NICHTLINEARE OPTIMIERUNG

Vorlesungsskript, Wintersemester 2015/16

Christian Clason

Stand vom 21. März 2016

Fakultät für Mathematik
Universität Duisburg-Essen

INHALTSVERZEICHNIS

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

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

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:

1. 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.
2. In *inversen Problemen* sucht man einen Parameter u (zum Beispiel Röntgenabsorption von Gewebe in der Computertomographie), hat aber nur eine (gestörte) Messung y^δ zur Verfügung. Ist ein Modell bekannt, das für gegebenen Parameter u die entsprechende

Messung $y = Ku$ liefert, so kann man den unbekannten Parameter rekonstruieren, indem man das Problem

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

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

3. 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 Zeitpunk $t \in [0, T]$, so gehorcht sie 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

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

In dieser Vorlesung behandeln wir *nichtlineare Optimierungsprobleme*, in denen $f : \mathbb{R}^n \to \mathbb{R}$ differenzierbar ist und U durch (differenzierbare) Gleichungen und Ungleichungen beschrieben werden kann. (Den Fall nichtdifferenzierbarer Funktionen werden wir im letzten Teil kurz behandeln.) 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 existieren; ansonsten wählen wir eine und gehen einen Schritt in dieser Richtung. Wie diese Richtungen und die Schrittlänge zu wählen sind, und was im Falle von $U \neq \mathbb{R}^n$ zu beachten ist, ist der Inhalt dieser Vorlesung.

Dieses Skriptum basiert vor allem auf den folgenden Werken:

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

[2] J. Dennis und R. Schnabel (1996). *Numerical methods for unconstrained optimization and nonlinear equations*. Bd. 16. Classics in Applied Mathematics. Society for Industrial und Applied Mathematics. DOI: 10.1137/1.9781611971200

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

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

[5] C. T. Kelley (1999). *Iterative methods for optimization*. Bd. 18. Frontiers in Applied Mathematics. Society for Industrial und Applied Mathematics (SIAM), Philadelphia, PA. DOI: 10.1137/1.9781611970920

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

Teil I

GRUNDLAGEN

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.

1

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:

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

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 $x^T y \leqslant \|x\| \|y\|$. In endlichdimensionalen Vektorräumen sind alle Normen äquivalent: Ist $\|\cdot\|_*$ eine weitere Norm, so existieren Konstanten $c_1, c_2 > 0$ mit

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

Die durch die Norm beschriebene offene Kugel um x mit Radius ε 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). Dies ist dann und nur dann der Fall, wenn $x_i^k \to x_i$ für alle $1 \leqslant i \leqslant n$ gilt.

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

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

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

Eine Matrix ist *symmetrisch*, wenn $A^T := (a_{ji})_{ij} = A$ 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

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

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

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

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 gilt

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

ABLEITUNGEN IM $\mathbb{R}^n$

Eine Funktion $F : \mathbb{R}^n \to \mathbb{R}^m$, $x \mapsto (F_1(x), \ldots, F_m(x))^T$, 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 $F_i : \mathbb{R}^n \to \mathbb{R}$, $1 \leqslant i \leqslant m$, stetig sind. Die Funktion F heißt *Lipschitz-stetig*, wenn es ein $L > 0$ gibt mit

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

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*.

Eine Funktion $F : \mathbb{R}^n \to \mathbb{R}^m$ heißt differenzierbar in x, wenn es eine lineare Abbildung $F'(x) : \mathbb{R}^n \to \mathbb{R}^m$ und ein $\varepsilon > 0$ sowie $\varphi : B_\varepsilon(0) \to \mathbb{R}^m$ gibt mit $\lim_{h\to 0} \frac{\varphi(h)}{\|h\|} = 0$ und

$$F(x + h) = F(x) + F'(x)h + \varphi(h) \qquad \text{für alle } h \in B_\varepsilon(0).$$

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

In diesem Fall nennt man $F'(x)$ die *Jacobi-Matrix* von F. 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 mit der *transponierten* Jacobi-Matrix arbeiten, weshalb wir ihr einen eigenen Namen geben: $\nabla F(x) := F'(x)^T \in \mathbb{R}^{n\times m}$. Ist die Abbildung $x \mapsto \nabla F(x)$ stetig, so heißt F *stetig differenzierbar*.

Speziell für den Fall $m = 1$ erhalten wir damit für $f : \mathbb{R}^n \to \mathbb{R}$ 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 $d \in \mathbb{R}^n$ 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 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 $\theta \in (0,1)$ mit*

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

für $\xi := y + \theta(x - y)$.

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 $\theta \in (0,1)$ mit*

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

für $\xi := y + \theta(x - y)$.

Für vektorwertige Funktionen $F : \mathbb{R}^n \to \mathbb{R}^m$ gelten diese Mittelwertsätze nicht direkt (die Schwierigkeit ist, für alle Kompenenten ein einheitliches θ 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*

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

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

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

(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

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

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. Offensichtlich ist jede gleichmäßig konvexe Funktion strikt konvex und jede strikt konvexe Funktion konvex.

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*

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

(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*

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

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

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

Grenzübergang $\lambda \to 0$ ergibt dann

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

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

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

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

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

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

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

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

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

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

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

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*

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

(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*

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

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 $y := x + td$

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

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

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

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)

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

und Division durch $\frac{t^2}{2}$ 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

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

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

Im Fall $n = 1$ entspricht dieses Resultat der bekannten Tatsache, dass eine Funktion genau dann (strikt) konvex ist, wenn 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.

GRUNDLEGENDE BEGRIFFE UND EXISTENZ

2

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

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

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* Minimierung, ansonsten von *restringierter Optimierung*. 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 $f(\bar{x}) \geqslant f(x)$ 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

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

(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

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

(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) lokaler 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

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

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

Zu (ii): 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 eine der beiden Aussagen*

(i) X *ist beschränkt,*

(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), müssen wir über einen Widerspruch argumentieren. Wir zeigen zuerst, dass f nach unten beschränkt ist. Angenommen, dies wäre nicht der Fall. Dann existiert eine Folge $\{x^k\}_{k\in\mathbb{N}} \subset X$ mit $f(x^k) \to -\infty$. Aus der Koerzivität von f folgt dann, dass diese Folge beschränkt ist (sonst müsste ja $f(x^k) \to \infty$ gelten), d. h. es gibt ein $M > 0$ mit $\|x^k\| \leqslant M$ für alle $k \in \mathbb{N}$. 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

$$f(\bar{x}) = \lim_{m\to\infty} f(x^{k_m}) = -\infty.$$

Dies ist aber ein Widerspruch zu $f(\bar{x}) \in \mathbb{R}$. Also ist f nach unten beschränkt, und daher existiert ein Infimum $M := \inf_{x\in X} f(x) \in \mathbb{R}$. Aus der Definition 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

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

Aus der Koerzivität von f folgt wieder, dass diese Folge beschränkt ist und daher eine konvergente Teilfolge $\{x^{k_m}\}_{m\in\mathbb{N}}$ mit Grenzwert $\bar{x} \in X$ besitzt. Auch für diese Teilfolge gilt nun $f(x^{k_m}) \to M$, woraus zusammen mit der Stetigkeit von f folgt

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

Das Infimum wird also in $\bar{x} \in X$ angenommen und 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

OPTIMALITÄTSBEDINGUNGEN

3

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*

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

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

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

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*

$$\nabla f(\bar{x}) = 0. \tag{3.2}$$

Ein Punkt $\bar{x}$, der (3.2) 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 $\xi_t = \bar{x} + \theta t d$ mit $\theta \in (0, 1)$. Da $\nabla^2 f$ nach Voraussetzung stetig in 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) $\nabla^2 f(\bar{x})$ *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 $\xi_t = \bar{x} + \theta t d$ mit $\theta \in (0, 1)$. Wegen (ii) ist $\nabla^2 f(\bar{x})$ auch gleichmäßig positiv definit, es existiert also ein $\mu > 0$ mit

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

Zusammen mit der produktiven Null erhalten wir daraus die Abschätzung

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

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

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

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. □

Diese Bedingung ist wiederum 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 folgt dann

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

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

4 ABSTIEGSVERFAHREN

Satz 3.2 legt folgendes iterative Verfahren zur Bestimmung eines Minimums nahe:

Algorithmus 4.1 : Allgemeines Abstiegsverfahren

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

Der Beweis von Satz 3.2 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 σ_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 σ_k nachprüfen können), für die also das Verfahren *konvergiert*. Dies wollen wir wie folgt verstehen: Ein Verfahren *konvergiert global*, wenn jeder Häufungspunkt $\bar{x}$ einer entsprechend erzeugten Folge $\{x^k\}_{k\in\mathbb{N}}$ für beliebigen Startpunkt $x^0 \in \mathbb{R}^n$ ein stationärer Punkt ist. Beachte, dass man nicht mehr von einem Verfahren, das nur erste Ableitungen verwendet, 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 durch Algorithmus 4.1 erzeugte Folge $\{s^k\}_{k\in\mathbb{N}} \subset \mathbb{R}^n$ nennen wir Folge von *zulässigen Suchrichtungen*, wenn gilt:

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

(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 $s^k := -\nabla f(x^k)$ 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 $\{s^k\}_{k\in\mathbb{N}} \subset \mathbb{R}^n \setminus \{0\}$ eine durch Algorithmus 4.1 erzeugte Folge von Suchrichtungen. Existiert ein $\eta > 0$ mit*

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

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

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

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

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

Die Bedingung (4.1) wird *Winkelbedingung* genannt.

Nun zu den Schrittweiten σ_k. Eine durch Algorithmus 4.1 erzeugte Folge $\{\sigma^k\}_{k\in\mathbb{N}} \subset \mathbb{R}_{>0}$ nennen wir Folge von *zulässigen Schrittweiten*, wenn gilt:

(i) Alle σ_k führen zu einem Abstieg, d. h. $f(x^k + \sigma_k s^k) \leqslant f(x^k)$ für alle $k \in \mathbb{N}$;

(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 zugehörige Folge von Schrittweiten $\{\sigma_k\}_{k\in\mathbb{N}} \subset (0, \infty)$ *effizient*, falls ein $\theta > 0$ existiert mit

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

Effiziente Schrittweiten sind zulässig.

Lemma 4.2. *Sei $f : \mathbb{R}^n \to \mathbb{R}$ stetig differenzierbar und sei $\{s^k\}_{k\in\mathbb{N}} \subset \mathbb{R}^n$ eine durch Algorithmus 4.1 erzeugte Folge von Suchrichtungen. Ist die Folge $\{\sigma^k\}_{k\in\mathbb{N}}$ effizient, so ist sie auch zulässig.*

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

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

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 nun $\bar{x}$ 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 der Zulässigkeit der Schrittweiten σ_k, Bedingung (ii), folgt

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

und die Zulässigkeit der Suchrichtungen s^k, Bedingung (ii), 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.

SCHRITTWEITENREGELN

5

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 $\sigma = \arg\min_{\sigma \geqslant 0} f(x + \sigma s)$

Leider ist diese Regel nur in Ausnahmefällen 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.

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

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

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 γ 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*

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

Beweis. Nach Definition der Richtungsableitung und Wahl von γ 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 Ungleichung offensichtlich erfüllt ist, erhalten wir die Aussage. □

Realisieren lässt sich die Armijo-Regel über eine einfache Rückwärtssuche, 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, die für eine streng monoton wachsende Funktion $\varphi : [0,\infty) \to [0,\infty)$ die Bedingung*

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

erfüllt. 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 negativ ist, ist die erste Bedingung für Zulässigkeit stets erfüllt. Die zweite Bedingun 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 (beachte $\nabla f(x^k)^T s^k < 0$)

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

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

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

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

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

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

Nun ist nach Annahme $\{x^k\}_{k\in\mathbb{N}}$ beschränkt und ∇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 $\sigma_k \leqslant \rho \|s^k\|^{-1}$ 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 $\sigma_k = 1$ oder $\sigma_k \leqslant \beta$ und $\sigma_k/\beta > \rho\|s^k\|^{-1}$. In beiden Fällen haben wir wegen $\|s^k\| \geqslant \delta$

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

Die Armijo-Bedingung (5.1) garantiert also

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

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

POWELL–WOLFE-REGEL

Die Powell–Wolfe-Regel (manchmal auch Wolfe–Powell-Regel genannt) soll garantieren, dass auch bei kurzen Suchrichtungen s^k der tatsächliche Schritt $\sigma_k s^k$ 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

$$\nabla f(x^k + \sigma s^k)^T s^k \geqslant \eta \nabla f(x^k)^T s^k. \tag{5.3}$$

Anschaulich bedeutet diese Bedingung, dass neben dem Funktionswert auch die negative Steigung von f in Richtung s hinreichend reduziert wird; man bezeichnet daher (5.3) auch als *Krümmungsbedingung* und (5.1) und (5.3) zusammen als *Powell–Wolfe-Bedingungen*. Üblicherweise wird dabei γ klein und η 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*

$$\begin{aligned} f(x + \sigma s) - f(x) &\leqslant \gamma\sigma\nabla f(x)^T s, \\ \nabla f(x + \sigma s)^T s &\geqslant \eta\nabla f(x)^T s. \end{aligned}$$

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

$$\psi(\sigma) := f(x + \sigma s) - f(x) - \sigma\gamma\nabla f(x)^T s, \tag{5.4}$$

die nach Annahme an f 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 ψ 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

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

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

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

d. h. $\bar{\sigma} > 0$ erfüllt auch die Krümmungsbedingung (5.3). □

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 (5.4) definierten Funktion ψ mit Hilfe des Bisektions-Verfahrens. Dazu suchen wir in einem ersten Schritt eine Untergrenze σ_- mit $\psi(\sigma_-) \leqslant 0$ (d. h. die Armijo-Bedingung ist erfüllt) und eine Obergrenze σ_+ mit $\psi(\sigma_+) > 0$ (d. h. die Armijo-Bedingung ist verletzt). In einem zweiten Schritt halbieren wir das Intervall $[\sigma_-, \sigma_+]$ so lange, bis σ_- nahe genug an $\bar{\sigma}$ liegt, dass auch die Krümmungsbedingung erfüllt ist. Der folgende Algorithmus setzt dieses Verfahren basierend auf der Armijo-Regel 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
```

1 Bestimme σ_- mit Algorithmus 5.2 für $\beta = \frac{1}{2}$

2 **if** $\sigma_- = 1$ **then** // $\sigma = 1$ erfüllt Armijo-Bedingung

3 **if** $\nabla f(x + \sigma_- s)^T s \geqslant \eta \nabla f(x)^T s$ **then** // $\sigma = 1$ erfüllt Krümmungsbedingung

4 **return** σ_- // Akzeptiere Schrittweite 1

5 **else**

6 **for** $\sigma \in \{2^k : k \in \mathbb{N}\}$ **do** // Finde kleinstes σ_+, das Armijo-Bedingung verletzt

7 **if** $f(x + \sigma s) - f(x) > \sigma\gamma\nabla f(x)^T s$ **then** // σ 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
```

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** $f(x + \sigma s) - f(x) \leqslant \sigma\gamma\nabla f(x)^T s$ **then** // σ erfüllt Armijo-Bedingung

15 Setze $\sigma_- = \sigma$

16 **else** // σ verletzt Armijo-Bedingung

17 Setze $\sigma_+ = \sigma$

18 **return** σ_- // Akzeptiere Schrittweite σ_-

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 σ groß genug verletzt. Die Intervall-Phase endet also nach endlich vielen Schritten mit einem Paar (σ_-, σ_+) mit $\sigma_- < \sigma_+$, so dass σ_- die Armijo-Bedingung erfüllt, σ_+ 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 $\psi(\sigma_-) < 0 < \psi(\sigma_+)$. Angenommen, die Schleife in der Bisektions-Phase bricht nicht nach endlich vielen Schritten ab. Da in jedem Schritt entweder σ_- vergrößert oder σ_+ verkleinert wird, muss dann ein $\bar{\sigma}$ existieren mit $\sigma_- \to \bar{\sigma}^-$ und $\sigma_+ \to \bar{\sigma}^+$. Da mit f und ∇f auch ψ stetig ist, folgt $\psi(\bar{\sigma}) = 0$. Wie im Beweis von Lemma 5.3 erhält man nun aus dem Vorzeichenwechsel von ψ in $\bar{\sigma}$, dass $\psi'(\bar{\sigma}) \geqslant 0$ und damit

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

gilt. Wegen der Stetigkeit von ∇f ist diese Ungleichung auch erfüllt für σ_- 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 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 ∇f, dass für alle $k \in \mathbb{N}$ gilt

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

Daraus folgt

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

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

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

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

DAS GRADIENTENVERFAHREN

6

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* $f : \mathbb{R}^n \to \mathbb{R}$ *stetig 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

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

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 $\|\nabla f(x^k)\| \leqslant \varepsilon$ für eine vorgegebene Toleranz $\varepsilon > 0$ (z. B. $\varepsilon = 10^{-8}$) erreicht ist.

Für die globale Konvergenz dieses Verfahrens können wir den abstrakten Konvergenzsatz 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 $c \in \mathbb{R}^n$ die Funktion

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

Man rechnet leicht nach, dass dann

$$\begin{aligned} \nabla f(x) &= Ax + c, \\ \nabla^2 f(x) &= A, \end{aligned}$$

ist. In diesem Fall ist sogar die Minimierungsregel 5.1 praktikabel. Für $x, s \in \mathbb{R}^n$ wird dabei σ 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 φ strikt konvex. Nach Satz 2.1 ist der eindeutige Minimierer $\bar{\sigma}$ von φ charakterisiert durch

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

Auflösen nach $\bar{\sigma}$ und Verwenden von $s = -\nabla f(x) = -(Ax + c)$ 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* $\kappa := \lambda_n/\lambda_1 \geqslant 1$ 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 für alle $x \in \mathbb{R}^n \setminus \{0\}$*

$$\frac{(x^TAx)(x^TA^{-1}x)}{(x^Tx)^2} \leqslant \frac{1}{4}\left(\kappa^{\frac{1}{2}} + \kappa^{-\frac{1}{2}}\right)^2.$$

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 $1 \leqslant i \leqslant n$. Da die Funktion $z \mapsto z + z^{-1}$ auf $(0, 1]$ und $[1, \infty)$ monoton und $\mu^{-1}\lambda_i \leqslant \mu^{-1}\lambda_n = \kappa^{1/2}$ ist, erhalten wir

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

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

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

Wir wenden jetzt auf die linke Seite die *Youngsche Ungleichung*

$$(ab)^{\frac{1}{2}} \leqslant \frac{1}{2}(\mu^{-1}a + \mu b)$$

für $a = x^TAx$ und $b = x^TA^{-1}x$ an, quadrieren beide Seiten, und erhalten die gewünschte 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 $c \in \mathbb{R}^n$. 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 Schrittweitenregel 5.1 anstelle von 5.2. Dann gilt*

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

sowie

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

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 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 $y = x^k$ für $x = x^{k+1} = x^k - \sigma_k s^k$ wegen $s^k = -\nabla f(x^k)$

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

Wir verwenden jetzt die Schrittweitenwahl (6.2) und erhalten

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

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

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

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 $x \in \mathbb{R}^n$

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

Damit erhalten wir

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

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* $c \in \mathbb{R}^n$. *Dann gilt für das Gradientenverfahren mit Minimierungsregel die Fehlerabschätzung*

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

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.

NEWTON-ARTIGE VERFAHREN

7

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 σ_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 $0 < \mu_1 \leqslant \mu_2$*, so dass für alle* $k \in \mathbb{N}$ *die Eigenwerte von* H_k *die Abschätzung*

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

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 $\nabla f(x^k) \neq 0$ daher auch $s^k = -H_k^{-1}\nabla f(x^k) \neq 0$, und unter Verwendung von (1.1) folgt

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

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

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

Zusammen erhalten wir

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

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 $\|x^k - \bar{x}\| \leqslant \varepsilon$ 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

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

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

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

(iii) *quadratisch*, falls ein $C > 0$ existiert mit

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

(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: $\|x^{k+1} - \bar{x}\| = o(\|x^k - \bar{x}\|)$ für superlineare Konvergenz bzw. $\|x^{k+1} - \bar{x}\| = \mathcal{O}(\|x^k - \bar{x}\|^2)$ für quadratische Konvergenz.

