}} (t, \mathbf{x}(t), \mathbf{u}(t)) - \left(\frac{\partial}{\partial \mathbf{x}} \mathbf{f} \right)^{\mathrm{T}} (t, \mathbf{x}(t), \mathbf{u}(t)) \boldsymbol{\lambda}(t) \\ \boldsymbol{\lambda}(t_1) &= \left(\frac{\mathrm{d}}{\mathrm{d}\mathbf{x}_1} \varphi \right)^{\mathrm{T}} (\mathbf{x}(t_1)) \end{aligned} \tag{4.101}$$

und aus (4.100c) die Differenzialgleichung (4.90). Um zu sehen, dass (4.100) tatsächlich genau die Gâteaux Ableitung $\delta J(\mathbf{u}; \boldsymbol{\xi}_u)$ gemäß (4.95) liefert, kann die Lösung $\boldsymbol{\lambda}(t)$ von (4.101) analog zu (4.97) mit Hilfe der zur Dynamikmatrix $-\left(\frac{\partial}{\partial \mathbf{x}} \mathbf{f}\right)^{\mathrm{T}} (t, \mathbf{x}(t), \mathbf{u}(t))$ und zum Zeitintervall $[\tau, t]$ gehörigen Transitionsmatrix $\boldsymbol{\Phi}^{\mathrm{T}}(\tau, t)$ formuliert und in (4.100a) eingesetzt werden. Ein Vergleich zwischen der hier beschriebenen Berechnung der Gâteaux Ableitung $\delta J(\mathbf{u}; \boldsymbol{\xi}_u)$ anhand des Lagrangefunktionals $L(\mathbf{x}, \mathbf{u}, \boldsymbol{\lambda})$ 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 (4.100a), (4.101) und (4.90) unter Verwendung von Lemma 4.2 (Fundamentallemma der Variationsrechnung) genau die Optimalitätsbedingungen (4.91) von Satz 4.9 (Euler-Lagrange Gleichungen).

Alternativ lassen sich diese mit Hilfe der *Hamiltonfunktion*

$$H(t, \mathbf{x}, \mathbf{u}, \boldsymbol{\lambda}) = l(t, \mathbf{x}, \mathbf{u}) + \boldsymbol{\lambda}^{\mathrm{T}}(t) \mathbf{f}(t, \mathbf{x}, \mathbf{u}) \tag{4.102}$$

auch in der Form

$$\dot{\mathbf{x}}^* = \left(\frac{\partial}{\partial \boldsymbol{\lambda}} H\right)^{\mathrm{T}} (t, \mathbf{x}^*(t), \mathbf{u}^*(t), \boldsymbol{\lambda}^*(t)), \qquad \mathbf{x}^*(t_0) = \mathbf{x}_0 \tag{4.103a}$$

$$\dot{\boldsymbol{\lambda}}^* = -\left(\frac{\partial}{\partial \mathbf{x}} H\right)^{\mathrm{T}} (t, \mathbf{x}^*(t), \mathbf{u}^*(t), \boldsymbol{\lambda}^*(t)), \qquad \boldsymbol{\lambda}^*(t_1) = \left(\frac{\partial}{\partial \mathbf{x}_1} \varphi\right)^{\mathrm{T}} (\mathbf{x}^*(t_1)) \tag{4.103b}$$

$$\mathbf{0} = \left(\frac{\partial}{\partial \mathbf{u}} H\right)^{\mathrm{T}} (t, \mathbf{x}^*(t), \mathbf{u}^*(t), \boldsymbol{\lambda}^*(t)) \tag{4.103c}$$

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 (4.26)) unterscheidet. Die Bedingung (4.103c) 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

$$\begin{aligned} \frac{\mathrm{d}}{\mathrm{d}t} H(t, \mathbf{x}, \mathbf{u}, \boldsymbol{\lambda}) &= \frac{\partial}{\partial t} H + \left(\frac{\partial}{\partial \mathbf{x}} H\right) \dot{\mathbf{x}}^* + \underbrace{\left(\frac{\partial}{\partial \mathbf{u}} H\right)}_{=\mathbf{0}} \dot{\mathbf{u}}^* + \left(\frac{\partial}{\partial \boldsymbol{\lambda}} H\right) \dot{\boldsymbol{\lambda}}^* \\ &= \frac{\partial}{\partial t} H - \left(\dot{\boldsymbol{\lambda}}^*\right)^{\mathrm{T}} \mathbf{f} + (\dot{\mathbf{x}}^*)^{\mathrm{T}} \dot{\boldsymbol{\lambda}}^* = \frac{\partial}{\partial t} H \ . \end{aligned} \tag{4.104}$$

Wenn daher weder $\mathbf{f}$ noch l explizit von der Zeit t abhängen, ist die Hamiltonfunktion H eine *Invariante* des Zweipunktrandwertproblems (4.103).

Im Weiteren muss ähnlich zur Legendre Bedingung gemäß Satz 4.3 für ein Minimum des Kostenfunktionals $J(\mathbf{u})$ die *notwendige Bedingung zweiter Ordnung*

$$\mathbf{d}^{\mathrm{T}} \frac{\partial^2}{\partial \mathbf{u}^2} H(t, \mathbf{x}^*(t), \mathbf{u}^*(t), \boldsymbol{\lambda}^*(t)) \mathbf{d} \geq 0 \qquad \forall\, \mathbf{d} \in \mathbb{R}^m,\ t \in [t_0, t_1] \ . \tag{4.105}$$

erfüllt sein. Sie wird auch *Legendre-Clebsch Bedingung* genannt.

In Satz 4.9 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 (4.91) 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 (4.90) 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 (4.89), (4.90), dann gelten für jeden Eckpunkt $c \in (t_0, t_1)$ die Bedingungen

$$\hat{\mathbf{x}}^*(c^-) = \hat{\mathbf{x}}^*(c^+) \tag{4.106a}$$

$$\hat{\boldsymbol{\lambda}}^*(c^-) = \hat{\boldsymbol{\lambda}}^*(c^+) \tag{4.106b}$$

$$H\left(c^-, \hat{\mathbf{x}}^*(c), \hat{\mathbf{u}}^*(c^-), \hat{\boldsymbol{\lambda}}^*(c)\right) = H\left(c^+, \hat{\mathbf{x}}^*(c), \hat{\mathbf{u}}^*(c^+), \hat{\boldsymbol{\lambda}}^*(c)\right) \ , \tag{4.106c}$$

wobei die Argumente c^- bzw. c^+ jeweils den links- bzw. rechtsseitigen Grenzwert liefern sollen. Man beachte, dass (4.106b) und (4.106c) 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 4.10 (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)*

$$J(\mathbf{u}, t_1) = \varphi(t_1, \mathbf{x}(t_1)) + \int_{t_0}^{t_1} l(t, \mathbf{x}(t), \mathbf{u}(t))\,\mathrm{d}t \tag{4.107}$$

unter den Gleichungsbeschränkungen

$$\dot{\mathbf{x}} - \mathbf{f}(t, \mathbf{x}(t), \mathbf{u}(t)) = \mathbf{0}, \quad \mathbf{x}(t_0) = \mathbf{x}_0 \tag{4.108a}$$

$$G_k(\mathbf{u}, t_1) = \psi_k(t_1, \mathbf{x}(t_1)) = 0, \quad k = 1, \ldots, p \tag{4.108b}$$

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* $\varphi(t_1, \mathbf{x}_1)$ *und* $\psi_k(t_1, \mathbf{x}_1)$, $k = 1, \ldots, p$ *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* $\mathbf{x}^* \in \left(C^1[t_0, T]\right)^n$ *die zugehörige Lösung des Anfangswertproblems* (4.108a)*. Darüber hinaus wird angenommen, dass für* p *linear unabhängige zulässige Richtungen* $\left(\boldsymbol{\xi}_{u,k}, \xi_{t_1,k}\right) \in (C[t_0, T])^m \times [t_0, T)$, $k = 1, \ldots, p$ *die* Regularitätsbedingung

$$\det\left(\begin{bmatrix} \delta G_1\left(\mathbf{u}^*, t_1^*; \boldsymbol{\xi}_{u,1}, \xi_{t_1,1}\right) & \cdots & \delta G_1\left(\mathbf{u}^*, t_1^*; \boldsymbol{\xi}_{u,p}, \xi_{t_1,p}\right) \\ \vdots & \ddots & \vdots \\ \delta G_p\left(\mathbf{u}^*, t_1^*; \boldsymbol{\xi}_{u,1}, \xi_{t_1,1}\right) & \cdots & \delta G_p\left(\mathbf{u}^*, t_1^*; \boldsymbol{\xi}_{u,p}, \xi_{t_1,p}\right) \end{bmatrix}\right) \neq 0 \tag{4.109}$$

gilt. Dann existieren ein $\boldsymbol{\lambda}^* \in \left(C^1[t_0, t_1^*]\right)^n$ *und ein* $\boldsymbol{\mu}^* \in \mathbb{R}^p$ *so, dass die Beziehungen*

$$\dot{\mathbf{x}}^* = \left(\frac{\partial}{\partial \boldsymbol{\lambda}} H\right)^{\mathrm{T}}(t, \mathbf{x}^*(t), \mathbf{u}^*(t), \boldsymbol{\lambda}^*(t)), \qquad \mathbf{x}^*(t_0) = \mathbf{x}_0 \tag{4.110a}$$

$$\begin{aligned} \dot{\boldsymbol{\lambda}}^* &= -\left(\frac{\partial}{\partial \mathbf{x}} H\right)^{\mathrm{T}}(t, \mathbf{x}^*(t), \mathbf{u}^*(t), \boldsymbol{\lambda}^*(t)) \\ \boldsymbol{\lambda}^*(t_1^*) &= \left(\frac{\partial}{\partial \mathbf{x}_1} \Phi\right)^{\mathrm{T}}(t_1^*, \mathbf{x}^*(t_1^*), \boldsymbol{\mu}^*) \end{aligned} \tag{4.110b}$$

$$\mathbf{0} = \left(\frac{\partial}{\partial \mathbf{u}} H\right)^{\mathrm{T}}(t, \mathbf{x}^*(t), \mathbf{u}^*(t), \boldsymbol{\lambda}^*(t)) \tag{4.110c}$$

für $t_0 \leq t \leq t_1$ *mit den* Transversalitätsbedingungen

$$\boldsymbol{\psi}(t_1^*, \mathbf{x}^*(t_1^*)) = \mathbf{0} \tag{4.111a}$$

$$\left(\frac{\partial}{\partial t_1}\Phi\right)(t_1^*, \mathbf{x}^*(t_1^*), \boldsymbol{\mu}^*) + H(t_1^*, \mathbf{x}^*(t_1^*), \mathbf{u}^*(t_1^*), \boldsymbol{\lambda}^*(t_1^*)) = 0 \ , \tag{4.111b}$$

und der Hamiltonfunktion $H = l + \boldsymbol{\lambda}^{\mathrm{T}}\mathbf{f}$ *sowie* $\Phi = \varphi + \boldsymbol{\mu}^{\mathrm{T}}\boldsymbol{\psi}$ *mit* $\boldsymbol{\psi} = \begin{bmatrix}\psi_1 & \psi_2 & \dots & \psi_p\end{bmatrix}^{\mathrm{T}}$ *erfüllt sind.*

Beweisskizze: Für das Kostenfunktional (4.107) mit den Gleichungsbeschränkungen (4.108) wird das *Lagrangefunktional*

$$\begin{aligned} &L(\mathbf{x}, \mathbf{u}, t_1, \boldsymbol{\lambda}, \boldsymbol{\mu}) \\ &\quad = J(\mathbf{u}, t_1) + \boldsymbol{\mu}^{\mathrm{T}}\boldsymbol{\psi}(t_1, \mathbf{x}(t_1)) + \int_{t_0}^{t_1} \boldsymbol{\lambda}^{\mathrm{T}}(t)(\mathbf{f}(t, \mathbf{x}(t), \mathbf{u}(t)) - \dot{\mathbf{x}}(t))\,\mathrm{d}t \\ &\quad = \varphi(t_1, \mathbf{x}(t_1)) + \boldsymbol{\mu}^{\mathrm{T}}\boldsymbol{\psi}(t_1, \mathbf{x}(t_1)) \\ &\qquad + \int_{t_0}^{t_1} l(t, \mathbf{x}(t), \mathbf{u}(t)) + \boldsymbol{\lambda}^{\mathrm{T}}(t)(\mathbf{f}(t, \mathbf{x}(t), \mathbf{u}(t)) - \dot{\mathbf{x}}(t))\,\mathrm{d}t \end{aligned} \tag{4.112}$$

formuliert. Die Gâteaux Ableitung $\delta J(\mathbf{u}, t_1; \boldsymbol{\xi}_u, \xi_{t_1})$ an einem allgemeinen Punkt $(\mathbf{u}, t_1) \in (C[t_0, T])^m \times [t_0, T)$ in Richtung $(\boldsymbol{\xi}_u(t), \xi_{t_1}) \in (C[t_0, T])^m \times [t_0, T)$ lautet

$$\delta J(\mathbf{u}, t_1; \boldsymbol{\xi}_u, \xi_{t_1}) = \delta_{\mathbf{u}} L(\mathbf{x}, \mathbf{u}, t_1, \boldsymbol{\lambda}, \boldsymbol{\mu}; \boldsymbol{\xi}_u) + \delta_{t_1} L(\mathbf{x}, \mathbf{u}, t_1, \boldsymbol{\lambda}, \boldsymbol{\mu}; \xi_{t_1}) \tag{4.113a}$$

mit

$$\begin{aligned} &\delta_{\mathbf{u}} L(\mathbf{x}, \mathbf{u}, t_1, \boldsymbol{\lambda}, \boldsymbol{\mu}; \boldsymbol{\xi}_u) \\ &\quad = \int_{t_0}^{t_1} \left[\frac{\partial}{\partial \mathbf{u}} l(t, \mathbf{x}(t), \mathbf{u}(t)) + \boldsymbol{\lambda}^{\mathrm{T}}(t)\frac{\partial}{\partial \mathbf{u}}\mathbf{f}(t, \mathbf{x}(t), \mathbf{u}(t))\right]\boldsymbol{\xi}_u(t)\,\mathrm{d}t \end{aligned} \tag{4.113b}$$

$$\begin{aligned} &\delta_{t_1} L(\mathbf{x}, \mathbf{u}, t_1, \boldsymbol{\lambda}, \boldsymbol{\mu}; \xi_{t_1}) \\ &\quad = \Bigg[\frac{\partial}{\partial t_1}\varphi(t_1, \mathbf{x}(t_1)) + \frac{\partial}{\partial \mathbf{x}_1}\varphi(t_1, \mathbf{x}(t_1))\dot{\mathbf{x}}(t_1) \\ &\qquad + \boldsymbol{\mu}^{\mathrm{T}}\left(\frac{\partial}{\partial t_1}\boldsymbol{\psi}(t_1, \mathbf{x}(t_1)) + \frac{\partial}{\partial \mathbf{x}_1}\boldsymbol{\psi}(t_1, \mathbf{x}(t_1))\dot{\mathbf{x}}(t_1)\right) \\ &\qquad + l(t_1, \mathbf{x}(t_1), \mathbf{u}(t_1)) + \boldsymbol{\lambda}^{\mathrm{T}}(t_1)(\mathbf{f}(t_1, \mathbf{x}(t_1), \mathbf{u}(t_1)) - \dot{\mathbf{x}}(t_1))\Bigg]\xi_{t_1} \ , \end{aligned} \tag{4.113c}$$

