Numerische Methoden der nichtlinearen Optimierung

Prof. Dr. Andreas Frommer

Sommersemester 2004

Bergische Universität Wuppertal
Fachbereich Mathematik

Inhaltsverzeichnis

Dank

Herr Dipl. Math. Stefan Borovac hat mehrere Versionen dieses Skriptes kritisch gelesen und an unzähligen Stellen zu Verbesserungen beigetragen. Frau Brigitte Schultz hat meine nicht immer leicht zu lesenden Vorlagen wie immer zuverlässig in LaTeX umgesetzt. Beiden möchte ich an dieser Stelle ganz herzlich danken.

Wuppertal, 17. April 2005 Andreas Frommer

Literatur

[1] Ortega, J., Rheinboldt, W.; Iterative Solution of Nonlinear Equation in Sveral Variables, Academic Press, New York (1970)

[2] Dennis Jr., J., Schnabel, R.: Numerical Methods for Unconstrained Optimization and Nonlinear Equations, Prentice Hall, Englewood Cliffs (1983), wiederaufgelegt bei SIAM, Philadelphia (1996)

[3] Geiger, C., Kanzow, C.: Numerische Verfahren zur Lösung unrestringierter Optimierungsaufgaben, Springer, Heidelberg (1999)

Kapitel 1

Vorbereitung

Abschnitt 1.1

Einleitung

In dieser Veranstaltung geht es um die Analyse und um Verfahren zur Lösung der nicht-restringierten Minimierungsaufgabe

(1.1.1) finde alle x^* mit $f(x^*) = \min\{f(x) : x \in \mathbb{R}^n\}$.

Dabei ist $f : \mathbb{R}^n \to \mathbb{R}$ eine nichtlineare Funktion, von der wir in der Regel wenigstens Differenzierbarkeit voraussetzen. Die Aufgabe (1.1.1) nennt man entsprechend *nicht-restringiertes nichtlineares Optimierungsproblem*. Zu seiner Behandlung gibt es spezifische Verfahren die Gegenstand dieser Veranstaltung sind.
Zur Abgrenzung: Wird in (1.1.1) nur $x \in D \subseteq \mathbb{R}^n$ zugelassen, so spricht man von einer *restringierten* Optimierungsaufgabe. Ist f eine affine Funktion, $f(x) = c^T x + \gamma$ und wird D durch ein System von linearen Gleichungen und Ungleichungen beschrieben, so spricht man von einem linearen Optimierungsproblem; sind f und D konvex von einer *konvexen* Optimierungsaufgabe. In diesen Fällen kann man die zusätzliche Struktur der Aufgabe bei der Entwicklung geeigneter Verfahren ausnutzen. Sie sind nicht Gegenstand dieser Veranstaltung.

1.1.1 Definition

(i) Jede Lösung x^* von (1.1.1) heißt *globale Minimalstelle* von f; der Wert $f(x^*)$ heißt *globales Minimum* von f.

(ii) Ein Punkt $x^* \in \mathbb{R}^n$ heißt *lokale Minimalstelle* von f und der Wert $f(x^*)$ *lokales Minimum*, falls gilt

$$f(x^*) \leq f(x) \quad \text{für alle } x \text{ in einer Umgebung von } x^*.$$

(iii) Ein globales Minimum heißt *streng* und die zugehörige Minimalstelle x^* eine *strenge globale Minimalstelle*, wenn x^* eindeutig ist.

(iv) Eine lokale Minimalstelle x^* von f heißt *strenge lokale Minimalstelle*, falls gilt

$$f(x^*) < f(x) \quad \text{für alle } x \neq x^* \text{ in einer Umgebung von } x^*.$$

1.1.2 Beispiel

(i) Die Funktion $f(x) = x^2$ besitzt die strenge globale Mimimalstelle $x^* = 0$ und keine weiteren lokalen Minimalstellen.

(ii) Die Funktion $f(x) = \max\{0, x^2 - 1\}$ besitzt als globale Mimimalstellen das ganze Intervall $[-1, 1]$. Sie hat keine strengen Minimalstellen und keine weiteren lokalen Minimalstellen.

(iii) Die Funktion $f(x) = -e^{-x^2}\cos(x)$ besitzt eine strenge globale Minimalstelle bei $x = 0$ und unendlich viele weitere strenge lokale Minimalstellen.

Abschnitt 1.2

Grundlagen aus der Analysis

Konventionen zur Notation: Im ganzen Skript wird f stets eine Funktion von $D \subseteq \mathbb{R}^n$ nach $\mathbb{R}$ bezeichnen. Groß geschriebene Symbole (z.B. F) reservieren wir für Funktionen von D nach $\mathbb{R}^m$. Vektoren sind prinzipiell Spaltenvektoren mit Ausnahme des Gradienten

$$\nabla f(x) = \left(\frac{\partial f(x)}{\partial x_1}, \ldots, \frac{\partial f(x)}{\partial x_n} \right),$$

den wir stets als Zeilenvektor auffassen. Dies hat den Vorteil, dass wir Innenprodukte des Gradienten mit einem Vektor d immer einfach als $\nabla f(x)d$ notieren können.
Unser Differenzierbarkeitsbegriff ist stets die Fréchet-Differenzierbarkeit (totale Differenzierbarkeit).

1.2.1 Definition
$f : D \subseteq \mathbb{R}^n \to \mathbb{R}$ heißt in dem Punkt x im Inneren $\overset{\circ}{D}$ von D differenzierbar, wenn es einen Zeilenvektor $df(x)$ gibt mit der Eigenschaft

$$|f(x+h) - f(x) - df(x)h| = o(\|h\|).$$

Aus der Analysis wissen wir:

1.2.2 Satz
Ist $f : D \subseteq \mathbb{R}^n \to \mathbb{R}$ in $x \in \overset{\circ}{D}$ differenzierbar, so existieren die partiellen Ableitungen und es gilt

$$df(x) = \nabla f(x).$$

Ist umgekehrt $\nabla f(x)$ stetig in x, so ist f differenzierbar in x. Wir schreiben $f \in \mathcal{C}^1(D)$, wenn $\nabla f(x)$ stetig ist in allen Punkten $x \in D$.

1.2.3 Definition
$f : D \subseteq \mathbb{R}^n \to \mathbb{R}$ sei differenzierbar in $x \in \overset{\circ}{D}$. Dann heißt x *stationärer Punkt* von f, falls $\nabla f(x) = 0$.

1.2.4 Satz
$f : D \subseteq \mathbb{R}^n \to \mathbb{R}$ sei differenzierbar in einer (lokalen) Minimalstelle $x \in \overset{\circ}{D}$. Dann ist x stationärer Punkt von f.

Definiert man für festes $x, d \in \mathbb{R}^n$ die Funktion $\varphi : \mathbb{R} \to \mathbb{R}, \varphi(t) = f(x+td)$, so ist $\varphi'(t) = \nabla f(x+td)d$, sofern f an der Stelle $x+td$ differenzierbar ist. Mit dem (eindimensionalen) Mittelwertsatz für φ erhält man so die folgenden Mittelwertsätze im Mehrdimensionalen, die sich in der Zukunft noch als sehr nützlich erweisen werden.

1.2.5 Lemma
Sei $f : D \subseteq \mathbb{R}^n \to \mathbb{R}$, und D enthalte die gesamte Verbindungsstrecke $x + \theta(y-x),\ \theta \in [0,1]$ zwischen zwei Punkten x und $y \in D$ im Inneren. f sei an allen diesen Stellen $x + \theta(y-x)$ differenzierbar. Dann gilt

(i) (Mittelwertsatz)

$$f(y) - f(x) = \nabla f(\xi)(y-x) \text{ mit } \xi = x + \theta(y-x),\ \theta \in (0,1),$$

(ii) (Integralform des Mittelwertsatzes)

$$f(y) - f(x) = \int_0^1 \nabla f(x + t(y-x)) \cdot (y-x) dt.$$

Man beachte, dass sich nur die Integralform auf Funktionen $F : D \to \mathbb{R}^m$ übertragen lässt.
Ist f zweimal differenzierbar, so schreiben wir $\nabla^2 f(x)$ für die Hesse-Matrix

$$\nabla^2 f(x) = \begin{pmatrix} \frac{\partial^2 f(x)}{\partial x_1 \partial x_1} & \cdots & \frac{\partial^2 f(x)}{\partial x_1 \partial x_n} \\ \vdots & & \vdots \\ \frac{\partial^2 f(x)}{\partial x_n \partial x_1} & & \frac{\partial^2 f(x)}{\partial x_n \partial x_n} \end{pmatrix}.$$

Die Hesse-Matrix ist symmetrisch, wenn die zweiten partiellen Ableitungen alle stetig in x sind. Trift dies für alle $x \in D$ zu, so schreiben wir $f \in \mathcal{C}^2(D)$.

1.2.6 Satz
Sei $f : D \subseteq \mathbb{R}^n \to \mathbb{R}$, $f \in \mathcal{C}^2(D)$ und $x^* \in \overset{\circ}{D}$ ein stationärer Punkt von f. Dann gilt

(i) Ist $\nabla^2 f(x^*)$ positiv definit, so ist x^* strikte lokale Minimalstelle von f.

(ii) Ist x^* lokale Minimalstelle von f, so ist $\nabla^2 f(x^*)$ positiv semidefinit.

Man beachte, dass aus Teil (ii) auch folgt: Ist $\nabla^2 f(x^*)$ indefinit an einem stationären Punkt x^*, so ist x^* keine lokale Minimalstelle.
Wir beenden diesen Abschnitt mit einem sehr nützlichen Resultat über die Norm von Inversen.

1.2.7 Lemma (Banach-Lemma)
Seien $A, H \in \mathbb{R}^n$. Es sei A regulär und $\|A^{-1}H\| < 1$. Dann ist auch $A - H$ regulär und es gilt

$$\|(A-H)^{-1}\| \leq \frac{\|A^{-1}\|}{1 - \|A^{-1}H\|}.$$

Beweis: Es ist

$$A - H = A(I - A^{-1}H).$$

Wegen $\|A^{-1}H\| < 1$ konvergiert die Neumann-Reihe

$$\sum_{k=0}^{\infty} (A^{-1}H)^k$$

gegen $(I - A^{-1}H)^{-1}$, und es gilt

$$\|(I - A^{-1}H)^{-1}\| \leq \sum_{k=0}^{\infty} (\|A^{-1}H\|)^k = \frac{1}{1 - \|A^{-1}H\|}.$$

□

Abschnitt 1.3

Konvexe Funktionen

Konvexe Funktionen sind in der Optimierung besonders wichtig, denn hier ist $\nabla f(x^*) = 0$ äquivalent zu 'x^* ist globale Minimalstelle', s. Satz 1.3.12 weiter unten. Außerdem ist die Struktur der Menge aller globalen Minimalstellen besonders einfach.

1.3.1 Definition
$D \subseteq \mathbb{R}^n$ heißt *konvex*, falls gilt

$$x, y \in D,\ \lambda \in [0,1] \Rightarrow \lambda x + (1-\lambda)y \in D.$$

1.3.2 Definition
$D \subseteq \mathbb{R}^n$ sei konvex, $f : D \to \mathbb{R}$.

(i) f heißt *(strikt) konvex*, falls gilt

$$x, y \in D, x \neq y, \lambda \in (0,1) \Rightarrow f(\lambda x + (1-\lambda)y) \leq (<)\, \lambda f(x) + (1-\lambda) f(y).$$

(ii) f heißt *gleichmäßig konvex*, falls $\mu > 0$ existiert mit

$$\begin{aligned} &x, y \in D, x \neq y, \lambda \in (0,1) \\ &\quad \Rightarrow \quad f(\lambda x + (1-\lambda)y) + \mu \cdot \lambda(1-\lambda)\|x-y\|^2 \leq \lambda f(x) + (1-\lambda) f(y). \end{aligned}$$

Klar: f glm. konvex $\Rightarrow$ f strikt konvex $\Rightarrow$ f konvex.

1.3.3 Beispiel
(i) $f(x) = x^2$ ist glm. konvex (mit $\mu = 1$).

(ii) $f(x) = x^4$ ist strikt konvex, aber nicht gleichmäßig konvex. (Begründung s. Satz 1.3.7).

(iii) $f(x) = \max\{x^4 - 1, 0\}$ ist konvex aber nicht strikt konvex.

Konvexität ist bei differenzierbaren Funktionen dadurch charakterisiert, dass die Funktion oberhalb jeder Tangentialhyperebene verläuft:

1.3.4 Satz
$D \subseteq \mathbb{R}^n$ sei konvex und offen, $f \in \mathcal{C}^1(D)$. Dann gilt

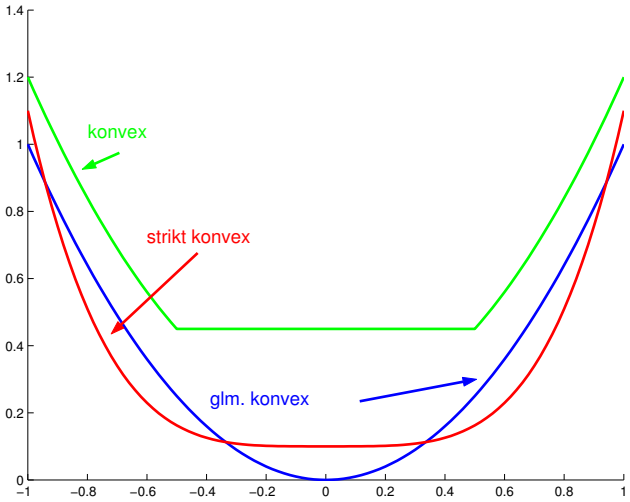

Abbildung 1.1: geometrische Interpretation der Konvexität

(i) f ist konvex auf D genau dann wenn für alle $x \neq y \in D$ gilt:

$$f(x) - f(y) \geq \nabla f(y)(x - y).$$

(ii) f ist strikt konvex auf D genau dann wenn für alle $x \neq y \in D$ gilt:

$$f(x) - f(y) > \nabla f(y)(x - y).$$

(iii) f ist glm. konvex auf D mit Konstante μ genau dann wenn für alle $x, y \in D$ gilt:

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

Beweis: Wir beweisen zuerst (iii).
„$\Leftarrow$“: Setze $z = \lambda x + (1 - \lambda)y$ für ein $\lambda \in (0, 1)$. Wir haben

$$\begin{aligned} f(x) - f(z) &\geq \nabla f(z)(x - z) + \mu \|x - z\|^2, \\ f(y) - f(z) &\geq \nabla f(z)(y - z) + \mu \|y - z\|^2. \end{aligned}$$

Multiplikation der ersten Ungleichung mit λ, der zweiten mit $1 - \lambda$ und anschließende Addition ergibt

$$\lambda f(x) + (1 - \lambda) f(y) - f(z) \geq \mu \underbrace{\left[(1 - \lambda)^2 \lambda + \lambda^2 (1 - \lambda)\right]}_{= \lambda(\lambda - 1)} \|x - y\|^2.$$

Also ist f glm. konvex.

„⇒“: Da f gleichmäßig konvex ist, gilt

$$\begin{aligned}\frac{f(y+\lambda(x-y))-f(y)}{\lambda} &\leq \frac{\lambda f(x)+(1-\lambda)f(y)-f(y)-\lambda(1-\lambda)\cdot\mu\cdot\|x-y\|^2}{\lambda}\\ &= f(x)-f(y)-(1-\lambda)\cdot\mu\cdot\|x-y\|^2.\end{aligned}$$

Für $\lambda \to 0$ erhalten wir so

$$\nabla f(y)(x-y) \;\leq\; f(x)-f(y)-\mu\cdot\|x-y\|^2.$$

Teil (i) wird wie (iii) bewiesen mit $\mu = 0$. Zum Beweis von (ii) geht „⇐“ ebenfalls wie für (iii).
„⇒“ muss für (ii) anders bewisen werden, da die strenge Ungleichung beim Grenzübergang „verloren“ geht. Sei also $x \neq y$ und $z = (1/2)\cdot x + (1/2)\cdot y$. Da f konvex ist, gilt nach (i)

$$f(z)-f(y) \geq \nabla f(y)(z-y) = \frac{1}{2}\nabla f(y)(x-y).$$

Es ist aber $f(z) < (1/2)\cdot f(x) + (1/2)\cdot f(y)$, also

$$\frac{1}{2}f(x)-\frac{1}{2}f(y) > \frac{1}{2}\nabla f(y)(x-y).$$

□

Im eindimensionalen sind differenzierbare Funktionen genau dann konvex, wenn f' monoton wächst. Bei geeigneter Begriffsbildung überträgt sich dies auf den $\mathbb{R}^n$.

1.3.5 Definition
Eine Abbildung $F : D \subseteq \mathbb{R}^n \to \mathbb{R}^n$ heißt

(i) *(strikt) monoton*, falls für alle $x \neq y \in D$ gilt

$$(F(y)-F(x))^T\cdot(y-x) \geq\; (>)0,$$

(ii) *gleichmäßig monoton mit Modulus* $\mu > 0$, falls

$$(F(y)-F(x))^T\cdot(y-x) \geq \mu\|x-y\|^2.$$

Klar: F glm. monoton ⇒ F strikt monoton ⇒ F monoton.

1.3.6 Satz
Sei $D \subseteq \mathbb{R}^n$ offen und konvex, $f : D \to \mathbb{R}$, $f \in \mathcal{C}^1(D)$. Dann gilt:

(i) f konvex $\Longleftrightarrow \nabla f$ ist monoton.

(ii) f strikt konvex $\Longleftrightarrow \nabla f$ ist strikt monoton.

(iii) f glm. konvex mit Konstante $\mu \Longleftrightarrow \nabla f$ ist glm. monoton mit Modulus 2μ.

Beweis: zu (iii), „$\Rightarrow$“: Nach Addition der beiden Ungleichungen

$$\begin{aligned} f(x) - f(y) &\geq \nabla f(y)(x-y) + \mu \cdot \|x-y\|^2 \\ f(y) - f(x) &\geq \nabla f(x)(y-x) + \mu \cdot \|x-y\|^2 \end{aligned}$$

erhalten wir

$$0 \geq (\nabla f(y) - \nabla f(x)) \cdot (x-y) + 2\mu \cdot \|x-y\|^2.$$

(ii) „$\Rightarrow$“ und (i) „$\Rightarrow$“ gehen analog.
(i) „$\Leftarrow$“: Nach dem Mittelwertsatz (Lemma 1.2.5 ist

$$f(x) - f(y) = \nabla f(\xi)(x-y), \;\; \xi = \theta x + (1-\theta)y, \theta \in (0,1).$$

da ∇f monoton ist, folgt

$$(\nabla f(\xi) - \nabla f(y)) \cdot (\xi - y) = \theta \, (\nabla f(\xi) - \nabla f(y)) \cdot (x-y) \geq 0,$$

also

$$f(x) - f(y) = \nabla f(\xi)(x-y) \geq \nabla f(y)(x-y).$$

(ii) „$\Leftarrow$“ geht genauso.
(iii) „$\Leftarrow$“ ist etwas aufwendiger: Sei $m \in \mathbb{N}$ fest und $x_k = y + \frac{k}{m}(x-y), k = 0, 1, \ldots, m$. Dann ist

$$\begin{aligned} f(y) - f(x) &= \sum_{k=0}^{m-1} f(x_k) - f(x_{k+1}) \\ &\overset{MWS}{=} \sum_{k=0}^{m-1} \nabla f(\xi_k)(\underbrace{x_k - x_{k+1}}_{=(1/m)(y-x)}), \; \xi_k = y + \theta_k(x-y), \theta_k \in \left(\frac{k}{m}, \frac{k+1}{m}\right) \\ &= \frac{1}{m} \sum_{k=0}^{m-1} \nabla f(\xi_k) \cdot \frac{1}{\theta_k} \cdot (y - \xi_k) \\ &\leq \frac{1}{m} \sum_{k=0}^{m-1} \frac{1}{\theta_k} \cdot \left(\nabla f(y)(y-\xi_k) - 2\mu \cdot \|y - \xi_k\|^2\right) \end{aligned}$$

$$
\begin{aligned}
&= \nabla f(y)(y-x) - 2\mu \underbrace{\sum_{k=0}^{m-1} \frac{\theta_k}{m}}_{\geq \sum_{k=0}^{m-1} \frac{k}{m^2}} \cdot \|y-x\|^2 \\
&\leq \nabla f(y)(y-x) - \mu \cdot \frac{m-1}{m} \cdot \|y-x\|^2.
\end{aligned}
$$

Für $m \to \infty$ ergibt sich schließlich

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

□

Im Falle $f \in \mathcal{C}^2(D)$ sind konvexe Funktionen vollständig durch die Hesse-Matrix charakterisierbar.

1.3.7 Satz
Sei $D \subseteq \mathbb{R}^n$ offen und konvex, $f : D \to \mathbb{R}$, $f \in \mathcal{C}^2(D)$. Dann gilt:

(i) f konvex $\iff \nabla^2 f(x)$ ist positiv semidefinit für alle $x \in D$.

(ii) f strikt konvex $\iff \nabla^2 f(x)$ ist positiv definit für alle $x \in D$.

(iii) f glm. konvex mit Konstante $\mu \iff \nabla^2 f(x)$ ist positiv definit mit $d^T \nabla^2 f(x) d \geq 2\mu d^T d$ für alle $d \in \mathbb{R}^n$ und $x \in D$.

Beweis: Zu (iii), „⇒“: Nach Satz 1.3.6 ist f glm. konvex mit Konstante μ, genau dann wenn ∇f glm. monoton mit Modulus 2μ. Wir haben deshalb

$$
\begin{aligned}
d^T \nabla^2 f(x) d &= d^T \lim_{t \to 0} \frac{1}{t} \cdot (\nabla f(x+td) - \nabla f(x))^T \\
&= \lim_{t \to 0} \frac{1}{t^2} \left((t \cdot d)^T (\nabla f(x+td) - \nabla f(x) \right)^T \\
&\geq \lim_{t \to 0} \frac{1}{t^2} 2\mu \|t \cdot d\|^2 \\
&= 2\mu \cdot \|d\|^2.
\end{aligned}
$$

Zu (iii) „⇐“: Nach dem Mittelwertsatz 1.2.5 gilt

$$
\begin{aligned}
(x-y)^T (\nabla f(x) - \nabla f(y))^T &= \int_0^1 (x-y)^T \nabla^2 f(x + t(y-x))(y-x) dt \\
&\geq \int_0^1 2\mu \|y-x\|^2 dt
\end{aligned}
$$

$$= \quad 2\mu\|y-x\|^2.$$

Also ist f glm. konvex mit Konstante μ nach Satz 1.3.6.
Zu (i): Hier folgen beide Richtungen aus (iii) mit $\mu = 0$.
Zu (ii): Auch hier geht der Beweis analog zu (iii). □

Nach diesen Charakterisierungen konvexer Funktionen wenden wir uns jetzt wieder der Minimierungsaufgabe zu, ausnahmsweise sogar z.T. mit Restriktionen.

1.3.8 Satz
Sei $f : \mathbb{R}^n \to \mathbb{R}$, $f \in \mathcal{C}^1(\mathbb{R}^n)$, $D \subseteq \mathbb{R}^n$ konvex. Wir betrachten die restringierte Minimierungsaufgabe

$$\min_{x \in D} f(x).$$

Die Menge aller Minimalstellen heiße L. Dann gilt

(i) L ist konvex (evtl. $L = \emptyset$).

(ii) f strinkt konvex auf $D \Rightarrow |L| \leq 1$.

(iii) f glm. konvex und D nicht leer und abgeschlossen $\Rightarrow |L| = 1$, d.h. es existiert eine eindeutige Lösung.

Beweis: Zu (i): $x_1, x_2 \in L \Rightarrow f^* = f(x_1) = f(x_2) \leq f(\lambda x_1 + (1-\lambda)x_2) \leq \lambda_1 f(x_1) + (1-\lambda) f(x_2) = f^*$ für alle $\lambda \in (0,1)$. Also ist $\lambda x_1 + (1-\lambda)x_2 \in L$.
Zu (ii): Angenommen, $x_1 \neq x_2 \in L$ und $f^* = f(x_1) = f(x_2)$. Dann gilt wegen der strikten Konvexität

$$f^* \leq f\left(\frac{1}{2}x_1 + \frac{1}{2}x_2\right) < \frac{1}{2}f(x_1) + \frac{1}{2}f(x_2) = f^*,$$

ein Widerspruch.
Zu (iii): Die Eindeutigkeit folgt aus (ii). Zum Nachweis der Existenz sei $x^0 \in D$ beliebig. Für alle $y \in D$ gilt dann nach Satz 1.3.4 (iii)

$$\begin{aligned} f(y) \quad &\geq \quad f(x^0) + \nabla f(x^0)(y-x^0) + 2\mu \cdot \|y-x^0\|^2 \\ &\overset{CSU}{\geq} \quad f(x^0) - \|\nabla f(x^0)\| \cdot \|y-x^0\| + 2\mu \cdot \|y-x^0\|^2. \end{aligned}$$

Für $\|y-x^0\| > r = \|\nabla f(x^0)\|/(2\mu)$ ist also $f(y) > f(x^0)$. Wir können die Minimierungsaufgabe deshalb auf den kompakten Bereich $D \cap \{y \in \mathbb{R}^n,\ \|y - x^0\| \leq r\}$ einschränken, auf welchem die stetige Funktion f ihr Minimum annimmt. □

1.3.9 Definition
Sei $f : D \subseteq \mathbb{R}^n \to \mathbb{R}$, $x^0 \in D$. Dann heißt die Menge

$$\mathcal{L}(x^0) = \{x \in D,\ f(x) \leq f(x^0)\}$$

die *Levelmenge von f bzgl. x^0*.

Soeben haben wir im Beweis von Satz 1.3.8 gezeigt:

1.3.10 Satz
Sei $f : \mathbb{R}^n \to \mathbb{R}$, $f \in \mathcal{C}^1(\mathbb{R}^n)$, $x^0 \in \mathbb{R}^n$. Weiter sei $\mathcal{L}(x^0)$ konvex und f glm. konvex auf $\mathcal{L}(x^0)$. Dann ist $\mathcal{L}(x^0)$ kompakt.

Für später halten wir noch fest:

1.3.11 Satz
Unter den Voraussetzungen von Satz 1.3.10 sei x^* die nach Satz 1.3.8(iii) eindeutige Minimalstelle von f auf $\mathcal{L}(x^0)$. Dann gilt für alle $x \in \mathcal{L}(x^0)$

$$f(x) \geq f(x^*) + \mu \cdot \|x - x^*\|^2$$

(μ aus glm. Konvexität).

Beweis: Auf Grund von Satz 1.2.4 ist $\nabla f(x^*) = 0$, wegen Satz 1.3.4(iii) gilt also

$$f(x) - f(x^*) \geq \nabla f(x^*)(x - x^*) + \mu \cdot \|x - x^*\|^2 = \mu \cdot \|x - x^*\|^2.$$

□

Schließlich haben wir noch das folgende, besonders eingängliche Resultat.

1.3.12 Satz
Sei $f : \mathbb{R}^n \to \mathbb{R}$, $f \in \mathcal{C}^1(\mathbb{R}^n)$ sei konvex auf ganz $\mathbb{R}^n$. Dann gilt

$$\nabla f(x^*) = 0 \Rightarrow x^* \text{ ist globale Minimalstelle von } f.$$

Beweis: Für alle $x \in \mathbb{R}^n$ ist nach Satz 1.3.4(i)

$$f(x) - f(x^*) \geq \nabla f(x^*)(x - x^*) = 0.$$

□

Kapitel 2

Abstiegsverfahren

Abschnitt 2.1

Allgemeine Formulierung und Konvergenz

Naheliegende Idee: Wähle vom aktuellen Punkt x eine Richtung d aus, in welcher f (zunächst) abnimmt. Gehe einen geeignet langen Schritt in diese Richtung.

2.1.1 Definition
Sei $f : \mathbb{R}^n \to \mathbb{R}, x \in \mathbb{R}^n$. Ein Vektor $d \in \mathbb{R}^n$ heißt *Abstiegsrichtung* (für f in x), falls $\overline{t} > 0$ existiert mit

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

Aus der Analysis ist folgendes Resultat bekannt.

2.1.2 Satz
Sei $f : \mathbb{R}^n \to \mathbb{R}$, $f \in \mathcal{C}^1(\mathbb{R}^n)$, $x \in \mathbb{R}^n$. Dann ist $d \in \mathbb{R}^n$ genau dann eine Abstiegsrichtung (für f in x), wenn

$$\nabla f(x) \cdot d < 0.$$

2.1.3 Beispiel
Ist $\nabla f(x) \neq 0$, $B \in \mathbb{R}^{n \times n}$ spd, so ist $d = -B \cdot \nabla f(x)$ eine Abstiegsrichtung.

Es ergibt sich das folgende allgemeine Gerüst für ein Abstiegsverfahren:

2.1.4 Algorithmus
wähle $x^0 \in \mathbb{R}^n$
for $k = 0, 1, \ldots$ **do**
bestimme Abstiegsrichtung d^k für f in x^k
bestimme Schrittweite $t^k > 0$ (mit $f(x^k + t^k d^k) < f(x^k)$)
setze $x^{k+1} = x^k + t^k d^k$
stoppe, wenn x^{k+1} geeignetem Abbruchkriterium genügt
end for

Ziel jetzt: Finde allgemeine Bedingungen an d^k und t^k, so dass „Konvergenz“ nachgewiesen werden kann.
Was bedeutet „Konvergenz“? Ideal wäre: $\lim_{k\to\infty} x^k = x^*$, x^* globale Minimalstelle von f. Tatsächlich kann man i.d.R. nur viel weniger nachweisen. Uns genügt, wenn wir zeigen können: Jeder Häufungspunkt der Folge $\{x^k\}$ ist stationärer Punkt von f ($f \in \mathcal{C}^1(\mathbb{R}^n)$).
Nicht jedes Abstiegsverfahren muss konvergieren:

2.1.5 Beispiel
$D = \mathbb{R}, f(x) = x^2$.

(i) Wir nehmen $x^0 = 2$, $d^k = (-1)^{k-1}$, $t^k = 2 + \frac{3}{2^{k+1}}$.
Es ist dann also

$$x^0 = 2,\ x^1 = -\frac{3}{2},\ x^2 = \frac{5}{4},\ x^3 = -\frac{9}{8}, \ldots, x^k = (-1)^k(1 + 2^{-k})$$

Es ist $\lim_{k\to\infty} f(x^k) = 1, \lim_{k\to\infty} x^{2k} = +1, \lim_{k\to\infty} x^{2k+1} = -1$.
Problem: Relativ zu $\|t^k d^k\|$ ist $f(x^k) - f(x^{k+1})$ zu klein, s. Abb. 2.1.

(ii) Wir nehmen $d^k = -1$, $t^k = 2^{-(k+1)}$, $x^0 = 2$. Dann ist $x^k = 1 + 2^{-k}$, also $\lim_{k\to\infty} x^k = +1$.
Problem: $\|t^k d^k\|$ ist zu klein (im Vergleich zu $f'(x^k)$), s. Abb. 2.2.

Wir formulieren eine „Regel“, die beide Probleme aus dem Beispiel ausschließt.

2.1.6 Definition
Eine Schrittweite t für die Abstiegsrichtung d in x heißt *effizient*, falls mit $\theta > 0$ unabhängig von x, d gilt

$$f(x + td) \le f(x) - \theta \left(\frac{\nabla f(x) d}{\|d\|} \right)^2 . \tag{2.1.1}$$

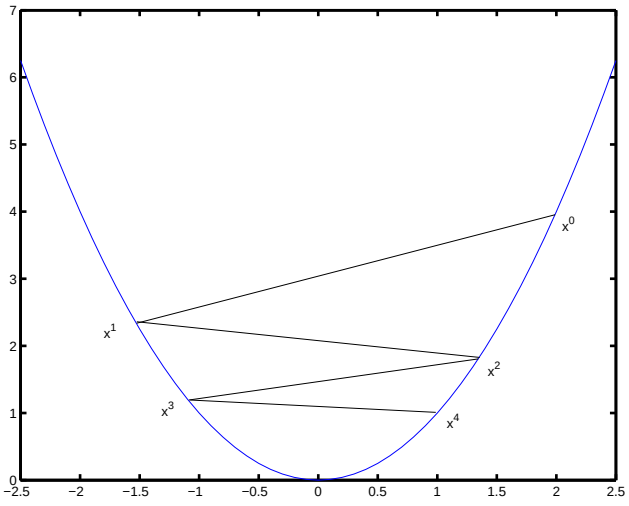

Abbildung 2.1: Illustration zu (i) aus Beispiel 2.1.5

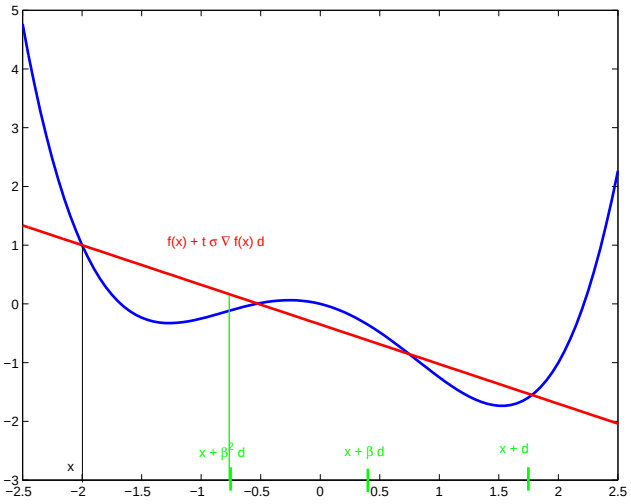

Abbildung 2.2: Illustration zu (ii) aus Beispiel 2.1.5

(Warth, Werner, 1972).

Wann ist dies erreichbar? Dies besprechen wir im nächsten Abschnitt. Zuerst kümmern wir uns um die daraus folgenden Konvergenzaussagen.

2.1.7 Satz
$f : \mathbb{R}^n \to \mathbb{R}$, $f \in \mathcal{C}^1(\mathbb{R}^n)$. Im Abstiegsverfahren 2.1.4 seien alle Schrittweiten t^k effizient und es gelte für alle k mit einer festen Zahl $c > 0$ die *Winkelbedingung*

$$-(\nabla f(x^k) \cdot d^k) \geq c \cdot \|\nabla f(x^k)\| \cdot \|d^k\| \tag{2.1.2}$$

Dann gilt

(i) Jeder Häufungspunkt der Folge $\{x^k\}$ ist ein stationärer Punkt von f.

(ii) Ist f nach unten beschränkt, so ist gilt $\lim_{k\to\infty} \nabla f(x^k) = 0$.

Beweis: Wegen t^k effizient, gilt

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

Wegen der Winkelbedingung folgt weiter

$$f(x^{k+1}) \leq f(x^k) - \theta c^2 \|\nabla f(x^k)\|^2. \tag{2.1.3}$$

Für (i) sei nun x^* Häufungspunkt von $\{x^k\}$. Die Folge $\{f(x^k)\}$ fällt monoton und konvergiert auf einer Teilfolge gegen $f(x^*)$. Also gilt

$$\lim_{k\to\infty} f(x^k) = f(x^*),$$

insbesondere $\lim_{k\to\infty} f(x^{k+1}) - f(x^k) = 0$, so dass mit (2.1.3) folgt: $\nabla f(x^*) = 0$. Für (ii) sei f nach unten beschränkt, weshalb die monoton fallende Folge $f(x^k)$ konvergiert. Aus (2.1.3) folgt dann sofort $\lim_{k\to\infty} \|\nabla f(x^k)\| = 0$. □

2.1.8 Korollar
Ist f außerdem konvex, so ist jeder Häufungspunkt von $\{x^k\}$ eine globale Minimalstelle. Ist f sogar strikt konvex, so konvergiert $\{x^k\}$ gegen die eindeutige Minimalstelle x^* von f.

Die Bedingung (2.1.2) sagt, dass der Winkel zwischen $\nabla f(x^*)$ und d^k gleichmäßig von $\pi/2$ weg beschränkt ist. Dies kann man abschwächen zur sog. Zoutendijk-Bedingung,

$$\sum_{k=0}^{\infty} \left(\frac{\nabla f(x^k) d^k}{\|\nabla f(x^k)\| \cdot \|d^k\|} \right)^2 = +\infty, \tag{2.1.4}$$

wenn man stärkere Voraussetzungen für f hat.

2.1.9 Satz
Sei $f : \mathbb{R}^n \to \mathbb{R}, f \in \mathcal{C}^1(\mathbb{R}^n)$, f nach unten beschränkt. Im Abstiegsverfahren 2.1.4 seien die Schrittweiten t^k effizient; die d^k erfüllen die Zoutendijk-Bedingung (2.1.4). Dann gilt für die Iterierten x^k

$$\liminf_{k\to\infty} \|\nabla f(x^k)\| = 0.$$

Beweis: Wegen der Effizienz der Schrittweiten haben wir

$$\begin{aligned} f(x^{k+1}) - f(x^k) &\leq -\theta \left(\frac{\nabla f(x^k) d^k}{\|d^k\|} \right)^2 \\ &= -\theta \|\nabla f(x^k)\|^2 \delta_k \end{aligned} \tag{2.1.5}$$

mit

$$\delta_k = \left(\frac{\nabla f(x^k) d^k}{\|\nabla f(x^k)\| \cdot \|d^k\|} \right)^2.$$

Angenommen, es existiert $\varepsilon > 0$ so dass $\|\nabla f(x^k)\| \geq \varepsilon$ für alle $k \geq k_\varepsilon$. Durch Summmation über alle k und mit f^* als eine untere Schranke für f erhalten wir dann

$$f^* - f(x^0) \leq -\theta \cdot \varepsilon \cdot \sum_{k=0}^{\infty} \delta_k = -\infty,$$

ein Widerspruch. □

2.1.10 Bemerkung
Ein zu Teil (i) aus Satz 2.1.7 analoges Resultat für die Zoutendijk-Bedingung können wir nicht zeigen. Der folgende Satz überträgt aber wenigstens die zweite Aussage aus Korollar 2.1.8.

2.1.11 Satz
Sei $f : \mathbb{R}^n \to \mathbb{R}, f \in \mathcal{C}^1(\mathbb{R}^n)$. Die Levelmenge $\mathcal{L}(x^0)$ sei konvex und f sei glm. konvex auf $\mathcal{L}(x^0)$. Im Abstiegsverfahren 2.1.4 seien die Schrittweiten t^k effizient; die d^k erfüllen die Zoutendijk-Bedingung (2.1.4). Dann konvergiert die Folge $\{x^k\}$ gegen die eindeutige globale Minimalstelle x^* von f.

Beweis: Es gilt $x^* \in \mathcal{L}(x^0)$. Wie vorher haben wir

$$\begin{aligned} f(x^{k+1}) &\leq f(x^k) - \theta \left(\frac{\nabla f(x^k) d^k}{\|d^k\|} \right)^2 \\ &= f(x^k) - \theta \|\nabla f(x^k)\|^2 \delta_k \end{aligned} \tag{2.1.6}$$

mit

$$\delta_k = \left(\frac{\nabla f(x^k) d^k}{\|\nabla f(x^k)\| \cdot \|d^k\|} \right)^2.$$

Da f glm. konvex auf $\mathcal{L}(x^0) \supseteq \{x^0, x^1, \ldots\}$, gilt nach Satz 1.3.11

$$\begin{aligned} \mu\|x^k - x^*\|^2 \leq f(x^k) - f(x^*) &\leq \nabla f(x^k)(x^k - x^*) \\ &\leq \|\nabla f(x^k)\| \cdot \|x^k - x^*\|, \end{aligned}$$

also insbesondere $\mu\|x^k - x^*\| \leq \|\nabla f(x^k)\|$ und damit

$$f(x^k) - f(x^*) \leq \frac{1}{\mu}\|\nabla f(x^k)\|^2.$$

Eingesetzt in (2.1.6) ergibt sich

$$f(x^{k+1}) - f(x^k) \leq -\delta_k \theta\mu \left(f(x^k) - f(x^*)\right).$$

Daraus folgt

$$\begin{aligned} 0 \leq f(x^{k+1}) - f(x^*) &\leq (1 - \delta_k\theta\mu)\left(f(x^k) - f(x^*)\right) \\ &\leq \ldots \\ &\leq \prod_{\ell=0}^{k}(1 - \delta_\ell\theta\mu) \cdot \left(f(x^0) - f(x^*)\right). \end{aligned}$$

Nun gilt

$$\begin{aligned} \log\left(\prod_{\ell=0}^{k}(1 - \delta_\ell\theta\mu)\right) &= \sum_{\ell=0}^{k}\log(1 - \delta_\ell\theta\mu) \\ &\leq -\theta\mu \cdot \sum_{\ell=0}^{k}\delta_\ell. \end{aligned}$$

Für $k \to \infty$ folgt wegen (2.1.4) somit

$$\begin{aligned} & \lim_{k\to\infty} \log\left(\prod_{\ell=0}^{k}(1 - \delta_\ell\theta\mu)\right) = -\infty \\ \iff \quad & \lim_{k\to\infty} \prod_{\ell=0}^{k}(1 - \delta_\ell\theta\mu) = 0. \end{aligned}$$

□

Die Sätze 2.1.7 und 2.1.11 sind unsere beiden wichtigsten „globalen“ Konvergenzsätze.

Abschnitt 2.2

Armijo-Schrittweiten

Wir wollen drei Strategien zur Wahl der Schrittweite kennen lernen, die jeweils effiziente Schrittweiten liefern.

2.2.1 Definition (Armijo-Regel)
Sei $f : \mathbb{R}^n \to \mathbb{R}, f \in \mathcal{C}^1(\mathbb{R}^n)$. Sei $\sigma \in (0,1)$ und $\beta \in (0,1)$ fest. Dann wählt man als Schrittweite t für die Abstiegsrichtung d im Punkt x:

$$t_A = \max\{\beta^\ell : \ell = 0, 1, \ldots, \quad f(x + \beta^\ell d) \leq f(x) + \sigma\beta^\ell \nabla f(x) d\}$$

2.2.2 Algorithmus (Armijo-Schrittweite)
$t = 1$
while $f(x + td) > f(x) + \sigma t \nabla f(x) d$ **do**
$t = t \cdot \beta$
end while
$t_A = t$

Existiert die Armijo-Schrittweite überhaupt immer?

2.2.3 Satz
$f : \mathbb{R}^n \to \mathbb{R}, f \in \mathcal{C}^1(\mathbb{R}^n), \sigma \in (0,1), \beta \in (0,1)$ fest. Zu $x \in \mathbb{R}^n$, Abstiegsrichtung $d \in \mathbb{R}^n$ existiert $\ell_0 \in \mathbb{N}_0$ mit

$$f(x + \beta^\ell d) \leq f(x) + \sigma\beta^\ell \nabla f(x) d \quad \text{für alle } \ell \geq \ell_0$$

Beweis: Angenommen, für unendlich viele ℓ wäre

$$f(x + \beta^\ell d) > f(x) + \sigma\beta^\ell \nabla f(x) d.$$

Dann folgt für $\ell \to \infty$ aus

$$\frac{f(x + \beta^\ell d) - f(x)}{\beta^\ell} > \sigma \nabla f(x) d$$

die Beziehung

$$\nabla f(x) \cdot d \geq \sigma \nabla f(x) d,$$

was wegen $\sigma \in (0,1)$ und $\nabla f(x)d < 0$ unmöglich ist.
Die Armijo-Regel kuriert Problem (i) aus Beispiel 2.1.5, aber nicht Problem (ii) (zu kleine Schritte). Die Armijo-Regel führt nicht notwendig zu effizienten Schrittweiten. □
Variante: Skalierte Armijo-Regel: Wähle $s = s(x,d) > 0$ geeignet.

$$t_{sA} = \max\{s\beta^\ell : \ell = 0,1,\dots, \quad f(x+s\beta^\ell d) \le f(x) + \sigma s\beta^\ell \nabla f(x)d\} \tag{2.2.1}$$

Für s groß genug kann man zu kleine Schritte ausschließen. Genauer gilt:

2.2.4 Satz
Sei $f : \mathbb{R}^n \to \mathbb{R}, f \in \mathcal{C}^1(\mathbb{R}^n), \sigma \in (0,1), \beta \in (0,1), c > 0$ fest. Sei $x \in \mathbb{R}^n$, d Abstiegsrichtung in x. Dann gilt

(i) t_{sA} ist definiert, d.h. in (2.2.1) rechts steht nicht die leere Menge.

(ii) Verlangt man
$$s = -c\frac{\nabla f(x)d}{\|d\|^2},$$
so ist t_{sA} sogar effizient für alle $x \in \mathcal{L}(x^0)$, vorausgesetzt ∇f ist Lipschitz-stetig auf $\mathcal{L}(x^0)$.

Beweis: (i) geht wie Satz 2.2.3.
(ii): Es gilt

$$\begin{aligned} f(x+t_{sA}d) &\le f(x) + \sigma t_{sA}\nabla f(x)d \\ &= f(x) + \sigma s\beta^\ell \nabla f(x)d \\ &\le f(x) - \sigma \cdot c\beta^\ell \left(\frac{\nabla f(x)d}{\|d\|}\right)^2. \end{aligned}$$

Es ist also nun noch zu zeigen, das für alle x, d der Exponent ℓ beschränkt ist.
∇f Lipschitz-stetig $\iff \|\nabla f(x) - \nabla f(y)\| \le \gamma\|x-y\|$ für alle $x, y \in \mathcal{L}(x^0)$.
Wir müssen zeigen, dass ℓ unabhängig von x, d existiert mit

$$f(x+s\beta^\ell d) \le f(x) + \sigma s\beta^\ell \nabla f(x)d, \quad x \in \mathcal{L}(x^0).$$

Dazu

$$\begin{aligned} f(x+s\beta^\ell d) - f(x) &= s\beta^\ell \cdot \nabla f(\xi)d, \quad \xi = x + \theta\cdot\beta^\ell d \\ &= s\beta^\ell \cdot (\nabla f(x)d + (\nabla f(\xi) - \nabla f(x))d) \\ &\le s\beta^\ell \cdot \nabla f(x)d + s\beta^\ell \|\nabla f(\xi) - \nabla f(x)\| \cdot \|d\| \end{aligned}$$

$$\leq \quad s\beta^\ell \cdot \left(\nabla f(x)d + s\beta^\ell\gamma \cdot \|d\|^2\right).$$

Also ist $f(x + s\beta^\ell d) - f(x) \leq \sigma s\beta^\ell \cdot \nabla f(x)d$, sobald

$$s\beta^\ell\gamma\|d\|^2 \leq (\sigma - 1)(\nabla f(x)d). \tag{2.2.2}$$

Nach Voraussetzung gilt

$$(\sigma - 1)\nabla f(x)d = -(\sigma - 1) \cdot \frac{s}{c} \cdot \|d\|^2,$$

also gilt (2.2.2), sobald

$$\begin{aligned} & s\beta^\ell\gamma\|d\|^2 \leq (1 - \sigma) \cdot \frac{s}{c} \cdot \|d\|^2 \\ \iff \quad & \ell \geq \frac{1}{\log\beta} \cdot \log\left(\frac{1-\sigma}{c\gamma}\right), \end{aligned}$$

unabhängig von x, d. □

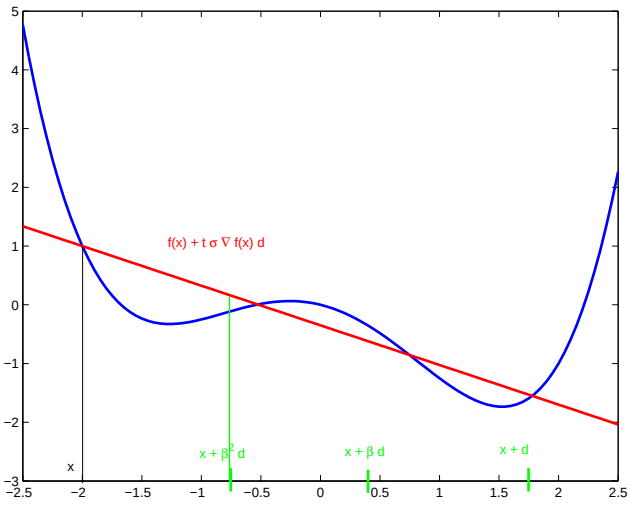

Abbildung 2.3: Die Armijo-Schrittweite

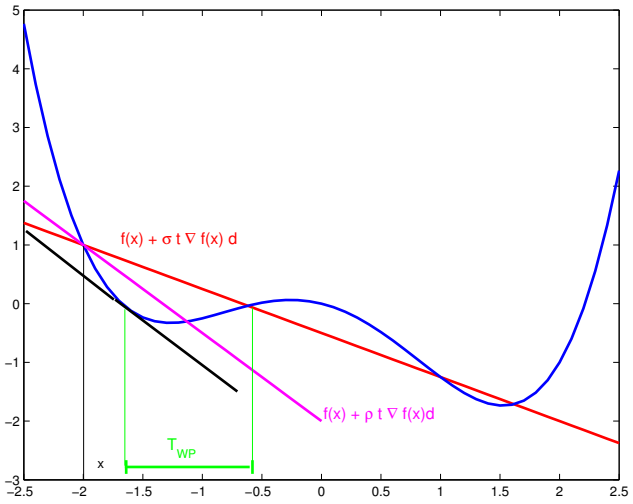

Abbildung 2.4: Die Wolfe-Powell-Schrittweiten

Abschnitt 2.3

Wolfe-Powell-Schrittweiten

Eine andere wichtige Schrittweitenstrategie ist die Wolfe-Powell-Strategie.

2.3.1 Definition
$f : \mathbb{R}^n \to \mathbb{R}, f \in \mathcal{C}^1(\mathbb{R}^n), \sigma \in (0, \frac{1}{2}), \rho \in [\sigma, 1)$ fest gegeben. Eine *Wolfe-Powell-Schrittweite* t_{WP} für $x \in \mathbb{R}^n$ mit Abstiegsrichtung d ist ein $t > 0$ mit

$$f(x + td) \leq f(x) + \sigma t \nabla f(x) d \tag{2.3.1}$$

und

$$\nabla f(x + td) d \geq \rho \nabla f(x) d. \tag{2.3.2}$$

Abb. 2.4 illustriert, dass die zweite Bedingung nun zu kleine Schrittweiten verhindert. Existiert t_{WP} immer?

2.3.2 Satz
$f : \mathbb{R}^n \to \mathbb{R}, f \in \mathcal{C}^1(\mathbb{R}^n), \sigma \in (0, \frac{1}{2}), \rho \in [\sigma, 1]$. Zu $x \in \mathbb{R}^n$ und $d \in \mathbb{R}^n$ mit $\nabla f(x) d < 0$ sei

$$T_{WP}(x, d) = \{t > 0 : t \text{ erfüllt } (2.3.1) \text{ und } (2.3.2)\}.$$

(i) Ist f nach unten beschränkt, so ist $T_{WP} \neq \emptyset$.

(ii) Ist ∇f auf $\mathcal{L}(x^0)$ Lipschitz-stetig, so sind für alle $x \in \mathcal{L}(x^0), d \in \mathbb{R}^n$ alle $t \in T_{WP}$ effizient.

Beweis:

(i) Setze

$$\begin{aligned} \varphi(t) &= f(x+td), \\ \psi(t) &= f(x+td) - f(x) - \sigma t \nabla f(x)d \;=\; \varphi(t) - \varphi(0) - \sigma t \varphi'(0). \end{aligned}$$

Dann ist $\varphi'(t) = \nabla f(x+td)d$, $\psi'(t) = \varphi'(t) - \sigma\varphi'(0)$.

Die Wolfe-Powell-Bedingungen (2.3.1) und (2.3.2) lauten so

$$\psi(t) \leq 0 \tag{2.3.3}$$

$$\varphi'(t) \geq \rho\varphi'(0). \tag{2.3.4}$$

Die Funktion $\psi(t)$ erfüllt $\psi(0) = 0, \lim_{t\to\infty} \psi(t) = \infty$ (denn f ist beschränkt), $\psi'(0) < 0$ und damit für $\psi'(t) < 0$ für $t \in [0, \bar{t})$ mit $\bar{t} > 0$. Also existiert $t^* > 0$ mit $\psi(t^*) = 0$, $\psi(t) < 0$ für $t \in (0, t^*)$. Es gilt $\psi'(t^*) \geq 0$, denn andernfalls gäbe es ein noch kleineres t^{**} mit $\psi(t^{**}) = 0$. Die Zahl t^* erfüllt (2.3.3) auf Grund ihrer Definition. Unter Verwendung von $\psi'(t^*) \geq 0$ folgt außerdem $\varphi'(t^*) \geq \sigma\varphi'(0)$, woraus wegen $\rho \geq \sigma > 0$ und $\varphi'(0) < 0$ die Beziehung (2.3.4) folgt.

