Nichtlineare Optimierung

Skript zur Vorlesung
im
Frühjahrsemester 2011

Helmut Harbrecht

Stand: 25. Mai 2011

Vorwort

Diese Mitschrift kann und soll nicht ganz den Wortlaut der Vorlesung wiedergeben. Sie soll als Lernhilfe dienen und das Nacharbeiten des Inhalts der Vorlesung erleichtern. Neben den unten genannten Büchern, diente mir auch das Vorlesungsskript *Numerik nichtlinearer Optimierung* von Gerhard Starke (Uni Hannover) als fruchtbare Quelle.

Literatur zur Vorlesung:

- C. Geiger und C. Kanzow: *Numerische Verfahren zur Lösung unrestringierter Optimierungsaufgaben*, Springer-Verlag
- C. Geiger und C. Kanzow: *Theorie und Numerik restringierter Optimierungsaufgaben*, Springer-Verlag
- F. Jarre und J. Stoer: *Optimierung*, Springer-Verlag

Inhaltsverzeichnis

1. Grundlagen

1.1 Einführung

Optimierungsaufgaben treten in zahlreichen Anwendungsproblemen in den Natur- und Ingenieurwissenschaften, der Wirtschaft oder der Industrie auf. Beispielsweise versuchen Transportunternehmen, die Fahrt- oder Flugkosten zu minimieren und dabei sicherzustellen, dass alle Aufträge ausgeführt werden. Ebenso führt die numerische Simulation vieler physikalischer Vorgänge in den Naturwissenschaften auf Optimierungsprobleme, da das zugrundeliegende mathematische Modell oftmals auf dem Prinzip der Energieminimierung beruht.

Unter einem endlichdimensionalen *Minimierungsproblem* wird die folgende Aufgabe verstanden: Gegeben sei eine *Zielfunktion* $f : \mathbb{R}^n \to \mathbb{R}$. Gesucht ist ein Punkt $\mathbf{x}^\star \in \mathbb{R}$, so dass

$$f(\mathbf{x}^\star) \leq f(\mathbf{x}) \quad \text{für alle } \mathbf{x} \in \mathbb{R}^n.$$

Dabei ist es ausreichend, sich nur mit Minimierungsproblemen zu beschäftigen, da ein Maximierungsproblem für f immer einem Minimierungsproblem für $-f$ entspricht.

Definition 1.1 Es sei $f : \mathbb{R}^n \to \mathbb{R}$. Ein Punkt $\mathbf{x}^\star \in \mathbb{R}^n$ heißt **globales Minimum**, falls gilt

$$f(\mathbf{x}^\star) \leq f(\mathbf{x}) \quad \text{für alle } \mathbf{x} \in \mathbb{R}^n.$$

Das Minimum ist ein **lokales Minimum**, wenn es eine Umgebung $U \subset \mathbb{R}^n$ von $\mathbf{x}^\star$ gibt, so dass

$$f(\mathbf{x}^\star) \leq f(\mathbf{x}) \quad \text{für alle } \mathbf{x} \in U.$$

Das Minimim heißt **strikt**, wenn im Fall $\mathbf{x} \neq \mathbf{x}^\star$ jeweils die strenge Ungleichung $f(\mathbf{x}^\star) < f(\mathbf{x})$ gilt.

In der Regel ist es mit vertretbarem Aufwand nur möglich, ein lokales Minimum von f in einer Umgebung eines Startwertes $\mathbf{x}_0$ zu bestimmen.

1.2 Optimalitätskriterien

Um ein lokales Minimum numerisch zu finden, versucht man iterativ die Gleichung $\nabla f(\mathbf{x}) = \mathbf{0}$ zu lösen.

Definition 1.2 Seien $U \subset \mathbb{R}^n$ eine offene Menge und $f : U \to \mathbb{R}$ eine stetig differenzierbare Funktion. Ein Punkt $\mathbf{x}^\star \in U$ heißt **stationärer Punkt**, falls gilt

$$\nabla f(\mathbf{x}^\star) = \mathbf{0}.$$

Wir wiederholen einige bekannte Eigenschaften lokaler Minima aus der Analysis:

Satz 1.3 (notwendige Bedingung 1. Ordnung) Ist $\mathbf{x}^\star$ ein lokales Minimum von f und ist f stetig differenzierbar in einer Umgebung von $\mathbf{x}^\star$, dann gilt $\nabla f(\mathbf{x}^\star) = \mathbf{0}$. Der Punkt $\mathbf{x}^\star$ ist also ein stationärer Punkt.

Satz 1.4 (notwendige Bedingung 2. Ordnung) Ist $\mathbf{x}^\star$ ein lokales Minimum von f und ist die Hesse-Matrix $\nabla^2 f$ stetig in einer Umgebung von $\mathbf{x}^\star$, dann gilt $\nabla f(\mathbf{x}^\star) = \mathbf{0}$ und $\nabla^2 f(\mathbf{x}^\star)$ ist eine positiv semidefinite Matrix.

Satz 1.5 (hinreichende Bedingung 2. Ordnung) Die Hesse-Matrix $\nabla^2 f$ sei stetig in einer Umgebung von $\mathbf{x}^\star$ mit $\nabla f(\mathbf{x}^\star) = \mathbf{0}$. Ist $\nabla^2 f(\mathbf{x}^\star)$ eine positiv definite Matrix, dann ist $\mathbf{x}^\star$ ein striktes lokales Minimum.

1.3 Konvexität

Wir wenden uns einem wichtigen und in der Praxis oft auftretenden Spezialfall zu, bei dem wir mit einem lokalen zugleich ein globales Minimum gefunden haben. Dazu sei angemerkt, dass eine Menge $D \subset \mathbb{R}^n$ konvex ist, falls aus $\mathbf{x}, \mathbf{y} \in D$ auch $\lambda\mathbf{x} + (1-\lambda)\mathbf{y} \in D$ folgt für alle $\lambda \in (0,1)$.

Definition 1.6 Es sei $D \subset \mathbb{R}^n$ eine konvexe Menge. Die Funktion $f : D \to \mathbb{R}$ heißt **konvex auf D**, wenn für alle $\lambda \in (0,1)$ und alle $\mathbf{x}, \mathbf{y} \in D$ gilt

$$f\big(\lambda\mathbf{x} + (1-\lambda)\mathbf{y}\big) \leq \lambda f(\mathbf{x}) + (1-\lambda) f(\mathbf{y}).$$

Gilt für $\mathbf{x} \neq \mathbf{y}$ sogar stets die strikte Ungleichung, dann heißt die Funktion **strikt konvex**. Gibt es ein $\mu > 0$, so dass

$$f\big(\lambda\mathbf{x} + (1-\lambda)\mathbf{y}\big) + \mu\lambda(1-\lambda)\|\mathbf{x}-\mathbf{y}\|_2^2 \leq \lambda f(\mathbf{x}) + (1-\lambda) f(\mathbf{y})$$

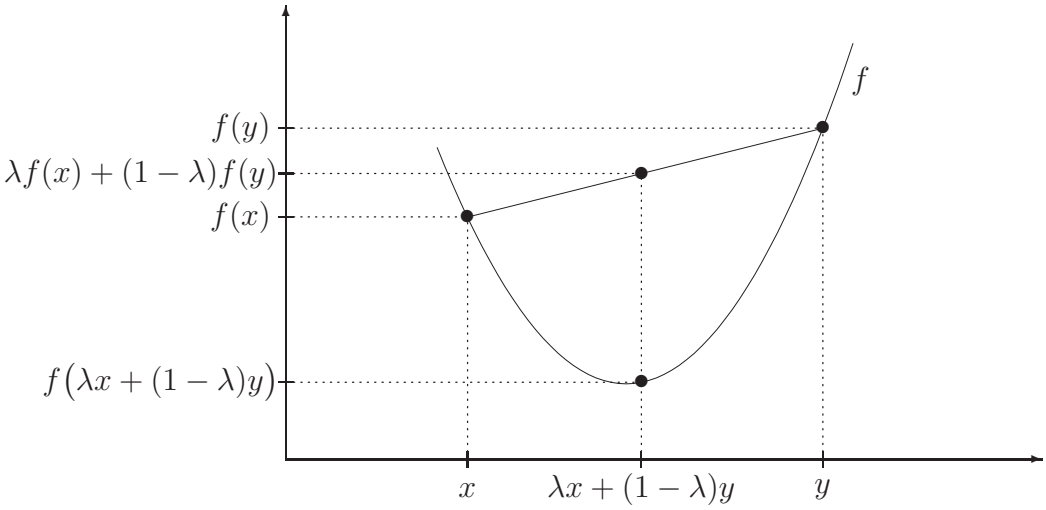
für alle $\lambda \in (0,1)$ und alle $\mathbf{x}, \mathbf{y} \in D$, dann heißt die Funktion f **gleichmäßig konvex**.

Beispiele 1.7

1. Die Gerade $f(x) := x$ ist konvex auf $\mathbb{R}$, aber nicht strikt konvex.

2. Die Exponentialfunktion $f(x) := \exp(x)$ ist strikt konvex auf $\mathbb{R}$, dort aber nicht gleichmäßig konvex.
3. Die Parabel $f(x) := x^2$ ist gleichmäßig konvex auf $\mathbb{R}$. Hingegen ist die sehr ähnlich aussehende Funktion $f(x) := x^4$ zwar strikt konvex auf $\mathbb{R}$, aber nicht gleichmäßig konvex.

△



Bei einer eindimensionalen konvexen Funktion liegt die Verbindungslinie zweier Punkte oberhalb des Graphen.

Bemerkung Sei $f : \mathbb{R}^n \to \mathbb{R}$ eine *quadratische Funktion*, das heißt

$$f(\mathbf{x}) = \frac{1}{2}\mathbf{x}^T\mathbf{A}\mathbf{x} + \mathbf{b}^T\mathbf{x} + c$$

mit einer symmetrischen Matrix $\mathbf{A} \in \mathbb{R}^{n\times n}$, $\mathbf{b} \in \mathbb{R}^n$ und $c \in \mathbb{R}$. Die Funktion f ist genau dann konvex, wenn $\mathbf{A}$ positiv semidefinit ist. Ist die Matrix $\mathbf{A}$ sogar positiv definit, so ist f sogar gleichmäßig konvex. △

Satz 1.8 Seien $D \subset \mathbb{R}^n$ eine offene und konvexe Menge und $f : D \to \mathbb{R}$ stetig differenzierbar. Die Funktion f ist genau dann konvex auf D, wenn für alle $\mathbf{x}, \mathbf{y} \in D$ gilt

$$f(\mathbf{x}) - f(\mathbf{y}) \geq \nabla f(\mathbf{y})^T(\mathbf{x} - \mathbf{y}). \tag{1.1}$$

Ist diese Ungleichung strikt für alle $\mathbf{x} \neq \mathbf{y}$, dann ist f sogar strikt konvex. Die Funktion f ist genau dann gleichmäßig konvex, wenn ein $\mu > 0$ existiert, so dass

$$f(\mathbf{x}) - f(\mathbf{y}) \geq \nabla f(\mathbf{y})^T(\mathbf{x} - \mathbf{y}) + \mu\|\mathbf{x} - \mathbf{y}\|_2^2 \tag{1.2}$$

für alle $\mathbf{x}, \mathbf{y} \in D$.

Beweis. Es gelte zuächst (1.2). Für $\mathbf{x}, \mathbf{y} \in D$ und beliebiges $\lambda \in (0, 1)$ ergibt sich dann mit $\mathbf{z} := \lambda\mathbf{x} + (1 - \lambda)\mathbf{y}$

$$f(\mathbf{x}) - f(\mathbf{z}) \geq \nabla f(\mathbf{z})^T(\mathbf{x} - \mathbf{z}) + \mu\|\mathbf{x} - \mathbf{z}\|_2^2,$$
$$f(\mathbf{y}) - f(\mathbf{z}) \geq \nabla f(\mathbf{z})^T(\mathbf{y} - \mathbf{z}) + \mu\|\mathbf{y} - \mathbf{z}\|_2^2.$$

Multipliziert man diese Gleichungen mit λ beziehungsweise $1-\lambda$ und addiert sie anschließend, dann folgt

$$\lambda f(\mathbf{x}) + (1-\lambda) f(\mathbf{y}) - f\big(\lambda\mathbf{x} + (1-\lambda)\mathbf{y}\big) \geq 2\mu\lambda(1-\lambda)\|\mathbf{x}-\mathbf{y}\|_2^2,$$

das heißt, f ist gleichmäßig konvex.
Sei f nun als gleichmäßig konvex auf D vorausgesetzt. Für alle $\mathbf{x}, \mathbf{y} \in D$ und $\lambda \in (0,1)$ gilt dann mit einem $\mu > 0$

$$\begin{aligned} f\big(\mathbf{y} + \lambda(\mathbf{x}-\mathbf{y})\big) &= f\big(\lambda\mathbf{x} + (1-\lambda)\mathbf{y}\big) \\ &\leq \lambda f(\mathbf{x}) + (1-\lambda) f(\mathbf{y}) - \mu\lambda(1-\lambda)\|\mathbf{x}-\mathbf{y}\|_2^2 \end{aligned}$$

und daher

$$\frac{f\big(\mathbf{y} + \lambda(\mathbf{x}-\mathbf{y})\big) - f(\mathbf{y})}{\lambda} \leq f(\mathbf{x}) - f(\mathbf{y}) - \mu(1-\lambda)\|\mathbf{x}-\mathbf{y}\|_2^2.$$

Aufgrund der stetigen Differenzierbarkeit von f folgt somit für $\lambda \to 0+$

$$\nabla f(\mathbf{y})^T(\mathbf{x}-\mathbf{y}) \leq f(\mathbf{x}) - f(\mathbf{y}) - \mu\|\mathbf{x}-\mathbf{y}\|_2^2,$$

dies bedeutet, es gilt (1.2). Da der soeben geführte Beweis auch im Fall $\mu = 0$ seine Gültigkeit behält, folgt die Äquivalenz von (1.1) zur Konvexität von f.
Es verbleibt zu zeigen, dass die strikte Konvexität von f die strikte Ungleichung

$$f(\mathbf{x}) - f(\mathbf{y}) > \nabla f(\mathbf{y})^T(\mathbf{x}-\mathbf{y})$$

für alle $\mathbf{x}, \mathbf{y} \in D$ mit $\mathbf{x} \neq \mathbf{y}$ impliziert. Als strikt konvexe Funktion ist f insbesondere konvex, das heißt, es gilt (1.1). Für

$$\mathbf{z} := \frac{1}{2}(\mathbf{x}+\mathbf{y}) = \frac{1}{2}\mathbf{x} + \left(1 - \frac{1}{2}\right)\mathbf{y}$$

ergibt sich daher

$$\nabla f(\mathbf{y})^T(\mathbf{x}-\mathbf{y}) = 2\nabla f(\mathbf{y})^T(\mathbf{z}-\mathbf{y}) \leq 2\big\{f(\mathbf{z}) - f(\mathbf{y})\big\}. \tag{1.3}$$

Ist $\mathbf{x} \neq \mathbf{y}$, dann folgt wegen der strikten Konvexität

$$f(\mathbf{z}) < \frac{1}{2}f(\mathbf{x}) + \frac{1}{2}f(\mathbf{y}).$$

Dies eingesetzt in (1.3) liefert die Behauptung

$$\nabla f(\mathbf{y})^T(\mathbf{x}-\mathbf{y}) < f(\mathbf{x}) - f(\mathbf{y}).$$

□

Satz 1.9 Die Funktion $f : D \subset \mathbb{R}^n \to \mathbb{R}$ sei konvex. Dann ist jedes lokale Minimum $\mathbf{x}^\star$ auch ein globales Minimum von f. Ist f zusätzlich differenzierbar, so ist jeder stationäre Punkt $\mathbf{x}^\star$ ein globales Minimum.

Beweis. Angenommen, der Punkt $\mathbf{x}^\star$ ist ein lokales, aber kein globales Minimum. Dann gibt es einen Punkt $\mathbf{y}^\star \in \mathbb{R}^n$ mit $f(\mathbf{y}^\star) < f(\mathbf{x}^\star)$. Für alle

$$\mathbf{x} = \lambda \mathbf{x}^\star + (1-\lambda)\mathbf{y}^\star, \quad \lambda \in (0,1) \tag{1.4}$$

gilt aufgrund der Konvexität

$$f(\mathbf{x}) \leq \lambda f(\mathbf{x}^\star) + (1-\lambda) f(\mathbf{y}^\star) < f(\mathbf{x}^\star).$$

Da in jeder Umgebung von $\mathbf{x}^\star$ Punkte der Form (1.4) liegen, steht dies im Widerspruch zur Annahme, dass $\mathbf{x}^\star$ ein lokales Minimum ist. Folglich ist jedes lokale Minimum auch ein globales Minimum.

Wir zeigen nun die zweite Aussage. Dazu sei f differenzierbar vorausgesetzt und $\mathbf{x}^\star$ ein stationärer Punkt. Wir führen den Beweis wieder per Widerspruch und nehmen an, dass $\mathbf{x}^\star$ kein lokales Minimum ist. Dann können wir ein $\mathbf{y}^\star$ wie oben wählen und erhalten aufgrund der Konvexität

$$\begin{aligned}
\nabla f(\mathbf{x}^\star)^T(\mathbf{y}^\star - \mathbf{x}^\star) &= \frac{\mathrm{d}}{\mathrm{d}s} f\big(\mathbf{x}^\star + t(\mathbf{y}^\star - \mathbf{x}^\star)\big)\Big|_{t=0} \\
&= \lim_{t \to 0+} \frac{f\big(\mathbf{x}^\star + t(\mathbf{y}^\star - \mathbf{x}^\star)\big) - f(\mathbf{x}^\star)}{t} \\
&\leq \lim_{t \to 0+} \frac{(1-t) f\big(\mathbf{x}^\star) + t f(\mathbf{y}^\star) - f(\mathbf{x}^\star)}{t} \\
&= f(\mathbf{y}^\star) - f(\mathbf{x}^\star) < 0.
\end{aligned}$$

Deshalb ist $\nabla f(\mathbf{x}^\star) \neq \mathbf{0}$ und folglich ist $\mathbf{x}^\star$ kein stationärer Punkt. □

2. Gradientenverfahren

2.1 Idealisierte Variante

Im folgenden setzen wir stets voraus, dass $f : D \subset \mathbb{R}^n \to \mathbb{R}$ stetig differenzierbar ist. Zunächst wollen wir das *Gradientenverfahren* betrachten, das auch *Verfahren des steilsten Abstiegs* genannt wird. Die Idee dabei ist, die Iterierte $\mathbf{x}_k$ in Richtung des Antigradienten $-\nabla f(\mathbf{x}_k)$ aufzudatieren

$$\mathbf{x}_{k+1} := \mathbf{x}_k - \alpha_k \nabla f(\mathbf{x}_k),$$

so dass $f(\mathbf{x}_{k+1}) < f(\mathbf{x}_k)$ ist. Dass dies im Fall $\nabla f(\mathbf{x}_k) \neq \mathbf{0}$ immer möglich ist, zeigt uns das nächste Lemma.

Lemma 2.1 Vorausgesetzt es ist $\nabla f(\mathbf{x}_k) \neq \mathbf{0}$, dann gibt es ein $\delta > 0$, so dass die Funktion

$$\varphi(\alpha) = f\big(\mathbf{x}_k - \alpha \nabla f(\mathbf{x}_k)\big)$$

für alle $0 \leq \alpha \leq \delta$ streng monoton fällt. Insbesondere gilt

$$f\big(\mathbf{x}_k - \delta \nabla f(\mathbf{x}_k)\big) < \varphi(0) = f(\mathbf{x}_k).$$

Beweis. Die Funktion φ ist stetig differenzierbar und es gilt

$$\varphi'(0) = \frac{\mathrm{d}}{\mathrm{d}\alpha} f\big(\mathbf{x}_k - \alpha \nabla f(\mathbf{x}_k)\big)\Big|_{\alpha=0} = -\|\nabla f(\mathbf{x}_k)\|_2^2 < 0.$$

Aus Stetigkeitsgründen folgt die Existenz eines $\delta > 0$ mit $\varphi'(\alpha) < 0$ für alle $0 \leq \alpha \leq \delta$ und damit die Behauptung. □

Algorithmus 2.2 (idealisiertes Gradientenverfahren)
input: Funktion $f : \mathbb{R}^n \to \mathbb{R}$ und Startnäherung $\mathbf{x}_0 \in \mathbb{R}^n$
output: Folge von Iterierten $\{\mathbf{x}_k\}_{k\in\mathbb{N}}$

① Initialisierung: setze $k := 0$
② berechne den Antigradienten $\mathbf{d}_k = -\nabla f(\mathbf{x}_k)$
③ löse

$$\alpha_k = \operatorname*{argmin}_{\alpha\in\mathbb{R}} f(\mathbf{x}_k + \alpha \mathbf{d}_k) \tag{2.1}$$

④ setze $\mathbf{x}_{k+1} := \mathbf{x}_k + \alpha_k \mathbf{d}_k$

⑤ erhöhe $k := k + 1$ und gehe nach ②

Satz 2.3 Die Funktion $f : D \subset \mathbb{R}^n \to \mathbb{R}$ sei gleichmäßig konvex. Weiter sei f differenzierbar mit Lipschitz-stetigem Gradienten:

$$\|\nabla f(\mathbf{x}) - \nabla f(\mathbf{y})\|_2 \leq L\|\mathbf{x} - \mathbf{y}\|_2 \quad \text{für alle } \mathbf{x}, \mathbf{y} \in D.$$

Dann konvergiert das idealisierte Gradientenverfahren für beliebige Startnäherungen $\mathbf{x}_0 \in D$ gegen das eindeutige globale Minimum $\mathbf{x}^\star$ und es gilt

$$f(\mathbf{x}_{k+1}) - f(\mathbf{x}^\star) \leq \left(1 - \frac{\mu^2}{L^2}\right)\big\{f(\mathbf{x}_k) - f(\mathbf{x}^\star)\big\}, \quad k = 1, 2, \ldots.$$

Beweis. Aufgrund von (2.1) gilt für ein beliebiges $\beta > 0$, dass

$$\begin{aligned}
f(\mathbf{x}_{k+1}) &\leq f\big(\mathbf{x}_k - \beta\nabla f(\mathbf{x}_k)\big) \\
&= f(\mathbf{x}_k) - \beta \int_0^1 \nabla f\big(\mathbf{x}_k - t\beta\nabla f(\mathbf{x}_k)\big)^T \nabla f(\mathbf{x}_k)\,\mathrm{d}t \\
&= f(\mathbf{x}_k) - \beta\|\nabla f(\mathbf{x}_k)\|_2^2 - \beta \int_0^1 \big\{\nabla f\big(\mathbf{x}_k - t\beta\nabla f(\mathbf{x}_k)\big) - \nabla f(\mathbf{x}_k)\big\}^T \nabla f(\mathbf{x}_k)\,\mathrm{d}t \\
&\leq f(\mathbf{x}_k) - \beta\|\nabla f(\mathbf{x}_k)\|_2^2 + \beta \int_0^1 \underbrace{\big\|\nabla f\big(\mathbf{x}_k - t\beta\nabla f(\mathbf{x}_k)\big) - \nabla f(\mathbf{x}_k)\big\|_2}_{\leq t\beta L\|\nabla f(\mathbf{x}_k)\|_2} \|\nabla f(\mathbf{x}_k)\|_2\,\mathrm{d}t \\
&\leq f(\mathbf{x}_k) - \left(\beta - \beta^2\frac{L}{2}\right)\|\nabla f(\mathbf{x}_k)\|_2^2.
\end{aligned}$$

Für $\beta = 1/L$ schließen wir also

$$f(\mathbf{x}_{k+1}) - f(\mathbf{x}^\star) \leq f(\mathbf{x}_k) - f(\mathbf{x}^\star) - \frac{1}{2L}\|\nabla f(\mathbf{x}_k) - \underbrace{\nabla f(\mathbf{x}^\star)}_{=\mathbf{0}}\|_2^2. \tag{2.2}$$

Die gleichmäßige Konvexität impliziert

$$\mu\|\mathbf{x}_k - \mathbf{x}^\star\|_2^2 \leq \big\{\nabla f(\mathbf{x}_k) - \nabla f(\mathbf{x}^\star)\big\}^T(\mathbf{x}_k - \mathbf{x}^\star) \leq \|\nabla f(\mathbf{x}_k) - \nabla f(\mathbf{x}^\star)\|_2\|\mathbf{x}_k - \mathbf{x}^\star\|_2.$$

Dies eingesetzt in (2.2) liefert

$$f(\mathbf{x}_{k+1}) - f(\mathbf{x}^\star) \leq f(\mathbf{x}_k) - f(\mathbf{x}^\star) - \frac{\mu^2}{2L}\|\mathbf{x}_k - \mathbf{x}^\star\|_2^2.$$

Weiterhin erhalten wir aus der Lipschitz-Bedingung

$$\begin{aligned} f(\mathbf{x}_k) - f(\mathbf{x}^\star) &= \int_0^1 \nabla f\big(\mathbf{x}^\star + t(\mathbf{x}_k - \mathbf{x}^\star)\big)^T (\mathbf{x}_k - \mathbf{x}^\star)\,\mathrm{d}t \\ &= \int_0^1 \big\{\nabla f\big(\mathbf{x}^\star + t(\mathbf{x}_k - \mathbf{x}^\star)\big) - \nabla f(\mathbf{x}^\star)\big\}^T (\mathbf{x}_k - \mathbf{x}^\star)\,\mathrm{d}t \\ &\le \int_0^1 \underbrace{\big\|\nabla f\big(\mathbf{x}^\star + t(\mathbf{x}_k - \mathbf{x}^\star)\big) - \nabla f(\mathbf{x}^\star)\big\|_2}_{\le tL\|\mathbf{x}_k - \mathbf{x}^\star\|_2} \|\mathbf{x}_k - \mathbf{x}^\star\|_2\,\mathrm{d}t \\ &= \frac{L}{2}\|\mathbf{x}_k - \mathbf{x}^\star\|_2^2, \end{aligned}$$

womit sich die gewünschte Abschätzung ergibt. □

Bemerkung Im obigen Beweis haben wir

$$f(\mathbf{x}_k) - f(\mathbf{x}^\star) \le \frac{L}{2}\|\mathbf{x}_k - \mathbf{x}^\star\|_2^2$$

gezeigt. Umgekehrt folgt aus der gleichmäßigen Konvexität aber auch

$$\begin{aligned} f(\mathbf{x}_k) - f(\mathbf{x}^\star) &\ge \int_0^1 \nabla f\big(\mathbf{x}^\star + t(\mathbf{x}_k - \mathbf{x}^\star)\big)^T (\mathbf{x}_k - \mathbf{x}^\star)\,\mathrm{d}t \\ &\ge \int_0^1 t\mu\|\mathbf{x}_k - \mathbf{x}^\star\|_2^2\,\mathrm{d}t \\ &= \frac{\mu}{2}\|\mathbf{x}_k - \mathbf{x}^\star\|_2^2. \end{aligned}$$

Daher folgt aus Satz 2.3, dass der Fehler des idealisierten Gradientenverfahrens

$$\begin{aligned} \|\mathbf{x}_k - \mathbf{x}^\star\|_2^2 &\le \frac{2}{\mu}\big\{f(\mathbf{x}_k) - f(\mathbf{x}^\star)\big\} \\ &\le \frac{2}{\mu}\left(1 - \frac{\mu^2}{L^2}\right)^k \big\{f(\mathbf{x}_0) - f(\mathbf{x}^\star)\big\} \\ &\le \frac{L}{\mu}\left(1 - \frac{\mu^2}{L^2}\right)^k \|\mathbf{x}_0 - \mathbf{x}^\star\|_2^2. \end{aligned}$$

Wir erhalten demnach eine *lineare* Konvergenzordnung

$$\|\mathbf{x}_k - \mathbf{x}^\star\|_2 \le c\rho^k \|\mathbf{x}_0 - \mathbf{x}^\star\|_2$$

mit $\rho = \sqrt{1 - \mu^2/L^2} < 1$. △

2.2 Praktische Variante

Um einen praktikablen Algorithmus für das Gradientenverfahren zu erhalten, müssen wir noch das eindimensionale Minimierungsproblem (2.1) lösen. Die Bedingung

$$\frac{\mathrm{d}}{\mathrm{d}\alpha} f(\mathbf{x}_k + \alpha\mathbf{d}_k) = \nabla f(\mathbf{x}_k + \alpha\mathbf{d}_k)^T \mathbf{d}_k = 0$$

stellt eine nichtlineare Gleichung für $\alpha \in \mathbb{R}$ dar, deren exaktes Lösen viel zu teuer ist. Daher halbiert man die Schrittweite ausgehend von $\alpha_k = 1$ solange, bis ein Abstieg erzielt wurde.

Bemerkung Ist $f(\mathbf{x}) = \frac{1}{2}\mathbf{x}^T\mathbf{A}\mathbf{x} + \mathbf{b}^T\mathbf{x} + c$ eine quadratische Funktion, dann kann (2.1) exakt gelöst werden:

$$\alpha_k = \frac{\mathbf{d}_k^T\mathbf{d}_k}{\mathbf{d}_k\mathbf{A}\mathbf{d}_k}, \quad \text{wobei } \mathbf{d}_k := \mathbf{b} - \mathbf{A}\mathbf{x}_k.$$

In diesem Fall ist das idealisierte Gradientenverfahren leicht durchführbar und konvergiert gemäß Satz 2.3 linear. Der Kontraktionsfaktor ρ hängt dabei stark von der Kondition der Matrix $\mathbf{A}$ ab, denn es gilt

$$\rho = \frac{\mathrm{cond}_2(A) - 1}{\mathrm{cond}_2(A) + 1}.$$

△

Algorithmus 2.4 (Gradientenverfahren)

input: Funktion $f : \mathbb{R}^n \to \mathbb{R}$ und Startnäherung $\mathbf{x}_0 \in \mathbb{R}^n$

output: Folge von Iterierten $\{\mathbf{x}_k\}_{k\in\mathbb{N}}$

① Initialisierung: wähle $\sigma \in (0,1)$ und setze $k := 0$

② berechne den Antigradienten $\mathbf{d}_k := -\nabla f(\mathbf{x}_k)$ und setze $\alpha_k := 1$

③ solange

$$f(\mathbf{x}_k + \alpha_k\mathbf{d}_k) > f(\mathbf{x}_k) - \sigma\alpha_k\|\mathbf{d}_k\|_2^2 \tag{2.3}$$

setze $\alpha_k := \alpha_k/2$

④ setze $\mathbf{x}_{k+1} := \mathbf{x}_k + \alpha_k\mathbf{d}_k$

⑤ erhöhe $k := k + 1$ und gehe nach ②

Bemerkung Die Liniensuche ③ wird als *Armijo-Schrittweitenregel* bezeichnet. Sie garantiert nicht nur Abbruchbedingung $f(\mathbf{x}_{k+1}) < f(\mathbf{x}_k)$, sondern die *Armijo-Goldstein-Bedingung*

$$f\big(\mathbf{x}_k - \alpha_k\nabla f(\mathbf{x}_k)\big) \le f(\mathbf{x}_k) - \sigma\alpha_k\|\nabla f(\mathbf{x}_k)\|_2^2. \tag{2.4}$$

Dabei bricht die Liniensuche mit einem $\alpha_k > 0$ ab, denn für $\varphi(\alpha) = f\big(\mathbf{x}_k - \alpha\nabla f(\mathbf{x}_k)\big)$ gilt nämlich

$$f\big(\mathbf{x}_k - \alpha\nabla f(\mathbf{x}_k)\big) = \varphi(\alpha) = \varphi(0) + \alpha\varphi'(0) + o(\alpha) = f(\mathbf{x}_k) - \alpha\|\nabla f(\mathbf{x}_k)\|_2^2 + o(\alpha).$$

△

Der nachfolgende Satz liefert ein globales Konvergenzresultat für das Gradientenverfahren unter der Verwendung der Armijo-Schrittweitenregel. Er benötigt nur die Lipschitzstetigkeit von f, aber keine Konvexität.

Satz 2.5 Es sei $D \subset \mathbb{R}^n$ eine offene Menge, in der f stetig differenzierbar, nach unten beschränkt und ∇f zudem Lipschitz-stetig ist. Ferner sei neben $\mathbf{x}_0$ auch die gesamte Niveaumenge $\{\mathbf{x} \in \mathbb{R}^n : f(\mathbf{x}) \le f(\mathbf{x}_0)\}$ in D enthalten. Dann gilt für die Iterierten $\{\mathbf{x}_k\}_{k\ge0}$ aus Algorithmus 2.4

$$\nabla f(\mathbf{x}_k) \overset{k\to\infty}{\longrightarrow} \mathbf{0}.$$

Beweis. Da D die gesamte Niveaumenge enthält, ist sichergestellt, dass die Iterierten $\{\mathbf{x}_k\}_{k\geq 0}$ alle in D enthalten sind. Nach Konstruktion ist dann die Folge $\{f(\mathbf{x}_k)\}_{k\geq 0}$ monoton fallend und nach unten beschränkt. Daher folgt aus der Armijo-Goldstein-Bedingung (2.4), dass

$$f(\mathbf{x}_0) \geq f(\mathbf{x}_1) + \sigma\alpha_0\|\nabla f(\mathbf{x}_0)\|_2^2 \geq \cdots \geq f(\mathbf{x}_{k+1}) + \sigma\sum_{\ell=0}^{k}\alpha_\ell\|\nabla f(\mathbf{x}_\ell)\|_2^2 \geq 0.$$

Da die Reihe auf der rechten Seite notwendigerweise für $k \to \infty$ konvergent ist, folgt

$$\alpha_k\|\nabla f(\mathbf{x}_k)\|_2^2 \overset{k\to\infty}{\longrightarrow} 0. \tag{2.5}$$

Es verbleibt zu zeigen, dass $\alpha_k > \varepsilon$ für ein $\varepsilon > 0$.
Für festes k ist aufgrund von Algorithmus 2.4 $\alpha_k = 1$ oder die Armijo-Goldstein-Bedingung ist für $2\alpha_k$ verletzt:

$$2\sigma\alpha_k\|\nabla f(\mathbf{x}_k)\|_2^2 > f(\mathbf{x}_k) - f\big(\mathbf{x}_k - 2\alpha_k\nabla f(\mathbf{x}_k)\big) = 2\alpha_k\|\nabla f(\mathbf{x}_k)\|_2^2 - R_2(\mathbf{x}_k, 2\alpha_k)$$

mit dem Taylor-Restglied

$$R_2(\mathbf{x}_k, 2\alpha_k) = 2\alpha_k\Big(\|\nabla f(\mathbf{x}_k)\|_2^2 - \nabla f\big(\mathbf{x}_k - \xi\nabla f(\mathbf{x}_k)\big)^T\nabla f(\mathbf{x}_k)\Big), \quad \xi \in (0, 2\alpha_k).$$

Da ∇f Lipschitz-stetig ist, ist das Taylor-Restglied $R_2(\mathbf{x}_k, 2\alpha_k)$ durch $\nu\alpha_k^2\|\nabla f(\mathbf{x}_k)\|_2^2$ für ein geeignetes $\nu > 0$ beschränkt. Daher ist

$$\sigma\alpha_k^2\|\nabla f(\mathbf{x}_k)\|_2^2 > 2\alpha_k\|\nabla f(\mathbf{x}_k)\|_2^2 - 2\nu\alpha_k\|\nabla f(\mathbf{x}_k)\|_2^2 = 2\alpha_k(1-\nu)\|\nabla f(\mathbf{x}_k)\|_2^2,$$

dies bedeutet

$$\alpha_k > \frac{2(1-\nu)}{\sigma} =: \varepsilon > 0.$$

Somit bleibt α_k für alle $k \geq 0$ größer als $\min\{1, \varepsilon\}$ und daher folgt die Behauptung aus (2.5). □

Beachte: Satz 2.5 besagt nicht, dass die Folge $\{\mathbf{x}_k\}_{k\geq 0}$ selber konvergiert. Selbst wenn die Folge $\{\mathbf{x}_k\}_{k\geq 0}$ konvergiert, braucht der Grenzwert darüberhinaus kein Minimum von f zu sein.

Beispiel 2.6 Gegeben sei die Funktion

$$f(\xi, \eta) = \xi^2 + (\eta^2 - 1)^2 + \xi^2(\eta^2 - 1)^2 \to \min.$$

Das Minimum von f ist 0 und wird offensichtlich für $\xi = 0$ und $\eta = \pm 1$ angenommen. Hat eine Iterierte $\mathbf{x}_k$ von Algorithmus 2.4 die Form $\mathbf{x}_k = [\xi_k, 0]^T$, dann gilt

$$\nabla f(\mathbf{x}_k) = \begin{bmatrix} 2\xi + 2\xi(\eta^2-1)^2 \\ 2\eta(\eta^2-1)(1+\xi^2) \end{bmatrix}\Bigg|_{(\xi,\eta)=(\xi_k,0)} = \begin{bmatrix} 4\xi_k \\ 0 \end{bmatrix}.$$

Daher hat die nächste Iterierte zwangsläufig wieder die Form $\mathbf{x}_{k+1} = [\xi_{k+1}, 0]^T$ und nach Satz 2.5 konvergiert $\nabla f(\mathbf{x}_k) = [4\xi_k, 0]^T$ gegen Null. Deshalb streben auch ξ_k und $\mathbf{x}_k$ gegen Null für $k \to \infty$. Dennoch ist $[0, 0]^T$ lediglich ein Sattelpunkt von f, da $f(0, \eta)$ für $\eta = 0$ ein lokales Maximum aufweist. △

3. Newton-Verfahren

3.1 Lokales Newton-Verfahren

Sei $f : \mathbb{R}^n \to \mathbb{R}$ zweimal stetig differenzierbar. Beim *Newton-Verfahren* wird das Minimierungsproblem $f(\mathbf{x}) \to \min$ gelöst, indem sukzessive die quadratischen Näherungen

$$q_k(\mathbf{x}) := f(\mathbf{x}_k) + \nabla f(\mathbf{x}_k)^T(\mathbf{x} - \mathbf{x}_k) + \frac{1}{2}(\mathbf{x} - \mathbf{x}_k)^T \nabla^2 f(\mathbf{x}_k)(\mathbf{x} - \mathbf{x}_k)$$

minimiert. Ist die Hesse-Matrix $\nabla^2 f(\mathbf{x}_k)$ positiv definit, so ist die neue Iterierte $\mathbf{x}_{k+1}$ genau die Lösung der Gleichung

$$\nabla q_k(\mathbf{x}) = \nabla f(\mathbf{x}_k) + \nabla^2 f(\mathbf{x}_k)(\mathbf{x} - \mathbf{x}_k) \stackrel{!}{=} \mathbf{0}.$$

Hieraus ergibt sich die Iterationsvorschrift

$$\mathbf{x}_{k+1} = \mathbf{x}_k - \left(\nabla^2 f(\mathbf{x}_k)\right)^{-1} \nabla f(\mathbf{x}_k), \quad k = 0, 1, 2, \ldots. \tag{3.1}$$

Algorithmus 3.1 (Newton-Verfahren)
input: Funktion $f : \mathbb{R}^n \to \mathbb{R}$ und Startnäherung $\mathbf{x}_0 \in \mathbb{R}^n$
output: Folge von Iterierten $\{\mathbf{x}_k\}_{k \in \mathbb{N}}$

① Initialisierung: setze $k := 0$

② berechne die Newton-Richtung durch Lösen des linearen Gleichungssystems

$$\nabla^2 f(\mathbf{x}_k)\mathbf{d}_k = -\nabla f(\mathbf{x}_k) \tag{3.2}$$

③ setze $\mathbf{x}_{k+1} := \mathbf{x}_k + \mathbf{d}_k$

④ erhöhe $k := k + 1$ und gehe nach ②

Der mächtigste Satz zum Newton-Verfahren ist der Satz von Newton-Kantorovich. Er liefert nicht nur die Konvergenz des Newton-Verfahrens, sondern auch die *Existenz* eines stationären Punkts.

Satz 3.2 (Newton-Kantorovich) Sei $D \subset \mathbb{R}^n$ offen und konvex und $f : D \subset \mathbb{R}^n \to \mathbb{R}$ zweimal stetig differenzierbar mit

$$\|\nabla^2 f(\mathbf{x}) - \nabla^2 f(\mathbf{y})\|_2 \leq L\|\mathbf{x} - \mathbf{y}\|_2$$

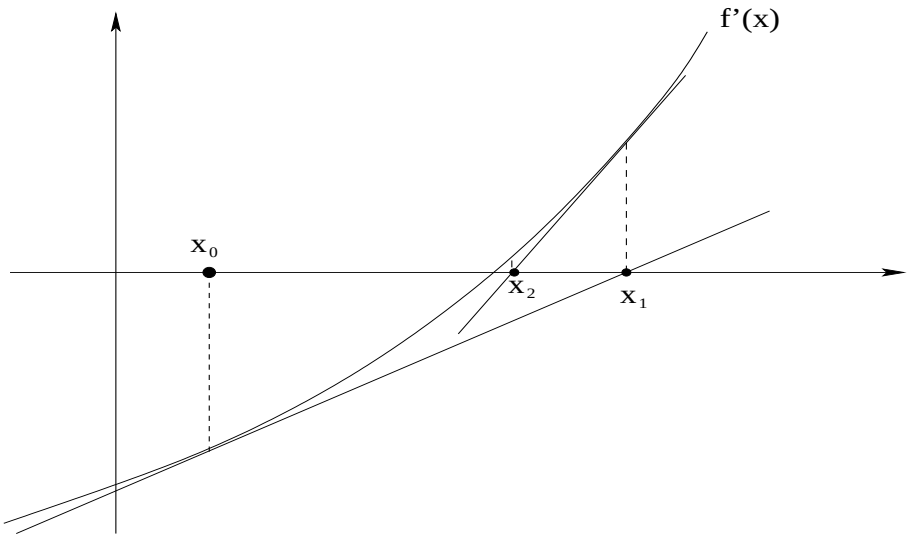


Abbildung 3.1: Geometrische Interpretation des Newton-Verfahrens.

für alle $\mathbf{x}, \mathbf{y} \in D$. Für ein gegebenes $\mathbf{x}_0 \in D$ nehmen wir an, dass die Hesse-Matrix $\nabla^2 f(\mathbf{x}_0)$ invertierbar ist mit $\alpha := \left\| \left(\nabla^2 f(\mathbf{x}_0)\right)^{-1} \right\|_2$. Ferner gelte

$$\beta := \left\| \left(\nabla^2 f(\mathbf{x}_0)\right)^{-1} \nabla f(\mathbf{x}_0) \right\|_2 \leq \frac{1}{2\alpha L}$$

und für

$$\gamma^{\pm} := \frac{1}{\alpha L}\left(1 \pm \sqrt{1 - 2\alpha\beta L}\right) \geq 0$$

sei $S := \{\mathbf{x} \in \mathbb{R}^n : \|\mathbf{x} - \mathbf{x}_0\|_2 \leq \gamma^-\} \subset D$. Dann sind die Iterierten des Newton-Verfahrens (3.1) wohldefiniert, liegen alle in S und konvergieren gegen ein $\mathbf{x}^\star \in S$ mit $\nabla f(\mathbf{x}^\star) = \mathbf{0}$, welches eindeutig ist in $D \cap \{\mathbf{x} \in \mathbb{R}^n : \|\mathbf{x} - \mathbf{x}_0\|_2 \leq \gamma^+\}$. Insbesondere ist die Konvergenz quadratisch, falls $2\alpha\beta L < 1$ ist. Dies bedeutet, es existiert ein $c > 0$ so dass gilt

$$\|\mathbf{x}^\star - \mathbf{x}_{k+1}\|_2 \leq c\|\mathbf{x}^\star - \mathbf{x}_k\|_2^2.$$

Bevor wir den Beweis dieses Satzes erbringen, stellen wir drei hilfreiche Lemmata zur Verfügung. Dabei gelten stets die Voraussetzungen des Satzes von Newton-Kantorovich.

