# OPTIMIERUNG 1

Vorlesungsskript, Sommersemester 2022

Christian Clason

Stand vom 3. Oktober 2022

Institut für Mathematik und wissenschaftliches Rechnen
Universität Graz

# INHALTSVERZEICHNIS

## III LINEARE OPTIMIERUNG

# ÜBERBLICK

Die mathematische Optimierung beschäftigt sich mit der Aufgabe, Minima bzw. Maxima von Funktionen zu bestimmen. Konkret seien eine Menge $X$, eine (nicht notwendigerweise echte) Teilmenge $U \subset X$ und eine Funktion $f : X \to \mathbb{R}$ gegeben. Gesucht ist ein $\bar{x} \in U$ mit

$$f(\bar{x}) \leq f(x) \quad \text{für alle } x \in U,$$

geschrieben

$$f(\bar{x}) = \min_{x \in U} f(x).$$

Die Fragen, die wir uns dabei stellen müssen, sind:

1. Hat dieses Problem eine Lösung?
2. Gibt es eine intrinsische Charakterisierung von $\bar{x}$, d. h. ohne Vergleich mit allen anderen $x \in U$?
3. Wie kann dieses $\bar{x}$ (effizient) berechnet werden?

Aus der Vielzahl der möglichen Beispiele sollen nur kurz folgende erwähnt werden:

(i) In *Transport- und Produktionsproblemen* sollen Kosten für Transport minimiert bzw. Gewinn aus Produktion maximiert werden. Dabei beschreibt $x \in \mathbb{R}^n$ die Menge der zu transportierenden bzw. produzierenden verschiedenen Güter und $f(x)$ die dafür nötigen Kosten bzw. aus dem Verkauf erzielten Gewinne. Die Nebenbedingung $x \in U$ beschreibt dabei, dass ein Mindestbedarf gedeckt werden muss bzw. nur endlich viele Rohstoffe zur Produktion zur Verfügung stehen.

(ii) In *inversen Problemen* sucht man einen Parameter $u$ (zum Beispiel Röntgenabsorption von Gewebe in der Computertomographie), hat aber nur eine (gestörte) Messung $y^\delta$ zur Verfügung. Ist ein Modell bekannt, das für gegebenen Parameter $u$ die entsprechende Messung $y = Ku$ liefert, so kann man den unbekannten Parameter näherungsweise rekonstruieren, indem man das Problem

$$\min_{u \in U} \|Ku - y^\delta\|^2 + \alpha \|u\|^2$$

für geeignet gewählte Normen und $\alpha > 0$ löst. Die Menge $U$ kann dabei bekannte Einschränkungen an den Parameter (z. B. Positivität) beschreiben.

(iii) In der *optimalen Steuerung* ist man zum Beispiel daran interessiert, ein Auto oder eine Raumsonde möglichst effizient von einem Punkt $x_0$ zu einem anderen Punkt $x_1$ zu steuern. Beschreibt $x(t) \in \mathbb{R}^3$ die Position zum Zeitpunkt $t \in [0, T]$, so gehorcht $x(t)$ der Differenzialgleichung

$$\begin{cases} x'(t) = f(t, x(t), u(t)), \\ x(0) = x_0, \end{cases} \tag{1}$$

wobei $u(t)$ die Rolle der Steuerung spielt. Will man dabei den Treibstoffverbrauch (der proportional zu $|u(t)|$ ist) minimieren, führt das auf das Problem

$$\min_{(x,u)\in U} \int_0^T |u(t)|\, dt \quad \text{mit} \quad U = \{(x, u) : (1) \text{ ist erfüllt und } x(T) = x_1\}\,.$$

In dieser Vorlesung behandeln wir zuerst *nichtlineare Optimierungsprobleme ohne Restriktionen*, in denen $f : \mathbb{R}^n \to \mathbb{R}$ differenzierbar und $U = \mathbb{R}^n$ ist. In diesem Fall ist die Antwort auf die Frage nach der Existenz relativ einfach zu beantworten; uns werden daher vor allem die beiden restlichen Fragen beschäftigen. Dabei wird das Konzept der *Abstiegsrichtung* in beiden Fällen fundamental sein: Grob gesprochen befinden wir uns in einem Minimum, falls keine Abstiegsrichtung existiert; ansonsten wählen wir eine und folgen ihr einen Schritt weit.

Im Anschluß behandeln wir *lineare Optimierungsprobleme*, in denen $f : \mathbb{R}^n \to \mathbb{R}$ linear und $U$ ein Polyeder ist, d. h. durch endlich viele lineare Ungleichungen beschrieben werden kann. Allgemein wird solch ein Problem beschrieben durch einen Vektor $c \in \mathbb{R}^n$, eine Matrix $A \in \mathbb{R}^{m\times n}$ und einen Vektor $b \in \mathbb{R}^m$. Wir suchen dann $x \in \mathbb{R}^n$ als Lösung von

$$\begin{cases} \min\limits_{x\in\mathbb{R}^n} c^T x \\ \text{mit } Ax \le b. \end{cases} \tag{LP}$$

Solche Probleme tauchen sehr häufig in Wirtschaft und Finanzen auf (siehe Beispiel 1); aber auch viele nichtlineare Probleme können auf lineare Probleme zurückgeführt werden. (Vergleiche die zentrale Rolle von linearen Gleichungssystemem in der numerischen Mathematik.)

Dieses Skriptum basiert vor allem auf den folgenden Werken:

[i] W. Alt (2011), *Nichtlineare Optimierung. Eine Einführung in Theorie, Verfahren und Anwendungen*, 2. Aufl., Vieweg+Teubner, Wiesbaden

[ii] P. Gritzmann (2014), *Grundlagen der Mathematischen Optimierung*, Springer, Berlin, doi: 10.1007/978-3-8348-2011-2

[iii] M. Grötschel (2010), Lineare und Ganzzahlige Programmierung (ADM II), Vorlesungsskript, Institut für Mathematik, Technische Universität Berlin, URL: http://www3.math.tu-berlin.de/Vorlesungen/WS09/LinOpt/index.de.html

[iv] J. Dennis & R. Schnabel (1996), *Numerical Methods for Unconstrained Optimization and Nonlinear Equations*, Bd. 16, Classics in Applied Mathematics, Society for Industrial & Applied Mathematics, DOI: 10.1137/1.9781611971200

[v] C. Geiger & C. Kanzow (1999), *Numerische Verfahren zur Lösung unrestringierter Optimierungsaufgaben*, Springer, Berlin, DOI: 10.1007/978-3-642-58582-1

[vi] C. Geiger & C. Kanzow (2002a), *Theorie und Numerik restringierter Optimierungsaufgaben*, Springer, Berlin, DOI: 10.1007/978-3-642-56004-0

[vii] C. T. Kelley (1999), *Iterative Methods for Optimization*, Bd. 18, Frontiers in Applied Mathematics, Society for Industrial & Applied Mathematics (SIAM), Philadelphia, PA, DOI: 10.1137/1.9781611970920

[viii] R. Schultz (2013), Optimierung 1, Vorlesungsskript, Fakultät für Mathematik, Universität Duisburg-Essen

[ix] M. Ulbrich & S. Ulbrich (2012), *Nichtlineare Optimierung*, Birkhäuser, Basel, DOI: 10.1007/978-3-0346-0654-7

# Teil I

# GRUNDLAGEN

# 1 GRUNDLAGEN DER LINEAREN ALGEBRA UND ANALYSIS

In diesem Kapitel stellen wir zunächst die wesentlichen Begriffe und Resultate aus der linearen Algebra und der Analysis im $\mathbb{R}^n$ zusammen, die in dieser Vorlesung benötigt werden.

## VEKTOREN, NORMEN, MATRIZEN

Vektoren im $\mathbb{R}^n$ sind stets Spaltenvektoren; der zu $x \in \mathbb{R}^n$ zugehörige Zeilenvektor wird mit $x^T$ bezeichnet. Für die Komponenten eines Vektors (bezüglich der Standardbasis aus Einheitsvektoren, die wir mit $e_i$ bezeichnen) verwenden wir Indizes:

$$x = (x_1, \ldots, x_n)^T \in \mathbb{R}^n.$$

Das Produkt aus Zeilen- und Spaltenvektoren ist das Skalarprodukt im $\mathbb{R}^n$: Für $x, y \in \mathbb{R}^n$ ist $x^T y := \sum_{i=1}^n x_i y_i$. Für die Norm verwenden wir zumeist die Euklidische Vektornorm:

$$\|x\| := \|x\|_2 := \sqrt{x^T x} = \left( \sum_{i=1}^n x_i^2 \right)^{1/2}.$$

Diese wird vom Skalarprodukt induziert und gehorcht daher der Cauchy–Schwarz-Ungleichung: Für alle $x, y \in \mathbb{R}^n$ ist $x^T y \leq \|x\| \|y\|$. In endlichdimensionalen Vektorräumen sind alle Normen äquivalent: Ist $\|\cdot\|_*$ eine weitere Norm, so existieren Konstanten $c_1, c_2 > 0$ mit

$$c_1 \|x\| \leq \|x\|_* \leq c_2 \|x\| \qquad \text{für alle } x \in \mathbb{R}^n.$$

Die durch die Norm beschriebene offene Kugel um $x$ mit Radius $\varepsilon$ bezeichnen wir mit

$$B_\varepsilon(x) := \{ y \in \mathbb{R}^n : \|x - y\| < \varepsilon \}.$$

Durch die Norm wird ein Konvergenzbegriff vermittelt: Eine Folge $\{x^k\}_{k \in \mathbb{N}} \subset \mathbb{R}^n$ konvergiert gegen $x \in \mathbb{R}^n$, geschrieben $x^k \to x$, falls gilt $\|x^k - x\| \to 0$ (im Sinne der Konvergenz reeller Zahlenfolgen). Da $\mathbb{R}^n$ endlichdimensional ist, ist das dann und nur dann der Fall, wenn $x_i^k \to x_i$ für alle $1 \leq i \leq n$ gilt.

Für Matrizen $A = (a_{ij})_{i,j} \in \mathbb{R}^{m\times n}$ verwenden wir die induzierte *Spektralnorm*

$$\|A\| = \max_{\|x\|=1} \|Ax\|,$$

für die $\|Ax\| \le \|A\|\|x\|$ für alle $x \in \mathbb{R}^n$ gilt.

Eine Matrix ist *symmetrisch*, wenn $A^T := (a_{ji})_{i,j} = A$ (und damit zwingend $m = n$) gilt; *positiv definit*, wenn gilt

$$x^T Ax > 0 \qquad \text{für alle } x \in \mathbb{R}^n \setminus \{0\};$$

und *gleichmäßig positiv definit*, wenn ein $\mu > 0$ existiert mit

$$x^T Ax \ge \mu\|x\|^2 \qquad \text{für alle } x \in \mathbb{R}^n.$$

Eine *symmetrische* Matrix ist positiv definit genau dann, wenn alle ihre Eigenwerte $\lambda_1 \le \cdots \le \lambda_n$ (strikt) positiv sind; in diesem Fall gilt nach dem Satz von Courant–Fischer[1]

$$\lambda_1 x^T x \le x^T Ax \le \lambda_n x^T x \qquad \text{für alle } x \in \mathbb{R}^n. \tag{1.1}$$

Eine symmetrische Matrix ist also positiv definit genau dann, wenn sie gleichmäßig positiv definit ist! Schließlich halten wir noch fest, dass eine symmetrische und positiv definite Matrix stets invertierbar ist, und dass für die Spektralnorm gilt

$$\|A\| = \lambda_n, \qquad \|A^{-1}\| = \lambda_1^{-1}.$$

### ABLEITUNGEN IM $\mathbb{R}^n$

Eine Funktion $F : \mathbb{R}^n \to \mathbb{R}^m$, $x \mapsto (F_1(x), \dots, F_m(x))^T$, heißt *stetig* in $x$, wenn $F(x^k) \to F(x)$ für alle konvergenten Folgen $\{x^k\}_{k\in\mathbb{N}} \subset \mathbb{R}^n$ gilt. Dies ist genau dann der Fall, wenn alle Komponenten $F_i : \mathbb{R}^n \to \mathbb{R}, 1 \le i \le m$, stetig sind. Die Funktion $F$ heißt *Lipschitz-stetig*, wenn es eine *Lipschitz-Konstante* $L > 0$ gibt mit

$$\|F(x^1) - F(x^2)\| \le L\|x^1 - x^2\| \qquad \text{für alle } x^1, x^2 \in \mathbb{R}^n.$$

(Stetig aber nicht Lipschitz-stetig ist zum Beispiel $f : \mathbb{R} \to \mathbb{R}$, $x \mapsto \sqrt{|x|}$.) Gilt dies nur für alle $x^1, x^2 \in B_\varepsilon(x)$ für ein $x \in \mathbb{R}^n$ und $\varepsilon > 0$, so heißt $F$ *lokal Lipschitz-stetig* in $x$.

Eine Funktion $F : \mathbb{R}^n \to \mathbb{R}^m$ heißt (Fréchet-)*differenzierbar* in $x$, wenn es eine lineare Abbildung $F'(x) : \mathbb{R}^n \to \mathbb{R}^m$ gibt mit

$$\lim_{\|h\|\to 0} \frac{\|F(x+h) - F(x) - F'(x)h\|}{\|h\|} = 0.$$

[1] siehe z. B. [Hanke-Bourgeois 2009, Satz 23.4]

Die Matrixdarstellung der linearen Abbildung $F'(x)$ (bezüglich der kanonischen Basis) nennt man die *Jacobi-Matrix* von $F$ (in $x$); wir werden diese nicht in der Notation unterscheiden. Die Einträge der Jacobi-Matrix bestehen aus den *partiellen Ableitungen*, d. h.

$$F'(x) = \left(\frac{\partial F_i(x)}{\partial x_j}\right)_{i,j} \in \mathbb{R}^{m\times n}, \qquad \frac{\partial F_i(x)}{\partial x_j} := \lim_{t\to 0} \frac{F_i(x+te_j) - F_i(x)}{t}.$$

Tatsächlich werden wir öfter Ist die Abbildung $x \mapsto \nabla F(x)$ stetig, so heißt $F$ *stetig differenzierbar*.

Speziell für den Fall $m = 1$ und $f : \mathbb{R}^n \to \mathbb{R}$ werden wir öfter mit der *transponierten* Jacobi-Matrix arbeiten, weshalb wir ihr einen eigenen Namen geben: den *Gradient*

$$\nabla f(x) = \left(\frac{\partial f(x)}{\partial x_1}, \ldots, \frac{\partial f(x)}{\partial x_n}\right)^T \in \mathbb{R}^n.$$

(Der Gradient ist also im Gegensatz zur Ableitung $f'(x)$ ebenfalls ein Spaltenvektor!) Für einen gegebenen Vektor $d \in \mathbb{R}^n$ (mit $\|d\| = 1$) gibt $\nabla f(x)^T d$ die Steigung von $f$ in $x$ in Richtung $d$ an; es handelt sich also um eine *Richtungsableitung*, die wir (mit Hilfe der Kettenregel) auch berechnen können über

$$\nabla f(x)^T d = \lim_{t\to 0} \frac{f(x+td) - f(x)}{t}.$$

Sind alle partiellen Ableitungen stetig differenzierbar, so heißt $f$ zweimal stetig differenzierbar; in diesem Fall bezeichnet

$$\nabla^2 f(x) := \left(\frac{\partial^2 f(x)}{\partial x_i \partial x_j}\right)_{i,j} \in \mathbb{R}^{n\times n}$$

die *Hesse-Matrix* von $f$. Diese ist wegen der Stetigkeit der zweiten Ableitungen nach dem Satz von Schwarz symmetrisch.

Ein Kern-Resultat in der mehrdimensionalen Analysis ist die *Taylor-Entwicklung*. Wir werden ihn in Gestalt der folgenden Spezialfälle benötigen, die man als *Mittelwertsätze* bezeichnen kann.

Satz 1.1 (Mittelwertsatz I). *Seien $f : \mathbb{R}^n \to \mathbb{R}$ stetig differenzierbar und $x, y \in \mathbb{R}^n$ gegeben. Dann existiert ein $\xi := y + \theta(x-y)$ mit $\theta \in (0,1)$ und*

$$f(x) = f(y) + \nabla f(\xi)^T (x-y).$$

Satz 1.2 (Mittelwertsatz II). *Seien $f : \mathbb{R}^n \to \mathbb{R}$ zweimal stetig differenzierbar und $x, y \in \mathbb{R}^n$ gegeben. Dann existiert ein $\xi := y + \theta(x-y)$ mit $\theta \in (0,1)$ und*

$$f(x) = f(y) + \nabla f(y)^T (x-y) + \frac{1}{2}(x-y)^T \nabla^2 f(\xi)(x-y).$$

Für vektorwertige Funktionen $F : \mathbb{R}^n \to \mathbb{R}^m$ gelten diese Mittelwertsätze nicht direkt (die Schwierigkeit ist, für alle Kompenenten ein einheitliches $\theta$ zu finden). Es gilt aber der folgende Mittelwertsatz in Integralform (den wir zumeist auf $F(x) = \nabla f(x)$ anwenden werden).

Satz 1.3 (Mittelwertsatz III). *Seien $F : \mathbb{R}^n \to \mathbb{R}^m$ stetig differenzierbar und $x, y \in \mathbb{R}^n$ gegeben. Dann gilt*

$$F(x) = F(y) + \int_0^1 \nabla F(y + \theta(x - y))^T (x - y)\, d\theta.$$

### KONVEXE FUNKTIONEN

Von besonderer Bedeutung in der Optimierung sind konvexe Mengen und Funktionen. Zur Erinnerung: Eine Menge $X \subset \mathbb{R}^n$ heißt *konvex*, wenn für alle $x, y, \in X$ und alle $\lambda \in (0, 1)$ gilt $\lambda x + (1 - \lambda) y \in X$. Anschaulich bedeutet dies, dass für je zwei Punkte in $X$ auch ihre Verbindungsstrecke in $X$ liegt.

Sei nun $X \subset \mathbb{R}^n$ konvex. Dann heißt eine Funktion $f : X \to \mathbb{R}$

(i) *konvex* (auf $X$), wenn für alle $x, y \in X$ und alle $\lambda \in [0, 1]$ gilt

$$f(\lambda x + (1 - \lambda) y) \leq \lambda f(x) + (1 - \lambda) f(y).$$

(ii) *strikt konvex* (auf $X$), wenn für alle $x, y \in X$ mit $x \neq y$ und alle $\lambda \in (0, 1)$ gilt

$$f(\lambda x + (1 - \lambda) y) < \lambda f(x) + (1 - \lambda) f(y).$$

(iii) *gleichmäßig konvex* (auf $X$), wenn es ein *Konvexitätsmodul* $\mu > 0$ gibt, so dass für alle $x, y \in X$ und alle $\lambda \in [0, 1]$ gilt

$$f(\lambda x + (1 - \lambda) y) + \mu\lambda(1 - \lambda)\|x - y\|^2 \leq \lambda f(x) + (1 - \lambda) f(y).$$

Anschaulich bedeutet dies, dass für eine konvexe Funktion kein Punkt einer Verbindungsstrecke von zwei Punkten auf dem Graphen der Funktion unterhalb des Graphen liegt; für strikt konvexe Funktionen darf die Strecke darüber hinaus nicht mit dem Graphen zusammenfallen. Für eine gleichmäßig konvexe Funktion muss zwischen Verbindungsstrecke und Graph sogar noch eine Parabel passen. Offensichtlich ist jede gleichmäßig konvexe Funktion strikt konvex und jede strikt konvexe Funktion konvex; die Umkehrung gilt allerdings nicht. Zum Beispiel ist

(i) $f : \mathbb{R} \to \mathbb{R}$, $x \mapsto x$, konvex aber nicht strikt konvex,

(ii) $f : \mathbb{R} \to \mathbb{R}$, $x \mapsto e^x$, strikt konvex aber nicht gleichmäßig konvex,

(iii) $f : \mathbb{R} \to \mathbb{R}$, $x \mapsto x^2$, gleichmäßig konvex,

(iv) $f : \mathbb{R} \to \mathbb{R}$, $x \mapsto x^4$, strikt konvex aber nicht gleichmäßig konvex.

Für stetig differenzierbare Funktionen lässt sich Konvexität über den Gradienten charakterisieren.

Satz 1.4. *Sei $X \subset \mathbb{R}^n$ offen und konvex und sei $f : X \to \mathbb{R}$ stetig differenzierbar. Dann ist*

*(i) $f$ genau dann konvex, wenn für alle $x, y \in X$ gilt*

$$\nabla f(y)^T (x - y) \leq f(x) - f(y),$$

*(ii) $f$ genau dann strikt konvex, wenn für alle $x, y \in X$ mit $x \neq y$ gilt*

$$\nabla f(y)^T (x - y) < f(x) - f(y),$$

*(iii) $f$ genau dann gleichmäßig konvex, wenn ein $\mu > 0$ existiert so dass für alle $x, y \in X$ gilt*

$$\nabla f(y)^T (x - y) + \mu \|x - y\|^2 \leq f(x) - f(y).$$

*Beweis.* Zu (i): Sei $f$ konvex. Dann gilt für alle $x, y \in X$ und $\lambda \in (0, 1)$ nach Definition

$$\frac{f(y + \lambda(x - y)) - f(y)}{\lambda} \leq \frac{\lambda f(x) + (1 - \lambda) f(y) - f(y)}{\lambda} = f(x) - f(y).$$

Grenzübergang $\lambda \to 0$ ergibt dann

$$\nabla f(y)^T (x - y) = \lim_{\lambda \to 0} \frac{f(y + \lambda(x - y)) - f(y)}{\lambda} \leq f(x) - f(y).$$

Gilt umgekehrt diese Ungleichung, so folgt daraus mit $x_\lambda := \lambda x + (1 - \lambda) y$ und der produktiven Null

$$\begin{aligned} \lambda f(x) + (1 - \lambda) f(y) - f(x_\lambda) &= \lambda (f(x) - f(x_\lambda)) + (1 - \lambda)(f(y) - f(x_\lambda)) \\ &\geq \lambda \nabla f(x_\lambda)^T (x - x_\lambda) + (1 - \lambda) \nabla f(x_\lambda)^T (y - x_\lambda) \\ &= \nabla f(x_\lambda)^T (\lambda x + (1 - \lambda) y - x_\lambda) = 0. \end{aligned}$$

Also ist $f$ konvex.

Zu (ii): Sei $f$ strikt konvex. Da beim Grenzübergang die strikte Ungleichung nicht erhalten bleibt, müssen wir anders als für (i) vorgehen. Für $x, y \in X$ mit $x \neq y$ setze $z := \frac{1}{2}(x + y) \in X$ (denn $X$ ist konvex). Da strikt konvexe Funktionen insbesondere konvex sind, können wir (i) verwenden und erhalten

$$\nabla f(y)^T (x - y) = 2 \nabla f(y)^T (z - y) \leq 2(f(z) - f(y)).$$

Aus der strikten Konvexität folgt weiterhin $f(z) < \frac{1}{2}(f(x) + f(y))$. Zusammen ergibt das

$$\nabla f(y)^T(x-y) < f(x) + f(y) - 2f(y) = f(x) - f(y).$$

Die umgekehrte Richtung geht dagegen genau wie für (i), nur mit strikter Ungleichung.

Zu (iii): Sei $f$ gleichmäßig konvex. Wie für (i) verwenden wir die Charakterisierung der Richtungsableitung und die Definition der gleichmäßigen Konvexität und schätzen ab

$$\begin{aligned}\nabla f(y)^T(x-y) &= \lim_{\lambda\to 0} \frac{f(y+\lambda(x-y)) - f(y)}{\lambda} \\ &\le \lim_{\lambda\to 0} \frac{\lambda f(x) + (1-\lambda)f(y) - \mu\lambda(1-\lambda)\|x-y\|^2 - f(y)}{\lambda} \\ &= f(x) - f(y) - \mu\|x-y\|^2.\end{aligned}$$

Gilt umgekehrt diese Ungleichung, so gehen wir analog zu (i) vor mit $x_\lambda := \lambda x + (1-\lambda)y \in X$ (denn $X$ ist konvex) und schätzen ab

$$\begin{aligned}\lambda f(x) + (1-\lambda)f(y) - f(x_\lambda) &= \lambda(f(x) - f(x_\lambda)) + (1-\lambda)(f(y) - f(x_\lambda)) \\ &\ge \lambda\left(\nabla f(x_\lambda)^T(x-x_\lambda) + \mu\|x-x_\lambda\|^2\right) \\ &\quad + (1-\lambda)\left(\nabla f(x_\lambda)^T(y-x_\lambda) + \mu\|y-x_\lambda\|^2\right) \\ &= \nabla f(x_\lambda)^T(\lambda x + (1-\lambda)y - x_\lambda) \\ &\quad + \mu\left(\lambda\|x-x_\lambda\|^2 + (1-\lambda)\|y-x_\lambda\|^2\right).\end{aligned}$$

Der erste Term auf der rechten Seite verschwindet wieder; für den zweiten verwenden wir

$$\|x-x_\lambda\| = (1-\lambda)\|x-y\|, \qquad \|y-x_\lambda\| = \lambda\|x-y\|,$$

und erhalten

$$\begin{aligned}\lambda f(x) + (1-\lambda)f(y) - f(x_\lambda) &\ge \mu\left(\lambda\|x-x_\lambda\|^2 + (1-\lambda)\|y-x_\lambda\|^2\right) \\ &= \mu(\lambda(1-\lambda)^2 + \lambda^2(1-\lambda))\|x-y\|^2 \\ &= \mu\lambda(1-\lambda)\|x-y\|^2,\end{aligned}$$

was zu zeigen war. □

Ist $f$ sogar zweimal stetig differenzierbar, lässt sich die Konvexität auch über die Hesse-Matrix charakterisieren.

Satz 1.5. *Sei $X \subset \mathbb{R}^n$ offen und konvex und sei $f : X \to \mathbb{R}$ zweimal stetig differenzierbar. Dann ist*

(i) *$f$ genau dann konvex, wenn $\nabla^2 f(x)$ für alle $x \in X$* positiv semidefinit *ist, d. h. wenn gilt*

$$d^T \nabla^2 f(x) d \geq 0 \qquad \text{für alle } x \in X, d \in \mathbb{R}^n,$$

(ii) *$f$ strikt konvex, wenn $\nabla^2 f(x)$ für alle $x \in X$ positiv definit ist, d. h. wenn gilt*

$$d^T \nabla^2 f(x) d > 0 \qquad \text{für alle } x \in X, d \in \mathbb{R}^n \setminus \{0\},$$

(iii) *$f$ genau dann gleichmäßig konvex, wenn $\nabla^2 f(x)$ für alle $x \in X$ gleichmäßig positiv definit ist, d. h. wenn ein $\mu > 0$ existiert mit*

$$d^T \nabla^2 f(x) d \geq \mu \|d\|^2 \qquad \text{für alle } x \in X, d \in \mathbb{R}^n,$$

*Beweis.* Zu (i): Sei $f$ konvex und seien $x \in X$ und $d \in \mathbb{R}^n$ beliebig. Da $X$ offen ist, existiert ein $\tau > 0$ mit $x + td \in X$ für alle $t \in (0, \tau)$. Aus Satz 1.4 (i) erhalten wir nun zusammen mit Satz 1.2 für $y := x + td$

$$0 \leq f(x + td) - f(x) - t\nabla f(x)^T d = \frac{t^2}{2} d^T \nabla^2 f(\xi_t) d$$

mit $\xi_t = (x + td) + \theta_t(x - (x + td)) = x + t(1 - \theta_t)d$ für ein $\theta_t \in (0, 1)$. Da $(1 - \theta_t) \in (0, 1)$ gleichmäßig beschränkt ist, konvergiert für $t \to 0$ also $\xi_t \to x$ und damit, da $f$ zweimal stetig differenzierbar ist, auch $\nabla^2 f(\xi_t) \to \nabla^2 f(x)$. Division durch $\frac{t^2}{2} > 0$ und Grenzübergang $t \to 0$ ergibt daher die Aussage. Umgekehrt folgt aus Satz 1.2 zusammen mit der positiven Semidefinitheit

$$f(x) - f(y) = \nabla f(y)^T (x - y) + \frac{1}{2}(x - y)^T \nabla^2 f(\xi_t)(x - y) \geq \nabla f(y)^T (x - y)$$

und daher mit Satz 1.4 (i) die Konvexität.

Analog zeigt man für (ii) die strikte Konvexität mit Satz 1.4 (ii) und strikter Ungleichung. (Für die andere Richtung geht das nicht, da die strikte Ungleichung beim Grenzübergang nicht erhalten bleibt.)

Zu (iii): Sei $f$ gleichmäßig konvex mit Konvexitätsmodul $\tilde{\mu} > 0$. Genau wie für (i) erhält man dann aus Satz 1.4 (iii)

$$0 \leq f(x + td) - f(x) - t\nabla f(x)^T d - \tilde{\mu}\|td\|^2 = \frac{t^2}{2} d^T \nabla^2 f(\xi_t) d - t^2 \tilde{\mu} \|d\|^2,$$

und Division durch $\frac{t^2}{2} > 0$ und Grenzübergang $t \to 0$ ergibt wieder die Aussage (mit $\mu := 2\tilde{\mu}$). Umgekehrt folgt aus Satz 1.2 zusammen mit der gleichmäßigen Definitheit

$$\begin{aligned} f(x) - f(y) &= \nabla f(y)^T (x - y) + \frac{1}{2}(x - y)^T \nabla^2 f(\xi_t)(x - y) \\ &\geq \nabla f(y)^T (x - y) + \frac{\mu}{2}\|x - y\|^2, \end{aligned}$$

und daher mit Satz 1.4 (iii) die gleichmäßige Konvexität mit Modul $\tilde{\mu} := \frac{\mu}{2}$. □

Im Fall $n = 1$ entsprechen diese Resultate der bekannten Tatsache, dass eine Funktion genau dann (strikt) konvex ist, wenn ihre erste Ableitung (strikt) monoton wachsend bzw. ihre zweite Ableitung (strikt) positiv ist. Beachte, dass die Bedingung in (ii) nur *hinreichend* ist, was man sich am Beispiel $f : \mathbb{R} \to \mathbb{R}$, $x \mapsto x^4$, leicht überlegen kann.

# 2 GRUNDLEGENDE BEGRIFFE UND EXISTENZ

Wir beginnen mit einigen elementaren Definitionen. Sei im folgenden stets $X \subset \mathbb{R}^n$ eine (nicht notwendigerweise echte) Teilmenge und $f : X \to \mathbb{R}$. Gesucht ist ein $\bar{x} \in X$ mit

$$f(\bar{x}) \leq f(x) \quad \text{für alle } x \in X,$$

geschrieben

$$f(\bar{x}) = \min_{x \in X} f(x).$$

Man nennt $X$ *zulässige Menge* und einen Punkt $x \in X$ *zulässigen Punkt*; die Forderung $\bar{x} \in X$ wird *Nebenbedingung* genannt. Ist $X = \mathbb{R}^n$, so spricht man auch von *unrestringierter Optimierung* (d. h. *ohne Nebenbedingungen*), ansonsten von *restringierter Optimierung* (d. h. *mit Nebenbedingungen*). Oft wird $f$ als *Zielfunktion* bezeichnet. Der optimale Wert $f(\bar{x})$ wird als *Minimum* bezeichnet, $\bar{x}$ selber als *Minimierer*, geschrieben $\bar{x} = \arg\min_{x \in X} f(x)$. Analog spricht man von *Maximum* und *Maximierer*, wenn $f(\bar{x}) \geq f(x)$ für alle $x \in X$ ist. Da gilt

$$\max_{x \in X} f(x) = -\min_{x \in X} -f(x),$$

werden wir in der Regel ohne Beschränkung der Allgemeinheit Minimierer suchen, können aber, wenn es bequemer ist, auch das äquivalente Maximierungsproblem betrachten.

Wir unterscheiden weiter: Die Funktion $f$ hat in $\bar{x} \in X$

(i) ein *globales Minimum*, falls gilt $\bar{x} \in X$ und

$$f(\bar{x}) \leq f(x) \qquad \text{für alle } x \in X,$$

(ii) ein *striktes globales Minimum*, falls gilt $\bar{x} \in X$ und

$$f(\bar{x}) < f(x) \qquad \text{für alle } x \in X \setminus \{\bar{x}\},$$

(iii) ein *lokales Minimum*, falls $\bar{x} \in X$ gilt und ein $\varepsilon > 0$ existiert mit

$$f(\bar{x}) \leq f(x) \qquad \text{für alle } x \in X \cap B_\varepsilon(\bar{x}),$$

(iv) ein *striktes lokales Minimum*, falls $\bar{x} \in X$ gilt und ein $\varepsilon > 0$ existiert mit

$$f(\bar{x}) < f(x) \qquad \text{für alle } x \in (X \cap B_\varepsilon(\bar{x})) \setminus \{\bar{x}\}.$$

Entsprechend spricht man von (strikten) lokalen oder globalen Minimierern. Offensichtlich ist jedes (strikte) globale Minimum auch ein (striktes) lokales Minimum, jedoch nicht umgekehrt. Dabei sind strikte globale Minima eindeutig, während strikte lokale Minima lediglich isoliert sein müssen.

Wir werden sehen, dass wir nur lokale Minima mit vertretbarem Aufwand finden können. Eine Ausnahme bilden konvexe Funktionen, was die Bedeutung dieser Funktionenklasse in der Optimierung unterstreicht.

Satz 2.1. *Sei $X \subset \mathbb{R}^n$ eine konvexe Menge und sei $f : X \to \mathbb{R}$ eine konvexe Funktion. Dann gilt:*

*(i) Jedes lokale Minimum von $f$ ist auch ein globales Minimum.*

*(ii) Ist $f$ strikt konvex, so besitzt $f$ höchstens ein lokales Minimum, das dann sogar ein striktes globales Minimum ist.*

*Beweis.* Zu (i): Angenommen, $f$ hätte in $\bar{x} \in X$ kein globales Minimum. Dann existiert ein $x \in X$ mit $f(x) < f(\bar{x})$. Für alle $t \in (0, 1]$ ist dann wegen der Konvexität von $X$ auch $\bar{x} + t(x - \bar{x}) \in X$, und aus der Konvexität von $f$ folgt

$$f(\bar{x} + t(x - \bar{x})) \leq tf(x) + (1 - t)f(\bar{x}) < tf(\bar{x}) + (1 - t)f(\bar{x}) = f(\bar{x}).$$

Also existiert für jedes $\varepsilon > 0$ ein $t \in (0, 1]$ mit $x_t := \bar{x} + t(x - \bar{x}) \in B_\varepsilon(\bar{x})$ und $f(x_t) < f(\bar{x})$, d. h. $f$ hat in $\bar{x}$ auch kein lokales Minimum.

Zu (ii): Sei nun $f$ strikt konvex. Angenommen, es gäbe zwei verschiedene lokale Minimierer $\bar{x}, \bar{y} \in X$. Dann sind nach (i) sowohl $\bar{x}$ als auch $\bar{y}$ globale Minimierer, d. h. es muss $f(\bar{x}) = f(\bar{y})$ gelten. Setze nun $z := \frac{1}{2}(\bar{x} + \bar{y}) \in X$. Aus der strikten Konvexität von $f$ und $\bar{x} \neq \bar{y}$ folgt dann aber

$$f(z) < \frac{1}{2}\left(f(\bar{x}) + f(\bar{y})\right) = f(\bar{x})$$

im Widerspruch dazu, dass $\bar{x}$ ein globaler Minimierer ist. Die Funktion $f$ kann daher höchstens ein lokales (und damit auch globales) Minimum haben, das damit eindeutig sein muss. □

Wir betrachten nun die Frage der Existenz von Minimierern, die wir mit Hilfe des folgenden recht allgemeinen Satzes beantworten.

Satz 2.2. *Sei $X \subset \mathbb{R}^n$ nichtleer und abgeschlossen und $f : X \to \mathbb{R}$ stetig. Gilt*

*(i) $X$ ist beschränkt* oder

*(ii) $f$ ist* koerziv *auf $X$, d. h. für jede Folge $\{x^k\}_{k\in\mathbb{N}} \subset X$ mit $\|x^k\| \to \infty$ gilt $f(x^k) \to \infty$,*

*so besitzt $f$ einen globalen Minimierer $\bar{x} \in X$.*

*Ist $X$ konvex und $f$ strikt konvex, so ist der Minimierer eindeutig.*

*Beweis.* Gilt (i), so ist $X$ nach dem Satz von Heine–Borel kompakt, und nach dem Satz von Weierstrass nimmt daher die stetige Funktion $f$ auf $X$ ihr Minimum an.

Gilt dagegen (ii), gehen wir anders vor. Da die Menge der Funktionswerte $\{f(x) : x \in X\}$ reell und nach Voraussetzung nichtleer ist, existiert ein Infimum $M := \inf_{x\in X} f(x) \in \mathbb{R} \cup \{-\infty\}$ (letzteren Fall werden wir später ausschließen). Aus den Eigenschaften des Infimums folgt dann, dass eine Folge $\{y^k\}_{k\in\mathbb{N}} \subset \{f(x) : x \in X\} \subset \mathbb{R}$ existiert mit $y^k \to M$, d. h. es existiert eine Folge $\{x^k\}_{k\in\mathbb{N}} \subset X$ mit

$$f(x^k) \to M = \inf_{x\in X} f(x) < \infty.$$

Aus der Koerzivität von $f$ folgt dann, dass diese Folge beschränkt ist (sonst müsste ja $f(x^k) \to \infty$ gelten). Aus dem Satz von Bolzano–Weierstrass folgt dann die Existenz einer konvergenten Teilfolge $\{x^{k_m}\}_{m\in\mathbb{N}} \subset X$, für deren Grenzwert $\bar{x} \in X$ (denn $X$ ist abgeschlossen) wegen der Stetigkeit von $f$ gilt

$$-\infty < f(\bar{x}) = \lim_{m\to\infty} f(x^{k_m}) = M = \inf_{x\in X} f(x).$$

Das Infimum ist also endlich und wird in $\bar{x} \in X$ angenommen; es ist daher ein (globales) Minimum.

Die Eindeutigkeit für $X$ konvex und $f$ strikt konvex folgt nun sofort aus Satz 2.1. □

Ab nun werden wir stillschweigend voraussetzen, dass ein Minimierer existiert, und uns auf die Frage nach der Charakterisierung und Berechnung konzentrieren.

# Teil II

# OPTIMIERUNG OHNE NEBENBEDINGUNGEN

# 3 OPTIMALITÄTSBEDINGUNGEN

Wir betrachten in Folge unrestringierte Optimierungsprobleme, d. h. für $X = \mathbb{R}^n$, und leiten zunächst notwendige und hinreichende Bedingungen dafür her, dass ein Punkt $\bar{x} \in \mathbb{R}^n$ ein Minimierer ist. Die fundamentale Einsicht ist dabei, dass wir uns in einem Minimum befinden, wenn der Funktionswert bei Bewegung in jeder Richtung zunehmen würde, d. h. wenn die Steigung in jeder Richtung positiv ist.

Satz 3.1. *Sei $f : \mathbb{R}^n \to \mathbb{R}$ differenzierbar auf der offenen Menge $U \subset \mathbb{R}^n$ und sei $\bar{x} \in U$ ein lokaler Minimierer von $f$. Dann gilt*

$$\nabla f(\bar{x})^T d \geq 0 \qquad \text{für alle } d \in \mathbb{R}^n. \tag{3.1}$$

*Beweis.* Angenommen, es gibt eine Richtung $d \in \mathbb{R}^n$ mit

$$0 > \nabla f(\bar{x})^T d = \lim_{t \to 0} \frac{f(\bar{x} + td) - f(\bar{x})}{t}. \tag{3.2}$$

Da der Grenzwert strikt negativ ist, muss der Differenzenquotient auf der rechten Seite für $t$ klein genug auch strikt negativ sein. Es gibt also ein $\tau > 0$ mit $\bar{x} + td \in U$ und

$$\frac{f(\bar{x} + td) - f(\bar{x})}{t} < 0 \qquad \text{für alle } t \in (0, \tau],$$

d. h.

$$f(\bar{x} + td) < f(\bar{x}) \qquad \text{für alle } t \in (0, \tau].$$

Für alle $\varepsilon > 0$ ist aber $\bar{x} + td \in B_\varepsilon(\bar{x})$ für $t$ klein genug, und daher kann $\bar{x}$ kein lokaler Minimierer sein. □

Gilt (3.1), so folgt durch Einsetzen von $-d \in \mathbb{R}^n$ sofort $\nabla f(\bar{x})^T d = 0$ für alle $d \in \mathbb{R}^n$, was nur für $\nabla f(\bar{x}) = 0$ möglich ist. Wir erhalten also die folgende Optimalitätsbedingung.

Satz 3.2 (notwendige Optimalitätsbedingung 1. Ordnung). *Sei $f : \mathbb{R}^n \to \mathbb{R}$ differenzierbar auf der offenen Menge $U \subset \mathbb{R}^n$ und sei $\bar{x} \in U$ ein lokaler Minimierer von $f$. Dann gilt*

$$\nabla f(\bar{x}) = 0. \tag{3.3}$$

Ein Punkt $\bar{x}$, der (3.3) erfüllt, heißt *stationärer Punkt*. Man spricht von einer *Bedingung 1. Ordnung*, da sie nur erste Ableitungen verwendet; die Bedingung ist lediglich notwendig, da auch Maximierer oder Sattelpunkte stationäre Punkte sind. Um diese auszuschließen, muss man zweite Ableitungen zu Rate ziehen.

Satz 3.3 (notwendige Optimalitätsbedingung 2. Ordnung). *Sei $f : \mathbb{R}^n \to \mathbb{R}$ zweimal stetig differenzierbar auf der offenen Menge $U \subset \mathbb{R}^n$ und sei $\bar{x} \in U$ ein lokaler Minimierer von $f$. Dann ist $\nabla^2 f(\bar{x})$ positiv semidefinit.*

*Beweis.* Angenommen, $\nabla^2 f(\bar{x})$ ist nicht positiv semidefinit, d. h. es gibt eine Richtung $d \in \mathbb{R}^n$ mit

$$d^T \nabla^2 f(\bar{x}) d < 0.$$

Sei nun $t > 0$ klein genug, dass $\bar{x} + td \in U$ gilt. Aus Satz 1.2 folgt dann zusammen mit Satz 3.2

$$f(\bar{x} + td) = f(\bar{x}) + \frac{t^2}{2} d^T \nabla^2 f(\xi_t)^T d$$

für ein $\xi_t = \bar{x} + \theta_t td$ mit $\theta_t \in (0, 1)$. Da $\nabla^2 f$ nach Voraussetzung stetig auf $U$ ist, ist auch $d^T \nabla^2 f(\xi_t)^T d < 0$ für alle $t \in (0, \tau]$ für ein $\tau > 0$ klein genug. Also gilt

$$f(\bar{x} + td) = f(\bar{x}) + \frac{t^2}{2} d^T \nabla^2 f(\xi_t)^T d < f(\bar{x}) \qquad \text{für alle } t \in (0, \tau],$$

und wieder kann $\bar{x}$ daher kein lokaler Minimierer sein. □

Auch diese Bedingung ist nur notwendig, da sie auch in Sattelpunkten erfüllt sein kann (betrachte $f : \mathbb{R} \to \mathbb{R}$, $x \mapsto x^3$). Um diese auszuschließen, müssen wir die Bedingung verschärfen.

Satz 3.4 (hinreichende Optimalitätsbedingung 2. Ordnung). *Sei $f : \mathbb{R}^n \to \mathbb{R}$ zweimal stetig differenzierbar auf der offenen Menge $U \subset \mathbb{R}^n$ und sei $\bar{x} \in U$ mit*

*(i)* $\nabla f(\bar{x}) = 0$ *und*

*(ii)* $\nabla^2 f(\bar{x})$ *gleichmäßig positiv definit.*

*Dann hat $f$ in $\bar{x}$ ein striktes lokales Minimum.*

*Beweis.* Betrachte wieder $\bar{x} + td \in U$ für $d \in \mathbb{R}^n$ und $t > 0$ klein genug. Aus Satz 1.2 folgt dann zusammen mit (i)

$$f(\bar{x} + td) = f(\bar{x}) + \frac{t^2}{2} d^T \nabla^2 f(\xi_t)^T d$$

für ein $\xi_t = \bar{x} + \theta_t t d$ mit $\theta_t \in (0,1)$. Wegen (ii) existiert weiter ein $\mu > 0$ mit

$$d^T \nabla^2 f(\bar{x})^T d \geq \mu \|d\|^2.$$

Zusammen mit der produktiven Null erhalten wir daraus die Abschätzung

$$\begin{aligned} f(\bar{x} + td) &= f(\bar{x}) + \frac{t^2}{2} d^T \nabla^2 f(\bar{x})^T d + \frac{t^2}{2} d^T (\nabla^2 f(\xi_t) - \nabla^2 f(\bar{x}))^T d \\ &\geq f(\bar{x}) + \frac{t^2}{2} \left(\mu - \|\nabla^2 f(\xi_t) - \nabla^2 f(\bar{x})\|\right) \|d\|^2, \end{aligned}$$

wobei wir im letzten Schritt die "umgekehrte" Cauchy–Schwarz-Ungleichung $x^T y \geq -\|x\|\|y\|$ zusammen mit der Definition der Matrixnorm verwendet haben.

Aus der Stetigkeit von $\nabla^2 f$ folgt nun, dass ein $\tau > 0$ existiert mit $\|\nabla^2 f(\xi_t) - \nabla^2 f(\bar{x})\| < \mu$ für alle $t \in (0, \tau]$. Also gilt

$$f(\bar{x} + td) > f(\bar{x}) \qquad \text{für alle } d \in \mathbb{R}^n \setminus \{0\}, t \in (0, \tau],$$

d. h. für alle $x := \bar{x} + td \in B_\tau(\bar{x})$. Damit hat $f$ in $\bar{x}$ nach Definition ein striktes lokales Minimum. □

Dabei ist (ii) (nur) im Endlichdimensionalen genau dann erfüllt, wenn $\nabla^2 f(\bar{x})$ positiv definit ist. Diese Bedingung ist umgekehrt nur hinreichend, aber nicht notwendig, wie das Beispiel $f : \mathbb{R} \to \mathbb{R}, x \mapsto x^4$, zeigt.

Beachte, dass Ableitungen immer nur lokale Informationen liefern und alle diese Bedingungen daher nur *lokale* Minimierer charakterisieren; ähnliche Bedingungen sind für *globale* Minimierer in der Regel nicht möglich! Eine Ausnahme bilden (mal wieder) konvexe Funktionen.

Satz 3.5 (notwendige und hinreichende Bedingung für konvexe Funktionen). *Sei $f : \mathbb{R}^n \to \mathbb{R}$ konvex und auf der offenen Menge $U \subset \mathbb{R}^n$ differenzierbar. Dann hat $f$ in $\bar{x} \in U$ ein globales Minimum genau dann, wenn $\nabla f(\bar{x}) = 0$ ist.*

*Beweis.* Dass die Bedingung notwendig ist, folgt aus Satz 3.2. Sei nun $\bar{x} \in U$ ein stationärer Punkt. Aus Satz 1.4 (i) folgt dann

$$f(x) - f(\bar{x}) \geq \nabla f(\bar{x})^T (x - \bar{x}) = 0 \qquad \text{für alle } x \in \mathbb{R}^n,$$

d. h. $\bar{x}$ ist ein globaler Minimierer. □

# 4 ABSTIEGSVERFAHREN

Satz 3.1 legt folgendes iterative Verfahren zur Bestimmung eines Minimierers nahe:

**Algorithmus 4.1 :** Allgemeines Abstiegsverfahren

1 Wähle einen *Startpunkt* $x^0 \in \mathbb{R}^n$, setze $k = 0$
2 **while** $\nabla f(x^k) \neq 0$ **do**
3 Wähle eine *Suchrichtung* $s^k \in \mathbb{R}^n$ mit $\nabla f(x^k)^T s^k < 0$
4 Wähle eine *Schrittweite* $\sigma_k > 0$ mit $f(x^k + \sigma_k s^k) < f(x^k)$
5 Setze $x^{k+1} = x^k + \sigma_k s^k$, $k \leftarrow k+1$

Der Beweis von Satz 3.1 zeigt, dass wir stets eine Suchrichtung und eine Schrittweite mit den gewünschten Eigenschaften finden können, solange $x^k$ kein stationärer Punkt ist. Das einzige, was noch schief gehen kann, ist, dass wir vor Erreichen eines stationären Punkts „verhungern“, d. h. dass entweder $s^k$ oder $\sigma_k$ zu schnell zu klein werden. Wir suchen also Bedingungen, die das verhindern (und die wir für konkrete Vorschriften zur Berechnung von $s^k$ und $\sigma_k$ nachprüfen können), für die also das Verfahren *konvergiert*. Dies wollen wir immer wie folgt verstehen: Ein Verfahren *konvergiert global*, wenn jeder Häufungspunkt $\bar{x}$ einer entsprechend erzeugten Folge $\{x^k\}_{k\in\mathbb{N}}$ für jeden beliebigen Startpunkt $x^0 \in \mathbb{R}^n$ ein stationärer Punkt ist. Beachte, dass man von einem Verfahren, das nur erste Ableitungen verwendet, nicht mehr erwarten kann. Insbesondere bedeutet globale Konvergenz *nicht*, dass das Verfahren gegen einen *globalen* Minimierer konvergiert! Dagegen sprechen wir von *lokaler Konvergenz*, wenn dies nur für Startwerte $x^0 \in B_\varepsilon(\bar{x})$ für ein $\varepsilon > 0$ und einen stationären Punkt $\bar{x}$ gelten muss.