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

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

(ii) quadratisch konvergent genau dann, wenn gilt

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

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 und der Definition der superlinearen Konvergenz folgt sofort

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

□

Der Nutzen dieser Eigenschaft liegt – neben der Verwendung in den folgenden Beweisen – darin, dass wir (nur!) für superlinear konvergente Folgen das "optimale" Gütekriterium $\|x^k - \bar{x}\| \leqslant \varepsilon$ durch das tatsächlich überprüfbare Kriterium $\|x^{k+1} - x^k\| \leqslant \varepsilon$ 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 Index* $k_0 \in \mathbb{N}$ *und eine Konstante* $\beta > 0$ *so, dass gilt*

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

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

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

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

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

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

Dies rechtfertigt, als Abbruchkriterium für Algorithmen die Bedingung $\|\nabla f(x^k)\| \leqslant \varepsilon$ anstelle von $\nabla f(x^k) = 0$ einzusetzen, denn für hinreichend kleines $\varepsilon > 0$ ist 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*

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

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*

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

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

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

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

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

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

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

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

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

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*

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

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

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

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

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

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

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

Damit ist auch

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

für alle $k \geqslant k_0$. 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

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

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

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

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

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

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

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

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 $\bar{x} \in \mathbb{R}$, 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) $\|\nabla f(x^k) + \nabla^2 f(x^k)(x^k - x^{k+1})\| = o(\|x^{k+1} - x^k\|)$,

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

Beweis. *(iii)* $\Rightarrow$ *(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

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

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

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

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

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

Zusammen mit (7.7) erhalten wir daraus

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

und daher

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

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

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

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

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 $\|\nabla^2 f(x)\| \neq 0$ 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 $L > 0$ an, und mit Satz 1.3 und $x^k \to \bar{x}$ folgt für alle $k \in \mathbb{N}$ groß genug

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

Wegen $\nabla f(\bar{x}) = 0$ erhalten wir daraus durch großzügiges Erweitern

$$\|\nabla f(x^{k+1})\| \leqslant \left(L \frac{\|x^{k+1} - \bar{x}\|}{\|x^k - \bar{x}\|} \cdot \frac{\|x^k - \bar{x}\|}{\|x^{k+1} - x^k\|} \right) \|x^{k+1} - x^k\|.$$

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, dass

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

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 sogenannte *Dennis–Moré-Bedingung* für die superlineare Konvergenz von Algorithmus 4.1.

Folgerung 7.10. *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) $\|\nabla f(x^k) + \nabla^2 f(x^k)(x^{k+1} - x^k)\| = \mathcal{O}(\|x^{k+1} - x^k\|^2)$,

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

Beweis. *(ii)* $\Rightarrow$ *(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

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

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

$$\frac{\|x^{k+1} - \bar{x}\|}{\|x^k - \bar{x}\|^2} \leqslant c \left(\frac{\|\nabla f(x^k) + \nabla^2 f(x^k)(x^{k+1} - \bar{x})\|}{\|x^{k+1} - x^k\|^2} \cdot \frac{\|x^{k+1} - \bar{x}\|^2}{\|x^k - \bar{x}\|^2} + \frac{\|\nabla f(x^k) - \nabla f(\bar{x}) - \nabla^2 f(x^k)(x^k - \bar{x})\|}{\|x^k - \bar{x}\|^2} \right).$$

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) $\Rightarrow$ *(ii):* Aus der Identität (7.9) folgt auch die Abschätzung

$$\frac{\|\nabla f(x^k) + \nabla^2 f(x^k)(x^{k+1} - x^k)\|}{\|x^k - \bar{x}\|^2} \leqslant \frac{\|\nabla f(x^k) - \nabla f(\bar{x}) - \nabla^2 f(x^k)(x^k - \bar{x})\|}{\|x^k - \bar{x}\|^2} + \|\nabla^2 f(x^k)\| \frac{\|x^{k+1} - \bar{x}\|}{\|x^k - \bar{x}\|^2}.$$

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) $\Leftrightarrow$ *(iii)* zeigt man analog zu Satz 7.9 unter Verwendung von Lemma 7.2. □

NEWTON-VERFAHREN

8

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 $\nabla f(x) = 0$.

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-artigen 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

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

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

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

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 folgt

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

Durch Induktion folgt daraus

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

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*.

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** $\nabla f(x^k)^T d^k \leqslant -\rho\|d\|^p$ **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 setzen dafür

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

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

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

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 $\|\nabla^2 f(x^k)\| \leqslant C$ für alle $k \in K$. Die Definition des Newton-Schritts ergibt nun

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

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

$$\|\nabla f(x^k)\|^{p-1} \leqslant \big(C\|s^k\|\big)^{p-1} \leqslant \frac{C^{p-1}}{\rho} \frac{-\nabla f(x^k)^T s^k}{\|s^k\|} \to 0$$

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

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

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

Also ist

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

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. □

Wir zeigen als nächstes, dass unter geeigneten Voraussetzungen die ganze Folge $\{x^k\}_{k\in\mathbb{N}}$ gegen einen Minimierer konvergiert. Dafür verwenden wir das folgende nützliche Lemma, das wir auch im weiteren Verlauf der Vorlesung heranziehen werden.

Lemma 8.3. *Sei $\bar{x} \in X$ 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 $\{x^k\}_{k\in K}$. Dann konvergiert die gesamte Folge $\{x^k\}_{k\in\mathbb{N}}$ gegen $\bar{x}$.*

Beweis. Da $\bar{x} \in X$ 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 $\|x^{l(k)} - \bar{x}\| \leqslant \frac{\varepsilon}{2}$ für alle $k \geqslant k_0$. Daraus folgt aber

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

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 ∇f und $\sigma_k \in (0, 1]$ nach Definition der Armijo-Regel

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

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

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

Damit ist Lemma 8.3 anwendbar und liefert die Aussage. □

Natürlich soll auch das globalisierte Newton-Verfahren superlinear konvergieren, damit sich der ganze Mehraufwand gegenüber dem Gradientenverfahren lohnt. Dazu zeigen wir, dass Algorithmus 8.2 irgendwann in das 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 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

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

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

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

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

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

für alle $k \geqslant k_0$.

Nun gilt für alle $k \geqslant k_0$

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

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

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

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

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

woraus die Aussage folgt. □

Damit haben wir nun alles beisammen, 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

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

sowie (im Falle eines Newton-Schrittes)

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

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

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

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

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

Also ist für alle $k \geqslant k_1$

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

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. □

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. Die Dennis–Moré-Bedingung garantiert aber die superlineare Konvergenz, solange wir nur bei Annäherung an einen stationären Punkt den Fehler 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 η_k bestimmen wir s^k mit

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

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 $\|\nabla^2 f(x^k)s^k + \nabla f(x^k)\| \leqslant \eta_k \|\nabla f(x^k)\|$
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 η_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

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

Dies definiert wegen

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

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 $\eta_k \leqslant \bar{\eta}$ für ein beliebiges $\bar{\eta} \in (0, 1)$, so konvergiert Algorithmus 8.1 lokal linear bezüglich $\|\cdot\|_*$ 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

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

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

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

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

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

Da $\bar{x}$ ein stationärer Punkt ist, ist nach dem Mittelwertsatz 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 $\|\cdot\|_*$ gilt daher für alle $x \in B_{\varepsilon_2}(\bar{x})$ die Abschätzung

$$\begin{aligned}\|\nabla f(x)\| &\leqslant \frac{\delta}{\mu_2}\|x-\bar{x}\| + \|\nabla^2 f(\bar{x})(x-\bar{x})\| \\ &\leqslant \frac{\delta}{\mu_2}\left(\mu_2\|x-\bar{x}\|_*\right) + \|x-\bar{x}\|_* = (1+\delta)\|x-\bar{x}\|_*.\end{aligned}$$

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

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

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

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

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 $\mu_1 \leqslant \mu_2$ 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

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

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

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

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

$$\|x^1 - \bar{x}\| \leqslant \mu_2 \|x^1 - \bar{x}\|_* \leqslant \mu_2 \eta \|x^0 - \bar{x}\|_* \leqslant \eta \frac{\mu_2}{\mu_1} \|x^0 - \bar{x}\| < \varepsilon.$$

Mit Induktion folgt daraus

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

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 $\|\cdot\|_*$. □

Aus der Konvergenz bezüglich $\|\cdot\|_*$ 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 $\eta_k = \mathcal{O}(\|\nabla f(x^k)\|)$, 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

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

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 $\|\nabla^2 f(x^k)\| \leqslant C$ für alle $k \geqslant k_0$. Daraus folgt

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

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 Kapitel 8.2 zeigen; für Details sei auf [Geiger und Kanzow 1999, Kapitel 10.2] verwiesen.

QUASI-NEWTON-VERFAHREN

9

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:

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

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.

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

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

Weiterhin gilt für jede Orthonormalbasis $\{v_1, \dots, v_n\}$ 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 $A^T A$, siehe z. B. [Geiger und 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 $\{s/\|s\|, v_2, \dots, v_n\}$ des $\mathbb{R}^n$ ergänzen. Dann ist

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

Wegen der Quasi-Newton-Gleichung muss der erste Term gleich $\|y - Hs\|^2$ sein, dort haben wir also keine Freiheit für die Minimierung. Der zweite Term ist jedoch minimal, wenn $H_+ v_i = Hv_i$ für $i = 2, \dots, n$ gilt. Dies erreichen wir durch die Wahl

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

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 $\mathrm{SPD}(n)$.

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

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

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 $M \in \mathrm{SPD}(n)$, die die Quasi-Newton-Gleichung erfüllt: Gilt nämlich $s^T y \leqslant 0$, so folgt aus $Ms = y$ sofort $s^T Ms = s^T y \leqslant 0$, 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, dass H_+ invertierbar ist, 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 und Kanzow 1999, Satz 11.8]). *Seien* $B \in \mathrm{SPD}(n)$ *und* $s, y \in \mathbb{R}^n$ *mit* $s^T y > 0$ *gegeben. Dann existiert eine Matrix* $W \in \mathrm{SPD}(n)$ *mit* $W^2 s = y$*, und die Lösung von*

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

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

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

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*

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

Durch zweimaliges Anwenden erhält man daraus die folgenden Updates

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

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

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

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

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

Cauchy–Schwarz-Ungleichung

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

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:

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

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.

Algorithmus 9.1 : Lokales Broyden-Verfahren

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

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

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 $\nabla f(\bar{x})$ 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*

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

Beweis. Aus der Definition des Broyden-Updates folgt

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

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

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

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 $\bar{x}$

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

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

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

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

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

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

womit auch die Aussage bewiesen wäre. Es gelte nun (9.7) und (9.8) für alle $i = 0, \dots, k-1$. Wir zeigen zuerst, dass (9.7) für k gilt. Aus den beiden Induktionsvoraussetzungen zusammen mit Lemma 9.6 folgt

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

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

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

Nach Wahl von ε gilt daher

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

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

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

Nach dem Banach-Lemma ist also H_k invertierbar mit

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

Aus dem Iterationsschritt folgt 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

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

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

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

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

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

und damit die gewünschte Aussage. □

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) $\|A(I - \frac{vv^T}{v^Tv})\|_F \leqslant \|A\|_F - \frac{1}{2\|A\|_F}\left(\frac{\|Av\|}{\|v\|}\right)^2$ 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 nun aus (9.5) und (9.6) sowie (9.8)

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

Durch Umformen erhalten wir daraus

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

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

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

Nun gilt nach Voraussetzung $\|E_0\|_F - \|E_{m+1}\|_F \leqslant \|E_0\|_F \leqslant \sqrt{n}\delta$ sowie wegen (9.8) und der geometrischen Reihe

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

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

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

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

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

und aus Folgerung 7.10 erhalten wir die superlineare Konvergenz. □

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

Algorithmus 9.2 : Lokales BFGS-Verfahren

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

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 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 und 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 Lemma 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 und 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 $B_0 \in \mathrm{SPD}(n)$ der Algorithmus 9.2 superlinear gegen $\bar{x}$ konvergiert.*

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

Setze $k = 0$
while $\|\nabla f(x^k)\| > 0$ **do**
  Setze $s^k = -B_k \nabla f(x^k)$
  Bestimme $\sigma_k > 0$ mit Algorithmus 5.3 für $\gamma \in (0, 1/2)$
  Setze $x^{k+1} = x^k + \sigma_k s^k$
  Setze $d^k := \sigma_k s^k$, $y^k = \nabla f(x^{k+1}) - \nabla f(x^k)$
  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 $\nabla f(x^k) \neq 0$, $B_k \in \mathrm{SPD}(n)$, und $\sigma_k > 0$ nach der Powell–Wolfe-Regel gewählt, so ist auch $B_{k+1} \in \mathrm{SPD}(n)$.*

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

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

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

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

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 $0 < \mu_1 \leqslant \mu_2$, so dass für alle $k \in \mathbb{N}$ die Eigenwerte von B_k die Abschätzung*

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

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 9.8 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 und 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 $B_k = B_0$).

TRUST-REGION-VERFAHREN

10

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*) $K_\Delta := \overline{B_\Delta(0)} = \{x \in \mathbb{R}^n : \|x\| \leqslant \Delta\}$ 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_Δ kompakt 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 Δ im Verlauf der Iteration richtig zu wählen. Der Ansatz ist hier, dies nicht in jedem Schritt neu zu tun, sondern Δ 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.

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

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

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

$$\min_{d \in K_{\Delta_k}} q_k(d) \tag{10.3}$$

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

$$\rho_k := \frac{f(x^k) - f(x^k + s^k)}{f(x^k) - q_k(s^k)} \tag{10.4}$$

und machen eine Fallunterscheidung:

(i) Ist ρ_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 ρ_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 ρ_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 (10.3)
4 Berechne ρ_k nach (10.4)
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** $\eta_1 \leqslant \rho_k < \eta_2$ **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** $\rho_k \geqslant \eta_2$ **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 (10.3). Schiefgehen kann also nur Schritt 4, und zwar wenn der Nenner von (10.4) 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* (10.3). *Dann gilt*

$$f(x^k) - q_k(s^k) \geqslant \frac{1}{2}\|\nabla f(x^k)\| \min\left\{\Delta_k, \tfrac{\|\nabla f(x^k)\|}{\|\nabla^2 f(x^k)\|}\right\}.$$

Beweis. Da s^k als globaler Minimierer von q_k über K_{Δ_k} gewählt ist, gilt für alle $d \in K_{\Delta_k}$

$$\begin{aligned} f(x^k) - q_k(s^k) &\geqslant f(x^k) - q_k(d) \\ &= -\nabla f(x^k)^T d - \frac{1}{2} d^T \nabla^2 f(x^k) d \\ &\geqslant -\nabla f(x^k)^T d - \frac{1}{2}\|\nabla^2 f(x^k)\|\,\|d\|^2. \end{aligned}$$

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

$$f(x^k) - q_k(x^k) \geqslant \frac{\|\nabla f(x^k)\|^2}{\|\nabla^2 f(x^k)\|} - \frac{1}{2}\frac{\|\nabla f(x^k)\|^2}{\|\nabla^2 f(x^k)\|} = \frac{1}{2}\frac{\|\nabla f(x^k)\|^2}{\|\nabla^2 f(x^k)\|}.$$

(ii) $\|\nabla f(x^k)\| \geqslant \Delta_k \|\nabla^2 f(x^k)\|$: 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

$$f(x^k) - q_k(x^k) \geqslant \Delta_k \|\nabla f(x^k)\| - \frac{1}{2}\Delta_k^2 \|\nabla^2 f(x^k)\| \geqslant \frac{1}{2}\Delta_k \|\nabla f(x^k)\|.$$

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 $\bar{x} \in \mathbb{R}^n$ ein stationärer Punkt. 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 $k \geqslant k_0$ verworfen werden. Dies setzt aber voraus, dass gilt

$$\rho_k < \eta_1 < 1 \qquad \text{für alle } k \in K, k \geqslant k_0. \tag{10.5}$$

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

$$\|\nabla f(x^k)\| \geqslant \beta_1 \qquad \text{für alle } k \in K.$$

Weiter ist $\nabla^2 f$ stetig und $\{x^k\}_{k\in K}$ konvergent und daher beschränkt, es gibt also ein $\beta_2 > 0$ mit

$$\|\nabla^2 f(x^k)\| \leqslant \beta_2 \qquad \text{für alle } k \in K.$$

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

$$\begin{aligned} |q_k(s^k) - f(x^k + s^k)| &= \left|\nabla f(x^k)^T s^k + \tfrac{1}{2}(s^k)^T \nabla^2 f(x^k) s^k - \nabla f(\xi^k)^T s^k\right| \\ &\leqslant \|\nabla f(x^k) - \nabla f(\xi^k)\| \|s^k\| + \frac{\beta_2}{2}\|s^k\|^2. \end{aligned}$$

Nach Lemma 10.1 gilt nun wegen $\|s^k\| \leqslant \Delta_k \to 0$ für alle hinreichend großen $k \in K$, dass

$$\begin{aligned} f(x^k) - q_k(s^k) &\geqslant \frac{1}{2}\|\nabla f(x^k)\| \min\left\{\Delta_k, \tfrac{\|\nabla f(x^k)\|}{\|\nabla^2 f(x^k)\|}\right\} \\ &\geqslant \frac{1}{2}\beta_1 \min\left\{\Delta_k, \tfrac{\beta_1}{\beta_2}\right\} \\ &= \frac{1}{2}\beta_1 \Delta_k \geqslant \frac{1}{2}\beta_1 \|s^k\|. \end{aligned}$$

Zusammen erhalten wir also

$$|\rho_k - 1| \leqslant \frac{1}{\beta_1}\left(2\|\nabla f(x^k) - \nabla f(\xi^k)\| + \beta_2 \|s^k\|\right).$$

Der zweite Term in Klammern konvergiert wegen $\|s^k\| \leqslant \Delta_k \to 0$, der erste wegen der Stetigkeit von ∇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 (10.5). □

Mit Hilfe von Lemma 10.1 und Lemma 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 $k \geqslant k_0$ der Schritt verworfen wird. Dann wird aber auch für alle $k \geqslant k_0$ 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 ∇f und $\nabla^2 f$ existieren nun wieder Konstanten $\beta_1, \beta_2 > 0$ mit

$$\|\nabla f(x^k)\| \geqslant \beta_1, \quad \text{und} \quad \|\nabla^2 f(x^k)\| \leqslant \beta_2 \qquad \text{für alle } k \in K.$$

Da alle Schritte erfolgreich sind, muss außerdem $\rho_k \geqslant \eta_1$ für alle $k \in K$ gelten. Mit Lemma 10.1 folgt daher für alle $k \in K$

$$
\begin{aligned}
f(x^k) - f(x^k + s^k) &\geqslant \eta_1 (f(x^k) - q_k(s^k)) \\
&\geqslant \eta_1 \frac{1}{2} \|\nabla f(x^k)\| \min\left\{\Delta_k, \tfrac{\|\nabla f(x^k)\|}{\|\nabla^2 f(x^k)\|}\right\} \\
&\geqslant \frac{\eta_1 \beta_1}{2} \min\left\{\Delta_k, \tfrac{\beta_1}{\beta_2}\right\}.
\end{aligned} \tag{10.6}
$$

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 (10.6) 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

$$
(s^k)^T \nabla^2 f(x^k) s^k \geqslant \mu \|s^k\| \qquad \text{für alle } k \in K, k \geqslant k_0.
$$

Weiter folgt aus Lemma 10.1 insbesondere $f(x^k) - q_k(s^k) \geqslant 0$ und daher

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

Zusammen erhalten wir für alle $k \in K$ mit $k \geqslant k_0$

$$
\frac{\mu}{2}\|s^k\|^2 \leqslant \frac{1}{2}(s^k)^T \nabla^2 f(x^k) s^k \leqslant -\nabla f(x^k)^T s^k \leqslant \|\nabla f(x^k)\| \|s^k\|.
$$

Da aber die Teilfolge $\{x^k\}_{k\in K}$ nach Satz 10.3 gegen den stationären Punkt $\bar{x}$ konvergiert und ∇f stetig ist, folgt daraus

$$\|x^{k+1} - x^k\| = \|s^k\| \leqslant \frac{2}{\mu}\|\nabla f(x^k)\| \to 0.$$

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 $k \geqslant k_0$ 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

$$\|s^k\| \leqslant \frac{2}{\mu}\|\nabla f(x^k)\| \qquad \text{für alle } k \geqslant k_0. \tag{10.7}$$

Wegen der Stetigkeit von $\nabla^2 f$ existiert weiter eine Konstante $c > 0$ mit

$$\|\nabla^2 f(x^k)\| \leqslant c \qquad \text{für alle } k \geqslant k_0.$$

Aus Lemma 10.1 zusammen mit $\|s^k\| \leqslant \Delta_k$ folgt nun

$$\begin{aligned} f(x^k) - q_k(s^k) &\geqslant \frac{1}{2}\|\nabla f(x^k)\| \min\left\{\Delta_k, \tfrac{\|\nabla f(x^k)\|}{\|\nabla^2 f(x^k)\|}\right\} \\ &\geqslant \frac{\mu}{4}\|s^k\| \min\left\{\|s^k\|, \tfrac{\mu}{2c}\|s^k\|\right\} \\ &= \kappa\|s^k\|^2 \end{aligned}$$

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

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

Zusammen erhalten wir

$$|\rho_k - 1| = \frac{|q_k(s^k) - f(x^k + s^k)|}{|f(x^k) - q_k(s^k)|} \leqslant \frac{1}{2\kappa}\|\nabla^2 f(\xi^k) - \nabla^2 f(x^k)\| \to 0,$$

da wegen $x^k \to \bar{x}$ mit $\nabla f(\bar{x}) = 0$ und (10.7) auch $\xi^k \to \bar{x}$ konvergiert. Also ist wegen $\eta_1 < 1$ für $k \in \mathbb{N}$ hinreichend groß stets $\rho_k \geqslant \eta_1$, 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 $k \geqslant k_0$. Also ist das quadratische Modell q_k für alle $k \geqslant k_0$ strikt konvex, und die nach Satz 3.5 notwendige und hinreichende Optimalitätsbedingung $\nabla q_k(\bar{s}) = 0$ ist identisch mit dem Newton-Schritt

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

Nach Lemma 7.5 existiert weiterhin ein $k_1 \in \mathbb{N}$ und ein $c > 0$ mit

$$\|\nabla^2 f(x^k)^{-1}\| \leqslant c \qquad \text{für alle } k \geqslant k_1.$$

Da x^k gegen den stationären Punkt $\bar{x}$ konvergiert, folgt daraus für $k \geqslant \max\{k_0, k_1\}$

$$\|\bar{s}^k\| \leqslant c\|\nabla f(x^k)\| \to 0.$$

Also muss ein $k_2 \in \mathbb{N}$ existieren mit $\|\bar{s}^k\| \leqslant \Delta_{\min}$ für alle $k \geqslant k_2$.

