Universität Basel

Iterative Verfahren der Numerik

Skript zur Vorlesung
im
HS 2025

Dozent: Marc Schmidlin

Skriptauthor: Helmut Harbrecht

Version vom 15. Oktober 2025

Informationen ...

zu den Übungen

Die Übungen bestehen aus wöchentlichen Übungsblättern mit Theorieaufgaben, die in den Übungsstunden besprochen werden, und drei Programmierblättern (in MATLAB), welche abgenommen werden.

Für die Übungsblätter mit den Theorieaufgaben gibt es keine verpflichtende Abgabe. Die Theorieaufgaben können zur Korrektur freiwillig physisch an der Spiegelgasse 1 in dem entsprechenden Fach abgegeben werden; jeweils montags bis **12:00**. Die Rückgabe der korrigierten Abgaben erfolgt dann physisch in der Übungsstunde, weswegen Sie jeweils den **Tutor** der besuchten Übungsstunde auf Ihren Abgaben vermerken sollen.

Die Programmierblätter müssen verpflichtend erfolgreich bearbeitet, auf ADAM abgegeben und vorgezeigt werden. Die Abnahmen geschehen dabei in den drei Wochen: **06.10–10.10**, **03.11–07.11** und **01.12–05.12**.

zu den Leistungsüberprüfungen

Die Vorlesung *Iterative Verfahren der Numerik* ist eine Hauptvorlesung und wird somit mit einem Examen in der Form einer mündlichen Prüfung geprüft. Diese mündlichen Prüfungen werden im Zeitrahmen **09.–13.02.2026** stattfinden.

Die Übungen *Iterative Verfahren der Numerik* sind Übungen mit lehrveranstaltungsbegleitender Leistungsüberprüfung. Diese Leistungsüberprüfung besteht aus den zwei Elementen:

- Dem erfolgreichen Bearbeiten und Vorzeigen der Programmierblättern (s.o.).
- Dem Bestehen einer schriftlichen Klausur am **10.12.2025** um **14:15–16:00**.

Vorwort

Zur Mitschrift vom HS 2023

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 speziell auch das Vorlesungsskript *Numerik nichtlinearer Optimierung* von Gerhard Starke (Uni Hannover) als fruchtbare Quelle.

Helmut Harbrecht

Zur Mitschrift vom HS 2025

Diese Variante der ursprünglichen Mitschrift vom HS 2023 wurde hauptsächlich neu gesetzt, um ein A5-formatiges Skript aus einem A4-formatigen zu erstellen. Dabei wurden auch die Grafiken überarbeitet und weitere, kleine Änderungen vorgenommen, sowie gefundene Typos eliminiert. Daher sind beide Mitschriften inhaltlich fast deckungsgleich.

Marc Schmidlin

Literatur zur Vorlesung:

- M. Hanke-Bourgeois: *Grundlagen der Numerischen Mathematik und des Wissenschaftlichen Rechnens*, Teubner-Verlag
- 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 Eigenwerte

Erinnerung: Ist $\mathbf{A} \in \mathbb{K}^{n\times n}$, dann ist $p(\lambda) = \det(\mathbf{A}-\lambda\mathbf{I})$ ein (komplexwertiges) Polynom über $\mathbb{C}$ vom Grad n. Jede der n Nullstellen von p ist ein Eigenwert von $\mathbf{A}$, das heißt, zu einer solchen Nullstelle λ gibt es einen Eigenvektor $\mathbf{0} \neq \mathbf{x} \in \mathbb{C}^n$ mit $\mathbf{A}\mathbf{x} = \lambda\mathbf{x}$; umgekehrt ist auch jeder Eigenwert eine Nullstelle von p. Die Menge aller Eigenwerte nennt man das *Spektrum* $\sigma(\mathbf{A})$ von $\mathbf{A}$. Aus $\lambda \in \sigma(\mathbf{A})$ folgt $\overline{\lambda} \in \sigma(\mathbf{A}^\star)$.

Eigenwerte sind selbst bei reellen Matrizen im allgemeinen nicht reell. Ist aber $\mathbf{A} \in \mathbb{R}^{n\times n}$ und $\lambda \in \sigma(\mathbf{A})$, so ist auch $\overline{\lambda} \in \sigma(\mathbf{A})$, denn aus $\mathbf{A}\mathbf{x} = \lambda\mathbf{x}$ folgt

$$\mathbf{A}\overline{\mathbf{x}} = \overline{\mathbf{A}\mathbf{x}} = \overline{\lambda\mathbf{x}} = \overline{\lambda}\overline{\mathbf{x}}.$$

1.1 Eigenwerteinschließungen

Die Eigenwertgleichung $\mathbf{A}\mathbf{x} = \lambda\mathbf{x}$ ist nichtlinear bezüglich der gemeinsamen Unbekannten $(\lambda, \mathbf{x})$. Daher sind die meisten numerischen Verfahren zur Berechnung von $\sigma(\mathbf{A})$ iterativ und manchmal auch nur lokal konvergent. Aus diesem Grund ist die folgende Sammlung relativ einfacher Ergebnisse über die Lage der Eigenwerte einer Matrix von Bedeutung.

Satz 1.1 (Gerschgorin) *Sei $\mathbf{A} = [a_{i,j}] \in \mathbb{K}^{n\times n}$ und λ ein beliebiger Eigenwert von $\mathbf{A}$. Dann gilt*

$$\lambda \in \bigcup_{i=1}^{n} K_i = \bigcup_{i=1}^{n} \left\{ z \in \mathbb{C} : |z - a_{i,i}| \leq \sum_{\substack{j=1 \\ j\neq i}}^{n} |a_{i,j}| \right\}. \tag{1.1}$$

Beweis. Sei $\mathbf{A}\mathbf{x} = \lambda\mathbf{x}$ mit $\mathbf{x} \neq \mathbf{0}$. Dann existiert ein x_i mit $|x_j| \leq |x_i|$ für $j \neq i$. Folglich ist

$$\lambda x_i = [\mathbf{A}\mathbf{x}]_i = \sum_{j=1}^{n} a_{i,j} x_j$$

und weiter

$$|\lambda - a_{i,i}| = \left| \sum_{\substack{j=1 \\ j\neq i}}^{n} a_{i,j} \frac{x_j}{x_i} \right| \leq \sum_{\substack{j=1 \\ j\neq i}}^{n} |a_{i,j}| \underbrace{\left| \frac{x_j}{x_i} \right|}_{\leq 1} \leq \sum_{\substack{j=1 \\ j\neq i}}^{n} |a_{i,j}|.$$

Also ist $\lambda \in K_i \subset \bigcup_{j=1}^{n} K_j$. ♠

Wegen $\overline{\lambda} \in \sigma(\mathbf{A}^\star)$ gilt entsprechend

$$\overline{\lambda} \in \bigcup_{i=1}^{n} \left\{ z \in \mathbb{C} : |z - \overline{a_{i,i}}| \leq \sum_{\substack{j=1 \\ j\neq i}}^{n} |a_{j,i}| \right\}$$

beziehungsweise

$$\lambda \in \bigcup_{i=1}^{n} \widetilde{K}_i := \bigcup_{i=1}^{n} \left\{ z \in \mathbb{C} : |z - a_{i,i}| \leq \sum_{\substack{j=1 \\ j\neq i}}^{n} |a_{j,i}| \right\}. \tag{1.2}$$

Dies ist der Satz von Gerschgorin angewendet auf $\mathbf{A}^\star$.

Ist $\mathbf{A}$ eine beliebige $(n \times n)$-Matrix, dann ist $(\mathbf{A} + \mathbf{A}^\star)/2$ hermitesch und $(\mathbf{A} - \mathbf{A}^\star)/2$ *schiefhermitesch*, dies bedeutet

$$\left(\frac{1}{2}(\mathbf{A} - \mathbf{A}^\star) \right)^\star = -\frac{1}{2}(\mathbf{A} - \mathbf{A}^\star).$$

Offensichtlich gilt

$$\mathbf{A} = \frac{1}{2}(\mathbf{A} + \mathbf{A}^\star) + \frac{1}{2}(\mathbf{A} - \mathbf{A}^\star).$$

Definition 1.2 Unter dem **Wertebereich** einer Matrix $\mathbf{A} \in \mathbb{K}^{n\times n}$ versteht man die Menge aller **Rayleigh-Quotienten** $\mathbf{x}^\star\mathbf{A}\mathbf{x}/(\mathbf{x}^\star\mathbf{x})$ mit $\mathbf{x} \in \mathbb{C}^n \setminus \{\mathbf{0}\}$:

$$\mathcal{W}(\mathbf{A}) = \left\{ \frac{\mathbf{x}^\star\mathbf{A}\mathbf{x}}{\mathbf{x}^\star\mathbf{x}} \,:\, \mathbf{x} \in \mathbb{C}^n \setminus \{\mathbf{0}\} \right\} \subset \mathbb{C}.$$

Lemma 1.3

1. *$\mathcal{W}(\mathbf{A})$ ist zusammenhägend.*
2. *Ist $\mathbf{A}$ hermitesch, dann ist $\mathcal{W}(\mathbf{A})$ das reelle Intervall $[\lambda_{\min}, \lambda_{\max}]$.*
3. *Ist $\mathbf{A}$ schiefhermitesch, dann ist $\mathcal{W}(\mathbf{A})$ ein rein imaginäres Intervall, nämlich die konvexe Hülle aller Eigenwerte.*

Beweis. 1. Sei $\xi_0 \neq \xi_1 \in \mathcal{W}(\mathbf{A})$ mit

$$\xi_0 = \frac{\mathbf{x}_0^\star\mathbf{A}\mathbf{x}_0}{\mathbf{x}_0^\star\mathbf{x}_0}, \qquad \xi_1 = \frac{\mathbf{x}_1^\star\mathbf{A}\mathbf{x}_1}{\mathbf{x}_1^\star\mathbf{x}_1}, \qquad \mathbf{x}_0, \mathbf{x}_1 \in \mathbb{C}^n \setminus \{\mathbf{0}\}.$$

Offensichtlich ist $\mathbf{x}_0 \neq \lambda\mathbf{x}_1$, da sonst $\xi_0 = \xi_1$. Für $t \in [0,1]$ ist

$$\xi_t := \frac{\mathbf{x}_t^\star\mathbf{A}\mathbf{x}_t}{\mathbf{x}_t^\star\mathbf{x}_t} \in \mathcal{W}(\mathbf{A}) \quad \text{mit} \quad \mathbf{x}_t := \mathbf{x}_0 + t(\mathbf{x}_1 - \mathbf{x}_0) \in [\mathbf{x}_0, \mathbf{x}_1] \not\ni \mathbf{0}$$

eine stetige Kurve, die ξ_0 mit ξ_1 verbindet.

2. Ist die Matrix $\mathbf{A}$ hermitesch, dann sind alle Eigenwerte reell und es gilt

$$\min_{\mathbf{x}\in\mathbb{C}^n} \mathbf{x}^\star\mathbf{A}\mathbf{x} = \lambda_{\min}\|\mathbf{x}\|_2^2, \quad \max_{\mathbf{x}\in\mathbb{C}^n} \mathbf{x}^\star\mathbf{A}\mathbf{x} = \lambda_{\max}\|\mathbf{x}\|_2^2.$$

Hieraus ergibt sich die Behauptung.

3. Wegen $\mathbf{A}^\star = -\mathbf{A}$ ist $i\mathbf{A}$ hermitesch:

$$(i\mathbf{A})^\star = \bar{i}\mathbf{A}^\star = -i\mathbf{A}^\star = i\mathbf{A}.$$

Da $\mathcal{W}(i\mathbf{A}) = i\mathcal{W}(\mathbf{A})$ und $\sigma(i\mathbf{A}) = i\sigma(\mathbf{A})$ ist, folgt die Behauptung aus der zweiten Aussage. ♠

Klar: Es gilt immer $\sigma(\mathbf{A}) \subset \mathcal{W}(\mathbf{A})$.

Satz 1.4 (Bendixson) *Sei* $\mathbf{A} \in \mathbb{K}^{n\times n}$ *beliebig. Dann liegt das Spektrum von* $\mathbf{A}$ *in dem Rechteck*

$$\sigma(\mathbf{A}) \subset R := \mathcal{W}\left(\frac{1}{2}(\mathbf{A}+\mathbf{A}^\star)\right) + \mathcal{W}\left(\frac{1}{2}(\mathbf{A}-\mathbf{A}^\star)\right).$$

Beweis. Wir zeigen

$$\mathcal{W}(\mathbf{A}) \subset \mathcal{W}\left(\frac{1}{2}(\mathbf{A}+\mathbf{A}^\star)\right) + \mathcal{W}\left(\frac{1}{2}(\mathbf{A}-\mathbf{A}^\star)\right).$$

Sei $\mathbf{x} \in \mathbb{C}^n \setminus \{\mathbf{0}\}$, dann folgt

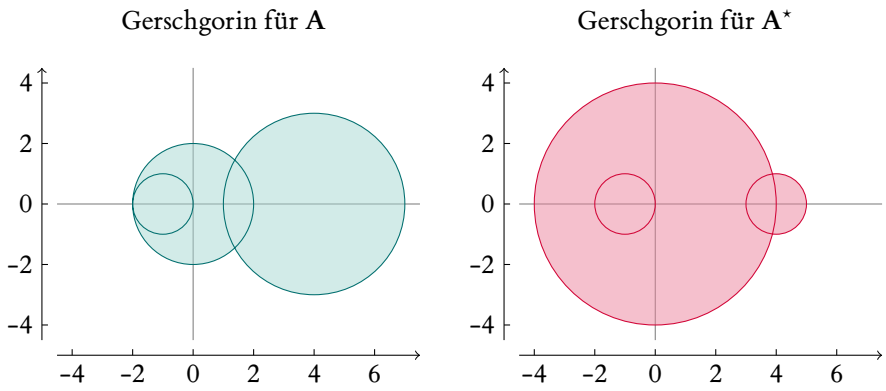
$$\begin{aligned}\frac{\mathbf{x}^\star\mathbf{A}\mathbf{x}}{\mathbf{x}^\star\mathbf{x}} &= \frac{\mathbf{x}^\star\left[\frac{1}{2}(\mathbf{A}+\mathbf{A}^\star)+\frac{1}{2}(\mathbf{A}-\mathbf{A}^\star)\right]\mathbf{x}}{\mathbf{x}^\star\mathbf{x}}\\ &= \frac{1}{2}\frac{\mathbf{x}^\star(\mathbf{A}+\mathbf{A}^\star)\mathbf{x}}{\mathbf{x}^\star\mathbf{x}} + \frac{1}{2}\frac{\mathbf{x}^\star(\mathbf{A}-\mathbf{A}^\star)\mathbf{x}}{\mathbf{x}^\star\mathbf{x}}\\ &\in \mathcal{W}\left(\frac{1}{2}(\mathbf{A}+\mathbf{A}^\star)\right) + \mathcal{W}\left(\frac{1}{2}(\mathbf{A}-\mathbf{A}^\star)\right).\end{aligned}$$

♠

Beispiel 1.5 Für die Matrix

$$\mathbf{A} = \begin{bmatrix} 4 & 0 & 3\\ 0 & -1 & 1\\ 1 & 1 & 0\end{bmatrix}$$

ergeben sich nach (1.1) beziehungsweise (1.2) die folgenden Einschließungen:



Der symmetrische und schiefsymmetrische Anteil von $\mathbf{A}$ ist

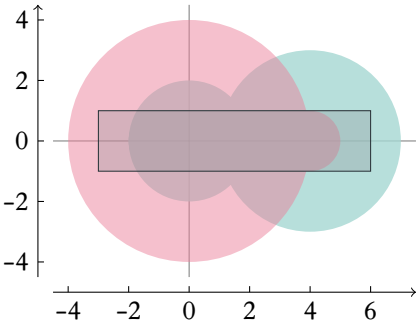
$$\mathbf{H} = \frac{1}{2}(\mathbf{A} + \mathbf{A}^\star) = \begin{bmatrix} 4 & 0 & 2 \\ 0 & -1 & 1 \\ 2 & 1 & 0 \end{bmatrix}, \qquad \mathbf{S} = \frac{1}{2}(\mathbf{A} - \mathbf{A}^\star) = \begin{bmatrix} 0 & 0 & 1 \\ 0 & 0 & 0 \\ -1 & 0 & 0 \end{bmatrix}.$$

Zur Anwendung des Satzes von Bendixson schließen wir die Spektren von $\mathbf{H}$ und $\mathbf{S}$ wieder mit dem Satz von Gerschgorin ein, was auf das Rechteck

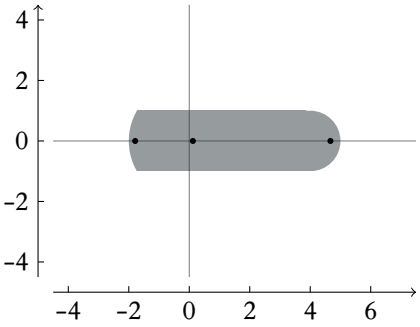
$$R = [-3, 6] + [-i, i]$$

führt. Das Spektrum von $\mathbf{A}$ muss im Schnitt aller drei Einschlussmengen liegen:

Einschlussmengen mit Bendixson



Schnitt & Spektrum



Tatsächlich ist das Spektrum $\sigma(\mathbf{A}) = \{-1.7878, 0.1198, 4.6679\}$. ♣

1.2 Kondition des Eigenwertproblems

Betrachte

$$\mathbf{A} = \begin{bmatrix} 0 & & & & & -a_0 \\ 1 & 0 & & & & -a_1 \\ & 1 & 0 & & & -a_2 \\ & & \ddots & \ddots & & \vdots \\ & & & 1 & 0 & -a_{n-2} \\ & & & & 1 & -a_{n-1} \end{bmatrix}.$$

Entwickelt man die Determinante nach der letzten Spalte, so gilt

$$(-1)^n \det(\mathbf{A} - \lambda \mathbf{I}) = \lambda^n + a_{n-1}\lambda^{n-1} + \ldots + a_1\lambda + a_0 =: p(\lambda).$$

Definition 1.6 Die Matrix $\mathbf{A}$ heißt **Frobenius-Begleitmatrix** von $p(\lambda)$. Die Eigenwerte von $\mathbf{A}$ sind die Nullstellen von p.

Das spezielle Polynom $p_0(\lambda) = (\lambda - a)^n$ hat die n-fache Nullstelle $\widehat{\lambda} = a$, während $p_\varepsilon(\lambda) = (\lambda - a)^n + \varepsilon$ für $\varepsilon > 0$ die Nullstellen

$$\lambda_k = a - \varepsilon^{1/n} e^{i2\pi k/n}, \quad k = 0, 1, \ldots, n-1,$$

besitzt. Die zugehörigen Frobenius-Begleitmatrizen unterscheiden sich nur um ε in der ∞, 1, 2 und der Frobeniusnorm-Norm, denn es gilt

$$\Delta\mathbf{A} = \mathbf{A}_0 - \mathbf{A}_\varepsilon = \begin{bmatrix} 0 & \cdots & 0 & \varepsilon \\ 0 & \cdots & 0 & 0 \\ \vdots & & \vdots & \vdots \\ 0 & \cdots & 0 & 0 \end{bmatrix}.$$

Allerdings haben die Eigenwerte der Begleitmatrizen den Abstand

$$|\Delta\lambda| = |\widehat{\lambda} - \lambda_k| = \varepsilon^{1/n}. \tag{1.3}$$

Für $a \neq 0$ folgt daher

$$\frac{|\Delta\lambda|}{|\lambda|} = \frac{\varepsilon^{1/n}}{|a|} = \underbrace{\frac{\|\mathbf{A}\|}{|a|}\frac{\varepsilon^{1/n}}{\varepsilon}}_{\to \infty \text{ für } \varepsilon \to 0} \frac{\|\Delta\mathbf{A}\|}{\|\mathbf{A}\|},$$

dies bedeutet, die Kondition des Eigenwertproblems kann ohne Zusatzvoraussetzungen an die Matrix $\mathbf{A}$ beliebig groß werden.

Man kann jedoch zeigen, dass die Eigenwerte stetig von den Einträgen der Matrix abhängen. Der gefundene Exponent $1/n$ in (1.3) ist schlimmstmöglich.

Definition 1.7 Eine Matrix $\mathbf{A}$ heißt **diagonalisierbar**, falls es eine Basis aus Eigenvektoren gibt. Sind $(\lambda_1, \mathbf{v}_1), (\lambda_2, \mathbf{v}_2), \ldots, (\lambda_n, \mathbf{v}_n)$ die Eigenpaare, dann gilt

$$\mathbf{A} = \mathbf{V}\mathbf{D}\mathbf{V}^{-1}, \qquad \mathbf{V} = [\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_n], \qquad \mathbf{D} = \operatorname{diag}(\lambda_1, \lambda_2, \ldots, \lambda_n).$$

Beachte: In der Regel ist eine Matrix *nicht* diagonalisierbar. Dann treten Hauptvektoren und Jordan-Kästchen auf.

Satz 1.8 (Bauer und Fike) *Sei* $\mathbf{A} = \mathbf{V}\mathbf{D}\mathbf{V}^{-1}$ *diagonalisierbar und* λ *ein Eigenwert von* $\mathbf{A} + \mathbf{E}$. *Dann existiert ein Eigenwert* λ_i *von* $\mathbf{A}$ *mit*

$$|\lambda - \lambda_i| \leq \operatorname{cond}(\mathbf{V})\|\mathbf{E}\|.$$

Hiebei bezeichne $\|\cdot\|$ *entweder die* 1, 2 *oder* ∞*-Norm und* $\operatorname{cond}(\mathbf{V})$ *die entsprechende Konditionszahl von* $\mathbf{V}$.

Beweis. Falls $\lambda \in \sigma(\mathbf{A})$ ist die Behauptung trivial. Andernfalls existiert $(\lambda\mathbf{I} - \mathbf{A})^{-1}$, und wir wählen einen Eigenvektor $\mathbf{v}$ von $\mathbf{A} + \mathbf{E}$ zum Eigenwert λ. Wir erhalten

$$\mathbf{E}\mathbf{v} = (\mathbf{A} + \mathbf{E} - \mathbf{A})\mathbf{v} = (\lambda\mathbf{I} - \mathbf{A})\mathbf{v},$$

und daher

$$(\lambda\mathbf{I} - \mathbf{A})^{-1}\mathbf{E}\mathbf{v} = \mathbf{v}.$$

Folglich ist

$$\begin{aligned} 1 &\leq \left\|(\lambda\mathbf{I} - \mathbf{A})^{-1}\mathbf{E}\right\| = \left\|\mathbf{V}(\lambda\mathbf{I} - \mathbf{D})^{-1}\mathbf{V}^{-1}\mathbf{E}\right\| \\ &\leq \|\mathbf{V}\|\left\|(\lambda\mathbf{I} - \mathbf{D})^{-1}\right\|\left\|\mathbf{V}^{-1}\right\|\|\mathbf{E}\| = \|\mathbf{E}\| \operatorname{cond}(\mathbf{V}) \max_{i=1}^{n}\{|\lambda - \lambda_i|^{-1}\}. \end{aligned}$$ ♠

Definition 1.9 Eine Matrix $\mathbf{A}$ heißt **normal**, falls gilt $\mathbf{A}\mathbf{A}^\star = \mathbf{A}^\star\mathbf{A}$. Insbesondere sind normale Matrizen diagonalisierbar mit $\mathbf{V}^{-1} = \mathbf{V}^\star$.

Bemerkung Hermitesche Matrizen sind normal. ♦

Korollar 1.10 *Sei* $\mathbf{E}$ *eine beliebige Matrix. Ist* $\mathbf{A}$ *normal und* λ *ein Eigenwert von* $\mathbf{A} + \mathbf{E}$, *dann existiert ein* $\lambda_i \in \sigma(\mathbf{A})$ *mit*

$$|\lambda - \lambda_i| \leq \|\mathbf{E}\|_2.$$

Beweis. Falls $\mathbf{A}$ normal ist, ist $\mathbf{V}$ unitär und daher $\operatorname{cond}_2(\mathbf{V}) = 1$. ♠

Die "normweise Konditionszahl" des einzelnen Eigenwerts ist also über $\operatorname{cond}(\mathbf{V})$ bestimmt worden, wobei $\mathbf{V}$ die Eigenvektormatrix bezeichnet. Eine lokalere Aussage lässt sich durch Differenzieren bestimmen:

Lemma 1.11 *Sei $\mathbf{A} \in \mathbb{K}^{n\times n}$ und λ_i ein einfacher Eigenwert zum Rechtseigenvektor $\mathbf{v}_i$, das heißt, $\mathbf{A}\mathbf{v}_i = \lambda_i \mathbf{v}_i$, und Linkseigenvektor $\mathbf{u}_i$, das heißt, $\mathbf{u}_i^\star \mathbf{A} = \mathbf{u}_i^\star \lambda_i$. Dann besitzt die Matrix $\mathbf{A} + \varepsilon\mathbf{C}$ für genügend kleines $|\varepsilon| > 0$ einen einfachen Eigenwert $\lambda(\varepsilon)$ und es gilt*

$$\lambda(\varepsilon) = \lambda_i + \varepsilon \frac{\mathbf{u}_i^\star \mathbf{C}\mathbf{v}_i}{\mathbf{u}_i^\star \mathbf{v}_i} + \mathcal{O}(\varepsilon^2), \quad \varepsilon \to 0. \tag{1.4}$$

Beweis. Da der Eigenwert λ_i einfach ist, folgt aus der Jordanschen Normalform

$$\mathbf{A} = \mathbf{T} \begin{bmatrix} \lambda_i & \mathbf{0} \\ \mathbf{0} & \mathbf{B} \end{bmatrix} \mathbf{T}^{-1} \quad \text{mit} \quad \mathbf{v}_i = \alpha \mathbf{T}\mathbf{e}_1, \quad \mathbf{u}_i = \beta \mathbf{T}^{-\star}\mathbf{e}_1, \quad \alpha, \beta \neq 0, \tag{1.5}$$

dies bedeutet,

$$\mathbf{u}_i^\star \mathbf{v}_i = \alpha\overline{\beta}\mathbf{e}_1^\star \mathbf{T}^{-1}\mathbf{T}\mathbf{e}_1 = \alpha\overline{\beta} \neq 0. \tag{1.6}$$

Wir betrachten nun die analytische Funktion

$$\mathbf{F} : \mathbb{C}^{n+1} \times \mathbb{C} \to \mathbb{C}^{n+1}, \quad \mathbf{F}(\mathbf{v}, \lambda, \varepsilon) = \begin{bmatrix} (\mathbf{A} + \varepsilon\mathbf{C} - \lambda\mathbf{I})\mathbf{v} \\ \mathbf{v}_i^\star \mathbf{v} - 1 \end{bmatrix},$$

für die offensichtlich gilt

$$\mathbf{F}(\mathbf{v}_i, \lambda_i, 0) = \mathbf{0}, \qquad \mathbf{M} := \frac{\partial \mathbf{F}}{\partial(\mathbf{v}, \lambda)}(\mathbf{v}_i, \lambda_i, 0) = \begin{bmatrix} \mathbf{A} - \lambda_i \mathbf{I} & -\mathbf{v}_i \\ \mathbf{v}_i^\star & 0 \end{bmatrix}.$$

Wir zeigen zunächst, dass die Matrix $\mathbf{M} \in \mathbb{C}^{(n+1)\times(n+1)}$ invertierbar ist. Dies folgt, wenn das Gleichungssystem

$$\mathbf{M} \begin{bmatrix} \mathbf{v} \\ \lambda \end{bmatrix} = \begin{bmatrix} \mathbf{A}\mathbf{v} - \lambda_i \mathbf{v} - \lambda \mathbf{v}_i \\ \mathbf{v}_i^\star \mathbf{v} \end{bmatrix} = \mathbf{0}$$

nur die Lösung $\mathbf{v} = \mathbf{0}$ und $\lambda = 0$ besitzt. Multiplizieren wir die erste Gleichung von links mit $\mathbf{u}_i^\star$, dann folgt

$$\underbrace{\mathbf{u}_i^\star \mathbf{A}\mathbf{v}}_{=\lambda_i \mathbf{u}_i^\star \mathbf{v}} - \lambda_i \mathbf{u}_i^\star \mathbf{v} - \lambda \mathbf{u}_i^\star \mathbf{v}_i = -\lambda \mathbf{u}_i^\star \mathbf{v}_i = 0,$$

was wegen (1.6) $\lambda = 0$ impliziert und weiter $\mathbf{A}\mathbf{v} = \lambda_i \mathbf{v}$. Folglich ist $\mathbf{v}$ ein Vielfaches von $\mathbf{v}_i$ und wegen der zweiten Gleichung gilt $\mathbf{v} = \mathbf{0}$.