Wir beginnen mit Bedingungen an die Suchrichtungen $s^k$. Eine Folge $\{s^k\}_{k\in\mathbb{N}} \subset \mathbb{R}^n$ nennen wir Folge von *zulässigen Suchrichtungen*, wenn gilt:

(i) $\nabla f(x^k)^T s^k < 0$ für alle $k \in \mathbb{N}$ (d. h. alle $s^k$ sind *Abstiegsrichtungen*);

(ii) aus $\dfrac{\nabla f(x^k)^T s^k}{\|s^k\|} \to 0$ folgt $\nabla f(x^k) \to 0$.

Zum Beispiel erzeugt die Wahl $s^k := -\nabla f(x^k)$ wegen $-\nabla f(x^k)^T s^k = \|s^k\|^2 = \|\nabla f(x^k)\|^2$ stets zulässige Suchrichtungen.

Bedingung (ii) besagt dabei gerade, dass die Steigung von $f$ in $x^k$ in Richtung $s^k$ nur verschwinden darf, wenn wir einen stationären Punkt erreichen. Eine hinreichende Bedingung dafür ist, dass der Winkel zwischen $\nabla f(x^k)$ und $s^k$ vom rechten Winkel weg beschränkt ist.

Lemma 4.1. *Sei $f : \mathbb{R}^n \to \mathbb{R}$ stetig differenzierbar und sei $\{s^k\}_{k\in\mathbb{N}} \subset \mathbb{R}^n \setminus \{0\}$. Existiert ein $\eta > 0$ mit*

$$\frac{-\nabla f(x^k)^T s^k}{\|\nabla f(x^k)\|\|s^k\|} \geq \eta \qquad \text{für alle } k \in \mathbb{N}, \tag{4.1}$$

*so ist $\{s^k\}_{k\in\mathbb{N}}$ eine Folge von zulässigen Suchrichtungen.*

*Beweis.* Für $\nabla f(x^k), s^k \neq 0$ folgt aus (4.1) sofort

$$-\nabla f(x^k)^T s^k \geq \eta \|\nabla f(x^k)\|\|s^k\| > 0,$$

d. h. $s^k$ ist eine Abstiegsrichtung. Es gelte nun $\frac{\nabla f(x^k)^T s^k}{\|s^k\|} \to 0$. Aus (4.1) folgt dann sofort

$$\|\nabla f(x^k)\| \leq \eta^{-1} \frac{-\nabla f(x^k)^T s^k}{\|s^k\|} \to 0. \qquad \square$$

Die Bedingung (4.1) wird *Winkelbedingung* genannt.

Nun zu den Schrittweiten $\sigma_k$. Eine Folge $\{\sigma_k\}_{k\in\mathbb{N}} \subset \mathbb{R}_{>0}$ nennen wir Folge von *zulässigen Schrittweiten* für $\{s^k\}_{k\in\mathbb{N}}$, wenn gilt:

(i) $f(x^k + \sigma_k s^k) \leq f(x^k)$ für alle $k \in \mathbb{N}$ (d. h. alle $\sigma_k s^k$ sind Abstiegsschritte);

(ii) aus $f(x^k + \sigma_k s^k) - f(x^k) \to 0$ folgt $\dfrac{\nabla f(x^k)^T s^k}{\|s^k\|} \to 0$.

Konkrete Beispiele von Schrittweiten werden wir im nächsten Kapitel untersuchen. Beachte, dass die Bedingungen Bezug auf die Suchrichtungen nehmen – die Zulässigkeit der Schrittweiten hängt also von den verwendeten Suchrichtungen ab! Bedingung (ii) besagt dabei gerade, dass die Reduktion im Funktionswert nicht beliebig klein werden darf, ohne dass die Steigung verschwindet (was bei zulässigen Suchrichtungen nur in der Nähe von stationären Punkten passieren kann). Wir müssen also einen *ausreichenden Abstieg* garantieren. Dies liefert die folgende Definition: Sei $\{s^k\}_{k\in\mathbb{N}}$ eine Folge von Suchrichtungen für $f$. Dann heißt die Folge $\{\sigma_k\}_{k\in\mathbb{N}} \subset (0, \infty)$ *effizient* (für $\{s^k\}_{k\in\mathbb{N}}$), falls ein $\theta > 0$ existiert mit

$$f(x^k + \sigma_k s^k) \leq f(x^k) - \theta \left( \frac{\nabla f(x^k)^T s^k}{\|s^k\|} \right)^2 \qquad \text{für alle } k \in \mathbb{N}.$$

Effiziente Schrittweiten sind zulässig.

Lemma 4.2. *Sei $f : \mathbb{R}^n \to \mathbb{R}$ stetig differenzierbar und sei $\{s^k\}_{k\in\mathbb{N}} \subset \mathbb{R}^n \setminus \{0\}$. Ist die Folge $\{\sigma^k\}_{k\in\mathbb{N}}$ effizient für $\{s^k\}_{k\in\mathbb{N}}$, so ist sie auch zulässig für diese.*

*Beweis.* Wegen $\theta > 0$ folgt aus der Effizienz sofort die Bedingung (i). Für Bedingung (ii) gelte $f(x^k + \sigma_k s^k) - f(x^k) \to 0$. Aus der Effizienz der Schrittweiten folgt dann

$$\left(\frac{\nabla f(x^k)^T s^k}{\|s^k\|}\right)^2 \leq \theta^{-1}\left(f(x^k) - f(x^k + \sigma_k s^k)\right) \to 0. \qquad \square$$

Wir können nun zeigen, dass Algorithmus 4.1 für zulässige Suchrichtungen und Schrittweiten global konvergiert.

Satz 4.3. *Sei $f : \mathbb{R}^n \to \mathbb{R}$ stetig differenzierbar. Dann bricht Algorithmus 4.1 entweder nach endlich vielen Schritten ab, oder er erzeugt Folgen $\{x^k\}_{k\in\mathbb{N}}$, $\{s^k\}_{k\in\mathbb{N}}$, und $\{\sigma_k\}_{k\in\mathbb{N}}$, die nicht endlich sind. Sind die Suchrichtungen $\{s^k\}_{k\in\mathbb{N}}$ und die Schrittweiten $\{\sigma_k\}_{k\in\mathbb{N}}$ zulässig, so ist jeder Häufungspunkt von $\{x^k\}_{k\in\mathbb{N}}$ ein stationärer Punkt von $f$.*

*Beweis.* Wir müssen nur den Fall betrachten, dass der Algorithmus nicht nach endlich vielen Schritten abbricht. Sei dafür $\bar{x}$ ein Häufungspunkt von $\{x^k\}_{k\in\mathbb{N}}$. Dann existiert eine Teilfolge, die wir der Übersichtlichkeit halber mit $\{x^k\}_{k\in K}$ mit $K \subset \mathbb{N}$ unendlich bezeichnen, mit $x^k \to \bar{x}$ für $K \ni k \to \infty$. Aus der Zulässigkeit der Schrittweiten folgt mit Bedingung (i), dass die Folge $\{f(x^k)\}_{k\in\mathbb{N}}$ monoton fallend ist und daher gegen ein $M \in \mathbb{R} \cup \{-\infty\}$ konvergiert. Da $f$ stetig ist, konvergiert die Teilfolge $\{f(x^k)\}_{k\in K}$ gegen $f(\bar{x}) \in \mathbb{R}$, und damit muss die gesamte Folge gegen $f(\bar{x})$ konvergieren. Unter Verwendung der Teleskopsumme erhalten wir daraus

$$f(\bar{x}) - f(x^0) = \lim_{k\to\infty}\left(f(x^k) - f(x^0)\right) = \sum_{k=0}^{\infty}\left(f(x^{k+1}) - f(x^k)\right).$$

Wegen $f(\bar{x}) - f(x^0) \in \mathbb{R}$ hat die Reihe auf der rechten Seite also einen endlichen Wert, und daher muss $\{f(x^{k+1}) - f(x^k)\}_{k\in\mathbb{N}}$ eine Nullfolge sein. Wegen $x^{k+1} = x^k + \sigma_k s^k$ und Bedingung (ii) für die Zulässigkeit der Schrittweiten $\sigma_k$ folgt

$$\frac{\nabla f(x^k)^T s^k}{\|s^k\|} \to 0,$$

und Bedingung (ii) für die Zulässigkeit der Suchrichtungen $s^k$ ergibt dann zusammen mit der stetigen Differenzierbarkeit von $f$

$$\nabla f(\bar{x}) = \lim_{K\ni k\to\infty} \nabla f(x^k) = 0. \qquad \square$$

Bedingung (i) für zulässige Suchrichtungen wurde hier nicht explizit verwendet, wird aber benötigt, um überhaupt einen Abstieg (und damit Bedingung (i) für zulässige Schrittweiten) erhalten zu können.

# 5 SCHRITTWEITENREGELN

Wir betrachten nun einige *Schrittweitenregeln*, d. h. Vorschriften, die für eine gegebene Folge von Suchrichtungen eine Folge von zulässigen Schrittweiten generieren. Eine naheliegende „Vorschrift“ ist die folgende:

**Algorithmus 5.1 :** Minimierungsregel

**Input :** $x, s \in \mathbb{R}^n$

1 Bestimme $\sigma = \arg\min_{\sigma \geq 0} f(x + \sigma s)$

Leider ist diese Regel nur in Ausnahmefällen (z. B. $f$ quadratisch) praktisch durchführbar und nicht einmal in allen Fällen effizient. Wir betrachten daher beispielhaft zwei durchführbare Schrittweitenregeln, die beide in den folgenden Kapiteln zur Berechnung zulässiger Schrittweiten verwendet werden.

## 5.1 ARMIJO-REGEL

Die Armijo-Regel generiert Schrittweiten $\sigma \in (0, 1]$, die über die *Armijo-Bedingung* einen hinreichenden Abstieg garantieren sollen: Für eine gegebene Richtung $s$ und $\gamma \in (0, 1)$ soll gelten

$$f(x + \sigma s) - f(x) \leq \gamma \sigma \nabla f(x)^T s. \tag{5.1}$$

Anschaulich bestimmt diese Regel die größte Schrittweite zwischen 0 und 1, die mindestens den gleichen Abstieg wie die linearisierte Funktion $\varphi(\sigma) = f(x) + \sigma\gamma\nabla f(x)^T s$ erreicht. Üblicherweise wird $\gamma$ klein gewählt, z. B. $\gamma = 10^{-2}$.

Wir müssen zuerst sicherstellen, dass so eine Schrittweite stets existiert.

Lemma 5.1. *Sei $f : \mathbb{R}^n \to \mathbb{R}$ stetig differenzierbar auf der offenen Menge $U \subset \mathbb{R}^n$ und sei $\gamma \in (0, 1)$ gegeben. Ist $s \in \mathbb{R}^n$ eine Abstiegsrichtung für $f$ in $x \in U$, so existiert ein $\bar{\sigma} \in (0, 1]$ mit*

$$f(x + \sigma s) - f(x) \leq \sigma\gamma\nabla f(x)^T s \qquad \text{für alle } \sigma \in [0, \bar{\sigma}].$$

*Beweis.* Nach Definition der Richtungsableitung und Wahl von $\gamma < 1$ gilt

$$\lim_{\sigma \to 0^+} \frac{f(x+\sigma s) - f(x)}{\sigma} - \gamma \nabla f(x)^T s = (1-\gamma)\nabla f(x)^T s < 0.$$

Wegen der strikten Ungleichung im Grenzwert existiert also ein $\bar{\sigma} \in (0,1]$ mit

$$\frac{f(x+\sigma s) - f(x)}{\sigma} - \gamma \nabla f(x)^T s < 0 \qquad \text{für alle } \sigma \in (0, \bar{\sigma}].$$

Da für $\sigma = 0$ die behauptete Ungleichung trivialerweise erfüllt ist, erhalten wir die Aussage. □

Realisieren lässt sich die Armijo-Regel über eine einfache Rückwärtssuche (engl. „backtracking"), deren Konvergenz durch Lemma 5.1 garantiert wird.

**Algorithmus 5.2 :** Armijo-Regel

**Input :** $\beta \in (0,1)$, $\gamma \in (0,1)$, $x, s \in \mathbb{R}^n$

```
1 Setze σ = 1
2 while f(x + σs) − f(x) > σγ∇f(x)^T s do
3     Setze σ ← βσ
```

Die nächste Frage ist nach der Zulässigkeit der Armijo-Schrittweiten. Diese ist nicht in jedem Fall gegeben (betrachte zum Beispiel $f : \mathbb{R} \to \mathbb{R}$, $x \mapsto \frac{1}{8}x^2$, mit den Suchrichtungen $s^k = -2^{-k}\nabla f(x^k)$). Wir müssen daher die zugrundeliegenden Suchrichtungen einschränken.

Satz 5.2. *Sei $f : \mathbb{R}^n \to \mathbb{R}$ stetig differenzierbar, sei $\{x^k\}_{k\in\mathbb{N}} \subset \mathbb{R}^n$ beschränkt, und sei $\{s^k\}_{k\in\mathbb{N}} \subset \mathbb{R}^n$ eine Folge von Abstiegsrichtungen mit*

$$\|s^k\| \geq \varphi\left(\frac{-\nabla f(x^k)^T s^k}{\|s^k\|}\right) \qquad \text{für alle } k \in \mathbb{N} \tag{5.2}$$

*für eine streng monoton wachsende Funktion $\varphi : [0,\infty) \to [0,\infty)$. Dann erzeugt Algorithmus 5.2 eine Folge $\{\sigma_k\}_{k\in\mathbb{N}}$ von zulässigen Schrittweiten.*

*Beweis.* Da für eine Abstiegsrichtung die rechte Seite von (5.1) und damit auch die linke Seite negativ ist, ist die erste Bedingung für Zulässigkeit stets erfüllt. Die zweite Bedingung zeigen wir durch Kontraposition: Angenommen, $\frac{\nabla f(x^k)^T s^k}{\|s^k\|}$ konvergiert für $k \to \infty$ *nicht* gegen 0. Dann muss es eine Teilfolge – die wir wieder mit $k \in \mathbb{N}$ indizieren – sowie ein $\varepsilon > 0$ geben, so dass gilt (beachte $\nabla f(x^k)^T s^k < 0$)

$$-\frac{\nabla f(x^k)^T s^k}{\|s^k\|} \geq \varepsilon \qquad \text{für alle } k \in \mathbb{N}.$$

Aus der Bedingung (5.2) an die $\{s^k\}_{k\in\mathbb{N}}$ folgt nun mit der strikten Monotonie von $\varphi$

$$\|s^k\| \geq \varphi\left(-\frac{\nabla f(x^k)^T s^k}{\|s^k\|}\right) \geq \varphi(\varepsilon) =: \delta > \varphi(0) \geq 0. \tag{5.3}$$

Nach Satz 1.1 existiert nun für alle $k \in \mathbb{N}$ ein $\theta_k \in (0,1)$ mit $\tau_k := \theta_k \sigma_k \in (0, \sigma_k)$ und

$$\begin{aligned}\frac{f(x^k + \sigma_k s^k) - f(x^k)}{\|\sigma_k s^k\|} - \frac{\sigma_k \gamma \nabla f(x^k)^T s^k}{\|\sigma_k s^k\|} &= \frac{\nabla f(x^k + \tau_k s^k)^T s^k}{\|s^k\|} - \frac{\gamma \nabla f(x^k)^T s^k}{\|s^k\|} \\ &\leq \|\nabla f(x^k + \tau_k s^k) - \nabla f(x^k)\| + (1-\gamma)\frac{\nabla f(x^k)^T s^k}{\|s^k\|} \\ &\leq \|\nabla f(x^k + \tau_k s^k) - \nabla f(x^k)\| - (1-\gamma)\varepsilon,\end{aligned}$$

wobei wir im zweiten Schritt die produktive Null und die Cauchy–Schwarz-Ungleichung eingesetzt haben.

Weiter ist nach Annahme $\{x^k\}_{k\in\mathbb{N}}$ beschränkt und $\nabla f$ stetig, es gibt also ein $\rho > 0$ so dass für alle $d \in B_\rho(0)$ gilt

$$\|\nabla f(x^k + d) - \nabla f(x^k)\| < (1-\gamma)\varepsilon \qquad \text{für alle } k \in \mathbb{N}.$$

Die Armijo-Bedingung (5.1) ist also erfüllt, sobald $\sigma_k \leq \rho \|s^k\|^{-1}$ ist.

Nun ist die Armijo-Schrittweite stets als die maximale Schrittweite der Form $\sigma_k = \beta^{m-1}$, $m \in \mathbb{N}$, gewählt, die die Armijo-Bedingung (5.1) erfüllt. Also ist entweder $\sigma_k = 1$ und damit wegen (5.3) $\|s^k\| \geq \delta > 0$, oder $\sigma_k \leq \beta$ und $\sigma_k/\beta > \rho\|s^k\|^{-1} > 0$. In beiden Fällen haben wir

$$\sigma_k \|s^k\| \geq \min\{\delta, \beta\rho\} =: \eta > 0 \qquad \text{für alle } k \in \mathbb{N}.$$

Die Armijo-Bedingung (5.1) garantiert also

$$f(x^k) - f(x^k + \sigma_k s^k) \geq -\sigma_k \gamma \nabla f(x^k)^T s^k = \gamma\left(-\frac{\nabla f(x^k)^T s^k}{\|s^k\|}\right)(\sigma_k \|s^k\|) \geq \gamma\varepsilon\eta > 0$$

für alle $k \in \mathbb{N}$, und damit kann auch $f(x^k + \sigma_k s^k) - f(x^k)$ nicht gegen 0 konvergieren. □

## 5.2 POWELL–WOLFE-REGEL

Die Powell–Wolfe-Regel (manchmal auch Wolfe–Powell-Regel genannt) soll garantieren, dass auch bei kurzen Suchrichtungen $s$ der tatsächliche Schritt $\sigma s$ hinreichend groß ist. Dafür wird neben der Armijo-Bedingung (5.1) für $\gamma \in (0, \frac{1}{2})$ (beachte die Einschränkung!) zusätzlich gefordert, dass für ein $\eta \in (\gamma, 1)$ gilt

$$\nabla f(x + \sigma s)^T s \geq \eta \nabla f(x)^T s. \tag{5.4}$$

Anschaulich bedeutet diese Bedingung, dass neben dem Funktionswert auch der Betrag der (negativen) Steigung von $f$ in Richtung $s$ hinreichend reduziert wird; man bezeichnet daher (5.4) auch als *Krümmungsbedingung* und (5.1) und (5.4) zusammen als *Powell–Wolfe-Bedingungen*. Üblicherweise wird dabei $\gamma$ klein und $\eta$ groß gewählt, z. B. $\gamma = 10^{-2}, \eta = 0.9$.

Wir zeigen zuerst wieder, dass eine Schrittweite, die diese Bedingungen erfüllt, unter sinnvollen Annahmen stets existiert.

Lemma 5.3. *Sei $f : \mathbb{R}^n \to \mathbb{R}$ stetig differenzierbar und nach unten beschränkt, und sei $s \in \mathbb{R}^n$ eine Abstiegsrichtung für $f$ in $x \in \mathbb{R}^n$. Dann existiert für alle $\gamma \in (0, \frac{1}{2})$ und $\eta \in (\gamma, 1)$ ein $\sigma > 0$, so dass gilt*

$$f(x + \sigma s) - f(x) \leq \gamma\sigma\nabla f(x)^T s,$$
$$\nabla f(x + \sigma s)^T s \geq \eta\nabla f(x)^T s.$$

*Beweis.* Wir beginnen mit der Armijo-Bedingung und betrachten dafür die Funktion

$$\psi(\sigma) := f(x + \sigma s) - f(x) - \sigma\gamma\nabla f(x)^T s, \tag{5.5}$$

die nach Annahme an $f$ ebenfalls stetig differenzierbar ist. Wegen $\psi(0) = 0$ und $\psi'(0) = (1 - \gamma)\nabla f(x)^T s < 0$ ($s$ ist Abstiegsrichtung) ist nun $\psi(\sigma) < 0$ für $\sigma > 0$ klein genug. Da $f$ nach unten beschränkt und $s$ eine Abstiegsrichtung ist, gilt $\psi(\sigma) \to \infty$ für $\sigma \to \infty$. Wegen der Stetigkeit von $\psi$ existiert daher ein $\bar{\sigma} > 0$ mit $\psi(\sigma) < 0$ für alle $\sigma \in (0, \bar{\sigma})$ und

$$0 = \psi(\bar{\sigma}) = f(x + \bar{\sigma} s) - f(x) - \bar{\sigma}\gamma\nabla f(x)^T s,$$

d. h. $\bar{\sigma} > 0$ erfüllt die Armijo-Bedingung (mit Gleichheit).

Diese Wahl von $\bar{\sigma}$ garantiert auch, dass gilt

$$\nabla f(x + \bar{\sigma} s)^T s - \gamma\nabla f(x)^T s = \psi'(\bar{\sigma}) = \lim_{t \to 0^+} \frac{\psi(\bar{\sigma}) - \psi(\bar{\sigma} - t)}{t} \geq 0,$$

und mit der Wahl $\eta > \gamma$ folgt nun

$$\nabla f(x + \bar{\sigma} s)^T s \geq \gamma\nabla f(x)^T s > \eta\nabla f(x)^T s, \tag{5.6}$$

d. h. $\bar{\sigma} > 0$ erfüllt auch die Krümmungsbedingung (5.4). □

Der Beweis von Lemma 5.3 ist konstruktiv und gibt uns daher ein Verfahren zur Hand, eine Powell–Wolfe-Schrittweite zu bestimmen: Wir suchen die Nullstelle $\bar{\sigma}$ der durch (5.5) definierten Funktion $\psi$ mit Hilfe des Bisektions-Verfahrens. Dazu suchen wir in einem ersten Schritt eine Untergrenze $\sigma_-$ mit $\psi(\sigma_-) \leq 0$ (d. h. die Armijo-Bedingung ist erfüllt) und eine Obergrenze $\sigma_+$ mit $\psi(\sigma_+) > 0$ (d. h. die Armijo-Bedingung ist verletzt). In einem zweiten Schritt halbieren wir das Intervall $[\sigma_-, \sigma_+]$ so lange, bis $\sigma_-$ nahe genug an $\bar{\sigma}$ liegt, dass auch die Krümmungsbedingung erfüllt ist. Der folgende Algorithmus setzt dieses Verfahren basierend auf der Armijo-Regel (Algorithmus 5.2) mit $\beta = \frac{1}{2}$ um. Dabei ist zu beachten, dass wir auch Schrittweiten $\sigma_- > 1$ zulassen müssen.

**Algorithmus 5.3 :** Powell–Wolfe-Regel

**Input :** $\gamma \in (0, \frac{1}{2})$, $\eta \in (\gamma, 1)$, $x, s \in \mathbb{R}^n$

`// Intervall-Phase: finde` $\sigma_-$ `mit` $\psi(\sigma_-) \leq 0$ `und` $\sigma_+$ `mit` $\psi(\sigma_+) > 0$

1 Bestimme $\sigma_-$ mit Algorithmus 5.2 für $\beta = \frac{1}{2}$
2 **if** $\sigma_- = 1$ **then** `// σ = 1 erfüllt Armijo-Bedingung`
3  **if** $\nabla f(x + \sigma_- s)^T s \geq \eta \nabla f(x)^T s$ **then** `// σ = 1 erfüllt Krümmungsbedingung`
4  **return** $\sigma_-$ `// Akzeptiere Schrittweite 1`
5  **else**
6  **for** $\sigma \in \{2^k : k \in \mathbb{N}\}$ **do** `// kleinstes` $\sigma_+$`, das Armijo-Bedingung verletzt`
7  **if** $f(x + \sigma s) - f(x) > \sigma\gamma \nabla f(x)^T s$ **then** `//` $\sigma$ `verletzt Armijo-Bedingung`
8  Setze $\sigma_+ = \sigma, \quad \sigma_- = \frac{1}{2}\sigma_+$
9  **break** `// Beende Intervall-Phase`
10 **else**
11  Setze $\sigma_+ = 2\sigma_-$

`// Bisektions-Phase: finde` $\sigma_-$ `mit` $\psi(\sigma_-) \leq 0$ `und Krümmungsbedingung`

12 **while** $\nabla f(x + \sigma_- s)^T s < \eta \nabla f(x)^T s$ **do** `// Krümmungsbedingung verletzt`
13  Setze $\sigma = \frac{1}{2}(\sigma_- + \sigma_+)$
14  **if** $f(x + \sigma s) - f(x) \leq \sigma\gamma \nabla f(x)^T s$ **then** `//` $\sigma$ `erfüllt Armijo-Bedingung`
15  Setze $\sigma_- = \sigma$
16  **else** `//` $\sigma$ `verletzt Armijo-Bedingung`
17  Setze $\sigma_+ = \sigma$
18 **return** $\sigma_-$ `// Akzeptiere Schrittweite` $\sigma_-$

Dieses Verfahren erzeugt unter sinnvollen Annahmen stets Schrittweiten, die den Powell–Wolfe-Bedingungen genügen.

Satz 5.4. *Sei $f : \mathbb{R}^n \to \mathbb{R}$ stetig differenzierbar und nach unten beschränkt, und sei $s \in \mathbb{R}^n$ eine Abstiegsrichtung für $f$ in $x \in \mathbb{R}^n$. Seien weiter $\gamma \in (0, \frac{1}{2})$ und $\eta \in (\gamma, 1)$. Dann bricht Algorithmus 5.3 nach endlich vielen Schritten ab mit einer Schrittweite $\sigma > 0$, die die Powell–Wolfe-Bedingungen erfüllt.*

*Beweis.* Wir betrachten zuerst die Intervall-Phase. Schritt 1 ruft Algorithmus 5.2 auf, der wegen Lemma 5.1 nach endlich vielen Schritten ein Ergebnis liefert. Da $f$ nach unten beschränkt und $s$ eine Abstiegsrichtung ist, gilt

$$\psi(\sigma) = f(x + \sigma s) - f(x) - \sigma\gamma \nabla f(x)^T s \to \infty$$

für $\sigma \to \infty$. Also ist die Armijo-Bedingung (5.1) für $\sigma$ groß genug verletzt. Die Intervall-Phase endet also nach endlich vielen Schritten mit einem Paar $(\sigma_-, \sigma_+)$ mit $\sigma_- < \sigma_+$, so dass $\sigma_-$ die Armijo-Bedingung erfüllt, $\sigma_+$ aber nicht.

Die Bisektions-Phase halbiert nun in jeder Iteration die Länge des Intervalls $[\sigma_-, \sigma_+]$, wobei diese Eigenschaften nach Konstruktion erhalten bleiben. Insbesondere gilt stets $\psi(\sigma_-) \leq 0 < \psi(\sigma_+)$. Angenommen, die Schleife in der Bisektions-Phase bricht nicht nach endlich vielen Schritten ab. Da in jedem Schritt entweder $\sigma_-$ vergrößert oder $\sigma_+$ verkleinert wird, muss dann ein $\bar{\sigma}$ existieren mit $\sigma_- \to \bar{\sigma}$ (von links) und $\sigma_+ \to \bar{\sigma}$ (von rechts). Da mit $f$ und $\nabla f$ auch $\psi$ stetig ist, folgt $\psi(\bar{\sigma}) = 0$. Wie im Beweis von Lemma 5.3 erhält man nun aus dem Vorzeichenwechsel von $\psi$ in $\bar{\sigma}$, dass $\psi'(\bar{\sigma}) \geq 0$ und damit

$$\nabla f(x + \bar{\sigma}s)^T s \geq \gamma \nabla f(x)^T s > \eta \nabla f(x)^T s$$

gilt. Wegen der Stetigkeit von $\nabla f$ ist diese strikte(!) Ungleichung auch erfüllt für $\sigma_-$ hinreichend nahe bei $\bar{\sigma}$, für das die Iteration aber abbrechen würde im Widerspruch zur Annahme. □

Die erzeugten Schrittweiten sind auch zulässig; unter etwas stärkeren Annahmen an die Funktion (aber nicht an die Suchrichtungen, vergleiche Satz 5.2) können wir sogar Effizienz zeigen.

Satz 5.5. *Sei $f : \mathbb{R}^n \to \mathbb{R}$ Lipschitz-stetig differenzierbar und nach unten beschränkt und sei $\{s^k\}_{k\in\mathbb{N}} \subset \mathbb{R}^n$ eine Folge von Abstiegsrichtungen. Dann erzeugt Algorithmus 5.3 eine Folge $\{\sigma_k\}_{k\in\mathbb{N}}$ von effizienten Schrittweiten.*

*Beweis.* Unter den gegebenen Voraussetzungen an $f$ ist nach Satz 5.4 die Powell–Wolfe-Regel stets durchführbar. Die Schrittweiten $\{\sigma_k\}_{k\in\mathbb{N}}$ erfüllen also alle die Powell–Wolfe-Bedingungen. Insbesondere impliziert die Krümmungsbedingung zusammen mit der Lipschitz-Stetigkeit von $\nabla f$, dass für alle $k \in \mathbb{N}$ gilt

$$\begin{aligned}(\eta - 1)\nabla f(x^k)^T s^k &\leq \left(\nabla f(x^k + \sigma_k s^k) - \nabla f(x^k)\right)^T s^k \\ &\leq \|\nabla f(x^k + \sigma_k s^k) - \nabla f(x^k)\| \|s^k\| \leq L\sigma_k \|s^k\|^2.\end{aligned}$$

Daraus folgt

$$\sigma_k \geq \frac{\eta - 1}{L} \frac{\nabla f(x^k)^T s^k}{\|s^k\|^2}$$

und damit wegen der Armijo-Bedingung und $\nabla f(x^k)^T s^k < 0$

$$f(x^k + \sigma_k s^k) \leq f(x^k) + \gamma\sigma_k \nabla f(x^k)^T s^k \leq f(x^k) - \theta \left(\frac{\nabla f(x^k)^T s^k}{\|s^k\|}\right)^2$$

für $\theta := (1 - \eta)\gamma L^{-1} > 0$, d. h. $\sigma_k$ ist effizient. □

# 6 DAS GRADIENTENVERFAHREN

Als Prototypen eines Abstiegsverfahrens betrachten wir nun das *Gradientenverfahren*, das auf der Wahl des negativen Gradienten als Suchrichtung $s^k := -\nabla f(x^k)$ beruht. Offensichtlich führt diese Wahl stets auf eine Abstiegsrichtung; tatsächlich handelt es sich um die Richtung des steilsten Abstiegs (weshalb man im Englischen oft auch von der *method of steepest descent* spricht).

Lemma 6.1. *Sei $f : \mathbb{R}^n \to \mathbb{R}$ differenzierbar und $x \in \mathbb{R}^n$ mit $\nabla f(x) \neq 0$. Dann gilt für $s := -\frac{\nabla f(x)}{\|\nabla f(x)\|}$*

$$\nabla f(x)^T s = \min_{\|d\|=1} \nabla f(x)^T d.$$

*Beweis.* Aus der Cauchy–Schwarz-Ungleichung folgt, dass für alle $d \in \mathbb{R}^n$ mit $\|d\| = 1$ gilt

$$\nabla f(x)^T d \geq -\|\nabla f(x)\|\|d\| = -\|\nabla f(x)\| = \nabla f(x)^T s,$$

mit Gleichheit für $d = s$. □

Ergänzt wird diese Wahl der Suchrichtung durch die Armijo-Regel für die Schrittweite.

**Algorithmus 6.1 :** Gradientenverfahren

```
1 Wähle x^0 ∈ ℝ^n,    setze k = 0
2 while ‖∇f(x^k)‖ > 0 do
3     Setze s^k := −∇f(x^k)
4     Bestimme σ_k > 0 mit Algorithmus 5.2
5     Setze x^{k+1} = x^k + σ_k s^k,    k ← k + 1
```

In der Praxis stoppt man bereits, wenn $\|\nabla f(x^k)\| \leq \varepsilon$ für eine vorgegebene Toleranz $\varepsilon > 0$ (z. B. $\varepsilon = 10^{-8}$) erreicht ist.

Für die globale Konvergenz dieses Verfahrens können wir den abstrakten Satz 4.3 heranziehen.

Satz 6.2. *Sei $f : \mathbb{R}^n \to \mathbb{R}$ stetig differenzierbar. Dann bricht Algorithmus 6.1 entweder nach endlich vielen Schritten ab oder erzeugt eine Folge $\{x^k\}_{k\in\mathbb{N}}$, von der jeder Häufungspunkt ein stationärer Punkt von $f$ ist.*

*Beweis.* Wir müssen lediglich nachweisen, dass diese Wahl von Suchrichtungen und Schrittweiten zulässig ist. Solange $x^k$ kein stationärer Punkt ist, gilt $s^k = -\nabla f(x^k) \neq 0$ und daher gilt für alle $k \in \mathbb{N}$

$$\frac{-\nabla f(x^k)^T s^k}{\|\nabla f(x^k)\|\|s^k\|} = \frac{\|\nabla f(x^k)^T\|^2}{\|\nabla f(x^k)^T\|^2} = 1 > 0,$$

d. h. die Winkelbedingung (4.1) ist für $\eta = 1$ erfüllt. Damit sind die negativen Gradienten nach Lemma 4.1 zulässige Suchrichtungen sowie die Armijo-Schrittweiten nach Satz 5.2 (mit $\varphi(t) = t$) zulässige Schrittweiten. Die Aussage folgt nun aus Satz 4.3. □

Zwar sind die Abstiegsrichtungen lokal optimal, das bedeutet aber noch nicht, dass dies auch *global* (d. h. in Hinblick auf die minimale Anzahl von Iterationen) der Fall ist. Tatsächlich kann die Konvergenz des Gradientenverfahrens beliebig langsam sein – und das bereits im "Idealfall" einer strikt konvexen, quadratischen Zielfunktion!

Wir betrachten in Folge für $A \in \mathbb{R}^{n\times n}$ symmetrisch und positiv definit und $b \in \mathbb{R}^n$ die Funktion

$$f : \mathbb{R}^n \to \mathbb{R}, \qquad x \mapsto \frac{1}{2}x^T Ax + b^T x. \tag{6.1}$$

Man rechnet leicht nach, dass dann gilt

$$\nabla f(x) = Ax + b,$$
$$\nabla^2 f(x) = A.$$

Insbesondere ist daher nach Annahme und Satz 1.5 (ii) $f$ strikt konvex. In diesem Fall ist sogar Algorithmus 5.1 praktikabel. Für $x, s \in \mathbb{R}^n$ wird dabei $\sigma$ bestimmt als Minimierer $\bar{\sigma}$ der Funktion

$$\varphi : (0, \infty) \to \mathbb{R}, \qquad \sigma \mapsto f(x + \sigma s).$$

Da $f$ strikt konvex ist, ist auch $\varphi$ strikt konvex. Nach Satz 2.1 ist der eindeutige Minimierer $\bar{\sigma}$ von $\varphi$ charakterisiert durch

$$0 = \varphi'(\bar{\sigma}) = \nabla f(x + \bar{\sigma}s)^T s = (A(x + \bar{\sigma}s) + b)^T s.$$

Auflösen nach $\bar{\sigma}$ und Verwenden von $s = -\nabla f(x) = -(Ax + b)$ ergibt dann

$$\bar{\sigma} = \frac{\|s\|^2}{s^T As}. \tag{6.2}$$

Selbst mit dieser Wahl konvergiert das Gradientenverfahren beliebig langsam. Wie wir sehen werden, hängt die Konvergenzgeschwindigkeit von den Eigenwerten der Matrix $A$ ab, und zwar speziell vom Verhältnis des größten zum kleinsten Eigenwert, das als *Konditionszahl* $\kappa := \lambda_n/\lambda_1 \geq 1$ bezeichnet wird. Um dies zu zeigen, verwenden wir die folgende nützliche Ungleichung.

Lemma 6.3 (Kantorovich-Ungleichung). *Sei $A \in \mathbb{R}^{n\times n}$ symmetrisch und positiv definit. Dann gilt*

$$\frac{(x^TAx)(x^TA^{-1}x)}{(x^Tx)^2} \leq \frac{1}{4}\left(\kappa^{\frac{1}{2}} + \kappa^{-\frac{1}{2}}\right)^2 \quad \text{für alle } x \in \mathbb{R}^n \setminus \{0\}.$$

*Beweis.* Sei $\mu = (\lambda_n\lambda_1)^{\frac{1}{2}}$ das geometrische Mittel der beiden extremalen Eigenwerte und setze $B := \mu^{-1}A + \mu A^{-1}$. Dann ist $B$ symmetrisch und positiv definit und besitzt die Eigenwerte $\mu^{-1}\lambda_i + \mu\lambda_i^{-1}$ für $1 \leq i \leq n$. Weiter ist für alle $1 \leq i \leq n$

$$\kappa^{-\frac{1}{2}} = \mu^{-1}\lambda_n \leq \mu^{-1}\lambda_i \leq \mu^{-1}\lambda_1 = \kappa^{\frac{1}{2}}.$$

Da die Funktion $z \mapsto z + z^{-1}$ auf $(0,1]$ und $[1,\infty)$ monoton (fallend bzw. steigend) ist, erhalten wir aus der ersten bzw. zweiten Ungleichung (je nachdem, ob $\mu^{-1}\lambda_i < 1$ oder nicht)

$$\mu^{-1}\lambda_i + \mu\lambda_i^{-1} \leq \kappa^{\frac{1}{2}} + \kappa^{-\frac{1}{2}}, \quad 1 \leq i \leq n.$$

Mit dieser Abschätzung der Eigenwerte von $B$ und (1.1) erhalten wir daher

$$\mu^{-1}(x^TAx) + \mu(x^TA^{-1}x) = x^TBx \leq (\kappa^{\frac{1}{2}} + \kappa^{-\frac{1}{2}})(x^Tx).$$

Wir wenden jetzt auf die linke Seite die *(verallgemeinerte) Youngsche Ungleichung*

$$(ab)^{\frac{1}{2}} = (\mu^{-1}a)^{\frac{1}{2}}(\mu b)^{\frac{1}{2}} \leq \frac{1}{2}\left(\mu^{-1}a + \mu b\right)$$

für $a = x^TAx$ und $b = x^TA^{-1}x$ an, quadrieren beide Seiten, und erhalten die Kantorovich-Ungleichung. □

Mit dieser Ungleichung bekommen wir eine Abschätzung der Distanz der Iterierten $x^k$ zum eindeutigen Minimierer $\bar{x}$ von $f$.

Satz 6.4. *Sei $f : \mathbb{R}^n \to \mathbb{R}$ gegeben durch (6.1) für $A \in \mathbb{R}^{n\times n}$ symmetrisch und positiv definit und $b \in \mathbb{R}^n$. Sei $\bar{x} \in \mathbb{R}^n$ der eindeutige Minimierer von $f$. Weiter sei die Folge $\{x^k\}_{k\in\mathbb{N}}$ erzeugt durch Algorithmus 6.1 mit Algorithmus 5.1 anstelle von Algorithmus 5.2. Dann gilt*

$$f(x^{k+1}) - f(\bar{x}) \leq \left(\frac{\kappa-1}{\kappa+1}\right)^2 (f(x^k) - f(\bar{x})),$$

*sowie*

$$\|x^{k+1} - \bar{x}\| \leq \sqrt{\kappa}\left(\frac{\kappa-1}{\kappa+1}\right)\|x^k - \bar{x}\|.$$

*Beweis.* Da die Hesse-Matrix $\nabla^2 f$ konstant ist, erhalten wir aus Satz 1.2 für $y = \bar{x}$ und beliebige $x \in \mathbb{R}^n$

$$f(x) - f(\bar{x}) = \nabla f(\bar{x})^T (x - \bar{x}) + \frac{1}{2}(x - \bar{x})^T A(x - \bar{x}) = \frac{1}{2}(x - \bar{x})^T A(x - \bar{x}), \tag{6.3}$$

da in $\bar{x}$ die notwendige Optimalitätsbedingung $\nabla f(\bar{x}) = 0$ erfüllt ist. Durch Ableiten beider Seiten nach $x$ folgt (wieder mit $\nabla f(\bar{x}) = 0$)

$$\nabla f(x) = A(x - \bar{x}). \tag{6.4}$$

Analog folgt aus Satz 1.2 für $y = x^k$ und $x = x^{k+1} = x^k - \sigma_k s^k$ wegen $s^k = -\nabla f(x^k)$

$$\begin{aligned} f(x^{k+1}) &= f(x^k) + \sigma_k \nabla f(x^k)^T s^k + \frac{\sigma_k}{2}(s^k)^T A s^k \\ &= f(x^k) - \sigma_k \|s^k\|^2 + \frac{\sigma_k^2}{2}(s^k)^T A s^k. \end{aligned}$$

Wir verwenden jetzt die Schrittweitenwahl (6.2) und erhalten

$$\begin{aligned} f(x^{k+1}) - f(\bar{x}) &= f(x^k) - f(\bar{x}) - \sigma_k \|s^k\|^2 + \frac{\sigma_k^2}{2}(s^k)^T A s^k \\ &= f(x^k) - f(\bar{x}) - \frac{\|s^k\|^4}{(s^k)^T A s^k} + \frac{1}{2}\frac{\|s^k\|^4}{(s^k)^T A s^k} \\ &= f(x^k) - f(\bar{x}) - \frac{1}{2}\frac{\|s^k\|^4}{(s^k)^T A s^k}. \end{aligned} \tag{6.5}$$

Andererseits können wir mit Hilfe von (6.3), $I = A^{-1}A$ und $A^T = A$ schreiben

$$\begin{aligned} f(x^k) - f(\bar{x}) &= \frac{1}{2}(x^k - \bar{x})^T A(x^k - \bar{x}) = \frac{1}{2}(x^k - \bar{x})^T A A^{-1} A(x^k - \bar{x}) \\ &= \frac{1}{2}\left(A(x^k - \bar{x})\right)^T A^{-1}\left(A(x^k - \bar{x})\right) \\ &= \frac{1}{2}(s^k)^T A^{-1} s^k, \end{aligned}$$

wobei wir im letzten Schritt (6.4) und $s^k = -\nabla f(x^k) = -A(x^k - \bar{x})$ verwendet haben. Zusammen mit (6.5) erhalten wir

$$\begin{aligned} f(x^{k+1}) - f(\bar{x}) &= f(x^k) - f(\bar{x}) - \frac{1}{2}\frac{\|s^k\|^4}{(s^k)^T A s^k} \\ &= f(x^k) - f(\bar{x}) - \frac{1}{2}\frac{\|s^k\|^4}{(s^k)^T A s^k}\frac{f(x^k) - f(\bar{x})}{\frac{1}{2}(s^k)^T A^{-1} s^k} \\ &= \left(1 - \frac{\|s^k\|^4}{\left((s^k)^T A s^k\right)\left((s^k)^T A^{-1} s^k\right)}\right)(f(x^k) - f(\bar{x})). \end{aligned}$$

Auf den Bruch wenden wir nun die Kantorovich-Ungleichung an und bringen die Klammer auf einen Nenner. Dies ergibt die erste Abschätzung.

Die zweite Abschätzung folgt aus der ersten, denn (6.3) in Verbindung mit (1.1) ergibt für alle $x \in \mathbb{R}^n$

$$\frac{\lambda_1}{2}\|x-\bar{x}\|^2 \le \frac{1}{2}(x-\bar{x})^T A(x-\bar{x}) = f(x) - f(\bar{x}) \le \frac{\lambda_n}{2}\|x-\bar{x}\|^2.$$

Damit erhalten wir

$$\begin{aligned}\|x^{k+1}-\bar{x}\|^2 &\le \frac{2}{\lambda_1}\left(f(x^{k+1}) - f(\bar{x})\right)\\ &\le \frac{2}{\lambda_1}\left(\frac{\kappa-1}{\kappa+1}\right)^2\left(f(x^k) - f(\bar{x})\right)\\ &\le \frac{\lambda_n}{\lambda_1}\left(\frac{\kappa-1}{\kappa+1}\right)^2\|x^k-\bar{x}\|^2,\end{aligned}$$

und Wurzelziehen ergibt die Aussage. □

Durch Induktion erhalten wir daraus die folgende Abschätzung.

Folgerung 6.5. *Sei $f : \mathbb{R}^n \to \mathbb{R}$ gegeben durch (6.1) für $A \in \mathbb{R}^{n\times n}$ symmetrisch und positiv definit und $b \in \mathbb{R}^n$. Dann gilt für das Gradientenverfahren mit Minimierungsregel die Fehlerabschätzung*

$$\|x^{k+1}-\bar{x}\| \le \sqrt{\kappa}\left(\frac{\kappa-1}{\kappa+1}\right)^k\|x^0-\bar{x}\|.$$

*Beweis.* Aus dem Beweis von Satz 6.4 folgt sofort

$$\|x^{k+1}-\bar{x}\|^2 \le \frac{2}{\lambda_1}\left(\frac{\kappa-1}{\kappa+1}\right)^{2k}\left(f(x^0) - f(\bar{x})\right) \le \frac{\lambda_n}{\lambda_1}\left(\frac{\kappa-1}{\kappa+1}\right)^{2k}\|x^0-\bar{x}\|^2$$

und damit die Behauptung. □

Je größer also die Konditionszahl der Matrix $A$, desto näher ist der Bruch auf der rechten Seite an 1, und desto langsamer konvergiert die Folge auf der linken Seite gegen Null. Wir brauchen also Verfahren, die die schlechte Kondition der Matrix $A$ bei der Wahl der Suchrichtungen berücksichtigen (und hoffentlich kompensieren) können.

# 7 NEWTON-ARTIGE VERFAHREN

Die zweite große Klasse von Optimierungsverfahren basiert auf der Idee, die notwendige Optimalitätsbedingung $\nabla f(x) = 0$ durch eine (präkonditionierte) Fixpunktiteration zu berechnen: Offensichtlich ist $\bar{x}$ Nullstelle von $\nabla f(x)$ genau dann, wenn für eine beliebige *invertierbare* Matrix $B = B(\bar{x})$ gilt

$$\bar{x} = \bar{x} - B(\bar{x})\nabla f(\bar{x}).$$

Die zugehörige Fixpunktiteration ist

$$x^{k+1} = x^k - B(x^k)\nabla f(x^k).$$

Schreibt man diese Iteration um unter Verwendung der Inversen $H_k := B(x^k)^{-1}$ und Einführen von $s^k := x^{k+1} - x^k$, führt das auf den folgenden abstrakten Algorithmus.

**Algorithmus 7.1 :** Allgemeines Newton-artiges Verfahren

1 Wähle einen *Startpunkt* $x^0 \in \mathbb{R}^n$, setze $k = 0$
2 **while** $\|\nabla f(x^k)\| > 0$ **do**
3 Wähle eine invertierbare Matrix $H_k \in \mathbb{R}^{n\times n}$
4 Berechne $s^k$ als Lösung von $H_k s^k = -\nabla f(x^k)$
5 Setze $x^{k+1} = x^k + s^k$, $k \leftarrow k+1$

Die Wahl der Schrittweite $\sigma_k$ steckt dabei als Skalierung in der Wahl der Matrix $H_k$. Motivation ist hier natürlich das Newton-Verfahren (mit $H_k := \nabla^2 f(x^k)$), das wir im nächsten Kapitel eingehend untersuchen werden.

Unter gewissen Voraussetzungen an die Matrizen $H_k$ sind die so berechneten Suchrichtungen $s^k$ zulässig.

Lemma 7.1. *Seien die Matrizen $H_k \in \mathbb{R}^{n\times n}$ so gewählt, dass gilt:*

*(i) $H_k$ ist symmetrisch und positiv definit für alle $k \in \mathbb{N}$;*

*(ii) es gibt Konstanten $0 < \mu_1 \le \mu_2$, so dass für alle $k \in \mathbb{N}$ die Eigenwerte von $H_k$ die Abschätzung*

$$\mu_1 \le \lambda_{k,1} \le \lambda_{k,n} \le \mu_2$$

*erfüllen.*

*Dann erzeugt Algorithmus 7.1 eine zulässige Folge von Suchrichtungen* $\{s^k\}_{k\in\mathbb{N}}$.

