Numerische Verfahren der restringierten Optimierung

Stefan Volkwein

28. Januar 2009

Ao. Univ.-Prof. Dr. Dipl.-Math. techn. Stefan Volkwein, Institut für Mathematik und Wissenschaftliches Rechnen, Karl-Franzens Universität Graz, Österreich

E-mail address: stefan.volkwein@uni-graz.at

Das Manuskript entstand während der vierstündigen Vorlesung Optimierung II im Wintersemester 2004/05 am Institut für Mathematik und Wissenschaftliches Rechnen der Karl-Franzens Universität Graz und wurde im Wintersemester 2008/09 erweitert.

Danksagung

Ich danke Frau A. Griesbacher sowie den Herren B. Gotthardt, M. Kahlbacher, I. Kopacka und H. Müller für die gründliche Durchsicht des Skriptes, welches dadurch deutlich verbessert werden konnte.

Inhaltsverzeichnis

KAPITEL 1

Optimalitätsbedingungen für die restringierte Optimierung

Das Ziel in diesem Abschnitt ist die Herleitung notwendiger und hinreichender Optimalitätsbedingungen erster und zweiter Ordnung. Für mehr Details und weitere Beispiele verweisen wir auf [**26**, § 10].

1. Nebenbedingungen

Wir betrachten

$$\textbf{(P)} \qquad \min J(x) \quad \text{u.d.N.} \quad x \in \mathbb{R}^n,\ e(x) = 0 \text{ und } g(x) \leq 0,$$

wobei $J : \mathbb{R}^n \to \mathbb{R}$ das *Kosten-* oder *Zielfunktional* ist, $e = (e_1, \ldots, e_m)^T : \mathbb{R}^n \to \mathbb{R}^m$, $m \leq n$, die *Gleichungs-Nebenbedingungen* und $g = (g_1, \ldots, g_p)^T : \mathbb{R}^n \to \mathbb{R}^p$ die *Ungleichungs-Nebenbedingungen* sind. In (**P**) wie im weiteren steht die Abkürzung "u.d.N." für "unter der Nebenbedingung". Wir schreiben $g(x) \leq 0$ genau dann, wenn $g_i(x) \leq 0$ für alle Komponenten $i = 1, \ldots, p$ von g gilt. Ein Punkt $x \in \mathbb{R}^n$ heißt *zulässig* für (**P**), sofern $e(x) = 0$ und $g(x) \leq 0$ erfüllt sind. Die *zulässige Menge* für (**P**) bezeichnen wir mit

$$\mathcal{F}(\mathbf{P}) = \big\{x \in \mathbb{R}^n \,|\, e(x) = 0 \text{ und } g(x) \leq 0\big\}.$$

Bei den Ungleichungen unterscheiden wir zwei Fälle. An $x \in \mathcal{F}(\mathbf{P})$ heißt eine Ungleichungs-Nebenbedingung $g_i(x)$ *aktiv* für ein $i \in \{1, \ldots, p\}$, wenn $g_i(x) = 0$ gilt, und *inaktiv*, wenn $g_i(x) < 0$ erfüllt ist.

In der folgenden Definition wollen wir den Begriff einer Lösung von (**P**) einführen.

Definition 1.1. *Sei x^* ein Punkt im $\mathbb{R}^n$.*

1) *Der Punkt x^* wird* lokale Lösung *von* (**P**) *genannt, wenn $x^* \in \mathcal{F}(\mathbf{P})$ und $J(x^*) \leq J(x)$ für alle $x \in U(x^*) \cap \mathcal{F}(\mathbf{P})$ gelten, wobei $U(x^*) \subseteq \mathbb{R}^n$ eine Umgebung des Punktes x^* bezeichnet.*
2) *Der Punkt x^* ist eine* strikte lokale Lösung *von* (**P**)*, wenn $x \in \mathcal{F}(\mathbf{P})$ und $J(x^*) < J(x)$ für alle $x \in U(x^*) \cap \mathcal{F}(\mathbf{P})$ erfüllt sind, wobei $U(x^*) \subseteq \mathbb{R}^n$ wieder eine Umgebung des Punktes x^* bezeichnet.*
3) *Der Punkt x^* ist eine* globale Lösung *von* (**P**)*, wenn $x^* \in \mathcal{F}(\mathbf{P})$ und $J(x^*) \leq J(x)$ für alle $x \in \mathcal{F}(\mathbf{P})$ gelten. Wir nennen x^* eine* strikte globale Lösung *von* (**P**)*, wenn $x \in \mathcal{F}(\mathbf{P})$ und $J(x^*) < J(x)$ für alle $x^* \in \mathcal{F}(\mathbf{P})$ erfüllt sind.*

Globale Lösungen sind im allgemeinen schwieriger zu bestimmen als lokale.

Beispiel 1.2. Wir betrachten das Problem

$$\min \|x\|_2^2 \quad \text{u.d.N.} \quad x \in \mathbb{R}^n \text{ und } \|x\|_2^2 \geq 1, \tag{1.1}$$

wobei $\|\cdot\|_2$ die Euklidische Norm bezeichnet. Offenbar ist (1.1) ohne Ungleichungs-Nebenbedingungen ein konvexes, quadratisches Problem mit der eindeutigen Lösung $x^* = 0$. Wegen der Bedingung $\|x\|_2 \geq 1$ löst (**P**) jedes $x \in \mathbb{R}^n$ mit $\|x\|_2 = 1$, insbesondere gibt es unendlich viele Lösungen für $n \geq 2$. $\Diamond$

Wenn wir a-priori wissen, dass eine Lösung von (**P**) aktiv ist in allen Komponenten $i \in \{1, \ldots, p\}$, so können wir die Ungleichungs-Nebenbedingungen ignorieren. Wir erhalten dann ein Problem mit Gleichungs-Nebenbedingungen. Diese Probleme werden wir zunächst genauer untersuchen.

2. Die Tangentialebene

Wir setzen nun voraus, dass die beiden Funktionen J und e stetig differenzierbar seien.

Um die *Tangentialebene* einzuführen, werden wir den Begriff der Kurven auf Hyperflächen verwenden. Eine *Kurve* in einer Hyperfläche $\mathcal{H} \subset \mathbb{R}^n$ ist eine Familie von Punkten $x(t) \in \mathcal{H}$, wobei $x : [a, b] \to \mathcal{H}$ eine stetige Abbildung ist. Die Kurve heißt *differenzierbar in* t, wenn $\dot{x}(t) = \frac{\mathrm{d}x}{\mathrm{d}t}(t)$ existiert, und sie *zweimal differenzierbar in* t, wenn $\ddot{x}(t) = \frac{\mathrm{d}^2x}{\mathrm{d}t^2}(t)$ existiert. Wir sagen, die Kurve $x(t)$ *geht durch einen Punkt* $\bar{x} \in \mathcal{H}$, wenn für ein $\bar{t} \in [a, b]$ die Bedingung $x(\bar{t}) = \bar{x}$ gilt. Die Tangentialebene an $\bar{x} \in \mathcal{H}$ ist die Menge der Tangentialvektoren $\dot{x}(\bar{t})$ aller durch den Punkt $\bar{x}$ gehenden differenzierbaren Kurven in $\mathcal{H}$.

Wir definieren die (glatte) Fläche

$$\mathcal{E} = \big\{x \in \mathbb{R}^n \mid e(x) = 0\big\} \tag{1.2}$$

in $\mathbb{R}^n$. Unser Ziel ist es nun, die Tangentialebene an einem Punkt $\bar{x} \in \mathcal{E}$ mit Hilfe der Gradienten von e_i, $1 \leq i \leq m$, zu beschreiben. Daher betrachten wir die Menge

$$\operatorname{Kern} \nabla e(\bar{x}) = \big\{v \in \mathbb{R}^n \mid \nabla e(\bar{x})v = 0\big\},$$

wobei $\nabla e(\bar{x}) \in \mathbb{R}^{m \times n}$ die Funktional-Matrix

$$\nabla e(\bar{x}) = \begin{pmatrix} \nabla e_1(\bar{x}) \\ \vdots \\ \nabla e_m(\bar{x}) \end{pmatrix}$$

bezeichnet und $\nabla e_i(\bar{x}) \in \mathbb{R}^{1 \times n}$ der Gradient von e_i an $\bar{x}$ ist, $1 \leq i \leq m$.

Wir wollen in Satz 1.4 zeigen, dass $\operatorname{Kern} \nabla e(\bar{x})$ die Tangentialebene an $\mathcal{E}$ im Punkt $\bar{x}$ darstellt. Dazu muss die Funktional-Matrix ∇e am Punkt $\bar{x}$ aber folgende Eigenschaften haben:

Definition 1.3. *Ein Punkt $\bar{x} \in \mathcal{E}$ heißt* regulärer Punkt *(oder einfach* regulär*) bezüglich der Nebenbedingung $e(x) = 0$, wenn die Gradienten $\nabla e_1(\bar{x}), \ldots, \nabla e_m(\bar{x})$ linear unabhängig in $\mathbb{R}^n$ sind.*

Nun können wir folgenden Satz beweisen.

Satz 1.4. *Sei $\bar{x} \in \mathcal{E}$ ein regulärer Punkt. Dann ist die Tangentialebene an $\bar{x}$ gleich der Menge* $\operatorname{Kern} \nabla e(\bar{x})$.

Beweis. Sei T die Tangentialebene am Punkt $\bar{x}$. Die Inklusion $T \subset \operatorname{Kern} \nabla e(\bar{x})$ ist erfüllt, auch ohne dass $\bar{x}$ ein regulärer Punkt ist; denn ist $x(t)$ eine Kurve, die durch $\bar{x}$ geht mit $x(\bar{t}) = \bar{x}$ und $\nabla e(\bar{x})\dot{x}(\bar{t}) \neq 0$, so liegt die Kurve $x = x(t)$ nicht in $\mathcal{E}$.

Nun zu $\operatorname{Kern} \nabla e(\bar{x}) \subset T$. Zu zeigen ist, dass für beliebiges $v \in \operatorname{Kern} \nabla e(\bar{x})$ eine Kurve $x(t)$ existiert, die durch $\bar{x}$ geht mit dem Tangentialvektor v. Wir betrachten dazu die Gleichung

$$e(\underbrace{\bar{x} + tv + \nabla e(\bar{x})^T u(t)}_{=x(t)}) = 0, \tag{1.3}$$

wobei $u(t) \in \mathbb{R}^m$ zu bestimmen ist. Offenbar ist (1.3) ein nicht-lineares Gleichungs-System mit m Gleichungen für die m Unbekannten $u(t)$, und die Abhängigkeit von t ist stetig. Wegen $e(\bar{x}) = 0$, löst $u(0) = 0$ die Gleichung (1.3) für $t = 0$. Ferner ist die Jacobi-Matrix von (1.3) bezüglich der Unbekannten u

$$\nabla e(\bar{x}) \nabla e(\bar{x})^T \in \mathbb{R}^{m \times m}$$

invertierbar, da $\bar{x}$ ein regulärer Punkt ist. Aufgrund des Satzes über Implizite Funktionen existieren ein $\varepsilon > 0$ und eine stetig differenzierbare Abbildung $u : [-\varepsilon, \varepsilon] \to \mathbb{R}^m$, so dass (1.3) für alle $t \in [-\varepsilon, \varepsilon]$ erfüllt ist. Wegen (1.3) liegt $x(t) = \bar{x} + tv + \nabla e(\bar{x})^T u(t)$, $t \in [-\varepsilon, \varepsilon]$, in $\mathcal{E}$ mit $x(0) = \bar{x}$. Wir differenzieren die Gleichung (1.3) bezüglich der Variablen t an $t = 0$ und erhalten

$$\begin{aligned} 0 = \frac{\mathrm{d}}{\mathrm{d}t} e(x(t))\big|_{t=0} &= \big(\nabla e(x(t))\big(v + \nabla e(\bar{x})^T \dot{u}(t)\big)\big)\big|_{t=0} \\ &= \nabla e(\bar{x}) v + \nabla e(\bar{x}) \nabla e(\bar{x})^T \dot{u}(0). \end{aligned}$$

Wegen $v \in \operatorname{Kern} \nabla e(\bar{x})$ folgt $\nabla e(\bar{x}) v = 0$, Da $\bar{x}$ ein regulärer Punkt ist, folgen $\dot{u}(0) = 0$ und damit

$$\dot{x}(0) = \frac{\mathrm{d}}{\mathrm{d}t}\big(\bar{x} + tv + \nabla e(\bar{x})^T u(t)\big)\big|_{t=0} = v.$$

Wir haben damit eine Kurve $x(t)$ gefunden, die in $\mathcal{E}$ liegt und deren Tangentialvektor an $x(0) = \bar{x}$ gleich $v \in \operatorname{Kern} \nabla e(\bar{x})$ ist. Damit liegt v in der Tangentialebene an $\mathcal{E}$ im Punkt $\bar{x}$, was zu zeigen war. □

Bemerkung 1.5. Die Voraussetzung, dass $\bar{x}$ regulär ist, stellt keine Voraussetzung an die Menge $\mathcal{E}$ dar, sondern an die Darstellung mittels der Funktion e. Die Tangentialebene an $\bar{x}$ ist unabhängig von der Repräsentation von $\mathcal{E}$ durch die Abbildung e, die Menge $\operatorname{Kern} \nabla e(\bar{x})$ aber offensichtlich nicht. ◊

Beispiel 1.6. Seien $n = 2$, $m = 1$ und $e(x_1, x_2) = x_1$. Dann ist die Menge $\mathcal{E}$ die x_2-Achse. Wegen $\nabla e(0, x_2) = (1, 0) \neq (0, 0)$ sind alle Punkte in $\mathcal{E}$ regulär. Repräsentieren wir aber die Menge $\mathcal{E}$ mit der Funktion $e(x_1, x_2) = x_1^2$, erhalten wir allerdings $\nabla e(0, x_2) = (0, 0)$ für alle Punkte aus $\mathcal{E}$. Damit ist in diesem Fall kein Punkt aus $\mathcal{E}$ regulär. ◊

3. Notwendige Bedingungen 1. Ordnung für Gleichungs-Restriktionen

Wir betrachten nun an der Stelle von (**P**) das (in der Regel einfachere) Problem

$$\min J(x) \quad \text{u.d.N.} \quad x \in \mathbb{R}^n \text{ und } e(x) = 0. \tag{$\mathbf{P}_{Gl}$}$$

Das Haupt-Resultat dieses Abschnittes lautet wie folgt: Sei $x^* \in \mathbb{R}^n$ ein lokales Minimum von ($\mathbf{P}_{Gl}$) und ein regulärer Punkt. Dann existiert ein *Lagrange-Multiplikator* $\lambda^* = (\lambda_1^*, \ldots, \lambda_m^*)^T \in \mathbb{R}^m$ mit

$$\nabla J(x^*) + \sum_{i=1}^{m} \lambda_i^* \nabla e_i(x^*) = 0. \tag{1.4}$$

Die Gleichung (1.4) lässt sich wie folgt interpretieren:

1) Der Gradient von J an x^* liegt im Span der Gradienten der Nebenbedingung:
$$\nabla J(x^*) \in \operatorname{Span}\big\{\nabla e_1(x^*), \dots, \nabla e_m(x^*)\big\}.$$
2) Schreiben wir
$$\sum_{i=1}^{m} \lambda_i^* \nabla e_i(x^*) = (\lambda^*)^T \nabla e(x^*)$$
und transponieren die Gleichung (1.4), so erhalten wir
$$\nabla e(x^*)^T \lambda^* = -\nabla J(x^*)^T.$$
Damit gilt $\nabla J(x^*)^T \in \operatorname{Bild} \nabla e(x^*)^T$. Der transponierte Gradient $\nabla J(x^*)^T$ liegt damit im Bild der adjungierten/transponierten Matrix $\nabla e(x^*)^T$. Da x^* regulär ist, ist $\nabla e(x^*)$ surjektiv und daher $\nabla e(x^*)^T$ injektiv. Es kann daher also höchstens nur ein λ^* geben.
3) Multiplizieren wir Gleichung (1.4) mit $v \in \operatorname{Kern} \nabla e(x^*)$, so folgt sofort
$$\nabla J(x^*) v = 0 \quad \text{für alle } v \in \operatorname{Kern} \nabla e(x^*).$$
Für Variationen $v \in \operatorname{Kern} \nabla e(x^*)$ erfüllt $x = x^* + v$ die Nebenbedingung $e(x) = 0$ bis zur ersten Ordnung. Damit darf J bezüglich dieser Variation bis zur ersten Ordnung nicht wachsen oder fallen.

Beispiel 1.7. Betrachte das Optimierungs-Problem
$$\min J(x) = x_1 + x_2 \quad \text{u.d.N.} \quad x_1^2 + x_2^2 = 2.$$
Wie wir uns leicht grafisch überlegen können, ist $x^* = (-1, -1)$ die eindeutige Lösung. Wir berechnen $\nabla J(x^*) = (1, 1)$ und $\nabla e(x^*) = (-2, -2) \neq (0, 0)$. Damit ist x^* ein regulärer Punkt, und mit $\lambda^* = 1/2$ folgt (1.4). ◊

Beispiel 1.8. Wir wollen nun das Problem
$$\min J(x) = x_1 + x_2 \quad \text{u.d.N.} \quad \begin{pmatrix} e_1(x) \\ e_2(x) \end{pmatrix} = \begin{pmatrix} (x_1 - 1)^2 + x_2^2 - 1 \\ (x_1 - 2)^2 + x_2^2 - 4 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$$
untersuchen. Hier ist die Menge $\mathcal{E}$ ein-elementig, und $x^* = (0, 0)$ der einzige zulässige Punkt und damit die (triviale) Lösung des Minimierungs-Problems. Der Gradient von J ist wie im vorangegangenen Beispiel gegeben durch $\nabla J(x^*) = (1, 1)$. Ferner ergeben sich wegen $\nabla e_1(x_1, x_2) = (2(x_1 - 1), 2x_2)$ und $\nabla e_2(x_1, x_2) = (2(x_1 - 2), 2x_2)$ die Gradienten von e_1 und e_2 an x^* zu $\nabla e_1(x^*) = (-2, 0)$ beziehungsweise $\nabla e_2(x^*) = (-4, 0)$. Damit sind die Gradienten $\nabla e_1(x^*)$ und $\nabla e_2(x^*)$ linear abhängig in $\mathbb{R}^2$, der Punkt x^* kann also nicht regulär sein. Offenbar löst x^* löst die Minimierungs-Aufgabe, es gibt aber kein $\lambda^* \in \mathbb{R}^2$, so dass (1.4) erfüllt ist. ◊

Wir kommen nun zur Formulierung des Haupt-Resultates von diesem Abschnitt.

Satz 1.9 (Notwendige Bedingungen 1. Ordnung). *Sei der Punkt x^* eine lokale Lösung von $(\mathbf{P}_{Gl})$ und ein regulärer Punkt. Dann existiert genau ein $\lambda^* = (\lambda_1^*, \dots, \lambda_m^*)^T \in \mathbb{R}^m$, der sogenannte Lagrange-Multiplikator, so dass*
$$\nabla J(x^*) + \sum_{i=1}^{m} \lambda_i^* \nabla e_i(x^*) = \nabla J(x^*) + (\lambda^*)^T \nabla e(x^*) = 0. \tag{1.5}$$

BEWEIS. Offenbar ergibt $m = n$ ein triviales Resultat, da $\nabla J(x^*) \in \mathbb{R}^n$ und, nach Voraussetzung, $\text{Span}\{\nabla e_1(x^*), \dots, \nabla e_m(x^*)\} = \mathbb{R}^n$ gelten. Sei also $m < n$. Wir sortieren den Vektor x um, indem wir $x = (x_B, x_R) \in \mathbb{R}^n$ schreiben mit $x_B \in \mathbb{R}^m$, $x_R \in \mathbb{R}^{n-m}$, so dass $\nabla_B e(x^*) \in \mathbb{R}^{m\times m}$ invertierbar ist. Hierbei bezeichnet ∇_B den Gradient mit den partiellen Ableitungen bezüglich des Vektors x_B. Da $x^* = (x_B^*, x_R^*)$ Lösung von $(\mathbf{P}_{Gl})$ ist, haben wir $e(x_B^*, x_R^*) = 0$. Wir werden den Satz über Implizite Funktionen anwenden, um die Variable x_B durch x_R auszudrücken. Da $\nabla_B e(x^*)$ invertierbar ist, existieren ein Radius $r > 0$ und eine stetig differenzierbare Funktion

$$\Phi : B_r(x_R^*) = \big\{x_R \in \mathbb{R}^{n-m} \,|\, \|x_R - x_R^*\| < r\big\} \to \mathbb{R}^m$$

mit

$$\Phi(x_R^*) = x_B^* \quad \text{und} \quad e(\Phi(x_R), x_R) = 0 \text{ für alle } x_R \in B_r(x_R^*).$$

Ferner ist der Gradient von Φ bezüglich des Vektors x_R gegeben durch

$$\nabla_R \Phi(x_R) = -\big(\nabla_B e(\Phi(x_R), x_R)\big)^{-1} \nabla_R e(\Phi(x_R), x_R) \quad \text{für alle } x_R \in B_r(x_R^*),$$

was wir sofort aus der Identität $e(\Phi(x_R), x_R) = 0$ durch Differenzieren nach x_R erhalten. Unter Verwendung der Abbildung Φ betrachten wir das *reduzierte Problem*

$$\min \hat{J}(x_R) \quad \text{u.d.N.} \quad x_R \in B_r(x_R^*) \subset \mathbb{R}^{n-m} \tag{1.6}$$

mit $\hat{J} : \mathbb{R}^{n-m} \to \mathbb{R}$, $\hat{J}(x_R) = J(\Phi(x_R), x_R)$. Die Abbildung $\hat{J}$ wird auch als reduziertes Kostenfunktional bezeichnet. Offenbar löst x_R^* lokal das (unrestringierte) Problem (1.6). Daher folgt die notwendige Optimalitätsbedingung

$$\nabla_R \hat{J}(x_R^*) = 0. \tag{1.7}$$

Für der Gradienten von $\hat{J}$ erhalten wir

$$\begin{aligned}
\nabla_R \hat{J}(x_R) &= \underbrace{\nabla_B J(\Phi(x_R), x_R)}_{\in \mathbb{R}^{1\times m}} \underbrace{\nabla_R \Phi(x_R)}_{\in \mathbb{R}^{m\times(n-m)}} + \underbrace{\nabla_R J(\Phi(x_R), x_R)}_{\in \mathbb{R}^{1\times(n-m)}} \\
&= -\underbrace{\nabla_B J(\Phi(x_R), x_R)}_{\in \mathbb{R}^{1\times m}} \underbrace{\big(\nabla_B e(\Phi(x_R), x_R)\big)^{-1}}_{\in \mathbb{R}^{m\times m}} \underbrace{\nabla_R e(\Phi(x_R), x_R)}_{\in \mathbb{R}^{m\times(n-m)}} \\
&\quad + \nabla_R J(\Phi(x_R), x_R),
\end{aligned}$$

wobei wir den Gradienten von Φ nach dem Satz über Implizite Funktionen ersetzt haben. Wir führen nun den Lagrange-Multiplikator $\lambda^* \in \mathbb{R}^m$ durch die Gleichung

$$\lambda^* = -\Big(\nabla_B J(\Phi(x_R^*), x_R^*) \big(\nabla_B e(\Phi(x_R^*), x_R^*)\big)^{-1}\Big)^T = -\big(\nabla_B e(x^*)\big)^{-T} \nabla_B J(x^*)^T$$

ein. Wir erhalten somit aus (1.7)

$$\nabla_R J(x^*) + (\lambda^*)^T \nabla_R e(x^*) = 0 \tag{1.8a}$$

Ferner löst λ^* die *adjungierte/duale Gleichung*

$$\nabla_B e(x^*)^T \lambda^* = -\nabla_B J(x^*)^T$$

beziehungsweise die Gleichung

$$\nabla_B J(x^*) + (\lambda^*)^T \nabla_B e(x^*) = 0. \tag{1.8b}$$

Aus (1.8) folgt (1.5), was zu zeigen war. □

BEMERKUNG 1.10. a) Die notwendigen Optimalitätsbedingungen erster Ordnung

$$\nabla J(x^*) + (\lambda^*)^T \nabla e(x^*) = 0$$

zusammen mit der Nebenbedingung

$$e(x^*) = 0$$

ergeben ein (nichtlineares) Gleichungs-System mit $n+m$ Gleichungen für die $n+m$ Unbekannten $x^* \in \mathbb{R}^n$ und $\lambda^* \in \mathbb{R}^m$.

b) Der Beweis von Satz 1.9 zeigt auch, wie der Lagrange Multiplikator λ^* berechnet werden kann. Für gegebene optimale Lösung x^* von $(\mathbf{P}_{Gl})$ löst λ^* das lineare System

$$\nabla_B e(x^*)^T \lambda^* = -\nabla_B J(x^*)^T$$

Offenbar ist $\lambda^* = 0$ im Fall von $\nabla_B J(x^*) = 0$. ◊

Wir bezeichnen mit $\langle \cdot\,, \cdot \rangle_{\mathbb{R}^m}$ das Euklidische Skalarprodukt im $\mathbb{R}^m$. Mit der Einführung der *Lagrange-Funktion*

$$L(x, \lambda) = J(x) + \lambda^T e(x) = J(x) + \langle \lambda, e(x) \rangle_{\mathbb{R}^m} \tag{1.9}$$

lassen sich die Optimalitätsbedingungen (1.8) kompakt schreiben mit Hilfe des Gradienten von L:

$$\nabla_x L(x^*, \lambda^*) = \nabla J(x^*) + (\lambda^*)^T \nabla e(x^*) = 0, \tag{1.10a}$$

$$\nabla_\lambda L(x^*, \lambda^*) = e(x^*)^T = 0, \tag{1.10b}$$

also insgesamt $\nabla L(x^*, \lambda^*) = 0$.

BEISPIEL 1.11. Wir betrachten das Problem

$$\min x_1 x_2 + x_2 x_3 + x_1 x_3 \quad \text{u.d.N.} \quad x_1 + x_2 + x_3 = 3.$$

Um die Form $(\mathbf{P}_{Gl})$ zu erhalten, setzen wir $n = 3$, $m = 1$, $J(x) = x_1x_2 + x_2x_3 + x_1x_3$ und $e(x) = x_1 + x_2 + x_3 - 3$ für $x = (x_1, x_2, x_3) \in \mathbb{R}^3$. Wegen $\nabla e(x) = (1, 1, 1)$ ist jeder Punkt $x \in \mathbb{R}^3$ regulär. Zum Aufstellen des Gleichungssystems (1.10) führen wir die Lagrange-Funktion gemäß (1.9) ein:

$$L(x, \lambda) = J(x) + \lambda e(x) = x_1x_2 + x_2x_3 + x_1x_3 + \lambda(x_1 + x_2 + x_3 - 3).$$

Dann folgen die beiden Gleichungen:

$$\nabla_x L(x, \lambda) = (x_2 + x_3 + \lambda, x_1 + x_3 + \lambda, x_1 + x_2 + \lambda) = 0,$$
$$\nabla_\lambda L(x, \lambda) = x_1 + x_2 + x_3 - 3 = 0.$$

Nun lassen sich $x^* = (x_1^*, x_2^*, x_3^*) \in \mathbb{R}^3$ und $\lambda^* \in \mathbb{R}$ als Lösung des Gleichungs-Systems

$$\begin{pmatrix} 0 & 1 & 1 & 1 \\ 1 & 0 & 1 & 1 \\ 1 & 1 & 0 & 1 \\ 1 & 1 & 1 & 0 \end{pmatrix} \begin{pmatrix} x_1^* \\ x_2^* \\ x_3^* \\ \lambda^* \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ 0 \\ 3 \end{pmatrix}$$

berechnen. Wir erhalten $x_i^* = 1$ für $i = 1, 2, 3$ und $\lambda^* = -2$. Somit haben wir eine Lösung von (1.10) gefunden. Wir wissen aber nicht, ob es sich bei dem Punkt x^* um eine Minimum, Maximum oder einen Sattelpunkt handelt. ◊

4. Bedingungen 2. Ordnung für Gleichungs-Restriktionen

Seien sowohl das Zielfunktional J als auch die Gleichungs-Nebenbedingung e zweimal stetig differenzierbar.

SATZ 1.12 (Notwendige Bedingungen 2. Ordnung). *Sei x^* ein lokales Minimum von J unter der Nebenbedingung $e(x) = 0$. Ferner sei x^* ein regulärer Punkt. Dann ist die $n \times n$-Matrix*

$$\nabla_{xx} L(x^*, \lambda^*) = \nabla^2 J(x^*) + (\lambda^*)^T \nabla^2 e(x^*) = \nabla^2 J(x^*) + \sum_{i=1}^{m} \lambda_i^* \nabla^2 e_i(x^*)$$

positiv semi-definit auf $\operatorname{Kern} \nabla e(x^*)$*, das heißt,*

$$v^T \nabla_{xx} L(x^*, \lambda^*) v \geq 0 \quad \textit{für alle } v \in \operatorname{Kern} \nabla e(x^*).$$

wobei λ^* den nach Satz 1.9 eindeutig bestimmten Lagrange-Multiplikator zu x^* bezeichnet und ∇^2 für die zweite Ableitung steht.

BEWEIS. Nach Satz 1.4 ist die Tangentialebene an $\mathcal{E}$ im Punkt x^* durch den Unterraum $\operatorname{Kern} \nabla e(x^*)$ gegeben. Sei $x(t)$ eine zweimal stetig differenzierbare Kurve in $\mathcal{E}$ mit $x(0) = x^*$. Dann gilt

$$\frac{\mathrm{d}^2}{\mathrm{d}t^2} J(x(t))\big|_{t=0} \geq 0. \tag{1.11}$$

Wir berechnen die zweite Ableitung von J und bekommen

$$\begin{aligned} \frac{\mathrm{d}^2}{\mathrm{d}t^2} J(x(t))\big|_{t=0} &= \frac{\mathrm{d}}{\mathrm{d}t} \big(\nabla J(x(t)) \dot{x}(t)\big)\big|_{t=0} \\ &= \big(\dot{x}(t)^T \nabla^2 J(x(t)) \dot{x}(t) + \nabla J(x(t)) \ddot{x}(t)\big)\,\big|_{t=0} \\ &= \dot{x}(0)^T \nabla^2 J(x^*) \dot{x}(0) + \nabla J(x^*) \ddot{x}(0). \end{aligned} \tag{1.12}$$

Differenzieren wir die Gleichung $(\lambda^*)^T e(x(t)) = 0$ zweimal nach t und werten die zweite Ableitung an $t = 0$ aus, so folgt

$$\begin{aligned} 0 = \frac{\mathrm{d}^2}{\mathrm{d}t^2} \big((\lambda^*)^T e(x(t))\big)\big|_{t=0} &= \frac{\mathrm{d}}{\mathrm{d}t} \big((\lambda^*)^T \nabla e(x(t)) \dot{x}(t)\big)\big|_{t=0} \\ &= \big(\dot{x}(t)^T (\lambda^*)^T \nabla^2 e(x(t)) \dot{x}(t)\big)\,\big|_{t=0} \\ &\quad + \big((\lambda^*)^T \nabla e(x(t) \ddot{x}(t)\big)\,\big|_{t=0} \\ &= \dot{x}(0)^T (\lambda^*)^T \nabla^2 e(x^*) \dot{x}(0) + (\lambda^*)^T \nabla e(x^*) \ddot{x}(0). \end{aligned} \tag{1.13}$$

Wir addieren nun (1.13) zu (1.12) und verwenden (1.11) sowie (1.5). Dann ergibt sich die Ungleichung

$$\begin{aligned} 0 \leq \frac{\mathrm{d}^2}{\mathrm{d}t^2} J(x(t))\big|_{t=0} &= \dot{x}(0)^T \left(\nabla^2 J(x^*) + (\lambda^*)^T \nabla^2 e(x^*)\right) \dot{x}(0) \\ &\quad + \left(\nabla J(x^*) + (\lambda^*)^T \nabla e(x^*)\right) \ddot{x}(0) \\ &= \dot{x}(0)^T \nabla_{xx} L(x^*, \lambda^*) \dot{x}(0). \end{aligned}$$

Da $\dot{x}(t)$ ein beliebiger Tangentialvektor in $\operatorname{Kern} \nabla e(x^*)$ ist, erhalten wir die Behauptung. □

SATZ 1.13 (Hinreichende Bedingungen 2. Ordnung). *Seien $x^* \in \mathcal{E} \subset \mathbb{R}^n$ und $\lambda^* \in \mathbb{R}^m$ zwei Punkte mit*

$$\nabla J(x^*) + (\lambda^*)^T \nabla e(x^*) = 0. \tag{1.14}$$

Ferner seien x^* ein regulärer Punkt und die Matrix $\nabla_{xx} L(x^*, \lambda^*)$ positiv definit auf $\operatorname{Kern} \nabla e(x^*)$*:*

$$v^T \nabla_{xx} L(x^*, \lambda^*) v > 0 \quad \textit{für alle } v \in \operatorname{Kern} \nabla e(x^*) \setminus \{0\}.$$

Dann ist x^* ein striktes lokales Minimum von J unter der Nebenbedingung $e(x) = 0$, das heißt, x^* ist eine strikte lokale Lösung von $(\mathbf{P}_{Gl})$.

BEWEIS. Angenommen, x^* ist kein striktes lokales Minimum von J unter der Nebenbedingung $e(x) = 0$. Dann gibt es eine Folge $\{x^k\}_{k=0}^{\infty}$ mit

$$\lim_{k\to\infty} x^k = x^*, \quad J(x^k) \le J(x^*) \text{ und } e(x^k) = 0 \text{ für alle } k \ge 0.$$

Wir schreiben $x^k = x^* + s_k\, \delta x^k$, wobei $s_k > 0$ für alle k und $\delta x^k \in \mathbb{R}^n$ mit $\|\delta x^k\| = 1$ gelten. Damit folgt $s_k \to 0$ für $k \to \infty$. Ferner besitzt nach dem Satz von Bolzano-Weierstrass die Folge $\{\delta x^k\}_{k=0}^{\infty}$ eine konvergente Teilfolge, die wir wieder mit $\{\delta x^k\}_{k=0}^{\infty}$ bezeichnen. Sei $\delta x^* \in \mathbb{R}^n$ das Grenzelement der Teilfolge, das heißt, $\delta x^k \to \delta x^*$ für $k \to \infty$. Wegen $e(x^k) = e(x^*) = 0$ erhalten wir

$$0 = \lim_{k\to\infty} \frac{e(x^k) - e(x^*)}{s_k} = \lim_{k\to\infty} \frac{e(x^k) - e(x^*)}{\|s_k\, \delta x^k\|} = \nabla e(x^*) \delta x^*.$$

Damit liegt δx^* in $\operatorname{Kern} \nabla e(x^*)$.
Eine Taylorentwicklung der i-ten Nebenbedingung $e_i(x^k)$ am Entwicklungspunkt x^* ergibt wegen $e_i(x^*) = 0$

$$0 = e_i(x^k) = s_k \nabla e_i(x^*) \delta x^k + \frac{s_k^2}{2} (\delta x^k)^T \nabla^2 e_i(\eta_i^k) \delta x^k, \quad i = 1, \dots, m, \tag{1.15}$$

und für das Kostenfunktional $J(x^k)$ erhalten wir

$$0 \ge J(x^k) - J(x^*) = s_k \nabla J(x^*) \delta x^k + \frac{s_k^2}{2} (\delta x^k)^T \nabla^2 J(\eta_0^k) \delta x^k. \tag{1.16}$$

In (1.15) und (1.16) bezeichnen die η_i^k, $i = 0, \dots, m$, Zwischenstellen aus der Verbindungsstrecke $\{x \in \mathbb{R}^n \,|\, x = \tau x^* + (1-\tau) x^k \text{ mit } \tau \in [0,1]\}$. Multiplizieren wir (1.15) mit λ_i^*, addieren die m Gleichungen und fügen diese zu (1.16) hinzu, so erhalten wir die Ungleichung

$$0 \ge s_k \left(\nabla J(x^*) + (\lambda^*)^T \nabla e(x^*) \right) \delta x^k + \frac{s_k^2}{2} (\delta x^k)^T \left(\nabla^2 J(\eta_0^k) + \sum_{i=1}^{m} \lambda_i^* \nabla^2 e_i(\eta_i^k) \right) \delta x^k.$$

Aus (1.14) schließen wir, dass der erste Term auf der rechten Seite gleich null ist. Da die Zahlen s_k für alle k positiv sind, folgt

$$0 \ge (\delta x^k)^T \left(\nabla^2 J(\eta_0^k) + \sum_{i=1}^{m} \lambda_i^* \nabla^2 e_i(\eta_i^k) \right) \delta x^k$$

Wegen $\lim_{k\to\infty} \eta_i^k = x_i^*$, $0 \le i \le m$, bekommen wir beim Grenzübergang $k \to \infty$ die Ungleichung

$$0 \ge (\delta x^*)^T \left(\nabla^2 J(x^*) + \sum_{i=1}^{m} \lambda_i^* \nabla^2 e_i(x^*) \right) \delta x^* = (\delta x^*)^T \nabla_{xx} L(x^*, \lambda^*) \delta x^*.$$

Da $\delta x^* \in \operatorname{Kern} \nabla e(x^*)$ gilt, haben wir einen Widerspruch zu der Voraussetzung, dass $\nabla_{xx} L(x^*, \lambda^*)$ positiv definit auf $\operatorname{Kern} \nabla e(x^*)$ ist. Daher ist die Annahme falsch und x^* eine strikte lokale Lösung von $(\mathbf{P}_{Gl})$. □

BEISPIEL 1.14. Wir wenden uns nun noch einmal dem Beispiel 1.11 zu. Die Lösung der notwendigen Optimalitätsbedingungen sind $x^* = (1, 1, 1)$ und $\lambda^* = -2$. Dann folgt

$$\nabla_{xx} L(x^*, \lambda^*) = \begin{pmatrix} 0 & 1 & 1 \\ 1 & 0 & 1 \\ 1 & 1 & 0 \end{pmatrix}.$$

Die symmetrische Matrix $\nabla_{xx} L(x^*, \lambda^*)$ ist indefinit mit den drei Eigenwerten $\mu_1 = -1$, $\mu_2 = -1$ und $\mu_3 = 2$. Um die hinreichenden Bedingungen 2. Ordnung nachzuprüfen, wählen wir $v = (v_1, v_2, v_3) \in \operatorname{Kern} \nabla e(x^*) \setminus \{0\}$ beliebig. Dann folgen $v_1 + v_2 + v_3 = 0$ und

$$v^T \nabla_{xx} L(x^*, \lambda^*) v = v_1(v_2 + v_3) + v_2(v_1 + v_3) + v_3(v_1 + v_2) = -v_1^2 - v_2^2 - v_3^2 < 0.$$

Damit ist $\nabla_{xx} L(x^*, \lambda^*)$ negativ definit und an x^* liegt kein Minimum, sondern ein Maximum vor. ◊

5. Sensitivitätsanalyse

Sei $x^* \in \mathbb{R}^n$ ein regulärer Punkt und eine lokale Lösung von

$$\min J(x) \quad \text{u.d.N.} \quad e(x) = 0. \tag{1.17}$$

Ferner seien J und e zweimal stetig differenzierbar. Mit $\lambda^* \in \mathbb{R}^m$ bezeichnen wir den nach Satz 1.9 eindeutig bestimmten Lagrange-Multiplikator zu x^* Wir wollen nun das Problem

$$\min J(x) \quad \text{u.d.N.} \quad e(x) = c \tag{1.18}$$

mit $c \in \mathbb{R}^m$ lösen.

LEMMA 1.15. *Seien $Q \in \mathbb{R}^{n \times n}$ und $A \in \mathbb{R}^{m \times n}$ gegeben, wobei* $\operatorname{Rang} A = m$ *gilt und Q positiv definit auf dem Teilraum* $\operatorname{Kern} A$ *ist. Dann ist die Matrix*

$$\begin{pmatrix} Q & A^T \\ A & 0 \end{pmatrix}$$

invertierbar.

BEWEIS. Angenommen, das Paar $(x, \lambda) \in \mathbb{R}^n \times \mathbb{R}^m$ löst das System

$$Qx + A^T \lambda = 0, \tag{1.19a}$$

$$Ax \qquad\quad = 0. \tag{1.19b}$$

Zu zeigen ist, dass $x = 0$ und $\lambda = 0$ gelten muß. Wir multiplizieren (1.19a) mit x^T von links und erhalten die Gleichung

$$x^T Q x + x^T A^T \lambda = 0.$$

Mit (1.19b) folgen $x^T A^T = 0$ und daher $x^T Q x = 0$. Wegen (1.19b) haben wir aber $x \in \operatorname{Kern} A$, so dass $x = 0$ ist. Aus (1.19a) ergibt sich mit $x = 0$ nun $A^T \lambda = 0$. Nach Voraussetzung gilt $\operatorname{Rang} A = m$. Damit ist A surjektiv und deshalb A^T injektiv. Das bedeutet aber, dass $\lambda = 0$ sein muß. Wir haben damit gezeigt, dass $(x, \lambda) = (0, 0)$ erfüllt ist. □

Nun können wir folgenden Satz beweisen.