wobei die Bedingungen

$$\begin{aligned}\mathbf{0} &= \delta_{\mathbf{x}} L(\mathbf{x},\mathbf{u},t_1,\boldsymbol{\lambda},\boldsymbol{\mu};\boldsymbol{\xi}_x)\\ &= \frac{\partial}{\partial \mathbf{x_1}}\varphi(t_1,\mathbf{x}(t_1))\boldsymbol{\xi}_x(t_1) + \boldsymbol{\mu}^{\mathrm{T}}\frac{\partial}{\partial \mathbf{x_1}}\boldsymbol{\psi}(t_1,\mathbf{x}(t_1))\boldsymbol{\xi}_x(t_1)\\ &\quad + \int_{t_0}^{t_1}\frac{\partial}{\partial \mathbf{x}} l(t,\mathbf{x}(t),\mathbf{u}(t))\boldsymbol{\xi}_x(t) + \boldsymbol{\lambda}^{\mathrm{T}}(t)\left(\frac{\partial}{\partial \mathbf{x}}\mathbf{f}(t,\mathbf{x}(t),\mathbf{u}(t))\boldsymbol{\xi}_x(t) - \dot{\boldsymbol{\xi}}_x(t)\right)\mathrm{d}t\\ &= \left[\frac{\partial}{\partial \mathbf{x}_1}\varphi(t_1,\mathbf{x}(t_1)) + \boldsymbol{\mu}^{\mathrm{T}}\frac{\partial}{\partial \mathbf{x_1}}\boldsymbol{\psi}(t_1,\mathbf{x}(t_1)) - \boldsymbol{\lambda}^{\mathrm{T}}(t_1)\right]\boldsymbol{\xi}_x(t_1) + \boldsymbol{\lambda}^{\mathrm{T}}(t_0)\boldsymbol{\xi}_x(t_0)\\ &\quad + \int_{t_0}^{t_1}\left[\frac{\partial}{\partial \mathbf{x}} l(t,\mathbf{x}(t),\mathbf{u}(t)) + \boldsymbol{\lambda}^{\mathrm{T}}(t)\frac{\partial}{\partial \mathbf{x}}\mathbf{f}(t,\mathbf{x}(t),\mathbf{u}(t)) + \dot{\boldsymbol{\lambda}}^{\mathrm{T}}(t)\right]\boldsymbol{\xi}_x(t)\,\mathrm{d}t\end{aligned} \tag{4.113d}$$

$$\mathbf{0} = \delta_{\boldsymbol{\lambda}} L(\mathbf{x},\mathbf{u},\boldsymbol{\lambda},\boldsymbol{\mu};\boldsymbol{\xi}_\lambda) = \int_{t_0}^{t_1}\boldsymbol{\xi}_\lambda^{\mathrm{T}}(t)[\mathbf{f}(t,\mathbf{x}(t),\mathbf{u}(t)) - \dot{\mathbf{x}}(t)]\,\mathrm{d}t \tag{4.113e}$$

$$\mathbf{0} = \left(\frac{\partial}{\partial \boldsymbol{\mu}} L\right)^{\mathrm{T}}(\mathbf{x},\mathbf{u},t_1,\boldsymbol{\lambda},\boldsymbol{\mu}) = \boldsymbol{\psi}(t_1,\mathbf{x}(t_1)) \tag{4.113f}$$

für alle zulässigen Richtungen $\boldsymbol{\xi}_x(t) \in (C^1[t_0,T])^n$ mit $\boldsymbol{\xi}_x(t_0) = \mathbf{0}$ zufolge der Anfangsbedingung $\mathbf{x}(t_0) = \mathbf{x}_0$ und für alle zulässigen Richtungen $\boldsymbol{\xi}_\lambda(t) \in (C^1[t_0,T])^n$ einzuhalten sind. Es ist an dieser Stelle zu beachten, dass $\boldsymbol{\xi}_x(t_1)$ grundsätzlich beliebige Werte annehmen kann. Dies gilt auch, wenn Endbedingungen für $\mathbf{x}(t_1)$ einzuhalten sind, da diese durch entsprechende Gleichungsbeschränkungen $\psi_k(t_1,\mathbf{x}(t_1)) = 0$ und zugehörige Lagrange-Multiplikatoren μ_k berücksichtigt werden. Gemäß Fundamentallemma der Variationsrechnung Lemma 4.2 folgt aus (4.113d)

$$\dot{\boldsymbol{\lambda}} = -\left(\frac{\partial}{\partial \mathbf{x}} l\right)^{\mathrm{T}}(t,\mathbf{x}(t),\mathbf{u}(t)) - \left(\frac{\partial}{\partial \mathbf{x}}\mathbf{f}\right)^{\mathrm{T}}(t,\mathbf{x}(t),\mathbf{u}(t))\boldsymbol{\lambda}(t) \tag{4.114a}$$

$$\boldsymbol{\lambda}(t_1) = \left(\frac{\partial}{\partial \mathbf{x}_1}\varphi\right)^{\mathrm{T}}(t_1,\mathbf{x}(t_1)) + \boldsymbol{\mu}^{\mathrm{T}}\frac{\partial}{\partial \mathbf{x_1}}\boldsymbol{\psi}(t_1,\mathbf{x}(t_1)) \tag{4.114b}$$

und aus (4.113e) die Differenzialgleichung (4.108a). Wegen (4.114b) vereinfacht sich (4.113c) zu

$$\begin{aligned}&\delta_{t_1} L(\mathbf{x},\mathbf{u},t_1,\boldsymbol{\lambda},\boldsymbol{\mu};\xi_{t_1})\\ &\qquad = \left[\frac{\partial}{\partial t_1}\varphi(t_1,\mathbf{x}(t_1)) + \boldsymbol{\mu}^{\mathrm{T}}\frac{\partial}{\partial t_1}\boldsymbol{\psi}(t_1,\mathbf{x}(t_1))\right.\\ &\qquad\quad \left. + l(t_1,\mathbf{x}(t_1),\mathbf{u}(t_1)) + \boldsymbol{\lambda}^{\mathrm{T}}(t_1)\mathbf{f}(t_1,\mathbf{x}(t_1),\mathbf{u}(t_1))\right]\xi_{t_1}\ .\end{aligned} \tag{4.115}$$

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 (4.113a), (4.113b), (4.113f), (4.114), (4.115) und (4.108a)

unter Verwendung von Lemma 4.2 genau die Optimalitätsbedingungen (4.110) und (4.111) von Satz 4.10.

Für den Beweis der Notwendigkeit der Regularitätsbedingung (4.109) 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 (4.108b) erfüllt. □

Aufgabe 4.4. Beweisen Sie Satz 4.10 ohne Verwendung des Lagrangefunktionals (4.112). **Hinweis:** Orientieren Sie sich dabei am Beweis von Satz 4.9.

Zur Berechnung der $m + 2n + p + 1$ *unbekannten Größen* $(\mathbf{u}^*(t), \mathbf{x}^*(t), \boldsymbol{\lambda}^*(t), \boldsymbol{\mu}^*, t_1^*)$ stehen mit Satz 4.10 $m + 2n + p + 1$ *Bedingungen* zur Verfügung. Das sind m algebraische Gleichungen (4.110c), $p + 1$ algebraische Gleichungen in Form der Transversalitätsbedingungen (4.111) und $2n$ Differentialgleichungen (4.110a) und (4.110b) für $\mathbf{x}^*$ und $\boldsymbol{\lambda}^*$. Zu diesen Differentialgleichungen gehören die in (4.110a) und (4.110b) angegebenen $2n$ Randbedingungen. Aus den genannten Gleichungen lassen sich eindeutig die unbekannten Größen $(\mathbf{u}^*(t), \mathbf{x}^*(t), \boldsymbol{\lambda}^*(t), \boldsymbol{\mu}^*, t_1^*)$ bestimmen.

Für die folgende kurze Diskussion des Satzes 4.10 werden nur sogenannte partielle Endbedingungen der Form

$$\psi_j = x_k(t_1) - \overline{x}_k, \quad j = 1, \ldots, p \tag{4.116}$$

mit $\overline{x}_k = \text{konst.}$ als fixem Endwert der Komponente $x_k(t)$ von $\mathbf{x}(t)$ betrachtet. In diesem Spezialfall vereinfacht sich die Endbedingung für $\boldsymbol{\lambda}^*(t_1^*)$ in (4.110b) unter Berücksichtigung von (4.114b) und (4.115) wie folgt:

(a) Wenn die *Endzeit* t_1 *fest ist* und damit $\xi_{t_1} = 0$ gilt, wird (4.115) automatisch erfüllt. Es liegt somit keine Transversalitätsbedingung gemäß (4.111b) vor.

(i) Wenn für eine Komponente $x_k(t)$ von $\mathbf{x}(t)$ gilt, dass *deren Endwert fest ist*, so wird dies über eine Gleichungsbedingung $\psi_j = x_k(t_1) - \overline{x}_k = 0$ berücksichtigt und es sollte $\partial\varphi/\partial x_{1,k} = 0$ gelten. Daraus folgt gemäß (4.114b) die Endbedingung

$$\lambda_k^*(t_1) = \mu_j \tag{4.117}$$

für den zugehörigen adjungierten Zustand $\lambda_k^*(t_1)$. Der Wert μ_j ist zunächst unbekannt.

(ii) Wenn für eine Komponente $x_k(t)$ von $\mathbf{x}(t)$ gilt, dass *deren Endwert frei ist*, so existiert für diesen Endwert keine Gleichungsbedingung und es folgt gemäß (4.114b) die Endbedingung

$$\lambda_k^*(t_1) = \frac{\partial}{\partial x_{1,k}} \varphi(t_1, \mathbf{x}^*(t_1)) \tag{4.118}$$

für den zugehörigen adjungierten Zustand $\lambda_k^*(t_1)$.

(b) Wenn *die Endzeit frei ist* und ξ_{t_1} somit beliebige Werte annehmen kann, muss die Transversalitätsbedingung

$$\frac{\partial}{\partial t_1} \Phi(t_1^*, \mathbf{x}^*(t_1^*), \boldsymbol{\mu}^*) + H(t_1^*, \mathbf{x}^*(t_1^*), \mathbf{u}^*(t_1^*), \boldsymbol{\lambda}^*(t_1^*)) = 0, \quad H = l + \boldsymbol{\lambda}^{\mathrm{T}} \mathbf{f} \tag{4.119}$$

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 angewendet werden.

Wenn in Satz 4.10 die Gleichungsbeschränkungen (4.108b) durch *Ungleichungsbeschränkungen* der Form

$$\psi_k(t_1, \mathbf{x}(t_1)) \leq 0, \quad k = 1, \dots, p \tag{4.120}$$

ersetzt werden, so ändert sich lediglich (4.111a) zu

$$\psi_k(t_1^*, \mathbf{x}^*(t_1^*)) \leq 0 \tag{4.121a}$$

$$\boldsymbol{\mu}^* \geq \mathbf{0} \tag{4.121b}$$

$$(\boldsymbol{\mu}^*)^{\mathrm{T}} \boldsymbol{\psi}(t_1^*, \mathbf{x}^*(t_1^*)) = 0\ , \tag{4.121c}$$

wobei (4.121c) auch als *complementary slackness condition* bezeichnet wird.

Aufgabe 4.5. Gesucht ist eine Lösung des Optimierungsproblems

$$\min_{u(\cdot)} \quad \int_0^1 \frac{1}{2} u^2 + \frac{a}{2} x^2 \ \mathrm{d}t, \quad a > 0 \tag{4.122a}$$

$$\text{u.B.v.} \quad \dot{x} = u, \quad x(0) = 1, \quad x(1) = 0\,. \tag{4.122b}$$

Zeigen Sie, dass die Lösung durch

$$x^*(t) = \frac{1}{1 - \mathrm{e}^{2\sqrt{a}}} \left(\mathrm{e}^{\sqrt{a}t} - \mathrm{e}^{\sqrt{a}(2-t)} \right), \quad u^*(t) = \frac{\sqrt{a}}{1 - \mathrm{e}^{2\sqrt{a}}} \left(\mathrm{e}^{\sqrt{a}t} + \mathrm{e}^{\sqrt{a}(2-t)} \right) \tag{4.123}$$

gegeben ist und interpretieren Sie die Ergebnisse, die in Abbildung 4.4 für verschiedene Parameterwerte a dargestellt sind.

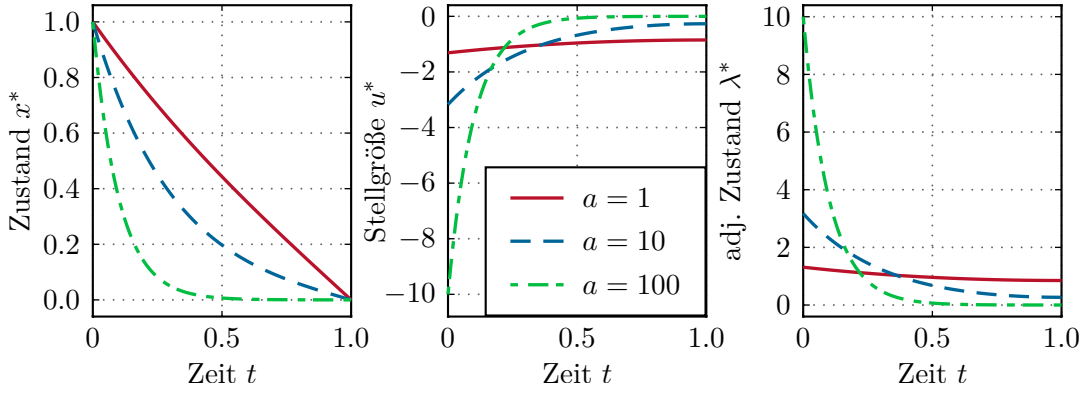


Abbildung 4.4: Optimale Trajektorien in Aufgabe 4.5.

Aufgabe 4.6. Gesucht ist eine Lösung des Optimierungsproblems

$$\min_{u(\cdot)} \quad \frac{a}{2} x_2^2(1) + \int_0^1 \frac{1}{2} u^2 \ \mathrm{d}t, \quad a \geq 0 \tag{4.124a}$$

$$\text{u.B.v.} \quad \dot{x}_1 = x_2, \quad x_1(0) = 1, \quad x_1(1) = 0 \tag{4.124b}$$
$$\dot{x}_2 = u, \quad x_2(0) = 0. \tag{4.124c}$$

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

$$x_1^*(t) = \frac{2(1+a)}{4+a}\, t^3 - \frac{3(2+a)}{a+4}\, t^2 + 1, \quad u^*(t) = \frac{12(1+a)}{4+a}\, t - \frac{6(2+a)}{a+4} \tag{4.125}$$

gegeben ist. Interpretieren Sie die Ergebnisse in Abbildung 4.5.

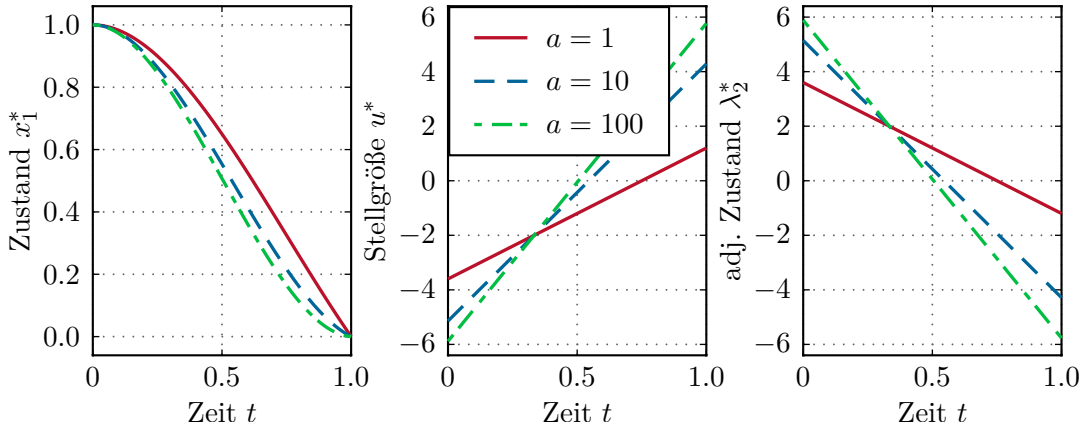


Abbildung 4.5: Optimale Trajektorien in Aufgabe 4.6.