*Beweis.* Zunächst impliziert die positive Definitheit die Invertierbarkeit aller $H_k$. Also gilt für $\nabla f(x^k) \neq 0$ auch $s^k = -H_k^{-1}\nabla f(x^k) \neq 0$, und unter Verwendung von (1.1) folgt

$$-\nabla f(x^k)^T s^k = (s^k)^T H_k s^k \geq \lambda_{k,1}\|s^k\|^2 \geq \mu_1\|s^k\|^2.$$

Nun gilt wegen $\|H_k\| = \lambda_{k,n}$ stets

$$\|\nabla f(x^k)\| = \|H_k H_k^{-1}\nabla f(x^k)\| \leq \lambda_{k,n}\|H_k^{-1}\nabla f(x^k)\| \leq \mu_2\|H_k^{-1}\nabla f(x^k)\|.$$

Zusammen erhalten wir

$$-\nabla f(x^k)^T s^k \geq \mu_1\|s^k\|^2 = \mu_1\|H_k^{-1}\nabla f(x^k)\|\|s^k\| \geq \frac{\mu_1}{\mu_2}\|\nabla f(x^k)\|\|s^k\|,$$

d. h. die Winkelbedingung (4.1) ist erfüllt für $\eta = \frac{\mu_1}{\mu_2}$. Nach Lemma 4.1 ist die Folge $\{s^k\}_{k\in\mathbb{N}}$ also zulässig. □

Newton-artige Verfahren sind im allgemeinen deutlich aufwendiger als Abstiegsverfahren, da in jedem Schritt ein lineares Gleichungssystem gelöst werden muss. Damit sich der Aufwand lohnt, sollten also deutlich weniger Iterationen notwendig sein, um einen vorgegebenen Abstand $\|x^k - \bar{x}\| \leq \varepsilon$ zu einem stationären Punkt $\bar{x}$ zu erreichen. Dies kann mit dem Begriff der *Konvergenzgeschwindigkeit* einer Folge mathematisch präzisiert werden.

Wir sagen, eine Folge $\{x^k\}_{k\in\mathbb{N}} \subset \mathbb{R}^n$ konvergiert gegen $\bar{x} \in \mathbb{R}^n$

(i) *linear*, falls ein $c \in (0,1)$ existiert mit

$$\|x^{k+1} - \bar{x}\| \leq c\|x^k - \bar{x}\| \qquad \text{für alle } k \in \mathbb{N} \text{ hinreichend groß,}$$

(ii) *superlinear*, falls eine Nullfolge $\{\varepsilon_k\}_{k\in\mathbb{N}}$ existiert mit

$$\|x^{k+1} - \bar{x}\| \leq \varepsilon_k\|x^k - \bar{x}\| \qquad \text{für alle } k \in \mathbb{N} \text{ hinreichend groß,}$$

(iii) *quadratisch*, falls $x^k \to \bar{x}$ und ein $C > 0$ existiert mit

$$\|x^{k+1} - \bar{x}\| \leq C\|x^k - \bar{x}\|^2 \qquad \text{für alle } k \in \mathbb{N} \text{ hinreichend groß.}$$

(Diese Begriffe werden in der Literatur auch als *q-lineare* (-superlineare, -quadratische) Konvergenz – im Gegensatz zu der weniger häufig verwendeten *r-linearen* (-superlinearen, -quadratischen) Konvergenz – bezeichnet.) Die letzten beiden Bedingungen kann man auch mit Hilfe der *Landau-Symbole* formulieren: Gilt $x^k \neq \bar{x}$ für alle $k \in \mathbb{N}$, so ist $\{x^k\}_{k\in\mathbb{N}}$

(i) superlinear konvergent genau dann, wenn gilt

$$\lim_{k\to\infty} \frac{\|x^{k+1} - \bar{x}\|}{\|x^k - \bar{x}\|} = 0,$$

geschrieben $\|x^{k+1} - \bar{x}\| = o(\|x^k - \bar{x}\|)$;

(ii) quadratisch konvergent genau dann, wenn gilt

$$\limsup_{k\to\infty} \frac{\|x^{k+1} - \bar{x}\|}{\|x^k - \bar{x}\|^2} < \infty,$$

geschrieben $\|x^{k+1} - \bar{x}\| = O(\|x^k - \bar{x}\|^2)$.

Der Rest dieses Kapitels ist der (eher technischen) Herleitung von Bedingungen gewidmet, unter denen eine durch Algorithmus 7.1 erzeugte Folge (mindestens) superlinear konvergiert.

Wir zeigen zuerst eine nützliche Eigenschaft der superlinearen Konvergenz.

Lemma 7.2. *Konvergiert die Folge $\{x^k\}_{k\in\mathbb{N}} \subset \mathbb{R}^n$ superlinear gegen $\bar{x} \in \mathbb{R}^n$ mit $x^k \neq \bar{x}$ für alle $k \in \mathbb{N}$, so gilt*

$$\lim_{k\to\infty} \frac{\|x^{k+1} - x^k\|}{\|x^k - \bar{x}\|} = 1.$$

*Beweis.* Mit Hilfe der umgekehrten Dreiecksungleichung $|\|x\| - \|y\|| \leq \|x - y\|$ und der Definition der superlinearen Konvergenz folgt sofort

$$\begin{aligned} 0 &\leq \lim_{k\to\infty} \left| \frac{\|x^{k+1} - x^k\|}{\|x^k - \bar{x}\|} - 1 \right| = \lim_{k\to\infty} \left| \frac{\|x^{k+1} - x^k\| - \|x^k - \bar{x}\|}{\|x^k - \bar{x}\|} \right| \\ &\leq \lim_{k\to\infty} \frac{\|x^{k+1} - \bar{x}\|}{\|x^k - \bar{x}\|} = 0. \end{aligned}$$

□

Der Nutzen dieser Eigenschaft liegt – neben der Verwendung in den folgenden Beweisen – darin, dass wir (nur!) für superlinear konvergente Folgen das "optimale" Gütekriterium $\|x^k - \bar{x}\| \leq \varepsilon$ durch das tatsächlich überprüfbare Kriterium $\|x^{k+1} - x^k\| \leq \tilde{\varepsilon}$ (für $\tilde{\varepsilon}$ hinreichend klein) ersetzen können.

Einen ähnlichen Nutzen hat das folgende Lemma.

Lemma 7.3. *Sei $f : \mathbb{R}^n \to \mathbb{R}$ zweimal stetig differenzierbar mit $\nabla f(\bar{x}) = 0$ und $\nabla^2 f(\bar{x})$ invertierbar in $\bar{x} \in \mathbb{R}$, und sei $\{x^k\}_{k\in\mathbb{N}} \subset \mathbb{R}^n$ eine Folge mit $x^k \to \bar{x}$. Dann existiert ein $k_0 \in \mathbb{N}$ und eine Konstante $\beta > 0$ so, dass gilt*

$$\|\nabla f(x^k)\| \geq \beta \|x^k - \bar{x}\| \qquad \text{für alle } k \geq k_0.$$

*Beweis.* Die Differenzierbarkeit von $\nabla f$ in $\bar{x}$ bedeutet nach Definition, dass zu jedem $\varepsilon > 0$ ein $k_0 \in \mathbb{N}$ existiert mit

$$\|\nabla f(x^k) - \nabla f(\bar{x}) - \nabla^2 f(\bar{x})(x^k - \bar{x})\| \leq \varepsilon \|x^k - \bar{x}\| \qquad \text{für alle } k \geq k_0.$$

Sei nun $\varepsilon < \|\nabla^2 f(\bar{x})^{-1}\|^{-1}$ gewählt. Dann folgt für alle $k \geq k_0(\varepsilon)$ mit Hilfe von $\nabla f(\bar{x}) = 0$, der umgekehrten Dreiecksungleichung, und der Ungleichung $\|A\|^{-1}\|Ay\| \leq \|y\|$ für $A = \nabla^2 f(\bar{x})^{-1}$ und $y = \nabla^2 f(\bar{x})(x^k - \bar{x})$

$$\begin{aligned}
\|\nabla f(x^k)\| &\geq \Big| \|\nabla^2 f(\bar{x})(x^k - \bar{x})\| - \|\nabla f(x^k) - \nabla f(\bar{x}) - \nabla^2 f(\bar{x})(x^k - \bar{x})\| \Big| \\
&\geq \Big| \|\nabla^2 f(\bar{x})^{-1}\|^{-1} \|x^k - \bar{x}\| - \varepsilon \|x^k - \bar{x}\| \Big| \\
&= \beta \|x^k - \bar{x}\|
\end{aligned}$$

mit $\beta := \|\nabla^2 f(\bar{x})^{-1}\|^{-1} - \varepsilon > 0$. □

Dies rechtfertigt, als Abbruchkriterium für Algorithmen die Bedingung $\|\nabla f(x^k)\| \leq \varepsilon$ anstelle von $\nabla f(x^k) = 0$ einzusetzen, denn für hinreichend kleines $\varepsilon > 0$ ist (unter den genannten Voraussetzungen!) $x^k$ bereits eine gute Näherung an $\bar{x}$.

Die nächsten Hilfssätze betreffen die Invertierbarkeit der Hesse-Matrix unter kleinen Störungen. Der Beweis beruht auf einem fundamentalen Störungslemma für Matrizen.[1]

Lemma 7.4 (Banach-Lemma). *Seien $A, B \in \mathbb{R}^{n\times n}$ mit $\|I - BA\| < 1$. Dann sind $A$ und $B$ invertierbar, und es gilt*

$$\|A^{-1}\| \leq \frac{\|B\|}{1 - \|I - BA\|}.$$

*Eine analoge Abschätzung gilt für $B^{-1}$.*

Lemma 7.5. *Sei $f : \mathbb{R}^n \to \mathbb{R}$ zweimal stetig differenzierbar und sei $\bar{x} \in \mathbb{R}^n$ mit $\nabla^2 f(\bar{x})$ invertierbar. Dann existieren Konstanten $\delta > 0$ und $c > 0$, so dass gilt*

$$\|\nabla^2 f(x)^{-1}\| \leq c \qquad \text{für alle } x \in B_\delta(\bar{x}).$$

*Insbesondere ist $\nabla^2 f(x)$ invertierbar für alle $x \in B_\delta(\bar{x})$.*

*Beweis.* Da $\nabla^2 f$ stetig und $\nabla^2 f(\bar{x})$ invertierbar ist, existiert für $\varepsilon := \frac{1}{2}\|\nabla^2 f(\bar{x})^{-1}\|^{-1} > 0$ ein $\delta > 0$ mit

$$\|\nabla^2 f(\bar{x}) - \nabla^2 f(x)\| \leq \frac{1}{2}\|\nabla^2 f(\bar{x})^{-1}\|^{-1} \quad \text{für alle } x \in B_\delta(\bar{x}).$$

[1] siehe z. B. [Geiger & Kanzow 1999, Lemma B.8]

Also gilt für alle $x \in B_\delta(\bar{x})$ die Abschätzung

$$\|I - \nabla^2 f(\bar{x})^{-1}\nabla^2 f(x)\| \leq \|\nabla^2 f(\bar{x})^{-1}\|\|\nabla^2 f(\bar{x}) - \nabla^2 f(x)\| \leq \frac{1}{2} < 1.$$

Nach dem Banach-Lemma ist daher auch $\nabla^2 f(x)$ invertierbar für alle $x \in B_\delta(\bar{x})$, und es gilt

$$\|\nabla^2 f(x)^{-1}\| \leq \frac{\|\nabla^2 f(\bar{x})^{-1}\|}{1 - \|I - \nabla^2 f(\bar{x})^{-1}\nabla^2 f(x)\|} \leq 2\|\nabla^2 f(\bar{x})^{-1}\| =: c. \qquad \square$$

Eine ähnliche Aussage gilt für die positive Definitheit.

Lemma 7.6. *Sei $f : \mathbb{R}^n \to \mathbb{R}$ zweimal stetig differenzierbar und sei $\bar{x} \in \mathbb{R}^n$ mit $\nabla^2 f(\bar{x})$ positiv definit. Dann existieren Konstanten $\delta > 0$ und $\mu > 0$, so dass gilt*

$$d^T \nabla^2 f(x) d \geq \mu \|d\|^2 \qquad \text{für alle } x \in B_\delta(\bar{x}), d \in \mathbb{R}^n.$$

*Beweis.* Angenommen, die Ungleichung gelte nicht. Dann existieren Folgen $\{x^k\}_{k\in\mathbb{N}} \subset \mathbb{R}^n$ und $\{d^k\}_{k\in\mathbb{N}} \subset \mathbb{R}^n$ mit $x^k \to \bar{x}$ sowie

$$(d^k)^T \nabla^2 f(x^k) d^k < \frac{1}{k}\|d^k\|^2 \qquad \text{für alle } k \in \mathbb{N}, \tag{7.1}$$

wobei wir ohne Einschränkung $\|d^k\| = 1$ annehmen können. Die Folge $\{d^k\}_{k\in\mathbb{N}}$ ist also beschränkt und enthält daher eine konvergente Teilfolge, deren Grenzwert $\bar{d}$ ebenfalls $\|\bar{d}\| = 1$ erfüllt. Da $\nabla^2 f$ stetig ist, können wir in (7.1) zum Grenzwert dieser Teilfolge übergehen und erhalten

$$\bar{d}^T \nabla^2 f(\bar{x}) \bar{d} \leq 0 \qquad \text{für } \bar{d} \neq 0.$$

Also kann $\nabla^2 f(\bar{x})$ nicht positiv definit sein, und Kontraposition ergibt die Aussage. $\square$

Wir benötigen noch die folgenden technischen Hilfssätze.

Lemma 7.7. *Sei $f : \mathbb{R}^n \to \mathbb{R}$ zweimal stetig differenzierbar, und sei $\{x^k\}_{k\in\mathbb{N}} \subset \mathbb{R}^n$ eine Folge mit $x^k \to \bar{x}$ für $k \to \infty$. Dann gilt*

$$\lim_{k\to\infty} \int_0^1 \|\nabla^2 f(x^k + t(x^{k+1} - x^k)) - \nabla^2 f(\bar{x})\|\, dt = 0 \tag{7.2}$$

*sowie*

$$\lim_{k\to\infty} \int_0^1 \|\nabla^2 f(\bar{x} + t(x^k - \bar{x})) - \nabla^2 f(\bar{x})\|\, dt = 0. \tag{7.3}$$

*Beweis.* Aus der Konvergenz $x^k \to \bar{x}$ folgt wegen der Kompaktheit von $[0,1]$ sofort die *gleichmäßige* Konvergenz

$$x^k + t(x^{k+1} - x^k) \to \bar{x} \qquad \text{für alle } t \in [0,1].$$

Wegen der Stetigkeit von $\nabla^2 f$ existiert daher für alle $\varepsilon > 0$ ein $k_0 \in \mathbb{N}$ mit

$$\|\nabla^2 f(x^k + t(x^{k+1} - x^k)) - \nabla^2 f(\bar{x})\| \leq \varepsilon \qquad \text{für alle } k \geq k_0, t \in [0,1].$$

Damit ist auch

$$\int_0^1 \|\nabla^2 f(x^k + t(x^{k+1} - x^k)) - \nabla^2 f(\bar{x})\| \, dt \leq \int_0^1 \varepsilon \, dt = \varepsilon$$

für alle $k \geq k_0$. Da $\varepsilon > 0$ beliebig war, folgt daraus (7.2). Analog beweist man (7.3). □

Lemma 7.8. *Sei $f : \mathbb{R}^n \to \mathbb{R}$ zweimal stetig differenzierbar, und sei $\{x^k\}_{k\in\mathbb{N}} \subset \mathbb{R}^n$ eine Folge mit $x^k \to \bar{x}$ für $k \to \infty$. Dann gilt*

$$\|\nabla f(x^k) - \nabla f(\bar{x}) - \nabla^2 f(x^k)(x^k - \bar{x})\| = o(\|x^k - \bar{x}\|). \tag{7.4}$$

*Ist zusätzlich $\nabla^2 f$ lokal Lipschitz-stetig, so gilt*

$$\|\nabla f(x^k) - \nabla f(\bar{x}) - \nabla^2 f(x^k)(x^k - \bar{x})\| = O(\|x^k - \bar{x}\|^2). \tag{7.5}$$

*Beweis.* Nach Voraussetzung ist $\nabla f$ differenzierbar in $\bar{x}$, so dass nach Definition eine Nullfolge $\{\varepsilon_k^1\}_{k\in\mathbb{N}}$ existiert mit

$$\|\nabla f(x^k) - \nabla f(\bar{x}) - \nabla^2 f(\bar{x})(x^k - \bar{x})\| \leq \varepsilon_k^1 \|x^k - \bar{x}\|$$

für alle $k \in \mathbb{N}$ groß genug. Ebenso folgt aus der Stetigkeit von $\nabla^2 f$ in $\bar{x}$, dass gilt

$$\varepsilon_k^2 := \|\nabla^2 f(x^k) - \nabla^2 f(\bar{x})\| \to 0$$

für $k \to \infty$. Zusammen folgt

$$\begin{aligned}\|\nabla f(x^k) - \nabla f(\bar{x}) - \nabla^2 f(x^k)(x^k - \bar{x})\| &\leq \|\nabla f(x^k) - \nabla f(\bar{x}) - \nabla^2 f(\bar{x})(x^k - \bar{x})\| \\ &\quad + \|\nabla^2 f(x^k) - \nabla^2 f(\bar{x})\| \|x^k - \bar{x}\| \\ &\leq (\varepsilon_k^1 + \varepsilon_k^2)\|x^k - \bar{x}\|\end{aligned}$$

für alle $k \in \mathbb{N}$ hinreichend groß. Da $\varepsilon_k := \varepsilon_k^1 + \varepsilon_k^2$ ebenfalls eine Nullfolge bildet, erhalten wir (7.4).

Aus Satz 1.3 folgt

$$\begin{aligned}\nabla f(x^k) - \nabla f(\bar{x}) - \nabla^2 f(x^k)(x^k - \bar{x}) &= \int_0^1 \nabla^2 f(\bar{x} + t(x^k - \bar{x}))(x^k - \bar{x})\, dt \\ &\quad - \nabla^2 f(x^k)(x^k - \bar{x}) \\ &= \int_0^1 \left[\nabla^2 f(\bar{x} + t(x^k - \bar{x})) - \nabla^2 f(x^k)\right](x^k - \bar{x})\, dt.\end{aligned}$$

Mit der lokalen Lipschitz-Konstante $L > 0$ von $\nabla^2 f$ in einer Umgebung von $\bar{x}$ und der gleichmäßigen Konvergenz $\bar{x} + t(x^k - \bar{x}) \to \bar{x}$ für $k \to \infty$ und $t \in [0,1]$ folgt, dass für alle $k \in \mathbb{N}$ groß genug gilt

$$\begin{aligned}\|\nabla f(x^k) - \nabla f(\bar{x}) - \nabla^2 f(x^k)(x^k - \bar{x})\| &\le \int_0^1 \|\nabla^2 f(\bar{x} + t(x^k - \bar{x})) - \nabla^2 f(x^k)\|\|x^k - \bar{x}\|\, dt \\ &\le \int_0^1 L(1-t)\|x^k - \bar{x}\|\, dt \cdot \|x^k - \bar{x}\| \\ &= \frac{L}{2}\|x^k - \bar{x}\|^2,\end{aligned}$$

woraus (7.5) folgt. □

Wir kommen nun zur versprochenen Bedingung für die superlineare Konvergenz von Algorithmus 7.1.

Satz 7.9. *Sei $f : \mathbb{R}^n \to \mathbb{R}$ zweimal stetig differenzierbar mit $\nabla^2 f(\bar{x})$ invertierbar in $\bar{x} \in \mathbb{R}^N$, und sei $\{x^k\}_{k\in\mathbb{N}} \subset \mathbb{R}^n \setminus \{\bar{x}\}$ eine Folge mit $x^k \to \bar{x}$ für $k \to \infty$. Dann sind äquivalent:*

*(i)* $\{x^k\}_{k\in\mathbb{N}}$ *konvergiert superlinear gegen* $\bar{x}$ *und* $\nabla f(\bar{x}) = 0$,

*(ii)* $\|\nabla f(x^k) + \nabla^2 f(x^k)(x^{k+1} - x^k)\| = o(\|x^{k+1} - x^k\|)$,

*(iii)* $\|\nabla f(x^k) + \nabla^2 f(\bar{x})(x^{k+1} - x^k)\| = o(\|x^{k+1} - x^k\|)$.

*Beweis. (iii)⇒(i):* Aus der produktiven Null zusammen mit Satz 1.3 folgt zunächst die Identität

$$\begin{aligned}\nabla f(x^{k+1}) &= \nabla f(x^{k+1}) - \nabla f(x^k) - \nabla^2 f(\bar{x})(x^{k+1} - x^k) \\ &\quad + \nabla f(x^k) + \nabla^2 f(\bar{x})(x^{k+1} - x^k) \\ &= \int_0^1 \left[\nabla^2 f(x^k + t(x^{k+1} - x^k)) - \nabla^2 f(\bar{x})\right](x^{k+1} - x^k)\, dt \\ &\quad + \nabla f(x^k) + \nabla^2 f(\bar{x})(x^{k+1} - x^k)\end{aligned} \tag{7.6}$$

und daraus die Abschätzung

$$\begin{aligned}\|\nabla f(x^{k+1})\| \le \int_0^1 \|\nabla^2 f(x^k + t(x^{k+1} - x^k)) - \nabla^2 f(\bar{x})\|\, dt \cdot \|x^{k+1} - x^k\| \\ + \|\nabla f(x^k) + \nabla^2 f(\bar{x})(x^{k+1} - x^k)\|.\end{aligned}$$

Nach Lemma 7.7 und Voraussetzung (iii) existiert also eine Nullfolge $\{\varepsilon_k\}_{k\in\mathbb{N}}$ mit

$$\|\nabla f(x^{k+1})\| \le \varepsilon_k \|x^{k+1} - x^k\|. \tag{7.7}$$

Daraus folgt $\nabla f(x^k) \to 0$ und somit, wegen der Stetigkeit von $\nabla f$, auch $\nabla f(\bar{x}) = 0$. Nach Lemma 7.3 existiert daher ein $\beta > 0$ mit

$$\|\nabla f(x^{k+1})\| \ge \beta \|x^{k+1} - \bar{x}\| \qquad \text{für alle } k \in \mathbb{N} \text{ groß genug.}$$

Zusammen mit (7.7) erhalten wir daraus

$$\beta \|x^{k+1} - \bar{x}\| \le \varepsilon_k \|x^{k+1} - x^k\| \le \varepsilon_k (\|x^{k+1} - \bar{x}\| + \|x^k - \bar{x}\|)$$

und daher

$$\|x^{k+1} - \bar{x}\| \le \frac{\varepsilon_k}{\beta - \varepsilon_k} \|x^k - \bar{x}\| \qquad \text{für alle } k \in \mathbb{N} \text{ groß genug,}$$

d. h. die superlineare Konvergenz von $\{x^k\}_{k\in\mathbb{N}}$.

*(i)⇒(iii):* Aus der Identität (7.6) folgt auch die Abschätzung

$$\begin{aligned}&\|\nabla f(x^k) + \nabla^2 f(\bar{x})(x^{k+1} - x^k)\| \\ &\qquad \le \|\nabla f(x^{k+1})\| + \int_0^1 \|\nabla^2 f(x^k + t(x^{k+1} - x^k)) - \nabla^2 f(\bar{x})\|\, dt \cdot \|x^{k+1} - x^k\|.\end{aligned} \tag{7.8}$$

Das Integral bildet nach Lemma 7.7 ein Nullfolge; wir müssen daher nur noch den ersten Term geeignet abschätzen. Da $\nabla^2 f(\bar{x})$ invertierbar ist, existiert nach Lemma 7.5 ein $\varepsilon > 0$ mit $\|\nabla^2 f(x)\| > 0$ für alle $x \in B_\varepsilon(\bar{x})$. Die stetige Funktion $x \mapsto \|\nabla^2 f(x)\|$ nimmt daher auf der kompakten Menge $\overline{B_\varepsilon(\bar{x})}$ ihr Maximum $M > 0$ an, und mit Satz 1.3 und $x^k \to \bar{x}$ folgt für alle $k \in \mathbb{N}$ groß genug

$$\begin{aligned}\|\nabla f(x^{k+1})\| = \|\nabla f(x^{k+1}) - \nabla f(\bar{x})\| &\le \int_0^1 \|\nabla^2 f(\bar{x} + t(x^{k+1} - \bar{x}))\|\, dt \cdot \|x^{k+1} - \bar{x}\| \\ &\le M \|x^{k+1} - \bar{x}\| \\ &\le \left( M \frac{\|x^{k+1} - \bar{x}\|}{\|x^k - \bar{x}\|} \cdot \frac{\|x^k - \bar{x}\|}{\|x^{k+1} - x^k\|} \right) \|x^{k+1} - x^k\|.\end{aligned}$$

Da $x^k \to \bar{x}$ superlinear konvergiert, geht der erste Bruch nach Definition gegen 0 und der zweite Bruch nach Lemma 7.2 gegen 1. Also ist der gesamte Term in Klammern eine Nullfolge, und aus (7.8) folgt (iii).

*(ii)⇒(iii):* Aus (ii) folgt mit der Dreiecksungleichung sofort

$$\begin{aligned}\|\nabla f(x^k) + \nabla^2 f(\bar{x})(x^{k+1} - x^k)\| &\leq \|\nabla f(x^k) + \nabla^2 f(x^k)(x^{k+1} - x^k)\| \\ &\quad + \|\nabla^2 f(x^k) - \nabla^2 f(\bar{x})\|\|x^{k+1} - \bar{x}\| \\ &\leq (\varepsilon_k + \|\nabla^2 f(x^k) - \nabla^2 f(\bar{x})\|)\|x^{k+1} - \bar{x}\|.\end{aligned}$$

Wegen der Stetigkeit von $\nabla^2 f$ und $x^k \to \bar{x}$ ist auch der zweite Term in der Klammer eine Nullfolge, und wir erhalten (iii). Analog zeigt man die Implikation *(iii)⇒ (ii).* □

Daraus erhält man durch Einsetzen die sogenannten *Dennis–Moré-Bedingungen* für die superlineare Konvergenz von Algorithmus 7.1.

Folgerung 7.10 (Dennis–Moré-Bedingungen). *Sei $f : \mathbb{R}^n \to \mathbb{R}$ zweimal stetig differenzierbar und sei $\{x^k\}_{k\in\mathbb{N}}$ eine durch Algorithmus 4.1 erzeugte Folge, die gegen ein $\bar{x} \in \mathbb{R}^n$ mit $\nabla^2 f(\bar{x})$ invertierbar und $x^k \neq \bar{x}$ für alle $k \in \mathbb{N}$ konvergiert. Dann sind äquivalent:*

*(i)* $\{x^k\}_{k\in\mathbb{N}}$ *konvergiert superlinear gegen* $\bar{x}$ *und* $\nabla f(\bar{x}) = 0$,

*(ii)* $\|(H_k - \nabla^2 f(x^k))(x^{k+1} - x^k)\| = o(\|x^{k+1} - x^k\|)$,

*(iii)* $\|(H_k - \nabla^2 f(\bar{x}))(x^{k+1} - x^k)\| = o(\|x^{k+1} - x^k\|)$.

*Beweis.* Nach Iterationsvorschrift gilt

$$\nabla f(x^k) = -H_k(x^{k+1} - x^k) \qquad \text{für alle } k \in \mathbb{N},$$

und Einsetzen in Satz 7.9 ergibt die Behauptung. □

Für die superlineare Konvergenz muss also $H_k$ für $k \to \infty$ hinreichend gut die Hesse-Matrix $\nabla^2 f(x^k)$ annähern, und dafür reicht es aus, dass die Anwendung auf die jeweilige Suchrichtung hinreichend gut übereinstimmt.

Unter etwas stärkeren Bedingungen an $f$ erhält man sogar quadratische Konvergenz.

Satz 7.11. *Sei $f : \mathbb{R}^n \to \mathbb{R}$ zweimal stetig differenzierbar mit $\nabla^2 f(\bar{x})$ invertierbar in $\bar{x} \in \mathbb{R}$, und sei $\{x^k\}_{k\in\mathbb{N}} \subset \mathbb{R}^n \setminus \{\bar{x}\}$ eine Folge mit $x^k \to \bar{x}$ für $k \to \infty$. Ist $\nabla^2 f$ darüber hinaus lokal Lipschitz-stetig, dann sind äquivalent:*

*(i)* $\{x^k\}_{k\in\mathbb{N}}$ *konvergiert quadratisch gegen* $\bar{x}$ *und* $\nabla f(\bar{x}) = 0$,

*(ii)* $\|\nabla f(x^k) + \nabla^2 f(x^k)(x^{k+1} - x^k)\| = O(\|x^{k+1} - x^k\|^2)$,

*(iii)* $\|\nabla f(x^k) + \nabla^2 f(\bar{x})(x^{k+1} - x^k)\| = O(\|x^{k+1} - x^k\|^2)$.

*Beweis.* *(ii)⇒(i):* Die quadratische Konvergenz impliziert insbesondere die superlineare, und daher folgt aus Satz 7.9 sowohl $\nabla f(\bar{x}) = 0$ als auch $x^k \to \bar{x}$ superlinear. Es bleibt also nur noch die quadratische Konvergenz zu zeigen. Dafür verwenden analog zu (7.6) die Identität

$$\begin{aligned}\nabla^2 f(x^k)(x^{k+1} - \bar{x}) = \nabla f(x^k) + \nabla^2 f(x^k)(x^{k+1} - x^k) \\ - \nabla f(x^k) + \nabla f(\bar{x}) + \nabla^2 f(x^k)(x^k - \bar{x}).\end{aligned} \tag{7.9}$$

Da $\nabla^2 f(\bar{x})$ invertierbar ist und $x^k \to \bar{x}$ konvergiert, existiert nach Lemma 7.5 ein $c > 0$ mit

$$\|x^{k+1} - \bar{x}\| \leq \|\nabla^2 f(x^k)^{-1}\| \|\nabla^2 f(x^k)(x^{k+1} - \bar{x})\| \leq c \|\nabla^2 f(x^k)(x^{k+1} - \bar{x})\|$$

für alle $k \in \mathbb{N}$ groß genug. Dividiert man nun durch $\|x^k - \bar{x}\|^2 \neq 0$ und schätzt die rechte Seite durch (7.9) ab, so folgt damit

$$\begin{aligned}\frac{\|x^{k+1} - \bar{x}\|}{\|x^k - \bar{x}\|^2} \leq c \Bigg( \frac{\|\nabla f(x^k) + \nabla^2 f(x^k)(x^{k+1} - x^k)\|}{\|x^{k+1} - x^k\|^2} \cdot \frac{\|x^{k+1} - x^k\|^2}{\|x^k - \bar{x}\|^2} \\ + \frac{\|\nabla f(x^k) - \nabla f(\bar{x}) - \nabla^2 f(x^k)(x^k - \bar{x})\|}{\|x^k - \bar{x}\|^2} \Bigg).\end{aligned}$$

Nun ist der erste Summand auf der rechten Seite beschränkt nach Voraussetzung (ii) und Lemma 7.2 (konvergente Folgen sind beschränkt), der zweite wegen Lemma 7.8 und der lokalen Lipschitz-Stetigkeit von $\nabla^2 f$. Die Folge $\{x^k\}_{k \in \mathbb{N}}$ konvergiert also nach Definition quadratisch gegen $\bar{x}$.

*(i)⇒(ii):* Aus der Identität (7.9) folgt auch die Abschätzung

$$\begin{aligned}\frac{\|\nabla f(x^k) + \nabla^2 f(x^k)(x^{k+1} - x^k)\|}{\|x^k - \bar{x}\|^2} \leq \frac{\|\nabla f(x^k) - \nabla f(\bar{x}) - \nabla^2 f(x^k)(x^k - \bar{x})\|}{\|x^k - \bar{x}\|^2} \\ + \|\nabla^2 f(x^k)\| \frac{\|x^{k+1} - \bar{x}\|}{\|x^k - \bar{x}\|^2}.\end{aligned}$$

Der erste Summand ist wieder beschränkt nach Lemma 7.8. Aus der Stetigkeit von $\nabla^2 f$ und der Konvergenz $x^k \to \bar{x}$ folgt weiter, dass $\{\|\nabla^2 f(x^k)\|\}_{k \in \mathbb{N}}$ beschränkt ist. Also ist auch der zweite Summand wegen der quadratischen Konvergenz $x^k \to \bar{x}$ beschränkt, und zusammen folgt (ii).

*(ii)⇔(iii)* zeigt man analog zu Satz 7.9 unter Verwendung von Lemma 7.2. □

Analog zu Folgerung 7.10 folgen daraus Dennis–Moré-Bedingungen für die quadratische Konvergenz von Newton-artigen Verfahren.

# 8 NEWTON-VERFAHREN

Folgerung 7.10 legt die Wahl $H_k = \nabla^2 f(x^k)$ nahe; dies führt auf das bekannte Newton-Verfahren für die Lösung des nichtlinearen Gleichungssystems $\nabla f(x) = 0$.

## 8.1 LOKALES NEWTON-VERFAHREN

Wir kommen schnell zur Sache, denn wir sind gut vorbereitet. Algorithmus 7.1 hat nun die Form

**Algorithmus 8.1 :** Lokales Newton-Verfahren

**Input :** $x^0 \in \mathbb{R}^n$

1 Setze $k = 0$
2 **while** $\|\nabla f(x^k)\| > 0$ **do**
3  Berechne $s^k$ als Lösung von $\nabla^2 f(x^k) s^k = -\nabla f(x^k)$
4  Setze $x^{k+1} = x^k + s^k, \quad k \leftarrow k+1$

Wegen Folgerung 7.10 ist nur noch zu beweisen, dass dieses Verfahren durchführbar ist (d. h. dass $\nabla^2 f(x^k)$ stets invertierbar ist) und dass die Iteration überhaupt konvergiert. Der folgende Beweis ist der Prototyp eines Konvergenzbeweises für Newton-artige Verfahren.

Satz 8.1. *Sei $f : \mathbb{R}^n \to \mathbb{R}$ zweimal stetig differenzierbar und sei $\bar{x} \in \mathbb{R}^n$ ein stationärer Punkt von $f$ mit $\nabla^2 f(\bar{x})$ invertierbar. Dann existiert ein $\varepsilon > 0$, so dass Algorithmus 8.1 für alle Startwerte $x^0 \in B_\varepsilon(\bar{x})$ superlinear gegen $\bar{x}$ konvergiert. Ist $\nabla^2 f$ darüber hinaus lokal Lipschitz-stetig, so ist die Konvergenz sogar quadratisch.*

*Beweis.* Wir beginnen mit der Durchführbarkeit des Verfahrens. Nach Lemma 7.5 existiert ein Radius $\varepsilon_1 > 0$ und eine Konstante $c > 0$ mit

$$\|\nabla^2 f(x)^{-1}\| \leq c \qquad \text{für alle } x \in B_{\varepsilon_1}(\bar{x}).$$

Die Hesse-Matrix ist also in einer Umgebung von $\bar{x}$ invertierbar. Wir müssen nun garantieren, dass die Iterierten $x^k$ diese Umgebung nicht verlassen. Dafür verwenden wir Lemma 7.8, welches einen Radius $\varepsilon_2 > 0$ liefert, so dass (mit $\nabla f(\bar{x}) = 0$) gilt

$$\|\nabla f(x) - \nabla^2 f(x)(x - \bar{x})\| \leq \frac{1}{2c}\|x - \bar{x}\| \qquad \text{für alle } x \in B_{\varepsilon_2}(\bar{x}).$$

Setze nun $\varepsilon := \min\{\varepsilon_1, \varepsilon_2\}$ und wähle $x^0 \in B_\varepsilon(\bar{x})$. Dann ist $\nabla^2 f(x^0)$ invertierbar, und aus der Iterationsvorschrift, aufgelöst nach $x^1 = x^0 - \nabla^2 f(x^0)^{-1}\nabla f(x^0)$, folgt

$$\begin{aligned}\|x^1 - \bar{x}\| &= \|x^0 - \bar{x} - \nabla^2 f(x^0)^{-1}\nabla f(x^0)\| \\ &\leq \|\nabla^2 f(x^0)^{-1}\| \, \|\nabla^2 f(x^0)(x^0 - \bar{x}) - \nabla f(x^0)\| \\ &\leq c\,\frac{1}{2c}\|x^0 - \bar{x}\| = \frac{1}{2}\|x^0 - \bar{x}\| < \varepsilon.\end{aligned}$$

Durch Induktion folgt daraus

$$\|x^k - \bar{x}\| \leq \left(\frac{1}{2}\right)^k \|x^0 - \bar{x}\| < \varepsilon \qquad \text{für alle } k \in \mathbb{N}.$$

Also ist $x^k \in B_\varepsilon(\bar{x})$ für alle $k \in \mathbb{N}$ und damit $\nabla^2 f(x^k)$ invertierbar für alle $k \in \mathbb{N}$; außerdem folgt $x^k \to \bar{x}$ für $k \to \infty$. Da nach Iterationsvorschrift gilt

$$\nabla f(x^k) + \nabla^2 f(x^k)(x^k - x^{k+1}) = 0 \qquad \text{für alle } k \in \mathbb{N},$$

folgt die superlineare bzw. quadratische Konvergenz sowie $\nabla f(\bar{x}) = 0$ nun aus Satz 7.9 bzw. Satz 7.11. □

Das lokale Newton-Verfahren hat zwei entscheidende Nachteile: Es konvergiert nur *lokal*, d. h. falls der Startwert $x^0$ bereits hinreichend nahe an $\bar{x}$ liegt. Außerdem ist $\nabla^2 f(\bar{x})$ lediglich als regulär vorausgesetzt; das Newton-Verfahren kann daher genauso gerne gegen einen *Maximierer* konvergieren. Beide Probleme kann man durch Kombination mit einem Abstiegsverfahren behandeln; man spricht dabei von *Globalisierung*.

## 8.2 GLOBALISIERTES NEWTON-VERFAHREN

Die Idee ist, in einem Abstiegsverfahren solange Gradientenschritte zu machen, bis man nahe genug an einem stationären Punkt ist, um das Newton-Verfahren durchführen zu können. Durch eine Schrittweitensuche wird dabei garantiert, dass der Funktionswert stets abnimmt (und dadurch der Grenzwert kein Maximierer sein kann.) Dies führt auf den folgenden Algorithmus.

**Algorithmus 8.2 :** Globalisiertes Newton-Verfahren

**Input :** $\rho > 0$, $p > 2$, $\gamma \in (0, 1/2)$, $x^0 \in \mathbb{R}^n$

1 Setze $k = 0$
2 **while** $\|\nabla f(x^k)\| > 0$ **do**
3  Versuche, Newton-Schritt $d^k$ mit $\nabla^2 f(x^k)d^k = -\nabla f(x^k)$ zu berechnen
4  **if** $\nabla f(x^k)^T d^k \leq -\rho \|d^k\|^p$ **then**
5  Setze $s^k = d^k$
6  **else**
7  Setze $s^k = -\nabla f(x^k)$
8  Bestimme $\sigma_k > 0$ mit Algorithmus 5.2 für $\gamma \in (0, 1/2)$
9  Setze $x^{k+1} = x^k + \sigma_k s^k, \quad k \leftarrow k + 1$

Die Bedingung in Schritt 4 setzt dabei stillschweigend voraus, dass eine Lösung des Newton-Systems $\nabla^2 f(x^k)d^k = -\nabla f(x^k)$ gefunden wurde. Beachte auch die Einschränkung $\gamma < \frac{1}{2}$ gegenüber dem Gradientenverfahren; dies wird später wichtig sein, um superlineare Konvergenz zu erhalten.

Wir zeigen zuerst, dass Algorithmus 8.2 tatsächlich ein konvergentes Abstiegsverfahren ist.

Satz 8.2. *Sei $f : \mathbb{R}^n \to \mathbb{R}$ zweimal stetig differenzierbar. Dann bricht Algorithmus 8.2 entweder nach endlich vielen Schritten ab, oder jeder Häufungspunkt von $\{x^k\}_{k\in\mathbb{N}}$ ist ein stationärer Punkt von $f$.*

*Beweis.* Im Falle eines endlichen Abbruchs ist nichts zu zeigen; sei daher $\nabla f(x^k) \neq 0$ für alle $k \in \mathbb{N}$ und sei $\bar{x}$ ein Häufungspunkt von $\{x^k\}_{k\in\mathbb{N}}$. Dann existiert eine gegen $\bar{x}$ konvergente Teilfolge $\{x^k\}_{k\in K}$ mit $K \subset \mathbb{N}$ unendlich. Um zu zeigen, dass $\bar{x}$ in stationärer Punkt ist, genügt nach Satz 4.3 der Nachweis, dass die erzeugten Suchrichtungen $\{s^k\}_{k\in K}$ und Schrittweiten $\{\sigma_k\}_{k\in K}$ zulässig sind. Wir unterscheiden dafür Gradienten- und Newtonschritte und setzen

$$K_G := \left\{k \in K : s^k = -\nabla f(x^k)\right\},$$
$$K_N := \left\{k \in K : s^k = -\nabla^2 f(x^k)^{-1}\nabla f(x^k)\right\}.$$

Wegen $\nabla f(x^k) \neq 0$ gilt dann

$$\frac{-\nabla f(x^k)^T s^k}{\|s^k\|} = \|\nabla f(x^k)\| > 0 \qquad \text{falls } k \in K_G, \tag{8.1}$$

und, wegen $s^k = -\nabla^2 f(x^k)^{-1}\nabla f(x^k) \neq 0$, auch

$$\frac{-\nabla f(x^k)^T s^k}{\|s^k\|} \geq \rho \|s^k\|^{p-1} > 0 \qquad \text{falls } k \in K_N. \tag{8.2}$$

In jedem Fall ist also $s^k$ eine Abstiegsrichtung. Es gelte nun

$$\frac{\nabla f(x^k)^T s^k}{\|s^k\|} \to 0 \qquad \text{für } K \ni k \to \infty. \tag{8.3}$$

Für $k \in K_G$ folgt dann aus (8.1)

$$\|\nabla f(x^k)\| = \frac{-\nabla f(x^k)^T s^k}{\|s^k\|} \to 0 \qquad \text{für } K_G \ni k \to \infty.$$

Für $k \in K_N$ verwenden wir, dass wegen der Beschränktheit der konvergenten Folge $\{x^k\}_{k\in K}$ und der Stetigkeit von $\nabla^2 f$ ein $C > 0$ existiert mit $\|\nabla^2 f(x^k)\| \le C$ für alle $k \in K$. Die Definition des Newton-Schritts ergibt nun

$$\|\nabla f(x^k)\| = \|\nabla^2 f(x^k) s^k\| \le C\|s^k\| \qquad \text{für alle } k \in K_N. \tag{8.4}$$

Also folgt wegen $p > 2$ aus (8.2) und (8.3)

$$\|\nabla f(x^k)\|^{p-1} \le \Big(C\|s^k\|\Big)^{p-1} \le \frac{C^{p-1}}{\rho} \frac{-\nabla f(x^k)^T s^k}{\|s^k\|} \to 0 \qquad \text{für } K_N \ni k \to \infty,$$

und damit ebenfalls $\|\nabla f(x^k)\| \to 0$. Damit sind die Suchrichtungen zulässig.

Nun zu den Schrittweiten. Für $k \in K_G$ folgt aus (8.1)

$$\|s^k\| = \|\nabla f(x^k)\| = \frac{-\nabla f(x^k)^T s^k}{\|s^k\|},$$

für $k \in K_N$ folgt aus (8.4) und der Cauchy–Schwarz-Ungleichung

$$\|s^k\| \ge \frac{1}{C}\|\nabla f(x^k)\| \ge \frac{1}{C} \frac{-\nabla f(x^k)^T s^k}{\|s^k\|}.$$

Also ist

$$\|s^k\| \ge \varphi\left(\frac{-\nabla f(x^k)^T s^k}{\|s^k\|}\right) \qquad \text{für alle } k \in K$$

für $\varphi : t \mapsto \min\{t, C^{-1}t\}$ stetig und streng monoton wachsend mit $\varphi(0) = 0$. Nach Satz 5.2 erzeugt die Armijo-Regel also zulässige Schrittweiten.

Die Behauptung folgt nun aus Satz 4.3. □

Natürlich soll auch das globalisierte Newton-Verfahren superlinear konvergieren, damit sich der ganze Mehraufwand gegenüber dem Gradientenverfahren lohnt. Wir wollen also, dass das globalisierte Newton-Verfahren in das superlinear konvergente lokale Newton-Verfahren übergeht, sobald wir entsprechend nahe an den stationären Punkt herangekommen sind. Dafür zeigen wir zuerst, dass unter geeigneten Voraussetzungen die ganze Folge $\{x^k\}_{k\in\mathbb{N}}$ gegen einen Minimierer konvergiert (und wir damit tatsächlich irgendwann nahe genug sind). Wesentliches Hilfsmittel wird das folgende nützliche Lemma sein, das wir auch im weiteren Verlauf der Vorlesung heranziehen werden.

Lemma 8.3. *Sei $\bar{x} \in \mathbb{R}^n$ ein isolierter Häufungspunkt der Folge $\{x^k\}_{k\in\mathbb{N}} \subset \mathbb{R}^n$ und es gelte $\|x^{k+1} - x^k\| \to 0$ für jede gegen $\bar{x}$ konvergente Teilfolge $\{x^k\}_{k\in K}$.[1] Dann konvergiert die gesamte Folge $\{x^k\}_{k\in\mathbb{N}}$ gegen $\bar{x}$.*

*Beweis.* Da $\bar{x} \in \mathbb{R}^n$ ein isolierter Häufungspunkt ist, existiert ein $\varepsilon > 0$ so, dass $\bar{x}$ der einzige Häufungspunkt von $\{x^k\}_{k\in\mathbb{N}}$ in $\overline{B_\varepsilon(\bar{x})}$ ist. Sei nun $\{x^k\}_{k\in K}$ eine Teilfolge mit $x^k \to \bar{x}$, und angenommen, $\{x^k\}_{k\in\mathbb{N}}$ konvergiert nicht gegen $\bar{x}$. Dann müssen unendlich viele Folgenglieder existieren mit $x^k \notin \overline{B_\varepsilon(\bar{x})}$. Wir können deshalb aus $\{x^k\}_{k\in\mathbb{N}}$ eine weitere Teilfolge $\{x^{l(k)}\}_{k\in K}$ durch Wahl von $l = l(k)$ auswählen, so dass für alle $k \in K$ gilt

$$x^{l(k)} \in \overline{B_\varepsilon(\bar{x})}, \qquad x^{l(k)+1} \notin \overline{B_\varepsilon(\bar{x})}$$

(d. h. $x^{l(k)+1}$ ist das *erste* Folgenglied von $\{x^k\}_{k\in\mathbb{N}}$, das wieder aus $\overline{B_\varepsilon(\bar{x})}$ herausspringt). Die Folge $\{x^{l(k)}\}_{k\in K}$ ist also beschränkt und besitzt daher (mindestens) einen Häufungspunkt; nach Annahme kommt dafür aber nur der Häufungspunkt $\bar{x}$ in Frage, weshalb $x^{l(k)} \to \bar{x}$ konvergieren muss. Es gibt also ein $k_0 \in K$ mit $\|x^{l(k)} - \bar{x}\| \le \frac{\varepsilon}{2}$ für alle $k \ge k_0$. Daraus folgt aber

$$\|x^{l(k)+1} - x^{l(k)}\| \ge \|x^{l(k)+1} - \bar{x}\| - \|x^{l(k)} - \bar{x}\| \ge \frac{\varepsilon}{2} \qquad \text{für alle } k \ge k_0.$$

Da $\{x^{l(k)}\}_{k\in K}$ als Teilfolge von $\{x^k\}_{k\in K}$ gewählt war, kann also $\|x^{k+1} - x^k\|$ nicht gegen 0 gehen. □

Daraus folgt die Konvergenz des globalisierten Newton-Verfahrens gegen Minimierer.

Satz 8.4. *Sei $f : \mathbb{R}^n \to \mathbb{R}$ zweimal stetig differenzierbar, und sei $\bar{x} \in \mathbb{R}^n$ ein Häufungspunkt der durch Algorithmus 8.2 erzeugten Folge $\{x^k\}_{k\in\mathbb{N}}$ mit $\nabla^2 f(\bar{x})$ positiv definit. Dann ist $\bar{x}$ ein strikter lokaler Minimierer und $\{x^k\}_{k\in\mathbb{N}}$ konvergiert gegen $\bar{x}$.*

*Beweis.* Nach Satz 8.2 ist jeder Häufungspunkt von $\{x^k\}_{k\in\mathbb{N}}$ ein stationärer Punkt und damit gilt $\nabla f(\bar{x}) = 0$. Zusammen mit der positiven Definitheit von $\nabla^2 f(\bar{x})$ folgt aus Satz 3.4, dass $\bar{x}$ ein strikter lokaler Minimierer ist.

Weiter ist $\nabla^2 f(\bar{x})$ insbesondere regulär; wegen Lemma 7.3 gilt daher $\nabla f(x^k) \neq 0$ für alle $x^k \neq \bar{x}$ hinreichend nahe bei $\bar{x}$. Also ist $\bar{x}$ ein isolierter Häufungspunkt (denn jeder weitere Häufungspunkt wäre nach Satz 8.2 wieder ein stationärer Punkt). Sei nun $\{x^k\}_{k\in K}$ eine Teilfolge mit $x^k \to \bar{x}$. Wir unterscheiden wieder Gradientenschritte (für $k \in K_G \subset K$) und Newton-Schritte (für $k \in K_N \subset K$). Für $k \in K_G$ gilt wegen der Stetigkeit von $\nabla f$ und $\sigma_k \in (0, 1]$ nach Definition der Armijo-Regel

$$\|x^{k+1} - x^k\| = \sigma_k \|\nabla f(x^k)\| \le \|\nabla f(x^k)\| \to \|\nabla f(\bar{x})\| = 0 \qquad \text{für } K_G \ni k \to \infty.$$

[1] Beachte, dass für $k \in K$ im Allgemeinen $k + 1 \notin K$ ist – in anderen Worten, $x^{k+1}$ ist das auf $x^k$ folgende Glied in der *gesamten* Folge, nicht der betrachteten konvergenten Teilfolge.