Lemma 3.3 Sei $\{\mathbf{y}_k\}_{k\geq 0}$ eine Folge im $\mathbb{R}^n$ und $\{t_k\}_{k\geq 0}$ eine Folge nichtnegativer, monoton wachsender, reeller Zahlen mit $t_k \to t^\star < \infty$. Gilt

$$\|\mathbf{y}_{k+1} - \mathbf{y}_k\|_2 \leq t_{k+1} - t_k, \quad k = 0, 1, \ldots,$$

dann existiert ein eindeutig bestimmtes $\mathbf{y}^\star \in \mathbb{R}^n$ mit $\mathbf{y}_k \to \mathbf{y}^\star$ und

$$\|\mathbf{y}^\star - \mathbf{y}_k\|_2 \leq t^\star - t_k, \quad k = 0, 1, \ldots.$$

Dies bedeutet, die Folge $\{t_k\}_{k\geq 0}$ majorisiert $\{\mathbf{y}_k\}_{k\geq 0}$.

Beweis. Es gilt

$$\begin{aligned}\|\mathbf{y}_{k+p}-\mathbf{y}_k\|_2 &= \Big\|\sum_{\ell=k}^{k+p-1}(\mathbf{y}_{\ell+1}-\mathbf{y}_\ell)\Big\|_2\\ &\le \sum_{\ell=k}^{k+p-1}\|\mathbf{y}_{\ell+1}-\mathbf{y}_\ell\|_2\\ &\le \underbrace{\sum_{\ell=k}^{k+p-1} t_{\ell+1}-t_\ell}_{=t_{k+p}-t_k}\\ &\le t^\star - t_k.\end{aligned}$$

Folglich ist $\{\mathbf{y}_k\}_{k\ge 0}$ eine Cauchy-Folge im $\mathbb{R}^n$, woraus die Behauptung folgt. □

Lemma 3.4 Für alle $\mathbf{x}\in T := \{\mathbf{x}\in\mathbb{R}^n : \|\mathbf{x}-\mathbf{x}_0\|_2 < 1/(\alpha L)\}$ ist die Hesse-Matrix $\nabla^2 f(\mathbf{x})$ invertierbar mit

$$\big\|\big(\nabla^2 f(\mathbf{x})\big)^{-1}\big\|_2 \le \frac{\alpha}{1-\alpha L\|\mathbf{x}-\mathbf{x}_0\|_2}. \tag{3.3}$$

Ist außerdem $N(\mathbf{x}) := \mathbf{x} - \big(\nabla^2 f(\mathbf{x})\big)^{-1}\nabla f(\mathbf{x}) \in T$, dann gilt

$$\big\|N\big(N(\mathbf{x})\big)-N(\mathbf{x})\big\|_2 \le \frac{1}{2}\frac{\alpha L\|\mathbf{x}-N(\mathbf{x})\|_2}{1-\alpha L\|\mathbf{x}_0-N(\mathbf{x})\|_2}.$$

Beweis. Für $\mathbf{x}\in T$ gilt

$$\big\|\big(\nabla^2 f(\mathbf{x}_0)\big)^{-1}\big(\nabla^2 f(\mathbf{x}_0)-\nabla^2 f(\mathbf{x})\big)\big\|_2 \le \underbrace{\big\|\big(\nabla^2 f(\mathbf{x}_0)\big)^{-1}\big\|_2}_{\le\alpha}\underbrace{\big\|\nabla^2 f(\mathbf{x}_0)-\nabla^2 f(\mathbf{x})\big\|_2}_{\le L\|\mathbf{x}_0-\mathbf{x}\|_2} < 1.$$

Die Satz von der Neumann-Reihe liefert daher die Existenz von

$$\big\{I-\big(\nabla^2 f(\mathbf{x}_0)\big)^{-1}\big(\nabla^2 f(\mathbf{x}_0)-\nabla^2 f(\mathbf{x})\big)\big\}^{-1} = \sum_{k=0}^{\infty}\big\{\big(\nabla^2 f(\mathbf{x}_0)\big)^{-1}\big(\nabla^2 f(\mathbf{x}_0)-\nabla^2 f(\mathbf{x})\big)\big\}^k.$$

Speziell folgt die Anschätzung

$$\big\|\big\{I-\big(\nabla^2 f(\mathbf{x}_0)\big)^{-1}\big(\nabla^2 f(\mathbf{x}_0)-\nabla^2 f(\mathbf{x})\big)\big\}^{-1}\big\|_2 \le \sum_{k=0}^{\infty}\big\{\alpha L\|\mathbf{x}_0-\mathbf{x}\|_2\big\}^k = \frac{1}{1-\alpha L\|\mathbf{x}_0-\mathbf{x}\|_2}.$$

Die Abschätzung (3.3) ergibt sich nun aus der Identität

$$\big(\nabla^2 f(\mathbf{x})\big)^{-1} = \big\{I-\big(\nabla^2 f(\mathbf{x}_0)\big)^{-1}\big(\nabla^2 f(\mathbf{x}_0)-\nabla^2 f(\mathbf{x})\big)\big\}^{-1}\nabla^2 f(\mathbf{x}_0).$$

Die zweite Aussage des Lemmas sieht man wie folgt. Mit dem soeben gezeigten gilt

$$\begin{aligned}\big\|N\big(N(\mathbf{x})\big) - N(\mathbf{x})\big\|_2 &= \Big\|N(\mathbf{x}) - \Big(\nabla^2 f\big(N(\mathbf{x})\big)\Big)^{-1}\nabla f\big(N(\mathbf{x})\big) - N(\mathbf{x})\Big\|_2 \\ &\le \Big\|\Big(\nabla^2 f\big(N(\mathbf{x})\big)\Big)^{-1}\Big\|_2 \big\|\nabla f\big(N(\mathbf{x})\big)\big\|_2 \\ &\le \frac{\alpha}{1-\alpha L\|N(\mathbf{x})-\mathbf{x}_0\|_2}\big\|\nabla f\big(N(\mathbf{x})\big)\big\|_2.\end{aligned}$$

Wir müssen also nur noch $\big\|\nabla f\big(N(\mathbf{x})\big)\big\|_2$ abschätzen. Aufgrund der Identität

$$\nabla f(\mathbf{x}) - \nabla^2 f(\mathbf{x})\big(N(\mathbf{x}) - \mathbf{x}\big) = \mathbf{0}$$

schließen wir

$$\big\|\nabla f\big(N(\mathbf{x})\big)\big\|_2 = \big\|\nabla f\big(N(\mathbf{x})\big) - \nabla f(\mathbf{x}) - \nabla^2 f(\mathbf{x})\big(N(\mathbf{x}) - \mathbf{x}\big)\big\|_2.$$

Mit

$$\nabla f(\mathbf{y}) - \nabla f(\mathbf{x}) = \int_0^1 \nabla^2 f\big(\mathbf{x} + t(\mathbf{y}-\mathbf{x})\big)(\mathbf{y}-\mathbf{x})\,\mathrm{d}t$$

und $\mathbf{y} := N(\mathbf{x})$ folgt

$$\begin{aligned}\big\|\nabla f\big(N(\mathbf{x})\big)\big\|_2 &= \Big\|\int_0^1 \big\{\nabla^2 f\big(\mathbf{x}+t(\mathbf{y}-\mathbf{x})\big) - \nabla^2 f(\mathbf{x})\big\}(\mathbf{y}-\mathbf{x})\,\mathrm{d}t\Big\|_2 \\ &\le \int_0^1 \underbrace{\big\|\nabla^2 f\big(\mathbf{x}+t(\mathbf{y}-\mathbf{x})\big) - \nabla^2 f(\mathbf{x})\big\|_2}_{\le Lt\|\mathbf{y}-\mathbf{x}\|_2}\|\mathbf{y}-\mathbf{x}\|_2\,\mathrm{d}t \\ &\le \frac{L}{2}\|N(\mathbf{x})-\mathbf{x}\|_2^2.\end{aligned}$$

Damit ergibt sich schließlich die Behauptung. □

Lemma 3.5 Die Folge der Newton-Iterierten $\{\mathbf{x}_k\}_{k\ge 0}$ ist wohldefiniert und durch die Folge $\{t_k\}_{k\ge 0}$ mit

$$t_0 := 0, \qquad t_{k+1} := \frac{\beta - \frac{\alpha L}{2}t_k^2}{1-\alpha L t_k}, \quad k = 1, 2, \ldots \tag{3.4}$$

majorisiert. Insbesondere konvergiert t_k monoton von unten gegen $t^\star := \gamma^-$.

Beweis. Wir zeigen zunächst induktiv, dass $0 = t_0 < t_1 < \cdots < t_k < t^\star$ gilt. Da für $k = 0$ nichts zu zeigen ist, können wir annehmen, dass die Aussage für ein $k \in \mathbb{N}_0$ bewiesen ist. Mit quadratischer Ergänzung folgt die Gleichung

$$\begin{aligned}t_{k+1} - t_k &= \frac{\beta - \frac{\alpha L}{2}t_k^2 - t_k + \alpha L t_k^2}{1-\alpha L t_k} \\ &= \frac{\frac{\alpha L}{2}(t_k - t_{k-1})^2 + \overbrace{\beta - \frac{\alpha L}{2}t_{k-1}^2 - (1-\alpha L t_{k-1})t_k}^{=0}}{1-\alpha L t_k} \\ &= \frac{1}{2}\frac{\alpha L(t_k - t_{k-1})^2}{1-\alpha L t_k}.\end{aligned}$$

Wegen $1 - \alpha L t_k > 0$ und $t_{k-1} < t_k$ schließen wir $t_k < t_{k+1}$. Um $t_{k+1} < t^\star$ zu zeigen, betrachten wir das Polynom $p(t) := \frac{\alpha L}{2} t^2 - t + \beta$, welches die zwei Nullstellen $\gamma^- = t^\star$ und γ^+ besitzt. Wegen

$$\begin{aligned} t^\star - t_{k+1} &= \frac{t^\star - \alpha L t_k t^\star - \beta + \frac{\alpha L}{2} t_k^2}{1 - \alpha L t_k} \\ &= \frac{\frac{\alpha L}{2}(t^\star - t_k)^2 - \overbrace{p(t^\star)}^{=0}}{1 - \alpha L t_k} \\ &= \frac{1}{2} \frac{\alpha L (t^\star - t_k)^2}{\underbrace{1 - \alpha L t_k}_{>0}} > 0 \end{aligned}$$

ist in der Tat $t_{k+1} < t^\star$, womit ist der Induktionsschritt $k \mapsto k+1$ erbracht ist. Insbesondere folgt zwingend die Konvergenz $t_k \to t^\star$, da Fixpunkte der Iteration (3.4) der Bedingung $0 = p(t)$ genügen müssen und $t^\star = \gamma^- < \gamma^+$ gilt.

Wir zeigen nun, dass alle $\mathbf{x}_k$ existieren, in S liegen und jeweils $\|\mathbf{x}_k - \mathbf{x}_{k-1}\|_2 \leq t_k - t_{k-1}$ erfüllen. Für $k = 0$ ist dies sicherlich richtig. Es möge also die Folge $\mathbf{x}_1, \ldots, \mathbf{x}_k$ das Behauptete erfüllen. Wegen $S \subset T$ folgt der Induktionsschritt $k \mapsto k+1$ dann aus Lemma 3.4 gemäß

$$\begin{aligned} \|\mathbf{x}_{k+1} - \mathbf{x}_k\|_2 &= \big\| N\big(N(\mathbf{x}_{k-1})\big) - N(\mathbf{x}_{k-1}) \big\|_2 \\ &\leq \frac{1}{2} \frac{\alpha L \|\mathbf{x}_k - \mathbf{x}_{k-1}\|_2}{1 - \alpha L \|\mathbf{x}_k - \mathbf{x}_0\|_2} \\ &\leq \frac{1}{2} \frac{\alpha L (t_k - t_{k-1})}{1 - \alpha L t_k} \\ &= t_{k+1} - t_k \end{aligned}$$

und

$$\|\mathbf{x}_{k+1} - \mathbf{x}_0\|_2 \leq \sum_{\ell=0}^{k} \|\mathbf{x}_{\ell+1} - \mathbf{x}_\ell\|_2 \leq \sum_{\ell=0}^{k} t_{\ell+1} - t_\ell = t_{k+1} - t_0 \leq t^\star = \gamma^-.$$

□

Wir sind nun in der Lage, den Satz von Newton-Kantorovich beweisen zu können, wobei wir nur die Existenz eines stationären Punktes $\mathbf{x}^\star$ nachweisen und nicht dessen Eindeutigkeit.

Beweis des Satzes von Newton-Kantorovich. Nach Lemma 3.5 werden die Newton-Iterierten $\{\mathbf{x}_k\}_{k \geq 0}$ von $\{t_k\}_{k \geq 0}$ majorisiert und konvergieren gemäß Lemma 3.3 gegen ein eindeutig bestimmtes $\mathbf{x}^\star \in S$. Wegen

$$\begin{aligned} \|\nabla f(\mathbf{x}_k)\|_2 &= \|\nabla^2 f(\mathbf{x}_k)(\mathbf{x}_{k+1} - \mathbf{x}_k)\|_2 \\ &\leq \big\| \big\{ \nabla^2 f(\mathbf{x}_k) - \nabla^2 f(\mathbf{x}_0) \big\} (\mathbf{x}_{k+1} - \mathbf{x}_k) \big\|_2 + \|\nabla^2 f(\mathbf{x}_0)(\mathbf{x}_{k+1} - \mathbf{x}_k)\|_2 \\ &\leq \underbrace{\big\{ L t^\star + \|\nabla^2 f(\mathbf{x}_0)\|_2 \big\}}_{<\infty} \underbrace{\|\mathbf{x}_{k+1} - \mathbf{x}_k\|_2}_{\to 0} \\ &\to 0 \end{aligned}$$

und der Stetigkeit von ∇f, ist auch die Gleichung $\nabla f(\mathbf{x}^\star) = \mathbf{0}$ erfüllt. Ist ferner $2\alpha\beta L < 1$, dann konvergiert aufgrund der Abschätzung

$$t^\star - t_k = \frac{\frac{\alpha L}{2}(t^\star - t_{k-1})^2}{1 - \alpha L t_{k-1}} \leq \frac{\alpha L}{2 - 2\alpha L t^\star}(t^\star - t_{k-1})^2$$

die Folge $\{t_k\}_{k\geq 0}$ quadratisch. Dies impliziert sofort die quadratische Konvergenz der durch $\{t_k\}_{k\geq 0}$ majorisierten Folge $\{\mathbf{x}_k\}_{k\geq 0}$. □

Bemerkungen

1. Die Konvergenz des Newton-Verfahrens ist im allgemeinen nur lokal.
2. Der Satz von Newton-Kantorovich benutzt nirgendwo Konvexität. Das Newton-Verfahren konvergiert folglich auch im Fall von Sattelpunkten.

△

3.2 Globalisiertes Newton-Verfahren

Um das Newton-Verfahren zu globalisieren, müssen wir sicherstellen, dass die Suchrichtung auch dann wohldefiniert ist, wenn sich das Gleichungssystem (3.2) nicht lösen lässt oder es gar keine Abstiegsrichtung beschreibt. In beiden Fällen machen wir dann einfach einen Gradientenschritt:

Algorithmus 3.6 (globalisiertes Newton-Verfahren)
input: Funktion $f : \mathbb{R}^n \to \mathbb{R}$ und Startnäherung $\mathbf{x}_0 \in \mathbb{R}^n$
output: Folge von Iterierten $\{\mathbf{x}_k\}_{k\in\mathbb{N}}$

① Initialisierung: wähle $\sigma \in (0,1)$, $\rho > 0$ und setze $k := 0$

② berechne, falls möglich, die Newton-Richtung durch Lösen des linearen Gleichungssystems

$$\nabla^2 f(\mathbf{x}_k)\mathbf{d}_k = -\nabla f(\mathbf{x}_k)$$

③ ist das Gleichungssystem nicht lösbar oder ist die Bedingung

$$\nabla f(\mathbf{x}_k)^T \mathbf{d}_k \leq -\rho \|\nabla f(\mathbf{x}_k)\|_2^2 \tag{3.5}$$

nicht erfüllt, dann setze $\mathbf{d}_k := -\nabla f(\mathbf{x}_k)$

④ setze $\alpha_k := 1$

⑤ solange

$$f(\mathbf{x}_k + \alpha_k \mathbf{d}_k) > f(\mathbf{x}_k) + \sigma\alpha_k \nabla f(\mathbf{x}_k)^T \mathbf{d}_k \tag{3.6}$$

setze $\alpha_k := \alpha_k/2$

⑥ setze $\mathbf{x}_{k+1} := \mathbf{x}_k + \alpha_k \mathbf{d}_k$

⑦ erhöhe $k := k + 1$ und gehe nach ②

Die Bedingungen (3.5) und (3.6) garantieren zusammen die Armijo-Goldstein-Bedingung (2.4) und somit einen hinreichenden Abstieg. Die Konvergenz $\nabla f(\mathbf{x}_k) \to \mathbf{0}$ folgt daher analog zum entsprechenden Satz 3.6 über das Gradientenverfahren. Aus dem nachfol-

genden Satz ergibt sich, dass das globalisierte Newton-Verfahren in der Umgebung eines Minimums zum lokalen Newton-Verfahren wird.

Satz 3.7 Seien $f : \mathbb{R}^n \to \mathbb{R}$ zweimal stetig differenzierbar und $\mathbf{x}^\star \in \mathbb{R}^n$ ein stationärer Punkt, an dem die Hesse-Matrix $\nabla^2 f(\mathbf{x}^\star)$ positiv definit ist. Konvergiert die Folge $\{\mathbf{x}_k\}_{k\geq 0}$ der Newton-Iterierten gegen $\mathbf{x}^\star$, dann exitistiert ein $k_0 \in \mathbb{N}$ mit

$$f(\mathbf{x}_k + \mathbf{d}_k) > f(\mathbf{x}_k) + \sigma \nabla f(\mathbf{x}_k)^T \nabla^2 f(\mathbf{x}_k) \nabla f(\mathbf{x}_k) \tag{3.7}$$

für alle $k \geq k_0$ und jedes feste $\sigma \in (0, 1/2)$.

Beweis. Aus den Voraussetzung folgt aus Stetigkeitsgründen zunächst

$$\|\nabla f(\mathbf{x}_k)\|_2 \overset{k\to\infty}{\longrightarrow} \mathbf{0}. \tag{3.8}$$

Ferner ist $\nabla^2 f(\mathbf{x}^\star)$ invertierbar und aufgrund der Stetigkeit schließen wir auf die Existenz einer Umgebung $U \subset \mathbb{R}^n$ von $\mathbf{x}^\star$ mit

$$\left\| \left(\nabla^2 f(\mathbf{x})\right)^{-1} \right\|_2 \leq \overline{c} \quad \text{für alle } \mathbf{x} \in U.$$

Daher existiert ein $k_0 \in \mathbb{N}$, so dass

$$\left\| \left(\nabla^2 f(\mathbf{x}_k)\right)^{-1} \right\|_2 \leq \overline{c} \quad \text{für alle } k \geq k_0.$$

Aus der Newton-Gleichung $\mathbf{d}_k := -\left(\nabla^2 f(\mathbf{x}_k)\right)^{-1} \nabla f(\mathbf{x}_k)$ folgt deshalb

$$\|\mathbf{d}_k\|_2 \leq \left\| \left(\nabla^2 f(\mathbf{x}_k)\right)^{-1} \right\|_2 \|\nabla f(\mathbf{x}_k)\|_2 \leq \overline{c} \|\nabla f(\mathbf{x}_k)\|_2 \overset{k\to\infty}{\longrightarrow} 0.$$

Da mit $\nabla^2 f(\mathbf{x}^\star)$ auch $\left(\nabla^2 f(\mathbf{x}^\star)\right)^{-1}$ positiv definit ist, können wir k_0 vergrößern, so dass

$$\nabla f(\mathbf{x}_k)^T \left(\nabla^2 f(\mathbf{x}_k)\right)^{-1} \nabla f(\mathbf{x}_k) \geq \underline{c} \|\nabla f(\mathbf{x}_k)\|_2^2 \quad \text{für alle } k \geq k_0. \tag{3.9}$$

Nach dem Taylorschen Satz exitiert ein $\boldsymbol{\xi}_k$ auf der Verbindungsstrecke von $\mathbf{x}_k$ und $\mathbf{x}_k + \mathbf{d}_k$, so dass gilt

$$f(\mathbf{x}_k + \mathbf{d}_k) = f(\mathbf{x}_k) + \nabla f(\mathbf{x}_k)^T \mathbf{d}_k + \frac{1}{2} \mathbf{d}_k^T \nabla^2 f(\boldsymbol{\xi}_k) \mathbf{d}_k.$$

Wegen $\mathbf{x}_k \to \mathbf{x}^\star$ und $\mathbf{d}_k \to \mathbf{0}$ gilt auch $\boldsymbol{\xi}_k \to \mathbf{x}^\star$ und folglich

$$\|\nabla^2 f(\boldsymbol{\xi}_k) - \nabla^2 f(\mathbf{x}_k)\|_2 \overset{k\to\infty}{\longrightarrow} 0.$$

Wegen $\sigma \in (0, 1/2)$ ist daher

$$\underbrace{\left(\sigma - \frac{1}{2}\right)}_{<0} \underline{c} + \frac{1}{2} \overline{c}^2 \|\nabla^2 f(\boldsymbol{\xi}_k) - \nabla^2 f(\mathbf{x}_k)\|_2 \leq 0 \tag{3.10}$$

für alle k genügend groß. Für großes k folgt die Behauptung nun aus der Cauchy-Schwarzschen Ungleichung, (3.8),(3.9) und (3.10):

$$\begin{aligned}
f(\mathbf{x}_k+\mathbf{d}_k) &= f(\mathbf{x}_k)+\nabla f(\mathbf{x}_k)^T\mathbf{d}_k+\frac{1}{2}\underbrace{\mathbf{d}_k^T\nabla^2 f(\mathbf{x}_k)\mathbf{d}_k}_{=-\nabla f(\mathbf{x}_k)^T\mathbf{d}_k}+\frac{1}{2}\mathbf{d}_k^T\{\nabla^2 f(\boldsymbol{\xi}_k)-\nabla^2 f(\mathbf{x}_k)\}\mathbf{d}_k\\
&\le f(\mathbf{x}_k)+\frac{1}{2}\nabla f(\mathbf{x}_k)^T\mathbf{d}_k+\frac{1}{2}\|\mathbf{d}_k\|_2^2\|\nabla^2 f(\boldsymbol{\xi}_k)-\nabla^2 f(\mathbf{x}_k)\|_2\\
&= f(\mathbf{x}_k)+\sigma\nabla f(\mathbf{x}_k)^T\mathbf{d}_k+\left(\sigma-\frac{1}{2}\right)\underbrace{\mathbf{d}_k^T\nabla^2 f(\mathbf{x}_k)\mathbf{d}_k}_{=\nabla f(\mathbf{x}_k)^T(\nabla^2 f(\mathbf{x}_k))^{-1}\nabla f(\mathbf{x}_k)}\\
&\qquad+\frac{1}{2}\|\mathbf{d}_k\|_2^2\|\nabla^2 f(\boldsymbol{\xi}_k)-\nabla^2 f(\mathbf{x}_k)\|_2\\
&\le f(\mathbf{x}_k)+\sigma\nabla f(\mathbf{x}_k)^T\mathbf{d}_k\\
&\qquad+\Bigg\{\underbrace{\left(\sigma-\frac{1}{2}\right)\underline{c}+\frac{1}{2}\overline{c}^2\|\nabla^2 f(\boldsymbol{\xi}_k)-\nabla^2 f(\mathbf{x}_k)\|_2}_{\le 0}\Bigg\}\|\nabla f(\mathbf{x}_k)\|_2^2\\
&\le f(\mathbf{x}_k)+\sigma\nabla f(\mathbf{x}_k)^T\mathbf{d}_k.
\end{aligned}$$

□

Sind die Voraussetzungen des Satzes 3.7 erfüllt, dann zeigt Abschätzung (3.10), dass (3.5) für ein genügend kleines ρ erfüllt ist. Daher folgt dann aus (3.7), dass (3.6) stets gilt.

3.3 Inexaktes Newton-Verfahren

Beim inexakten Newton-Verfahren wird das Gleichungssystem (3.2) nicht exakt, sondern nur näherungsweise gelöst. Dies entspricht natürlich der numerischen Praxis.

Algorithmus 3.8 (inexaktes lokales Newton-Verfahren)
input: Funktion $f:\mathbb{R}^n\to\mathbb{R}$ und Startnäherung $\mathbf{x}_0\in\mathbb{R}^n$
output: Folge von Iterierten $\{\mathbf{x}_k\}_{k\in\mathbb{N}}$

① Initialisierung: setze $k:=0$

② wähle eine Toleranz $\eta_k>0$ und bestimme eine Suchrichtung $\mathbf{d}_k$ mit

$$\|\nabla f(\mathbf{x}_k)+\nabla^2 f(\mathbf{x}_k)\mathbf{d}_k\|_2\le\eta_k\|\nabla f(\mathbf{x}_k)\|_2 \tag{3.11}$$

③ setze $\mathbf{x}_{k+1}:=\mathbf{x}_k+\mathbf{d}_k$

④ erhöhe $k:=k+1$ und gehe nach ②

Die Konvergenzrate des inexakten Newton-Verfahrens ergibt sich aus dem nachfolgenden Satz. Speziell gibt er uns eine Bedingung an die Wahl der Toleranzen $\{\eta_k\}$ an, so dass das Verfahren dennoch quadratisch konvergiert.

Satz 3.9 Seien $f : \mathbb{R}^n \to \mathbb{R}$ zweimal stetig differenzierbar und $\mathbf{x}^\star \in \mathbb{R}^n$ ein stationärer Punkt, an dem die Hesse-Matrix $\nabla^2 f(\mathbf{x}^\star)$ positiv definit ist. Dann existiert eine Umgebung $U \subset \mathbb{R}^n$ von $\mathbf{x}^\star$, so dass für alle $\mathbf{x}_0 \in U$ gelten:

(*i.*) Ist $\eta_k \leq \eta$ für ein hinreichend kleines η, dann ist Algorithmus 3.8 wohldefiniert und die durch ihn erzeugte Folge $\{\mathbf{x}_k\}_{k\geq 0}$ konvergiert mindestens linear gegen $\mathbf{x}^\star$.

(*ii.*) Die Konvergenzrate ist superlinear, falls $\eta_k \to 0$ gilt.

(*iii.*) Die Konvergenzrate ist quadratisch, falls $\eta_k = \mathcal{O}(\|\nabla f(\mathbf{x}_k)\|_2)$ gilt und $\nabla^2 f$ lokal Lipschitz-stetig ist.

Beweis. Da f zweimal stetig differenzierbar ist, ist ∇f lokal Lipschitz-stetig. Daher existiert eine Umgebung $U \subset \mathbb{R}^n$ von $\mathbf{x}^\star$ und ein $L > 0$ derart, dass

$$\|\nabla f(\mathbf{x})\|_2 = \|\nabla f(\mathbf{x}) - \nabla f(\mathbf{x}^\star)\|_2 \leq L\|\mathbf{x} - \mathbf{x}^\star\|_2 \quad \text{für alle } \mathbf{x} \in U.$$

Da $\nabla^2 f(\mathbf{x}^\star)$ invertierbar ist, folgt, indem wir U eventuell verkleinern, aufgrund der Stetigkeit

$$\left\|\left(\nabla^2 f(\mathbf{x})\right)^{-1}\right\|_2 \leq c \quad \text{für alle } \mathbf{x} \in U.$$

Aus der Definition der Ableitung einer Funktion schließen wir, indem wir wieder U eventuell verkleinern, dass

$$\underbrace{\|\nabla f(\mathbf{x}^\star) - \nabla f(\mathbf{x}) - \nabla^2 f(\mathbf{x})(\mathbf{x} - \mathbf{x}^\star)\|_2}_{=\mathcal{O}(\|\mathbf{x}-\mathbf{x}^\star\|_2)} \leq \frac{1}{4c}\|\mathbf{x} - \mathbf{x}^\star\|_2.$$

Setze nun $\eta := 1/(4cL)$ und wähle $\mathbf{x}_0 \in U$. Dann ist $\nabla^2 f(\mathbf{x}_0)$ regulär und es lässt sich ein $\mathbf{d}_0$ mit (3.11) berechnen. Somit ist $\mathbf{x}_1$ wohldefiniert und wegen

$$\begin{aligned}
\|\mathbf{x}_1 - \mathbf{x}^\star\|_2 &= \left\|\mathbf{x}_0 - \mathbf{x}^\star - \left(\nabla^2 f(\mathbf{x}_0)\right)^{-1}\nabla f(\mathbf{x}_0) + \left(\nabla^2 f(\mathbf{x}_0)\right)^{-1}\left\{\nabla^2 f(\mathbf{x}_0)\mathbf{d}_0 + \nabla f(\mathbf{x}_0)\right\}\right\|_2 \\
&\leq \left\|\left(\nabla^2 f(\mathbf{x}_0)\right)^{-1}\right\|_2 \Big\{\left\|\nabla^2 f(\mathbf{x}_0)(\mathbf{x}_0 - \mathbf{x}^\star) - \nabla f(\mathbf{x}_0) + \nabla f(\mathbf{x}^\star)\right\|_2 \\
&\qquad\qquad + \underbrace{\left\|\nabla^2 f(\mathbf{x}_0)\mathbf{d}_0 + \nabla f(\mathbf{x}_0)\right\|_2}_{\leq \eta\|\nabla^2 f(\mathbf{x}_0)\|_2}\Big\} \\
&\leq c\left(\frac{1}{4c} + \eta L\right)\|\mathbf{x}_0 - \mathbf{x}^\star\|_2 \\
&= \frac{1}{2}\|\mathbf{x}_0 - \mathbf{x}^\star\|_2
\end{aligned}$$

wieder in U enthalten. Mit vollständiger Induktion schließen wir

$$\|\mathbf{x}_k - \mathbf{x}^\star\|_2 \leq \frac{1}{2^k}\|\mathbf{x}_0 - \mathbf{x}^\star\|_2,$$

was die lineare Konvergenz $\mathbf{x}_k \to \mathbf{x}^\star$ nachweist.

Zum Nachweis von (*ii.*) bemerken wir, dass analog zur obigen Ungleichungskette auch

$$\begin{aligned}
\|\mathbf{x}_{k+1} - \mathbf{x}^\star\|_2 &\leq c\Big\{\left\|\nabla^2 f(\mathbf{x}_k)(\mathbf{x}_k - \mathbf{x}^\star) - \nabla f(\mathbf{x}_k) + \nabla f(\mathbf{x}^\star)\right\|_2 + \eta_k\|\nabla^2 f(\mathbf{x}_k)\|_2\Big\} \\
&\leq c\Big\{\left\|\nabla^2 f(\mathbf{x}_k)(\mathbf{x}_k - \mathbf{x}^\star) - \nabla f(\mathbf{x}_k) + \nabla f(\mathbf{x}^\star)\right\|_2 + \eta_k L\|\mathbf{x}_k - \mathbf{x}^\star\|_2\Big\}
\end{aligned}$$

bewiesen werden kann. Für $\eta_k \to 0$ ist daher

$$\|\mathbf{x}_{k+1} - \mathbf{x}^\star\|_2 = o(\|\mathbf{x}_k - \mathbf{x}^\star\|_2),$$

das heißt, $\{\mathbf{x}_k\}_{k\geq 0}$ konvergiert superlinear gegen $\mathbf{x}^\star$.
Die lokal quadratische Konvergenz wird ganz ähnlich nachgewiesen: Mit

$$\nabla f(\mathbf{x}_k) - \nabla f(\mathbf{x}^\star) = \int_0^1 \nabla^2 f\big(\mathbf{x}^\star + t(\mathbf{x}_k - \mathbf{x}^\star)\big)(\mathbf{x}_k - \mathbf{x}^\star)\,\mathrm{d}t$$

folgt mittels der lokalen Lipschitzstetigkeit

$$\begin{aligned}
&\big\|\nabla^2 f(\mathbf{x}_k)(\mathbf{x}_k - \mathbf{x}^\star) - \nabla f(\mathbf{x}_k) + \nabla f(\mathbf{x}^\star)\big\|_2 \\
&\quad = \left\| \int_0^1 \big\{\nabla^2 f(\mathbf{x}_k) - f\big(\mathbf{x}^\star + t(\mathbf{x}_k - \mathbf{x}^\star)\big)\big\}(\mathbf{x}_k - \mathbf{x}^\star)\,\mathrm{d}t \right\|_2 \\
&\quad \leq \int_0^1 \underbrace{\big\|\nabla^2 f(\mathbf{x}_k) - \nabla^2 f\big(\mathbf{x}^\star + t(\mathbf{x}_k - \mathbf{x}^\star)\big)\big\|_2}_{\leq tL\|\mathbf{x}_k - \mathbf{x}^\star\|_2} \|\mathbf{x}_k - \mathbf{x}^\star\|_2\,\mathrm{d}t \\
&\quad \leq \frac{L}{2}\|\mathbf{x}_k - \mathbf{x}^\star\|_2^2.
\end{aligned}$$

Setzt man dies zusammen mit $\eta_k = \mathcal{O}(\|\nabla f(\mathbf{x}_k)\|_2) = \mathcal{O}(\|\mathbf{x}_k - \mathbf{x}^\star\|_2)$ in die obige Abschätzung ein, dann ergibt sich

$$\|\mathbf{x}_{k+1} - \mathbf{x}^\star\|_2 = \mathcal{O}(\|\mathbf{x}_k - \mathbf{x}^\star\|_2^2),$$

das ist Aussage (*iii.*). □

3.4 Trust-Region-Verfahren

Die quadratische Approximation

$$f(\mathbf{x}_k + \mathbf{d}) \approx f(\mathbf{x}_k) + \nabla f(\mathbf{x}_k)^T\mathbf{d} + \frac{1}{2}\mathbf{d}^T\nabla^2 f(\mathbf{x}_k)\mathbf{d} = q_k(\mathbf{d})$$

ist nur für kleines $\|\mathbf{d}\|_2$ genau. Man schränkt daher den Bereich ein, in dem man der quadratischen Approximation $q_k(\mathbf{d})$ vertraut. Dies erklärt den Begriff *trust region*. Im Zusammenhang mit dem Newton-Verfahren wird das Update $\mathbf{d}_k$ beispielsweise aus

$$\min_{\mathbf{d}\in\mathbb{R}^n} q_k(\mathbf{d}) \text{ unter der Nebenbedingung } \|\mathbf{d}\|_2 \leq \Delta_k \tag{3.12}$$

bestimmt, wobei der Radius Δ_k noch geeignet zu bestimmen ist. Wir haben also in jedem Schritt ein quadratisches Minimierungsproblem mit Nebenbedingungen zu lösen.
Wir wenden uns der Wahl des Radius Δ_k zu. Dieser wird so gesteuert, dass eine möglichst gute Übereinstimmung der quadratischen Näherung q_k mit der Funktion $f(\mathbf{x}_k + \cdot)$ sichergestellt ist. Liegt das Qualitätsmaß

$$\rho_k = \frac{f(\mathbf{x}_k) - f(\mathbf{x}_k + \mathbf{d}_k)}{q_k(\mathbf{0}) - q_k(\mathbf{d}_k)}$$

nahe bei 1, stimmt die tatsächliche Reduktion (Zähler) mit der vorhergesagten Reduktion (Nenner) gut überein und der Radius Δ_k kann sogar vergrößert werden. Ist das Qualitätsmaß klein, dann sollte der Radius Δ_k verkleinert und der Schritt wiederholt werden. Ansonsten behält man Δ_k bei.

Algorithmus 3.10 (Trust-Region-Verfahren)
input: Funktion $f : \mathbb{R}^n \to \mathbb{R}$ und Startnäherung $\mathbf{x}_0 \in \mathbb{R}^n$
output: Folge von Iterierten $\{\mathbf{x}_k\}_{k\in\mathbb{N}}$

① Initialisierung: wähle $0 < \rho^- < \rho^+ < 1$ und setze $\Delta_0 := 1$ und $k := 0$

② bestimme $\mathbf{d}_k$ durch näherungsweise Lösung von (3.12) und berechne

$$\rho_k := \frac{f(\mathbf{x}_k) - f(\mathbf{x}_k + \mathbf{d}_k)}{q_k(\mathbf{0}) - q_k(\mathbf{d}_k)}$$

③ falls $\rho_k > \rho^-$ setze $\mathbf{x}_{k+1} := \mathbf{x}_k + \mathbf{d}_k$, sonst setze $\Delta_k := \Delta_k/2$ und gehe nach ②

④ falls $\rho_k > \rho^+$ setze $\Delta_{k+1} := 2\Delta_k$, sonst setze $\Delta_{k+1} := \Delta_k$

⑤ erhöhe $k := k + 1$ und gehe nach ②

Ausgangspunkt zur näherungsweisen Lösung von (3.12) ist der *Cauchy-Punkt*, dessen Konstruktion in zwei Schritten erfolgt.

1. Bestimme $\mathbf{d}_k$ aus

$$\min_{\mathbf{d}\in\mathbb{R}^n} f(\mathbf{x}_k) + \nabla f(\mathbf{x}_k)^T\mathbf{d} \text{ unter der Nebenbedingung } \|\mathbf{d}\|_2 \le \Delta_k.$$

Dieser Richtungsvektor lässt sich explizit angeben:

$$\mathbf{d}_k = -\frac{\Delta_k}{\|\nabla f(\mathbf{x}_k)\|_2}\nabla f(\mathbf{x}_k).$$

2. Bestimme τ_k aus

$$\min_{\tau>0} q_k(\tau\mathbf{d}_k) \text{ unter der Nebenbedingung } \|\tau\mathbf{d}_k\|_2 \le \Delta_k$$

und setze $\mathbf{d}_k^\star := \tau_k\mathbf{d}_k$. Für die Berechnung von τ_k sind zwei Fälle zu unterscheiden. Falls $\nabla f(\mathbf{x}_k)^T\nabla^2 f(\mathbf{x}_k)\nabla f(\mathbf{x}_k) \le 0$ gilt, so folgt $\tau_k = 1$. Ansonsten ist

$$\tau_k = \min\left\{1, \frac{1}{\Delta_k}\frac{\|\nabla f(\mathbf{x}_k)\|_2^3}{\nabla f(\mathbf{x}_k)^T\nabla^2 f(\mathbf{x}_k)\nabla f(\mathbf{x}_k)}\right\}.$$

Unser Ziel ist es nun, eine Kombination zu finden, die für kleine Werte Δ_k mit dem Cauchy-Punkt übereinstimmt und für größere Werte Δ_k in das Newton-Verfahren übergeht. Wir stellen dazu die sogenannte *Dogleg-Strategie* vor und verwenden zur Abkürzung statt (3.12)

$$q(\mathbf{d}) := f + \mathbf{g}^T\mathbf{d} + \frac{1}{2}\mathbf{d}^T\mathbf{H}\mathbf{d} \text{ unter der Nebenbedingung } \|\mathbf{d}\|_2 \le \Delta_k.$$

Für kleine Werte Δ_k fällt der quadratische Term nicht ins Gewicht und wir gehen in Richtung des Cauchy-Punktes $\mathbf{d}_C := -\frac{\|\mathbf{g}\|_2^2}{\mathbf{g}^T\mathbf{H}\mathbf{g}}\mathbf{g}$, für große Werte Δ_k wollen wir hingegen die Newton-Richtung $\mathbf{d}_N := -\mathbf{H}^{-1}\mathbf{g}$ verwenden:

$$\mathbf{d}(\tau) = \begin{cases} \tau\mathbf{d}_C, & \text{falls } 0 \le \tau \le 1, \\ \mathbf{d}_C + (\tau - 1)(\mathbf{d}_N - \mathbf{d}_C), & \text{falls } 1 \le \tau \le 2. \end{cases}$$

Lemma 3.11 Es sei $\mathbf{H}$ postitiv definit und $\mathbf{g} \neq \mathbf{0}$. Dann ist
(*i.*) die Funktion $\|\mathbf{d}(\tau)\|_2$ streng monoton wachsend in τ, falls $\mathbf{d}_N \neq \mathbf{d}_C$, und
(*ii.*) die Funktion $q\big(\mathbf{d}(\tau)\big)$ monoton fallend in τ.

Beweis. Beide Aussagen im Fall $0 \leq \tau \leq 1$ offensichtlich.
Für $1 \leq \tau := \sigma + 1 \leq 2$ betrachten wir die Funktion

$$\begin{aligned}\varphi(\sigma) &= \frac{1}{2}\|\mathbf{d}(1+\sigma)\|_2^2 \\ &= \frac{1}{2}\|\mathbf{d}_C + \sigma(\mathbf{d}_N - \mathbf{d}_C)\|_2^2 \\ &= \frac{1}{2}\|\mathbf{d}_C\|_2^2 + \sigma\mathbf{d}_C^T(\mathbf{d}_N - \mathbf{d}_C) + \frac{1}{2}\sigma^2\|\mathbf{d}_N - \mathbf{d}_C\|_2^2\end{aligned}$$

und zeigen $\varphi'(\sigma) > 0$ für alle $\sigma \in (0,1)$:

$$\begin{aligned}\varphi'(\sigma) &= \mathbf{d}_C^T(\mathbf{d}_N - \mathbf{d}_C) + \sigma \underbrace{\|\mathbf{d}_N - \mathbf{d}_C\|_2^2}_{>0 \text{ da } \mathbf{d}_N \neq \mathbf{d}_C} \\ &> \mathbf{d}_C^T(\mathbf{d}_N - \mathbf{d}_C) \\ &= \frac{\mathbf{g}^T\mathbf{g}}{\mathbf{g}^T\mathbf{H}\mathbf{g}}\mathbf{g}^T\left(\mathbf{H}^{-1}\mathbf{g} - \frac{\mathbf{g}^T\mathbf{g}}{\mathbf{g}^T\mathbf{H}\mathbf{g}}\mathbf{g}\right) \\ &= \frac{\mathbf{g}^T\mathbf{g}}{\mathbf{g}^T\mathbf{H}\mathbf{g}}\mathbf{g}^T\mathbf{H}^{-1}\mathbf{g}\left(1 - \frac{(\mathbf{g}^T\mathbf{g})^2}{(\mathbf{g}^T\mathbf{H}\mathbf{g})(\mathbf{g}^T\mathbf{H}^{-1}\mathbf{g})}\right) \\ &\geq 0,\end{aligned}$$

wobei sich die letzte Ungleichung aus

$$(\mathbf{g}^T\mathbf{g})^2 = (\mathbf{g}^T\mathbf{H}^{1/2}\mathbf{H}^{-1/2}\mathbf{g})^2 \leq \|\mathbf{H}^{1/2}\mathbf{g}\|_2^2\|\mathbf{H}^{-1/2}\mathbf{g}\|_2^2 = (\mathbf{g}^T\mathbf{H}\mathbf{g})(\mathbf{g}^T\mathbf{H}^{-1}\mathbf{g})$$

ergibt. Damit ist Aussage (*i.*) gezeigt.
Um Aussage (*ii.*) nachzuweisen, bilden wir $\psi(\sigma) = q\big(\mathbf{d}(1+\sigma)\big)$ und erhalten

$$\begin{aligned}\psi'(\sigma) &= (\mathbf{g} + \mathbf{H}\mathbf{d}_C)^T(\mathbf{d}_N - \mathbf{d}_C) + \sigma \underbrace{(\mathbf{d}_N - \mathbf{d}_C)^T\mathbf{H}(\mathbf{d}_N - \mathbf{d}_C)}_{\geq 0} \\ &\leq \big\{\mathbf{g} + \mathbf{H}\mathbf{d}_C + \mathbf{H}(\mathbf{d}_N - \mathbf{d}_C)\big\}^T(\mathbf{d}_N - \mathbf{d}_C) \\ &= (\mathbf{g} + \mathbf{H}\mathbf{d}_N)^T(\mathbf{d}_N - \mathbf{d}_C) = \mathbf{0}.\end{aligned}$$

□

Bemerkung Aus diesem Lemma folgt, dass es genau einen Punkt gibt, an dem der Dogleg-Pfad den Kreis mit Radius Δ_k schneidet, falls $\|\mathbf{d}_N\|_2 \geq \Delta_k$. Da $q\big(\mathbf{d}(\tau)\big)$ monoton fällt, nimmt man diesen Schittpunkt als Update. Ist hingegen $\|\mathbf{d}_N\|_2 \leq \Delta_k$, dann verläuft der Pfad ganz im Innern des Trust-Region-Bereichs. Folglich wählt man als Update

$$\mathbf{d}_k = \begin{cases}\mathbf{d}_N, & \text{falls } \|\mathbf{d}_N\|_2 \leq \Delta_k, \\ \mathbf{d}_S, & \text{falls } \|\mathbf{d}_C\|_2 < \Delta_k < \|\mathbf{d}_N\|_2, \\ \frac{\Delta_k}{\|\mathbf{d}_C\|_2}\mathbf{d}_C, & \text{falls } \|\mathbf{d}_C\|_2 \geq \Delta_k,\end{cases}$$

wobei

$$\mathbf{d}_S := \mathbf{d}_C + (\tau^\star - 1)(\mathbf{d}_N - \mathbf{d}_C) \text{ mit } \tau^\star = \operatorname*{argmin}_{\tau>0} \left\{ (\|\mathbf{d}_C + (\tau - 1)(\mathbf{d}_N - \mathbf{d}_C)\|_2 - \Delta_k)^2 \right\}.$$

Bei Implementierung muss man jedoch zuallererst überprüfen, ob überhaupt

$$q_k(\mathbf{d}_N) < q_k(\mathbf{0})$$

gilt. Denn ist $\nabla^2 f(\mathbf{x}_k)$ nicht positiv definit, muss dem nicht so sein. In diesem Fall ist die Newton-Richtung von vornherein zu verwerfen und der Cauchy-Punkt zu nehmen. △

Unser Ziel ist der Nachweis globaler Konvergenz des Newton-Verfahrens mit der vorgestellten Trust-Region-Technik. Ausgangspunkt dafür ist die Reduktion des Funktionals durch die Verwendung des Cauchy-Punktes.