(ii) Für $x \in \mathcal{L}(x^0), t \in T_{WP}(x, d)$ ist $x + td \in \mathcal{L}(x^0)$. Sei γ die Lipschitz-Konstante für $\nabla f(x)$. Wir haben nun wegen (2.3.1)

$$f(x+td) \leq f(x) + \sigma t \nabla f(x)d. \tag{2.3.5}$$

Andererseits gilt wegen (2.3.2)

$$\rho\nabla f(x)d \leq \nabla f(x+td)d,$$

also

$$\begin{aligned} (\rho-1)\nabla f(x)d &\leq (\nabla f(x+td) - \nabla f(x))d \\ &\leq \|\nabla f(x+td) - \nabla f(x)\| \cdot \|d\| \\ &\leq \gamma \cdot t \cdot \|d\|^2, \end{aligned}$$

also

$$t \geq \frac{(\rho-1)\nabla f(x)d}{\gamma \cdot \|d\|^2} \geq 0$$

Durch Einsetzen in (2.3.5) ergibt sich

$$f(x+td) \leq f(x) + \frac{\sigma(\rho-1)\nabla f(x)d}{\gamma \cdot \|d\|^2} \cdot \nabla f(x)d.$$

Man nehme also $\theta = -\frac{\sigma(\rho-1)}{\gamma}$ in der Definition für effiziente Schrittweite.

□

Eine Variante sind die strengen Wolfe-Powell-Schrittweiten.

2.3.3 Definition
$f : \mathbb{R}^n \to \mathbb{R}, f \in \mathcal{C}^1(\mathbb{R}^n), \sigma \in (0, \frac{1}{2}), \rho \in [\sigma, 1)$ fest gegeben. Eine *strenge Wolfe-Powell-Schrittweite* t_{sWP} für $x \in \mathbb{R}^n$ mit Abstiegsrichtung d ist ein $t > 0$ mit

$$f(x+td) \leq f(x) + \sigma t \nabla f(x)d \tag{2.3.6}$$

und

$$|\nabla f(x+td)d| \leq -\rho \nabla f(x)d \tag{2.3.7}$$

(2.3.6) ist identisch zu (2.3.1). Die zweite Bedingung (2.3.7) ist stärker als (2.3.2), denn wegen $\nabla f(x)d < 0$ ist (2.3.7) äquivalent zu

$$\rho \nabla f(x)d \leq \nabla f(x+td)d \leq -\rho \nabla f(x)d.$$

Auf Grund von Satz (2.3.6) (ii) wissen wir bereits, dass t_{sWP} effizient ist, vorausgesetzt, t_{sWP} existiert.
Zur weiteren Diskussion setzen wir wie im Beweis zu Satz 2.3.2 (bei festen x, d, σ, ρ)

$$\begin{aligned} \varphi(t) &= f(x+td), \\ \psi(t) &= \varphi(t) - \varphi(0) - \sigma t \varphi'(0). \end{aligned}$$

Die strenge Wolfe-Powell-Bedingung lautet also

$$\psi(t) \leq 0 \tag{2.3.8}$$

$$|\varphi'(t)| \geq \rho |\varphi'(0)|. \tag{2.3.9}$$

2.3.4 Satz
f, σ, ρ seien wie in Definition 2.3.3. Ist f nach unten beschränkt, so ist $T_{sWP} \neq \emptyset$.

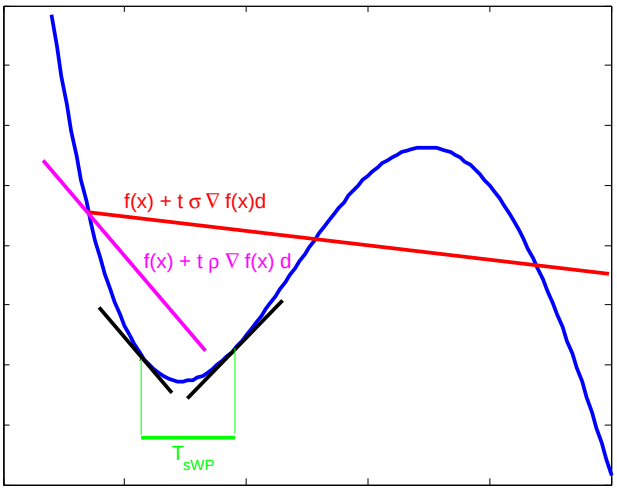

Abbildung 2.5: Die strengen Wolfe-Powell-Schrittweiten

Beweis: Die Argumentation ist ähnlich wie für Satz 2.3.2(i): es existiert ein Intervall $(0, t^*)$, auf dem die differenzierbare Funktion $\psi(t)$ die Beziehungen $\psi(t) < 0$ und $\psi(t^*) = 0$ erfüllt. Wegen $\psi(0) = 0$ existiert ein $\tilde{t} \in (0, t^*)$ mit $\psi'(\tilde{t}) = 0$. Aus der letzten Beziehung folgt $|\varphi'(t)| = \sigma|\varphi'(0)| \leq \rho|\varphi'(0)|$. □

Die konkrete Bestimmung einer (strengen) Wolfe-Powell-Schrittweite ist komplizierter als bei der Armijo-Schrittweite. Wir formulieren dazu jetzt Algorithmen, die jeweils auf einer einfachen geometrischen Idee beruhen.
Zur weiteren Diskussion setzen wir wie im Beweis zu Satz 2.3.2 (bei festen x, d, σ, ρ)

$$\begin{aligned} \varphi(t) &= f(x + td), \\ \psi(t) &= \varphi(t) - \varphi(0) - \sigma t\varphi'(0). \end{aligned}$$

Die Wolfe-Powell-Bedingungen lauten also

$$\psi(t) \leq 0 \tag{2.3.10}$$

$$\varphi'(t) \geq \rho\varphi'(0) \tag{2.3.11}$$

Wir nehmen an, dass f nach unten beschränkt ist. Dann ist $\lim_{t\to\infty} \psi(t) = +\infty$, mit $\psi(0) = 0, \psi'(0) = (1 - \sigma)\varphi'(0) < 0$. Es gibt also ein Intervall $[a, b] \subseteq [0, \infty)$, auf welchem gilt: $\psi(a) < 0, \psi(b) > 0, \psi'(t^*) = 0$ für ein $t^* \in [a, b]$. Es ist

$$\begin{aligned} & \psi'(t^*) = 0 \\ \iff\ & \varphi'(t^*) - \sigma\varphi'(0) = 0 \\ \iff\ & \varphi'(t^*) = \sigma\varphi'(0) \end{aligned}$$

Wegen $\rho > \sigma$ und $\varphi'(0) < 0$ gilt also $\varphi'(t^*) \geq \rho\varphi'(0)$, d.h. $t^* \in T_{WP}$.
Damit **Idee**: Finde zuerst a und b, suche danach ein geeignetes t in $[a, b]$ durch Halbierungsstrategie. Beachte dabei: Nach dem MWS existiert $t \in [a, b]$ mit $\psi'(t) = (\psi(b) - \psi(a))/(b - a) \geq 0$, d.h. $\psi'(t) \geq \sigma\varphi'(0) \Rightarrow \varphi'(t) \geq \rho\varphi'(0)$. Das t findet man im folgenden Algorithmus durch „Zufall“ oder weil $b - a$ immer kleiner wird.

2.3.5 Algorithmus (Wolfe-Powell-Schrittweite)

wähle $t > 0, \gamma > 1$,
while $(\psi(t) \leq 0) \wedge (\varphi'(t) < \rho\varphi'(0))$ **do**
  $t = \gamma \cdot t$
end while
$\{\psi(t) > 0$ oder $\varphi'(t) \geq \rho\varphi'(0)\}$
if $\psi(t) \leq 0$ **then**
$\{$dann ist $\varphi'(t) \geq \rho\varphi'(0)$, Glück gehabt!$\}$
  $t_{WP} = t$
else
$\{\psi(t) > 0$ und $\varphi'(t) \geq \rho\varphi'(0)\}$
  setze $a = 0, b = t$
  setze $t = \frac{1}{2}(a + b)$
$\{\varphi'(a) < \rho\varphi'(0)\}$
  while $(\psi(t) > 0) \vee (\varphi'(t) < \rho\varphi'(0))$ **do**
    if $\psi(t) > 0$ **then**
$\{t$ ist zu groß$\}$
      setze $b = t$
    else
$\{\psi(t) \leq 0, \varphi'(t) < \rho\varphi'(0)\}$
      setze $a = t$
    end if
    setze $t = \frac{1}{2}(a + b)$
  end while
  $t_{WP} = t$
end if

Der Algorithmus bedarf einer Analyse. Es genügt zu zeigen, dass er terminiert, denn das berechnete t_{WP} erfüllt (2.3.10) und (2.3.11) auf Grund der Abbruchbedingung der zweiten while-Schleife.

2.3.6 Satz
Sei $f : \mathbb{R}^n \to \mathbb{R}$, $f \in \mathcal{C}^1(\mathbb{R}^n)$, f nach unten beschränkt. Seien $x, d \in \mathbb{R}^n$ fest $\nabla f(x)d < 0$. In Algorithmus 2.3.5 sei $0 < \sigma < \rho < 1$. Dann terminiert der Algorithmus.

Beweis: Die erste while-Schleife bricht ab, da wegen $\gamma > 1$ die Zahl t beliebig groß wird und $\lim_{t\to\infty} \psi(t) = \infty$, d.h. es wird $\psi(t) \geq 0$ erreicht.
Die zweite while-Schleife bestimmt eine Folge von Intervallen $[a_i, b_i] \subseteq [a_{i-1}, b_{i-1}]$, $a_i \leq b_i$, mit den Eigenschaften

$$\psi(a_i) \leq 0, \quad \psi(b_i) > 0, \quad \varphi'(a_i) < \rho\varphi'(0),$$

$$|b_i - a_i| = \frac{1}{2}|b_{i-1} - a_{i-1}|.$$

Angenommen, die Schleife bricht nicht ab. Dann gilt $\lim_{i\to\infty} a_i = t = \lim_{i\to\infty} b_i$ mit $\psi(t) = 0, \psi'(t) = \lim_{i\to\infty}(\psi(b_i) - \psi(a_i))/(b_i - a_i) \geq 0$. Es gilt also $\psi(t) = 0, \varphi'(t) \geq \sigma\varphi'(0) > \rho\varphi'(0)$. Aus Stetigkeitsgründen gilt deshalb für i groß genug $\varphi'(a_i) \geq \rho\varphi'(0)$ (und sowieso $\psi(a_i) \leq 0$). Ein solches a_i ist aber ein t in der zweiten while-Schleife des Algorithmus, d.h. der Algorithmus bricht ab. □

2.3.7 Bemerkung
Algorithmus 2.3.5 bleibt auch dann noch korrekt, wenn man statt $t = \frac{1}{2}(a+b)$ irgendein $t \in [a + \tau_1(b-a), b - \tau_2(b-a)]$ mit $\tau_1, \tau_2 > 0, \tau_1 + \tau_2 \leq 1$ nimmt. Denn auch dann hätte man, falls die zweite while-Schleife nicht abbricht, $\lim_{i\to\infty} a_i = t = \lim_{i\to\infty} b_i$.

Folge: Man kann τ_1, τ_2 relativ klein wählen, z.B. $\tau_1, \tau_2 = 10^{-2}$, und stellt ein einfaches Modell für φ auf $[a, b]$ auf, z. B.

$$\varphi(t) \approx p(t),$$

p Interpolationspolynom vom Grad 3 mit $p(a) = \varphi(a), p'(a) = \varphi'(a), p(b) = \varphi(b), p'(b) = \varphi'(b)$.
Bestimme dann die Minimalstelle t^* von p auf $[a, b]$ (geht explizit!). Falls $t^* \in [a + \tau_1(b-a), b - \tau_2(b-a)]$, setze $t = t^*$, sonst $t = a + \tau_1(b-a)$ (z. B.).
Alternativ kann man für p auch das Interpolationspolynom 2. Grades nehmen mit $p(a) = \varphi(a), p'(a) = \varphi'(a), p(b) = \varphi(b)$. Dies erspart die Berechnung einer Ableitung.

Für die strenge Wolfe-Powell-Schrittweiten entwerfen wir einen ähnlichen Algorithmus. Die strenge Wolfe-Powell-Bedingung ist äquivalent zu

$$\psi(t) \leq 0, \quad |\varphi'(t)| \leq \rho|\varphi'(0)| = -\rho\varphi'(0).$$

Die zweite Bedingung ist dabei erfüllt, wenn $|\psi'(t)|$ klein genug ist, denn es gilt

$$\begin{aligned}
& |\psi'(t)| \leq (\rho - \sigma) \cdot |\varphi'(0)| \\
\Rightarrow \quad & -(\rho - \sigma) \cdot |\varphi'(0)| \leq \varphi'(t) - \sigma \cdot \varphi'(0) \leq (\rho - \sigma) \cdot |\varphi'(0)| \\
\Longleftrightarrow \quad & -\rho|\varphi'(0)| \leq \varphi'(t) \leq \underbrace{(\rho - 2\sigma)}_{\leq \rho} |\varphi'(0)| \\
\Rightarrow \quad & |\varphi'(t)| \leq \rho|\varphi'(0)|.
\end{aligned}$$

Idee des Algorithmus: Finde Folge von immer kleineren Intervallen, so dass in jedem Intervall ein t^* liegt mit $\psi(t^*) < 0$ und $\psi'(t^*) = 0$. In einer Umgebung

dieses t^* gilt dann $\psi(t) \leq 0$ und $|\psi'(t)| \leq (\rho - \sigma) \cdot |\varphi'(0)|$. Hierzu bemerken wir:
Für $a < b$ existiert ein solches $t^* \in (a, b)$, falls gilt

$$\psi(a) \leq \psi(b), \psi(a) < 0 \text{ und } \psi'(a) < 0$$

oder

$$\psi(b) \leq \psi(a), \psi(b) < 0 \text{ und } \psi'(b) > 0.$$

Eine Möglichkeit, beide Bedingungen in einer zu formulieren, erhält man, wenn man auch $b < a$ zulässt:

$$\psi(a) \leq \psi(b), \psi(a) < 0 \text{ und } \psi'(a)(b - a) < 0 \tag{2.3.12}$$

Im folgenden Algorithmus wird dies jetzt auch verwendet. (a, b sind jetzt die Grenzen eines Intervalls $[a, b]$ oder $[b, a]$).

2.3.8 Algorithmus (strenge Wolfe-Powell-Schrittweite)

wähle $t > 0, \gamma > 1$,
while $(\psi(t) < 0) \wedge (\psi'(t) \leq 0)$ **do**
  $t = \gamma t$
end while
$\{\psi(t) \geq 0$ oder $\psi'(t) > 0\}$
if $(\psi(t) < 0) \wedge (|\psi'(t)| \leq (\rho - \sigma)|\varphi'(0)|)$ **then**
  $t_{sWP} = t$ {Glück gehabt!}
else
$\{\psi(t) \geq 0$ oder $(\psi(t) < 0$ und $|\psi'(t)| > (\rho - \sigma)|\varphi'(0)|)\}$
  if $(\psi(t) < 0) \wedge (\psi'(t) > 0)$ **then**
$\{$es ist sogar $\psi'(t) > (\rho - \sigma)|\varphi'(0)|\}$
    setze $a = t, b = 0$ $\{a, b$ erfüllen (2.3.12)$\}$
  else
$\{\psi(t) \geq 0\}$
    setze $a = 0, b = t$ $\{a, b$ erfüllen (2.3.12)$\}$
  end if
$\{$hier ist außerdem $|\psi'(a)| > (\rho - \sigma)|\varphi'(0)|$ $\}$
$\{a, b$ erfüllen (2.3.12), $|\psi'(a)| > (\rho - \sigma)|\varphi'(0)|\}$
  setze $t = \frac{1}{2}(a + b)$
  while $(\psi(t) \geq \psi(a)) \vee (|\psi'(t)| > (\rho - \sigma)|\varphi'(0)|)$ **do**
    if $\psi(t) \geq \psi(a)$ **then**
      setze $b = t$ $\{$ a, b erfüllen (2.3.12)$\}$
    else
$\{|\psi'(t)| > (\rho - \sigma)|\varphi'(0)|, \psi(t) < \psi(a)\}$
      if $\psi'(t)(t - a) < 0$ **then**
        setze $a = t$ $\{a, b$ erfüllen (2.3.12)$\}$
      else
        setze $b = a$, $a = t$
      end if
$\{a, b$ erfüllen (2.3.12), $|\psi'(a)| > (\rho - \sigma)|\varphi'(0)|\}$
    end if
    $t = \frac{1}{2}(a + b)$
  end while
  $t_{sWP} = t$
end if

Zum Nachweis der Korrektheit müssen wir wieder nur die Terminierung zeigen.

2.3.9 Satz
Sei $f : \mathbb{R}^n \to \mathbb{R}$, $f \in \mathcal{C}^1(\mathbb{R}^n)$, f nach unten beschränkt. Seien $x, d \in \mathbb{R}^n$ fest $\nabla f(x)d < 0$. In 2.3.8 sei $0 < \sigma < \rho < 1$. Dann terminiert der Algorithmus.

Beweis: Die Terminierung der ersten while-Schleife (wegen $\psi(t) \geq 0$) folgt wie in Satz 2.3.6. Zur Terminierung der zweiten while-Schleife notieren wir $\langle a, b\rangle$ für das Intervall

$$\langle a, b\rangle = [\min\{a, b\}, \max\{a, b\}].$$

Angenommen, die while-Schleife terminiert nicht. Dann bestimmt sie Folgen $\{a_i\}, \{b_i\}$ mit $\langle a_i, b_i\rangle \subseteq \langle a_{i-1}, b_{i-1}\rangle$ und $|b_i - a_i| = \frac{1}{2}|b_{i-1} - a_{i-1}|$. Also gilt $\lim_{i\to\infty} a_i = t^* = \lim_{i\to\infty} b_i$. Da alle Paare $\{a_i, b_i\}$ die Bedingung (2.3.12) erfüllen, existiert stets ein $t_i \in \langle a_i, b_i\rangle$ mit $\psi'(t_i) = 0$. Also gilt auch $\psi'(t^*) = 0$. Andererseits ist die Beziehung $|\psi'(a)| > (\rho - \sigma)|\varphi'(0)$ aber eine Schleifeninvariante der while-Schleife, so dass $|\psi'(t^*)| \geq (\rho - \sigma)|\varphi'(0)| > 0$ folgt, ein Widerspruch. □

2.3.10 Bemerkung
Wie bei der gewöhnlichen Wolfe-Powell-Schrittweite kann man t aus einem Intervall $\langle a+\tau_1(b-a), b-\tau_2(b-a)\rangle$ wählen und dann z.B. wieder die Minmalstelle eines kubischen oder quadratischen Interpolationpolynoms verwenden.

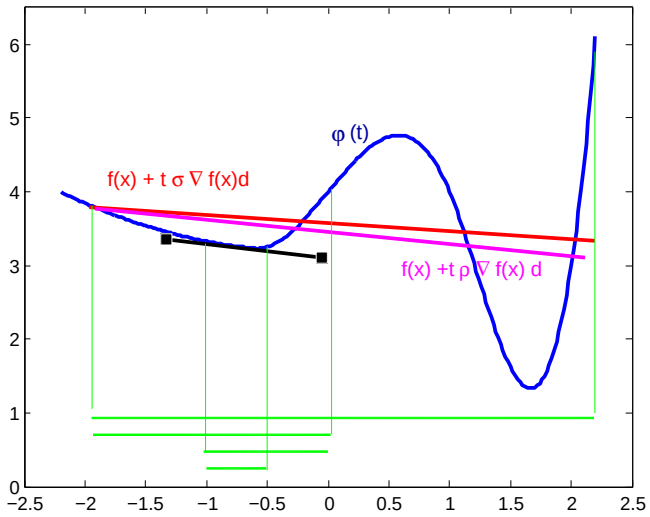

Abbildung 2.6: zur Bestimmung der Wolfe-Powell-Schrittweite

Abbildung 2.7: zur Bestimmung der strengen Wolfe-Powell-Schrittweiten

Kapitel 3

Lokale Konvergenz

Ein konvergentes Minimierungsverfahren durchläuft typischerweise zwei Phasen. In der ersten Phase wird sichergestellt, dass es in die Nähe einer Minimalstelle gerät. Diese erste Phase konvergiert in der Regel langsam. In der zweiten Phase wird die Minimalstelle dann schnell gefunden.
Die zweite Phase beruht in der Regel auf einem nur lokal (d.h. bei genügend genauer Anfangsapproximation) konvergenten Verfahren.
In diesem Kapitel führen wir Größen ein, mit denen die Konvergenzgeschwindigkeit der zweiten Phase gut beschrieben und verglichen werden kann. Dann werden wir das Newton-Verfahren untersuchen und erläutern, weshalb es für die zweite Phase so grundlegend ist.

Abschnitt 3.1

Konvergenzordnungen

Sei $\{x^k\} \subseteq \mathbb{R}^n$ eine Folge mit $\lim_{k\to\infty} x^k = x^* \in \mathbb{R}^n$ und $\|\cdot\|$ eine Norm. Man stelle sich $\{x^k\}$ als Iteriertenfolge eines lokal gegen x^* konvergenten Iterationsverfahrens vor.

3.1.1 Definition
$\{x^k\}$ konvergiert gegen x^*

(i) *q-linear*, falls

$$\|x^{k+1} - x^*\| \leq c \cdot \|x^k - x^*\| \text{ für } k \geq k_0 \text{ mit } c \in [0, 1)$$

(ii) *q-überlinear*, falls

$$\|x^{k+1} - x^*\| \leq c_k \|x^k - x^*\| \text{ für } k \geq 0 \text{ mit } \lim_{k\to\infty} c_k = 0$$

(iii) mit *q-Ordnung* $p(> 1)$, falls

$$\|x^{k+1} - x^*\| \leq c \cdot \|x^k - x^*\|^p, \text{ für } k \geq 0, c \geq 0$$

- $p = 2$: *q-quadratisch*
- $p = 3$: *q-kubisch*

3.1.2 Definition
$\{x^k\}$ konvergiert gegen x^*

(i) *r-linear*, falls

$$\|x^k - x^*\| \leq c \cdot t^k \text{ für } k \geq 0 \text{ mit } c \geq 0, t \in [0, 1).$$

(ii) *r-überlinear*, falls

$$\|x^k - x^*\| \leq c_k t^k \text{ für } k \geq 0 \text{ mit } \lim_{k\to\infty} c_k = 0 \text{ für jedes } t \in (0, 1).$$

(iii) mit *r-Ordnung* $p(> 1)$, falls

$$\|x^k - x^*\| \leq c \cdot t^{p^k} \text{ für } k \geq 0 \text{ mit } c \geq 0, t \in [0, 1).$$

q-Ordungen sind aussagekräftiger ('besser') als r-Ordungen, aber häufig kann man nur r-Ordnungen nachweisen.
Der folgende Satz diskutiert den Zusammenhang zwischen den eingeführten Begriffen.

3.1.3 Satz

(i) Alle Begriffe aus Definition 3.1.1 und 3.1.2 (bis auf q-lineare Konvergenz) sind unabhängig von $\|\cdot\|$.

(ii) q-Ordnung $p > 1 \Rightarrow$ q-Ordnung p' mit $1 < p' < p$
$\Rightarrow$ q-überlinear
$\Rightarrow$ q-linear

analog mit r-Ordnung

(iii) q-Ordnung $p > 1 \Rightarrow$ r-Ordnung p' für jedes $p' \in (1, p)$

Ansonsten Vorsicht, s. Beispiel 3.1.4 und Übung.

Beweis: Übungsaufgabe. □

3.1.4 Beispiel

(i)

$$x^k = \begin{cases} \left(\frac{1}{2}\right)^k & k \text{ gerade} \\ \left(\frac{1}{3}\right)^k & k \text{ ungerade} \end{cases}$$

$\{x^k\}$ konvergiert r-linear gegen 0, denn

$$|x^k| \leq \left(\frac{1}{2}\right)^k \quad k \in \mathbb{N}.$$

$\{x^k\}$ konvergiert *nicht* q-linear gegen 0, denn für k ungerade gilt

$$\frac{|x^{k+1}|}{|x^k|} = \frac{\left(\frac{1}{2}\right)^k}{\left(\frac{1}{3}\right)^{k+1}} = 3 \cdot \left(\frac{3}{2}\right)^k \to \infty.$$

(ii) Sei $p > 1$.

$$x^k = \begin{cases} \left(\frac{1}{2}\right)^{p^k} & k \text{ gerade} \\ \left(\frac{1}{3}\right)^{p^k} & k \text{ ungerade} \end{cases}$$

$\{x^k\}$ konvergiert mit r-Ordnung p gegen 0, denn

$$|x^k| \leq \left(\frac{1}{2}\right)^{p^k} \quad k \in \mathbb{N}.$$

$\{x^k\}$ konvergiert *nicht* mit q-Ordnung p gegen 0, denn für k ungerade gilt

$$\frac{|x^{k+1}|}{|x^k|} = \frac{\left(\frac{1}{2}\right)^{p^k}}{\left(\frac{1}{3}\right)^{p^{k+1}}} = 3^p \cdot \left(\frac{3}{2}\right)^{p^k} \to \infty.$$

Zur Bestimmung von r-Konvergenzordnungen aus einfacheren Abschätzungen formulieren wir den nachfolgenden Satz 3.1.6, den wir mit einem Hilfsresultat vorbereiten.

3.1.5 Lemma

Für $m \in \mathbb{N}$ besitzt das Polynom $p_m(t) = t^{m+1} - t^m - 1$ eine eindeutige positive Nullstelle τ mit $\tau \in (1, 2)$.

Beweis: $p_m(0) = -1$, $p_m(2) = 2^{m+1} - 2^m - 1 > 0$. Also besitzt p_m eine Nullstelle in $(0, 2)$. Weiter ist $p'_m(t) = t^{m-1}((m+1)t - m)$. Also fällt p_m monoton in $[0, \frac{m}{m+1}]$ und wächst in $[\frac{m}{m+1}, +\infty)$. Also besitzt p in (0,2) (und sogar in $(0, \infty)$) eine eindeutige Nulstelle τ, welche wegen und wegen $p_m(1) < 0$ größer als 1 ist. □

3.1.6 Satz
Sei $\gamma_0, \ldots, \gamma_m \geq 0$, $\{x^k\} \subseteq \mathbb{R}^n$ mit $\lim_{k\to\infty} x^k = x^*$ und

$$\|x^{k+1} - x^*\| \leq \|x^k - x^*\| \cdot \left(\sum_{j=0}^{m} \gamma_j \|x^{k-j} - x^*\| \right), k = m, m+1, \ldots .$$

Dann konvergiert $\{x^k\}$ mit r-Ordnung τ, wobei τ die nach Lemma 3.1.5 eindeutige Nullstelle von $p_m(t) = t^{m+1} - t^m - 1$ im Intervall $(1, 2)$ ist.

Beweis: Wir können $\gamma = \sum_{j=0}^{m} \gamma_j > 0$ voraussetzen, denn der Fall $\gamma = 0$ ist trivial. Setze

$$\varepsilon_k = \|x^k - x^*\|, \ \eta_k = \gamma \varepsilon_k, \ \delta_j = \gamma_j / \gamma.$$

Es gilt

$$\sum_{j=0}^{m} \delta_j = 1, \quad \eta_{k+1} \leq \eta_k \sum_{j=0}^{m} \delta_j \eta_{k-j}.$$

Wegen $\lim_{k\to\infty} \varepsilon_k = 0$ gilt $\eta_k \leq \beta < 1$ für $k \geq k_0$. Damit haben wir dann

$$\begin{aligned} \eta_{k_0+m+1} &\leq \beta \sum_{j=0}^{m} \delta_j \beta = \beta^2, \\ \eta_{k_0+m+2} &\leq \beta^2 \delta_0 \beta^2 + \beta^2 \sum_{j=1}^{m} \delta_j \beta \leq \beta^3, \\ &\vdots \end{aligned}$$

und allgemein (einfache Induktion)

$$\eta_{k_0+i} \leq \beta^{\mu_i}, \quad i = 0, 1, \ldots,$$

wobei

$$\mu_0 = \mu_1 = \ldots = \mu_m = 1, \quad \mu_{i+1} = \mu_i + \mu_{i-m}, \quad i \geq m.$$

Es bleibt noch zu zeigen:

$$\mu_i \geq \alpha \cdot \tau^i, \quad i = 0, 1, \ldots . \tag{3.1.1}$$

Setze dazu $\alpha = \tau^{-m}$. Dann ist wenigstens schon mal

$$\mu_0 = \mu_1 = \ldots = \mu_m = 1 \geq \alpha\tau^i = \tau^{i-m} \text{ für } i = 0, \ldots, m.$$

Angenommen, (3.1.1) gilt bis zu einem $i \geq m$. Dann ist

$$\begin{aligned} \mu_{i+1} = \mu_i + \mu_{i-m} &\geq \alpha(\tau^i + \tau^{i-m}) \\ &= \alpha\tau^{i+1}(\tau^{-1} + \tau^{-m-1}). \end{aligned}$$

Es ist aber $\tau^{-1} + \tau^{-m-1} = 1$, da $\tau^{m+1} - \tau^m - 1 = \tau^{m+1}(1 - \tau^{-1} - \tau^{-m-1}) = 0$.
□

Abschnitt 3.2

Lokale Konvergenz von Newton-Verfahren

Wir betrachten zuerst Folgen $\{x^k\} \subseteq \mathbb{R}^n$, die das Ergebnis eines Iterationsprozesses sind:

$$x^{k+1} = H(x^k) \tag{3.2.1}$$

und charakterisieren die Konvergenzrate durch Eigenschaften von H. Danach besprechen wir das Newton-Verfahren im Detail.

3.2.1 Definition
Für $x \in \mathbb{R}^n, \delta > 0, \|\cdot\|$ in $\mathbb{R}^n$, fest, bezeichnet

$$B_\delta(x) = \{y \in \mathbb{R}^n : \|y - x\| \leq \delta\}$$

die abgeschlossene Kugel um x mit Radius δ bzgl. $\|\cdot\|$.

3.2.2 Satz
Sei $H : D \subseteq \mathbb{R}^n \to \mathbb{R}^n$. Weiter sei H in $x^* \in \overset{\circ}{D}$ mit $H(x^*) = x^*$ differenzierbar mit $\rho(H'(x^*)) =: \sigma < 1$. Dann gilt für die Iterierten $x^{k+1} = H(x^k), k = 0, 1, \ldots$

(i) Für x^0 nahe genug an x^* sind alle x^k definiert.

(ii) $x^k \to x^*$ q-linear bzgl. einer geeigneten Norm $\|\cdot\|$. (Ostrowski 1960)

(iii) Ist $\rho(H'(x^*)) = 0$, so konvergiert $x^k \to x^*$ r-überlinear.

Beweis: Wir halten zuerst fest: H ist stetig im Fixpunkt x^*.
Aus Numerik I ist bekannt: Zu jedem $\varepsilon > 0$ existiert eine Norm $\|\cdot\|_\varepsilon$, so dass für die zugehörige Operatornorm $\|\cdot\|_\varepsilon$ gilt

$$\|H'(x^*)\|_\varepsilon \leq \sigma + \varepsilon.$$

H differenzierbar bedeutet, dass ein $\delta_\varepsilon > 0$ existiert mit

$$\|H(x) - H(x^*) - H'(x^*)(x - x^*)\|_\varepsilon \leq \varepsilon \cdot \|x - x^*\|_\varepsilon \text{ für } \|x - x^*\|_\varepsilon \leq \delta_\varepsilon.$$

Wähle ε fest, so klein, dass $\sigma + 2\varepsilon < 1$. Verkleinere δ_ε evtl., so dass $B_{\delta_\varepsilon}(x^*) \subseteq D$. Für $x^k \in B_{\delta_\varepsilon}(x^*)$ haben wir dann

$$\begin{aligned} \|x^{k+1} - x^*\|_\varepsilon &= \|H(x^k) - x^*\|_\varepsilon \\ &\leq \|H(x^k) - H(x^*) - H'(x^*)(x^k - x^*)\|_\varepsilon + \|H'(x^*)(x^k - x^*)\|_\varepsilon \\ &\leq \varepsilon \cdot \|x^k - x^*\|_\varepsilon + (\sigma + \epsilon) \cdot \|x^k - x^*\|_\varepsilon \\ &= (2\varepsilon + \sigma) \cdot \|x^k - x^*\|_\varepsilon \qquad (3.2.2) \\ &\leq \|x^k - x^*\|_\varepsilon \end{aligned}$$

Mit x^k liegt also auch x^{k+1} in B_{δ_ε}, d.h. die Folge $\{x^k\}$ ist definiert, sobald $x^0 \in B_{\delta_\varepsilon}(x^*)$. Dies beweist (i). Außerdem beweist (3.2.2) direkt auch (ii), denn die Konvergenz ist q-linear bezüglich $\|\cdot\|_\varepsilon$.
Zu (iii): Im Beweis zu (i) hat man jetzt $\sigma = 0$. Nehme $r \in (0,1)$ beliebig und $\varepsilon < r$. Dann gilt für $k \geq k(\varepsilon)$ wegen (3.2.2) die Beziehung

$$\begin{aligned} \|x^{k+1} - x^*\|_\varepsilon &\leq \varepsilon \|x^k - x^*\|_\varepsilon \leq \varepsilon^{k+1} \cdot a, \\ a &= \varepsilon^{-k(\varepsilon)} \prod_{\ell=0}^{k(\varepsilon)-1} \|x^\ell - x^*\|_\varepsilon, \end{aligned}$$

woraus $\lim_{k\to\infty} \frac{\|x^{k+1}-x^*\|_\varepsilon}{r^{k+1}} = 0$ folgt. Die Konvergenz ist also r-überlinear. □
Man beachte, dass wir q-überlineare Konvergenz im Fall (iii) nicht nachweisen konnten.
Bei Minimierungsproblemen suchen wir stationäre Punkte für $f : \mathbb{R}^n \to \mathbb{R}$, d.h. Nullstellen von $F(x) = \nabla f(x)$. Die Mutter aller Verfahren hierzu ist das Newton-Verfahren.

3.2.3 Definition
Es sei $F : D \subseteq \mathbb{R}^n \to \mathbb{R}^n$ differenzierbar. Das Newton-Verfahren bestimmt Iterierte x^k mit

$$x^{k+1} = x^k - F'(x^k)^{-1} F(x^k) \qquad (3.2.3)$$

Interpretation: Die Taylor-Entwicklung

$$F(x) = F(x^k) + F'(x^k)(x - x^k) + \ldots$$

wird nach dem linearen Glied abgebrochen. Dann wird die Nullstelle des *linaeren Modells*

$$M(x) = F(x^k) + F'(x^k)(x - x^k)$$

als x^{k+1} bestimmt.
Das Newton-Verfahren ist eine Iteration der Gestalt (3.2.1) mit $H(x) = x - F'(x)^{-1}F(x)$.
Durch Anwendung von Satz 3.2.2 erhalten wir

3.2.4 Satz
Sei $F; D \subseteq \mathbb{R}^n \to \mathbb{R}^n$ in einer Umgebung von $x^* \in \overset{\circ}{D}$ differenzierbar, $F(x^*) = 0$, F' stetig in x^* und und $F'(x^*)$ regulär. Dann existieren die Newton-Iterierten (3.2.3), sobald x^0 nahe genug an x^* gewählt wird und die Iterierten konvergieren r-überlinear gegen x^*.

Beweis: Wir zeigen, dass für die Iterationsfunktion $H(x) = x - F'(x)^{-1}F(x)$ gilt

$$H'(x^*) = 0.$$

Dann ist auch $\rho(H'(x^*)) = 0$ und die Behauptung folgt aus Satz 3.2.2 (iii). Nun gilt

$$\begin{aligned} H(x) - H(x^*) &= x - F'(x)^{-1}F(x) - x^* \\ &= x - F'(x)^{-1}\left(F(x^*) + F'(x^*)(x - x^*)\right. \\ &\quad \left. + F(x) - F(x^*) - F'(x^*)(x - x^*)\right) - x^*, \end{aligned}$$

also

$$\begin{aligned} \|H(x) - H(x^*)\| \leq\ & \|(I - F'(x)^{-1}F'(x^*)\| \cdot \|x - x^*\| \\ & + \|F(x) - F(x^*) - F'(x^*)(x - x^*)\| \cdot \|F'(x)^{-1}\|. \end{aligned}$$

Auf Grund der Stetigkeit von F' an der Stelle x^* ist $\lim_{x \to x^*} \|(I - F'(x)^{-1}F'(x^*)\| = 0$ und $\|F'(x)^{-1}\| \leq \beta$ in einer Umgebung von x^*. Auf Grund der Differenzierbarkeit ist $\|F(x) - F(x^*) - F'(x^*)(x - x^*)\| = o(\|x - x^*\|)$. Insgesamt ergibt sich so

$$\|H(x) - H(x^*)\| = o(\|x - x^*\|),$$

d.h. $H'(x^*) = 0$. □

Die folgenden Definitionen und Hilfsresultate sind nützlich bei einer weitergehenden Konvergenzanalyse des Newton-Verfahrens.

3.2.5 Definition
Für $F : D \subseteq \mathbb{R}^m \to \mathbb{R}^n$ notieren wir $F \in \mathrm{Lip}_\gamma(D)$, falls F auf D Lipschitzstetig ist mit Lipschitz-Konstante γ gemäß

$$\|F(x) - F(y)\| \leq \gamma \cdot \|x - y\| \text{ für alle } x, y \in D.$$

Beachte: γ hängt von den gewählten Normen in $\mathbb{R}^m$ und $\mathbb{R}^n$ ab.

3.2.6 Lemma
Sei $D \subseteq \mathbb{R}^n$ offen, konvex, $F : D \to \mathbb{R}^n, F \in \mathcal{C}^1(D), F' \in \mathrm{Lip}_\gamma(D), x, y, z \in D$. Dann gilt

(i) $\|F(y) - F(z) - F'(x)(y - z)\| \leq \frac{\gamma}{2} \cdot \|y - z\| \cdot (\|y - x\| + \|z - x\|)$

(ii) $\|F(y) - F(x) - F'(x)(y-x)\| \leq \frac{\gamma}{2} \cdot \|y-x\|^2$

Beweis: (ii) folgt aus (i) mit $z = x$.
Zu (ii): Es ist

$$F(y) - F(z) - F'(x)(y-z) = \int_0^1 \left(F'(z + t(y-z)) - F'(x)\right)(y-z)dt,$$

also

$$\begin{aligned}
&\|F(y) - F(z) - F'(x)(y-z)\| \\
&\leq \int_0^1 \|F'(z + t(y-z)) - F'(x)\| \cdot \|y-z\| dt \\
&\leq \gamma \cdot \|y-z\| \int_0^1 \|z + t(y-z) - x\| dt \\
&\leq \gamma \cdot \|y-z\| \int_0^1 t \cdot \|y-x\| + (1-t) \cdot \|z-x\| dt \\
&= \frac{\gamma}{2} \cdot \|y-z\| \cdot (\|y-x\| + \|z-x\|).
\end{aligned}$$

□

Mit dieser Vorbereitung können wir jetzt die q-quadratische Konvergenz des Newton-Verfahrens nachweisen.

3.2.7 Satz
Sei $D \subseteq \mathbb{R}^n$ offen, $F : D \to \mathbb{R}^n, F \in \mathcal{C}^1(D), F' \in \mathrm{Lip}_\gamma(D), F(x^*) = 0$ mit $x^* \in D$, $F'(x^*)$ regulär. Dann existiert $\eta > 0$, so dass im Falle $\|x^0 - x^*\| < \eta$ die Newton-Iterierten

$$x^{k+1} = x^k - F'(x^k)^{-1} F(x^k)$$

definiert sind und

$$\|x^{k+1} - x^*\| \leq d \cdot \|x^k - x^*\|^2 \leq c \cdot \|x^k - x^*\|, \quad c < 1,\ c, d > 0 \tag{3.2.4}$$

erfüllen.

Beweis: Wähle $\widetilde{\eta}$ so, dass $B_{\widetilde{\eta}}(x^*) \subseteq D$. Indem man $\widetilde{\eta}$ evtl. verkleinert, kann außderdem

$$\|F'(x)^{-1}\| \leq \sigma < \infty \text{ für alle } x \in B_{\widetilde{\eta}}(x^*))$$

vorausgesetzt werden. Wähle nun η so klein, dass $\eta \leq \widetilde{\eta}$ und $\frac{1}{2}\gamma\sigma\eta < 1$. Dann gilt:

$$\begin{aligned}
\|x^1 - x^*\| &= \|x^0 - x^* - F'(x^0)^{-1}(F(x^0) - \underbrace{F(x^*)}_{=0})\| \\
&\leq \|F'(x^0)^{-1}\| \cdot \|F'(x^0)(x^0 - x^*) - F(x^0) + F(x^*)\| \\
&\leq \sigma \cdot \frac{\gamma}{2} \cdot \|x^0 - x^*\|^2 \\
&\leq \left(\sigma \cdot \frac{\gamma}{2} \cdot \eta\right) \cdot \|x^0 - x^*\|.
\end{aligned}$$

Dies beweist (3.2.4) für $k = 0$ mit $d = \sigma\frac{\gamma}{2}$, $c = d\eta$.
Wegen $\sigma\frac{\gamma}{2}\eta < 1$ folgt insbesondere $\|x^1 - x^*\| \leq \eta$. Deshalb folgt auf genau dieselbe Weise (3.2.4) induktiv für alle k. □
Häufig steht $F'(x^k)$ nicht direkt zur Verfügung oder/und $d^k = -F'(x^k)^{-1}F(x^k)$ wird nur approximativ als Lösung des LGS

$$F'(x^k)d^k = -F(x^k)$$

bestimmt. Es ist also praktisch wichtig, *inexakte Newton-Verfahren*

$$x^{k+1} = x^k + d^k \quad \text{mit} \quad d^k \approx -F'(x^k)^{-1}F(x^k)$$

zu betrachten.

3.2.8 Satz
Unter der Voraussetzung von Satz (3.2.7) gelte für ein inexaktes Newton-Verfahren

$$x^{k+1} = x^k + d^k,$$

die Genauigkeitsforderung

$$\|F'(x^k)d^k + F(x^k)\| < c_k \cdot \|F(x^k)\|, \quad c_k < c < 1. \tag{3.2.5}$$

Mit $\|\cdot\|_*$ wird die Norm

$$\|y\|_* = \|F'(x^*)y\|, \;\; y \in \mathbb{R}^n$$

bezeichnet. Dann existiert $\eta > 0$, so dass für $x^0 \in B^*_\eta(x^*)$ (Kugel bzgl. $\|\cdot\|_*$) die Iterierten alle existieren, wobei $\alpha > 0$ und $\beta > 0$ mit $(1+\alpha)c+\beta\eta \leq c' < 1$ existieren mit

$$\|x^{k+1} - x^*\|_* \leq \left((1+\alpha)c_k + \beta\|x^k - x^*\|_*\right) \cdot \|x^k - x^*\|_* \text{ für alle } k. \tag{3.2.6}$$

Beweis: Der wesentliche Kniff bei diesem Beweis liegt in der Verwendung der Norm $\|\cdot\|_*$. Wir wählen $\widetilde{\eta} = \widetilde{\eta}(\varepsilon) > 0$ so klein, dass $B^*_{\widetilde{\eta}}(x^*) \subseteq D$, $\|F'(x^*)^{-1}F'(x)\| \leq 1+\varepsilon$, $\|F'(x)^{-1}F'(x^*)\| \leq 1+\varepsilon$ gilt für alle $x \in B^*_{\widetilde{\eta}}(x^*)$. Sei außerdem $\sigma = \max\{\|F'(x^*)\|, \|F'(x^*)^{-1}\|\}$. Dann erhalten wir für $x^k \in B^*_{\widetilde{\eta}}(x^*)$ mit der Bezeichnung $r^k = F'(x^k)d^k + F(x^k)$

$$\begin{aligned} x^{k+1} - x^* &= x^k + d^k - x^* \\ &= x^k - x^* + F'(x^k)^{-1}(r^k - F(x^k)) \\ &= F'(x^k)^{-1}\left[r^k - F(x^k) + F(x^*) + F'(x^k)(x^k - x^*)\right] \end{aligned}$$

und somit

$$\begin{aligned} & F'(x^*)(x^{k+1} - x^*) \\ (3.2.7) \quad &= F'(x^*)F'(x^k)^{-1}\left[r^k - F(x^k) + F(x^*) + F'(x^k)(x^k - x^*)\right]. \end{aligned}$$

Wir schätzen die Terme rechts ab. Nach Voraussetzung ist

$$\|r^k\| \leq c_k \cdot \|F(x^k)\|,$$

wobei

$$\begin{aligned} F(x^k) &= \int_0^1 F'(x^* + t(x^k - x^*)) \cdot (x^k - x^*)dt \\ &= \int_0^1 F'(x^* + t(x^k - x^*)) \cdot F'(x^*)^{-1} \cdot F'(x^*)(x^k - x^*)dt, \end{aligned}$$

also

$$\|F(x^k)\| \leq (1+\varepsilon) \cdot \|x^k - x^*\|_*.$$

Für den anderen Term haben wir

$$\begin{aligned} & \| - F(x^k) + F(x^*) + F'(x^k)(x^k - x^*)\| \\ &\leq \| - F(x^k) + F(x^*) - F'(x^k)(x^* - x^k)\| \\ &\leq \frac{\gamma}{2} \cdot \|x^k - x^*\|^2 \\ &\leq \sigma^2 \cdot \frac{\gamma}{2} \cdot \|x^k - x^*\|_*^2 \end{aligned}$$

Insgesamt ergibt sich so aus (3.2.7)

$$\|x^{k+1} - x^*\|_* \leq (1+\varepsilon)\left[(1+\varepsilon)c_k \cdot \|x^k - x^*\|_* + \sigma^2\frac{\gamma}{2}\|x^k - x^*\|_*^2\right]$$

$$= \left[(1+\varepsilon)^2 c_k + (1+\varepsilon)\sigma^2 \frac{\gamma}{2}\|x^k - x^*\|_*\right] \cdot \|x^k - x^*\|_*. \tag{3.2.8}$$

Durch Verkleinerung von ε und $\widetilde{\eta} = \widetilde{\eta}(\varepsilon)$ kann man erreichen, dass $(1+\varepsilon)^2 c_k + (1+\varepsilon)\sigma^2\frac{\gamma}{2}\widetilde{\eta} < 1$ gilt.
Per Induktion folgt so aus (3.2.8), dass alle Iterierten in $B^*_{\widetilde{\eta}}(x^*)$ liegen. Mit $1+\alpha = (1+\varepsilon)^2$ und $\beta = (1+\varepsilon)\sigma^3\frac{\gamma}{2}$ folgt außerdem (3.2.6) □

3.2.9 Korollar
Unter den Voraussetzungen von Satz 3.2.8 konvergiert die Folge x^k

(i) q-linear bzgl. $\|\cdot\|_*$

(ii) q-überlinear falls $\lim_{k\to\infty} c_k = 0$

(iii) mit r-Ordnung $\frac{1+\sqrt{5}}{2}$, falls $c_k \le \theta \cdot \|x^k - x^{k-1}\|$

(iv) q-quadratisch, falls $c_k \le \theta \cdot \|F(x^k)\|$,

vorausgesetzt, x^0 liegt nahe genug bei x^*.

Beweis: (i) ist gerade die Aussage von Satz 3.2.8. Auch (ii) folgt sofort aus (3.2.6). Im Fall (iii) ergibt sich aus (3.2.6)

$$\|x^{k+1} - x^*\|_* \le \left((1+\alpha)\|x^{k-1} - x^*\|_* + (\alpha+\beta)\|x^k - x^*\|_*\right) \cdot \|x^k - x^*\|_*,$$

so dass nach Lemma 3.2.6 die r-Ordnung τ mit $\tau^2 - \tau - 1 = 0$, d.h. $\tau = \frac{1+\sqrt{5}}{2}$ folgt.
Im Fall (iv) beachten wir, dass aus $F(x^k) = F(x^k) - F(x^*) = \int_0^1 F'(x^k + t(x^k - x^*))(x^k - x^*)dt$ folgt

$$\|F(x^k)\|_* \le \theta_2 \cdot \|x^k - x^*\|_*,$$

wobei θ_2 so definiert ist, dass $\|F'(x)\|_* \le \theta_2$ für alle x in $B^*_{\widetilde{\eta}}(x^*)$. Aus (3.2.6) folgt also

$$\|x^{k+1} - x^*\|_* \le \left[(1+\alpha)\theta\theta_2 + \beta\right]\|x^k - x^*\|_*^2.$$

□

Inexakte Newton-Verfahren treten praktisch in zwei Situationen, die wir als Beispiele aufführen.

3.2.10 Beispiel
Zur Lösung der Newton-Gleichung (diese ist ein LGS für d_N^k mit $x^{k+1} = x^k + d_N^k$)

$$F'(x^k)d_N^k = -F'(x^k) \tag{3.2.9}$$

wird ein Iterationsverfahren herangezogen. Es bestimmt Iterierte $d^{(i)}, i = 0, 1, \ldots$ mit Residuen

$$r^{(i)} = -F'(x^k) - F'(x^k)d^i.$$

Die Genauigkeitsforderung (3.2.5) ist also erfüllt, falls in einem Iterationsschritt

$$r^{(i)} \leq c_k \|F(x^k)\|$$

gilt, was explizit überprüft werden kann, wenn $r^{(i)}$ berechnet wird. Hierzu zwei konkrete Fälle

(i) Ist $F'(x^k)$ spd und nimmt man das CG-Verfahren zur Lösung des LGS (3.2.9), so wird $r^{(i)}$ explizit berechnet, denn das CG-Verfahren berechnet für $i = 0, 1, \ldots$ (richtige Initialisierungen werden weggelassen):

$$\begin{aligned} \alpha_i &= \frac{\|r^{(i)}\|^2}{{p^{(i)}}^T F'(x^k) p^{(i)}} \\ d^{(i+1)} &= d^{(i)} + \alpha_i p^{(i)} \\ r^{(i+1)} &= r^{(i)} - \alpha_i F'(x^k) p^{(i)} \\ \beta_i &= \frac{\|r^{(i+1)}\|^2}{\|r^{(i)}\|^2} \\ p^{(i+1)} &= p^{(i)} + \beta_i r^{(i+1)} \end{aligned}$$

(ii) Ist $D \in \mathbb{R}^{n \times n}$ der Diagonalteil von $F'(x^k)$ und $F'(x^k) = D - B$, so kann man die Jacobi-Iteration für (3.2.9)

$$d^{(i+1)} = D^{-1}\left(Bd^{(i)} - F(x^k)\right), \; i = 0, 1, \ldots$$

äquivalent umformulieren in

$$r^{(i)} = -F(x^k) - F'(x^k)d^{(i)}, \;\; d^{(i+1)} = d^{(i)} + D^{-1}r^{(i)}.$$

3.2.11 Beispiel
Steht $F'(x)$ nicht explizit zur Verfügung, so kann man die Einträge durch Differenzenquotienten mit der Schrittweite h approximieren. Wir notieren

$$DF_h(x) \in \mathbb{R}^{n \times n}$$

mit

$$(DF_h(x))_{ij} = \frac{F_i(x_1, \ldots, x_{j-1}, x_j + h, x_{j+1}, \ldots, x_n) - F(x)}{h}.$$

Statt (3.2.9) löst man nun (exakt!)

$$DF_{h^k}(x^k)d^k = -F(x^k).$$

Dann haben wir

$$\begin{aligned}
\|F'(x^k)d^k + F(x^k)\| &= \| \left(F'(x^k) - DF_{h^k}\right) d^k + \underbrace{DF_{h^k}(x^k)d^k + F(x^k)}_{=0} \| \\
&\leq \|F'(x^k) - DF_{h^k}\| \cdot \|d^k\| \\
&\leq \|F'(x^k) - DF_{h^k}\| \cdot \|DF_{h^k}^{-1}\| \cdot \|F(x^k)\| \\
&\leq c \cdot |h^k| \cdot \|F(x^k)\|,
\end{aligned}$$

wobei die letzte Ungleichung (mit einer i.d.R. unbekannten Konstanten c) aus Stetigkeitsgründen für gegnügend kleines h und x^k nahe genug an x^* gilt.
Aus Korollar 3.2.9 erhalten wir damit unter den üblichen Voraussetzungen, dass die Folge x^k gegen x^*

- q-linear konvergiert, vorausgesetzt $|h^k| \leq \tilde{h}$ für $\tilde{h}$ genügend klein
- q-überlinear konvergiert, falls $\lim_{k\to\infty} h^k = 0$
- mit r-Ordnung $\frac{1+\sqrt{5}}{2}$ konvergiert, falls $h^k \leq \theta\|x^k - x^{k-1}\|$
- q-quadratisch konvergiert, falls $h^k \leq \theta \cdot \|F(x^k)\|$.