Andererseits sind nach Lemma 10.5 für ein $k_3 \in \mathbb{N}$ alle Schritte $k \geqslant k_3$ erfolgreich; aufgrund der Iterationsvorschrift gilt daher $\Delta_k \geqslant \Delta_{k_3} \geqslant \Delta_{\min} > 0$ für alle $k \geqslant k_3$. Also stimmt für alle $k \geqslant \max\{k_2, k_3\}$ der Newton-Schritt $\bar{s}^k \in K_{\Delta_k}$ mit der Lösung s^k von (10.2) überein. Damit ist für alle $k \geqslant \max\{k_2, k_3\}$ der Algorithmus 10.1 identisch mit dem Newton-Verfahren und hat daher die selbe Konvergenzgeschwindigkeit. □

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_{Δ_k} berechnen kann. Prinzipiell handelt es sich dabei um ein beschränktes Optimierungsproblem, das mit den in den folgenden Kapiteln vorgestellten Verfahren behandelt werden kann (wobei die quadratische Struktur von q_k und die Kugelgestalt von K_{Δ_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 σ_k Lösung ist des eingeschränkten Problems

$$\min_{\sigma \geqslant 0} q_k(-\sigma \nabla f(x^k)) \qquad \text{mit } \|\sigma \nabla f(x^k)\| \leqslant \Delta_k.$$

Da q_k quadratisch ist, ist die globale Lösung dieses Problems gegeben durch

$$\sigma^k = \begin{cases} \frac{\Delta_k}{\|\nabla f(x^k)\|} & \text{falls } \nabla f(x^k)^T \nabla^2 f(x^k) \nabla f(x^k) \leqslant 0, \\ \min\left\{ \frac{\|\nabla f(x^k)\|^2}{\nabla f(x^k)^T \nabla^2 f(x^k) \nabla f(x^k)}, \frac{\Delta_k}{\|\nabla f(x^k)\|} \right\} & \text{sonst,} \end{cases} \tag{10.8}$$

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 σ_k gegeben durch (10.8). Dann gilt*

$$f(x^k) - q_k(s^k) \geqslant \frac{1}{2} \|\nabla f(x^k)\| \min\left\{ \Delta_k, \frac{\|\nabla f(x^k)\|}{\|\nabla^2 f(x^k)\|} \right\}.$$

Daraus folgt wie zuvor die globale Konvergenz von Algorithmus 10.1, wenn für x^{k+1} der Cauchy-Punkt anstelle von $x^k + s^k$ gewählt wird.

Satz 10.8. *Sei $f : \mathbb{R}^n \to \mathbb{R}$ zweimal stetig differenzierbar. Dann bricht Algorithmus 10.1 mit x_C anstelle von $x^k + s^k$ 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

$$d(\tau) := \begin{cases} \tau d_G & \text{für } \tau \in [0,1], \\ d_G + (\tau - 1)(d_N - d_G) & \text{für } \tau \in [1,2], \end{cases}$$

unter der Beschränkung $\|d(\tau)\| \leqslant \Delta_k$. Der Pfad $d(\tau)$, $\tau \in [0,2]$, 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 $\varphi'(t) \geqslant 0$ für $t \in (0,1)$. Ausmultiplizieren und Ableiten ergibt

$$\begin{aligned} \varphi'(t) &= d_G^T(d_N - d_G) + t\|d_N - d_G\|^2 \geqslant d_G^T(d_N - d_G) \\ &= -\frac{\|g\|^2}{g^T H g}\left(-g^T H^{-1} g + \frac{\|g\|^4}{g^T H g}\right) = \|g\|^2 \frac{g^T H^{-1} g}{g^T H g}\left(1 - \frac{\|g\|^4}{(g^T H^{-1} g)(g^T H g)}\right). \end{aligned}$$

Da H und damit auch H^{-1} positiv definit sind, ist der Term vor der Klammer positiv. Mit Hilfe der Cholesky-Zerlegung $H = R^T R$ mit R invertierbar erhalten wir auch

$$\begin{aligned}\|g\|^2 &= g^T g = g^T R^{-1} R g = (R^{-T} g)^T (Rg) \\ &\leqslant \|R^{-T} g\| \|Rg\| = \left((R^{-T} g)^T (R^{-T} g)\right)^{1/2} \left((Rg)^T (Rg)\right)^{1/2} \\ &= \left(g^T (R^{-1})(R^{-T}) g\right)^{1/2} \left(g^T R^T R g\right)^{1/2} = (g^T H^{-1} g)^{1/2} (g^T H g)^{1/2}.\end{aligned}$$

Also ist

$$\frac{\|g\|^4}{(g^T H^{-1} g)(g^T H g)} \leqslant 1$$

und damit $\varphi'(t) \geqslant 0$.

Zu (ii): Wir betrachten analog für $t \in [0, 1]$ die Funktion

$$\begin{aligned}\psi(t) &= q_k(d(1+t)) \\ &= f(x^k) + g^T (d_G + t(d_N - d_G)) + \frac{1}{2}(d_G + t(d_N - d_G))^T H (d_G + t(d_N - d_G))\end{aligned}$$

und zeigen $\psi'(t) \leqslant 0$ für $t \in (0, 1)$. Ausmultiplizieren und Ableiten ergibt wegen $t \leqslant 1$ und der positiven Definitheit von H

$$\begin{aligned}\psi'(t) &= g^T (d_N - d_G) + d_G^T H (d_N - d_G) + t(d_N - d_G)^T H (d_N - d_G) \\ &\leqslant g^T (d_N - d_G) + d_G^T H (d_N - d_G) + (d_N - d_G)^T H (d_N - d_G) \\ &= (g + H d_N)^T (d_N - d_G) \\ &= 0,\end{aligned}$$

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.

Im zweiten Fall unterscheiden wir noch weiter:

(iia) $\tau^* \leqslant 1$: 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

$$\begin{aligned}r(t) &= \|d_G + t(d_N - d_G)\|^2 - \Delta^2 \\ &= \|d_N - d_G\|^2 t^2 + 2 d_G^T (d_N - d_G) t + \|d_G\|^2 - \Delta^2.\end{aligned}$$

Wir können also den *Dogleg-Schritt* wie folgt bestimmen.

Algorithmus 10.2 : Dogleg-Schritt

if $\|d_G\| \geqslant \Delta_k$ **then** // Gradientenschritt nicht in Vertrauensbereich
  Setze $s^k = \frac{\Delta_k}{\|d_G\|} d_G$ // Akzeptiere Cauchy-Punkt
else if $\|d_N\| \leqslant \Delta_k$ **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 $\nabla f(x^k)^T \nabla^2 f(x^k) \nabla f(x^k) \leqslant 0$ statt dem Dogleg-Schritt der Cauchy-Punkt verwendet wird. Auch in diesem Fall kann man globale Konvergenz und lokal superlineare Konvergenz zeigen; siehe [Kelley 1999, Abschnitt 3.3.6].

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 $\|\nabla^2 f(x^k)s^m + \nabla f(x^k)\| \leqslant \eta_k \|\nabla f(x^k)\|$ für eine vorgegebene Toleranz $\eta_k > 0$, so setze $s^k := s^m$ und beende die Iteration.

(ii) Ist $\|s^m\| \geqslant \Delta_k$ (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 und Kanzow 1999, Kapitel 14.7] verwiesen.

Teil III

OPTIMIERUNG MIT NEBENBEDINGUNGEN

OPTIMALITÄTSBEDINGUNGEN

11

Wir betrachten nun für $X \subsetneq \mathbb{R}^n$ und $f : X \to \mathbb{R}$ das *restringierte* Optimierungsproblem

$$\min_{x \in X} f(x)$$

und leiten dafür zunächst Optimalitätsbedingungen her. Wie schon für den Fall $n = 1$ bekannt, ist das Verschwinden des Gradienten *keine* notwendige Optimalitätsbedingung für einen lokalen Minimierer, wenn dieser auf dem Rand der zulässigen Menge liegt. Für (vernünftige) $X \subset \mathbb{R}$ ist dies durch einfaches Nachprüfen endlich vieler Randpunkte noch leicht zu handhaben. Ist $n > 1$, aber $X \subset \mathbb{R}^n$ ein Polyeder (d. h. durch endlich viele *lineare* Gleichungs- und Ungleichungsnebenbedingungen beschrieben), so hat der Rand von X ebenfalls eine spezielle Struktur (was in der linearen Optimierung weidlich ausgenutzt wird). Für allgemeine Teilmengen $X \subsetneq \mathbb{R}^n$ kann der Rand aber beliebig kompliziert sein, was die Schwierigkeit der restringierten Optimierung ausmacht.

TANGENTIALKEGEL

Wir orientieren uns an Satz 3.1, der besagt, dass für unrestringierte Probleme in einem lokalen Minimierer $\bar{x}$ alle Richtungen Abstiegsrichtungen sein müssen, d. h. $\nabla f(\bar{x})^T d \geqslant 0$ für alle $d \in \mathbb{R}^n$ gilt. Für ein restringiertes Problem spielen dagegen nur die Richtungen eine Rolle, die nicht sofort(!) aus X hinausführen; für solch eine Richtung $d \in \mathbb{R}^n$ muss also $x := \bar{x} + td \in X$ für alle $t > 0$ klein genug gelten. Dies motiviert die folgende Definition: Wir nennen $d \in \mathbb{R}^n$ *Tangentialrichtung* an X in x, falls Folgen $\{x^k\}_{k\in\mathbb{N}} \subset X$ und $\{t_k\}_{k\in\mathbb{N}} \subset (0, \infty)$ existieren mit

$$x^k \to x, \qquad t_k \to 0, \qquad \frac{x^k - x}{t_k} \to d.$$

Die Menge aller Tangentialrichtungen bezeichnen wir als *Tangentialkegel*

$$T_X(x) := \{d \in \mathbb{R}^n : d \text{ ist Tangentialrichtung an } X \text{ in } x\}.$$

Ist x ein innerer Punkt von X, so gilt $T_X(x) = \mathbb{R}^n$ (da wir in dem Fall $\{x^k\}_{k\in\mathbb{N}} \subset B_\varepsilon(x) \subset X$ beliebig wählen können). Weiter gilt stets $0 \in T_X(x)$ (wähle $x^k := x$) sowie für $d \in T_X(x)$ und $\alpha > 0$ auch $\tilde{d} := \alpha d \in T_X(x)$ (wähle $\tilde{t}^k := \alpha^{-1} t_k$), was die Bezeichnung *Kegel* rechtfertigt.

Dass wir bei der Konstruktion von Tangentialrichtungen Folgen von Punkten in X und nicht nur feste Punkte zulassen, liegt daran, dass der so definierte Tangentialkegel stets abgeschlossen ist. Dies wird später von Bedeutung sein.

Lemma 11.1. *Seien $X \subset \mathbb{R}^n$ nichtleer und $x \in X$. Dann ist $T_X(x)$ abgeschlossen.*

Beweis. Wir müssen zeigen, dass für $\{d^k\}_{k\in\mathbb{N}} \subset T_X(x)$ mit $d^k \to d$ auch $d \in T_X(x)$ gilt. Für jede Tangentialrichtung d^k existieren gemäß Definition Folgen $\{x^{k,l}\}_{l\in\mathbb{N}} \subset X$ und $\{t_{k,l}\}_{l\in\mathbb{N}} \subset (0,\infty)$ so, dass für alle $k \in \mathbb{N}$ ein $l(k) \in \mathbb{N}$ existiert mit

$$\|x^{k,l(k)} - x\| \leqslant \frac{1}{k}, \qquad t_{k,l(k)} \leqslant \frac{1}{k}, \qquad \left\| \frac{x^{k,l(k)} - x}{t_{k,l(k)}} - d^k \right\| \leqslant \frac{1}{k}.$$

Für die entsprechenden Diagonalfolgen $\{x^{k,l(k)}\}_{k\in\mathbb{N}} \subset X$ und $\{t_{k,l(k)}\}_{k\in\mathbb{N}} \subset (0,\infty)$ gilt daher wegen $d^k \to d$

$$\left\| \frac{x^{k,l(k)} - x}{t_{k,l(k)}} - d \right\| \leqslant \left\| \frac{x^{k,l(k)} - x}{t_{k,l(k)}} - d^k \right\| + \|d^k - d\| \to 0$$

für $k \to \infty$. Also ist auch d eine Tangentialrichtung. □

Analog zu Satz 3.1 erhalten wir eine abstrakte notwendige Optimalitätsbedingung erster Ordnung.

Satz 11.2. *Seien $X \subset \mathbb{R}^n$ eine nichtleere Menge und $f : X \to \mathbb{R}$ stetig differenzierbar. Hat f in $\bar{x} \in X$ ein lokales Minimum, so gilt*

$$\nabla f(\bar{x})^T d \geqslant 0 \qquad \textit{für alle } d \in T_X(\bar{x}). \tag{11.1}$$

Beweis. Sei $d \in T_X(\bar{x})$ beliebig. Dann existieren nach Definition Folgen $\{x^k\}_{k\in\mathbb{N}} \subset X$ und $\{t_k\}_{k\in\mathbb{N}} \subset (0,\infty)$ mit $x^k \to \bar{x}$ und $\frac{x^k - \bar{x}}{t_k} \to d$. Weiter existiert nach Satz 1.1 ein $\xi^k = \bar{x} + \theta_k(x^k - \bar{x})$ mit $\theta_k \in (0,1)$ und

$$\nabla f(\xi^k)^T (x^k - \bar{x}) = f(x^k) - f(\bar{x}) \geqslant 0$$

für $k \in \mathbb{N}$ hinreichend groß (denn $\bar{x}$ ist lokaler Minimierer und $x^k \in X$). Da mit $x^k \to \bar{x}$ auch $\xi^k \to \bar{x}$ konvergiert, können wir durch $t_k > 0$ dividieren und erhalten wegen der Stetigkeit von ∇f durch Grenzübergang

$$\nabla f(\bar{x})^T d = \lim_{k\to\infty} \nabla f(\xi^k)^T \left(\frac{x^k - \bar{x}}{t_k} \right) \geqslant 0,$$

was zu beweisen war. □

Ist $K \subset \mathbb{R}^n$ ein nichtleerer Kegel, so bezeichnet man die Menge

$$K^o := \{x \in \mathbb{R}^n : x^T d \leqslant 0 \text{ für alle } d \in K\}$$

als *Polarkegel* von K. Die notwendige Optimalitätsbedingung (11.1) kann man damit kompakt schreiben als

$$-\nabla f(\bar{x}) \in T_X(\bar{x})^o. \tag{11.2}$$

REGULARITÄTSBEDINGUNGEN

Der Rest des Kapitels ist nun der Aufgabe gewidmet, konkrete Darstellung sowohl des Tangentialkegels $T_X(x)$ als auch der notwendigen Optimalitätsbedingung (11.1) herzuleiten für Mengen der speziellen Form

$$X := \{x \in \mathbb{R}^n : g_i(x) \leqslant 0,\ 1 \leqslant i \leqslant m, \quad h_j(x) = 0,\ 1 \leqslant j \leqslant p\} \tag{11.3}$$

für $g_i, h_j : \mathbb{R}^n \to \mathbb{R}$ stetig differenzierbar.

UNGLEICHUNGSNEBENBEDINGUNGEN

Wir betrachten zuerst den Fall von reinen Ungleichungsnebenbedingungen, d. h.

$$X = \{x \in \mathbb{R}^n : g_i(x) \leqslant 0, \quad 1 \leqslant i \leqslant m\}$$

für $g_i : \mathbb{R}^n \to \mathbb{R}$ stetig differenzierbar. Da wir Abstiegsrichtungen durch die Ableitung (und damit Linearisierung) der Zielfunktion charakterisieren können, ist es naheliegend zu versuchen, Tangentialrichtungen über die Ableitung der Nebenbedingungen zu charakterisieren. Weiterhin ist zu erwarten, dass bei der Charakterisierung eines lokalen(!) Minimierers $\bar{x}$ nur die *aktiven* Nebenbedingungen mit $g_i(\bar{x}) = 0$ eine Rolle spielen. Tatsächlich können wir folgendes zeigen.

Lemma 11.3. *Für* $x \in X$ *und* $d \in T_X(x)$ *gilt*

$$\nabla g_i(x)^T d \leqslant 0 \qquad \textit{für alle } i \textit{ mit } g_i(x) = 0.$$

Beweis. Für $d \in T_X(x)$ existieren nach Definition Folgen $\{x^k\}_{k\in\mathbb{N}} \subset X$ und $\{t_k\}_{k\in\mathbb{N}} \subset (0, \infty)$ mit $x^k \to \bar{x}$ und $\frac{x^k - \bar{x}}{t_k} \to d$. Sei nun $i \in \{1, \dots, m\}$ mit $g_i(x) = 0$ beliebig. Dann folgt für alle $k \in \mathbb{N}$ wegen $x^k \in X$ und Satz 1.1

$$0 \geqslant g_i(x^k) - g_i(x) = \nabla g_i(\xi_k)^T (x^k - x)$$

für $\xi_k := x^k + \theta_k(x - x^k)$ mit $\theta_k \in (0, 1)$. Wieder folgt aus $x^k \to x$ auch $\xi^k \to x$. Division durch $t_k > 0$ und Grenzübergang $k \to \infty$ ergibt dann wegen der stetigen Differenzierbarkeit von g_i die Behauptung. □

Definieren wir die *Menge der aktiven Nebenbedingungen*

$$\mathcal{A}_X(x) := \{i \in \{1, \dots, m\} : g_i(x) = 0\}$$

und den *Linearisierungskegel*

$$L_X(x) := \left\{ d \in \mathbb{R}^n : \nabla g_i(x)^T d \leqslant 0 \text{ für alle } i \in \mathcal{A}_X(x) \right\},$$

so haben wir gerade gezeigt, dass $T_X(x) \subset L_X(x)$ gilt. Für Satz 11.2 ist das aber die falsche Richtung: Da $L_X(x)$ *größer* ist als $T_X(x)$, ist die Bedingung $\nabla f(x)^T d \geqslant 0$ für alle $d \in L_X(x)$ *stärker* als (11.1) und daher nicht unbedingt für jeden lokalen Minimierer erfüllt – es ist also keine notwendige Optimalitätsbedingung mehr (und eine hinreichende sowieso nicht). Beachte auch, dass der Tangentialkegel nur von der Menge X abhängt und damit unabhängig ist von ihrer Beschreibung durch konkrete Ungleichungen $g_i(x) \leqslant 0$, der Linearisierungskegel dagegen sehr wohl von der Wahl der g_i abhängt.

Leider gilt die umgekehrte Inklusion $L_X(x) \subset T_X(x)$ im Allgemeinen nicht. Als Beispiel betrachten wir die Menge

$$X := \left\{ x \in \mathbb{R}^2 : g_1(x) = x_2 - x_1^3 \leqslant 0,\ g_2(x) = -x_2 \leqslant 0 \right\}$$

sowie den zulässigen Punkt $x = (0,0)^T$, in dem beide Nebenbedingungen aktiv sind. Der zugehörige Linearisierungskegel ist

$$L_X(0) = \left\{ d \in \mathbb{R}^2 : d_2 \leqslant 0,\ -d_2 \leqslant 0 \right\} = \left\{ d \in \mathbb{R}^2 : d_2 = 0 \right\}.$$

Der Tangentialkegel ist nach Lemma 11.3 eine Teilmenge dieses Kegels. Allerdings gilt für alle $x \in X$ stets $x_1^3 \geqslant x_2 \geqslant 0$ und damit auch $x_1 \geqslant 0$. Also gilt nach Definition von Tangentialrichtungen in $x = 0$ auch $d_1 = \lim_{k\to\infty} x_1^k / t_k \geqslant 0$. Der Tangentialkegel ist daher die echte Teilmenge

$$T_X(0) = \left\{ d \in \mathbb{R}^2 : d_1 \geqslant 0,\ d_2 = 0 \right\}.$$

Das Problem ist hier, dass die Nebenbedingungen in $x = (0,0)^T$ lokal nicht von der Bedingung $x_2 = 0$ unterscheidbar sind (X ist dort zu "spitz"); um durch Linearisierung genau den Tangentialkegel zu bekommen, brauchen wir also mehr "Luft" in X.