Für $k \in K_N$ hinreichend groß folgt dagegen aus Lemma 7.5

$$\begin{aligned}\|x^{k+1} - x^k\| &= \sigma_k \|\nabla^2 f(x^k)^{-1} \nabla f(x^k)\| \\ &\leq c\|\nabla f(x^k)\| \to c\|\nabla f(\bar{x})\| = 0 \qquad \text{für } K_N \ni k \to \infty.\end{aligned}$$

Damit ist Lemma 8.3 anwendbar und liefert die Aussage. □

Wir zeigen nun, dass Algorithmus 8.2 tatsächlich irgendwann in das lokale Newton-Verfahren übergeht. Ein wesentliches Hilfsresultat ist dabei, dass die Armijo-Regel für Newton-Schritte ab einem gewissen Schritt stets die Schrittweite $\sigma^k = 1$ akzeptiert. Dafür ist die Einschränkung $\gamma < \frac{1}{2}$ wesentlich.

Lemma 8.5. *Sei $f : \mathbb{R}^n \to \mathbb{R}$ zweimal stetig differenzierbar und $\bar{x} \in \mathbb{R}^n$ ein stationärer Punkt von $f$ mit $\nabla^2 f(\bar{x})$ positiv definit. Seien weiter $\{x^k\}_{k\in\mathbb{N}}$ eine Folge mit $x^k \to \bar{x}$ und $\{s^k\}_{k\in\mathbb{N}}$ gegeben durch*

$$s^k = -\nabla^2 f(x^k)^{-1} \nabla f(x^k).$$

*Dann gilt für $k \in \mathbb{N}$ hinreichend groß und $\gamma \in (0, \frac{1}{2})$ beliebig*

$$f(x^k + s^k) \leq f(x^k) + \gamma \nabla f(x^k)^T s^k.$$

*Beweis.* Wegen Lemmata 7.5 und 7.6 existieren $k_0 \in \mathbb{N}$, $c > 0$ und $\mu > 0$ so dass für alle $k \geq k_0$ gilt

$$\|\nabla^2 f(x^k)^{-1}\| \leq c \qquad \text{und} \qquad d^T \nabla^2 f(x^k) d \geq \mu \|d\|^2 \quad \text{für alle } d \in \mathbb{R}^n.$$

Weiter liefert Satz 1.2 ein $\xi^k = x^k + \theta_k s^k$, $\theta_k \in (0,1)$, mit

$$f(x^k + s^k) = f(x^k) + \nabla f(x^k)^T s^k + \frac{1}{2}(s^k)^T \nabla^2 f(\xi^k) s^k.$$

Daraus folgt unter Verwendung des Newton-Schritts $\nabla^2 f(x^k) s^k = -\nabla f(x^k)$ sowie der gleichmäßigen positiven Definitheit

$$\begin{aligned}f(x^k + s^k) - f(x^k) - \gamma \nabla f(x^k)^T s^k &= (1-\gamma)\nabla f(x^k)^T s^k + \frac{1}{2}(s^k)^T \nabla^2 f(\xi^k) s^k \\ &= -(1-\gamma)(s^k)^T \nabla^2 f(x^k) s^k + \frac{1}{2}(s^k)^T \nabla^2 f(\xi^k) s^k \\ &\leq -\left(\frac{1}{2} - \gamma\right) \mu \|s^k\|^2 + \frac{1}{2}\|\nabla^2 f(\xi^k) - \nabla^2 f(x^k)\| \|s^k\|^2\end{aligned}$$

für alle $k \geq k_0$.

Nun gilt für alle $k \geq k_0$

$$\|s^k\| = \|\nabla^2 f(x^k)^{-1} \nabla f(x^k)\| \leq c\|\nabla f(x^k)\| \to c\|\nabla f(\bar{x})\| = 0 \qquad \text{für } k \to \infty,$$

woraus $\xi^k = x^k + \theta_k s^k \to \bar{x}$ für $k \to \infty$ gleichmäßig in $\theta_k \in [0,1]$ folgt. Wegen der Stetigkeit von $\nabla^2 f$ können wir daher ein $k_1 \in \mathbb{N}$ finden mit

$$\|\nabla^2 f(x^k) - \nabla^2 f(\xi^k)\| \leq 2\left(\frac{1}{2} - \gamma\right)\mu \qquad \text{für alle } k \geq \max\{k_0, k_1\}.$$

(Die Klammer ist wegen $\gamma < \frac{1}{2}$ positiv!) Für $k \geq \max\{k_0, k_1\}$ ist daher

$$f(x^k + s^k) - f(x^k) - \gamma \nabla f(x^k)^T s^k \leq 0,$$

woraus die Aussage folgt. □

Damit haben wir nun alles zur Hand, um die superlineare Konvergenz des globalisierten Newton-Verfahrens zu zeigen. Hier wird nun die Wahl $p > 2$ wichtig.

Satz 8.6. *Sei $f : \mathbb{R}^n \to \mathbb{R}$ zweimal stetig differenzierbar, und sei $\bar{x} \in \mathbb{R}^n$ ein Häufungspunkt der durch Algorithmus 8.2 erzeugten Folge $\{x^k\}_{k\in\mathbb{N}}$ mit $\nabla^2 f(\bar{x})$ positiv definit. Dann ist $\bar{x}$ ein strikter lokaler Minimierer und $\{x^k\}_{k\in\mathbb{N}}$ konvergiert gegen $\bar{x}$ superlinear. Ist $\nabla^2 f$ darüber hinaus lokal Lipschitz-stetig, so ist die Konvergenz quadratisch.*

*Beweis.* Die Konvergenz der gesamten Folge gegen einen strikten lokalen Minimierer folgt aus Lemma 8.3. Für die superlineare bzw. quadratische Konvergenz ist nur noch zu zeigen, dass Algorithmus 8.2 irgendwann in das lokale Newton-Verfahren übergeht.

Die Voraussetzungen von Lemma 8.5 sind erfüllt, und wie in dessen Beweis gezeigt, existiert ein $k_0 \in \mathbb{N}$ mit

$$\|\nabla^2 f(x^k)^{-1}\| \leq c \qquad \text{und} \qquad d^T \nabla^2 f(x^k) d \geq \mu \|d\|^2 \quad \text{für alle } d \in \mathbb{R}^n$$

sowie (im Falle eines Newton-Schrittes)

$$\|s^k\| = \|\nabla^2 f(x^k)^{-1} \nabla f(x^k)\| \leq c\|\nabla f(x^k)\| \qquad \text{für alle } k \geq k_0.$$

Insbesondere ist $\nabla^2 f(x^k)$ invertierbar und es gilt $s^k \to 0$ für $k \to \infty$. Aus der Definition des Newton-Schrittes folgt nun

$$-\nabla f(x^k)^T s^k = (s^k)^T \nabla^2 f(x^k) s^k \geq \mu \|s^k\|^2.$$

Wegen $s^k \to 0$ existiert nun für beliebiges $\rho > 0$ und $p > 2$ ein $k_1 \geq k_0$ so dass gilt

$$\|s^k\| \leq \left(\frac{\mu}{\rho}\right)^{\frac{1}{p-2}} \qquad \text{für alle } k \geq k_1.$$

Also ist für alle $k \geq k_1$ insbesondere $-\mu \leq -\rho \|s^k\|^{p-2}$ und damit

$$\nabla f(x^k)^T s^k \leq -\mu \|s^k\|^2 \leq -\rho \|s^k\|^p,$$

d. h. der Newton-Schritt wird akzeptiert. Nach Lemma 8.5 wird für Newton-Schritte zudem irgendwann stets die Schrittweite $\sigma_k = 1$ akzeptiert. Ab diesem Punkt stimmt Algorithmus 8.2 mit Algorithmus 8.1 überein und hat damit die gleiche Konvergenzgeschwindigkeit. □

## 8.3 INEXAKTE NEWTON-VERFAHREN

Das Lösen der Newton-Gleichung $\nabla^2 f(x^k)s = -\nabla f(x^k)$ ist oft aufwendig, und eine exakte Lösung (z. B. aufgrund von Rundungsfehlern) in der Regel nicht möglich. Satz 7.9 garantiert aber die superlineare Konvergenz, solange wir nur bei Annäherung an einen stationären Punkt das Residuum in der Newton-Gleichung beliebig klein bekommen können; insbesondere ist zu Beginn der Iteration eine genaue Lösung gar nicht nötig. Die Idee ist dabei, den Fehler über das *relative Residuum* der Newton-Gleichung zu steuern: Für eine vorgegebene Toleranz $\eta_k$ bestimmen wir $s^k$ mit

$$\frac{\|\nabla^2 f(x^k)s^k + \nabla f(x^k)\|}{\|\nabla f(x^k)\|} \leq \eta_k \qquad \text{für alle } k \in \mathbb{N}.$$

Dies führt auf das *inexakte Newton-Verfahren.*

**Algorithmus 8.3 :** Inexaktes Newton-Verfahren

**Input :** $x^0 \in \mathbb{R}^n$, $\varepsilon > 0$

1 Setze $k = 0$
2 **while** $\|\nabla f(x^k)\| > 0$ **do**
3  Wähle Toleranz $\eta_k > 0$
4  Berechne $s^k$ mit $\|\nabla^2 f(x^k)s^k + \nabla f(x^k)\| \leq \eta_k \|\nabla f(x^k)\|$
5  Setze $x^{k+1} = x^k + s^k, \quad k \leftarrow k+1$

Da im Newton-Verfahren $x^k$ gegen einen stationären Punkt konvergiert, sollte der Fehler im Verlauf der Iteration automatisch kleiner werden. Beispielsweise kann man dafür die Newton-Gleichung mit einem iterativen Verfahren (Gauß–Seidel, konjugierte Gradienten) lösen, wobei in jeder Iteration mehr Schritte gemacht werden. Eine andere Möglichkeit ist, die Newton-Gleichung durch ein "grobes Modell" $[\nabla^2 f(x^k)]_{h_k} s = -[\nabla f(x^k)]_{h_k}$ zu ersetzen, dessen Lösung $s^k_{h_k}$ für $h_k \to 0$ gegen $s^k$ konvergiert. (Dies ist insbesondere für die Optimierung mit Differenzialgleichungen relevant.)

Die wesentliche Frage ist nun, wie die Toleranz $\eta_k$ zu wählen ist, um Konvergenz und insbesondere superlineare Konvergenz zu erhalten. Wir zeigen zuerst die lokale lineare Konvergenz. Eine Schwierigkeit ist dabei, dass diese nicht bezüglich der Euklidischen Norm gezeigt werden kann. Wir betrachten stattdessen für einen stationären Punkt $\bar{x} \in \mathbb{R}^n$ mit $\nabla^2 f(\bar{x})$ invertierbar die Konvergenz bezüglich

$$\|x\|_H := \|\nabla^2 f(\bar{x})x\| \qquad \text{für alle } x \in \mathbb{R}^n.$$

Dies definiert wegen

$$\|\nabla^2 f(\bar{x})\|^{-1}\|x\|_H \leq \|x\| \leq \|\nabla^2 f(\bar{x})^{-1}\|\|x\|_H \qquad \text{für alle } x \in \mathbb{R}^n \tag{8.5}$$

eine äquivalente Norm auf $\mathbb{R}^n$. Der Übersichtlichkeit halber setzen wir in Folge $\mu_1 := \|\nabla^2 f(\bar{x})\|^{-1}$ und $\mu_2 := \|\nabla^2 f(\bar{x})^{-1}\|$.

Satz 8.7. *Sei $f : \mathbb{R}^n \to \mathbb{R}$ zweimal stetig differenzierbar und sei $\bar{x} \in \mathbb{R}^n$ ein stationärer Punkt von $f$ mit $\nabla^2 f(\bar{x})$ invertierbar. Ist $\eta_k \leq \bar{\eta}$ für ein beliebiges $\bar{\eta} \in (0,1)$, so konvergiert Algorithmus 8.3 lokal linear bezüglich $\|\cdot\|_H$ gegen $\bar{x}$.*

*Beweis.* Wir gehen im Prinzip analog zum Beweis der entsprechenden Aussage in Satz 8.1 vor, wobei wir in den Abschätzungen wegen der Toleranz etwas genauer aufpassen müssen. Zunächst gilt wieder, dass nach Lemma 7.5 ein Radius $\varepsilon_1 > 0$ und eine Konstante $c > 0$ existieren mit

$$\|\nabla^2 f(x)^{-1}\| \leq c \qquad \text{für alle } x \in B_{\varepsilon_1}(\bar{x}).$$

Wähle nun ein $\eta \in (\bar{\eta}, 1)$ sowie ein hinreichend kleines $\delta > 0$ mit

$$\frac{c\delta}{\mu_1} + \bar{\eta}(1+\delta)(1+c\delta) \leq \eta.$$

(Dies ist wegen $\bar{\eta} < \eta$ stets möglich.) Nach Lemma 7.7 existiert weiter ein $\varepsilon_2 > 0$ mit

$$\int_0^1 \|\nabla^2 f(\bar{x} + t(x-\bar{x})) - \nabla^2 f(\bar{x})\|\, dt \leq \frac{\delta}{\mu_2} \qquad \text{für alle } x \in B_{\varepsilon_2}(\bar{x}).$$

Da $\bar{x}$ ein stationärer Punkt ist, ist nach dem Satz 1.3 für alle $x \in \mathbb{R}^n$

$$\begin{aligned}\nabla f(x) &= \nabla f(\bar{x}) + \int_0^1 \nabla^2 f(\bar{x} + t(x-\bar{x}))(x-\bar{x})\, dt \\ &= \int_0^1 \left[\nabla^2 f(\bar{x} + t(x-\bar{x})) - \nabla^2 f(\bar{x})\right](x-\bar{x})\, dt + \nabla^2 f(\bar{x})(x-\bar{x}).\end{aligned}$$

Nach Definition von $\|\cdot\|_H$ gilt daher für alle $x \in B_{\varepsilon_2}(\bar{x})$ die Abschätzung

$$\begin{aligned}\|\nabla f(x)\| &\leq \frac{\delta}{\mu_2}\|x-\bar{x}\| + \|\nabla^2 f(\bar{x})(x-\bar{x})\| \\ &\leq \frac{\delta}{\mu_2}(\mu_2\|x-\bar{x}\|_H) + \|x-\bar{x}\|_H = (1+\delta)\|x-\bar{x}\|_H.\end{aligned}$$

Nach Lemma 7.8 existiert außerdem ein $\varepsilon_3 > 0$ mit

$$\|\nabla f(x) - \nabla^2 f(x)(x-\bar{x})\| \leq \frac{\delta}{\mu_2}\|x-\bar{x}\| \leq \delta\|x-\bar{x}\|_H \qquad \text{für alle } x \in B_{\varepsilon_3}(\bar{x}).$$

Schließlich existiert wegen der Stetigkeit von $\nabla^2 f$ ein $\varepsilon_4 > 0$ mit

$$\|\nabla^2 f(x) - \nabla^2 f(\bar{x})\| \leq \delta \qquad \text{für alle } x \in B_{\varepsilon_4}(\bar{x}).$$

Setze nun $\varepsilon := \min\{\varepsilon_1, \varepsilon_2, \varepsilon_3, \varepsilon_4\}$ und wähle $x^0 \in B_{\varepsilon\frac{\mu_1}{\mu_2}}(\bar{x})$ (beachte $\mu_1 \leq \mu_2$ wegen (8.5)). Dann ist $\nabla^2 f(x^0)$ invertierbar, und wegen der exakten Lösbarkeit der Newton-Gleichung existiert daher ein $s^0 \in \mathbb{R}^n$ mit

$$\|r^0\| := \|\nabla^2 f(x^0)s^0 + \nabla f(x^0)\| \leq \eta_0\|\nabla f(x^0)\|.$$

Aus der Iterationsvorschrift folgt nun für $\eta_0 \leq \bar{\eta}$ mit Hilfe der obigen Abschätzungen

$$\begin{aligned}
\|x^1 - \bar{x}\|_H &= \|\nabla^2 f(\bar{x}) \left[x^0 - \bar{x} - \nabla^2 f(x^0)^{-1}(\nabla f(x^0) + r^0)\right] \| \\
&\leq \|\nabla^2 f(\bar{x})\| \, \|x^0 - \bar{x} - \nabla^2 f(x^0)^{-1} \nabla f(x^0)\| \\
&\quad + \|r^0 + \left[\nabla^2 f(\bar{x}) - \nabla^2 f(x^0)\right] \nabla^2 f(x^0)^{-1} r^0\| \\
&\leq \|\nabla^2 f(\bar{x})\| \, \|\nabla^2 f(x^0)^{-1}\| \, \|\nabla^2 f(x^0)(x^0 - \bar{x}) - \nabla f(x^0)\| \\
&\quad + \|r^0\| + \|\nabla^2 f(x^0) - \nabla^2 f(\bar{x})\| \, \|\nabla^2 f(x^0)^{-1}\| \, \|r^0\| \\
&\leq \frac{c\delta}{\mu_1} \|x^0 - \bar{x}\|_H + (1 + c\delta)\|r^0\| \\
&\leq \frac{c\delta}{\mu_1} \|x^0 - \bar{x}\|_H + \eta_0 (1 + c\delta)\|\nabla f(x^0)\| \\
&\leq \left(\frac{c\delta}{\mu_1} + \bar{\eta}(1+\delta)(1+c\delta)\right) \|x^0 - x^*\|_H \\
&\leq \eta \|x^0 - \bar{x}\|_H.
\end{aligned}$$

Nach Wahl von $x_0$ gilt daher wegen $\eta < 1$

$$\|x^1 - \bar{x}\| \leq \mu_2 \|x^1 - \bar{x}\|_H \leq \mu_2 \eta \|x^0 - \bar{x}\|_H \leq \eta \frac{\mu_2}{\mu_1} \|x^0 - \bar{x}\| < \varepsilon.$$

Mit Induktion folgt daraus $\|x^k - \bar{x}\|_H \leq \eta^k \|x^0 - \bar{x}\|_H$ und damit auch

$$\|x^k - \bar{x}\| \leq \eta^k \frac{\mu_2}{\mu_1} \|x^0 - \bar{x}\| < \varepsilon \qquad \text{für alle } k \in \mathbb{N}.$$

Also ist $x^k \in B_\varepsilon(\bar{x})$ für alle $k \in \mathbb{N}$ und damit $\nabla^2 f(x^k)$ invertierbar für alle $k \in \mathbb{N}$; außerdem folgt wegen $\eta < 1$ die lineare Konvergenz von $\{x^k\}_{k\in\mathbb{N}}$ bezüglich $\|\cdot\|_H$. □

Aus der Konvergenz bezüglich $\|\cdot\|_H$ folgt auch die Konvergenz bezüglich der äquivalenten Norm $\|\cdot\|$ (allerdings nicht die *lineare* Konvergenz, da für $\eta < 1$ nicht unbedingt auch $\eta \frac{\mu_2}{\mu_1} < 1$ sein muss!) Reduziert man die Toleranz im Laufe der Iteration schnell genug, folgt daraus mit Hilfe der Dennis–Moré-Bedingung die superlineare Konvergenz.

Satz 8.8. *Sei $f : \mathbb{R}^n \to \mathbb{R}$ zweimal stetig differenzierbar und sei $\bar{x} \in \mathbb{R}^n$ ein stationärer Punkt von $f$ mit $\nabla^2 f(\bar{x})$ invertierbar. Ist $\{\eta_k\}_{k\in\mathbb{N}} \subset (0,1)$ eine Nullfolge, so konvergiert Algorithmus 8.3 lokal superlinear. Ist $\nabla^2 f$ darüber hinaus lokal Lipschitz-stetig und $\eta_k = O(\|\nabla f(x^k)\|)$, so ist die Konvergenz sogar quadratisch.*

*Beweis.* Da die Voraussetzungen von Satz 8.7 erfüllt sind, konvergiert $x^k \to \bar{x}$. Aus der produktiven Null, der Dreiecksungleichung und der Iterationsvorschrift folgt nun

$$\begin{aligned}
\|\nabla f(x^k)\| &\leq \|\nabla f(x^k) + \nabla^2 f(x^k)(x^{k+1} - x^k)\| + \|\nabla^2 f(x^k)(x^{k+1} - x^k)\| \\
&\leq \eta_k \|\nabla f(x^k)\| + \|\nabla^2 f(x^k)(x^{k+1} - x^k)\|.
\end{aligned}$$

Wegen der Stetigkeit von $\nabla^2 f$ und $x^k \to \bar{x}$ existiert ein $C > 0$ und ein $k_0 \in \mathbb{N}$ mit $\|\nabla^2 f(x^k)\| \leq C$ für alle $k \geq k_0$. Daraus folgt, wieder nach Iterationsvorschrift,

$$\begin{aligned}\|\nabla f(x^k) + \nabla^2 f(x^k)(x^{k+1} - x^k)\| \leq \eta_k \|\nabla f(x^k)\| &\leq \frac{\eta_k}{1-\eta_k} \|\nabla^2 f(x^k)(x^{k+1} - x^k)\| \\ &\leq C \frac{\eta_k}{1-\eta_k} \|x^{k+1} - x^k\| =: \varepsilon_k \|x^{k+1} - x^k\|.\end{aligned}$$

Da mit $\{\eta_k\}_{k\in\mathbb{N}}$ auch $\{\varepsilon_k\}_{k\in\mathbb{N}}$ eine Nullfolge ist, folgt die superlineare Konvergenz mit Satz 7.9.

Analog zeigt man mit Hilfe von Satz 7.11 die quadratische Konvergenz. □

Auch das inexakte Newton-Verfahren lässt sich wie in Algorithmus 8.2 globalisieren. Die globale Konvergenz sowie den Übergang zu superlinearer Konvergenz kann man durch eine analoge Modifikation der entsprechenden Beweise in Abschnitt 8.2 zeigen; für Details sei auf [Geiger & Kanzow 1999, Kapitel 10.2] verwiesen.

# 9 QUASI-NEWTON-VERFAHREN

In der Praxis ist das Aufstellen der Hesse-Matrix oft aufwendig oder sogar überhaupt nicht möglich (wenn die zu minimierende Funktion zwar zweimal differenzierbar ist, aber die zweiten Ableitungen nicht mit vertretbarem Aufwand berechenbar sind). Die Dennis–Moré-Bedingung besagt aber, dass es ausreicht, eine hinreichend gute Näherung der Hesse-Matrix zu verwenden. Für Funktionen $f : \mathbb{R} \to \mathbb{R}$ wäre ein Ansatz, statt der zweiten Ableitung einen Differenzenquotienten zu verwenden:

$$f''(x^{k+1}) \approx \frac{f'(x^{k+1}) - f'(x^k)}{x^{k+1} - x^k}.$$

Der für $f : \mathbb{R}^n \to \mathbb{R}$ analoge Ansatz führt auf die *Quasi-Newton-Gleichung*

$$H_{k+1}(x^{k+1} - x^k) = \nabla f(x^{k+1}) - \nabla f(x^k). \tag{9.1}$$

Diese Gleichung für $H_{k+1}$ ist allerdings unterbestimmt, da durch sie nur die Wirkung auf eine Richtung $s^k = x^{k+1} - x^k$ festgelegt wird, wir für den nächsten Schritt aber $H_{k+1}s^{k+1}$ benötigen. Wir brauchen also noch eine weitere Forderung. Dazu betrachten wir wieder die Dennis–Moré-Bedingung, nach der der folgende Term für $x^k \to \bar{x}$ superlinear in $\|s^k\|$ sein soll:

$$\begin{aligned}\|(H_k - \nabla^2 f(x^k))s^k\| &\leq \|(H_k - H_{k+1})s^k\| + \|(H_{k+1} - \nabla^2 f(x^k))s^k\| \\ &= \|(H_k - H_{k+1})s^k\| + \|\nabla f(x^{k+1}) - \nabla f(x^k) - \nabla^2 f(x^k)s^k\|,\end{aligned}$$

wobei wir im zweiten Schritt die Quasi-Newton-Gleichung verwendet haben. Der zweite Term auf der rechten Seite ist nun für $f$ zweimal differenzierbar nach Definition $o(\|s^k\|)$; für die superlineare Konvergenz genügt also die Forderung

$$\lim_{k\to\infty} \|H_{k+1} - H_k\| = 0$$

für eine beliebige Matrixnorm.

Wir gehen daher wie folgt vor: Ausgehend von einer Startmatrix $H_0$ wählen wir für alle $k \in \mathbb{N}$ die neue Näherung $H_{k+1}$ so, dass

(i) die Quasi-Newton-Gleichung (9.1) erfüllt ist;

(ii) der Abstand $\|H_{k+1} - H_k\|$ minimiert wird.

Man spricht dabei von einem *Update* der Matrix $H_k$ auf $H_{k+1}$. Diese Wahl von $H_k$ in Algorithmus 7.1 führt auf die Klasse der *Quasi-Newton-Verfahren*, die sich als mit die leistungsfähigsten Verfahren zur unrestringierten Optimierung herausgestellt haben. Verschiedene Wahlen der Matrix-Norm führen dabei auf verschiedene Verfahren in dieser Klasse.

## 9.1 QUASI-NEWTON-UPDATES

Die Kernidee der verbreiteten Quasi-Newton-Verfahren ist, für die Minimierung nicht die induzierte Norm $\|A\| = \|A\|_2$, sondern die *Frobenius-Norm*

$$\|A\|_F := \left( \sum_{i=1}^{n} \sum_{j=1}^{n} a_{ij}^2 \right)^{1/2}$$

zu verwenden. Dies ist eine äquivalente Norm mit

$$n^{-1/2}\|A\|_F \leq \|A\| \leq \|A\|_F \qquad \text{für alle } A \in \mathbb{R}^{n\times n}. \tag{9.2}$$

Weiterhin gilt für jede Orthonormalbasis $\{v_1, \ldots, v_n\}$ von $\mathbb{R}^n$

$$\|A\|_F^2 = \sum_{i=1}^{n} \|Av_i\|^2. \tag{9.3}$$

(Beide Eigenschaften folgen direkt aus einer äquivalenten Charakterisierung der Frobenius-Norm über die Spur von $A^TA$, siehe z. B. [Geiger & Kanzow 1999, Lemma B.1].)

Wir bestimmen nun für gegebenes $H_k$ ein $H_{k+1}$, das die Quasi-Newton-Gleichung (9.1) erfüllt und $\|H_{k+1} - H_k\|_F$ minimiert. Dafür setzen wir der Kürze halber

$$H := H_k, \qquad H_+ := H_{k+1}, \qquad s := x^{k+1} - x^k, \qquad y := \nabla f(x^{k+1}) - \nabla f(x^k).$$

Die Quasi-Newton-Gleichung kann nun kurz geschrieben werden als $H_+s = y$. Für die Minimierung verwenden wir (9.3), und zwar indem wir $s/\|s\|$ zu einer Orthonormalbasis $\{s/\|s\|, v_2, \ldots, v_n\}$ des $\mathbb{R}^n$ ergänzen. Dann ist

$$\|H_+ - H\|_F^2 = \frac{1}{\|s\|^2}\|H_+s - Hs\|^2 + \sum_{i=2}^{n} \|H_+v_i - Hv_i\|^2.$$

Die Quasi-Newton-Gleichung erzwingt nun $H_+s = y$; im ersten Term haben wir also keine Freiheit für die Minimierung. Der zweite Term ist jedoch offensichtlich minimal, wenn $H_+v_i = Hv_i$ für $i = 2, \ldots, n$ gilt. Dies erreichen wir durch die Wahl

$$H_+ = H + \frac{(y - Hs)s^T}{s^Ts}, \tag{9.4}$$

denn dann ist wegen der Orthonormalität der Basis $s^T v_i = 0$ für alle $i = 2, \dots, n$ und damit

$$H_+ s = Hs + (y - Hs)\frac{s^T s}{s^T s} = y,$$

sowie

$$H_+ v_i = Hv_i + (y - Hs)\frac{s^T v_i}{s^T s} = Hv_i \quad \text{für alle } i = 2, \dots, n.$$

Die Wahl (9.4) wird *Broyden-Update* genannt; Matrizen von der Form $M = uv^T$ für $u, v \in \mathbb{R}^n$ heißen *Rang-1-Matrizen*, weshalb man auch von einem *Rang-1-Update* spricht.

Nachteil des Broyden-Updates ist, dass die neue Matrix $H_+$ weder symmetrisch noch positiv definit sein muss, auch wenn dies für $H$ der Fall ist. Ein Newton-artiges Verfahren mit dieser Wahl von $H_k$ würde daher auch gegen lokale Maximierer konvergieren (und tatsächlich wird das *Broyden-Verfahren* vor allem zur Lösung von nichtlinearen Gleichungen eingesetzt). Eine symmetrische Matrix erhält man durch einen *symmetrischen Rang-1-Update*, kurz *SR1-Update*,

$$H_+^{SR1} = H + \frac{(y - Hs)(y - Hs)^T}{(y - Hs)^T s},$$

der jedoch ebenfalls nicht positiv definit ist. Dafür benötigt man Rang-2-Updates (d. h. eine Summe von zwei Rang-1-Updates), die man als Minimierung einer *gewichteten* Frobenius-Norm erhält. Dies ist relativ technisch, weshalb wir hier auf Beweise verzichten. Wir bezeichnen in Folge die Menge der symmetrischen und positiv definiten Matrizen kurz als $SPD(n)$.

Satz 9.1 ([Geiger & Kanzow 1999, Satz 11.6]). *Seien $H \in SPD(n)$ und $s, y \in \mathbb{R}^n$ mit $s^T y > 0$ gegeben. Dann existiert eine Matrix $W \in SPD(n)$ mit $W^2 s = y$, und die Lösung von*

$$\min_{M \in SPD(n)} \|W^{-1}(M - H)W^{-1}\|_F \quad \text{mit} \quad Ms = y$$

*ist gegeben durch den* Davidon–Fletcher–Powell-Update, *kurz* DFP-Update,

$$H_+^{DFP} = H + \frac{(y - Hs)y^T + y(y - Hs)^T}{y^T s} - \frac{(y - Hs)^T s}{(y^T s)^2} yy^T.$$

Die Voraussetzung $s^T y > 0$ ist dabei sogar notwendig für die Existenz einer Matrix $M \in SPD(n)$, die die Quasi-Newton-Gleichung erfüllt: Gilt nämlich $s^T y \leq 0$, so folgt aus $Ms = y$ sofort $s^T Ms = s^T y \leq 0$, und damit ist $M$ nicht positiv definit.

Nun ist man eigentlich an der Lösung des Gleichungssystems $H_k s^k = -\nabla f(x^k)$ interessiert, was bei Kenntnis von $H_k^{-1}$ durch einfache Matrixmultiplikation möglich wäre – schön wäre daher ein *inverser Update* von $B := H^{-1}$ auf $B_+ := H_+^{-1}$. Dazu verwendet man einfach, dass unter der sehr sinnvollen Forderung der Invertierbarkeit von $H_+$ die Quasi-Newton-Gleichung $H_+ s = y$ äquivalent ist zur *inversen Quasi-Newton-Gleichung* $B_+ y = s$. Ganz analog wie oben erhält man daraus den folgenden Update.

Satz 9.2 ([Geiger & Kanzow 1999, Satz 11.8]). *Seien $B \in SPD(n)$ und $s, y \in \mathbb{R}^n$ mit $s^T y > 0$ gegeben. Dann existiert eine Matrix $W \in SPD(n)$ mit $W^2 s = y$, und die Lösung von*

$$\min_{M \in SPD(n)} \|W(M - B)W\|_F \quad \text{mit} \quad My = s$$

*ist gegeben durch den* inversen Broyden–Fletcher–Goldfarb–Shanno-Update, *kurz* BFGS-Update,

$$B_+^{BFGS} = B + \frac{(s - By)s^T + s(s - By)^T}{s^T y} - \frac{(s - By)^T y}{(s^T y)^2} ss^T.$$

Um daraus ein direktes BFGS-Update zu erhalten (bzw. aus Satz 9.1 ein inverses DFP-Update), verwenden wir die Tatsache, dass die Inverse einer Rang-1-Matrix wieder eine Rang-1-Matrix ist. Das folgende Lemma verifiziert man durch einfaches aber lästiges Ausrechnen.

Lemma 9.3 (Sherman–Morrison–Woodbury-Formel). *Seien $A \in \mathbb{R}^{n \times n}$ invertierbar und $u, v \in \mathbb{R}^n$. Ist $1 + v^T A^{-1} u \neq 0$, dann ist $A + uv^T$ invertierbar mit*

$$(A + uv^T)^{-1} = A^{-1} - \frac{A^{-1} u v^T A^{-1}}{1 + v^T A^{-1} u}.$$

Durch zweimaliges Anwenden erhält man daraus die folgenden Updates

Satz 9.4. *Seien $H \in SPD(n)$, $B = H^{-1}$, und $y, s \in \mathbb{R}^n$ mit $y^T s > 0$. Dann gilt mit $B_+ := H_+^{-1}$*

$$B_+^{DFP} = B + \frac{ss^T}{y^T s} - \frac{Byy^T B}{y^T By},$$
$$H_+^{BFGS} = H + \frac{yy^T}{s^T y} - \frac{Hss^T H}{s^T Hs}.$$

Anstelle der ausgelassenen Beweise weisen wir nun nach, dass das BFGS-Update tatsächlich die geforderten Eigenschaften hat.

Satz 9.5. *Seien $H \in SPD(n)$ und $y, s \in \mathbb{R}^n$ mit $y^T s > 0$. Dann ist $H_+^{BFGS} \in SPD(n)$ und erfüllt die Quasi-Newton-Gleichung.*

*Beweis.* Die Symmetrie von $H_+^{BFGS}$ für symmetrische $H$ ist direkt aus der Definition ersichtlich.

Für die positive Definitheit verwenden wir, dass $H \in SPD(n)$ eine Cholesky-Zerlegung $H = R^T R$ mit $R \in \mathbb{R}^{n \times n}$ invertierbar besitzt.[1] Für beliebige $d \in \mathbb{R}^n \setminus \{0\}$ folgt dann aus der Cauchy–Schwarz-Ungleichung

$$\begin{aligned} d^T H_+^{BFGS} d &= d^T H d + \frac{(d^T y)^2}{y^T s} - \frac{(d^T H s)^2}{s^T H s} \\ &= \|Rd\|^2 + \frac{(d^T y)^2}{y^T s} - \frac{((Rd)^T (Rs))^2}{\|Rs\|^2} \\ &\geq \|Rd\|^2 + \frac{(d^T y)^2}{y^T s} - \frac{\|Rd\|^2 \|Rs\|^2}{\|Rs\|^2} \\ &= \frac{(d^T y)^2}{y^T s} \geq 0. \end{aligned}$$

Also ist $H_+^{BFGS}$ zumindest semidefinit. Für die positive Definitheit genügt es, dass eine der beiden Ungleichungen strikt ist. Angenommen, die erste Ungleichung ist nicht strikt, d. h. die Cauchy–Schwarz-Ungleichung gilt mit Gleichheit. Dies ist nur möglich, falls $Rd$ und $Rs$ linear abhängig sind, d. h. es gilt $Rd = tRs$ für ein $t \in \mathbb{R} \setminus \{0\}$. Dann folgt aus der Invertierbarkeit von $R$ aber auch $d = ts$ und damit

$$\frac{(d^T y)^2}{y^T s} = t^2 \frac{(s^T y)^2}{y^T s} = t^2 (s^T y) > 0$$

wegen $t \neq 0$ und $s^T y > 0$.

Für die Quasi-Newton-Gleichung rechnen wir einfach nach, dass dann gilt

$$H_+^{BFGS} s = Hs + \frac{y^T s}{s^T y} y - \frac{s^T H s}{s^T H s} Hs = y. \qquad \square$$

## 9.2 LOKALE KONVERGENZ

Wir untersuchen nun die lokale superlineare Konvergenz von Quasi-Newton-Verfahren. Da dies sehr technisch ist, beschränken wir uns für den Beweis auf das einfacher zu analysierende (wenn auch für die Optimierung weniger relevante) lokale Broyden-Verfahren. Einsetzen der Definition des Broyden-Updates in das allgemeine Newton-artige Verfahren ergibt den folgenden Algorithmus.

[1] siehe z. B. [Hanke-Bourgeois 2009, Satz 5.4]

**Algorithmus 9.1 :** Lokales Broyden-Verfahren

**Input :** $x^0 \in \mathbb{R}^n$, $H_0 \in \mathbb{R}^{n\times n}$

Setze $k = 0$

**while** $\|\nabla f(x^k)\| > 0$ **do**

  Berechne $s^k$ als Lösung von $H_k s^k = -\nabla f(x^k)$

  Setze $x^{k+1} = x^k + s^k$

  Setze $y^k = \nabla f(x^{k+1}) - \nabla f(x^k)$

  Setze $H_{k+1} = H_k + \frac{(y^k - H_k s^k)(s^k)^T}{(s^k)^T(s^k)}$, $\quad k \leftarrow k+1$

Wir zeigen wieder zuerst die lineare Konvergenz. Dafür benötigen wir das folgende Lemma, das garantiert, dass der Broyden-Update den Abstand zur exakten Hesse-Matrix $\nabla^2 f(\bar{x})$ nicht zu sehr vergrößert.

Lemma 9.6. *Sei $f : \mathbb{R}^n \to \mathbb{R}$ zweimal stetig differenzierbar mit $\nabla^2 f$ lokal Lipschitz-stetig, und sei $\{x^k\}_{k\in\mathbb{N}}$ eine von Algorithmus 9.1 erzeugte Folge mit $x^k \to \bar{x}$. Dann gilt für alle $k \in \mathbb{N}$ groß genug*

$$\|H_{k+1} - \nabla^2 f(\bar{x})\| \leq \|H_k - \nabla^2 f(\bar{x})\| + \frac{L}{2}\left(\|x^{k+1} - \bar{x}\| + \|x^k - \bar{x}\|\right).$$

*Beweis.* Aus der Definition des Broyden-Updates folgt

$$\begin{aligned} H_{k+1} - \nabla^2 f(\bar{x}) &= H_k - \nabla^2 f(\bar{x}) + \frac{(y^k - H_k s^k)(s^k)^T}{(s^k)^T(s^k)} \\ &= H_k - \nabla^2 f(\bar{x}) + \frac{(\nabla^2 f(\bar{x})s^k - H_k s^k)(s^k)^T}{(s^k)^T(s^k)} + \frac{(y^k - \nabla^2 f(\bar{x})s^k)(s^k)^T}{(s^k)^T(s^k)} \\ &= (H_k - \nabla^2 f(\bar{x}))\left(I - \frac{s^k(s^k)^T}{(s^k)^T s^k}\right) + \frac{(y^k - \nabla^2 f(\bar{x})s^k)(s^k)^T}{(s^k)^T(s^k)}. \end{aligned} \tag{9.5}$$

Aus der Definition der induzierten Matrix-Norm folgt nun $\|uv^T\| = \|u\|\|v\|$ für alle $u, v \in \mathbb{R}^n$ sowie $\|I - \frac{uu^T}{u^T u}\| = 1$ für alle $u \in \mathbb{R}^n$. Wir erhalten also

$$\|H_{k+1} - \nabla^2 f(\bar{x})\| \leq \|H_k - \nabla^2 f(\bar{x})\| + \frac{\|y^k - \nabla^2 f(\bar{x})s^k\|}{\|s^k\|}.$$

Um den zweiten Summanden abzuschätzen, verwenden wir Satz 1.3 sowie die lokale

Lipschitz-Stetigkeit von $\nabla^2 f$ und erhalten für $x^k$ hinreichend nahe an $\bar{x}$

$$\begin{aligned}(9.6)\qquad \|y^k - \nabla^2 f(\bar{x})s^k\| &= \|\nabla f(x^{k+1}) - \nabla f(x^k) - \nabla^2 f(\bar{x})(x^{k+1} - x^k)\| \\ &\leq \int_0^1 \|\nabla^2 f(x^k + t(x^{k+1} - x^k)) - \nabla^2 f(\bar{x})\|\, dt \cdot \|x^{k+1} - x^k\| \\ &\leq L \int_0^1 t\|x^{k+1} - \bar{x}\| + (1-t)\|x^k - \bar{x}\|\, dt \cdot \|x^{k+1} - x^k\| \\ &= \frac{L}{2}\left(\|x^{k+1} - \bar{x}\| + \|x^k - \bar{x}\|\right)\|s^k\|,\end{aligned}$$

woraus die Aussage folgt. □

Wir können nun die lokale lineare Konvergenz zeigen.

Satz 9.7. *Sei $f : \mathbb{R}^n \to \mathbb{R}$ zweimal stetig differenzierbar mit $\nabla^2 f$ lokal Lipschitz-stetig und sei $\bar{x} \in \mathbb{R}^n$ ein stationärer Punkt von $f$ mit $\nabla^2 f(\bar{x})$ invertierbar. Dann existieren Konstanten $\delta, \varepsilon > 0$, so dass für $x_0 \in B_\varepsilon(\bar{x})$ und $H_0 \in B_\delta(\nabla^2 f(\bar{x}))$ die von Algorithmus 9.1 erzeugte Folge $\{x^k\}_{k\in\mathbb{N}}$ linear gegen $\bar{x}$ konvergiert.*

*Beweis.* Setze $c := \|\nabla^2 f(\bar{x})^{-1}\|$ und wähle $\varepsilon, \delta > 0$ mit

$$\delta \leq \frac{1}{6c}, \qquad \varepsilon \leq \frac{2\delta}{3L},$$

und seien $x_0 \in B_\varepsilon(\bar{x})$ und $H_0 \in B_\delta(\nabla^2 f(\bar{x}))$ beliebig. Wir zeigen nun per starker Induktion, dass für alle $k \in \mathbb{N}$ gilt

$$\|H_k - \nabla^2 f(\bar{x})\| \leq (2 - 2^{-k})\delta, \tag{9.7}$$

$$\|x^{k+1} - \bar{x}\| \leq \frac{1}{2}\|x^k - \bar{x}\|. \tag{9.8}$$

Es sei nun $k \in \mathbb{N}$ beliebig und gelte (9.7) und (9.8) für alle $i = 0, \ldots, k-1$. Wir zeigen zuerst, dass (9.7) für $k$ gilt. Aus den beiden Induktionsvoraussetzungen zusammen mit Lemma 9.6 folgt

$$\|H_k - \nabla^2 f(\bar{x})\| \leq (2 - 2^{-(k-1)})\delta + \frac{3L}{4}\|x^{k-1} - \bar{x}\|.$$

Weiter folgt aus der Induktionsvoraussetzung (9.8) und $x_0 \in B_\varepsilon(\bar{x})$

$$\|x^{k-1} - \bar{x}\| \leq 2^{-(k-1)}\|x^0 - \bar{x}\| \leq 2^{-(k-1)}\varepsilon. \tag{9.9}$$

Nach Wahl von $\varepsilon$ gilt daher

$$\begin{aligned}\|H_k - \nabla^2 f(\bar{x})\| &\leq (2 - 2^{-(k-1)})\delta + \frac{3L}{4}2^{-(k-1)}\varepsilon \leq (2 - 2^{-(k-1)} + 2^{-k})\delta \\ &= (2 - 2^{-k})\delta\end{aligned}$$

und damit der Induktionsschritt für (9.7).

Als nächstes zeigen wir, dass $H_k$ invertierbar ist. Aus der Wahl von $\delta$ und der Induktionsvoraussetzung (9.7) folgt

$$\|I - \nabla^2 f(\bar{x})^{-1} H_k\| = \|\nabla^2 f(\bar{x})^{-1}(H_k - \nabla^2 f(\bar{x}))\| \leq c(2 - 2^{-k})\delta \leq 2c\delta \leq \frac{1}{3}.$$

Nach dem Banach-Lemma ist also $H_k$ invertierbar mit

$$\|H_k^{-1}\| \leq \frac{\|\nabla^2 f(\bar{x})^{-1}\|}{1 - \|I - \nabla^2 f(\bar{x})^{-1} H_k\|} \leq \frac{c}{1 - \frac{1}{3}} = \frac{3}{2}c.$$

Es bleibt der Induktionsschritt für (9.8). Aus dem Iterationsschritt folgt zunächst mit Hilfe der produktiven Null

$$H_k(x^{k+1} - \bar{x}) = -\nabla f(x^k) + \nabla f(\bar{x}) + \nabla^2 f(\bar{x})(x^k - \bar{x}) + (H_k - \nabla^2 f(\bar{x}))(x^k - \bar{x})$$

und damit

$$\|x^{k+1} - \bar{x}\| \leq \|H_k^{-1}\| \left( \|\nabla f(x^k) - \nabla f(\bar{x}) - \nabla^2 f(\bar{x})(x^k - \bar{x})\| + \|H_k - \nabla^2 f(\bar{x})\| \|x^k - \bar{x}\| \right).$$

Für den ersten Term in der Klammer gilt analog zu (9.6) mit (9.9) und der Wahl von $\varepsilon$

$$\begin{aligned} \|\nabla f(x^k) - \nabla f(\bar{x}) - \nabla^2 f(\bar{x})(x^k - \bar{x})\| &\leq \frac{L}{2}\|x^k - \bar{x}\|^2 \leq 2^{-(k+1)} \varepsilon L \|x^k - \bar{x}\| \\ &\leq \frac{2^{-k}}{3}\delta \|x^k - \bar{x}\|. \end{aligned}$$

Für den zweiten Term verwenden wir natürlich die Induktionsvoraussetzung (9.8) und erhalten nach Wahl von $\delta$

$$\|x^{k+1} - \bar{x}\| \leq \frac{3}{2}c \left( \frac{2^{-k}}{3} + 2 - 2^{-k} \right) \delta \|x^k - \bar{x}\| \leq 3c\delta \|x^k - \bar{x}\| \leq \frac{1}{2}\|x^k - \bar{x}\|$$

und damit den Induktionsschritt für (9.8), woraus auch die behauptete lineare Konvergenz folgt. □

Da also $x^k \to \bar{x}$ gilt, können wir die superlineare Konvergenz wieder mit Hilfe der Dennis–Moré-Bedingung zeigen.

Satz 9.8. *Sei $f : \mathbb{R}^n \to \mathbb{R}$ zweimal stetig differenzierbar mit $\nabla^2 f$ lokal Lipschitz-stetig und sei $\bar{x} \in \mathbb{R}^n$ ein stationärer Punkt von $f$ mit $\nabla^2 f(\bar{x})$ invertierbar. Dann konvergiert Algorithmus 9.1 lokal superlinear gegen $\bar{x}$.*

*Beweis.* Wir knüpfen an den Beweis von Satz 9.7 an, wobei wir nun für den Approximationsfehler in $H_k$ die Frobenius-Norm verwenden. Dafür benötigen wir die folgenden Eigenschaften, die man mit Hilfe von (9.3) beweisen kann:

(i) $\|uv^T\|_F = \|u\|\|v\|$ für alle $u, v \in \mathbb{R}^n$,

(ii) $\|A(I - \frac{vv^T}{v^Tv})\|_F \leq \|A\|_F - \frac{1}{2\|A\|_F}\left(\frac{\|Av\|}{\|v\|}\right)^2$ für alle $A \in \mathbb{R}^{n\times n}$, $v \in \mathbb{R}^n \setminus \{0\}$.

Wir setzen in Folge

$$e^k := x^k - \bar{x}, \qquad E_k := H_k - \nabla^2 f(\bar{x}).$$

Wie im Beweis von Lemma 9.6 folgt zuerst aus (9.5) und (9.6) sowie (9.8)

$$\begin{aligned}\|E_{k+1}\|_F &\leq \|E_k(I - \tfrac{s^k(s^k)^T}{(s^k)^Ts^k})\|_F + \frac{L}{2}\left(\|e^{k+1}\| + \|e^k\|\right)\\ &\leq \|E_k\|_F - \frac{\|E_ks^k\|^2}{2\|E_k\|_F\|s^k\|^2} + \frac{3}{4}L\|e^k\|.\end{aligned}$$

Durch Umformen erhalten wir daraus

$$\begin{aligned}\frac{\|E_ks^k\|^2}{\|s^k\|^2} &\leq 2\|E_k\|_F\left(\|E_k\|_F - \|E_{k+1}\|_F + \frac{3}{4}L\|e^k\|\right)\\ &\leq 4\sqrt{n}\delta\left(\|E_k\|_F - \|E_{k+1}\|_F + \frac{3}{4}L\|e^k\|\right),\end{aligned}$$

wobei wir im letzten Schritt (9.2) sowie (9.7) verwendet haben. Wir summieren diese Gleichung nun über alle $k = 0, \ldots, m$ für $m \in \mathbb{N}$ beliebig und erhalten als Teleskopsumme

$$\sum_{k=0}^{m}\frac{\|E_ks^k\|^2}{\|s^k\|^2} \leq 4\sqrt{n}\delta\left(\|E_0\|_F - \|E_{m+1}\|_F + \frac{3}{4}L\sum_{k=0}^{m}\|e^k\|\right). \tag{9.10}$$

Nun gilt nach Voraussetzung an $H_0$ Die Ungleichung $\|E_0\|_F - \|E_{m+1}\|_F \leq \|E_0\|_F \leq \sqrt{n}\delta$, und daraus sowie (9.8) und der geometrischen Reihe folgt

$$\sum_{k=0}^{m}\|e^k\| \leq \sum_{k=0}^{m}\left(\tfrac{1}{2}\right)^k\|e_0\| \leq (2 - 2^{-m})\varepsilon.$$

Einsetzen in (9.10) und Grenzübergang $m \longrightarrow \infty$ ergibt daher

$$\sum_{k=0}^{\infty}\frac{\|E_ks^k\|^2}{\|s^k\|^2} \leq 4\sqrt{n}\delta\left(\sqrt{n}\delta + \tfrac{3}{2}L\varepsilon\right) < \infty.$$

Also gilt $\frac{\|E_ks^k\|^2}{\|s^k\|^2} \longrightarrow 0$ und damit auch

$$\frac{\|E_ks^k\|}{\|s^k\|} = \frac{\|(H_k - \nabla^2 f(\bar{x}))(x^{k+1} - x^k)\|}{\|x^{k+1} - x^k\|} \longrightarrow 0,$$

und aus Folgerung 7.10 erhalten wir die superlineare Konvergenz. □

Das am weitesten verbreitete Quasi-Newton-Verfahren ist jedoch das BFGS-Verfahren mit inverser Aufdatierung.