SATZ 1.16. *Sei $x^* \in \mathbb{R}^n$ eine lokale Lösung von (1.17) und sei x^* ein regulärer Punkt. Der Vektor $\lambda^* \in \mathbb{R}^m$ bezeichne den Lagrange-Multiplikator zu x^*. Es gelte*

$$v^T \nabla_{xx} L(x^*, \lambda^*) v > 0 \quad \text{für alle } v \in \operatorname{Kern} \nabla e(x^*) \setminus \{0\}. \tag{1.20}$$

Dann gibt es eine Umgebung $U(0) \subset \mathbb{R}^m$ von $0 \in \mathbb{R}^m$, so dass (1.18) eine Lösung $x(c)$ für alle $c \in U(0)$ besitzt, die stetig von c abhängt mit $x(0) = x^*$. Ferner gilt

$$\nabla_c J(x(c))\big|_{c=0} = -(\lambda^*)^T.$$

BEWEIS. Wir betrachten das Gleichungs-System

$$\nabla J(x) + \lambda^T \nabla e(x) = 0, \tag{1.21a}$$

$$e(x) = c. \tag{1.21b}$$

Nach Voraussetzung löst (x^*, λ^*) das System (1.21) für $c = 0$. Die Jacobi-Matrix der Abbildung

$$F(x, \lambda, c) = \begin{pmatrix} \nabla J(x)^T + \nabla e(x)^T \lambda \\ e(x) - c \end{pmatrix}$$

am Punkt $(x^*, \lambda^*, 0) \in \mathbb{R}^n \times \mathbb{R}^m \times \mathbb{R}^m$ bezüglich der Variablen (x, λ) ist

$$\nabla_{(x,\lambda)} F(x^*, \lambda^*, 0) = \begin{pmatrix} \nabla_{xx} L(x^*, \lambda^*) & \nabla e(x^*)^T \\ \nabla e(x^*) & 0 \end{pmatrix} = \nabla^2 L(x^*, \lambda^*).$$

Wegen (1.20) und der Tatsache, dass x^* ein regulärer Punkt ist, ist $\nabla^2 L(x^*, \lambda^*)$ nach Lemma 1.15 invertierbar. Nach dem Satz über Implizite Funktionen gibt es eine Lösung $(x(c), \lambda(c))$ von (1.21) in einer Umgebung $U(0) \subset \mathbb{R}^m$ von $0 \in \mathbb{R}^m$. Diese Lösung hängt sogar stetig differenzierbar von c ab. Da $x \mapsto \nabla e(x)$ stetig ist, ist $\nabla e(x(c))$ surjektiv für alle c in einer eventuell kleineren Umgebung $\tilde{U}(0) \subseteq U(0)$ von 0. Da $\nabla_{xx} L(x, \lambda)$ stetig ist, lässt sich auch zeigen, dass eine Umgebung $\hat{U}(0) \subset \tilde{U}(0)$ existiert mit

$$v^T \nabla_{xx} L(x(c), \lambda(c)) v > 0 \quad \text{für alle } v \in \operatorname{Kern} \nabla e(x(c)) \setminus \{0\}$$

für alle $c \in \hat{U}(c)$, siehe [**35**, Lemma 2.12]. Damit erfüllen die Punkte $(x(c), \lambda(c))$, $c \in \hat{U}(0)$, die hinreichenden Bedingungen 2. Ordnung. Also ist $x(c)$ eine strikte lokale Lösung von (1.18).
Mit der Kettenregel erhalten wir

$$\nabla_c J(x(c))\big|_{c=0} = \nabla J(x^*) \nabla_c x(0)$$

und

$$\nabla_c e(x(c))\big|_{c=0} = \nabla e(x^*) \nabla_c x(0).$$

Wegen (1.21b) folgen $\nabla e(x^*) \nabla_c x(0) = I \in \mathbb{R}^{m \times m}$ und mit (1.21a)

$$\nabla_c J(x(c))\big|_{c=0} = -(\lambda^*)^T,$$

was zu zeigen war. □

6. Ungleichungs-Nebenbedingungen

Seien J und e zweimal stetig differenzierbar. Wir betrachten das Optimierungs-Problem

$$\min J(x) \quad \text{u.d.N.} \quad e(x) = 0 \text{ und } g(x) \leq 0. \tag{P}$$

DEFINITION 1.17. *Sei $x^* \in \mathbb{R}^n$ ein Punkt mit*

$$e(x^*) = 0 \quad \textit{und} \quad g(x^*) \leq 0. \tag{1.22}$$

Mit $\mathcal{A} \subseteq \{1, \ldots, p\}$ bezeichnen wir die Menge der aktiven Indizes *$i \in \mathcal{A}$, für die $g_i(x^*) = 0$ gilt. Der Punkt x^* heißt* regulärer Punkt für (1.22), *wenn $\nabla e_i(x^*)$, $1 \leq i \leq m$, und $\nabla g_i(x^*)$, $i \in \mathcal{A}$, linear unabhängig sind.*

SATZ 1.18 (Karush-Kuhn-Tucker). *Sei $x^* \in \mathbb{R}^n$ ein lokales Minimum von* (P) *und sei x^* ein regulärer Punkt für* (1.22)*. Dann existieren Vektoren $\lambda^* \in \mathbb{R}^m$ und $\mu^* \in \mathbb{R}^p$ mit $\mu^* \geq 0$, so dass*

$$\nabla J(x^*) + (\lambda^*)^T \nabla e(x^*) + (\mu^*)^T \nabla g(x^*) = 0, \tag{1.23a}$$

$$(\mu^*)^T g(x^*) = 0. \tag{1.23b}$$

BEMERKUNG 1.19. Die Gleichung (1.23b) wird *Komplementaritäts-Bedingung* genannt. Aus $\mu^* \geq 0$ und $g(x^*) \leq 0$ folgen, dass (1.23b) äquivalent mit der Tatsache ist, dass $\mu^* > 0$ nur dann gelten kann, wenn $i \in \mathcal{A}$ erfüllt ist. ◊

BEWEIS VON SATZ 1.18. Nach Voraussetzung ist x^* eine lokale Lösung von (**P**). Dann existiert eine Umgebung $U \subset \mathbb{R}^n$ von x^*, so dass $g_i(x) < 0$ für alle $i \in \{1, \ldots, p\} \setminus \mathcal{A}$ und alle $x \in U$ gilt. Damit löst x^* auch das Problem

$$\min J(x) \quad \text{u.d.N.} \quad e(x) = 0,\ g_i(x) = 0 \text{ für } i \in \mathcal{A}. \tag{1.24}$$

Sei $\mathcal{A} = \{i_1, \ldots, i_\ell\}$ mit $i_j \in \{1, \ldots, p\}$, $i_1 < \ldots < i_\ell$ und $\ell \leq p$. Da x^* ein regulärer Punkt ist, garantiert Satz 1.9 die Existenz eines Multiplikators $\xi^* \in \mathbb{R}^{m+\ell}$, so dass

$$\nabla J(x^*) + (\xi^*)^T \begin{pmatrix} \nabla e(x^*) \\ \nabla g_{i_1}(x^*) \\ \vdots \\ \nabla g_{i_\ell}(x^*) \end{pmatrix} = 0.$$

Wir setzen $\lambda_j^* = \xi_j^*$ für $j = 1, \ldots, m$ und $\lambda^* = (\lambda_1^*, \ldots, \lambda_m^*)^T \in \mathbb{R}^m$. Ferner definieren wir für $j = 1, \ldots, p$

$$\mu_j^* = \begin{cases} \xi_{m+k}^* & \text{falls } j = i_k \text{ für ein } k \in \{1, \ldots, \ell\}, \\ 0 & \text{sonst.} \end{cases}$$

und $\mu^* = (\mu_1^*, \ldots, \mu_p^*)^T$. Offenbar gelten dann sowohl (1.23a) als auch (1.23b). Zu zeigen bleibt noch, dass $\mu^* \geq 0$ erfüllt ist. Wegen $\mu_i^* = 0$ für $i \notin \mathcal{A}$ brauchen wir nur $\mu_i^* \geq 0$ für $i \in \mathcal{A}$ zu zeigen. Angenommen, es gibt ein $i_k \in \mathcal{A}$ mit $\mu_{i_k}^* < 0$. Wir setzen

$$\mathcal{S} = \big\{x \in \mathbb{R}^n \,|\, e(x) = 0,\ g_j(x) = 0 \text{ für } j \in \mathcal{A} \setminus \{i_k\}\big\}.$$

Da x^* ein regulärer Punkt für (1.22) ist, sind $\nabla e_1(x^*), \ldots, \nabla e_m(x^*)$, $\nabla g_{i_1}(x^*), \ldots,$ $\nabla g_{i_\ell}(x^*)$ linear unabhängig. Insbesondere gilt $\nabla g_{i_k}(x^*) \neq 0$ und x^* ist auch ein regulärer Punkt für $\mathcal{S}$, so dass wir die Tangentialebene an $\mathcal{S}$ im Punkt x^* durch den Kern der Matrix mit den Zeilen $\nabla e_i(x^*)$, $i \in \{1, \ldots, m\}$, und $\nabla g_i(x^*)$, $i \in \mathcal{A} \setminus \{i_k\}$

dartellen können. Da die Menge

$$\text{Kern}\begin{pmatrix} \nabla e(x^*) \\ \nabla g_{i_1}(x^*) \\ \vdots \\ \nabla g_{i_{k-1}}(x^*) \\ \nabla g_{i_{k+1}}(x^*) \\ \vdots \\ \nabla g_{i_\ell}(x^*) \end{pmatrix} = \left\{v \in \mathbb{R}^n \mid \nabla e(x^*)v = 0,\, \nabla g_i(x^*)v = 0 \text{ für } i \in \mathcal{A} \setminus \{i_k\}\right\}$$

größer als die Menge

$$\text{Kern}\begin{pmatrix} \nabla e(x^*) \\ \nabla g_{i_1}(x^*) \\ \vdots \\ \nabla g_{i_\ell}(x^*) \end{pmatrix} = \left\{v \in \mathbb{R}^n \mid \nabla e(x^*)v = 0,\, \nabla g_i(x^*)v = 0 \text{ für } i \in \mathcal{A}\right\}$$

ist, gibt es ein Element $v \in \mathbb{R}^n$ mit $\nabla e(x^*)v = 0$, $\nabla g_i(x^*)v = 0$ für $i \in \mathcal{A} \setminus \{i_k\}$ und $\nabla g_{i_k}(x^*)v < 0$. Der Vektor v liegt in der Tangentialebene an $\mathcal{S}$ im Punkt x^*. Es gibt daher eine Kurve $x(t)$ in $\mathcal{S}$ mit $x(0) = x^*$ und $\dot{x}(0) = v$. Damit gelten $e(x(t)) = 0$ und $g_i(x(t)) = 0$ für alle $i \in \mathcal{A} \setminus \{i_k\}$. Aus $g_i(x(0)) = g_i(x^*) < 0$ für alle $i \notin \mathcal{A}$ folgt die Existenz von $\varepsilon > 0$ mit $g_i(x(t)) < 0$ für alle $t \in (-\varepsilon, \varepsilon)$. Wegen

$$\frac{\mathrm{d}}{\mathrm{d}t}\, g_{i_k}(x(t))\,\big|_{t=0} = \nabla g_{i_k}(x^*)v < 0$$

und $g_{i_k}(x^*) = 0$ gibt es ein $\epsilon \geq \varepsilon > 0$, so dass $g_{i_k}(x(t)) < 0$ für alle $t \in [0, \epsilon)$. Deshalb schließen wir, dass die Kurve $x(t)$ für alle $t \in [0, \epsilon)$ zulässig für (**P**) ist. Wir betrachten die Ableitung des Zielfunktionals entlang der Kurve $x(t)$ an $t = 0$ und verwenden (1.23a). Es ergibt sich

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}t}\, J(x(t))\,\big|_{t=0} &= \nabla J(x^*)v = -(\lambda^*)^T \nabla e(x^*)v - (\mu^*)^T \nabla g(x^*)v \\ &= -\sum_{i \in \mathcal{A}\setminus\{i_k\}} \mu_i^* \nabla g_i(x^*)v - \mu_{i_k}^* \nabla g_{i_k}(x^*)v - \sum_{i \notin \mathcal{A}} \mu_i^* \nabla g_i(x^*)v < 0,\end{aligned}$$

da $\nabla e(x^*)v = 0$, $\nabla g_i(x^*)v = 0$ für alle $i \in \mathcal{A} \setminus \{i_k\}$, $\mu_{i_k}^* < 0$, $\nabla g_{i_k}(x^*)v < 0$ und $\mu_i^* = 0$ für alle $i \notin \mathcal{A}$. Damit fällt die Zielfunktion an x^* in Richtung v und x^* kann keine lokale Lösung von (**P**) sein. Das ergibt einen Widerspruch. Damit kann es kein $i_k \in \mathcal{A}$ mit $\mu_{i_k}^* < 0$ geben. □

Bei Ungleichungs-Restriktionen führen wir eine gegenüber (1.9) erweiterte Lagrange-Funktion wie folgt ein:

$$L(x, \lambda, \mu) = J(x) + (\lambda)^T e(x) + (\mu)^T g(x) \quad \text{für } (x, \lambda, \mu) \in \mathbb{R}^n \times \mathbb{R}^m \times \mathbb{R}^p.$$

Offenbar lässt sich (1.23a) in der Form $\nabla_x L(x^*, \lambda^*, \mu^*)$ schreiben.

Beispiel 1.20. Wir betrachten das folgende Beispiel:

$$\min \frac{1}{2}\,(x_1^2 + x_2^2 + x_3^2) \quad \text{u.d.N.} \quad x_1 + x_2 + x_3 \leq -3.$$

Aus (1.23a) erhalten wir die drei Gleichungen

$$x_1^* + \mu^* = 0, \quad x_2^* + \mu^* = 0, \quad x_3^* + \mu^* = 0$$

für ein lokales Minimum $x^* = (x_1^*, x_2^*, x_3^*)$. Es gibt zwei Möglichkeiten:

1) Die Nebenbedingung ist an x^* inaktiv, d.h.,
$$x_1^* + x_2^* + x_3^* < -3.$$
Wegen (1.23b) folgt sofort $\mu^* = 0$. Also erhalten wir direkt $x^* = (0, 0, 0)$, was aber der Ungleichungs-Nebenbedingung widerspricht. Also entfällt diese Möglichkeit.
2) Die Nebenbedingung ist aktiv. Dann haben wir vier Gleichungen zur Verfügung, um die Variablen $x_1^*, x_2^*, x_3^*, \mu^*$ zu berechnen. Wir bekommen das lineare Gleichungs-System
$$\left(\begin{array}{ccc|c} 1 & 0 & 0 & 1 \\ 0 & 1 & 0 & 1 \\ 0 & 0 & 1 & 1 \\ \hline 1 & 1 & 1 & 0 \end{array}\right) \left(\begin{array}{c} x_1^* \\ x_2^* \\ x_3^* \\ \mu^* \end{array}\right) = \left(\begin{array}{c} 0 \\ 0 \\ 0 \\ -3 \end{array}\right)$$
Die Koeffizienten-Matrix ist nach Lemma 1.15 invertierbar. Wir berechnen die Lösung $x_i^* = -1$, $1 \le i \le 3$, und $\mu^* = 1 \ge 0$. Es gibt damit nur einen Kandidaten x^* für ein lokales Minimum. ◊

Notwendige und hinreichende Bedingungen zweiter Ordung für Probleme mit Ungleichungs-Restriktionen werden im wesentlichen dadurch hergeleitet, dass nur die aktiven Nebenbedingungen betrachtet werden.

Satz 1.21. *Seien J, e, g zweimal stetig differenzierbar und x^* ein regulärer Punkt von* (1.22). *Ferner nehmen wir an, dass x^* ein relatives Minimum für* (**P**) *ist. Dann existieren Lagrange-Multiplikatoren $\lambda^* \in \mathbb{R}^m$ und $\mu^* \in \mathbb{R}^p$ mit $\mu^* \ge 0$, so dass* (1.23) *gilt und die Hesse-Matrix*
$$\begin{aligned} \nabla_{xx} L(x^*, \lambda^*, \mu^*) &= \nabla^2 J(x^*) + (\lambda^*)^T \nabla^2 e(x^*) + (\mu^*)^T \nabla^2 g(x^*) \\ &= \nabla^2 J(x^*) + \sum_{i=1}^{m} \lambda_i^* \nabla^2 e_i(x^*) + \sum_{i=1}^{p} \mu_i^* \nabla^2 g_i(x^*) \end{aligned}$$
positiv semi-definit ist auf dem Tangentialraum zu den aktiven Nebenbedingungen.

Beweis. Wenn x^* ein lokales Minimum bezüglich der Nebenbedingung (1.22) ist und die Menge der aktiven Indizes mit $\mathcal{A} = \{i_1, \ldots, i_\ell\}$ bezeichnet wird, ist es auch eines für das Problem
$$\min J(x) \quad \text{u.d.N.} \quad e(x) = 0,\ g_{i_1}(x) = \ldots = g_{i_\ell}(x) = 0$$
(siehe auch im Beweis von Satz 1.18). Daher folgt die Aussage aus Satz 1.12. □

Um hinreichende Kriterien herzuleiten, müssen wir den Fall berücksichtigen, dass die zu aktiven Ungleichungen assozierten Lagrange-Multiplikatoren den Wert Null haben können. Daher muß $\nabla_{xx} L(x^*, \lambda^*, \mu^*)$ positiv definit auf einem größerem Unterraum sein.

Satz 1.22. *Seien J, e, g zweimal stetig differenzierbar. Hinreichende Bedingung, dass $x^* \in \mathbb{R}^n$ ein striktes lokales Minimum von* (**P**) *ist, ist die Existenz von Vektoren $\lambda^* \in \mathbb{R}^m$ und $\mu^* \in \mathbb{R}^p$, so dass*

(1.25a) $$\mu^* \ge 0,$$

(1.25b) $$(\mu^*)^T g(x^*) = 0,$$

(1.25c) $$\nabla J(x^*) + (\lambda^*)^T \nabla e(x^*) + (\mu^*)^T \nabla g(x^*) = 0$$

gelten und die Hesse-Matrix

$$\nabla_{xx} L(x^*, \lambda^*, \mu^*) = \nabla^2 J(x^*) + (\lambda^*)^T \nabla^2 e(x^*) + (\mu^*)^T \nabla^2 g(x^*)$$

positiv definit auf dem Unterraum

$$\tilde{\mathcal{K}} = \left\{ v \in \mathbb{R}^n \,|\, \nabla e(x^*)v = 0,\ \nabla g_i(x^*)v = 0 \text{ für } i \in \tilde{\mathcal{A}} \right\}$$

mit $\tilde{\mathcal{A}} = \{ i \in \mathcal{A} \,|\, \mu_i^* > 0 \}$.

Bemerkung 1.23. Wegen $\tilde{\mathcal{A}} \subset \mathcal{A}$ folgt

$$\mathcal{K} = \left\{ v \in \mathbb{R}^n \,|\, \nabla e(x^*)v = 0,\ \nabla g_i(x^*)v = 0 \text{ für } i \in \mathcal{A} \right\} \subset \tilde{\mathcal{K}},$$

das heißt, die Menge $\tilde{\mathcal{K}}$ ist im allgemeinen größer als die Menge $\mathcal{K}$. ◊

Beweis von Satz 1.22. Wir nehmen wie im Beweis von Satz 1.13 an, dass x^* kein striktes lokales Minimum ist. Sei $\{x^k\}_{k=0}^{\infty}$ eine Folge von zulässigen Punkten mit

$$\lim_{k\to\infty} x^k = x^* \quad \text{und} \quad J(x^k) \le J(x^*) \text{ für alle } k \in \mathbb{N}.$$

Wir schreiben wieder $x^k = x^* + s_k \delta x^k$ mit $s_k > 0$ und $\|\delta x^k\| = 1$ für alle k. Eventuell nach Übergang zu einer Teilfolge setzen wir voraus, dass ein Element $\delta x^* \in \mathbb{R}^n$ existiert mit

$$\lim_{k\to\infty} s_k = 0 \quad \text{und} \quad \lim_{k\to\infty} \delta x^k = \delta x^*.$$

Wegen

$$\frac{J(x^k) - J(x^*)}{s_k \,\|\delta x^k\|} \le 0$$

erhalten wir $\nabla J(x^*)\delta x^* \le 0$. Ferner schließen wir aus $e_i(x^*) = e_i(x^k) = 0$ für $1 \le i \le m$, dass

$$\nabla e_i(x^*)\delta x^* = \lim_{k\to\infty} \frac{e_i(x^k) - e_i(x^*)}{s_k \|\delta x^k\|} = 0, \quad 1 \le i \le m,$$

erfüllt ist. Damit gilt $\delta x^* \in \operatorname{Kern} \nabla e(x^*)$. Analog ergibt sich für die aktiven Nebenbedingungen

$$\nabla g_i(x^*)\delta x^* = \lim_{k\to\infty} \frac{g_i(x^k) - g_i(x^*)}{s_k \,\|\delta x^k\|} \le 0, \quad i \in \mathcal{A}.$$

Im Fall von $\nabla g_i(x^*)\delta x^* = 0$ beziehungsweise $\mu_i^* = 0$, $i \in \mathcal{A}$, das heißt, $i \in \mathcal{A} \setminus \tilde{\mathcal{A}}$, verläuft der Beweis genauso wie in dem von Satz 1.13. Hier kann ich die positive Definitheit von $\nabla_{xx} L(x^*, \lambda^*, \mu^*)$ auf $\mathcal{K}$ verwenden. Gilt $\nabla g_i(x^*)\delta x^* < 0$ für mindestens ein $i \in \tilde{\mathcal{A}}$. Dann bekommen wir mit (1.25c), und $\mu_i^* = 0$ für alle $i \notin \mathcal{A} \setminus \tilde{\mathcal{A}}$,

$$\begin{aligned} 0 \ge \nabla J(x^*)\delta x^* &= -(\lambda^*)^T \nabla e(x^*)\delta x^* - (\mu^*)^T \nabla g(x^*)\delta x^* \\ &= -\sum_{i \notin \mathcal{A}} \mu_i^* \nabla g_i(x^*)\delta x^* - \sum_{i \in \tilde{\mathcal{A}}} \mu_i^* \nabla g_i(x^*)\delta x^* - \sum_{i \in \mathcal{A}\setminus\tilde{\mathcal{A}}} \mu_i^* \nabla g_i(x^*)\delta x^* \\ &= -\sum_{i \in \tilde{\mathcal{A}}} \underbrace{\mu_i^*}_{>0} \underbrace{\nabla g_i(x^*)\delta x^*}_{\le 0} > 0, \end{aligned}$$

da mindestens ein $i \in \tilde{\mathcal{A}}$ existiert mit $\nabla g_i(x^*)\delta x^* < 0$. Damit haben wir aber einen Widerspruch. □

BEMERKUNG 1.24. Gilt $\tilde{\mathcal{A}} = \mathcal{A}$, so ist die Voraussetzung, dass $\nabla_{xx} L(x^*, \lambda^*, \mu^*)$ positiv definit auf

$$\left\{ v \in \mathbb{R}^n \mid \nabla e(x^*) v = 0,\ \nabla g_i(x^*) v = 0 \text{ für } i \in \mathcal{A} \right\}$$

ist, eine hinreichende Bedingung für ein striktes lokales Minimum an x^*. ◊

BEISPIEL 1.25. Wir setzen das Beispiel 1.20 fort und berechnen

$$\nabla^2 J(x^*) = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} \quad \text{und} \quad \nabla^2 g(x^*) = \begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix}.$$

Also ist $\nabla_{xx} L(x^*, \lambda^*, \mu^*)$ gleich der Identität in $\mathbb{R}^{3\times 3}$. Da die Nebenbedingung aktiv ist und $\mu^* = 1 > 0$ gilt, ist eine hinreichende Bedingung für ein striktes lokales Minimum an x^* nach Satz 1.22, dass $\nabla_{xx} L(x^*, \lambda^*, \mu^*)$ positiv definit auf dem Unterraum $\operatorname{Kern} \nabla g(x^*)$ ist. Offenbar ist aber $\nabla_{xx} L(x^*, \lambda^*, \mu^*)$ positiv auf $\mathbb{R}^3$, und damit liegt an x^* ein striktes lokales Minimum vor. ◊

Wir geben noch ein Sensitivitätsresultat an. Ein Beweis basiert auf ähnlichen Argumenten wie denen im Beweis von Satz 1.16.

SATZ 1.26. *Seien J, e, g zweimal stetig differenzierbar. Für $(c, d) \in \mathbb{R}^m \times \mathbb{R}^p$ betrachten wir die Familie von Problemen*

$$\min J(x) \quad \text{u.d.N.} \quad e(x) = c \text{ und } g(x) \leq d. \tag{1.26}$$

Der Punkt $x^* \in \mathbb{R}^n$ sei regulär und eine lokale Lösung von (1.26) für $(c, d) = (0, 0)$. Ferner erfülle x^* zusammen mit den nach Satz 1.18 zugehörigen Lagrange-Multiplikatoren $\lambda^* \in \mathbb{R}^m$ und $\mu^* \in \mathbb{R}^p$ mit $\mu^* \geq 0$ die hinreichenden Bedingungen 2. Ordnung für ein striktes lokales Minimum (siehe Satz 1.22). Ferner gelte $\mu_i^* > 0$ für alle aktiven Ungleichungs-Restriktionen. Dann existiert eine Umgebung $U \subset \mathbb{R}^{m+p}$ von $(0, 0)$, so dass (1.26) eine lokale Lösung $x = x(c, d)$ zu jedem $(c, d) \in U$ besitzt. Diese Lösung $x(c, d)$ hängt stetig von (c, d) ab mit $x(0, 0) = x^*$. Ferner gelten

$$\nabla_c J(x(c,d))\big|_{(c,d)=(0,0)} = -(\lambda^*)^T \quad \text{und} \quad \nabla_d J(x(c,d))\big|_{(c,d)=(0,0)} = -(\mu^*)^T.$$

KAPITEL 2

Lineare Programmierung: Innere-Punkte Verfahren

In diesem Abschnitt werden wir uns mit Innere Punkte Verfahren zur Behandlung von lineare Ungleichungs-Nebenbedingungen beschäftigen. Dabei werden wir *Innere-Punkte Verfahren* verwenden.

1. Primal-Duale Verfahren

Wir betrachten das Problem

$$\min c^T x \quad \text{u.d.N.} \quad Ax = b \text{ und } x \geq 0 \tag{2.1}$$

mit $c \in \mathbb{R}^n$, $A \in \mathbb{R}^{m \times n}$ sowie $b \in \mathbb{R}^m$. Da sowohl die Zielfunktion sowie die Nebenbedingungen linear sind, wird (2.1) auch als Problem der *Linearen Programmierung* bezeichnet.

Im Kontext von Abschnitt 1 setzen wir

$$J(x) = c^T x,\ e(x) = b - Ax,\ g(x) = -x \text{ und } p = n.$$

Dann erhalten wir $\nabla J(x) = c^T$, $\nabla e(x) = -A$ sowie $\nabla g(x) = -I$. Sei $x^* \in \mathbb{R}^n$ eine lokale Lösung von (2.1). Da alle Nebenbedingungen, d.h., e und g, linear sind, existieren Lagrange-Multiplikatoren $\lambda^* \in \mathbb{R}^m$ und $\mu^* \in \mathbb{R}^n$ mit $\mu^* \geq 0$, so dass die notwendigen Optimalitäts-Bedingungen erster Ordnung erfüllt sind:

$$\nabla J(x^*) + (\lambda^*)^T \nabla e(x^*) + (\mu^*)^T \nabla g(x^*) = 0. \tag{2.2}$$

Einen Beweis finden wir z.B. in [**29**, S. 351-353]. Für unser Beispiel lautet (2.2):

$$A^T \lambda^* + \mu^* = c. \tag{2.3a}$$

Ferner erfüllt x^* die Gleichungs-Restriktionen:

$$Ax^* = b. \tag{2.3b}$$

Aus der Komplementaritäts-Bedingung $(\mu^*)^T g(x^*) = 0$ folgt

$$\sum_{i=1}^{n} \mu_i^* g_i(x^*) = 0.$$

Nun gelten $\mu_i^* \geq 0$ und $g_i(x^*) \leq 0$ für alle $i = 1, \ldots, n$. Daher können wir die Komplementaritäts-Bedingung auch in der Form

$$\mu_i^* g_i(x^*) = 0 \quad \text{für } i = 1, \ldots, n \tag{2.3c}$$

schreiben. Schließlich sind x^* und μ^* nicht-negativ. Wir drücken das wie folgt aus:

$$(x^*, \mu^*) \geq 0. \tag{2.3d}$$

Primal-Duale Verfahren bestimmen eine Lösung (x^*, λ^*, μ^*) von (2.3) durch Anwendung des Newton-Algothmus auf (2.3a)-(2.3c), wobei die Suchrichtung so modifiziert wird, dass (2.3d) in jeder Iteration strikt erfüllt ist. Daher werden diese Methoden auch *Innere-Punkte Verfahren* genannt.

Die Bedingung (2.3d) macht das Problem (2.3) deutlich schwieriger und ist Ursache für Entwicklung von unterschiedlichen Varianten der Inneren-Punkte Verfahren. Wir führen die Abbildung $F : \mathbb{R}^n \times \mathbb{R}^m \times \mathbb{R}^n \to \mathbb{R}^n \times \mathbb{R}^m \times \mathbb{R}^n$ durch

$$F(x, \lambda, \mu) = \begin{pmatrix} A^T\lambda + \mu - c \\ Ax - b \\ XMe \end{pmatrix}, \quad (x, \lambda, \mu) \in \mathbb{R}^n \times \mathbb{R}^m \times \mathbb{R}^n$$

ein, wobei

$$X = \operatorname{diag}(x_1, \ldots, x_n) \in \mathbb{R}^{n\times n},\ M = \operatorname{diag}(\mu_1, \ldots, \mu_n) \in \mathbb{R}^{n\times n},\ e = \begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix} \in \mathbb{R}^n$$

gesetzt worden sind. Dann lässt sich das Problem (2.3) in der Form

$$F(x, \lambda, \mu) = 0, \tag{2.4a}$$

$$(x, \mu) \geq 0 \tag{2.4b}$$

schreiben. Wie wir bereits oben erwähnt haben, generieren Primal-Duale Verfahren Iterierte (x^k, λ^k, μ^k), die (2.4b) strikt erfüllen, das heißt, es gelten $x^k > 0$ und $\mu^k > 0$ für alle k. Damit werden insbesondere keine Iterierten erzeugt, die (2.4b) nicht erfüllen.

Zulässige Innere-Punkte Verfahren fordern, dass sowohl (2.3a) als auch (2.3b) für alle Iterationen gelten. Wir führen aus diesem Grund zwei Mengen ein:

$$\mathcal{F} = \big\{(x, \lambda, \mu) \in \mathbb{R}^n \times \mathbb{R}^m \times \mathbb{R}^n \mid A^T\lambda + \mu = c,\ Ax = b,\ (x, \mu) \geq 0\big\},$$

$$\mathcal{F}^\circ = \big\{(x, \lambda, \mu) \in \mathbb{R}^n \times \mathbb{R}^m \times \mathbb{R}^n \mid A^T\lambda + \mu = c,\ Ax = b,\ (x, \mu) > 0\big\}$$

und bezeichnen $\mathcal{F}$ als *zulässige Menge* und $\mathcal{F}^\circ$ als *strikt zulässige Menge*. Die Forderung, dass die Iterierte (x^k, λ^k, μ^k) strikt zulässig ist, lässt sich wie folgt schreiben:

$$(x^k, \lambda^k, \mu^k) \in \mathcal{F}^\circ.$$

Wir wollen nun einen Newton-Schritt für das nicht-lineare System (2.4a) betrachten. Sei die Iterierte (x^k, λ^k, μ^k), $k \geq 0$, gegeben. Dann berechnen wir $(\Delta x^k, \Delta\lambda^k, \Delta\mu^k)$ als Lösung des linearen Gleichungs-Systems

$$\nabla F(x^k, \lambda^k, \mu^k) \begin{pmatrix} \Delta x^k \\ \Delta\lambda^k \\ \Delta\mu^k \end{pmatrix} = -F(x^k, \lambda^k, \mu^k)$$

mit der nichtsymmetrischen Funktional-Matrix

$$\nabla F(x, \lambda, \mu) = \begin{pmatrix} 0 & A^T & I \\ A & 0 & 0 \\ M & 0 & X \end{pmatrix}.$$

Wir bemerken an dieser Stelle, dass die Funktionalmatrix durch die Diagonalmatrizen M und X von den Argumenten x sowie μ, aber nicht von λ abhängen. Die neue Iterierte ist nun gegeben durch

$$(x^{k+1}, \lambda^{k+1}, \mu^{k+1}) = (x^k, \lambda^k, \mu^k) + (\Delta x^k, \Delta\lambda^k, \Delta\mu^k).$$

Im Allgemeinen werden wir keinen vollen Newton-Schritt ausführen können, ohne (2.4b) zu verletzen. Daher bestimmen wir einen Schrittweiten-Parameter $\alpha_k \in (0,1]$, so dass

$$(x^{k+1}, \lambda^{k+1}, \mu^{k+1}) = (x^k, \lambda^k, \mu^k) + \alpha_k(\Delta x^k, \Delta\lambda^k, \Delta\mu^k)$$

die Bedingung (2.4b) erfüllt. Oft muss aber dann der Parameter α_k sehr klein gewählt werden, um (2.4b) für die nächste Iterierte zu garantieren.

Primal-Duale Verfahren modifizieren daher wie folgt:

1) Sie "zwingen" die Suchrichtung in das Innere von $\mathcal{F}^\circ$, so dass wir ein größeres α_k wählen können, ohne (2.4b) zu verletzen.
2) Sie verhindern, dass die Komponenten von x und μ "zu nahe" an die Null kommen.

Wir führen den *zentralen Pfad* $\mathcal{C}$ ein, eine durch $\tau > 0$ parametrisierte Kurve in $\mathbb{R}^n \times \mathbb{R}^m \times \mathbb{R}^n$, wobei die Punkte $(x^\tau, \lambda^\tau, \mu^\tau) \in \mathcal{C}$ Lösungen des folgenden nichtlinearen Gleichungs-Systems sind:

$$A^T\lambda + \mu = c, \tag{2.5a}$$

$$Ax = b, \tag{2.5b}$$

$$x_i\mu_i = \tau, \quad i \in \{1, \dots, n\}, \tag{2.5c}$$

$$(x, \mu) > 0. \tag{2.5d}$$

In (2.5d) bedeutet $(x, \mu) > 0$, dass sowohl $x > 0$ als auch $\mu > 0$ gelten. Anstatt von (2.3c) verlangen wir, dass alle Produkte von x_i^τ und μ_i^τ gleich $\tau > 0$ sind. Der zentrale Pfad ist daher die Menge

$$\mathcal{C} = \big\{(x^\tau, \lambda^\tau, \mu^\tau) \in \mathbb{R}^n \times \mathbb{R}^m \times \mathbb{R}^n \,|\, (x^\tau, \lambda^\tau, \mu^\tau) \text{ löst (2.5) für ein } \tau > 0\big\}.$$

Es kann gezeigt werden, dass für jedes $\tau > 0$ genau eine Lösung $(x^\tau, \lambda^\tau, \mu^\tau)$ existiert, wenn $\mathcal{F}^\circ \neq \emptyset$ gilt. Wir können (2.5) in der Form

$$\tilde{F}(x^\tau, \lambda^\tau, \mu^\tau) = F(x^\tau, \lambda^\tau, \mu^\tau) - \begin{pmatrix} 0 \\ 0 \\ \tau e \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix} \tag{2.6}$$

schreiben. Damit approximiert (2.5) das System (2.3) zunehmend besser für $\tau \to 0$. Wenn für eine Folge $\{\tau_n\}_{n=0}^\infty$ mit $\tau_n > 0$ für alle n und $\lim_{n\to\infty} \tau_n = 0$ die Folge $\{(x^{\tau_n}, \lambda^{\tau_n}, \mu^{\tau_n})\}_{n=0}^\infty$ für $n \to \infty$ gegen ein Grenzelement (x^*, λ^*, μ^*) konvergiert, so löst (x^*, λ^*, μ^*) das System (2.3).

Primal-Duale Verfahren wählen für $\tau > 0$ Newton-Schritte zum zentralen Pfad $\mathcal{C}$. Sei $\sigma \in [0, 1]$, und die sogenannte *gewichtete Dualitäts-Lücke* definiert durch

$$\eta = \frac{1}{n}\sum_{i=1}^n x_i\mu_i = \frac{x^T\mu}{n}.$$

Im Englischen werden die Parameter σ und η *centering parameter* beziehungsweise *duality measure* genannt. Wir schreiben $\tau = \sigma\eta$ und wenden bei festem τ einen Newton-Schritt auf (2.6) an, das heißt, auf das System $\tilde{F}(x^\tau, \lambda^\tau, \mu^\tau) = 0$:

$$\begin{pmatrix} 0 & A^T & I \\ A & 0 & 0 \\ M^k & 0 & X^k \end{pmatrix} \begin{pmatrix} \Delta x^k \\ \Delta\lambda^k \\ \Delta\mu^k \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ -X^kM^ke + \sigma\eta e \end{pmatrix},$$

wobei $X^k = \operatorname{diag}(x_1^k, \ldots, x_n^k)$ und $M^k = \operatorname{diag}(\mu_1^k, \ldots, \mu_n^k)$ gelten. Das Lösungs-Tripel $(\Delta x^k, \Delta\lambda^k, \Delta\mu^k)$ ist damit ein Newton-Schritt in Richtung des Punktes $(x^{\sigma\eta}, \lambda^{\sigma\eta}, \mu^{\sigma\eta})$ mit $x_i^{\sigma\eta}\mu_i^{\sigma\eta} = \sigma\eta$ bei festen Werten von σ und η. Im Fall von $\sigma = 0$ erhalten wir wieder den Newton-Schritt für (2.4a), für $\sigma = 1$ hingegen einen Schritt in Richtung von $(x^\eta, \lambda^\eta, \mu^\eta)$. Oft wird $\sigma \in (0, 1)$ gewählt.

Wir wollen nun den Primal-Dualen Algorithmus formulieren.

ALGORITHMUS 2.1 (Primal-Duales Verfahren).

1) *Wähle* $(x^0, \lambda^0, \mu^0) \in \mathcal{F}^\circ$ *und setze* $k = 0$.
2) *Löse das lineare Gleichungs-System*

$$\begin{pmatrix} 0 & A^T & I \\ A & 0 & 0 \\ M^k & 0 & X^k \end{pmatrix} \begin{pmatrix} \Delta x^k \\ \Delta\lambda^k \\ \Delta\mu^k \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ -X^k M^k e + \sigma_k \eta_k e \end{pmatrix} \tag{2.7}$$

mit $\sigma_k \in [0, 1]$ *und* $\eta_k = (x^k)^T \mu^k / n$. *Setze*

$$(x^{k+1}, \lambda^{k+1}, \mu^{k+1}) = (x^k, \lambda^k, \mu^k) + \alpha_k(\Delta x^k, \Delta\lambda^k, \Delta\mu^k), \tag{2.8}$$

wobei $\alpha_k \in (0, 1]$ *derart bestimmt wird, dass* $(x^{k+1}, \mu^{k+1}) > 0$ *erfüllt ist.*
3) *Sofern kein Abbruch-Kriterium eintritt, setze* $k = k + 1$ *und gehe zurück zu Schritt* 2).

BEMERKUNG 2.2.