Abschnitt 3.3

Der Satz von Dennis und Moré

Das Newton-Verfahren ist auch deshalb so wichtig, weil im Wesentlichen jedes q-überlinear konvergente Verfahren das Newton-Verfahren imitieren *muss*. Dazu formulieren wir den wichtigen Satz von Dennis und Moré, den wir mit einigen Hilfsresultaten vorbereiten.

3.3.1 Lemma
Die Folge $\{x^k\} \subseteq \mathbb{R}^n$ konvergiere q-überlinear gegen x^*, $x^k \neq x^*$ für alle k. Dann gilt

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

Beweis: Es ist

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

□

Folge: Bei überlinearer Konvergenz ist $\|x^{k+1} - x^k\| \leq \varepsilon$ ein gutes Abbruchkriterium, denn für ε klein genug ist $\|x^k - x^*\| \approx \|x^{k+1} - x^k\| \leq \varepsilon$.

3.3.2 Lemma
Es sei $F : D \subseteq \mathbb{R}^n \to \mathbb{R}^n$, D offen, $x^* \in D, F(x^*) = 0$ und F sei differenzierbar in x^* mit $F'(x^*)$ regulär. Dann existieren $\delta, \beta > 0$ mit der Eigenschaft, dass für alle $x \in D$ mit $\|x - x^*\| \leq \delta$ gilt

$$\beta \cdot \|x - x^*\| \leq \|F(x)\|.$$

Beweis: Da F differenzierbar in x^*, gilt

$$\|F(x) - F(x^*) - F'(x^*)(x - x^*)\| = o(\|x - x^*\|).$$

Wir haben damit

$$\|F'(x^*)(x - x^*)\| \leq \|F(x) - \underbrace{F(x^*)}_{=0} - F'(x^*)(x - x^*)\| + \|F(x)\| \tag{3.3.1}$$

und

$$\|F'(x^*)(x-x^*)\| \geq \frac{1}{\|F'(x^*)^{-1}\|} \cdot \|x-x^*\|.$$

Wähle nun z.B. δ so klein, dass für alle x mit $\|x-x^*\| \leq \delta$

$$\|F(x)-F(x^*)-F'(x^*)(x-x^*)\| \leq \frac{1}{2\cdot\|F'(x^*)^{-1}\|} \cdot \|x-x^*\|$$

gilt.
Dann folgt für diese x insgesamt aus (3.3.1)

$$\frac{1}{2\cdot\|F'(x^*)^{-1}\|} \cdot \|x-x^*\| \leq \|F(x)\|.$$

Man nehme also $\beta = 1/(2\cdot\|F'(x^*)^{-1}\|)$. □

Folge: Auch das Abbruchkriterium $\|F(x^k)\| < \varepsilon$ ist ein vernünftiges Abbruckriterium, denn es impliziert $\|x^k - x^*\| < (1/\beta)\varepsilon$, mit allerdings in der Regel unbekanntem β.

3.3.3 Satz (Dennis & Moré)
Sei $F : \mathbb{R}^n \to \mathbb{R}^n, F \in \mathcal{C}^1(\mathbb{R}^n)$ mit Nullstelle x^* mit $F'(x^*)$ regulär. Die Folge $\{x^k\}$ konvergiere gegen x^*. Dann sind äquivalent

(i) $x^k \to x^*$ q-überlinear

(ii) $\|F(x^k) + F'(x^k)(x^{k+1}-x^k)\| = o(\|x^{k+1}-x^k\|)$

(iii) $\|F(x^k) + F'(x^*)(x^{k+1}-x^k)\| = o(\|x^{k+1}-x^k\|)$

Beweis: (ii) ist äquivalent zu (iii), denn $\|F'(x^k)-F'(x^*)\| \cdot \|x^{k+1}-x^k\| = o(\|x^{k+1}-x^k\|)$ auf Grund der Stetigkeit von F'.
„(iii) ⇒ (i)“

$$\begin{aligned} F(x^{k+1}) &= F(x^{k+1}) - F(x^k) - F'(x^*)(x^{k+1}-x^k) \\ &\quad + F(x^k) + F'(x^*)(x^{k+1}-x^k) \\ &= \int_0^1 \left(F'(x^k + t(x^{k+1}-x^k)) - F'(x^*)\right)(x^{k+1}-x^k)dt \\ &\quad + F(x^k) + F'(x^*)(x^{k+1}-x^k). \end{aligned}$$

Nun ist aber

$$\left\| \int_0^1 F'(x^k + t(x^{k+1}-x^k)) - F'(x^*)dt \right\| \leq \int \|F'(x^k + t(x^{k+1}-x^k)) - F'(x^*)\|dt$$

$$=: \varepsilon_k \ \to 0 \ (k \to \infty)$$

und damit

$$\begin{aligned} \|F(x^{k+1})\| &\leq \varepsilon_k \cdot \|x^{k+1} - x^k\| + \|F(x^k) + F'(x^*)(x^{k+1} - x^k)\| \\ &= o(\|x^{k+1} - x^k\|). \end{aligned}$$

Wegen Lemma 3.3.2 folgt so

$$\|x^{k+1} - x^*\| = o(\|x^{k+1} - x^k\|)$$

und wegen $\|x^{k+1} - x^k\| \leq \|x^{k+1} - x^*\| + \|x^* - x^k\|$ schließlich auch

$$\|x^{k+1} - x^*\| = o(\|x^k - x^*\|).$$

„(i) $\Rightarrow$ (ii)“: Es ist

$$\|F(x^{k+1}) - F(x^*)\| = \|F(x^{k+1})\| \leq \int_0^1 \|F'(x^* + t(x^{k+1} - x^*))\| dt \cdot \|x^{k+1} - x^*\|.$$

Wegen $\lim_{k\to\infty} x^k = x^*$ und F' stetig existiert $L > 0$ mit

$$\int_0^1 \|F'(x^* + t(x^{k+1} - x^*))\| dt \leq L \text{ für alle } k.$$

Wir haben also unter Verwendung von Lemma 3.3.1

$$\|F(x^{k+1})\| \leq L\|x^{k+1} - x^*\| = L \cdot \underbrace{\frac{\|x^{k+1} - x^*\|}{\|x^k - x^*\|}}_{\to 0} \cdot \underbrace{\frac{\|x^k - x^*\|}{\|x^{k+1} - x^k\|}}_{\to 1} \cdot \|x^{k+1} - x^k\|.$$

Hieraus folgt

$$\|F(x^{k+1})\| = o(\|x^{k+1} - x^k\|)$$

und damit

$$\begin{aligned} &\|F(x^k) + F'(x^*)(x^{k+1} - x^k)\| \\ &\leq \|F(x^{k+1})\| + \int_0^1 \|F'(x^k + t(x^{k+1} - x^k)) - F'(x^*)\| dt \cdot \|x^{k+1} - x^k\| \\ &= o(\|x^{k+1} - x^k\|), \end{aligned}$$

denn $\lim_{k\to\infty} \int_0^1 \|F'(x^k + t(x^{k+1} - x^k) - F'(x^*)\| dt = 0$, da F' stetig. □

Ist $F'(x^*)$ regulär, so ist $\|F'(x)^{-1}\| \leq \beta$ in einer Umgebung von x^*. Es gilt deshalb das folgende

3.3.4 Korollar

Unter den Voraussetzungen von Satz 3.3.3 sind auch äquivalent

(i) $x^k \to x^*$ q-überlinear

(ii) $\|F'(x^k)^{-1} \cdot F(x^k) + (x^{k+1} - x^k)\| = o(\|x^{k+1} - x^k\|)$

(iii) $\|F'(x^*)^{-1} \cdot F(x^k) + (x^{k+1} - x^k)\| = o(\|x^{k+1} - x^k\|)$

Teil (ii) des Korollars besagt, dass q-überlinear konvergente Verfahren das Newton-Verfahren immer besser imitieren müssen, denn in $x^{k+1} = x^k + d^k$ approximiert d^k die Newton-Korrektur $-F'(x^k)^{-1} \cdot F(x^k)$ in dem Sinne

$$\lim_{k\to\infty} \|\frac{1}{\|d^k\|} \cdot F'(x^k)^{-1} \cdot F(x^k) + \frac{1}{\|d^k\|} d^k\| \;\to\; 0 \;\; (k \to \infty).$$

Daraus folgt einerseits $\lim_{k\to\infty} \|\frac{1}{\|d^k\|} \cdot F'(x^k)^{-1} \cdot F(x^k)\| = 1$, d.h. die Länge von d^k konvergiert gegen die Länge der Newton-Korrektur. Zusätzlich konvergiert auch die *Richtung* (i.S. der normierten Vektoren) von d^k gegen die Richtung der Newton-Korrektur.
In der angegebenen Formulierung stellt der Satz von Dennis und Moré keinen direkten Zusammenhang mit unserer Analyse der lokalen Konvergenz von inexakten Newton-Verfahren aus Satz 3.2.8 her. Das folgende Korollar leistet dies.

3.3.5 Korollar

Unter den Voraussetzungen von Satz 3.3.3 sind auch äquivalent

(i) $x^k \to x^*$ q-überlinear

(ii) $\|F'(x^k)^{-1} \cdot F(x^k) + (x^{k+1} - x^k)\| = o(\|F(x^k)\|)$

(iii) $\|F'(x^*)^{-1} \cdot F(x^k) + (x^{k+1} - x^k)\| = o(\|F(x^k)\|)$

Beweis: Wir zeigen, dass (ii) und (iii) jeweils äquivalent sind zu der entsprechenden Aussage (ii) und (iii) aus Korollar 3.3.4. Wir starten mit (ii) und nehmen an, es gilt

$$\|F'(x^k)^{-1} \cdot F(x^k) + (x^{k+1} - x^k)\| = o(\|F(x^k)\|).$$

In einer Umgebung von x^* ist $\|F'(x)\| \le \beta$, $\|F'(x)^{-1}\| \le \beta$. Dann gilt für x^k in dieser Umgebung

$$\frac{1}{\beta}\|x^{k+1} - x^k\| - \|F(x^k)\| = o(\|F(x^k)\|).$$

Für k groß genug ist $o(\|F(x^k)\|) \geq -(1/(2\beta)) \cdot \|F(x^k)\|$, so dass für diese k gilt

$$\|F(x^k)\| \leq 2\beta \|x^{k+1} - x^k\|.$$

Daraus ergibt sich $o(\|F(x^k)\|) = o(\|x^{k+1} - x^k\|)$ und damit (ii) aus Korollar 3.3.4.
Für die umgekehrte Richtung leiten wir in analoger Weise aus

$$\|F'(x^k)^{-1} \cdot F(x^k) + (x^{k+1} - x^k)\| = o(\|x^{k+1} - x^k\|)$$

für alle gegnügend großen k mit $o(\|x^{k+1} - x^k\|) \leq (1/2) \cdot (\|x^{k+1} - x^k\|)$ die Beziehung

$$\|x^{k+1} - x^k\| \leq 2\beta \|F(x^k)\|$$

her. Es ist also wieder $o(\|F(x^k)\|) = o(\|x^{k+1} - x^k\|)$, und es ergibt sich (ii) des vorliegenden Korollars.
Für den Beweis von (iii) geht alles analog. □
Das Korollar gilt auch wieder, wenn wir auf den linken Seiten 'mit $F'(x^k)$ bzw. $F'(x^*)$ multiplizieren', was wir jetzt aber nicht mehr extra festhalten.

Abschnitt 3.4

Newton-Verfahren zur Optimierung

Ist $f : \mathbb{R}^n \to \mathbb{R}^n \in \mathcal{C}^2(\mathbb{R}^n)$ und x^* ein Minimum von f, so kann man wegen $\nabla f(x^*) = 0$ die Stelle x^* dadurch approximieren, dass man das Newton-Verfahren auf $F(x) = \nabla f(x)^T$ anwendet.
Die entscheidende Frage ist, wie sich ein solches Newton-Verfahren in ein (global konvergentes) Verfahren zur Optimierung einbauen lässt.
In diesem Abschnitt zeigen wir, dass unter gewissen Bedingungen die Newton-Richtung $d = -\nabla^2 f(x)^{-1} \cdot \nabla f(x)^T$ eine Abstiegsrichtung ist und dass nahe am Minimum x^* bereits die Schrittweite $t = 1$ die Armijo-Bedingung erfüllt. Dasselbe gilt sogar für alle q-überlinear konvergenten Verfahren.

3.4.1 Lemma
Für $\nabla^2 f(x)$ spd ist die Newton-Richtung $d = -\nabla^2 f(x)^{-1} \cdot \nabla f(x)^T$ eine Abstiegsrichtung für f in d.

Beweis: Beispiel 2.1.3 □

Ist $f \in \mathcal{C}^2(\mathbb{R}^n)$, so ist $d = -\nabla^2 f(x)^{-1} \nabla f(x)^T$ also auch eine Abstiegsrichtung, sobald x nahe genug an einem x^* liegt mit $\nabla^2 f(x^*)$ spd .

3.4.2 Lemma
Sei $f : \mathbb{R}^n \to \mathbb{R}, f \in \mathcal{C}^2(\mathbb{R}^n)$ und $\nabla^2 f(x^*)$ spd. Dann existiert $\delta > 0, \alpha > 0$, so dass

$$d^T \nabla^2 f(x) d \geq \alpha \cdot \|d\|^2 \text{ für alle } x \in B_\delta(x^*) \text{ und alle } d \in \mathbb{R}^n$$

Beweis: Die Funktion

$$g(x) = \min_{d \in \mathbb{R}^n, \|d\|=1} \left\{ d^T \nabla^2 f(x) d \right\}$$

ist stetig in x mit $g(x^*) > 0$. □

3.4.3 Satz
Sei $f : \mathbb{R}^n \to \mathbb{R}, f \in \mathcal{C}^2(\mathbb{R}^n), x^* \in \mathbb{R}^n$ mit $\nabla f(x^*) = 0$ und $\nabla^2 f(x^*)$ spd (d.h. x^* ist lokales Minimum). Weiter sei $\sigma \in (0, \frac{1}{2})$ fest. Dann existiert $\delta > 0$, so dass für $x \in B_\delta(x^*)$ die Hesse-Matrix $\nabla^2 f(x)$ spd ist und

$$f(x + d_N) < f(x) + \sigma \nabla f(x) d_N, \quad d_N = -\nabla^2 f(x)^{-1} \cdot \nabla f(x)^T$$

Beweis: $\nabla f(x)$ spd für δ klein genug, folgt direkt aus Lemma 3.4.2. Für $x \in B_\delta(x^*)$ mit einem solchen δ erhalten wir dann mit einer Taylor-Entwicklung 2. Ordnung für $\varphi(t) = f(x + td_N)$

$$f(x+d_N) = f(x) + \nabla f(x)d_N + \frac{1}{2}d_N^T \nabla^2 f(\xi) d_N, \quad \xi = x + \vartheta \cdot d_N, \vartheta \in (0,1). \tag{3.4.1}$$

Für den quadratischen Term gilt

$$d_N^T \nabla^2 f(\xi) d_N = -\nabla f(x) d_N + d_N(\nabla^2 f(\xi) - \nabla^2 f(x)))d_N$$

Da $\nabla^2 f$ stetig ist und $d_N = -\nabla^2 f(x)^{-1} \cdot \nabla f(x)^T \to 0$ für $x \to x^*$, können wir zu jedem $\varepsilon > 0$ die Zahl $\delta > 0$ so klein wählen, dass für $x \in B_\delta(x^*)$

$$|d_N^T \left(\nabla^2 f(\xi) - \nabla^2 f(x)\right) d_N| \leq \varepsilon \cdot \|d_N\|^2 \tag{3.4.2}$$

gilt. Indem wir δ eventuell verkleinern, erhalten wir nach Lemma 3.4.2 ein $\alpha > 0$ mit

$$\alpha \|d_N\|^2 \leq d_N^T \nabla^2 f(x) d_N = -\nabla f(x) d_N \quad \text{für alle } x \in B_\delta(x^*). \tag{3.4.3}$$

Durch Einsetzen von (3.4.2) erhalten wir für diese x also

$$\begin{aligned} f(x+d_N) &\leq f(x) + \frac{1}{2}\nabla f(x) d_N + \frac{1}{2}\varepsilon \cdot \|d_N\|^2 \\ &\leq f(x) + \frac{1}{2}\left(1 - \frac{\varepsilon}{\alpha}\right) \nabla f(x) d_N. \end{aligned}$$

Für gegebenes $\sigma \in (0, \frac{1}{2})$ muss man also $\varepsilon < \alpha(1-2\sigma)$ wählen. □

Allgemeiner gilt, dass nicht nur das Newton-Verfahren, sondern jedes q-überlinear konvergente Verfahren auf Abstiegsrichtungen beruht, sobald sich die Iterierten nahe genug an der Lösung befinden. Genauer formuliert dies das folgende Resultat.

3.4.4 Lemma
Sei $f : \mathbb{R}^n \to \mathbb{R}, f \in \mathcal{C}^2(\mathbb{R}^n), x^* \in \mathbb{R}^n$ mit $\nabla f(x^*) = 0$, $\nabla^2 f(x^*)$ spd. Für $x \in \mathbb{R}^n$ sei $d = d(x)$ eine Richtung mit

$$\|\nabla^2 f(x) d + \nabla f(x)^T\| \leq \varepsilon \cdot \|\nabla f(x)\|.$$

Dann ist d eine Abstiegsrichtung für f in x, vorausgesetzt, $\|x - x^*\|$ und ε sind genügend klein.

Beweis: Es ist

$$\begin{aligned}\nabla f(x)d &= \nabla f(x)\cdot\nabla^2 f(x)^{-1}\left(\nabla^2 f(x)d+\nabla f(x)^T\right)-\nabla f(x)\nabla^2 f(x)^{-1}\nabla f(x)^T\\ &\leq \|\nabla f(x)\|\cdot\|\nabla^2 f(x)^{-1}\|\cdot\varepsilon\cdot\|\nabla f(x)\|-\nabla f(x)\nabla^2 f(x)^{-1}\nabla f(x)^T.\end{aligned}$$

Nach Lemma 3.4.2 existieren $\alpha,\delta>0$ mit

$$\nabla f(x)\nabla^2 f(x)^{-1}\nabla f(x)^T\geq\alpha\cdot\|\nabla f(x)\|^2 \text{ für } x\in B_\delta(x^*).$$

Indem wir δ evtl. weiter verkleinern, gilt außerdem $\|\nabla^2 f(x)^{-1}\|\leq\beta$ für alle $x\in B_\delta(x^*)$. Für $\varepsilon<\alpha/\beta$ ist also $\nabla f(x)d<0$ für $x\in B_\delta(x^*)$. □

3.4.5 Korollar

Sei $f:\mathbb{R}^n\to\mathbb{R}, f\in\mathcal{C}^2(\mathbb{R}^n), x^*\in\mathbb{R}^n$ mit $\nabla f(x^*)=0$, $\nabla^2 f(x^*)$ spd. Für $x\in\mathbb{R}^n$ sei $d=d(x)$ eine Richtung mit

$$\|\nabla^2 f(x)d+\nabla f(x)^T\|=o(\|d\|) \text{ für } x\to x^*.$$

Dann ist d eine Abstiegsrichtung für f in x, vorausgesetzt, $\|x-x^*\|$ ist genügend klein.

Beweis: Es existiert ein $\delta>0$ so dass für $x\in B_\delta(x^*)$ die Hesse-Matrix $\nabla^2 f(x)$ positiv definit ist und

$$\|\nabla^2 f(x)^{-1}\|\leq\beta \text{ für } x\in B_\delta(x^*).$$

Aus der Voraussetzung erhalten wir mit der Dreiecksungleichung

$$\frac{1}{\beta}\|d\|-\|\nabla f(x)\|\leq\|\nabla^2 f(x)d+\nabla f(x)\|=o(\|d\|),$$

woraus sich

$$\frac{1}{2\beta}\cdot\|d\|\leq\|\nabla f(x)\|$$

für genügend kleines $\|d\|$, also genügend kleines $\|x-x^*\|$ ergibt. Für jedes $\varepsilon>0$ existiert damit $\eta>0$, so dass wegen der Voraussetzung

$$\|\nabla^2 f(x)d+\nabla f(x)\|\leq\varepsilon\|\nabla f(x)\| \text{ für } x\in B_\eta(x^*)$$

gilt, d.h. wir können Lemma 3.4.4 anwenden. □

Interpretation: Konvergiert eine Folge $\{x^k\}$ q-überlinear gegen x^*, so sind die Richtungen $d^k=x^{k+1}-x^k$ für große k alle Abstiegsrichtungen für f in x^k, denn auf Grund des Satzes von Dennis und Moré (Korollar 3.3.4) ist dann $\|\nabla^2 f(x^k)d^k+\nabla f(x^k)^T\|=o(\|d^k\|)$.
Als nächstes weisen wir in Analogie zu Satz 3.4.3 nach, dass auch allgemein bei q-überlinear konvergenten Verfahren bei der Wahl $\sigma<1/2$ die Armijo-Schrittweite $t=1$ erreicht wird, vorausgesetzt, x liegt bereits nahe genug an x^*.

3.4.6 Satz
Sei $f : \mathbb{R}^n \to \mathbb{R}, f \in \mathcal{C}^2(\mathbb{R}^n), x^* \in \mathbb{R}^n$ mit $\nabla f(x^*) = 0$ und $\nabla^2 f(x^*)$ spd. Weiter sei $\sigma \in (0, \frac{1}{2})$ fest und zu x werde eine Suchrichtung $d_{aN} = d_{aN}(x)$ verwendet mit

$$\|\nabla^2 f(x) d_{aN} + \nabla f(x)^T\| = o(\|d_{aN}\|) \text{ für } x \to x^*. \tag{3.4.4}$$

(approximative Newton-Richtung). Dann existiert $\delta > 0$, so dass für $x \in B_\delta(x^*)$ die Hesse-Matrix $\nabla^2 f(x)$ spd ist, d_{aN} eine Abstiegsrichtung ist und

$$f(x + d_{aN}) \leq f(x) + \sigma \nabla f(x) d_{aN}. \tag{3.4.5}$$

Beweis: Wir wählen $\delta > 0$ zunächst einmal wenigstens so klein, dass $\nabla^2 f(x)$ spd ist für $x \in B_\delta(x^*)$ und dass d_{aN} eine Abstiegsrichtung ist für $x \in B_\delta(x^*)$ (Korollar 3.4.5).
Zum Beweis der zentralen Aussage (3.4.5) verwenden wir wieder

$$f(x + d_{aN}) = f(x) + \nabla f(x) d_{aN} + \frac{1}{2} d_{aN}^T \nabla^2 f(\xi) d_{aN}, \tag{3.4.6}$$

wobei wir diesmal den letzten Term ausdrücken als

$$\begin{aligned} d_{aN}^T \nabla^2 f(\xi) d_{aN} &= -\nabla f(x) d_{aN} + \left(d_{aN}^T \nabla^2 f(\xi) + \nabla f(x)\right) d_{aN} \\ &= -\nabla f(x) d_{aN} + d_{aN}^T \left(\nabla^2 f(\xi) - \nabla^2 f(x)\right) d_{aN} \\ &\quad + \left(d_{aN}^T \nabla^2 f(x) + \nabla f(x)\right) d_{aN}. \end{aligned} \tag{3.4.7}$$

Nach Lemma 3.4.2 existiert, evtl. nach Verkleinerung von δ, eine Zahl $\alpha > 0$ mit

$$d^T \nabla^2 f(x) d \geq \alpha \cdot \|d\|^2 \text{ für } x \in B_\delta(x^*),\ d \in \mathbb{R}^n. \tag{3.4.8}$$

Auf Grund der Stetigkeit von $\nabla^2 f(x)$ existiert zu jedem $\varepsilon > 0$ ein eventuell noch kleineres $\delta > 0$, so dass

$$|d_{aN}^T \left(\nabla^2 f(\xi) - \nabla^2 f(x)\right) d_{aN}| \leq \varepsilon \cdot \|d_{aN}\|^2 \text{ für } x \in B_\delta(x^*).$$

Indem wir δ eventuell nochmals verkleinern, folgt wegen der Voraussetzung (3.4.4) auch

$$\|\nabla^2 f(x) d_{aN} + \nabla f(x)^T\| \leq \varepsilon \cdot (\|d_{aN}\|).$$

Somit ergibt sich aus (3.4.7)

$$\begin{aligned} d_{aN}^T \nabla^2 f(\xi) d_{aN} &\leq -\nabla f(x) d_{aN} + \varepsilon \cdot \|d_{aN}\|^2 + \varepsilon \cdot \|d_{aN}\|^2 \\ &= -\nabla f(x) d_{aN} + 2\varepsilon \cdot \|d_{aN}\|^2. \end{aligned} \tag{3.4.9}$$

Wir müssen noch zeigen, dass $\|d_{aN}\| = \mathcal{O}(|\nabla f(x) d_{aN}|)$. Dazu beachten wir

$$\nabla f(x) d_{aN} = \left(\nabla f(x) + (\nabla^2 f(x) d_{aN})^T\right) d_{aN} - d_{aN}^T \nabla^2 f(x) d_{aN},$$

also für $x \in B_\delta(x^*)$

$$\begin{aligned} |\nabla f(x) d_{aN}| &\geq |d_{aN}^T \nabla^2 f(x) d_{aN}| - \|\nabla f(x) + (\nabla^2 f(x) d_{aN})^T\| \cdot \|d_{aN}\| \\ &\geq \alpha \|d_{aN}\|^2 - \varepsilon \|d_{aN}\|^2. \end{aligned}$$

Hierin ist $|\nabla f(x) d_{aN}| = -\nabla f(x) d_{aN}$. Indem wir ε eventuell verkleinern, können wir $\alpha - \varepsilon > 0$ annehmen und erhalten

$$\|d_{aN}\|^2 \leq \frac{1}{\alpha - \varepsilon} \cdot (-\nabla f(x) d_{aN}).$$

Eingesetzt in die Taylorentwicklung (3.4.6) unter Berücksichtigung von (3.4.9) erhalten wir so

$$f(x + d_{aN}) \leq f(x) + \left(1 + \frac{1}{2}\left(-1 - \frac{2\varepsilon}{\alpha - \varepsilon}\right)\right) \nabla f(x) d_{aN}.$$

Man wähle für $\sigma \in (0, \frac{1}{2})$ also ε (und damit δ) so klein, dass $\frac{2\varepsilon}{\alpha - \varepsilon} < 1 - 2\sigma$). □

Abschnitt 3.5

Globalisierung des Newton-Verfahrens

Ziel ist es jetzt, das Newton-Verfahren in ein global konvergentes Abstiegsverfahren einzubauen. Dies ist nicht trivial, denn die Newton-Richtung ist nicht immer eine Abstiegsrichtung.

Bemerkung: Der Begriff *Globalisierung* eines Verfahrens bedeutet, dass man es so modifiziert, dass globale Konvergenz vorliegt, *nicht notwendig aber* Konvergenz gegen ein globales Minimum!

Zur Vorbereitung betrachten wir ein Abstiegsverfahren, welches nur den Gradienten benutzt: das *Verfahren des steilsten Abstiegs* oder auch *Gradientenverfahren.*

3.5.1 Algorithmus (Gradientenverfahren)
wähle $x^0 \in \mathbb{R}^n, \sigma \in (0,1), \beta \in (0,1), \varepsilon > 0$
for $k = 0, 1, \ldots$ **do**
  $d^k = -\nabla f(x^k)^T$
  if $\|d^k\| < \varepsilon$ **then**
    STOP
  else
    bestimme $t^k = \max\{\beta^\ell, \ \ell = 0, 1, \ldots : f(x^k + \beta^\ell d^k) \leq f(x^k) + \sigma\beta^\ell \nabla f(x^k) d\}$
    {Armijo-Schrittweite}
    $x^{k+1} = x^k + t^k d^k$
  end if
end for

3.5.2 Bemerkung
Man sagt: Der Algorithmus *akzeptiert* die Schrittweite t^k, wenn er denn diese Schrittweite verwendet.

Da wir die nicht-notwendig effiziente Armijo-Schrittweiten verwenden, ist die Konvergenz dieses Verfahrens nicht durch Satz 2.1.7 gesichert. Zur Vorbereitung einer Analyse formulieren wir zunächst folgendes Hilfsresultat.

3.5.3 Lemma
Es sei $f : \mathbb{R}^n \to \mathbb{R}, f \in \mathcal{C}^1(\mathbb{R}^n), x, d \in \mathbb{R}^n, \{x^k\}, \{d^k\} \subseteq \mathbb{R}^n$ mit $\lim_{k\to\infty} x^k = x, \lim_{k\to\infty} d^k = d$ und $\{t^k\} \subseteq \mathbb{R}$ mit $\lim_{k\to\infty} t^k = 0$. Dann gilt

$$\lim_{k\to\infty} \frac{f(x^k + t^k d^k) - f(x^k)}{t^k} = \nabla f(x)d.$$

Beweis: Nach dem Mittelwertsatz ist

$$\frac{f(x^k + t^k d^k) - f(x^k)}{t^k} = \nabla f(x^k + \vartheta_k t^k d^k)d^k, \ \vartheta_k \in (0, 1).$$

Für $k \to \infty$ konvergiert $x^k + \vartheta_k t^k d^k$ gegen x. □

3.5.4 Satz
Sei $f : \mathbb{R}^n \to \mathbb{R}, f \in \mathcal{C}^1(\mathbb{R}^n)$. Dann ist jeder Häufungspunkt der vom Gradientenverfahren (Algorithmus 3.5.1) erzeugten Folge $\{x^k\}$ ein stationärer Punkt von f.

Beweis: Sei x^* ein Häufungspunkt von $\{x^k\}$. Sei $\lim_{i\to\infty} x^{k_i} = x^*$. Dann gilt $\lim_{i\to\infty} f(x^{k_i}) = f(x^*)$, und wegen $\{f(x^k)\} \searrow$ sogar $\lim_{k\to\infty} f(x^k) = f(x^*)$. Insbesondere folgt $\lim_{k\to\infty} f(x^{k+1}) - f(x^k) = 0$ und wegen

$$f(x^k) - f(x^{k+1}) \geq \sigma t^k \cdot \|\nabla f(x^k)\|^2 \geq 0$$

folgt $\lim_{k\to\infty} t^k \cdot \|\nabla f(x^k)\| = 0$. Angenommen, es wäre x^* kein stationärer Punkt, also $\|\nabla f(x^*)\| \neq 0$. Dann folgt $\lim_{i\to\infty} t^{k_i} = 0$, d.h.

$$\lim_{i\to\infty} \ell_{k_i} = \infty, \quad \ell_{k_i} \text{ Exponent in Armijo-Schritt: } t^{k_i} = \beta^{\ell_{k_i}}.$$

Also gilt für alle i

$$f(x^{k_i} + \beta^{\ell_{k_i}-1} d^{k_i}) > f(x^{k_i}) + \sigma\beta^{\ell_{k_i}-1}\nabla f(x^{k_i})d^{k_i},$$

und damit für $i \to \infty$ nach Lemma 3.5.3

$$\nabla f(x^*) \cdot d \geq \sigma \nabla f(x^*) \cdot d$$

mit $d = \lim_{i\to\infty} d^{k_i} = -\nabla f(x^*)^T$, ein Widerspruch, da $\nabla f(x^*) \neq 0$. □

Im Beweis zu Satz 3.5.4 haben wir folgendes allgemeinere, für später wichtige Resultat bewiesen.

3.5.5 Korollar
Sei $f : \mathbb{R}^n \to \mathbb{R}, f \in \mathcal{C}^1(\mathbb{R}^n)$. Die Iterierten $\{x^k\}$ werden gemäß

$$x^{k+1} = x^k + t^k d^k, \quad k = 0, 1, 2, \ldots$$

berechnet mit $t^k > 0$. Es gelte

$$\begin{aligned} f(x^{k+1}) &\leq f(x^k), \quad k = 0, 1, 2, \ldots \\ \lim_{i \to \infty} x^{k_i} &= x^* \\ d^{k_i} &= -\nabla f(x^{k_i}) \end{aligned}$$

und für alle i werde t^{k_i} als Armijo-Schrittweite bestimmt.
Dann ist $\nabla f(x^*) = 0$.

3.5.6 Beispiel
In den Übungen haben wir das Verfahren des steilsten Abstiegs für

$$f(x) = \frac{1}{2} x^T Q x + c^T x + \gamma, \quad Q \in \mathbb{R}^{n \times n} \text{ spd}$$

analysiert bei Verwendung der optimalen Schrittweite, welche die 1-dimensionale Minimierungsaufgabe

$$\text{minimiere } f(x + td)$$

löst. Wir hatten auch gesehen: Bei geeigneter Wahl des Startvektors x^0 gilt für die Konvergenz

$$\|x^{2k} - x^*\| = \left(\frac{\lambda_{\max} - \lambda_{\min}}{\lambda_{\max} + \lambda_{\min}} \right)^{2k} \|x^0 - x^*\|,$$

$\lambda_{\min}, \lambda_{\max}$ kleinster bzw. größter Eigenwert von Q. Dies ist eine *langsame* Konvergenz.
Es ist $\nabla f(x) = x^T Q + c^T, \nabla^2 f(x) = Q$, so dass sich als Newton-Richtung

$$d = -Q^{-1}(Qx + c)$$

ergibt. Die Newton-Iterierte (nach der ersten Iteration)

$$\begin{aligned} x^1 = x^0 + \nabla^2 f(x^0)^{-1}(-\nabla f(x^0)^T) &= x^0 - Q^{-1}(Qx^0 + c) \\ &= -Q^{-1}c \end{aligned}$$

ist dagegen bereits die Minimalstelle von f, d.h. die Konvergenz des Newton-Verfahrens ist besonders schnell.

Das Beispiel zeigt, dass es nützlich ist, Newton-Schritte – sofern möglich – in ein Abstiegsverfahren einzubauen.
Dazu formulieren wir den folgenden Algorithmus:

3.5.7 Algorithmus (globalisiertes Newton-Verfahren)
wähle $x^0 \in \mathbb{R}^n, \rho > 0, p > 2, \beta \in (0,1), \sigma \in (0, \frac{1}{2}), \varepsilon > 0$
for $k = 0, 1, \ldots$ **do**
  if $\|\nabla f(x^k)\| \leq \varepsilon$ **then**
    STOP
  else
    löse $\nabla^2 f(x^k) d^k = -\nabla f(x^k)^T$ {Newton-Gleichung}
    if $\nabla^2 f(x^k)$ singulär oder $\nabla f(x^k) d^k > -\rho \|d^k\|^p$ **then**
      $d^k = -\nabla f(x^k)^T$ {steilster Abstieg}
    end if
    bestimme $t^k = \max\{\beta^\ell, \ell = 0, 1, \ldots : f(x^k + \beta^\ell d^k) \leq f(x^k) + \sigma \beta^\ell \nabla f(x^k) d^k\}$
    {Armijo-Schrittweite}
    setze $x^{k+1} = x^k + t^k d^k$
  end if
end for

3.5.8 Bemerkung
Man sagt: Der Algorithmus *akzeptiert* die Newton-Richtung, wenn er sie denn verwendet.

In dem Algorithmus werden nicht-zufriedenstellende Newton-Richtungen d^k (wenn sie nämlich $\nabla f(x^k) d^k > -\rho \|d^k\|^p$ erfüllen) durch steilste Abstiegsrichtungen ersetzt. Dass die gewählte Strategie vernünftig ist, klärt der Beweis des folgenden Satzes.

3.5.9 Satz
Sei $f : \mathbb{R}^n \to \mathbb{R}, f \in \mathcal{C}^2(\mathbb{R}^n)$. Dann ist jeder Häufungspunkt x^* der vom globalisierten Newton-Verfahren (Algorithmus 3.5.7) erzeugten Folge $\{x^k\}$ ein stationärer Punkt von f.

Beweis: Sei $x^* = \lim_{i \to \infty} x^{k_i}$ ein Häufungspunkt mit zugehöriger konvergenter Teilfolge $\{x^{k_i}\}$. Falls $d^{k_i} = -\nabla f(x^{k_i})^T$ für unendlich viele i, so ist x^* stationärer Punkt wegen Korollar 3.5.5. Wir nehmen also ab jetzt an, dass

$$\nabla^2 f(x^{k_i}) d^{k_i} = -\nabla f(x^{k_i})^T \text{ für alle } i \geq i_0. \tag{3.5.1}$$

Wie im Beweis zum vorangegangenen Satz folgt aus $\lim_{k \to \infty} f(x^{k+1}) - f(x^k) = 0$ die Beziehung

$$\lim_{i \to \infty} t^{k_i} \nabla f(x^{k_i}) d^{k_i} = 0.$$

Wir nehmen an, dass x^* kein stationärer Punkt ist, also $\nabla f(x^*) \neq 0$ und werden einen Widerspruch herleiten. Wenn wir zeigen können, dass

$$t^{k_i} \geq \bar{t} > 0 \text{ für eine Teilfolge der } t^{k_i} \tag{3.5.2}$$

gilt, so folgt (wir bezeichnen die Teilfolge wieder mit t^{k_i}) $\lim_{i\to\infty} \nabla f(x^{k_i}) d^{k_i} = 0$, was wegen der Bedingung

$$\nabla f(x^{k_i}) d^{k_i} \leq -\rho \|d^{k_i}\|^p$$

aber $\lim_{i\to\infty} d^{k_i} = 0$ zur Folge hätte. Dies kann aber nicht sein, da (3.5.1) für $i \to \infty$ die Beziehung

$$\nabla^2 f(x^*) \cdot 0 = -\nabla f(x^*)^T \neq 0$$

liefern würde.
Zur Erledigung des Widerspruchsbeweises müssen wir also noch (3.5.2) zeigen. Dazu stellen wir zunächst fest, dass 0 kein Häufungspunkt der Folge $\{d^{k_i}\}$ ist (Beweis wie gerade eben). Da für alle $i \geq i_0$ außerdem

$$-\|\nabla f(x^{k_i})\| \cdot \|d^{k_i}\| \leq \nabla f(x^{k_i}) d^{k_i} \leq -\rho \cdot \|d^{k_i}\|^p$$

gilt, folgt

$$\|d^{k_i}\| \leq \left(\frac{1}{\rho}\|\nabla f(x^{k_i}\|\right)^{1/(p-1)}.$$

Also existieren zwei Konstanten c_1, c_2 mit

$$c_1 \leq \|d^{k_i}\| \leq c_2 \text{ für alle } i = 0, 1, \ldots.$$

Indem wir zu einer Teilfolge der $\{k_i\}$ übergehen, die wir wieder mit $\{k_i\}$ bezeichnen, können wir deshalb

$$\lim_{i\to\infty} d^{k_i} = d^* \neq 0$$

annehmen.
Angenommen, 0 ist Häufungspunkt der Folge t^{k_i}. Dann existiert eine konvergente Teilfolge, die wir der Einfachheit halber wieder mit t^{k_i} bezeichnen mit $\lim_{i\to\infty} t^{k_i} = 0$, d.h. in der Armijo-Schrittweitenbestimmung ist

$$t^{k_i} = \beta^{\ell_{k_i}} \text{ mit } \lim_{i\to\infty} \ell_{k_i} = +\infty.$$

Aus

$$f(x^{k_i} + \beta^{\ell_{k_i}-1} d^{k_i}) > f(x^{k_i}) + \sigma\beta^{\ell_{k_i}-1} \nabla f(x^{k_i}) d^{k_i}$$

folgt für $i \to \infty$ dann

$$\nabla f(x^*)d^* \geq \sigma \nabla f(x^*)d^*$$

und damit $\nabla f(x^*)d^* \geq 0$. Andererseits folgt aus dem Algorithmus aber auch

$$\nabla f(x^*)d^* \leq -\rho \|d^*\|^p < 0,$$

ein Widerspruch! □

3.5.10 Bemerkung
Für den soeben gezeigten Konvergenzsatz wurde nur $p > 1$ und $\sigma \in (0, 1)$ benötigt.

Aus dem Beweis zu Satz 3.4.3 wissen wir bereits, dass $\sigma < \frac{1}{2}$ nötig ist um zu erreichen, dass in der Nähe eines stationären Punktes x^* die Armijo-Schrittweite $t = 1$ für die Newton-Richtung akzeptiert wird. Die nun folgende Diskussion wird klären, weshalb $p > 2$ eine gute Wahl für zusätzliche Aussagen zur globalen Konvergenz ist.

3.5.11 Lemma
$\{x^k\} \subseteq \mathbb{R}^n$ sei eine Folge und x^* ein isolierter Häufungspunkt dieser Folge. Weiter gelte für jede gegen x^* konvergente Teilfolge $\{x^{k_i}\}$ die Beziehung

$$\lim_{i \to \infty} \|x^{k_i+1} - x^{k_i}\| = 0.$$

Dann gilt sogar $\lim_{k \to \infty} x^k = x^*$, d.h. die Folge $\{x^k\}$ konvergiert gegen x^*.

Beweis: Da x^* ein isolierter Häufungspunkt ist, existiert $\varepsilon > 0$ so, dass $B_\varepsilon(x^*)$ keine weiteren Häufungspunkte von $\{x^k\}$ enthält. Angenommen, $\{x^k\}$ konvergiert nicht gegen x^*. Dann sind für alle $i \geq i_0$ die Zahlen

$$\ell(k_i) = \max\{\ell : \|x^m - x^*\| \leq \varepsilon \text{ für alle } m = k_i, k_i + 1, \ldots, \ell\}$$

wohldefiniert, und wir haben

$$\|x^{\ell(k_i)} - x^*\| \leq \varepsilon, \quad \|x^{\ell(k_i)+1} - x^*\| > \varepsilon, k = 0, 1, \ldots$$

Die Teilfolge $\{x^{\ell(k_i)}\} \subseteq B_\varepsilon(x^*)$ konvergiert gegen x^*, denn in $B_\varepsilon(x^*)$ existiert kein weiterer Häufungspunkt. Damit folgt aber

$$\|x^{\ell(k_i)+1} - x^*\| \leq \|x^{\ell(k_i)} - x^*\| + \|x^{\ell(k_i)+1} - x^{\ell(k)}\| \to 0,$$

im Widerspruch zu $\|x^{\ell(k_i)+1} - x^*\| > \varepsilon$. □

3.5.12 Satz
Es gelten die Voraussetzungen von Satz 3.5.9 und x^* sei ein isolierter Häufungspunkt der Iterierten $\{x^k\}$ des globalisierten Newton-Verfahrens (Algorithmus 3.5.7). Dann konvergiert bereits die ganze Folge $\{x^k\}$ gegen x^*.

Beweis: Sei $\{x^{k_i}\}$ eine gegen x^* konvergente Teilfolge. Wir zeigen

$$\lim_{i\to\infty} \|x^{k_i+1} - x^{k_i}\| = 0$$

und wenden dann Lemma 3.5.11 an. Es ist

$$\|x^{k_i+1} - x^{k_i}\| = t^{k_i}\|d^{k_i}\| \leq \|d^{k_i}\| \tag{3.5.3}$$

und damit

$$\rho\|d^{k_i}\|^p \leq -\nabla f(x^{k_i})d^{k_i} \leq \|\nabla f(x^{k_i})\| \cdot \|d^{k_i}\|,$$

sofern im Algorithmus die Newton-Richtung als Suchrichtung akzeptiert wird. In diesen Fällen gilt also

$$\rho\|d^{k_i}\|^{p-1} \leq \|\nabla f(x^{k_i})\|.$$

Wurde statt dessen auf die Richtung des steilsten Abstiegs zurückgegriffen, so gilt trivialerweise

$$\|d^{k_i}\| = \|\nabla f(x^{k_i})\|.$$

Wegen $\lim_{i\to\infty} \|\nabla f(x^{k_i})\| = \|\nabla f(x^*)\| = 0$ (s. Satz 3.5.9) folgt also

$$\lim_{i\to\infty} \|x^{k_i+1} - x^{k_i}\| = 0.$$

□

Der vorangehende Satz motiviert $p > 1$. Dass $p > 2$ die richtige Wahl ist, ergibt sich aus

3.5.13 Satz
Sei $f : \mathbb{R}^n \to \mathbb{R}, f \in \mathcal{C}^2(\mathbb{R}^n), \{x^k\}$ die vom globalisierten Newton-Verfahren Algorithmus 3.5.7 erzeugte Folge, wobei $p > 2$. Weiter sei x^* ein Häufungspunkt dieser Folge und $\nabla^2 f(x^*)$ sei positiv definit. Dann gilt

(i) $\lim_{k\to\infty} x^k = x^*$ und x^* ist Stelle eines striktes Minimum von f

(ii) für k genügend groß ist d^k stets die Newton-Richtung

(iii) für k groß genug ist stets $t^k = 1$

(iv) $x^k \to x^*$ q-überlinear; im Falle $\nabla^2 f \in \mathrm{Lip}_\gamma(B_\varepsilon(x^*))$ sogar q-quadratisch.

Beweis: Zu (i): Wegen $f(x^k) \searrow f(x^*)$ gilt für *jeden* Häufungspunkt y^* der Folge $\{x^k\}$ die Beziehung $f(y^*) = f(x^*)$. Nach Satz 3.5.9 ist $\nabla f(x^*) = 0$, also ist x^* *strikte* lokale Minmalstelle für f. Also liegt y^* nicht beliebig nahe an x^*, d.h. x^* ist isolierter Häufungspunkt. Damit folgt (i) aus Satz 3.5.12.
Zu (ii): Wir zeigen, dass $\widetilde{\rho} > 0$ existiert mit

$$\nabla f(x^k)d^k \leq -\widetilde{\rho}\|d^k\|^2 \text{ für alle } k \geq k_0, \tag{3.5.4}$$

wobei $d^k = -\nabla^2 f(x^k)^{-1}\nabla f(x^k)^T$ die Newton-Richtung ist.
Wir wissen nach Lemma 3.4.2, dass für k groß genug

$$-\nabla f(x^k)d^k = (d^k)^T\nabla^2 f(x^k)d^k \geq \alpha \cdot \|d^k\|^2$$

gilt mit einer Konstanten $\alpha > 0$. Dies beweist (3.5.4).
Im Algorithmus wird die Newton-Richtung $d^k = -\nabla^2 f(x^k)^{-1}\nabla f(x^k)^T$ akzeptiert, falls

$$\nabla f(x^k)d^k \leq -\rho \cdot \|d^k\|^p, \quad p > 2.$$

Wegen $\lim_{k\to\infty} d^k = 0$ existiert ein k_0, so dass für alle $k \geq k_0$ gilt

$$\widetilde{\rho} \cdot \|d^k\|^{2-p} \geq \rho,$$

d.h. aus (3.5.4) folgt

$$\nabla f(x^k)d^k \leq -\tilde{\rho}\|d^k\|^2 = -\tilde{\rho}\|d^k\|^{2-p} \cdot \|d^k\|^p \leq -\rho \cdot \|d^k\|^p \text{ für } k \geq k_1.$$

Teil (iii) wurde bereits in Satz 3.4.3 bewiesen, der Teil (iv) folgt sofort aus den Sätzen 3.2.4 und 3.2.7. □

Wir übertragen die Resultate für das Newton-Verfahren jetzt noch auf ein globalisiertes inexaktes Newton-Verfahren.

3.5.14 Algorithmus (globalisiertes inexaktes Newton-Verfahren)
wähle $x^0 \in \mathbb{R}^n, \rho > 0, p > 2, \beta \in (0,1), \sigma \in (0, \frac{1}{2}), \varepsilon > 0, \bar{\eta} \in (0,1)$
for $k = 0, 1, \ldots$ **do**
  if $\|\nabla f(x^k)\| \leq \varepsilon$ **then**
    STOP
  else
    wähle $\eta_k \in [0, \bar{\eta}]$
    bestimme d^k mit $\|\nabla^2 f(x^k)d^k + \nabla f(x^k)^T\| \leq \eta_k \|\nabla f(x^k)\|$
      {inexakte Lösung der Newton-Gleichung}
    if $\nabla^2 f(x^k)$ singulär oder $\nabla f(x^k)d^k > -\rho\|d^k\|^p$ **then**
      $d^k = -\nabla f(x^k)^T$ {steilster Abstieg}
    end if
    bestimme $t^k = \max\{\beta^\ell, \ell = 0, 1, \ldots : f(x^k + \beta^\ell d^k) \leq f(x^k) + \sigma\beta^\ell \nabla f(x^k)d^k\}$
      {Armijo-Schrittweite}
    setze $x^{k+1} = x^k + t^k d^k$
  end if
end for

In Analogie zu Satz 3.5.9 gilt

3.5.15 Satz
Sei $f : \mathbb{R}^n \to \mathbb{R}, f \in \mathcal{C}^2(\mathbb{R}^n)$. Dann ist jeder Häufungspunkt x^* der vom globalisierten inexakten Newton-Verfahren (Algorithmus 3.5.14 erzeugten Folge $\{x^k\}$ ein stationärer Punkt von f.

Beweis: Sei $x^* = \lim_{i\to\infty} x^{k_i}$ ein Häufungspunkt mit zugehöriger konvergenter Teilfolge $\{x^{k_i}\}$. Falls $d^{k_i} = -\nabla f(x^{k_i})^T$ für unendlich viele i, so ist x^* stationärer Punkt wegen Korollar 3.5.5. Wir nehmen also ab jetzt an, dass

$$\|\nabla^2 f(x^{k_i})d^{k_i} + \nabla f(x^{k_i})^T\| \leq \bar{\eta}\|\nabla f(x^{k_i})\| \text{ für alle } i \geq i_0. \tag{3.5.5}$$

Wegen $\lim_{k\to\infty} f(x^k) \searrow f(x^*)$ gilt $\lim_{k\to\infty} f(x^{k+1}) - f(x^k) = 0$ und damit die Beziehung

$$\lim_{i\to\infty} t^{k_i} \nabla f(x^{k_i})d^{k_i} = 0. \tag{3.5.6}$$

Wir nehmen an, dass x^* kein stationärer Punkt ist, also $\nabla f(x^*) \neq 0$. Dann ergibt sich aus 3.5.5 zunächst

$$\|\nabla f(x^{k_i})\| - \|\nabla^2 f(x^{k_i})d^{k_i}\| \leq \bar{\eta} \cdot \|\nabla f(x^{k_i})\|,$$

und daher

$$\frac{1-\bar{\eta}}{\alpha} \cdot \|\nabla f(x^{k_i})\| \leq \|d^{k_i}\|$$

mit $\alpha = \sup\{\|\nabla^2 f(x^{k_i})\|, i = 0, 1, \ldots\} < \infty$.
Damit konvergieren die d^{k_i} nicht gegen 0, und weil die Folge $\{d^{k_i}\}$ beschränkt ist (dies folgt aus der Beziehung $\nabla f(x^k)d^k > -\rho\|d^k\|^p$), besitzt sie eine gegen ein $d^* \neq 0$ konvergente Teilfolge, die wir wieder mit k_i indizieren. Die Folge der t^{k_i} hat 0 nicht als Häufungspunkt, denn andernfalls würde, wieder nach Übergang auf eine Teilfolge, aus der Armijo-Schrittweitenwahl folgen

$$f(x^{k_i} + \beta^{\ell(k_i)-1}d^{k_i}) \geq f(x^{k_i}) + \sigma\beta^{\ell(k_i)-1}\nabla f(x^{k_i})d^{k_i},$$

woraus für $i \to \infty$ wegen Lemma 3.5.3

$$\nabla f(x^*)d^* \geq \sigma\nabla f(x^*)d^* \leq 0$$

folgt. Dies bedeutet aber $\nabla f(x^*)d^* = 0$, woraus mit der Bedingung

$$\nabla f(x^k)d^k \leq -\rho\|d^k\|^p$$

aus dem Verfahren jedoch $d^* = 0$ folgen würde, ein Widerspruch. Also besitzt die Folge der $\{t^{k_i}\}$ nicht den Häufungspunkt 0, weshalb aus (3.5.6) wieder $\nabla f(x^*)d^* = 0$ folgt. Hieraus ergibt sich wie gerade eben wieder der Widerspruch $d^* = 0$. Die anfängliche Annahme $\nabla f(x^*) \neq 0$ war also falsch. □

Auch die anderen Aussagen zum globalisierten Newton-Verfahren lassen sich übertragen, vorausgesetzt, das inexakte Verfahren ist q-überlinear konvergent.

3.5.16 Satz
Es gelten die Voraussetzungen von Satz 3.5.15. Weiter sei $\lim_{k\to\infty} \eta_k = 0$ in Algorithmus 3.5.14 und x^* sei ein isolierter Häufungspunkt der Iterierten $\{x^k\}$ des globalisierten inexakten Newton-Verfahrens (Algorithmus 3.5.14). Dann konvergiert bereits die ganze Folge $\{x^k\}$ gegen x^*.

