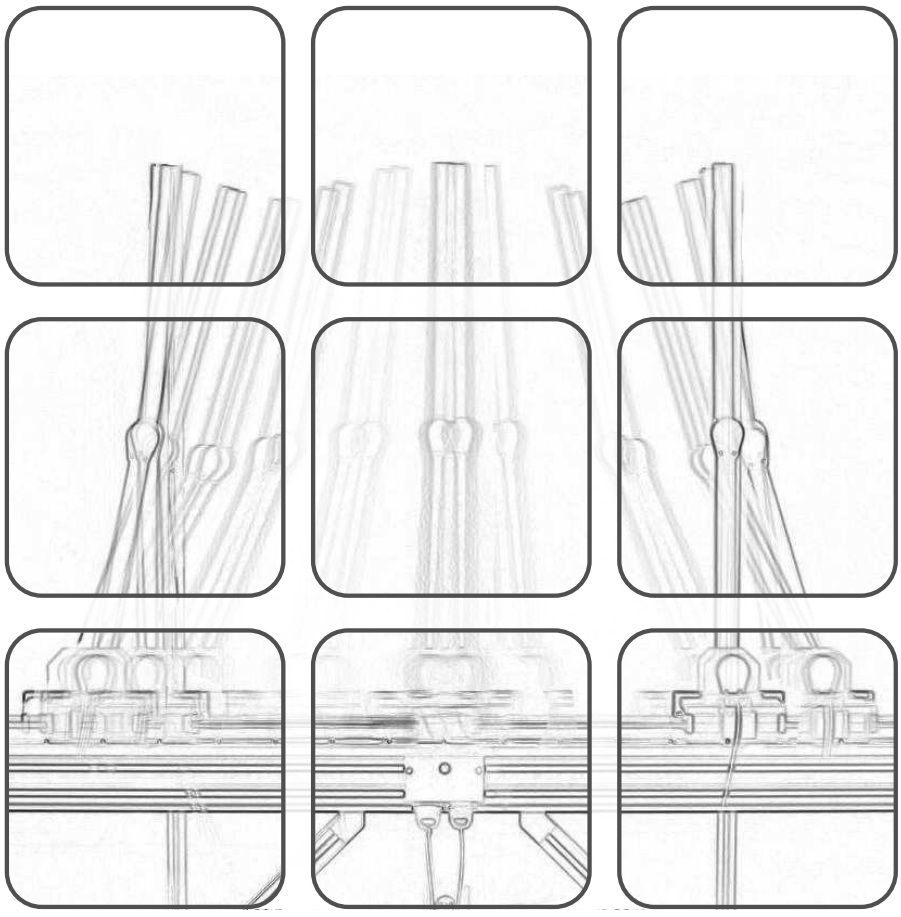
TECHNISCHE UNIVERSITÄT WIEN

AUTOMATION & CONTROL INSTITUTE
INSTITUT FÜR AUTOMATISIERUNGS- & REGELUNGSTECHNIK

Vorlesung und Übung
WS 2016/2017

Andreas STEINBÖCK
Univ.-Prof. Dr. techn. Andreas KUGI

OPTIMIERUNG



Optimierung

Vorlesung und Übung
WS 2016/2017

Andreas STEINBÖCK
Univ.-Prof. Dr. techn. Andreas KUGI

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

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

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)–(1.1c), handelt es sich um ein *beschränktes Optimierungsproblem*.

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

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

wird als *zulässiger Bereich* (Englisch: *admissible or feasible region*) und jedes $\mathbf{x} \in \mathcal{X}_{\text{a}\lceil}$ als *zulässiger Punkt* bezeichnet. Damit lässt sich das statische Optimierungsproblem (1.1) auch in der äquivalenten Form

$$\min_{\mathbf{x} \in \mathcal{X}_{\text{a}\lceil}} \ f(\mathbf{x}) \tag{1.3}$$

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

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

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

$$\max_{\mathbf{x} \in \mathbb{R}^n} \ f(\mathbf{x}) = \min_{\mathbf{x} \in \mathbb{R}^n} \ -f(\mathbf{x})\,.$$

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

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

Unterschieden werden bei statischen Optimierungsproblemen häufig folgende Klassen:

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

1.1.2 Beispiele

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

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

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

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

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

$$x_3 = 1 - x_1 - x_2$$

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

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

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

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

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

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

ausdrücken lässt.

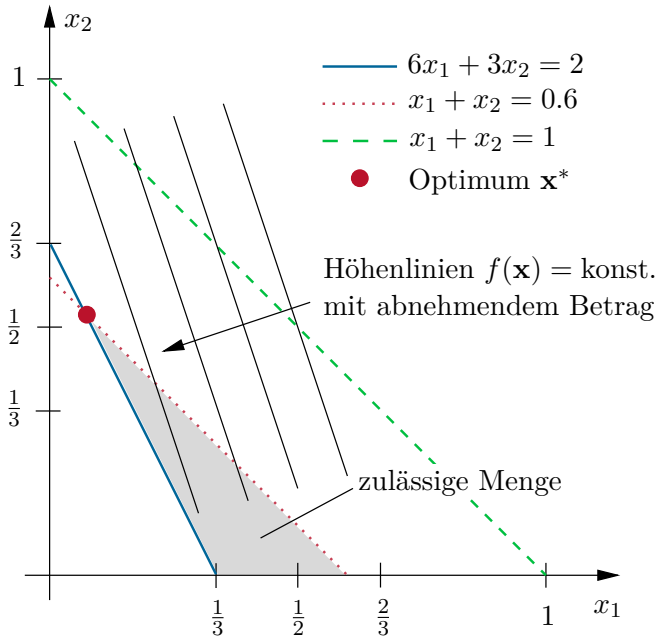


Abbildung 1.1: Veranschaulichung der Portfolio-Optimierung in Beispiel 1.1.

Somit kann das statische Optimierungsproblem in der Form

$$\begin{aligned} \min_{\mathbf{x}\in\mathbb{R}^2} \quad & f(\mathbf{x}) = 1 + 3x_1 + x_2 && \text{(1.7a)} \\ \text{u.B.v.} \quad & 6x_1 + 3x_2 \geq 2 && \text{(1.7b)} \\ & x_1 + x_2 \leq 0.6 && \text{(1.7c)} \\ & x_1 + x_2 \leq 1 && \text{(1.7d)} \\ & x_1, x_2 \geq 0 && \text{(1.7e)} \end{aligned}$$

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

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

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

Beispiel 1.2. Betrachtet wird das (zunächst) unbeschränkte Problem

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

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

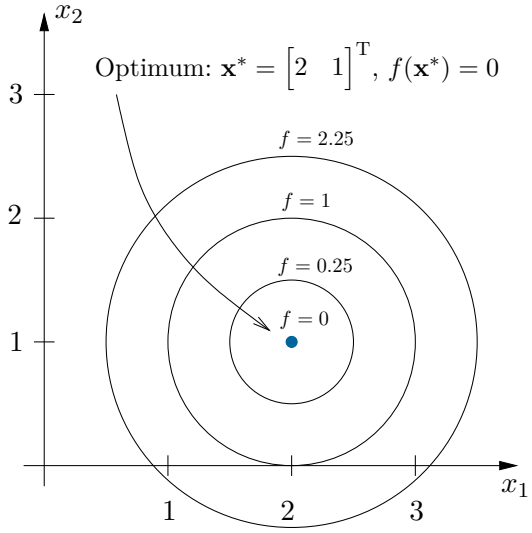


Abb. 1.2: Geometrische Darstellung des unbeschränkten Optimierungsproblems (1.9).

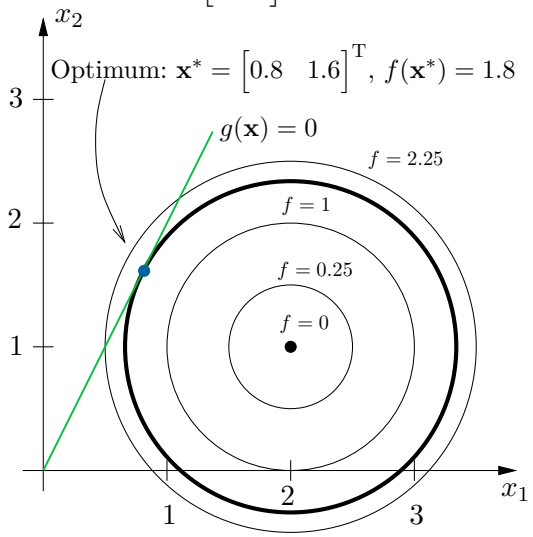


Abb. 1.3: Geometrische Darstellung des beschränkten Optimierungsproblems (1.9), (1.10).

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

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

Die Gleichungsbeschränkung entspricht einer algebraischen Zwangsbedingung, wodurch lediglich noch eine Optimierungsvariable frei wählbar ist. Geometrisch interpretiert bedeutet dies, dass eine mögliche Lösung auf der Gerade liegen muss, die durch (1.10) 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 (1.10) wird nun die Ungleichungsbeschränkung

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

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

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

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

betrachtet werden, durch die sich die Menge der zulässigen Punkte weiter verkleinert (Abbildung 1.5). Der optimale Punkt $\mathbf{x}^* = \begin{bmatrix} 1 & 1 \end{bmatrix}^{\mathrm{T}}$ 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 (1.11) und (1.12) sind aktiv.

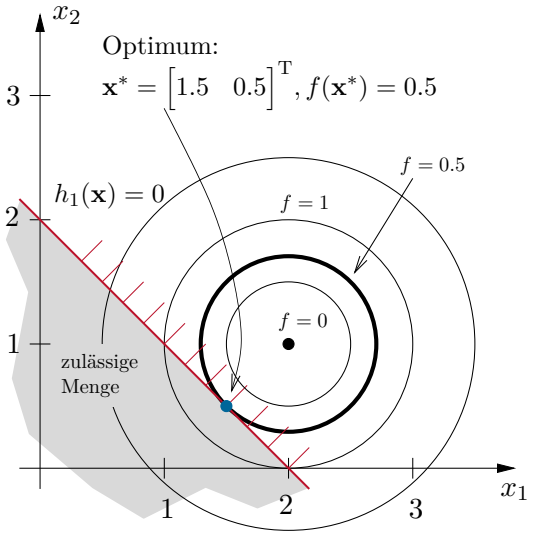


Abb. 1.4: Geometrische Darstellung des beschränkten Optimierungsproblems (1.9), (1.11).

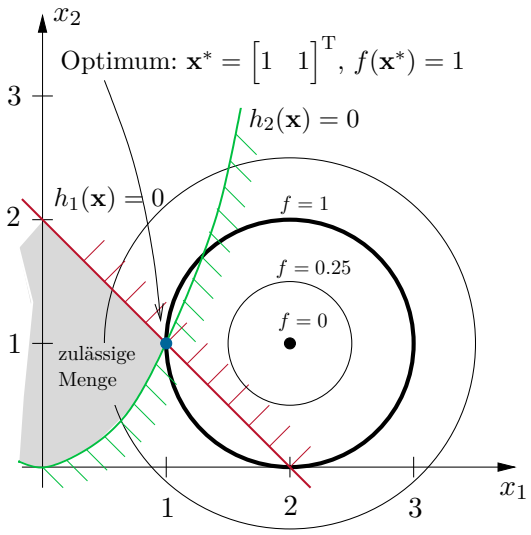


Abb. 1.5: Geometrische Darstellung des beschränkten Optimierungsproblems (1.9), (1.11), (1.12).

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

1.2 Dynamische Optimierungsprobleme

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

1.2.1 Mathematische Formulierung

Die generelle Struktur eines dynamischen Optimierungsproblems lautet

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

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

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

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

Dabei stellt $\mathbf{u} \in \mathbb{R}^m$ die Eingangsgröße des nichtlinearen Systems (1.13b) mit dem Zustand $\mathbf{x} \in \mathbb{R}^n$ dar. Zusätzlich zu den Anfangsbedingungen in (1.13b) sind häufig Endbedingungen der Form (1.13c) gegeben, um z. B. einen gewünschten Zustand $\mathbf{x}_\mathrm{f}$ 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 (1.13d) auf, die z. B. die Begrenzung einer Stellgröße oder Sicherheitsschranken eines Zustandes darstellen können.

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

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

1.2.2 Beispiele

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

$$\min_{u(\cdot)} \quad J(u) = \int_0^{t_1} 1 + c\,u^2 \;\mathrm{d}t\,, \tag{1.14a}$$

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

$$\mathbf{x}(0) = \begin{bmatrix} 0 & 0 & \pi & 0 \end{bmatrix}^\mathrm{T}, \quad \mathbf{x}(t_1) = \begin{bmatrix} 1 & 0 & \pi & 0 \end{bmatrix}^\mathrm{T}, \tag{1.14c}$$

$$-1 \leq u \leq 1\,. \tag{1.14d}$$

Die vereinfachten Bewegungsgleichungen (1.14b) für die Zustände $\mathbf{x} = \begin{bmatrix} x & \dot{x} & \theta & \dot{\theta} \end{bmatrix}^\mathrm{T}$ sind normiert. Der Eingang u stellt die am Wagen angreifende Kraft dar und ist durch (1.14d) beschränkt. Die Masse des Pendels wird mit m, diejenige des Wagens

mit M bezeichnet. Abbildung 1.6 zeigt exemplarisch das seitliche Versetzen des Pendels, um die Bewegung zu verdeutlichen.

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

$$J(u) = \int_0^{t_1} 1 \ \mathrm{d}t = t_1 \,. \tag{1.15}$$

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

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

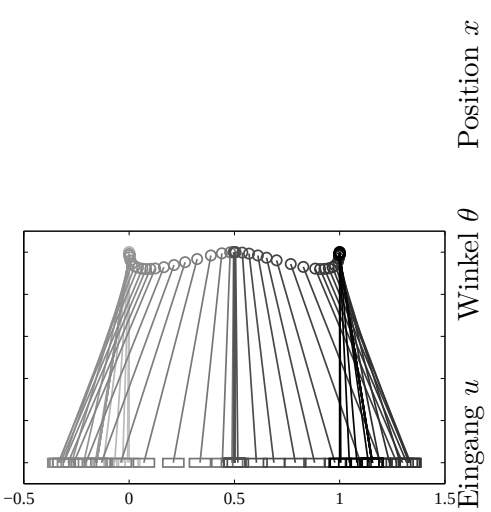


Abb. 1.6: Momentaufnahmen beim Versetzens des inversen Pendels.

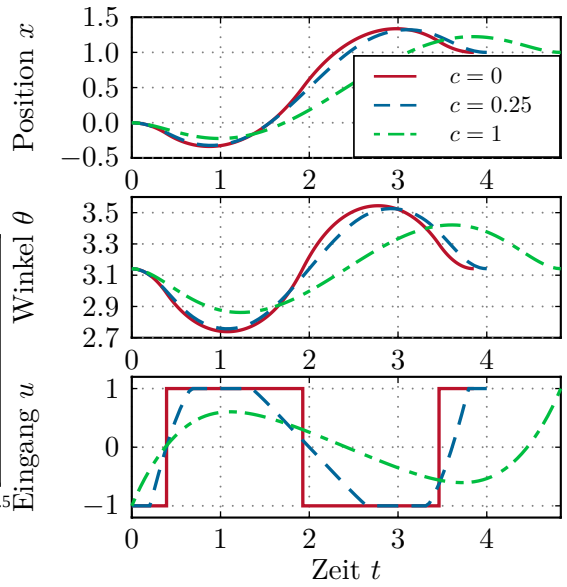


Abb. 1.7: Optimale Trajektorien beim Versetzen des inversen Pendels.

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

$$J(u) = \int_0^{t_1} u^2 \ \mathrm{d}t \,, \tag{1.16}$$

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

Beispiel 1.4 (Goddard-Rakete [1, 2]). Ein klassisches Optimierungsproblem aus der Raumfahrt ist die Maximierung der Flughöhe einer Rakete unter dem Einfluss von Luftreibung und Erdbeschleunigung. Dieses Problem wurde von dem amerikanischen Raketenpionier Robert H. Goddard im Jahr 1919 aufgestellt und kann in der normierten Form

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

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

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

$$0 \leq u \leq 3.5 \tag{1.17d}$$

geschrieben werden.

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

$$D(h,v) = D_0\, v^2 \mathrm{e}^{[\beta(1-h)]} \tag{1.18}$$

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

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

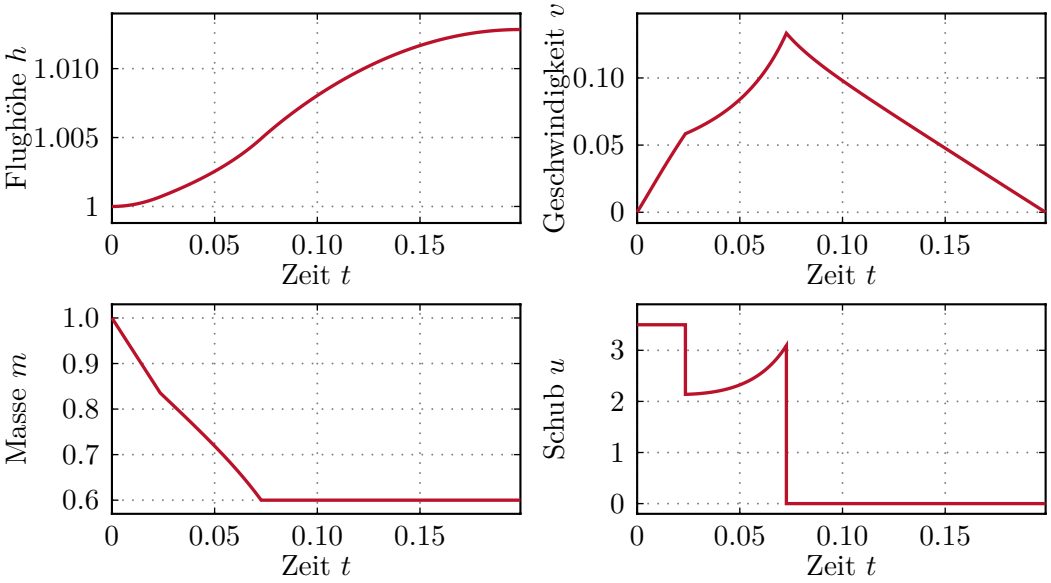


Abbildung 1.8: Trajektorien für die Goddard-Rakete in Beispiel 1.4.

Beispiel 1.5 (Ökonomisches Modell [3, 4]). Ein weiterer Anwendungszweig der dynamischen Optimierung sind wirtschaftliche Prozesse. Das folgende Beispiel beschreibt das Verhalten eines typischen Konsumenten, der Konsum, Freizeit und Bildung über die Lebensdauer maximieren will. Der Bildungsgrad B und das Kapital K eines durchschnittlichen Konsumenten lassen sich durch folgendes Modell beschreiben

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

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

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

$$u_1 > 0, \quad 0 \le u_2 \le 1, \quad 0 \le u_3 < 1. \tag{1.20}$$

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

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

ausgedrückt ist. Der Integralanteil

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

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

Die optimalen Zeitverläufe des Bildungsgrades $B(t)$ und des Kapitals $K(t)$ sind in Abbildung 1.9 dargestellt. Die Funktion $g(u_3)$ in (1.19b) ist durch die Parabel $g(u_3) = 1 - (1 - a)u_3 - au_3^2$ 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.

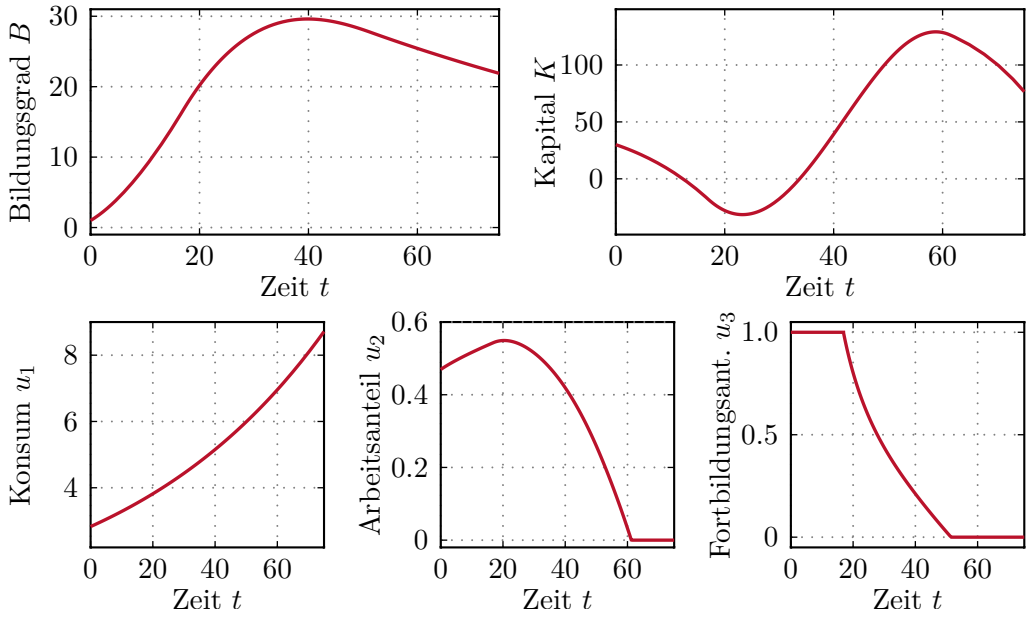


Abbildung 1.9: Optimale Trajektorien für das Konsumentenverhalten in Beispiel 1.5.

1.3 Mathematische Grundlagen

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

1.3.1 Infimum, Supremum, Minimum und Maximum

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

(a) $x \geq \alpha$ für alle $x \in \mathcal{X}$

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

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

(a) $x \leq \alpha$ für alle $x \in \mathcal{X}$

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

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

Für die folgende Definition wird angenommen, dass $\mathcal{X}_{\mathrm{a}\lceil} \subset \mathbb{R}^n$ den zulässigen Bereich des betrachteten Optimierungsproblems gemäß (1.2) bezeichnet.

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

(a) ein *lokales Minimum*, falls ein $\varepsilon > 0$ so existiert, dass gilt $f(\mathbf{x}^*) \leq f(\mathbf{x})$ für alle $\mathbf{x} \in \mathcal{U}_\varepsilon \cap \mathcal{X}_{\mathrm{a}\lceil}$, wobei $\mathcal{U}_\varepsilon$ eine hinreichend kleine ε-Umgebung von $\mathbf{x}^*$ bezeichnet,

(b) ein *striktes lokales Minimum*, falls ein $\varepsilon > 0$ so existiert, dass gilt $f(\mathbf{x}^*) < f(\mathbf{x})$ für alle $\mathbf{x} \in \mathcal{U}_\varepsilon \backslash \{\mathbf{x}^*\} \cap \mathcal{X}_{\mathrm{a}\lceil}$

(c) ein *globales (absolutes) Minimum*, falls $f(\mathbf{x}^*) \leq f(\mathbf{x})$ für alle $\mathbf{x} \in \mathcal{X}_{\mathrm{a}\lceil}$,

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

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

An dieser Stelle sollte nochmals betont werden, dass im Falle eines Minimums (Maximums) der Wert $\min\left\{f(\mathbf{x}) \,|\mathbf{x} \in \mathcal{X}_{\mathrm{a}\lceil}\right\}$ bzw. $\max\left\{f(\mathbf{x}) \,|\mathbf{x} \in \mathcal{X}_{\mathrm{a}\lceil}\right\}$ in $\mathcal{X}_{\mathrm{a}\lceil}$ enthalten sein muss, wohingegen der Wert $\inf\left\{f(\mathbf{x}) \,|\mathbf{x} \in \mathcal{X}_{\mathrm{a}\lceil}\right\}$ bzw. $\sup\left\{f(\mathbf{x}) \,|\mathbf{x} \in \mathcal{X}_{\mathrm{a}\lceil}\right\}$ nicht unbedingt ein zulässiger Punkt ist.

Die Menge aller Minima wird oftmals auch in der Form

$$\mathcal{G} = \arg\min\left\{f(\mathbf{x}) \mid \mathbf{x} \in \mathcal{X}_{\mathrm{a}\lceil}\right\} \tag{1.23}$$

angeschrieben, wobei die Menge sowohl leer sein kann als auch aus endlich oder unendlich vielen Punkten bestehen kann. Im Falle eines strikten globalen Minimums in $\mathcal{X}_{\mathrm{a}\lceil}$ versteht

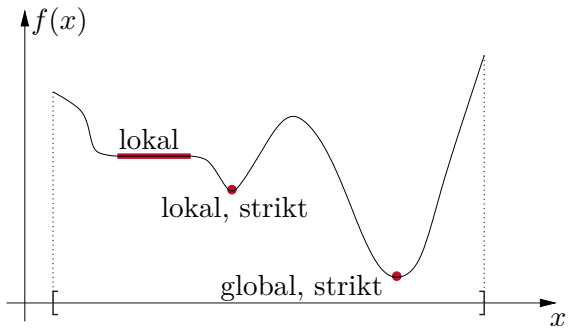


Abbildung 1.10: Verschiedene Minima einer Funktion $f(x)$ mit $x \in \mathbb{R}$.

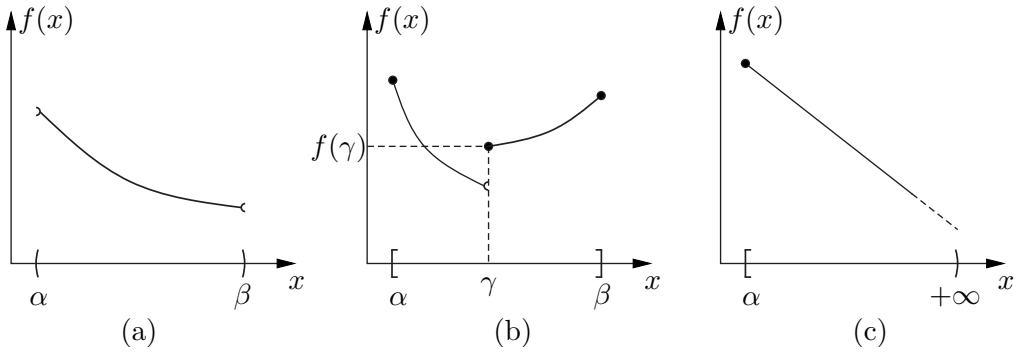


Abbildung 1.11: Nichtexistenz von Minima.

man unter dem Ausdruck $\bar{\mathbf{x}} = \arg\min\left\{ f(\mathbf{x}) \mid \mathbf{x} \in \mathcal{X}_{\mathrm{a}\lceil} \right\}$ meist jene Funktion, die gerade den Punkt $\bar{\mathbf{x}} \in \mathcal{X}_{\mathrm{a}\lceil}$ zurückgibt, der die Funktion $f(\mathbf{x})$ global minimiert.

1.3.2 Existenz von Minima und Maxima

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

Satz 1.1 (Weierstrass). *Es sei $\mathcal{X}$ eine* nichtleere *und* kompakte (abgeschlossene und beschränkte) *Menge und $f : \mathcal{X} \to \mathbb{R}$* 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 [5, 6] zu finden. Es sei an dieser Stelle

jedoch betont, dass Satz 1.1 nur eine hinreichende Bedingung für die Existenz einer optimalen Lösung angibt. Als Beispiel dazu betrachte man die Minimierungsaufgabe $\min_{x\in(-1,1)} x^2$, die zeigt, dass mit $x = 0$ ein Minimum gegeben ist, obwohl die Menge $\mathcal{X} = \{x \in \mathbb{R} | -1 < x < 1\}$ offen und damit nicht kompakt ist.

1.3.3 Gradient und Hessematrix

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

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

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

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

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

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

die Hessematrix (also die 2. partielle Ableitung) von $f(\mathbf{x})$ an der Stelle $\mathbf{x} = \begin{bmatrix} x_1 & \dots & x_n \end{bmatrix}^{\mathrm{T}}$.

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

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

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

Somit ist die Hessematrix symmetrisch ($(\nabla^2 f)(\mathbf{x}) = \left(\nabla^2 f\right)^{\mathrm{T}}(\mathbf{x})$) und hat 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} die Elemente der i-ten Zeile und j-ten Spalte der Matrix $\mathbf{A}$ bezeichnen. Um die Definitheit einer symmetrischen Matrix $\mathbf{A}$ zu bestimmen, muss lediglich eine der drei Bedingungen (a)–(c) ausgewertet werden, da jede für sich notwendig und hinreichend ist. Das Kriterium (c) wird auch *Sylvester-Kriterium* genannt und kann nicht für semi-definite Matrizen verwendet werden.

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

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

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

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*

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

gilt.

1.3.4 Konvexität

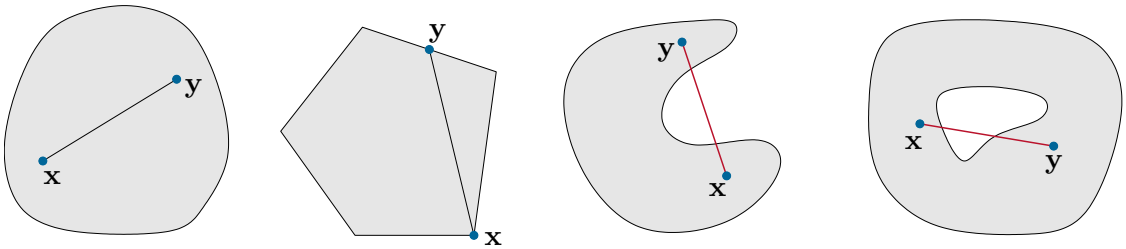
Die Eigenschaft der Konvexität ist von großer Bedeutung in der Optimierung und erlaubt häufig eine einfache (numerische) Lösung des Optimierungsproblems. Der Begriff *konvex* kann sowohl auf Mengen als auch auf Funktionen angewandt werden.

1.3.4.1 Konvexe Mengen

Definition 1.5 (Konvexe Menge)**.** Eine Menge $\mathcal{X} \subset \mathbb{R}^n$ nennt man *konvex*, falls für alle $\mathbf{x}, \mathbf{y} \in \mathcal{X}$ und alle reellen Zahlen α mit $0 < \alpha < 1$ gilt

$$(\alpha\,\mathbf{x} + (1-\alpha)\,\mathbf{y}) \in \mathcal{X}\,. \tag{1.28}$$

Eine geometrische Interpretation dieser Definition ist, dass eine Menge $\mathcal{X} \subset \mathbb{R}^n$ 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 1.12 zeigt einige Beispiele konvexer und nicht-konvexer Mengen.



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

Abbildung 1.12: Beispiele von konvexen und nicht-konvexen Mengen.

Konvexe Mengen besitzen folgende Eigenschaften:

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

(b) Wenn $\mathcal{X}$ eine konvexe Menge ist und α eine feste reelle Zahl, dann ist die Menge

$$\{\alpha\mathbf{x} \,|\, \mathbf{x} \in \mathcal{X}\}$$

ebenfalls konvex.

(c) Das Bild einer konvexen Menge unter einer linearen Transformation ist konvex.

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

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

ebenfalls konvex.

Diese Eigenschaften sind u. a. bei der Charakterisierung der Konvexität der zulässigen Menge $\mathcal{X}_{\mathrm{a}\lceil}$ von Optimierungsproblemen von Bedeutung.

1.3.4.2 Konvexe Funktionen

Definition 1.6 (Konvexe und konkave Funktionen)**.** Es sei $\mathcal{X} \subset \mathbb{R}^n$ 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

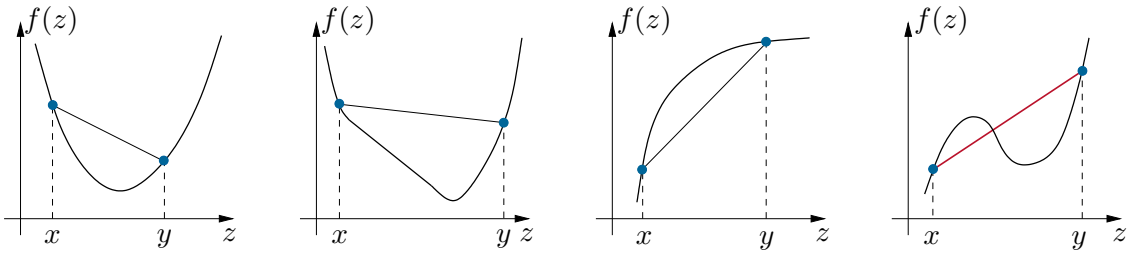
$$f(\alpha\,\mathbf{x} + (1-\alpha)\,\mathbf{y}) \leq \alpha\, f(\mathbf{x}) + (1-\alpha)\, f(\mathbf{y}). \tag{1.29}$$

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

$$f(\alpha\,\mathbf{x} + (1-\alpha)\,\mathbf{y}) < \alpha\, f(\mathbf{x}) + (1-\alpha)\, f(\mathbf{y}). \tag{1.30}$$

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

Die Definition 1.6 kann wie folgt geometrisch interpretiert werden: Eine Funktion f ist konvex (konkav), falls für alle $\mathbf{x} \in \mathcal{X}$, $\mathbf{y} \in \mathcal{X}$ und $0 < \alpha < 1$ alle Funktionswerte $f(\alpha\,\mathbf{x} + (1-\alpha)\,\mathbf{y})$ unterhalb (oberhalb) oder auf der Verbindungslinie zwischen $f(\mathbf{x})$ und $f(\mathbf{y})$ liegen. Abbildung 1.13 zeigt einige Beispiele konvexer und konkaver Funktionen. Es ist direkt ersichtlich, dass affine Funktionen sowohl konkav als auch konvex sind.



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

Abbildung 1.13: Beispiele von konvexen und konkaven Funktionen.

Konvexe Funktionen besitzen folgende Eigenschaften:

(a) Die Summenfunktion

$$f(\mathbf{x}) = \sum_{i=1}^{k} a_i\, f_i(\mathbf{x}) \tag{1.31}$$

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.

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

$$\mathcal{S} = \{\mathbf{x} \in \mathcal{X} \mid f(\mathbf{x}) \leq c\} \tag{1.32}$$

für alle reellen Zahlen $c \in \mathbb{R}$ ebenfalls konvex, siehe Abbildung 1.14.

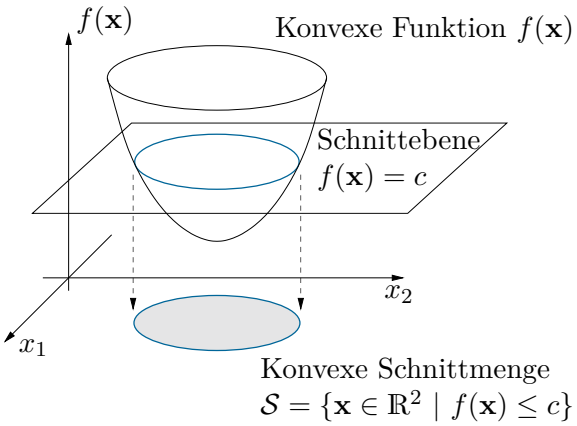


Abb. 1.14: Konvexe Menge $\mathcal{S}$, die durch den Schnitt einer konvexen Funktion $f(\mathbf{x})$ mit der Ebene $f(\mathbf{x}) = \text{konst.}$ entsteht.

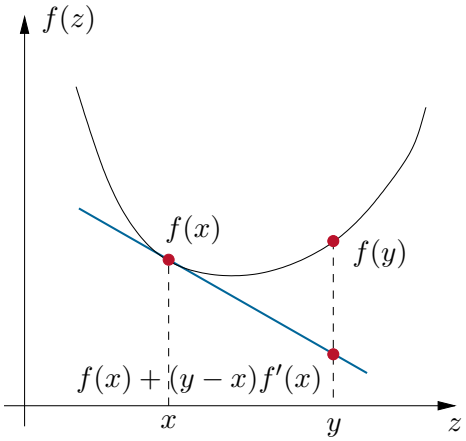


Abb. 1.15: Stützende Tangente einer konvexen Funktion $f(z)$.

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

$$f(\mathbf{y}) \geq f(\mathbf{x}) + (\mathbf{y} - \mathbf{x})^{\mathrm{T}}(\nabla f)(\mathbf{x}) \tag{1.33}$$

erfüllt ist. Die geometrische Interpretation der Ungleichung (1.33) 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 1.15 veranschaulicht.

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

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

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

1.4 Literatur

[1] A. E. Bryson, Jr., *Dynamic Optimization*. Addison-Wesley, 1999.

[2] R. H. Goddard, „A method of reaching extreme altitudes“, *Smithsonian Miscellaneous Collections*, Jg. 71, Nr. 2, 1919.

[3] K. Pohmer, *Mikroökonomische Theorie der personellen Einkommens- und Vermögensverteilung*, Ser. Studies in Contemporary Economics. Springer, 1985, Bd. 16.

[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.

[5] M. Bazaraa, H. Sherali und C. Shetty, *Nonlinear Programming: Theory and Algorithms*, 3. Aufl. John Wiley & Sons, 2006.

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

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

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

[9] M. Papageorgiou, M. Leibold und M. Buss, *Optimierung: Statische, dynamische, stochastische Verfahren für die Anwendung*, 3. Aufl. Springer, 2012.

[10] D. P. Bertsekas, *Nonlinear Programming*, 2. Aufl. Athena Scientific, 1999.

[11] D. G. Luenberger und Y. Ye, *Linear and Nonlinear Programming*, 3. Aufl., Ser. International Series in Operations Research & Management Science. Springer, 2008, Bd. 116.

[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.

2 Statische Optimierung: Unbeschränkter Fall

2.1 Optimalitätsbedingungen

Bevor in den Abschnitten 2.2–2.6 numerische Verfahren zur Lösung unbeschränkter statischer Optimierungsprobleme der Art

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

behandelt werden, sollen in diesem Abschnitt die Optimalitätsbedingungen für ein allgemeines beschränktes Optimierungsproblem der Form (1.3)

$$\min_{\mathbf{x}\in\mathcal{X}_{\mathrm{a}\lceil}} \; f(\mathbf{x}) \tag{2.2}$$

mit dem zulässigen Bereich $\mathcal{X}_{\mathrm{a}\lceil}$ diskutiert werden. Zur Definition der Begriffe lokaler und globaler Minima sei auf Abschnitt 1.3.1 und im Speziellen auf Definition 1.2 verwiesen.

Um die *notwendigen* Bedingungen für ein lokales Minimum $\mathbf{x}^*$ der Optimierungsaufgabe (2.2) zu formulieren, führt man den Begriff einer *zulässigen Richtung* ein. Für $\mathbf{x} \in \mathcal{X}_{\mathrm{a}\lceil}$ ist der Vektor $\mathbf{d}$ eine zulässige Richtung am Punkt $\mathbf{x}$, wenn ein $\bar{\alpha} > 0$ so existiert, dass $\mathbf{x} + \alpha\mathbf{d} \in \mathcal{X}_{\mathrm{a}\lceil}$ für alle α, $0 \leq \alpha \leq \bar{\alpha}$.

Satz 2.1 (Notwendige Optimalitätsbedingungen erster Ordnung)**.** *Es sei $\mathcal{X}_{\mathrm{a}\lceil} \subset \mathbb{R}^n$ die zulässige Menge des Optimierungsproblems (2.2) und $f \in C^1$ eine Funktion definiert auf $\mathcal{X}_{\mathrm{a}\lceil}$. Wenn $\mathbf{x}^*$ ein lokales Minimum von f auf $\mathcal{X}_{\mathrm{a}\lceil}$ ist, dann gilt für jede zulässige Richtung $\mathbf{d}$ am Punkt $\mathbf{x}^*$ die Ungleichungsbedingung*

$$\mathbf{d}^{\mathrm{T}}(\nabla f)(\mathbf{x}^*) \geq 0. \tag{2.3}$$

Gilt darüberhinaus, dass $\mathbf{x}^*$ im Inneren von $\mathcal{X}_{\mathrm{a}\lceil}$ liegt, dann folgt die Bedingung

$$(\nabla f)(\mathbf{x}^*) = \mathbf{0}. \tag{2.4}$$

Beweis. Da $\mathbf{d}$ eine zulässige Richtung am Punkt $\mathbf{x}^*$ ist, gilt für jedes α, $0 \leq \alpha \leq \bar{\alpha}$, dass der Punkt $\mathbf{x}(\alpha) = \mathbf{x}^* + \alpha\mathbf{d} \in \mathcal{X}_{\mathrm{a}\lceil}$. Nun definiert man für $0 \leq \alpha \leq \bar{\alpha}$ die Funktion $g(\alpha) = f(\mathbf{x}(\alpha))$, die am Punkt $\alpha = 0$ ein lokales Minimum besitzt. Entwickelt man $g(\alpha)$ um den Punkt $\alpha = 0$ in eine Taylorreihe und bricht diese nach dem linearen Glied ab, erhält man

$$g(\alpha) = g(0) + g'(0)\alpha + o(\alpha) \;, \tag{2.5}$$

wobei $o(\alpha)$ den Restterm bezeichnet, der schneller nach Null abklingt als α. Wäre

nun $g'(0) < 0$, dann würde für ein hinreichend kleines $\alpha > 0$ gelten $g(\alpha) - g(0) < 0$, was ein Widerspruch zur Annahme ist, 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 $\mathcal{X}_{\mathrm{a}\lceil}$ liegt, dann ist *jede Richtung* am Punkt $\mathbf{x}^*$ zulässig, d. h. $\mathbf{d}^{\mathrm{T}}(\nabla f)(\mathbf{x}^*) \geq 0$ für alle $\mathbf{d} \in \mathbb{R}^n$. Dies kann aber nur für alle $\mathbf{d}$ erfüllt sein, wenn $(\nabla f)(\mathbf{x}^*) = \mathbf{0}$ ist. □

Beispiel 2.1. Man betrachte das Optimierungsproblem

$$\min_{\mathbf{x}\in\mathbb{R}^2} \; f(x_1, x_2) = x_1^2 - x_1 x_2 + x_2^2 - 3x_2 \; . \tag{2.6}$$

Berechnet man nun die notwendige Optimalitätsbedingung erster Ordnung gemäß (2.4)

$$2x_1 - x_2 = 0 \tag{2.7a}$$
$$-x_1 + 2x_2 = 3 \; , \tag{2.7b}$$

dann erkennt man, dass $\mathbf{x}^* = \begin{bmatrix} 1 & 2 \end{bmatrix}^{\mathrm{T}}$ eine eindeutige Lösung von (2.7) ist, welche in diesem Fall sogar das globale Minimum beschreibt.

Beispiel 2.2. In einem weiteren Beispiel betrachte man die Optimierungsaufgabe

$$\min_{\mathbf{x}\in\mathcal{X}_{\mathrm{a}\lceil}} \; f(x_1, x_2) = x_1^2 - x_1 + x_2 + x_1 x_2 \tag{2.8}$$

mit der zulässigen Menge

$$\mathcal{X}_{\mathrm{a}\lceil} = \left\{ \mathbf{x} \in \mathbb{R}^2 \mid x_1 \geq 0, \; x_2 \geq 0 \right\} \; . \tag{2.9}$$

Das Problem hat an der Stelle $\mathbf{x}^* = \begin{bmatrix} \frac{1}{2} & 0 \end{bmatrix}^{\mathrm{T}}$ ein globales Minimum. Wertet man den Gradienten an der Stelle $\mathbf{x}^*$ aus, so erhält man

$$\frac{\partial}{\partial x_1} f(\mathbf{x}^*) = \quad 2x_1^* - 1 + x_2^* = 0 \tag{2.10a}$$
$$\frac{\partial}{\partial x_2} f(\mathbf{x}^*) = \qquad 1 + x_1^* = \frac{3}{2} \; . \tag{2.10b}$$

Wie man erkennt, verschwindet in diesem Fall der Gradient an der Stelle $\mathbf{x}^*$ nicht, aber die notwendige Bedingung (2.3) ist für alle zulässigen Richtungen $\mathbf{d}$ erfüllt, da die zweite Komponente von $\mathbf{d}$ wegen der Definition von $\mathcal{X}_{\mathrm{a}\lceil}$ gemäß (2.9) größer gleich Null sein muss.

Die notwendige Optimalitätsbedingung erster Ordnung für einen inneren Punkt (2.4) gemäß Satz 2.1 gibt lediglich an, dass es sich bei diesem Punkt um einen *Extremalpunkt* (auch als *stationären Punkt* bezeichnet) handelt, die Bedingung wird aber von einem

Minimum, Maximum oder Sattelpunkt gleichermaßen erfüllt, siehe Abbildung 2.1.

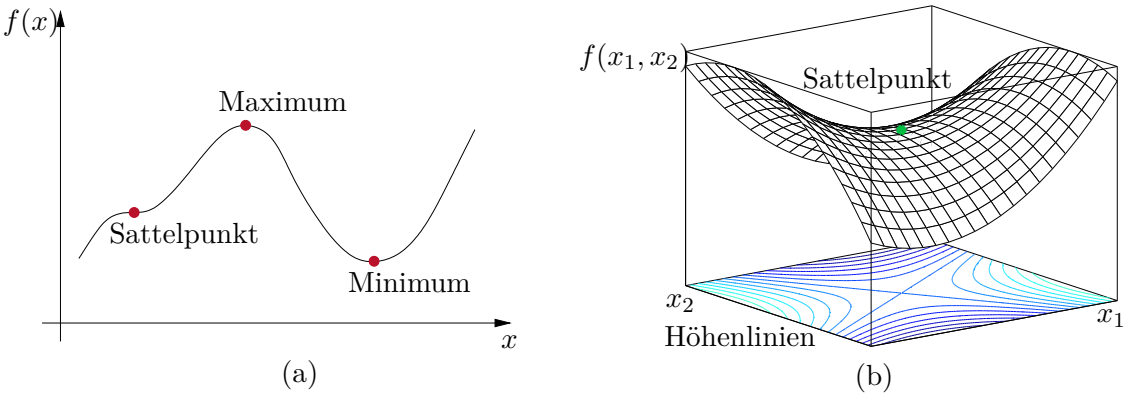


Abbildung 2.1: Beispiele von stationären Punkten im ein- und zweidimensionalen Fall.

Man kann nun Satz 2.1 weiter präzisieren, indem man bei der Taylorreihenentwicklung (2.5) Terme höherer Ordnung in α hinzunimmt.

Satz 2.2 (Notwendige Optimalitätsbedingungen zweiter Ordnung)**.** *Es sei $\mathcal{X}_{\mathrm{a}\lceil} \subset \mathbb{R}^n$ die zulässige Menge des Optimierungsproblems (2.2) und $f \in C^2$ eine Funktion definiert auf $\mathcal{X}_{\mathrm{a}\lceil}$. Wenn $\mathbf{x}^*$ ein lokales Minimum von f auf $\mathcal{X}_{\mathrm{a}\lceil}$ ist, dann gelten für jede zulässige Richtung $\mathbf{d}$ am Punkt $\mathbf{x}^*$ die Bedingungen*

$$(a) \quad \mathbf{d}^{\mathrm{T}}(\nabla f)(\mathbf{x}^*) \geq 0 \tag{2.11a}$$

$$(b) \quad \text{wenn } \mathbf{d}^{\mathrm{T}}(\nabla f)(\mathbf{x}^*) = 0, \text{ dann } \mathbf{d}^{\mathrm{T}}\left(\nabla^2 f\right)(\mathbf{x}^*)\mathbf{d} \geq 0 \ . \tag{2.11b}$$

Gilt darüberhinaus, dass $\mathbf{x}^*$ im Inneren von $\mathcal{X}_{\mathrm{a}\lceil}$ liegt, dann folgen die Bedingungen

$$(a) \quad (\nabla f)(\mathbf{x}^*) = \mathbf{0} \tag{2.12a}$$

$$(b) \quad \text{für alle } \mathbf{d} \text{ gilt } \mathbf{d}^{\mathrm{T}}\left(\nabla^2 f\right)(\mathbf{x}^*)\mathbf{d} \geq 0 \ . \tag{2.12b}$$

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

Die Bedingung (2.12b) entspricht der Forderung, dass die Hessematrix $(\nabla^2 f)(\mathbf{x})$ an der Stelle $\mathbf{x} = \mathbf{x}^*$ positiv semi-definit ist.

Aufgabe 2.2. Betrachten Sie die Optimierungsaufgabe

$$\min_{\mathbf{x} \in \mathcal{X}_{\mathrm{a}\lceil}} \ f(x_1, x_2) = x_1^3 - x_1^2 x_2 + 2x_2^2 \tag{2.13}$$

mit der zulässigen Menge

$$\mathcal{X}_{\mathrm{a}\lceil} = \left\{ \mathbf{x} \in \mathbb{R}^2 \mid x_1 \geq 0, \ x_2 \geq 0 \right\} \ . \tag{2.14}$$

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

Die Optimalitätsbedingungen von Satz 2.2 sind lediglich notwendig, wie man sich einfach anhand der Funktion $f(x) = x^3$ überzeugen kann. Die 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, sondern ein Sattelpunkt (siehe Abbildung 2.1).

Für Punkte im Inneren von $\mathcal{X}_{\mathrm{a}\lceil}$ lassen sich ferner hinreichende Optimalitätsbedingungen angeben.

Satz 2.3 (Hinreichende Optimalitätsbedingungen zweiter Ordnung)**.** *Es sei $\mathcal{X}_{\mathrm{a}\lceil} \subset \mathbb{R}^n$ die zulässige Menge des Optimierungsproblems (2.2) und $f \in C^2$ eine Funktion*

definiert auf $\mathcal{X}_{\mathrm{a}\lceil}$. *Wenn* $\mathbf{x}^*$ *ein* innerer Punkt *von* $\mathcal{X}_{\mathrm{a}\lceil}$ *ist und folgende Bedingungen*

$$(a) \quad (\nabla f)(\mathbf{x}^*) = \mathbf{0} \tag{2.15a}$$

$$(b) \quad \left(\nabla^2 f\right)(\mathbf{x}^*) > 0 \quad \textit{(positiv definite Hessematrix am Punkt } \mathbf{x}^*\textit{)} \tag{2.15b}$$

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

Aufgabe 2.3. Beweisen Sie Satz 2.3.

Beispiel 2.3. Für das Optimierungsproblem

$$\min_{\mathbf{x}\in\mathbb{R}^2} f(\mathbf{x}) = \; x_1^2 + a x_2^2 - x_1 x_2 \tag{2.16}$$

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

$$(\nabla f)(\mathbf{x}) = \begin{bmatrix} 2x_1 - x_2 \\ 2ax_2 - x_1 \end{bmatrix}, \quad \left(\nabla^2 f\right)(\mathbf{x}) = \begin{bmatrix} 2 & -1 \\ -1 & 2a \end{bmatrix}. \tag{2.17}$$

Aus $(\nabla f)(\mathbf{x}^*) = \mathbf{0}$ folgt $\mathbf{x}^* = \begin{bmatrix} 0 & 0 \end{bmatrix}^{\mathrm{T}}$ 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

$$D_1 = 2, \quad D_2 = 4a - 1\,. \tag{2.18}$$

Somit ist $(\nabla^2 f)(\mathbf{x}^*)$ positiv definit für $a > \frac{1}{4}$ und $\mathbf{x}^* = \begin{bmatrix} 0 & 0 \end{bmatrix}^{\mathrm{T}}$ 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 $\mathbf{x}^* = \begin{bmatrix} 0 & 0 \end{bmatrix}^{\mathrm{T}}$ ein *Sattelpunkt*, wie er in Abbildung 2.1(b) für $a = -1$ dargestellt ist.

Wenn die Funktion $f(\mathbf{x})$ (strikt) konvex ist, dann lassen sich stärkere Aussagen im Vergleich zu den bisherigen Sätzen treffen. Der Grund dafür liegt darin, dass aus der Konvexität von $f(\mathbf{x})$ unmittelbar die positive Semi-Definitheit der Hessematrix von $f(\mathbf{x})$ folgt.

Satz 2.4 (Minimierung konvexer Funktionen — Menge der Minima)**.** *Es sei* $f(\mathbf{x})$ *eine konvexe Funktion auf der konvexen Menge* $\mathcal{X}_{\mathrm{a}\lceil}$. *Dann ist die Menge aller Minima* $\mathcal{G} = \arg\min\left\{ f(\mathbf{x}) \mid \mathbf{x} \in \mathcal{X}_{\mathrm{a}\lceil} \right\}$ *ebenfalls konvex und* jedes lokale Minimum *von* $f(\mathbf{x})$ *ist ein* globales Minimum.

Beweis. Angenommen c_0 beschreibt das Minimum von f. Dann ist die Menge $\mathcal{G} = \{\mathbf{x} \mid \mathbf{x} \in \mathcal{X}_{\mathrm{a}\lceil},\ f(\mathbf{x}) \leq c_0\}$ gemäß (1.32) ebenfalls konvex, womit der erste Teil des

Satzes gezeigt ist.

Im Weiteren nehme man an, dass $\mathbf{x}^* \in \mathcal{X}_{\mathrm{a}\lceil}$ ein lokales Minimum von f ist und ein weiterer Punkt $\mathbf{y} \in \mathcal{X}_{\mathrm{a}\lceil}$ so existiert, dass gilt $f(\mathbf{y}) < f(\mathbf{x}^*)$. Auf Grund der Konvexität von f folgt nach Definition 1.6, im Speziellen (1.29), für alle α mit $0 < \alpha < 1$ die Ungleichung

$$f(\alpha\,\mathbf{y} + (1-\alpha)\,\mathbf{x}^*) \leq \alpha\,f(\mathbf{y}) + (1-\alpha)\,f(\mathbf{x}^*) < f(\mathbf{x}^*)\,. \tag{2.19}$$

Da α hinreichend klein sein kann, folgt aber aus (2.19), dass ein weiterer Punkt $\mathbf{z} = \alpha\,\mathbf{y} + (1-\alpha)\,\mathbf{x}^*$ in einer hinreichend kleinen Umgebung von $\mathbf{x}^*$ existiert, der die Funktion f noch kleiner macht, was gemäß Definition 1.2 ein Widerspruch dazu ist, dass $\mathbf{x}^*$ ein lokales Minimum von f ist. □

Der nächste Satz zeigt, dass für eine stetig differenzierbare und konvexe Funktion f die notwendigen Optimalitätsbedingungen erster Ordnung notwendig und hinreichend für die Existenz eines globalen Minimums sind.

Satz 2.5 (Minimierung konvexer Funktionen — globales Minimum)**.** *Es sei $f \in C^1$ eine konvexe Funktion auf der konvexen Menge $\mathcal{X}_{\mathrm{a}\lceil}$. Existiert ein Punkt $\mathbf{x}^* \in \mathcal{X}_{\mathrm{a}\lceil}$ so, dass für alle $\mathbf{y} \in \mathcal{X}_{\mathrm{a}\lceil}$ gilt*

$$(\mathbf{y} - \mathbf{x}^*)^{\mathrm{T}}(\nabla f)(\mathbf{x}^*) \geq 0\ , \tag{2.20}$$

dann ist $\mathbf{x}^*$ ein globales Minimum *von f auf $\mathcal{X}_{\mathrm{a}\lceil}$. Gilt darüberhinaus, dass $\mathbf{x}^*$ im Inneren von $\mathcal{X}_{\mathrm{a}\lceil}$ liegt, dann kann die Ungleichung (2.20) durch die Bedingung $(\nabla f)(\mathbf{x}^*) = \mathbf{0}$ ersetzt werden.*

Beweis. Da $\mathbf{d} = \mathbf{y} - \mathbf{x}^*$ eine zulässige Richtung am Punkt $\mathbf{x}^*$ ist, entspricht (2.20) der notwendigen Optimalitätsbedingung erster Ordnung (2.3) von Satz 2.1. Auf Grund der Konvexität von f folgt nach (1.33) die Ungleichung

$$f(\mathbf{y}) \geq f(\mathbf{x}^*) + (\mathbf{y} - \mathbf{x}^*)^{\mathrm{T}}(\nabla f)(\mathbf{x}^*) \geq f(\mathbf{x}^*) \tag{2.21}$$

für alle $\mathbf{y} \in \mathcal{X}_{\mathrm{a}\lceil}$, womit der Satz bewiesen ist. □

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 (2.15a) (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.

$$f(\mathbf{x}_{k+1}) < f(\mathbf{x}_k), \quad k = 0, 1, 2, \ldots, \tag{2.22}$$

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

$$\lim_{k\to\infty} \mathbf{x}_k = \mathbf{x}^*. \tag{2.23}$$

In der englischsprachigen Literatur werden solche Algorithmen auch als *iterative descent algorithms* bezeichnet. Neben der anhand von (2.23) 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 Euklidische Abstand

$$e(\mathbf{x}) = ||\mathbf{x} - \mathbf{x}^*|| \geq 0 \tag{2.24a}$$

oder die Kostendifferenz

$$e(\mathbf{x}) = f(\mathbf{x}) - f(\mathbf{x}^*) \geq 0 \tag{2.24b}$$

verwendet werden [1]. Das Konvergenzverhalten eines Algorithmus kann nun anhand der zu $\{\mathbf{x}_k\}$ gehörenden Folge $\{e_k\}$ mit $e_k = e(\mathbf{x}_k)$ analysiert werden. Zunächst sollen dazu die Begriffe *Konvergenzordnung* und *Konvergenzrate* einer Folge von Skalaren definiert werden.

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

$$0 \leq \lim_{k\to\infty} \frac{|e_{k+1}|}{|e_k|^p} = \mu < \infty. \tag{2.25}$$

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 2.4. 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

$$\lim_{k\to\infty} \frac{a^{k+1}}{a^{kp}} = \lim_{k\to\infty} a^{1+k(1-p)} < \infty \tag{2.26}$$

erfüllt ist. Mit $p = 1$ folgt dann

$$\lim_{k\to\infty} \frac{a^{k+1}}{a^{k}} = a = \mu \tag{2.27}$$

für die Konvergenzrate μ.

Aufgabe 2.4. Zeigen Sie, dass die Folge $\{a^{2^k}\}$ mit $0 < a < 1$ mit der Konvergenzordnung 2 und der Konvergenzrate 1 gegen 0 konvergiert.

Beispiel 2.5. Die Folge $\{\frac{1}{k^k}\}$ hat eine lineare Konvergenzordnung, da nur für $p \leq 1$ die Bedingung

$$\lim_{k\to\infty} \frac{k^{kp}}{(k+1)^{k+1}} < \infty \tag{2.28}$$

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

$$\lim_{k\to\infty} \frac{k^{k}}{(k+1)^{k+1}} = \lim_{k\to\infty} \frac{1}{k+1}\left(\frac{k}{k+1}\right)^{k} = \mu = 0. \tag{2.29}$$

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]), dass die Konvergenzordnung eines Optimierungsalgorithmus von der Wahl der Fehlerfunktion $e(\mathbf{x})$ weitgehend unabhängig ist. Dies gilt nicht für die Konvergenzrate.