**Algorithmus 9.2 :** Lokales BFGS-Verfahren

**Input :** $x^0 \in \mathbb{R}^n$, $B_0 \in SPD(n)$

1 Setze $k = 0$

2 **while** $\|\nabla f(x^k)\| > 0$ **do**

3 Setze $s^k = -B_k \nabla f(x^k)$

4 Setze $x^{k+1} = x^k + s^k$

5 Setze $y^k = \nabla f(x^{k+1}) - \nabla f(x^k)$

6 Setze $B_{k+1} = B_k + \frac{(s^k - B_k y^k)(s^k)^T + s^k (s^k - B_k y^k)^T}{(s^k)^T y^k} - \frac{(s^k - B_k y^k)^T y^k}{((s^k)^T y^k)^2} s^k (s^k)^T, \quad k \leftarrow k+1$

Als Start-Matrix kann z. B. $B_0 = I$ gewählt werden. In der Praxis werden dabei anstelle von $B_k$ die Vektoren $s^k$, $y^k$ sowie die im inversen BFGS-Update auftauchenden Skalarprodukte gespeichert. Damit lässt sich das Produkt $B_k v$ für gegebenes $v \in \mathbb{R}^n$ durch eine rekursive Prozedur effizient berechnen; siehe etwa [Kelley 1999, Kapitel 4.2.1]. Für sehr große $n$ bringt dies eine erhebliche Ersparnis mit sich. In den sogenannten *limited-memory-BFGS-Verfahren* werden darüber hinaus nur die letzten $m$ (z. B. $m = 30$) Vektoren und Skalare aufbewahrt; siehe [Geiger & Kanzow 1999, Kapitel 13]. Diese gehören zu den derzeit effizientesten Optimierungsverfahren für große Probleme.

Auf ähnliche Weise (wenn auch mit deutlich mehr Aufwand) wie für das Broyden-Verfahren zeigt man die lokal superlineare Konvergenz von Algorithmus 9.2, wobei man eine entsprechend Satz 9.2 gewichtete Frobenius-Norm sowie eine "inverse Dennis–Moré-Bedingung" für den Fehler $B_k - \nabla^2 f(\bar{x})^{-1}$ verwenden muss.

Satz 9.9 ([Geiger & Kanzow 1999, Satz 11.33]). *Sei $f : \mathbb{R}^n \to \mathbb{R}$ zweimal stetig differenzierbar mit $\nabla^2 f$ lokal Lipschitz-stetig und sei $\bar{x} \in \mathbb{R}^n$ ein stationärer Punkt von $f$ mit $\nabla^2 f(\bar{x})$ positiv definit. Dann existieren Konstanten $\delta, \varepsilon > 0$, so dass für $x_0 \in B_\varepsilon(\bar{x})$ und $B_0 \in B_\delta(\nabla^2 f(\bar{x})^{-1})$ mit $B_0 \in SPD(n)$ der Algorithmus 9.2 superlinear gegen $\bar{x}$ konvergiert.*

## 9.3 GLOBALE KONVERGENZ

Die Globalisierung von Quasi-Newton-Verfahren erfolgt analog zum Newton-Verfahren, wobei hier die Powell–Wolfe-Regel verwendet werden muss, um positiv definite Updates zu garantieren.

**Algorithmus 9.3 :** Globalisiertes BFGS-Verfahren

**Input :** $\gamma \in (0,1/2)$, $\eta \in (\gamma,1)$, $x^0 \in \mathbb{R}^n$, $B_0 \in SPD(n)$

1 Setze $k = 0$

2 **while** $\|\nabla f(x^k)\| > 0$ **do**

3 Setze $s^k = -B_k \nabla f(x^k)$

4 Bestimme $\sigma_k > 0$ mit Algorithmus 5.3 für $\gamma \in (0,1/2)$ und $\eta \in (\gamma,1)$

5 Setze $x^{k+1} = x^k + \sigma_k s^k$

6 Setze $d^k := \sigma_k s^k$, $y^k = \nabla f(x^{k+1}) - \nabla f(x^k)$

7 Setze $B_{k+1} = B_k + \frac{(d^k - B_k y^k)(d^k)^T + d^k (d^k - B_k y^k)^T}{(d^k)^T y^k} - \frac{(d^k - B_k y^k)^T y^k}{((d^k)^T y^k)^2} d^k (d^k)^T, \quad k \leftarrow k+1$

Wir zeigen zuerst, dass die Powell–Wolfe-Regel in der Tat zu positiv definiten Updates führt.

Lemma 9.10. *Sei $f : \mathbb{R}^n \to \mathbb{R}$ stetig differenzierbar. Ist $\nabla f(x^k) \neq 0$, $B_k \in SPD(n)$, und $\sigma_k > 0$ nach der Powell–Wolfe-Regel gewählt, so ist auch $B_{k+1} \in SPD(n)$.*

*Beweis.* Eine Matrix $A$ ist genau dann positiv definit, wenn $A^{-1}$ positiv definit ist. Nach Satz 9.5 genügt daher, $(y^k)^T d^k > 0$ zu zeigen. Da $\sigma_k > 0$ nach Voraussetzung die Krümmungsbedingung

$$\nabla f(x^{k+1})^T s^k \geq \eta \nabla f(x^k)^T s^k$$

für ein $\eta < 1$ erfüllt, gilt nach Definition von $d^k$ und $y^k$

$$\begin{aligned}(y^k)^T d^k &= \sigma_k (\nabla f(x^{k+1}) s^k - \nabla f(x^k)^T s^k) \\ &\geq -\sigma_k (1-\eta) \nabla f(x^k)^T s^k \\ &= \sigma_k (1-\eta) \nabla f(x^k)^T B_k \nabla f(x^k) > 0\end{aligned}$$

weil $\eta < 1$ und $B_k$ positiv definit ist, und damit die Behauptung. □

Die globale Konvergenz folgt nun aus den abstrakten Konvergenzresultaten.

Satz 9.11. *Sei $f : \mathbb{R}^n \to \mathbb{R}$ Lipschitz-stetig differenzierbar und nach unten beschränkt. Existieren Konstanten $0 < \mu_1 \leq \mu_2$, so dass für alle $k \in \mathbb{N}$ die Eigenwerte von $B_k$ die Abschätzung*

$$\mu_1 \leq \lambda_1^k \leq \lambda_n^k \leq \mu_2$$

*erfüllen, dann konvergiert Algorithmus 9.3 global.*

*Beweis.* Solange $\nabla f(x^k) \neq 0$ und $B_k$ positiv definit ist, gilt

$$\nabla f(x^k)^T s^k = -\nabla f(x^k)^T B_k \nabla f(x^k) < 0,$$

d. h. $s^k$ ist eine Abstiegsrichtung. Nach Satz 5.4 liefert Algorithmus 5.3 eine Schrittweite, die die Powell–Wolfe–Bedingung erfüllt. Lemma 9.10 liefert dann die positive Definitheit von $B_{k+1}$, woraus per Induktion die Durchführbarkeit von Algorithmus 9.3 folgt.

Da mit $B_k$ auch alle $H_k = B_k^{-1}$ symmetrisch und positiv definit sind, sind nach Lemma 7.1 die BFGS-Schrittweiten zulässig und nach Satz 5.5 die Powell–Wolfe-Schrittweiten effizient. Aus Satz 4.3 folgt nun die globale Konvergenz. □

Ähnlich wie für das Newton-Verfahren kann man zeigen, dass unter den Voraussetzungen von Satz 9.9 in Algorithmus 9.3 irgendwann stets die Schrittweite $\sigma_k = 1$ akzeptiert wird und damit lokal superlineare Konvergenz erreicht wird, siehe [Spellucci 1993, Satz 3.1.13]. Die Eigenwertbedingung kann allerdings nur für gleichmäßig konvexe Funktionen garantiert werden (siehe z. B. [Geiger & Kanzow 1999, Kapitel 11.5]); im Allgemeinen wird man daher in jeder Iteration überprüfen, ob $s^k$ eine Abstiegsrichtung ist, und falls nicht, die Matrix $B_k$ neu initialisieren (etwa mit $B_k = B_0$).

# 10 TRUST-REGION-VERFAHREN

Wir haben in den letzten Kapiteln gesehen, dass lokal konvergente Newton-artige Verfahren mit Hilfe einer Schrittweitensuche globalisiert werden können. In diesem Kapitel untersuchen wir eine Alternative, die auch ohne die Voraussetzung der positiven Definitheit auskommt und daher direkt auf das Newton-Verfahren (ohne Rückfall auf Gradientenschritte) angewendet werden kann. Ausgangspunkt ist die Beobachtung, dass in Newton-artigen Verfahren die Charakterisierung

$$H_k s^k = -\nabla f(x^k) \tag{10.1}$$

der Suchrichtung der (für $H_k$ symmetrisch) notwendigen (und für $H_k$ positiv definit auch hinreichenden) Optimalitätsbedingung entspricht für das Problem

$$\min_{d \in \mathbb{R}^n} f(x^k) + \nabla f(x^k)^T d + \frac{1}{2} d^T H_k d.$$

Ist nun $H_k$ sehr schlecht konditioniert oder gar nicht positiv definit, so kann die Lösung von (10.1) beliebig schlecht sein oder gar nicht existieren. Statt zu versuchen, dies durch eine Schrittweitensuche (oder einen Gradientenschritt) zurechtzubiegen, ist die Idee nun, die Minimierung auf einen *Vertrauensbereich* (Englisch: *trust region*) $K_\Delta := \overline{B_\Delta(0)} = \{x \in \mathbb{R}^n : \|x\| \leq \Delta\}$ für einen gegebenen *Trust-Region-Radius* $\Delta > 0$ einzuschränken. Man betrachtet also das Problem

$$\min_{d \in K_\Delta} f(x^k) + \nabla f(x^k)^T d + \frac{1}{2} d^T H_k d. \tag{10.2}$$

Da $K_\Delta$ kompakt und die zu minimierende Funktion stetig ist, existiert in jedem Fall ein Minimierer $s^k := \bar{d} \in K_\Delta$, auch wenn $H_k$ nicht positiv definit ist. Hierbei spielt der Trust-Region-Radius die Rolle der Schrittweite, und für die globale Konvergenz ist wichtig, den Radius $\Delta$ im Verlauf der Iteration richtig zu wählen. Der Ansatz ist hier, dies nicht in jedem Schritt komplett neu zu tun, sondern $\Delta$ in Abhängigkeit vom Erfolg des Schrittes geeignet zu vergrößern oder zu verkleinern.

Wir untersuchen zunächst diese Anpassung für den konkreten Fall $H_k = \nabla^2 f(x^k)$, bevor wir am Ende des Kapitels auf die Berechnung des Trust-Region-Schrittes $s^k \in K_\Delta$ eingehen.

## 10.1 DAS TRUST-REGION-NEWTON-VERFAHREN

Mit der Wahl $H_k = \nabla^2 f(x^k)$ entspricht die quadratische Funktion

$$q_k(d) := f(x^k) + \nabla f(x^k)^T d + \frac{1}{2} d^T \nabla^2 f(x^k) d$$

der Taylor-Entwicklung von $f$ im Punkt $x^k$; es gilt also

$$f(x^k + d) = q_k(d) + o(\|d\|^2). \tag{10.3}$$

Es handelt sich bei $q_k$ daher um ein *quadratisches Modell* von $f$, das für kleine $\|d\|$ gut mit der Funktion $f$ übereinstimmt (dies motiviert die Bezeichnung "Vertrauensbereich" für $K_\Delta$ – nur innerhalb dieses Bereichs "trauen" wir dem Modell). Umgekehrt können wir den Grad der Übereinstimmung heranziehen, um einzuschätzen, ob der Radius $\Delta_k$ gut gewählt war. Haben wir einen Schritt $s^k$ als Lösung von

$$\min_{d \in K_{\Delta_k}} q_k(d) \tag{10.4}$$

berechnet, können wir die *tatsächliche Reduktion* $f(x^k) - f(x^k + s^k)$ mit der *vorausgesagten Reduktion* $f(x^k) - q_k(s^k)$ vergleichen. Konkret betrachten wir den Quotient

$$\rho_k := \frac{f(x^k) - f(x^k + s^k)}{f(x^k) - q_k(s^k)} \tag{10.5}$$

und machen eine Fallunterscheidung:

(i) Ist $\rho_k$ sehr klein (und insbesondere negativ), so war $q_k$ kein gutes Modell für $f$ im Vertrauensbereich bzw. der Vertrauensbereich zu groß; wir verwerfen also den Schritt und versuchen es erneut mit einem kleineren Radius, d. h. wir setzen $x^{k+1} = x^k$ und $\Delta_{k+1} < \Delta_k$.

(ii) Ist $\rho_k$ klein aber nicht sehr klein, so stimmt das Modell im Vertrauensbereich hinreichend gut mit der Funktion überein. Wir können also den Schritt $x^{k+1} = x^k + s^k$ akzeptieren und den Radius beibehalten

(iii) Ist $\rho_k$ ungefähr 1, so ist die Übereinstimmung sogar sehr gut. Wir können also nicht nur den Schritt $x^{k+1} = x^k + s^k$ akzeptieren, sondern auch im nächsten Schritt einen noch größeren Radius $\Delta_{k+1} > \Delta_k$ versuchen.

Ein Schritt mit $x^{k+1} = x^k + s^k$ heißt *erfolgreich*. Der Fall (ii) soll dabei verhindern, dass irgendwann stets der Radius vergrößert wird, nur um ihn im nächsten Schritt wieder zu reduzieren. Der folgende Algorithmus präzisiert das Vorgehen, wobei wir für die Konvergenz sicherstellen müssen, dass der Trust-Region-Radius bei *erfolgreichen* Iterationen einen vorgegebenen Minimalradius $\Delta_{\min} > 0$ nicht unterschreitet.

**Algorithmus 10.1 :** Trust-Region-Newton-Verfahren

**Input :** $\eta_1 \in (0,1)$, $\eta_2 \in (\eta_1, 1)$, $\sigma_1 \in (0,1)$, $\sigma_2 \in (1,\infty)$, $\Delta_{\min} > 0$

1 Wähle $x^0 \in \mathbb{R}^n$, $\Delta_0 > 0$
2 **while** $\|\nabla f(x^k)\| > 0$ **do**
3  Berechne $s^k$ als Lösung von (10.4)
4  Berechne $\rho_k$ nach (10.5)
5  **if** $\rho_k < \eta_1$ **then** // Schritt nicht erfolgreich, Modell schlecht
6  Setze $x^{k+1} = x^k$ // verwerfe Schritt
7  Setze $\Delta_{k+1} = \sigma_1 \Delta_k$ // verkleinere Radius
8  **else if** $\eta_1 \le \rho_k < \eta_2$ **then** // Schritt erfolgreich, Modell OK
9  Setze $x^{k+1} = x^k + s^k$ // akzeptiere Schritt
10  Setze $\Delta_{k+1} = \max\{\Delta_{\min}, \Delta_k\}$ // behalte Radius
11  **else if** $\rho_k \ge \eta_2$ **then** // Schritt erfolgreich, Modell gut
12  Setze $x^{k+1} = x^k + s^k$ // akzeptiere Schritt
13  Setze $\Delta_{k+1} = \max\{\Delta_{\min}, \sigma_2 \Delta_k\}$ // vergrößere Radius
14  Setze $k \leftarrow k+1$

Da der Vertrauensbereich wegen $\Delta_k > 0$ (was durch die Radius-Anpassung gewährleistet bleibt) nichtleer, abgeschlossen und beschränkt und das quadratische Modell $q_k$ stetig ist, existiert nach Satz 2.2 stets eine Lösung $s^k$ von (10.4). Schiefgehen kann also nur Schritt 4, und zwar wenn der Nenner von (10.5) gleich Null ist, d. h. der vorausgesagte Abstieg gleich Null ist. Unangenehm wäre auch, wenn der Nenner *negativ* ist, denn dann würde Algorithmus 10.1 einen *Aufstiegsschritt* akzeptieren. Das nächste Lemma garantiert, dass beides erst bei Erreichen eines stationären Punkts eintreten kann, und ist fundamental für die Konvergenz des Trust-Region-Verfahrens (vergleiche die Armijo-Bedingung für Schrittweiten).

Lemma 10.1. *Sei $s^k \in \mathbb{R}^n$ eine Lösung von (10.4). Dann gilt*

$$f(x^k) - q_k(s^k) \ge \frac{1}{2}\|\nabla f(x^k)\| \min\left\{\Delta_k, \frac{\|\nabla f(x^k)\|}{\|\nabla^2 f(x^k)\|}\right\}.$$

*Beweis.* Da $s^k$ als globaler Minimierer von $q_k$ über $K_{\Delta_k}$ gewählt ist, gilt für alle $d \in K_{\Delta_k}$

$$\begin{aligned} f(x^k) - q_k(s^k) &\ge f(x^k) - q_k(d) \\ &= -\nabla f(x^k)^T d - \frac{1}{2} d^T \nabla^2 f(x^k) d \\ &\ge -\nabla f(x^k)^T d - \frac{1}{2}\|\nabla^2 f(x^k)\|\,\|d\|^2. \end{aligned}$$

Die gewünschte Abschätzung folgt nun durch Einsetzen einer geeigneten Wahl von $d$. Dafür machen wir eine Fallunterscheidung:

(i) $\|\nabla f(x^k)\| < \Delta_k \|\nabla^2 f(x^k)\|$: In diesem Fall wählen wir $d = -\frac{1}{\|\nabla^2 f(x^k)\|}\nabla f(x^k) \in K_{\Delta_k}$ und erhalten

$$f(x^k) - q_k(x^k) \geq \frac{\|\nabla f(x^k)\|^2}{\|\nabla^2 f(x^k)\|} - \frac{1}{2}\frac{\|\nabla f(x^k)\|^2}{\|\nabla^2 f(x^k)\|} = \frac{1}{2}\frac{\|\nabla f(x^k)\|^2}{\|\nabla^2 f(x^k)\|}.$$

(ii) $\|\nabla f(x^k)\| \geq \Delta_k \|\nabla^2 f(x^k)\|$: In diesem Fall wählen wir $d = -\frac{\Delta_k}{\|\nabla f(x^k)\|}\nabla f(x^k) \in K_{\Delta_k}$ und erhalten unter Verwendung der Annahme

$$f(x^k) - q_k(x^k) \geq \Delta_k \|\nabla f(x^k)\| - \frac{1}{2}\Delta_k^2 \|\nabla^2 f(x^k)\| \geq \frac{1}{2}\Delta_k \|\nabla f(x^k)\|.$$

Schätzen wir in beiden Fällen durch das Minimum der rechten Seiten ab, erhalten wir die Aussage. □

Algorithmus 10.1 ist also stets durchführbar, könnte aber "leer laufen", indem irgendwann keine Schritte mehr akzeptiert werden. Das folgende, technische, Lemma garantiert, dass das nicht eintritt. (Beachte, dass nur bei nicht erfolgreichen Schritten der Radius verkleinert wird.)

Lemma 10.2. *Sei $f : \mathbb{R}^n \to \mathbb{R}$ zweimal stetig differenzierbar und sei $\{x^k\}_{k\in\mathbb{N}}$ eine durch Algorithmus 10.1 erzeugte Folge. Ist $\bar{x} \in \mathbb{R}^n$ kein stationärer Punkt von $f$, so gilt für jede gegen $\bar{x}$ konvergente Teilfolge $\{x^k\}_{k\in K}$*

$$\liminf_{K\ni k\to\infty} \Delta_k > 0.$$

*Beweis.* Sei $\bar{x} \in \mathbb{R}^n$ mit $\nabla f(\bar{x}) \neq 0$. Angenommen, es gibt eine Teilfolge $\{x^k\}_{k\in K}$, so dass die entsprechende Folge $\{\Delta_k\}_{k\in K}$ den Häufungspunkt 0 hat. Durch eventuellen Übergang zu einer weiteren Teilfolge – immer noch mit $k \in K$ bezeichnet – können wir annehmen, dass gilt

$$\lim_{K\ni k\to\infty} \Delta_k = 0.$$

Da jeder erfolgreiche Schritt mindestens den Radius auf $\Delta_{\min} > 0$ zurücksetzt, ist dies nur möglich, wenn ein $k_0 \in \mathbb{N}$ existiert, so dass alle Schritte für $k \geq k_0$ verworfen werden. Dies setzt aber voraus, dass gilt

$$\rho_k < \eta_1 < 1 \qquad \text{für alle } k \in K, k \geq k_0. \tag{10.6}$$

Diese Ungleichung führen wir nun zum Widerspruch. Dazu betrachten wir

$$|\rho_k - 1| = \frac{|q_k(s^k) - f(x^k + s^k)|}{|f(x^k) - q_k(s^k)|}$$

und zeigen, dass die rechte Seite für $k \to \infty$ gegen 0 geht. Dafür verwenden wir, dass $\bar{x}$ nach Voraussetzung kein stationärer Punkt ist, d. h. ein $\beta_1 > 0$ existiert mit

$$\|\nabla f(x^k)\| \geq \beta_1 \qquad \text{für alle } k \in K.$$

Weiter ist $\{x^k\}_{k\in K}$ konvergent und $\nabla^2 f$ stetig und daher beschränkt, es gibt also ein $\beta_2 > 0$ mit

$$\|\nabla^2 f(x^k)\| \leq \beta_2 \qquad \text{für alle } k \in K.$$

Nach Satz 1.1 existiert nun für alle $k \in K$ ein $\xi^k = x^k + \theta_k s^k$ mit $\theta_k \in (0,1)$ und $f(x^k + s^k) = f(x^k) + \nabla f(\xi^k)^T s^k$. Daraus folgt

$$\begin{aligned} |q_k(s^k) - f(x^k + s^k)| &= \left|\nabla f(x^k)^T s^k + \tfrac{1}{2}(s^k)^T \nabla^2 f(x^k) s^k - \nabla f(\xi^k)^T s^k\right| \\ &\leq \|\nabla f(x^k) - \nabla f(\xi^k)\| \|s^k\| + \frac{\beta_2}{2}\|s^k\|^2. \end{aligned}$$

Nach Lemma 10.1 gilt nun wegen $\|s^k\| \leq \Delta_k \to 0$ für alle hinreichend großen $k \in K$

$$\begin{aligned} f(x^k) - q_k(s^k) &\geq \frac{1}{2}\|\nabla f(x^k)\| \min\left\{\Delta_k, \tfrac{\|\nabla f(x^k)\|}{\|\nabla^2 f(x^k)\|}\right\} \\ &\geq \frac{1}{2}\beta_1 \min\left\{\Delta_k, \tfrac{\beta_1}{\beta_2}\right\} \\ &= \frac{1}{2}\beta_1 \Delta_k \geq \frac{1}{2}\beta_1 \|s^k\|. \end{aligned}$$

Zusammen erhalten wir also

$$|\rho_k - 1| \leq \frac{1}{\beta_1}\left(2\|\nabla f(x^k) - \nabla f(\xi^k)\| + \beta_2 \|s^k\|\right).$$

Der zweite Term in Klammern konvergiert wegen $\|s^k\| \leq \Delta_k \to 0$, der erste wegen der Stetigkeit von $\nabla f$ und $x^k \to \bar{x}$ sowie $\xi^k = x^k + \theta_k s^k \to \bar{x} + 0 = \bar{x}$. Also gilt $\rho_k \to 1$, im Widerspruch zu (10.6). □

Mit Hilfe von Lemmata 10.1 und 10.2 können wir nun die globale Konvergenz von Algorithmus 10.1 zeigen.

Satz 10.3. *Sei $f : \mathbb{R}^n \to \mathbb{R}$ zweimal stetig differenzierbar. Dann bricht Algorithmus 10.1 entweder nach endlich vielen Schritten ab, oder jeder Häufungspunkt von $\{x^k\}_{k\in\mathbb{N}}$ ist ein stationärer Punkt von $f$.*

*Beweis.* Im Falle eines endlichen Abbruchs ist nichts zu zeigen. Sei daher $\nabla f(x^k) \neq 0$ für alle $k \in \mathbb{N}$ und sei $\{x^k\}_{k\in K}$ eine gegen $\bar{x} \in \mathbb{R}^n$ konvergente Teilfolge, so dass $\bar{x}$ kein stationärer Punkt ist. Wir zeigen zuerst, dass in dieser Teilfolge unendlich viele erfolgreiche Iterationsschritte enthalten sind: Wäre dies nicht der Fall, gäbe es ein $k_0 \in K$ so dass für

alle $k \geq k_0$ der Schritt verworfen wird. Dann wird aber auch für alle $k \geq k_0$ der Radius verkleinert, woraus $\Delta_k \to 0$ folgt, im Widerspruch zu Lemma 10.2. Also ist entweder $\bar{x}$ doch ein stationärer Punkt (und wir sind fertig), oder es sind unendlich viele Schritte erfolgreich. Da für nicht erfolgreiche Schritte $x^{k+1} = x^k$ gilt, können wir (durch Übergang zu einer weiteren Teilfolge) sogar annehmen, dass alle Schritte $x^k$, $k \in K$, erfolgreich sind.

Aufgrund der Stetigkeit von $\nabla f$ und $\nabla^2 f$ existieren nun wieder Konstanten $\beta_1, \beta_2 > 0$ mit

$$\|\nabla f(x^k)\| \geq \beta_1, \quad \text{und} \quad \|\nabla^2 f(x^k)\| \leq \beta_2 \qquad \text{für alle } k \in K.$$

Da alle Schritte erfolgreich sind, muss außerdem $\rho_k \geq \eta_1$ für alle $k \in K$ gelten. Mit Lemma 10.1 folgt daher für alle $k \in K$

$$\begin{aligned} f(x^k) - f(x^k + s^k) &\geq \eta_1 (f(x^k) - q_k(s^k)) \\ &\geq \eta_1 \frac{1}{2} \|\nabla f(x^k)\| \min\left\{\Delta_k, \frac{\|\nabla f(x^k)\|}{\|\nabla^2 f(x^k)\|}\right\} \\ &\geq \frac{\eta_1 \beta_1}{2} \min\left\{\Delta_k, \frac{\beta_1}{\beta_2}\right\}. \end{aligned} \tag{10.7}$$

Weiterhin ist nach Konstruktion die Folge $\{f(x^k)\}_{k\in\mathbb{N}}$ monoton fallend, da Algorithmus 10.1 nur Abstiegsschritte akzeptiert. Nun konvergiert nach Annahme die Teilfolge $x^k \to \bar{x}$ und damit wegen der Stetigkeit von $f$ auch $f(x^k) \to f(\bar{x})$. Wegen der Monotonie ist das aber der einzige Häufungspunkt von $\{f(x^k)\}_{k\in\mathbb{N}}$, so dass die gesamte Folge konvergiert. Also gilt

$$f(x^k) - f(x^k + s^k) = f(x^k) - f(x^{k+1}) \to 0$$

und damit wegen (10.7) auch $\Delta_k \to 0$ für $K \ni k \to \infty$, im Widerspruch zu Lemma 10.2. Also muss $\bar{x}$ ein stationärer Punkt sein. □

Analog zu Satz 8.4 können wir unter einer Optimalitätsbedingung zweiter Ordnung Konvergenz der ganzen Folge zeigen.

Satz 10.4. *Sei $f : \mathbb{R}^n \to \mathbb{R}$ zweimal stetig differenzierbar, und sei $\bar{x} \in \mathbb{R}^n$ ein Häufungspunkt der durch Algorithmus 10.1 erzeugten Folge $\{x^k\}_{k\in\mathbb{N}}$ mit $\nabla^2 f(\bar{x})$ positiv definit. Dann ist $\bar{x}$ ein strikter lokaler Minimierer und $\{x^k\}_{k\in\mathbb{N}}$ konvergiert gegen $\bar{x}$.*

*Beweis.* Nach Satz 10.3 ist jeder Häufungspunkt von $\{x^k\}_{k\in\mathbb{N}}$ ein stationärer Punkt und damit insbesondere $\nabla f(\bar{x}) = 0$. Zusammen mit der positiven Definitheit von $\nabla^2 f(\bar{x})$ folgt aus Satz 3.4, dass $\bar{x}$ ein strikter lokaler Minimierer ist.

Weiter ist $\nabla^2 f(\bar{x})$ insbesondere regulär; wegen Lemma 7.3 gilt daher $\nabla f(x) \neq 0$ für alle $x \neq \bar{x}$ hinreichend nahe bei $\bar{x}$. Also ist $\bar{x}$ ein isolierter Häufungspunkt (denn jeder weitere Häufungspunkt wäre nach Satz 10.3 wieder ein stationärer Punkt). Sei nun $\{x^k\}_{k\in K}$ eine Teilfolge mit $x^k \to \bar{x}$. Wegen Lemma 7.6 existieren $k_0 \in \mathbb{N}$ und $\mu > 0$ mit

$$(s^k)^T \nabla^2 f(x^k) s^k \geq \mu \|s^k\|^2 \qquad \text{für alle } k \in K, k \geq k_0.$$

Weiter folgt aus Lemma 10.1 insbesondere $f(x^k) - q_k(s^k) \geq 0$ und daher

$$f(x^k) + \nabla f(x^k)^T s^k + \frac{1}{2}(s^k)^T \nabla^2 f(x^k) s^k = q_k(s^k) \leq f(x^k).$$

Zusammen erhalten wir für alle $k \in K$ mit $k \geq k_0$

$$\frac{\mu}{2}\|s^k\|^2 \leq \frac{1}{2}(s^k)^T \nabla^2 f(x^k) s^k \leq -\nabla f(x^k)^T s^k \leq \|\nabla f(x^k)\| \|s^k\|.$$

Da aber die Teilfolge $\{x^k\}_{k\in K}$ nach Satz 10.3 gegen den stationären Punkt $\bar{x}$ konvergiert und $\nabla f$ stetig ist, folgt daraus

$$\|x^{k+1} - x^k\| = \|s^k\| \leq \frac{2}{\mu}\|\nabla f(x^k)\| \to 0.$$

Nach Lemma 8.3 impliziert dies die Konvergenz der gesamten Folge $\{x^k\}_{k\in\mathbb{N}}$ gegen $\bar{x}$. □

Für die lokale superlineare Konvergenz zeigen wir wieder, dass Algorithmus 10.1 irgendwann in das Newton-Verfahren übergeht. Dafür beweisen wir zuerst, dass ab einem gewissen Punkt *alle* Schritte (und nicht nur unendlich viele) erfolgreich sind.

Lemma 10.5. *Sei $f : \mathbb{R}^n \to \mathbb{R}$ zweimal stetig differenzierbar, und sei $\bar{x} \in \mathbb{R}^n$ ein Häufungspunkt der durch Algorithmus 10.1 erzeugten Folge $\{x^k\}_{k\in\mathbb{N}}$ mit $\nabla^2 f(\bar{x})$ positiv definit. Dann existiert ein $k_0 \in \mathbb{N}$, so dass alle Schritte $k \geq k_0$ erfolgreich sind.*

*Beweis.* Wir gehen ähnlich vor wie im Beweis von Lemma 10.2 und zeigen $\rho_k \to 1$, nur dass wir diesmal die Annahme $\Delta_k \to 0$ natürlich nicht verwenden können. Nach Satz 10.4 konvergiert unter den genannten Voraussetzungen die gesamte Folge gegen $\bar{x}$, und wie in dessen Beweis gezeigt existieren $\mu > 0$ und $k_0 \in \mathbb{N}$ mit

$$\|s^k\| \leq \frac{2}{\mu}\|\nabla f(x^k)\| \qquad \text{für alle } k \geq k_0. \tag{10.8}$$

Wegen der Stetigkeit von $\nabla^2 f$ existiert weiter eine Konstante $c > 0$ mit

$$\|\nabla^2 f(x^k)\| \leq c \qquad \text{für alle } k \geq k_0.$$

Aus Lemma 10.1 zusammen mit $\|s^k\| \leq \Delta_k$ folgt nun

$$\begin{aligned} f(x^k) - q_k(s^k) &\geq \frac{1}{2}\|\nabla f(x^k)\| \min\left\{\Delta_k, \tfrac{\|\nabla f(x^k)\|}{\|\nabla^2 f(x^k)\|}\right\} \\ &\geq \frac{\mu}{4}\|s^k\| \min\left\{\|s^k\|, \tfrac{\mu}{2c}\|s^k\|\right\} \\ &= \kappa\|s^k\|^2 \end{aligned}$$

mit $\kappa := \frac{\mu}{4}\min\{1, \frac{\mu}{2c}\}$. Außerdem existiert nach Satz 1.2 für alle $k \in K$ ein $\xi^k = x^k + \theta_k s^k$ mit $\theta_k \in (0,1)$ so dass gilt

$$\begin{aligned}|q_k(s^k) - f(x^k + s^k)| &= \frac{1}{2}\left|(s^k)^T(\nabla^2 f(\xi^k) - \nabla^2 f(x^k))s^k\right| \\ &\leq \frac{1}{2}\|s^k\|^2\|\nabla^2 f(\xi^k) - \nabla^2 f(x^k)\|.\end{aligned}$$

Zusammen erhalten wir

$$|\rho_k - 1| = \frac{|q_k(s^k) - f(x^k + s^k)|}{|f(x^k) - q_k(s^k)|} \leq \frac{1}{2\kappa}\|\nabla^2 f(\xi^k) - \nabla^2 f(x^k)\| \to 0,$$

da wegen $x^k \to \bar{x}$ mit $\nabla f(\bar{x}) = 0$ und (10.8) auch $\xi^k \to \bar{x}$ konvergiert. Also ist wegen $\eta_1 < 1$ für $k \in \mathbb{N}$ hinreichend groß stets $\rho_k \geq \eta_1$, d. h. alle Schritte sind erfolgreich. □

Wir können nun die lokale superlineare Konvergenz beweisen.

Satz 10.6. *Sei $f : \mathbb{R}^n \to \mathbb{R}$ zweimal stetig differenzierbar, und sei $\bar{x} \in \mathbb{R}^n$ ein Häufungspunkt der durch Algorithmus 10.1 erzeugten Folge $\{x^k\}_{k\in\mathbb{N}}$ mit $\nabla^2 f(\bar{x})$ positiv definit. Dann konvergiert $x^k \to \bar{x}$ lokal superlinear. Ist $\nabla^2 f$ darüber hinaus lokal Lipschitz-stetig, so konvergiert $x^k \to \bar{x}$ quadratisch.*

*Beweis.* Es bleibt nur noch zu zeigen, dass irgendwann der beschränkte Minimierer des quadratischen Modells mit dem Newton-Schritt übereinstimmt. Nach Satz 10.4 konvergiert die gesamte Folge $x^k \to \bar{x}$; wegen Lemma 7.6 existiert daher ein $k_0 \in \mathbb{N}$, so dass $\nabla^2 f(x^k)$ positiv definit ist für alle $k \geq k_0$. Also ist das quadratische Modell $q_k$ für alle $k \geq k_0$ strikt konvex, und die nach Satz 3.5 notwendige und hinreichende Optimalitätsbedingung $\nabla q_k(\bar{s}^k) = 0$ ist identisch mit dem Newton-Schritt

$$\bar{s}^k = -\nabla^2 f(x^k)^{-1}\nabla f(x^k).$$

Nach Lemma 7.5 existiert weiterhin ein $k_1 \in \mathbb{N}$ und ein $c > 0$ mit

$$\|\nabla^2 f(x^k)^{-1}\| \leq c \qquad \text{für alle } k \geq k_1.$$

Da $x^k$ gegen den stationären Punkt $\bar{x}$ konvergiert, folgt daraus für $k \geq \max\{k_0, k_1\}$

$$\|\bar{s}^k\| \leq c\|\nabla f(x^k)\| \to 0.$$

Also muss ein $k_2 \in \mathbb{N}$ existieren mit $\|\bar{s}^k\| \leq \Delta_{\min}$ für alle $k \geq k_2$.

Andererseits sind nach Lemma 10.5 für ein $k_3 \in \mathbb{N}$ alle Schritte $k \geq k_3$ erfolgreich; aufgrund der Iterationsvorschrift gilt daher $\Delta_k \geq \Delta_{k_3} \geq \Delta_{\min} > 0$ für alle $k \geq k_3$. Also stimmt für alle $k \geq \max\{k_2, k_3\}$ der Newton-Schritt $\bar{s}^k \in K_{\Delta_k}$ mit der Lösung $s^k$ von (10.2) überein. Damit ist für alle $k \geq \max\{k_2, k_3\}$ der Algorithmus 10.1 identisch mit dem Newton-Verfahren und hat daher die selbe Konvergenzgeschwindigkeit. □

## 10.2 ZUR BERECHNUNG DES TRUST-REGION-SCHRITTES

Noch offen ist die Frage, wie man in jeder Iteration von Algorithmus 10.1 den globalen Minimierer des quadratischen Modells $q_k$ über den Vertrauensbereich $K_{\Delta_k}$ berechnen kann. Prinzipiell handelt es sich dabei um ein beschränktes Optimierungsproblem, wobei die quadratische Struktur von $q_k$ und die Kugelgestalt von $K_{\Delta_k}$ eine effiziente Lösung erlaubt. Andererseits haben wir gesehen, dass die Konvergenz von Algorithmus 10.1 lediglich voraussetzt, dass der Schritt $s^k$ einen ausreichend großen vorausgesagten Abstieg produziert; siehe Lemma 10.1. Es genügt also, für $s^k$ eine *Näherungslösung* von (10.2) zu verwenden, die diese Bedingung erfüllt. Im Rest dieses Kapitel sollen drei der am häufigsten verwendeten Ansätze kurz vorgestellt werden.

### DER CAUCHY-PUNKT

Da wir im Beweis von Lemma 10.1 nur zulässige Richtungen der Form $d = -\lambda \nabla f(x^k)$, $\lambda \in (0, \infty)$, gewählt haben, genügt es offenbar, das quadratische Modell nur entlang dieser Richtung zu minimieren. Statt (10.2) setzen wir also $s^k = -\sigma_k \nabla f(x^k)$, wobei $\sigma_k$ Lösung ist des eingeschränkten Problems

$$\min_{\sigma \geq 0} q_k(-\sigma \nabla f(x^k)) \qquad \text{mit } \|\sigma \nabla f(x^k)\| \leq \Delta_k.$$

Da $q_k$ quadratisch ist, ist die globale Lösung dieses Problems gegeben durch

$$\sigma_k = \begin{cases} \dfrac{\Delta_k}{\|\nabla f(x^k)\|} & \text{falls } \nabla f(x^k)^T \nabla^2 f(x^k) \nabla f(x^k) \leq 0, \\ \min\left\{ \dfrac{\|\nabla f(x^k)\|^2}{\nabla f(x^k)^T \nabla^2 f(x^k) \nabla f(x^k)}, \dfrac{\Delta_k}{\|\nabla f(x^k)\|} \right\} & \text{sonst,} \end{cases} \tag{10.9}$$

vergleiche das Gradientenverfahren mit exakter Schrittweite (6.2). Der Punkt

$$x_C := x^k - \sigma_k \nabla f(x^k)$$

wird *Cauchy-Punkt* genannt.

Wörtlich wie in Lemma 10.1 beweist man nun

Lemma 10.7. *Sei $s^k = -\sigma_k \nabla f(x^k)$ mit $\sigma_k$ gegeben durch (10.9). Dann gilt*

$$f(x^k) - q_k(s^k) \geq \frac{1}{2} \|\nabla f(x^k)\| \min\left\{ \Delta_k, \frac{\|\nabla f(x^k)\|}{\|\nabla^2 f(x^k)\|} \right\}.$$

Daraus folgt wie zuvor die globale Konvergenz von Algorithmus 10.1, wenn für $x^{k+1}$ der Cauchy-Punkt gewählt wird.

Satz 10.8. *Sei $f : \mathbb{R}^n \to \mathbb{R}$ zweimal stetig differenzierbar. Dann bricht Algorithmus 10.1 mit $x^{k+1} := x_C$ entweder nach endlich vielen Schritten ab, oder jeder Häufungspunkt von $\{x^k\}_{k\in\mathbb{N}}$ ist ein stationärer Punkt von $f$.*

Dieser Ansatz ist einfach zu implementieren, hat aber den Nachteil, dass der Cauchy-Punkt einem (möglicherweise gedämpften) Gradientenschritt entspricht. Im besten Fall wird das Verfahren daher in das Gradientenverfahren übergehen, weshalb man auch nur dessen (niedrige) Konvergenzgeschwindigkeit erwarten kann.

### DER DOGLEG-SCHRITT

Um lokal superlineare Konvergenz zu erhalten, müssen wir also irgendwie den Newton-Schritt ins Spiel bringen. Ein Ansatz dafür ist, den Minimierer des quadratischen Modells nicht nur entlang des Gradientenschrittes zu suchen, sondern entlang eines "geknickten" Schritts, der vom optimalen (unbeschränkten) Gradientenschritt

$$d_G := -\frac{\|\nabla f(x^k)\|^2}{\nabla f(x^k)^T \nabla^2 f(x^k) \nabla f(x^k)} \nabla f(x^k)$$

weiter entlang des Newton-Schritts

$$d_N := -\nabla^2 f(x^k)^{-1} \nabla f(x^k)$$

(falls existent) geht. Wir suchen daher den Minimierer von $q_l$ entlang des Pfades

$$d : [0,2] \to \mathbb{R}^n, \qquad d(\tau) := \begin{cases} \tau d_G & \text{für } \tau \in [0,1], \\ d_G + (\tau - 1)(d_N - d_G) & \text{für } \tau \in [1,2], \end{cases}$$

unter der Beschränkung $\|d(\tau)\| \leq \Delta_k$. Der Pfad $d(\tau)$ wird *Dogleg-Pfad* genannt, da er wegen seinem Knick (manche) an das Bein eines Hundes erinnert.

Das folgende Lemma zeigt, dass eine Suche entlang dieses Pfades sinnvoll ist.

Lemma 10.9. *Sei $\nabla^2 f(x^k)$ positiv definit. Dann gilt:*

*(i) die Funktion $\tau \mapsto \|d(\tau)\|$ ist auf $[0,2]$ monoton steigend;*

*(ii) die Funktion $\tau \mapsto q_k(d(\tau))$ ist auf $[0,2]$ monoton fallend.*

*Beweis.* Beide Eigenschaften sind für $\tau \in [0,1]$ erfüllt, da $\|d(\tau)\| = \tau \|d_G\|$ offensichtlich monoton und $d_G$ der unbeschränkte Minimierer von $q_k$ entlang $\nabla f(x^k)$ ist. Es ist daher nur das Intervall $[1,2]$ zu untersuchen. Wir setzen im Folgenden kurz $g := \nabla f(x^k)$ und $H := \nabla^2 f(x^k)$.

Zu (i): Wir betrachten für $t \in [0,1]$ die Funktion

$$\varphi(t) = \frac{1}{2}\|d(1+t)\|^2 = \frac{1}{2}\|d_G + t(d_N - d_G)\|^2$$

und zeigen $\varphi'(t) \geq 0$ für $t \in (0,1)$. Ausmultiplizieren und Ableiten ergibt

$$\begin{aligned}\varphi'(t) &= d_G^T(d_N - d_G) + t\|d_N - d_G\|^2 \geq d_G^T(d_N - d_G) \\ &= -\frac{\|g\|^2}{g^THg}\left(-g^TH^{-1}g + \frac{\|g\|^4}{g^THg}\right) = \|g\|^2\frac{g^TH^{-1}g}{g^THg}\left(1 - \frac{\|g\|^4}{(g^TH^{-1}g)(g^THg)}\right).\end{aligned}$$

Da $H$ und damit auch $H^{-1}$ positiv definit sind, ist der Term vor der Klammer positiv. Mit Hilfe der Cholesky-Zerlegung $H = R^TR$ mit $R$ invertierbar erhalten wir wegen $H^{-1} = R^{-1}R^{-T}$ auch

$$\begin{aligned}\|g\|^2 &= g^Tg = g^TR^{-1}Rg = (R^{-T}g)^T(Rg) \\ &\leq \|R^{-T}g\|\|Rg\| = \left((R^{-T}g)^T(R^{-T}g)\right)^{1/2}\left((Rg)^T(Rg)\right)^{1/2} \\ &= \left(g^T(R^{-1})(R^{-T})g\right)^{1/2}\left(g^TR^TRg\right)^{1/2} = (g^TH^{-1}g)^{1/2}(g^THg)^{1/2}.\end{aligned}$$

Also ist

$$\frac{\|g\|^4}{(g^TH^{-1}g)(g^THg)} \leq 1$$

und damit $\varphi'(t) \geq 0$.

Zu (ii): Wir betrachten analog für $t \in [0,1]$ die Funktion

$$\begin{aligned}\psi(t) &= q_k(d(1+t)) \\ &= f(x^k) + g^T(d_G + t(d_N - d_G)) + \frac{1}{2}(d_G + t(d_N - d_G))^TH(d_G + t(d_N - d_G))\end{aligned}$$

und zeigen $\psi'(t) \leq 0$ für $t \in (0,1)$. Ausmultiplizieren und Ableiten ergibt wegen der Symmetrie und positiven Definitheit von $H$ sowie $t \leq 1$

$$\begin{aligned}\psi'(t) &= g^T(d_N - d_G) + d_G^TH(d_N - d_G) + t(d_N - d_G)^TH(d_N - d_G) \\ &\leq g^T(d_N - d_G) + d_G^TH(d_N - d_G) + (d_N - d_G)^TH(d_N - d_G) \\ &= (g + Hd_N)^T(d_N - d_G) \\ &= 0,\end{aligned}$$

da nach Definition des Newton-Schrittes gilt $d_N = -H^{-1}g$. □

Aus dem Lemma folgt, dass genau zwei Fälle auftreten können:

(i) Der Newton-Schritt $d_N$ liegt im Vertrauensbereich; in diesem Fall ist $d_N$ der eindeutige Minimierer von $q_k$ entlang dem Dogleg-Pfad $d(\tau)$;

(ii) Der Newton-Schritt $d_N$ liegt außerhalb des Vertrauensbereichs; in diesem Fall gibt es (wegen der Stetigkeit von $\tau \mapsto \|d(\tau)\|$) genau einen Punkt $d(\tau^*)$, in dem der Dogleg-Pfad den Rand des Vertrauensbereichs kreuzt.

Wir unterscheiden in diesem Fall weiter

(iia) $\tau^* \leq 1$: dann ist $x^k + d(\tau^*)$ genau der Cauchy-Punkt;

(iib) $\tau^* > 1$: dann ist $\tau^* = 1 + t^*$, wobei $t^*$ gewählt ist als die einzige positive Nullstelle des quadratischen Polynoms

$$\begin{aligned} r(t) &= \|d_G + t(d_N - d_G)\|^2 - \Delta_k^2 \\ &= \|d_N - d_G\|^2 t^2 + 2d_G^T(d_N - d_G)t + \|d_G\|^2 - \Delta_k^2 \end{aligned}$$

(wofür eine geschlossene Formel existiert).

Wir können also den *Dogleg-Schritt* wie folgt bestimmen.

**Algorithmus 10.2 :** Dogleg-Schritt

**if** $\|d_G\| \geq \Delta_k$ **then** `// Gradientenschritt nicht in Vertrauensbereich`
  Setze $s^k = \frac{\Delta_k}{\|d_G\|} d_G$ `// Akzeptiere Cauchy-Punkt`
**else if** $\|d_N\| \leq \Delta_k$ **then** `// Newton-Schritt im Vertrauensbereich`
  Setze $s^k = d_N$ `// Akzeptiere Newton-Schritt`
**else** `// Pfad schneidet Rand zwischen` $d_G$ `und` $d_N$
  Bestimme positive Nullstelle $t^* \in (0, 1)$ von $r(t)$
  Setze $s^k = d(1 + t^*)$ `// Gehe zum Rand des Vertrauensbereichs`

Da nach Lemma 10.9 der Funktionswert des quadratischen Modells entlang des Dogleg-Pfads monoton abfällt und wir nach Konstruktion auf dem Pfad mindestens bis zum Cauchy-Punkt gehen, ist der vorausgesagte Abstieg für den Dogleg-Schritt stets mindestens so groß wie für den Cauchy-Punkt. Aus Lemma 10.7 folgt daher die globale Konvergenz des Dogleg-Trust-Region-Verfahrens. Genau wie im Beweis von Satz 10.6 zeigt man nun, dass deshalb irgendwann der Newton-Schritt stets im Vertrauensbereich liegt und daher als Dogleg-Schritt akzeptiert wird, woraus die lokale superlineare Konvergenz folgt.