Da $\mathbf{M}$ invertierbar ist, kann nach dem Satz über implizite Funktionen $\mathbf{F} = \mathbf{0}$ in einer Umgebung von $(\mathbf{v}_i, \lambda_i, 0)$ eindeutig aufgelöst werden. Für hinreichend kleines ε gibt es demnach eindeutige analyische Funktionen $\mathbf{v}(\varepsilon)$ und $\lambda(\varepsilon)$ mit

$$\mathbf{v}(0) = \mathbf{v}_i, \quad \lambda(0) = \lambda_i, \quad \text{und} \quad \big(\mathbf{A} + \varepsilon\mathbf{C} - \lambda(\varepsilon)\mathbf{I}\big)\mathbf{v}(\varepsilon) = \mathbf{0}, \quad \mathbf{v}_i^\star \mathbf{v}(\varepsilon) = 1.$$

Offensichtlich ist $\mathbf{v}(\varepsilon)$ nicht Null und somit ein Eigenvektor von $\mathbf{A} + \varepsilon\mathbf{C}$. Differenziert man $\big(\mathbf{A} + \varepsilon\mathbf{C} - \lambda(\varepsilon)\mathbf{I}\big)\mathbf{v}(\varepsilon) = \mathbf{0}$ nach ε und setzt $\varepsilon = 0$, dann ergibt sich

$$\big(\mathbf{C} - \lambda'(0)\mathbf{I}\big)\mathbf{v}_i + (\mathbf{A} - \lambda_i\mathbf{I})\mathbf{v}'(0) = \mathbf{0}.$$

Multiplikation von links mit $\mathbf{u}_i^\star$ liefert

$$\mathbf{u}_i^\star \mathbf{C}\mathbf{v}_i - \lambda'(0)\mathbf{u}_i^\star \mathbf{v}_i = 0$$

und daraus $\lambda'(0)$ wie behauptet. ♠

Korollar 1.12 *Sei $\mathbf{A} \in \mathbb{K}^{n\times n}$ und λ_i ein einfacher Eigenwert zum Rechtseigenvektor $\mathbf{v}_i$ und Linkseigenvektor $\mathbf{u}_i$. Dann ist die Kondition der Berechnung dieses Eigenwerts*

$$\kappa_{abs}(\lambda_i) = \frac{1}{\big|\cos\big(\sphericalangle(\mathbf{u}_i, \mathbf{v}_i)\big)\big|}, \qquad \kappa_{rel}(\lambda_i) = \frac{\|\mathbf{A}\|_2}{\big|\lambda_i \cos\big(\sphericalangle(\mathbf{u}_i, \mathbf{v}_i)\big)\big|}.$$

Beweis. Aus (1.4) ergibt sich für die absolute Kondition

$$\kappa_{\text{abs}}(\lambda_i) = \sup_{\substack{\mathbf{C}\in\mathbb{K}^{n\times n}\\ \|\mathbf{C}\|_2=1}} \frac{\mathbf{u}_i^\star \mathbf{C}\mathbf{v}_i}{\mathbf{u}_i^\star \mathbf{v}_i} = \frac{\|\mathbf{u}_i\|_2\|\mathbf{v}_i\|_2}{\|\mathbf{u}_i\|_2\|\mathbf{v}_i\|_2\big|\cos\big(\sphericalangle(\mathbf{u}_i, \mathbf{v}_i)\big)\big|} = \frac{1}{\big|\cos\big(\sphericalangle(\mathbf{u}_i, \mathbf{v}_i)\big)\big|},$$

woraus für die relative Kondition folgt

$$\kappa_{\text{rel}}(\lambda_i) = \frac{\|\mathbf{A}\|_2}{\big|\lambda_i \cos\big(\sphericalangle(\mathbf{u}_i, \mathbf{v}_i)\big)\big|}.$$ ♠

Bemerkung Die Kondition wird demnach groß, wenn $\mathbf{u}_i^\star \mathbf{v}_i \approx 0$. Bei normalen Matrizen fallen Rechts- und Linkseigenvektor zusammen, weshalb sich hier $\kappa_{\text{abs}}(\lambda_i) = 1$ ergibt. Offensichtlich ist dann auch die relative Kondition durch

$$\kappa_{\text{rel}}(\lambda_i) = \frac{\lambda_{|\max|}}{|\lambda_i|}.$$

gegben und damit fast 1 für betragsmässig grosse Eigenwerte; kann aber für betragsmässig kleine Eigenwerte gross werden. ♦

1.3 Potenzmethode

Als konstruktives Verfahren zur näherungsweisen Bestimmung einzelner Eigenwerte und Eigenvektoren betrachten wir die *Potenzmethode* oder *von Mises-Verfahren*. Um die Ideen des Verfahrens so klar wie möglich herauszustellen, beschränken wir uns grundsätzlich im folgenden auf reelle, diagonalisierbare Matrizen $\mathbf{A} \neq \mathbf{0}$ mit n betragsmäßig verschiedenen Eigenwerten

$$|\lambda_1| > |\lambda_2| > \ldots > |\lambda_n| \geq 0 \qquad \text{(beachte: alle } \lambda_i \in \mathbb{R}\text{)}.$$

Alle Ergebnisse können mit entsprechendem technischen Aufwand auf den allgemeinen Fall übertragen werden. Sind $\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_n$ mit $\|\mathbf{v}_i\| = 1$ die zugehörigen Eigenvektoren von $\mathbf{A}$, dann gibt es für jeden Vektor $\mathbf{x} \in \mathbb{R}^n$ eine Entwicklung

$$\mathbf{x} = \sum_{i=1}^{n} \xi_i \mathbf{v}_i, \tag{1.7}$$

und folglich ist

$$\mathbf{A}^k \mathbf{x} = \sum_{i=1}^{n} \lambda_i^k \xi_i \mathbf{v}_i. \tag{1.8}$$

Die Potenzmethode beruht nun auf der asymptotischen Identität

$$\mathbf{A}^k \mathbf{x} \approx \lambda_1^k \xi_1 \mathbf{v}_1.$$

Algorithmus 1.13 (von Mises-Potenzmethode)
input: Matrix $\mathbf{A} \in \mathbb{K}^{n \times n}$ und Startvektor $\mathbf{x} \in \mathbb{K}^n$
output: Folge von Iterierten $\{\mathbf{z}_k\}_{k \geq 0}$

① Initialisierung: setze $\mathbf{z}_0 := \mathbf{x}$ und $k := 0$

② berechne

$$\widetilde{\mathbf{z}}_{k+1} := \mathbf{A}\mathbf{z}_k, \qquad \mathbf{z}_{k+1} := \frac{\widetilde{\mathbf{z}}_{k+1}}{\|\widetilde{\mathbf{z}}_{k+1}\|} \tag{1.9}$$

③ erhöhe $k := k + 1$ und gehe nach ②

Die Potenzmethode besitzt die folgenden Eigenschaften:

Satz 1.14 *Ist $\xi_1 \neq 0$ in (1.7), dann gilt mit $q := |\lambda_2/\lambda_1| < 1$:*

1. *Es ist*

$$\|\widetilde{\mathbf{z}}_k\| = |\lambda_1| + \mathcal{O}(q^k), \quad k \to \infty.$$

2. *Ist $\lambda_1 > 0$, dann gilt*

$$\|\mathbf{z}_k - \operatorname{sign}(\xi_1)\mathbf{v}_1\| = \mathcal{O}(q^k), \quad k \to \infty.$$

3. *Ist $\lambda_1 < 0$, dann gilt*

$$\|(-1)^{k+1}\mathbf{z}_k - \operatorname{sign}(\xi_1)\mathbf{v}_1\| = \mathcal{O}(q^k), \quad k \to \infty.$$

Beweis. Aus (1.8) folgt

$$\mathbf{A}^k\mathbf{x} = \lambda_1^k\xi_1\Bigg(\mathbf{v}_1 + \underbrace{\sum_{i=2}^{n}\left[\frac{\lambda_i}{\lambda_1}\right]^k\frac{\xi_i}{\xi_1}\mathbf{v}_i}_{=:\mathbf{w}_k}\Bigg) \tag{1.10}$$

mit

$$\|\mathbf{w}_k\| \le q^k\underbrace{\sum_{i=2}^{n}\left|\frac{\xi_i}{\xi_1}\right|}_{=:C} = Cq^k.$$

Damit ist

$$\mathbf{z}_{k+1} = \frac{\mathbf{A}^k\mathbf{x}}{\|\mathbf{A}^k\mathbf{x}\|} = \operatorname{sign}(\lambda_1^k\xi_1)\,\frac{\mathbf{v}_1 + \mathbf{w}_k}{\|\mathbf{v}_1 + \mathbf{w}_k\|}. \tag{1.11}$$

Wegen

$$1 - Cq^k \le \|\mathbf{v}_1\| - \|\mathbf{w}_k\| \le \|\mathbf{v}_1 + \mathbf{w}_k\| \le \|\mathbf{v}_1\| + \|\mathbf{w}_k\| \le 1 + Cq^k$$

folgt also

$$\mathbf{z}_k = \operatorname{sign}(\lambda_1^{k-1}\xi_1)\,\mathbf{v}_1 + \mathcal{O}(q^k), \quad k \to \infty,$$

und

$$\widetilde{\mathbf{z}}_{k+1} = \mathbf{A}\mathbf{z}_k = \lambda_1\operatorname{sign}(\lambda_1^{k-1}\xi_1)\,\mathbf{v}_1 + \mathcal{O}(q^k), \quad k \to \infty.$$

Hiervon lassen sich die Eigenschaften 1–3 leicht ablesen. ♠

Bemerkungen

1. Die Normierung $\widetilde{\mathbf{z}}_k \mapsto \mathbf{z}_k$ in (1.9) ist notwendig, um overflow/underflow zu vermeiden. Die Wahl der Norm ist dabei unerheblich.

2. Aus Eigenschaft 1 ergibt sich $|\lambda_1|$, aus dem Vorzeichenverhalten von $\mathbf{z}_k$ das Vorzeichen von λ_1: alternieren die Vorzeichen von $\mathbf{z}_k$, dann folgt $\lambda_1 < 0$, ansonsten ist $\lambda_1 > 0$.

3. Die Voraussetzung $\xi_1 \neq 0$ kann natürlich nicht a priori überprüft werden. Wegen Rundungsfehlereinflüssen wird jedoch in der Regel eine Komponente von $\mathbf{z}_k$ längs $\mathbf{v}_1$ im Verlauf der Iteration eingeschleppt. ♦

Varianten: Die Potenzmethode kann in dieser Form nur verwendet werden, um λ_1 zu bestimmen. Zur Berechnung anderer Eigenwerte von $\mathbf{A}$ kann man jedoch $\mathbf{A}$ zunächst geeignet transformieren:

1. Gilt $\lambda_n \neq 0$, so kann man $\mathbf{A}^{-1}$ statt $\mathbf{A}$ in (1.9) verwenden. Dies ist dann die *inverse Iteration*. Da $\mathbf{A}^{-1}$ die Eigenwerte

$$|\lambda_n^{-1}| > |\lambda_{n-1}^{-1}| > \ldots > |\lambda_1^{-1}|$$

   mit den gleichen Eigenvektoren besitzt, approximiert die inverse Iteration $|\lambda_n^{-1}|$ und den entsprechenden Eigenvektor $\mathbf{v}_n$.

2. Bei der *gebrochenen Iteration* von Wielandt verwendet man $(\mathbf{A} - \lambda\mathbf{I})^{-1}$ statt $\mathbf{A}$ in (1.9), $\lambda \notin \sigma(\mathbf{A})$. Die Matrix $(\mathbf{A} - \lambda\mathbf{I})^{-1}$ besitzt die Eigenwerte $\{(\lambda_j - \lambda)^{-1}\}_{j=1}^n$ und die gebrochene Iteration approximiert den Eigenvektor zu $\lambda_i \in \sigma(\mathbf{A})$, der am nächsten zu λ liegt.

Bemerkungen

1. Bei beiden Varianten muss ein lineares Gleichungssystem pro Schritt gelöst werden!

2. Ist λ_i der nächste Eigenwert zu λ, dann konvergiert die gebrochene Iteration umso schneller, je näher λ ist, da der Konvergenzfaktor

$$q = \max_{j \neq i} \frac{|\lambda_j - \lambda|^{-1}}{|\lambda_i - \lambda|^{-1}} = \max_{j \neq i} \frac{|\lambda_i - \lambda|}{|\lambda_j - \lambda|}$$

   dann am kleinsten ist. ♦

Ist $\mathbf{A} = \mathbf{A}^\mathsf{T}$ reell symmetrisch (oder aber auch $\mathbf{A} = \mathbf{A}^\star$ komplex hermitesch) und $\|\cdot\| = \|\cdot\|_2$, dann kann man die Näherung $\|\widetilde{\mathbf{z}}_k\| \approx |\lambda_1|$ verbessern, indem man statt dessen die Näherung

$$\lambda_1 \approx \frac{\mathbf{z}_k^\star \mathbf{A}\mathbf{z}_k}{\underbrace{\mathbf{z}_k^\star \mathbf{z}_k}_{=1}} = \mathbf{z}_k^\star \mathbf{A}\mathbf{z}_k = \mathbf{z}_k^\star \widetilde{\mathbf{z}}_{k+1} \tag{1.12}$$

verwendet.

Satz 1.15 *Ist $\mathbf{A} = \mathbf{A}^\star$, dann gilt*

$$\left|\lambda_1 - \mathbf{z}_k^\star \widetilde{\mathbf{z}}_{k+1}\right| = \mathcal{O}(q^{2k}), \quad k \to \infty.$$

Beweis. Der Beweis verwendet, dass $\mathbf{w}_k$ aus (1.10) senkrecht auf $\mathbf{v}_1$ steht. Mit

$$\gamma_k := \operatorname{sign}(\lambda_1^{k-1}\xi_1) \underbrace{\|\mathbf{v}_1 + \mathbf{w}_{k-1}\|_2^{-1}}_{\geq 1}$$

folgt aus (1.11) nämlich

$$(\lambda_1 \mathbf{I} - \mathbf{A})\mathbf{z}_k = \gamma_k(\lambda_1 \mathbf{I} - \mathbf{A})\mathbf{w}_{k-1} \perp \mathbf{v}_1.$$

In Anbetracht von $\mathbf{z}_k^\star = \gamma_k \mathbf{v}_1^\star + \gamma_k \mathbf{w}_{k-1}^\star$ und $|\gamma_k| \leq 1$ bedeutet dies

$$|\mathbf{z}_k^\star \underbrace{(\lambda_1 \mathbf{I} - \mathbf{A})\mathbf{z}_k}_{\perp \mathbf{v}_1}| = \gamma_k^2 |\mathbf{w}_{k-1}^\star (\lambda_1 \mathbf{I} - \mathbf{A})\mathbf{w}_{k-1}| \leq \underbrace{\|\lambda_1 \mathbf{I} - \mathbf{A}\|_2}_{\leq 2\|\mathbf{A}\|} \|\mathbf{w}_{k-1}\|_2^2.$$

Wegen

$$\lambda_1 - \mathbf{z}_k^\star \widetilde{\mathbf{z}}_{k+1} = \lambda_1 \mathbf{z}_k^\star \mathbf{z}_k - \mathbf{z}_k^\star \widetilde{\mathbf{z}}_{k+1} = \mathbf{z}_k^\star(\lambda_1 \mathbf{z}_k - \widetilde{\mathbf{z}}_{k+1}) = \mathbf{z}_k^\star(\lambda_1 \mathbf{I} - \mathbf{A})\mathbf{z}_k$$

ergibt sich daher

$$|\lambda_1 - \mathbf{z}_k^\star \widetilde{\mathbf{z}}_{k+1}| \leq 2\|\mathbf{A}\|_2 \|\mathbf{w}_{k-1}\|_2^2 = \mathcal{O}(q^{2k}), \quad k \to \infty. \qquad \spadesuit$$

Entsprechend kann man für "innere" Eigenwerte verfahren:

Algorithmus 1.16 (Rayleigh-Quotient-Iteration)
input: Matrix $\mathbf{A} \in \mathbb{K}^{n\times n}$ und Näherungseigenpaar $(\mu_0, \mathbf{z}_0)$ mit $\|\mathbf{z}_0\|_2 = 1$ (bestimmt etwa durch Potenzmethode)
output: Folge von Iterierten $\{(\mu_k, \mathbf{z}_k)\}_{k>0}$

① setze $k := 0$

② berechne

$$\mu_{k+1} := \mathbf{z}_k^\star \mathbf{A}\mathbf{z}_k, \qquad \widetilde{\mathbf{z}}_{k+1} := (\mu_{k+1}\mathbf{I} - \mathbf{A})^{-1}\mathbf{z}_k, \qquad \mathbf{z}_{k+1} := \frac{\widetilde{\mathbf{z}}_{k+1}}{\|\widetilde{\mathbf{z}}_{k+1}\|_2}$$

③ erhöhe $k := k + 1$ und gehe nach ②

Bemerkung Die Iteration bricht man ab, wenn $\|\widetilde{\mathbf{z}}_{k+1}\|_2$ sehr groß wird, was ein Zeichen dafür ist, dass $\mu_{k+1}\mathbf{I} - \mathbf{A}$ fast singulär ist. ♦

Nach obiger Konvergenzdiskussion wird man vermuten, dass die Konvergenz dieses Algorithmus superlinear ist. Tatsächlich ist die Konvergenz sogar lokal kubisch, vorausgesetzt es gilt $\mathbf{A} = \mathbf{A}^\star$.

Satz 1.17 *Sei* $\mathbf{A} = \mathbf{A}^\star$ *und* $(\mu_0, \mathbf{z}_0)$ *eine hinreichend gute Näherung an ein Eigenpaar* $(\lambda, \mathbf{v})$ *von* $\mathbf{A}$ $(\|\mathbf{v}\|_2 = \|\mathbf{z}_0\|_2 = 1)$. *Dann konvergiert* μ_k *aus der Rayleigh-Quotient-Iteration lokal kubisch gegen* λ, *das heißt, es existiert ein* $C > 0$ *mit*

$$|\lambda - \mu_{k+1}| \le C|\lambda - \mu_k|^3, \quad k = 0, 1, 2, \ldots.$$

Beweis. Im weiteren bezeichne $\widehat{\lambda}$ denjenigen Eigenwert von $\mathbf{A}$, der am nächsten an λ ist. Ferner zerlegen wir

$$\mathbf{z}_{k-1} = \mathbf{x}_{k-1} + \mathbf{y}_{k-1} \quad \text{mit} \quad \mathbf{A}\mathbf{x}_{k-1} = \lambda\mathbf{x}_{k-1}, \quad \mathbf{x}_{k-1} \perp \mathbf{y}_{k-1} \tag{1.13}$$

in seinen Anteil $\mathbf{x}_{k-1}$ im Eigenraum zu λ und den Anteil $\mathbf{y}_{k-1}$ in den anderen Eigenräumen. Wir präzisieren die "Nähebedingung" wie folgt:

$$\|\mathbf{y}_{k-1}\|_2 \le \frac{1}{2}\|\mathbf{z}_{k-1}\|_2 = \frac{1}{2}, \tag{1.14}$$

$$|\lambda - \mu_k| \le \frac{1}{3}|\lambda - \widehat{\lambda}| = \frac{1}{3}\min_{\widetilde{\lambda}\in\sigma(\mathbf{A})\setminus\{\lambda\}} |\lambda - \widetilde{\lambda}|. \tag{1.15}$$

In diesem Fall gilt für jedes $\widetilde{\lambda} \in \sigma(\mathbf{A}) \setminus \{\lambda\}$

$$|\mu_k - \widetilde{\lambda}| = |\mu_k - \lambda + \lambda - \widetilde{\lambda}| \geq |\lambda - \widetilde{\lambda}| - |\lambda - \mu_k| \geq \frac{2}{3}|\lambda - \widetilde{\lambda}|. \tag{1.16}$$

Es gilt

$$|\lambda - \mu_{k+1}| = |\lambda - \mathbf{z}_k^\star \mathbf{A}\mathbf{z}_k| = |\mathbf{z}_k^\star(\lambda\mathbf{I} - \mathbf{A})\mathbf{z}_k| = \frac{|\widetilde{\mathbf{z}}_k^\star(\lambda\mathbf{I} - \mathbf{A})\widetilde{\mathbf{z}}_k|}{\|\widetilde{\mathbf{z}}_k\|_2^2}, \tag{1.17}$$

und Einsetzen von $\widetilde{\mathbf{z}}_k = (\mu_k\mathbf{I} - \mathbf{A})^{-1}\mathbf{z}_{k-1}$ ergibt

$$|\lambda - \mu_{k+1}| = \left| \frac{\mathbf{z}_{k-1}^\star(\mu_k\mathbf{I} - \mathbf{A})^{-2}(\lambda\mathbf{I} - \mathbf{A})\mathbf{z}_{k-1}}{\mathbf{z}_{k-1}^\star(\mu_k\mathbf{I} - \mathbf{A})^{-2}\mathbf{z}_{k-1}} \right|. \tag{1.18}$$

Wegen (1.13) gilt auch $(\mu_k\mathbf{I} - \mathbf{A})^{-2}\mathbf{x}_{k-1} = (\mu_k - \lambda)^{-2}\mathbf{x}_{k-1} \perp \mathbf{y}_{k-1}$, und daher folgt

$$\begin{aligned}
\mathbf{z}_{k-1}^\star(\mu_k\mathbf{I} - \mathbf{A})^{-2}\mathbf{z}_{k-1} &= \frac{1}{(\mu_k - \lambda)^2}\|\mathbf{x}_{k-1}\|_2^2 + \underbrace{\mathbf{y}_{k-1}^\star(\mu_k\mathbf{I} - \mathbf{A})^{-2}\mathbf{y}_{k-1}}_{\geq 0} \\
&\geq \frac{1}{(\mu_k - \lambda)^2}\|\mathbf{x}_{k-1}\|_2^2 \overset{(1.14)}{\geq} \frac{1}{4(\mu_k - \lambda)^2}.
\end{aligned}$$

Andererseits ist

$$\begin{aligned}
|\mathbf{z}_{k-1}^\star(\mu_k\mathbf{I} - \mathbf{A})^{-2}(\lambda\mathbf{I} - \mathbf{A})\mathbf{z}_{k-1}| &= |\mathbf{z}_{k-1}^\star(\mu_k\mathbf{I} - \mathbf{A})^{-2}(\lambda\mathbf{I} - \mathbf{A})\mathbf{y}_{k-1}| \\
&\overset{\text{Ortho-gonalität}}{=} |\mathbf{y}_{k-1}^\star(\mu_k\mathbf{I} - \mathbf{A})^{-2}(\lambda\mathbf{I} - \mathbf{A})\mathbf{y}_{k-1}| \\
&\leq \max_{\widetilde{\lambda} \in \sigma(\mathbf{A}) \setminus \{\lambda\}} \frac{1}{(\mu_k - \widetilde{\lambda})^2} \underbrace{|\mathbf{y}_{k-1}^\star(\lambda\mathbf{I} - \mathbf{A})\mathbf{y}_{k-1}|}_{= |\mathbf{z}_{k-1}^\star(\lambda\mathbf{I} - \mathbf{A})\mathbf{z}_{k-1}|, \text{ da } (\lambda\mathbf{I} - \mathbf{A})\mathbf{x}_{k-1} = \mathbf{0}} \\
&\overset{(1.16)}{\leq} \left(\frac{3}{2}\right)^2 \frac{1}{(\lambda - \widehat{\lambda})^2}|\mathbf{z}_{k-1}^\star(\lambda\mathbf{I} - \mathbf{A})\mathbf{z}_{k-1}| \\
&\overset{(1.17)}{=} \frac{9}{4}\frac{1}{(\lambda - \widehat{\lambda})^2}|\lambda - \mu_k|.
\end{aligned}$$

Im Hinblick auf (1.18) ergibt sich daher

$$|\lambda - \mu_{k+1}| \leq \frac{9}{4}\frac{|\lambda - \mu_k|}{(\lambda - \widehat{\lambda})^2} \cdot 4(\lambda - \mu_k)^2 = 9\frac{1}{(\lambda - \widehat{\lambda})^2}|\lambda - \mu_k|^3,$$

dies ist die kubische Konvergenzrate.

Wir müssen nur noch zeigen, dass die Iterierte $(\mu_{k+1}, \mathbf{z}_k)$ wieder den Abschätzungen (1.14) und (1.15) genügt. Die Bedingung (1.15) folgt aus

$$|\lambda - \mu_{k+1}| \le \frac{9}{(\lambda - \widehat{\lambda})^2} \frac{(\lambda - \widehat{\lambda})^2}{9} |\lambda - \mu_k| = |\lambda - \mu_k| \overset{(1.15)}{\le} \frac{1}{3} |\lambda - \widehat{\lambda}|.$$

Ferner folgt (1.14) aus

$$\widetilde{\mathbf{z}}_k = (\mu_k \mathbf{I} - \mathbf{A})^{-1} \mathbf{z}_{k-1} = \underbrace{\frac{1}{\mu_k - \lambda} \mathbf{x}_{k-1}}_{=\widetilde{\mathbf{x}}_k} + \underbrace{(\mu_k \mathbf{I} - \mathbf{A})^{-1} \mathbf{y}_{k-1}}_{=\widetilde{\mathbf{y}}_k},$$

da wegen

$$\|\widetilde{\mathbf{y}}_k\|_2 = \|(\mu_k \mathbf{I} - \mathbf{A})^{-1} \mathbf{y}_{k-1}\|_2 \overset{(1.16)}{\le} \frac{3}{2} \frac{1}{|\lambda - \widehat{\lambda}|} \|\mathbf{y}_{k-1}\|_2 \overset{(1.15)}{\le} \frac{1}{2} \frac{1}{|\lambda - \mu_k|} \|\mathbf{y}_{k-1}\|_2$$

der Anteil von $\|\mathbf{y}_k\|_2$ an $\|\mathbf{z}_k\|_2$ sogar noch kleiner ist als der von $\|\mathbf{y}_{k-1}\|_2$ an $\|\mathbf{z}_{k-1}\|_2$. ♠

Bemerkung Die Rayleigh-Quotient-Iteration kann auch bei nichtsymmetrischen Matrizen eingesetzt werden. Die Konvergenz ist dann lokal quadratisch. ♦

1.4 QR-Zerlegung

Im folgenden sei $\mathbf{A} \in \mathbb{R}^{m \times n}$, $m \ge n$, eine gegebene Matrix mit $\operatorname{rang} \mathbf{A} = n$. Die Grundidee der QR-Zerlegung ist eine Faktorisierung $\mathbf{A} = \mathbf{QR}$ in eine rechte obere Dreiecksmatrix $\mathbf{R} \in \mathbb{R}^{m \times n}$ und eine orthogonale Matrix $\mathbf{Q} \in \mathbb{R}^{m \times m}$.