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

2.3 Liniensuchverfahren

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

$$\mathbf{x}_{k+1} = \mathbf{x}_k + \alpha_k \mathbf{s}_k \tag{2.30}$$

mit einer geeigneten *Schrittweite* $\alpha_k > 0$, die Abstiegsbedingung (2.22), d. h.

$$f(\mathbf{x}_{k+1}) = f(\mathbf{x}_k + \alpha_k \mathbf{s}_k) < f(\mathbf{x}_k) \tag{2.31}$$

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.

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

$$\min_{\alpha_k>0} g(\alpha_k) = f(\mathbf{x}_k + \alpha_k \mathbf{s}_k) \tag{2.32}$$

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.

2.3.1 Wahl der Schrittweite

2.3.1.1 Intervallschachtelungsverfahren („Goldener Schnitt“)

Das *Intervallschachtelungsverfahren* generiert für das skalare Optimierungsproblem (2.32) eine konvergierende Folge von Intervallschachtelungen, um das Minimum von $g(\alpha_k)$ einzugrenzen.

Zunächst muss ein Intervall $[l_0, r_0]$ gefunden werden, in dem die Funktion $g(\alpha_k)$ ein Minimum aufweist, siehe Abbildung 2.3. Dies kann z. B. dadurch erreicht werden, dass mit einem hinreichend kleinen l_0 gestartet und r_0 (ausgehend von l_0) sukzessive vergrößert wird, bis der Funktionswert $g(r_0)$ anfängt zuzunehmen. Für das Folgende wird vorausgesetzt, dass die Funktion $g(\alpha_k)$ stetig und *unimodal* im Intervall $[l_0, r_0]$ ist, d. h. die Funktion $g(\alpha_k)$ hat ein eindeutiges lokales Minimum im offenen Intervall (l_0, r_0).

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

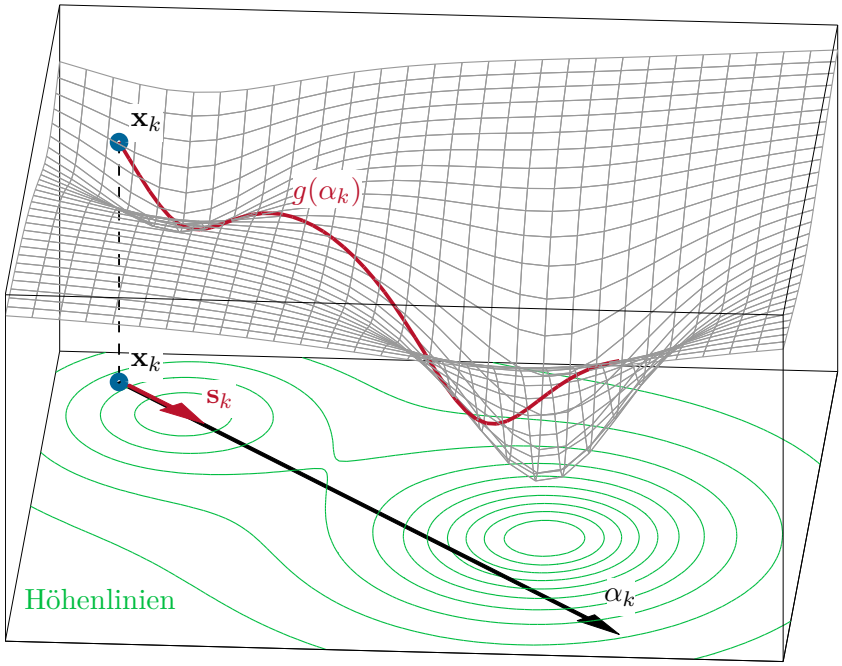


Abbildung 2.2: Veranschaulichung der Wahl der Schrittweite gemäß (2.32).

$$\hat{l}_j = l_j + (1-a)(r_j - l_j) = a l_j + (1-a) r_j \tag{2.33a}$$
$$\hat{r}_j = l_j + a(r_j - l_j) \qquad = (1-a) l_j + a r_j \tag{2.33b}$$

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

Satz 2.6 (Zur Intervallschachtelung). *Es sei $l_j < \hat{l}_j < \hat{r}_j < r_j$ und $g(\alpha_k)$ eine stetige unimodale Funktion auf dem Intervall $[l_j, r_j]$. Bezeichnet man mit α_k^* das Minimum auf (l_j, r_j), dann gilt $\alpha_k^* \in [l_j, \hat{r}_j]$, wenn $g(\hat{l}_j) \leq g(\hat{r}_j)$ bzw. $\alpha_k^* \in \left[\hat{l}_j, r_j\right]$, wenn $g(\hat{l}_j) \geq g(\hat{r}_j)$.*

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

Gemäß Satz 2.6 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 $l_{j+1} = \hat{l}_j,\ r_{j+1} = r_j$.

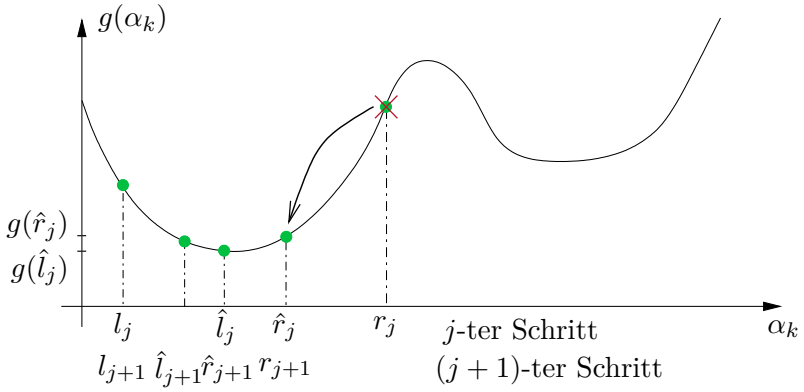


Abbildung 2.3: Veranschaulichung des Intervallschachtelungsverfahrens nach dem Prinzip des „Goldenen Schnittes“.

Für die weitere Betrachtung nehme man an, dass, wie in Abbildung 2.3 dargestellt, $g(\hat{l}_j) \leq g(\hat{r}_j)$ ist. Die nachfolgenden Schritte lassen sich direkt auf den anderen Fall übertragen. Man führt zunächst eine weitere Iteration zur Berechnung der Zwischenpunkte gemäß (2.33) in der Form

$$\hat{l}_{j+1} = al_{j+1} + (1-a)r_{j+1} = (1-a+a^2)l_j + (1-a)ar_j \tag{2.34a}$$

$$\hat{r}_{j+1} = (1-a)l_{j+1} + ar_{j+1} = (1-a^2)l_j + a^2 r_j \tag{2.34b}$$

durch. Aus einem Koeffizientenvergleich von (2.33a) und (2.34b) folgt, dass $\hat{r}_{j+1} = \hat{l}_j$ genau dann erreicht wird, wenn $a^2 = 1 - a$ gilt, d. h. wenn

$$a = \frac{\sqrt{5}-1}{2} \approx 0.618. \tag{2.35}$$

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 $1/a = 1 + a = 1.618$ 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* zwischen den kleinsten drei der vier Funktionswerte gewonnen werden. Das Intervallschachtelungsverfahren ist ein *einfaches und robustes* Verfahren, das allerdings im Vergleich zu anderen Verfahren meist mehr Iterationen benötigt.

2.3.1.2 Quadratische Interpolation

Eine sehr effiziente Methode zur Lösung des skalaren Optimierungsproblems (2.32) besteht darin, durch drei Punkte eine quadratische Interpolationsfunktion zu legen. Dazu nehme man an, dass die voneinander paarweise verschiedenen Punkte $\alpha_{k,1}$, $\alpha_{k,2}$ und $\alpha_{k,3}$ sowie

Initialisierung: l_0, r_0 (Startintervall mit Minimum im Inneren)
$j = 0$ (Startindex)
$a = 0.618$ (Goldener Schnitt-Parameter)
ε_{lr}, ε_g (Schranken 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 $g(\hat{l}_j) > g(\hat{r}_j)$ **do**
    $l_{j+1} \leftarrow \hat{l}_j$
    $r_{j+1} \leftarrow r_j$
    $\hat{l}_{j+1} \leftarrow \hat{r}_j$
    $\hat{r}_{j+1} \leftarrow (1-a) l_{j+1} + a r_{j+1}$
  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 $|r_j - l_j| \leq \varepsilon_{lr}$ **or** $|g(r_j) - g(l_j)| \leq \varepsilon_g$

Tabelle 2.2: Intervallschachtelungsverfahren („Goldener Schnitt“).

deren Funktionswerte $g_1 = g(\alpha_{k,1})$, $g_2 = g(\alpha_{k,2})$ und $g_3 = g(\alpha_{k,3})$ bekannt sind. Die quadratische Interpolationsfunktion $q(\alpha_k)$ durch diese Punkte lautet dann

$$q(\alpha_k) = \sum_{i=1}^{3} g_i \frac{\prod_{j \neq i} (\alpha_k - \alpha_{k,j})}{\prod_{j \neq i} (\alpha_{k,i} - \alpha_{k,j})} \tag{2.36}$$

und der optimale Wert α_k^* errechnet sich zu

$$\alpha_k^* = \frac{1}{2} \frac{g_1\left(\alpha_{k,2}^2 - \alpha_{k,3}^2\right) + g_2\left(\alpha_{k,3}^2 - \alpha_{k,1}^2\right) + g_3\left(\alpha_{k,1}^2 - \alpha_{k,2}^2\right)}{g_1(\alpha_{k,2} - \alpha_{k,3}) + g_2(\alpha_{k,3} - \alpha_{k,1}) + g_3(\alpha_{k,1} - \alpha_{k,2})} \,. \tag{2.37}$$

Aufgabe 2.5. Zeigen Sie die Gültigkeit von (2.37).

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 (2.32),

$$\min_{\alpha_k>0} g(\alpha_k) = f(\mathbf{x}_k + \alpha_k \mathbf{s}_k) \tag{2.38}$$

zugrunde.

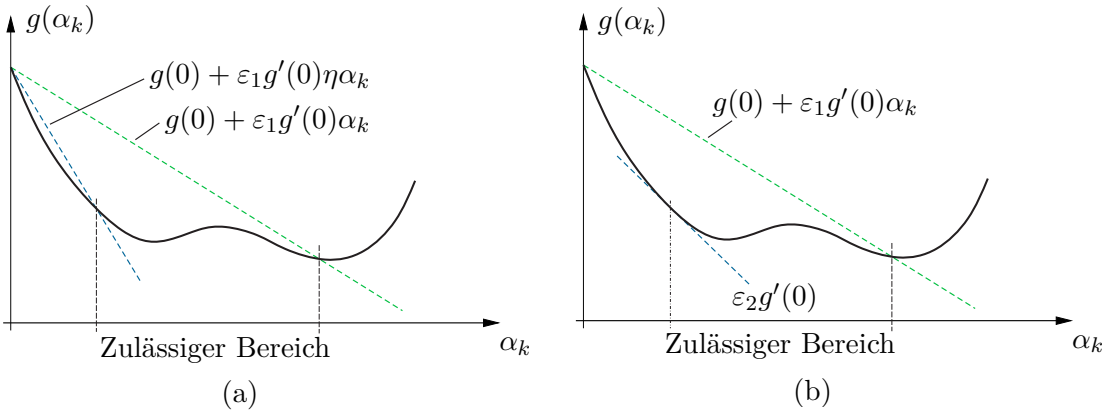


Abbildung 2.4: Veranschaulichung der Armijo- und Wolfe-Bedingung.

Armijo-Bedingung: Entwickelt man $g(\alpha_k)$ um $\alpha_k = 0$ in eine Taylorreihe und bricht nach dem linearen Glied ab, so erhält man

$$g(\alpha_k) \approx g(0) + g'(0)\alpha_k \ . \tag{2.39}$$

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

$$g(\alpha_k) \leq g(0) + \varepsilon_1 g'(0)\alpha_k \tag{2.40}$$

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 einen Parameter $\eta > 1$ ein und fordert zusätzlich, dass die Schrittweite α_k der Ungleichung

$$g(\alpha_k) > g(0) + \varepsilon_1 g'(0)\eta\alpha_k \tag{2.41}$$

genügt. Abbildung 2.4(a) gibt eine grafische Veranschaulichung dieses Sachverhaltes. In der Praxis geht man häufig so vor, dass man in einem ersten Schritt einen (weitgehend beliebigen) Startwert für α_k wählt. Ist für dieses α_k die Ungleichung (2.40) erfüllt, dann erhöht man α_k sukzessive um den Faktor η solange, bis die Ungleichung (2.40) erstmals verletzt wird. Der vorletzte Wert von α_k wird dann als geeignete Schrittweite gewählt. Umgekehrt, wenn der Startwert von α_k die Ungleichung (2.40) nicht erfüllt, dann wird α_k sukzessive durch den Faktor η dividiert, bis erstmals die Ungleichung (2.40) 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-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 (2.40) und

$$g'(\alpha_k) \geq \varepsilon_2 g'(0) \tag{2.42}$$

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 nichtlineare 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 (2.30) den *negativen Gradienten* an der Stelle $\mathbf{x}_k$, d. h. die Abstiegsrichtung

$$\mathbf{s}_k = -(\nabla f)(\mathbf{x}_k) \ . \tag{2.43}$$

Entwickelt man $g(\alpha_k) = f(\mathbf{x}_k + \alpha_k \mathbf{s}_k)$ um den Punkt $\alpha_k = 0$ in eine Taylorreihe mit $\mathbf{s}_k$ gemäß (2.43)

$$g(\alpha_k) = f(\mathbf{x}_k + \alpha_k \mathbf{s}_k) = f(\mathbf{x}_{k+1}) = f(\mathbf{x}_k) - \alpha_k \|(\nabla f)(\mathbf{x}_k)\|_2^2 + o(\alpha_k) \ , \tag{2.44}$$

wobei $o(\alpha_k)$ den Restterm bezeichnet, der schneller nach Null abklingt als α_k, dann erkennt man unmittelbar, dass für hinreichend kleines α_k die Ungleichungsbedingung (2.22) für $(\nabla f)(\mathbf{x}_k) \neq \mathbf{0}$ erfüllt ist.

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

$$\min_{\mathbf{x} \in \mathbb{R}^n} f(\mathbf{x}) = \frac{1}{2} \mathbf{x}^{\mathrm{T}} \mathbf{Q} \mathbf{x} - \mathbf{x}^{\mathrm{T}} \mathbf{b} \tag{2.45}$$

mit der positiv definiten Matrix $\mathbf{Q} \in \mathbb{R}^{n \times n}$. Da die Hessematrix $(\nabla^2 f)(\mathbf{x}) = \mathbf{Q}$ von $f(\mathbf{x})$ positiv definit ist, folgt aus der Eigenschaft (d) konvexer Funktionen von Abschnitt 1.3.4.2 die strikte Konvexität von $f(\mathbf{x})$. Auf Grund von Satz 2.5 errechnet sich daher das globale eindeutige Minimum $\mathbf{x}^*$ von $f(\mathbf{x})$ aus der Beziehung

$$(\nabla f)(\mathbf{x}^*) = \mathbf{g}(\mathbf{x}^*) = \mathbf{Q}\mathbf{x}^* - \mathbf{b} = \mathbf{0} \tag{2.46}$$

zu

$$\mathbf{x}^* = \mathbf{Q}^{-1} \mathbf{b} \ . \tag{2.47}$$

Die Iterationsvorschrift gemäß Gradientenmethode lautet in diesem Fall, siehe (2.30)

$$\mathbf{x}_{k+1} = \mathbf{x}_k - \alpha_k \mathbf{g}_k \qquad \text{mit} \qquad \mathbf{g}_k = \mathbf{g}(\mathbf{x}_k) = \mathbf{Q}\mathbf{x}_k - \mathbf{b}. \tag{2.48}$$

Die optimale Schrittweite α_k^* kann durch explizites Lösen des Optimierungsproblems gemäß (2.32)

$$\min_{\alpha_k > 0} f(\mathbf{x}_k + \alpha_k \mathbf{s}_k) = \frac{1}{2}(\mathbf{x}_k - \alpha_k \mathbf{g}_k)^{\mathrm{T}} \mathbf{Q}(\mathbf{x}_k - \alpha_k \mathbf{g}_k) - (\mathbf{x}_k - \alpha_k \mathbf{g}_k)^{\mathrm{T}} \mathbf{b} \tag{2.49}$$

in der Form

$$\alpha_k^* = \frac{\mathbf{g}_k^{\mathrm{T}} \mathbf{g}_k}{\mathbf{g}_k^{\mathrm{T}} \mathbf{Q} \mathbf{g}_k} \tag{2.50}$$

berechnet werden.

Aufgabe 2.6. Zeigen Sie die Gültigkeit von (2.50).

Zusammenfassend lässt sich damit die Gradientenmethode für die quadratische Kostenfunktion (2.45) wie folgt

$$\mathbf{x}_{k+1} = \mathbf{x}_k - \frac{\mathbf{g}_k^{\mathrm{T}} \mathbf{g}_k}{\mathbf{g}_k^{\mathrm{T}} \mathbf{Q} \mathbf{g}_k} \mathbf{g}_k \ , \quad \mathbf{g}_k = \mathbf{Q}\mathbf{x}_k - \mathbf{b} \tag{2.51}$$

anschreiben.

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

$$F(\mathbf{x}) = f(\mathbf{x}) + \frac{1}{2}(\mathbf{x}^*)^{\mathrm{T}} \mathbf{Q} \mathbf{x}^* = \frac{1}{2}(\mathbf{x} - \mathbf{x}^*)^{\mathrm{T}} \mathbf{Q} (\mathbf{x} - \mathbf{x}^*) \tag{2.52}$$

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

Lemma 2.1 (Zur Konvergenzrate des Gradientenverfahrens)**.** *Mit der Iterationsvorschrift des Gradientenverfahrens* (2.51) *gilt für die Kostenfunktion* $F(\mathbf{x})$ *die Beziehung*

$$F(\mathbf{x}_{k+1}) = \left(1 - \frac{(\mathbf{g}_k^{\mathrm{T}} \mathbf{g}_k)^2}{\mathbf{g}_k^{\mathrm{T}} \mathbf{Q} \mathbf{g}_k \mathbf{g}_k^{\mathrm{T}} \mathbf{Q}^{-1} \mathbf{g}_k}\right) F(\mathbf{x}_k) \ . \tag{2.53}$$

Beweis. Aus (2.47) und (2.51) erhält man

$$\mathbf{g}_k = \mathbf{Q}\mathbf{x}_k - \mathbf{b} = \mathbf{Q}(\mathbf{x}_k - \mathbf{x}^*). \tag{2.54}$$

Folglich gilt

$$F(\mathbf{x}_k) = \frac{1}{2} \mathbf{g}_k^{\mathrm{T}} \mathbf{Q}^{-1} \mathbf{g}_k. \tag{2.55}$$

Aus diesen Beziehungen und der Iterationsvorschrift (2.51) lässt sich nun direkt (2.53)

berechnen.

$$\begin{aligned} F(\mathbf{x}_{k+1}) &= \frac{1}{2}\left(\mathbf{x}_k - \frac{\mathbf{g}_k^{\mathrm{T}}\mathbf{g}_k}{\mathbf{g}_k^{\mathrm{T}}\mathbf{Q}\mathbf{g}_k}\mathbf{g}_k - \mathbf{x}^*\right)^{\mathrm{T}}\mathbf{Q}\left(\mathbf{x}_k - \frac{\mathbf{g}_k^{\mathrm{T}}\mathbf{g}_k}{\mathbf{g}_k^{\mathrm{T}}\mathbf{Q}\mathbf{g}_k}\mathbf{g}_k - \mathbf{x}^*\right) \\ &= \frac{1}{2}(\mathbf{x}_k - \mathbf{x}^*)^{\mathrm{T}}\mathbf{Q}(\mathbf{x}_k - \mathbf{x}^*) - \frac{1}{2}\frac{(\mathbf{g}_k^{\mathrm{T}}\mathbf{g}_k)^2}{\mathbf{g}_k^{\mathrm{T}}\mathbf{Q}\mathbf{g}_k} \\ &= \left(1 - \frac{(\mathbf{g}_k^{\mathrm{T}}\mathbf{g}_k)^2}{\mathbf{g}_k^{\mathrm{T}}\mathbf{Q}\mathbf{g}_k\mathbf{g}_k^{\mathrm{T}}\mathbf{Q}^{-1}\mathbf{g}_k}\right)F(\mathbf{x}_k) \end{aligned} \tag{2.56}$$

□

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

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

$$\frac{\left(\mathbf{x}^{\mathrm{T}}\mathbf{x}\right)^2}{(\mathbf{x}^{\mathrm{T}}\mathbf{Q}\mathbf{x})(\mathbf{x}^{\mathrm{T}}\mathbf{Q}^{-1}\mathbf{x})} \geq \frac{4\lambda_{\min}\lambda_{\max}}{(\lambda_{\min}+\lambda_{\max})^2}\ , \tag{2.57}$$

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

Aufgabe 2.7. Beweisen Sie Lemma 2.2. **H**inweis: Dieser Beweis ist z. B. in [2] skizziert.

Damit lässt sich folgender Satz angeben.

Satz 2.7 (Konvergenz der Gradientenmethode — Quadratische Kostenfunktion)**.** *Für jeden Anfangswert $\mathbf{x}_0 \in \mathbb{R}^n$ konvergiert die Iterationsvorschrift (2.51) der Gradientenmethode gegen das eindeutige globale Minimum $\mathbf{x}^*$ der Kostenfunktion $f(\mathbf{x})$ gemäß (2.45) bzw. $F(\mathbf{x})$ gemäß (2.52)* linear mit der Konvergenzrate

$$F(\mathbf{x}_{k+1}) \leq \left(\frac{\kappa-1}{\kappa+1}\right)^2 F(\mathbf{x}_k)\ , \tag{2.58}$$

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

Beweis. Aus den Lemmas 2.1 und 2.2 folgt unmittelbar

$$F(\mathbf{x}_{k+1}) \leq \left\{1 - \frac{4\lambda_{\min}\lambda_{\max}}{(\lambda_{\min}+\lambda_{\max})^2}\right\}F(\mathbf{x}_k) = \left(\frac{\lambda_{\max}-\lambda_{\min}}{\lambda_{\max}+\lambda_{\min}}\right)^2 F(\mathbf{x}_k)\ . \tag{2.59}$$

□

Satz 2.7 lässt sich nun wie folgt interpretieren. Auf Grund der positiven Definitheit der

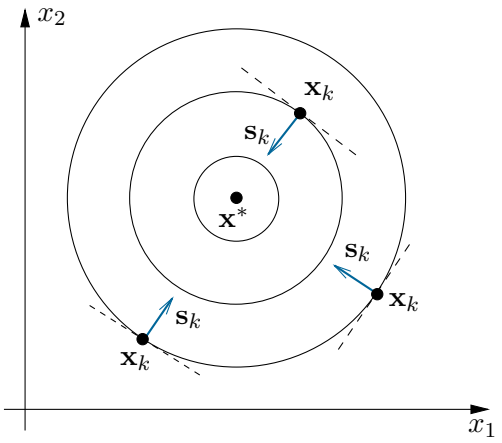


Abb. 2.5: Beispiel eines ideal konditionierten Problems für die Gradientenmethode.

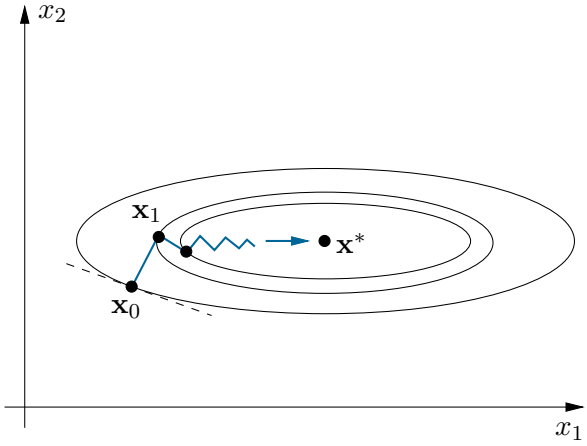


Abb. 2.6: Beispiel eines schlecht konditionierten Problems für die Gradientenmethode.

Matrix $\mathbf{Q}$ sind die Höhenlinien ($f(\mathbf{x}) = \text{konst.}$) der Kostenfunktion (2.45) n-dimensionale Ellipsoide, deren Achsen mit den Richtungen der n paarweise orthogonalen Eigenvektoren der Matrix $\mathbf{Q}$ zusammenfallen und deren Längen invers proportional zum jeweiligen (positiv reellen) Eigenwert sind. Der Gradient $(\nabla f)(\mathbf{x}_k)$ steht orthogonal zur Höhenlinie durch den Punkt $\mathbf{x}_k$, siehe Abbildungen 2.5 und 2.6. Wenn die Eigenwerte von $\mathbf{Q}$ in (2.45) 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) konvergiert die Gradientenmethode sehr langsam, siehe Abbildung 2.6.

Die Gradientenmethode kann natürlich auch auf nichtquadratische Kostenfunktionen angewandt werden. Für diesen Fall beschreibt der nachfolgende Satz das Konvergenzverhalten der Gradientenmethode. Sein Beweis findet sich z. B. in [2].

Satz 2.8 (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*

$$\mathbf{x}_{k+1} = \mathbf{x}_k - \alpha_k (\nabla f)(\mathbf{x}_k) \tag{2.60}$$

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

Schlecht konditionierte Problemstellungen bei der Gradientenmethode können mitunter durch eine geeignete *Skalierung* oder *Transformation* verbessert werden. Die Idee beruht

darauf, dass die Aufgabe, ein Minimum der Kostenfunktion $f(\mathbf{x})$ zu finden, äquivalent dazu ist, für die Funktion $h(\mathbf{y}) = f(\mathbf{T}\mathbf{y})$ mit $\mathbf{x} = \mathbf{T}\mathbf{y}$ und der regulären Matrix $\mathbf{T} \in \mathbb{R}^{n\times n}$ ein Minimum zu suchen. Entwickelt man die Funktion $h(\mathbf{y})$ um den optimalen Punkt $\mathbf{y}^* = \mathbf{T}^{-1}\mathbf{x}^*$ in eine Taylorreihe

$$\begin{aligned} h(\mathbf{y}) &= h(\mathbf{y}^*) + (\nabla h)^{\mathrm{T}}(\mathbf{y}^*)(\mathbf{y}-\mathbf{y}^*) + \frac{1}{2}(\mathbf{y}-\mathbf{y}^*)^{\mathrm{T}}\Big(\nabla^2 h\Big)(\mathbf{y}^*)(\mathbf{y}-\mathbf{y}^*) + \dots \\ &= h(\mathbf{y}^*) + (\nabla f)^{\mathrm{T}}(\mathbf{x}^*)\mathbf{T}(\mathbf{y}-\mathbf{y}^*) + \frac{1}{2}(\mathbf{y}-\mathbf{y}^*)^{\mathrm{T}}\mathbf{T}^{\mathrm{T}}\Big(\nabla^2 f\Big)(\mathbf{x}^*)\mathbf{T}(\mathbf{y}-\mathbf{y}^*) + \dots \ , \end{aligned} \tag{2.61}$$

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

$$\Big(\nabla^2 h\Big)(\mathbf{y}^*) = \mathbf{T}^{\mathrm{T}}\Big(\nabla^2 f\Big)(\mathbf{x}^*)\mathbf{T} \tag{2.62}$$

gegenüber jener der Eigenwerte von $(\nabla^2 f)(\mathbf{x}^*)$ verbessert werden kann. Aus (2.62) folgt, dass mit der idealen Wahl $\mathbf{T} = (\nabla^2 f)^{-\frac{1}{2}}(\mathbf{x}^*)$ für die Hessematrix $(\nabla^2 h)(\mathbf{y}^*) = \mathbf{E}$ mit der Einheitsmatrix $\mathbf{E} \in \mathbb{R}^{n\times n}$ folgen würde und das Gradientenverfahren bei quadratischen Optimierungsproblemen nach einem Schritt konvergieren würde (vgl. Abbildung 2.5). Praktisch ist diese Vorgehensweise sehr ähnlich zur nachfolgend beschriebenen Newton-Methode. Sie hat aber den Nachteil, dass die Hessematrix explizit berechnet werden muss. Um diesen Rechenaufwand zu vermeiden, wird alternativ häufig eine Diagonalmatrix $\mathbf{T}$ verwendet, deren Diagonaleinträge beispielsweise in der Form $T_{ii} = \big((\nabla^2 f)_{ii}(\mathbf{x}^*)\big)^{-\frac{1}{2}}$, $i = 1, \dots, n$ gewählt werden können. Wird eine Diagonalmatrix $\mathbf{T}$ verwendet, so führt dies zu einer reinen Skalierung oder Normierung der Optimierungsvariablen.

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

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

2.3.2.2 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

$$f(\mathbf{x}) \approx f(\mathbf{x}_k) + \mathbf{s}_k^{\mathrm{T}}(\nabla f)(\mathbf{x}_k) + \frac{1}{2}\mathbf{s}_k^{\mathrm{T}}\Big(\nabla^2 f\Big)(\mathbf{x}_k)\mathbf{s}_k \ . \tag{2.63}$$

Die so genannte *Newton-Richtung* $\mathbf{s}_k$ ergibt sich unmittelbar durch Minimierung der rechten Seite von (2.63) bezüglich $\mathbf{s}_k$ in der Form

$$\mathbf{s}_k = -\Big(\nabla^2 f\Big)^{-1}(\mathbf{x}_k)(\nabla f)(\mathbf{x}_k) \ . \tag{2.64}$$

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 (2.64) 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 [2–4] zu finden.

Satz 2.9 (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

$$\mathbf{x}_{k+1} = \mathbf{x}_k - \Big(\nabla^2 f\Big)^{-1}(\mathbf{x}_k)(\nabla f)(\mathbf{x}_k) \tag{2.65}$$

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

Für die praktische Anwendung der Newton-Iteration (2.65) führt man noch eine geeignete Schrittweite $\alpha_k \leq 1$ gemäß (2.31) ein

$$\mathbf{x}_{k+1} = \mathbf{x}_k - \alpha_k \Big(\nabla^2 f\Big)^{-1}(\mathbf{x}_k)(\nabla f)(\mathbf{x}_k) \ . \tag{2.66}$$

Im Zusammenhang mit der Newton-Methode wird α_k auch als *Dämpfungsparameter* bezeichnet. Es ist zu erwarten, dass in der Nähe des Minimums $\alpha_k \approx 1$ ist, weshalb man typischerweise die Iteration 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 häufig auftritt, besteht in dem Verlust der positiven Definitheit von $(\nabla^2 f)(\mathbf{x}_k)$, wenn man zu weit vom Minimum entfernt ist. Dadurch wird die Sinnhaftigkeit der zugrunde liegenden Minimierung der rechten Seite von (2.63) in Frage gestellt und unter Umständen geht die Invertierbarkeit von $(\nabla^2 f)(\mathbf{x}_k)$ verloren. Aus diesem Grund ersetzt man (2.65) gelegentlich durch

$$\mathbf{x}_{k+1} = \mathbf{x}_k - \alpha_k \mathbf{M}_k (\nabla f)(\mathbf{x}_k) \ , \quad \mathbf{M}_k = \Big(\Big(\nabla^2 f\Big)(\mathbf{x}_k) + \varepsilon_k \mathbf{E}\Big)^{-1} \tag{2.67}$$

mit einem geeigneten positiven Parameter ε_k. Man erkennt unmittelbar, dass (2.67) für $\varepsilon_k = 0$ in die Newton-Methode gemäß (2.65) und für sehr große ε_k in die Gradientenmethode gemäß (2.60) übergeht. Eine geeignete Wahl von ε_k erweist sich jedoch als nicht sehr einfach. Typischerweise wird beginnend bei einem Startwert $\varepsilon_k > 0$ sukzessive erhöht, bis die Matrix $(\nabla^2 f)(\mathbf{x}_k) + \varepsilon_k \mathbf{E}$ positiv definit ist. Dies kann sowohl basierend auf den Eigenwerten als auch beispielsweise über die *Cholesky-Faktorisierung* überprüft werden. Für die Cholesky-Faktorisierung gilt nämlich, dass eine Matrix $\mathbf{A}$ genau dann positiv definit ist, wenn sich die Matrix in der Form $\mathbf{A} = \mathbf{G}\mathbf{G}^{\mathrm{T}}$ faktorisieren lässt, wobei $\mathbf{G}$ eine untere Dreiecksmatrix mit positiven Diagonaleinträgen ist.

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

$$\min_{\mathbf{x} \in \mathbb{R}^n} \ \frac{1}{2} \mathbf{x}^{\mathrm{T}} \mathbf{Q} \mathbf{x} - \mathbf{b}^{\mathrm{T}} \mathbf{x} \tag{2.68}$$

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

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

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

2.3.2.3 Konjugierte Gradientenmethode

Die konjugierte Gradientenmethode (Englisch: *conjugate gradient method* oder kurz *C-G method*) versucht nun, die Vorteile der schnellen Konvergenz der Newton-Methode und der Recheneffizienz der Gradientenmethode zu kombinieren. Ursprünglich wurde diese Methode für hochdimensionale quadratische Probleme der Form (siehe auch (2.45))

$$\min_{\mathbf{x}\in\mathbb{R}^n} f(\mathbf{x}) = \frac{1}{2}\mathbf{x}^{\mathrm{T}}\mathbf{Q}\mathbf{x} - \mathbf{x}^{\mathrm{T}}\mathbf{b} \tag{2.69}$$

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

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

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

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

Aufgabe 2.9. Beweisen Sie das Lemma 2.3.

Für das Folgende sei angenommen, dass $\mathbf{d}_0, \mathbf{d}_1, \ldots, \mathbf{d}_{n-1}$ nichttriviale $\mathbf{Q}$-orthogonale Vektoren der Matrix $\mathbf{Q}$ der Kostenfunktion des Optimierungsproblems (2.69) sind. Nach Lemma 2.3 sind die n Vektoren linear unabhängig und spannen daher den $\mathbb{R}^n$ auf. Die optimale Lösung $\mathbf{x}^*$ des Optimierungsproblems (2.69) lässt sich somit als Linearkombination der $\mathbf{Q}$-orthogonalen Vektoren in der Form

$$\mathbf{x}^* = \underbrace{\begin{bmatrix}\mathbf{d}_0 & \mathbf{d}_1 & \ldots & \mathbf{d}_{n-1}\end{bmatrix}}_{=\mathbf{D}} \boldsymbol{\eta} \tag{2.70}$$

mit $\boldsymbol{\eta} \in \mathbb{R}^n$ darstellen. Alternativ zur direkten Berechnung von $\boldsymbol{\eta}$ aus (2.70) erhält man mit $\mathbf{Q}\mathbf{x}^* = \mathbf{b}$ aus (2.46) durch Vormultiplikation von (2.70) mit $\mathbf{D}^{\mathrm{T}}\mathbf{Q}$

$$\mathbf{D}^{\mathrm{T}}\mathbf{Q}\mathbf{x}^* = \mathbf{D}^{\mathrm{T}}\mathbf{b} = \mathbf{D}^{\mathrm{T}}\mathbf{Q}\mathbf{D}\boldsymbol{\eta}. \tag{2.71}$$

Der sich daraus ergebende Wert für $\boldsymbol{\eta}$ wird wieder in (2.70) eingesetzt und man erhält für die optimale Lösung

$$\mathbf{x}^* = \mathbf{D}(\mathbf{D}^{\mathrm{T}}\mathbf{Q}\mathbf{D})^{-1}\mathbf{D}^{\mathrm{T}}\mathbf{b} = \sum_{i=0}^{n-1} \frac{\mathbf{d}_i^{\mathrm{T}}\mathbf{b}}{\mathbf{d}_i^{\mathrm{T}}\mathbf{Q}\mathbf{d}_i}\mathbf{d}_i, \tag{2.72}$$

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

Satz 2.10 (Konjugierte Richtung)**.** *Die Vektoren $\mathbf{d}_0, \mathbf{d}_1, \ldots, \mathbf{d}_{n-1}$ seien nichttriviale $\mathbf{Q}$-orthogonale Vektoren der Matrix $\mathbf{Q}$ der Kostenfunktion des Optimierungsproblems (2.69). Für jeden Anfangswert $\mathbf{x}_0$ konvergiert die Folge*

$$\mathbf{x}_{k+1} = \mathbf{x}_k + \alpha_k\mathbf{d}_k, \quad \alpha_k = -\frac{\mathbf{g}_k^{\mathrm{T}}\mathbf{d}_k}{\mathbf{d}_k^{\mathrm{T}}\mathbf{Q}\mathbf{d}_k}, \quad \mathbf{g}_k = \mathbf{Q}\mathbf{x}_k - \mathbf{b} \tag{2.73}$$

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

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

$$\mathbf{x}^* - \mathbf{x}_0 = \alpha_0\mathbf{d}_0 + \alpha_1\mathbf{d}_1 + \ldots + \alpha_{n-1}\mathbf{d}_{n-1}\ . \tag{2.74}$$

Multipliziert man (2.74) von links mit $\mathbf{Q}$ und bildet man das Skalarprodukt mit $\mathbf{d}_k$, dann erhält man

$$\alpha_k = \frac{\mathbf{d}_k^{\mathrm{T}}\mathbf{Q}(\mathbf{x}^* - \mathbf{x}_0)}{\mathbf{d}_k^{\mathrm{T}}\mathbf{Q}\mathbf{d}_k}\ . \tag{2.75}$$

Aus (2.73) folgt durch rekursives Einsetzen

$$\mathbf{x}_k - \mathbf{x}_0 = \alpha_0\mathbf{d}_0 + \alpha_1\mathbf{d}_1 + \ldots + \alpha_{k-1}\mathbf{d}_{k-1} \tag{2.76}$$

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

$$\mathbf{d}_k^{\mathrm{T}}\mathbf{Q}(\mathbf{x}_k - \mathbf{x}_0) = 0\ . \tag{2.77}$$

Setzt man (2.77) in (2.75) ein, so erhält man unmittelbar das Ergebnis von (2.73)

$$\alpha_k = \frac{\mathbf{d}_k^{\mathrm{T}}\mathbf{Q}(\mathbf{x}^* - \mathbf{x}_k)}{\mathbf{d}_k^{\mathrm{T}}\mathbf{Q}\mathbf{d}_k} = -\frac{\mathbf{d}_k^{\mathrm{T}}(\mathbf{Q}\mathbf{x}_k - \mathbf{b})}{\mathbf{d}_k^{\mathrm{T}}\mathbf{Q}\mathbf{d}_k} = -\frac{\mathbf{d}_k^{\mathrm{T}}\mathbf{g}_k}{\mathbf{d}_k^{\mathrm{T}}\mathbf{Q}\mathbf{d}_k}\ . \tag{2.78}$$

□

Für eine geometrische Interpretation der konjugierten Gradientenmethode betrachte man die linearen Unterräume $\mathcal{B}_k = \text{span}\{\mathbf{d}_0, \mathbf{d}_1, \ldots, \mathbf{d}_{k-1}\}$. Man kann nun mit Hilfe der vollständigen Induktion zeigen, dass der Gradient $\mathbf{g}_k$ zum Iterationsschritt k orthogonal auf den Unterraum $\mathcal{B}_k$ ist. Da $\mathcal{B}_0 = \{\ \}$, ist die Aussage trivialerweise für $k = 0$ erfüllt. Angenommen es gilt $\mathbf{g}_k \perp \mathcal{B}_k$, dann soll im nächsten Schritt gezeigt werden, dass $\mathbf{g}_{k+1} \perp \mathcal{B}_{k+1}$ ist. Aus (2.73) folgt

$$\mathbf{g}_{k+1} = \mathbf{Q}\mathbf{x}_{k+1} - \mathbf{b} = \mathbf{Q}\mathbf{x}_k - \mathbf{b} + \alpha_k \mathbf{Q}\mathbf{d}_k = \mathbf{g}_k + \alpha_k \mathbf{Q}\mathbf{d}_k \tag{2.79}$$

und damit gilt wegen der Definition von α_k gemäß (2.73)

$$\mathbf{d}_k^{\mathrm{T}} \mathbf{g}_{k+1} = \mathbf{d}_k^{\mathrm{T}} \mathbf{g}_k + \alpha_k \mathbf{d}_k^{\mathrm{T}} \mathbf{Q}\mathbf{d}_k = 0\ . \tag{2.80}$$

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

$$\mathbf{d}_j^{\mathrm{T}} \mathbf{g}_{k+1} = \mathbf{d}_j^{\mathrm{T}} \mathbf{g}_k + \alpha_k \mathbf{d}_j^{\mathrm{T}} \mathbf{Q}\mathbf{d}_k = 0 \tag{2.81}$$

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

Die Frage, die es nun noch zu klären gilt, ist, wie die $\mathbf{Q}$-orthogonalen Vektoren $\mathbf{d}_0, \mathbf{d}_1, \ldots, \mathbf{d}_{n-1}$ festgelegt werden. Diese werden bei der konjugierten Gradientenmethode sukzessive bestimmt, wie dies im folgenden Satz beschrieben ist. Seine Herleitung findet sich z. B. in [4].

Satz 2.11 (Konjugierte Gradientenmethode)**.** *Für jeden Anfangswert $\mathbf{x}_0$ konvergiert die Folge*

$$\mathbf{x}_{k+1} = \mathbf{x}_k + \alpha_k \mathbf{d}_k \tag{2.82a}$$

$$\alpha_k = -\frac{\mathbf{d}_k^{\mathrm{T}} \mathbf{g}_k}{\mathbf{d}_k^{\mathrm{T}} \mathbf{Q}\mathbf{d}_k} \tag{2.82b}$$

$$\mathbf{d}_{k+1} = -\mathbf{g}_{k+1} + \beta_k \mathbf{d}_k \tag{2.82c}$$

$$\beta_k = \frac{\mathbf{g}_{k+1}^{\mathrm{T}} \mathbf{Q}\mathbf{d}_k}{\mathbf{d}_k^{\mathrm{T}} \mathbf{Q}\mathbf{d}_k} \tag{2.82d}$$

mit $\mathbf{g}_k = \mathbf{Q}\mathbf{x}_k - \mathbf{b}$ und $\mathbf{d}_0 = -\mathbf{g}_0 = \mathbf{b} - \mathbf{Q}\mathbf{x}_0$ in höchstens n Iterationsschritten gegen die eindeutige optimale Lösung $\mathbf{x}^*$ des Optimierungsproblems (2.69).

Es kann sehr einfach gezeigt werden, dass die iterativ bestimmten Vektoren $\mathbf{d}_i$ die Eigenschaft der $\mathbf{Q}$-Orthogonalität aufweisen. Man erkennt, dass im ersten Iterationsschritt mit $\mathbf{d}_0 = -\mathbf{g}_0$ ein reiner Gradientenschritt durchgeführt wird und anschließend die neue Suchrichtung $\mathbf{d}_{k+1}$ über eine Linearkombination des momentanen Gradienten $\mathbf{g}_{k+1}$ und der vorigen Suchrichtung $\mathbf{d}_k$ bestimmt wird.

Für viele praktische Fragestellungen zeigt die sogenannte *partielle konjugierte Gradientenmethode* große Vorteile. Dabei wird die konjugierte Gradientenmethode von Satz 2.11

lediglich für $m+1 < n$ Iterationsschritte ausgeführt ehe sie mit dem so erhaltenen Punkt als Startlösung neu gestartet wird. Bei jedem Aufruf des Verfahrens werden $m+1$ Iterationen durchgeführt. In diesem Zusammenhang kann folgender Satz angegeben werden, welcher z. B. in [2] bewiesen wird.

Satz 2.12 (Partielle konjugierte Gradientenmethode)**.** *Gegeben ist das Optimierungsproblem* (2.69) *mit der Kostenfunktion* $f(\mathbf{x})$ *oder äquivalent dazu mit der Kostenfunktion* $F(\mathbf{x})$ *gemäß* (2.52)*. Wenn nun die positiv definite Matrix* $\mathbf{Q}$ $n-m$ *Eigenwerte im Intervall* $[l, r]$ $(l > 0)$ *und* m *Eigenwerte größer als* r *besitzt, dann zeigt die partielle konjugierte Gradientenmethode, welche alle* $m+1$ *Schritte neu gestartet wird, das Konvergenzverhalten*

$$F(\mathbf{x}_{k+1}) \leq \left(\frac{r-l}{r+l}\right)^2 F(\mathbf{x}_k) \ . \tag{2.83}$$

Man beachte, dass der Punkt $\mathbf{x}_{k+1}$ durch $(m+1)$-fache Zwischeniteration nach Satz 2.11 mit dem Anfangswert $\mathbf{x}_k$ entsteht. Satz 2.12 zeigt, dass durch Anwendung der partiellen konjugierten Gradientenmethode das schlechte Konvergenzverhalten der Gradientenmethode bei schlecht konditionierten Systemen (vergleiche dazu Satz 2.8) umgangen werden kann.

Für *nichtquadratische Kostenfunktionen* $f(\mathbf{x})$ müssen in Satz 2.11 lediglich die Substitutionen

$$\mathbf{g}_k \leftrightarrow (\nabla f)(\mathbf{x}_k) \quad \text{und} \quad \mathbf{Q} \leftrightarrow \left(\nabla^2 f\right)(\mathbf{x}_k) \tag{2.84}$$

vorgenommen werden. An dieser Stelle ist jedoch zu erwähnen, dass der Algorithmus im Allgemeinen nicht wie im quadratischen Fall in n Schritten terminieren wird. Um die aufwändige Berechnung der Hessematrix $(\nabla^2 f)(\mathbf{x}_k)$ zu vermeiden, kann die Bestimmung von α_k in (2.82b) nach Satz 2.11 über ein Verfahren aus Abschnitt 2.3.1 erfolgen und β_k in (2.82d) wird beispielsweise durch die sogenannte *Formel von Fletcher-Reeves*

$$\beta_k = \frac{\mathbf{g}_{k+1}^{\mathrm{T}} \mathbf{g}_{k+1}}{\mathbf{g}_k^{\mathrm{T}} \mathbf{g}_k} \tag{2.85}$$

ersetzt.

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

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

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

$$\mathbf{g}_{k+1} - \mathbf{g}_k \approx \Big(\nabla^2 f\Big)(\mathbf{x}_k)\mathbf{p}_k \tag{2.86}$$

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

$$\mathbf{q}_k = \mathbf{g}_{k+1} - \mathbf{g}_k = \mathbf{K}\mathbf{p}_k \ . \tag{2.87}$$

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

$$\mathbf{K} = \begin{bmatrix}\mathbf{q}_0 & \mathbf{q}_1 & \ldots & \mathbf{q}_{n-1}\end{bmatrix}\begin{bmatrix}\mathbf{p}_0 & \mathbf{p}_1 & \ldots & \mathbf{p}_{n-1}\end{bmatrix}^{-1} \tag{2.88}$$

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

$$\mathbf{H}_{k+1}\mathbf{q}_j = \mathbf{p}_j, \quad j = 0, \ldots, k \tag{2.89}$$

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

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

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 (2.90) in (2.89) ein, so erhält man für $j = k$

$$\mathbf{p}_k = \mathbf{H}_{k+1}\mathbf{q}_k = \mathbf{H}_k\mathbf{q}_k + \gamma_k \mathbf{z}_k \mathbf{z}_k^{\mathrm{T}}\mathbf{q}_k \ . \tag{2.91}$$

Mit

$$(\mathbf{p}_k - \mathbf{H}_k\mathbf{q}_k)(\mathbf{p}_k - \mathbf{H}_k\mathbf{q}_k)^{\mathrm{T}} = \gamma_k^2 \mathbf{z}_k \mathbf{z}_k^{\mathrm{T}}\mathbf{q}_k\Big(\mathbf{z}_k\mathbf{z}_k^{\mathrm{T}}\mathbf{q}_k\Big)^{\mathrm{T}} = \gamma_k^2 \mathbf{z}_k \underbrace{\mathbf{z}_k^{\mathrm{T}}\mathbf{q}_k\mathbf{q}_k^{\mathrm{T}}\mathbf{z}_k}_{\left(\mathbf{z}_k^{\mathrm{T}}\mathbf{q}_k\right)^2}\mathbf{z}_k^{\mathrm{T}} \tag{2.92}$$

lässt sich (2.90) in der Form

$$\mathbf{H}_{k+1} = \mathbf{H}_k + \frac{(\mathbf{p}_k - \mathbf{H}_k\mathbf{q}_k)(\mathbf{p}_k - \mathbf{H}_k\mathbf{q}_k)^{\mathrm{T}}}{\gamma_k\left(\mathbf{z}_k^{\mathrm{T}}\mathbf{q}_k\right)^2} \tag{2.93}$$

anschreiben. Bildet man von (2.91) das Skalarprodukt mit $\mathbf{q}_k$

$$\mathbf{q}_k^{\mathrm{T}}\mathbf{p}_k = \mathbf{q}_k^{\mathrm{T}}\mathbf{H}_k\mathbf{q}_k + \gamma_k(\mathbf{z}_k^{\mathrm{T}}\mathbf{q}_k)^2 \ , \tag{2.94}$$

dann kann (2.93) wie folgt

$$\mathbf{H}_{k+1} = \mathbf{H}_k + \frac{(\mathbf{p}_k - \mathbf{H}_k\mathbf{q}_k)(\mathbf{p}_k - \mathbf{H}_k\mathbf{q}_k)^{\mathrm{T}}}{\mathbf{q}_k^{\mathrm{T}}(\mathbf{p}_k - \mathbf{H}_k\mathbf{q}_k)} \tag{2.95}$$

geschrieben werden. 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*

$$\mathbf{H}_{j+1} = \mathbf{H}_j + \frac{(\mathbf{p}_j - \mathbf{H}_j\mathbf{q}_j)(\mathbf{p}_j - \mathbf{H}_j\mathbf{q}_j)^{\mathrm{T}}}{\mathbf{q}_j^{\mathrm{T}}(\mathbf{p}_j - \mathbf{H}_j\mathbf{q}_j)} \tag{2.96}$$

die Beziehung

$$\mathbf{p}_j = \mathbf{H}_{k+1}\mathbf{q}_j, \quad j = 0, \ldots, k\ . \tag{2.97}$$

Aufgabe 2.10. Beweisen Sie Satz 2.13 mit vollständiger Induktion. **H**inweis: Dieser Beweis ist z. B. in [4] 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)

$$\mathbf{H}_{k+1} = \mathbf{H}_k + \frac{\mathbf{p}_k\mathbf{p}_k^{\mathrm{T}}}{\mathbf{q}_k^{\mathrm{T}}\mathbf{p}_k} - \frac{\mathbf{H}_k\mathbf{q}_k\mathbf{q}_k^{\mathrm{T}}\mathbf{H}_k}{\mathbf{q}_k^{\mathrm{T}}\mathbf{H}_k\mathbf{q}_k} \tag{2.98}$$

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

$$\mathbf{H}_{k+1} = \left(\mathbf{E} - \frac{\mathbf{p}_k\mathbf{q}_k^{\mathrm{T}}}{\mathbf{q}_k^{\mathrm{T}}\mathbf{p}_k}\right)\mathbf{H}_k\left(\mathbf{E} - \frac{\mathbf{q}_k\mathbf{p}_k^{\mathrm{T}}}{\mathbf{q}_k^{\mathrm{T}}\mathbf{p}_k}\right) + \frac{\mathbf{p}_k\mathbf{p}_k^{\mathrm{T}}}{\mathbf{q}_k^{\mathrm{T}}\mathbf{p}_k}\ . \tag{2.99}$$

Beide werden als *Rang* 2 *Korrekturformeln* bezeichnet, da die aktuelle Approximation der inversen Hessematrix jeweils durch eine Matrix mit Rang 2 korrigiert wird. Linearkombinationen der obigen beiden Formeln in der Art $\mathbf{H}_{k+1} = \phi\mathbf{H}_{k+1}^{\mathrm{DFP}} + (1-\phi)\mathbf{H}_{k+1}^{\mathrm{BFGS}}$ 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 [1, 3, 5]), dass sie die positive Definitheit von $\mathbf{H}_k$ erhalten, wenn

$$\mathbf{q}_k^{\mathrm{T}}\mathbf{p}_k > 0 \tag{2.100}$$

erfüllt ist.

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

$$\mathbf{s}_k = -\mathbf{H}_k(\nabla f)(\mathbf{x}_k) \tag{2.101}$$

gewählt. Man beachte, dass hierfür, anders als bei der Newton-Methode (siehe (2.64)), 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.

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_f, ε_x	(Schwellwerte für Abbruchkriterien)

repeat

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

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

Schritt 3:
$$\mathbf{x}_{k+1} = \mathbf{x}_k + \alpha_k \mathbf{s}_k$$
$$\mathbf{p}_k = \mathbf{x}_{k+1} - \mathbf{x}_k = \alpha_k \mathbf{s}_k$$
$$\mathbf{g}_{k+1} = (\nabla f)(\mathbf{x}_{k+1})$$
$$\mathbf{q}_k = \mathbf{g}_{k+1} - \mathbf{g}_k$$

Schritt 4: BFGS-Korrekturformel

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

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

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

Die Erfüllung der Bedingung (2.100) lässt sich durch eine geeignete Wahl der Schrittweite α_k in $\mathbf{x}_{k+1} = \mathbf{x}_k + \alpha_k \mathbf{s}_k$ sicherstellen. Genügt die Wahl der Schrittweite beispielsweise der Wolfe-Bedingung gemäß Abschnitt 2.3.1.3, so ist (2.100) automatisch erfüllt. Um dies zu sehen, beachte man, dass aus (2.42)

$$g'(\alpha_k) = \mathbf{g}_{k+1}^{\mathrm{T}} \mathbf{s}_k \geq \varepsilon_2 g'(0) = \varepsilon_2 \mathbf{g}_k^{\mathrm{T}} \mathbf{s}_k \tag{2.102}$$

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

$$(\mathbf{g}_{k+1} - \mathbf{g}_k)^{\mathrm{T}} \mathbf{s}_k = \mathbf{q}_k^{\mathrm{T}} \mathbf{s}_k \geq \underbrace{(\varepsilon_2 - 1) \mathbf{g}_k^{\mathrm{T}} \mathbf{s}_k}_{> 0} , \tag{2.103}$$

wobei die rechte Seite dieser Ungleichung (abseits des optimalen Punktes $\mathbf{x}^*$) strikt positiv sein muss, da $\mathbf{s}_k$ eine Abstiegsrichtung ist. Mit $\mathbf{p}_k = \mathbf{x}_{k+1} - \mathbf{x}_k = \alpha_k \mathbf{s}_k$ folgt schließlich aus (2.103), dass

$$\mathbf{q}_k^{\mathrm{T}} \mathbf{p}_k = \alpha_k \mathbf{q}_k^{\mathrm{T}} \mathbf{s}_k \geq \alpha_k (\varepsilon_2 - 1) \mathbf{g}_k^{\mathrm{T}} \mathbf{s}_k > 0 \tag{2.104}$$

gilt und (2.100) somit erfüllt ist.

Für unbeschränkte nichtlineare Optimierungsprobleme mit konvexer Kostenfunktion $f(\mathbf{x})$ konvergiert die Quasi-Newton-Methode mit superlinearer Konvergenzordnung. Für das unbeschränkte quadratische Optimierungsproblem (2.69) konvergiert die Quasi-Newton-Methode nach spätestens n Iterationsschritten. Konvergiert die Methode in

diesem Fall genau nach n Iterationsschritten, so kann gezeigt werden (siehe [1]), dass $\mathbf{H}_n = \left(\nabla^2 f\right)^{-1}(\mathbf{x}^*) = \mathbf{Q}^{-1}$, d. h. der Algorithmus liefert die exakte inverse Hessematrix.

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

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

2.4 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 (2.43) bei der Gradientenmethode oder die *Newton-Richtung* gemäß (2.64) bei der Newton-Methode) gewählt und anschließend wird über das skalare Optimierungsproblem (2.32) 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

$$f(\mathbf{x}_k + \mathbf{s}_k) \approx m_k(\mathbf{s}_k) = f(\mathbf{x}_k) + \mathbf{s}_k^{\mathrm{T}}(\nabla f)(\mathbf{x}_k) + \frac{1}{2}\mathbf{s}_k^{\mathrm{T}}\mathbf{B}_k\mathbf{s}_k \tag{2.105}$$

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

$$\begin{aligned} \min_{\mathbf{s}_k \in \mathbb{R}^n} \quad & m_k(\mathbf{s}_k) = f(\mathbf{x}_k) + \mathbf{s}_k^{\mathrm{T}}(\nabla f)(\mathbf{x}_k) + \frac{1}{2}\mathbf{s}_k^{\mathrm{T}}\mathbf{B}_k\mathbf{s}_k \\ \text{u.B.v.} \quad & \|\mathbf{s}_k\| \leq \Delta_k \end{aligned} \tag{2.106}$$

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

$$\rho_k(\mathbf{s}_k) = \frac{f(\mathbf{x}_k) - f(\mathbf{x}_k + \mathbf{s}_k)}{m_k(\mathbf{0}) - m_k(\mathbf{s}_k)} \tag{2.107}$$

Initialisierung: $\bar{\Delta}, \Delta_0 \in (0, \bar{\Delta})$ (Vertrauensbereich: Grenz- & Startwert)
$\eta \in [0, \frac{1}{4})$ (Parameter)
$k \leftarrow 0$ (Iterationsindex)
ε_x, ε_f (Abbruchkriterien)

repeat

$m_k(\mathbf{s}_k)$ nach (2.105) (Modell)
$\mathbf{s}_k$ Lösung von (2.106) (evtl. approximativ gelöst)
ρ_k nach (2.107) (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** $||\mathbf{s}_k|| = \Delta_k$
$\Delta_{k+1} \leftarrow \min\{2\Delta_k, \bar{\Delta}\}$ (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)
$\mathbf{B}_{k+1} \leftarrow \mathbf{B}_k + \ldots$ (Aktualisierung der Ansatzfunktion)
else
$\mathbf{x}_{k+1} \leftarrow \mathbf{x}_k$ (Schritt mit $\Delta_{k+1} < \Delta_k$ wiederholen)
end if

$k \leftarrow k + 1$

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

Tabelle 2.4: Methode der Vertrauensbereiche.

berechnet wird. Der Zählerterm in (2.107) beschreibt die *tatsächliche Reduktion* der Kostenfunktion während der Nennerterm die *prädizierte Reduktion* wiedergibt. Der Nennerterm von $\rho_k(\mathbf{s}_k)$ ist stets größer gleich Null, da $\mathbf{s}_k$ die Funktion m_k gemäß (2.106) innerhalb des Vertrauensbereiches minimiert und der Punkt $\mathbf{s}_k = \mathbf{0}$ im Vertrauensbereich liegt. Ist nun $\rho_k(\mathbf{s}_k) < 0$, so bedeutet dies, dass der Wert der Kostenfunktion am nächsten Iterationspunkt $f(\mathbf{x}_k + \mathbf{s}_k)$ größer als am vorigen Iterationspunkt $f(\mathbf{x}_k)$ ist, weshalb dieser Iterationsschritt verworfen und der Vertrauensbereich verkleinert werden muss. Andererseits kann bei $\rho_k(\mathbf{s}_k) \approx 1$ der Vertrauensbereich vergrößert werden, da die Kostenfunktion $f(\mathbf{x}_k)$ in diesem Fall gut von der Ansatzfunktion beschrieben wird. Für den Fall, dass $\rho_k(\mathbf{s}_k)$ positiv und deutlich kleiner als 1 ist, wird der Vertrauensbereich im nächsten Schritt verkleinert.

