Inhaltsverzeichnis

Abbildungsverzeichnis

Kapitel 1

Einleitung

Optimale Steuerungsprobleme sind aufgrund ihrer Bedeutung für viele technische, physikalische, wirtschaftliche und biologische Prozesse seit Jahrzehnten von großem Interesse und stehen auch im Mittelpunkt dieser Arbeit.
Grundlage der vorliegenden Diplomarbeit bildet ein in Grüne (2004) vorgestellter Algorithmus zur Lösung optimaler Steuerungsprobleme. Für ein darin auftretendes Maximierungsproblem werden in Jarzcyk (2005) das Brent-Verfahren, das erweiterte Brent-Verfahren und die Methode der rekursiven Suche sowohl untereinander als auch mit der ursprünglich verwendeten äquidistanten Diskretisierung verglichen. Die Untersuchungen haben ergeben, dass das Brent-Verfahren im Hinblick auf Genauigkeit und Aufwand den anderen Methoden vorzuziehen ist.
Das Ziel dieser Arbeit besteht darin, die Bundle-Newton-Methode, die für die Ermittlung des globalen Maximums einer konkaven, nicht notwendigerweise differenzierbaren Problemstellung geeignet ist, zu implementieren und anhand der Testbeispiele aus Jarczyk (2005) zu überprüfen, ob dieses Verfahren zu besseren Ergebnissen führt. Obwohl sich diese neue komplexe Optimierungsstrategie im Gegensatz zu den Verfahren aus Jarzcyk (2005) auch die Informationen der Ableitung und der Hessematrix zunutze macht und mit den Nichtdifferenzierbarkeitsstellen der vorliegenden Probleme umgehen kann, blieb sie hinter den Erwartungen zurück und konnte in den meisten Fällen trotz des sehr hohen Aufwands die Ergebnisse der jeweils genauesten Methode aus Jarzcyk (2005) nicht übertreffen. Dies lässt darauf schließen, dass das Maximierungsproblem innerhalb des Algorithmus zur Lösung des optimalen Steuerungsproblems die für eine erfolgreiche Anwendung der Bundle-Newton-Methode erforderlichen Eigenschaften nicht besitzt.

Die Arbeit ist in drei Teile untergliedert.
Der erste Teil widmet sich der Herleitung des Algorithmus zur Lösung optimaler Steuerungsprobleme. Bevor wir in Kapitel 3 optimale Steuerungsprobleme definieren und deren Eigenschaften analysieren, werden in Kapitel 2 die für die Formulierung solcher Probleme

notwendigen Kontrollsysteme untersucht. In Kapitel 4 wird die Diskretisierung optimaler Steuerungsprobleme bezüglich der Zeit als auch des Raumes durchgeführt, welche die Basis für die Herleitung des Algorithmus bildet.
Im zweiten Teil der Arbeit steht die Bundle-Newton-Methode im Mittelpunkt. Nach der Darstellung der Grundlagen in Kapitel 5 gelangen wie über die Vorgängermethoden (Kapitel 6) zu der eigentlichen Herleitung der Bundle-Newton-Methode in Kapitel 7.
Im dritten Teil der Arbeit gehen wir schließlich auf die konkrete Implementierung der Bundle-Newton-Methode (Kapitel 8) ein und vergleichen in Kapitel 9 die Ergebnisse, die sich ergeben, wenn diese Methode in den Algorithmus zur Lösung eines optimalen Steuerungsalgorithmus eingebaut wird, mit den Resultaten aus Jarzcyk (2005). Es schließt sich im letzten Kapitel eine Zusammenfassung und Beurteilung der Ergebnisse an.
Im Anhang findet sich eine Auflistung der verwendeten Routinen an. Des Weiteren wird die Programmbedienung erläutert und der CD-ROM-Inhalt beschrieben.

Ich möchte mich bei Herrn Prof. Dr. Grüne sowie Herrn Prof. Dr. Gerdts für die ausgezeichnete Betreuung und Unterstützung während der Erstellung dieser Arbeit bedanken.

Teil I

Numerische Lösung optimaler Steuerungsprobleme

In Teil I der vorliegenden Arbeit leiten wir einen Algorithmus zur Lösung eines optimalen Steuerungsproblems her. Wir werden nicht zu allen Aussagen Beweise anführen und verweisen für ausführliche Informationen auf Grüne (2004). Bei der Formulierung eines optimalen Steuerungsproblems spielt der Begriff des Kontrollsystems eine wichtige Rolle, weshalb sich Kapitel 2 kontinuierlichen und diskreten Kontrollsystemen widmet. Mit Hilfe dieser werden im darauffolgenden Kapitel optimale Steuerungsprobleme in kontinuierlicher und diskreter Zeit definiert. Die Grundlage für die numerische Berechnung der Lösung solcher Probleme bildet die Diskretisierung, welche in Kapitel 4 sowohl in Bezug auf die Zeit (falls es sich nicht schon um ein diskretes System handelt) als auch bezüglich des Raumes durchgeführt wird.

Kapitel 2

Kontrollsysteme

Nach der Definition kontinuierlicher und diskreter Kontrollsysteme werden wir analysieren, unter welchen Bedingungen eindeutige Lösungen dieser Systeme existieren. Anschließend gehen wir darauf ein, wie ein kontinuierliches Kontrollsystem durch ein diskretes approximiert werden kann.

2.1 Definition

Ein Kontrollsystem ist folgendermaßen definiert:

Definition 2.1 (Kontrollsystem). *(i) Ein Kontrollsystem in kontinuierlicher Zeit* $\mathbb{T} = \mathbb{R}$ *im* $\mathbb{R}^d, d \in \mathbb{N}$*, ist gegeben durch die gewöhnliche Differentialgleichung*

$$\frac{d}{dt}x(t) = f(x(t), u(t)), \tag{2.1}$$

wobei $f\colon \mathbb{R}^d \times U \to \mathbb{R}^d$ *ein parameterabhängiges stetiges Vektorfeld darstellt.*

(ii) Ein Kontrollsystem in diskreter Zeit $\mathbb{T} = h\mathbb{Z} = \{hk \mid k \in \mathbb{Z}\}$ *für ein* $h > 0$ *im* $\mathbb{R}^d, d \in \mathbb{N}$*, ist gegeben durch die Differenzengleichung*

$$x(t+h) = f_h(x(t), u(t)), \tag{2.2}$$

wobei $f_h\colon \mathbb{R}^d \times U \to \mathbb{R}^d$ *eine stetige Abbildung ist.*

(iii) Die Menge $U \subseteq \mathbb{R}^m$ *wird Kontrollwertebereich genannt und beinhaltet die Werte, die* $u(t)$ *für* $t \in \mathbb{R}$ *annehmen darf.*

(iv) Mit $\mathcal{U}$ *bzw.* $\mathcal{U}_h$ *bezeichnen wir den Raum der zulässigen Kontrollfunktionen, also*

$$\mathcal{U} := \{u\colon \mathbb{R} \to U \mid u \text{ zulässig}\} \text{ bzw. } \mathcal{U}_h := \{u_h\colon h\mathbb{Z} \to U \mid u_h \text{ zulässig}\}.$$

2.2 Existenz- und Eindeutigkeitssatz

Der Raum der zulässigen Kontrollfunktionen wird so gewählt, dass er zum einen ausreichend allgemein ist und zum anderen eindeutige Lösungen der Kontrollsysteme existieren. Im zeitdiskreten Fall können alle Funktionen $u_h \colon h\mathbb{Z} \to U$ zugelassen werden. Der Nachweis erfolgt über Induktion. Im kontinuierlichen Fall erfüllen essentiell beschränkte, messbare Funktionen die genannten Kriterien, stetige Kontrollfunktionen erweisen sich als zu einschränkend.

Definition 2.2 ((Lebesgue-)messbar). *Sei $I = [a, b] \subset \mathbb{R}$ ein abgeschlossenes Intervall.*

(i) Eine Funktion $f \colon I \to \mathbb{R}^n$ heißt stückweise konstant, falls eine Zerlegung von I in endlich viele Teilintervalle $I_j, j = 1, \dots, m$, existiert, so dass f auf I_j konstant ist für alle $j = 1, \dots, m$.

(ii) Eine Funktion $f \colon I \to \mathbb{R}^n$ heißt (Lebesgue-)messbar, falls eine Folge von stückweise konstanten Funktionen $f_i \colon I \to \mathbb{R}^n, i \in \mathbb{N}$, existiert mit $lim_{i\to\infty} f_i(x) = f(x)$ für fast alle $x \in I$, also für alle $x \in I$ bis auf eine Nullmenge.

(iii) Eine Funktion $f \colon \mathbb{R} \to \mathbb{R}^n$ heißt (Lebesgue-)messbar, falls für jedes abgeschlossene Teilintervall $I = [a, b] \subset \mathbb{R}$ die Einschränkung $f|_I$ messbar im Sinne von (ii) ist.

Definition 2.3 (essentiell beschränkt). *Eine Funktion $f \colon \mathbb{R} \to \mathbb{R}^n$ heißt essentiell beschränkt, wenn f außerhalb einer Nullmenge beschränkt ist.*

Der Satz von Carathéodory liefert eine Existenz- und Eindeutigkeitsaussage für ein kontinuierliches Kontrollsystem.

Satz 2.1 (Satz von Carathéodory). *Es wird ein kontinuierliches Kontrollsystem (2.1) mit den folgenden Eigenschaften betrachtet:*

(i) Der Raum der Kontrollfunktionen ist gegeben durch

$$\mathcal{U} = L_\infty(\mathbb{R}, U) := \{u \colon \mathbb{R} \to U \mid u \text{ ist messbar und essentiell beschränkt}\}.$$

(ii) Das Vektorfeld $f \colon \mathbb{R}^d \times U \to \mathbb{R}^d$ ist stetig.

(iii) Für jedes $R > 0$ existiert eine Konstante $L_R > 0$, so dass die Abschätzung

$$\|f(x_1, u) - f(x_2, u)\| \leq L_R \|x_1 - x_2\|$$

für alle $x_1, x_2 \in \mathbb{R}^d$ und alle $u \in U$ mit $\|x_1\|, \|x_2\|, \|u\| \leq R$ erfüllt ist.

Dann gibt es für jeden Punkt $x_0 \in \mathbb{R}^d$ und jede Kontrollfunktion $u \in \mathcal{U}$ ein (maximales) offenes Intervall I mit $0 \in I$ und genau eine absolut stetige Funktion $x(t)$, welche die Integralgleichung

$$x(t) = x_0 + \int_0^t f(x(\tau), u(\tau))d\tau$$

für alle $t \in I$ erfüllt.

Bemerkung 2.2. *Die Funktion $x(t)$ aus Satz 2.1 wird mit $\Phi(t, x_0, u)$ bezeichnet und stellt die Lösung von (2.1) zum Anfangswert $x_0 \in \mathbb{R}^d$ und zur Kontrollfunktion $u \in \mathcal{U}$ dar.*

2.3 Approximation eines kontinuierlichen Kontrollsystems

Im Hinblick auf die numerische Berechnung der Lösung eines kontinuierlichen Kontrollsystems ist es erforderlich, die gewöhnliche Differentialgleichung (2.1) durch ein diskretes System der Form (2.2) zu approximieren. Dazu verwenden wir in dieser Arbeit ein einfaches Einschrittverfahren, das Euler-Verfahren.

Definition 2.4 (Euler-Verfahren für Kontrollsysteme). *Für einen Zeitschritt $h > 0$ und einen Kontrollwert $u \in U$ können wir mit Hilfe des Euler-Verfahrens ein diskretes Kontrollsystem der Form (2.2) mit*

$$f_h(x, u) := x + hf(x, u)$$

definieren. Die Lösungen bezeichnen wir mit $\tilde{\phi}_h(t, x_0, u_h)$.

Der folgende Satz zeigt die Eigenschaften des in Definition 2.4 vorgestellten Verfahrens auf.

Satz 2.3. *Wir betrachten ein Kontrollsystem, für das die Voraussetzungen (i)-(iii) aus Satz 2.1 gelten. $\bar{B}_R(0)$ sei eine abgeschlossene Kugel mit Radius R um 0 im $\mathbb{R}^d$. Dann gilt für das Euler-Verfahren aus Definition 2.4 und jede Konstante $R > 0$:*

(i) Es gibt eine (von R unabhängige) Konstante $K > 0$, so dass für jede Kontrollfunktion $u \in \mathcal{U}$ mit $\|u\|_\infty \leq R$, jeden Anfangswert $x_0 \in \bar{B}_R(0)$ und für alle $t \in hN_0$, für welche die Lösungen in $\bar{B}_R(0)$ liegen, eine diskrete Kontrollfunktion $u_h \in \mathcal{U}_h$ mit der Abschätzung

$$\|\tilde{\Phi}_h(t, x_0, u_h) - \Phi(t, x_0, u)\| \leq K\sqrt{h}e^{Lt}$$

existiert.

(ii) Umgekehrt existiert eine von R abhängige Konstante $K > 0$, so dass für jedes $x_0 \in \bar{B}_R(0)$, jede diskrete Kontrollfunktion $u_h \in \mathcal{U}_h$ mit $\|u_h\|_\infty \leq R$ und die durch

$$u(\tau) = u_h(t), \quad \tau \in [t, t+h), \quad t \in h\mathbb{N}_0$$

definierte stückweise konstante (also messbare) Kontrollfunktion die Abschätzung

$$\|\tilde{\Phi}_h(t, x_0, u_h) - \Phi(t, x_0, u)\| \leq Kh(e^{Lt} - 1)$$

für alle $t \in h\mathbb{N}_0$, für welche die Lösungen in $\bar{B}_R(0)$ liegen, gilt.

Beweis. Der allgemeine Beweis kann in Gonzáles und Tidball (1991) nachgelesen werden. Für den konvexen Fall verweisen wir auf Grüne (2004), Satz 1.13. □

Kapitel 3

Optimale Steuerungsprobleme

3.1 Problemstellung

Nun betrachten wir das optimale Steuerungsproblem, für dessen Lösung wir einen Algorithmus herleiten werden.
Die Aufgabe der optimalen Steuerung besteht darin, in Abhängigkeit vom Anfangswert x_0 eine Kontrollfunktion derart zu wählen, dass der Wert, den man erhält, wenn man die Ertragsfunktion entlang einer Trajektorie integriert bzw. aufsummiert, maximiert wird. Formal lässt sich ein optimales Steuerungsproblem in der folgenden Weise definieren:

Definition 3.1 (optimales Steuerungsproblem). *Es sei ein Kontrollsystem (2.1) bzw. (2.2) gegeben. Für eine Funktion $g\colon \mathbb{R}^d \times U \to \mathbb{R}$ und einen Parameter $\delta > 0$ definieren wir das diskontierte Funktional auf unendlichem Zeithorizont in kontinuierlicher Zeit als*

$$J(x,u) := \int_0^\infty e^{-\delta t} g(\Phi(t,x,u),u(t))dt$$

und in diskreter Zeit als

$$J_h(x,u_h) := h\sum_{j=0}^{\infty}(1-\delta h)^j g(\Phi_h(jh,x,u_h),u_h(jh)).$$

Das optimale Steuerungsproblem lässt sich nun in folgender Weise formulieren:
Gesucht ist die optimale Wertefunktion

$$v(x) := \sup_{u\in\mathcal{U}} J(x,u) \;\; \textit{bzw.} \;\; v_h := \sup_{u_h\in\mathcal{U}_h} J_h(x,u_h).$$

Wir gehen von den nachstehenden Annahmen aus:

(i) Der Kontrollwertebereich U sei kompakt.

(ii) In kontinuierlicher Zeit erfülle das Kontrollsystem (2.1) die Voraussetzungen (i)-(iii) aus Satz 2.1, wobei die Lipschitz-Konstante $L_R = L$ unabhängig von R sei.
Für den zeit-diskreten Fall existiere eine Konstante $L > 0$, so dass die Lipschitz-Abschätzung

$$\|f_h(x_1, u) - f_h(x_2, u)\| \leq (1 + Lh)\|x_1 - x_2\|$$

für alle $x_1, x_2 \in \mathbb{R}^d$ und alle $u \in U$ gilt.

(iii) Die Funktion g sei stetig und erfülle

$$|g(x, u)| \leq M_g \text{ und } |g(x_1, u) - g(x_2, u)| \leq L_g\|x_1 - x_2\|$$

für alle $x, x_1, x_2 \in \mathbb{R}^d$, alle $u \in U$ und geeigneten Konstanten $M_g, L_g > 0$.

3.2 Eigenschaften der optimalen Wertefunktion

Wir weisen nun eine wichtige Eigenschaft der optimalen Wertefunktion nach.
Der Ausdruck $e^{-\delta h}$ bzw. $1 - \delta h$ mit Diskontrate $\delta > 0$ stellt den Diskontfaktor dar. Diese Größe berücksichtigt den Einfluss der Verzinsung. Seine mathematische Bedeutung wird im folgenden Lemma deutlich.

Lemma 3.1. *Es sei $\delta h < 1$ (dies ist im kontinuierlichen Fall immer gegeben, da "$h = 0$" gilt). Dann ist das diskontierte Funktional endlich:*

$$|J(x, u)| \leq \frac{M_g}{\delta} \text{ und } |J_h(x, u_h)| \leq \frac{M_g}{\delta}$$

Insbesondere gilt auch: $|v(x)| \leq \frac{M_g}{\delta}$ und $|v_h(x)| \leq \frac{M_g}{\delta}$

Beweis. In kontinuierlicher Zeit gilt:

$$\begin{aligned} |J(x, u)| &= \left| \int_0^\infty e^{-\delta t} g(\Phi(t, x, u), u(t)) dt \right| \leq \int_0^\infty e^{-\delta t} |g(\Phi(t, x, u), u(t))| dt \\ &= \int_0^\infty e^{-\delta t} M_g dt \leq M_g \int_0^\infty e^{-\delta t} dt = M_g \left[-\frac{1}{\delta} e^{-\delta t} \right]_0^\infty = \frac{M_g}{\delta} \end{aligned}$$

In diskreter Zeit erhalten wir:

$$\begin{aligned}
|J(x,u)| &= \left| h\sum_{j=0}^{\infty}(1-\delta h)^j g(\Phi_h(jh,x,u_h),u_h(jh)) \right| \\
&\leq h\sum_{j=0}^{\infty}\left|(1-\delta h)^j g(\Phi_h(jh,x,u_h),u_h(jh))\right| \\
&= h\sum_{j=0}^{\infty}(1-\delta h)^j \left|g(\Phi_h(jh,x,u_h),u_h(jh))\right| \\
&\leq h\sum_{j=0}^{\infty}(1-\delta h)^j M_g = hM_g\sum_{j=0}^{\infty}(1-\delta h)^j = hM_g\frac{1}{\delta h} = \frac{M_g}{\delta}
\end{aligned}$$

Die geometrische Reihe konvergiert wegen der Voraussetzung $\delta h < 1$. □

Für die optimale Wertefunktion kann nicht in allen Fällen Lipschitzstetigkeit nachgewiesen werden, aber eine Abschwächung, die Hölderstetigkeit.

Satz 3.2. *Wir betrachten das optimale Steuerungsproblem aus Definition 3.1. Für $\delta > L$ ist die optimale Wertefunktion lipschitzstetig mit Konstante $\frac{L_g}{\delta - L}$. Wenn $\delta \leq L$ gilt, dann ist sie hölderstetig, d.h. es existieren Konstanten $K, \gamma > 0$, so dass für alle $x, y \in \mathbb{R}^d$ die Abschätzung*

$$|v(x) - v(y)| \leq K\|x-y\|^\gamma$$

erfüllt ist. Hierbei ist $\gamma = \frac{\delta}{L}$, falls $\delta < L$ und $\gamma \in (0,1)$ beliebig, falls $\delta = L$.

Beweis. Grüne (2004), Satz 2.6 □

Die optimale Wertefunktion erfüllt das sogenannte Bellman'sche Optimalitätsprinzip, welches die Grundlage für die Herleitung des Algorithmus liefert. Das Bellman'sche Optimalitätsprinzip besagt, dass Endstücke optimaler Trajektorien wieder optimale Trajektorien darstellen. Formal ausgedrückt erhalten wir:

Satz 3.3 (Bellman'sches Optimalitätsprinzip). *Das optimale Steuerungsproblem aus Definition 3.1 sei gegeben. Dann erfüllt die optimale Wertefunktion $v(x)$ für jedes $x \in \mathbb{R}^d$ und jedes $T > 0$ die Gleichung*

$$v(x) = \sup_{u\in\mathcal{U}}\left\{\int_0^T e^{-\delta h} g(\Phi(t,x,u),u(t))dt + e^{-\delta T}v(\Phi(T,x,u))\right\}.$$

Analog gilt im diskreten Fall für jedes $k \in \mathbb{N}$ und jedes $x \in \mathbb{R}^d$:

$$v_h(x) = \sup_{u_h \in \mathcal{U}_h} \left\{ h \sum_{j=0}^{k} (1-\delta h)^j g(\Phi_h(jh, x, u_h), u_h(jh)) \right. \\ \left. + (1-\delta h)^{k+1} v_h(\Phi_h((k+1)h, x, u_h)) \right\}$$

Beweis. Wir beweisen nur den kontinuierlichen Fall, der diskrete Fall verläuft analog.
"$\leq$": Seien $x \in \mathbb{R}^d, T > 0$ und $u \in \mathcal{U}$ beliebig. Dann gilt

$$\begin{aligned} J(x,u) &= \int_0^\infty e^{-\delta t} g(\Phi(t,x,u), u(t)) dt \\ &= \int_0^T e^{-\delta t} g(\Phi(t,x,u), u(t)) dt + \int_T^\infty e^{-\delta t} g(\Phi(t,x,u), u(t)) dt \\ &\leq \int_0^T e^{-\delta t} g(\Phi(t,x,u), u(t)) dt + e^{-\delta T} g(\Phi(T,x,u)). \end{aligned}$$

Diese Ungleichung ist für jedes beliebige $u \in \mathcal{U}$ erfüllt und somit auch für das Supremum.
"$\geq$": Seien $x \in \mathbb{R}^d$, $T > 0$ und $\varepsilon > 0$ beliebig. Wir wählen $u_1 \in \mathcal{U}$ so, dass die Ungleichung

$$\begin{aligned} &\sup_{u \in \mathcal{U}} \left\{ \int_0^T e^{-\delta t} g(\Phi(t,x,u), u(t)) dt + e^{-\delta T} v(\Phi(T,x,u)) \right\} \\ &\quad \leq \int_0^T e^{-\delta t} g(\Phi(t,x,u_1), u_1(t)) dt + e^{-\delta T} v(\Phi(T,x,u_1)) + \varepsilon \end{aligned}$$

gilt. Dadurch ist u_1 auf $[0,T]$ festgelegt. Nun wählen wir $u_1|_{(T,\infty)}$ mit

$$J(\Phi(T,x,u_1), u_1(T+\cdot)) \geq v(\Phi(T,x,u_1)) - \varepsilon.$$

Daher lässt sich auf

$$\begin{aligned} &\sup_{u \in \mathcal{U}} \left\{ \int_0^T e^{-\delta t} g(\Phi(t,x,u), u(t)) dt + e^{-\delta T} v(\Phi(T,x,u)) \right\} \\ &\quad \leq \int_0^T e^{-\delta t} g(\Phi(t,x,u_1), u_1(t)) dt + e^{-\delta T} v(\Phi(T,x,u_1)) + \varepsilon \\ &\quad \leq \int_0^T e^{-\delta t} g(\Phi(t,x,u_1), u_1(t)) dt + e^{-\delta T} J(\Phi(T,x,u_1), u_1, u_1(T+\cdot)) dt + (1+e^{-\delta T})\varepsilon \\ &\quad = \int_0^T e^{-\delta t} g(\Phi(t,x,u_1), u_1(t)) dt + \int_T^\infty e^{-\delta t} g(\Phi(t,x,u_1), u_1(t)) dt + (1+e^{-\delta T})\varepsilon \\ &\quad = J(x,u_1) + (1+e^{-\delta T})\varepsilon \leq v(x) + (1+e^{-\delta T})\varepsilon \end{aligned}$$

schließen. Da $\varepsilon > 0$ beliebig klein gewählt werden kann, ergibt sich hieraus

$$\sup_{u\in\mathcal{U}} \left\{ \int_0^T e^{-\delta t} g(\Phi(t,x,u),u(t))dt + e^{-\delta T} v(\Phi(T,x,u)) \right\} \leq v(x).$$

Insgesamt ist damit die Behauptung bewiesen. □

Mit Hilfe des folgenden Lemmas kann gezeigt werden, dass die optimale Wertefunktion durch das Optimalitätsprinzip eindeutig bestimmt ist.

Lemma 3.4. *Sei B eine beliebige Menge und seien $a_1, a_2 \colon B \to \mathbb{R}$ zwei Abbildungen. Dann gilt die Abschätzung*

$$\left| \sup_{b_1\in B} a_1(b_1) - \sup_{b_2\in B} a_2(b_2) \right| \leq \sup_{b\in B} |a_1(b) - a_2(b)|.$$

Beweis. O.B.d.A. sei

$$\left| \sup_{b_1\in B} a_1(b_1) - \sup_{b_2\in B} a_2(b_2) \right| = \sup_{b_1\in B} a_1(b_1) - \sup_{b_2\in B} a_2(b_2).$$

Sei $\varepsilon > 0$ beliebig. Wir wählen ein $b_\varepsilon \in B$ so, dass

$$a_1(b_\varepsilon) \geq \sup_{b_1\in B} a_1(b_1) - \varepsilon$$

gilt. Wegen $b_\varepsilon \in B$ schließen wir auf

$$\sup_{b_2\in B} a_2(b_2) \geq a_2(b_\varepsilon)$$

und somit auf

$$\begin{aligned} \sup_{b_1\in B} a_1(b_1) - \sup_{b_2\in B} a_2(b_2) &\leq a_1(b_\varepsilon) - a_2(b_\varepsilon) + \varepsilon \\ &\leq |a_1(b_\varepsilon) - a_2(b_\varepsilon)| + \varepsilon \\ &\leq \sup_{b\in B} |a_1(b) - a_2(b)| + \varepsilon. \end{aligned}$$

Da $\varepsilon > 0$ beliebig klein gewählt werden kann, folgt die Behauptung.

□

Satz 3.5. *Wir betrachten das optimale Steuerungsproblem aus Definition 3.1 mit optimaler Wertefunktion v. Sei $T > 0$ gegeben und sei $w \colon \mathbb{R}^d \to \mathbb{R}$ eine beschränkte Funktion, die das Optimalitätsprinzip*

$$w(x) = \sup_{u\in\mathcal{U}} \left\{ \int_0^T e^{-\delta t} g(\Phi(t,x,u),u(t))dt + e^{-\delta T} w(\Phi(T,x,u)) \right\}$$

für alle $x \in \mathbb{R}^d$ erfüllt. Dann gilt w=v. Diese Aussage erhalten wir auch im diskreten Fall.

Beweis. Wir setzen

$$a_1(u) = \int_0^T e^{-\delta t} g(\Phi(t,x,u), u(t))dt + e^{-\delta T} w(\Phi(T,x,u))$$

und

$$a_2(u) = \int_0^T e^{-\delta t} g(\Phi(t,x,u), u(t))dt + e^{-\delta T} v(\Phi(T,x,u))$$

für alle $x \in \mathbb{R}^d$. Unter Anwendung von Lemma 3.4 erhalten wir

$$\begin{aligned} |w(x) - v(x)| &\leq \sup_{u\in\mathcal{U}} \Bigg| \int_0^T e^{-\delta t} g(\Phi(t,x,u), u(t))dt + e^{-\delta T} w(\Phi(T,x,u)) \\ &\qquad - \int_0^T e^{-\delta t} g(\Phi(t,x,u), u(t))dt + e^{-\delta T} v(\Phi(T,x,u)) \Bigg| \\ &= \sup_{u\in\mathcal{U}} \left\{ e^{-\delta T} |w(\Phi(T,x,u)) - v(\Phi(T,x,u))| \right\} \\ &\leq e^{-\delta T} \sup_{y\in\mathbb{R}^d} |w(y) - v(y)|, \end{aligned}$$

da $\{\Phi(T,x,u) | u \in \mathcal{U}\} \subseteq \mathbb{R}^d$. Da dies für alle $x \in \mathbb{R}^d$ gilt, folgt

$$\sup_{y\in\mathbb{R}^d} |w(y) - v(y)| \leq e^{-\delta T} \sup_{y\in\mathbb{R}^d} |w(y) - v(y)|$$

und damit

$$(1 - e^{-\delta T}) \sup_{y\in\mathbb{R}^d} |w(y) - v(y)| \leq 0.$$

Mit $1 - e^{-\delta T} > 0$ können wir auf $\sup_{y\in\mathbb{R}^d} |w(y) - v(y)| = 0$, also auf $w = v$ schließen.

□

Kapitel 4

Diskretisierung

Für die Herleitung des Algorithmus ist es erforderlich, das optimale Steuerungsproblem sowohl in Bezug auf die Zeit als auch in Bezug auf den Raum zu diskretisieren.

4.1 Diskretisierung in der Zeit

Bei der Diskretisierung bezüglich der Zeit wird ein zeitkontinuierliches Kontrollsystem durch ein zeitdiskretes Kontrollsystem ersetzt. Falls das Kontrollproblem bereits diskret ist, so ist dieser Schritt natürlich nicht notwendig. Die Diskretisierung erfolgt wie in Abschnitt 2.3 mit Hilfe des Eulerverfahrens. Diese Approximation liefert für jeden Anfangswert $x \in \mathbb{R}^d$ und jede diskrete Kontrollfunktion $u_h \colon h\mathbb{Z} \to U$ eine diskrete Näherungslösung $\tilde{\Phi}_h(t, x, u_h)$, mit der wir das zeitdiskrete optimale Steuerungsproblem

$$\tilde{v}_h(x) := \sup_{u_h \in \mathcal{U}_h} \tilde{J}_h(x, u_h) \text{ mit } \tilde{J}_h(x, u_h) := h \sum_{j=0}^{\infty} (1 - \delta h)^j g(\tilde{\Phi}_h(jh, x, u_h), u_h(jh)) \tag{4.1}$$

erhalten. Der nächste Satz zeigt, dass die Approximation $\tilde{v}_h$ für $h \to 0$ gegen v konvergiert.

Satz 4.1. *Wir betrachten das optimale Steuerungsproblem aus Definition 3.1 und das dazugehörige Euler-diskretisierte optimale Steuerungsproblem (4.1). Wir nehmen an, dass das zugrundeliegende Kontrollsystem die Voraussetzungen von Satz 2.1 und Satz 2.3 erfüllt. Dann gelten für die optimalen Wertefunktionen v und $\tilde{v}_h$ und alle $h \in [0, 1/\delta]$ die folgenden Abschätzungen für alle $x \in \mathbb{R}^d, \gamma \in (0, 1]$ aus Satz 3.2 und eine passende Konstante $K > 0$:*

(i) $v(x) \leq \tilde{v}_h(x) + K(h^{\frac{\gamma}{2}} + h)$

(ii) $\tilde{v}_h(x) \leq v_h(x) + K(h^{\gamma} + h)$

Insbesondere gilt also für eine geeignete Konstante $\tilde{K} > 0$ und alle $x \in \mathbb{R}^d$ die Abschätzung

$$|v(x) - \tilde{v}_h(x)| \leq \tilde{K} h^{\frac{\gamma}{2}}.$$

Beweis. Der allgemeine Beweis dieses Satzes kann in González und Tidball (1991) nachgelesen werden. Der Beweis für ein konvexes optimales Steuerungsproblem findet sich in Grüne (2004).

□

Die Approximation v_h erfüllt nach Satz 3.3 das Optimalitätsprinzip

$$\tilde{v}_h(x) = \sup_{u_h \in \mathcal{U}_h} \left\{ h \sum_{i=0}^{k} \beta^i g(\tilde{\Phi}_h(ih, x, u_h), u_h(ih)) + \beta^{k+1} \tilde{v}_h(\tilde{\Phi}_h((k+1)hx, u_h)) \right\}. \qquad (4.2)$$

Da $\tilde{\Phi}_h$ und g stetig in $(u_h(0), ..., u_h(k))$ sind, $\tilde{v}_h$ stetig ist und U eine kompakte Menge darstellt, kann das Supremum durch die Maximumsbildung ersetzt werden. Für $k = 0$ gibt es somit zu jedem $x \in \mathbb{R}^d$ mindestens ein $u_x^* \in U$ gibt, so dass das Supremum in (4.2) für ein $u_h \in \mathcal{U}_h$ mit $u(0) = u_x^*$ angenommen wird. Auf dieser Grundlage kann ein Iterationsverfahren für zeitdiskrete optimale Steuerungsprobleme formuliert werden.

Definition 4.1. *Iterativ werden Funktionen $v_h^i : \mathbb{R}^d \to \mathbb{R}, i = 0, 1, \ldots$ über die Berechnungsvorschrift $v_h^0(x) = 0$ und $v_h^{i+1}(x) = T_h(v_h^i)(x)$ für alle $x \in \mathbb{R}^d$ definiert, wobei der Operator $T_h \colon C(\mathbb{R}^d, \mathbb{R}) \to C(\mathbb{R}^d, \mathbb{R})$ durch*

$$T_h(w)(x) := \max_{u \in U} \{ hg(x, u) + \beta w(f_h(x, u)) \}$$

mit $\beta = 1 - \delta h$ gegeben ist. Hierbei bezeichnet $C(\mathbb{R}^d, \mathbb{R})$ die Menge der stetigen Funktionen von $\mathbb{R}^d$ nach $\mathbb{R}$.

Im Folgenden wird gezeigt, dass die durch die Iteration erzeugte Funktionenfolge gegen v_h konvergiert.

Satz 4.2. *Wir betrachten das zeitdiskretisierte optimale Steuerungsproblem 3.1 mit optimaler Wertefunktion v_h. Es sei $\delta h < 1$. Dann gilt für die Funktionen aus Definition 4.1 die Abschätzung*

$$\|v_h^i - v_h\|_\infty \leq \beta^i \frac{M_g}{\delta}.$$

Aufgrund von $\beta < 1$ folgt insbesondere die Konvergenz $v_h^i(x) \to v_h(x)$ gleichmäßig für alle $x \in \mathbb{R}^d$.