Definition 1.18 Eine Matrix $\mathbf{Q} \in \mathbb{R}^{n \times n}$ heißt **orthogonal**, falls

$$\mathbf{Q}^\mathsf{T} \mathbf{Q} = \mathbf{I},$$

das heißt, falls die Spalten von $\mathbf{Q}$ eine Orthonormalbasis bilden.

Eigenschaften orthogonaler Matrizen:

1. Wegen
$$\|\mathbf{Qx}\|_2^2 = (\mathbf{Qx})^\mathsf{T} \mathbf{Qx} = \mathbf{x}^\mathsf{T} \mathbf{Q}^\mathsf{T} \mathbf{Qx} = \mathbf{x}^\mathsf{T} \mathbf{Ix} = \mathbf{x}^\mathsf{T} \mathbf{x} = \|\mathbf{x}\|_2^2$$
gilt $\|\mathbf{Qx}\|_2 = \|\mathbf{x}\|_2$ für alle $\mathbf{x} \in \mathbb{R}^n$.

2. Es gilt $\operatorname{cond}_2(\mathbf{Q}) = 1$, da

$$\|\mathbf{Q}\|_2 = \max_{\|\mathbf{x}\|_2=1} \|\mathbf{Q}\mathbf{x}\|_2 = 1 \quad \text{und} \quad \|\mathbf{Q}^{-1}\|_2 = \|\mathbf{Q}^\mathsf{T}\|_2 = \|\mathbf{Q}\|_2 = 1.$$

3. Mit $\mathbf{P}, \mathbf{Q} \in \mathbb{R}^{n\times n}$ orthogonal ist auch $\mathbf{PQ}$ orthogonal, da

$$(\mathbf{PQ})^\mathsf{T}\mathbf{PQ} = \mathbf{Q}^\mathsf{T}\mathbf{P}^\mathsf{T}\mathbf{PQ} = \mathbf{Q}^\mathsf{T}\mathbf{IQ} = \mathbf{Q}^\mathsf{T}\mathbf{Q} = \mathbf{I}.$$

Definition 1.19 Sei $\mathbf{v} \in \mathbb{R}^n \setminus \{\mathbf{0}\}$. Die Matrix

$$\mathbf{P} = \mathbf{I} - \frac{2}{\|\mathbf{v}\|_2^2}\mathbf{v}\mathbf{v}^\mathsf{T} \in \mathbb{R}^{n\times n}$$

heißt **Householder-Transformation**.

Lemma 1.20 $\mathbf{P}$ *ist eine symmetrische, orthogonale Matrix und es gilt*

$$\mathbf{Pv} = -\mathbf{v} \quad \textit{und} \quad \mathbf{Pw} = \mathbf{w}$$

für alle $\mathbf{w} \in \mathbb{R}^n$ *mit* $\mathbf{w} \perp \mathbf{v}$.

Beweis. Aus der Definition von $\mathbf{P}$ folgt unmittelbar, dass $\mathbf{P}$ symmetrisch ist. Weiter gilt

$$\begin{aligned}\mathbf{P}^\mathsf{T}\mathbf{P} = \mathbf{P}^2 &= \left(\mathbf{I} - \frac{2}{\|\mathbf{v}\|_2^2}\mathbf{v}\mathbf{v}^\mathsf{T}\right)\left(\mathbf{I} - \frac{2}{\|\mathbf{v}\|_2^2}\mathbf{v}\mathbf{v}^\mathsf{T}\right)\\ &= \mathbf{I} - \frac{4}{\|\mathbf{v}\|_2^2}\mathbf{v}\mathbf{v}^\mathsf{T} + \frac{4}{\|\mathbf{v}\|_2^4}\mathbf{v}\underbrace{\mathbf{v}^\mathsf{T}\mathbf{v}}_{=\|\mathbf{v}\|_2^2}\mathbf{v}^\mathsf{T}\\ &= \mathbf{I} - \frac{4}{\|\mathbf{v}\|_2^2}\mathbf{v}\mathbf{v}^\mathsf{T} + \frac{4}{\|\mathbf{v}\|_2^2}\mathbf{v}\mathbf{v}^\mathsf{T} = \mathbf{I}.\end{aligned}$$

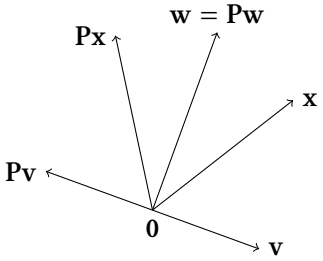
Außerdem ergibt sich für den Vektor $\mathbf{v}$ aus der Definition von $\mathbf{P}$

$$\mathbf{Pv} = \mathbf{Iv} - \frac{2}{\|\mathbf{v}\|_2^2}\mathbf{v}\underbrace{\mathbf{v}^\mathsf{T}\mathbf{v}}_{=\|\mathbf{v}\|_2^2} = \mathbf{v} - 2\mathbf{v} = -\mathbf{v}$$

und für beliebiges $\mathbf{w} \perp \mathbf{v}$

$$\mathbf{Pw} = \mathbf{Iw} - \frac{2}{\|\mathbf{v}\|_2^2}\mathbf{v}\underbrace{\mathbf{v}^\mathsf{T}\mathbf{w}}_{=0} = \mathbf{w}.$$

♠



Householder-Transformationen sind Spiegelungen!

Eine QR-Zerlegung kann erzeugt werden, indem man schrittweise die Matrix $\mathbf{A}$ durch Multiplikation mit geeigneten Householder-Transformationen $\mathbf{Q}_1, \mathbf{Q}_2, \ldots, \mathbf{Q}_n$ auf rechte obere Dreicksgestalt bringt. Das nächste Lemma erlaubt uns, solche Householder-Transformationen zu konstruieren, indem es für jedes $\mathbf{x} \in \mathbb{R}^n \setminus \{\mathbf{0}\}$ eine Householder-Transformation $\mathbf{Q}$ angibt, so dass

$$\mathbf{Q}\mathbf{x} = \sigma \mathbf{e}_1 \quad \text{mit} \quad \sigma \in \mathbb{R} \setminus \{0\}.$$

Lemma 1.21 *Gegeben sei* $\mathbf{x} \in \mathbb{R}^n \setminus \{\mathbf{0}\}$. *Für* $\mathbf{x} \notin \operatorname{span}\{\mathbf{e}_1\}$ *und*

$$\mathbf{v} = \mathbf{x} + \sigma \mathbf{e}_1 \quad \textit{mit} \quad \sigma = \pm \begin{cases} \operatorname{sign}(x_1)\|\mathbf{x}\|_2, & \textit{falls } x_1 \neq 0, \\ \|\mathbf{x}\|_2, & \textit{falls } x_1 = 0, \end{cases} \tag{1.19}$$

oder $\mathbf{x} \in \operatorname{span}\{\mathbf{e}_1\}$ *mit* $\sigma = \operatorname{sign}(x_1)\|\mathbf{x}\|_2$ *gilt*

$$\left(\mathbf{I} - \frac{2}{\|\mathbf{v}\|_2^2}\mathbf{v}\mathbf{v}^\intercal\right)\mathbf{x} = -\sigma \mathbf{e}_1.$$

Beweis. In beiden Fällen, $\mathbf{x} \notin \operatorname{span}\{\mathbf{e}_1\}$ oder $\mathbf{x} \in \operatorname{span}\{\mathbf{e}_1\}$, ist $\mathbf{v} \neq \mathbf{0}$. Weiter gilt

$$\|\mathbf{x} + \sigma \mathbf{e}_1\|_2^2 = \|\mathbf{x}\|_2^2 + 2\sigma \mathbf{x}^\intercal \mathbf{e}_1 + \sigma^2 = 2(\mathbf{x} + \sigma \mathbf{e}_1)^\intercal \mathbf{x}.$$

Daraus erhält man

$$2\mathbf{v}^\intercal \mathbf{x} = 2(\mathbf{x} + \sigma \mathbf{e}_1)^\intercal \mathbf{x} = \|\mathbf{x} + \sigma \mathbf{e}_1\|_2^2 = \|\mathbf{v}\|_2^2,$$

was zusammen mit (1.19) die Darstellung

$$\frac{2}{\|\mathbf{v}\|_2^2}\mathbf{v}(\mathbf{v}^\intercal \mathbf{x}) = \mathbf{x} + \sigma \mathbf{e}_1$$

liefert. Dies impliziert die Behauptung. ♠

Bemerkung Damit im Fall $\mathbf{x} \notin \operatorname{span}\{\mathbf{e}_1\}$ mit $x_1 \neq 0$ bei der Berechnung von $\mathbf{v}$ keine Auslöschung auftritt, wählen wir σ mit dem oberen Vorzeichen, das heißt

$$\mathbf{v} = \mathbf{x} + \frac{x_1}{|x_1|}\|\mathbf{x}\|_2\mathbf{e}_1, \qquad \|\mathbf{v}\|_2^2 = 2\|\mathbf{x}\|_2^2 + 2|x_1|\|\mathbf{x}\|_2. \tag{1.20}$$

♦

Satz 1.22 *Sei $\mathbf{A} \in \mathbb{R}^{m\times n}$ mit $\operatorname{rang}(\mathbf{A}) = n$ (also $m \geq n$). Dann existiert eine orthogonale Matrix $\mathbf{Q} \in \mathbb{R}^{m\times m}$ und eine obere Dreiecksmatrix $\mathbf{R} \in \mathbb{R}^{m\times n}$ mit*

$$\mathbf{A} = \mathbf{Q}\cdot\mathbf{R} = \mathbf{Q}\cdot\underbrace{\left[\begin{array}{ccc} \star & \cdots & \star \\ & \ddots & \vdots \\ & & \star \\ \hline & \mathbf{0} & \end{array}\right]}_{n}\begin{matrix}\rceil\\ n\\ \rfloor\\ \rceil\\ m-n\\ \rfloor\end{matrix}.$$

Beweis. Wir bestimmen die gesuchte Zerlegung, indem wir in jedem Schritt eine Householder-Transformation an $\mathbf{A}$ heranmultiplizieren, um sukzessive die Spalten von 1 bis n von $\mathbf{R}$ zu erhalten:

$$\mathbf{Q}_n\mathbf{Q}_{n-1}\cdots\mathbf{Q}_1\mathbf{A} = \mathbf{R}. \tag{1.21}$$

Wegen der Symmetrie der $\mathbf{Q}_i$, $1 \leq i < n$, ist $\mathbf{Q}$ dann gegeben durch

$$\mathbf{Q} = \mathbf{Q}_1\mathbf{Q}_2\cdots\mathbf{Q}_n.$$

Im ersten Schritt setzen wir $\mathbf{A}_1 := \mathbf{A}$ und $\mathbf{x} = \mathbf{a}_1$ (erste Spalte von $\mathbf{A}_1$) und bestimmen die Householder-Transformation $\mathbf{Q}_1 \in \mathbb{R}^{m\times m}$ gemäß (1.20). Es folgt

$$\mathbf{Q}_1\mathbf{a}_1 = r_{1,1}\mathbf{e}_1 \quad \text{mit} \quad |r_{1,1}| = \|\mathbf{a}_1\|_2 \neq 0,$$

beziehungsweise

$$\mathbf{Q}_1\mathbf{A}_1 = \left[\begin{array}{c|c} r_{1,1} & \mathbf{r}_1 \\ \hline \mathbf{0} & \mathbf{A}_2 \end{array}\right], \qquad \mathbf{A}_2 \in \mathbb{R}^{(m-1)\times(n-1)}, \qquad \mathbf{r}_1^{\mathsf{T}} \in \mathbb{R}^{n-1}.$$

Im nächsten Schritt setzen wir $\mathbf{x} = \mathbf{a}_2 \in \mathbb{R}^{m-1}$ (erste Spalte von $\mathbf{A}_2$) und wählen wiederum die Householder-Matrix $\widetilde{\mathbf{Q}}_2 \in \mathbb{R}^{(m-1)\times(m-1)}$ gemäß (1.20). Wir erhalten

$$\widetilde{\mathbf{Q}}_2\mathbf{A}_2 = \left[\begin{array}{c|c} r_{2,2} & \mathbf{r}_2 \\ \hline \mathbf{0} & \mathbf{A}_3 \end{array}\right], \quad |r_{2,2}| = \|\mathbf{a}_2\|_2 \neq 0, \quad \mathbf{A}_3 \in \mathbb{R}^{(m-2)\times(n-2)}, \quad \mathbf{r}_2^\mathsf{T} \in \mathbb{R}^{n-2},$$

beziehungsweise

$$\underbrace{\left[\begin{array}{c|c} 1 & \mathbf{0} \\ \hline \mathbf{0} & \widetilde{\mathbf{Q}}_2 \end{array}\right]}_{=:\mathbf{Q}_2} \mathbf{Q}_1\mathbf{A} = \left[\begin{array}{c|c} r_{1,1} & \mathbf{r}_1 \\ \hline \mathbf{0} & \widetilde{\mathbf{Q}}_2\mathbf{A}_2 \end{array}\right] = \left[\begin{array}{c|cc} r_{1,1} & \multicolumn{2}{c}{\mathbf{r}_1} \\ \hline & r_{2,2} & \mathbf{r}_2 \\ \mathbf{0} & \mathbf{0} & \mathbf{A}_3 \end{array}\right].$$

Die erste Zeile $\mathbf{r}_1$ verändert sich nicht mehr. Die Matrix $\mathbf{Q}_2$ kann ebenfalls als $(m \times m)$-Householder-Transformation aufgefasst werden mit $\mathbf{v} = \left[\begin{smallmatrix} 0 \\ \widetilde{\mathbf{v}} \end{smallmatrix}\right]$.

Auf diese Weise erhalten wir sukzessive die gewünschte Zerlegung (1.21). Man beachte, dass $|r_{i,i}| = \|\mathbf{a}_i\|_2$ $(1 \leq i \leq n)$ immer von Null verschieden ist, da ansonsten $\mathbf{A}_i$ und damit auch $\mathbf{A}$ einen Rangdefekt hätte. ♠

Bemerkungen

1. Bei der Implementierung ist darauf zu achten, dass Householder-Transformationen $\mathbf{P}$ *niemals* explizit gebildet werden, denn sonst kostet die Berechnung $\mathbf{P} \cdot \mathbf{A}$ $m^2 n$ Multiplikationen. Besser ist

   $$\mathbf{P}\mathbf{A} = \mathbf{A} - \frac{2}{\|\mathbf{v}\|_2^2}\mathbf{v}\underbrace{\mathbf{v}^\mathsf{T}\mathbf{A}}_{=\mathbf{w}^\mathsf{T}} = \mathbf{A} - \frac{2}{\|\mathbf{v}\|_2^2}\mathbf{v}\mathbf{w}^\mathsf{T}, \quad \mathbf{w} = \mathbf{A}^\mathsf{T}\mathbf{v}$$

   mit $\mathcal{O}(mn)$ Multiplikationen. Wenn man $\mathbf{P}$ später verwenden will, speichert man den Vektor $\mathbf{v}$ ab.

2. Die während der QR-Zerlegung anfallenden Vektoren

   $$\mathbf{v}_i = [0, \ldots, 0, 1, v_{i,i+1}, \ldots, v_{i,m}]^\mathsf{T}$$

   lassen sich analog zur LR-Zerlegung wieder in der freiwerdenden linken unteren Dreiecksmatrix von $\mathbf{A}$ speichern. Die Matrix $\mathbf{Q}$ ist dann wie folgt gegeben

   $$\mathbf{Q} = \prod_{i=1}^{n} \left(\mathbf{I} - \frac{2}{\|\mathbf{v}_i\|_2^2}\mathbf{v}_i\mathbf{v}_i^\mathsf{T}\right).$$ ♦

Algorithmus 1.23
input: Matrix $\mathbf{A} \in \mathbb{R}^{m\times n}$
output: Zerlegung $\mathbf{A} = \mathbf{QR}$ mit einer orthogonalen Matrix $\mathbf{Q} \in \mathbb{R}^{m\times m}$ und einer rechten oberen Dreiecksmatrix $\mathbf{R} \in \mathbb{R}^{m\times n}$

① Initialisierung: setze $\mathbf{A}_1 := \mathbf{A}$ und $i := 1$

② mit Hilfe der ersten Spalte $\mathbf{x}$ von $\mathbf{A}_i$ berechne

$$\mathbf{v}_i := \mathbf{x} + \frac{x_1}{|x_1|}\|\mathbf{x}\|_2\mathbf{e}_1$$

③ setze gemäß (1.20)

$$\beta_i := \frac{2}{\|\mathbf{v}\|_2^2} = \frac{1}{\|\mathbf{x}\|_2^2 + |x_1|\|\mathbf{x}\|_2}$$

④ berechne $\mathbf{w}_i := \beta_i \mathbf{A}_i^\mathsf{T}\mathbf{v}_i$

⑤ ersetze $\mathbf{A}_i$ durch $\mathbf{A}_i - \mathbf{v}_i\mathbf{w}_i^\mathsf{T}$

⑥ erhalte $\mathbf{A}_{i+1}$ aus $\mathbf{A}_i$ durch Streichen der ersten Zeile und Spalte

⑦ falls $i < n$ erhöhe $i := i + 1$ und gehe nach ②

Aufwand: Wir bilanzieren den Aufwand im i-ten Schritt. Wie man leicht einsieht benutzt der i-te Schritt:

$\mathbf{v}_i$:	$m - i + 3$ Multiplikationen
β_i:	2 Multiplikationen
$\mathbf{w}_i$:	$(m - i + 2)(n - i + 1)$ Multiplikationen
$\mathbf{A}_i$:	$(m - i)(n - i + 1)$ Multiplikationen
	$\approx 2(m - i + 1)(n - i + 1)$ Multiplikationen

Es ist weiter einfach einzusehen, dass pro Addition auch mindestens eine Multiplikation stattfindet, und auch, dass jeweils im i-ten Schritt nur eine skalare Inverse berechnet werden muss. Wir beschränken uns daher auf den Gesamtaufwand der Multiplikationen.

Für den Gesamtaufwand ergibt sich daher nun

$$\begin{aligned} 2\sum_{i=1}^{n}(m-i+1)(n-i+1) &\overset{j:=n-i+1}{=} 2\sum_{j=1}^{n} j(m-n+j) \\ &= 2\sum_{j=1}^{n} j^2 + 2(m-n)\sum_{j=1}^{n} j \\ &= \frac{2}{3}n^3 + (m-n)n^2 + \mathcal{O}(mn) \\ &= mn^2 - \frac{1}{3}n^3 + \mathcal{O}(mn), \end{aligned}$$

dies bedeutet dass der Aufwand etwa doppelt so hoch ist wie bei der *LR*-Zerlegung.

Die *QR*-Zerlegung kann wie die *LR*-Zerlegung zur Lösung eines nichtsingulären linearen Gleichungssystems $\mathbf{Ax} = \mathbf{b}$ (also $m = n$) verwendet werden. Dies geschieht in folgender Weise: Zerlege $\mathbf{A} = \mathbf{QR}$, und löse $\mathbf{QRx} = \mathbf{b}$ durch Rückwärtssubstitution

$$\mathbf{Rx} = \underbrace{\mathbf{Q}^\mathsf{T}\mathbf{b}}_{\mathcal{O}(n^2)\text{ Operationen}}.$$

Bemerkung Die *QR*-Zerlegung gehört zu den "stabilsten" Algorithmen in der numerischen linearen Algebra. Der Grund dafür ist, dass Orthogonaltransformationen keine Fehlerverstärkung bringen, da $\mathrm{cond}_2(\mathbf{Q}) = 1$. Die abschließende Rückwärtssubstitution hat die gleiche Kondition wie das Ausgangsproblem, da wegen $\|\mathbf{Qx}\|_2 = \|\mathbf{x}\|_2$ folgt

$$\begin{aligned} \|\mathbf{A}\|_2\|\mathbf{A}^{-1}\|_2 &= \left(\max_{\|\mathbf{x}\|_2=1}\|\mathbf{Ax}\|_2\right)\left(\max_{\|\mathbf{x}\|_2}\|\mathbf{A}^{-1}\mathbf{x}\|_2\right) \\ &= \left(\max_{\|\mathbf{x}\|_2=1}\|\mathbf{QRx}\|_2\right)\left(\max_{\|\mathbf{x}\|_2=1}\|\mathbf{R}^{-1}\mathbf{Q}^\mathsf{T}\mathbf{x}\|_2\right) \\ &= \left(\max_{\|\mathbf{x}\|_2=1}\|\mathbf{Rx}\|_2\right)\left(\max_{\|\mathbf{y}\|_2=1}\|\mathbf{R}^{-1}\mathbf{y}\|_2\right) \\ &= \|\mathbf{R}\|_2\|\mathbf{R}^{-1}\|_2, \end{aligned}$$

das heißt

$$\mathrm{cond}_2(\mathbf{R}) = \mathrm{cond}_2(\mathbf{A}).$$ ♦

Beispiel 1.24 Gesucht ist die *QR*-Zerlegung von

$$\mathbf{A} = \mathbf{A}_1 = \begin{bmatrix} 1 & -11/2 \\ -2 & 0 \\ 2 & -1 \end{bmatrix}.$$

Die erste Spalte von $\mathbf{A}$ ist

$$\mathbf{a}_1 = \begin{bmatrix} 1 \\ -2 \\ 2 \end{bmatrix}, \qquad \|\mathbf{a}_1\|_2 = \sqrt{1+4+4} = 3.$$

Also ist

$$\mathbf{v}_1 = \mathbf{a}_1 + \operatorname{sign}(a_{1,1})\|\mathbf{a}_1\|_2 \mathbf{e}_1 = \begin{bmatrix} 1 \\ -2 \\ 2 \end{bmatrix} + 3 \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix} = \begin{bmatrix} 4 \\ -2 \\ 2 \end{bmatrix}.$$

Somit folgt

$$\beta_1 = \frac{1}{\|\mathbf{a}_1\|_2^2 + |a_{1,1}|\|\mathbf{a}_1\|_2} = \frac{1}{12}$$

woraus sich

$$\mathbf{w}_1 = \beta \mathbf{A}_1^\mathsf{T} \mathbf{v}_1 = \frac{1}{12} \begin{bmatrix} 1 & -2 & 2 \\ -11/2 & 0 & -1 \end{bmatrix} \begin{bmatrix} 4 \\ -2 \\ 2 \end{bmatrix} = \begin{bmatrix} 1 \\ -2 \end{bmatrix}$$

ergibt. Dies bedeutet

$$\mathbf{Q}_1 \mathbf{A}_1 = \mathbf{A}_1 - \mathbf{v}_1 \mathbf{w}_1^\mathsf{T} = \begin{bmatrix} 1 & -11/2 \\ -2 & 0 \\ 2 & -1 \end{bmatrix} - \begin{bmatrix} 4 & -8 \\ -2 & 4 \\ 2 & -4 \end{bmatrix} = \begin{bmatrix} -3 & 5/2 \\ 0 & -4 \\ 0 & 3 \end{bmatrix}.$$

Die erste Spalte stimmt dabei mit $-\sigma \mathbf{e}_1$ überein, so war die Householder-Transformation schließlich konstruiert. Sie ist übrigens gegeben durch

$$\mathbf{Q}_1 = \mathbf{I} - \beta \mathbf{v}_1 \mathbf{v}_1^\mathsf{T} = \frac{1}{3} \begin{bmatrix} -1 & 2 & -2 \\ 2 & 2 & 1 \\ -2 & 1 & 2 \end{bmatrix}.$$

Nun ist

$$\mathbf{A}_2 = \mathbf{a}_2 = \begin{bmatrix} -4 \\ 3 \end{bmatrix}, \qquad \|\mathbf{a}_2\|_2 = \sqrt{16+9} = 5.$$

Mit

$$\mathbf{v}_2 = \mathbf{a}_2 + \operatorname{sign}(a_{2,1})\|\mathbf{a}_2\|_2 \mathbf{e}_1 = \begin{bmatrix} -4 \\ 3 \end{bmatrix} - 5 \begin{bmatrix} 1 \\ 0 \end{bmatrix} = \begin{bmatrix} -9 \\ 3 \end{bmatrix}$$

und

$$\beta_2 = \frac{1}{\|\mathbf{a}_2\|_2^2 + |a_{2,1}|\|\mathbf{a}_2\|_2} = \frac{1}{45}$$

folgt

$$\mathbf{w}_2 = \beta \mathbf{A}_2^\mathsf{T}\mathbf{v}_2 = \frac{1}{45}\begin{bmatrix}-4 & 3\end{bmatrix}\begin{bmatrix}-9\\3\end{bmatrix} = \begin{bmatrix}1\end{bmatrix}.$$

Damit ergibt sich

$$\widetilde{\mathbf{Q}}_2\mathbf{A}_2 = \mathbf{A}_2 - \mathbf{v}_2\mathbf{w}_2^\mathsf{T} = \begin{bmatrix}-4\\3\end{bmatrix} - \begin{bmatrix}-9\\3\end{bmatrix} = \begin{bmatrix}5\\0\end{bmatrix}.$$

Dabei hat die Matrix $\mathbf{Q}_2$ die Form

$$\mathbf{Q}_2 = \left[\begin{array}{c|c} 1 & \mathbf{0} \\ \hline \mathbf{0} & \mathbf{I} - \beta\mathbf{v}_2\mathbf{v}_2^\mathsf{T} \end{array}\right] = \frac{1}{5}\begin{bmatrix}5 & 0 & 0\\ 0 & -4 & 3\\ 0 & 3 & 4\end{bmatrix}.$$

Für $\mathbf{Q}$ erhalten wir schließlich

$$\mathbf{Q} = \mathbf{Q}_1\mathbf{Q}_2 = \frac{1}{15}\begin{bmatrix}-1 & 2 & -2\\ 2 & 2 & -1\\ -2 & 1 & 2\end{bmatrix}\begin{bmatrix}5 & 0 & 0\\ 0 & -4 & 3\\ 0 & 3 & 4\end{bmatrix} = \frac{1}{15}\begin{bmatrix}-5 & -14 & -2\\ 10 & -5 & 10\\ -10 & 2 & 11\end{bmatrix},$$

während $\mathbf{R}$ gegeben ist durch

$$\mathbf{R} = \begin{bmatrix}-3 & 5/2\\ 0 & 5\\ 0 & 0\end{bmatrix}. \qquad \clubsuit$$

Bemerkung Algorithmus 1.23 bricht zusammen, wenn $\operatorname{rang}(\mathbf{A}) = p < n$. In diesem Fall muss man Spalten von $\mathbf{A}$ permutieren (ähnlich zur Pivotsuche) und erhält eine Faktorisierung der Art

$$\mathbf{Q}^\mathsf{T}\mathbf{A}\mathbf{P} = \begin{bmatrix}\mathbf{R}_1 & \mathbf{R}_2\\ \mathbf{0} & \mathbf{0}\end{bmatrix}$$

mit einer Permutationsmatrix $\mathbf{P} \in \mathbb{R}^{n\times n}$, einer oberen Dreiecksmatrix $\mathbf{R}_1 \in \mathbb{R}^{p\times p}$, und einer eventuell vollbesetzten Matrix $\mathbf{R}_2 \in \mathbb{R}^{p\times(n-p)}$. $\blacklozenge$

1.5 QR-Verfahren

Das QR–Verfahren ist *das* in der Praxis eingesetzte Verfahren, wenn alle Eigenwerte benötigt werden. Das Verfahren an sich ist sehr einfach:

Algorithmus 1.25 (QR-Verfahren)
input: Matrix $\mathbf{A} \in \mathbb{K}^{n\times n}$
output: Folge von Iterierten $\{\mathbf{A}_k\}_{k\geq 0}$

① setze $\mathbf{A}_0 := \mathbf{A}$ und $k := 0$

② berechne die QR-Zerlegung

$$\mathbf{A}_k = \mathbf{Q}_k\mathbf{R}_k \tag{1.22}$$

und setze

$$\mathbf{A}_{k+1} := \mathbf{R}_k\mathbf{Q}_k \tag{1.23}$$

③ erhöhe $k := k + 1$ und gehe nach ②

Lemma 1.26 *Die Matrizen $\mathbf{A}_k$ aus Algorithmus 1.25 besitzen folgende Eigenschaften:*

1. $\mathbf{A}_{k+1} = \mathbf{Q}_k^{\mathsf{T}}\mathbf{A}_k\mathbf{Q}_k$
2. $\mathbf{A}_{k+1} = (\mathbf{Q}_0\mathbf{Q}_1\cdots\mathbf{Q}_k)^{\mathsf{T}}\mathbf{A}(\mathbf{Q}_0\mathbf{Q}_1\cdots\mathbf{Q}_k)$
3. $\mathbf{A}^{k+1} = (\mathbf{Q}_0\mathbf{Q}_1\cdots\mathbf{Q}_k)(\mathbf{R}_k\mathbf{R}_{k-1}\cdots\mathbf{R}_0)$

Beweis.