Beispiel 4.10. Betrachtet wird eine Punktmasse mit der Masse m in der (x, y)-Ebene, auf die eine konstante Schubkraft $F = ma$ wirkt. Die Stellgröße u des Problems ist der Winkel zwischen der Kraftrichtung (Schubrichtung) und der x-Achse, siehe Abbildung 4.6. Ziel ist es, die Punktmasse im Zeitraum $[t_0 = 0, t_1^*]$ mit *minimaler Endzeit* t_1^* zu einem *fest vorgegebenen Zielpunkt* $(\bar{x}_1, \bar{y}_1)$ zu steuern. Unter der Annahme, dass außer F keine weiteren Kräfte auftreten, kann das Optimalsteuerungsproblem wie folgt formuliert werden

$$\min_{u(\cdot)} \quad t_1 \tag{4.126a}$$
$$\text{u.B.v.} \quad \dot{x} = v\,, \qquad x(0) = x_0\,, \quad x(t_1) = \bar{x}_1 \tag{4.126b}$$
$$\dot{v} = a\cos(u)\,, \quad v(0) = v_0 \tag{4.126c}$$
$$\dot{y} = w\,, \qquad y(0) = y_0\,, \quad y(t_1) = \bar{y}_1 \tag{4.126d}$$
$$\dot{w} = a\sin(u)\,, \quad w(0) = w_0\,. \tag{4.126e}$$

Man beachte, dass der Endzustand nur für die Position (x, y) aber nicht für die Geschwindigkeiten (v, w) vorgegeben ist.

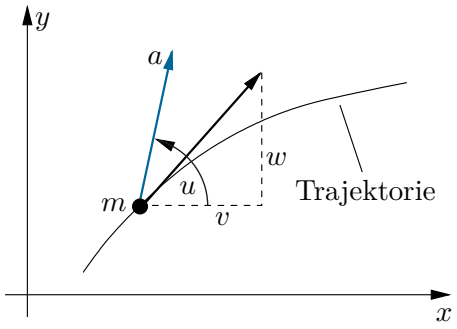


Abbildung 4.6: Bewegung einer Punktmasse der Masse m in der (x, y)-Ebene.

Die beiden fest vorgegebenen Endwerte für x und y können als Gleichungsbeschränkungen gemäß (4.108b) in der Form

$$\psi_1(t_1, \mathbf{x}(t_1)) = x(t_1) - \bar{x}_1, \quad \psi_2(t_1, \mathbf{x}(t_1)) = y(t_1) - \bar{y}_1 \tag{4.127}$$

formuliert werden. Die Hamiltonfunktion H und die Funktion Φ gemäß Satz 4.10 lauten dann für das vorliegende Optimierungsproblem

$$H(\mathbf{x}, u, \boldsymbol{\lambda}) = \lambda_x v + \lambda_v a\cos(u) + \lambda_y w + \lambda_w a\sin(u) \tag{4.128a}$$
$$\Phi(t_1, \mathbf{x}(t_1), \boldsymbol{\mu}) = \varphi + \mu_x\psi_1 + \mu_y\psi_2 = t_1 + \mu_x(x(t_1) - \bar{x}_1) + \mu_y(y(t_1) - \bar{y}_1) \tag{4.128b}$$

mit $\mathbf{x} = [x \quad v \quad y \quad w]^{\mathrm{T}}$, dem adjungierten Zustand $\boldsymbol{\lambda} = [\lambda_x \quad \lambda_v \quad \lambda_y \quad \lambda_w]^{\mathrm{T}}$ und dem *konstanten Lagrange-Multiplikator* $\boldsymbol{\mu} = [\mu_x \quad \mu_y]^{\mathrm{T}}$. Die Randbedingungen für den adjungierten Zustand errechnen sich gemäß (4.110b) zu

$$\lambda_x^*(t_1^*) = \left(\frac{\partial}{\partial x_1}\Phi\right)(t_1^*, \mathbf{x}^*(t_1^*), \boldsymbol{\mu}^*) = \mu_x^* \tag{4.129a}$$
$$\lambda_v^*(t_1^*) = \left(\frac{\partial}{\partial v_1}\Phi\right)(t_1^*, \mathbf{x}^*(t_1^*), \boldsymbol{\mu}^*) = 0 \tag{4.129b}$$
$$\lambda_y^*(t_1^*) = \left(\frac{\partial}{\partial y_1}\Phi\right)(t_1^*, \mathbf{x}^*(t_1^*), \boldsymbol{\mu}^*) = \mu_y^* \tag{4.129c}$$
$$\lambda_w^*(t_1^*) = \left(\frac{\partial}{\partial w_1}\Phi\right)(t_1^*, \mathbf{x}^*(t_1^*), \boldsymbol{\mu}^*) = 0\ . \tag{4.129d}$$

Damit lautet das adjungierte System

$$\dot{\lambda}_x^* = -\left(\frac{\partial H}{\partial x}\right)(\mathbf{x}^*, u^*, \boldsymbol{\lambda}^*) = 0\,, \qquad \lambda_x^*(t_1^*) = \mu_x^* \tag{4.130a}$$
$$\dot{\lambda}_v^* = -\left(\frac{\partial H}{\partial v}\right)(\mathbf{x}^*, u^*, \boldsymbol{\lambda}^*) = -\lambda_x^*\,, \quad \lambda_v^*(t_1^*) = 0 \tag{4.130b}$$
$$\dot{\lambda}_y^* = -\left(\frac{\partial H}{\partial y}\right)(\mathbf{x}^*, u^*, \boldsymbol{\lambda}^*) = 0\,, \qquad \lambda_y^*(t_1^*) = \mu_y^* \tag{4.130c}$$

$$\dot{\lambda}_w^* = -\left(\frac{\partial H}{\partial w}\right)(\mathbf{x}^*, u^*, \boldsymbol{\lambda}^*) = -\lambda_y^*\,, \quad \lambda_w^*(t_1^*) = 0\,, \tag{4.130d}$$

woraus direkt

$$\lambda_x^* = \mu_x^*\,, \quad \lambda_v^* = \mu_x^*(t_1^* - t)\,, \quad \lambda_y^* = \mu_y^*\,, \quad \lambda_w^* = \mu_y^*(t_1^* - t) \tag{4.131}$$

folgt. Des Weiteren muss die Hamiltonfunktion H gemäß (4.110) extremal sein, weshalb die Bedingung

$$\left(\frac{\partial H}{\partial u}\right)(\mathbf{x}^*, u^*, \boldsymbol{\lambda}^*) = -\lambda_v^* a \sin(u^*) + \lambda_w^* a \cos(u^*) = 0\,, \tag{4.132}$$

erfüllt sein muss. Unter Verwendung von (4.131) ergibt sich also

$$\frac{\sin(u^*)}{\cos(u^*)} = \frac{\lambda_w^*}{\lambda_v^*} = \frac{\mu_y^*(t_1^* - t)}{\mu_x^*(t_1^* - t)} = \frac{\mu_y^*}{\mu_x^*} = \text{konst.} \tag{4.133}$$

Daraus kann u^* noch nicht eindeutig berechnet werden. Aus der Legendre-Clebsch Bedingung (4.105) folgt

$$\left(\frac{\partial^2 H}{\partial u^2}\right)(\mathbf{x}^*, u^*, \boldsymbol{\lambda}^*) = -\lambda_v^* a \cos(u^*) - \lambda_w^* a \sin(u^*) \geq 0 \tag{4.134}$$

und damit unter Verwendung von (4.131)

$$\mu_x^* \cos(u^*) + \mu_y^* \sin(u^*) \leq 0\,. \tag{4.135}$$

Wegen (4.133) und (4.135) müssen die Vektoren $[\cos(u^*), \sin(u^*)]^{\mathrm{T}}$ und $[\mu_x, \mu_y]^{\mathrm{T}}$ parallel und in entgegengesetzte Richtung orientiert sein. Dies führt unter Einhaltung der Bedingung $\cos^2(u^*) + \sin^2(u^*) = 1$ auf die Lösungen

$$\cos(u^*) = -\frac{\mu_x^*}{\sqrt{(\mu_x^*)^2 + (\mu_y^*)^2}}\,, \qquad \sin(u^*) = -\frac{\mu_y^*}{\sqrt{(\mu_x^*)^2 + (\mu_y^*)^2}}\,. \tag{4.136}$$

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 (4.126) und Einsetzen der Anfangsbedingungen in der Form

$$x^*(t) = g_x(\mu_x^*, \mu_y^*, t) = x_0 + v_0 t + \frac{1}{2} a \cos(u^*) t^2 \tag{4.137a}$$

$$v^*(t) = g_v(\mu_x^*, \mu_y^*, t) = v_0 + a \cos(u^*) t \tag{4.137b}$$

$$y^*(t) = g_y(\mu_x^*, \mu_y^*, t) = y_0 + w_0 t + \frac{1}{2} a \sin(u^*) t^2 \tag{4.137c}$$

$$w^*(t) = g_w(\mu_x^*, \mu_y^*, t) = w_0 + a \sin(u^*) t \tag{4.137d}$$

bestimmt werden. Da die Endzeit t_1 frei ist, muss zusätzlich die Transversalitätsbedingung (4.111b) gelten, wobei sich die Hamiltonfunktion (4.128a) aufgrund der Endbedingungen $\lambda_v^*(t_1^*) = \lambda_w^*(t_1^*) = 0$ entsprechend vereinfacht

$$\begin{aligned} 0 &= \left(\frac{\partial}{\partial t_1}\Phi\right)(t_1^*, \mathbf{x}^*(t_1^*), \boldsymbol{\mu}^*) + H(\mathbf{x}^*(t_1^*), u^*(t_1^*), \boldsymbol{\lambda}^*(t_1^*)) \\ &= 1 + \mu_x^*\, g_v(\mu_x^*, \mu_y^*, t_1^*) + \mu_y^*\, g_w(\mu_x^*, \mu_y^*, t_1^*)\ . \end{aligned} \tag{4.138}$$

Mit Hilfe der zwei Gleichungsbeschränkungen (4.127) und der Transversalitätsbedingung (4.138) lässt sich ein Gleichungssystem für die verbleibenden drei Unbekannten μ_x^*, μ_y^* und t_1^* in der Form

$$\begin{bmatrix} g_x(\mu_x^*, \mu_y^*, t_1^*) \\ g_y(\mu_x^*, \mu_y^*, t_1^*) \\ \mu_x^*\, g_v(\mu_x^*, \mu_y^*, t_1^*) + \mu_y^*\, g_w(\mu_x^*, \mu_y^*, t_1^*) \end{bmatrix} - \begin{bmatrix} \bar{x}_1 \\ \bar{y}_1 \\ -1 \end{bmatrix} = \mathbf{0} \tag{4.139}$$

formulieren, welches auf *numerischem Wege* gelöst werden kann, z. B. mit der MATLAB-Funktion fsolve. Diese Möglichkeit ist in der Code-Auflistung 4.1 gezeigt. 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 $(v_0, w_0) = (0, 1)$ 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 repräsentieren die (jeweils konstante) optimale Richtung der angreifenden Kraft ma definiert durch $\cos(u^*)$ und $\sin(u^*)$ gemäß (4.136).

Code-Auflistung 4.1: MATLAB-Code für das Punktmasse-Problem.

```
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 = -mux/sqrt(mux^2+muy^2);
x = p.x0 + p.v0*t + p.a/2*cosu*t.^2;
function v = vfct(mux,muy,t,p)
cosu = -mux/sqrt(mux^2+muy^2);
v = p.v0 + p.a*cosu*t;
% ------------------------------------------------
function y = yfct(mux,muy,t,p)              % Funktionen für y und w
sinu = -muy/sqrt(mux^2+muy^2);
y = p.y0 + p.w0*t + p.a/2*sinu*t.^2;
function w = wfct(mux,muy,t,p)
sinu = -muy/sqrt(mux^2+muy^2);
w = p.w0 + p.a*sinu*t;
```

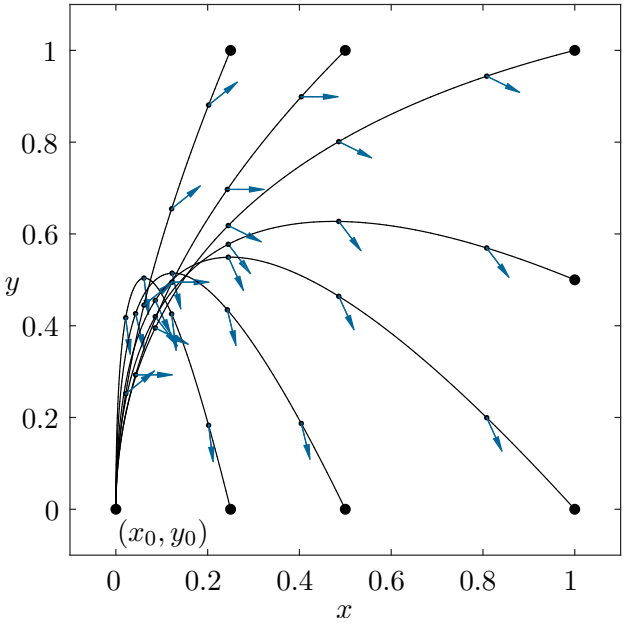


Abbildung 4.7: Zeitoptimale Bewegung einer Punktmasse.

Aufgabe 4.7. Gegeben ist ein lineares zeitvariantes Mehrgrößensystem der Form

$$\dot{\mathbf{x}} = \mathbf{A}(t)\mathbf{x} + \mathbf{B}(t)\mathbf{u}, \quad \mathbf{x}(t_0) = \mathbf{x}_0 \tag{4.140}$$

mit dem Zustand $\mathbf{x} \in \mathbb{R}^n$ und dem Stelleingang $\mathbf{u} \in \mathbb{R}^m$. Zeigen Sie, dass das *zeitvariante Zustandsregelgesetz*

$$\mathbf{u}^*(t) = -\mathbf{R}^{-1}(t)\mathbf{B}^{\mathrm{T}}(t)\mathbf{S}(t)\mathbf{x}^*(t) \tag{4.141}$$

mit $\mathbf{S}(t)$ als Lösung der *Matrix-Riccati-Differentialgleichung*

$$\dot{\mathbf{S}} = -\mathbf{S}\mathbf{A} - \mathbf{A}^{\mathrm{T}}\mathbf{S} + \mathbf{S}\mathbf{B}\mathbf{R}^{-1}\mathbf{B}^{\mathrm{T}}\mathbf{S} - \mathbf{Q}, \quad \mathbf{S}(t_1) = \mathbf{S}_1, \quad t_0 \le t \le t_1 \tag{4.142}$$

das Kostenfunktional

$$J(\mathbf{u}) = \frac{1}{2}\int_{t_0}^{t_1} \mathbf{x}^{\mathrm{T}}(t)\mathbf{Q}(t)\mathbf{x}(t) + \mathbf{u}^{\mathrm{T}}(t)\mathbf{R}(t)\mathbf{u}(t)\,\mathrm{d}t + \frac{1}{2}\mathbf{x}^{\mathrm{T}}(t_1)\mathbf{S}_1\mathbf{x}(t_1) \tag{4.143}$$

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 semi-definiten Matrix $\mathbf{Q}(t)$ und der positiv semi-definiten Matrix $\mathbf{S}_1$ 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 Kostenfunktionals

$$J(\mathbf{u}) = \int_{t_0}^{t_1} l(\mathbf{x}(t), \mathbf{u}(t))\,\mathrm{d}t \tag{4.144}$$

mit der freien Endzeit t_1 und einem festen Endzustand $\mathbf{x}(t_1) = \mathbf{x}_1$ unter der Gleichungsbeschränkung des dynamischen Systems

$$\dot{\mathbf{x}} = \mathbf{f}(\mathbf{x}(t), \mathbf{u}(t)), \quad \mathbf{x}(t_0) = \mathbf{x}_0 \tag{4.145}$$

für die zulässigen Stellgrößen

$$\mathbf{u} \in \left(\hat{C}_U[t_0, T]\right)^m := \left\{ \mathbf{u} \in \left(\hat{C}[t_0, T]\right)^m \,\middle|\, \mathbf{u}(t) \in U,\, \forall t_0 \leq t \leq T \right\} \tag{4.146}$$

mit einer hinreichend großen Zeit $T \gg t_1$ und der nichtleeren Menge der Stellgrößenbeschränkungen U.