Satz 11.4. *Sei $x \in X$. Existiert ein $v \in \mathbb{R}^n$ mit*

$$\nabla g_i(x)^T v < 0 \qquad \textit{für alle } i \in \mathcal{A}_X(x), \tag{11.4}$$

so ist $T_X(x) = L_X(x)$.

Beweis. Nach Lemma 11.3 ist nur die Inklusion $L_X(x) \subset T_X(x)$ zu zeigen. Sei dazu $d \in L_X(x)$ beliebig. Wir konstruieren nun geeignete Folgen $\{x^k\}_{k\in\mathbb{N}} \subset X$ und $\{t_k\}_{k\in\mathbb{N}} \subset (0,\infty)$. Dafür betrachten wir für festes $\alpha > 0$ und $t > 0$ beliebig den Vektor

$$x_t := x + t(d + \alpha v).$$

Wir zeigen zuerst, dass für t hinreichend klein $x_t \in X$ gilt, d. h. $g_i(x_t) \leqslant 0$ für alle $1 \leqslant i \leqslant m$ gilt. Dafür machen wir eine Fallunterscheidung:

(i) $i \in \mathcal{A}_X(x)$: Wegen $d \in L_X(x)$ gilt zunächst $\nabla g_i(x)^T d \leqslant 0$ und daher

$$\lim_{t \to 0^+} \frac{g_i(x_t) - g_i(x)}{t} = \nabla g_i(x)^T (d + \alpha v) = \nabla g_i(x)^T d + \alpha \nabla g_i(x)^T v < 0.$$

Da der Grenzwert strikt negativ ist, muss wegen $g_i(x) = 0$ für $i \in \mathcal{A}_X(x)$ auch gelten

$$g_i(x_t) = g_i(x_t) - g_i(x) < 0$$

für $t > 0$ klein genug.

(ii) $i \notin \mathcal{A}_X(x)$: Dann ist $g_i(x) < 0$, und wegen der Stetigkeit von g_i und $x_t \to x$ für $t \to 0$ gilt auch $g_i(x_t) < 0$ für $t > 0$ klein genug.

Für alle $1 \leqslant i \leqslant m$ existiert also ein $t_i > 0$ mit $g_i(x_t) \leqslant 0$ für alle $t < t_i$. Also existiert ein $s := \min\{t_i : 1 \leqslant i \leqslant m\} > 0$ mit $x_t \in X$ für alle $t < s$.

Wir setzen nun $t_k := \frac{1}{k}$ und $x^k := x_{t_k}$. Für alle $k \in \mathbb{N}$ groß genug ist dann $x^k \in X$; außerdem gilt sowohl $x^k \to x$ für $k \to \infty$ als auch

$$\frac{x^k - x}{t_k} = d + \alpha v.$$

Also ist nach Definition $d + \alpha v \in T_X(x)$ für alle $\alpha > 0$. Da nach Lemma 11.1 Tangentialkegel stets abgeschlossen sind, folgt durch Grenzübergang $\alpha \to 0$ auch $d \in T_X(x)$. □

Ein Punkt $x \in X$, für den $T_X(x) = L_X(x)$ gilt, heißt *regulär*; eine Bedingung wie (11.4), die die Regularität eines Punktes garantiert, nennt man *Regularitätsbedingung* oder (auch im Deutschen) *constraint qualification*. (Die "triviale" Regularitätsbedingung $T_X(x) = L_X(x)$ wird auch als *Adabie constraint qualification* bezeichnet.)

Eine stärkere, aber leichter überprüfbare, Bedingung ist die sogenannte *linear independence constraint qualification* (LICQ).

Folgerung 11.5. *Sei $x \in X$. Ist die Menge $\{\nabla g_i(x) : i \in \mathcal{A}_X(x)\} \subset \mathbb{R}^n$ linear unabhängig, so ist x regulär.*

Beweis. Unter dieser Voraussetzung hat das lineare Gleichungssystem

$$\nabla g_i(x)^T d = b_i, \qquad i \in \mathcal{A}_X(x),$$

vollen (Zeilen-)Rang und damit für beliebige $b_i \in \mathbb{R}$ eine (nicht unbedingt eindeutige) Lösung. Wählt man $b_i < 0$ für alle $i \in \mathcal{A}_X(x)$, ist somit (11.4) erfüllt und die Aussage folgt aus Satz 11.4. □

Wieder sind für konvexe Funktionen stärkere Aussagen möglich.

Folgerung 11.6. *Sei* g_i *konvex für alle* $1 \leqslant i \leqslant m$. *Existiert ein* $\tilde{x} \in X$ *mit*

$$g_i(\tilde{x}) < 0 \qquad \text{für alle } 1 \leqslant i \leqslant m, \tag{11.5}$$

so ist jeder *Punkt* $x \in X$ *regulär.*

Beweis. Für $x = \tilde{x}$ ist wegen (11.5) die aktive Menge $\mathcal{A}_X(x)$ leer und damit (11.4) trivialerweise erfüllt. Sei daher $x \neq \tilde{x}$. Aus der Konvexität folgt mit Satz 1.4 (i) für alle $i \in \mathcal{A}_X(x)$

$$\nabla g_i(x)^T(\tilde{x} - x) \leqslant g_i(\tilde{x}) - g_i(x) = g_i(\tilde{x}) < 0$$

und daraus (11.4) mit $v := \tilde{x} - x$. □

Die Bedingung (11.5) wird *Slater-Bedingung* genannt.

Für lineare Nebenbedingungen könnte man erwarten, dass automatisch $T_X(x) = L_X(x)$ gilt, und in der Tat stellt die Linearität an sich eine Reguläritsbedingung dar.

Satz 11.7. *Seien alle* g_i *affin-linear für alle* $1 \leqslant i \leqslant m$, *d. h. es existieren* $a_i \in \mathbb{R}^n$ *und* $\alpha_i \in \mathbb{R}$ *mit* $g_i(x) = a_i^T x - \alpha_i$. *Dann ist jeder Punkt* $x \in X$ *regulär.*

Beweis. Der Beweis ist ein stark vereinfachter Spezialfall von Satz 11.4. Für $x \in X$ beliebig und $d \in L_X(x)$ setze $t_k := \frac{1}{k}$ und $x^k := x + t_k d$. Wieder machen wir die Fallunterscheidung

(i) $i \in \mathcal{A}_X(x)$: Dann gilt $a_i^T x = \alpha_i$ und zusammen mit $a_i^T d \leqslant 0$ wegen $d \in L_X(x)$ folgt

$$a_i^T x^k = a_i^T x + t_k a_i^T d \leqslant \alpha_i.$$

(ii) $i \notin \mathcal{A}_X(x)$: Dann gilt $a_i^T x < \alpha_i$ und damit auch

$$a_i^T x^k = a_i^T x + t_k a_i^T d < \alpha_i$$

für t_k hinreichend klein.

Also ist $x^k \in X$ für $k \in \mathbb{N}$ groß genug sowie

$$\frac{x^k - x}{t_k} = d,$$

d. h. $d \in T_X(x)$. □

Auch Kombinationen dieser Regularitätsbedingungen sind möglich – es reicht zum Beispiel, dass nur diejenigen g_i, die nicht affin-linear sind, die Slater-Bedingung erfüllen.

GLEICHUNGSNEBENBEDINGUNGEN

Wir betrachten nun denn Fall von reinen Gleichungsnebenbedingungen, d. h.

$$X = \{x \in \mathbb{R}^n : h_i(x) = 0, \quad 1 \leqslant i \leqslant p\}$$

für $h_i : \mathbb{R}^n \to \mathbb{R}$ stetig differenzierbar. Da Gleichungsnebenbedingungen stets aktiv sind, ist der entsprechende Linearisierungskegel für $x \in X$ definiert durch

$$L_X(x) = \left\{d \in \mathbb{R}^n : \nabla h_i(x)^T d = 0,\ 1 \leqslant i \leqslant p\right\}.$$

Völlig analog zu Lemma 11.3 (nur mit Gleichheit an Stelle der Ungleichung) beweist man nun die folgende Inklusion.

Lemma 11.8. *Für alle $x \in X$ gilt $T_X(x) \subset L_X(x)$.*

Für die umgekehrte Inklusion benötigt man wieder eine Regularitätsbedingung.

Satz 11.9. *Sei $x \in X$. Ist die Menge $\{\nabla h_i(x) : 1 \leqslant i \leqslant p\} \subset \mathbb{R}^n$ linear unabhängig, so ist x regulär.*

Beweis. Der Beweis folgt im Prinzip dem von Satz 11.4; die Schwierigkeit besteht dabei darin, dass wir mit unserer "Tangentenfolge" $\{x^k\}_{k\in\mathbb{N}}$ den (nichtlinearen) Gleichungen $h_i(x) = 0$ folgen müssen. Dazu addieren wir zu der üblichen "linearen" Konstruktion $x_t = x + td$ einen nichtlinearen (in t) Korrekturterm, den wir – unter der genannten Voraussetzung – mit Hilfe des Satzes über implizite Funktionen erhalten.

Sei dafür $d \in L_X(x)$ beliebig. Ziel ist zu zeigen, dass für $\varepsilon > 0$ klein genug eine Kurve $x : (-\varepsilon, \varepsilon) \to \mathbb{R}^n$ existiert mit $x(0) = x$, $x'(0) = d$ und $h_i(x(t)) = 0$ für alle $t \in (-\varepsilon, \varepsilon)$ und $1 \leqslant i \leqslant p$. Zu diesem Zweck definieren wir zunächst die Funktion

$$h : \mathbb{R}^n \to \mathbb{R}^p, \qquad x \mapsto (h_1(x), \ldots, h_p(x))^T$$

und konstruieren mit ihrer Hilfe eine Funktion $H : \mathbb{R}^{p+1} \to \mathbb{R}^p$ komponentenweise durch

$$H_j(y, t) := h_j(x + td + h'(x)^T y), \qquad 1 \leqslant i \leqslant p,$$

wobei $h'(x)$ die Jacobi-Matrix von h bezeichnet (deren Spalten genau die $\nabla h_i(x)$ sind). Wir betrachten nun das nichtlineare Gleichungssystem $H(y, t) = 0$, das wegen $x \in X$ offensichtlich die Lösung $(\bar{y}, \bar{t}) = (0, 0)$ besitzt. Auf diese Gleichung wenden wir nun den Satz über implizite Funktionen an, um eine Kurve $y(t)$ zu erhalten. Zunächst hat die Funktion $y \mapsto H(y, t)$ für $t > 0$ fest nach der Kettenregel die Jacobi-Matrix

$$H_y(0, 0) = h'(x)h'(x)^T \in \mathbb{R}^{p\times p}.$$

Nach Voraussetzung hat $h'(x)$ vollen Rang, und damit ist $H_y(0,0)$ invertierbar. Nach dem Satz über implizite Funktionen existiert daher ein $\varepsilon > 0$ und eine stetig differenzierbare Funktion $y : (-\varepsilon, \varepsilon) \to \mathbb{R}^p$ mit $y(0) = 0$ und $H(y(t), t) = 0$ für alle $t \in (-\varepsilon, \varepsilon)$ sowie

$$y'(t) = -H_y(y(t), t)^{-1} H_t(y(t), t) \qquad \text{für alle } t \in (-\varepsilon, \varepsilon).$$

Analog hat die Funktion $t \mapsto H(y, t)$ für $y \in \mathbb{R}^p$ fest die Ableitung $H_t(y, t) = h'(x + td + h'(x)^T y)d$, und damit folgt

$$y'(0) = -H_y(0,0)^{-1} h'(x)d = 0$$

wegen $\nabla h_i(x)^T d = 0$ für alle $1 \leqslant i \leqslant m$ nach Annahme an d.

Wir definieren nun die gesuchte Kurve durch

$$x(t) := x + td + h'(x)^T y(t).$$

Dann gilt nach Konstruktion von $y(t)$

$$h(x(t)) = H(y(t), t) = 0 \qquad \text{für alle } t \in (-\varepsilon, \varepsilon)$$

sowie $x(0) = 0$ und $x'(0) = d + h'(x)^T y'(0) = d$. Setzen wir also $t_k := \frac{1}{k}$ und $x^k := x(t_k)$, so gilt $x^k \in X$ für $k \in \mathbb{N}$ groß genug sowie

$$\lim_{k\to\infty} \frac{x^k - x}{t_k} = \lim_{k\to\infty} \frac{x(t_k) - x}{t_k} = x'(0) = d,$$

d. h. $d \in T_X(x)$. □

Wieder ist die Linearität an sich eine Regularitätsbedingung.

Satz 11.10. *Seien alle h_j affin-linear für alle $1 \leqslant i \leqslant p$, d. h. es existieren $b_j \in \mathbb{R}^n$ und $\beta_j \in \mathbb{R}$ mit $h_j(x) = b_j^T x - \beta_j$. Dann ist jeder Punkt $x \in X$ regulär.*

Beweis. Der Beweis ist völlig analog zu dem von Satz 11.7: Wir wählen wieder $t_k := \frac{1}{k}$ und $x^k := x + t_k d$. Für $x \in X$ und $d \in L_X(x)$ folgt dann mit $b_j^T x = h(x) + \beta_j = \beta_j$ und $b_j^T d = \nabla h_j(x)^T d = 0$ sofort

$$b_j^T x^k = b_j^T x + t_k b_j^T d = \beta_j.$$

Also ist $x^k \in X$ für alle $k \in \mathbb{N}$ sowie $\frac{x^k - x}{t_k} = d$, d. h. $d \in T_X(x)$. □

GEMISCHTE NEBENBEDINGUNGEN

Wir kommen nun zum allgemeinen Fall,

$$X = \{x \in \mathbb{R}^n : g_i(x) \leqslant 0,\ 1 \leqslant i \leqslant m, \quad h_j(x) = 0,\ 1 \leqslant j \leqslant p\}$$

für $g_i, h_j : \mathbb{R}^n \to \mathbb{R}$ stetig differenzierbar, für den der Linearisierungskegel gegeben ist durch

$$L_X(x) = \left\{d \in \mathbb{R}^n : \nabla g_i(x)^T d \leqslant 0,\ i \in \mathcal{A}_X(x), \quad \nabla h_j(x)^T d = 0,\ 1 \leqslant j \leqslant p\right\}.$$

Die "einfache" Inklusion beweist man wieder völlig analog zu Lemma 11.3, indem man die Nebenbedingungen einzeln betrachtet.

Lemma 11.11. *Für alle $x \in X$ gilt $T_X(x) \subset L_X(x)$.*

Für die andere Richtung kombiniert man nun die obigen Ansätze, wobei man nur darauf achten muss, dass sich Gleichungs- und Ungleichungsrestriktionen nicht in die Quere kommen. Die entsprechende Regularitätsbedingung nennt man *Mangasarian–Fromowitz constraint qualification* (MFCQ).

Satz 11.12. *Sei $x \in X$. Gilt*

(i) die Menge $\{\nabla h_j(x) : 1 \leqslant j \leqslant p\} \subset \mathbb{R}^n$ ist linear unabhängig,

(ii) es gibt ein $v \in \mathbb{R}^n$ mit

$$\begin{aligned} \nabla g_i(x)^T v &< 0 \qquad \text{für alle } i \in \mathcal{A}_X(x), \\ \nabla h_i(x)^T v &= 0 \qquad \text{für alle } i \in \{1, \dots, p\}, \end{aligned}$$

so ist x regulär.

Beweis. Sei $d \in L_X(x)$ beliebig. Dann gilt insbesondere $\nabla h_j(x)^T d = 0$ und damit nach Annahme (ii) auch $\nabla h_j(x)^T(d + \alpha v) = 0$ für alle $\alpha > 0$ und $1 \leqslant i \leqslant p$. Wie im Beweis von Satz 11.9 konstruiert man nun mit Hilfe von Annahme (i) für $\tilde{d} := d + \alpha v$ eine Kurve $x : (-\varepsilon, \varepsilon) \to \mathbb{R}^n$ mit $x(0) = x$, $x'(0) = \tilde{d} = d + \alpha v$ und $h_j(x(t)) = 0$ für alle $1 \leqslant j \leqslant p$ und $t \in (-\varepsilon, \varepsilon)$.

Weiter ist $\nabla g_i(x)^T d \leqslant 0$ und damit nach Annahme (ii) auch $\nabla g_i(x)^T(d + \alpha v) < 0$ für alle $i \in \mathcal{A}_X(x)$. Also gilt für $x(t)$ nach Konstruktion

$$\lim_{t \to 0^+} \frac{g_i(x(t)) - g_i(x)}{t} = \nabla g_i(x)^T x'(0) = \nabla g_i(x)^T(d + \alpha v) < 0.$$

Wie im Beweis von Satz 11.4 existiert daher ein $s \in (0, \varepsilon)$ mit $g_i(x(t)) < 0$ für alle $t \in (0, s)$ und $1 \leqslant i \leqslant m$.

Durch Wahl von $t_k := \frac{1}{k}$ und $x^k := x(t_k)$ folgt nun wie zuvor, dass $d + \alpha v \in T_X(x)$ für alle $\alpha > 0$ ist. Aus der Abgeschlossenheit von Tangentialkegeln erhält man nun $d \in T_X(x)$. □

Weitere Regularitätsbedingungen erhält man ebenfalls durch geeignete Kombinationen, etwa die "volle" LICQ.

Folgerung 11.13. *Sei* $x \in X$. *Ist die Menge*

$$\{\nabla g_i(x), \nabla h_j : i \in \mathcal{A}_X(x), 1 \leqslant j \leqslant p\} \subset \mathbb{R}^n$$

linear unabhängig, so ist x *regulär.*

Beweis. Unter dieser Voraussetzung hat das lineare Gleichungssystem

$$\begin{aligned} \nabla g_i(x)^T d &= b_i, \qquad i \in \mathcal{A}_X(x), \\ \nabla h_j(x)^T d &= c_j, \qquad j \in \{1, \dots, p\}, \end{aligned}$$

vollen (Zeilen-)Rang und damit für beliebige $b_i, c_j \in \mathbb{R}$ eine (nicht unbedingt eindeutige) Lösung. Wählt man $b_i < 0$ und $c_j = 0$, ist somit die MFCQ erfüllt, und die Aussage folgt aus Satz 11.12. □

Ebenso kombiniert man die "globalen" Regularitätsbedingungen.

Folgerung 11.14. *Seien alle* g_i *konvex und alle* h_j *affin-linear, d. h. es existieren* $b_j \in \mathbb{R}^n$ *und* $\beta_j \in \mathbb{R}$ *mit* $h_j(x) = b_j^T x - \beta_j$. *Gilt:*

(i) die Menge $\{b_j : 1 \leqslant j \leqslant p\}$ *ist linear unabhängig,*

(ii) es existiert ein $\tilde{x} \in X$ *mit*

$$g_i(\tilde{x}) < 0 \qquad \textit{für alle } 1 \leqslant i \leqslant m,$$

so ist jeder *Punkt* $x \in X$ *regulär.*

Beweis. Wir müssen lediglich noch Annahme (ii) von Satz 11.12 nachprüfen. Dafür betrachten wir wieder für $x \in X$ beliebig die Richtung $v := \tilde{x} - x$. Genau wie im Beweis von Folgerung 11.6 erhält man aus der Konvexität von g_i

$$\nabla g_i(x)^T v \leqslant g_i(\tilde{x}) - g_i(x) = g_i(\tilde{x}) < 0 \qquad \text{für alle } i \in \mathcal{A}_X(x).$$

Außerdem gilt für alle $x \in X$ wegen $\tilde{x} \in X$ und damit insbesondere $h_j(\tilde{x}) = 0$ auch

$$\nabla h_j(x)^T v = b_j^T \tilde{x} - b_j^T x = \beta_j - \beta_j = 0 \qquad \text{für alle } 1 \leqslant j \leqslant p.$$

Also ist die MFCQ erfüllt, und die Aussage folgt aus Satz 11.12. □

Durch Kombination der Beweise von Satz 11.7 und Satz 11.10 erhält man schließlich den folgenden Satz, der erklärt, warum in der linearen Optimierung keine Regularitätsbedingungen notwendig war.

Satz 11.15. *Seien alle* g_i *und* h_j *affin-linear. Dann ist jeder Punkt* $x \in X$ *regulär.*

DIE KKT-BEDINGUNGEN

Ist ein lokaler Minimierer $\bar{x}$ regulär, so kann man aus der abstrakten Optimalitätsbedingung (11.1) eine explizite Charakterisierung von $\bar{x}$ gewinnen. Um dies zu motivieren, betrachten wir den Fall einer einzigen Gleichungsnebenbedingung, $X = \{x \in \mathbb{R}^n : h(x) = 0\}$ für $h : \mathbb{R}^n \to \mathbb{R}$ stetig differenzierbar. Dann ist $L_X(x) = \left\{d \in \mathbb{R}^n : \nabla h(x)^T d = 0\right\} = \ker \nabla h(x)^T$ ein linearer Unterraum (nämlich der *Tangentialraum* von X). Da für Unterräume der Polarkegel mit dem orthogonalen Komplement übereinstimmt, erhalten wir für einen regulären Minimierer $\bar{x}$ aus der äquivalenten Schreibweise (11.2) die Bedingung

$$-\nabla f(\bar{x}) \in T_X(\bar{x})^\circ = L_X(\bar{x})^\circ = (\ker \nabla h(x)^T)^\perp = \operatorname{ran} \nabla h(x),$$

denn für jede lineare Abbildung A gilt $(\ker A)^\perp = \operatorname{ran} A^T$. Nach Definition des Bildes existiert also ein $\bar{\lambda} \in \mathbb{R}$ mit

$$-\nabla f(\bar{x}) = \nabla h(\bar{x})\bar{\lambda},$$

womit wir (im Fall $n = 1$) die klassische Lagrange-Multiplikator-Regel für die Minimierung reeller Funktionen unter Gleichungsnebenbedingungen wiedergewonnen haben. (Die dafür "offensichtlich" hinreichende Bedingung $T_X(x)^\circ = L_X(x)^\circ$ wird auch *Guignard constraint qualification* genannt.)