1. Mit (1.22) und (1.23) ergibt sich
$$\mathbf{A}_{k+1} = \mathbf{R}_k\mathbf{Q}_k = \mathbf{Q}_k^{\mathsf{T}}\mathbf{Q}_k\mathbf{R}_k\mathbf{Q}_k = \mathbf{Q}_k^{\mathsf{T}}(\mathbf{Q}_k\mathbf{R}_k)\mathbf{Q}_k = \mathbf{Q}_k^{\mathsf{T}}\mathbf{A}_k\mathbf{Q}_k.$$
2. Die Behauptung ergibt sich sofort aus der ersten Aussage wegen $\mathbf{A}_0 = \mathbf{A}$.
3. Für $k = 0$ entspricht die Aussage genau (1.22). Der Induktionsschritt $k \mapsto k + 1$ folgt nun aus
$$\mathbf{Q}_{k+1}\mathbf{R}_{k+1} = \mathbf{A}_{k+1} = (\mathbf{Q}_0\mathbf{Q}_1\cdots\mathbf{Q}_k)^{\mathsf{T}}\mathbf{A}(\mathbf{Q}_0\mathbf{Q}_1\cdots\mathbf{Q}_k)$$

zusammen mit der Induktionsannahme

$$(\mathbf{Q}_0 \cdots \mathbf{Q}_k \underbrace{\mathbf{Q}_{k+1})(\mathbf{R}_{k+1}}_{=\mathbf{A}_{k+1}} \mathbf{R}_k \cdots \mathbf{R}_0) = \mathbf{A} \underbrace{(\mathbf{Q}_0\mathbf{Q}_1 \cdots \mathbf{Q}_k)(\mathbf{R}_k\mathbf{R}_{k-1} \cdots \mathbf{R}_0)}_{=\mathbf{A}^{k+1}} = \mathbf{A}^{k+2}.$$

♠

Bemerkungen

1. Wegen der ersten Aussage von Lemma 1.26 sind alle Matrizen $\mathbf{A}_k$ ähnlich zueinander und besitzen daher dieselben Eigenwerte.
2. Anstelle der *QR*-Zerlegung kann man auch eine *LR*-Zerlegung verwenden: berechne in (1.22) die *LR*-Zerlegung $\mathbf{A}_k = \mathbf{L}_k\mathbf{R}_k$ und setze in (1.23) $\mathbf{A}_{k+1} := \mathbf{R}_k\mathbf{L}_k$. Dies ist das *LR-Verfahren*, das jedoch instabil ist. ♦

Die Konvergenz des *QR*-Verfahrens zeigen wir nur für den einfachen Fall, dass $\mathbf{A}$ diagonalisierbar ist mit betragsmäßig verschiedenen Eigenwerten.

Satz 1.27 *Sei* $\mathbf{A} = \mathbf{V}\mathbf{D}\mathbf{V}^{-1} \in \mathbb{R}^{n\times n}$ *reell diagonalisierbar mit betragsmäßig verschiedenen Eigenwerten*

$$|\lambda_1| > |\lambda_2| > \ldots > |\lambda_n| > 0, \qquad \mathbf{D} = \operatorname{diag}(\lambda_1, \lambda_2, \ldots, \lambda_n)$$

und $\mathbf{V} = [\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_n]$ *die zugehörige Eigenvektormatrix. Existiert eine LR-Zerlegung von* $\mathbf{V}^{-1}$*, dann sind die Matrizen* $\mathbf{A}_k$ *asymptotisch rechte, obere Dreiecksmatrizen und ihr Diagonalanteil* $\operatorname{diag}(\mathbf{A}_k)$ *konvergiert für* $k \to \infty$ *mindestens linear gegen* $\mathbf{D}$.

Beweis. Die dritte Aussage aus Lemma 1.26 liefert die *QR*-Zerlegung von $\mathbf{A}^k$,

$$\mathbf{A}^k = (\mathbf{Q}_0\mathbf{Q}_1 \cdots \mathbf{Q}_{k-1})(\mathbf{R}_{k-1}\mathbf{R}_{k-2} \cdots \mathbf{R}_0).$$

Andererseits ist wegen der Existenz der *LR*-Zerlegung von $\mathbf{V}^{-1} = \mathbf{L}\mathbf{U}$

$$\mathbf{A}^k = (\mathbf{V}\mathbf{D}\mathbf{V}^{-1})^k = \mathbf{V}\mathbf{D}^k\mathbf{V}^{-1} = \mathbf{V}\mathbf{D}^k\mathbf{L}\mathbf{U} = (\underbrace{\mathbf{V}\mathbf{D}^k\mathbf{L}\mathbf{D}^{-k}}_{=:\mathbf{B}_k})\mathbf{D}^k\mathbf{U} = \mathbf{B}_k\mathbf{D}^k\mathbf{U}. \tag{1.24}$$

Die Matrix $\mathbf{B}_k$ ist regulär und besitzt eine *QR*-Zerlegung $\mathbf{B}_k = \mathbf{P}_k\mathbf{S}_k$ mit einer invertierbaren oberen Dreiecksmatrix $\mathbf{S}_k$. Damit ist

$$\mathbf{A}^k = \mathbf{P}_k(\underbrace{\mathbf{S}_k\mathbf{D}^k\mathbf{U}}_{\substack{\text{rechte, obere}\\ \text{Dreiecksmatrix}}})$$

eine weitere QR-Zerlegung von $\mathbf{A}^k$. Wegen der Eindeutigkeit der QR-Zerlegung folgt

$$\mathbf{Q}_0\mathbf{Q}_1\cdots\mathbf{Q}_{k-1} = \mathbf{P}_k\mathbf{T}_k^\intercal, \qquad \mathbf{R}_{k-1}\mathbf{R}_{k-2}\cdots\mathbf{R}_0 = \mathbf{T}_k\mathbf{S}_k\mathbf{D}^k\mathbf{U} \tag{1.25}$$

für eine Diagonalmatrix $\mathbf{T}_k$ mit den Einträgen ± 1. Wegen

$$\mathbf{Q}_k = (\mathbf{Q}_0\mathbf{Q}_1\cdots\mathbf{Q}_{k-1})^{-1}(\mathbf{Q}_0\mathbf{Q}_1\cdots\mathbf{Q}_k) \overset{(1.25)}{=} \mathbf{T}_k\mathbf{P}_k^\intercal\mathbf{P}_{k+1}\mathbf{T}_{k+1}^\intercal$$

und

$$\begin{aligned}\mathbf{R}_k &= (\mathbf{R}_k\mathbf{R}_{k-1}\cdots\mathbf{R}_0)(\mathbf{R}_{k-1}\mathbf{R}_{k-2}\cdots\mathbf{R}_0)^{-1}\\ &\overset{(1.25)}{=} \mathbf{T}_{k+1}\mathbf{S}_{k+1}\mathbf{D}^{k+1}\mathbf{U}\mathbf{U}^{-1}\mathbf{D}^{-k}\mathbf{S}_k^{-1}\mathbf{T}_k^\intercal\\ &= \mathbf{T}_{k+1}\mathbf{S}_{k+1}\mathbf{D}\mathbf{S}_k^{-1}\mathbf{T}_k^\intercal\end{aligned}$$

ergibt sich aus (1.22)

$$\begin{aligned}\mathbf{A}_k = \mathbf{Q}_k\mathbf{R}_k &= \mathbf{T}_k\mathbf{P}_k^\intercal\mathbf{P}_{k+1}\mathbf{T}_{k+1}^\intercal\mathbf{T}_{k+1}\mathbf{S}_{k+1}\mathbf{D}\mathbf{S}_k^{-1}\mathbf{T}_k^\intercal\\ &= \mathbf{T}_k\mathbf{S}_k\underbrace{\mathbf{S}_k^{-1}\mathbf{P}_k^\intercal}_{=\mathbf{B}_k^{-1}}\underbrace{\mathbf{P}_{k+1}\mathbf{S}_{k+1}}_{=\mathbf{B}_{k+1}}\mathbf{D}\mathbf{S}_k^{-1}\mathbf{T}_k^\intercal\\ &= \mathbf{T}_k\mathbf{S}_k\mathbf{B}_k^{-1}\mathbf{B}_{k+1}\mathbf{D}\mathbf{S}_k^{-1}\mathbf{T}_k^\intercal.\end{aligned} \tag{1.26}$$

Bezeichnen wir mit $\ell_{i,j}$ die Einträge von $\mathbf{L}$, dann ist insbesondere $\ell_{i,i} = 1$ und aus der Anordnung der Eigenwerte in $\mathbf{D}$ ergibt sich

$$[\mathbf{D}^k\mathbf{L}\mathbf{D}^{-k}]_{i,j} = \lambda_i^k\ell_{i,j}\lambda_j^{-k} = \begin{cases}0, & \text{falls } i<j,\\ 1, & \text{falls } i=j,\\ \mathcal{O}\left(\left|\frac{\lambda_i}{\lambda_j}\right|^k\right), & \text{falls } i>j.\end{cases} \tag{1.27}$$

Setzen wir $q := \max_{i>j}\{|\lambda_i/\lambda_j|\} < 1$, so folgt (vgl. (1.24)) hieraus

$$\mathbf{B}_k = \mathbf{V}\mathbf{D}^k\mathbf{L}\mathbf{D}^{-k} = \mathbf{V} + \mathbf{E}_k \quad \text{mit} \quad \|\mathbf{E}_k\|_2 = \mathcal{O}(q^k), \quad k\to\infty.$$

Demzufolge erhalten wir

$$\mathbf{B}_k^{-1}\mathbf{B}_{k+1} = (\mathbf{V}+\mathbf{E}_k)^{-1}(\mathbf{V}+\mathbf{E}_{k+1}) = \mathbf{I}+\mathbf{F}_k \quad \text{mit} \quad \|\mathbf{F}_k\|_2 = \mathcal{O}(q^k), \quad k\to\infty$$

und eingesetzt in (1.26) ergibt sich

$$\mathbf{A}_k = \mathbf{T}_k\mathbf{S}_k\mathbf{D}\mathbf{S}_k^{-1}\mathbf{T}_k^\intercal + \mathbf{T}_k\mathbf{S}_k\mathbf{F}_k\mathbf{D}\mathbf{S}_k^{-1}\mathbf{T}_k^\intercal.$$

Da $\mathbf{P}_k$ und $\mathbf{T}_k$ orthogonale Matrizen sind, lässt sich der zweite Term abschätzen gemäß

$$\begin{aligned}\|\mathbf{T}_k\mathbf{S}_k\mathbf{F}_k\mathbf{D}\mathbf{S}_k^{-1}\mathbf{T}_k^\intercal\|_2 &\le \underbrace{\|\mathbf{T}_k\|_2}_{=1}\ \underbrace{\|\mathbf{S}_k\|_2}_{=\|\mathbf{P}_k^\intercal\mathbf{B}_k\|_2=\|\mathbf{B}_k\|_2}\ \|\mathbf{F}_k\|_2\ \underbrace{\|\mathbf{D}\|_2}_{=|\lambda_1|}\ \underbrace{\|\mathbf{S}_k^{-1}\|_2}_{=\|\mathbf{B}_k^{-1}\mathbf{P}_k\|_2=\|\mathbf{B}_k^{-1}\|_2}\ \underbrace{\|\mathbf{T}_k^\intercal\|_2}_{=1}\\ &\le \operatorname{cond}_2(\mathbf{B}_k)|\lambda_1|\|\mathbf{F}_k\|_2 = \mathcal{O}(q^k).\end{aligned}$$

Die Matrix $\mathbf{A}_k$ konvergiert also wegen

$$\|\mathbf{A}_k - \mathbf{T}_k\mathbf{S}_k\mathbf{D}\mathbf{S}_k^{-1}\mathbf{T}_k^\intercal\|_2 = \|\mathbf{T}_k\mathbf{S}_k\mathbf{F}_k\mathbf{D}\mathbf{S}_k^{-1}\mathbf{T}_k^\intercal\|_2 = \mathcal{O}(q^k), \quad k\to\infty$$

mindestens linear gegen eine obere Dreiecksmatrix und es gilt

$$\begin{aligned}\operatorname{diag}\left(\mathbf{A}_k - \mathbf{T}_k\mathbf{S}_k\mathbf{D}\mathbf{S}_k^{-1}\mathbf{T}_k^\intercal\right) &= \operatorname{diag}(\mathbf{A}_k) - \underbrace{\mathbf{T}_k\operatorname{diag}(\mathbf{S}_k)\mathbf{D}\operatorname{diag}(\mathbf{S}_k)^{-1}\mathbf{T}_k^\intercal}_{=\mathbf{D}}\\ &= \operatorname{diag}(\mathbf{A}_k) - \mathbf{D} \to \mathbf{0}.\end{aligned}$$

♠

1.6 Implementierung des QR-Verfahrens

Reduktion auf Hessenberg-Form: Für beliebige Matrizen $\mathbf{A}\in\mathbb{R}^{n\times n}$ wäre das QR-Verfahren viel zu aufwendig ($\mathcal{O}(n^3)$ Operationen pro Iteration). Statt dessen transformiert man $\mathbf{A}$ zunächst auf obere Hessenberg-Form.

Definition 1.28 Eine Matrix $\mathbf{H} = [h_{i,j}]_{i,j=1}^n$ besitzt **obere Hessenberg-Form**, wenn $h_{i,j} = 0$ für $j < i-1$, das heißt

$$\mathbf{H} = \begin{bmatrix} h_{1,1} & h_{1,2} & \cdots & \cdots & h_{1,n}\\ h_{2,1} & h_{2,2} & & & h_{2,n}\\ 0 & h_{3,2} & h_{3,3} & & h_{3,n}\\ \vdots & \ddots & \ddots & \ddots & \vdots\\ 0 & \dots & 0 & h_{n,n-1} & h_{n,n}\end{bmatrix}.$$

Ziel ist es also, zunächst $\mathbf{A}$ durch Ähnlichkeitstransformationen auf obere Hessenberg-Form zu bringen. Dazu gehen wir wie in Abschnitt 1.4 bei der QR-Zerlegung vor: Wähle Householder-Spiegelungen $\mathbf{P}_1, \mathbf{P}_2, \dots, \mathbf{P}_{n-2}$ und bilde

$$\mathbf{A} \mapsto \mathbf{P}_{n-2}\mathbf{P}_{n-3}\cdots\mathbf{P}_1\mathbf{A}\mathbf{P}_1\mathbf{P}_2\cdots\mathbf{P}_{n-2}$$

gemäß dem folgenden Schema:

$$
\mathbf{A} = \begin{bmatrix} \times & \times & \times & \times & \times \\ \times & \times & \times & \times & \times \\ \times & \times & \times & \times & \times \\ \times & \times & \times & \times & \times \\ \times & \times & \times & \times & \times \end{bmatrix}
\xrightarrow[\mathbf{P}_1\cdot]{}
\left[\begin{array}{ccccc} \times & \times & \times & \times & \times \\ \hline + & + & + & + & + \\ 0 & + & + & + & + \\ 0 & + & + & + & + \\ 0 & + & + & + & + \end{array}\right]
\xrightarrow[\cdot\mathbf{P}_1]{}
\left[\begin{array}{c|cccc} \times & + & + & + & + \\ \times & + & + & + & + \\ 0 & + & + & + & + \\ 0 & + & + & + & + \\ 0 & + & + & + & + \end{array}\right]
$$

$$
\xrightarrow[\mathbf{P}_2\cdot]{}
\left[\begin{array}{ccccc} \times & \times & \times & \times & \times \\ \times & \times & \times & \times & \times \\ \hline 0 & + & + & + & + \\ 0 & 0 & + & + & + \\ 0 & 0 & + & + & + \end{array}\right]
\xrightarrow[\cdot\mathbf{P}_2]{}
\left[\begin{array}{cc|ccc} \times & \times & + & + & + \\ \times & \times & + & + & + \\ 0 & \times & + & + & + \\ 0 & 0 & + & + & + \\ 0 & 0 & + & + & + \end{array}\right]
$$

$$
\xrightarrow[\mathbf{P}_3\cdot]{}
\left[\begin{array}{ccccc} \times & \times & \times & \times & \times \\ \times & \times & \times & \times & \times \\ 0 & \times & \times & \times & \times \\ \hline 0 & 0 & + & + & + \\ 0 & 0 & 0 & + & + \end{array}\right]
\xrightarrow[\cdot\mathbf{P}_3]{}
\left[\begin{array}{ccc|cc} \times & \times & \times & + & + \\ \times & \times & \times & + & + \\ 0 & \times & \times & + & + \\ 0 & 0 & \times & + & + \\ 0 & 0 & 0 & + & + \end{array}\right]
$$

Allgemein gilt im i-ten Schritt

$$
\mathbf{A}_i = \left[\begin{array}{c|c|c} \mathbf{H}_i & & \star \\ \hline \mathbf{0} & \mathbf{c} & \star \end{array}\right],
$$

wobei $\mathbf{H}_i \in \mathbb{R}^{i\times i}$ eine obere Hessenberg-Matrix und $\mathbf{c} \in \mathbb{R}^{n-i}$ ein Vektor ist. Die Householder-Spiegelung $\mathbf{P}_i$ ist nun so zu wählen, dass der Vektor $\mathbf{c}$ auf $\sigma\mathbf{e}_{i+1}$ abgebildet wird, dies bedeutet

$$
\mathbf{P}_i\mathbf{A}_i = \left[\begin{array}{c|c} \mathbf{I} & \mathbf{0} \\ \hline \mathbf{0} & \mathbf{I} - \beta\mathbf{v}\mathbf{v}^\star \end{array}\right] \left[\begin{array}{c|c|c} \mathbf{H}_i & & \star \\ \hline \mathbf{0} & \mathbf{c} & \star \end{array}\right] = \left[\begin{array}{c|c|c} \mathbf{H}_i & & \star \\ \hline \mathbf{0} & \sigma\mathbf{e}_{i+1} & \star \end{array}\right].
$$

Der Aufwand zur Reduktion einer Matrix $\mathbf{A} \in \mathbb{R}^{n\times n}$ auf obere Hessenberg-Form beträgt

$$
\sum_{i=1}^{n-1} 2(i^2 + in) \approx \frac{2}{3}n^3 + n^3 = \mathcal{O}(n^3).
$$

QR-Zerlegung einer Hessenberg-Matrix: Für Hessenberg-Matrizen lässt sich die QR-Zerlegung besonders effizient mit sogenannten *Givens-Rotationen* berechnen.

Definition 1.29 Eine Matrix $\mathbf{G} = \mathbf{G}(i, j, \theta)$ mit

$$\mathbf{G} = \begin{bmatrix} 1 & & & & & & & & & \\ & \ddots & & & & & & & & \\ & & 1 & & & & & & & \\ & & & c & & & & s & & \leftarrow i \\ & & & & 1 & & & & & \\ & & & & & \ddots & & & & \\ & & & & & & 1 & & & \\ & & & -s & & & & c & & \leftarrow j \\ & & & & & & & & 1 & \\ & & & & & & & & & \ddots \\ & & & & & & & & & 1 \end{bmatrix}$$

mit

$$c = \cos(\theta), \qquad s = \sin(\theta)$$

heißt **Givens-Rotation**.

Givens-Rotationen sind orthogonale Matrizen (alle Zeilen sind paarweise zueinander senkrecht) und die Operation $\mathbf{G}(i, j, \theta)\mathbf{A}$ ersetzt die Zeilen i und j von $\mathbf{A}$ durch Linearkombinationen

$$c[a_{i,1}, a_{i,2}, \ldots, a_{i,n}] + s[a_{j,1}, a_{j,2}, \ldots, a_{j,n}]$$

beziehungsweise

$$-s[a_{i,1}, a_{i,2}, \ldots, a_{i,n}] + c[a_{j,1}, a_{j,2}, \ldots, a_{j,n}],$$

während $\mathbf{AG}(i, j, \theta)$ die Spalten i und j von $\mathbf{A}$ durch entsprechende Linearkombinationen ersetzt.

Man kann daher eine Givens-Rotation so wählen, dass *ein* Element von $\mathbf{A}$ zu 0 transformiert wird. Soll etwa $a_{j,k}$ zu 0 gesetzt werden, dann liefert der Ansatz

$$-sa_{i,k} + ca_{j,k} \stackrel{!}{=} 0$$

die Lösung

$$c = \frac{a_{i,k}}{\sqrt{a_{i,k}^2 + a_{j,k}^2}}, \qquad s = \frac{a_{j,k}}{\sqrt{a_{i,k}^2 + a_{j,k}^2}}.$$

Numerisch stabiler ist es im Fall $|a_{i,k}| \geq |a_{j,k}|$ die Rechnung

$$t = a_{j,k}/a_{i,k}, \qquad c = \frac{1}{\sqrt{1+t^2}}, \qquad s = \frac{t}{\sqrt{1+t^2}}$$

und entsprechend im Fall $|a_{i,k}| < |a_{j,k}|$

$$t = a_{i,k}/a_{j,k}, \qquad c = \frac{t}{\sqrt{1+t^2}}, \qquad s = \frac{1}{\sqrt{1+t^2}}.$$

Die QR-Zerlegung einer Hessenberg-Matrix $\mathbf{A}$ erhalten wir nun durch sukzessives Anwenden der Givens-Rotationen $\mathbf{G}(i, i+1, \theta_i)$, $i = 1, 2, \ldots, n-1$, um jeweils das $(i+1, i)$-Element zu Null zu machen:

$$\mathbf{R} = \mathbf{G}(n-1, n, \theta_{n-1})\mathbf{G}(n-2, n-1, \theta_{n-2}) \ldots \mathbf{G}(1, 2, \theta_1)\mathbf{A}.$$

gemäß dem Schema:

$$\begin{bmatrix} \times & \times & \times & \times & \times \\ \times & \times & \times & \times & \times \\ 0 & \times & \times & \times & \times \\ 0 & 0 & \times & \times & \times \\ 0 & 0 & 0 & \times & \times \end{bmatrix} \xrightarrow[i=1]{} \left[\begin{array}{ccccc} + & + & + & + & + \\ 0 & + & + & + & + \\ \hline 0 & \times & \times & \times & \times \\ 0 & 0 & \times & \times & \times \\ 0 & 0 & 0 & \times & \times \end{array}\right] \xrightarrow[i=2]{} \left[\begin{array}{ccccc} \times & \times & \times & \times & \times \\ 0 & + & + & + & + \\ 0 & 0 & + & + & + \\ \hline 0 & 0 & \times & \times & \times \\ 0 & 0 & 0 & \times & \times \end{array}\right]$$

$$\xrightarrow[i=3]{} \left[\begin{array}{ccccc} \times & \times & \times & \times & \times \\ 0 & \times & \times & \times & \times \\ \hline 0 & 0 & + & + & + \\ 0 & 0 & 0 & + & + \\ \hline 0 & 0 & 0 & \times & \times \end{array}\right] \xrightarrow[i=4]{} \left[\begin{array}{ccccc} \times & \times & \times & \times & \times \\ 0 & \times & \times & \times & \times \\ 0 & 0 & \times & \times & \times \\ \hline 0 & 0 & 0 & + & + \\ 0 & 0 & 0 & 0 & + \end{array}\right]$$

Im Teilschritt (1.23) muss das Produkt der Givens-Rotationen von rechts an $\mathbf{R}_k$ heranmultipliziert werden. Wegen

$$\mathbf{Q}_k = \big(\mathbf{G}(n-1, n, \theta_{n-1})\mathbf{G}(n-2, n-1, \theta_{n-2}) \cdots \mathbf{G}(1, 2, \theta_1)\big)^\mathsf{T}$$

ergibt sich

$$\mathbf{A}_{k+1} = \mathbf{R}_k\mathbf{G}(1, 2, \theta_1)^\mathsf{T}\mathbf{G}(2, 3, \theta_2)^\mathsf{T} \cdots \mathbf{G}(n-1, n, \theta_{n-1})^\mathsf{T}.$$

Die Multiplikation von rechts mit $\mathbf{G}(i, i+1, \theta_i)$ verändert die Spalten i und $i+1$. Offensichtlich besitzt $\mathbf{A}_{k+1}$ dann wieder Hessenberg-Gestalt.

Als Aufwand für die beiden Teilschritte (1.22) und (1.23) ergeben sich

$$2\sum_{i=1}^{n-1} 4(n-i+1) = 2\sum_{i=2}^{n} 4i \sim 4n^2 = \mathcal{O}(n^2)$$

Multiplikationen.

Shifts und Deflation: Liegen zwei Eigenwerte λ_i und λ_{i+1} betragsmäßig dicht beieinander, so konvergiert das QR-Verfahren nur sehr langsam. Dies kann mit Hilfe einer *Shift-Strategie* verbessert werden. Im Prinzip versucht man, die beiden Eigenwerte dichter an die Null zu schieben und so den Quotienten $|\lambda_i/\lambda_{i+1}|$ zu vergrößern. Dazu verwendet man für jeden Iterationsschritt k einen Shift-Parameter μ_k und definiert die Folge $\{\mathbf{A}_k\}_{k\geq 0}$ gemäß dem folgenden Algorithmus:

Algorithmus 1.30 (QR-Verfahren mit Shift)
input: Matrix $\mathbf{A} \in \mathbb{K}^{n\times n}$
output: Folge von Iterierten $\{\mathbf{A}_k\}_{k\geq 0}$

① setze $\mathbf{A}_0 := \mathbf{A}$ und $k := 0$

② berechne die QR-Zerlegung

$$\mathbf{A}_k - \mu_k \mathbf{I} = \mathbf{Q}_k \mathbf{R}_k$$

und setze

$$\mathbf{A}_{k+1} := \mathbf{R}_k \mathbf{Q}_k + \mu_k \mathbf{I}$$

③ erhöhe $k := k + 1$ und gehe nach ②

Für das QR-Verfahren mit Shift bleiben die ersten beiden Aussagen von Lemma 1.26 erhalten, während die dritte nun lautet

$$\prod_{\ell=0}^{k} (\mathbf{A} - \mu_\ell \mathbf{I}) = (\mathbf{Q}_0 \mathbf{Q}_1 \cdots \mathbf{Q}_k)(\mathbf{R}_k \mathbf{R}_{k-1} \cdots \mathbf{R}_0).$$

Ferner ist die für $i > j$ in (1.27) beschriebene Konvergenzgeschwindigkeit nun durch den Faktor

$$\mathcal{O}\left(\left| \frac{\lambda_i - \mu_0}{\lambda_j - \mu_0} \right| \cdot \left| \frac{\lambda_i - \mu_1}{\lambda_j - \mu_1} \right| \cdots \left| \frac{\lambda_i - \mu_{k-1}}{\lambda_j - \mu_{k-1}} \right| \right)$$

bestimmt.