Der Algorithmus der Methode der Vertrauensbereiche ist in Tabelle 2.4 aufgelistet. Man beachte, dass hier $\bar{\Delta}$ die obere Grenze des zulässigen Vertrauensbereiches beschreibt und dass eine Vergrößerung des Vertrauensbereiches im nächsten Iterationsschritt nur dann stattfindet, wenn $\mathbf{s}_k$ durch die Grenze des Vertrauensbereiches beschränkt wurde (Bedingung $\|\mathbf{s}_k\| = \Delta_k$). Die genaue praktische Ausführung des Algorithmus, insbesondere die Iterationsvorschrift für die Matrix $\mathbf{B}_k$, wird z. B. in [1, 3, 4] beschrieben.

2.5 Direkte Suchverfahren

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

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

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

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

$$\begin{aligned} f(\mathbf{x}_{k,\min}) &= \min_{i=0,\ldots,n} f(\mathbf{x}_{k,i}) \\ f(\mathbf{x}_{k,\max}) &= \max_{i=0,\ldots,n} f(\mathbf{x}_{k,i}) \ . \end{aligned} \tag{2.108}$$

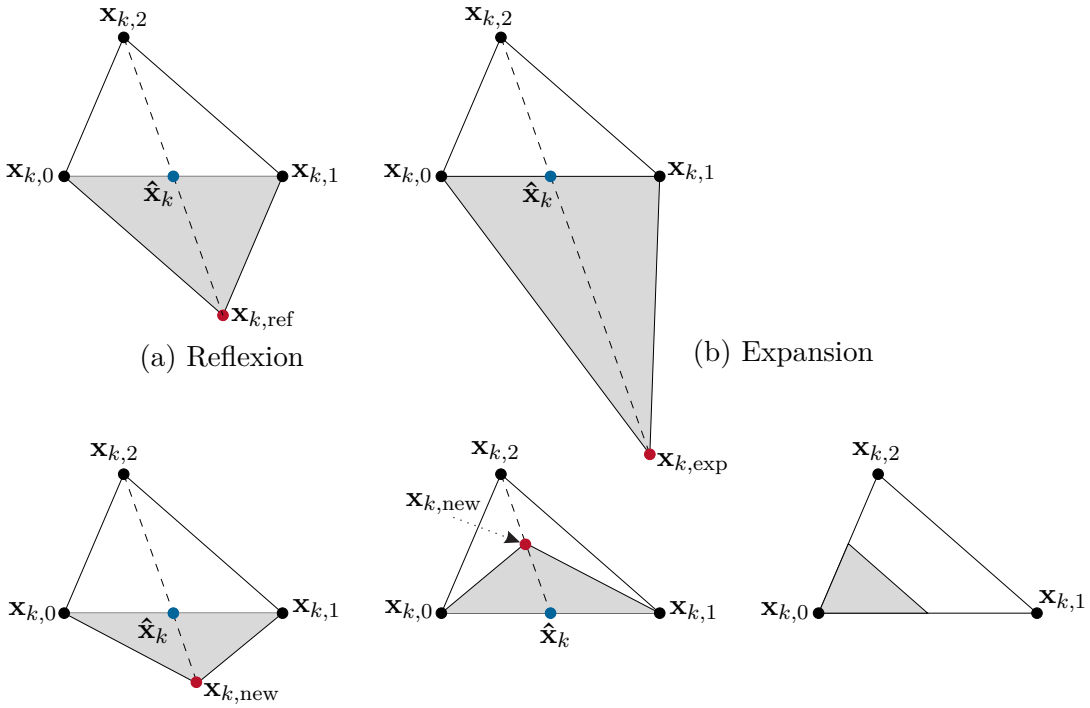
Der *Schwerpunkt* oder Mittelpunkt des Simplex $\hat{\mathbf{x}}_k$ gebildet durch alle Punkte außer $\mathbf{x}_{k,\max}$ errechnet sich zu

$$\hat{\mathbf{x}}_k = \frac{1}{n}\left(\sum_{i=0}^{n} \mathbf{x}_{k,i} - \mathbf{x}_{k,\max}\right) \ . \tag{2.109}$$

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

$$\mathbf{x}_{k,\text{ref}} = \hat{\mathbf{x}}_k + (\hat{\mathbf{x}}_k - \mathbf{x}_{k,\max}) \ , \tag{2.110}$$

der auf einer Geraden durch die Punkte $\mathbf{x}_{k,\max}$ und $\hat{\mathbf{x}}_k$ liegt und symmetrisch bezüglich $\hat{\mathbf{x}}_k$ zu $\mathbf{x}_{k,\max}$ ist, siehe Abbildung 2.7(a). Abhängig von $f(\mathbf{x}_{k,\text{ref}})$ im Vergleich zu den



(a) Reflexion (b) Expansion

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

Abbildung 2.7: Operationen des Simplex-Verfahrens nach Nelder und Mead für $\beta = 1$, $\gamma = 1$ und $\theta = 1/2$.

Funktionswerten der anderen Punkte des Simplex außer $\mathbf{x}_{k,\max}$ wird ein neuer Punkt $\mathbf{x}_{k,\text{new}}$ konstruiert, der im nächsten Iterationsschritt den Punkt $\mathbf{x}_{k,\max}$ ersetzt. Der Algorithmus ist in seiner Grundfunktion in Tabelle 2.5 aufgelistet und die unterschiedlichen Operationen sind grafisch in Abbildung 2.7 dargestellt. Man beachte, dass die Schrumpfung im Algorithmus von Tabelle 2.5 stets bezüglich des minimierenden Punktes $\mathbf{x}_{k,\min}$ ausgeführt wird, d. h. in Abbildung 2.7(e) gilt $\mathbf{x}_{k,1} = \mathbf{x}_{k,\min}$. Während der Iteration wandert der Simplex in Richtung des Optimums und zieht sich sukzessive zusammen. Allerdings ist die Konvergenz im Allgemeinen nicht garantiert und es kann vorkommen, dass das Simplex-Verfahren zu einem *nicht-optimalen Punkt* konvergiert. In der Praxis führt das Simplex-Verfahren dennoch häufig zu guten Ergebnissen und akzeptablem Konvergenzverhalten.

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 $\theta = 1/2$)
$\varepsilon_x,\ \varepsilon_f$ (Abbruchkriterien)

repeat

$\mathbf{x}_{k,\min}$, $\mathbf{x}_{k,\max}$ gemäß (2.108) (Punkte mit min. und max. Kostenfunktionswert)

$\hat{\mathbf{x}}_k$ gemäß (2.109) (Schwerpunkt)

$\mathbf{x}_{k,\text{ref}} = \hat{\mathbf{x}}_k + \beta(\hat{\mathbf{x}}_k - \mathbf{x}_{k,\max})$ (Reflexionsschritt, Abb. 2.7(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. 2.7(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. 2.7(d))

else

$\mathbf{x}_{k,\text{new}} = \theta\mathbf{x}_{k,\text{ref}} + (1-\theta)\hat{\mathbf{x}}_k$ (Äußere Kontraktion, Abb. 2.7(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. 2.7(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 $\|\mathbf{x}_k - \mathbf{x}_{k-1}\| \leq \varepsilon_x$ **or** $|f(\mathbf{x}_k) - f(\mathbf{x}_{k-1})| \leq \varepsilon_f$

Tabelle 2.5: Simplex-Verfahren nach Nelder und Mead.

2.6 Beispiel: Rosenbrock's „Bananenfunktion“

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

$$\min_{\mathbf{x}\in\mathbb{R}^2} \; f(\mathbf{x}) \quad \text{mit} \quad f(\mathbf{x}) = 100(x_2 - x_1^2)^2 + (x_1 - 1)^2 \,. \tag{2.111}$$

Abbildung 2.8 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.

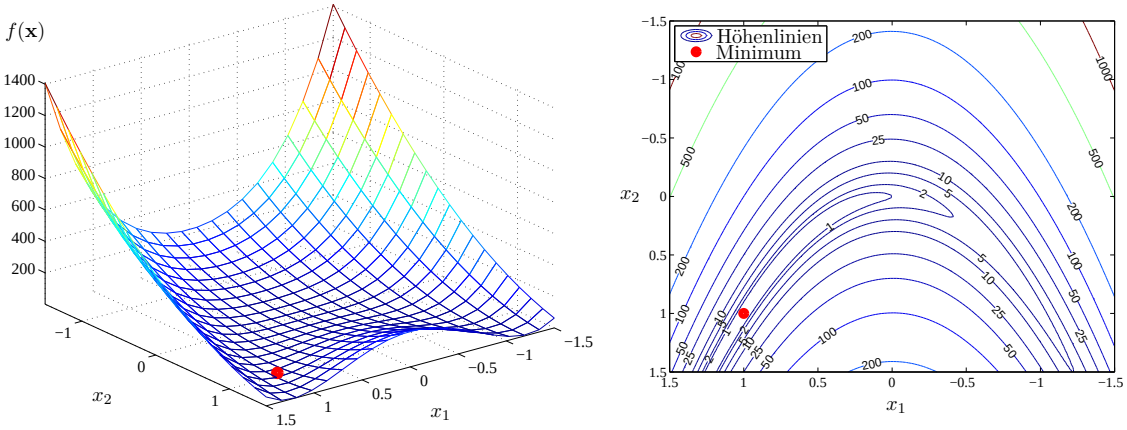


Abbildung 2.8: Profil und Höhenlinien von Rosenbrock's Bananenfunktion.

> *Aufgabe* 2.11. Verifizieren Sie, dass der Punkt $\mathbf{x}^* = \begin{bmatrix} 1 & 1 \end{bmatrix}^{\mathrm{T}}$ ein Minimum darstellt. Ist das Minimum $\mathbf{x}^*$ global und eindeutig? Sind die Funktion $f(\mathbf{x})$ und das Optimierungsproblem (2.111) konvex?

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

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

Eine empfehlenswerte Alternative ist die frei zugängliche MATLAB-Funktion minFunc [6], die eine große Auswahl an Liniensuchverfahren bietet. Tabelle 2.6 zeigt einige Vergleichsdaten für die numerische Lösung des Rosenbrock-Problems (ausgehend vom Startwert $\mathbf{x}_0 = \begin{bmatrix} -1 & -1 \end{bmatrix}^{\mathrm{T}}$), die mit Hilfe von fminunc, fminsearch und minFunc berechnet wurden.

Abbildung 2.10 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 (2.111) angegeben, um zu verdeutlichen, wie die einzelnen Optimierungsverfahren mit fminunc und fminsearch angesprochen werden können.

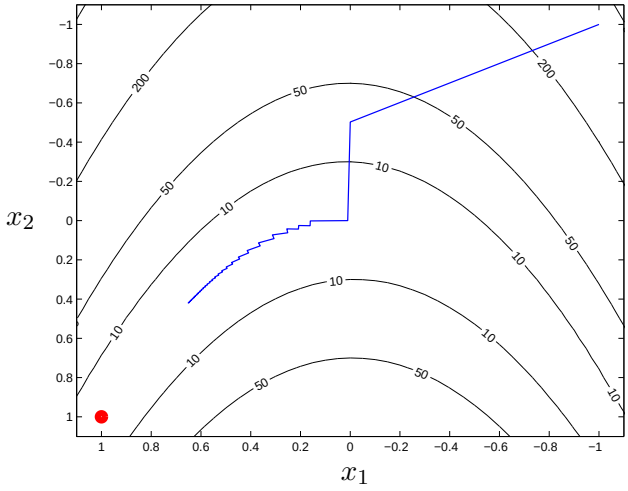


Abbildung 2.9: Darstellung der Iterationen des Gradientenverfahrens.

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: Gradientenverfahren	fminunc	57	0.1232	1.1978	200	200	–
LS: Konj. Gradientenm.	minFunc	28	$6.9 \cdot 10^{-18}$	$9.6 \cdot 10^{-8}$	68	68	–
LS: Newton-Verfahren	minFunc	20	$3.8 \cdot 10^{-16}$	$7.3 \cdot 10^{-7}$	32	32	26
LS: Quasi-Newton (BFGS)	fminunc	23	$5.4 \cdot 10^{-12}$	$9.2 \cdot 10^{-6}$	29	29	–
VB: Newton-Verfahren	fminunc	25	$2.2 \cdot 10^{-18}$	$2.1 \cdot 10^{-8}$	26	26	26
Nelder-Mead Simplex-Verf.	fminsearch	67	$5.3 \cdot 10^{-10}$	–	125	–	–

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

Beim *Gradientenverfahren* fällt die *langsame Konvergenz* auf, weil auch nach dem Erreichen der maximalen Anzahl an Funktionsauswertungen von 200 das Minimum noch immer nicht erreicht ist. In Abbildung 2.9 ist der Iterationsverlauf des Gradientenverfahrens über den Höhenlinien der Rosenbrock-Funktion (2.111) dargestellt. Erkennbar ist, dass sich das Gradientenverfahren an der maximalen Abstiegsrichtung orientiert, die orthogonal zur jeweiligen Höhenlinie verläuft. In Richtung des Minimums werden die Iterationsschritte immer kleiner. Die niedrige Konvergenzgeschwindigkeit wurde bereits in Abbildung 2.6 veranschaulicht und soll in der folgenden Aufgabe näher untersucht werden.

Aufgabe 2.12. Berechnen Sie für das Minimum $\mathbf{x}^* = \begin{bmatrix} 1 & 1 \end{bmatrix}^{\mathrm{T}}$ des Rosenbrock-Problems (2.111) die Konvergenzrate des Gradientenverfahrens gemäß Satz 2.8.

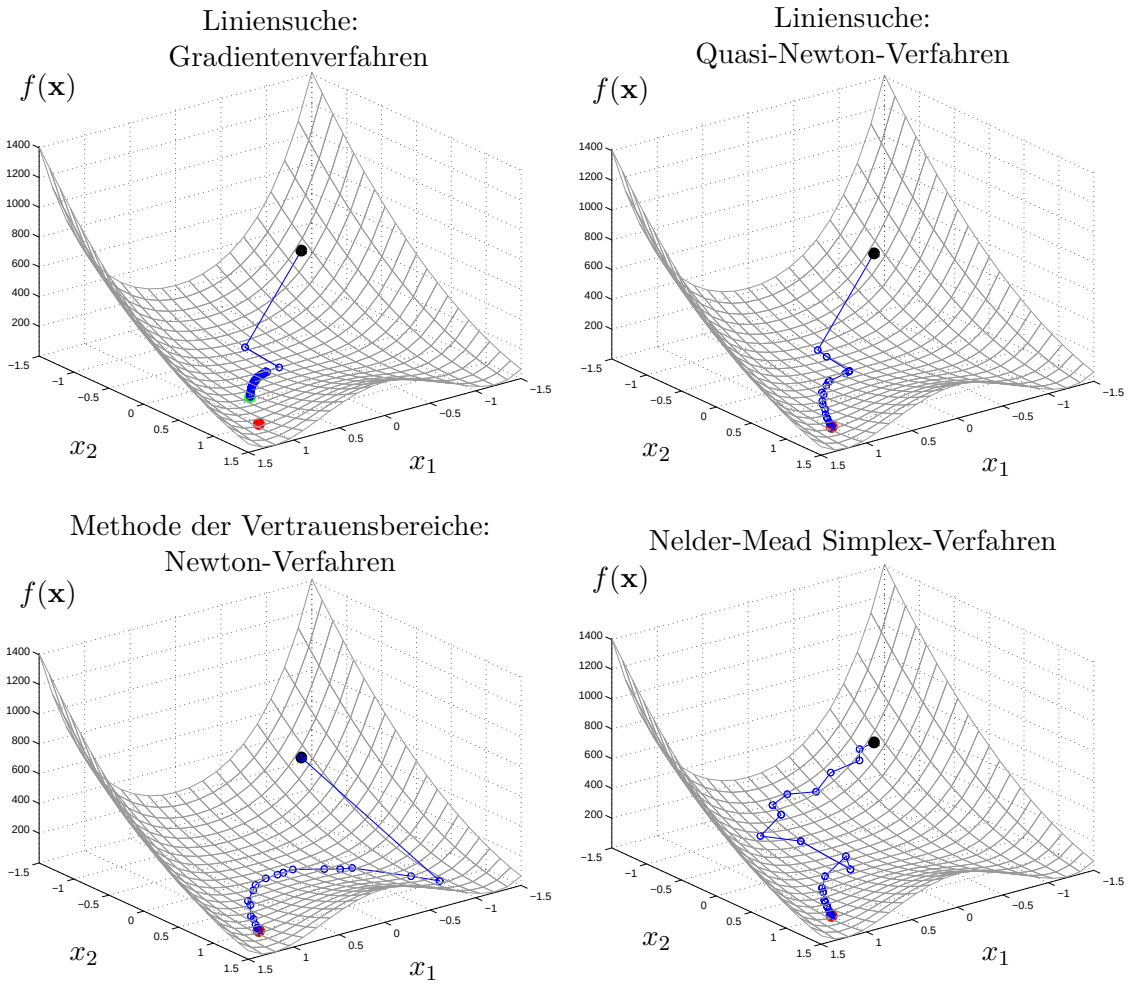


Abbildung 2.10: Rosenbrock's Bananenfunktion: Vergleich der numerischen Verfahren mit fminunc und fminsearch.

Das Konvergenzverhalten des *Quasi-Newton-Verfahrens* in Abbildung 2.10 ist wesentlich besser als beim Gradientenverfahren. Das *Newton-Verfahren (Methode der Vertrauensbereiche)* in Abbildung 2.10 startet zunächst in die „falsche“ Richtung, was durch die *quadratische Approximation* (2.105) am Startpunkt $\mathbf{x}_0 = \begin{bmatrix} -1 & -1 \end{bmatrix}^{\mathrm{T}}$ zu erklären ist, deren Minimum in der Nähe von $\mathbf{x} \approx \begin{bmatrix} -1 & 1 \end{bmatrix}^{\mathrm{T}}$ liegt. Allerdings sind die einzelnen Schritte entlang des Tales der Rosenbrock-Funktion deutlich größer, da das Newton-Verfahren die *Hessematrix explizit verwendet* und nicht auf eine Approximation angewiesen ist wie im Fall des Quasi-Newton-Verfahrens.

Zusätzlich sind in Tabelle 2.6 und Abbildung 2.10 die Ergebnisse für das *Simplex-Verfahren von Nelder-Mead* angegeben, die mit der MATLAB-Funktion fminsearch erzielt wurden. Allerdings bietet die Grafikausgabe unter fminsearch nicht die Möglichkeit, die

einzelnen Simplexe darzustellen. In der nächsten Aufgabe soll das Simplex-Verfahren deshalb näher untersucht werden, um einen Eindruck von den Simplex-Operationen und der Robustheit des Verfahrens zu erhalten.

Aufgabe 2.13. Schreiben Sie eine MATLAB-Funktion, die das Rosenbrock-Problem (2.111) mit Hilfe des Simplex-Verfahrens nach Nelder-Mead numerisch löst (siehe den Algorithmus von Tabelle 2.5). Konstruieren Sie den ersten Simplex mit den Eckpunkten

$$\mathbf{x}_{0,1} = \begin{bmatrix} -1 \\ -1 \end{bmatrix}, \quad \mathbf{x}_{0,2} = \mathbf{x}_{0,1} + s \begin{bmatrix} 1 \\ 0 \end{bmatrix}, \quad \mathbf{x}_{0,3} = \mathbf{x}_{0,1} + s \begin{bmatrix} 0 \\ 1 \end{bmatrix} \tag{2.112}$$

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

Aufgabe 2.14. Schreiben Sie eine MATLAB-Funktion, die das Rosenbrock-Problem (2.111) mit Hilfe des Newton-Verfahrens (Liniensuche) löst, siehe Abschnitt 2.3.2.2. 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 2.6. Stellen Sie die einzelnen Iterationen grafisch dar und variieren Sie die Startpunkte $\mathbf{x}_0$.

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

```
function Xopt = rosenbrock_problem(Xinit,methodQ)
% ------------------------------------------------------------
% Xinit:  Startpunkt
% methodQ: 1 - Liniensuche: Gradientenverfahren
%          2 - Liniensuche: Quasi-Newton
%          3 - Methode der Vertrauensbereiche: Newton-Verfahren
%          4 - Nelder-Mead Simplex-Verfahren

global old
old = [Xinit; rosenbrock(Xinit)];

% Optionen für Ausgabe
opt_fminu = optimoptions('fminunc','Display','iter','PlotFcns',@plot_iterates);
opt_fmins = optimset('Display','iter','PlotFcns',@plot_iterates);

switch methodQ
 case 1,     % Liniensuche: Gradientenverfahren
  opt_fminu = optimoptions(opt_fminu,'Algorithm','quasi-newton',...
                           'HessUpdate','steepdesc','GradObj','on');
 case 2,     % Liniensuche: Quasi-Newton
  opt_fminu = optimoptions(opt_fminu,'Algorithm','quasi-newton',...
                           'HessUpdate','bfgs','GradObj','on');
 case 3,     % Methode der Vertrauensbereiche: Newton-Verfahren
  opt_fminu = optimoptions(opt_fminu,'Algorithm','trust-region','Hessian','on',...
                           'GradObj','on');
```

```
end

% numerische Lösung mit...
if methodQ<4,  Xopt = fminunc(@rosenbrock,Xinit,opt_fminu);    % fminunc
else           Xopt = fminsearch(@rosenbrock,Xinit,opt_fmins); % fminsearch
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

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];

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');                                % figure settings
set(gca,'Xdir','reverse','Ydir','reverse');
```

2.7 Literatur

[1] D. P. Bertsekas, *Nonlinear Programming*, 2. Aufl. Athena Scientific, 1999.

[2] D. G. Luenberger und Y. Ye, *Linear and Nonlinear Programming*, 3. Aufl., Ser. International Series in Operations Research & Management Science. Springer, 2008, Bd. 116.

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

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

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

[6] M. Schmidt, „minFunc“, abrufbar unter http://www.cs.ubc.ca/~schmidtm/Software/minFunc.html, University of British Columbia, Vancouver, 2005.

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

[8] M. Papageorgiou, M. Leibold und M. Buss, *Optimierung: Statische, dynamische, stochastische Verfahren für die Anwendung*, 3. Aufl. Springer, 2012.

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

[10] 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.

[11] 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 $\mathbf{g}(\mathbf{x}) = \begin{bmatrix} g_1(\mathbf{x}) & \dots & g_p(\mathbf{x}) \end{bmatrix}^{\mathrm{T}}$ und $\mathbf{h}(\mathbf{x}) = \begin{bmatrix} h_1(\mathbf{x}) & \dots & h_q(\mathbf{x}) \end{bmatrix}^{\mathrm{T}}$ zusammen, so kann das beschränkte Optimierungsproblem (3.1) in der Form

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

angeschrieben werden.

3.1 Optimalitätsbedingungen

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

3.1.1 Gleichungsbeschränkungen

Treten ausschließlich Gleichungsbeschränkungen auf, so soll mit

$$\mathcal{S} = \{\mathbf{x} \in \mathbb{R}^n | g_i(\mathbf{x}) = 0,\ i = 1,\dots,p\} \tag{3.3}$$

die *Menge der zulässigen Punkte* beschrieben werden. $\mathcal{S}$ wird auch *zulässige Menge* genannt.

Definition 3.1 (Regulärer Punkt bei Gleichungsbeschränkungen, LICQ)**.** Ein zulässiger Punkt $\mathbf{x} \in \mathcal{S}$ der Optimierungsaufgabe (3.2) mit ausschließlich Gleichungsbeschränkungen ($q = 0$) ist *regulär*, wenn die Gradientenvektoren $(\nabla g_i)(\mathbf{x})$, $i = 1, \ldots, p$ linear unabhängig sind. D. h. die Bedingung

$$\operatorname{rang}((\nabla \mathbf{g})(\bar{\mathbf{x}})) = \operatorname{rang}\Big(\Big[(\nabla g_1)(\bar{\mathbf{x}}) \quad (\nabla g_2)(\bar{\mathbf{x}}) \quad \ldots \quad (\nabla g_p)(\bar{\mathbf{x}})\Big]\Big) = p, \tag{3.4}$$

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

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

Man beachte, dass die Regularität eines Punktes gemäß Definition 3.1 direkt von der Formulierung der Gleichungsbeschränkungen abhängt. Als Beispiel dazu betrachte man die zunächst äquivalent erscheinenden Gleichungsbeschränkungen $g(x_1, x_2) = x_1 = 0$ und $g(x_1, x_2) = x_1^2 = 0$ im $\mathbb{R}^2$. Beide Beschränkungen definieren die gleiche zulässige Menge $\mathcal{S}$, nämlich die gesamte x_2-Achse. Für $g(x_1, x_2) = x_1$ ist jeder Punkt von $\mathcal{S}$ ein regulärer Punkt gemäß der Definition 3.1. Für $g(x_1, x_2) = x_1^2$ jedoch ist kein Punkt von $\mathcal{S}$ regulär.

Die zulässige Menge $\mathcal{S} = \{\mathbf{x} \in \mathbb{R}^n | g_i(\mathbf{x}) = 0,\ i = 1, \ldots, p\}$ mit den stetig differenzierbaren Funktionen $g_i(\mathbf{x})$, $i = 1, \ldots, p$ beschreibt eine $(n-p)$-*dimensionale* C^1-*Mannigfaltigkeit* (siehe dazu Anhang A des Skriptums Regelungssysteme 2). Der Tangentialraum $\mathcal{T}_{\bar{\mathbf{x}}}\mathcal{S}$ an einem regulären Punkt $\bar{\mathbf{x}}$ wird durch $n-p$ linear unabhängige Vektoren $\mathbf{d}_j(\bar{\mathbf{x}})$, $j = 1, \ldots, n-p$ aufgespannt. $\mathcal{T}_{\bar{\mathbf{x}}}\mathcal{S}$ lässt sich nun als Annullator des Kotangentialraumes, welcher durch die exakten Differentiale $\mathrm{d}g_i : \mathcal{T}_{\bar{\mathbf{x}}}\mathcal{S} \to \mathbb{R}$, $i = 1, \ldots, p$ gebildet wird, definieren. Das heißt, es gilt

$$\mathcal{T}_{\bar{\mathbf{x}}}\mathcal{S} = \left\{\mathbf{d} \;\middle|\; \mathrm{d}g_i(\mathbf{d})|_{\mathbf{x}=\bar{\mathbf{x}}} = \mathrm{L}_{\mathbf{d}} g_i(\bar{\mathbf{x}}) = \underbrace{\left(\frac{\partial}{\partial \mathbf{x}} g_i(\bar{\mathbf{x}})\right)}_{(\nabla g_i)^{\mathrm{T}}(\bar{\mathbf{x}})} \mathbf{d} = 0,\ i = 1, \ldots, p\right\}. \tag{3.5}$$

Als Vorbereitung für die Formulierung notwendiger Optimalitätsbedingungen des Optimierungsproblems (3.1) mit reinen Gleichungsnebenbedingungen ($q = 0$) sei folgendes Lemma angegeben.

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

$$(\nabla \mathbf{g})^{\mathrm{T}}(\mathbf{x}^*)\mathbf{d} = \mathbf{0} \tag{3.6}$$

erfüllen, muss auch gelten

$$(\nabla f)^{\mathrm{T}}(\mathbf{x}^*)\mathbf{d} = 0\ . \tag{3.7}$$

Beweisskizze: Da $\mathbf{x}^*$ ein regulärer Punkt von $\mathcal{S}$ ist, liegt jedes $\mathbf{d}$, das (3.6) erfüllt, gemäß (3.5) im Tangentialraum $\mathcal{T}_{\mathbf{x}^*}\mathcal{S}$. Mit $\mathbf{x}(t)$, $t \in (-a, a)$, $a > 0$ bezeichne man im Weiteren eine stetig differenzierbare Kurve parametriert in t durch den Punkt $\mathbf{x}^*$ mit dem Tangentialvektor $\mathbf{d}$ so, dass gilt $\mathbf{x}(0) = \mathbf{x}^*$ und $\left(\frac{\mathrm{d}}{\mathrm{d}t}\mathbf{x}\right)(0) = \mathbf{d}$. Da nun $\mathbf{x}^*$ ein lokaler Extremalpunkt ist, muss die Beziehung

$$\frac{\mathrm{d}}{\mathrm{d}t} f(\mathbf{x}(t))\bigg|_{t=0} = \underbrace{\left(\frac{\partial}{\partial \mathbf{x}} f\right)(\mathbf{x}(0))}_{(\nabla f)^{\mathrm{T}}(\mathbf{x}^*)} \underbrace{\left(\frac{\mathrm{d}}{\mathrm{d}t}\mathbf{x}\right)(0)}_{\mathbf{d}} = 0 \tag{3.8}$$

gelten. □

Lemma 3.1, speziell (3.7), besagt also, dass $(\nabla f)(\mathbf{x}^*)$ *orthogonal auf den Tangentialraum* $\mathcal{T}_{\mathbf{x}^*}\mathcal{S}$ steht. Dies bedeutet für einen regulären Punkt $\mathbf{x}^*$, dass $(\nabla f)(\mathbf{x}^*)$ sich als Linearkombination von $(\nabla g_i)(\mathbf{x}^*)$, $i = 1, \ldots, p$ darstellen lassen muss. Dies motiviert die Einführung der so genannten *Lagrange-Multiplikatoren* $\boldsymbol{\lambda}$ im folgenden Satz.

Satz 3.1 (Notwendige Optimalitätsbedingungen erster Ordnung)**.** *Es sei* $\mathbf{x}^* \in \mathcal{S}$ *mit* $\mathcal{S} = \{\mathbf{x} \in \mathbb{R}^n | g_i(\mathbf{x}) = 0,\ i = 1, \ldots, p\}$ *ein regulärer Punkt und ein lokaler Extremalpunkt von* $f(\mathbf{x})$ *unter den Gleichungsnebenbedingungen* $\mathbf{g}(\mathbf{x}) = \mathbf{0}$ *mit* $f, g_1, \ldots, g_p \in C^1$*. Dann existiert ein eindeutiges* $\boldsymbol{\lambda}^* \in \mathbb{R}^p$ *so, dass gilt*

$$(\nabla f)(\mathbf{x}^*) + (\nabla \mathbf{g})(\mathbf{x}^*)\boldsymbol{\lambda}^* = \mathbf{0}\ . \tag{3.9}$$

Die notwendigen Optimalitätsbedingungen (3.9) von Satz 3.1 gemeinsam mit den 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*

$$L(\mathbf{x}, \boldsymbol{\lambda}) = f(\mathbf{x}) + \boldsymbol{\lambda}^{\mathrm{T}}\mathbf{g}(\mathbf{x}) \tag{3.10}$$

einführen und die notwendigen Optimalitätsbedingungen lassen sich in der Form

$$\left(\frac{\partial}{\partial \mathbf{x}} L\right)^{\mathrm{T}}(\mathbf{x}^*, \boldsymbol{\lambda}^*) = (\nabla f)(\mathbf{x}^*) + (\nabla \mathbf{g})(\mathbf{x}^*)\boldsymbol{\lambda}^* = \mathbf{0} \tag{3.11a}$$

$$\left(\frac{\partial}{\partial \boldsymbol{\lambda}} L\right)^{\mathrm{T}}(\mathbf{x}^*, \boldsymbol{\lambda}^*) = \mathbf{g}(\mathbf{x}^*) = \mathbf{0} \tag{3.11b}$$

schreiben.

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

$$\min_{\mathbf{x} \in \mathbb{R}^2} \quad f(\mathbf{x}) = (x_1 - 2)^2 + (x_2 - 1)^2 \tag{3.12a}$$

$$\text{u.B.v.} \quad g(\mathbf{x}) = x_2 - 2x_1 = 0\,. \tag{3.12b}$$

Abbildung 3.1 stellt die Gerade $g(\mathbf{x}) = 0$ und die Höhenlinien von $f(\mathbf{x})$ dar.

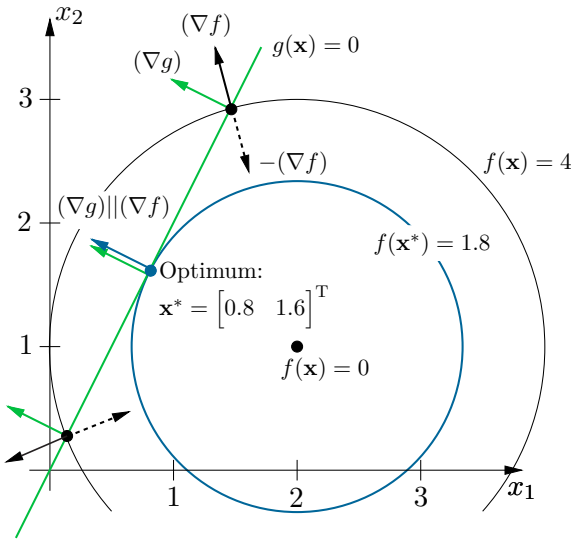


Abbildung 3.1: Veranschaulichung von Beispiel 3.1 mit einer Gleichungsbeschränkung.

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

$$\mathbf{x}^* = \begin{bmatrix}0.8 & 1.6\end{bmatrix}^{\mathrm{T}}, \quad f(\mathbf{x}^*) = 1.8 \tag{3.13}$$

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

$$(\nabla f)(\mathbf{x}) = \begin{bmatrix}2(x_1 - 2)\\ 2(x_2 - 1)\end{bmatrix}, \quad (\nabla g)(\mathbf{x}) = \begin{bmatrix}-2\\ 1\end{bmatrix} \tag{3.14}$$

und damit errechnen sich die notwendigen Optimalitätsbedingungen (3.11) 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

$$\frac{\partial}{\partial x_1} L(x_1, x_2, \lambda) = 2(x_1 - 2) - 2\lambda = 0 \tag{3.15a}$$

$$\frac{\partial}{\partial x_2} L(x_1, x_2, \lambda) = 2(x_2 - 1) + \lambda = 0 \tag{3.15b}$$

$$\frac{\partial}{\partial \lambda} L(x_1, x_2, \lambda) = x_2 - 2x_1 = 0\ . \tag{3.15c}$$

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

Aufgabe 3.1. Gegeben ist eine quaderförmige Box mit der Oberfläche A. Zeigen Sie, dass unter allen möglichen Quadern der Würfel mit der Seitenlänge $\sqrt{A/6}$ das größte Volumen besitzt.

Aufgabe 3.2. Gegeben ist ein nichtlineares zeitvariantes Abtastsystem der Form

$$x_{k+1} = \varphi_k(x_k, u_k), \quad x_0 = x(0) \tag{3.16}$$

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

$$J = \sum_{k=0}^{N} \psi_k(x_k, u_k) \tag{3.17}$$

minimiert wird und die Endbedingungen $g(x_{N+1}) = 0$ erfüllt ist. Nehmen Sie dabei an, dass sämtliche partielle Ableitungen erster Ordnung aller auftretenden Funktionen stetig sind und die LICQ Bedingung gemäß Definition 3.1 erfüllt ist. Zeigen Sie, dass mit der optimalen Lösung die Gleichungen

$$\lambda_{k-1} = \lambda_k \left(\frac{\partial}{\partial x}\varphi_k\right)(x_k, u_k) + \left(\frac{\partial}{\partial x}\psi_k\right)(x_k, u_k), \quad k = 1, \ldots, N \tag{3.18a}$$

$$\lambda_N = \mu \left(\frac{\partial}{\partial x} g\right)(x_{N+1}) \tag{3.18b}$$

$$0 = \lambda_k \left(\frac{\partial}{\partial u}\varphi_k\right)(x_k, u_k) + \left(\frac{\partial}{\partial u}\psi_k\right)(x_k, u_k), \quad k = 0, \ldots, N \tag{3.18c}$$

mit einem geeigneten Wert μ verbunden sind.

Satz 3.2 (Notwendige Optimalitätsbedingungen zweiter Ordnung)**.** *Es sei $\mathbf{x}^* \in \mathcal{S}$ mit $\mathcal{S} = \{\mathbf{x} \in \mathbb{R}^n | g_i(\mathbf{x}) = 0,\ i = 1, \ldots, p\}$ 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*

$$\left(\frac{\partial}{\partial \mathbf{x}} L\right)^{\mathrm{T}} (\mathbf{x}^*, \boldsymbol{\lambda}^*) = (\nabla f)(\mathbf{x}^*) + (\nabla \mathbf{g})(\mathbf{x}^*)\boldsymbol{\lambda}^* = \mathbf{0} \tag{3.19}$$

und

$$\mathbf{d}^{\mathrm{T}} \left(\nabla^2 L\right)(\mathbf{x}^*, \boldsymbol{\lambda}^*)\mathbf{d} = \mathbf{d}^{\mathrm{T}} \left(\left(\nabla^2 f\right)(\mathbf{x}^*) + \sum_{i=1}^{p} \lambda_i^* \left(\nabla^2 g_i\right)(\mathbf{x}^*)\right)\mathbf{d} \geq 0 \tag{3.20}$$

für alle $\mathbf{d} \in \mathcal{T}_{\mathbf{x}^*}\mathcal{S}$ mit der Lagrangefunktion L gemäß (3.10).

Die Sätze 3.1 und 3.2 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 3.3 (Hinreichende Optimalitätsbedingungen zweiter Ordnung)**.** *Gesucht ist das Minimum der Kostenfunktion $f(\mathbf{x})$ unter den Gleichungsnebenbedingungen $g_i(\mathbf{x}) =$*

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

$$\left(\frac{\partial}{\partial \mathbf{x}} L\right)^{\mathrm{T}} (\mathbf{x}^*, \boldsymbol{\lambda}^*) = \mathbf{0} \tag{3.21a}$$

$$\left(\frac{\partial}{\partial \boldsymbol{\lambda}} L\right)^{\mathrm{T}} (\mathbf{x}^*, \boldsymbol{\lambda}^*) = \mathbf{0} \tag{3.21b}$$

und

$$\mathbf{d}^{\mathrm{T}} \Big(\nabla^2 L\Big)(\mathbf{x}^*, \boldsymbol{\lambda}^*)\mathbf{d} > 0 \qquad \forall\, \mathbf{d} \in \mathcal{T}_{\mathbf{x}^*}\mathcal{S},\ \mathbf{d} \neq \mathbf{0} \tag{3.22}$$

mit der Lagrangefunktion L *gemäß* (3.10)*, dann ist* $\mathbf{x}^*$ *ein striktes lokales Minimum.*

Man erkennt aus Satz 3.3, dass die Matrix $\mathbf{L} = (\nabla^2 L)(\mathbf{x}^*, \boldsymbol{\lambda}^*)$ eine ähnliche Rolle wie die Hessematrix $(\nabla^2 f)(\mathbf{x}^*)$ der Kostenfunktion $f(\mathbf{x})$ im unbeschränkten Fall spielt, siehe die Sätze 2.2 und 2.3. In der Tat wird das *Konvergenzverhalten des beschränkten Optimierungsproblems durch die Eigenwerte der Matrix* $\mathbf{L}$ *eingeschränkt auf den Tangentialraum* $\mathcal{T}_{\mathbf{x}^*}\mathcal{S}$ *bestimmt*. Um nun die Erfüllung von (3.22) zu überprüfen, kann die Matrix $\mathbf{L}$ auf den Tangentialraum $\mathcal{T}_{\mathbf{x}^*}\mathcal{S}$ 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 $\mathcal{T}_{\mathbf{x}^*}\mathcal{S}$ bilden. Es lässt sich nun für jedes $\mathbf{d} \in \mathcal{T}_{\mathbf{x}^*}\mathcal{S}$ 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 (3.22) ein, so erhält man

$$\mathbf{d}^{\mathrm{T}} \Big(\nabla^2 L\Big)(\mathbf{x}^*, \boldsymbol{\lambda}^*)\mathbf{d} = \mathbf{d}^{\mathrm{T}}\mathbf{L}\mathbf{d} = \mathbf{z}^{\mathrm{T}}\mathbf{T}^{\mathrm{T}}\mathbf{L}\mathbf{T}\mathbf{z} > 0 \qquad \forall\, \mathbf{z} \in \mathbb{R}^{n-p},\ \mathbf{z} \neq \mathbf{0}. \tag{3.23}$$

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

$$\mathbf{L}_S = \mathbf{T}^{\mathrm{T}}\mathbf{L}\mathbf{T}, \tag{3.24}$$

und die Überprüfung der Erfüllung von (3.22) reduziert sich auf die Prüfung der positiven Definitheit von $\mathbf{L}_S$. Es sei nun λ ein Eigenwert von $\mathbf{L}_S$ und $\mathbf{v}$ der zugehörige Eigenvektor (zugleich Links- und Rechtseigenvektor). Folglich gilt

$$0 = \mathbf{v}^{\mathrm{T}}(\lambda\mathbf{E} - \mathbf{L}_S)\mathbf{v} = \mathbf{v}^{\mathrm{T}}(\lambda\mathbf{T}^{\mathrm{T}}\mathbf{T} - \mathbf{T}^{\mathrm{T}}\mathbf{L}\mathbf{T})\mathbf{v} = \mathbf{v}^{\mathrm{T}}\mathbf{T}^{\mathrm{T}}(\lambda\mathbf{E} - \mathbf{L})\mathbf{T}\mathbf{v}. \tag{3.25}$$

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

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

$$\min_{\mathbf{x} \in \mathbb{R}^3} \quad f(\mathbf{x}) = x_1 + x_2^2 + x_2 x_3 + 2x_3^2 \tag{3.26a}$$

$$\text{u.B.v.} \quad g(\mathbf{x}) = x_1^2 + x_2^2 + x_3^2 - 1 = 0\,. \tag{3.26b}$$

Die notwendigen Optimalitätsbedingungen erster Ordnung nach Satz 3.1 lauten

$$\frac{\partial}{\partial x_1} L(\mathbf{x}^*, \lambda^*) = 1 + 2\lambda^* x_1^* = 0 \tag{3.27a}$$

$$\frac{\partial}{\partial x_2} L(\mathbf{x}^*, \lambda^*) = 2x_2^* + x_3^* + 2\lambda^* x_2^* = 0 \tag{3.27b}$$

$$\frac{\partial}{\partial x_3} L(\mathbf{x}^*, \lambda^*) = x_2^* + 4x_3^* + 2\lambda^* x_3^* = 0 \tag{3.27c}$$

$$\frac{\partial}{\partial \lambda} L(\mathbf{x}^*, \lambda^*) = (x_1^*)^2 + (x_2^*)^2 + (x_3^*)^2 - 1 = 0. \tag{3.27d}$$

Man kann sich selbst einfach davon überzeugen, dass mit $x_1^* = 1$, $x_2^* = 0$, $x_3^* = 0$ und $\lambda^* = -1/2$ eine Lösung von (3.27) gegeben ist. Für die notwendige Optimalitätsbedingung zweiter Ordnung gemäß Satz 3.2 muss die Matrix

$$\mathbf{L} = \left(\nabla^2 L\right)(\mathbf{x}^*, \boldsymbol{\lambda}^*) = \begin{bmatrix} -1 & 0 & 0 \\ 0 & 1 & 1 \\ 0 & 1 & 3 \end{bmatrix} \tag{3.28}$$

berechnet werden.

Um den Tangentialraum $\mathcal{T}_{\mathbf{x}^*}\mathcal{S}$ der Mannigfaltigkeit $\mathcal{S} = \{\mathbf{x} \in \mathbb{R}^3 | x_1^2 + x_2^2 + x_3^2 = 1\}$ (Einheitskugel) gemäß (3.5) zu berechnen, bestimme man vorerst folgenden Ausdruck

$$\left(\frac{\partial}{\partial \mathbf{x}} g\right)(\mathbf{x}^*) = \begin{bmatrix} 2x_1 & 2x_2 & 2x_3 \end{bmatrix}\Big|_{\mathbf{x}=\mathbf{x}^*} = \begin{bmatrix} 2 & 0 & 0 \end{bmatrix}. \tag{3.29}$$

Man erkennt aus (3.29), dass gemäß (3.5) die erste Komponente sämtlicher Vektorfelder $\mathbf{d} \in \mathcal{T}_{\mathbf{x}^*}\mathcal{S}$ identisch Null sein muss. Damit ist aber für alle $\mathbf{d} \in \mathcal{T}_{\mathbf{x}^*}\mathcal{S}$ Beziehung (3.22) von Satz 3.3 erfüllt, da die Submatrix $\mathbf{L}_{[2..3,2..3]}$ positiv definit ist. Alternativ erhält man dieses Ergebnis durch Projektion der Matrix $\mathbf{L}$. Wählt man dazu zwei orthogonale Basisvektoren des Tangentialraumes $\mathcal{T}_{\mathbf{x}^*}\mathcal{S}$ und fasst diese in der Matrix $\mathbf{T}$ als Spaltenvektoren zusammen, z. B.

$$\mathbf{T} = \begin{bmatrix} 0 & 0 \\ 1 & 0 \\ 0 & 1 \end{bmatrix}, \tag{3.30}$$

dann kann die Matrix $\mathbf{L}$ von (3.28) gemäß (3.24) wie folgt

$$\mathbf{L}_S = \mathbf{T}^{\mathrm{T}}\mathbf{L}\mathbf{T} = \begin{bmatrix} 1 & 1 \\ 1 & 3 \end{bmatrix} \tag{3.31}$$

in den Tangentialraum projiziert werden. Da die Matrix $\mathbf{L}_S$ positiv definit ist, folgt auf Basis von Satz 3.3, dass der Punkt $\mathbf{x}^* = \begin{bmatrix} 1 & 0 & 0 \end{bmatrix}^{\mathrm{T}}$ ein striktes lokales Minimum des beschränkten Optimierungsproblems (3.26) ist.

3.1.2 Sensitivitätsbetrachtung

Angenommen, der Punkt $\mathbf{x}^*$ ist eine Lösung des beschränkten Optimierungsproblems

$$\min_{\mathbf{x}\in\mathbb{R}^n} \quad f(\mathbf{x}) \tag{3.32a}$$
$$\text{u.B.v.} \quad \mathbf{g}(\mathbf{x}) = \mathbf{0} \tag{3.32b}$$

mit dem zugehörigen Lagrange-Multiplikator $\boldsymbol{\lambda}^*$. Dann lässt sich der Lagrange-Multiplikator wie folgt interpretieren.

Satz 3.4 (Sensitivitätstheorem des Lagrange-Multiplikators)**.** *Für $f, g_1, \ldots, g_p \in C^2$ betrachte man folgende Familie beschränkter Optimierungsprobleme*

$$\min_{\mathbf{x}\in\mathbb{R}^n} \quad f(\mathbf{x}) \tag{3.33a}$$
$$\text{u.B.v.} \quad \mathbf{g}(\mathbf{x}) = \mathbf{c} \tag{3.33b}$$

mit $\mathbf{c} \in \mathbb{R}^p$. Angenommen, für $\mathbf{c} = \mathbf{0}$ sei $\mathbf{x}^*$ ein regulärer Punkt und erfülle gemeinsam mit dem Lagrange-Multiplikator $\boldsymbol{\lambda}^*$ die hinreichenden Optimalitätsbedingungen zweiter Ordnung von Satz 3.3 für ein striktes lokales Minimum. Dann existiert für jedes $\mathbf{c} \in \mathbb{R}^p$ in einer Umgebung von $\mathbf{0}$ für das Optimierungsproblem (3.33) ein lokales Minimum an einer Stelle $\mathbf{x}(\mathbf{c})$, welches stetig von $\mathbf{c}$ abhängt mit $\mathbf{x}(\mathbf{0}) = \mathbf{x}^*$. Ferner gilt die Beziehung

$$\frac{\mathrm{d}}{\mathrm{d}\mathbf{c}} f(\mathbf{x}(\mathbf{c}))\bigg|_{\mathbf{c}=\mathbf{0}} = -(\boldsymbol{\lambda}^*)^{\mathrm{T}} \,. \tag{3.34}$$

Beweisskizze: Die notwendigen Optimalitätsbedingungen erster Ordnung für das Optimierungsproblem (3.33) lauten

$$(\nabla f)(\mathbf{x}) + (\nabla \mathbf{g})(\mathbf{x})\boldsymbol{\lambda} = \mathbf{0} \tag{3.35a}$$
$$\mathbf{g}(\mathbf{x}) = \mathbf{c} \,. \tag{3.35b}$$

Berechnet man die Jacobimatrix von (3.35) an der Stelle $(\mathbf{x}^*, \boldsymbol{\lambda}^*)$, also für $\mathbf{c} = \mathbf{0}$, dann erhält man (siehe auch (3.20))

$$\begin{bmatrix} (\nabla^2 L)(\mathbf{x}^*, \boldsymbol{\lambda}^*) & (\nabla \mathbf{g})(\mathbf{x}^*) \\ (\nabla \mathbf{g})^{\mathrm{T}}(\mathbf{x}^*) & \mathbf{0} \end{bmatrix} . \tag{3.36}$$

Da nach den Voraussetzungen von Satz 3.3 die Matrix $(\nabla^2 L)(\mathbf{x}^*, \boldsymbol{\lambda}^*)$ positiv definit auf $\mathcal{T}_{\mathbf{x}^*}\mathcal{S}$ sein muss (striktes lokales Minimum an der Stelle $\mathbf{x}^*$) und die Matrix $(\nabla \mathbf{g})(\mathbf{x}^*)$ regulär ist ($\mathbf{x}^*$ ist ein regulärer Punkt), ist die Jacobimatrix (3.36) regulär auf $\mathcal{T}_{\mathbf{x}^*}\mathcal{S}$.

> *Aufgabe 3.3.* Beweisen Sie diese Behauptung.

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

Die Ableitung von (3.35b) nach $\mathbf{c}$ am Punkt $\mathbf{c} = \mathbf{0}$ liefert

$$\left.\frac{\mathrm{d}\mathbf{g}(\mathbf{x}(\mathbf{c}))}{\mathrm{d}\mathbf{c}}\right|_{\mathbf{c}=\mathbf{0}} = (\nabla \mathbf{g})^{\mathrm{T}}(\mathbf{x}^*) \left.\frac{\mathrm{d}\mathbf{x}}{\mathrm{d}\mathbf{c}}\right|_{\mathbf{c}=\mathbf{0}} = \mathbf{E}. \tag{3.37}$$

Wird die Transponierte von (3.35a) rechtsseitig mit $\frac{\mathrm{d}\mathbf{x}}{\mathrm{d}\mathbf{c}}$ multipliziert, so ergibt sich am Punkt $\mathbf{c} = \mathbf{0}$

$$(\nabla f)^{\mathrm{T}}(\mathbf{x}^*) \left.\frac{\mathrm{d}\mathbf{x}}{\mathrm{d}\mathbf{c}}\right|_{\mathbf{c}=\mathbf{0}} + (\boldsymbol{\lambda}^*)^{\mathrm{T}} \underbrace{(\nabla \mathbf{g})^{\mathrm{T}}(\mathbf{x}^*) \left.\frac{\mathrm{d}\mathbf{x}}{\mathrm{d}\mathbf{c}}\right|_{\mathbf{c}=\mathbf{0}}}_{\mathbf{E}} = \mathbf{0}. \tag{3.38}$$

Gemeinsam mit (3.37) folgt daraus unmittelbar

$$\left.\frac{\mathrm{d}}{\mathrm{d}\mathbf{c}} f(\mathbf{x}(\mathbf{c}))\right|_{\mathbf{c}=\mathbf{0}} + (\boldsymbol{\lambda}^*)^{\mathrm{T}} = \mathbf{0} \tag{3.39}$$

also (3.34). □

3.1.3 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

$$J = \{ j \in \mathbb{N} | 1 \leq j \leq q,\ h_j(\mathbf{x}) = 0 \} \tag{3.40}$$

definiert. Wieder soll

$$\mathcal{S} = \{ \mathbf{x} \in \mathbb{R}^n | g_i(\mathbf{x}) = 0,\ i = 1, \ldots, p,\ h_j(\mathbf{x}) \leq 0,\ j = 1, \ldots, q \} \tag{3.41}$$

die *zulässige Menge* genannt werden. Ferner beschreibt

$$\bar{\mathcal{S}} = \{ \mathbf{x} \in \mathcal{S} |\ h_j(\mathbf{x}) = 0,\ j \in J \} \tag{3.42}$$

die durch die Gleichungs- und aktiven Ungleichungsbeschränkungen definierte Mannigfaltigkeit.

Definition 3.2 (Regulärer Punkt bei Gleichungs- und Ungleichungsbeschränkungen, LICQ)**.** Ein zulässiger Punkt $\mathbf{x} \in \mathcal{S}$ 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 (3.40) linear unabhängig sind. D. h. die Bedingung

$$\operatorname{rang}\Big(\Big[[(\nabla g_i)(\bar{\mathbf{x}})]_{i=1,\ldots,p} \quad [(\nabla h_j)(\bar{\mathbf{x}})]_{j \in J} \Big] \Big) = p + |J|, \tag{3.43}$$

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

Satz 3.5 (Karush-Kuhn-Tucker (KKT) notwendige Optimalitätsbedingungen erster Ordnung)**.** *Angenommen,* $\mathbf{x}^*$ *sei ein lokales Minimum des Optimierungsproblems* (3.1) *bzw.* (3.2)

$$\begin{aligned} \min_{\mathbf{x}\in\mathbb{R}^n} \quad & f(\mathbf{x}) && \textit{Kostenfunktion} && (3.44a)\\ \text{u.B.v.} \quad & g_i(\mathbf{x}) = 0, \quad i = 1,\dots,p && \textit{Gleichungsbeschränkungen} && (3.44b)\\ & h_i(\mathbf{x}) \le 0, \quad i = 1,\dots,q && \textit{Ungleichungsbeschränkungen} && (3.44c) \end{aligned}$$

mit $f, g_1, \dots, g_p, h_1, \dots, h_q \in C^1$*. Im Weiteren sei* $\mathbf{x}^*$ *ein regulärer Punkt der Beschränkungen* (3.44b) *und* (3.44c)*. Dann existiert ein eindeutiger Lagrange-Multiplikator* $((\boldsymbol{\lambda}^*)^\mathrm{T}, (\boldsymbol{\mu}^*)^\mathrm{T})$ *mit* $\boldsymbol{\lambda}^* \in \mathbb{R}^p$ *und* $\boldsymbol{\mu}^* \in \mathbb{R}^q$ *so, dass die Bedingungen*

$$(\nabla f)(\mathbf{x}^*) + (\nabla \mathbf{g})(\mathbf{x}^*)\boldsymbol{\lambda}^* + (\nabla \mathbf{h})(\mathbf{x}^*)\boldsymbol{\mu}^* = \mathbf{0} \tag{3.45a}$$

$$\boldsymbol{\mu}^* \ge \mathbf{0} \tag{3.45b}$$

$$\mathbf{h}^\mathrm{T}(\mathbf{x}^*)\boldsymbol{\mu}^* = 0 \tag{3.45c}$$

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: Da $\boldsymbol{\mu}^* \ge \mathbf{0}$ und $\mathbf{h}(\mathbf{x}^*) \le \mathbf{0}$ folgt aus (3.45c), dass eine Komponente μ_j^* von $\boldsymbol{\mu}^*$ nur dann von Null verschieden sein kann, wenn die zugehörige Ungleichungsbedingung aktiv ist, d. h. $h_j(\mathbf{x}^*) = 0$. Diese so genannte *complementary slackness condition* hat zur Folge, dass $h_j(\mathbf{x}^*) < 0$ stets $\mu_j^* = 0$ und $\mu_j^* > 0$ stets $h_j(\mathbf{x}^*) = 0$ impliziert.

Da $\mathbf{x}^*$ ein lokales Minimum des beschränkten Optimierungsproblems (3.44) beschreibt, ist es auch ein lokales Minimum für jenes Optimierungsproblem, bei dem alle aktiven Ungleichungsbeschränkungen durch Gleichungsbeschränkungen ersetzt werden. Dann gibt (3.45a) mit $\mu_j^* = 0$ falls $h_j(\mathbf{x}^*) < 0$ exakt die Bedingung (3.9) des bei gleichungsbeschränkten Problemen anwendbaren Satzes 3.1 wieder.

Um zu zeigen, dass $\boldsymbol{\mu}^* \ge \mathbf{0}$ gelten muss, schreibt man (3.44c) in

$$\mathbf{h}(\mathbf{x}) \le \mathbf{c} \le \mathbf{0}$$

mit $\mathbf{c} \in \mathbb{R}^q$ um, wobei am optimalen Punkt $\mathbf{c} = \mathbf{0}$ gelten soll. Ähnlich zum Beweis von Satz 3.4 folgt, dass $\mathbf{x}(\mathbf{c})$, $\boldsymbol{\lambda}(\mathbf{c})$ und $\boldsymbol{\mu}(\mathbf{c})$ stetig differenzierbare Funktionen von $\mathbf{c}$ sind und

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

gilt. Hierbei gilt $\mu_j^* = 0 \;\forall\, j \notin J$, d. h. für alle inaktiven Ungleichungsbeschränkungen

$h_j(\mathbf{x}^*) < 0$. Die Entwicklung von f in eine Taylorreihe um den optimalen Punkt liefert

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

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

□

An dieser Stelle sei betont, dass *nicht* jedes lokale Minimum die KKT-Bedingungen (3.45) erfüllt. Dies ist nur der Fall, wenn die Beschränkungen (3.44b) und (3.44c) gewisse Voraussetzungen erfüllen, die im Englischen auch als *constraint qualification (CQ)* bezeichnet werden, vgl. [1–3]. Diese Voraussetzungen sind jedenfalls erfüllt, wenn $\mathbf{x}^*$ ein regulärer Punkt gemäß Definition 3.2 ist, d. h. wenn die LICQ Bedingung erfüllt ist. Die Erfüllung der LICQ Bedingung garantiert ferner, dass der Lagrange-Multiplikator $((\boldsymbol{\lambda}^*)^{\mathrm{T}}, (\boldsymbol{\mu}^*)^{\mathrm{T}})$ *eindeutig* ist. Es existieren auch andere CQ Bedingungen, die diese Eindeutigkeit nicht garantieren.