Beweis. Wir betrachten zwei beliebige Funktionen $w_1, w_2 \colon \mathbb{R}^d \to \mathbb{R}$. Aufgrund von Lemma

3.4 und $\{f_h(x,u) \mid u \in \mathcal{U}\} \subseteq \mathbb{R}^d$ ergibt sich

$$\begin{aligned}
&|T_h(w_1)(x) - T_h(w_2)(x)| = \\
&= \left|\max_{u\in U}\{hg(x,u) + \beta w_1(f_h(x,u))\} - \max_{u\in U}\{hg(x,u) + \beta w_2(f_h(x,u))\}\right| \\
&= \left|\sup_{u\in U}\{hg(x,u) + \beta w_1(f_h(x,u))\} - \sup_{u\in U}\{hg(x,u) + \beta w_2(f_h(x,u))\}\right| \\
&\leq \sup_{u\in U} |\beta w_1(f_h(x,u)) - \beta w_2(f_h(x,u))| \\
&\leq \beta \|w_1 - w_2\|_\infty.
\end{aligned}$$

Da die Ungleichung für alle $x \in \mathbb{R}^d$ erfüllt ist, folgt

$$\|T_h(w_1)(x) - T_h(w_2)(x)\|_\infty \leq \beta \|w_1 - w_2\|_\infty. \tag{4.3}$$

Gemäß dem Optimalitätsprinzip (4.2) für $k = 0$ gilt die Gleichung

$$v_h = T_h(v_h). \tag{4.4}$$

Mit (4.3) und (4.4) sowie mit der Definition von v_h^{i+1} erhalten wir

$$\|v_h - v_h^{i+1}\|_\infty = \|T_h(v_h) - T_h(v_h^i)\|_\infty \leq \beta \|v_h - v_h^i\|_\infty. \tag{4.5}$$

Nach Lemma 3.1 sowie der Iterationsvorschrift aus Definition 4.1 ergibt sich

$$\|v_h - v_h^0\|_\infty = \|v_h\|_\infty \leq \frac{M_g}{\delta} = \beta^0 \frac{M_g}{\delta}. \tag{4.6}$$

Der Beweis ist nun mit Induktion leicht zu Ende zu führen, indem wir die Ungleichung (4.6) als Induktionsanfang nutzen und (4.5) als Induktionsschritt heranziehen. □

4.2 Diskretisierung im Raum

Die Diskretisierung in Bezug auf die Zeit ist nicht ausreichend, um einen Algorithmus zur Ermittlung der optimalen Wertefunktion anzugeben, da die Funktionen v_h^i für unendlich viele Punkte berechnet werden müssten. Wir gehen deswegen wie folgt vor: Wir schränken den Definitionsbereich auf ein Rechteck Ω ein und legen über diese Menge ein Gitter. Die Funktionen v_h^i werden dann mit Hilfe von Funktionen approximiert, die eindeutig durch ihre Werte an den Eckpunkten des Gitters definiert sind und somit eine Berechnung in endlich vielen Schritten möglich machen. Ein regelmäßiges Rechteckgitter auf der Menge Ω ist folgendermaßen definiert:

Definition 4.2. $\Omega \subset \mathbb{R}^2$ *sei gegeben durch* $\Omega = [a_1, b_1] \times [a_2, b_2]$ *mit Werten* $a_1 < b_1$ *und* $a_2 < b_2$. *Ein (regelmäßiges) Rechteckgitter* Γ *auf* Ω *ist eine Menge von Rechtecken* $R_i, i = 0, ..., P-1, P = P_1 P_2$, *mit Kantenlängen* $k_1 = \frac{(b_1 - a_1)}{P_1}$ *und* $k_2 = \frac{(b_2 - a_2)}{P_2}$, *so dass*

$$\bigcup_{i=0}^{P-1} R_i = \Omega \text{ und } int\, R_i \cap int\, R_j = \emptyset \text{ für alle } i, j = 0, ..., P-1, i \neq j.$$

Mit $E_i, i = 0, ..., N-1, N = (P_1+1)(P_2+1)$, *bezeichnen wir die Eckpunkte (oder Knotenpunkte) des Gitters. Der Wert* $k = \sqrt{k_1^2 + k_2^2}$ *stellt den maximalen Durchmesser eines Rechtecks dar.*

Ein Beispiel eines solchen Gitters ist in Abbildung (4.1) zu sehen.

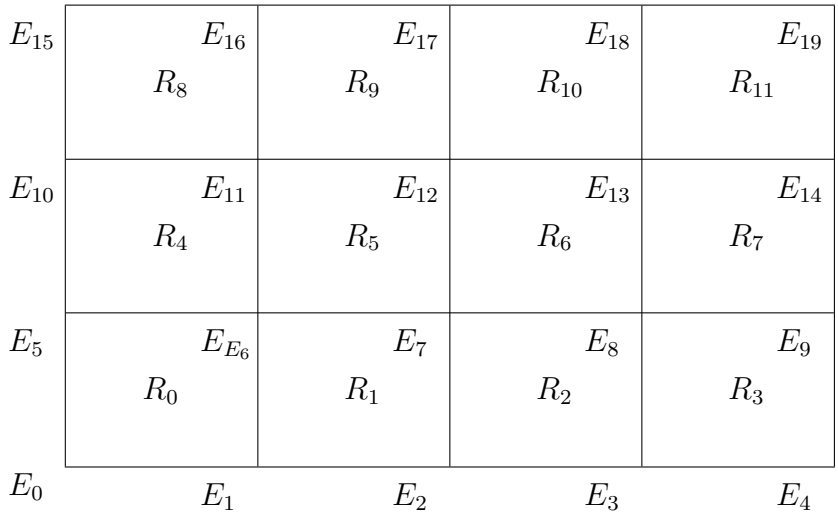

Abbildung 4.1: Beispielgitter

Bevor wir einen für die Approximation geeigneten Funktionenraum definieren, sei noch erwähnt, dass nur solche Lösungen von Φ_h bzw. $\tilde{\Phi}_h$ herangezogen werden, welche die Menge Ω für $t > 0$ nicht verlassen. Eine geeignete Menge Ω erhält man meistens schon aufgrund der Modelleigenschaften. Für das Verhalten der Lösungen gibt es drei Möglichkeiten:

(1) Die Menge Ω wird stark invariant genannt, wenn gilt:

$$f_h(x, u) \in \Omega \text{ für alle } x \in \Omega \text{ und für alle } u \in U$$

Die Einschränkung auf eine kleinere Menge bereitet hier keine Schwierigkeiten.

(2) Die Menge Ω wird schwach invariant genannt, wenn gilt:

$$\text{Für alle } x \in \Omega \text{ gibt es mindestens ein } u \in U \text{ mit } f_h(x,u) \in \Omega$$

In diesem Fall wird nicht über alle, sondern nur über die $u \in U$ optimiert, die diese Bedingung erfüllen. Die so erhaltene optimale Wertefunktion kann daher auch kleinere Werte annehmen als die ursprüngliche.

(3) Die Menge Ω wird nicht invariant genannt, wenn gilt:

$$\text{Es gibt ein } x \in \Omega, \text{ so dass für alle } u \in U \text{ gilt: } f_h(x,u) \notin \Omega$$

Es gibt auch in diesem Fall Methoden zur Berechnung einer Lösung, allerdings weicht diese unter Umständen erheblich von der tatsächlichen optimalen Wertefunktion ab.

Bei den in Kapitel 9 behandelten optimalen Steuerungsproblemen tritt der letzte Fall nicht ein. Des Weiteren beschränken wir uns im Folgenden auf $d = 2$, da die Beispiele zweidimensional sind und die verwendeten Ideen leicht auf höhere Dimensionen verallgemeinert werden können.
Einen geeigneten endlichdimensionalen Funktionenraum für unser Problem stellt der Raum der stetigen und stückweise affin bilinearen Funktionen auf einem Rechteck Ω bezüglich des Rechteckgitters Γ dar.

Definition 4.3. *(i) Sei $A \subset \mathbb{R}^2$. Eine Funktion $w\colon A \to \mathbb{R}$ heißt affin bilinear, falls es Konstanten $\alpha_0, \ldots, \alpha_3$ gibt, so dass für alle $x = (x_1, x_2)^\top \in A$ die Identität $w(x) = \alpha_0 + \alpha_1 x_1 + \alpha_2 x_2 + \alpha_3 x_1 x_2$ gilt.*

(ii) Wir betrachten eine rechteckförmige Menge $\Omega \subset \mathbb{R}^2$ mit Rechteckgitter Γ und definieren den Raum der stetigen und stückweise affin bilinearen Funktionen auf Ω bezüglich Γ als

$$\mathcal{W} := \{w\colon \Omega \to \mathbb{R} \,|\, w \text{ ist stetig und } w|_{R_i} \text{ ist affin bilinear für jedes } i = 0, \ldots, P-1\}.$$

Das nachstehende Lemma liefert die wichtige Eigenschaft, dass sich jede Funktion $w \in \mathcal{W}$ eindeutig über die Ecken des Gitters darstellen lässt. Da die Anzahl der Ecken, d.h. die Anzahl der Basiselemente, endlich ist, handelt es sich beim Funktionenraum $\mathcal{W}$ um einen endlichdimensionalen Vektorraum über $\mathbb{R}$.

Lemma 4.3. *(i) Jede Funktion $w \in \mathcal{W}$ ist eindeutig durch ihre Werte $w(E_i)$ an den Eckpunkten des Gitters bestimmt.*
(ii) Für jedes Rechteck $R_i = [c_1, d_1] \times [c_2, d_2]$ mit den Eckpunkten

$$E_{i_0} = (c_1, c_2)^\top,\ E_{i_1} = (d_1, c_2)^\top,\ E_{i_2} = (c_1, d_2)^\top,\ E_{i_3} = (d_1, d_2)^\top$$

lässt sich $w|_{R_i}$ *für* $x = (x_1, x_2)^\top \in R_i$ *schreiben als*

$$w(x) = \sum_{j=0}^{3} \mu_j(x) w(E_{i_j})$$

mit

$$\mu_0(x) = (1 - y_1(x))(1 - y_2(x)), \qquad \mu_1(x) = y_1(x)(1 - y_2(x))$$
$$\mu_2(x) = (1 - y_1(x))y_2(x), \qquad \mu_3(x) = y_1(x)y_2(x)$$

und

$$y_l = \frac{x_l - c_l}{d_l - c_l} \text{ für } l = 1, 2.$$

Insbesondere gilt hierbei $\mu_j(x) \geq 0$ *für* $j = 0, ..., 3$ *und* $\sum_{j=0}^{3} \mu_j(x) = 1$.

Beweis. Die Gültigkeit der obigen Aussagen ist leicht nachzurechnen. Wir verweisen auf Grüne (2004), Lemma 3.10. □

4.3 Vollständige Diskretisierung

Mit Hilfe der räumlichen Diskretisierung kann nun der iterative Algorithmus aus Definition 4.1, der bisher nur die Diskretisierung bezüglich der Zeit ausnutzt, so formuliert werden, dass eine Berechnung der optimalen Wertefunktion möglich ist. Denn der Operator T_h muss nur noch an endlich vielen Punkten, nämlich an den Eckpunkten des Gitters, ausgewertet werden. Es wird also eine Folge von Funktionen $\hat{v}_h^j \in \mathcal{W}$ über die Vorschrift

$$\hat{v}_h^{j+1}(E_i) = \max_{u \in U} \left\{ hg(E_i, u) + \beta \hat{v}_h^j(\Phi_h(E_i, u)) \right\},$$

mit $\beta = 1 - \delta h$ bestimmt. Schreiben wir $V^j = (V_1^j, \ldots, V_N^j) \in \mathbb{R}^N$ mit $V_i^j = \hat{v}_h^j(E_i)$, so kann diese Iteration auf $\mathcal{W}$ nun als eine Iteration auf N-dimensionalen Vektoren gemäß der folgenden Definition formuliert werden.

Definition 4.4. *Wir betrachten ein zeitdiskretes optimales Steuerungsproblem und ein Rechteckgitter* Γ *mit* P *Rechtecken und* N *Eckpunkten. Zu jedem* $u \in U$ *und jedem* $i = 0, \ldots, N-1$ *sei* $B(i, u)$ *der* N*-dimensionale Zeilenvektor, für den für jedes* $w \in \mathcal{W}$ *und* $W = (w(E_0), \ldots, w(E_{N-1}))^\top \in \mathbb{R}^N$ *mit der üblichen Matrixmultiplikation*

$$w(\Phi_h(E_i, u)) = B(i, u)W$$

gilt. Außerdem sei $G(i,u) = hg(E_i,u)$. *Dann berechnen wir Vektoren* V^j *iterativ durch* $V^0 = (0,\dots,0)^\top$ *und dem Gesamtschrittverfahren*

$$V_i^{j+1} := \max_{u\in U}\{G(i,u) + \beta B(i,u)V^j\} \text{ für } i = 0,\dots,N-1$$

oder dem Einzelschrittverfahren

$$V^{j+1} := V^j, \quad V_i^{j+1} := \max_{u\in U}\{\underbrace{G(i,u) + \beta B(i,u)V^{j+1}}_{=\tilde{h}(u)}\} \text{ für } i = 0,\dots,N-1. \tag{4.7}$$

$B(i,u)$ enthält für jede Ecke des Gitters die μ_j aus Lemma 4.3 in globaler Notierung. Nur die Einträge der aktiven Ecke sind ungleich Null. Da $\sum_{j=0}^{3}\mu_j = 1$ gilt, können auch die Einträge von $B(i,u)$ zu 1 aufsummiert werden.

Das Einzelschrittverfahren ist dem Gesamtschrittverfahren im Allgemeinen vorzuziehen, da beim Einzelschrittverfahren für jedes $i > 0$ bereits die aktuellen Werte V_k^{j+1} für $0 \le k \le i$ verwendet werden, was ein etwas besseres Konvergenzverhalten zur Folge hat. Ein weiterer Vorteil des Einzelschrittverfahrens besteht darin, dass in jedem Iterationsschritt nur ein Vektor abgespeichert werden muss, während das Gesamtschrittverfahren jeweils die Abspeicherung der zwei Vektoren V^j und V^{j+1} erfordert. Im Algorithmus aus Jarczyk (2005), der dieser Arbeit zugrunde liegt, wird daher das Einschrittverfahren verwendet.

Das folgende Lemma enthält ein geeignetes Abbruchkriterium für den vorgestellten Algorithmus.

Lemma 4.4. *Wir betrachten die Iterationsvorschrift aus Definition 4.4. Sei* $\delta h < 1$. *Dann konvergieren die Vektoren* V^j *für* $j \to \infty$ *komponentenweise gegen den Vektor* V, *der eindeutig durch*

$$V_i = \max_{u\in U}\{G(i,u) + \beta B(i,u)V\} \text{ für } i = 0,\dots,N-1$$

bestimmt ist. Für die mit $\hat{v}_h^j, j = 1,\dots,\infty$ *und* $\hat{v}_h$ *bezeichneten zugehörigen Funktionen aus* $\mathcal{W}$ *gilt außerdem: Falls* $|V_i^j - V_i^{j+1}| \le \varepsilon$ *für alle* $i = 0,\dots,N-1$, *so folgt*

$$\|\hat{v}_h^j - \hat{v}_h\| \le \frac{\varepsilon}{h\delta}.$$

Beweis. Für $X \in \mathbb{R}^N$ definieren wir

$$f(X) = \begin{pmatrix} f_1(X) \\ \vdots \\ f_{N-1}(X) \end{pmatrix} = \begin{pmatrix} \max_{u\in U}\{G(0,u) + \beta B(0,u)X\} \\ \vdots \\ \max_{u\in U}\{G(N-1,u) + \beta B(N-1,u)X\} \end{pmatrix}.$$

Nach Lemma 3.4 folgt für beliebige Vektoren $Z, W \in \mathbb{R}^N$ und alle $i = 0, \ldots, N-1$ aufgrund von $B(i,u)_k \in [0,1], k = 0, \ldots, N-1$,

$$\begin{aligned}|f_i(Z) - f_i(W)| &\leq |\max_{u\in U}\{G(i,u) + \beta B(i,u)Z\} - \max_{u\in U}\{G(i,u) + \beta B(i,u)W\}| \\ &\leq \beta \max_{u\in U} |B(i,u)Z - B(i,u)W| \\ &\leq \beta \|Z - W\|_\infty.\end{aligned}$$

Da die Ungleichung für alle $i = 0, \ldots, N-1$ gilt, ergibt sich

$$\|f(Z) - f(W)\|_\infty \leq \beta \|Z - W\|_\infty. \tag{4.8}$$

Wegen $\beta < 1$ liegt eine Kontraktion auf dem $\mathbb{R}^N$ bezüglich $\|.\|_\infty$ vor. Somit existiert ein Vektor V. Zum Nachweis der Eindeutigkeit nehmen wir an, es gebe einen weiteren Fixpunkt W. Dann erhalten wir mit (4.8)

$$\| \underbrace{f(V)}_{=V} - \underbrace{f(W)}_{=W} \|_\infty \leq \beta \|V - W\|_\infty < \|V - W\|_\infty.$$

Es liegt ein Widerspruch vor, der Vektor V ist folglich auch eindeutig.
Mehrmaliges Anwenden von (4.8) liefert

$$\begin{aligned}\|V^{j+1} - V\|_\infty = \|f(V^j) - f(V)\|_\infty &\leq \beta \|V^j - V\|_\infty = \beta \|f(V^{j-1}) - f(V)\|_\infty \\ &\leq \cdots \leq \beta^j \|V^0 - V\|_\infty = \beta^j \|V\|_\infty\end{aligned}$$

und somit

$$\|V^{j+1} - V\|_\infty \leq \beta^j \|V\|_\infty.$$

Daraus folgt die Konvergenzaussage $V^j \to V$ für $j \to \infty$.
Darüber hinaus können wir

$$\begin{aligned}\|V^j - V\|_\infty &\leq \|V^j - V^{j+1}\|_\infty + \|V^{j+1} - V\|_\infty \\ &\leq \varepsilon + \|f(V^j) - f(V)\|_\infty \\ &\leq \varepsilon + \beta \|V^j - V\|_\infty\end{aligned}$$

herleiten. Durch Umformung ergibt sich

$$\|V^j - V\|_\infty \leq \frac{\varepsilon}{1-\beta} = \frac{\varepsilon}{h\delta}.$$

Unter Verwendung der Definition von V^j und Lemma 4.3(ii) folgt damit für alle $x \in R_i$

$$\begin{aligned}\|\hat{v}_h(x) - \hat{v}_h^j(x)\| = \|\sum_{j=0}^{3} \mu_j(x)(\hat{v}_h(E_{i_j}) - \hat{v}_h^j(E_{i_j}))\| &\leq \|\hat{v}_h(E_{i_j}) - \hat{v}_h^j(E_{i_j})\| \\ = \|V_k^j - V_k\| \leq \frac{\varepsilon}{h\delta}.\end{aligned}$$

Hierbei stellt k den zu i_j gehörigen globalen Iterationsindex dar. Da die Ungleichung für alle $R_i, i = 0, \ldots, P-1$, und alle $x \in R_i$ erfüllt ist, gilt

$$\|\hat{v}_h - \hat{v}_h^j\|_\infty \leq \frac{\varepsilon}{h\delta}.$$

□

Der folgende Satz liefert eine Abschätzung für den Fehler, der durch die räumliche Diskretisierung entsteht.

Satz 4.5. *Wir betrachten das zeitdiskrete optimale Steuerungsproblem aus Definition 3.1 auf einer kompakten Rechteckmenge Ω mit Zeitschritt h. Sei v_h die zugehörige optimale Wertefunktion. Darüber hinaus betrachten wir ein Gitter Γ auf Ω mit Durchmesser k. Dann gilt für die Funktion $\hat{v}_h$ aus Lemma 4.4 die Abschätzung*

$$\|v_h - \hat{v}_h\| \leq K \left(\frac{k}{h}\right)^\gamma,$$

wobei $K > 0$ eine geeignete Konstante ist und $\gamma = 1$, wenn $\delta > L, \gamma \in (0,1)$ beliebig, falls $\delta = L$ und $\gamma = \frac{\delta}{L}$ für $\delta < L$.

Beweis. Grüne (2004), Satz 3.17 □

Kombinieren wir Satz 4.5 und Satz 4.1, so erhalten wir eine Abschätzung für den durch die räumliche und die zeitliche Diskretisierung entstandenen Fehler.

Satz 4.6. *Wir betrachten das optimale Steuerungsproblem aus Definition 3.1 auf einer kompakten Rechteckmenge Ω mit optimaler Wertefunktion v, das dazugehörige zeitdiskrete optimale Steuerungsproblem für ein $h > 0$ sowie ein Gitter Γ auf Ω mit Durchmesser k. Dann gilt für die Funktion $\hat{v}_h$ aus Lemma 4.4 die Abschätzung*

$$\|v - \hat{v}_h\|_\infty \leq K h^{\frac{\gamma}{2}} + K \left(\frac{k}{h}\right)^\gamma.$$

Hierbei ist $K > 0$ eine geeignete Konstante, $\gamma = 1$, falls $\delta > L, \gamma \in (0,1)$ beliebig, wenn $\delta = L$ und $\gamma = \frac{\delta}{L}$ für $\delta < L$.

Teil II

Die Bundle-Newton-Methode

In diesem Teil der Arbeit stellen wir das Optimierungsverfahren vor, das wir für die Lösung des Optimierungsproblems (4.7) im optimalen Steuerungsalgorithmus verwenden: die Bundle-Newton-Methode. Diese Optimierungsstrategie wurde in Luksan und Vlček (1998) für den unrestringierten Fall veröffentlicht. Sie ist für die Minimierung von nicht notwendig differenzierbaren, lokal lipschitz-stetigen Funktionen geeignet. Im konvexen Fall wird eine Folge von Punkten x_k erzeugt, die gegen das globale Minimum konvergiert, falls dieses existiert. Im nichtkonvexen Fall müssen wir uns mit der Ermittlung eines stationären Punktes zufrieden geben.
Das Verfahren gehört zu den Bundle-Methoden und nutzt die Schnittebenenidee. Das Besondere besteht darin, dass die auftretenden Funktionen im Gegensatz zu einigen anderen Varianten der Bundle-Methode nicht stückweise linear, sondern stückweise quadratisch approximiert werden, was sich unter bestimmten zusätzlichen Annahmen positiv auf das Konvergenzverhalten auswirkt. Der Umgang mit der Nichtdifferenzierbarkeit erfordert neue Konzepte. Die gewöhnliche Ableitung, die beim Gradientenverfahren der glatten Optimierung noch zum Ziel führte, ist nicht mehr ausreichend, es wird auf das Konstrukt des Subdifferentials zurückgegriffen. Im nächsten Kapitel werden wir sowohl diese Verallgemeinerung der Ableitung als auch alle anderen für die Herleitung der Bundle-Newton-Methode notwendigen Grundlagen vorstellen. Die Bundle-Methoden können als Antwort auf die Unzulänglichkeiten der Vorgängermethoden verstanden werden. Deswegen geben wir im darauffolgenden Kapitel einen Überblick über diese Vorstufen. Erst danach erfolgt die eigentliche Herleitung der Bundle-Newton-Methode.

Kapitel 5

Grundlagen

Es bedarf einiger Vorarbeit, um die Konzepte, auf denen die Bundle-Newton-Methode aufbaut, darzulegen. Nach der Darstellung wichtiger Eigenschaften konvexer und lipschitzstetiger Funktionen führen wir im Hinblick auf den Umgang mit Nichtdifferenzierbarkeitsstellen Verallgemeinerungen der Richtungsableitung und der Ableitung ein. Es schließen sich zwei Abschnitte über die notwendigen Bedingungen für restringierte und unrestringierte Probleme an. Die wichtigsten Ergebnisse der Dualitätstheorie, die in vielen Standardwerken der Optimierung ausführlich behandelt werden, sprechen wir in dieser Arbeit nur kurz an. Den letzten Abschnitt bildet die Herleitung der Lagrange-Newton-Methode, die für das Bundle-Newton-Verfahren eine entscheidende Rolle spielt. Die Aussagen dieses Kapitels sind Clarke (1983), Geiger und Kanzow (2002), Gerdts (2003) sowie Mäkela und Neittaanmäki (1992) entnommen.

5.1 Konvexe und lipschitzstetige Funktionen

Im Rahmen der Bundle-Newton-Methode betrachten wir lokal lipschitzstetige Funktionen.

Definition 5.1 (lokale Lipschitzstetigkeit). *Eine Funktion $f\colon \mathbb{R}^n \to \mathbb{R}$ heißt lokal lipschitzstetig in $x \in \mathbb{R}^n$ mit Konstante L, falls es ein $\varepsilon > 0$ gibt mit*

$$|f(y) - f(z)| \leq L\|y - z\| \text{ für alle } y, z \in U_\varepsilon(x).$$

Konvexe Funktionen stellen einen wichtigen Spezialfall lokal lipschitzstetiger Funktionen dar.

Definition 5.2. *Eine Funktion $f\colon \mathbb{R}^n \to \mathbb{R}$ heißt konvex, falls*

$$f(\lambda x_1 + (1-\lambda)x_2) \leq \lambda f(x_1) + (1-\lambda) f(x_2)$$

für alle $\lambda \in [0,1]$ und alle $x_1, x_2 \in \mathbb{R}^n$.

Satz 5.1. *Sei $f\colon \mathbb{R}^n \to \mathbb{R}$ eine konvexe Funktion. Dann ist f lokal lipschitzstetig in jedem $x \in \mathbb{R}^n$.*

Beweis. Mäkela und Neittaanmäki (1992), Theorem 2.1.2 □

Der nachstehende Satz verdeutlicht den Zusammenhang zwischen Lipschitzstetigkeit und Differenzierbarkeit:

Satz 5.2 (Rademacher). *Eine lipschitzstetige Funktion $f\colon \mathbb{R}^n \to \mathbb{R}$ ist fast überall differenzierbar.*

Beweis. Mäkela und Neittaanmäki (1992), Theorem 3.2.15 □

5.2 Richtungsableitungen

Die Richtungsableitung im gewöhnlichen Sinne ist in folgender Weise definiert:

Definition 5.3 (Richtungsdifferenzierbarkeit). *Eine Funktion $f\colon \mathbb{R}^n \to \mathbb{R}$ heißt richtungsdifferenzierbar in $x \in \mathbb{R}^n$ in Richtung $h \in \mathbb{R}^n$, falls*

$$f'(x;h) = \lim_{t \downarrow 0} \frac{f(x+th) - f(x)}{t}$$

existiert. $f'(x;h)$ heißt Richtungsableitung von f in x in Richtung h.

Ist f stetig differenzierbar in $x \in \mathbb{R}^n$, so gilt

$$f'(x;h) = \nabla f(x)^\top h, \quad \nabla f(x) = \left(\frac{\partial f}{\partial x_1}(x), \ldots, \frac{\partial f}{\partial x_n}(x) \right)^\top. \tag{5.1}$$

Die Richtungsableitung existiert auch für konvexe Funktionen.

Satz 5.3. *Sei $f\colon \mathbb{R}^n \to \mathbb{R}$ eine konvexe Funktion. Dann existiert die Richtungsableitung von f in x in jede Richtung $h \in \mathbb{R}^n$.*

Beweis. Mäkela und Neittaanmäki (1992), Theorem 2.1.3 □

Für lipschitzstetige Funktionen existiert sie allerdings nicht notwendigerweise, so dass in diesem Fall eine Verallgemeinerung der gewöhnlichen Richtungsableitung wie beispielsweise die Richtungsableitung nach Clarke Verwendung findet.

Definition 5.4 (verallgemeinerte Richtungsableitung nach Clarke). *Die Funktion $f\colon \mathbb{R}^n \to \mathbb{R}$ sei lokal lipschitzstetig in $x \in \mathbb{R}^n$. Die verallgemeinerte Richtungsableitung von f in x in Richtung $h \in \mathbb{R}^n$ ist definiert als*

$$f^\circ(x;h) = \limsup_{\substack{y \to x \\ t \downarrow 0}} \frac{f(y+th) - f(y)}{t}. \tag{5.2}$$

Diese verallgemeinerte Richtungsableitung existiert für alle lokal lipschitzstetigen Funktionen und stimmt im konvexen Fall mit der gewöhnlichen Richtungsableitung überein.

Satz 5.4. *Sei $f\colon \mathbb{R}^n \to \mathbb{R}$ eine konvexe Funktion. Dann gilt:*

$$f'(x;h) = f^\circ(x;h) \text{ für alle } h \in \mathbb{R}^n$$

Beweis. Mäkela und Neittaanmäki (1992), Theorem 3.1.8 □

Die Herleitung der Bundle-Methoden basiert auf einer Abschwächung der herkömmlichen Richtungsableitung.

Definition 5.5 (ε-Richtungsableitung für konvexe Funktionen). *Sei $f\colon \mathbb{R}^n \to \mathbb{R}$ konvex. Die ε-Richtungsableitung von f in x in Richtung $h \in \mathbb{R}^n$ ist definiert als*

$$f'_\varepsilon(x;h) = \inf_{t>0} \frac{f(x+th) - f(x) + \varepsilon}{t}.$$

5.3 Subdifferentiale

Da es sich bei den zu optimierenden Funktionen nicht notwendigerweise um differenzierbare Funktionen handelt, ist es erforderlich, eine Verallgemeinerung der Ableitung einzuführen.

Definition 5.6 (Subdifferential, verallgemeinerter Gradient). *Sei $f^*(x;h)$ eine beliebige Richtungsableitung. Die Menge*

$$\partial_* f(x) = \{\xi \in \mathbb{R}^n \mid f^*(x;h) \geq \xi^\top h \text{ für alle } h \in \mathbb{R}^n\}$$

heißt Subdifferential oder verallgemeinerter Gradient.

Für die Clarke'sche Richtungsableitung ist das Subdifferential folgendermaßen definiert:

Definition 5.7 (Subdifferential (verallgemeinerter Gradient) nach Clarke). *Sei $f\colon \mathbb{R}^n \to \mathbb{R}$ lokal lipschitzstetig in $x \in \mathbb{R}^n$. Das Subdifferential nach Clarke von f in x ist definiert als die Menge*

$$\partial_\circ f(x) = \{\xi \in \mathbb{R}^n \mid f^\circ(x;h) \geq \xi^\top h \text{ für alle } h \in \mathbb{R}^n\}.$$

Jedes Element $\xi \in \partial_\circ f(x)$ heißt Subgradient von f in x.

Wichtige Eigenschaften des Clarke'schen Subdifferentials fasst der nächste Satz zusammen.

Satz 5.5. *Sei $f\colon \mathbb{R}^n \to \mathbb{R}$ lokal lipschitzstetig. Dann gelten:*

(i) $\partial_\circ f(x)$ *ist eine nichtleere, konvexe und kompakte Menge mit* $\partial_\circ f(x) \subseteq U_L(0)$*, wobei* L *die Lipschitzkonstante von* f *in* x *bezeichnet.*

(ii) $f^\circ(x;h) = \max\{\xi^\top h \,|\, \xi \in \partial_\circ f(x)\}$ *für alle* $h \in \mathbb{R}^n$

(iii) Die Abbildung $\partial f(\cdot)$*, die jedem Punkt* $x \in \mathbb{R}^n$ *eine Menge* $A \subset \mathbb{R}^n$ *zuordnet, ist lokal beschränkt, d.h. die Beschränktheit von* $B \subset \mathbb{R}^n$ *impliziert die Beschränktheit der Menge* $\{g_f \in \partial f(y) \,|\, y \in B\}$*.*

Beweis. (i), (ii): Mäkela und Neittaanmäki (1992), Theorem 3.1.4, (iii): Kiwiel (1985), Lemma 2.2 □

Mit Hilfe des Satzes von Rademacher kann das Subdifferential nach Clarke auch folgendermaßen dargestellt werden:

Satz 5.6. *Sei* $f\colon \mathbb{R}^n \to \mathbb{R}$ *lokal lipschitzstetig in* $x \in \mathbb{R}^n$*. Dann gilt*

$$\partial_\circ f(x) = conv\{\xi \in \mathbb{R}^n \,|\, \exists (x_i) \subset \mathbb{R}^n \backslash \Omega_f \text{ mit } x_i \to x \text{ und } \nabla f(x_i) \to \xi\}.$$

Hierbei bezeichnet Ω_f *die Menge der Nichtdifferenzierbarkeitsstellen.*

Beweis. Mäkela und Neittaanmäki (1992), Theorem 3.2.16 □

Der nächste Satz drückt aus, dass das Subdifferential tatsächlich als eine Verallgemeinerung des Gradienten angesehen werden kann.

Satz 5.7. *Sei* $f\colon \mathbb{R}^n \to \mathbb{R}$ *stetig differenzierbar in* x*, dann gilt*

$$\partial_\circ f(x) = \{\nabla f(x)\}.$$

Beweis. Mäkela und Neittaanmäki (1992), Theorem 3.1.7 □

Nun definieren wir das Subdifferential für konvexe Funktionen.

Definition 5.8 (Subdifferential für konvexe Funktionen). *Sei* $f\colon \mathbb{R}^n \to \mathbb{R}$ *konvex. Das Subdifferential von* f *in* x *ist definiert als die Menge*

$$\partial_c f(x) = \{\xi \in \mathbb{R}^n \,|\, f(y) \geq f(x) + \xi^\top (y - x) \text{ für alle } y \in \mathbb{R}^n\}. \tag{5.3}$$

Jedes Element $\xi \in \partial_c f(x)$ *heißt Subgradient von* f *in* x*.*

Wir erhalten eine zu Satz 5.4 analoge Aussage:

Satz 5.8. *Sei* $f\colon \mathbb{R}^n \to \mathbb{R}$ *konvex. Dann gilt*

$$\partial_c f(x) = \{\xi \in \mathbb{R}^n \,|\, f'(x;h) \geq \xi^\top h \text{ für alle } h \in \mathbb{R}^n\} = \partial_\circ f(x)$$

Beweis. Mäkela und Neittaanmäki (1992), Theorem 2.1.5 (ii) □

Da das Clarke'sche Subdifferential und das Subdifferential für konvexe Funktionen im konvexen Fall übereinstimmen, behält die Aussage von Satz 5.5 ihre Gültigkeit.

Satz 5.9. *Sei $f\colon \mathbb{R}^n \to \mathbb{R}$ konvex. Dann gelten:*

(i) $\partial_c f(x)$ ist eine nichtleere, konvexe und kompakte Menge mit $\partial_c f(x) \subseteq U_L(0)$, wobei L die Lipschitzkonstante von f in x bezeichnet.

(ii) $f'(x;h) = \max\{\xi^\top h \,|\, \xi \in \partial_c f(x)\}$ für alle $h \in \mathbb{R}^n$

Beweis. Satz 5.9 folgt direkt aus Satz 5.5. □

Mit Hilfe der ε-Richtungsableitung kann das Subdifferential für konvexe Funktionen abgeschwächt werden.

Definition 5.9 (ε-Subdifferential für konvexe Funktionen). *Sei $\varepsilon \geq 0$ und $f\colon \mathbb{R}^n \to \mathbb{R}$ konvex. Das ε-Subdifferential von f in $x \in \mathbb{R}^n$ ist definiert als die Menge*

$$\partial_\varepsilon f(x) := \{\xi \in \mathbb{R}^n \,|\, f(y) \geq f(x) + \xi^\top (y - x) - \varepsilon \text{ für alle } y \in \mathbb{R}^n\}$$

Jedes Element $\xi \in \partial_\varepsilon f(x)$ heißt ε-Subgradient von f in x.

Auch für das ε-Subdifferential erhalten wir entsprechend zu Satz 5.5 und Satz 5.9 die Aussage:

Satz 5.10. *Sei $f\colon \mathbb{R}^n \to \mathbb{R}$ konvex. Dann gelten:*

(i) $\partial_\varepsilon f(x)$ ist nichtleer, konvex und kompakt mit $\partial_\varepsilon f(x) \subseteq U_L(0)$, wobei L die Lipschitzkonstante von f in x bezeichnet.

(ii) $f'_\varepsilon(x;h) = \max\{\xi^\top h \,|\, \xi \in \partial_\varepsilon f(x)\}$ für alle $h \in \mathbb{R}^n$

Beweis. Mäkela und Neittaanmäki (1992), Theorem 3.3.1.4 □

Das ε-Subdifferential liefert Informationen über die Subgradienten in einer Umgebung von x.

Satz 5.11. *Sei $f\colon \mathbb{R}^n \to \mathbb{R}$ konvex mit Lipschitzkonstante L in x und $\varepsilon \geq 0$. Dann gilt:*

$$\partial_c f(y) \subseteq \partial_\varepsilon f(x) \text{ für alle } y \in U_{\frac{\varepsilon}{2L}}(x)$$

Beweis. Mäkelä und Neittaanmäki (1992), Theorem 3.3.1.5 □

Der nachfolgende Satz ist von großer Bedeutung für die Herleitung der verallgemeinerten Fritz-John-Bedingungen in Kapitel 5.4.

Satz 5.12. *Die Funktionen $f_i\colon \mathbb{R}^n \to \mathbb{R}, i = 1, \ldots, m$, seien lokal lipschitzstetig in x. Dann ist auch*

$$f(x) := \max\{f_i(x) \,|\, i = 1, \ldots, m\}$$

lokal lipschitzstetig in x und es gilt

$$\partial_\circ f(x) \subseteq conv\{\partial_\circ f_i(x) \,|\, i \in I(x)\} \text{ mit } I(x) := \{i \,|\, f_i(x) = f(x), 1 \leq i \leq m\}.$$

Beweis. Mäkelä und Neittaanmäki (1992), Theorem 3.2.13 □

5.4 Notwendige Bedingungen für unrestringierte Optimierungsprobleme

In der differenzierbaren Optimierung stellt "$\nabla f(x) = 0$" eine notwendige Bedingung für ein lokales Minimum im unrestringierten Fall dar. Der folgende Satz zeigt, dass wir eine ähnliche Aussage auch für lokal lipschitzstetige, nicht notwendig differenzierbare Funktionen erhalten.