In der Praxis bietet sich die Wahl $\mu_k = a_{n,n}^{(k)}$ zur Konvergenzbeschleunigung an. Als noch erfolgreicher hat sich erwiesen, denjenigen Eigenwert von

$$\begin{bmatrix} a_{n-1,n-1}^{(k)} & a_{n-1,n}^{(k)} \\ a_{n,n-1}^{(k)} & a_{n,n}^{(k)} \end{bmatrix}$$

als Shift-Parameter zu wählen, der am nächsten an $a_{n,n}^{(k)}$ liegt. In beiden Fällen konvergiert $a_{n,n}^{(k)}$ sehr schnell gegen den exakten Eigenwert und $a_{n,n-1}^{(k)}$ gegen Null, dies bedeutet

$$\mathbf{A}_k \to \left[\begin{array}{ccccc} \star & \star & \dots & \star & \star \\ \star & \star & \dots & \star & \star \\ & \ddots & \ddots & \vdots & \vdots \\ & & \star & \star & \star \\ \hline 0 & \dots & 0 & 0 & \lambda_n \end{array}\right] = \left[\begin{array}{ccccc} & & & & \star \\ & & \mathbf{B} & & \vdots \\ & & & & \star \\ \hline 0 & \dots & 0 & & \lambda_n \end{array}\right].$$

Ab diesem Zeitpunkt reicht es aus, nur noch das kleinere Teilproblem mit der Hessenberg-Matrix $\mathbf{B}$ zu betrachten. Diese Reduktion wird *Deflation* genannt.

Das QR-Verfahren mit Shift konvergiert in der Regel quadratisch, im Falle symmetrischer Matrizen sogar kubisch. Insgesamt benötigt man daher $\mathcal{O}(n)$ Iterationen, so dass der Gesamtaufaufwand des QR-Verfahrens kubisch ist.

Eigenvektoren: Gemäß Lemma 1.26 liefert das QR-Verfahren asymptotisch die Faktorisierung

$$\mathbf{A} = \mathbf{Q}^\mathsf{T}\mathbf{R}\mathbf{Q}, \quad \mathbf{Q} \text{ orthogonal}, \quad \mathbf{R} \text{ obere Dreiecksmatrix},$$

wobei die Diagonaleinträge $r_{i,i}$ der Marix $\mathbf{R}$ genau den Eigenwerten λ_i entsprechen. Sind $\mathbf{v}_i$ die Eigenvektoren von $\mathbf{R}$, so ergeben sich die entsprechenden Eigenvektoren von $\mathbf{A}$ gemäß $\mathbf{Q}^\mathsf{T}\mathbf{v}_i$. Dabei bestimmt man die Eigenvektoren $\mathbf{v}_i$ dadurch, dass man $v_i^{(i)} = 1$ und $v_j^{(i)} = 0$ für $j > i$ setzt, und $v_j^{(i)}$ für $j < i$ durch Rückwärtssubstitution aus dem Gleichungssystem $(\mathbf{R} - \lambda_i\mathbf{I})\mathbf{v}_i = \mathbf{0}$ bildet.

1.7 Lanczos-Verfahren

Ist $\mathbf{A} \in \mathbb{R}^{n\times n}$ eine symmetrische Matrix, dann sind die extremalen Eigenwerte genau die Extremalwerte des Rayleigh-Quotienten

$$\lambda_{\max} = \max_{\mathbf{x}\in\mathbb{R}^n\setminus\{\mathbf{0}\}} \frac{\mathbf{x}^\mathsf{T}\mathbf{A}\mathbf{x}}{\mathbf{x}^\mathsf{T}\mathbf{x}}, \qquad \lambda_{\min} = \min_{\mathbf{x}\in\mathbb{R}^n\setminus\{\mathbf{0}\}} \frac{\mathbf{x}^\mathsf{T}\mathbf{A}\mathbf{x}}{\mathbf{x}^\mathsf{T}\mathbf{x}}.$$

Anstelle einer Maximierung beziehungsweise Minimierung des Rayleigh-Quotienten über dem ganzen $\mathbb{R}^n$ wollen wir diese nur über dem *Krylov-Raum*

$$\mathcal{K}_k(\mathbf{A}, \mathbf{z}) = \operatorname{span}\{\mathbf{z}, \mathbf{A}\mathbf{z}, \dots, \mathbf{A}^{k-1}\mathbf{z}\}$$

durchführen. Dazu wählen wir eine Orthonormalbasis $\{\mathbf{w}_1, \mathbf{w}_2, \dots, \mathbf{w}_k\}$ von $\mathcal{K}_k(\mathbf{A}, \mathbf{z})$ und setzen

$$\mathbf{W}_k = \left[\begin{array}{c|c|c|c} \mathbf{w}_1 & \mathbf{w}_2 & \dots & \mathbf{w}_k \end{array} \right] \in \mathbb{R}^{n \times k}.$$

Dann gilt

$$\begin{aligned} \lambda_{\max} \geq \mu_{\max}^{(k)} &= \max_{\mathbf{x} \in \mathcal{K}_k(\mathbf{A},\mathbf{z}) \setminus \{\mathbf{0}\}} \frac{\mathbf{x}^\mathsf{T} \mathbf{A} \mathbf{x}}{\mathbf{x}^\mathsf{T} \mathbf{x}} \\ &= \max_{\mathbf{y} \in \mathbb{R}^k \setminus \{\mathbf{0}\}} \frac{\mathbf{y}^\mathsf{T} \mathbf{W}_k^\mathsf{T} \mathbf{A} \mathbf{W}_k \mathbf{y}}{\mathbf{y}^\mathsf{T} \underbrace{\mathbf{W}_k^\mathsf{T} \mathbf{W}_k}_{=\mathbf{I}} \mathbf{y}} = \max_{\mathbf{y} \in \mathbb{R}^k \setminus \{\mathbf{0}\}} \frac{\mathbf{y}^\mathsf{T} \mathbf{W}_k^\mathsf{T} \mathbf{A} \mathbf{W}_k \mathbf{y}}{\mathbf{y}^\mathsf{T} \mathbf{y}} \end{aligned}$$

und analog

$$\lambda_{\min} \leq \mu_{\min}^{(k)} = \min_{\mathbf{y} \in \mathbb{R}^k \setminus \{\mathbf{0}\}} \frac{\mathbf{y}^\mathsf{T} \mathbf{W}_k^\mathsf{T} \mathbf{A} \mathbf{W}_k \mathbf{y}}{\mathbf{y}^\mathsf{T} \mathbf{y}}.$$

Dabei sind $\mu_{\max}^{(k)}$ beziehungsweise $\mu_{\min}^{(k)}$ genau die extremalen Eigenwerte der $(k \times k)$-Matrix $\mathbf{W}_k^\mathsf{T} \mathbf{A} \mathbf{W}_k$.

Satz 1.31 *Sei $\mathbf{A} \in \mathbb{R}^{n \times n}$ eine symmetrische Matrix mit absteigend sortierten Eigenwerten $\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_n$ und $\mathbf{v}_1, \mathbf{v}_2, \dots, \mathbf{v}_n$ die zugehörigen orthonormalen Eigenvektoren. Seien weiter $\mu_1^{(k)} \geq \mu_2^{(k)} \geq \dots \geq \mu_k^{(k)}$ die Eigenwerte von $\mathbf{W}_k^\mathsf{T} \mathbf{A} \mathbf{W}_k$. Dann gilt*

$$\lambda_1 \geq \mu_1^{(k)} \geq \lambda_1 - \frac{(\lambda_1 - \lambda_n) \tan^2(\phi_1)}{T_{k-1}^2(1 + 2\rho_1)},$$

wobei $T_{k-1} \in \Pi_{k-1}$ das k-te Tschebyscheff-Polynom bezeichnet und

$$\rho_1 = \frac{\lambda_1 - \lambda_2}{\lambda_2 - \lambda_n}, \quad \cos(\phi_1) = \frac{|\mathbf{v}_1^\mathsf{T} \mathbf{z}|}{\|\mathbf{z}\|_2}$$

gilt.

Beweis. Sei ohne Beschränkung der Allgemeinheit $\|\mathbf{z}\|_2 = 1$. Da

$$\mathcal{K}_k(\mathbf{A}, \mathbf{z}) = \operatorname{span}\{p(\mathbf{A})\mathbf{z} \,:\, p \in \Pi_{k-1}\}$$

gilt

$$\mu_1^{(k)} = \max_{\mathbf{x} \in \mathcal{K}_k(\mathbf{A},\mathbf{z}) \setminus \{\mathbf{0}\}} \frac{\mathbf{x}^\mathsf{T} \mathbf{A} \mathbf{x}}{\mathbf{x}^\mathsf{T} \mathbf{x}} = \max_{0 \neq p \in \Pi_{k-1}} \frac{\big(p(\mathbf{A})\mathbf{z}\big)^\mathsf{T} \mathbf{A} p(\mathbf{A})\mathbf{z}}{\big(p(\mathbf{A})\mathbf{z}\big)^\mathsf{T} p(\mathbf{A})\mathbf{z}}.$$

Stellen wir $\mathbf{z}$ bezüglich der Orthonormalbasis $\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_n$ dar,

$$\mathbf{z} = \sum_{i=1}^{n} (\mathbf{v}_i^\mathsf{T}\mathbf{z})\mathbf{v}_i = \sum_{i=1}^{n} \xi_i \mathbf{v}_i,$$

so folgt

$$\big(p(\mathbf{A})\mathbf{z}\big)^\mathsf{T}\mathbf{A}p(\mathbf{A})\mathbf{z} = \sum_{i=1}^{n} |\xi_i|^2 p^2(\lambda_i)\lambda_i, \qquad \big(p(\mathbf{A})\mathbf{z}\big)^\mathsf{T}p(\mathbf{A})\mathbf{z} = \sum_{i=1}^{n} |\xi_i|^2 p^2(\lambda_i).$$

Wir erhalten

$$\begin{aligned}\max_{0\neq p\in\Pi_{k-1}} \frac{\big(p(\mathbf{A})\mathbf{z}\big)^\mathsf{T}\mathbf{A}p(\mathbf{A})\mathbf{z}}{\big(p(\mathbf{A})\mathbf{z}\big)^\mathsf{T}p(\mathbf{A})\mathbf{z}} &= \lambda_1 + \max_{0\neq p\in\Pi_{k-1}} \frac{\sum_{i=2}^{n} |\xi_i|^2 p^2(\lambda_i)(\lambda_i-\lambda_1)}{\sum_{i=1}^{n} |\xi_i|^2 p^2(\lambda_i)} \\ &\geq \lambda_1 + (\lambda_n-\lambda_1) \min_{0\neq p\in\Pi_{k-1}} \frac{\sum_{i=2}^{n} |\xi_i|^2 p^2(\lambda_i)}{|\xi_1|^2 p^2(\lambda_1) + \sum_{i=2}^{n} |\xi_i|^2 p^2(\lambda_i)}.\end{aligned}$$

Um eine möglichst scharfe Abschätzung zu erhalten, müssen wir ein Polynom $p \in \Pi_{k-1}$ einsetzen, das innerhalb des Intervalls $[\lambda_n, \lambda_2]$ möglichst klein ist. Wir wählen das transformierte Tschebyscheff-Polynom

$$p(\lambda) := T_{k-1}\left(1 + 2\frac{\lambda-\lambda_2}{\lambda_2-\lambda_n}\right)$$

mit der Eigenschaft $|p(\lambda_i)| \leq 1$ für $i = 2, 3, \ldots, n$. Damit gilt dann wegen

$$\sum_{i=1}^{n} |\xi_i|^2 = \|\mathbf{z}\|_2^2 = 1,$$

dass

$$\mu_1^{(k)} \geq \lambda_1 + (\lambda_n-\lambda_1)\frac{1-|\xi_1|^2}{|\xi_1|^2 T_{k-1}^2(1+2\rho_1)}$$

und die Behauptung folgt aus der Tatsache, dass

$$\frac{1-|\xi_1|^2}{|\xi_1|^2} = \frac{1-\cos^2(\phi_1)}{\cos^2(\phi_1)} = \tan^2(\phi_1). \qquad \spadesuit$$

Ein analoges Resultat erhalten wir für den kleinsten Eigenwert, indem wir Satz 1.31 auf $-\mathbf{A}$ anwenden.

Korollar 1.32 *Unter den Voraussetzungen von Satz 1.31 gilt*

$$\lambda_n \le \mu_k^{(k)} \le \lambda_n - \frac{(\lambda_1 - \lambda_n)\tan^2(\phi_n)}{T_{k-1}^2(1+2\rho_n)}$$

mit $\rho_n = (\lambda_{n-1} - \lambda_n)/(\lambda_1 - \lambda_{n-1})$ *und* $\cos(\phi_n) = |\mathbf{v}_n^\mathsf{T}\mathbf{z}|/\|\mathbf{z}\|_2$.

Bemerkungen

1. Da der Krylov-Raum $\mathcal{K}_k(\mathbf{A},\mathbf{z})$ den Vektor $\mathbf{A}^{k-1}\mathbf{z}$ enthält, ist durch den betragsgrößten Eigenwert von $\mathbf{W}_k^\mathsf{T}\mathbf{A}\mathbf{W}_k$ eine bessere Näherung an den betragsgrößten Eigenwert von $\mathbf{A}$ gegeben als durch den entsprechenden Rayleigh-Quotienten des Vektors $\mathbf{z}_{k-1} = \mathbf{A}^{k-1}\mathbf{z}_0/\|\mathbf{A}^{k-1}\mathbf{z}_0\|_2$ der Potenzmethode, siehe (1.12).

2. Auch die anderen Eigenwerte von $\mathbf{W}_k^\mathsf{T}\mathbf{A}\mathbf{W}_k$ können als Näherungen an die Eigenwerte von $\mathbf{A}$ herangezogen werden: Mit wachsendem k fällt der j-kleinste Eigenwert von $\mathbf{W}_k^\mathsf{T}\mathbf{A}\mathbf{W}_k$ monoton von oben gegen den j-kleinsten Eigenwert von $\mathbf{A}$, während der j-größte Eigenwert von $\mathbf{W}_k^\mathsf{T}\mathbf{A}\mathbf{W}_k$ monoton von unten gegen den j-größten Eigenwert von $\mathbf{A}$ wächst.

3. Ist $\mathbf{u}_1$ der Eigenvektor von $\mathbf{W}_k^\mathsf{T}\mathbf{A}\mathbf{W}_k$ zum Eigenwert $\mu_1^{(k)}$, dann ist gemäß Konstruktion $\mathbf{W}_k\mathbf{u}_1$ eine Näherung an den Eigenvektor zum größten Eigenwert λ_1 von $\mathbf{A}$. Entsprechendes gilt für die anderen Eigenpaare. ♦

Die Orthonormalbasis von $\mathcal{K}_k(\mathbf{A},\mathbf{z})$ bildet man billig mit dem *Lanzcos-Prozess*:

Algorithmus 1.33 (Lanczos-Prozess)
input: Matrix $\mathbf{A} \in \mathbb{R}^{n\times n}$ und Startvektor $\mathbf{z} \in \mathbb{R}^n$
output: Orthonormalbasis $\{\mathbf{w}_k\}_{k\ge 1}$

① Initialisierung: setze $\mathbf{w}_0 := \mathbf{0}$, $\mathbf{u}_1 := \mathbf{z}$ und $k := 1$

② berechne

$$\beta_{k-1} := \|\mathbf{u}_k\|_2 \tag{1.28}$$

$$\mathbf{w}_k := \frac{\mathbf{u}_k}{\beta_{k-1}}, \text{ falls } \beta_{k-1} \neq 0 \tag{1.29}$$

$$\alpha_k := \mathbf{w}_k^\mathsf{T}\mathbf{A}\mathbf{w}_k \tag{1.30}$$

$$\mathbf{u}_{k+1} := (\mathbf{A} - \alpha_k\mathbf{I})\mathbf{w}_k - \beta_{k-1}\mathbf{w}_{k-1} \tag{1.31}$$

③ erhöhe $k := k+1$ und gehe nach ②

Satz 1.34 *Die durch Algorithmus 1.33 gebildete Folge $\{\mathbf{w}_i\}_{i=1}^k$ ist eine Orthonormalbasis des Krylov-Raums $\mathcal{K}_k(\mathbf{A},\mathbf{z})$.*

Beweis. Wir beweisen die Aussage mittels vollständiger Induktion über k. Für $k=1$ ist die Aussage offensichtlich. Sei also $\{\mathbf{w}_i\}_{i=1}^k$ ist eine Orthonormalbasis des Krylov-Raums $\mathcal{K}_k(\mathbf{A},\mathbf{z})$. Dann folgt aus Algorithmus 1.33, dass

$$\mathbf{u}_{k+1} = \mathbf{A}\mathbf{w}_k + \mathbf{p}_k \quad \text{mit} \quad \mathbf{p}_k \in \mathcal{K}_k(\mathbf{A},\mathbf{z}) \quad \text{und} \quad \mathbf{A}\mathbf{w}_k \in \mathcal{K}_{k+1}(\mathbf{A},\mathbf{z}). \tag{1.32}$$

Folglich ist $\mathbf{w}_{k+1} \in \mathcal{K}_{k+1}(\mathbf{A},\mathbf{z})$ mit $\|\mathbf{w}_{k+1}\|_2 = 1$.

Wir zeigen nun, dass gilt $\mathbf{w}_{k+1} \perp \mathcal{K}_k(\mathbf{A},\mathbf{z})$. Für $1 \le i < k-1$ ergibt sich

$$\mathbf{w}_i^\mathsf{T}\mathbf{u}_{k+1} \overset{(1.31)}{=} \mathbf{w}_i^\mathsf{T}(\mathbf{A}-\alpha_k\mathbf{I})\mathbf{w}_k - \beta_{k-1}\underbrace{\mathbf{w}_i^\mathsf{T}\mathbf{w}_{k-1}}_{=0} = \underbrace{\big((\mathbf{A}-\alpha_k\mathbf{I})\mathbf{w}_i\big)^\mathsf{T}}_{\in\mathcal{K}_{i+1}(\mathbf{A},\mathbf{z})=\operatorname{span}\{\mathbf{w}_1,\ldots,\mathbf{w}_{i+1}\}}\mathbf{w}_k = 0.$$

Ferner gilt

$$\begin{aligned}\mathbf{w}_{k-1}^\mathsf{T}\mathbf{u}_{k+1} &\overset{(1.31)}{=} \mathbf{w}_{k-1}^\mathsf{T}\mathbf{A}\mathbf{w}_k - \alpha_k\underbrace{\mathbf{w}_{k-1}^\mathsf{T}\mathbf{w}_k}_{=0} - \beta_{k-1}\underbrace{\mathbf{w}_{k-1}^\mathsf{T}\mathbf{w}_{k-1}}_{=1} \\ &\overset{(1.32)}{=} (\mathbf{u}_k - \underbrace{\mathbf{p}_{k-1}}_{\in\mathcal{K}_{k-1}(\mathbf{A},\mathbf{z})\perp\mathbf{w}_k})^\mathsf{T}\mathbf{w}_k - \beta_{k-1} = \mathbf{u}_k^\mathsf{T}\mathbf{w}_k - \beta_{k-1} \overset{(1.28)}{=} 0,\end{aligned}$$

sowie

$$\mathbf{w}_k^\mathsf{T}\mathbf{u}_{k+1} \overset{(1.31)}{=} \mathbf{w}_k^\mathsf{T}(\mathbf{A}-\alpha_k\mathbf{I})\mathbf{w}_k - \beta_{k-1}\underbrace{\mathbf{w}_k^\mathsf{T}\mathbf{w}_{k-1}}_{=0} = \mathbf{w}_k^\mathsf{T}\mathbf{A}\mathbf{w}_k - \alpha_k \overset{(1.30)}{=} 0.$$

Also ist $\mathbf{u}_{k+1} \perp \mathcal{K}_k(\mathbf{A},\mathbf{z})$, und der Induktionsschritt vollständig $k \mapsto k+1$ bewiesen. ♠

Bemerkung Der Lanczos-Prozess bricht zusammmen, falls sich in (1.29) $\beta_k = 0$ ergibt. Dies ist genau dann der Fall, wenn

$$\mathbf{A}\mathbf{w}_k = \alpha_k\mathbf{w}_k + \beta_{k-1}\mathbf{w}_{k-1} \in \mathcal{K}_k(\mathbf{A},\mathbf{z}),$$

dies bedeutet $\mathcal{K}_{k+1}(\mathbf{A},\mathbf{z}) \subset \mathcal{K}_k(\mathbf{A},\mathbf{z})$. Somit ist $\mathcal{K}_k(\mathbf{A},\mathbf{z})$ ein invarianter Unterraum der Matrix $\mathbf{A}$ und *alle* Eigenwerte von $\mathbf{W}_k^\mathsf{T}\mathbf{A}\mathbf{W}_k$ sind auch Eigenwerte von $\mathbf{A}$. Um weiter Eigenwerte von $\mathbf{A}$ zu bestimmen, muss der Lanczos-Prozess mit einem anderen Startvektor neu gestartet werden. ♦

Der Lanczos-Prozess ist nur unwesentlich teurer als die Potenzmethode, zumal wenn $\mathbf{A}$ deutlich mehr als n Elemente $\neq 0$ enthält. Die Eigenwertberechnung von $\mathbf{W}_k^\mathsf{T}\mathbf{A}\mathbf{W}_k$ benötigt nur $\mathcal{O}(k^2)$ Operationen, da $\mathbf{W}_k^\mathsf{T}\mathbf{A}\mathbf{W}_k$ eine Tridiagonalmatrix ist:

Proposition 1.35 *Es ist*

$$\mathbf{T}_k := \mathbf{W}_k^\mathsf{T}\mathbf{A}\mathbf{W}_k = \begin{bmatrix} \alpha_1 & \beta_1 & & 0 \\ \beta_1 & \alpha_2 & \ddots & \\ & \ddots & \ddots & \beta_{k-1} \\ 0 & & \beta_{k-1} & \alpha_k \end{bmatrix}$$

mit α_i, β_i wie in Algorithmus 1.33.

Beweis. Der (i,j)-Eintrag von $\mathbf{T}_k$ ist $\mathbf{w}_i^\mathsf{T}\mathbf{A}\mathbf{w}_j$. Damit folgen sofort die Diagonalelemente α_i und die Nullelemente, da $\mathbf{A}\mathbf{w}_j \in \mathcal{K}_{j+1}(\mathbf{A},\mathbf{z})$, und damit

$$\mathbf{w}_i^\mathsf{T}\mathbf{A}\mathbf{w}_j = 0 \quad \text{für} \quad i > j+1.$$

Der Fall $i < j-1$ ergibt sich dann automatisch wegen der Symmetrie von $\mathbf{T}_k$. Die Nebendiagonalelemente berechnet man wie folgt:

$$[\mathbf{T}_k]_{i+1,i} = \mathbf{w}_{i+1}^\mathsf{T}\mathbf{A}\mathbf{w}_i \overset{(1.32)}{=} \mathbf{w}_{i+1}^\mathsf{T}(\mathbf{u}_{i+1} - \mathbf{p}_i) = \mathbf{w}_{i+1}^\mathsf{T}\mathbf{u}_{i+1} \overset{(1.29)}{=} \beta_i. \qquad \spadesuit$$

Die Eigenwerte von $\mathbf{T}_k$ lassen sich dann sehr schnell mit dem QR-Verfahren ermitteln.

Ist $\mathbf{A}$ nicht symmetrisch, dann kann man den Lanczos-Prozess nicht verwenden. Stattdessen benötigt man den *Arnoldi-Prozess*, einer stabilen Variante des Gram-Schmidtschen Verfahrens zur Orthogonalisierung des Krylov-Raums $\mathcal{K}_k(\mathbf{A},\mathbf{z})$. Er basiert auf der Hessenberg-Reduktion $\mathbf{Q}^\mathsf{T}\mathbf{A}\mathbf{Q} = \mathbf{H}$. Setzen wir $\mathbf{Q} = [\mathbf{q}_1, \mathbf{q}_2, \ldots, \mathbf{q}_n]$ und vergleichen $\mathbf{A}\mathbf{Q} = \mathbf{Q}\mathbf{H}$, dann folgt

$$\mathbf{A}\mathbf{q}_k = \sum_{i=1}^{k+1} h_{i,k}\mathbf{q}_i, \qquad 1 \le k \le n-1.$$

Auflösen nach dem letzten Term der Summe ergibt

$$h_{k+1,k}\mathbf{q}_{k+1} = \mathbf{A}\mathbf{q}_k - \sum_{i=1}^{k} h_{i,k}\mathbf{q}_i = \mathbf{r}_k,$$

wobei $h_{i,k} = \mathbf{q}_i^\mathsf{T}\mathbf{A}\mathbf{q}_k$ für alle $i = 1, 2, \ldots, k$. Falls $\mathbf{r}_k \neq \mathbf{0}$, dann folgt

$$\mathbf{q}_{k+1} = \frac{1}{h_{k+1,k}}\mathbf{r}_k, \qquad h_{k+1,k} = \|\mathbf{r}_k\|_2.$$

Diese Gleichungen führen auf den Arnoldi-Prozess:

Algorithmus 1.36 (Arnoldi-Prozess)
input: Matrix $\mathbf{A} \in \mathbb{R}^{n\times n}$ und Startvektor $\mathbf{z} \in \mathbb{R}^n$
output: Orthonormalbasis $\{\mathbf{w}_k\}_{k\geq 1}$

① Initialisierung: setze $\mathbf{w}_1 := \mathbf{z}/\|\mathbf{z}\|_2$ und $k := 1$

② berechne

$$\begin{aligned}
\mathbf{r}_k &:= \mathbf{A}\mathbf{w}_k \\
h_{1,k} &:= \mathbf{w}_1^\intercal \mathbf{r}_k, \ldots, h_{k,k} := \mathbf{w}_k^\intercal \mathbf{r}_k \\
\mathbf{r}_k &:= \mathbf{r}_k - \sum_{i=1}^{k} h_{i,k}\mathbf{w}_i \\
h_{k+1,k} &:= \|\mathbf{r}_k\|_2 \\
\mathbf{w}_{k+1} &:= \frac{1}{h_{k+1,k}}\mathbf{r}_k, \text{ falls } h_{k+1,k} \neq 0
\end{aligned}$$

③ erhöhe $k := k + 1$ und gehe nach ②

Anschließend berechnet man mit dem QR-Verfahren die Eigenwerte der oberen Hessenberg-Matrix $\mathbf{W}_k^\intercal \mathbf{A}\mathbf{W}_k$.