Beweis: Der Beweis geht eigentlich wörtlich so wie für das exakte Newton-Verfahren: Sei $\{x^{k_i}\}$ eine gegen x^* konvergente Teilfolge. Wir zeigen

$$\lim_{i\to\infty} \|x^{k_i+1} - x^{k_i}\| = 0$$

und wenden dann Lemma 3.5.11 an. Es ist

$$\|x^{k_i+1} - x^{k_i}\| = t^{k_i}\|d^{k_i}\| \leq \|d^{k_i}\| \tag{3.5.7}$$

und damit

$$\rho\|d^{k_i}\|^p \leq -\nabla f(x^{k_i})d^{k_i} \leq \|\nabla f(x^{k_i})\| \cdot \|d^{k_i}\|,$$

sofern im Algorithmus die inexakte Newton-Richtung als Suchrichtung akzeptiert wird. In diesen Fällen gilt also

$$\rho\|d^{k_i}\|^{p-1} \leq \|\nabla f(x^{k_i})\|.$$

Wurde statt dessen auf die Richtung des steilsten Abstiegs zurückgegriffen, so gilt trivialerweise

$$\|d^{k_i}\| = \|\nabla f(x^{k_i})\|.$$

Wegen $\lim_{i\to\infty} \|\nabla f(x^{k_i})\| = \|\nabla f(x^*)\| = 0$ (s. Satz 3.5.15) folgt also

$$\lim_{i\to\infty} \|x^{k_i+1} - x^{k_i}\| = 0.$$

□

3.5.17 Satz
Sei $f : \mathbb{R}^n \to \mathbb{R}, f \in \mathcal{C}^2(\mathbb{R}^n), \{x^k\}$ die vom globalisierten inexakten Newton-Verfahren (Algorithmus 3.5.14) erzeugte Folge. In dem Verfahren gelte wieder $\lim_{k\to\infty} \eta_k = 0$. Weiter sei x^* ein Häufungspunkt dieser Folge und $\nabla^2 f(x^*)$ sei positiv definit. Dann gilt

(i) $\lim_{k\to\infty} x^k = x^*$ und x^* ist striktes lokale Minimalstelle von f

(ii) für k genügend groß ist d^k stets die inexakte Newton-Richtung

(iii) für k groß genug ist stets $t^k = 1$

(iv) $x^k \to x^*$ q-überlinear; im Falle $\nabla^2 f \in \mathrm{Lip}_\gamma(B_\varepsilon(x^*))$ und $\eta_k = \mathcal{O}(\|F(x^k)\|)$ sogar q-quadratisch.

Beweis: Zu (i): Der erste Teil folgt wie beim exakten Newton-Verfahren: Wegen $f(x^k) \searrow f(x^*)$ gilt für *jeden* Häufungspunkt y^* der Folge $\{x^k\}$ die Beziehung $f(y^*) = f(x^*)$. Nach Satz 3.5.15 ist $\nabla f(x^*) = 0$, also ist x^* *strikte* lokale Minmalstelle für f. Also liegt y^* nicht beliebig nahe an x^*, d.h. x^* ist isolierter Häufungspunkt. Damit folgt (i) aus Satz 3.5.16.
Zu (ii): Wir zeigen, dass $\widetilde{\rho} > 0$ und k_0 existieren mit

$$\nabla f(x^k)d^k \leq -\widetilde{\rho}\|d^k\|^2 \text{ für alle } k \geq k_0, \tag{3.5.8}$$

wobei d^k die inexakte Newton-Richtung ist mit

$$\|\nabla^2 f(x^k)d^k + \nabla f(x^k)^T\| \leq \eta_k\|\nabla f(x^k)\|.$$

Wir müssen jetzt etwas anders argumentieren als beim exakten Verfahren. Für $k \geq k_0$ ist nach Lemma 3.4.2 $\nabla f(x^k)\nabla^2 f(x^k)\nabla f(x^k) \geq \alpha\|\nabla f(x^k)^T\|^2$ mit einem festen $\alpha > 0$. Für diese k haben wir dann

$$\begin{aligned}\nabla f(x^k)d^k &= \nabla f(x^k)\nabla^2 f(x^k)^{-1}\left(\nabla^2 f(x^k)d^k + \nabla f(x^k)^T\right)\\ &\quad - \nabla f(x^k)\nabla^2 f(x^k)^{-1}\nabla f(x^k)^T\\ &\leq \nabla f(x^k)\nabla^2 f(x^k)^{-1}\left(\nabla^2 f(x^k)d^k + \nabla f(x^k)^T\right) - \alpha\|\nabla f(x^k)\|^2\\ &\leq (\eta_k\gamma - \alpha)\|\nabla f(x^k)\|^2,\end{aligned}$$

mit $\gamma = \sup\{\|\nabla^2 f(x^k)^{-1}\|, k \geq k_0\} < \infty$. Wie im Beweis zu Korollar 3.3.5 existiert eine weitere Konstante $\beta > 0$ mit $\|\nabla f(x^k)\| \geq \beta\|d^k\|$ für alle k. Nimmt man also k_0 so groß, dass auch noch $\eta_k\gamma \leq \alpha/2$ für alle $k \geq k_0$, so kann man $\widetilde{\rho} = \beta^2 \cdot \alpha/2$ nehmen.
Im Algorithmus wird die inexakte Newton-Richtung d^k akzeptiert, falls

$$\nabla f(x^k)d^k \leq -\rho \cdot \|d^k\|^p, \quad p > 2.$$

Wegen $\lim_{k\to\infty} d^k = 0$ existiert ein k_1, so dass für alle $k \geq k_1$ gilt

$$\widetilde{\rho} \cdot \|d^k\|^{2-p} \geq \rho,$$

d.h. aus (3.5.8) folgt

$$\nabla f(x^k)d^k \leq -\tilde{\rho}\|d^k\|^2 = -\tilde{\rho}\|d^k\|^{2-p} \cdot \|d^k\|^p \leq -\rho \cdot \|d^k\|^p \text{ für } k \geq k_1.$$

Teil (iii) wurde bereits in Satz 3.4.6 bewiesen, der Teil (iv) folgt sofort aus Korollaren 3.3.5 und 3.2.9. □

Abschnitt 3.6

Praktisch relevante Modifikationen

Die Praxis zeigt, dass das nicht-gobalisierte Newton-Verfahren häufig auch dann eine 'gute' Iterierte bestimmt, wenn die Newton-Richtung keine Abstiegsrichtung ist oder wenn die Schrittweite $t = 1$ nicht die Armijo-Bedingung erfüllt. Unsere bisherigen Globalisierungen sind in diesem Sinne 'zu streng', d.h. sie verwerfen die Newton-Korrektur zu häufig zu Gunsten der Richtung des steilsten Abstiegs oder sie machen nur einen 'kleinen' Schritt in die Newton-Richtung. Zwar ist ein möglicher Rückzug auf die Richtung des steilsten Abstiegs fast immer eine notwendige Zutat bei der Globalisierung, aber man kann darauf abzielen, die steilste Abstiegsrichtung nur in möglichst wenig Fällen zu verwenden oder möglichst häufig die volle Schrittweite bei Newton-Schritten zu erreichen.
In diesem Abschnitt besprechen wir zwei Strategien mit dieser Zielsetzung.

3.6.1 Nicht-monotone Armijo-Regel

Die Idee ist, häufiger als bisher einen vollen Schritt (Schrittweite $t = 1$) in die Newton-Richtungen zu akzeptieren.

3.6.1 Definition
Sei $f : \mathbb{R}^n \to \mathbb{R}$, $f \in \mathcal{C}^1(\mathbb{R}^n)$. Es sei $\{x^k\} \subset \mathbb{R}^n$ eine Folge und $\sigma \in (0,1)$, $\beta \in (0,1)$ gegeben. Wir sagen, die Iterierte x^k mit Suchrichtung d^k genügt der *nicht monotonen Armijo-Bedingung* (mit Tiefe $m_k \in \mathbb{N}_0$), falls gilt

$$x^{k+1} = x^k + \beta^{\ell(k)} d^k$$

mit

$$\beta^{\ell(k)} = \max\{\beta^\ell : f(x^k + \beta^\ell d^k) \leq \max_{i=k-m_k}^{k} f(x^i) + \sigma\beta^\ell \nabla f(x^k) d^k\}. \tag{3.6.1}$$

Für $m_k = 0$ erhält man die normale Armijo-Bedingung. Die Bedingung (3.6.1) erlaubt also ein zwischenzeitliches Anwachsen des Funktionswertes, 'im Mittel' fällt er aber trotzdem ab.
Zum Einbau in ein globalisiertes Newton-Verfahren braucht man noch eine Vorschrift, ob und wie man die m_k aufdatiert. Folgendes Vorgehen hat sich in der Praxis bewährt.

3.6.2 Algorithmus (global. Newton, nicht-monot. Armijo-Regel)
wähle $x^0 \in \mathbb{R}^n, \rho > 0, p > 2, \beta \in (0,1), \sigma \in (0, \frac{1}{2}), \varepsilon > 0$
wähle $\bar{m} \in \mathbb{N}_0$, setze $m_0 = 0$
for $k = 0, 1, \ldots$ **do**
  if $\|\nabla f(x^k)\| \leq \varepsilon$ **then**
    STOP
  else
    löse $\nabla^2 f(x^k) d^k = -\nabla f(x^k)^T$ {Newton-Gleichung}
    setze $m_k = \min\{m_{k-1} + 1, \bar{m}\}$
    if $\nabla^2 f(x^k)$ singulär oder $\nabla f(x^k) d^k > -\rho \|d^k\|^p$ **then**
      $d^k = -\nabla f(x^k)^T$ {steilster Abstieg}
      $m_k = 0$
    end if
    bestimme t^k über die nicht-montone Armijo-Regel 3.6.1 {Armijo-Schrittweite}
    setze $x^{k+1} = x^k + t^k d^k$
  end if
end for

Der Rückgriff auf $m_k = 0$ (und damit die normale Armijo-Regel) im Falle der Richtung des steilsten Abstiegs ist dafür verantwortlich, dass man auch für dieses Verfahren globale Konvergenz zeigen kann.[1] Die Vorschrift $m_k = \min\{m_{k-1} + 1, \bar{m}\}$ im Falle der Newton-Richtung bedeutet: Wir berücksichtigen nun auch den neuesten Funktionswert in der nichtmonotonen Armijo-Regel. Wird dabei die Maximalzahl $\bar{m}$ erreicht, entfernen wir dafür den ältesten Funktionswert.
In der Praxis hat sich $\bar{m} \approx 10$ bewährt und $m_k = 0$ für $k = 0, 1, \ldots, 5$. Die Praxis zeigt auch: die nicht-monotone Armijo-Regel ist weniger günstig bei inexakten Newton-Verfahren.

3.6.2 Modifikation der Cholesky-Zerlegung

Die Newton-Richtung ist nicht notwendig eine Abstiegsrichtung, wenn $\nabla^2 f(x)$ nicht spd ist. Das LGS

$$\nabla^2 f(x) d = -\nabla f(x)^T \tag{3.6.2}$$

kann man mittels der Cholesky-Faktorisierung $\nabla^2 f(x) = LL^T$ lösen. Die Cholesky-Faktorisierung existiert genau dann, wenn $\nabla^2 f(x)$ spd ist. Die Idee

[1] s. Grippo, L., Lampariello, F. und Lucidi, S.: A nonmonotone line search technique for Newton's method, SIAM J. Numer. Anal. **23**, 707 - 716 (1986)

ist nun, während der Berechnung der Cholesky-Faktoriserung im Falle dass $\nabla^2 f(x)$ nicht spd ist, die Matrix so abzuändern, dass die Matrix spd wird und ihre Cholesky-Zerlegung bestimmt wird. Statt (3.6.2) löst man dann das System mit der modfizierten Matrix. Da diese spd ist, ist das berechnete d eine Abstiegsrichtung, wahrscheinlich eine bessere als die Richtung des steilsten Abstiegs.
Zur Präzisierung des Vorgehens wiederholen wir den Algorithmus zur Berechnung der Cholesky-Zerlegung $A = LL^T$ einer spd Matrix A. Hierin ist L eine untere Dreiecksmatrix, so dass wir durch Gleichsetzen erhalten

$$a_{ij} = \sum_{k=1}^{j} \ell_{ik} \cdot \ell_{jk} \quad i = 1, \ldots, n, \; j = 1, \ldots, i.$$

Löst man diese Beziehung sukzessive nach den Zeilen von L auf, so erhält man

3.6.3 Algorithmus (Cholesky-Zerlegung, zeilenorientiert)
for $i = 1, \ldots, n$ **do**
  for $j = 1, \ldots, i-1$ **do**
    $\ell_{ij} = \left(a_{ij} - \sum_{k=1}^{j-1} \ell_{ik} \cdot \ell_{jk}\right) / \ell_{jj}$
  end for
  $\ell_{ii} = \sqrt{a_{ii} - \sum_{k=1}^{i-1} \ell_{ik}^2}$
end for

Ist A nicht positiv definit, so scheitert dieser Algorithmus, weil für ein i die Beziehung $a_{ii} - \sum_{k=1}^{i-1} \ell_{ik}^2 \leq 0$ erreicht wird. Ist A spd, so gilt stets $a_{ii} - \sum_{k=1}^{i-1} \ell_{ik}^2 > 0$.
Die modifizierte Cholesky-Zerlegung ändert nun einfach $a_{ii} - \sum_{k=1}^{i-1} \ell_{ik}^2$ in eine (ausreichend) positive Größe ab, wenn dies notwendig sein sollte.

3.6.4 Algorithmus (modifiz. Cholesky-Zerlegung, zeilenorient.)
wähle $\mu > 0$ $\{$z.B. $\mu = 10^{-7}\}$

for $i = 1, \ldots, n$ **do**
  for $j = 1, \ldots, i-1$ **do**
    $\ell_{ij} = \left(a_{ij} - \sum_{k=1}^{j-1} \ell_{ik} \cdot \ell_{jk}\right) / \ell_{jj}$
  end for
  $aux = a_{ii} - \sum_{k=1}^{i-1} \ell_{ik}^2$
  if $aux < \mu$ **then**
    $aux = \mu$ $\{aux$ wird ausreichend positiv gemacht$\}$
  end if
  $\ell_{ii} = \sqrt{aux}$
end for

3.6.5 Lemma
Für die mit der modifizierten Cholesky-Zerlegung berechnete Matrix L gilt

$$LL^T = A + D, \; D = \operatorname{diag}(d_1, \ldots, d_n)$$

mit

$$d_i = \begin{cases} 0 & \text{falls } a_{ii} - \sum_{k=1}^{i-1} \ell_{ik}^2 \geq \mu \\ \mu - (a_{ii} - \sum_{k=1}^{i-1} \ell_{ik}^2) & \text{sonst} \end{cases}$$

Insbesondere ist $A + D$ spd.

Beweis: Man verifiziert sofort, dass die gewöhnliche Cholesky-Zerlegung nach Algorithmus 3.6.3 für $A + D$ äquivalent ist zur modifizierten Cholesky-Zerlegung nach Algorithmus 3.6.4 für A. □

Das globalisierte Newton-Verfahren (Algorithmus 3.5.7 wird nun wie folgt abgeändert: In jedem Schritt wird zunächst die Gleichung

$$\nabla^2 f(x^k) d^k = -\nabla f(x^k)^T$$

durch Bestimmung der modifizierten Cholesky-Zerlegung von $\nabla^2 f(x^k)$ mit anschließendem Lösen der Dreieckssysteme 'gelöst'. Der Rest des Verfahrens bleibt unverändert. Die resultierende Abstiegsrichtung d^k wird nun häufig akzeptiert werden, so dass nur selten auf die Richtung des tiefsten Abstiegs zurück gegriffen werden muss. Die Konvergenztheorie zu dieser Modifikation besprechen wir hier nicht.

Kapitel 4

Quasi-Newton-Verfahren

Weil sowohl die Berechnung von $\nabla^2 f(x)$ (n^2 Einträge), wie auch die Lösung von $\nabla^2 f(x)d = -\nabla f(x)^T$ ($\mathcal{O}(n^3)$ Operationen bei Faktorisierung von $\nabla^2 f(x)$ als dicht besetzte Matrix) aufwendig sind, sucht man nach billigeren Alternativen zum Newton-Verfahren, welche trotzdem noch überlinear konvergent sind. Die wichtigste Klasse solcher Verfahren sind die *Quasi-Newton-Verfahren*, welche wir hier behandeln.

Abschnitt 4.1

Motivation und Definition

Wir untersuchen Verfahren von der Bauart

$$x^{k+1} = x^k - H_k^{-1}\nabla f(x^k)^T. \tag{4.1.1}$$

Statt H_k als $\nabla^2 f(x^k)$ immer wieder neu zu berechnen, suchen wir nun nach einer Aufdatierungsvorschrift, mit der man H_{k+1} einfach aus H_k berechnen kann.
Wir nehmen an, dass $f \in \mathcal{C}^2(\mathbb{R}^n)$. Der Satz von Dennis und Moré (Satz 3.3.3) zeigt, dass wir überlineare Konvergenz genau dann erreichen können, wenn

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

gilt. Auf Grund der Differenzierbarkeit von ∇f gilt

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

Wir streben deshalb

$$\|H_k(x^{k+1} - x^k) - \left(\nabla f(x^{k+1})^T - \nabla f(x^k)^T\right)\| = o(\|x^{k+1} - x^k\|)$$

an. Dies ist zunächst noch keine brauchbare Beziehung, denn sie verbindet das zu bestimmende H_k mit dem über H_k zu bestimmenden x^{k+1}.
Fordern wir allerdings mit H_{k+1} statt H_k die *Sekantenbedingung*

$$H_{k+1}(x^{k+1} - x^k) = \nabla f(x^{k+1})^T - \nabla f(x^k)^T, \tag{4.1.2}$$

so haben wir eine „praktikable“ Vorschrift für H_{k+1}. Vorsorglich halten wir dazu schon einmal fest:

4.1.1 Satz
Sei $f : \mathbb{R}^n \to \mathbb{R}$, $f \in \mathcal{C}^1(\mathbb{R}^n)$. Erfüllt ein Verfahren der Bauart 4.1.1 die Sekantenbedingung 4.1.2 und konvergiert $\{x^k\}$ gegen einen stationären Punkt x^* von f, so konvergiert die Folge sogar q-überlinear, falls

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

gilt.

Beweis: Wegen $H_k(x^{k+1} - x^k) = \nabla f(x^k)^T$ ist obige Bedingung die klassische Bedingung von Dennis-Moré für q-überlineare Konvergenz gemäß Satz 3.3.3. □

4.1.2 Definition
Sei $f : \mathbb{R}^n \to \mathbb{R}$, $f \in \mathcal{C}^1(\mathbb{R}^n)$. Ein *Quasi-Newton-Verfahren* zur Berechnung einer Minimalstelle von f ist ein Verfahren der Gestalt

$$x^{k+1} = x^k - H_k^{-1}\nabla f(x^k),\ k = 0, 1, \ldots,$$

wobei die H_k der Sekantenbedingung

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

genügen.

Man beachte, dass die Sekantenbedingung die Matrix H_{k+1} nicht vollständig festlegt, denn sie definiert nur das Bild von H_{k+1} auf einem eindimensionalen Teilraum. Je nachdem, wie H_{k+1} sonst noch definiert wird, ergeben sich verschiedene Quasi-Newton-Verfahren. Der prinzipielle Ansatz ist aber immer, H_{k+1} „einfach“ aus H_k aufzudatieren.
Im Folgenden wird es primär um die Diskussion solcher Aufdatierungsvorschriften gehen. Zur Vereinfachung der Notation sparen wir uns dazu den Index k, und notieren ein „+“ für den Index $k + 1$. Die Sekantenbedingung lautet so

$$H_+(x^+ - x) = \nabla f(x^+)^T - \nabla f(x)^T. \tag{4.1.3}$$

Abschnitt 4.2

PSB-, DFP- und BFGS-Verfahren

In diesem Abschnitt leiten wir einige konkrete Quasi-Newton-Verfahren her. Wir benötigen dazu einige Hilfsresultate, die so einfach zu zeigen sind, dass wir auf den Beweis verzichten. Im ganzen Kapitel bezeichnet jetzt $\|\cdot\|$ die ℓ_2-Norm.

4.2.1 Lemma
Sei $w \in \mathbb{R}^n$. Dann gilt

$$\|w\| = \max_{\|x\|=1} \{x^T w\}.$$

4.2.2 Lemma
Für alle $v, w \in \mathbb{R}^n$ gilt

$$\|vw^T\| = \|v\| \cdot \|w\|.$$

Hier ist $vw^T \in \mathbb{R}^{n\times n}$, links steht also die ℓ_2-Operatornorm.

4.2.3 Definition
Für $A \in \mathbb{R}^{n\times m}$ ist die *Frobenius-Norm* gegeben durch

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

4.2.4 Lemma
Für jede orthogonale Matrix $Q \in \mathbb{R}^{n\times n}$ und $A \in \mathbb{R}^{n\times n}$ gilt

$$\|AQ\|_F = \|A\|_F = \|Q^T A\|_F.$$

Wir starten nun mit der *PSB (Powell symmetric Broyden) Formel*.

4.2.5 Satz
Sei $H \in \mathbb{R}^{n\times n}$ symmetrisch und $y, s \in \mathbb{R}^n$, $s \neq 0$. Die eindeutig bestimmte Lösung der Aufgabe

$$\text{minimiere } \|H_+ - H\|_F^2 \text{ unter der Nebenbedingung } H_+ = H_+^T,\ H_+ s = y$$

ist gegeben durch die *PSB-Formel*

$$H_+^{PSB} = H + \frac{1}{s^T s} \cdot ((y - Hs)s^T + s(y - Hs)^T) - \frac{(y - Hs)^T s}{(s^T s)^2} ss^T.$$

Beweis: Klar ist zunächst: H_+^{PSB} ist symmetrisch und $H_+^{PSB} s = y$. Die gestellte Aufgabe ist die Minimierung einer streng konvexen Funktion auf einer konvexen Teilmenge von $\mathbb{R}^{n\times n}$ und besitzt demnach eine eindeutige Lösung.
Sei $H_+ \in \mathbb{R}^{n\times n}$ eine symmetrische Matrix mit $H_+ s = y$. Sei $v_1 = (1/\|s\|)s$, und $v_2, \ldots, v_n$ mögen v_1 zu einer Orthonomalbasis von $\mathbb{R}^n$ ergänzen. Sei $Q = [v_1|\ldots|v_n]$ die orthonormale Matrix mit den Spalten v_i. Dann gilt

$$\begin{aligned}
\|H_+^{PSB} - H\|_F^2 &\overset{\text{L. 4.2.4}}{=} \|(H_+^{PSB} - H)Q\|_F^2 \\
&= \sum_{i=1}^{n} \|(H_+^{PSB} - H)v_i\|^2 \\
&= \|(H_+^{PSB} - H)v_1\|^2 + \sum_{i=2}^{n} \|(H_+^{PSB} - H)v_i\|^2 \\
&= \|(H_+ - H)v_1\|^2 + \sum_{i=2}^{n} \|(H_+^{PSB} - H)v_i\|^2. \qquad (4.2.1)
\end{aligned}$$

Für die Terme in der Summe haben wir aber wegen $v_i^T s = 0, i = 2, \ldots, n$

$$\begin{aligned}
&\|(H_+^{PSB} - H)v_i\| \\
&= \|\frac{1}{s^T s} \cdot ((y - Hs)s^T + s(y - Hs)^T)v_i - \frac{(y - Hs)^T s}{(s^T s)^2} ss^T v_i\| \\
&= \|\frac{1}{s^T s} \cdot s(y - Hs)^T)v_i\| \\
&= \|\frac{1}{s^T s} \cdot ss^T (H_+ - H)^T v_i\| \\
&\leq \frac{1}{s^T s} \cdot \|ss^T\| \cdot \|(H_+ - H)v_i\| \\
&\overset{\text{L. 4.2.2}}{=} \|(H_+ - H)v_i\|.
\end{aligned}$$

Aus (4.2.1) folgt damit

$$\|H_+^{PSB} - H\|_F^2 \leq \sum_{i=1}^{n} \|(H_+ - H)v_i\|^2 = \|H_+ - H\|_F^2,$$

d.h. H_+^{PSB} löst die Minimierungsaufgabe. □

4.2.6 Bemerkung
Der Beweis hat gezeigt, dass H_+^{PSB} sich auf dem orthogonalen Komplement von s so wenig wie möglich von H unterscheidet.

Den PSB-Ansatz kann man auf gewichtete Frobenius-Normen übertragen.

4.2.7 Satz
Sei $W \in \mathbb{R}^{n\times n}$ regulär. Weiter sei $H \in \mathbb{R}^{n\times n}$ symmetrisch und $y, s \in \mathbb{R}^n$, $s \neq 0$. Die eindeutig bestimmte Lösung der Aufgabe

$$\text{minimiere } \|W(H_+ - H)W^T\|_F^2 \text{ unter der Nebenbed. } H_+ = H_+^T,\ H_+ s = y$$

ist gegeben durch

$$H_+^{PSB} = H + \frac{1}{s_W^T s} \cdot ((y - Hs)s_W^T + s_W(y - Hs)^T) - \frac{(y - Hs)^T s}{(s_W^T s)^2} s_W s_W^T.$$

mit

$$s_W = (W^T W)^{-1} s.$$

Beweis: Die gewichtete Minimierungsaufgabe wird von H_+ genau dann gelöst, wenn die Matrix $\tilde{H}_+ = WH_+W^T$ die ungewichtete Aufgabe aus Satz 4.2.5 löst mit neuen Vektoren $\tilde{s} = W^{-T}s$ und $\tilde{y} = Wy$ sowie $\tilde{H} = WHW^T$. Damit ergibt sich die Behauptung nach Rücktransformation der Lösung aus Satz 4.2.5. Details als Übungsaufgabe. □

Man beachte, dass die Lösung nur von W^TW, nicht aber von W selbst abhängt.
Die PSB-Formel liefert, auch wenn H spd ist, nicht notwendig eine spd Matrix H_+. Für beliebiges s und y braucht ein solches H_+ gar nicht zu existieren.

4.2.8 Lemma
Seien $y, s \in \mathbb{R}^n$ mit $s \neq 0$. Genau dann existiert eine spd Matrix Q mit $Qs = y$, wenn $y^T s > 0$.

Beweis: „⇒“: Wegen Q spd gilt $0 < s^T Qs = y^T s$.
„⇐“: Setze

$$v = \sqrt{\frac{y^T s}{s^T s}} \cdot s.$$

und

$$W = I + \frac{1}{v^T v} \cdot (y - v)v^T,\ Q = WW^T.$$

Dann ist Q symmetrisch, positiv semidefinit und $Qs = y$. Außerdem ist W regulär, denn aus $Wx = 0$ folgt

$$x = -\frac{1}{v^T v} \cdot (y - v)v^T x = -\frac{v^T x}{v^T v} \cdot (y - v),$$

woraus sich

$$x = \alpha \cdot (y - v)$$

ergibt, und damit sogar dann

$$\alpha = -\alpha \frac{v^T(y-v)}{v^T v}.$$

Dies bedeutet aber $\alpha = 0$, denn andernfalls wäre

$$1 = -\frac{v^T(y-v)}{v^T v} \iff v^T y = 0,$$

was wegen $v^T y = y^T s / \sqrt{s^T s}$ und $y^T s > 0$ aber unmöglich ist. Also ist W regulär und Q damit positiv definit. □
Nimmt man nun im Falle, dass H symmetrisch ist und $s^T y > 0$ für Q die nach Lemma 4.2.8 existierende spd Matrix mit $Qs = y$ und nimmt in Satz 4.2.7 für W einen Cholesky-Faktor von Q, so ist dort $s_W = y$ und es ergibt sich die DFP-Formel (Davidon, Fletcher, Powell).

4.2.9 Definition
Sei $H \in \mathbb{R}^{n\times n}$ symmetrisch und $y, s \in \mathbb{R}^n$, $s \neq 0$ mit $y^T s > 0$. Die *DFP-Formel* bestimmt die Matrix

$$H_+^{DFP} = H + \frac{1}{y^T s} \cdot ((y - Hs)y^T + y(y - Hs)^T) - \frac{(y - Hs)^T s}{(y^T s)^2} yy^T$$

mit

$$H_+^{DFP} s = y.$$

4.2.10 Bemerkung
Mit H ist auch H_+^{DFP} spd, denn

$$x^T H_+^{DFP} x = \frac{1}{(y^T s)^2} \cdot \Big((x^T y)^2 (y^T s) + \big[(y^T s)x - (y^T x)s\big]^T H \big[(y^T s)x - (y^T x)s\big] \Big).$$

Wir formulieren beispielhaft das DFP-Verfahren (= Quasi-Newton-Verfahren mit DFP-Formel). Dabei vereinfachen wir in der DFP-Formel die Terme $y - Hs$, denn in einem Quasi-Newton-Verfahren mit der Gestalt 4.1.1 gilt wegen der Sekantenbedingung 4.1.3

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

4.2.11 Algorithmus (DFP-Verfahren, Davidon, Fletcher, Powell)
wähle x^0
wähle H_0, z.B. $H_0 = \nabla^2 f(x^0)$
setze $u^{-1} = \nabla f(x^0)^T$
for $k = 0, 1, \ldots$ **do**
  löse $H_k s^k = -u^{k-1}$
  setze $x^{k+1} = x^k + s^k$
  setze $u^k = \nabla f(x^{k+1})^T$, $y^k = u^k - u^{k-1}$
  setze $H_{k+1} = H_k + \frac{1}{(y^k)^T s^k} \cdot ((u^k)(y^k)^T + y^k(u^k)^T) - \frac{(u^k)^T s^k}{((y^k)^T s^k)^2} y^k (y^k)^T$.
  {DFP-Update}
end for

Im DFP-Verfahren müssen wir in jedem Schritt ein LGS mit der Matrix H_k lösen. Das kann man vermeiden, wenn man immer gleich mit den Inversen arbeitet, denn das nächste Lemma zeigt, dass sich mit den Matrizen auch die Inversen einfach aufdatieren lassen.

4.2.12 Lemma (Sherman-Morrison-Woodbury-Formel)
Sei $A \in \mathbb{R}^{n\times n}$ regulär und $U, V \in \mathbb{R}^{n\times m}$ mit $m \leq n$. Dann ist die Matrix $A + UV^T$ genau dann regulär, wenn $I + V^T A^{-1} U$ dies ist, und es gilt

$$\underbrace{(A + UV^T)^{-1}}_{n\times n} = A^{-1} - A^{-1} \cdot U \underbrace{\left(I + V^T A^{-1} U\right)^{-1}}_{m\times m} \cdot V^T A^{-1}.$$

Beweis: s. Übung. □
Die Sherman-Morrison-Woodbury-Formel ist besonders nützlich, wenn m sehr klein ist. Z. B. ergibt sich sowohl beim PSB- wie auch beim DFP-Update

$$H_+ = H + UV^T \text{ mit } U, V \in \mathbb{R}^{n\times 2}.$$

Bei bekanntem $B = H^{-1}$ ergibt sich dann für $B_+ = H_+^{-1}$

$$B_+ = B - BU \cdot \underbrace{(I + V^T BU)^{-1}}_{2\times 2} \cdot V^T B,$$

also eine billige Aufdatierungsformel für B. Man kann also Algorithmus 4.2.11 entsprechend umformulieren.
Man kann aber auch noch einen Schritt weiter gehen, und von vornherein *inverse* Quasi-Newton-Formeln untersuchen. Es wird dann B_+ aus B aufdatiert mit der Bedingung

$$B_+ y = s.$$

Das folgende Resultat ist lediglich Satz 4.2.7 mit anderen Bezeichnungen.

4.2.13 Lemma
Sei $W \in \mathbb{R}^{n\times n}$ regulär. Weiter sei $B \in \mathbb{R}^{n\times n}$ symmetrisch und $y, s \in \mathbb{R}^n$, $y \neq 0$. Die eindeutig bestimmte Lösung der Aufgabe
(4.2.2)
minimiere $\|W(B_+ - B)W^T\|_F^2$ unter der Nebenbed. $B_+ = B_+^T,\ s = B_+ y$

ist gegeben durch

$$B_+ = B + \frac{1}{y_W^T y} \cdot ((s - By)y_W^T + y_W(s - By)^T) - \frac{(s - By)^T y}{(y_W^T y)^2} y_W y_W^T.$$

mit

$$y_W = (W^T W)^{-1} y.$$

Nimmt man in diesem Lemma in Analogie zur DFP-Formel für $W^T W$ die spd Matrix Q mit $Qs = y$, so ergibt sich die wichtige BFGS-Formel (Broyden, Fletcher, Goldfarb, Shanno):

4.2.14 Definition
Sei $B \in \mathbb{R}^{n\times n}$ symmetrisch und $y, s \in \mathbb{R}^n$ mit $y^T s > 0$. Die *BFGS-Formel* bestimmt die Matrix

$$B_+^{BFGS} = B + \frac{1}{y^T s} \cdot ((s - By)s^T + s(s - By)^T) - \frac{(s - By)^T y}{(y^T s)^2} ss^T,$$

wobei

$$B_+^{BFGS} y = s.$$

4.2.15 Bemerkung
Die Matrix B_+^{BFGS} löst also die Minimierungsaufgabe (4.2.2), wobei W so gewählt ist, dass $W^T W s = y$.

Es ergibt sich das wichtigste Quasi-Newton-Verfahren überhaupt. Dabei vereinfachen wir in der BFGS-Formel die Terme $s - By$, denn wegen der Sekantenbedingung 4.1.3 gilt diesmal

$$s - By = -B\nabla f(x^+).$$

4.2.16 Algorithmus (BFGS-Verfahren)
{**Broyden, Fletcher, Goldfarb, Shanno**}

wähle x^0
wähle B_0, z.B. $B_0 = \nabla^2 f(x^0)^{-1}$
for $k = 0, 1, \ldots$ **do**
  berechne $s^k = -B_k \nabla f(x^k)^T$
  setze $x^{k+1} = x^k + s^k$
  setze $y^k = \nabla f(x^{k+1})^T - \nabla f(x^k)^T$
  berechne $v^k = -B_k \nabla f(x^{k+1})^T$
  setze $B_{k+1} = B_k + \frac{1}{(y^k)^T s^k} \cdot ((v^k)(s^k)^T + s^k(v^k)^T) - \frac{(v^k)^T y^k}{((y^k)^T s^k)^2} s^k (s^k)^T$.
  {BFGS-Update}
end for

4.2.17 Bemerkung
Vorsicht: Es ist $B_+^{BFGS} \neq \left(H_+^{DFP}\right)^{-1}$, denn $\left(H_+^{DFP}\right)^{-1}$ erfüllt nicht die Minimierungsbedingung aus Lemma 4.2.13. Mit den Bezeichnungen

$$B_+^{DFP} = \left(H_+^{DFP}\right)^{-1}, \quad H_+^{BFGS} = \left(B_+^{BFGS}\right)^{-1}$$

gelten vielmehr die folgenden Zusammenhänge (s. Übung):

$$\begin{aligned}
H^{DFP} &= H + \frac{1}{y^T s} \cdot ((y - Hs)y^T + y(y - Hs)^T) - \frac{(y - Hs)^T s}{(y^T s)^2} yy^T \\
B^{DFP} &= B + \frac{1}{y^T s} ss^T - \frac{1}{y^T By} (By)(By)^T \\
H^{BFGS} &= H + \frac{1}{y^T s} yy^T - \frac{1}{s^T Hs} (Hs)(Hs)^T \\
B^{BFGS} &= B + \frac{1}{y^T s} \cdot ((s - By)s^T + s(s - By)^T) - \frac{(s - By)^T y}{(y^T s)^2} ss^T
\end{aligned}$$

Man beachte die Symmetrien: Ersetzt man (H, y, s) durch (B, s, y), kommt man von den DFP-Formeln zu den BFGS-Formeln.

Abschnitt 4.3

Resultate zur Frobenius-Norm

Bei der Konvergenzanalyse von Quasi-Newton-Verfahren wird es entscheidend darauf ankommen, Frobenius-Normen von (Rang -1-) modifizierten Matrizen gut genug abschätzen zu können. Wir behandeln deshalb hier einige nützliche Eigenschaften der Frobenius-Norm.

4.3.1 Lemma
Sei $A \in \mathbb{R}^{n\times n}$. Dann gilt

$$\|A\| \leq \|A\|_F.$$

($\|A\|$: ℓ_2-Operatornorm)

Beweis: Es bezeichne a_i^T die i-te Zeile von A. Dann ist für jedes x mit $\|x\| = 1$

$$\|Ax\|^2 = \sum_{i=1}^{n} \left(a_i^T x\right)^2 \overset{CSU}{\leq} \left(\sum_{i=1}^{n} \|a_i\|^2\right) \cdot \|x\|^2 = \|A\|_F^2,$$

also

$$\|A\| = \max_{\|x\|=1} \|Ax\| \leq \|A\|_F.$$

□

4.3.2 Lemma
Sei $u, v \in \mathbb{R}^n$. Dann ist

$$\|uv^T\|_F = \|u\|\|v\|$$

und

$$\text{spur}(uv^T) = v^T u.$$

Beweis: Übung. □

Von besonderem Interesse sind bei Quasi-Newton-Verfahren symmetrische und unsymmetrische Rang-1 Modifikationen der Identität. Wir starten mit einer Aussage für den symmetrischen Fall.

4.3.3 Lemma
Sei $s \in \mathbb{R}^n$, $u \neq 0$. Dann gilt für $P = I - \frac{1}{u^T u} uu^T$

$$\|P\| = 1.$$

Beweis: Sei $v_1 = (1/\|u\|)u$, $v_2, \dots v_n$ so, dass $v_1, \dots, v_n$ eine Orthonormalbasis von $\mathbb{R}^n$. Dann ist $Pv_1 = 0$, $Pv_i = v_i$ für $i = 2, \dots, n$. Damit ist $\text{spek}(P) = \{0, 1\}$. Weil P symmetrisch ist, gilt also

$$\|P\| = \max \text{spek}(P) = 1.$$

□

4.3.4 Lemma
Sei $A, B \in \mathbb{R}^{n\times n}$. Dann gelten die Abschätzungen

$$\|AB\|_F \leq \|A\| \cdot \|B\|_F \text{ und } \|AB\|_F \leq \|A\|_F \cdot \|B\|.$$

Beweis: Es bezeichne b_j die j-te Spalte von B. Dann ist Ab_j die j-te Spalte von AB und

$$\|AB\|_F^2 = \sum_{j=1}^{n} \|Ab_j\|_2^2 \leq \|A\|^2 \sum_{j=1}^{n} \|b_j\|_2^2 = \|A\|^2 \|B\|_F^2.$$

Dies beweist die erste Ungleichung. Die zweite ergibt sich, indem man überall die transponierten Matrizen verwendet ($\|\cdot\|$ und $\|\cdot\|_F$ sind invariant gegenüber Transposition). □

Aus dem letzten Lemma folgt insbesondere für $B = P = I - \frac{1}{u^T u} uu^T$

$$\|AP\|_F \leq \|A\|_F.$$

Diese Abschätzung werden wir (auch für unsymmetrische Rang-1-Modifikationen der Identität) noch in einer verbesserten Form benötigen.

4.3.5 Lemma
Sei $A \in \mathbb{R}^{n\times n}, u, v \in \mathbb{R}^n$, $v^T u \neq 0$. Dann gilt

(i)

$$\left\| A\left(I - \frac{1}{u^T v} uv^T\right)\right\|_F^2 = \|A\|_F^2 - \frac{2}{u^T v}(Av)^T(Au) + \frac{1}{(u^T v)^2}\|Au\|^2 \cdot \|v\|^2.$$

(ii) Im Falle $v = u \neq 0$ ist

$$\left\| A(I - \frac{1}{u^T u} uu^T)\right\|_F^2 = \|A\|_F^2 - \frac{1}{u^T u}\|Au\|^2.$$

Beweis: Wir schreiben $\alpha = 1/(u^T v)$. Dann ist

$$\begin{aligned}
&\left\|A\left(I - \alpha u v^T\right)\right\|_F^2 \\
&= \operatorname{spur}\left((I - \alpha v u^T) A^T A (I - \alpha u v^T)\right) \\
&= \operatorname{spur}(A^T A) - \alpha \cdot \operatorname{spur}\left(v u^T (A^T A)\right) \\
&\quad - \alpha \cdot \operatorname{spur}\left((A^T A) u v^T\right) + \alpha^2 \cdot \operatorname{spur}\left(v (u^T A^T A u) v^T\right) \\
&= \|A\|_F^2 - \alpha \cdot (u^T A^T A) v - \alpha \cdot v^T (A^T A u) + \alpha^2 \cdot (u^T A^T A u) \cdot v^T v \\
&= \|A\|_F^2 - 2\alpha \cdot (Av)^T (Au) + \alpha^2 \cdot (Au)^T (Au) \cdot v^T v.
\end{aligned}$$

Dies beweist (i). Teil (ii) ist ein Spezialfall. □

Es wird später darauf ankommen, die rechten Seiten oben explizit ins Verhältnis zu $\|A\|_F$ zu setzen. Im Falle (ii) ergibt sich sofort

$$\left\|A\left(I - \frac{1}{u^T u} u u^T\right)\right\|_F^2 \leq \left(1 - \Theta^2\right) \cdot \|A\|_F^2$$

mit

$$\Theta = \begin{cases} \frac{\|Au\|}{\|A\|_F \|u\|} & \text{falls } A \neq 0 \\ 0 & \text{sonst.} \end{cases}.$$

Im unsymmetrischen Fall $u \neq v$ kann man ein entsprechendes Resultat nur erwarten, wenn der Unterschied zwischen u und v in die Abschätzung mit aufgenommen wird. Wir formulieren dies präzise in dem folgenden wichtigen Lemma.

4.3.6 Lemma
Sei $A \in \mathbb{R}^{n \times n}, u, v \in \mathbb{R}^n$, $u \neq 0$. Weiter sei

$$\|u - v\| \leq \beta \cdot \|u\| \text{ mit } \beta \in [0, 1/3].$$

Dann gelten mit

$$\Theta = \begin{cases} \frac{\|Au\|}{\|A\|_F \|u\|} & \text{falls } A \neq 0 \\ 0 & \text{sonst.} \end{cases}.$$

die Abschätzungen

(i)

$$\begin{aligned}
\left\|A\left(I - \frac{1}{u^T v} u v^T\right)\right\|_F &\leq \left(1 - \frac{1-\beta}{2(1+\beta)}\Theta^2 + \frac{1}{1-\beta} \cdot \frac{\|u - v\|}{\|u\|}\right) \cdot \|A\|_F \\
&\leq \left(1 - \frac{1}{4}\Theta^2 + \frac{3}{2} \cdot \frac{\|u - v\|}{\|u\|}\right) \cdot \|A\|_F
\end{aligned}$$

(ii)
$$\left\| A(I - \frac{1}{u^T u} u u^T) \right\|_F \leq \left(1 - \frac{1}{2}\Theta^2\right) \cdot \|A\|_F$$

Beweis: Offensichtlich ist (ii) der Spezialfall für (i) mit $v = u$, $\beta = 0$. Die zweite Zeile von (i) folgt sofort aus der ersten, da für $\beta \in [0, 1/3]$ die Abschätzungen
$$\frac{1-\beta}{2(1+\beta)} \geq \frac{1}{4}, \quad \frac{1}{1-\beta} \leq \frac{3}{2}$$
gelten. Zum Beweis der ersten Zeile von (i) stellen wir zunächst einmal fest, dass wegen $\|u - v\| \leq \beta \cdot \|u\|$ aus der CSU die Abschätzung
$$|v^T u - u^T u| = |(v-u)^T u| \leq \beta \cdot \|u\|^2$$
folgt und damit
$$(1-\beta) \cdot \|u\|^2 \leq v^T u \leq (1+\beta) \cdot \|u\|^2. \tag{4.3.1}$$
Insbesondere ist $v^T u$ stets nichtnegativ.
Auf Grund von Lemma 4.3.5 haben wir
$$\begin{aligned} \left\| A\left(I - \frac{1}{u^T v} u v^T\right) \right\|_F^2 &= \|A\|_F^2 - \frac{2}{v^T u}(Av)^T(Au) + \frac{1}{(v^T u)^2}\|Au\|^2 \cdot \|v\|^2 \\ &= \|A\|_F^2 + \left(\frac{\|v\|^2}{(v^T u)^2} - \frac{2}{v^T u}\right) \cdot \|Au\|^2 \\ &\quad - \frac{2}{v^T u}\left(A(v-u)\right)^T (Au). \end{aligned}$$

Für den dritten Summanden auf der rechten Seite ergibt sich durch wenig raffiniertes Abschätzen
$$\begin{aligned} -\frac{2}{v^T u}\left(A(v-u)\right)^T (Au) &\leq \frac{2}{v^T u}\|A\|^2 \cdot \|v-u\| \cdot \|u\| \\ &\overset{(4.3.1)}{\leq} \frac{2}{(1-\beta)} \cdot \frac{\|v-u\|}{\|u\|} \cdot \|A\|^2 \\ &\leq \frac{2}{(1-\beta)} \cdot \frac{\|v-u\|}{\|u\|} \cdot \|A\|_F^2. \end{aligned}$$

Für den zweiten Summanden haben wir
$$\frac{\|v\|^2}{(v^T u)^2} - \frac{2}{v^T u} = \frac{v^T v - 2v^T u}{(v^T u)^2} = \frac{(v-u)^T(v-u) - u^T u}{(v^T u)^2}$$

und damit

$$\frac{\|v\|^2}{(v^Tu)^2} - \frac{2}{v^Tu} \leq \frac{(\beta^2-1)\cdot\|u\|^2}{(v^Tu)^2} \overset{(4.3.1)}{\leq} \frac{\beta^2-1}{(1+\beta)^2}\cdot\frac{1}{\|u\|^2} = -\frac{1-\beta}{1+\beta}\cdot\frac{1}{\|u\|^2}.$$

Im Falle $A \neq 0$ erhalten wir so insgesamt

$$\left\|A\left(I - \frac{1}{u^Tv}uv^T\right)\right\|_F^2 \leq \left(1 - \frac{1-\beta}{1+\beta}\Theta^2 + \frac{2}{1-\beta}\cdot\frac{\|v-u\|}{\|u\|}\right)\cdot\|A\|_F^2,$$

was auch für $A = 0$ richtig ist. Unter Verwendung von $(1+t)^{1/2} \leq 1 + t/2$ erhalten wir nach Ziehen der Wurzel schließlich die erste Zeile von (i). $\square$

Abschnitt 4.4

Konvergenz des PSB-Verfahrens

In diesem Abschnitt führen wir, vor allem als Vorbereitung für die entsprechenden Resultate beim BFGS-Verfahren, eine lokale Konvergenzanalyse für das PSB-Verfahren durch.
Das PSB-Verfahren ist das Quasi-Newton-Verfahren, bei welchem die Matrix H gemäß der PSB-Formel aus Satz 4.2.5 aufdatiert wird.

4.4.1 Algorithmus (PSB-Verfahren)
wähle x^0
wähle H_0, z.B. $H_0 = \nabla^2 f(x^0)$
for $k = 0, 1, \ldots$ **do**
löse $H_k s^k = -\nabla f(x^k)^T$
setze $x^{k+1} = x^k + s^k$
setze $y^k = \nabla f(x^{k+1})^T - \nabla f(x^k)^T$
setze $H_{k+1} = H_k + \frac{1}{(s^k)^T s^k} \cdot ((y^k - H_k s^k)(s^k)^T + s^k(y^k - H_k s^k)^T) - \frac{(y^k - H_k s^k)^T s^k}{((s^k)^T s^k)^2} s^k (s^k)^T$.
{PSB-Update}
end for

Unser Ziel ist eine lokale Konvergenzaussage der Art: Ist x^* lokale Minimalstelle und ist x^0 nahe genug an x^* sowie H^0 nahe genug an $\nabla^2 f(x^0)$, so konvergieren die Iterierten q-linear. Daraus wird sich dann sogar die q-überlineare Konvergenz ergeben.
Im Folgenden wird es wichtig sein, aus $H - \nabla^2 f(x)$ auf $H_+^{PSB} - \nabla^2 f(x)$ schließen zu können. Dazu formulieren wir

4.4.2 Lemma
Sei $H \in \mathbb{R}^{n\times n}$ symmetrisch, $A \in \mathbb{R}^{n\times n}$*symmetrisch* sowie $s, y \in \mathbb{R}^n$ mit $s \neq 0$. Dann gilt für die PSB-aufdatierte Matrix H_+^{PSB} mit $H_+^{PSB} s = y$

$$H_+^{PSB} - A = P^T(H - A)P + \frac{1}{s^T s} \cdot (y - As)s^T + \frac{1}{s^T s} \cdot s(y - As)^T P$$

mit

$$P = I - \frac{1}{s^T s} ss^T.$$

Beweis: Das Resultat ergibt sich durch Ausmultiplizieren und Zusammenfassen. □
Auf der Grundlage dieser Darstellung werden wir nun entscheidende Abschätzungen vornehmen.

4.4.3 Lemma
Seien $A, H \in \mathbb{R}^{n\times n}$ symmetrisch, $y, s \in \mathbb{R}^n$, $s \neq 0$. Dann gilt für die PSB-Matrix H_+^{PSB} mit $H_+^{PSB}s = y$ die Abschätzung

$$\|H_+^{PSB} - A\|_F \leq \|H - A\|_F + 2\frac{\|y - As\|}{\|s\|}.$$

Beweis: Auf Grund von Lemma 4.4.2 haben wir

$$\|H_+^{PSB}-A\|_F \leq \|P^T(H-A)P\|_F+\frac{1}{s^Ts}\cdot\|(y-As)s^T\|_F+\frac{1}{s^Ts}\cdot\|s(y-As)^TP\|_F.$$

mit $P = I - \frac{1}{s^Ts}ss^T$. Nach Lemma 4.3.3 ist $\|P\| = 1$. Für den ersten Term auf der rechten Seite erhalten wir so durch zweimalige Anwendung von Lemma 4.3.4

$$\|P^T(H - A)P\|_F \leq \|(H - A)P\|_F \leq \|H - A\|_F.$$

Nach Lemma 4.3.2 erhalten wir außerdem für den zweiten Term

$$\|(y - As)s^T\|_F = \|s(y - As)^T\|_F = \|s\| \cdot \|y - As\|,$$

woraus sich die behauptete Abschätzung ergibt. □

Das letzte Resultat werden wir gleich für $A = \nabla^2 f(x^*)$ mit lokal Lipschitzstetigem $\nabla^2 f$ anwenden.

4.4.4 Definition
Eine Funktion $F : \mathbb{R}^n \to \mathbb{R}^m$ heißt *lokal Lipschitz stetig* in $x \in \mathbb{R}^n$ mit Lipschitzkonstante γ ($F \in \mathrm{Lip}_\gamma(x)$), falls es eine Umgebung U von x gibt, so dass gilt

$$\|F(y) - F(z)\| \leq \gamma\|y - z\| \text{ für alle } y, z \in U.$$

4.4.5 Lemma
Sei $f : \mathbb{R}^n \to \mathbb{R}$, $f \in \mathcal{C}^2(\mathbb{R}^n)$, $x^* \in \mathbb{R}^n$ und $\nabla^2 f \in \mathrm{Lip}_\gamma(x^*)$. Weiter sei $H \in \mathbb{R}^{n\times n}$ symmetrisch.
Dann existiert eine Umgebung U von x^*, so dass für $x, x^+ \in U$ und die PSB-Matrix H_+^{PSB} mit $H_+^{PSB}s = y$, $s = x^+ - x$, $y = \nabla f(x^+)^T - \nabla f(x)^T$ gilt

$$\|H_+^{PSB} - \nabla^2 f(x^*)\|_F \leq \|H - \nabla^2 f(x^*)\|_F + \gamma \cdot (\|x^+ - x^*\| + \|x - x^*\|).$$

Beweis: Nach Lemma 4.4.3 ist mit $A = \nabla^2 f(x^*)$

$$\|H_+^{PSB} - A\|_F \leq \|H - A\|_F + 2\frac{\|y - As\|}{\|s\|}.$$

Jetzt wenden wir das alte Lemma 3.2.6 an, wonach

$$\begin{aligned} \|y - As\| &= \|\nabla f(x^+)^T - \nabla f(x)^T - \nabla^2 f(x^*)(x^+ - x)\| \\ &\leq \frac{\gamma}{2} \cdot \|\underbrace{x^+ - x}_{=s}\| \cdot (\|x^+ - x^*\| + \|x - x^*\|) \end{aligned}$$

gilt, woraus die Behauptung folgt. □

Jetzt haben wir alle Elemente für den Beweis eines lokalen Konvergenzsatzes beisammen.