1) Die Wahl der Parameter σ_k und α_k führt zu unterschiedlichen Varianten von Algorithmus 2.1.
2) Für den Startwert haben wir $(x^0, \lambda^0, \mu^0) \in \mathcal{F}^\circ$. Mittels Induktion können wir zeigen, dass $(x^k, \lambda^k, \mu^k) \in \mathcal{F}^\circ$ für alle $k \geq 1$ gilt.
3) Bei *nicht-zulässigen Innere-Punkte Verfahren* fordern wir nur $(x^0, \mu^0) > 0$. Mit der Notation

$$r_b(x) = Ax - b \quad \text{und} \quad r_c(\lambda, \mu) = A^T\lambda + \mu - c$$

lösen wir dann an der Stelle von (2.7) das lineare System

$$\begin{pmatrix} 0 & A^T & I \\ A & 0 & 0 \\ M^k & 0 & X^k \end{pmatrix} \begin{pmatrix} \Delta x^k \\ \Delta\lambda^k \\ \Delta\mu^k \end{pmatrix} = \begin{pmatrix} -r_c(\lambda^k, \mu^k) \\ -r_b(x^k) \\ -X^k M^k e + \sigma_k \eta_k e \end{pmatrix}.$$

Gilt $\alpha_{\bar{k}} = 1$ für ein $\bar{k} \geq 0$, so folgen $r_b(x^k) = 0$ und $r_c(\lambda^k, \mu^k) = 0$ für alle $k > \bar{k}$. Das liegt an der Linearität der beiden Gleichungen (2.5a) und (2.5b). ◊

2. Pfad-Verfolgungs Verfahren

Pfad-Verfolgungs Verfahren restringieren die Iterierten auf eine Umgebung des zentralen Pfades $\mathcal{C}$ und folgen der Kurve $\mathcal{C}$ zu einer Lösung des linearen Optimierungs-Problems. Dabei wird der Ausdruck

$$\eta_k = \frac{1}{n} \sum_{i=1}^{n} x_i^k \mu_i^k$$

sukzessive für $k \to \infty$ verkleinert. Für $\theta, \gamma \in (0, 1]$ definieren wir folgende Umgebungen des zentralen Pfades $\mathcal{C}$:

$$\mathcal{U}_2(\theta) = \left\{(x, \lambda, \mu) \in \mathcal{F}^\circ \mid \|XMe - \eta e\|_2 \leq \theta\eta\right\}$$
$$\mathcal{U}_{-\infty}(\gamma) = \left\{(x, \lambda, \mu) \in \mathcal{F}^\circ \mid x_i\mu_i \geq \gamma\eta \text{ für } i = 1, \ldots, n\right\}$$

Typische Werte sind $\theta = 0.5$ und $\gamma = 10^{-3}$. In $\mathcal{U}_{-\infty}(\gamma)$ fordern wir $x_i\mu_i \geq \gamma\eta$ für jede Komponente des Vektors XMe. Wir nähern uns im Grenzfall $\gamma \to 0$ der Menge $\mathcal{F}$. In $\mathcal{U}_2(\theta)$ sind die Anforderungen im allgemeinen restriktiver.

BEISPIEL 2.3. Wir wählen als Beispiel im $\mathbb{R}^2$ die Vektoren $x = (11, 1)^T$ und $\mu = (1, 1)^T$. Dann gilt $(x, \mu) > 0$. Ferner bekommen wir

$$\eta = \frac{11 \cdot 1 + 1 \cdot 1}{2} = 6.$$

Mit $\gamma = 1/6$ folgt $x_i\mu_i \geq \gamma\eta$ für $i = 1, 2$. Für die Euklidische Norm hingegen erhalten wir

$$\|XMe - \eta e\|_2 = \sqrt{(11-6)^2 + (1-6)^2} = \sqrt{50} > 7 > \theta\eta \quad \text{für alle } \theta \in (0, 1].$$

Damit erfüllt das Paar (x, μ) die Ungleichungs-Bedingung in $\mathcal{U}_{-\infty}(\gamma)$ für $\gamma = 1/6 \in (0, 1]$, allerdings nicht die entsprechende Bedingung in $\mathcal{U}_2(\theta)$ für ein $\theta \in (0, 1]$. $\Diamond$

ALGORITHMUS 2.4 (Zulässiges Pfad-Verfolgungs Verfahren).

1) *Wähle* $\gamma \in (0, 1)$, $0 < \underline{\sigma} < \overline{\sigma} < 1$, $\varepsilon \in (0, 1)$, $w^0 = (x^0, \lambda^0, \mu^0) \in \mathcal{U}_{-\infty}(\gamma)$ *und setze* $k = 0$.
2) *Ist* $\eta_k = (x^k)^T\mu^k/n \leq \varepsilon$ *erfüllt, dann STOPP.*
3) *Wähle* $\sigma_k \in [\underline{\sigma}, \overline{\sigma}]$ *und bestimme eine Lösung*

$$\Delta w^k = (\Delta x^k, \Delta\lambda^k, \Delta\mu^k)$$

des linearen Gleichungs-Systems (2.7). *Sei* α_k *die größte Schrittweite* $\alpha \in (0, 1]$ *mit*

$$w^k(\alpha) = (x^k + \alpha\Delta x^k, \lambda^k + \alpha\Delta\lambda^k, \mu^k + \alpha\Delta\mu^k) \in \mathcal{U}_{-\infty}(\gamma).$$

4) *Setze* $w^{k+1} = w^k(\alpha_k)$, $k = k + 1$ *und gehe zurück zu Schritt* 2).

BEMERKUNG 2.5.

1) Algorithmus 2.4 ist ein Spezialfall von Algorithmus 2.1. Bei der Wahl von σ_k bestehen weiter Freiheitsgrade, allerdings sind $\sigma_k = 0$ und $\sigma_k = 1$ ausgeschlossen. Ferner haben wir gleichmäßige Schranken für die σ_k's: $0 < \underline{\sigma} \leq \sigma_k \leq \overline{\sigma} < 1$ für alle $k \geq 0$.
2) Die Wahl von α_k ist in Algorithmus 2.4 fest vorgeschrieben. Die Konvergenzeigenschaften des Verfahrens ändern sich aber nicht, wenn wir geeignete Backtracking-Strategien verwenden:

$i_{\max} \in \mathbb{N}$; $\epsilon \in (0, 1)$; $i = 0$; $\alpha_k^{(0)} = 1$;
while $\left(w^k(\alpha_k^{(i)}) \notin \mathcal{U}_{-\infty}(\gamma)$ **and** $i \leq i_{\max}$ **and** $\alpha_k^{(i)} > \epsilon\right)$
$\alpha_k^{(i+1)} = \beta\alpha_k^{(i)}$; $i = i + 1$;
end;
$\alpha_k = \alpha_k^{(i)}$;

mit einem Parameter $\beta \in (0, 1)$. Bei einer Wahl $\beta \approx 1$ lässt sich das α_k aus Algorithmus 2.4, Schritt 2), recht gut bestimmen, es sind aber eventuell viele Iterationen in der while-Schleife der Backtracking-Strategie notwendig. Ist $\beta \approx 0$, so ist die while-Schleife schnell beendet, der Schrittweiten-Parameter α_k wird aber in der Regel deutlich kleiner sein als der von Algorithmus 2.4 in Schritt 3). $\Diamond$

3. Konvergenz-Analyse für Algorithmus 2.4

In diesem Abschnitt beschäftigen wir uns mit der Untersuchung der Konvergenz von Algorithmus 2.4. Es wird sich herausstellen, dass der Algorithmus folgende beiden Eigenschaften besitzt:

- das Verfahren bricht nach endlich vielen Iterationen mit Schritt 2.) ab,
- die Anzahl der Iterationen hängt polynomial von der Anzahl der Unbekannten ab (*polynomiale Komplexität*).

Wir bemerken an der Stelle, dass das Simplex-Verfahren aus der Linearen Programmierung kein polynomiales Verfahren ist. Es gibt dazu das Gegenbeispiel von Klee und Minty.

Zur Untersuchung der Konvergenz-Eigenschaften sind einige Hilfsresultate notwendig.

LEMMA 2.6. *Seien $u, v \in \mathbb{R}^n$ zwei Vektoren mit $u^T v \geq 0$. Dann gilt*

$$\|UVe\|_2 \leq 2^{-3/2} \|u + v\|_2^2$$

wobei $U = \operatorname{diag}(u_1, \ldots, u_n)$, $V = \operatorname{diag}(v_1, \ldots, v_n)$ Diagonalmatrizen aus $\mathbb{R}^{n \times n}$ sind und $e = (1, \ldots, 1)^T \in \mathbb{R}^n$ gilt.

BEWEIS. Zunächst gilt für beliebige Zahlen $\alpha, \beta \in \mathbb{R}$:

$$\frac{1}{4}(\alpha + \beta)^2 = \frac{1}{4}(\alpha - \beta)^2 + \alpha\beta \geq \alpha\beta. \tag{2.9}$$

Wegen $u^T v \geq 0$ erhalten wir

$$0 \leq u^T v = \sum_{u_i v_i \geq 0} u_i v_i + \sum_{u_i v_i < 0} u_i v_i = \sum_{i \in \mathcal{P}} u_i v_i - \sum_{i \in \mathcal{M}} |u_i v_i| \tag{2.10}$$

mit der Menge $\mathcal{P} = \{i \in \{1, \ldots, n\} \mid u_i v_i \geq 0\}$ der nicht-negativen Indizes und der Menge $\mathcal{M} = \{i \in \{1, \ldots, n\} \mid u_i v_i < 0\}$ der negativen Indizes. Weiter haben wir die Ungleichung

$$\|x\|_2 \leq \|x\|_1 = \sum_{i=1}^{n} |x_i| \quad \text{für alle } x \in \mathbb{R}^n \tag{2.11}$$

zur Verfügung. Die Beziehung (2.11) sehen wir wie folgt: Für $x = 0$ ist nichts zu zeigen. Sei nun $\alpha = \|x\|_1 > 0$. Dann gilt $|x_i|/\alpha \leq 1$ für alle $1 \leq i \leq n$ und wir erhalten

$$\frac{\|x\|_2^2}{\|x\|_1^2} = \sum_{i=1}^{n} \left(\frac{x_i}{\alpha}\right)^2 \leq \sum_{i=1}^{n} \frac{|x_i|}{\alpha} = 1,$$

woraus direkt (2.11) folgt. Bevor wir nun die Aussage des Lemmas beweisen können, führen wir noch eine Notation ein: Für den Vektor $w \in \mathbb{R}^n$ mit den Komponenten $w_i = u_i v_i$, $i = 1, \ldots, n$, schreiben wir $[u_i v_i]_{i \in \mathcal{P}}$ für den Vektor, der nur die nicht-negativen Komponenten von w_i enthält und $[u_i v_i]_{i \in \mathcal{M}}$ für den Vektor, der nur die

negativen Komponenten enthält. Nun ergibt sich aus (2.9)-(2.11):

$$\begin{aligned}\|UVe\|_2 &= \Big(\|[u_iv_i]_{i\in\mathcal{P}}\|_2^2 + \|[u_iv_i]_{i\in\mathcal{M}}\|_2^2\Big)^{1/2}\\ &\overset{(2.11)}{\leq} \Big(\|[u_iv_i]_{i\in\mathcal{P}}\|_1^2 + \|[u_iv_i]_{i\in\mathcal{M}}\|_1^2\Big)^{1/2}\\ &\overset{(2.10)}{\leq} \Big(2\,\|[u_iv_i]_{i\in\mathcal{P}}\|_1^2\Big)^{1/2} = \sqrt{2}\,\|[u_iv_i]_{i\in\mathcal{P}}\|_1\\ &\overset{(2.9)}{\leq} \Big(2\,\|\big[\tfrac{1}{2}(u_i+v_i)^2\big]_{i\in\mathcal{P}}\|_1^2\Big)^{1/2} = 2^{-3/2}\sum_{i\in\mathcal{P}}(u_i+v_i)^2\\ &\leq 2^{-3/2}\sum_{i=1}^{n}(u_i+v_i)^2 = 2^{-3/2}\,\|u+v\|_2^2,\end{aligned}$$

was zu zeigen war. □

Seien $\Delta x^k = (\Delta x_1^k, \ldots, \Delta x_n^k)^T$ und $\Delta\mu^k = (\Delta\mu_1^k, \ldots, \Delta\mu_n^k)^T$ die Lösungen von (2.7). Wir führen die beiden folgenden $n \times n$-Diagonal-Matrizen ein:

$$\Delta X^k = \operatorname{diag}(\Delta x_1^k, \ldots, \Delta x_n^k) \quad \text{und} \quad \Delta M^k = \operatorname{diag}(\Delta\mu_1^k, \ldots, \Delta\mu_n^k).$$

LEMMA 2.7. *Sei $(x^k, \lambda^k, \mu^k) \in \mathcal{U}_{-\infty}(\gamma)$. Dann folgt*

$$\|\Delta X^k \Delta M^k e\|_2 \leq 2^{-3/2}\left(1 + \frac{1}{\gamma}\right) n\eta_k,$$

wobei $\eta_k = (x^k)^T\mu^k/n$ die aktuelle gewichtete Dualitätslücke bezeichnet.

BEWEIS. Nach Voraussetzung gelten $x_i^k\mu_i^k \geq \gamma\eta_k > 0$, so dass alle Komponenten auf den Diagonalen von X^k und M^k positiv sind. Aus der dritten Blockzeile von (2.7) ergibt sich

$$M^k\Delta x^k + X^k\Delta\mu^k = -X^kM^ke + \sigma_k\eta_ke.$$

Multiplizieren wir diese Gleichung mit $(X^kM^k)^{-1/2}$ von links und verwenden die Abkürzung $D^k = (X^k)^{1/2}(M^k)^{-1/2}$ so erhalten wir

$$\begin{aligned}&(M^k)^{-1/2}(X^k)^{-1/2}M^k\Delta x^k + (M^k)^{-1/2}(X^k)^{1/2}\Delta\mu^k\\ &\qquad = (X^kM^k)^{-1/2}\big(-X^kM^ke + \sigma_k\eta_ke\big).\end{aligned}$$

Da X^k und M^k Diagonalmatrizen sind, erhalten wir

$$(D^k)^{-1}\Delta x^k + D^k\Delta\mu^k = (X^kM^k)^{-1/2}\big(-X^kM^ke + \sigma_k\eta_ke\big). \tag{2.12}$$

Weiter haben wir

$$(\Delta x^k)^T\Delta\mu^k = 0. \tag{2.13}$$

Denn aus der ersten Blockzeile in (2.7) folgen $A^T\Delta\lambda^k + \Delta\mu^k = 0$ und daher

$$(\Delta x^k)^TA^T\Delta\lambda^k + (\Delta x^k)^T\Delta\mu^k = 0.$$

Wegen der zweiten Blockzeile in (2.7) gilt $(\Delta x^k)^TA^T = 0$ und damit bekommen wir (2.13). Setzen wir $u = (D^k)^{-1}\Delta x^k$ und $v = D^k\Delta\mu^k$, so bekommen wir mit

Lemma 2.6 und (2.12)

$$\begin{aligned}\|\Delta X^k \Delta M^k e\|_2 &= \|((D^k)^{-1}\Delta X^k)(D^k \Delta M^k)e\|_2 \\ &\le 2^{-3/2}\,\|(D^k)^{-1}\Delta x^k + D^k \Delta\mu^k\|_2^2 \\ &= 2^{-3/2}\,\|(X^k M^k)^{-1/2}(-X^k M^k e + \sigma_k\eta_k e)\|_2^2.\end{aligned} \tag{2.14}$$

Aus $(x^k)^T\mu^k = n\eta_k$, $e^T e = n$, $x_i^k\mu_i^k \ge \gamma\eta_k$ für $i = 1,\dots,n$ und $\sigma_k \in (0,1)$ erhalten wir

$$\begin{aligned}\|\Delta X^k \Delta M^k e\|_2 &\le 2^{-3/2}\,\|-(X^k M^k)^{1/2}e + \sigma_k\eta_k (X^k M^k)^{-1/2}e)\|_2^2 \\ &= 2^{-3/2}\left((x^k)^T\mu^k - 2\sigma_k\eta_k e^T e + \sigma_k^2\eta_k^2 \sum_{i=1}^n \frac{1}{x_i^k\mu_i^k}\right) \\ &\le 2^{-3/2}\left((x^k)^T\mu^k - 2\sigma_k\eta_k e^T e + \sigma_k^2\eta_k \frac{n}{\gamma}\right) \\ &= 2^{-3/2}\left(n\eta_k - 2\sigma_k\eta_k n + \frac{n\sigma_k^2\eta_k}{\gamma}\right) \\ &= 2^{-3/2} n\eta_k\left(1 - 2\sigma_k + \frac{\sigma_k^2}{\gamma}\right) < 2^{-3/2} n\eta_k\left(1 + \frac{1}{\gamma}\right),\end{aligned}$$

was zu zeigen war. □

Wir geben als nächstes eine obere Schranke für die Schrittweite α_k an. Dieses Resultat kann als der wesentliche Schritt zum Nachweis der polynomialen Komplexität von Algorithmus 2.4 angesehen werden.

LEMMA 2.8. *Sei die Iterierte $(x^k,\lambda^k,\mu^k) \in \mathcal{U}_{-\infty}(\gamma)$ gegeben. Dann gilt*

$$(x^k(\alpha),\lambda^k(\alpha),\mu^k(\alpha)) \in \mathcal{U}_{-\infty}(\gamma)$$

für alle $\alpha \in (0,\bar\alpha_k]$ mit

$$\bar\alpha_k = 2^{3/2}\gamma\,\frac{\sigma_k}{n}\,\frac{1-\gamma}{1+\gamma}.$$

BEWEIS. Aus der dritten Blockzeile von (2.7) ergibt sich

$$\mu_i^k\Delta x_i^k + x_i^k\Delta\mu_i^k = -x_i^k\mu_i^k + \sigma_k\eta_k \quad \text{für } i = 1,\dots,n. \tag{2.15}$$

Anwendung von Lemma 2.7 führt auf

$$|\Delta x_i^k\Delta\mu_i^k| \le \|\Delta X^k\Delta M^k e\|_2 \le 2^{-3/2}\left(1+\frac{1}{\gamma}\right)n\eta_k \quad \text{für } i = 1,\dots,n. \tag{2.16}$$

Mit $x_i^k\mu_i^k \ge \gamma\eta_k$ für $i = 1,\dots,n$, (2.15) und (2.16) folgt

$$\begin{aligned}x_i^k(\alpha)\mu_i^k(\alpha) &= (x_i^k + \alpha\Delta x_i^k)(\mu_i^k + \alpha\Delta\mu_i^k) \\ &= x_i^k\mu_i^k + \alpha\big(x_i^k\Delta\mu_i^k + \mu_i^k\Delta x_i^k\big) + \alpha^2\Delta x_i^k\Delta\mu_i^k \\ &\overset{(2.15)}{\ge} (1-\alpha)x_i^k\mu_i^k + \alpha\sigma_k\eta_k - \alpha^2|\Delta x_i^k\Delta\mu_i^k| \\ &\overset{(2.16)}{\ge} (1-\alpha)\gamma\eta_k + \alpha\sigma_k\eta_k - 2^{-3/2}\alpha^2 n\eta_k\left(1+\frac{1}{\gamma}\right)\end{aligned} \tag{2.17}$$

für $i = 1,\dots,n$ und für $\alpha \in [0,1]$. Die dritte Blockzeile von (2.7) lautet in Matrix-Schreibweise

$$M^k\Delta x^k + X^k\Delta\mu^k = -X^k M^k e + \sigma_k\eta_k e.$$

Summation über die n Komponenten dieser Gleichung liefert

$$(\mu^k)^T \Delta x^k + (x^k)^T \Delta \mu^k = -(1-\sigma_k)(x^k)^T \mu^k.$$

Zusammen mit (2.13) ergibt sich daher

$$\begin{aligned}(x^k(\alpha))^T \mu^k(\alpha) &= (x^k)^T \mu^k + \alpha\big((\mu^k)^T \Delta x^k + (x^k)^T \Delta \mu^k\big) \\ &= (x^k)^T \mu^k (1 - \alpha(1-\sigma_k)).\end{aligned} \tag{2.18}$$

Daher ist die Bedingung

$$x_i^k(\alpha)\mu_i^k(\alpha) \geq \gamma\eta_k = \gamma\,\frac{(x^k(\alpha))^T\mu^k(\alpha)}{n} \overset{(2.18)}{=} \gamma\big(1-\alpha(1-\sigma_k)\big)\eta_k \tag{2.19}$$

für $i = 1, \ldots, n$ erfüllt, sofern wegen (2.17)

$$\gamma(1-\alpha)\eta_k + \alpha\sigma_k\eta_k - 2^{-3/2}\alpha^2\left(1+\frac{1}{\gamma}\right)n\eta_k \geq \gamma\big(1-\alpha(1-\sigma_k)\big)\eta_k$$

erfüllt ist. Eine Umformung der Terme zeigt, dass dies äquivalent ist zu

$$\alpha\sigma_k\eta_k(1-\gamma) \geq 2^{-3/2}\alpha^2 n\eta_k\left(1+\frac{1}{\gamma}\right).$$

Letzteres führt auf die Bedingung

$$\alpha \leq 2^{3/2}\gamma\,\frac{\sigma_k}{n}\,\frac{1-\gamma}{1+\gamma} = \bar{\alpha}_k.$$

Nun ist nur noch $(x^k(\alpha), \lambda^k(\alpha), \mu^k(\alpha)) \in \mathcal{F}^\circ$ für alle $\alpha \in [0, \bar{\alpha}_k]$ zu zeigen. Nach Voraussetzung haben wir $Ax^k = b$ und $A^T\lambda^k + \mu^k = c$. Wegen (2.7) gelten dann offenbar

$$Ax^k(\alpha) = b \quad \text{und} \quad A^T\lambda^k(\alpha) + \mu^k(\alpha) = c$$

für alle $\alpha \geq 0$. Aus $\gamma \in (0,1)$ bekommen wir $\gamma(1-\gamma) \leq 1/4$. Also

$$\alpha \leq 2^{3/2}\gamma\,\frac{\sigma_k}{n}\,\frac{1-\gamma}{1+\gamma} \leq 2^{3/2}\,\frac{1}{4}\,\frac{\overline{\sigma}}{n}\,\frac{1}{1+\gamma} < \frac{1}{\sqrt{2}n} < 1.$$

Wegen $x^k = x^k(0) > 0$ und $\mu^k = \mu^k(0) > 0$ folgt aus (2.19) und dem gerade bewiesenem Teil für alle $\alpha \in (0, \bar{\alpha}_k] \subsetneq [0,1]$:

$$x_i^k(\alpha)\mu_i^k(\alpha) \geq \gamma\big(1 - \underbrace{\alpha(1-\sigma_k)}_{<1}\big)\eta_k > 0 \quad \text{für } i = 1, \ldots, n.$$

Also kann kein Index $i \in \{1, \ldots, n\}$ und kein $\alpha \in [0, \bar{\alpha}_k]$ existieren mit $x_i^k(\alpha) = 0$ oder $\mu_i^k(\alpha) = 0$. □

Nun können wir die Reduktion von η_k abschätzen.

Satz 2.9. *Sei $\{(x^k, \lambda^k, \mu^k)\}_{k=0}^\infty$ eine durch Algorithmus 2.4 erzeugte Folge. Dann gilt*

$$\eta_{k+1} \leq \left(1 - \frac{\delta}{n}\right)\eta_k \quad \textit{für alle } k \geq 0$$

für eine von k unabhängige Konstante $\delta > 0$.

BEWEIS. Wegen Lemma 2.8 erhalten wir

$$\alpha_k \geq \bar{\alpha}_k = 2^{3/2}\gamma \frac{\sigma_k}{n} \frac{1-\gamma}{1+\gamma} \quad \text{für alle } k \geq 0.$$

Aus (2.18) ergibt sich daher

$$\begin{aligned} \eta_{k+1} &= \eta_k(\alpha_k) = \frac{(x^k(\alpha_k))^T \mu^k(\alpha_k)}{n} \\ &= \big(1 - \alpha_k(1-\sigma_k)\big)\eta_k \leq \left(1 - \frac{2^{3/2}}{n}\,\gamma\,\frac{1-\gamma}{1+\gamma}\sigma_k(1-\sigma_k)\right)\eta_k. \end{aligned} \tag{2.20}$$

Die quadratische Funktion $\sigma \mapsto \sigma(1-\sigma)$ ist strikt konkav. Daher nimmt sie ihr Minimum in dem kompakten Intervall $[\underline{\sigma}, \overline{\sigma}] \subset (0,1)$ an einem der Endpunkte an. Also gilt

$$\sigma_k(1-\sigma_k) \geq \min\big\{\underline{\sigma}(1-\underline{\sigma}), \overline{\sigma}(1-\overline{\sigma})\big\} \quad \text{für alle } \sigma_k \in [\underline{\sigma}, \overline{\sigma}].$$

Setzen wir

$$\delta = 2^{3/2}\,\gamma\,\frac{1-\gamma}{1+\gamma}\,\min\big\{\underline{\sigma}(1-\underline{\sigma}), \overline{\sigma}(1-\overline{\sigma})\big\} > 0,$$

so folgt die Behauptung des Satzes aus (2.20). □

Nun sind wir in der Lage, das wesentliche Konvergenzresultat für Algorithmus 2.4 zu beweisen. Dieses besagt, dass der Algorithmus 2.4 nach $O(n|\log(\varepsilon)|)$ Iterationen dem Abbruchkriterium aus Schritt 2) genügt, wobei die in der O-Notation steckende Konstante von der Qualität des Startvektors abhängt.

SATZ 2.10. *Sei $\{(x^k, \lambda^k, \mu^k)\}_{k=0}^{\infty}$ eine durch Algorithmus 2.4 erzeugte Folge, wobei der Startvektor (x^0, λ^0, μ^0) der Bedingung*

$$\eta_0 \leq \frac{1}{\varepsilon^{\varrho}} \tag{2.21}$$

für eine positive Konstante $\varrho > 0$ genügt. Dann existiert ein $K \in \mathbb{N}$ mit $K = O(n|\log(\varepsilon)|)$ und

$$\eta_k \leq \varepsilon \quad \textit{für alle } k \geq K.$$

BEWEIS. Nach Satz 2.9 haben wir

$$\eta_{k+1} \leq \left(1 - \frac{\delta}{n}\right)\eta_k.$$

Also folgt

$$\log \eta_{k+1} \leq \log\left(1 - \frac{\delta}{n}\right) + \log \eta_k.$$

Wiederholte Anwendung ergibt mit (2.21)

$$\begin{aligned} \log \eta_k &\leq \log\left(1 - \frac{\delta}{n}\right) + \log \eta_{k-1} \leq 2\log\left(1 - \frac{\delta}{n}\right) + \log \eta_{k-2} \\ &\leq k \log\left(1 - \frac{\delta}{n}\right) + \log \eta_0 \leq k \log\left(1 - \frac{\delta}{n}\right) + \varrho \log \frac{1}{\varepsilon}. \end{aligned}$$

Wegen $1 + \beta \leq e^{\beta}$ gilt

$$\log(1+\beta) \leq \beta \quad \text{für alle } \beta > -1.$$

Daher erhalten wir

$$\log \eta_k \leq k\left(-\frac{\delta}{n}\right) + \varrho \log \frac{1}{\varepsilon}.$$

Es gilt also $\eta_k \leq \varepsilon$, sofern

$$k\left(-\frac{\delta}{n}\right) + \varrho \log \frac{1}{\varepsilon} \leq \log \varepsilon$$

erfüllt ist. Also bekommen wir die Bedingung

$$k \geq K = (1+\varrho)\frac{n}{\delta} \log \frac{1}{\varepsilon},$$

was zu zeigen war. □

Bemerkung 2.11. Wir betonen abschließend noch, dass die beiden Sätze 2.9 und 2.10 auch noch gelten, wenn wir α_k in Algorithmus 2.4 durch die in Lemma 2.8 gegebene explizite Schranke $\bar{\alpha}_k$ ersetzen. ◊

Nun kommen wir zu einem nicht-zulässigen Verfahren. Dazu haben wir bereits die Residuenvektoren

$$r_b(x) = Ax - b \in \mathbb{R}^m \quad \text{und} \quad r_c(\lambda, \mu) = A^T\lambda + \mu - c \in \mathbb{R}^n$$

eingeführt. Wir werden die Kurzschreibweise $r_b^k = r_b(x^k)$ sowie $r_c^k = r_c(\lambda^k, \mu^k)$ verwenden. In Algorithmus 2.4 galten stets $r_b^k = 0$ und $r_c^k = 0$ für alle $k \in \mathbb{N}$. Die Wahl eines Startwertes $(x^0, \lambda^0, \mu^0) \in \mathcal{U}_{-\infty}(\gamma)$ kann aber unter Umständen nicht so einfach sein. Im weiteren müssen die beiden Residuenvektoren r_b^k und r_c^k nicht mehr notwendig gleich null sein, so dass im Hinblick auf die Wahl des Startwertes mehr Freiheiten zugelassen sind. Allerdings muss die Menge $\mathcal{U}_{-\infty}(\gamma)$ modifiziert werden:

$$\mathcal{U}_{-\infty}(\gamma,\beta) = \Bigg\{ (x,\lambda,\mu) \in \mathbb{R}^n \times \mathbb{R}^m \times \mathbb{R}^n \mid x_i\,\mu_i \geq \gamma\eta \text{ für } 1 \leq i \leq n \text{ und } \\ \|(r_b(x), r_c(\lambda,\mu))\|_2 \leq \frac{\|(r_b^0, r_c^0)\|_2}{\eta_0}\,\beta\eta \Bigg\}$$

mit $\gamma \in (0,1)$, $\beta \geq 1$ und $\eta = x^T\mu/n$. Offenbar ist $\beta \geq 1$ notwendig dafür, dass auch der Startwert (x^0, λ^0, μ^0) in $\mathcal{U}_{-\infty}(\gamma, \beta)$ liegt. In $\mathcal{U}_{-\infty}(\gamma, \beta)$ habe wir die zusätzliche Forderung

$$\|(r_b(x), r_c(\lambda,\mu)\|_2 \leq \frac{\|(r_b^0, r_c^0)\|_2}{\eta_0}\,\beta\eta,$$

um die Verletztheit der beiden linearen Gleichungen $Ax = b$ und $A^T\lambda + \mu = c$ zu messen.

Gilt $\lim_{k\to\infty} \eta_k = 0$, so folgt auch für nicht-zulässiges (x^k, λ^k, μ^k)

$$\lim_{k\to\infty} r_b^k = 0 \quad \text{und} \quad \lim_{k\to\infty} r_c^k = 0.$$

Jeder Häufungspunkt der Folge $\{(x^k, \lambda^k, \mu^k)\}_{k=0}^{\infty}$ erfüllt daher die Optimalitätsbedingungen

$$\begin{aligned} A^T\lambda + \mu &= c, \\ Ax &= b, \\ x_i\mu_i &= 0, \quad i = 1,\ldots,n, \\ (x,\mu) &\geq 0. \end{aligned}$$

Algorithmus 2.12 (Nicht-zulässiges Pfad-Verfolgungs Verfahren).

1) *Wähle* $\gamma \in (0,1)$, $\beta \geq 1$, $0 < \underline{\sigma} < \overline{\sigma} \leq 1/2$, $\varepsilon \in (0,1)$, $w^0 = (x^0, \lambda^0, \mu^0)$ *mit* $(x^0, \mu^0) > 0$ *und* $x_i^0\mu_i^0 \geq \gamma\eta_0$ *für* $i = 1,\ldots,n$ *und setze* $k = 0$.

2) *Ist* $\eta_k = (x^k)^T \mu^k / n \leq \varepsilon$ *erfüllt, dann STOPP.*
3) *Wähle* $\sigma_k \in [\underline{\sigma}, \overline{\sigma}]$ *und bestimme eine Lösung*

$$\Delta w^k = (\Delta x^k, \Delta \lambda^k, \Delta \mu^k)$$

des linearen Gleichungs-Systems

$$\begin{pmatrix} 0 & A^T & I \\ A & 0 & 0 \\ M^k & 0 & X^k \end{pmatrix} \begin{pmatrix} \Delta x \\ \Delta \lambda \\ \Delta \mu \end{pmatrix} = \begin{pmatrix} -r_c^k \\ -r_b^k \\ -X^k M^k e + \sigma_k \eta_k e \end{pmatrix}. \tag{2.22}$$

Sei α_k *die größte Schrittweite* $\alpha \in (0, 1]$ *mit*

$$w^k(\alpha) = (x^k + \alpha \Delta x^k, \lambda^k + \alpha \Delta \lambda^k, \mu^k + \alpha \Delta \mu^k) \in \mathcal{U}_{-\infty}(\gamma, \beta) \tag{2.23}$$

und

$$\eta_k(\alpha) \leq (1 - 0.01\alpha)\eta_k. \tag{2.24}$$

4) *Setze* $w^{k+1} = w^k(\alpha_k)$, $k = k + 1$ *und gehe zurück zu Schritt* 2).

Bemerkung 2.13. 1) Wegen $\beta \geq 1$ folgt $(x^0, \lambda^0, \mu^0) \in \mathcal{U}_{-\infty}(\gamma, \beta)$. Daher liegt wegen (2.23) die gesamte Folge $\{w^k\}_{k=0}^{\infty}$ in $\mathcal{U}_{-\infty}(\gamma, \beta)$.
2) Die Bedingung (2.24) garantiert eine hinreichende Abnahme der gewichteten Dualitätslücke η_k.
3) Die schwer zu berechnende Schrittweite α_k kann wieder durch einen expliziten Ausdruck ersetzt werden. Für Details verweisen wir an dieser Stelle auf [**13**]. ◊

Wir zitieren hier den folgenden Satz aus [**13**, S. 159].

Satz 2.14. *Sei* $\{w^k\}_{k=0}^{\infty}$ *eine durch Algorithmus* 2.12 *erzeugte Folge. Dann gelten folgende Aussagen:*

1) *Die Folge* $\{\eta_k\}_{k=0}^{\infty}$ *konvergiert linear gegen Null.*
2) *Die Folge* $\{\|(r_b^k, r_c^k)\|\}_{k=0}^{\infty}$ *konvergiert r-linear gegen Null, das heißt, die Folge* $\{\|(r_b^k, r_c^k)\|\}_{k=0}^{\infty}$ *nicht-negativer Zahlen wird durch eine linear gegen Null konvergierende Folge* $\{c_k\}_{k=0}^{\infty}$ *majorisiert (*$c_k \geq 0$ *für alle* $k \in \mathbb{N}$, $\lim_{k\to\infty} c_k = 0$ *und* $\|(r_b^k, r_c^k)\|_2 \leq c_k$ *für alle* $k \in \mathbb{N}$*).*

4. Der Prädiktor-Korrektor Algorithmus von Mehrotra

Die meisten Innere-Punkte Verfahren in Programm-Bibliotheken basieren auf einer von Mehrotra vorgeschlagenen Variante, die im wesentlichen zwei Gesichtspunkte hat:

1) Hinzufügen eines Korrektur-Schrittes bei der Berechnung der Suchrichtung;
2) adaptive Wahl des Parameters σ_k.

Motivieren kann man das Verfahren, in dem der zentrale Pfad $\mathcal{C}$ so verschoben wird, dass er an der aktuellen Iterierten (x^k, λ^k, μ^k) beginnt und weiterhin in der Menge der zulässigen Lösungen endet. Daher haben wir eine modifizierte Kurve

$$\mathcal{H} = \left\{(\hat{x}(s), \hat{\lambda}(s), \hat{\mu}(s)) \in \mathbb{R}^n \times \mathbb{R}^m \times \mathbb{R}^n : s \in [0, 1)\right\}$$

mit $(\hat{x}(0), \hat{\lambda}(0), \hat{\mu}(0)) = (x^k, \lambda^k, \mu^k)$ und, sofern der Grenzwert existiert,

$$\lim_{s \nearrow 1} (\hat{x}(s), \hat{\lambda}(s), \hat{\mu}(s)) \in \mathcal{C}$$

Damit startet die Kurve $\mathcal{H}$ von einem Punkt, der nicht auf $\mathcal{C}$ liegt, endet aber in einem Punkt, der zur Menge $\mathcal{C}$ gehört.

Das Verfahren kombiniert drei Schritte zur Bestimmung der Suchrichtung:

1) *Prädiktor-Schritt*, der erlaubt, σ_k zu berechnen;
2) *Korrektur-Schritt*, der Information zweiter Ordnung von $\mathcal{H}$ (d.h., Information über die Krümmung), ausnutzt, um näher an die Lösung zu kommen;
3) *Zentrierender Schritt*, in dem σ_k in die dritte Blockgleichung in (2.22) eingesetzt wird.

Konkret lösen wir zunächst (2.22) mit der Wahl $\sigma_k = 0$, d.h.,

$$\begin{pmatrix} 0 & A^T & I \\ A & 0 & 0 \\ M^k & 0 & X^k \end{pmatrix} \begin{pmatrix} \Delta x^{\text{aff}} \\ \Delta\lambda^{\text{aff}} \\ \Delta\mu^{\text{aff}} \end{pmatrix} = \begin{pmatrix} -r_c^k \\ -r_b^k \\ -X^k M^k e \end{pmatrix}. \tag{2.25}$$

Die Richtung $(\Delta x^{\text{aff}}, \Delta\lambda^{\text{aff}}, \Delta\mu^{\text{aff}})$ wird im Englischen *affine scaling direction* genannt, daher die Bezeichnung mit dem Index aff.

Dann bestimmen wir die größtmögliche Schrittweiten $\alpha^{\text{aff}}_{\text{prim}}$, $\alpha^{\text{aff}}_{\text{dual}} \in (0, 1]$, so dass

$$x^k(\alpha^{\text{aff}}_{\text{prim}}) = x^k + \alpha^{\text{aff}}_{\text{prim}} \Delta x^{\text{aff}} \geq 0, \quad \mu^k(\alpha^{\text{aff}}_{\text{dual}}) = \mu^k + \alpha^{\text{aff}}_{\text{dual}} \Delta\mu^{\text{aff}} \geq 0.$$

Wir haben explizite Formeln für die Schrittweiten zur Verfügung:

$$\alpha^{\text{aff}}_{\text{prim}} = \min\left\{1, \min_{i\,:\,\Delta x_i^{\text{aff}} < 0} \frac{-x_i^k}{\Delta x_i^{\text{aff}}}\right\}, \tag{2.26a}$$

$$\alpha^{\text{aff}}_{\text{dual}} = \min\left\{1, \min_{i\,:\,\Delta \mu_i^{\text{aff}} < 0} \frac{-\mu_i^k}{\Delta \mu_i^{\text{aff}}}\right\}. \tag{2.26b}$$

Dann folgen offenbar

$$x_i^k + \alpha^{\text{aff}}_{\text{prim}} \Delta x_i^{\text{aff}} \geq x_i^k - \frac{x_i^k}{\Delta x_i^{\text{aff}}} \Delta x_i^{\text{aff}} = 0,$$

$$\mu_i^k + \alpha^{\text{aff}}_{\text{dual}} \Delta \mu_i^{\text{aff}} \geq \mu_i^k - \frac{\mu_i^k}{\Delta \mu_i^{\text{aff}}} \Delta \mu_i^{\text{aff}} = 0.$$

Mit den berechneten Schrittweiten berechnen wir

$$\eta^{\text{aff}} = \frac{1}{n} \left(x^k(\alpha^{\text{aff}}_{\text{prim}})\right)^T \mu^k(\alpha^{\text{aff}}_{\text{dual}}) \tag{2.27}$$

und mit $\eta_k = (x^k)^T \mu^k / n$ setzen wir

$$\sigma = \left(\frac{\eta^{\text{aff}}}{\eta_k}\right)^3.$$

Ist nun $\eta^{\text{aff}} \ll \eta_k$, so ist σ klein (und umgekehrt).

Im Korrektur-Schritt wählen wir auf der rechten Seite von (2.25) den Vektor $(0, 0, -\Delta X^{\text{aff}} \Delta M^{\text{aff}} e)^T$ und im letzten Schritt $(0, 0, \sigma\eta_k e)^T$. Insgesamt haben wir dann das System

$$\begin{pmatrix} 0 & A^T & I \\ A & 0 & 0 \\ M^k & 0 & X^k \end{pmatrix} \begin{pmatrix} \Delta x \\ \Delta\lambda \\ \Delta\mu \end{pmatrix} = \begin{pmatrix} -r_c^k \\ -r_b^k \\ -X^k M^k e - \Delta X^{\text{aff}} \Delta M^{\text{aff}} e + \sigma\eta_k e \end{pmatrix} \tag{2.28}$$

und setzen dann

$$\alpha_k^{\text{prim}} = \min\left\{1, \min_{i\,:\,\Delta x_i<0} \frac{-x_i^k}{\Delta x_i}\right\}, \tag{2.29a}$$

$$\alpha_k^{\text{dual}} = \min\left\{1, \min_{i\,:\,\Delta \mu_i<0} \frac{-\mu_i^k}{\Delta \mu_i}\right\}. \tag{2.29b}$$

Bemerkung 2.15. Um den Korrekturschritt zu motivieren, betrachten wir $(x_i^k + \Delta x_i^{\text{aff}})(\mu_i^k + \Delta\mu_i^{\text{aff}}) = x_i^k\mu_i^k + x_i^k\Delta\mu_i^{\text{aff}} + \Delta x_i^{\text{aff}}\mu_i^k + \Delta x_i^{\text{aff}}\Delta\mu_i^{\text{aff}} = \Delta x_i^{\text{aff}}\Delta\mu_i^{\text{aff}}$, wobei wir die dritte Blockzeile von (2.25) verwendet haben. Wird also ein voller Schritt gewählt, d.h., $\alpha_{\text{prim}}^{\text{aff}} = \alpha_{\text{dual}}^{\text{aff}} = 1$, so geht das Produkt $x_i^k\mu_i^k$ im Prädiktor-Schritt statt auf den Wert 0 über in das Produkt $\Delta x_i^{\text{aff}}\Delta\mu_i^{\text{aff}}$, $i = 1, \ldots, n$. Der Korrektur-Schritt versucht dieses zu kompensieren, so dass das Produkt der Komponenten $x^k + \Delta x$ mit $\mu^k + \Delta\mu$ näher an Null ist. ◊

Algorithmus 2.16 (Mehrotra Prädiktor-Korrektur Verfahren).