2 Lineare Ausgleichsprobleme

2.1 Normalengleichungen revisited

Im folgenden sei $\mathbf{A} \in \mathbb{R}^{m \times n}$ und $\mathbf{b} \in \mathbb{R}^m$. Gesucht ist ein Vektor $\mathbf{x} \in \mathbb{R}^n$ mit

$$\mathbf{A}\mathbf{x} \approx \mathbf{b}.$$

Da wir m Gleichungen für n Unbekannte haben, ist das lineare Gleichungssystem im allgemeinen nicht — oder nicht eindeutig — lösbar. Ist $m > n$, dann nennen wir das lineare Gleichungssystem *überbestimmt*, ist $m < n$, dann nennen wir es *unterbestimmt*. Überbestimmte Probleme treten häufig in den Anwendungen auf, wenn es darum geht, Modellparameter an Messdaten anzupassen.

Da wir für $m \neq n$ die m Gleichungen im allgemeinen nicht alle exakt erfüllen können, suchen wir nun nach Vektoren $\mathbf{x} \in \mathbb{R}^n$, für die das *Residuum*

$$\mathbf{r} = \mathbf{b} - \mathbf{A}\mathbf{x} \tag{2.1}$$

möglichst klein ist.

Definition 2.1 Für eine Matrix $\mathbf{A} \in \mathbb{R}^{m \times n}$ und ein $\mathbf{b} \in \mathbb{R}^m$ heißt das Problem

$$\|\mathbf{b} - \mathbf{A}\mathbf{x}\|_2 \to \min \tag{2.2}$$

ein lineares **Ausgleichsproblem**. Eine Lösung $\mathbf{x} \in \mathbb{R}^n$ des Ausgleichsproblems heißt **Ausgleichslösung** oder **kleinste-Quadrate-Lösung**.

Bemerkung Der Lösungsbegriff in (2.2) ist eine Verallgemeinerung der klassischen Lösung. Ist nämlich $m = n$ und ist $\mathbf{x} \in \mathbb{R}^n$ eine klassische Lösung, das heißt, gilt $\mathbf{A}\mathbf{x} = \mathbf{b}$, dann ist offensichtlich $\mathbf{x}$ ebenfalls eine Lösung von (2.2). ♦

Satz 2.2 *Die Lösungen von* (2.2) *sind genau die Lösungen der* ***Gaußschen Normalengleichungen***

$$\mathbf{A}^\mathsf{T}\mathbf{A}\mathbf{x} = \mathbf{A}^\mathsf{T}\mathbf{b}, \tag{2.3}$$

insbesondere existiert eine Lösung $\mathbf{x}$. *Ist* $\mathbf{z}$ *eine weitere Lösung, so gilt* $\mathbf{A}\mathbf{x} = \mathbf{A}\mathbf{z}$. *Das Residuum* (2.1) *ist eindeutig bestimmt und genügt der Gleichung* $\mathbf{A}^\mathsf{T}\mathbf{r} = \mathbf{0}$.

Beweis. Betrachte das Funktional

$$\phi(\mathbf{x}) = \|\mathbf{b} - \mathbf{A}\mathbf{x}\|_2^2 = \mathbf{b}^\mathsf{T}\mathbf{b} - 2\mathbf{x}^\mathsf{T}\mathbf{A}^\mathsf{T}\mathbf{b} + \mathbf{x}^\mathsf{T}\mathbf{A}^\mathsf{T}\mathbf{A}\mathbf{x} \geq 0.$$

Da das Funktional stetig und nach unten beschränkt ist, gibt es ein Minimum. Dieses Minimum von ϕ erfüllt

$$\nabla\phi(\mathbf{x}) = -2\mathbf{A}^\mathsf{T}\mathbf{b} + 2\mathbf{A}^\mathsf{T}\mathbf{A}\mathbf{x} \stackrel{!}{=} \mathbf{0}.$$

Dies ist genau dann der Fall, wenn $\mathbf{x}$ den Normalengleichungen (2.3) genügt. Insbesondere gilt dann $\mathbf{A}^\mathsf{T}(\mathbf{b} - \mathbf{A}\mathbf{x}) = \mathbf{A}^\mathsf{T}\mathbf{r} = \mathbf{0}$. Außerdem folgt $\phi(\mathbf{z}) = \phi(\mathbf{x})$ für jede weitere Lösung $\mathbf{z}$ der Normalengleichungen. Wegen

$$\phi(\mathbf{z}) = \|(\mathbf{b} - \mathbf{A}\mathbf{x}) + (\mathbf{A}\mathbf{x} - \mathbf{A}\mathbf{z})\|_2^2 = \phi(\mathbf{x}) + 2\underbrace{(\mathbf{b} - \mathbf{A}\mathbf{x})}_{=\mathbf{r}}{}^\mathsf{T}\mathbf{A}(\mathbf{x} - \mathbf{z}) + \|\mathbf{A}\mathbf{x} - \mathbf{A}\mathbf{z}\|_2^2$$

ergibt sich schließlich $\|\mathbf{A}\mathbf{x} - \mathbf{A}\mathbf{z}\|_2^2 = 0$. ♠

Bemerkung Aus $\mathbf{A}^\mathsf{T}\mathbf{r} = \mathbf{0}$ folgt, dass das Residuum senkrecht auf den Spalten von $\mathbf{A}$ steht. Das Residuum $\mathbf{r}$ ist folglich Normale zum von den Spalten der Matrix $\mathbf{A}$ aufgespannten Raum. Daher erklärt sich die Bezeichnung Normalengleichungen. ♦

Satz 2.3 *Die Matrix* $\mathbf{A}^\mathsf{T}\mathbf{A} \in \mathbb{R}^{n\times n}$ *ist symmetrisch und positiv semidefinit. Darüber hinaus ist* $\mathbf{A}^\mathsf{T}\mathbf{A}$ *genau dann positiv definit, wenn der Kern von* $\mathbf{A}$ *trivial ist, das heißt, wenn* $\operatorname{kern}(\mathbf{A}) = \{\mathbf{0}\}$. *Dies ist genau dann der Fall, wenn die Spalten von* $\mathbf{A}$ *linear unabhängig sind.*

Beweis. Offensichtlich ist $\mathbf{A}^\mathsf{T}\mathbf{A}$ symmetrisch und wegen

$$\mathbf{x}^\mathsf{T}\mathbf{A}^\mathsf{T}\mathbf{A}\mathbf{x} = \|\mathbf{A}\mathbf{x}\|_2^2 \geq 0 \quad \text{für alle } \mathbf{x} \in \mathbb{R}^n$$

auch positiv semidefinit. Ist $\operatorname{kern}\mathbf{A} = \{\mathbf{0}\}$, so gilt Gleichheit ("=") nur im Falle $\mathbf{x} = \mathbf{0}$, das heißt, $\mathbf{A}^\mathsf{T}\mathbf{A}$ ist positiv definit. ♠

2.2 Singulärwertzerlegung und Pseudoinverse

Offensichtlich spielt die Matrix $\mathbf{A}^\mathsf{T}\mathbf{A}$ eine große Rolle beim linearen Ausgleichsproblem. Im folgenden seien $\lambda_1, \lambda_2, \dots, \lambda_p$ die von Null verschiedenen Eigenwerte von $\mathbf{A}^\mathsf{T}\mathbf{A}$

$$\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p > \lambda_{p+1} = \cdots = \lambda_n = 0, \quad (p \leq n)$$

und $\mathbf{v}_1, \mathbf{v}_2, \dots, \mathbf{v}_n$ zugehörige orthonormale Eigenvektoren. Bezeichnen wir ferner mit

$$\mathbf{u}_i := \frac{1}{\sqrt{\lambda_i}}\mathbf{A}\mathbf{v}_i, \qquad i = 1, 2 \dots, p, \tag{2.4}$$

so folgt für alle $1 \leq i, j \leq p$ dass

$$\mathbf{u}_i^\mathsf{T}\mathbf{u}_j = \frac{1}{\sqrt{\lambda_i}}\frac{1}{\sqrt{\lambda_j}}(\mathbf{A}\mathbf{v}_i)^\mathsf{T}(\mathbf{A}\mathbf{v}_j) = \frac{1}{\sqrt{\lambda_i}\sqrt{\lambda_j}}\mathbf{v}_i^\mathsf{T}(\mathbf{A}^\mathsf{T}\mathbf{A}\mathbf{v}_j) = \frac{\lambda_j}{\sqrt{\lambda_i}\sqrt{\lambda_j}}\mathbf{v}_i^\mathsf{T}\mathbf{v}_j = \delta_{i,j}.$$

Die Vektoren $\{\mathbf{u}_i\}_{i=1}^p$ bilden folglich eine Orthonormalbasis von $\mathrm{img}(\mathbf{A})$ und können durch weitere $m - p$ Vektoren $\mathbf{u}_{p+1}, \dots, \mathbf{u}_n$ zu einer von $\mathbb{R}^m$ ergänzt werden.

Es gilt wegen (2.4)

$$\mathbf{A}^\mathsf{T}\mathbf{u}_i = \frac{1}{\sqrt{\lambda_i}}\mathbf{A}^\mathsf{T}\mathbf{A}\mathbf{v}_i = \sqrt{\lambda_i}\mathbf{v}_i, \quad i = 1, 2, \dots, p,$$

während weiter

$$\mathbf{A}^\mathsf{T}\mathbf{u}_i = \mathbf{0}, \quad i = p + 1, \dots, m$$

gilt, da $\{\mathbf{u}_{p+1}, \dots, \mathbf{u}_m\} \subset \mathrm{img}(\mathbf{A})^\perp = \mathrm{kern}(\mathbf{A}^\mathsf{T})$.

Wir fassen zusammen:

Satz 2.4 *Zu jeder Matrix $\mathbf{A} \in \mathbb{R}^{m\times n}$ existiert eine* ***Singulärwertzerlegung*** *(SVD = singular value decomposition), das ist ein Tripel $(\{\sigma_i\}_{i=1}^p, \{\mathbf{u}_i\}_{i=1}^m, \{\mathbf{v}_i\}_{i=1}^n)$ mit*

$$\begin{aligned}
&\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_p > 0,\\
&\mathbf{u}_i \in \mathbb{R}^m, \qquad \mathbf{u}_i^\mathsf{T}\mathbf{u}_j = \delta_{i,j}, \qquad i, j = 1, 2, \dots, m,\\
&\mathbf{v}_i \in \mathbb{R}^n, \qquad \mathbf{v}_i^\mathsf{T}\mathbf{v}_j = \delta_{i,j}, \qquad i, j = 1, 2, \dots, n,
\end{aligned}$$

und

$$\begin{aligned} &\mathbf{A}\mathbf{v}_i = \sigma_i \mathbf{u}_i, \qquad &&\mathbf{A}^\mathsf{T}\mathbf{u}_i = \sigma_i \mathbf{v}_i, \qquad &&i = 1, 2, \dots, p, \\ &\mathbf{A}\mathbf{v}_k = \mathbf{0}, \qquad &&\mathbf{A}^\mathsf{T}\mathbf{u}_\ell = \mathbf{0}, \qquad &&k, \ell > p. \end{aligned}$$

Ferner sind σ_i^2 entsprechend ihrer Vielfachheit genau die von Null verschiedenen Eigenwerte von $\mathbf{A}^\mathsf{T}\mathbf{A}$.

In Matrixnotation lässt sich Satz 2.4 kürzer schreiben. Wir setzen

$$\mathbf{U} := \left[\mathbf{u}_1, \mathbf{u}_2, \dots, \mathbf{u}_m\right] \in \mathbb{R}^{m\times m}, \quad \mathbf{V} := \left[\mathbf{v}_1, \mathbf{v}_2, \dots, \mathbf{v}_n\right] \in \mathbb{R}^{n\times n},$$

$$\boldsymbol{\Sigma} := \left[\begin{array}{ccc|c} \sigma_1 & & & \\ & \ddots & & \mathbf{0} \\ & & \sigma_p & \\ \hline & \mathbf{0} & & \mathbf{0} \end{array}\right] \in \mathbb{R}^{m\times n}$$

und erhalten

$$\mathbf{A} = \mathbf{U}\boldsymbol{\Sigma}\mathbf{V}^\mathsf{T}, \quad \mathbf{A}^\mathsf{T} = \mathbf{V}\boldsymbol{\Sigma}^\mathsf{T}\mathbf{U}^\mathsf{T}. \tag{2.5}$$

Dabei sind die Matrizen $\mathbf{U}$ und $\mathbf{V}$ orthogonal.

Alternativ zu (2.5) gelten die Summendarstellungen

$$\mathbf{A} = \sum_{i=1}^{p} \sigma_i \mathbf{u}_i \mathbf{v}_i^\mathsf{T}, \quad \mathbf{A}^\mathsf{T} = \sum_{i=1}^{p} \sigma_i \mathbf{v}_i \mathbf{u}_i^\mathsf{T}.$$

Definition 2.5 Sei $\mathbf{U}\boldsymbol{\Sigma}\mathbf{V}^\mathsf{T}$ die Singulärwertzerlegung von $\mathbf{A}$ und

$$\boldsymbol{\Sigma}^+ := \left[\begin{array}{ccc|c} \sigma_1^{-1} & & & \\ & \ddots & & \mathbf{0} \\ & & \sigma_p^{-1} & \\ \hline & \mathbf{0} & & \mathbf{0} \end{array}\right] \in \mathbb{R}^{n\times m}.$$

Dann heißt die Matrix

$$\mathbf{A}^+ = \mathbf{V}\boldsymbol{\Sigma}^+\mathbf{U}^\mathsf{T} \in \mathbb{R}^{n\times m}$$

Pseudoinverse oder **Moore-Penrose-Inverse** von $\mathbf{A}$.

Auch für die Pseudoinverse gilt eine ensprechende Summendarstellung

$$\mathbf{A}^+ = \sum_{i=1}^{p} \frac{1}{\sigma_i} \mathbf{v}_i \mathbf{u}_i^\intercal, \tag{2.6}$$

aus der sofort folgt

$$\operatorname{kern}(\mathbf{A}^+) = \operatorname{kern}(\mathbf{A}^\intercal) = \operatorname{img}(\mathbf{A})^\perp, \quad \operatorname{img}(\mathbf{A}^+) = \operatorname{img}(\mathbf{A}^\intercal) = \operatorname{kern}(\mathbf{A})^\perp. \tag{2.7}$$

Beispiel 2.6 Für die Matrix

$$\mathbf{A} = \begin{bmatrix} 1 & 1 \\ 1 & 1 \\ 1 & 1 \\ 1 & 1 \end{bmatrix}$$

gilt

$$\operatorname{img}(\mathbf{A}) = \operatorname{span}\left\{[1,1,1,1]^\intercal\right\} \quad \operatorname{kern}(\mathbf{A}) = \operatorname{span}\left\{[-1,1]^\intercal\right\}.$$

Hieraus folgt

$$\operatorname{img}(\mathbf{A}^+) = \operatorname{kern}(\mathbf{A})^\perp = \operatorname{span}\left\{[1,1]^\intercal\right\}$$

und daher

$$\mathbf{A}^+ = \begin{bmatrix} \alpha & \beta & \gamma & \delta \\ \alpha & \beta & \gamma & \delta \end{bmatrix}.$$

Da für alle $\mathbf{u}$ mit $[1,1,1,1]^\intercal \perp \mathbf{u}$ gilt $\mathbf{A}^+\mathbf{u} = \mathbf{0}$ ergibt sich zwangsläufig $\alpha = \beta = \gamma = \delta$, das heißt,

$$\mathbf{A}^+ = \alpha \begin{bmatrix} 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 \end{bmatrix}.$$

Der Parameter α berechnet sich wie folgt: Es ist $p = 1$ und

$$\mathbf{v}_1 = \frac{1}{\sqrt{2}} \begin{bmatrix} 1 \\ 1 \end{bmatrix}, \quad \mathbf{u}_1 = \frac{1}{2} \begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \end{bmatrix}, \quad \mathbf{A}\mathbf{v}_1 = \sqrt{2} \begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \end{bmatrix} = 2\sqrt{2}\mathbf{u}_1,$$

dies bedeutet, $\sigma_1 = 2\sqrt{2}$. Wegen

$$\frac{\alpha}{2}\begin{bmatrix}4\\4\end{bmatrix} = \mathbf{A}^+\mathbf{u}_1 \overset{!}{=} \frac{1}{2\sqrt{2}}\mathbf{v}_1 = \frac{1}{4}\begin{bmatrix}1\\1\end{bmatrix}$$

ergibt sich $\alpha = 1/8$. ♣

Der Name "Pseudoinverse" beruht auf folgendem Resultat:

Satz 2.7 *Die Pseudoinverse $\mathbf{A}^+$ von $\mathbf{A} \in \mathbb{R}^{m\times n}$ ist die eindeutige Lösung der vier Gleichungen*

$$\begin{array}{ll} (i) & \mathbf{AXA} = \mathbf{A} \\ (ii) & \mathbf{XAX} = \mathbf{X} \end{array} \qquad \begin{array}{ll} (iii) & (\mathbf{AX})^\mathsf{T} = \mathbf{AX} \\ (iv) & (\mathbf{XA})^\mathsf{T} = \mathbf{XA} \end{array}$$

Beweis. Wir weisen zunächst nach, dass die Pseudoinverse $\mathbf{X} = \mathbf{A}^+$ alle vier Gleichungen erfüllt. Wegen

$$\boldsymbol{\Sigma}\boldsymbol{\Sigma}^+ = \left[\begin{array}{ccc|c} \sigma_1 & & & \\ & \ddots & & \mathbf{0} \\ & & \sigma_p & \\ \hline & \mathbf{0} & & \mathbf{0} \end{array}\right]\left[\begin{array}{ccc|c} \sigma_1^{-1} & & & \\ & \ddots & & \mathbf{0} \\ & & \sigma_p^{-1} & \\ \hline & \mathbf{0} & & \mathbf{0} \end{array}\right] = \left[\begin{array}{c|c} \mathbf{I} & \mathbf{0} \\ \hline \mathbf{0} & \mathbf{0} \end{array}\right] \in \mathbb{R}^{m\times m} \tag{2.8}$$

folgen die ersten beiden Gleichungen

$$\begin{aligned} \mathbf{AA}^+\mathbf{A} &= \mathbf{U}\boldsymbol{\Sigma}\mathbf{V}^\mathsf{T}\mathbf{V}\boldsymbol{\Sigma}^+\mathbf{U}^\mathsf{T}\mathbf{U}\boldsymbol{\Sigma}\mathbf{V}^\mathsf{T} = \mathbf{U}\boldsymbol{\Sigma}\boldsymbol{\Sigma}^+\boldsymbol{\Sigma}\mathbf{V}^\mathsf{T} = \mathbf{U}\boldsymbol{\Sigma}\mathbf{V}^\mathsf{T} = \mathbf{A}, \\ \mathbf{A}^+\mathbf{AA}^+ &= \mathbf{V}^\mathsf{T}\boldsymbol{\Sigma}^+\mathbf{U}\mathbf{U}^\mathsf{T}\boldsymbol{\Sigma}\mathbf{V}^\mathsf{T}\mathbf{V}\boldsymbol{\Sigma}\mathbf{U}^\mathsf{T} = \mathbf{V}^\mathsf{T}\boldsymbol{\Sigma}^+\boldsymbol{\Sigma}\boldsymbol{\Sigma}^+\mathbf{U}^\mathsf{T} = \mathbf{V}\boldsymbol{\Sigma}^+\mathbf{U}^\mathsf{T} = \mathbf{A}^+. \end{aligned}$$

Weiter ist $\mathbf{AA}^+ = \mathbf{U}\boldsymbol{\Sigma}\boldsymbol{\Sigma}^+\mathbf{U}^\mathsf{T}$ und somit wegen (2.8) symmetrisch. Entsprechend ist auch $\mathbf{A}^+\mathbf{A}$ symmetrisch, womit auch die beiden letzten Gleichungen gezeigt sind.

Es verbleibt noch zu zeigen, dass die vier Gleichungen nur die eine Lösung $\mathbf{X} = \mathbf{A}^+$ haben. Wegen (i) ist

$$\mathbf{0} = \mathbf{AXA}\mathbf{v}_i - \mathbf{A}\mathbf{v}_i = \mathbf{A}(\mathbf{X}\underbrace{\mathbf{A}\mathbf{v}_i}_{=\sigma_i\mathbf{u}_i} - \mathbf{v}_i), \quad i = 1, 2, \ldots, p.$$

Dies bedeutet, dass

$$\mathbf{X}\mathbf{u}_i = \frac{1}{\sigma_i}\mathbf{v}_i + \mathbf{w}_i \quad \text{für ein} \quad \mathbf{w}_i \in \operatorname{kern}(\mathbf{A}).$$

Wegen $\mathbf{w}_i \in \ker(\mathbf{A}) \subset \ker(\mathbf{XA})$ und (iv) folgt für jedes $i = 1, 2, \dots, p$

$$0 = (\mathbf{XAw}_i)^\mathsf{T}\left(\frac{1}{\sigma_i}\mathbf{v}_i\right) = \mathbf{w}_i^\mathsf{T}\mathbf{XA}\left(\frac{1}{\sigma_i}\mathbf{v}_i\right) = \mathbf{w}_i^\mathsf{T}\mathbf{Xu}_i = \mathbf{w}_i^\mathsf{T}\left(\frac{1}{\sigma_i}\mathbf{v}_i + \mathbf{w}_i\right)$$
$$= \underbrace{\frac{1}{\sigma_i}\mathbf{w}_i^\mathsf{T}\mathbf{v}_i}_{\substack{=0,\text{ da}\\ \mathbf{v}_i \perp \operatorname{span}\{\mathbf{v}_{p+1},\dots,\mathbf{v}_n\} = \ker(\mathbf{A})}} + \mathbf{w}_i^\mathsf{T}\mathbf{w}_i = \|\mathbf{w}_i\|^2.$$

Dies bedeutet $\mathbf{w}_i = \mathbf{0}$ und daher

$$\mathbf{Xu}_i = \frac{1}{\sigma_i}\mathbf{v}_i, \quad i = 1, 2, \dots, p. \tag{2.9}$$

Hieraus folgt die Inklusion

$$\begin{aligned}\operatorname{img}(\mathbf{AX}) \supset \operatorname{span}\{\mathbf{AXu}_i : i = 1, 2, \dots, p\} &= \operatorname{span}\{\mathbf{Av}_i : i = 1, 2, \dots, p\}\\ &= \operatorname{span}\{\mathbf{u}_i : i = 1, 2, \dots, p\} = \operatorname{img}(\mathbf{A}).\end{aligned}$$

Da andererseits trivialerweise $\operatorname{img}(\mathbf{AX}) \subset \operatorname{img}(\mathbf{A})$ ist, ergibt dies

$$\operatorname{img}(\mathbf{AX}) = \operatorname{img}(\mathbf{A}).$$

Aus (iii) folgt damit

$$\ker(\mathbf{AX}) = \operatorname{img}(\mathbf{AX})^\perp = \operatorname{img}(\mathbf{A})^\perp = \operatorname{span}\{\mathbf{u}_i : i = p+1, \dots, m\}.$$

Demnach ist

$$\mathbf{AXu}_i = \mathbf{0} \quad \text{bzw.} \quad \mathbf{Xu}_i = \mathbf{w}_i \in \ker(\mathbf{A}), \quad i = p+1, \dots, m.$$

Gleichung (ii) impliziert jedoch $\mathbf{w}_i = \mathbf{0}$, denn

$$\mathbf{w}_i = \mathbf{Xu}_i = \mathbf{XAXu}_i = \mathbf{XAw}_i = \mathbf{0}, \quad i = p+1, \dots, m. \tag{2.10}$$

Ein Vergleich von (2.6) mit (2.9) und (2.10) zeigt, dass $\mathbf{X}$ und $\mathbf{A}^+$ übereinstimmen, also $\mathbf{A}^+$ die einzige Lösung der Gleichungen (i)–(iv) ist. ♠

Bemerkung Ist $\mathbf{A}$ invertierbar, dann ist $\mathbf{A}^+ = \mathbf{A}^{-1}$ wegen Gleichung (i) bzw. (ii). ♦

Den Zusammenhang zwischen Pseudoinverse und linearem Ausgleichsproblem beschreibt der folgende Satz.

Satz 2.8 *Der Vektor* $\mathbf{A}^+\mathbf{b}$ *ist die eindeutige Lösung des linearen Ausgleichsproblems* (2.2) *mit minimaler* $\|\cdot\|_2$*-Norm.*

Beweis. Nach Satz 2.7 (i) ist

$$\mathbf{A}\mathbf{A}^+\mathbf{b} - \mathbf{b} \in \operatorname{kern}(\mathbf{A}^+) \overset{(2.7)}{=} \operatorname{img}(\mathbf{A})^\perp = \operatorname{kern}(\mathbf{A}^\intercal).$$

Also erfüllt $\mathbf{A}^+\mathbf{b}$ die Normalengleichungen (2.3)

$$\mathbf{A}^\intercal\mathbf{A}(\mathbf{A}^+\mathbf{b}) = \mathbf{A}^\intercal\mathbf{b}$$

und ist daher eine Lösung des linearen Ausgleichsproblems.

Ist $\mathbf{z}$ eine zweite Lösung der Normalengleichungen, dann gilt gemäß Satz 2.2

$$\mathbf{w} := \mathbf{A}^+\mathbf{b} - \mathbf{z} \in \operatorname{kern}(\mathbf{A}).$$

Da $\mathbf{A}^+\mathbf{b} \in \operatorname{img}(\mathbf{A}^+) \overset{(2.7)}{=} \operatorname{kern}(\mathbf{A})^\perp$ haben wir $\mathbf{z} = \mathbf{A}^+\mathbf{b} - \mathbf{w}$ orthogonal zerlegt, und nach dem Satz des Pythagoras gilt

$$\|\mathbf{z}\|_2^2 = \|\mathbf{A}^+\mathbf{b}\|_2^2 + \|\mathbf{w}\|_2^2 \geq \|\mathbf{A}^+\mathbf{b}\|_2^2. \qquad \spadesuit$$

Korollar 2.9 *Hat* $\mathbf{A} \in \mathbb{R}^{m\times n}$ *vollen Rang* $\operatorname{rang}(\mathbf{A}) = n \leq m$*, dann gilt*

$$\mathbf{A}^+ = (\mathbf{A}^\intercal\mathbf{A})^{-1}\mathbf{A}^\intercal.$$

Beispiel 2.10 (Fortsetzung von Beispiel 2.6) Zu lösen sei das lineare Ausgleichsproblem $\|\mathbf{A}\mathbf{x} - \mathbf{b}\|_2 \to \min$ mit

$$\mathbf{A} = \begin{bmatrix} 1 & 1 \\ 1 & 1 \\ 1 & 1 \\ 1 & 1 \end{bmatrix}, \quad \mathbf{b} = \begin{bmatrix} 2 \\ 0 \\ 0 \\ -1 \end{bmatrix}.$$

Die Lösung $\mathbf{x}$ mit minimaler Euklidnorm ist

$$\mathbf{x} = \mathbf{A}^+\mathbf{b} = \frac{1}{8}\begin{bmatrix} 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 \end{bmatrix}\begin{bmatrix} 2 \\ 0 \\ 0 \\ -1 \end{bmatrix} = \frac{1}{8}\begin{bmatrix} 1 \\ 1 \end{bmatrix}.$$

Alle anderen Lösungen haben wegen $\operatorname{kern}(\mathbf{A}) = [1, -1]^\intercal$ die Form

$$\mathbf{x} = \frac{1}{8}\begin{bmatrix} 1 \\ 1 \end{bmatrix} + \alpha\begin{bmatrix} 1 \\ -1 \end{bmatrix}, \quad \alpha \in \mathbb{R}. \qquad \clubsuit$$

2.3 CG- und CGLS-Verfahren

Es sei $\mathbf{A} \in \mathbb{R}^{n\times n}$ eine symmetrische, 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 $\mathbf{x}$ eindeutiges Minimum $\phi(\mathbf{x}) = 0$ des Funktionals

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

ist.