4.4.6 Satz
Sei $f : \mathbb{R}^n \to \mathbb{R}$, $f \in \mathcal{C}^2(\mathbb{R}^n)$. $x^* \in \mathbb{R}^n$ sei ein stationärer Punkt von f und $\nabla^2 f(x^*)$ sei regulär. Weiter sei $f \in \mathrm{Lip}_\gamma(x^*)$. Dann existieren Zahlen $\varepsilon > 0$ und $\delta > 0$, so dass das PSB-Verfahren (Algorithmus 4.4.1) für jeden Startvektor $x^0 \in B_\varepsilon(x^*)$ und jede Wahl einer symmetrischen Matrix H_0 mit $\|H_0 - \nabla^2 f(x^*)\| < \delta$ durchgeführt werden kann und eine Iteriertenfolge $\{x^k\}$ generiert, welche q-linear gegen x^* konvergiert.

Beweis: Wir starten mit einer Vorüberlegung. Es ist

$$\begin{aligned} \|x^{k+1} - x^*\| &= \|x^k - x^* - H_k^{-1}\nabla f(x^k)^T\| \\ &\leq \|H_k^{-1}\| \cdot \|H_k(x^k - x^*) - \nabla f(x^k)^T - \nabla f(x^*)^T\| \\ &\leq \|H_k^{-1}\| \cdot \big(\|\nabla^2 f(x^*)(x^k - x^*) - \nabla f(x^k)^T - \nabla f(x^*)^T\| \\ &\quad + \|(H_k - \nabla^2 f(x^*)\| \cdot \|(x^k - x^*)\|\big) \end{aligned} \tag{4.4.1}$$

Eine Aussage über lineare Konvergenz erhalten wir also, wenn wir zeigen können, dass $\|H_k^{-1}\| \leq \eta$, $\|H_k - \nabla^2 f(x^*)\| \leq \alpha$ und $\|\nabla^2 f(x^*)(x^k - x^*) - \nabla f(x^k)^T - \nabla f(x^*)^T\| < \beta \cdot \|x^k - x^*\|$ für alle k mit $\eta(\alpha + \beta) \leq \rho < 1$. Hierin wird β beliebig klein, wenn ε gegen 0 geht. Auch α wird hoffentlich beliebig klein, wenn wir δ klein genug machen.
Wir setzen $\mu = \|\nabla^2 f(x^*)^{-1}\|$. Für festes $\rho \in (0, 1)$ wählen wir nun $\varepsilon, \delta > 0$ so klein, dass in $B_\varepsilon(x^*)$ die lokale Lipschitzbedingung gilt und dass

$$\frac{2\mu}{1 - \rho}(\gamma\epsilon + \delta) \leq \rho, \quad 2\delta \cdot \mu \leq \rho, \quad \frac{2\gamma\epsilon}{1 - \rho} \leq \delta \tag{4.4.2}$$

und zeigen, dass für alle k gilt

(i) $\|H_k - \nabla^2 f(x^*)\|_F \leq 2\delta$,

(ii) H_k ist regulär mit $\|H_k^{-1}\| \leq \frac{\mu}{1-\rho}$,

(iii) $\|x^{k+1} - x^k\| \leq \rho \cdot \|x^k - x^*\|$.

Wir beweisen alle drei Punkte per Induktion. Für $k = 0$ ist (i) erfüllt, denn wir haben ja sogar $\|H_0 - \nabla^2 f(x^*)\|_F \leq \delta$. Aus

$$H_0 = \nabla^2 f(x^*) + (H_0 - \nabla^2 f(x^*))$$

mit $\|\nabla^2 f(x^*)^{-1} (H_0 - \nabla^2 f(x^*))\| \leq \mu \cdot \|H_0 - \nabla^2 f(x^*)\|_F \leq \mu\delta$ folgt nach dem Banach-Lemma 1.2.7 sofort, dass H_0 regulär ist mit

$$\|H_0^{-1}\| \leq \frac{\mu}{1-\mu\delta} \overset{(4.4.2)}{\leq} \frac{\mu}{1-\rho}.$$

Schließlich erhalten wir aus (4.4.1) für $k = 0$ und unter Verwendung von Lemma 3.2.6 die Beziehung

$$\|x^1 - x^*\| \leq \frac{\mu}{1-\rho} \cdot \left(\frac{\gamma}{2} \cdot \|x^0 - x^*\|^2 + 2\delta \cdot \|x^0 - x^*\| \right),$$

woraus durch großzügiges Abschätzen nach oben

$$\|x^1 - x^*\| \leq \frac{2\mu}{1-\rho} \cdot (\gamma\varepsilon + \delta) \cdot \|x^0 - x^*\| \leq \rho \cdot \|x^0 - x^*\|$$

folgt.
Zum Beweis des Induktionsschrittes seien (i) bis (iii) als richtig angenommen bis hin zu einem Index k. Wenn wir die Gültigkeit von (i) für $k+1$ gezeigt haben, sind wir fertig, denn (ii) und (iii) folgen genauso wie beim Induktionsanfang.
Zum Beweis von (i) erhalten wir aus Lemma 4.4.5 für alle $j = 0, \ldots, k$ die Beziehung

$$\|H_{j+1} - \nabla^2 f(x^*)\|_F - \|H_j - \nabla^2 f(x^*)\|_F \leq \gamma \cdot (\|x^{j+1} - x^*\| + \|x^j - x^*\|)$$

und damit nach mehrfacher Verwendung von (iii)

$$\|H_{j+1} - \nabla^2 f(x^*)\|_F - \|H_j - \nabla^2 f(x^*)\|_F \leq 2\gamma\rho^j \|x^0 - x^*\|.$$

Durch Summation ergibt sich hieraus

$$\|H_{k+1} - \nabla^2 f(x^*)\|_F - \|H_0 - \nabla^2 f(x^*)\|_F$$

$$\begin{aligned} &\leq 2\gamma \cdot \|x^0 - x^*\| \sum_{j=0}^{k} \rho^j \\ &= 2\gamma \cdot \|x^0 - x^*\| \cdot \frac{1-\rho^{k+1}}{1-\rho} \leq \frac{2\gamma}{1-\rho} \cdot \varepsilon. \end{aligned}$$

Also ist

$$\|H_{k+1} - \nabla^2 f(x^*)\|_F \leq \delta + \frac{2\gamma}{1-\rho} \cdot \varepsilon \overset{(4.4.2)}{\leq} 2\delta.$$

□

Der vorangegangene Satz zeigt die lineare Konvergenz des PSB-Verfahrens. Tatsächlich ist die Konvergenz dann sogar q-überlinear, was wir in dem nun folgenden Satz nachweisen werden. Bemerkenswerterweise – und im Gegensatz zu Satz 4.4.6 – benötigt dieser Satz nur eine Aussage über die Konvergenz der Iterierten x^k, aber keinerlei Voraussetzung über die Qualität von B_0. Wir bereiten den Satz mit einer gegenüber Lemma 4.4.3 verbesserten Abschätzung vor.

4.4.7 Lemma
Seien $A, H \in \mathbb{R}^{n\times n}$ symmetrisch, $y, s \in \mathbb{R}^n$, $s \neq 0$. Dann gilt für die PSB-Matrix H_+^{PSB} mit $H_+^{PSB}s = y$ die Abschätzung

$$\|H_+^{PSB} - A\|_F \leq \|H - A\|_F \cdot \left(1 - \frac{\Theta^2}{2}\right) + 2\frac{\|y - As\|}{\|s\|},$$

wobei

$$\Theta = \begin{cases} \frac{\|(H-A)s\|}{\|H-A\|_F\|s\|} & \text{falls } H \neq A \\ 0 & \text{sonst.} \end{cases}$$

Beweis: Auf Grund von Lemma 4.4.2 haben wir

$$\|H_+^{PSB} - A\|_F \leq \|P^T(H-A)P\|_F + \|\frac{1}{s^Ts}\cdot(y-As)s^T\|_F + \frac{1}{s^Ts}\cdot\|s(y-As)^TP\|_F.$$

mit $P = I - \frac{1}{s^Ts}ss^T$. Nach Lemma 4.3.3 ist $\|P\| = 1$. Für den ersten Term auf der rechten Seite erhalten wir so aus Lemma 4.3.4

$$\|P^T(H-A)P\|_F \leq \|(H-A)P\|_F$$

und daraus mit Lemma 4.3.6 (ii) sogar

$$\|P^T(H-A)P\|_F \leq \|H-A\|_F \cdot \left(1 - \frac{\Theta^2}{2}\right).$$

Nach Lemma 4.3.2 erhalten wir außerdem

$$\|(y - As)s^T\|_F = \|s(y - As)^T\| = \|s\| \cdot \|y - As\|,$$

woraus sich die behauptete Abschätzung ergibt. □

4.4.8 Satz
Sei $f : \mathbb{R}^n \to \mathbb{R}$, $f \in \mathcal{C}^2(\mathbb{R}^n)$. $x^* \in \mathbb{R}^n$ sei ein stationärer Punkt von f und $\nabla^2 f(x^*)$ sei regulär. Weiter sei $f \in \mathrm{Lip}_\gamma(x^*)$. Die Iterierten x^k des PSB-Verfahrens (Algorithmus 4.4.1) mögen gegen x^* konvergieren mit

$$\sum_{k=0}^{\infty} \|x^k - x^*\| \leq \xi < \infty.$$

Dann konvergiert die Folge $\{x^k\}$ bereits q-überlinear gegen x^*.

Beweis: Wir können $x^k \neq x^*$ für alle k annehmen und auch $s^k = x^{k+1} - x^k \neq 0$ für alle k (sonst wäre H_{k+1} nicht definiert).
Gemäß Satz 4.1.1 müssen wir zeigen

$$\| \left(H_k - \nabla^2 f(x^*)\right) \left(x^{k+1} - x^k\right) \| = o(\|x^{k+1} - x^k\|). \tag{4.4.3}$$

Es sei zunächst $H_k \neq \nabla^2 f(x^*)$. Aus Lemma 4.4.7 erhalten wir mit

$$s^k = x^{k+1} - x^k,\ y^k = \nabla f(x^{k+1})^T - \nabla f(x^k)^T,\ \Theta_k = \frac{\|(H_k - \nabla^2 f(x^*))s^k\|}{\|(H_k - \nabla^2 f(x^*)\|_F \|s^k\|}.$$

die Abschätzung

$$\|H_{k+1} - \nabla^2 f(x^*)\|_F \leq \|H_k - \nabla^2 f(x^*)\|_F \cdot \left(1 - \frac{\Theta_k^2}{2}\right) + \frac{2}{\|s^k\|} \cdot \left(\|y^k - \nabla^2 f(x^*) s^k\|\right).$$

Durch Anwendung des Lemmas 3.2.6 auf den zweiten Term erhalten wir

$$\begin{aligned} \|H_{k+1} - \nabla^2 f(x^*)\|_F \quad &\leq \quad \|H_k - \nabla^2 f(x^*)\|_F \cdot \left(1 - \frac{\Theta_k^2}{2}\right) \\ &\quad + \gamma \cdot (\|x^{k+1} - x^*\| + \|x^k - x^*\|). \end{aligned} \tag{4.4.4}$$

Daraus folgt durch wiederholte Anwendung (und mit $1 - \Theta^2/2 \leq 1$)

$$\begin{aligned} &\|H_{k+1} - \nabla^2 f(x^*)\|_F \\ &\leq \quad \|H_0 - \nabla^2 f(x^*)\|_F + \gamma \sum_{j=0}^{k} (\|x^{j+1} - x^*\| + \|x^j - x^*\|) \end{aligned}$$

$$\leq \quad \|H_0 - \nabla^2 f(x^*)\|_F + 2\gamma\xi.$$

Also sind die $\sigma_k := \|H_k - \nabla^2 f(x^*)\|_F$ beschränkt. Aus (4.4.4) folgt nun außerdem

$$\frac{\Theta_k^2}{2}\sigma_k \leq \sigma_k - \sigma_{k+1} + \gamma \cdot (\|x^{k+1} - x^*\| + \|x^k - x^*\|).$$

Diese Ungleichung ist auch im Fall $\sigma_k = 0$ (dann mit $\Theta_k = 0$) richtig. Summieren wir für alle k mit $\sigma_k \neq 0$, so erkennen wir

$$\sum_{k=0,\sigma_k\neq 0}^{\infty} \frac{\Theta_k^2}{2}\sigma_k \leq \sigma_0 + 2\gamma\xi < \infty,$$

woraus insbesondere $\lim_{k\to\infty} \Theta_k^2\sigma_k = 0$ folgt. Weil die σ_k beschränkt sind, folgt daraus $\lim_{k\to\infty} \Theta_k^2\sigma_k^2 = 0$ und damit auch

$$\Theta_k\sigma_k = \frac{\|(H_k - \nabla^2 f(x^*))s^k\|}{\|s^k\|} \to 0 \ (k \to \infty),$$

was wegen $s^k = x^{k+1} - x^k$ gleichbedeutend mit (4.4.3) ist. □

Abschnitt 4.5

Konvergenz des BFGS-Verfahrens

Wir gehen im Prinzip ähnlich vor wie beim PSB-Verfahren. Eine zusätzliche Schwierigkeit entsteht durch die unsymmetrische Aufdatierungsformel.
Das BFGS-Verfahren wurde als Algorithmus 4.2.16 beschrieben. Die wesentlichen Schritte sind

$$\begin{aligned} x^{k+1} &= x^k - B_k \nabla f(x^k)^T \\ B_{k+1} &= B_k + \tfrac{1}{(y^k)^T s^k}\left((s^k - B_k y^k)(s^k)^T + s^k(s^k - B_k y^k)^T\right) \\ &\quad - \tfrac{(s^k - B_k y^k)^T y_k}{\left((y^k)^T s^k\right)^2} s^k (s^k)^T \\ s^k &= x^{k+1} - x^k,\ y^k = \nabla f(x^{k+1})^T - \nabla f(x^k)^T. \end{aligned}$$

Beim BFGS-Verfahren arbeitet man also mit unsymmetrischen Rang-1 Modifikationen der Identität von der Bauart

$$I - \frac{1}{y^T s} s y^T$$

mit $s = x^+ - x$, $y = \nabla f(x^+)^T - \nabla f(x)^T$. In Lemma 4.3.6 hatten wir gesehen, dass wir eine zu Lemma 4.4.7 beim PSB-Verfahren analoge Abschätzung nur erreichen können, wenn sich s und y relativ wenig unterscheiden. Dies ist von vornherein nicht gegeben, aber wir können es durch eine Skalierung erreichen. Es ist nämlich

$$y = \nabla^2 f(x) s + o(\|s\|)$$

und damit in der Nähe einer Nullstelle x^* von ∇f auch

$$y \approx \nabla^2 f(x^*) s.$$

Faktorisieren wir $\nabla^2 f(x^*) = W^T W$, so haben wir

$$W^{-T} y \approx W s.$$

Ab sofort notieren wir deshalb zur Abkürzung

$$y_W = W^{-T} y,\ s_W = W s.$$

Man beachte, dass $y^T s = y_W^T s_W$ und damit

$$I - \frac{1}{y^T s} s y^T = W^{-1} \cdot \left(I - \frac{1}{y_W^T s_W} s_W y_W^T \right) \cdot W.$$

Wir präziseren unsere Überlegungen im folgenden Hilfssatz.

4.5.1 Lemma
Sei $f : \mathbb{R}^n \to \mathbb{R}, f \in \mathcal{C}^2(\mathbb{R}^n), x^* \in \mathbb{R}^n$ mit $\nabla f(x^*) = 0$ und $\nabla^2 f(x^*)$ spd, $\nabla^2 f \in \mathrm{Lip}_\gamma(x^*)$. Weiter sei $\nabla^2 f(x^*) = W^T W$ und $\beta \in (0, 1/3]$. Dann existiert $\varepsilon > 0$, so dass für $x, x^+ \in B_\varepsilon(x^*)$ gilt

$$\|s_W - y_W\| \leq \beta \cdot \|y_W\|, \; s_W = W(x^+ - x), \; y_W = W^{-T}(\nabla f(x^+)^T - \nabla f(x)^T).$$

Beweis: $\nabla^2 f$ sei Lipschitz-stetig in $B_\varepsilon(x^*)$ mit Konstante γ. Damit haben wir für $x, x^+ \in B_\varepsilon(x^*)$ nach Lemma 3.2.6

$$\begin{aligned} \|s_W - y_W\| &= \|Ws - W^{-T}y\| \\ &= \|W^{-T}(W^T W s - y)\| \\ &\leq \|W^{-T}\| \cdot \|\nabla^2 f(x^*)s - y\| \\ &\leq \|W^{-T}\| \cdot \frac{\gamma}{2} \cdot \|x^+ - x\| \cdot (\|x^+ - x^*\| + \|x - x^*\|). \end{aligned}$$

Indem wir ε notfalls verkleinern, erreichen wir, dass $\nabla^2 f$ glm. positiv definit ist in $B_\varepsilon(x^*)$. Also ist ∇f glm. monoton in $B_\varepsilon(x^*)$ und es existiert $\mu > 0$ mit

$$(\nabla f(x^+) - \nabla f(x))(x^+ - x) \geq \mu \|x^+ - x\|^2 \text{ für } x^+, x \in B_\varepsilon(x^*).$$

Über die CSU ergibt sich daher

$$\|y\| = \|\nabla f(x^+) - \nabla f(x)\| \geq \mu \|x^+ - x\|. \tag{4.5.1}$$

Durch Einsetzen erhalten wir so

$$\begin{aligned} \|s_W - y_W\| &\leq \frac{\gamma}{2\mu} \cdot (\|x^+ - x^*\| + \|x - x^*\|) \cdot \|W^{-T}\| \cdot \|y\| \\ &\leq \frac{\gamma}{\mu} \cdot \varepsilon \cdot \|W^{-T}\| \cdot \|y\| \\ &\leq \frac{\gamma}{\mu} \varepsilon \cdot \|W^{-T}\| \cdot \|W\| \cdot \|y_W\|. \end{aligned} \tag{4.5.2}$$

Man hat also ε eventuell noch so weit zu verkleinern, dass $\frac{\gamma}{\mu}\varepsilon \cdot \|W^{-T}\| \cdot \|W\| \leq \beta$ gilt. □

Im Folgenden kann man sich unter W also schon immer eine reguläre Matrix mit $\nabla^2 f(x^*) = W^T W$ vorstellen, auch wenn die Resultate erst einmal für beliebiges reguläres W gelten.
Wir fahren fort mit einem Pendant zu Lemma 4.4.2.

4.5.2 Lemma
Sei $A, B \in \mathbb{R}^{n\times n}$ symmetrisch, $W \in \mathbb{R}^{n\times n}$ regulär, $y, s \in \mathbb{R}^n$ mit $y^T s > 0$. Dann gilt für die bzgl. s und y aufdatierte Matrix B_+^{BFGS}

$$\begin{aligned} W(B_+^{BFGS} - A)W^T &= P^T(W(B-A)W^T)P + \frac{1}{y^T s}(W(s-Ay)(Ws)^T) \\ &\quad + \frac{1}{y^T s}(Ws(s-Ay)^T W^T)P \end{aligned}$$

mit

$$P = I - \frac{1}{y^T s} y_W s_W^T = I - \frac{1}{y_W^T s_W} y_W s_W^T.$$

Beweis: Ausmultiplizieren und zusammenfassen. □

Mit dieser Darstellung erhalten wir wie beim PSB-Verfahren eine Abschätzung in der Frobenius-Norm.

4.5.3 Lemma
Seien $B, A \in \mathbb{R}^{n\times n}$ symmetrisch, B_+ die BFGS-aufdatierte Matrix bzgl. $s, y \in \mathbb{R}^n$ mit $y^T s > 0$. Weiter sei $W \in \mathbb{R}^{n\times n}$ regulär und

$$\|s_W - y_W\| \leq \beta \cdot \|y_W\| \text{ mit } \beta \in [0, \tfrac{1}{3}]. \tag{4.5.3}$$

Dann gilt

$$\begin{aligned} \|W(B_+ - A)W^T\|_F &\leq \left((1 - \frac{\alpha}{2}\Theta^2) + \frac{5 \cdot \|s_W - y_W\|}{2 \cdot (1-\beta) \cdot \|y_W\|}\right) \cdot \|W(B-A)W^T\|_F \\ &\quad + 2(1 + 2\sqrt{n}) \cdot \|W\|_F \cdot \frac{\|s - Ay\|}{\|y_W\|} \end{aligned}$$

mit

$$\begin{aligned} \alpha &= \frac{1-\beta}{1+\beta} \in [\tfrac{1}{2}, 1], \\ \Theta &= \begin{cases} \frac{\|W(B-A)y\|}{\|W(B-A)W^T\|_F \cdot \|y_W\|} & \text{falls } W(B-A)W^T \neq 0 \\ 0 & \text{sonst} \end{cases}. \end{aligned}$$

Beweis: Wir halten zunächst fest, dass wie im Beweis von Lemma 4.3.6 aus der Voraussetzung (4.5.3) die Beziehung

$$\frac{2}{3} \cdot \|y_W\|^2 \leq (1-\beta) \cdot \|y_W\|^2 \leq y^T s = y_W^T s_W \leq (1+\beta) \cdot \|y_W\|^2 \leq \frac{4}{3} \cdot \|y_W\|^2 \tag{4.5.4}$$

folgt. Insbesondere ist also $y^T s > 0$. Aus Lemma 4.5.2 erhalten wir

$$\begin{aligned}\|W(B_+ - A)W^T\|_F &\leq \|P^T W(B-A)W^T P\|_F + \frac{1}{s^T y}\cdot\|W(s-Ay)(Ws)^T\|_F \\ &\quad + \frac{1}{s^T y}\cdot\|Ws(s-Ay)^T W^T P\|_F, \end{aligned} \tag{4.5.5}$$

wobei

$$P = I - \frac{1}{s_W^T y_W} y_W s_W^T.$$

Zur Vereinfachung notieren wir $E = W(B-A)W^T$. Aus Lemma 4.3.6 (i) erhalten wir unter Verwendung von $1 - \frac{\alpha}{2}\Theta^2 \leq 1$

$$\|P^T E P\|_F \leq \left(1 + \frac{1}{1-\beta}\cdot\frac{\|s_W - y_W\|}{\|y_W\|}\right)\|P^T E\|_F, \tag{4.5.6}$$

und nach nochmaliger Anwendung (auf $P^T E = E^T P = EP$ mit $\|P^T E\|_F = \|EP\|_F$)

$$\|P^T E\|_F \leq \left(1 - \frac{\alpha}{2}\Theta^2 + \frac{1}{1-\beta}\frac{\|s_W - y_W\|}{\|y_W\|}\right)\|E\|_F. \tag{4.5.7}$$

Aus (4.5.6) und (4.5.7) erhalten wir so unter Ausnutzung von (4.5.3)

$$\|P^T E P\|_F \leq \left(1 - \frac{\alpha}{2}\Theta^2 + \frac{5}{2(1-\beta)}\cdot\frac{\|s_W - y_W\|}{\|y_W\|}\right)\cdot\|E\|_F.$$

Es bleiben die beiden übrigen Terme aus (4.5.5) abzuschätzen. Dazu werden wir immer wieder auf (4.5.4) zurückgreifen. Nach Lemma 4.3.4 und 4.3.2 haben wir unter Verwendung von $\|s_W\| \leq (1+\beta)\cdot\|y_W\|$

$$\begin{aligned}\frac{1}{s^T y}\|W(s-Ay)s_W^T\| &\leq \frac{1}{s^T y}\|W\|_F\cdot\|s-Ay\|\cdot\|s_W\| \\ &\leq \underbrace{\frac{1+\beta}{1-\beta}}_{\leq 2}\cdot\|W\|_F\cdot\frac{\|s-Ay\|}{\|y_W\|}.\end{aligned}$$

Für den letzten Term von (4.5.5) verwenden wir zusätzlich

$$\begin{aligned}\|P\| \leq \|P\|_F &\leq \|I\|_F + \frac{1}{s^T y}\|y_W s_W^T\|_F \\ &= \sqrt{n} + \frac{1}{s^T y}\|y_W\|\cdot\|s_W\|\end{aligned}$$

$$\begin{aligned} &\leq \sqrt{n} + \frac{1+\beta}{1-\beta} \\ &\leq \sqrt{n} + 2 \\ &\leq 2\sqrt{n}, \end{aligned}$$

und erhalten so wie vorhin

$$\begin{aligned} \frac{1}{s^T y}\|s_W(s-Ay)^T W^T P\|_F &\leq \frac{1}{s^T y}\|s_W(s-Ay)^T W^T\|_F \cdot 2\sqrt{n} \\ &\leq 2 \cdot \|W\|_F \frac{\|s-Ay\|}{\|y_W\|} \cdot 2\sqrt{n}. \end{aligned}$$

□

Wie beim PSB-Verfahren wird es für den Nachweis der linearen Konvergenz (Satz 4.5.5 unten) genügen, im obigen Resultat statt $(1 - \frac{\alpha}{2}\Theta^2)$ einfach 1 zu verwenden. Die verbesserte Abschätzung wird erst für den Nachweis der überlinearen Konvergenz wichtig.
Mit $A = \nabla^2 f(x^*)$ erhalten wir aus Lemma 4.5.3 jetzt die zentrale Abschätzung für die BFGS-Aufdatierungen.

4.5.4 Lemma
Sei $B \in \mathbb{R}^{n\times n}$ spd, $f : \mathbb{R}^n \to \mathbb{R}, f \in \mathcal{C}^2(\mathbb{R}^n), x^* \in \mathbb{R}^n$ mit $\nabla f(x^*) = 0$, $\nabla^2 f(x^*)$ spd, $\nabla^2 f \in \mathrm{Lip}_\gamma(x^*)$. Sei $\nabla^2 f(x^*) = W^T W, \beta \in [0, \frac{1}{3}]$. Dann existiert $\varepsilon > 0$, so dass für $x, x^+ \in B_\varepsilon(x^*)$ gilt

$$\begin{aligned} \|W(B_+ - \nabla^2 f(x^*)^{-1})W^T\|_F \leq\ & \left(1 - \frac{\alpha}{2}\Theta^2\right) \cdot \|W(B - \nabla^2 f(x^*)^{-1})W^T\|_F \\ &+ \left(\alpha_1 \cdot \|W(B - \nabla^2 f(x^*)^{-1})W^T\|_F + \alpha_2\right) \\ &\cdot \left(\|x^+ - x^*\| + \|x - x^*\|\right) \end{aligned}$$

mit α, Θ wie gehabt und $\alpha_1, \alpha_2 > 0$.

Beweis: Wie immer sei $s = x^+ - x, y = \nabla f(x^+)^T - \nabla f(x)^T$ und $s_W = Ws$, $y_W = W^{-T}y$.
Wir bauen auf Lemma 4.5.3 (mit $A = \nabla^2 f(x^*)^{-1}$) auf, wonach für ε klein genug gilt

$$\begin{aligned} &\|W(B_+ - \nabla^2 f(x^*)^{-1})W^T\|_F \\ (4.5.8)\leq\ & \left((1 - \frac{\alpha}{2}\Theta^2) + \frac{5 \cdot \|s_W - y_W\|}{2(1-\beta)\cdot\|y_W\|}\right) \cdot \|W(B - \nabla^2 f(x^*)^{-1})W^T\|_F \\ &+ 2(1 + 2\sqrt{n})\|W\|_F \cdot \frac{\|s - \nabla^2 f(x^*)^{-1}y\|}{\|y_W\|}. \end{aligned}$$

Hierin haben wir nach (4.5.2) aus dem Beweis von Lemma 4.5.1, evtl. nach Verkleinerung von ε,

$$\frac{\|s_W - y_W\|}{\|y_W\|} \leq \frac{\gamma}{2\mu} \cdot \|W^{-T}\| \cdot \|W\| \cdot (\|x^+ - x^*\| + \|x - x^*\|),$$

μ definiert wie dort. Ebenfalls aus dem Beweis von Lemma 4.5.1 wissen wir (s. (4.5.1))

$$\|y\| \geq \mu \cdot \|\underbrace{x^+ - x}_{s}\| \Rightarrow \|y_W\| \geq \frac{\mu}{\|W^T\|} \cdot \|x^+ - x\|$$

Außerdem ist nach Lemma 3.2.6

$$\begin{aligned}
&\|s - \nabla^2 f(x^*)^{-1} y\| \\
&\leq \|\nabla^2 f(x^*)^{-1}\| \cdot \|\nabla^2 f(x^*) s - y\| \\
&\leq \|\nabla^2 f(x^*)^{-1}\| \cdot \frac{\gamma}{2} \cdot \|x^+ - x\| \cdot (\|x^+ - x^*\| + \|x - x^*\|) \\
&\leq \|\nabla^2 f(x^*)^{-1}\| \cdot \frac{\gamma}{2} \cdot \frac{\|W^T\|}{\mu} \cdot \|y_W\| \cdot (\|x^+ - x^*\| + \|x - x^*\|).
\end{aligned}$$

Damit erhalten wir durch Einsetzen aus (4.5.8) die behauptete Ungleichung mit

$$\alpha_1 = \frac{5\gamma}{4(1-\beta)\mu}\|W^{-T}\| \cdot \|W\|, \alpha_2 = \frac{\gamma}{\mu}(1+2\sqrt{n}) \cdot \|W\|_F \cdot \|W^T\| \cdot \|\nabla^2 f(x^*)^{-1}\|.$$

□

Jetzt können wir endlich die lineare Konvergenz des BFGS-Verfahrens nachweisen.

4.5.5 Satz
Sei $f : \mathbb{R}^n \to \mathbb{R}, f \in \mathcal{C}^2(\mathbb{R}^n)$ mit $\nabla f(x^*) = 0$, $\nabla^2 f \in \mathrm{Lip}_\gamma(x^*), \nabla^2 f(x^*)$ spd. Dann existieren $\varepsilon, \overline{\delta} > 0$, so dass die Iterierten des BFGS-Verfahrens x^k (Algorithmus 4.2.16) für jede Wahl von x^0 mit $\|x^0 - x^*\| \leq \varepsilon$ und jede Wahl von B_0 spd mit $\|B_0 - \nabla^2 f(x^*)^{-1}\|_F \leq \overline{\delta}$ definiert sind und q-linear gegen x^* konvergieren.

Beweis: Sei $\nabla^2 f(x^*) = W^T W$. Wähle $\beta \in (0, \frac{1}{3}]$. Auf Grund der Norm-Äquivalenz in $\mathbb{R}^{n\times n}$ existiert $\eta > 0$ mit

$$\|A\|_F \leq \eta \|W A W^T\|_F \text{ für alle } A \in \mathbb{R}^{n\times n}.$$

Setze weiter

$$\mu = \|\nabla^2 f(x^*)^{-1}\|_F, \ \sigma = \|\nabla^2 f(x^*)\|_F$$

und wähle $\rho \in (0,1)$ fest. Wähle nun ε so klein, dass die Aussage von Lemma 4.5.4 gilt. Durch Verkleinerung von ε und δ erreicht man außerdem

$$\begin{aligned}(2\eta\delta + \mu) \cdot \frac{\gamma}{2} \cdot \varepsilon + 2\sigma\eta\delta &\leq \rho, \\ \frac{2\alpha_1\delta + \alpha_2}{1-\rho} \cdot 2\varepsilon &\leq \delta.\end{aligned}$$

Es sei nun

$$\|x^0 - x^*\| \leq \varepsilon,\ \|W(B_0 - \nabla^2 f(x^*)^{-1})W^T\|_F \leq \delta.$$

Wir zeigen, dass dann gilt

(i) $\left\|W\left(B_k - \nabla^2 f(x^*)^{-1}\right)W^T\right\|_F \leq 2\delta$,

(ii) $\|B_k\|_F \leq 2\eta\delta + \mu$,

(iii) B_k ist regulär mit $\|B_k^{-1}\| \leq \frac{\sigma}{1-\rho}$,

(iv) $\|x^{k+1} - x^*\| \leq \rho \cdot \|x^k - x^*\|$.

Der Satz ist dann bewiesen (mit $\overline{\delta} = \delta/\eta$). Wir verwenden Induktion.
$k = 0$:
(i) gilt sogar mit δ statt 2δ.
Zu (ii): Wir haben

$$\begin{aligned}\|B_0\|_F &\leq \|B_0 - \nabla^2 f(x^*)^{-1}\|_F + \|\nabla^2 f(x^*)^{-1}\|_F \\ &\leq \eta\|W(B_0 - \nabla^2 f(x^*)^{-1})W^T\|_F + \mu \\ &\overset{(i)}{\leq} 2\eta\delta + \mu.\end{aligned}$$

Zu (iii): Aus

$$\begin{aligned}\|B_0 - \nabla^2 f(x^*)^{-1}\| &\leq \|B_0 - \nabla^2 f(x^*)^{-1}\|_F \\ &\leq \eta \cdot \|W(B_0 - \nabla^2 f(x^*)^{-1}W^T\|_F \overset{(i)}{\leq} 2\eta\delta\end{aligned}$$

mit $2\eta\delta\sigma < 1$ folgt nach dem Banach-Lemma 1.2.7, dass die Matrix

$$\nabla^2 f(x^*)^{-1} - (B_0 - \nabla^2 f(x^*)^{-1}) = -B_0$$

regulär ist mit

$$\|B_0^{-1}\| \leq \frac{\sigma}{1 - 2\sigma\eta\delta} \leq \frac{\sigma}{1-\rho}.$$

Zu (iv): Wir haben

$$\begin{aligned} x^1 - x^* &= x^0 - B_0 \nabla f(x^0)^T - x^* \\ &= x^0 - x^* - B_0 \left(\nabla f(x^0)^T - \nabla f(x^*)^T - \nabla^2 f(x^*)(x^0 - x^*) \right) \\ &\quad - B_0 \nabla^2 f(x^*)(x^0 - x^*). \end{aligned}$$

Damit erhalten wir

$$\begin{aligned} \|x^1 - x^*\| \leq\ & \|B_0\| \cdot \|\nabla f(x^0)^T - \nabla f(x^*)^T - \nabla^2 f(x^*)(x^0 - x^*)\| \\ & + \|I - B_0 \nabla^2 f(x^*)\| \cdot \|x^0 - x^*\| \end{aligned}$$

und damit nach Lemma 3.2.6

$$\|x^1 - x^*\| \leq \|B_0\|_F \cdot \frac{\gamma}{2} \cdot \|x^0 - x^*\|^2 + \|I - B_0 \nabla^2 f(x^*)\| \cdot \|x^0 - x^*\|.$$

Hierin ist $\|x^0 - x^*\| \leq \varepsilon$ und

$$\begin{aligned} \|I - B_0 \nabla^2 f(x^*)\| &\leq \|\nabla^2 f(x^*)^{-1} - B_0\| \cdot \|\nabla^2 f(x^*)\| \\ &\leq \|\nabla^2 f(x^*)^{-1} - B_0\|_F \cdot \sigma \\ &\leq \eta \cdot \|W(\nabla^2 f(x^*)^{-1} - B_0) W^T\|_F \cdot \sigma \\ &\overset{(i)}{\leq} 2\eta\delta\sigma. \end{aligned}$$

Insgesamt ergibt sich so mit (ii)

$$\begin{aligned} \|x^1 - x^*\| &\leq \left(\frac{\gamma}{2} \cdot (2\mu\delta + \mu)\varepsilon + 2\eta\delta\sigma \right) \cdot \|x^0 - x^*\| \\ &\leq \rho \cdot \|x^0 - x^*\|. \end{aligned}$$

Damit ist der Induktionsanfang geschafft. Im Induktionsschritt genügt es, (i) nachzuweisen, denn (ii)-(iv) folgen genauso wie für $k = 0$.
$k \curvearrowright k+1$, (i): Auf Grund von Lemma 4.5.4 und unter Verwendung von $1 - \frac{\alpha}{2}\Theta^2 \leq 1$ gilt für $j = 0, 1, \ldots, k$

$$\begin{aligned} &\|W(B_{j+1} - \nabla^2 f(x^*)^{-1}) W^T\|_F - \|W(B_j - \nabla^2 f(x^*)^{-1}) W^T\|_F \\ &\quad \leq (2\alpha_1\delta + \alpha_2) \cdot (\|x^{j+1} - x^*\| + \|x^j - x^*\|) \\ &\quad \leq (2\alpha_1\delta + \alpha_2) \cdot (\rho^{j+1} - \rho^j) \cdot \|x^0 - x^*\| \\ &\quad \leq (2\alpha_1\delta + \alpha_2) \cdot 2\rho^j \cdot \|x^0 - x^*\|. \end{aligned}$$

Summation für $j = 0, \ldots, k$ ergibt

$$\|W(B_{k+1} - \nabla^2 f(x^*)^{-1}) W^T\|_F - \|W(B_0 - \nabla^2 f(x^*)^{-1}) W^T\|_F$$

$$\leq (2\alpha_1\delta + \alpha_2) \cdot \frac{2}{1-\rho} \cdot \|x^0 - x^*\|$$
$$\leq (2\alpha_1\delta + \alpha_2) \cdot \frac{2}{1-\rho} \cdot \varepsilon,$$

und damit

$$\|W(B_{k+1} - \nabla^2 f(x^*)^{-1})W^T\|_F \leq \delta + (2\alpha_1\delta + \alpha_2)\frac{2}{1-\rho}\varepsilon \leq 2\delta$$

□

Der Nachweis der überlinearen Konvergenz geht wieder über den Satz von Dennis-Moré (Satz 3.3.3), d. h. wir weisen

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

nach.

4.5.6 Satz
Es gelten die Voraussetzungen von Satz 4.5.5. Weiter seien $\varepsilon, \bar{\delta}, x^0$ und B_0 wie dort. Dann gilt für die BFGS-Iterierten x^k sogar

(i) $\| (B_k - \nabla^2 f(x^*)^{-1}) (\nabla f(x^{k+1})^T - \nabla f(x^k)^T)\| = o(\|\nabla f(x^{k+1}) - \nabla f(x^k)^T\|)$

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

Beweis: Wie im Beweis zu Satz 4.5.5 sei $\beta \in [0, \frac{1}{3}]$ fest, $\nabla^2 f(x^*) = W^T W$ und $\rho \in (0,1)$, so dass $\|x^{k+1} - x^*\| \leq \rho \cdot \|x^k - x^*\|$ für $\|x^0 - x^*\| \leq \varepsilon$, $\|B_0 - \nabla^2 f(x^*)^{-1}\|_F \leq \delta$. Nach Satz 4.5.5 sind die Größen

$$\sigma_k = \|W(B_k - \nabla^2 f(x^*)^{-1})W\|_F$$

beschränkt. Außerdem ist wegen Lemma 4.5.4 im Falle $\sigma_k \neq 0$

$$\sigma_{k+1} \leq \left(1 - \frac{\alpha}{2}\Theta_k^2\right)\sigma_k + (\alpha_1\sigma_k + \alpha_2) \cdot (\|x^{k+1} - x^*\| + \|x^k - x^*\|)$$

mit $\alpha, \alpha_1, \alpha_2 > 0$ und

$$\Theta_k = \frac{\|W(B_k - \nabla^2 f(x^*)^{-1})y^k\|}{\sigma_k \cdot \|W^{-T}y^k\|}, \quad y^k = \nabla f(x^{k+1})^T - \nabla f(x^k)^T.$$

Hieraus folgt

$$\frac{\alpha}{2}\Theta_k^2\sigma_k \leq \sigma_k - \sigma_{k+1} + (\alpha_1\sigma_k + \alpha_2) \cdot (\rho^{k+1} + \rho^k) \cdot \|x^0 - x^*\|$$

Im Falle $\sigma_k = 0$ ist dies wegen $\Theta_k = 0$ ebenfalls richtig. Durch Summation ergibt sich mit $\sigma = \sup_k \sigma_k < \infty$

$$\frac{\alpha}{2} \sum_{k=0}^{\infty} \Theta_k^2 \sigma_k \leq \sigma + 2(\alpha_1 \sigma + \alpha_2) \frac{1}{1-\rho} \|x^0 - x^*\|.$$

Also gilt $\lim_{k\to\infty} \Theta_k^2 \sigma_k = 0$ und damit auch $\lim_{k\to\infty} \Theta_k^2 \sigma_k^2 = 0$, also

$$\lim_{k\to\infty} \Theta_k \sigma_k = 0$$

Dies ist aber gerade (i).
Zum Nachweis von (ii) verwenden wir die aus Lemma 3.2.6 resultierende Darstellung

$$\nabla f(x^{k+1})^T - \nabla f(x^k)^T = \nabla^2 f(x^*)(x^{k+1} - x^k) + z^k \tag{4.5.9}$$

mit $\|z^k\| \leq \frac{\gamma}{2} \cdot \|x^{k+1} - x^k\| \cdot (\|x^{k+1} - x^*\| + \|x^k - x^*\|) = o(\|x^{k+1} - x^k\|)$.
Damit erhalten wir

$$\begin{aligned}
&(B_k - \nabla^2 f(x^*)^{-1})(\nabla f(x^{k+1})^T - \nabla f(x^k)^T) \\
&= (B_k - \nabla^2 f(x^*)^{-1}) \left(\nabla^2 f(x^*)(x^{k+1} - x^k) + z^k\right) \\
&= B_k \nabla^2 f(x^*)(x^{k+1} - x^k) - (x^{k+1} - x^k) + (B_k - \nabla^2 f(x^*)^{-1}) \cdot z^k
\end{aligned}$$

woraus wegen $x^{k+1} - x^k = -B_k \nabla f(x^k)^T$

$$\begin{aligned}
&(B_k - \nabla^2 f(x^*)^{-1})(\nabla f(x^{k+1})^T - \nabla f(x^k)^T) \\
&= B_k(\nabla^2 f(x^*)(x^{k+1} - x^k) + \nabla f(x^k)^T) \\
&\quad + (B_k - \nabla^2 f(x^*)^{-1}) z^k
\end{aligned} \tag{4.5.10}$$

folgt.
Nach (i) und (4.5.9) erhalten wir so

$$\begin{aligned}
&\|B_k(\nabla^2 f(x^*)(x^{k+1} - x^k) + \nabla f(x^k)) + (B_k - \nabla^2 f(x^*)^{-1}) z^k\| \\
&= o(\|\nabla^2 f(x^*) \cdot (x^{k+1} - x^k) + z^k\|),
\end{aligned}$$

worin

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

und

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

gilt. Dies bedeutet

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

und wegen

$$\|x^{k+1}-x^k\| \le \|\nabla^2 f(x^*)^{-1}\| \cdot \|\nabla^2 f(x^*)(x^{k+1}-x^k)\|$$

sowie

$$\|B_k(\nabla^2 f(x^*)(x^{k+1}-x^k)+\nabla f(x^k))\| \le \|B_k\| \cdot \|\nabla^2 f(x^*)(x^{k+1}-x^k)+\nabla f(x^k)\|$$

mit $\|B_k\| \le$ const nach Satz 4.5.5 (iii) auch

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

□

4.5.7 Bemerkung
Dieser Satz zur überlinearen Konvergenz unterscheidet sich von dem entsprechenden Resultat für PSB, Satz 4.4.8. Es reicht diesmal nicht, die genügend schnelle Konvergenz im Sinne von

$$\sum_{k=0}^{\infty} \|x^k - x^*\| \le \infty \tag{4.5.11}$$

zu fordern; jedenfalls kommt man damit im Beweis zu Teil (i) nicht durch. Der entscheidende Punkt ist, dass wir die Beschränktheit der σ_k benötigen und diese diesmal *nicht* aus (4.5.11) gefolgert werden kann.

4.5.8 Bemerkung
Wie für viele andere Quasi-Newton-Verfahren kann man auch für BFGS zeigen, dass es im Falle einer quadratischen Funktion $f(x) = \frac{1}{2}x^T Q x + c^T x + \gamma$ und der Wahl einer exakten Schrittweite ein endliches Verfahren ist, welches nach spätestens n Schritten mit der exakten Lösung $x^* = Q^{-1}c$ abbricht und zudem $B_n = Q^{-1}$ bestimmt. Details s. Übung.

4.5.9 Bemerkung
In der Praxis beobachtet man auch bei allgemeinerem f teilweise $\lim_{k\to\infty} B_k = \nabla^2 f(x^*)^{-1}$, aber keinesfalls immer.

Abschnitt 4.6

Globalisierung des BFGS-Verfahrens

In diesem Abschnitt beschreiben wir eine geeignete Globalisierung des BFGS-Verfahrens. Wir werden dann zeigen, dass für streng konvexe Funktionen immer Konvergenz vorliegt. Interessanterweise sind analoge Resultate für PSB oder DFP nicht bekannt.
Die Globalisierung bezieht sich diesmal nur auf die Berechnung einer Schrittweite, für die wir die Wolfe-Powell-Regel verwenden.

4.6.1 Algorithmus (globalisiertes BFGS-Verfahren)
wähle $\sigma \in (0, 1/2),\ \rho \in (\sigma, 1)$
wähle x^0
wähle B_0 spd, z.B. $B_0 = \nabla^2 f(x^0)^{-1}$
for $k = 0, 1, \ldots$ **do**
berechne $d^k = -B_k \nabla f(x^k)^T$
bestimme $t^k > 0$ so, dass {Wolfe-Powell-Schrittweite}
$f(x^k + t^k d^k) \leq f(x^k) + \sigma t^k \nabla f(x^k) d^k$ und
$\nabla f(x^k + t^k d^k) d^k \geq \rho \nabla f(x^k) d^k$ {s. Alg. 2.3.5}
setze $s^k = t^k d^k$, $x^{k+1} = x^k + s^k$, $y^k = \nabla f(x^{k+1})^T - \nabla f(x^k)^T$
setze $B_{k+1} = B_k + \frac{1}{(y^k)^T s^k} \cdot ((s^k - B_k y^k)(s^k)^T + s^k (s^k - B_k y^k)^T) - \frac{(s^k - B_k y^k)^T y^k}{((y^k)^T s^k)^2} s^k (s^k)^T$
{BFGS-Update}
end for

Für die nun folgende Analyse ist es einfacher, mit den Inversen $H_k = B_k^{-1}$ zu arbeiten. Bereits in Bemerkung 4.2.17 hatten wir dazu festgehalten:

$$H_+ = H + \frac{1}{y^T s} y y^T - \frac{1}{s^T H s} (Hs)(Hs)^T. \tag{4.6.1}$$

Wir beginnen mit einer Diskussion der Wohldefiniertheit des globalisierten BFGS-Verfahrens. In Analogie zu Bemerkung 4.2.10 für DFP haben wir

4.6.2 Bemerkung
Im Falle $y^T s > 0$ ist im globalisierten BFGS-Verfahren mit B auch B_+ spd, denn für alle $x \neq 0$ ist

$$x^T B_+ x = \frac{1}{(y^T s)^2} \cdot \Big((x^T s)^2 (y^T s) + \left[(y^T s)x - (s^T x)y \right]^T B \left[(y^T s)x - (s^T x)y \right] \Big) .$$

Natürlich ist dann auch die Inverse H_+ spd.
Das folgende Lemma beruht nun entscheidend darauf, dass wir die Wolfe-Powell-Schrittweite verwenden.

4.6.3 Lemma
Im globalisierten BFGS-Verfahren (Algorithmus 4.6.1) sei H_k spd. Dann ist $(y^k)^T s^k > 0$, solange $\nabla f(x^k) \neq 0$.

Beweis: Wir lassen den Index k weg. Wir haben

$$\begin{aligned} s^T y &= (x^+ - x)^T \left(\nabla f(x^+)^T - \nabla f(x)^T \right) \\ &= t \cdot d^T \left(\nabla f(x^+)^T - \nabla f(x)^T \right) \\ &\geq t \cdot (\rho - 1) \cdot d^T \nabla f(x)^T \\ &= t \cdot (1 - \rho) \cdot \nabla f(x) B \nabla f(x)^T \\ &> 0. \end{aligned}$$

□

Zusammenfassend haben wir damit das folgende Resultat.

4.6.4 Satz
Sei $f : \mathbb{R}^n \to \mathbb{R}$, $f \in \mathcal{C}^1(\mathbb{R}^n)$ nach unten beschränkt. Dann gilt für das globalisiserte BFGS-Verfahren (Algorithmus 4.6.1)

(i) $(s^k)^T y^k > 0$ für alle k.

(ii) B_k ist spd für alle k.

(iii) Das Verfahren ist wohldefiniert, d.h. alle Iterierten existieren solange $\nabla f(x^k) \neq 0$, und es ist $f(x^{k+1}) \leq f(x^k)$ für alle k.

Beweis: (i) und (ii) wurden soeben gezeigt. Für (iii) bleibt nur nachzuweisen, dass eine Wolfe-Powell-Schrittweite immer existiert. Dies ist aber wegen $\nabla f(x^k) d^k < 0$ und f nach unten beschränkt nach Satz 2.3.2 tatsächlich der Fall. □

Das globalisierte BFGS-Verfahren berechnet also immer Abstiegsrichtungen d^k. Trotzdem ist die globale Konvergenz (im selben Sinne wie bei den globalisierten Newton-Verfahren aus Abschnitt 3.5) noch nicht garantiert. Um dies

zu erreichen, muss man garantieren, dass z.B. die Winkelbedingung (2.1.2) oder die Zoutendijk-Bedingung (2.1.4) gilt. Man kann dies wieder durch einen Rückzug auf Gradientenschritte erreichen. Es stellt sich dann allerdings die Frage, wie B in einem solchen Falle aufdatiert werden soll. Eine Diskussion findet man im Buch von Geiger und Kanzow.
Für streng konvexe Funktionen ist die in Algoritgmus 4.6.1 verwendete Globalisierung allerdings ausreichend. Dies wollen wir im Folgenden noch nachweisen. Zur Vorbereitung brauchen wir zwei Hilfssätze zu Determinanten und ein weiteres technisches Resultat.

4.6.5 Lemma
Seien $u, v \in \mathbb{R}^n$. Dann ist

$$\det(I + uv^T) = 1 + u^T v.$$

Beweis: s. Übung. □

Das nächste Lemma bezieht sich auf die Aufdatierung der inversen BFGS-Matrizen H, s. (4.6.1).

4.6.6 Lemma
Es sei $H \in \mathbb{R}^{n\times n}$ spd und es seien $s, y \in \mathbb{R}^n$ mit $s^T y > 0$. Dann gilt

$$\det H_+ = \frac{y^T s}{s^T H s} \cdot \det H.$$

Beweis: Wir machen das in zwei Etappen und setzen dazu

$$A = H + \frac{1}{y^T s} y y^T \quad \Rightarrow H_+ = A - \frac{1}{s^T H s}(Hs)(Hs)^T.$$

Wir haben

$$A = H \cdot \left(I + \frac{1}{y^T s} H^{-1} y y^T \right), \; H_+ = A \cdot \left(I - \frac{1}{s^T H s}(A^{-1} H s)(Hs)^T \right),$$

woraus mit Lemma 4.6.5 folgt

$$\det A = \left(1 + \frac{y^T H^{-1} y}{y^T s} \right) \cdot \det H, \; \det H_+ = \left(1 - \frac{s^T H A^{-1} H s}{s^T H s} \right) \cdot \det A.$$

Man beachte, dass hieraus insbesondere $\det A \neq 0$ folgt, d.h. A^{-1} existiert tatsächlich. Mit der Sherman-Morrison-Woodbury-Formel (Lemma 4.2.12) erhalten wir

$$A^{-1} = H^{-1} - \frac{1}{y^T s + y^T H^{-1} y}(H^{-1} y)(H^{-1} y)^T.$$

Insgesamt ergibt sich so

$$\begin{aligned}\det H_+ &= \left(1 - \frac{s^T H A^{-1} H s}{s^T H s}\right) \cdot \left(1 + \frac{y^T H^{-1} y}{y^T s}\right) \det H \\ &= \left(1 - \frac{s^T H s}{s^T H s} + \frac{1}{y^T s + y^T H^{-1} y} \cdot \frac{s^T H (H^{-1} y)(H^{-1} y)^T H s}{s^T H s}\right) \\ &\quad \cdot \left(1 + \frac{y^T H^{-1} y}{y^T s}\right) \det H \\ &= \frac{y^T s}{s^T H s} \cdot \det H.\end{aligned}$$

□

Ein letzter Hilfssatz wird uns für den Nachweis der Zoutendijk-Bedingung nützlich sein.