Satz 5.13. *Sei $f\colon \mathbb{R}^n \to \mathbb{R}$ lokal lipschitzstetig in x und sei x ein lokales Minimum von f. Dann gelten:*

(i) $0 \in \partial_\circ f(x)$

(ii) $f^\circ(x; h) \geq 0$ *für alle* $h \in \mathbb{R}^n$

Beweis. Mäkelä und Neittaanmäki (1992), Theorem 5.1.1, Theorem 3.2.5 □

Für konvexe Funktionen sind diese Bedingungen sogar hinreichend:

Satz 5.14. *Sei $f\colon \mathbb{R}^n \to \mathbb{R}$ konvex. Dann sind folgende Aussagen äquivalent:*

(i) f hat in x ein globales Minimum

(ii) $0 \in \partial_c f(x)$

(iii) $f'(x; h) \geq 0$ *für alle* $h \in \mathbb{R}^n$

Beweis. Mäkelä und Neittaanmäki (1992), Theorem 5.1.2 □

Eine hinreichende Bedingung für $\varepsilon-$Optimalität im konvexen Fall liefert der nächste Satz:

Satz 5.15. *Sei $f\colon \mathbb{R}^n \to \mathbb{R}$ konvex. Dann sind die folgenden Aussagen äquivalent:*

(i) $0 \in \partial_\varepsilon f(x)$

(ii) Es gilt $f(x) \leq f(y) + \varepsilon$ *für alle* $y \in \mathbb{R}^n$.

Beweis. Mäkelä und Neittaanmäki (1992), Theorem 5.1.4 □

5.5 Notwendige Bedingungen für restringierte Optimierungsprobleme

Wir betrachten das Optimierungsproblem

$$\begin{array}{ll} \min & f(x) \\ \text{u.d.N.} & F_i(x) \leq 0,\ i = 1, \ldots, m, \end{array} \tag{5.4}$$

wobei $f\colon \mathbb{R}^n \to \mathbb{R}$ und $F_i\colon \mathbb{R}^n \to \mathbb{R}$ lokal lipschitzstetige Funktionen sind. In diesem Abschnitt zeigen wir, dass die notwendigen Bedingungen für ein Minimum von Problemstellung (5.4) eine Verallgemeinerung der Fritz-John-Bedingungen bzw. KKT-Bedingungen der differenzierbaren Optimierung darstellen. Die Idee besteht darin, das restringierte Problem (5.4) in ein dazu äquivalentes unrestringiertes Problem zu transformieren. Wir definieren

$$F(x) := \max\{F_i(x) \,|\, i = 1, \ldots, m\} \text{ für alle } x \in \mathbb{R}^n$$

sowie die Improvementfunktion in x

$$H(y, x) := \max\{f(y) - f(x), F(y)\} \text{ für alle } y \in \mathbb{R}^n.$$

Die lokale Lipschitzstetigkeit der Funktionen $f_i, i = 1, \ldots, m$ impliziert nach Satz 5.12 auch die lokale Lipschitzstetigkeit von F und demzufolge von H. Sei $\hat{x}$ ein lokales Minimum des Optimierungsproblems (5.4). Dann stellt $\hat{x}$ auch ein lokales Minimum des unrestringierten Problems $\min H(\cdot, \hat{x})$ mit $H(\hat{x}, \hat{x}) = 0$ dar. Andernfalls könnte man in einer Umgebung von $\hat{x}$ ein x mit $H(x, \hat{x}) = \max\{f(x) - f(\hat{x}), F(x)\} < 0$ finden. Daher läge in x ein zulässiger Punkt mit $f(x) < f(\hat{x})$ vor, was der Annahme widerspricht. Nach Satz 5.13 gilt somit

$$0 \in \partial_\circ H(\hat{x}, \hat{x}).$$

Wenden wir Satz 5.12 an, so folgt

$$0 \in \partial_\circ H(\hat{x}, \hat{x}) \subseteq \begin{cases} \partial_\circ f(\hat{x}), & \text{falls } F(\hat{x}) < 0 \\ conv\{\partial_\circ f(\hat{x}) \cup \partial_\circ F(\hat{x})\}, & \text{falls } F(\hat{x}) = 0. \end{cases}$$

Daraus ergibt sich wiederum nach Satz 5.12

$$0 \in \begin{cases} \partial_\circ f(\hat{x}), & \text{falls } F(\hat{x}) < 0, \\ conv\{\partial_\circ f(\hat{x}) \cup conv\{\partial_\circ F_i(\hat{x}) \,|\, i \in I(\hat{x})\}\}, & \text{falls } F(\hat{x}) = 0, \end{cases} \tag{5.5}$$

wobei *conv* die konvexe Hülle bezeichnet. Wir wissen außerdem aufgrund von Satz 5.5, dass $\partial_\circ f(\hat{x})$ und $\partial_\circ F_i(\hat{x}), i = 1, \ldots, m$ nichtleere konvexe Mengen sind. Für nichtleere konvexe Mengen $C_1, \ldots, C_k$ gilt die Beziehung

$$conv(C_1 \cup \cdots \cup C_k) = \left\{ \sum_{i=1}^{k} \lambda_i C_i \,|\, \lambda_i \geq 0,\ i = 1, \ldots, k,\ \sum_{i=1}^{k} \lambda_i = 1 \right\}.$$

Somit ist Aussage (5.5) gleichbedeutend mit: Es existieren Multiplikatoren $\eta_i \geq 0$, $i = 0, 1, \ldots, m$ mit

$$0 \in \eta_0 \partial_\circ f(\hat{x}) + \sum_{i=1}^{m} \eta_i \partial_\circ F_i(\hat{x}),$$
$$\sum_{i=0}^{m} \eta_i = 1,$$
$$\eta_i F_i(\hat{x}) = 0,\ i = 1, 2, \ldots, m.$$

Dies kann leicht nachgewiesen werden:
Aus $F(\hat{x}) < 0$ folgt nach (5.5) $0 \in \partial_0 f(\hat{x})$. Anders ausgedrückt erhalten wir: $0 \in \eta_0 \partial_\circ f(\hat{x}) + \sum_{i=1}^{m} \eta_i \partial_\circ F_i(\hat{x})$ mit $\eta_0 = 1$ und $\eta_i = 0$ für $i = 1, \ldots, m$. Somit gilt $\eta_i F_i(\hat{x}) = 0, i = 1, 2, \ldots, m$. Im Fall $F(\hat{x}) = 0$ gilt nach (5.5) $0 \in conv\{\partial_\circ f(\hat{x}) \cup conv\{\partial_\circ F_i(\hat{x}) \,|\, i \in I(\hat{x})\}\}$. Dies kann auch geschrieben werden als $0 \in \eta_0 \partial_\circ f(\hat{x}) + \sum_{i=1}^{m} \eta_i \partial_\circ F_i(\hat{x})$ mit $\eta_0 \geq 0$, $\eta_i \geq 0$ für $i = I(\hat{x})$, $\eta_0 + \sum_{i \in I(\hat{x})} \eta_i = 1$ sowie $\eta_i = 0$ für $i \in I(\hat{x})$. Die Bedingung $\eta_i F_i(\hat{x}) = 0, i = 1, 2, \ldots, m$ ist also erfüllt. Zusammenfassend erhalten wir das folgende Ergebnis:

Satz 5.16 (verallgemeinerte Fritz-John-Bedingungen). *Sei $\hat{x}$ ein lokales Minimum des Optimierungsproblems (5.4). Dann existieren Multiplikatoren $\eta_i \geq 0, i = 0, 1, \ldots, m$ mit*

$$0 \in \eta_0 \partial_\circ f(\hat{x}) + \sum_{i=1}^{m} \eta_i \partial_\circ F_i(\hat{x}),$$
$$\sum_{i=0}^{m} \eta_i = 1,$$
$$\eta_i F_i(\hat{x}) = 0,\ i = 1, 2, \ldots, m.$$

Analog zum differenzierbaren Fall kann η_0 auch hier den Wert Null annehmen, wodurch die Informationen über die Zielfunktion verloren gehen. Wir sind deswegen an Bedingungen interessiert, unter denen $\eta_0 \neq 0$ gewählt werden kann.

Definition 5.10 (Cottle-Regularitätsbedingung). *F erfüllt die Cottle-Regularitätsbedingung in x, falls*

$$F(x) < 0 \text{ oder } 0 \notin \partial_\circ F(x).$$

Ist die Cottle-Regularitätsbedingung im lokalen Minimum $\hat{x}$ erfüllt, kann in den Fritz-John-Bedingungen o.B.d.A. $\eta_0 = 1$ gesetzt werden und wir erhalten die KKT-Bedingungen.

Satz 5.17 (verallgemeinerte KKT-Bedingungen). *Sei $\hat{x}$ ein lokales Minimum des Optimierungsproblems (5.4). In $\hat{x}$ gelte die Cottle-Regularitätsbedingung. Dann existieren Multiplikatoren $\eta_i \geq 0, i = 1, \ldots, m$ mit*

$$0 \in \partial_\circ f(\hat{x}) + \sum_{i=1}^{m} \eta_i \partial_\circ F_i(\hat{x}),$$
$$\eta_i F_i(\hat{x}) = 0,\ i = 1, 2, \ldots, m.$$

Beweis. Wir betrachten zunächst den Fall $F(\hat{x}) < 0$. Es gilt also $F_i(\hat{x}) < 0$ für alle $i = 1, \ldots, m$ und aufgrund von (5.5) $0 \in \partial_\circ f(\hat{x})$. Daher können wir $\eta_i = 0, i = 1, \ldots, m$ und $\eta_0 = 1$ setzen.
Nun nehmen wir $0 \notin \partial_\circ F(\hat{x})$ und $F(\hat{x}) = 0$ an. Nach (5.5) existiert ein Multiplikator $\mu_0 \in [0, 1]$ mit $0 \in \mu_0 \partial_\circ f(\hat{x}) + (1 - \mu_0)\partial_\circ F(\hat{x})$. Wegen der Voraussetzung $0 \notin \partial_\circ F(\hat{x})$ können wir $\mu_0 = 0$ ausschließen. Somit gilt $\mu_0 > 0$. Aufgrund von Satz 5.12 erhalten wir außerdem

$$0 \in \mu_0 \partial_\circ f(\hat{x}) + (1 - \mu_0)\partial_\circ F(\hat{x}) \subseteq \mu_0 \partial_\circ f(\hat{x}) + (1 - \mu_0) conv\{\partial_\circ F_i(\hat{x}) \,|\, i = I(\hat{x})\}.$$

Daher gibt es Multiplikatoren $\mu_i \geq 0, i \in I(\hat{x}), \sum_{i \in I(\hat{x})} \mu_i = 1$ mit

$$0 \in \mu_0 \partial_\circ f(\hat{x}) + (1 - \mu_0) \sum_{i \in I(\hat{x})} \mu_i \partial_\circ F_i(\hat{x}).$$

Da $\mu_0 \neq 0$, ist die Division durch μ_0 möglich. Mit $\eta_i := (1 - \mu_0)\mu_i/\mu_0 \geq 0$ für $i \in I(\hat{x})$ und $\eta_i = 0$ für $i \notin I(\hat{x})$ folgt $0 \in \partial_\circ f(\hat{x}) + \sum_{i=1}^{m} \eta_i \partial_\circ F_i(\hat{x})$. □

Wenn konvexe Restriktionsfunktionen vorliegen, besteht eine Verbindung der Cottle-Regularitätsbedingung zur Slater-Bedingung:

Satz 5.18. *Die Funktionen $F_i, i = 1, \ldots, m$ seien konvex. Dann ist die Cottle-Regularitätsbedingung in einem zulässigen Punkt x äquivalent zur Slater-Bedingung*

$$\exists\, y \in \mathbb{R}^n : F(y) < 0. \tag{5.6}$$

Beweis. Mäkelä und Neittaanmäki (1992), Theorem 5.3.4 □

Der nächste Satz besagt, dass die KKT-Bedingungen hinreichend sind für das konvexe restringierte Optimierungsproblem

$$\begin{array}{ll} \min & f(x) \\ \text{u.d.N.} & F_i(x) \leq 0,\ i = 1, \ldots, m, \end{array} \tag{5.7}$$

wobei $f \colon \mathbb{R}^n \to \mathbb{R}$ und $F_i \colon \mathbb{R}^n \to \mathbb{R}$ konvexe Funktionen darstellen.

Satz 5.19 (verallgemeinerte KKT-Bedingungen im konvexen Fall). *Wir betrachten Problemstellung (5.7). Dann sind die folgenden Aussagen äquivalent:*

(i) $\hat{x}$ ist zulässig für (5.7) und es existieren Multiplikatoren $\eta_i \geq 0, i = 0, 1, \ldots, m$ mit

$$0 \in \partial_c f(\hat{x}) + \sum_{i=1}^{m} \eta_i \partial_c F_i(\hat{x}),$$
$$\eta_i F_i(\hat{x}) = 0,\ i = 1, 2, \ldots, m.$$

(ii) $\hat{x}$ ist globale Minimalstelle von (5.7).

Beweis. $(ii) \Rightarrow (i)$: nach Satz 5.17
$(i) \Rightarrow (ii)$:
Aufgrund von $0 \in \partial_c f(\hat{x}) + \sum_{i=1}^{m} \eta_i \partial_c g_i(\hat{x})$ gibt es Subgradienten $\xi \in \partial_c f(\hat{x})$ und $\xi_i \in \partial_c g_i(\hat{x})$ mit

$$\xi + \sum_{i=1}^{m} \eta_i \xi_i = 0. \tag{5.8}$$

Nach Definition der Subgradienten im konvexen Fall gilt

$$f(y) \geq f(\hat{x}) + \xi^\top (y - \hat{x}),\ y \in \mathbb{R}^n$$
$$F_i(y) \geq F_i(\hat{x}) + \xi_i^\top (y - \hat{x}),\ y \in \mathbb{R}^n,\ i = 1, \ldots, m.$$

Multipliziert man die zweite Ungleichung mit $\eta_i \geq 0$, summiert das Ergebnis über alle $i = 1, \ldots, m$ auf und addiert anschließend die erste Ungleichung, so ergibt sich

$$f(y) + \sum_{i=1}^{m} \eta_i F_i(y) \geq f(\hat{x}) + \sum_{i=1}^{m} \eta_i F_i(\hat{x}) + (y - \hat{x})^\top (\xi + \sum_{i=1}^{m} \eta_i \xi_i),\ y \in \mathbb{R}^n. \tag{5.9}$$

Mit $\eta_i F_i(\hat{x}) = 0$, (5.8) und der Tatsache, dass für zulässige Punkte y die Ungleichung $\eta_i F_i(y) \leq 0$ gilt, können wir mittels (5.9) auf

$$f(y) \geq f(\hat{x}) \text{ für alle } y \text{ mit } F(y) \leq 0$$

schließen. □

5.6 Dualität

Da bei der Herleitung der Bundle-Methoden der starke Dualitätssatz zur Anwendung kommt, soll dieser im vorliegenden Abschnitt in Erinnerung gerufen werden. Wir betrachten ein

allgemeines Optimierungsproblem der Form

$$\begin{array}{ll} \min & f(x) \\ \text{u.d.N.} & g_i(x) \leq 0 \text{ für } i = 1, \ldots, m, \\ & h_j(x) = 0 \text{ für } j = 1, \ldots, p, \\ & x \in X, \end{array}$$

welches primales Problem genannt wird. Dabei seien $f\colon \mathbb{R}^n \to \mathbb{R}$, $g_i\colon \mathbb{R}^n \to \mathbb{R}$ und $h_j\colon \mathbb{R}^n \to \mathbb{R}$ nicht notwendig konvexe, nicht notwendig differenzierbare beliebige Funktionen und $X \subset \mathbb{R}^n$ eine beliebige nichtleere Menge. Zu dieser Problemstellung können wir ein dazugehöriges Problem konstruieren, das sogenannte duale Problem

$$\begin{array}{ll} \max & \theta(\lambda, \mu) \\ \text{u.d.N.} & \lambda \geq 0. \end{array} \tag{5.10}$$

Hierbei ist die duale Zielfunktion durch

$$\theta(\lambda, \mu) := \inf_{x \in X} L(x, \lambda, \mu)$$

und die Lagrangefunktion durch

$$L(x, \lambda, \mu) := f(x) + \lambda^\top g(x) + \mu^\top h(x).$$

mit $\lambda \in \mathbb{R}^m$ und $\mu \in \mathbb{R}^p$ gegeben. Die Dualitätssätze beschreiben den Zusammenhang zwischen dem primalen und dem dualen Problem. Der optimale Zielfunktionswert des primalen Problems stimmt nicht immer mit dem des dualen Problems überein, allerdings gilt nach dem sogenannten schwachen Dualitätssatz, dass der Optimalwert des Dualproblems eine untere Schranke für den Optimalwert des primalen Problems darstellt. Das Primalproblem kann indirekt über die Lösung des dualen Problems gelöst werden, wenn keine Dualitätslücke auftritt, d.h. wenn der optimale Zielfunktionswert des primalen Problems mit dem optimalen Zielfunktionswert des dualen Problems übereinstimmt. Hinreichende Bedingungen dafür werden im starken Dualitätssatz formuliert:

Satz 5.20 (starker Dualitätssatz). *Die Menge $X \subseteq \mathbb{R}^n$ sei nichtleer und konvex. Die Funktionen f und $g_i, i = 1, \ldots, m$ seien konvex, die Funktionen $h_j, j = 1, \ldots, p$ seien affin linear. Die optimale Lösung des primalen Problems sei endlich und es gebe ein y aus dem relativen Inneren von X mit*

$$\begin{array}{l} g_i(y) < 0 \textit{ für } i = 1, \ldots, m \\ h_i(y) = 0 \textit{ für } j = 1, \ldots, p. \end{array} \tag{5.11}$$

Dann ist das duale Problem lösbar und es gilt:

$$\inf\{f(x) \,|\, x \in X, g(x) \leq 0, h(x) = 0\} = \sup\{\theta(\lambda, \mu) \,|\, \lambda \geq 0\}$$

Beweis. Geiger und Kanzow (2002), Satz 6.13 □

Bemerkung 5.21. *Die Voraussetzung (5.11) wird nicht benötigt, wenn wir fordern, dass die Funktionen g_i affin linear sind und die Menge X durch endlich viele lineare Ungleichungen beschrieben wird. Ein Beweis hierfür findet sich in Bertsekas (1999). Der genannte Spezialfall ist für die Anwendung der Dualitätstheorie in Kapitel 7.1.3 relevant.*

5.7 Herleitung des SQP-Verfahrens

In diesem Abschnitt werden wir das SQP-Verfahren vorstellen, das wir für die Bestimmung der Suchrichtung heranziehen. Das SQP-Verfahren ist für die Lösung von stetig differenzierbaren restringierten Optimierungsproblemen geeignet und stützt sich auf die Ideen des Newton- sowie des Lagrange-Newton-Verfahrens, die wir in den folgenden Unterabschnitten vor der eigentlichen Herleitung des SQP-Verfahrens vorstellen.

5.7.1 Newton-Verfahren

Zunächst skizzieren wir das Newton-Verfahren. Es kann zum Auffinden einer Lösung $x^* \in \mathbb{R}^n$ des Gleichungssystems

$$F(x) = 0,$$

wobei $F : \mathbb{R}^n \to \mathbb{R}^n$ eine stetig differenzierbare Abbildung darstellt, verwendet werden. In der k-ten Iteration sei eine Nährung x_k von x^* vorhanden. Wir ersetzen $F(x)$ durch die Linearisierung

$$F_k(x) := F(x_k) + F'(x_k)(x - x_k)$$

der Funktion F um x_k. Die Lösung des linearen Gleichungssystems

$$F_k(x_{k+1}) = F(x_k) + F'(x_k)(x - x_k) = 0$$

liefert den neuen Iterationspunkt

$$x_{k+1} = x_k - F'(x_k)^{-1}F(x_k),$$

falls die Inverse $F'(x_k)^{-1}$ existiert. Allerdings wird die Inverse in der Praxis nicht direkt berechnet. Vielmehr bestimmen wir einen Korrekturvektor d_k über das lineare Gleichungssystem

$$F'(x_k)d = -F(x_k)$$

und setzen

$$x_{k+1} := x_k + d_k.$$

Insgesamt erhalten wir den folgenden

Newton-Algorithmus:

S0 Wähle $x_0 \in \mathbb{R}^n$ und setze $k := 0$.

S1 Ist $F(x_k) = 0$: STOPP.

S2 Bestimme $d_k \in \mathbb{R}^n$ über die Lösung des Gleichungssystems

$$F'(x_k)d = -F(x_k).$$

S3 Setze $x_{k+1} := x_k + d_k$, erhöhe k um 1 und gehe zu **S0**.

Das Newton-Verfahren können wir zur Lösung von unrestringierten Optimierungsproblemen

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

mit einer zweimal stetig differenzierbaren Funktion $f\colon \mathbb{R}^n \to \mathbb{R}$ benutzen, indem wir es auf

$$\nabla f(x) = 0$$

anwenden. Der Algorithmus berechnet den neuen Iterationspunkt in diesem Fall über die Vorschrift

$$x_{k+1} := x_k - \nabla^2 f(x_k)^{-1} \nabla f(x_k).$$

Diese Idee wird beim Lagrange-Newton-Verfahren auf gleichsheitrestringierte Optimierungsprobleme übertragen.

5.7.2 Lagrange-Newton-Verfahren

Wir betrachten nun die Problemstellung

$$\begin{array}{ll} \min & f(x) \\ \text{u.d.N.} & h_j(x) = 0,\ j = 1, \ldots, p, \end{array}$$

wobei $f\colon \mathbb{R}^n \to \mathbb{R}$ und $h_j\colon \mathbb{R}^n \to \mathbb{R}, j = 1, \ldots, p$ zweimal stetig differenzierbare Funktionen sind. Die zugehörigen KKT-Bedingungen sind durch

$$\Phi(x, \mu) := \begin{pmatrix} \nabla_x L(x, \mu) \\ h(x) \end{pmatrix} = 0 \tag{5.12}$$

mit der Lagrangefunktion

$$L(x,\mu) := f(x) + \sum_{j=1}^{p} \mu_j h_j(x).$$

gegeben. Dabei ist $\mu := (\mu_1, \ldots, \mu_p)$. Des Weiteren setzen wir $h := (h_1, \ldots, h_p)$. Gehen wir zur Bestimmung der Lösung von (5.12) analog zum Newton-Verfahren vor, so erhalten wir den folgenden

Lagrange-Newton-Algorithmus:

S0 Wähle $(x_0, \mu_0) \in \mathbb{R}^n \times \mathbb{R}^p$ und setze $k := 0$.

S1 Ist $\Phi(x_k, \mu_k) = 0$: STOPP.

S2 Bestimme $(\Delta x_k, \Delta\mu_k) \in \mathbb{R}^n \times \mathbb{R}^p$ über die Lösung des Gleichungssystems

$$\Phi'(x_k, \mu_k) \begin{pmatrix} \Delta x \\ \Delta\mu \end{pmatrix} = -\Phi(x_k, \mu_k). \tag{5.13}$$

S3 Setze $(x_{k+1}, \mu_{k+1}) := (x_k, \mu_k) + (\Delta x_k, \Delta\mu_k)$, erhöhe k um 1 und gehe zu **S1**.

Im nächsten Unterabschnitt, der das sogenannte SQP-Verfahren behandelt, wird neben dem Newton-Verfahren auch die spezielle Struktur von Φ ausgenutzt.

5.7.3 SQP-Verfahren

Das lineare Gleichungssystem (5.13) kann auch geschrieben werden als

$$\begin{aligned} &\nabla^2_{xx} L(x_k, \mu_k)\Delta x + h'(x_k)^\top \Delta\mu = -\nabla_x L(x_k, \mu_k), \\ &\nabla h_j(x_k)^\top \Delta x = -h_j(x_k) \text{ für } j = 1, \ldots, p. \end{aligned} \tag{5.14}$$

Mit $\mu^+ := \mu_k + \Delta\mu$ lässt sich das System (5.14) umformen zu

$$\begin{aligned} &\nabla^2_{xx} L(x_k, \mu_k)\Delta x + h'(x_k)^\top \mu^+ = -\nabla f(x_k), \\ &\nabla h_j(x_k)^\top \Delta x = -h_j(x_k) \text{ für } j = 1, \ldots, p. \end{aligned}$$

Diese Gleichungen können als KKT-Bedingungen des quadratischen Optimierungsproblems

$$\begin{aligned} \min \quad & \nabla f(x_k)^\top \Delta x + \tfrac{1}{2}\Delta x^\top \nabla^2_{xx} L(x_k, \mu_k)\Delta x \\ \text{u.d.N.} \quad & h_j(x_k) + \nabla h_j(x_k)^\top \Delta x = 0 \text{ für } j = 1, \ldots, p \end{aligned}$$

interpretiert werden, welches äquivalent ist zu

$$\begin{array}{ll} \min & f(x_k) + \nabla f(x_k)^\top \Delta x + \frac{1}{2}\Delta x^\top \nabla^2_{xx} L(x_k, \mu_k)\Delta x \\ \text{u.d.N.} & h_j(x_k) + \nabla h_j(x_k)^\top \Delta x = 0 \text{ für } j = 1, \ldots, p. \end{array}$$

Diese Erkenntnis legt es nahe, das quadratische Teilproblem

$$\begin{array}{ll} \min & f(x_k) + \nabla f(x_k)^\top \Delta x + \frac{1}{2}\Delta x^\top \nabla^2_{xx} L(x_k, \mu_k)\Delta x \\ \text{u.d.N.} & g_i(x_k) + \nabla g_i(x_k)^\top \Delta x \leq 0 \text{ für } i = 1, \ldots, m, \\ & h_j(x_k) + \nabla h_j(x_k)^\top \Delta x = 0 \text{ für } j = 1, \ldots, p \end{array} \tag{5.15}$$

zur Bestimmung von $x_{k+1} := x_k + \Delta x_k$ für das Optimierungsproblem

$$\begin{array}{ll} \min & f(x) \\ \text{u.d.N.} & g_i(x) \leq 0 \text{ für } i = 1, \ldots, m, \\ & h_j(x) = 0 \text{ für } j = 1, \ldots, p \end{array} \tag{5.16}$$

heranzuziehen, wobei $f\colon \mathbb{R}^n \to \mathbb{R}$, $g_i\colon \mathbb{R}^n \to \mathbb{R}, i = 1, \ldots, m$ und $h_j\colon \mathbb{R}^n \to \mathbb{R}, j = 1, \ldots, p$ stetig differenzierbare Funktionen darstellen. Die Lagrangefunktion für dieses Problem ist gegeben durch $L(x, \lambda, \mu) := f(x) + \sum_{j=1}^m \lambda_j g_j(x) + \sum_{j=1}^p \mu_j h_j(x)$ mit $\lambda := (\lambda_1, \ldots, \lambda_m)$. Diese Idee führt zum folgenden

SQL-Algorithmus:

S0 Wähle $(x_0, \mu_0, \lambda_0) \in \mathbb{R}^n \times \mathbb{R}^m \times \mathbb{R}^p$ und setze $k := 0$.

S1 Ist (x_k, μ_k, λ_k) ein KKT-Punkt von (5.16): STOPP.

S2 Berechne eine Lösung $\Delta x_k \in \mathbb{R}^n$ des Problems (5.15) und ermittle die zugehörigen Lagrange-Multiplikatoren λ_{k+1} und μ_{k+1}.

S3 Setze $x_{k+1} := x_k + \Delta x_k$, erhöhe k um 1 und gehe zu **S1**.

Da in jedem Iterationsschritt ein quadratisches Problem gelöst wird, nennt man die soeben beschriebene Methode SQP-Verfahren (*Sequential Quadratic Programming*).

Der folgende Satz zeigt, unter welchen Bedingungen das Verfahren superlinear bzw. sogar quadratisch konvergiert.

Satz 5.22. *Sei $(x^*, \lambda^*, \mu^*) \in (\mathbb{R}^n, \mathbb{R}^m, \mathbb{R}^p)$ ein KKT-Punkt von (5.16) mit den folgenden Eigenschaften:*

(a) Es ist $g_i(x^*) + \lambda_i^* \neq 0$ für alle $i = 1, \ldots, m$.

(b) *Die Gradienten* $\nabla h_j(x^*), j = 1, \ldots, p$ *und* $\nabla g_i(x^*), i \in I(x^*)$ *sind linear unabhängig, wobei* $I(x^*) = \{i \mid g_i(x^*) = 0\}$.

(c) *Es gilt* $d^\top \nabla^2_{xx} L(x^*, \lambda^*, \mu^*) d > 0$ *für alle* $d \neq 0$ *mit* $\nabla h_j(x^*)^\top d = 0, j = 1, \ldots, p$ *und* $\nabla g_i(x^*)^\top d = 0, i \in I(x^*)$.

Dann existiert ein $\varepsilon > 0$*, so dass für jeden Startvektor* $(x_0, \lambda_0, \mu_0) \in \mathcal{U}_\varepsilon(x^*, \lambda^*, \mu^*)$ *und jede durch den SQL-Algorithmus erzeugte Folge* $(x_k, \lambda_k, \mu_k)_{k \in \mathbb{N}}$ *gilt:*
Die Folge $(x_k, \lambda_k, \mu_k)_{k \in \mathbb{N}}$ *konvergiert gegen* (x^*, λ^*, μ^*) *mit superlinearer Konvergenzrate, d.h. es existiert eine Nullfolge* $c_k \subseteq \mathbb{R}^+$ *mit*

$$\|(x_{k+1}, \lambda_{k+1}, \mu_{k+1}) - (x^*, \lambda^*, \mu^*)\| \leq c_k \|(x_k, \lambda_k, \mu_k) - (x^*, \lambda^*, \mu^*)\| \text{ für alle } k \in \mathbb{N}.$$

Falls $\nabla^2 f, \nabla^2 g_i, i = 1, \ldots, m$ *und* $\nabla^2 h_j, j = 1, \ldots, p$ *lokal lipschitzstetig sind, so ist die Konvergenzrate sogar quadratisch, d.h. es existiert eine Konstante* $K > 0$ *mit*

$$\|(x_{k+1}, \lambda_{k+1}, \mu_{k+1}) - (x^*, \lambda^*, \mu^*)\| \leq K \|(x_k, \lambda_k, \mu_k) - (x^*, \lambda^*, \mu^*)\|^2 \text{ für alle } k \in \mathbb{N}.$$

Beweis. Geiger und Kanzow (2002), Satz 5.31 □

Bemerkung 5.23. *Es ist zu beachten, dass die Hessematrix auch durch eine positiv definite Näherung dieser ersetzt werden kann. Es ist möglich, dass dies zu einem Verlust der superlinearen bzw. quadratischen Konvergenz führt, allerdings wird die Existenz einer eindeutigen Lösung des quadratischen Problems garantiert.*

Kapitel 6

Vorgängermethoden der Bundle-Newton-Methode

Im Folgenden möchten wir die wesentlichen Ideen aufzeigen, die ausgehend von der Gradientenmethode über die Subgradienten-Methode und die ε-Subgradientenmethode zu den Bundle-Methoden geführt haben. Zur Darstellung der konzeptionellen Struktur dieser iterativen Verfahren betrachten wir das allgemeine unrestringierte Problem

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

Auf die speziellen Eigenschaften der Zielfunktion gehen wir bei der Darstellung der einzelnen Verfahren ein.

Konzeptioneller Algorithmus:

S0 Initialisierung
Wähle einen Startpunkt $x_1 \in \mathbb{R}^n$ und setze $k := 0$.

S1 Bestimmung der Suchrichtung
Ermittle eine Abstiegsrichtung $d_k \in \mathbb{R}^n$ mit $f(x_k + td_k) < f(x_k)$ für irgendein $t > 0$.

S2 Abbruchkriterium
Wenn x_k sich nahe genug an der Lösung befindet: STOPP.

S3 Schrittweitenbestimmung
Ermittle eine Schrittweite $t_k > 0$, für die $t_k \approx \arg\min_{t>0}\{f(x_k + td_k)\}$ gilt.

S4 Update
Setze $x_{k+1} := x_k + t_k d_k$, erhöhe k um 1 und gehe zu Schritt **S1**.

Wir werden die Methoden nicht vollständig erläutern, sondern nur auf die Hauptunterscheidungsmerkmale, die vor allem in der Bestimmung der Suchrichtung und der Auswahl eines Abbruchkriteriums zu sehen sind, eingehen. Zu Beginn möchten wir das Gradientenverfahren skizzieren, um die Verbindung der nichtglatten Optimierung zur glatten Optimierung aufzuzeigen.

6.1 Gradientenverfahren

Die Anwendung des Gradientenverfahrens setzt eine zweimal stetig differenzierbare Zielfunktion voraus. Wenn x_k keinen lokalen Optimalpunkt darstellt, ist $\nabla f(x_k) \neq 0$ und es gibt eine Richtung $d \in \mathbb{R}^n$ mit $f'(x_k, d) < 0$. Um die Richtung des steilsten Abstiegs zu finden, muss das Optimierungsproblem

$$\min_{d \in \mathbb{R}^n, \|d\|=1} f'(x_k; d) \tag{6.1}$$

mit $f'(x_k; d) = \nabla f(x_k)^\top d$ gelöst werden. Als Ergebnis erhalten wir

$$d_k = \frac{-\nabla f(x_k)}{\|\nabla f(x_k)\|}.$$

Aufgrund der Stetigkeit des Gradienten kann die Bedingung $\|\nabla f(x_k)\| \leq \varepsilon$ als Abbruchkriterium herangezogen werden, wobei $\varepsilon \geq 0$ eine vorgebene Toleranz ist.

Bei den folgenden auf Subgradienten basierenden Verfahren setzen wir voraus, dass wir für einen gegebenen Punkt irgendeinen beliebigen Subgradienten aus dem Subdifferential berechnen können.

6.2 Subgradientenverfahren

Beim Subgradientenverfahren wird die Idee des Gradientenverfahrens auf konvexe, nicht notwendig differenzierbare Funktionen übertragen. Im nichtdifferenzierbaren Fall erweist sich das Konstrukt des Subdifferentials zwar prinzipiell als geeigneter Ersatz für den Gradienten, allerdings ruft das pauschale Ersetzen der Ableitung im Gradientenverfahren durch einen beliebigen Subgradienten ξ_k in der Weise, dass sich die Suchrichtung nun als $d_k = -\xi_k/\|\xi_k\|$ ergibt und "$\|\xi_k\| \leq 0$" als Abbruchkriterium verwendet wird, Probleme hervor. Denn zum einen sagt die Tatsache, dass "$\|\xi_k\| > 0$" gilt, aufgrund der willkürlichen Auswahl des Subgradienten noch lange nicht aus, dass das Abbruchkriterium nicht doch für einen anderen Subgradienten des Subdifferentials erfüllt ist. Zum anderen kann wegen der Unstetigkeit des Gradienten in einer Nichtdifferenzierbarkeitsstelle nicht davon ausgegangen werden, dass die

Gradienten bei Annäherung an eine optimale Nichtdifferenzierbarkeitsstelle gegen Null streben. Ein einfaches Beispiel hierfür stellt die Funktion $f(x) = |x|$ dar. Im Punkt 0 liegt zwar ein Minimum vor, in allen Punkten $x_k \neq 0$ gilt aber $|\nabla f(x_k)| = 1$. Des Weiteren stellt ein beliebiger Subgradient nicht unbedingt eine Abstiegsrichtung dar.