Für den allgemeinen Fall von beliebig vielen Gleichungs- und Ungleichungsnebenbedingungen verwenden wir statt der Gleichheit $(\ker A)^\perp = \operatorname{ran} A^T$ das *Farkas-Lemma* aus der linearen Optimierung.[1]

Lemma 11.16 (Farkas-Lemma)**.** *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*

$$\begin{cases} Ax + By \leqslant c, \\ Cx + Dy = d, \\ x \geqslant 0. \end{cases}$$

(ii) Es existieren u, v *mit*

$$\begin{cases} u^T A + v^T C \geqslant 0, \\ u^T B + v^T D = 0, \\ u \geqslant 0, \\ u^T c + v^T d < 0. \end{cases}$$

[1] siehe z. B. [Clason 2014, Folgerung 1.7]

Satz 11.17. *Seien* $f : \mathbb{R}^n \to \mathbb{R}$ *stetig differenzierbar und* $X \subset \mathbb{R}^n$ *von der Form* (11.3). *Sei* $\bar{x} \in X$ *ein lokaler Minimierer von* f. *Ist* $\bar{x}$ *regulär, so existieren* $\bar{\mu} \in \mathbb{R}^m$ *und* $\bar{\lambda} \in \mathbb{R}^p$ *mit*

$$\text{(11.6)} \qquad \begin{cases} \nabla f(\bar{x}) + \sum_{i=1}^{m} \bar{\mu}_i \nabla g_i(\bar{x}) + \sum_{j=1}^{p} \bar{\lambda}_j \nabla h_j(\bar{x}) = 0, \\ \qquad\qquad\qquad\qquad\qquad\qquad h_j(\bar{x}) = 0, \quad 1 \leqslant j \leqslant p, \\ \bar{\mu}_i \geqslant 0, \quad g_i(\bar{x}) \leqslant 0, \quad \bar{\mu}_i g_i(\bar{x}) = 0, \quad 1 \leqslant i \leqslant m. \end{cases}$$

Beweis. Aus (11.1) und der Regularität von $\bar{x}$ folgt

$$\nabla f(\bar{x})^T d \geqslant 0 \qquad \text{für alle } d \in L_X(\bar{x}).$$

Also existiert *kein* $v \in \mathbb{R}^n$ mit

$$\nabla f(\bar{x})^T v < 0, \qquad \nabla g_i(\bar{x})^T v \leqslant 0, \qquad \nabla h_j(\bar{x})^T v = 0$$

für alle $i \in \mathcal{A}_X(\bar{x})$ und $1 \leqslant j \leqslant p$. Für

- $A = 0$ und $B = 0$,
- $C = -\nabla g_{\mathcal{A}}(\bar{x})$ (die Matrix, deren Spalten durch $\nabla g_i(\bar{x})$, $i \in \mathcal{A}_X(\bar{x})$ gegeben ist) und $D = -\nabla h(\bar{x})$ (analog aus den Spalten $\nabla h_j(\bar{x})$, $1 \leqslant j \leqslant p$ bestehend),
- $c = 0$ und $d = \nabla f(\bar{x})$,

ist also Aussage (ii) in Lemma 11.16 für beliebige $u \geqslant 0$ *nicht* erfüllt. Es muss daher Aussage (i) gelten, d. h. es gibt $x \geqslant 0$ und y mit $Cx + Dy = d$. Setzen wir

$$\bar{\lambda}_j := y_j \quad \text{für } 1 \leqslant j \leqslant p, \qquad \bar{\mu}_i := \begin{cases} x_i & i \in \mathcal{A}_X(\bar{x}), \\ 0 & i \notin \mathcal{A}_X(\bar{x}), \end{cases}$$

ist dies genau die erste Zeile von (11.6). Nach Definition gilt auch $\bar{\mu}_i g_i(\bar{x}) = 0$ für alle $1 \leqslant i \leqslant m$, woraus zusammen mit der Zulässigkeit von $\bar{x} \in X$ die restlichen Zeilen folgen. □

Analog zur Minimierung reeller Funktionen nennt man $\bar{\lambda}$, $\bar{\mu}$ *Lagrange-Multiplikatoren*; die Bedingungen (11.6) werden *Karush–Kuhn–Tucker-* oder kurz *KKT-Bedingungen* genannt. Die letzte Zeile ist dabei eine *Komplementaritätsbedingung*; man spricht von *strikter Komplementarität*, falls für alle $i \in \mathcal{A}_X(\bar{x})$ gilt $\bar{\mu}_i = 0$ genau dann, wenn $g_i(\bar{x}) < 0$. Die erste Zeile kann man auch mit Hilfe der *Lagrange-Funktion*

$$\text{(11.7)} \qquad L : \mathbb{R}^n \times \mathbb{R}^m \times \mathbb{R}^p \to \mathbb{R}, \qquad (x, \mu, \lambda) \mapsto f(x) + \sum_{i=1}^{m} \mu_i g_i(x) + \sum_{j=1}^{p} \lambda_j h_j(x),$$

auch knapp schreiben als $\nabla_x L(\bar{x}, \bar{\mu}, \bar{\lambda}) = 0$, wobei ∇_x den Gradienten nur bezüglich der ersten Variablen bezeichnet. (Die zweite Zeile kann man dann entsprechend schreiben als $\nabla_\lambda L(\bar{x}, \bar{\mu}, \bar{\lambda}) = 0$.)

Unter stärkeren Regularitätsbedingungen kann man mehr über die Lagrange-Multiplikatoren aussagen.

Folgerung 11.18. *Sei $\bar{x} \in X$ ein lokaler Minimierer, der die LICQ erfüllt. Dann sind die zugehörigen Lagrange-Multiplikatoren $\bar{\lambda} \in \mathbb{R}^p$, $\bar{\mu} \in \mathbb{R}^m$ eindeutig bestimmt.*

Beweis. Zunächst muss für alle $i \notin \mathcal{A}_X(\bar{x})$ wegen der Komplementaritätsbedingung $\bar{\mu}_i = 0$ gelten. Damit reduziert sich die erste Zeile von (11.6) auf

$$\sum_{i \in \mathcal{A}_X(\bar{x})} \bar{\mu}_i \nabla g_i(\bar{x}) + \sum_{j=1}^{p} \bar{\lambda}_j \nabla h_j(\bar{x}) = -\nabla f(\bar{x}).$$

Da nach Voraussetzung die Vektoren $\nabla g_i(\bar{x})$, $i \in \mathcal{A}_X(\bar{x})$ und $\nabla h_j(\bar{x})$, $1 \leqslant j \leqslant p$, linear unabhängig sind, hat dieses Gleichungssystem eine eindeutige Lösung $\bar{\mu}_i$, $i \in \mathcal{A}_X(\bar{x})$, $\bar{\lambda}_j$, $1 \leqslant j \leqslant p$. Also sind die Lagrange-Multiplikatoren eindeutig. □

Für affin-lineare bzw. konvexe Gleichungs- bzw. Ungleichungsrestriktionen sind die KKT-Bedingungen sogar hinreichend. Dafür ist nicht mal eine Slater-Bedingung erforderlich.

Folgerung 11.19. *Seien f und alle g_i konvex und alle h_j affin-linear, d. h. es existieren $b_j \in \mathbb{R}^n$ und $\beta_j \in \mathbb{R}$ mit $h_j(x) = b_j^T x - \beta_j$. Erfüllt ein Tripel $(\bar{x}, \bar{\mu}, \bar{\lambda}) \in \mathbb{R}^n \times \mathbb{R}^m \times \mathbb{R}^p$ die KKT-Bedingungen (11.6), so ist $\bar{x}$ ein globaler Minimierer von f in X.*

Beweis. Aus den KKT-Bedingungen folgt direkt die Zulässigkeit von $\bar{x}$. Sei nun $x \in X$ beliebig. Aus der Konvexität von g_i folgt dann mit Satz 1.4

$$\nabla g_i(\bar{x})^T (x - \bar{x}) \leqslant g_i(x) - g_i(\bar{x}) \leqslant 0 \qquad \text{für alle } i \in \mathcal{A}_X(\bar{x}).$$

Ebenso folgt aus der Konvexität von f mit Satz 1.4 und der ersten Zeile von (11.6)

$$\begin{aligned} f(x) &\geqslant f(\bar{x}) + \nabla f(\bar{x})^T (x - \bar{x}) \\ &= f(\bar{x}) - \sum_{i=1}^{m} \bar{\mu}_i \nabla g_i(\bar{x})^T (x - \bar{x}) - \sum_{j=1}^{p} \bar{\lambda}_j b_j^T (x - \bar{x}) \\ &= f(\bar{x}) - \sum_{i \in \mathcal{A}_X(\bar{x})} \bar{\mu}_i \nabla g_i(\bar{x})^T (x - \bar{x}) \\ &\geqslant f(\bar{x}), \end{aligned}$$

denn wegen $x, \bar{x} \in X$ ist $b_j^T x = \beta_j = b_j^T \bar{x}$ für alle $1 \leqslant j \leqslant p$. Also ist $\bar{x}$ ein globaler Minimierer von f in X. □

Sind schließlich auch f und alle g_i affin-linear, so entsprechen die Lagrange-Multiplikatoren genau den dualen Variablen in der linearen Optimierung, und die KKT-Bedingungen liefern eine weitere Herleitung der dort zentralen schwachen Komplementarität.

HINREICHENDE BEDINGUNGEN

Zum Abschluss betrachten wir Optimalitätsbedingungen zweiter Ordnung, wobei wir uns auf die in der Praxis wesentlich relevanteren hinreichenden Bedingungen beschränken. Während im Falle der unrestringierten Optimierung die Krümmung der Zielfunktion (über die positive Definitheit der Hessematrix) ausschlaggebend ist, müssen wir hier gegebenfals auch die Krümmung der Nebenbedingungen berücksichtigen. Wir brauchen dafür einen weiteren Kegel. Dafür zerlegen wir für einen KKT-Punkt $(\bar{x}, \bar{\mu}, \bar{\lambda})$ die aktive Menge $\mathcal{A}_X(\bar{x})$ in

$$\begin{aligned}\mathcal{A}_+(\bar{x}, \bar{\mu}) &:= \{i \in \mathcal{A}_X(\bar{x}) : \bar{\mu}_i > 0\},\\ \mathcal{A}_0(\bar{x}, \bar{\mu}) &:= \{i \in \mathcal{A}_X(\bar{x}) : \bar{\mu}_i = 0\},\end{aligned}$$

d. h. in die Menge der aktiven Nebenbedingungen, für die strikte Komplementarität gilt bzw. nicht. Wir definieren nun den *kritischen Kegel*

$$K_X(\bar{x}, \bar{\mu}) := \left\{ d \in \mathbb{R}^n : \begin{array}{c} \nabla g_i(\bar{x})^T d = 0,\ i \in \mathcal{A}_+(\bar{x}, \bar{\mu}),\\ \nabla g_i(\bar{x})^T d \leqslant 0,\ i \in \mathcal{A}_0(\bar{x}, \bar{\mu}),\\ \nabla h_j(\bar{x})^T d = 0,\ 1 \leqslant j \leqslant p \end{array} \right\} \subset L_X(\bar{x}).$$

Für die hinreichende Bedingung verwenden wir dann statt der Hesse-Matrix $\nabla^2 f$ die zweite Ableitung $\nabla^2_{xx} L$ der Lagrange-Funktion (11.7) nach x (die ja zusätzlich zu f auch die Nebenbedingungen g_i, h_j enthält), aber dafür nur entlang der kritischen Richtungen.

Satz 11.20. *Seien f, g_i, h_j zweimal stetig differenzierbar. Erfüllt $(\bar{x}, \bar{\mu}, \bar{\lambda}) \in \mathbb{R}^n \times \mathbb{R}^m \times \mathbb{R}^p$ die KKT-Bedingungen (11.6) und gilt*

$$d^T \nabla^2_{xx} L(\bar{x}, \bar{\mu}, \bar{\lambda}) d > 0 \qquad \text{für alle } d \in K_X(\bar{x}, \bar{\mu}) \setminus \{0\}, \tag{11.8}$$

so ist $\bar{x}$ ein strikter lokaler Minimierer von f in X.

Beweis. Angenommen, die Voraussetzungen sind erfüllt, aber $\bar{x} \in X$ ist kein strikter lokaler Minimierer. Dann existiert eine Folge $\{x^k\}_{k\in\mathbb{N}} \subset X \setminus \{\bar{x}\}$ mit $x^k \to \bar{x}$ und $f(x^k) \leqslant f(\bar{x})$ für alle $k \in \mathbb{N}$. Wir definieren nun eine Folge $\{d^k\}_{k\in\mathbb{N}}$ durch

$$d^k := \frac{x^k - \bar{x}}{\|x^k - \bar{x}\|}.$$

Da wegen $\|d^k\| = 1$ diese Folge beschränkt ist, existiert eine konvergente Teilfolge $\{d^k\}_{k\in K}$ mit $d^k \to d$ für ein $d \in \mathbb{R}^n$ mit $\|d\| = 1$, d. h. $d \neq 0$. Wir zeigen nun, dass $d \in L_X(\bar{x})$ gilt. Für beliebige $i \in \mathcal{A}_X(\bar{x})$ und $k \in \mathbb{N}$ folgt aus Satz 1.1 die Existenz eines $\xi^{i,k}$ mit

$$\nabla g_i(\xi^{i,k})^T (x^k - \bar{x}) = g_i(x^k) - g_i(\bar{x}) \leqslant 0$$

wegen $g_i(x^k) \leqslant 0$ für $x^k \in X$ und $g_i(\bar{x}) = 0$ für $i \in \mathcal{A}_X(\bar{x})$. Wie üblich folgt aus $x^k \to \bar{x}$ auch $\xi^{i,k} \to \bar{x}$, und Division durch $\|x^k - \bar{x}\|$ und Grenzübergang $K \ni k \to \infty$ liefert

$$\nabla g_i(\bar{x})^T d \leqslant 0 \qquad \text{für alle } i \in \mathcal{A}_X(\bar{x}). \tag{11.9}$$

Analog folgt für beliebige $1 \leqslant j \leqslant p$ wegen $x^k, x \in X$ auch

$$\nabla h_j(\xi^k)^T(x^k - \bar{x}) = h_j(x^k) - h_j(\bar{x}) = 0$$

und damit nach Division durch $\|x^k - \bar{x}\|$ und Grenzübergang

$$\nabla h_j(\bar{x})^T d = 0 \qquad \text{für alle } 1 \leqslant j \leqslant p. \tag{11.10}$$

Also ist $d \in L_X(\bar{x}) \setminus \{0\}$.

Wir machen nun eine Fallunterscheidung.

(i) $d \in K_X(\bar{x}, \bar{\mu})$. In diesem Fall zeigen wir, dass d die Bedingung (11.8) verletzt. Zunächst gilt für alle x^k nach Konstruktion

$$f(\bar{x}) \geqslant f(x^k) \geqslant f(x^k) + \sum_{i=1}^{m} \bar{\mu}_i g_i(x^k) + \sum_{j=1}^{p} \bar{\lambda}_j h_j(x^k) = L(x^k, \bar{\mu}, \bar{\lambda}),$$

da x^k zulässig und $\bar{\mu}_i \geqslant 0$ ist für alle $1 \leqslant i \leqslant m$. Wir wenden nun Satz 1.2 auf die Funktion $x \mapsto L(x, \bar{\mu}, \bar{\lambda})$ an und erhalten

$$\begin{aligned} f(\bar{x}) &\geqslant L(x^k, \bar{\mu}, \bar{\lambda}) \\ &= L(\bar{x}, \bar{\mu}, \bar{\lambda}) + \nabla_x L(\bar{x}, \bar{\mu}, \bar{\lambda})^T(x^k - \bar{x}) + \frac{1}{2}(x^k - \bar{x})^T \nabla^2_{xx} L(\xi^k, \bar{\mu}, \bar{\lambda})^T(x^k - \bar{x}) \\ &= f(\bar{x}) + \frac{1}{2}(x^k - \bar{x})^T \nabla^2_{xx} L(\xi^k, \bar{\mu}, \bar{\lambda})^T(x^k - \bar{x}), \end{aligned}$$

da aufgrund der KKT-Bedingungen $L(\bar{x}, \bar{\mu}, \bar{\lambda}) = f(\bar{x})$ sowie $\nabla_x L(\bar{x}, \bar{\mu}, \bar{\lambda}) = 0$ gilt. Wieder gilt $\xi^k \to \bar{x}$ wegen $x^k \to \bar{x}$, und Division durch $\|x^k - \bar{x}\|^2$ und Grenzübergang $K \ni k \to \infty$ liefert

$$d^T \nabla^2_{xx} L(\bar{x}, \bar{\mu}, \bar{\lambda}) d \leqslant 0 \qquad \text{für } d \in K_X(\bar{x}, \bar{\mu}),$$

im Widerspruch zu (11.8).

(ii) $d \notin K_X(\bar{x}, \bar{\mu})$. Dann existiert ein $i_+ \in \mathcal{A}_+$ mit $\nabla g_{i_+}(\bar{x})^T d < 0$. Analog zu oben folgt mit $f(x^k) \leqslant f(\bar{x})$ aus

$$\nabla f(\xi^k)^T(x^k - \bar{x}) = f(x^k) - f(\bar{x}) \leqslant 0$$

durch Division und Grenzübergang zusammen mit $\nabla_x L(\bar{x}, \bar{\mu}, \bar{\lambda}) = 0$ und (11.9), (11.10) die Ungleichung

$$0 \geqslant \nabla f(\bar{x})^T d = -\sum_{i=1}^{m} \bar{\mu}_i \nabla g_i(\bar{x})^T d - \sum_{j=1}^{p} \bar{\lambda}_j \nabla h_j(\bar{x})^T d \geqslant -\bar{\mu}_{i_+} \nabla g_{i_+}(\bar{x})^T d > 0$$

und damit ebenfalls ein Widerspruch. □

Gilt andererseits (11.8) mit größer oder gleich und zusätzlich eine geeignete Regularitätsbedingung (z. B. die LICQ), so kann man zeigen, dass dies eine notwendige Optimalitätsbedingung zweiter Ordnung darstellt; siehe z. B. [Geiger und Kanzow 2002, Satz 2.54].

STRAF- UND BARRIEREVERFAHREN

12

Wir kommen nun zu Verfahren zur numerischen Lösung von Optimierungsproblemen mit Nebenbedingungen. Ein klassischer Ansatz besteht darin, das Problem durch eine Folge von unrestringierten Problemen zu approximieren, indem man die Zielfunktion so modifiziert, dass das Verlassen des zulässigen Bereichs X zunehmend "teuer" wird. Bei *Strafverfahren* (auch *Penalty-Verfahren*) wird dabei eine *Straffunktion* $\pi : \mathbb{R}^n \to \mathbb{R}$ mit $\pi(x) = 0$ für $x \in X$ und $\pi(x) > 0$ für $x \notin X$ addiert, die mit einem *Penalty-Parameter* $\alpha > 0$ gewichtet wird. Man betrachtet also das Problem

$$\min_{x \in \mathbb{R}^n} f(x) + \alpha\pi(x).$$

Je größer α gewählt ist, desto mehr Wert wird auf die (näherungsweise) Erfüllung der Nebenbedingung $x \in X$ gelegt. Man hofft daher, dass für $\alpha \to \infty$ die Folge $\{x_\alpha\}_{\alpha>0}$ der entsprechenden Minimierer gegen einen Minimierer $\bar{x} \in X$ von f konvergiert.

QUADRATISCHE STRAFVERFAHREN

Wir betrachten wieder den konkreten Fall

$$X = \{x \in \mathbb{R}^n : g_i(x) \leqslant 0,\ 1 \leqslant i \leqslant m, \quad h_j(x) = 0,\ 1 \leqslant j \leqslant p\}$$

für $g_i, h_j : \mathbb{R}^n \to \mathbb{R}$ stetig differenzierbar, und bestrafen die Nebenbedingungen einzeln. Im *quadratischen Strafverfahren* wählt man dazu quadratische Funktionen, und zwar

(i) für die Gleichungsnebenbedingung $h_j(x) = 0$ die Funktion $x \mapsto \frac{1}{2}|h_j(x)|^2$;

(ii) für die Ungleichungsnebenbedingung $g_i(x) \leqslant 0$ die Funktion $x \mapsto \frac{1}{2}|(g_i)^+|^2$, wobei $(t)^+ := \max\{0, t\}$ bezeichnet.