Ein Problem bei der in Satz 3.5 gezeigten Vorgehensweise ist, dass man *a priori* nicht weiß, welche Ungleichungsbeschränkungen aktiv sind. Eigentlich müsste man sämtliche Kombinationen aktiver und inaktiver Ungleichungsbeschränkungen überprüfen, um mögliche (lokale) Minima zu finden.

Beispiel 3.3. Man betrachte das Optimierungsproblem (3.1) mit zwei Ungleichungsbeschränkungen in der Form

$$\begin{aligned}
\min_{\mathbf{x}\in\mathbb{R}^3} \quad & f(\mathbf{x}) = \frac{1}{2}(x_1^2 + x_2^2 + x_3^2) && \text{(3.46a)}\\
\text{u.B.v.} \quad & h_1(\mathbf{x}) = x_1 + x_2 + x_3 + 3 \leq 0 && \text{(3.46b)}\\
& h_2(\mathbf{x}) = x_1 \leq 0 \ . && \text{(3.46c)}
\end{aligned}$$

Mit $(\nabla h_1)(\mathbf{x}) = \begin{bmatrix} 1 & 1 & 1 \end{bmatrix}^{\mathrm{T}}$ und $(\nabla h_2)(\mathbf{x}) = \begin{bmatrix} 1 & 0 & 0 \end{bmatrix}^{\mathrm{T}}$ erkennt man, dass jeder zulässige Punkt $\bar{\mathbf{x}}$ ein regulärer Punkt ist. Die KKT-Bedingungen (3.45) lauten in diesem Fall

$$\begin{aligned}
\underbrace{\begin{bmatrix} x_1^* \\ x_2^* \\ x_3^* \end{bmatrix}}_{(\nabla f)(\mathbf{x}^*)} + \underbrace{\begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix}}_{(\nabla h_1)(\mathbf{x}^*)} \mu_1^* + \underbrace{\begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix}}_{(\nabla h_2)(\mathbf{x}^*)} \mu_2^* = \mathbf{0} && \text{(3.47a)}\\
\mu_1^* \geq 0 && \text{(3.47b)}\\
\mu_2^* \geq 0 && \text{(3.47c)}\\
\mu_1^*(x_1^* + x_2^* + x_3^* + 3) + \mu_2^* x_1^* = 0 && \text{(3.47d)}\\
x_1^* + x_2^* + x_3^* + 3 \leq 0 && \text{(3.47e)}\\
x_1^* \leq 0. && \text{(3.47f)}
\end{aligned}$$

Nun können vier Fälle unterschieden werden.

a.) Die Ungleichungsbeschränkungen sind beide 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 aber $\mu_1^* = \mu_2^* = 0$ und $x_1^* = x_2^* = x_3^* = 0$ wäre die einzige Lösung, die aber nicht zulässig ist, da sie nicht die Ungleichungsbedingung $h_1(\mathbf{x}^*)$ erfüllt.

b.) Die Ungleichungsbeschränkung $h_1(\mathbf{x}^*)$ ist inaktiv und $h_2(\mathbf{x}^*)$ ist aktiv, d. h. $h_1(\mathbf{x}^*) = x_1^* + x_2^* + x_3^* + 3 < 0$, $h_2(\mathbf{x}^*) = x_1^* = 0$, $\mu_1^* = 0$ und $\mu_2^* > 0$. Die Lösung $x_2^* = x_3^* = 0$ ist wiederum kein zulässiger Punkt.

c.) Die Ungleichungsbeschränkung $h_1(\mathbf{x}^*)$ ist aktiv und $h_2(\mathbf{x}^*)$ ist inaktiv, d. h. $h_1(\mathbf{x}^*) = x_1^* + x_2^* + x_3^* + 3 = 0$, $h_2(\mathbf{x}^*) = x_1^* < 0$, $\mu_1^* > 0$ und $\mu_2^* = 0$. Die Lösung $x_1^* = x_2^* = x_3^* = -1$ und $\mu_1^* = 1$, $\mu_2^* = 0$ ist damit ein zulässiger Kandidat für ein (lokales) Minimum.

d.) Die Ungleichungsbeschränkungen sind beide aktiv, d. h. $h_1(\mathbf{x}^*) = x_1^* + x_2^* + x_3^* + 3 = 0$, $h_2(\mathbf{x}^*) = x_1^* = 0$, $\mu_1^* > 0$ und $\mu_2^* > 0$. Da wegen $x_1^* = 0$ die Beziehung $\mu_1^* = -\mu_2^*$ gelten muss, widerspricht dies der Forderung $\mu_1^* > 0$ und $\mu_2^* > 0$.

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

Satz 3.6 (KKT notwendige Optimalitätsbedingungen zweiter Ordnung)**.** *Angenommen,* $\mathbf{x}^*$ *sei ein lokales Minimum des Optimierungsproblems* (3.1) *bzw.* (3.2)

$$\min_{\mathbf{x}\in\mathbb{R}^n} \quad f(\mathbf{x}) \qquad \text{Kostenfunktion} \tag{3.48a}$$

$$\text{u.B.v.} \quad g_i(\mathbf{x}) = 0, \quad i = 1, \dots, p \qquad \text{Gleichungsbeschränkungen} \tag{3.48b}$$

$$h_i(\mathbf{x}) \leq 0, \quad i = 1, \dots, q \qquad \text{Ungleichungsbeschränkungen} \tag{3.48c}$$

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* (3.48b) *und* (3.48c)*. Dann existiert ein eindeutiger Lagrange-Multiplikator* $((\boldsymbol{\lambda}^*)^{\mathrm{T}}, (\boldsymbol{\mu}^*)^{\mathrm{T}})$ *mit* $\boldsymbol{\lambda}^* \in \mathbb{R}^p$ *und* $\boldsymbol{\mu}^* \in \mathbb{R}^q$ *so, dass die Bedingungen*

$$(\nabla f)(\mathbf{x}^*) + (\nabla \mathbf{g})(\mathbf{x}^*)\boldsymbol{\lambda}^* + (\nabla \mathbf{h})(\mathbf{x}^*)\boldsymbol{\mu}^* = \mathbf{0} \tag{3.49a}$$

$$\boldsymbol{\mu}^* \geq \mathbf{0} \tag{3.49b}$$

$$\mathbf{h}^{\mathrm{T}}(\mathbf{x}^*)\boldsymbol{\mu}^* = 0 \tag{3.49c}$$

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*

$$\mathbf{d}^{\mathrm{T}} \underbrace{\left(\left(\nabla^2 f\right)(\mathbf{x}^*) + \sum_{i=1}^{p} \lambda_i^* \left(\nabla^2 g_i\right)(\mathbf{x}^*) + \sum_{j \in J} \mu_j^* \left(\nabla^2 h_j\right)(\mathbf{x}^*) \right)}_{(\nabla^2 L)(\mathbf{x}^*, \boldsymbol{\lambda}^*, \boldsymbol{\mu}^*)} \mathbf{d} \geq 0 \tag{3.50}$$

für alle $\mathbf{d} \in \mathcal{T}_{\mathbf{x}^*}\bar{\mathcal{S}}$ *mit* $\bar{\mathcal{S}} = \{\mathbf{x} \in \mathbb{R}^n | g_i(\mathbf{x}) = 0, i = 1, \dots, p,\ h_j(\mathbf{x}) = 0,\ j \in J\}$ *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 aktiven Ungleichungsbeschränkungen bezeichnet, d. h. es gilt* $h_j(\mathbf{x}^*) = 0,\ j \in J$.

Satz 3.7 (KKT hinreichende Optimalitätsbedingungen zweiter Ordnung)**.** *Gesucht ist das (lokale) Minimum des Optimierungsproblems* (3.1) *bzw.* (3.2)

$$\begin{aligned}
&\min_{\mathbf{x} \in \mathbb{R}^n} && f(\mathbf{x}) && \text{Kostenfunktion} && \text{(3.51a)}\\
&\text{u.B.v.} && g_i(\mathbf{x}) = 0, \quad i = 1, \dots, p && \text{Gleichungsbeschränkungen} && \text{(3.51b)}\\
&&& h_i(\mathbf{x}) \leq 0, \quad i = 1, \dots, q && \text{Ungleichungsbeschränkungen} && \text{(3.51c)}
\end{aligned}$$

mit $f, g_1, \dots, g_p, h_1, \dots, h_q \in C^2$. *Wenn für einen regulären Punkt* $\mathbf{x}^*$ *der Beschränkungen* (3.51b) *und* (3.51c)*, Größen* $\mathbf{x}^* \in \mathbb{R}^n$, $\boldsymbol{\lambda}^* \in \mathbb{R}^p$ *und* $\boldsymbol{\mu}^* \in \mathbb{R}^q$ *so existieren, dass gilt*

$$(\nabla f)(\mathbf{x}^*) + (\nabla \mathbf{g})(\mathbf{x}^*)\boldsymbol{\lambda}^* + (\nabla \mathbf{h})(\mathbf{x}^*)\boldsymbol{\mu}^* = \mathbf{0} \tag{3.52a}$$

$$\boldsymbol{\mu}^* \geq \mathbf{0} \tag{3.52b}$$

$$\mathbf{h}^{\mathrm{T}}(\mathbf{x}^*)\boldsymbol{\mu}^* = 0 \tag{3.52c}$$

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* $(\nabla \mathbf{h})(\mathbf{x}^*) = \begin{bmatrix}(\nabla h_1)(\mathbf{x}^*) & \dots & (\nabla h_q)(\mathbf{x}^*)\end{bmatrix}$ *und*

$$\mathbf{d}^{\mathrm{T}} \underbrace{\left(\left(\nabla^2 f\right)(\mathbf{x}^*) + \sum_{i=1}^{p} \lambda_i^* \left(\nabla^2 g_i\right)(\mathbf{x}^*) + \sum_{j \in J} \mu_j^* \left(\nabla^2 h_j\right)(\mathbf{x}^*) \right)}_{(\nabla^2 L)(\mathbf{x}^*, \boldsymbol{\lambda}^*, \boldsymbol{\mu}^*)} \mathbf{d} > 0 \tag{3.53}$$

für alle $\mathbf{d} \in \mathcal{T}_{\mathbf{x}^*}\bar{\mathcal{S}}$, $\mathbf{d} \neq \mathbf{0}$ *und* $\bar{\mathcal{S}} = \{\mathbf{x} \in \mathbb{R}^n | g_i(\mathbf{x}) = 0, i = 1, \dots, p,\ h_j(\mathbf{x}) = 0,\ j \in J\}$ *und der Lagrangefunktion* $L(\mathbf{x}^*, \boldsymbol{\lambda}^*, \boldsymbol{\mu}^*) = f(\mathbf{x}^*) + (\boldsymbol{\lambda}^*)^{\mathrm{T}} \mathbf{g}(\mathbf{x}^*) + (\boldsymbol{\mu}^*)^{\mathrm{T}} \mathbf{h}(\mathbf{x}^*)$, *dann ist* $\mathbf{x}^*$ *ein striktes (lokales) Minimum. Mit* J *wird dabei wieder die Menge der Indizes der aktiven Ungleichungsbeschränkungen bezeichnet, d. h. es gilt* $h_j(\mathbf{x}^*) = 0,\ j \in J$.

Aufgabe 3.4. Zeigen Sie, dass der Punkt $x_1^* = x_2^* = x_3^* = -1$ von Beispiel 3.3 ein globales Minimum ist.

3.2 Rechnergestützte Optimierungsverfahren

Als Ausgangspunkt betrachte man wiederum das beschränkte Optimierungsproblem (3.2)

$$\min_{\mathbf{x}\in\mathbb{R}^n} \quad f(\mathbf{x}) \tag{3.54a}$$

$$\text{u.B.v.} \quad \mathbf{g}(\mathbf{x}) = \mathbf{0} \tag{3.54b}$$

$$\mathbf{h}(\mathbf{x}) \leq \mathbf{0} \tag{3.54c}$$

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 $x_1, \ldots, x_n$. Zur Lösung dieses Problems können im Rahmen der *Methode der aktiven Beschränkungen* aktive und inaktive Ungleichungsbeschränkungen unterschiedlich behandelt werden. Mit der *Gradienten-Projektionsmethode* wird während der iterativen Lösungssuche eine Fortbewegung in einem zulässigen Gebiet sichergestellt. Alternativ kann das Problem auch durch *sequentielle quadratische Programmierung* gelöst werden, wobei hier die Optimierungsaufgabe durch eine Sequenz 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. Im Folgenden werden einige ausgewählte Methoden näher erläutert.

3.2.1 Methode der aktiven Beschränkungen

Ohne Einschränkung der Allgemeinheit sei für die folgenden Betrachtungen angenommen, dass keine Gleichungsbeschränkungen vorhanden sind. Aus Satz 3.5 weiß man, dass die notwendigen Optimalitätsbedingungen für ein lokales Minimum durch

$$(\nabla f)(\mathbf{x}^*) + \sum_{i\in J} \mu_i^* (\nabla h_i)(\mathbf{x}^*) = \mathbf{0} \tag{3.55a}$$

$$h_i(\mathbf{x}^*) = 0, \quad i \in J \tag{3.55b}$$

$$h_i(\mathbf{x}^*) < 0, \quad i \notin J \tag{3.55c}$$

$$\mu_i^* \geq 0, \quad i \in J \tag{3.55d}$$

$$\mu_i^* = 0, \quad i \notin J \tag{3.55e}$$

gegeben sind. Mit J wird dabei wieder die Menge der Indizes der aktiven Ungleichungsbeschränkungen bezeichnet. Wenn man nun das beschränkte Optimierungsproblem für eine angenommene Menge der aktiven Ungleichungsbeschränkungen löst und diese Lösung sowohl die nichtaktiven Ungleichungsbeschränkungen erfüllt als auch ausschließlich nichtnegative Lagrange-Multiplikatoren μ_i^* beinhaltet, dann kann die Lösung als Kandidat für das Minimum angesehen werden.

Die Idee der Methode der aktiven Beschränkungen (Englisch: *active set methods*) beruht darauf, in jedem Iterationsschritt eine *Arbeitsmenge* festzulegen, die für den

aktuellen Iterationspunkt $\mathbf{x}_k$ die aktiven Ungleichungsbeschränkungen beinhaltet. Sind Gleichungsbeschränkungen vorhanden, können diese auf analoge Art und Weise in der Arbeitsmenge berücksichtigt werden. Der aktuelle Iterationspunkt $\mathbf{x}_k$ ist daher zulässig im Hinblick auf die Arbeitsmenge. Um zum nächsten Iterationspunkt $\mathbf{x}_{k+1}$ zu gelangen, bewegt man sich entlang der durch die Arbeitsmenge definierten Mannigfaltigkeit. Ziel ist es, so zu einem hinsichtlich der Kostenfunktion verbesserten Punkt zu gelangen. Die verschiedenen Verfahren unterscheiden sich dadurch, wie man sich entlang der Mannigfaltigkeit, die durch die Arbeitsmenge definiert ist, bewegt, wobei dies auch wesentlich das Konvergenzverhalten des Verfahrens bestimmt.

Angenommen, W bezeichne die Arbeitsmenge, also die Indexmenge der aktiven Ungleichungsbeschränkungen zu einem Iterationsschritt. Dann besteht die Aufgabe in diesem Iterationsschritt darin, für das Optimierungsproblem

$$\min_{\mathbf{x}\in\mathbb{R}^n} \quad f(\mathbf{x}) \qquad \text{Kostenfunktion} \tag{3.56a}$$

$$\text{u.B.v.} \quad h_i(\mathbf{x}) = 0, \quad i \in W \qquad \text{angenommene aktive Ungleichungsbeschr.} \tag{3.56b}$$

eine Lösung $\mathbf{x}^*_W$ zu finden, für die gilt $h_i(\mathbf{x}^*_W) < 0$ für alle $i \in \{1,\dots,q\}\backslash W$. Dazu löse man das Gleichungssystem

$$(\nabla f)(\mathbf{x}^*_W) + \sum_{i\in W} \mu^*_{i,W}(\nabla h_i)(\mathbf{x}^*_W) = \mathbf{0} \tag{3.57a}$$

$$h_i(\mathbf{x}^*_W) = 0, \quad i \in W \tag{3.57b}$$

nach $\mathbf{x}^*_W$ und $\mu^*_{i,W},\ i \in W$. Wenn nun $\mu^*_{i,W} \geq 0$ für alle $i \in W$, dann ist $\mathbf{x}^*_W$ eine mögliche lokale Lösung des beschränkten Optimierungsproblems. Existiert hingegegen ein $k \in W$, für das gilt $\mu^*_{k,W} < 0$, dann kann die Kostenfunktion weiter reduziert werden, indem die Beschränkung $h_k(\mathbf{x})$ inaktiv gesetzt wird, d. h. der Index k wird aus der Indexmenge W entfernt. Dies erkennt man unmittelbar aus dem Beweis von Satz 3.5, denn wenn man statt der aktiven Ungleichungsbeschränkung $h_k(\mathbf{x}) = 0$ die Gleichungsbeschränkung $h_k(\mathbf{x}) = c$ mit einem kleinen Wert $c < 0$ verwendet, d. h. $\mathbf{x}$ wird vom Rand in das Innere des Gebietes der Ungleichungsbeschränkung $h_k(\mathbf{x}) < 0$ bewegt, dann ändert sich für $\mu^*_{k,W} < 0$ und $\mathbf{x}(0) = \mathbf{x}^*_W$ die Kostenfunktion in der Form

$$f(\mathbf{x}(c)) \approx f(\mathbf{x}^*_W) + \underbrace{\frac{\mathrm{d}}{\mathrm{d}c} f(\mathbf{x}(c))\Big|_{c=0}}_{-\mu^*_{k,W}>0} \underbrace{c}_{<0} < f(\mathbf{x}^*_W)\ . \tag{3.58}$$

Dies zeigt also, dass durch eine Bewegung in das Innere des Gebietes der Ungleichungsbeschränkung $h_k(\mathbf{x}) < 0$ die Kostenfunktion weiter minimiert werden kann. Abbildung 3.2 veranschaulicht diesen Sachverhalt grafisch 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 als inaktiv erachtete Ungleichungsbeschränkung verletzt wird, d. h. $\exists k \in \{1,\dots,q\}\backslash W\ :\ h_k(\mathbf{x}^*_W) > 0$. In diesem Fall muss die Indexmenge W um den Index k dieser Ungleichungsbeschränkung erweitert werden.

Die Methode der aktiven Beschränkungen beruht im Wesentlichen auf den obigen Überlegungen, wobei die Konvergenz durch den nachfolgenden Satz sichergestellt werden kann.

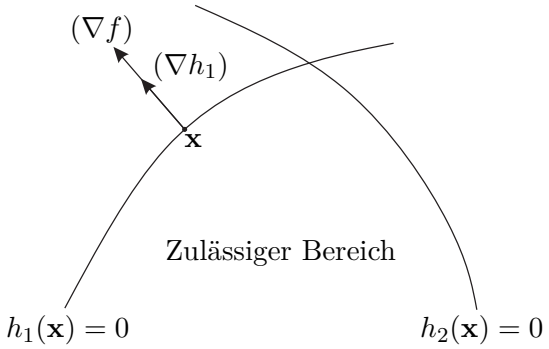


Abbildung 3.2: Zur Deaktivierung von Ungleichungsbeschränkungen.

Satz 3.8 (Konvergenz der Methode der aktiven Beschränkungen)**.** *Angenommen, für aufeinanderfolgende Arbeitsmengen W ist das Optimierungsproblem (3.56) wohldefiniert und hat eine eindeutige nichtdegenerierte Lösung (d. h. für alle $i \in W$, $\mu^*_{i,W} \neq 0$), dann konvergiert die Methode der aktiven Beschränkungen gegen die Lösung des zugrundeliegenden beschränkten Optimierungsproblems.*

Eine Schwierigkeit in der praktischen Anwendung ist, dass die Iterationslösungen exakte Lösungen des unterlagerten Minimierungsproblems sein müssen. Ansonsten können Vorzeichen der Lagrange-Multiplikatoren falsch sein und man muss verhindern, dass zwischen gleichen Arbeitsmengen ständig hin- und hergesprungen wird. Dazu gibt es eine Vielzahl unterschiedlicher Strategien, die sich mehr oder weniger von dem vorgestellten Basisalgorithmus unterscheiden.

3.2.2 Gradienten-Projektionsmethode

Die Grundidee ist, dass bei dieser Methode die Bewegungsrichtung auf der Mannigfaltigkeit, die durch die Arbeitsmenge definiert ist, als Projektion des negativen Gradienten auf diese Mannigfaltigkeit festgelegt wird. Dazu betrachte man vorerst folgendes Optimierungsproblem mit linearen Gleichungs- und Ungleichungsbeschränkungen

$$\min_{\mathbf{x}\in\mathbb{R}^n} \quad f(\mathbf{x}) \tag{3.59a}$$

$$\text{u.B.v.} \quad \mathbf{a}_i^{\mathrm{T}}\mathbf{x} = b_i, \qquad i \in J_1 \tag{3.59b}$$

$$\mathbf{a}_i^{\mathrm{T}}\mathbf{x} \leq b_i, \qquad i \in J_2\ . \tag{3.59c}$$

Zu einem Iterationsschritt sei nun angenommen, dass in Summe $q < n$ Gleichungs- und Ungleichungsbeschränkungen aktiv sind und die aktuelle Arbeitsmenge definieren. Fasst man nun alle Vektoren $\mathbf{a}_i^{\mathrm{T}}$ der Gleichungs- und aktiven Ungleichungsbeschränkungen als

Zeilenvektoren der Matrix $\mathbf{A}_q \in \mathbb{R}^{q\times n}$ zusammen, d. h.

$$\mathbf{A}_q = \begin{bmatrix} \mathbf{a}_1^{\mathrm{T}} \\ \mathbf{a}_2^{\mathrm{T}} \\ \vdots \\ \mathbf{a}_q^{\mathrm{T}} \end{bmatrix} , \tag{3.60}$$

dann weiß man zufolge von (3.5), dass der Tangentialraum der Mannigfaltigkeit definiert durch alle aktuellen aktiven Beschränkungen durch die Vektoren des Nullraums Kern($\mathbf{A}_q$) aufgespannt wird, d. h. $\mathcal{T}\bar{\mathcal{S}} = \{\mathbf{d} \in \mathbb{R}^n \,|\, \mathbf{A}_q\mathbf{d} = \mathbf{0}\}$. Da die LICQ Bedingung angenommen wurde, ist die Matrix $\mathbf{A}_q$ zeilenregulär und es gilt $\mathrm{rang}(\mathbf{A}_q) = q$, also $\dim(\mathrm{Kern}(\mathbf{A}_q)) = n - q$. Der Gradientenvektor $(\nabla f)(\mathbf{x}_k)$ im Iterationsschritt k steht nun im Allgemeinen nicht orthogonal auf $\mathcal{T}\bar{\mathcal{S}}$. Aufgrund von $\mathbb{R}^n = \mathrm{Kern}(\mathbf{A}_q) \oplus \mathrm{Bild}\left(\mathbf{A}_q^{\mathrm{T}}\right)$ mit Bild($\mathbf{A}_q$) als dem Bild von $\mathbf{A}_q$ lässt sich der negative Gradient $-(\nabla f)(\mathbf{x}_k)$ immer in der Form

$$-(\nabla f)(\mathbf{x}_k) = \mathbf{d}_k + \mathbf{A}_q^{\mathrm{T}}\boldsymbol{\sigma}_k \tag{3.61}$$

für ein geeignetes $\mathbf{d}_k \in \mathcal{T}\bar{\mathcal{S}}$ und $\boldsymbol{\sigma}_k \in \mathbb{R}^q$ anschreiben. Multipliziert man (3.61) mit $\mathbf{A}_q$ von links, so erhält man aufgrund der Zeilenregularität von $\mathbf{A}_q$ unmittelbar die Beziehung

$$\boldsymbol{\sigma}_k = -\left(\mathbf{A}_q\mathbf{A}_q^{\mathrm{T}}\right)^{-1}\mathbf{A}_q(\nabla f)(\mathbf{x}_k) \; . \tag{3.62}$$

Setzt man (3.62) in (3.61) ein, so errechnet sich $\mathbf{d}_k$ zu

$$\mathbf{d}_k = -\mathbf{P}_k(\nabla f)(\mathbf{x}_k) \quad \text{mit} \quad \mathbf{P}_k = \mathbf{E} - \mathbf{A}_q^{\mathrm{T}}\left(\mathbf{A}_q\mathbf{A}_q^{\mathrm{T}}\right)^{-1}\mathbf{A}_q \tag{3.63}$$

mit der $n\times n$ Einheitsmatrix $\mathbf{E}$. Die Matrix $\mathbf{P}_k$ wird dabei als *Projektionsmatrix* bezeichnet. Sie projiziert den negativen Gradienten $-(\nabla f)(\mathbf{x}_k)$ in den Tangentialraum $\mathcal{T}\bar{\mathcal{S}}$.

> *Aufgabe 3.5.* Zeigen Sie, dass eine Projektionsmatrix $\mathbf{P}$ die Eigenschaften $\mathbf{P}^{\mathrm{T}} = \mathbf{P}$ sowie $\mathbf{P}^2 = \mathbf{P}$ erfüllt.

> *Aufgabe 3.6.* Zeigen Sie, dass man den projizierten Gradienten $\mathbf{d}_k$ auch als Lösung des beschränkten Optimierungsproblems
>
> $$\min_{\mathbf{d}_k\in\mathbb{R}^n} \quad \|(\nabla f)(\mathbf{x}_k) + \mathbf{d}_k\|_2^2 \tag{3.64a}$$
> $$\text{u.B.v.} \quad \mathbf{A}_q\mathbf{d}_k = \mathbf{0} \tag{3.64b}$$
>
> erhalten kann.

Man erkennt aus (3.61), dass $(\nabla f)(\mathbf{x}_k)+\mathbf{d}_k$ orthogonal auf $\mathcal{T}\bar{\mathcal{S}}$ und damit $\mathbf{d}_k$ steht. Wenn $\mathbf{d}_k \neq \mathbf{0}$ ist, dann ist damit unmittelbar eine *zulässige Abstiegsrichtung* des beschränkten Optimierungsproblems gegeben, denn es gilt

$$(\nabla f)^{\mathrm{T}}(\mathbf{x}_k)\mathbf{d}_k = \left((\nabla f)^{\mathrm{T}}(\mathbf{x}_k) + \mathbf{d}_k^{\mathrm{T}} - \mathbf{d}_k^{\mathrm{T}}\right)\mathbf{d}_k = -\|\mathbf{d}_k\|_2^2 < 0 \; . \tag{3.65}$$

Da nun eine zulässige Abstiegsrichtung $\mathbf{d}_k$ festliegt, 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 (2.32) bestimmt werden. Wenn $\mathbf{d}_k = \mathbf{0}$ gilt, dann folgt aus (3.61)

$$(\nabla f)(\mathbf{x}_k) + \mathbf{A}_q^{\mathrm{T}} \boldsymbol{\sigma}_k = \mathbf{0} \ . \tag{3.66}$$

Man beachte, dass (3.66) exakt der KKT-Bedingung (3.45a) von Satz 3.5 für das Optimierungsproblem (3.59) mit $\boldsymbol{\sigma}_k$ als den zugehörigen Lagrange-Multiplikatoren entspricht. Wenn alle Einträge $\sigma_{k,j},\ j \in J_2$ nichtnegativ sind, dann erfüllt der Punkt $\mathbf{x}_k$ die notwendigen KKT-Bedingungen für ein (lokales) Minimum. Für den Fall, dass eine Komponente $\sigma_{k,j} < 0,\ j \in J_2$, kann die Kostenfunktion weiter verkleinert werden, indem die Ungleichungsbeschränkung $\mathbf{a}_j^{\mathrm{T}} \mathbf{x} \leq b_j$ inaktiv gesetzt wird. Im Weiteren bezeichne man mit $\mathbf{A}_{\bar{q}}$ die Matrix $\mathbf{A}_q$ mit der j-ten Zeile gestrichen. Die Projektionsmatrix $\mathbf{P}_k$ von (3.63) muss nun nicht jedes Mal komplett neu berechnet werden, siehe dazu folgende Aufgabe.

Aufgabe 3.7. Angenommen die zeilenreguläre Matrix $\mathbf{A}_q$ hat die Form

$$\mathbf{A}_q = \begin{bmatrix} \mathbf{a}^{\mathrm{T}} \\ \mathbf{A}_{\bar{q}} \end{bmatrix} . \tag{3.67}$$

Zeigen Sie, dass die Beziehung

$$\left(\mathbf{A}_q \mathbf{A}_q^{\mathrm{T}}\right)^{-1} = \begin{bmatrix} \theta & -\theta \mathbf{a}^{\mathrm{T}} \mathbf{A}_{\bar{q}}^{\mathrm{T}} \left(\mathbf{A}_{\bar{q}} \mathbf{A}_{\bar{q}}^{\mathrm{T}}\right)^{-1} \\ -\theta \left(\mathbf{A}_{\bar{q}} \mathbf{A}_{\bar{q}}^{\mathrm{T}}\right)^{-1} \mathbf{A}_{\bar{q}} \mathbf{a} & \left(\mathbf{A}_{\bar{q}} \mathbf{A}_{\bar{q}}^{\mathrm{T}}\right)^{-1} \left(\mathbf{E} + \theta \mathbf{A}_{\bar{q}} \mathbf{a} \mathbf{a}^{\mathrm{T}} \mathbf{A}_{\bar{q}}^{\mathrm{T}} \left(\mathbf{A}_{\bar{q}} \mathbf{A}_{\bar{q}}^{\mathrm{T}}\right)^{-1}\right) \end{bmatrix} \tag{3.68}$$

mit

$$\theta = \frac{1}{\mathbf{a}^{\mathrm{T}} \mathbf{a} - \mathbf{a}^{\mathrm{T}} \mathbf{A}_{\bar{q}}^{\mathrm{T}} \left(\mathbf{A}_{\bar{q}} \mathbf{A}_{\bar{q}}^{\mathrm{T}}\right)^{-1} \mathbf{A}_{\bar{q}} \mathbf{a}} \tag{3.69}$$

gilt. **Hinweis:** Sie können unterstützend die Ergebnisse aus Kapitel 1.3.4 des Skriptums Regelungssysteme 1 verwenden.

Der Algorithmus der Gradienten-Projektionsmethode ist in Tabelle 3.1 zusammengefasst.

Betrachtet man nun anstelle des Optimierungsproblems mit linearen Beschränkungen (3.59) das Optimierungsproblem mit nichtlinearen Beschränkungen

$$\min_{\mathbf{x} \in \mathbb{R}^n} \quad f(\mathbf{x}) \tag{3.70a}$$

$$\text{u.B.v.} \quad h_i(\mathbf{x}) = 0, \qquad i \in J_1 \tag{3.70b}$$

$$h_i(\mathbf{x}) \leq 0, \qquad i \in J_2 \ , \tag{3.70c}$$

dann ergeben sich zusätzliche Schwierigkeiten, die im Folgenden kurz erläutert werden sollen. Die Gleichungs- und aktiven Ungleichungsbeschränkungen werden wiederum in $\mathbf{h}(\mathbf{x}) = \begin{bmatrix} h_1(\mathbf{x}) & \dots & h_q(\mathbf{x}) \end{bmatrix}^{\mathrm{T}}$ zusammengefasst und beschreiben eine Mannigfaltigkeit $\bar{\mathcal{S}}$. Der negative Gradient $-(\nabla f)(\mathbf{x}_k)$ an einem Punkt $\mathbf{x}_k$ wird mit Hilfe der Projektionsmatrix

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 $\bar{\mathcal{S}}$) mit der zugehörigen Indexmenge 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 (3.63) in den Tangentialraum $\mathcal{T}\bar{\mathcal{S}}$.

Schritt 3:

if $\mathbf{d}_k \neq \mathbf{0}$ **do**

Berechne

$$\begin{aligned} \alpha_{k,1} &= \arg\max\{\alpha_k \,|\, \mathbf{x}_k + \alpha_k \mathbf{d}_k \text{ ist zulässig}\} \\ \alpha_{k,2} &= \arg \min_{0<\alpha_k<\alpha_{k,1}} f(\mathbf{x}_k + \alpha_k \mathbf{d}_k) \ , \end{aligned}$$

setze $\mathbf{x}_{k+1} = \mathbf{x}_k + \alpha_{k,2}\mathbf{d}_k$ und gehe zu Schritt 1 in der nächsten Iteration, $k \leftarrow k+1$.

else (d. h. $\mathbf{d}_k = \mathbf{0}$)

Berechne $\boldsymbol{\sigma}_k = -\left(\mathbf{A}_q \mathbf{A}_q^{\mathrm{T}}\right)^{-1} \mathbf{A}_q (\nabla f)(\mathbf{x}_k)$ (siehe (3.62))

1. Wenn $\sigma_{k,j} \geq 0$ für alle $j \in J_2$, dann erfüllt $\mathbf{x}_k$ die KKT-Bedingungen, setze stop=1.

2. Wenn nicht alle $\sigma_{k,j},\ j \in J_2$ nichtnegativ sind, dann streiche jene Ungleichungsbeschränkung, die zur negativsten Komponente $\sigma_{k,j},\ j \in J_2$ gehört, passe die Indexmenge W und die Matrix $\mathbf{A}_q$ entsprechend an und gehe zu Schritt 2 in der nächsten Iteration.

end

until stop $== 1$

Tabelle 3.1: Gradienten-Projektionsmethode.

(vergleiche (3.63))

$$\mathbf{P}_k = \mathbf{E} - (\nabla \mathbf{h})(\mathbf{x}_k)\Big((\nabla \mathbf{h})^{\mathrm{T}}(\mathbf{x}_k)(\nabla \mathbf{h})(\mathbf{x}_k)\Big)^{-1}(\nabla \mathbf{h})^{\mathrm{T}}(\mathbf{x}_k) \tag{3.71}$$

mit $(\nabla \mathbf{h})(\mathbf{x}) = \Big[(\nabla h_1)(\mathbf{x}) \quad \ldots \quad (\nabla h_q)(\mathbf{x})\Big]$ in den Tangentialraum $\mathcal{T}_{\mathbf{x}_k}\bar{\mathcal{S}}$ projiziert.

Abbildung 3.3 gibt eine grafische Veranschaulichung dieses Sachverhaltes und es ist unmittelbar ersichtlich, dass im Gegensatz zum bisher diskutierten Problem mit linearen

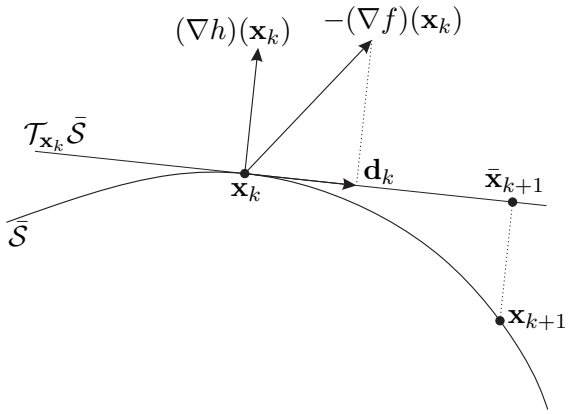


Abbildung 3.3: Gradienten-Projektionsmethode.

Beschränkungen der Punkt

$$\bar{\mathbf{x}}_{k+1} = \mathbf{x}_k + \alpha_k \mathbf{d}_k \quad \text{mit} \quad \mathbf{d}_k = -\mathbf{P}_k(\nabla f)(\mathbf{x}_k) \tag{3.72}$$

auch für hinreichend kleines α_k nicht mehr auf der Mannigfaltigkeit $\bar{\mathcal{S}}$ zu liegen kommt. Man muss deshalb eine weitere Bewegung vom Punkt $\bar{\mathbf{x}}_{k+1}$ orthogonal zu $\mathbf{d}_k$ ausführen, um auf die Mannigfaltigkeit $\bar{\mathcal{S}}$ zu gelangen. Die Idee dabei besteht darin, ein $\boldsymbol{\eta}_k \in \mathbb{R}^q$ so zu suchen, dass gilt

$$\mathbf{x}_{k+1} = \bar{\mathbf{x}}_{k+1} + (\nabla \mathbf{h})(\mathbf{x}_k)\boldsymbol{\eta}_k \quad \text{mit} \quad \mathbf{h}(\mathbf{x}_{k+1}) = \mathbf{0} \ . \tag{3.73}$$

Man beachte, dass so ein $\boldsymbol{\eta}_k$ nicht immer existieren muss, siehe beispielsweise Abbildung 3.4.

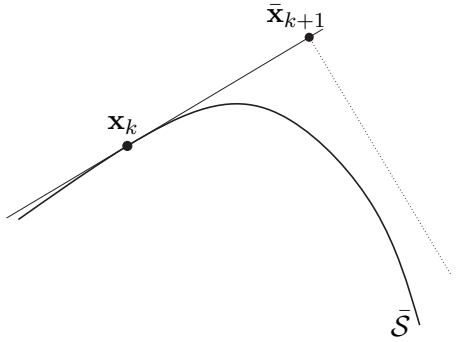


Abbildung 3.4: Nichtexistenz von $\boldsymbol{\eta}_k$, um auf $\bar{\mathcal{S}}$ zu kommen.

Eine einfache Möglichkeit, $\boldsymbol{\eta}_k$ im Sinne einer Approximation erster Ordnung zu berechnen, besteht darin, den Ausdruck $\mathbf{h}(\mathbf{x}_{k+1})$ bezüglich $\boldsymbol{\eta}_k$ zu linearisieren und nach dem linearen Glied abzubrechen, d. h.

$$\mathbf{0} = \mathbf{h}(\mathbf{x}_{k+1}) = \mathbf{h}(\bar{\mathbf{x}}_{k+1} + (\nabla \mathbf{h})(\mathbf{x}_k)\boldsymbol{\eta}_k) \approx \mathbf{h}(\bar{\mathbf{x}}_{k+1}) + (\nabla \mathbf{h})^{\mathrm{T}}(\mathbf{x}_k)(\nabla \mathbf{h})(\mathbf{x}_k)\boldsymbol{\eta}_k \ . \tag{3.74}$$

Daraus lassen sich $\boldsymbol{\eta}_k$ und $\mathbf{x}_{k+1}$ wie folgt berechnen

$$\boldsymbol{\eta}_k = -\Big[(\nabla \mathbf{h})^{\mathrm{T}}(\mathbf{x}_k)(\nabla \mathbf{h})(\mathbf{x}_k)\Big]^{-1}\mathbf{h}(\bar{\mathbf{x}}_{k+1}) \tag{3.75a}$$

$$\mathbf{x}_{k+1} = \bar{\mathbf{x}}_{k+1} - (\nabla \mathbf{h})(\mathbf{x}_k)\Big[(\nabla \mathbf{h})^{\mathrm{T}}(\mathbf{x}_k)(\nabla \mathbf{h})(\mathbf{x}_k)\Big]^{-1}\mathbf{h}(\bar{\mathbf{x}}_{k+1})\ . \tag{3.75b}$$

Gleichung (3.75) beinhaltet die gleichen Matrizen wie die Projektionsmatrix $\mathbf{P}_k$ von (3.71) und kann iterativ aufgerufen werden. Man beachte, dass für hinreichend kleines α_k in (3.72) die Iterationsvorschrift (3.75) auch tatsächlich konvergiert.

Ein weiteres Problem, das bei der Gradienten-Projektionsmethode mit nichtlinearen Beschränkungen auftreten kann, besteht darin, dass inaktive Ungleichungsbeschränkungen verletzt werden können, wenn man sich in Richtung des projizierten negativen Gradienten am Punkt $\mathbf{x}_k$ bewegt, siehe Abbildung 3.5. Typischerweise müssen in diesem Zusammenhang Verfahren zur Interpolation zwischen den Punkten $\mathbf{x}_k$ und $\bar{\mathbf{x}}_{k+1}$ eingesetzt werden, um zu einem Punkt $\bar{\bar{\mathbf{x}}}_{k+1}$ bzw. nach der Projektion $\mathbf{x}_{k+1}$ zu gelangen, der die ursprünglich inaktiven Ungleichungsbeschränkungen nicht verletzt.

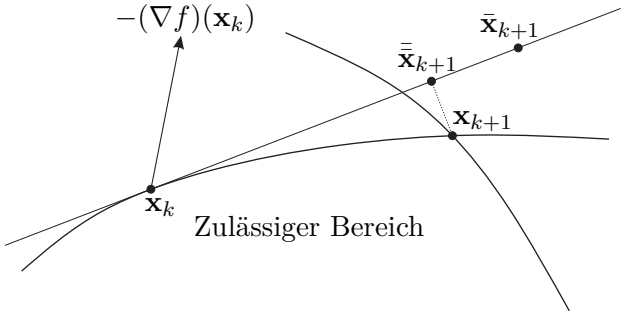


Abbildung 3.5: Interpolation zur Einhaltung der ursprünglich inaktiven Ungleichungsbeschränkungen.

Die nachfolgend beschriebenen Methoden fordern nun nicht mehr, dass die Gleichungs- und aktiven Ungleichungsbeschränkungen exakt eingehalten werden, sondern nur innerhalb eines vorgegebenen Toleranzbandes.

3.2.3 Methode der Straf- und Barrierefunktionen

Mit Hilfe von Straf- und Barrierefunktionen lassen sich beschränkte in unbeschränkte Optimierungsprobleme überführen.

3.2.3.1 Straffunktionen

Die grundlegende Idee der Methode der Straffunktionen besteht darin, das *beschränkte Optimierungsproblem*

$$\min_{\mathbf{x}\in\mathcal{X}_{\mathrm{a}\lceil}} f(\mathbf{x}) \tag{3.76}$$

mit dem zulässigen Bereich $\mathcal{X}_{\mathrm{a}\lceil}$ in ein *unbeschränktes Optimierungsproblem* der Form

$$\min_{\mathbf{x}\in\mathbb{R}^n} f(\mathbf{x}) + cP(\mathbf{x}) \tag{3.77}$$

mit einem positiven Parameter c und der stetigen Funktion $P(\mathbf{x})$ überzuführen. Die Funktion $P(\mathbf{x})$ wird als *Straffunktion* bezeichnet und besitzt die Eigenschaft, dass $P(\mathbf{x}) \geq 0$ für alle $\mathbf{x} \in \mathbb{R}^n$ und $P(\mathbf{x}) = 0$ genau dann, wenn $\mathbf{x} \in \mathcal{X}_{\mathrm{a}\lceil}$. Man beachte, dass der zulässige Bereich $\mathcal{X}_{\mathrm{a}\lceil}$ typischerweise implizit über Gleichungs- und Ungleichungsbeschränkungen definiert ist.

Für $\mathcal{X}_{\mathrm{a}\lceil} = \{\mathbf{x} \in \mathbb{R}^n \,|\, h_i(\mathbf{x}) \leq 0,\ i = 1, \ldots, q\}$ kann als Straffunktion beispielsweise

$$P(\mathbf{x}) = \frac{1}{2} \sum_{i=1}^{q} (\max\{0, h_i(\mathbf{x})\})^2 \tag{3.78}$$

verwendet werden. Abbildung 3.6 zeigt den Verlauf der Straffunktionen $cP(x)$ für $h_1(x) = x - b$ und $h_2(x) = a - x$ und unterschiedliche Werte von $c > 0$.

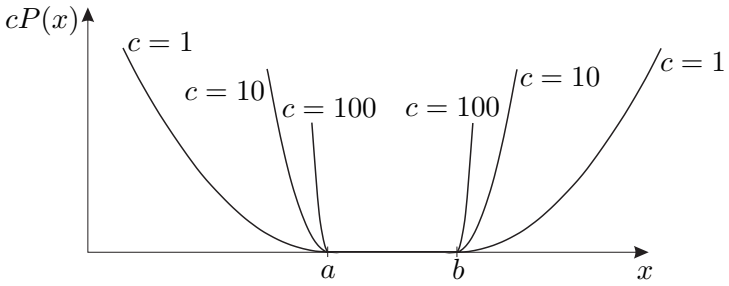


Abbildung 3.6: Verlauf der Straffunktionen $cP(x)$.

Für größere Werte von c ist zu erwarten, dass die Lösung des unbeschränkten Optimierungsproblems (3.77) zumindest in der Nähe des zulässigen Bereiches $\mathcal{X}_{\mathrm{a}\lceil}$ zu liegen kommt und für $c \to \infty$ wird die Lösung von (3.77) gegen jene von (3.76) 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 (3.77) gelöst wird und man jeweils einen optimalen Punkt $\mathbf{x}_l^*$ erhält. Für die Lösung des Optimierungsproblems (3.77) 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 3.2 (Ungleichungen bei der Methode der Straffunktionen)**.** *Für $c_{l+1} > c_l > 0$ und den zugehörigen Lösungen $\mathbf{x}_l^*$ und $\mathbf{x}_{l+1}^*$ des unbeschränkten Optimierungsproblems (3.77) gelten folgende Ungleichungen*

$$f(\mathbf{x}_l^*) + c_l P(\mathbf{x}_l^*) \leq f(\mathbf{x}_{l+1}^*) + c_{l+1} P(\mathbf{x}_{l+1}^*) \tag{3.79a}$$

$$P(\mathbf{x}_l^*) \geq P(\mathbf{x}_{l+1}^*) \tag{3.79b}$$

$$f(\mathbf{x}_l^*) \leq f(\mathbf{x}_{l+1}^*)\ . \tag{3.79c}$$

Beweis. Aufgrund von $c_{l+1} > c_l$ und der Definitionen von $\mathbf{x}_l^*$ und $\mathbf{x}_{l+1}^*$ gilt unmittelbar

$$f(\mathbf{x}_{l+1}^*) + c_{l+1} P(\mathbf{x}_{l+1}^*) \geq f(\mathbf{x}_{l+1}^*) + c_l P(\mathbf{x}_{l+1}^*) \geq f(\mathbf{x}_l^*) + c_l P(\mathbf{x}_l^*), \tag{3.80}$$

womit (3.79a) gezeigt ist. Aus

$$-f(\mathbf{x}_{l+1}^*) - c_l P(\mathbf{x}_{l+1}^*) \leq -f(\mathbf{x}_l^*) - c_l P(\mathbf{x}_l^*) \tag{3.81a}$$
$$f(\mathbf{x}_{l+1}^*) + c_{l+1} P(\mathbf{x}_{l+1}^*) \leq f(\mathbf{x}_l^*) + c_{l+1} P(\mathbf{x}_l^*) \tag{3.81b}$$

folgt

$$(c_{l+1} - c_l) P(\mathbf{x}_{l+1}^*) \leq (c_{l+1} - c_l) P(\mathbf{x}_l^*) \tag{3.82}$$

und daher (3.79b). Aus (3.80) erhält man

$$f(\mathbf{x}_{l+1}^*) + c_l \underbrace{\left(P(\mathbf{x}_{l+1}^*) - P(\mathbf{x}_l^*)\right)}_{\leq 0} \geq f(\mathbf{x}_l^*), \tag{3.83}$$

woraus sich schließlich (3.79c) ergibt. □

Lemma 3.3 (Methode der Straffunktionen). *Wenn $\mathbf{x}^*$ die Lösung des beschränkten Optimierungsproblems (3.76) ist, dann gilt für jedes l der Folge $\{c_l\}$*

$$f(\mathbf{x}^*) \geq f(\mathbf{x}_l^*) + c_l P(\mathbf{x}_l^*) \geq f(\mathbf{x}_l^*) \ . \tag{3.84}$$

Aufgabe 3.8. Beweisen Sie Lemma 3.3.

Satz 3.9 (Konvergenz der Methode der Straffunktionen). *Angenommen, $\{\mathbf{x}_l^*\}$ sei eine Folge von Punkten, die durch die Lösung des unbeschränkten Optimierungsproblems (3.77) 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 (3.76).*

3.2.3.2 Barrierefunktionen

Barrieremethoden sind auf das beschränkte Optimierungsproblem (3.76) dann anwendbar, wenn der zulässige Bereich $\mathcal{X}_{\mathrm{a}\lceil}$ eine *robuste Menge* ist, d. h. jeder Punkt am Rand von $\mathcal{X}_{\mathrm{a}\lceil}$ kann über das Innere der Menge erreicht werden, siehe Abbildung 3.7.

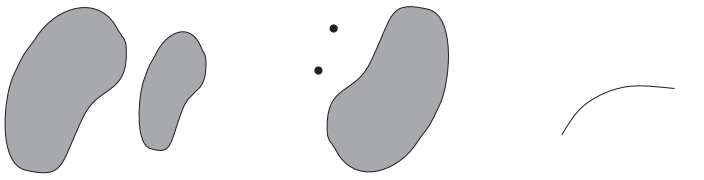


Abbildung 3.7: Robuste und nicht robuste Mengen.

Eine *Barrierefunktion* $B(\mathbf{x})$ ist im Inneren von $\mathcal{X}_{\mathrm{a}\lceil}$ definiert und hat die Eigenschaften, dass sie stetig ist, für alle $\mathbf{x} \in \mathcal{X}_{\mathrm{a}\lceil}$ nichtnegativ und $B(\mathbf{x}) \to \infty$ wenn $\mathbf{x}$ sich dem Rand

von $\mathcal{X}_{\mathrm{a}\lceil}$ nähert.

Angenommen, $\mathcal{X}_{\mathrm{a}\lceil} = \{\mathbf{x} \in \mathbb{R}^n \,|\, h_i(\mathbf{x}) \leq 0,\ i = 1, \ldots, q\}$ sei eine robuste Menge und im Inneren ist $h_i(\mathbf{x}) < 0, i = 1, \ldots, q$, dann kann als Barrierefunktion beispielsweise

$$B(\mathbf{x}) = -\sum_{i=1}^{q} \frac{1}{h_i(\mathbf{x})} \tag{3.85}$$

verwendet werden. Abbildung 3.8 zeigt den Verlauf der Barrierefunktionen $\frac{1}{c}B(x)$ für $h_1(x) = x - b$ und $h_2(x) = a - x$ und unterschiedliche Werte des Parameters $c > 0$.

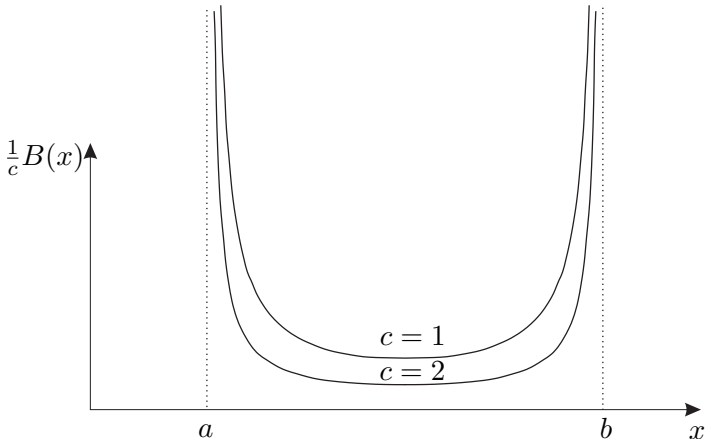


Abbildung 3.8: Verlauf der Barrierefunktionen $\frac{1}{c}B(x)$.

Eine alternative Möglichkeit, eine Barrierefunktion zu konstruieren, bietet beispielsweise die Funktion

$$B(\mathbf{x}) = -\sum_{i=1}^{q} \log(-h_i(\mathbf{x}))\ . \tag{3.86}$$

Die Vorgehensweise ist nun ähnlich zur Methode der Straffunktionen. Es wird wieder 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

$$\min_{\mathbf{x} \in \mathcal{X}_{\mathrm{a}\lceil}} f(\mathbf{x}) + \frac{1}{c_l} B(\mathbf{x}) \tag{3.87}$$

gelöst und mit $\mathbf{x}_l^*$ der jeweilige optimale Punkt bezeichnet. Mit $\mathbf{x} \in \mathcal{X}_{\mathrm{a}\lceil}$ ist der Umstand angedeutet, dass $\mathbf{x}$ auf Grund des Definitionsbereiches der Barrierefunktion stets im Inneren des zulässigen Bereiches liegen muss. Auch hier bietet es sich zur numerischen Lösung an, $\mathbf{x}_l^*$ als Startpunkt für die Optimierung mit $c_{l+1} > c_l$ zu verwenden. Es gilt nun folgender Satz.

Satz 3.10 (Konvergenz der Methode der Barrierefunktionen)**.** *Angenommen, $\{\mathbf{x}_l^*\}$ sei eine Folge von Punkten, die durch die Lösung des (unbeschränkten) Optimierungsproblems (3.87) 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 (3.76).

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

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

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

mit $f \in C^2$ und $p < n$ Gleichungsbeschränkungen $g_1(\mathbf{x}), \ldots, g_p(\mathbf{x}) \in C^2$. Nach Satz 3.5 lauten die notwendigen Optimalitätsbedingungen (KKT-Bedingungen) erster Ordnung für einen optimalen Punkt $(\mathbf{x}^*, \boldsymbol{\lambda}^*)$ mit der Lagrangefunktion $L(\mathbf{x}, \boldsymbol{\lambda}) = f(\mathbf{x}) + \boldsymbol{\lambda}^{\mathrm{T}}\mathbf{g}(\mathbf{x})$

$$\begin{bmatrix} \left(\frac{\partial}{\partial \mathbf{x}} L\right)^{\mathrm{T}}(\mathbf{x}^*, \boldsymbol{\lambda}^*) \\ \left(\frac{\partial}{\partial \boldsymbol{\lambda}} L\right)^{\mathrm{T}}(\mathbf{x}^*, \boldsymbol{\lambda}^*) \end{bmatrix} = \begin{bmatrix} (\nabla f)(\mathbf{x}^*) + (\nabla \mathbf{g})(\mathbf{x}^*)\boldsymbol{\lambda}^* \\ \mathbf{g}(\mathbf{x}^*) \end{bmatrix} = \mathbf{0}\ , \tag{3.89}$$

wobei gilt $(\nabla \mathbf{g})(\mathbf{x}^*) = \begin{bmatrix} (\nabla g_1)(\mathbf{x}^*) & \ldots & (\nabla g_p)(\mathbf{x}^*) \end{bmatrix}$. Eine Möglichkeit, das Gleichungssystem (3.89) mit den $n+p$ Gleichungen in den $n+p$ Unbekannten $(\mathbf{x}^*, \boldsymbol{\lambda}^*)$ rekursiv numerisch zu lösen, ist das Newton Verfahren mit der Iterationsvorschrift

$$\begin{bmatrix} \mathbf{x}_{k+1} \\ \boldsymbol{\lambda}_{k+1} \end{bmatrix} = \begin{bmatrix} \mathbf{x}_k \\ \boldsymbol{\lambda}_k \end{bmatrix} + \begin{bmatrix} \mathbf{p}_{\mathbf{x},k} \\ \mathbf{p}_{\boldsymbol{\lambda},k} \end{bmatrix} \tag{3.90a}$$

$$\underbrace{\begin{bmatrix} \mathbf{L}(\mathbf{x}_k, \boldsymbol{\lambda}_k) & (\nabla \mathbf{g})(\mathbf{x}_k) \\ (\nabla \mathbf{g})^{\mathrm{T}}(\mathbf{x}_k) & \mathbf{0} \end{bmatrix}}_{\mathbf{M}_k} \begin{bmatrix} \mathbf{p}_{\mathbf{x},k} \\ \mathbf{p}_{\boldsymbol{\lambda},k} \end{bmatrix} = -\begin{bmatrix} (\nabla f)(\mathbf{x}_k) + (\nabla \mathbf{g})(\mathbf{x}_k)\boldsymbol{\lambda}_k \\ \mathbf{g}(\mathbf{x}_k) \end{bmatrix} \tag{3.90b}$$

und $\mathbf{L}(\mathbf{x}_k, \boldsymbol{\lambda}_k) = \left(\frac{\partial^2}{\partial \mathbf{x}^2} L\right)(\mathbf{x}_k, \boldsymbol{\lambda}_k)$ (vgl. auch das Newton Verfahren für unbeschränkte Optimierungsprobleme gemäß Abschnitt 2.3.2.2). Die Matrix $\mathbf{M}_k$ in (3.90b) hat vollen Rang und kann damit invertiert werden, wenn sowohl die LICQ Bedingung erfüllt ist (d. h. die Matrix $(\nabla \mathbf{g})(\mathbf{x}_k)$ besitzt linear unabhängige Spaltenvektoren) als auch für alle $\mathbf{d} \neq \mathbf{0}$ mit der Eigenschaft $(\nabla \mathbf{g})^{\mathrm{T}}\mathbf{d} = \mathbf{0}$ gilt $\mathbf{d}^{\mathrm{T}}\mathbf{L}(\mathbf{x}_k, \boldsymbol{\lambda}_k)\mathbf{d} > 0$. Letztere Bedingung muss nach Satz 3.7 für ein striktes lokales Minimum $(\mathbf{x}^*, \boldsymbol{\lambda}^*)$ erfüllt sein und gilt wegen der getroffenen Differenzierbarkeitsannahmen auch für Punkte $(\mathbf{x}_k, \boldsymbol{\lambda}_k)$ in einer hinreichend kleinen Umgebung des Optimums. Wenn in einem Iterationsschritt $\mathbf{p}_{\mathbf{x},k} = \mathbf{0}$ gilt, ist aus (3.90b) ersichtlich, dass damit auch ein Punkt $\mathbf{x}^* = \mathbf{x}_k$, $\boldsymbol{\lambda}^* = \boldsymbol{\lambda}_{k+1}$ gefunden wurde, der die KKT-Bedingungen (3.89) des ursprünglichen Optimierungsproblems (3.88) erfüllt.