6.3 $\varepsilon-$Subgradientenverfahren

Das $\varepsilon-$Subgradientenverfahren orientiert sich ebenfalls am Gradientenverfahren, wählt aber eine andere Herangehensweise als die zuvor erläuterte Methode. Um auch im konvexen, nichtdiffererenzierbaren Fall eine Abstiegsrichtung zu finden, wird analog zu (6.1) das Problem

$$\min_{d\in\mathbb{R}^n,\|d\|\leq 1} f'(x^k;d) \tag{6.2}$$

gelöst. Die Richtungsableitung existiert gemäß Satz 5.3. Nach Satz 5.9 (ii) gilt für die Richtungsableitung im konvexen Fall

$$f'(x_k;d) = \max_{\xi\in\partial_c f(x_k)} \xi^\top d.$$

Damit wird (6.2) zu

$$\min_{d\in\mathbb{R}^n,\|d\|\leq 1} \max_{\xi\in\partial_c f(x_k)} \xi^\top d.$$

Da die Mengen $\{d \,|\, \|d\| \leq 1\}$ und $\partial_c f(x_k)$ nichtleer, konvex und kompakt sind, können wir min und max vertauschen und erhalten das äquivalente Problem

$$\max_{\xi\in\partial_c f(x_k)} \min_{d\in\mathbb{R}^n,\|d\|\leq 1} \xi^\top d. \tag{6.3}$$

Für gegebenes ξ hat das innere Problem die Lösung $d = -\xi/\|\xi\|$. Somit kann (6.3) umgeformt werden zu $\max_{\xi\in\partial_c f(x_k)} -\|\xi\|$ bzw.

$$\min_{\xi\in\partial_c f(x_k)} \|\xi\|. \tag{6.4}$$

Die Aufgabe besteht also darin, einen Subgradienten $\xi_k \in \partial_c f(x_k)$ mit minimaler euklidischer Norm zu finden. Da $\partial_c f(x_k)$ gemäß Satz 5.9 nichtleer, konvex und kompakt ist, existiert eine Lösung ξ_k des obigen Problems. Als optimale Suchrichtung erhalten wir somit $d_k = -\xi_k/\|\xi_k\|$. Das Verfahren führt trotz der Bestimmung einer Abstiegsrichtung in manchen Fällen zu falschen Ergebnisse, weil neben dem Subdifferential im aktuellen Punkt auch noch Informationen über die Subgradienten in Nachbarpunkten notwendig sind (siehe Gerdts (2003), Beispiel 4.5). Man greift deswegen auf das $\varepsilon-$Subdifferential zurück, das nach

Satz 5.11 Informationen über die Subgradienten in einer gewissen Umgebung von x_k enthält. Die Abstiegsrichtung erhalten wir demzufolge, indem wir $\xi_k \in \partial_\varepsilon f(x_k)$ mit

$$\|\xi_k\| = \min_{\xi \in \partial_\varepsilon f(x_k)} \|\xi\| \tag{6.5}$$

berechnen und $d_k = -\xi_k/\|\xi_k\|$ setzen. Da darüber hinaus der optimale Subgradient derjenige mit der kleinsten Norm ist, stellt $\|\xi_k\| \leq \varepsilon$ im Gegensatz zum Subgradientenverfahren entsprechend Satz 5.15 ein sinnvolles Abbruchkriterium dar. Unter der Voraussetzung der Endlichkeit des optimalen Funktionswertes kann gezeigt werden, dass das $\varepsilon-$Subgradientenverfahren nach endlich vielen Schritten in einem $\varepsilon-$optimalen Punkt x_k (d.h. $f(x_k) \leq f(y) + \varepsilon$ für alle $y \in \mathbb{R}^n$) abbricht. Für die Lösung des Problems ist im Allgemeinen die Ermittlung des ganzen ε-Subdifferentials notwendig, was in der Praxis in der Regel nicht realisiert werden kann. Die numerische Approximation führt schließlich auf die Bundle-Methoden.

6.4 Bundle-Methode

Wir gehen weiterhin davon aus, dass die Zielfunktion f konvex, aber nicht notwendigerweise differenzierbar ist. Die grundlegende Idee der Bundle-Methoden besteht darin, das $\varepsilon-$Subdifferential in einem Punkt unter Zuhilfenahme eines "Bündels" von Subgradienten in benachbarten Punkten anzunähern. Genauer gesagt: In der k-ten Iteration sollen Konvexkombinationen auf Basis des aktuellen Subgradienten $\xi_k \in \partial_c f(x_k)$ sowie der Subgradienten $\xi_j \in \partial f_c(x_j)$ aus den vergangenen Iterationen $j = 0, \ldots, k-1$ so gewählt werden, dass das dadurch entstehende Polytop im $\varepsilon-$Subdifferential enthalten ist. Für die Definition eines geeigneten Polytops betrachten wir neben den Konvexkombinationen der Form

$$\sum_{j=0}^{k} \lambda_j \xi_j \text{ mit } \sum_{j=0}^{k} \lambda_j = 1,\ \lambda_j \geq 0,\ j = 0, 1, \ldots, k$$

die Linearisierung von f in $y \in \mathbb{R}^n$ für $\xi \in \partial_c f(y)$, die durch

$$\bar{f}(x; y, \xi) := f(y) + \xi^\top (x - y)$$

gegeben ist. Die Linearisierungsfehler in x_k für die Linearisierungen in $y = x_j$ und $\xi = \xi_j$ für $j = 0, \ldots, k$ werden mit

$$\alpha_j^k := \alpha(x_k, x_j, \xi_j) = f(x_k) - f(x_j) - \xi_j^\top (x_k - x_j) \tag{6.6}$$

bezeichnet. Aufgrund der Definition eines Subgradienten im konvexen Fall ergibt sich

$$\alpha_j^k \geq 0,\ j = 0, 1, \ldots, k-1 \text{ und } \alpha_k^k = 0.$$

Definition 6.1. *Es seien Punkte $x_j \in \mathbb{R}^n$ sowie dazugehörige Subgradienten $\xi_j \in \partial_c f(x_j)$ gegeben. Des Weiteren sei $\varepsilon > 0$. Dann definieren wir*

$$P_\varepsilon^k := \{\xi \,|\, \xi = \sum_{j=0}^{k} \lambda_j \xi_j,\ \sum_{j=0}^{k} \lambda_j \alpha_j^k \leq \varepsilon,\ \sum_{j=0}^{k} \lambda_j = 1,\ \lambda_i \geq 0,\ j = 0, 1, \ldots, k\}.$$

Der folgende Satz zeigt, dass dieses Polytop die gewünschten Eigenschaften besitzt.

Satz 6.1. *Für P_ε^k aus Definition 6.1 gilt*

$$P_\varepsilon^k \subseteq \partial_\varepsilon f(x_k).$$

Beweis. Es seien Zahlen $\lambda_j \geq 0, j = 0, \ldots, k$ mit

$$\sum_{j=0}^{k} \lambda_j \alpha_j^k \leq \varepsilon \text{ und } \sum_{j=0}^{k} \lambda_j = 1$$

gegeben. Dann ist die Aussage

$$\sum_{j=0}^{k} \lambda_j \xi_j \in \partial_\varepsilon f(x_k)$$

zu zeigen. Gemäß der Definition des Subgradienten sowie des Linearisierungsfehlers gilt

$$\begin{aligned} \xi_j^\top (x - x_k) &= \xi_j^\top (x - x_j) - \xi_j^\top (x_k - x_j) \\ &\leq f(x) - f(x_j) - \xi_j^\top (x_k - x_j) \\ &= f(x) - f(x_k) + \alpha_j^k \end{aligned}$$

für alle $x \in \mathbb{R}^n$ und alle $j = 0, \ldots, k$. Multiplizieren wir die erhaltenen Ungleichungen mit λ_j und summieren diese für $j = 0, \ldots, k$ auf, so erhalten wir

$$\sum_{j=0}^{k} \lambda_j \xi_j^\top (x - x_k) \leq \underbrace{\sum_{j=0}^{k} \lambda_j}_{=1} f(x) - \underbrace{\sum_{j=0}^{k} \lambda_j}_{=1} f(x_k) + \underbrace{\sum_{j=0}^{k} \lambda_j \alpha_j^k}_{\leq \varepsilon} \text{ für alle } x \in \mathbb{R}^n$$

und somit

$$f(x_k) + \sum_{j=0}^{k} \lambda_j \xi_j^\top (x - x_k) - \varepsilon \leq f(x) \text{ für alle } x \in \mathbb{R}^n.$$

Aufgrund der Definition des ε-Subdifferentials können wir nun auf $\sum_{j=0}^{k} \lambda_j \xi_j \in \partial_\varepsilon f(x_k)$ schließen. □

Für die Bestimmung einer Suchrichtung ist die folgende Aussage von Bedeutung.

Satz 6.2. *Die Menge P_ε^k aus Definition 6.1 ist nichtleer, konvex und kompakt.*

Beweis. Setzen wir $\lambda_i = 0, i \neq k$ und $\lambda_k = 1$, so gilt $\sum_{j=0}^k \lambda_j \alpha_j^k = \lambda_k \alpha_k^k = 0 \leq \varepsilon$. Daher ist P_ε^k aufgrund von $0 \in P_\varepsilon^k$ nichtleer. Wegen der Kompaktheit von $\partial_\varepsilon f(x_k)$ nach Satz 5.10 (i) folgt mit Satz 6.1 die Kompaktheit von P_ε^k. Die Konvexität ist offensichtlich. □

Aufgrund von Satz (6.2) sind die Bedingungen, die für die Herleitung einer Abstiegsrichtung im $\varepsilon-$Subgradientenverfahren notwendig waren, erfüllt. Wir bestimmen deswegen analog zu (6.5) $z_k \in P_\varepsilon^k$ mit

$$\|z_k\| = \min_{z \in P_\varepsilon^k} \|z\|. \tag{6.7}$$

Wegen der besonderen Struktur des Polytops aus Definition 6.1 ist Problem (6.7) äquivalent zu dem konvexen quadratischen Optimierungsproblem

$$\begin{aligned} \min \quad & \frac{1}{2}\Big\| \sum_{j \in J_k} \lambda_j \xi_j \Big\|^2 \\ \text{u.d.N.} \quad & \sum_{j \in J_k} \lambda_j \alpha_j^k \leq \varepsilon, \\ & \sum_{j \in J_k} \lambda_j = 1, \\ & \lambda_j \geq 0 \text{ für alle } j \in J_k, \end{aligned} \tag{6.8}$$

wobei $J_k := \{0, 1, \ldots, k\}$ die Indexmenge ist. Stellt $\lambda_j, j \in J_k$ die Lösung des obigen Problems dar, so erhalten wir die Lösung von Problemstellung (6.7) durch $z_k = \sum_{j=0}^k \lambda_j \xi_j$. Die Suchrichtung $d_k = -z_k$ ist im Allgemeinen keine Abstiegsrichtung. Aus diesem Grund werden später Nullschritte eingeführt. Die Hauptschwierigkeit des obigen Problems besteht in der Wahl der Toleranz ε. Auf der einen Seite erzeugt ein großes ε eine schlechte Approximation von $\partial f_\varepsilon(x_k)$, auf der anderen Seite kann ein kleines ε aufgrund der Ungleichung $f(x_{k+1}) \leq f(x_k) - \varepsilon$ keinen großen Abstieg in der Zielfunktion garantieren. Aufgrund der großen Sensibilität von Problem (6.8) in Bezug auf die Wahl von ε wurden andere äquivalente quadratische Optimierungsprobleme für die Bestimmung der Suchrichtung hergeleitet, die aus numerischer Sicht besser geeignet sind. Wir werden nun eine Variante, die die Bundle-Idee mit dem Schnittebenenkonzept und dem Trust-Region-Ansatz verknüpft, näher betrachten. Eine konvexe Funktion kann entsprechend dem Schnittebenenverfahren durch die stückweise lineare Funktion

$$\bar{f}_k := \max_{j \in J_k} \{f(x_j) + \xi_j^\top (x - x_j)\} = f(x_k) + \max_{j \in J_k} \{-\alpha_j^k + \xi_j^\top (x - x_k)\}$$

angenähert werden, wobei $\xi_j \in \partial_c f(x_j), j \in J_k$ gegebene Subgradienten sind. Nach der Definition des Subdifferentials gilt

$$f(x) \geq f(x_j) + \xi_j^\top (x - x_j) \text{ für alle } j \in J_k,\ x \in \mathbb{R}^n,$$

woraus

$$f(x) \geq \max_{j \in J_k} \{f(x_j) + \xi_j^\top (x - x_j)\} = \bar{f}_k(x) \text{ für alle } x \in \mathbb{R}^n$$

folgt. Die Funktion f wird somit durch $\bar{f}_k$ von unten approximiert. Gemäß dem Trust-Region-Konzept nehmen wir an, dass die Approximation auf einem bestimmten "Vertrauensbereich" eine akzeptable Genauigkeit besitzt. Dieser Bereich wird durch $\frac{1}{2}\|d\|^2 \leq \rho$ mit einem Parameter $\rho > 0$ und $d := x - x_k$ beschrieben. Die optimale Suchrichtung erhalten wir über die Lösung des nichtdifferenzierbaren Optimierungsproblems

$$\min_{\frac{1}{2}\|d\| \leq \rho} \bar{f}_k(x_k + d),$$

welches äquivalent zum differenzierbaren Optimierungsproblem

$$\begin{array}{ll} \min & v \\ \text{bzgl.} & v, d \\ \text{u.d.N.} & -\alpha_j^k + \xi_j^\top d \leq v \text{ für alle } j \in J_k, \\ & \frac{1}{2}\|d\|^2 \leq \rho \end{array} \tag{6.9}$$

ist. Der nächste Hilfssatz zeigt auf, dass zwischen den Problemen (6.8) und (6.9) eine Dualitätsbeziehung vorhanden ist.

Hilfssatz 6.3. *Es gilt:*

(i) Sei $\hat{\lambda}_j, j \in J_k$ eine Optimallösung von (6.8) und $\hat{\mu}$ ein zugehöriger Lagrangemultiplikator der ersten Nebenbedingung in (6.8), für den $\hat{\mu} > 0$ gelte. Dann ist $(\hat{v}, \hat{d})$ mit

$$\hat{d} = -\frac{1}{\hat{\mu}} \sum_{j \in J_k} \hat{\lambda}_j \xi_j, \quad \hat{v} = \bar{f}_k(x_k + \hat{d})$$

Optimallösung von (6.9) für

$$\rho = \frac{1}{2\hat{\mu}^2} \left\| \sum_{j \in J_k} \hat{\lambda}_j \xi_j \right\|^2.$$

(ii) Sei $(\hat{v}, \hat{d})$ eine Optimallösung von (6.9) mit Lagrangemultiplikatoren $\hat{\lambda}_j \geq 0, j \in J_k$ der ersten $k+1$ Restriktionen in Problem (6.9). Für den Lagrangemultiplikator $\hat{\mu}$ der letzten Nebenbedingung in (6.9) gelte $\hat{\mu} > 0$. Dann ist $\hat{\lambda}_j > 0, j \in J_k$ Optimallösung von (6.8) für

$$\varepsilon = \textstyle\sum_{j \in J^k} \hat{\lambda}_j \alpha_j^k.$$

Beweis. Schramm (1989), Lemma 4.4 □

Aufgrund der quadratischen Nebenbedingung ist die Lösung von (6.9) aufwendig. Zur Erleichterung der numerischen Behandlung nehmen wir diese Restriktion mit Hilfe des Parameters $\mu \geq 0$ in die Zielfunktion auf und erhalten

$$\begin{array}{ll} \min & v + \frac{1}{2}\mu\|d\|^2 \\ \text{bzgl.} & v, d \\ \text{u.d.N.} & -\alpha_j^k + \xi_j^\top d \leq v \text{ für alle } j \in J_k. \end{array} \tag{6.10}$$

Auch zwischen diesem Optimierungsproblem und (6.9) bestehen Zusammenhänge:

Hilfssatz 6.4. *Es gilt:*

(i) Ist $(\hat{v}, \hat{d})$ eine Optimallösung von (6.9) und $\hat{\mu}$ ein zugehöriger Lagrangemultiplikator der letzten Restriktion in (6.9), so ist $(\hat{v}, \hat{d})$ eine Optimallösung von (6.10) für $\mu = \hat{\mu}$.

(ii) Ist $(\hat{v}, \hat{d})$ eine Optimallösung von (6.10) mit $\mu \geq 0$, so ist $(\hat{v}, \hat{d})$ eine Optimallösung von (6.9) für

$$\rho = \frac{1}{2}\|\hat{d}\|^2.$$

Beweis. Schramm(1989), Lemma 4.2 □

Die Hilfssätze 6.3 und 6.4 verdeutlichen, dass alle drei Optimierungsprobleme zur Bestimmung der Suchrichtung herangezogen werden können. Im Folgenden werden wir uns auf Problemstellung (6.10) konzentrieren. Mit $u = v + f(x_k)$ ist dieses Problem äquivalent zu

$$\begin{array}{ll} \min & u + \frac{1}{2}\mu\|d\|^2 \\ \text{u.d.N.} & f_j^k + \xi_j^\top d \leq u \text{ für alle } j \in J_k, \end{array} \tag{6.11}$$

wobei wir $f_j^k := f(x_j) + \xi^\top (x_k - x_j)$ setzen. Gemäß Unterabschnitt 5.6 lautet das zu (6.11) duale Problem:

$$\begin{aligned} \min \quad & \frac{1}{2\mu}\|\lambda_j \xi_j\|^2 - \sum_{j \in J_k} f_j^k \\ \text{u.d.N.} \quad & \lambda_j \geq 0 \text{ für alle } j \in J_k, \\ & \sum_{j \in J_k} \lambda_j = 1 \end{aligned} \tag{6.12}$$

Da wir in Unterabschnitt 7.1.3 ein duales Problem in analoger Weise bestimmen, verzichten wir hier auf eine ausführliche Herleitung. Problemstellung (6.11) führt in der Praxis nach einer großen Anzahl von Iterationen zu Speicher- und Berechnungsproblemen, da die Anzahl der Nebenbedingungen und die Größe der Dimension von Iteration zu Iteration zunimmt. Eine Möglichkeit, die Anzahl der Nebenbedingungen zu begrenzen, stellt die Subgradienten-aggregationsstrategie dar. Die Idee besteht darin, die Informationen der Restriktionen aus den vergangenen Iterationen in einer aggregierten Nebenbedingung zu bündeln. Genauer ausgedrückt: Nach der Lösung des aktuellen Richtungsbestimmungsproblems (6.11) wird in einem ersten Schritt ein reduziertes Hilfsproblem konstruiert, das die gleiche Lösung besitzt. Das neue Problem zur Bestimmung der Suchrichtung erhalten wir in einem zweiten Schritt, indem wir zu diesem reduzierten Hilfsproblem die Nebenbedingung, die durch den aktuellen Iterationspunkt erzeugt wurde, hinzufügen, um den Fortschritt der Methode zu garantieren. Für die Ermittlung des Ersatzproblems werden Konvexkombinationen der vergangenen Subgradienten auf der Basis der dazugehörigen Lagrangemultiplikatoren, die über die Lösung des dualen Problems (6.12) ermittelt werden, gebildet:

$$(p^k, \tilde{f}_p^k) \quad := \quad \textstyle\sum_{j \in J_k} \lambda_j^k (\xi_j, f_j^k) \tag{6.13}$$

Für den Nachweis der Äquivalenz des Ausgangsproblems und des konstruierten Problems benötigen wir den nachfolgenden Satz:

Satz 6.5. *Es gilt:*

(i) Es existiert immer eine eindeutige Lösung (d_k, u_k) des Optimierungsproblems (6.11).

(ii) Der Vektor $(d_k, u_k) \in \mathbb{R}^n \times \mathbb{R}$ löst Problem (6.11) genau dann, wenn ein Vektor p^k sowie reelle Lagrangemultiplikatoren λ_j^k, $j \in J_k$ mit den folgenden Eigenschaften existieren:

(a) $\sum\limits_{j \in J_k} \lambda_j^k = 1, \quad \lambda_j^k \geq 0,\ j \in J_k$

(b) $\lambda_j^k [f_j^k + \xi_j^\top d_k - u_k] = 0,\ j \in J_k$

(c) $p^k = \sum\limits_{j \in J_k} \lambda_j^k \xi_j$

(d) $d_k = -\frac{1}{\mu} p^k$

(e) $u_k = -\frac{1}{\mu}\|p^k\|^2 + \sum_{j\in J_k} \lambda_j^k f_j^k$

(f) $f_j^k + \xi_j^\top d_k \leq u_k,\ j \in J_k$

(iii) Die Multiplikatoren $\lambda_j^k, j \in J_k$ erfüllen die Bedingungen (a)-(f) genau dann, wenn sie Lösung des Problems (6.12) sind.

Beweis. (i) Wir definieren die Funktion ϕ durch

$$\phi(d) := \frac{1}{2}\mu\|d\|^2 + \varphi(d),\ d \in \mathbb{R}^n$$

mit $\varphi(d) := \max\{f_j^k + \xi_j^\top d \,|\, j \in J_k\}$. Die Funktion ϕ ist strikt konvex. Des Weiteren gilt $\phi(d) \to +\infty$ für $\|d\| \to \infty$, da $\varphi(d) \geq f_j^k + \xi_j^\top d \geq f_j^k - \|\xi_j\|\|d\|$ für $j \in J_k$ und $d \in \mathbb{R}^n$ nach der Cauchy-Schwarzschen Ungleichung. Deswegen besitzt ϕ das eindeutige Minimum d_k. Ebenfalls auf eindeutige Weise ergibt sich $u_k = \varphi(d_k)$. Somit erhalten wir die Behauptung.

(ii) Problem (6.11) ist konvex und erfüllt für den Punkt $(\tilde{d}, \tilde{u}) := (0, \max_{j\in J_k} f_j^k + 1)$ die Slater-Bedingung (5.6). Deswegen können wir Satz 5.19 anwenden: Die Tatsache, dass (d_k, u_k) die Problemstellung (6.11) löst, ist äquivalent zu der Bedingung, dass es Multiplikatoren $\lambda_j^k \geq 0$ gibt, so dass $\lambda_j^k[f_j^k + \xi_j d_k - u_k] = 0$ für $j \in J_k$ und

$$0 \in \partial(u_k + \frac{1}{2}\mu\|d_k\|^2) + \sum_{j\in J_k} \lambda_j^k \partial(f_j^k + \xi_j d_k - u_k) \tag{6.14}$$

gilt. Da die auftretenden Funktionen in (6.14) stetig differenzierbar sind, fallen die Subdifferentiale mit den Gradienten zusammen und (6.14) ist äquivalent zu

$$0 = (\mu d_k, 1) + \sum_{j\in J_k} \lambda_j^k (\xi_j, -1).$$

Die Bedingungen (a)-(d) sind somit bewiesen. Für den Nachweis der Aussage (e) werden die Gleichungen in (b) addiert und (a), (c) und (d) angewendet:

$$\underbrace{\sum_{j\in J_k} \lambda_j^k}_{=1} u_k = \sum_{j\in J_k} \lambda_j^k \xi_j^\top d_k + \sum_{j\in J_k} \lambda_j^k f_j^k$$

$$= -p^k(\frac{1}{\mu}p^k) + \sum_{j\in J_k} \lambda_j^k f_j^k = -\frac{1}{\mu}\|p^k\|^2 + \sum_{j\in J_k} \lambda_j^k f_j^k$$

Die Zulässigkeit der optimalen Lösung führt unmittelbar auf Ungleichung (f).

(iii) Kiwiel (1985), Lemma 2.2.1 (iii)

□

Wir betrachten nun das reduzierte Problem

$$\begin{aligned} \min \quad & u + \frac{1}{2}\mu\|d\|^2 \\ \text{u.d.N.} \quad & f_j^k + \xi_j^\top d \le u,\ j \in \hat{J}_k, \\ & \tilde{f}_p^k + (p^k)^\top d \le u, \end{aligned} \tag{6.15}$$

wobei $\hat{J}_k$ eine beliebige Teilmenge von J_k ist.

Satz 6.6. *Problem (6.11) ist äquivalent zum reduzierten Problem (6.15).*

Beweis. Wir werden zeigen, dass eine Lösung des ursprünglichen Problems (6.11) auch eine Lösung des reduzierten Problems (6.15) darstellt. Darüber hinaus wissen wir aufgrund von 6.5(i), dass für (6.11) eine eindeutige Lösung existiert. Dieselbe Aussage gilt auch für (6.15), da beide Probleme die gleiche Struktur besitzen. Insgesamt erhalten wir somit die Behauptung.
Wir nehmen an, dass (d_k, u_k) eine Lösung von (6.11) darstellt. Nach Satz 6.5(ii) existieren für diesen Punkt somit Lagrangemultiplikatoren $\lambda_k^j, j \in \mathbb{R}^n$ mit den Eigenschaften $(a)-(f)$. Mit Hilfe dieser Eigenschaften, bezogen auf Problem (6.11), leiten wir her, dass auch Lagrangemultiplikatoren $\tilde{\lambda}_p^k, \tilde{\lambda}_j^k, j \in \hat{J}_k$ existieren, die die Bedingungen $(a)-(f)$ für das reduzierte Problem (6.15) erfüllen. Wir setzen $\tilde{\lambda}_p^k := 1, \tilde{\lambda}_j^k := 0, j \in \hat{J}_k$. Dann folgt:

(a)

$$\sum_{j \in \hat{J}_k} \tilde{\lambda}_j^k + \tilde{\lambda}_p^k = 1$$

(b)

$$\tilde{\lambda}_j^k [f_j^k + \xi_j^\top d_k - u_k] = 0 \text{ für alle } j \in \hat{J}_k, \text{ da } \tilde{\lambda}_j^k = 0 \text{ für alle } j \in \hat{J}_k$$

$$\begin{aligned} & \underbrace{\tilde{\lambda}_p^k}_{=1} [\tilde{f}_p^k + (p^k)^\top d_k - u_k] = \\ &= \sum_{j \in J_k} \lambda_j^k f_j^k + \left(\sum_{j \in J_k} \lambda_j^k \xi_j \right)^\top d_k - \underbrace{\sum_{j \in J_k} \lambda_j^k}_{=1} u_k \\ &= \sum_{j \in J_k} \lambda_j^k [f_j^k + \xi_j^\top d_k - u_k] = 0 \end{aligned}$$

(c)

$$\underbrace{\tilde{\lambda}_p^k}_{=1} p^k + \underbrace{\sum_{j\in\hat{J}_k} \tilde{\lambda}_j^k \xi_j}_{=0} = p^k$$

(d) folgt direkt aus (c)

(e)

$$-\frac{1}{\mu}\|p^k\|^2 + \underbrace{\sum_{j\in\hat{J}_k} \tilde{\lambda}_j^k f_j^k}_{=0} + \tilde{\lambda}_p^k \tilde{f}_p^k = -\frac{1}{\mu}\|p^k\|^2 + \sum_{j\in J_k} \lambda_j^k f_j^k = u_k$$

(f)

$$\begin{aligned}
&\tilde{f}_p^k + (p^k)^\top d_k - u_k = \\
&= \sum_{j\in J_k} \lambda_j^k f_j^k + \left(\sum_{j\in J_k} \lambda_j^k \xi_j\right)^\top d_k - \sum_{j\in J_k} \lambda_j^k u_k = \sum_{j\in J_k} \lambda_j^k (\underbrace{f_j^k + \xi_j^\top d_k - u_k}_{\leq 0}) \leq 0
\end{aligned}$$

□

Um das k+1-te quadratische Hilfsproblem zur Bestimmung der Suchrichtung zu bilden, müssen wir das reduzierte Problem noch updaten und um die Nebenbedingung, die durch den neuen Iterationspunkt x_{k+1} entsteht, ergänzen. Aufgrund von

$$\begin{aligned}
f_j(x) &= f(y_j) + \xi_j^\top (x - y_j) \\
&= f(y_j) + \xi_j^\top (x_k - y_j) + \xi_j^\top (x - x_k) \\
&= f_j^k + \xi_j^\top (x - x_k)
\end{aligned} \tag{6.16}$$

für alle $j \in J_k$ müssen die Punkte y_j, $j \in J_k$ nicht gespeichert werden und man erhält die Updateformel

$$f_j^{k+1} = f_j(x_{k+1}) = f_j^k + \xi_j^\top (x_{k+1} - x_k),\ j \in J_k.$$

Die aggregierten Größen können analog upgedatet werden. Mit

$$\begin{aligned}
f_p^k(x) &= \sum_{j\in J_k} \lambda_j^k f_j(x) \\
&= \sum_{j\in J_k} \lambda_j^k (f_j^k + \xi_j^\top (x - x_k)) \\
&= \tilde{f}_p^k + (p^k)^\top (x - x_k)
\end{aligned} \tag{6.17}$$

ergibt sich

$$f_p^{k+1} = f_p(x_{k+1}) = \tilde{f}_p^k + (p^k)^\top (x_{k+1} - x_k). \tag{6.18}$$

Aus (6.18) folgt

$$\tilde{f}_p^k = f_p^{k+1} - (p^k)^\top (x_{k+1} - x_k). \tag{6.19}$$

Setzen wir Gleichung (6.19) in die aggregierte Nebenbedingung aus Problem (6.15) ein, so erhalten wir

$$\begin{aligned} & \tilde{f}_p^k + (p^k)^\top d \leq u \\ \Rightarrow \quad & f_p^{k+1} - (p^k)^\top (x_{k+1} - x_k) + (p^k)^\top (x - x_k) \quad \leq u \\ \Rightarrow \quad & f_p^{k+1} + (p^k)^\top (x - x_{k+1}) \leq u. \end{aligned}$$

Um die neue Nebenbedingung zu erzeugen, berechnen wir $\xi_{k+1} \in \partial f(y_{k+1})$. Insgesamt führen die obigen Ausführungen zu dem folgenden k+1-ten Teilproblem:

$$\begin{aligned} \min \quad & \frac{1}{2}\mu \|d\|^2 + u \\ \text{u.d.N.} \quad & f_j^{k+1} + (\xi_j)^\top d \leq u,\ j \in J_{k+1} = \hat{J}_k \cup \{k+1\}, \\ & f_p^{k+1} + (p^k)^\top d \leq u \end{aligned} \tag{6.20}$$

Es ist allerdings zu beachten, dass wir die Methode der Subgradientenaggregation außer in der ersten Iteration auf Teilprobleme anwenden, bei denen schon in vorhergehenden Iterationen Aggregation stattgefunden hat. Wir müssen (6.13) deswegen leicht modifizieren:

$$(p^k, \tilde{f}_p^k) = \sum_{j \in J^k} \lambda_j^k (g^j, f_j^k) + \lambda_p^k (p^{k-1}, f_p^k)$$

Die Subgradientenaggregationsstrategie ermöglicht uns eine beliebige Wahl der Indexmenge im Problem (6.20). Besitzt diese Menge eine große Mächtigkeit, so ist jede Iteration sehr effizient, allerdings wird viel Speicherplatz benötigt und ein beträchtlicher Rechenaufwand ist erforderlich. Umgekehrt verhält es sich, wenn die Menge $\hat{J}_k$ aus wenigen Elementen besteht. Dem Nutzer fällt die Aufgabe zu, einen Ausgleich zwischen beiden Aspekten zu finden.

Kapitel 7

Herleitung der Bundle-Newton-Methode

Nun kommen wir zu der eigentlichen Herleitung der Bundle-Newton-Methode. Das Besondere besteht darin, dass die auftretenden Funktionen im Gegensatz zu den in Kapitel 6.4 betrachteten Bundle-Methoden nicht stückweise linear, sondern stückweise quadratisch approximiert werden. Wir werden die Bundle-Newton-Methode aus Luksan und Vlček (1998) für den restringierten Fall in Anlehnung an Kiwiel (1985) und Mäkelä und Neittaanmäki (1992) herleiten. Wir betrachten also das Problem

$$\begin{array}{ll} \min & f(x) \\ \text{u.d.N.} & F_i \leq 0,\ i = 0, \ldots, m, \end{array} \tag{7.1}$$

wobei die Funktionen $f\colon \mathbb{R} \to \mathbb{R}^n$ und $F_i\colon \mathbb{R} \to \mathbb{R}^n, i = 0, \ldots, m$ lokal lipschitzstetig sind. Die Funktion F sei durch

$$F(x) = \max\{F_i(x) \,|\, i = 1, \ldots, m\} \text{ für } x \in \mathbb{R}^n$$

gegeben.
Wie bereits bei den bisherigen auf Subgradienten basierenden Verfahren nehmen wir an, dass wir in jedem Punkt $y \in \mathbb{R}^n$ einen beliebigen Subgradienten $g_f(y) \in \partial f(y)$ bzw. $g_F(y) \in \partial F(y)$ ermitteln können. Da der Algorithmus auf quadratischer Approximation beruht, benötigen wir neben den Subgradienten auch die Hessematrix. Allerdings reicht es aus, wenn wir in jedem Punkt eine symmetrische $n \times n-$Matrix $G_f(y)$ [bzw. $G_F(y)$] als Näherung von $\nabla\nabla f(y)$ [bzw. $\nabla\nabla F(y)$] angeben können. Für die Problemfunktionen wird keine Differenzierbarkeit, sondern nur lokale Lipschitzstetigkeit gefordert. Nach dem Satz von Rademacher (Satz 5.2) sind die Funktionen aber fast überall, also überall bis auf eine Lebesgue-Nullmenge, differenzierbar. An Nichtdifferenzierbarkeitsstellen kann statt eines Elementes des Subdifferentials der Gradient in einem Punkt y aus einer infinitesimal kleinen ε-Umgebung $\mathcal{U}_\varepsilon(y)$ herangezogen werden. Analog erhalten wir im Falle einer stückweise zweimal differenzierbaren Funktion

eine Approximation der Hessematrix an einer Nichtdifferenzierbarkeitsstelle y, indem wir auf die Hessematrix in einem Punkt $x \in U_\varepsilon(y)$ zurückgreifen.
Die Bundle-Newton-Methode stützt die Bestimmmung der Suchrichtung entsprechend Variante (6.9) aus dem vorhergehenden Kapitel auf das Schnittebenenkonzept. Wir gehen davon aus, dass in der k-ten Iteration des Algorithmus neben dem aktuellen Iterationspunkt x_k auch die Hilfspunkte $y_j \in \mathbb{R}^n$, $j \in J_k$ sowie beliebige Subgradienten $\xi_j^f \in \partial f(y_j)$ und $\xi_j^F \in \partial F(y_j), j \in J_k$ vorhanden sind. Wir setzen vorerst $J_k := \{1, \ldots, k\}$. Auf eine sinnvollere Wahl der Indexmenge J_k gehen wir in Unterabschnitt 7.1.4 näher ein. Des Weiteren sei in jedem Punkt y_j eine Näherung G_j^f der Hessematrix $\nabla\nabla f(y_j)$ bzw. eine Näherung G_j^F der Hessematrix $\nabla\nabla F(y_j)$ gegeben. Außerdem benötigen wir die Damping-Parameter $\varrho_j^f \in [0,1]$ und $\varrho_j^F \in [0,1]$, $j \in J_k$, deren Zweck ebenfalls in Unterabschnitt 7.1.4 verdeutlicht wird. Wir definieren nun mit Hilfe von

$$f_j(x) := f(y_j) + (\xi_j^f)^\top (x - y_j) + \frac{1}{2}\varrho_j^f (x - y_j)^\top G_j^f (x - y_j),\ j \in J_k,$$
$$F_j(x) := F(y_j) + (\xi_j^F)^\top (x - y_j) + \frac{1}{2}\varrho_j^F (x - y_j)^\top G_j^F (x - y_j),\ j \in J_k$$

die stückweise quadratischen Approximationen von f, F und H

$$\hat{f}_k(x) := \max\{f_j(x) \,|\, j \in J_k\},$$
$$\hat{F}_k(x) := \max\{F_j(x) \,|\, j \in J_k\},$$
$$\hat{H}_k(x, x_k) := \max\{\hat{f}_k(x) - f(x_k), \hat{F}_k(x)\}$$

für alle $x \in \mathbb{R}^n$, wobei $H(x, x_k)$ die in Kapitel 5.5 eingeführte Improvement-Funktion darstellt. Im Folgenden werden die Bezeichnungen

$$f_j^k := f_j(x_k) := f(y_j) + (\xi_j^f)^\top (x_k - y_j) + \frac{1}{2}\varrho_j^f (x_k - y_j)^\top G_j^f (x_k - y_j),\ j \in J_k,$$
$$F_j^k := F_j(x_k) := F(y_j) + (\xi_j^F)^\top (x_k - y_j) + \frac{1}{2}\varrho_j^F (x_k - y_j)^\top G_j^F (x_k - y_j),\ j \in J_k$$

verwendet.

7.1 Bestimmung der Suchrichtung

Ausgehend vom aktuellen Iterationspunkt x_k möchten wir eine zulässige Abstiegsrichtung ermitteln.

Definition 7.1. *Die Richtung $d \in \mathbb{R}^n$ heißt eine zulässige Abstiegsrichtung von Problem (7.1) in x_k, falls es ein $\tilde{t} > 0$ gibt, so dass*

$$f(x_k + \tilde{t}d) < f(x) \text{ und } F(x_k + \tilde{t}d) \leq 0 \text{ für alle } t \in (0, \tilde{t}].$$

Es kann gezeigt werden, dass eine Abstiegsrichtung der Improvement-Funktion eine zulässige Abstiegsrichtung für die Problemstellung (7.1) darstellt. Eine Abstiegsrichtung der Improvementfunktion erhalten wir über die Lösung des Problems

$$\min H(x_k + d, x_k) \text{ bzgl. } d \in \mathbb{R}^n. \tag{7.2}$$

Wir werden (7.2) allerdings nicht direkt verwenden, sondern $\hat{H}$ heranziehen, so dass sich

$$\min \hat{H}(x, x_k) \text{ bzgl. } x \in \mathbb{R}^n \tag{7.3}$$

mit $x := x_k + d$ ergibt.