Lemma 3.12 Für den Cauchy-Punkt $\mathbf{d}_k^\star$ gilt

$$q_k(\mathbf{d}_k^\star) \leq q_k(\mathbf{0}) - \frac{1}{2}\|\nabla f(\mathbf{x}_k)\|_2 \min\left\{\Delta_k, \frac{\|\nabla f(\mathbf{x}_k)\|_2}{\|\nabla^2 f(\mathbf{x}_k)\|_2}\right\}.$$

Beweis. Wir müssen drei Fälle unterscheiden.
(*i.*) Falls $\nabla f(\mathbf{x}_k)^T \nabla^2 f(\mathbf{x}_k) \nabla f(\mathbf{x}_k) \leq 0$ folgt

$$\begin{aligned} q_k(\mathbf{d}_k^\star) - q_k(\mathbf{0}) &= -\frac{\Delta_k}{\|\nabla f(\mathbf{x}_k)\|_2}\|\nabla f(\mathbf{x}_k)\|_2^2 + \frac{1}{2}\frac{\Delta_k^2 \nabla f(\mathbf{x}_k)^T \nabla^2 f(\mathbf{x}_k) \nabla f(\mathbf{x}_k)}{\|\nabla f(\mathbf{x}_k)\|_2^2} \\ &\leq -\Delta_k \|\nabla f(\mathbf{x}_k)\|_2 \\ &\leq -\frac{1}{2}\|\nabla f(\mathbf{x}_k)\|_2 \min\left\{\Delta_k, \frac{\|\nabla f(\mathbf{x}_k)\|_2}{\|\nabla^2 f(\mathbf{x}_k)\|_2}\right\}. \end{aligned}$$

(*ii.*) Im Fall $0 < \nabla f(\mathbf{x}_k)^T \nabla^2 f(\mathbf{x}_k) \nabla f(\mathbf{x}_k) \leq \|\nabla f(\mathbf{x}_k)\|_2^3 / \Delta_k$ ergibt sich

$$\begin{aligned} q_k(\mathbf{d}_k^\star) - q_k(\mathbf{0}) &= -\frac{\Delta_k}{\|\nabla f(\mathbf{x}_k)\|_2}\|\nabla f(\mathbf{x}_k)\|_2^2 + \frac{1}{2}\frac{\Delta_k^2 \nabla f(\mathbf{x}_k)^T \nabla^2 f(\mathbf{x}_k) \nabla f(\mathbf{x}_k)}{\|\nabla f(\mathbf{x}_k)\|_2^2} \\ &\leq -\Delta_k \|\nabla f(\mathbf{x}_k)\|_2 + \frac{1}{2}\frac{\Delta_k^2}{\|\nabla f(\mathbf{x}_k)\|_2^2}\frac{\|\nabla f(\mathbf{x}_k)\|_2^3}{\Delta_k} \\ &= -\frac{1}{2}\Delta_k \|\nabla f(\mathbf{x}_k)\|_2 \\ &\leq -\frac{1}{2}\|\nabla f(\mathbf{x}_k)\|_2 \min\left\{\Delta_k, \frac{\|\nabla f(\mathbf{x}_k)\|_2}{\|\nabla^2 f(\mathbf{x}_k)\|_2}\right\}. \end{aligned}$$

(*iii.*) Ist schließlich $\nabla f(\mathbf{x}_k)^T \nabla^2 f(\mathbf{x}_k) \nabla f(\mathbf{x}_k) > \|\nabla f(\mathbf{x}_k)\|_2^3 / \Delta_k$, so liegt der Cauchy-Punkt im Innern des Trust-Region-Bereichs und ist durch

$$\mathbf{d}_k^\star = -\frac{\|\nabla f(\mathbf{x}_k)\|_2^2}{\nabla f(\mathbf{x}_k)^T \nabla^2 f(\mathbf{x}_k) \nabla f(\mathbf{x}_k)} \nabla f(\mathbf{x}_k)$$

gegeben. Somit gilt

$$\begin{aligned} q_k(\mathbf{d}_k^\star) - q_k(\mathbf{0}) &= -\frac{\|\nabla f(\mathbf{x}_k)\|_2^4}{\nabla f(\mathbf{x}_k)^T \nabla^2 f(\mathbf{x}_k) \nabla f(\mathbf{x}_k)} \\ &\quad + \frac{1}{2} \frac{\|\nabla f(\mathbf{x}_k)\|_2^4}{(\nabla f(\mathbf{x}_k)^T \nabla^2 f(\mathbf{x}_k) \nabla f(\mathbf{x}_k))^2} \nabla f(\mathbf{x}_k)^T \nabla^2 f(\mathbf{x}_k) \nabla f(\mathbf{x}_k) \\ &= -\frac{1}{2} \frac{\|\nabla f(\mathbf{x}_k)\|_2^4}{\nabla f(\mathbf{x}_k)^T \nabla^2 f(\mathbf{x}_k) \nabla f(\mathbf{x}_k)} \\ &\leq -\frac{1}{2} \frac{\|\nabla f(\mathbf{x}_k)\|_2^4}{\|\nabla f(\mathbf{x}_k)\|_2^2 \|\nabla^2 f(\mathbf{x}_k)\|_2} \\ &\leq -\frac{1}{2} \|\nabla f(\mathbf{x}_k)\|_2 \min \left\{ \Delta_k, \frac{\|\nabla f(\mathbf{x}_k)\|_2}{\|\nabla^2 f(\mathbf{x}_k)\|_2} \right\}. \end{aligned}$$

□

Bemerkung Mit der Dogleg-Strategie erreicht man mindestens eine Reduktion des Funktionals $q_k(\mathbf{d})$ durch den Cauchy-Punkt, das heißt, es gilt

$$q_k(\mathbf{d}_k) \leq q_k(\mathbf{0}) - \frac{1}{2} \|\nabla f(\mathbf{x}_k)\|_2 \min \left\{ \Delta_k, \frac{\|\nabla f(\mathbf{x}_k)\|_2}{\|\nabla^2 f(\mathbf{x}_k)\|_2} \right\}.$$

△

Satz 3.13 Auf der Niveaumenge $N := \{\mathbf{x} \in \mathbb{R}^n : f(\mathbf{x}) \leq f(\mathbf{x}_0)\}$ sei f zweimal stetig differenzierbar und nach unten beschränkt. Ist $\nabla^2 f(\mathbf{x})$ beschränkt auf der gesamten Niveaumenge N, dann erfüllen die vom Trust-Region-Verfahren 3.10 mit der Dogleg-Strategie berechneten Iterierten $\{\mathbf{x}_k\}_{k \geq 0}$

$$\nabla f(\mathbf{x}_k) \overset{k \to \infty}{\longrightarrow} \mathbf{0}.$$

Beweis. Wir vollziehen den Beweis des Satzes in vier Schritten.
(*i.*) Für das Maß ρ_k zur Anpassung des Trust-Region-Radius gilt

$$|\rho_k - 1| = \left| \frac{f(\mathbf{x}_k) - f(\mathbf{x}_k + \mathbf{d}_k) - \{q_k(\mathbf{0}) - q_k(\mathbf{d}_k)\}}{q_k(\mathbf{0}) - q_k(\mathbf{d}_k)} \right| = \left| \frac{f(\mathbf{x}_k + \mathbf{d}_k) - q_k(\mathbf{d}_k)}{q_k(\mathbf{0}) - q_k(\mathbf{d}_k)} \right|.$$

Die Taylor-Entwicklung von f um $\mathbf{x}_k$ liefert

$$f(\mathbf{x}_k + \mathbf{d}_k) = f(\mathbf{x}_k) + \nabla f(\mathbf{x}_k)^T \mathbf{d}_k + \frac{1}{2} \mathbf{d}_k^T \nabla^2 f(\mathbf{x}_k + \xi \mathbf{d}_k) \mathbf{d}_k$$

für ein $\xi \in (0, 1)$. Damit ergibt sich

$$\begin{aligned} |q_k(\mathbf{d}_k) - f(\mathbf{x}_k + \mathbf{d}_k)| &= \left| \frac{1}{2} \mathbf{d}_k^T \nabla^2 f(\mathbf{x}_k) \mathbf{d}_k - \frac{1}{2} \mathbf{d}_k^T \nabla^2 f(\mathbf{x}_k + \xi \mathbf{d}_k) \mathbf{d}_k \right| \\ &\leq \frac{1}{2} \left\{ \|\nabla^2 f(\mathbf{x}_k)\|_2 + \max_{0 \leq t \leq 1} \|\nabla^2 f(\mathbf{x}_k + t \mathbf{d}_k)\|_2 \right\} \|\mathbf{d}_k\|_2^2 \\ &\leq \overline{c} \|\mathbf{d}_k\|_2^2, \end{aligned}$$

wobei $\overline{c} > 0$ eine obere Schranke für $\nabla^2 f(\mathbf{x})$ auf der Niveau-Menge N sei. Da nach Lemma 3.12 gilt

$$q_k(\mathbf{d}_k) \leq q_k(\mathbf{0}) - \frac{1}{2}\|\nabla f(\mathbf{x}_k)\|_2 \min\left\{\Delta_k, \frac{\|\nabla f(\mathbf{x}_k)\|_2}{\overline{c}}\right\}, \tag{3.13}$$

erhalten wir folglich die Abschätzung

$$|\rho_k - 1| \leq \frac{\overline{c}\|\mathbf{d}_k\|_2^2}{\|\nabla f(\mathbf{x}_k)\|_2 \min\{\Delta_k, \frac{\|\nabla f(\mathbf{x}_k)\|_2}{\overline{c}}\}} \leq \frac{2\overline{c}\Delta_k^2}{\|\nabla f(\mathbf{x}_k)\|_2\{\Delta_k, \frac{\|\nabla f(\mathbf{x}_k)\|_2}{\overline{c}}\}}. \tag{3.14}$$

(*ii.*) Bei erfolgreichem Iterationsschritt ist $f(\mathbf{x}_k) - f(\mathbf{x}_k + \mathbf{d}_k) \geq \rho^-\big\{q_k(\mathbf{0}) - q_k(\mathbf{d}_k)\big\}$, das heißt, gemäß (3.13) gilt

$$f(\mathbf{x}_k) - f(\mathbf{x}_k + \mathbf{d}_k) \geq \frac{\rho^-}{2}\|\nabla f(\mathbf{x}_k)\|_2 \min\left\{\Delta_k, \frac{\|\nabla f(\mathbf{x}_k)\|_2}{\overline{c}}\right\}. \tag{3.15}$$

Da die Folge $\{f(\mathbf{x}_k)\}_{k\geq 0}$ nach Konstruktion streng monoton fällt und nach unten beschränkt ist, muss gelten

$$\min\left\{\Delta_k, \frac{\|\nabla f(\mathbf{x}_k)\|_2}{\overline{c}}\right\} \overset{k\to\infty}{\longrightarrow} 0. \tag{3.16}$$

(*iii.*) Wir zeigen nun, dass $\{\|\nabla f(\mathbf{x}_k)\|_2\}_{k\geq 0}$ für eine unendliche Teilfolge $\{k_\ell\}_{\ell\geq 0}$ gegen Null konvergiert. Angenommen, die Behauptung gilt nicht, dann folgt

$$\|\nabla f(\mathbf{x}_k)\|_2 \geq \varepsilon > 0 \quad \text{für alle } k \geq K(\varepsilon).$$

Aus (3.16) ergibt sich unmittelbar

$$\Delta_k \overset{k\to\infty}{\longrightarrow} 0,$$

weshalb (3.14) impliziert $|\rho_k - 1| \to 0$ für $k \to \infty$, beziehungsweise

$$\rho_k \overset{k\to\infty}{\longrightarrow} 1.$$

Demnach existiert ein $M(\varepsilon) \geq K(\varepsilon)$, so dass $\rho_k \geq \rho^+$ für alle $k \geq M(\varepsilon)$. Ab dem $M(\varepsilon)$-ten Schritt wird folglich Δ_k in jedem Schritt von Algorithmus 3.10 verdoppelt, was jedoch im Widerspruch zu (3.16) steht.

(*iv.*) Wir beweisen nun die Aussage des Satzes. Dazu nehmen wir an, dass eine Teilfolge von $\{\|\nabla f(\mathbf{x}_k)\|_2\}_{k\geq 0}$ nicht gegen Null konvergiert. Nach Aussage (*iii.*) existiert dann ein $\varepsilon > 0$ und zwei Indizes $\ell < m$, so dass

$$\|\nabla f(\mathbf{x}_\ell)\|_2 \geq 2\varepsilon, \qquad \|\nabla f(\mathbf{x}_m)\|_2 \leq \varepsilon, \qquad \|\nabla f(\mathbf{x}_k)\|_2 > \varepsilon, \quad k = \ell+1, \ldots, m-1.$$

Da $\{f(\mathbf{x}_k)\}_{k\geq 0}$ eine Cauchy-Folge ist, kann ℓ dabei so groß gewählt werden, dass

$$f(\mathbf{x}_\ell) - f(\mathbf{x}_m) < \frac{\varepsilon^2\rho^-}{2\overline{c}}. \tag{3.17}$$

Wegen $\|\mathbf{x}_{k+1} - \mathbf{x}_k\|_2 = \|\mathbf{d}_k\|_2 \leq \Delta_k$, folgt aus (3.15), dass

$$f(\mathbf{x}_k) - f(\mathbf{x}_{k+1}) \geq \frac{\varepsilon\rho^-}{2} \min\left\{\|\mathbf{x}_{k+1} - \mathbf{x}_k\|_2, \frac{\varepsilon}{\overline{c}}\right\}, \quad k = \ell, \ell+1, \ldots, m-1.$$

Summation ergibt wegen (3.17)

$$\frac{\varepsilon\rho^-}{2}\sum_{k=\ell}^{m-1}\min\left\{\|\mathbf{x}_{k+1}-\mathbf{x}_k\|_2,\frac{\varepsilon}{\overline{c}}\right\}\leq f(\mathbf{x}_\ell)-f(\mathbf{x}_m)<\frac{\varepsilon^2\rho^-}{2\overline{c}},$$

was nur erfüllt sein kann, wenn

$$\min\left\{\|\mathbf{x}_{k+1}-\mathbf{x}_k\|_2,\frac{\varepsilon}{\overline{c}}\right\}=\|\mathbf{x}_{k+1}-\mathbf{x}_k\|_2,\quad k=\ell,\ell+1,\ldots,m-1,$$

und insgesamt

$$\sum_{k=\ell}^{m-1}\|\mathbf{x}_{k+1}-\mathbf{x}_k\|_2<\frac{\varepsilon}{\overline{c}}$$

gilt. Da $\nabla f(\mathbf{x})$ auf der Niveaumenge N Lipschitz-stetig zur Konstante $\overline{c}$ ist, ergibt dies schließlich

$$\|\nabla f(\mathbf{x}_m)-\nabla f(\mathbf{x}_\ell)\|_2\leq\overline{c}\|\mathbf{x}_m-\mathbf{x}_\ell\|_2\leq\overline{c}\sum_{k=\ell}^{m-1}\|\mathbf{x}_{k+1}-\mathbf{x}_k\|_2<\varepsilon$$

im Widerspruch zur Annahme. Damit ist der Satz bewiesen. □

4. Quasi-Newton-Verfahren

Beim Newton-Verfahren ist das Update $\mathbf{d}_k$ durch die Newton-Gleichung $\nabla^2 f(\mathbf{x}_k)\mathbf{d}_k = -\nabla f(\mathbf{x}_k)$ gegeben. Da das Berechnen der Hesse-Matrix und das Lösen dieses Gleichungssystems oftmals zu teuer ist, versucht man, $\big(\nabla^2 f(\mathbf{x}_k)\big)^{-1}$ durch einfach zu berechnende Matrizen $\mathbf{H}_k$ zu ersetzen und die Suchrichtung

$$\mathbf{d}_k := -\mathbf{H}_k \nabla f(\mathbf{x}_k)$$

zu benutzen. Man spricht von einem *Quasi-Newton-Verfahren*, wenn für alle $k \geq 0$ die Matrix $\mathbf{H}_{k+1}$ der *Quasi-Newton-Gleichung*

$$\mathbf{H}_{k+1}\big\{\nabla f(\mathbf{x}_{k+1}) - \nabla f(\mathbf{x}_k)\big\} = \mathbf{x}_{k+1} - \mathbf{x}_k \tag{4.1}$$

genügt. Diese Bedingung stellt sicher, dass sich $\mathbf{H}_{k+1}$ in der Richtung $\mathbf{x}_{k+1} - \mathbf{x}_k$ ähnlich wie die Newton-Matrix $\big(\nabla^2 f(\mathbf{x}_k)\big)^{-1}$ verhält, für die gilt

$$\nabla f(\mathbf{x}_{k+1}) - \nabla f(\mathbf{x}_k) = \nabla^2 f(\mathbf{x}_k)(\mathbf{x}_{k+1} - \mathbf{x}_k) + \mathcal{O}(\|\mathbf{x}_{k+1} - \mathbf{x}_k\|_2^2).$$

Für eine quadratische Funktion $q(\mathbf{x}) = \frac{1}{2}\mathbf{x}^T\mathbf{A}\mathbf{x} + \mathbf{b}^T\mathbf{x} + c$ mit positiv definiter Matrix $\mathbf{A}$ gilt (4.1) wegen $\nabla q(\mathbf{x}) = \mathbf{A}\mathbf{x} + \mathbf{b}$ sogar exakt. Ferner erscheint es sinnvoll, als $\mathbf{H}_k$ nur positiv definite Matrizen zu wählen. Dies garantiert, dass für $\nabla f(\mathbf{x}_k) \neq \mathbf{0}$ die Richtung $\mathbf{d}_k = -\mathbf{H}_k \nabla f(\mathbf{x}_k)$ eine Abstiegsrichtung von f wird

$$\nabla f(\mathbf{x}_k)^T \mathbf{d}_k = -\nabla f(\mathbf{x}_k)^T \mathbf{H}_k \nabla f(\mathbf{x}_k) < 0.$$

Beide Forderungen lassen sich erfüllen: Mit den Abkürzungen

$$\mathbf{p}_k = \mathbf{x}_{k+1} - \mathbf{x}_k, \quad \mathbf{q}_k = \nabla f(\mathbf{x}_{k+1}) - \nabla f(\mathbf{x}_k)$$

und frei wählbaren Parametern

$$\gamma_k > 0, \quad \nu_k \geq 0$$

ist $\mathbf{H}_{k+1}$ rekursiv gegeben durch

$$\begin{aligned} \mathbf{H}_{k+1} &:= \Phi(\mathbf{H}_k, \mathbf{p}_k, \mathbf{q}_k, \gamma_k, \nu_k), \\ \Phi(\mathbf{H}, \mathbf{p}, \mathbf{q}, \gamma, \nu) &:= \gamma\mathbf{H} + \left(1 + \gamma\nu\frac{\mathbf{q}^T\mathbf{H}\mathbf{q}}{\mathbf{p}^T\mathbf{q}}\right)\frac{\mathbf{p}\mathbf{p}^T}{\mathbf{p}^T\mathbf{q}} - \gamma\frac{1-\nu}{\mathbf{q}^T\mathbf{H}\mathbf{q}}\mathbf{H}\mathbf{q}\mathbf{q}^T\mathbf{H} \\ &\quad - \frac{\gamma\nu}{\mathbf{p}^T\mathbf{q}}(\mathbf{p}\mathbf{q}^T\mathbf{H} + \mathbf{H}\mathbf{q}\mathbf{p}^T). \end{aligned} \tag{4.2}$$

Die Update-Funktion Φ ist nur für $\mathbf{p}^T\mathbf{q} \neq 0$ und $\mathbf{q}^T\mathbf{H}\mathbf{q} \neq 0$ erklärt. Man beachte, dass man $\mathbf{H}_{k+1}$ aus $\mathbf{H}_k$ dadurch erhält, dass man zur Matrix $\gamma_k\mathbf{H}_k$ eine Korrekturmatrix vom Rang ≤ 2 addiert:

$$\operatorname{rang}(\mathbf{H}_{k+1} - \gamma_k\mathbf{H}_k) \leq 2.$$

Man nennt dieses Verfahren daher auch *Rang-2-Verfahren*.

Folgende Spezialfälle sind in (4.2) enthalten:

1. $\gamma_k \equiv 1$, $\nu_k \equiv 0$: Verfahren von Davidon, Fletcher und Powell (*DFP-Verfahren*).
2. $\gamma_k \equiv 1$, $\nu_k \equiv 1$: Rang-2-Verfahren von Broydon, Fletcher, Goldfarb und Shanno (*BFGS-Verfahren*).
3. $\gamma_k \equiv 1$, $\nu_k = \mathbf{p}_k^T\mathbf{q}_k/(\mathbf{p}_k^T\mathbf{q}_k - \mathbf{p}_k^T\mathbf{H}_k\mathbf{q}_k)$: *symmetrisches Rang-1-Verfahren von Broydon.*

Letzteres Verfahren ist nur für $\mathbf{p}_k^T\mathbf{q}_k \neq \mathbf{p}_k^T\mathbf{H}_k\mathbf{q}_k$ definiert; $\nu_k < 0$ ist möglich: in diesem Fall kann $\mathbf{H}_{k+1}$ auch indefinit werden, auch wenn $\mathbf{H}_k$ positiv definit ist (vergleiche Satz 4.2). Setzt man den gewählten Wert in (4.2) ein, erhält man für $\mathbf{H}_k$ eine Rekursionformel, die den Namen Rang-1-Verfahren erklärt:

$$\mathbf{H}_{k+1} := \mathbf{H}_k + \frac{\mathbf{z}_k\mathbf{z}_k^T}{\alpha_k}, \quad \mathbf{z}_k := \mathbf{p}_k - \mathbf{H}_k\mathbf{q}_k, \quad \alpha_k := \mathbf{p}_k^T\mathbf{q}_k - \mathbf{q}_k^T\mathbf{H}_k\mathbf{q}_k.$$

Algorithmus 4.1 (Quasi-Newton-Verfahren)
input: Funktion $f : \mathbb{R}^n \to \mathbb{R}$ und Startnäherung $\mathbf{x}_0 \in \mathbb{R}^n$
output: Folge von Iterierten $\{\mathbf{x}_k\}_{k\in\mathbb{N}}$

① Initialisierung: setze $\mathbf{H}_0 := \mathbf{I}$ und $k := 0$
② berechne die Quasi-Newton-Richtung $\mathbf{d}_k = -\mathbf{H}_k\nabla f(\mathbf{x}_k)$
③ löse

$$\alpha_k \approx \operatorname*{argmin}_{\alpha\in\mathbb{R}} f(\mathbf{x}_k + \alpha\mathbf{d}_k)$$

④ setze $\mathbf{x}_{k+1} := \mathbf{x}_k + \alpha_k\mathbf{d}_k$, $\mathbf{p}_k := \mathbf{x}_{k+1} - \mathbf{x}_k$ und $\mathbf{q}_k := \nabla f(\mathbf{x}_{k+1}) - \nabla f(\mathbf{x}_k)$
⑤ wähle $\gamma_k > 0$, $\nu_k \geq 0$ und berechne $\mathbf{H}_{k+1} := \Phi(\mathbf{H}_k, \mathbf{p}_k, \mathbf{q}_k, \gamma_k, \nu_k)$ gemäß (4.2)
⑥ erhöhe $k := k + 1$ und gehe nach ②

Das Verfahren ist eindeutig durch die Wahl der Parameter γ_k, ν_k und die Minimierung in Schritt ③ fixiert. Die Minimierung $\mathbf{x}_k \mapsto \mathbf{x}_{k+1}$ und ihre Qualität kann man mit Hilfe eines Parameters σ_k beschreiben, der durch

$$\nabla f(\mathbf{x}_{k+1})^T\mathbf{d}_k = \sigma_k\nabla f(\mathbf{x}_k)^T\mathbf{d}_k = -\sigma_k\nabla f(\mathbf{x}_k)^T\mathbf{H}_k\nabla f(\mathbf{x}_k)$$

definiert ist. Falls $\mathbf{d}_k$ eine Abstiegsrichtung ist, das heißt $\nabla f(\mathbf{x}_k)^T\mathbf{d}_k < 0$, dann ist σ_k eindeutig bestimmt. Bei exakter Liniensuche ist $\sigma_k = 0$ wegen

$$\nabla f(\mathbf{x}_{k+1})^T\mathbf{d}_k = \varphi_k'(\alpha_k) = 0, \quad \text{wobei} \quad \varphi_k(\alpha) := f(\mathbf{x}_k + \alpha\mathbf{d}_k).$$

Wir setzen für das folgende

$$\sigma_k < 1 \tag{4.3}$$

voraus. Falls $\nabla f(\mathbf{x}_k) \neq \mathbf{0}$ und $\mathbf{H}_k$ positiv definit ist, folgt aus (4.3) $\alpha_k > 0$ und deshalb

$$\begin{aligned}\mathbf{q}_k^T\mathbf{p}_k &= \alpha_k\{\nabla f(\mathbf{x}_{k+1}) - \nabla f(\mathbf{x}_k)\}^T\mathbf{d}_k \\ &= \alpha_k(\sigma_k - 1)\nabla f(\mathbf{x}_k)^T\mathbf{d}_k \\ &= -\alpha_k(\sigma_k - 1)\nabla f(\mathbf{x}_k)^T\mathbf{H}_k\nabla f(\mathbf{x}_k) \\ &> 0,\end{aligned}$$

also auch $\mathbf{q}_k \neq \mathbf{0}$ und $\mathbf{q}_k^T\mathbf{H}_k\mathbf{q}_k > 0$. Die Matrix $\mathbf{H}_{k+1}$ ist damit durch (4.2) wohldefiniert.

Die Forderung (4.3) kann nur dann nicht erfüllt werden, wenn

$$\varphi_k'(\alpha) = \nabla f(\mathbf{x}_k + \alpha \mathbf{d}_k)^T \mathbf{d}_k \leq \nabla f(\mathbf{x}_k)^T \mathbf{d}_k = \varphi_k'(0) < 0$$

für alle $\alpha \geq 0$ gilt. Dann ist aber

$$f(\mathbf{x}_k + \alpha \mathbf{d}_k) - f(\mathbf{x}_k) = \int_0^\alpha \varphi_k'(t)\,\mathrm{d}t \leq \alpha \nabla f(\mathbf{x}_k)^T \mathbf{d}_k < 0 \quad \text{für alle } \alpha \geq 0,$$

so dass $f(\mathbf{x}_k + \alpha \mathbf{d}_k)$ für $\alpha \to \infty$ nicht nach unten beschränkt ist. Die Forderung (4.3) bedeutet also keine wesentliche Einschränkung. Damit ist bereits der erste Teil des folgenden Satzes gezeigt, der besagt, dass das Quasi-Newton-Verfahren 4.1 unsere oben aufgestellten Forderungen erfüllt.

Satz 4.2 Falls im Quasi-Newton-Verfahren 4.1 die Matrix $\mathbf{H}_k$ für ein $k \geq 0$ positiv definit ist, $\nabla f(\mathbf{x}_k) \neq \mathbf{0}$ und $\sigma_k < 1$ ist, dann ist für alle $\gamma_k > 0$, $\nu_k \geq 0$ die Matrix $\mathbf{H}_{k+1} := \Phi(\mathbf{H}_k, \mathbf{p}_k, \mathbf{q}_k, \gamma_k, \nu_k)$ wohldefiniert und wieder positiv definit. Insbesondere erfüllt sie die Quasi-Newton-Gleichung

$$\mathbf{H}_{k+1}\mathbf{q}_k = \mathbf{p}_k.$$

Beweis. Die Wohldefiniertheit von $\mathbf{H}_{k+1}$ haben wir bereits gezeigt, so dass wir nur noch die positive Definitheit nachweisen müssen. Sei $\mathbf{y} \in \mathbb{R}^n \setminus \{\mathbf{0}\}$ ein beliebiger Vektor und $\mathbf{H}_k = \mathbf{L}\mathbf{L}^T$ die Cholesky-Zerlegung von $\mathbf{H}_k$. Mit Hilfe der Vektoren

$$\mathbf{u} := \mathbf{L}^T\mathbf{y}, \quad \mathbf{v} := \mathbf{L}^T\mathbf{q}_k$$

lässt sich $\mathbf{y}^T\mathbf{H}_{k+1}\mathbf{y}$ wegen (4.2) so schreiben:

$$\begin{aligned}
\mathbf{y}^T\mathbf{H}_{k+1}\mathbf{y} &= \gamma_k \mathbf{u}^T\mathbf{u} + \left(1 + \gamma_k\nu_k \frac{\mathbf{v}^T\mathbf{v}}{\mathbf{p}_k^T\mathbf{q}_k}\right)\frac{(\mathbf{p}_k^T\mathbf{y})^2}{\mathbf{p}_k^T\mathbf{q}_k} - \gamma_k \frac{1-\nu_k}{\mathbf{v}^T\mathbf{v}}(\mathbf{u}^T\mathbf{v})^2 - \frac{2\gamma_k\nu_k}{\mathbf{p}_k^T\mathbf{q}_k}(\mathbf{p}_k^T\mathbf{y})(\mathbf{u}^T\mathbf{v}) \\
&= \gamma_k\left(\mathbf{u}^T\mathbf{u} - \frac{(\mathbf{u}^T\mathbf{v})^2}{\mathbf{v}^T\mathbf{v}}\right) + \frac{(\mathbf{p}_k^T\mathbf{y})^2}{\mathbf{p}_k^T\mathbf{q}_k} + \gamma_k\nu_k\left(\sqrt{\mathbf{v}^T\mathbf{v}}\frac{\mathbf{p}_k^T\mathbf{y}}{\mathbf{p}_k^T\mathbf{q}_k} - \frac{\mathbf{u}^T\mathbf{v}}{\sqrt{\mathbf{v}^T\mathbf{v}}}\right)^2 \\
&\geq \gamma_k\left(\mathbf{u}^T\mathbf{u} - \frac{(\mathbf{u}^T\mathbf{v})^2}{\mathbf{v}^T\mathbf{v}}\right) + \frac{(\mathbf{p}_k^T\mathbf{y})^2}{\mathbf{p}_k^T\mathbf{q}_k}.
\end{aligned}$$

Die Cauchy-Schwarzsche Ungleichung ergibt

$$\mathbf{u}^T\mathbf{u} - \frac{(\mathbf{u}^T\mathbf{v})^2}{\mathbf{v}^T\mathbf{v}} \geq 0,$$

mit Gleichheit genau dann, wenn $\mathbf{u} = \lambda\mathbf{v}$ für ein $\lambda \neq 0$ (wegen $\mathbf{y} \neq \mathbf{0}$). Für $\mathbf{u} \neq \lambda\mathbf{v}$ ist also $\mathbf{y}\mathbf{H}_{k+1}\mathbf{y} > 0$. Für $\mathbf{u} = \lambda\mathbf{v}$ folgt aus der Nichtsingularität von $\mathbf{H}_k$ und $\mathbf{L}$ auch $\mathbf{0} \neq \mathbf{y} = \lambda\mathbf{q}_k$, so dass

$$\mathbf{y}\mathbf{H}_{k+1}\mathbf{y} \geq \frac{(\mathbf{p}_k^T\mathbf{y})^2}{\mathbf{p}_k^T\mathbf{q}_k} = \lambda^2\mathbf{p}_k^T\mathbf{q}_k > 0.$$

Da $\mathbf{y} \in \mathbb{R}^n \setminus \{\mathbf{0}\}$ beliebig war, muss $\mathbf{H}_{k+1}$ positiv definit sein.

Die Quasi-Newton-Gleichung $\mathbf{H}_{k+1}\mathbf{q}_k = \mathbf{p}_k$ verifiziert man schließlich sofort mittels (4.2). □

Ein wesentliches Resultat ist, dass das Quasi-Newton-Verfahren im Fall einer quadratischen Funktion $f : \mathbb{R}^n \to \mathbb{R}$ das Minimum nach höchstens n Schritten liefert, sofern die Minimierung in ③ exakt sind. Da sich jede genügend oft differenzierbare Funktion f in der Nähe ihres Minimums beliebig genau durch eine quadratische Funktion approximieren lässt, lässt diese Eigenschaft vermuten, dass das Verfahren auch bei der Anwendung auf nichtquadratische Funktionen rasch konvergiert.

Satz 4.3 Sei

$$f(\mathbf{x}) = \frac{1}{2}\mathbf{x}^T\mathbf{A}\mathbf{x} + \mathbf{b}^T\mathbf{x} + c$$

eine quadratische Funktion mit einer positiv definiten Matrix $\mathbf{A} \in \mathbb{R}^{n\times n}$. Wendet man das Quasi-Newton-Verfahren 4.1 zur Minimierung von f mit den Startwerten $\mathbf{x}_0$ und $\mathbf{H}_0$ an, wobei man die Minimierungen in ③ exakt durchführt, so liefert das Verfahren Folgen $\{\mathbf{x}_k\}_{k\geq 0}$, $\{\mathbf{H}_k\}_{k\geq 0}$, $\{\nabla f(\mathbf{x}_k)\}_{k\geq 0}$, $\{\mathbf{p}_k\}_{k\geq 0}$ und $\{\mathbf{q}_k\}_{k\geq 0}$ mit den Eigenschaften:

(*i.*) Es gibt ein kleinstes $m \leq n$ mit $\mathbf{x}_m = \mathbf{x}^\star = -\mathbf{A}^{-1}\mathbf{b}$, das heißt, $\mathbf{x}_m$ ist das eindeutige Minimum von f, insbesondere gilt also $\nabla f(\mathbf{x}_m) = \mathbf{0}$.

(*ii.*) Es ist $\mathbf{p}_k^T\mathbf{q}_k > 0$ und $\mathbf{p}_k^T\mathbf{q}_\ell = \mathbf{p}_k^T\mathbf{A}\mathbf{p}_\ell = 0$ für alle $0 \leq k \neq \ell < m$. Die Vektoren $\mathbf{p}_k$ sind demnach $\mathbf{A}$-*konjugiert*.

(*iii.*) Es gilt $\mathbf{p}_k^T\nabla f(\mathbf{x}_\ell) = 0$ für alle $0 \leq k < \ell \leq m$.

(*iv.*) Es ist $\mathbf{H}_\ell\mathbf{q}_k = \lambda_{k,\ell}\mathbf{p}_k$ für alle $0 \leq k < \ell \leq m$ mit

$$\gamma_{k,\ell} := \begin{cases} \gamma_k\gamma_{k+1}\cdots\gamma_{\ell-1}, & \text{für } k < \ell - 1, \\ 1, & \text{für } k = \ell - 1. \end{cases}$$

(*v.*) Falls $m = n$, so gilt zusätzlich

$$\mathbf{H}_m = \mathbf{H}_n = \mathbf{P}\mathbf{D}\mathbf{P}^{-1}\mathbf{A}^{-1},$$

wobei

$$\mathbf{D} = \operatorname{diag}(\gamma_{0,n}, \gamma_{1,n}, \ldots, \gamma_{n-1,n}), \quad \mathbf{P} = [\mathbf{p}_0, \mathbf{p}_1, \ldots, \mathbf{p}_{n-1}].$$

Für $\gamma_k \equiv 1$ folgt $\mathbf{H}_n = \mathbf{A}^{-1}$.

Beweis. Wir zeigen zunächst induktiv, dass die Bedigungen (*ii.*)–(*iv.*) für ein beliebiges $m \geq 0$ gelten, falls für alle $j < m$ $\mathbf{H}_j$ positiv definit und $\nabla f(\mathbf{x}_j) \neq \mathbf{0}$ ist. Da die Aussagen für $m = 0$ trivialerweise erfüllt sind, können wir annehmen, dass sie für ein beliebiges $m \geq 0$ gelten. Der Induktionsschritt $m \mapsto m + 1$ ergibt sich nun wie folgt.
Da $\mathbf{H}_m$ positiv definit ist, folgt aus $\nabla f(\mathbf{x}_m) \neq \mathbf{0}$ sofort $\mathbf{d}_m = -\mathbf{H}_m\nabla f(\mathbf{x}_m) \neq \mathbf{0}$ und $\nabla f(\mathbf{x}_m)^T\mathbf{H}_m\nabla f(\mathbf{x}_m) > 0$. Weil exakt minimiert wird, ist α_m die Nullstelle von

$$0 = \nabla f(\mathbf{x}_{m+1})^T\mathbf{d}_m = \left\{\nabla f(\mathbf{x}_m) + \alpha_m\mathbf{A}\mathbf{d}_m\right\}^T\mathbf{d}_m, \quad \alpha_m = \frac{\nabla f(\mathbf{x}_m)\mathbf{H}_m\nabla f(\mathbf{x}_m)}{\mathbf{d}^T\mathbf{A}\mathbf{d}_m},$$

also $\mathbf{p}_m = \alpha_m\mathbf{d}_m$ und

$$\nabla f(\mathbf{x}_{m+1})^T\mathbf{p}_m = \alpha_m\nabla f(\mathbf{x}_{m+1})^T\mathbf{d}_m = 0. \tag{4.4}$$

Deshalb gilt

$$\begin{aligned}\mathbf{p}_m^T\mathbf{q}_m &= \alpha_m\mathbf{d}_m^T\big\{\nabla f(\mathbf{x}_{m+1}) - \nabla f(\mathbf{x}_m)\big\}\\ &= -\alpha_m\mathbf{d}_m^T\nabla f(\mathbf{x}_m)\\ &= \alpha_m\nabla f(\mathbf{x}_m)^T\mathbf{H}_m\nabla f(\mathbf{x}_m)\\ &> 0\end{aligned}$$

und folglich ist $\mathbf{H}_{m+1}$ nach Satz 4.2 positiv definit. Weiter ist für $k < m$ wegen $\mathbf{A}\mathbf{p}_k = \mathbf{q}_k$

$$\mathbf{p}_k^T\mathbf{q}_m = \mathbf{p}_k^T\mathbf{A}\mathbf{p}_m = \mathbf{q}_k^T\mathbf{p}_m = -\alpha_m\mathbf{q}_k^T\mathbf{H}_m\nabla f(\mathbf{x}_m) \overset{(iv.)}{=} -\alpha_m\gamma_{k,m}\mathbf{p}_k^T\nabla f(\mathbf{x}_m) \overset{(iii.)}{=} 0. \quad (4.5)$$

Das ist der Induktionsschritt für Aussage (*ii.*).
Weiter gilt für $k < m$

$$\mathbf{p}_k^T\nabla f(\mathbf{x}_{m+1}) = \mathbf{p}_k^T\left(\nabla f(\mathbf{x}_{k+1}) + \sum_{j=k+1}^{m}\mathbf{q}_j\right) = 0$$

nach dem eben bewiesenen und Aussage (*iii.*). Zusammen mit (4.4) ergibt dies Aussage (*iii.*) für $m+1$.
Den Induktionsschritt für Aussage (*iv.*) sieht man wie folgt ein. Anhand von (4.2) verifiziert man sofort

$$\mathbf{H}_{m+1}\mathbf{q}_m = \mathbf{p}_m.$$

Wegen Aussage (*ii.*) für $m+1$ und der Induktionsvoraussetzung hat man ferner für $k < m$

$$\mathbf{p}_m^T\mathbf{q}_k \overset{(ii.)}{=} \mathbf{0}, \quad \mathbf{q}_m^T\mathbf{H}_m\mathbf{q}_k \overset{(iv.)}{=} \gamma_{k,m}\mathbf{q}_m^T\mathbf{p}_m \overset{(ii.)}{=} \mathbf{0},$$

so dass für $k < m$ aus (4.2) folgt

$$\mathbf{H}_{m+1}\mathbf{q}_k = \gamma_m\mathbf{H}_m\mathbf{q}_k \overset{(iv.)}{=} \gamma_m\gamma_{k,m}\mathbf{p}_k = \gamma_{k,m+1}\mathbf{p}_k.$$

Der restliche Beweis ist nun einfach. Die Aussagen (*ii.*)–(*iv.*) können nur für $m \le n$ richtig sein, da die Vektoren $\mathbf{p}_0, \mathbf{p}_1, \ldots, \mathbf{p}_{m-1}$ linear unabhängig sind. Aus $\mathbf{0} = \sum_{\ell=0}^{m-1}\lambda_\ell\mathbf{p}_\ell$ folgt nämlich durch Multiplikation mit $\mathbf{p}_k^T\mathbf{A}$, $k = 0, 1, \ldots, m-1$, wegen Aussage (*ii.*) $\lambda_k\mathbf{p}_k^T\mathbf{A}\mathbf{p}_k = 0$, das heißt, $\lambda_k = 0$.
Da wir bewiesen haben, dass die Aussagen (*ii.*)–(*iv.*) für beliebiges m gelten, solange $\nabla f(\mathbf{x}_m) \neq \mathbf{0}$ ist, muss es also einen ersten Index $m \le n$ geben mit

$$\nabla f(\mathbf{x}_m) = \mathbf{0}, \quad \mathbf{x}_m = -\mathbf{A}^{-1}\mathbf{b},$$

dies bedeutet, es gilt Aussage (*i.*).
Für den Fall $m = n$ gilt wegen Aussage (*iv.*) zusätzlich $\mathbf{H}_n\mathbf{Q} = \mathbf{P}\mathbf{D}$ für die Matrizen

$$\mathbf{D} = \operatorname{diag}(\gamma_{0,n}, \gamma_{1,n}, \ldots, \gamma_{n-1,n}), \quad \mathbf{P} = [\mathbf{p}_0, \mathbf{p}_1, \ldots, \mathbf{p}_{n-1}], \quad \mathbf{Q} = [\mathbf{q}_0, \mathbf{q}_1, \ldots, \mathbf{q}_{n-1}].$$

Wegen $\mathbf{A}\mathbf{P} = \mathbf{Q}$ ergibt sich schließlich wegen der Nichtsingularität der Matrix $\mathbf{P}$ die Beziehung

$$\mathbf{H}_n = \mathbf{P}\mathbf{D}\mathbf{P}^{-1}\mathbf{A}^{-1},$$

Damit ist der Satz vollständig bewiesen. □

Es stellt sich nun die Frage, wie man die Parameter γ_k und ν_k wählen soll, um ein möglichst gutes Verfahren zu erhalten. Aussage $(v.)$ aus Satz 4.3 legt die Wahl $\gamma_k \equiv 1$ nahe, weil dies $\mathbf{D} = \mathbf{I}$ und folglich $\lim_m \mathbf{H}_m = \left(\nabla^2 f(\mathbf{x}^\star)\right)^{-1}$ vermuten lässt, weshalb das Verfahren voraussichtlich ähnlich schnell wie ein Newton-Verfahren konvergiert. Im allgemeinen ist diese Vermutung für nichtquadratische Funktionen aber nur unter zusätzlichen Voraussetzungen richtig. Nach praktischen Erfahrungen ist die Wahl

$$\gamma_k \equiv 1, \quad \nu_k \equiv 1 \quad \text{(BFGS-Verfahren)}$$

am besten.