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

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

Aus

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

folgt daher

$$\alpha = \frac{\mathbf{d}^\mathsf{T}(\mathbf{b}-\mathbf{Az})}{\mathbf{d}^\mathsf{T}\mathbf{Ad}}. \tag{2.11}$$

Lemma 2.11 *Angenommen die Vektoren $\mathbf{d}_0, \mathbf{d}_1, \ldots, \mathbf{d}_k$ seien* **A-konjugiert**, *das heißt, es gelte $\mathbf{d}_i^\mathsf{T}\mathbf{Ad}_j = 0$ für alle $i \neq j$. Ist*

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

und setzt man

$$\mathbf{x}_{k+1} = \mathbf{x}_k + \alpha_k\mathbf{d}_k, \qquad \alpha_k = \frac{\mathbf{d}_k^\mathsf{T}\mathbf{r}_k}{\mathbf{d}_k^\mathsf{T}\mathbf{Ad}_k}, \qquad \mathbf{r}_k = \mathbf{b}-\mathbf{Ax}_k, \tag{2.12}$$

so folgt

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

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

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

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 (2.11) genau dann der Fall, wenn

$$\alpha_k = \frac{\mathbf{d}_k^\mathsf{T}(\mathbf{b} - \mathbf{A}\mathbf{x}_k)}{\mathbf{d}_k^\mathsf{T}\mathbf{A}\mathbf{d}_k} = \frac{\mathbf{d}_k^\mathsf{T}(\mathbf{b} - \mathbf{A}\mathbf{x}_0)}{\mathbf{d}_k^\mathsf{T}\mathbf{A}\mathbf{d}_k}. \tag{2.13}$$

♠

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

$$\mathbf{r}_0 = \mathbf{b} - \mathbf{A}\mathbf{x}_0, \qquad \mathbf{r}_{k+1} = \mathbf{b} - \mathbf{A}\mathbf{x}_{k+1} \overset{(2.12)}{=} \mathbf{r}_k - \alpha_k\mathbf{A}\mathbf{d}_k, \quad k \ge 0, \tag{2.14}$$

erfüllt dann für alle $\ell < k$

$$\begin{aligned}\mathbf{d}_\ell^\mathsf{T}\mathbf{r}_k = \mathbf{d}_\ell^\mathsf{T}(\mathbf{b} - \mathbf{A}\mathbf{x}_k) &= \mathbf{d}_\ell^\mathsf{T}\left(\mathbf{b} - \mathbf{A}\mathbf{x}_0 - \sum_{i=0}^{k-1}\alpha_i\mathbf{A}\mathbf{d}_i\right)\\ &= \mathbf{d}_\ell^\mathsf{T}(\mathbf{b} - \mathbf{A}\mathbf{x}_0) - \alpha_\ell\mathbf{d}_\ell^\mathsf{T}\mathbf{A}\mathbf{d}_\ell \overset{(2.13)}{=} 0.\end{aligned} \tag{2.15}$$

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 2.12 *Für beliebiges $\mathbf{d}_0 = \mathbf{r}_0$ erzeugt die Rekursion*

$$\mathbf{r}_{k+1} = \mathbf{r}_k - \alpha_k\mathbf{A}\mathbf{d}_k, \qquad \mathbf{d}_{k+1} = \mathbf{r}_{k+1} - \beta_k\mathbf{d}_k, \qquad \beta_k = \frac{\mathbf{d}_k^\mathsf{T}\mathbf{A}\mathbf{r}_{k+1}}{\mathbf{d}_k^\mathsf{T}\mathbf{A}\mathbf{d}_k} \tag{2.16}$$

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

Beweis. Sei

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

Wir zeigen zunächst induktiv, dass stets gilt

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

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

$$\mathbf{r}_k \overset{(2.16)}{=} \underbrace{\mathbf{r}_{k-1}}_{\in \mathcal{K}_k(\mathbf{A}, \mathbf{r}_0)} - \alpha_{k-1} \underbrace{\mathbf{A}\mathbf{d}_{k-1}}_{\in \mathcal{K}_{k+1}(\mathbf{A}, \mathbf{r}_0)} \in \mathcal{K}_{k+1}(\mathbf{A}, \mathbf{r}_0).$$

Gemäß (2.15) ist $\mathbf{r}_k \perp \operatorname{span}\{\mathbf{d}_0, \mathbf{d}_1, \dots, \mathbf{d}_{k-1}\} = \mathcal{K}_k(\mathbf{A}, \mathbf{r}_0)$, dies bedeutet

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

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

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

Aus $\mathbf{r}_k = \mathbf{d}_k - \beta_{k-1}\mathbf{d}_{k-1}$ folgt

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

Insbesondere muss aus Dimensionsgründen $\mathbf{d}_k \neq \mathbf{0}$ sein.

Es verbleibt die $\mathbf{A}$-Konjugiertheit zu zeigen: Angenommen, $\mathbf{d}_0, \mathbf{d}_1, \dots, \mathbf{d}_k$ sind $\mathbf{A}$-konjugiert. Der Induktionsschritt folgt dann aus

$$\mathbf{d}_k^\mathsf{T}\mathbf{A}\mathbf{d}_{k+1} \overset{(2.16)}{=} \mathbf{d}_k^\mathsf{T}\mathbf{A}(\mathbf{r}_{k+1} - \beta_k\mathbf{d}_k) \overset{(2.16)}{=} \mathbf{d}_k^\mathsf{T}\mathbf{A}\mathbf{r}_{k+1} - \frac{\mathbf{d}_k^\mathsf{T}\mathbf{A}\mathbf{r}_{k+1}}{\mathbf{d}_k^\mathsf{T}\mathbf{A}\mathbf{d}_k}\mathbf{d}_k^\mathsf{T}\mathbf{A}\mathbf{d}_k = 0$$

und für alle $\ell < k$

$$\mathbf{d}_\ell^\mathsf{T}\mathbf{A}\mathbf{d}_{k+1} \overset{(2.16)}{=} \mathbf{d}_\ell^\mathsf{T}\mathbf{A}\mathbf{r}_{k+1} - \frac{\mathbf{d}_k^\mathsf{T}\mathbf{A}\mathbf{r}_{k+1}}{\mathbf{d}_k^\mathsf{T}\mathbf{A}\mathbf{d}_k} \underbrace{\mathbf{d}_\ell^\mathsf{T}\mathbf{A}\mathbf{d}_k}_{=0} = (\mathbf{A}\mathbf{d}_\ell)^\mathsf{T}\mathbf{r}_{k+1} = 0$$

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

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

$$\alpha_k \overset{(2.13)}{=} \frac{\mathbf{d}_k^\intercal \mathbf{r}_k}{\mathbf{d}_k^\intercal \mathbf{A}\mathbf{d}_k} \overset{(2.16)}{=} \frac{(\mathbf{r}_k - \beta_{k-1}\mathbf{d}_{k-1})^\intercal \mathbf{r}_k}{\mathbf{d}_k^\intercal \mathbf{A}\mathbf{d}_k} \overset{(2.15)}{=} \frac{\|\mathbf{r}_k\|_2^2}{\mathbf{d}_k^\intercal \mathbf{A}\mathbf{d}_k}. \tag{2.17}$$

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

$$\mathbf{d}_k^\intercal \mathbf{A}\mathbf{r}_{k+1} = (\mathbf{A}\mathbf{d}_k)^\intercal \mathbf{r}_{k+1} \overset{(2.16)}{=} \frac{1}{\alpha_k}(\mathbf{r}_k - \mathbf{r}_{k+1})^\intercal \mathbf{r}_{k+1} \overset{(2.17)}{=} -\frac{\mathbf{d}_k^\intercal \mathbf{A}\mathbf{d}_k}{\|\mathbf{r}_k\|_2^2}\|\mathbf{r}_{k+1}\|_2^2$$

und damit

$$\beta_k = \frac{\mathbf{d}_k^\intercal \mathbf{A}\mathbf{r}_{k+1}}{\mathbf{d}_k^\intercal \mathbf{A}\mathbf{d}_k} = -\frac{\|\mathbf{r}_{k+1}\|_2^2}{\|\mathbf{r}_k\|_2^2}. \tag{2.18}$$

Kombination von (2.12) und (2.16)–(2.18) liefert schließlich:

Algorithmus 2.13 (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 $\{\mathbf{x}_k\}_{k>0}$

① Initialisierung: setze $\mathbf{d}_0 = \mathbf{r}_0 := \mathbf{b} - \mathbf{A}\mathbf{x}_0$ und $k := 0$

② berechne

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

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.

Satz 2.14 *Die Iterierten $\{\mathbf{x}_k\}$ des CG-Verfahrens genügen bezüglich der* ***Energienorm***

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

der Fehlerabschätzung

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

Beweis. Wegen $\mathbf{r}_k = \mathbf{b} - \mathbf{A}\mathbf{x}_k = \mathbf{A}(\mathbf{x} - \mathbf{x}_k)$ folgt

$$\begin{aligned}
\|\mathbf{x} - \mathbf{x}_k\|_{\mathbf{A}}^2 &= (\mathbf{b} - \mathbf{A}\mathbf{x}_k)^{\mathsf{T}}\mathbf{A}^{-1}(\mathbf{b} - \mathbf{A}\mathbf{x}_k) \\
&= \min_{\mathbf{z} \in \mathbf{x}_0 + \mathcal{K}_k(\mathbf{A}, \mathbf{r}_0)} \|\mathbf{x} - \mathbf{z}\|_{\mathbf{A}}^2 \\
&= \min_{\mathbf{c} \in \mathbb{R}^k} \|\mathbf{x} - \mathbf{x}_0 - c_1\mathbf{r}_0 - c_2\mathbf{A}\mathbf{r}_0 - \cdots - c_k\mathbf{A}^{k-1}\mathbf{r}_0\|_{\mathbf{A}}^2 \\
&= \min_{\mathbf{c} \in \mathbb{R}^k} \|(\mathbf{x} - \mathbf{x}_0) - c_1\mathbf{A}(\mathbf{x} - \mathbf{x}_0) - c_2\mathbf{A}^2(\mathbf{x} - \mathbf{x}_0) - \cdots - c_k\mathbf{A}^k(\mathbf{x} - \mathbf{x}_0)\|_{\mathbf{A}}^2 \\
&= \min_{\mathbf{c} \in \mathbb{R}^k} \|(\underbrace{\mathbf{I} - c_1\mathbf{A} - c_2\mathbf{A}^2 - \cdots - c_k\mathbf{A}^k}_{=p(\mathbf{A})})(\mathbf{x} - \mathbf{x}_0)\|_{\mathbf{A}}^2 \\
&= \min_{p \in \Pi_k \,:\, p(0)=1} \|p(\mathbf{A})(\mathbf{x} - \mathbf{x}_0)\|_{\mathbf{A}}^2.
\end{aligned}$$

Da $\mathbf{A} \in \mathbb{R}^{n\times n}$ symmetrisch und positiv definit ist, existieren n Eigenwerte $0 < \lambda_1 \le \lambda_2 \le \cdots \le \lambda_n$ und zugehörige orthonormale Eigenvektoren $\{\mathbf{v}_i\}_{i=1}^n$. Hiermit ergibt sich

$$\mathbf{x} - \mathbf{x}_0 = \sum_{i=1}^n \mathbf{v}_i^{\mathsf{T}}(\mathbf{x} - \mathbf{x}_0)\mathbf{v}_i = \sum_{i=1}^n \gamma_i \mathbf{v}_i$$

und

$$\|\mathbf{x} - \mathbf{x}_0\|_{\mathbf{A}}^2 = \sum_{i,j=1}^n \gamma_i\gamma_j \mathbf{v}_i^{\mathsf{T}}\mathbf{A}\mathbf{v}_j = \sum_{i=1}^n |\gamma_i|^2\lambda_i.$$

Für ein beliebiges Polynom p folgt daher

$$\begin{aligned}
\|p(\mathbf{A})(\mathbf{x} - \mathbf{x}_0)\|_{\mathbf{A}}^2 &= \left\|\sum_{i=1}^n \gamma_i p(\mathbf{A})\mathbf{v}_i\right\|_{\mathbf{A}}^2 = \left\|\sum_{i=1}^n \gamma_i p(\lambda_i)\mathbf{v}_i\right\|_{\mathbf{A}}^2 = \sum_{i=1}^n |\gamma_i|^2 |p(\lambda_i)|^2\lambda_i \\
&\le \left(\max_{i=1}^n |p(\lambda_i)|^2\right) \sum_{i=1}^n |\gamma_i|^2\lambda_i = \max_{i=1}^n |p(\lambda_i)|^2 \|\mathbf{x} - \mathbf{x}_0\|_{\mathbf{A}}^2.
\end{aligned}$$

Wir werden nun ein spezielles Polynom $q \in \{p \in \Pi_k : p(0) = 1\}$ angeben, für das sich die gewünschte Fehlerabschätzung ergibt. Dazu wählen wir das $(k+1)$-te *Tschebyscheff-Polynom*

$$T_k(t) = \begin{cases} \cos(k \arccos t), & |t| \le 1 \\ \frac{1}{2}\left\{(t + \sqrt{t^2-1})^k + (t + \sqrt{t^2-1})^{-k}\right\}, & |t| > 1 \end{cases}$$

und setzen

$$q(\lambda) = \frac{T_k\big((\lambda_n + \lambda_1 - 2\lambda)/(\lambda_n - \lambda_1)\big)}{T_k\big((\lambda_n + \lambda_1)/(\lambda_n - \lambda_1)\big)}.$$

Wegen

$$\lambda \in [\lambda_1, \lambda_n] \quad \Longleftrightarrow \quad \frac{\lambda_n + \lambda_1 - 2\lambda}{\lambda_n - \lambda_1} \in [-1, 1]$$

und $\max_{|t|\le 1} |T_k(t)| = 1$ folgt daher

$$\max_{i=1}^{n} |q(\lambda_i)| = \frac{1}{T_k\big((\lambda_n + \lambda_1)/(\lambda_n - \lambda_1)\big)} = \frac{2}{c^k + c^{-k}}$$

mit

$$c = \frac{\lambda_n + \lambda_1}{\lambda_n - \lambda_1} + \sqrt{\left(\frac{\lambda_n + \lambda_1}{\lambda_n - \lambda_1}\right)^2 - 1} = \frac{\sqrt{\frac{\lambda_n}{\lambda_1}} + 1}{\sqrt{\frac{\lambda_n}{\lambda_1}} - 1} = \frac{\sqrt{\operatorname{cond}_2 \mathbf{A}} + 1}{\sqrt{\operatorname{cond}_2 \mathbf{A}} - 1}.$$

Zusammengefasst haben wir damit schließlich gezeigt, dass

$$\|\mathbf{x} - \mathbf{x}_k\|_{\mathbf{A}} \le \max_{i=1}^{n} |p(\lambda_i)| \|\mathbf{x} - \mathbf{x}_0\|_{\mathbf{A}} \le \frac{2c^{-k}}{1 + c^{-2k}} \|\mathbf{x} - \mathbf{x}_0\|_{\mathbf{A}} \le 2c^{-k} \|\mathbf{x} - \mathbf{x}_0\|_{\mathbf{A}}. \qquad \spadesuit$$

Bemerkung Aus der Approximationstheorie ist bekannt, dass das im obigen Beweis verwendete Polynom auf die kleinstmögliche obere Schranke führt. ♦

Wir wollen das CG-Verfahren nun dazu verwenden, die Normalengleichungen

$$\mathbf{A}^\mathsf{T}\mathbf{A}\mathbf{x} = \mathbf{A}^\mathsf{T}\mathbf{b}$$

zu lösen im Falle rechteckiger Matrizen $\mathbf{A} \in \mathbb{R}^{m\times n}$ mit $m \ge n = \operatorname{rang}(\mathbf{A})$. Allerdings soll das explizite Ausmultiplizieren der Matrix $\mathbf{A}^\mathsf{T}\mathbf{A}$ vermieden werden, da deren Kondition

wesentlich schlechter ist als die von $\mathbf{A}$. Zudem ist mit der Matrix $\mathbf{A}$ nicht notwendig auch $\mathbf{A}^\mathsf{T}\mathbf{A}$ dünnbesetzt.

Es bezeichne $\mathbf{r}_k = \mathbf{b} - \mathbf{A}\mathbf{x}_k$ das Residuum und

$$\mathbf{s}_k = \mathbf{A}^\mathsf{T}\mathbf{b} - \mathbf{A}^\mathsf{T}\mathbf{A}\mathbf{x}_k = \mathbf{A}^\mathsf{T}\mathbf{r}_k$$

das Residuum der Normalengleichungen. Die Berechnung von $\mathbf{s}_{k+1}$ geschieht dann im Algorithmus am besten in zwei Schritten

$$\mathbf{r}_{k+1} := \mathbf{r}_k - \alpha_k \mathbf{A}\mathbf{d}_k, \qquad \mathbf{s}_{k+1} = \mathbf{A}^\mathsf{T}\mathbf{r}_{k+1}.$$

Benutzt man ferner

$$\alpha_k = \frac{\|\mathbf{s}_k\|_2^2}{\mathbf{d}_k^\mathsf{T}(\mathbf{A}^\mathsf{T}\mathbf{A})\mathbf{d}_k} = \frac{\|\mathbf{s}_k\|_2^2}{\|\mathbf{A}\mathbf{d}_k\|_2^2},$$

so kann die explizite Multiplikation $\mathbf{A}^\mathsf{T}\mathbf{A}$ vollständig vermieden werden:

Algorithmus 2.15 (CGLS-Verfahren)
input: Matrix $\mathbf{A} \in \mathbb{R}^{m\times n}$, rechte Seite $\mathbf{b} \in \mathbb{R}^m$ und Startnäherung $\mathbf{x}_0 \in \mathbb{R}^n$
output: Folge von Iterierten $\{\mathbf{x}_k\}_{k>0}$

① Initialisierung: setze $\mathbf{r}_0 = \mathbf{b} - \mathbf{A}\mathbf{x}_0$, $\mathbf{d}_0 = \mathbf{s}_0 := \mathbf{A}^\mathsf{T}\mathbf{r}_0$ und $k := 0$

② berechne

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

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

Proposition 2.16 *Die k-te Iterierte $\mathbf{x}_k$ des CGLS-Verfahrens liegt im verschobenen Krylov-Raum*

$$\mathbf{x}_0 + \mathcal{K}_k(\mathbf{A}^\mathsf{T}\mathbf{A}, \mathbf{A}^\mathsf{T}\mathbf{r}_0) = \mathbf{x}_0 + \operatorname{span}\{\mathbf{A}^\mathsf{T}\mathbf{r}_0, (\mathbf{A}^\mathsf{T}\mathbf{A})\mathbf{A}^\mathsf{T}\mathbf{r}_0, \dots, (\mathbf{A}^\mathsf{T}\mathbf{A})^{k-1}\mathbf{A}^\mathsf{T}\mathbf{r}_0\}.$$

Unter allen diesen Elementen $\mathbf{z}$ dieses affinen Raums minimiert $\mathbf{x}_k$ die Residuennorm $\|\mathbf{b}-\mathbf{A}\mathbf{z}\|_2$.

Beweis. Gemäß der Konstruktion des CG-Verfahrens minimiert die Iterierte $\mathbf{x}_k$ das Funktional

$$\begin{aligned}\phi(\mathbf{z}) &= (\mathbf{A}^\mathsf{T}\mathbf{b} - \mathbf{A}^\mathsf{T}\mathbf{A}\mathbf{z})^\mathsf{T}(\mathbf{A}^\mathsf{T}\mathbf{A})^{-1}(\mathbf{A}^\mathsf{T}\mathbf{b} - \mathbf{A}^\mathsf{T}\mathbf{A}\mathbf{z}) \\ &= (\mathbf{b} - \mathbf{A}\mathbf{z})^\mathsf{T} \underbrace{\mathbf{A}(\mathbf{A}^\mathsf{T}\mathbf{A})^{-1}\mathbf{A}^\mathsf{T}}_{=\mathbf{I}}(\mathbf{b} - \mathbf{A}\mathbf{z}) \\ &= \|\mathbf{b} - \mathbf{A}\mathbf{z}\|_2^2\end{aligned}$$

unter allen Elementen $\mathbf{z} \in \mathbf{x}_0 + \mathcal{K}_k(\mathbf{A}^\mathsf{T}\mathbf{A}, \mathbf{A}^\mathsf{T}\mathbf{r}_0)$. ♠

Bemerkung Man kann das CGLS-Verfahren sogar auf beliebige Matrizen $\mathbf{A} \in \mathbb{R}^{m\times n}$ anwenden, insbesondere auf Matrizen ohne vollen Rang, falls man einen Startvektor $\mathbf{x}_0 \in \operatorname{img}(\mathbf{A}^\mathsf{T}\mathbf{A}) = \operatorname{kern}(\mathbf{A}^\mathsf{T}\mathbf{A})^\perp$ wählt, beispielsweise $\mathbf{x}_0 = \mathbf{0}$. Da

$$\mathbf{A}^\mathsf{T}\mathbf{r}_0 = \mathbf{A}^\mathsf{T}(\mathbf{b} - \mathbf{A}\mathbf{x}_0) \in \operatorname{img}(\mathbf{A}^\mathsf{T}\mathbf{A}) \perp \operatorname{kern}(\mathbf{A}^\mathsf{T}\mathbf{A}),$$

gilt für die Iterierten stets

$$\mathbf{x}_k \in \mathbf{x}_0 + \mathcal{K}_k(\mathbf{A}^\mathsf{T}\mathbf{A}, \mathbf{A}^\mathsf{T}\mathbf{r}_0) \in \operatorname{img}(\mathbf{A}^\mathsf{T}\mathbf{A}) \perp \operatorname{kern}(\mathbf{A}^\mathsf{T}\mathbf{A}).$$

Mit anderen Worten, man iteriert nur orthogonal zum Kern von $\mathbf{A}^\mathsf{T}\mathbf{A}$. Daher folgt insbesondere, dass $\mathbf{x}_k$ auch stets die Lösung mit der kleinsten Euklid-Norm ist. ♦

3 Nichtlineare Ausgleichsprobleme

3.1 Gradientenverfahren

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

$$\mathbf{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_m \end{bmatrix} \in \mathbb{R}^m, \qquad \mathbf{f}(\mathbf{z}) = \begin{bmatrix} f_1(z_1, z_2, \dots, z_n) \\ f_2(z_1, z_2, \dots, z_n) \\ \vdots \\ f_m(z_1, z_2, \dots, x_n) \end{bmatrix} : \mathbb{R}^n \to \mathbb{R}^m$$

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

$$\phi(\mathbf{z}) := \frac{1}{2}\|\mathbf{y} - \mathbf{f}(\mathbf{z})\|_2^2 = \frac{1}{2}\sum_{i=1}^{m} |y_i - f_i(z_1, z_2, \dots, z_n)|^2 \to \min_{\mathbf{z}\in\mathbb{R}^n} \tag{3.1}$$

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.

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

$$-\nabla\phi(\mathbf{x}_k) = \big(\mathbf{f}'(\mathbf{x}_k)\big)^\mathsf{T}\big(\mathbf{y} - \mathbf{f}(\mathbf{x}_k)\big), \quad \mathbf{f}'(\mathbf{x}_k) = \begin{bmatrix} \frac{\partial f_1}{\partial z_1}(\mathbf{x}_k) & \frac{\partial f_1}{\partial z_2}(\mathbf{x}_k) & \dots & \frac{\partial f_1}{\partial z_n}(\mathbf{x}_k) \\ \frac{\partial f_2}{\partial z_1}(\mathbf{x}_k) & \frac{\partial f_2}{\partial z_2}(\mathbf{x}_k) & \dots & \frac{\partial f_2}{\partial z_n}(\mathbf{x}_k) \\ \vdots & \vdots & & \vdots \\ \frac{\partial f_m}{\partial z_1}(\mathbf{x}_k) & \frac{\partial f_m}{\partial z_2}(\mathbf{x}_k) & \dots & \frac{\partial f_m}{\partial z_n}(\mathbf{x}_k) \end{bmatrix}$$

so aufzudatieren

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

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

Lemma 3.1 *Vorausgesetzt es ist $\phi'(\mathbf{x}_k) \neq \mathbf{0}$, dann gibt es ein $\delta > 0$, so dass die Funktion*

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

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

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

Beweis. Die Funktion φ ist stetig differenzierbar und es gilt

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

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

Algorithmus 3.2 (Gradientenverfahren)
input: Funktion $\phi : \mathbb{R}^n \to \mathbb{R}$ und Startnäherung $\mathbf{x}_0 \in \mathbb{R}^n$
output: Folge von Iterierten $\{\mathbf{x}_k\}_{k>0}$

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

② berechne den Gradienten $\nabla\phi(\mathbf{x}_k)$ und setze $\alpha_k := 1$

③ solange

$$\phi\big(\mathbf{x}_k - \alpha_k\nabla\phi(\mathbf{x}_k)\big) > \phi(\mathbf{x}_k) - \sigma\alpha_k\|\nabla\phi(\mathbf{x}_k)\|_2^2 \tag{3.2}$$

setze $\alpha_k := \alpha_k/2$

④ setze $\mathbf{x}_{k+1} := \mathbf{x}_k - \alpha_k\nabla\phi(\mathbf{x}_k)$

⑤ erhöhe $k := k+1$ und gehe nach ②

Bemerkung Man beachte die modifizierte Abbruchbedingung der Liniensuche in (3.2), die nicht nur $\phi(\mathbf{x}_{k+1}) < \phi(\mathbf{x}_k)$ garantiert, sondern die *Armijo-Goldstein-Bedingung*

$$\phi\big(\mathbf{x}_k - \alpha_k \nabla\phi(\mathbf{x}_k)\big) \le \phi(\mathbf{x}_k) - \sigma\alpha_k \|\nabla\phi(\mathbf{x}_k)\|_2^2. \tag{3.3}$$

Dass die Liniensuche mit einem $\alpha_k > 0$ abbricht, folgt aus

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

♦

Satz 3.3 *Es sei $D \subset \mathbb{R}^n$ eine offene Menge, in der $\mathbf{f}$ stetig differenzierbar und $\mathbf{f}'$ zudem Lipschitz-stetig ist. Ferner sei neben $\mathbf{x}_0$ auch die gesamte Niveaumenge $\{\mathbf{z} \in \mathbb{R}^n : \phi(\mathbf{z}) \le \phi(\mathbf{x}_0)\}$ in D enthalten. Dann gilt für die Iterierten $\{\mathbf{x}_k\}_{k\ge 0}$ aus Algorithmus 3.2*

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

Beweis. Da D die gesamte Niveaumenge enthält, ist sichergestellt, dass die Iterierten $\{\mathbf{x}_k\}_{k\ge 0}$ alle in D enthalten ist. Nach Konstruktion ist dann die Folge $\{\phi(\mathbf{x}_k)\}_{k\ge 0}$ monoton fallend und nach unten beschränkt. Daher folgt aus der Armijo-Goldstein-Bedingung (3.3), dass

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

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

$$\alpha_k \|\nabla\phi(\mathbf{x}_k)\|_2^2 \to 0, \quad k \to \infty. \tag{3.4}$$

Es verbleibt zu zeigen, dass $\alpha_k > \varepsilon$ für ein $\varepsilon > 0$.