Bemerkung 7.1. *Da $\hat{H}$ allerdings nur eine Approximation von H darstellt, ist es möglich, dass durch die direkte Wahl $x_{k+1} := x_k + d_k$ gar nicht der größte bzw. kein ausreichend großer Abstieg erzielt wird. Zur Klärung dieser Problematik verweisen wir auf die Ausführungen zur Schrittweitensteuerung in Unterabschnitt 7.2.*

Problem (7.3) ist äquivalent zu

$$\begin{aligned} \min \quad & \hat{v} \\ \text{u.d.N.} \quad & \hat{f}_k(x) - f(x_k) \leq \hat{v}, \\ & \hat{F}_k(x) \leq \hat{v}, \end{aligned} \tag{7.4}$$

welches auch in der Form

$$\begin{aligned} \min \quad & \hat{v} \\ \text{u.d.N.} \quad & f(y_j) + (\xi_j^f)^\top (x - y_j) + \frac{1}{2}\varrho_j^f (x - y_j)^\top G_j^f (x - y_j) - f(x_k) \leq \hat{v},\ j \in J_k, \\ & F(y_j) + (\xi_j^F)^\top (x - y_j) + \frac{1}{2}\varrho_j^F (x - y_j)^\top G_j^F (x - y_j) \leq \hat{v},\ j \in J_k, \end{aligned} \tag{7.5}$$

geschrieben werden kann.

7.1.1 Anwendung des SQP-Verfahrens

Problem (7.5) kann mit Hilfe der Lagrange-Newton-Iteration, die in Kapitel 5.7.2 behandelt wurde, approximativ gelöst werden. Für die Herleitung des Hilfsproblems, das in diesem Verfahren in jedem Iterationsschritt auftritt, werden die folgenden Größen benötigt:

- $z = (x, \hat{v})$
- $f(z) = f(x, \hat{v}) = \hat{v}$
- $\nabla_z f(z) = (0, 1)$

- $\nabla_{zz} f(z) = 0^{(n+1)\times(n+1)}$
- $g_j^f(z) = f(y_j) + (\xi_j^f)^\top (x - y_j) + \frac{1}{2}\varrho_j^f (x - y_j)^\top G_j^f (x - y_j) - f(x_k) - \hat{v},\ j \in J_k$
- $g_j^F(z) = F(y_j) + (\xi_j^F)^\top (x - y_j) + \frac{1}{2}\varrho_j^F (x - y_j)^\top G_j^F (x - y_j) - \hat{v},\ j \in J_k$
- $\nabla_z g_j^f(z) = ((\xi_j^f + \varrho_j^f G_j^f (x - y_j))^\top, -1),\ j \in J_k$
- $\nabla_z g_j^F(z) = ((\xi_j^F + \varrho_j^F G_j^F (x - y_j))^\top, -1),\ j \in J_k$
- $\tilde{G}_j^f := \begin{pmatrix} & & 0 \\ & G_j^f & \vdots \\ 0 & \dots & 0 \end{pmatrix}$
- $\tilde{G}_j^F := \begin{pmatrix} & & 0 \\ & G_j^F & \vdots \\ 0 & \dots & 0 \end{pmatrix}$
- $\nabla_{zz} g_j^f(z) = \varrho_j^f \tilde{G}_j^f,\ j \in J_k$
- $\nabla_{zz} g_j^F(z) = \varrho_j^F \tilde{G}_j^F,\ j \in J_k$
- $\begin{aligned} \nabla_{zz} L(z, \mu) &= \nabla_{zz} f(z) + \textstyle\sum_{j \in J_{k-1}} \lambda_{f,j}^{k-1} \nabla_{zz} g_j^f(z) + \sum_{j \in J_{k-1}} \lambda_{F,j}^{k-1} \nabla_{zz} g_j^F(z) \\ &= 0 + \textstyle\sum_{j \in J_{k-1}} \lambda_{f,j}^{k-1} \varrho_j^f \tilde{G}_j^f + \sum_{j \in J_{k-1}} \lambda_{F,j}^{k-1} \varrho_j^F \tilde{G}_j^F \end{aligned}$

In der k-ten Iteration erhalten wir somit gemäß (5.15) das quadratische Teilproblem

$$
\begin{aligned}
\min \quad & \hat{v} + (0,1)^\top \Delta z + \frac{1}{2} \Delta z^\top \Bigg(\sum_{j \in J_{k-1}} \lambda_{f,j}^{k-1} \varrho_j^f \tilde{G}_j^f + \sum_{j \in J_{k-1}} \lambda_{F,j}^{k-1} \varrho_j^F \tilde{G}_j^F \Bigg) \Delta z \\
\text{u.d.N.} \quad & f(y_j) + (\xi_j^f)^\top (x_k - y_j) + \frac{1}{2} \varrho_j^f (x_k - y_j)^\top G_j^f (x_k - y_j) - f(x_k) - \hat{v} \\
& \quad + ((\xi_j^f + \varrho_j^f G_j^f (x_k - y_j))^\top, -1) \Delta z \leq 0,\ j \in J_k, \\
& F(y_j) + (\xi_j^F)^\top (x_k - y_j) + \frac{1}{2} \varrho_j^F (x_k - y_j)^\top G_j^F (x_k - y_j) - \hat{v} \\
& \quad + ((\xi_j^F + \varrho_j^F G_j^F (x_k - y_j))^\top, -1) \Delta z \leq 0,\ j \in J_k,
\end{aligned}
$$

was vereinfacht dargestellt werden kann als

$$\begin{aligned}
\min \quad & \hat{v} + \Delta\hat{v} + \frac{1}{2}\Delta x^\top \Bigg(\sum_{j\in J_{k-1}} \lambda_{f,j}^{k-1} \varrho_j^f G_j^f + \sum_{j\in J_{k-1}} \lambda_{F,j}^{k-1} \varrho_j^F G_j^F \Bigg) \Delta x \\
\text{u.d.N.} \quad & f(y_j) + (\xi_j^f)^\top (x_k - y_j) + \frac{1}{2}\varrho_j^f (x_k - y_j)^\top G_j^f (x_k - y_j) - f(x_k) - \hat{v} \\
& \quad + (\xi_j^f + \varrho_j^f G_j^f (x_k - y_j))^\top \Delta x - \Delta\hat{v} \leq 0,\ j \in J_k, \\
& F(y_j) + (\xi_j^F)^\top (x_k - y_j) + \frac{1}{2}\varrho_j^F (x_k - y_j)^\top G_j^F (x_k - y_j) - \hat{v} \\
& \quad + (\xi_j^F + \varrho_j^F G_j^F (x_k - y_j))^\top \Delta x - \Delta\hat{v} \leq 0,\ j \in J_k.
\end{aligned}$$

Mit $v := \hat{v} + \Delta\hat{v}$, $\tilde{W}_k := \sum_{j\in J_{k-1}} \lambda_j^{f,k-1} \varrho_j^f G_j^f + \sum_{j\in J_{k-1}} \lambda_j^{F,k-1} \varrho_j^F G_j^F$, $\Delta x := x - x_k$ und $x := x_k + d$ erhalten wir

$$\begin{aligned}
\min \quad & v + \frac{1}{2} d^\top \tilde{W}_k d \\
\text{u.d.N.} \quad & f(y_j) + (\xi_j^f)^\top (x_k - y_j) + \frac{1}{2}\varrho_j^f (x_k - y_j)^\top G_j^f (x_k - y_j) - f(x_k) \\
& \quad + (\xi_j^f + \varrho_j^f G_j^f (x_k - y_j))^\top d \leq v,\ j \in J_k, \\
& F(y_j) + (\xi_j^F)^\top (x_k - y_j) + \frac{1}{2}\varrho_j^F (x_k - y_j)^\top G_j^F (x_k - y_j) \\
& \quad + (\xi_j^F + \varrho_j^F G_j^F (x_k - y_j))^\top d \leq v,\ j \in J_k.
\end{aligned} \tag{7.6}$$

Hierbei ist zu beachten, dass die Matrix $\tilde{W}_k$ im Folgenden durch ihre positiv definite Approximation W_k ersetzt wird (siehe dazu auch Bemerkung 5.23). Dies gewährleistet die Existenz einer Lösung des quadratischen Teilproblems und erweist sich auch für die Ermittlung des dualen Problems als nützlich, wie wir in Abschnitt 7.1.3 sehen werden.

Bemerkung 7.2. *In Abschnitt 6.4 wurde bei Variante (6.10) der Term $\frac{1}{2}\mu\|d\|^2$ addiert, um zu verhindern, dass das Minimum der stückweise linear approximierten Zielfunktion nicht existiert bzw. zu weit vom Minimum der ursprünglichen Funktion f entfernt ist. Im Falle der quadratischen Approximation ist ein solcher Stabilisierungsterm durch $\frac{1}{2}d^\top \tilde{W}_k d$ automatisch gegeben.*

7.1.2 Approximationsfehler

Die in der k-ten Iteration aufgrund der quadratischen Approximation entstandenen Fehler sind durch

$$\begin{aligned}
\beta_{f,j}^k &= f(x_k) - f(y_j) - (\xi_j^f)^\top (x_k - y_j) - \frac{1}{2}\varrho_j^f (x_k - y_j)^\top G_j^f (x_k - y_j),\ j \in J_k, \\
\beta_{F,j}^k &= -F(y_j) - (\xi_j^F)^\top (x_k - y_j) - \frac{1}{2}\varrho_j^F (x_k - y_j)^\top G_j^F (x_k - y_j),\ j \in J_k
\end{aligned} \tag{7.7}$$

gegeben. Diese Größen müssen allerdings noch geeignet modifiziert werden, um die Ideen, die im konvexen, linearisierten Fall angewandt werden, auf den nichtkonvexen Fall mit quadratischer Approximation übertragen zu können. Denn im Gegensatz zum Linearisierungsfehler α_j^k (siehe (6.6)) können die Fehler $\beta_{f,j}^k$ und $\beta_{F,j}^k$ sogar im konvexen Fall negativ werden. Dies bringt Schwierigkeiten mit sich. Die Funktionen werden nicht mehr notwendigerweise von unten approximiert. Des Weiteren geht die Eigenschaft, dass die Nebenbedingungen, die zu Hilfspunkten mit großem Linearisierungsfehler gehören, das Ergebnis weniger stark beeinflussen, verloren (Vergleiche hierzu Problemstellung (6.10)). Um die genannten Probleme zu beheben und somit den Weg für die Anwendung der Konzepte aus Abschnitt (6.4), die Konvexität voraussetzen und auf Linearisierung basieren, zu ebnen, führen wir die "Gewichte"

$$\begin{aligned} \alpha_{f,j}^k &= \max\{|\beta_{f,j}^k|, \gamma_f \|x_k - y_j\|^\omega\},\ j \in J_k, \\ \alpha_{F,j}^k &= \max\{|\beta_{F,j}^k|, \gamma_F \|x_k - y_j\|^\omega\},\ j \in J_k \end{aligned} \tag{7.8}$$

ein, wobei $\gamma_f > 0$, $\gamma_F > 0$ und $\omega \geq 1$ Parameter darstellen. Der Term $\|x_k - y_j\|$ sorgt dafür, dass $\alpha_{f,j}^k, j \in J_k$ und $\alpha_{F,j}^k, j \in J_k$ nichtnegativ sind und dass die Approximationen, die die dazugehörigen Problemfunktionen nicht von unten annähern (für die also $\beta_{f,j}^k < 0$ bzw. $\beta_{F,j}^k < 0$ gilt), schwächer eingehen, falls $\|x_k - y_j\|$ "groß" ist. Die Parameter γ_f und γ_F werden "klein" gewählt, da nur sichergestellt werden soll, dass $\alpha_{f,j}^k, j \in J_k$ und $\alpha_{F,j}^k, j \in J_k$ keine negativen Werte annehmen. Die Betragsstriche um $\beta_{f,j}^k$ bzw. $\beta_{F,j}^k$ sind nicht notwendig, verbessern jedoch die numerischen Ergebnisse erheblich. Des Weiteren liefert der Algorithmus für $\omega = 1$ und $\omega = 2$ gute Resultate. Die obige Definition von $\alpha_{f,j}^k$ bzw. $\alpha_{F,j}^k$ macht die Speicherung der Hilfspunkte y_j für alle $j \in J_k$ erforderlich. Dies kann umgangen werden, indem wir $\|x_k - y_j\|$ durch die Größen

$$s_j^k := \begin{cases} \|x_j - y_j\| + \sum_{i=j}^{k-1} \|x_{i+1} - x_i\| \geq \|x_k - y_j\|,\ j \in J_k \backslash \{k\} \\ \|x_k - y_k\|, \end{cases}$$

approximieren, welche rekusiv nach

$$s_j^{k+1} := s_j^k + \|x_{k+1} - x_k\|,\ j \in J_k \tag{7.9}$$

upgedatet werden können, so dass wir als Gewichte

$$\alpha_{f,j}^k = \max\{|\beta_{f,j}^k|, \gamma_f (s_j^k)^\omega\},\ j \in J_k \text{ bzw. } \alpha_{F,j}^k = \max\{|\beta_{F,j}^k|, \gamma_F (s_j^k)^\omega\},\ j \in J_k \tag{7.10}$$

erhalten. Statt (7.6) verwenden wir nun

$$\begin{aligned} \min \quad & v + \frac{1}{2} d^\top W_k d \\ \text{u.d.N.} \quad & -\alpha_{f,j}^k + d^\top g_{f,j}^k \leq v,\ j \in J_k, \\ & -\alpha_{F,j}^k + d^\top g_{F,j}^k \leq v,\ j \in J_k, \end{aligned} \tag{7.11}$$

wobei $g_{f,j}^k = g_{f,j}(x_k) = \nabla f_j(x_k) = \xi_j^f + \varrho_j^f G_j^f (x_k - y_j)$ und $g_{F,j}^k = g_{F,j}(x_k) = \nabla F_j(x_k) = \xi_j^F + \varrho_j^F G_j^F (x_k - y_j)$.

7.1.3 Duales Problem

Es soll nun das zu (7.11) duale Problem hergeleitet werden, wobei wir wie in Kapitel (5.6) vorgehen. Die duale Zielfunktion ist gegeben durch

$$\theta(d, v, \lambda_f, \lambda_F) = \inf_{(d,v)\in\mathbb{R}^{n+1}} L(d, v, \lambda_f, \lambda_F)$$

mit

$$\begin{aligned} L(d, v, \lambda_f, \lambda_F) = & \\ & v + \frac{1}{2}d^\top W_k d + \sum_{j\in J_k} \lambda^k_{f,j}(-\alpha^k_{f,j} + d^\top g^k_{f,j} - v) + \sum_{j\in J_k} \lambda^k_{F,j}(-\alpha^k_{F,j} + d^\top g^k_{F,j} - v) \\ = & \frac{1}{2}d^\top W_k d + d^\top \sum_{j\in J_k}(\lambda^k_{f,j} g^k_{f,j} + \lambda^k_{F,j} g^k_{F,j}) + v(1 - \sum_{j\in J_k}(\lambda^k_{f,j} + \lambda^k_{F,j})) \\ & - \sum_{j\in J_k}(\lambda^k_{f,j}\alpha^k_{f,j} + \lambda^k_{F,j}\alpha^k_{F,j}). \end{aligned}$$

Um die duale Zielfunktion zu erhalten, müssen wir demnach das Infimum der Lagrange-Funktion berechnen. Damit dieses angenommen wird, ist die Einführung der Nebenbedingung $\sum_{j\in J_k}(\lambda^k_{f,j} + \lambda^k_{F,j}) = 1$ erforderlich. Das Minimum der Lagrangefunktion bezüglich d ermitteln wir über das notwendige und aufgrund der positiven Definitheit der Matrix W_k auch hinreichende Kriterium

$$\nabla_d L(d, \lambda_f, \lambda_F) = W_k d + \sum_{j\in J_k}(\lambda^k_{f,j} g^k_{f,j} + \lambda^k_{F,j} g^k_{F,j}) = 0.$$

Da positiv definite Matrizen invertierbar sind, folgt

$$d = -W_k^{-1} \sum_{j\in J_k}(\lambda^k_{f,j} g^k_{f,j} + \lambda^k_{F,j} g^k_{F,j}). \tag{7.12}$$

Setzen wir (7.12) in $L(d, \lambda_f, \lambda_F)$ ein, so ergibt sich

$$\begin{aligned} L(\lambda_f, \lambda_F) = & \\ = & \frac{1}{2}\big((W_k^{-1} \sum_{j\in J_k}(\lambda^k_{f,j} g^k_{f,j} + \lambda^k_{F,j} g^k_{F,j}))^\top W_k W_k^{-1} \sum_{j\in J_k}(\lambda^k_{f,j}\alpha^k_{f,j} + \lambda^k_{F,j}\alpha^k_{F,j})\big) \\ & - \big(W_k^{-1} \sum_{j\in J_k}(\lambda^k_{f,j} g^k_{f,j} + \lambda^k_{F,j} g^k_{F,j})\big)^\top \sum_{j\in J_k}(\lambda^k_{f,j} g^k_{f,j} + \lambda^k_{F,j} g^k_{F,j}) \\ & - \sum_{j\in J_k}(\lambda^k_{f,j}\alpha^k_{f,j} + \lambda^k_{F,j}\alpha^k_{F,j}) \end{aligned}$$

$$
\begin{aligned}
&= -\frac{1}{2}((W_k^{-1}\sum_{j\in J_k}(\lambda_{f,j}^k g_{f,j}^k + \lambda_{F,j}^k g_{F,j}^k))^\top \sum_{j\in J_k}(\lambda_{f,j}^k \alpha_{f,j}^k + \lambda_{F,j}^k \alpha_{F,j}^k)) \\
&\quad - \sum_{j\in J_k}(\lambda_{f,j}^k \alpha_{f,j}^k + \lambda_{F,j}^k \alpha_{F,j}^k) \\
&= -\frac{1}{2}\big\|W_k^{-\frac{1}{2}}\sum_{j\in J_k}(\lambda_{f,j}^k g_{f,j}^k + \lambda_{F,j}^k g_{F,j}^k)\big\|^2 - \sum_{j\in J_k}(\lambda_{f,j}^k \alpha_{f,j}^k + \lambda_{F,j}^k \alpha_{F,j}^k),
\end{aligned}
$$

wobei im letzten Schritt ausgenutzt wurde, dass die Inverse einer positiv definiten Matrix ebenfalls positiv definit ist und dass jede positiv definite Matrix eine eindeutig bestimmte positiv definite Wurzel besitzt (siehe Horn und Johnson (1985), Kapitel 7). Wird die Zielfunktion noch mit (-1) multipliziert, um die Maximierung in eine Minimierung umzuwandeln, so erhalten wir das zu (7.11) duale Problem

$$
\begin{aligned}
\min \quad & \frac{1}{2}\big\|W_k^{-\frac{1}{2}}\sum_{j\in J_k}(\lambda_{f,j}^k g_{f,j}^k + \lambda_{F,j}^k g_{F,j}^k)\big\|^2 + \sum_{j\in J_k}(\lambda_{f,j}^k \alpha_{f,j}^k + \lambda_{F,j}^k \alpha_{F,j}^k) \\
\text{u.d.N.} \quad & \sum_{j\in J_k}\left(\lambda_{f,j}^k + \lambda_{F,j}^k\right) = 1, \\
& \lambda_{f,j}^k \geq 0 \text{ und } \lambda_{F,j}^k \geq 0,\ j \in J_k.
\end{aligned}
$$

Die Notwendigkeit der ersten Nebenbedingung wurde bereits erläutert, die Positivitätsforderungen ergeben sich aus dem Dualitätsprinzip (siehe (5.10)). Da die Voraussetzungen des in Bemerkung 5.21 genannten Spezialfalls des starken Dualitätssatzes erfüllt sind, können wir v ermitteln, indem wir den primalen und den dualen Zielfunktionswert gleichsetzen und die Darstellung von d gemäß (7.12) berücksichtigen:

$$
\begin{aligned}
& v + \frac{1}{2}d^\top W_k d + \frac{1}{2}(W_k^{-1}\sum_{j\in J_k}(\lambda_{f,j}^k g_{f,j}^k + \lambda_{F,j}^k g_{F,j}^k))^\top \sum_{j\in J_k}(\lambda_{f,j}^k g_{f,j}^k + \lambda_{F,j}^k g_{F,j}^k) \\
& \quad + \sum_{j\in J_k}(\lambda_{f,j}^k \alpha_{f,j}^k + \lambda_{F,j}^k \alpha_{F,j}^k) = 0 \\
\Rightarrow\ & v + \frac{1}{2}d^\top W_k d + \frac{1}{2}(W_k^{-1}\sum_{j\in J_k}(\lambda_{f,j}^k g_{f,j}^k + \lambda_{F,j}^k g_{F,j}^k))^\top W_k W_k^{-1}\sum_{j\in J_k}(\lambda_{f,j}^k g_{f,j}^k + \lambda_{F,j}^k g_{F,j}^k) \\
& \quad + \sum_{j\in J_k}(\lambda_{f,j}^k \alpha_{f,j}^k + \lambda_{F,j}^k \alpha_{F,j}^k) = 0 \\
\Rightarrow\ & v + d^\top W_k d + \sum_{j\in J_k}(\lambda_{f,j}^k \alpha_{f,j}^k + \lambda_{F,j}^k \alpha_{F,j}^k) = 0 \\
\Rightarrow\ & v = -d^\top W_k d - \sum_{j\in J_k}(\lambda_{f,j}^k \alpha_{f,j}^k + \lambda_{F,j}^k \alpha_{F,j}^k)
\end{aligned}
$$

7.1.4 Subgradientenaggregation

Auch bei der Bundle-Newton-Methode bedienen wir uns des Konzeptes der Subgradientenaggregation, um den Speicher- und Berechnungsaufwand zu reduzieren. Allerdings müssen wir die Strategie aus Abschnitt 6.4 modifizieren. Denn wir haben es nicht mehr mit einer unrestringierten, konvexen, sondern mit einer restringierten, lipschitzstetigen Problemstellung zu tun. Ein weiterer Unterschied besteht darin, dass wir für die Approximation anstelle von stückweise linearen Funktionen stückweise quadratische verwenden. Zur Behebung der Schwierigkeiten, die durch die fehlende Konvexität entstehen, wurden die Approximationsfehler (7.7) durch die Gewichte (7.10) ersetzt. Im Folgenden werden wir zeigen, wie die aus Abschnitt 6.4 bekannte Subgradientenaggregationsmethode abzuändern ist, um angemessen auf die quadratische Approximation sowie die Existenz von Nebenbedingungen zu reagieren. Dazu nehmen wir an, dass das Problem (7.11) unter Berücksichtigung der Subgradientenaggregationsmethode in der k-ten Iteration die Form

$$\begin{aligned} \min \quad & v + \frac{1}{2} d^\top \bar{G}_p^k d \\ \text{u.d.N.} \quad & -\alpha_{f,j}^k + d^\top g_{f,j}^k \leq v,\ j \in J_k, \\ & -\alpha_{f,p}^k + d^\top g_{f,p}^k \leq v, \\ & -\alpha_{F,j}^k + d^\top g_{F,j}^k \leq v,\ j \in J_k, \\ & -\alpha_{F,p}^k + d^\top g_{F,p}^k \leq v \end{aligned} \tag{7.13}$$

mit

$$\begin{aligned} \alpha_{f,j}^k &:= \max[|f(x_k) - f_j^k|, \gamma_f (s_j^k)^\omega],\ j \in J_k, \\ \alpha_{f,p}^k &:= \max[|f(x_k) - f_p^k|, \gamma_f (s_{f,p}^k)^\omega], \\ \alpha_{F,j}^k &:= \max[|F_j^k|, \gamma_F (s_j^k)^\omega],\ j \in J_k, \\ \alpha_{F,p}^k &:= \max[|F_p^k|, \gamma_F (s_{F,p}^k)^\omega] \end{aligned}$$

besitzt und konstruieren ausgehend von (7.13) das k+1-te Teilproblem. Dabei sei $\bar{G}_p^k$ eine positiv definite Matrix. Auf die gleiche Weise wie bei Problem (7.11) kann auch das zu (7.13) duale Problem

$$\begin{aligned} \min \quad & \frac{1}{2} \Big\| (\bar{G}_p^k)^{-\frac{1}{2}} \sum_{j \in J_k} (\lambda_{f,j}^k g_{f,j}^k + \lambda_{f,p}^k g_{f,p}^k + \lambda_{F,j}^k g_{F,j}^k + \lambda_{F,p}^k g_{F,p}^k) \Big\|^2 \\ & \quad + \sum_{j \in J_k} (\lambda_{f,j}^k \alpha_{f,j}^k + \lambda_{f,p}^k \alpha_{f,p}^k + \lambda_{F,j}^k \alpha_{F,j}^k + \lambda_{F,p}^k \alpha_{F,p}^k) \\ \text{u.d.N.} \quad & \sum_{j \in J_k} (\lambda_{f,j}^k + \lambda_{f,p}^k + \lambda_{F,j}^k + \lambda_{F,p}^k) = 1, \\ & \lambda_{f,j}^k \geq 0,\ \lambda_{F,j}^k \geq 0,\ j \in J_k,\ \lambda_{f,p}^k \geq 0 \text{ und } \lambda_{F,p}^k \geq 0 \end{aligned} \tag{7.14}$$

hergeleitet werden. Die Lösung des primalen Problems ergibt sich über

$$d_k = -(\bar{G}_p^k)^{-1}\big(\sum_{j\in J_k}\lambda_{f,j}^k g_{f,j}^k + \lambda_{f,p}^k g_{f,p}^k + \sum_{j\in J_k}\lambda_{F,j}^k g_{F,j}^k + \lambda_{F,p}^k g_{F,p}^k\big),$$
$$v_k = -d_k^\top \bar{G}_p^k d_k - \sum_{j\in J_k}\lambda_{f,j}^k\alpha_{f,j}^k - \lambda_{F,p}^k\alpha_{F,p}^k - \sum_{j\in J_k}\lambda_{F,j}^k\alpha_{F,j}^k - \lambda_{f,p}^k\alpha_{F,p}^k.$$

Im Gegensatz zum unrestringierten Fall liegt im restringierten Fall das Problem vor, dass die Lagrange-Multiplikatoren $\lambda_{f,j}^k$ von f und $\lambda_{F,j}^k$ von F getrennt betrachtet keine Konvexkombinationen bilden. Weil dies jedoch für die Anwendung der Methode erforderlich ist, setzen wir $\lambda_f^k := \lambda_{f,p}^k + \sum_{j\in J_k}\lambda_{f,j}^k$ und $\lambda_F^k := \lambda_{F,p}^k + \sum_{j\in J_k}\lambda_{F,j}^k$ und definieren für alle $j \in J_k$ die skalierten Multiplikatoren

$$\tilde{\lambda}_{f,j}^k := \begin{cases} \lambda_{f,j}^k/\lambda_f^k, & \text{wenn } \lambda_f^k > 0 \\ 1/|J_{k+1}|, & \text{wenn } \lambda_f^k = 0, \end{cases} \qquad \tilde{\lambda}_{f,p}^k := \begin{cases} \lambda_{f,p}^k/\lambda_f^k, & \text{wenn } \lambda_f^k > 0 \\ 1/|J_{k+1}|, & \text{wenn } \lambda_f^k = 0, \end{cases}$$

$$\tilde{\lambda}_{F,j}^k := \begin{cases} \lambda_{F,j}^k/\lambda_F^k, & \text{wenn } \lambda_F^k > 0 \\ 1/|J_{k+1}|, & \text{wenn } \lambda_F^k = 0, \end{cases} \qquad \tilde{\lambda}_{F,p}^k := \begin{cases} \lambda_{F,p}^k/\lambda_F^k, & \text{wenn } \lambda_F^k > 0 \\ 1/|J_{k+1}|, & \text{wenn } \lambda_F^k = 0 \end{cases}$$

sowie die aggregierten Subgradienten

$$\begin{aligned}(\tilde{g}_{f,p}^k, \tilde{f}_p^k, G_{f,p}^{k+1}, \tilde{s}_{f,p}^k) &:= \sum_{j\in J_k}\tilde{\lambda}_{f,j}^k(g_{f,j}^k, f_j^k, \varrho_{f,j}G_{f,j}, s_j^k) + \tilde{\lambda}_{f,p}^k(g_p^k, f_p^k, G_{f,p}^k, s_{f,p}^k),\\ (\tilde{g}_{F,p}^k, \tilde{F}_p^k, G_{F,p}^{k+1}, \tilde{s}_{F,p}^k) &:= \sum_{j\in J_k}\tilde{\lambda}_{F,j}^k(g_{F,j}^k, F_j^k, \varrho_{F,j}G_{F,j}, s_j^k) + \tilde{\lambda}_{F,p}^k(g_p^k, F_p^k, G_p^k, s_{F,p}^k)\end{aligned} \tag{7.15}$$

und darüber hinaus die Größen

$$\begin{aligned}\tilde{\alpha}_{f,p}^k &:= \max[|f(x_k) - \tilde{f}_p^k|, \gamma_f(\tilde{s}_{f,p}^k)^\omega],\\ \tilde{\alpha}_{F,p}^k &:= \max[|F(x_k) - \tilde{F}_p^k|, \gamma_F(\tilde{s}_{F,p}^k)^\omega],\\ \tilde{\alpha}_p^k &:= \lambda_f^k\tilde{\alpha}_{f,p}^k + \lambda_F^k\tilde{\alpha}_{F,p}^k.\end{aligned} \tag{7.16}$$

Es ist leicht zu erkennen, dass es sich bei den aggregierten Subgradienten um Konvexkombinationen handelt.
Wie in Abschnitt 6.4 möchten wir vermeiden, die Hilfspunkte $y_j, j \in J_k$ aus den vergangenen Iterationen zu speichern. Analog zu (7.9) gilt

$$\begin{aligned}s_{f,p}^{k+1} &:= \tilde{s}_{f,p}^k + \|x_{k+1} - x_k\|,\\ s_{F,p}^{k+1} &:= \tilde{s}_{F,p}^k + \|x_{k+1} - x_k\|.\end{aligned}$$

Entsprechend der Herleitungen (6.16) und (6.17) erhalten wir

$$\begin{aligned}g_{f,j}^{k+1} &= g_{f,j}^k + \varrho_{f,j}G_{f,j}(x_{k+1} - x_k),\ j \in J_k,\\ g_{F,j}^{k+1} &= g_{F,j}^k + \varrho_{F,j}G_{F,j}(x_{k+1} - x_k),\ j \in J_k\end{aligned}$$

sowie

$$g_{f,p}^{k+1} = \tilde{g}_{f,p}^{k+1} + G_{f,p}^{k+1}(x_{k+1} - x_k),$$
$$g_{F,p}^{k+1} = \tilde{g}_{F,p}^{k+1} + G_{F,p}^{k+1}(x_{k+1} - x_k).$$

Darüber hinaus verwenden wir die Update-Regeln

$$\begin{aligned}
f_j^{k+1} &:= f_j^k + (x_{k+1} - x_k)^\top g_{f,j}^k + \frac{1}{2}\varrho_{f,j}(x_{k+1} - x_k)^\top G_{f,j}(x_{k+1} - x_k),\ j \in J_k,\\
F_j^{k+1} &:= F_j^k + (x_{k+1} - x_k)^\top g_{F,j}^k + \frac{1}{2}\varrho_{F,j}(x_{k+1} - x_k)^\top G_{F,j}(x_{k+1} - x_k),\ j \in J_k,\\
f_p^{k+1} &:= \tilde{f}_p^k + (x_{k+1} - x_k)^\top \tilde{g}_{f,p}^k + \frac{1}{2}(x_{k+1} - x_k)^\top G_{f,p}^{k+1}(x_{k+1} - x_k),\\
F_p^{k+1} &:= \tilde{F}_p^k + (x_{k+1} - x_k)^\top \tilde{g}_{F,p}^k + \frac{1}{2}(x_{k+1} - x_k)^\top G_{F,p}^{k+1}(x_{k+1} - x_k.
\end{aligned} \tag{7.17}$$

Allerdings werden wir beim Versuch, die Regeln in (7.17) analog zu (6.16) und (6.17) nachzurechnen, aufgrund der quadratischen Terme scheitern. Diese Regeln sind also nur als Näherungen anzusehen. Mit Hilfe der Damping-Parameter $\varrho_{f,j} \in [0,1], j \in J_k$, und $\varrho_{F,j} \in [0,1]$, $j \in J_k$ wird der Einfluß der quadratischen Terme "gedämpft". Falls in drei aufeinanderfolgenden Iterationsschritten (die Anzahl wurde empirisch ermittelt) kein wesentlicher Abstieg in der Zielfunktion erreicht wurde, so wird $\varrho_{f,k+1} := \varrho_{F,k+1} = 0$ gesetzt, ansonsten gilt $\varrho_{f,k+1} := \min[1, C_G/\|G_{f,k+1}\|_S]$ bzw. $\varrho_{F,k+1} := \min[1, C_G/\|G_{F,k+1}\|_S]$. Aufgrund der Wahl von $\varrho_{f,k}$ ist die Beschränktheit von $\{\varrho_f G_{f,k}\}$ sichergestellt, da

$$\varrho_{f,k}\|G_k\|_S \leq C_G \tag{7.18}$$

gilt. Die gleiche Aussage erhalten wir auch in Bezug auf F.
Für die Nebendingungen, die in der Iteration k+1 neu hinzukommen, berechnen wir $f_{k+1} = f(y_{k+1}), F_{k+1} = F(y_{k+1}), g_{f,k+1} \in \partial f(y_{k+1}), g_{F,k+1} \in \partial F(y_{k+1})$, Näherungen $G_{f,k+1}$ von $\nabla\nabla f(y_{k+1})$ und $G_{F,k+1}$ von $\nabla\nabla F(y_{k+1})$ und anschließend die Größen

$$\begin{aligned}
s_{k+1}^{k+1} &:= \|x_{k+1} - y_{k+1}\|,\\
f_{k+1}^{k+1} &:= f_{k+1} + (x_{k+1} - y_{k+1})^\top g_{f,k+1} + \frac{1}{2}\varrho_{k+1}(x_{k+1} - y_{k+1})^\top G_{f,k+1}(x_{k+1} - y_{k+1}),\\
F_{k+1}^{k+1} &:= F_{k+1} + (x_{k+1} - y_{k+1})^\top g_{F,k+1} + \frac{1}{2}\varrho_{k+1}(x_{k+1} - y_{k+1})^\top G_{F,k+1}(x_{k+1} - y_{k+1}),\\
g_{f,k+1}^{k+1} &:= \tilde{g}_{f,p}^k + G_{f,p}^{k+1}(x_{k+1} - x_k),\\
g_{F,k+1}^{k+1} &:= \tilde{g}_{F,p}^k + G_{F,p}^{k+1}(x_{k+1} - x_k).
\end{aligned}$$

Die Ermittlung des Punktes y_{k+1} wird in Abschnitt 7.2 behandelt. In Abschnitt 7.4 werden wir auf die Bestimmung von $\bar{G}_p^{k+1}$ eingehen und eine geeignete Wahl der Indexmenge J_{k+1}

vorstellen. Nun können wir das Teilproblem für die k+1-te Iteration aufstellen:

$$\begin{aligned}
\min \quad & v + \frac{1}{2} d^\top \bar{G}_p^{k+1} d \\
\text{u.d.N.} \quad & -\alpha_{f,j}^{k+1} + d^\top g_{f,j}^{k+1} \leq v \text{ für alle } j \in J_{k+1}, \\
& -\alpha_{f,p}^{k+1} + d^\top g_{f,p}^{k+1} \leq v, \\
& -\alpha_{F,j}^{k+1} + d^\top g_{F,j}^{k+1} \leq v \text{ für alle } j \in J_{k+1}, \\
& -\alpha_{F,p}^{k+1} + d^\top g_{F,p}^{k+1} \leq v
\end{aligned}$$

mit

$$\begin{aligned}
\alpha_{f,j}^{k+1} &:= \max[|f(x_{k+1}) - f_j^{k+1}|, \gamma_f (s_j^{k+1})^\omega],\ j \in J_{k+1}, \\
\alpha_{f,p}^{k+1} &:= \max[|f(x_{k+1}) - f_p^{k+1}|, \gamma_f (s_{f,p}^{k+1})^\omega], \\
\alpha_{F,j}^{k+1} &:= \max[|F_j^{k+1}|, \gamma_F (s_j^{k+1})^\omega],\ j \in J_{k+1}, \\
\alpha_{F,p}^{k+1} &:= \max[|F_p^{k+1}|, \gamma_F (s_{F,p}^{k+1})^\omega].
\end{aligned}$$