Man minimiert nun anstelle von f für $\alpha > 0$ die Funktion

$$\begin{aligned} P_\alpha(x) &:= f(x) + \frac{\alpha}{2}\sum_{i=1}^{m} |(g_i)^+|^2 + \frac{\alpha}{2}\sum_{j=1}^{p} |h_j(x)|^2 \\ &= f(x) + \frac{\alpha}{2}\|(g(x))^+\|^2 + \frac{\alpha}{2}\|h(x)\|^2, \end{aligned}$$

wobei wir wieder die Nebenbedingungen zu vektorwertigen Funktionen $g : \mathbb{R}^n \to \mathbb{R}^m$ und $h : \mathbb{R}^n \to \mathbb{R}^p$ zusammengesetzt haben, und $(v)^+$ für $v \in \mathbb{R}^m$ komponentenweise zu verstehen ist. Offensichtlich gilt $P_\alpha(x) = f(x)$ für alle $x \in X$ und $\alpha > 0$. Wir setzen in Folge kurz

$$\pi(x) := \frac{1}{2}\left(\|(g(x))^+\|^2 + \|h(x)\|^2\right).$$

Wir fragen uns zuerst, wann das penalisierte Problem

$$\min_{x\in\mathbb{R}^n} P_\alpha(x) \tag{12.1}$$

eine Lösung besitzt. Dafür müssen wir annehmen, dass f auf ganz $\mathbb{R}^n$ wohldefiniert ist. Außerdem brauchen wir eine Annahme an die Darstellung der Menge X.

Satz 12.1. *Sei $f : \mathbb{R}^n \to \mathbb{R}$ stetig, $X \subset \mathbb{R}^n$ nichtleer und abgeschlossen, und entweder*

(i) f koerziv oder

(ii) X beschränkt, π koerziv, und f nach unten beschränkt.

Dann existiert für alle $\alpha > 0$ eine Lösung $x_\alpha \in \mathbb{R}^n$ von (12.1).

Beweis. Gilt (i) oder (ii), so ist $P_\alpha = f + \alpha\pi$ die Summe einer nach unten beschränkten und einer koerziven Funktion und damit koerziv. Da mit g und h (und $t \mapsto (t)^+$) auch $\pi(x)$ stetig ist, folgt die Existenz aus Satz 2.2. □

Wir nehmen in Folge an, dass die Bedinungen von Satz 12.1 erfüllt sind. Ein Strafverfahren hat dann die Form von

Algorithmus 12.1 : Quadratisches Strafverfahren

Input : $\alpha_0 > 0$, $x^0 \in \mathbb{R}^n$

for $k = 0, \dots$ **do**
  Berechne x^{k+1} als globalen Minimierer von (12.1) mit $\alpha = \alpha_k$ (und Startwert x^k)
  if $x^{k+1} \in X$ **then**
    return x^{k+1}
  else
    Wähle $\alpha_{k+1} > \alpha_k$

Um die Konvergenz dieses Verfahrens zu untersuchen, zeigen wir zuerst nützliche Eigenschaften der so erzeugten Folge $\{x^k\}_{k\in\mathbb{N}}$.

Lemma 12.2. *Sei* $f : \mathbb{R}^n \to \mathbb{R}$ *stetig und* $X \subset \mathbb{R}^n$ *nichtleer. Angenommen, Algorithmus 12.1 erzeugt für eine streng monoton wachsende, unbeschränkte Folge* $\{\alpha_k\}_{k\in\mathbb{N}} \subset (0,\infty)$ *eine unendliche Folge* $\{x^k\}_{k\in\mathbb{N}}$. *Dann gilt*

(i) $\{P_{\alpha_k}(x^k)\}_{k\in\mathbb{N}}$ *ist monoton wachsend;*

(ii) $\{\pi(x^k)\}_{k\in\mathbb{N}}$ *ist monoton fallend;*

(iii) $\{f(x^k)\}_{k\in\mathbb{N}}$ *ist monoton wachsend;*

(iv) $\{(g_i(x^k))^+\}_{k\in\mathbb{N}}$, $1 \leqslant i \leqslant m$, *und* $\{h_j(x^k)\}_{k\in\mathbb{N}}$, $1 \leqslant j \leqslant p$, *sind Nullfolgen.*

Beweis. Zu (i): Aus der globalen Optimalität von x^k und $\alpha_k < \alpha_{k+1}$ folgt

$$\begin{aligned} P_{\alpha_k}(x^k) \leqslant P_{\alpha_k}(x^{k+1}) &= f(x^{k+1}) + \alpha_k \pi(x^{k+1}) \\ &\leqslant f(x^{k+1}) + \alpha_{k+1}\pi(x^{k+1}) = P_{\alpha_{k+1}}(x^{k+1}). \end{aligned}$$

Zu (ii): Aus $P_{\alpha_k}(x^k) \leqslant P_{\alpha_k}(x^{k+1})$ und $P_{\alpha_{k+1}}(x^{k+1}) \leqslant P_{\alpha_{k+1}}(x^k)$ folgt durch Addition

$$\alpha_k \pi(x^k) + \alpha_{k+1}\pi(x^{k+1}) \leqslant \alpha_k \pi(x^{k+1}) + \alpha_{k+1}\pi(x^k),$$

was durch Umformen

$$(\alpha_k - \alpha_{k+1})\left(\pi(x^k) - \pi(x^{k+1})\right) \leqslant 0$$

ergibt. Aus $\alpha_k < \alpha_{k+1}$ folgt nun $\pi(x^k) \geqslant \pi(x^{k+1})$.

Zu (iii): Aus der Optimalität von x^k und (ii) folgt sofort

$$f(x^k) + \alpha_k \pi(x^k) \leqslant f(x^{k+1}) + \alpha_k \pi(x^{k+1}) \leqslant f(x^{k+1}) + \alpha_k \pi(x^k)$$

und damit $f(x^k) \leqslant f(x^{k+1})$.

Zu (iv): Da X nichtleer ist, existiert ein $\hat{x} \in X$. Aus der Optimalität von x^k und (iii) folgt dann

$$f(\hat{x}) = f(\hat{x}) + \alpha_k \pi(\hat{x}) \geqslant f(x^k) + \alpha_k \pi(x^k) \geqslant f(x^0) + \alpha_k \pi(x^k).$$

Wegen $\alpha_k \to \infty$ folgt daraus nun

$$\pi(x^k) \leqslant \frac{1}{\alpha_k}\left(f(\hat{x}) - f(x^0)\right) \to 0.$$

Nach Definition von π müssen daher auch $(g_i(x^k))^+ \to 0$ und $h(x^k) \to 0$ gehen. □

Damit können wir nun die Konvergenz zeigen.

Satz 12.3. *Sei $f : \mathbb{R}^n \to \mathbb{R}$ stetig und $X \subset \mathbb{R}^n$ nichtleer. Dann bricht Algorithmus 12.1 entweder nach endlich vielen Schritten in einem globalen Minimierer ab oder erzeugt für eine streng monoton wachsende, unbeschränkte Folge $\{\alpha_k\}_{k\in\mathbb{N}} \subset (0,\infty)$ eine unendliche Folge $\{x^k\}_{k\in\mathbb{N}}$, für die jeder Häufungspunkt ein globaler Minimierer ist.*

Beweis. Nach Satz 12.1 kann das Strafverfahren nur im Fall $x^k \in X$ abbrechen. In diesem Fall gilt aber

$$f(x^k) = f(x^k) + \alpha_k \pi(x^k) \leqslant f(x) + \alpha_k \pi(x) = f(x)$$

für alle $x \in X$, d. h. x^k ist globaler Minimierer von f in X

Andernfalls sei $\bar{x} \in \mathbb{R}^n$ ein Häufungspunkt von $\{x^k\}_{k\in\mathbb{N}}$ und $\{x^k\}_{k\in K}$ eine gegen $\bar{x}$ konvergente Teilfolge. Da mit g und h (und $t \mapsto (t)^+$) auch $\pi(x)$ stetig ist, folgt aus Lemma 12.2 (iv)

$$\pi(\bar{x}) = \lim_{k\to\infty} \pi(x^k) = 0.$$

Nach Definition von π impliziert dies $\bar{x} \in X$. Für alle $x \in X$ und $k \in K$ gilt wegen $\pi(x) \geqslant 0$ daher

$$f(x^k) \leqslant f(x^k) + \alpha_k \pi(x^k) \leqslant f(x) + \alpha_k \pi(x) = f(x).$$

Durch Grenzübergang $k \to \infty$ auf beiden Seiten folgt daraus $f(\bar{x}) \leqslant f(x)$ für alle $x \in X$, d. h. $\bar{x}$ ist globaler Minimierer von f in X. □

Um die in Schritt 3 benötigte Lösung von (12.1) zu berechnen, möchten wir Verfahren aus Teil II einsetzen, für die P_α zumindest stetig differenzierbar sein muss. Problematisch ist dabei nur die Penalisierung der Ungleichungsnebenbedingungen. Durch Fallunterscheidung $t > 0$, $t < 0$ und $t = 0$ in der Definition der reellen Ableitung vergewissert man sich aber schnell, dass gilt

$$\frac{d}{dt}\left(\tfrac{1}{2}|(t)^+|^2\right) = (t)^+.$$

Aus der Summen- und Kettenregel folgt daher

$$\nabla P_\alpha(x) = \nabla f(x) + \alpha \sum_{i=1}^{m} (g_i(x))^+ \nabla g_i(x) + \alpha \sum_{j=1}^{p} h_j(x) \nabla h_j(x). \tag{12.2}$$

Vergleicht man dies mit (11.7), so erhält man

$$\nabla P_\alpha(x) = \nabla_x L(x, \mu, \lambda) \qquad \text{für} \quad \mu := \alpha(g(x))^+, \quad \lambda := \alpha h(x).$$

Tatsächlich kann man (unter einer Regularitätsbedingung) zeigen, dass die zu $\{x^k\}_{k\in\mathbb{N}}$ gehörenden Folgen $\{\mu^k\}_{k\in\mathbb{N}}$, $\{\lambda^k\}_{k\in\mathbb{N}}$ gegen die entsprechenden Lagrange-Multiplikatoren des restringierten Problems konvergieren.

Satz 12.4. *Seien* $f : \mathbb{R}^n \to \mathbb{R}$, $g : \mathbb{R}^n \to \mathbb{R}^m$ *und* $h : \mathbb{R}^n \to \mathbb{R}^p$ *stetig differenzierbar. Konvergiert die durch Algorithmus 12.1 erzeugte Folge* $\{x^k\}_{k\in\mathbb{N}}$ *gegen einen Punkt* $\bar{x} \in X$*, der die LICQ erfüllt, so konvergieren auch*

$$\mu^k := \alpha_k (g(x^k))^+ \to \bar{\mu}, \qquad \lambda^k := \alpha_k h(x^k) \to \bar{\lambda},$$

und $(\bar{x}, \bar{\mu}, \bar{\lambda})$ *erfüllt die KKT-Bedingungen* (11.6).

Beweis. Wir zeigen zuerst die Konvergenz von $\{\mu^k\}_{k\in\mathbb{N}}$ und $\{\lambda^k\}_{k\in\mathbb{N}}$. Wir unterscheiden wieder aktive und inaktive Nebenbedingungen: Für $i \notin \mathcal{A}_X(\bar{x})$ gilt $g_i(\bar{x}) < 0$. Aus der Stetigkeit von g_i folgt dann $g_i(x^k) < 0$ und damit auch $(g_i(x^k))^+ = 0$ für $k \in \mathbb{N}$ groß genug. Also gilt

$$\mu_i^k = \alpha_k (g(x^k))^+ \to 0 =: \bar{\mu}_i \qquad \text{für alle } i \notin \mathcal{A}(\bar{x}). \tag{12.3}$$

Für die restlichen Komponenten verwenden wir die LICQ. Wir bezeichnen mit A_k diejenige Matrix, die aus den Spalten $\nabla g_i(x^k)$, $i \in \mathcal{A}_X(\bar{x})$, und $\nabla h_j(x^k)$, $1 \leqslant j \leqslant p$, gebildet wird. Da nach Voraussetzung g und h stetig differenzierbar sind, konvergiert die Folge dieser Matrizen gegen die Matrix $\bar{A}$, die entsprechend aus den Spalten $\nabla g_i(\bar{x})$ und $\nabla h_j(\bar{x})$ gebildet wird. Weiterhin hat $\bar{A}$ aufgrund der LICQ vollen Spaltenrang, und damit ist $\bar{A}^T\bar{A}$ invertierbar. Nach Lemma 7.4 ist damit auch $A_k^T A_k$ für $k \in \mathbb{N}$ hinreichend groß invertierbar, und es gilt $(A_k^T A_k)^{-1} \to (\bar{A}^T\bar{A})^{-1}$.

Nun ist x^k ein unrestringierter Minimierer von P_{α_k}, erfüllt also die notwendige Optimalitätsbedingung $\nabla P_{\alpha_k}(x^k) = 0$. Für $k \in \mathbb{N}$ mit $g_i(x^k) < 0$ folgt daraus (vergleiche (12.2))

$$0 = A_k^T \nabla P_{\alpha_k}(x^k) = A_k^T \nabla f(x^k) + A_k^T A_k \begin{pmatrix} \mu_{\mathcal{A}}^k \\ \lambda^k \end{pmatrix},$$

wobei $\mu_{\mathcal{A}}^k$ den Vektor bezeichnet, der aus den Komponenten μ_i^k, $i \in \mathcal{A}_X(\bar{x})$, besteht. Aufgrund der Stetigkeit von ∇f und der Konvergenz $A_k \to \bar{A}$ erhalten wir damit

$$\begin{pmatrix} \mu_{\mathcal{A}}^k \\ \lambda^k \end{pmatrix} = -(A_k^T A_k)^{-1} A_k^T \nabla f(x^k) \to -(\bar{A}^T\bar{A})^{-1}\bar{A}\nabla f(\bar{x}) =: \begin{pmatrix} \bar{\mu}_{\mathcal{A}} \\ \bar{\lambda} \end{pmatrix}.$$

Damit ist die Konvergenz der Folgen $\{\mu^k\}_{k\in\mathbb{N}}$ und $\{\lambda^k\}_{k\in\mathbb{N}}$ gezeigt.

Für die KKT-Bedingungen folgt aus der Optimalität der x^k, der Definition der μ^k und λ^k sowie der Stetigkeit von ∇f, ∇g und ∇h

$$\nabla_x L(\bar{x}, \bar{\mu}, \bar{\lambda}) = \lim_{k\to\infty} \nabla_x L(x^k, \mu^k, \lambda^k) = \lim_{k\to\infty} \nabla P_{\alpha^k}(x^k) = 0$$

und damit die erste Relation von (11.6). Aus Satz 12.3 folgt insbesondere $\bar{x} \in X$ und damit $h_j(\bar{x}) = 0$ für $1 \leqslant j \leqslant p$, d. h. die zweite Relation. Ebenso gilt $g_i(\bar{x}) \leqslant 0$ sowie nach Definition

$$\bar{\mu}_i = \lim_{k\to\infty} \alpha_k (g_i(x^k))^+ \geqslant 0$$

für alle $1 \leqslant i \leqslant m$. Schließlich folgt aus (12.3) auch die Komplementaritätsbedingung $\bar{\mu}_i g_i(\bar{x}) = 0$ für alle $1 \leqslant i \leqslant m$ und damit die dritte Relation. □

Ein Nachteil der quadratischen Penalisierung ist, dass in der Regel $x_\alpha \notin X$ gilt. Aus (12.2) folgt nämlich $\nabla P_\alpha(x) = \nabla f(x)$ für alle $x \in X$. Wäre also $x_\alpha \in X$ ein Minimierer von P_α, so würde aus der notwendigen Optimalitätsbedingung folgen

$$0 = \nabla P_\alpha(x_\alpha) = \nabla f(x_\alpha),$$

was nur möglich ist, wenn x_α stationärer Punkt von f im Inneren von X ist. Man muß also tatsächlich $\alpha \to \infty$ streben lassen; erschwerend kommt hinzu, dass in der Regel die Minimierung von P_α mit wachsendem α zunehmend schwieriger wird (z. B. durch wachsende Konditionszahl der Newton-Systeme).

EXAKTE STRAFVERFAHREN

Dieser Nachteil kann durch eine unterschiedliche Wahl der Straffunktion vermieden werden. Eine Straffunktion $\pi : \mathbb{R}^n \to \mathbb{R}$ heißt *exakt* in einem Minimierer $\bar{x} \in X$, wenn ein endliches $\bar{\alpha} > 0$ existiert, so dass $\bar{x}$ auch ein unrestringierter Minimierer von $f + \alpha\pi$ für alle $\alpha > \bar{\alpha}$ ist.

Eine Variante besteht darin, anstelle der Quadrate den Absolutbetrag der Nebenbedingungen zu penalisieren. Man betrachtet also die Straffunktion

$$\pi_1(x) := \sum_{i=1}^{m} (g_i(x))^+ + \sum_{j=1}^{p} |h_j(x)| = \|(g(x))^+\|_1 + \|h(x)\|_1.$$

Für konvexe Probleme kann man zeigen, dass π_1 für α groß genug in der Tat exakt ist.

Satz 12.5. *Seien $f : \mathbb{R}^n \to \mathbb{R}$ und $g_i : \mathbb{R}^n \to \mathbb{R}$, $1 \leqslant i \leqslant m$, stetig differenzierbar und konvex, und sei $h_j : \mathbb{R}^n \to \mathbb{R}$, $1 \leqslant j \leqslant p$, affin-linear. Es sei weiter die Slater-Bedingung (11.5) erfüllt. Dann ist π_1 exakt in jedem Minimierer $\bar{x} \in X$.*

Beweis. Unter den Annahmen an die Nebenbedingung sind alle Punkte $x \in X$ regulär; jeder Minimierer $\bar{x} \in X$ von f erfüllt also die KKT-Bedingungen (11.6). Insbesondere existieren Lagrange-Multiplikatoren $\bar{\mu} \in \mathbb{R}^m$ und $\bar{\lambda} \in \mathbb{R}^p$. Wegen $\bar{\mu} \geqslant 0$ und der Konvexität aller Funktionen ist auch die Abbildung $x \mapsto L(x, \bar{\mu}, \bar{\lambda})$ konvex. Aus Satz 1.4 und der ersten Relation in (11.6) folgt daher für alle $x \in \mathbb{R}^n$

$$L(x, \bar{\mu}, \bar{\lambda}) \geqslant L(\bar{x}, \bar{\mu}, \bar{\lambda}) + \nabla_x L(\bar{x}, \bar{\mu}, \bar{\lambda})^T (x - \bar{x}) = L(\bar{x}, \bar{\mu}, \bar{\lambda}).$$

Weiterhin folgt aus der Definition der Straffunktion sowie den restlichen KKT-Bedingungen

$$f(\bar{x}) + \alpha\pi_1(\bar{x}) = f(\bar{x}) = f(\bar{x}) + \sum_{i=1}^{m} \bar{\mu}_i g_i(\bar{x}) + \sum_{j=1}^{p} \bar{\lambda}_j h_j(\bar{x}) = L(\bar{x}, \bar{\mu}, \bar{\lambda}).$$

Wir wählen nun

$$\bar{\alpha} := \max\{\bar{\mu}_1, \dots, \bar{\mu}_m, |\bar{\lambda}_1|, \dots, |\bar{\lambda}_p|\}. \tag{12.4}$$

Dann gilt für alle $\alpha \geqslant \bar{\alpha}$ und $x \in \mathbb{R}^n$ wegen $\bar{\mu}_i \geqslant 0$

$$\begin{aligned} f(\bar{x}) + \alpha\pi_1(\bar{x}) &= f(\bar{x}) + \sum_{i=1}^{m} \bar{\mu}_i g_i(\bar{x}) + \sum_{j=1}^{p} \bar{\lambda}_j h_j(\bar{x}) \\ &\leqslant f(x) + \sum_{i=1}^{m} \bar{\mu}_i g_i(x) + \sum_{j=1}^{p} \bar{\lambda}_j h_j(x) \\ &\leqslant f(x) + \sum_{i=1}^{m} \bar{\mu}_i (g_i(x))^+ + \sum_{j=1}^{p} |\bar{\lambda}_j| |h_j(x)| \\ &\leqslant f(x) + \alpha \sum_{i=1}^{m} (g_i(x))^+ + \alpha \sum_{j=1}^{p} |h_j(x)| \\ &= f(x) + \alpha\pi_1(x). \end{aligned}$$

Also ist $\bar{x}$ globaler Minimierer von $f + \alpha\pi_1$. □

Allerdings gibt es nichts geschenkt; da der Betrag (bzw. die Funktion $t \mapsto (t)^+$) nicht differenzierbar ist, ist auch $f + \alpha\pi_1$ nicht differenzierbar. Die für das Strafverfahren benötigten Minimierer können also nicht mit den Methoden aus Teil II berechnet werden. Tatsächlich sind exakte Straffunktionen notwendig nicht differenzierbar (außer in Minimierern, die im Inneren von X liegen).