4.6.7 Lemma
Seien $\beta_j \geq 1, i = 0, 1, \ldots$ Zahlen so, dass für ein $b > 1$ gilt

$$\prod_{j=0}^{k} \beta_j \leq b^{k+1}, \; k = 0, 1, \ldots.$$

Dann ist

$$\sum_{j=0}^{\infty} \frac{1}{\beta_j} = \infty.$$

Beweis: Es sei k ungerade. Dann gilt für mindestens $(k+1)/2$ Indizes $j \in \{0, \ldots, k\}$ die Beziehung $\beta_j \leq b^2$, denn andernfalls wäre $\prod_{j=0}^{k} \beta_j > (b^2)^{(k+1)/2} = b^{k+1}$. Damit gilt aber auch

$$\sum_{j=0}^{k} \frac{1}{\beta_j} \geq \frac{k+1}{2} \cdot \frac{1}{b^2} \to \infty \; (k \to \infty).$$

□

4.6.8 Satz
Sei $f : \mathbb{R}^n \to \mathbb{R}$, $f \in \mathcal{C}^1(\mathbb{R}^n)$, sei $x^0 \in \mathbb{R}^n$, die Levelmenge $\mathcal{L}(x^0) = \{x \in \mathbb{R}^n : f(x) \leq f(x^0)\}$ konvex und f glm. konvex auf $\mathcal{L}(x^0)$. Dann konvergieren die mit dem globalisierten BFGS-Verfahren (Algorithmus 4.6.1) erzeugten Iterierten x^k für jede Wahl von B_0 spd gegen die eindeutige Minimalstelle x^* von f.

Beweis: Wir können davon ausgehen, dass der Algorithmus nicht nach endlich vielen Schritten mit $\nabla f(x^k) = 0$ abbricht, denn ansonsten ist $x^* = x^k$ und die Aussage des Satzes trivial.
Wir halten zunächst fest, dass nach Satz 4.6.4 alle x^k in $\mathcal{L}(x^0)$ liegen. Weil f glm. konvex ist, ist ∇f glm. monoton auf $\mathcal{L}(x^0)$. Also existiert $\mu > 0$ mit

$$(\nabla f(x) - \nabla f(y))\,(x - y) \geq \mu \cdot \|x - y\|^2 \text{ für } x, y \in \mathcal{L}(x^0).$$

Nach Satz 1.3.10 ist $\mathcal{L}(x^0)$ kompakt. Deshalb ist ∇f sogar Lipschitz-stetig auf $\mathcal{L}(x^0)$, denn aus dem Mittelwertsatz (Lemma 1.2.5 (ii)) folgt

$$\|\nabla f(x) - \nabla f(y)\| \leq \gamma \|x - y\| \text{ mit } \gamma = \max_{z \in \mathcal{L}(x^0)} \|\nabla^2 f(z)\|, \quad x, y \in \mathcal{L}(x^0).$$

Wegen Satz 2.1.11 sind wir fertig, sobald wir die Zoutendijk-Bedingung

$$\sum_{j=0}^{\infty} \delta_j = +\infty \quad \text{mit } \delta_j = \left(\frac{\nabla f(x^j) d^j}{\|\nabla f(x^j)\| \cdot \|d^j\|} \right)^2 \tag{4.6.2}$$

nachgeweisen haben. Dies wird uns über Lemma 4.6.7 gelingen, d.h. wir weisen

$$\prod_{j=0}^{k} \frac{1}{\delta_j} \leq b^{k+1}$$

mit einer geeigneten Konstante b nach. Wegen $s^j = t^j d^j$, $t_j \nabla f(x^j)^T = -H_j s^j$ haben wir

$$\begin{aligned} \prod_{j=0}^{k} \frac{1}{\delta_j} &= \prod_{j=0}^{k} \left(\frac{\|\nabla f(x^j)\| \cdot \|d^j\|}{\nabla f(x^j) d^j} \right)^2 \\ &= \prod_{j=0}^{k} \frac{\|H_j s^j\|^2 \cdot \|s^j\|^2}{((s^j)^T H_j s^j)^2} \\ &\leq \frac{1}{\mu^{k+1}} \cdot \prod_{j=0}^{k} \frac{\|H_j s^j\|^2}{(s^j)^T H_j s^j} \cdot \prod_{j=0}^{k} \frac{(y^j)^T s^j}{(s^j)^T H_j s^j}. \end{aligned} \tag{4.6.3}$$

Hierin folgt die letzte Ungleichheit aus der glm. Monotonie von ∇f. Wir schätzen jetzt beide Produkte auf der rechten Seite getrennt ab. Wesentliches Mittel dazu ist die Ungleichung vom geometrisch-arithmetischen Mittel, wonach für nichtnegative Zahlen $\alpha_0, \ldots \alpha_k$ stets

$$\prod_{j=0}^{k} \alpha_j \leq \left(\frac{1}{k+1} \sum_{j=0}^{k} \alpha_j \right)^{k+1}$$

gilt. (Man beweist dies z.B. über die Konkavität des Logarithmus, den man auf beide Seiten anwendet.)
Wir erreichen unser Ziel jetzt durch Untersuchung von spurH_j. Es ist

$$0 < \text{spur}H_{j+1} = \text{spur}H_j - \frac{\|H_j s^j\|^2}{(s^j)^T H_j s^j} + \frac{\|y^j\|^2}{(y^j)^T s^j} \leq \text{spur}H_j + \frac{\|y^j\|^2}{(y^j)^T s^j} \tag{4.6.4}$$

und damit

$$0 < \text{spur}H_{k+1} + \sum_{j=0}^{k} \frac{\|H_j s^j\|^2}{(s^j)^T H_j s^j} = \text{spur}H_0 + \sum_{j=0}^{k} \frac{\|y^j\|^2}{(y^j)^T s^j}.$$

Wegen $s^j = x^{j+1} - x^j$, $y^j = \nabla f(x^{j+1})^T - \nabla f(x^j)^T$ können wir die letzte Summe mit der glm. Monotonie und der Lipschitz-Stetigkeit weiter nach oben abschätzen und erhalten

$$0 < \text{spur}H_{k+1} + \sum_{j=0}^{k} \frac{\|H_j s^j\|^2}{(s^j)^T H_j s^j} \leq \text{spur}H_0 + (k+1) \cdot \frac{\gamma^2}{\mu} = (k+1)b_1$$

mit geeigneter Wahl von b_1. Hieraus erhalten wir

$$0 < \text{spur}H_{k+1} \leq (k+1)b_1, \tag{4.6.5}$$

$$0 < \sum_{j=0}^{k} \frac{\|H_j s^j\|^2}{(s^j)^T H_j s^j} \leq (k+1)b_1. \tag{4.6.6}$$

Mit der Ungleichung vom geometrisch-arithmetischen Mittel erhalten wir so für das erste Produkt in (4.6.3) aus (4.6.6) die Abschätzung

$$\prod_{j=0}^{k} \frac{\|H_j s^j\|^2}{(s^j)^T H_j s^j} \leq (b_1)^{k+1} \tag{4.6.7}$$

Auf Grund von Lemma 4.6.6 haben wir auch

$$\det H_{k+1} = \prod_{j=0}^{k} \frac{(y^j)^T s^j}{(s^j)^T H_j s^j} \cdot \det H_0.$$

Bezeichnen wir mit $\lambda_1, \ldots \lambda_n$ die (positiven) Eigenwerte von H_{k+1}, so gilt

$$\det H_{k+1} = \prod_{i=0}^{n} \lambda_i, \quad \text{spur}H_{k+1} = \sum_{i=0}^{n} \lambda_i.$$

Die Ungleichung vom geometrischen-arithmetischen Mittel liefert deshalb

$$\det H_{k+1} \leq \left(\frac{1}{n} \cdot \text{spur} H_{k+1} \right)^n ,$$

und damit wegen (4.6.5)

$$\prod_{j=0}^{k} \frac{(y^j)^T s^j}{(s^j)^T H_j s^j} \cdot \det H_0 \leq \left(\frac{1}{n} \cdot \text{spur} H_{k+1} \right)^n \leq \left(\frac{k+1}{n} \cdot b_1 \right)^n . \tag{4.6.8}$$

Mit (4.6.7) und (4.6.8) können wir nun beide Produkte in (4.6.3) abschätzen und erhalten

$$\prod_{j=0}^{k} \frac{1}{\delta_j} \leq \frac{1}{\mu^{k+1}} \cdot \frac{1}{\det H_0} \cdot (b_1)^{k+1} \cdot \left(\frac{k+1}{b} \cdot b_1 \right)^n \leq b^{k+1}$$

mit geeignet gewähltem $b > 1$. □

Kapitel 5

CG-Verfahren

Das CG-Verfahren wurde ursprünglich für lineare Gleichungssysteme entwickelt. Es gibt inzwischen aber diverse Übertragungen auf Minimierungsaufgaben mit nicht-quadratischer Zielfunktion, von denen wir hier einige besprechen.

Abschnitt 5.1

Wiederholung: CG für lineare Gleichungssysteme

Dieser Abschnitt ist kurz gehalten, da sein Inhalt aus einführenden Veranstaltungen bekannt sein sollte.
Sei $A \in \mathbb{R}^{n\times n}$ spd, $b \in \mathbb{R}^n$ und

$$f : \mathbb{R}^n \to \mathbb{R},\ f(x) = \frac{1}{2}x^T Ax - b^T x. \tag{5.1.1}$$

Es ist bekannt:

(i) $\nabla f(x) = (Ax - b)^T$

(ii) f ist glm. konvex

(iii) $x^* = A^{-1}b$ ist die (eindeutige) globale Minimalstelle von f.

Idee des CG-Verfahrens: Finde, ausgehend von $x^0 \in \mathbb{R}^n$, Richtungen $d^0, d^1, \ldots$ so, dass mit

$$x^{k+1} = x^k + t_k d^k, \; t_k \text{ minimiert } f(x^k + td^k) \text{ für } t \in \mathbb{R}$$

sogar gilt

$$f(x^{k+1}) = \min_{v \in \text{span}\{d^0,\ldots,d^k\}} f(x^0 + v). \tag{5.1.2}$$

Wir halten dazu fest:

5.1.1 Lemma
Für $x, d \in \mathbb{R}^n$, $d \neq 0$ löst t^k mit

$$t^k = \frac{-1}{d^T A d} \cdot \nabla f(x) d = \frac{1}{d^T A d}(b - Ax)^T d$$

die Minimierungsaufgabe

$$\text{minimiere } f(x + td) \text{ für } t \in \mathbb{R}.$$

Beweis: Es ist

$$\begin{aligned} f(x+td) &= \frac{1}{2}x^T Ax + tx^T Ad + \frac{1}{2}t^2 d^T Ad - b^T x - tb^T d \\ &= \frac{1}{2}d^T Ad \left(t - \frac{b^T d - x^T Ad}{d^T Ad} \right)^2 + x^T Ax - \frac{(b^T d - x^T Ad)^2}{d^T Ad} \end{aligned}$$

□

5.1.2 Lemma
Die Eigenschaft (5.1.2) wird erreicht, wenn die Richtungen d^k alle A-konjugiert sind, d. h. es gilt für $k = 0, 1, \ldots$

$$(d^k)^T A d^j = 0, \; j = 0, 1, \ldots, k-1.$$

Beweis: bekannt □

Das CG-Verfahren setzt diese Idee nun algorithmisch effizient um.

5.1.3 Algorithmus (CG-Verfahren)
wähle $x^0 \in \mathbb{R}^n$, setze $r^0 = b - Ax^0, d^0 = r^0$
for $k = 0, 1, \ldots$ **do**
$\alpha_k = \frac{(r^k)^T d^k}{(d^k)^T A d^k}$
$x^{k+1} = x^k + \alpha_k d^k$
$r^{k+1} = r^k - \alpha_k A d^k$
$\beta_k = \frac{\|r^{k+1}\|^2}{\|r^k\|^2}$
$d^{k+1} = r^{k+1} + \beta_k d^k$
end for

5.1.4 Satz
Im CG-Verfahren ist $\|r^k\| \neq 0$ für $k = 0, \ldots, m-1$ und $r^m = 0$ für ein $m \leq n$. Es gilt

(i) $(d^k)^T A d^j = 0$ für $j = 0, \ldots, k-1$, $k = 0, \ldots, m$, wobei alle $d^k \neq 0$.

(ii) $r^k = b - Ax^k = -\nabla f(x^k)^T$, $k = 0, \ldots, m$

(iii) x^k erfüllt (5.1.2) für $k = 0, \ldots, m$

(iv) $x^m = x^* = A^{-1}b$

(v) $r^k, d^k \in \text{span}\{r^0, Ar^0, \ldots, A^k r^0\}$, $k = 0, \ldots, m$

(vi) $(r^k)^T r^j = 0, \quad j = 0, \ldots, k-1$, $k = 0, \ldots, m$

Beweis: Im Prinzip folgt alles per vollständiger Induktion. Sobald $r^k = 0$ gilt, ist $x^k = x^*$. □

Man beachte, dass α_k aus dem CG-Verfahren nach Lemma 5.1.1 die Minimierungsaufgabe

$$\text{minimiere } f(x^k + td^k) \text{ für } t \in \mathbb{R}$$

löst. Man kann zeigen (Übungsaufgabe), dass sich α_k auch darstellen lässt als

$$\alpha_k = \frac{\|r^k\|^2}{(d^k)^T A d^k},$$

was im CG-Verfahren die Berechnung eines Innenproduktes spart.
Nach Satz 5.1.4 kann man das CG-Verfahren als ein direktes Verfahren auffassen, welches nach spätestens n Schritten die Lösung berechnet. Praktisch viel wichtiger ist es, CG als Iterationsverfahren zu betrachten und die Qualität der Iterierten x^k geeignet zu messen.

5.1.5 Definition
Für $r^0 \in \mathbb{R}^n$ bezeichnet

$$K_k(A, r^0) = \text{span}\{r^0, Ar^0, \ldots, A^{k-1}r^0\}$$

den Krylov-Unterraum der Stufe k bzgl. A und r^0.

Die Aussage (v) von Satz 5.1.4 liefert in Verbindung mit der Eigenschaft (5.1.2) deshalb folgende Charakterisierung der CG-Iterierten.

5.1.6 Lemma
Für die CG-Iterierten x^k gilt

$$f(x^k) = \min_{v \in K_k(A, x^0)} f(x^0 + v).$$

Verwenden wir die Tatsache, dass $f(x^*) = \frac{1}{2}b^T x^*$ (wegen $Ax^* = b$), so erhalten wir nach kurzer Rechnung

$$f(x) - f(x^*) = \frac{1}{2}(x - x^*)^T A(x - x^*).$$

Weil A spd ist, ist $\langle \cdot, \cdot \rangle_A$ mit $\langle x, y \rangle_A = x^T Ay$ ein Innenprodukt auf $\mathbb{R}^n$. Mit $\| \cdot \|_A$ bezeichnen wir die zugehörige Norm, d.h. $\|x\|_A^2 = \langle x, x \rangle_A$. Ist $A = A^{1/2} \cdot A^{1/2}$ so haben wir

$$\|x\|_A = \|A^{1/2}x\|.$$

Weil $f(x^*)$ konstant ist, gilt nach Lemma 5.1.6 für die CG-Iterierten x^k also

$$\|x^k - x^*\|_A = \min_{x \in x^0 + K_k(A, r^0)} \|x - x^*\|_A. \tag{5.1.3}$$

Jedes $x \in x^0 + K_k(A, r^0)$ hat die Gestalt

$$x = x^0 + \sum_{j=0}^{k-1} \alpha_j A^j r^0,$$

woraus

$$x^* - x = p_k(A) \cdot (x^* - x)$$

folgt mit dem Polynom $p_k(t) = 1 - \sum_{j=1}^{k} \alpha_{j-1} t^j$. Wir notieren

$$\overline{\Pi}_k = \{p : p \text{ ist Polynom mit Grad } \leq k \text{ und } p(0) = 1\}.$$

Wir erhalten so für $x \in K_k(A, r^0)$

$$\|x^* - x\|_A = \|A^{1/2} p_k(A)(x^0 - x^*)\| = \|p_k(A) A^{1/2}(x^* - x^0)\|.$$

Die Charakterisierung (5.1.3) der CG-Iterierten x^k können wir dehalb auch ausdrücken als

$$\|x^k - x^*\|_A = \min_{p \in \overline{\Pi}_k} \|p(A) A^{1/2}(x - x^*)\|$$

woraus folgt

$$\|x^k - x^*\|_A \leq \left(\min_{p \in \overline{\Pi}_k} \|p(A)\| \right) \cdot \|(x - x^*)\|_A. \tag{5.1.4}$$

Abschätzungen für $\min_{p \in \overline{\Pi}_k} \|p(A)\|$ ergeben also Abschätzungen für den Fehler $x^k - x^*$ der CG-Iterierten.

5.1.7 Satz
Es seien $\lambda_1 < \lambda_2 \ldots < \lambda_m$ die *verschiedenen* Eigenwerte von A und $0 < a \leq \lambda_1,\ \lambda_m \leq b$. Dann gilt

(i) Es ist $x^k = x^*$ für einen Index $k \leq m$.

(ii) Solange $x^k \neq x^*$ gilt

$$\|x^k - x^*\|_A \leq 2 \cdot c^k \cdot \|x^0 - x^*\|_A$$

mit $c = \frac{\sqrt{\kappa}-1}{\sqrt{\kappa}+1},\ \kappa = \frac{b}{a}$.

Beweis: zu (i): Wir nehmen für p_m das skalierte Minimalpolynom der symmetrischen Matrix A

$$p_m(t) = \prod_{j=1}^{m} \left(1 - \frac{t}{\lambda_j}\right) \in \overline{\Pi}_m.$$

Es ist $p_m(A) = 0$, also

$$\min_{p \in \overline{\Pi}_k} \|p(A)\| = 0,$$

woraus nach (5.1.4) dann $x^m = x^*$ folgt.
zu (ii): S. Übung: Man kann explizit Polynome q_k konstruieren mit

$$\|q_k(A)\| = \frac{2}{c^k + c^{-k}} \leq 2c^k.$$

Dabei verwendet man, dass für jedes Polynom q gilt

$$\|q(A)\| = \max_{\lambda \in \text{spek}(A)} |q(\lambda)| \leq \max_{t \in [a,b] \supseteq \text{spek}(A)} |q(t)|.$$

□

Abschnitt 5.2

Das Fletcher-Reeves-Verfahren

Das Fletcher-Reeves-Verfahren ist eine erste mögliche Übertragung des CG-Verfahrens (Algorithmus 5.1.3) auf nicht notwendig quadratische Zielfunktionen.

Sei $f : \mathbb{R}^n \to \mathbb{R}$, $f \in \mathcal{C}^1(\mathbb{R}^n)$. In Algorithmus 5.1.3 ist $r^k = b - Ax^k = -\nabla f(x^k)^T$ für die quadratische Zielfunktion $f(x) = \frac{1}{2}x^T Ax - x^T b$. Für eine allgemeinere Zielfunktion ersetzen wir also in Algorithmus 5.1.3 jedes Vorkommen von r^k durch $-\nabla f(x^k)^T$.
Außerdem ist bei einer nicht-quadratischen Zielfunktion die Minimierungsaufgabe

$$\text{minimiere } f(x + td) \text{ für } t \in \mathbb{R}$$

nicht mehr explizit lösbar, d.h. die Bestimmung von α_k in Algorithmus 5.1.3 muss durch eine geeignete Schrittweitenstrategie ersetzt werden. Die Theorie wird zeigen, dass eine strenge Wolfe-Powell-Schreite diesmal das Richtige ist. Wir schreiben ab jetzt wieder t_k statt α_k.

5.2.1 Algorithmus (Fletcher-Reeves-Verfahren)
wähle $x^0 \in \mathbb{R}^n$, wähle $0 < \sigma < \rho < 1/2$, setze $d^0 = -\nabla f(x^0)^T$
for $k = 0, 1, \ldots$ **do**
bestimme Schrittweite $t_k > 0$ mit
$f(x^k + t_k d^k) \leq f(x^k) + \sigma t_k \cdot \nabla f(x^k) d^k$
$|\nabla f(x^k + t_k d^k) d^k| \leq -\rho \cdot \nabla f(x^k) d^k$
{strenge Wolfe-Powell-Schrittweite}
$x^{k+1} = x^k + t_k d^k$
$\beta_k^{FR} = \|\nabla f(x^{k+1})\|^2 / \|\nabla f(x^k)\|^2$
$d^{k+1} = -\nabla f(x^{k+1})^T + \beta_k^{FR} d^k$
end for

5.2.2 Bemerkung
Bei der Definition der Wolfe-Powell-Schrittweiten hatten wir $0 < \sigma < \rho < 1$ zugelassen. Dass wir $\rho < 1/2$ brauchen, zeigt schon der nächste Satz.

5.2.3 Satz
Sie $f : \mathbb{R}^n \to \mathbb{R}$, $f \in \mathcal{C}^1(\mathbb{R}^n)$ und nach unten beschränkt. Dann sind für jede Wahl von x^0 alle Iterierten x^k definiert und es gilt für $k = 0, 1, \ldots$

$$\left| \frac{\nabla f(x^k) d^k}{\|\nabla f(x^k)\|^2} + 1 \right| \leq \sum_{j=0}^{k} \rho^j - 1. \tag{5.2.1}$$

Beweis: Wegen $\rho \in (0, 1/2)$ ist die rechte Seite in (5.2.1) kleiner als 1, d.h. es ist dann

$$\frac{\nabla f(x^k) d^k}{\|\nabla f(x^k)\|^2} < 0,$$

so dass d^k eine Abstiegsrichtung ist. Da für eine Abstiegsrichtung immer eine strenge Wolfe-Powell-Schrittweite existiert ist mit (5.2.1) die Definiertheit der Iterierten gezeigt.
Für $k = 0$ ist (5.2.1) richtig, da $d^0 = -\nabla f(x^0)^T$. Für den Induktionsschritt $k \to k + 1$ folgt aus der Schrittweitenwahl

$$|\nabla f(x^{k+1}) d^k| \leq -\rho \nabla f(x^k) d^k.$$

Aus der Aufdatierung für d^{k+1} folgt außerdem

$$\frac{\nabla f(x^{k+1}) d^{k+1}}{\|\nabla f(x^{k+1})\|^2} = -1 + \frac{\nabla f(x^{k+1}) d^k}{\|\nabla f(x^k)\|^2}$$

und damit

$$\begin{aligned}
\left| \frac{\nabla f(x^{k+1}) d^{k+1}}{\|\nabla f(x^{k+1})\|^2} + 1 \right| &= \left| \frac{\nabla f(x^{k+1}) d^k}{\|\nabla f(x^k)\|^2} \right| \\
&\leq \rho \left| \frac{\nabla f(x^k) d^k}{\|\nabla f(x^k)\|^2} \right| \\
&\leq \rho \left| \frac{\nabla f(x^k) d^k}{\|\nabla f(x^k)\|^2} - 1 \right| + \rho \\
&\leq \rho \left(\sum_{j=0}^{k} \rho^j - 1 \right) + \rho \\
&\leq \sum_{j=0}^{k+1} \rho^j - 1.
\end{aligned}$$

□

5.2.4 Bemerkung
Insbesondere gilt im Falle exakter Schrittweiten ($\rho = 0$) also

$$\nabla f(x^k)d^k = -\|\nabla f(x^k)\|^2.$$

Bevor wir gleich in Satz 5.2.6 ein wesentliches Konvergenzresultat für das Fletcher-Reeves-Verfahren beweisen werden, formulieren wir zuerst ein nützliches Hilfsresultat zu Situationen, in welchen der Gradient nicht unbeschränkt wachsen kann.

5.2.5 Lemma
Sei $f : \mathbb{R}^n \to \mathbb{R}, f \in \mathcal{C}^1(\mathbb{R}^n), \nabla f \in \mathrm{Lip}_\gamma(\mathbb{R}^n)$. Weiter sei $x^0 \in \mathbb{R}^n$ und f auf $\mathcal{L}(x^0)$ nach unten beschränkt. Dann ist ∇f auf $\mathcal{L}(x^0)$ beschränkt.

Beweis: Für $x \in \mathbb{R}^n$ gilt

$$\begin{aligned}
\frac{d}{dt} f(x - t\nabla f(x)^T) &= -\nabla f(x - t\nabla f(x)^T) \cdot \nabla f(x)^T \\
&= -\left(\nabla f(x - t\nabla f(x)^T) - \nabla f(x)\right) \nabla f(x)^T - \|\nabla f(x)\|^2 \\
&\leq \|\nabla f\left(x - t\nabla f(x)^T\right) - \nabla f(x)\| \cdot \|\nabla f(x)\| - \|\nabla f(x)\|^2 \\
&\leq (\gamma \cdot t - 1) \cdot \|\nabla f(x)\|^2.
\end{aligned}$$

Danach erhalten wir mit $y = x - \frac{1}{2\gamma}\nabla f(x)$

$$\begin{aligned}
f(y) - f(x) &= \int_0^{1/2\gamma} \frac{d}{dt} f\left(x - t\nabla f(x)\right) dt \\
&\leq -\frac{1}{2} \int_0^{1/2\gamma} \|\nabla f(x)\|^2 dt = -\frac{1}{4\gamma} \cdot \|\nabla f(x)\|^2.
\end{aligned}$$

Daraus folgt für $x \in \mathcal{L}(x^0)$ und mit $\alpha \leq f(y)$ für alle $y \in \mathbb{R}^n$

$$\|\nabla f(x)\|^2 \leq 4\gamma \left(f(x) - f(y)\right) \leq 4\gamma \left(f(x^0) - \alpha\right).$$

□

5.2.6 Satz
Sei $f : \mathbb{R}^n \to \mathbb{R}$, $f \in \mathcal{C}^1(\mathbb{R}^n)$ und nach unten beschränkt. Es sei $x^0 \in \mathbb{R}^n$ und es sei $\nabla f \in \mathrm{Lip}_\gamma(\mathcal{L}(x^0))$. Dann gilt für die Iterierten x^k des Fletcher-Reeves-Verfahrens (Algorithmus 5.2.1)

$$\liminf_{k\to\infty} \|\nabla f(x^k)\| = 0.$$

Beweis: Wir nehmen an, dass die Aussage nicht richtig ist und deshalb $\varepsilon > 0$ und $k_\varepsilon \in \mathbb{N}$ existieren mit

$$\|\nabla f(x^k)\| \geq \varepsilon \quad \text{für alle } k \geq k_\varepsilon.$$

Wir weisen nach, dass dann die Zoutendijk-Bedingung

$$\sum_{k=0}^{\infty} \delta_k = \infty \quad \text{mit } \delta_k = \left(\frac{\nabla f(x^k) d^k}{\|\nabla f(x^k)\| \cdot \|d^k\|} \right)^2$$

erfüllt ist, so dass aus Satz 2.1.9 dann doch wieder

$$\liminf_{k \to \infty} \|\nabla f(x^k)\| = 0$$

folgt, ein Widerspruch.
Nach Satz 5.2.3 wissen wir bereits

$$\left| \frac{\nabla f(x^k) d^k}{\|\nabla f(x^k)\|^2} \right| \geq 1 - \left(\sum_{j=0}^{k} \rho^j - 1 \right) \geq 2 - \frac{1}{1-\rho} = \frac{1-2\rho}{1-\rho} > 0 \tag{5.2.2}$$

sowie

$$\left| \frac{\nabla f(x^k) d^k}{\|\nabla f(x^k)\|^2} \right| \leq \sum_{j=0}^{k} \rho^j \leq \frac{1}{1-\rho}. \tag{5.2.3}$$

Für die Größen $\gamma_k = \|d^k\|^2 / \|\nabla f(x^k)\|^4$ erhalten wir wegen der strengen Wolfe-Powell-Schrittweiten

$$\begin{aligned}
\gamma_{k+1} = \frac{(d^{k+1})^T d^{k+1}}{\|\nabla f(x^{k+1})\|^4} &= \frac{(-\nabla f(x^{k+1})^T + \beta_k^{FR} d^k)^T (-\nabla f(x^{k+1})^T + \beta_k^{FR} d^k)}{\|\nabla f(x^{k+1})\|^4} \\
&= \frac{1}{\|\nabla f(x^{k+1})\|^2} - \beta_k^{FR} \cdot \frac{2 \nabla f(x^{k+1}) d^k}{\|\nabla f(x^{k+1})\|^4} + \gamma_k \\
&\leq \frac{1}{\|\nabla f(x^{k+1})\|^2} - 2\rho \frac{\nabla f(x^k) d^k}{\|\nabla f(x^k)\|^2 \|\nabla f(x^{k+1})\|^2} + \gamma_k \\
&\leq \left(1 + \frac{2\rho}{1-\rho} \right) \cdot \frac{1}{\|\nabla f(x^{k+1})\|^2} + \gamma_k \\
&\leq \frac{1+\rho}{1-\rho} \cdot \frac{1}{\varepsilon^2} + \gamma_k, \ k \geq k_\varepsilon.
\end{aligned}$$

Durch wiederholtes Einsetzen ergibt sich daraus

$$\gamma_{k+1} \leq (k - k_\varepsilon) \cdot \frac{1+\rho}{(1-\rho)\varepsilon^2} + \gamma_{k_\varepsilon} \leq (k - k_\varepsilon) \cdot c$$

mit einer geeigneten Konstanten c, z.B.

$$c = \left(\frac{1+\rho}{(1-\rho)\varepsilon^2} \right).$$

Damit haben wir auch

$$\frac{1}{\gamma_k} \geq \frac{1}{c} \cdot \frac{1}{k - k_\varepsilon} \quad \text{für } k > k_\varepsilon.$$

Nach Lemma 5.2.5 existiert eine Konstante D mit $\|\nabla f(x)\| \leq D$ für alle $x \in \mathcal{L}(x^0)$. Zusammen mit (5.2.2) erhalten wir deshalb

$$\begin{aligned} \delta_k &= \left(\frac{\nabla f(x^k) d^k}{\|\nabla f(x^k)\| \cdot \|d^k\|} \right)^2 \\ &= \left| \frac{\nabla f(x^k) d^k}{\|\nabla f(x^k)\|^2} \right|^2 \cdot \frac{1}{\|\nabla f(x^k)\|^2} \cdot \frac{1}{\gamma_k} \\ &\geq \frac{1}{cD^2} \cdot \frac{1}{k - k_\varepsilon} \text{ für } k > k_\varepsilon. \end{aligned}$$

Weil die harmonische Reihe divergiert folgt daraus die Zoutendijk-Bedingung. □

5.2.7 Satz
Sei $f : \mathbb{R}^n \to \mathbb{R}$, $f \in \mathcal{C}^2(\mathbb{R}^n)$. Sei $x^0 \in \mathbb{R}^n$, die Levelmenge $\mathcal{L}(x^0)$ sei konvex und f sei glm. konvex auf $\mathcal{L}(x^0)$. Dann konvergiert die durch das Fletcher-Reeves-Verfahren erzeugte Folge x^k gegen die eindeutige globale Minimalstelle von f.

Beweis: Nach Satz 1.3.10 ist die Levelmenge $\mathcal{L}(x^0)$ kompakt. Damit ist ∇f sogar Lipschitz-stetig auf $\mathcal{L}(x^0)$, denn $\nabla^2 f$ ist auf $\mathcal{L}(x^0)$ beschränkt. Nach Satz 5.2.6 existiert eine Teilfolge der x^{k_i} mit $\lim_{i\to\infty} \nabla f(x^{k_i}) = 0$. Diese Teilfolge besitzt mindestens einen Häufungspunkt x^*, gegen welchen eine Teilfolge konvergiert, die wir der Einfachheit halber wieder mit x^{k_i} bezeichnen. Es ist $\nabla f(x^*) = 0$, so dass x^* die eindeutige Minimalstelle von f ist. Auf Grund der gleichmäßigen Konvexität existiert $\mu > 0$ mit

$$\mu \|x^k - x^*\|^2 \leq f(x^k) - f(x^*).$$

Die $f(x^k)$ fallen monoton und konvergieren gegen $f(x^*)$ (denn $\lim_{i\to\infty} x^{k_i} = x^*$). Also gilt auch $\lim_{k\to\infty} x^k = x^*$. □

Abschnitt 5.3

Das Polak-Ribière-Verfahren

Auf Grund der Orthogonalitätsbeziehungen im CG-Verfahren für quadratische Funktionen f (Satz 5.1.4), kann man β_k in Algorithmus 5.1.3 auch ausdrücken als

$$\beta_k = \frac{\|r^{k+1}\|^2}{\|r^k\|^2} = \frac{(r^{k+1} - r^k)^T r^{k+1}}{\|r^k\|^2}.$$

Eine Übertragung auf den nicht-quadratischen Fall ergibt

$$\beta_k = \frac{\big(\nabla f(x^{k+1}) - \nabla f(x^k)\big)\, \nabla f(x^{k+1})^T}{\|\nabla f(x^k)\|^2},$$

was dann aber nicht mehr äquivalent zu der im Fletcher-Reeves-Verfahren verwendeten Größe

$$\beta_k^{FR} = \|\nabla f(x^{k+1})\|^2 / \|\nabla f(x^k)\|^2$$

ist. Polak und Ribière hatten festgestellt, dass mit dieser alternativen Übertragung (und einer weiter verschärften Schrittweiten-Strategie) in der Praxis häufig noch bessere Resultate erzielt werden als mit dem Fletcher-Reeves-Verfahren.

5.3.1 Algorithmus (Polak-Ribière-Verfahren)
wähle $x^0 \in \mathbb{R}^n$, setze $d^0 = -\nabla f(x^0)^T$
for $k = 0, 1, \ldots$ **do**
bestimme Schrittweite $t_k > 0$ mit
$t_k = \min\{t > 0 : \nabla f(x^k + t^k d^k) d^k = 0\}$ {exakte Schrittweite!}
$x^{k+1} = x^k + t_k d^k$
$\beta_k^{PR} = \big((\nabla f(x^{k+1}) - \nabla f(x^k)) \nabla f(x^{k+1})^T\big) / \|\nabla f(x^k)\|^2$
$d^{k+1} = -\nabla f(x^{k+1})^T + \beta_k^{PR} d^k$
end for

Man beachte, dass die Berechnung einer exakten Schrittweite in der Praxis gar nicht möglich ist. Man behilft sich deshalb mit einer strengen Wolfe-Powell-Schrittweite mit einem kleinen Wert für ρ (z.B. $\rho = 0.1$). Ein so

implementiertes Polak-Ribière Verfahren zeigt in der Praxis bessere Konvergenzeigenschaften als das Fletcher-Reeves-Verfahren. Die möglichen theoretischen Resultate sind dagegen schwächer als bei Fletcher-Reeves und setzen insbesondere die exakten Schrittweiten voraus.

5.3.2 Bemerkung
Da die Schrittweiten t_k exakt sind, ist für alle $k = 0, 1, \ldots$
(5.3.1)
$$\nabla f(x^{k+1})d^{k+1} = \nabla f(x^{k+1})(-\nabla f(x^{k+1})^T + \beta_{k+1}^{PR} d^k) = -\|\nabla f(x^{k+1})\|^2,$$

vgl. Bemerkung 5.2.4. Die Richtung d^k ist also eine Abstiegsrichtung, solange $\nabla f(x^k) \neq 0$, d.h. das Verfahren ist wohldefiniert.

5.3.3 Satz
Sei $f : \mathbb{R}^n \to \mathbb{R}$, $f \in \mathcal{C}^1(\mathbb{R}^n)$, $x^0 \in \mathbb{R}^n$ sowie $\nabla f \in \mathrm{Lip}_\gamma(\mathcal{L}(x^0))$. Weiter sei f auf $\mathbb{R}^n$ nach unten beschränkt. Die Folge $\{x^k\}$ entstehe aus dem Polak-Ribière-Verfahren (Algorithmus 5.3.1) und es gelte

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

Dann gilt

$$\liminf_{k\to\infty} \|\nabla f(x^k)\| = 0.$$

Beweis: Nach Aufgabe 7 (!) ist die Schrittweite t_k effizient, d.h. es existiert eine Konstante $\theta > 0$ mit

$$(5.3.2)\quad f(x^k) - f(x^{k+1}) \geq \theta \cdot \frac{1}{\gamma_k} \quad \text{mit } \gamma_k = \frac{\|d^k\|^2}{(\nabla f(x^k)d^k)^2} \overset{(5.3.1)}{=} \frac{\|d^k\|^2}{\|\nabla f(x^k)\|^4}.$$

Angenommen, es existiert $\varepsilon > 0$ und $k_\epsilon \in \mathbb{N}$, so dass $\|\nabla f(x^k)\| \geq \epsilon$ für alle $k \geq k_\epsilon$. Wir haben dann

$$\begin{aligned}
\gamma_{k+1} &= \frac{\|d^{k+1}\|^2}{\|\nabla f(x^{k+1})\|^4} \\
&= \frac{(-\nabla f(x^{k+1})^T + \beta_k^{PR} d^k)^T(-\nabla f(x^{k+1})^T + \beta_k^{PR} d^k)}{\|\nabla f(x^{k+1})\|^4} \\
&= \frac{1}{\|\nabla f(x^{k+1})\|^2} - 2\beta_k^{PR} \cdot \frac{\nabla f(x^{k+1})d^k}{\|\nabla f(x^{k+1})\|^4} + (\beta_k^{PR})^2 \cdot \frac{\|d^k\|^2}{\|\nabla f(x^{k+1})\|^4} \\
&= \frac{1}{\|\nabla f(x^{k+1})\|^2} + (\beta_k^{PR})^2 \cdot \frac{\|d^k\|^2}{\|\nabla f(x^{k+1})\|^4} \\
&= \frac{1}{\|\nabla f(x^{k+1})\|^2} + \gamma_k \cdot (\beta_k^{PR})^2 \cdot \frac{\|\nabla f(x^k)\|^4}{\|\nabla f(x^{k+1})\|^4}
\end{aligned}$$

mit

$$\begin{aligned}(\beta_k^{PR})^2 \cdot \frac{\|\nabla f(x^k)\|^4}{\|\nabla f(x^{k+1})\|^4} &\leq \frac{\gamma^2 \cdot \|x^{k+1}-x^k\|^2 \cdot \|\nabla f(x^{k+1})\|^2}{\|\nabla f(x^k)\|^4} \cdot \frac{\|\nabla f(x^k)\|^4}{\|\nabla f(x^{k+1})\|^4} \\ &\leq \gamma^2 \|x^{k+1}-x^k\|^2 \cdot \frac{1}{\varepsilon^2} \quad (k \geq k_\varepsilon) \\ &\to 0 \quad (k \to \infty).\end{aligned}$$

Es existiert deshalb ein $k_1 \in \mathbb{N}$, so dass für alle $k \geq k_1$ gilt

$$\gamma_{k+1} \leq \frac{1}{\varepsilon^2} + \gamma_k$$

und damit auch

$$\frac{1}{\gamma_k} \leq \frac{1}{\varepsilon^2} \cdot (k-k_1) + \gamma_{k_1} \leq c \cdot (k-k_1), \; c > 0.$$

Für die Kehrwerte gilt dann

$$\frac{1}{\gamma_k} \geq \frac{1}{c} \cdot \frac{1}{k-k_1}, \; k > k_1.$$

Also divergiert die Summe

$$\sum_{k=0}^{\infty} \frac{1}{\gamma_k}.$$

Andererseits folgt aus (5.3.2) mit einer unteren Schranke α für $f(x)$ auf $\mathbb{R}^n$

$$f(x^0) - \alpha \geq \theta \cdot \sum_{k=0}^{\infty} \frac{1}{\gamma_k},$$

ein Widerspruch. □

Störend an diesem Resultat ist vor allem die Voraussetzung an $\|x^{k+1}-x^k\|$, welche im Beweis aber entscheidend eingeht. Es gibt tatsächlich Beispiele, die zeigen, dass ohne diese Voraussetzung der Satz nicht gilt. Zufriedenstellender ist die Situation im Falle einer gleichmäßig konvexen Funktion.

5.3.4 Satz
Sei $f : \mathbb{R}^n \to \mathbb{R}$, $f \in \mathcal{C}^2(\mathbb{R}^n)$, $x^0 \in \mathbb{R}^n$ sowie f gleichmäßig konvex auf der konvexen Levelmenge $\mathcal{L}(x^0)$. Die Folge $\{x^k\}$ entstehe aus dem Polak-Ribière-Verfahren (Algorithmus 5.3.1). Dann gilt

$$\lim_{k\to\infty} x^k = x^*,$$

wobei x^* die eindeutige globale Minimalstelle von f ist.

Beweis: Wir weisen die Winkelbedingung (2.1.2) nach. Da exakte Schrittweiten effizient sind (Aufgabe 7), folgt die Behauptung aus Korollar 2.1.8. Wir haben

$$\frac{-\nabla f(x^k)d^k}{\|f(x^k)\| \cdot \|d^k\|} = \frac{\|\nabla f(x^k)\|^2}{\|f(x^k)\| \cdot \|d^k\|} = \frac{\|\nabla f(x^k)\|}{\|d^k\|}.$$

Wir müssen also noch $\|d^k\|$ relativ zu $\|\nabla f(x^k)\|$ abschätzen. Es gilt

$$\|d^k\| = \| - \nabla f(x^k) + \beta_{k-1}^{PR} d^{k-1}\| \leq \|\nabla f(x^k)\| + |\beta_{k-1}^{PR}| \cdot \|d^{k-1}\|.$$

Hierin haben wir

$$|\beta_{k-1}^{PR}| \leq \frac{\gamma \cdot t_{k-1} \cdot \|d^{k-1}\| \cdot \|\nabla f(x^k)\|}{\|\nabla f(x^{k-1})\|^2} = \frac{\gamma \cdot t_{k-1} \cdot \|d^{k-1}\|}{\|\nabla f(x^{k-1})\|^2} \cdot \|\nabla f(x^k)\|.$$

Wir sind also fertig, wenn wir

$$t_{k-1} \leq c \cdot \frac{\|\nabla f(x^{k-1})\|^2}{\|d^{k-1}\|^2}$$

zeigen können. Auf Grund der gleichmäßigen Konvexität existiert $\mu > 0$ mit

$$(\nabla f(y) - \nabla f(x))(y - x) \geq \mu \cdot \|y - x\|^2.$$

Damit erhalten wir

$$\begin{aligned} 0 &= \nabla f(x^k)d^{k-1} \\ &= \left(\nabla f(x^k) - \nabla f(x^{k-1})\right)(x^k - x^{k-1}) + \nabla f(x^{k-1})(x^k - x^{k-1}) \\ &\geq \mu \cdot t_{k-1}^2 \|d^{k-1}\|^2 + t_{k-1} \cdot \nabla f(x^{k-1})d^{k-1} \end{aligned}$$

woraus sich wie gewünscht

$$t_{k-1} \leq \frac{1}{\mu} \cdot \frac{-\nabla f(x^{k-1})d^{k-1}}{\|d^{k-1}\|^2} = \frac{1}{\mu} \cdot \frac{\|\nabla f(x^{k-1})\|^2}{\|d^{k-1}\|^2}$$

ergibt. □

Im Jahre 1997 wurde von Grippo und Lucidi eine Modifikation des Polak-Ribière-Verfahrens vorgeschlagen, das in der Praxis ähnlich effizient ist wie das nicht-modifizierte Verfahren, dessen Konvergenztheorie aber wesentlich befriedigender ist. Dabei wird die Schrittweite mit einer Armijo-ähnlichen Regel bestimmt.

5.3.5 Algorithmus (modifiziertes Polak-Ribière-Verfahren)
wähle $x^0 \in \mathbb{R}^n$, wähle $0 < \sigma < 1/2$, setze $d^0 = -\nabla f(x^0)^T$
wähle $\beta \in (0,1)$, $0 < \delta_1 < 1 < \delta_2$
for $k = 0, 1, \ldots$ **do**
  setze $\rho_k = |\nabla f(x^k) d^k| / \|d^k\|^2$
  bestimme die größtmögliche Schrittweite $t_k \in \{\rho_k \beta^\ell : \ell = 0, 1, \ldots\}$
  so dass für $x^{k+1} = x^k + t_k d^k$, $d^{k+1} = -\nabla f(x^{k+1})^T + \beta_k^{PR} d^k$ gilt
    $f(x^{k+1}) \le f(x^k) - t_k^2 \sigma \|d^k\|^2$ und
    $-\delta_2 \cdot \|\nabla f(x^{k+1})\|^2 \le \nabla f(x^{k+1}) d^{k+1} \le -\delta_1 \cdot \|\nabla f(x^{k+1})\|^2$
    {zwei Bedingungen für t_k!}
end for

Wir weisen zuerst nach, dass die Iterierten des modifizierten Verfahrens wohldefiniert sind.

5.3.6 Lemma
Sei $f : \mathbb{R}^n \to \mathbb{R}$, $f \in \mathcal{C}^1(\mathbb{R}^n)$. Weiter sei $x \in \mathbb{R}^n, \sigma, \beta \in (0,1), 0 < \delta_1 < 1 < \delta_2$ und d sei eine Abstiegsrichtung für f in x. Dann existiert ein $\ell \in \mathbb{N}_0$, so dass für $t_\ell = \frac{|\nabla f(x) d|}{\|d\|^2} \cdot \beta^\ell$ mit

$$x^+ = x + t_\ell d,\ d^+ = -\nabla f(x^+)^T + \beta^{PR} d,\ \beta^{PR} = \frac{(\nabla f(x^+) - \nabla f(x)) \nabla f(x^+)^T}{\|\nabla f(x)\|^2}$$

gilt

$$f(x^+) \le f(x) - \sigma t_\ell^2 \|d\|^2, \tag{5.3.3}$$

$$-\delta_2 \|\nabla f(x^+)\|^2 \le \nabla f(x^+) d^+ \le -\delta_1 \|\nabla f(x^+)\|^2. \tag{5.3.4}$$

Insbesondere ist d^+ eine Abstiegsrichtung für f in x^+, solange $\nabla f(x^+) \neq 0$.

Beweis: Wir notieren y^ℓ für $x + t_\ell d$. Angenommen, für unendlich viele t_ℓ gilt

$$f(y^\ell) > f(x) - \sigma t_\ell^2 \|d\|^2.$$

Dann folgt für die Teilfolge t_{ℓ_i} mit dieser Eigenschaft

$$\nabla f(x) d = \lim_{i \to \infty} \frac{f(y^\ell) - f(x)}{t_{\ell_i}} \ge \lim_{i \to \infty} -t_{\ell_i} \cdot \sigma \|d\|^2 = 0,$$

im Widerspruch dazu, dass d Abstiegsrichtung ist. Also ist (5.3.3) für alle $\ell \ge \ell_0$ erfüllt.

Auch zum Beweis von (5.3.4) nehmen wir zunächst an, dass für unendlich viele ℓ die Beziehung nicht erfüllt ist. Die zugehörige Teilfolge der t_ℓ heiße wieder t_{ℓ_i}. Für alle i gilt also

$$-\delta_2\|\nabla f(y^{\ell_i})\|^2 > \nabla f(y^{\ell_i})\left(-\nabla f(y^{\ell_i})^T + \beta_i^{PR} d\right)$$

oder

$$\nabla f(y^{\ell_i})\left(-\nabla f(y^{\ell_i})^T + \beta_i^{PR} d\right) > -\delta_1\|\nabla f(y^{\ell_i})\|^2$$

mit

$$\beta_i^{PR} = \frac{\left(\nabla f(y^{\ell_i}) - \nabla f(x)\right)\nabla f(y^{\ell_i})^T}{\|\nabla f(x)\|^2}.$$

Daraus folgt für $i \to \infty$ wegen $\lim\limits_{i\to\infty} \beta_i^{PR} = 0$, dass eine der beiden Beziehungen

$$-\delta_2\|\nabla f(x)\|^2 \geq -\|\nabla f(x)\|^2$$

oder

$$-\|\nabla f(x)\|^2 \geq -\delta_1\|\nabla f(x)\|^2$$

gilt, woraus wegen $\delta_2 > 1, \delta_1 < 1$ dann $\nabla f(x) = 0$ folgt. Dies widerspricht aber der Voraussetzung, den notwendig für das Vorliegen einer Abstiegsrichtung ist $\nabla f(x) \neq 0$. Also ist auch (5.3.4) für $\ell \geq \ell_1$ erfüllt. Für ℓ groß genug, sind demnach beide Bedingungen erfüllt; der Algorithmus verwendet dann das kleinste aller solchen ℓ. □

Wir kommen jetzt zum entscheidenden Konvergenzsatz für das modifizierte Polak-Ribière-Verfahren.

5.3.7 Satz
Sei $f : \mathbb{R}^n \to \mathbb{R}, f \in \mathcal{C}^1(\mathbb{R}^n)$. Für $x^0 \in \mathbb{R}^n$ sei die Levelmenge $\mathcal{L}(x^0)$ kompakt und es sei $\nabla f \in \mathrm{Lip}_\gamma(B)$ mit einer Menge $B \supseteq \{y \in \mathbb{R}^n :$ es existiert $x \in \mathcal{L}(x^0)$ mit $\|x-y\| \leq r\} \supseteq \mathcal{L}(x^0)$ mit einem $r > 0$. Dann gilt für die Iterierten x^k des modifizierten Polak-Ribière-Verfahrens (Algorithmus 5.3.5)

$$\lim_{k\to\infty} \|\nabla f(x^k)\| = 0.$$

Insbesondere ist $\nabla f(x^*) = 0$ für jeden Häufungspunkt der Folge $\{x^k\}$.