7.2 Schrittweitenbestimmung

In diesem Abschnitt werden wir die Gründe für die Notwendigkeit der Anwendung einer Schrittweitenstrategie verdeutlichen und einen für die Bundle-Newton-Methode geeigneten Schrittweitenalgorithmus vorstellen. Wir gehen davon aus, dass wir in der k-ten Iteration eine Lösung (d_k, v_k) des quadratischen Teilproblems (7.13) gefunden haben. Auch wenn $\hat{H}_k(x_k + d, x_k)$ in d_k ein Minimum annimmt, ist der größte Abstieg nicht unbedingt durch die direkte Wahl $x_{k+1} := x_k + d_k$ gegeben, denn $\hat{H}$ ist nur eine Approximation der Improvementfunktion H. Man kann versuchen, eine Schrittweite $t_k \in (0, 1]$ zu ermitteln, die den Bedingungen

$$t_k \approx \arg\min_{t \in (0,1]} f(x_k + t d_k) \text{ und } F(x_k + t_k d_k) \leq 0$$

genügt. Allerdings garantiert auch die Wahl $x_{k+1} := x_k + t_k d_k$ nicht, dass der Abstieg in der Zielfunktion ausreichend groß ist. Es kann sogar vorkommen, dass d_k nicht einmal eine Abstiegsrichtung darstellt. In solchen Fällen ist es möglich, dass unendlich viele Iterationen durchgeführt werden, ohne dass der Zielfunktionswert in bedeutendem Maße abnimmt, was sich negativ auf das Konvergenzverhalten auswirkt. Um die genannten Schwierigkeiten zu beheben ermitteln wir zwei Schrittweiten t_L^k und t_R^k mit $0 \leq t_L^k \leq t_R^k \leq 1$ nach der folgenden Strategie. Wir versuchen, die größte Schrittweite $t_L^k \in [0, 1]$, welche die Bedingungen

(a) $f(x_k + t_L^k d_k) \leq f(x_k) + m_L t_L^k \tilde{v}_k$,

(b) $F(x_k + t_L^k d_k) \leq 0$,

(c) $t_L^k \geq \bar{t}$

erfüllt, zu finden, wobei $m_L \in (0, \frac{1}{2})$ ein Liniensuchparameter und $\bar{t} \in (0,1)$ eine Schranke ist. Existiert ein solcher Parameter, so führen wir einen "wesentlichen Schritt" von x_k nach $x_{k+1} := x_k + t_L^k d_k$ durch und setzen $y_{k+1} := x_{k+1}$. Die Größe $\tilde{v}_k = -\|(\bar{G}_p^k)^{-\frac{1}{2}} \tilde{g}_p^k\|^2 - \tilde{\alpha}_p^k$ stellt den vorhergesagten Abstieg von f in x_k dar. Wir verweisen auf die Verbindung zur Variable w_k in Abschnitt 7.3. Es gilt $\tilde{v}_k \leq 0$. Im Falle von $\tilde{v}_k = 0$ stoppt der Algorithmus (auch dies wird in Abschnitt 7.3 ersichtlich), weswegen wir von $\tilde{v}_k < 0$ ausgehen können. Somit gilt aufgrund von (a) bei Vorliegen eines wesentlichen Schrittes $f(x_{k+1}) < f(x_k)$. Falls die Forderungen (a) und (b) erfüllt sind, aber $0 < t_L^k < \bar{t}$ gilt, so erfolgt nur ein "kleiner Schritt" mit $x_{k+1} := x_k + t_L^k d_k$ und $y_{k+1} := y_k + t_R^k d_k$. Falls die Kriterien (a) und (b) sogar nur für $t_L = 0$ eingehalten werden können, so setzen wir $x_{k+1} := x_k$ und $y_{k+1} := y_k + t_R^k d_k$. Es handelt sich hierbei um einen Nullschritt. Wenn ein kleiner Schritt oder ein Nullschritt vorliegen, dann sorgt entweder

(d) $-\alpha_{f,k+1}^{k+1} + d_k^\top g_{f,k+1}^{k+1} \geq m_R \tilde{v}_k$, wenn $F(y_{k+1}) \leq 0$

oder

(e) $-\alpha_{F,k+1}^{k+1} + d_k^\top g_{F,k+1}^{k+1} \geq m_R \tilde{v}_k$, wenn $F(y_{k+1}) > 0$

dafür, dass mindestens einer der beiden neuen Subgradienten die nächste stückweise quadratische Approximation von H^{k+1} signifikant verändert. Dies gewährleistet, dass der Algorithmus nicht in Nichtdifferenzierbarkeitsstellen hängenbleibt. Bei $m_R \in (m_L, 1)$ handelt es sich um einen Liniensuchparameter. Des Weiteren wird

(f) $\|x_{k+1} - y_{k+1}\| \leq C_S$ für $C_S > 0$

gefordert, um zu vermeiden, dass unnötige Subgradienteninformationen herangezogen werden. Die Bedingung (f) ist Teil einer sogenannten Reset-Strategie, auf die wir in Abschnitt 7.4 noch näher eingehen. Nun stellen wir einen Liniensuchalgorithmus vor, der Schrittweiten t_R^k und t_L^k erzeugt, die die Forderungen (a)-(f) erfüllen.

Liniensuchalgorithmus:

S0 Setze $t_L^k := 0$ und $t := t_U := 1$. Wähle $\zeta \in (0, \frac{1}{2}), \vartheta \geq 1$.

S1 Wenn $f(x_k + td_k) \leq f(x_k) + m_L t v_k$ und $F(x_k + td_k) \leq 0$, so setze $t_L^k := t$, ansonsten setze $t_U := t$.

S2 Wenn $t_L^k \geq t_0$, setze $t_R^k := t_L^k$ und STOPP.

S3 Falls $F(x_k+td_k) \leq 0$, ermittle einen Subgradienten $g \in \partial f(x_k+td_k)$, eine symmetrische Matrix G als Näherung der Hessematrix von f in $x_k + td_k$ und

$$\begin{aligned} \varrho &:= \begin{cases} \min[1, C_G/\|G\|_S], & \text{wenn } i_n \leq 3, \\ 0 & \text{sonst,} \end{cases} \\ f &:= f(x_k + td_k) + (t_L^k - t)g^\top d_k + \frac{1}{2}\varrho(t_L^k - t)^2 d_k^\top G d_k, \\ \beta &:= \max[|f - f(x_k + t_L^k d_k)|, \gamma_f |t_L^k - t|^\omega \|d_k\|^\omega], \end{aligned}$$

ansonsten ermittle einen Subgradienten $g \in \partial F(x_k + td_k)$, eine symmetrische Matrix G als Näherung der Hessematrix von F in $x_k + td_k$ und

$$\begin{aligned} \varrho &:= \begin{cases} \min[1, C_G/\|G\|_S], & \text{wenn } i_n \leq 3, \\ 0 & \text{sonst,} \end{cases} \\ F &:= F(x_k + td_k) + (t_L^k - t)g^\top d_k + \frac{1}{2}\varrho(t_L^k - t)^2 d_k^\top G d_k, \\ \beta &:= \max[|F|, \gamma_F |t_L^k - t|^\omega \|d_k\|^\omega]. \end{aligned}$$

(Wenn der Algorithmus abbricht, dann ist $x_k + t_L d_k$ als x_{k+1} und $x_k + td_k$ als y_{k+1} anzusehen)

S4 Wenn $-\beta + d_k^\top(g + \varrho(t_L^k - t)Gd_k) \geq m_R\tilde{v}_k$ und $(t - t_L^k)\|d_k\| \leq C_S$, dann setze $t_R^k := t$ und STOPP.

S5 Wähle $t \in [t_L^k + \zeta(t_U - t_L^k)^\vartheta, t_U - \zeta(t_U - t_L^k)^\vartheta]$ mit Hilfe einer Interpolationsmethode und gehe zu **S1**.

Es sei angemerkt, dass die Bedingungen in **S4** äquivalent zu

$$\alpha_{k+1}^{k+1} + d_k^\top g_{k+1}^{k+1} \geq m_R\tilde{v}_k,\ \|x_{k+1} - y_{k+1}\| \leq C_S$$

sind. Wir werden nur die Größe f nachrechnen (die restlichen Berechnungen erfolgen analog). Zum Nachweis ziehen wir die Darstellung von f_{k+1}^{k+1} aus Unterabschnitt 7.1.4 heran und setzen anschließend die entsprechenden Größen ein:

$$\begin{aligned} f &= f_{k+1} + (x_{k+1} - y_{k+1})^\top g_{f,k+1} + \frac{1}{2}\varrho_{k+1}(x_{k+1} - y_{k+1})^\top G_{f,k+1}(x_{k+1} - y_{k+1}), \\ &= f(x_k + td_k) + (x_k + t_L d_k - (x_k + td_k))^\top g + \\ &\quad + \frac{1}{2}\varrho(x_k + t_L d_k - (x_k + td_k))^\top G(x_k + t_L d_k - (x_k + td_k)) \\ &= f(x_k + td_k) + (t_L - t)d_k^\top g + \frac{1}{2}\varrho(t_L - t)d_k^\top G(t_L - t)d_k \end{aligned}$$

Nun untersuchen wir die Konvergenz des Liniensuchalgorithmus.

Satz 7.3. *Für die Funktionen f und F seien die folgenden "semismoothness"-Bedingungen erfüllt:*

(i) Für $x \in \mathbb{R}^n, d \in \mathbb{R}^n$ und Folgen $\{\bar{g}_i\} \subset \mathbb{R}^n$ und $\{t^i\} \subset \mathbb{R}^+$ mit $\bar{g}_i \in \partial f(x + t^i d)$ und $t^i \downarrow 0$ gelte:

$$\limsup_{i\to\infty} \bar{g}_i^\top d \geq \liminf_{i\to\infty} \frac{[f(x + t^i d) - f(x)]}{t^i}$$

(ii) Für $x \in \mathbb{R}^n, d \in \mathbb{R}^n$ und Folgen $\{\bar{g}_i\} \subset \mathbb{R}^n$ und $\{t^i\} \subset \mathbb{R}^+$ mit $\bar{g}_i \in \partial F(x + t^i d)$, $F(x + t^i d) > 0$, $F(x) = 0$ und $t^i \downarrow 0$ gelte:

$$\limsup_{i\to\infty} \bar{g}_i^\top d \geq \liminf_{i\to\infty} \frac{[F(x + t^i d) - F(x)]}{t^i}$$

Dann stoppt der Algorithmus mit $t_L^k = t_L$, $t_R^k = t$ und die Bedingungen $(a) - (d)$ sind erfüllt.

Beweis. Beim Eintritt in die Liniensuchiteration ist $\tilde{v}_k < 0$. Wir führen einen Widerspruchsbeweis und nehmen deswegen an, dass die Liniensuche nicht terminiert. Es seien t^i, t_U^i, g^i, ϱ^i, G^i und β^i die Werte, die von $t, t_L, t_U, g, \varrho, G$ und β in der i-ten Iteration angenommen werden. Es ist leicht zu erkennen, dass $t^i \in \{t_L^i, t_U^i\}$ für alle i gilt. Mit Hilfe von Induktion kann man $(t_U^i - t_L^i)^{(\vartheta-1)} \leq 1$ und $t_L^i \leq t_L^{i+1} \leq t_U^{i+1} \leq t_U^i$ für alle i nachweisen. Somit ist $\{t_U^i\}$ eine monoton fallende nach unten beschränkte und $\{t_U^i\}$ eine monoton steigende nach oben beschränkte Folge. Es existiert also ein $\tilde{t} \geq 0$ mit $t_L^i \uparrow \tilde{t}$ und $t_U^i \downarrow \tilde{t}$. Wir definieren die Menge

$$S = \{t \geq 0 \,|\, f(x_k + td_k) \leq f(x_k) + m_L t v_k \text{ und } F(x_k + td_k) \leq 0\}.$$

Da nach S1 $\{t_L^i\}$ eine Teilmenge von S ist, außerdem $t_L^i \uparrow \tilde{t}$ gilt und sowohl f als auch F stetig sind, erhalten wir

$$f(x_k + \tilde{t}d_k) - f(x_k) \leq m_L \tilde{t} v_k \text{ und } F(x_k + \tilde{t}d_k) \leq 0, \tag{7.19}$$

d.h. $\tilde{t} \in S$.
Aufgrund von $t^i \to \tilde{t}$ und $t_L^i \uparrow \tilde{t}$ gilt $(t^i - t_L^i)\|d_k\| \leq C_S$ für hinreichend große i. Daraus folgt

$$-\beta^i + d_k^\top (g^i + \varrho^i (t_L^i - t^i) G^i d_k) < m_R v_k \tag{7.20}$$

für hinreichend große i, denn laut Voraussetzung bricht die Liniensuche nicht ab. Wegen $t^i \to \tilde{t}$, $t_L^i \downarrow \tilde{t}$, der Stetigkeit von f und F, der lokalen Beschränktheit der Abbildung g^i nach Satz 5.5(iii) sowie der Beschränktheit von $\{\varrho^i G^i\}$ nach (7.18) erhalten wir $\beta^i \to 0$, $(t_L^i - t^i)\varrho^i d_k^\top G^i d_k \to 0$, so dass sich aus (7.20)

$$\limsup_{i\to\infty} d_k^\top g^i \leq m_R \tilde{v}_k \tag{7.21}$$

ergibt.
Betrachte nun die Menge $I = \{i \,|\, t^i \notin S\}$. Wir zeigen, dass diese Menge unendlich viele Elemente enthält. Dazu nehmen wir an, dass sie endlich ist, d.h. S enthalte unendlich viele Elemente. Dann gibt es ein $i_0 \in I$, so dass $t^i \in S$ für alle $i > i_0$ (Beachte, dass $I \neq \emptyset$, da auf jeden Fall $t^0 \in I$. Wäre dies nicht der Fall, so wird in S1 $t_L = 1$ gesetzt, was in Schritt S2 aufgrund von $t_0 \in (0, 1)$ zum Abbruch führt; dies stellt aber einen Widerspruch zur Annahme, dass die Liniensuche nicht terminiert, dar). Das bedeutet, dass t_U^i für $i > i_0$ konstant gleich $t_U^{i_0}$ bleibt, d.h. $t_U^i = t_U^{i_0}$ für alle $i > i_0$. Da aber $t_U^i = t_U^{i_0} \downarrow \tilde{t}$ und $i_0 \in I$ gilt, folgt $\tilde{t} \notin S$, so dass sich ein Widerspruch ergibt. Es handelt sich also bei I um eine Menge mit unendlich vielen Elementen.
Für alle $i \in I$ gilt entweder $f(x+t^i d) > f(x)+m_L t^i \tilde{v}_k$ oder $F(x+t^i d) > 0$. Zunächst nehmen wir an, dass es eine unendliche Teilmenge $\tilde{I} \subset I$ mit

$$f(x + t^i d) > f(x) + m_L t^i \tilde{v}_k \text{ für alle } i \in \tilde{I} \tag{7.22}$$

gibt. Mit (7.19) sowie der "semismoothness"-Annahme folgt

$$m_L v_k \leq \liminf_{i\to\infty} \inf_{i\in I} \frac{f(x_k + \tilde{t}d_k + (t^i - \tilde{t})d_k) - f(x_k + \tilde{t}d_k)}{t^i - \tilde{t}} \leq \limsup_{i\to\infty} \sup_{i\in I} d_k^\top g^i, \tag{7.23}$$

wobei $g^i \in \partial f(x_k + t^i d_k)$. Mit (7.21) folgt $m_L \tilde{v}_k \leq m_R \tilde{v}_k$, was einen Widerspruch zu $0 < m_L < m_R < 1$ und $\tilde{v}_k < 0$ darstellt.
Da es auch möglich ist, dass (7.22) nicht für unendlich viele $i \in I$ erfüllt ist, nehmen wir nun an, dass

$$F(x + t^i d) > 0 \tag{7.24}$$

für alle $i \in I$ gilt. Mit $t^i \downarrow \tilde{t}$, $F(x + \tilde{t}d) \leq 0$ sowie der Stetigkeit von F folgt aus (7.24)

$$F(x + \tilde{t}d) = 0$$

und somit

$$\liminf_{i\to\infty, i\in\tilde{I}} \frac{F(x + \tilde{t}d + (t^i - \tilde{t})d) - F(x + \tilde{t}d)}{t^i - \tilde{t}} \geq 0 > m_L \tilde{v}_k, \tag{7.25}$$

weil $m_L \tilde{v}_k < 0$. Mit (7.21) und der "semismoothness"-Annahme erhalten wir

$$\liminf_{i\to\infty, i\in\tilde{I}} \frac{F(x + \tilde{t}d + (t^i - \tilde{t})d) - F(x + \tilde{t}d)}{t^i - \tilde{t}} \leq m_R \tilde{v}_k,$$

so dass wir mit (7.25) auch im vorliegenden Fall analog zu oben $m_L \tilde{v}_k \leq m_R \tilde{v}_k$ erhalten, was wiederum einen Widerspruch zu $0 < m_L < m_R < 1$ und $\tilde{v}_k < 0$ darstellt. Es ergibt sich somit insgesamt ein Widerspruch, d.h. die Liniensuche terminiert. Es ist leicht zu zeigen, dass die Bedingungen (a)-(f) erfüllt sind. □

7.3 Abbruchkriterium

Falls eine stetig differenzierbare Funktion f vorliegt, so stellt

$$\|\nabla f(x)\| = 0 \tag{7.26}$$

eine notwendige Bedingung für ein Minimum dar. Aufgrund der Stetigkeit wird der Gradient immer kleiner, wenn wir uns einer Minimalstelle nähern. Somit ist die Verwendung des Abbruchkriteriums $\|\nabla f(x)\| < \varepsilon$ für ein $\varepsilon > 0$ gerechtfertigt. In Kapitel 6 wurde bereits erläutert, dass wir im nichtdifferenzierbaren Fall kein geeignetes Abbruchkriterium erhalten, wenn wir den Gradienten in (7.26) pauschal durch einen beliebigen Subgradienten ersetzen. Durch

$$0 \in \partial H(x, x) \tag{7.27}$$

ist nach Satz 5.14 das zu (7.26) analoge Kriterium für die nicht notwendig differenzierbare, aber lipschitzstetige Funktion H gegeben. Ein beliebiger Subgradient aus $\partial H(x, x)$ erfüllt diese Bedingung aber nicht unbedingt, so dass auch (7.27) nicht zum Ziel führt. Allerdings liefert uns die Anwendung der Subgradientenaggregationsmethode eine Näherung des Gradienten, nämlich den aggregierten Subgradienten $\tilde{g}_p^k$. Wir können den Test $\|\tilde{g}_p^k\| \leq \varepsilon$ für ein $\varepsilon > 0$ nicht direkt verwenden, da die stückweise quadratische Approximation oft nicht ausreichend genau ist. Eine Verbesserung erreichen wir mit Hilfe der Größe $\tilde{\alpha}_p^k$ aus (7.16), die als Maß für die Qualität der Approximation anzusehen ist, und der Stabilisierungsmatrix $\bar{G}_p^k$. In der k-ten Iteration verwenden wir deswegen den Parameter

$$w_k := \frac{1}{2}\|(\bar{G}_p^k)^{-\frac{1}{2}}\tilde{g}_p^k\|^2 + \tilde{\alpha}_p^k.$$

Der Algorithmus stoppt, wenn $w_k < \varepsilon$ für ein gegebenes $\varepsilon > 0$, denn dann ist nicht nur die Norm des approximierten Gradienten, sondern auch der Approximationsfehler hinreichend klein.
Wir gehen abschließend auf die Verbindung zwischen $\tilde{v}_k$ aus Unterabschnitt 7.2 und w_k ein. Es gilt $\tilde{v}_k = -\{\|(\bar{G}_p^k)^{-\frac{1}{2}}\tilde{g}_p^k\|^2 + \tilde{\alpha}_p^k\} \leq -\{\frac{1}{2}\|(\bar{G}_p^k)^{-\frac{1}{2}}\tilde{g}_p^k\|^2 + \tilde{\alpha}_p^k\} = -w_k$. Somit können wir für $\tilde{v} = 0$ auf $w_k = 0$ schließen. Folglich ist $\tilde{v} < 0$, solange noch kein Minimum gefunden wurde.

7.4 Der Bundle-Newton-Algorithmus

Nun stellen wir den vollständigen Bundle-Newton-Algorithmus vor. Im Anschluss daran sind allerdings noch einige Bemerkungen notwendig. Auf die konkrete Vorgehensweise bei der Implementierung werden wir in einem eigens dafür vorgesehenen Abschnitt eingehen.

Bundle-Newton-Algorithmus:

S0 Initialisierung
Wähle einen Startpunkt $x_1 \in \mathbb{R}^n$ sowie die Parameter $\varepsilon \geq 0$, $\gamma_f > 0$, $\gamma_F > 0$, die Anzahl $M \geq 2$ der verwendeten Subgradienten, $t_0 \in (0,1)$, $C_S > 0$, $C_G > 0$, einen Matrix-Auswahl-Paramater $i_m \geq 0$, einen Bundle-Reset-Parameter $i_r \geq 0$ und $\omega \geq 1$. Setze $y_1 := x_1$ und berechne $f(y_1), F(y_1), g_{f,1} \in \partial f(y_1), g_{F,1} \in \partial F(y_1)$ sowie symmetrische Matrizen $G_{f,1}$ und $G_{F,1}$ als Näherungen der Hessematrizen von f bzw. F im Punkt y_1. Initialisiere den Iterationenzähler $k := 1$, die Zahl $i_n := 0$ der aufeinanderfolgenden Null- und kleinen Schritte, die Zahl $i_s := 0$ der wesentlichen Schritte seit dem letzten Bundle-Reset, die Indexmenge $J_1 := \{1\}$, $\varrho_{f,1} := \varrho_{F,1} := 1, s^1_{f,p} := s^1_{F,p} := s^1 := 0$, $g^1_{f,p} := g_{f,1}, g^1_{F,p} := g_{F,1}, f^1_p := f^1_1 := f(y_1), F^1_p := F^1_1 := F(y_1)$, $G^1_{f,p} := G_{f,1}$ und $G^1_{F,p} := G_{F,1}$.

S1 Bestimmung der Suchrichtung
Wenn die Schritte $k-1$ und $k-2$ beide wesentlich sind und $\lambda^{k-1}_{f,k-1} := \lambda^{k-1}_{F,k-1} := 1$ oder wenn $i_s > i_r$, dann setze $G := G_{f,k} + G_{F,k}$, ansonsten setze $G := G^k_{f,p} + G^k_{F,p}$. Wenn $i_n \leq i_m$, dann ersetze G durch ihre positiv definite Approximation $\bar{G}^k_p$, ansonsten setze $\bar{G}^k_p := \bar{G}^{k-1}_p$. Finde die Lösung (d_k, v_k) des k-ten quadratischen Teilproblems

$$
\begin{aligned}
\text{min} \quad & v + \frac{1}{2} d^\top \bar{G}^k_p d \\
\text{bzgl.} \quad & (d, v) \in \mathbb{R}^n \times \mathbb{R} \\
\text{unter} \quad & -\alpha^k_{f,j} + d^\top g^k_{f,j} \leq v, \; j \in J_k, \\
& -\alpha^k_{f,p} + d^\top g^k_{f,p} \leq v, \text{ wenn } i_s \leq i_r, \\
& -\alpha^k_{F,j} + d^\top g^k_{F,j} \leq v, \; j \in J_k, \\
& -\alpha^k_{F,p} + d^\top g^k_{F,p} \leq v, \text{ wenn } i_s \leq i_r,
\end{aligned}
$$

wobei

$$
\begin{aligned}
&\alpha^k_{f,j} := \max[|f^k_j - f(x_k)|, \gamma_f (s^k_j)^\omega], \; j \in J_k, \\
&\alpha^k_{f,p} := \max[|f^k_p - f(x_k)|, \gamma_f (s^k_{f,p})^\omega], \text{ wenn } i_s \leq i_r, \\
&\alpha^k_{F,j} := \max[|F^k_j|, \gamma_F (s^k_j)^\omega], \; j \in J_k, \\
&\alpha^k_{F,p} := \max[|F^k_p|, \gamma_F (s^k_{F,p})^\omega], \text{ wenn } i_s \leq i_r,
\end{aligned}
$$

welche sich über die Lösung des dazugehörigen Dualproblems

$$
\begin{aligned}
\min \quad & \frac{1}{2}\|H_k \sum_{j\in J_k}(\lambda_{f,j}^k g_{f,j}^k + \lambda_{f,p}^k g_{f,p}^k + \lambda_{F,j}^k g_{F,j}^k + \lambda_{F,p}^k g_{F,p}^k)\|^2 \\
& \quad + \sum_{j\in J_k}(\lambda_{f,j}^k \alpha_{f,j}^k + \lambda_{f,p}^k \alpha_{f,p}^k + \lambda_{F,j}^k \alpha_{F,j}^k + \lambda_{F,p}^k \alpha_{F,p}^k) \\
\text{bzgl.} \quad & \lambda_{f,j}^k,\ j\in J_k,\ \lambda_{f,p}^k,\ \lambda_{F,j}^k, j\in J_k,\ \lambda_{F,p}^k \\
\text{unter} \quad & \sum_{j\in J_k}(\lambda_{f,j}^k + \lambda_{f,p}^k + \lambda_{F,j}^k + \lambda_{F,p}^k) = 1, \\
& \lambda_{f,j}^k \geq 0, j\in J_k, \lambda_{f,p}^k \geq 0, \lambda_{F,j}^k \geq 0, j\in J_k, \lambda_{F,p}^k \geq 0, \\
& \lambda_{f,p}^k = \lambda_{F,p}^k = 0,\ \text{wenn } i_s > i_r
\end{aligned}
\tag{7.28}
$$

mit

$$
\begin{aligned}
d_k &:= -H_k^2(\sum_{j\in J_k}\lambda_{f,j}^k g_{f,j}^k + \lambda_{f,p}^k g_{f,p}^k + \sum_{j\in J_k}\lambda_{F,j}^k g_{F,j}^k + \lambda_{F,p}^k g_{F,p}^k), \\
v_k &:= -d_k^\top \bar{G}_p^k d_k - \sum_{j\in J_k}\lambda_{f,j}^k \alpha_{f,j}^k - \lambda_{f,p}^k \alpha_{f,p}^k - \sum_{j\in J_k}\lambda_{F,j}^k \alpha_{F,j}^k - \lambda_{F,p}^k \alpha_{F,p}^k
\end{aligned}
$$

ergibt, wobei $H_k := (\bar{G}_p^k)^{-\frac{1}{2}}$ ist. Wenn $i_s > i_r$, so setze $i_s := 0$. Berechne

$$
\tilde{\lambda}_{f,j}^k := \begin{cases} \lambda_{f,j}^k/\lambda_f^k, & \text{wenn } \lambda_f^k > 0 \\ 1/|J_{k+1}|, & \text{wenn } \lambda_f^k = 0, \end{cases} \qquad \tilde{\lambda}_{f,p}^k := \begin{cases} \lambda_{f,p}^k/\lambda_f^k, & \text{wenn } \lambda_f^k > 0 \\ 1/|J_{k+1}|, & \text{wenn } \lambda_f^k = 0, \end{cases}
$$

$$
\tilde{\lambda}_{F,j}^k := \begin{cases} \lambda_{F,j}^k/\lambda_F^k, & \text{wenn } \lambda_F^k > 0 \\ 1/|J_{k+1}|, & \text{wenn } \lambda_F^k = 0, \end{cases} \qquad \tilde{\lambda}_{F,p}^k := \begin{cases} \lambda_{F,p}^k/\lambda_F^k, & \text{wenn } \lambda_F^k > 0 \\ 1/|J_{k+1}|, & \text{wenn } \lambda_F^k = 0, \end{cases}
$$

$$
(\tilde{g}_{f,p}^k, \tilde{f}_p^k, G_{f,p}^{k+1}, \tilde{s}_{f,p}^k) := \sum_{j\in J_k}\tilde{\lambda}_{f,j}^k(g_{f,j}^k, f_j^k, \varrho_{f,j}G_{f,j}, s_j^k) + \tilde{\lambda}_{f,p}^k(g_p^k, f_p^k, G_{f,p}^k, s_{f,p}^k),
$$

$$
(\tilde{g}_{F,p}^k, \tilde{F}_p^k, G_{F,p}^{k+1}, \tilde{s}_{F,p}^k) := \sum_{j\in J_k}\tilde{\lambda}_{F,j}^k(g_{F,j}^k, F_j^k, \varrho_{F,j}G_{F,j}, s_j^k) + \tilde{\lambda}_{F,p}^k(g_p^k, F_p^k, G_p^k, s_{F,p}^k),
$$

$$
\begin{aligned}
\tilde{\alpha}_{f,p}^k &:= \max[|\tilde{f}_p^k - f(x_k)|, \gamma_f(\tilde{s}_{f,p}^k)^\omega], \\
\tilde{\alpha}_{F,p}^k &:= \max[|\tilde{F}_p^k - F(x_k)|, \gamma_F(\tilde{s}_{F,p}^k)^\omega], \\
\tilde{\alpha}_p^k &:= \lambda_f^k\tilde{\alpha}_{f,p}^k + \lambda_F^k\tilde{\alpha}_{F,p}^k, \\
\tilde{g}_p^k &:= \lambda_f^k\tilde{g}_{f,p}^k + \lambda_F^k\tilde{g}_{F,p}^k, \\
\tilde{v}_k &:= -\|H_k\tilde{g}_p^k\|^2 - \tilde{\alpha}_p^k, \\
w_k &:= \frac{1}{2}\|H_k\tilde{g}_p^k\|^2 + \tilde{\alpha}_p^k.
\end{aligned}
$$

S2 Abbruchkriterium
Wenn $w_k \leq \varepsilon$, STOPP.

S3 Schrittweitenbestimmung
Ermittle mit Hilfe des Liniensuchalgorithmus aus Abschnitt 7.2 Schrittweiten t_L^k und t_R^k und setze $x_{k+1} := x_k + t_L^k d_k$ und $x_{k+1} := x_k + t_R^k d_k$. Berechne $f_{k+1} := f(y_{k+1})$, $F_{k+1} := F(y_{k+1})$, $g_{f,k+1} \in \partial f(y_{k+1})$, $g_{F,k+1} \in \partial F(y_{k+1})$ sowie symmetrische Matrizen $G_{f,k+1}$ und $G_{F,k+1}$ als Näherungen für die Hessematrizen von f und F in y_{k+1}. Falls $t_L^k < t_0$, setze $i_n := i_n + 1$, ansonsten setze $i_n := 0$ und $i_s := i_s + 1$.

S4 Update
Falls $i_n \leq 3$, setze $\varrho_{f,k+1} := \min[1, C_G/\|G_{f,k+1}\|_S]$ und
$\varrho_{F,k+1} := \min[1, C_G/\|G_{F,k+1}\|_S]$, ansonsten setze $\varrho_{f,k+1} := \varrho_{F,k+1} = 0$.
Berechne

$$
\begin{aligned}
s_j^{k+1} &= s_j^k + \|x_{k+1} - x_k\|,\ j \in J_k,\\
s_{k+1}^{k+1} &= \|x_{k+1} - y_{k+1}\|,\\
s_{f,p}^{k+1} &= \tilde{s}_{f,p}^k + \|x_{k+1} - x_k\|,\\
s_{F,p}^{k+1} &= \tilde{s}_{F,p}^k + \|x_{k+1} - x_k\|,\\
f_j^{k+1} &= f_j^k + (x_{k+1} - x_k)^\top g_{f,j}^k + \frac{1}{2}\varrho_{f,j}(x_{k+1} - x_k)^\top G_{f,j}(x_{k+1} - x_k),\ j \in J_k,\\
F_j^{k+1} &= F_j^k + (x_{k+1} - x_k)^\top g_{F,j}^k + \frac{1}{2}\varrho_{F,j}(x_{k+1} - x_k)^\top G_{F,j}(x_{k+1} - x_k),\ j \in J_k,\\
f_p^{k+1} &= \tilde{f}_p^k + (x_{k+1} - x_k)^\top \tilde{g}_{f,p}^k + \frac{1}{2}(x_{k+1} - x_k)^\top G_{f,p}^{k+1}(x_{k+1} - x_k),\\
F_p^{k+1} &= \tilde{F}_p^k + (x_{k+1} - x_k)^\top \tilde{g}_{F,p}^k + \frac{1}{2}(x_{k+1} - x_k)^\top G_{F,p}^{k+1}(x_{k+1} - x_k),\\
g_{f,j}^{k+1} &= g_{f,j}^k + \varrho_{f,j} G_{f,j}(x_{k+1} - x_k),\ j \in J_k,\\
g_{F,j}^{k+1} &= g_{F,j}^k + \varrho_{F,j} G_{F,j}(x_{k+1} - x_k),\ j \in J_k,\\
g_{f,k+1}^{k+1} &= g_{f,k+1} + \varrho_{f,k+1} G_{f,k+1}(x_{k+1} - y_{k+1}),\\
g_{F,k+1}^{k+1} &= g_{F,k+1} + \varrho_{F,k+1} G_{F,k+1}(x_{k+1} - y_{k+1}),\\
g_{f,p}^{k+1} &= \tilde{g}_{f,p}^{k+1} + G_{f,p}^{k+1}(x_{k+1} - x_k),\\
g_{F,p}^{k+1} &= \tilde{g}_{F,p}^{k+1} + G_{F,p}^{k+1}(x_{k+1} - x_k).
\end{aligned}
$$

S5 Wähle eine Indexmenge J_{k+1}, die in $\{k - M, \ldots, k+1\} \cap \{1, 2, \ldots\}$ enthalten ist und $k+1$ enthält. Erhöhe k um 1 und gehe zu **S1**.

Im obigen Algorithmus findet die Bundle-Reset-Strategie Anwendung, d.h. wenn $i_s > i_r$, dann wird $\lambda_{f,p}^k := \lambda_{F,p}^k := 0$ gesetzt. Anders ausgedrückt: Falls die Zahl der wesentlichen

Schritte seit dem letzten Bundle-Reset größer ist als eine vom Nutzer festgelegte Schranke i_r, dann findet eine Veränderung des quadratischen Teilproblems in S1 dahingehend statt, dass die aggregierte Nebenbedingung bei der Lösung des Problems nicht einbezogen wird. Diese Strategie sowie die erste Maßnahme in Schritt S1 werden für den Beweis der superlinearen Konvergenz benötigt (siehe Satz 7.5). Die zweite "wenn, dann"-Bedingung in Schritt S1 ist notwendig, um die Konvergenz allgemein zu beweisen (siehe Satz 7.4).

7.5 Konvergenzanalyse

In diesem Kapitel formulieren wir Konvergenzaussagen. Die Beweise können für den unrestringierten Fall in Luksan und Vlček (1998) nachgelesen werden.
Der folgende Satz zeigt die allgemeine Konvergenz des Bundle-Newton-Algorithmus auf.

Satz 7.4. *Wir nehmen an, dass $\{x_k\}$ und H_k beschränkt sind. Dann ist jeder Häufungspunkt von $\{x_k\}$ stationär für f.*

Unter bestimmten zusätzlichen Forderungen an die Problemfunktionen erzeugt der Algorithmus Newton-Iterationen und die Konvergenz ist superlinear.