Die Iterationsvorschrift (3.90) kann nun auch als sukzessives *Lösen eines quadratischen Programms* der Form

$$\min_{\mathbf{p}_k \in \mathbb{R}^n} \quad f(\mathbf{x}_k) + (\nabla f)^{\mathrm{T}}(\mathbf{x}_k)\mathbf{p}_k + \frac{1}{2}\mathbf{p}_k^{\mathrm{T}}\mathbf{L}(\mathbf{x}_k, \boldsymbol{\lambda}_k)\mathbf{p}_k \tag{3.91a}$$

$$\text{u.B.v.} \quad (\nabla \mathbf{g})^{\mathrm{T}}(\mathbf{x}_k)\mathbf{p}_k + \mathbf{g}(\mathbf{x}_k) = \mathbf{0} \tag{3.91b}$$

aufgefasst werden. Die KKT-Bedingungen für (3.91) lauten

$$\begin{bmatrix} \mathbf{L}(\mathbf{x}_k, \boldsymbol{\lambda}_k) & (\nabla \mathbf{g})(\mathbf{x}_k) \\ (\nabla \mathbf{g})^{\mathrm{T}}(\mathbf{x}_k) & \mathbf{0} \end{bmatrix} \begin{bmatrix} \mathbf{p}_k^* \\ \bar{\boldsymbol{\lambda}}_k^* \end{bmatrix} = - \begin{bmatrix} (\nabla f)(\mathbf{x}_k) \\ \mathbf{g}(\mathbf{x}_k) \end{bmatrix} \tag{3.92}$$

mit den Lagrange-Multiplikatoren $\bar{\boldsymbol{\lambda}}_k$. Durch Einsetzen von $\mathbf{p}_{\boldsymbol{\lambda},k} = \boldsymbol{\lambda}_{k+1} - \boldsymbol{\lambda}_k$ in (3.90b) erhält man

$$\begin{bmatrix} \mathbf{L}(\mathbf{x}_k, \boldsymbol{\lambda}_k) & (\nabla \mathbf{g})(\mathbf{x}_k) \\ (\nabla \mathbf{g})^{\mathrm{T}}(\mathbf{x}_k) & \mathbf{0} \end{bmatrix} \begin{bmatrix} \mathbf{p}_{\mathbf{x},k} \\ \boldsymbol{\lambda}_{k+1} \end{bmatrix} = - \begin{bmatrix} (\nabla f)(\mathbf{x}_k) \\ \mathbf{g}(\mathbf{x}_k) \end{bmatrix}. \tag{3.93}$$

Ein Vergleich von (3.92) mit (3.93) bestätigt, dass statt der Lösung des Gleichungssystems (3.93) auch das Minimum des quadratischen Programms (3.91) berechnet werden kann. Die eigentliche Iterationsvorschrift lautet $\boldsymbol{\lambda}_{k+1} = \bar{\boldsymbol{\lambda}}_k^*$ und $\mathbf{p}_{\mathbf{x},k} = \mathbf{p}_k^*$ bzw. $\mathbf{x}_{k+1} = \mathbf{x}_k + \mathbf{p}_k^*$. Mit ihr wird implizit durch sukzessive Lösung der Optimierungsaufgabe (3.91) die Newton-Iteration (3.90) durchgeführt. Wenn in einem Iterationsschritt $\mathbf{p}_k^* = \mathbf{0}$ gilt, ist aus (3.92) ersichtlich, dass damit auch ein Punkt $\mathbf{x}^* = \mathbf{x}_k$, $\boldsymbol{\lambda}^* = \bar{\boldsymbol{\lambda}}_k^*$ gefunden wurde, der die KKT-Bedingungen (3.89) des ursprünglichen Optimierungsproblems (3.88) erfüllt. Da sukzessive 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

$$\min_{\mathbf{x} \in \mathbb{R}^n} \quad f(\mathbf{x}) \tag{3.94a}$$

$$\text{u.B.v.} \quad \mathbf{g}(\mathbf{x}) = \mathbf{0} \tag{3.94b}$$

$$\mathbf{h}(\mathbf{x}) \leq \mathbf{0} \tag{3.94c}$$

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 (3.94) wird in jedem Iterationsschritt durch das *quadratische Programm*

$$\min_{\mathbf{p}_k \in \mathbb{R}^n} \quad f(\mathbf{x}_k) + (\nabla f)^{\mathrm{T}}(\mathbf{x}_k)\mathbf{p}_k + \frac{1}{2}\mathbf{p}_k^{\mathrm{T}}\mathbf{L}(\mathbf{x}_k, \boldsymbol{\lambda}_k, \boldsymbol{\mu}_k)\mathbf{p}_k \tag{3.95a}$$

$$\text{u.B.v.} \quad (\nabla \mathbf{g})^{\mathrm{T}}(\mathbf{x}_k)\mathbf{p}_k + \mathbf{g}(\mathbf{x}_k) = \mathbf{0} \tag{3.95b}$$

$$(\nabla \mathbf{h})^{\mathrm{T}}(\mathbf{x}_k)\mathbf{p}_k + \mathbf{h}(\mathbf{x}_k) \leq \mathbf{0} \tag{3.95c}$$

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 (3.95) lauten (siehe Satz 3.5)

$$(\nabla f)(\mathbf{x}_k) + \mathbf{L}(\mathbf{x}_k, \boldsymbol{\lambda}_k, \boldsymbol{\mu}_k)\mathbf{p}_k^* + (\nabla \mathbf{g})(\mathbf{x}_k)\bar{\boldsymbol{\lambda}}_k^* + (\nabla \mathbf{h})(\mathbf{x}_k)\bar{\boldsymbol{\mu}}_k^* = \mathbf{0} \tag{3.96a}$$

$$\bar{\boldsymbol{\mu}}_k^* \geq \mathbf{0} \tag{3.96b}$$

$$\left((\nabla \mathbf{h})^{\mathrm{T}}(\mathbf{x}_k)\mathbf{p}_k^* + \mathbf{h}(\mathbf{x}_k)\right)^{\mathrm{T}} \bar{\boldsymbol{\mu}}_k^* = 0 \tag{3.96c}$$

$$(\nabla \mathbf{g})^{\mathrm{T}}(\mathbf{x}_k)\mathbf{p}_k^* + \mathbf{g}(\mathbf{x}_k) = \mathbf{0} \tag{3.96d}$$

$$(\nabla \mathbf{h})^{\mathrm{T}}(\mathbf{x}_k)\mathbf{p}_k^* + \mathbf{h}(\mathbf{x}_k) \leq \mathbf{0} \tag{3.96e}$$

mit den Lagrange-Multiplikatoren $\bar{\boldsymbol{\lambda}}_k$ und $\bar{\boldsymbol{\mu}}_k$. Die Iterationsvorschrift des SQP-Verfahrens lautet analog zum gleichungsbeschränkten Fall $\mathbf{x}_{k+1} = \mathbf{x}_k + \mathbf{p}_k^*$, $\boldsymbol{\lambda}_{k+1} = \bar{\boldsymbol{\lambda}}_k^*$ und $\boldsymbol{\mu}_{k+1} = \bar{\boldsymbol{\mu}}_k^*$. Gilt in einem Iterationsschritt $\mathbf{p}_k^* = \mathbf{0}$, kann man sich wieder leicht anhand von (3.96) überzeugen, dass damit ein Punkt $\mathbf{x}^* = \mathbf{x}_k$, $\boldsymbol{\lambda}^* = \bar{\boldsymbol{\lambda}}_k^*$, $\boldsymbol{\mu}^* = \bar{\boldsymbol{\mu}}_k^*$ gefunden wurde, der die KKT-Bedingungen des ursprünglichen Optimierungsproblems (3.94) erfüllt.

Unter bestimmten Voraussetzungen kann eine quadratische Konvergenzordnung (vergleiche Satz 2.9) des SQP-Verfahrens gegen $(\mathbf{x}^*, \boldsymbol{\lambda}^*, \boldsymbol{\mu}^*)$ gezeigt werden. Allerdings gilt diese Aussage im Allgemeinen nur für Startwerte $(\mathbf{x}_0, \boldsymbol{\lambda}_0, \boldsymbol{\mu}_0)$, die in einer hinreichend kleinen Umgebung um $(\mathbf{x}^*, \boldsymbol{\lambda}^*, \boldsymbol{\mu}^*)$ liegen. Man spricht deshalb auch vom *lokalen SQP-Verfahren*, welches in Tabelle 3.2 formuliert 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 $\mathbf{p}_k^*$, $\bar{\boldsymbol{\lambda}}_k^*$, $\bar{\boldsymbol{\mu}}_k^*$ durch Lösen des Optimierungsproblems (3.95).	
	Schritt 3: Setze $\mathbf{x}_{k+1} \leftarrow \mathbf{x}_k + \mathbf{p}_k^*, \boldsymbol{\lambda}_{k+1} \leftarrow \bar{\boldsymbol{\lambda}}_k^*, \boldsymbol{\mu}_{k+1} \leftarrow \bar{\boldsymbol{\mu}}_k^*, k \leftarrow k+1$.	
until	$\|\mathbf{p}_k^*\| \leq \varepsilon$	

Tabelle 3.2: Lokaler SQP-Algorithmus.

Das quadratische Programm (3.95) kann beispielsweise über die Methode der aktiven Beschränkungen (siehe Abschnitt 3.2.1) gelöst werden. Zur Veranschaulichung betrachte man das nachfolgende Beispiel.

Beispiel 3.4. Gegeben ist das quadratische Optimierungsproblem

$$\min_{\mathbf{p}_k \in \mathbb{R}^2} \quad \frac{1}{2}\mathbf{p}_k^{\mathrm{T}}\mathbf{H}\mathbf{p}_k + \mathbf{c}^{\mathrm{T}}\mathbf{p}_k \tag{3.97a}$$

$$\text{u.B.v.} \quad \mathbf{a}_1^{\mathrm{T}}\mathbf{p}_k - b_1 = -p_{k,1} + 2p_{k,2} - 2 \leq 0 \qquad \text{Ungleichungsbeschr. U1} \tag{3.97b}$$

$$\mathbf{a}_2^{\mathrm{T}}\mathbf{p}_k - b_2 = p_{k,1} + 2p_{k,2} - 6 \leq 0 \qquad \text{Ungleichungsbeschr. U2} \tag{3.97c}$$

$$\mathbf{a}_3^{\mathrm{T}}\mathbf{p}_k - b_3 = p_{k,1} - 2p_{k,2} - 2 \leq 0 \qquad \text{Ungleichungsbeschr. U3} \tag{3.97d}$$

$$\mathbf{a}_4^{\mathrm{T}}\mathbf{p}_k - b_4 = -p_{k,1} \leq 0 \qquad \text{Ungleichungsbeschr. U4} \tag{3.97e}$$

$$\mathbf{a}_5^{\mathrm{T}}\mathbf{p}_k - b_5 = -p_{k,2} \leq 0 \qquad \text{Ungleichungsbeschr. U5} \tag{3.97f}$$

mit $\mathbf{H} = \begin{bmatrix} 2 & 0 \\ 0 & 2 \end{bmatrix}$ und $\mathbf{c}^{\mathrm{T}} = \begin{bmatrix} -2 & -5 \end{bmatrix}$. Als Startpunkt für die Methode der aktiven Beschränkungen wählt man den zulässigen Punkt $\mathbf{p}_{k,0} = \begin{bmatrix} 2 & 0 \end{bmatrix}^{\mathrm{T}}$, an dem die Ungleichungsbeschränkungen U3 und U5 aktiv sind. Die Indexmenge W_0 der aktiven Ungleichungsbeschränkungen lautet damit $W_0 = \{3, 5\}$. Der nächste Iterationspunkt $\mathbf{p}_{k,j+1}$ wird in der Form $\mathbf{p}_{k,j+1} = \mathbf{p}_{k,j} + \mathbf{s}_j^*$ angesetzt. Der optimale Schritt $\mathbf{s}_j^*$ folgt aus

$$\min_{\mathbf{s}_j \in \mathbb{R}^2} \quad \frac{1}{2}(\mathbf{p}_{k,j} + \mathbf{s}_j)^{\mathrm{T}}\mathbf{H}(\mathbf{p}_{k,j} + \mathbf{s}_j) + \mathbf{c}^{\mathrm{T}}(\mathbf{p}_{k,j} + \mathbf{s}_j) \tag{3.98a}$$

$$\text{u.B.v.} \quad \mathbf{a}_w^{\mathrm{T}}(\mathbf{p}_{k,j} + \mathbf{s}_j) - b_w = \mathbf{a}_w^{\mathrm{T}}\mathbf{s}_j = 0, \ \forall w \in W_j. \tag{3.98b}$$

Die KKT-Bedingungen für (3.98) lauten mit der Matrix $\mathbf{A}_j$, deren Spalten durch $\mathbf{a}_w$, $w \in W_j$, gegeben sind, und den Lagrange-Multiplikatoren $\boldsymbol{\nu}_j$

$$\mathbf{H}\mathbf{s}_j^* + \mathbf{A}_j\boldsymbol{\nu}_j^* = -\mathbf{H}\mathbf{p}_{k,j} - \mathbf{c} \tag{3.99a}$$

$$\mathbf{a}_w^{\mathrm{T}}\mathbf{s}_j^* = 0, \quad \forall w \in W_j. \tag{3.99b}$$

Für $j = 0$ erhält man als Lösung von (3.99)

$$\begin{bmatrix} 2 & 0 & 1 & 0 \\ 0 & 2 & -2 & -1 \\ 1 & -2 & 0 & 0 \\ 0 & -1 & 0 & 0 \end{bmatrix} \begin{bmatrix} \mathbf{s}_0^* \\ \boldsymbol{\nu}_0^* \end{bmatrix} = \begin{bmatrix} -2 \\ 5 \\ 0 \\ 0 \end{bmatrix} \tag{3.100}$$

die Größen $\mathbf{s}_0^* = \mathbf{0}$ und $\boldsymbol{\nu}_0^* = \begin{bmatrix} -2 & -1 \end{bmatrix}^{\mathrm{T}}$ und somit $\mathbf{p}_{k,1} = \mathbf{p}_{k,0} = \begin{bmatrix} 2 & 0 \end{bmatrix}^{\mathrm{T}}$. Nun wird die Ungleichung U3 (Lagrange-Multiplikator mit negativstem Wert) inaktiv gesetzt

($W_1 = \{5\}$) und (3.99) erneut mit

$$\begin{bmatrix} 2 & 0 & 0 \\ 0 & 2 & -1 \\ 0 & -1 & 0 \end{bmatrix} \begin{bmatrix} \mathbf{s}_1^* \\ \nu_1^* \end{bmatrix} = \begin{bmatrix} -2 \\ 5 \\ 0 \end{bmatrix} \tag{3.101}$$

zu $\mathbf{s}_1^* = \begin{bmatrix} -1 & 0 \end{bmatrix}^{\mathrm{T}}$ und $\nu_1^* = -5$ gelöst. Damit kann der neue Iterationspunkt $\mathbf{p}_{k,2}$ zu $\mathbf{p}_{k,2} = \mathbf{p}_{k,1} + \mathbf{s}_1^*$ berechnet werden, vorausgesetzt es werden keine inaktiven Ungleichungsbeschränkungen verletzt. Um diesen Fall zu berücksichtigen, wird eine Schrittweite $\beta_j > 0$ in der Form $\mathbf{p}_{k,j+1} = \mathbf{p}_{k,j} + \beta_j \mathbf{s}_j^*$ verwendet. Die Wahl der Schrittweite β_j erfolgt nun auf Basis folgender Überlegungen. Wenn für alle inaktiven (affinen) Ungleichungsbeschränkungen gilt $\mathbf{a}_i^{\mathrm{T}} \mathbf{s}_j^* \leq 0$, dann kann $\beta_j > 0$ beliebig gewählt werden ohne eine Ungleichung $\mathbf{a}_i^{\mathrm{T}}(\mathbf{p}_{k,j} + \beta_j \mathbf{s}_j^*) \leq b_i,\ i \notin W_j$ zu verletzen. Falls hingegen $\mathbf{a}_i^{\mathrm{T}} \mathbf{s}_j^* > 0$, dann ist die Schrittweite durch $\beta_j \leq \frac{b_i - \mathbf{a}_i^{\mathrm{T}} \mathbf{p}_{k,j}}{\mathbf{a}_i^{\mathrm{T}} \mathbf{s}_j^*}$ begrenzt. Da andererseits das Minimum für $\beta_j = 1$ erreicht wird, folgt die Wahl der Schrittweite zu

$$\beta_j = \min\left\{1, \min_{i \notin W_j, \mathbf{a}_i^{\mathrm{T}} \mathbf{s}_j^* > 0} \frac{b_i - \mathbf{a}_i^{\mathrm{T}} \mathbf{p}_{k,j}}{\mathbf{a}_i^{\mathrm{T}} \mathbf{s}_j^*}\right\}. \tag{3.102}$$

Im vorliegenden Fall gilt $\beta_1 = \min\{1, \underbrace{4}_{\text{U1}}, \underbrace{2}_{\text{U4}}\} = 1$ und damit $\mathbf{p}_{k,2} = \begin{bmatrix} 1 & 0 \end{bmatrix}^{\mathrm{T}}$. Da die optimale Schrittweite $\beta_1 = 1$ möglich ist, muss (3.99) nicht erneut gelöst werden. Es kann direkt die Ungleichung U5 (Lagrange-Multiplikator ist negativ) inaktiv gesetzt ($W_2 = \{\ \}$) und das unbeschränkte Optimierungsproblem zu $\mathbf{s}_2^* = \begin{bmatrix} 0 & 2.5 \end{bmatrix}^{\mathrm{T}}$ gelöst werden. Die maximale Schrittweite gemäß (3.102) errechnet sich zu $\beta_2 = \min\{1, \underbrace{0.6}_{\text{U1}}, \underbrace{1}_{\text{U2}}\} = 0.6$ und damit folgt $\mathbf{p}_{k,3} = \begin{bmatrix} 1 & 1.5 \end{bmatrix}^{\mathrm{T}}$. Man erkennt nun, dass die Ungleichung U1 aktiv ist (der Wert $\beta_2 = 0.6$ wurde gerade durch die Ungleichungsbeschränkung U1 bestimmt), weshalb die Indexmenge der aktiven Beschränkungen zu $W_3 = \{1\}$ gesetzt wird. Aus (3.99) folgt mit

$$\begin{bmatrix} 2 & 0 & -1 \\ 0 & 2 & 2 \\ -1 & 2 & 0 \end{bmatrix} \begin{bmatrix} \mathbf{s}_3^* \\ \nu_3^* \end{bmatrix} = \begin{bmatrix} 0 \\ 2 \\ 0 \end{bmatrix} \tag{3.103}$$

die Lösung $\mathbf{s}_3^* = \begin{bmatrix} 0.4 & 0.2 \end{bmatrix}^{\mathrm{T}}$ und $\nu_3^* = 0.8$. Die Schrittweite β_3 folgt aus der Beziehung (3.102) zu $\beta_3 = \min\{1, \underbrace{2.5}_{\text{U2}}\} = 1$. Daher und auf Grund des positiven Lagrange-Multiplikators $\nu_3^* = 0.8$ stellt $\mathbf{p}_k^* = \mathbf{p}_{k,3} + \beta_3 \mathbf{s}_3^* = \begin{bmatrix} 1.4 & 1.7 \end{bmatrix}^{\mathrm{T}}$ die optimale Lösung von

(3.97) dar. Der Vektor der Lagrange-Multiplikatoren für das quadratische Programm (3.97) lautet $\bar{\boldsymbol{\mu}}_k^* = \begin{bmatrix} \nu_3^* & 0 & 0 & 0 & 0 \end{bmatrix}^{\mathrm{T}}$.

Für die Formulierung des quadratischen Programms (3.95) wird in jedem Iterationsschritt die Hessematrix der Lagrangefunktion $\mathbf{L}(\mathbf{x}_k, \boldsymbol{\lambda}_k, \boldsymbol{\mu}_k)$ benötigt. Dies birgt folgende Probleme. Zum einen ist die exakte Hessematrix in vielen Anwendungen nicht bekannt. Eine numerische Approximation mit finiten Differenzen ist sehr aufwändig und ungenau und kommt daher meist auch nicht infrage. Zum anderen kann die Hessematrix indefinit sein, was insbesondere dann auftritt, wenn das Verfahren nicht in einer hinreichend kleinen Umgebung um $(\mathbf{x}^*, \boldsymbol{\lambda}^*, \boldsymbol{\mu}^*)$ gestartet wird. Eine indefinite Hessematrix erschwert die Lösung des quadratischen Programms erheblich. Aus diesen Gründen ersetzt man in der Praxis die Hessematrix $\mathbf{L}(\mathbf{x}_k, \boldsymbol{\lambda}_k, \boldsymbol{\mu}_k)$ durch eine geeignete *positiv definite Approximation* $\mathbf{H}_k$. Für die Berechnung von $\mathbf{H}_k$ in jedem Iterationsschritt kann in Analogie zur Quasi-Newton-Methode die *modifizierte BFGS Methode* (siehe z. B. [4])

$$\mathbf{H}_{k+1} = \mathbf{H}_k + \frac{\mathbf{q}_k \mathbf{q}_k^{\mathrm{T}}}{\mathbf{q}_k^{\mathrm{T}} \mathbf{d}_k} - \frac{\mathbf{H}_k \mathbf{d}_k \mathbf{d}_k^{\mathrm{T}} \mathbf{H}_k}{\mathbf{d}_k^{\mathrm{T}} \mathbf{H}_k \mathbf{d}_k} \tag{3.104a}$$

mit

$$\mathbf{d}_k = \mathbf{x}_{k+1} - \mathbf{x}_k \tag{3.104b}$$

$$\mathbf{y}_k = \left(\frac{\partial}{\partial \mathbf{x}} L\right)^{\mathrm{T}} (\mathbf{x}_{k+1}, \boldsymbol{\lambda}_k, \boldsymbol{\mu}_k) - \left(\frac{\partial}{\partial \mathbf{x}} L\right)^{\mathrm{T}} (\mathbf{x}_k, \boldsymbol{\lambda}_k, \boldsymbol{\mu}_k) \tag{3.104c}$$

$$\theta_k = \begin{cases} 1, & \text{wenn } \mathbf{d}_k^{\mathrm{T}} \mathbf{y}_k \geq 0.2 \mathbf{d}_k^{\mathrm{T}} \mathbf{H}_k \mathbf{d}_k \\ \frac{0.8 \mathbf{d}_k^{\mathrm{T}} \mathbf{H}_k \mathbf{d}_k}{\mathbf{d}_k^{\mathrm{T}} \mathbf{H}_k \mathbf{d}_k - \mathbf{d}_k^{\mathrm{T}} \mathbf{y}_k} & \text{sonst} \end{cases} \tag{3.104d}$$

$$\mathbf{q}_k = \theta_k \mathbf{y}_k + (1 - \theta_k) \mathbf{H}_k \mathbf{d}_k \tag{3.104e}$$

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 (3.104) ist garantiert, dass $\mathbf{H}_{k+1}$ symmetrisch und positiv definit bleibt, wenn $\mathbf{H}_k$ symmetrisch und positiv definit war. Damit kann der lokale SQP-Algorithmus gemäß Tabelle 3.2 dahingehend modifiziert werden, dass ausgehend von einer symmetrischen, positiv definiten Matrix $\mathbf{H}_0$ in Schritt 1 $\mathbf{H}_k$ statt $\mathbf{L}(\mathbf{x}_k, \boldsymbol{\lambda}_k, \boldsymbol{\mu}_k)$ berechnet wird. Dabei verliert man allerdings die quadratische Konvergenzordnung des Verfahrens. Unter Verwendung von (3.104) für die Berechnung von $\mathbf{H}_{k+1}$ kann aber noch superlineare Konvergenz in einer hinreichend kleinen Umgebung um $(\mathbf{x}^*, \boldsymbol{\lambda}^*, \boldsymbol{\mu}^*)$ nachgewiesen werden.

3.2.4.2 Globalisierung des SQP-Verfahrens

Im Allgemeinen konvergiert der SQP-Algorithmus gemäß Tabelle 3.2 nur für Startwerte $(\mathbf{x}_0, \boldsymbol{\lambda}_0, \boldsymbol{\mu}_0)$, die in einer hinreichend kleinen Umgebung um $(\mathbf{x}^*, \boldsymbol{\lambda}^*, \boldsymbol{\mu}^*)$ liegen. Um zu erreichen, dass das SQP-Verfahren (idealerweise) für beliebige Startwerte konvergiert, führt man eine Globalisierung des Verfahrens durch. In Analogie zur Newton-Methode wird dies durch die Einführung einer Schrittweite $\alpha_k > 0$ erzielt. In Schritt 3 des Algorithmus

berechnet man $\mathbf{x}_{k+1}$ daher zu

$$\mathbf{x}_{k+1} = \mathbf{x}_k + \alpha_k \mathbf{p}_k^*. \tag{3.105}$$

Die Schrittweite α_k folgt aus einem geeignet formulierten Liniensuchproblem. Die Kostenfunktion dieses Liniensuchproblems muss eine Bewertung ermöglichen, ob $\mathbf{x}_{k+1}$ „besser“ als $\mathbf{x}_k$ ist. Während diese Bewertung im unbeschränkten Fall direkt anhand der Kostenfunktionswerte an den Punkten $\mathbf{x}_k$ und $\mathbf{x}_{k+1}$ möglich ist, trifft dies für ein beschränktes Optimierungsproblem im Allgemeinen nicht mehr zu. Insbesondere gilt beim SQP-Verfahren im Allgemeinen nicht mehr strikt $\mathbf{x}_k \in \mathcal{X}_{\mathrm{a}\lceil}$. Damit kann $\mathbf{x}_{k+1}$ zwar den Kostenfunktionswert verbessern aber möglicherweise zu einer stärkeren Verletzung der Beschränkungen führen. Um für die Wahl von α_k einen guten Kompromiss zu finden, verwendet man daher eine so genannte *Bewertungsfunktion* (Englisch: *merit function*). Eine beliebte Wahl ist die l_1-*Bewertungsfunktion* in der Form

$$l_1(\mathbf{x}, \eta) = f(\mathbf{x}) + \eta \left(\sum_{i=1}^{p} |g_i(\mathbf{x})| + \sum_{i=1}^{q} \max\{0, h_i(\mathbf{x})\} \right) \tag{3.106}$$

mit $\eta > 0$. Damit folgt die optimale Schrittweite α_k aus dem Liniensuchproblem

$$\alpha_k = \arg\min_{\alpha} \; l_1(\mathbf{x}_k + \alpha \mathbf{p}_k^*, \eta). \tag{3.107}$$

In der Praxis wird α_k so gewählt, dass mit $l_1(\mathbf{x}_k + \alpha_k \mathbf{p}_k^*, \eta)$ eine hinreichende Verbesserung gegenüber $l_1(\mathbf{x}_k, \eta)$ erreicht wird. Dies kann beispielsweise mit einem der Verfahren zur Schrittweitenwahl aus Abschnitt 2.3.1 erfolgen.

In seltenen Fällen besitzt die l_1-Bewertungsfunktion gemäß (3.106) die gute Eigenschaft, dass ein lokales Minimum $\mathbf{x}^*$ von (3.94) auch ein lokales Minimum von $l_1(\mathbf{x}, \eta)$ ist. Man spricht dann auch von einer *exakten Bewertungsfunktion*. Üblicherweise ist eine Anpassung von η in jedem Iterationsschritt des SQP-Verfahrens erforderlich (vgl. [5]).

3.3 Beispiel: Rosenbrock's „Bananenfunktion“

Es wird das beschränkte Optimierungsproblem (vgl. (2.111))

$$\min_{\mathbf{x} \in \mathbb{R}^2} \quad f(\mathbf{x}) = 100(x_2 - x_1^2)^2 + (x_1 - 1)^2 \tag{3.108a}$$

$$\text{u.B.v.} \quad x_1^2 + x_2^2 \geq 0.5^2 \tag{3.108b}$$

betrachtet. Die Kostenfunktion („Bananenfunktion“) ist gemeinsam mit dem Rand des zulässigen Bereiches $\mathcal{X}_{\mathrm{a}\lceil}$ in Abbildung 3.9 dargestellt.

Zur Lösung von beschränkten Optimierungsproblemen bietet sich in MATLAB der Befehl `fmincon` an. In diesem Befehl sind die folgenden vier Algorithmen implementiert:

1. `interior-point`: Verwendet logarithmische Barrierefunktionen.
2. `active-set`: Verwendet die sequentielle quadratische Programmierung mit unterlagerter Lösung des quadratischen Programms nach der Methode der aktiven Beschränkungen.

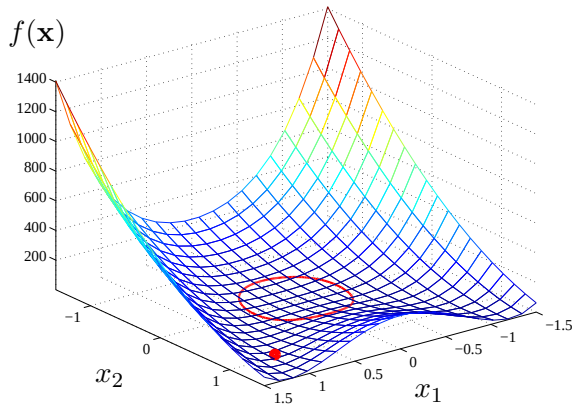


Abbildung 3.9: Profil der Rosenbrock Bananenfunktion und Rand des zulässigen Bereiches.

3. `sqp`: Ähnlich `active-set`, unterscheidet sich aber in den unterlagerten Programmroutinen sowie den Eigenschaften der Iteration zum Minimum.

4. `trust region reflective`: Methode der Vertrauensbereiche, erweitert auf Optimierungsprobleme mit entweder Beschränkungen der Form $\mathbf{A}\mathbf{x} = \mathbf{b}$ *oder* Beschränkungen der Form $\mathbf{l} \leq \mathbf{x} \leq \mathbf{u}$, wobei $\mathbf{l}$ bzw. $\mathbf{u}$ untere bzw. obere Schranken von $\mathbf{x}$ bezeichnen.

Die Lösung des Optimierungsproblems (3.108) ist damit nur mit den Methoden `active-set`, `interior-point` und `sqp` möglich, weil `trust region reflective` keine nichtlinearen Ungleichungsbeschränkungen verarbeiten kann. Die Ergebnisse der drei angesprochenen Algorithmen sind in Abbildung 3.10 dargestellt. Die Algorithmen `active-set` und `sqp` finden das globale Minimum, `interior-point` konvergiert zu einem anderen lokalen Minimum. Die gewählten Einstellungen der einzelnen Algorithmen sind in der MATLAB-Implementierung in Code-Auflistung 3.1 ersichtlich.

Listing 3.1: MATLAB-Code für die beschränkte Optimierung der Rosenbrock'schen Bananenfunktion.

```
function [Xopt,fopt,exitflag,output] = rosenbrock_problem_constrained(Xinit,testCase)
% ----------------------------------------------------------
% Xinit:  Startpunkt
% testCase: 1 - Active-Set
%           2 - SQP
%           3 - Interior Point

old = [Xinit; rosenbrock(Xinit)];
opt = optimoptions('fmincon','Display','iter','PlotFcns',@plot_iterates); % Optionen für die Ausgabe

switch testCase
 case 1,     % Active-Set mit SQP
  opt = optimoptions(opt,'Algorithm','active-set','GradObj','on','MaxFunEvals',1000,'TolFun',1e-12);
  [Xopt,fopt,exitflag,output] = fmincon(@rosenbrock,Xinit,[],[],[],[],[],[],@nonlconstr1,opt);
 case 2,     % SQP
  opt = optimoptions(opt,'Algorithm','sqp','GradObj','on','MaxFunEvals',2000,'TolX',1e-18);
  [Xopt,fopt,exitflag,output] = fmincon(@rosenbrock,Xinit,[],[],[],[],[],[],@nonlconstr1,opt);
 case 3,     % Interior-Point
  opt = optimoptions(opt,'Algorithm','interior-point','GradObj','on','TolFun',1e-12);
```

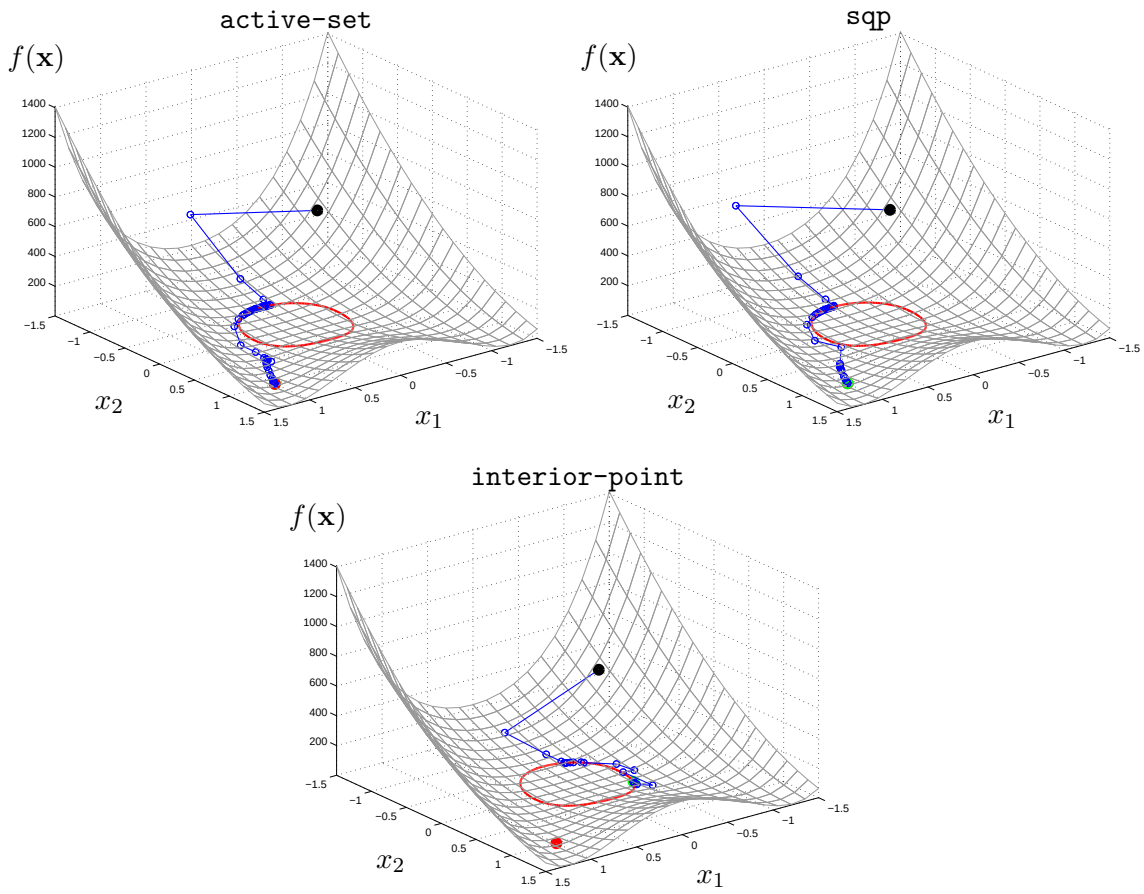


Abbildung 3.10: Rosenbrock Bananenfunktion: Vergleich der numerischen Verfahren aus `fmincon`.

```
  [Xopt,fopt,exitflag,output] = fmincon(@rosenbrock,Xinit,[],[],[],[],[],[],@nonlconstr1,opt);
end

function [c,ceq] = nonlconstr1(x) % Nichtlineare Beschränkungsfunktion
c = 0.5^2 - (x(1))^2 - (x(2))^2; ceq = [];

function stop = plot_iterates(x,info,state)
global old
f = rosenbrock(x);
switch state                                          % 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];
```

```
function plot_constraints % Zeichnen der eingestellten Beschränkung
[X1,X2] = meshgrid(-1.5:0.15:1.5);
x1_values = [-1.5:0.15:1.5]; x2_values = [-1.5:0.15:1.5]; r = 0.5;
x1_plot_values = [-r:0.01:r]; x2_plot_values1 = sqrt(r^2-(x1_plot_values).^2);
x2_plot_values2 = -sqrt(r^2-(x1_plot_values).^2);
z_values1 = 100*(x2_plot_values1-x1_plot_values.^2).^2 + (x1_plot_values-1).^2;
z_values2 = 100*(x2_plot_values2-x1_plot_values.^2).^2 + (x1_plot_values-1).^2;
line(x1_plot_values,x2_plot_values1,z_values1,'LineWidth',2,'color','r');
line(x1_plot_values,x2_plot_values2,z_values2,'LineWidth',2,'color','r');

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

function [f, grad, H] = rosenbrock(x)
grad = {}; H = {};
f = 100*(x(2)-x(1)^2)^2 + (x(1)-1)^2;                   % Rosenbrock-Funktion
if 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
```

3.4 Software-Übersicht

Im Folgenden ist eine Auswahl an Software zur Lösung von statischen Optimierungsproblemen zusammengestellt.

Lineare Optimierung

- linprog: MATLAB Optimization Toolbox (kostenpflichtig)
- CPLEX (kostenpflichtig)
  `http://www.ilog.com/products/cplex`
- GLPK: „GNU Linear Programming Kit“ (frei zugänglich)
  `http://www.gnu.org/software/glpk`
- lp_solve: Mixed-Integer Lineare Optimierung (frei zugänglich)
  `http://lpsolve.sourceforge.net`

Quadratische Optimierung

- quadprog: MATLAB Optimization Toolbox (kostenpflichtig)
- CPLEX (kostenpflichtig)
  `http://www.ilog.com/products/cplex`

- OOQP (frei zugänglich)
  `http://pages.cs.wisc.edu/~swright/ooqp`
- qpOASES (frei zugänglich)
  `https://projects.coin-or.org/qpOASES`
- CVX (frei zugänglich)
  `http://cvxr.com/cvx/`
- LOQO (kostenpflichtig)
  `http://www.princeton.edu/~rvdb`

Nichtlineare Optimierung

- fmincon: MATLAB Optimization Toolbox (kostenpflichtig)
- LOQO (kostenpflichtig)
  `http://www.princeton.edu/~rvdb`
- MINOS (kostenpflichtig)
  `http://www.sbsi-sol-optimize.com/asp/sol_product_minos.htm`
- 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

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

[2] D. P. Bertsekas, *Nonlinear Programming*, 2. Aufl. Athena Scientific, 1999.

[3] M. Bazaraa, H. Sherali und C. Shetty, *Nonlinear Programming: Theory and Algorithms*, 3. Aufl. John Wiley & Sons, 2006.

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

[5] L. Biegler, *Nonlinear Programming: Concepts, Algorithms, and Applications to Chemical Processes*. Society for Industrial und Applied Mathematics, 2010.

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

[7] M. Papageorgiou, M. Leibold und M. Buss, *Optimierung: Statische, dynamische, stochastische Verfahren für die Anwendung*, 3. Aufl. Springer, 2012.

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

[9] D. G. Luenberger und Y. Ye, *Linear and Nonlinear Programming*, 3. Aufl., Ser. International Series in Operations Research & Management Science. Springer, 2008, Bd. 116.

[10] 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.

[11] P. E. Gill, W. Murray und M. H. Wright, *Practical Optimization*. Academic Press, 1981.

[12] S.-P. Han, „A Globally Convergent Method for Nonlinear Programming“, *Journal of Optimization Theory and Applications*, Jg. 22, Nr. 3, S. 297–309, 1977.

[13] 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.

[14] 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* $J : \mathcal{X} \to \mathbb{R}$ bezüglich einer (reellen vektorwertigen) Funktion $\mathbf{x}(t)$ aus einem geeigneten *Funktionenraum* $\mathcal{X}$ 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. Typischerweise hat das Kostenfunktional die Form (*Lagrange Problem der Variationsrechnung*)

$$J(\mathbf{x}) = \int_{t_0}^{t_1} l(t, \mathbf{x}(t), \dot{\mathbf{x}}(t))\,\mathrm{d}t \tag{4.1}$$

oder (*Bolza Problem der Variationsrechnung*)

$$J(\mathbf{x}) = \varphi(t_0, \mathbf{x}(t_0), t_1, \mathbf{x}(t_1)) + \int_{t_0}^{t_1} l(t, \mathbf{x}(t), \dot{\mathbf{x}}(t))\,\mathrm{d}t\ . \tag{4.2}$$

Dabei wird $\mathbf{x}(t) = \begin{bmatrix} x_1(t) & \dots & x_n(t) \end{bmatrix}^{\mathrm{T}} : [t_0, t_1] \to \mathbb{R}^n$ 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 3.1.1 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 $\mathcal{X}_{\mathrm{a}\lceil}$ bezeichnet. Die einfachste Form solcher Beschränkungen ist, dass beide Endpunkte fixiert sind, d. h. $\mathbf{x}(t_0) = \mathbf{x}_0$ und $\mathbf{x}(t_1) = \mathbf{x}_1$ bzw. $\mathcal{X}_{\mathrm{a}\lceil} = \{\mathbf{x}(t) \in \mathcal{X} \,|\, \mathbf{x}(t_0) = \mathbf{x}_0,\, \mathbf{x}(t_1) = \mathbf{x}_1\}$. Eine weitere Möglichkeit besteht darin, dass die Trajektorie zwar an einem festen Punkt $(t_0, \mathbf{x}_0)$ startet aber zu einem *freien* Zeitpunkt $t_1 \in [T_0, T]$ auf einer vorgegebenen Kurve Γ definiert durch $\mathbf{x} = \mathbf{g}(t)$ mit $t_0 \le t \le T$ zu liegen kommen muss. In diesem Fall ist die freie Endzeit t_1 eine zu optimierende Größe und die zulässige Menge lautet $\mathcal{X}_{\mathrm{a}\lceil} = \{\mathbf{x}(t) \in \mathcal{X} \,|\, \mathbf{x}(t_0) = \mathbf{x}_0,\ \mathbf{x}(t_1) = \mathbf{g}(t_1)\}$. Andere mögliche Beschränkungen sind so genannte *Pfadbeschränkungen* (Englisch: *path constraints*) oder *Beschränkungen der Lagrangeschen Form*

$$\psi(t, \mathbf{x}(t), \dot{\mathbf{x}}(t)) = 0 \quad \text{bzw.} \quad \psi(t, \mathbf{x}(t), \dot{\mathbf{x}}(t)) \le 0, \quad t \in I \subseteq [t_0, t_1] \tag{4.3}$$

und *isoperimetrische Beschränkungen* der Form

$$\int_{t_0}^{t_1} \psi(t, \mathbf{x}(t), \dot{\mathbf{x}}(t))\, \mathrm{d}t = C \ . \tag{4.4}$$

Die bei der Variationsrechnung typischerweise betrachteten Funktionenräume sind die im Intervall $[t_0, t_1]$ *stetig differenzierbaren Funktionen* $(C^1[t_0, t_1])^n$ und die *stückweise stetig differenzierbaren Funktionen*, im Weiteren als $(\hat{C}^1[t_0, t_1])^n$ bezeichnet. Elemente des Funktionenraumes $(\hat{C}^1[t_0, t_1])^n$ werden dabei im Folgenden auch als global stetig angenommen. Die Definition des *globalen Minimums* $\mathbf{x}^*$ eines Kostenfunktionals $J(\mathbf{x})$ lässt sich ohne Angabe einer Norm direkt in der Form

$$J(\mathbf{x}^*) \leq J(\mathbf{x}), \quad \forall \mathbf{x} \in \mathcal{X}_{\mathrm{a}\lceil} \tag{4.5}$$

angeben. Die Beschreibung des *lokalen* Verhaltens in der Umgebung des Minimums $\mathbf{x}^*$ hingegen verlangt die Definition einer Norm. Eine zulässige Lösung $\mathbf{x}^*$ ist ein *lokales Minimum* in $\mathcal{X}_{\mathrm{a}\lceil}$ bezüglich der Norm $\|\cdot\|$, wenn gilt

$$\exists \gamma > 0 \text{ so, dass gilt } J(\mathbf{x}^*) \leq J(\mathbf{x}), \quad \forall \mathbf{x} \in \mathcal{X}_{\mathrm{a}\lceil} \cap B_\gamma(\mathbf{x}^*) \tag{4.6}$$

mit $B_\gamma(\mathbf{x}^*) = \{\mathbf{x} \in \mathcal{X} \,|\, \|\mathbf{x} - \mathbf{x}^*\| < \gamma\}$. Da in infinit-dimensionalen Vektorräumen Normen grundsätzlich nicht äquivalent sind, kann $\mathbf{x}^*$ zwar bezüglich einer Norm ein lokales Minimum sein, aber bezüglich einer anderen Norm nicht. Im Funktionenraum $(C^1[t_0, t_1])^n$ werden häufig die Normen

$$\|\mathbf{x}(t)\|_\infty := \max_{t_0 \leq t \leq t_1} \|\mathbf{x}(t)\| \quad \text{und} \quad \|\mathbf{x}(t)\|_{1,\infty} := \max_{t_0 \leq t \leq t_1} \|\mathbf{x}(t)\| + \max_{t_0 \leq t \leq t_1} \|\dot{\mathbf{x}}(t)\| \tag{4.7}$$

verwendet, wobei $\|\mathbf{x}(t)\|$ eine Norm im finit-dimensionalen Vektorraum $\mathbb{R}^n$ beschreibt.

4.1.2 Optimalitätsbedingungen

Zur Herleitung notwendiger Optimalitätsbedingungen benötigt man den Begriff der *Variation eines Funktionals*.

Definition 4.1 (Variation eines Funktionals, Gâteaux Ableitung)**.** Die *erste Variation des Funktionals* $J(\mathbf{x})$ am Punkt $\mathbf{x} \in \mathcal{X}$ in Richtung $\boldsymbol{\xi} \in \mathcal{X}$, auch als *Gâteaux Ableitung* von $J(\mathbf{x})$ bezüglich $\boldsymbol{\xi}$ am Punkt $\mathbf{x}$ bezeichnet, ist in der Form

$$\delta J(\mathbf{x}; \boldsymbol{\xi}) := \lim_{\eta \to 0} \frac{J(\mathbf{x} + \eta\boldsymbol{\xi}) - J(\mathbf{x})}{\eta} = \frac{\mathrm{d}}{\mathrm{d}\eta} J(\mathbf{x} + \eta\boldsymbol{\xi}) \bigg|_{\eta=0} \tag{4.8}$$

definiert. Falls $\delta J(\mathbf{x}; \boldsymbol{\xi})$ für alle $\boldsymbol{\xi} \in \mathcal{X}$ definiert ist, dann nennt man $J(\mathbf{x})$ *Gâteaux differenzierbar* am Punkt $\mathbf{x}$.

Für die Existenz der Gâteaux Ableitung muss nicht nur das Funktional $J(\mathbf{x})$ definiert sein, sondern auch die Ableitung von $J(\mathbf{x}+\eta\boldsymbol{\xi})$ bezüglich 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

$$\begin{aligned}\delta J(x;\xi) &= \lim_{\eta\to 0} \frac{\int_{t_0}^{t_1} (x(t)+\eta\xi(t))^2\,\mathrm{d}t - \int_{t_0}^{t_1} x^2(t)\,\mathrm{d}t}{\eta} \\ &= \lim_{\eta\to 0}\left(\int_{t_0}^{t_1} 2x(t)\xi(t)\,\mathrm{d}t + \eta\int_{t_0}^{t_1}\xi^2(t)\,\mathrm{d}t\right) = 2\int_{t_0}^{t_1} x(t)\xi(t)\,\mathrm{d}t\end{aligned} \tag{4.9}$$

für alle $\xi \in C^1[t_0, t_1]$, weshalb $J(x)$ an jedem Punkt $x \in C^1[t_0, t_1]$ Gâteaux differenzierbar ist.

Beispiel 4.2. Man betrachte das Funktional $J(x) = \int_0^1 |x(t)|\,\mathrm{d}t$, $x \in C^1[0, 1]$, welches für jedes $x \in C^1[0, 1]$ im endlichen Intervall $[0, 1]$ einen finiten Wert liefert. Für $x_0(t) = 0$ und $\xi_0(t) = t$ lautet die Gâteaux Ableitung (4.8)

$$\delta J(x_0;\xi_0) = \lim_{\eta\to 0}\frac{1}{\eta}\left(\int_0^1 |x_0+\eta\xi_0|\,\mathrm{d}t - \int_0^1 |x_0|\,\mathrm{d}t\right) = \tag{4.10a}$$

$$= \lim_{\eta\to 0}\operatorname{sgn}(\eta)\int_0^1 |t|\,\mathrm{d}t = \begin{cases}\frac{1}{2}, & \eta\to+0\\ -\frac{1}{2}, & \eta\to-0\end{cases}. \tag{4.10b}$$

Dabei erkennt man, dass in Richtung $\xi_0 = t$ an der Stelle $x_0 = 0$ die Gâteaux Ableitung nicht existiert.

Die Gâteaux Ableitung ist eine *lineare Operation*, weshalb gilt

$$\delta(J_1 + J_2)(\mathbf{x};\boldsymbol{\xi}) = \delta J_1(\mathbf{x};\boldsymbol{\xi}) + \delta J_2(\mathbf{x};\boldsymbol{\xi}) \tag{4.11}$$

und für jedes reelle α gilt die Beziehung

$$\delta J(\mathbf{x};\alpha\boldsymbol{\xi}) = \alpha\delta J(\mathbf{x};\boldsymbol{\xi})\ . \tag{4.12}$$

Basierend auf der Gâteaux Ableitung lässt sich nun der Begriff der *zulässigen Richtung* eines Funktionals definieren.

Definition 4.2 (Zulässige Richtung)**.** $J : \mathcal{X}_{\mathrm{a}\lceil} \to \mathbb{R}$ sei ein Funktional welches in einer Teilmenge $\mathcal{X}_{\mathrm{a}\lceil}$ eines normierten linearen Vektorraums $(\mathcal{X}, \|\cdot\|)$ definiert ist. An einem (zulässigen) Punkt $\mathbf{x}$ im Inneren von $\mathcal{X}_{\mathrm{a}\lceil}$ bezeichnet man $\boldsymbol{\xi} \in \mathcal{X}$ 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 $\mathbf{x} + \eta\boldsymbol{\xi} \in \mathcal{X}_{\mathrm{a}\lceil}$ für alle $\eta \in (-\varepsilon, \varepsilon)$ gilt.

Die Bedingung (b) verlangt natürlich, dass $\mathbf{x}$ im Inneren von $\mathcal{X}_{\mathrm{a}\lceil}$ liegt. Eine zulässige Richtung $\boldsymbol{\xi}$ am Punkt $\bar{\mathbf{x}}$ für die gilt $\delta J(\bar{\mathbf{x}};\boldsymbol{\xi}) < 0$ wird *Abstiegsrichtung* des Funktionals J am Punkt $\bar{\mathbf{x}}$ bezeichnet. Dies stellt eine Generalisierung der Abstiegsrichtung $\mathbf{d}$ der Kostenfunktion $f(\mathbf{x})$ im finit-dimensionalen Fall mit $\mathbf{d}^{\mathrm{T}}(\nabla f)(\bar{\mathbf{x}}) < 0$ am Punkt $\bar{\mathbf{x}}$ gemäß Satz 2.1 dar. Es gilt nun folgendes Lemma.

Lemma 4.1 (Ausschluss eines Minimums). *Wenn J ein Funktional in einem normierten linearen Vektorraum $(\mathcal{X}, \|\cdot\|)$ beschreibt und an einem Punkt $\bar{\mathbf{x}} \in \mathcal{X}_{a\lceil}$ eine zulässige Richtung $\boldsymbol{\xi} \in \mathcal{X}$ so existiert, dass gilt $\delta J(\bar{\mathbf{x}}; \boldsymbol{\xi}) < 0$, dann kann $\bar{\mathbf{x}}$ kein lokales Minimum sein.*

Beweisskizze: Gemäß Definition 4.1 gilt

$$\delta J(\bar{\mathbf{x}}; \boldsymbol{\xi}) = \lim_{\eta \to 0} \frac{J(\bar{\mathbf{x}} + \eta\boldsymbol{\xi}) - J(\bar{\mathbf{x}})}{\eta} < 0 \tag{4.13}$$

und es existiert ein $\gamma > 0$ so, dass

$$J(\bar{\mathbf{x}} + \eta\boldsymbol{\xi}) < J(\bar{\mathbf{x}}), \quad \forall \eta \in (0, \gamma) \ . \tag{4.14}$$

Da nun $\boldsymbol{\xi}$ eine zulässige Richtung gemäß Definition 4.2 ist, kann das Funktional J am Punkt $\bar{\mathbf{x}}$ in Richtung $\eta\boldsymbol{\xi}$ mit beliebigem $\eta \in (0, \gamma)$ weiter verkleinert werden. Da unabhängig von der verwendeten Norm $\|\bar{\mathbf{x}} + \eta\boldsymbol{\xi} - \bar{\mathbf{x}}\| = \|\eta\boldsymbol{\xi}\| \to 0$ für $\eta \to 0$ gilt, findet man stets einen hinreichend kleinen Wert $\eta \in (0, \gamma)$, so dass $\bar{\mathbf{x}} + \eta\boldsymbol{\xi}$ im Sinne der Norm $\|\cdot\|$ in der Umgebung von $\bar{\mathbf{x}}$ liegt. Folglich kann $\bar{\mathbf{x}}$ kein lokales Minimum sein. □

Die notwendigen Bedingungen erster Ordnung für ein lokales Minimum eines Funktionals lassen sich nun wie folgt formulieren [1].

Satz 4.1 (Notwendige Bedingungen erster Ordnung). *Angenommen $\mathbf{x}^* \in \mathcal{X}_{a\lceil}$ ist ein (lokales) Minimum des Funktionals J, welches in einer Teilmenge $\mathcal{X}_{a\lceil}$ eines normierten linearen Vektorraums $(\mathcal{X}, \|\cdot\|)$ definiert ist. Dann gilt*

$$\delta J(\mathbf{x}^*; \boldsymbol{\xi}) = 0 \tag{4.15}$$

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 (4.1) mit festem Anfangs- und Endpunkt.

Satz 4.2 (Euler-Lagrange Gleichungen). *Gegeben sei das Funktional*

$$J(\mathbf{x}) = \int_{t_0}^{t_1} l(t, \mathbf{x}(t), \dot{\mathbf{x}}(t)) \, \mathrm{d}t \tag{4.16}$$

mit der zulässigen Menge $\mathcal{X}_{a\lceil} = \left\{ \mathbf{x}(t) \in \left(C^1[t_0, t_1]\right)^n \,|\, \mathbf{x}(t_0) = \mathbf{x}_0, \, \mathbf{x}(t_1) = \mathbf{x}_1 \right\}$ 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 $\mathcal{X}_{a\lceil}$ bezeichnet, dann erfüllt $\mathbf{x}^*(t)$ die Euler-Lagrange Gleichungen

$$\frac{\mathrm{d}}{\mathrm{d}t}\left(\frac{\partial}{\partial \dot{x}_i} l\right)(t, \mathbf{x}^*(t), \dot{\mathbf{x}}^*(t)) - \left(\frac{\partial}{\partial x_i} l\right)(t, \mathbf{x}^*(t), \dot{\mathbf{x}}^*(t)) = 0 \tag{4.17}$$

für alle $t \in [t_0, t_1]$ und $i = 1, \ldots, n$.

Beweis. Da $\mathbf{x}^*$ ein Minimum ist, muss wegen Satz 4.1 gelten

$$\begin{aligned}\delta J(\mathbf{x}^*; \boldsymbol{\xi}) &= \frac{\mathrm{d}}{\mathrm{d}\eta} J(\mathbf{x}^* + \eta\boldsymbol{\xi})\Big|_{\eta=0} = \int_{t_0}^{t_1} \frac{\mathrm{d}}{\mathrm{d}\eta} l\Big(t, \mathbf{x}^*(t) + \eta\boldsymbol{\xi}(t), \dot{\mathbf{x}}^*(t) + \eta\dot{\boldsymbol{\xi}}(t)\Big)\,\mathrm{d}t\Big|_{\eta=0} \\ &= \int_{t_0}^{t_1} \left[\left(\frac{\partial}{\partial \mathbf{x}} l\right)(t, \mathbf{x}^*(t), \dot{\mathbf{x}}^*(t))\boldsymbol{\xi} + \left(\frac{\partial}{\partial \dot{\mathbf{x}}} l\right)(t, \mathbf{x}^*(t), \dot{\mathbf{x}}^*(t))\dot{\boldsymbol{\xi}}\right] \mathrm{d}t = 0\ .\end{aligned} \tag{4.18}$$

Wegen der stetigen Differenzierbarkeit der Lagrangeschen Dichte l und da $\boldsymbol{\xi} \in \left(C^1[t_0, t_1]\right)^n$ ist der Integrand von (4.18) im Intervall $[t_0, t_1]$ stetig und daher ist das Funktional $J(\mathbf{x})$ an allen Punkten $\mathbf{x} \in \left(C^1[t_0, t_1]\right)^n$ 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 (4.18) eine partielle Integration durch, so erhält man

$$\int_{t_0}^{t_1} \left(\frac{\partial}{\partial \mathbf{x}} l\right)\boldsymbol{\xi} + \left(\frac{\partial}{\partial \dot{\mathbf{x}}} l\right)\dot{\boldsymbol{\xi}}\,\mathrm{d}t = \int_{t_0}^{t_1} \left[\left(\frac{\partial}{\partial \mathbf{x}} l\right) - \frac{\mathrm{d}}{\mathrm{d}t}\left(\frac{\partial}{\partial \dot{\mathbf{x}}} l\right)\right]\boldsymbol{\xi}\,\mathrm{d}t + \underbrace{\left[\left(\frac{\partial}{\partial \dot{\mathbf{x}}} l\right)\boldsymbol{\xi}\right]_{t_0}^{t_1}}_{=0} = 0\ . \tag{4.19}$$

Wählt man nun für ein festes $i = 1, \ldots, n$ die Richtung $\boldsymbol{\xi} = \begin{bmatrix}\xi_1 & \ldots & \xi_n\end{bmatrix}^{\mathrm{T}} \in \left(C^1[t_0, t_1]\right)^n$ 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

$$\int_{t_0}^{t_1} \left[\left(\frac{\partial}{\partial x_i} l\right) - \frac{\mathrm{d}}{\mathrm{d}t}\left(\frac{\partial}{\partial \dot{x}_i} l\right)\right]\xi_i\,\mathrm{d}t = 0\ . \tag{4.20}$$

Gemäß dem nachfolgend angeführten *Fundamentallemma der Variationsrechnung* folgt aus (4.20) unmittelbar das Ergebnis (4.17). □

Lemma 4.2 (Fundamentallemma der Variationsrechnung)**.** *Angenommen $g(t)$ ist eine stückweise stetige Funktion auf dem Intervall $[t_0, t_1]$ und es gilt*

$$\int_{t_0}^{t_1} g(t)\xi_i(t)\,\mathrm{d}t = 0 \tag{4.21}$$

für alle stückweise stetigen Funktionen $\xi_i(t)$ im Intervall $[t_0, t_1]$, dann folgt fast überall (abgesehen von einer abzählbaren Menge von Punkten) $g(t) = 0$, $t \in [t_0, t_1]$.