1) *Wähle* $(x^0, \lambda^0, \mu^0) \in \mathbb{R}^n \times \mathbb{R}^m \times \mathbb{R}^n$ *mit* $(x^0, \mu^0) > 0$, $\varepsilon \in (0, 1)$ *und setze* $k = 0$.
2) *Ist* $\eta_k = (x^k)^T\mu^k/n \leq \varepsilon$ *erfüllt, dann STOPP.*
3) *Löse* (2.25) *für* $(\Delta x^{\text{aff}}, \Delta\lambda^{\text{aff}}, \Delta\mu^{\text{aff}})$.
4) *Berechne die Schrittweiten* $\alpha_{\text{prim}}^{\text{aff}}$, $\alpha_{\text{dual}}^{\text{aff}}$, η^{aff} *gemäß* (2.26) *sowie* (2.27) *und setze* $\sigma = (\eta^{\text{aff}}/\eta_k)^3$.
5) *Löse* (2.28) *für* $(\Delta x, \Delta\lambda, \Delta\mu)$.
6) *Berechne die Schrittweiten* α_k^{prim} *und* α_k^{dual} *mit* (2.29).
7) *Setze* $x^{k+1} = x^k + \alpha_k^{\text{prim}}\Delta x$, $(\lambda^{k+1}, \mu^{k+1}) = (\lambda^k, \mu^k) + \alpha_k^{\text{dual}}(\Delta\lambda, \Delta\mu)$, $k = k + 1$, *und gehe zurück zu Schritt* 2).

Bemerkung 2.17. Für Algorithmus 2.16 ist keine Konvergenzanalysis vorhanden. Es gibt auch Beispiele, in denen Algorithmus 2.16 divergiert, was allerdings durch kleine Modifikationen verhindert werden kann. In vielen Anwendungen ist aber das Konvergenzverhalten des Prädiktor-Korrektor Verfahrens sehr gut. ◊

KAPITEL 3

Quadratische Programmierung

In diesem Kapitel beschäftigen wir uns mit der Quadratischen Programmierung (im Englischen *quadratic programming*), wo die Zielfunktion quadratisch ist und lineare Nebenbedingungen vorliegen. Diese Problemklasse ist insbesondere auch deshalb von großer Bedeutung, da sie uns im Kapitel 4 als Teilproblem von einem iterativem Verfahren, dem SQP-Verfahren, wieder beschäftigen wird.

Das Standard-Problem in diesem Abschnitt lautet wie folgt

$$(\mathbf{QP}) \quad \min q(x) = \frac{1}{2}x^T Qx + x^T d \quad \text{u.d.N.} \quad \begin{cases} a_i^T x = b_i, & i = 1, \ldots, m, \\ a_i^T x \geq b_i, & i = m+1, \ldots, m+p. \end{cases}$$

Wir setzen voraus, dass $Q \in \mathbb{R}^{n \times n}$ symmetrisch und positiv semi-definit ist. Ferner gelte $a_i \in \mathbb{R}^n$ für $i = 1, \ldots, m+p$. Dann ist (**QP**) ein konvexes Optimierungsproblem, da die Nebenbedingungen eine konvexe Menge beschreiben und die Zielfunktion konvex ist.

1. Gleichungsrestringierte Probleme

In diesem Abschnitt beschränken wir uns auf Probleme ohne Ungleichungen. Daher betrachten wir

$$(\mathbf{QP}_{Gl}) \qquad \min q(x) = \frac{1}{2}x^T Qx + x^T d \quad \text{u.d.N.} \quad Ax = b,$$

wobei $A \in \mathbb{R}^{m \times n}$ gegeben ist durch

$$A = \begin{pmatrix} a_1^T \\ \vdots \\ a_m^T \end{pmatrix}$$

mit $\operatorname{Rang} A = m$, d.h., A hat vollen Rang. Ferner gelte $m \leq n$.

Um die notwendigen Bedingungen erster Ordnung (siehe Satz 1.9) aufzustellen, führen wir die affin-lineare Abbildung $e : \mathbb{R}^n \to \mathbb{R}^m$ durch $e(x) = Ax - b$ für $x \in \mathbb{R}^n$ ein. Wegen $\nabla e(x) = A$ folgt, dass die Jacobi-Matrix von e vollen Rang besitzt für alle $x \in \mathbb{R}^n$. Damit sind alle Punkte in $\mathbb{R}^n$ reguläre Punkte bezüglich der Nebenbedingung $e(x) = 0$. Ist also $x^* \in \mathbb{R}^n$ eine lokale Lösung von $(\mathbf{QP}_{Gl})$, so existiert ein Lagrange-Multiplikator $\lambda^* \in \mathbb{R}^m$ mit

$$\nabla q(x^*) + (\lambda^*)^T \nabla e(x^*) = 0.$$

Speziell für $(\mathbf{QP}_{Gl})$ bekommen wir daher die notwendige Bedingung:

$$Qx^* + d + A^T \lambda^* = 0.$$

Mit der Gleichungs-Nebenbedingung erhalten wir das lineare System

$$\begin{pmatrix} Q & A^T \\ A & 0 \end{pmatrix} \begin{pmatrix} x^* \\ \lambda^* \end{pmatrix} = \begin{pmatrix} -d \\ b \end{pmatrix}. \tag{3.1}$$

Da (3.1) die notwendigen Bedingungen erster Ordung darstellen, wird die Koeffizienten-Matrix in (3.1) auch *KKT-Matrix* bezeichnet, wobei die Abkürzung KKT für *Karush-Kuhn-Tucker* steht (vergleiche Satz 1.18).

Mit $x^* = x + \Delta x$, $e(x) = Ax - b$ und $\nabla q(x) = (Qx + d)^T$ lässt sich (3.1) auch in der Form

$$\begin{pmatrix} Q & A^T \\ A & 0 \end{pmatrix} \begin{pmatrix} \Delta x \\ \lambda^* \end{pmatrix} = - \begin{pmatrix} \nabla q(x)^T \\ e(x) \end{pmatrix}. \tag{3.2}$$

Wir erhalten bei der Wahl eines beliebigen $x \in \mathbb{R}^n$ die Lösung (x^*, λ^*) von (3.1), indem wir (3.2) lösen und dann $x^* = x + \Delta x$ setzen.

Um hinreichende Bedingungen für die Invertierbarkeit der KKT-Matrix anzugeben, führen wir eine Matrix $Z \in \mathbb{R}^{n \times (n-m)}$ ein, deren Spalten eine Basis für $\operatorname{Kern} \nabla e(x)$ bilden. Damit gilt

$$AZ = 0 \in \mathbb{R}^{m \times (n-m)}. \tag{3.3}$$

LEMMA 3.1. *Die Matrix A habe vollen Rang m, und Z^TQZ sei positiv definit. Dann ist die KKT-Matrix*

$$K = \begin{pmatrix} Q & A^T \\ A & 0 \end{pmatrix}$$

invertierbar. Insbesondere existiert ein eindeutiges Paar (x^*, λ^*), welches (3.1) *löst.*

BEWEIS. Der Beweis folgt bereits aus Lemma 1.15. Wir wollen ihn aber hier noch einmal mit etwas anderen Argumenten durchführen. Seien $(x, \lambda) \in \mathbb{R}^n \times \mathbb{R}^m$ beliebig gewählt mit

$$\begin{pmatrix} Q & A^T \\ A & 0 \end{pmatrix} \begin{pmatrix} x \\ \lambda \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}. \tag{3.4}$$

Aus der zweiten Blockzeile in (3.4) erhalten wir $Ax = 0$, das heißt, $x \in \operatorname{Kern} A$. Damit existiert ein $w \in \mathbb{R}^{n-m}$, so dass wir x in der Form $x = Zw$ schreiben können. Aus $Ax = 0$ folgt $x^TA^T = 0$. Daher führt (3.4) auf die skalare Gleichung

$$\begin{aligned} 0 &= \begin{pmatrix} x \\ \lambda \end{pmatrix}^T \begin{pmatrix} Q & A^T \\ A & 0 \end{pmatrix} \begin{pmatrix} x \\ \lambda \end{pmatrix} = \begin{pmatrix} x \\ \lambda \end{pmatrix}^T \begin{pmatrix} Qx + A^T\lambda \\ 0 \end{pmatrix} = x^TQx \\ &= w^TZ^TQZw. \end{aligned}$$

Nach Voraussetzung ist Z^TQZ positiv definit, also muss $w = 0$ gelten. Damit ist auch $x = Zw = 0$. Es bleibt also nur noch zu zeigen, dass auch $\lambda = 0$ gilt. Aus der ersten Blockzeile in (3.4) ergibt sich mit $x = 0$ die Gleichung $A^T\lambda = 0$. Da A vollen Rang besitzt, ist A surjektiv. Deshalb ist die Matrix A^T injektiv. Das bedeutet aber $\lambda = 0$. □

BEMERKUNG 3.2. Die Matrix Z^TQZ wird *reduzierte Hesse-Matrix* genannt. Wir werden später noch auf diese Matrix zurückkommen. ◊

BEISPIEL 3.3. Wir betrachten das Problem

$$\min q(x) \quad \text{u.d.N.} \quad x_1 + x_3 = 3,\ x_2 + x_3 = 0$$

mit $x = (x_1, x_2, x_3) \in \mathbb{R}^3$ und

$$q(x) = 3x_1^2 + 2x_1x_2 + x_1x_3 + \frac{5}{2}\,x_2^2 + 2x_2x_3 + 2x_3^2 - 8x_1 - 3x_2 - 3x_3.$$

Zuerst schreiben wir das Problem in der Form $(\mathbf{QP}_{Gl})$. Wir setzen daher $n = 3$, $m = 2$ und

$$Q = \begin{pmatrix} 6 & 2 & 1 \\ 2 & 5 & 2 \\ 1 & 2 & 4 \end{pmatrix},\ d = \begin{pmatrix} -8 \\ -3 \\ -3 \end{pmatrix},\ A = \begin{pmatrix} 1 & 0 & 1 \\ 0 & 1 & 1 \end{pmatrix},\ b = \begin{pmatrix} 3 \\ 0 \end{pmatrix}.$$

Eine Basis von Kern A ist gegeben durch $Z = (-1, -1, 1)^T \in \mathbb{R}^{3\times 1}$. Dann folgt $Z^TQZ = 13 > 0$. Nach Lemma 3.1 existiert genau eine Lösung (x^*, λ^*) von (3.1), und zwar

$$x^* = \begin{pmatrix} 2 \\ -1 \\ 1 \end{pmatrix} \quad \text{und} \quad \lambda^* = \begin{pmatrix} 3 \\ -2 \end{pmatrix}.$$

In diesem Beispiel ist Q selbst positiv definit. ◊

Ist (x^*, λ^*) eine Lösung von (3.1) unter den Voraussetzungen von Lemma 3.1, so gelten auch die hinreichenden Bedingungen zweiter Ordnung, das heißt, x^* ist ein striktes lokales Minimum von $(\mathbf{QP}_{Gl})$. Wir können diese Tatsache aber auch auf einem anderen, direktem Weg beweisen.

Satz 3.4. *Es seien die Voraussetzungen von Lemma* 3.1 *erfüllt. Dann ist die eindeutige Lösung x^* von* (3.1) *auch eine eindeutige globale Lösung von* $(\mathbf{QP}_{Gl})$.

Beweis. Sei $x \in \mathbb{R}^n$ ein zulässiger Punkt, das heißt, es gilt $Ax = b$. Ferner sei $\Delta x = x^* - x$. Dann folgen $Ax^* = Ax = b$ und daher $A\Delta x = 0$. Somit liegt Δx im Kern von A. Ferner erhalten wir

$$\begin{aligned} q(x) &= \frac{1}{2}\,(x^* - \Delta x)^T Q(x^* - \Delta x) + d^T(x^* - \Delta x) \\ &= \frac{1}{2}\,\Delta x^T Q \Delta x - \Delta x^T Q x^* - d^T \Delta x + q(x^*). \end{aligned} \tag{3.5}$$

Aus (3.1) schließen wir $Qx^* = -d - A^T\lambda^*$. Also bekommen wir mit $\Delta x \in \operatorname{Kern} A$

$$\Delta x^T Q x^* = \Delta x^T \big(-d - A^T\lambda^* \big) = -\Delta x^T d.$$

Einsetzen in (3.5) liefert

$$q(x) = \frac{1}{2}\,\Delta x^T Q \Delta x + q(x^*).$$

Wegen $\Delta x \in \operatorname{Kern} A$ ergibt sich die Darstellung $\Delta x = Zu$ für ein $u \in \mathbb{R}^{n-m}$, wobei die Spalten der Matrix $Z \in \mathbb{R}^{n\times(n-m)}$ eine Basis des Kerns von A bilden. Damit lässt sich q an der Stelle x schreiben als

$$q(x) = \frac{1}{2}\,u^T Z^T Q Z u + q(x^*).$$

Da Z^TQZ positiv definit ist, gilt $q(x) > q(x^*)$ für alle $u \in \mathbb{R}^{n-m} \setminus \{0\}$ und daher für alle $x \in \mathbb{R}^n \setminus \{x^*\}$ mit $Ax = b$. Das bedeutet, dass x^* das eindeutige globale Minimum von $(\mathbf{QP}_{Gl})$ bezeichnet. □

BEMERKUNG 3.5. Wenn die reduzierte Hesse-Matrix nicht-positive Eigenwerte hat, so besitzt $(\mathbf{QP}_{Gl})$ keine beschränkte Lösung, ausgenommen in einem Spezialfall. Angenommen, (x^*, λ^*) lösen (3.1). Sei $u \in \mathbb{R}^{n-m}$ ein Vektor mit $u^T Z^T Q Z u \leq 0$. Wir setzen $\Delta x = Zu$. Dann folgt für alle $\alpha > 0$

$$A(x^* + \alpha \Delta x) = b,$$

so dass $x^* + \alpha \Delta x$ für alle $\alpha > 0$ zulässig ist, aber

$$q(x^* + \alpha \Delta x) = q(x^*) + \alpha \Delta x^T \big(Qx^* + d\big) + \frac{\alpha^2}{2} \Delta x^T Q \Delta x = q(x^*) + \frac{\alpha^2}{2} \Delta x^T Q \Delta x$$

gilt, wobei wir die Beziehungen $Qx^* + d = -A^T \lambda^*$ und $\Delta x^T A^T \lambda^* = u^T Z^T A^T \lambda^* = 0$ genutzt haben. Damit können wir zu jedem x^*, das die KKT-Bedingungen (3.1) erfüllt, eine Richtung Δx finden, in die q nicht wächst. Es existiert sogar im Fall, wenn $Z^T Q Z$ mindestens einen negativen Eigenwert besitzt, eine Richtung, in die q sogar streng monoton fallend ist. Der einzige Fall, in dem $(\mathbf{QP}_{Gl})$ eine Lösung besitzt, tritt ein, wenn $Z^T Q Z$ positiv semidefinit ist. Aber dann ist x^* auch kein striktes lokales Minimum. ◊

2. Lösung des KKT-Systems

Zunächst wollen wir bemerken, dass im Fall $m \geq 1$ die KKT-Matrix stets indefinit ist. Es gilt sogar das folgende Resultat (ohne Beweis):

LEMMA 3.6. *Die Matrix A habe vollen Rang m, und $Z^T Q Z$ sei positiv definit. Dann hat die nach Lemma* 3.1 *reguläre KKT-Matrix genau n positive und m negative Eigenwerte.*

Wir wollen hier zwei Methoden zum Lösen des KKT-Systems besprechen, die im Englischen mit *range space method* und *null space method* bezeichnet werden.

Range space method. Ist Q symmetrisch und positiv definit, so können wir folgende Blockelimination beim KKT-System durchführen: Wir multiplizieren die erste Zeile von (3.2) mit AQ^{-1} von links und erhalten

$$AQ^{-1} Q \Delta x + AQ^{-1} A^T \lambda^* = -AQ^{-1} \nabla q(x)^T.$$

Subtraktion der zweiten Zeile von (3.2) führt auf

$$AQ^{-1} A^T \lambda^* = -AQ^{-1}(Qx + d) + e(x) = e(x) - Ax - AQ^{-1} d. \tag{3.6}$$

Offenbar ist die Matrix $AQ^{-1}A^T \in \mathbb{R}^{m \times m}$ symmetrisch und positiv definit, da wir vorausgesetzt haben, dass Q symmetrisch und positiv definit ist. Damit können wir das lineare Gleichungs-System (3.6) mit dem CG-Verfahren oder mit Hilfe der Cholesky-Zerlegung lösen. Ist λ^* berechnet, so bekommen wir für Δx das System

$$Q \Delta x = -(Qx + d) - A^T \lambda^*$$

Auch hier können wir das CG-Vefahren oder die Cholesky-Zerlegung zur Bestimmung von Δx verwenden.

Erforderlich ist bei der Anwendung der Range-Space-Methode die Realisierung von Q^{-1}. Daher wird dieses Verfahren zur Lösung des KKT-Systems angewendet, wenn

- Q gut konditioniert ist,
- Q^{-1} ohne viel Aufwand zu invertieren ist, explizit bekannt ist oder durch Quasi-Newton Updates approximiert wird,
- im Fall von $m \ll n$.

Null space method. Für diese Strategie ist $\det Q \neq 0$ nicht erforderlich, so dass dieses Verfahren im allgemeinen öfter angewendet werden kann. Vorausgesetzt werden die Annahmen von Lemma 3.1: Rang $A = m$ und $Z^T QZ$ ist positiv definit, wobei $Z \in \mathbb{R}^{n\times(n-m)}$ eine Matrix ist, deren Spalten eine Basis des Nullraums von A bilden. Wir schreiben Δx in

$$\begin{pmatrix} Q & A^T \\ A & 0 \end{pmatrix} \begin{pmatrix} \Delta x \\ \lambda^* \end{pmatrix} = - \begin{pmatrix} Qx + d \\ Ax - b \end{pmatrix} \tag{3.7}$$

in der Form

$$\Delta x = Y \Delta x_Y + Z \Delta x_Z, \tag{3.8}$$

wobei wir die Matrix Z bereits eingeführt haben und $Y \in \mathbb{R}^{n\times m}$ derart gewählt ist, dass die zusammengesetzte Matrix $[Y|Z] \in \mathbb{R}^{n\times n}$ regulär ist. Ferner gelten $\Delta x_Y \in \mathbb{R}^m$ und $\Delta x_Z \in \mathbb{R}^{n-m}$. Offenbar bilden die Spalten von Y eine Basis von Bild A^T. Weiter bekommen wir $A[Y|Z] = [AY|0]$, $AY \in \mathbb{R}^{m\times m}$ und Rang $(AY) = m$. Wir schließen aus der zweiten Blockzeile in (3.7)

$$(AY)\Delta x_Y = -(Ax - b). \tag{3.9}$$

Das System (3.9) besitzt genau eine Lösung $\Delta x_Y \in \mathbb{R}^m$. Wir setzen nun die Zerlegung (3.8) in die erste Blockzeile von (3.7) ein:

$$QY\Delta x_Y + QZ\Delta x_Z + A^T\lambda^* = -Qx - d.$$

Multiplikation mit Z^T von links führt wegen $AZ = 0 \in \mathbb{R}^{m\times(n-m)}$ auf

$$\begin{aligned} (Z^T QZ)\Delta x_Z &= -Z^T QY\Delta x_Y - Z^T A^T\lambda^* - Z^T(Qx + d) \\ &= -Z^T\big(QY\Delta x_Y + Qx + d\big). \end{aligned} \tag{3.10}$$

Unter den Voraussetzungen von Lemma 3.1 ist die Matrix $Z^T QZ$ positiv definit. Daher können wir zur Berechnung von Δx_Z aus (3.10) das CG-Verfahren oder die Cholesky-Faktorisierung verwendet werden. Damit ist Δx aus (3.8) mittels (3.9) und (3.10) berechenbar. Multiplizieren wir die erste Blockzeile in (3.7) mit Y^T von links, so erhalten wir das lineare System

$$(AY)^T\lambda^* = -Y^T(Qx + d + Q\Delta x). \tag{3.11}$$

Wegen $\det(AY) \neq 0$ ist λ^* durch (3.11) eindeutig bestimmt.

Beispiel 3.7. Wir betrachten das Problem von Beispiel 3.3. Als Matrix Z wählen wir $Z = (-1, -1, 1)^T$. Damit können wir die Matrix

$$Y = \frac{1}{3}\begin{pmatrix} 2 & -1 \\ -1 & 2 \\ 1 & 1 \end{pmatrix}$$

wählen. Es folgt

$$AY = \frac{1}{3}\begin{pmatrix} 1 & 0 & 1 \\ 0 & 1 & 1 \end{pmatrix}\begin{pmatrix} 2 & -1 \\ -1 & 2 \\ 1 & 1 \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}.$$

Wir wählen $x = 0$ in (3.7). Damit folgen

$$e(x) = Ax - b = -b = \begin{pmatrix} -3 \\ 0 \end{pmatrix} \quad \text{und} \quad \nabla q(x)^T = Qx + d = d = \begin{pmatrix} -8 \\ -3 \\ -3 \end{pmatrix}.$$

Aus (3.9) bekommen wir $\Delta x_Y = b = (3,0)^T$. Wir berechnen die reche Seite von (3.10):

$$\begin{aligned}
&-Z^TQY\begin{pmatrix}3\\0\end{pmatrix} - Z^T\begin{pmatrix}-8\\-3\\-3\end{pmatrix}\\
&\quad= \left(\frac{1}{3},\frac{1}{3},-\frac{1}{3}\right)\begin{pmatrix}6&2&1\\2&5&2\\1&2&4\end{pmatrix}\begin{pmatrix}2&-1\\-1&2\\1&1\end{pmatrix}\begin{pmatrix}3\\0\end{pmatrix} - (1,1,-1)\begin{pmatrix}-8\\-3\\-3\end{pmatrix}\\
&\quad= \left(\frac{7}{3},\frac{5}{3},-\frac{1}{3}\right)\begin{pmatrix}2&-1\\-1&2\\1&1\end{pmatrix}\begin{pmatrix}3\\0\end{pmatrix} + 8 = (7,5,-1)\begin{pmatrix}2\\-1\\1\end{pmatrix} = 0.
\end{aligned}$$

Damit ist die Lösung von (3.10) durch $\Delta x_Z = 0$ gegeben. Also folgt

$$\Delta x = Y\Delta x_Y + Z\Delta x_Z = Y\begin{pmatrix}3\\0\end{pmatrix} + \begin{pmatrix}-1\\-1\\1\end{pmatrix}0 = \begin{pmatrix}2\\-1\\1\end{pmatrix}.$$

Nun zur Berechnung von λ^* gemäß (3.11): Wegen $AY = I$ folgt

$$\lambda^* = -Y^T\begin{pmatrix}-8\\-3\\-3\end{pmatrix} - Y^TQ\begin{pmatrix}2\\-1\\1\end{pmatrix} = \begin{pmatrix}-3\\2\end{pmatrix}.$$

Wegen $x^* = x + \Delta x$ bekommen wir wegen $x = 0$ die Lösung

$$x^* = \begin{pmatrix}2\\-1\\1\end{pmatrix} \quad \text{und} \quad \lambda^* = \begin{pmatrix}-3\\2\end{pmatrix}.$$

als Lösung von (3.7). ◊

Ist $n - m$ klein, so ist die Null-Space-Methode oft sehr effizient. Allerdings ist die Berechnung von Z notwendig. Die Matrix Z ist nicht eindeutig bestimmt und die Matrix Z^TQZ eine schlecht konditionierte Matrix. Sind allerdings die Spalten von Z orthonormal, so folgt für die Konditionszahl $\kappa_2(Z^TQZ) = \kappa_2(Q)$.

3. Ungleichungsrestringierte Probleme

Wir wollen die Optimalitätsbedingungen für das Problem (**QP**). Dazu setzen wir

$$e(x) = Ax - b \quad \text{und} \quad g(x) = r - Cx,$$

wobei

$$A = \begin{pmatrix}a_1^T\\\vdots\\a_m^T\end{pmatrix} \in \mathbb{R}^{m\times n},\ C = \begin{pmatrix}a_{m+1}^T\\\vdots\\a_{m+p}^T\end{pmatrix} \in \mathbb{R}^{p\times n}$$

und

$$b = \begin{pmatrix}b_1\\\vdots\\b_m\end{pmatrix} \in \mathbb{R}^m,\ r = \begin{pmatrix}b_{m+1}\\\vdots\\b_{m+p}\end{pmatrix} \in \mathbb{R}^p$$

gelten. Die Lagrange-Funktion ist

$$L(x,\lambda,\mu) = \frac{1}{2}x^T Qx + x^T d + \langle Ax-b,\lambda\rangle_{\mathbb{R}^m} + \langle r - Cx,\mu\rangle_{\mathbb{R}^p}.$$

Die KKT-Bedingungen lauten daher

$$Qx^* + d + A^T\lambda^* - C^T\mu^* = 0 \qquad \text{in } \mathbb{R}^n, \tag{3.12a}$$

$$Ax^* = b \qquad \text{in } \mathbb{R}^m, \tag{3.12b}$$

$$Cx^* \geq r \qquad \text{in } \mathbb{R}^p, \tag{3.12c}$$

$$\mu^* \geq 0 \qquad \text{in } \mathbb{R}^p, \tag{3.12d}$$

$$(\mu^*)^T(r - Cx^*) = 0. \tag{3.12e}$$

Für konvexe quadratische Optimierungsprobleme, wenn also Q positive semidefinite ist, sind die notwendigen Optimalitätsbedingungen bereits hinreichend dafür, dass x^* eine globale Lösung von (**QP**) ist.

SATZ 3.8. *Wenn x^* die Bedingungen* (3.12) *erfüllt zusammen mit einem $\lambda^* \in \mathbb{R}^m$ und einem $\mu^* \in \mathbb{R}^p$ mit $\mu^* \geq 0$ und wenn Q positiv semidefinit auf* $\operatorname{Ker}\nabla e(x^*)$ *ist, dann ist x^* eine globale Lösung von* (**QP**).

PROOF. Sei x ein zulässiger Punkt für (**QP**). Dann gelten $Ax = b$ in $\mathbb{R}^m$ und $Cx \geq r$ in $\mathbb{R}^p$. Wir setzen $\Delta x = x - x^*$. Dann folgen $A\Delta x = 0$ und

$$\big(C\Delta x\big)_i = \big(Cx - Cx^*\big)_i \geq r_i - \big(Cx^*\big)_i = 0 \quad \text{für alle } i \in \mathcal{A}(x^*),$$

wobei $A(x^*) \subset \{1,\dots,p\}$ die Menge der an x^* aktiven Indizes bezeichnet. Es gilt weiter $\mu_i^* = 0$ für alle $i \in \mathcal{I}(x^*) = \{1,\dots,p\} \setminus \mathcal{A}(x^*)$. Zusammen mit (3.12a) und (3.12d) erhalten wir

$$\begin{aligned}\Delta x^T(Qx^* + d) &= -\Delta x^T A^T\lambda^* + \Delta x^T C^T\mu^* = \Delta x^T C^T\mu^* \\ &= \sum_{i\in\mathcal{A}(x^*)} \big(C\Delta x\big)_i\mu_i^* + \sum_{i\in\mathcal{I}(x^*)} \big(C\Delta x\big)_i\mu_i^* \geq 0.\end{aligned} \tag{3.13}$$

Da Q positiv semidefinit auf $\operatorname{Ker}\nabla e(x^*)$ ist, schließen wir aus (3.13), dass

$$q(x) = q(x^*) + \Delta x^T(Qx^* + d) + \frac{1}{2}\Delta x^T Q\Delta x \geq q(x^*) + \frac{1}{2}\Delta x^T Q\Delta x \geq q(x^*),$$

so dass x^* eine globale Lösung von (**QP**) ist. □

BEMERKUNG 3.9.
1) Eine kleine Modifikation des Beweises von Satz 3.8 zeigt, dass x^* die eindeutige, globale Lösung von (**QP**) ist, wenn Q positiv definit auf $\operatorname{Ker}\nabla e(x^*)$ ist.
2) Wenn Q nicht positiv definit auf $\operatorname{Ker}\nabla e(x^*)$ ist, dann kann (**QP**) mehrere Lösungen besitzen.

Verfahren zum Lösen der KKT-Bedingungen sind zum Beispiel *aktive Mengenstrategien*, *projezierte Gradienten-Verfahren* oder *Innere Punkte Methoden.*

4. Innere-Punkte Verfahren für Quadratische Programmierung

Wir betrachten

$$\min q(x) = \frac{1}{2}\,x^T Qx + x^T d \quad \text{u.d.N.} \quad Ax \geq b \tag{3.14}$$

wobei Q symmetrisch und positiv semi-definit ist und $d \in \mathbb{R}^n$, $A \in \mathbb{R}^{m\times n}$ sowie $b \in \mathbb{R}^m$ gelten.

Notwendige Optimalitäts-Bedingungen erster Ordnung: Sei x^* eine Lösung von (3.14). Dann löst (x^*, μ^*), μ^* der assozierte Lagrange-Multiplikator, das System

$$\begin{aligned} Qx - A^T\mu + d &= 0 && \text{in } \mathbb{R}^n,\\ b - Ax &\le 0 && \text{in } \mathbb{R}^m,\\ (b - Ax)_i\mu_i &= 0 && \text{für } i = 1,\dots,m,\\ \mu &\ge 0 && \text{in } \mathbb{R}^m. \end{aligned}$$

Mit Einführung der *Slack-Variablen* $y = Ax - b \in \mathbb{R}^m$ folgt

$$Qx - A^T\mu + d = 0 \quad \text{in } \mathbb{R}^n, \tag{3.15a}$$

$$b - Ax + y = 0 \quad \text{in } \mathbb{R}^m, \tag{3.15b}$$

$$y_i\mu_i = 0 \quad \text{für } i = 1,\dots,m, \tag{3.15c}$$

$$y \ge 0 \quad \text{in } \mathbb{R}^m, \tag{3.15d}$$

$$\mu \ge 0 \quad \text{in } \mathbb{R}^m. \tag{3.15e}$$

Das System (3.15) ist auch hinreichend, da die Zielfunktion und die Menge der zulässigen Punkte konvex sind. Wie in Abschnitt 2 schreiben wir (3.15) in der folgenden Form

$$F(x,y,\mu) = \begin{pmatrix} Qx - A^T\mu + d \\ b - Ax + y \\ YMe \end{pmatrix} = 0 \quad \text{und} \quad (y,\mu) \ge 0,$$

wobei

$$Y = \operatorname{diag}(y_1,\dots,y_m),\ M = \operatorname{diag}(\mu_1,\dots,\mu_m) \quad \text{und} \quad e = (1,\dots,1)^T \in \mathbb{R}^m$$

Sei $(x,y,\mu) \in \mathbb{R}^n \times \mathbb{R}^m \times \mathbb{R}^m$ eine aktuelle Iterierte. Dann ist die gewichtete Dualitätslücke η definiert durch

$$\eta = \frac{1}{m}\sum_{i=1}^{m} y_i\mu_i = \frac{y^T\mu}{m}.$$

Der zentrale Pfad $\mathcal{C}$ besteht aus der Menge von Punkten $(x_\tau, y_\tau, \mu_\tau)$, $\eta > 0$, so dass

$$F(x,y,\mu) = \begin{pmatrix} 0 \\ 0 \\ \tau e \end{pmatrix} \quad \text{und} \quad (y_\tau,\mu_\tau) > 0$$

gilt.

Ein Newton-Schritt, ausgehend von (x,y,μ) auf den Punkt $(x_{\sigma\eta}, y_{\sigma\eta}, \mu_{\sigma\eta}) \in \mathcal{C}$ zu mit $\sigma \in [0,1]$, genügt dem lineaen Gleichungs-System

$$\begin{pmatrix} Q & 0 & -A^T \\ -A & I & 0 \\ 0 & M & Y \end{pmatrix} \begin{pmatrix} \Delta x \\ \Delta y \\ \Delta\mu \end{pmatrix} = \begin{pmatrix} -r_d(x,\mu) \\ -r_b(x,y) \\ -YMe + \sigma\eta e \end{pmatrix}, \tag{3.16}$$

wobei

$$r_d(x,\mu) = Qx - A^T\mu + d \quad \text{und} \quad r_b(x,y) = b - Ax + y \tag{3.17}$$

gelten. Die nächste Iterierte ist dann für $\alpha \in (0,1]$ gegeben durch

$$(x^+, y^+, \mu^+) = (x,y,\mu) + \alpha(\Delta x, \Delta y, \Delta\mu), \tag{3.18}$$

so dass $(y^+, \mu^+) > 0$ erfüllt ist.

Lösung des primal-dualen Systems. Der Hauptaufwand der Inneren-Punkte-Verfahren besteht in der Regel in der Lösung des Systems (3.16). Aufgrund der Hessematrix Q in der Koeffizientenmatrix kann die Lösung von (3.16) viel aufwendiger sein als die Lösung des KKT-System in Abschnitt 2 im Zusammenhang mit Inneren-Punkte-Verfahren in der Linearen Programmierung. Daher ist es wichtig, die spezielle Struktur von (3.16) auszunutzen, indem wir eine geeignete direkte Zerlegung oder einen passenden Vorkonditionierer im Kontext von iterativen Verfahren verwenden.

Aus der dritten Blockzeile von (3.16) erhalten wir

$$\begin{aligned}\Delta y &= M^{-1}\big(-MYe + \sigma\eta e - Y\Delta\mu\big) = -Ye + \sigma\eta M^{-1}e - M^{-1}Y\Delta\mu \\ &= -y + \sigma\eta M^{-1}e - M^{-1}Y\Delta\mu.\end{aligned}$$

Einsetzen von Δy in die zweite Blockzeile von (3.16) ergibt

$$\begin{aligned}A\Delta x - \Delta y &= A\Delta x + y - \sigma\eta M^{-1}e + M^{-1}Y\Delta\mu \\ &= \big(A \,|\, M^{-1}Y\big)\begin{pmatrix}\Delta x\\ \Delta\mu\end{pmatrix} - \big(-y + \sigma\eta M^{-1}e\big).\end{aligned}$$

Damit können wir das System (3.16) in der Form

$$\begin{pmatrix}Q & -A^T\\ A & M^{-1}Y\end{pmatrix}\begin{pmatrix}\Delta x\\ \Delta\mu\end{pmatrix} = \begin{pmatrix}-r_d(x,\mu)\\ -r_b(x,y) + \big(-y + \sigma\eta M^{-1}e\big)\end{pmatrix}. \tag{3.19}$$

Mit der zweiten Blockzeile in (3.19) erhalten wir

$$\Delta\mu = Y^{-1}M\big(-r_b - y + \sigma\eta M^{-1}e - A\Delta x\big)$$

so dass die erste Blockzeile von (3.19) auf das System

$$\big(Q + A^TY^{-1}MA\big)\Delta x = -r_d + A^TY^{-1}M\big(-r_b(x,y) - y + \sigma\eta M^{-1}e\big). \tag{3.20}$$

Das System (3.20) kann mit einem (modifizierten) Cholesky-Verfahren (siehe [**29**]) gelöst werden. Diese Vorgangsweise ist insbesondere geeignet, wenn die Matrix $A^TY^{-1}MA$ im Vergleich zur Matrix Q nicht so dicht besetzt ist und das System (3.20) deutlich kleiner als das in (3.19) ist. Als iteratives Verfahren kann ein projeziertes CG-Verfahren eingesetzt werden (siehe [**29**]), in dem nur Matrix-Vektor-Produkte auszuwerten sind.

Das System (3.16) kann auch in der Form

$$\begin{pmatrix}Q & 0 & -A^T\\ 0 & M & Y\\ A & -I & 0\end{pmatrix}\begin{pmatrix}\Delta x\\ \Delta y\\ \Delta\mu\end{pmatrix} = \begin{pmatrix}-r_d(x,\mu)\\ -YMe + \sigma\eta e\\ r_b(x,y)\end{pmatrix}$$

geschrieben werden. Das sind aber die KKT-Bedingungen für das konvexe, quadratische Optimierungsproblem

$$\begin{aligned}\min\ &\begin{pmatrix}\Delta x\\ \Delta y\end{pmatrix}^T\begin{pmatrix}Q & 0\\ 0 & Y^{-1}M\end{pmatrix}\begin{pmatrix}\Delta x\\ \Delta y\end{pmatrix} + \begin{pmatrix}\Delta x\\ \Delta y\end{pmatrix}^T\begin{pmatrix}r_d(x,\mu)\\ YMe - \sigma\eta e\end{pmatrix}\\ \text{u.d.N. }&\big(A \,|\, -I\big)\begin{pmatrix}\Delta x\\ \Delta y\end{pmatrix} = r_b(x,y),\end{aligned}$$

welches unter Verwendung geeigneter Optimierungsverfahren gelöst werden kann, zum Beispiel mit dem projezierten CG-Verfahren.

Schrittlängen-Bestimmung. Innere Punkte-Verfahren für die Lineare programmierung sind effizienter, wenn für die primalen und dualen Variablen unterschiedliche Schrittweiten-Parameter (α^{pri} beziehungsweise α^{dual}) verwendet werden.

Seien die nächsten Iterierten durch

$$(x^+, y^+) = (x, y) + \alpha^{\text{pri}}(\Delta x, \Delta y) \quad \text{und} \quad \mu^+ = \mu + \alpha^{\text{dual}}\Delta\mu, \tag{3.21}$$

wobei $\alpha^{\text{pri}} > 0$ und $\alpha^{\text{dual}} > 0$ so gewählt sind dass $(y^+, \mu^+) > 0$ erfüllt ist. Aus (3.16) und (3.17) folgen

$$\begin{aligned} r_b^+ &= A\big(x + \alpha^{\text{pri}}\Delta x\big) - \big(y + \alpha^{\text{pri}}\Delta y\big) - b = r_b(x,y) + \alpha^{\text{pri}}\big(A\Delta x - \Delta y\big) \\ &= \big(1 - \alpha^{\text{pri}}\big) r_b(x,y) \end{aligned} \tag{3.22a}$$

sowie

$$\begin{aligned} r_d^+ &= Q\big(x + \alpha^{\text{pri}}\Delta x\big) - A^T\big(\mu + \alpha^{\text{dual}}\Delta\mu\big) + d \\ &= r_d + \alpha^{\text{pri}} Q\Delta x - \alpha^{\text{dual}} A^T\Delta\mu \\ &= r_d + \alpha^{\text{dual}}\big(Q\Delta x - A^T\Delta\mu\big) + \big(\alpha^{\text{pri}} - \alpha^{\text{dual}}\big)Q\Delta x \\ &= \big(1 - \alpha^{\text{dual}}\big) r_d + \big(\alpha^{\text{pri}} - \alpha^{\text{dual}}\big)Q\Delta x. \end{aligned} \tag{3.22b}$$

Gilt $\alpha^{\text{pri}} = \alpha^{\text{dual}} = \alpha$, so fallen beide Residuen linear für alle $\alpha \in (0,1)$. Allerdings für unterschiedliche Schrittweiten α^{pri} und α^{dual} kann $\|r_d^+\|_2$ anwachsen, so dass das Innere-Punkte-Verfahren divergiert. Eine Möglichkeit besteht darin Schrittweiten gemäß (3.18) zu wählen, wobei wir $\alpha = \min\{\alpha_\tau^{\text{pri}}, \alpha_\tau^{\text{dual}}\}$ setzen mit

$$\begin{aligned} \alpha_\tau^{\text{pri}} &= \max\big\{\alpha \in (0,1) \,|\, \mu + \alpha\Delta\mu \geq (1-\tau)y\big\}, \\ \alpha_\tau^{\text{dual}} &= \max\big\{\alpha \in (0,1) \,|\, \mu + \alpha\Delta\mu \geq (1-\tau)\mu\big\}. \end{aligned} \tag{3.23}$$

Dabei steuert $\tau \in (0,1)$, wie weit wir von dem maximalen Schritt, der die Bedingungen $\mu + \alpha\Delta\mu$ und $\mu + \alpha\Delta\mu$ erfüllt, (relativ) entfernt sind.