Satz 12.6. *Sei $f : \mathbb{R}^n \to \mathbb{R}$ stetig differenzierbar und sei $\bar{x} \in X$ ein Minimierer mit $\nabla f(\bar{x}) \neq 0$. Ist $\pi : \mathbb{R}^n \to \mathbb{R}$ eine differenzierbare Straffunktion, so ist sie nicht exakt in $\bar{x}$.*

Beweis. Angenommen, $\pi : \mathbb{R}^n \to \mathbb{R}$ wäre eine differenzierbare, exakte Straffunktion. Dann existiert ein $\bar{\alpha} > 0$, so dass $\bar{x}$ ein unrestringierter Minimierer von $f + \alpha\pi$ für alle $\alpha \geqslant \bar{\alpha}$ ist. Da $f + \alpha\pi$ nach Annahme differenzierbar ist, folgt aus der notwendigen Optimalitätsbedingung

$$\nabla f(\bar{x}) + \alpha \nabla \pi(\bar{x}) = 0. \tag{12.5}$$

Für beliebige $\alpha_1, \alpha_2 \geqslant \bar{\alpha}$ mit $\alpha_1 \neq \alpha_2$ gilt daher

$$\nabla f(\bar{x}) + \alpha_1 \nabla \pi(\bar{x}) = 0 = \nabla f(\bar{x}) + \alpha_2 \nabla \pi(\bar{x}),$$

was durch Umformen

$$(\alpha_1 - \alpha_2) \nabla \pi(\bar{x}) = 0$$

ergibt. Daraus folgt $\nabla \pi(\bar{x}) = 0$ und damit wegen (12.5) auch $\nabla f(\bar{x}) = 0$, im Widerspruch zur Annahme. □

BARRIEREVERFAHREN

Strafverfahren sind ungeeignet, wenn die Zielfunktion f für $x \notin X$ gar nicht definiert ist. In *Barriereverfahren* wird dagegen der Minimierer $\bar{x} \in X$ durch eine Folge von *inneren* Punkten angenähert. (Man spricht daher auch von *innere-Punkte-Verfahren*.) Da für Gleichungsnebenbedingungen der zulässige Bereich keine inneren Punkte besitzt, betrachten wir hier nur reine Ungleichungsnebenbedingungen

$$X = \{x \in \mathbb{R}^n : g_i(x) \leqslant 0,\ 1 \leqslant i \leqslant m\}$$

für $g_i : \mathbb{R}^n \to \mathbb{R}$ stetig differenzierbar. (Zusätzliche Gleichungsnebenbedingungen werden entweder durch eine Straffunktion oder den im folgenden Kapitel vorgestellten Ansatz behandelt.) Anstelle einer Straffunktion, die erst außerhalb von X zu wirken beginnt, verwendet man nun eine *Barrierefunktion*, die bereits bei Annäherung an den Rand von X gegen unendlich strebt; verbreitet ist dabei die logarithmische Barrierefunktion $x \mapsto -\ln(-g_i(x))$. Anstelle von f minimiert man daher für $\alpha > 0$ die Funktion

$$B_\alpha(x) := f(x) - \alpha \sum_{i=1}^{m} \ln(-g_i(x)) =: f(x) + \alpha\beta(x).$$

Ist diese Funktion nicht definiert wegen $g_i(x) \geqslant 0$ für ein $1 \leqslant i \leqslant m$, so setzen wir $B_\alpha(x) := \infty$. Damit also eine Lösung von

$$\min_{x \in \mathbb{R}^n} B_\alpha(x) \tag{12.6}$$

existiert, muss es einen *strikt* zulässigen Punkt mit $g_i(x) < 0$ für alle $1 \leqslant i \leqslant m$ geben – dies ist genau die Slater-Bedingung (11.6).

Satz 12.7. *Sei $f : X \to \mathbb{R}$ stetig, $X \subset \mathbb{R}^n$, beschränkt, nichtleer und abgeschlossen. Gilt die Slater-Bedingung (11.6), dann existiert für alle $\alpha > 0$ eine Lösung $x_\alpha \in X$ von (12.6).*

Beweis. Die Slater-Bedingung garantiert die Existenz eines strikt zulässigen Punktes $\tilde{x} \in X$, für den

$$M := B_\alpha(\tilde{x}) < \infty$$

gilt. Wir betrachten nun die Menge

$$X_M := \{x \in X : B_\alpha(x) \leqslant M\}.$$

Offensichtlich muss ein Minimierer von B_α (falls existent) in X_M liegen, d. h. auch Lösung sein von

$$\min_{x \in X_M} B_\alpha(x). \tag{12.7}$$

Es genügt also zu zeigen, dass dieses Problem eine Lösung hat.

Wir zeigen zuerst, dass X_M kompakt ist. Mit X ist auch $X_M \subset X$ beschränkt. Für die Abgeschlossenheit sei $\{x^k\}_{k\in\mathbb{N}} \subset X_M$ eine konvergente Folge mit Grenzwert $\bar{x} \in X$. Wegen der Stetigkeit von f, g_i, $1 \leqslant i \leqslant m$, sowie $t \mapsto \ln(t)$ ist B_α stetig auf X_M (beachte, dass $g_i(x) < 0$ für alle $x \in X_M$ gelten muss). Durch Grenzübergang folgt daher auch $B_\alpha(\bar{x}) \leqslant M$, d. h. $\bar{x} \in X_M$. Also ist X_M kompakt, und aus der Stetigkeit von B_α folgt mit Satz 2.2 die Existenz einer Lösung von (12.7). □

Man kann nun hoffen, dass für $\alpha \to 0$ die Folge $\{x_\alpha\}_{\alpha>0}$ der entsprechenden Minimierer von (12.6) gegen einen Minimierer $\bar{x} \in X$ von f konvergiert. Entsprechend hat das Barriereverfahren nun die Form von

Algorithmus 12.2 : Barriereverfahren

Input : $\alpha_0 > 0$, $x^0 \in X$ strikt zulässig

1 Setze $k = 0$
2 **for** $k = 0, \dots$ **do**
3 Berechne x^{k+1} als globalen Minimierer von (12.6) mit $\alpha = \alpha_k$ (und Startwert x^k)
4 Wähle $\alpha_{k+1} < \alpha_k$

Wieder wird man in Schritt 3 Verfahren aus Teil II anwenden, wobei man diesmal (etwa bei der Schrittweitensuche) aufpassen muss, dass nur strikt zulässige Iterierte erzeugt werden.

Auch hier kann man Monotonieeigenschaften der so erzeugten Folge zeigen.

Lemma 12.8. *Angenommen, Algorithmus 12.2 erzeugt für eine streng monoton fallende Nullfolge* $\{\alpha_k\}_{k\in\mathbb{N}} \subset (0,\infty)$ *eine unendliche Folge* $\{x^k\}_{k\in\mathbb{N}} \subset X$. *Dann gilt*

1. $\{\beta(x^k)\}_{k\in\mathbb{N}}$ *ist monoton wachsend;*
2. $\{f(x^k)\}_{k\in\mathbb{N}}$ *ist monoton fallend.*

Beweis. Zu (i): Wir verwenden wieder die globale Optimalität zusammen mit der strikten Zulässigkeit aller x^k. Aus $B_{\alpha_k}(x^k) \leqslant B_{\alpha_k}(x^{k+1})$ und $B_{\alpha_{k+1}}(x^{k+1}) \leqslant B_{\alpha_{k+1}}(x^k)$ folgt durch Addition und Umformen

$$(\alpha_k - \alpha_{k+1})\left(\beta(x^k) - \beta(x^{k+1})\right) \leqslant 0.$$

Aus $\alpha_k > \alpha_{k+1}$ folgt nun $\beta(x^k) \leqslant \beta(x^{k+1})$.

Zu (ii): Aus der Optimalität folgt zusammen mit (i)

$$\begin{aligned} 0 &\leqslant B_{\alpha_{k+1}}(x^k) - B_{\alpha_{k+1}}(x^{k+1}) = f(x^k) - f(x^{k+1}) + \alpha_{k+1}\left(\beta(x^k) - \beta(x^{k+1})\right) \\ &\leqslant f(x^k) - f(x^{k+1}). \end{aligned}$$

□

Um die Konvergenz von Algorithmus 12.2 zeigen zu können, bedarf es einer Art Regularitätsbedingung. Dafür definieren wir das *strikte Innere*

$$X^\circ := \{x \in \mathbb{R}^n : g_i(x) < 0,\ 1 \leqslant i \leqslant m\}$$

der zulässigen Menge. Beachte, dass dies im Allgemeinen nur eine Teilmenge des topologischen Inneren sein muss! (Das strikte Innere hängt ja, im Gegensatz zum topologischen Innerern, von der konkreten Beschreibung von X durch die g_i ab.)

Satz 12.9. *Seien* $f : X \to \mathbb{R}$ *und* $g_i : \mathbb{R}^n \to \mathbb{R}$, $1 \leqslant i \leqslant m$, *stetig. Es gelte* $X^\circ \neq \emptyset$ *sowie* $\overline{X^\circ} = X$. *Erzeugt Algorithmus 12.2 für eine streng monoton fallende Nullfolge* $\{\alpha_k\}_{k\in\mathbb{N}} \subset (0, \infty)$ *eine unendliche Folge* $\{x^k\}_{k\in\mathbb{N}} \subset X$, *so ist jeder Häufungspunkt ein globaler Minimierer von* f *in* X.

Beweis. Sei $\{x^k\}_{k\in K} \subset X^\circ$ eine konvergente Teilfolge mit Grenzwert $\bar{x}$, der wegen $\overline{X^\circ} = X$ in X liegt. Angenommen, $\bar{x}$ wäre kein globaler Minimierer von f in X. Dann existiert ein $x \in X$ mit $f(x) < f(\bar{x})$. Wegen $\overline{X^\circ} = X$ und der Stetigkeit von f existiert daher ein $\hat{x} \in X^\circ$ nahe genug an x, so dass ebenfalls $f(\hat{x}) < f(\bar{x})$ gilt.

Aus Lemma 12.8 (i) und der Optimalität von x^k folgt nun

$$f(x^k) + \alpha_k \beta(x^0) \leqslant f(x^k) + \alpha_k \beta(x^k) \leqslant f(\hat{x}) + \alpha_k \beta(\hat{x})$$

für alle $k \geqslant 0$. Stetigkeit von f und $\alpha_k \to 0$ ergibt nun

$$f(\bar{x}) = \lim_{K \ni k \to \infty} f(x^k) \leqslant f(\hat{x}) + \lim_{K \ni k \to \infty} \alpha_k \left(\beta(\hat{x}) - \beta(x^0)\right) = f(\hat{x}),$$

im Widerspruch zu $f(\hat{x}) < f(\bar{x})$. Also ist $\bar{x} \in X$ ein globaler Minimierer. □

Für stetig differenzierbare g_i folgt aus $\frac{d}{dt} \ln(t) = \frac{1}{t}$ mit der Summen- und Kettenregel

$$\nabla B_\alpha(x) = \nabla f(x) - \alpha \sum_{i=1}^{m} \frac{\nabla g_i(x)}{g_i(x)}.$$

Vergleicht man dies wieder mit (11.7), so erhält man

$$\nabla B_\alpha(x) = \nabla_x L(x, \mu) \qquad \text{für} \quad \mu := -\frac{\alpha}{g(x)},$$

und in der Tat kann man zeigen, dass für $x^k \to \bar{x}$ die entsprechenden μ^k gegen den Lagrange-Multiplikator $\bar{\mu}$ konvergieren. Dies wird im *primal-dualen innere-Punkte-Verfahren* ausgenutzt. Dafür schreibt man die notwendige Optimalitätsbedingung $\nabla B_\alpha(x_\alpha) = 0$ um in

$$\begin{cases} \nabla_x L(x_\alpha, \mu_\alpha) = 0, \\ -(\mu_\alpha)_i g_i(x_\alpha) = \alpha, \quad 1 \leqslant i \leqslant m, \end{cases}$$

(vergleiche die KKT-Bedingungen (11.6) – die Komplementaritätsbedingungen $\bar{\mu}_i g_i(\bar{x}) = 1$ werden hier "von innen" angenähert). Auf dieses System wird nun ein Newton-Verfahren angewendet, wobei nach jedem Newton-Schritt der Penalty-Parameter α geeignet reduziert wird; eine Schrittweitenregel sorgt dabei dafür, dass die neuen Iterierten sich nicht zu weit von (x_α, μ_α) entfernen. Für konvexe Optimierungsprobleme haben sich diese Verfahren als sehr leistungsfähig erwiesen; siehe z. B. [Boyd und Vandenberghe 2004, Kapitel 11.7].

SQP-VERFAHREN

13

Eine der leistungsfähigsten und flexibelsten Klassen von Verfahren für Optimierungsprobleme mit Nebenbedingungen sind die sogenannten *sequential quadratic programming-Verfahren*, kurz *SQP-Verfahren*. Dabei handelt es sich um die Erweiterung von (Quasi-)Newton-Verfahren für unrestringierte Probleme auf Gleichungs- und Ungleichungsnebenbedingungen. Wir führen diese zuerst für den Fall reiner Gleichungsnebenbedingungen ein, und gehen dann kurz auf den Fall von gemischten Nebenbedingungen ein.

LAGRANGE–NEWTON-VERFAHREN FÜR GLEICHUNGSNEBENBEDINGUNGEN

Wir betrachten das Problem

$$\min_{x \in \mathbb{R}^n} f(x) \qquad \text{mit} \quad h(x) = 0$$

für zweimal stetig differenzierbare Funktionen $f : \mathbb{R}^n \to \mathbb{R}$ und $h : \mathbb{R}^n \to \mathbb{R}^p$. Gilt eine Regularitätsbedingung, so erfüllt ein lokaler Minimierer $\bar{x} \in \mathbb{R}^n$ zusammen mit einem Lagrange-Multiplikator $\bar{\lambda} \in \mathbb{R}^p$ die KKT-Bedingungen

$$\begin{cases} \nabla f(\bar{x}) + \bar{\lambda}^T \nabla h(\bar{x}) = 0, \\ h(\bar{x}) = 0. \end{cases}$$

Dies sind $n + p$ nichtlineare Gleichungen für die $n + p$ unbekannten Komponenten von $(\bar{x}, \bar{\lambda})$, die wir mit Hilfe der Lagrange-Funktion

$$L : \mathbb{R}^n \times \mathbb{R}^p \to \mathbb{R}, \qquad (x, \lambda) \mapsto f(x) + \lambda^T h(x),$$

schreiben können als

$$\nabla L(\bar{x}, \bar{\lambda}) = \begin{pmatrix} \nabla_x L(\bar{x}, \bar{\lambda}) \\ \nabla_\lambda L(\bar{x}, \bar{\lambda}) \end{pmatrix} = 0.$$

Auf diese Gleichung wenden wir nun das Newton-Verfahren an: Für gegebene (x^k, λ^k) berechnen wir $s^k \in \mathbb{R}^{n+p}$ als Lösung von

$$\nabla^2 L(x^k, \lambda^k) s = -\nabla L(x^k, \lambda^k),$$

und setzen (nach Zerlegung von $s^k \in \mathbb{R}^{n+p}$ in $s_x^k \in \mathbb{R}^n$ und $s_\lambda^k \in \mathbb{R}^p$)

$$x^{k+1} := x^k + s_x^k, \qquad \lambda^{k+1} := \lambda^k + s_\lambda^k.$$

Für die lokale superlineare Konvergenz müssen wir zunächst nachweisen, dass die Hesse-Matrix $\nabla^2 L$ in einem KKT-Punkt $(\bar{x}, \bar{\lambda})$ invertierbar ist. Da die Lagrange-Funktion linear ist in λ, hat diese stets die Form

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

Diese Struktur können wir ausnutzen, um hinreichende Bedingungen für die Invertierbarkeit anzugeben.

Lemma 13.1. *Seien $f : \mathbb{R}^n \to \mathbb{R}$ und $h : \mathbb{R}^n \to \mathbb{R}^p$ zweimal stetig differenzierbar. Gilt in $(x, \lambda) \in \mathbb{R}^n \times \mathbb{R}^p$:*

(i) $\nabla h(x) \in \mathbb{R}^{n \times p}$ *hat vollen Spaltenrang* p*;*

(ii) $d^T \nabla_{xx}^2 L(x, \lambda) d > 0$ *für alle* $d \in \mathbb{R}^n \setminus \{0\}$ *mit* $\nabla h(x)^T d = 0$*,*

so ist $\nabla^2 L(x, \lambda) \in \mathbb{R}^{(n+p)\times(n+p)}$ *invertierbar.*

Beweis. Da $\nabla^2 L(x, \lambda)$ quadratisch ist, genügt es zu zeigen, dass $\nabla^2 L(x, \lambda)$ injektiv ist. Seien daher (s_x, s_λ) mit

$$\begin{pmatrix} \nabla_{xx}^2 L(x, \lambda) & \nabla h(x) \\ \nabla h(x)^T & 0 \end{pmatrix} \begin{pmatrix} s_x \\ s_\lambda \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}.$$

Aus der zweiten Zeile folgt sofort $\nabla h(x)^T s_x = 0$. Multiplizieren der ersten Zeile von links mit s_x^T ergibt dann

$$0 = s_x^T \nabla_{xx}^2 L(x, \lambda) s_x + s_x^T \nabla h(x) s_\lambda = s_x^T \nabla_{xx}^2 L(x, \lambda) s_x.$$

Wegen Annahme (ii) und $\nabla h(x)^T s_x = 0$ ist dies aber nur möglich, falls $s_x = 0$ gilt. Damit reduziert sich die erste Zeile zu $\nabla h(x) s_\lambda = 0$, was aber nach Annahme (i) nur für $s_\lambda = 0$ gilt. Also ist $\nabla^2 L(x, \lambda)$ injektiv und deshalb invertierbar. □

Für einen KKT-Punkt $(\bar{x}, \bar{\lambda})$ entspricht Annahme (i) genau der LICQ, während Annahme (ii) die hinreichende Bedingung zweiter Ordnung ist. Analog zu Lemma 7.5 mit L an Stelle von f folgt daraus, dass $\nabla^2 L(x, \lambda)$ für (x, λ) in einer hinreichend kleinen Umgebung eines solchen KKT-Punktes ebenfalls invertierbar ist. Ebenso folgt daraus wie in Satz 8.1 die lokal superlineare Konvergenz des Lagrange–Newton-Verfahrens.

Satz 13.2. *Seien* $f : \mathbb{R}^n \to \mathbb{R}$ *und* $h : \mathbb{R}^n \to \mathbb{R}^p$ *zweimal stetig differenzierbar, und sei* $(\bar{x}, \bar{\lambda})$ *ein KKT-Punkt, in dem die LICQ und die hinreichende Optimalitätsbedingung zweiter Ordnung gilt. Dann existiert ein* $\delta > 0$ *so, dass für alle Startwerte* (x^0, λ^0) *mit*

$$\|x^0 - \bar{x}\| + \|\lambda^0 - \bar{\lambda}\| \leqslant \delta$$

das Lagrange–Newton-Verfahren Folgen $\{x^k\}_{k\in\mathbb{N}}$ *und* $\{\lambda^k\}_{k\in\mathbb{N}}$ *erzeugt, die superlinear gegen* $\bar{x}$ *bzw.* $\bar{\lambda}$ *konvergieren. Ist* $\nabla^2 L$ *lokal Lipschitz-stetig, so ist die Konvergenz sogar quadratisch.*

Wie im unrestringierten Fall hat dieses Verfahren den Nachteil der lokalen Konvergenz (die darüberhinaus auch einen Maximierer als Grenzwert haben kann), den man mit einer Globalisierung in den Griff bekommen möchte. Die Frage ist auch, wie man dieses Verfahren auf Ungleichungen anwenden kann. Dafür ist eine eine alternative Sichtweise auf das Verfahren nützlich.

SQP-VERFAHREN FÜR GEMISCHTE NEBENBEDINGUNGEN

Wir erinnern uns aus der Herleitung der Trust-Region-Verfahren für unrestringierte Verfahren, dass der Newton-Schritt äquivalent als Minimierer einer geeigneten quadratischen Funktion charakterisiert werden kann. Analog kann man für Gleichungsnebenbedingungen die Lösung (s_x^k, s_λ^k) des Lagrange–Newton-Schritts

$$\begin{pmatrix} \nabla^2_{xx} L(x^k, \lambda^k) & \nabla h(x^k) \\ \nabla h(x^k)^T & 0 \end{pmatrix} \begin{pmatrix} s_x \\ s_\lambda \end{pmatrix} = \begin{pmatrix} -\nabla_x L(x^k, \lambda^k) \\ -h(x^k) \end{pmatrix}. \tag{13.1}$$

über ein quadratisches Minimierungsproblem mit *linearen* Beschränkungen charakterisieren. Wir betrachten für $(x^k, \lambda^k) \in \mathbb{R}^n \times \mathbb{R}^p$ das Problem

$$\begin{cases} \min\limits_{s\in\mathbb{R}^n} \nabla f(x^k)^T s + \frac{1}{2} s^T \nabla^2_{xx} L(x^k, \lambda^k) s \\ \text{mit } \; h(x^k) + \nabla h(x^k)^T s = 0. \end{cases} \tag{13.2}$$

Da die Linearität der Nebenbedingungen eine Regularitätsbedingung darstellt (Satz 11.10), existiert für jeden Minimierer $\bar{s}$ ein Lagrange-Multiplikator $\bar{\lambda}_{lin} \in \mathbb{R}^p$, so dass die KKT-Bedingungen

$$\begin{cases} \nabla f(x^k) + \nabla^2_{xx} L(x^k, \lambda^k)\bar{s} + \nabla h(x^k)\bar{\lambda}_{lin} = 0, \\ h(x^k) + \nabla h(x^k)^T \bar{s} = 0, \end{cases}$$

erfüllt sind. Setzen wir $s_x^k := \bar{s}$ und $s_\lambda^k := \bar{\lambda}_{lin} - \lambda^k$ und bringen alle Terme, die kein s^k enthalten, auf die rechte Seite, so erhalten wir genau (13.1). Umgekehrt ist für einen Lagrange–Newton-Schritt (s_x^k, s_λ^k) das Paar $(s_x^k, \lambda^k + s_\lambda^k)$ ein KKT-Punkt von (13.2). Anstelle eines Updates berechnet man hier also direkt die neue Näherung für den Lagrange-Multiplikator.