Eine Funktion $\bar{\mathbf{x}}(t)$, die die Euler-Lagrange Gleichungen (4.17) erfüllt, wird auch als *stationäre Funktion der Lagrangeschen Dichte l* bezeichnet. In manchen Literaturstellen werden diese Funktionen auch als *extremale Funktionen* oder nur *Extremale* bezeichnet, obwohl es sein kann, dass sie weder ein Minimum noch ein Maximum des Kostenfunktionals beschreiben.

Mit Satz 4.2 ist es also gelungen, die Minimierung des Funktionals (4.1) in ein *Zweipunktrandwertproblem* mit den Euler-Lagrange Gleichungen umzuformulieren. Das erhaltene

Randwertproblem kann meist mit gängigen numerischen Methoden [2–4], wie z. B. Einfach-Schießverfahren, Mehrfach-Schießverfahren und Kollokationsverfahren, gelöst werden. Die Lösung der Euler-Lagrange Gleichungen (4.17) kann für Spezialfälle auch mit Hilfe so genannter *erster Integrale* formuliert werden:

(a) Die Lagrangesche Dichte hängt nicht von der unabhängigen Variablen t ab, d. h. $l = l(\mathbf{x}, \dot{\mathbf{x}})$. Mit der *Hamiltonfunktion*

$$H = \left(\frac{\partial}{\partial \dot{\mathbf{x}}} l\right)\dot{\mathbf{x}} - l(\mathbf{x}, \dot{\mathbf{x}}) \tag{4.22}$$

folgt aus den Euler-Lagrange Gleichungen (4.17), dass

$$\begin{aligned} \frac{\mathrm{d}}{\mathrm{d}t} H &= \frac{\mathrm{d}}{\mathrm{d}t}\left(\frac{\partial}{\partial \dot{\mathbf{x}}} l\right)\dot{\mathbf{x}} + \left(\frac{\partial}{\partial \dot{\mathbf{x}}} l\right)\ddot{\mathbf{x}} - \left(\frac{\partial}{\partial \mathbf{x}} l\right)\dot{\mathbf{x}} - \left(\frac{\partial}{\partial \dot{\mathbf{x}}} l\right)\ddot{\mathbf{x}} \\ &= \left(\frac{\mathrm{d}}{\mathrm{d}t}\left(\frac{\partial}{\partial \dot{\mathbf{x}}} l\right) - \left(\frac{\partial}{\partial \mathbf{x}} l\right)\right)\dot{\mathbf{x}} = 0 \ . \end{aligned} \tag{4.23}$$

D. h. die Hamiltonfunktion H ist entlang von stationären Funktionen konstant und bildet damit eine *Invariante* des Systems.

(b) Die Lagrangesche Dichte hängt nicht von $\mathbf{x}$ ab, d. h. $l = l(t, \dot{\mathbf{x}})$. Dann folgt aus den Euler-Lagrange Gleichungen (4.17), dass $\frac{\partial}{\partial \dot{x}_i} l$, $i = 1, \ldots, n$ *Invarianten* des Systems sind, denn es gilt

$$\frac{\mathrm{d}}{\mathrm{d}t}\left(\frac{\partial}{\partial \dot{x}_i} l\right) = 0 \ . \tag{4.24}$$

Aufgabe 4.1. Nehmen Sie an, dass $l(\mathbf{x}, \dot{\mathbf{x}})$ die Lagrangefunktion eines Starrkörpersystems ist (siehe Skriptum Fachvertiefung: Automatisierungs- und Regelungstechnik oder Regelungssysteme 2) und $\mathbf{x}$ bzw. $\dot{\mathbf{x}}$ die generalisierten Lagekoordinaten und deren Geschwindigkeiten bezeichnen. Geben Sie eine physikalische Interpretation der Hamiltonfunktion H von (4.22) und der darin auftretenden Größen $\frac{\partial}{\partial \dot{x}_i} l$, $i = 1, \ldots, n$ an.

Bemerkung 4.1. Konservative Starrkörpersysteme erfüllen die Euler-Lagrange Gleichungen (4.17). Dies gilt im Allgemeinen nicht für nicht-konservative Starrkörpersysteme. Für sie lauten die Euler-Lagrange Gleichungen

$$\frac{\mathrm{d}}{\mathrm{d}t}\left(\frac{\partial}{\partial \dot{x}_i} l\right) - \left(\frac{\partial}{\partial x_i} l\right) = \tau_i \tag{4.25}$$

für alle $t \in [t_0, t_1]$ und $i = 1, \ldots, n$ mit den externen generalisierten Kräften τ_j (siehe Skriptum Fachvertiefung: Automatisierungs- und Regelungstechnik oder Regelungssysteme 2).

Beispiel 4.3 (Elastischer Zugstab belastet durch Eigengewicht). Ein gerader, linear elastischer Stab habe die Zugsteifigkeit k und im unbelasteten Zustand die Masse pro Längeneinheit $\bar{m}$ und die Länge x_1. Der Stab wird am Punkt $x = x_0 = 0$ senkrecht befestigt und durch sein Eigengewicht (Erdbeschleunigung g) belastet. Es soll das Verschiebungsfeld $y(x)$ zufolge der Eigengewichtsbelastung berechnet werden. Die Längskoordinate x sei materialfest, d. h. sie wird im unbelasteten Zustand gemessen.

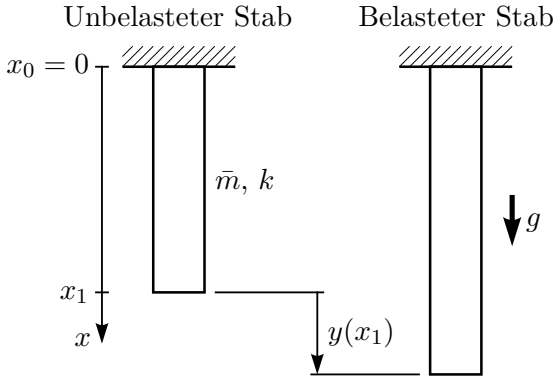


Abbildung 4.1: Elastischer Zugstab belastet durch Eigengewicht.

Zur Lösung dieser Aufgabe kann das *Hamiltonsche Prinzip* der Mechanik [5, 6] verwendet werden. Angewandt auf den Sonderfall der hier rein statischen Beanspruchung besagt es, dass die potentielle Energie des Stabes im statischen Gleichgewicht extremal sein muss [7]. Für ein stabiles statisches Gleichgewicht muss sie minimal sein.

Die bis auf einen konstanten Term definierte potentielle Energie

$$J(y) = \int_0^{x_1} \frac{k(y'(x))^2}{2} - \bar{m}gy(x)\ \mathrm{d}x \tag{4.26}$$

des Stabes setzt sich aus der Dehnungsenergie mit der Längsdehnung $y'(x)$ und der potentiellen Höhenenergie zusammen. Am Befestigungspunkt $x = x_0 = 0$ des Stabes darf keine Verschiebung auftreten und es gilt

$$y(0) = 0\ . \tag{4.27a}$$

Da am freien Ende $x = x_1$ des Stabes die Zugkraft 0 beträgt, muss dort auch die Dehnung verschwinden, d. h.

$$y'(x_1) = 0\ . \tag{4.27b}$$

Die Minimierung des Funktionals (4.26) unter Berücksichtigung der Randbedingungen (4.27) kann mit Hilfe der Variationsrechnung erfolgen. Da in der Lagrangeschen Dichte $l(y, y') = k(y')^2/2 - \bar{m}gy$ die unabhängige Variable x nicht explizit auftritt,

muss gemäß (4.23) die Hamiltonfunktion eine Invariante des Systems sein, d. h.

$$H = \left(\frac{\partial}{\partial y'} l\right)(y, y')y' - l(y, y') = \frac{k(y')^2}{2} + \bar{m}gy = c_1 = \text{konst.} \tag{4.28}$$

Die Integration dieser Differentialgleichung liefert

$$\left[-\sqrt{2k}\frac{\sqrt{c_1 - \bar{m}gy}}{\bar{m}g}\right]_{y(0)}^{y(x)} = x \ . \tag{4.29}$$

Die Werte c_1 und $y(0)$ folgen schließlich aus den Randbedingungen (4.27) und für die Lösung ergibt sich

$$y(x) = \frac{\bar{m}g}{k}\left(x_1 x - \frac{x^2}{2}\right) \ . \tag{4.30}$$

Alternativ kann diese Aufgabe natürlich auch direkt mit Satz 4.2 gelöst werden. Aus der Euler-Lagrange Gleichung (4.17) folgt

$$\frac{\partial}{\partial x}\left(\frac{\partial}{\partial y'} l\right)(y, y') - \frac{\partial}{\partial y} l(y, y') = ky'' + \bar{m}g = 0 \ . \tag{4.31}$$

Die Integration dieser Differentialgleichung liefert bei Berücksichtigung der Randbedingungen (4.27) ebenfalls die Lösung (4.30).

Analog zum finit-dimensionalen Fall, siehe Satz 2.2, können auch für die Minimierung von Funktionalen notwendige Bedingungen zweiter Ordnung formuliert werden.

Satz 4.3 (Notwendige Bedingungen zweiter Ordnung - Legendre Bedingung)**.** *Angenommen* $\mathbf{x}^* \in \mathcal{X}_{a\lceil}$ *ist ein lokales Minimum des Funktionals*

$$J(\mathbf{x}) = \int_{t_0}^{t_1} l(t, \mathbf{x}(t), \dot{\mathbf{x}}(t))\,\mathrm{d}t \tag{4.32}$$

mit der zulässigen Menge $\mathcal{X}_{a\lceil} = \{\mathbf{x}(t) \in (C^1[t_0, t_1])^n \,|\, \mathbf{x}(t_0) = \mathbf{x}_0, \mathbf{x}(t_1) = \mathbf{x}_1\}$ *und der zweifach stetig differenzierbaren Lagrangeschen Dichte* $l : \mathbb{R} \times \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}$, *dann erfüllt* $\mathbf{x}^*$ *die Euler-Lagrange Gleichungen* (4.17) *und die so genannte* Legendre Bedingung

$$\mathbf{d}^{\mathrm{T}}\left(\frac{\partial^2 l}{\partial \dot{\mathbf{x}}^2}\right)(t, \mathbf{x}^*(t), \dot{\mathbf{x}}^*(t))\mathbf{d} \geq 0 \qquad \forall\, \mathbf{d} \in \mathbb{R}^n,\ t \in [t_0, t_1] \ . \tag{4.33}$$

Satz 4.4 (Hinreichende Bedingungen zweiter Ordnung - Konvexitätsbedingung)**.** *Gegeben sei das Funktional*

$$J(\mathbf{x}) = \int_{t_0}^{t_1} l(t, \mathbf{x}(t), \dot{\mathbf{x}}(t))\,\mathrm{d}t \tag{4.34}$$

mit der zulässigen Menge $\mathcal{X}_{\mathrm{a}\lceil} = \{\mathbf{x}(t) \in \left(C^1[t_0, t_1]\right)^n \,|\, \mathbf{x}(t_0) = \mathbf{x}_0,\ \mathbf{x}(t_1) = \mathbf{x}_1\}$ *und der zweifach stetig differenzierbaren Lagrangeschen Dichte* $l : \mathbb{R} \times \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}$. *Erfüllt eine Funktion* $\mathbf{x}^* \in \mathcal{X}_{\mathrm{a}\lceil}$ *die Euler-Lagrange Gleichungen* (4.17) *und die sogenannte* Konvexitätsbedingung

$$\mathbf{d}^{\mathrm{T}} \begin{bmatrix} \left(\frac{\partial^2 l}{\partial \mathbf{x}^2}\right)(t, \mathbf{x}^*(t), \dot{\mathbf{x}}^*(t)) & \left(\frac{\partial^2 l}{\partial \mathbf{x} \partial \dot{\mathbf{x}}}\right)(t, \mathbf{x}^*(t), \dot{\mathbf{x}}^*(t)) \\ \left(\frac{\partial^2 l}{\partial \dot{\mathbf{x}} \partial \mathbf{x}}\right)(t, \mathbf{x}^*(t), \dot{\mathbf{x}}^*(t)) & \left(\frac{\partial^2 l}{\partial \dot{\mathbf{x}}^2}\right)(t, \mathbf{x}^*(t), \dot{\mathbf{x}}^*(t)) \end{bmatrix} \mathbf{d} \geq 0 \quad \forall\, \mathbf{d} \in \mathbb{R}^{2n},\ t \in [t_0, t_1]\ , \tag{4.35}$$

d. h. $l(t, \mathbf{x}(t), \dot{\mathbf{x}}(t))$ *ist lokal konvex in* $\mathbf{x}(t)$ *und* $\dot{\mathbf{x}}(t)$, *dann ist* $\mathbf{x}^*(t)$ *ein lokales Minimum des Funktionals* J. *Ist* $l(t, \mathbf{x}(t), \dot{\mathbf{x}}(t))$ *sogar strikt lokal konvex in* $\mathbf{x}(t)$ *und* $\dot{\mathbf{x}}(t)$ *(Gleichung* (4.35) *ist dann nur für* $\mathbf{d} = \mathbf{0}$ *mit Gleichheit erfüllt), dann ist* $\mathbf{x}^*(t)$ *ein striktes lokales Minimum.*

Der Beweis zu Satz 4.4 findet sich z. B. in [1, 8].

Satz 4.2 behandelt das Lagrange Problem der Variationsrechnung (4.1). Im nächsten Schritt soll das *Bolza Problem der Variationsrechnung* (4.2) *mit freier Endzeit* näher untersucht werden.

Satz 4.5 (Euler-Lagrange Gleichungen für freie Endzeit)**.** *Gegeben sei das Funktional*

$$J(t_1, \mathbf{x}) = \varphi(t_1, \mathbf{x}(t_1)) + \int_{t_0}^{t_1} l(t, \mathbf{x}(t), \dot{\mathbf{x}}(t))\, \mathrm{d}t \tag{4.36}$$

mit der zulässigen Menge $\mathcal{X}_{\mathrm{a}\lceil} = \left\{(t_1, \mathbf{x}(t)) \in (t_0, T) \times \left(C^1[t_0, T]\right)^n \,|\, \mathbf{x}(t_0) = \mathbf{x}_0\right\}$, *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* $\mathcal{X}_{\mathrm{a}\lceil}$ *bezeichnet, dann erfüllt* $\mathbf{x}^*(t)$ *die* Euler-Lagrange Gleichungen (4.17) *im Intervall* $[t_0, t_1^*]$ *und es gelten die Anfangsbedingung* $\mathbf{x}^*(t_0) = \mathbf{x}_0$ *sowie die* Transversalitätsbedingungen

$$\left[\frac{\partial}{\partial \dot{\mathbf{x}}} l + \frac{\partial}{\partial \mathbf{x}} \varphi\right]_{t=t_1^*,\ \mathbf{x}=\mathbf{x}^*} = \mathbf{0}^{\mathrm{T}} \tag{4.37a}$$

$$\left[l - \left(\frac{\partial}{\partial \dot{\mathbf{x}}} l\right)\dot{\mathbf{x}} + \frac{\partial}{\partial t} \varphi\right]_{t=t_1^*,\ \mathbf{x}=\mathbf{x}^*} = 0\ . \tag{4.37b}$$

Beweis. Wenn man die Endzeit t_1^* fixiert, dann folgt aus Satz 4.2 unmittelbar, dass die optimale Lösung $\mathbf{x}^*(t)$ im Intervall $[t_0, t_1^*]$ die Euler-Lagrange Gleichungen (4.17) erfüllt. Um die optimale Endzeit t_1^* zu berechnen, nimmt man an, dass $\mathbf{x}(t)$ in einem hinreichend großen Intervall $[t_0, T]$, $T \gg t_1^*$ definiert ist und betrachtet den linearen Funktionenraum $\mathbb{R} \times \left(C^1[t_0, T]\right)^n$. Die Gâteaux Ableitung gemäß Definition 4.1 wird

dann in der Form

$$\begin{aligned}\delta J(t_1,\mathbf{x};\tau,\boldsymbol{\xi}) &:= \lim_{\eta\to 0}\frac{J(t_1+\eta\tau,\mathbf{x}+\eta\boldsymbol{\xi})-J(t_1,\mathbf{x})}{\eta}\\ &= \frac{\mathrm{d}}{\mathrm{d}\eta}J(t_1+\eta\tau,\mathbf{x}+\eta\boldsymbol{\xi})\Big|_{\eta=0}\end{aligned} \tag{4.38}$$

erweitert und die notwendige Bedingung für ein Minimum (4.15) von Satz 4.1 ausgewertet. Wendet man nun (4.38) für beliebiges η auf (4.36) an, so erhält man

$$\begin{aligned}&\frac{\mathrm{d}}{\mathrm{d}\eta}J(t_1^*+\eta\tau,\mathbf{x}^*+\eta\boldsymbol{\xi}) =\\ &=\frac{\mathrm{d}}{\mathrm{d}\eta}\varphi(t_1^*+\eta\tau,\mathbf{x}^*(t_1^*+\eta\tau)+\eta\boldsymbol{\xi}(t_1^*+\eta\tau))+\frac{\mathrm{d}}{\mathrm{d}\eta}\int_{t_0}^{t_1^*+\eta\tau} l\Big(t,\mathbf{x}^*+\eta\boldsymbol{\xi},\dot{\mathbf{x}}^*+\eta\dot{\boldsymbol{\xi}}\Big)\,\mathrm{d}t\\ &=\Big[\Big(\frac{\partial}{\partial t}\varphi\Big)\tau+\frac{\partial}{\partial\mathbf{x}}\varphi\Big[\underbrace{\frac{\partial}{\partial t}\mathbf{x}}_{\dot{\mathbf{x}}}+\eta\underbrace{\frac{\partial}{\partial t}\boldsymbol{\xi}}_{\dot{\boldsymbol{\xi}}}\Big]\tau+\frac{\partial}{\partial\mathbf{x}}\varphi\boldsymbol{\xi}\Big]_{t=t_1^*+\eta\tau,\ \mathbf{x}=\mathbf{x}^*+\eta\boldsymbol{\xi}}\\ &\quad+\int_{t_0}^{t_1^*+\eta\tau}\frac{\mathrm{d}}{\mathrm{d}\eta}l\Big(t,\mathbf{x}^*+\eta\boldsymbol{\xi},\dot{\mathbf{x}}^*+\eta\dot{\boldsymbol{\xi}}\Big)\,\mathrm{d}t+\tau\Big[l\Big(t,\mathbf{x}^*+\eta\boldsymbol{\xi},\dot{\mathbf{x}}^*+\eta\dot{\boldsymbol{\xi}}\Big)\Big]_{t=t_1^*+\eta\tau}\\ &=\Big[\Big(\frac{\partial}{\partial t}\varphi\Big)\tau+\frac{\partial}{\partial\mathbf{x}}\varphi\Big(\dot{\mathbf{x}}+\eta\dot{\boldsymbol{\xi}}\Big)\tau+\frac{\partial}{\partial\mathbf{x}}\varphi\boldsymbol{\xi}\Big]_{t=t_1^*+\eta\tau,\ \mathbf{x}=\mathbf{x}^*+\eta\boldsymbol{\xi}}\\ &\quad+\tau\Big[l\Big(t,\mathbf{x}^*+\eta\boldsymbol{\xi},\dot{\mathbf{x}}^*+\eta\dot{\boldsymbol{\xi}}\Big)\Big]_{t=t_1^*+\eta\tau}+\Big[\Big(\frac{\partial}{\partial\dot{\mathbf{x}}}l\Big)\Big(t,\mathbf{x}^*+\eta\boldsymbol{\xi},\dot{\mathbf{x}}^*+\eta\dot{\boldsymbol{\xi}}\Big)\boldsymbol{\xi}\Big]_{t_0}^{t_1^*+\eta\tau}\\ &\quad+\int_{t_0}^{t_1^*+\eta\tau}\Big(\Big(\frac{\partial}{\partial\mathbf{x}}l\Big)\Big(t,\mathbf{x}^*+\eta\boldsymbol{\xi},\dot{\mathbf{x}}^*+\eta\dot{\boldsymbol{\xi}}\Big)-\frac{\mathrm{d}}{\mathrm{d}t}\Big(\frac{\partial}{\partial\dot{\mathbf{x}}}l\Big)\Big(t,\mathbf{x}^*+\eta\boldsymbol{\xi},\dot{\mathbf{x}}^*+\eta\dot{\boldsymbol{\xi}}\Big)\Big)\boldsymbol{\xi}\,\mathrm{d}t\ .\end{aligned} \tag{4.39}$$

Wertet man (4.39) für $\eta=0$ aus, so lautet die notwendige Optimalitätsbedingung

$$\begin{aligned}\delta J(t_1^*,\tau;\mathbf{x}^*,\boldsymbol{\xi}) &= \frac{\mathrm{d}}{\mathrm{d}\eta}J(t_1^*+\eta\tau,\mathbf{x}^*+\eta\boldsymbol{\xi})\Big|_{\eta=0}\\ &=\Big[\Big(\frac{\partial}{\partial t}\varphi\Big)\tau+\frac{\partial}{\partial\mathbf{x}}\varphi(\dot{\mathbf{x}}\tau+\boldsymbol{\xi})+\tau l\Big]_{t=t_1^*,\mathbf{x}=\mathbf{x}^*}+\Big[\Big(\frac{\partial}{\partial\dot{\mathbf{x}}}l\Big)(t,\mathbf{x}^*,\dot{\mathbf{x}}^*)\boldsymbol{\xi}\Big]_{t_0}^{t_1^*}\\ &\quad+\int_{t_0}^{t_1^*}\Big[\Big(\frac{\partial}{\partial\mathbf{x}}l\Big)(t,\mathbf{x}^*,\dot{\mathbf{x}}^*)-\frac{\mathrm{d}}{\mathrm{d}t}\Big(\frac{\partial}{\partial\dot{\mathbf{x}}}l\Big)(t,\mathbf{x}^*,\dot{\mathbf{x}}^*)\Big]\boldsymbol{\xi}\,\mathrm{d}t=0\ .\end{aligned} \tag{4.40}$$

Da der Anfangswert mit $\mathbf{x}(t_0)=\mathbf{x}_0$ festgelegt ist, muss für eine zulässige Richtung $\boldsymbol{\xi}$ die Bedingung $\boldsymbol{\xi}(t_0)=\mathbf{0}$ gelten. Im Weiteren erfüllt die optimale Lösung $\mathbf{x}^*(t)$ im

Intervall $[t_0, t_1^*]$ die Euler-Lagrange Gleichungen (4.17), weshalb sich (4.40) zu

$$\begin{aligned}&\left[\left(\frac{\partial}{\partial t}\varphi\right)\tau + \frac{\partial}{\partial \mathbf{x}}\varphi(\dot{\mathbf{x}}\tau + \boldsymbol{\xi}) + \tau l\right]_{t=t_1^*, \mathbf{x}=\mathbf{x}^*} + \left[\left(\frac{\partial}{\partial \dot{\mathbf{x}}} l\right)(\boldsymbol{\xi} + \dot{\mathbf{x}}\tau - \dot{\mathbf{x}}\tau)\right]_{t=t_1^*, \mathbf{x}=\mathbf{x}^*} \\ &= \tau\left[l - \left(\frac{\partial}{\partial \dot{\mathbf{x}}} l\right)\dot{\mathbf{x}} + \frac{\partial}{\partial t}\varphi\right]_{t=t_1^*,\ \mathbf{x}=\mathbf{x}^*} + \left[\frac{\partial}{\partial \mathbf{x}}\varphi + \frac{\partial}{\partial \dot{\mathbf{x}}} l\right]_{t=t_1^*,\ \mathbf{x}=\mathbf{x}^*} (\dot{\mathbf{x}}^*(t_1^*)\tau + \boldsymbol{\xi}(t_1^*)) = 0\end{aligned} \tag{4.41}$$

vereinfacht. Wenn die Endzeit t_1 und der Endwert $\mathbf{x}(t_1)$ frei sind, dann sind τ und $\boldsymbol{\xi}(t_1^*)$ unabhängig voneinander frei wählbar, weshalb (4.41) nur dann Null ist, wenn die *Transversalitätsbedingungen* (4.37) erfüllt sind. □

Das Ergebnis von Satz 4.5 lässt sich nun wie folgt verallgemeinern.

(a) Wenn die *Endzeit fest ist*, dann gilt $t_1 = t_1^*$ und damit $\tau = 0$, womit automatisch der erste Term von (4.41) verschwindet. Es liegt somit keine Transversalitätsbedingung (4.37b) vor.

(i) Wenn für eine Komponente $x_j(t),\ j \in \{1, \dots, n\}$ von $\mathbf{x}(t)$ gilt, dass *der Endwert* $x_j(t_1^*) = x_j^*(t_1^*) = x_{j1}$ *fest ist*, dann muss für diese Komponente $\xi_j(t_1^*) = 0$ gelten, womit der zugehörige Eintrag im zweiten Term von (4.41) automatisch verschwindet und keine Transversalitätsbedingung für diese Komponente vorliegt. Dieser Fall entspricht dem Ergebnis von Satz 4.2.

(ii) Wenn für eine Komponente $x_j(t),\ j \in \{1, \dots, n\}$ von $\mathbf{x}(t)$ gilt, dass *der Endwert* $x_j(t_1^*)$ *frei ist*, dann lautet die *Transversalitätsbedingung* gemäß (4.41) für diese Komponente

$$\left[\frac{\partial}{\partial \dot{x}_j} l + \frac{\partial}{\partial x_j}\varphi\right]_{t=t_1^*,\ \mathbf{x}=\mathbf{x}^*} = 0\ . \tag{4.42}$$

(b) Wenn *die Endzeit frei ist*, dann muss die Transversalitätsbedingung (4.37b)

$$\left[l - \left(\frac{\partial}{\partial \dot{\mathbf{x}}} l\right)\dot{\mathbf{x}} + \frac{\partial}{\partial t}\varphi\right]_{t=t_1^*,\ \mathbf{x}=\mathbf{x}^*} = 0 \tag{4.43}$$

gelten.

(i) Wenn für eine Komponente $x_j(t),\ j \in \{1, \dots, n\}$ von $\mathbf{x}(t)$ gilt, dass *der Endwert* $x_j^*(t_1^*) = x_{j1}$ *fest ist*, dann muss für diese Komponente eine zulässige Richtung (τ, ξ_j) die Bedingung

$$x_{j1} = x_j^*(t_1^* + \eta\tau) + \eta\xi_j(t_1^* + \eta\tau) \tag{4.44}$$

bzw.

$$0 = \frac{\partial}{\partial \eta} x_{j1}\bigg|_{\eta=0} = \xi_j(t_1^*) + \tau\dot{x}_j^*(t_1^*) \tag{4.45}$$

erfüllen. Damit verschwindet der zugehörige Eintrag im zweiten Term von (4.41) und es liegt keine weitere Transversalitätsbedingung für diese Komponente vor.

(ii) Wenn für eine Komponente $x_j(t),\ j \in \{1,\dots,n\}$ von $\mathbf{x}(t)$ gilt, dass *der Endwert* $x_j^*(t_1^*)$ *frei ist*, dann lautet, analog zum Fall (a)(ii), die *Transversalitätsbedingung* für diese Komponente

$$\left[\frac{\partial}{\partial \dot{x}_j} l + \frac{\partial}{\partial x_j}\varphi\right]_{t=t_1^*,\ \mathbf{x}=\mathbf{x}^*} = 0\ . \tag{4.46}$$

4.1.3 Stückweise stetig differenzierbare Extremale

Bei den bisherigen Betrachtungen, siehe im Speziellen die Sätze 4.2 bis 4.5, wurde stets angenommen, dass $\mathbf{x}(t)$ im Funktionenraum der im Intervall $[t_0, t_1]$ (vektorwertigen) stetig differenzierbaren Funktionen $\left(C^1[t_0,T]\right)^n$ 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 $(c_k, c_{k+1}),\ k = 0,\dots,N$ stetig differenzierbar ist, siehe Abbildung 4.2. Die inneren Punkte $c_1,\dots,c_N$ werden als *Eckpunkte von* $x(t)$ bezeichnet. Für stückweise stetig differenzierbare Funktionen $\hat{x}(t) \in \hat{C}^1[t_0,t_1]$ lauten die

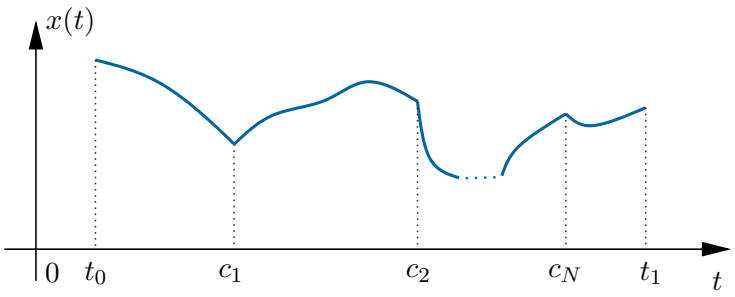


Abbildung 4.2: Beispiel einer Funktion $x(t) \in \hat{C}^1[t_0,t_1]$.

Normen gemäß (4.7)

$$\|\hat{\mathbf{x}}(t)\|_\infty := \max_{t_0\le t\le t_1} \|\hat{\mathbf{x}}(t)\| \quad \text{und} \quad \|\hat{\mathbf{x}}(t)\|_{1,\infty} := \max_{t_0\le t\le t_1} \|\hat{\mathbf{x}}(t)\| + \sup_{t\in\cup_{k=0}^{N}(c_k,c_{k+1})} \left\|\frac{\mathrm{d}}{\mathrm{d}t}\hat{\mathbf{x}}(t)\right\| . \tag{4.47}$$

Es gilt nun folgender Satz, welcher z. B. in [9] bewiesen wird.

Satz 4.6 (Stückweise stetig vs. stetig differenzierbare Extremale)**.** *Angenommen* $\mathbf{x}^* \in \mathcal{X}_{\mathrm{a}\lceil}$ *ist ein (lokales) Minimum des Funktionals*

$$J(\mathbf{x}) = \int_{t_0}^{t_1} l(t,\mathbf{x}(t),\dot{\mathbf{x}}(t))\,\mathrm{d}t \tag{4.48}$$

mit der zulässigen Menge $\mathcal{X}_{\mathrm{a}\lceil} = \{\mathbf{x}(t) \in \left(C^1[t_0,t_1]\right)^n \,|\, \mathbf{x}(t_0) = \mathbf{x}_0, \mathbf{x}(t_1) = \mathbf{x}_1\}$ *und der stetig differenzierbaren Lagrangeschen Dichte* $l : \mathbb{R}\times\mathbb{R}^n\times\mathbb{R}^n \to \mathbb{R}$*, dann ist*

$\mathbf{x}^* \in \hat{\mathcal{X}}_{\text{ad}}$ auch ein (lokales) Minimum des Funktionals (4.48) in der zulässigen Menge $\hat{\mathcal{X}}_{\text{ad}} = \{\mathbf{x}(t) \in (\hat{C}^1[t_0, t_1])^n \,|\, \mathbf{x}(t_0) = \mathbf{x}_0, \mathbf{x}(t_1) = \mathbf{x}_1\}$. Handelt es sich um ein lokales Minimum, so ist der Begriff lokal *bezüglich der gleichen Norm $\|\cdot\|_\infty$ bzw. $\|\cdot\|_{1,\infty}$ zu verstehen.*

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 (4.17) und die Legendre-Bedingung (4.33) erfüllt. Die Transversalitätsbedingungen (4.42), (4.43) und (4.46) bleiben im Falle stückweise stetig differenzierbarer Extremale *unverändert*. Die Unstetigkeiten von $\frac{\mathrm{d}}{\mathrm{d}t}\hat{\mathbf{x}}^*(t)$ an den Eckpunkten $t = c_k,\ k = 1, \ldots, N$ unterliegen nun folgenden Einschränkungen:

Satz 4.7 (Weierstrass-Erdmann Bedingungen)**.** *Angenommen $\hat{\mathbf{x}}^* \in \hat{\mathcal{X}}_{\text{ad}}$ ist ein (lokales) Minimum des Funktionals*

$$J(\hat{\mathbf{x}}) = \int_{t_0}^{t_1} l\Big(t, \hat{\mathbf{x}}(t), \dot{\hat{\mathbf{x}}}(t)\Big)\,\mathrm{d}t \tag{4.49}$$

mit der zulässigen Menge $\hat{\mathcal{X}}_{\text{ad}} = \Big\{\hat{\mathbf{x}}(t) \in (\hat{C}^1[t_0, t_1])^n \,|\, \hat{\mathbf{x}}(t_0) = \hat{\mathbf{x}}_0, \hat{\mathbf{x}}(t_1) = \hat{\mathbf{x}}_1\Big\}$, 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, $\hat{vecx}(t)$ und $\dot{\hat{vecx}}(t)$ sind. Dann gilt für jeden Eckpunkt $c \in (t_0, t_1)$ von $\hat{\mathbf{x}}^*(t)$, dass die Bedingungen

$$\left(\frac{\partial}{\partial \dot{\hat{\mathbf{x}}}} l\right)\Big(c, \hat{\mathbf{x}}^*(c), \dot{\hat{\mathbf{x}}}^*(c^-)\Big) = \left(\frac{\partial}{\partial \dot{\hat{\mathbf{x}}}} l\right)\Big(c, \hat{\mathbf{x}}^*(c), \dot{\hat{\mathbf{x}}}^*\Big(c^+\Big)\Big) \tag{4.50a}$$

$$H\Big(c, \hat{\mathbf{x}}^*(c), \dot{\hat{\mathbf{x}}}^*(c^-)\Big) = H\Big(c, \hat{\mathbf{x}}^*(c), \dot{\hat{\mathbf{x}}}^*\Big(c^+\Big)\Big) \tag{4.50b}$$

mit der Hamiltonfunktion

$$H\Big(t, \hat{\mathbf{x}}(t), \dot{\hat{\mathbf{x}}}(t)\Big) = \left(\frac{\partial}{\partial \dot{\hat{\mathbf{x}}}} l\right)\Big(t, \hat{\mathbf{x}}(t), \dot{\hat{\mathbf{x}}}(t)\Big)\dot{\hat{\mathbf{x}}}(t) - l\Big(t, \hat{\mathbf{x}}(t), \dot{\hat{\mathbf{x}}}(t)\Big) \tag{4.51}$$

erfüllt sind, wobei $\dot{\hat{\mathbf{x}}}^*(c^-)$ und $\dot{\hat{\mathbf{x}}}^*(c^+)$ den links- bzw. rechtsseitigen Grenzwert von $\dot{\hat{\mathbf{x}}}^*(t)$ an der Stelle $t = c$ bezeichnen.

Die Weierstrass-Erdmann Bedingungen besagen also, dass an den Eckpunkten einer (lokal) extremalen Trajektorie $\hat{\mathbf{x}}^*(t) \in (\hat{C}^1[t_0, t_1])^n$ nur jene Unstetigkeiten von $\dot{\hat{\mathbf{x}}}^*$ erlaubt sind, die die zeitliche Stetigkeit von $\frac{\partial}{\partial \dot{\hat{\mathbf{x}}}} l$ und die zeitliche Stetigkeit der Hamiltonfunktion H erhalten. Die Weierstrass-Erdmann Bedingungen sind *notwendige* Optimalitätsbedingungen. Ihre Herleitung findet sich z. B. in [8].

Beispiel 4.4. Gesucht ist ein (lokales) Minimum $x^* \in \mathcal{X}_{\text{a}\lceil}$ des Funktionals

$$J(x) = \int_{-1}^{1} x^2(t)(1 - \dot{x}(t))^2\,\mathrm{d}t \tag{4.52}$$

in der zulässigen Menge $\mathcal{X}_{a\lceil} = \{x(t) \in C^1[-1,1] \,|\, x(-1) = 0, x(1) = 1\}$. Da die Lagrangesche Dichte nicht explizit von der Zeit t abhängt, ist die Hamiltonfunktion

$$H = \left(\frac{\partial}{\partial \dot{x}} l\right)\dot{x} - l = -2x^2(1-\dot{x})\dot{x} - x^2(1-\dot{x})^2 = x^2\left(\dot{x}^2 - 1\right) = -k_1 \tag{4.53}$$

für alle Zeiten $t \in [-1,1]$ konstant mit der Konstanten k_1 und damit eine Invariante des Systems, siehe auch (4.23). Ersetzt man $x^2(t) = z(t)$ und $2x(t)\dot{x}(t) = \dot{z}(t)$ in (4.53), so ergibt sich

$$z(t) - \frac{1}{4}\dot{z}^2(t) = k_1 \ . \tag{4.54}$$

Die Lösung von (4.54) lautet

$$z(t) = (t + k_2)^2 + k_1 \tag{4.55}$$

mit der Konstanten k_2. Mit $x(-1) = 0$ und $x(1) = 1$ sowie $z(t) = x^2(t)$ folgen die Konstanten k_1 und k_2 zu $k_1 = -\left(\frac{3}{4}\right)^2$ und $k_2 = \frac{1}{4}$ und die mögliche stationäre Lösung $\bar{x}(t)$ des Kostenfunktionals (4.52) lautet

$$\bar{x}(t) = \pm\sqrt{\left(t + \frac{1}{4}\right)^2 - \left(\frac{3}{4}\right)^2} \ . \tag{4.56}$$

Die Wurzel liefert nur für $t \geq \frac{1}{2}$ und $t < -1$ ein reellwertiges Ergebnis, weshalb $\bar{x}(t)$ keine stationäre Lösung von (4.52) in der zulässigen Menge $\mathcal{X}_{a\lceil} = \{x(t) \in C^1[-1,1] \,|\, x(-1) = 0,\ x(1) = 1\}$ darstellt.

Im nächsten Schritt soll das Kostenfunktional (4.52) in der zulässigen Menge $\hat{\mathcal{X}}_{ad} = \{\hat{x}(t) \in \hat{C}^1[-1,1] \,|\, \hat{x}(-1) = 0,\ \hat{x}(1) = 1\}$ minimiert werden. Die Weierstrass-Erdmann Bedingung (4.50a) besagt nun, dass an einem Eckpunkt $c \in (-1,1)$ gilt

$$-2\hat{x}^2(c)\left[1 - \dot{\hat{x}}(c^-)\right] = -2\hat{x}^2(c)\left[1 - \dot{\hat{x}}\left(c^+\right)\right] \tag{4.57}$$

und folglich

$$\hat{x}^2(c)\left[\dot{\hat{x}}\left(c^+\right) - \dot{\hat{x}}(c^-)\right] = 0 \ . \tag{4.58}$$

Da an einem Eckpunkt $t = c$ gilt $\dot{\hat{x}}(c^+) \neq \dot{\hat{x}}(c^-)$, muss zur Erfüllung von (4.58) die Bedingung $\hat{x}(c) = 0$ eingehalten werden. D. h. Eine Unstetigkeit in $\dot{\hat{x}}(t)$ kann nur an Stellen auftreten, an denen der Wert von $\hat{x}(t)$ selbst identisch Null ist. Den minimalen Wert des Kostenfunktionals (4.52), nämlich den Wert Null, erhält man, wenn $\hat{x}(t) = 0$ oder $\dot{\hat{x}}(t) = 1$ für alle t in $[-1,1]$ gilt. Außerdem sind die Randbedingungen $\hat{x}(-1) = 0$ und $\hat{x}(1) = 1$ zu erfüllen. Aus diesen Überlegungen folgt, dass für die optimale Lösung $\hat{x}(t) = 0\ \forall t \in [-1,c]$ und $\dot{\hat{x}}(t) = 1\ \forall t \in (c,1]$ gelten muss. Daraus ergibt sich der Umschaltpunkt $c = 0$ und die optimale Lösung

$$\hat{x}^*(t) = \begin{cases} 0 & \text{für } -1 \leq t \leq 0 \\ t & \text{für } 0 < t \leq 1. \end{cases} \tag{4.59}$$

Diese Lösung ist das eindeutige globale Minimum des Kostenfunktionals (4.52) in der zulässigen Menge $\hat{\mathcal{X}}_{\text{ad}} = \{\hat{x}(t) \in \hat{C}^1[-1,1] \,|\, \hat{x}(-1) = 0,\ \hat{x}(1) = 1\}$.

4.2 Entwurf von Optimalsteuerungen

4.2.1 Problemformulierung

Eine typische Optimalsteuerungsaufgabe besteht darin, für ein dynamisches System beschrieben durch die Differenzialgleichungen

$$\dot{\mathbf{x}} = \mathbf{f}(t, \mathbf{x}(t), \mathbf{u}(t)) \tag{4.60a}$$

mit der Zeit $t \in \mathbb{R}$, dem *Zustand* $\mathbf{x} \in \mathbb{R}^n$, dem *Anfangszustand*

$$\mathbf{x}(t_0) = \mathbf{x}_0 \tag{4.60b}$$

und dem *Stelleingang* $\mathbf{u} \in \mathbb{R}^m$ eine geeignete *Steuertrajektorie* $\mathbf{u}(t), t \in [t_0, t_1]$ so zu finden, dass ein *Kostenfunktional* der Form (siehe auch (4.2))

$$J(\mathbf{u}) = \varphi(t_0, \mathbf{x}(t_0), t_1, \mathbf{x}(t_1)) + \int_{t_0}^{t_1} l(t, \mathbf{x}(t), \mathbf{u}(t))\,\mathrm{d}t \tag{4.61}$$

bezüglich $\mathbf{u}(t)$ minimiert wird und dabei allfällige Beschränkungen für $\mathbf{x}(t)$ und $\mathbf{u}(t)$ eingehalten werden. Beim Kostenfunktional unterscheidet man im Allgemeinen zwischen der *Bolza-Form* (4.61), der *Lagrange-Form*

$$J(\mathbf{u}) = \int_{t_0}^{t_1} l(t, \mathbf{x}(t), \mathbf{u}(t))\,\mathrm{d}t \tag{4.62}$$

und der *Mayer-Form*

$$J(\mathbf{u}) = \varphi(t_0, \mathbf{x}(t_0), t_1, \mathbf{x}(t_1))\ . \tag{4.63}$$

Die Abhängigkeit der Funktion φ von t_0 und $\mathbf{x}(t_0)$ ist natürlich nur relevant, wenn die Anfangsbedingung (4.60b) nicht vorhanden ist.

Aufgabe 4.2. Zeigen Sie, dass die Lagrange-Form in die Mayer-Form übergeführt werden kann, indem man einen zusätzlichen Zustand

$$\dot{x}_{n+1} = l(t, \mathbf{x}, \mathbf{u}), \quad x_{n+1}(t_0) = 0 \tag{4.64}$$

einführt und das Kostenfunktional in der Form $J(\mathbf{u}) = x_{n+1}(t_1)$ anschreibt.

Zeigen Sie, dass die Mayer-Form in die Lagrange-Form übergeführt werden kann, indem man einen zusätzlichen Zustand

$$\dot{x}_{n+1} = 0, \quad x_{n+1}(t_0) = \frac{1}{t_1 - t_0}\varphi(t_0, \mathbf{x}(t_0), t_1, \mathbf{x}(t_1)) \tag{4.65}$$

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 zwischen *Punktbeschränkungen*, beispielsweise *Endpunktbeschränkungen* der Form

$$\psi(t_1, \mathbf{x}(t_1)) \leq 0, \tag{4.66}$$

Pfadbeschränkungen

$$\psi(t, \mathbf{x}(t), \mathbf{u}(t)) \leq 0, \quad \forall t \in I \subset [t_0, t_1], \tag{4.67}$$

und *isoperimetrischen Beschränkungen*

$$\int_{t_0}^{t_1} \psi(t, \mathbf{x}(t), \mathbf{u}(t))\, \mathrm{d}t \leq C\ . \tag{4.68}$$

Häufig sind Pfadbeschränkungen (4.67) schwieriger zu berücksichtigen, wenn sie nicht von der Stellgröße abhängen.

4.2.2 Existenz einer Lösung

Im Skriptum Regelungssysteme 2 (Satz 2.13) wurden hinreichende Bedingungen für die lokale Eindeutigkeit und Existenz der Lösung eines *Anfangswertproblems* angegeben. Sie besagen, wenn $\mathbf{g}(t, \mathbf{x})$ stückweise stetig in t ist und der Lipschitz-Bedingung

$$\|\mathbf{g}(t, \mathbf{x}) - \mathbf{g}(t, \mathbf{y})\| \leq L\|\mathbf{x} - \mathbf{y}\|, \quad 0 < L < \infty \tag{4.69}$$

für alle $\mathbf{x}, \mathbf{y} \in B_\gamma = \{\mathbf{x} \in \mathbb{R}^n \mid \|\mathbf{x} - \mathbf{x}_0\| \leq \gamma\}$ und alle $t \in [t_0, t_0 + \tau]$ genügt, dann existiert ein $\delta \in (0, \tau]$ so, dass das Anfangswertproblem

$$\dot{\mathbf{x}} = \mathbf{g}(t, \mathbf{x}), \quad \mathbf{x}(t_0) = \mathbf{x}_0 \tag{4.70}$$

für $t \in [t_0, t_0 + \delta]$ *genau eine Lösung* besitzt. Da hier nur stückweise Stetigkeit von $\mathbf{g}(t, \mathbf{x})$ in t gefordert wird, sind entsprechend (4.60) mit $\mathbf{g}(t, \mathbf{x}) := \mathbf{f}(t, \mathbf{x}(t), \mathbf{u}(t))$ für die Stellgrößen $\mathbf{u}(t)$ auch *stückweise stetige Funktionen* zugelassen, d. h. $\mathbf{u}(t) \in (\hat{C}[t_0, t_1])^m$. Man nennt eine reellwertige Funktion $u(t) \in \hat{C}[t_0, t_1]$ *stückweise stetig*, wenn eine *Partitionierung* $t_0 = c_0 < c_1 < \ldots < c_{N+1} = t_1$ 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 (4.60) stückweise stetig differenzierbar, d. h. $\mathbf{x}(t) \in (\hat{C}^1[t_0, t_1])^n$, wobei die Eckpunkte mit den Unstetigkeitsstellen der Stellgrößen übereinstimmen. Zur Erinnerung sei angemerkt, dass gemäß Satz 2.14 des Skriptums Regelungssysteme 2 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 ist, dass $\mathbf{g}(t, \mathbf{x})$ die Lipschitz-Bedingung (4.69) lokal erfüllt.

Bei den meisten praktischen Anwendungen unterliegen die Stellgrößen gewissen Beschränkungen, d. h. $\mathbf{u}(t) \in U \subset \mathbb{R}^m$. Häufig auftretende Beschränkungen sind z. B. sogenannte *box constraints*

$$u_i^- \leq u_i \leq u_i^+, \quad i = 1, \ldots, m\ . \tag{4.71}$$

Eine stückweise stetige Stellgröße $\mathbf{u}(t)$ im Intervall $t_0 \leq t \leq t_1$ mit $\mathbf{u}(t) \in U$ für alle $t \in [t_0, t_1]$ bezeichnet man im Weiteren als *zulässige Stellgröße*. Für das Folgende sei angenommen, dass $\bar{\mathbf{x}}(t; \mathbf{x}_0, \mathbf{u}(t))$ die Lösung von (4.60) zum Zeitpunkt t für den Anfangswert $\mathbf{x}(t_0) = \mathbf{x}_0$ und die Stellgröße $\mathbf{u}(\tau),\ t_0 \leq \tau \leq t$ bezeichnet. Dann nennt man eine zulässige Stellgröße *realisierbar*, wenn $\bar{\mathbf{x}}(t; \mathbf{x}_0, \mathbf{u}(t))$ im gesamten Intervall $t_0 \leq t \leq t_1$ definiert ist und sämtliche Beschränkungen einhält.

Das Problem bei der Existenz einer Lösung des *Optimalsteuerungsproblems* besteht häufig darin, dass die Menge der realisierbaren Lösungen *nicht kompakt* ist. Es könnte beispielsweise passieren, dass die Lösung von (4.60) innerhalb des Optimierungsintervalls $[t_0, t_1]$ nach Unendlich strebt und in Folge das Kostenfunktional unendlich wird. Dieses Phänomen ist auch unter dem Namen *finite escape time* bekannt. Um derartige Fälle auszuschließen, fordert man oft, dass die Lösungen des dynamischen Systems (4.60) beschränkt sind, also dass gilt

$$\|\bar{\mathbf{x}}(t; \mathbf{x}_0, \mathbf{u}(t))\| \leq \alpha, \quad \forall t \in [t_0, t_1] \tag{4.72}$$

für ein finites $\alpha > 0$. Man beachte, dass bezüglich $\mathbf{x}$ affine Systeme der Form

$$\dot{\mathbf{x}} = \mathbf{A}(t, \mathbf{u})\mathbf{x} + \mathbf{b}(t, \mathbf{u}), \quad \mathbf{x}(t_0) = \mathbf{x}_0 \tag{4.73}$$

diese Eigenschaft stets erfüllen und in endlicher Zeit nicht nach Unendlich streben können.

Wenn das Optimierungsintervall selbst unendlich ist, dann ist die Menge der realisierbaren Stellgrößen unbeschränkt und nicht kompakt, weshalb stets ein kompaktes (finites) Zeitintervall $[t_0, T]$ mit hinreichend großem $T > t_1$ gewählt werden sollte. Dies wird anhand des nachfolgenden Beispiels näher erläutert.

Beispiel 4.5. Gegeben ist eine Punktmasse m, die über eine Kraft $u(t)$ mit $0 \leq u(t) \leq 1$ für alle t im Optimierungsintervall $[t_0, t_1]$ beschleunigt wird. Die Aufgabe besteht nun darin, die Stellgröße $u(t)$ so zu bestimmen, dass ausgehend von der Anfangsposition $x(t_0) = x_0$ die Masse nach der Zeit $t = t_1$ die Position $x(t_1) = x_1$ erreicht und dabei das Kostenfunktional

$$J(u) = \int_{t_0}^{t_1} u^2(t)\,\mathrm{d}t \tag{4.74}$$

minimiert. Man erkennt unmittelbar, dass für $x_1 > x_0$ die Stellgröße $u(t) \equiv 0$ nicht realisierbar ist und somit für jede realisierbare Stellgröße $J(u) > 0$ gelten muss. Betrachtet man nun die Folge der realisierbaren konstanten Stellgrößen $u_k(t) = \frac{1}{k}$, $k \geq 1$ für alle $t \geq t_0$, so erhält man als Lösung des Differentialgleichungssystems

$$\dot{x}_k = v_k, \quad x_k(t_0) = x_0 \tag{4.75a}$$
$$m\dot{v}_k = u_k, \quad v_k(t_0) = 0 \tag{4.75b}$$

das Ergebnis

$$x_k(t) = x_0 + \frac{1}{2mk}(t - t_0)^2 \tag{4.76a}$$
$$v_k(t) = \frac{1}{mk}(t - t_0)\ . \tag{4.76b}$$

Die Zeit t_1, nach der der Zustand $x_k(t_1) = x_1$ erreicht wird, errechnet sich direkt aus (4.76a) zu

$$t_{1,k} = t_0 + \sqrt{2mk}\sqrt{x_1 - x_0} \; . \tag{4.77}$$

Damit erhält man für $u_k(t) = \frac{1}{k}$ im Optimierungsintervall $[t_0, t_{1,k}]$ den Wert des Kostenfunktionals zu

$$J(u_k) = \int_{t_0}^{t_{1,k}} \frac{1}{k^2} \, \mathrm{d}t = \sqrt{\frac{2m}{k^3}}\sqrt{x_1 - x_0} \; . \tag{4.78}$$

Es gilt nun $J(u_k) \to 0$ für $k \to \infty$ und damit $t_{1,k} \to \infty$. Man erkennt also, dass gilt $\inf J(u_k) = 0$, d. h. das Problem hat *kein Minimum*.

Damit die Existenz einer Lösung des *Optimalsteuerungsproblems* auch tatsächlich gewährleistet ist, müssen weitere Einschränkungen der zulässigen Steuerungen vorgenommen werden. Zwei Möglichkeiten sollen im Folgenden kurz aufgezeigt werden. Einerseits besteht die Möglichkeit, zu fordern, dass die Stellgröße der *zusätzlichen Lipschitz-Bedingung*

$$\|\mathbf{u}(t) - \mathbf{u}(s)\| \leq L_u |t - s|, \quad 0 < L_u < \infty, \quad \forall s, t \in [t_0, t_1] \tag{4.79}$$

genügt und andererseits kann die Klasse der zulässigen Stellgrößen auf die *stückweise konstanten Stellgrößen* mit einer finiten Anzahl an Unstetigkeitsstellen eingeschränkt werden.

4.2.3 Variationsformulierung

Im Folgenden werden die notwendigen Bedingungen erster Ordnung für ein Optimalsteuerungsproblem mit fester Endzeit und freiem Endwert formuliert und hergeleitet.