Die numerische Erfahrung hat allerdings gezeigt, dass die Wahl unterschiedlicher Schrittweiten für die primalen und dualen Variablen oft zu schnellerer Konvergenz der Innere-Punkte-Verfahrens führt. Eine Möglichkeit zur Wahl unterschiedlicher Schrittweiten is, $(\alpha^{\text{pri}}, \alpha^{\text{dual}})$ als (näherungsweise) Lösung der Minimierungsaufgabe

$$\begin{aligned} &\min \|Qx^+ - A^T\mu^+ + d\|_2^2 + \|Ax^+ - y^+ - b\|_2^2 + \big(y^+\big)^T\mu^+ \\ &\text{u.d.N. } 0 \leq \alpha^{\text{pri}} \leq \alpha_\tau^{\text{pri}},\ 0 \leq \alpha^{\text{dual}} \leq \alpha_\tau^{\text{dual}} \text{ und } (x^+, y^+, \mu^+) \text{ gemäß (3.18).} \end{aligned}$$

Ein praktischer Primal-dualer Algorithmus. Die am meisten verwendete Variante des Innere-Punkte-Verfahrens basiert auf dem Pädiktor-Korrektor Algorithmus von Mehrotra; vergleiche Abschnitt 4. Zuerst wird ein affiner Skalierungsschritt $(\Delta x^{\text{aff}}, \Delta y^{\text{aff}}, \Delta\mu^{\text{aff}})$ bestimmt, indem in (3.16) mit $\sigma = 0$. Die erhaltene Richtung wird dann in einem nachfolgenden Korrekturschritt verbessert, wobei $\sigma = (\eta^{\text{aff}}/\eta)^3$ gesetzt wird. Insgesamt lösen wir im Korrekturschritt das System

$$\begin{pmatrix} Q & 0 & -A^T \\ A & I & 0 \\ 0 & M & Y \end{pmatrix} \begin{pmatrix} \Delta x \\ \Delta y \\ \Delta\mu \end{pmatrix} = \begin{pmatrix} -r_d(x,\mu) \\ -r_b(x,y) \\ -MYe - \Delta M^{\text{aff}}\Delta Y^{\text{aff}}e + \sigma\eta e \end{pmatrix}. \tag{3.24}$$

Algorithmus 3.10 (Prädiktor-Korrektor Verfahren für quadratische Probleme).

1) *Wähle* (x^0, y^0, μ^0) *mit* $(y^+, \mu^+) > 0$ *und setze* $k = 0$.

2) *Setze* $x, y, \mu) = (x^k, y^k, \mu^k)$ *und löse* (3.16) *für* $(\Delta x^{\text{aff}}, \Delta y^{\text{aff}}, \Delta \mu^{\text{aff}})$ *mit* $\sigma = 0$.
3) *Berechne* $\eta = y^T \mu / m$.
4) *Setze* $\hat{\alpha}^{\text{aff}} = \max\{\alpha \in (0,1] \,|\, (y, \mu) + \alpha(\Delta y^{\text{aff}}, \Delta \mu^{\text{aff}}) \geq 0\}$
5) *Bestimme* $\eta^{\text{aff}} = (y + \alpha^{\text{aff}} \Delta y)^T (\mu + \alpha^{\text{aff}} \Delta \mu)/m$ *und wähle* $\sigma = (\eta^{\text{aff}}/\eta)^3$.
6) *Löse* (3.24) *für* $(\Delta x, \Delta y, \Delta \mu)$.
7) *Wähle* $\tau_k \in (0,1)$ *und setze* $\hat{\alpha} = \min\{\alpha_{\tau_k}^{\text{pri}}, \alpha_{\tau_k}^{\text{dual}}\}$; *vergleiche* (3.23).
8) *Setze* $(x^{k+1}, y^{k+1}, \mu^{k+1}) = (x^k, y^k, \mu^k) + \hat{\alpha}(\Delta x, \Delta y, \Delta \mu)$, $k = k + 1$ *und gehe zurück zu Schritt* 2).

Im Algorithmus 3.10 können wir $t_k \to 1$ wählen, wenn die Iterierten konvergieren, um die Konvergenz zu beschleunigen. Wie im Fall der Linearen Programmierung hängt die Effizienz von Algorithmus 3.10 von der Wahl geeigneter Startwerte ab. Eine mögliche Heuristik verwendet einen gegebenen Startwert $(\bar{x}, \bar{y}, \bar{\mu})$, um diesen hinreichend weit weg von dem Rand der durch die Bedingung $(y, \mu) \geq 0$ definierten Menge zu verschieben, so dass zu Beginn von Algorithmus 3.10 große Schrittweiten möglich sind.

KAPITEL 4

SQP-Verfahren

In diesem Abschnitt werden wir uns mit einem der effizientesten Verfahren der nichtlinearen restringierten Optimierung beschäftigen, mit dem *SQP-Verfahren*. Dabei steht SQP für *sequential quadratic programming*.

1. Das lokale SQP-Verfahren

Wir betrachten

$$\textbf{(P)} \qquad \min J(x) \qquad \text{u.d.N.} \quad e(x) = 0,$$

wobei $J : \mathbb{R}^n \to \mathbb{R}$ und $e : \mathbb{R}^n \to \mathbb{R}^m$ zweimal stetig differenzierbar sind und die zweiten Ableitungen Lipschitz-stetig sind.

Wesentliche Idee des SQP-Verfahrens ist es, dass (**P**) an jeder Iterierten x^k durch ein quadratisches Modell ersetzt wird und der Minimierer dazu benutzt wird, die neue Iterierte x^{k+1} zu berechnen.

Die Lagrange-Funktion zu (**P**) lautet

$$L(x, \lambda) = J(x) + \lambda^T e(x).$$

Wir bezeichnen mit

$$A(x) = \begin{pmatrix} \nabla e_1(x) \\ \vdots \\ \nabla e_m(x) \end{pmatrix} \in \mathbb{R}^{m \times n} \tag{4.1}$$

die Jacobi-Matrix von e am Punkt x. Die KKT-Bedingungen $\nabla L(x, \lambda) = 0$ ergeben

$$F(x, \lambda) = \begin{pmatrix} \nabla J(x)^T + A(x)^T \lambda \\ e(x) \end{pmatrix} \stackrel{!}{=} 0 \tag{4.2}$$

Hat $A(x^*)$ vollen Rang m, so ist x^* ein regulärer Punkt und es existiert zu jeder Lösung $x^* \in \mathbb{R}^n$ von (**P**) ein zugehöriger Lagrange-Multiplikator $\lambda^* \in \mathbb{R}^m$ mit $F(x^*, \lambda^*) = 0$. Wir lösen (4.2) mit dem Newton-Verfahren. Die Jacobi-Matrix von F ist

$$\nabla^2 L(x, \lambda) = \begin{pmatrix} \nabla_{xx} L(x, \lambda) & A(x)^T \\ A(x) & 0 \end{pmatrix}. \tag{4.3}$$

Damit ist der Newton-Schritt von der Iterierten (x^k, λ^k) gegeben durch

$$\begin{pmatrix} x^{k+1} \\ \lambda^{k+1} \end{pmatrix} = \begin{pmatrix} x^k \\ \lambda^k \end{pmatrix} + \begin{pmatrix} \Delta x^k \\ \Delta \lambda^k \end{pmatrix}, \tag{4.4a}$$

wobei

$$\nabla^2 L(x^k, \lambda^k) \begin{pmatrix} \Delta x^k \\ \Delta \lambda^k \end{pmatrix} = -\nabla L(x^k, \lambda^k) \tag{4.4b}$$

gilt. Das Verfahren (4.4) heißt daher oft auch *Lagrange-Newton-SQP Verfahren*, da es als Newton-Verfahren in den Variablen (x^k, λ^k) interpretiert werden kann. Ist die Hesse-Matrix $\nabla^2 L(x^k, \lambda^k)$ invertierbar, so ist die Iteration (4.4) wohldefiniert. Voraussetzungen für die Regularität von $\nabla^2 L(x^k, \lambda^k)$ sind in Lemma 3.1 gegeben.

VORAUSSETZUNG 4.1. 1) *Die Jacobi-Matrix* $A_k = \nabla e(x^k)$ *hat vollen Rang.*
2) *Die Hesse-Matrix* $\nabla_{xx} L(x^k, \lambda^k)$ *ist positiv definit auf dem Kern von* $A(x^k)$.

BEMERKUNG 4.2. 1) Gilt

$$\Delta x^T \nabla_{xx} L(x^*, \lambda^*) \Delta x \geq \kappa \|\Delta x\|^2 \quad \text{für alle } \Delta x \in \operatorname{Kern} A(x^*)$$

für ein $\kappa > 0$ (hinreichende Bedingungen zweiter Ordnung), so folgt die Voraussetzung 4.1-2) in einer Umgebung von (x^*, λ^*).

2) Aufgrund der Theorie des Newton-Verfahrens ergibt sich für das SQP-Verfahren lokal quadratische Konvergenz in (x, λ), das heißt, es gibt ein $C > 0$ mit

$$\|(x^{k+1}, \lambda^{k+1}) - (x^*, \lambda^*)\| \leq C \|(x^k, \lambda^k) - (x^*, \lambda^*)\|^2 \quad \text{für alle } k \geq 0,$$

sofern $\|(x^0, \lambda^0) - (x^*, \lambda^*)\|$ hinreichend klein sind. ◊

Wir wollen nun eine andere Motivation für (4.4) geben. Dazu betrachten wir das quadratische Problem

$$\min_{\Delta x^k} \frac{1}{2} (\Delta x^k)^T \nabla_{xx} L(x^k, \lambda^k) \Delta x^k + \nabla J(x^k) \Delta x^k \tag{4.5a}$$

$$\text{u.d.N. } A_k \Delta x^k + e(x^k) = 0. \tag{4.5b}$$

Hier erkennen wir, warum das Verfahren SQP-Algorithmus heißt: Es sind in jeder Iteration quadratische Probleme zu lösen.

Die Optimalitäts-Bedingungen für das quadratische Problem (4.5) lauten

$$\nabla_{xx} L(x^k, \lambda^k) \Delta x^k + \nabla J(x^k)^T + A_k^T \mu^k = 0, \tag{4.6a}$$

$$A_k \Delta x^k = -e(x^k) \tag{4.6b}$$

mit einem Lagrange-Multiplikator $\mu^k \in \mathbb{R}^m$. Die Voraussetzung 4.1 garantiert, dass es eine eindeutige Lösung $(\Delta x^k, \mu^k)$ von (4.6) existiert. Wenn wir $A_k^T \lambda^k$ in der ersten Blockzeile auf beiden Seiten von (4.4b) addieren, so erhalten wir

$$\nabla_{xx} L(x^k, \lambda^k) \Delta x^k + A_k^T (\underbrace{\lambda^{k+1} - \lambda^k}_{=\Delta\lambda^k}) + A_k^T \lambda^k = -\nabla J(x^k)^T - A_k^T \lambda^k + A_k^T \lambda^k.$$

Insgesamt ergibt damit (4.4b)

$$\begin{pmatrix} \nabla_{xx} L(x^k, \lambda^k) & A_k^T \\ A_k & 0 \end{pmatrix} \begin{pmatrix} \Delta x^k \\ \lambda^{k+1} \end{pmatrix} = - \begin{pmatrix} \nabla J(x^k)^T \\ e(x^k) \end{pmatrix}. \tag{4.7}$$

Unter der Voraussetzung $\det(\nabla^2 L(x^k, \lambda^k)) \neq 0$ folgt $\lambda^{k+1} = \mu^k$. Damit sind das Newton- und das SQP-Verfahren äquivalent. Unter der Voraussetzung 4.1 an (x^k, λ^k) kann die nächste Iterierte (x^{k+1}, λ^{k+1}) als Lösung des quadratischen Problems (4.5) oder als Iterierte des Newton-Verfahrens (4.4) berechnet werden.

Aus der Sicht des Newton-Verfahrens lassen sich eher theoretische Resultate nachweisen (zum Beispiel die quadratische Konvergenz), während die Interpretation als SQP-Algorithmus es ermöglicht, praktische Verfahren zu entwerfen und auch Ungleichungen zu berücksichtigen.

ALGORITHMUS 4.3 (Lokales SQP-Verfahren).

1) *Wähle Startwerte* $x^0 \in \mathbb{R}^n$, $\lambda^0 \in \mathbb{R}^m$ *und* $k_{\max} \in \mathbb{N}$.
2) *For* $k = 0, 1, \ldots, k_{\max}$
   - *Berechne* $J_k = J(x^k)$, $\nabla J_k = \nabla J(x^k)$, $\nabla_{xx} L(x^k, \lambda^k)$, $e_k = e(x^k)$ *und* $A_k = \nabla e(x^k)$;
   - *Löse* (4.5) *für* $(\Delta x^k, \mu^k)$;
   - *Setze* $x^{k+1} = x^k + \Delta x^k$ *und* $\lambda^{k+1} = \mu^k$;
   - *Prüfe die Abbruchkriterien;*

   end (For).

Wir haben bereits erwähnt, dass die lokal quadratische Konvergenz in (x^k, λ^k) von Algorithmus 4.3 aus der lokalen Äquivalenz mit dem Newton-Verfahren für die Gleichung (4.2) folgt.

Das Zielfunktional in (4.5a)

$$\frac{1}{2}(\Delta x^k)^T \nabla_{xx} L(x^k, \lambda^k) \Delta x^k + \nabla J(x^k) \Delta x^k$$

können wir wegen (4.5b) durch

$$\frac{1}{2}(\Delta x^k)^T \nabla_{xx} L(x^k, \lambda^k) \Delta x^k + \nabla_x L(x^k, \lambda^k) \Delta x^k$$

ersetzen; denn es gilt

$$\begin{aligned} \nabla_x L(x^k, \lambda^k) \Delta x^k &= \nabla J(x^k) \Delta x^k + (\lambda^k)^T \nabla e(x^k) \Delta x^k \\ &= \nabla J(x^k) \Delta x^k + (\lambda^k)^T (-e(x^k)) \\ &= \nabla J(x^k) \Delta x^k - (\lambda^k)^T e(x^k) \end{aligned}$$

und der Term $-(\lambda^k)^T e(x^k)$ ist konstant, beeinflußt daher die Minimierung nicht Damit können wir (4.5) auch durch

$$\begin{aligned} &\min_{\Delta x^k} \frac{1}{2}(\Delta x^k)^T \nabla_{xx} L(x^k, \lambda^k) \Delta x^k + \nabla_x L(x^k, \lambda^k) \Delta x^k \\ &\text{u.d.N. } A_k \Delta x^k + e(x^k) = 0. \end{aligned}$$

ersetzen.

Das SQP-Verfahren kann einfach auf nichtlineare Probleme mit Ungleichungs-Nebenbedingungen erweitert werden. Wir betrachten

$$\min J(x) \quad \text{u.d.N.} \quad e(x) = 0 \text{ in } \mathbb{R}^m \text{ und } g(x) \leq 0 \text{ in } \mathbb{R}^p. \tag{4.8}$$

Zur Lösung von (4.8) linearizieren wir sowohl die Gleichungs- als auch die Ungleichungs-Nebenbedingungen. Dann erhalten wir

$$\begin{cases} \min\limits_{\Delta x^k \in \mathbb{R}^n} J_k + \nabla J_k \Delta x + \dfrac{1}{2}\left(\Delta x^k\right)^T \nabla_{xx} L(x^k, \lambda^k) \Delta x^k \\ \text{u.d.N. } \nabla e(x^k) \Delta x^k + e(x^k) = 0 \text{ in } \mathbb{R}^m, \ \nabla g(x^k) \Delta x^k + g(x^k) \leq 0 \text{ in } \mathbb{R}^p. \end{cases} \tag{4.9}$$

Nun können wir Algorithmen für quadratische Programme verwenden, z.B., das Innere-Punkte-Verfahren aus Abschnitt 4. Die neue Iterierte ist durch das Paar $(x^k + \Delta x^k, \lambda^{k+1})$ gegeben, wobei δx^k und λ^{k+1} die Lösung beziehungsweise der assoziierte Lagrange-Multiplikator von (4.9) sind. Ein lokales SQP-Verfahren für (4.8) hat damit die Form von Algorithmus 4.3 mit der Modifikation, dass der Schritt durch die Lösung von (4.9) bestimmt wird.

2. Berechnung des SQP-Schrittes

Wie im vorigen Abschnitt wollen wir uns auch hier auf Gleichungs-Restriktionen beschränken.

a) Als erste Alternative können wir das System (4.7) entweder mit einem direktem Verfahren (zum Beispiel einer LR-Zerlegung) oder einem iterativen Gleichungslöser (zum Beispiel GMRES-, QMR-, MINRES- oder SYMMLQ-Verfahren, siehe [**21**]) lösen. Die Vorkonditionierung der iterativen Verfahren ist ein aktives Forschungsgebiet.

b) Wenn $\nabla_{xx} L(x^k, \lambda^k)$ positiv definit ist, können wir das System (4.7) entkoppeln und gemäß der Range-Space Methode in Abschnitt 2 lösen. Dann erhalten wir die beiden Gleichungen

$$\begin{aligned} \big(A_k \nabla_{xx} L(x^k, \lambda^k)^{-1} A_k^T\big)\lambda^{k+1} &= -A_k \nabla_{xx} L(x^k, \lambda^k)^{-1} \nabla J_k^T + e_k, \\ \nabla_{xx} L(x^k, \lambda^k)\Delta x^k &= -\nabla J_k^T - A_k^T \lambda^{k+1} \end{aligned}$$

mit $A_k = \nabla e(x^k)$, $\nabla J_k = \nabla J(x^k)$ und $e_k = e(x^k)$. Dieser Lösungsweg ist insbesondere dann sehr effizient, wenn $\nabla_{xx} L(x^k, \lambda^k)$ durch positiv definite Approximationen ersetzt wird, zum Beispiel durch Quasi-Newton Updates.

c) Die letzte Möglichkeit wird bei sehr vielen SQP-Verfahren genutzt. Hier wird die Idee der Null-Space Methode aus dem dritten Abschnitt angewendet. Wir müssen also Matrizen Z_k und Y_k bestimmen, wobei die Spalten von Z_k eine Basis des Nullraums von A_k und die Spalten von Y_k eine des Bildraums von A_k^T bilden. Mit

$$\Delta x^k = Y_k \Delta x_Y^k + Z_k \Delta x_Z^k$$

ergeben sich aus (4.6) die beiden Gleichungen

$$(A_k Y_k)\Delta x_Y^k = -e_k, \tag{4.10a}$$

$$\big(Z_k^T \nabla_{xx} L(x^k, \lambda^k) Z_k\big)\Delta x_Z^k = -Z_k^T \nabla_{xx} L(x^k, \lambda^k) Y_k \Delta x_Y^k - Z_k^T \nabla J_k^T. \tag{4.10b}$$

Der Lagrange-Multiplikator zu Problem (4.5) ergibt sich dann aus

$$(A_k Y_k)^T \lambda^{k+1} = -Y_k^T (\nabla J_k^T + \nabla_{xx} L(x^k, \lambda^k)\Delta x^k). \tag{4.10c}$$

Für diesen Lösungsweg benötigen wir nur, dass die reduzierte Hesse-Matrix $Z_k^T \nabla_{xx} L(x^k, \lambda^k) Z_k$ positiv definit ist. Eine Variante berechnet λ^{k+1} in (4.10c), indem auf der rechten Seite der Term $\nabla_{xx} L(x^k, \lambda^k)\Delta x^k$ weggelassen wird. Wegen $\Delta x^k \to 0$ macht dieses auch Sinn. Weiters können wir im Fall von $\operatorname{Rang} A_k^T = m$ die Wahl $Y_k = A_k^T$ treffen, was auf

$$\hat{\lambda}^{k+1} = -(A_k A_k^T)^{-1} A_k \nabla J_k^T \tag{4.11}$$

führt. Dieser Lagrange-Multiplikator wird als *Least-Squares Multiplikator* bezeichnet; denn er ergibt sich als Lösung des Least-Squares-Problems

$$\min_{\hat{\lambda} \in \mathbb{R}^m} \|\nabla J_k^T + A_k^T \hat{\lambda}\|_2. \tag{4.12}$$

Offenbar sind nämlich die Optimalitäts-Bedingungen für (4.12)

$$\big(\nabla J_k^T + A_k^T \hat{\lambda}\big)^T \big(A_k^T v\big) = 0 \quad \text{für alle } v \in \mathbb{R}^n,$$

was auf die Normalgleichungen

$$A_k A_k^T \hat{\lambda} = -A_k \nabla J_k^T$$

führt. Auch wenn x^k weit von der Lösung x^* von (**P**) entfernt ist, macht (4.11) Sinn, da in jeder Iteration die Optimalitäts-Bedingungen

$$\nabla_x L(x, \hat{\lambda})^T = \nabla J(x)^T + A(x)^T \hat{\lambda} = 0$$

erfüllt ist, vergleiche (4.2). Wir berechnen daher $\hat{\lambda}^{k+1}$ durch (4.11), wobei wir auf der rechten Seite die Ausdrücke bereits an der neuen Iterierten auswerten:

$$\hat{\lambda}^{k+1} = -(A_{k+1} A_{k+1}^T)^{-1} A_{k+1} \nabla J_{k+1}^T.$$

Damit wird das SQP-Verfahren in ein Verfahren transformiert, das nur auf der primalen Variablen x^k arbeitet, denn $\hat{\lambda}^k$ hängt nur von x^k, nicht aber von $\hat{\lambda}^{k-1}$ ab.

Eine weitere Variante vernachlässigt $Z_k^T \nabla_{xx} L(x^k, \lambda^k) Y_k \Delta x_Y^k$ auf der rechten Seite von (4.10b). Es wird also nur das System

$$\big(Z_k^T \nabla_{xx} L(x^k, \lambda^k) Z_k\big) \Delta x_Z^k = -Z_k^T \nabla J_k^T \tag{4.13}$$

gelöst. Die Konvergenz dieser sogenannten *reduced SQP methods* wurde in [**24**] untersucht.

3. Die Hesse-Matrix des quadratischen Modells

In Abschnitt 1 haben wir über die Äquivalenz des SQP- mit dem Newton-Verfahren gesprochen. Unter sinnvollen Voraussetzungen erhalten wir daher lokal quadratische Konvergenz. Unter Umständen ist aber die Matrix

$$\nabla_{xx} L(x^k, \lambda^k) = \nabla^2 J(x^k) + \sum_{i=1}^{m} \lambda_i^k \nabla^2 e_i(x^k)$$

schwer zu berechnen oder nicht positiv definit auf dem Kern der linearisierten Nebenbedingungen. Eine Alternative ist daher, $\nabla_{xx} L(x^k, \lambda^k)$ durch eine Quasi-Newton Approximation B_k zu ersetzen. Die Quasi-Newton Updates haben sich bereits sehr effizient in der unrestringierten Optimierung erwiesen. Wir werden sie daher jetzt hier anwenden. Der Update für B_k beim Schritt von k nach $k+1$ verwendet die Vektoren

$$s^k = x^{k+1} - x^k \quad \text{und} \quad y^k = \nabla_x L(x^{k+1}, \lambda^{k+1})^T - \nabla_x L(x^k, \lambda^k)^T \tag{4.14}$$

zum Beispiel beim BFGS-Update wie folgt

$$B_{k+1} = B_k - \frac{B_k s^k (s^k)^T B_k}{(s^k)^T B_k s^k} + \frac{y^k (y^k)^T}{(y^k)^T s^k}.$$

Diese Variante können wir als Quasi-Newton Update für den Fall deuten, dass die Zielfunktion durch $L(x, \lambda)$ bei fixiertem λ gegeben ist. Das macht die Stärken, aber auch die Schwächen dieser Variante klar. Ist $\nabla_{xx} L(x^k, \lambda^k)$ in der Region, wo die Minimierung durchgeführt wird, positiv definit, so geben die Quasi-Newton Approximationen B_k gute Informationen über die Krümmung und das Verfahren konvergiert schnell und robust gegen die Minimalstelle. Besitzt hingegen $\nabla_{xx} L(x^k, \lambda^k)$

negative Eigenwerte, so sind die positiv definiten Approximationen nicht sehr geeignet. Die Bedingung $(s^k)^T y^k > 0$ braucht noch nicht einmal in einer kleinen Umgebung der Lösung zu gelten. Diese Beobachtungen haben zu folgenden gedämpften BFGS-Update Formeln für SQP-Verfahren geführt:

1) Definiere die Vektoren s^k und y^k gemäß (4.14) und setze

$$r^k = \theta_k y^k + (1-\theta_k) B_k s^k,$$

wobei der Skalar $\theta_k \in [0,1]$ gegeben ist durch

$$\theta_k = \begin{cases} 1 & \text{falls } (s^k)^T y^k \geq 0.2 (s^k)^T B_k s^k, \\ \frac{0.8 (s^k)^T B_k s^k}{(s^k)^T B_k s^k - (s^k)^T y^k} & \text{falls } (s^k)^T y^k < 0.2 (s^k)^T B_k s^k. \end{cases} \tag{4.15}$$

2) Berechne B_{k+1} mit der Update-Formel

$$B_{k+1} = B_k - \frac{B_k s^k (s^k)^T B_k}{(s^k)^T B_k s^k} + \frac{r^k (r^k)^T}{(r^k)^T s^k}. \tag{4.16}$$

Die Formel (4.16) ist die BFGS-Update Formel, wobei y^k durch den Vektor r^k ersetzt worden ist. Für $\theta_k = 1$ folgt $r^k = y^k$. Im Fall von $\theta_k \neq 1$ erhalten wir mit (4.15) die Abschätzung

$$\begin{aligned} (s^k)^T r^k &= (s^k)^T \big(\theta_k y^k + (1-\theta_k) B_k s^k\big) = \theta_k (s^k)^T y^k + (1-\theta_k)(s^k)^T B_k s^k \\ &= \frac{0.8 (s^k)^T B_k s^k \, (s^k)^T y^k}{(s^k)^T B_k s^k - (s^k)^T y^k} + \frac{0.2 (s^k)^T B_k s^k - (s^k)^T y^k}{(s^k)^T B_k s^k - (s^k)^T y^k} \, (s^k)^T B_k s^k \\ &= \left(\frac{0.8 (s^k)^T y^k}{(s^k)^T B_k s^k - (s^k)^T y^k} + \frac{0.2 (s^k)^T B_k s^k - (s^k)^T y^k}{(s^k)^T B_k s^k - (s^k)^T y^k} \right) (s^k)^T B_k s^k \\ &= 0.2 (s^k)^T B_k s^k > 0. \end{aligned}$$

Damit ist B_{k+1} positiv definit. Für $\theta_k = 0$ folgt $B_k = B_{k+1}$. Andererseits führt $\theta_k = 1$ auf eine möglicherweise indefinite Matrix, die sich aus den unmodifizierten BFGS-Formeln ergibt Mit $\theta_k \in (0,1)$ erhalten wir eine Interpolation der beiden Extremfälle.

Eine andere Variante bietet sich dadurch an, die reduzierte Hesse-Matrix der Lagrange-Funktion $Z_k^T \nabla_{xx} L(x^k, \lambda^k) Z_k$ zu approximieren, insbesondere dann, wenn die Dimension dieser Matrix klein ist. Die Herangehensweise ist in *Reduced-Hessian Quasi-Newton Methods* realisiert. Die Suchrichtung erfüllt

$$\lambda_k = -(A_k A_k^T)^{-1} A_k \nabla J_k^T, \tag{4.17a}$$

$$(A_k Y_k) \Delta x_Y^k = -e_k, \tag{4.17b}$$

$$M_k \Delta x_Z^k = -Z_k^T \nabla J_k^T, \tag{4.17c}$$

wobei wir in (4.17c) im Vergleich mit (4.13) eine Quasi-Newton Approximation für die reduzierte Hesse-Matrix $Z_k^T \nabla_{xx} L(x^k, \lambda^k) Z_k$ verwendet haben. Im Folgenden wollen wir diskutieren, wie die Matrizen M_k konstruiert werden können. Sei $\alpha_k \Delta x^k$ der Schritt von (x^k, λ^k) nach (x^{k+1}, λ^{k+1}). Wegen des Satzes von Taylor folgt

$$\nabla_{xx} L(x^{k+1}, \lambda^{k+1}) \big(\alpha_k \Delta x^k\big) \approx \nabla_x L(x^k + \alpha_k \Delta x^k, \lambda^{k+1})^T - \nabla_x L(x^k, \lambda^{k+1})^T$$

mit $\Delta x^k = x^{k+1} - x^k = Z_k \Delta x_Z^k + Y_k \Delta x_Y^k$. Multiplikation mit Z_k^T ergibt

$$\begin{aligned} & Z_k^T \nabla_{xx} L(x^{k+1}, \lambda^{k+1}) Z_k \big(\alpha_k \Delta x_Z^k\big) \\ & \quad \approx -Z_k^T \nabla_{xx} L(x^{k+1}, \lambda^{k+1}) \big(\alpha_k Y_k \Delta x_Y^k\big) \\ & \qquad + Z_k^T \big(\nabla_x L(x^{k+1}, \lambda^{k+1})^T - \nabla_x L(x^k, \lambda^{k+1})^T\big). \end{aligned} \tag{4.18}$$

Vernachlässigen des ersten Terms auf der rechten Seite von (4.18) führt auf

$$M_{k+1} s^k = y^k$$

mit

$$s^k = \alpha_k \Delta x_Z^k \quad \text{und} \quad y^k = Z_k^T \big(\nabla_x L(x^{k+1}, \lambda^{k+1})^T - \nabla_x L(x^k, \lambda^{k+1})^T\big). \tag{4.19}$$

Damit können wir die BFGS-Formeln

$$M_{k+1} = M_k - \frac{M_k s^k (s^k)^T M_k}{(s^k)^T M_k s^k} + \frac{y^k (y^k)^T}{(y^k)^T s^k}$$

verwenden, um die neue Approximation M_{k+1} zu berechnen. Es gibt Varianten von (4.19), zum Beispiel

$$y^k = Z_k^T \big(\nabla J_{k+1}^T - \nabla J_k^T\big)$$

oder

$$y^k = Z_k^T \big(\nabla_x L(x^k + Z_k \Delta x_Z^k, \lambda^{k+1})^T - \nabla_x L(x^k, \lambda^{k+1})^T\big) \tag{4.20}$$

(Cole und Coleman). In (4.20) ist die zusätzliche Auswertung des Terms $\nabla_x L(x^k + Z_k \Delta x_Z^k, \lambda^{k+1})$ notwendig. In einer Umgebung der Lösung gilt für (4.20) die Abschätzung

$$\begin{aligned} (y^k)^T s^k &= \alpha_k \big(\nabla_x L(x^k + Z_k \Delta x_Z^k, \lambda^{k+1}) - \nabla_x L(x^k, \lambda^{k+1})\big) Z_k \Delta x_Z^k \\ &= \alpha_k (\Delta x_Z^k)^T \left(\int_0^1 Z_k^T \nabla_{xx} L(x^k + s Z_k \Delta x_Z^k, \lambda^{k+1}) Z_k \, \mathrm{d}s \right) \Delta x_Z^k > 0 \end{aligned}$$

für $(x^k + s Z_k \Delta x_Z^k, \lambda^{k+1}) \in U(x^*, \lambda^*)$. Damit sind die BFGS-Formeln wohldefiniert.

4. Merit- oder Straffunktionen

Um zu garantieren, dass das SQP-Verfahren von Startwerten, die weit weg von Lösungen liegen, konvergiert, wird häufig eine Merit- oder Straffunktion verwendet, um bei Liniensuch-Verfahren die Schrittweite zu kontrollieren oder bei Trust-Region Verfahren den Trust-Region zu modifizieren. Im unrestringierten Fall haben wir die Zielfunktion verwendet. Wir wollen hier nur zwei Meritfunktionen diskutieren: die nicht-differenzierbare ℓ_1-*Meritfunktion* sowie Fletchers exakte und differenzierbare *augmentierte Lagrange Funktion*.

Ziel der Meritfunktion ist die Garantie globaler Konvergenz ohne Schritte zu verwerfen, die zur Lösung führen. Die ℓ_1-Meritfunktion für Probleme mit Gleichungs-Restriktionen lautet

$$\Phi_1(x; \mu) = J(x) + \frac{1}{\mu} \|e(x)\|_1 = J(x) + \frac{1}{\mu} \sum_{i=1}^m |e_i(x)|, \tag{4.21}$$

wobei $\mu > 0$ ein Strafparameter ist. Die Abbildung Φ_1 ist insbesondere für Punkte x mit $e_i(x) = 0$ für mindestens ein $i \in \{1, \ldots, m\}$ nicht differenzierbar. Eine Richtungs-Ableitung von Φ_1 existiert dagegen immer.

Beispiel 4.4. Wir berechnen für $J(x) = \|x\|_1$ die Richtungs-Ableitung

$$D(J(x);\Delta x) = \lim_{\varepsilon \searrow 0} \frac{1}{\varepsilon} \big(J(x+\varepsilon\Delta x) - J(x)\big)$$

in eine Richtung $\Delta x \in \mathbb{R}^n$. Wir erhalten

$$D(J(x);\Delta x) = \lim_{\varepsilon \searrow 0} \frac{1}{\varepsilon} \left(\|x+\varepsilon\Delta x\|_1 - \|x\|_1\right) = \lim_{\varepsilon \searrow 0} \frac{1}{\varepsilon} \sum_{i=1}^{n} \big(|x_i+\varepsilon\Delta x_i| - |x_i|\big).$$

Gilt $x_i > 0$ für ein $i \in \{1,\dots,n\}$, so folgt $|x_i+\varepsilon\Delta x_i| = x_i+\varepsilon\Delta x_i$ für ε hinreichend klein. Ist hingegen $x_i < 0$ für ein $i \in \{1,\dots,n\}$, so erhalten wir $|x_i+\varepsilon\Delta x_i| = |x_i| - \varepsilon\Delta x_i$ für ε klein genug. Im Fall von $x_i = 0$ für ein $i \in \{1,\dots,n\}$ gilt $|x_i+\varepsilon\Delta x_i| = \varepsilon\,|\Delta x_i|$. Insgesamt berechnen wir daher

$$D(J(x);\Delta x) = \sum_{i:x_i>0} \Delta x_i - \sum_{i:x_i<0} \Delta x_i + \sum_{i:x_i=0} |\Delta x_i|$$

als Richtungs-Ableitung von J am Punkt x in Richtung Δx. ◊

Lemma 4.5. *Seien Δx^k und λ^{k+1} Lösungen von (4.7). Dann gilt für die Richtungs-Ableitung von Φ_1 in Richtung Δx^k die Abschätzung*
(4.22)

$$D(\Phi_1(x^k;\mu);\Delta x^k) \le -(\Delta x^k)^T \nabla_{xx} L(x^k,\lambda^k)\Delta x^k - \left(\frac{1}{\mu} - \|\lambda^{k+1}\|_\infty\right) \|e(x^k)\|_1.$$

Bemerkung 4.6. Ist $\nabla_{xx}L(x^k,\lambda^k)$ positiv definit, so folgt aus (4.22), dass Δx^k eine Abstiegs-Richtung für Φ_1 an x^k ist, wenn μ hinreichend klein gewählt wird. Es lässt sich zeigen dass dieser Schluß auch gilt, wenn die reduzierte Hesse-Matrix $Z_k^T \nabla_{xx} L(x^k,\lambda^k) Z_k$ positiv definit ist. In der Praxis wird

$$\mu = \frac{1}{\|\lambda^{k+1}\|_\infty + \delta}$$

mit einem $\delta > 0$ gewählt. Eine andere Möglichkeit ist es zu fordern, dass die Richtungsableitung von Φ_1 hinreichend negative ist:

$$D\Phi_1(x^k;\mu);\Delta x^k) = \nabla J(x^k)\Delta x^k - \frac{1}{\mu}\|e(x^k)\|_1 \le -\frac{\varrho}{\mu}\|e(x^k)\|_1$$

for some $\varrho \in (0,1)$. This inequality hold if

$$\frac{1}{\mu} \ge \frac{\nabla J(x^k)\Delta x^k}{(1-\varrho)\|e(x^k)\|_1} \quad \text{beziehingsweise} \quad \mu \le \frac{(1-\varrho)\|e(x^k)\|_1}{\nabla J(x^k)\Delta x^k}. \tag{4.23}$$

Diese Wahl hängt nicht vom Lagrange-Multiplikator ab und wird in der Praxis häufig verwendet. ◊

Beweis von Lemma 4.5. Wir wenden den Satz von Taylor an:

$$\begin{aligned}
\Phi_1(x^k+\alpha\Delta x^k;\mu) - \Phi_1(x^k;\mu) &= J(x^k+\alpha\Delta x^k) - J(x^k) \\
&\quad + \frac{1}{\mu}\left(\|e(x^k+\alpha\Delta x^k)\|_1 - \|e(x^k)\|_1\right) \\
&\le \alpha\nabla J(x^k)\Delta x^k + \gamma\alpha^2\|\Delta x^k\|^2 \\
&\quad + \frac{1}{\mu}\left(\|e(x^k)+\alpha A_k\Delta x^k\|_1 - \|e(x^k)\|_1\right)
\end{aligned}$$

wobei $\gamma > 0$ eine Schranke für die zweiten Ableitungen von J und e bezeichnet. Für Δx^k aus (4.7) gilt $A_k \Delta x^k = -e(x^k)$. Also gilt für $\alpha \in [0,1]$

$$\Phi_1(x^k + \alpha\Delta x^k;\mu) - \Phi_1(x^k;\mu) \leq \alpha\left(\nabla J(x^k)\Delta x^k - \frac{1}{\mu}\left\|e(x^k)\right\|_1\right) + \alpha^2\gamma\left\|\Delta x^k\right\|^2.$$

Analog schließen wir

$$\Phi_1(x^k + \alpha\Delta x^k;\mu) - \Phi_1(x^k;\mu) \geq \alpha\left(\nabla J(x^k)\Delta x^k - \frac{1}{\mu}\left\|e(x^k)\right\|_1\right) - \alpha^2\gamma\left\|\Delta x^k\right\|^2.$$

Daher folgt

$$D(\Phi_1(x^k;\mu), \Delta x^k) = \nabla J(x^k)\Delta x^k - \frac{1}{\mu}\left\|e(x^k)\right\|_1.$$

Aus der ersten Blockzeile in (4.7) bekommen wir

$$D(\Phi_1(x^k;\mu), \Delta x^k) = -(\Delta x^k)^T \nabla_{xx} L(x^k,\lambda^k)\Delta x^k - (\Delta x^k)^T A_k^T \lambda^{k+1} - \frac{1}{\mu}\left\|e(x^k)\right\|_1,$$

und wegen der zweiten Blockzeile in (4.7) gilt

$$D(\Phi_1(x^k;\mu), \Delta x^k) = -(\Delta x^k)^T \nabla_{xx} L(x^k,\lambda^k)\Delta x^k + e(x^k)^T \lambda^{k+1} - \frac{1}{\mu}\left\|e(x^k)\right\|_1.$$

Aufgrund der Abschätzung

$$e(x^k)^T\lambda^{k+1} = \sum_{i=1}^m e_i(x^k)\lambda_i^{k+1} \leq \|\lambda^{k+1}\|_\infty \sum_{i=1}^m |e_i(x^k)| = \|\lambda^{k+1}\|_\infty \|e(x^k)\|_1$$

folgt (4.22). □

Eine weitere sehr effektive Strategie, um den Strafparameter μ zu wählen, wird sowohl im Kontext von Liniensuch- oder Trust-Region-Verfahren verwendet. Das Vorgehen basiert auf einem quadratischen Modell für Φ_1:

$$q_\mu(\Delta x) = J(x^k) + \nabla J(x^k)\Delta x \frac{\sigma}{2}\,\Delta x^T \nabla_{xx} L(x^k,\lambda^k)\Delta x + \frac{1}{\mu}\, m(\Delta x), \tag{4.24}$$

wobei

$$m(\Delta x) = \left\|e(x^k) + A_k\Delta x\right\|_1$$

gilt und σ ein Parameter ist, den wir später definieren. Haben wir einen Schritt Δx^k berechnet, so wählen wir μ hinreichend klein, so dass

$$q_\mu(0) - q_\mu(\Delta x^k) \geq \frac{\varrho}{\mu}\left(m(0) - m(\Delta x^k)\right) \tag{4.25}$$

erfüllt ist mit einem $\varrho \in (0,1)$. Es folgt aus (4.24), dass die Ungleichung (4.25) bfür

$$\frac{1}{\mu} \geq \frac{\nabla J(x^k)\Delta x^k + \frac{\sigma}{2}(\Delta x^k)^T \nabla_{xx} L(x^k,\lambda^k)\Delta x^k}{(1-\varrho)\|e(x^k)\|_1} \tag{4.26}$$

gilt. Erfüllt der Parameter μ aus der vorangegangenen SQP-Iteration die Bedingung (4.26), so ändern wir μ nicht. Andernfalls wird μ verkleinert, so dass (4.26) gilt. Die Konstante σ ermöglicht es, den Fall zu behandeln, wenn die Hesse-Matrix $\nabla_{xx} L(x^k,\lambda^k)$ nicht positiv definit ist. Wir setzen daher

$$\sigma = \begin{cases} 1 & \text{falls } (\Delta x^k)^T \nabla_{xx} L(x^k,\lambda^k)\Delta x^k > 0, \\ 0 & \text{andernfalls.} \end{cases}$$

Erfüllt μ (4.26), so garantiert die Wahl für σ die Ungleichung $D(\Phi_1(x^k;\mu);\Delta x^k) \leq -(\varrho/\mu)\,\|e(x^k)\|_1$. Damit ist Δx^k eine Abstiegsrichtung für Φ_1. Dies ist nicht immer

erfüllt, wenn $\sigma = 1$ und $\nabla_{xx} L(x^k, \lambda^k)$ nicht positiv definit ist. Ein Vergleich von (4.23) und (4.26) zeigt, dass wir im Fall $\sigma > 0$ in der Strategie, die (4.24) verwendet, einen größeren Strafparameter zulassen. Damit wird mehr Gewicht auf die Erfüllung der Nebenbedingungen gelegt. Das ist sinnvoll bei Schritten, die eine Reduktion der Nebenbedingungen, aber einen Anstieg im Zielfunktional bewirken. Diese Schritte werden dann durch die Meritfunktion eher akzeptiert.