Das Verfahren funktioniert in der Form allerdings nur, falls $\nabla^2 f(x^k)$ immer positiv definit ist; man kann es aber so modifizieren, dass für $\nabla f(x^k)^T \nabla^2 f(x^k) \nabla f(x^k) \leq 0$ statt dem Dogleg-Schritt der Cauchy-Punkt verwendet wird. Auch in diesem Fall kann man globale Konvergenz und – falls $\nabla^2 f(\bar{x})$ positiv definit ist – lokal superlineare Konvergenz zeigen; siehe [Kelley 1999, Abschnitt 3.3.6].

### 10.2.1 INEXAKTE TRUST-REGION-VERFAHREN

Eine besonders effiziente Variante dieser Idee verbindet den Trust-Region-Ansatz mit dem inexakten Newton-Verfahren, in dem die Newton-Gleichung $\nabla^2 f(x^k)s = -\nabla f(x^k)$ näherungsweise mit Hilfe eines iterativen Verfahrens gelöst wird. Im Trust-Region-Verfahren wird dieses iterative Verfahren nun so modifiziert, dass die Iteration den Vertrauensbereich nicht verlässt. Produziert das iterative Verfahren eine Folge $\{s^m\}_{m\in\mathbb{N}}$ von Näherungslösungen der Newton-Gleichung, so wird grob vereinfacht für jeden Schritt überprüft:

(i) Ist $\|\nabla^2 f(x^k)s^m + \nabla f(x^k)\| \leq \eta_k \|\nabla f(x^k)\|$ für eine vorgegebene Toleranz $\eta_k > 0$, so setze $s^k := s^m$ und beende die Iteration.

(ii) Ist $\|s^m\| \geq \Delta_k$ (oder kann $s^m$ aus irgendeinem Grund nicht als Näherungslösung vertraut werden, etwa wenn $\nabla^2 f(x^k)$ als nicht positiv definit erkannt wird), so bestimme analog zum Dogleg-Schritt $s^k$ als denjenigen Punkt zwischen $s^{m-1}$ und $s^m$, für den $\|s^k\| = \Delta_k$ gilt.

(iii) Ansonsten fahre mit der Iteration fort.

Verwendet man als iteratives Verfahren das CG-Verfahren, so kann man zeigen, dass dadurch $s^k$ stets einen mindestens so großen vorausgesagten Abstieg erzeugt wie der Cauchy-Punkt, woraus die globale Konvergenz folgt. Ähnlich wie für den Dogleg-Schritt zeigt man dann $\|s^k\| \to 0$, so dass das inexakte Trust-Region-Verfahren irgendwann in das inexakte Newton-Verfahren übergeht, welches für $\eta_k \to 0$ lokal superlinear konvergiert. Für Details (die auf spezifischen Eigenschaften des CG-Verfahrens beruhen) sei auf [Geiger & Kanzow 1999, Kapitel 14.7] verwiesen.

# Teil III

# LINEARE OPTIMIERUNG

# 11 KONVEXE MENGEN UND POLYEDER

Wir wenden uns nun der Optimierung *mit* Nebenbedingungen zu; konkret der Minimierung einer *linearen* Funktion unter *linearen* Nebenbedingungen, siehe (LP). Es ist klar, dass die Schwierigkeit dabei in der Nebenbedingung $Ax \leq b$ liegt; das Studium der zulässigen Menge $X := \{x \in \mathbb{R}^n : Ax \leq b\}$ ist daher von fundamentaler Wichtigkeit. Hier und in Folge schreiben wir $c \geq 0$ für einen Vektor $c \in \mathbb{R}^n$ genau dann, wenn $c_i \geq 0$ für alle $1 \leq i \leq n$ gilt.

## 11.1 TRENNUNG KONVEXER MENGEN

Ähnlich wie konvexe Funktionen eine besonders günstige Klasse von Zielfunktionen darstellen, spielen konvexe Mengen als zulässige Mengen eine fundamentale Rolle in der restringierten Optimierung. Zur Erinnerung: eine Menge $M \subset \mathbb{R}^n$ heißt *konvex*, falls gilt

$$\lambda x + (1-\lambda)y \in M \qquad \text{für alle } x, y \in M \text{ und } \lambda \in [0,1].$$

Ein fundamentales Hilfsmittel ist das folgende prototypische restringierte konvexe Optimierungsproblem.

Lemma 11.1. *Sei $M \subset \mathbb{R}^n$ nichtleer, konvex, und abgeschlossen und $z \in \mathbb{R}^n$ beliebig. Dann hat das Problem*

$$\min_{x \in M} \|x - z\| \tag{11.1}$$

*eine eindeutige Lösung $\bar{x} \in M$, genannt* Projektion von $z$ auf $M$.

*Beweis.* Da $M$ nichtleer ist, existiert ein $x_0 \in M$. Setze nun $C := \|x_0 - z\|$. Offensichtlich ist $\bar{x}$ Lösung von (11.1) genau dann, wenn gilt

$$\|\bar{x} - z\|^2 \leq \|x - z\|^2 =: f(x) \qquad \text{für alle } x \in M.$$

Auf dieses Problem wenden wir Satz 2.2 an. Die Stetigkeit und Koerzivität von $f$ folgt sofort aus der Definition. Bleibt nur noch die strikte Konvexität von $f$ zu zeigen. Seien dafür $x, y \in \mathbb{R}^n$ mit $x \neq y$ und $\lambda \in (0,1)$ beliebig. Dann gilt

$$\begin{aligned}
\|\lambda x + (1-\lambda)y\|^2 &= (\lambda x + (1-\lambda)y)^T(\lambda x + (1-\lambda)y) \\
&= \lambda^2 x^T x + 2\lambda(1-\lambda)x^T y + (1-\lambda)^2 y^T y \\
&= \lambda\Big(\lambda x^T x - (1-\lambda)(x-y)^T x + (1-\lambda)y^T y\Big) \\
&\quad + (1-\lambda)\Big(\lambda x^T x + \lambda(x-y)^T y + (1-\lambda)y^T y\Big) \\
&= (\lambda + (1-\lambda))\Big(\lambda x^T x + (1-\lambda)y^T y\Big) - \lambda(1-\lambda)(x-y)^T(x-y) \\
&= \lambda\|x\|^2 + (1-\lambda)\|y\|^2 - \lambda(1-\lambda)\|x-y\|^2 \\
&< \lambda\|x\|^2 + (1-\lambda)\|y\|^2.
\end{aligned}$$

Unter den genannten Annahmen an $M$ liefert Satz 2.2 nun einen eindeutigen Minimierer $\bar{x} \in M$ von $f$ und damit die gesuchte Projektion. □

Die folgende Charakterisierung der Projektion ist unser erstes Beispiel einer Optimalitätsbedingung für restringierte Optimierungsprobleme (vergleiche die Aussage von Satz 3.1).

Satz 11.2 (Projektionssatz). *Sei $M \subset \mathbb{R}^n$ nichtleer, konvex, und abgeschlossen und $z \in \mathbb{R}^n$ beliebig. Dann ist $\bar{x} \in M$ die Projektion von $z$ auf $M$ genau dann, wenn gilt*

$$(\bar{x} - z)^T(x - \bar{x}) \geq 0 \qquad \text{für alle } x \in M. \tag{11.2}$$

*Beweis.* Sei $\bar{x}$ die Projektion von $z$ auf $M$ aus Lemma 11.1 und $x \in M$ beliebig. Da $M$ konvex ist, ist dann auch $x_\lambda := \bar{x} + \lambda(x - \bar{x}) \in M$ für alle $\lambda \in (0,1)$. Aus der Optimalität von $\bar{x}$ folgt daher

$$\|\bar{x} - z\|^2 \leq \|x_\lambda - z\|^2 = \|(\bar{x} - z) + \lambda(x - \bar{x})\|^2.$$

Aus der binomischen Formel folgt dann nach Umsortieren

$$0 \leq 2\lambda(\bar{x} - z)^T(x - \bar{x}) + \lambda^2\|x - \bar{x}\|^2.$$

Division durch $2\lambda > 0$ und Grenzübergang $\lambda \longrightarrow 0$ ergibt nun (11.2).

Sei umgekehrt $\bar{x} \in M$ so, dass (11.2) gilt, und $x \in M$ beliebig. Aus (11.2) folgt dann zusammen mit der Cauchy–Schwarz-Ungleichung

$$\begin{aligned}
0 \geq (z - \bar{x})^T(x - \bar{x}) = (z - \bar{x})^T(x - z + z - \bar{x}) &= \|z - \bar{x}\|^2 + (z - \bar{x})^T(x - z) \\
&\geq \|z - \bar{x}\|^2 - \|z - \bar{x}\|\|x - z\|.
\end{aligned}$$

Ist nun $\bar{x} \neq z$ (ansonsten gilt (11.2) trivialerweise), so folgt daraus

$$\|z - \bar{x}\| \leq \|x - z\| \qquad \text{für alle } x \in M,$$

d. h. $\bar{x} \in M$ ist nach Definition die Projektion von $z$ auf $M$. □

Wir können nun den folgenden fundamentalen Satz über konvexe Mengen beweisen, der einen Spezialfall der *Trennungssätze von Hahn–Banach* darstellt. Man zeigt leicht durch einfache geometrische Beispiele, dass keine der Voraussetzungen fallengelassen werden kann.

Satz 11.3 (Trennungssatz). *Sei $M \subset \mathbb{R}^n$ nichtleer, abgeschlossen, und konvex und $x_0 \in \mathbb{R}^n \setminus M$. Dann existieren ein $a \in \mathbb{R}^n \setminus \{0\}$ und ein $\alpha \in \mathbb{R}$ mit*

$$a^T x_0 > \alpha \geq a^T x \qquad \text{für alle } x \in M. \tag{11.3}$$

*Beweis.* Wir verwenden die Projektion $\bar{x} \in M$ von $x_0$ auf $M$ – die nach Lemma 11.1 unter den genannten Voraussetzungen existiert – und setzen $a := x_0 - \bar{x} \neq 0$ (wegen $x_0 \notin M$) und $\alpha := a^T \bar{x}$. Dann gilt

$$0 < \|a\|^2 = a^T(x_0 - \bar{x}) = a^T x_0 - \alpha$$

und damit die erste Ungleichung in (11.3). Für die zweite Ungleichung verwenden wir, dass nach Satz 11.2 für alle $x \in M$ gilt

$$0 \leq (\bar{x} - x_0)^T (x - \bar{x}) = -a^T(x - \bar{x}) = \alpha - a^T x. \qquad \square$$

Für die lineare Optimierung besonders relevant ist ein weiterer Spezialfall unter einer zusätzlichen Voraussetzung. Eine Menge $K \subset \mathbb{R}^n$ heißt *Kegel*, wenn gilt

$$\lambda x \in K \qquad \text{für alle } x \in K \text{ und } \lambda > 0.$$

Ein elementares Beispiel für einen Kegel, der darüber hinaus konvex und abgeschlossen ist, ist der *positive Orthant*

$$\mathbb{R}^n_+ := \{x \in \mathbb{R}^n : x_i \geq 0,\ 1 \leq i \leq n\}.$$

Folgerung 11.4. *Sei $K \subset \mathbb{R}^n$ ein nichtleerer, abgeschlossener, und konvexer Kegel und $x_0 \in \mathbb{R}^n \setminus K$. Dann existiert ein $a \in \mathbb{R}^n \setminus \{0\}$ mit*

$$a^T x_0 > 0 \geq a^T x \qquad \text{für alle } x \in K. \tag{11.4}$$

*Beweis.* Wir müssen lediglich zeigen, dass für einen Kegel in Satz 11.3 stets $\alpha = 0$ gewählt werden kann. Sei dafür zunächst $x \in K$ beliebig. Da $K$ ein Kegel ist, gilt dann auch $\lambda x \in K$ für alle $\lambda > 0$. Daraus folgt einerseits wegen der Abgeschlossenheit von $K$

$$0 = \lim_{\lambda \to 0} \lambda x \in K$$

und damit aus der zweiten Ungleichung in (11.3) insbesondere $a^T x_0 > \alpha \geq a^T 0 = 0$. Andererseits folgt aus $\lambda x \in K$ auch

$$\lambda(a^T x) = a^T(\lambda x) \leq \alpha \qquad \text{für alle } \lambda > 0,$$

was mit Grenzübergang $\lambda \to \infty$ nur möglich ist, falls $a^T x \leq 0$ ist. Zusammen erhalten wir (11.4). $\square$

## 11.2 POLYEDER UND IHRE DARSTELLUNGEN

Wir betrachten nun die zulässige Menge von (LP) und beginnen mit einigen elementaren Notationen und Eigenschaften. Für eine Matrix $A \in \mathbb{R}^{m\times n}$ schreiben wir $a_j \in \mathbb{R}^m$ für die $j$-te Spalte; die $i$-te Zeile (aufgefasst als Zeilenvektor) bezeichnen wir mit $A_i \in \mathbb{R}^n$. Damit lässt sich die $i$-te Zeile des linearen Ungleichungssystems $Ax \leq b$ kurz schreiben als $A_i x \leq b_i$.

Eine Teilmenge $G \subset \mathbb{R}^n$ heißt *Hyperebene*, falls gilt

$$G = \left\{x \in \mathbb{R}^n : a^T x = \alpha\right\} \qquad \text{für ein } a \in \mathbb{R}^n, \alpha \in \mathbb{R},$$

man bezeichnet $a$ dann als *Normalenvektor* zu $G$ (vergleiche Satz 11.3). Eine Teilmenge $H \subset \mathbb{R}^n$ heißt (abgeschlossener) *Halbraum*, falls gilt

$$H = \left\{x \in \mathbb{R}^n : a^T x \leq \alpha\right\} \qquad \text{für ein } a \in \mathbb{R}^n, \alpha \in \mathbb{R}.$$

Eine Teilmenge $P \subset \mathbb{R}^n$ heißt *Polyeder*, falls gilt

$$P = P(A, b) := \{x \in \mathbb{R}^n : Ax \leq b\} \qquad \text{für ein } A \in \mathbb{R}^{m\times n}, b \in \mathbb{R}^m,$$

und *Polytop*, falls $P$ ein beschränkter Polyeder ist, d. h. $P \subset \{x \in \mathbb{R}^n : \|x\| \leq M\}$ für ein $M > 0$. Man rechnet leicht nach, dass Polyeder stets konvex und abgeschlossen sind.

Offensichtlich ist jeder Halbraum ein Polyeder, aber auch die leere Menge (denn es gilt $\left\{x \in \mathbb{R}^n : 0^T x \leq -1\right\} = \emptyset$). Damit ist jeder Polyeder Durchschnitt endlich vieler Halbräume, denn

$$P(A, b) = \bigcap_{i=1}^{m} \{x \in \mathbb{R}^n : A_i x \leq b_i\}.$$

Jedes Paar $(A, b)$ definiert einen eindeutigen Polyeder, aber nicht umgekehrt: Da man Ungleichungen mit positiven(!) Skalaren multiplizieren und addieren darf, ist

$$P(A, b) = P(A, b) \cap \{x \in \mathbb{R}^n : (\alpha A_i)x \leq \alpha b_i\} \cap \left\{x \in \mathbb{R}^n : (A_i + A_j)x \leq (b_i + b_j)\right\}$$

für beliebige $\alpha > 0$ und $i, j \in \{1, \ldots, m\}$. Jeder Polyeder hat also unendlich viele verschiedene Darstellungen.

Oft tauchen in linearen Optimierungsproblemen zusätzlich Gleichheitsbedingungen auf; diese kann man aber einfach als weitere Ungleichungen aufnehmen: Es gilt $a^T x = \beta$ genau dann, wenn $\beta \leq a^T x \leq \beta$ gilt, d. h. wir fügen die Ungleichungen

$$\begin{aligned} a^T x &\leq \beta, \\ -a^T x &\leq -\beta \end{aligned}$$

hinzu. Häufig sind auch Vorzeichenbedingungen an $x$, die wir ebenfalls als zusätzliche Ungleichungen einfügen können.

Lemma 11.5. *Für dimensionsverträgliche Vektoren $c, d, x, y$ und Matrizen $A, B, C, D$ ist die Lösungsmenge des linearen Systems*

$$\begin{cases} Ax + By \leq c, \\ Cx + Dy = d, \\ x \geq 0 \end{cases} \tag{11.5}$$

*ein Polyeder.*

*Beweis.* Setze

$$A' = \begin{pmatrix} A & B \\ C & D \\ -C & -D \\ -I & 0 \end{pmatrix}, \qquad b' = \begin{pmatrix} c \\ d \\ -d \\ 0 \end{pmatrix}, \qquad x' = \begin{pmatrix} x \\ y \end{pmatrix},$$

dann ist (11.5) äquivalent zu $A'x' \leq b'$. □

Insbesondere ist

$$P^{=}(A, b) := \{x \in \mathbb{R}^n : Ax = b, x \geq 0\}$$

ein Polyeder, den man wie in Lemma 11.5 in die Form $P(A', b')$ transformieren kann. Umgekehrt kann man jeden Polyeder $P(A, b)$ in die Form $P^{=}(A', b')$ bringen, indem man folgende Transformationen verwendet:

(i) Erfüllt $x \in \mathbb{R}^n$ die Ungleichung

$$A_i x \leq b_i,$$

so gilt stets

$$A_i x + y_i = b_i \quad \text{mit} \quad y_i = b_i - A_i x \geq 0.$$

Man bezeichnet $y_i \in \mathbb{R}$ als *Schlupfvariable* (englisch „slack variable“), ebenso den Vektor $y = (y_1, \ldots, y_m)^T \in \mathbb{R}^m$.

Erfüllt umgekehrt $x' = (x, y_i)$ die Bedingungen

$$\begin{aligned} A_i x + y_i &= b_i, \\ y_i &\geq 0, \end{aligned}$$

so gilt

$$A_i x \leq b_i.$$

(ii) Jedes $x \in \mathbb{R}^n$ kann man schreiben als

$$x = x^+ - x^- \qquad \text{mit} \qquad x^+, x^- \geq 0,$$

für

$$x_i^+ := \begin{cases} x_i & x_i \geq 0, \\ 0 & x_i < 0, \end{cases} \qquad x_i^- := \begin{cases} 0 & x_i \geq 0, \\ -x_i & x_i < 0. \end{cases}$$

Dann ist $P(A,b)$ äquivalent zu $P^=(A',b')$ mit

$$A' = (A, -A, I), \qquad b' = b$$

in dem Sinne, dass für $x \in P(A,b)$ stets $(x^+, x^-, b - Ax)^T \in P^=(A',b')$ und für $x' := (u,v,w)^T \in P^=(A',b')$ stets $x := u - v \in P(A,b)$ gilt. Wir können also je nach Bedarf zwischen beiden Formen wechseln.

## 11.3 DAS FARKAS-LEMMA

Von zentraler Frage ist nun, wann eine zulässige Menge in der Form $P^=(A,b)$ nichtleer ist, d. h. ob für gegebenes $A$ und $b$ ein $x \geq 0$ existiert mit $Ax = b$. Wir werden dies mit Hilfe des Trennungssatzes (in Form von Folgerung 11.4) entscheiden; dafür benötigen wir das folgende Lemma.

Lemma 11.6. *Sei $A \in \mathbb{R}^{m\times n}$. Dann ist*

$$K := \{Ax : x \geq 0 \in \mathbb{R}^n\} \subset \mathbb{R}^m$$

*ein nichtleerer, konvexer, und abgeschlossener Kegel.*

*Beweis.* Dass $K$ nichtleer, konvex, und ein Kegel ist, überprüft man direkt anhand der entsprechenden Definitionen. Bleibt die Abgeschlossenheit zu zeigen, welche der wesentliche technische Schritt in diesem Kapitel ist. (Beachte, dass $A$ nicht quadratisch sein und vollen Rang haben muss!)

Dafür verwenden wir, dass für $x \in \mathbb{R}^n$ das Matrix-Vektor-Produkt $Ax$ als Linearkombination der Spalten $a_1, \ldots, a_n \in \mathbb{R}^m$ von $A$ aufgefasst werden kann; es ist also

$$K = \left\{ \sum_{i=1}^n \xi_i a_i : \xi_i \geq 0,\ 1 \leq i \leq n \right\}$$

(die *konische Hülle* von $\{a_1, \ldots, a_n\}$). Wir zeigen nun mit Induktion nach $n$, dass die konische Hülle von $n$ beliebigen Vektoren abgeschlossen ist. Für $n = 1$ ist $K = \{\xi a_1 : \xi \geq 0\}$, wofür die Abgeschlossenheit leicht ersichtlich ist.

Sei nun $n > 1$ beliebig, und die konische Hülle von $n-1$ beliebigen Vektoren abgeschlossen.

Sei also $\{y^k\}_{k\in\mathbb{N}} \subset K$ eine Folge mit $y^k \to y^* \in \mathbb{R}^m$. Zunächst gilt wegen $y^k \in K \subset \mathrm{ran}(A)$ und der Abgeschlossenheit des Bildes von linearen Operatoren, dass auch $y^* \in \mathrm{ran}(A)$ liegt, d. h.

$$y^* = \sum_{i=1}^n \xi_i^* a_i \qquad \text{für } \xi_i^* \in \mathbb{R},\ 1 \leq i \leq n.$$

Sind alle $\xi_i^* \geq 0$, so ist die Aussage bereits gezeigt. Ansonsten existiert (mindestens) ein $i_0$ mit $\xi_{i_0}^* < 0$. Wir definieren nun für alle $k \in \mathbb{N}$

$$\beta_k := \min \left\{ \frac{\xi_i^k}{\xi_i^k - \xi_i^*} : \xi_i^* < 0 \right\} \in [0,1]. \tag{11.6}$$

Dann gilt für alle $i$ mit $\xi_i^* < 0$ und $k \in \mathbb{N}$

$$r_{ik} := \beta_k \xi_i^* + (1 - \beta_k)\xi_i^k \geq 0$$

wegen $\beta_k(\xi_i^k - \xi_i^*) \leq \xi_i^k$. Wir wählen nun für jedes $k \in \mathbb{N}$ einen Index $i_k \in \{1, \dots, n\}$, für den das Minimum in (11.6) angenommen wird und daher $r_{i_k k} = 0$ gilt. Da die Folge $\{i_k\}_{k \in \mathbb{N}}$ nur endlich viele Werte annimmt, muss eine konstante Teilfolge $\{i^*\}_{k \in N}$ mit $N \subset \mathbb{N}$ unendlich existieren.

Definiere nun für $k \in N$

$$z^k := y^k + \beta_k(y^* - y^k) = \sum_{i=1}^{n} (\beta_k \xi_i^* + (1 - \beta_k)\xi_i^k) a_i = \sum_{i=1}^{n} r_{ik} a_i.$$

Dann gilt wegen $y^k \to y^*$ und $\beta_k \in [0,1]$ auch $z^k \to y^*$. Weiterhin ist wegen $r_{ik} \geq 0$ und $r_{i^* k} = 0$

$$z^k \in K' := \left\{ \sum_{i \neq i^*} \xi_i a_i : \xi_i \geq 0,\ i \neq i^* \right\} \subset K.$$

Nach Induktionsvoraussetzung ist aber $K'$ als konische Hülle von $n - 1$ Vektoren abgeschlossen, woraus auch $y^* \in K' \subset K$ folgt. Also ist $K$ abgeschlossen. □

Damit sind wir nun in der Lage, das zentrale *Farkas-Lemma* zu beweisen, das eine äquivalente Charakterisierung von nichtleeren Polyedern liefert.

Lemma 11.7 (Farkas). *Seien $A \in \mathbb{R}^{m \times n}$ und $b \in \mathbb{R}^m$. Dann sind äquivalent:*

*(i) Es existiert ein $x \in \mathbb{R}^n$ mit $Ax = b$ und $x \geq 0$.*

*(ii) Für alle $d \in \mathbb{R}^n$ mit $A^T d \geq 0$ gilt $b^T d \geq 0$.*

*Beweis.* Sei $d \in \mathbb{R}^n$ mit $A^T d \geq 0$. Existiert nun ein $x \in \mathbb{R}^n$ mit $Ax = b$ und $x \geq 0$, so folgt sofort

$$b^T d = (Ax)^T d = x^T (A^T d) \geq 0,$$

da das Skalarprodukt zweier Vektoren mit nichtnegativen Einträgen ebenfalls nichtnegativ ist.

Die andere Richtung zeigen wir mit Kontraposition. Angenommen, es existiert *kein* $x \in \mathbb{R}^m$ mit $Ax = b$ und $x \geq 0$. Dann ist $b \notin K$ für den nichtleeren, konvexen, und abgeschlossenen

Kegel aus Lemma 11.6. Wir können also Folgerung 11.4 anwenden und erhalten ein $a \in \mathbb{R}^m \setminus \{0\}$ mit

$$a^T b > 0 \geq a^T y \qquad \text{für alle } y = Ax \text{ mit } x \geq 0.$$

Dann gilt für $d := -a$ und $y = Ae_i \in K$ (wobei $e_i \geq 0$ den $i$-ten Einheitsvektor bezeichnet)

$$(e_i)^T (A^T d) = d^T (Ae_i) \geq 0 \qquad \text{für alle } 1 \leq i \leq n$$

und damit $A^T d \geq 0$, aber $b^T d = d^T b < 0$. □

Aus dem Beweis wird ersichtlich, dass Lemma 11.7 auch als *Alternativsatz* aufgefasst werden kann: Entweder das System $Ax = b, x \geq 0$ *oder* das System $A^T d \geq 0, b^T d < 0$ hat eine Lösung. Wir werden diese Alternative in verschiedenen Formulierungen benötigen.

Folgerung 11.8 (Alternativsatz). *Für dimensionsverträgliche Vektoren $c, d, x, y, u, v$ und Matrizen $A, B, C, D$ gilt genau eine der beiden Aussagen:*

*(i) Es existieren $x, y$ mit*

$$\begin{cases} Ax + By \leq c, \\ Cx + Dy = d, \\ x \geq 0. \end{cases}$$

*(ii) Es existieren $u, v$ mit*

$$\begin{cases} u^T A + v^T C \geq 0, \\ u^T B + v^T D = 0, \\ u \geq 0, \\ u^T c + v^T d < 0. \end{cases}$$

*Beweis.* Wie zuvor führen wir für die erste Ungleichung in (i) eine Schlupfvariable $w \geq 0$ ein und schreiben $y = y^+ - y^-$ mit $y^+, y^- \geq 0$. Dann ist das System in (i) äquivalent zu $A'x' = b'$ und $x' \geq 0$ für

$$A' := \begin{pmatrix} A & B & -B & I \\ C & D & -D & 0 \end{pmatrix}, \qquad x' := \begin{pmatrix} x \\ y^+ \\ y^- \\ w \end{pmatrix}, \qquad b' := \begin{pmatrix} c \\ d \end{pmatrix}.$$

Nach Lemma 11.7 hat dieses System genau dann keine Lösung, wenn ein $d' =: (u, v)$ existiert mit (nach Transponieren)

$$(d')^T A' \geq 0, \qquad (d')^T b' < 0.$$

Ausmultipliziert bedeutet dies

$$\begin{cases} u^T A + v^T C \geq 0, \\ u^T B + v^T D \geq 0, \\ -u^T B - v^T D \geq 0, \\ u \geq 0, \\ u^T c + v^T d < 0, \end{cases}$$

was äquivalent zu dem System in (ii) ist. □

Durch Spezialisierung erhalten wir daraus eine Familie von Aussagen.

Folgerung 11.9. *Für $A \in \mathbb{R}^{m \times n}$ und $b \in \mathbb{R}^m$ gilt jeweils genau eine der folgenden Aussagen:*

*a) Es existiert $x$ mit $Ax \leq b$ oder es existiert $u \geq 0$ mit $u^T A = 0$ und $u^T b < 0$.*

*b) Es existiert $x \geq 0$ mit $Ax \leq b$ oder es existiert $u \geq 0$ mit $u^T A \geq 0$ und $u^T b < 0$.*

*c) Es existiert $x \geq 0$ mit $Ax = b$ oder es existiert $u$ mit $u^T A \geq 0$ und $u^T b < 0$.*

*d) Es existiert $x$ mit $Ax = b$ oder es existiert $u$ mit $u^T A = 0$ und $u^T b < 0$.*

Hier sehen wir unser erstes *Dualitätsresultat*: Eine Vorzeichenbedingung für $x$ (bzw. $u$) taucht genau dann auf, wenn eine Ungleichungsbedingung für $u$ (bzw. $x$) auftaucht.

# 12 FUNDAMENTALSATZ DER LINEAREN OPTIMIERUNG

Wir können nun entscheiden, wann die zulässige Menge $P(A,b)$ eines linearen Optimierungsproblems nichtleer ist. Da jeder Polyeder abgeschlossen und das lineare Funktional $c^T x$ stetig ist, kann nur noch schief gehen, dass $c^T x$ auf $P(A,b)$ nach unten unbeschränkt ist, also $\inf_{x\in P(A,b)} c^T x = -\infty$ ist. In Folge seien stets $A \in \mathbb{R}^{m\times n}$, $b \in \mathbb{R}^m$, und $c \in \mathbb{R}^n$.

## 12.1 DUALITÄT

Wir fragen uns also, ob eine der Ungleichungen aus $Ax \leq b$ erlaubt, eine untere Schranke für $c^T x$ anzugeben. Zunächst gilt für alle $x \in P(A,b)$ und $y \geq 0$ die Ungleichung

$$-y^T Ax \geq -y^T b.$$

Können wir also ein $\bar{y} \geq 0$ finden mit $A^T \bar{y} = -c$, dann ist für alle $x \in P(A,b)$

$$c^T x = (-A^T \bar{y})^T x = -\bar{y}^T Ax \geq -\bar{y}^T b = -b^T \bar{y} \tag{12.1}$$

und wir haben eine untere Schranke gefunden. Damit haben wir auch schon unser erstes Existenzkriterium: Ist $P^=(A^T, -c)$ nichtleer, so hat das Problem

$$\begin{cases} \min\limits_{x\in\mathbb{R}^n} c^T x \\ \text{mit } Ax \leq b. \end{cases} \tag{P}$$

eine Lösung $\bar{x} \in P(A,b)$. Wir können sogar noch mehr sagen: Gilt Gleichheit in (12.1) für ein $\bar{x}$ (d. h. $A\bar{x} = b$), so ist $\bar{x}$ Lösung von (P), denn für alle $x \in P(A,b)$ folgt dann

$$c^T x \geq -b^T \bar{y} = -\bar{y}^T A\bar{x} = c^T \bar{x}.$$

Umgekehrt muss dann $\bar{y}$ die beste (d. h. größte) untere Schranke ergeben, löst also das *duale Problem*

$$\begin{cases} \max\limits_{y\in\mathbb{R}^m} -b^T y \\ \text{mit } A^T y = -c, \\ \qquad\quad y \geq 0. \end{cases} \tag{D}$$

Entsprechend wird (P) das *primale Problem* genannt. Diese einfache Beobachtung ist so zentral, dass wir sie festhalten.

Satz 12.1 (schwache Dualität). *Für alle $x \in P(A,b)$ und $y \in P^=(A^T,-c)$ ist*

$$c^T x \geq -b^T y. \tag{12.2}$$

*Gilt Gleichheit für ein $\bar{x} \in P(A,b)$ und ein $\bar{y} \in P^=(A^T,-c)$, so ist $\bar{x}$ Lösung von (P) und $\bar{y}$ Lösung von (D).*

Eine äquivalente Aussage gilt auch für allgemeinere lineare Probleme mit Nebenbedingungen der Form (11.5).

Lemma 12.2. *Für dimensionsverträgliche Vektoren $c, d, e, f, x, y$ und Matrizen $A, B, C, D$ ist das duale Problem zu*

$$\text{(PA)} \qquad \begin{cases} \min\limits_{x,y} c^T x + d^T y \\ \text{mit } Ax + By \geq e, \\ \quad Cx + Dy = f, \\ \quad x \geq 0, \end{cases}$$

*gegeben durch*

$$\text{(DA)} \qquad \begin{cases} \max\limits_{u,v} e^T u + f^T v \\ \text{mit } A^T u + C^T v \leq c, \\ \quad B^T u + D^T v = d, \\ \quad u \geq 0. \end{cases}$$

*Beweis.* Wir schreiben (PA) analog zum Beweis von Lemma 11.5 als $\min_{x' \in P(A',b')} (c')^T x'$ mit

$$x' := \begin{pmatrix} x \\ y \end{pmatrix}, \qquad c' := \begin{pmatrix} c \\ d \end{pmatrix}, \qquad A' := \begin{pmatrix} -A & -B \\ C & D \\ -C & -D \\ -I & 0 \end{pmatrix}, \qquad b' := \begin{pmatrix} -e \\ f \\ -f \\ 0 \end{pmatrix}.$$

Das duale Problem dazu ist dann $\max_{y' \in P^=((A')^T, -c')} -(b')^T y'$ mit $y' = (u, v^+, v^-, w)^T$. Einsetzen von $v := -(v^+ - v^-)$ und Auffassen von $w$ als Schlupfvariable wie im Beweis von Folgerung 11.8 ergibt dann (DA). □

Damit erkennt man auch, dass das duale Problem zu (DA) wieder (PA) ist. Wie im Farkas-Lemma gehört zu jeder primalen Nebenbedingung eine duale Variable, die nicht-negativ ist genau dann, wenn es sich um eine Ungleichungsnebenbedingung handelt.

Das zentrale Resultat dieses Kapitels – daher auch als *Fundamentalsatz der linearen Optimierung* bezeichnet – ist die Tatsache, dass Gleichheit in (12.1) *immer* gilt, solange die beiden zulässigen Mengen nichtleer sind.

Satz 12.3 (starke Dualität). *Beide Probleme* (P) *und* (D) *haben eine Lösung* $\bar{x}$ *bzw.* $\bar{y}$ *genau dann, wenn die zulässigen Mengen* $P(A, b)$ *bzw.* $P^=(A^T, -c)$ *nichtleer sind. In diesem Fall gilt*

$$c^T \bar{x} = -b^T \bar{y}.$$

*Beweis.* Lösungen müssen natürlich zulässig sein. Seien also $P(A, b)$ und $P^=(A^T, -c)$ nichtleer. Nach Satz 12.1 genügt es zu zeigen, dass zulässige Punkte $\bar{x}$ und $\bar{y}$ mit $c^T \bar{x} \leq -b^T \bar{y}$ existieren, d. h. dass das System

$$\begin{cases} \begin{pmatrix} 0 \\ b^T \end{pmatrix} y + \begin{pmatrix} A \\ c^T \end{pmatrix} x \leq \begin{pmatrix} b \\ 0 \end{pmatrix}, \\ A^T y = -c, \\ y \geq 0, \end{cases} \tag{12.3}$$

eine Lösung hat. Dies ist nach Folgerung 11.8 äquivalent dazu, dass für $u = (w, \gamma) \geq 0$ und $v$ das System

$$\begin{cases} \gamma b^T + v^T A^T \geq 0, \\ w^T A + \gamma c^T = 0, \\ w^T b - v^T c < 0, \end{cases} \tag{12.4}$$

keine Lösung hat. Dies zeigen wir durch Widerspruch. Wir nehmen an, (12.4) hat eine Lösung, und machen eine Fallunterscheidung nach $\gamma \in \mathbb{R}$.

(i) Fall: $\gamma = 0$. Einsetzen in (12.4) und Anwenden von Folgerung 11.8 „rückwärts“ auf die ersten beiden (Un-)Gleichungen ergibt dann, dass das System

$$\begin{cases} Ax \leq b, \\ A^T y = -c, \\ y \geq 0. \end{cases}$$

keine Lösung hat. Dafür müsste aber entweder $P(A, b)$ oder $P^=(A^T, -c)$ leer sein, im Widerspruch zur Annahme.

(ii) Fall: $\gamma > 0$. Dann folgt aus (12.4), dass gilt

$$0 > \gamma(w^T b - v^T c) = (\gamma b^T)w - (\gamma c^T)v \geq (-v^T A^T)w + (w^T A)v = 0$$

und damit ein Widerspruch.

Also kann (12.4) keine Lösung haben. Damit hat (12.3) eine Lösung, was zu zeigen war. □

Eine analoge Aussage gilt für (PA) und (DA). Tatsächlich reicht für die starke Dualität aus, dass *eines* der beiden Probleme eine Lösung hat. Manchmal ist es auch einfacher, Lösbarkeit des dualen Problems zu zeigen als Existenz eines primal zulässigen Punkts.

Satz 12.4. *Die folgenden Aussagen sind äquivalent:*

*(i)* (P) *hat eine Lösung* $\bar{x}$*;*

*(ii)* (D) *hat eine Lösung* $\bar{y}$*;*

*(iii)* (P) *und* (D) *haben beide eine Lösung, für die gilt* $c^T\bar{x} = -b^T\bar{y}$.

*Beweis.* Aus (iii) folgt trivialerweise (i) und (ii). Sei also $\bar{x}$ Lösung von (P), d. h. gelte $\bar{x} \in P(A,b)$ und $c^T\bar{x} \leq c^Tx$ für alle $x \in P(A,b)$. Es gibt also kein $x \in P(A,b)$ mit $c^Tx \leq \gamma$ für beliebiges $\gamma < c^T\bar{x}$, d. h. das System

$$\begin{cases} Ax \leq b, \\ c^Tx \leq \gamma \end{cases}$$

hat keine Lösung. Nach Folgerung 11.9 a) gibt es daher $(u,\beta) \geq 0$ mit

$$\begin{cases} A^Tu + \beta c = 0, \\ u^Tb + \beta\gamma < 0. \end{cases}$$

Ist $\beta = 0$, so folgt daraus – wieder mit Folgerung 11.9 a) – dass $Ax \leq b$ keine Lösung hat, im Widerspruch zur Annahme $\bar{x} \in P(A,b)$. Also muss $\beta > 0$ gelten, und $y := \beta^{-1}u$ erfüllt

$$\begin{cases} A^Ty = -c, \\ y^Tb < -\gamma, \\ y \geq 0. \end{cases}$$

Insbesondere ist damit $y \in P^=(A^T,-c)$ und aus Satz 12.3 folgt die Behauptung (iii).

Man argumentiert analog, falls eine Lösung $\bar{y}$ von (D) existiert. □

Durch Kontraposition erhält man daraus Bedingungen, wann ein Problem *keine* Lösung hat. Wir nennen dafür das primale Problem (P) *unbeschränkt*, falls gilt

$$\inf_{x \in P(A,b)} c^Tx = -\infty,$$

und *nicht zulässig*, falls gilt $P(A,b) = \emptyset$. Analog nennen wir das duale Problem (D) unbeschränkt, falls gilt

$$\sup_{y \in P^=(A^T,-c)} -b^Ty = \infty,$$

und nicht zulässig, falls gilt $P^=(A^T,-c) = \emptyset$.

Folgerung 12.5. *Es gilt:*

*(i) Ist* (P) *unbeschränkt, dann ist* (D) *nicht zulässig.*

*(ii)* *Ist* (D) *unbeschränkt, dann ist* (P) *nicht zulässig.*

*(iii)* *Ist* (P) *nicht zulässig, dann ist* (D) *unbeschränkt oder nicht zulässig.*

*(iv)* *Ist* (D) *nicht zulässig, dann ist* (P) *unbeschränkt oder nicht zulässig.*

*Beweis.* Für (i) sei $y \in P^{=}(A^T, -c)$. Dann folgt aus Satz 12.1

$$\inf_{x \in P(A,b)} c^T x \geq -b^T y > -\infty,$$

im Widerspruch zur Unbeschränktheit. Analog argumentiert man für (ii).

Für (iii) nehmen wir an, dass (P) nicht zulässig und (D) nach oben beschränkt ist, d. h. das Maximum wird angenommen (sonst wären wir bereits fertig). Ist nun (D) zulässig, dann hat (P) nach Satz 12.4 eine Lösung, im Widerspruch zur Annahme. Also ist (D) nicht zulässig. Analog argumentiert man für (iv). □

Dass tatsächlich beide Probleme unzulässig sein können, zeigt das folgende einfache Beispiel.

Beispiel 12.6. Sei

$$A = \begin{pmatrix} 1 & -1 \\ -1 & 1 \end{pmatrix}, \qquad b = \begin{pmatrix} 0 \\ -1 \end{pmatrix}, \qquad c = \begin{pmatrix} -1 \\ -1 \end{pmatrix}.$$

Dann ist nach Folgerung 11.9

- $P(A, b) = \emptyset$, da das System $u^T A = 0$, $u^T b < 0$ die Lösung $\bar{u} = (1, 1)^T \geq 0$ hat;
- $P^{=}(A^T, -c) = \emptyset$, da das System $u^T A^T \geq 0$, $u^T(-c) < 0$ die Lösung $\bar{u} = (-1, -1)$ hat.

## 12.2 KOMPLEMENTARITÄT

Wir beschäftigen uns nun mit Optimalitätsbedingungen für (P). Dazu verwenden wir, dass eine zulässige Lösung $\bar{x}$ die beste untere Schranke in (12.2) annehmen wird – und diese ist genau durch die Lösung $\bar{y}$ des dualen Problems gegeben.

Satz 12.7 (schwache Komplementarität). *Es ist $\bar{x}$ Lösung von* (P) *und $\bar{y}$ Lösung von* (D) *genau dann, wenn gilt*

$$\text{(12.5)} \qquad \begin{cases} A\bar{x} \leq b, & \textit{(primale Zulässigkeit)} \\ A^T \bar{y} = -c, \quad \bar{y} \geq 0, & \textit{(duale Zulässigkeit)} \\ \bar{y}_i (b_i - A_i \bar{x}) = 0 \quad \textit{für alle } i = 1, \ldots, m. & \textit{(Komplementarität)} \end{cases}$$

Man nennt (12.5) auch *Karush–Kuhn–Tucker-Bedingungen* (oder kurz: *KKT-Bedingungen*).

*Beweis.* Ist $\bar{x}$ Lösung von (P) und $\bar{y}$ Lösung von (D), so sind natürlich insbesondere $\bar{x}$ und $\bar{y}$ zulässig. Nach Satz 12.4 (iii) ist dann wegen der dualen Zulässigkeit

$$0 = c^T\bar{x} - (-b^T\bar{y}) = (-A^T\bar{y})^T\bar{x} + b^T\bar{y} = (b - A\bar{x})^T\bar{y} = \sum_{i=1}^{m}(b_i - A_i\bar{x})\,\bar{y}_i.$$

Aus primaler und dualer Zulässigkeit folgt, dass alle Summanden nicht-negativ sind und daher einzeln verschwinden müssen. Daraus folgt die Komplementarität.

Sind umgekehrt die KKT-Bedingungen erfüllt, dann ist $\bar{x} \in P(A, b)$ und $\bar{y} \in P^=(A^T, -c)$ und aus der Komplementarität folgt wie oben

$$0 = (b - A\bar{x})^T\bar{y} = c^T\bar{x} - (-b^T\bar{y}),$$

und damit nach Satz 12.1, dass $\bar{x}$ Lösung von (P) und $\bar{y}$ Lösung von (D) ist. □

Schwache Komplementarität kann man mit Hilfe von Lemma 12.2 auch für allgemeine lineare Optimierungsprobleme zeigen.

Folgerung 12.8. *Seien $A, B, C, D$ und $c, d, e, f$ wie in Lemma 12.2. Dann ist $(\bar{x}, \bar{y})$ Lösung von (PA) und $(\bar{u}, \bar{v})$ Lösung von (DA) genau dann, wenn $(\bar{x}, \bar{y})$ und $(\bar{u}, \bar{v})$ zulässig sind und die Komplementaritätsbedingungen*

$$\begin{cases} \bar{u}_i\,(e_i - [A\bar{x} + B\bar{y}]_i) = 0 & \text{für alle } i = 1, \dots, m, \\ \bar{x}_i\left(c_i - [A^T\bar{u} + C^T\bar{v}]_i\right) = 0 & \text{für alle } i = 1, \dots, n, \end{cases}$$

*erfüllen.*

*Beweis.* Wie im Beweis von Lemma 12.2 transformiert man (PA) auf die Form (P) für geeignetes $A', b', c', x'$. Dann ist wie im Beweis von Satz 12.7 die Optimalität von $(\bar{x}, \bar{y})$ und $(\bar{u}, \bar{v})$ äquivalent zu

$$\begin{aligned} 0 &= c^T\bar{x} + d^T\bar{y} - (e^T\bar{u} + f^T\bar{v}) \\ &= c^T\bar{x} + (B^T\bar{u} + D^T\bar{v})^T\bar{y} - e^T\bar{u} - (C\bar{x} + D\bar{y})^T\bar{v} + (\bar{u}^TA\bar{x} - \bar{u}^TA\bar{x}) \\ &= (c - A^T\bar{u} - C^T\bar{v})^T\bar{x} - (e - A\bar{x} - B\bar{y})^T\bar{u}. \end{aligned}$$

Aus den Ungleichungsbedingungen folgt, dass die erste Klammer (und wegen $\bar{x} \geq 0$ der ganze Term) nicht-negativ und die zweite Klammer (und wegen $\bar{u} \geq 0$ der ganze Term) nicht-positiv ist. Die Differenz kann also nur gleich Null sein, wenn beide Terme separat verschwinden, und daraus folgt wie zuvor die Komplementarität. □

Die Komplementaritätsbedingung sagt, dass $\bar{y}_i = 0$ oder $b_i - A_i\bar{x} = 0$ für jedes $1 \le i \le m$ gilt. Dies ist aber kein exklusives Oder – es ist also zugelassen, dass sowohl $\bar{y}_i$ als auch $b_i - A_i\bar{x}$ verschwinden. Man kann jedoch zeigen, dass unter allen Lösungen auch ein Paar existiert, für das immer nur genau eine Gleichheit gilt.

Satz 12.9 (strikte Komplementarität). *Sind $P(A, b)$ und $P^=(A^T, -c)$ nichtleer, so existiert eine Lösung $\bar{x}$ von (P) und eine Lösung $\bar{y}$ von (D) mit*

$$\bar{y}_i = 0 \quad \textit{genau dann, wenn} \quad A_i\bar{x} < b_i \qquad \textit{für alle } i = 1, \dots, m.$$

*Beweis.* Wir zerlegen die Menge der Ungleichungen wie folgt. Definiere für $A \in \mathbb{R}^{m\times n}$ die Zeilenmenge $M := \{1, \dots, m\}$ und

$$\begin{aligned} N &:= \{i \in M : A_i\bar{x} < b_i \text{ für eine Lösung } \bar{x} \text{ von (P)}\}, \\ B &:= \{i \in M : \bar{y}_i > 0 \quad \text{für eine Lösung } \bar{y} \text{ von (D)}\}. \end{aligned}$$

(Beachten Sie, dass nicht gefordert ist, dass alle Ungleichungen in $N$ bzw. $B$ für das *gleiche* $\bar{x}$ bzw. $\bar{y}$ strikt sind!) Die zu beweisende Aussage ist dann äquivalent mit der Behauptung, dass $N \cap B = \emptyset$ und $M = N \cup B$ ist.

Ersteres folgt direkt aus der schwachen Komplementarität: Wäre $i \in N \cap B$, so wäre $A_i\bar{x} < b_i$ für eine Lösung $\bar{x}$ des primalen Problems und $\bar{y}_i > 0$ für eine Lösung $\bar{y}$ des dualen Problems, im Widerspruch zu Satz 12.7.

Für die zweite Aussage zeigen wir zuerst, dass es Lösungen gibt, die alle strikten Ungleichungen gleichzeitig erfüllen. Wir wählen zu jedem $i \in N$ eine der Lösungen $\bar{x}^{(i)}$, die nach Definition die entsprechende strikte Ungleichung erfüllt, und bilden die Konvexkombination

$$\bar{x} := \frac{1}{|N|} \sum_{i\in N} \bar{x}^{(i)}.$$

Da die Lösungsmenge eines linearen Optimierungsproblems konvex ist, ist $\bar{x}$ wieder Lösung von (P). Weiterhin gilt für alle $j \in N$

$$A_j\bar{x} = \sum_{i\in N} \frac{1}{|N|} A_j\bar{x}^{(i)} = \frac{1}{|N|} A_j\bar{x}^{(j)} + \frac{1}{|N|} \sum_{i\neq j} A_i\bar{x}^{(i)} < \frac{1}{|N|} b_j + \frac{|N|-1}{|N|} b_j = b_j,$$

da alle $\bar{x}^{(i)}$ zulässig sind und für $i = j$ die Ungleichung nach Konstruktion sogar strikt ist. Genauso definieren wir $\bar{y}$ mit $\bar{y}_i > 0$ für alle $i \in B$.