Satz 4.8 (Steuerungsproblem: Endzeit fest/Endwert frei)**.** *Gesucht ist die Stellgröße* $\mathbf{u} \in (C[t_0, t_1])^m$ *so, dass das Kostenfunktional (Lagrange-Form)*

$$J(\mathbf{u}) = \int_{t_0}^{t_1} l(t, \mathbf{x}(t), \mathbf{u}(t)) \, \mathrm{d}t \tag{4.80}$$

unter der Gleichungsbeschränkung (dynamisches System)

$$\dot{\mathbf{x}} = \mathbf{f}(t, \mathbf{x}(t), \mathbf{u}(t)), \quad \mathbf{x}(t_0) = \mathbf{x}_0 \tag{4.81}$$

mit fester Anfangszeit t_0 *und fester Endzeit* $t_1 > t_0$ *minimiert wird. Dabei wird angenommen, dass* l *und* $\mathbf{f}$ *stetig in* t *und stetig differenzierbar bezüglich* $\mathbf{x}$ *und* $\mathbf{u}$ *für alle* $(t, \mathbf{x}, \mathbf{u}) \in [t_0, t_1] \times \mathbb{R}^n \times \mathbb{R}^m$ *sind. 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 (4.81) *ist, dann existiert ein* $\boldsymbol{\lambda}^*(t) \in \left(C^1[t_0, t_1]\right)^n$ *so, dass gilt*

$$\dot{\mathbf{x}}^* = \mathbf{f}(t, \mathbf{x}^*(t), \mathbf{u}^*(t)), \quad \mathbf{x}^*(t_0) = \mathbf{x}_0 \tag{4.82a}$$

$$\dot{\boldsymbol{\lambda}}^* = -\left(\frac{\partial}{\partial \mathbf{x}} l\right)^{\mathrm{T}}(t, \mathbf{x}^*(t), \mathbf{u}^*(t)) - \left(\frac{\partial}{\partial \mathbf{x}} \mathbf{f}\right)^{\mathrm{T}}(t, \mathbf{x}^*(t), \mathbf{u}^*(t))\boldsymbol{\lambda}^*(t) \tag{4.82b}$$

$$\boldsymbol{\lambda}^*(t_1) = \mathbf{0}$$

$$\mathbf{0} = \left(\frac{\partial}{\partial \mathbf{u}} l\right)^{\mathrm{T}}(t, \mathbf{x}^*(t), \mathbf{u}^*(t)) + \left(\frac{\partial}{\partial \mathbf{u}} \mathbf{f}\right)^{\mathrm{T}}(t, \mathbf{x}^*(t), \mathbf{u}^*(t))\boldsymbol{\lambda}^*(t) \tag{4.82c}$$

für $t_0 \leq t \leq t_1$. *Die Gleichungen* (4.82) *werden als die* Euler-Lagrange Gleichungen des Optimalsteuerungsproblems *und* $\boldsymbol{\lambda}^*(t)$ *als der* adjungierte Zustand *oder der* Kozustand *bezeichnet.*

Beweis. Man betrachte dazu die einparametrische Familie der zulässigen Stellgrößen $\mathbf{v}(t;\eta) = \mathbf{u}^*(t) + \eta\boldsymbol{\xi}_u(t)$ mit $\boldsymbol{\xi}_u(t) \in (C[t_0, t_1])^m$ und dem skalaren Parameter η. Aufgrund der Stetigkeits- und Differenzierbarkeitsannahmen für $\mathbf{f}$ existiert ein $\eta \neq 0$ so, dass die zu $\mathbf{v}(t)$ zugehörige Lösung $\mathbf{y}(t;\eta)$ des Anfangswertproblems (4.81) für alle $t \in [t_0, t_1]$ eindeutig und bezüglich η differenzierbar ist. Für $\eta = 0$ gilt offensichtlich $\mathbf{y}(t;0) = \mathbf{x}^*(t),\ t_0 \leq t \leq t_1$. Das um die Gleichungsbeschränkungen (4.81) erweiterte Kostenfunktional (4.80) für $\mathbf{v}(t;\eta)$ lautet

$$\begin{aligned}\bar{J}(\mathbf{v}(t;\eta)) &= \int_{t_0}^{t_1} l(t, \mathbf{y}(t;\eta), \mathbf{v}(t;\eta)) + \boldsymbol{\lambda}^{\mathrm{T}}(t)(\mathbf{f}(t, \mathbf{y}(t;\eta), \mathbf{v}(t;\eta)) - \dot{\mathbf{y}})\,\mathrm{d}t \\ &= \int_{t_0}^{t_1} l(t, \mathbf{y}(t;\eta), \mathbf{v}(t;\eta)) + \dot{\boldsymbol{\lambda}}^{\mathrm{T}}\mathbf{y}(t;\eta) + \boldsymbol{\lambda}^{\mathrm{T}}(t)\mathbf{f}(t, \mathbf{y}(t;\eta), \mathbf{v}(t;\eta))\,\mathrm{d}t - \left[\boldsymbol{\lambda}^{\mathrm{T}}\mathbf{y}\right]_{t_0}^{t_1}\end{aligned} \tag{4.83}$$

für jedes $\boldsymbol{\lambda}(t) \in \left(C^1[t_0, t_1]\right)^n$. Gleichung (4.83) kann auch als Lagrangefunktion mit den Lagrangemultiplikatoren $\boldsymbol{\lambda}(t)$ interpretiert werden. Gemäß Satz 4.1 lautet die notwendige Bedingung für ein Minimum

$$\begin{aligned}\delta\bar{J}(\mathbf{u}^*;\boldsymbol{\xi}_u) &= \frac{\mathrm{d}}{\mathrm{d}\eta}\bar{J}(\mathbf{v}(t;\eta))\bigg|_{\eta=0} = 0 \\ &= \int_{t_0}^{t_1}\left(\frac{\partial}{\partial \mathbf{x}} l\right)(t, \mathbf{x}^*, \mathbf{u}^*)\boldsymbol{\xi}_y + \dot{\boldsymbol{\lambda}}^{\mathrm{T}}(t)\boldsymbol{\xi}_y + \boldsymbol{\lambda}^{\mathrm{T}}(t)\left(\frac{\partial}{\partial \mathbf{x}}\mathbf{f}\right)(t, \mathbf{x}^*, \mathbf{u}^*)\boldsymbol{\xi}_y\,\mathrm{d}t \\ &\quad + \int_{t_0}^{t_1}\left(\frac{\partial}{\partial \mathbf{u}} l\right)(t, \mathbf{x}^*, \mathbf{u}^*)\boldsymbol{\xi}_u + \boldsymbol{\lambda}^{\mathrm{T}}(t)\left(\frac{\partial}{\partial \mathbf{u}}\mathbf{f}\right)(t, \mathbf{x}^*, \mathbf{u}^*)\boldsymbol{\xi}_u\,\mathrm{d}t - \left[\boldsymbol{\lambda}^{\mathrm{T}}(t)\boldsymbol{\xi}_y(t)\right]_{t_0}^{t_1}\end{aligned} \tag{4.84}$$

mit

$$\boldsymbol{\xi}_y(t) = \left(\frac{\mathrm{d}}{\mathrm{d}\eta}\mathbf{y}\right)(t;0)\ . \tag{4.85}$$

Nachdem der Anfangszustand $\mathbf{x}(t_0) = \mathbf{y}(t_0;\eta) = \mathbf{x}_0$ festgelegt ist, gilt $\boldsymbol{\xi}_y(t_0) = \mathbf{0}$. Da die Auswirkung von $\boldsymbol{\xi}_u(\tau)$ für $\tau \in [t_0, t]$ auf den Wert $\boldsymbol{\xi}_y(t)$ nur aufwendig zu

berechnen ist, wählt man $\boldsymbol{\lambda}(t) = \boldsymbol{\lambda}^*(t)$ so, dass

$$\dot{\boldsymbol{\lambda}}^* = -\left(\frac{\partial}{\partial \mathbf{x}}\mathbf{f}\right)^{\mathrm{T}}(t, \mathbf{x}^*(t), \mathbf{u}^*(t))\boldsymbol{\lambda}^*(t) - \left(\frac{\partial}{\partial \mathbf{x}}l\right)^{\mathrm{T}}(t, \mathbf{x}^*(t), \mathbf{u}^*(t)) \tag{4.86}$$

mit der Endbedingung $\boldsymbol{\lambda}^*(t_1) = \mathbf{0}$ gilt. Diese Endbedingung folgt aus dem letzten Term von (4.84), da $\mathbf{x}(t_1)$ frei und $\boldsymbol{\xi}_y(t_1)$ somit im Allgemeinen von Null verschieden ist. Die adjungierte Differentialgleichung (4.86) mit der Endbedingung $\boldsymbol{\lambda}^*(t_1) = \mathbf{0}$ ist *linear*. Aufgrund der Differenzierbarkeitsannahmen für l und $\mathbf{f}$ existiert die Lösung $\boldsymbol{\lambda}^*(t)$ und ist im Intervall $[t_0, t_1]$ eindeutig. Damit verbleibt in (4.84) der Ausdruck

$$\int_{t_0}^{t_1} \boldsymbol{\xi}_u^{\mathrm{T}} \left[\left(\frac{\partial}{\partial \mathbf{u}}l\right)^{\mathrm{T}}(t, \mathbf{x}^*(t), \mathbf{u}^*(t)) + \left(\frac{\partial}{\partial \mathbf{u}}\mathbf{f}\right)^{\mathrm{T}}(t, \mathbf{x}^*(t), \mathbf{u}^*(t))\boldsymbol{\lambda}^*(t)\right] \mathrm{d}t = 0, \tag{4.87}$$

welcher aufgrund des Fundamentallemmas der Variationsrechnung Lemma 4.2 die Bedingung

$$\left(\frac{\partial}{\partial \mathbf{u}}l\right)^{\mathrm{T}}(t, \mathbf{x}^*(t), \mathbf{u}^*(t)) + \left(\frac{\partial}{\partial \mathbf{u}}\mathbf{f}\right)^{\mathrm{T}}(t, \mathbf{x}^*(t), \mathbf{u}^*(t))\boldsymbol{\lambda}^*(t) = \mathbf{0} \tag{4.88}$$

für alle $t \in [t_0, t_1]$ impliziert. □

Wie man aus (4.82) erkennen kann, setzen sich die notwendigen Optimalitätsbedingungen für das Optimalsteuerungsproblem (4.80) und (4.81) aus $2n$ *Differentialgleichungen* in $\mathbf{x}^*$ und $\boldsymbol{\lambda}^*$ und m *algebraischen Gleichungen* zusammen. Da für die Differentialgleichung in $\mathbf{x}^*$ der Wert zum Anfangszeitpunkt $t = t_0$ und für die Differentialgleichung in $\boldsymbol{\lambda}^*$ der Wert zum Endzeitpunkt $t = t_1$ gegeben ist, handelt es sich um ein *Zweipunktrandwertproblem*. Analog zu den Lagrange-Multiplikatoren von Abschnitt 3.1.2 lässt sich der adjungierte Zustand $\boldsymbol{\lambda}(t)$ in der Form interpretieren, dass $\boldsymbol{\lambda}(t_0)$ der *Sensitivität des Kostenfunktionals* (4.80) bezüglich einer Änderung des Anfangswertes $\mathbf{x}_0$ entspricht.

Die Euler-Lagrange Gleichungen (4.82) lassen sich mit Hilfe der *Hamiltonfunktion*

$$H(t, \mathbf{x}, \mathbf{u}, \boldsymbol{\lambda}) = l(t, \mathbf{x}, \mathbf{u}) + \boldsymbol{\lambda}^{\mathrm{T}}(t)\mathbf{f}(t, \mathbf{x}, \mathbf{u}) \tag{4.89}$$

auch in der Form

$$\dot{\mathbf{x}}^* = \left(\frac{\partial}{\partial \boldsymbol{\lambda}}H\right)^{\mathrm{T}}(t, \mathbf{x}^*(t), \mathbf{u}^*(t), \boldsymbol{\lambda}^*(t)), \qquad \mathbf{x}^*(t_0) = \mathbf{x}_0 \tag{4.90a}$$

$$\dot{\boldsymbol{\lambda}}^* = -\left(\frac{\partial}{\partial \mathbf{x}}H\right)^{\mathrm{T}}(t, \mathbf{x}^*(t), \mathbf{u}^*(t), \boldsymbol{\lambda}^*(t)), \qquad \boldsymbol{\lambda}^*(t_1) = \mathbf{0} \tag{4.90b}$$

$$\mathbf{0} = \left(\frac{\partial}{\partial \mathbf{u}}H\right)^{\mathrm{T}}(t, \mathbf{x}^*(t), \mathbf{u}^*(t), \boldsymbol{\lambda}^*(t)) \tag{4.90c}$$

für $t_0 \leq t \leq t_1$ anschreiben. Man beachte, dass sich die hier in der dynamischen Optimierung verwendete Hamiltonfunktion H im Vorzeichen von jener der Variationsrechnung (siehe (4.22)) unterscheidet. Die Bedingung (4.90c) zeigt, dass $\mathbf{u}^*$ ein *stationärer Punkt* der Hamiltonfunktion H sein muss. Leitet man die Hamiltonfunktion entlang der optimalen

Lösung $(\mathbf{x}^*(t), \mathbf{u}^*(t), \boldsymbol{\lambda}^*(t))$ nach der Zeit ab, so erhält man

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}t}H &= \frac{\partial}{\partial t}H + \left(\frac{\partial}{\partial \mathbf{x}}H\right)\dot{\mathbf{x}}^* + \underbrace{\left(\frac{\partial}{\partial \mathbf{u}}H\right)}_{=\mathbf{0}}\dot{\mathbf{u}}^* + \left(\frac{\partial}{\partial \boldsymbol{\lambda}}H\right)\dot{\boldsymbol{\lambda}}^* \\ &= \frac{\partial}{\partial t}H - \left(\dot{\boldsymbol{\lambda}}^*\right)^{\mathrm{T}}\mathbf{f} + (\dot{\mathbf{x}}^*)^{\mathrm{T}}\dot{\boldsymbol{\lambda}}^* = \frac{\partial}{\partial t}H \ .\end{aligned} \tag{4.91}$$

Wenn daher weder $\mathbf{f}$ noch l explizit von der Zeit t abhängen, ist die Hamiltonfunktion H eine *Invariante* des Zweipunktrandwertproblems (4.90). Im Weiteren muss ähnlich zur Legendre Bedingung gemäß Satz 4.3 für ein Minimum des Kostenfunktionals $\bar{J}$ die *notwendige Bedingung zweiter Ordnung*

$$\mathbf{d}^{\mathrm{T}}\left(\frac{\partial^2}{\partial \mathbf{u}^2}H\right)(t, \mathbf{x}^*(t), \mathbf{u}^*(t), \boldsymbol{\lambda}^*(t))\mathbf{d} \geq 0 \qquad \forall\, \mathbf{d} \in \mathbb{R}^m,\ t \in [t_0, t_1] \ . \tag{4.92}$$

erfüllt sein. Sie wird auch *Legendre-Clebsch Bedingung* genannt.

In Satz 4.8 wurde angenommen, dass die optimale Stellgröße $\mathbf{u}^*$ stetig ist, d. h. $\mathbf{u}^*(t) \in (C[t_0, t_1])^m$. Für manche Beispiele findet man keine Lösung der Euler-Lagrange Gleichungen (4.82) in der Klasse der stetigen Stellgrößen. Aus diesem Grund sucht man Extremale in der erweiterten Klasse der stückweise stetigen Stellgrößen $(\hat{C}[t_0, t_1])^m$. Wie bereits im Abschnitt 4.2.2 diskutiert, sind für stückweise stetige Stellgrößen $\mathbf{u}(t)$ die zugehörigen Zustandsgrößen $\mathbf{x}(t)$ von (4.81) stückweise stetig differenzierbar, d. h. $\mathbf{x}(t) \in (\hat{C}^1[t_0, t_1])^n$ wobei die Eckpunkte mit den Unstetigkeitsstellen der Stellgrößen übereinstimmen. Bezeichnet man mit $\hat{\mathbf{u}}^*(t) \in (\hat{C}[t_0, t_1])^m$ die optimale Stellgröße und mit $\hat{\mathbf{x}}^*(t)$ und $\hat{\boldsymbol{\lambda}}^*(t)$ den zugehörigen Zustand und den adjungierten Zustand des Optimalsteuerungsproblems (4.80), (4.81), dann gelten für jeden Eckpunkt $c \in (t_0, t_1)$ die Bedingungen

$$\hat{\mathbf{x}}^*(c^-) = \hat{\mathbf{x}}^*\left(c^+\right) \tag{4.93a}$$

$$\hat{\boldsymbol{\lambda}}^*(c^-) = \hat{\boldsymbol{\lambda}}^*\left(c^+\right) \tag{4.93b}$$

$$H\left(c^-, \hat{\mathbf{x}}^*(c), \hat{\mathbf{u}}^*(c^-), \hat{\boldsymbol{\lambda}}^*(c)\right) = H\left(c^+, \hat{\mathbf{x}}^*(c), \hat{\mathbf{u}}^*\left(c^+\right), \hat{\boldsymbol{\lambda}}^*(c)\right) \ , \tag{4.93c}$$

wobei c^- bzw. c^+ den jeweiligen links- bzw. rechtsseitigen Grenzwert angeben. Man beachte, dass (4.93b) und (4.93c) 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 und hergeleitet.

Satz 4.9 (Steuerungsproblem: Endzeit frei/Endbeschränkung)**.** *Gesucht ist die Stellgröße* $\mathbf{u} \in (C[t_0, t_1])^m$ *so, dass das Kostenfunktional (Bolza-Form)*

$$J(\mathbf{u}, t_1) = \varphi(t_1, \mathbf{x}(t_1)) + \int_{t_0}^{t_1} l(t, \mathbf{x}(t), \mathbf{u}(t))\,\mathrm{d}t \tag{4.94}$$

unter den Gleichungsbeschränkungen

$$\dot{\mathbf{x}} - \mathbf{f}(t, \mathbf{x}(t), \mathbf{u}(t)) = \mathbf{0}, \quad \mathbf{x}(t_0) = \mathbf{x}_0 \tag{4.95a}$$

$$G_k(\mathbf{u}, t_1) = \psi_k(t_1, \mathbf{x}(t_1)) = 0, \quad k = 1, \ldots, p \tag{4.95b}$$

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 sowie die Funktionen φ und ψ_k, $k = 1, \ldots, p$ 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 $(\mathbf{u}^*, t_1^*) \in (C[t_0, t_1])^m \times [t_0, T)$ die optimale Lösung des Optimierungsproblems und $\mathbf{x}^* \in (C^1[t_0, T])^n$ die zugehörige Lösung des Anfangswertproblems (4.95a). Darüber hinaus wird angenommen, dass für p unabhängige zulässige Richtungen $(\boldsymbol{\xi}_k, \tau_k) \in (C[t_0, t_1])^m \times [t_0, T)$, $k = 1, \ldots, p$ die Regularitätsbedingung

$$\det\left(\begin{bmatrix} \delta G_1(\mathbf{u}^*, t_1^*; \boldsymbol{\xi}_1, \tau_1) & \cdots & \delta G_1(\mathbf{u}^*, t_1^*; \boldsymbol{\xi}_p, \tau_p) \\ \vdots & \ddots & \vdots \\ \delta G_p(\mathbf{u}^*, t_1^*; \boldsymbol{\xi}_1, \tau_1) & \cdots & \delta G_p(\mathbf{u}^*, t_1^*; \boldsymbol{\xi}_p, \tau_p) \end{bmatrix}\right) \neq 0 \tag{4.96}$$

gilt. Dann existiert ein $\boldsymbol{\lambda}^* \in (C^1[t_0, t_1^*])^n$ und ein $\boldsymbol{\mu}^* \in \mathbb{R}^p$ so, dass die Beziehungen

$$\dot{\mathbf{x}}^* = \left(\frac{\partial}{\partial \boldsymbol{\lambda}} H\right)^{\mathrm{T}}(t, \mathbf{x}^*(t), \mathbf{u}^*(t), \boldsymbol{\lambda}^*(t)), \qquad \mathbf{x}^*(t_0) = \mathbf{x}_0 \tag{4.97a}$$

$$\dot{\boldsymbol{\lambda}}^* = -\left(\frac{\partial}{\partial \mathbf{x}} H\right)^{\mathrm{T}}(t, \mathbf{x}^*(t), \mathbf{u}^*(t), \boldsymbol{\lambda}^*(t)), \quad \boldsymbol{\lambda}^*(t_1^*) = \left(\frac{\partial}{\partial \mathbf{x}_1} \Phi\right)^{\mathrm{T}}(t_1^*, \mathbf{x}^*(t_1^*), \boldsymbol{\mu}^*) \tag{4.97b}$$

$$\mathbf{0} = \left(\frac{\partial}{\partial \mathbf{u}} H\right)^{\mathrm{T}}(t, \mathbf{x}^*(t), \mathbf{u}^*(t), \boldsymbol{\lambda}^*(t)) \tag{4.97c}$$

für $t_0 \leq t \leq t_1$ mit den Transversalitätsbedingungen

$$\boldsymbol{\psi}(t_1^*, \mathbf{x}^*(t_1^*)) = \mathbf{0} \tag{4.98a}$$

$$\left(\frac{\partial}{\partial t_1} \Phi\right)(t_1^*, \mathbf{x}^*(t_1^*), \boldsymbol{\mu}^*) + H(t_1^*, \mathbf{x}^*(t_1^*), \mathbf{u}^*(t_1^*), \boldsymbol{\lambda}^*(t_1^*)) = 0\ , \tag{4.98b}$$

und der Hamiltonfunktion $H = l + \boldsymbol{\lambda}^{\mathrm{T}} \mathbf{f}$ *sowie* $\Phi = \varphi + \boldsymbol{\mu}^{\mathrm{T}} \boldsymbol{\psi}$ *mit* $\boldsymbol{\psi}^{\mathrm{T}} = \begin{bmatrix} \psi_1 & \psi_2 & \ldots & \psi_p \end{bmatrix}$ *erfüllt sind.*

Beweis. Man betrachte dazu wieder die einparametrische Familie der zulässigen Stellgrößen $\mathbf{v}(t; \eta) = \mathbf{u}^*(t) + \eta \boldsymbol{\xi}_u(t)$ mit $\boldsymbol{\xi}_u \in (C[t_0, t_1])^m$ und dem skalaren Parameter η. Aufgrund der Stetigkeits- und Differenzierbarkeitsannahmen für $\mathbf{f}$ existiert ein $\eta \neq 0$ so, dass die zu $\mathbf{v}(t)$ zugehörige Lösung $\mathbf{y}(t; \eta)$ des Anfangswertproblems (4.95a) für alle $t \in [t_0, T]$ eindeutig und bezüglich η differenzierbar ist. Für $\eta = 0$ gilt offensichtlich $\mathbf{v}(t; 0) = \mathbf{u}^*(t)$, $t_0 \leq t \leq t_1^*$, und $\mathbf{y}(t; 0) = \mathbf{x}^*(t)$, $t_0 \leq t \leq T$. Da der Anfangswert

$\mathbf{x}(t_0) = \mathbf{x}_0$ fest ist, muss die Beziehung $\mathbf{y}(t_0;\eta) = \mathbf{x}_0$ sowie $(\frac{\partial}{\partial\eta}\mathbf{y})(t_0;0) = \boldsymbol{\xi}_y(t_0) = \mathbf{0}$, siehe auch (4.85), gelten. Das um die Gleichungsbeschränkungen (4.95) erweiterte Kostenfunktional (4.94) für $\mathbf{v}(t;\eta)$ und der Endzeit $\bar{t}_1 = t_1^* + \eta\tau$ lautet dann

$$\begin{aligned}\bar{J}(\mathbf{v}(t;\eta),\bar{t}_1) &= \int_{t_0}^{\bar{t}_1} l(t,\mathbf{y}(t;\eta),\mathbf{v}(t;\eta)) + \boldsymbol{\lambda}^{\mathrm{T}}(t)(\mathbf{f}(t,\mathbf{y}(t;\eta),\mathbf{v}(t;\eta)) - \dot{\mathbf{y}})\,\mathrm{d}t \\ &\quad + \varphi(\bar{t}_1,\mathbf{y}(\bar{t}_1;\eta)) + \boldsymbol{\mu}^{\mathrm{T}}\boldsymbol{\psi}(\bar{t}_1,\mathbf{y}(\bar{t}_1;\eta)) \\ &= \int_{t_0}^{\bar{t}_1} l(t,\mathbf{y}(t;\eta),\mathbf{v}(t;\eta)) + \boldsymbol{\lambda}^{\mathrm{T}}(t)\mathbf{f}(t,\mathbf{y}(t;\eta),\mathbf{v}(t;\eta)) + \dot{\boldsymbol{\lambda}}^{\mathrm{T}}(t)\mathbf{y}(t;\eta)\,\mathrm{d}t \\ &\quad - \Big[\boldsymbol{\lambda}^{\mathrm{T}}(t)\mathbf{y}(t;\eta)\Big]_{t_0}^{\bar{t}_1} + \varphi(\bar{t}_1,\mathbf{y}(\bar{t}_1;\eta)) + \boldsymbol{\mu}^{\mathrm{T}}\boldsymbol{\psi}(\bar{t}_1,\mathbf{y}(\bar{t}_1;\eta))\end{aligned} \tag{4.99}$$

für jedes $\boldsymbol{\lambda} \in \left(C^1[t_0,\bar{t}_1]\right)^n$ und $\boldsymbol{\mu} \in \mathbb{R}^p$. Gemäß Satz 4.1 berechnet sich die notwendige Bedingung für ein Minimum aus

$$\delta\bar{J}(\mathbf{u}^*,t_1^*;\boldsymbol{\xi}_u,\tau) = \frac{\mathrm{d}}{\mathrm{d}\eta}\bar{J}(\mathbf{v}(t;\eta),\bar{t}_1)\Big|_{\eta=0} = 0 \tag{4.100}$$

zu

$$\begin{aligned}0 =& \int_{t_0}^{t_1^*}\left(\frac{\partial}{\partial\mathbf{x}}l\right)(t,\mathbf{x}^*,\mathbf{u}^*)\boldsymbol{\xi}_y + \left(\frac{\partial}{\partial\mathbf{u}}l\right)(t,\mathbf{x}^*,\mathbf{u}^*)\boldsymbol{\xi}_u + \boldsymbol{\lambda}^{\mathrm{T}}(t)\left(\frac{\partial}{\partial\mathbf{x}}\mathbf{f}\right)(t,\mathbf{x}^*,\mathbf{u}^*)\boldsymbol{\xi}_y\,\mathrm{d}t \\ &+ \int_{t_0}^{t_1^*}\boldsymbol{\lambda}^{\mathrm{T}}(t)\left(\frac{\partial}{\partial\mathbf{u}}\mathbf{f}\right)(t,\mathbf{x}^*,\mathbf{u}^*)\boldsymbol{\xi}_u + \dot{\boldsymbol{\lambda}}^{\mathrm{T}}(t)\boldsymbol{\xi}_y\,\mathrm{d}t + l(t_1^*,\mathbf{y}(t_1^*;0),\mathbf{v}(t_1^*;0))\tau \\ &+ \boldsymbol{\lambda}^{\mathrm{T}}(t_1^*)\mathbf{f}(t_1^*,\mathbf{y}(t_1^*;0),\mathbf{v}(t_1^*;0))\tau + \dot{\boldsymbol{\lambda}}^{\mathrm{T}}(t_1^*)\mathbf{y}(t_1^*;0)\tau - \Big[\boldsymbol{\lambda}^{\mathrm{T}}(t)\boldsymbol{\xi}_y(t)\Big]_{t_0}^{t_1^*} \\ &- \dot{\boldsymbol{\lambda}}^{\mathrm{T}}(t_1^*)\mathbf{y}(t_1^*;0)\tau - \boldsymbol{\lambda}^{\mathrm{T}}(t_1^*)\dot{\mathbf{y}}(t_1^*;0)\tau + \left(\frac{\partial}{\partial\mathbf{x}_1}\varphi\right)(t_1^*,\mathbf{y}(t_1^*;0))\boldsymbol{\xi}_y(t_1^*) \\ &+ \left(\frac{\partial}{\partial t_1}\varphi\right)(t_1^*,\mathbf{y}(t_1^*;0))\tau + \left(\frac{\partial}{\partial\mathbf{x}_1}\varphi\right)(t_1^*,\mathbf{y}(t_1^*;0))\dot{\mathbf{y}}(t_1^*;0)\tau \\ &+ \boldsymbol{\mu}^{\mathrm{T}}\left(\frac{\partial}{\partial\mathbf{x}_1}\boldsymbol{\psi}\right)(t_1^*,\mathbf{y}(t_1^*;0))\boldsymbol{\xi}_y(t_1^*) + \boldsymbol{\mu}^{\mathrm{T}}\left(\frac{\partial}{\partial t_1}\boldsymbol{\psi}\right)(t_1^*,\mathbf{y}(t_1^*;0))\tau \\ &+ \boldsymbol{\mu}^{\mathrm{T}}\left(\frac{\partial}{\partial\mathbf{x}_1}\boldsymbol{\psi}\right)(t_1^*,\mathbf{y}(t_1^*;0))\dot{\mathbf{y}}(t_1^*;0)\tau\end{aligned} \tag{4.101}$$

und mit $\mathbf{y}(t_1^*;0)=\mathbf{x}^*(t_1^*)$, $\mathbf{v}(t_1^*;0)=\mathbf{u}^*(t_1^*)$ folgt

$$\begin{aligned}0=&\int_{t_0}^{t_1^*}\left\{\left(\frac{\partial}{\partial\mathbf{x}}l\right)(t,\mathbf{x}^*,\mathbf{u}^*)+\boldsymbol{\lambda}^{\mathrm{T}}(t)\left(\frac{\partial}{\partial\mathbf{x}}\mathbf{f}\right)(t,\mathbf{x}^*,\mathbf{u}^*)+\dot{\boldsymbol{\lambda}}^{\mathrm{T}}(t)\right\}\boldsymbol{\xi}_y\,\mathrm{d}t\\&+\int_{t_0}^{t_1^*}\left\{\left(\frac{\partial}{\partial\mathbf{u}}l\right)(t,\mathbf{x}^*,\mathbf{u}^*)+\boldsymbol{\lambda}^{\mathrm{T}}(t)\left(\frac{\partial}{\partial\mathbf{u}}\mathbf{f}\right)(t,\mathbf{x}^*,\mathbf{u}^*)\right\}\boldsymbol{\xi}_u\,\mathrm{d}t+\\&+\left\{\left(\frac{\partial}{\partial\mathbf{x}_1}\varphi(t_1^*,\mathbf{x}^*(t_1^*))+\boldsymbol{\mu}^{\mathrm{T}}\frac{\partial}{\partial\mathbf{x}_1}\boldsymbol{\psi}(t_1^*,\mathbf{x}^*(t_1^*))\right)-\boldsymbol{\lambda}^{\mathrm{T}}(t_1^*)\right\}(\boldsymbol{\xi}_y(t_1^*)+\dot{\mathbf{x}}^*(t_1^*)\tau)\\&+\left\{\frac{\partial}{\partial t_1}\varphi(t_1^*,\mathbf{x}^*(t_1^*))+\boldsymbol{\mu}^{\mathrm{T}}\frac{\partial}{\partial t_1}\boldsymbol{\psi}(t_1^*,\mathbf{x}^*(t_1^*))+l(t_1^*,\mathbf{x}^*(t_1^*),\mathbf{u}^*(t_1^*))\right.\\&\quad\left.+\boldsymbol{\lambda}^{\mathrm{T}}\mathbf{f}(t_1^*,\mathbf{x}^*(t_1^*),\mathbf{u}^*(t_1^*))\right\}\tau\ .\end{aligned}\tag{4.102}$$

Da die Auswirkung von $\boldsymbol{\xi}_u(\tilde{\tau})$ für $\tilde{\tau}\in[t_0,t]$ auf den Wert $\boldsymbol{\xi}_y(t)$ nur aufwendig zu berechnen ist, wählt man $\boldsymbol{\lambda}(t)=\boldsymbol{\lambda}^*(t)$ so, dass die erste Zeile in (4.102) identisch verschwindet, d. h.

$$\dot{\boldsymbol{\lambda}}^*=-\left(\frac{\partial}{\partial\mathbf{x}}\mathbf{f}\right)^{\mathrm{T}}(t,\mathbf{x}^*(t),\mathbf{u}^*(t))\boldsymbol{\lambda}^*(t)-\left(\frac{\partial}{\partial\mathbf{x}}l\right)^{\mathrm{T}}(t,\mathbf{x}^*(t),\mathbf{u}^*(t))\tag{4.103}$$

mit der zugehörigen Endbedingung so, dass die dritte Zeile in (4.102) zu Null wird, also

$$\boldsymbol{\lambda}^*(t_1^*)=\left(\frac{\partial}{\partial\mathbf{x}_1}\varphi(t_1^*,\mathbf{x}^*(t_1^*))+(\boldsymbol{\mu}^*)^{\mathrm{T}}\frac{\partial}{\partial\mathbf{x}_1}\boldsymbol{\psi}(t_1^*,\mathbf{x}^*(t_1^*))\right)^{\mathrm{T}}.\tag{4.104}$$

Die Transversalitätsbedingung (4.98b) folgt unmittelbar aus der letzten Zeile von (4.102) und die Extremalbedingung für die Hamiltonfunktion in (4.97) ergibt sich wiederum direkt aufgrund des Fundamentallemmas der Variationsrechnung angewandt auf die zweite Zeile von (4.102). □

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 4.9 $m+2n+p+1$ *Bedingungen* zur Verfügung. Das sind m algebraische Gleichungen (4.97c), $p+1$ algebraische Gleichungen in Form der Transversalitätsbedingungen (4.98) und $2n$ Differentialgleichungen (4.97a) und (4.97b) für $\mathbf{x}^*$ und $\boldsymbol{\lambda}^*$. Zu diesen Differentialgleichungen gehören die in (4.97a) und (4.97b) 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 eine Zusammenfassung der Ergebnisse von Satz 4.9 werden in weiterer Folge nur sogenannte partielle Endbedingungen der Form

$$\psi_j=x_k(t_1)-\overline{x}_k,\quad j=1,\dots,p\tag{4.105}$$

mit $\overline{x}_k=$ konst. als fixem Endwert der Komponente x_k von $\mathbf{x}$ betrachtet. Für diesen vereinfachenden Spezialfall kann die Endbedingung für $\boldsymbol{\lambda}^*(t_1^*)$ von (4.97b) ersetzt werden und unter Berücksichtigung von (4.102) gilt Folgendes:

(a) Wenn die *Endzeit fest ist*, d. h. $t_1 = t_1^*$ und damit $\tau = 0$, verschwindet automatisch der zugehörige Eintrag in der vierten Zeile von (4.102). Es liegt somit keine Transversalitätsbedingung gemäß (4.98b) vor.

(i) Wenn für eine Komponente x_k von $\mathbf{x}$ gilt, dass *der Endwert fest ist*, dann gilt $y_k(t_1^*;\eta) = x_k(t_1^*) = x_k^*(t_1^*) = \overline{x}_k$. Daraus folgt $\xi_{y,k}(t_1^*) = 0$ womit automatisch der zugehörige Eintrag in der dritten Zeile von (4.102) verschwindet. Damit liegt für diese Komponente keine Endbedingung für den zugehörigen adjungierten Zustand $\lambda_k^*(t_1^*)$ vor.

(ii) Wenn für eine Komponente x_k von $\mathbf{x}$ gilt, dass *der Endwert frei ist*, dann muss, wie man aus der dritten Zeile von (4.102) erkennen kann, die Komponente des zugehörigen adjungierten Zustands $\lambda_k^*(t)$ folgende Endbedingung

$$\lambda_k^*(t_1^*) = \left(\frac{\partial}{\partial x_{1,k}}\varphi\right)(t_1^*, \mathbf{x}^*(t_1^*)) \tag{4.106}$$

erfüllen.

(b) Wenn *die Endzeit frei ist*, dann muss die Transversalitätsbedingung

$$\left(\frac{\partial}{\partial t_1}\Phi\right)(t_1^*, \mathbf{x}^*(t_1^*), \boldsymbol{\mu}^*) + H(t_1^*, \mathbf{x}^*(t_1^*), \mathbf{u}^*(t_1^*), \boldsymbol{\lambda}^*(t_1^*)) = 0, \quad H = l + \boldsymbol{\lambda}^{\mathrm{T}}\mathbf{f} \tag{4.107}$$

gelten. In Abhängigkeit davon, ob der Endwert einer Komponente x_k von $\mathbf{x}$ fest oder frei ist, können die Unterpunkte (i) und (ii) vom Fall (a) auch direkt hier angewandt werden.

Wenn in Satz 4.9 die Gleichungsbeschränkungen (4.95b) durch *Ungleichungsbeschränkungen* der Form

$$G_k(\mathbf{u}, t_1) = \psi_k(t_1, \mathbf{x}(t_1)) \leq 0, \quad k = 1, \ldots, p \tag{4.108}$$

ersetzt werden, so ändert sich lediglich (4.98a) zu

$$\psi_k(t_1^*, \mathbf{x}^*(t_1^*)) \leq 0 \tag{4.109a}$$

$$\boldsymbol{\mu}^* \geq \mathbf{0} \tag{4.109b}$$

$$(\boldsymbol{\mu}^*)^{\mathrm{T}}\boldsymbol{\psi}(t_1^*, \mathbf{x}^*(t_1^*)) = 0\,. \tag{4.109c}$$

Gleichung (4.109) wird auch als *complementary slackness condition* bezeichnet.

Aufgabe 4.3. Gesucht ist eine Lösung des Optimierungsproblems

$$\min_{u(\cdot)} \quad \int_0^1 \frac{1}{2}u^2 + \frac{a}{2}x^2 \;\mathrm{d}t, \quad a > 0 \tag{4.110a}$$

$$\text{u.B.v.} \quad \dot{x} = u, \quad x(0) = 1, \quad x(1) = 0\,. \tag{4.110b}$$

Zeigen Sie, dass die Lösung durch

$$x^*(t) = \frac{1}{1 - \mathrm{e}^{2\sqrt{a}}}\left(\mathrm{e}^{\sqrt{a}t} - \mathrm{e}^{\sqrt{a}(2-t)}\right), \quad u^*(t) = \frac{\sqrt{a}}{1 - \mathrm{e}^{2\sqrt{a}}}\left(\mathrm{e}^{\sqrt{a}t} + \mathrm{e}^{\sqrt{a}(2-t)}\right) \tag{4.111}$$

gegeben ist und interpretieren Sie die Ergebnisse, die in Abbildung 4.3 für verschiedene Parameterwerte a dargestellt sind.

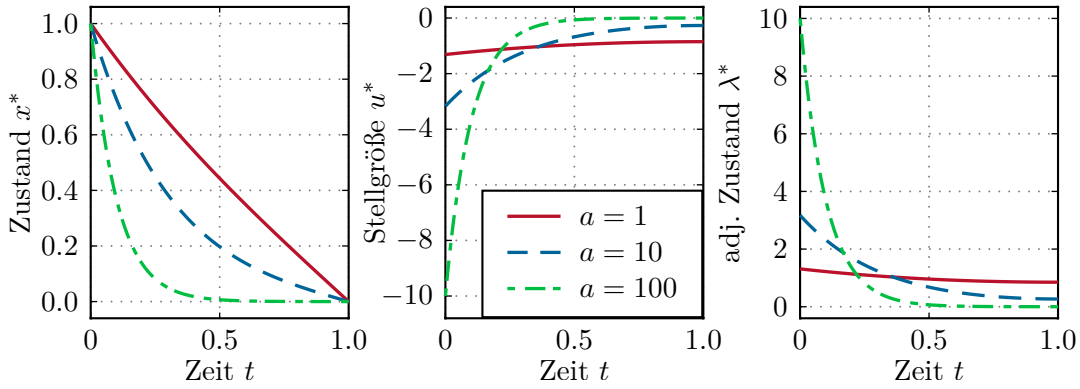


Abbildung 4.3: Optimale Trajektorien in Aufgabe 4.3.

Aufgabe 4.4. Gesucht ist eine Lösung des Optimierungsproblems

$$\min_{u(\cdot)} \quad \frac{a}{2}\, x_2^2(1) + \int_0^1 \frac{1}{2}\, u^2 \;\mathrm{d}t, \quad a \geq 0 \tag{4.112a}$$

$$\text{u.B.v.} \quad \dot{x}_1 = x_2, \quad x_1(0) = 1, \quad x_1(1) = 0 \tag{4.112b}$$

$$\dot{x}_2 = u, \quad x_2(0) = 0. \tag{4.112c}$$

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

$$x_1^*(t) = \frac{2(1+a)}{4+a}\, t^3 - \frac{3(2+a)}{a+4}\, t^2 + 1, \quad u^*(t) = \frac{12(1+a)}{4+a}\, t - \frac{6(2+a)}{a+4} \tag{4.113}$$

gegeben ist. Interpretieren Sie die Ergebnisse in Abbildung 4.4.

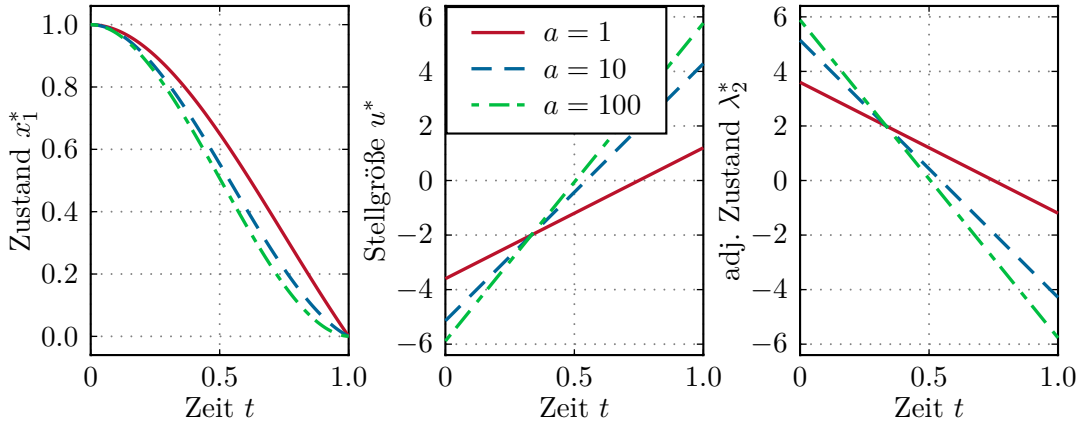


Abbildung 4.4: Optimale Trajektorien in Aufgabe 4.4.

Beispiel 4.6. Betrachtet wird eine Punktmasse mit der Masse m in der (x, y)-Ebene, auf die eine konstante Kraft $F = ma$ wirkt. Die Stellgröße u des Problems ist der Winkel zwischen der Schubrichtung und der x-Achse, siehe Abbildung 4.5. Ziel ist es, die Punktmasse in *minimaler Zeit* $[t_0 = 0, t_1^*]$ zu einem *fest vorgegebenen Zielpunkt* $(\bar{x}_1, \bar{y}_1)$ zu steuern. Unter der Annahme, dass außer dem Schub keine weiteren Kräfte auftreten, kann das Optimalsteuerungsproblem wie folgt formuliert werden

$$\min_{u(\cdot)} \quad t_1 \tag{4.114a}$$

$$\text{u.B.v.} \quad \dot{x} = v\,, \qquad x(0) = x_0\,, \quad x(t_1) = \bar{x}_1 \tag{4.114b}$$

$$\dot{v} = a\cos(u)\,, \quad v(0) = v_0 \tag{4.114c}$$

$$\dot{y} = w\,, \qquad y(0) = y_0\,, \quad y(t_1) = \bar{y}_1 \tag{4.114d}$$

$$\dot{w} = a\sin(u)\,, \quad w(0) = w_0\,. \tag{4.114e}$$

Man beachte, dass der Endzustand nur für die Position (x, y) aber nicht für die Geschwindigkeiten (v, w) vorgegeben ist.

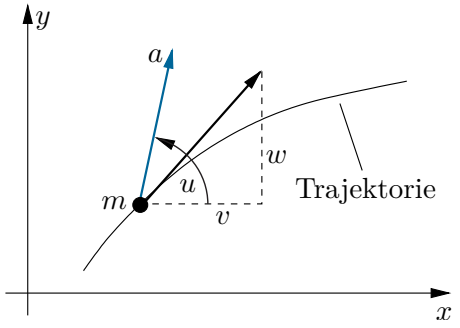


Abbildung 4.5: Bewegung einer Punktmasse der Masse m in der (x, y)-Ebene.

Die beiden fest vorgegebenen Endwerte für x und y können als Gleichungsbeschränkungen gemäß (4.95b) in der Form

$$\psi_1(t_1, \mathbf{x}(t_1)) = x(t_1) - \bar{x}_1, \quad \psi_2(t_1, \mathbf{x}(t_1)) = y(t_1) - \bar{y}_1 \tag{4.115}$$

formuliert werden. Die Hamiltonfunktion H und die Funktion Φ gemäß Satz 4.9 lauten dann für das vorliegende Optimierungsproblem

$$H(\mathbf{x}, u, \boldsymbol{\lambda}) = \lambda_x v + \lambda_v a\cos(u) + \lambda_y w + \lambda_w a\sin(u) \tag{4.116a}$$
$$\Phi(t_1, \mathbf{x}(t_1), \boldsymbol{\mu}) = \varphi + \mu_x\psi_1 + \mu_y\psi_2 = t_1 + \mu_x(x(t_1) - \bar{x}_1) + \mu_y(y(t_1) - \bar{y}_1) \tag{4.116b}$$

mit $\mathbf{x} = \begin{bmatrix} x & v & y & w \end{bmatrix}^{\mathrm{T}}$, den adjungierten Zuständen $\boldsymbol{\lambda} = \begin{bmatrix} \lambda_x & \lambda_v & \lambda_y & \lambda_w \end{bmatrix}^{\mathrm{T}}$ und den *konstanten Lagrange-Multiplikatoren* $\boldsymbol{\mu} = \begin{bmatrix} \mu_x & \mu_y \end{bmatrix}^{\mathrm{T}}$. Die Randbedingungen für den adjungierten Zustand errechnen sich gemäß (4.97b) zu

$$\lambda_x^*(t_1^*) = \left(\frac{\partial}{\partial x_1}\Phi\right)(t_1^*, \mathbf{x}^*(t_1^*), \boldsymbol{\mu}^*) = \mu_x^*, \quad \lambda_v^*(t_1^*) = \left(\frac{\partial}{\partial v_1}\Phi\right)(t_1^*, \mathbf{x}^*(t_1^*), \boldsymbol{\mu}^*) = 0 \tag{4.117a}$$

$$\lambda_y^*(t_1^*) = \left(\frac{\partial}{\partial y_1}\Phi\right)(t_1^*, \mathbf{x}^*(t_1^*), \boldsymbol{\mu}^*) = \mu_y^*, \quad \lambda_w^*(t_1^*) = \left(\frac{\partial}{\partial w_1}\Phi\right)(t_1^*, \mathbf{x}^*(t_1^*), \boldsymbol{\mu}^*) = 0\ . \tag{4.117b}$$

Damit lautet das adjungierte System

$$\dot{\lambda}_x^* = -\left(\frac{\partial H}{\partial x}\right)(\mathbf{x}^*, \mathbf{u}^*, \boldsymbol{\lambda}^*) = 0\,, \qquad \lambda_x^*(t_1^*) = \mu_x^* \tag{4.118a}$$
$$\dot{\lambda}_v^* = -\left(\frac{\partial H}{\partial v}\right)(\mathbf{x}^*, \mathbf{u}^*, \boldsymbol{\lambda}^*) = -\lambda_x^*\,, \quad \lambda_v^*(t_1^*) = 0 \tag{4.118b}$$
$$\dot{\lambda}_y^* = -\left(\frac{\partial H}{\partial y}\right)(\mathbf{x}^*, \mathbf{u}^*, \boldsymbol{\lambda}^*) = 0\,, \qquad \lambda_y^*(t_1^*) = \mu_y^* \tag{4.118c}$$
$$\dot{\lambda}_w^* = -\left(\frac{\partial H}{\partial w}\right)(\mathbf{x}^*, \mathbf{u}^*, \boldsymbol{\lambda}^*) = -\lambda_y^*\,, \quad \lambda_w^*(t_1^*) = 0\,, \tag{4.118d}$$

woraus direkt

$$\lambda_x^* = \mu_x^*\,, \quad \lambda_v^* = \mu_x^*(t_1^* - t)\,, \quad \lambda_y^* = \mu_y^*\,, \quad \lambda_w^* = \mu_y^*(t_1^* - t) \tag{4.119}$$

folgt. Des Weiteren muss die Hamiltonfunktion H gemäß (4.97) extremal sein, weshalb die Bedingung

$$\left(\frac{\partial H}{\partial u}\right)(\mathbf{x}^*, \mathbf{u}^*, \boldsymbol{\lambda}^*) = -\lambda_v^* a\sin(u^*) + \lambda_w^* a\cos(u^*) = 0\,, \tag{4.120}$$

erfüllt sein muss und u^* sich in der Form

$$\begin{aligned} \tan(u^*) &= \frac{\lambda_w^*}{\lambda_v^*} \overset{(4.119)}{=} \frac{\mu_y^*(t_1^* - t)}{\mu_x^*(t_1^* - t)} = \frac{\mu_y^*}{\mu_x^*} = \text{konst.} \\ \Rightarrow \quad u^* &= \arctan\left(\frac{\mu_y^*}{\mu_x^*}\right) \quad \text{mit} \quad -\frac{\pi}{2} < u^* < \frac{\pi}{2} \end{aligned} \tag{4.121}$$

berechnen lässt. Die optimale Steuerung u^* ist also auf dem gesamten Zeitintervall $[t_0, t_1^*]$ konstant und die zugehörigen optimalen Zustandstrajektorien $\mathbf{x}^*(t)$ können durch Lösen der Differentialgleichungen (4.114) und Einsetzen der Anfangsbedingungen in der Form

$$x^*(t) = g_x(\mu_x^*, \mu_y^*, t) = x_0 + v_0 t + \frac{1}{2} a \cos(u^*) t^2 \,, \quad \cos(u^*) = \frac{1}{\sqrt{1 + (\mu_y^*/\mu_x^*)^2}} \tag{4.122a}$$

$$v^*(t) = g_v(\mu_x^*, \mu_y^*, t) = v_0 + a \cos(u^*) t \tag{4.122b}$$

$$y^*(t) = g_y(\mu_x^*, \mu_y^*, t) = y_0 + w_0 t + \frac{1}{2} a \sin(u^*) t^2 \,, \quad \sin(u^*) = \frac{\mu_y^*}{\mu_x^* \sqrt{1 + (\mu_y^*/\mu_x^*)^2}} \tag{4.122c}$$

$$w^*(t) = g_w(\mu_x^*, \mu_y^*, t) = w_0 + a \sin(u^*) t \,. \tag{4.122d}$$

bestimmt werden. Dabei wurden die trigonometrischen Beziehungen

$$\sin(\arctan(b)) = \frac{b}{\sqrt{1 + b^2}}, \quad \cos(\arctan(b)) = \frac{1}{\sqrt{1 + b^2}} \tag{4.123}$$

verwendet. Da die Endzeit t_1 frei ist, muss zusätzlich die Transversalitätsbedingung (4.98b) gelten, wobei sich die Hamiltonfunktion (4.116a) aufgrund der Endbedingungen $\lambda_v^*(t_1^*) = \lambda_w^*(t_1^*) = 0$ entsprechend vereinfacht

$$0 = \left(\frac{\partial}{\partial t_1} \Phi\right)(t_1^*, \mathbf{x}^*(t_1^*), \boldsymbol{\mu}^*) + H(\mathbf{x}^*(t_1^*), \mathbf{u}^*(t_1^*), \boldsymbol{\lambda}^*(t_1^*)) \tag{4.124a}$$

$$= 1 + \mu_x^* \, g_v(\mu_x^*, \mu_y^*, t_1^*) + \mu_y^* \, g_w(\mu_x^*, \mu_y^*, t_1^*) \,. \tag{4.124b}$$

Mit Hilfe der zwei Gleichungsbeschränkungen (4.115) und der Transversalitätsbedingung (4.124) lässt sich ein Gleichungssystem für die verbleibenden drei Unbekannten μ_x^*, μ_y^* und t_1^* in der Form

$$\begin{bmatrix} g_x(\mu_x^*, \mu_y^*, t_1^*) \\ g_y(\mu_x^*, \mu_y^*, t_1^*) \\ \mu_x^* \, g_v(\mu_x^*, \mu_y^*, t_1^*) + \mu_y^* \, g_w(\mu_x^*, \mu_y^*, t_1^*) \end{bmatrix} - \begin{bmatrix} \bar{x}_1 \\ \bar{y}_1 \\ -1 \end{bmatrix} = \mathbf{0} \tag{4.125}$$

formulieren, welches auf *numerischem Wege* gelöst werden kann. Eine geeignete MATLAB-Funktion zur Lösung von nichtlinearen Gleichungen ist mit dem Befehl *fsolve*

aus der Optimization Toolbox gegeben. Die Funktion fsolve verwendet standardmäßig die Methode der Vertrauensbereiche (siehe Abschnitt 2.4), um ein Gleichungssystem in Residuenform $\mathbf{F}(\mathbf{x}) = \mathbf{0}$ als Minimierungsproblem in $\mathbf{x}$ zu lösen. Als Beispiel ist in der Code-Auflistung 4.1 der MATLAB-Code dargestellt, wie fsolve zur Lösung von (4.125) verwendet werden kann. Der gewünschte Endpunkt $(\bar{x}_1, \bar{y}_1)$ wird beim Aufruf der Funktion punktmasse(x1,y1) übergeben, wobei angenommen wird, dass die Punktmasse am Punkt $(x_0, y_0) = (0, 0)$ mit der Geschwindigkeit $(v_0, w_0) = (0, 1)$ in vertikale Richtung startet. Abbildung 4.6 stellt die optimalen Bahnen $x^*(t)$, $y^*(t)$ der Punktmasse in der (x, y)-Ebene für verschiedene Endpunkte $(\bar{x}_1, \bar{y}_1)$ dar. Die Pfeile zeigen die (jeweils konstante) Richtung $u^* = \arctan(\mu_y^*/\mu_x^*)$ der angreifenden Kraft ma an.

Listing 4.1: MATLAB-Code für das Punktmasse-Problem unter Verwendung von fsolve.

```
function [t,x,y,p] = punktmasse(x1,y1)
% ------------------------------------------------------
% (x1,y1): gewünschter Endpunkt
% (t,x,y): Trajektorien der Punktmasse
% p:       Parameterstruktur

p.a = 1;                                    % Parameter
p.x0=0; p.v0=0; p.y0=0; p.w0=1;             % Anfangsbedingungen
p.x1=x1; p.y1=y1;                           % Endbedingungen (Übergabe aus Funktionsaufruf)

opt  = optimoptions('fsolve','Display','iter'); % Optionen
X0   = [-1,0,1];                            % Startwert
Xopt = fsolve(@eqns,X0,opt,p);              % Numerische Lösung mit fsolve
p.mux=Xopt(1); p.muy=Xopt(2); p.t1=Xopt(3); % Lösung

t = linspace(0,p.t1,100);                   % Trajektorien
x = xfct(p.mux,p.muy,t,p);
y = yfct(p.mux,p.muy,t,p);
% ------------------------------------------------------
function res = eqns(X,p)                    % Gleichungen in Residuenform
mux=X(1); muy=X(2); t1=X(3);
res = [ xfct(mux,muy,t1,p) - p.x1;
        yfct(mux,muy,t1,p) - p.y1;
        mux*vfct(mux,muy,t1,p) + muy*wfct(mux,muy,t1,p) + 1 ];
% ------------------------------------------------------
function x = xfct(mux,muy,t,p)              % Funktionen für x und v
cosu = 1/sqrt(1+(muy/mux)^2);
x = p.x0 + p.v0*t + p.a/2*cosu*t.^2;        % 't.^2' steht für komponentenweise Auswertung
function v = vfct(mux,muy,t,p)
cosu = 1/sqrt(1+(muy/mux)^2);
v = p.v0 + p.a*cosu*t;
% ------------------------------------------------------
function y = yfct(mux,muy,t,p)              % Funktionen für y und w
sinu = muy/(mux*sqrt(1+(muy/mux)^2));
y = p.y0 + p.w0*t + p.a/2*sinu*t.^2;        % 't.^2' steht für komponentenweise Auswertung
function w = wfct(mux,muy,t,p)
sinu = muy/(mux*sqrt(1+(muy/mux)^2));
w = p.w0 + p.a*sinu*t;
```

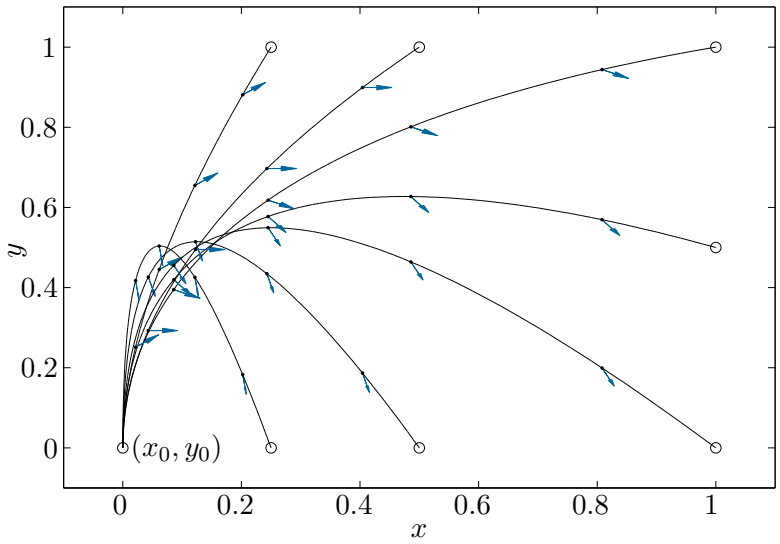


Abbildung 4.6: Zeitoptimale Steuerung einer Punktmasse zu verschiedenen Endpunkten.

Aufgabe 4.5. Gegeben ist ein lineares zeitvariantes Mehrgrößensystem der Form

$$\dot{\mathbf{x}} = \mathbf{A}(t)\mathbf{x} + \mathbf{B}(t)\mathbf{u}, \quad \mathbf{x}(t_0) = \mathbf{x}_0 \tag{4.126}$$

mit dem Zustand $\mathbf{x} \in \mathbb{R}^n$ und dem Stelleingang $\mathbf{u} \in \mathbb{R}^m$. Zeigen Sie, dass das *zeitvariante Zustandsregelgesetz*

$$\mathbf{u}^*(t) = -\mathbf{R}^{-1}(t)\mathbf{B}^{\mathrm{T}}(t)\mathbf{S}(t)\mathbf{x}^*(t) \tag{4.127}$$

mit $\mathbf{S}(t)$ als Lösung der *Matrix-Riccati-Differentialgleichung*

$$\dot{\mathbf{S}} = -\mathbf{S}\mathbf{A} - \mathbf{A}^{\mathrm{T}}\mathbf{S} + \mathbf{S}\mathbf{B}\mathbf{R}^{-1}\mathbf{B}^{\mathrm{T}}\mathbf{S} - \mathbf{Q}, \quad \mathbf{S}(t_1) = \mathbf{S}_1, \quad t_0 \le t \le t_1 \tag{4.128}$$

das Kostenfunktional

$$J(\mathbf{u}) = \frac{1}{2}\int_{t_0}^{t_1} \mathbf{x}^{\mathrm{T}}(t)\mathbf{Q}(t)\mathbf{x}(t) + \mathbf{u}^{\mathrm{T}}(t)\mathbf{R}(t)\mathbf{u}(t)\,\mathrm{d}t + \frac{1}{2}\mathbf{x}^{\mathrm{T}}(t_1)\mathbf{S}_1\mathbf{x}(t_1) \tag{4.129}$$

mit der für alle Zeiten $t_0 \le t \le t_1$ positiv definiten Matrix $\mathbf{R}(t)$, der für alle Zeiten $t_0 \le t \le t_1$ positiv semidefiniten Matrix $\mathbf{Q}(t)$ und der positiv semidefiniten 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

$$J(\mathbf{u}) = \int_{t_0}^{t_1} l(\mathbf{x}(t), \mathbf{u}(t))\,\mathrm{d}t \tag{4.130}$$

mit der freien Endzeit t_1 und einem festen Endzustand $\mathbf{x}(t_1) = \mathbf{x}_1$ unter der Gleichungsbeschränkung des dynamischen Systems

$$\dot{\mathbf{x}} = \mathbf{f}(\mathbf{x}(t), \mathbf{u}(t)), \quad \mathbf{x}(t_0) = \mathbf{x}_0 \tag{4.131}$$

für die zulässigen Stellgrößen

$$\mathbf{u} \in \left(\hat{C}_U[t_0, T]\right)^m := \left\{ \mathbf{u} \in \left(\hat{C}[t_0, T]\right)^m \,\middle|\, \mathbf{u}(t) \in U,\, \forall t_0 \le t \le t_1 \right\} \tag{4.132}$$

mit einer hinreichend großen Zeit $T \gg t_1$ und der nichtleeren Menge der Stellgrößenbeschränkungen U.

Man kann nun das Optimierungsproblem bestehend aus (4.130) und (4.131) durch Erweiterung des Zustandsvektors in der Form $\bar{\mathbf{x}}^{\mathrm{T}} = \begin{bmatrix} \mathbf{x}^{\mathrm{T}} & x_{n+1} \end{bmatrix}$ mit

$$x_{n+1}(t) := \int_{t_0}^{t} l(\mathbf{x}(\tau), \mathbf{u}(\tau))\,\mathrm{d}\tau \tag{4.133}$$

wie folgt umformulieren: Gesucht wird eine zulässige Stellgröße $\mathbf{u}(t) \in (\hat{C}_U[t_0, T])^m$ und eine Endzeit t_1 so, dass die Lösung des erweiterten Systems

$$\underbrace{\begin{bmatrix} \dot{\mathbf{x}} \\ \dot{x}_{n+1} \end{bmatrix}}_{\dot{\bar{\mathbf{x}}}} = \underbrace{\begin{bmatrix} \mathbf{f}(\mathbf{x}, \mathbf{u}) \\ l(\mathbf{x}, \mathbf{u}) \end{bmatrix}}_{\bar{\mathbf{f}}(\mathbf{x}, \mathbf{u})}, \quad \bar{\mathbf{x}}(t_0) = \begin{bmatrix} \mathbf{x}_0 \\ 0 \end{bmatrix} \tag{4.134}$$

beim Punkt $\bar{\mathbf{x}}^{\mathrm{T}}(t_1) = \begin{bmatrix} \mathbf{x}_1^{\mathrm{T}} & x_{n+1}(t_1) \end{bmatrix}$ terminiert und dabei $x_{n+1}(t_1)$ möglichst klein gemacht wird. Abbildung 4.7 veranschaulicht diesen Sachverhalt.

Die Linie durch den Punkt $(\mathbf{x}_1, 0)$ parallel zur x_{n+1}-Achse beschreibt alle Punkte einer Familie von Trajektorien $\bar{\mathbf{x}}(t)$ des erweiterten Systems (4.134), die die Bedingung $\mathbf{x}(t_1) = \mathbf{x}_1$ erfüllen 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 sind auch der Ausgangspunkt für die Herleitung des Minimumsprinzips von Pontryagin. Diese Herleitung wird z. B. in [10, 11] gezeigt.