Nun kommen wir zu Fletchers augmentierter Lagrange-Funktion:

$$\Phi_F(x;\mu) = J(x) + \lambda(x)^T e(x) + \frac{1}{2\mu} \|e(x)\|^2, \tag{4.27}$$

wobei $\mu > 0$ einen Penalty-Parameter bezeichnet und

$$\lambda(x) = -\big(A(x)A(x)^T\big)^{-1} A(x) \nabla J(x)^T \tag{4.28}$$

der Least-Squares Multiplikator ist. Offenbar ist Φ_F differenzierbar. Es folgt

$$\nabla \Phi_F(x^k;\mu) = \nabla J(x^k) + \left(A_k^T \lambda^k + \nabla \lambda(x^k)^T e(x^k) + \frac{1}{\mu} A_k^T e(x^k) \right)^T.$$

Löst Δx^k das System (4.7), so gilt

$$\nabla \Phi_F(x^k;\mu) \Delta x^k = \nabla J(x^k) \Delta x^k - (\lambda^k)^T e(x^k) + e(x^k)^T \nabla \lambda(x^k) \Delta x^k - \frac{1}{\mu} \|e(x^k)\|^2.$$

Wir schreiben $\Delta x^k = Z_k \Delta x_Z^k + Y_k \Delta x_Y^k$, wobei Z_k eine Basis des Nullraums von A_k ist und $Y_k = A_k^T$ gesetzt ist. Wegen (4.10a) erhalten wir

$$A_k^T \Delta x_Y^k = -A_k^T \big(A_k A_k^T\big)^{-1} e(x^k),$$

und wegen (4.28) gilt

$$\nabla J(x^k) A_k^T \Delta x_Y^k = -\nabla J(x^k) A_k^T \big(A_k A_k^T\big)^{-1} e(x^k) = (\lambda^k)^T e(x^k).$$

Damit folgt

$$\begin{aligned} \nabla \Phi_F(x^k;\mu) \Delta x^k &= \nabla J(x^k) Z_k \Delta x_Z^k + \nabla J(x^k) A_k^T \Delta x_Y^k - (\lambda^k)^T e(x^k) \\ &\quad + e(x^k)^T \nabla \lambda(x^k) \Delta x^k - \frac{1}{\mu} \|e(x^k)\|^2 \\ &= \nabla J(x^k) Z_k \Delta x_Z^k + e(x^k)^T \nabla \lambda(x^k) \Delta x^k - \frac{1}{\mu} \|e(x^k)\|^2. \end{aligned}$$

Aus der ersten Blockzeile in (4.7) folgt

$$\begin{aligned} \nabla_{xx} L(x^k, \lambda^k) \Delta x^k &= \nabla_{xx} L(x^k, \lambda^k) Z_k \Delta x_Z^k + \nabla_{xx} L(x^k, \lambda^k) A_k^T \Delta x_Y^k \\ &= -\nabla J(x^k)^T - A_k^T \lambda^{k+1}. \end{aligned}$$

Mit $\Delta x^k = A_k^T \Delta x_Y^k + Z_k \Delta x_Z^k$ erhalten wir

$$\begin{aligned} \nabla \Phi_F(x^k;\mu) \Delta x^k &= -(\Delta x_Z^k)^T Z_k^T \nabla_{xx} L(x^k, \lambda^k) Z_k \Delta x_Z^k \\ &\quad - (\Delta x_Y^k)^T A_k \nabla_{xx} L(x^k, \lambda^k) Z_k \Delta x_Z^k + e(x^k)^T \nabla \lambda(x^k) \Delta x^k \\ &\quad - \frac{1}{\mu} \|e(x^k)\|^2. \end{aligned}$$

Damit is Δx^k eine Abstiegsrichtung für die Abbildung Φ_F, wenn die reduzierte Hesse-Matrix $Z_k^T \nabla_{xx} L(x^k, \lambda^k) Z_k$ positiv definit ist und μ der Bedingung

$$\frac{1}{\mu} > \frac{-\frac{1}{2}(\Delta x^k)^T Z_k^T \nabla_{xx} L(x^k, \lambda^k) Z_k \Delta x^k + e(x^k)^T \nabla \lambda(x^k) \Delta x^k}{\|e(x^k)\|^2} + \frac{-(\Delta x_Y^k)^T A_k \nabla_{xx} L(x^k, \lambda^k) Z_k \Delta x_Z^k}{\|e(x^k)\|^2} + \delta \tag{4.29}$$

genügt für ein $\delta > 0$. Im Falle von $e(x^k) = 0$ ist Δx^k eine Abstiegsrichtung für jedes $\mu > 0$, siehe (4.10a).

Satz 4.7. *Angenommen, x^k ist kein stationärer Punkt des Problem* (**P**) *und die Hesse-Matrix $Z_k^T \nabla_{xx} L(x^k, \lambda^k) Z_k$ ist positiv definit. Dann ist die Suchrichtung Δx^k aus* (4.7) *eine Abstiegsrichtung für Φ_1, wenn* (4.22) *gilt, und für Φ_F, wenn* (4.29) *erfüllt ist*

5. Ein SQP-Verfahren mit Liniensuche

Es gibt viele Varianten für SQP-Verfahren. Sie können sich zum Beispiel durch folgende Aspekte unterscheiden:

- Approximation der Hesse-Matrix,
- Wahl der Merit-Funktion,
- Berechnung des Schrittweiten-Parameters,
- Update für den Multiplikator λ^k,
- unterschiedliche Formeln für die Quasi-Newton Approximation,
- andere Parameter,
- Globalisierung mit Trust-Region oder Liniensuch-Strategien,
- Berechnung des SQP-Schrittes.

Wir wollen nun ein Beispiel für ein SQP-Verfahren angeben.

Algorithmus 4.8 (SQP-Verfahren für nichtlineare Optimierung).

1) *Wähle $\eta \in (0, 1/2)$, $\tau \in (0, 1)$, $(x^0, \lambda^0) \in \mathbb{R}^n \times \mathbb{R}^m$, $k_{\max} \in \mathbb{N}$;*
2) *Wähle positiv definite und symmetrische Startmatrix $B_0 \in \mathbb{R}^{n \times n}$ als Approximation der reduzierten Hesse-Matrix; berechne J_0, ∇J_0, e_0 sowie A_0;*
3) **for** $k = 0$ **to** $k_{\max}$
   - **if** *Konvergenz, breche ab;*
   - *Berechne den SQP-Schritt Δx^k;*
   - *Bestimme $\mu_k > 0$, so dass Δx^k eine Abstiegsrichtung für die Merit-Funktion Φ ist;*
   - *Setze $\alpha^{(0)} = 1$ und $i = 0$;*
   - **while** *($\Phi(x^k + \alpha^{(i)} \Delta x^k; \mu_k) > \Phi(x^k; \mu_k) + \eta \alpha^{(i)} D(\Phi(x^k; \mu); \Delta x^k)$*
     Setze $\alpha^{(i+1)} = \tau_\alpha \alpha^{(i)}$ mit $\tau_\alpha \in (0, \tau)$ und $i = i + 1$;
   - **end**
   - *Setze $\alpha_k = \alpha^{(i)}$ und $x^{k+1} = x^k + \alpha_k \Delta x^k$;*
   - *Berechne J_{k+1}, ∇J_{k+1}, e_{k+1} sowie A_{k+1};*
   - *Bestimme Least-Squares Multiplikator λ^{k+1}:*

$$\lambda^{k+1} = -\left(A_{k+1} A_{k+1}^T\right)^{-1} A_{k+1} \nabla J_{k+1}^T;$$

   - *Setze $s^k = \alpha_k \Delta x^k$, $y^k = Z_k^T (\nabla_x L(x^{k+1}, \lambda^{k+1})^T - \nabla_x L(x^k, \lambda^{k+1})^T)$;*
   - *Berechne B_{k+1} aus B_k mittels BFGS-Update;*

end (for)

Wir können bei der Lösung der quadratischen Teilprobleme durch die Verwendung von Warm-up Strategien deutlich effizienter werden. Ferner kann ein Limited-Memory BFGS-Verfahren [**29**] verwendet werden, insbesondere im Kontext von hoch-dimensionalen Optimierungsaufgaben. Wird die Hesse-Matrix $\nabla_{xx} L_k$ verwendet, so gehen wir davon aus, dass eine Modifikation der Hesse-Matrix durchgeführt wird, sofern die Matrix nicht positiv definit auf dem $\ker A_k$ ist.

6. Trust-Region SQP-Verfahren

Trust-Region SQP-Verfahren besitzen mehrere Vorteile. Auch wenn die Hesse-Matrix $\nabla_{xx} L_k$ positiv definit auf dem Unterraum $\ker A_k$ oder gar singulär ist, kann eine Strategie verfolgt werden, die globale Konvergenz garantiert. Die einfachste Weise, einen Trust-Region Algorithmus zu entwerfen, besteht darin, zum quadratischen Teilproblem (4.9) eine Trust-Region-Restriktion dazuzufügen:

$$\min_{\Delta x^k \in \mathbb{R}^n} J_k + \nabla J_k \Delta x + \frac{1}{2} \left(\Delta x^k\right)^T \nabla_{xx} L(x^k, \lambda^k) \Delta x^k \tag{4.30a}$$

$$\text{u.d.N.} \quad \nabla e(x^k) \Delta x^k + e(x^k) = 0 \text{ in } \mathbb{R}^m, \tag{4.30b}$$

$$\nabla g(x^k) \Delta x^k + g(x^k) \leq 0 \text{ in } \mathbb{R}^p, \tag{4.30c}$$

$$\|\Delta x^k\|_2 \leq \Delta_k. \tag{4.30d}$$

Selbst wenn die Bedingungen (4.30b) und (4.30c) kompatibel sind, kann es sein, dass das Problem (4.30) keine Lösung besitzt, da aufgrund von der Restriktion (4.30) die Menge der zulässigen Lösungen leer ist. Um den Konflikt der Bedingungen (4.30b)-(4.30d) zu lösen, kann nicht einfach der Rust-Region Radius Δ_k vergrößert werden, denn sonst kann keine globale Konvergenz garantiert werden. Daher wird die folgede Strategie verwendet: Es besteht keine Notwendigkeit, die Gleichungsnebenbedingung (4.30b) exakt zu erfüllen, sondern im Laufe der SQP-Iterationen die Gleichungsnebenbedingung zunehmend besser zu garantieren. Es gibt hier drei unterschiedliche Strategien: Ralaxierungsmethoden, Penalty-Verfahren oder Filter-Algorithmen.

Wir wollen hier kurz auf Relaxierungsmethoden eingehen. Dabei beschränken wir uns auf (**P**), das heißt, auf Optimierungsprobleme mit Gleichungsrestriktionen. Erweiterungen auf Probleme mit Ungleichungs-Nebenbedingungen basieren auf Innere-Punkte-Verfahren. Sei die Iterierte x^k gegeben. Wir berechnen im SQP Schritt die Lösung des Teilproblems

$$\min_{\Delta x^k \in \mathbb{R}^n} J_k + \nabla J_k \Delta x + \frac{1}{2} \left(\Delta x^k\right)^T \nabla_{xx} L(x^k, \lambda^k) \Delta x^k \tag{4.31a}$$

$$\text{u.d.N.} \quad \nabla e(x^k) \Delta x^k + e(x^k) = r_k \text{ in } \mathbb{R}^m, \tag{4.31b}$$

$$\|\Delta x^k\|_2 \leq \Delta_k. \tag{4.31c}$$

Die Wahl des Vektors r_k erfordert eine gute Strategie, da die Effizienz des Verfahrens wesentlich davon abhängt. Wir wählen r_k als kleinsten Vektor, so dass (4.31b) und (4.31c) erfüllt sind für einen leicht reduzierten Trust-Region Radius Δ_k. Daher lösen wir zunächst das Teilproblem

$$\min_{v \in \mathbb{R}^n} \|A_k v + e_k\|_2^2 = v^T A_k^T A_k v + 2 e_k A_k v + \|e_k\|_2^2 \quad \text{u.d.N.} \quad \|v\|_2 \leq 0.8\, \Delta_k. \tag{4.32}$$

Sei v_k die Lösung von (4.32). Dann definieren wir

$$r_k = A_k v_k + e_k. \tag{4.33}$$

Nun lösen wir (4.31), bestimmen den Schritt Δx^k und setzen $x^{k+1} = x^k + \Delta x^k$. Nun kann λ^{k+1} mit Hilfe der Least-Squares Formel berechnet werden. Wir bemerken, dass nun (4.31b) und (4.31c) konsistent sind, da sie für $\Delta x^k = v_k$ erfüllt sind.

Auf den ersten Blick erscheint die Vorgangsweise nicht effizient zu sein, da in der Regel die Probleme (4.30) und (4.32) nicht einfach zu lösen sind, insbesondere wenn $\nabla_{xx} L_k$ indefinite ist. Es sind aber sehr effiziente Verfahren entwickelt worden, die beiden Optimierungsaufgaben inexakt zu lösen.

Zur Lösung von (4.32) verwenden wir ein Dogleg-Verfahren [**29**]. Dazu benötigen wir den Cauchy-Punkt v^{CP}, welcher der Minimierer des Zielfunktionals in (4.31a) entlang der Richtung $-A_k e_k$ ist, und — im Falle der Existenz — den Newton-Punkt v^{NP}, den unrestringierten Minimierer von (4.31a). Da die Hesse-Matrix von (4.32) singulär ist, gibt es unendlich viele Möglichkeiten zur Wahl von v^{NP}, die alle die Gleichung $A_k v^{\mathrm{NP}} + e_k = 0$ erfüllen. Wir wählen die Lösung mit der minimalen Euklidischen Norm, indem wir die Singulärwertzerlegung zur Lösung verwenden. Nun sei v_k der Minimierer von (4.32) entlang des Pfads, der durch v^{CP} und v^{NP} definiert wird:

$$\tilde{v}(\tau) = \begin{cases} \tau v^{\mathrm{CP}} & \text{für } 0 \le \tau \le 1, \\ v^{\mathrm{CP}} + (\tau - 1)\left(v^{\mathrm{NP}} - v^{\mathrm{CP}}\right) & \text{für } 1 \le \tau \le 2. \end{cases}$$

Eine bevorzugte Technik zur Berechnung einer approximativen Lösung δx^k für (4.31) ist das projezierte konjugierte Gradienten-Verfahren. Wir wenden dieses Verfahren zur Lösung des problems (4.31a)-(4.31b) an, wobei wir darauf achten, dass die Trust-Region-Bedingung (4.31c) erfüllt ist, und brechen ab, sobald der Trust-Region-Rand oder eine Richtung negativer Krümmung erreicht wird.

Eine Meritfunktion für die präsentierte Vorgangsweise ist zum Beispiel die nichtglatte ℓ_2-Funktion

$$\Phi_2(x;\mu) = J(x) + \frac{1}{\mu}\,\|e(x)\|_2, \quad \mu > 0.$$

Für Φ_2 verwenden wir das quadratische Modell

$$q_\mu(\Delta x) = J(x^k) + \nabla J(x^k)\Delta x + \frac{1}{2}\,\Delta x^T \nabla_{xx} L(x^k, \lambda^k)\Delta x + \frac{1}{\mu}\,m(\Delta x)$$

wobei wir

$$m(\delta x) = \|e_k + A_k \Delta x\|_2$$

setzen. Wir wählen den Strafparameter μ hinreichend klein, so dass die Ungleichung

$$q_\mu(0) - q_\mu(\Delta x^k) \ge \frac{\varrho}{\mu}\left(m(0) - m(\Delta x^k)\right), \quad \varrho \in (0,1), \tag{4.34}$$

erfüllt ist. Die Entscheidung, ob ein Schritt Δx^k akzeptiert wird, wird anhand des Quotienten

$$\varrho_k = \frac{\mathrm{ared}_k}{\mathrm{pred}_k} = \frac{\Phi_2(x^k;\mu) - \Phi_2(x^k + \Delta x^k;\mu)}{q_\mu(0) - q_\mu(\Delta x^k)}$$

durchgeführt.

Algorithmus 4.9 (Byrd-Omojokun Trust-Region SQP-Verfahren).

1) *Wähle Konstanten* $k_{\max} \in \mathbb{N}$, $\varepsilon > 0$ *und* $\eta, \gamma \in (0,1)$*;*
2) *Wähle einen Startwert* x^0 *und einen Trust-Radius Radius* Δ^0*;*
3) **for** $k = 0$ **to** $k_{\max}$
   Berechne J_k, e_k, ∇J_k *und* A_k*;*

Bestimme den Least-Squares Multiplikator $\hat{\lambda} = (A_k A_k^T)^T A_k \nabla J_k$*;*
if $\|\nabla J_k - A_k^T \lambda_k\|_\infty < \varepsilon$ **and** $\|e_k\|_\infty < \varepsilon$
    Abbruch mit der approximativen Lösung x^k*;*
end (if)
Löse das Teilproblem (4.32) *zur Berechnung von* v_k *und setze* r_k *gemäß* (4.33)*;*
Berechne $\nabla_{xx} L(x^k, \lambda^k)$ *oder eine Quasi-Newton Approximation der Hesse-Matrix;*
Löse das Problem (4.31) *unter Verwendung eines projezierten konjugierten Gradienten-Verfahren;*
Bestimme einen Strafparameter μ_k*, der* (4.34) *erfüllt;*
Berechne den Quotienten $\varrho_k = \mathrm{ared}_k / \mathrm{pred}_k$*;*
if $\varrho_k > \eta$
    Setze $x^{k+1} = x^k + \Delta x^k$ *und wähle einen neuen Trust-Region Radius mit* $\Delta_{k+1} \geq \Delta_k$*;*
else
    Setze $x^{k+1} = x^k$ *und wähle einen neuen Trust-Region Radius mit* $\Delta_{k+1} \leq \gamma \|\Delta x^k\|_2$*;*
end (if)
end (for)

KAPITEL 5

Grundlagen der multikriteriellen Optimierung

Wir wollen in diesem ersten Abschnitt einige wesentliche Grundlagen der Vektoroptimierung zusammenstellen und einen Überblick über bestehende Verfahren geben. Dabei orientieren wir uns an dem Buch [**17**] von C. Hillermeier. Als weitere Literatur verweisen wir zum Beispiel auf [**28, 30, 33, 34**].

1. Das Konzept der Pareto-Optimalität

Gegeben sei eine vektorwertige Abbildung $\mathbf{f} : \mathbb{R}^n \to \mathbb{R}^k$ durch $x \mapsto \mathbf{f}(x) = (f_1(x), \ldots, f_k(x))$. Da die Minimierung einer Komponentenfunktion f_i, $1 \leq i \leq k$, von f die Vergrößerung eines anderen f_j, $j \neq i$, bedeuten kann, müssen wir das Problem der Vektoroptimierung in geeigneter Weise interpretieren. Dazu benötigen wir eine Ordnung auf $\mathbb{R}^k$, die zu dem Problem passt. Offenbar gibt es keine totale Ordnung auf $\mathbb{R}^k$. Seien zum Beispiel $k = 2$ und $y^1 = (4, 2)$ sowie $y^2 = (2, 4)$. Wenn es keine Gewichtung der beiden Komponenten von y^1 und y^2 gibt, lassen sich beide Vektoren nicht mit einer Ordnungsrelation vergleichen. Aus diesem Grund führen wir die folgende Halbordnung auf $\mathbb{R}^k$ ein.

Definition 5.1. *Mit $\leq$ bezeichnen wir eine* Ordnungsrelation im $\mathbb{R}^k$*, das heißt, eine Teilmenge von $\mathbb{R}^k \times \mathbb{R}^k$ bestehend aus allen geordneten Paaren von Elementen aus $\mathbb{R}^k$. An der Stelle von $(y^1, y^2) \in \leq$ schreiben wir auch einfach $y^1 \leq y^2$. Die Ordnungsrelation ist wie folgt definiert:*

$$y^1 \leq y^2 \iff y^2 - y^1 \in \mathbb{R}^k_+,$$

wobei $\mathbb{R}^k_+ = \{y = (y_1, \ldots, y_k) \in \mathbb{R}^k \mid y_i \geq 0$ für alle $i = 1, \ldots, k\}$ der nicht-negative Orthant in $\mathbb{R}^k$ ist und $\mathbb{R}_+ = \mathbb{R}^1_+$ gesetzt ist.

Anhand von Abbildung 5.1 ist Definition 5.1 erläutert. Für einen Vektor y^1, der ungleich y^2 und kleiner als y^2 im Sinne von $\leq$ ist, gelten $y^1_i \leq y^2_i$, $1 \leq i \leq k$, und $y^1_j < y^2_j$ für mindestens ein $j \in \{1, \ldots, k\}$. Mit diesem Ordnungskonzept lassen sich zwei Vektoren vergleichen.

Bemerkung 5.2. Wesentliche Eigenschaften der Ordnungsrelation $\leq$ sind wie folgt:

(1) Es gibt Paare $(y^1, y^2) \in \mathbb{R}^k \times \mathbb{R}^k$, die nicht vergleichbar bezüglich der Ordnungsrelation sind. Damit gilt weder $y^1 \leq y^2$ noch $y^2 \leq y^1$. Ein Beispiel haben wir bereits mit $y^1 = (4, 2)$ und $y^2 = (2, 4)$ gegeben. Wir erhalten $y^1 - y^2 = (2, -2) \notin \mathbb{R}^k_+$ und $y^2 - y^1 = (-2, 2) \notin \mathbb{R}^k_+$. Unterschiedliche Vektoren können von unterschiedlicher Signifikanz sein. Das ist ein wesentlicher Unterschied zur skalarwertigen Optimierung, wo der Raum, in den die Zielfunktion abbildet (nämlich $\mathbb{R}$), eine totale Ordnung besitzt.

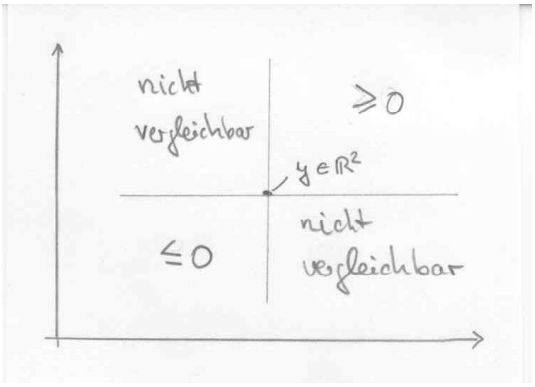

ABBILDUNG 5.1. Bei gegebenem Vektor y ergibt sich in $\mathbb{R}^2$ die in der Abbildung gezeigte Situation. Dabei ist $z \geq y$ äquivalent mit $y \leq z$, das heißt, $z - y \in \mathbb{R}^k_+$.

(2) Die Ordnungsrelation $\leq$ ist eine partielle Ordnung im $\mathbb{R}^k$, denn es gelten folgende Aussagen:
   - $y \leq y$ für alle $y \in \mathbb{R}^k$ (Reflexivität),
   - $y^1 \leq y^2$ und $y^2 \leq y^3$ implizieren $y^1 \leq y^3$ für $y^1, y^2, y^3 \in \mathbb{R}^k$ (Transitivität),
   - $y^1 \leq y^2$ und $y^2 \leq y^1$ ergeben $y^1 = y^2$ für $y^1, y^2 \in \mathbb{R}^k$ (Antisymmetrie),
   - $y^1 \leq y^2$ und $y^3 \leq y^4$ implizieren $y^1 + y^3 \leq y^2 + y^3$ für $y^1, y^2, y^3, y^4 \in \mathbb{R}^k$ (Verträglichkeit mit der Addition),
   - $y^1 \leq y^2$, $\alpha \in \mathbb{R}_+$ ergeben $\alpha y^1 \leq \alpha y^2$ für $y^1, y^2 \in \mathbb{R}^k$ (Verträglichkeit mit skalarer Multiplikation).

(3) Da der nicht-negative Orthant $\mathbb{R}^k_+$ ein konvexer Kegel ist, wird $\leq$ auch Kegel-Halbordnung genannt. Insbesondere gelten:
   - für $y^1, y^2 \in \mathbb{R}^k_+$, $y^1 \leq y^2$, $\lambda \in \mathbb{R}$, $\lambda \geq 0$ folgt $\lambda y^1 \leq \lambda y^2$,
   - für $y^1, y^2, y^3 \in \mathbb{R}^k$ mit $y^1 \leq y^2$ folgt $y^1 + y^3 \leq y^2 + y^3$. ◊

Auf der Grundlage dieser Ordnungsrelation können wir nun das Ziel der Vektoroptimierung formulieren: Gesucht sind die Punkte $x^* \in \mathbb{R}^n$, so dass die Zielfunktions-Vektoren $\mathbf{f}(x^*)$ minimal bezüglich der Ordnungsrelation sind. Dabei wird das Ziel mit Hilfe der Einführung von effizienten Punkten $y^* \in \mathbb{R}^k$ formuliert. Weitere Literatur zu diesem Thema finden wir zum Beispiel in [**15, 31**].

DEFINITION 5.3 (Effizienter Punkt, optimaler Pareto-Punkt, dominierender Punkt). *Sei $\mathbf{f}(R)$ das Bild der zulässigen Menge $R \subset \mathbb{R}^n$ unter der Abbildung $\mathbf{f} : \mathbb{R}^n \to \mathbb{R}^k$. Ein Punkt $y^* \in \mathbf{f}(R)$ heißt* (global) effizient *bezüglich der Ordnungsrelation '$\leq$' im $\mathbb{R}^k$ genau dann, wenn kein $y \in \mathbf{f}(R)$ existiert mit $y \neq y^*$ und $y \leq y^*$. Ein Punkt $x^* \in R$ mit $y^* = \mathbf{f}(x^*)$ heißt* (global) Pareto-optimal *genau dann, wenn y^* ein effizienter Punkt ist. Ein Punkt $x^1 \in R$* dominiert *einen Punkt $x^2 \in R$ genau dann, wenn $\mathbf{f}(x^1) \neq \mathbf{f}(x^2)$ und $\mathbf{f}(x^1) \leq \mathbf{f}(x^2)$ gelten.*

Damit besteht das Ziel der Vektoroptimierung, effiziente Punkte $y^* \in \mathbf{f}(R)$ und optimale Pareto-Punkte $x^* \in R$ mit $y^* = \mathbf{f}(x^*)$ zu finden. Sind die Komponenten

f_j, $1 \leq j \leq k$, der Zielfunktion von der gleichen Wertigkeit, so können wir apriori nicht einen effizienten Punkt von einem anderen unterscheiden. An dieser Stelle bleibt uns nur, alle effizienten Punkte $y^* \in \mathbf{f}(R) \subset \mathbb{R}^k$ zusammen mit den Pareto-Punkten $x^* \in R \subset \mathbb{R}^n$ zu bestimmen. Von dieser effizienten Menge und der Pareto-Menge muß dann in der Anwendung entschieden werden, welche spezielle Lösung realisiert werden soll.

Wir wollen hier auch das folgende Konzept einführen, bei dem der Vergleich nur lokal vorgenommen wird.

Definition 5.4 (Lokaler optimaler Pareto-Punkt). *Ein Punkt $x^* \in R$ heißt* lokal Pareto-optimal *genau dann, wenn eine Umgebung $U(x^*) \subset \mathbb{R}^n$ von x^* existiert, so dass $y^* = \mathbf{f}(x^*)$ effizient bezüglich des Bildraumes $f(R \cap U(x^*))$ ist. Analog definieren wir* lokal effiziente Punkte $y^* \in \mathbf{f}(R)$.

Bemerkung 5.5. Manchmal wird der Begriff des effizienten Punktes auch in der Urbildmenge $\mathbb{R}^n$ von $\mathbf{f}$ definiert: Ein Punkt $x^* \in \mathbb{R}^n$ heißt *effizienter Punkt* oder *Pareto-Punkt*, wenn kein $x \in \mathbb{R}^n$ existiert mit

$$\mathbf{f}(x) \neq \mathbf{f}(x^*) \quad \text{und} \quad \mathbf{f}(x) \leq \mathbf{f}(x^*). \tag{5.1}$$

Gilt (5.1) nur in einer Umgebung $U^* \subset \mathbb{R}^n$ von x^*, so heißt x^* *lokal effizient* oder *lokaler Pareto-Punkt*. Ein Punkt $x_1 \in \mathbb{R}^n$ wird durch einen Punkt $x_2 \in \mathbb{R}^n$ *dominiert*, wenn $\mathbf{f}(x_1) \neq \mathbf{f}(x_2)$ und $\mathbf{f}(x_1) \geq \mathbf{f}(x_2)$ gelten. ◊

Das Problem, dass Optimierungs-Verfahren lokale Lösungen berechnen, die Anwendung aber oft nach globalen Lösungen fragt, haben wir in der Vektoroptimierung genauso wie in der skalarwertigen Optimierung. Eine mögliche Strategie, um das Problem zu lösen, ist es, sowohl in der skalarwetigen als auch in der Vektor-Optimierung stochastische Suchverfahren einzusetzen. Wir verweisen hier zum Beispiel auf [**32**]. Ein Nachteil der stochastischen Verfahren ist die große Anzahl an erforderlichen Auswertungen der Zielfunktion. Insbesondere bei industriellen Anwendungen ist das oft ein Problem, da hinter jeder Auswertung der Zielfunktion die Simulation des Systems stehen kann.

2. Beziehung zur skalarwertigen Optimierung

Wir betrachten das Problem

$$\min \mathbf{f}(x) \quad \text{u.d.N} \quad x \in R, \tag{5.2}$$

wobei $\mathbf{f} : \mathbb{R}^n \to \mathbb{R}^k$ gilt und die Menge $R \subset \mathbb{R}^n$ durch

$$R = \left\{ x \in \mathbb{R}^n \,|\, h_i(x) = 0,\ i = 1, \ldots, m \text{ und } h_j(x) \leq 0,\ j = m+1, \ldots, m+q \right\} \tag{5.3}$$

gegeben ist. Wir setzen voraus, dass die Funktionen $\mathbf{f}$ und $h_i : \mathbb{R}^n \to \mathbb{R}$ stetig differenzierbar sind.

Zum Nachweis des nächsten Resultats verweisen wir auf [**15**].

Satz 5.6 (Notwendige Bedingungen für optimalen Pareto-Punkt). *Sei $x^* \in \mathbb{R}$ ein optimaler Pareto-Punkt für (5.2). Wir bezeichnen mit $\mathcal{A}^* \subset \{1, \ldots, m+q\}$ die Menge der aktiven Indizes, für die $h_l(x^*) = 0$ für alle $l \in \mathcal{A}^*$ gilt. Dann existieren*

Vektoren

$$\alpha^* \in \mathbb{R}^k_+ \text{ mit } \sum_{i=1}^{k} \alpha_i^* = 1, \tag{5.4}$$

$$\lambda^* \in \mathbb{R}^{m+q}, \tag{5.5}$$

so dass

$$\sum_{i=1}^{k} \alpha_i^* \nabla f_i(x^*) + \sum_{j=1}^{m+q} \lambda_j^* \nabla h_j(x^*) = 0, \tag{5.6a}$$

$$h_i(x^*) = 0, \quad i = 1, \ldots, m, \tag{5.6b}$$

$$\lambda_j^* \geq 0,\ h_j(x^*) \leq 0,\ \lambda_j^* h_j(x^*) = 0, \quad j = m+1, \ldots, m+q \tag{5.6c}$$

gelten, wobei die Gradienten ∇f_i, $1 \leq i \leq k$, und ∇h_j, $1 \leq j \leq m+q$, Spaltenvektoren sind. Sind die Abbildungen f_i und h_i konvex, so sind die Bedingungen (5.6) *auch hinreichend.*

Bemerkung 5.7. Die notwendigen Optimalitätsbedingungen erster Ordnung werden Karush-Kuhn-Tucker oder auch kurz KKT-Bedingungen genannt. ◊

Für $\alpha \in \mathbb{R}^k_+$ definieren wir die skalarwertige Funktion

$$g_\alpha : \mathbb{R}^n \to \mathbb{R}, \quad x \mapsto \sum_{i=1}^{k} \alpha_i f_i(x) \tag{5.7}$$

und beobachten, dass $\sum_{i=1}^{k} \alpha_i \nabla f_i(x) = \nabla g_\alpha(x)$ gilt. Damit sind die Gleichungen (5.6) äquivalent damit, dass der Punkt x^* ein KKT-Punkt für

$$\min g_\alpha(x) \quad \text{u.d.N.} \quad x \in R$$

ist mit $\alpha = \alpha^*$ (vergleiche [**20, 23**] und [**18**]). Aus (5.4) und (5.7) schliessen wir, dass g_α eine Konvex-Kombination der Komponenten f_i der Zielfunktion ist, wobei die α_i's die entsprechenden relativen Gewichte sind. Auf dieser Beobachtung basiert die Wichtungsmethode (siehe Abschnitt 2.1).

Bedingungen zweiter Ordnung sind bei der Vektoroptimierung nicht vorhanden. Das ist der Preis dafür, dass wir bei der Vektoroptimierung keine totale, sondern nur eine Halb-Ordnung vorliegen haben. Notwendige Bedingungen zweiter Ordnung für einen Punkt x^*, der g_α lokal minimiert, sind nicht notwendig dafür, dass x^* ein optimaler Pareto-Punkt zu sein. Hier liegt ein wesentlicher Unterschied zwischen der skalarwertigen und der Vektoroptimierung. Im Prinzip können auch Sattelpunkte von einer konvexen Kombination g_α optimale Pareto-Punkte sein.

Aus Satz 5.6 folgt, dass ein optimaler Pareto-Punkt x^* ein Karush-Kuhn-Tucker Punkt für die Funktion g_α ist. Diese Eigenschaft führt dazu, dass der Gewichtungsvektor $\alpha \in \mathbb{R}^k_+$ in (5.4) wichtige Informationen über die lokale Geometrie der Menge der effizienten Punkte enthält.

Wir betrachten dazu das bikriterielle Beispiel in die Abbildung 5.2. Seien also $k = 2$ und x^* ein optimaler Pareto-Punkt von $\mathbf{f} = (f_1, f_2)$. Weiter sei x^* ein Minimum von g_α, das heißt, es gibt keinen anderen Punkt $\bar{x} \in R$ mit

$$g_\alpha(\bar{x}) < g_\alpha(x^*) =: c^*. \tag{5.8}$$

Die Menge von allen $y \in \mathbb{R}^2$ im Bildraum

$$\text{Bild } \mathbf{f}(R) = \{z \in \mathbb{R}^k \,|\, \text{es gibt ein } x \in R \text{ mit } z = \mathbf{f}(x)\}$$

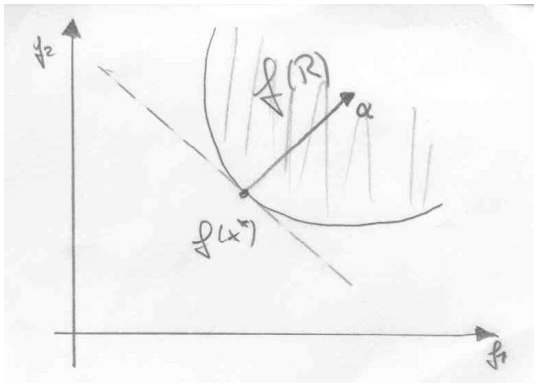

ABBILDUNG 5.2. Detail der Bildmenge $\mathbf{f}(R)$ und der anliegenden Geraden $\alpha^T y = c$, siehe Text.

von $\mathbf{f}$, für die gilt

$$\mathbf{f}(x) = y \quad \Longrightarrow \quad g_\alpha(x) = c^*,$$

bilden die Gerade $\alpha^T y = c$, die durch den Normalvektor α definiert ist, siehe Abbildung 5.2. Punkte im Bildbereich mit kleinerem Wert von g_α liegen links/unterhalb der Geraden, Punkte mit größerem Wert hingegen rechts/oberhalb der Geraden. Die Menge $\mathbf{f}(R)$ muß komplett auf der rechten Seite der Geraden liegen, damit es keinen Punkt $\bar{x} \in R$ mit (5.8) gibt. Ist der Rand der Menge $\mathbf{f}(R)$ durch eine stetig differenzierbare Kurve parametrisierbar, so ist der Winkel zwischen der Tangente an den Rand der effizienten Punkte und der Geraden g_α gleich null. Damit stimmen die Tangente und die Gerade überein. Also ist α der Normalvektor der Tangente an die Menge der effizienten Punkte, siehe [**5, 6**].

Die geometrische Bedeutung von α lässt sich vom bikriteriellen auf den multikriteriellen Fall übertragen. Dazu zitieren wir den folgenden Satz (siehe [**17**, Theorem 4.2]) für unrestringierte Vektoroptimierungs-Probleme

$$\min \mathbf{f}(x) \quad \text{u.d.N.} \quad x \in \mathbb{R}^n \tag{5.9}$$

mit $\mathbf{f} = (f_1, \ldots, f_k) : \mathbb{R}^n \to \mathbb{R}^k$.

SATZ 5.8. *Sei $y^* \in \mathbb{R}^k$ ein lokaler effizienter Punkt von* (5.9) *und x^* ein assozierter lokaler optimaler Pareto-Punkt, das heißt, $y^* = \mathbf{f}(x^*)$. Wir setzen $g_{\alpha^*}(x) = \sum_{i=1}^k \alpha_i^* f_i(x)$, wobei $\alpha^* \in \mathbb{R}^k_+$ den Gewichtsvektor aus* (5.5) *bezeichnet. Dann liegt α^* im orthogonalen Komplement von* Bild $\mathbf{f}'(x^*) \subset \mathbb{R}^k$, *wobei $\mathbf{f}'(x^*)$ die Jacobi- oder Funktionalmatrix von $\mathbf{f}$ am Punkt x^* bezeichnet.*

PROOF. Von den notwendigen Optimalitätsbedingungen erster Ordnung erhalten wir

$$\begin{aligned} \nabla g_{\alpha^*}(x^*) = 0 &\Longleftrightarrow \sum_{i=1}^k \alpha_i^* \nabla f_i(x^*) = 0 \\ &\Longleftrightarrow \left(\alpha^*\right)^T \begin{pmatrix} \nabla f_1(x^*) \\ \vdots \\ \nabla f_k(x^*) \end{pmatrix} = 0 \Longleftrightarrow \left(\alpha^*\right)^T \nabla f(x^*) = 0. \end{aligned} \tag{5.10}$$

Aus (5.10) schliessen wir, dass α^* orthogonal zu den Spalten der Jacobi-Matrix $\nabla f(x^*)$ ist und

$$\big(\alpha^*\big)^T\big(\nabla f(x^*)v\big) = 0 \quad \text{für alle } v \in \mathbb{R}^n,$$

woraus die Behauptung folgt. □

Offenbar folgt aus Satz 5.8, dass der Rang der Jacobimatrix $\mathbf{f}'(x^*)$ kleiner als m sein muß. Damit ist die Abbildung $\mathbf{f}'(x^*) : \mathbb{R}^n \to \mathbb{R}^k$ nicht surjektiv. Gilt insbesondere $\dim \text{Bild } \mathbf{f}'(x^*) = k-1$, so ist der Vektor α eindeutig bestimmt. Das Resultat aus Satz 5.8 lässt sich auf den restringierten Fall übertragen. Sei $y^* \in \mathbb{R}^k$ ein effizienter Punkt mit $\mathbf{f}(x^*) = y^*$ und seien im Punkt x^* die Nebenbedingungen h_i, $i \in \mathcal{A}^*$ aktiv: $h_i(x^*) = 0$ für $i \in \mathcal{A}^*$. Mit $p \le m+q$ bezeichnen wir die Anzahl der Indizes in $\mathcal{A}^*$. Wir nehmen an, dass die Vektoren $\{\nabla h_i(x^*)\}_{i\in\mathcal{A}^*}$ linear unabhängig sind (*constrained qualification*). Wir betrachten die $n-p$-dimensionale Menge

$$\mathcal{N}^* = \big\{x \in U^* \,|\, h_i(x) = 0 \text{ for } i \in \mathcal{A}^*\big\} \subset \mathbb{R}^n$$

in einer Umgebung $U^* \subset \mathbb{R}^n$ von x^*. Sei $s : T \to V$ eine lokale, stetig differenzierbare Parametrisierung der Menge $\mathcal{N}^*$, wobei $T \subset \mathbb{R}^{n-p}$ offen und $V \subset N$ eine offene Umgebung von x^* sind. Dann liegt α im orthogonalen Komplement von Bild $\nabla\tilde{\mathbf{f}}(t^*)$ mit $\tilde{\mathbf{f}}(t) = \mathbf{f}\circ s(t)$ (lokale Parametrisierung von $\mathbf{f}$ auf der Menge $\mathcal{N}^*$) und $s(t^*) = x^*$.