Beweis: Der Beweis ist etwas aufwendiger. Wir halten zunächst fest, dass eine Konstante $c > 0$ existiert mit

$$\|\nabla f(x)\| \leq c \text{ für } x \in \mathcal{L}(x^0).$$

Wir zeigen nun die folgenden Aussagen

(i) $x^k \in \mathcal{L}(x^0)$ für alle k,

(ii) die Folge $\{f(x^k)\}$ ist konvergent,

(iii) $\lim\limits_{k\to\infty} t_k\|d^k\| = 0$,

(iv) $t_k\|d^k\|^2 \leq \delta_2 c^2$ für alle k,

(v) es existiert eine Konstante $\theta > 0$ mit

$$t_k \geq \theta \cdot \frac{|\nabla f(x^k)d^k|}{\|d^k\|} \text{ für alle } k.$$

zu (i): Dies ist trivial, da wir ein Abstiegsverfahren vorliegen haben.

zu (ii): Dies ist ebenfalls trivial, denn die Funktion f nimmt auf der kompakten Menge $\mathcal{L}(x^0)$ ihr Minimum aus f^* an. Es gilt also

$$f^* \leq f(x^{k+1}) \leq f(x^k) \text{ für alle } k,$$

d.h. $\lim\limits_{k\to\infty} f(x^k)$ existiert (und ist $\geq f^*$).

zu (iii): Aus dem Verfahren folgt

$$f(x^{k+1}) - f(x^k) \leq -\sigma t_k^2\|d^k\|^2 \leq 0 \text{ für alle } k.$$

Wegen (ii) folgt für $k \to \infty$

$$\lim_{k\to\infty} t_k^2\|d^k\|^2 = 0,$$

was (iii) beweist.

zu (iv): Nach dem Verfahren gilt für alle k

$$t_k\|d^k\|^2 \leq \rho_k\|d^k\|^2 = |\nabla f(x^k)d^k| \leq \delta_2\|\nabla f(x^k)\|^2.$$

Daraus folgt sofort $t_k\|d^k\|^2 \leq \delta_2 c^2$.

zu (v): Wir benötigen eine mehrfach geschachtelte Fallunterscheidung.
Fall 1: $t_k = \rho_k$. Dann ist

$$t_k = \rho_k = \frac{|\nabla f(x^k)d^k|}{\|d^k\|^2} \geq \theta \cdot \frac{|\nabla f(x^k)d^k|}{\|d^k\|^2} \quad \text{für alle } \theta \in (0,1].$$

Fall 2: $t_k < \rho_k$. Dann war $t = t_k/\beta$ noch keine zulässige Schrittweite, d.h. für $y = x^k + \frac{t_k d^k}{\beta}$ gilt

$$f(y) > f(x^k) - \sigma\left(\frac{t_k}{\beta}\right)^2\|d^k\|^2 \tag{5.3.5}$$

oder

$$\nabla f(y)d^{k+1} > -\delta_1\|\nabla f(y)\|^2 \tag{5.3.6}$$

oder

$$-\delta_2\|\nabla f(y)\|^2 > \nabla f(y)d^{k+1} \tag{5.3.7}$$

Wir betrachten ab jetzt nur $k \geq k_0$, wobei k_0 so gewählt ist, dass

$$x^k + \frac{t_k}{\beta} \cdot \|d^k\| \in B \text{ für } k \geq k_0.$$

Ein solches k_0 existiert wegen (iii).
Fall 2a: Es gilt (5.3.5). Nach dem Mittelwertsatz ist

$$f(y) = f(x^k) + \nabla f(\xi)(y - x^k),\ \xi = x^k + \vartheta(y - x^k),\ \vartheta \in (0,1).$$

Mit (5.3.5) und der Lipschitz-Stetigkeit folgt dann

$$\begin{aligned} -\sigma\left(\frac{t_k}{\beta}\right)^2\|d^k\|^2 &< \nabla f(\xi)\left(\frac{t_k}{\beta}d^k\right) \\ &= (\nabla f(\xi) - \nabla f(x^k))\left(\frac{t_k}{\beta}d^k\right) + \nabla f(x^k)\left(\frac{t_k}{\beta}d^k\right) \\ &\leq \gamma \cdot \underbrace{\vartheta}_{\leq 1}\left(\frac{t_k}{\beta}\|d^k\|\right)^2 + \nabla f(x^k)\left(\frac{t_k}{\beta}d^k\right), \end{aligned}$$

und damit

$$t_k \geq \frac{\beta}{\gamma+\sigma} \cdot \frac{|\nabla f(x^k)d^k|}{\|d^k\|^2} \geq \theta \cdot \frac{|\nabla f(x^k)d^k|}{\|d^k\|^2} \text{ für alle } \theta \in \left(0, \frac{\beta}{\gamma+\sigma}\right).$$

Fall 2b: Es gilt (5.3.6). Unter Verwendung der Definition von β_k^{PR} gilt also

$$\nabla f(y)\left(-\nabla f(y) + \frac{\nabla f(y)(\nabla f(y) - \nabla f(x^k))^T}{\|\nabla f(x^k)\|^2}d^k\right) > -\delta_1\|\nabla f(y)\|^2.$$

Mit mehrfacher Anwendung der CSU ergibt sich

$$-\|\nabla f(y)\|^2 + \frac{\|\nabla f(y)\|^2 \cdot \|\nabla f(y) - \nabla f(x^k)\| \cdot \|d^k\|}{\|\nabla f(x^k)\|^2} > -\delta_1\|\nabla f(y)\|^2$$

und deshalb

$$-1 + \frac{\|\nabla f(y) - \nabla f(x^k)\| \cdot \|d^k\|}{\|\nabla f(x^k)\|^2} > -\delta_1.$$

Mit der Lipschitz-Stetigkeit ergibt sich

$$-1 + \frac{\gamma \cdot \frac{t_k}{\beta} \cdot \|d^k\|^2}{\|\nabla f(x^k)\|^2} > -\delta_1$$

und damit

$$t_k > \frac{\beta(1-\delta_1)}{\gamma} \cdot \frac{\|\nabla f(x^k)\|^2}{\|d^k\|^2}.$$

Weil x^k aus dem Verfahren gewonnen wurde, haben wir $|\nabla f(x^k)d^k| \leq \delta_2 \|\nabla f(x^k)\|^2$. Damit ergibt sich

$$t_k > \frac{\beta(1-\delta_1)}{\gamma\delta_2} \cdot \frac{|\nabla f(x^k)d^k|}{\|d^k\|^2} \geq \theta \cdot \frac{|\nabla f(x^k)d^k|}{\|d^k\|^2} \quad \text{für alle } \theta \in \left(0, \frac{\beta(1-\delta_1)}{\gamma\delta_2}\right).$$

Fall 2c: Es gilt (5.3.7), also

$$\nabla f(y)\left(-\nabla f(y) + \frac{\nabla f(y)^T(\nabla f(y) - \nabla f(x^k))}{\|\nabla f(x^k)\|^2} d^k\right) < -\delta_2 \|\nabla f(y)\|^2.$$

Analog zu Fall 2b ergibt sich hier

$$t_k \geq \frac{\beta(\delta_2 - 1)}{\gamma\delta_2} \cdot \frac{|\nabla f(x^k)d^k|}{\|d^k\|^2}.$$

Nimmt man also

$$\theta \leq \min\left\{1, \frac{\beta}{\gamma+\sigma}, \frac{\beta(1-\delta_1)}{\gamma\delta_2}, \frac{\beta(\delta_2-1)}{\gamma\delta_2}\right\},$$

so ist (v) jedenfalls für $k \geq k_0$ erfüllt. Durch weitere Verkleinerung von θ erreicht man, das dann (v) sogar für alle k gilt.
Nach diesen vorbereitenden Folgerungen beweisen wir jetzt schließlich die Aussage des Satzes und nehmen dazu an, $\lim_{k\to\infty} \|\nabla f(x^k)\| = 0$ gilt nicht. Dann existiert $\varepsilon > 0$ und eine Teilfolge $\{x^{k_i}\}$ mit $\|\nabla f(x^{k_i})\| \geq \varepsilon$ für alle i. Dann haben wir

$$\begin{aligned} \|d^{k_i+1}\| &\leq \|\nabla f(x^{k_i+1})\| + \frac{\|\nabla f(x^{k_i+1})\| \cdot \|\nabla f(x^{k_i+1}) - \nabla f(x^{k_i})\|}{\|\nabla f(x^{k_i})\|^2} \|d^{k_i}\| \\ &\leq c + \delta_2 \cdot \frac{\gamma c^3}{\varepsilon^2}. \end{aligned}$$

Aus (iii) folgt deshalb

$$\lim_{i\to\infty} t_{k_i+1} \|d^{k_i+1}\|^2 = 0,$$

und damit aus (v)

$$\lim_{i\to\infty} |\nabla f(x^{k_i+1})d^{k_i+1}| = 0,$$

woraus wegen der Schrittweitenwahl

$$\lim_{i\to\infty} \|\nabla f(x^{k_i+1})\| = 0$$

folgt. Hieraus ergibt sich mit (iii)

$$\begin{aligned}
\|\nabla f(x^{k_i})\| &\leq \|\nabla f(x^{k_i+1}) - \nabla f(x^{k_i})\| + \|\nabla f(x^{k_i+1})\| \\
&\leq \gamma \cdot \|x^{k_i+1} - x^{k_i}\| + \|\nabla f(x^{k_i+1})\| \\
&= \gamma \cdot t_{k_i} d^{k_i} + \|\nabla f(x^{k_i+1})\| \\
&\to 0 \ \ (k \to \infty),
\end{aligned}$$

im Widerspruch zur Annahme $\|\nabla f(x^{k_i})\| \geq \varepsilon$. □

Kapitel 6

Trust-Region Verfahren

Trust-Region: Vertrauensgebiet. Trust-Region Verfahren sind *keine* Abstiegsverfahren und deshalb eine völlig neue Verfahrensklasse. Abstiegsverfahren sind dadurch charakterisiert, dass zuerst eine Abstiegsrichtung d gewählt wird und dann eine geeignete Schrittweite t bestimmt wird. Im Newton-Verfahren ist z.B.

$$d = -\nabla^2 f(x)^{-1} \nabla f(x)^T$$

und t wird anschließend z.B. über die Armijo-Regel bestimmt.
Bei den Trust-Region Verfahren ist der Ansatz genau umgekehrt. Es wird eine (maximale) Schrittweite Δ vorgegeben, dann wird eine geeignete Richtung (und damit die nächste Iterierte) gesucht. Hierzu wird zuerst ein einfaches, quadratisches Modell der Funktion erstellt, z.B. durch Taylor-Entwicklung um x

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

Jetzt nimmt man das Minimum von q unter der Nebenbedinqung $\|y-x\| \leq \Delta$ als nächste Iterierte. Die Bestimmung dieses Minimums bezeichnen wir als *Trust-Region Teilproblem*. Ist Δ groß genug und $\nabla^2 f(x)$ spd erhält man im Beispiel wieder die Newton-Iterierte.
Wir untersuchen zunächst Trust-Region Verfahren so, als ob wir das Trust-Region Teilproblem exakt lösen können. Dann werden wir zeigen, dass sich die Theorie aber auch auf geeignete approximative Lösungen übertragen wird. Schließlich werden wir uns sehr ausführlich mit der exakten Lösung des Trust-Region Teilproblems beschäftigen, auch als Ausblick auf typische Techniken für restringierte Aufgaben.

Abschnitt 6.1

Trust-Region Newton-Verfahren

Wir formulieren und analysieren in diesem Abschnitt ein vollständiges Trust-Region-Verfahren.
Es sei $f : \mathbb{R}^n \to \mathbb{R}$, $f \in \mathcal{C}^2(\mathbb{R}^n)$. Das quadratische Modell q_x entsteht durch Taylor-Entwicklung zweiter Ordnung um den Punkt $x \in \mathbb{R}^n$, also

$$q_x(d) = f(x) + \nabla f(x)d + \frac{1}{2}d^T \nabla^2 f(x)d,$$

das zugehörige Trust-Region Teilproblem lautet

(6.1.1) bestimme d^x mit $q_x(d^x) \leq q_x(d)$ für alle d mit $\|d\| \leq \Delta$.

In einem Trust-Region Algorithmus müssen wir in jedem Schritt entscheiden, wie groß der Radius Δ der Trust-Region gewählt wird. Dazu vergleichen wir den durch das Modell vorhergesagten Abstieg $q_x(0) - q_x(d)$ mit dem tatsächlich erreichten Abstieg $f(x) - f(x^+)$ (wobei $x^+ = x + d, q_x(0) = f(x)$) und setzen dazu

$$r = \frac{f(x^+) - f(x)}{q_x(d) - f(x)}.$$

Ist r nahe bei 0, so stimmt das quadratische Modell nur wenig mit der Funktion überein und der Radius Δ der Trust Region muss verkleinert werden. Ist r nahe bei 1 oder ist r sogar größer als 1, so kann Δ für die Lösung des nächsten Trust-Region Teilproblems vorsorglich vergrößert werden.
Diese Strategie wird in dem im folgenden beschriebenen Algorithmus formuliert. Wenn r zu klein ist, wird die neue Iterierte nicht akzeptiert, und die neue Iterierte muss mit einem kleineren Δ nochmal berechnet werden. Wir verwenden die Parameter $\sigma_1 \in (0, 1)$ und $\sigma_2 \in (1, \infty)$ um Δ zu verkleinern bzw. zu vergrößern, und die Parameter $0 < \rho_1 < \rho_2 < 1$ um zu entscheiden, ob Δ verkleinert, beibehalten oder vergrößert wird. Außerdem verhindern wir, dass Δ zu klein wird.

6.1.1 Algorithmus (Trust-Region Newton-Verfahren)

wähle $x^0 \in \mathbb{R}^n$, $\Delta_0 \geq \Delta_{min} > 0$, $0 < \rho_1 < \rho_2 < 1$, $0 < \sigma_1 < 1 < \sigma_2$, $\varepsilon \geq 0$
for $k = 0, 1, \ldots$ **do**
  repeat
    bestimme eine Lösung d^k des Trust-Region Teilproblems (6.1.1)
    (mit $x = x^k$, $\Delta = \Delta_k$) {z.B. mit Algorithmus 6.5.1}
    berechne $r_k = \frac{f(x^k+d^k)-f(x^k)}{q_{x^k}(d^k)-f(x^k)}$
    $\Delta_k = \sigma_1 \cdot \Delta_k$
  until $r^k \geq \rho_1$
  $\overline{Delta}_k = \frac{1}{\sigma_1}\Delta_k$ {mit $\overline{\Delta}_k$ wurden r^k, d^k berechnet}
  $x^{k+1} = x^k + d^k$
  if $r^k \leq \rho_2$ **then**
    setze $\Delta_{k+1} = \max\{\Delta_{min}, \Delta_k\}$ $\{r^k \in [\rho_1, \rho_2]\}$
  else
    setze $\Delta_{k+1} = \max\{\Delta_{min}, \sigma_2 \cdot \Delta_k\}$ $\{r^k > \rho_2\}$
  end if
end for

Der Deutlichkeit halber haben wir den im Algorithmus zur Berechnung von x^{k+1} tatsächlich verwendeten Radius für die Trust-Region mit $\overline{\Delta}_k$ bezeichnet. Als ein erstes Resultat zur Analyse von Algorithmus 6.1.1 zeigen wir, dass im Falle $\nabla f(x^k) \neq 0$ stets $q_{x^k}(d^k) < q_{x^k}(0)$ gilt. Deshalb sind alle r^k und damit der ganze Algorithmus wohldefiniert.

6.1.2 Lemma
Es sei $x \in \mathbb{R}^n$, $\Delta > 0$ und d löse das Trust-Region Teilproblem 6.1.1. Dann gilt

$$f(x) - q_x(d) \geq \frac{\|\nabla f(x)\|}{2} \cdot \min\left\{\Delta, \frac{\|\nabla f(x^k)\|}{\|\nabla^2 f(x)\|}\right\}.$$

(Im Falle $\nabla^2 f(x) = 0$ ist das Minimum $= \Delta_{min}$ zu setzen.)

Beweis: Wir setzen $\varphi(t) = q_x(-t \cdot \frac{\Delta}{\|\nabla f(x)\|} \cdot \nabla f(x)^T)$, $t \in [0, 1]$ und notieren $H = \nabla^2 f(x)$. Dann gilt

$$\begin{aligned}
\varphi(0) - \varphi(t) &= t \cdot \Delta\|\nabla f(x)\| - \frac{t^2}{2} \cdot \frac{\Delta^2}{\|\nabla f(x)\|^2} \cdot \nabla f(x) H \nabla f(x)^T \\
&\geq t \cdot \Delta\|\nabla f(x)\| - \frac{t^2}{2} \cdot \Delta^2 \cdot \|H\| \\
&=: \psi(t).
\end{aligned}$$

Die Funktion $\psi(t)$ hat ihr Maximum bei $t^* = \|\nabla f(x)\|/(\Delta\|H\|)$ mit Wert $\psi(t^*) = \|\nabla f(x)\|^2/(2\|H\|)$. Im Intervall $[0, t^*]$ ist ψ monoton wachsend. Ist $t^* > 1$ so wird das Maximum auf $[0, 1]$ für $t = 1$ angenommen und es gilt

$$\begin{aligned}\psi(1) &= \Delta\|\nabla f(x)\| - \frac{1}{2}\Delta^2 \cdot \|H\| \\ &\overset{t^*>1}{\geq} \Delta\|\nabla f(x)\| - \frac{\Delta}{2}\|\nabla f(x)\| \\ &= \frac{1}{2}\Delta\|\nabla f(x)\|.\end{aligned}$$

Insgesamt erhalten wir damit

$$\max_{t\in[0,1]}\{\varphi(0) - \varphi(t)\} \geq \frac{1}{2}\|\nabla f(x)\| \min\left\{\Delta, \frac{\|\nabla f(x^k)\|}{\|\nabla^2 f(x)\|}\right\}.$$

Dies gilt dann erst recht für die Größe

$$\max_{d\in B_\Delta(0)}\{q_x(0) - q_x(d)\},$$

welche an der Lösung des Trust-Region Teilproblems angenommen wird. □

Das nächste Resultat zeigt, dass die Repeat-Schleife stets abbricht, solange $\nabla f(x^k) \neq 0$. Dabei interessiert jetzt bei festen x das Trust-Region Teilproblem in Abhängigkeit vom Radius Δ, so dass wir dessen Lösung d_Δ mit dem Index Δ versehen.

6.1.3 Lemma
Sei $f \in \mathcal{C}^2(\mathbb{R}^n)$ und $x \in \mathbb{R}^n$ mit $\nabla f(x) \neq 0$. Für $\Delta > 0$ bezeichne d_Δ die Lösung des Trust-Region Teilproblems (6.1.1) mit der Funktion q_x und Radius Δ für die Trust-Region. Dann gilt

$$\lim_{\Delta\to 0}\frac{f(x) - f(x + d_\Delta)}{f(x) - q_x(d_\Delta)} = 1.$$

(Wegen $\rho_1 \in (0, 1)$ bedeutet das, dass die Repeat-Schleife in Algorithmus 6.1.1 abbricht.)

Beweis: Mit einer Taylor-Entwicklung haben wir einerseits

$$f(x + d_\Delta) = f(x) + \nabla f(\xi_\Delta)d_\Delta,\ \xi_\Delta = x + \theta d_\Delta,\ \theta \in (0, 1).$$

Andererseits haben wir nach Lemma 6.1.2 für Δ genügend klein

$$f(x) - q_x(d_\Delta) \geq \frac{\|\nabla f(x)\|}{2} \cdot \min\left\{\Delta, \frac{\|\nabla f(x^k)\|}{\|\nabla^2 f(x)\|}\right\} = \frac{\|\nabla f(x)\|}{2} \cdot \Delta. \tag{6.1.2}$$

Für diese Δ gilt dann

$$\begin{aligned}
&\left|\frac{f(x)-f(x+d_\Delta)}{f(x)-q_x(d_\Delta)}-1\right| \\
&= \left|\frac{q_x(d_\Delta)-f(x+d_\Delta)}{f(x)-q_x(d_\Delta)}\right| \\
&\leq \left|\frac{q_x(d_\Delta)-f(x+d_\Delta)}{\Delta}\cdot\frac{2}{\|\nabla f(x)\|}\right| \\
&= \left|\frac{2}{\Delta\|\nabla f(x)\|}\left(f(x)+\nabla f(x)d_\Delta+\frac{1}{2}d_\Delta^T\nabla^2 f(x)d_\Delta-(f(x)+\nabla f(\xi_\Delta)d_\Delta)\right)\right| \\
&= \left|\frac{2}{\Delta\|\nabla f(x)\|}\left((\nabla f(x)-\nabla f(\xi_\Delta))d_\Delta+\frac{1}{2}d_\Delta^T\nabla^2 f(x)d_\Delta)\right)\right| \\
&\leq \frac{2\|d_\Delta\|}{\Delta\|\nabla f(x)\|}\left(\|\nabla f(x)-\nabla f(\xi_\Delta)\|+\|d_\Delta\|\cdot\|\nabla^2 f(x)\|\right) \\
&\leq \frac{2}{\|\nabla f(x)\|}\left(\|\nabla f(x)-\nabla f(\xi_\Delta)\|+\Delta\|\nabla^2 f(x)\|\right) \\
&\quad \to 0\ (\Delta\to 0,\ \text{denn } \xi_\Delta\to x).
\end{aligned}$$

□

6.1.4 Korollar
Sei $f\in\mathcal{C}^2(\mathbb{R}^n)$ und $\{x^k\}$ sei eine Folge von Iterierten von Algorithmus 6.1.1 und $\{x^{k_i}\}$ eine gegen einen Punkt x^* mit $\nabla f(x^*)\neq 0$ konvergente Teilfolge. Dann gilt

$$\liminf_{i\to\infty}\overline{\Delta}_{k_i}>0.$$

Beweis: Der Beweis ergibt sich in Analogie zu Lemma 6.1.3: Für $i\geq i_0$ ist $\|\nabla f(x^{k_i})\|\geq\beta>0$. Statt (6.1.2) haben wir deshalb für alle solchen i und für $\Delta\leq\Delta_0$

$$f(x^{k_i})-q_{x^{k_i}}(d_\Delta^i)\geq\frac{\beta}{2}\cdot\Delta.$$

Hierein ist d_Δ^i die Lösung der Trust-Region Teilaufgabe für $x=x^{k_i}$ und Radius Δ. Außerdem erhalten wir wie in Lemma 6.1.3

$$\begin{aligned}
&\left|\frac{f(x^{k_i})-f(x^{k_i}+d_\Delta^i)}{f(x^{k_i})-q_{x^{k_i}}(d_\Delta^i)}-1\right| \\
&\leq \frac{2}{\beta}\left(\|\nabla f(x^{k_i})-\nabla f(\xi_\Delta^i)\|+\Delta\|\nabla^2 f(x^{k_i})\|\right) \qquad (6.1.3)
\end{aligned}$$

mit $\xi_\Delta^i=x^{k_i}+\theta_i d_\Delta^i$, $\theta_i\in(0,1)$.

Wir nehmen nun an, es sei

$$\liminf_{i\to\infty} \overline{\Delta}_{k_i} = 0,$$

so dass wir durch Übergang zu einer Teilfolge, die wir wieder mit k_i indizieren, sogar

$$\lim_{i\to\infty} \overline{\Delta}_{k_i} = 0$$

annehmen können. Für alle großen i bedeutet dies, dass die Repeat-Schleife für k_i mehrfach durchlaufen wird. Setzen wir also $\hat{\Delta}_{k_i} = \frac{1}{\sigma_1}\overline{\Delta}_{k_i}$, so gilt

$$\lim_{i\to\infty} \hat{\Delta}_{k_i} = 0$$

sowie

$$\frac{f(x^{k_i}) - f(x^{k_i} + d_{\hat{\Delta}_{k_i}})}{f(x^{k_i}) - q_{x^{k_i}}(d_{\hat{\Delta}_{k_i}})} < \rho_1. \tag{6.1.4}$$

Weil $\nabla^2 f$ stetig ist, existiert eine Konstante η mit $\|\nabla^2 f(x^{k_i})\| \le \eta$ für alle i. Weil ∇f in einer Umgebung von x^* Lipschitz-stetig ist (denn $\nabla^2 f$ ist stetig), existieren Konstanten $\gamma, \delta > 0$ mit

$$\|\nabla f(y) - \nabla f(z)\| \le \gamma \cdot \|y - z\| \quad \text{für alle } y, z \in B_\delta(x^*).$$

Wegen $\lim_{i\to\infty} x^{k_i} = x^*$ und $\lim_{i\to\infty} d_{\hat{\Delta}_{k_i}} = 0$ (denn $\|d_{\hat{\Delta}_{k_i}}\| \le \hat{\Delta}_{k_i}$) folgt aus (6.1.3) für $i \ge i_2$

$$\left| \frac{f(x^{k_i}) - f(x^{k_i} + d_{\hat{\Delta}_{k_i}})}{f(x^{k_i}) - q_{x^{k_i}}(d_{\hat{\Delta}_{k_i}})} - 1 \right| \le \frac{2}{\beta}\left(\gamma \cdot \hat{\Delta}_{k_i} + \hat{\Delta}_{k_i} \cdot \eta\right) \to 0 \ (i \to \infty),$$

im Widerspruch zu (6.1.4). □

Nach diesen Vorbereitungen sind wir nun in der Lage, die Konvergenzanalyse des Trust-Region Newton-Verfahrens voranzutreiben.

6.1.5 Satz
Sei $f : \mathbb{R}^n \to \mathbb{R}, f \in \mathcal{C}^2(\mathbb{R}^n)$. Die Folge $\{x^k\}$ sei durch das Trust-Region Newton-Verfahren (Algorithmus 6.1.1) erzeugt mit $\nabla f(x^k) \neq 0$ für alle k. Dann ist jeder Häufungspunkt der Folge $\{x^k\}$ ein stationärer Punkt von f.

Beweis: Sei x^* ein Häufungspunkt mit $\lim_{i\to\infty} x^{k_i} = x^*$. Angenommen, es gilt $\nabla f(x^*) \neq 0$. Dann existieren $\alpha, \beta > 0$, so dass für alle i gilt

$$\|\nabla f(x^{k_i})\| \geq \alpha, \ \|\nabla^2 f(x^{k_i})\| \leq \beta.$$

Nach dem Algorithmus ist $r^{k_i} \geq \rho_1$ für alle i. Mit Lemma 6.1.2 gilt dann

$$\begin{aligned} f(x^{k_i}) - f(x^{k_i+1}) &\geq \rho_1 \left(f(x^{k_i}) - q_{x^{k_i}}(d^{k_i})\right) \\ &\geq \frac{\rho_1}{2} \cdot \|\nabla f(x^{k_i})\| \cdot \min\left\{\overline{\Delta}_{k_i}, \frac{\|\nabla f(x^{k_i})\|}{\|\nabla^2 f(x^{k_i})\|}\right\} \\ &\geq \frac{\rho_1}{2} \cdot \alpha \min\left\{\overline{\Delta}_{k_i}, \frac{\alpha}{\beta}\right\}. \end{aligned} \tag{6.1.5}$$

Die Folge $\{f(x^k)\}$ fällt monoton und es ist $\lim_{i\to\infty} f(x^{k_i}) = f(x^*)$. Damit folgt

$$\lim_{i\to\infty} f(x^{k_i}) - f(x^{k_i+1}) = 0,$$

so dass aus (6.1.5) die Beziehung $\lim_{i\to\infty} \overline{\Delta}_{k_i} = 0$ folgt, im Widerspruch zu Korollar 6.1.4. □

Das vorangehende Resultat kann im folgenden Sinne zu einem echten Konvergenzresultat verbessert werden.

6.1.6 Satz
Sei $f : \mathbb{R}^n \to \mathbb{R}, f \in \mathcal{C}^2(\mathbb{R}^n)$. Die Folge $\{x^k\}$ sei durch das Trust-Region Newton-Verfahren (Algorithmus 6.1.1) erzeugt mit $\nabla f(x^k) \neq 0$ für alle k. Es sei x^* ein Häufungspunkt der Folge $\{x^k\}$ mit $\nabla^2 f(x^*)$ spd. Dann gilt

(i) $\lim_{k\to\infty} x^k = x^*$ mit $\nabla f(x^*) = 0$.

(ii) Es existiert k_0, so dass für alle $k \geq k_0$ die Repeat-Schleife nur einmal durchlaufen wird.

(iii) Es existiert $\overline{\Delta} > 0$ mit $\overline{\Delta}_k \geq \overline{\Delta}$ für alle k.

(iv) $\lim_{k\to\infty} x^k = x^*$ r-überlinear.

(v) Ist $\nabla^2 f(x)$ Lipschitz-stetig in x^*, so ist die Konvergenz sogar q-quadratisch.

Beweis: Zu (i): Die Argumentation ist ähnlich wie beim Newton-Verfahren aus Abschnitt 3.5. Wir bemerken zuerst, dass x^* isolierter Häufungspunkt ist, denn $\nabla^2 f(x^*)$ ist spd, und jeder weitere Häufungspunkt von $\{x^k\}$ ist nach Satz 6.1.5 ein stationärer Punkt. Es gelte

$$\lim_{i\to\infty} x^{k_i} = x^*.$$

Es existiert eine Zahl $\alpha > 0$, so dass für alle $i \geq i_0$ gilt

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

(s. Lemma 3.4.2). Damit haben wir

$$\begin{aligned} 0 &\geq q_{x^{k_i}}(d^{k_i}) - f(x^{k_i}) \\ &= \nabla f(x^{k_i}) d^{k_i} + \frac{1}{2}(d^{k_i})^T \nabla^2 f(x^{k_i}) d^{k_i} \\ &\geq -\|\nabla f(x^{k_i})\| \cdot \|d^{k_i}\| + \frac{\alpha}{2}\|d^{k_i}\|^2, \end{aligned}$$

und damit

$$\|d^{k_i}\| \leq \frac{2}{\alpha} \cdot \|\nabla f(x^{k_i})\|. \tag{6.1.6}$$

Nach Satz 6.1.5 gilt $\lim_{i\to\infty} \|\nabla f(x^{k_i})\| = 0$, so dass $\lim_{i\to\infty} d^{k_i} = 0$. Wegen

$$x^{k_i+1} - x^{k_i} = t_{k_i} d^{k_i} \text{ mit } |t_{k_i}| \leq 1$$

erhalten wir

$$\lim_{i\to\infty} x^{k_i+1} - x^{k_i} = 0$$

und damit nach Lemma 3.5.11 sogar

$$\lim_{k\to\infty} x^k = x^*.$$

Zu (ii): Wir halten zunächst fest, dass wegen $\lim_{k\to\infty} x^k = x^*$ jetzt sogar für alle k die (6.1.6) entsprechende Beziehung

$$\|d^k\| \leq \frac{2}{\alpha} \cdot \|\nabla f(x^k)\|$$

gilt und damit insbesondere

$$\lim_{k\to\infty} d^k = 0.$$

– Wir werden $r^k \to 1$ zeigen. Für alle großen k wird dann die Repeat-Schleife nur einmal durchlaufen, denn $\rho_1 < 1$. Es ist

$$r^k - 1 = \frac{f(x^k + d^k) - q_{x^k}(d^k)}{f(x^k) - q_{x^k}(d^k)}.$$

Nach (i) konvergiert die Folge $\{x^k\}$ gegen x^*. Also existiert $c > 0$ mit

$$\|\nabla^2 f(x^k)\| \leq c \text{ für alle } k.$$

Nach Lemma 6.1.2 haben wir deshalb für alle k

$$\begin{aligned} f(x^k) - q_{x^k}(d^k) &\geq \frac{1}{2}\|\nabla f(x^k)\| \min\left\{\Delta_k, \frac{\|\nabla f(x^k)\|}{\|\nabla^2 f(x^k)\|}\right\} \\ &\overset{(6.1.6)}{\geq} \frac{\alpha}{4}\|d^k\| \min\left\{\|d^k\|, \frac{\alpha}{2c}\|d^k\|\right\} \\ &= \kappa \cdot \|d^k\|^2 \end{aligned}$$

mit

$$\kappa = \frac{1}{4}\alpha \min\left\{1, \frac{\alpha}{2c}\right\}.$$

Um den Zähler in $r^k - 1$ abzuschätzen verwenden wir die wegen des Mittelwertsatzes gültige Darstellung

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

mit $\xi^k = x^k + \theta_k d^k$, $\theta_k \in (0,1)$. Damit ist dann

$$\begin{aligned} |f(x^k + d^k) - q_{x^k}(d^k)| &= \frac{1}{2} \cdot \left|(d^k)^T \left(\nabla^2 f(\xi^k) - \nabla^2 f(x^k)\right) d^k\right| \\ &\leq \frac{1}{2}\|d^k\|^2 \cdot \|\nabla^2 f(\xi^k) - \nabla^2 f(x^k)\|. \end{aligned}$$

Dies ergibt für $|r^k - 1|$ die Abschätzung

$$\begin{aligned} |r^k - 1| &= \left|\frac{f(x^k + d^k) - q_{x^k}(d^k)}{f(x^k) - q_{x^k}(d^k)}\right| \\ &\leq \frac{1}{2\kappa}\|\nabla^2 f(\xi^k) - \nabla^2 f(x^k)\| \\ &\to 0 \ (k \to \infty) \end{aligned}$$

Zu (iii): Nach Teil (ii) und den Zeilen 11 und 13 des Algorithmus folgt direkt $\Delta_k \geq \Delta_{min}$.
Zu (iv) und (v): Beide Teile folgen aus den bekannten Sätzen zum Newton-Verfahren (Sätze 3.2.4 und 3.2.7), wenn wir gezeigt haben, dass für große k Algorithmus 6.1.1 in das Newton-Verfahren übergeht. Dazu beachten wir: für k groß genug ist $\nabla^2 f(x^k)$ spd. Die Newton-Korrektur $n^k = -(\nabla^2 f(x^k))^{-1}\nabla f(x^k)^T$ ist dann globales Minimum von q_{x^k}. Nach Teil (ii) ist ist $\overline{\Delta}_k > \overline{\Delta} > 0$ für alle k. Wegen (i) gilt

$$\lim_{k\to\infty} n^k = 0.$$

Für k groß genug gilt also $\|n^k\| \leq \overline{\Delta} \leq \overline{\Delta}_k$, so dass Algorithmus 6.1.1 als Lösung des Trust-Region Teilproblems immer $d^k = n^k$ bestimmt. Es wird also tatsächlich das Newton-Verfahren durchgeführt. □

Abschnitt 6.2

Teilräume und das doppelte Hundebein

Im Trust-Region Newton-Verfahren (Algorithmus 6.1.1) ist die wiederholte Lösung des Trust-Region Teilproblems der vom Aufwand her bei weitem dominierende Faktor. Man versucht deshalb, mit approximativen Lösungen auszukommen, ohne dabei theoretische Resultate preiszugeben. Dies gelingt erstaunlich gut.
Wie immer sei $f : \mathbb{R}^n \to \mathbb{R}$, $f \in \mathcal{C}^2(\mathbb{R}^n)$ und für $x \in \mathbb{R}^n$ sei $q = q_x$ das quadratische Modell

$$q_x(d) = f(x) + \nabla f(x)d + \frac{1}{2}d^T\nabla^2 f(x)d.$$

6.2.1 Definition
Sei $\Delta > 0$ ein Radius für eine Trust Region. Dann ist der Cauchy-Punkt d_c von q gegeben als die Minimalstelle von

$$\varphi : [0,1] \to \mathbb{R}, \quad \varphi(t) = q_x\left(-\tfrac{t\cdot\Delta}{\|\nabla f(x)\|} \cdot \nabla f(x)^T\right).$$

Es ist also

$$d_c = -\frac{t_c}{\|\nabla f(x)^T\|}\nabla f(x) \text{ mit } \varphi'(t_c) = 0 \text{ oder } t_c = 1.$$

Der Cauchy-Punkt war als "Vergleichspunkt" im Beweis von Lemma 6.1.2 herangezogen worden. Deshalb gilt auch für den Cauchy-Punkt

6.2.2 Lemma
Sei $x \in \mathbb{R}^n, \Delta > 0$ und d_c der Cauchy-Punkt. Dann ist

$$f(x) - q_x(d_c) \geq \frac{\|\nabla f(x)\|}{2} \cdot \min\left\{\Delta, \frac{\|\nabla f(x)\|}{\|\nabla^2 f(x)\|}\right\}. \tag{6.2.1}$$

Diese Tatsache hat enorme Konsequenzen: Eine Untersuchung der Beweise zu Lemma 6.1.3, Korollar 6.1.4, Satz 6.1.5 und Satz 6.1.6 (i), (ii) und (iii) zeigt nämlich, dass dort *nie* verwendet wurde, dass d^k das Trust-Region Teilproblem löst. Es wurde vielmehr nur die (6.2.1) entsprechende Eigenschaft von d^k herangezogen.
Also gilt folgende wichtige Bemerkung

6.2.3 Bemerkung

Das Trust-Region Newton-Verfahren werde so modifiziert, dass statt der exakten Lösung d^k des Trust-Region Teilproblems jeweils nur eine approximative Lösung d^k bestimmt wird mit $q_{x^k}(d^k) \leq q_{x^k}(d_c^k)$, d_c Cauchy-Punkt zu x^k und Δ_k. Dann gelten die Konvergenzaussagen aus Satz 6.1.5 und Satz 6.1.6 (i), (ii), (iii) unverändert weiter.

In diesem Sinne könnte man also stets $d^k = d_c^k$ nehmen. Allerdings: Auch die Aussagen (iv) und (v) von Satz 6.1.6 sind sehr wesentlich, denn sie zeigen die schnelle (nämlich überlineare oder gar quadratische) Konvergenz. Eine Inspektion des Beweises zeigt, dass dies daran liegt, dass für große k im "exakten" Trust-Region Verfahren die Newton-Richtung

$$d^k = -(\nabla^2 f(x^k))^{-1} \nabla f(x^k)^T$$

erreicht wird. Für eine schnelle Konvergenz bei approximativer Bestimmung der Lösung des Trust-Region Teilproblems sollte man also dafür sorgen, dass auch dann die Newton-Richtung zur Verfügung steht. Wir besprechen hierfür jetzt zwei Möglichkeiten.

Teilraum Trust-Region Hier wird das Trust-Region Teilproblem auf einem Teilraum $V_x \subseteq \mathbb{R}^n$ gelöst, d.h. man sucht

$$\hat{d} \in \mathbb{R}^n \text{ mit } q_x(\hat{d}) = \min\{q_x(d) : \|d\| \leq \Delta,\, d \in V_x\}. \tag{6.2.2}$$

Nimmt man

$$V_x = \operatorname{span}\{\nabla f(x)^T, (\nabla^2 f(x))^{-1} \nabla f(x)^T\}, \tag{6.2.3}$$

so ist $d_c \in V_x$ und damit $q(\hat{d}) \leq q(d_c)$. Deshalb gelten die Konvergenzaussagen von Satz 6.1.5 und Satz 6.1.6 (i), (ii), (iii). Weil auch die Newton-Richtung in V_x liegt, gelten auch die Teile (iv) und (v) von Satz 6.1.6.
Diese Diskussion bleibt richtig, falls $V_x \supseteq \operatorname{span}\{\nabla f(x)^T, (\nabla^2 f(x))^{-1} \nabla f(x)^T\}$ gewählt wird. Man könnte als weiteren erzeugenden Vektor z.B. einen Eigenvektor zu einem negativen Eigenwert von $\nabla^2 f(x)$ nehmen, sofern ein solcher existiert.
Im Falle (6.2.3) reduziert sich die Minimierungsaufgabe (6.2.2) auf eine zweidimensionale Aufgabe. Sind die Spalten von $Q = [q_1|q_2] \in \mathbb{R}^{n\times 2}$ eine Orthonormalbasis für V_x, so ist

$$\begin{aligned} \widetilde{q} &: \mathbb{R}^2 \to \mathbb{R} \\ \widetilde{q}_x(s,t) &= f(x) + (\nabla f(x) Q) \cdot \begin{pmatrix} s \\ t \end{pmatrix} + \frac{1}{2} (s\ ,\ t)\, Q^T \nabla^2 f(x)^T Q \begin{pmatrix} s \\ t \end{pmatrix} \end{aligned}$$

zu minimieren unter der Nebenbedingung $s^2 + t^2 \leq \Delta^2$. Hierzu kann man direkt die stationären Punkte von $\widetilde{q}_x$ bestimmen und dann zusätzlich den Rand $s^2 + t^2 = \Delta^2$ z.B. mit der Parametrisierung $s = \sin\varphi, t = \cos\varphi$ untersuchen.

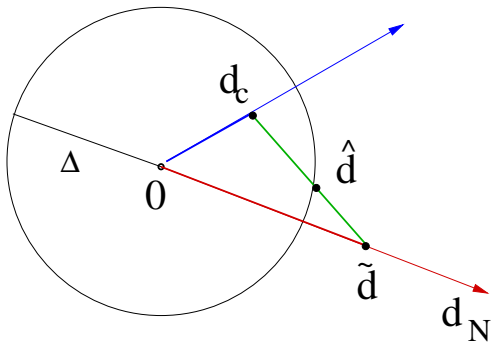

Abbildung 6.1: Illustration zum doppelten Hundebein

Das doppelte Hundebein Auch bei Einschränkung auf Teilräume ist es eigentlich unnötig, *exakte* Lösungen zu bestimmen. Wesentlich für die Übertragung der Konvergenzsätze 6.1.5 und 6.1.6 ist nur, dass der Abstieg mindestens so groß ist wie der mit Cauchy-Punkt, und dass im Falle $\nabla^2 f(x)$ spd die Newton-Richtung herangezogen wird. Der folgende Algorithmus beschreibt einen solchen Ansatz. Gegeben sind $x \in \mathbb{R}^n$ und ein Trust-Region Radius Δ.

6.2.4 Algorithmus (double dog leg step)

1: bestimme den Cauchy-Punkt d_c
2: **if** $\nabla^2 f(x)$ nicht spd **then**
3: setze $\hat{d} = d_c$
4: **else**
5: bestimme Newton-Punkt $d_N = -(\nabla^2 f(x))^{-1} \nabla f(x)^T$
6: **if** $\|d_N\| \leq \Delta$ **then**
7: setze $\hat{d} = d_N$
8: **else**
9: **if** $\|d_c\| = \Delta$ **then**
10: setze $\hat{d} = d_c$
11: **else**
12: setze $\widetilde{d}_N = \gamma d_N$ mit $\gamma \in \left(\frac{\Delta}{\|(\nabla^2 f(x)^{-1}) \nabla f(x)^T\|}, 1 \right)$
13: setze $\hat{d} = d_c + t(\widetilde{d}_N - d_c)$, t so, dass $\|\hat{d}\| = \Delta$.
14: **end if**
15: **end if**
16: **end if**

Die Zeilen 12 und 13 stellen das „doppelte Hundebein“ dar. Für $\gamma = 1$ hat man das „einfache“ Hundebein. Die Motivation dafür ist, dass q entlang der

Strecke von d_c nach $\widetilde{d}_N$ weiter abnimmt, wie unsere beiden nächsten Resultate zeigen.

Wir zeigen zuerst, dass (wie in der Abbildung verwendet) der Newton-Punkt weiter vom Zentrum entfernt liegt als der Cauchy-Punkt. Im Folgenden verwenden wir die Bezeichnungen

$$g^T = \nabla f(x), \;\; H = \nabla^2 f(x).$$

6.2.5 Lemma
Sei $\nabla^2 f(x)$ spd. Dann gilt $\|d_c\| \leq \|d_N\|$, wobei

$$d_c = -\frac{g^T g}{g^T H g} \cdot g, \;\; d_N = -H^{-1} g. \tag{6.2.4}$$

Beweis: Wir zeigen zunächst

$$\eta := \frac{(g^T g)^2}{g^T H g \cdot g^T H^{-1} g} \leq 1. \tag{6.2.5}$$

Sei dazu $v_1, \ldots, v_n$ eine Orthonormalbasis des $\mathbb{R}^n$ aus Eigenvektoren zu den Eigenwerten $\lambda_1, \ldots, \lambda_n > 0$ von H und sei $g = \sum_{i=1}^n \alpha_i v_i$. Dann gilt unter Verwendung von $(t + 1/t) \geq 2$ für $t > 0$

$$\begin{aligned} g^T H^{-1} g \cdot g^T H g &= \left(\sum_{i=1}^n \frac{1}{\lambda_i} \alpha_i^2 \right) \left(\sum_{j=1}^n \lambda_j \alpha_j^2 \right) \\ &= \sum_{i,j=1, i>j}^n \left(\frac{\lambda_i}{\lambda_j} + \frac{\lambda_j}{\lambda_i} \right) \alpha_i^2 \alpha_j^2 + \sum_{i=1}^n \alpha_i^4 \\ &\geq \sum_{i,j=1, i>j}^n 2\alpha_i^2 \alpha_j^2 + \sum_{i=1}^n \alpha_i^4 \\ &= (g^T g)^2, \end{aligned}$$

was (6.2.5) beweist. Damit erhalten wir nun

$$\begin{aligned} \|d_c\| &= \frac{(g^T g)^{3/2}}{g^T H g} \\ &\overset{CSU}{\leq} \frac{(g^T g)^{3/2}}{g^T H g} \cdot \frac{\|H^{-1} g\| \cdot \|g\|}{g^T H^{-1} g} \\ &= \eta \cdot \|d_N\| \\ &\overset{(6.2.5)}{\leq} \|d_N\|, \end{aligned}$$

was zu beweisen war. □

6.2.6 Lemma
Ist $\nabla^2 f(x)$ spd, so fällt q_x monoton entlang der Strecke von 0 nach d_c und auch entlang der Strecke von d_c nach $\widetilde{d}_N$.

Beweis: Das Abfallen auf der Strecke von 0 nach d_c ist bereits bekannt. Wir zeigen jetzt, dass wie behauptet $\varphi(t) = q_x(d_c + t(\widetilde{d}_N - d_c))$ auf dem Intervall $[0,1]$ monton fällt. Es ist (unter Verwendung von (6.2.4))

$$\begin{aligned}\varphi'(t) &= \nabla q_x\left(d_c + t(\gamma d_N - d_c)\right)(\gamma d_N - d_c)\\ &= g^T\left(\frac{g^Tg}{g^THg}\cdot g - \gamma H^{-1}g\right)\\ &\quad + (g + Hd_c)^T(\gamma d_N - d_c) + t\cdot(\gamma d_N - d_c)^T H(\gamma d_N - d_c).\end{aligned}$$

Also wächst φ' monoton in t, und wir müssen lediglich

$$(g + Hd_c)^T(\gamma d_N - d_c) + (\gamma d_N - d_c)^T H(\gamma d_N - d_c) \le 0 \tag{6.2.6}$$

zeigen. Dazu berechnen wir

$$\begin{aligned}&(g + Hd_c)^T(\gamma d_N - d_c) + (\gamma d_N - d_c)^T H(\gamma d_N - d_c)\\ &= \left(g + Hd_c + H(\gamma d_N - d_c)\right)^T(\gamma d_N - d_c)\\ &= (1-\gamma)g^T(\gamma d_N - d_c)\\ &= (1-\gamma)g^T\left(-\gamma H^{-1}g + \frac{g^Tg}{g^THg}g\right)\\ &= (1-\gamma)(\eta-\gamma)g^TH^{-1}g\\ &\le 0 \quad \text{für } \eta \le \gamma \le 1.\end{aligned}$$

□

6.2.7 Bemerkung
Empfohlen wird z.B. (s. Buch von Dennis und Schnabel)

$$\gamma = 0.8\eta + 0.2.$$

Abschnitt 6.3

Analyse des Trust-Region Teilproblems

Wir starten jetzt die Betrachtungen zur exakten Lösung des Trust-Region Teilproblems. Es folgen deshalb zunächst ein Abschnitt zur Analyse des Trust-Region-Teilproblems und dann ein Abschnitt zu numerischen Methoden zu dessen Lösung.
Das Trust-Region Teilproblem ist die Aufgabe

$$(6.3.1) \qquad \begin{cases} \text{gegeben:} & H \in \mathbb{R}^{n\times n}, \text{ symmetrisch} \\ & g \in \mathbb{R}^n, \\ & \Delta > 0, \\ & q(d) = g^T d + \frac{1}{2} d^T H d \\ \text{gesucht:} & d^* \text{ mit } q(d^*) \leq q(d) \\ & \text{für alle } d \text{ mit } \|d\| \leq \Delta. \end{cases}$$

Das Trust-Region Teilproblem ist eine restringierte Minimierungsaufgabe; die Zielfunktion q ist nur dann konvex, wenn H spd. Bevor wir gleich in einem Satz die Lösungen von (6.3.1) charkterisieren werden, benötigen wir Vorbereitungen zu Halbräumen in $\mathbb{R}^n$.

6.3.1 Definition
Sei $d \in \mathbb{R}^n, d \neq 0$. Mit $\mathcal{H}(d)$ bezeichen wir den abgeschlossenen Halbraum

$$\mathcal{H}(d) = \{y \in \mathbb{R}^n : y^T d \geq 0\},$$

mit $\overset{\circ}{\mathcal{H}}(d)$ den offenen Teilraum

$$\overset{\circ}{\mathcal{H}}(d) = \{y \in \mathbb{R}^n : y^T d > 0\}.$$

6.3.2 Lemma
Seien $d, d' \in \mathbb{R}^n, d, d' \neq 0$. Dann gilt

(i) $\mathbb{R}^n = \mathcal{H}(d) \cup \mathcal{H}(-d)$.

(ii) $\overset{\circ}{\mathcal{H}}(d) = \overset{\circ}{\mathcal{H}}(d') \Longrightarrow d = \lambda d'$ mit $\lambda > 0$.

Beweis: Teil (i) ist trivial. Zum Beweis von (ii) merken wir zunächst an, dass aus Stetigkeitsgründen auch $\mathcal{H}(d) = \mathcal{H}(d')$ gilt. Sei nun $d' = \lambda d + e$ mit $e^T d = 0$. Also ist $e \in \mathcal{H}(d)$ und deshalb auch $e \in \mathcal{H}(d')$. Dabei ist sogar $e^T d' = 0$, denn wäre $e^T d' > 0$ so wäre auch $(e - td)^T d' \in \mathcal{H}(d')$ für $t > 0$ klein genug, was aber $e - td \notin \mathcal{H}(d)$ zur Folge hätte, denn $(e - td)^T d = -td^T d < 0$. Aus $e^T d = e^T d' = 0$ folgt wegen $d' = \lambda d + e$ nun sofort $e^T e = 0$, also $e = 0$. Hierin ist $\lambda > 0$, denn im Falle $\lambda < 0$ wäre $\mathcal{H}(d) = -\mathcal{H}(d') \neq \mathcal{H}(d)$. □

6.3.3 Satz
Der Vektor d^* löst das Trust-Region-Teilproblem (6.3.1) genau dann, wenn $\lambda^* \in \mathbb{R}$ existiert mit

(i) $\lambda^* \cdot (\Delta - \|d^*\|) = 0$, $\lambda^* \geq 0$, $\Delta - \|d^*\| \geq 0$,

(ii) $(H + 2\lambda^* I)d^* = -g$,

(iii) $(H + 2\lambda^* I)$ ist positiv semidefinit.

Beweis: „⇒“: Eine Lösung von (6.3.1) erfüllt natürlich $\Delta - \|d^*\| \geq 0$.
Fall 1: $\Delta - \|d^*\| > 0$. Dann nehmen wir $\lambda^* = 0$. Dann ist (i) erfüllt, und da d^* (unrestringierte) lokale Minimalstelle von q ist, auch (ii) und (iii).
Fall 2: $\Delta - \|d^*\| = 0$. Sei v ein Vektor mit $v^T d^* < 0$ und $t \in [0, t^*]$ mit $t^* = -(2v^T d^*)/\|v\|^2$. Dann ist

$$\|d^* + tv\|^2 = \|d^*\|^2 + 2t \cdot v^T d^* + t^2 \|v\|^2 \leq \|d^*\|^2.$$

Also gilt für diese t

$$0 \leq q(d^* + tv) - q(d^*) = t(g + Hd^*)^T v + \frac{t^2}{2} v^T H v, \tag{6.3.2}$$

und damit, nach Division durch t und $t \to 0$,

$$(g + Hd^*)^T v \geq 0.$$

Dies gilt für alle v mit $v^T d^* < 0$, d.h. wir haben $\overset{\circ}{\mathcal{H}}(-d^*) = \overset{\circ}{\mathcal{H}}(g + Hd^*)$. Nach Lemma 6.3.2 ist $(g + Hd^*) = -2\lambda^* d^*$ für ein $\lambda^* > 0$ (der Faktor 2 hat kosmetische Gründe für später). Hieraus folgt (ii). Aus (ii) und (6.3.2) folgt mit der speziellen Wahl $t = t^*$ außerdem

$$0 \leq (t^*)^2 \cdot v^T (H + 2\lambda^* I) v \text{ für alle } v \text{ mit } v^T d^* < 0.$$

Damit gilt sogar $v^T (H + 2\lambda^* I) v \geq 0$ für alle v mit $v^T d^* \neq 0$ (man nehme notfalls $-v$ statt v) und aus Stetigkeitsgründen schließlich für alle $v \in \mathbb{R}^n$.

„⇐“: Das Paar (d^*, λ^*) erfülle (i) bis (iii). Wir zeigen, dass d^* globale Minimalstelle ist. Sei dazu $d \in \mathbb{R}^n$ mit $\|d\| \leq \Delta$. Dann erhalten wir

$$\begin{aligned}
q(d) - q(d^*) &= (g + Hd^*)^T(d - d^*) + \frac{1}{2}(d - d^*)^T H(d - d^*) \\
&\overset{(ii)}{=} -2\lambda^* \cdot (d - d^*)^T d^* + \frac{1}{2}\underbrace{(d - d^*)^T(H + 2\lambda^* I)(d - d^*)}_{\geq 0 \text{ wegen } (iii)} \\
&\quad - \lambda^* \cdot \|d - d^*\|^2 \\
&\geq \lambda^* \cdot (\|d^*\|^2 - \|d\|^2) \qquad (6.3.3) \\
&\overset{(i)}{=} \lambda^* \cdot (\|d^*\|^2 - \|d\|^2 + \Delta^2 - \|d^*\|^2) \\
&= \lambda^* \cdot (\Delta^2 - \|d\|^2) \\
&\geq 0.
\end{aligned}$$

Also ist d^* tatsächlich globale Minimalstelle. □

6.3.4 Korollar
Ist $H + 2\lambda^* I$ in Satz 6.3.3 spd, so ist d^* eindeutig.

Beweis: Für jedes $d \neq d^*$ gilt in (6.3.3) die strikte Ungleichung, d.h. es ist $q(d) > q(d^*)$. □

6.3.5 Korollar
Ist d^* Lösung von (6.3.1), so sind äquivalent

(i) $q(d^*) = 0$

(ii) $g = 0$ und H ist positiv semidefinit.

Beweis: “(i) ⇒ (ii)“: Es ist $q(d^*) = 0 = q(0)$. Also ist auch 0 eine Lösung von (6.3.1) und deshalb $\lambda^* = 0$ wegen Satz 6.3.3 (i). Nach den Teilen (ii) und (iii) folgt dann auch $g = -(H) \cdot 0 = 0$ sowie H ist positiv semidefinit.
“(ii) ⇒ (i)“: Mit $\lambda^* = 0$ und $d^* = 0$ gelten (i) bis (iii) von Satz 6.3.3 und $q(d^*) = 0$. Für jede weitere Lösung $\widetilde{d}$ ist dann ebenfalls $q(\widetilde{d}) = q(d^*) = 0$. □

Zur Bestimmung von d^*, λ^* ist es nützlich, die sog. KKT (Karush-Kuhn-Tucker)-Punkte zu betrachten.

6.3.6 Definition
Ein Punkt $(d, \lambda) \in \mathbb{R}^n \times \mathbb{R}$ heißt *KKT-Punkt* für das Trust-Region Teilproblem (6.3.1), falls gilt

(i) $\lambda \cdot (\Delta - \|d\|) = 0$, $\lambda \geq 0$, $\|d\| \leq \Delta$,

(ii) $(H + \lambda I)d = -g$.

Der Parameter λ heißt dann *Lagrange-Multiplikator*.