Bemerkungen

1. Sowohl das DFP-Verfahren als auch das BFGS-Verfahren konvergieren superlinear in der Umgebung eines lokalen Minimus $\mathbf{x}^\star$, falls $f : \mathbb{R}^n \to \mathbb{R}$ zweimal stetig differenzierbar ist und die Hesse-Matrix in der Umgebung von $\mathbf{x}^\star$ Lipschitz-stetig ist.
2. Eine andere Startmatrix $\mathbf{H}_0 \neq \mathbf{I}$ ist denkbar, solange sie symmetrisch und positiv definit ist.
3. In der Praxis macht man gelegentlich Restarts, setzt also $\mathbf{H}_k := \mathbf{H}_0$, falls $k \in m\mathbb{Z}$ mit festem $m \in \mathbb{N}$, beispielsweise $m = 100$.
4. Gerade bei großen Optimierungsproblemen stellt man die Matrix $\mathbf{H}_k$ nicht direkt auf, sondern berechnet sie rekursiv aus den Vektoren $\{(\gamma_k, \nu_k, \mathbf{p}_k, \mathbf{q}_k)\}_{k\geq 0}$. Damit auch bei vielen Schritten der Speicherplatz nicht überhand nimmt, speichert man nur die höchstens letzten m Vektoren. Man erlaubt also ein "Gedächtnis" von m Updates und ersetzt die unbekannte Matrix $\mathbf{H}_{k-m}$ durch $\mathbf{H}_0$. Man spricht von einem *Limited-Memory-Quasi-Newton-Verfahren*.

△

5. Verfahren der konjugierten Gradienten

5.1 CG-Verfahren für lineare Gleichungssysteme

Es sei $\mathbf{A} \in \mathbb{R}^{n\times n}$ eine symmetrische und positive definite Matrix und $\mathbf{b} \in \mathbb{R}^n$. Das *Verfahren der konjugierten Gradienten* oder *CG-Verfahren* zur Lösung des linearen Gleichungssystems $\mathbf{Ax} = \mathbf{b}$ geht davon aus, dass die Lösung $\mathbf{x}^\star$ eindeutiges Minimum $\phi(\mathbf{x}^\star) = 0$ des Funktionals

$$\phi(\mathbf{x}) = \frac{1}{2}(\mathbf{b}-\mathbf{Ax})^T\mathbf{A}^{-1}(\mathbf{b}-\mathbf{Ax}) = \frac{1}{2}\mathbf{x}^T\mathbf{Ax} - \mathbf{x}^T\mathbf{b} + \frac{1}{2}\mathbf{b}^T\mathbf{A}^{-1}\mathbf{b} \geq 0$$

ist.

Ausgehend von einer Startnäherung $\mathbf{x}$ wollen wir ϕ in die Richtung $\mathbf{d}$ minimieren

$$\phi(\mathbf{x}+\alpha\mathbf{d}) = \phi(\mathbf{x}) + \frac{\alpha^2}{2}\mathbf{d}^T\mathbf{Ad} - \alpha\mathbf{d}^T(\mathbf{b}-\mathbf{Ax}) \to \min_{\alpha\in\mathbb{R}}.$$

Aus

$$\frac{\partial\phi(\mathbf{x}+\alpha\mathbf{d})}{\partial\alpha} = \alpha\mathbf{d}^T\mathbf{Ad} - \mathbf{d}^T(\mathbf{b}-\mathbf{Ax}) \stackrel{!}{=} 0$$

folgt daher

$$\alpha = \frac{\mathbf{d}^T(\mathbf{b}-\mathbf{Ax})}{\mathbf{d}^T\mathbf{Ad}}. \tag{5.1}$$

Lemma 5.1 Die Vektoren $\mathbf{d}_0, \mathbf{d}_1, \ldots, \mathbf{d}_k$ seien $\mathbf{A}$-*konjugiert*, das heißt, es gelte $\mathbf{d}_i^T\mathbf{Ad}_j = 0$ für alle $i \neq j$. Ist

$$\mathbf{x}_k = \arg\min_{\mathbf{x}\in\mathbf{x}_0+\operatorname{span}\{\mathbf{d}_0,\mathbf{d}_1,\ldots,\mathbf{d}_{k-1}\}} \phi(\mathbf{x})$$

und setzt man

$$\mathbf{x}_{k+1} = \mathbf{x}_k + \alpha_k\mathbf{d}_k, \qquad \alpha_k = \frac{\mathbf{d}_k^T\mathbf{r}_k}{\mathbf{d}_k^T\mathbf{Ad}_k}, \qquad \mathbf{r}_k = \mathbf{b}-\mathbf{Ax}_k, \tag{5.2}$$

so folgt

$$\mathbf{x}_{k+1} = \arg\min_{\mathbf{x}\in\mathbf{x}_0+\operatorname{span}\{\mathbf{d}_0,\mathbf{d}_1,\ldots,\mathbf{d}_k\}} \phi(\mathbf{x}).$$

Beweis. Die $\mathbf{A}$-Konjugiertheit der Vektoren $\{\mathbf{d}_\ell\}$ impliziert $\mathbf{d}_k^T\mathbf{A}(\mathbf{x}_k - \mathbf{x}_0) = 0$. Daher

folgt

$$\begin{aligned}\phi(\mathbf{x}_k + \alpha\mathbf{d}_k) &= \phi(\mathbf{x}_k) + \frac{\alpha^2}{2}\mathbf{d}_k^T\mathbf{A}\mathbf{d}_k - \alpha\mathbf{d}_k^T(\mathbf{b} - \mathbf{A}\mathbf{x}_k)\\ &= \phi(\mathbf{x}_k) + \frac{\alpha^2}{2}\mathbf{d}_k^T\mathbf{A}\mathbf{d}_k - \alpha\mathbf{d}_k^T(\mathbf{b} - \mathbf{A}\mathbf{x}_0)\\ &=: \phi(\mathbf{x}_k) + \varphi(\alpha),\end{aligned}$$

das heißt, das Minimierungsproblem entkoppelt. Da nach Voraussetzung $\mathbf{x}_k$ das Funktional ϕ über $\mathbf{x}_0 + \operatorname{span}\{\mathbf{d}_0, \mathbf{d}_1 \ldots, \mathbf{d}_{k-1}\}$ minimiert, wird das eindeutige Minimum angenommen, wenn $\varphi(\alpha)$ minimal ist. Dies ist aber nach (5.1) genau dann der Fall, wenn

$$\alpha_k = \frac{\mathbf{d}_k^T(\mathbf{b} - \mathbf{A}\mathbf{x}_k)}{\mathbf{d}_k^T\mathbf{A}\mathbf{d}_k} = \frac{\mathbf{d}_k^T(\mathbf{b} - \mathbf{A}\mathbf{x}_0)}{\mathbf{d}_k^T\mathbf{A}\mathbf{d}_k}. \tag{5.3}$$

□

Die Idee des CG-Verfahrens ist es nun, ausgehend von einer Startnäherung $\mathbf{x}_0$, sukzessive über die konjugierten Richtungen $\mathbf{d}_k$ zu minimieren. Die Folge der Residuen

$$\mathbf{r}_0 = \mathbf{b} - \mathbf{A}\mathbf{x}_0, \qquad \mathbf{r}_{k+1} = \mathbf{b} - \mathbf{A}\mathbf{x}_{k+1} \overset{(5.2)}{=} \mathbf{r}_k - \alpha_k\mathbf{A}\mathbf{d}_k, \quad k \geq 0, \tag{5.4}$$

erfüllt dann für alle $\ell < k$

$$\mathbf{d}_\ell^T\mathbf{r}_k = \mathbf{d}_\ell^T(\mathbf{b}-\mathbf{A}\mathbf{x}_k) = \mathbf{d}_\ell^T\left(\mathbf{b}-\mathbf{A}\mathbf{x}_0-\sum_{i=0}^{k-1}\alpha_i\mathbf{A}\mathbf{d}_i\right) = \mathbf{d}_\ell^T(\mathbf{b}-\mathbf{A}\mathbf{x}_0)-\alpha_\ell\mathbf{d}_\ell^T\mathbf{A}\mathbf{d}_\ell \overset{(5.3)}{=} 0. \tag{5.5}$$

Da die Richtungen $\mathbf{d}_k$ paarweise $\mathbf{A}$-konjugiert und folglich linear unabhängig sind, ergibt sich $\mathbf{r}_n = \mathbf{0}$, das heißt, das CG-Verfahren liefert die Lösung $\mathbf{A}^{-1}\mathbf{b}$ nach höchstens n Schritten. Zu beantworten bleibt daher nur die Frage, wie die Suchrichtungen $\mathbf{d}_k$ geschickt gewählt werden können.

Lemma 5.2 Für beliebiges $\mathbf{d}_0 = \mathbf{r}_0$ erzeugt die Rekursion

$$\mathbf{r}_{k+1} = \mathbf{r}_k - \alpha_k\mathbf{A}\mathbf{d}_k, \qquad \mathbf{d}_{k+1} = \mathbf{r}_{k+1} - \beta_k\mathbf{d}_k, \qquad \beta_k = \frac{\mathbf{d}_k^T\mathbf{A}\mathbf{r}_{k+1}}{\mathbf{d}_k^T\mathbf{A}\mathbf{d}_k} \tag{5.6}$$

solange eine Folge nichtverschwindender $\mathbf{A}$-konjugierter Vektoren $\mathbf{d}_0, \mathbf{d}_1, \ldots, \mathbf{d}_{k+1}$ bis $\mathbf{r}_{k+1} = \mathbf{0}$ ist.

Beweis. Sei

$$\mathcal{K}_k(\mathbf{A}, \mathbf{r}_0) := \operatorname{span}\{\mathbf{r}_0, \mathbf{A}\mathbf{r}_0, \ldots, \mathbf{A}^{k-1}\mathbf{r}_0\}.$$

Wir zeigen zunächst induktiv, dass stets gilt

$$\mathcal{K}_k(\mathbf{A}, \mathbf{r}_0) = \operatorname{span}\{\mathbf{r}_0, \mathbf{r}_1, \ldots, \mathbf{r}_{k-1}\} = \operatorname{span}\{\mathbf{d}_0, \mathbf{d}_1, \ldots, \mathbf{d}_{k-1}\}.$$

Da für $k = 1$ die Aussage klar ist, nehmen wir an, sie gilt für ein $k \geq 1$. Dann folgt

$$\mathbf{r}_k \overset{(5.6)}{=} \underbrace{\mathbf{r}_{k-1}}_{\in\mathcal{K}_k(\mathbf{A},\mathbf{r}_0)} - \alpha_k \underbrace{\mathbf{A}\mathbf{d}_k}_{\in\mathcal{K}_{k+1}(\mathbf{A},\mathbf{r}_0)} \in \mathcal{K}_{k+1}(\mathbf{A}, \mathbf{r}_0).$$

Gemäß (5.5) ist $\mathbf{r}_k \perp \operatorname{span}\{\mathbf{d}_0, \mathbf{d}_1, \ldots, \mathbf{d}_{k-1}\} = \mathcal{K}_k(\mathbf{A}, \mathbf{r}_0)$, dies bedeutet

$$\mathcal{K}_k(\mathbf{A}, \mathbf{r}_0) \subsetneq \operatorname{span}\{\mathbf{r}_0, \mathbf{r}_1, \ldots, \mathbf{r}_k\} \subseteq \mathcal{K}_{k+1}(\mathbf{A}, \mathbf{r}_0).$$

Da die Dimension von $\mathcal{K}_{k+1}(\mathbf{A}, \mathbf{r}_0)$ höchstens um 1 höher ist als die von $\mathcal{K}_k(\mathbf{A}, \mathbf{r}_0)$, muss gelten

$$\mathcal{K}_{k+1}(\mathbf{A}, \mathbf{r}_0) = \operatorname{span}\{\mathbf{r}_0, \mathbf{r}_1, \ldots, \mathbf{r}_k\}.$$

Aus $\mathbf{r}_k \overset{(5.6)}{=} \mathbf{d}_k + \beta_{k-1}\mathbf{d}_{k-1}$ folgt

$$\begin{aligned}\operatorname{span}\{\mathbf{d}_0, \mathbf{d}_1, \ldots, \mathbf{d}_k\} &= \operatorname{span}\{\mathbf{d}_0, \mathbf{d}_1, \ldots, \mathbf{d}_{k-1}, \mathbf{r}_k\} \\ &= \operatorname{span}\{\mathbf{r}_0, \mathbf{r}_1, \ldots, \mathbf{r}_{k-1}, \mathbf{r}_k\} = \mathcal{K}_{k+1}(\mathbf{A}, \mathbf{r}_0).\end{aligned}$$

Insbesondere muss aus Dimensionsgründen $\mathbf{d}_k \neq \mathbf{0}$ sein.
Es verbleibt, die $\mathbf{A}$-Konjugiertheit zu zeigen: Angenommen, $\mathbf{d}_0, \mathbf{d}_1, \ldots, \mathbf{d}_k$ sind $\mathbf{A}$-konjugiert. Der Induktionsschritt folgt dann aus

$$\mathbf{d}_k^T\mathbf{A}\mathbf{d}_{k+1} \overset{(5.6)}{=} \mathbf{d}_k^T\mathbf{A}(\mathbf{r}_{k+1} - \beta_k\mathbf{d}_k) \overset{(5.6)}{=} \mathbf{d}_k^T\mathbf{A}\mathbf{r}_{k+1} - \frac{\mathbf{d}_k^T\mathbf{A}\mathbf{r}_{k+1}}{\mathbf{d}_k^T\mathbf{A}\mathbf{d}_k}\mathbf{d}_k^T\mathbf{A}\mathbf{d}_k = 0$$

und für alle $\ell < k$

$$\mathbf{d}_\ell^T\mathbf{A}\mathbf{d}_{k+1} \overset{(5.6)}{=} \mathbf{d}_\ell^T\mathbf{A}\mathbf{r}_{k+1} - \frac{\mathbf{d}_k^T\mathbf{A}\mathbf{r}_{k+1}}{\mathbf{d}_k^T\mathbf{A}\mathbf{d}_k}\underbrace{\mathbf{d}_\ell^T\mathbf{A}\mathbf{d}_k}_{=0} = (\mathbf{A}\mathbf{d}_\ell)^T\mathbf{r}_{k+1} = 0$$

wegen $\mathbf{A}\mathbf{d}_\ell \in \mathcal{K}_{k+1}(\mathbf{A}, \mathbf{r}_0) \perp \mathbf{r}_{k+1}$. □

Bemerkung Der Raum $\mathcal{K}_k(\mathbf{A}, \mathbf{r}_0)$ heißt *Krylov-Raum*. Die Iterierte $\mathbf{x}_k$ des CG-Verfahrens minimiert demnach das Funktional $\phi(\mathbf{x})$ unter allen $\mathbf{x}$ aus dem verschobenen Krylov-Raum $\mathbf{x}_0 + \mathcal{K}_k(\mathbf{A}, \mathbf{r}_0)$. △

Um den CG-Algorithmus endgültig zu formulieren, bemerken wir zunächst, dass gilt

$$\alpha_k \overset{(5.3)}{=} \frac{\mathbf{d}_k^T\mathbf{r}_k}{\mathbf{d}_k^T\mathbf{A}\mathbf{d}_k} \overset{(5.6)}{=} \frac{(\mathbf{r}_k - \beta_{k-1}\mathbf{d}_{k-1})^T\mathbf{r}_k}{\mathbf{d}_k^T\mathbf{A}\mathbf{d}_k} \overset{(5.5)}{=} \frac{\|\mathbf{r}_k\|_2^2}{\mathbf{d}_k^T\mathbf{A}\mathbf{d}_k}. \tag{5.7}$$

Wegen $\mathbf{r}_k \in \mathcal{K}_{k+1}(\mathbf{A}, \mathbf{r}_0) \perp \mathbf{r}_{k+1}$ folgt ferner

$$\mathbf{d}_k^T\mathbf{A}\mathbf{r}_{k+1} = (\mathbf{A}\mathbf{d}_k)^T\mathbf{r}_{k+1} \overset{(5.6)}{=} \frac{1}{\alpha_k}(\mathbf{r}_k - \mathbf{r}_{k+1})^T\mathbf{r}_{k+1} \overset{(5.7)}{=} -\frac{\mathbf{d}_k^T\mathbf{A}\mathbf{d}_k}{\|\mathbf{r}_k\|_2^2}\|\mathbf{r}_{k+1}\|_2^2$$

und damit

$$\beta_k \overset{(5.6)}{=} \frac{\mathbf{d}_k^T\mathbf{A}\mathbf{r}_{k+1}}{\mathbf{d}_k^T\mathbf{A}\mathbf{d}_k} = -\frac{\|\mathbf{r}_{k+1}\|_2^2}{\|\mathbf{r}_k\|_2^2}. \tag{5.8}$$

Die Kombination von (5.2) und (5.6)–(5.8) liefert schließlich:

Algorithmus 5.3 (CG-Verfahren)
input: Matrix $\mathbf{A} \in \mathbb{R}^{n\times n}$, rechte Seite $\mathbf{b} \in \mathbb{R}^n$ und Startnäherung $\mathbf{x}_0 \in \mathbb{R}^n$
output: Folge von Iterierten $\{\mathbf{x}_k\}_{k\in\mathbb{N}}$
① Initialisierung: setze $\mathbf{d}_0 = \mathbf{r}_0 := \mathbf{b} - \mathbf{A}\mathbf{x}_0$ und $k := 0$

② berechne

$$\alpha_k := \frac{\|\mathbf{r}_k\|_2^2}{\mathbf{d}_k^T \mathbf{A}\mathbf{d}_k}$$
$$\mathbf{x}_{k+1} := \mathbf{x}_k + \alpha_k \mathbf{d}_k$$
$$\mathbf{r}_{k+1} := \mathbf{r}_k - \alpha_k \mathbf{A}\mathbf{d}_k$$
$$\beta_k := \frac{\|\mathbf{r}_{k+1}\|_2^2}{\|\mathbf{r}_k\|_2^2}$$
$$\mathbf{d}_{k+1} := \mathbf{r}_{k+1} + \beta_k \mathbf{d}_k$$

③ falls $\|\mathbf{r}_{k+1}\|_2 > \varepsilon$ erhöhe $k := k+1$ und gehe nach ②

Das CG-Verfahren wird generell als Iterationsverfahren verwendet, das heißt, man bricht die Iteration ab, falls die Residuennorm $\|\mathbf{r}_k\|_2$ kleiner als eine Fehlertoleranz ε ist. Pro Iterationsschritt wird nur eine Matrix-Vektor-Multiplikation benötigt. Allerdings hängt die Konvergenz des Verfahrens stark von der Kondition der Matrix ab. Man kann zeigen, dass die Iterierten $\{\mathbf{x}_k\}$ des CG-Verfahrens bezüglich der *Energienorm*

$$\|\mathbf{x}\|_{\mathbf{A}} := \sqrt{\langle \mathbf{x}, \mathbf{x} \rangle_{\mathbf{A}}} = \sqrt{\mathbf{x}^T \mathbf{A}\mathbf{x}}$$

der Fehlerabschätzung

$$\|\mathbf{x}_k - \mathbf{x}^\star\|_{\mathbf{A}} \leq 2\left(\frac{\sqrt{\operatorname{cond}_2 \mathbf{A}} - 1}{\sqrt{\operatorname{cond}_2 \mathbf{A}} + 1}\right)^k \|\mathbf{x}_0 - \mathbf{x}^\star\|_{\mathbf{A}}$$

genügen.

5.2 Nichtlineares CG-Verfahren

In Anlehnung an das CG-Verfahren 5.3 ist das *nichtlineare CG-Verfahren* zur Lösung von nichtlinearen Optimierungsproblemen $f(\mathbf{x}) \to \min$ definiert.

Algorithmus 5.4 (Nichtlineares CG-Verfahren)
input: Funktion $f : \mathbb{R}^n \to \mathbb{R}$ und Startnäherung $\mathbf{x}_0 \in \mathbb{R}^n$
output: Folge von Iterierten $\{\mathbf{x}_k\}_{k\in\mathbb{N}}$

① Initialisierung: setze $\mathbf{d}_0 = -\nabla f(\mathbf{x}_0)$ und $k := 0$

② löse

$$\alpha_k \approx \operatorname*{argmin}_{\alpha \in \mathbb{R}} f(\mathbf{x}_k + \alpha \mathbf{d}_k)$$

③ berechne

$$\mathbf{x}_{k+1} := \mathbf{x}_k + \alpha_k \mathbf{d}_k$$

$$\underbrace{\beta_k := \frac{\|\nabla f(\mathbf{x}_{k+1})\|_2^2}{\|\nabla f(\mathbf{x}_k)\|_2^2}}_{\text{Verfahren von Fletcher und Reeves}} \quad \text{oder} \quad \underbrace{\beta_k := \frac{\nabla f(\mathbf{x}_{k+1})^T \{\nabla f(\mathbf{x}_{k+1}) - \nabla f(\mathbf{x}_k)\}}{\|\nabla f(\mathbf{x}_k)\|_2^2}}_{\text{Verfahren von Polak und Ribière}}$$

$$\mathbf{d}_{k+1} := -\nabla f(\mathbf{x}_{k+1}) + \beta_k \mathbf{d}_k$$

④ erhöhe $k := k + 1$ und gehe nach ②

Bemerkung Ist $f(\mathbf{x}) = \frac{1}{2}\mathbf{x}^T\mathbf{A}\mathbf{x} - \mathbf{b}^T\mathbf{x} + c$ eine quadratische Funktion, dann fallen bei exakter Minimierung in ② sowohl das Verfahren von Fletcher und Reeves als auch das Verfahren von Polak und Ribière mit dem CG-Verfahren zusammen. Ersteres folgt aus $\nabla f(\mathbf{x}_k) = \mathbf{A}\mathbf{x}_k - \mathbf{b} = -\mathbf{r}_k$, zweiteres aus $\nabla f(\mathbf{x}_k)^T \nabla f(\mathbf{x}_{k+1}) = \mathbf{r}_k^T \mathbf{r}_{k+1} = 0$. △

Lemma 5.5 Die Funktion $f : D \subset \mathbb{R}^n \to \mathbb{R}$ sei gleichmäßig konvex. Weiter sei f differenzierbar mit Lipschitz-stetigem Gradienten:

$$\|\nabla f(\mathbf{x}) - \nabla f(\mathbf{y})\|_2 \leq L\|\mathbf{x} - \mathbf{y}\|_2 \quad \text{für alle } \mathbf{x}, \mathbf{y} \in D.$$

Dann gilt für das Verfahren von Polak und Ribière bei exakter Liniensuche in ②

$$-\mathbf{d}_k^T \nabla f(\mathbf{x}_k) \geq \frac{\mu}{\mu + L}\|\nabla f(\mathbf{x}_k)\|_2 \|\mathbf{d}_k\|_2.$$

Beweis. Bei exakter Liniensuche gilt im k-ten Schritt des Verfahrens von Polak und Ribière

$$0 = \nabla f(\mathbf{x}_\ell + \alpha_\ell \mathbf{d}_\ell)^T \mathbf{d}_\ell = \nabla f(\mathbf{x}_{\ell+1})^T \mathbf{d}_\ell \quad \text{für alle } \ell \leq k. \tag{5.9}$$

Daher können wir den Nenner in

$$\beta_k = \frac{\nabla f(\mathbf{x}_{k+1})^T \{\nabla f(\mathbf{x}_{k+1}) - \nabla f(\mathbf{x}_k)\}}{\|\nabla f(\mathbf{x}_k)\|_2^2}$$

folgendemaßen umformen:

$$\begin{aligned}
\|\nabla f(\mathbf{x}_k)\|_2^2 &= (\beta_{k-1}\mathbf{d}_{k-1} - \mathbf{d}_k)^T \nabla f(\mathbf{x}_k) \\
&\overset{(5.9)}{=} -\mathbf{d}_k^T \nabla f(\mathbf{x}_k) \\
&\overset{(5.9)}{=} \mathbf{d}_k^T \big\{\nabla f(\mathbf{x}_{k+1}) - \nabla f(\mathbf{x}_k)\big\} \\
&= \frac{1}{\alpha_k}(\mathbf{x}_{k+1} - \mathbf{x}_k)^T \big\{\nabla f(\mathbf{x}_{k+1}) - \nabla f(\mathbf{x}_k)\big\}.
\end{aligned}$$

Aufgrund der gleichmäßigen Konvexität und der Lipschitz-Bedingung erhalten wir

$$|\beta_k| \leq \frac{L}{\mu}|\alpha_k| \frac{\|\nabla f(\mathbf{x}_{k+1})\|_2 \|\mathbf{x}_{k+1} - \mathbf{x}_k\|_2}{\|\mathbf{x}_{k+1} - \mathbf{x}_k\|_2^2} = \frac{L}{\mu}\frac{\|\nabla f(\mathbf{x}_{k+1})\|_2}{\|\mathbf{d}_k\|_2}.$$

Dies führt auf

$$\|\mathbf{d}_{k+1}\|_2 \leq \|\nabla f(\mathbf{x}_{k+1})\|_2 + |\beta_k|\|\mathbf{d}_k\|_2 \leq \left(1 + \frac{L}{\mu}\right)\|\nabla f(\mathbf{x}_{k+1})\|_2,$$

woraus dann die Behauptung folgt

$$
\begin{aligned}
-\frac{\mathbf{d}_{k+1}^T \nabla f(\mathbf{x}_{k+1})}{\|\mathbf{d}_{k+1}\|_2 \|\nabla f(\mathbf{x}_{k+1})\|_2} &= \frac{\{\nabla f(\mathbf{x}_{k+1}) - \beta_k \mathbf{d}_k\}^T \nabla f(\mathbf{x}_{k+1})}{\|\mathbf{d}_{k+1}\|_2 \|\nabla f(\mathbf{x}_{k+1})\|_2} \\
&\overset{(5.9)}{=} \frac{\|\nabla f(\mathbf{x}_{k+1})\|_2^2}{\|\mathbf{d}_{k+1}\|_2 \|\nabla f(\mathbf{x}_{k+1})\|_2} \\
&\geq \frac{\mu}{\mu + L}.
\end{aligned}
$$

□

Bemerkung Die geometrische Interpretation von Lemma 5.5 ist, dass beim Verfahren von Polak und Ribière die Suchrichtung $\mathbf{d}_k$ und die Richtung des steilsten Abstiegs $-\nabla f(\mathbf{x}_k)$ stets den Winkel θ mit $\cos\theta > \mu/(\mu + L)$ einschließen. △

Satz 5.6 Die Funktion $f : D \subset \mathbb{R}^n \to \mathbb{R}$ sei gleichmäßig konvex. Weiter sei f differenzierbar mit Lipschitz-stetigem Gradienten:

$$\|\nabla f(\mathbf{x}) - \nabla f(\mathbf{y})\|_2 \leq L\|\mathbf{x} - \mathbf{y}\|_2 \quad \text{für alle } \mathbf{x}, \mathbf{y} \in D.$$

Dann konvergiert das Verfahren von Polak und Ribière mit exakter Liniensuche in ② für beliebige Startnäherungen $\mathbf{x}_0 \in D$ gegen das eindeutige globale Minimum $\mathbf{x}^\star$ und es gilt

$$f(\mathbf{x}_{k+1}) - f(\mathbf{x}^\star) \leq \left(1 - \frac{\mu^4}{L^2(\mu + L)^2}\right)\{f(\mathbf{x}_k) - f(\mathbf{x}^\star)\}, \quad k = 1, 2, \ldots.$$

Beweis. Der Beweis ist ähnlich zu dem von Satz 2.3. Aufgrund der Minimierungsbedingung gilt wieder für ein $\gamma > 0$

$$
\begin{aligned}
f(\mathbf{x}_{k+1}) &\leq f(\mathbf{x}_k + \gamma \mathbf{d}_k) \\
&= f(\mathbf{x}_k) + \gamma \int_0^1 \nabla f(\mathbf{x}_k + \gamma t \mathbf{d}_k)^T \mathbf{d}_k \, \mathrm{d}t \\
&= f(\mathbf{x}_k) + \gamma \nabla f(\mathbf{x}_k)^T \mathbf{d}_k + \gamma \int_0^1 \{\nabla f(\mathbf{x}_k + \gamma t \mathbf{d}_k) - \nabla f(\mathbf{x}_k)\}^T \mathbf{d}_k \, \mathrm{d}t \\
&\leq f(\mathbf{x}_k) + \gamma \nabla f(\mathbf{x}_k)^T \mathbf{d}_k + \gamma \int_0^1 \underbrace{\|\nabla f(\mathbf{x}_k + \gamma t \mathbf{d}_k) + \nabla f(\mathbf{x}_k)\|_2}_{\leq t\gamma L \|\mathbf{d}_k\|_2} \|\mathbf{d}_k\|_2 \, \mathrm{d}t \\
&\leq f(\mathbf{x}_k) + \gamma \nabla f(\mathbf{x}_k)^T \mathbf{d}_k + \gamma^2 \frac{L}{2} \|\mathbf{d}_k\|_2^2.
\end{aligned}
$$

Für die Wahl

$$\gamma := -\frac{\nabla f(\mathbf{x}_k)^T \mathbf{d}_k}{L\|\mathbf{d}_k\|_2^2}$$

folgern wir mit Lemma 5.5

$$\begin{aligned} f(\mathbf{x}_{k+1}) - f(\mathbf{x}^\star) &\leq f(\mathbf{x}_k) - f(\mathbf{x}^\star) - \frac{(\nabla f(\mathbf{x}_k)^T \mathbf{d}_k)^2}{2L\|\mathbf{d}_k\|_2^2} \\ &\leq f(\mathbf{x}_k) - f(\mathbf{x}^\star) - \frac{\mu^2}{2L(\mu+L)^2} \|\nabla f(\mathbf{x}_k) - \underbrace{\nabla f(\mathbf{x}^\star)}_{=\mathbf{0}}\|_2^2. \end{aligned}$$

Aus der gleichmäßigen Konvexität ergibt sich

$$\mu\|\mathbf{x}_k - \mathbf{x}^\star\|_2^2 \leq \big\{\nabla f(\mathbf{x}_k) - \nabla f(\mathbf{x}^\star)\big\}^T (\mathbf{x}_k - \mathbf{x}^\star) \leq \|\nabla f(\mathbf{x}_k) - \nabla f(\mathbf{x}^\star)\|_2 \|\mathbf{x}_k - \mathbf{x}^\star\|_2,$$

während die Lipschitz-Stetigkeit impliziert

$$f(\mathbf{x}_k) - f(\mathbf{x}^\star) = \int_0^1 \nabla f\big(t\mathbf{x}_k + (1-t)\mathbf{x}^\star\big)^T (\mathbf{x}_k - \mathbf{x}^\star)\, \mathrm{d}t \leq \frac{L}{2}\|\mathbf{x}_k - \mathbf{x}^\star\|_2^2.$$

Setzen wir diese beiden Abschätzungen in die obige ein, so erhalten wir das Behauptete. □

Bemerkungen

1. Die Konvergenzabschätzung aus Satz 5.6 ist schlechter als die das Gradientenverfahren betreffende aus Satz 2.3. Dies spiegelt jedoch nicht die Tatsache wieder, dass das CG-Verfahren im Fall quadratischer Funktionen viel schneller konvergiert als das Gradientenverfahren. Die Ursache liegt in der Beweistechnik begründet, die das Verfahren von Polak und Ribière als Störung des Gradientenverfahrens auffasst.
2. Das Verfahren von Polak und Ribière konvergiert im allgemeinen schneller als das Verfahren von Fletcher und Reeves.
3. In der Praxis verwendet man Restarts: Wird der Winkel zwischen dem Antigradienten und der Suchrichtung zu groß, etwa

   $$-\frac{\nabla f(\mathbf{x}_k)^T \mathbf{d}_k}{\|\nabla f(\mathbf{x}_k)\|_2 \|\mathbf{d}_k\|_2} < \gamma$$

   für kleines $\gamma \in (0,1)$, dann startet das Verfahren durch einen Gradientenschritt neu.

△

5.3 Modifiziertes Verfahren von Polak und Ribière

Nichtlineare CG-Verfahren fallen, wie auch das BFGS-Verfahren, im Fall einer konvexen quadratischen Funktion mit dem CG-Verfahren zusammen. Allerdings sind die Quasi-Newton-Verfahren robuster hinsichtlich der Schrittweitensteuerung. Die nichtlinearen CG-Verfahren funktionieren umso besser, je genauer die Liniensuche in ② von Algorithmus 5.4 durchgeführt wird. Eine direkt zu implementierende Schrittweitensteuerung stellen wir im folgenden modifizierten Verfahren von Polak und Ribière vor.

Algorithmus 5.7 (modifiziertes Verfahren von Polak und Ribière)
input: Funktion $f : \mathbb{R}^n \to \mathbb{R}$ und Startnäherung $\mathbf{x}_0 \in \mathbb{R}^n$
output: Folge von Iterierten $\{\mathbf{x}_k\}_{k\in\mathbb{N}}$

① Initialisierung: wähle $\sigma \in (0,1)$, $0 < \underline{\gamma} < 1 < \overline{\gamma}$ und setze $\mathbf{d}_0 = -\nabla f(\mathbf{x}_0)$, $k := 0$

② setze

$$\alpha_k := \frac{|\nabla f(\mathbf{x}_k)^T \mathbf{d}_k|}{\|\mathbf{d}_k\|_2^2}$$

③ berechne

$$\begin{aligned} \mathbf{x}_{k+1} &:= \mathbf{x}_k + \alpha_k \mathbf{d}_k \\ \beta_k &:= \frac{\nabla f(\mathbf{x}_{k+1})^T \{\nabla f(\mathbf{x}_{k+1}) - \nabla f(\mathbf{x}_k)\}}{\|\nabla f(\mathbf{x}_k)\|_2^2} \\ \mathbf{d}_{k+1} &:= -\nabla f(\mathbf{x}_{k+1}) + \beta_k \mathbf{d}_k \end{aligned}$$

④ ist eine der Bedingungen

$$f(\mathbf{x}_{k+1}) \le f(\mathbf{x}_k) - \sigma \alpha_k^2 \|\mathbf{d}_k\|_2^2 \tag{5.10}$$

$$-\overline{\gamma} \|\nabla f(\mathbf{x}_{k+1})\|_2^2 \le \nabla f(\mathbf{x}_{k+1})^T \mathbf{d}_{k+1} \le -\underline{\gamma} \|\nabla f(\mathbf{x}_{k+1})\|_2^2 \tag{5.11}$$

verletzt, dann halbiere α_k und gehe nach ③

⑤ ist $\nabla f(\mathbf{x}_{k+1}) \neq \mathbf{0}$, dann erhöhe $k := k+1$ und gehe nach ②

Lemma 5.8 Ist $f : \mathbb{R}^n \to \mathbb{R}$ stetig differenzierbar, so ist Algorithmus 5.7 wohldefiniert.

Beweis. Wir bemerken zunächst, dass stets $\mathbf{d}_k \neq \mathbf{0}$ ist und somit der Faktor α_k in ② existiert. Wäre nämlich $\mathbf{d}_k = \mathbf{0}$ für ein $k \in \mathbb{N}_0$, so würde aus ② im Fall $k = 0$ beziehungsweise aus (5.11) im Fall $k > 0$ sofort $\nabla f(\mathbf{x}_k) = \mathbf{0}$ folgen.

Es ist also nur zu zeigen, dass die Liniensuche ②–④ in jedem Iterationschritt $k \in \mathbb{N}_0$ erfolgreich ist. Zu diesem Zweck nehmen wir an, dass $k \in \mathbb{N}_0$ ein fester Iterationsindex mit $\nabla f(\mathbf{x}_k)^T \mathbf{d}_k < 0$ ist. Als erstes stellen wir fest, dass die Bedingung (5.10) wegen

$$f(\mathbf{x}_k + \alpha \mathbf{d}_k) = f(\mathbf{x}_k) + \alpha \nabla f(\mathbf{x}_k)^T \mathbf{d}_k + o(\alpha)$$

nach endlich vielen erfolglosen Schritten der Liniensuche immer erfüllt ist.

Als nächstes zeigen wir, dass die Bedingung (5.11) ebenfalls nach endlich vielen erfolglosen Schritten stets erfüllt ist. Denn angenommen, dem ist nicht so. Dann gibt es eine Teilfolge $\{k_\ell\}_{\ell>0}$, so dass für jedes

$$\mathbf{y}_\ell = \mathbf{x}_k + 2^{-k_\ell} \frac{|\nabla f(\mathbf{x}_k)^T \mathbf{d}_k|}{\|\mathbf{d}_k\|_2^2} \mathbf{d}_k, \quad \ell \in \mathbb{N}$$

zumindest eine der beiden Bedingungen

$$\nabla f(\mathbf{y}_\ell)^T \left\{ -\nabla f(\mathbf{y}_\ell) + \frac{\nabla f(\mathbf{y}_\ell)^T \{\nabla f(\mathbf{y}_\ell) - \nabla f(\mathbf{x}_k)\}}{\|\nabla f(\mathbf{x}_k)\|_2^2} \mathbf{d}_k \right\} > -\underline{\gamma} \|\nabla f(\mathbf{y}_\ell)\|_2^2,$$

$$\nabla f(\mathbf{y}_\ell)^T \left\{ -\nabla f(\mathbf{y}_\ell) + \frac{\nabla f(\mathbf{y}_\ell)^T \{\nabla f(\mathbf{y}_\ell) - \nabla f(\mathbf{x}_k)\}}{\|\nabla f(\mathbf{x}_k)\|_2^2} \mathbf{d}_k \right\} < -\overline{\gamma} \|\nabla f(\mathbf{y}_\ell)\|_2^2$$

nicht erfüllt ist. Der Grenzübergang $\ell \to \infty$ liefert $\mathbf{y}_\ell \to \mathbf{x}_k$ und folglich gilt

$$-\|\nabla f(\mathbf{x}_k)\|_2^2 \geq -\underline{\gamma}\|\nabla f(\mathbf{x}_k)\|_2^2 \quad \text{oder} \quad -\|\nabla f(\mathbf{x}_k)\|_2^2 \leq -\overline{\gamma}\|\nabla f(\mathbf{x}_k)\|_2^2.$$

Aus $0 < \underline{\gamma} < 1 < \overline{\gamma}$ folgt dann aber $\|\nabla f(\mathbf{x}_k)\|_2 = 0$ im Widerspruch zu unserer Voraussetzung $\nabla f(\mathbf{x}_k)^T \mathbf{d}_k < 0$.
Damit ist gezeigt, dass Algorithmus 5.7 wohldefiniert ist, sofern die Abstiegsbedingung $\nabla f(\mathbf{x}_k)^T \mathbf{d}_k < 0$ für alle $k \in \mathbb{N}_0$ erfüllt ist. Für $k = 0$ gilt sie aber nach Definition von $\mathbf{d}_0$ und für $k > 0$ folgt sie dann aus Bedingung (5.11). □

Lemma 5.9 Es sei $D \subset \mathbb{R}^n$ eine offene, beschränkte und konvexe Menge, in der f stetig differenzierbar, nach unten beschränkt und ∇f zudem Lipschitz-stetig ist. Ferner sei neben $\mathbf{x}_0$ auch die gesamte Niveaumenge $\mathrm{N} := \{\mathbf{x} \in \mathbb{R}^n : f(\mathbf{x}) \leq f(\mathbf{x}_0)\}$ in D enthalten. Dann gelten die folgenden Aussagen:

(*i.*) Alle Iterierten $\mathbf{x}_k$ liegen in der Niveaumenge N.

(*ii.*) Die Folge $\{f(\mathbf{x}_k)\}_{k\in\mathbb{N}}$ ist konvergent.

(*iii.*) Es gilt $\lim_{k\to\infty} \alpha_k \|\mathbf{d}_k\|_2 = 0$.

(*iv.*) Es ist $\alpha_k \|\mathbf{d}_k\|_2^2 \leq \overline{\gamma} c^2$, wobei $c < \infty$ eine obere Schranke von $\|\nabla f(\mathbf{x})\|_2$ auf der Niveaumenge N sei.

(*v.*) Es existiert eine Konstante $\theta > 0$ mit

$$\alpha_k \geq \theta \frac{|\nabla f(\mathbf{x}_k)^T \mathbf{d}_k|}{\|\mathbf{d}_k\|_2^2}$$

für alle $k \in \mathbb{N}_0$.

Beweis.

(*i.*) Diese Aussage ergibt sich unmittelbar aus der Bedingung (5.10).

(*ii.*) Die Folge $\{f(\mathbf{x}_k)\}_{k\in\mathbb{N}}$ ist streng monoton fallend und aufgrund der Voraussetzung nach unten beschränkt. Hieraus ergibt sich das Behauptete.

(*iii.*) Aus (5.10) folgt

$$\sigma \alpha_k^2 \|\mathbf{d}_k\|_2^2 \leq f(\mathbf{x}_k) - f(\mathbf{x}_{k+1})$$

für alle $k \in \mathbb{N}_0$. Der Grenzübergang $k \to \infty$ liefert daher unter Berücksichtigung der schon bewiesenen Aussage (*ii.*) die Behauptung.

(*iv.*) Aus den Schritten ②–④ von Algorithmus 5.7 folgt

$$\alpha_k \|\mathbf{d}_k\|_2^2 \leq \frac{|\nabla f(\mathbf{x}_k)^T \mathbf{d}_k|}{\|\mathbf{d}_k\|_2^2} \|\mathbf{d}_k\|_2^2 \leq \overline{\gamma} \|\nabla f(\mathbf{x}_k)\|_2^2 \leq \overline{\gamma} c^2$$

für alle $k \in \mathbb{N}_0$.

(*v.*) Zum Nachweis dieser Aussage führen wir eine Fallunterscheidung durch.
Fall 1: $\alpha_k = |\nabla f(\mathbf{x}_k)^T \mathbf{d}_k| / \|\mathbf{d}_k\|_2^2$.
Dann ist offensichtlich

$$\alpha_k \geq \frac{|\nabla f(\mathbf{x}_k)^T \mathbf{d}_k|}{\|\mathbf{d}_k\|_2^2}. \tag{5.12}$$

Fall 2: $\alpha_k < |\nabla f(\mathbf{x}_k)^T \mathbf{d}_k| / \|\mathbf{d}_k\|_2^2$.

Dann verletzt die Schrittweite $2\alpha_k$ zumindest eine der Bedingungen in ④. Der Punkt $\mathbf{z}_k := \mathbf{x}_k + 2\alpha_k \mathbf{d}_k$ genügt also (5.10) oder (5.11) nicht. Nach Aussage (*iii.*) existiert ein $K \in \mathbb{N}$, so dass $\mathbf{z}_k \in D$ für alle $k \geq K$. Im folgenden zeigen wir Aussage (*v.*) zunächst nur für solche k.

Fall 2A: Der Punkt $\mathbf{z}_k \in D$ verletzt (5.10).