Die Lagrangesche Dichte l in (4.144) und die rechte Seite $\mathbf{f}$ in (4.145) hängen nicht explizit von der Zeit t ab, d. h. es handelt sich um eine *zeitinvariante* Problemformulierung. Der Abschnitt 4.2.5 wird zeigen, wie zeitvariante Problemformulierungen behandelt werden können.

Das Optimierungsproblem bestehend aus (4.144) und (4.145) kann durch Erweiterung des Zustandsvektors auf die Form $\bar{\mathbf{x}} = [\mathbf{x}^{\mathrm{T}} \quad x_{n+1}]^{\mathrm{T}}$ mit

$$x_{n+1}(t) := \int_{t_0}^{t} l(\mathbf{x}(\tau), \mathbf{u}(\tau))\,\mathrm{d}\tau \tag{4.147}$$

wie folgt umformuliert werden: Gesucht werden 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

$$\underbrace{\begin{bmatrix} \dot{\mathbf{x}} \\ \dot{x}_{n+1} \end{bmatrix}}_{\dot{\bar{\mathbf{x}}}} = \underbrace{\begin{bmatrix} \mathbf{f}(\mathbf{x}, \mathbf{u}) \\ l(\mathbf{x}, \mathbf{u}) \end{bmatrix}}_{\bar{\mathbf{f}}(\mathbf{x},\mathbf{u})}, \quad \bar{\mathbf{x}}(t_0) = \begin{bmatrix} \mathbf{x}_0 \\ 0 \end{bmatrix} \tag{4.148}$$

beim Punkt $\bar{\mathbf{x}}(t_1) = [\mathbf{x}_1^{\mathrm{T}} \quad x_{n+1}(t_1)]^{\mathrm{T}}$ terminiert und dabei $x_{n+1}(t_1)$ möglichst klein ist. D. h. das neue Kostenfunktional lautet $J(\mathbf{u}) = x_{n+1}(t_1)$ (Mayer-Form). Abbildung 4.8 veranschaulicht diesen Sachverhalt für $n = 2$.

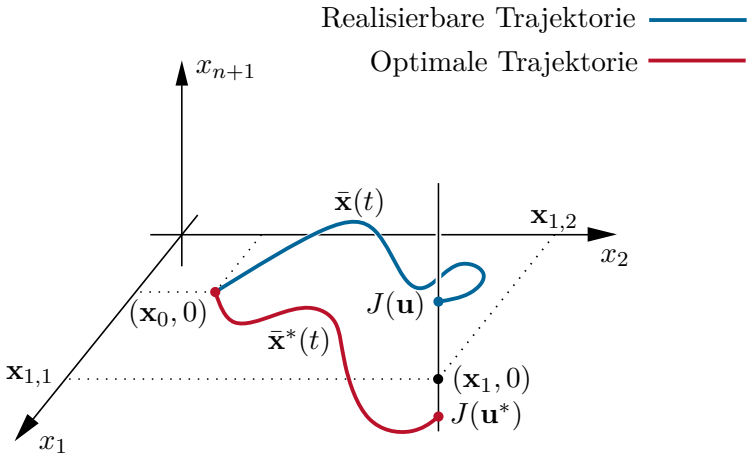


Abbildung 4.8: Zum Minimumsprinzip von Pontryagin.

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 (4.148), die die Bedingung $\mathbf{x}(t_1) = \mathbf{x}_1$ erfüllen (realisierbare Lösungen) und unterschiedliche Werte des Kostenfunktionals $x_{n+1}(t_1)$ 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 werden auch in der Herleitung des Minimumsprinzips von Pontryagin verwendet. Auf diese Herleitung wird hier verzichtet; sie wird z. B. in [4.13, 4.17] gezeigt.

Satz 4.11 (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)*

$$J(\mathbf{u}) = \int_{t_0}^{t_1} l(\mathbf{x}(t), \mathbf{u}(t))\,\mathrm{d}t \tag{4.149}$$

unter den Gleichungsbeschränkungen (dynamisches System)

$$\dot{\mathbf{x}} - \mathbf{f}(\mathbf{x}(t), \mathbf{u}(t)) = \mathbf{0}, \quad \mathbf{x}(t_0) = \mathbf{x}_0, \quad \mathbf{x}(t_1) = \mathbf{x}_1 \tag{4.150}$$

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* (4.150)*. Dann existiert ein* $\bar{\boldsymbol{\lambda}}^* = [\bar{\lambda}^*_1 \quad \dots \quad \bar{\lambda}^*_{n+1}]^{\mathrm{T}} \in (\hat{C}^1[t_0, t^*_1])^{n+1}$ *so, dass die Beziehung*

$$\dot{\bar{\boldsymbol{\lambda}}}^* = -\left(\frac{\partial}{\partial \bar{\mathbf{x}}} H\right)^{\mathrm{T}} \left(\mathbf{x}^*(t), \mathbf{u}^*(t), \bar{\boldsymbol{\lambda}}^*(t)\right) \tag{4.151}$$

für $t_0 \leq t \leq t_1$ *mit* $H\Big(\mathbf{x}^*(t), \mathbf{u}^*(t), \bar{\boldsymbol{\lambda}}^*(t)\Big) = \bar{\boldsymbol{\lambda}}^{*\mathrm{T}} \bar{\mathbf{f}}(\mathbf{x}^*(t), \mathbf{u}^*(t))$ *erfüllt ist, wobei* $\bar{\mathbf{f}}$ *gemäß* (4.148) *definiert ist, und folgende Eigenschaften gelten:*

(a) Die optimale Lösung $\mathbf{u}^*(t)$ *minimiert die Funktion* $H\Big(\mathbf{x}^*(t), \mathbf{u}(t), \bar{\boldsymbol{\lambda}}^*(t)\Big)$ *für alle Zeiten* $t_0 \leq t \leq t_1^*$ *in der Menge der Stellgrößenbeschränkungen* U*, d. h.*

$$H\Big(\mathbf{x}^*(t), \mathbf{v}, \bar{\boldsymbol{\lambda}}^*(t)\Big) \geq H\Big(\mathbf{x}^*(t), \mathbf{u}^*(t), \bar{\boldsymbol{\lambda}}^*(t)\Big), \quad \forall \mathbf{v} \in U. \tag{4.152}$$

(b) Es gilt für alle Zeiten $t_0 \leq t \leq t_1^*$

$$\bar{\boldsymbol{\lambda}}^*(t) \neq \mathbf{0} \tag{4.153a}$$

$$\bar{\lambda}_{n+1}^*(t) = \text{konst.} \geq 0 \tag{4.153b}$$

$$H\Big(\mathbf{x}^*(t), \mathbf{u}^*(t), \bar{\boldsymbol{\lambda}}^*(t)\Big) = \text{konst.} \tag{4.153c}$$

(c) Es gilt die folgende Transversalitätsbedingung *(für* t_1 *frei)*

$$H\Big(\mathbf{x}^*(t_1^*), \mathbf{u}^*(t_1^*), \bar{\boldsymbol{\lambda}}^*(t_1^*)\Big) = 0 \ . \tag{4.154}$$

Zur Berechnung der $m + 2n + 3$ *unbekannten 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 (4.152) 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 (4.154) und $2n + 2$ Differentialgleichungen für den erweiterten Zustand $\bar{\mathbf{x}}^*$ gemäß (4.148) und den erweiterten adjungierten Zustand $\bar{\boldsymbol{\lambda}}^*$ gemäß (4.151). Zu diesen Differentialgleichungen gehören $n + 1$ Anfangsbedingungen $\bar{\mathbf{x}}^*(t_0) = [\mathbf{x}_0^{\mathrm{T}} \quad 0]^{\mathrm{T}}$, n Endbedingungen $\mathbf{x}^*(t_1^*) = \mathbf{x}_1$ sowie die Bedingung $\bar{\lambda}_{n+1}^*(t_1^*) \geq 0$. Daraus können jedoch die unbekannten Größen $(\mathbf{u}^*(t), \bar{\mathbf{x}}^*(t), \bar{\boldsymbol{\lambda}}^*(t), t_1^*)$ noch nicht eindeutig bestimmt werden. Um dies zu sehen, werden die Zeilen 1 bis n und die Zeile $n + 1$ von (4.151) getrennt voneinander angeschrieben.

$$\dot{\boldsymbol{\lambda}}^* = -\left(\frac{\partial}{\partial \mathbf{x}} \mathbf{f}\right)^{\mathrm{T}} (\mathbf{x}^*(t), \mathbf{u}^*(t)) \boldsymbol{\lambda}^*(t) - \left(\frac{\partial}{\partial \mathbf{x}} l\right)^{\mathrm{T}} (\mathbf{x}^*(t), \mathbf{u}^*(t)) \bar{\lambda}_{n+1}^*(t) \tag{4.155a}$$

$$\dot{\bar{\lambda}}_{n+1}^* = 0 \tag{4.155b}$$

Für $\boldsymbol{\lambda}^*$ existieren keine Randbedingungen und für $\bar{\lambda}_{n+1}^*$ nur die Endbedingung

$$\bar{\lambda}_{n+1}^*(t_1^*) \geq 0 \ . \tag{4.156}$$

Ist $\bar{\boldsymbol{\lambda}}^*(t)$ eine Lösungstrajektorie von (4.155) und (4.156), so gilt mit jeder beliebigen Konstante $c > 0$, dass auch $c\bar{\boldsymbol{\lambda}}^*(t)$ diese Gleichungen erfüllt. Dabei sind $\bar{\mathbf{x}}^*(t)$ und $\mathbf{u}^*(t)$ unverändert, da in (4.152)

$$\arg\min_{\mathbf{u}} \bar{\boldsymbol{\lambda}}^{\mathrm{T}} \bar{\mathbf{f}} = \arg\min_{\mathbf{u}} c\bar{\boldsymbol{\lambda}}^{\mathrm{T}} \bar{\mathbf{f}} \tag{4.157}$$

gilt. Folglich ist die optimale Lösung gemäß Satz 4.11 nur bis auf einen multiplikativen Faktor bei $\bar{\boldsymbol{\lambda}}^*$ definiert. Es sind nun drei Fälle von extremalen Lösungen zu unterscheiden [4.18–4.20]:

(i) Ein Lösung $(\mathbf{u}^*(t), \mathbf{x}^*(t), t_1^*)$ wird als *normal* bezeichnet, wenn sie erlaubt die Bedingungen von Satz 4.11 mit $\bar{\lambda}_{n+1}^* > 0$ zu erfüllen. In diesem Fall kann aufgrund des uneindeutigen multiplikativen Faktors bei $\bar{\boldsymbol{\lambda}}^*$, ohne Beschränkung der Allgemeinheit, der (praktisch etablierte) Wert $\bar{\lambda}_{n+1}^* = 1$ verwendet werden. Diese Wahl liefert genau jene Hamiltonfunktion die auch schon in Abschnitt 4.2.3 verwendet wurde, siehe (4.102).

(ii) Ein Lösung $(\mathbf{u}^*(t), \mathbf{x}^*(t), t_1^*)$ wird als *abnormal* bezeichnet, wenn sie erlaubt die Bedingungen von Satz 4.11 mit $\bar{\lambda}_{n+1}^* = 0$ zu erfüllen. In diesem Fall hängen also die Hamiltonfunktion H und folglich auch die Minimierungsbedingung (4.152) nicht von der Lagrangeschen Dichte l ab.

(iii) Ein Lösung $(\mathbf{u}^*(t), \mathbf{x}^*(t), t_1^*)$ wird als *strikt abnormal* bezeichnet, wenn sie zur Erfüllung der Bedingungen von Satz 4.11 $\bar{\lambda}_{n+1}^* = 0$ erfordert.

Es können Lösungen $(\mathbf{u}^*(t), \mathbf{x}^*(t), t_1^*)$ existieren, die sowohl *normal* als auch *abnormal* sind. Lösungen die nur dem Fall (i) genügen werden dennoch einfach als *normal* bezeichnet. In Beispiel 4.13 wird sich zeigen, dass, trotz ihrer Bezeichnung und ihrer bemerkenswerten Unabhängigkeit von der Lagrangeschen Dichte l, *abnormale* und *strikt abnormale* Lösungen tatsächlich auch praktisch vorkommen können. Abgesehen von Beispiel 4.13 werden jedoch in dieser Vorlesung nur *normale* Lösungen behandelt.

Die notwendige Bedingung (4.152), also dass die Hamiltonfunktion $H(\mathbf{x}, \mathbf{u}, \bar{\boldsymbol{\lambda}})$ bezüglich $\mathbf{u}$ minimal ist, entspricht im Falle von nicht aktiven Beschränkungen der Stellgröße $\mathbf{u}$ den notwendigen Bedingungen erster und zweiter Ordnung (4.103c) und (4.105) aus der Variationsrechnung, d. h.

$$\left(\frac{\partial}{\partial \mathbf{u}} H\right)(\mathbf{x}^*, \mathbf{u}^*, \boldsymbol{\lambda}^*) = \mathbf{0} \tag{4.158a}$$

$$\mathbf{d}^{\mathrm{T}} \left(\frac{\partial^2}{\partial \mathbf{u}^2} H\right)(\mathbf{x}^*, \mathbf{u}^*, \boldsymbol{\lambda}^*)\mathbf{d} \geq 0 \quad \forall\, \mathbf{d} \in \mathbb{R}^m,\, t \in [t_0, t_1^*] \,. \tag{4.158b}$$

Trotz dieser Analogie ist das Minimumsprinzip von Pontryagin nach Satz 4.11 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 4.9.

Beispiel 4.11. Gesucht ist das Minimum des Kostenfunktionals

$$J(u) = \frac{1}{2} \int_0^1 u^2(t)\, \mathrm{d}t \tag{4.159}$$

für das dynamische System

$$\dot{x} = -x + u, \quad x(0) = 1, \quad x(1) = 0 \tag{4.160}$$

unter Berücksichtigung der Stellgrößenbeschränkung $-0.6 \leq u(t) \leq 0$ für alle $0 \leq t \leq 1$. Die Hamiltonfunktion H von Satz 4.11 für dieses Beispiel lautet

$$H(x, u, \bar{\boldsymbol{\lambda}}) = \bar{\lambda}_1(-x + u) + \bar{\lambda}_2 \frac{1}{2} u^2 \tag{4.161}$$

und die adjungierten Zustände $\bar{\boldsymbol{\lambda}}$ erfüllen gemäß (4.151) und (4.153b) die Gleichungen

$$\frac{\mathrm{d}}{\mathrm{d}t}\bar{\lambda}_1^* = -\left(\frac{\partial}{\partial x} H\right)\left(x^*, u^*, \bar{\boldsymbol{\lambda}}^*\right) = \bar{\lambda}_1^* \tag{4.162a}$$

$$\frac{\mathrm{d}}{\mathrm{d}t}\bar{\lambda}_2^* = 0 \ . \tag{4.162b}$$

Daraus folgt die Lösung

$$\bar{\lambda}_1^*(t) = C_1 \mathrm{e}^t \quad \text{und} \quad \bar{\lambda}_2^*(t) = C_2 \tag{4.163}$$

mit geeigneten Konstanten C_1 und $C_2 \geq 0$.