Ein KKT-Punkt ist "fast" eine Minimalstelle; die Semidfinitheit von $H + \lambda I$ wird aber nicht gefordert. Algorithmen zur Lösung des Trust-Region Teilproblems haben die Tendenz, KKT-Punkte zu finden; ähnlich wie globale Verfahren evtl. nur stationäre Punkte finden. Es ist deshalb wichtig, KKT-Punkte genauer zu untersuchen.
KKT-Punkte mit gleichem λ haben gleiche Funktionswerte.

6.3.7 Lemma
Seien (d_1, λ) und (d_2, λ) zwei KKT-Punkte. Dann ist $q(d_1) = q(d_2)$.

Beweis: Es ist für $i = 1, 2$

$$q(d_i) = \frac{1}{2} g^T d_i - \lambda \|d_i\|^2$$

und damit wegen $\|d_1\| = \|d_2\| = \Delta$ (oder $\lambda^* = 0$)

$$\begin{aligned} q(d_1) = \frac{1}{2} g^T d_1 - \lambda \|d_1\|^2 &= -\frac{1}{2} d_2^T (H + 2\lambda I) d_1 - \lambda \Delta^2 \\ &= -\frac{1}{2} d_2^T g - \lambda \|d_s\|^2 \\ &= q(d_2). \end{aligned}$$

Dieselbe Darstellung gilt auch für $q(d_2)$. □

Tatsächlich gibt es nur wenige Lagrange-Multiplikatoren.

6.3.8 Satz
$H \in \mathbb{R}^{n \times n}$ sei symmetrisch, $m \leq n$ sei die Zahl der verschiedenen negativen Eigenwerte von H. Dann besitzen die KKT-Punkte des Trust-Region Teilproblems (6.3.1) höchstens $2m + 2$ verschiedene Lagrange-Multiplikatoren.

Beweis: Sei $H = VDV^T$ mit V orthogonal, $D = \operatorname{diag}(\delta_1, \ldots, \delta_n)$. Für einen KKT-Punkt (d, λ) gilt

$$(H + 2\lambda I)d = -g \Longleftrightarrow (D + 2\lambda I)V^T d = -V^T g.$$

Wir notieren $\overline{g}$ für $V^T g$ und $\overline{d}$ für $V^T d$. Wir erhalten so die n Gleichungen

$$(\delta_i + 2\lambda)\overline{d}_i = -\overline{g}_i, \quad i = 1, \ldots, n. \tag{6.3.4}$$

Unter allen KKT-Punkten betrachten wir zunächst einmal die, für welche der Lagrange-Multiplikator λ für ein i die Beziehung $\delta_i + 2\lambda = 0$ erfüllt.

Die Anzahl der verschiedenen solchen Multiplikatoren sei k, die Menge der Indizes i mit $\delta_i + 2\lambda = 0$ sei I. Es ist dann $|I| \geq k$, wobei ">" bei mehrfachen Eigenwerten δ_i auftritt.
Für $i \in I$ ist dann $\overline{g}_i = 0$ in (6.3.4). Für die KKT-Punkte mit $\delta_i + 2\lambda \neq 0$ für $i = 1, \ldots, n$ folgt wegen $\|d\|^2 = \|\overline{d}\|^2$ aus $\lambda(\|\overline{d}\|^2 - \Delta^2) = 0$ dann

$$\lambda s(\lambda) = 0 \tag{6.3.5}$$

mit

$$s(\lambda) = \sum_{i=1, i \notin I}^{n} \left(\frac{-\overline{g}_i}{(\delta_i + 2\lambda)} \right)^2 - \Delta^2.$$

Es sei $\{\delta_i : \overline{g}_i \neq 0\} = \{\widetilde{\delta}_j,\ j = 1, \ldots, n-k\} \subseteq \{\delta_i : i \notin I\}$ mit

$$\widetilde{\delta}_1 \leq \widetilde{\delta}_2 \leq \ldots \widetilde{\delta}_p < 0 \leq \widetilde{\delta}_{p+1} \leq \ldots \leq \widetilde{\delta}_{\widetilde{n}},\ \widetilde{n} \leq n-k,\ p \leq m-k.$$

Dann ist

$$s(\lambda) = \sum_{i=1}^{\widetilde{n}} \frac{\widetilde{g}_j^2}{(\widetilde{\delta}_i + 2\lambda)^2} - \Delta^2.$$

Auf jedem der Intervalle

$$\left(-\infty, -\frac{\widetilde{\delta}_{\widetilde{n}}}{2}\right), \left(-\frac{\widetilde{\delta}_{\widetilde{n}}}{2}, -\frac{\widetilde{\delta}_{\widetilde{n}-1}}{2}\right), \ldots, \left(-\frac{\widetilde{\delta}_2}{2}, -\frac{\widetilde{\delta}_1}{2}\right) \left(-\frac{\widetilde{\delta}_1}{2}, +\infty\right)$$

ist $s(\lambda)$ streng konvex, auf den Randintervallen $\left(-\infty, -\frac{\widetilde{\delta}_{\widetilde{n}}}{2}\right), \left(-\frac{\widetilde{\delta}_1}{2}, +\infty\right)$ außerdem streng monoton. Auf jeden Randintervall gibt es deshalb höchstens eine Nullstelle von $s(\lambda)$, auf den anderen höchstens 2. Nichtnegative Lösungen von $s(\lambda) = \Delta^2$ liegen in den Intervallen

$$\left(-\frac{\widetilde{\delta}_{p+1}}{2}, -\frac{\widetilde{\delta}_p}{2}\right), \left(-\frac{\widetilde{\delta}_p}{2}, -\frac{\widetilde{\delta}_{p-1}}{2}\right), \ldots, \left(-\frac{\widetilde{\delta}_1}{2}, +\infty\right),$$

von denen außer dem Randintervall höchstens $p \leq m-k$ weitere nicht leer sind. Es gibt also maximal $2p+1$ nichtnegative Lösungen von $s(\lambda) = \Delta^2$. Damit besitzt (6.3.5) außer $\lambda = 0$ höchstens noch $2(m-k)+1+k \leq 2m+1$ weitere nichtnegative Lösungen. □

6.3.9 Bemerkung
Im Falle, dass H negativ definit ist, gilt Satz 6.3.8 mit $2m+1$ statt $2m+2$.

Beweis: Übung. □

Aus einem KKT-Punkt, der noch keine globale Minimalstelle von (6.3.1) ist, kann man explizit einen neuen Punkt mit kleinerem Wert für q bestimmen. Dies formuliert der nächste Satz.

6.3.10 Satz
Sei (d^*, λ^*) ein KKT-Punkt für das Trust-Region Teilproblem (6.3.1), so dass d^* noch keine Lösung ist. Dann ist für den nachfolgend definierten Punkt $\widehat{d}$ sowohl $\|\widehat{d}\| \leq \Delta$ wie auch $q(\widehat{d}) < q(d^*)$. Dabei ist $\widehat{d}$ wie folgt definiert

(i) falls $g^T d^* > 0$: $\widehat{d} = -\frac{\Delta}{\|d^*\|} d^*$

(ii) falls $g^T d^* \leq 0$: Wähle $z \in \mathbb{R}^n$ mit $z^T(H + 2\lambda^* I)z < 0$ und $g^T z \leq 0$. (Ein solches z existiert, denn (d^*, λ^*) erfüllt (6.3.1) (iii) nicht, d. h. $H + 2\lambda^* I$ besitzt einen Eigenvektor z zu einem negativen Eigenwert; z oder $-z$ erfüllt auch $g^T z \leq 0$).

(a) falls $\|d^*\| < \Delta$: $\widehat{d} = d^* + \alpha z$, wobei α die betragsgrößere Lösung von

$$\|z\|^2 \alpha^2 + 2 z^T d^* \alpha + (\Delta^2 - \|d^*\|^2) = 0$$

ist.

b) falls $\|d^*\| = \Delta$ und $z^T d^* \neq 0$: $\widehat{d} = d^* - 2\frac{z^T d^*}{\|z\|^2} z$

c) falls $\|d^*\| = \Delta$ und $z^T d^* = 0$: $\widehat{d} = d^* - 2\frac{\Delta^2}{\Delta^2 + \alpha^2 \|z\|^2}(d^* + \alpha z)$, wobei α so gewählt ist, dass

$$\omega(\alpha) = 2\left(\frac{\Delta^2}{\Delta^2 + \alpha^2 \|z\|^2}\right)^2 (-\alpha^2 |z^T(H + 2\lambda^* I)z| - 2\alpha g^T z + |g^T d^*|)$$

negativ (und möglichst klein) ist.

Beweis: Für einen KKT-Punkt (d^*, λ^*) gilt

$$(H + 2\lambda^* I) d^* = -g. \tag{6.3.6}$$

Für einen beliebigen Punkt y haben wir deshalb

$$\begin{aligned} q(d^* + y) &= -(d^*)^T (H + 2\lambda^* I)(d^* + y) + \frac{1}{2}(d^* + y)^T H (d^* + y) \\ &= q(d^*) - 2\lambda^* (d^*)^T y + \frac{1}{2} y^T H y \\ &= q(d^*) - \lambda^* (2(d^*)^T y + y^T y) + \frac{1}{2} y^T (H + 2\lambda^* I) y. \end{aligned} \tag{6.3.7}$$

Zu (i): $\|\widehat{d}\| = \Delta$ ist klar. Mit $\alpha = \frac{\Delta}{\|d^*\|} \geq 1$ gilt $\widehat{d} = d^* + (-\alpha - 1)d^*$ und damit nach (6.3.7) und (6.3.6)

$$\begin{aligned} q(\widehat{d}) &= q(d^*) - \lambda^*(-2(\alpha+1) + (\alpha+1)^2) \cdot \|d^*\|^2 - \frac{1}{2}(\alpha+1)^2 g^T d^* \\ &= q(d^*) - \lambda^*(\alpha^2 - 1) \cdot \|d^*\|^2 - \frac{1}{2}(\alpha+1)^2 g^T d^* \\ &< q(d^*). \end{aligned}$$

Zu (ii) a): In diesem Fall ist $\lambda^* = 0$. Für $\hat{d}_\alpha = d^* + \alpha z$ gilt dann nach (6.3.7)

$$q(\hat{d}_\alpha) = q(d^*) + \frac{\alpha^2}{2} z^T H z < q(d^*).$$

Für den angegeben Wert von α ist dann α^2 maximal unter allen $\hat{d}_\alpha$ mit $\|\hat{d}_\alpha\| \leq \Delta$ (und es ist sogar $\|\hat{d}_\alpha\| = \Delta$).
Zu (ii) b): s. Übung.
Zu (ii) c): Für beliebiges $\alpha \in \mathbb{R}$ sei $d_\alpha = d^* + \alpha z$. Dann ist

$$\|d_\alpha\|^2 = \Delta^2 + \alpha^2 \|z\|^2 \quad \text{und} \quad d_\alpha^T d^* = \|d^*\|^2 = \Delta^2.$$

Wir setzen

$$\hat{d}_\alpha = d^* - 2\frac{d_\alpha^T d^*}{\|d_\alpha\|^2} d_\alpha = d^* - 2\frac{\Delta^2}{\Delta^2 + \alpha^2\|z\|^2} \cdot d_\alpha.$$

Dann ist $\|\hat{d}_\alpha\| = \Delta$ und nach (6.3.7) gilt

$$q(\hat{d}_\alpha) = q(d^*) + \frac{1}{2} \cdot \left(2\frac{\Delta^2}{\Delta^2 + \alpha^2\|z\|^2}\right)^2 d_\alpha^T (H + 2\lambda^* I) d_\alpha.$$

Für den letzten Term ergibt sich auf Grund der KKT-Bedingungen

$$\begin{aligned} d_\alpha^T (H + 2\lambda^* I) d_\alpha &= d^*(H + 2\lambda^* I)d^* + \alpha^2 z^T (H + 2\lambda^* I) z \\ &\quad + 2\alpha \cdot (d^*)^T (H + 2\lambda^* I) z \\ &= -\alpha^2 \left| z^T (H + 2\lambda^* I) z \right| - 2\alpha g^T z + |g^T d^*|, \end{aligned}$$

d.h. wir haben

$$q(\hat{d}_\alpha) = q(d^*) + \omega(\alpha)$$

mit der angegebenen Funktion ω, und $\omega(\alpha) < 0$ für alle α außerhalb eines Intervalls. □

Abschnitt 6.4

Penalty-Funktionen für das Teilproblem

Dieser Abschnitt sollte wohl eher weggelassen werden ...
Wir wollen das globale Minimum des Trust-Region Teilproblems algorithmisch finden. Dazu stellen wir mit Hilfe einer sog. Penalty-Funktion zunächst eine äquivalente Formulierung als *unrestringiertes* Problem vor, worauf wir dann ein Newton-ähnliches Verfahren anwenden.
Wir erinnern an das Trust-Region Teilproblem als die Aufgabe

$$(6.4.1) \qquad \begin{cases} \text{gegeben:} & H \in \mathbb{R}^{n\times n}, \text{ symmetrisch} \\ & g \in \mathbb{R}^n, \\ & \Delta > 0, \\ & q(d) = g^T d + \frac{1}{2} d^T H d \\ \text{gesucht:} & d^* \text{ mit } q(d^*) \leq q(d) \\ & \text{für alle } d \text{ mit } \|d\| \leq \Delta. \end{cases}$$

6.4.1 Definition
Die zu (6.4.1) gehörige *Multiplikator-Funktion* $\lambda : \mathbb{R}^n \to \mathbb{R}$ ist gegeben als

$$\lambda(d) = -\frac{1}{2\Delta^2} \cdot (d^T H d + g^T d).$$

6.4.2 Lemma
Für die Multiplikator-Funktion λ gilt

(i) λ ist differenzierbar mit

$$\nabla\lambda(d) = -\frac{1}{2\Delta^2} \cdot (2d^T H + g^T).$$

(ii) Für jeden KKT-Punkt (d^*, λ^*) des Trust-Region Teilproblems (6.4.1) gilt

$$\lambda(d^*) = \lambda^*.$$

Beweis: Teil (i) ist bekannt.
Ein KKT-Punkt (d^*, λ^*) erfüllt $(H + 2\lambda^* I)d^* = -g^T$. Hieraus folgt nach Multiplikation mit $(d^*)^T$

$$\lambda(d^*) = -\frac{1}{2\Delta^2} \cdot (-2\lambda^*)\|d^*\|^2.$$

Ist $\lambda^* \neq 0$, so gilt $\|d^*\| = \Delta$. Dann ist also $\lambda(d^*) = \lambda^*$. Andernfalls ist $\lambda(d^*) = 0 = \lambda^*$. Dies beweist (ii). □

Teil (ii) erklärt die Namensgebung.
Mit Hilfe der Multiplikatorfunktion definieren wir nun eine Familie von Penalty-Funktionen. Penalty-Funktionen p_α sind so gestaltet, dass sie auf dem zulässigen Bereich ($\|d\| \leq \Delta, \lambda > 0$) gut mit der minimierenden Funktion übereinstimmen und im Grenzwert $\alpha \to 0$ sogar mit ihr identisch sind. Außerhalb des zulässigen Bereichs wachsen Penalty-Funktionen schnell an. Die Idee ist es damit, die restringierte Minimierungsaufgabe durch eine Folge unrestringierter Aufgaben mit den Penalty-Funktionen zu ersetzen.
Wir werden im Verlaufe dieses Abschnitt sehen, dass für das Trust-Region Teilproblem sogar eine einzige Penalty-Funktion ausreicht. Trotzdem starten wir mit einer ganzen Familie und passen erst später den Paramater α geeignet an. Zunächst bringen wir die Beschreibung des zulässigen Bereichs zusammen mit Teil (i) der KKT-Bedingungen geeignet „in einer Funktion unter“.

6.4.3 Lemma
Für $\alpha > 0$ sind äquivalent:

(i) $\|d^*\| \leq \Delta$, $\lambda(d^*) \geq 0$, $\lambda(d^*)(\Delta - \|d^*\|) = 0$

(ii) $\max\{\|d^*\|^2 - \Delta^2, -\frac{\alpha}{2}\lambda(d^*)\} = 0$

(iii) $\max\{0, \frac{2}{\alpha}(\|d^*\|^2 - \Delta^2) + \lambda(d^*)\} = \lambda(d^*)$

Beweis: Übung. □

6.4.4 Definition
Sei $\alpha > 0$. Die *Penalty-Funktion* $p_\alpha : \mathbb{R}^n \to \mathbb{R}$ für (6.4.1) ist gegeben durch

$$p_\alpha(d) \quad = \quad q(d) + \frac{\alpha}{4}\left(\left[\max\{0, \tfrac{2}{\alpha}(\|d\|^2 - \Delta^2) + \lambda(d)\}\right]^2 - \lambda^2(d)\right).$$

Die Abbildungen 6.2 und 6.3 zeigen Penalty-Funktionen für zwei typische Situationen im Eindimensionalen.
Einen Eidruck im Fall höherer Dimension vermitteln die Abbildungen 6.4 und 6.5.
Eine alternative Darstellung für p_α wird nachher nützlich sein.

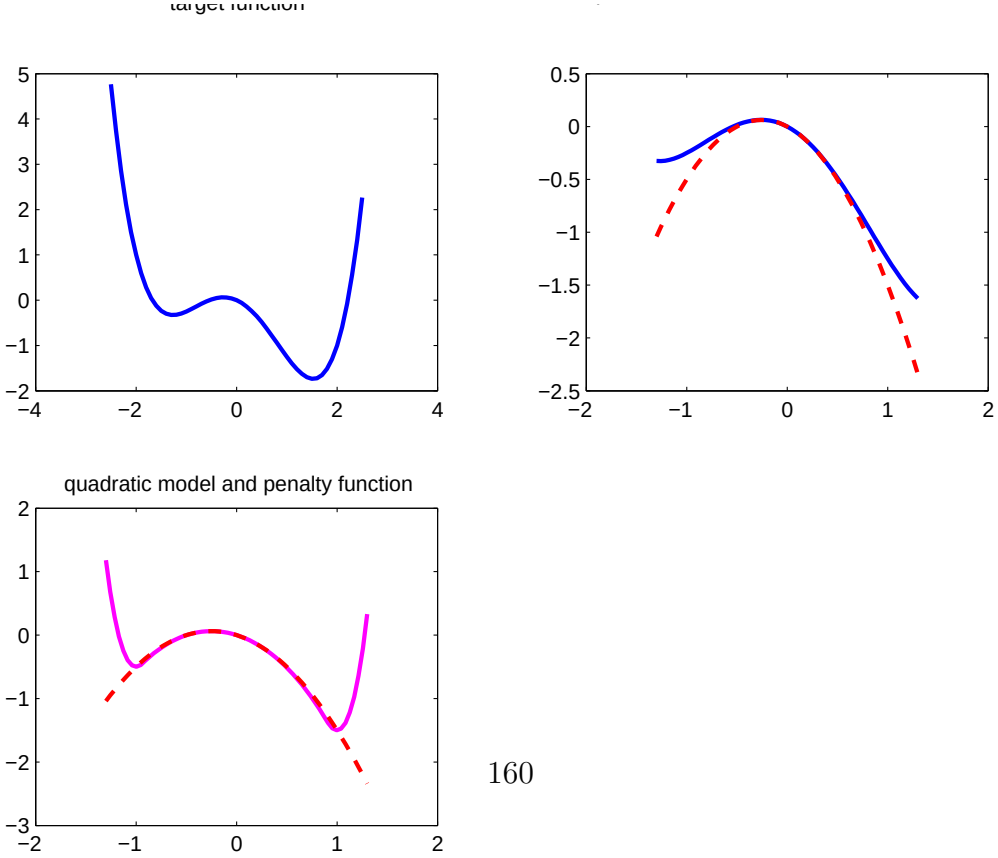

Abbildung 6.2: Quadratisches Modell und Penalty-Funktion ($\Delta = 1$). Das quadratische Modell ist nicht positiv definit

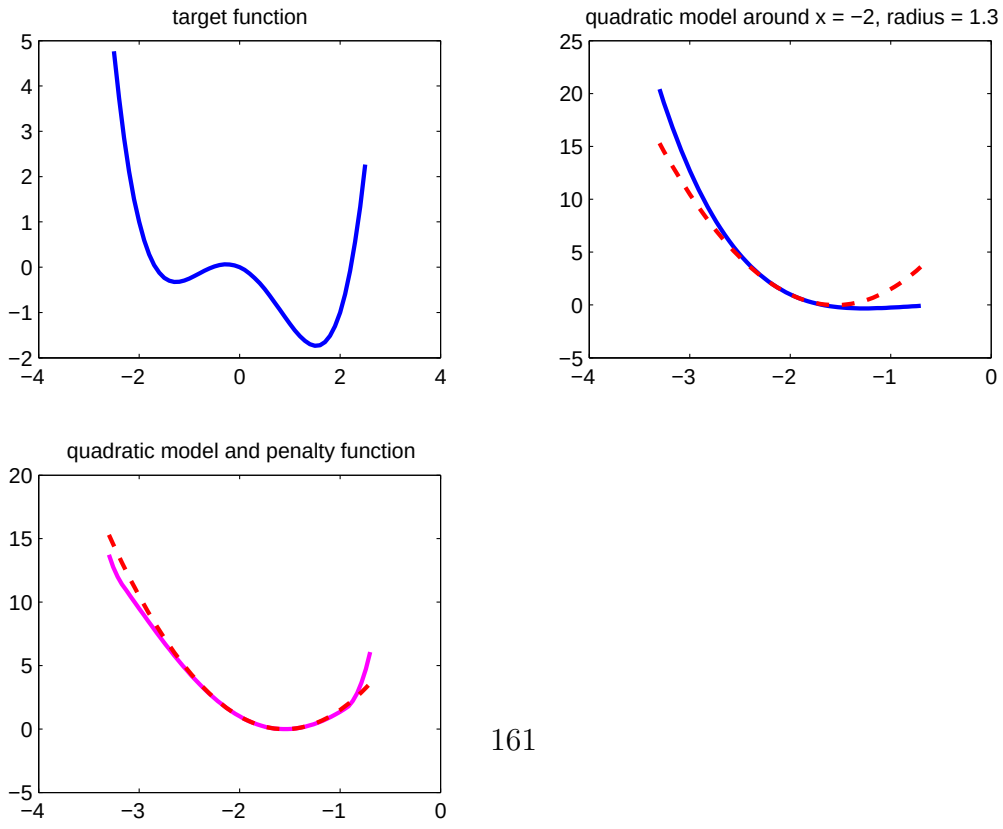

Abbildung 6.3: Quadratisches Modell und Penalty-Funktion ($\Delta = 1$). Das quadratische Modell ist positiv definit

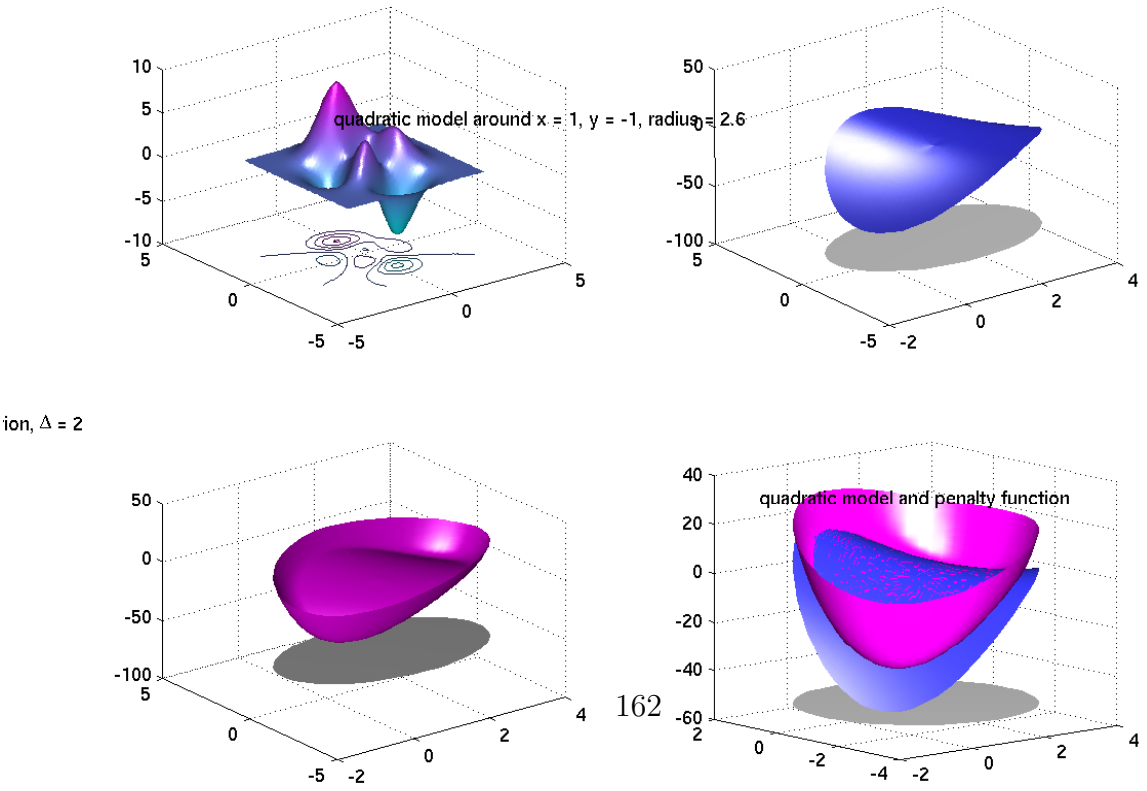

Abbildung 6.4: Quadratisches Modell und Penalty-Funktion ($\Delta = 1$). Das quadratische Modell ist nicht positiv definit

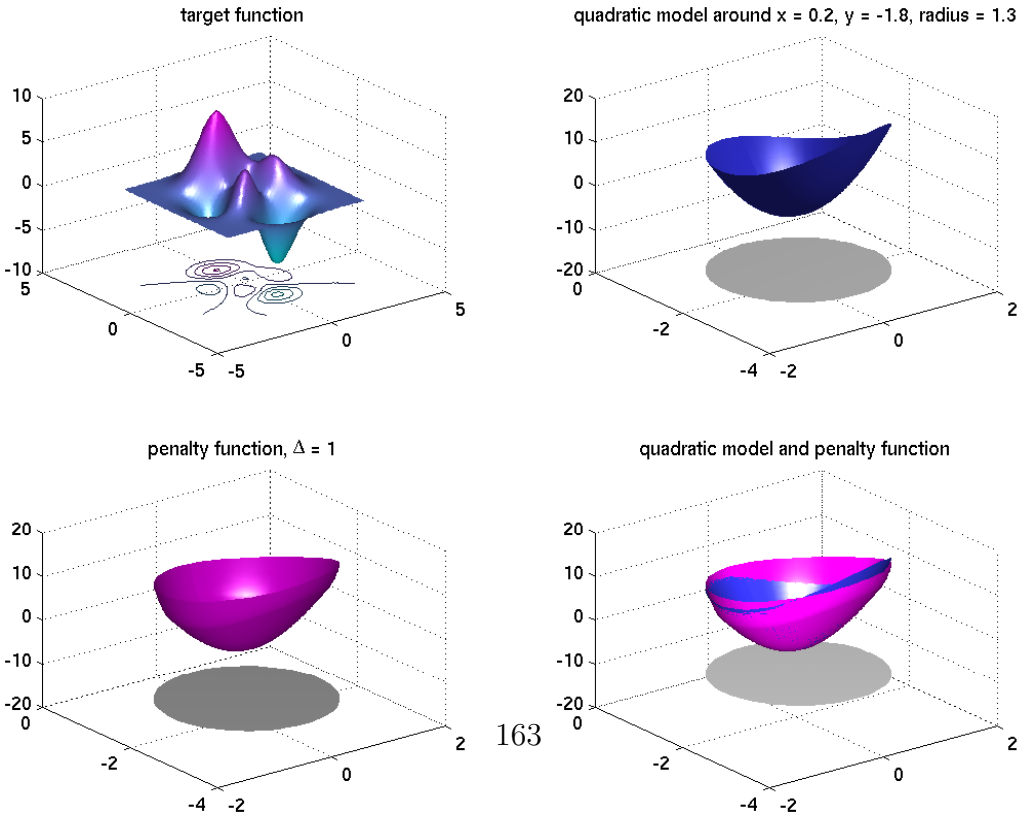

Abbildung 6.5: Quadratisches Modell und Penalty-Funktion ($\Delta = 1$). Das quadratische Modell ist positiv definit

6.4.5 Lemma

(i) Es ist

$$\begin{aligned} p_\alpha(d) &= q(d) + \lambda(d)\max\{\|d\|^2 - \Delta^2, -\frac{\alpha}{2}\lambda(d)\} \\ &\quad + \frac{1}{\alpha}\left[\max\{\|d\|^2 - \Delta^2, -\frac{\alpha}{2}\lambda(d)\}\right]^2 \end{aligned} \tag{6.4.2}$$

(ii) Die Funktionen p_α sind differenzierbar und es ist

$$\begin{aligned} (\nabla p_\alpha(d))^T &= Hd + g - \frac{\alpha}{2}\lambda(d)(\nabla\lambda(d))^T \\ &\quad + \frac{\alpha}{2}\cdot\max\{0, \tfrac{2}{\alpha}(\|d\|^2 - \Delta^2) + \lambda(d)\}\cdot\left(\frac{4}{\alpha}d + (\nabla\lambda(d))^T\right) \\ &= (\nabla q(d))^T + 2\lambda(d)\cdot d \\ &\quad + \left(\nabla\lambda(d)^T + \frac{4}{\alpha}d\right)\cdot\max\{\|d\|^2 - \Delta^2, -\frac{\alpha}{2}\lambda(d)\}. \end{aligned}$$

Beweis: Übung. □

6.4.6 Lemma

Sei $\alpha \in \left(0, \frac{2\Delta^2}{\|H\|}\right)$. Dann gilt

(i) Für $d \in \mathbb{R}^n$ mit $\|d\| \le \Delta$ gilt

$$p_\alpha(d) \le q(d).$$

(ii) Für jedes $c \in \mathbb{R}$ sind die Levelmengen

$$\mathcal{L}_c = \{d \in \mathbb{R}^n : p_\alpha(d) \le c\}$$

kompakt.

(iii) Die Penalty-Funktion p_α besitzt mindestens ein globales Minimum.

Beweis: Zu (i): Wir müssen

$$\left[\max\{0, \tfrac{2}{\alpha}(\|d\|^2 - \Delta^2) + \lambda(d)\}\right]^2 - \lambda^2(d) \le 0 \tag{6.4.3}$$

nachweisen.

Fall 1: Es ist

$$\frac{2}{\alpha}(\|d\|^2 - \Delta^2) + \lambda(d) \le 0 \text{ und } \|d\| \le \Delta.$$

Daraus folgt (6.4.3) direkt.
Fall 2: Es ist

$$\frac{2}{\alpha}(\|d\|^2 - \Delta^2) + \lambda(d) > 0 \quad \text{und } \|d\| \leq \Delta.$$

Dann haben wir

$$\begin{aligned}
&\left[\tfrac{2}{\alpha}(\|d\|^2 - \Delta^2) + \lambda(d)\right]^2 - \lambda^2(d) \\
&= \frac{4}{\alpha}\underbrace{(\|d\|^2 - \Delta^2)}_{\leq 0}\left(\frac{1}{\alpha}\cdot\underbrace{(\|d\|^2 - \Delta^2)}_{\leq 0} + \lambda(d)\right) \\
&\leq \frac{4}{\alpha}\cdot(\|d\|^2 - \Delta^2)\underbrace{\left(\frac{2}{\alpha}\cdot(\|d\|^2 - \Delta^2) + \lambda(d)\right)}_{>0 \text{ wg. Fallunterscheidung}} \\
&\leq 0,
\end{aligned}$$

also (6.4.3).
Zu (ii): Hier müssen wir etwas rechnen. Angenommen, für ein festes $c \in \mathbb{R}$ existiert eine unbeschränkte Folge $\{d^k\} \subseteq \mathcal{L}_c$. Auf Grund der Voraussetzung an α gilt

$$\frac{2}{\alpha} - \frac{1}{2\Delta^2}\cdot\|H\| > 0.$$

Aus

$$\frac{2}{\alpha}(\|d\|^2 - \Delta^2) + \lambda(d) \geq \left(\frac{2}{\alpha} - \frac{1}{2\Delta^2}\|H\|\right)\cdot\|d\|^2 - \frac{1}{2\Delta^2}\|g\|\cdot\|d\| - \frac{2\Delta^2}{\alpha}$$

folgt, dass für genügend großes k die Beziehung

$$\frac{2}{\alpha}\cdot(\|d^k\|^2 - \Delta^2) + \lambda(d^k) > 0$$

gilt. Für diese k gilt also

$$\begin{aligned}
p_\alpha(d^k) &= q(d^k) + \frac{\alpha}{4}\left(\left[\tfrac{2}{\alpha}(\|d^k\|^2 - \Delta^2) + \lambda(d^k)\right]^2 - \lambda^2(d^k)\right) \\
&= (d^k)^T H d^k + \frac{3}{2}g^T d^k + \frac{1}{\alpha}\|d^k\|^4 - \frac{2\Delta^2}{\alpha}\|d^k\|^2 + \frac{1}{\alpha}\Delta^4 \\
&\quad - \frac{1}{2\Delta^2}(d^k)^T H d^k\|d^k\|^2 - \frac{1}{2\Delta^2}g^T d^k\|d^k\|^2 \\
&\geq -\|H\|\cdot\|d^k\|^2 - \frac{3}{2}\|g\|\cdot\|d^k\| + \frac{1}{\alpha}\cdot\|d^k\|^4 - \frac{2\Delta^2}{\alpha}\|d^k\|^2 + \frac{\Delta^4}{\alpha} \\
&\quad - \frac{1}{2\Delta^2}\|H\|\cdot\|d^k\|^4 - \frac{1}{2\Delta^2}\|g\|\cdot\|d^k\|^3
\end{aligned}$$

$$= \left(\frac{1}{\alpha} - \frac{1}{2\Delta^2}\|H\|\right)\|d^k\|^4 - \frac{\|g\|}{2\Delta^2}\|d^k\|^3 - \left(\frac{2\Delta^2}{\alpha} + \|H\|\right)\cdot\|d^k\|^2 - \frac{3}{2}\|g\|\cdot\|d^k\| + \frac{\Delta^4}{\alpha}.$$

Im letzten Ausdruck dominiert der Term $\|d^k\|^4$ mit Koeffizient $\left(\frac{1}{\alpha} - \frac{1}{2\Delta^2}\|H\|\right) > 0$. Es gilt also $\lim_{k\to\infty} p_\alpha(d^k) = \infty$, im Widerspruch zu $d^k \in \mathcal{L}_c$.
Zu (iii): Dies folgt sofort aus (ii), da p_α stetig ist. □

Mit den nun folgenden Resultaten zeigen wir, dass für kleine Werte von α die Penalty-Funktion P_α *exakt* ist, d.h. ihre Minimalstellen stimmen mit denen von q überein.

6.4.7 Satz
Es sei

$$\alpha \in \left(0, \frac{16\Delta^4}{\Delta^2(8\|H\| + 3) + 5\|g\|^2}\right). \tag{6.4.4}$$

(i) Genau dann ist $d^* \in \mathbb{R}^n$ stationärer Punkt von p_α, wenn $(d^*, \lambda(d^*))$ ein KKT-Punkt von (6.4.1) ist.

(ii) Ist d^* stationärer Punkt von p_α, so gilt $p_\alpha(d^*) = q(d^*)$.

Beweis: Zu (i):
„⇐“: Für einen KKT-Punkt (d^*, λ^*) gilt nach Lemma 6.4.3

$$\max\{\|d^*\|^2 - \Delta^2, -\frac{\alpha}{2}\lambda(d^*)\} = 0$$

sowie

$$(H + 2\lambda(d^*)I)d^* = -g.$$

Mit der in Lemma 6.4.5 angegebenen zweiten Darstellung für ∇p_α folgt damit

$$\nabla p_\alpha(d^*) = \nabla q(d^*) + 2\lambda(d^*)d^* = (d^*)^T(H + 2\lambda(d^*) + g)d^* = 0.$$

„⇒“: Diese Richtung ist relativ technisch; insbesondere benötigt man hier (6.4.4). Wir verzichten auf diesen Teil und verweisen auf die insgesamt 4 Seiten im Buch von Kanzow und Geiger (Lemmata 14.12 und 14.16).
Zu (ii): Nach (i) ist $(d^*, d(\lambda^*))$ auch KKT-Punkt von (6.4.1) und deshalb

$$\max\{\|d^*\|^2 - \Delta^2, -\frac{\alpha}{2}\lambda(d^*)\} = 0.$$

Aus der Darstellung (6.4.2) für p_α folgt daraus sofort $p_\alpha(d^*) = q(d^*)$. □

6.4.8 Satz
Für α gelte (6.4.4). Dann ist $d^* \in \mathbb{R}^n$ ein globales Minimum von p_α genau dann, wenn d^* globale Minimalstelle des Trust-Region Teilproblems (6.4.1).

Beweis: d^* sei globales Minimum von p_α. Dann ist d^* stationärer Punkt von p_α und nach Satz 6.4.7 auch von q. Sei $\hat{d}$ globale Minimalstelle des Trust-Region Teilproblems (6.4.1) und $(\hat{d}, \hat{\lambda})$ der zugehörige KKT-Punkt. Dann ist nach Lemma 6.4.2 $\hat{\lambda} = \lambda(\hat{d})$, also ist $\hat{d}$ nach Satz 6.4.7 auch stationärer Punkt von p_α. Es ist dann nach diesem Satz auch

$$p_\alpha(d^*) = q(d^*), \;\; p_\alpha(\hat{d}) = q(\hat{d})$$

mit

$$p_\alpha(d^*) \leq p_\alpha(\hat{d}), \;\; q(\hat{d}) \leq q(d^*).$$

Also gilt

$$p_\alpha(d^*) = p_\alpha(\hat{d}) = q(\hat{d}),$$

d.h. d^* ist auch globale Minimalstelle von q, $\hat{d}$ ist auch globale Minimalstelle von p_α. □

Für lokale Minimalstellen haben wir das folgende Resultat.

6.4.9 Satz
Für α aus dem Intervall aus (6.4.4) ist jede lokale Minimalstelle d^* von p_α auch eine lokale Minimalstelle des Trust-Region Teilproblems (6.4.1).

Beweis: Übung. □

Abschnitt 6.5

Ein Algorithmus für das Teilproblem

Jetzt gehen wir zum algorithmischen Teil über und entwickeln ein (globalisiertes) Verfahren zur Minimierung von p_α.

Problem: $p_\alpha \notin \mathcal{C}^2(\mathbb{R}^n)$. Deshalb ist das Newton-Verfahren (z.B.) nicht direkt anwendbar.
Alternative: Wir bestimmen die Lösungen von

$$F(d,\lambda) = \begin{pmatrix} (H+2\lambda I)d+g \\ \max\{\|d\|^2-\Delta^2, -\frac{\alpha}{2}\lambda\} \end{pmatrix} = 0, \tag{6.5.1}$$
$$F: \quad \mathbb{R}^n \times \mathbb{R} \to \mathbb{R}^n \times \mathbb{R},$$

und „globalisieren“ das Verfahren dadurch, dass in p_α ein ausreichender Abstieg verlangt wird.
Im Prinzip wenden wir auf (6.5.1) nun eine Variante des Newton-Verfahrens für nichtlineare Gleichungssysteme an. Im Falle $\|d\|^2 - \Delta^2 > -\frac{\alpha}{2}\lambda$ ist F differenzierbar mit

$$F'(d,\lambda) = \begin{pmatrix} H & 2d \\ 2d^T & 0 \end{pmatrix},$$

so dass sich als Newton-Richtung

$$\begin{pmatrix} z \\ \zeta \end{pmatrix} = -\begin{pmatrix} H & 2d \\ 2d^T & 0 \end{pmatrix}^{-1} \begin{pmatrix} (H+2\lambda I)d+g \\ \|d\|^2-\Delta^2 \end{pmatrix}$$

ergibt mit $z \in \mathbb{R}, \zeta \in \mathbb{R}$. Statt der üblichen Newton-Korrektur

$$d^+ = d+z, \ \lambda^+ = \lambda+\zeta$$

setzen wir jetzt die zusätzliche Forderung $\lambda = \lambda(d)$ mit ein. Ein Iterationsschritt lautet deshalb

$$\begin{cases} d^+ = d+tz, & t>0 \text{ geeignet} \\ \lambda^+ = \max\{0, \lambda(d^*)\} \end{cases}. \tag{6.5.2}$$

Im Falle $\|d\|^2 - \Delta^2 < -\frac{\alpha}{2}\lambda$ ist die Nebenbedingung des Trust-Region Teilproblems „inaktiv“. Deshalb ignorieren wir die zweite Zeile in (6.5.1) und

erhalten die neue Iterierte wie in (6.5.2), wobei wir aber diesmal z als Lösung von

$$(H + 2\lambda I)z = -((H + 2\lambda I)d + g)$$

erhalten. Die „geeignete“ Schrittweite finden wir mir der Armijo-Regel. Unser Algorithmus lehnt sich außerdem an das globalisierte Newton-Verfahren 3.5.7 an.

6.5.1 Algorithmus (Lösung des Trust-Region Teilproblems)

wähle $\rho > 0, p > 2, \beta \in (0,1), \sigma \in (0, \frac{1}{2}), \varepsilon \geq 0, \alpha$ wie in (6.4.4)
wähle $d^0 \in \mathbb{R}^n$, setze $\lambda^0 = \max\{0, \lambda(d^*)\}$
for $k = 0, 1, \ldots$ **do**
  if $\|\nabla p_\alpha(d^k)\| \leq \varepsilon$ **then**
    if $H + 2\lambda^k I$ ist psd **then**
      STOP
    else
      bestimme d^{k+1} mit $\|d^{k+1}\| \leq \Delta$, $q(d^{k+1}) < q(d^k)$ {Satz 6.3.10}
    end if
  else
    if $\|d^k\|^2 - \Delta^2 \geq -\frac{\alpha}{2}\lambda^k$ **then**
      löse $\begin{pmatrix} H & 2d^k \\ 2(d^k)^T & 0 \end{pmatrix} \begin{pmatrix} z^k \\ \zeta^k \end{pmatrix} = - \begin{pmatrix} (H + 2\lambda^k I)d^k + g \\ \|d^k\|^2 - \Delta^2 \end{pmatrix}$
    else
      löse $(H + 2\lambda^k I)z^k = -((H + 2\lambda^k I)d^k + g)$
    end if
    if eines der Systeme nicht lösbar oder $\nabla p_\alpha(d^k)z^k > -\rho\|z^k\|^p$ **then**
      setze $z^k = -\nabla p_\alpha(d^k)$
    end if
    bestimme $t_k = \max\{\beta^\ell : \ell = 0, 1, \ldots : p_\alpha(d^k + \beta^\ell z^k) \leq$
      $p_\alpha(d^k) + \sigma \cdot \beta^\ell \nabla p_\alpha(d^k)z^k\}$
      {Armijo-Regel}
    setze $d^{k+1} = d^k + t_k z^k$, $\lambda^{k+1} = \max\{\lambda(d^{k+1}), 0\}$
  end if
end for

6.5.2 Bemerkung

(i) In der Praxis sollte man statt auf $\|\nabla p_\alpha(d^k)\| = 0$ auf $\|\nabla p_\alpha(d^k)\| \leq \varepsilon$ testen.

(ii) Ist $\|\nabla_{p_\alpha}(d^k)\| = 0$, so ist nach Satz 6.4.7 $(d^k, \lambda(d^k))$ ein KKT-Punkt des Trust-Region Teilproblems. Also ist $\lambda(d^k) \geq 0$ und damit $\lambda^k = \lambda(d^k)$ (λ^k wird im Algorithmus gesetzt.)

(iii) Terminiert der Algorithmus wegen $\|\nabla p_\alpha(d^k)\| = 0$ und $H + 2\lambda_k I$ psd, so ist nach Satz 6.3.3 d^k globale Minimalstelle des Trust-Region Teilproblems.

Es folgt der entscheidende Konvergenzsatz für Algorithmus 6.5.1.

6.5.3 Satz
In Algorithmus 6.5.1 sei $\varepsilon = 0$. Es sei$\nabla p_\alpha(d^k) \neq 0$ für $k = 0, 1, \ldots$. Dann besitzt die Folge $\{(d^k, \lambda^k)\}$ mindestens einen Häufungspunkt, und jeder Häufungspunkt (d^*, λ^*) ist ein KKT-Punkt des Trust-Region Teilproblems (6.3.1).

Beweis: Wir zeigen zuerst, dass $\{d^k\}$ beschränkt ist. Weil $\lambda(d)$ stetig ist, ist dann auch $\max\{0, \lambda(d)\}$ und damit $\{(d^k, \lambda^k)\}$ beschränkt, was dann die Existenz einer konvergenten Teilfolge garantiert.
Nach dem Algorithmus gilt

$$d^k \in \mathcal{L}(d^0) = \{d \in \mathbb{R}^n : p_\alpha(d) \leq p_\alpha(d^0)\},$$

welche nach Lemma 6.4.6 (ii) aber kompakt ist.
Nun sei (d^*, λ^*) ein Häufungspunkt von $\{(d^k, \lambda^k)\}$ und zur Vereinfachung der Notation sei $\{(d^k, \lambda^k)\}$ auch eine gegen (d^*, λ^*) konvergente Teilfolge.
Wir müssen $\nabla p_\alpha(d^*) = 0$ zeigen, denn nach Satz 6.4.7 ist dann $(d^*, \lambda(d^*))$ ein KKT-Punkt. Insbesondere ist also $\lambda(d^*) \geq 0$ und aus Stetigkeitsgründen dann $\lambda^* = \max\{0, \lambda(d^*)\} = \lambda(d^*)$.
Angenommen, es ist $\nabla p_\alpha(d^*) \neq 0$.
Fall 1: $z^k = -\nabla p_\alpha(d^k)$ für unendlich viele k. Dann ist nach Korollar 3.5.5 $\nabla p_\alpha(d^*) = 0$.
Fall 2: $z^k \neq -\nabla_{p_\alpha}(d^k)$ für $k \geq k_0$. Wir werden zeigen, dass es Konstanten $c_1, c_2 > 0$ gibt mit

$$c_1 \leq \|z^k\| \leq c_2 \text{ für alle } k \geq k_0.$$

Die Existenz von c_2 ergibt sich aus der durch den Algorithmus in diesem Fall gewährleisteten Beziehung

$$\nabla p_\alpha(d^k)^T z^k \leq -\rho \|z^k\|^p,$$

denn daraus folgt mit der CSU

$$\|\nabla p_\alpha(d^k)\| \geq \rho \|z^k\|^{p-1},$$

und $\|\nabla p_\alpha(d^k)\|$ ist beschränkt und $p - 1 > 0$.
Zum Nachweis der Existenz von c_1 nehmen wir für einen Widerspruchsbeweis an, es sei 0 ein Häufungspunkt von $\{z^k\}$, so dass wir $\lim\limits_{i\to\infty} z^{k_i} = 0$ notieren können.

Wir unterscheiden zwei Unterfälle.
Fall a: Für unendlich viele i ist $\|d^{k_i}\|^2 - \Delta^2 < -\frac{\alpha}{2}\lambda^{k_i}$. Wieder indizieren wir die entsprechende Teilfolge einfach mit k_i. Es gilt

$$\lambda^* = \max\{0, \lambda(d^*)\} = \max\{0, -\frac{1}{2\Delta^2}((d^*)^T H d^* + g^T d^*)\}.$$

Da (z^{k_i}, ζ^{k_i}) nach Zeile 12 von Algorithmus 6.5.1 bestimmt werden und die Matrizen $H + 2\lambda^{k_i} I$ beschränkt sind, folgt für $k_i \to \infty$ außerdem

$$(H + 2\lambda^*)d^* = -g, \tag{6.5.3}$$

woraus sich

$$(d^*)^T H d^* + g^T d^* = -2\lambda^* \|d^*\|^2$$

ergibt. Also haben wir

$$\lambda^* = \max\left\{0, \frac{\lambda^* \|d^*\|^2}{\Delta^2}\right\}$$

und deshalb $\lambda^* = 0$ oder $\|d^* = \|^2 \Delta^2$. Außerdem ist in Fall a

$$\|d^*\|^2 - \Delta^2 = \lim_{i\to\infty} \|d^{k_i}\|^2 - \Delta^2 \leq -\frac{\alpha}{2} \lim_{i\to\infty} \lambda^{k_i} = -\frac{\alpha}{2}\lambda^*,$$

woraus wegen $\lambda^* \geq 0$ schließlich $\|d^*\|^2 \leq \Delta^2$ folgt. Der Punkt (d^*, λ^*) ist wegen (6.5.3) also KKT-Punkt, nach Satz 6.4.7 also stationärer Punkt von p_α im Widerspruch zur Annahme $\nabla p_\alpha(d^*) \neq 0$.
Fall b: Für unendlich viele i ist $\|d^{k_i}\|^2 - \Delta^2 \geq -\frac{\alpha}{2}\lambda^{k_i}$. Die entsprechende Teilfolge wird wieder mit k_i indiziert. Es gilt dann

$$\begin{aligned} (H + 2\lambda^{k_i} I)z^{k_i} + 2d^{k_i}\zeta^{k_i} &= -(H + 2\lambda^{k_i} I)d^{k_i} - g, \\ 2(d^{k_i})^T z^{k_i} &= -(\|d^{k_i}\|^2 - \Delta^2). \end{aligned}$$

Wegen $\lim_{i\to\infty} z^{k_i} = 0$ und $\lim_{i\to\infty} \lambda^{k_i} = \lambda^* \geq 0$ folgt aus der zweiten Gleichung $\|d^*\| = \Delta$; aus der ersten folgt die Existenz von $\zeta^* = \lim_{i\to\infty} \zeta^{k_i}$ mit

$$2d^*\zeta^* = -(H + 2\lambda^* I)d^* - g. \tag{6.5.4}$$

Multiplikation mit $(d^*)^T$ liefert

$$2\|d^*\|^2 \cdot \zeta^* = -(d^*)^T(H + 2\lambda^* I)d^* - (d^*)^T g. \tag{6.5.5}$$

Im Fall $\lambda^* > 0$ ist $\lambda^* = -\frac{1}{2\Delta^2}((d^*)^T H d^* + g^T d^*)$, weshalb aus (6.5.5) durch Einsetzen folgt

$$2\|d^*\|^2 \cdot \zeta^* = 0, \text{ also } \zeta^* = 0.$$

Wegen (6.5.3) ist dann (d^*, λ^*) ein KKT-Punkt. Im Fall $\lambda^* = 0$ gilt

$$-g = (H + 2\zeta^*)d^*,$$

d.h. (d^*, ζ^*) ist KKT-Punkt und deshalb (Satz 6.4.7) $\nabla p_\alpha(d^*) = 0$, im Widerspruch zu Annahme $\nabla p_\alpha(d^*) \neq 0$.
Also existiert keine Teilfolge mit $\lim_{i\to\infty} z^{k_i} = 0$, d.h. es existiert $c_1 > 0$ mit $\|z^k\| \geq c_1$ für alle k.
Des Rest des Beweises geht jetzt vollständig analog zu dem beim Konvergenzsatz für das globalisierte Newton-Verfahren (Satz 3.5.9) mit $p_\alpha \widehat{=} f, z^k \widehat{=} d^k$, $d^k \widehat{=} x^k$. □

6.5.4 Bemerkung
Der soeben gezeigte Satz garantiert nicht die Konvergenz gegen eine globale Minimalstelle (bzw. dass eine solche ein Häufungspunkt ist). Dies liegt daran, dass wir $\nabla p_\alpha(d^k) \neq 0$ für alle k und $\varepsilon = 0$ vorausgesetzt haben, so dass die Zeilen 4 - 8 im Algorithmus nicht erreicht werden. Nimmt man jedoch $\varepsilon > 0$, so ergibt sich aus dem Beweis des Satzes, dass für ein k erreicht wird, dass $\|\nabla p_\alpha(d^k)\| \leq \varepsilon$ ist. In diesem Fall wird im Algorithmus d^k als lokale Minimalstelle angesehen; im Falle, dass d^k noch kein globales Minimum ist, wird d^k in Zeile 8 weiter verbessert, und wir können auf die nun erzeugte Folge wieder den Satz anwenden. Da es nach Satz 6.3.8 nur endlich viele lokale Minimalstellen gibt, wird man irgendwann beim globalen Minimum landen. Diese Überlegung ist allerdings nur heuristisch, denn weil d^k nur $\|\nabla p_\alpha(d^k)\| \leq \varepsilon$ und nicht $\nabla p_\alpha(d^k) = 0$ erfüllt, kann Satz 6.3.8 eigentlich nicht angewendet werden, d.h. es ist nicht klar, dass Zeile 8 im Algorithmus überhaupt funktioniert.