Dann gilt

$$f(\mathbf{z}_k) > f(\mathbf{x}_k) - \sigma(2\alpha_k)^2 \|\mathbf{d}_k\|_2^2. \tag{5.13}$$

Aufgrund des Mittelwertsatzes existiert ein $\boldsymbol{\xi}_k$ auf der Verbindungsstrecke von $\mathbf{x}_k$ und $\mathbf{z}_k$, so dass

$$f(\mathbf{z}_k) = f(\mathbf{x}_k) + \nabla f(\boldsymbol{\xi}_k)^T (\mathbf{z}_k - \mathbf{x}_k) = f(\mathbf{x}_k) + 2\alpha_k \nabla f(\boldsymbol{\xi}_k)^T \mathbf{d}_k. \tag{5.14}$$

Aus (5.13) und (5.14) folgt daher

$$f(\mathbf{x}_k) + 2\alpha_k \nabla f(\mathbf{x}_k)^T \mathbf{d}_k + 2\alpha_k \big\{ \nabla f(\boldsymbol{\xi}_k)^T \mathbf{d}_k - \nabla f(\mathbf{x}_k)^T \mathbf{d}_k \big\} > f(\mathbf{x}_k) - \sigma(2\alpha_k)^2 \|\mathbf{d}_k\|_2^2.$$

Aus der Lipschitz-Stetigkeit von ∇f in D ergibt sich

$$2\alpha_k \nabla f(\mathbf{x}_k)^T \mathbf{d}_k + 2\alpha_k L \underbrace{\|\boldsymbol{\xi}_k - \mathbf{x}_k\|_2}_{=2\alpha_k \|\mathbf{d}_k\|_2} \|\mathbf{d}_k\|_2 > -\sigma(2\alpha_k)^2 \|\mathbf{d}_k\|_2^2.$$

Dies liefert unmittelbar

$$\alpha_k \geq \frac{1}{2(L+\sigma)} \frac{|\nabla f(\mathbf{x}_k)^T \mathbf{d}_k|}{\|\mathbf{d}_k\|_2^2}. \tag{5.15}$$

Fall 2B: Der Punkt $\mathbf{z}_k \in D$ verletzt die linke Ungleichung in (5.11).

Wegen

$$\nabla f(\mathbf{z}_k)^T \left\{ -\nabla f(\mathbf{z}_k) + \frac{\nabla f(\mathbf{z}_k)^T \{\nabla f(\mathbf{z}_k) - \nabla f(\mathbf{x}_k)\}}{\|\nabla f(\mathbf{x}_k)\|_2^2} \mathbf{d}_k \right\} < -\overline{\gamma} \|\nabla f(\mathbf{z}_k)\|_2^2$$

ergibt sich mit Hilfe der Cauchy-Schwarzschen Ungleichung

$$-\|\nabla f(\mathbf{z}_k)\|_2^2 - \|\nabla f(\mathbf{z}_k)\|_2^2 \frac{\|\nabla f(\mathbf{z}_k) - \nabla f(\mathbf{x}_k)\|_2}{\|\nabla f(\mathbf{x}_k)\|_2^2} \|\mathbf{d}_k\|_2 < -\overline{\gamma} \|\nabla f(\mathbf{z}_k)\|_2^2,$$

das heißt, es ist

$$1 + \frac{\|\nabla f(\mathbf{z}_k) - \nabla f(\mathbf{x}_k)\|_2}{\|\nabla f(\mathbf{x}_k)\|_2^2} \|\mathbf{d}_k\|_2 > \overline{\gamma}.$$

Aus der Lipschitz-Stetigkeit von ∇f und der Tatsache, dass die Schrittweite α_k der Bedingung (5.11) genügt, folgt daher nach kurzer Rechnung

$$\alpha_k \geq \frac{(\overline{\gamma} - 1)\|\nabla f(\mathbf{x}_k)\|_2^2}{2L\|\mathbf{d}_k\|_2^2} \geq \frac{\overline{\gamma} - 1}{2\overline{\gamma} L} \frac{|\nabla f(\mathbf{x}_k)^T \mathbf{d}_k|}{\|\mathbf{d}_k\|_2^2}. \tag{5.16}$$

Fall 2C: Der Punkt $\mathbf{z}_k \in D$ verletzt die linke Ungleichung in (5.11). Es ist

$$\nabla f(\mathbf{z}_k)^T \left\{ -\nabla f(\mathbf{z}_k) + \frac{\nabla f(\mathbf{z}_k)^T \{\nabla f(\mathbf{z}_k) - \nabla f(\mathbf{x}_k)\}}{\|\nabla f(\mathbf{x}_k)\|_2^2} \mathbf{d}_k \right\} > -\underline{\gamma} \|\nabla f(\mathbf{z}_k)\|_2^2.$$

Analog zum *Fall 2B* erhält man hieraus

$$\alpha_k \geq \frac{1-\underline{\gamma}}{2\overline{\gamma}L}\frac{|\nabla f(\mathbf{x}_k)^T\mathbf{d}_k|}{\|\mathbf{d}_k\|_2^2}. \tag{5.17}$$

Wegen (5.12), (5.15), (5.16), (5.17) folgt Aussage (*v.*) mit

$$\theta := \min\left\{1, \frac{1}{2(L+\sigma)}, \frac{\overline{\gamma}-1}{2\overline{\gamma}L}, \frac{1-\underline{\gamma}}{2\overline{\gamma}L}\right\},$$

und zwar zunächst für alle $k \geq K$. Da nur endlich viele k übrigbleiben, folgt Aussage (*v.*) nach eventueller Verkleinerung von θ aber auch für alle $k \in \mathbb{N}_0$.

□

Satz 5.10 Unter den Voraussetzungen von Lemma 5.9 gilt für die Iterierten $\{\mathbf{x}_k\}_{k\geq 0}$ des modifizierten Verfahren von Polak und Ribière 5.7

$$\nabla f(\mathbf{x}_k) \overset{k\to\infty}{\longrightarrow} \mathbf{0}.$$

Beweis. Angenommen, die Aussage des Satzes ist falsch. Dann existieren ein $\varepsilon > 0$ und eine Teilfolge $\{\mathbf{x}_{k_\ell}\}$, so dass

$$\|\nabla f(\mathbf{x}_{k_\ell-1})\|_2 > \varepsilon$$

für alle $\ell \in \mathbb{N}$. Aus der Aufdatierungsvorschrift für $\mathbf{d}_{k_\ell}$ und Lemma 5.9 (*iv.*) folgt dann

$$\|\mathbf{d}_{k_\ell}\|_2 \leq \|\nabla f(\mathbf{x}_{k_\ell})\|_2 + \frac{\|\nabla f(\mathbf{x}_{k_\ell})\|_2\|\nabla f(\mathbf{x}_{k_\ell}) - \nabla f(\mathbf{x}_{k_\ell-1})\|_2}{\|\nabla f(\mathbf{x}_{k_\ell-1})\|_2^2}\|\mathbf{d}_{k_\ell-1}\|_2 \leq c + \overline{\gamma}\frac{Lc^3}{\varepsilon^2}$$

für alle $\ell \in \mathbb{N}$. Zusammen mit Lemma 5.9 (*iii.*) ergibt sich hieraus

$$\lim_{\ell\to\infty} \alpha_{k_\ell}\|\mathbf{d}_{k_\ell}\|_2^2 = 0$$

und weiter aus Lemma 5.9 (*v.*)

$$\lim_{\ell\to\infty} |\nabla f(\mathbf{x}_{k_\ell})^T\mathbf{d}_{k_\ell}| = 0.$$

Die rechte Ungleichung in (5.11) liefert daher

$$\lim_{\ell\to\infty} \|\nabla f(\mathbf{x}_{k_\ell})\|_2 = 0.$$

Weil nach Lemma 5.9 (*iii.*) gilt

$$\lim_{\ell\to\infty} \|\mathbf{x}_{k_\ell} - \mathbf{x}_{k_\ell-1}\|_2 = \lim_{\ell\to\infty} \alpha_{k_\ell-1}\|\mathbf{d}_{k_\ell-1}\|_2 = 0,$$

schließen wir

$$\begin{aligned}\|\nabla f(\mathbf{x}_{k_\ell-1})\|_2 &\leq \|\nabla f(\mathbf{x}_{k_\ell}) - \nabla f(\mathbf{x}_{k_\ell-1})\|_2 + \|\nabla f(\mathbf{x}_{k_\ell})\|_2\\ &\leq L\|\mathbf{x}_{k_\ell} - \mathbf{x}_{k_\ell-1}\|_2 + \|\nabla f(\mathbf{x}_{k_\ell})\|_2\\ &\overset{\ell\to\infty}{\longrightarrow} 0.\end{aligned}$$

Dies steht aber im Widerspruch zur Annahme, dass $\{\|\nabla f(\mathbf{x}_{k_\ell-1})\|_2\}_{\ell>0}$ nicht nach Null konvergiert. □

6. Nichtlineare Ausgleichsprobleme

6.1 Gauß-Newton-Verfahren

Ein nichtlineares Ausgleichsproblem liegt vor, falls zu gegebenen Daten und Funktionen

$$\mathbf{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_m \end{bmatrix} \in \mathbb{R}^m, \qquad \mathbf{f}(\mathbf{x}) = \begin{bmatrix} f_1(x_1, x_2, \ldots, x_n) \\ f_2(x_1, x_2, \ldots, x_n) \\ \vdots \\ f_m(x_1, x_2, \ldots, x_n) \end{bmatrix} : \mathbb{R}^n \to \mathbb{R}^m$$

dasjenige $\mathbf{x}^\star = [x_1^\star, x_2^\star, \ldots, x_n^\star]^T$ gesucht wird, das die Minimierungsaufgabe

$$\phi(\mathbf{x}) := \frac{1}{2}\|\mathbf{y} - \mathbf{f}(\mathbf{x})\|_2^2 = \sum_{i=1}^{m} |y_i - f_i(x_1, x_2, \ldots, x_n)|^2 \to \min_{\mathbf{x}\in\mathbb{R}^n} \tag{6.1}$$

löst.

Nichtlineare Ausgleichsprobleme können im allgemeinen nur iterativ gelöst werden. Da hierzu Gradienteninformationen benötigt wird, setzen wir $\mathbf{f}$ als stetig differenzierbar voraus. Die Ableitung $\mathbf{f}'$ sei zusätzlich sogar Lipschitz-stetig.

Natürlich kann man das nichtlineare Ausgleichsproblem (6.1) mit dem Gradientenverfahren oder dem Quasi-Newton-Verfahren zu lösen. Besser ist es jedoch, es mit dem Newton-Verfahren zu lösen, was auf die Iteration

$$\mathbf{x}_{k+1} = \mathbf{x}_k - \phi''(\mathbf{x}_k)\nabla\phi(\mathbf{x}_k), \quad k = 0, 1, 2, \ldots$$

führt. Hierzu wird jedoch neben dem Gradienten

$$\nabla\phi(\mathbf{x}) = -\big(\mathbf{f}'(\mathbf{x})\big)^T\big(\mathbf{y} - \mathbf{f}(\mathbf{x})\big), \quad \mathbf{f}'(\mathbf{x}) = \begin{bmatrix} \frac{\partial f_1}{\partial x_1}(\mathbf{x}) & \frac{\partial f_1}{\partial x_2}(\mathbf{x}) & \ldots & \frac{\partial f_1}{\partial x_n}(\mathbf{x}) \\ \frac{\partial f_2}{\partial x_1}(\mathbf{x}) & \frac{\partial f_2}{\partial x_2}(\mathbf{x}) & \ldots & \frac{\partial f_2}{\partial x_n}(\mathbf{x}) \\ \vdots & \vdots & & \vdots \\ \frac{\partial f_m}{\partial x_1}(\mathbf{x}) & \frac{\partial f_m}{\partial x_2}(\mathbf{x}) & \ldots & \frac{\partial f_m}{\partial x_n}(\mathbf{x}) \end{bmatrix}$$

auch die Hesse-Matrix benötigt, das ist

$$\phi''(\mathbf{x}) = \big(\mathbf{f}'(\mathbf{x})\big)^T\mathbf{f}'(\mathbf{x}) - \big(\mathbf{y} - \mathbf{f}(\mathbf{x})\big)^T\mathbf{f}''(\mathbf{x}).$$

In der Praxis will man die Berechnung des Tensors $\mathbf{f}''(\mathbf{x}) \in \mathbb{R}^{m\times(n\times n)}$ jedoch vermeiden. Daher vernachlässigt man den Term $\big(\mathbf{y} - \mathbf{f}(\mathbf{x})\big)^T\mathbf{f}''(\mathbf{x})$ und erhält das *Gauß-Newton-Verfahren*.

Zu dessen Herleitung linearisieren wir die Funktion $\mathbf{f}$

$$\mathbf{f}(\mathbf{x}+\mathbf{d}) = \mathbf{f}(\mathbf{x}) + \mathbf{f}'(\mathbf{x})\mathbf{d} + \mathcal{O}(\|\mathbf{d}\|_2).$$

Ist nun $\mathbf{x}_k$ eine Näherungslösung des Ausgleichsproblems (6.1), so erwartet man, dass die Optimallösung $\mathbf{x}_{k+1} = \mathbf{x}_k + \mathbf{d}_k$ des linearisierten Problems

$$\min_{\mathbf{d}\in\mathbb{R}^n} \|\mathbf{y} - \mathbf{f}(\mathbf{x}_k) - \mathbf{f}'(\mathbf{x_k})\mathbf{d}\|_2^2 = \|\mathbf{r}_k - \mathbf{f}'(\mathbf{x}_k)\mathbf{d}_k\|_2^2, \qquad \mathbf{r}_k := \mathbf{y} - \mathbf{f}(\mathbf{x}_k)$$

eine bessere Lösung des Ausgleichsproblems ist. Gemäß Definition muss das Update $\mathbf{d}_k$ die Normalengleichungen

$$\big(\mathbf{f}'(\mathbf{x}_k)\big)^T\mathbf{f}'(\mathbf{x}_k)\mathbf{d}_k = \big(\mathbf{f}'(\mathbf{x}_k)\big)^T\mathbf{r}_k$$

lösen. Dies führt auf folgenden Algorithmus:

Algorithmus 6.1 (Gauß-Newton–Verfahren)
input: Funktion $\mathbf{f} : \mathbb{R}^n \to \mathbb{R}^m$, Datenvektor $\mathbf{y} \in \mathbb{R}^m$ und Startnäherung $\mathbf{x}_0 \in \mathbb{R}^n$
output: Folge von Iterierten $\{\mathbf{x}_k\}_{k\in\mathbb{N}}$

① Initialisierung: setze $k = 0$

② löse die Normalengleichungen

$$\big(\mathbf{f}'(\mathbf{x}_k)\big)^T\mathbf{f}'(\mathbf{x}_k)\mathbf{d}_k = \big(\mathbf{f}'(\mathbf{x}_k)\big)^T\big(\mathbf{y} - \mathbf{f}(\mathbf{x}_k)\big) \tag{6.2}$$

③ setze $\mathbf{x}_{k+1} := \mathbf{x}_k + \mathbf{d}_k$

④ erhöhe $k := k + 1$ und gehe nach ②

Satz 6.2 Sei $D \subset \mathbb{R}^n$ offen und $\mathbf{f} : D \to \mathbb{R}^m$ eine stetig differenzierbare Abbildung. Das Minimierungsproblem (6.1) habe eine Lösung $\mathbf{x}^\star \in D$ mit rang $\big(\mathbf{f}'(\mathbf{x}^\star)\big) = n \le m$. Sei $\lambda > 0$ der kleinste Eigenwert von $\big(\mathbf{f}'(\mathbf{x}^\star)\big)^T\mathbf{f}'(\mathbf{x}^\star)$. Ferner gelte die Lipschitz-Bedingung

$$\|\mathbf{f}'(\mathbf{x}) - \mathbf{f}'(\mathbf{z})\|_2 \le \alpha\|\mathbf{x} - \mathbf{z}\|_2 \tag{6.3}$$

und

$$\big\|\big(\mathbf{f}'(\mathbf{x}) - \mathbf{f}'(\mathbf{x}^\star)\big)^T\big(\mathbf{y} - \mathbf{f}(\mathbf{x}^\star)\big)\big\|_2 \le \beta\|\mathbf{x} - \mathbf{x}^\star\|_2 \tag{6.4}$$

mit $\beta < \lambda$ für alle $\mathbf{x}$, $\mathbf{z}$ aus einer Umgebung von $\mathbf{x}^\star$. Dann existiert ein $\varepsilon > 0$, so dass für jeden Startvektor $\mathbf{x}_0 \in B_\varepsilon(\mathbf{x}^\star) := \{\mathbf{x} \in \mathbb{R}^n : \|\mathbf{x} - \mathbf{x}^\star\|_2 < \varepsilon\}$ die Folge der Iterierten $\{\mathbf{x}_k\}_{k\ge0}$ mindestens linear gegen $\mathbf{x}^\star$ konvergiert.

Beweis. Nach Voraussetzung gibt es ein $\varepsilon_1 > 0$, so dass (6.3) und (6.4) für alle $\mathbf{x} \in B_{\varepsilon_1}(\mathbf{x}^\star)$ gelten. Aus Stetigkeitsgründen folgt außerdem die Existenz von $\gamma > 0$, so dass

$$\|\mathbf{f}'(\mathbf{x})\|_2 \le \gamma \quad \text{für alle } \mathbf{x} \in B_{\varepsilon_1}(\mathbf{x}^\star).$$

Wegen rang $\big(\mathbf{f}'(\mathbf{x}^\star)\big) = n$, ist die Matrix $\big(\mathbf{f}'(\mathbf{x}^\star)\big)^T\mathbf{f}'(\mathbf{x}^\star)$ regulär mit

$$\left\|\Big(\big(\mathbf{f}'(\mathbf{x}^\star)\big)^T\mathbf{f}'(\mathbf{x}^\star)\Big)^{-1}\right\|_2 = \frac{1}{\lambda}. \tag{6.5}$$

Daher gibt es zu beliebigem $\delta > 1$ ein $\varepsilon_2 > 0$ derart, dass $\big(\mathbf{f}'(\mathbf{x})\big)^T\mathbf{f}'(\mathbf{x})$ regulär ist und

$$\left\|\Big(\big(\mathbf{f}'(\mathbf{x})\big)^T\mathbf{f}'(\mathbf{x})\Big)^{-1}\right\|_2 \le \frac{\delta}{\lambda} \quad \text{für alle } \mathbf{x} \in B_{\varepsilon_2}(\mathbf{x}^\star). \tag{6.6}$$

Wir wählen $\delta > 1$ so, dass zusätzlich gilt

$$\delta < \frac{\lambda}{\beta}. \tag{6.7}$$

Weiter gibt es ein $\varepsilon_3 > 0$, so dass für jedes $\mathbf{x}_k \in B_{\varepsilon_3}(\mathbf{x}^\star)$ folgt

$$\begin{aligned}
&\|\mathbf{f}(\mathbf{x}^\star) - \mathbf{f}(\mathbf{x}_k) - \mathbf{f}'(\mathbf{x}_k)(\mathbf{x}^\star - \mathbf{x}_k)\|_2 \\
&\quad= \left\| \int_0^1 \mathbf{f}'\big(\mathbf{x}_k + t(\mathbf{x}^\star - \mathbf{x}_k)\big)(\mathbf{x}^\star - \mathbf{x}_k)\,\mathrm{d}t - \mathbf{f}'(\mathbf{x}_k)(\mathbf{x}^\star - \mathbf{x}_k)\right\|_2 \\
&\quad= \left\| \int_0^1 \Big(\mathbf{f}'\big(\mathbf{x}_k + t(\mathbf{x}^\star - \mathbf{x}_k)\big) - \mathbf{f}'(\mathbf{x}_k)\Big)(\mathbf{x}^\star - \mathbf{x}_k)\,\mathrm{d}t\right\|_2 \\
&\quad\le \int_0^1 \underbrace{\left\|\mathbf{f}'\big(\mathbf{x}_k + t(\mathbf{x}^\star - \mathbf{x}_k)\big) - \mathbf{f}'(\mathbf{x}_k)\right\|_2}_{\le \alpha t\|\mathbf{x}_k - \mathbf{x}^\star\|_2} \|\mathbf{x}_k - \mathbf{x}^\star\|_2\,\mathrm{d}t \\
&\quad\le \frac{\alpha}{2}\|\mathbf{x}_k - \mathbf{x}^\star\|_2^2.
\end{aligned}$$

Wir setzen nun

$$\varepsilon := \min\left\{\varepsilon_1, \varepsilon_2, \varepsilon_3, \frac{\lambda - \beta\delta}{\alpha\gamma\delta}\right\} > 0.$$

Aus (6.2) folgt dann für $\mathbf{x}_k \in B_\varepsilon(\mathbf{x}^\star)$ dass

$$\begin{aligned}
\mathbf{x}_{k+1} - \mathbf{x}^\star &= \mathbf{x}_k + \mathbf{d}_k - \mathbf{x}^\star \\
&= \Big(\big(\mathbf{f}'(\mathbf{x}_k)\big)^T\mathbf{f}'(\mathbf{x}_k)\Big)^{-1}\Big[\big(\mathbf{f}'(\mathbf{x}_k)\big)^T\big(\mathbf{y} - \mathbf{f}(\mathbf{x}_k)\big) - \big(\mathbf{f}'(\mathbf{x}_k)\big)^T\mathbf{f}'(\mathbf{x}_k)(\mathbf{x}^\star - \mathbf{x}_k)\Big] \\
&= \Big(\big(\mathbf{f}'(\mathbf{x}_k)\big)^T\mathbf{f}'(\mathbf{x}_k)\Big)^{-1}\Big[\big(\mathbf{f}'(\mathbf{x}_k)\big)^T\big(\mathbf{y} - \mathbf{f}(\mathbf{x}^\star)\big) \\
&\qquad\qquad + \big(\mathbf{f}'(\mathbf{x}_k)\big)^T\big(\mathbf{f}(\mathbf{x}^\star) - \mathbf{f}(\mathbf{x}_k) - \mathbf{f}'(\mathbf{x}_k)(\mathbf{x}^\star - \mathbf{x}_k)\big)\Big].
\end{aligned}$$

Hieraus folgt dann mit (6.5) und (6.6)

$$\begin{aligned}
\|\mathbf{x}_{k+1} - \mathbf{x}^\star\|_2 &\le \left\|\Big(\big(\mathbf{f}'(\mathbf{x}_k)\big)^T\mathbf{f}'(\mathbf{x}_k)\Big)^{-1}\right\|_2\Big[\big\|\big(\mathbf{f}'(\mathbf{x}_k)\big)^T\big(\mathbf{y} - \mathbf{f}(\mathbf{x}^\star)\big)\big\|_2 \\
&\qquad\qquad + \|\mathbf{f}'(\mathbf{x}_k)\|_2\|\mathbf{f}(\mathbf{x}^\star) - \mathbf{f}(\mathbf{x}_k) - \mathbf{f}'(\mathbf{x}_k)(\mathbf{x}^\star - \mathbf{x}_k)\|_2\Big] \\
&\le \frac{\delta}{\lambda}\left[\big\|\big(\mathbf{f}'(\mathbf{x}_k)\big)^T\big(\mathbf{y} - \mathbf{f}(\mathbf{x}^\star)\big)\big\|_2 + \frac{\alpha\gamma}{2}\|\mathbf{x}_k - \mathbf{x}^\star\|_2^2\right].
\end{aligned} \tag{6.8}$$

Wegen $\mathbf{0} = \nabla\phi(\mathbf{x}^\star) = -\big(\mathbf{f}'(\mathbf{x}^\star)\big)^T\big(\mathbf{y} - \mathbf{f}(\mathbf{x}^\star)\big)$ ergibt sich mit (6.4)

$$\big\|\big(\mathbf{f}'(\mathbf{x}_k)\big)^T\big(\mathbf{y} - \mathbf{f}(\mathbf{x}^\star)\big)\big\|_2 = \big\|\big(\mathbf{f}'(\mathbf{x}_k) - \mathbf{f}'(\mathbf{x}^\star)\big)^T\big(\mathbf{y} - \mathbf{f}(\mathbf{x}^\star)\big)\big\|_2 \le \beta\|\mathbf{x}_k - \mathbf{x}^\star\|_2,$$

dies bedeutet,

$$\|\mathbf{x}_{k+1} - \mathbf{x}^\star\|_2 \le \frac{\delta}{\lambda}\left[\beta + \frac{\alpha\gamma}{2}\underbrace{\|\mathbf{x}_k - \mathbf{x}^\star\|_2}_{\le\varepsilon\le(\lambda-\beta\delta)/(\alpha\gamma\delta)}\right]\|\mathbf{x}_k - \mathbf{x}^\star\|_2 \le \Bigg[\underbrace{\frac{\beta\delta}{\lambda} + \frac{\lambda - \beta\delta}{2\lambda}}_{=(\lambda+\beta\delta)/(2\lambda)}\Bigg]\|\mathbf{x}_k - \mathbf{x}^\star\|_2.$$

Wegen (6.7) ist die Konstante $(\lambda + \beta\delta)/(2\lambda) < 1$, das heißt, alle Iterierten $\{\mathbf{x}_k\}_{k\geq 0}$ liegen in $B_\varepsilon(\mathbf{x}^\star)$ und konvergieren mindestens linear gegen $\mathbf{x}^\star$. □

Bemerkung Die Voraussetzung (6.4) besagt im Prinzip, dass das Residuum $\mathbf{r}(\mathbf{x}^\star) = \mathbf{y} - \mathbf{f}(\mathbf{x}^\star)$ klein genug sein soll. Dies sieht man insbesondere, wenn man sie durch die stärkere Bedingung

$$\|\mathbf{y} - \mathbf{f}(\mathbf{x}^\star)\|_2 \leq \frac{\lambda}{\alpha}$$

ersetzt. Dann folgt nämlich (6.4):

$$\left\|\left(\mathbf{f}'(\mathbf{x}) - \mathbf{f}'(\mathbf{x}^\star)\right)^T\left(\mathbf{y} - \mathbf{f}(\mathbf{x}^\star)\right)\right\|_2 \leq \underbrace{\|\mathbf{f}'(\mathbf{x}) - \mathbf{f}'(\mathbf{x}^\star)\|_2}_{\leq\alpha\|\mathbf{x}-\mathbf{x}^\star\|_2}\underbrace{\|\mathbf{y} - \mathbf{f}(\mathbf{x}^\star)\|_2}_{\leq\lambda/\alpha} \leq \lambda\|\mathbf{x} - \mathbf{x}^\star\|_2.$$

△

Korollar 6.3 Zusätzlich zu den Voraussetzungen aus Satz 6.2 gelte $\mathbf{f}(\mathbf{x}^\star) = \mathbf{y}$. Dann existiert ein $\varepsilon > 0$, so dass für jeden Startvektor $\mathbf{x}_0 \in B_\varepsilon(\mathbf{x}^\star)$ die Folge der Iterierten $\{\mathbf{x}_k\}_{k\geq 0}$ quadratisch gegen $\mathbf{x}^\star$ konvergiert.

Beweis. Aus Satz 6.2 folgt die lineare Konvergenz der Iterierten $\{\mathbf{x}_k\}_{k\geq 0}$ gegen $\mathbf{x}^\star$. Zum Nachweis der quadratischen Konvergenz bemerken wir, dass aufgrund der Voraussetzung (6.4) mit $\beta = 0$ gilt. Daher folgt aus (6.8)

$$\|\mathbf{x}_{k+1} - \mathbf{x}^\star\|_2 \leq \frac{\alpha\gamma\delta}{2\lambda}\|\mathbf{x}_k - \mathbf{x}^\star\|_2^2.$$

□

6.2 Levenberg-Marquardt-Verfahren

Das Levenberg-Marquardt-Verfahren ist ein *Trust-Region-Verfahren*, also ein Verfahren, bei dem der Linearisierung nur im Bereich $\|\mathbf{d}_k\|_2 \leq \Delta_k$ vertraut wird. Demnach wollen wir das restringierte Optimierungsproblem

$$\min_{\mathbf{d}\in\mathbb{R}^n:\|\mathbf{d}\|_2\leq\Delta_k} \psi(\mathbf{d}) = \min_{\mathbf{d}\in\mathbb{R}^n:\|\mathbf{d}\|_2\leq\Delta_k} \frac{1}{2}\|\mathbf{r}_k - \mathbf{f}'(\mathbf{x_k})\mathbf{d}\|_2^2, \tag{6.9}$$

lösen, um die Iterierte dann mit der Lösung $\mathbf{d}_k$ aufzudatieren: $\mathbf{x}_{k+1} = \mathbf{x}_k + \mathbf{d}_k$.
Da die Menge $\overline{B_{\Delta_k}(\mathbf{0})}$ kompakt ist, existiert ein Minimum $\mathbf{d}_k$. Es treten dabei zwei Fälle auf:

1. Das Minimum $\mathbf{d}_k$ liegt im Inneren der Kugel $B_{\Delta_k}(\mathbf{0})$ und erfüllt

$$\nabla\psi(\mathbf{d}_k) = \left(\mathbf{f}'(\mathbf{x}_k)\right)^T\left(\mathbf{f}'(\mathbf{x}_k)\mathbf{d}_k - \mathbf{r}_k\right) = \mathbf{0}. \tag{6.10}$$

2. Das Minimum liegt auf dem Rand, erfüllt also $\|\mathbf{d}_k\|_2 = \Delta_k$. In diesem Fall muss die Höhenlinie von ψ in $\mathbf{d}_k$ genau den Kreis $\partial B_{\Delta_k}(\mathbf{0})$ tangieren, das heißt, der Gradient $\nabla\psi(\mathbf{d}_k)$ zeigt in Richtung des Nullpunkts:

$$\nabla\psi(\mathbf{d}_k) = \left(\mathbf{f}'(\mathbf{x}_k)\right)^T\left(\mathbf{f}'(\mathbf{x}_k)\mathbf{d}_k - \mathbf{r}_k\right) = -\lambda_k\mathbf{d}_k \quad \text{für ein} \quad \lambda_k \geq 0. \tag{6.11}$$

Der Parameter λ_k wird *Lagrange-Parameter* genannt. Da die Gleichung (6.10) als Spezialfall $\lambda_k = 0$ von (6.11) angesehen werden kann, erhalten wir:

Lemma 6.4 Die Lösung $\mathbf{d}_k$ des restringierten Problems (6.9) genügt der Gleichung

$$\Big(\big(\mathbf{f}'(\mathbf{x}_k)\big)^T\mathbf{f}'(\mathbf{x}_k) + \lambda_k\mathbf{I}\Big)\mathbf{d}_k = \big(\mathbf{f}'(\mathbf{x}_k)\big)^T\mathbf{r}_k \tag{6.12}$$

für ein $\lambda_k \geq 0$. Dabei ist der Wert λ_k genau dann positiv, wenn $\nabla\psi(\mathbf{d}_k) \neq \mathbf{0}$ und daher die Nebenbedingung $\|\mathbf{d}_k\|_2 = \Delta_k > 0$ gilt.

Der Vorteil des regularisierten Systems (6.12) liegt darin, dass es stets eindeutig lösbar ist, sofern $\lambda_k > 0$ ist. Die zugehörige Lösung

$$\mathbf{d}_k = \Big(\big(\mathbf{f}'(\mathbf{x}_k)\big)^T\mathbf{f}'(\mathbf{x}_k) + \lambda_k\mathbf{I}\Big)^{-1}\big(\mathbf{f}'(\mathbf{x}_k)\big)^T\mathbf{r}_k = -\Big(\big(\mathbf{f}'(\mathbf{x}_k)\big)^T\mathbf{f}'(\mathbf{x}_k) + \lambda_k\mathbf{I}\Big)^{-1}\nabla\phi(\mathbf{x}_k)$$

erfüllt offenbar die Abstiegsbedingung $\mathbf{d}_k^T\nabla\phi(\mathbf{x}_k) < 0$, es sei denn, der Punkt $\mathbf{x}_k$ ist stationär.

Bemerkung Ist $\lambda_k = 0$, dann ergibt sich ein Gauß-Newton-Schritt, während $\mathbf{d}_k$ für $\lambda_k \to \infty$ der Richtung des steilsten Abstiegs entspricht. △

Wir müssen uns noch ein Kriterium überlegen, wie wir Δ_k wählen. Eine neue Näherung $\mathbf{x}_{k+1} = \mathbf{x}_k + \mathbf{d}_k$ können wir anhand der Armijo-Goldstein-Bedingung (vergleiche (2.4)) bewerten:

$$\rho = \frac{\phi(\mathbf{x}_k + \mathbf{d}_k) - \phi(\mathbf{x}_k)}{\mathbf{d}_k^T\nabla\phi(\mathbf{x}_k)} = \frac{1}{2}\frac{\|\mathbf{y} - \mathbf{f}(\mathbf{x}_k)\|_2^2 - \|\mathbf{y} - \mathbf{f}(\mathbf{x}_k + \mathbf{d}_k)\|_2^2}{\mathbf{d}_k^T\big(\mathbf{f}'(\mathbf{x}_k)\big)^T\big(\mathbf{y} - \mathbf{f}(\mathbf{x}_k)\big)}.$$

Wir wählen zwei Toleranzgrenzen $0 < \rho^- < \rho^+ < 1$ und akzeptieren den Iterationsschritt, wenn $\rho > \rho^-$ gilt. Dann war der Trust-Region-Radius Δ_k geeignet gewählt. Ist $\rho \geq \rho^+$, so können wir Δ_k sogar vergrößern. Ist hingegen $\rho \leq \rho^-$, dann verwerfen wir den Iterationsschritt und verkleinern Δ_k. Dies führt auf den folgenden Algorithmus:

Algorithmus 6.5 (Levenberg-Marquardt-Verfahren)
input: Funktion $\mathbf{f} : \mathbb{R}^n \to \mathbb{R}^m$, Datenvektor $\mathbf{y} \in \mathbb{R}^m$ und Startnäherung $\mathbf{x}_0 \in \mathbb{R}^n$
output: Folge von Iterierten $\{\mathbf{x}_k\}_{k\in\mathbb{N}}$

① Initialisierung: wähle $0 < \rho^- < \rho^+ < 1$ und setze $\Delta_0 := 1$ und $k := 0$

② bestimme die Lösung $\mathbf{d}_k$ des restringierten Optimierungsproblems (6.9)

③ berechne

$$\rho_k = \frac{1}{2}\frac{\|\mathbf{y} - \mathbf{f}(\mathbf{x}_k)\|_2^2 - \|\mathbf{y} - \mathbf{f}(\mathbf{x}_k + \mathbf{d}_k)\|_2^2}{\mathbf{d}_k^T\big(\mathbf{f}'(\mathbf{x}_k)\big)^T\big(\mathbf{y} - \mathbf{f}(\mathbf{x}_k)\big)} \tag{6.13}$$

④ falls $\rho_k > \rho^-$ setze $\mathbf{x}_{k+1} := \mathbf{x}_k + \mathbf{d}_k$, sonst setze $\Delta_k := \Delta_k/2$ und gehe nach ②

⑤ falls $\rho_k > \rho^+$ setze $\Delta_{k+1} := 2\Delta_k$, sonst setze $\Delta_{k+1} := \Delta_k$

⑥ erhöhe $k := k + 1$ und gehe nach ②

Satz 6.6 Es sei $D \subset \mathbb{R}^n$ eine kompakte Menge, in der $\mathbf{f}$ stetig differenzierbar und $\mathbf{f}'$ zudem Lipschitz-stetig ist. Ferner sei neben $\mathbf{x}_0$ auch die gesamte Niveaumenge $\{\mathbf{x} \in \mathbb{R}^n : \phi(\mathbf{x}) \le \phi(\mathbf{x}_0)\}$ in D enthalten. Dann gilt für die Iterierten $\{\mathbf{x}_k\}_{k\ge 0}$ aus Algorithmus 6.5

$$\nabla\phi(\mathbf{x}_k) \overset{k\to\infty}{\longrightarrow} \mathbf{0}.$$

Beweis. (*i*) Zunächst beweisen wir eine obere Schranke für den Lagrange-Parameter λ_k aus (6.12). Dazu nehmen wir ohne Beschränkung der Allgemeinheit an, dass $\lambda_k > 0$ und daher $\|\mathbf{d}_k\|_2 = \Delta_k$ ist. Aus (6.12) folgt

$$\mathbf{d}_k^T\Big(\big(\mathbf{f}'(\mathbf{x}_k)\big)^T\mathbf{f}'(\mathbf{x}_k) + \lambda_k\mathbf{I}\Big)\mathbf{d}_k = \mathbf{d}_k^T\big(\mathbf{f}'(\mathbf{x}_k)\big)^T\mathbf{r}_k \le \|\mathbf{d}_k\|_2\big\|\big(\mathbf{f}'(\mathbf{x}_k)\big)^T\mathbf{r}_k\big\|_2.$$

Da $\big(\mathbf{f}'(\mathbf{x}_k)\big)^T\mathbf{f}'(\mathbf{x}_k)$ positiv semidefinit ist, kann die linke Seite nach unten durch λ_k abgeschätzt werden, so dass folgt

$$\lambda_k \le \frac{\big\|\big(\mathbf{f}'(\mathbf{x}_k)\big)^T\mathbf{r}_k\big\|_2}{\Delta_k} = \frac{\|\nabla\phi(\mathbf{x}_k)\|_2}{\Delta_k}. \tag{6.14}$$

(*ii*) Als nächstes leiten wir eine untere Schranke für den Nenner ν_k in (6.13) her. Im Fall $\lambda_k > 0$ ist $\big(\mathbf{f}'(\mathbf{x}_k)\big)^T\mathbf{f}'(\mathbf{x}_k) + \lambda_k\mathbf{I}$ positiv definit, weshalb eine Cholesky-Zerlegung $\mathbf{L}\mathbf{L}^T$ existiert. Dabei gilt

$$\|\mathbf{L}^T\mathbf{L}\|_2 = \|\mathbf{L}\mathbf{L}^T\|_2 = \big\|\big(\mathbf{f}'(\mathbf{x}_k)\big)^T\mathbf{f}'(\mathbf{x}_k) + \lambda_k\mathbf{I}\big\|_2 = \underbrace{\|\mathbf{f}'(\mathbf{x}_k)\|_2^2}_{\le c \text{ für alle } \mathbf{x}\in D} + \lambda_k \overset{(6.14)}{\le} c + \frac{\|\nabla\phi(\mathbf{x}_k)\|_2}{\Delta_k}.$$

Setzen wir $\mathbf{w} = \mathbf{L}^{-1}\nabla\phi(\mathbf{x}_k)$, so folgt unter Beachtung von (6.12) hieraus

$$\begin{aligned}\nu_k &= \big(\nabla\phi(\mathbf{x}_k)\big)^T(\mathbf{L}\mathbf{L}^T)^{-1}\nabla\phi(\mathbf{x}_k) = \mathbf{w}^T\mathbf{w}\frac{\|\nabla\phi(\mathbf{x}_k)\|_2^2}{\big(\nabla\phi(\mathbf{x}_k)\big)^T\nabla\phi(\mathbf{x}_k)} = \frac{\|\mathbf{w}\|_2^2\|\nabla\phi(\mathbf{x}_k)\|_2^2}{\mathbf{w}^T\mathbf{L}^T\mathbf{L}\mathbf{w}} \\ &\ge \frac{\|\mathbf{w}\|_2^2\|\nabla\phi(\mathbf{x}_k)\|_2^2}{\|\mathbf{w}\|_2^2\big(c + \|\nabla\phi(\mathbf{x}_k)\|_2/\Delta_k\big)} \ge \frac{\|\nabla\phi(\mathbf{x}_k)\|_2}{1+c}\min\{\Delta_k, \|\nabla\phi(\mathbf{x}_k)\|_2\}.\end{aligned} \tag{6.15}$$

Im Fall $\lambda_k = 0$ folgt

$$\nu_k = \mathbf{r}_k^T\mathbf{f}'(\mathbf{x}_k)\big(\mathbf{f}'(\mathbf{x}_k)\big)^+\mathbf{r}_k = \|\mathbf{P}_k\mathbf{r}_k\|_2^2,$$

wobei $\mathbf{P}_k = \mathbf{f}'(\mathbf{x}_k)\big(\mathbf{f}'(\mathbf{x}_k)\big)^+$ den Orthogonalprojektor auf $\operatorname{img}\big(\mathbf{f}'(\mathbf{x}_k)\big)$ bezeichnet. Wegen $\operatorname{img}\big(\mathbf{f}'(\mathbf{x}_k)\big) = \operatorname{kern}\Big(\big(\mathbf{f}'(\mathbf{x}_k)\big)^T\Big)^\perp$ folgt daher

$$\|\nabla\phi(\mathbf{x}_k)\|_2^2 = \big\|\big(\mathbf{f}'(\mathbf{x}_k)\big)^T\mathbf{r}_k\big\|_2^2 = \big\|\big(\mathbf{f}'(\mathbf{x}_k)\big)^T\mathbf{P}_k\mathbf{r}_k\big\|_2^2 \le \big\|\big(\mathbf{f}'(\mathbf{x}_k)\big)^T\big\|_2^2\|\mathbf{P}_k\mathbf{r}_k\|_2^2 \le c\nu_k,$$

das heißt, (6.15) ist auch im Fall $\lambda_k = 0$ gültig.

(*iii*) Wir beweisen nun, dass die rechte Seite von (6.15) gegen Null konvergiert. Bei erfolgreichem Iterationsschritt ist $\rho_k > \rho^-$ und aus (6.13) und (6.15) folgt

$$\phi(\mathbf{x}_k) - \phi(\mathbf{x}_{k+1}) > \rho^-\nu_k \ge \frac{\rho^-\|\nabla\phi(\mathbf{x}_k)\|_2}{1+c}\min\{\Delta_k, \|\nabla\phi(\mathbf{x}_k)\|_2\}. \tag{6.16}$$

Da nach Konstruktion $\{\phi(\mathbf{x}_k)\}_{k\geq 0}$ eine monoton fallende, nach unten beschränkte Folge ist, muss gelten

$$\min\{\Delta_k, \|\nabla\phi(\mathbf{x}_k)\|_2\} \overset{k\to\infty}{\longrightarrow} 0. \tag{6.17}$$

(iv) Als nächstes zeigen wir, dass $\|\nabla\phi(\mathbf{x}_k)\|_2$ für eine Teilfolge $\{k_n\}_{n\in\mathbb{N}}$ mit $k_n \to \infty$ gegen Null konvergiert. Angenommen, die Behauptung gilt nicht, dann folgt

$$\|\nabla\phi(\mathbf{x}_k)\|_2 \geq \varepsilon > 0 \quad \text{für alle} \quad k \geq K(\varepsilon).$$

Aus (6.17) ergibt sich damit unmittelbar

$$\Delta_k \overset{k\to\infty}{\longrightarrow} 0. \tag{6.18}$$

Taylor-Entwicklung von ρ_k liefert jedoch

$$\begin{aligned}\rho_k &= \frac{\phi(\mathbf{x}_{k+1}) - \phi(\mathbf{x}_k)}{\mathbf{d}_k^T\nabla\phi(\mathbf{x}_k)} = \frac{\mathbf{d}_k^T\nabla\phi(\mathbf{x}_k) + \mathcal{O}(\|\mathbf{d}_k\|_2^2)}{\mathbf{d}_k^T\nabla\phi(\mathbf{x}_k)} \\ &= 1 + \mathcal{O}\left(\frac{\Delta_k^2}{\nu_k}\right) \overset{(6.15)}{=} 1 + \mathcal{O}\Bigg(\frac{\Delta_k}{\underbrace{\|\nabla\phi(\mathbf{x}_k)\|_2}_{\geq\varepsilon>0}}\Bigg) = 1 + \mathcal{O}(\Delta_k) \quad \text{für } k \to \infty.\end{aligned}$$

Demnach existiert ein $M(\varepsilon) \geq K(\varepsilon)$, so dass $\rho_k > \rho^+$ für alle $k \geq M(\varepsilon)$. Ab dem $M(\varepsilon)$-ten Schritt wird folglich Δ_k in jedem Schritt von Algorithmus 6.5 verdoppelt, was jedoch im Widerspruch zu (6.18) steht.