Zunächst wird $C_2 = 0$ angenommen (Prüfung ob *abnormaler* Fall vorliegt). Die optimale Lösung u^* folgt aus der Minimierungsbedingung (4.152)

$$H(x^*(t), v, \bar{\boldsymbol{\lambda}}^*(t)) \geq H(x^*(t), u^*(t), \bar{\boldsymbol{\lambda}}^*(t)) \qquad \forall\, v \in [-0.6, 0] \ , \ t \in [0, 1] \tag{4.164}$$

in der Form

$$u^*(t) = \begin{cases} 0 & \text{für} \quad \bar{\lambda}_1^* < 0 \\ -0.6 & \text{für} \quad \bar{\lambda}_1^* > 0 \ . \end{cases} \tag{4.165}$$

Da $\bar{\lambda}_1^*(t)$ gemäß (4.163) im Zeitraum $[0, 1]$ keinen Nulldurchgang besitzt, ist die optimale Stellgröße konstant mit dem Wert $u^* = 0$ oder $u^* = -0.6$. Aus einer einfachen Rechnung folgt, dass keiner dieser beiden Lösungskandidaten das Zweipunktrandwertproblem (4.160) löst (keine realisierbare Lösung). Folglich kann kein *abnormaler* Fall vorliegen und es wird im Weiteren $C_2 = 1$ verwendet (*normaler* Fall). Die Minimierungsbedingung (4.152) liefert nun unter Berücksichtigung von

$$\left(\frac{\partial}{\partial u} H\right)(x^*, u^*, \bar{\boldsymbol{\lambda}}^*) = \bar{\lambda}_1^* + \bar{\lambda}_2^* u^* = \bar{\lambda}_1^* + u^* \tag{4.166}$$

und der Stellgrößenbeschränkung $u(t) \in [-0.6, 0]$ die optimale Stellgröße

$$u^*(t) = \begin{cases} 0 & \text{für} \quad \bar{\lambda}_1^* \leq 0 \\ -\bar{\lambda}_1^* = -C_1 \mathrm{e}^t & \text{für} \quad 0 < \bar{\lambda}_1^* < 0.6 \\ -0.6 & \text{für} \quad \bar{\lambda}_1^* \geq 0.6 \ . \end{cases} \tag{4.167}$$

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 $x^*(1) = \mathrm{e}^{-1} \neq 0$ keine realisierbare Lösung resultiert. Mit $C_1 > 0$ bzw. $\bar{\lambda}_1^* > 0$ muss deshalb die optimale Stellgröße $u^*(t)$ zwischen $-C_1\mathrm{e}^t$ und -0.6 umschalten. Da $\bar{\lambda}_1^*(t) = C_1\mathrm{e}^t$ streng monoton steigend in t ist, setzt man eine stückweise stetige Steuerung

$$u^*(t) = \begin{cases} -C_1\mathrm{e}^t & \text{falls } t \in [0, c^*] \\ -0.6 & \text{falls } t \in (c^*, 1] \end{cases} \tag{4.168}$$

mit einem Umschaltzeitpunkt $t = c^*$ (Eckpunkt für x^*) an. Für das Zeitintervall $[0, c^*]$ errechnet sich die Lösung von

$$\dot{x}_{(1)}^*(t) = -x_{(1)}^*(t) + u^*(t), \quad x_{(1)}^*(0) = 1 \tag{4.169}$$

zu

$$x_{(1)}^*(t) = -\frac{1}{2}C_1\mathrm{e}^t + \mathrm{e}^{-t}\left(\frac{1}{2}C_1 + 1\right) \tag{4.170}$$

und für das Zeitintervall $[c^*, 1]$ folgt aus

$$\dot{x}_{(2)}^*(t) = -x_{(2)}^*(t) + u^*(t), \quad x_{(2)}^*(1) = 0 \tag{4.171}$$

die Lösung zu

$$x_{(2)}^*(t) = 0.6\left(\mathrm{e}^{1-t} - 1\right) . \tag{4.172}$$

Nach (4.153c) muss die Hamiltonfunktion $H(x^*(t), u^*(t), \bar{\boldsymbol{\lambda}}^*(t))$ im gesamten Zeitintervall konstant sein. Daraus folgt

$$\begin{aligned} &\bar{\lambda}_1^*(\tau_1)\left(-x_{(1)}^*(\tau_1) + u^*(\tau_1)\right) + \frac{1}{2}(u^*(\tau_1))^2 = \\ &\qquad \bar{\lambda}_1^*(\tau_2)\left(-x_{(2)}^*(\tau_2) + u^*(\tau_2)\right) + \frac{1}{2}(u^*(\tau_2))^2 \quad \forall\, \tau_1 \in [0, c^*],\ \tau_2 \in (c^*, 1] \end{aligned} \tag{4.173a}$$

und nach Einsetzen

$$-C_1\left(1 + \frac{1}{2}C_1\right) = -C_1 0.6\mathrm{e} + \frac{1}{2}0.6^2 . \tag{4.173b}$$

Hieraus ergeben sich die zwei möglichen Lösungen $C_{1,1} = 0.436$ und $C_{1,2} = 0.826$. Der Zeitpunkt der Umschaltung $t = c^*$ folgt aus der Stetigkeitsbedingung der Zustandsgröße

$$x_{(1)}^*(c^*) = x_{(2)}^*(c^*) \tag{4.174}$$

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 4.11 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

$$\psi_k(\mathbf{x}) = 0, \quad k = 1, \ldots, p \tag{4.175}$$

beschrieben. D. h. es gilt $\mathcal{X}_1 = \{\mathbf{x} \in \mathbb{R}^n \,|\, \psi_k(\mathbf{x}) = 0,\ k = 1, \ldots, p\}$. Der Tangentialraum $\mathcal{T}_{\check{\mathbf{x}}}\mathcal{X}_1$ an der Stelle $\mathbf{x} = \check{\mathbf{x}}$ ist dann in der Form

$$\mathcal{T}_{\check{\mathbf{x}}}\mathcal{X}_1 = \left\{ \mathbf{d} \,\middle|\, \left(\frac{\partial}{\partial \mathbf{x}}\psi_k(\check{\mathbf{x}})\right)\mathbf{d} = 0,\ k = 1, \ldots, p \right\} \tag{4.176}$$

definiert (siehe auch Abschnitt 3.1.2.1).

Satz 4.12 (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)*

$$J(\mathbf{u}) = \int_{t_0}^{t_1} l(\mathbf{x}(t), \mathbf{u}(t))\,\mathrm{d}t \tag{4.177}$$

unter den Gleichungsbeschränkungen (dynamisches System)

$$\dot{\mathbf{x}} - \mathbf{f}(\mathbf{x}(t), \mathbf{u}(t)) = \mathbf{0}, \quad \mathbf{x}(t_0) = \mathbf{x}_0, \quad \mathbf{x}(t_1) \in \mathcal{X}_1 \tag{4.178}$$

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* (4.178)*. Dann existiert ein* $\bar{\boldsymbol{\lambda}}^* = [\bar{\lambda}_1^* \quad \ldots \quad \bar{\lambda}_{n+1}^*]^{\mathrm{T}} \in (\hat{C}^1[t_0, t_1^*])^{n+1}$ *so, dass die Beziehungen* (4.151)*–*(4.154) *von Satz 4.11 erfüllt sind und* $\boldsymbol{\lambda}^*(t_1^*) = [\lambda_1^*(t_1^*) \quad \ldots \quad \lambda_n^*(t_1^*)]^{\mathrm{T}}$ orthogonal zum Tangentialraum $\mathcal{T}_{\mathbf{x}^*(t_1^*)}\mathcal{X}_1$ *ist, d. h. es gelten die* Transversalitätsbedingungen

$$(\boldsymbol{\lambda}^*)^{\mathrm{T}}(t_1^*)\mathbf{d} = 0, \quad \forall \mathbf{d} \in \mathcal{T}_{\mathbf{x}^*(t_1^*)}\mathcal{X}_1\ . \tag{4.179}$$

Nach Satz 4.12 und (4.176) muss $\boldsymbol{\lambda}^*(t_1^*)$ sich also als Linearkombination der Gradienten $\left(\frac{\partial}{\partial \mathbf{x}}\psi_k\right)(\mathbf{x}_1^*)$ mit $k = 1, \ldots, p$ darstellen lassen, d. h. in der Form

$$\boldsymbol{\lambda}^*(t_1^*) = \sum_{k=1}^{p} \mu_k \left(\frac{\partial}{\partial \mathbf{x}}\psi_k\right)^{\mathrm{T}}(\mathbf{x}_1^*), \quad \mathbf{x}_1^* = \mathbf{x}^*(t_1^*) \tag{4.180}$$

mit dem Lagrange-Multiplikator $\boldsymbol{\mu} = [\mu_1 \quad \ldots \quad \mu_p]^{\mathrm{T}} \in \mathbb{R}^p$. Die Bedingung

$$\operatorname{rang}\left(\left(\frac{\partial}{\partial \mathbf{x}}\boldsymbol{\psi}\right)(\mathbf{x}_1^*)\right) = p, \quad \boldsymbol{\psi} = \begin{bmatrix}\psi_1 & \ldots & \psi_p\end{bmatrix}^{\mathrm{T}} \tag{4.181}$$

entspricht der LICQ (linear independence constraint qualification) Bedingung der statischen Optimierung mit Gleichungsbeschränkungen, siehe auch Definition 3.2.

4.2.5 Anwendung des Minimumsprinzips auf zeitvariante Problemformulierungen

Für *zeitvariante* Lagrangesche Dichten oder *zeitvariante* Systeme der Art

$$J(\mathbf{u}) = \int_{t_0}^{t_1} l(t, \mathbf{x}(t), \mathbf{u}(t))\,\mathrm{d}t \tag{4.182}$$

$$\dot{\mathbf{x}} = \mathbf{f}(t, \mathbf{x}(t), \mathbf{u}(t)), \quad \mathbf{x}(t_0) = \mathbf{x}_0 \tag{4.183}$$

führt man eine weitere Zustandsgröße der Form $x_0 = t$ mit dem Anfangswert $x_0(t_0) = t_0$ ein und entwirft für das erweiterte System

$$J(\mathbf{u}) = \int_{t_0}^{t_1} \underbrace{l(x_0, \mathbf{x}(t), \mathbf{u}(t))}_{\tilde{l}(\tilde{\mathbf{x}}, \mathbf{u})}\,\mathrm{d}t \tag{4.184}$$

$$\frac{\mathrm{d}}{\mathrm{d}t}\underbrace{\begin{bmatrix} x_0 \\ \mathbf{x} \end{bmatrix}}_{\tilde{\mathbf{x}}} = \underbrace{\begin{bmatrix} 1 \\ \mathbf{f}(x_0, \mathbf{x}, \mathbf{u}) \end{bmatrix}}_{\tilde{\mathbf{f}}(\tilde{\mathbf{x}}, \mathbf{u})} \tag{4.185}$$

eine optimale Steuerung gemäß Satz 4.11 oder Satz 4.12. Dabei wird vorausgesetzt, dass $\mathbf{f}$ und l stetig differenzierbar in t sind.

4.2.6 Minimumsprinzip für eingangsaffine Systeme

Den weiteren Betrachtungen liege das *eingangsaffine System*

$$\dot{\mathbf{x}} = \mathbf{f}(\mathbf{x}, \mathbf{u}) = \mathbf{f}_0(\mathbf{x}) + \sum_{i=1}^{m} \mathbf{f}_i(\mathbf{x}) u_i \tag{4.186}$$

mit den Stellgrößenbeschränkungen der Form

$$\mathbf{u} \in U = [\mathbf{u}^-, \mathbf{u}^+] \quad \text{bzw.} \quad u_i \in [u_i^-, u_i^+], \quad i = 1, \ldots, m \tag{4.187}$$

(englisch: *box constraints*) zugrunde.

4.2.6.1 Kostenfunktional mit verbrauchsoptimalem Anteil

Im Zusammenhang mit dem Entwurf von *verbrauchsoptimalen Steuerungen* werden häufig Kostenfunktionale der Form

$$J(\mathbf{u}) = \int_{t_0}^{t_1} l_0(\mathbf{x}) + \sum_{i=1}^{m} r_i |u_i|\,\mathrm{d}t\,, \quad r_i > 0 \tag{4.188}$$

verwendet. Die zu (4.186) und (4.188) passende Hamiltonfunktion lautet (mit $\bar{\lambda}_{n+1} = 1$)

$$H(\mathbf{x}, \mathbf{u}, \boldsymbol{\lambda}) = l_0(\mathbf{x}) + \sum_{i=1}^{m} r_i |u_i| + \boldsymbol{\lambda}^{\mathrm{T}} \left(\mathbf{f}_0(\mathbf{x}) + \sum_{i=1}^{m} \mathbf{f}_i(\mathbf{x}) u_i \right). \tag{4.189}$$

Da der Anteil $l_0(\mathbf{x})+\boldsymbol{\lambda}^{\mathrm{T}}\mathbf{f}_0(\mathbf{x})$ in H unabhängig von $\mathbf{u}$ ist, kann er im Minimierungsproblem (4.152) vernachlässigt werden. Die Minimierung von H kann nun für jedes $u_i,\ i=1,\dots,m$, separat durchgeführt werden, d. h.

$$\min_{u_i\in[u_i^-,u_i^+]} H_i(u_i) = r_i|u_i| + q_i(\mathbf{x},\boldsymbol{\lambda})u_i, \quad q_i(\mathbf{x},\boldsymbol{\lambda}) = \boldsymbol{\lambda}^{\mathrm{T}}\mathbf{f}_i(\mathbf{x})\ . \tag{4.190}$$

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)–c) mit denen die optimale Stellgröße

$$u_i^* = \begin{cases} u_i^- & \text{falls}\ \ q_i(\mathbf{x},\boldsymbol{\lambda}) > r_i \\ 0 & \text{falls}\ \ q_i(\mathbf{x},\boldsymbol{\lambda}) \in (-r_i, r_i)\,, \\ u_i^+ & \text{falls}\ \ q_i(\mathbf{x},\boldsymbol{\lambda}) < -r_i \end{cases} \qquad \forall\, i=1,\dots,m \tag{4.191}$$

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 (4.190) bestimmbar. Dieser Fall wird als *singulär* bezeichnet und im Abschnitt 4.2.7 näher erläutert.

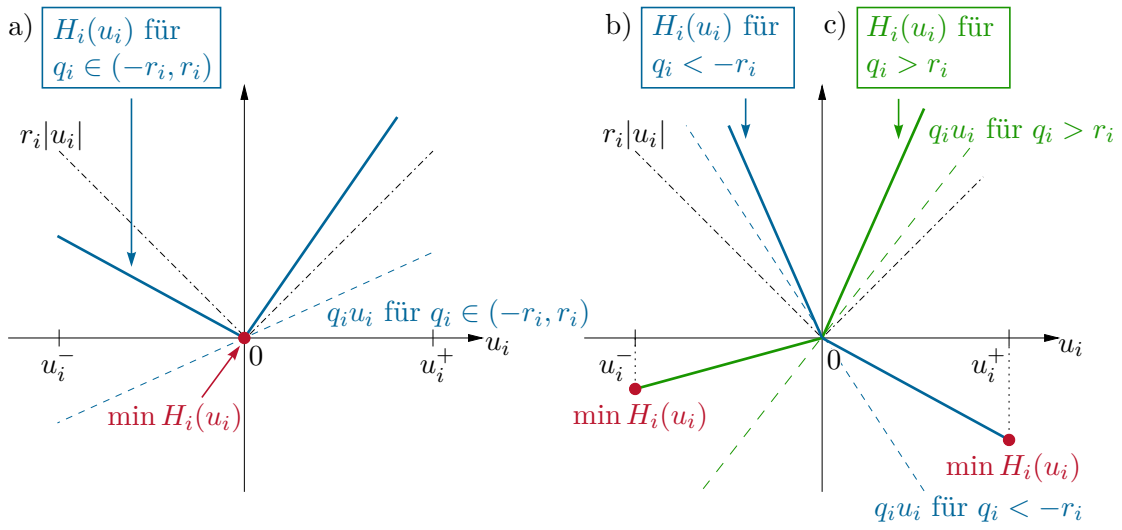


Abbildung 4.9: Verbrauchsoptimaler Fall, grafische Veranschaulichung von (4.190).