Satz 7.5. *Unter den Voraussetzungen, dass*

(a) die Folge der Hilfspunkte $\{y_k\}$ gegen $\bar{x}$ konvergiert,

(b) die Problemfunktionen f und F streng konvex mit Modulus $C > 0$ sind, d.h. $f(x) - (C/2)\|x\|^2$ und $F(x) - (C/2)\|x\|^2$ konvex sind,

(c) f und F stetige Ableitungen zweiter Ordnung in einer Umgebung $B(\bar{x})$ von $\bar{x}$ besitzen,

(d) die Zahl der wesentlichen Schritte unbegrenzt ist,

(e) $\omega = 1$ gewählt wird und

(f) $G_{f,k}$ und $G_{F,k}$ die Hessematrizen von f und F in y_k darstellen,

erzeugt der Bundle-Newton-Algorithmus nach einer genügend großen Anzahl von Schritten nur noch Newton-Iterationen und $\{x_k\}$ konvergiert superlinear gegen $\bar{x}$.

Teil III

Implementierung und numerische Beispiele

Dieser letzte Teil der Arbeit ist der praktischen Umsetzung der vorangegangenen Theorie gewidmet. Im Rahmen dieser Arbeit wurde die in Teil II hergeleitete Bundle-Newton-Methode implementiert und in den optimalen Steuerungsalgorithmus, der in Teil I vorgestellt wurde, eingesetzt. Ausgewählte Aspekte der Vorgehensweise werden in Kapitel 8 erläutert. In Kapitel 9 stellen wir die Ergebnisse vor, die wir unter Verwendung der Bundle-Newton-Methode erhalten, und vergleichen diese mit den Resultaten aus Jarzcyk (2005). Eine Beurteilung der Ergebnisse findet sich im letzten Kapitel.

Kapitel 8

Implementierung

Das vorliegende Kapitel greift wichtige konkrete Umsetzungen des in Abschnitt 7.4 vorgestellten Algorithmus heraus.
Das Ziel dieser Arbeit besteht darin, das Bundle-Newton-Programm in den Algorithmus zur Lösung von optimalen Steuerungsproblemen, den wir aus Jarczyk (2005) entnommen haben, einzubauen und diesen auf die Testbeispiele aus eben genannter Arbeit anzuwenden. Da bei der Implementierung Besonderheiten dieser optimalen Steuerungsprobleme ausgenutzt werden, möchten wir deren Struktur betrachten.
Die Kontrollsysteme sind durch

$$\dot{x}(t) = f(x(t), u(t))$$

gegeben, wobei $f\colon \mathbb{R}^2 \times U \to \mathbb{R}^2$ ein stetiges Vektorfeld darstellt. Die Zustandsvariable $x(t) = (x_1(t), x_2(t))$ stellt also einen zweidimensionalen Vektor dar. Als Kontrollwertebereich wird ein kompaktes, eindimensionales Intervall $U = [a, b], a, b \in \mathbb{R}$, herangezogen. Die Ertragsfunktion g bildet von $\mathbb{R}^2 \times U$ nach $\mathbb{R}$ ab. Das Maximierungsproblem (4.7) besitzt demgemäß die Form

$$\begin{array}{ll} \max & \tilde{h}(u) \\ \text{unter} & u \in [a, b], \end{array} \tag{8.1}$$

wobei $\tilde{h}\colon U \to \mathbb{R}$ definiert ist wie in (4.7). Es muss allerdings für jeden Gitterpunkt x herausgefunden werden, für welche Werte $u \in U$ die Bedingung $f_h(x, u) \in \Omega$ erfüllt ist (siehe hierzu Abschnitt 4.2). Dazu greifen wir auf die Funktion admit_intervall() aus Jarczyk (2005) zurück, welche folgendermaßen vorgeht: Zuerst werden die ursprünglichen Intervallgrenzen auf Zulässigkeit überprüft. Ist dies der Fall, so kann das Ausgangsintervall als zulässiger Bereich verwendet werden, wobei wir annehmen, dass dieser zusammenhängend ist.
Wenn einer der beiden Eckpunkte nicht zulässig ist, so wird mit Hilfe von Bisektion ein zulässiger zweiter Eckpunkt $c \in [a, b]$ ermittelt und $[c, b]$ bzw. $[a, c]$ zurückgegeben.

Gilt sowohl $f_h(x,a) \notin \Omega$ als auch $f_h(x,b) \notin \Omega$, so wird $[a,b]$ in 20 äquidistante Abschnitte unterteilt. Findet man auf diese Weise einen zulässigen Punkt c, so wird über Bisektion ein Intervall $[d,e]$ ermittelt, so dass $c \in [d,e] \subset [a,b]$ gilt, ansonsten wird abgebrochen.
Sei nun $\tilde{U} = [\tilde{a}, \tilde{b}]$ das so erzeugte zulässige Intervall. Mit $H_1(u) = \tilde{a} - u$ und $H_2(u) = u - \tilde{b}$ definieren wir $H(u) = \max\{H_1(u), H_2(u)\}$, so dass sich unter Vewendung von $h(x) = -\tilde{h}(x)$ das Minimierungsproblem

$$\begin{array}{ll} \min & h(u) \\ \text{unter} & H(u) \leq 0 \end{array} \tag{8.2}$$

ergibt. Es ist zu beachten, dass die konvexe Funktion H im Punkt $u = \frac{1}{2}(\tilde{a} + \tilde{b})$ nicht differenzierbar ist. An dieser Stelle erhalten wir ein Element des Subdifferentials, indem wir die Ableitung in einer Umgebung dieses Punktes, also entweder -1 oder $+1$, heranziehen.
Nun wenden wir uns der Zielfunktion h zu. Aufgrund der Diskretisierung im Raum können an den Kanten der einzelnen Segmente Nichtdifferenzierbarkeitsstellen auftreten. Auch weisen die numerischen Ergebnisse darauf hin, dass die Funktion nicht immer konvex ist. Die Testphase hat ergeben, dass die besten Resultate erzielt werden, wenn ein Subgradient in einem Punkt $u \in \mathbb{R}$ über

$$\frac{f(u + \tilde{h}) - f(u)}{\tilde{h}} \tag{8.3}$$

für ein geeignetes $\tilde{h} > 0$ approximiert wird. Analog sind wir zur Berechnung der Hessematrix in einem Punkt u vorgegangen.

Zur Berechnung einer positiv definiten Näherung einer Matrix wurde ein modifiziertes Choleskyverfahren nach Gill, Murray und Wright (1984) sowie Gill und Murray (1974) programmiert.

Die Problemstellung (7.14) ist äquivalent zu

$$\begin{array}{ll} \min & \frac{1}{2}\lambda^\top Q \lambda + \alpha^\top \lambda \\ \text{u.d.N.} & 1 \geq e^\top \lambda \geq 1, \\ & \lambda \geq 0, \end{array} \tag{8.4}$$

Dabei ist $\lambda = (\lambda^k_{f,j\,j\in J_k}, \lambda^k_{f,p}, \lambda^k_{F,j\,j\in J_k}, \lambda^k_{F,p})$, $\alpha = (\alpha^k_{f,j\,j\in J_k}, \alpha^k_{f,p}, \alpha^k_{F,j\,j\in J_k}, \alpha^k_{F,p})$ und $e = \underbrace{(1,\ldots,1)}_{|J_k|+1-\text{mal}}$.
Aufgrund der Symmetrie der Matrix $Q = ((\bar{G}^k_p)^{-\frac{1}{2}}T)^\top(\bar{G}^k_p)^{-\frac{1}{2}}T = T^\top(\bar{G}^k_p)^{-1}T$ mit $T = (g^k_{f,j\,j\in J_k}, g^k_{f,p}, g^k_{F,j\,j\in J_k}, g^k_{F,p})$ kann dieses quadratische Teilproblem mit Hilfe der Funktion e04ncc

aus der NAG-Bibliothek gelöst werden. Um zu garantieren, dass jede Nebenbedingung nur einmal auftritt und zur Durchführung der Reset-Strategie wird ein Array restrikt verwendet, das für jede Nebenbedingung einen Eintrag besitzt. Der Wert 1 wird gesetzt, wenn die Nebenbedingung bei der Lösung des Problems nicht einbezogen werden soll, ansonsten steht 0 an dieser Stelle.

Für die Invertierung der Matrix $(\bar{G}_p^k)$ führen wir zuerst mittels der NAG-Routine f07adf eine LU-Zerlegung durch und wenden anschließend die NAG-Funktion f07ajf an.

Die Berechnung der Spektralnorm erfordert die Ermittlung der Eigenwerte. Dazu machen wir uns die NAG-Routine f02acc zunutze.

Für ausführliche Beschreibungen der eingebundenen Funktionen sei auf die Dokumentation zur NAG-Bibliothek verwiesen.

Kapitel 9

Testbeispiele

Nun vergleichen wir die Ergebnisse aus Jarczyk (2005) mit den Resultaten, die mit Hilfe der Bundle-Newton-Methode erzeugt wurden. In der erwähnten Diplomarbeit werden die ableitungsfreien Verfahren

- äquidistante Diskretisierung (diskr[I])
- Brent-Verfahren (brent[I])
- rekursive Suche (rekur[I])
- erweitertes Brent-Verfahren (erw-b[I])

zur Lösung von (4.7) benutzt, wobei I die Anzahl der Teilintervalle, in die $\tilde{U}$ bei den entsprechenden Verfahren eingeteilt wird, bezeichnet. Für eine Erklärung dieser Optimierungsmethoden sei auf Jarczyk (2005) und Grüne (2004) verwiesen. Wir führen die folgenden Abkürzungen ein:

Iter: Anzahl der Iterationen des optimalen Steuerungsalgorithmus
Min: globales Minimum der optimalen Wertefunktion
Max: globales Maximum der optimalen Wertefunktion

Der Algorithmus zur Lösung des optimalen Steuerungsalgorithmus, der in Teil I hergeleitet wurde, kann über die folgenden Parameter auf das konkrete Problem angepasst werden:

$a[0]$: x_0-Koordinate der linken unteren Ecke des Rechteckgitters Ω
$a[1]$: x_1-Koordinate der linken unteren Ecke des Rechteckgitters Ω
$b[0]$: x_0-Koordinate der rechten oberen Ecke des Rechteckgitters Ω
$b[1]$: x_1-Koordinate der rechten oberen Ecke des Rechteckgitters Ω
$n[0]$: Anzahl der Unterteilungen des Gitters in x_0- Richtung
$n[1]$: Anzahl der Unterteilungen des Gitters in x_1- Richtung

h: Schrittweite
δ: Diskontrate
eps: Genauigkeitsschranke für Steuerungsalgorithmus
U: Bereich, über den in (4.7) maximiert wird

Bei der Bundle-Newton-Methode steuern wir die Qualität der Ergebnisse über die spezielle Wahl von $\gamma_f, \gamma_F, \omega, \tilde{h}, \varepsilon$ und S. Mit Hilfe von γ_f, γ_F und ω können wir die in Unterabschnitt 7.1.2 definierten Gewichte beeinflussen. Den Startpunkt der Methode bestimmen wir gemäß $S = \tilde{a} + \frac{(\tilde{b}-\tilde{a})}{q}$ mit $q \in [1, \infty)$. Der Parameter h wird zur Berechnung einer Approximation des Subdifferentials (siehe (8.3)) herangezogen. Darüber hinaus stellt ε die in der Bundle-Newton-Methode verwendete Genauigkeitsschranke dar. Wir benutzen im Folgenden die kompakte Schreibweise bundle-newton[γ_f, γ_F, ω, $\tilde{h}$, ε, q]. Wenn wir die Anzahl der Iterationen künstlich erhöhen, geben wir diese mit einem Index (anzahl) an, beispielsweise bundle-newton$[\gamma_f, \gamma_F, \omega, \tilde{h}, \varepsilon, q]_{anzahl}$.

Die folgende Tabelle gibt einen Überblick über die Belegung weiterer Parameter im Bundle-Newton-Algorithmus:

m_L	m_R	t_0	ϑ	ζ	C_S	C_G	i_m	i_r
0.01	0.5	0.001	1.0	0.01	10^{50}	10^{50}	100	100

Da der Zweck und der Einfluss der einzelnen Parameter im Zusammenhang mit der Herleitung der Bundle-Newton-Methode einsichtiger ist, verweisen wir für ausführlichere Informationen auf Kapitel 7.

Beim Vergleich der Methoden legen wir im Folgenden den Schwerpunkt auf die Genauigkeit der Ergebnisse. Der Aufwand der Bundle-Newton-Methode ist bei allen Beispielen um ein Vielfaches größer als der der übrigen Verfahren. Diesen werden wir für das erste Modell über die Laufzeit abschätzen.

9.1 Beispiel 1: Einfaches Modell

Zuerst wenden wir uns einem einfachen optimalen Steuerungsproblem zu, das durch das Kontrollsystem

$$\dot{x}_1(t) = x_1(t) + u(t)x_2(t)$$
$$\dot{x}_2(t) = -x_2(t)$$

sowie durch die Ertragsfunktion

$$g(x_1, x_2, u) = -1.0 + x_1 - 0.3u$$

gegeben ist. Mit den Parametern

$a[0]$	$a[1]$	$b[0]$	$b[1]$	$n[0]$	$n[1]$	h	δ	ε	U
0.0	0.0	1.0	1.0	50	50	0.05	0.5	10^{-4}	[0,1]

erhalten wir:

Verfahren	**Iter**	**Min**	**Max**
bundle-newton$[0.5, 0.5, 1.0, 10^{-6}, 10^{-5}, 1.0]$	247	-1.9978	-1.2040
bundle-newton$[0.5, 0.5, 1.0, 10^{-6}, 10^{-5}, 1.0]_{600}$	600	-2.0001	-1.2050
brent[3]	247	-1.9977	-1.2040
brent[6]	247	-1.9977	-1.2040
brent$[3]_{600}$	600	-2.0000	-1.2049
diskr[10]	247	-1.9977	-1.2040
rekur[6]	247	-1.9977	-1.2040
rekur[15]	247	-1.9977	-1.2040
erw-brent[3]	247	-1.9977	-1.2040
erw-brent[6]	247	-1.9977	-1.2040

Die neue Optimierungsstrategie liefert sowohl nach 247 als auch nach 600 Iterationen etwas genauere Ergebnisse als die in Jarczyk (2005) verwendeten Verfahren. Die Abbildungen (9.1) und (9.2) zeigen die optimale Wertefunktion sowie die optimale Steuerung für bundle-newton$[0.5, 0.5, 1.0, 10^{-6}, 10^{-5}, 1.0]$. Der Sprung in der optimalen Steuerung hat einen Knick in der optimalen Wertefunktion zur Folge, an dem diese nicht differenzierbar ist. Die Abbildungen (9.3) und (9.4) verdeutlichen den Unterschied zwischen brent[3] und bundle-newton$[0.5, 0.5, 1.0, 10^{-6}, 10^{-5}, 1.0]$ (bundle-newton[ ]-brent[ ]). Die Differenz zwischen den optimalen Steuerungen ist in der Nähe der Sprungstellen besonders hoch.

Für dieses Modell möchten wir den Aufwand grob über die Laufzeit abschätzen.
Das Verfahren brent[3] führt die notwendigen Berechnungen 75-mal schneller durch als bundle-newton$[0.5, 0.5, 1.0, 10^{-6}, 10^{-5}, 1.0]$. Auch im Vergleich zu brent[10] müssen wir bei bundle-newton$[0.5, 0.5, 1.0, 10^{-6}, 10^{-5}, 1.0]$ 34-mal länger auf die Ergebnisse warten. Selbst rekur[15] benötigt nur ungefähr 15% der Zeit von bundle-newton$[0.5, 0.5, 1.0, 10^{-6}, 10^{-5}, 1.0]$.

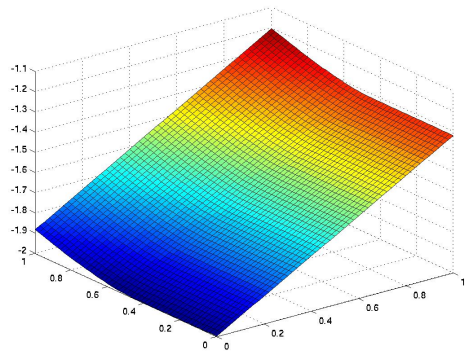

Abbildung 9.1: Wertefunktion Bsp1

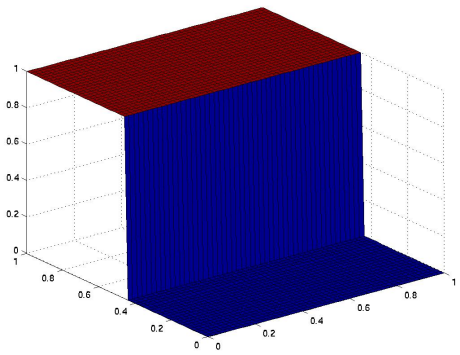

Abbildung 9.2: Steuerung Bsp1

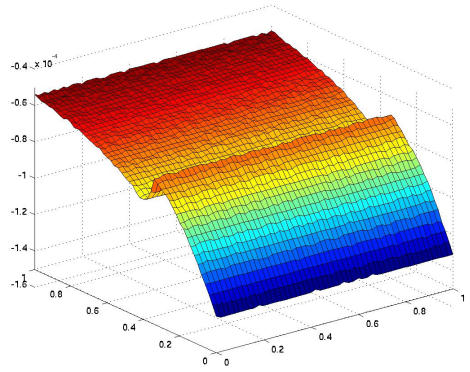

Abbildung 9.3: Wertefunktion$_{Differenz}$ Bsp1

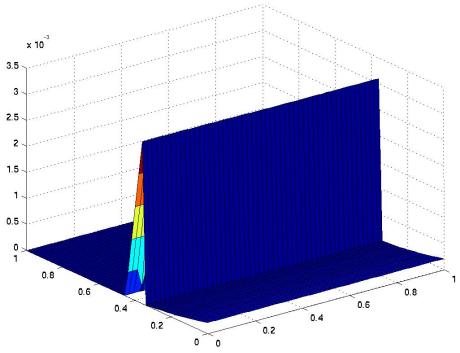

Abbildung 9.4: Steuerung$_{Differenz}$ Bsp1

9.2 Beispiel 2: Investitionsmodell

Das Kontrollsystem

$$\begin{aligned} \dot{x}_1(t) &= x_2(t) - \sigma x_1(t) \\ \dot{x}_2(t) &= u(t) \end{aligned} \tag{9.1}$$

und die Ertragsfunktion

$$g(x_1, x_2, u) = k_1\sqrt{x_1} - \frac{x_1}{1 + k_2 x_1^4} - c_1 x_2 - \frac{c_2}{2}x_2^2 - \frac{\alpha}{2}u^2 \tag{9.2}$$

charakterisieren ein optimales Steuerungsproblem. Dabei modelliert das System (9.1) Kapitalströme, (9.2) stellt den diskontierten cash flow dar.
Legen wir die Parameter

$a[0]$	$a[1]$	$b[0]$	$b[1]$	$n[0]$	$n[1]$	h	δ	ε	U
0.0	0.0	6.0	1.5	50	50	0.05	0.04	10^{-3}	[-1, 1]

zugrunde, ergibt sich:

Verfahren	**Iter**	**Min**	**Max**
bundle-newton$[10^{-4}, 10^{-4}, 1.0, 10^{-6}, 10^{-4}, 3.0]$	1857	12.957	30.552
bundle-newton$[0.5, 0.5, 1.0, 10^{-6}, 10^{-4}, 2.5]$	1867	12.980	30.560
bundle-newton$[0.5, 0.5, 1.0, 10^{-6}, 10^{-4}, 1.5]$	1881	12.986	30.577
bundle-newton$[0.5, 0.5, 1.0, 10^{-6}, 10^{-4}, 2.5]_{3600}$	3600	13.391	30.994
brent[3]	1836	12.962	30.551
brent[6]	1837	12.961	30.551
brent$[6]_{3600}$	3600	13.393	31.008
diskr[10]	1847	12.835	30.470
diskr[100]	1834	12.958	30.546
rekur[6]	1834	12.960	30.550
erw-b[3]	1838	12.958	30.559
erw-b[6]	1834	12.959	30.548

Während die Ergebnisse von bundle-newton$[10^{-4}, 10^{-4}, 1.0, 10^{-6}, 10^{-4}, 3.0]$ noch hinter denen von brent[3] zurückbleiben, kann mit bundle-newton$[0.5, 0.5, 1.0, 10^{-6}, 10^{-4}, 2.5]$ und bundle-newton$[0.5, 0.5, 1.0, 10^{-6}, 10^{-4}, 1.5]$ eine Verbesserung erzielt werden. Da aber brent[6] nach 3600 Iterationen die Resultate von bundle-newton$[0.5, 0.5, 1.0, 10^{-6}, 10^{-4}, 2.5]_{3600}$ übertrifft, muss davon ausgegangen werden, dass die soeben erwähnte Verbesserung nur auf die höhere

Iterationenzahl zurückzuführen ist. Die optimale Wertefunktion sowie die optimale Steuerung sind für bundle-newton$[0.5, 0.5, 1.0, 10^{-6}, 10^{-4}, 2.5]$ in den Abbildungen (9.5) und (9.6) visualisiert worden. Wiederum korrespondiert der Sprung in der optimalen Steuerung mit einem Knick in der optimalen Wertefunktion. Diese Linie trennt die Einzugsbereiche der zwei stabilen optimalen Gleichgewichte und wird Skiba-Linie genannt. Erstaunlich ist die enorme

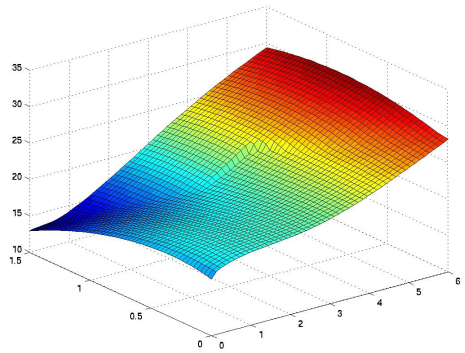

Abbildung 9.5: Wertefunktion Bsp2

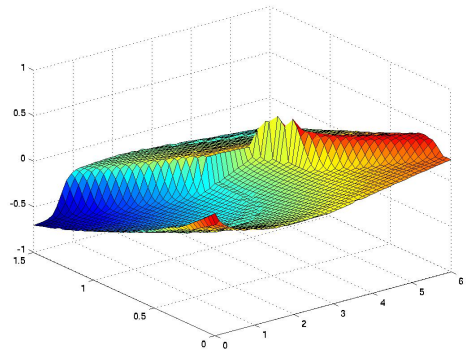

Abbildung 9.6: Steuerung Bsp2

Veränderung des globalen Maximums und Minimums, welche mittels einer Erhöhung auf 3600 Iterationen bewirkt wird.
Die Differenz zwischen bundle-newton$[0.5, 0.5, 1.0, 10^{-6}, 10^{-4}, 2.5]$ und brent[3] sowohl in Bezug auf die optimale Wertefunktion als auch bezüglich der optimalen Steuerung ist in den Abbildungen (9.7) und (9.8) zu sehen (bundle-newton[ ]-brent[ ]). Diese ist in der Nähe der Skibalinie besonders hoch.

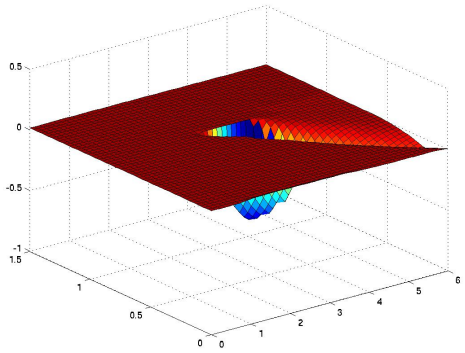

Abbildung 9.7: Wertefunktion$_{Differenz}$ Bsp2

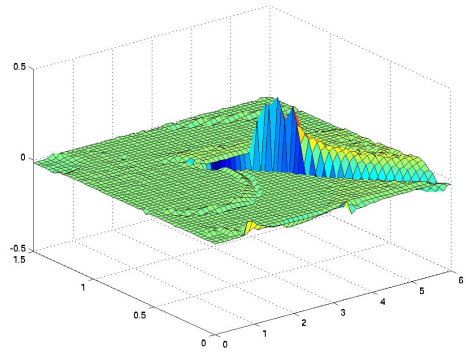

Abbildung 9.8: Steuerung$_{Differenz}$ Bsp2

9.3 Beispiel 3: Makroökonomisches Modell

Das makroökonische Modell basiert auf dem Kontrollsystem

$$\dot{x}_1(t) = u - 0.55x_1 + \frac{x_1^2}{1 + x_1^2}$$
$$\dot{x}_2(t) = 0$$

sowie der Ertragsfunktion

$$g(x_1, x_2, u) = 2\sqrt{u} - \frac{\lambda}{2}x_1^2,\ \lambda \geq 0.$$

Dieses kontinuierliche optimale Steuerungsproblem wurde von W. Brock und D. Starret entwickelt und kann folgendermaßen interpretiert werden:
Ein Unternehmen produziert in der Nähe eines Sees und leitet das dadurch entstandene phosphathaltige Abwasser in dieses Gewässer. Die Variable x_1 stellt den Phosphatgehalt des Sees dar und die Gleichung

$$\dot{x_1}(t) = -0.55x_1 + \frac{x_1^2}{1 + x_1^2}$$

beschreibt, wie schnell das Phosphat abgebaut werden kann.
Die Steuerung u gibt an, wieviel das Werk produziert. Aus Vereinfachungsgründen wird sie so normiert, dass u der Menge der Phosphateinleitung entspricht. Das Ziel besteht nun darin, den Gewinn durch die Produktion zu maximieren, ohne den See zu sehr zu verschmutzen. In die Ertragsfunktion geht die Verunreinigung des Sees somit negativ ein, während die Produktion den Gewinn erhöht. Da das vorliegende Beispiel eigentlich als eindimensionales Modell konstruiert wurde, setzen wir die Variable $x_2 = 0$.
Die Bundle-Newton-Methode setzt voraus, dass die auftretenden Funktionen auf ganz $\mathbb{R}$ definiert sind, da es möglich ist, dass der Algorithmus Hilfspunkte y_k erzeugt, die nicht zulässig sind, für die aber eine Berechnung eines Subgradienten oder einer Näherung der Hessematrix erforderlich ist. Da obige Bedingung für die Wurzelfunktion nicht vorliegt, hat es sich beim makroökonomischen Modell als sinnvoll erwiesen, u durch $\tilde{u} = e^u$ zu substituieren und die Ergebnisse anschließend zurückzutransformieren.
Mit Hilfe der Parameter

$a[0]$	$a[1]$	$b[0]$	$b[1]$	$n[0]$	$n[1]$	h	δ	ε	U
0.0	0.0	3.0	1.0	50	50	0.05	0.1	10^{-3}	[0,1]

erhalten wir die nachfolgenden Resultate:

Verfahren	**Iter**	**Min**	**Max**
bundle-newton$[1.0, 1.0, 1.0, 10^{-6}, 10^{-4}, 1]$	557	-6.472	5.878
bundle-newton$[0.5, 0.5, 1.0, 10^{-6}, 10^{-5}, 1.01]$	575	-6.483	5.895
bundle-newton$[10^{-5}, 10^{-3}, 1.0, 10^{-6}, 10^{-4}, 1.0]$	569	-6.491	5.879
bundle-newton$[10^{-5}, 10^{-5}, 1.0, 10^{-6}, 10^{-4}, 1.7]$	579	-6.497	5.887
bundle-newton$[10^{-3}, 10^{-5}, 1.0, 10^{-6}, 10^{-4}, 1.0]$	976	-6.579	6.041
bundle-newton$[0.5, 0.5, 1.0, 10^{-6}, 10^{-4}, 1.0]_{1200}$	1200	-6.525	6.060
brent[6]	638	-6.504	5.909
brent[10]	546	-6.458	5.871
brent[25]	546	-6.458	5.871
diskr[10]	491	-6.427	5.679
diskr[20]	534	-6.460	5.823
diskr[100]	544	-6.457	5.863
rekur[6]	546	-6.458	5.871
erw-b[3]	642	-6.506	5.910
erw-b$[3]_{1200}$	1200	-6.572	6.017
erw-b[6]	543	-6.456	5.867
erw-b[25]	546	-6.458	5.871

Eine Verbesserung gegenüber erw-b[3] kann erst bei einer Parameterkonstellation, die zu 976 Iterationen führt, erzielt werden. Es fällt auf, dass das durch bundle-newton$[10^{-3}, 10^{-5}, 1.0, 10^{-6}, 10^{-4}, 1]$ erzeugte globale Minimum nach 976 Iterationen kleiner ist als der entsprechende Wert bei bundle-newton$[0.5, 0.5, 1.0, 10^{-6}, 10^{-4}, 1.0]$ nach 1200 Iterationen und dem von erw-b$[3]_{1200}$ errechneten Minimum fast entspricht. Auch das approximierte globale Maximum liegt bei bundle-newton$[10^{-3}, 10^{-5}, 1.0, 10^{-6}, 10^{-4}, 1.0]$ näher bei dem durch erw-b$[3]_{1200}$ ermittelten Wert als bei bundle-newton$[0.5, 0.5, 1.0, 10^{-6}, 10^{-4}, 1.0]$.
Die zu bundle-newton$[10^{-3}, 10^{-5}, 1.0, 10^{-6}, 10^{-4}, 1.0]$ gehörigen Plots sind in den Abbildungen (9.9) und (9.10) zu finden.
Die Differenz zwischen erw-b[3] und bundle-newton$[10^{-3}, 10^{-5}, 1.0, 10^{-6}, 10^{-4}, 1.0]$ kann für die optimale Wertefunktion in Abbildung (9.11), für die optimale Steuerung in Abbildung (9.12) betrachtet werden (bundle-newton[ ]-erw-b[ ]). Der Unterschied ist im Bereich des Knicks der optimalen Wertefunktion erheblich. Die Abbildungen (9.13) und (9.14) zeigen, dass die Differenz zwischen brent[6] und bundle-newton$[10^{-3}, 10^{-5}, 1.0, 10^{-6}, 10^{-4}, 1.0]$ etwas geringer ausfällt.

Über den Parameter λ ist es möglich, den Einfluss der Verschmutzung in der Ertragsfunktion zu gewichten. Um die Ergebnisse für alle λ aus $[0, 1]$ zu erhalten, setzen wir $x_2 = \lambda$.

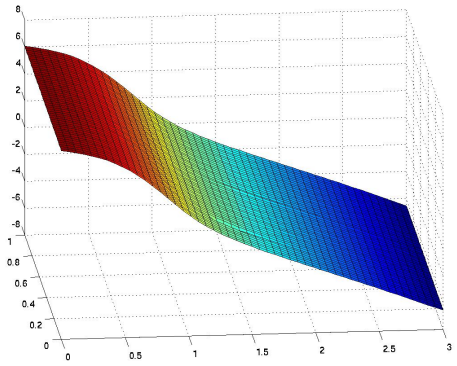

Abbildung 9.9: Wertefunktion Bsp3a

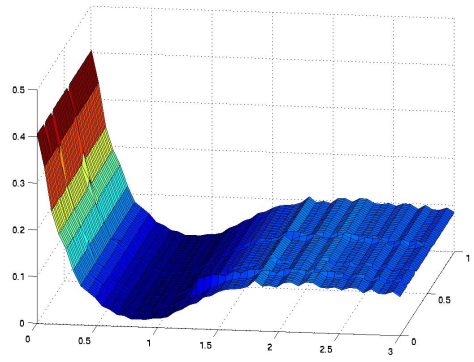

Abbildung 9.10: Steuerung Bsp3a

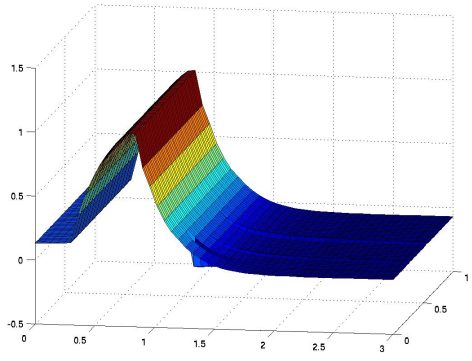

Abbildung 9.11: Wertefunktion$_{Differenz1}$ Bsp3a

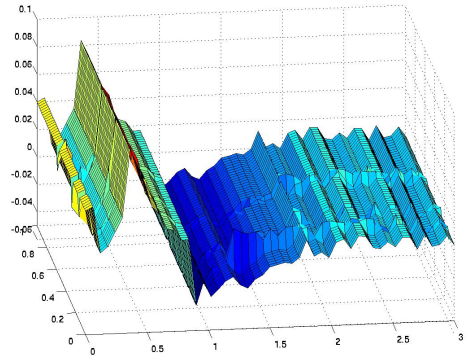

Abbildung 9.12: Steuerung$_{Differenz1}$ Bsp3a

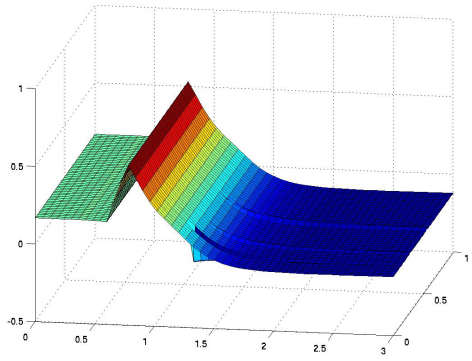

Abbildung 9.13: Wertefunktion$_{Differenz2}$ Bsp3a

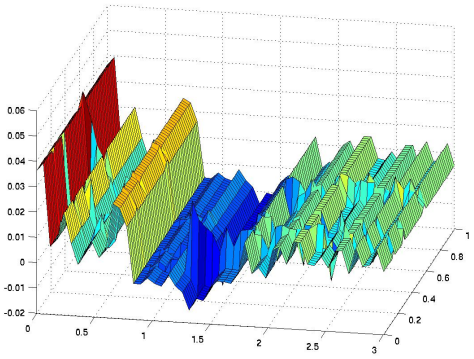

Abbildung 9.14: Steuerung$_{Differenz2}$ Bsp3a

Die Ertragsfunktion nimmt deswegen die Form

$$g(x_1, x_2, u) = 2\sqrt{u} - \frac{x_2}{2}x_1^2$$

an. Eine Übersicht über die Ergebnisse findet sich in der folgenden Aufstellung:

Verfahren	**Iter**	**Min**	**Max**
bundle-newton$[10^{-5}, 10^{-5}, 1.0, 10^{-6}, 10^{-4}, 1.0]$	874	-9.284	18.076
bundle-newton$[0.5, 0.5, 1.0, 10^{-6}, 10^{-4}, 1.0]$	874	-9.265	18.076
bundle-newton$[0.5, 0.5, 1.0, 10^{-6.0}, 10^{-4}, 1.0]_{1500}$	1500	-9.234	18.266
brent[6]	874	-9.378	18.077
brent[10]	874	-9.297	18.077
brent[25]	874	-9.257	18.077
diskr[10]	858	-9.787	18.019
diskr[13]	856	-9.589	18.010
diskr[20]	874	-9.452	18.077
diskr[100]	874	-9.262	18.077
rekur[9]	874	-9.257	18.077
rekur$[9]_{1500}$	1500	-9.225	18.266
erw-b[6]	874	-9.378	18.077
erw-b[10]	874	-9.298	18.077
erw-b[25]	874	-9.257	18.077

Es ist zu erkennen, dass die Maßnahme $x_2 = \lambda$ zu einem größeren globalen Maximum bzw. zu einem kleineren globalen Minimum führt. Dies ist darauf zurückzuführen, dass nicht nur $\lambda = 0.8$, sondern alle $\lambda \in [0, 1]$ berücksichtigt werden. Die optimale Wertefunktion setzt sich aus den zu dem jeweiligen $\lambda \in [0, 1]$ gehörigen Linien zusammen.
Es wurde keine Parameterkonstellation gefunden, die zu einer Verbesserung der durch rekur[9] erzeugten Resultate führt. Nach 1500 Iterationen liefert rekur[9] einen größeren Wert für das globale Minimum als bundle-newton$[0.5, 0.5, 1.0, 10^{-6.0}, 10^{-4}, 1.0]$. Die approximierten globalen Maxima stimmen überein.
Die optimale Wertefunktion für bundle-newton$[0.5, 0.5, 1.0, 10^{-6}, 10^{-4}, 1.0]$ wird in Abbildung (9.15), die optimale Steuerung in Abbildung (9.16) dargestellt. Die Differenz zwischen bundle-newton$[0.5, 0.5, 1.0, 10^{-6}, 10^{-4}, 1.0]$ und rekur[9] wird in den Abbildungen (9.17) und (9.18) veranschaulicht (bundle-newton[ ]-rekur[ ]). Wiederum ist der Unterschied an der Sprungstelle besonders hoch.