(v) Wir beweisen nun die Aussage des Satzes. Dazu nehmen wir an, dass eine Teilfolge von $\{\|\nabla\phi(\mathbf{x}_k)\|_2\}_{k\geq 0}$ nicht gegen Null konvergiert. Nach Aussage (iv) existiert dann ein $\varepsilon > 0$ und zwei Indizes $\ell < m$, so dass

$$\|\nabla\phi(\mathbf{x}_\ell)\|_2 \geq 2\varepsilon, \qquad \|\nabla\phi(\mathbf{x}_m)\|_2 \leq \varepsilon, \qquad \|\nabla\phi(\mathbf{x}_k)\|_2 > \varepsilon, \quad k = \ell+1, \ldots, m-1.$$

Da $\{\phi(\mathbf{x}_k)\}_{k\geq 0}$ eine Cauchy-Folge ist, kann ℓ dabei so groß gewählt werden, dass

$$\phi(\mathbf{x}_\ell) - \phi(\mathbf{x}_m) < \frac{\varepsilon^2\rho^-}{(1+c)L}, \tag{6.19}$$

wobei $L > 1$ eine Lipschitz-Konstante von $\nabla\phi$ in D bezeichne. Wegen $\|\mathbf{x}_{k+1} - \mathbf{x}_k\|_2 \leq \Delta_k$ folgt aus (6.16) dass

$$\phi(\mathbf{x}_k) - \phi(\mathbf{x}_{k+1}) \geq \frac{\varepsilon\rho^-}{1+c}\min\{\|\mathbf{x}_{k+1} - \mathbf{x}_k\|_2, \varepsilon\}, \quad k = \ell, \ell+1, \ldots, m-1.$$

Summation ergibt

$$\frac{\varepsilon\rho^-}{1+c}\sum_{k=\ell}^{m-1}\min\{\|\mathbf{x}_{k+1} - \mathbf{x}_k\|_2, \varepsilon\} \leq \phi(\mathbf{x}_\ell) - \phi(\mathbf{x}_m) \overset{(6.19)}{<} \frac{\varepsilon^2\rho^-}{(1+c)L},$$

was wegen $L > 1$ nur erfüllt sein kann, wenn

$$\min\{\|\mathbf{x}_{k+1} - \mathbf{x}_k\|_2, \varepsilon\} = \|\mathbf{x}_{k+1} - \mathbf{x}_k\|_2, \quad k = \ell, \ell+1, \ldots, m-1,$$

und insgesamt

$$\sum_{k=\ell}^{m-1}\|\mathbf{x}_{k+1} - \mathbf{x}_k\|_2 < \frac{\varepsilon}{L}$$

gilt. Dies ergibt

$$\|\nabla\phi(\mathbf{x}_m) - \nabla\phi(\mathbf{x}_\ell)\|_2 \le L\|\mathbf{x}_m - \mathbf{x}_\ell\|_2 \le L\sum_{k=\ell}^{m-1}\|\mathbf{x}_{k+1} - \mathbf{x}_k\|_2 < \varepsilon$$

im Widerspruch zur Annahme. Damit ist der Satz bewiesen. □

Wir kommen nun zur Implementierung. Um das restringierte Minimierungsproblem (6.9) zu lösen, berechnen wir zunächst die Lösung $\mathbf{d}_k$ bezüglich des unrestringierten Minimierungsproblems und akzeptieren den Schritt, falls $\|\mathbf{d}_k\|_2 \le \Delta_k$. Ist hingegen $\|\mathbf{d}_k\|_2 > \Delta_k$, so wissen wir, dass das Minimum von (6.9) auf dem Rand liegt. Wir suchen dann dasjenige Tupel $(\lambda_k, \mathbf{d}_k)$, das (6.12) und $\|\mathbf{d}_k\|_2 = \Delta_k$ löst.

Es bezeichne $z_1 \ge \cdots \ge z_n \ge 0$ die Eigenwerte von $\big(\mathbf{f}'(\mathbf{x}_k)\big)^T\mathbf{f}'(\mathbf{x}_k)$ und $\{\mathbf{v}_i\}_{i=1}^n$ die zugehörigen orthonormalen Eigenvektoren. Entwickeln wir die rechte Seite von (6.12) in diese Eigenbasis

$$\big(\mathbf{f}'(\mathbf{x}_k)\big)^T\mathbf{r}_k = \sum_{i=1}^n \xi_i\mathbf{v}_i,$$

dann folgt

$$\mathbf{d}_k(\lambda_k) = \Big(\big(\mathbf{f}'(\mathbf{x}_k)\big)^T\mathbf{f}'(\mathbf{x}_k) + \lambda_k\mathbf{I}\Big)^{-1}\sum_{i=1}^n \xi_i\mathbf{v}_i = \sum_{i=1}^n \frac{\xi_i}{z_i+\lambda_k}\mathbf{v}_i.$$

Die Forderung $\|\mathbf{d}_k(\lambda_k)\|_2 = \Delta_k$ führt auf die nichtlineare Gleichung

$$r(\lambda_k) := \sum_{i=1}^n \frac{|\xi_i|^2}{|z_i+\lambda_k|^2} \stackrel{!}{=} \Delta_k^2.$$

Diese kann mit dem *Hebden-Verfahren* gelöst werden, einem Newton-Verfahren für die Gleichung

$$\frac{1}{\sqrt{r(\lambda)}} - \frac{1}{\Delta_k} \stackrel{!}{=} 0.$$

Ausgehend vom Startwert $\lambda^{(0)} = 0$ konvergiert die zugehörige Iteration

$$\lambda^{(i+1)} = \lambda^{(i)} + 2\frac{r^{3/2}(\lambda^{(i)})}{r'(\lambda^{(i)})}\left(r^{-1/2}(\lambda^{(i)}) - \frac{1}{\Delta_k}\right), \quad i = 0,1,2,\ldots.$$

sehr schnell gegen die Lösung λ_k. Die explizite Spektralzerlegung kann vermieden werden, indem man $r(\lambda) = \|\mathbf{d}_k(\lambda)\|_2^2$ und

$$r'(\lambda) = -2\mathbf{r}_k^T\mathbf{f}'(\mathbf{x}_k)\Big(\big(\mathbf{f}'(\mathbf{x}_k)\big)^T\mathbf{f}'(\mathbf{x}_k) + \lambda\mathbf{I}\Big)^{-3}\big(\mathbf{f}'(\mathbf{x}_k)\big)^T\mathbf{r}_k = -2\mathbf{d}_k(\lambda)^T\mathbf{g}(\lambda)$$
$$\text{mit} \quad \Big(\big(\mathbf{f}'(\mathbf{x}_k)\big)^T\mathbf{f}'(\mathbf{x}_k) + \lambda\mathbf{I}\Big)\mathbf{g}(\lambda) = \mathbf{d}_k(\lambda)$$

benutzt.

Für jede Hebden-Iterierte sind zwei Gleichungssysteme mit derselben Systemmatrix zu lösen. Diese entsprechen genau den Normalengleichungen zu den Ausgleichsproblemen

$$\left\|\begin{bmatrix}\mathbf{f}'(\mathbf{x}_k)\\ \sqrt{\lambda}\mathbf{I}\end{bmatrix}\mathbf{d}_k(\lambda) - \begin{bmatrix}\mathbf{r}_k\\ \mathbf{0}\end{bmatrix}\right\|_2^2 \to \min, \qquad \left\|\begin{bmatrix}\mathbf{f}'(\mathbf{x}_k)\\ \sqrt{\lambda}\mathbf{I}\end{bmatrix}\mathbf{g}(\lambda) - \begin{bmatrix}\mathbf{0}\\ \mathbf{d}_k(\lambda)/\sqrt{\lambda}\end{bmatrix}\right\|_2^2 \to \min.$$

Verwendet man die QR-Zerlegung $\mathbf{QR} = \mathbf{f}'(\mathbf{x}_k)$, so können letztere durch Anwendung von jeweils $n(n+1)/2$ Givens-Rotationen effizient gelöst werden.

7. Optimierungsprobleme mit Nebenbedingungen

7.1 Optimalitätsbedingungen erster Ordnung

Wir betrachten im folgenden Optimierungsprobleme mit Nebenbedingungen, das heißt, Probleme der Form

$$\begin{aligned} &\min_{\mathbf{x}\in\mathbb{R}^n} f(\mathbf{x}) \text{ unter den Nebenbedingungen} \\ &\qquad c_i(\mathbf{x}) = 0 \quad \text{für alle } i \in \mathcal{J}_G, \\ &\qquad c_i(\mathbf{x}) \geq 0 \quad \text{für alle } i \in \mathcal{J}_U. \end{aligned} \tag{7.1}$$

Dabei seien $\mathcal{J}_G$ und $\mathcal{J}_U$ Indexmengen für Gleichungs- und Ungleichungsnebenbedingungen.

Definition 7.1 Es bezeichne

$$G := \{\mathbf{x} \in \mathbb{R}^n : c_i(\mathbf{x}) = 0 \text{ für } i \in \mathcal{J}_G \text{ und } c_i(\mathbf{x}) \geq 0 \text{ für } i \in \mathcal{J}_U\} \subset \mathbb{R}^n$$

die **Menge der zulässigen Punkte** des Minimierungsproblems unter Nebenbedingungen (7.1). Dann verstehen wir unter einem **lokalen Minimum** des Problems (7.1) einen Punkt $\mathbf{x}^\star \in G$, für den für alle $\mathbf{x} \in U \cap G$ gilt $f(\mathbf{x}^\star) \leq f(\mathbf{x})$ mit einer Umgebung $U \subset \mathbb{R}^n$ von $\mathbf{x}^\star$.

Als generelle Voraussetzung für das folgende seien sowohl f als auch c_i für alle $i \in \mathcal{J}_G \cup \mathcal{J}_U$ stetig differenzierbar im betrachteten Bereich.

Definition 7.2 Es sei $\mathbf{x}^\star \in G$ ein lokales Minimum des Minimierungsproblems (7.1). Die Folge $\{\mathbf{x}_k\}_{k\in\mathbb{N}} \subset \mathbb{R}^n$ heißt **zulässige Folge** für das Problem (7.1), falls folgende Eigenschaften erfüllt sind:

(*i.*) Es gilt $\lim_{k\to\infty} \mathbf{x}_k = \mathbf{x}^\star \in G$.

(*ii.*) Es ist $\mathbf{x}_k \neq \mathbf{x}^\star$ für alle $k \in \mathbb{N}$.

(*iii.*) Es existiert ein $K \in \mathbb{N}$, so dass $\mathbf{x}_k \in G$ für alle $k \geq K$.

Die Richtung $\mathbf{d} \in \mathbb{R}^n$ heißt **Grenzrichtung** einer zulässigen Folge, falls

$$\lim_{\ell\to\infty} \frac{\mathbf{x}_{k_\ell} - \mathbf{x}^\star}{\|\mathbf{x}_{k_\ell} - \mathbf{x}^\star\|_2} = \mathbf{d} \tag{7.2}$$

für eine Teilfolge $\{\mathbf{x}_{k_\ell}\}_{\ell\in\mathbb{N}}$ gilt.

Satz 7.3 Ist $\mathbf{x}^\star \in G$ ein lokales Minimum des Minimierungsproblems (7.1), so gilt

$$\nabla f(\mathbf{x}^\star)^T\mathbf{d} \geq 0$$

für alle Grenzrichtungen $\mathbf{d} \in \mathbb{R}^n$ zu zulässigen Folgen mit Grenzwert $\mathbf{x}^\star$.

Beweis. Angenommen, es gibt eine zulässige Folge $\{\mathbf{x}_k\}_{k\in\mathbb{N}}$ mit $\nabla f(\mathbf{x}^\star)^T\mathbf{d} < 0$ für eine Grenzrichtung $\mathbf{d}$. Es sei $\{\mathbf{x}_{k_\ell}\}_{\ell\in\mathbb{N}}$ eine zur Grenzrichtung gehörende Teilfolge mit (7.2). Dann gilt

$$\begin{aligned} f(\mathbf{x}_{k_\ell}) &= f(\mathbf{x}^\star) + \nabla f(\mathbf{x}^\star)^T(\mathbf{x}_{k_\ell} - \mathbf{x}^\star) + o(\|\mathbf{x}_{k_\ell} - \mathbf{x}^\star\|_2) \\ &= f(\mathbf{x}^\star) + \|\mathbf{x}_{k_\ell} - \mathbf{x}^\star\|_2 \nabla f(\mathbf{x}^\star)^T\mathbf{d} + o(\|\mathbf{x}_{k_\ell} - \mathbf{x}^\star\|_2). \end{aligned}$$

Wegen $\nabla f(\mathbf{x}^\star)^\mathbf{d} < 0$ gibt es ein $L \in \mathbb{N}$, so dass

$$f(\mathbf{x}_{k_\ell}) < f(\mathbf{x}^\star) + \frac{1}{2}\|\mathbf{x}_{k_\ell} - \mathbf{x}^\star\|_2 \nabla f(\mathbf{x}^\star)^T\mathbf{d}$$

für alle $\ell \geq L$ gilt. In jeder Umgebung von $\mathbf{x}^\star$ finden wir daher ein $\mathbf{x}_{k_\ell}$ mit $f(\mathbf{x}_{k_\ell}) < f(\mathbf{x}^\star)$ im Widerspruch zur Voraussetzung. □

Mit Hilfe dieses Satzes können wir den Begriff des stationären Punktes auf das Minimierungsproblem mit Nebenbedingungen (7.1) verallgemeinern.

Definition 7.4 Der Punkt $\mathbf{x}^\star \in G$ wird als **stationärer Punkt** des Minimierungsproblems mit Nebenbedingungen (7.1) bezeichnet, falls

$$\nabla f(\mathbf{x}^\star)^T\mathbf{d} \geq 0 \tag{7.3}$$

für alle Grenzrichtungen $\mathbf{d} \in \mathbb{R}^n$ zulässiger Folgen mit Grenzwert $\mathbf{x}^\star$ gilt.

Man beachte, dass die Bedingung (7.3) im Fall eines unrestringierten Minimierungsproblems mit der üblichen Bedingung $\nabla f(\mathbf{x}^\star) = \mathbf{0}$ zusammenfällt.

Definition 7.5 Zu gegeben Punkt $\mathbf{x}^\star \in G$ sei

$$\mathcal{J}_A(\mathbf{x}^\star) := \mathcal{J}_G \cup \{i \in \mathcal{J}_U : c_i(\mathbf{x}^\star) = 0\}.$$

die **Menge der aktiven Nebenbedingungen**. Sind die zu dieser Menge gehörigen Gradienten

$$\{\nabla c_i(\mathbf{x}^\star) : i \in \mathcal{J}_A(\mathbf{x}^\star)\}$$

linear unabhängig, so sagen wir, der Punkt $\mathbf{x}^\star$ erfüllt die **LICQ-Bedingung**.

Bemerkung LICQ steht abkürzend für *linear independence constraint qualification.* △

Lemma 7.6 Sei $\mathbf{x}^\star$ ein lokales Minimum des Minimierungsproblems (7.1). Ist $\mathbf{d} \in \mathbb{R}^n$ eine Grenzrichtung einer zulässigen Folge, so gilt

$$\nabla c_i(\mathbf{x}^\star)^T \mathbf{d} = 0 \text{ für alle } i \in \mathcal{J}_G,$$
$$\nabla c_i(\mathbf{x}^\star)^T \mathbf{d} \geq 0 \text{ für alle } i \in \mathcal{J}_U \cap \mathcal{J}_A.$$

Gelten umgekehrt für eine Richtung $\mathbf{d} \in \mathbb{R}^n$ mit $\|\mathbf{d}\|_2 = 1$ diese beiden Bedingungen und ist zusätzlich die LICQ-Bedingung für $\mathbf{x}^\star$ erfüllt, dann ist $\mathbf{d}$ eine Grenzrichtung zu $\mathbf{x}^\star$.

Beweis. (*i.*) Es sei $\{\mathbf{x}_k\}_{k\in\mathbb{N}}$ eine zulässige Folge, für die $\mathbf{d}$ eine Grenzrichtung ist. Wir erhalten, gegebenfalls durch Übergang zu einer Teilfolge,

$$\lim_{k\to\infty} \frac{\mathbf{x}_k - \mathbf{x}^\star}{\|\mathbf{x}_k - \mathbf{x}^\star\|_2} = \mathbf{d}.$$

Daraus folgt

$$\mathbf{x}_k = \mathbf{x}^\star + \|\mathbf{x}_k - \mathbf{x}^\star\|_2 \mathbf{d} + o(\|\mathbf{x}_{k_\ell} - \mathbf{x}^\star\|_2).$$

Ist $i \in \mathcal{J}_G$, so haben wir für alle $k \geq K$

$$\begin{aligned} 0 &= \frac{c_i(\mathbf{x}_k)}{\|\mathbf{x}_k - \mathbf{x}^\star\|_2} \\ &= \frac{c_i(\mathbf{x}^\star) + \|\mathbf{x}_k - \mathbf{x}^\star\|_2 \nabla c_i(\mathbf{x}^\star)^T \mathbf{d} + o(\|\mathbf{x}_k - \mathbf{x}^\star\|_2)}{\|\mathbf{x}_k - \mathbf{x}^\star\|_2} \\ &= \nabla c_i(\mathbf{x}^\star)^T \mathbf{d} + \frac{o(\|\mathbf{x}_k - \mathbf{x}^\star\|_2)}{\|\mathbf{x}_k - \mathbf{x}^\star\|_2}. \end{aligned}$$

Durch den Grenzübergang $k \to \infty$ schließen wir in diesem Fall $\nabla c_i(\mathbf{x}^\star)^T \mathbf{d} = 0$.
Ist $i \in \mathcal{J}_U \cup \mathcal{J}_A(\mathbf{x}^\star)$, dann erhalten wir für alle $k \geq K$

$$0 \leq \frac{c_i(\mathbf{x}_k)}{\|\mathbf{x}_k - \mathbf{x}^\star\|_2} = \frac{o(\|\mathbf{x}_k - \mathbf{x}^\star\|_2)}{\|\mathbf{x}_k - \mathbf{x}^\star\|_2}$$

und analog zum obigen Vorgehen $\nabla c_i(\mathbf{x}^\star)^T \mathbf{d} \geq 0$. Damit ist die erste Aussage des Satzes gezeigt.
(*ii.*) Wir setzen $\mathbf{A} := [\nabla c_i(\mathbf{x}^\star)^T]_{i\in\mathcal{J}_A(\mathbf{x}^\star)}$ und beachten, dass diese Matrix aufgrund der LICQ-Bedingung vollen Rang $m := |\mathcal{J}_A(\mathbf{x}^\star)| \leq n$ besitzt. Ohne Beschränkung der Allgemeinheit können wir annehmen, dass alle Nebenbedingungen aktiv sind, also $\mathcal{J}_G \cup \mathcal{J}_U = \mathcal{J}_A(\mathbf{x}^\star)$ ist. Zur Matrix $\mathbf{A} \in \mathbb{R}^{m\times n}$ können wir eine Matrix $\mathbf{B} \in \mathbb{R}^{n\times(n-m)}$ mit vollem Rand $n - m$ finden, so dass $\mathbf{AB} = \mathbf{0}$ gilt. Dabei bilden die Spalten von $\mathbf{B}$ eine Basis des Kerns von $\mathbf{A}$.
Für $t > 0$ und $\mathbf{d} \in \mathbb{R}^n$ mit $\|\mathbf{d}\|_2 = 1$ definieren wir die Abbildung $\phi : \mathbb{R}^n \times \mathbb{R} \to \mathbb{R}^n$ durch

$$\phi(\mathbf{x}, t) = \begin{bmatrix} \mathbf{c}(\mathbf{x}) - t\mathbf{A}\mathbf{d} \\ \mathbf{B}^T(\mathbf{x} - \mathbf{x}^\star - t\mathbf{d}) \end{bmatrix} \quad \text{mit} \quad \mathbf{c}(\mathbf{x}) = [c_i(\mathbf{x})]_{i\in\mathcal{J}_A(\mathbf{x}^\star)}$$

und betrachten die Lösung des Gleichungssystems $\phi(\mathbf{x}, t) = \mathbf{0}$. Für $t = 0$ ist $\mathbf{x} = \mathbf{x}^\star$ die Lösung dieses Gleichungssystems und die Jakobi-Matrix

$$\nabla_{\mathbf{x}}\phi(\mathbf{x}^\star, 0) = \begin{bmatrix} \nabla \mathbf{c}(\mathbf{x}^\star)^T \\ \mathbf{B}^T \end{bmatrix} = \begin{bmatrix} \mathbf{A} \\ \mathbf{B}^T \end{bmatrix} \in \mathbb{R}^{n \times n}$$

ist nichtsingulär. Nach dem Satz über implizite Funktionen besitzt dieses Gleichungssystem also für hinreichend kleines $t > 0$ eine eindeutige Lösung $\mathbf{x} = \mathbf{x}(t)$.

Sei $\{t_k\}_{k \in \mathbb{N}} \subset (0, 1]$ eine Folge mit $t_k \to 0$ und für $k \geq K$ sei $\mathbf{x}_k$ die eindeutige Lsoung von $\phi(\mathbf{x}, t_k) = \mathbf{0}$. Wir zeigen, dass $\{\mathbf{x}_k\}_{k \in \mathbb{N}}$ eine zulässige Folge mit Grenzrichtung $\mathbf{d}$ ist. Zunächst folgern wir $\mathbf{x}_k \in G$ für alle $k \in \mathbb{N}$ aus

$$\begin{aligned} i \in \mathcal{J}_G &\Rightarrow c_i(\mathbf{x}_k) = t_k \nabla c_i(\mathbf{x}^\star)^T \mathbf{d} = 0, \\ i \in \mathcal{J}_U \cap \mathcal{J}_A &\Rightarrow c_i(\mathbf{x}_k) = t_k \nabla c_i(\mathbf{x}^\star)^T \mathbf{d} \geq 0. \end{aligned}$$

Aufgrund des Satzes über implizite Funktionen hängt die Lösung $\mathbf{x}_k$ des Gleichungssystems $\phi(\mathbf{x}, t) = \mathbf{0}$ stetig von t_k ab. Dies impliziert

$$\lim_{k \to \infty} \mathbf{x}_k = \mathbf{x}^\star.$$

Als nächstes zeigen wir, dass $\mathbf{x}_k \neq \mathbf{x}^\star$ gilt. Angenommen, es ist $\mathbf{x}_k = \mathbf{x}^\star$ für ein $k \in \mathbb{N}$, dann würde gelten

$$\phi(\mathbf{x}^\star, t_k) = \begin{bmatrix} \mathbf{c}(\mathbf{x}^\star) - t_k \mathbf{A}\mathbf{d} \\ \mathbf{B}^T(\mathbf{x}^\star - \mathbf{x}^\star - t_k \mathbf{d}) \end{bmatrix} = \mathbf{0}.$$

Da alle Nebenbedingungen als aktiv angenommen wurden, gilt $\mathbf{c}(\mathbf{x}^\star) = \mathbf{0}$ und es folgt

$$\begin{bmatrix} \mathbf{A} \\ \mathbf{B}^T \end{bmatrix} \mathbf{d} = \mathbf{0}.$$

Daraus ergibt sich $\mathbf{d} = \mathbf{0}$ im Widerspruch zu $\|\mathbf{d}\|_2 = 1$. Folglich ist $\{\mathbf{x}_k\}_{k \in \mathbb{N}}$ eine zulässige Folge.

Wir müssen noch nachweisen, dass $\mathbf{d}$ eine Grenzrichtung ist.

$$\begin{aligned} \mathbf{0} &= \phi(\mathbf{x}_k, t_k) \\ &= \begin{bmatrix} \mathbf{c}(\mathbf{x}_k) - t_k \mathbf{A}\mathbf{d} \\ \mathbf{B}^T(\mathbf{x}_k - \mathbf{x}^\star - t_k \mathbf{d}) \end{bmatrix} \\ &= \begin{bmatrix} \mathbf{c}(\mathbf{x}^\star) + \mathbf{A}(\mathbf{x}_k - \mathbf{x}^\star) + o(\|\mathbf{x}_k - \mathbf{x}^\star\|_2) - t_k \mathbf{A}\mathbf{d} \\ \mathbf{B}^T(\mathbf{x}_k - \mathbf{x}^\star - t_k \mathbf{d}) \end{bmatrix} \\ &= \begin{bmatrix} \mathbf{A} \\ \mathbf{B}^T \end{bmatrix} (\mathbf{x}_k - \mathbf{x}^\star - t_k \mathbf{d}) + o(\|\mathbf{x}_k - \mathbf{x}^\star\|_2). \end{aligned}$$

Durch die spezielle Wahl von

$$\mathbf{d}_k := \frac{\mathbf{x}_k - \mathbf{x}^\star}{\|\mathbf{x}_k - \mathbf{x}^\star\|_2}$$

erhalten wir hieraus

$$\lim_{k \to \infty} \left\{ \mathbf{d}_k - \frac{t_k \mathbf{d}}{\|\mathbf{x}_k - \mathbf{x}^\star\|_2} \right\} = \mathbf{0}.$$

Wegen $\|\mathbf{d}_k\|_2 = 1$ für alle $k \in \mathbb{N}$ und $\|\mathbf{d}\|_2 = 1$ folgt

$$\lim_{k \to \infty} \frac{t_k}{\|\mathbf{x}_k - \mathbf{x}^\star\|_2} = 1$$

und damit $\lim_{k \to \infty} \mathbf{d}_k = \mathbf{d}$. Dies beweist die zweite Aussage des Satzes. □

Definition 7.7 Zu einem Punkt $\mathbf{x}^\star \in G$ und aktiven Nebenbedingungen $\mathcal{J}_A(\mathbf{x}^\star)$ bezeichnen wir

$$K(\mathbf{x}^\star) := \{\mathbf{d} \in \mathbb{R}^n : \nabla c_i(\mathbf{x}^\star)^T\mathbf{d} = \mathbf{0} \text{ für } i \in \mathcal{J}_G \text{ und} \\ \nabla c_i(\mathbf{x}^\star)^T\mathbf{d} \geq \mathbf{0} \text{ für } i \in \mathcal{J}_U \cap \mathcal{J}_A(\mathbf{x}^\star)\}$$

als **Linearisierungskegel** an $\mathbf{x}^\star$.

Lemma 7.8 Genau dann gibt es keine Richtung $\mathbf{d} \in K(\mathbf{x}^\star)$ mit $\nabla f(\mathbf{x}^\star)^T\mathbf{d} < 0$, wenn ein Vektor $\boldsymbol{\lambda} = [\lambda_i]_{i\in\mathcal{J}_A(\mathbf{x}^\star)}$ existiert mit

$$\nabla f(\mathbf{x}^\star) = \sum_{i\in\mathcal{J}_A(\mathbf{x}^\star)} \lambda_i \nabla c_i(\mathbf{x}^\star) = \mathbf{A}^T\boldsymbol{\lambda}$$

und $\lambda_i \geq 0$ für $i \in \mathcal{J}_U \cap \mathcal{J}_A(\mathbf{x}^\star)$.

Beweis. (*i.*) Wir betrachten den Kegel

$$\widehat{K} := \left\{\mathbf{y} \in \mathbb{R}^n : \mathbf{y} = \sum_{i\in\mathcal{J}_A(\mathbf{x}^\star)} \lambda_i \nabla c_i(\mathbf{x}^\star),\ \lambda_i \geq 0 \text{ für } i \in \mathcal{J}_U \cap \mathcal{J}_A(\mathbf{x}^\star)\right\}.$$

Da $\widehat{K}$ offensichtlich abgeschlossen ist, lautet die Aussage des Lemmas also:

$$\nabla f(\mathbf{x}^\star)^T\mathbf{d} \geq 0 \text{ für alle Richtungen } \mathbf{d} \in K(\mathbf{x}^\star) \iff \nabla f(\mathbf{x}^\star) \in \widehat{K}.$$

(*ii.*) Seien $\nabla f(\mathbf{x}^\star) \in \widehat{K}$ und $\mathbf{d} \in K(\mathbf{x}^\star)$, dann gilt

$$\begin{aligned}\nabla f(\mathbf{x}^\star)^T\mathbf{d} &= \sum_{i\in\mathcal{J}_A(\mathbf{x}^\star)} \lambda_i \nabla c_i(\mathbf{x}^\star)^T\mathbf{d} \\ &= \sum_{i\in\mathcal{J}_G} \lambda_i \nabla c_i(\mathbf{x}^\star)^T\mathbf{d} + \sum_{i\in\mathcal{J}_U\cap\mathcal{J}_A(\mathbf{x}^\star)} \lambda_i \nabla c_i(\mathbf{x}^\star)^T\mathbf{d} \\ &\geq 0\end{aligned}$$

wegen $\lambda_i \geq 0$ und $\nabla c_i(\mathbf{x}^\star)^T\mathbf{d} \geq 0$ für alle $i \in \mathcal{J}_U \cap \mathcal{J}_A(\mathbf{x}^\star)$.
(*iii.*) Angenommen, es gilt $\nabla f(\mathbf{x}^\star) \notin \widehat{K}$, dann sei $\widehat{\mathbf{y}} \in \widehat{K}$ der Punkt mit minimalen Abstand zu $\nabla f(\mathbf{x}^\star)$, das heißt, die Lösung des Minimierungsproblems

$$\min_{\mathbf{y}\in\mathbb{R}^n} \|\mathbf{y} - \nabla f(\mathbf{x}^\star)\|_2 \text{ unter der Nebenbedingung } \mathbf{y} \in \widehat{K}.$$

Da $\widehat{K}$ ein Kegel ist, liegt mit $\widehat{\mathbf{y}}$ auch $t\widehat{\mathbf{y}}$ für alle $t \geq 0$ in $\widehat{K}$. Insbesondere muss $\widehat{\mathbf{y}}$ also erfüllen:

$$\begin{aligned}0 &= \frac{\mathrm{d}}{\mathrm{d}t}\|t\widehat{\mathbf{y}} - \nabla f(\mathbf{x}^\star)\|_2^2\Big|_{t=1} \\ &= 2t\|\widehat{\mathbf{y}}\|_2^2 - 2\widehat{\mathbf{y}}^T\nabla f(\mathbf{x}^\star)\big|_{t=1} \\ &= 2\widehat{\mathbf{y}}^T\big(\widehat{\mathbf{y}} - \nabla f(\mathbf{x}^\star)\big).\end{aligned}$$

Für beliebiges $\mathbf{y} \in \widehat{K}$ folgt aus der Konvexität von $\widehat{K}$, dass

$$\|\widehat{\mathbf{y}} + \theta(\mathbf{y} - \widehat{\mathbf{y}}) - \nabla f(\mathbf{x}^\star)\|_2^2 \geq \|\widehat{\mathbf{y}} - \nabla f(\mathbf{x}^\star)\|_2^2 \quad \text{für alle } \theta \in (0,1).$$

Deshalb ist

$$2\theta(\mathbf{y} - \widehat{\mathbf{y}})^T\big(\widehat{\mathbf{y}} - \nabla f(\mathbf{x}^\star)\big) + \theta^2\|\mathbf{y} - \widehat{\mathbf{y}}\|_2^2 \geq 0,$$

was für $\theta \to 0$ auf

$$0 \leq (\mathbf{y} - \widehat{\mathbf{y}})^T\big(\widehat{\mathbf{y}} - \nabla f(\mathbf{x}^\star)\big) = \mathbf{y}^T\big(\widehat{\mathbf{y}} - \nabla f(\mathbf{x}^\star)\big) \tag{7.4}$$

führt.

Wir zeigen nun, dass $\mathbf{d} := \widehat{\mathbf{y}} - \nabla f(\mathbf{x}^\star) \neq \mathbf{0}$ die Bedingungen $\mathbf{d} \in K(\mathbf{x}^\star)$ und $\nabla f(\mathbf{x}^\star)^T\mathbf{d} < 0$ erfüllt. Zweiteres folgt sofort aus

$$\mathbf{d}^T\nabla f(\mathbf{x}^\star) = \mathbf{d}^T(\widehat{\mathbf{y}} - \mathbf{d}) = \big(\widehat{\mathbf{y}} - \nabla f(\mathbf{x}^\star)\big)^T\widehat{\mathbf{y}} - \mathbf{d}^T\mathbf{d} = -\|\mathbf{d}\|_2^2 < 0.$$

Ersteres sieht man hingegen wie folgt ein. Durch geeignete Wahl von λ_i, $i \in \mathcal{J}_A(\mathbf{x}^\star)$, erreicht man

$$\begin{aligned} i \in \mathcal{J}_G &\Longrightarrow \nabla c_i(\mathbf{x}^\star) \in \widehat{K} \text{ und } -\nabla c_i(\mathbf{x}^\star) \in \widehat{K}, \\ i \in \mathcal{J}_U \cap \mathcal{J}_A(\mathbf{x}^\star) &\Longrightarrow \nabla c_i(\mathbf{x}^\star) \in \widehat{K}. \end{aligned}$$

Setzen wir in (7.4) die spezielle Wahl $\mathbf{y} = \pm\nabla c_i(\mathbf{x}^\star)$ falls $i \in \mathcal{J}_G$ und $\mathbf{y} = \nabla c_i(\mathbf{x}^\star)$ falls $i \in \mathcal{J}_U \cap \mathcal{J}_A(\mathbf{x}^\star)$ ein, so erhalten wir

$$\begin{aligned} i \in \mathcal{J}_G &\Longrightarrow \pm\nabla c_i(\mathbf{x}^\star)^T\mathbf{d} \geq 0 \text{ also } \nabla c_i(\mathbf{x}^\star)^T\mathbf{d} = 0, \\ i \in \mathcal{J}_U \cap \mathcal{J}_A(\mathbf{x}^\star) &\Longrightarrow \nabla c_i(\mathbf{x}^\star)^T\mathbf{d} \geq 0. \end{aligned}$$

Daraus folgt $\mathbf{d} \in K(\mathbf{x}^\star)$. □

Definition 7.9 Die zum Minmierungsproblem mit Nebenbedingungen (7.1) gehörige **Lagrange-Funktion** lautet

$$L(\mathbf{x}, \boldsymbol{\lambda}) := f(\mathbf{x}) - \sum_{i \in \mathcal{J}_G \cup \mathcal{J}_U} \lambda_i c_i(\mathbf{x}).$$

Die Parameter $\boldsymbol{\lambda} = [\lambda_i]_{i \in \mathcal{J}_G \cup \mathcal{J}_G}$ werden **Lagrange-Parameter** genannt.

Mit Hilfe der Lagrange-Funktion kann man nur ein Minimium von (7.1) charakterisieren.

Satz 7.10 (Karush, Kuhn und Tucker) Es sei $\mathbf{x}^\star \in G$ eine lokales Minimum des Minimierungsproblems mit Nebenbedingungen (7.1) und $\mathbf{x}^\star$ erfülle die LICQ-Bedingung. Dann gibt es genau einen Vektor von Lagrange-Parametern $\boldsymbol{\lambda}^\star$, so dass die folgenden **KKT-Bedingungen** erfüllt sind:

$$\begin{aligned} \nabla_{\mathbf{x}} L(\mathbf{x}^\star, \boldsymbol{\lambda}^\star) &= \mathbf{0}, \\ c_i(\mathbf{x}^\star) &= 0 \text{ für alle } i \in \mathcal{J}_G, \\ c_i(\mathbf{x}^\star) &\geq 0 \text{ für alle } i \in \mathcal{J}_U, \\ \lambda_i^\star &\geq 0 \text{ für alle } i \in \mathcal{J}_U, \\ \lambda_i^\star c_i(\mathbf{x}^\star) &= 0 \text{ für alle } i \in \mathcal{J}_G \cup \mathcal{J}_U. \end{aligned}$$

Beweis. (*i.*) Angenommen, $\mathbf{x}^\star \in G$ ist ein lokales Minimum von (7.1), an dem die LICQ-Bedingung erfüllt ist. Dann gilt nach Satz 7.3 $\nabla f(\mathbf{x}^\star)^T \mathbf{d} \geq 0$ für alle dazugehörigen Grenzrichtungen $\mathbf{d}$. Nach Lemma 7.6 sind alle Richtungen, für die die Bedingungen

$$\begin{aligned} &\nabla c_i(\mathbf{x}^\star)^T \mathbf{d} = 0 \text{ für alle } i \in \mathcal{J}_G, \\ &\nabla c_i(\mathbf{x}^\star)^T \mathbf{d} \geq 0 \text{ für alle } i \in \mathcal{J}_U \cap \mathcal{J}_A(\mathbf{x}^\star). \end{aligned}$$

efüllt sind, auch Grenzrichtungen und erfüllen somit $\nabla f(\mathbf{x}^\star)^T \mathbf{d} \geq 0$. Mit anderen Worten, für alle $\mathbf{d} \in K(\mathbf{x}^\star)$ gilt $\nabla f(\mathbf{x}^\star)^T \mathbf{d} \geq 0$. Dies impliziert nach Lemma 7.8 die Existenz von $\lambda_i \in \mathbb{R}$, so dass

$$\nabla f(\mathbf{x}^\star) = \sum_{i \in \mathcal{J}_A(\mathbf{x}^\star)} \lambda_i \nabla c_i(\mathbf{x}^\star) \text{ mit } \lambda_i \geq 0 \text{ für } i \in \mathcal{J}_U \cap \mathcal{J}_A(\mathbf{x}^\star).$$

Hierin sind die Lagrange-Parameter λ_i aufgrund der *LICQ*-Bedingung eindeutig bestimmt.
(*ii.*) Wir definieren

$$\lambda_i^\star := \begin{cases} \lambda_i, & i \in \mathcal{J}_A(\mathbf{x}^\star) \\ 0, & \text{sonst} \end{cases}$$

und weisen nach, dass damit die KKT-Bedingungen erfüllt sind. Es gilt nach (*i.*)

$$\begin{aligned} \nabla_{\mathbf{x}} L(\mathbf{x}^\star, \boldsymbol{\lambda}^\star) &= \nabla f(\mathbf{x}^\star) - \sum_{i \in \mathcal{J}_G \cup \mathcal{J}_U} \lambda_i \nabla c_i(\mathbf{x}^\star) \\ &= \nabla f(\mathbf{x}^\star) - \sum_{i \in \mathcal{J}_A(\mathbf{x}^\star)} \lambda_i \nabla c_i(\mathbf{x}^\star) \\ &= \mathbf{0}. \end{aligned}$$

Die Bedingungen $c_i(\mathbf{x}^\star) = 0$ für alle $i \in \mathcal{J}_G$ und $c_i(\mathbf{x}^\star) \geq 0$ für alle $i \in \mathcal{J}_U$ folgen aus $\mathbf{x}^\star \in G$. Weiter gilt $\lambda_i^\star = 0$ für $i \in \mathcal{J}_U \setminus \mathcal{J}_A(\mathbf{x}^\star)$ und $\lambda_i^\star \geq 0$ für $i \in \mathcal{J}_U \cap \mathcal{J}_A(\mathbf{x}^\star)$, das heißt, es ist $\lambda_i^\star \geq 0$ für $i \in \mathcal{J}_U$. Schließlich gilt $\lambda_i^\star = 0$ für $i \in \mathcal{J}_U \setminus \mathcal{J}_A(\mathbf{x}^\star)$, sowie $c_i(\mathbf{x}^\star) = 0$ für alle $i \in \mathcal{J}_A(\mathbf{x}^\star)$, und somit $\lambda_i^\star c_i(\mathbf{x}^\star) = 0$ für alle $i \in \mathcal{J}_G \cup \mathcal{J}_U$. Weil $c_i(\mathbf{x}^\star) > 0$ ist für alle $i \in \mathcal{J}_U \setminus \mathcal{J}_A$, folgt die Eindeutigkeit der $\lambda_i^\star$ aus der letzten KKT-Bedingung. □

7.2 Optimalitätsbedingungen zweiter Ordnung

Wir wenden uns nun Bedingungen zweiter Ordnung zu, wobei wir voraussetzen, dass f und c_i für alle $i \in \mathcal{J}_G \cup \mathcal{J}_U$ zweimal stetig differenzierbar seien.

Definition 7.11 Zu gegebenen Lagrange-Parametern $\boldsymbol{\lambda}^\star$ sei

$$\overline{K}(\mathbf{x}^\star, \boldsymbol{\lambda}^\star) = \{\mathbf{d} \in K(\mathbf{x}^\star) : \nabla c_i(\mathbf{x}^\star)^T \mathbf{d} = 0 \text{ für alle } i \in \mathcal{J}_U \cap \mathcal{I}_A(\mathbf{x}^\star) \text{ mit } \lambda_i^\star > 0\}.$$

Bemerkung Wegen $\overline{K}(\mathbf{x}^\star, \boldsymbol{\lambda}^\star) \subset K(\mathbf{x}^\star)$ gilt offensichtlich

$$\mathbf{d} \in \overline{K}(\mathbf{x}^\star, \boldsymbol{\lambda}^\star) \Longleftrightarrow \begin{cases} \nabla c_i(\mathbf{x}^\star)^T \mathbf{d} = 0, & i \in \mathcal{J}_G, \\ \nabla c_i(\mathbf{x}^\star)^T \mathbf{d} = 0, & i \in \mathcal{J}_U \cap \mathcal{J}_A(\mathbf{x}^\star) \text{ mit } \lambda_i^\star > 0, \\ \nabla c_i(\mathbf{x}^\star)^T \mathbf{d} \geq 0, & i \in \mathcal{J}_U \cap \mathcal{J}_A(\mathbf{x}^\star) \text{ mit } \lambda_i^\star = 0. \end{cases}$$

Deshalb folgt aus der ersten KKT-Bedingung aus Satz 7.10 die Aussage

$$\mathbf{d} \in \overline{K}(\mathbf{x}^\star, \boldsymbol{\lambda}^\star) \Longrightarrow \nabla f(\mathbf{x}^\star)^T \mathbf{d} = \sum_{i \in \mathcal{J}_G \cup \mathcal{J}_U} \lambda_i^\star \nabla c_i(\mathbf{x}^\star)^T \mathbf{d} = 0.$$

△

Satz 7.12 (hinreichende Bedingung 2. Ordnung) Für einen zulässigen Punkt $\mathbf{x}^\star \in G$ gebe es einen Vektor von Lagrange-Parametern $\boldsymbol{\lambda}^\star$, so dass die KKT-Bedingungen erfüllt sind. Weiter sei

$$\mathbf{d}^T \nabla_{\mathbf{x}}^2 L(\mathbf{x}^\star, \boldsymbol{\lambda}^\star) \mathbf{d} > 0 \text{ für alle } \mathbf{d} \in \overline{K}(\mathbf{x}^\star, \boldsymbol{\lambda}^\star) \setminus \{\mathbf{0}\}.$$

Dann ist $\mathbf{x}^\star$ Lösung des Minimierungsproblems mit Nebenbedingungen (7.1).