Für Ungleichungsnebenbedingungen betrachtet man analog für $(x^k, \mu^k, \lambda^k) \in \mathbb{R}^n \times \mathbb{R}^m \times \mathbb{R}^p$ das Problem

$$\begin{cases} \min\limits_{s \in \mathbb{R}^n} \nabla f(x^k)^T s + \frac{1}{2} s^T \nabla^2_{xx} L(x^k, \mu^k, \lambda^k) s \\ \text{mit } g(x^k) + \nabla g(x^k)^T s \leqslant 0, \\ \qquad h(x^k) + \nabla h(x^k)^T s = 0, \end{cases} \tag{13.3}$$

berechnet einen KKT-Punkt $(\bar{s}, \bar{\mu}_{\text{lin}}, \bar{\lambda}_{\text{lin}})$, und setzt dann

$$x^{k+1} = x^k + \bar{s}, \qquad \mu^{k+1} = \bar{\mu}_{\text{lin}}, \qquad \lambda^{k+1} = \bar{\lambda}_{\text{lin}}.$$

Der Beweis der lokalen Konvergenz ist deutlich komplizierter als im Fall reiner Gleichungsbeschränkungen; wir beschränken uns hier auf die grobe Idee. Angenommen, zusätzlich zur Regularitätsbedingung gilt im KKT-Punkt $(\bar{x}, \bar{\mu}, \bar{\lambda})$ des nichtlinearen Problems die strikte Komplementarität, d. h. $\bar{\mu}_i = 0$ genau dann, wenn $g_i(\bar{x}) < 0$. Dann sind die KKT-Bedingungen äquivalent zum *Gleichungs*system

$$\begin{cases} \nabla_x L(\bar{x}, \bar{\mu}, \bar{\lambda}) = 0, & \\ \qquad h_j(\bar{x}) = 0, & 1 \leqslant j \leqslant p, \\ \qquad g_i(\bar{x}) = 0, & i \in \mathcal{A}_X(\bar{x}), \\ \qquad \bar{\mu}_i = 0, & i \notin \mathcal{A}_X(\bar{x}). \end{cases}$$

Auf dieses System wird nun das Newton-Verfahren angesetzt. Da die g_i stetig differenzierbar sind, ist $\mathcal{A}_X(x^k) = \mathcal{A}_X(\bar{x})$ für x^k nahe genug an $\bar{x}$; ebenso gilt $g_i(x^k) + \nabla g_i(x^k)^T(x^k - \bar{x}) < 0$ für alle $i \in \mathcal{A}_X(\bar{x})$ für x^k nahe genug an $\bar{x}$. In diesem Fall kann man wieder zeigen, dass der Newton-Schritt genau den KKT-Bedingungen für (13.3) entspricht. Schließlich zeigt man noch, dass die Newton-Matrix in $(\bar{x}, \bar{\mu}, \bar{\lambda})$ invertierbar ist. Startet man also in einer hinreichend kleinen Umgebung an $\bar{x}$, kann man die lokal superlineare Konvergenz des SQP-Verfahrens aus der des Newton-Verfahrens (für eine *feste* aktive Menge $\mathcal{A}_X(\bar{x})$ anstelle der aktiven Menge in x^k für die linearisierten Ungleichungen!) folgern; siehe [Geiger und Kanzow 2002, Satz 5.31].

Ähnlich knapp gehen wir auf weitere praktische Details ein:

- Analog zum unrestringierten Newton-Verfahren kann man das SQP-Verfahren durch eine Schrittweitensuche globalisieren. Dabei ist es nicht ausreichend, nur die Zielfunktion f zu betrachten; es müssen auch die Nebenbedingungen berücksichtigt werden. Dies kann durch Verwendung der exakten Straffunktion

  $$f(x) + \alpha \pi_1(x)$$

  für α hinreichend groß geschehen; dabei wird in der Praxis $\alpha = \alpha_k$ während des Verfahrens angepasst (etwa mit Hilfe von (12.4) für μ^k, λ^k anstelle von $\bar{\mu}, \bar{\lambda}$). Obwohl π_1 nicht differenzierbar ist, kann man eine Armijo-Bedingung mit Hilfe von Richtungsableitungen formulieren; siehe [Geiger und Kanzow 2002, Kapitel 5.5.4].

Um trotzdem lokal superlineare Konvergenz zu erhalten, muss irgendwann stets die Schrittweite $\sigma_k = 1$ akzeptiert werden. Für Probleme mit Gleichungsnebenbedingung kann es aber sein, dass das nie der Fall ist. Dies ist als *Maratos-Effekt* bekannt, und kann durch einen zusätzlichen Korrektur-Schritt behandelt werden; siehe [Geiger und Kanzow 2002, Kapitel 5.5.6].

Eine Alternative stellen Trust-Region-SQP-Verfahren dar, die in letzter Zeit vermehrt untersucht werden.

- Für die Lösung der SQP-Teilprobleme (13.3) kann man analog zum Konvergenzbeweis die KKT-Bedingung als Gleichungssystem schreiben. Da die linearisierte aktive Menge

$$\mathcal{A}_{\text{lin}}^k(s^k) := \left\{ i : g_i(x^k) + \nabla g_i(x^k)^T s^k = 0 \right\}$$

nicht bekannt ist, geht man iterativ vor: Man wählt eine Startschätzung $\mathcal{A}^0 \subset \mathcal{A}_{\text{lin}}^k(\bar{s})$, löst das entsprechende Gleichungssystem mit $\mathcal{A}^0$ anstelle von $\mathcal{A}_{\text{lin}}^k(\bar{s})$, und wählt (falls man nicht bereits einen KKT-Punkt zu (13.3) gefunden hat) eine geeignete neue Näherung $\mathcal{A}^1$. Dies wird als *aktive-Mengen-Strategie* bezeichnet; siehe [Geiger und Kanzow 2002, Kapitel 5.1.2].

- Schließlich kann auch in (13.3) anstelle der exakten Hesse-Matrix $\nabla_{xx}^2 L(x^k, \mu^k, \lambda^k)$ eine Quasi-Newton-Näherung H_k verwendet werden. Bei der Wahl des Updates sind dabei einige Schwierigkeiten zu beachten, um die positive Definitheit zu gewährleisten; siehe [Geiger und Kanzow 2002, Kapitel 5.5.5].

Teil IV

ABLEITUNGSFREIE VERFAHREN

DAS NELDER–MEAD-SIMPLEX-VERFAHREN

14

Zum Abschluss gehen wir noch kurz auf Verfahren ein, die keine Kenntnis der Ableitung der zu minimierenden Funktion voraussetzen, d. h. ausschließlich Funktionswerte in vorgegebenen oder geeignet gewählten Punkten verwenden. Solche Verfahren sind insbesondere in Anwendungen relevant, in denen ein komplexes System optimiert werden soll, für das kein hinreichend analysierbares mathematisches Modell existiert; insbesondere wenn es genügt, lediglich einen *besseren* Funktionswert als vorher bekannt zu finden.

Dabei unterscheidet man

- *modellbasierte Verfahren*, in denen die Funktionswerte benutzt werden, um – etwa durch Interpolation – ein lokales Modell zu erstellen, das dann mit Methoden aus Teil II oder Teil III minimiert wird, und
- *direkte Suchverfahren*, in denen die zulässige Menge mehr oder weniger systematisch durchprobiert und jeweils der beste Funktionswert gespeichert wird.

Letztere unterscheidet man weiter in

- *stochastische Suchverfahren*, bei denen die Auswahl der zu probierenden Kandidaten (zum Teil) auf Zufall beruht, und
- *deterministische Suchverfahren*, bei denen dies nicht der Fall ist.

Die Verfahren der ersten Klasse lehnen sich oft an physikalische oder biologische Prozesse an und haben klingende Namen wie *simulated annealing*, *genetische Algorithmen*, *particle swarm*- oder *firefly*-Verfahren. Eine Konvergenztheorie dafür existiert in der Regel nicht, abgesehen von Aussagen in der Form "Wenn man lange genug wartet, wird irgendwann einmal jeder Punkt durchprobiert." (Entsprechend spricht man auch von *heuristischen Verfahren*.) Für deterministische (und modellbasierte) Verfahren sind stärkere Aussagen möglich, wenn man annimmt, dass Ableitungen existieren (auch wenn sie nicht zur Verfügung stehen).

Wir betrachten hier beispielhaft das bekannteste deterministische Suchverfahren, das *Nelder–Mead-Simplex-Verfahren*.[1] Die Idee dabei ist, eine Funktion $f : \mathbb{R}^n \to \mathbb{R}$ in $n + 1$ Punkten

[1] nicht zu verwechseln mit dem Simplex-Verfahren der linearen Optimierung

$x_1, \ldots, x_{n+1} \in \mathbb{R}^n$ auszuwerten, die einen *Simplex*

$$S := \left\{ x \in \mathbb{R}^n : x = \sum_{i=1}^{n+1} \alpha_i x_i, \alpha_i \geqslant 0, \sum_{i=1}^{n+1} \alpha_i = 1 \right\}$$

bilden (daher der Name). Der Punkt mit dem größten Funktionswert wird nun durch einen neuen Punkt ausgetauscht, und zwar so, dass der dadurch entstehende neue Simplex (hoffentlich) in Richtung eines Minimierers wandert und sich um ihn zusammenzieht. Um den Schritt konkret zu beschreiben, nehmen wir an, dass die Punkte des aktuellen Simplex nach aufsteigendem Funktionswert angeordnet sind:

$$f(x_1) \leqslant f(x_2) \leqslant \cdots \leqslant f(x_{n+1}),$$

d. h. x_1 ist der aktuell beste Kandidat für einen Minimierer, und x_{n+1} ist der Punkt, der ausgetauscht werden soll. Für den Ersatzpunkt machen wir dabei den Ansatz

$$x(t) := \bar{x} + t(x_{n+1} - \bar{x}) \qquad \text{mit} \quad \bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i t \in \mathbb{R} \quad \text{und} \quad t \in \mathbb{R},$$

d. h. einen Punkt auf der Verbindungsgerade zwischen dem entfernten Punkt x_{n+1} und dem Schwerpunkt der restlichen Punkte. Konkret probieren wir folgende Punkte durch:

Algorithmus 14.1 : Nelder–Mead-Simplex-Schritt

```
1  Berechne f(x(−1))
2  if f(x_1) ⩽ f(x(−1)) < f(x_n) then        // neuer Punkt weder bester noch schlechtester
3  |  Ersetze x_{n+1} durch x(−1), return                          // Reflektionsschritt
4  else if f(x(−1)) < f(x_1) then                                  // neuer Punkt bester
5  |  Berechne f(x(−2))                                  // versuche Richtung weiter
6  |  if f(x(−2)) < f(x(−1)) then                          // neuer Punkt noch besser
7  |  |  Ersetze x_{n+1} durch x(−2), return                      // Expansionsschritt
8  |  else                                              // neuer Punkt wieder schlechter
9  |  |  Ersetze x_{n+1} durch x(−1), return                      // Reflektionsschritt
10 else if f(x(−1)) ⩾ f(x_n) then                          // neuer Punkt schlechtester
11 |  if f(x_n) ⩽ f(x(−1)) < f(x_{n+1}) then     // neuer Punkt zumindest besser als alter
12 |  |  Berechne f(x(−1/2))                          // versuche Richtung nicht so weit
13 |  |  if f(x(−1/2)) ⩽ f(x(−1)) then                    // neuer Punkt noch besser
14 |  |  |  Ersetze x_{n+1} durch x(−1/2), return      // äußerer Kontraktionsschritt
15 |  else                                  // neuer Punkt noch schlechter als alter
16 |  |  Berechne f(x(1/2))
17 |  |  if f(x(1/2)) < f(x_{n+1}) then                // neuer Punkt nicht mehr schlechter
18 |  |  |  Ersetze x_{n+1} durch x(1/2), return      // innerer Kontraktionsschritt
19 |  Ersetze x_i durch 1/2(x_1 + x_i), 2 ⩽ i ⩽ n + 1   // keine Verbesserung, Schrumpfschritt
```

Das Nelder–Mead-Simplex-Verfahren kann man als eine Art Finite-Differenzen-Version des Gradientenverfahrens auffassen. Dazu bedarf es etwas Notation. Für einen durch die (wieder nach aufsteigendem Funktionswert angeordneten) Punkte $x_1, \dots, x_{n+1} \in \mathbb{R}^n$ aufgespannten Simplex S sei $\Delta_S x \in \mathbb{R}^{n\times n}$ diejenige Matrix, deren Spalten durch die Richtungen $x_2 - x_1, \dots, x_{n+1} - x_1 \in \mathbb{R}^n$ gegeben sind. Dann gilt für die *Länge* von S

$$\sigma(S) := \max_{2\leqslant i\leqslant n+1} \|x_i - x_1\| \leqslant \|\Delta_S x\|.$$

Wir nennen S *singulär*, falls $\Delta_S x$ singulär ist. Andernfalls bezeichne $\kappa(S) := \kappa(\Delta_S x)$ die *Kondition* von S. Wir definieren weiter

$$\Delta_S f := (f(x_2) - f(x_1), \dots, f(x_{n+1}) - f(x_1))^T \in \mathbb{R}^n.$$

Der *Simplex-Gradient* ist dann gegeben durch

$$\nabla_S f := (\Delta_S x)^{-T} \Delta_S f \in \mathbb{R}^n.$$

Für Lipschitz-stetig differenzierbare Funktionen ist der Simplex-Gradient eine Näherung an den Gradienten, die umso besser ist, je kleiner der Simplex ist.

Lemma 14.1. *Sei $f : \mathbb{R}^n \to \mathbb{R}$ Lipschitz-stetig differenzierbar mit Lipschitz-Konstante L und sei S ein nicht-singulärer Simplex. Dann existiert eine Konstante $C > 0$ unabhängig von S mit*

$$\|\nabla_S f - \nabla f(x_1)\| \leqslant C\kappa(S)\sigma(S).$$

Beweis. Aus der Lipschitz-Stetigkeit von ∇f folgt ähnlich wie im Beweis von (7.5) aus Satz 1.3 für alle $2 \leqslant i \leqslant n+1$

$$|f(x_i) - f(x_1) - \nabla f(x_1)^T (x_i - x_1)| \leqslant \frac{L}{2}\|x_i - x_1\|^2 \leqslant \frac{L}{2}\sigma(S)^2.$$

Quadrieren, Summieren über alle i und Wurzelziehen ergibt dann

$$\|\Delta_S f - (\Delta_S x)^T \nabla f(x_1)\| \leqslant \sqrt{n}\frac{L}{2}\sigma(S)^2.$$

Daraus folgt

$$\begin{aligned}\|(\Delta_S x)^{-T}\Delta_S f - \nabla f(x_1)\| &\leqslant \|(\Delta_S x)^{-T}\|\|\Delta_S f - (\Delta_S x)^T \nabla f(x_1)\| \\ &\leqslant \|(\Delta_S x)^{-T}\|\sqrt{n}\frac{L}{2}\sigma(S)^2 \\ &\leqslant \sqrt{n}\frac{L}{2}\kappa(S)\sigma(S)\end{aligned}$$

wegen $\sigma(S) \leqslant \|\Delta_S x\|$ und $\|A^{-T}\|\|A\| = \|A^{-1}\|\|A\| = \kappa(A)$ für $A \in \mathbb{R}^{n\times n}$. Aus der Definition von $\nabla_S f$ folgt nun die Aussage für $C := \sqrt{n}\frac{L}{2}$. □

Sei nun $\{(x_1^k, \dots, x_{n+1}^k\}_{k\in\mathbb{N}}$ eine durch Algorithmus 14.1 erzeugte Folge von Punkten und $\{S^k\}_{k\in\mathbb{N}}$ die Folge der entsprechenden Simplices. Zwar verbessert Algorithmus 14.1 in der Regel nicht in jedem Schritt den zur Zeit kleinsten Funktionswert $f(x_1^k)$; da jedoch in jeder Iteration – außer im Fall eines Schrumpfschrittes – *einer* der $n+1$ Funktionswerte verringert wird, gilt dies zumindest für den durchschnittlichen Funktionswert

$$\bar{f}^k := \frac{1}{n+1} \sum_{i=1}^{n+1} f(x_i^k).$$

Ist der Abstieg hinreichend groß, und bleibt die Kondition der Simplices beschränkt, so konvergiert das Nelder–Mead-Simplex-Verfahren.

Satz 14.2. *Sei $f : \mathbb{R}^n \to \mathbb{R}$ Lipschitz-stetig differenzierbar und nach unten beschränkt. Gilt für die durch Algorithmus 14.1 erzeugten Folgen*

(i) S^k *ist nicht-singulär für alle* $k \in \mathbb{N}$,

(ii) $\lim_{k\to\infty} \sigma(S^k)\kappa(S^k) = 0$,

(iii) $\bar{f}^{k+1} - \bar{f}^k < -\alpha\|\nabla_{S^k} f\|^2$ *für ein* $\alpha > 0$ *und alle bis auf endlich viele* $k \in \mathbb{N}$,

so ist jeder Häufungspunkt von $\{x_1^k\}_{k\in\mathbb{N}}$ *ein stationärer Punkt von* f.

Beweis. Mit f ist auch $\{\bar{f}^k\}_{k\in\mathbb{N}}$ nach unten beschränkt und – wegen der Abstiegsbedingung (iii) – strikt monoton fallend. Also konvergiert diese Folge, und aus (iii) folgt weiter

$$\|\nabla_{S^k} f\|^2 < \alpha^{-1} \left(\bar{f}^k - \bar{f}^{k+1}\right) \to 0.$$

Aus Lemma 14.1 folgt nun zusammen mit der umgekehrten Dreiecksungleichung und der Beschränktheitsbedingung (ii)

$$\|\nabla f(x_1^k)\| \leqslant C\kappa(S^k)\sigma(S^k) + \|\nabla_{S^k} f\| \to 0. \tag{14.1}$$

Sei nun $\{x_1^k\}_{k\in K}$ eine konvergente Teilfolge mit Grenzwert $\tilde{x}_1$. Dann folgt aus (14.1) und der Stetigkeit von ∇f

$$\|\nabla f(\tilde{x}_1)\| = \lim_{K\ni k\to\infty} \|\nabla f(x_1^k)\| = 0,$$

was zu zeigen war. □

Keine der Annahmen kann in der Praxis garantiert werden. Man wird daher bei (vermuteter) Verletzung einer der Bedingungen das Nelder–Mead-Verfahren mit einer geeigneten Wahl von Punkten $(\tilde{x}_1, \tilde{x}_2, \dots, \tilde{x}_n)$ mit $\tilde{x}_1 = x_1^k$ neu starten; siehe z. B. [Kelley 1999, Kapitel 8.1.4]. Dafür bleibt die Aussage von Satz 14.2 auch dann noch gültig, wenn die Funktionswerte $f(x_i)$ nur bis auf (kleine) Fehler bekannt sind; siehe z. B. [Kelley 1999, Satz 8.1.2].

LITERATUR

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

S. Boyd und L. Vandenberghe (2004). *Convex optimization*. Cambridge University Press, Cambridge. DOI: 10.1017/CBO9780511804441.

C. Clason (2014). „Optimierung I“. Vorlesungsskript, Fakultät für Mathematik, Universität Duisburg-Essen. URL: https://www.uni-due.de/~adf040p/skripte/Optim1Skript14.pdf.

J. Dennis und R. Schnabel (1996). *Numerical methods for unconstrained optimization and nonlinear equations*. Bd. 16. Classics in Applied Mathematics. Society for Industrial und Applied Mathematics. DOI: 10.1137/1.9781611971200.

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

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

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 und Applied Mathematics (SIAM), Philadelphia, PA. DOI: 10.1137/1.9781611970920.

P. Spellucci (1993). *Numerische Verfahren der nichtlinearen Optimierung*. Birkhäuser Verlag, Basel. DOI: 10.1007/978-3-0348-7214-0.

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