Wir schreiben nun $A_N$ für die Matrix, die nur die Zeilen $A_i$ mit $i \in N$ enthält, und $b_N$ für den Vektor mit Einträgen $b_i$ mit $i \in N$; wir definieren analog $A_B$ und $b_B$. Mit $N \cap B = \emptyset$ erhalten wir dann

$$A_N\bar{x} < b_N, \qquad \bar{y}_B > 0, \tag{12.6}$$

$$A_B\bar{x} = b_B, \qquad \bar{y}_N = 0. \tag{12.7}$$

Angenommen, $J := M \setminus (B \cup N)$ wäre nichtleer, d. h. es existiert ein $j \in J$. Nach Definition gilt dann $\bar{y}_J = 0$ und $[b - A\bar{x}]_J = 0$. Wir zeigen nun, dass das System

$$\begin{cases} A_{J\setminus\{j\}} x \leq 0, \\ A_B x = 0, \\ A_j x < 0, \end{cases} \tag{12.8}$$

keine Lösung haben kann. Wäre nämlich $\tilde{x}$ eine Lösung, so gäbe es wegen $A_i\bar{x} < b_i$ für alle $i \in N$ ein $\varepsilon > 0$ klein genug, so dass $A_i(\bar{x} + \varepsilon\tilde{x}) \leq b_i$ für alle $i \in N$ gilt. Da nach Annahme $A_i\tilde{x} \leq 0$ für $i \notin N$ gilt, wäre also auch $\bar{x} + \varepsilon\tilde{x}$ zulässig für (P). Wegen (12.7) und $\bar{y}_J = 0$ erfüllt $\bar{x} + \varepsilon\tilde{x}$ auch die schwachen Komplementaritätsbedingungen, ist also nach Satz 12.7 sogar eine Lösung. Da aus (12.8) sogar

$$A_j(\bar{x} + \varepsilon\tilde{x}) < A_j\bar{x} \leq b_j$$

folgt, gilt nach Definition $j \in N$, im Widerspruch zu $j \notin N$.

Also hat (12.8) keine Lösung, und aus dem Farkas-Lemma (Folgerung 11.8 mit $C^T = -A_{J\setminus\{j\}}$, $D^T = -A_B$, $b^T = A_j$) erhalten wir die Existenz von $v \geq 0$ und $w$ mit

$$-(A_{J\setminus\{j\}})^T v - (A_B)^T w = A_j. \tag{12.9}$$

Definiere nun $\tilde{y} \in \mathbb{R}^m$ mit

$$\tilde{y}_N = 0, \quad \tilde{y}_B = w, \quad \tilde{y}_{J\setminus\{j\}} = v, \quad \tilde{y}_j = 1.$$

Aus (12.9) und der Definition von $\tilde{y}$ folgt dann

$$A^T\tilde{y} = (A_{J\setminus\{j\}})^T v + (A_B)^T w + A_j = 0.$$

Außerdem ist $\tilde{y}_i \geq 0$ für alle $i \in J \cup N$ sowie $\tilde{y}_j > 0$. Da nach Definition $\bar{y}_B > 0$ gilt, erfüllt für $\varepsilon > 0$ klein genug also $\bar{y} + \varepsilon\tilde{y}$ das System

$$\begin{aligned} A^T(\bar{y} + \varepsilon\tilde{y}) &= -c, \\ (\bar{y} + \varepsilon\tilde{y}) &\geq 0, \end{aligned}$$

ist damit zulässig für (D). Wegen (12.7) und $(b - A\bar{x})_J = 0$ erfüllt $\bar{y} + \varepsilon\tilde{y}$ auch die schwachen Komplementaritätsbedingungen, ist also nach Satz 12.7 sogar eine Lösung. Wieder folgt aus $\tilde{y}_j > 0$ nach Definition $j \in B$, im Widerspruch zur Annahme. Also ist $J = \emptyset$ und damit $N \cup B = M$, was zu zeigen war. □

Auch dieses Resultat kann wie in Folgerung 12.8 auf allgemeine lineare Optimierungsprobleme erweitern, was wir hier aber nicht ausführen.

# 13 GEOMETRIE DER POLYEDER

Aus Satz 12.9 folgt, dass wenn ein lineares Optimierungsproblem eine Lösung hat, eine Lösung existiert für die in einigen Ungleichungen (nämlich denen, die einer strikt positiven dualen Lösung entsprechen) sogar Gleichheit gilt. Anschaulich heißt das, dass diese Lösung auf einer Seite des zulässigen Polyeders liegt. Wir wollen nun zeigen, dass unter bestimmten Bedingungen die Lösung sogar in einer Ecke liegt. Dafür müssen wir insbesondere den Begriff von Seite und Ecke mathematisch präzisieren.

Sei $P \subset \mathbb{R}^n$ ein Polyeder. Eine Ungleichung $a^T x \leq \beta$ mit $a \in \mathbb{R}^n$ und $\beta \in \mathbb{R}$ heißt *gültig* für $P$, falls gilt

$$P \subset \left\{x \in \mathbb{R}^n : a^T x \leq \beta\right\}.$$

Wir nennen $F \subset P$ *Seite* von $P$, falls es eine gültige Ungleichung $a^T x \leq \beta$ gibt mit

$$F = \left\{x \in P : a^T x = \beta\right\}.$$

Eine Seite $F$ heißt *nichttrivial*, falls weder $F = \emptyset$ noch $F = P$ gilt.

Mit dieser Definition erhalten wir sofort das Gewünschte.

Lemma 13.1. *Die Lösungsmenge von* $\min_{x \in P} c^T x$ *ist eine Seite von* $P$.

*Beweis.* Ist die Lösungsmenge leer, so gilt die Aussage trivialerweise. Ist die Menge nichtleer, so setzen wir $\eta := \min_{x \in P} c^T x$. Dann ist $-c^T x \leq -\eta$ gültig für $P$, und die Lösungsmenge ist darstellbar als $\left\{x \in P : c^T x = \eta\right\}$. □

Dieses Resultat ist noch nicht besonders überzeugend, da nicht klar ist, ob dann auch Gleichheit in einer Ungleichung aus $Ax \leq b$ gelten muss. Dies zeigt aber der nächste Satz.

Satz 13.2. *Sei* $P := P(A, b)$ *ein Polyeder und* $F \neq \emptyset$. *Dann ist* $F$ *Seite von* $P$ *genau dann, wenn es ein Teilsystem* $A'x \leq b'$ *von* $Ax \leq b$ *gibt mit* $F = \{x \in P : A'x = b'\}$.

*Beweis.* Sei $F$ Seite von $P$, d. h. es existieren $a \in \mathbb{R}^n$, $\beta \in \mathbb{R}$ mit

$$F = \{x \in P : a^T x = \beta\} \quad \text{und} \quad \beta = \max_{x \in P} a^T x.$$

Also ist $a^T x \leq \beta$ eine gültige Ungleichung für $P$. Die Seite $F$ ist daher Lösungsmenge von

$$\max_{x \in P(A,b)} a^T x = - \min_{x \in P(A,b)} (-a^T x). \tag{13.1}$$

Nach Satz 12.4 hat das duale Problem $\min_{y \in P^=(A^T,a)} -b^T y$ eine Lösung $\bar{y}$. Satz 12.7 ergibt dann, dass $A_i \bar{x} = b_i$ für alle $i$ mit $\bar{y}_i > 0$ gilt. Dies ergibt das gewünschte Teilsystem.

Sei umgekehrt $F = \{x \in P : A'x = b'\}$. Setze $\mathbb{1} := (1, 1, \ldots, 1)^T$ und $a := (A')^T \mathbb{1}$ (d. h. $a$ ist die Summe der Zeilen von $A'$), dann gilt für alle $x \in P$

$$a^T x = ((A')^T \mathbb{1})^T x = \mathbb{1}^T (A'x) \leq \mathbb{1}^T b' =: \beta.$$

Also ist $a^T x \leq \beta$ gültig für $P$. Weiter gilt für alle $x \in F$

$$a^T x = \mathbb{1}^T A' x = \mathbb{1}^T b' = \beta,$$

und damit ist $F$ Seite von $P$. □

Folgerung 13.3. *Seien $P$ ein Polyeder und $F \subset P$ eine Seite. Dann gilt:*

*(i) $P$ hat endlich viele Seiten.*

*(ii) $F$ ist wieder ein Polyeder.*

*(iii) $F' \subset F$ ist Seite von $P$ genau dann, wenn $F'$ Seite von $F$ ist.*

Dies ist noch nicht völlig befriedigend, da eine Seite durch mehrere Ungleichungen dargestellt werden kann, ohne dass klar ist ob ihre Anzahl etwas über die Dimension der Seite aussagt. Wir suchen daher eine sparsamere Darstellung. Eine nichttriviale Seite, die nicht Teilmenge einer anderen nichttrivialen Seite ist, nennen wir *Facette*. Diese charakterisieren wir nun. Dafür nennen wir eine Ungleichung $A_i x \leq b_i$ von $Ax \leq b$

- *implizite Gleichung* (in $Ax \leq b$), falls $A_i x = b_i$ für alle $x \in P(A, b)$ gilt;
- *redundant* (in $Ax \leq b$), falls sie durch positive Linearkombination anderer Ungleichungen in $Ax \leq b$ dargestellt werden kann.

Wir bezeichnen das System der impliziten Gleichungen mit $A^= x \leq b^=$ und das der verbleibenden nichtreduntanten Ungleichungen mit $A^+ x \leq b^+$. Wie im Beweis von Satz 12.9 zeigt man nun, dass alle diese Ungleichungen gleichzeitig strikt sein können.

Lemma 13.4. *Ist $P(A,b)$ nichtleer, so existiert ein $x \in P(A,b)$ mit $A^= x = b^=$ und $A^+ b < b^+$.*

*Beweis.* Ist $A^+$ die leere Matrix, so ist die Behauptung trivial. Sei also $A^+$ nichtleer. Für jede Zeile $A_i x \leq b_i$ aus $A^+ x \leq b^+$ muss dann ein $x^{(i)} \in P(A,b)$ existieren mit $A_i x^{(i)} < b_i$ und $A^= x^{(i)} = b^=$. Bezeichne wieder $N$ die Menge der strikten Ungleichungen und definiere $x := \frac{1}{|N|} \sum_{i \in N} x^{(i)} \in P(A,b)$. Dann gilt $A^+ x < b^+$ und $A^= x = b^=$. □

Jeder (nichttriviale) Polyeder besitzt also ein nichtleeres relatives Inneres.

Satz 13.5. *Eine Seite $F$ ist Facette von $P(A,b)$ genau dann, wenn gilt*

$$F = \{x \in P(A,b) : A_i x = b_i\}$$

*für eine Ungleichung aus $A^+ x \leq b^+$.*

*Beweis.* Sei $F = \{x \in P : A'x = b'\}$ eine Facette von $P := P(A,b)$, wobei $A'x \leq b'$ ein Teilsystem von $A^+ x \leq b^+$ ist (ein solches existiert, da $F \neq P$ nach Annahme), und sei $A_i x \leq b_i$ eine Ungleichung aus $A'x \leq b'$. Dann ist $F' = \{x \in P : A_i x = b_i\}$ nach Satz 13.2 eine Seite von $P$ mit $F \subset F' \subset P$. Weil $A_i x \leq b_i$ keine implizite Gleichung von $P$ ist, gilt $F' \neq P$, und aus der Maximalität von $F$ folgt $F = F'$.

Sei umgekehrt $A_i x \leq b_i$ eine Ungleichung aus $A^+ x \leq b^+$, und habe $F$ die angegebene Darstellung. Nach Satz 13.2 ist $F$ eine Seite von $P$; wir haben also zu zeigen, dass $F$ maximal ist. Seien dafür $A'x \leq b'$ die verbleibenden Ungleichungen. Nach Lemma 13.4 existiert dann ein $\bar{x}$ mit $A^= \bar{x} = b^=$ und $A^+ \bar{x} < b^+$. Da $A_i x \leq b_i$ nach Voraussetzung nichtredundant ist, gibt es – wieder mit Lemma 13.4 – ein $\hat{x}$ mit $A^= \hat{x} = b^=$, $A'\hat{x} < b'$ und $A_i \hat{x} > b_i$. Aus der Stetigkeit von $x \mapsto A_i x$ und der Konvexität von Polyedern folgt dann die Existenz eines $t \in (0,1)$ so dass für $x(t) := t\hat{x} + (1-t)\bar{x} \in P$ gilt

$$A^= x(t) = b^=, \qquad A_i x(t) = b_i, \qquad A'x(t) < b',$$

d. h. $x(t) \in F$. Sei nun $F' = \{x \in P : A''x = b''\}$ eine beliebige Seite von $P$ mit $F \subset F'$, wobei $A''x \leq b''$ ein Teilsystem von $A^+ x \leq b^+$ ist. Dann ist $x(t) \in F \subset F'$. Wegen $A'x(t) < b'$ ist aber $A_i x \leq b_i$ die einzige Ungleichung, für die Gleichheit in $x(t)$ gilt. Damit muss aber $A'' = A_i$ und $b'' = b_i$ sein, woraus $F' = F$ und damit Maximalität von $F$ folgt. Also ist $F$ Facette von $P$. □

Jede Seite von $P$ außer $P$ selbst ist also Schnitt von Facetten, den maximalen Seiten. Umgekehrt wollen wir Ecken charakterisieren als minimale Seiten. Dabei ist nicht klar, ob $P$ überhaupt Ecken hat (Halbräume sind ja auch Polyeder). Wir erinnern: Mengen von der Form $\{x \in \mathbb{R}^n : Ax = b\}$ heißen *affiner Unterraum* von $\mathbb{R}^n$.

Folgerung 13.6. *Ein Polyeder $P$ hat genau dann keine nichttrivialen Seiten, wenn $P$ ein affiner Unterraum ist.*

*Beweis.* Ein Polyeder $P = P(A, b)$ ist genau dann ein affiner Unterraum, wenn $A^+x \leq b^+$ leer ist. Dies ist nach Satz 13.5 dann und nur dann der Fall, wenn es keine Facetten gibt. Da jede Seite Schnitt von Facetten ist, ist dies aber äquivalent dazu, dass es keine nichttrivialen Seiten gibt. □

Daraus folgt sofort, dass eine Seite (die ja selber ein Polyeder ist) genau dann keine weiteren nichtleeren Seiten enthält – wir sagen: eine *minimale Seite* ist – wenn sie ein affiner Unterraum ist.

Satz 13.7. *Eine nichtleere Menge $F$ ist genau dann minimale Seite von $P(A, b)$, wenn $F \subset P(A, b)$ ist und die Form $F = \{x \in \mathbb{R}^n : A'x = b'\}$ hat für ein Teilsystem $A'x \leq b'$ von $Ax \leq b$.*

*Beweis.* Hat $F$ die angegebene Darstellung, so ist sie ein affiner Teilraum und damit minimale Seite. Sei umgekehrt $F \neq \emptyset$ eine minimale Seite. Dann ist

$$F = \{x \in \mathbb{R}^n : A''x \leq b'', A'x = b'\}$$

für zwei Teilsysteme von $A^+x \leq b^+$, wobei $A''x \leq b''$ minimal sein und insbesondere keine redundanten Ungleichungen in $A''x \leq b'', A'x = b'$, enthalten soll. Da $F$ minimal ist, enthält es keine nichttrivialen Seiten und ist daher nach Folgerung 13.6 ein affiner Unterraum. Also muss $A''x \leq b''$ leer sein, und $F$ daher die gewünschte Darstellung haben. □

Das nächste Lemma charakterisiert die minimalen Seiten. Wir erinnern: der Rang einer Matrix $\operatorname{rang} A$ ist die Anzahl der linear unabhängigen Zeilen.

Lemma 13.8. *Alle nichttrivialen minimalen Seiten von $P(A, b)$ haben die Dimension $n - \operatorname{rang} A$.*

*Beweis.* Sei wie in Satz 13.7 eine Seite $F = \{x \in \mathbb{R}^n : A'x = b'\}$ für ein Teilsystem $A'x \leq b'$ gegeben. Aus der linearen Algebra wissen wir, dass die Lösungsmenge eines linearen Gleichungssystems $A'x = b'$ die Dimension $n - \operatorname{rang} A'$ hat. Wir zeigen nun durch Widerspruch, dass $\operatorname{rang} A' = \operatorname{rang} A$ gilt. Angenommen, es wäre $\operatorname{rang} A > \operatorname{rang} A'$. (Da es sich bei $A'x \leq b'$ um ein Teilsystem handelt, kann der Rang ja nicht kleiner sein.) Dann existiert eine Zeile $A_i$ von $A$, die linear unabhängig zu allen Zeilen von $A'$ ist. Damit gilt aber wegen $F \subset P$, dass

$$F \subset \{x \in \mathbb{R}^n : A'x = b' \text{ und } A_i x \leq b_i\} \subsetneq \{x \in \mathbb{R}^n : A'x = b'\} = F,$$

ein Widerspruch. □

Damit kommen wir endlich zu der gesuchten Charakterisierung von Ecken: Eine Menge $F \subset P$ heißt *Ecke* von $P$, wenn $F$ eine (minimale) Seite der Dimension Null ist. Ein Polyeder, der Ecken hat, heißt *spitz*.

Satz 13.9. *Folgende Aussagen über ein Polyeder $P \subset \mathbb{R}^n$ sind äquivalent:*

*(i)* *$P$ ist spitz;*

*(ii)* $\operatorname{rang} A = n$;

*(iii)* *Jede nichtleere Seite von $P$ ist spitz.*

*Beweis.* Die Äquivalenz von (i) und (ii) folgt aus der Definition und Lemma 13.8. Hat eine nichtleere Seite $F$ von $P$ eine Ecke, so ist diese (als minimale Seite) nach Folgerung 13.3 (iii) auch Ecke von $P$, also ist $P$ spitz. Sei umgekehrt $P$ spitz und $F \subset P$ eine nichtleere Seite. Angenommen, eine minimale Seite $F_0$ von $F$ ist keine Ecke, d. h. $F_0$ ist ein affiner Teilraum der Dimension größer Null. Da $F_0$ nach Folgerung 13.3 (iii) auch Seite von $P$ ist, hat auch $P$ eine minimale Seite der Dimension größer Null. Nach Lemma 13.8 haben aber alle minimalen Seiten von $P$ den selben Rang, und damit kann $P$ keine Ecken haben. Damit ist (i) äquivalent zu (iii). □

Nach all der Vorarbeit erhalten wir nun das zentrale Resultat dieses Kapitels.

Satz 13.10. *Ist $P$ ein spitzer Polyeder und ist $\min_{x \in P} c^T x$ lösbar, so existiert eine Lösung, die Ecke von $P$ ist.*

*Beweis.* Nach Lemma 13.1 ist die Lösungsmenge des Optimierungsproblems eine (nichtleere) Seite von $P$, die nach Satz 13.9 spitz ist, also eine Ecke enthält. Diese ist die gesuchte Lösung. □

Dies ist auch der Grundgedanke der Simplex-Methode, die wir im nächsten Kapitel betrachten werden: Um eine Lösung zu finden, muss man nur an den Ecken von $P$ suchen – die dafür natürlich existieren müssen. Für Polyeder der Form $P^=(A, b) = \{x \in \mathbb{R}^n : Ax = b, x \geq 0\}$ ist das stets der Fall.

Satz 13.11. *Ein nichtleerer Polyeder $P^=(A, b)$ ist spitz.*

*Beweis.* Wir können $P^=(A, b)$ darstellen als $P(D, f)$ mit

$$D = \begin{pmatrix} A \\ -A \\ -I \end{pmatrix}, \qquad f = \begin{pmatrix} b \\ -b \\ 0 \end{pmatrix}.$$

Auf Grund des Identitätsblocks ist $\operatorname{rang} D = n$, und aus Satz 13.9 folgt die Behauptung. □

Schließlich charakterisieren wir die Ecken von $P^=(A,b)$. Bezeichne dafür für $x \in \mathbb{R}^n$

$$\operatorname{supp}(x) := \{j \in \{1,\dots,n\} : x_j \neq 0\}$$

den *Träger* von $x$.

Satz 13.12. *Die folgenden Aussagen sind äquivalent:*

*(i)* $\{x\} \subset \mathbb{R}^n$ *ist Ecke von* $P^=(A,b)$*;*

*(ii) Die Spaltenvektoren* $a_j$, $j \in \operatorname{supp}(x)$, *von* $A$ *sind linear unabhängig.*

*Beweis.* Wir schreiben $P^=(A,b)$ wie im Beweis von Satz 13.11 als $P(D,f)$. Dann ist $\{x\}$ nach Satz 13.7 und Satz 13.9 Ecke von $P^=(A,b)$ genau dann, wenn ein Teilsystem $D'x \leq f'$ existiert mit

$$\{x \in \mathbb{R}^n : D'x = f'\} = \{x\}$$

und $\operatorname{rang} D' = n$. Betrachten wir nun die Struktur von $D'$ genauer, so bedeutet das, dass $Ax = b$ diejenigen Komponenten von $x$, die im Träger von $x$ liegen, eindeutig festlegt. (Alle anderen sind ja nach Definition Null, und $-Ax = -b$ ist offensichtlich linear abhängig von $Ax = b$.) Dies ist aber genau dann der Fall, wenn die Spalten von $A$, die zu den Komponenten $x_j$ im Träger $\operatorname{supp}(x)$ gehören, linear unabhängig sind. (Für die anderen ist $x_j = 0$, so dass die entsprechenden Spalten keine Rolle spielen; es reicht also, dass die reduzierte Matrix vollen Rang hat.) □

# 14 DAS SIMPLEX-VERFAHREN

Wir betrachten nun das Standard-Verfahren zur Lösung linearer Optimierungsprobleme der Form $\min_{x \in P} c^T x$. Die Grundidee des Simplex-Verfahrens ist es, ausgehend von einer gegebenen Ecke des zulässigen Polyeders $P$ (die immer existiert, wenn $P = P^=(A, b)$ ist) eine neue Ecke zu finden, für die der Wert der Zielfunktion $c^T x$ strikt kleiner ist. Dabei vermeiden wir die historisch übliche Tableau-Form und leiten das (revidierte) Simplex-Verfahren gleich in Matrix-Schreibweise her.

Sei daher ein lineares Optimierungsproblem in *Normalform*

$$\text{(P)} \qquad \begin{cases} \min\limits_{x \in \mathbb{R}^n} \; c^T x \\ \text{mit } Ax = b, \\ \qquad\quad x \geq 0, \end{cases}$$

für $A \in \mathbb{R}^{m \times n}$ gegeben. Wir nehmen an, dass $n > m$ ist (sonst hat $Ax = b$ keine oder eine eindeutig bestimmte Lösung, und es gibt nichts zu optimieren), und dass $\operatorname{rang} A = m$ ist (sonst entfernen wir linear abhängige Zeilen). Das zu (P) duale Problem lautet nach Lemma 12.2

$$\text{(D)} \qquad \begin{cases} \max\limits_{y \in \mathbb{R}^m} \; b^T y \\ \text{mit } A^T y \leq c. \end{cases}$$

Nach Folgerung 12.8 ist $\bar{x}$ Lösung von (P) und $\bar{y}$ Lösung von (D) – man nennt dann $(\bar{x}, \bar{y})$ ein *primal-duales Paar* – genau dann, wenn $\bar{x}$ und $\bar{y}$ jeweils zulässig sind und die Komplementaritätsbedingungen

$$x_j (c_j - [A^T y]_j) = 0 \qquad \text{für alle } j = 1, \ldots, n \tag{14.1}$$

erfüllen.

## 14.1 HERLEITUNG DES VERFAHRENS

Aus Satz 13.10 wissen wir, dass falls (P) lösbar ist, eine Lösung $\bar{x}$ existiert, die Ecke von $P^=(A, b)$ ist. Nach Satz 13.12 sind in diesem Fall die Spalten von $A$, die zu strikt positiven

Komponenten von $\bar{x}$ gehören, linear unabhängig. Dies motiviert die folgende Definition: Ein Vektor $x \in P^=(A,b)$ heißt für $P^=(A,b)$ zulässiger *Basisvektor*, falls $x \in P^=(A,b)$ liegt und eine Indexmenge $B \subset \{1,\dots,n\}$ mit $|B| = m$ (die *Basis*) existiert, so dass gilt:

(i) $x_j = 0$ für alle $j \notin B$, und

(ii) die Spalten $a_j$, $j \in B$, sind linear unabhängig.

Wir schreiben $x_B$ für den Vektor mit Einträgen $x_j$, $j \in B$ (den *Basisvariablen*) sowie $A_B \in \mathbb{R}^{m\times m}$ für die *Basismatrix* mit den Spalten $a_j$ für $j \in B$, die nach Konstruktion invertierbar ist. Analog definieren wir $N := \{1,\dots,n\} \setminus B$ sowie die *Nichtbasisvariablen* $x_N$ und die *Nichtbasismatrix* $A_N$. Es gilt dann für alle $z \in \mathbb{R}^n$

$$Az = A_B z_B + A_N z_N,$$

und insbesondere für den zugehörigen Basisvektor $x$

$$A_B x_B = b, \qquad x_N = 0. \tag{14.2}$$

Wir zeigen nun, dass es sich bei den zulässigen Basisvektoren genau um die Ecken von $P^=(A,b)$ handelt.

Lemma 14.1. *Sei $A \in \mathbb{R}^{m\times n}$ mit $\operatorname{rang}(A) = m < n$. Dann ist $\{x\}$ genau dann Ecke von $P^=(A,b)$, wenn $x$ ein für $P^=(A,b)$ zulässiger Basisvektor ist.*

*Beweis.* Sei $\{x\} \subset P^=(A,b)$ eine Ecke und $I = \operatorname{supp}(x)$. Dann sind nach Satz 13.12 die Spalten $a_j$, $j \in I$, linear unabhängig. Wegen $\operatorname{rang} A = m$ gilt $|I| \le m$. Es ist daher entweder $|I| = m$ (dann sind wir fertig), oder es existieren $m - |I|$ weitere (zueinander und zu $A_I$) linear unabhängige Spalten $a_j$. Ergänzen wir $I$ mit diesen $j$ zu $B$, so ist $A_B \in \mathbb{R}^{m\times m}$ regulär und $x_j = 0$ für $j \notin B$ (da $\operatorname{supp}(x) = I \subset B$). Also ist $x$ ein Basisvektor.

Sei umgekehrt $x \in P^=(A,b)$ ein Basisvektor. Dann existiert eine Basis $B$ mit invertierbarer Basismatrix $A_B \in \mathbb{R}^{m\times m}$ und $x_j = 0$ für $j \notin B$. Also ist $\operatorname{supp}(x) \subset B$, und die Spalten $a_j$, $j \in \operatorname{supp}(x)$, sind linear unabhängig. Nach Satz 13.12 ist daher $\{x\}$ eine Ecke von $P^=(A,b)$. □

Sei nun ein Basisvektor $x$ mit Basis $B$ gegeben. Unser Ziel ist nun, einen neuen Vektor $z$ so zu konstruieren, dass gilt:

(i) $c^T z < c^T x$;

(ii) $z$ ist zulässig;

(iii) $z$ ist Basisvektor.

Dabei nutzen wir aus, dass die Basisvariablen von $z$ durch die Gleichungsnebenbedingungen eindeutig festgelegt werden: Damit $z$ zulässig ist, muss insbesondere $b = Az = A_B z_B + A_N z_N$ gelten; wegen der Invertierbarkeit von $A_B$ und mit $A_B x_B = b$ bedeutet dies[1]

$$z_B = A_B^{-1} b - A_B^{-1} A_N z_N = x_B - A_B^{-1} A_N z_N. \tag{14.3}$$

Wir überlegen nun, wie wir durch Wahl von $z_N$ die Bedingung (i) erreichen können. Aus (14.3) und $x_N = 0$ folgt.

$$\begin{aligned} c^T z &= c_B^T z_B + c_N^T z_N \\ &= c_B^T (x_B - A_B^{-1} A_N z_N) + c_N^T z_N \\ &= c_B^T x_B + (c_N^T - c_B^T A_B^{-1} A_N) z_N \\ &= c^T x + \left( c_N - A_N^T (A_B^T)^{-1} c_B \right)^T z_N. \end{aligned}$$

Setzen wir nun $y := (A_B^T)^{-1} c_B \in \mathbb{R}^m$, d. h. $y \in \mathbb{R}^m$ löst das lineare Gleichungssystem

$$A_B^T y = c_B, \tag{14.4}$$

und

$$u_N := c_N - A_N^T y \tag{14.5}$$

(der Übersichtlichkeit halber schleppen wir $u_B = 0$ mit, d. h. $u \in \mathbb{R}^n$), so erhalten wir

$$c^T z = c^T x + u_N^T z_N = c^T x + \sum_{j \in N} u_j z_j. \tag{14.6}$$

Um den Zielfunktionswert zu verkleinern, muss also die Summe auf der rechten Seite negativ sein. Ist nun $u_j \geq 0$ für alle $j \in N$, so müsste dafür mindestens ein $z_j$ negativ sein – aber damit wäre $z$ nicht zulässig. In diesem Falle ist $x$ also bereits der beste zulässige Vektor.

Lemma 14.2. *Gilt $u_N \geq 0$, so ist der zugehörige Basisvektor $x \in \mathbb{R}^n$ eine Lösung von (P) und $y := (A_B^T)^{-1} c_B \in \mathbb{R}^m$ eine Lösung von (D).*

*Beweis.* Nach Voraussetzung ist $x$ zulässig für (P). Aus $c_N - A_N^T y = u_N \geq 0$ folgt mit (14.4)

$$[A^T y]_j = (a_j)^T y \begin{cases} \leq c_j & j \in N, \\ = c_j & j \in B, \end{cases}$$

und damit Zulässigkeit von $y$ für das duale Problem (D). Insbesondere gilt $c_j - [A^T y]_j = 0$ für alle $j \in B$, und damit sind wegen $x_j = 0$ für $j \in N$ die Komplementaritätsbedingungen (14.1) erfüllt. Also ist $(x, y)$ nach Satz 12.7 ein primal-duales Paar. □

[1] In Folge ist stets $A_B^{-1} = (A_B)^{-1}$ bzw. $A_B^T = (A_B)^T$ zu lesen.

Wir nehmen nun an, es existiert ein $r \in N$ mit $u_r < 0$, und machen für $z_N$ den Ansatz

$$z_j = \begin{cases} t & j = r \\ 0, & j \in N \setminus \{r\}. \end{cases}$$

Für alle $t > 0$ gilt dann

$$c^T z = c^T x + t u_r < c^T x,$$

womit Bedingung (i) erfüllt wäre.

Nun zu Bedingung (ii), der Zulässigkeit. Weiterhin gilt

$$A_N z_N = \sum_{j \in N} z_j a_j = t a_r.$$

Damit $z$ zulässig ist, muss insbesondere $Az = b$ gelten; nach (14.3) bedeutet dass, dass wir

$$z_B = x_B - t A_B^{-1} a_r$$

wählen müssen. Führen wir $w_B$ ein als Lösung von

$$A_B w_B = a_r \tag{14.7}$$

(wieder mit $w_N = 0$, d. h. $w \in \mathbb{R}^n$), so lässt sich das schreiben als

$$z_B = x_B - t w_B.$$

Es bleibt noch, $z \geq 0$ zu erreichen. Offensichtlich ist $z_N = t e_r \geq 0$ für alle $t \geq 0$. Ist $w_B \leq 0$, so gilt wegen $x_B \geq 0$ auch $z_B \geq 0$ für alle $t \geq 0$; damit ist $z = z(t)$ für alle $t > 0$ zulässig, aber

$$c^T z(t) = c^T x + t u_r \to -\infty \qquad \text{für } t \to \infty,$$

das Problem (P) ist also unbeschränkt. Wir halten dies fest.

Lemma 14.3. *Gilt $w_i \leq 0$ für alle $i \in B$, so hat das Problem (P) keine Lösung.*

Wir nehmen nun an, es existiert (mindestens) ein $i \in B$ mit $w_i > 0$. Um $z_B = x_B - t w_B \geq 0$ zu gewährleisten, muss also gelten

$$t \leq \frac{x_i}{w_i} \qquad \text{für alle } i \in B \text{ mit } w_i > 0. \tag{14.8}$$

Jede Wahl von $t \geq 0$, die (14.8) erfüllt, ergibt also also einen zulässigen Vektor $z$.

Es bleibt die Bedingung (iii). Zunächst ist $z$ im Allgemeinen kein Basisvektor, denn für $r \in N$ ist ja $z_r = t > 0$ und $z_B > 0$ möglich; und damit haben wir mehr als die $n - m$ erlaubten Indizes im Träger von $z$. Wir müssen also $t$ so wählen, dass gilt $z_s = x_s - t w_s = 0$ für ein $s \in B$. Wir setzen daher

$$t = \min_{i \in B, w_i > 0} \frac{x_i}{w_i} = \frac{x_s}{w_s}. \tag{14.9}$$

Damit hat der Index $s$ die Basis verlassen, wofür wir den neuen Index $r$ hinzunehmen können, und wir haben einen neuen Basisvektor.

Lemma 14.4. *Der Vektor $z$ ist ein Basisvektor von $P^=(A,b)$ mit Basis $B' := (B \cup \{r\}) \setminus \{s\}$.*

*Beweis.* Nach Konstruktion ist $|B'| = |B| = m$ und $z_j = 0$ für $j \notin B'$. Es bleibt zu zeigen, dass die entsprechenden Spalten linear unabhängig sind. Seien $\alpha_i$, $i \in B'$, gegeben mit $\sum_{i\in B'} \alpha_i a_i = 0$. Wir zeigen nun $\alpha_i = 0$: Wegen (14.7) ist

$$\begin{aligned} 0 = \sum_{i\in B'} \alpha_i a_i &= \sum_{i\in B, i\neq s} \alpha_i a_i + \alpha_r a_r \\ &= \sum_{i\in B, i\neq s} \alpha_i a_i + \alpha_r A_B w_B \\ &= \sum_{i\in B, i\neq s} \alpha_i a_i + \alpha_r \sum_{i\in B} w_i a_i \\ &= \sum_{i\in B, i\neq s} (\alpha_i + \alpha_r w_i) a_i + \alpha_r w_s a_s. \end{aligned}$$

Da $B$ eine Basis ist, sind nach Voraussetzung die Spalten $a_i$, $i \in B$, linear unabhängig. Also gilt $\alpha_i + \alpha_r w_i = 0$ für alle $i \in B \setminus \{s\}$ und $\alpha_r w_s = 0$. Wegen $w_s > 0$ muss aber $\alpha_r = 0$ und damit auch $\alpha_i = 0$ für alle $i \in B \setminus \{s\}$ gelten. □

Wir fassen zusammen: Wählen wir $r$ mit $u_r < 0$ und $t > 0$ beliebig, so ist Bedingung (i) erfüllt; wählen wir $s$ und $t$ nach (14.9), so sind Bedingungen (ii) und (iii) erfüllt. Damit alle drei Bedingungen gleichzeitig gelten, muss daher $x_s/w_s > 0$ sein. Zwar ist stets $x_s \geq 0$ und $w_s > 0$, aber auch für $s \in B$ kann $x_s = 0$ sein (vergleiche den Beweis von Lemma 14.1 im Fall $|\operatorname{supp}(x)| < m$). In diesem Fall nennen wir $x$ einen *entarteten Basisvektor*. Ist $x$ aber nicht entartet (d. h. $x_B > 0$), so ist das Minimum in (14.9) strikt positiv, und wir erhalten eine echte Reduktion im Zielfunktionswert. Setzen wir nun $x = z$ und wiederholen die Prozedur, so ergibt dies das Simplex-Verfahren; die notwendigen Schritte sind in Algorithmus 14.1 zusammengefasst.

**Algorithmus 14.1 :** Simplex-Verfahren

**Input :** $A \in \mathbb{R}^{m\times n}$, $b \in \mathbb{R}^m$, $c \in \mathbb{R}^n$, Basisvektor $x \in P^=(A,b)$ mit Basis $B$
**Output :** Primal-duales Paar $(\bar{x}, \bar{y})$ oder Information „primales Problem unbeschränkt“

1 **while** *nicht fertig* **do**
2 Bestimme Nichtbasis $N$, Matrizen $A_B, A_N$
3 Löse $A_B^T y = c_B$
4 Setze $u_N = c_N - A_N^T y$
5 **if** $u_N \geq 0$ **then return** $(x, y)$ `// x, y optimal`
6 Wähle $r \in N$ mit $u_r < 0$
7 Löse $A_B w_B = a_r$
8 **if** $w_B \leq 0$ **then return** nicht lösbar `// Problem unbeschränkt`
9 Wähle $t = \min_{i\in B, w_i>0} \frac{x_i}{w_i} = \frac{x_s}{w_s}$
10 Setze $x_B = x_B - t w_B$
$x_r = t$
$B = (B \cup \{r\}) \setminus \{s\}$

Satz 14.5. *Sind alle in Algorithmus 14.1 auftretenden Basisvektoren nicht entartet, so bricht das Verfahren nach endlich vielen Schritten ab. Bei Abbruch in Zeile 5 ist der aktuelle Vektor $x$ Lösung von (P) und der in Schritt 3 berechnete Vektor $y$ Lösung von (D); bei Abbruch in Zeile 8 hat (P) keine Lösung.*

*Beweis.* Bezeichne $\{x^k\}_{k\in\mathbb{N}}$ die Folge der Iterierten im Simplex-Verfahren. Sind alle $x^k$ nicht entartet, so gilt nach Herleitung stets $c^T x^{k+1} < c^T x^k$. Insbesondere ist damit $x^{k+1} \neq x^j$ für alle $j \leq k$; ein Basisvektor wird also höchstens einmal vorkommen. Da es nach Lemma 14.1 nur endlich viele Basisvektoren gibt, muss das Verfahren abbrechen. Die weiteren Aussagen folgen aus Lemma 14.2 und Lemma 14.3. □

Aus praktischer Hinsicht bleiben aber noch Fragen offen:

(i) Wie berechnet man den ursprünglichen Basisvektor, von dem das Verfahren startet?

(ii) Wie wählt man die Indizes $r, s$ in Schritt 6 und Schritt 9, falls mehrere zur Auswahl stehen?

(iii) Wie löst man die Gleichungssysteme in Schritt 3 und Schritt 7 effizient?

Hier konzentrieren wir uns auf die ersten beiden Fragen; insbesondere wird die Antwort auf die zweite Frage erlauben, auch bei degenerierten Basisvektoren nach endlich vielen Schritten zum Optimum zu kommen. Die Antwort auf die letzte Frage, die auf die Aktualisierung einer $LU$- oder $QR$-Zerlegung einer Matrix bei Modifikation einer Spalte hinausläuft, überlassen wir der numerischen Linearen Algebra; siehe etwa [Golub & Van Loan 2013, Chapter 6.5] und [Geiger & Kanzow 2002b, Kapitel 3.4.3].

## 14.2 FINDEN EINES STARTVEKTORS

Ein zulässiger Basisvektor für Polyeder der Form $P^=(A, b)$ ist in der Regel nicht einfach zu erraten; eine Ausnahme sind Polyeder, die aus den Bedingungen $Ax \leq b$, $x \geq 0$ durch Einführung einer Schlupfvariablen $z$ erzeugt wurden. Ist nämlich $b \geq 0$, so sind diese Bedingungen äquivalent dazu, dass für $A' = (A, I)$ und $x' := (x, z)^T$ gilt

$$A'x' = b, \quad x' \geq 0,$$

und $(x, z) = (0, b)$ ist ein zulässiger Basisvektor mit Basis $B = \{n+1, \ldots, n+m\}$. Für den allgemeinen Fall kann man nun einen Basisvektor berechnen, indem man das Simplex-Verfahrens auf ein Hilfsproblem mit diesem Basisvektor anwendet. (Dies wird oft als „Phase I“ des Simplex-Verfahrens bezeichnet; der eigentliche Algorithmus 14.1 ist dann „Phase II“.) Tatsächlich taugt schon das obige Beispiel als Hilfsproblem. Sei wieder angenommen, dass $b \geq 0$ ist (da wir in (P) reine Gleichungsnebenbedingungen betrachten, ist die Annahme

$b \geq 0$ keine Einschränkung, denn wir können Zeilen $A_i$ mit $b_i < 0$ mit $-1$ multiplizieren, ohne das Problem zu verändern), und betrachte das Hilfsproblem

$$\text{(H)} \qquad \begin{cases} \min\limits_{\substack{x \in \mathbb{R}^n \\ z \in \mathbb{R}^m}} \mathbb{1}^T z \\ \text{mit } Ax + z = b, \\ \qquad x, z \geq 0 \end{cases}$$

Hier ist $z$ *keine* Schlupfvariable; man spricht von einer *künstlichen Variablen.*

Satz 14.6. *Für das Problem* (H) *gelten folgende Aussagen:*

*(i) Der Vektor $(x, z) = (0, b)$ ist zulässiger Basisvektor mit Basis $B = \{n+1, \ldots, n+m\}$.*

*(ii) Das Problem hat eine Lösung $(\bar{x}, \bar{z})$.*

*(iii) Ist $(\bar{x}, \bar{z})$ ein Basisvektor, so ist im Fall $\bar{z} \neq 0$ das ursprüngliche Problem* (P) *nicht zulässig; andernfalls ist $\bar{x}$ ein für $P^=(A, b)$ zulässiger Basisvektor.*

*Beweis.* Aussage (i) folgt wie oben aus $A'_B = I \in \mathbb{R}^{m \times m}$. Außerdem ist 0 zulässig für das zu (H) duale Problem, und damit folgt Aussage (ii) aus Satz 12.3.

Für Aussage (iii) sei die Lösung $(\bar{x}, \bar{z})$ ein Basisvektor. Sei $\bar{z} \neq 0$ und $\tilde{x}$ ein für (P) zulässiger Vektor. Dann ist aber auch $(x, z) = (\tilde{x}, 0)$ zulässig für (H) mit Funktionswert $\mathbb{1}^T z = 0 < \mathbb{1}^T \bar{z}$, im Widerspruch zur Optimalität von $(\bar{x}, \bar{z})$. Sei nun $\bar{z} = 0$. Dann ist $A\bar{x} = b$, und die zu positiven Komponenten von $\bar{x}$ gehörenden Spalten sind linear unabhängig. Sind dies $m$ Stück, so sind wir fertig; ansonsten können wir diese wegen $\operatorname{rang} A = m$ zu $m$ linear unabhängigen Spalten ergänzen und erhalten eine Basis für (P). □

Zu beachten ist, dass für $\bar{z} = 0$ zwar $\bar{x}$ ein Basisvektor für (P) ist, aber die zu $(\bar{x}, \bar{z})$ gehörige Basis $B$ nicht unbedingt eine Basis für $\bar{x}$ ist (nämlich falls in $B$ Indizes $i > n$ enthalten sind; dies entspricht einer degenerierten Basislösung). Vor Start der zweiten Phase muss also eventuell die Basis neu bestimmt werden. Dies wird in der im Folgenden beschriebenen Methode vermieden; hier wird das ursprüngliche Problem so modifiziert, dass der Start-Basisvektor für (H) auch für dieses Problem ein Basisvektor ist, die Lösung aber (unter bestimmten Voraussetzungen) mit der von (P) übereinstimmt. Wir betrachten für ein gegebenes $\gamma \geq 0$ das Problem

$$\text{(P}_\gamma\text{)} \qquad \begin{cases} \min\limits_{\substack{x \in \mathbb{R}^n \\ z \in \mathbb{R}^m}} c^T x + \gamma \mathbb{1}^T z \\ \text{mit } Ax + z = b, \\ \qquad x, z \geq 0 \end{cases}$$

Da die zulässige Menge die gleiche wie für (H) ist, ist natürlich $(0, b)$ wieder ein Basisvektor mit Basis $\{n+1, \ldots, n+m\}$. Wir können also das Simplex-Verfahren für $A' = (A, I)$ und

$x' = (x, z)^T$ mit diesem Vektor starten, und erhalten (falls das Verfahren abbricht) eine Lösung $(x_\gamma, z_\gamma)$ (die natürlich von $\gamma$ abhängt). Die überraschende Aussage ist nun, dass für $\gamma$ groß genug $\bar{x}$ bereits eine Lösung von (P) ist.

Lemma 14.7. *Ist $(x_\gamma, z_\gamma)$ Lösung von ($P_\gamma$) mit $z_\gamma = 0$, so ist $x_\gamma$ Lösung von (P).*

*Beweis.* Ist $(x_\gamma, z_\gamma)$ Lösung mit $z_\gamma = 0$, so ist der Vektor zulässig und erfüllt daher $Ax_\gamma = b$. Damit ist $x_\gamma$ auch zulässig für (P). Umgekehrt ist für jeden zulässigen Vektor $x \in P^=(A, b)$ der Vektor $(x, 0)$ zulässig für ($P_\gamma$). Aus der Optimalität von $(x_\gamma, z_\gamma)$ für ($P_\gamma$) folgt dann

$$c^T x_\gamma = c^T x_\gamma + \gamma \mathbb{1}^T z_\gamma \leq c^T x + \gamma \mathbb{1}^T 0 = c^T x,$$

d. h. $x_\gamma$ minimiert $c^T x$ über $P^=(A, b)$. □

Satz 14.8. *Es existiert ein $\bar{\gamma} \geq 0$, so dass für alle $\gamma > \bar{\gamma}$ das Problem ($P_\gamma$) lösbar ist und für jede Lösung $(x_\gamma, z_\gamma)$ gilt $z_\gamma = 0$.*

*Beweis.* Wir betrachten das zu ($P_\gamma$) duale Problem (vergleiche (D))

$$\begin{cases} \max\limits_{y \in \mathbb{R}^m} b^T y \\ \text{mit } \begin{pmatrix} A^T \\ I \end{pmatrix} y \leq \begin{pmatrix} c \\ \gamma \mathbb{1} \end{pmatrix}, \end{cases}$$

beziehungsweise

$$(D_\gamma) \qquad \begin{cases} \max\limits_{y \in \mathbb{R}^m} b^T y \\ \text{mit } A^T y \leq c \\ \qquad\quad y_i \leq \gamma, \quad 1 \leq i \leq m. \end{cases}$$

Nun ist (P) lösbar, und deshalb existiert nach Satz 12.4 eine Lösung $\bar{y}$ von (D); diese erfüllt insbesondere $A^T \bar{y} \leq c$. Setzen wir $\bar{\gamma} := \max_i \bar{y}_i$, so ist für alle $\gamma > \bar{\gamma}$ diese Lösung $\bar{y}$ auch zulässig für ($D_\gamma$). Da der zulässige Bereich für ($D_\gamma$) kleiner ist als der für (D), kann der optimale Zielfunktionswert $b^T y$ sicher nicht größer sein als $b^T \bar{y}$, d. h. $\bar{y}$ ist auch Lösung von ($D_\gamma$). Wieder nach Satz 12.4 hat deshalb ($P_\gamma$) eine Lösung $(x_\gamma, z_\gamma)$, und diese erfüllt die Komplementaritätsbedingungen

$$\begin{aligned} [x_\gamma]_j (c_j - [A^T \bar{y}]_j) &= 0, \quad 1 \leq j \leq n, \\ [z_\gamma]_i (\gamma - \bar{y}_i) &= 0, \quad 1 \leq i \leq m. \end{aligned}$$

Da $\gamma$ so gewählt war, dass $\bar{y}_i \leq \bar{\gamma} < \gamma$ für alle $1 \leq i \leq m$ gilt, folgt aus der zweiten Bedingung $z_\gamma = 0$, was zu zeigen war. □

In der Praxis startet man mit einer hinreichend großen Zahl $\gamma$ und erhöht diese z. B. durch Multiplikation mit einem festen Faktor, falls Algorithmus 14.1 mit einer Lösung $(x_\gamma, z_\gamma)$ mit $z_\gamma \neq 0$ (oder wegen Unzulässigkeit) abbricht. (Da $\gamma$ in den Nebenbedingungen nicht vorkommt, bleibt $(x_\gamma, z_\gamma)$ ein zulässiger Basisvektor, mit dem man den nächsten Durchlauf starten kann.)

## 14.3 VERMEIDUNG VON ZYKLEN

Wir untersuchen nun die Wahl von $r$ und $s$ in Schritt 6 und Schritt 9. Für beide Schritte ist eine Vielzahl von Strategien vorgeschlagen worden. Beginnen wir mit der Wahl des Index $r \in N$, der neu in die Basis aufgenommen wird. Wir erinnern, dass wir für $u_N$ ein $r$ suchen mit $u_r < 0$, da dann im nächsten Schritt gilt $c^T z = c^T x + u_N^T z_N < c^T x$. Mögliche Strategien sind die

(i) *Erster-Index-Regel*: wähle $r$ als den ersten Index $j \in N$, für den $u_j < 0$ ist;

(ii) *Kleinster-Index-Regel*: wähle $r$ als den *kleinsten* Index $j \in N$, für den $u_j < 0$ ist (da in der praktischen Implementierung die Einträge in $N$ nicht sortiert sein müssen, kann diese Regel zu einem anderen Ergebnis als (i) führen);

(iii) *Größter-Abstieg-Regel*: wähle $r$ so, dass $u_r = \min\{u_i : i \in N, u_i < 0\}$ ist;

(iv) *Steilster-Abstieg-Regel*: wähle $r$ so, dass für das entsprechende $w$ aus Schritt 7 der Wert $tu_r/\|w\|$ minimal wird. (Anstatt dem größten Abstieg suchen wir den größten Abstieg *pro Änderung* von $x$.)