Beweis. Der Satz ist bewiesen, wenn wir zeigen können, dass zu jeder zulässigen Folge $\{\mathbf{x}_k\}$ ein $M \in \mathbb{N}$ existiert, so dass $f(\mathbf{x}_k) > f(\mathbf{x}^\star)$ für alle $k \geq M$ ist.
Zu einer beliebigen zulässigen Folge $\{\mathbf{x}_k\}$ sei $\mathbf{d}$ eine beliebige Grenzrichtung. Nach Lemma 7.6 gilt $\mathbf{d} \in K(\mathbf{x}^\star)$. Nach der Definition der Grenzrichtung gibt es eine Teilfolge $\{\mathbf{x}_{k_\ell}\}$, welche

$$\mathbf{x}_{k_\ell} - \mathbf{x}^\star = \|\mathbf{x}_{k_\ell} - \mathbf{x}^\star\|_2 \mathbf{d} + o(\|\mathbf{x}_{k_\ell} - \mathbf{x}^\star\|_2) \tag{7.5}$$

erfüllt. Wegen den KKT-Bedingungen gilt für die Lagrange-Funktion

$$L(\mathbf{x}_{k_\ell}, \boldsymbol{\lambda}^\star) = f(\mathbf{x}_{k_\ell}) - \sum_{i \in \mathcal{J}_A(\mathbf{x}^\star)} \lambda_i^\star c_i(\mathbf{x}_{k_\ell}) \leq f(\mathbf{x}_{k_\ell}).$$

Taylor-Entwicklung ergibt

$$\begin{aligned} L(\mathbf{x}_{k_\ell}, \boldsymbol{\lambda}^\star) &= \underbrace{L(\mathbf{x}^\star, \boldsymbol{\lambda}^\star)}_{=f(\mathbf{x}^\star)} + \underbrace{\nabla_{\mathbf{x}} L(\mathbf{x}^\star, \boldsymbol{\lambda}^\star)}_{=0}{}^T (\mathbf{x}_{k_\ell} - \mathbf{x}^\star) \\ &\quad + \frac{1}{2} (\mathbf{x}_{k_\ell} - \mathbf{x}^\star)^T \nabla_{\mathbf{x}}^2 L(\mathbf{x}^\star, \boldsymbol{\lambda}^\star)(\mathbf{x}_{k_\ell} - \mathbf{x}^\star) + o(\|\mathbf{x}_{k_\ell} - \mathbf{x}^\star\|_2^2) \\ &= f(\mathbf{x}^\star) + \frac{1}{2} (\mathbf{x}_{k_\ell} - \mathbf{x}^\star)^T \nabla_{\mathbf{x}}^2 L(\mathbf{x}^\star, \boldsymbol{\lambda}^\star)(\mathbf{x}_{k_\ell} - \mathbf{x}^\star) + o(\|\mathbf{x}_{k_\ell} - \mathbf{x}^\star\|_2^2), \end{aligned}$$

wobei wieder die KKT-Bedingungen verwendet wurden.
Wir unterscheiden nun zwei Fälle: $\mathbf{d} \notin \overline{K}(\mathbf{x}^\star, \boldsymbol{\lambda}^\star)$ und $\mathbf{d} \in \overline{K}(\mathbf{x}^\star, \boldsymbol{\lambda}^\star)$.
Für $\mathbf{d} \notin \overline{K}(\mathbf{x}^\star, \boldsymbol{\lambda}^\star)$ gibt es ein $i^\star \in \mathcal{J}_U \cap \mathcal{I}_A(\mathbf{x}^\star)$, für welches

$$\lambda_{i^\star}^\star \nabla c_{i^\star}(\mathbf{x}^\star)^T \mathbf{d} > 0$$

gilt. Eine Taylor-Entwicklung führt auf

$$\begin{aligned} \lambda_{i^\star}^\star c_{i^\star}(\mathbf{x}_{k_\ell}) &= \underbrace{\lambda_{i^\star}^\star c_{i^\star}(\mathbf{x}^\star)}_{=0} + \lambda_{i^\star}^\star \nabla c_{i^\star}(\mathbf{x}^\star)^T (\mathbf{x}_{k_\ell} - \mathbf{x}^\star) + o(\|\mathbf{x}_{k_\ell} - \mathbf{x}^\star\|_2) \\ &\overset{(7.5)}{=} \lambda_{i^\star}^\star \|\mathbf{x}_{k_\ell} - \mathbf{x}^\star\|_2 \nabla c_{i^\star}(\mathbf{x}^\star)^T \mathbf{d} + o(\|\mathbf{x}_{k_\ell} - \mathbf{x}^\star\|_2). \end{aligned}$$

Für die Lagrange-Funktion gilt somit

$$\begin{aligned} L(\mathbf{x}_{k_\ell}, \boldsymbol{\lambda}^\star) &= f(\mathbf{x}_{k_\ell}) - \sum_{i \in \mathcal{J}_A(\mathbf{x}^\star)} \lambda_i^\star c_i(\mathbf{x}_{k_\ell}) \\ &\leq f(\mathbf{x}_{k_\ell}) - \lambda_{i^\star}^\star c_{i^\star}(\mathbf{x}_{k_\ell}) \\ &= f(\mathbf{x}_{k_\ell}) - \lambda_{i^\star}^\star \|\mathbf{x}_{k_\ell} - \mathbf{x}^\star\|_2 \nabla c_{i^\star}(\mathbf{x}^\star)^T \mathbf{d} + o(\|\mathbf{x}_{k_\ell} - \mathbf{x}^\star\|_2). \end{aligned}$$

Andererseits wurde oben gezeigt, dass

$$L(\mathbf{x}_{k_\ell}, \boldsymbol{\lambda}^\star) = f(\mathbf{x}^\star) + o(\|\mathbf{x}_{k_\ell} - \mathbf{x}^\star\|_2).$$

Daraus ergibt sich nun

$$f(\mathbf{x}_{k_\ell}) \geq f(\mathbf{x}^\star) + \lambda_{i^\star}^\star \|\mathbf{x}_{k_\ell} - \mathbf{x}^\star\|_2 \nabla c_{i^\star}(\mathbf{x}^\star)^T \mathbf{d} + o(\|\mathbf{x}_{k_\ell} - \mathbf{x}^\star\|_2)$$

und wegen $\lambda_{i^\star} \nabla c_{i^\star}(\mathbf{x}^\star)^T \mathbf{d} > 0$ impliziert dies $f(\mathbf{x}_{k_\ell}) > f(\mathbf{x}^\star)$, falls nur k_ℓ genügend groß ist.
Für $\mathbf{d} \in \overline{K}(\mathbf{x}^\star, \boldsymbol{\lambda}^\star)$ erhalten wir direkt

$$\begin{aligned} f(\mathbf{x}_{k_\ell}) &\geq L(\mathbf{x}_{k_\ell}, \boldsymbol{\lambda}^\star) \\ &= f(\mathbf{x}^\star) + \frac{1}{2}(\mathbf{x}_{k_\ell} - \mathbf{x}^\star)^T \nabla_{\mathbf{x}}^2 L(\mathbf{x}^\star, \boldsymbol{\lambda}^\star)(\mathbf{x}_{k_\ell} - \mathbf{x}^\star) + o(\|\mathbf{x}_{k_\ell} - \mathbf{x}^\star\|_2^2) \\ &\overset{(7.5)}{=} f(\mathbf{x}^\star) + \frac{1}{2}\|\mathbf{x}_{k_\ell} - \mathbf{x}^\star\|_2^2 \mathbf{d}^T \nabla_{\mathbf{x}}^2 L(\mathbf{x}^\star, \boldsymbol{\lambda}^\star)\mathbf{d} + o(\|\mathbf{x}_{k_\ell} - \mathbf{x}^\star\|_2^2), \end{aligned}$$

woraus sich wegen $\mathbf{d}^T \nabla_{\mathbf{x}}^2 L(\mathbf{x}^\star, \boldsymbol{\lambda}^\star)\mathbf{d} > 0$ wiederum $f(\mathbf{x}_{k_\ell}) > f(\mathbf{x}^\star)$ für genügend große k_ℓ ergibt. □

Bemerkung Die notwendige Bedingung zweiter Ordnung für ein Minimum von (7.1) lautet: Ist $\mathbf{x}^\star \in G$ ein Minimum von (7.1), an dem die LICQ-Bedingung erfüllt ist, und genügen die zugehörigen Lagrange-Parameter $\boldsymbol{\lambda}^\star$ den KKT-Bedingungen, dann gilt

$$\mathbf{d}^T \nabla_{\mathbf{x}}^2 L(\mathbf{x}^\star, \boldsymbol{\lambda}^\star)\mathbf{d} \geq 0 \text{ für alle } \mathbf{d} \in \overline{K}(\mathbf{x}^\star, \boldsymbol{\lambda}^\star).$$

△

8. Projiziertes Gradientenverfahren

8.1 Konvergenzeigenschaften

Ziel dieses Kapitels ist es, das Gradientenverfahren aus Kapitel 2 auf das Minimierungsproblem mit Nebenbedingungen (7.1) zu verallgmeinern. Dazu sei neben der stetigen Differenzierbarkeit der Funktion $f : \mathbb{R}^n \to \mathbb{R}$ vorausgesetzt, dass die zulässige Menge $G \subset \mathbb{R}^n$ konvex ist.

Grundlage des projizierten Gradientenverfahren ist die orthogonale Projektion:

Definition 8.1 Es sei $G \subset \mathbb{R}^n$ eine abgeschlossene, konvexe Menge. Dann ist die **orthogonale Projektion** $\mathbf{P}_G : \mathbb{R}^n \to G$ definiert durch die Bedingung

$$\|\mathbf{P}_G(\mathbf{x}) - \mathbf{x}\|_2 = \min_{\mathbf{y} \in G} \|\mathbf{y} - \mathbf{x}\|_2.$$

Der Punkt $\mathbf{P}_G(\mathbf{x}) \in G$ besitzt also die Eigenschaft, den kürzesten Abstand zu einem gegebenen Punkt $\mathbf{x} \in \mathbb{R}^n$ zu besitzen.

Die Berechnung des projizierten Punktes $\mathbf{P}_G(\mathbf{x})$ lässt sich für viele spezielle Nebenbedingungen relativ einfach umsetzen. Beispielsweise gilt dies für affine Nebenbedingungen, die später genauer behandelt werden.

Die Grundversion des projizierten Gradientenverfahren ist im folgenden Algorithmus beschrieben:

Algorithmus 8.2 (projiziertes Gradientenverfahren)

input: Funktion $f : \mathbb{R}^n \to \mathbb{R}$, konvexe zulässige Menge $G \subset \mathbb{R}^n$ und Startnäherung $\mathbf{x}_0 \in G$

output: Folge von Iterierten $\{\mathbf{x}_k\}_{k \in \mathbb{N}}$

① Initialisierung: wähle $\sigma \in (0, 1)$ und setze $k := 0$

② berechne den Antigradienten $\mathbf{d}_k := -\nabla f(\mathbf{x}_k)$ und setze $\alpha_k := 1$

③ solange

$$f\big(\mathbf{P}_G(\mathbf{x}_k + \alpha_k \mathbf{d}_k)\big) > f(\mathbf{x}_k) - \sigma \mathbf{d}_k^T \big(\mathbf{P}_G(\mathbf{x}_k + \alpha_k \mathbf{d}_k) - \mathbf{x}_k\big) \tag{8.1}$$

setze $\alpha_k := \alpha_k / 2$

④ setze $\mathbf{x}_{k+1} := \mathbf{P}_G(\mathbf{x}_k + \alpha_k \mathbf{d}_k)$

⑤ erhöhe $k := k + 1$ und gehe nach ②

Für den Fall der Minimierung ohne Nebenbedingungen, das heißt $G = \mathbb{R}^n$, stellt obiger Algorithmus genau das bekannte Gradientverfahren 2.4 dar. Insbesondere geht die Bedingung an die Reduktion des Funktionals über in die Armijo-Goldstein-Bedingung

$$f(\mathbf{x}_{k+1}) \leq f(\mathbf{x}_k) - \sigma\alpha_k \|\nabla f(\mathbf{x}_k)\|_2^2.$$

Ähnlich wie früher kann man zeigen, dass auch für das projizierte Gradientenverfahren ein $\alpha_k > 0$ existiert, für das die Reduktionsbedingung erfüllt ist.

Lemma 8.3 Ist die zulässige Menge $G \subset \mathbb{R}^n$ konvex, so erfüllt die orthogonale Projektion $\mathbf{P}_G$ die folgenden Eigenschaften:

(*i.*) Es gilt $\big(\mathbf{P}_G(\mathbf{x}) - \mathbf{x}\big)^T\big(\mathbf{P}_G(\mathbf{x}) - \mathbf{y}\big) \leq 0$ für alle $\mathbf{x} \in \mathbb{R}^n$, $\mathbf{y} \in G$.

(*ii.*) Es gilt $\big(\mathbf{P}_G(\mathbf{y}) - \mathbf{P}_G(\mathbf{x})\big)^T(\mathbf{y} - \mathbf{x}) \geq \|\mathbf{P}_G(\mathbf{y}) - \mathbf{P}_G(\mathbf{x})\|_2^2 \geq 0$ für alle $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$, das heißt, $\mathbf{P}_G$ ist monoton.

(*iii.*) Es gilt $\|\mathbf{P}_G(\mathbf{y}) - \mathbf{P}_G(\mathbf{x})\|_2 \leq \|\mathbf{y} - \mathbf{x}\|_2$ für alle $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$, das heißt, $\mathbf{P}_G$ ist nicht expandierend.

Beweis. (*i.*) Wegen der Konvexität folgt aus $\mathbf{y} \in G$ auch $\widehat{\mathbf{y}} := (1-t)\mathbf{P}_G(\mathbf{x}) + t\mathbf{y} \in G$ für alle $t \in [0,1]$. Aus

$$\begin{aligned}\|\widehat{\mathbf{y}} - \mathbf{x}\|_2^2 &= \|\widehat{\mathbf{y}} - \mathbf{P}_G(\mathbf{x}) + \mathbf{P}_G(\mathbf{x}) - \mathbf{x}\|_2^2 \\ &= \|\widehat{\mathbf{y}} - \mathbf{P}_G(\mathbf{x})\|_2^2 + \|\mathbf{P}_G(\mathbf{x}) - \mathbf{x}\|_2^2 - 2\big(\mathbf{P}_G(\mathbf{x}) - \mathbf{x}\big)^T\big(\mathbf{P}_G(\mathbf{x}) - \widehat{\mathbf{y}}\big)\end{aligned}$$

folgt aufgrund der Minimierungseigenschaft von $\mathbf{P}_G$, dass

$$\|\widehat{\mathbf{y}} - \mathbf{P}_G(\mathbf{x})\|_2^2 - 2\big(\mathbf{P}_G(\mathbf{x}) - \mathbf{x}\big)^T\big(\mathbf{P}_G(\mathbf{x}) - \widehat{\mathbf{y}}\big) = \|\widehat{\mathbf{y}} - \mathbf{x}\|_2^2 - \|\mathbf{P}_G(\mathbf{x}) - \mathbf{x}\|_2^2 \geq 0.$$

Einsetzen von $\mathbf{P}_G(\mathbf{x}) - \widehat{\mathbf{y}} = t\big(\mathbf{P}_G(\mathbf{x}) - \mathbf{y}\big)$ führt auf

$$t^2\|\mathbf{y} - \mathbf{P}_G(\mathbf{x})\|_2^2 - 2t\big(\mathbf{P}_G(\mathbf{x}) - \mathbf{x}\big)^T\big(\mathbf{P}_G(\mathbf{x}) - \mathbf{y}\big) \geq 0,$$

was für $t \to 0$ die gewünschte Aussage liefert.

(*ii.*) Die bereits bewiesene Aussage (*i.*) impliziert

$$\begin{aligned}\big(\mathbf{P}_G(\mathbf{x}) - \mathbf{x}\big)^T\big(\mathbf{P}_G(\mathbf{x}) - \mathbf{P}_G(\mathbf{y})\big) &\leq 0, \\ \big(\mathbf{P}_G(\mathbf{y}) - \mathbf{y}\big)^T\big(\mathbf{P}_G(\mathbf{y}) - \mathbf{P}_G(\mathbf{x})\big) &\leq 0.\end{aligned}$$

Zusammen führt dies auf

$$\big(\mathbf{P}_G(\mathbf{y}) - \mathbf{y} + \mathbf{x} - \mathbf{P}_G(\mathbf{x})\big)^T\big(\mathbf{P}_G(\mathbf{y}) - \mathbf{P}_G(\mathbf{x})\big) \leq 0,$$

das ist Aussage (*ii.*).

(*iii.*) Diese Aussage folgt sofort aus Aussage (*ii.*) durch Anwenden der Cauchy-Schwarzschen Ungleichung. □

Bemerkung Aus der Monotonieeigenschaft (*ii.*) folgt wegen $\mathbf{x}_k = \mathbf{P}_G(\mathbf{x}_k)$ für die neue Iterierte $\mathbf{x}_{k+1} = \mathbf{P}_G\big(\mathbf{x}_k - \alpha_k\nabla f(\mathbf{x}_k)\big)$ des projizierten Gradientenverfahrens, dass

$$\|\mathbf{x}_{k+1} - \mathbf{x}_k\|_2^2 \leq (\mathbf{x}_{k+1} - \mathbf{x}_k)^T\big(\mathbf{x}_k - \alpha_k\nabla f(\mathbf{x}_k) - \mathbf{x}_k\big).$$

Wir erhalten daher

$$-\nabla f(\mathbf{x}_k)^T(\mathbf{x}_{k+1}-\mathbf{x}_k) \geq \frac{1}{\alpha_k}\|\mathbf{x}_{k+1}-\mathbf{x}_k\|_2^2. \tag{8.2}$$

Dies bedeutet, dass durch die Abstiegsbedingung (8.1) des Algorithmus 8.2 tatsächlich $f(\mathbf{x}_{k+1}) < f(\mathbf{x}_k)$ erreicht wird. △

Lemma 8.4 Ist die zulässige Menge $G \subset \mathbb{R}^n$ konvex, so ist für beliebige $\mathbf{x} \in \mathbb{R}^n$ und $\mathbf{d} \in \mathbb{R}^n$ die Funktion

$$\varphi(\alpha) := \frac{\|\mathbf{P}_G(\mathbf{x}+\alpha\mathbf{d})-\mathbf{x}\|_2}{\alpha}$$

für alle $\alpha > 0$ monoton fallend.

Beweis. (*i.*) Für $0 < \alpha < \beta$ setzen wir

$$\mathbf{u} := \mathbf{P}_G(\mathbf{x}+\alpha\mathbf{d})-\mathbf{x}, \quad \mathbf{v} := \mathbf{P}_G(\mathbf{x}+\beta\mathbf{d})-\mathbf{x}$$

und erhalten unter Verwendung von Lemma 8.3 (*i.*)

$$\begin{aligned}\mathbf{u}^T(\mathbf{u}-\mathbf{v}) &= \{\mathbf{P}_G(\mathbf{x}+\alpha\mathbf{d})-(\mathbf{x}+\alpha\mathbf{d})+\alpha\mathbf{d}\}^T\{\mathbf{P}_G(\mathbf{x}+\alpha\mathbf{d})-\mathbf{P}_G(\mathbf{x}+\beta\mathbf{d})\}\\ &\leq \alpha\mathbf{d}^T\{\mathbf{P}_G(\mathbf{x}+\alpha\mathbf{d})-\mathbf{P}_G(\mathbf{x}+\beta\mathbf{d})\}\end{aligned}$$

und analog

$$\mathbf{v}^T(\mathbf{v}-\mathbf{u}) \leq \beta\mathbf{d}^T\{\mathbf{P}_G(\mathbf{x}+\beta\mathbf{d})-\mathbf{P}_G(\mathbf{x}+\alpha\mathbf{d})\}.$$

Zusammen ergibt dies

$$\frac{\mathbf{u}^T(\mathbf{u}-\mathbf{v})}{\alpha} \leq \frac{\mathbf{v}^T(\mathbf{u}-\mathbf{v})}{\beta}. \tag{8.3}$$

(*ii.*) Weiter erhalten wir mit Lemma 8.3 (*ii.*)

$$\begin{aligned}\mathbf{u}^T(\mathbf{u}-\mathbf{v}) &\leq \alpha\mathbf{d}^T\{\mathbf{P}_G(\mathbf{x}+\alpha\mathbf{d})-\mathbf{P}_G(\mathbf{x}+\beta\mathbf{d})\}\\ &= -\frac{\alpha}{\beta-\alpha}(\beta\mathbf{d}-\alpha\mathbf{d})^T\{\mathbf{P}_G(\mathbf{x}+\beta\mathbf{d})-\mathbf{P}_G(\mathbf{x}+\alpha\mathbf{d})\}\\ &\leq -\frac{\alpha}{\beta-\alpha}\|\mathbf{P}_G(\mathbf{x}+\beta\mathbf{d})-\mathbf{P}_G(\mathbf{x}+\alpha\mathbf{d})\|_2^2\\ &\leq 0.\end{aligned}$$

Aus der Cauchy-Schwarzschen Ungleichung ergibt sich

$$\mathbf{u}^T\mathbf{v}(\|\mathbf{u}\|_2+\|\mathbf{v}\|_2) \leq \|\mathbf{u}\|_2\|\mathbf{v}\|_2(\|\mathbf{u}\|_2+\|\mathbf{v}\|_2),$$

woraus

$$\|\mathbf{u}\|_2\mathbf{v}^T(\mathbf{u}-\mathbf{v}) = \|\mathbf{u}\|_2(\mathbf{u}^T\mathbf{v}-\|\mathbf{v}\|_2^2) \leq \|\mathbf{v}\|_2(\|\mathbf{u}\|_2^2-\mathbf{u}^T\mathbf{v}) = \|\mathbf{v}\|_2\mathbf{u}^T(\mathbf{u}-\mathbf{v}) \tag{8.4}$$

folgt.

(*iii.*) Wir unterscheiden nun zwei Fälle: Für $\mathbf{u}^T(\mathbf{u}-\mathbf{v}) = 0$ gilt $\mathbf{P}_G(\mathbf{x}+\alpha\mathbf{d}) = \mathbf{P}_G(\mathbf{x}+\beta\mathbf{d})$ und somit $\mathbf{u} = \mathbf{v}$. Hieraus folgt unmittelbar auch

$$\varphi(\alpha) = \frac{\|\mathbf{u}\|_2}{\alpha} \geq \frac{\|\mathbf{v}\|_2}{\beta} = \varphi(\beta).$$

Für den Fall $\mathbf{u}^T(\mathbf{u}-\mathbf{v}) < 0$ folgt aus (8.3)

$$\frac{\beta}{\alpha} \geq \frac{\mathbf{v}^T(\mathbf{u}-\mathbf{v})}{\mathbf{u}^T(\mathbf{u}-\mathbf{v})}$$

und aus (8.4)

$$\|\mathbf{v}\|_2 \leq \|\mathbf{u}\|_2 \frac{\mathbf{v}^T(\mathbf{u}-\mathbf{v})}{\mathbf{u}^T(\mathbf{u}-\mathbf{v})}.$$

Kombiniert man diese zwei Ungleichungen, so erhält man wieder

$$\varphi(\alpha) = \frac{\|\mathbf{u}\|_2}{\alpha} \geq \frac{\|\mathbf{v}\|_2}{\beta} = \varphi(\beta).$$

□

Satz 8.5 Die zulässige Menge $G \subset \mathbb{R}^n$ sei konvex und die Funktion $f : \mathbb{R}^n \to \mathbb{R}$ sei auf G stetig differenzierbar und nach unten beschränkt. Weiter sei ∇f auf G gleichmäßig stetig. Dann gilt für die Iterierten $\{\mathbf{x}_k\}_{k\in\mathbb{N}}$ des projizierten Gradientenverfahrens

$$\lim_{k\to\infty} \frac{\|\mathbf{x}_{k_\ell+1} - \mathbf{x}_k\|_2}{\alpha_k} = 0.$$

Beweis. Wir führen einen Widerspruchsbeweis. Angenommen, es existiert zu jedem $\varepsilon > 0$ eine unendliche Teilfolge $\{k_\ell\}_{\ell\in\mathbb{N}}$, so dass

$$\frac{\|\mathbf{x}_{k_\ell+1} - \mathbf{x}_{k_\ell}\|_2}{\alpha_{k_\ell}} \geq \varepsilon.$$

Dann gilt insbesondere auch

$$\frac{\|\mathbf{x}_{k_\ell+1} - \mathbf{x}_{k_\ell}\|_2^2}{\alpha_{k_\ell}} \geq \varepsilon \max\{\varepsilon\alpha_{k_\ell}, \|\mathbf{x}_{k_\ell+1} - \mathbf{x}_{k_\ell}\|_2\}. \tag{8.5}$$

Da die Folge $\{f(\mathbf{x}_{k_\ell})\}_{\ell\in\mathbb{N}}$ monoton fallend und nach unten beschränkt ist, folgt aus der Abstiegsbedingung (8.1) des projizierten Gradientenverfahrens

$$\lim_{\ell\to\infty} \nabla f(\mathbf{x}_{k_\ell})^T(\mathbf{x}_{k_\ell+1} - \mathbf{x}_{k_\ell}) = 0,$$

was wiederum gemäß (8.2)

$$\lim_{\ell\to\infty} \frac{\|\mathbf{x}_{k_\ell+1} - \mathbf{x}_{k_\ell}\|_2^2}{\alpha_{k_\ell}} = 0.$$

nach sich zieht. Aufgrund von (8.5) erhalten wir hieraus

$$\lim_{k\to\infty} \alpha_{k_\ell} = 0 \quad \text{und} \quad \lim_{k\to\infty} \|\mathbf{x}_{k_\ell+1} - \mathbf{x}_{k_\ell}\|_2 = 0.$$

Für $\mathbf{y}_{k_\ell+1} := \mathbf{P}_G(\mathbf{x}_{k_\ell} + 2\alpha_{k_\ell}\mathbf{d}_{k_\ell})$ gilt aufgrund der algorithmischen Umsetzung des projizierten Gradientenverfahrens

$$f(\mathbf{y}_{k_\ell+1}) > f(\mathbf{x}_{k_\ell}) + \sigma\nabla f(\mathbf{x}_{k_\ell})^T(\mathbf{y}_{k_\ell+1} - \mathbf{x}_{k_\ell}),$$

also auch

$$f(\mathbf{x}_{k_\ell}) - f(\mathbf{y}_{k_\ell+1}) < \sigma \nabla f(\mathbf{x}_{k_\ell})^T (\mathbf{x}_{k_\ell} - \mathbf{y}_{k_\ell+1}). \tag{8.6}$$

Aus Lemma 8.4 folgt

$$\frac{\|\mathbf{x}_{k_\ell+1} - \mathbf{x}_{k_\ell}\|_2^2}{\alpha_{k_\ell}} \geq \|\mathbf{x}_{k_\ell+1} - \mathbf{x}_{k_\ell}\|_2 \frac{\|\mathbf{y}_{k_\ell+1} - \mathbf{x}_{k_\ell}\|_2}{2\alpha_{k_\ell}} \geq \alpha_{k_\ell} \varepsilon \frac{\|\mathbf{y}_{k_\ell+1} - \mathbf{x}_{k_\ell}\|_2}{2\alpha_{k_\ell}} = \frac{\varepsilon}{2} \|\mathbf{y}_{k_\ell+1} - \mathbf{x}_{k_\ell}\|_2.$$

Weiter ergibt sich mit Lemma 8.3 (*ii.*)

$$(\mathbf{x}_{k_\ell+1} - \mathbf{x}_{k_\ell})^T \Big\{ \underbrace{\mathbf{x}_{k_\ell} - \alpha_{k_\ell} \nabla f(\mathbf{x}_{k_\ell}) - \mathbf{x}_{k_\ell}}_{= -\alpha_{k_\ell} \nabla f(\mathbf{x}_{k_\ell})} \Big\} \geq \|\mathbf{x}_{k_\ell+1} - \mathbf{x}_{k_\ell}\|_2^2,$$

$$(\mathbf{y}_{k_\ell+1} - \mathbf{x}_{k_\ell+1})^T \Big\{ \underbrace{\mathbf{x}_{k_\ell} - 2\alpha_{k_\ell} \nabla f(\mathbf{x}_{k_\ell}) - \big(\mathbf{x}_{k_\ell} - \alpha_{k_\ell} \nabla f(\mathbf{x}_{k_\ell})\big)}_{= -\alpha_{k_\ell} \nabla f(\mathbf{x}_{k_\ell})} \Big\} \geq 0.$$

Zusammen führt dies auf

$$\begin{aligned}
(\mathbf{x}_{k_\ell} - \mathbf{y}_{k_\ell+1})^T \nabla f(\mathbf{x}_{k_\ell}) &\geq (\mathbf{x}_{k_\ell} - \mathbf{x}_{k_\ell+1})^T \nabla f(\mathbf{x}_{k_\ell}) \\
&\geq \frac{\|\mathbf{x}_{k_\ell+1} - \mathbf{x}_{k_\ell}\|_2^2}{\alpha_{k_\ell}} \\
&\geq \frac{\varepsilon}{2} \|\mathbf{y}_{k_\ell+1} - \mathbf{x}_{k_\ell}\|_2,
\end{aligned}$$

woraus sich speziell auch $\|\mathbf{y}_{k_\ell+1} - \mathbf{x}_{k_\ell}\|_2 \to 0$ für $\ell \to \infty$ ergibt. Die gleichmäßige Stetigkeit von ∇f impliziert nun

$$\begin{aligned}
\left| 1 - \frac{f(\mathbf{x}_{k_\ell}) - f(\mathbf{y}_{k_\ell+1})}{(\mathbf{x}_{k_\ell} - \mathbf{y}_{k_\ell+1})^T \nabla f(\mathbf{x}_{k_\ell})} \right| &= \frac{\mathcal{O}(\|\mathbf{y}_{k_\ell+1} - \mathbf{x}_{k_\ell}\|_2)}{(\mathbf{x}_{k_\ell} - \mathbf{y}_{k_\ell+1})^T \nabla f(\mathbf{x}_{k_\ell})} \\
&\leq \frac{2}{\varepsilon} \frac{\mathcal{O}(\|\mathbf{y}_{k_\ell+1} - \mathbf{x}_{k_\ell}\|_2)}{\|\mathbf{x}_{k_\ell} - \mathbf{y}_{k_\ell+1}\|_2} \\
&\overset{\ell \to \infty}{\longrightarrow} 0.
\end{aligned}$$

Dies steht jedoch im Widerspruch zu der aus (8.6) folgenden Abschätzung

$$\frac{f(\mathbf{x}_{k_\ell}) - f(\mathbf{y}_{k_\ell+1})}{(\mathbf{x}_{k_\ell} - \mathbf{y}_{k_\ell+1})^T \nabla f(\mathbf{x}_{k_\ell})} < \sigma < 1.$$

□

Bemerkung Da $\alpha_k \leq 1$ ist, folgt aus Satz 8.5 speziell $\|\mathbf{x}_{k+1} - \mathbf{x}_k\|_2 \to 0$ für $k \to \infty$, das heißt, die Iterierten aus Algorithmus 8.2 konvergieren gegen einen Punkt $\mathbf{x}^\star \in G$. △

Definition 8.6 Die zulässige Menge $G \subset \mathbb{R}^n$ sei konvex. Für jeden Punkt $\mathbf{x} \in G$ ist der **Tangentialkegel** $T_G(\mathbf{x})$ definiert als kleinster abgeschlossener Kegel, der die Menge

$$\{\mathbf{d} = \mathbf{y} - \mathbf{x} : \mathbf{y} \in G\}$$

enthält.

Bemerkung Der Tangentialkegel $T_G(\mathbf{x}^\star)$ ist die Menge aller Grenzrichtungen zulässiger Folgen für das Minimierungsproblem (7.1), skaliert mit einem beliebigen positiven Faktor. Wir erinnern an den Begriff des Linearisierungskegels $K(\mathbf{x}^\star)$ aus dem vorigen Kapitel. Er besteht aus allen Richtungen $\mathbf{d} \in \mathbb{R}^n$ mit $\mathbf{d}^T \nabla c_i(\mathbf{x}^\star) = 0$ für $i \in \mathcal{J}_G$ und $\mathbf{d}^T \nabla c_i(\mathbf{x}^\star) \geq 0$ für $i \in \mathcal{J}_U \cap \mathcal{J}_A(\mathbf{x}^\star)$. Nach Lemma 7.6 gilt $T_G(\mathbf{x}^\star) \subset K(\mathbf{x}^\star)$ und für den Fall, dass die LICQ-Bedingung erfüllt ist, sogar $T_G(\mathbf{x}^\star) = K(\mathbf{x}^\star)$. △

Lemma 8.7 Die zulässige Menge $G \subset \mathbb{R}^n$ sei konvex. Für jeden Punkt $\mathbf{x} \in G$ erfüllt die orthogonale Projektion $\mathbf{P}_{T_G(\mathbf{x})}\big(-\nabla f(\mathbf{x})\big)$ der Richtung des steilsten Abstiegs auf den Tangentialkegel $T_G(\mathbf{x})$ die folgenden Eigenschaften:

(*i.*) Es gilt

$$\nabla f(\mathbf{x})^T \mathbf{P}_{T_G(\mathbf{x})}\big(-\nabla f(\mathbf{x})\big) = -\big\|\mathbf{P}_{T_G(\mathbf{x})}\big(-\nabla f(\mathbf{x})\big)\big\|_2^2.$$

(*ii.*) Es ist

$$\min\{\nabla f(\mathbf{x})^T \mathbf{d} : \mathbf{d} \in T_G(\mathbf{x}),\ \|\mathbf{d}\|_2 \leq 1\} = -\|\mathbf{P}_{T_G(\mathbf{x})}\big(-\nabla f(\mathbf{x})\big)\|_2.$$

(*iii.*) Der Punkt $\mathbf{x}$ ist genau dann ein stationärer Punkt des Minimierungsproblems mit Nebenbedingungen (7.1), wenn $\mathbf{P}_{T_G(\mathbf{x})}\big(-\nabla f(\mathbf{x})\big) = \mathbf{0}$.

Beweis. (*i.*) Nach Definition der Orthogonalprojektion besitzt die Funktion

$$g(\lambda) := \frac{1}{2}\big\|\lambda \mathbf{P}_{T_G(\mathbf{x})}\big(-\nabla f(\mathbf{x})\big) + \nabla f(\mathbf{x})\big\|_2^2$$

ein Minimum bei $\lambda = 1$. Daher gilt

$$g'(1) := \big\|\mathbf{P}_{T_G(\mathbf{x})}\big(-\nabla f(\mathbf{x})\big)\big\|_2^2 + \nabla f(\mathbf{x})^T \mathbf{P}_{T_G(\mathbf{x})}\big(-\nabla f(\mathbf{x})\big) = 0.$$

(*ii.*) Wegen Aussage (*i.*) gilt

$$\big\|\mathbf{P}_{T_G(\mathbf{x})}\big(-\nabla f(\mathbf{x})\big) + \nabla f(\mathbf{x})\big\|_2^2 = \|\nabla f(\mathbf{x})\|_2^2 - \big\|\mathbf{P}_{T_G(\mathbf{x})}\big(-\nabla f(\mathbf{x})\big)\big\|_2^2.$$

Für alle $\mathbf{d} \in T_G(\mathbf{x})$ mit $\|\mathbf{d}\|_2 \leq \big\|\mathbf{P}_{T_G(\mathbf{x})}\big(-\nabla f(\mathbf{x})\big)\big\|_2$ gilt nach Definition der orthogonalen Projektion

$$\begin{aligned}\big\|\mathbf{P}_{T_G(\mathbf{x})}\big(-\nabla f(\mathbf{x})\big) + \nabla f(\mathbf{x})\big\|_2^2 &\leq \|\mathbf{d} + \nabla f(\mathbf{x})\|_2^2 \\ &\leq \big\|\mathbf{P}_{T_G(\mathbf{x})}\big(-\nabla f(\mathbf{x})\big)\big\|_2^2 + 2\nabla f(\mathbf{x})^T \mathbf{d} + \|\nabla f(\mathbf{x})\|_2^2.\end{aligned}$$

Zusammen ergibt dies

$$\nabla f(\mathbf{x})^T \mathbf{d} \geq -\big\|\mathbf{P}_{T_G(\mathbf{x})}\big(-\nabla f(\mathbf{x})\big)\big\|_2^2.$$

Das Behauptete erhält man, indem man $\widehat{\mathbf{d}} = \mathbf{d}/\big\|\mathbf{P}_{T_G(\mathbf{x})}\big(-\nabla f(\mathbf{x})\big)\big\|_2$ setzt.

(*iii.*) Definitionsgemäß ist $\mathbf{x} \in G$ genau dann ein stationärer Punkt, wenn $\nabla f(\mathbf{x})^T \mathbf{d} \geq 0$ ist für alle Grenzrichtungen zulässiger Folgen mit Grenzwert $\mathbf{x}$. Dies ist gleichbedeutend damit, dass $\nabla f(\mathbf{x})^T \mathbf{d} \geq 0$ für alle $\mathbf{d} \in T_G(\mathbf{x})$ ist. Aussage (*ii.*) impliziert, dass dies genau dann der Fall ist, wenn $\mathbf{P}_{T_G(\mathbf{x})}\big(-\nabla f(\mathbf{x})\big) = \mathbf{0}$ erfüllt ist. □

Bemerkung Aussage (*ii.*) des obigen Lemmas kann auch als

$$\min\{\nabla f(\mathbf{x}^\star)^T\mathbf{d} : \mathbf{d} \in T_G(\mathbf{x}^\star),\ \|\mathbf{d}\|_2 = 1\} = -\|\mathbf{P}_{T_G(\mathbf{x}^\star)}\big(-\nabla f(\mathbf{x}^\star)\big)\|_2. \tag{8.7}$$

geschrieben werden, da das Minimum für $\|\mathbf{d}\|_2 = 1$ angenommen wird. △

Satz 8.8 Die zulässige Menge $G \subset \mathbb{R}^n$ sei konvex und die Funktion $f : \mathbb{R}^n \to \mathbb{R}$ sei auf G stetig differenzierbar und nach unten beschränkt. Weiter sei ∇f auf G gleichmäßig stetig. Dann gilt für die Iterierten $\{\mathbf{x}_k\}_{k\in\mathbb{N}}$ des projizierten Gradientenverfahrens

$$\lim_{k\to\infty} \mathbf{P}_{T_G(\mathbf{x}_k)}\big(-\nabla f(\mathbf{x}_k)\big) = \mathbf{0}.$$

Beweis. Zu beliebigen $\varepsilon > 0$ gibt es nach Lemma 8.7 (*ii.*) zu jeder Iterierten $\mathbf{x}_k$ ein $\mathbf{d}_k \in T_G(\mathbf{x}_k)$ mit $\|\mathbf{d}_k\|_2 = 1$, so dass

$$\nabla f(\mathbf{x}_k)^T\mathbf{d}_k \le -\|\mathbf{P}_{T_G(\mathbf{x}_k)}\big(-\nabla f(\mathbf{x}_k)\big)\|_2 + \varepsilon \tag{8.8}$$

gilt. Da $\mathbf{d}_k$ Grenzrichtung einer zulässigen Folge ist, gibt es ein $\mathbf{y}_k \in G$ mit

$$\left\| \frac{\mathbf{y}_k - \mathbf{x}_k}{\|\mathbf{y}_k - \mathbf{x}_k\|_2} - \mathbf{d}_k \right\|_2 \le \varepsilon.$$

Aus Lemma 8.3 (*i.*) folgt

$$\begin{aligned}
&\big\{\mathbf{x}_{k+1} - \big(\mathbf{x}_k - \alpha_k\nabla f(\mathbf{x}_k)\big)\big\}^T(\mathbf{x}_{k+1} - \mathbf{y}_{k+1}) \\
&\quad = \big\{\mathbf{P}_G\big(\mathbf{x}_k - \alpha_k\nabla f(\mathbf{x}_k)\big) - \big(\mathbf{x}_k - \alpha_k\nabla f(\mathbf{x}_k)\big)\big\}^T\big\{\mathbf{P}_G\big(\mathbf{x}_k - \alpha_k\nabla f(\mathbf{x}_k)\big) - \mathbf{y}_{k+1}\big\} \\
&\quad \le 0,
\end{aligned}$$

was auf

$$\alpha_k\nabla f(\mathbf{x}_k)^T(\mathbf{x}_{k+1} - \mathbf{y}_{k+1}) \le \|\mathbf{x}_{k+1} - \mathbf{x}_k\|_2\|\mathbf{x}_{k+1} - \mathbf{y}_{k+1}\|_2,$$

beziehungsweise

$$-\frac{\nabla f(\mathbf{x}_k)^T(\mathbf{y}_{k+1} - \mathbf{x}_{k+1})}{\|\mathbf{x}_{k+1} - \mathbf{y}_{k+1}\|_2} \le \frac{\|\mathbf{x}_{k+1} - \mathbf{x}_k\|_2}{\alpha_k},$$

führt. Insgesamt erhalten wir deshalb

$$\begin{aligned}
-\nabla f(\mathbf{x}_k)^T\mathbf{d}_{k+1} &\le \|\nabla f(\mathbf{x}_k)\|_2\left\| \frac{\mathbf{y}_{k+1} - \mathbf{x}_{k+1}}{\|\mathbf{y}_{k+1} - \mathbf{x}_{k+1}\|_2} - \mathbf{d}_{k+1} \right\|_2 - \frac{\nabla f(\mathbf{x}_k)^T(\mathbf{y}_{k+1} - \mathbf{x}_{k+1})}{\|\mathbf{y}_{k+1} - \mathbf{x}_{k+1}\|_2} \\
&\le \varepsilon\|\nabla f(\mathbf{x}_k)\|_2 + \frac{\|\mathbf{x}_{k+1} - \mathbf{x}_k\|_2}{\alpha_k}.
\end{aligned}$$

Die Kombination mit (8.8) ergibt

$$\begin{aligned}
&\|\mathbf{P}_{T_G(\mathbf{x}_{k+1})}\big(-\nabla f(\mathbf{x}_{k+1})\big)\|_2 \\
&\quad \le -\nabla f(\mathbf{x}_{k+1})^T\mathbf{d}_{k+1} + \varepsilon \\
&\quad \le -\nabla f(\mathbf{x}_k)^T\mathbf{d}_{k+1} + \|\nabla f(\mathbf{x}_{k+1}) - \nabla f(\mathbf{x}_k)\|_2\underbrace{\|\mathbf{d}_{k+1}\|_2}_{=1} + \varepsilon \\
&\quad \le \varepsilon\|\nabla f(\mathbf{x}_k)\|_2 + \frac{\|\mathbf{x}_{k+1} - \mathbf{x}_k\|_2}{\alpha_k} + \|\nabla f(\mathbf{x}_{k+1}) - \nabla f(\mathbf{x}_k)\|_2 + \varepsilon.
\end{aligned}$$

Weil $\varepsilon > 0$ beliebig war, folgt hieraus schließlich

$$\lim_{k\to\infty} \big\|\mathbf{P}_{T_G(\mathbf{x}_{k+1})}\big(-\nabla f(\mathbf{x}_{k+1})\big)\big\|_2 \le \lim_{k\to\infty} \frac{\|\mathbf{x}_{k+1}-\mathbf{x}_k\|_2}{\alpha_k} + \lim_{k\to\infty} \|\nabla f(\mathbf{x}_{k+1}) - \nabla f(\mathbf{x}_k)\|_2 = 0$$

wobei Satz 8.5 und die gleichmäßige Stetigkeit von ∇f zur Anwendung kommt. □

In der Regel folgt aus der Stetigkeit von ∇f nicht, dass auch $\mathbf{P}_{T_G(\mathbf{x})}\big(-\nabla f(\mathbf{x})\big)$ stetig ist. Um sicherzustellen, dass die Iterierten des projizierten Gradientenverfahrens tatsächlich gegen einen stationären Punkt konvergieren, benötigen wir daher das folgende Resultat.