Insgesamt lässt das Verhalten der Bundle-Newton-Methode vermuten, dass die Zielfunktionen, die in diesem Beispiel zugrunde liegen, nicht immer die erforderlichen Voraussetzungen aufweisen.

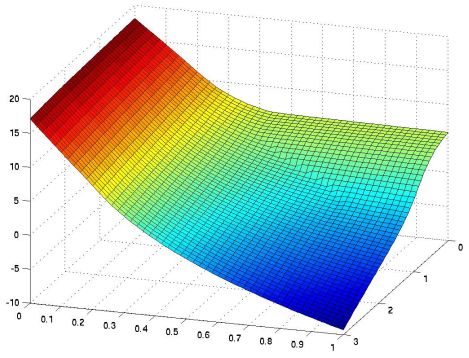

Abbildung 9.15: Wertefunktion Bsp3b

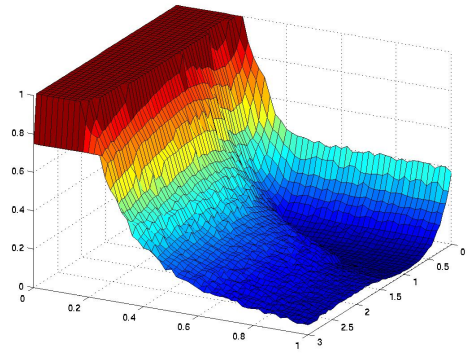

Abbildung 9.16: Steuerung Bsp3b

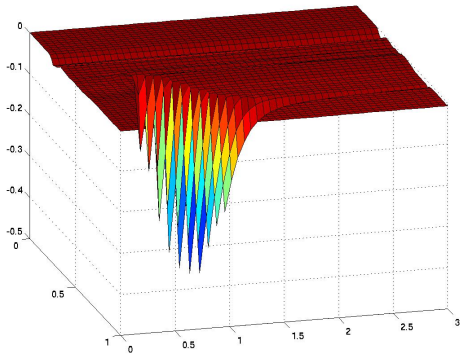

Abbildung 9.17: Wertefunktion$_{Differenz}$ Bsp3b

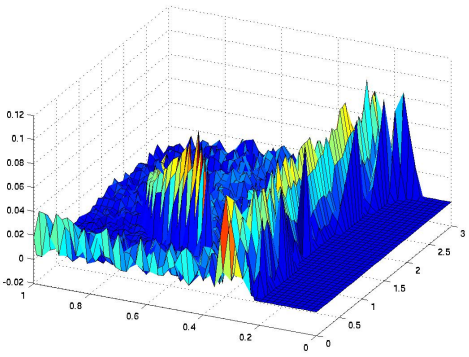

Abbildung 9.18: Steuerung$_{Differenz}$ Bsp3b

9.4 Beispiel 4: Ökonomisches Wachstumsmodell

Nun analysieren wir das ökonomische Wachstumsmodell, ein diskretes optimales Steuerungsproblem, für das eine exakte Lösung bekannt ist. Diesem Modell liegt das Kontrollsystem

$$x_1(t+1) = Ae^{x_2(t)}(x_1(t))^\alpha - u(t)$$
$$x_2(t+1) = \rho x_2(t)$$

sowie die Ertragsfunktion

$$g(x_1, x_2, u) = \ln u$$

mit $\rho, \alpha, A \in \mathbb{R}^+$ zugrunde. Wir wählen $\rho = 0.9, \alpha = 0.34, A = 5$ und $\beta = 1 - \delta h = 0.95$. Die weiteren benötigten Parameter entnehmen wir der nachstehenden Tabelle:

$a[0]$	$a[1]$	$b[0]$	$b[1]$	$n[0]$	$n[1]$	h	δ	ε	U
0.1	-0.4	10.0	0.4	50	50	1.0	0.05	10^{-5}	[0.1, 13.5]

Die exakte optimale Wertefunktion ist durch

$$V(x) = B + C \ln x_1 + Dx_2$$

mit

$$B = \frac{\ln((1-\beta\alpha)A) + \frac{\beta\alpha}{1-\beta\alpha}\ln(\beta\alpha A)}{1-\beta},\ C = \frac{\alpha}{1-\alpha\beta},\ D = \frac{1}{(1-\alpha\beta)(1-\rho\beta)},$$

die optimale Steuerung durch

$$u_{opt}(x) = (1-\alpha\beta)Ae^{x_2}x_1^\alpha$$

gegeben.
Die Anwendung der Bundle-Newton-Methode führt bei diesem Modell zu Schwierigkeiten. Sowohl die Bundle-Newton-Methode innerhalb der Iterationen als auch die Iterationen des optimalen Steuerungsalgorithmus müssen künstlich abgebrochen werden. Des Weiteren erweist sich für dieses Modell eine modifizierte Liniensuche als geeigneter:

S0 Setze $t_L := 0$ und $t := t_U := 1$. Wähle $\zeta \in (0, \frac{1}{2}), \vartheta \geq 1$.

S1 Wenn $f(x_k + td_k) \leq f(x_k) + m_L t v_k$, so setze $t_L := t$, ansonsten setze $t_U := t$.

S2 Wenn $t_L \geq t_0$, setze $t_R := t_L$ und STOPP.

S3 Ermittle einen Subgradienten $g_f \in \partial f(x_k + td_k)$,
eine symmetrische Matrix G und

$$\begin{aligned} \varrho &:= \begin{cases} \min[1, C_G/\|G_f\|_S], & \text{wenn } i_n \leq 3, \\ 0 & \text{sonst,} \end{cases} \\ f &:= f(x_k + td_k) + (t_L - t)g^\top d_k + \frac{1}{2}\varrho(t_L - t)^2 d_k^\top, \\ \beta &:= \max[|f - f(x_k + t_L d_k)|, \gamma_f |t_L - t|^\omega \|d_k\|^\omega]. \end{aligned}$$

S4 Wenn $-\beta + d_k^\top(g + \varrho(t_L - t)Gd_k) \geq m_R v_k$ und $(t - t_L)\|d_k\| \leq C_S$, dann setze $t_R := t$ und STOPP.

S5 Wähle $t \in [t_L + \zeta(t_U - t_L)^\vartheta, t_U - \zeta(t_U - t_L)^\vartheta]$ mit Hilfe einer Interpolationsmethode und gehe zu **S1**.

Eine weitere Modifikation des Bundle-Newton-Algorithmus ist von Vorteil. Für die Ermittlung der Größen

$$\begin{aligned} \tilde{v} &= -\|H_k \tilde{g}_p^k\|^2 - \tilde{\alpha}_p^k \\ \tilde{w} &= \frac{1}{2}\|H_k, \tilde{g}_p^k\|^2 + \tilde{\alpha}_p^k \end{aligned}$$

mit $H_k = (\bar{G}_p^k)^{-\frac{1}{2}}$ in Schritt S1 des Bundle-Newton-Algorithmus in Abschnitt 7.4 reicht es aus, die Inverse von $\bar{G}_p^k$ zu ermitteln. Allerdings hat es sich beim ökonomischen Wachstumsmodell gezeigt, dass die explizite Berechnung der inversen Matrixwurzel zu besseren Ergebnissen führt. Für die Bildung der Wurzel von $\bar{G}_p^k$ machen wir uns zunutze, dass $\bar{G}_p^k$ eine reelle positiv definite Matrix und somit diagonalisierbar ist. Wir können also eine invertierbare Matrix P und eine Diagonalmatrix D mit

$$P^{-1}\bar{G}_p^k P = D = \begin{pmatrix} \kappa_1 & & 0 \\ & \ddots & \\ 0 & & \kappa_n \end{pmatrix}$$

finden, wobei die Diagonalelemente $\kappa_i, i = 0, \ldots, n$ von D die Eigenwerte von $\bar{G}_p^k$ und die Spalten der Matrix P die zugehörigen Eigenvektoren darstellen. Die Wurzel von $\bar{G}_p^k$ lässt sich nun folgendermaßen bestimmen:

$$\begin{aligned} & \bar{G}_p^k = (\bar{G}_p^k)^{\frac{1}{2}}(\bar{G}_p^k)^{\frac{1}{2}} = PDP^{-1} = PD^{\frac{1}{2}}D^{\frac{1}{2}}P^{-1} = (PD^{\frac{1}{2}}P^{-1})(PD^{\frac{1}{2}}P^{-1}) \\ \Longrightarrow\ & (\bar{G}_p^k)^{\frac{1}{2}} = PD^{\frac{1}{2}}P^{-1} \end{aligned}$$

Für die Ermittlung der Eigenwerte und Eigenvektoren benutzen wir wiederum die NAG-Funktion f02acc. Da es sich bei D um eine Diagonalmatrix handelt, ist die Invertierung

leicht durchzuführen.
Mit Hilfe dieser speziellen Maßnahmen ergeben sich die folgenden Werte:

Verfahren	**Iter**	**Min**	**Max**	**Fehler**
exakt		23.7298	34.1921	0.00
bundle-newton$[0.5, 0.5, 1.0, 10^{-7}, 10^{-6}, 1.5]_{160}$	160	23.72402	34.18971	0.0033
bundle-newton$[0.5, 0.5, 1.0, 10^{-7}, 10^{-6}, 1.5]_{300}$	300	23.72397	34.18973	0.0032
brent[3]	153	23.72397	34.18973	0.0034
brent[6]	156	23.72398	34.18973	0.0034
brent[30]	149	23.72397	34.18972	0.0034
diskr[10]	186	21.55441	34.01367	0.2649
diskr[50]	146	23.67337	34.18219	0.0156
diskr[100]	134	23.71483	34.18543	0.0102
rekur[6]	153	23.72397	34.18973	0.0034
erw-b[3]	157	23.72398	34.18973	0.0034
erw-b[6]	158	23.72398	34.18973	0.0034
erw-b[30]	152	23.72399	34.18973	0.0033

Die Bundle-Newton-Methode liefert nach 160 Iterationen im Vergleich zu erw-b[30] einen etwas genaueren Wert für das globale Minimum, allerdings einen geringfügig schlechteren Wert für das globale Maximum. Auch eine Erhöhung auf 300 Iterationen bringt keine wesentliche Veränderung.
Für bundle-newton$[0.5, 0.5, 1.0, 10^{-7}, 10^{-6}, 1.5]_{160}$ sind die optimale Wertefunktion sowie die optimale Steuerung in (9.19) und (9.20) abgebildet, die Differenz zwischen erw-b[30] und bundle-newton$[0.5, 0.5, 1.0, 10^{-7}, 10^{-6}, 1.5]_{160}$ ist in (9.21) und (9.22) zu sehen.

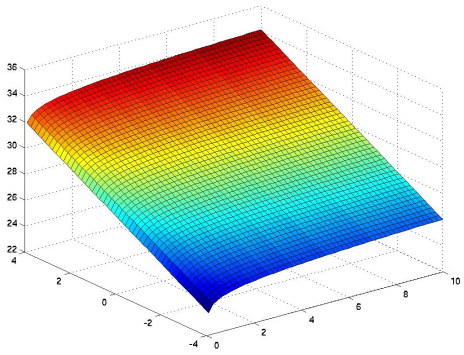

Abbildung 9.19: Wertefunktion Bsp4

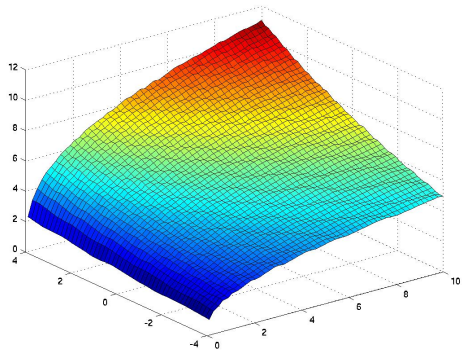

Abbildung 9.20: Steuerung Bsp4

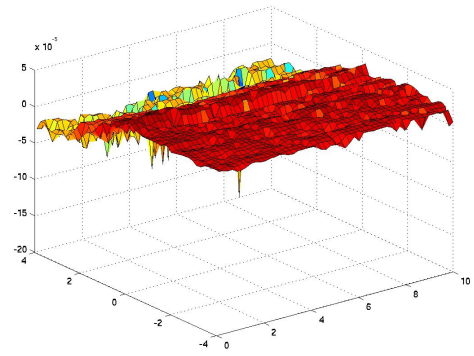

Abbildung 9.21: Wertefunktion$_{Differenz1}$ Bsp4

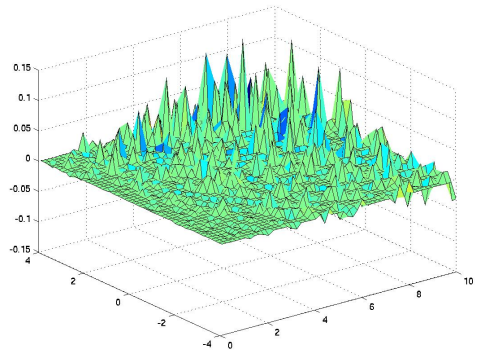

Abbildung 9.22: Steuerung$_{Differenz1}$ Bsp4

Da es bei diesem Beispiel möglich ist, die exakte Lösung anzugeben, bietet es sich an, die durch bundle-newton$[0.5, 0.5, 1.0, 10^{-7}, 10^{-6}, 1.5]_{160}$ erzeugten Ergebnisse mit dieser zu vergleichen. Eine Darstellung der Differenz findet sich in den Abbildungen (9.23) und (9.24).

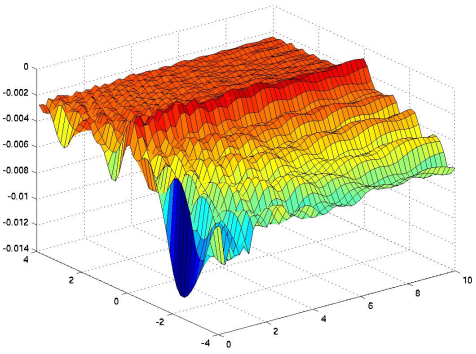

Abbildung 9.23: Wertefunktion$_{Differenz2}$ Bsp4

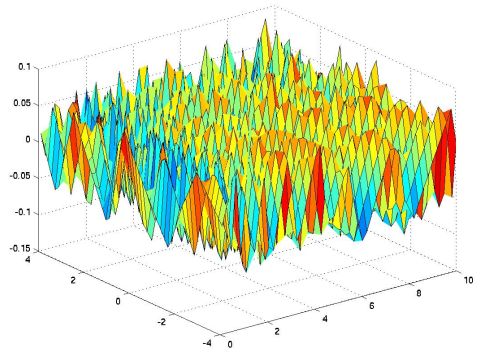

Abbildung 9.24: Steuerung$_{Differenz2}$ Bsp4

Außerdem kann der Fehler über die Formel

$$Fehler = \frac{\sum_{i=0}^{N} |exakt(i) - Verfahren(i)|}{N}$$

berechnet werden werden. Aus obiger Aufstellung wird ersichtlich, dass das Verfahren bundle-newton$[0.5, 0.5, 1.0, 10^{-7}, 10^{-6}, 1.5]_{160}$ gut abschneidet und den selben Fehler wie erw-b[30] aufweist. Bei einer Erhöhung auf 300 Iterationen sinkt der Fehler auf 0.0032.

Die Schwierigkeiten, die die sich aus der Anwendung der Bundle-Newton-Methode ergeben, legen die Vermutung nahe, dass die zu minimierenden Zielfunktionen nicht die für eine erfolgreiche Benutzung erforderlichen Eigenschaften besitzen.

9.5 Beispiel 5: Räuber-Beute-Modell

Als letztes untersuchen wir das kontinuierliche optimale Steuerungsproblem, das durch das Kontrollsystem

$$\dot{x}_1 = (a_0 - a_2 x_2 - a_1 x_1 - u)x_1$$
$$\dot{x}_2 = (b_1 x_1 - b_0 - b_2 x_2 - u)x_2$$

und die Ertragsfunktion

$$g(x_1, x_2, u) = \frac{1}{1 + x_1 u} x_1 u + \frac{1}{1 + x_2 u} x_2 u - \frac{u}{2}$$

gegeben ist. Dieses Beispiel lässt sich folgendermaßen interpretieren: x_1 kann als eine Population von Beutefischen und x_2 als eine Population von Räuberfischen angesehen werden. Das Ziel des optimalen Steuerungsproblems besteht darin, die Fangrate u unter Berücksichtigung des gegenseitigen Einflusses der beiden Fischarten so zu gestalten, dass der Gewinn maximiert wird.

Wir setzen $a_0 = 0.001, a_2 = 0.07, b_0 = 1.01, b_1 = 0.2, b_2 = 0.01$ und wählen außerdem:

$a[0]$	$a[1]$	$b[0]$	$b[1]$	$n[0]$	$n[1]$	h	δ	ε	U
0.0	0.0	20	40	50	50	0.05	5	10^{-5}	[0, 3]

Die Ergebnisse sind in der folgenden Übersicht aufgelistet:

Verfahren	**Iter**	**Min**	**Max**
bundle-newton$[10^{-7}, 10^{-3}, 1.0, 10^{-6}, 10^{-6}, 2.0]$	30	00.000	31.533
bundle-newton$[10^{-7}, 10^{-3}, 1.0, 10^{-6}, 10^{-6}, 2.0]_{60}$	60	00.000	31.533
brent[3]	30	00.000	31.528
brent[6]	30	00.000	31.532
brent[10]	30	00.000	31.533
brent[20]	30	00.000	31.533
diskr[10]	30	00.000	31.296
diskr[100]	30	00.000	31.531
diskr[1000]	30	00.000	31.533
rekur[6]	30	00.000	31.533
erw-b[3]	30	00.000	31.528
erw-b[6]	30	00.000	31.532
erw-b[10]	30	00.000	31.533
erw-b[20]	30	00.000	31.533

Die Bundle-Newton-Methode liefert die gleichen Ergebnisse wie brent[10], brent[20], diskr[1000], erw-b[10] und erw-b[20]. Eine Erhöhung der Iterationenzahl auf 60 bringt keine weitere Verbesserung, so dass wir annehmen können, dass die erhaltenen Resultate bereits eine sehr gute Näherung des globalen Maximums bzw. Minimums darstellen. Die optimale Wertefunktion und die optimale Steuerung von bundle-newton$[10^{-7}, 10^{-3}, 1.0, 10^{-6}, 10^{-6}, 2.0]$ sind in den Abbildungen (9.25) und (9.26) dargestellt.

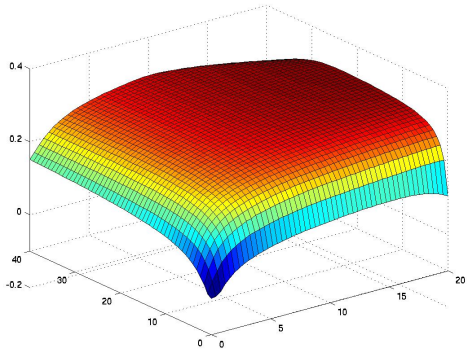

Abbildung 9.25: Wertefunktion Bsp5

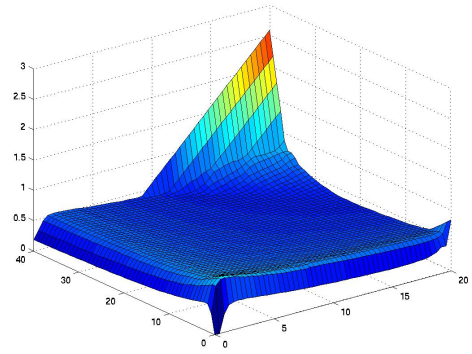

Abbildung 9.26: Steuerung Bsp5

Aus den Abbildungen (9.27) und (9.28) geht hervor, dass bundle-newton$[10^{-4}, 10^{-4}, 1.0, 10^{-6.0}, 10^{-7}, 1.5]$ und brent[10] annähernd die gleichen Ergebnisse liefern.

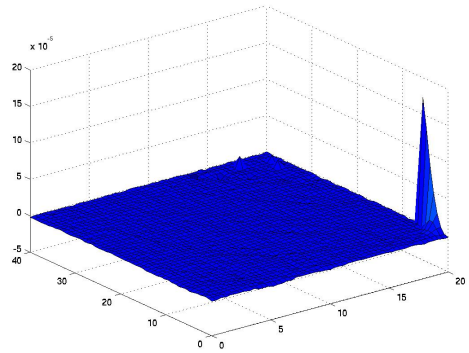

Abbildung 9.27: Wertefunktion$_{Differenz}$ Bsp5

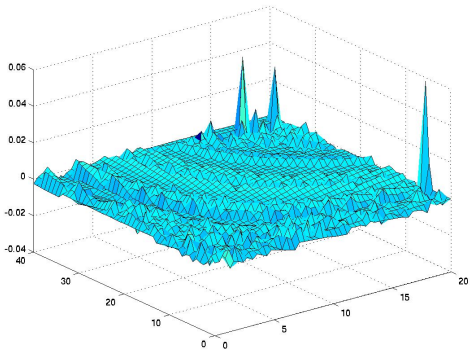

Abbildung 9.28: Steuerung$_{Differenz}$ Bsp5

Kapitel 10

Schlussbemerkungen

Die Aufgabenstellung dieser Arbeit bestand darin, die Bundle-Newton-Methode für das Maximierungsproblem, das innerhalb des betrachteten optimalen Steuerungsalgorithmus auftritt, zu verwenden und die so gewonnenen Ergebnisse mit den Resultaten aus Jarzcyk (2005) zu vergleichen. Im Vorfeld wurde angenommen, dass die Bundle-Newton-Methode wesentlich exaktere Ergebnisse liefert als die in Jarzcyk (2005) zugrundeliegenden Verfahren. Diese Erwartungen konnten jedoch nicht bestätigt werden.
Nur beim ersten einfachen Beispiel konnte bei gleicher Iterationenzahl des optimalen Steuerungsalgorithmus eine leichte Verbesserung erzielt werden. Dass der Aufwand pro Iteration unter Verwendung der Bundle-Newton-Methode deutlich größer ist, wird durch die enormen Laufzeitdifferenzen deutlich.
Schon Beispiel 2 zeigt, dass die Ergebnisse der Bundle-Newton-Methode hinter den mittels des Brent-Verfahrens erzeugten Werte zurückbleiben. Nach 3600 Iterationen liegt das globale Minimum gemäß der Bundle-Newton-Methode bei 13.391, das globale Maximum bei 30.994. Das Brent-Verfahren hingegen errechnet als globales Minimum den Wert 13.393 und als globales Maximum den Wert 31.008.
Beim dritten Beispiel verhält es sich ähnlich wie beim zuvor betrachteten Modell. Stimmt die Anzahl der Iterationen überein, so zeigt das Brent-Verfahren bzw. die rekursive Suche deutliche Vorteile.
Die Anwendung des Steuerungsalgorithmus auf das ökonomische Wachstumsmodell führt bei Zugrundeliegen der Bundle-Newton-Methode zu erheblichen Schwierigkeiten. Nur unter Zuhilfenahme von Zusatzmaßnahmen kann überhaupt ein Ergebnis erzielt werden.
Beim letzten Beispiel entsprechen die Resultate den besten Werten der in Jarczyk angewandten Verfahren.
Insgesamt legen die erhaltenen Testergebnisse die Vermutung nahe, dass die Zielfunktionen der betrachteten Optimierungsprobleme nicht immer den für eine erfolgreiche Anwendung der Bundle-Newton-Methode notwendigen Voraussetzungen genügen.

Anhang A

Verwendete Funktionen

Dieser Arbeit wurden drei Programme beigefügt.
Der Quellcode in programm.c ermöglicht die Berechnung der optimalen Wertefunktion sowie der optimalen Steuerung für die Beispiele 1, 2, 3a, 3b, 4 und 5 mittels den vier aus Jarczyk bekannten Verfahren. Des Weiteren können die Beispiele 1, 2, 3a, 3b und 5 mit Hilfe des Bundle-Newton-Verfahrens gelöst werden.
Da für Beispiel 4 keine Parameterkonstellation gefunden wurde, die zu einem selbständigen Abbruch des Programms führt, wurde diese Problemstellung isoliert in beispiel4.c betrachtet. Dieses Programm begrenzt die Iterationen des allgemeinen Steuerungsalgorithmus sowie der Bundle-Newton-Methode künstlich. Darüber hinaus wird die in Abschnitt 9.4 vorgestellte Schrittweitenstrategie verwendet und die inverse Matrixwurzel explizit berechnet.
Das Programm bundle_newton.c stellt eine Implementierung der Bundle-Newton-Methode für den eindimensionalen restringierten Fall dar, kann aber leicht auf höherdimensionale Problemstellungen übertragen werden.
Im Folgenden listen wir die verwendeten Routinen auf und erläutern sie kurz.

programm.c

Allgemeine Funktionen des optimalen Steuerungsalgorithmus:

- main()
  Hier erfolgt die Auswahl des gewünschten Beispiels sowie der Optimierungsstrategie. Die Ergebnisse werden im Anschluss in einer Datei abgespeichert.

- algorithmus()
  führt die Optimierung mit der gewählten Optimierungsmethode durch.

- beispiel()

Hier werden die für die Beispiele benötigten Parameter vorbelegt. Der Nutzer kann diese vor dem Programmstart ändern.

- Fkt_f()
  Hier finden sich die zu den Beispielen gehörigen Kontrollsysteme.
- Fkt_g()
  enthält die Ertragsfunktionen der zugrundeliegenden fünf Modellbeispiele.
- solution()
  berechnet die exakte Lösung für das vierte Beispiel.
- Euler()
  führt einen Euler-Schritt durch.
- Koordinaten()
  berechnet aus dem globalen Eckenindex die Koordinaten der Ecke.
- FindeRechteckKoord()
  ermittelt den Rechteckindex eines Rechtecks, in dem sich ein Punkt befindet, anhand der Koordinaten des Punktes.
- Gitter()
  liefert zum globalen Rechteckindex und zum lokalen Eckenindex den globalen Eckenindex.
- Wert()
  errechnet zu einem Punkt den approximierten Wert v durch eine affin bilineare Funktion.

Für das Bundle-Newton-Verfahren verwendete Routinen:

- bundle_newton()
  ist dafür geeignet, das Minimum einer konvexen Funktion zu berechnen. Um die Bundle-Newton-Methode auf das Maximierungsproblem (4.7) anwenden zu können, werden die Zielfunktionswerte in der Funktion funktion_bundle() mit (-1) multipliziert und später wieder zurücktransformiert.
- funktion_bundle()
  unterscheidet sich von der Funktion funktion() lediglich durch die Multiplikation des Zielfunktionswertes mit (-1).
- FunktionswertF()
  berechnet den Funktionswert der Nebenbedingungsfunktion H aus (8.2).

- ersteAbleitungf()
  approximiert die Ableitung bzw. den Subgradienten von h aus (8.2).
- ersteAbleitungF()
  ermittelt die Ableitung bzw. den Subgradienten von H aus (8.2).
- Hessematrixf()
  liefert eine Approximation der Hessematrix der Zielfunktion.
- HessematrixF()
  berechnet die Hessematrix der Nebenbedingungsfunktion.
- Choleskyerweiterung()
  wird verwendet, um zu einer positiv semidefiniten Matrix eine positiv definite Approximation dieser zu ermitteln.
- eukNorm()
  bestimmt die euklidische Norm eines Vektors.
- Invertierung()
  dient der Invertierung einer Matrix.
- MatrixMatrixMultiplikation()
  liefert das Produkt zweier Matrizen.
- MatrixVektorMultiplikation()
  multipliziert eine Matrix mit einem Vektor.
- VektorVektorMultiplikation()
  berechnet das Produkt aus zwei Vektoren.
- Transponierte()
  transponiert eine Matrix.
- QuadOpt()
  löst das quadratische Optimierungsproblem (7.28) aus Schritt S1 des Bundle-Newton-Algorithmus.
- Schrittweitenbestimmung()
  ermittelt geeignete Schrittweiten entsprechend des Algorithmus aus Abschnitt 7.2.
- spektralnorm()
  berechnet die Spektralnorm einer Matrix

- max()
  bildet das Maximum zweier reeller Zahlen.
- min()
  liefert das Minimum zweier reeller Zahlen.
- admit_intervall()
  berechnet zu einem Punkt x das zulässige Intervall für die Steuerungsvariable u.

Für die Erläuterung der Funktionen, die für die in Jarzcyk (2005) implementierten Verfahren benötigt werden, verweisen wir auf die genannte Arbeit.

beispiel4.c

Für die Bundle-Newton-Methode in diesem Programm wird die modifzierte Liniensuche benutzt. Des Weiteren wird statt der Funktion Invertierung() die Routine InverseWurzel() verwendet, die neben einer Invertierung auch die Wurzelbildung vornimmt.

bundle_newton.c

Die isolierte Bundle-Newton-Methode greift auf die Funktion Invertierung() und die Liniensuche aus Abschnitt 9.4 zurück.

Anhang B

Programmbedienung

Die drei Programme programm.c, beispiel4.c und bundle_newton.c wurden in C implementiert und binden Fortran-Funktionen ein. Sie können über den Befehl
$gcc \quad Programmname.c \quad -lnagc \quad -lnag \quad /usr/lib/libf2c.a \quad -lm$
$\qquad -I/share/lib/NAG/NAGC\_Mark6/include \quad -I/usr/include \quad -lpthread$
compiliert und mit *a.out* ausgeführt werden.

programm.c

Mit dem Programm programm.c können die optimalen Wertefunktionen sowie die optimalen Steuerungen der Beispiele 1, 2, 3a, 3b und 5 mittels der Bundle-Newton-Methode berechnet werden. Des Weiteren ist es dafür geeignet, die Verfahren von Jarzcyk (2005) auf alle behandelten Beispiele anzuwenden. Zuerst erfolgt eine Abfrage des gewünschten Beispiels. Anschließend kann das Verfahren ausgewählt werden. Hierbei müssen bei den Methoden aus Jarzcyk (2005) Intervallunterteilungen angegeben werden. Für Erläuterungen zu deren Bedeutung sei auf die genannte Arbeit verwiesen. Bei der Auswahl der Bundle-Newton-Methode sind keine weiteren Parameter beim Programmaufruf zu wählen. Alle Größen wurden bereits vorbelegt. In beispiel() können die zum optimalen Steuerungsalgorithmus gehörigen Parameter h, δ, ϵ, $n[0]$, $n[1]$, $a[0]$, $a[1]$, $b[0]$, $b[1]$, $u[0]$, $u[1]$ sowie einige Parameter der Bundle-Newton-Methode, nämlich γ_f, γ_F, ω, $\tilde{h}$, ε und q, verändert werden. Des Weiteren hat der Nutzer die Möglichkeit, die Werte von m_R, m_L, t_0, ϑ, ζ, C_S, C_G, i_m und i_r neu zu belegen. Nähere Erläuterungen zu den genannten Parametern finden sich in den Kapiteln 7 und 9. Die folgenden Übersichten zeigen, welche Parameterwerte bei den jeweiligen Beispielen herangezogen und welche Ergebnisse erzeugt werden:

m_L	m_R	t_0	ϑ	ζ	C_S	C_G	i_m	i_r
0.01	0.5	0.001	1.0	0.01	10^{50}	10^{50}	100	100

Beispiel	**Verfahren**	**Iter**	**Min**	**Max**
Beispiel 1	bundle-newton$[0.5, 0.5, 1.0, 10^{-6}, 10^{-5}, 1.0]$	247	-1.9978	-1.2040
Beispiel 2	bundle-newton$[0.5, 0.5, 1.0, 10^{-6}, 10^{-4}, 2.5]$	1867	12.980	30.560
Beispiel 3a	bundle-newton$[10^{-3}, 10^{-5}, 1.0, 10^{-6}, 10^{-4}, 1.0]$	976	-6.579	6.041
Beispiel 3b	bundle-newton$[0.5, 0.5, 1.0, 10^{-6}, 10^{-4}, 1.0]$	874	-9.265	18.076
Beispiel 5	bundle-newton$[10^{-7}, 10^{-3}, 1.0, 10^{-6}, 10^{-6}, 2.0]$	30	00.000	31.533

beispiel4.c

Im Programm beispiel4.c wurde das vierte Beispiel aufgrund der Besonderheiten bei der Implementierung im Vergleich zu den anderen Beispielen einzeln betrachtet. Nach der Ausführung berechnet der Algorithmus die optimale Steuerung sowie die optimale Wertefunktion für bundle-newton$[0.5, 0.5, 1.0, 10^{-7}, 10^{-6}, 1.5]_{160}$ ohne zusätzliche Eingaben des Nutzers. Als Ergebnis erhalten wir das globale Minimum 23.72402 sowie das globale Maximum 34.18971. Die Parameter können vor der Ausführung im Quellcode verändert werden.

bundle_newton.c

Das Programm bundle_newton.c ist eine isolierte Implementierung der Bundle-Newton-Methode für den restringierten eindimensionalen Fall. Als Beispiel haben wir die Problemstellung

$$\begin{aligned} \min \quad & x^2 \\ \text{u.d.N.} \quad & x \in [-2.5, 1.0] \end{aligned}$$

herangezogen. Sowohl das Optimierungsproblem als auch die Parameter können im Quellcode modifiziert werden.

Anhang C

CD-ROM Inhalt

Die beigefügte CD enthält unter dem Ordner Programme die Quellcodes der Programme programm.c, beispiel4.c und bundle_newton.c sowie unter dem Ordner Arbeit die Diplomarbeit als pdf-Datei.

Literaturverzeichnis

M. S. **Bertsekas**. Nonlinear Programming. Athena Scientific, Belmont, MA (2. Auflage), **1999**.

F. H. **Clarke**. Optimization and Nonsmooth Analysis. Wiley-Interscience, New York, **1983**.

C. **Geiger** und C. **Kanzow**. Theorie und Numerik restringierter Optimierungsaufgaben. Springer, **2002**.

M. **Gerdts**. Nichtdifferenzierbare Optimierung. Vorlesungsskript, Fakultät Mathematik und Physik, Universität Bayreuth, **2003**.

P. E. **Gill** und W. **Murray**. Newton-type methods for unconstrained and linearly constrained optimization. Mathematical Programming 28, 311-350, **1974**.

P. E. **Gill**, W. **Murray** und M. H. **Wright**. Practical Optimization. Academic Press, London (4. Auflage), **1984**.

R. L. **Gonzáles** und M. M. **Tidball**. On a Discrete Time Approximation of the Hamilton-Jacobi Equation of Dynamic Programming. INRIA Rapports de Recherche Nr. 1375, **1991**.

L. **Grüne**. Numerische Dynamik von Kontrollsystemen. Vorlesungsskript, Fakultät Mathematik und Physik, Universität Bayreuth, **2004**.

R. A. **Horn** und C. A. **Johnson**. Matrix Analysis. Cambridge University Press, **1985**.

M. **Jarczyk**. Lokale Optimierungsstrategien in der dynamischen Programmierung. Diplomarbeit, Fakultät Mathematik und Physik, Universität Bayreuth, **2005**.

K. C. **Kiwiel**. Methods of Descent for Nondifferentiable Optimization. volume 1133 of Lecture Notes in Mathematics. Springer, **1985**.

C. **Lemaréchal**, J. J. **Strodiot** und A. **Bihain**. On a Bundle Algorithm for Nonsmooth Optimization. In O. L. Mangasarian, R. R. Meyer and S. M. Robinson, editors, Nonlinear Programming 4, 245-282, Academic Press, New York, **1981**.

L. **Luksan** und J. **Vlček**. A Bundle-Newton method for nonsmooth unconstrained

minimization. Mathematical Programming 83, 373-391, **1998**.

M.M. **Mäkelä** und P. **Neittaanmäki**. Nonsmooth Optimiziation: Analysis and Algorithms with Applications to Optimal Control. World Scientific Publishing Co. Pte.Ltd., **1992**.

H. **Schramm**. Eine Kombination von Bundle- und Trust-Region-Verfahren zur Lösung nichtdifferenzierbarer Optimierungsprobleme. Bayreuther Mathematische Schriften, Heft 30, Bayreuth, **1989**.

E. D. **Sontag**. Mathematical Control Theory, Springer Verlag, New York (2nd Edition), **1998**.