Diese Regel sind nach wachsendem Aufwand sortiert; insbesondere kann man für (a) und (b) die Komponenten $u_j = c_j - A_j y$ nacheinander berechnen und abbrechen, sobald zum ersten Mal $u_j < 0$ ist (wobei für (b) die Menge $N$ zuerst sortiert werden muss). Bei (c) muss dagegen stets der gesamte Vektor $u_N$ berechnet werden, und (d) verlangt sogar jedesmal die Lösung eines Gleichungssystems. Dafür kann man erwarten, dass die aufwändigeren Regeln (die ja zu größerem Abstieg pro Schritt führen sollen) weniger Iterationen benötigen. Ob unter dem Strich weniger komplizierte Iterationen schneller sind als mehr einfache, hängt von der Problemgröße ab. (Erfahrungen favorisieren die einfachen Regeln für kleine bis mittlere Probleme (etwa $10^2$ Variablen) und die komplizierteren Regeln für größere Probleme).

Für die Wahl des Index $s$, der für $r$ die Basis verlässt, existieren weniger Regeln. Zur Erinnerung: Hier suchen wir zu diesem $s$ ein $t \geq 0$ so, dass gilt $z_s = x_s - tw_s = 0$. Verbreitet sind die

(i) *Kleinster-Index-Regel*: wähle $s$ als den kleinsten Index $j \in B$, für den gilt $\frac{x_j}{w_j} = \min_{i \in B, w_i > 0} \frac{x_i}{w_i}$;

(ii) *Lexikographische Regel*: wähle $s$ als den Index $i \in B$ mit $w_i > 0$, für den $\frac{(A_B^{-1}A)_i}{w_i}$ lexikographisch minimal ist (ein Vektor $x$ ist lexikographisch kleiner als $y$, falls $x \neq y$ und $x_i < y_i$ für den kleinsten Index $i$, für den $x_i \neq y_i$; z. B. ist $(1, 2)$ kleiner als $(2, 1)$).

Der Wert dieser Regeln ist, dass sie sicher stellen können, dass das Simplex-Verfahren auch bei Auftauchen von degenerierten Basisvektoren terminiert.

Dazu betrachten wir zuerst genauer, was einen degenerierten Basisvektor auszeichnet. Jeder Vektor $x \in \mathbb{R}^n$ wird durch $n$ linear unabhängige Gleichungen eindeutig bestimmt; dies gilt insbesondere für jeden Basisvektor. Für einen nichtdegenerierten Basisvektor kann man diese Gleichungen eindeutig zerlegen in $m$ Gleichungen $A_B x_B = b$ und $n - m$ Gleichungen $x_N = 0$. (Geometrisch bedeutet dies, dass die Ecke $\{x\}$ Schnitt von genau $n$ Facetten ist, von denen $n - m$ *Basisflächen* des positiven Orthanten $\{x \geq 0\}$ sind.) Ein degenerierter Basisvektor $x$ erfüllt immer noch $A_B x_B = b$ und $x_N = 0$, jedoch ist zusätzlich $x_i = 0$ für (mindestens) ein $i \in B$. Damit erfüllt $x$ die $n + 1$ Gleichungen $A_B x_B = b$, $x_B = 0$ und $x_i = 0$; $x$ ist also überbestimmt. (Geometrisch: Die Ecke $\{x\}$ ist Schnitt von mehr als $n$ Facetten.) Jede Wahl von $n$ dieser Gleichungen ist also eine unterschiedliche Beschreibung desselben Basisvektors (wobei man den Fall, dass $x_i = 0$ als Gleichung in $A_B x_B = b$ auftaucht, vom Fall $x_i = 0$ wegen Gleichheit in $x \geq 0$ unterscheiden muss).

Was bedeutet das für das Simplex-Verfahren? In Schritt 7 bestimmen wir das *minimale* $t$, so dass eine der neuen Basisvariablen $z_i = 0$ mit $i \in B$ erfüllt (nur so können wir garantieren, dass nicht mehrere Indizes die Basis verlassen). Da aber bereits eine alte Basisvariable $x_i = 0$ mit $i \in B$ erfüllt, passiert gar nichts – der Basisvektor ändert sich also nicht. Allerdings hat sich die Basis geändert (denn $s$ wird aus der alten Basis gewählt, die $r$ noch nicht enthält, und damit ist garantiert $r \neq s$). Insbesondere hat der störende Index $i$ in jedem Fall die Basis verlassen. Die Karten sind also neu gemischt, und es ist möglich, dass in der nächsten Iteration die degenerierte Ecke verlassen wird (wenn in der neuen Basis kein Index $i$ mit $x_i = 0$ mehr enthalten ist). Allerdings ist es genauso möglich, dass der gerade entfernte Index als $r$ wieder hinzugenommen wird – und das Spiel geht von vorne los. Da das Simplex-Verfahren deterministisch ist, wird es sich bei gleicher Ausgangsituation wieder gleich verhalten, und der Algorithmus gerät in einen *Zyklus*, in dem sich für den gleichen Basisvektor zwei oder mehr unterscheidliche Basen abwechseln. Wir müssen also verhindern, dass sich die Wahl von $(r, s)$ wiederholen kann.

Eine Möglichkeit ist, $r$ und $s$ jeweils nach der Kleinster-Index-Regel auszuwählen; dies wird als *Regel von Bland* bezeichnet. Für diese Regel kann man beweisen, dass das Simplex-Verfahren in jedem Fall terminiert.

Satz 14.9. *Unter Berücksichtigung der Regel von Bland terminiert Algorithmus 14.1 nach endlich vielen Schritten.*

*Beweis.* Wir nehmen an, dass in Algorithmus 14.1 ein Zyklus auftritt; es gibt also eine Folge von Iterierten $x^k = x^{k+1} = \cdots = x^{k+p}$ mit zugehörigen Basen $B^k, \ldots, B^{k+p}$ für die gilt

$B^{k+p} = B^k$ und $B^{k+l} \neq B^k$ für $1 \leq l < p$. Sei $r$ der *größte* Index, der während dieses Zyklus in die Basis aufgenommen wird. Um unnötige Notation zu vermeiden, bezeichnen wir die in diesem Schritt auftauchenden Größen einfach mit $x, B, y, u, w$, etc. Da $r$ nach der Kleinste-Index-Regel gewählt wurde, gilt dann $u_r < 0$ und $u_j \geq 0$ für alle $j \in B$ mit $j < r$. Da es sich um einen Zyklus handelt, muss $r$ die Basis irgendwann wieder zugunsten eines neuen Index verlassen. Die Größen in diesem Schritt bezeichnen wir mit $B', y', u', w'$, etc. (beachte, dass dann nach Annahme gilt $x' = x$). Insbesondere ist $s' = r$ und

$$t' = \min_{i \in B', w'_i > 0} \frac{x_i}{w'_i} = \frac{x_{s'}}{w'_{s'}} = 0$$

(denn sonst wäre $t' > 0$ und wir würden den Zyklus verlassen); es gilt also $x_{s'} = 0$. Da auch $s'$ nach der Kleinste-Index-Regel gewählt wurde, muss $w'_i \leq 0$ oder $x_i > 0$ für alle $i \in B'$ mit $i < s'$ gelten. Weiterhin gilt deshalb für

$$t'' := \begin{cases} 1 & \text{es gibt kein } i < s' \text{ mit } w'_i > 0, \\ \min\limits_{i \in B', w'_i > 0, i < s'} \dfrac{x_i}{w'_i} & \text{sonst,} \end{cases}$$

dass $t'' > t' = 0$ (sonst wäre $t'$ nicht der kleinste Index, für den das Minimum angenommen wird). Sei nun $r' \in N'$ mit $u'_{r'} < 0$ derjenige Index, der neu in die Basis aufgenommen wird. Wir definieren dann $z \in \mathbb{R}^n$ durch

$$z_i := \begin{cases} x_i - t'' w'_i & i \in B', \\ t'' & i = r', \\ 0 & \text{sonst.} \end{cases}$$

Dann ist $z_{r'} > 0$ und $z_i \geq 0$ für alle $i \in B'$ mit $i < s'$, da entweder $w'_i \leq 0$ oder $t''$ nach Konstruktion klein genug ist. Weiter gilt

$$\begin{aligned} Az &= \sum_{i \in B'} (x_i - t'' w'_i) a_i + t'' a_{r'} \\ &= \sum_{i \in B'} x_i a_i - t'' \left( \sum_{i \in B'} w'_i a_i - a_{r'} \right) \\ &= A_{B'} x_{B'} - t''(A_{B'} w'_{B'} - a_{r'}) \\ &= b, \end{aligned}$$

da $B'$ Basis zu $x$ ist und die Klammer nach Konstruktion von $w'_{B'}$ in Schritt 7 gleich Null ist.

Analog zur Herleitung von (14.6) (wofür die Bedingung $z \geq 0$ nicht verwendet wurde) erhalten wir daraus sowohl für das Paar $(x, B)$ als auch für $(x, B')$

$$c^T z - c^T x = u_N^T z_N = (u'_{N'})^T z_{N'}. \tag{14.10}$$

Wir untersuchen nun die einzelnen Summanden in $u_N^T z_N = \sum_{j \in N} u_j z_j$ auf ihr Vorzeichen und machen dabei eine Fallunterscheidung nach $j \in N$.

1. Fall: $j = r$: Dann ist $u_r < 0$ nach Wahl von $r$, und wegen $r = s' \in B'$ und $w'_{s'} > 0$ sowie $x_{s'} = 0$ gilt
$$z_r = x_r - t'' w'_r = 0 - t'' w'_{s'} < 0$$
und daher $u_r z_r > 0$.

2. Fall: $j < r$: Da $r$ nach der Kleinste-Index-Regel gewählt wurde, ist $u_j \geq 0$. Außerdem gilt nach Definition $z_j \geq 0$ (trivialerweise für $j \in N'$ und nach Bemerkung oben sonst) und damit $u_j z_j \geq 0$.

3. Fall: $j > r$: Da $r$ nach Annahme der größte Index ist, der während des Zyklus aus der Nichtbasis in die Basis genommen wurde, gilt $j \in N'$. Insbesondere gilt $r' \leq r < j$ und damit $z_j = 0$.

Zusammen erhalten wir $u_N^T z_N > 0$.

Für die zweite Summe erhalten wir aber sofort aus der Definition von $z$, dass

$$(u'_{N'})^T z_{N'} = \sum_{j \in N'} u'_j z_j = u'_{r'} t'' < 0$$

gilt, da $u'_{r'} < 0$ nach Wahl von $r'$ und $t'' > 0$ nach Konstruktion gilt. Damit haben wir einen Widerspruch, und es kann daher kein Zyklus auftreten. □

Wählt man $s$ nach der lexikographischen Regel, können unabhängig von der Wahlregel für $r$ ebenfalls keine Zyklen auftreten. Diese Regel entspricht der Störung der rechten Seite $b$ durch ein geeignetes $\eta \in \mathbb{R}^m$, $0 < \eta \ll 1$, so dass der gestörte Polyeder $P^=(A, b + \eta)$ keine degenerierten Ecken enthält; siehe [Gritzmann 2014, Kapitel 5.4].[2]

[2] In der Praxis treten Zyklen selten auf; man kann das Rechnen mit endlicher Genauigkeit als solch eine kleine Störung auffassen – da man Zahlen nur bis auf Maschinengenauigkeit vergleichen kann, wird in der Regel nie ein $x_j$ exakt gleich Null sein. Trotzdem muss man in Programmen diese Möglichkeit natürlich berücksichtigen.

# 15 DAS DUALE SIMPLEX-VERFAHREN

Das Simplex-Verfahren liefert gleichzeitig eine primale Lösung $\bar{x}$ und eine duale Lösung $\bar{y}$, indem so lange eine Folge zulässiger Vektoren $x$ und dazugehöriger komplementärer Vektoren $y := (A_B^T)^{-1} c_B$ erzeugt wird, bis $y$ auch dual zulässig (d. h. $u_N := c_N - A_N^T y \geq 0$) ist. Im *dualen* Simplex-Verfahren geht man umgekehrt vor: man hält $y$ dual zulässig und komplementär zu $x$, und versucht Zulässigkeit von $x$ zu erreichen. Eine Möglichkeit dafür ist die Anwendung von Algorithmus 14.1 auf das duale Problem (D), das wir dafür in Normalform transformieren müssen. Durch Einführen der Schlupfvariablen $z \geq 0$ erhalten wir

$$(\mathrm{D}_=) \qquad \begin{cases} \max\limits_{y \in \mathbb{R}^m, z \in \mathbb{R}^n} & b^T y \\ \text{mit } A^T y + z = c, \\ & z \geq 0. \end{cases}$$

Wir nehmen wieder an, dass $A$ vollen Zeilenrang hat, d. h. $\operatorname{rang}(A) = \operatorname{rang}(A^T) = m \leq n$. Um das Simplex-Verfahren auf ($\mathrm{D}_=$) anzuwenden, führen wir $D = (A^T, I) \in \mathbb{R}^{n \times (m+n)}$ ein. Ein Basisvektor $(y, z)$ hat als Basis also eine Teilmenge $H \subset \{1, \ldots, m+n\}$ mit $|H| = n$; die entsprechende Nichtbasis bezeichnen wir mit $M$. Da die Variablen $y \in \mathbb{R}^m$ keine Vorzeichenbedingungen erfüllen müssen, können sie nur durch Gleichheitsnebenbedingungen festgelegt werden. Existiert also ein optimaler Basisvektor $(\bar{y}, \bar{z})$, so muss in der dazugehörigen Basis $\bar{H}$ die Menge $\{1, \ldots, m\}$ enthalten sei; ansonsten wäre zumindest ein $y_i$ frei wählbar und damit $b^T y$ unbeschränkt. Wir schränken daher unsere Suche nach einer Lösung auf die entsprechenden Ecken ein. Es kommen also noch $n - m \geq 0$ Variablen aus $z \in \mathbb{R}^n$ für die Basis in Frage. Wir bezeichnen die entsprechenden Indizes aus $\{1, \ldots, n\}$ mit $N$ und setzen $B := \{1, \ldots, n\} \setminus N$, und erhalten damit die Basismatrix $D_H = (A^T, I_N)$ und die Nichtbasismatrix $D_M = I_B$. Die dualen Nichtbasisvariablen sind also genau die $z_B$. (Der Grund für diese vom primalen Simplex-Verfahren abweichende Notation wird später klar.)

Ein Basisvektor $(y, z)$ mit zugehöriger Basis $H$ ist genau dann zulässig, wenn $A^T y + z = c$ und $z \geq 0$ gilt. Die Gleichung können wir nun zerlegen in Zeilen $i \in B$ und Zeilen $i \in N$:

$$A_B^T y + z_B = c_B, \qquad A_N^T y + z_N = c_N.$$

Für die Nichtbasisvariablen gilt nun nach Definition $z_B = 0$. Weiter gilt, dass die Spalten von $D_H$ linear unabhängig sind; damit sind es insbesondere die Spalten von $A^T$ und auch

von $A_B^T$. Also sind die Zeilen von $A_B \in \mathbb{R}^{m\times m}$ linear unabhängig und damit $A_B$ regulär. Der Vektor $y$ ist daher eindeutig als Lösung von $A_B^T y = c_B$ bestimmt. Einsetzen in die zweite Gleichung liefert dann $z_N = c_N - A_N^T y$. Der Basisvektor $(y, z)$ ist also genau dann zulässig für ($D_=$), wenn $z_N \geq 0$ gilt.

Haben wir einen zulässigen Basisvektor $(y, z)$ gegeben, können wir Algorithmus 14.1 auf Problem ($D_=$) anwenden. Dabei wollen wir wegen $n \gg m$ vermeiden, mit der künstlich erweiterten Matrix $D$ und der zugehörigen Basis $H$ zu operieren. Wir betrachten dafür der Reihe nach die Schritte in Algorithmus 14.1:

**Schritt 3:** Löse $D_H^T x = (b, 0)^T$ für $x \in \mathbb{R}^n$. Nach Einsetzen von $D_H = (A^T, I_N)$ ist dies äquivalent zu $Ax = b$ und $x_N = 0$, womit sich die erste Gleichung vereinfacht zu $A_B x_B = b$.

**Schritt 4:** Setze $u_M = (b, 0)_M^T - D_M^T x$. Da $y$ stets zu den dualen Basisvariablen gehört, ist $(b, 0)_M = 0_B = 0$, und damit muss nur $u_B = -I_B x = -x_B$ betrachtet werden.

**Schritt 5:** Gilt $u_B \leq 0$ (im dualen Problem wird ja maximiert), so ist $x_B \geq 0$ und mit der Wahl $x_N = 0$ daher $x$ zulässig. Wegen $A_B^T y = c_B$ sind auch die Komplementaritätsbedingungen erfüllt, und damit ist $x$ Lösung von (P) und $y$ Lösung von (D).

**Schritt 6:** Andernfalls wähle ein $r \in B$ mit $x_B < 0$. Der Index $r$ tritt also in die duale Basis $N$ ein (und verlässt damit die primale Basis $B$). Beachte, dass der entsprechende Index in $M$ dann $m + r$ ist.

**Schritt 7:** Löse $D_H w_H = d_{m+r} = e_r$ für den Einheitsvektor $e_r \in \mathbb{R}^n$. Mit $w = (p, q)^T$ ist $w_H = (p, q_N)$. Weiter gilt $D_H = (A^T, I_N)$, und wegen $r \notin N$ kann die Gleichung zerlegt werden in

$$\begin{pmatrix} A_B^T & 0 \\ A_N^T & I \end{pmatrix} \begin{pmatrix} p \\ q_N \end{pmatrix} = \begin{pmatrix} [e_r]_B \\ 0 \end{pmatrix}.$$

Die erste Gleichung ergibt $A_B^T p = [e_r]_B$, und Einsetzen in die zweite $q_N = -A_N^T p$.

**Schritt 8:** Überprüfe, ob $w_H \leq 0$ gilt. Da wir stets alle Variablen in $y$ in der dualen Basis halten wollen, kommt für den Basistausch nur ein Index aus $N$ in Frage. Wir müssen also lediglich $q_N \leq 0$ überprüfen. Ist dies der Fall, ist das duale Problem unbeschränkt (und damit das primale Problem nach Folgerung 12.5 (ii) nicht zulässig).

**Schritt 9:** Wähle $t = \min_{i\in H, w_i>0} \frac{(y,z)_i}{w_i}$. Wegen der Beschränkung auf die Variable $z$ reduziert sich die Wahl auf

$$t = \min_{i\in N, q_i>0} \frac{z_i}{q_i} = \frac{z_s}{q_s}.$$

Der Index $s$ verlässt also die duale Basis $N$ und tritt damit in die primale Basis $B$ ein; der entsprechende Index in $H$ ist $m + s$.

**Schritt 10:** Setze $(y,z)_H = (y,z)_H - tw_H$, $(y,z)_{m+r} = t$, und $H = (H \cup \{m+r\}) \setminus \{m+s\}$.
Nach Festlegung von $H$ bedeutet dies

$$y = y - tp, \qquad z_N = z_N - tq_N, \qquad z_r = t, \qquad N = (N \cup \{r\}) \setminus \{s\}.$$

Damit erhalten wir das duale Simplex-Verfahren in Algorithmus 15.1. Verwendet man zur Wahl von $r$ und $s$ die Regel von Bland, folgt aus Satz 14.5 und den obigen Überlegungen zu den Abbruchbedingungen die Korrektheit des Verfahrens.

**Algorithmus 15.1 :** Duales Simplex-Verfahren

**Input :** $A \in \mathbb{R}^{m\times n}$, $b \in \mathbb{R}^m$, $c \in \mathbb{R}^n$,
Basisvektor $x \in \mathbb{R}^n$ (evtl. unzulässig) mit primaler Basis $B$, Nichtbasis $N$,
dualen Basisvektor $(y,z) \in \mathbb{R}^m \times \mathbb{R}^n$ mit $A_B^T y = c_B$ und $z_N := c_N - A_N^T y \geq 0$,
$z_B = 0$

**Output :** Primal-duales Paar $(\bar{x}, \bar{y})$ oder Information „primales Problem unzulässig“

1 **while** *nicht fertig* **do**
2 Bestimme primale Basis $B$, Matrizen $A_B, A_N$
3 Löse $A_B x_B = b$
4 **if** $x_B \geq 0$ **then return** $(x,y)$ mit $x_N = 0$ `// x, y optimal`
5 Wähle $r \in B$ mit $x_r < 0$
6 Löse $A_B^T p = [e_r]_B$ und setze $q_N = -A_N^T p$
7 **if** $q_N \leq 0$ **then return** nicht lösbar `// Problem unzulässig`
8 Wähle $t = \min\limits_{i\in N, q_i>0} \dfrac{z_i}{q_i} = \dfrac{z_s}{q_s}$
9 Setze $z_N = z_N - tq_N$
$z_r = t$
$y = y - tp$
$N = (N \cup \{r\}) \setminus \{s\}$

Der Vorteil des dualen Simplex-Verfahren gegenüber der primalen Variante ist, dass es einfacher sein kann, einen dual zulässigen Basisvektor zu finden als einen primal zulässigen. Dies ist insbesondere dann der Fall, wenn zu einem bereits gelösten Problem neue Restriktionen hinzugefügt werden: Angenommen, wir haben für das Problem (P) mit Hilfe des (primalen oder dualen) Simplex-Verfahrens einen optimalen Basisvektor $\bar{x} \in \mathbb{R}^n$ mit Basis $B$ und zugehöriger dualer Lösung $\bar{y} \in \mathbb{R}^m$ gefunden, und werden mit dem modifizierten Problem

$$\begin{cases} \min\limits_{x\in\mathbb{R}^n} c^T x \\ \text{mit } Ax = b, \\ \qquad x \geq 0, \\ \quad a^T x \leq \beta, \end{cases} \tag{15.1}$$

für $a \in \mathbb{R}^n$ und $\beta \in \mathbb{R}$ konfrontiert. Einführen einer Schlupfvariablen $\xi \in \mathbb{R}$ transformiert dieses Problem in die Normalform

$$(\mathrm{P}_+) \qquad \begin{cases} \min\limits_{x' \in \mathbb{R}^{n+1}} (c')^T x' \\ \quad \text{mit } A'x' = b', \\ \qquad\quad x' \geq 0 \end{cases}$$

mit

$$x' = \begin{pmatrix} x \\ \xi \end{pmatrix}, \qquad c' = \begin{pmatrix} c \\ 0 \end{pmatrix}, \qquad A' = \begin{pmatrix} A & 0 \\ a^T & 1 \end{pmatrix}, \qquad b' = \begin{pmatrix} b \\ \beta \end{pmatrix}.$$

Dann ist

$$x' := \begin{pmatrix} \bar{x} \\ \beta - a^T \bar{x} \end{pmatrix}$$

ein (möglicherweise degenerierter) Basisvektor zu ($\mathrm{P}_+$) mit Indexmenge $B' = B \cup \{n+1\}$, denn für $A_B$ invertierbar ist auch

$$A'_{B'} = \begin{pmatrix} A_B & 0 \\ a_B^T & 1 \end{pmatrix}$$

invertierbar. Ein zulässiger dualer Basisvektor ist dann

$$y' := \begin{pmatrix} \bar{y} \\ 0 \end{pmatrix}, \qquad z' := \begin{pmatrix} \bar{z} \\ 0 \end{pmatrix}$$

mit $\bar{z}_N = c_N - A_N^T \bar{y}$ und $\bar{z}_B = 0$, denn wegen $(n+1) \in B'$ und der Zulässigkeit von $(\bar{y}, \bar{z})$ gilt $z'_{N'} = \bar{z}_N \geq 0$, $z'_{B'} = (\bar{z}_B, 0)^T = 0$ sowie

$$(A'_{B'})^T y' = \begin{pmatrix} A_B^T & a_B \\ 0 & 1 \end{pmatrix} \begin{pmatrix} \bar{y} \\ 0 \end{pmatrix} = \begin{pmatrix} A_B^T \bar{y} \\ 0 \end{pmatrix} = \begin{pmatrix} c_B \\ 0 \end{pmatrix} = c'_{B'}.$$

Ist nun $a^T \bar{x} \leq \beta$, so ist $\bar{x}$ zulässig für (15.1) und damit optimal (denn durch die zusätzliche Restriktion kann der minimale Funktionswert nur grösser werden). Ansonsten ist $x'$ ein zulässiger und optimaler Basisvektor für ($\mathrm{P}_+$), und man stellt mit einem oder mehreren Schritten des dualen Simplexverfahrens, ausgehend von $N' := \{1, \ldots, n+1\} \setminus B'$ und $(y', z')$, die primale Zulässigkeit und damit die Optimalität von $x'$ her.

Ähnlich geht man bei Änderungen der rechten Seite vor: Angenommen, $\bar{x}$ ist ein optimaler Basisvektor für (P) mit Basis $B$, und in (P) wird $b$ durch ein $b'$ ersetzt. Gilt $x'_B := A_B^{-1} b' \geq 0$, so ist $x'$ mit $x'_N = 0$ zulässig. Da $b$ nicht in die Charakterisierung der dualen Zulässigkeit eingeht und sich die Basis nicht geändert hat, bleibt $\bar{y}$ zulässig und komplementär zu $x'$; damit ist $x'$ eine Lösung des modifizierten Problems. Ansonsten hat man einen zulässigen dualen Basisvektor $\bar{y}$ und kann mit dem dualen Simplex-Verfahren die primale Zulässigkeit von $x'$ erreichen.

# 16 PRIMAL-DUALE VERFAHREN

Die Grundidee des (primalen oder dualen) Simplex-Verfahrens ist, solange von Ecke zu Ecke des (primalen oder dualen) Zulässigkeitsbereiches zu springen, bis die Komplementaritätsbedingungen erfüllt sind und daher ein primal-duales Paar optimaler Lösungen gefunden ist. Nun kann man für jede Index-Wahl-Regel ein lineares Optimierungsproblem konstruieren – z. B. den sogenannten „Klee–Minty-Würfel“ – für den das entsprechende Simplex-Verfahren jede Ecke des zulässigen Polyeders absuchen muss, bevor das Optimum erreicht wird. Primal-duale Verfahren suchen dagegen für einen gegebenen (üblicherweise dual) zulässigen Vektor (der kein Basisvektor sein muss!) direkt einen zulässigen komplementären Vektor, der dann natürlich mit dem gegebenen Vektor ein primal-duales Lösungspaar ergibt. Existiert kein solcher Vektor, so wird der ursprüngliche Vektor geeignet modifiziert und das Verfahren wiederholt.

## 16.1 DAS PRIMAL-DUALE SIMPLEX-VERFAHREN

Wir betrachten wieder das primale Problem in Normalform

$$\text{(P)} \qquad \begin{cases} \min\limits_{x \in \mathbb{R}^n} \; c^T x \\ \text{mit } Ax = b \\ \qquad\quad x \geq 0 \end{cases}$$

mit zugehörigem dualen Problem

$$\text{(D)} \qquad \begin{cases} \max\limits_{y \in \mathbb{R}^n} \; b^T y \\ \text{mit } A^T y \leq c \end{cases}$$

und Komplementaritätsbedingungen

$$x_j (c_j - [A^T y]_j) = 0 \qquad \text{für alle } j = 1, \ldots, n \tag{16.1}$$

Wir nehmen wieder an, dass $A \in \mathbb{R}^{m \times n}$ vollen Rang hat und – ohne Einschränkung – dass $b \geq 0$ gilt.

Sei nun $y \in \mathbb{R}^m$ ein dual zulässiger Vektor und bezeichne mit

$$J := \{j \in \{1, \dots, n\} : [A^T y]_j = c_j\}$$

die in $y$ *aktive Menge* der Ungleichungen. (Da $J$ keine Basis sein muss, ist weder $|J| = m$ noch Regularität von $A_J$ gefordert!) Existiert nun ein $x \in P^=(A, b)$ mit $x_j = 0$ für alle $j \notin J$, so sind die Komplementaritätsbedingungen (16.1) erfüllt und damit $x$ Lösung von (P) und $y$ Lösung von (D). Fixieren wir $x_j = 0$ für $j \notin J$, müssen wir ein $x_J \in \mathbb{R}^{|J|}$ finden mit

$$A_J x_J = b, \qquad x_J \geq 0.$$

Wir gehen dafür analog zur Phase I im Simplex-Verfahren vor, indem wir das *reduzierte primale Problem*

$$(\mathrm{P}_J) \qquad \begin{cases} \min\limits_{x_J, z} \mathbb{1}^T z \\ \text{mit } A_J x_J + z = b \\ \qquad\qquad x_J \geq 0 \\ \qquad\qquad\ z \geq 0 \end{cases}$$

betrachten. Dieses Problem ist zulässig, da wegen $b \geq 0$ zum Beispiel $(x_J, z) = (0, b)$ ein zulässiger Vektor ist. Die Zielfunktion ist wegen $z \geq 0$ außerdem nach unten durch 0 beschränkt, so dass ($\mathrm{P}_J$) eine Lösung $(\bar{x}_J, \bar{z})$ mit Optimalwert $\bar{\mu} := \mathbb{1}^T \bar{z} = |\bar{z}|_1$ hat. Ist $\bar{\mu} = 0$, so gilt daher $\bar{z} = 0$ und damit $A_J \bar{x}_J = b$, d. h. $\bar{x}_J \geq 0$ ist unser gesuchter Vektor und $(\bar{x}, y)$ mit $\bar{x}_j = 0$ für $j \notin J$ ist ein primal-duales Lösungspaar.

Ist dagegen $\bar{\mu} > 0$, so existiert kein $x_J$ mit den gesuchten Eigenschaften, und $y$ kann nicht zu einem primal-dualen Paar ergänzt werden, d. h. $y$ kann keine Lösung von (D). Wir suchen daher einen neuen zulässigen Vektor $y'$ mit $b^T y' > b^T y$, wofür wir wie im Simplexverfahren den Ansatz $y' = y + tu$ mit $t > 0$ und $u \in \mathbb{R}^m$ geeignet machen. Da aber $y'$ nun kein Basisvektor sein muss, können wir dabei aggressiver vorgehen: Wir suchen $t$ und $u$ so, dass $b^T y' = b^T y + t b^T u$ maximal ist unter der Nebenbedingung

$$A^T y' = A^T y + t A^T u \leq c.$$

Die Richtung $u$ berechnen wir dafür als Lösung des *reduzierten dualen Problems*

$$(\mathrm{D}_J) \qquad \begin{cases} \max\limits_{u \in \mathbb{R}^m} b^T u \\ \text{mit } A_J^T u \leq 0 \\ \qquad\quad u \leq 1. \end{cases}$$

Dies ist gerade das duale Problem zu ($\mathrm{P}_J$). Da letzteres lösbar ist, hat ($\mathrm{D}_J$) nach Satz 12.4 ebenfalls eine Lösung $\bar{u}$ mit $b^T \bar{u} = \mathbb{1}^T \bar{z} = \bar{\mu} > 0$. Daher gilt für alle $t > 0$

$$b^T y' = b^T y + t b^T \bar{u} = b^T y + t\bar{\mu} > b^T y.$$

Nun müssen wir durch geeignete Wahl der Schrittlänge $t$ noch erreichen, dass $y + t\bar{u}$ zulässig ist. Ist $[A^T\bar{u}]_j \leq 0$ für alle $j \notin J$ (für aktive Indizes $j \in J$ folgt dies bereits aus der Zulässigkeit von $\bar{u}$), so gilt dies für jede Wahl von $t > 0$; wegen $b^T y' = b^T y + t\bar{\mu} \to \infty$ für $t \to \infty$ muss dann aber (D) unbeschränkt und daher (P) unzulässig sein. Ansonsten existiert

$$t = \min_{j \notin J, [A^T\bar{u}]_j > 0} \frac{c_j - [A^T y]_j}{[A^T\bar{u}]_j}, \tag{16.2}$$

für das nach Konstruktion $y' = y + t\bar{u}$ zulässig ist. Insbesondere gilt für jeden Index $j$, für den das Minimum angenommen wird, $[A^T y']_j = c_j$; diese Indizes tauchen also garantiert in der neuen aktiven Menge auf.

Damit ist unser primal-duales Verfahren im Prinzip schon vollständig spezifiziert; für die gleichzeitige Lösung der beiden reduzierten Hilfsprobleme bietet sich natürlich das (primale oder duale) Simplex-Verfahren an.

**Algorithmus 16.1 :** Primal-Duales-Simplex-Verfahren

**Input :** $A \in \mathbb{R}^{m\times n}$, $b \in \mathbb{R}^m$, $c \in \mathbb{R}^n$, zulässiger Vektor $y \in P(A^T, c)$
**Output :** Primal-duales Paar $(\bar{x}, \bar{y})$ oder Information „primales Problem unzulässig“

1 **while** *nicht fertig* **do**
2 Bestimme aktive Menge $J := \{j \in \{1, \dots, n\} : [A^T y]_j = c_j\}$
3 Bestimme Lösung $(x_J, z)$ von $(P_J)$ und Lösung $u$ von $(D_J)$ mit Simplex-Verfahren
4 **if** $b^T u = 0$ **then return** $(x, y)$ (mit $x_j = 0$ für $j \notin J$)
5 **else if** $[A^T u]_j \leq 0$ *für alle* $j \notin J$ **then return** Problem unbeschränkt
6 Wähle $t = \min_{j \notin J, [A^T\bar{u}]_j > 0} \frac{c_j - [A^T y]_j}{[A^T u]_j}$
7 Setze $y = y + tu$

Für den dual zulässigen Startvektor kann man im Fall $c \geq 0$ den Vektor $y = 0$ wählen; ansonsten löst man wie in Abschnitt 14.2 ein geeignet konstruiertes Hilfsproblem. Für die praktische Anwendung des (primalen) Simplex-Verfahrens in Schritt (3) kann man verwenden, dass eine Basislösung $(x^k_{J^k}, z^k)$ in Iteration $k$ bereits einen zulässigen Basisvektor für Iteration $k + 1$ liefert: Ist nämlich $j \in J^k$ mit $x^k_j > 0$ (sonst ist $x^k$ eine degenerierte Ecke, und Basiskomponenten mit $x^k_j = 0$ müssen ausgetauscht werden), so folgt aus der Komplementarität der reduzierten Probleme $[A^T u]_j = 0$ und damit

$$[A^T y^{k+1}]_j = [A^T y^k]_j + t[A^T u]_j = [A^T y^k]_j = c_j,$$

d. h. $j \in J^{k+1}$. Setzen wir also $x^{k+1}_j = x^k_j$ für $j \in J^k$ und $x^{k+1}_j = 0$ für $j \in J^{k+1} \setminus J^k$, so ist $(x^{k+1}, z^k)$ zulässig für das reduzierte Problem in Iteration $k + 1$ und $B^{k+1} = B^k$ weiterhin

eine Basis, da sich die Gleichungs-Nebenbedingungen nicht ändern. (In der ersten Iteration kann man für $b \geq 0$ mit dem Basisvektor $(x_J, z) = (0, b)$ starten.)

Allerdings ist dies kein optimaler Basisvektor, da für den Index $s \notin J$, für den das Minimum in (16.2) angenommen wird, $[A^T \bar{u}]_s > 0$ gilt; analog zur Herleitung vom Simplexverfahren kann man nun zeigen, dass durch Hinzunahme von $s$ in die Basis $B^{k+1}$ die Kostenfunktion strikt reduziert wird. Die *optimale* Basis $B^{k+1}$ ist also sicher nicht gleich $B^k$. Da es nur endlich viele verschiedene Basen gibt, muss (bei Vermeidung von Zyklen, z. B. durch Wahl von $s$ durch lexikographische Regel) das primal-duale Simplex-Verfahren nach endlich vielen Schritten terminieren.

## 16.2 KOMBINATORISCHE PRIMAL-DUALE ALGORITHMEN

Das primal-duale Simplex-Verfahren ist natürlich nur dann von Vorteil, wenn sich die reduzierten Probleme ($P_J$) und ($D_J$) deutlich schneller lösen lassen als die vollen Probleme (P) und (D). Dies ist insbesondere dann der Fall, wenn für die reduzierten Probleme eine Lösung direkt konstruiert werden kann, was für kombinatorische Probleme aus der Graphentheorie oft möglich ist.

Wir betrachten hier ein einfaches aber weitverbreitetes Problem: Das Finden kürzester Wege in gerichteten Graphen. Dabei ist ein *gerichteter Graph* ein Paar $G = (V, K)$, wobei $V$ eine endliche Menge (der *Knoten*) und $K$ eine Teilmenge der Menge aller geordneten Paare $(u, v)$ mit $u, v \in V$ (der *Kanten*, wobei $(u, v)$ der Kante entspricht, die $u$ mit $v$ verbindet). Beachte, dass $(u, v) \neq (v, u)$ und insbesondere für $(u, v) \in K$ nicht unbedingt auch $(v, u) \in K$ ist. Jeder Kante $k_j = (u, v) \in K$ sei ein *Kantengewicht* $c_j \geq 0$ zugeordnet (das die reale „Länge“ des Weges von $u$ nach $v$ angibt). Weiter seien zwei Knoten $s \in V$ (die *Quelle*) und $t \in V$ (die *Senke*) ausgezeichnet. Wir suchen nun den kürzesten $(s, t)$-Weg, d. h. denjenigen Weg von $s$ nach $t$, für den die Summe der Gewichte entlang der durchlaufenen Kanten minimal ist. Um dies als lineares Optimierungsproblem schreiben zu können, repräsentieren wir einen Weg $w$ durch einen Vektor $x \in \mathbb{R}^{|K|}$ mittels der Definition

$$x_j = \begin{cases} 1 & k_j \in w, \\ 0 & k_j \notin w. \end{cases}$$

Die Länge eines Weges $w$ ist dann gegeben durch

$$\sum_{k_j \in w} c_j = c^T x.$$

Die Bedingung, dass $w$ ein $(s, t)$-Weg ist, lässt sich wie folgt ausdrücken: Für jeden Knoten $v_i \in V$ gilt

$$\sum_{k_j=(p,v_i)\in K} x_j - \sum_{k_j=(v_i,q)\in K} x_j = b_i := \begin{cases} -1 & \text{falls } v_i = s, \\ 1 & \text{falls } v_i = t, \\ 0 & \text{sonst.} \end{cases}$$

In Worten: Die einzigen Knoten, die in $w$ genau einmal auftauchen, sind $s$ und $t$ – alle anderen Knoten kommen entweder gar nicht vor oder zweimal (einmal als Start- und einmal als Endknoten einer inneren Kante). Mit Hilfe der Inzidenzmatrix $A \in \mathbb{R}^{n\times m}$,

$$A_{ij} = \begin{cases} -1 & k_j = (v_i, q) \text{ für ein } q \in V, \\ 1 & k_j = (p, v_i) \text{ für ein } p \in V, \\ 0 & \text{sonst,} \end{cases}$$

können wir diese Bedingungen kurz schreiben als $Ax = b$. Ersetzen wir die Bedingung $x \in \{0,1\}^n$ durch $x \geq 0$, liegt also ein lineares Optimierungsproblem in Normalform (P) vor, auf das wir das primal-duale Simplex-Verfahren anwenden können. Allerdings hat die Matrix $A$ keinen vollen Rang (da sich alle Zeilen zu Null addieren); wir streichen daher die Zeile, die zum Knoten $s$ gehört; mit der Konvention $v_1 = s$ und $v_2 = t$ ist dies also die erste Zeile. Der Übersichtlichkeit halber passen wir die Nummerierung *nicht* an. Damit erhalten wir das duale Problem

$$\text{(D)} \qquad \begin{cases} \max\limits_{y\in\mathbb{R}^{|V|}} \; y_2 \\ \text{mit } y_r - y_i \leq c_j, \qquad k_j = (v_i, v_r) \in K, \\ \qquad\qquad y_1 = 0, \end{cases}$$

wobei wir $y_1 = 0$ gesetzt haben, da die zugehörige Ungleichung im primalen Problem eliminiert wurde.

Da $c \geq 0$ vorausgesetzt wurde, haben wir mit $y = 0$ einen dual zulässigen Startvektor. Die zu $y$ gehörende aktive Menge $J$ ist dann gegeben durch alle Indizes $j$ mit

$$c_j = [A^T y]_j = y_r - y_i \qquad \text{für } k_j = (v_i, v_r).$$

Das reduzierte duale Problem ist dann

$$\text{(D}_J\text{)} \qquad \begin{cases} \max\limits_{u\in\mathbb{R}^{|V|}} \; u_2 \\ \text{mit } u_r - u_i \leq 0, \qquad k_j = (v_i, v_r) \in K, j \in J, \\ \qquad\qquad u \leq 1, \\ \qquad\qquad u_1 = 0. \end{cases}$$

Der wesentliche Punkt ist nun, dass für dieses Problem eine Lösung direkt angegeben werden kann. Einen Knoten $v \in V$ nennen wir dafür *$J$-erreichbar*, wenn es einen $(s, v)$-Weg gibt, der nur Kanten $k_j$ mit $j \in J$ enthält.

Lemma 16.1. *Sei $u \in \mathbb{R}^{|V|}$ mit*

$$u_i = \begin{cases} 0 & v_i \textit{ ist } J\textit{-erreichbar}, \\ 1 & \textit{sonst}, \end{cases}$$

*und bezeichne $\bar{\mu}$ den optimalen Wert in* (D$_J$)*. Dann gilt:*

*(i)* $u$ *ist zulässig für* ($D_J$);

*(ii)* $\bar{\mu} = 0$ *genau dann, wenn* $t$ *ein* $J$*-erreichbarer Knoten ist;*

*(iii) ist* $u_2 \neq 0$*, so ist* $u_2 = 1$ *und* $u$ *Lösung von* ($D_J$).

*Beweis.* Zu (i): Angenommen, $u$ wäre unzulässig. Dann gibt es eine Kante $k_j = (v_i, v_r)$ für $j \in J$ mit $u_i = 0$ und $u_r = 1$ (die mit dieser Wahl von $u$ einzige Möglichkeit für $u_r - u_i > 0$). Also ist $v_i$, aber nicht $v_r$, $J$-erreichbar von $s$. Dies ist aber ein Widerspruch zu $j \in J$, denn jeden $(s, v_i)$-Weg in $J$ kann man mit $k_j$ zu einem $(s, v_r)$-Weg in $J$ verlängern.

Zu (ii): Ist $t$ nicht $J$-erreichbar, so ist $u$ ein zulässiger Vektor mit $u_2 = 1 > 0$, d. h. $\bar{\mu} > 0$. Ist $t$ aber $J$-erreichbar, so ergibt Addition der zu den entsprechenden Kanten gehörenden Ungleichungen in ($D_J$) für jeden zulässigen Vektor $u$ die Bedingung

$$0 \geq u_2 - u_1 = u_2,$$

und damit ist $\bar{u} = 0$ eine Lösung mit $\bar{\mu} = \bar{u}_2 = 0$.

Zu (iii): Ist $\bar{\mu} > 0$, kann $t$ nach (ii) nicht $J$-erreichbar sein. Also ist $u_2 = 1$, und wegen $0 < \bar{\mu} = \bar{u}_2 \leq 1$ wird in $u$ das Maximum angenommen. □

Ist $\bar{\mu} = 0$, so ist $y$ Lösung von (D); außerdem haben wir festgestellt, dass $t$ ein $J$-erreichbarer Knoten ist. Es gibt also einen $(s, t)$-Weg $w$, der nur Kanten $k_j$ mit $j \in J$ enthält. Definieren wir $x \in \mathbb{R}^n$ mit

$$x_j = \begin{cases} 1 & k_j \in w, \\ 0 & \text{sonst}, \end{cases}$$

so ist dieser Vektor zulässig und erfüllt wegen $x_j = 0$ für alle $j \notin J$ die Komplementaritätsbedingungen (16.1). Ansonsten wählen wir (falls das primale Problem zulässig ist, d. h. ein $(s, t)$-Weg existiert)

$$t = \min_{\substack{j \notin J, k_j = (v_i, v_r) \\ u_i = 0, u_r = 1}} c_j - (y_r - y_i),$$

setzen $y = y + tu$, und wiederholen die Prozedur. Man kann zeigen, dass dieses Verfahren nach endlich vielen Schritten terminiert, indem man nachweist, dass

(i) Kantenindizes, die in einer Iteration in die aktive Menge $J$ eintreten, diese in späteren Iterationen nicht mehr verlassen;

(ii) in jeder Iteration mindestens ein Kantenindex (nämlich der zu $t$ gehörende) in die aktive Menge eintritt.

Damit muss nach endlich vielen Iterationen $t$ ein $J$-erreichbarer Knoten sein (falls er überhaupt erreichbar ist) und das Verfahren abbrechen.

Die zur Konstruktion von $u$ notwendige Erreichbarkeitsbestimmung kann über eine Breitensuche durchgeführt werden. Dabei ist $U$ als *Warteschlange* zu verstehen, d. h. wird $v_r$ in Schritt 17 hinzugefügt, so kommt er an das Ende der Schlange, während er in Schritt 18 vom Anfang der Schlange genommen wird.

**Algorithmus 16.2 :** Primal-duales Verfahren für kürzeste Wege

**Input :** $(V, K)$, $c \in \mathbb{R}^n$, $y = 0$
**Output :** Kürzester Weg $x$ oder Information „kein $(s,t)$-Weg“

1 **while** *nicht fertig* **do**
2  Bestimme aktive Menge $J := \{j \in \{1,\dots,n\} : y_r - y_i = c_j, k_j = (v_i, v_r)\}$
3  Setze $U = \{v_1\}$, $u = \mathbb{1}$
4  **repeat** // Markiere $J$-erreichbare Knoten
5  **for** $v_i \in U$ **do** // Prüfe Knoten
6  **for** $k_j = (v_i, v_r) \in K$ *mit* $j \in J$ **do** // Prüfe alle $J$-erreichbaren Knoten
7  **if** $r = 2$ **then** // $t$ ist $J$-erreichbar
8  Setze $x = 0$
9  **repeat** // Gehe $(s,t)$-Weg zurück
10  Setze $x_j = 1$, $i = |\sigma_r|$
11  Finde $j$ mit $k_j = (v_i, v_r)$
12  **until** $i = 1$
13  **return** $x$ // $y$ ist optimal, $x$ kürzester Weg
14  **else if** $u_r = 1$ **then** // $v_r$ noch nicht besucht
15  Setze $u_r = 0$ // $v_r$ ist $J$-erreichbar
16  Setze $\sigma_r = i$ // Vorgänger
17  Setze $U := U \cup \{v_r\}$ // $v_r$ in Warteschlange
18  Setze $U := U \setminus \{v_i\}$ // $v_i$ besucht
19  **until** $U = \emptyset$
 // alle $J$-erreichbaren Knoten markiert
20  **if** $u_r - u_i \leq 0$ *für alle* $k_j = (v_i, v_r)$ *mit* $j \notin J$ **then**
21  **return** kein $(s,t)$-Weg
22  **else**
23  Wähle $t = \min_{\substack{j \notin J, k_j = (v_i, v_r) \\ u_i = 0, u_r = 1}} c_j - (y_r - y_i)$
24  Setze $y = y + tu$

# LITERATUR


W. Alt (2011), *Nichtlineare Optimierung. Eine Einführung in Theorie, Verfahren und Anwendungen*, 2. Aufl., Vieweg+Teubner, Wiesbaden.

J. Dennis & R. Schnabel (1996), *Numerical Methods for Unconstrained Optimization and Nonlinear Equations*, Bd. 16, Classics in Applied Mathematics, Society for Industrial & Applied Mathematics, doi: 10.1137/1.9781611971200.

C. Geiger & C. Kanzow (1999), *Numerische Verfahren zur Lösung unrestringierter Optimierungsaufgaben*, Springer, Berlin, doi: 10.1007/978-3-642-58582-1.

C. Geiger & C. Kanzow (2002a), *Theorie und Numerik restringierter Optimierungsaufgaben*, Springer, Berlin, doi: 10.1007/978-3-642-56004-0.

C. Geiger & C. Kanzow (2002b), *Theorie und Numerik restringierter Optimierungsaufgaben*, Springer, Berlin, doi: 10.1007/978-3-642-56004-0.

G. H. Golub & C. F. Van Loan (2013), *Matrix Computations*, 4. Aufl., Johns Hopkins University Press, Baltimore, MD.

P. Gritzmann (2014), *Grundlagen der Mathematischen Optimierung*, Springer, Berlin, doi: 10.1007/978-3-8348-2011-2.

M. Grötschel (2010), Lineare und Ganzzahlige Programmierung (ADM II), Vorlesungsskript, Institut für Mathematik, Technische Universität Berlin, url: http://www3.math.tu-berlin.de/Vorlesungen/WS09/LinOpt/index.de.html.

M. Hanke-Bourgeois (2009), *Grundlagen der Numerischen Mathematik und des Wissenschaftlichen Rechnens*, Vieweg+Teubner, Wiesbaden, doi: 10.1007/978-3-8348-9309-3.

C. T. Kelley (1999), *Iterative Methods for Optimization*, Bd. 18, Frontiers in Applied Mathematics, Society for Industrial & Applied Mathematics (SIAM), Philadelphia, PA, doi: 10.1137/1.9781611970920.

R. Schultz (2013), Optimierung 1, Vorlesungsskript, Fakultät für Mathematik, Universität Duisburg-Essen.

P. Spellucci (1993), *Numerische Verfahren der nichtlinearen Optimierung*, Birkhäuser Verlag, Basel, doi: 10.1007/978-3-0348-7214-0.

M. Ulbrich & S. Ulbrich (2012), *Nichtlineare Optimierung*, Birkhäuser, Basel, doi: 10.1007/978-3-0346-0654-7.