4.2.6.2 Kostenfunktional mit energieoptimalem Anteil

Unter dem Begriff *energieoptimale Steuerung* wird häufig die Minimierung eines Kostenfunktionals der Form

$$J(\mathbf{u}) = \int_{t_0}^{t_1} l_0(\mathbf{x}) + \frac{1}{2}\sum_{i=1}^{m} r_i u_i^2\ \mathrm{d}t\,, \quad r_i > 0 \tag{4.192}$$

verstanden. Analog zum vorherigen Fall kann die Minimierung der Hamiltonfunktion H wieder für jedes $u_i,\ i = 1, \ldots, m$, separat erfolgen, d. h.

$$\min_{u_i \in [u_i^-, u_i^+]} H_i(u_i) = \frac{1}{2} r_i u_i^2 + q_i(\mathbf{x}, \boldsymbol{\lambda}) u_i, \quad q_i(\mathbf{x}, \boldsymbol{\lambda}) = \boldsymbol{\lambda}^{\mathrm{T}} \mathbf{f}_i(\mathbf{x}) \ . \tag{4.193}$$

Durch den quadratischen Term mit $r_i > 0$ hätte die Funktion $H_i(u_i)$ im unbeschränkten Fall stets ein *Minimum* an der Stelle

$$u_i^0 = -\frac{1}{r_i} q_i(\mathbf{x}, \boldsymbol{\lambda}) \, . \tag{4.194}$$

Falls u_i^0 innerhalb des zulässigen Intervalls $[u_i^-, u_i^+]$ liegt, ist die optimale Lösung von (4.186), (4.187) und (4.192) durch $u_i^* = u_i^0$ gegeben. Falls u_i^0 links (rechts) von $[u_i^-, u_i^+]$ liegt, so befindet sich das Minimum von $H_i(u_i)$ an der Schranke u_i^- (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

$$u_i^* = \begin{cases} u_i^- & \text{falls} \quad u_i^0 \le u_i^- \\ u_i^0 & \text{falls} \quad u_i^0 \in (u_i^-, u_i^+) \, , \\ u_i^+ & \text{falls} \quad u_i^0 \ge u_i^+ \end{cases} \qquad i = 1, \ldots, m \, . \tag{4.195}$$

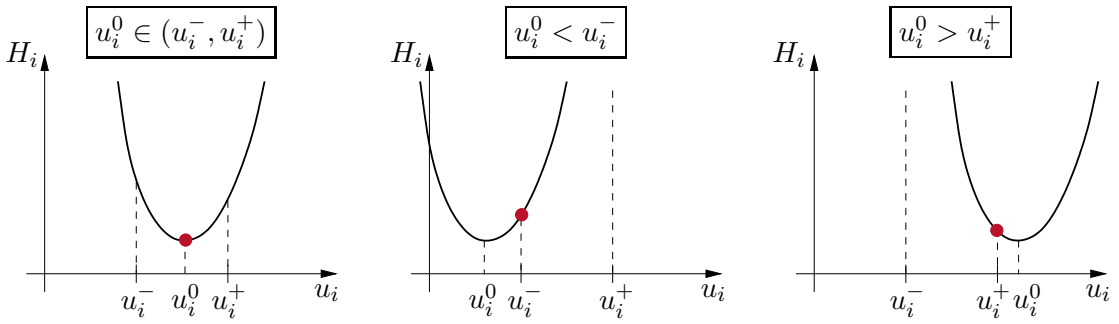


Abbildung 4.10: Energieoptimaler Fall, grafische Veranschaulichung von (4.193).

4.2.6.3 Zeitoptimales Kostenfunktional

Für zeitoptimale Probleme lautet das Kostenfunktional

$$J(\mathbf{u}) = \int_{t_0}^{t_1} 1 \ \mathrm{d}t = t_1 - t_0 \tag{4.196}$$

und die Hamiltonfunktion lässt sich in der Form (mit $\bar{\lambda}_{n+1} = 1$)

$$H(\mathbf{x}, \mathbf{u}, \boldsymbol{\lambda}) = 1 + \boldsymbol{\lambda}^{\mathrm{T}} \left(\mathbf{f}_0(\mathbf{x}) + \sum_{i=1}^{m} \mathbf{f}_i(\mathbf{x}) u_i \right) \tag{4.197}$$

anschreiben. Minimiert man die Hamiltonfunktion H wieder für jedes u_i, $i = 1, \dots, m$ separat

$$\min_{u_i \in [u_i^-, u_i^+]} H_i(u_i) = q_i(\mathbf{x}, \boldsymbol{\lambda}) u_i, \quad q_i(\mathbf{x}, \boldsymbol{\lambda}) = \boldsymbol{\lambda}^{\mathrm{T}} \mathbf{f}_i(\mathbf{x}) \,, \tag{4.198}$$

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

$$u_i^* = \begin{cases} u_i^- & \text{falls} \quad q_i(\mathbf{x}, \boldsymbol{\lambda}) > 0 \\ u_i^+ & \text{falls} \quad q_i(\mathbf{x}, \boldsymbol{\lambda}) < 0 \end{cases}, \quad i = 1, \dots, m \,. \tag{4.199}$$

Diese Steuerung wird häufig als *Bang-Bang-Steuerung* bezeichnet, da lediglich zwischen den Maximal- und Minimalwerten des Stellgrößenbereiches hin- und hergeschaltet wird. Ein singulärer Fall (siehe Abschnitt 4.2.7) 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 (4.198) ist damit zwar erfüllt, liefert aber keine Informationen über die Wahl von u_i. Alternativ zu der in Abschnitt 4.2.7 beschriebenen Vorgangsweise kann der singuläre Fall durch einen zusätzlichen *Regularisierungsterm*

$$J(\mathbf{u}) = \int_{t_0}^{t_1} 1 + \frac{1}{2} \sum_{i=1}^{m} r_i u_i^2 \, \mathrm{d}t \,, \quad r_i > 0 \tag{4.200}$$

vermieden werden, wobei r_i hinreichend klein gewählt wird, um annähernd Zeitoptimalität zu erzielen. Der Regularisierungsterm entspricht natürlich einem energieoptimalen Anteil, so dass (4.200) die Form (4.192) besitzt und die optimale Stellgröße $\mathbf{u}^*$ gemäß (4.195) berechnet werden kann.

Beispiel 4.12 (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 $\mathbf{x} = [x_1 \quad x_2]^{\mathrm{T}}$ kann wie folgt formuliert werden

$$\min_{u(\cdot), t_1} \quad \int_{t_0=0}^{t_1} \mathrm{d}t = t_1 \tag{4.201a}$$

$$\text{u.B.v.} \quad \dot{x}_1 = x_2, \quad \dot{x}_2 = u \tag{4.201b}$$

$$\mathbf{x}(0) = \mathbf{x}_0 \,, \quad \mathbf{x}(t_1) = \mathbf{0} \tag{4.201c}$$

$$|u(t)| \leq 1 \quad \forall t \in [0, t_1] \,. \tag{4.201d}$$

Mit der Hamiltonfunktion ($\bar{\lambda}_3 = 1$) $H(\mathbf{x}, u, \boldsymbol{\lambda}) = 1 + \lambda_1 x_2 + \lambda_2 u$ ergeben sich die adjungierten Zustände $\boldsymbol{\lambda}^* = [\lambda_1^* \quad \lambda_2^*]^{\mathrm{T}}$ aus (4.151) zu

$$\dot{\lambda}_1^* = 0 \quad \Rightarrow \quad \lambda_1^*(t) = C_1 \tag{4.202a}$$

$$\dot{\lambda}_2^* = -\lambda_1^* \quad \Rightarrow \quad \lambda_2^*(t) = -C_1 t + C_2 \,, \tag{4.202b}$$

wobei C_1 und C_2 Integrationskonstanten darstellen. Die Minimierungsbedingung (4.152) für die Hamiltonfunktion führt auf die optimale Stellgröße

$$u^*(t) = \begin{cases} +1 & \text{falls} \quad \lambda_2^* < 0 \\ -1 & \text{falls} \quad \lambda_2^* > 0\,. \end{cases} \tag{4.203}$$

Der *singuläre Fall*, d. h. $\lambda_2^*(t) = 0$ auf einem nicht verschwindenden Subintervall $t \in I_s \subseteq [0, t_1]$, kann hier nicht auftreten, da dann aufgrund von (4.202) $\lambda_2^*(t) = 0$ und $\lambda_1^*(t) = 0$ auf dem Gesamtintervall $[0, t_1]$ gelten müsste. Dies widerspricht aber der Transversalitätsbedingung (4.154) für die *freie Endzeit* t_1

$$H(\mathbf{x}^*, u^*, \boldsymbol{\lambda}^*)\big|_{t=t_1^*} = 1 + \lambda_1^*(t_1^*)x_2^*(t_1^*) + \lambda_2^*(t_1^*)\,u^*(t_1^*) = 0\,. \tag{4.204}$$

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 Zustandstrajektorien in der (x_1, x_2)-Ebene *Parabeln* dar.

> *Aufgabe 4.8.* Zeigen Sie, dass die Lösung von (4.201b) für $u(t) = \pm 1 = \text{konst.}$ die folgende Parabelgleichung erfüllt
>
> $$x_1 = \frac{x_2^2}{2u} + c\,, \quad u = \pm 1\,. \tag{4.205}$$

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 $u = 1$ und $x_2 < 0$ bzw. $x_1 = -x_2^2/2$ für $u = -1$ und $x_2 > 0$ erreichbar, siehe Abbildung 4.11. Diese beiden Fälle können mit der Funktion

$$S(x_2) = -\frac{1}{2} x_2 |x_2| \tag{4.206}$$

unterschieden werden. Da lediglich die Schaltsequenzen $\{+1\}$, $\{-1\}$, $\{+1, -1\}$, $\{-1, +1\}$ für u in Frage kommen, gibt es jeweils nur eine Möglichkeit, um den Systemzustand mit $\mathbf{x}(0) = \mathbf{x}_0 = [x_{1,0} \quad x_{2,0}]^{\mathrm{T}}$ 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.

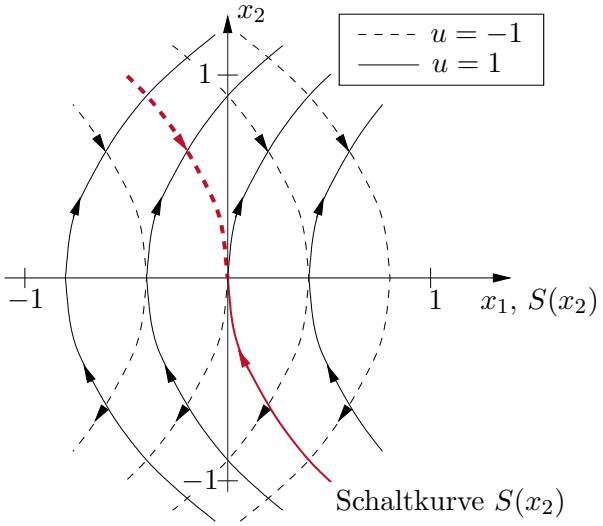


Abbildung 4.11: Mögliche Trajektorien des Doppelintegrators in der (x_1, x_2)-Ebene.

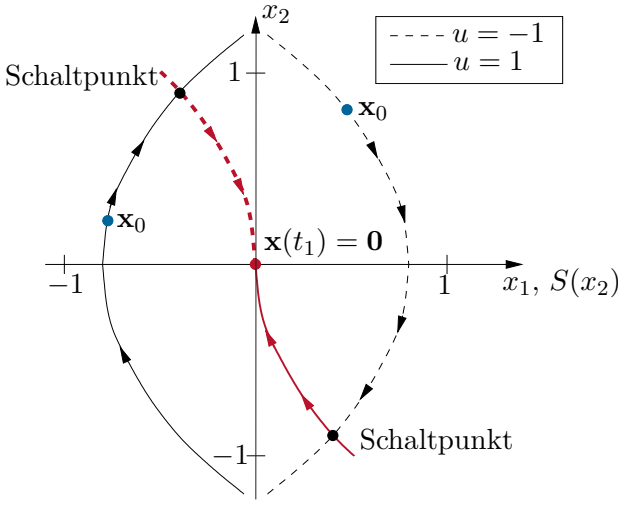


Abbildung 4.12: Optimale Umschaltung für verschiedene Anfangspunkte $\mathbf{x}_0$.

Das *optimale Stellgesetz* lautet also

$$u^*(t) = \begin{cases} +1 & \text{falls} \quad x_1 < S(x_2) \\ +1 & \text{falls} \quad x_1 = S(x_2) \quad \text{und} \quad x_1 > 0 \\ -1 & \text{falls} \quad x_1 > S(x_2) \\ -1 & \text{falls} \quad x_1 = S(x_2) \quad \text{und} \quad x_1 < 0\ . \end{cases} \tag{4.207}$$

Daraus können auch der Schaltzeitpunkt t_s und die minimale Endzeit t_1^* berechnet werden.

Aufgabe 4.9. Verifizieren Sie, dass der optimale Umschaltzeitpunkt t_s und die minimale Endzeit t_1^* wie folgt definiert sind

$$t_s = \begin{cases} x_{2,0} + \sqrt{\frac{1}{2}x_{2,0}^2 + x_{1,0}} & \text{falls} \quad x_{1,0} > S(x_{2,0}) \\ -x_{2,0} + \sqrt{\frac{1}{2}x_{2,0}^2 - x_{1,0}} & \text{falls} \quad x_{1,0} < S(x_{2,0}) \end{cases} \tag{4.208}$$

$$t_1^* = \begin{cases} x_{2,0} + \sqrt{2x_{2,0}^2 + 4x_{1,0}} & \text{falls} \quad x_{1,0} > S(x_{2,0}) \\ -x_{2,0} + \sqrt{2x_{2,0}^2 - 4x_{1,0}} & \text{falls} \quad x_{1,0} < S(x_{2,0}) \\ |x_{2,0}| & \text{falls} \quad x_{1,0} = S(x_{2,0}). \end{cases} \tag{4.209}$$

Abbildung 4.13 zeigt die zeitoptimalen Trajektorien für den Doppelintegrator für verschiedene Anfangswerte $\mathbf{x}_0 = [x_{1,0} \quad x_{2,0}]^{\mathrm{T}}$.

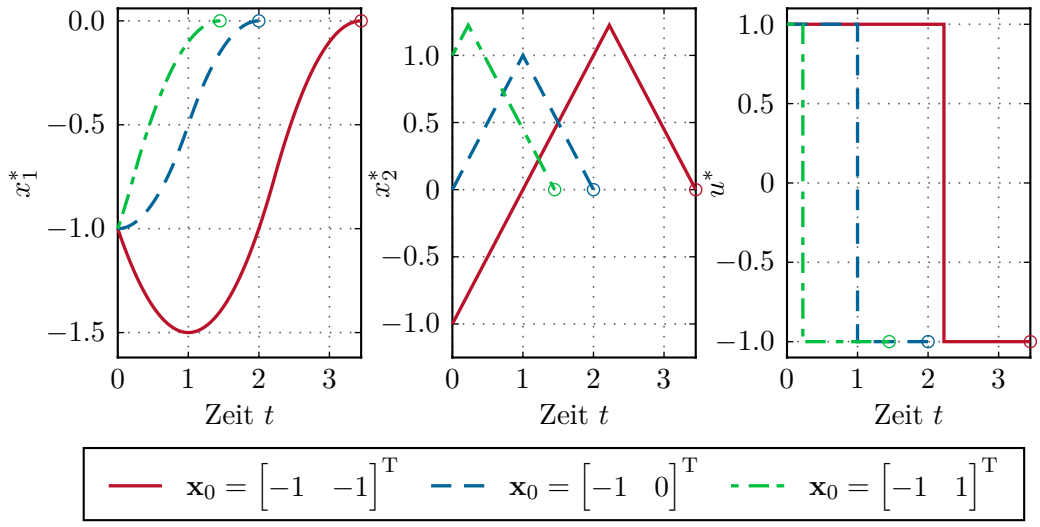


Abbildung 4.13: Zeitoptimale Trajektorien des Doppelintegrators für verschiedene Anfangswerte $\mathbf{x}_0$.

Aufgabe 4.10. Implementieren Sie das System (4.201b) mit dem optimalen Stellgesetz (4.207) in MATLAB/SIMULINK und verifizieren Sie die Ergebnisse in Abbildung 4.13 für verschiedene Anfangswerte $\mathbf{x}_0$. Verwenden Sie t_1^* gemäß (4.209) als Zeithorizont für die Simulation.