3. Die KKT-Punkte als differenzierbare Mannigfaltigkeit

Für jeden optimalen Pareto-Punkt gibt es nach Satz 5.6 einen Gewichtsvektor $\alpha^* \in \mathbb{R}^k_+$, so dass x^* ein KKT-Punkt für die skalarwertige Funktion $g_{\alpha^*}(x) = \sum_{i=1}^k \alpha_i^* f_i(x)$ ist. Wenn die zulässige Menge des zugrundeliegenden Vektoroptimierungs-Problems durch m Gleichungen gegeben ist, gilt die folgende Aussage: Für jeden optimalen Pareto-Punkt x^*, für den $\{\nabla h_i(x^*)\}_{i=1}^m$ linear unabhängig sind, gibt es Vektoren $\lambda^* = (\lambda_1^*, \ldots, \lambda_m^*)^T \in \mathbb{R}^m$ und $\alpha^* = (\alpha_1^*, \ldots, \alpha_k^*)^T \in \mathbb{R}^k_+$ mit

$$\sum_{i=1}^k \alpha_i^* \nabla f_i(x^*) + \sum_{i=1}^m \lambda_i^* \nabla h_i(x^*) = 0 \qquad (n \text{ Gleichungen}) \tag{5.11a}$$

$$\mathbf{h}(x^*) = 0 \qquad (m \text{ Gleichungen}) \tag{5.11b}$$

$$\sum_{i=1}^k \alpha_i^* = 1 \qquad (1 \text{ Gleichung}) \tag{5.11c}$$

In (5.11b) haben wir $\mathbf{h} = (h_1, \ldots, h_m)$ gesetzt. Nun definieren wir die Abbildung $\mathbf{F} : \mathbb{R}^{n+m+1} \to \mathbb{R}^{n+m+1}$ durch

$$\mathbf{F}(x,\lambda,\alpha) = \begin{pmatrix} \sum_{i=1}^k \alpha_i \nabla f_i(x) + \sum_{i=1}^m \lambda_i \nabla h_i(x) \\ \mathbf{h}(x) \\ \sum_{i=1}^k \alpha_i - 1 \end{pmatrix}.$$

Wir können dann (5.11) in der Form

$$\mathbf{F}(x^*, \lambda^*, \alpha^*) = 0 \tag{5.12}$$

schreiben. Andererseits sind Lösungen $(x^*, \lambda^*, \alpha^*)$ von (5.12) mit $\alpha^* \in \mathbb{R}^k_+$ Kandidaten für optimale Pareto-Punkte. In [**17**, Theorem 5.1] finden wir das folgende Resultat. Der Beweis beruht auf dem Satz über implizite Funktionen.

SATZ 5.9. *Seien* $\bar{\mathbb{R}}^k_+ = \{\alpha \in \mathbb{R}^k \mid \alpha_i > 0 \text{ für } i = 1, \dots, k\}$ *und*

$$M = \big\{(x, \lambda, \alpha) \in \mathbb{R}^{n+m+1} \mid \mathbf{F}(x, \lambda, \alpha) = 0 \text{ und } \alpha \in \bar{\mathbb{R}}^k_+\big\}.$$

Gilt

$$\text{Rang } \mathbf{F}'(x, \lambda, \alpha) = n + m + 1 \quad \text{für alle } (x, \lambda, \alpha) \in \mathbb{R}^{n+m+1}, \tag{5.13}$$

so ist M eine $k-1$ dimensionale Fläche in $\mathbb{R}^{n+m+1}$, die sich durch eine differenzierbare Funktion parametrisieren lässt.

In [**17**, Abschnitt 5.2] finden wir notwendige und hinreichende Bedingungen für die Voraussetzung (5.13) von Satz 5.9. Insbesondere gilt das folgende Resultat (siehe [**17**, Corollary 5.5]:

KOROLLAR 5.10. *Sei $(x^*, \lambda^*, \alpha^*) \in M$, das heißt, x^* ist ein KKT-Punkt für die skalarwertige Funktion g_{α^*} unter der Nebenbedingung $\mathbf{h}(x^*) = 0$. Ferner seien die Vektoren $\{\nabla h_i(x^*)\}_{i=1}^m$ linear unabhängig (constrained qualification). Wir setzten*

$$\mathcal{K}^* = \text{Kern } \mathbf{h}'(x^*).$$

Weiter sei x^* ein lokales Minimum von g_{α^*}, an dem die hinreichenden Bedingungen zweiter Ordnung gelten:

$$\nabla^2 L_\alpha(x^*, \lambda^*) = \sum_{i=1}^k \alpha_i^* \nabla^2 f_i(x^*) + \sum_{i=1}^m \lambda_i^* \nabla^2 h_i(x^*) \text{ ist positiv definit auf } \mathcal{K}^*$$

oder x^* sei ein Sattelpunkt von g_α und $\nabla^2 L_\alpha(x^*, \lambda^*)$ — eingeschränkt auf $\mathcal{K}^*$ — ist regular mit negativen und positiven Eigenwerten. Dann hat $\mathbf{F}'(x^*, \lambda^*, \alpha^*)$ vollen Rang $n + m + 1$.

In Homotopie-Strategien werden ausgehend von einem Punkt $(x^*, \lambda^*, \alpha^*) \in M$ weitere Punkte (x, λ, α) in M bestimmt. Auf diese Weise erhalten wir Kandidaten für optimale Pareto-Punkte.

4. Ein Überblick über Verfahren

In diesem Abschnitt wollen wir kurz die am häufigsten verwendeten Verfahren beschreiben, die zur Berechnung der Menge der effizienten Punkte (oder meist einer Teilmenge davon) für Probleme der Vektoroptimierung eingesetzt werden.

Zuerst wenden wir uns den deterministischen Methoden zu. Diese basieren (fast alle) darauf, das Problem der Vektoroptimierung auf ein Problem der skalarwertigen Optimierung zurückzuführen. Die Transformation des Vektoroptimierungs-Problem in ein Problem der skalarwertigen Optimierung erfolgt unter der Verwendung von Parametern. Da wir beim Lösen des skalarwertigen Optimierungs-Problems in der Regel jeweils nur einen effizienten Punkt berechnen, erhalten wir die Menge (oder letztlich eine Teilmenge) der effizienten Punkte des Vektoroptimierungs-Problems durch Variation der Parameter bei der Transformation des Vektoroptimierungs- in das skalarwertige Optimierungs-Problem.

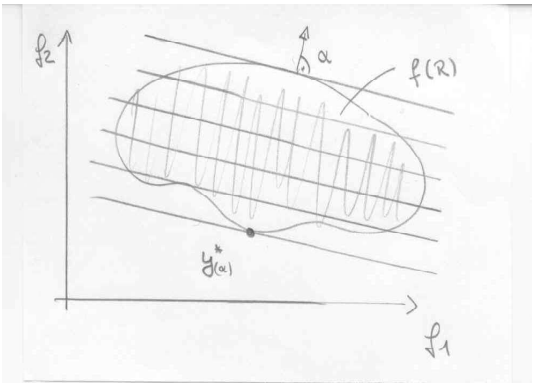

ABBILDUNG 5.3. Das Wichtungsverfahren für die Lösung eines Vektoroptimierungs-Problems in $\mathbb{R}^k$ mit $k = 2$.

4.1. Die Wichtungsmethode. Dieses Verfahren wurde zuerst von Zadeh im Jahre 1963 eingeführt und wird am häufigsten verwendet, um Probleme der Vektoroptimierung zu lösen. Dabei wird jeder Komponente der Zielfunktion wird ein Gewicht $\alpha_i \geq 0$, $1 \leq i \leq k$, zugeordnet mit der Eigenschaft $\sum_{i=1}^{k} \alpha_i = 1$. Wir lösen dann das Problem

$$\min_{x \in R} \sum_{i=1}^{k} \alpha_i f_i(x). \tag{5.14}$$

Der Transformations-Parameter ist der Gewichtsvektor $\alpha = (\alpha_1, \ldots, \alpha_k)^T$, der die relative Signifikanz der einzelnen Komponenten der Zielfunktion ausdrückt. Unter Variation von α können wir eine Teilmenge der Menge der effizienten Punkte berechnen. Globale Minimierer der skalarwertigen Zielfunktion $\sum_{i=1}^{k} \alpha_i f_i(x) = \alpha^T \mathbf{f}(x)$ sind notwendigerweise auch globale optimale Pareto-Lösungen (bei $\alpha_i > 0$ für alle $i \in \{1, \ldots, k\}$, sonst nur bei eindeutigen globalen Lösungen von (5.14)). Ebenso sind auch lokale Minimierer von (5.14) notwendigerweise lokale Pareto-Lösungen.

Eine geometrische Interpretation des Verfahrens kann durch die Betrachtung der skalarwertigen Funktion $g_\alpha(x) = \alpha^T \mathbf{f}(x)$ durchgeführt werden. Durch die Gleichung $g_\alpha(x) = c \in \mathbb{R}$ wird eine Ebene im Raum $\mathbb{R}^k$ aufgespannt, die durch den Normalvektor $\alpha \in \mathbb{R}^k$ charakterisiert ist. Jede Wahl für α führt auf unterschiedliche Ebenen, siehe Abbildung 5.3. Die Methode hat folgende Nachteile (siehe [**6**]):

- Bei Vektoroptimierungs-Problemen ist die Menge $\mathbf{f}(R)$ im allgemeineinen nicht konvex in $\mathbb{R}^k$. In diesem Fall gibt es Beispiele, bei denen die effizienten Punkte keine Minimierer einer skalarwertigen Zielfunktion der Bauart $\alpha^T \mathbf{f}(x)$ sind. Diese Punkte lassen sich nicht mit diesem Verfahren bestimmen.
- Bei jedem numerischen Verfahren zur Lösung eines Vektoroptimierungs-Problems lassen sich nur endlich-viele effiziente Punkte berechnen. Eine gleichmäßige Verteilung dieser Punkte in der Menge $\mathbb{R}^k$ (Zielfunktionsraum) kann durch die Gewichtungs-Methode nicht erreicht werden. Die Abstände zwischen zwei effizienten Punkten kann nicht direkt durch das Verfahren gesteuert werden.

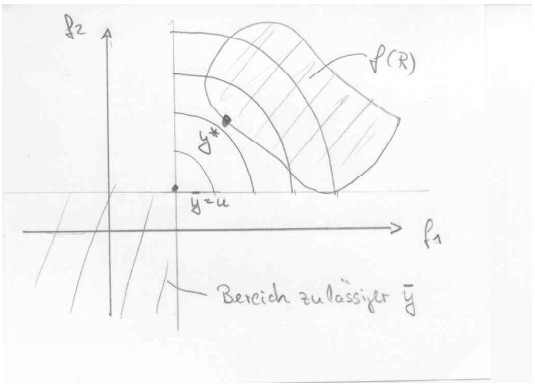

ABBILDUNG 5.4. Die gewichtete ℓ^p-Methode mit der Zielvorgabe $\bar{y} = u$, dem Gewichtsvektor $\omega = (1,1)^T$, der Dimension $k = 2$ und der ℓ^p-Norm für $p = 2$.

4.2. Das gewichtete ℓ^p-Verfahren. Bei diesem Verfahren werden ein Referenz-Punkt $\bar{y} \in \mathbb{R}^k$ gewählt und dann effiziente Punkte gesucht, die $\bar{y}$ in einer ℓ^p-Norm möglichst nahe kommen. Daher wird diese Methode auch Zieloptimierung und Normskalierung genannt (vergleiche [**15**, Abschnitt 3.2.2]). Dabei können sowohl die Wahl des Referenz-Punktes als auch die verwendete Metrik, in der der Abstand einer Lösung $\mathbf{f}(\bar{x})$ zu $\bar{y}$ gemessen wird, variiert werden. Im Fall der Wahl der gewichteten ℓ^p-Norm

$$d_p(y) = \left(\sum_{i=1}^{k} \omega_i \, |y_i|^p \right)^{1/p}, \quad p \in [1, \infty), \quad y = (y_1, \ldots, y_k) \in \mathbb{R}^k$$

mit $\omega_i > 0$ für $1 \leq i \leq k$ und $\sum_{i=1}^{k} \omega_i = 1$ erhalten wir das folgende skalarwertige Optimierungsproblem

$$\min_{x \in R} \sum_{i=1}^{k} \omega_i \, |f_i(x) - \bar{y}_i|^p. \tag{5.15}$$

Im Fall von $p = \infty$ definieren wir $d_\infty(y) = \max\{\omega_1 \, |y_1|, \ldots, \omega_k \, |y_k|\}$ als gewichtete Maximumnorm in $\mathbb{R}^k$. In Abbildung 5.4 haben wir das Vorgehen dieser Methode graphisch erläutert. Der Referenz-Punkt $\bar{y}$, der Gewichtungsvektor ω sowie der Exponent p sind die Transformations-Parameter bei der Überführung des Vektoroptimierungs–Problems in (5.15). Dabei muß $\bar{y}$ der Forderung $\bar{y}_i \leq f_i(x)$ für alle $x \in R$ und $i \in \{1, \ldots, k\}$. Das ist beispielsweise der Fall, wenn $\bar{y}_i = \operatorname{argmin}\{f_i(x) : x \in R\}$ für $1 \leq i \leq k$ gesetzt wird. Im Fall von $p \in [1, \infty)$ sind die erhaltenen Lösungen von (5.15) notwendigerweise effiziente Punkte. Im Fall von $p = \infty$ können alle effizienten Punkte ebenfalls berechnet werden, wobei aber eine sinnvolle Variation der Parameter $(\bar{y}, \omega, p)$ durchgeführt werden muß. Das ist das Problem dieser Methode. Insbesondere kann nicht gesagt werden, mit welcher Strategie zur Variation der Parameter $(\bar{y}, \omega, p)$ alle effizienten Punkte berechnet werden können, siehe [**15**].

4.3. Die ε-Methode oder Methode der Beschränkungen. Dieses Verfahren geht zurück auf Marglin [**27**] und Haimes [**16**]. Dabei wird eine Komponente

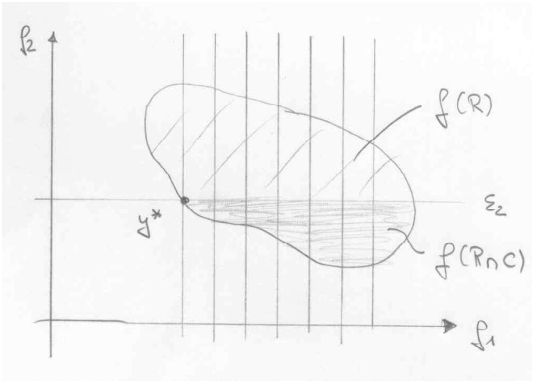

ABBILDUNG 5.5. Die ε-Methode oder Methode der Beschränkungen mit $j = 1$ und $k = 2$.

$j \in \{1, \ldots, k\}$ der Zielfunktion ausgewählt und diese dann minimiert. Für alle anderen Komponenten wird eine obere Schranke festgelegt, die nicht überschritten werden darf. Damit hat das skalarwertige Problem die folgende Darstellung

$$\begin{aligned} &\min\big\{f_j(x) \,:\, x \in R \cap C\big\}, \quad j \in \{1, \ldots, k\} \\ &C = \{x \in \mathbb{R}^n \mid f_i(x) \leq \varepsilon_i \text{ für alle } 1 \leq i \leq k \text{ mit } i \neq j\}. \end{aligned} \tag{5.16}$$

Ein eindeutiger globaler Minimierer von (5.16) ist notwendigerweise auch eine globale optimale Pareto-Lösung des Vektoroptimierungs-Problems. Bei dieser Methode sind der Index $j \in \{1, \ldots, k\}$ sowie die oberen Schranken ε_i die Transformations-Parameter bei der Überführung des Vektoroptimierungs-Problem in (5.16). Prinzipiell sind daher alle effizienten Punkte berechenbar, wenn die Transformations-Parameter variiert werden. Das Hauptproblem dieses Verfahrens besteht darin, den Bereich von sinnvollen Werten für die ε_i festzulegen. Werden die ε_i zu klein gewählt, dann kann (5.16) unter Umständen keine Lösung besitzen. Zu starke Restriktionen erzeugen dann leere zulässige Mengen. Andererseits, wenn eine obere Schranke zu groß gewählt ist, so geht die dazugehörige Komponente der Zielfunktion bei Variation des zugehörigen ε_i's kaum in die Problemstellung (5.16) ein. Das kann dazuführen, dass keine neuen effizienten Punkte berechnet werden, obwohl die Parameter variiert werden.

4.4. Das Verfahren mit Gleichungs-Nebenbedingungen. Diese Strategie ist von [**25**] entwickelt worden. Analog wie bei der vorigen Methode wird wieder ein Index $j \in \{1, \ldots, k\}$ ausgewählt und die entsprechenede Komponente der Zielfunktion minimiert. Hier gehen aber die verbleibenden Komponenten f_i, $i \in \{1, \ldots, k\} \setminus \{j\}$ durch Gleichungs-Nebenbedingungen ein:

$$\begin{aligned} &\min\big\{f_j(x) \,:\, x \in (R \cap D)\big\}, \quad j \in \{1, \ldots, k\} \\ &C = \{x \in \mathbb{R}^n \mid f_i(x) = \varepsilon_i \text{ für alle } 1 \leq i \leq k \text{ mit } i \neq j\}. \end{aligned} \tag{5.17}$$

Prinzipiell können bei diesem Verfahren wieder alle effizienten Punkte durch Variation der Transformations-Parameter (j und ε_i für $i \in \{1, \ldots, k\} \setminus \{j\}$) berechnet werden. Andererseits zeigt Abbildung 5.5, dass nicht jede Lösung von (5.17) optimaler Pareto-Punkt ist. Das muss durch Überprüfen von notwendigen und hinreichenden Kriterien festgestellt werden, siehe [**25**]. Analog zur ε-Methode hat dieses

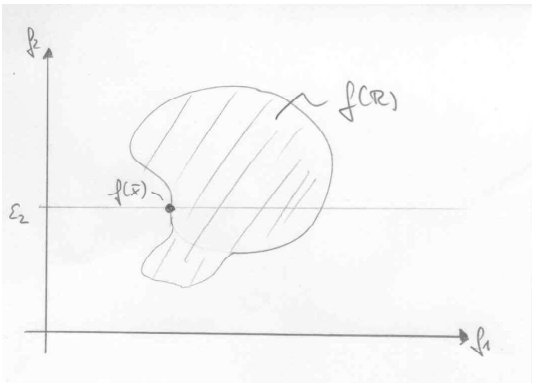

ABBILDUNG 5.6. Illustration der Situation bei dem Verfahren mit Gleichungs-Nebenbedingungen, wo der globale Minimierer $\bar{x}$ von (5.17) bei gewähltem ε_2 kein optimaler Pareto-Punkt ist mit $j = 1$ und $k = 2$.

Verfahren auch den Nachteil, dass (5.17) unter Umständen für viele Parameterwerte (für j und ε_i für $i \in \{1, \ldots, k\} \setminus \{j\}$) keine zulässigen Lösungen besitzt.

4.5. Die lexikographische Methode. Gemäß der Wertigkeit der Komponenten der Zielfunktion $\mathbf{f} = (f_1, \ldots, f_k)$ wird eine Anordnung der Komponenten angegeben:

$$i_j \in \{1, \ldots, k\} \text{ für } j \in \{1, \ldots, k\}, \quad \text{und} \quad i_j \neq i_l \text{ für } j \neq l.$$

Dann werden für $j = 1, \ldots, k$ die Aufgaben

$$\min_{x \in R_j} f_{i_j}(x) \tag{5.18}$$

gelöst, wobei $R_1 = R \subset \mathbb{R}^n$ die zulässige Menge für das zugrundeliegende Vektoroptimierungs-Problem ist und $R_j = \operatorname{argmin}\{f_{i_{j-1}}(x) \,|\, x \in R_{j-1}\}$ für $2 \leq j \leq k$ gilt. Das Verfahren endet bereits für $j < k$, wenn die Menge R_j einelementig ist; denn in diesem Fall gilt $R_j = R_{j+1} = \ldots = R_k$. Es lässt sich zeigen, dass wir im Fall der Lösbarkeit von (5.18) optimale Pareto-Punkte berechnen, das heißt, die Menge R_k ist eine Teilmenge der Menge aller optimalen Pareto-Punkte des Vektoroptimierungs-Problems. Wir können das Verfahren mit der Methode der Beschränkungen kombinieren, indem wir wie folgt vorgehen:

(1) Löse
$$\min f_{i_1}(x) \quad \text{u.d.N.} \quad x \in R$$
und setze $f^*_{i_1} = \min_{x \in R} f_{i_1}(x)$, $j = 0$ und $R_1 = R$.

(2) Setze $j = j + 1$ und gebe ein $\varepsilon_j > 0$ vor, um den der minimale Wert der Komponentenfunktion f_{i_j} vergrößert werden darf. Löse dann
$$\min_{x \in R_j} f_{i_j}(x) \quad \text{u.d.N.} \quad x \in R_{j+1} := \left\{x \in R_j \,|\, f_{i_j}(x) \leq \varepsilon_j - f^*_{i_j}\right\}.$$
Gilt $j + 1 = k$, so breche ab. Andernfalls setze $f^*_{i_{j+1}} = \min_{x \in R_{j+1}} f_{i_j}(x)$ und wiederhole Schritt (2).

Die lexikographische Methode berechnet im allgemeinen nicht alle optimalen Pareto-Punkte. Dazu betrachten wir als Beispiel für $k = 2$ die Abbildung 5.7. Offenbar ist

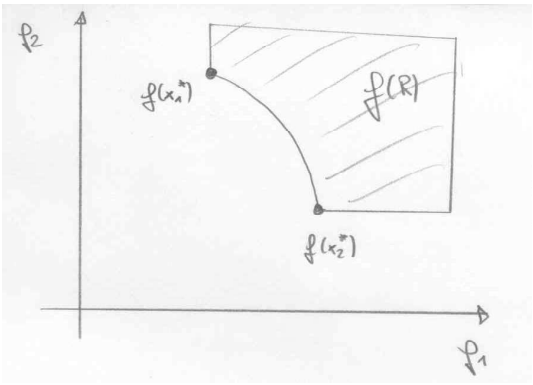

ABBILDUNG 5.7. Illustration der lexikographischen-Methode für ein konkretes Beispiel mit $k = 2$.

der Punkt $f(x_1^*)$ die Lösung der lexikografischen Methode für

$$\min_{x \in R} \begin{pmatrix} f_1(x) \\ f_2(x) \end{pmatrix}.$$

Andererseits erhalten wir die Lösung $f(x_2^*)$, wenn wir

$$\min_{x \in R} \begin{pmatrix} f_2(x) \\ f_1(x) \end{pmatrix}$$

betrachten. Es besteht aber die gesamte Kurve zwischen den beiden Punkten $f(x_1^*)$ und $f(x_2^*)$ aus optimalen Pareto-Punkten.

4.6. Das Goal-Attainment Verfahren. In diesem Abschnitt widmen wir uns der Goal-Attainment Methode, die in der MATLAB Optimierungs-Toolbox als Routine `fgoalattain` zur Verfügung steht (siehe Abschnitt 3.5). Wir betrachten das Problem

$$\min_{x \in R} \mathbf{f}(x), \tag{5.19}$$

wobei die zulässige Menge R gegeben ist durch

$$R = \left\{ x \in X_{\text{ad}} \,\middle|\, h_i(x) = 0,\ i = 1, \ldots, m,\ h_j(x) \leq 0,\ j = m+1, \ldots, m+q \right\}$$

mit $X_{\text{ad}} = \{x \in \mathbb{R}^n \,|\, \underline{x}_i \leq x_i \leq \overline{x}_i,\ i = 1, \ldots, n\}$. Zur Lösung von (5.19) verwenden wir das Goal-Attainment Verfahren, siehe [**14**]. Dazu wird ein Vektor $\mathbf{f}^* = (f_1^*, \ldots, f_k^*) \in \mathbb{R}^k$ (*Goal*) und ein Gewichts-Vektor $\omega = (\omega_1, \ldots, \omega_k) \in \mathbb{R}^k$ gewählt. Dann betrachten wir das Problem

$$\min_{(\gamma, x) \in \mathbb{R} \times R} \gamma \quad \text{u.d.N.} \quad f_i(x) - \omega_i \gamma \leq f_i^* \text{ für } i = 1, \ldots, k. \tag{5.20}$$

Die Problemstellung (5.20) lässt zu, dass die Komponenten f_i, $i = 1, \ldots, k$, der Zielfunktion $\mathbf{f}$ an der optimalen Lösung $x^* \in R$ sowohl größer als auch kleiner als die jeweils vorgegebene i-te Komponente des Goals $\mathbf{f}^*$ sein könner. Die maximale relative Abweichung ist durch den Gewichtsvektor $\gamma\omega$ gegeben, der auch als Slack-Variable interpretiert werden kann. Wählen wir $\omega_i = f_i^* \neq 0$, so folgt für zulässige

Punkte $x \in R$ mit $f_i(x) - \omega_i\gamma \leq f_i^*$, $i = 1, \ldots, k$

$$\frac{f_i(x)}{f_i^*} - \gamma \leq 1 \quad \Longleftrightarrow \quad \frac{f_i(x)}{f_i^*} \leq 1 + \gamma,$$

das heißt, die relative Abweichung zwischen $f_i(x)$ und f_i^*, $i \in \{1, \ldots, k\}$, ist für jede Komponente durch eine von i unabhängige Zahl beschränkt. Wählen wir hingegen $\omega_i = 0$ für $i = 1, \ldots, k$, so erhalten wir $f_i(x) \leq f_i^*$, $i = 1, \ldots, k$.

Zur Lösung von (5.20) kann ein Verfahren der skalarwertigen Optimierung verwendet werden. Beispiele finden wir z.B. in [**10, 11**]. In der MATLAB Optimization Toolbox wird in der Routine `fgoalattain` wird ein SQP-Verfahren verwendet. Dabei wird eine spezielle Merit-Funktion eingesetzt. Dazu schreiben wir (5.20) als Min-Max-Problem

$$\min_{x \in R} \max_{1 \leq i \leq k} \Gamma_i \quad \text{u.d.N.} \quad \Gamma_i = \frac{f_i(x) - f_i^*}{\omega_i}, \ i = 1, \ldots, k. \tag{5.21}$$

Als Merit-Funktion wird

$$\psi(x) = \begin{cases} \mu_i \max\left\{0, f_i(x) - \gamma\omega_i - f_i^*\right\} & \text{für } \omega_i = 0, \\ \max\limits_{1 \leq i \leq k} \Gamma_i & \text{sonst,} \end{cases}$$

eingesetzt, siehe [**4**]. Weiter ist die Hessematrix der Zielfunktion in (5.20) indefinite, da die zweite Ableitung bezüglich der Variable γ verschwindet. In dem Quasi-Newton Verfahren wird die Approximation der Hessematrix so gewählt, dass die zweite Ableitung nach γ auf einen kleinen Wert $\varepsilon = 10^{-8}$ gesetzt.

KAPITEL 6

Kontrolltheorie

Ziel diese Abschnitts ist eine kurze Einführung in das Gebiet der Kontrolltheorie. Wir verwenden hier die Vorgangsweise aus [**9**].

1. Hamilton-Jacobi-Bellman (HJB) Gleichung

Betrachte das Problem

$$\begin{cases} \dot{x}(s) = f(x(s), u(s)) & \text{für } t < s \leq T, \\ x(t) = x_\circ, \end{cases} \tag{6.1}$$

wobei $\dot{x} = \frac{\mathrm{d}x}{\mathrm{d}s}$ bezeichnet und $T > 0$ (Endzeit), $x_\circ \in \mathbb{R}^n$ (Anfangsbedingung), $t \geq 0$ (Anfangszeit) gegeben sind. Ferner sind $U \subset \mathbb{R}^m$ eine kompakte Menge und $f : \mathbb{R}^n \times U \to \mathbb{R}^n$ eine gegebene beschränkte, Lipschitz-stetige Abbildung. Wir bezeichnen $x = x(s)$ den *Zustand* und $\alpha = \alpha(s)$ die *Steuerung* oder *Kontrollvariable*. Mit

$$\mathcal{U} = \big\{ u : [0,T] \to U \,\big|\, u \text{ ist messbar} \big\} \tag{6.2}$$

führen wir die *Menge der zulässigen Steuerungen* ein. Wegen

$$\big|f(x,u)\big| \leq C \text{ und } \big|f(x,u) - f(y,u)\big| \leq C\,|x-y| \text{ für alle } x, y \in \mathbb{R}^n,\ u \in U \tag{6.3}$$

mit einer Konstanten $C > 0$ hat (6.1) für jede Steuerung $u \in \mathcal{U}$ genau eine Lösung $x = x(\cdot\,; u)$ auf $[t,T]$, die stetig differenzierbar ist. Manchmal wird x auch als *Response* des Systems (6.1) auf den *Input* u bezeichnet.

Ziel ist es, eine Steuerung $u \in \mathcal{U}$ zu finden, die (6.1) *optimal* beeinflusst. Dabei ist *optimal* bezüglich eines noch zu definierenden Zielfunktionals gemeint.

Gegeben seien $x_\circ \in \mathbb{R}^n$ und $0 \leq t \leq T$. Wir definieren

$$J(u; x_\circ, t) = \int_t^T h(x(s), u(s))\,\mathrm{d}s + g(x(T)), \tag{6.4}$$

wobei x das Problem (6.1) löst und h sowie g gegebene Funktionen sind mit den Eigenschaften

$$\begin{aligned} &|h(x,u)| \leq C, \quad |h(x,u) - h(y,u)| \leq C\,\|x-y\|_2, \\ &|g(x)| \leq C, \qquad |g(x) - g(y)| \leq C\,\|x-y\|_2 \end{aligned} \tag{6.5}$$

für alle $x, y \in \mathbb{R}^n$, $u \in U$. In (6.5) bezeichnet $C \geq 0$ eine Konstante. Ziel ist es, zu gegebener Anfangsbedingung $x_\circ$ zur Anfangszeit t eine optimale Steuerung u^* zu bestimmen, die das Zielfunktional J unter allen zulässigen Steurungen $u \in \mathcal{U}$ minimiert.

Um das Problem zu lösen, führen wir für $x_\circ \in \mathbb{R}^n$ und $t \in [0,T]$ die *minimale Wertefunktion*

$$V(x_\circ, t) = \inf_{u \in \mathcal{U}} J(u; x_\circ, t) \tag{6.6}$$

ein. Wir betrachten V als Abbildung in der Anfangsbedingung und der Anfangszeit. Auf diese Weise wird unser Problem, das Zielfunktional (6.4) unter der Nebenbedingung (6.1) für fixierte Werte von $x_\circ$ und t zu minimieren, in ein allgemeineres Problem überführt.

SATZ 6.1 (Optimalitätsbedingungen). *Seien $x_\circ \in \mathbb{R}^n$ und $t \in [0,T]$. Für jedes hinreichend kleine $\Delta t > 0$ mit $t + \Delta t \le T$ gilt dann*

$$V(x_\circ, t) = \inf_{u \in \mathcal{U}} \left\{ \int_t^{t+\Delta t} h(x(s), u(s))\,\mathrm{d}t + V(x(t+\Delta t, t+\Delta t) \right\}, \tag{6.7}$$

wobei $x = x(\cdot\,; u)$ das Problem (6.1) mit der Steuerung u löst.

PROOF. Der Beweis erfolgt in zwei Schritten. Zunächst wähle eine Steuerung $u_1 \in \mathcal{U}$ und löse

$$\begin{cases} \dot{x}_1(s) = f(x_1(s), u_1(s)) \quad \text{for } s \in (t, t+\Delta t], \\ x_1(t) = x_\circ. \end{cases} \tag{6.8}$$

Zu beliebig gewähltem $\varepsilon > 0$ wähle ein $u_2 \in \mathcal{U}$ mit

$$V(x_1(t+\Delta t), t+\Delta t) + \varepsilon \ge \int_{t+\Delta t}^{T} h(x_2(s), u_2(s))\,\mathrm{d}s + g(x_2(T)), \tag{6.9}$$

wobei x_2 das Anfangswertproblem

$$\begin{cases} \dot{x}_2(s) = f(x_2(s), u_2(s)) \quad \text{for } s \in (t+\Delta t, T], \\ x_2(t+\Delta t) = x_1(t+\Delta t) \end{cases} \tag{6.10}$$

löst. Nun definieren wir die Steuerung

$$u_3(s) = \begin{cases} u_1(s) & \text{für } t \le s \le t+\Delta t, \\ u_2(s) & \text{für } t+\Delta t < s < T \end{cases}$$

und bestimme die Lösung x_3 von

$$\begin{cases} \dot{x}_3(s) = f(x_3(s), u_3(s)) \quad \text{for } s \in (t, T], \\ x_3(t) = x_\circ \end{cases}$$

Wegen der eindeutigen Lösbarkeit von (6.1) gilt

$$x_3(s) = \begin{cases} x_1(s) & \text{für } t \le s \le t+\Delta t, \\ x_2(s) & \text{für } t+\Delta t < s < T \end{cases}$$

Aufgrund der Definition der minimalen Wertefunktion durch (6.6) und mit (6.9) folgt

$$\begin{aligned} V(x_\circ, t) &= \inf_{u \in \mathcal{U}} J(u; x_\circ, t) \\ &\le J(u_3; x_\circ, t) = \int_t^T h(x_3(s), u_3(s))\,\mathrm{d}s + g(x_3(T)) \\ &= \int_t^{t+\Delta t} h(x_1(s), u_1(s))\,\mathrm{d}s + \int_{t+\Delta t}^T h(x_2(s), u_2(s))\,\mathrm{d}s + g(x_2(T)) \\ &\le \int_t^{t+\Delta t} h(x_1(s), u_1(s))\,\mathrm{d}s + V(x_1(t+\Delta t), t+\Delta t) + \varepsilon. \end{aligned}$$

Da u_1 beliebig gewählt worden ist, erhalten wir somit

$$V(x_\circ, t) \le \inf_{u\in\mathcal{U}} \left\{ \int_t^{t+\Delta t} h(x(s), u(s))\,\mathrm{d}s + V(x(t+\Delta t), t+\Delta t) \right\} + \varepsilon, \tag{6.11}$$

wobei $x = x(\cdot\,; u)$ das Problem (6.1) löst.
Nun kommen wir zum zweiten Schritt. Wir fixieren wieder ein beliebiges $\varepsilon > 0$ und wählen nun ein $u_4 \in \mathcal{U}$ mit

$$\begin{aligned} V(x_\circ, t) + \varepsilon &\ge \int_t^T h(x_4(s), u_4(s))\,\mathrm{d}s + g(x_4(T)) \\ &= \int_t^{t+\Delta t} h(x_4(s), u_4(s))\,\mathrm{d}s + \int_{t+\Delta t}^T h(x_4(s), u_4(s))\,\mathrm{d}s + g(x_4(T)), \end{aligned} \tag{6.12}$$

wobei $x_4 = x_4(\cdot\,; u_4)$ das Problem (6.1) mit $u = u_4$ löst. Wegen (6.6) folgt

$$\begin{aligned} V(x_4(t+\Delta t), t+\Delta t) &= \inf_{u\in\mathcal{U}} J(u; x_4(t+\Delta t), t+\Delta t) \\ &\le \int_{t+\Delta t}^T h(x_4(s), u_4(s))\,\mathrm{d}s + g(x_4(T)). \end{aligned}$$

Also erhalten wir

$$V(x_\circ, t) + \varepsilon \ge \int_t^{t+\Delta t} h(x_4(s), u_4(s))\,\mathrm{d}s + V(x_4(t+\Delta t), t+\Delta t),$$

und daher

$$V(x_\circ, t) + \varepsilon \ge \inf_{u\in\mathcal{U}} \left\{ \int_t^{t+\Delta t} h(x(s), u(s))\,\mathrm{d}s + V(x(t+\Delta t), t+\Delta t) \right\}$$

mit der Lösung $x = x(\cdot\,; u)$ von (6.1). Diese Ungleichung zusammen mit (6.11) beweist die Behauptung. □

Ziel ist es nun, die Optimalitätsbedingung (6.7) in der Form einer eine partiellen Differentialgleichung zu schreiben. Dazu benötigen wir allerdings Abschätzungen für die minimale Wertefunktion V. Zum Beweis dieser Abschätzungen verwenden wir die *Gronwall-Ungleichung.*

LEMMA 6.2 (Gronwall-Ungleichung). 1) *Sei η eine nicht-negative, absolut-stetige Funktion auf $[0,T]$, die fast überall in $[0,T]$ die Ungleichung*

$$\eta'(t) \le \varphi(t)\eta(t) + \psi(t), \tag{6.13}$$

wobei φ und ψ nicht-negative, integrierbare Funktionen auf $[0,T]$ sind. Dann gilt

$$\eta(t) \le \exp\left(\int_0^t \varphi(s)\,\mathrm{d}s\right)\left(\eta(0) + \int_0^t \psi(s)\,\mathrm{d}s\right) \tag{6.14}$$

für alle $t \in [0,T]$.

2) *Insbesondere, wenn*

$$\eta'(t) \le \varphi(t)\eta(t) \text{ für fast alle } t \in [0,T] \quad \text{und} \quad \eta(0) = 0$$

gelten, dann folgt $\eta \equiv 0$ auf $[0,T]$.

PROOF. Aus (6.13) schließen wir

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}s}\Big(\eta(s)\exp\Big(-\int_0^s\varphi(r)\,\mathrm{d}r\Big)\Big) &= \exp\Big(-\int_0^s\varphi(r)\,\mathrm{d}r\Big)\big(\eta'(s)-\varphi(s)\eta(s)\big)\\ &\le \exp\Big(-\int_0^s\varphi(r)\,\mathrm{d}r\Big)\psi(s)\end{aligned}$$

für $s\in[0,T]$ fast überall. ferner gilt $\exp x\le 1$ für alle $x\in[0,\infty)$. Also erhalten wir nach Integration von 0 bis $t\in[0,T]$ die Beziehung

$$\begin{aligned}\eta(t)\exp\Big(-\int_0^t\varphi(s)\,\mathrm{d}s\Big) &\le \eta(0)+\int_0^t\exp\Big(-\int_0^s\varphi(r)\,\mathrm{d}r\Big)\psi(s)\,\mathrm{d}s\\ &\le \eta(0)+\int_0^t\psi(s)\,\mathrm{d}s,\end{aligned}$$

was die Ungleichung (6.14) beweist. □

Nun können wir folgendes Resultat beweisen.

LEMMA 6.3. *Es existiert eine Konstante $C>0$, so dass die Ungleichungen*

$$\big|V(x_\circ,t)\big|\le C,\quad \big|V(x_\circ,t)-V(\hat x_\circ,\hat t)\big|\le C\left(\|x_\circ-\hat x_\circ\|_2+|t-\hat t|\right)$$

für alle $x_\circ,\hat x_\circ\in\mathbb{R}^n$ und $t,\hat t\in[0,T]$ erfüllt ist.