An dieser Stelle wird jedoch auf einen Beweis verzichtet und auf die am Ende des Kapitels angegebene Literatur verwiesen.

Satz 4.10 (Minimumsprinzip von Pontryagin, vorgeschriebener Endzustand)**.** *Gesucht ist die Stellgröße* $\mathbf{u} \in (\hat{C}_U[t_0, t_1])^m$ *so, dass das Kostenfunktional (Lagrange-Form)*

$$J(\mathbf{u}) = \int_{t_0}^{t_1} l(\mathbf{x}(t), \mathbf{u}(t))\,\mathrm{d}t \tag{4.135}$$

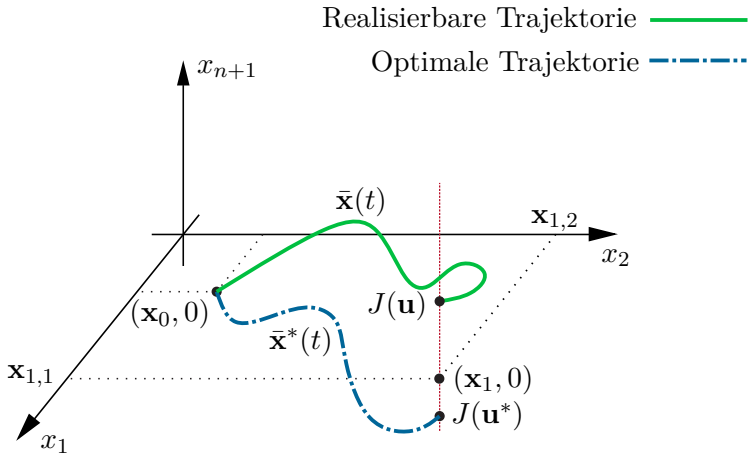


Abbildung 4.7: Zum Minimumsprinzip von Pontryagin.

unter den Gleichungsbeschränkungen (dynamisches System)

$$\dot{\mathbf{x}} - \mathbf{f}(\mathbf{x}(t), \mathbf{u}(t)) = \mathbf{0}, \quad \mathbf{x}(t_0) = \mathbf{x}_0, \quad \mathbf{x}(t_1) = \mathbf{x}_1 \tag{4.136}$$

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 $(\mathbf{u}^*, t_1^*) \in (\hat{C}_U[t_0, t_1^*])^m \times [t_0, T)$ die optimale Lösung des Optimierungsproblems und $\mathbf{x}^*$ die zugehörige Lösung von (4.136). Dann existiert ein $\bar{\boldsymbol{\lambda}}^* \in (\hat{C}^1[t_0, t_1^*])^{n+1}$, $\bar{\boldsymbol{\lambda}}^* = \begin{bmatrix} \bar{\lambda}_1^* & \dots & \bar{\lambda}_{n+1}^* \end{bmatrix}^{\mathrm{T}} \neq \mathbf{0}$ so, dass die Beziehung

$$\dot{\bar{\boldsymbol{\lambda}}}^* = -\left(\frac{\partial}{\partial \bar{\mathbf{x}}} H\right)^{\mathrm{T}} \left(\mathbf{x}^*(t), \mathbf{u}^*(t), \bar{\boldsymbol{\lambda}}^*(t)\right) \tag{4.137}$$

für $t_0 \leq t \leq t_1$ mit $H\left(\mathbf{x}^*(t), \mathbf{u}^*(t), \bar{\boldsymbol{\lambda}}^*(t)\right) = \bar{\boldsymbol{\lambda}}^{*\mathrm{T}} \bar{\mathbf{f}}(\mathbf{x}^*(t), \mathbf{u}^*(t))$ erfüllt ist, wobei $\bar{\boldsymbol{\lambda}}$ und $\bar{\mathbf{f}}$ gemäß (4.134) definiert sind, und folgende Eigenschaften gelten:

(a) Die optimale Lösung $\mathbf{u}^*(t)$ minimiert die Funktion $H\left(\mathbf{x}^*(t), \mathbf{u}(t), \bar{\boldsymbol{\lambda}}^*(t)\right)$ für alle Zeiten $t_0 \leq t \leq t_1^*$ in der Menge der Stellgrößenbeschränkungen U, d. h.

$$H\left(\mathbf{x}^*(t), \mathbf{v}, \bar{\boldsymbol{\lambda}}^*(t)\right) \geq H\left(\mathbf{x}^*(t), \mathbf{u}^*(t), \bar{\boldsymbol{\lambda}}^*(t)\right), \quad \forall \mathbf{v} \in U. \tag{4.138}$$

(b) Es gilt für alle Zeiten $t_0 \leq t \leq t_1^*$

$$\bar{\lambda}_{n+1}^*(t) = \text{konst.} \geq 0 \tag{4.139a}$$

$$H\left(\mathbf{x}^*(t), \mathbf{u}^*(t), \bar{\boldsymbol{\lambda}}^*(t)\right) = \text{konst.} \tag{4.139b}$$

(c) Es gilt die folgende Transversalitätsbedingung *(für t_1 frei)*

$$H\Big(\mathbf{x}^*(t_1^*), \mathbf{u}^*(t_1^*), \bar{\boldsymbol{\lambda}}^*(t_1^*)\Big) = 0 \ . \tag{4.140}$$

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 (4.138) 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 (4.140) und $2n + 2$ Differentialgleichungen für den erweiterten Zustand $\bar{\mathbf{x}}^*$ gemäß (4.134) und den erweiterten adjungierten Zustand $\bar{\boldsymbol{\lambda}}^*$ gemäß (4.137). Zu diesen Differentialgleichungen gehören $n + 1$ Anfangsbedingungen $\bar{\mathbf{x}}^*(t_0) = \begin{bmatrix}\mathbf{x}_0^{\mathrm{T}} & 0\end{bmatrix}^{\mathrm{T}}$, n Endbedingungen $\mathbf{x}^*(t_1^*) = \mathbf{x}_1$ sowie die Bedingung $\bar{\lambda}_{n+1}^*(t_1^*) \geq 0$. Man beachte, dass daraus noch nicht eindeutig die unbekannten Größen $(\mathbf{u}^*(t), \bar{\mathbf{x}}^*(t), \bar{\boldsymbol{\lambda}}^*(t), t_1^*)$ bestimmt werden können. Es sind nun zwei Fälle zu unterscheiden:

(i) Für $\bar{\lambda}_{n+1}^* = 0$ liegt ein *abnormaler Fall* vor, da dann wegen

$$H(\mathbf{x}^*, \mathbf{u}^*, \bar{\boldsymbol{\lambda}}^*) = \bar{\boldsymbol{\lambda}}^{*\mathrm{T}} \bar{\mathbf{f}}(\mathbf{x}^*, \mathbf{u}^*) = \boldsymbol{\lambda}^{*\mathrm{T}} \mathbf{f}(\mathbf{x}^*, \mathbf{u}^*) + \underbrace{\bar{\lambda}_{n+1}^*}_{=0} l(\mathbf{x}^*, \mathbf{u}^*) \tag{4.141}$$

die Hamiltonfunktion H und folglich auch die Optimalitätsbedingungen gemäß Satz 4.10 unabhängig von der Lagrangeschen Dichte l sind. In diesem Fall ist die Optimierungsaufgabe *nicht sinnvoll gestellt.*

(ii) Für $\bar{\lambda}_{n+1}^* > 0$ liegt ein *normaler Fall* vor und $\boldsymbol{\lambda}^*$ ist bis auf einen multiplikativen Faktor durch die genannten Gleichungen definiert. In der Praxis wählt man meist den Wert $\bar{\lambda}_{n+1}^* = 1$ und erhält damit genau die Hamiltonfunktion wie sie bereits in Abschnitt 4.2.3 verwendet wurde, siehe (4.89).

Die notwendigen Bedingungen dafür, dass die Hamiltonfunktion $H(\mathbf{x}, \mathbf{u}, \bar{\boldsymbol{\lambda}})$ für $\bar{\lambda}_{n+1}^* = 1$ (normaler Fall) bezüglich $\mathbf{u}$ minimal ist, wie in (4.138) gefordert, entsprechen den notwendigen Bedingungen erster und zweiter Ordnung (4.90c) und (4.92), d. h.

$$\left(\frac{\partial}{\partial \mathbf{u}} H\right)(\mathbf{x}^*, \mathbf{u}^*, \boldsymbol{\lambda}^*) = \mathbf{0} \tag{4.142a}$$

$$\mathbf{d}^{\mathrm{T}} \left(\frac{\partial^2}{\partial \mathbf{u}^2} H\right)(\mathbf{x}^*, \mathbf{u}^*, \boldsymbol{\lambda}^*) \mathbf{d} \geq 0 \quad \forall\, \mathbf{d} \in \mathbb{R}^m,\ t \in [t_0, t_1^*] \ , \tag{4.142b}$$

die im Rahmen der Variationsrechnung hergeleitet wurden. Trotz dieser Analogie ist das Minimumsprinzip von Pontryagin nach Satz 4.10 allgemeiner als die Ergebnisse der Variationsrechnung, da die Bedingung $\frac{\partial}{\partial \mathbf{u}} H = \mathbf{0}$ im Allgemeinen nicht mehr gültig ist, wenn das Minimum von H am Rand der Menge U der Stellgrößenbeschränkungen liegt. Im Weiteren fordert man beim Minimumsprinzip von Pontryagin lediglich die Stetigkeit von l und $\mathbf{f}$ bezüglich $\mathbf{u}$, wohingegen bei der Herleitung der Euler-Lagrange Gleichungen die stetige Differenzierbarkeit bezüglich $\mathbf{u}$ gefordert wurde, siehe Satz 4.8.

Beispiel 4.7 (Abnormaler Fall). Gegeben ist das Optimalsteuerungsproblem

$$\min_{u(\cdot)} \quad \int_0^1 l(x,u)\,\mathrm{d}t \tag{4.143a}$$

$$\text{u.B.v.} \quad \dot{x} = u\,, \quad x(0) = 0\,, \quad x(1) = 1 \tag{4.143b}$$

$$u(t) \in [0,1]\,. \tag{4.143c}$$

Es existiert nur eine realisierbare Steuerung $u^*(t) = 1$, die den Zustand $x^*(t) = t$ von $x^*(0) = 0$ nach $x^*(1) = 1$ überführt. Somit ist die optimale Lösung unabhängig von der Wahl der Kostenfunktion $l(x,u)$ und es liegt ein abnormaler Fall vor.

Beispiel 4.8. Gesucht ist das Minimum des Kostenfunktionals

$$J(u) = \frac{1}{2}\int_0^1 u^2(t)\,\mathrm{d}t \tag{4.144}$$

für das dynamische System

$$\dot{x} = -x + u, \quad x(0) = 1, \quad x(1) = 0 \tag{4.145}$$

unter Berücksichtigung der Stellgrößenbeschränkung $-0.6 \leq u(t) \leq 0$ für alle $0 \leq t \leq 1$. Die Hamiltonfunktion H von Satz 4.10 für dieses Beispiel lautet

$$H\left(x,u,\bar{\boldsymbol{\lambda}}\right) = \bar{\lambda}_1(-x+u) + \bar{\lambda}_2\frac{1}{2}u^2 \tag{4.146}$$

und die adjungierten Zustände $\bar{\boldsymbol{\lambda}}$ erfüllen gemäß (4.137) und (4.139a) die Gleichungen

$$\frac{\mathrm{d}}{\mathrm{d}t}\bar{\lambda}_1^* = -\left(\frac{\partial}{\partial x}H\right)\left(x^*,u^*,\bar{\boldsymbol{\lambda}}^*\right) = \bar{\lambda}_1^* \tag{4.147a}$$

$$\frac{\mathrm{d}}{\mathrm{d}t}\bar{\lambda}_2^* = 0\,. \tag{4.147b}$$

Daraus folgt die Lösung

$$\bar{\lambda}_1^*(t) = C_1\mathrm{e}^t \quad \text{und} \quad \bar{\lambda}_2^*(t) = C_2 \tag{4.148}$$

für geeignete Konstanten C_1 und $C_2 \geq 0$. Im Weiteren setzt man $C_2 = 1$, da ein *normaler Fall* vorliegt. Die optimale Lösung u^* mit $-0.6 \leq u^* \leq 0$ muss gemäß (4.138) der Ungleichung

$$H\left(x^*(t),v,\bar{\boldsymbol{\lambda}}^*(t)\right) \geq H\left(x^*(t),u^*(t),\bar{\boldsymbol{\lambda}}^*(t)\right), \quad \forall v \in [-0.6,0], \quad \forall t \in [0,1] \tag{4.149}$$

genügen. Aus

$$\left(\frac{\partial}{\partial u}H\right)\left(x^*, u^*, \bar{\boldsymbol{\lambda}}^*\right) = \bar{\lambda}_1^* + \bar{\lambda}_2^* u^* = \bar{\lambda}_1^* + u^* \tag{4.150}$$

folgt für die optimale Stellgröße unter Berücksichtigung der Stellgrößenbeschränkung

$$u^*(t) = \begin{cases} 0 & \text{für} \quad \bar{\lambda}_1^* \leq 0 \\ -\bar{\lambda}_1^* = -C_1 \mathrm{e}^t & \text{für} \quad 0 < \bar{\lambda}_1^* < 0.6 \\ -0.6 & \text{für} \quad \bar{\lambda}_1^* \geq 0.6 \ . \end{cases} \tag{4.151}$$

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 zulässige 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

$$u^*(t) = \begin{cases} -C_1\mathrm{e}^t & \text{falls } t \in [0, c^*] \\ -0.6 & \text{falls } t \in (c^*, 1] \end{cases} \tag{4.152}$$

mit einem Umschaltzeitpunkt (Eckpunkt) $t = c^*$ an. Für das Zeitintervall $[0, c^*]$ errechnet sich die Lösung von

$$\dot{x}_{(1)}^*(t) = -x_{(1)}^*(t) + u^*(t), \quad x_{(1)}^*(0) = 1 \tag{4.153}$$

zu

$$x_{(1)}^*(t) = -\frac{1}{2}C_1\mathrm{e}^t + \mathrm{e}^{-t}\left(\frac{1}{2}C_1 + 1\right) \tag{4.154}$$

und für das Zeitintervall $[c^*, 1]$ folgt aus

$$\dot{x}_{(2)}^*(t) = -x_{(2)}^*(t) + u^*(t), \quad x_{(2)}^*(1) = 0 \tag{4.155}$$

die Lösung zu

$$x_{(2)}^*(t) = 0.6\left(\mathrm{e}^{1-t} - 1\right) \ . \tag{4.156}$$

Nach (4.139b) muss die Hamiltonfunktion $H(x^*(t), u^*(t), \bar{\boldsymbol{\lambda}}^*(t))$ im gesamten Zeitintervall konstant sein. Daraus folgt

$$\begin{aligned} &\bar{\lambda}_1^*(\tau_1)\left(-x_{(1)}^*(\tau_1) + u^*(\tau_1)\right) + \frac{1}{2}(u^*(\tau_1))^2 = \\ &\qquad \bar{\lambda}_1^*(\tau_2)\left(-x_{(2)}^*(\tau_2) + u^*(\tau_2)\right) + \frac{1}{2}(u^*(\tau_2))^2 \quad \forall\, \tau_1 \in [0, c^*],\ \tau_2 \in (c^*, 1] \end{aligned} \tag{4.157a}$$

und nach Einsetzen

$$-C_1\left(1 + \frac{1}{2}C_1\right) = -C_1 0.6\mathrm{e} + \frac{1}{2}0.6^2 \ . \tag{4.157b}$$

Hieraus ergeben sich die zwei möglichen Lösungen $C_{1,1} = 0.436$ und $C_{1,2} = 0.826$. Der Zeitpunkt der Umschaltung (Eckpunkt) $t = c^*$ folgt aus der Stetigkeitsbedingung der Zustandsgröße

$$x^*_{(1)}(c^*) = x^*_{(2)}(c^*) \tag{4.158}$$

zu $c_1^* = 0.32$ für $C_{1,1}$ und $c_1^* = -0.32$ für $C_{1,2}$. Die für das betrachtete Zeitintervall $0 \leq t \leq 1$ relevante Lösung lautet daher $C_1 = C_{1,1} = 0.436$.

Im nächsten Schritt wird gezeigt, wie sich das Minimumsprinzip von Pontryagin nach Satz 4.10 ändert, wenn die Endbedingung $\mathbf{x}(t_1) = \mathbf{x}_1$ durch eine *Endbeschränkung* der Form $\mathbf{x}(t_1) \in \mathcal{X}_1$ mit einer glatten Mannigfaltigkeit $\mathcal{X}_1$ (siehe auch Abschnitt 3.1.1) der Dimension $n - p$ ersetzt wird. Diese $(n-p)$-dimensionale Mannigfaltigkeit wird durch p Gleichungen der Form

$$\psi_k(\mathbf{x}) = 0, \quad k = 1, \ldots, p \tag{4.159}$$

beschrieben. Der Tangentialraum $\mathcal{T}_{\check{\mathbf{x}}}\mathcal{X}_1$ an der Stelle $\mathbf{x} = \check{\mathbf{x}}$ ist dann in der Form

$$\mathcal{T}_{\check{\mathbf{x}}}\mathcal{X}_1 = \left\{ \mathbf{d} \,\middle|\, \left(\frac{\partial}{\partial \mathbf{x}} \psi_k(\check{\mathbf{x}})\right)\mathbf{d} = 0,\ k = 1, \ldots, p \right\} \tag{4.160}$$

definiert (siehe auch Abschnitt 3.1.1).

Satz 4.11 (Minimumsprinzip von Pontryagin, beschränkter Endzustand)**.** *Gesucht ist die Stellgröße* $\mathbf{u} \in (\hat{C}_U[t_0, t_1])^m$ *so, dass das Kostenfunktional (Lagrange-Form)*

$$J(\mathbf{u}) = \int_{t_0}^{t_1} l(\mathbf{x}(t), \mathbf{u}(t))\, \mathrm{d}t \tag{4.161}$$

unter den Gleichungsbeschränkungen (dynamisches System)

$$\dot{\mathbf{x}} - \mathbf{f}(\mathbf{x}(t), \mathbf{u}(t)) = \mathbf{0}, \quad \mathbf{x}(t_0) = \mathbf{x}_0, \quad \mathbf{x}(t_1) \in \mathcal{X}_1 \tag{4.162}$$

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* $(\mathbf{u}^*, t_1^*) \in (\hat{C}_U[t_0, t_1^*])^m \times [t_0, T)$ *die optimale Lösung des Optimierungsproblems und* $\mathbf{x}^*$ *die zugehörige Lösung von (4.162). Dann existiert ein* $\bar{\boldsymbol{\lambda}}^* \in (\hat{C}^1[t_0, t_1^*])^{n+1}$, $\bar{\boldsymbol{\lambda}}^* = \begin{bmatrix} (\boldsymbol{\lambda}^*)^{\mathrm{T}} & \bar{\lambda}^*_{n+1} \end{bmatrix}^{\mathrm{T}} = \begin{bmatrix} \bar{\lambda}^*_1 & \ldots & \bar{\lambda}^*_{n+1} \end{bmatrix}^{\mathrm{T}} \neq \mathbf{0}$ *so, dass die Beziehungen (4.137)–(4.140) von Satz 4.10 erfüllt sind und* $\boldsymbol{\lambda}^*(t_1^*) = \begin{bmatrix} \lambda_1^*(t_1^*) & \ldots & \lambda_n^*(t_1^*) \end{bmatrix}^{\mathrm{T}}$ orthogonal zum Tangentialraum $\mathcal{T}_{\mathbf{x}^*(t_1^*)}\mathcal{X}_1$ *ist, d. h. es gelten die* Transversalitätsbedingungen

$$(\boldsymbol{\lambda}^*)^{\mathrm{T}}(t_1^*)\mathbf{d} = 0, \quad \forall \mathbf{d} \in \mathcal{T}_{\mathbf{x}^*(t_1^*)}\mathcal{X}_1\ . \tag{4.163}$$

Nach Satz 4.11 und (4.160) 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

$$\boldsymbol{\lambda}^*(t_1^*) = \sum_{k=1}^{p} \mu_k \left(\frac{\partial}{\partial \mathbf{x}}\psi_k\right)^{\mathrm{T}}(\mathbf{x}_1^*), \quad \mathbf{x}_1^* = \mathbf{x}^*(t_1^*) \tag{4.164}$$

mit dem Lagrange-Multiplikator $\boldsymbol{\mu} = \begin{bmatrix}\mu_1 & \dots & \mu_p\end{bmatrix}^{\mathrm{T}} \in \mathbb{R}^p$. Die Bedingung

$$\operatorname{rang}\left(\left(\frac{\partial}{\partial \mathbf{x}}\boldsymbol{\psi}\right)(\mathbf{x}_1^*)\right) = p, \quad \boldsymbol{\psi} = \begin{bmatrix}\psi_1 & \dots & \psi_p\end{bmatrix}^{\mathrm{T}} \tag{4.165}$$

entspricht der LICQ (linear independence constraint qualification) Bedingung der statischen Optimierung mit Gleichungsbeschränkungen, siehe auch Definition 3.1.

Für *zeitvariante Systeme*

$$\dot{\mathbf{x}} = \mathbf{f}(t, \mathbf{x}(t), \mathbf{u}(t)), \quad \mathbf{x}(t_0) = \mathbf{x}_0 \tag{4.166}$$

führt man eine weitere Zustandsgröße der Form $x_{n+1} = t$ ein und entwirft für das erweiterte System

$$\frac{\mathrm{d}}{\mathrm{d}t}\underbrace{\begin{bmatrix}\mathbf{x} \\ x_{n+1}\end{bmatrix}}_{\mathbf{x}_e} = \underbrace{\begin{bmatrix}\mathbf{f}(x_{n+1}, \mathbf{x}, \mathbf{u}) \\ 1\end{bmatrix}}_{\mathbf{f}_e(\mathbf{x}_e, \mathbf{u})} \tag{4.167}$$

die optimale Steuerung gemäß Satz 4.10 oder Satz 4.11. Dabei wird vorausgesetzt, dass $\mathbf{f}$ und l stetig differenzierbar in t sind.

4.2.5 Minimumsprinzip für eingangsaffine Systeme

Den weiteren Betrachtungen liege das *eingangsaffine System*

$$\dot{\mathbf{x}} = \mathbf{f}(\mathbf{x}, \mathbf{u}) = \mathbf{f}_0(\mathbf{x}) + \sum_{i=1}^{m} \mathbf{f}_i(\mathbf{x}) u_i \tag{4.168}$$

mit den Stellgrößenbeschränkungen der Form

$$\mathbf{u} \in U = [\mathbf{u}^-, \mathbf{u}^+] \quad \text{bzw.} \quad u_i \in [u_i^-, u_i^+], \quad i = 1, \dots, m \tag{4.169}$$

(box constraints) zugrunde.

4.2.5.1 Kostenfunktional mit verbrauchsoptimalem Anteil

In der Literatur findet man im Zusammenhang mit dem Entwurf von *verbrauchsoptimalen Steuerungen* häufig Kostenfunktionale der Form

$$J(\mathbf{u}) = \int_{t_0}^{t_1} l_0(\mathbf{x}) + \sum_{i=1}^{m} r_i |u_i| \, \mathrm{d}t, \quad r_i > 0 \, . \tag{4.170}$$

Die zu (4.168) und (4.170) passende Hamiltonfunktion lautet (mit $\bar{\lambda}_{n+1} = 1$)

$$H(\mathbf{x}, \mathbf{u}, \boldsymbol{\lambda}) = l_0(\mathbf{x}) + \sum_{i=1}^{m} r_i |u_i| + \boldsymbol{\lambda}^{\mathrm{T}} \left(\mathbf{f}_0(\mathbf{x}) + \sum_{i=1}^{m} \mathbf{f}_i(\mathbf{x}) u_i\right) . \tag{4.171}$$

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 (4.138) vernachlässigt werden. Die Minimierung von H

$$\min_{u_i\in[u_i^-,u_i^+]} H_i(u_i) = r_i|u_i| + q_i(\mathbf{x},\boldsymbol{\lambda})u_i, \quad q_i(\mathbf{x},\boldsymbol{\lambda}) = \boldsymbol{\lambda}^{\mathrm{T}}\mathbf{f}_i(\mathbf{x}) \tag{4.172}$$

kann nun für jedes $u_i,\ i=1,\dots,m$, separat durchgeführt werden. Der Term $q_i(\mathbf{x},\boldsymbol{\lambda})$ spielt eine wichtige Rolle bei der Lösung dieses Problems. Im aktuellen Abschnitt wird davon ausgegangen, dass $u_i^- < 0 < u_i^+$ gilt. Sind diese Bedingungen nicht erfüllt, lassen sich analog zu den nachfolgenden Ausführungen sehr einfach Vorschriften zur Bestimmung der optimalen Stellgröße ableiten. Abbildung 4.8 illustriert die unterschiedlichen Fälle a)–d) mit denen die optimale Stellgröße

$$u_i^* = \begin{cases} u_i^- & \text{falls } \ q_i(\mathbf{x},\boldsymbol{\lambda}) > r_i \\ 0 & \text{falls } \ q_i(\mathbf{x},\boldsymbol{\lambda}) \in (-r_i, r_i)\,, \\ u_i^+ & \text{falls } \ q_i(\mathbf{x},\boldsymbol{\lambda}) < -r_i \end{cases} \qquad \forall\, i=1,\dots,m \tag{4.173}$$

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 (4.172) bestimmbar. Dieser Fall wird als *singulär* bezeichnet und im Abschnitt 4.2.6 näher erläutert.

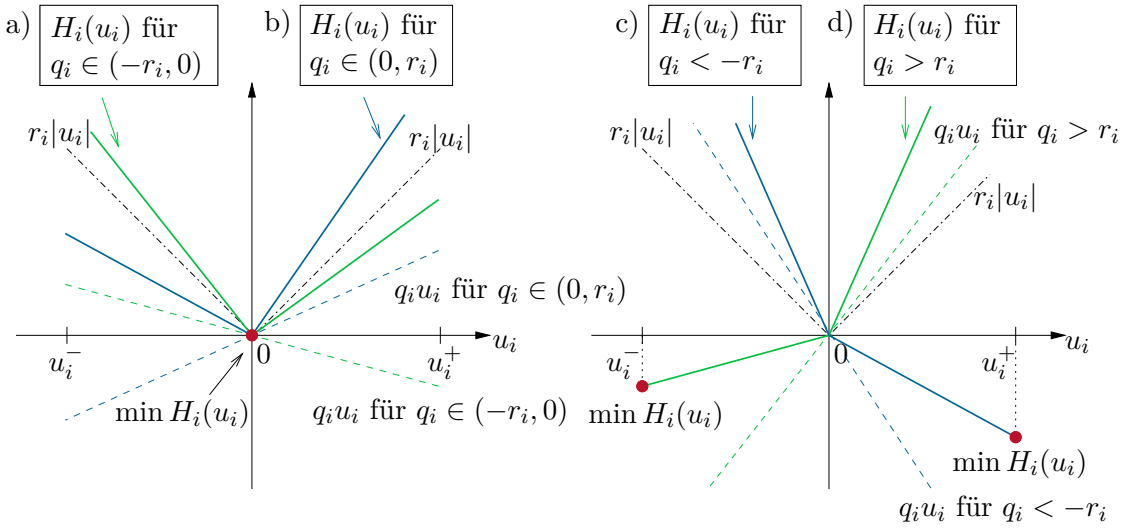


Abbildung 4.8: Verbrauchsoptimaler Fall, grafische Veranschaulichung von (4.172).

4.2.5.2 Kostenfunktional mit energieoptimalem Anteil

Unter dem Begriff *energieoptimale Steuerung* wird häufig die Minimierung eines Kostenfunktionals der Form

$$J(\mathbf{u}) = \int_{t_0}^{t_1} l_0(\mathbf{x}) + \frac{1}{2}\sum_{i=1}^{m} r_i u_i^2 \,\mathrm{d}t\,, \quad r_i > 0 \tag{4.174}$$

verstanden. Analog zum vorherigen Fall kann die Minimierung der Hamiltonfunktion H

$$\min_{u_i \in [u_i^-, u_i^+]} H_i(u_i) = \frac{1}{2} r_i u_i^2 + q_i(\mathbf{x}, \boldsymbol{\lambda}) u_i, \quad q_i(\mathbf{x}, \boldsymbol{\lambda}) = \boldsymbol{\lambda}^{\mathrm{T}} \mathbf{f}_i(\mathbf{x}) \tag{4.175}$$

wieder für jedes u_i, $i = 1, \ldots, m$, separat erfolgen. Durch den quadratischen Term mit $r_i > 0$ hätte die Funktion $H_i(u_i)$ im unbeschränkten Fall stets ein *Minimum* an der Stelle

$$u_i^0 = -\frac{1}{r_i} q_i(\mathbf{x}, \boldsymbol{\lambda}) \,. \tag{4.176}$$

Falls u_i^0 innerhalb des zulässigen Intervalls $[u_i^-, u_i^+]$ liegt, ist die optimale Lösung von (4.168), (4.169) und (4.174) durch $u_i^* = u_i^0$ gegeben. Falls u_i^0 außerhalb von $[u_i^-, u_i^+]$ liegt, so befindet sich das Minimum von $H_i(u_i)$ an der Schranke u_i^- oder u_i^+, da $H_i(u_i)$ für $u_i^0 < u_i^-$ (bzw. $u_i^0 > u_i^+$) im Intervall $[u_i^-, u_i^+]$ *streng monoton steigend (fallend)* ist, siehe Abbildung 4.9. Somit ist die optimale Stellgröße $\mathbf{u}^*(t)$ komponentenweise wie folgt definiert

$$u_i^* = \begin{cases} u_i^- & \text{falls} \quad u_i^0 \le u_i^- \\ u_i^0 & \text{falls} \quad u_i^0 \in (u_i^-, u_i^+) \,, \\ u_i^+ & \text{falls} \quad u_i^0 \ge u_i^+ \end{cases} \qquad i = 1, \ldots, m \,. \tag{4.177}$$

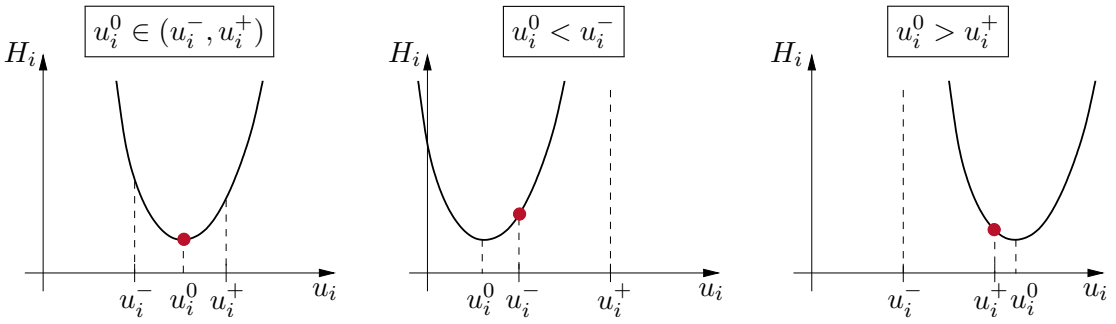


Abbildung 4.9: Energieoptimaler Fall, grafische Veranschaulichung von (4.175).

4.2.5.3 Zeitoptimales Kostenfunktional

Für zeitoptimale Probleme lautet das Kostenfunktional

$$J(\mathbf{u}) = \int_{t_0}^{t_1} 1 \, \mathrm{d}t = t_1 - t_0 \tag{4.178}$$

und die Hamiltonfunktion lässt sich in der Form (mit $\bar{\lambda}_{n+1} = 1$)

$$H(\mathbf{x}, \mathbf{u}, \boldsymbol{\lambda}) = 1 + \boldsymbol{\lambda}^{\mathrm{T}} \left(\mathbf{f}_0(\mathbf{x}) + \sum_{i=1}^{m} \mathbf{f}_i(\mathbf{x}) u_i \right) \tag{4.179}$$

anschreiben. Minimiert man die Hamiltonfunktion H wieder für jedes u_i, $i = 1, \dots, m$ separat

$$\min_{u_i \in [u_i^-, u_i^+]} H_i(u_i) = q_i(\mathbf{x}, \boldsymbol{\lambda}) u_i, \quad q_i(\mathbf{x}, \boldsymbol{\lambda}) = \boldsymbol{\lambda}^{\mathrm{T}} \mathbf{f}_i(\mathbf{x}) \ , \tag{4.180}$$

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

$$u_i^* = \begin{cases} u_i^- & \text{falls} \quad q_i(\mathbf{x}, \boldsymbol{\lambda}) > 0 \\ u_i^+ & \text{falls} \quad q_i(\mathbf{x}, \boldsymbol{\lambda}) < 0 \end{cases}, \quad i = 1, \dots, m \, . \tag{4.181}$$

Diese Steuerung wird in der Literatur häufig als *Bang-Bang-Steuerung* bezeichnet, da lediglich zwischen den Maximal- und Minimalwerten des Stellgrößenbereiches hin- und hergeschaltet wird. Ein singulärer Fall liegt vor, falls $q_i(\mathbf{x}(t), \boldsymbol{\lambda}(t))$ auf einem Subintervall $I_s \subset [t_0, t_1]$ identisch Null ist. Die Hamiltonfunktion H ist dann *unabhängig von* u_i, sodass H für jeden beliebigen Wert von u_i trivialerweise minimal ist. Die Minimumsforderung (4.180) ist damit zwar erfüllt, liefert aber keine Informationen über die Wahl von u_i.

In der Praxis wird der singuläre Fall oft durch einen zusätzlichen *Regularisierungsterm*

$$J(\mathbf{u}) = \int_{t_0}^{t_1} 1 + \frac{1}{2} \sum_{i=1}^{m} r_i u_i^2 \, \mathrm{d}t \, , \quad r_i > 0 \tag{4.182}$$

vermieden, wobei r_i hinreichend klein gewählt wird, um annähernd Zeitoptimalität zu erzielen. Der Regularisierungsterm entspricht natürlich einem energieoptimalen Anteil, so dass (4.182) die Form (4.174) besitzt und die optimale Stellgröße $\mathbf{u}^*$ gemäß (4.177) berechnet werden kann.

Beispiel 4.9 (Doppelintegrator). Zur Veranschaulichung des Minimumsprinzips von Pontryagin wird die *zeitminimale* Überführung eines Doppelintegrators mit beschränktem Eingang in den Ursprung $\mathbf{x}(t_1) = \mathbf{0}$ betrachtet. Das Optimalsteuerungsproblem mit dem Zustand $\mathbf{x} = \begin{bmatrix} x_1 & x_2 \end{bmatrix}^{\mathrm{T}}$ kann wie folgt formuliert werden

$$\min_{u(\cdot)} \quad \int_{t_0=0}^{t_1} \mathrm{d}t = t_1 \tag{4.183a}$$

$$\text{u.B.v.} \quad \dot{x}_1 = x_2, \quad \dot{x}_2 = u \tag{4.183b}$$

$$\mathbf{x}(0) = \mathbf{x}_0 \, , \quad \mathbf{x}(t_1) = \mathbf{0} \tag{4.183c}$$

$$|u(t)| \leq 1 \quad \forall t \in [0, t_1] \, . \tag{4.183d}$$

Mit der Hamiltonfunktion ($\bar{\lambda}_{n+1} = 1$) $H(\mathbf{x}, u, \boldsymbol{\lambda}) = 1 + \lambda_1 x_2 + \lambda_2 u$ ergeben sich die adjungierten Zustände $\boldsymbol{\lambda}^* = \begin{bmatrix} \lambda_1^* & \lambda_2^* \end{bmatrix}^{\mathrm{T}}$ aus (4.137) zu

$$\dot{\lambda}_1^* = 0 \quad \Rightarrow \quad \lambda_1^*(t) = C_1 \tag{4.184a}$$

$$\dot{\lambda}_2^* = -\lambda_1^* \quad \Rightarrow \quad \lambda_2^*(t) = -C_1 t + C_2 \, , \tag{4.184b}$$

wobei C_1 und C_2 Integrationskonstanten darstellen. Die Minimierungsbedingung (4.138) für die Hamiltonfunktion führt auf die optimale Stellgröße

$$u^*(t) = \begin{cases} +1 & \text{falls} \quad \lambda_2^* < 0 \\ -1 & \text{falls} \quad \lambda_2^* > 0\,. \end{cases} \tag{4.185}$$

Der *singuläre Fall*, d. h. $\lambda_2^*(t) = 0$ auf einem nicht verschwindenden Subintervall $t \in I_s \subseteq [t_0, t_1]$, kann hier nicht auftreten, da dann aufgrund von (4.184) $\lambda_2^*(t) = 0$ und $\lambda_1^*(t) = 0$ auf dem Gesamtintervall $[0, t_1]$ gelten müsste. Dies widerspricht aber der Transversalitätsbedingung (4.140) für die *freie Endzeit* t_1

$$H(\mathbf{x}^*, u^*, \boldsymbol{\lambda}^*)\big|_{t=t_1^*} = 1 + \lambda_1^*(t_1^*)x_2^*(t_1^*) + \lambda_2^*(t_1^*)\,u^*(t_1^*) = 0\,. \tag{4.186}$$

Der Fall $\lambda_2^*(t) = 0$ kann also nur zu diskreten Zeitpunkten auftreten. Zu diesen Zeitpunkten wird $u^*(t)$ zwischen -1 und $+1$ umgeschaltet. Da $\lambda_2^*(t) = -C_1 t + C_2$, gibt es *maximal einen Umschaltzeitpunkt* $t = t_s$ im Zeitintervall $[0, t_1^*]$, an dem $u^*(t)$ zwischen -1 und $+1$ wechselt. Somit existieren lediglich *vier mögliche Schaltsequenzen* $\{+1\}$, $\{-1\}$, $\{+1, -1\}$, $\{-1, +1\}$, die für eine optimale Lösung in Frage kommen. Da u stets auf einem Zeitintervall konstant ist, stellen die Trajektorien in der (x_1, x_2)-Ebene *Parabeln* dar.

Aufgabe 4.6. Zeigen Sie, dass die Lösung von (4.183b) für $u(t) = \pm 1 = \text{konst.}$ die folgende Parabelgleichung erfüllt

$$x_1 = \frac{x_2^2}{2u} + c\,, \quad u = \pm 1\,. \tag{4.187}$$

Das Optimierungsziel ist das *schnellstmögliche Erreichen des Ursprungs* $\mathbf{x}(t_1) = \mathbf{0}$ ausgehend von einem beliebigen Anfangspunkt $\mathbf{x}(0) = \mathbf{x}_0$. Der Ursprung ist aber nur über die *Schaltkurve* $x_1 = x_2^2/2$ für $u = 1$ bzw. $x_1 = -x_2^2/2$ für $u = -1$ erreichbar, siehe Abbildung 4.10. Diese beiden Fälle können mit der Funktion

$$S(x_2) = -\frac{1}{2} x_2 |x_2| \tag{4.188}$$

unterschieden werden. Da lediglich die Schaltsequenzen $\{+1\}$, $\{-1\}$, $\{+1, -1\}$, $\{-1, +1\}$ für u in Frage kommen, gibt es nur eine Möglichkeit, um den Systemzustand mit $\mathbf{x}(0) = \mathbf{x}_0 = \begin{bmatrix} x_{1,0} & x_{2,0} \end{bmatrix}^{\mathrm{T}}$ 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 4.10.

- 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 4.11.

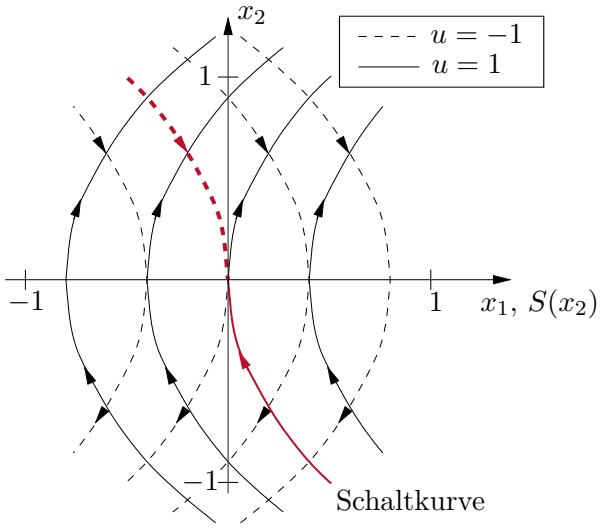


Abbildung 4.10: Mögliche Trajektorien des Doppelintegrators in der (x_1, x_2)-Ebene.

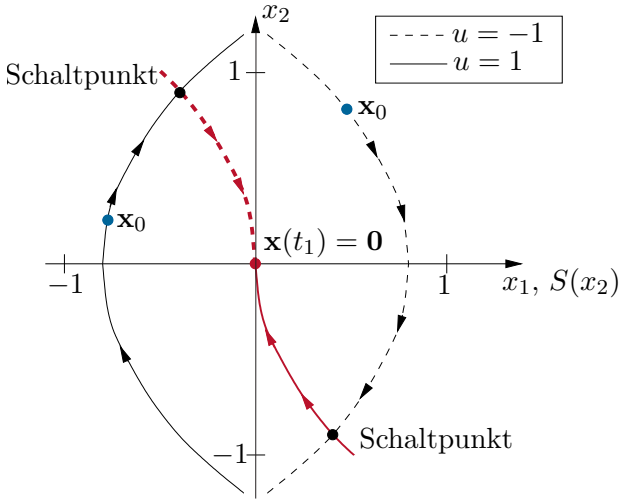


Abbildung 4.11: Optimale Umschaltung für verschiedene Anfangspunkte $\mathbf{x}_0$.

Das *optimale Stellgesetz* lautet also

$$u^*(t) = \begin{cases} +1 & \text{falls} \quad x_1 < S(x_2) \\ +1 & \text{falls} \quad x_1 = S(x_2) \quad \text{und} \quad x_1 > 0 \\ -1 & \text{falls} \quad x_1 > S(x_2) \\ -1 & \text{falls} \quad x_1 = S(x_2) \quad \text{und} \quad x_1 < 0\ . \end{cases} \tag{4.189}$$

Daraus können auch der Schaltzeitpunkt t_s und die minimale Endzeit t_1^* berechnet werden.

Aufgabe 4.7. Verifizieren Sie, dass der optimale Umschaltzeitpunkt t_s und die minimale Endzeit t_1^* wie folgt definiert sind

$$t_s = \begin{cases} x_{2,0} + \sqrt{\frac{1}{2}x_{2,0}^2 + x_{1,0}} & \text{falls} \quad x_{1,0} > S(x_{2,0}) \\ -x_{2,0} + \sqrt{\frac{1}{2}x_{2,0}^2 - x_{1,0}} & \text{falls} \quad x_{1,0} < S(x_{2,0}) \end{cases} \tag{4.190}$$

$$t_1^* = \begin{cases} x_{2,0} + \sqrt{2x_{2,0}^2 + 4x_{1,0}} & \text{falls} \quad x_{1,0} > S(x_{2,0}) \\ -x_{2,0} + \sqrt{2x_{2,0}^2 - 4x_{1,0}} & \text{falls} \quad x_{1,0} < S(x_{2,0}) \\ |x_{2,0}| & \text{falls} \quad x_{1,0} = S(x_{2,0}). \end{cases} \tag{4.191}$$

Abbildung 4.12 zeigt die zeitoptimalen Trajektorien für den Doppelintegrator für verschiedene Anfangswerte $\mathbf{x}_0 = \begin{bmatrix} x_{1,0} & x_{2,0} \end{bmatrix}^{\mathrm{T}}$.

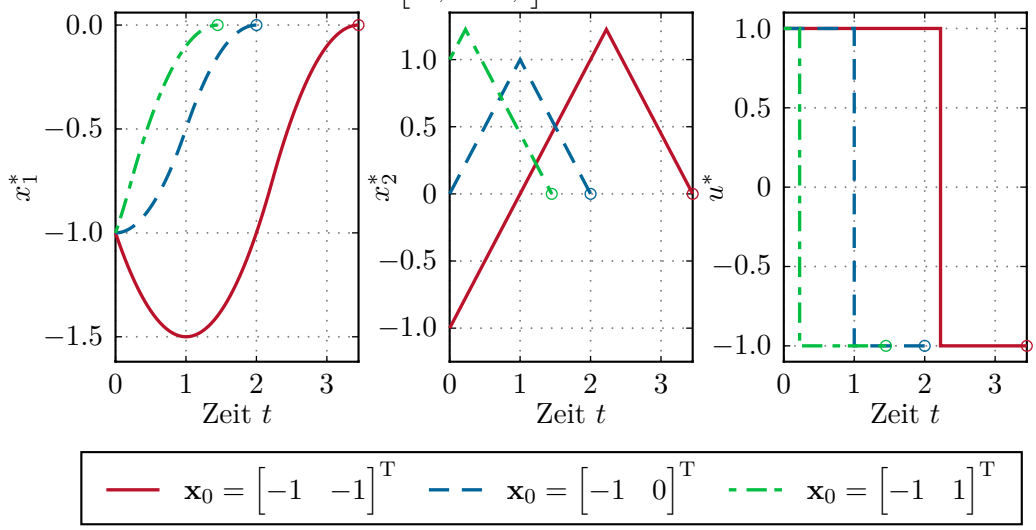


Abbildung 4.12: Zeitoptimale Trajektorien des Doppelintegrators für verschiedene Anfangswerte $\mathbf{x}_0$.

Aufgabe 4.8. Implementieren Sie das System (4.183b) mit dem optimalen Stellgesetz (4.189) in MATLAB/SIMULINK und verifizieren Sie die Ergebnisse in Abbildung 4.12 für verschiedene Anfangswerte $\mathbf{x}_0$. Verwenden Sie t_1^* gemäß (4.191) als Zeithorizont für die Simulation.

4.2.6 Der singuläre Fall

Wenn auf einem endlichen Subintervall $I_s \subseteq [t_0, t_1]$ die optimale Stellgröße $\mathbf{u}^*$ nicht aus der Minimierungsbedingung (4.138) bestimmt werden kann, so liegt ein singulärer Fall vor. Zur Verdeutlichung dieser Problematik soll im Weiteren das Optimalsteuerungsproblem

$$\min_{u(\cdot)} \quad J(u) = \int_{t_0}^{t_1} l_0(\mathbf{x}) + l_1(\mathbf{x})u \ \mathrm{d}t \tag{4.192a}$$

$$\text{u.B.v.} \quad \dot{\mathbf{x}} = \mathbf{f}_0(\mathbf{x}) + \mathbf{f}_1(\mathbf{x})u, \quad \mathbf{x}(t_0) = \mathbf{x}_0 \tag{4.192b}$$

$$u \in \hat{C}_U[t_0, t_1] \tag{4.192c}$$

mit skalarer und affin auftretender Stellgröße $u(t)$ betrachtet werden. Die grundsätzliche Vorgehensweise ist aber auch auf allgemeinere Optimalsteuerungsprobleme anwendbar. Die Hamiltonfunktion

$$H(\mathbf{x}, u, \boldsymbol{\lambda}) = l_0(\mathbf{x}) + \boldsymbol{\lambda}^{\mathrm{T}}\mathbf{f}_0(\mathbf{x}) + \Big(l_1(\mathbf{x}) + \boldsymbol{\lambda}^{\mathrm{T}}\mathbf{f}_1(\mathbf{x})\Big)u \tag{4.193}$$

ist affin in u. Die Funktion

$$\zeta(t) = \left(\frac{\partial}{\partial u}H\right)(\mathbf{x}, u, \boldsymbol{\lambda}) = l_1(\mathbf{x}) + \boldsymbol{\lambda}^{\mathrm{T}}\mathbf{f}_1(\mathbf{x}) \tag{4.194}$$

wird als *Schaltfunktion* bezeichnet und für sie gilt entlang der optimalen Lösung $\zeta^*(t) = \zeta(t)|_{\mathbf{x}=\mathbf{x}^*(t),\boldsymbol{\lambda}=\boldsymbol{\lambda}^*(t)}$. Wenn $\zeta^*(t) = 0$ auf einem endlichen Zeitintervall $I_s \subseteq [t_0, t_1]$ gilt, so liefert die Minimierungsbedingung (4.138) 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

$$\frac{\mathrm{d}^k}{\mathrm{d}t^k}(\zeta^*(t)) = 0 \qquad \forall\, k \in \mathbb{N},\, t \in I_s \ . \tag{4.195}$$

Entlang des singulären Pfades ist die Minimierungsbedingung (4.138) für alle zulässigen Stellgrößen erfüllt, d. h. es gilt nicht nur $\frac{\partial H}{\partial u} = 0$, sondern auch $\frac{\partial^2 H}{\partial u^2} = 0$. Um dennoch eine optimale Stellgröße $u^*(t)$ ermitteln zu können, sucht man die *kleinste positive natürliche Zahl* $\bar{k}$ so, dass gilt

$$\frac{\partial}{\partial u}\frac{\mathrm{d}^{\bar{k}}}{\mathrm{d}t^{\bar{k}}}(\zeta^*(t)) \neq 0 \ . \tag{4.196}$$

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

$$\frac{\mathrm{d}^k}{\mathrm{d}t^k}(\zeta^*(t)) = 0 \qquad \forall\, k = 0, \ldots, 2p-1 \tag{4.197}$$

zu liegen kommen. Ähnlich der Legendre-Clebsch Bedingung (4.92) muss entlang eines singulären Pfades für alle Zeiten $t \in I_s$ die sogenannte *generalisierte Legendre-Clebsch Bedingung*

$$(-1)^p \frac{\partial}{\partial u}\frac{\mathrm{d}^{2p}}{\mathrm{d}t^{2p}}(\zeta^*(t)) \geq 0 \tag{4.198}$$

erfüllt sein, vgl. [12].

Beispiel 4.10. Gesucht ist das Minimum des Kostenfunktionals

$$J(u) = \frac{1}{2}\int_0^2 x_1^2(t)\,\mathrm{d}t \tag{4.199}$$

für das dynamische System

$$\dot{x}_1 = x_2 + u \qquad x_1(0) = 1 \qquad x_1(2) = 0 \tag{4.200a}$$
$$\dot{x}_2 = -u \qquad x_2(0) = 1 \qquad x_2(2) = 0 \tag{4.200b}$$

unter Berücksichtigung der Stellgrößenbeschränkung $-10 \leq u(t) \leq 10$ für alle $t \in [0, 2]$. Die Hamiltonfunktion H für dieses Beispiel lautet (mit $\bar{\lambda}_{n+1} = \bar{\lambda}_3 = 1$)

$$H(\mathbf{x}, u, \boldsymbol{\lambda}) = \lambda_1(x_2 + u) - \lambda_2 u + \frac{1}{2}x_1^2 \tag{4.201}$$

und die adjungierten Zustände $\boldsymbol{\lambda}$ erfüllen gemäß (4.137) die Gleichungen

$$\frac{\mathrm{d}}{\mathrm{d}t}\lambda_1^* = -\left(\frac{\partial}{\partial x_1}H\right)(\mathbf{x}^*, u^*, \boldsymbol{\lambda}^*) = -x_1^* \tag{4.202a}$$
$$\frac{\mathrm{d}}{\mathrm{d}t}\lambda_2^* = -\left(\frac{\partial}{\partial x_2}H\right)(\mathbf{x}^*, u^*, \boldsymbol{\lambda}^*) = -\lambda_1^*\,. \tag{4.202b}$$

Die optimale Lösung u^* mit $-10 \leq u^* \leq 10$ muss der Ungleichung (4.138)

$$H(\mathbf{x}^*(t), v, \boldsymbol{\lambda}^*(t)) \geq H(\mathbf{x}^*(t), u^*(t), \boldsymbol{\lambda}^*(t)), \quad \forall v \in [-10, 10] \tag{4.203}$$

genügen. Damit folgt zunächst für die optimale Stellgröße unter Berücksichtigung der Stellgrößenbeschränkung

$$u^*(t) = \begin{cases} 10 & \text{für} \quad \lambda_1^* < \lambda_2^* \\ -10 & \text{für} \quad \lambda_1^* > \lambda_2^*\,. \end{cases} \tag{4.204}$$

Für $\lambda_1^* = \lambda_2^*$ tritt ein *singulärer Pfad* auf. Eine Auswertung von (4.196)

$$\left(\frac{\partial}{\partial u}H\right)(\mathbf{x}^*, u, \boldsymbol{\lambda}^*) = \lambda_1^* - \lambda_2^* = 0 \tag{4.205a}$$
$$\left(\frac{\mathrm{d}}{\mathrm{d}t}\frac{\partial}{\partial u}H\right)(\mathbf{x}^*, u, \boldsymbol{\lambda}^*) = \frac{\mathrm{d}}{\mathrm{d}t}\lambda_1^* - \frac{\mathrm{d}}{\mathrm{d}t}\lambda_2^* = -x_1^* + \lambda_1^* = 0 \tag{4.205b}$$
$$\left(\frac{\mathrm{d}^2}{\mathrm{d}t^2}\frac{\partial}{\partial u}H\right)(\mathbf{x}^*, u, \boldsymbol{\lambda}^*) = -x_2^* - u^* - x_1^* = 0 \tag{4.205c}$$
$$\left(\frac{\partial}{\partial u}\frac{\mathrm{d}^2}{\mathrm{d}t^2}\frac{\partial}{\partial u}H\right)(\mathbf{x}^*, u, \boldsymbol{\lambda}^*) = -1 \neq 0 \tag{4.205d}$$

liefert die Ordnung $p = 1$ des singulären Pfades. Aus (4.205c) folgt

$$u^* = -x_2^* - x_1^* \tag{4.206}$$

und aus (4.205d) folgt

$$(-1)^1 \left(\frac{\partial}{\partial u} \frac{\mathrm{d}^2}{\mathrm{d}t^2} \frac{\partial}{\partial u} H \right) (\mathbf{x}^*, u, \boldsymbol{\lambda}^*) = 1 > 0 \ . \tag{4.207}$$

Dies zeigt, dass die generalisierte Legendre-Clebsch Bedingung (4.198) hier erfüllt ist. Gemäß (4.139b) muss die Hamiltonfunktion $H(\mathbf{x}^*(t), u^*(t), \boldsymbol{\lambda}^*(t))$ im gesamten Zeitintervall konstant sein, d. h.

$$\lambda_1^* x_2^* + \frac{1}{2}(x_1^*)^2 + (\lambda_1^* - \lambda_2^*)u^* = C = \text{konst.} \tag{4.208}$$

Entlang des singulären Pfades müssen $\mathbf{x}^*(t)$ und $\boldsymbol{\lambda}^*(t)$ auf der durch (4.205a) und (4.205b) definierten Mannigfaltigkeit $\lambda_1^* = \lambda_2^* = x_1^*$ liegen (siehe auch (4.197)), weshalb sich für diesen Fall (4.208) zu

$$x_1^* x_2^* + \frac{1}{2}(x_1^*)^2 = C \tag{4.209}$$

vereinfacht.

Aufgabe 4.9. Zeigen Sie, dass die optimale Stellgröße durch

$$u^*(t) = \begin{cases} 10 & \text{für} \quad 0 \le t \le 0.299 \\ -x_2^* - x_1^* & \text{für} \quad 0.299 < t < 1.927 \\ -10 & \text{für} \quad 1.927 \le t \le 2 \end{cases} \tag{4.210}$$

gegeben ist.

4.3 Literatur

[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.

[2] H.R. Schwarz und N. Köckler, *Numerische Mathematik*, 6. Aufl. Wiesbaden: B.G. Teubner, 2006.

[3] M. Hermann, *Numerik gewöhnlicher Differentialgleichungen: Anfangs- und Randwertprobleme*. München: Oldenbourg, 2004.

[4] J. Stoer und R. Bulirsch, *Introduction to Numerical Analysis*, 3. Aufl., Ser. Texts in Applied Mathematics 12. New York, Berlin: Springer, 2002.

[5] C. Lanczos, *The Variational Principles of Mechanics*, 4. Aufl. New York: Dover, 1970.

[6] L. Meirovitch, *Methods of Analytical Dynamics*, Ser. Advanced Engineering Series. New York: McGraw-Hill, 1970.

[7] H. A. Mang und G. Hofstetter, *Festigkeitslehre*, 4. Aufl. Wien, New York: Springer, 2013.

[8] M. I. Kamien und N. L. Schwartz, *Dynamic Optimization: The Calculus of Variations and Optimal Control in Economics and Management*, 2. Aufl. Amsterdam: Elsevier, 1991.

[9] J. Troutman, *Variational Calculus and Optimal Control: Optimization with Elementary Convextity*, 2. Aufl., Ser. Undergraduate Texts in Mathematics. New York: Springer, 1996.

[10] J. Macki und A. Strauss, *Introduction to Optimal Control Theory*, Ser. Undergraduate Texts in Mathematics. New York: Springer, 1982.

[11] L. Pontryagin, V. Boltyanskii, R. Gamkrelidze und E. Mishchenko, *The Mathematical Theory of Optimal Processes*. Pergamon Press, 1964.

[12] A. E. Bryson, Jr. und Y.-C. Ho, *Applied Optimal Control: Optimization, Estimation, and Control*. John Wiley & Sons, 1975.

[13] R. F. Hartl, S. P. Sethi und R. G. Vickson, „A Survey of the Maximum Principles for Optimal Control Problems with State Constraints“, *SIAM Review*, Jg. 37, Nr. 2, S. 181–218, 1995.

[14] M. Papageorgiou, M. Leibold und M. Buss, *Optimierung: Statische, dynamische, stochastische Verfahren für die Anwendung*, 3. Aufl. Springer, 2012.

[15] M. Athans und P. L. Falb, *Optimal Control: An Introduction to the Theory and Its Applications*. New York: McGraw-Hill, 1966.

[16] B. van Brunt, *The Calculus of Variations*, Ser. Universitext. Springer, 2004.

[17] O. Föllinger, *Optimale Regelung und Steuerung*, Ser. Methoden der Regelungs- und Automatisierungstechnik. R. Oldenbourg Verlag, 1994.

[18] D. S. Naidu, *Optimal Control Systems*, Ser. Electrical Engineering Series. CRC Press, 2003.

[19] D. E. Kirk, *Optimal Control Theory: An Introduction*. Dover Publications, 2004.