Satz 8.9 Die zulässige Menge $G \subset \mathbb{R}^n$ sei konvex und die Funktion $f : \mathbb{R}^n \to \mathbb{R}$ sei auf G stetig differenzierbar. Dann folgt für jede Folge $\{\mathbf{x}_k\}_{k\in\mathbb{N}} \subset G$ mit $\mathbf{x}_k \to \mathbf{x}^\star \in G$

$$\big\|\mathbf{P}_{T_G(\mathbf{x}^\star)}\big(-\nabla f(\mathbf{x}^\star)\big)\big\|_2 \le \liminf_{k\to\infty} \big\|\mathbf{P}_{T_G(\mathbf{x}_k)}\big(-\nabla f(\mathbf{x}_k)\big)\big\|_2.$$

Beweis. Aus Lemma 8.7 (*ii.*) folgt für jedes $\mathbf{y} \in G$

$$-\nabla f(\mathbf{x}_k)^T(\mathbf{y}-\mathbf{x}_k) \le \big\|\mathbf{P}_{T_G(\mathbf{x}_k)}\big(-\nabla f(\mathbf{x}_k)\big)\big\|_2 \|\mathbf{y}-\mathbf{x}_k\|_2,$$

woraus sich für $k \to \infty$ die Ungleichung

$$-\nabla f(\mathbf{x}^\star)^T(\mathbf{y}-\mathbf{x}^\star) \le \big\|\mathbf{P}_{T_G(\mathbf{x}_k)}\big(-\nabla f(\mathbf{x}_k)\big)\big\|_2 \|\mathbf{y}-\mathbf{x}^\star\|_2$$

ergibt. Jedes $\mathbf{d} \in T_G(\mathbf{x}^\star)$ mit $\|\mathbf{d}\|_2 = 1$ ist Grenzrichtung einer zulässigen Folge $\{\mathbf{y}_k\}_{k\in\mathbb{N}} \subset G$, das heißt, es gilt

$$\mathbf{d} = \lim_{k\to\infty} \frac{\mathbf{y}_k - \mathbf{x}^\star}{\|\mathbf{y}_k - \mathbf{x}^\star\|_2} \quad \text{und} \quad \lim_{k\to\infty} \mathbf{y}_k = \mathbf{x}^\star.$$

Somit erhalten wir

$$-\nabla f(\mathbf{x}^\star)^T\mathbf{d} \le \liminf_{k\to\infty} \big\|\mathbf{P}_{T_G(\mathbf{x}_k)}\big(-\nabla f(\mathbf{x}_k)\big)\big\|_2$$

und daraus wegen (8.7) die Behauptung:

$$\begin{aligned}\big\|\mathbf{P}_{T_G(\mathbf{x}^\star)}\big(-\nabla f(\mathbf{x}^\star)\big)\big\|_2 &= \max\{-\nabla f(\mathbf{x}^\star)^T\mathbf{d} : \mathbf{d} \in T_G(\mathbf{x}^\star),\ \|\mathbf{d}\|_2 = 1\} \\ &\le \liminf_{k\to\infty} \big\|\mathbf{P}_{T_G(\mathbf{x}_k)}\big(-\nabla f(\mathbf{x}_k)\big)\big\|_2.\end{aligned}$$

□

Bemerkung Die Kombination der Sätze 8.5, 8.8 und 8.9 liefert die folgende Aussage: Ist die zulässige Menge $G \subset \mathbb{R}^n$ konvex und ist die Funktion $f : \mathbb{R}^n \to \mathbb{R}$ auf G stetig differenzierbar mit gleichmäßig stetigem Gradienten und nach unten beschränkt, dann konvergieren die Iterierten $\{\mathbf{x}_k\}_{k\in\mathbb{N}}$ des projizierten Gradientenverfahrens 8.2 gegen ein $\mathbf{x}^\star \in G$ mit

$$\mathbf{P}_{T_G(\mathbf{x}^\star)}\big(-\nabla f(\mathbf{x}^\star)\big) = \mathbf{0}.$$

Gemäß Lemma 8.7 (*iii.*) bedeutet dies, dass $\mathbf{x}^\star$ ein stationärer Punkt ist. △

8.2 Affine Nebenbedingungen

Bei Anwendungsproblemen treten häufig affine Ungleichungsnebenbedingungen auf, weshalb wir diesen Spezialfall eingehender untersuchen wollen. Wir nehmen demnach an, dass das Minimierungsproblem unter Nebenbedingungen von der speziellen Form

$$\begin{aligned} &\min_{\mathbf{x}\in\mathbb{R}^n} f(\mathbf{x}) \text{ unter den Nebenbedingungen} \\ &\mathbf{a}_i^T\mathbf{x} \le b_i \text{ für alle } i = 1, 2, \dots, m \end{aligned} \tag{8.9}$$

mit $\mathbf{a}_i \in \mathbb{R}^n$ und $b_i \in \mathbb{R}$ für alle $i = 1, 2, \dots, m$ ist. Dies ist also ein Spezialfall von (7.1), für den $\mathcal{J}_G = \emptyset$, $\mathcal{J}_U = \{1, 2, \dots, m\}$ und $\mathbf{c}_i(\mathbf{x}) := b_i - \mathbf{a}_i^T\mathbf{x}$ gesetzt wird. Die Menge der zulässigen Punkte ist in diesem Fall gegeben durch

$$G = \{\mathbf{x} \in \mathbb{R}^n : \mathbf{a}_i^T\mathbf{x} \le b_i \text{ für alle } i = 1, 2, \dots, m\}.$$

Für $\mathbf{x} \in G$ betrachten wir die Menge der aktiven Indizes

$$\mathcal{J}_A(\mathbf{x}) := \big\{i \in \{1, 2, \dots, m\} : \mathbf{a}_i^T\mathbf{x} = b_i\big\}.$$

Die LICQ-Bedingung bedeutet hier gerade die lineare Unabhängigkeit der Vektoren $\mathbf{a}_i$ für $i \in \mathcal{J}_A(\mathbf{x})$.
Wir wenden uns der Charakterisierung von stationären Punkten zu, die für den Spezialfall affiner Nebenbedingungen besonders handlich ist.

Lemma 8.10 Für den Punkt $\mathbf{x}^\star \in G$ sei die Menge $\{\mathbf{a}_i : i \in \mathcal{J}_A(\mathbf{x}^\star)\}$ linear unabhängig. Dann ist $\mathbf{x}^\star$ genau dann ein stationärer Punkt für das Minimierungsproblem unter affinen Nebenbedingungen (8.9), wenn es für jedes $i \in \mathcal{J}_A(\mathbf{x}^\star)$ eine eindeutige Zahl $\lambda_i^\star \ge 0$ gibt, so dass

$$\nabla f(\mathbf{x}^\star) + \sum_{i\in\mathcal{J}_A(\mathbf{x}^\star)} \lambda_i^\star \mathbf{a}_i = \mathbf{0} \tag{8.10}$$

gilt.

Beweis. Nach Lemma 8.7 (*iii.*) ist $\mathbf{x}^\star \in G$ genau dann ein stationärer Punkt des Problems (8.9), wenn für die orthogonale Projektion auf den Tangentialkegel $\mathbf{P}_{T_G(\mathbf{x}^\star)}\big(-\nabla f(\mathbf{x}^\star)\big) = \mathbf{0}$ gilt. Dies ist nach Lemma 8.7 (*ii.*) genau dann der Fall, wenn

$$-\nabla f(\mathbf{x}^\star)^T\mathbf{d} \le 0 \quad \text{für alle } \mathbf{d} \in T_G(\mathbf{x}^\star).$$

gilt. Da die LICQ-Bedingung erfüllt ist, gilt nun nach Lemma 7.6

$$T_G(\mathbf{x}^\star) = K(\mathbf{x}^\star) = \{\mathbf{d} \in \mathbb{R}^n : \mathbf{d}^T\mathbf{a}_i \le 0 \text{ für alle } i \in \mathcal{J}_A(\mathbf{x}^\star)\}.$$

Schließlich ist nach Lemma 7.8 die Aussage

$$-\nabla f(\mathbf{x}^\star)^T\mathbf{d} \le 0 \quad \text{für alle } \mathbf{d} \in K(\mathbf{x}^\star)$$

äquivalent zur Existenz und Eindeutigkeit nichtnegativer Zahlen $\lambda_i^\star \in \mathbb{R}$, $i \in \mathcal{J}_A(\mathbf{x}^\star)$, mit

$$\nabla f(\mathbf{x}^\star) = -\sum_{i\in\mathcal{J}_A(\mathbf{x}^\star)} \lambda_i^\star \mathbf{a}_i.$$

□

Bemerkung Die Notwendigkeit der Bedingung (8.10) ergibt sich auch direkt aus den KKT-Bedingungen (vergleiche Satz 7.10). Lemma 8.10 sagt jedoch zusätzlich aus, dass diese Bedingung bei Vorliegen affiner Nebenbedingungen auch hinreichend für das Vorliegen eines stationären Punktes ist. △

Definition 8.11 Ein stationärer Punkt heißt **nicht entartet**, wenn die Menge $\{\mathbf{a}_i : i \in \mathcal{J}_A(\mathbf{x}^\star)\}$ linear unabhängig ist und für die Lagrange-Parameter in (8.10) $\lambda_i^\star > 0$ für alle $i \in \mathcal{J}_A(\mathbf{x}^\star)$ gilt.

Satz 8.12 Die Funktion $f : \mathbb{R}^n \to \mathbb{R}$ sei stetig differenzierbar auf der zulässigen Menge

$$G = \{\mathbf{x} \in \mathbb{R}^n : \mathbf{a}_i^T \mathbf{x} \leq b_i \text{ für alle } i = 1, 2, \ldots, m\}.$$

Die Folge $\{\mathbf{x}_k\}_{k\in\mathbb{N}} \in G$ konvergiere gegen den nicht entarteten stationären Punkt $\mathbf{x}^\star \in G$ und es gelte

$$\lim_{k\to\infty} \mathbf{P}_{T_G(\mathbf{x}_k)}\big(-\nabla f(\mathbf{x}_k)\big) = \mathbf{0}.$$

Dann gibt es ein $L \in \mathbb{N}$ mit $\mathcal{J}_A(\mathbf{x}_k) = \mathcal{J}_A(\mathbf{x}^\star)$ für alle $k \geq L$, das heißt, die Menge der aktiven Indizes ändert sich für genügend große k nicht mehr.

Beweis. Aus $\mathbf{x}_k \to \mathbf{x}^\star$ und $\mathbf{x}_k \in G$ folgt $\mathcal{J}_A(\mathbf{x}_k) = \mathcal{J}_A(\mathbf{x}^\star)$ für genügend große k. Wir nehmen nun das Gegenteil der Aussage des Satzes an, nämlich dass eine unendliche Teilfolge $\{\mathbf{x}_{k_\ell}\}_{\ell\in\mathbb{N}}$ und ein Index $p \in \mathcal{J}_A(\mathbf{x}^\star)$ existieren, so dass $p \notin \mathcal{J}_A(\mathbf{x}_{k_\ell})$ für alle $\ell \in \mathbb{N}$. Wir verwenden den orthogonalen Projektor $\mathbf{P}_H$ auf den Unterraum

$$H := \big\{\mathbf{x} \in \mathbb{R}^n : \mathbf{a}_i^T \mathbf{x} = 0 \text{ für alle } i \in \mathcal{J}_A(\mathbf{x}^\star) \setminus \{p\}\big\}.$$

Für $i \in \mathcal{J}_A(\mathbf{x}^\star) \setminus \{p\}$ gilt insbesondere $\mathbf{a}_i^T \mathbf{P}_H \mathbf{a}_i = 0$, woraus $\mathbf{P}_H \mathbf{a}_i = \mathbf{0}$ folgt.
Aufgrund der Eigenschaft der orthogonalen Projektion steht $\mathbf{a}_p - \mathbf{P}_H \mathbf{a}_p$ senkrecht auf dem Unterraum H, was wiederum bedeutet, dass $\mathbf{a}_p - \mathbf{P}_H \mathbf{a}_p$ im Raum $\operatorname{span}\big\{\mathbf{a}_i : i \in \mathcal{J}_A(\mathbf{x}^\star) \setminus \{p\}\big\}$ enthalten ist. Angenommen, es würde $\mathbf{P}_H \mathbf{a}_p = \mathbf{0}$ gelten, dann ließe sich $\mathbf{a}_p$ als Linearkombination der Vektoren $\mathbf{a}_i$, $i \in \mathcal{J}_A(\mathbf{x}^\star) \setminus \{p\}$, schreiben. Die Menge $\{\mathbf{a}_i : i \in \mathcal{J}_A(\mathbf{x}^\star)\}$ ist aber linear unabhängig, da der stationäre Punkt $\mathbf{x}^\star$ nicht entartet ist. Somit muss $\mathbf{P}_H \mathbf{a}_p \neq \mathbf{0}$ gelten.
Wegen $\mathbf{P}_H \mathbf{a}_p \in H$, das heißt, $\mathbf{a}_i^T \mathbf{P}_H \mathbf{a}_p = 0$ für $i \in \mathcal{J}_A(\mathbf{x}^\star) \setminus \{p\} \supset \mathcal{J}_A(\mathbf{x}_{k_\ell})$ schließen wir insbesondere $\mathbf{P}_H \mathbf{a}_p \in T_G(\mathbf{x}_{k_\ell})$ für alle $\ell \in \mathbb{N}$. Aus Lemma 8.7 (*ii.*) erhalten wir

$$\nabla f(\mathbf{x}_k)^T \mathbf{P}_H \mathbf{a}_p \geq -\big\|\mathbf{P}_{T_G(\mathbf{x}_k)}\big(-\nabla f(\mathbf{x}_k)\big)\big\|_2 \|\mathbf{P}_H \mathbf{a}_p\|_2.$$

Für $k \to \infty$ ergibt sich wegen $\mathbf{x}_k \to \mathbf{x}^\star$ damit

$$\nabla f(\mathbf{x}^\star)^T \mathbf{P}_H \mathbf{a}_p \geq 0. \tag{8.11}$$

Andererseits gilt, da der stationäre Punkt $\mathbf{x}^\star$ nicht entartet ist,

$$\nabla f(\mathbf{x}^\star) + \sum_{i\in\mathcal{J}_A(\mathbf{x}^\star)} \lambda_i^\star \mathbf{a}_i = \mathbf{0}$$

mit $\lambda_i^\star > 0$ für $i \in \mathcal{J}_A(\mathbf{x}^\star)$. Dies impliziert

$$\begin{aligned}
\nabla f(\mathbf{x}^\star)^T \mathbf{P}_H \mathbf{a}_p &= \big(\mathbf{P}_H \nabla f(\mathbf{x}^\star)\big)^T \mathbf{a}_p \\
&= -\Bigg(\sum_{i \in \mathcal{J}_A(\mathbf{x}^\star)} \lambda_i^\star \mathbf{P}_H \mathbf{a}_i\Bigg)^T \mathbf{a}_p \\
&= -\lambda_p^\star (\mathbf{P}_H \mathbf{a}_p)^T \mathbf{a}_p \\
&= -\lambda_p^\star \|\mathbf{P}_H \mathbf{a}_p\|_2^2 \\
&< 0
\end{aligned}$$

im Widerspruch zur Abschätzung (8.11). □

9. SQP-Verfahren

9.1 Quadratische Minimierungsprobleme mit affinen Nebenbedingungen

Das Prinzip des *SQP-Verfahrens* (kurz für *sequential quadratic programs*) ist die Rückführung allgemeiner Minimierungsprobleme auf eine Folge von Minimierungsproblemen mit quadratischer Zielfunktion und affinen Nebenbedingungen.

Ersetzen wir das allgemeine Minimierungsproblem (7.1) durch eine quadratische Zielfunktion und affine Nebenbedingungen, so erhalten wir folgende Aufgabenstellung:

$$\begin{aligned} \min_{\mathbf{x}\in\mathbb{R}^n} \frac{1}{2}\mathbf{x}^T\mathbf{H}\mathbf{x} + \mathbf{g}^T\mathbf{x} \text{ unter den Nebenbedingungen} \\ \mathbf{a}_i^T\mathbf{x} = b_i \text{ für alle } i \in \mathcal{J}_G, \\ \mathbf{a}_i^T\mathbf{x} \leq b_i \text{ für alle } i \in \mathcal{J}_U. \end{aligned} \tag{9.1}$$

Dabei wird die quadratische Zielfunktion durch die quadratische Matrix $\mathbf{H} \in \mathbb{R}^{n\times n}$ und den Vektor $\mathbf{g} \in \mathbb{R}^n$ beschrieben. Eine eventuell hinzukommende additive Konstante hat keine Konstante auf das Minimierungsproblem. Für $i \in \mathcal{J}_G \cup \mathcal{J}_U$ sind weiter $\mathbf{a}_i \in \mathbb{R}^n$ und $b_i \in \mathbb{R}$ vorgegeben, um die Nebenbedingungen festzulegen.

Um (9.1) zu lösen, wird zunächst iterativ die Menge der aktiven Indizes bestimmt. Hierzu werden im nächsten Abschnitt geeignete Techniken vorgestellt. Als Teilaufgabe resultieren quadratische Minimierungsprobleme mit affinen Gleichungsnebenbedingungen. Wir betrachten also zuerst im Detail die Lösung solcher Probleme:

$$\begin{aligned} \min_{\mathbf{x}\in\mathbb{R}^n} \frac{1}{2}\mathbf{x}^T\mathbf{H}\mathbf{x} + \mathbf{g}^T\mathbf{x} \text{ unter den Nebenbedingungen} \\ \mathbf{a}_i^T\mathbf{x} = b_i \text{ für alle } i = 1, 2, \ldots, m. \end{aligned} \tag{9.2}$$

Um (9.2) zu lösen, stellen wir die KKT-Bedingungen auf. Die Lagrange-Funktion lautet

$$L(\mathbf{x}, \boldsymbol{\lambda}) := \frac{1}{2}\mathbf{x}^T\mathbf{H}\mathbf{x} + \mathbf{g}^T\mathbf{x} - \sum_{i=1}^{m} \lambda_i(b_i - \mathbf{a}_i^T\mathbf{x}).$$

Zur Abkürzung schreiben wir

$$\mathbf{A} = \left[\begin{array}{c|c|c|c} \mathbf{a}_1 & \mathbf{a}_2 & \cdots & \mathbf{a}_m \end{array}\right] \in \mathbb{R}^{n\times m}, \quad \mathbf{b} = \begin{bmatrix} b_1 \\ b_2 \\ \vdots \\ b_m \end{bmatrix} \in \mathbb{R}^m, \quad \boldsymbol{\lambda} = \begin{bmatrix} \lambda_1 \\ \lambda_2 \\ \vdots \\ \lambda_m \end{bmatrix} \in \mathbb{R}^m$$

und erhalten

$$L(\mathbf{x}, \boldsymbol{\lambda}) := \frac{1}{2}\mathbf{x}^T\mathbf{H}\mathbf{x} + \mathbf{g}^T\mathbf{x} - \boldsymbol{\lambda}^T(\mathbf{b} - \mathbf{A}^T\mathbf{x}).$$

Als notwendige Bedingung ergibt sich gemäß Satz 7.10

$$\begin{aligned}\nabla_{\mathbf{x}} L(\mathbf{x}, \boldsymbol{\lambda}) &= \mathbf{H}\mathbf{x} + \mathbf{g} + \mathbf{A}\boldsymbol{\lambda} = \mathbf{0},\\ \nabla_{\mathbf{y}} L(\mathbf{x}, \boldsymbol{\lambda}) &= \mathbf{A}^T\mathbf{x} - \mathbf{b} = \mathbf{0}.\end{aligned}$$

Daher erfüllt die Lösung $(\mathbf{x}^\star, \boldsymbol{\lambda}^\star)$ der notwendigen Bedingung erster Ordnung das Sattelpunktproblem

$$\begin{bmatrix}\mathbf{H} & \mathbf{A}\\ \mathbf{A}^T & \mathbf{0}\end{bmatrix}\begin{bmatrix}\mathbf{x}^\star\\ \boldsymbol{\lambda}^\star\end{bmatrix} = \begin{bmatrix}-\mathbf{g}\\ \mathbf{b}\end{bmatrix}. \tag{9.3}$$

Wir wenden uns nun der Frage zu, wann dieses lineare Gleichungssystem eindeutig lösbar ist.

Wir stellen fest, dass $\mathbf{A} \in \mathbb{R}^{n\times m}$ vollen Rang besitzen muss. Wegen $m < n$ ist jedoch $\operatorname{kern}(\mathbf{A}^T) \neq \{\mathbf{0}\}$. Sei $\mathbf{p} \in \mathbb{R}^n$ ein solcher Vektor, für den $\mathbf{A}^T\mathbf{p} = \mathbf{0}$ gilt. Dann ist für beliebiges $\mathbf{q} \in \mathbb{R}^m$

$$\begin{aligned}\begin{bmatrix}\mathbf{p}^T & \mathbf{q}^T\end{bmatrix}\begin{bmatrix}\mathbf{H} & \mathbf{A}\\ \mathbf{A}^T & \mathbf{0}\end{bmatrix}\begin{bmatrix}\mathbf{p}\\ \mathbf{q}\end{bmatrix} &= \mathbf{p}^T\mathbf{H}\mathbf{p} + \mathbf{q}^T\mathbf{A}^T\mathbf{p} + \mathbf{p}^T\mathbf{A}\mathbf{q}\\ &= \mathbf{p}^T\mathbf{H}\mathbf{p} + \underbrace{\mathbf{q}^T\mathbf{A}^T\mathbf{p}}_{=\mathbf{0}} + \underbrace{(\mathbf{A}^T\mathbf{p})^T}_{=\mathbf{0}}\mathbf{q}\\ &= \mathbf{p}^T\mathbf{H}\mathbf{p}.\end{aligned} \tag{9.4}$$

Folglich müssen wir vermeiden, dass der Nullraum der Matrix $\mathbf{H}$ einen Vektor des Nullraums von der Matrix $\mathbf{A}^T$ enthält, was die Voraussetzungen des nachfolgenden Satzes motiviert.

Satz 9.1 Es sei $\mathbf{H} \in \mathbb{R}^{n\times n}$ und $\mathbf{A} \in \mathbb{R}^{n\times m}$, wobei $m < n$. Besitzt $\mathbf{A}$ vollen Rang m und ist $\mathbf{H}$ auf dem Kern von $\mathbf{A}^T$ positiv definit, das heißt, ist $\mathbf{p}^T\mathbf{H}\mathbf{p} > 0$ für alle $\mathbf{p} \in \operatorname{kern}(\mathbf{A}^T) \setminus \{\mathbf{0}\}$, so ist die KKT-Matrix

$$\mathbf{K} := \begin{bmatrix}\mathbf{H} & \mathbf{A}\\ \mathbf{A}^T & \mathbf{0}\end{bmatrix}$$

regulär und folglich (9.3) eindeutig lösbar.

Beweis. Sei $\mathbf{p} \in \mathbb{R}^n$ und $\mathbf{q} \in \mathbb{R}^m$ derart, dass

$$\mathbf{K}\begin{bmatrix}\mathbf{p}\\ \mathbf{q}\end{bmatrix} = \begin{bmatrix}\mathbf{H} & \mathbf{A}\\ \mathbf{A}^T & \mathbf{0}\end{bmatrix}\begin{bmatrix}\mathbf{p}\\ \mathbf{q}\end{bmatrix} = \mathbf{0}$$

gilt. Aus der zweiten Zeile folgt $\mathbf{A}^T\mathbf{p} = \mathbf{0}$ und daher ist

$$\mathbf{0} = \begin{bmatrix}\mathbf{p}^T & \mathbf{q}^T\end{bmatrix}\begin{bmatrix}\mathbf{H} & \mathbf{A}\\ \mathbf{A}^T & \mathbf{0}\end{bmatrix}\begin{bmatrix}\mathbf{p}\\ \mathbf{q}\end{bmatrix} \overset{(9.4)}{=} \mathbf{p}^T\mathbf{H}\mathbf{p}.$$

Dies kann aber nach Voraussetzung nur für $\mathbf{p} = \mathbf{0}$ erfüllt sein. Wegen $\mathbf{H}\mathbf{p} + \mathbf{A}\mathbf{q} = \mathbf{0}$ folgt schließlich $\mathbf{A}\mathbf{q} = \mathbf{0}$ und aufgrund des vollen Spaltenrangs von $\mathbf{A}$ ist auch $\mathbf{q} = \mathbf{0}$. Folglich ist $\mathbf{K}$ regulär und das Behauptete bewiesen. □

Bemerkung Man kann sogar zeigen, dass die Matrix $\mathbf{K}$ genau n positive und m negative Eigenwerte besitzt. Die Lösung indefiniter Systeme kann mit geeigneten Erweiterungen der Cholesky-Zerlegung oder aber bei sehr großen Systemen auch iterativ, etwa mit dem MINRES-Verfahren oder dem Bramble-Pasciak-CG, geschehen. △

Der nächste Satz besagt, dass die Lösung von (9.3) die Lösung von (9.2) liefert.

Satz 9.2 Es gelten die Voraussetzungen von Satz 9.1. Dann ist die Lösung $\mathbf{x}^\star \in \mathbb{R}^n$ von (9.3) die eindeutige Lösung von (9.2).

Beweis. Sei $\mathbf{x} \neq \mathbf{x}^\star$ derart, dass $\mathbf{A}^T\mathbf{x} = \mathbf{b}$ und setze $\mathbf{p} := \mathbf{x}^\star - \mathbf{x}$. Dann ist $\mathbf{A}^T\mathbf{p} = \mathbf{0}$ und $\mathbf{p} \neq \mathbf{0}$. Für $q(\mathbf{x}) := \frac{1}{2}\mathbf{x}^T\mathbf{H}\mathbf{x} + \mathbf{g}^T\mathbf{x}$ gilt dann offensichtlich mit $\mathbf{x} = \mathbf{x}^\star - \mathbf{p}$

$$q(\mathbf{x}) = \frac{1}{2}\mathbf{p}^T\mathbf{H}\mathbf{p} - \mathbf{p}^T\mathbf{H}\mathbf{x}^\star - \mathbf{g}^T\mathbf{p} + q(\mathbf{x}^\star). \tag{9.5}$$

Aus (9.3) folgt, dass $-\mathbf{H}\mathbf{x}^\star = \mathbf{A}\boldsymbol{\lambda} + \mathbf{g}$, so dass

$$-\mathbf{p}^T\mathbf{H}\mathbf{x}^\star = \mathbf{p}^T(\mathbf{A}\boldsymbol{\lambda} + \mathbf{g}) = \mathbf{g}^T\mathbf{p}$$

ist. Eingesetzt in (9.5) ergibt sich

$$q(\mathbf{x}) = \frac{1}{2}\mathbf{p}^T\mathbf{H}\mathbf{p} + q(\mathbf{x}^\star) > q(\mathbf{x}^\star),$$

dies bedeutet, $\mathbf{x}^\star$ ist das eindeutige Minimum von (9.2). □

9.2 Bestimmung aktiver Nebenbedingungen

Wir wenden uns nun der Lösung der Minimierungsaufgabe (9.1) zu, wobei wir annehmen, dass die Vektoren $\mathbf{a}_i$ linear unabhängig sind. Indem wir wie bisher für jeden Punkt $\mathbf{x} \in \mathbb{R}^n$ die Menge der aktiven Indizes bezeichnen als

$$\mathcal{J}_A(\mathbf{x}) = \{i \in \mathcal{J}_G \cup \mathcal{J}_U : \mathbf{a}_i^T\mathbf{x} = b_i\} = \mathcal{J}_G \cup \{i \in \mathcal{J}_U : \mathbf{a}_i^T\mathbf{x} = b_i\},$$

können wir (9.1) umschreiben gemäß

$$\begin{aligned} \min_{\mathbf{x}\in\mathbb{R}^n} \frac{1}{2}\mathbf{x}^T\mathbf{H}\mathbf{x} + \mathbf{g}^T\mathbf{x} \text{ unter den Nebenbedingungen} \\ \mathbf{a}_i^T\mathbf{x} = b_i \text{ für alle } i \in \mathcal{J}_A(\mathbf{x}), \\ \mathbf{a}_i^T\mathbf{x} \leq b_i \text{ für alle } i \in \mathcal{J}_U \setminus \mathcal{J}_A(\mathbf{x}). \end{aligned}$$

Für jede lokale Lösung $\mathbf{x}^\star \in \mathbb{R}^n$ des Minimierungsproblems (9.1) gelten somit nach Satz 7.10 die folgenden KKT-Bedingungen

$$\begin{aligned} \mathbf{H}\mathbf{x}^\star + \mathbf{g} + \sum_{i\in\mathcal{J}_A(\mathbf{x}^\star)} \lambda_i^\star \mathbf{a}_i = \mathbf{0}, \\ \mathbf{a}_i^T\mathbf{x}^\star = b_i \text{ für } i \in \mathcal{J}_A(\mathbf{x}^\star), \\ \mathbf{a}_i^T\mathbf{x}^\star \leq b_i \text{ für } i \in \mathcal{J}_U \setminus \mathcal{J}_A(\mathbf{x}^\star), \\ \lambda_i \geq 0 \text{ für } i \in \mathcal{J}_U \setminus \mathcal{J}_A(\mathbf{x}^\star). \end{aligned}$$

Die Grundidee des folgenden Vorgehens ist die Beobachtung, dass bei Kenntnis der Indizes $\mathcal{J}_A(\mathbf{x}^\star)$ der aktiven Nebenbendigungen am Lösungspunkt die Lösung von (9.1) gleichzeitig die Lösung des quadratischen Minimierungsproblem unter Gleichungsnebenbedingungen

$$\begin{gathered}\min_{\mathbf{x}\in\mathbb{R}^n} \frac{1}{2}\mathbf{x}^T\mathbf{H}\mathbf{x} + \mathbf{g}^T\mathbf{x} \text{ unter den Nebenbedingungen} \\ \mathbf{a}_i^T\mathbf{x} = b_i \quad \text{für alle } i \in \mathcal{J}_A(\mathbf{x}^\star)\end{gathered}$$

ist. Die aktiven Indizes werden über ein iteratives Vorgehen gefunden, ausgehend von einem Sattelpunkt $\mathbf{x}_0$, der zulässiger Punkt für das Problem (9.1) ist, und einer Indexmenge $\mathcal{J}_0$ mit $\mathcal{J}_G \subset \mathcal{J}_0 \subset \mathcal{J}_A(\mathbf{x}_0)$.

Ausgehend von einem zulässigem Punkt $\mathbf{x}_k$ und einer aktuellen Indexmenge $\mathcal{J}_k$ mit $\mathcal{J}_G \subset \mathcal{J}_k \subset \mathcal{J}_A(\mathbf{x}_k)$ besteht ein Schritt dieses iterativen Vorgehens aus den folgenden beiden Halbschritten:

1. Löse das quadratische Minimierungsproblem mit Gleichungsnebenbedingungen an der aktuellen Indexmenge:

$$\begin{gathered}\min_{\mathbf{d}\in\mathbb{R}^n} \frac{1}{2}(\mathbf{x}_k+\mathbf{d})^T\mathbf{H}(\mathbf{x}_k+\mathbf{d}) + \mathbf{g}^T(\mathbf{x}_k+\mathbf{d}) \text{ unter den Nebenbedingungen} \\ \mathbf{a}_i^T(\mathbf{x}_k+\mathbf{d}) = b_i \text{ für alle } i \in \mathcal{J}_k.\end{gathered} \tag{9.6}$$

2. Konstruiere aus dem aus dem ersten Halbschritt erhaltenen Punkt $\mathbf{x}_k + \mathbf{d}_k$ einen Punkt $\mathbf{x}_{k+1}$, der in der zulässigen Menge

$$G = \{\mathbf{x} \in \mathbb{R}^n : \mathbf{a}_i^T\mathbf{x} = b_i \text{ für alle } i \in \mathcal{J}_G \text{ und } \mathbf{a}_i^T\mathbf{x} \le b_i \text{ für alle } i \in \mathcal{J}_U\}$$

enthalten ist.

Die Lösung von (9.6) ist offenbar äquivalent zu

$$\begin{gathered}\min_{\mathbf{d}\in\mathbb{R}^n} \frac{1}{2}\mathbf{d}^T\mathbf{H}\mathbf{d} + (\mathbf{g}+\mathbf{H}\mathbf{x}_k)^T\mathbf{d} + \frac{1}{2}\mathbf{x}_k^T\mathbf{H}\mathbf{x}_k + \mathbf{g}^T\mathbf{x}_k \text{ unter den Nebenbedingungen} \\ \mathbf{a}_i^T\mathbf{d} = b_i - \mathbf{a}_i^T\mathbf{x}_k \text{ für alle } i \in \mathcal{J}_k.\end{gathered}$$

Mit $\mathbf{g}_k := \mathbf{g} + \mathbf{H}\mathbf{x}_k$ und unter der Beachtung von $\mathbf{a}_i^T\mathbf{x}_k = b_i$ sowie der Tatsache, dass Konstanten in der zu minimierenden Funktion keine Rolle spielen, lässt sich dies auch umschreiben in

$$\begin{gathered}\min_{\mathbf{d}\in\mathbb{R}^n} \frac{1}{2}\mathbf{d}^T\mathbf{H}\mathbf{d} + \mathbf{g}_k^T\mathbf{d} \text{ unter den Nebenbedingungen} \\ \mathbf{a}_i^T\mathbf{d} = 0 \text{ für alle } i \in \mathcal{J}_k.\end{gathered} \tag{9.7}$$

Dieses quadratische Minimierungsproblem unter Gleichungsnebenbedingungen hat die Form (9.2) und kann entsprechend gelöst werden.

Bei der Durchführung des zweiten Halbschritts wird wie folgt vorgegangen. Erfüllt $\mathbf{x}_k+\mathbf{d}_k$ auch die Nebenbedingungen

$$\mathbf{a}_i^T(\mathbf{x}_k+\mathbf{d}_k) \le b_i \text{ für alle } i \in \mathcal{J}_U \setminus \mathcal{J}_k,$$

dann wird $\mathbf{x}_{k+1} := \mathbf{x}_k + \mathbf{d}_k$ gesetzt. Falls nicht alle Nebenbedingungen erfüllt sind, bestimmen wir $\mathbf{x}_{k+1} := \mathbf{x}_k + \alpha_k\mathbf{d}_k$ mit $\alpha_k \in [0,1)$ größtmöglich, so dass noch alle Nebenbedingungen

$$\mathbf{a}_i^T(\mathbf{x}_k+\alpha_k\mathbf{d}_k) \le b_i \text{ für alle } i \in \mathcal{J}_U \setminus \mathcal{J}_k$$

erfüllt sind. Die Bestimmung von α_k geschieht durch eine Fallunterscheidung:

1. Für alle $i \in \mathcal{J}_U \setminus \mathcal{J}_k$ mit $\mathbf{a}_i^T \mathbf{d}_k \leq 0$ gilt

$$\mathbf{a}_i^T(\mathbf{x}_k + \alpha_k \mathbf{d}_k) \leq \mathbf{a}_i^T \mathbf{x}_k \leq b_i$$

   für alle $\alpha_k \geq 0$. Somit verursachen diese Nebenbedingungen keine Einschränkung an α_k.
2. Für alle $i \in \mathcal{J}_U \setminus \mathcal{J}_k$ mit $\mathbf{a}_i^T \mathbf{d}_k > 0$ haben wir jedoch die Einschränkung

$$\alpha_k \leq \frac{b_i - \mathbf{a}_i^T \mathbf{x}_k}{\mathbf{a}_i^T \mathbf{d}_k}.$$

Insgesamt erhalten wir daher

$$\alpha_k = \min \left\{ 1, \min_{i \notin \mathcal{J}_k, \mathbf{a}_i^T \mathbf{d}_k > 0} \frac{b_i - \mathbf{a}_i^T \mathbf{x}_k}{\mathbf{a}_i^T \mathbf{d}_k} \right\}.$$

Alle Nebenbedingungen, für die das Minimum angenommen wird, bezeichnen wir als *blockierende Nebenbedingungen*. Falls $\alpha_k < 1$, legen wir zu $\mathbf{x}_{k+1} := \mathbf{x}_k + \alpha_k \mathbf{d}_k$ eine neue Indexmenge $\mathcal{J}_{k+1}$ fest, indem wir eine oder mehrere blockierende Nebenbedingungen hinzufügen.

Diese Iteration wird solange durchgeführt, bis wir an einem Punkt $\widehat{\mathbf{x}}$ angelangt sind, der das quadratische Minimierungsproblem mit Gleichungsnebenbedingungen an der aktuellen Indexmenge $\widehat{\mathcal{J}}$ löst. Dies ist leicht daran zu erkennen, dass das zugehörige quadratische Minimierungsproblem (9.7) die Lösung $\widehat{\mathbf{d}} = \mathbf{0}$ besitzt.

Aufgrund der Minimierungseigenschaft gilt nach Satz 7.10 für das Problem (9.7) die erste KKT-Bedingung

$$\mathbf{H}\widehat{\mathbf{x}} + \mathbf{g} + \sum_{i \in \widehat{\mathcal{J}}} \widehat{\lambda}_i \mathbf{a}_i = \mathbf{0}$$

mit Lagrange-Parametern $\widehat{\lambda}_i$, über deren Vorzeichen aber keine Aussage möglich ist. Setzen wir $\widehat{\lambda}_i = 0$ für $i \in \mathcal{J}_U \setminus \widehat{\mathcal{J}}$, so haben wir auch die erste KKT-Bedingung für das Minimierungsproblem (9.1) erfüllt:

$$\mathbf{H}\widehat{\mathbf{x}} + \mathbf{g} + \sum_{i \in \mathcal{J}_G \cup \mathcal{J}_U} \widehat{\lambda}_i \mathbf{a}_i = \mathbf{0}.$$

Die zweite und dritte KKT-Bedingungen lauten

$$\mathbf{a}_i^T \widehat{\mathbf{x}} = b_i \text{ für } i \in \mathcal{J}_G,$$
$$\mathbf{a}_i^T \widehat{\mathbf{x}} \leq b_i \text{ für } i \in \mathcal{J}_U.$$

Sie sind nach Konstruktion ebenfalls erfüllt. Ferner ist $\widehat{\lambda}_i = 0$ für alle $i \in \mathcal{J}_U \setminus \widehat{\mathcal{J}}$, weshalb auch die fünfte KKT-Bedingung

$$\widehat{\lambda}_i (b_i - \mathbf{a}_i^T \widehat{\mathbf{x}}) = 0 \text{ für } i \in \mathcal{J}_G \cup \mathcal{J}_U$$

gilt.

Für die vierte KKT-Bedingung ist noch das Vorzeichen der Lagrange-Parameter $\widehat{\lambda}_i$ für $i \in \widehat{\mathcal{J}} \cap \mathcal{J}_U$ zu bestimmen. Sind diese sämtlich nichtnegativ, dann ist auch die vierte KKT-Bedingung erfüllt und es gelten alle für das Minimierungsproblem (9.1) notwendigen

Bedingungen aus Satz 7.10. Für den Fall, dass $\mathbf{H}$ positiv definit ist, folgt insbesondere sogar, dass $\widehat{\mathbf{x}}$ tatsächlich eine lokale Lösung des Minimierungsproblems (9.1) ist.

Tritt dagegen ein negativer Lagrange-Parameter $\widehat{\lambda}_j$ mit $j \in \widehat{\mathcal{J}} \cap \mathcal{J}_U$ auf, so ist die vierte KKT-Bedingung verletzt. Daher kann $\widehat{\mathbf{x}}$ keine Lösung des Minimierungsproblems (9.1) sein. In diesem Fall lässt sich der Wert des quadratischen Funktionals verkleinern, indem man die zu j gehörige Nebenbedingung aus der Indexmenge $\widehat{\mathcal{J}}$ streicht. Dies folgt aus dem Beweis von Lemma 7.8. Man kann zeigen, dass sich auf diese Weise alle Nebenbedingungen mit negativen Lagrange-Parametern entfernen lassen und am Ende ein Punkt $\widehat{\mathbf{x}}$ resultiert, der alle KKT-Bedingungen erfüllt.

Beispiel 9.3 Wir illustrieren das beschriebene Vorgehen an folgendem konkreten Minimierungsproblem

$$\min_{(x_1,x_2)\in\mathbb{R}^2} (x_1-1)^2 + \left(x_2 - \frac{5}{2}\right)^2 \text{ unter den Nebenbedingungen}$$
$$\begin{aligned} x_1 - 2x_2 + 2 &\geq 0 \\ -x_1 - 2x_2 + 6 &\geq 0 \\ -x_1 + 2x_2 + 2 &\geq 0 \\ x_1 &\geq 0 \\ x_2 &\geq 0. \end{aligned}$$

In der Notation (9.1) haben wir also

$$\mathbf{H} = \begin{bmatrix} 2 & 0 \\ 0 & 2 \end{bmatrix}, \quad \mathbf{g} = \begin{bmatrix} -2 \\ -5 \end{bmatrix}, \quad \mathbf{x} = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}$$

$$\mathbf{a}_1 = \begin{bmatrix} -1 \\ 2 \end{bmatrix}, \quad \mathbf{a}_2 = \begin{bmatrix} 1 \\ 2 \end{bmatrix}, \quad \mathbf{a}_3 = \begin{bmatrix} 1 \\ -2 \end{bmatrix}, \quad \mathbf{a}_4 = \begin{bmatrix} -1 \\ 0 \end{bmatrix}, \quad \mathbf{a}_5 = \begin{bmatrix} 0 \\ -1 \end{bmatrix}$$

$$b_1 = 2, \quad b_2 = 6, \quad b_3 = 2, \quad b_4 = 0, \quad b_5 = 0.$$

Wir starten das Verfahren mit

$$\mathbf{x}_0 = \begin{bmatrix} 2 \\ 0 \end{bmatrix} \quad \text{und} \quad \mathcal{J}_0 = \{3, 5\}.$$

Das Minimierungsproblem (9.6) liefert für $k = 0$ als Lösung $\mathbf{x}_1 = \mathbf{x}_0$, da nur dieser Punkt in der zulässigen Menge liegt und die beiden Gleichungsnebenbedingungen erfüllt. Die zugehörigen Lagrange-Parameter erhalten wir aus dem linearen Gleichungssystem

$$\mathbf{0} = \mathbf{H}\mathbf{x}_1 + \mathbf{g} + \widehat{\lambda}_3\mathbf{a}_3 + \widehat{\lambda}_5\mathbf{a}_5 = \begin{bmatrix} 2 + \widehat{\lambda}_3 \\ -5 - 2\widehat{\lambda}_3 - \widehat{\lambda}_5 \end{bmatrix},$$

dessen Lösung $\widehat{\lambda}_3 = -2$ und $\widehat{\lambda}_5 = -1$ ist. Wir streichen die Nebenbedingungen zum kleineren Lagrange-Parameter und erhalten $\mathcal{J}_1 = \{5\}$.

Das Minimierungsproblem (9.6) für $k = 1$ liefert als Lösung

$$\mathbf{x}_2 = \begin{bmatrix} 1 \\ 0 \end{bmatrix}$$

und der Lagrange-Parameter $\widehat{\lambda}_5$ errechnet sich aus

$$\mathbf{0} = \mathbf{H}\mathbf{x}_2 + \mathbf{g} + \widehat{\lambda}_5 \mathbf{a}_5 = \begin{bmatrix} 0 \\ -5 - \widehat{\lambda}_5 \end{bmatrix},$$

das heißt, $\widehat{\lambda}_5 = -5$. Wir streichen folglich auch diese Nebenbedingung und landen bei $\mathcal{J}_2 = \emptyset$.
Nun lösen wir das Minimierungsproblem (9.7) für $k = 2$, das heißt, ein Minimierungsproblem ohne Nebenbedingungen, was auf

$$\mathbf{d}_2 = \begin{bmatrix} 0 \\ 5/2 \end{bmatrix} \quad \text{und somit} \quad \mathbf{x}_3 = \mathbf{x}_2 + \frac{3}{5}\mathbf{d}_2 = \begin{bmatrix} 1 \\ 3/2 \end{bmatrix}$$

führt. Wir addieren die blockierende Nebenbedingung zu unserer leeren Indexmenge $\mathcal{J}_2$ hinzu und erhalten $\mathcal{J}_3 = \{1\}$.
Die Lösung des Minimierungsproblems (9.7) für $k = 3$ ergibt

$$\mathbf{d}_3 = \begin{bmatrix} 2/5 \\ 1/5 \end{bmatrix} \quad \text{und somit} \quad \mathbf{x}_4 = \mathbf{x}_3 + \mathbf{d}_3 = \begin{bmatrix} 7/5 \\ 17/10 \end{bmatrix},$$

wobei keine weitere blockierende Nebenbedingung auftritt. Der zugehörige Lagrange-Parameter folgt aus

$$\mathbf{0} = \mathbf{H}\mathbf{x}_4 + \mathbf{g} + \widehat{\lambda}_1 \mathbf{a}_1 = \begin{bmatrix} 4/5 - \widehat{\lambda}_1 \\ -8/5 + 2\widehat{\lambda}_1 \end{bmatrix},$$

also $\widehat{\lambda}_1 = 4/5$. Folglich erfüllt $\mathbf{x}_4$ alle KKT-Bedingungen des Minimierungsproblems (9.1) und ist die gesuchte Lösung. △

Index