Beispiel 4.13 (Harmonischer Oszillator). Der harmonische Oszillator

$$\dot{x}_1 = x_2 \ , \quad \dot{x}_2 = -x_1 + u \tag{4.210}$$

mit dem Zustand $\mathbf{x} = [x_1 \quad x_2]^{\mathrm{T}}$ soll bei beschränktem Eingang $u \in [-1, 1]$ ausgehend vom Anfangszustand $\mathbf{x}(t_0) = \mathbf{0}$ mit $t_0 = 0$ in minimaler Zeit in einen gegebenen Endzustand $\mathbf{x}(t_1) = \mathbf{x}_1$ übergeführt werden. Das Optimalsteuerungsproblem kann daher wie folgt formuliert werden

$$\min_{u(\cdot), t_1} \quad \int_{t_0=0}^{t_1} \mathrm{d}t = t_1 \tag{4.211a}$$

$$\text{u.B.v.} \quad \dot{x}_1 = x_2 \ , \quad \dot{x}_2 = -x_1 + u \tag{4.211b}$$

$$\mathbf{x}(0) = \mathbf{0} \ , \quad \mathbf{x}(t_1) = \mathbf{x}_1 \tag{4.211c}$$

$$|u(t)| \leq 1 \quad \forall t \in [0, t_1] \ . \tag{4.211d}$$

Mit der Hamiltonfunktion $H(\mathbf{x}, u, \bar{\boldsymbol{\lambda}}) = \bar{\lambda}_1 x_2 + \bar{\lambda}_2(-x_1 + u) + \bar{\lambda}_3$ folgt die Differentialgleichung

$$\dot{\bar{\lambda}}_1^* = \bar{\lambda}_2^* \ , \quad \dot{\bar{\lambda}}_2^* = -\bar{\lambda}_1^* \tag{4.212}$$

für die adjungierten Zustände $\bar{\lambda}_1^*$ und $\bar{\lambda}_2^*$. Eine allgemeine Lösung dieser Gleichung lautet

$$\bar{\lambda}_1^*(t) = -A_1 \cos(t) + A_2 \sin(t) \ , \quad \bar{\lambda}_2^*(t) = A_1 \sin(t) + A_2 \cos(t) \tag{4.213}$$

mit den noch zu bestimmenden Konstanten A_1 und A_2. Die Minimierungsbedingung (4.152) für die Hamiltonfunktion führt auf die optimale Stellgröße

$$u^*(t) = \begin{cases} +1 & \text{falls} \quad \bar{\lambda}_2^* < 0 \\ -1 & \text{falls} \quad \bar{\lambda}_2^* > 0 \ . \end{cases} \tag{4.214}$$

Die Stellgröße wird also bei Nulldurchgängen von $\bar{\lambda}_2^*(t)$ umgeschaltet. Aus (4.213) folgt, dass diese Nulldurchgänge in regelmäßigen Zeitabständen von π auftreten. Wie viele solcher Nulldurchgänge auftreten, steht zunächst noch nicht fest.

Der *singuläre Fall*, d. h. $\bar{\lambda}_2^*(t) = 0$ auf einem nicht verschwindenden Subintervall $t \in I_s \subseteq [0, t_1]$, kann hier nicht auftreten, da dann aufgrund von (4.212) $\bar{\lambda}_2^*(t) = 0$ und $\bar{\lambda}_1^*(t) = 0$ auf dem Gesamtintervall $[0, t_1]$ gelten müsste. Die Transversalitätsbedingung (4.154) für die *freie Endzeit* t_1, d. h.

$$H(\mathbf{x}^*, u^*, \boldsymbol{\lambda}^*)\big|_{t=t_1^*} = \bar{\lambda}_1^*(t_1^*) x_2^*(t_1^*) + \bar{\lambda}_2^*(t_1^*)(-x_1^*(t_1^*) + u^*(t_1^*)) + \bar{\lambda}_3^* = \bar{\lambda}_3^* = 0 \tag{4.215}$$

würde dann jedoch auf $\bar{\boldsymbol{\lambda}}^*(t_1^*) = \bar{\boldsymbol{\lambda}}^*(t) = \mathbf{0}$ führen, was die Bedingung (4.153a) verletzt.

In Zeitintervallen mit der Stellgröße $u^*(t) = \pm 1$ sind die optimalen Zustandstrajektorien durch die Differentialgleichung

$$\dot{x}_1^* = x_2^* \ , \quad \dot{x}_2^* = -x_1^* \pm 1 \tag{4.216}$$

definiert. Im Zustandsraum beschreiben diese Trajektorien im Uhrzeigersinn durchlaufene Kreisbögen mit dem Mittelpunkt $(\pm 1, 0)$. Da die optimale Stellgröße in Zeitabständen von π zwischen 1 und -1 wechselt, muss die optimale Zustandstrajektorie stückweise aus Halbkreisbögen mit den abwechselnden Mittelpunkten $(1, 0)$ und $(-1, 0)$ zusammengesetzt werden. Am Beginn und Ende der optimalen Zustandstrajektorie können auch Kreisbögen mit kürzerer Länge (Zeitdauer) auftreten.

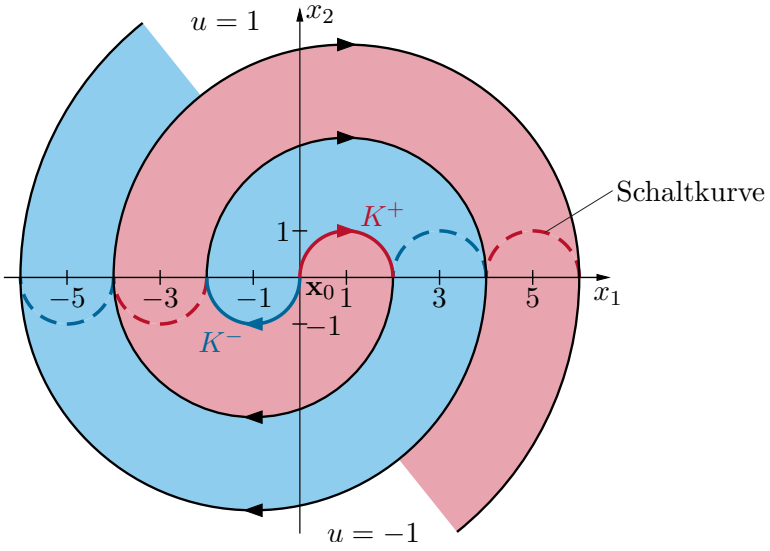


Abbildung 4.14: Mögliche Trajektorien des harmonischen Oszillators in der (x_1, x_2)-Ebene.

Einige dieser Kreisbögen sind in Abbildung 4.14 als durchgezogene Linien skizziert. Liegt der Endpunkt $\mathbf{x}_1$ im roten Gebiet, so beginnt die optimale Eingangstrajektorie mit 1 und die optimale Zustandstrajektorie folgt zunächst dem roten Kreisbogen K^+ und verbleibt dann im rot dargestellten Gebiet. Liegt der Endpunkt $\mathbf{x}_1$ im blauen Gebiet, so beginnt die optimale Eingangstrajektorie mit -1 und die optimale Zustandstrajektorie folgt zunächst dem blauen Kreisbogen K^- und verbleibt dann im blau dargestellten Gebiet. Beim Verlassen des ersten Kreisbogens K^+ oder K^- sowie jeweils beim Kreuzen einer strichlierten Kurve wird die Stellgröße umgeschaltet. Insgesamt ergibt sich damit die rot und blau dargestellte Schaltkurve, welche in Zeitabständen von π gekreuzt wird. Oberhalb dieser Kurve beträgt die Stellgröße 1, unterhalb -1.

Gemäß Satz 4.11 gilt $H(\mathbf{x}, u, \bar{\boldsymbol{\lambda}}) = \text{konst.}$ und da t_1 frei ist, muss zusätzlich $H(\mathbf{x}(t_1), u(t_1), \bar{\boldsymbol{\lambda}}(t_1)) = 0$ gelten. Es können nun drei Fälle unterschieden werden:

1. Liegt der Endpunkt $\mathbf{x}_1$ auf einer der in Abbildung 4.14 schwarz dargestellten Kreisbögen, so muss die optimale Zustandstrajektorie den ersten Kreisbogen K^- oder K^+ komplett durchlaufen und die Stellgrößenumschaltungen erfolgen genau zu den Zeitpunkten $t = \pi, 2\pi, 3\pi, \ldots$, was $\bar{\lambda}_2^*(t) = A_1 \sin(t)$ $(A_2 = 0)$ mit $A_1 \neq 0$ erfordert. Eine Auswertung der Hamiltonfunktion zum Zeitpunkt $t_0 = 0$

liefert daher $H(\mathbf{x}(t_0), u(t_0), \bar{\boldsymbol{\lambda}}(t_0)) = \bar{\lambda}_3^* = 0$. Es liegt also ein *strikt abnormaler* Fall vor.

2. Liegt der Endpunkt $\mathbf{x}_1$ auf der Kurve $K^+ \cup K^- \setminus \{(-2,0),(0,0),(2,0)\}$, so gilt $t_1 \in (0,\pi)$ und die optimale Stellgröße muss nie umgeschaltet werden. Damit sind die Konstanten A_1 und A_2 nur insoweit definiert als $\bar{\lambda}_2^*(t)$ im Zeitintervall $(0,\pi)$ keinen Nulldurchgang aufweisen darf. Eine Auswertung der Hamiltonfunktion zum Zeitpunkt $t_0 = 0$ liefert daher $H(\mathbf{x}(t_0), u(t_0), \bar{\boldsymbol{\lambda}}(t_0)) = \bar{\lambda}_3^* + \bar{\lambda}_2^*(t_0)u^*(t_0) = 0$. Es ist also möglich, aber nicht zwingend, $\bar{\lambda}_3^* = 0$ zu setzen. Es liegt folglich je nach Wahl von $\bar{\lambda}_3^*$ ein *abnormaler* oder *normaler* Fall vor.

3. Liegt der Endpunkt $\mathbf{x}_1$ abseits aller in Abbildung 4.14 durchgezogen dargestellten Kreisbögen, so gilt $\bar{\lambda}_3^* > 0$. Es liegt also jedenfalls ein *normaler* Fall vor und ohne Einschränkungen kann $\bar{\lambda}_3^* = 1$ gewählt werden.

Aufgabe 4.11. Berechnen Sie für den Endpunkt $\mathbf{x}_1 = [1 \quad 0]^\mathrm{T}$ die optimale Lösung sowie den Umschaltzeitpunkt und t_1. Zeigen Sie, dass es sich um einen normalen Fall handelt.

4.2.7 Der singuläre Fall

Wenn auf einem endlichen Subintervall $I_s \subseteq [t_0, t_1]$ die optimale Stellgröße $\mathbf{u}^*$ nicht oder nicht vollständig aus der Minimierungsbedingung (4.152) bestimmt werden kann, so liegt ein singulärer Fall vor. Zur Verdeutlichung dieser Problematik soll im Weiteren das Optimalsteuerungsproblem

$$\min_{u(\cdot)} \quad J(u) = \int_{t_0}^{t_1} l_0(\mathbf{x}) + l_1(\mathbf{x})u \ \mathrm{d}t \tag{4.217a}$$

$$\text{u.B.v.} \quad \dot{\mathbf{x}} = \mathbf{f}_0(\mathbf{x}) + \mathbf{f}_1(\mathbf{x})u, \quad \mathbf{x}(t_0) = \mathbf{x}_0 \tag{4.217b}$$

$$u \in \hat{C}_U[t_0, t_1] \tag{4.217c}$$

mit skalarer und affin auftretender Stellgröße $u(t)$ betrachtet werden. Die grundsätzliche Vorgehensweise ist aber auch auf allgemeinere Optimalsteuerungsprobleme anwendbar. Die Hamiltonfunktion

$$H(\mathbf{x}, u, \boldsymbol{\lambda}) = l_0(\mathbf{x}) + \boldsymbol{\lambda}^\mathrm{T}\mathbf{f}_0(\mathbf{x}) + \Big(l_1(\mathbf{x}) + \boldsymbol{\lambda}^\mathrm{T}\mathbf{f}_1(\mathbf{x})\Big)u \tag{4.218}$$

ist affin in u. Die Funktion

$$\zeta(t) = \left(\frac{\partial}{\partial u}H\right)(\mathbf{x}, u, \boldsymbol{\lambda}) = l_1(\mathbf{x}) + \boldsymbol{\lambda}^\mathrm{T}\mathbf{f}_1(\mathbf{x}) \tag{4.219}$$

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 unabhängig von $u^*(t)$ die Bedingung $\zeta^*(t) = 0$ auf einem endlichen Zeitintervall $I_s \subseteq [t_0, t_1]$ gilt, so liefert die Minimierungsbedingung (4.152) 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

$$\frac{\mathrm{d}^k}{\mathrm{d}t^k}(\zeta^*(t)) = 0 \qquad \forall\, k \in \mathbb{N},\, t \in I_s\ . \tag{4.220}$$

Entlang des singulären Pfades ist also die Minimierungsbedingung (4.152) für alle zulässigen Stellgrößen erfüllt, d. h. es gilt nicht nur $\frac{\partial H}{\partial u} = \zeta^*(t) = 0$, sondern auch $\frac{\partial^2 H}{\partial u^2} = \frac{\partial \zeta^*}{\partial u} = 0$. Um dennoch eine optimale Stellgröße $u^*(t)$ ermitteln zu können, sucht man die *kleinste positive natürliche Zahl* $\bar{k}$ so, dass gilt

$$\frac{\partial}{\partial u}\frac{\mathrm{d}^{\bar{k}}}{\mathrm{d}t^{\bar{k}}}(\zeta^*(t)) \neq 0\ . \tag{4.221}$$

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

$$\frac{\mathrm{d}^k}{\mathrm{d}t^k}(\zeta^*(t)) = 0 \qquad \forall\, k = 0, \dots, 2p-1 \tag{4.222}$$

zu liegen kommen. Ähnlich der Legendre-Clebsch Bedingung (4.105) muss entlang eines singulären Pfades für alle Zeiten $t \in I_s$ die sogenannte *generalisierte Legendre-Clebsch Bedingung*

$$(-1)^p \frac{\partial}{\partial u}\frac{\mathrm{d}^{2p}}{\mathrm{d}t^{2p}}(\zeta^*(t)) \geq 0 \tag{4.223}$$

erfüllt sein, vgl. [4.21].