PROOF. Offenbar impliziert (6.5), dass V auf $\mathbb{R}^n\times[0,T]$ beschränkt ist. Es ist also nur die Lispschitz-Stetigkeit zu zeigen. Wir fixieren dazu $x_\circ$, $\hat x_\circ\in\mathbb{R}^n$ und $0\le t\le T$. Zu beliebigem $\varepsilon>0$ wählen wir ein $\hat u\in\mathcal{U}$, so dass

$$V(\hat x_\circ,t)+\varepsilon\ge\int_t^T h(\hat x(s),\hat u(s))\,\mathrm{d}s+g(\hat x(T)),\tag{6.15}$$

wobei $\hat x$ das Problem

$$\begin{cases}\dot{\hat x}(s)=f(\hat x(s),\hat u(s)) & \text{for } s\in(t,T],\\ \hat x(t)=\hat x_\circ\end{cases}\tag{6.16}$$

Also ergibt sich wegen (6.15) die Abschätzung

$$\begin{aligned}&V(x_\circ,t)-V(\hat x_\circ,t)\\ &\quad\le\int_t^T h(x(s),\hat u(s))-h(\hat x(s),\hat u(s))\,\mathrm{d}s+g(x(T))-g(\hat x(T))+\varepsilon,\end{aligned}\tag{6.17}$$

wobei

$$\begin{cases}\dot x(s)=f(x(s),\hat u(s)) & \text{for } s\in(t,T],\\ x(t)=x_\circ\end{cases}\tag{6.18}$$

gilt. Aus (6.16) und (6.18) folgern wir

$$\dot x(s)-\dot{\hat x}(s)=f(x(s),\hat u(s))-f(\hat x(s),\hat u(s)).$$

Da die Abbildung f Lipschitz-stetig im ersten Argument ist (vergeiche (6.3)), bekommen wir die Ungleichung

$$\|f(x(s),\hat u(s))-f(\hat x(s),\hat u(s))\|_2\le C\|x(s)-\hat x(s)\|_2.$$

Damit folgern wir

$$\begin{aligned}\frac{1}{2}\frac{\mathrm{d}}{\mathrm{d}s}\|x(s)-\hat{x}(s)\|_2^2 &= \big(\dot{x}(s)-\dot{\hat{x}}(s)\big)^T\big(x(s)-\hat{x}(s)\big)\\ &= \big(f(x(s),\hat{u}(s))-f(\hat{x}(s),\hat{u}(s))\big)^T\big(x(s)-\hat{x}(s)\big)\\ &\le \|f(x(s),\hat{u}(s))-f(\hat{x}(s),\hat{u}(s))\|_2\|x(s)-\hat{x}(s)\|_2\\ &\le C\|x(s)-\hat{x}(s)\|_2^2.\end{aligned}$$

Nun wenden wir die Gronwall-Ungleichung (siehe Lemma 6.2) an mit $\eta = \|x(s)-\hat{x}(s)\|_2^2$, $\Phi(s)\equiv C$ und $\psi(s)\equiv 0$:

$$\begin{aligned}\|x(s)-\hat{x}(s)\|_2^2 &\le \exp\Big(\int_0^s 2C\,\mathrm{d}s\Big)\big(\|x(0)-\hat{x}(0)\|_2^2+0\big)\\ &\le \exp\big(2CT\big)\|x(0)-\hat{x}(0)\|_2^2.\end{aligned}$$

Setzen wir $C_1=\sqrt{\exp(2CT)}$, so gilt deshalb

$$\|x(s)-\hat{x}(s)\|_2 \le C_1\,\|x(0)-\hat{x}(0)\|_2 = C_1\,\|x_\circ-\hat{x}_\circ\|_2 \quad \text{für } s\in(t,T].$$

Damit folgt aus (6.5) und (6.17), dass

$$\begin{aligned}V(x_\circ,t)-V(\hat{x}_\circ,t) &\le \int_t^T C\,\|x(s)-\hat{x}(s)\|_2\,\mathrm{d}s + C\,\|x(T)-\hat{x}(T)\|_2+\varepsilon\\ &\le \int_t^T CC_1\,\|x_\circ-\hat{x}_\circ\|_2\,\mathrm{d}s + CC_1\,\|x_\circ-\hat{x}_\circ\|_2+\varepsilon\\ &\le C_2\,\|x_\circ-\hat{x}_\circ\|_2+\varepsilon\end{aligned}$$

mit $C_2 = CC_1(T+1)$. Vertauschen wir $x_\circ$ und $\hat{x}_\circ$, so ergibt sich

$$|V(x_\circ,t)-V(\hat{x}_\circ,t)| \le C_2\,\|x_\circ-\hat{x}_\circ\|_2 \quad \text{für } x_\circ,\hat{x}_\circ\in\mathbb{R}^n \text{ und } t\in[0,T],$$

das heisst, V ist Lipschitz-stetig im ersten Argument.
Nun seien $x_\circ\in\mathbb{R}^n$ sowie $0\le t<\hat{t}\le T$ beliebig gewählt. Zu $\varepsilon>0$ sei $u\in\mathcal{U}$ derart, dass

$$V(x_\circ,t)+\varepsilon \ge \int_t^T h(x(s),u(s))\,\mathrm{d}s + g(x(T)),$$

wobei x das Problem (6.1) löst. Wir definieren

$$\hat{u}(s) = u(s+t-\hat{t}) \quad \text{für } \hat{t}\le s\le T$$

und bezeichnen mit $\hat{x}$ die Lösung von

$$\begin{cases}\dot{\hat{x}}(s) = f(\hat{x}(s),\hat{u}(s)) & \text{für } s\in(\hat{t},T],\\ \hat{x}(\hat{t}) = x_\circ\end{cases}$$

Offenbar gilt $\hat{x}(s) = x(s+t-\hat{t})$, insbesondere $\hat{x}(\hat{t}) = x(t) = x_\circ$. Daher erhalten wir

$$\begin{aligned} V(x_\circ,\hat{t}) - V(x_\circ,t) &\le \int_{\hat{t}}^{T} h(\hat{x}(s),\hat{u}(s))\,\mathrm{d}s + g(\hat{x}(T)) \\ &\quad - \int_t^T h(x(s),u(s))\,\mathrm{d}s - g(x(T)) + \varepsilon \\ &= \int_{\hat{t}}^{T} h(x(s+t-\hat{t}),u(s+t-\hat{t}))\,\mathrm{d}s + g(x(T+t-\hat{t})) \\ &\quad - \int_t^T h(x(s),u(s))\,\mathrm{d}s - g(x(T)) + \varepsilon. \end{aligned}$$

Wir setzen $r = r(s) = s + t - \hat{t}$ für $s \in [\hat{t}, T]$. Dann $s = s(r) = r - t + \hat{t}$. Mit der Substituionsregel und (6.5) folgt

$$\begin{aligned} &V(x_\circ,\hat{t}) - V(x_\circ,t) \\ &\quad\le \int_t^{T+t-\hat{t}} h(x(r),u(r))\,\mathrm{d}r + g(x(T+t-\hat{t})) - \int_t^T h(x(s),u(s))\,\mathrm{d}s \\ &\qquad - g(x(T)) + \varepsilon \\ &\quad= C\|x(T+t-\hat{t}) - x(T)\|_2 - \int_{T+t-\hat{t}}^{T} h(x(r),u(r))\,\mathrm{d}r + \varepsilon \\ &\quad\le CC_3\,|t-\hat{t}| + \int_{T+t-\hat{t}}^{T} C\,\mathrm{d}r + \varepsilon = C(C_3+1)\,|t-\hat{t}| + \varepsilon \end{aligned} \tag{6.19}$$

für $x_\circ \in \mathbb{R}^n$ und $0 \le t < \hat{t} \le T$, wobei wir auch verwendet haben, dass die Lösung x Lipschitz-stetig ist, das heisst,

$$\|x(T+t-\hat{t}) - x(T)\|_2 \le C_3\,|t-\hat{t}|.$$

Nun wählen wir $\hat{u}$ so, dass

$$V(x_\circ,\hat{t}) + \varepsilon \ge \int_{\hat{t}}^{T} h(\hat{x}(s),\hat{u}(s))\,\mathrm{d}s + g(\hat{x}(T))$$

gilt, wobei $\hat{x}$ das Problem

$$\begin{cases} \dot{\hat{x}}(s) = f(\hat{x}(s),\hat{u}(s)) & \text{für } s \in (\hat{t},T], \\ \hat{x}(\hat{t}) = x_\circ \end{cases}$$

löst. Definiere

$$u(s) = \begin{cases} \hat{u}(s+\hat{t}-t) & \text{für } s \in [t, T+t-\hat{t}], \\ \hat{u}(T) & \text{für } s \in (T+t-\hat{t}, T] \end{cases}$$

und bezeichne mit x die Lösung von (6.1). Offenbar gelten die Beziehungen

$$u(s) = \hat{u}(s+\hat{t}-t) \text{ und } x(s) = \hat{x}(s+\hat{t}-t) \text{ für } s \in [t, T+\hat{t}-t].$$

Insbesondere haben wir $x(t) = \hat{x}(\hat{t}) = x_\circ$. Daher bekommen wir analog zur Herleitung der Ungleichung (6.19) die Beziehung

$$\begin{aligned}
V(x_\circ, t) - V(x_\circ, \hat{t}) &\le \int_t^T h(x(s), u(s))\,\mathrm{d}s + g(x(T)) \\
&\quad - \int_{\hat{t}}^T h(\hat{x}(s), \hat{u}(s))\,\mathrm{d}s - g(\hat{x}(T)) + \varepsilon \\
&= \int_t^T h(\hat{x}(s+\hat{t}-t), \hat{u}(s+\hat{t}-t))\,\mathrm{d}s + g(\hat{x}(T+\hat{t}-t)) \\
&\quad - \int_{\hat{t}}^T h(\hat{x}(s), \hat{u}(s))\,\mathrm{d}s - g(\hat{x}(T)) + \varepsilon \\
&\le \int_{\hat{t}}^{T+\hat{t}-t} h(\hat{x}(r), \hat{u}(r))\,\mathrm{d}r - \int_{\hat{t}}^T h(\hat{x}(s), \hat{u}(s))\,\mathrm{d}s \\
&\quad + C\,\|\hat{x}(T+\hat{t}-t) - \hat{x}(T)\|_2 + \varepsilon \\
&\le - \int_{T+\hat{t}-t}^T h(\hat{x}(s), \hat{u}(s))\,\mathrm{d}s + CC_2\,|\hat{t}-t| + \varepsilon \\
&\le C\big(1 + C_4\big)\,|\hat{t}-t| + \varepsilon,
\end{aligned}$$

wobei wir die Lipschitz-Stetigkeit

$$\|\hat{x}(T+\hat{t}-t) - \hat{x}(T)\|_2 \le C_4\,|\hat{t}-t|, \quad C_2 > 0,$$

verwendet haben. Diese Ungleichung zusammen mit (6.19) ergibt

$$|V(x_\circ, t) - V(x_\circ, \hat{t})| \le C_5\,|t-\hat{t}| \quad \text{für } x_\circ \in \mathbb{R}^n \text{ und } t, \hat{t} \in [0,T]$$

mit $C_5 = C(1 + \max(C_3, C_4))$. Die Dreiecksungleichung

$$\begin{aligned}
|V(x_\circ, t) - V(\hat{x}_\circ, \hat{t})| &\le |V(x_\circ, t) - V(\hat{x}_\circ, t)| + |V(\hat{x}_\circ, t) - V(\hat{x}_\circ, \hat{t})| \\
&\le C_6\left(\|x_\circ - \hat{x}_\circ\|_2 + |t-\hat{t}|\right)
\end{aligned}$$

mit $C_6 = \max(C_2, C_5)$ ergibt nun die Aussage des Lemmas. □

Mit Lemma 6.3 können wir zeigen, dass die minimale Wertefunktion einer partiellen Differentialgleichung genügt.

Satz 6.4 (HJB Gleichung). *Die minimale Wertefunktion V ist die eindeutige Viskositätslösung der Hamilton-Jacobi-Bellman Gleichung*

$$V_t(x,t) + \min_{u \in U}\left\{f(x,u)^T \nabla_x V(x,t) + h(x,t)\right\} = 0, \quad (x,t) \in \mathbb{R}^n \times (0,T), \tag{6.20a}$$

$$V(x,T) = g(x), \qquad x \in \mathbb{R}^n. \tag{6.20b}$$

Bemerkung 6.5. Wir führen für $(p,x) \in \mathbb{R}^n \times \mathbb{R}^n$ die Hamilton-Funktion

$$H(p,x) = \min_{u \in U}\left\{f(x,u)^T p + h(x,t)\right\}$$

ein. Dann lässt sich (6.20) in der kompakten Form

$$V_t(x,t) + H(\nabla_x V(x,t), x) = 0, \qquad (x,t) \in \mathbb{R}^n \times (0,T), \tag{6.21a}$$

$$V(x,T) = g(x), \qquad x \in \mathbb{R}^n \tag{6.21b}$$

schreiben. Aus der Theorie hyperbolischer Gleichungen ist bekannt, dass (6.21) keine glatten Lösungen für alle Zeiten $t \ge 0$ besitzen kann. Aus diesem Grund wird der Begriff *Viskositätslösung* eingeführt. ◊

Für ein $\varepsilon > 0$ betrachten wir zunächst das Problem

$$W_t^\varepsilon(x,t) - H(\nabla_x W^\varepsilon(x,t), x) = \varepsilon \Delta_x W^\varepsilon(x,t), \qquad (x,t) \in \mathbb{R}^n \times (0,T), \tag{6.22a}$$

$$W^\varepsilon(x,0) = g(x), \qquad x \in \mathbb{R}^n. \tag{6.22b}$$

Problem (6.22) ist ein quasilineares, parabolisches Problem, das (unter gewissen Voraussetzungen) eindeutig lösbar ist und glatte Lösungen besitzt. Der Term $\Delta_x W^\varepsilon$ regularisiert das hyperbolische Problem. Wenn W^ε für $\varepsilon \to 0$ gegen eine Funktion W (in einem gewissen Sinn) konvergiert, so interpretieren wir W als Lösung von

$$W_t(x,t) - H(\nabla_x W(x,t), x) = 0, \qquad (x,t) \in \mathbb{R}^n \times (0,T), \tag{6.23a}$$

$$W(x,0) = g(x), \qquad x \in \mathbb{R}^n. \tag{6.23b}$$

Bemerkung 6.6. Wenn V die Viskositätslösung von (6.21) ist, so ist $W(x,t) = V(x, T-t)$, $(x,t) \in \mathbb{R}^n \times [0,T]$, die Viskositätslösung von (6.23). ◊

Definition 6.7. *Eine beschränkte, gleichmäßig stetige Funktion V heißt* Viskositätslösung *von* (6.21)*, wenn folgende Eigenschaften erfüllt sind:*

1) $V(\cdot\,, T) = g$ *in* $\mathbb{R}^n$.
2) *Sei* $W \in C^\infty(\mathbb{R}^n \times (0,T))$ *beliebig. Hat* $V - W$ *ein lokales Maximum in* $(x_0, t_0) \in \mathbb{R}^n \times (0,T)$*, so gilt*
$$W_t(x_0,t_0) + H(\nabla_x W(x_0,t_0), x_0) \geq 0.$$
3) *Sei* $W \in C^\infty(\mathbb{R}^n \times (0,T))$*. Hat* $V - W$ *ein lokales Minimum in* (x_0,t_0)*, dann folgt*
$$W_t(x_0,t_0) + H(\nabla_x W(x_0,t_0), x_0) \leq 0.$$

Bemerkung 6.8. Um zu zeigen, dass die minimale Wertefunktion V eine Viskositätslösung von (6.21) ist, müssen wir die Eigenschaften 2) und 3) aus Definition 6.7 für beliebige Funktion $W \in C^\infty(\mathbb{R}^n \times (0,T))$ nachprüfen. ◊

Beweis von Satz 6.4. Nach Lemma 6.3 ist V beschränkt und Lipschitz-stetig. Ferner folgt aus (6.4) und (6.6) direkt

$$V(x_\circ, T) = \inf_{u \in \mathcal{U}} J(u; x_\circ, T) = g(x_\circ) \quad \text{für } (x_\circ, t) \in \mathbb{R}^n \times [0,T].$$

Damit gilt Teil 1) von Definition 6.7. Wir zeigen nun Teil 2). Sei $W \in C^\infty(\mathbb{R}^n \times [0,T])$ beliebig vorgegeben. Wir nehmen an, dass $V - W$ in (x_0,t_0) ein lokales Maximum hat. Zu zeigen ist, dass

$$W_t(x_0,t_0) + \min_{u \in U} \left\{ f(x_0,t_0)^T \nabla_x W(x_0,t_0) + h(x_0,u) \right\} \geq 0 \tag{6.24}$$

gilt. Angenommen, (6.24) ist nicht erfüllt. Dann existieren ein $u \in U$ und ein $\theta > 0$ mit

$$W_t(x,t) + f(x,t)^T \nabla_x W(x,t) + h(x,u) \leq -\theta < 0 \tag{6.25}$$

für alle (x,t) in einer hinreichend kleinen Umgebung von (x_0,t_0), das heisst, es gibt ein $\delta > 0$ mit

$$\|x - x_0\|_2 + |t - t_0| < \delta. \tag{6.26}$$

Da $V - W$ ein lokales Maximum besitzt, folgt — möglicherweise nach einer Reduktion von δ — die Ungleichung

$$(V-W)(x,t) \leq (V-W)(x_0,t_0) \quad \text{für alle } (x,t), \text{ die (6.26) erfüllen.} \tag{6.27}$$

Wir betrachten die konstante Steuerung $u(s) = u$ für alle $s \in [t_0, T]$ mit dem dazugehörigen Zustand

$$\begin{cases} \dot{x}(s) = f(x(s), u) & \text{für } s \in (t_0, T], \\ x(t_0) = x_0 \end{cases}$$

Wähle $\tau \in (0, \delta)$ so klein, dass $\|x(s) - x_0\|_2 + |s - t_0| < \delta$ für $s \in [t_0, t_0 + \tau]$. Dann folgt aus (6.25) und (6.25)

$$W_t(x(s), t) + f(x(s), t)^T \nabla_x W(x(s), t) + h(x(s), u) \leq -\theta, \quad s \in [t_0, t_0 + \tau]. \tag{6.28}$$

Unter Verwendung von (6.27) erhalten wir

$$\begin{aligned} V(x(t_0 + \tau), t_0 + \tau) - V(x_0, t_0) &\leq W(x(t_0 + \tau), t_0 + \tau) - W(x_0, t_0) \\ &= \int_{t_0}^{t_0+\tau} \frac{\mathrm{d}W}{\mathrm{d}s}(x(s), s)\,\mathrm{d}s = \int_{t_0}^{t_0+\tau} W_t(x(s), s) + \nabla_x W(x(s), s)\dot{x}(s)\,\mathrm{d}s \\ &= \int_{t_0}^{t_0+\tau} W_t(x(s), s) + f(x(s), u)^T \nabla_x W(x(s), s)\,\mathrm{d}s. \end{aligned} \tag{6.29}$$

Aufgrund der Optimalitätsbedingung (6.7) gilt

$$V(x_0, t_0) \leq \int_{t_0}^{t_0+\tau} h(x(s), u)\,\mathrm{d}s + V(x(t_0 + \tau), t_0 + \tau). \tag{6.30}$$

Kombinieren wir (6.28), (6.29) und (6.30), so erhalten wir

$$\begin{aligned} 0 &\leq \int_{t_0}^{t_0+\tau} W_t(x(s), s) + f(x(s), u)^T \nabla_x W(x(s), s)\,\mathrm{d}s \\ &\quad + V(x_0, t_0) - V(x(t_0 + \tau), t_0 + \tau) \\ &\leq \int_{t_0}^{t_0+\tau} W_t(x(s), s) + f(x(s), u)^T \nabla_x W(x(s), s) + h(x(s), u)\,\mathrm{d}s \\ &\leq -\int_{t_0}^{t_0+\tau} \theta\,\mathrm{d}s = -\theta\tau < 0. \end{aligned}$$

Dieser Widerspruch beweist (6.24).
Sei nun $W \in C^\infty(\mathbb{R}^n \times [0, T])$ wieder beliebig gewählt. Wir nehmen nun an, dass $V - W$ ein lokales Minimum im Punkt $(x_0, t_0) \in \mathbb{R}^n \times [0, T]$ hat. Dann müssen wir zeigen, dass

$$W_t(x_0, t_0) + \min_{u \in \mathcal{U}} \left\{ f(x_0, u)^T \nabla_x W(x_0, t_0) + h(x_0, u) \right\} \leq 0 \tag{6.31}$$

gilt. Angenommen, (6.31) ist nicht erfüllt. Dann existiert ein $\theta > 0$ mit

$$W_t(x, t) + f(x, u)^T \nabla_x W(x, t) + h(x, u) \geq \theta > 0 \quad \text{für alle } u \in U, \tag{6.32}$$

wobei (x, t) wieder hinreichend nahe an (x_0, t_0) liegen, das heisst, es gibt ein $\delta > 0$ mit

$$\|x - x_0\|_2 + |t - t_0| < \delta. \tag{6.33}$$

Da $V - W$ in (x_0, t_0) ein lokales Minimum besitzt, gilt — möglicherweise nach einer Reduktion von δ — die Ungleichung

$$(V - W)(x, t) \geq (V - W)(x_0, t_0) \quad \text{für alle } (x, t), \text{die (6.33) erfüllen.} \tag{6.34}$$

Sei $\tau \in (0, \delta)$ so gewählt, dass $\|x(s) - x_0\|_2 < \delta$ für $s \in [t_0, t_0 + \tau]$, wobei

$$\begin{cases} \dot{x}(s) = f(x(s), u(s)) \quad \text{für } s \in (t_0, T], \\ x(t_0) = x_0 \end{cases} \tag{6.35}$$

für eine Steuerung $u \in \mathcal{U}$ gelte. Diese Annahme können wir wegen (6.3) aufstellen. Wegen (6.34) und (6.35) haben wir für beliebiges $u \in \mathcal{U}$

$$\begin{aligned} V(x(t_0+\tau), t_0+\tau) - V(x_0, t_0) &\geq W(x(t_0+\tau), t_0+\tau) - W(x_0, t_0) \\ &= \int_{t_0}^{t_0+\tau} \frac{\mathrm{d}W}{\mathrm{d}s}\,\mathrm{d}s = \int_{t_0}^{t_0+\tau} W_t(x(s), s) + \nabla_x W(x(s), s)\dot{x}(s)\,\mathrm{d}s \\ &= \int_{t_0}^{t_0+\tau} W_t(x(s), s) + f(x(s), u(s))^T \nabla_x W(x(s), s)\,\mathrm{d}s. \end{aligned} \tag{6.36}$$

Andererseits können wir wegen der Optimalitätsbedingung (6.7) eine Steuerung $u \in \mathcal{U}$ wählen, so dass

$$V(x_0, t_0) \geq \int_{t_0}^{t_0+\tau} h(x(s), u(s))\,\mathrm{d}s + V(x(t_0+\tau), t_0+\tau) - \frac{\theta\tau}{2}$$

und somit

$$\frac{\theta\tau}{2} \geq V(x(t_0+\tau), t_0+\tau) - V(x_0, t_0) + \int_{t_0}^{t_0+\tau} h(x(s), u(s))\,\mathrm{d}s. \tag{6.37}$$

Kombination von (6.36) und (6.37) ergibt mit (6.32) die Beziehung

$$\frac{\theta\tau}{2} \geq \int_{t_0}^{t_0+\tau} W_t(x(s), s) + f(x(s), u(s))^T \nabla_x W(x(s), s) + h(x(s), u(s))\,\mathrm{d}s \geq \theta\tau.$$

Dieser Widerspruch beweist (6.31). □

Bemerkung 6.9. Wir kommen nun zur Wahl der optimalen Steuerungen. Gegeben seien eine Anfangszeit $t \in (0, T]$ und ein zugehöriger Anfangswert $x_\circ \in \mathbb{R}^n$. Wir betrachten das optimal gesteuerte System

$$\dot{x}^*(s) = f(x^*(s), u^*(s)) \quad \text{für alle } s \in (t, T], \tag{6.38a}$$

$$x^*(t) = x_\circ, \tag{6.38b}$$

wobei zu jeder Zeit s die Werte $u^*(s) \in U$ so gewählt sind, dass

$$\begin{aligned} f(x^*(s), u^*(s))^T \nabla_x V(x^*(s), s) + h(x^*(s), u^*(s)) \\ = H(\nabla_x V(x^*(s), s), x^*(s)) \\ = \min_{u \in U} \left\{ f(x^*(s), u)^T \nabla_x V(x^*(s), s) + h(x^*(s), u) \right\} \end{aligned} \tag{6.39}$$

erfüllt ist. Dann erhalten wir beim Lösen der HJB Gleichung neben der minimalen Wertefunktion für jedes $(x, t) \in \mathbb{R}^n \times (0, T)$ auch eine optimale Steuerung $u^* \in \mathcal{U}$, die sogenannte *Rückkopplungs-* oder *Feedback-Steuerung*. Natürlich müssen wir voraussetzen, dass (6.39) sinnvoll definiert ist, das heisst, u^* und V müssen hinreichend glatt sein. ◊

2. Linear-quadratische Steuerprobleme

In diesem Abschnitt beschäftigen wir uns mit linear-quadratischen Kontrollproblemen. Im Detail seien

$$\begin{aligned} f(x,u) &= Ax + Bu && \text{mit } A \in \mathbb{R}^{n\times n},\ B \in \mathbb{R}^{n\times m},\\ h(x,u) &= x^T Qx + u^T Ru && \text{mit } Q \in \mathbb{R}^{n\times n},\ R \in \mathbb{R}^{m\times m},\\ g(x) &= x^T Mx && \text{mit } M \in \mathbb{R}^{n\times n}, \end{aligned}$$

wobei Q sowie M positiv semidefinit und symmetrisch sowie R positiv definit und symmetrisch sind. Ferner seien die Anfangszeit $t = 0$ und $U = \mathbb{R}^m$. Also betrachten wir das Problem

$$\textbf{(LQR)} \qquad \begin{aligned} &\min J(u; x_\circ, 0) = \int_0^T x(t)^T Qx(t) + u(t)^T Ru(t)\,\mathrm{d}t + x(T)^T Mx(T)\\ &\text{u.d.N. } \dot{x}(t) = Ax(t) + Bu(t) \text{ für } t \in (0,T] \quad \text{und} \quad x(0) = x_\circ. \end{aligned}$$

Das Problem (**LQR**) wird *Linear-quadratisches Regulatorproblem* genannt. Der erste Schritt ist nun, dass wir bei der HJB Gleichung (6.20) die Minimierung betrachten. Hier müssen wir den Ausdruck

$$\varphi(u) = \big(Ax + Bu\big)^T \nabla_x V(x,t) + x^T Qx + u^T Ru$$

bezüglich $u \in \mathbb{R}^m$ minimieren. Daher betrachten wir die erste Ableitung von φ in eine beliebige Richtung $u_\delta \in \mathbb{R}^m$:

$$\nabla\varphi(u)u_\delta = \nabla_x V(x,t)^T Bu_\delta + u_\delta^T Ru + u^T Ru_\delta = \big(\nabla_x V(x,t)^T B + 2u^T R\big)\, u_\delta.$$

Aus $\nabla\varphi(u) = 0$ folgt daher $\nabla_x V(x,t)^T B + 2u^T R = 0$, das heisst, wir bekommen die Beziehung $2Ru = -B^T \nabla_x V(x,t)$. Wegen der Invertierbarkeit von R erhalten wir

$$u^* = -\frac{1}{2}\, R^{-1} B^T \nabla_x V(x,t). \tag{6.40}$$

Nun machen wir den folgenden Ansatz

$$V(x,t) = x^T P(t) x \quad \text{für } x \in \mathbb{R}^n, \tag{6.41}$$

wobei $P(t) \in \mathbb{R}^{n\times n}$ symmetrisch sei. Dann folgt $\nabla_x V(x,t) = 2P(t)x$. Einsetzen von (6.41) in (6.40) ergibt

$$u^*(t) = -K(t)x^*(t) \quad \text{für } t \in [0,T] \tag{6.42a}$$

mit

$$K(t) = R^{-1} B^T P(t) \in \mathbb{R}^{m\times n}. \tag{6.42b}$$

Einsetzen von (6.41) in die HJB Gleichung führt auf

$$x^T \dot{P}(t) x + \min_{u\in\mathbb{R}^m} \Big\{ \big(Ax + Bu\big)^T (2P(t)x) + x^T Qx + u^T Ru \Big\} = 0 \tag{6.43a}$$

für alle $(x,t) \in \mathbb{R}^n \times [0,T)$ und

$$P(T) = M \quad \text{in } \mathbb{R}^{n\times n}. \tag{6.43b}$$

Wir setzen (6.42a) in (6.43a) ein und erhalten

$$\begin{aligned} &x^T P(t)x + 2x^T A^T P(t)x + 2u^T B^T P(t)x + x^T Qx + u^T Ru \\ &\quad = x^T \left(\dot P(t) + A^T P(t) + P(t)A + Q\right) x - 2x^T P(t)BR^{-1}B^T P(t)x \\ &\qquad + x^T P(t)BR^{-1}RR^{-1}B^T P(t)x \\ &\quad = x^T \left(\dot P(t) + A^T P(t) + P(t)A + Q - x^T P(t)BR^{-1}B^T P(t)\right) x \end{aligned}$$

für alle $x \in \mathbb{R}^n$ und $t \in (0,T)$. Also folgt aus (6.43) ein Anfangswertproblem für $P(t)$ in $\mathbb{R}^{n\times n}$:

$$-\dot P(t) = A^T P(t) + P(t)A + Q - P(t)BR^{-1}B^T P(t) \qquad \text{für } t \in [0,T), \tag{6.44a}$$

$$P(T) = M. \tag{6.44b}$$

Zunächst lösen wir daher (6.44), die sogenannte (differentielle) *Matrix-Riccati Gleichung*, zur Bestimmung der Abbildung $t \mapsto P(t)\mathbb{R}^{n\times n}$. Dann berechnen wir die Feedback-Matrix $K = K(t) \in \mathbb{R}^{m\times n}$ gemäß (6.42b). Das Rückkopplungsgesetz ist dann durch (6.42a) gegeben. Dieses Feedback-Gesetz ist unabhängig von der Anfangsbedingung $x_\circ$ in (**LQR**). Die optimale Trajektorie $x^* = x^*(\cdot\,, u^*)$ erhalten wir durch das *Closed-Loop* System

$$\begin{aligned} \dot x^*(t) &= Ax^*(t) + Bu^*(t) = Ax^*(t) - BR^{-1}B^T P(t)x^*(t) \\ &= \left(A - BR^{-1}B^T P(t)\right) x^*(t) \qquad \text{für } t \in (0,T], \\ x^*(0) &= x_\circ. \end{aligned}$$

Bemerkung 6.10. Im Fall von $T = \infty$ (*Infinite Horizon Problem*) ist P unabhängig von t, das heisst, P ist konstant, und löst die *algebraische Matrix-Riccati Gleichung*

$$A^T P + PA + Q - PBR^{-1}B^T P = 0 \quad \text{in } \mathbb{R}^{n\times n}. \tag{6.45}$$

Es gibt in der System Control Toolbox von MATLAB Algorithmen zur Lösung von (6.45). ◊

Beispiel 6.11. Betrachte das Problem

$$\min J(u; x_\circ, 0) = \int_0^T x(t)^2 + u(t)^2 \,\mathrm{d}t$$

unter der Nebenbedingung

$$\dot x(t) = u(t) \text{ für } t \in (0,T] \quad \text{und} \quad x(0) = x_\circ.$$

Wir haben also $n = m = 1$, $A = M = 0$, $B = Q = R = 1$. Daher folgt aus (6.44)

$$-\dot P(t) = 1 - P(t)^2 \text{ für } t \in [0,T) \quad \text{und} \quad P(T) = 0. \tag{6.46}$$

Das Anfangswertproblem (6.46) lässt sich durch Separation der Variablen lösen. Seien $t = T - \tau$ und $p(\tau) = P(t)$. Dann erhalten wir $\dot P(t) = -\dot p(\tau)$ und

$$\dot p(\tau) = 1 - p(\tau)^2 \text{ für } \tau \in (0,T], \quad p(0) = 0.$$

Also mit $r(p) = 1 - p^2$ und $h(\tau) = 1$ gilt

$$\frac{\mathrm{d}p(\tau)}{\mathrm{d}\tau} = r(p(\tau))h(\tau) \quad \Longleftrightarrow \quad \frac{\mathrm{d}p(\tau)}{1 - p(\tau)^2} = \frac{\mathrm{d}\tau}{1} \tag{6.47}$$

Wegen

$$\frac{1}{1-p^2} = \frac{a}{1-p} + \frac{b}{1+p} \;\Leftrightarrow\; a(1+p) + b(1-p) = 1 \;\Leftrightarrow\; p(a-b) + a + b = 1$$

folgen $a = b = 1/2$. Damit gilt

$$\begin{aligned}\int_0^{p(\tau)} \frac{1}{1-p^2}\,\mathrm{d}p &= \int_0^{p(\tau)} \frac{1}{2(1-p)} + \frac{1}{2(1+p)}\,\mathrm{d}t \\ &= \frac{1}{2}\Big(\int_0^{p(\tau)} \frac{1}{(1-p)}\,\mathrm{d}t + \int_0^{p(\tau)} \frac{1}{(1+p)}\,\mathrm{d}t\Big) \\ &= \frac{1}{2}\big(\ln|p(\tau)+1| + |p(\tau)-1|\big).\end{aligned}$$

Also mit (6.47) schliessen wir

$$\frac{1}{2}\ln\left|\frac{p(\tau)+1}{p(\tau)-1}\right| = \tau \quad\Leftrightarrow\quad \left|\frac{p(\tau)+1}{p(\tau)-1}\right| = \exp(2\tau)$$

Die Bedingung $p(0) = 0$ führt auf $p(\tau) - 1 < 0$ für $\tau \in [0, T]$ hinreichend klein. Somit bekommen wir

$$\begin{aligned}\frac{p(\tau)+1}{1-p(\tau)} = \exp(2\tau) \;\Leftrightarrow\quad & 1 + p(\tau) = (1-p(\tau))\exp(2\tau) \\ \Leftrightarrow\quad & \big(1+\exp(2\tau)\big)p(\tau) = \exp(2\tau) - 1 \\ \Leftrightarrow\quad & \big(1+\exp(-2\tau)\big)p(\tau) = 1 - \exp(-2\tau) \\ \Leftrightarrow\quad & p(\tau) = \frac{1-\exp(-2\tau)}{1+\exp(-2\tau)}.\end{aligned}$$

Mit der Variablentransformation $\tau = T - t$ erhalten wir die Lösung von (6.46):

$$P(t) = \frac{1-\exp(2(t-T))}{1+\exp(2(t-T))}.$$

Ferner ist die optimale Steuerung durch

$$u^*(t) = -K(t)x^*(t) = -R^{-1}B^TP(t)x^*(t) = -P(t)x^*(t) = \frac{\exp(2(t-T))-1}{1+\exp(2(t-T))}x^*(t)$$

gegeben. ◊

Literaturverzeichnis

[1] D. P. Bertsekas. *Nonlinear Programming*. Athena Scientific, Belmont, Mass., 1995.

[2] J. F. Blowey, A. W. Craig, and T. Shardlow. *Frontiers in Numerical Analysis*. Springer-Verlag,, Berlin, 2003.

[3] J. F. Bonnans, J. C. Gilbert, C. Lemaréchal, and C. A. Segastizábal. *Numerical Optimization. Theoretical and Practical Aspects*. Springer-Verlag, Berlin, 2003.

[4] R. K. Brayton, S. W. Director, G. D. Hachtel, and L. Vidigal. A new algorithm for statistical circuit design based on quasi-Newton methods and function splitting. *IEEE Transactions on Circuits and Systems*, CAS-26:784-794, 1979

[5] I. Das. *Nonlinear Multicriteria Optimization and Robust Optimality*. Ph.D. Thesis, Rice University, Houston, Texas, 1997

[6] I. Das und J. Dennis. A closer look at drawbacks of minimizing weigthed sums of objectives for Pareto set generation in multicrieria optimization. S. 1–12, 1996

[7] J. E. Dennis and R. Schnabel. *Numerical Methods for Unconstrained Optimization and Nonlinear Equations*. SIAM, Philadelphia, 1996.

[8] P. Dorato, C. Abdallah, and V. Cerone. *Linear-Quadratic Control*. Prentice Hall, Englewood Cliffs, New Yersey 07632, 1995.

[9] L.C. Evans. *Partial Differential Equations*. Graduate Studies in Mathematics, vol. 19, American Mathematical Society, Providence, Rhode Island, 2002.

[10] P. J. Fleming. Application of Multiobjective Optimization to Compensator Design for SISO Control Systems. *Electronic Letters*, 22:258-259, 1986

[11] P. J. Fleming. Computer-Aided Control System Design of Regulators using a Multiobjective Optimization Approach. *Proc. IFAC Control Applications of Nonlinear Prog. and Optim.*, Capri, Italy, pp. 47-52, 1985

[12] C. Geiger and C. Kanzow. *Numerische Verfahren zur Lösung unrestringierter Optimierungsaufgaben*. Springer-Verlag, Berlin, 1999.

[13] C. Geiger and C. Kanzow. *Theorie und Numerik restringierter Optimierungsaufgaben*. Springer-Verlag, Berlin, 2002.

[14] F. R. Gembick. *Vector Optimization for Control with Performance and Parameter Sensitivity Indices*. Ph.D. Thesis, Case Western Reserve Univ., Cleveland, Ohio, 1974

[15] A. Göpfert und R. Nehse. *Vektoroptimierung*. Teubner Verlag, Leipzig, 1990

[16] Y. Haines. Integrated system identification and optimization. *Control and Dynamic Systems: Advances in Theory and Applications*, 10:435–518, 1973

[17] C. Hillermeier. *Nonlinear Multiobjective Optimization: A Generalized Homotopy Approach*. Birkhäuser Verlag, Basel, 2001

[18] J. Jahn. *Introduction to the Theory of Nonlinear Optimization*. Springer-verlag, Berlin, 1999

[19] F. Jarre und J. Stoer. *Optimierung*. Springer-Verlag, Berlin, 2004.

[20] W. Karush. Minima of functions of several variables with inequalities as side constraints. Master's dissertation, University of Chicago, 1939

[21] C. T. Kelley. *Iterative Methods for Linear and Nonlinear Equations*. Frontiers in Applied Mathematics. SIAM, Philadelphia, 1995.

[22] C. T. Kelley. *Iterative Methods for Optimization*. Frontiers in Applied Mathematics. SIAM, Philadelphia, 1999.

[23] H. Kuhn und A. Tucker. Nonlinear Programming. In J. Neyman, Editor, *Proceedings of the Second Berkeley Symposium on Mathematical Statistics and Probability*, University of California Press, Berkeley, pp. 481-492, 1951

[24] F.-S. Kupfer. An infinite-dimensional convergence theory for reduced SQP methods in Hilbert spaces. *SIAM Journal on Optimization*, 6:126-163, 1996.

[25] J. Liu. Multiple objective problems: Pareto optimal solutions by methods of properly equality constraints. *IEEE Transactions on Automatic Control*, 21:641–650, 1976

[26] D. G. Luenberger. *Linear and Nonlinear Progamming*. Addison-Wesley Publishing Company, Reading, Massachusetts, 1984.

[27] S. Marglin. *Public Investment Crietria*. MIT Press, Cambridge, Massachusetts, 1967

[28] K. Mietinen. *Nonlinear Multiobjective Opimization*. Kluwer Academic Publishers, 1999

[29] J. Nocedal and S. J. Wright. *Numerical Optimization*. 2. Auflage, Springer Series in Operation Research. Springer-Verlag, New York, 2006.

[30] B. Rustem. *Algorithms for Nonlinear Programming and Multiple Objective Decision*. Wiley, Chichester, 1998

[31] Y. Sawaragi, H. Nakagama and T. Tanino. *Theory of Multiobjective Optimization*. Academic Press, Orlando, Florida, USA, 1985

[32] S. Schäffler. *Global Optimization using Stochastic Integration*. S. Roderer Verlag, Regensburg, 1995

[33] W. Stadler (Editor). *Multicriteria Optimization in Engineering and in the Sciences*. Plenum Press, New York, 1988

[34] R. E. Steuer. *Multiple Criteria Optimization: Theory, Computations and Applications*. John Wiley & Sons, New York, 1986

[35] S. Volkwein. *Some remarks on augmented Lagrange-Newton SQP methods*. SFB-Preprint No. 256, 2003. Siehe http://www.uni-graz.at/imawww/volkwein/publist.html