Beispiel 4.14. Gesucht ist das Minimum des Kostenfunktionals

$$J(u) = \frac{1}{2}\int_0^2 x_1^2(t)\,\mathrm{d}t \tag{4.224}$$

für das dynamische System

$$\dot{x}_1 = x_2 + u \qquad x_1(0) = 1 \qquad x_1(2) = 0 \tag{4.225a}$$
$$\dot{x}_2 = -u \qquad x_2(0) = 1 \qquad x_2(2) = 0 \tag{4.225b}$$

unter Berücksichtigung der Stellgrößenbeschränkung $-10 \leq u(t) \leq 10$ für alle $t \in [0, 2]$. Die Hamiltonfunktion H für dieses Beispiel lautet (mit $\bar{\lambda}_{n+1} = \bar{\lambda}_3 = 1$)

$$H(\mathbf{x}, u, \boldsymbol{\lambda}) = \lambda_1(x_2 + u) - \lambda_2 u + \frac{1}{2}x_1^2 \tag{4.226}$$

und die adjungierten Zustände $\boldsymbol{\lambda}$ erfüllen gemäß (4.151) die Gleichungen

$$\frac{\mathrm{d}}{\mathrm{d}t}\lambda_1^* = -\left(\frac{\partial}{\partial x_1}H\right)(\mathbf{x}^*, u^*, \boldsymbol{\lambda}^*) = -x_1^* \tag{4.227a}$$
$$\frac{\mathrm{d}}{\mathrm{d}t}\lambda_2^* = -\left(\frac{\partial}{\partial x_2}H\right)(\mathbf{x}^*, u^*, \boldsymbol{\lambda}^*) = -\lambda_1^*\ . \tag{4.227b}$$

Die optimale Lösung u^* mit $-10 \leq u^* \leq 10$ muss der Ungleichung (4.152)

$$H(\mathbf{x}^*(t), v, \boldsymbol{\lambda}^*(t)) \geq H(\mathbf{x}^*(t), u^*(t), \boldsymbol{\lambda}^*(t)), \quad \forall v \in [-10, 10] \tag{4.228}$$

genügen. Damit folgt zunächst für die optimale Stellgröße unter Berücksichtigung der Stellgrößenbeschränkung

$$u^*(t) = \begin{cases} 10 & \text{für} \quad \lambda_1^* < \lambda_2^* \\ -10 & \text{für} \quad \lambda_1^* > \lambda_2^* \end{cases}. \tag{4.229}$$

Für $\lambda_1^* = \lambda_2^*$ tritt ein *singulärer Pfad* auf. Eine Auswertung von (4.221) liefert

$$\left(\frac{\partial}{\partial u} H\right)(\mathbf{x}^*, u, \boldsymbol{\lambda}^*) = \lambda_1^* - \lambda_2^* = 0 \tag{4.230a}$$

$$\left(\frac{\mathrm{d}}{\mathrm{d}t}\frac{\partial}{\partial u} H\right)(\mathbf{x}^*, u, \boldsymbol{\lambda}^*) = \frac{\mathrm{d}}{\mathrm{d}t}\lambda_1^* - \frac{\mathrm{d}}{\mathrm{d}t}\lambda_2^* = -x_1^* + \lambda_1^* = 0 \tag{4.230b}$$

$$\left(\frac{\mathrm{d}^2}{\mathrm{d}t^2}\frac{\partial}{\partial u} H\right)(\mathbf{x}^*, u, \boldsymbol{\lambda}^*) = -x_2^* - u^* - x_1^* = 0 \tag{4.230c}$$

$$\left(\frac{\partial}{\partial u}\frac{\mathrm{d}^2}{\mathrm{d}t^2}\frac{\partial}{\partial u} H\right)(\mathbf{x}^*, u, \boldsymbol{\lambda}^*) = -1 \neq 0 \tag{4.230d}$$

und damit die Ordnung $p = 1$ des singulären Pfades. Aus (4.230c) folgt

$$u^* = -x_2^* - x_1^* \tag{4.231}$$

und aus (4.230d) folgt

$$(-1)^1 \left(\frac{\partial}{\partial u}\frac{\mathrm{d}^2}{\mathrm{d}t^2}\frac{\partial}{\partial u} H\right)(\mathbf{x}^*, u, \boldsymbol{\lambda}^*) = 1 > 0 \,. \tag{4.232}$$

Dies zeigt, dass die generalisierte Legendre-Clebsch Bedingung (4.223) hier erfüllt ist. Gemäß (4.153c) muss die Hamiltonfunktion $H(\mathbf{x}^*(t), u^*(t), \boldsymbol{\lambda}^*(t))$ im gesamten Zeitintervall konstant sein, d. h.

$$\lambda_1^* x_2^* + \frac{1}{2}(x_1^*)^2 + (\lambda_1^* - \lambda_2^*)u^* = C = \text{konst.} \tag{4.233}$$

Entlang des singulären Pfades müssen $\mathbf{x}^*(t)$ und $\boldsymbol{\lambda}^*(t)$ auf der durch (4.230a) und (4.230b) definierten Mannigfaltigkeit $\lambda_1^* = \lambda_2^* = x_1^*$ liegen (siehe auch (4.222)), weshalb sich für diesen Fall (4.233) zu

$$x_1^* x_2^* + \frac{1}{2}(x_1^*)^2 = C \tag{4.234}$$

vereinfacht.

Aufgabe 4.12. Plausibilisieren Sie, dass die optimale Stellgröße durch

$$u^*(t) = \begin{cases} 10 & \text{für} \quad 0 \le t \le 0.299 \\ -x_2^* - x_1^* & \text{für} \quad 0.299 < t < 1.927 \\ -10 & \text{für} \quad 1.927 \le t \le 2 \end{cases} \tag{4.235}$$

gegeben ist.

Aufgabe 4.13. 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 4.15 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. $|\,\mathrm{d}h/\,\mathrm{d}y| \le s$. Wird das Gelände für den Straßenbau aufgeschüttet, so ergeben sich Kosten in Höhe von $k^+ > 0$ je aufgeschütteter Höheneinheit und je Längeneinheit y. Wird das Gelände für den Straßenbau abgegraben, so ergeben sich Kosten in Höhe von $k^- > 0$ je abgegrabener Höheneinheit und je Längeneinheit y. Brücken oder Tunnels sollen nicht gebaut werden.

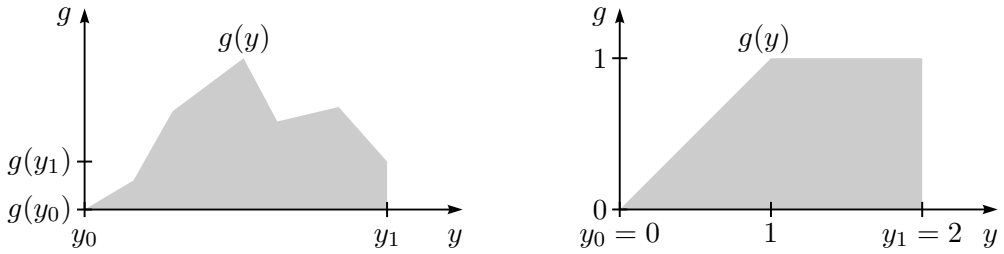


Abbildung 4.15: Geländeprofile.

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 4.15 skizzierte Geländeprofil $g(y)$ und die Parameterwerte $s = 1/2$ und $k^+ = k^- = 1$.

Lösung von Aufgabe 4.13. Unter Verwendung der Sprungfunktion

$$\sigma(t) = \begin{cases} 1 & \text{falls } t > 0 \\ 0 & \text{falls } t \le 0 \end{cases} \tag{4.236}$$

ergeben sich die Gesamtkosten für den Straßenbau in der Form

$$\int_{y_0}^{y_1} (h - g)\Big(k^+\sigma(h - g) - k^-\sigma(g - h)\Big)\,\mathrm{d}y\ . \tag{4.237}$$

Dementsprechend muss die Optimierungsaufgabe

$$\min_{u(\cdot)} \quad \int_{y_0}^{y_1} (h-g)\Big(k^+\sigma(h-g) - k^-\sigma(g-h)\Big)\,\mathrm{d}y \tag{4.238a}$$

$$\text{u.B.v.} \quad \frac{\mathrm{d}h}{\mathrm{d}y} = u \tag{4.238b}$$

$$-s \leq u \leq s \tag{4.238c}$$

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 $\mathbf{x} = [h \quad y]^{\mathrm{T}}$ und das erweiterte dynamische System

$$\frac{\mathrm{d}\mathbf{x}}{\mathrm{d}y} = \begin{bmatrix} u \\ 1 \end{bmatrix} \tag{4.239}$$

mit dem Anfangszustand $x_2(y_0) = y_0$ verwendet. Mit $\bar{\lambda}_{n+1} = 1$ (normaler Fall) folgt die Hamiltonfunktion

$$H(\mathbf{x}, u, \boldsymbol{\lambda}) = (h-g)\Big(k^+\sigma(h-g) - k^-\sigma(g-h)\Big) + \boldsymbol{\lambda}^{\mathrm{T}} \begin{bmatrix} u \\ 1 \end{bmatrix} . \tag{4.240}$$

Die optimale Lösung $u^*(y)$, die zugehörige Zustandstrajektorie $\mathbf{x}^*(y)$ und die zugehörige Kozustandstrajektorie $\boldsymbol{\lambda}^*(y)$ müssen (4.238c), (4.239), die Bedingung

$$H(\mathbf{x}^*(y), v, \boldsymbol{\lambda}^*(y)) \geq H(\mathbf{x}^*(y), u^*(y), \boldsymbol{\lambda}^*(y)) \quad \forall v \in [-s, s] \tag{4.241}$$

und die Differenzialgleichung

$$\frac{\mathrm{d}\boldsymbol{\lambda}^*}{\mathrm{d}y} = -\left(\frac{\partial}{\partial \mathbf{x}} H\right)^{\mathrm{T}} (\mathbf{x}^*, u^*, \boldsymbol{\lambda}^*) = \begin{bmatrix} -1 \\ \frac{\mathrm{d}g}{\mathrm{d}y} \end{bmatrix} \Big(k^+\sigma(h-g) - k^-\sigma(g-h)\Big) \tag{4.242}$$

erfüllen. Als Randbedingungen treten zusätzlich die Transversalitätsbedingungen

$$\lambda_1^*(y_0) = 0 \ , \qquad \lambda_1^*(y_1) = 0 \tag{4.243}$$

auf, da weder $h(y_0)$ noch $h(y_1)$ beschränkt sind (vgl. (4.179)).

> *Bemerkung* 4.3. Würden $h(y_0) = g(y_0)$ und $h(y_1) = g(y_1)$ als weitere Bedingungen in (4.238) auftreten, so könnten keine Randbedingungen für $\lambda_1^*(y_0)$ und $\lambda_1^*(y_1)$ formuliert werden.

Für die Schaltfunktion ergibt sich gemäß (4.219)

$$\zeta(y) = \left(\frac{\partial}{\partial u} H\right)(\mathbf{x}, u, \boldsymbol{\lambda}) = \lambda_1 \ . \tag{4.244}$$

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 (4.220) liefert für den singulären Pfad

$$\frac{\mathrm{d}\zeta^*(y)}{\mathrm{d}y} = \frac{\mathrm{d}\lambda_1^*}{\mathrm{d}y} = k^-\sigma(g - h^*) - k^+\sigma(h^* - g) = 0 \ , \tag{4.245}$$

d. h. am singulären Pfad muss $h^*(y) = g(y)$ und folglich $u^* = \frac{\mathrm{d}g}{\mathrm{d}y}$ gelten. Die optimale Lösung $u^*(y)$ kann daher in der Form

$$u^* = \begin{cases} -s & \text{falls } \lambda_1^* > 0 \\ s & \text{falls } \lambda_1^* < 0 \\ \frac{\mathrm{d}g}{\mathrm{d}y} & \text{falls } \lambda_1^* = 0 \end{cases} \tag{4.246}$$

zusammengefasst werden. Das optimale Profil $h^*(y)$ der Straße kann schließlich durch einfache Integration von (4.238b) berechnet werden.

Für das im rechten Teil der Abbildung 4.15 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

$$u^*(y) = \begin{cases} \frac{1}{2} & \text{falls } y \in [0, b) \\ 0 & \text{falls } y \in (b, 2] \end{cases} \ , \qquad h^*(y) = \begin{cases} h_0 + \frac{y}{2} & \text{falls } y \in [0, b) \\ 1 & \text{falls } y \in [b, 2] \end{cases} \tag{4.247}$$

mit den noch zu bestimmenden Konstanten h_0 und b (siehe Abbildung 4.16). Integration der ersten Zeile von (4.242) liefert unter Berücksichtigung von (4.243)

$$\lambda_1^*(y) = \begin{cases} -y & \text{falls } y \in [0, a) \\ y - 2a & \text{falls } y \in [a, b) \\ 0 & \text{falls } y \in [b, 2] \end{cases} \tag{4.248}$$

mit der noch zu bestimmenden Konstante a. Damit $\lambda_1^*(y)$ am Punkt $y = b$ stetig ist, muss

$$b - 2a = 0 \tag{4.249a}$$

gelten. Weiters müssen die Bedingungen

$$h^*(a) = h_0 + \frac{a}{2} = g(a) = a \ , \qquad h^*(b) = h_0 + \frac{b}{2} = g(b) = 1 \tag{4.249b}$$

erfüllt sein (siehe Abbildung 4.16). Aus (4.249) folgen unmittelbar

$$h_0 = \frac{1}{3} \ , \qquad a = \frac{2}{3} \ , \qquad b = \frac{4}{3} \tag{4.250}$$

und damit das in Abbildung 4.16 rot dargestellte optimale Höhenprofil $h^*(y)$ der Straße. Dieses Höhenprofil genügt den Optimalitätsbedingungen gemäß Satz 4.11 und der Bereich $[b, y_1]$ stellt einen singulären Pfad dar, da in diesem Bereich (4.220) erfüllt ist. Die Schaltsequenz $\{s, 0\}$ ist also optimal.

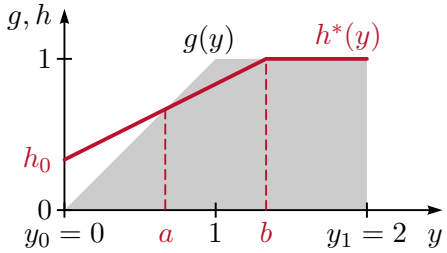


Abbildung 4.16: Optimales Höhenprofil der Straße.

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 30. 09. 2020).

[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* (Texts in Applied Mathematics 12), 3. Aufl. New York, Berlin: Springer, 2002.

[4.5] W. Kemmetmüller, *Skriptum zur VU Modellbildung (SS 2024)*, Institut für Automatisierungs- und Regelungstechnik, TU Wien, 2024. Adresse: https://www.acin.tuwien.ac.atbachelor/modellbildung/.

[4.6] C. Lanczos, *The Variational Principles of Mechanics*, 4. Aufl. New York: Dover, 1970.

[4.7] L. Meirovitch, *Methods of Analytical Dynamics* (Advanced Engineering Series). New York: McGraw-Hill, 1970.

[4.8] H. A. Mang und G. Hofstetter, *Festigkeitslehre*, 4. Aufl. Wien, New York: Springer, 2013.

[4.9] 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.

[4.10] J. Troutman, *Variational Calculus and Optimal Control: Optimization with Elementary Convextity* (Undergraduate Texts in Mathematics), 2. Aufl. New York: Springer, 1996.

[4.11] A. Kugi, *Skriptum zur VO Nichtlineare dynamische Systeme und Regelung (SS 2024)*, Institut für Automatisierungs- und Regelungstechnik, TU Wien, 2024. Adresse: https://www.acin.tuwien.ac.at/master/nichtlineare-dynamische-systeme-und-regelung/.

[4.12] M. Athans und P. L. Falb, *Optimal Control: An Introduction to the Theory and Its Applications*. New York: McGraw-Hill, 1966.

[4.13] J. Macki und A. Strauss, *Introduction to Optimal Control Theory* (Undergraduate Texts in Mathematics). New York: Springer, 1982.

[4.14] E.B. Lee und L. Markus., *Foundations of optimal control theory* (The SIAM Series in Applied Mathematics). New York: John Wiley & Sons, 1967.

[4.15] R. Vinter, *Optimal Control*. Boston: Birkhäuser, 2000.

[4.16] D. Liberzon, *Calculus of Variations and Optimal Control Theory, A Concise Introduction*. Princeton, Oxford: Princeton University Press, 2012.

[4.17] L. Pontryagin, V. Boltyanskii, R. Gamkrelidze und E. Mishchenko, *The Mathematical Theory of Optimal Processes*. Pergamon Press, 1964.

[4.18] H. Schättler und U. Ledzewicz, *Geometric Optimal Control, Theory, Methods and Examples* (Interdisciplinary Applied Mathematics 38). New York: Springer, 2012.

[4.19] M. Ross, *A Primer on Pontryagin's Principle in Optimal Control*, 2. Aufl. San Francisco: Collegiate Publishers, 2015.

[4.20] A. Lewis, *The Maximum Principle of Pontryagin in control and in optimal control*, Department of Methematics und Statistics, Queen's University, Kingston, Canada, 2006. Adresse: https://mast.queensu.ca/~andrew/teaching/pdf/maximum-principle.pdf.

[4.21] A. E. Bryson, Jr. und Y.-C. Ho, *Applied Optimal Control: Optimization, Estimation, and Control*. John Wiley & Sons, 1975.

[4.22] 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.

[4.23] M. Papageorgiou, M. Leibold und M. Buss, *Optimierung: Statische, dynamische, stochastische Verfahren für die Anwendung*, 4. Aufl. Springer, 2015.

[4.24] B. van Brunt, *The Calculus of Variations* (Universitext). Springer, 2004.

[4.25] O. Föllinger, *Optimale Regelung und Steuerung* (Methoden der Regelungs- und Automatisierungstechnik). R. Oldenbourg Verlag, 1994.

[4.26] D. S. Naidu, *Optimal Control Systems* (Electrical Engineering Series). CRC Press, 2003.

[4.27] D. E. Kirk, *Optimal Control Theory: An Introduction*. Dover Publications, 2004.