Für festes k ist aufgrund von Algorithmus 3.2 $\alpha_k = 1$ oder die Armijo-Goldstein-Bedingung ist für $2\alpha_k$ verletzt:

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

mit dem Taylor-Restglied

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

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

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

dies bedeutet

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

Somit bleibt α_k für alle $k \geq 0$ größer als $\min\{1, \varepsilon\}$ und daher folgt die Behauptung aus (3.4). ♠

Beachte: Satz 3.3 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 darüber hinaus kein Minimum von ϕ zu sein.

Beispiel 3.4 Gegeben sei die Funktion $\mathbf{f}(\xi, \eta) = [\xi, \eta^2 - 1, \xi(\eta^2 - 1)]^\mathsf{T}$ und der Datenvektor $\mathbf{y} = \mathbf{0}$. Wir betrachten das nichtlineare Ausgleichsproblem

$$\phi(\xi, \eta) = \|\mathbf{y} - \mathbf{f}(\xi, \eta)\|_2^2 = \xi^2 + (\eta^2 - 1)^2 + \xi^2(\eta^2 - 1)^2 \to \min.$$

Das Minimum von ϕ ist 0 und wird offensichtlich für $\xi = 0$ und $\eta = \pm 1$ angenommen. Hat eine Iterierte $\mathbf{x}_k$ von Algorithmus 3.2 die Form $\mathbf{x}_k = [\xi_k, 0]^\mathsf{T}$, dann gilt

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

Daher hat die nächste Iterierte zwangsläufig wieder die Form $\mathbf{x}_{k+1} = [\xi_{k+1}, 0]^\mathsf{T}$ und nach Satz 3.3 konvergiert $\nabla\phi(\mathbf{x}_k) = [4\xi_k, 0]^\mathsf{T}$ gegen Null. Deshalb streben auch ξ_k und $\mathbf{x}_k$ gegen Null für $k \to \infty$. Dennoch ist $[0, 0]^\mathsf{T}$ lediglich ein Sattelpunkt von ϕ, da $\phi(0, \eta)$ für $\eta = 0$ ein lokales Maximum aufweist. ♣

3.2 Gauß-Newton-Verfahren

Natürlich kann man das nichtlineare Ausgleichsproblem (3.1) auch mit dem Newton-Verfahren für die Gleichung

$$\nabla\phi(\mathbf{z}) = -\big(\mathbf{f}'(\mathbf{z})\big)^\mathsf{T}\big(\mathbf{y} - \mathbf{f}(\mathbf{z})\big) \stackrel{!}{=} \mathbf{0}$$

lösen, was der Iteration

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

entspricht. Hierzu wird jedoch neben dem Gradienten $\nabla\phi$ auch die Hesse-Matrix benötigt, das ist

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

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

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

$$\mathbf{f}(\mathbf{z}+\mathbf{h}) = \mathbf{f}(\mathbf{z}) + \mathbf{f}'(\mathbf{z})\mathbf{h} + o(\|\mathbf{h}\|_2).$$

Ist nun $\mathbf{x}_k$ eine Näherungslösung des Ausgleichsproblems (3.1), so erwartet man, dass die Optimallösung $\mathbf{x}_{k+1} = \mathbf{x}_k + \mathbf{d}_k$ des linearisierten Problems

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

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

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

lösen. Dies führt auf folgenden Algorithmus:

Algorithmus 3.5 (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 $\{\mathbf{x}_k\}_{k>0}$

① Initialisierung: setze $k = 0$

② löse die Normalengleichungen

$$\big(\mathbf{f}'(\mathbf{x}_k)\big)^{\mathsf{T}}\mathbf{f}'(\mathbf{x}_k)\mathbf{d}_k = \big(\mathbf{f}'(\mathbf{x}_k)\big)^{\mathsf{T}}\big(\mathbf{y} - \mathbf{f}(\mathbf{x}_k)\big) \tag{3.5}$$

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

④ erhöhe $k := k + 1$ und gehe nach ②

Satz 3.6 *Sei $D \subset \mathbb{R}^n$ offen und $\mathbf{f} : D \to \mathbb{R}^m$ eine stetig differenzierbare Abbildung. Das Minimierungsproblem* (3.1) *habe eine Lösung $\mathbf{x} \in D$ mit* $\operatorname{rang}\big(\mathbf{f}'(\mathbf{x})\big) = n \leq m$. *Sei $\lambda > 0$ der kleinste Eigenwert von $\big(\mathbf{f}'(\mathbf{x})\big)^{\mathsf{T}}\mathbf{f}'(\mathbf{x})$. Ferner gelte die Lipschitz-Bedingung*

$$\|\mathbf{f}'(\mathbf{z}) - \mathbf{f}'(\mathbf{x})\|_2 \leq \alpha\|\mathbf{z} - \mathbf{x}\|_2 \tag{3.6}$$

und

$$\big\|\big(\mathbf{f}'(\mathbf{z}) - \mathbf{f}'(\mathbf{x})\big)^{\mathsf{T}}\big(\mathbf{y} - \mathbf{f}(\mathbf{x})\big)\big\|_2 \leq \beta\|\mathbf{z} - \mathbf{x}\|_2 \tag{3.7}$$

mit $\beta < \lambda$ für alle $\mathbf{z}$ aus einer Umgebung von $\mathbf{x}$. Dann existiert ein $\varepsilon > 0$, so dass für jeden Startvektor $\mathbf{x}_0 \in B_\varepsilon(\mathbf{x})$ die Folge der Iterierten $\{\mathbf{x}_k\}_{k\geq 0}$ mindestens linear gegen $\mathbf{x}$ konvergiert.

Beweis. Nach Voraussetzung gibt es ein $\varepsilon_1 > 0$, so dass (3.6) und (3.7) für alle $\mathbf{z} \in B_{\varepsilon_1}(\mathbf{x})$ gelten. Aus Stetigkeitsgründen folgt außerdem die Existenz von $\gamma > 0$, so dass

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

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

$$\left\|\Big(\big(\mathbf{f}'(\mathbf{x})\big)^{\mathsf{T}}\mathbf{f}'(\mathbf{x})\Big)^{-1}\right\|_2 = \frac{1}{\lambda}. \tag{3.8}$$

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

$$\left\|\Big(\big(\mathbf{f}'(\mathbf{z})\big)^{\mathsf{T}}\mathbf{f}'(\mathbf{z})\Big)^{-1}\right\|_2 \leq \frac{\delta}{\lambda} \quad \text{für alle} \quad \mathbf{z} \in B_{\varepsilon_2}(\mathbf{x}). \tag{3.9}$$

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

$$\delta < \frac{\lambda}{\beta}. \tag{3.10}$$

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

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

und damit

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

Wir setzen nun

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

Aus (3.5) folgt für $\mathbf{x}_k \in B_\varepsilon(\mathbf{x})$ dann

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

Hieraus ergibt sich mit (3.8) und (3.9)

$$\begin{aligned}
\|\mathbf{x}_{k+1} - \mathbf{x}\|_2 &\le \left\|\Big(\big(\mathbf{f}'(\mathbf{x}_k)\big)^\mathsf{T}\mathbf{f}'(\mathbf{x}_k)\Big)^{-1}\right\|_2 \Big[\big\|\big(\mathbf{f}'(\mathbf{x}_k)\big)^\mathsf{T}\big(\mathbf{y} - \mathbf{f}(\mathbf{x})\big)\big\|_2 \\
&\qquad\qquad + \|\mathbf{f}'(\mathbf{x}_k)\|_2 \|\mathbf{f}(\mathbf{x}) - \mathbf{f}(\mathbf{x}_k) - \mathbf{f}'(\mathbf{x}_k)(\mathbf{x} - \mathbf{x}_k)\|_2\Big] \\
&\le \frac{\delta}{\lambda}\left[\big\|\big(\mathbf{f}'(\mathbf{x}_k)\big)^\mathsf{T}\big(\mathbf{y} - \mathbf{f}(\mathbf{x})\big)\big\|_2 + \frac{\alpha\gamma}{2}\|\mathbf{x}_k - \mathbf{x}\|_2^2\right]. \qquad (3.11)
\end{aligned}$$

Wegen $\mathbf{0} = \nabla\phi(\mathbf{x}) = -\big(\mathbf{f}'(\mathbf{x})\big)^\mathsf{T}\big(\mathbf{y} - \mathbf{f}(\mathbf{x})\big)$ erhalten wir aufgrund von (3.7)

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

dies bedeutet,

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

Wegen (3.10) ist die Konstante $(\lambda + \beta\delta)/(2\lambda) < 1$, das heißt, alle Iterierten $\{\mathbf{x}_k\}_{k\geq 0}$ liegen in $B_\varepsilon(\mathbf{x})$ und konvergieren mindestens linear gegen $\mathbf{x}$. ♠

Bemerkung Die Voraussetzung (3.7) besagt im Prinzip, dass das Residuum

$$\mathbf{r}(\mathbf{x}) = \mathbf{y} - \mathbf{f}(\mathbf{x})$$

klein genug sein soll. Dies sieht man insbesondere, wenn man sie durch die stärkere Bedingung

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

ersetzt. Dann folgt nämlich (3.7):

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

Korollar 3.7 *Zusätzlich zu den Voraussetzungen aus Satz 3.6 gelte* $\mathbf{f}(\mathbf{x}) = \mathbf{y}$. *Dann existiert ein* $\varepsilon > 0$, *so dass für jeden Startvektor* $\mathbf{x}_0 \in B_\varepsilon(\mathbf{x})$ *die Folge der Iterierten* $\{\mathbf{x}_k\}_{k\geq 0}$ *quadratisch gegen* $\mathbf{x}$ *konvergiert.*

Beweis. Aus Satz 3.6 folgt die lineare Konvergenz der Iterierten $\{\mathbf{x}_k\}_{k\geq 0}$ gegen $\mathbf{x}$. Zum Nachweis der quadratischen Konvergenz bemerken wir, dass aufgrund der Voraussetzung (3.7) mit $\beta = 0$ gilt. Daher folgt aus (3.11)

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

3.3 Levenberg-Marquardt-Verfahren

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

$$\min_{\mathbf{h}\in\mathbb{R}^n\,:\,\|\mathbf{h}\|_2\leq\Delta} \psi(\mathbf{h}) = \min_{\mathbf{h}\in\mathbb{R}^n\,:\,\|\mathbf{h}\|_2\leq\Delta} \frac{1}{2}\|\mathbf{r}_k - \mathbf{f}'(\mathbf{x}_k)\mathbf{h}\|_2^2, \tag{3.12}$$

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 $\overline{B_\Delta(\mathbf{0})}$ 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 $B_\Delta(\mathbf{0})$ und erfüllt
$$\nabla\psi(\mathbf{d}_k) = \big(\mathbf{f}'(\mathbf{x}_k)\big)^\mathsf{T}\big(\mathbf{f}'(\mathbf{x}_k)\mathbf{d}_k - \mathbf{r}_k\big) = \mathbf{0}. \tag{3.13}$$
2. Das Minimum liegt auf dem Rand, erfüllt also $\|\mathbf{h}\|_2 = \Delta$. In diesem Fall muss die Höhenlinie von ψ in $\mathbf{d}_k$ genau den Kreis $\partial B_\Delta(\mathbf{0})$ tangieren, das heißt, der Gradient $\nabla\psi(\mathbf{d}_k)$ zeigt in Richtung des Nullpunkts:
$$\nabla\psi(\mathbf{d}_k) = \big(\mathbf{f}'(\mathbf{x}_k)\big)^\mathsf{T}\big(\mathbf{f}'(\mathbf{x}_k)\mathbf{d}_k - \mathbf{r}_k\big) = -\lambda_k\mathbf{d}_k \quad \text{für ein} \quad \lambda_k \geq 0. \tag{3.14}$$

Da die Gleichung (3.13) als Grenzfall $\lambda_k = 0$ von (3.14) angesehen werden kann, erhalten wir:

Lemma 3.8 *Die Lösung $\mathbf{d}_k$ des restringierten Problems (3.12) genügt der Gleichung*
$$\Big(\big(\mathbf{f}'(\mathbf{x}_k)\big)^\mathsf{T}\mathbf{f}'(\mathbf{x}_k) + \lambda_k\mathbf{I}\Big)\mathbf{d}_k = \big(\mathbf{f}'(\mathbf{x}_k)\big)^\mathsf{T}\mathbf{r}_k \tag{3.15}$$
für ein $\lambda_k \geq 0$. Dabei ist der Wert λ_k genau dann positiv, wenn $\|\mathbf{d}_k\|_2 = \Delta > 0$ gilt.

Der Vorteil des regularisierten Systems (3.15) liegt darin, dass es stets eindeutig lösbar ist, sofern $\lambda_k > 0$ ist. Die zugehörige Lösung
$$\mathbf{d}_k = \Big(\big(\mathbf{f}'(\mathbf{x}_k)\big)^\mathsf{T}\mathbf{f}'(\mathbf{x}_k)+\lambda_k\mathbf{I}\Big)^{-1}\big(\mathbf{f}'(\mathbf{x}_k)\big)^\mathsf{T}\mathbf{r}_k = -\Big(\big(\mathbf{f}'(\mathbf{x}_k)\big)^\mathsf{T}\mathbf{f}'(\mathbf{x}_k)+\lambda_k\mathbf{I}\Big)^{-1}\nabla\phi(\mathbf{x}_k)$$
erfüllt offenbar die *Abstiegsbedingung* $\mathbf{d}_k^\mathsf{T}\nabla\phi(\mathbf{x}_k) < 0$, 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 wählen. Eine neue Näherung $\mathbf{x}_{k+1} = \mathbf{x}_k + \mathbf{d}_k$ können wir anhand der Armijo-Goldstein-Bedingung (vergleiche (3.3)) bewerten:
$$\mu = \frac{\phi(\mathbf{x}_k + \mathbf{d}_k) - \phi(\mathbf{x}_k)}{\mathbf{d}_k^\mathsf{T}\nabla\phi(\mathbf{x}_k)} = \frac{1}{2}\frac{\|\mathbf{y} - \mathbf{f}(\mathbf{x}_k)\|_2^2 - \|\mathbf{y} - \mathbf{f}(\mathbf{x}_k + \mathbf{d}_k)\|_2^2}{\mathbf{d}_k^\mathsf{T}\big(\mathbf{f}'(\mathbf{x}_k)\big)^\mathsf{T}\big(\mathbf{y} - \mathbf{f}(\mathbf{x}_k)\big)}.$$

Wir wählen zwei Toleranzgrenzen $0 < \mu^- < \mu^+ < 1$ und akzeptieren den Iterationsschritt, wenn $\mu > \mu^-$ gilt. Dann war der Trust-Region-Radius Δ geeignet gewählt. Ist $\mu > \mu^+$, so können wir Δ sogar vergrößern. Ist hingegen $\mu \leq \mu^-$, dann verwerfen wir den Iterationsschritt und verkleinern Δ. Damit erhalten wir schließlich den folgenden Algorithmus:

Algorithmus 3.9 (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 $\{\mathbf{x}_k\}_{k>0}$

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

② bestimme die Lösung $\mathbf{d}_k$ des restringierten Optimierungsproblems (3.12)

③ berechne

$$\mu_k = \frac{1}{2} \frac{\|\mathbf{y} - \mathbf{f}(\mathbf{x}_k)\|_2^2 - \|\mathbf{y} - \mathbf{f}(\mathbf{x}_k + \mathbf{d}_k)\|_2^2}{\mathbf{d}_k^\mathsf{T} \big(\mathbf{f}'(\mathbf{x}_k)\big)^\mathsf{T} \big(\mathbf{y} - \mathbf{f}(\mathbf{x}_k)\big)} \tag{3.16}$$

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

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

⑥ erhöhe $k := k + 1$ und gehe nach ②

Satz 3.10 *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 $\{\mathbf{z} \in \mathbb{R}^n : \phi(\mathbf{z}) \leq \phi(\mathbf{x}_0)\}$ in D enthalten. Dann gilt für die Iterierten $\{\mathbf{x}_k\}_{k\geq 0}$ aus Algorithmus 3.9*

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

Beweis. (*i*) Zunächst beweisen wir eine obere Schranke für den Lagrange-Parameter λ_k aus (3.15). Dazu nehmen wir ohne Beschränkung der Allgemeinheit an, dass $\lambda_k > 0$ und daher $\|\mathbf{d}_k\|_2 = \Delta_k$ ist. Aus (3.15) folgt

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

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

abgeschätzt werden, so dass folgt

$$\lambda_k \le \frac{\left\|\left(\mathbf{f}'(\mathbf{x}_k)\right)^{\mathsf{T}}\mathbf{r}_k\right\|_2}{\Delta_k} = \frac{\|\nabla\phi(\mathbf{x}_k)\|_2}{\Delta_k}. \tag{3.17}$$

(*ii*) Als nächstes leiten wir eine untere Schranke für den Nenner ν_k in (3.16) her. Im Fall $\lambda_k > 0$ ist $\left(\mathbf{f}'(\mathbf{x}_k)\right)^{\mathsf{T}}\mathbf{f}'(\mathbf{x}_k) + \lambda_k\mathbf{I}$ positiv definit, weshalb eine Cholesky-Zerlegung $\mathbf{L}\mathbf{L}^{\mathsf{T}}$ existiert. Dabei gilt

$$\|\mathbf{L}^{\mathsf{T}}\mathbf{L}\|_2 = \|\mathbf{L}\mathbf{L}^{\mathsf{T}}\|_2 = \left\|\left(\mathbf{f}'(\mathbf{x}_k)\right)^{\mathsf{T}}\mathbf{f}'(\mathbf{x}_k) + \lambda_k\mathbf{I}\right\|_2 = \underbrace{\|\mathbf{f}'(\mathbf{x}_k)\|_2^2}_{\le c \text{ für alle } \mathbf{z}\in D} + \lambda_k \overset{(3.17)}{\le} c + \frac{\|\nabla\phi(\mathbf{x}_k)\|_2}{\Delta_k}.$$

Setzen wir $\mathbf{w} = \mathbf{L}^{-1}\nabla\phi(\mathbf{x}_k)$, so folgt unter Beachtung von (3.15) hieraus

$$\begin{aligned}\nu_k &= \left(\nabla\phi(\mathbf{x}_k)\right)^{\mathsf{T}}(\mathbf{L}\mathbf{L}^{\mathsf{T}})^{-1}\nabla\phi(\mathbf{x}_k) = \mathbf{w}^{\mathsf{T}}\mathbf{w}\frac{\|\nabla\phi(\mathbf{x}_k)\|_2^2}{\left(\nabla\phi(\mathbf{x}_k)\right)^{\mathsf{T}}\nabla\phi(\mathbf{x}_k)} = \frac{\|\mathbf{w}\|_2^2\|\nabla\phi(\mathbf{x}_k)\|_2^2}{\mathbf{w}^{\mathsf{T}}\mathbf{L}^{\mathsf{T}}\mathbf{L}\mathbf{w}} \\ &\ge \frac{\|\mathbf{w}\|_2^2\|\nabla\phi(\mathbf{x}_k)\|_2^2}{\|\mathbf{w}\|_2^2\left(c + \|\nabla\phi(\mathbf{x}_k)\|_2/\Delta_k\right)} \ge \frac{\|\nabla\phi(\mathbf{x}_k)\|_2}{1+c}\min\{\Delta_k, \|\nabla\phi(\mathbf{x}_k)\|_2\}. \end{aligned} \tag{3.18}$$

Im Fall $\lambda_k = 0$ folgt

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

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

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

das heißt, (3.18) ist auch im Fall $\lambda_k = 0$ gültig.

(*iii*) Wir beweisen nun, dass die rechte Seite von (3.18) gegen Null konvergiert. Bei erfolgreichem Iterationsschritt ist $\mu_k > \mu^-$ und aus (3.16) und (3.18) folgt

$$\phi(\mathbf{x}_k) - \phi(\mathbf{x}_{k+1}) > \mu^-\nu_k \ge \frac{\mu^-\|\nabla\phi(\mathbf{x}_k)\|_2}{1+c}\min\{\Delta_k, \|\nabla\phi(\mathbf{x}_k)\|_2\}. \tag{3.19}$$

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

$$\min\{\Delta_k, \|\nabla\phi(\mathbf{x}_k)\|_2\} \to 0, \quad k\to\infty. \tag{3.20}$$

(*iv*) Als nächstes zeigen wir, dass $\|\nabla\phi(\mathbf{x}_k)\|_2$ für eine Teilfolge $\{k_n\}_{n\in\mathbb{N}}$ gegen Null konvergiert für $k_n \to \infty$. 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 (3.20) ergibt sich damit unmittelbar

$$\Delta_k \to 0, \quad k \to \infty. \tag{3.21}$$

Taylor-Entwicklung von μ_k liefert jedoch

$$\begin{aligned}\mu_k &= \frac{\phi(\mathbf{x}_{k+1}) - \phi(\mathbf{x}_k)}{\mathbf{d}_k^\mathsf{T}\nabla\phi(\mathbf{x}_k)} = \frac{\mathbf{d}_k^\mathsf{T}\nabla\phi(\mathbf{x}_k) + \mathcal{O}(\|\mathbf{d}_k\|_2^2)}{\mathbf{d}_k^\mathsf{T}\nabla\phi(\mathbf{x}_k)} \\ &= 1 + \mathcal{O}\left(\frac{\Delta_k^2}{\nu_k}\right) \overset{(3.18)}{=} 1 + \mathcal{O}\Bigg(\underbrace{\frac{\Delta_k}{\|\nabla\phi(\mathbf{x}_k)\|_2}}_{\geq\varepsilon>0}\Bigg) = 1 + \mathcal{O}(\Delta_k), \quad k \to \infty.\end{aligned}$$

Demnach existiert ein $M(\varepsilon) \geq K(\varepsilon)$, so dass $\mu_k > \mu^+$ für alle $k \geq M(\varepsilon)$. Ab dem $M(\varepsilon)$-ten Schritt wird folglich Δ_k in jedem Schritt von Algorithmus 3.9 verdoppelt, was jedoch im Widerspruch zu (3.21) 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

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

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

$$\phi(\mathbf{x}_\ell) - \phi(\mathbf{x}_m) < \frac{\varepsilon^2\mu^-}{(1+c)L}, \tag{3.22}$$

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 (3.19) dass

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

Summation ergibt

$$\frac{\varepsilon\mu^-}{1+c}\sum_{k=\ell}^{m-1}\min\{\|\mathbf{x}_{k+1} - \mathbf{x}_k\|_2, \varepsilon\} \leq \phi(\mathbf{x}_\ell) - \phi(\mathbf{x}_m) \overset{(3.22)}{<} \frac{\varepsilon^2\mu^-}{(1+c)L},$$

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

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

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 (3.12) 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 (3.12) auf dem Rand liegt. Wir suchen dann dasjenige Tupel $(\lambda_k, \mathbf{d}_k)$, das (3.15) und $\|\mathbf{d}_k\|_2 = \Delta_k$ löst.

Es bezeichne $z_1 \ge \cdots \ge z_n \ge 0$ die Eigenwerte von $\big(\mathbf{f}'(\mathbf{x}_k)\big)^\mathsf{T}\mathbf{f}'(\mathbf{x}_k)$ und $\{\mathbf{v}_i\}_{i=1}^n$ die zugehörigen orthonormalen Eigenvektoren. Entwickeln wir die rechte Seite von (3.15) in diese Eigenbasis

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

dann folgt

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

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

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

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.

4 Nichtlineare Optimierung

4.1 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 $\mathbf{x}^\star \in \mathbb{R}^n$, 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 4.1 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.

4.2 Optimalitätskriterien

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

Definition 4.2 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 4.3 (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 4.4 (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 4.5 (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.*

4.3 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 4.6 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

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

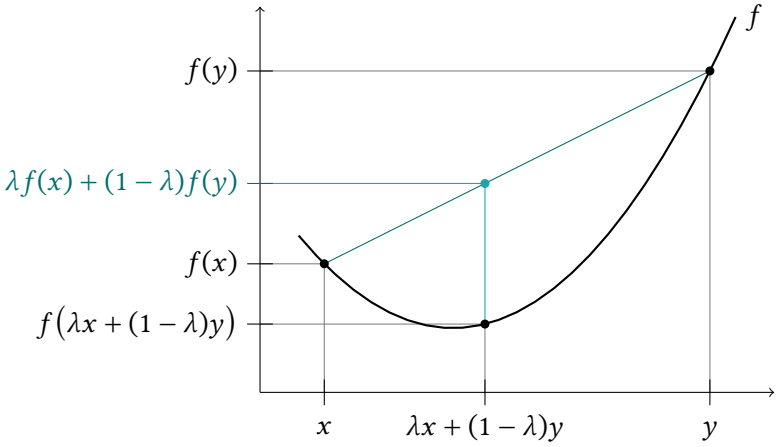
Gilt für $\mathbf{x} \neq \mathbf{y}$ sogar stets die strikte Ungleichung, dann heißt die Funktion **strikt konvex**. Gibt es ein $\mu > 0$, so dass

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

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

Beispiele 4.7

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

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

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 4.8 *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*

$$f(\mathbf{x}) - f(\mathbf{y}) \geq \nabla f(\mathbf{y})^\mathsf{T}(\mathbf{x} - \mathbf{y}). \tag{4.1}$$

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

$$f(\mathbf{x}) - f(\mathbf{y}) \geq \nabla f(\mathbf{y})^\mathsf{T}(\mathbf{x} - \mathbf{y}) + \mu\|\mathbf{x} - \mathbf{y}\|_2^2 \tag{4.2}$$

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

Beweis. Es gelte zuächst (4.2). Für $\mathbf{x}, \mathbf{y} \in D$ und beliebiges $\lambda \in (0, 1)$ ergibt sich dann mit $\mathbf{z} := \lambda\mathbf{x} + (1 - \lambda)\mathbf{y}$

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

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

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

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 $\lambda \to 0+$

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

dies bedeutet, es gilt (4.2). Da der soeben geführte Beweis auch im Fall $\mu = 0$ seine Gültigkeit behält, folgt die Äquivalenz von (4.1) zur Konvexität von f.

Es verbleibt zu zeigen, dass die strikte Konvexität von f die strikte Ungleichung

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

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 (4.1). 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

$$\nabla f(\mathbf{y})^\intercal(\mathbf{x}-\mathbf{y}) = 2\nabla f(\mathbf{y})^\intercal(\mathbf{z}-\mathbf{y}) \le 2\Big\{f(\mathbf{z}) - f(\mathbf{y})\Big\}. \tag{4.3}$$

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 (4.3) liefert die Behauptung

$$\nabla f(\mathbf{y})^\intercal(\mathbf{x}-\mathbf{y}) < f(\mathbf{x}) - f(\mathbf{y}). \qquad \spadesuit$$

Satz 4.9 *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 $\mathbf{y}^\star \in D$ mit $f(\mathbf{y}^\star) < f(\mathbf{x}^\star)$. Für alle

$$\mathbf{x} = \lambda\mathbf{x}^\star + (1-\lambda)\mathbf{y}^\star, \quad \lambda \in (0,1) \tag{4.4}$$

gilt aufgrund der Konvexität

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

Da in jeder Umgebung von $\mathbf{x}^\star$ Punkte der Form (4.4) 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 gemäß (4.1)

$$\nabla f(\mathbf{x}^\star)^\mathsf{T}(\mathbf{y}^\star - \mathbf{x}^\star) \leq f(\mathbf{y}^\star) - f(\mathbf{x}^\star) < 0.$$

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

4.4 Quasi-Newton-Verfahren

Im folgenden setzen wir stets voraus, dass $f : D \subset \mathbb{R}^n \to \mathbb{R}$ stetig differenzierbar ist. Beim *Gradientenverfahren* ist die Idee, 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. Dieses Vorgehen haben wir bereits in Abschnitt 3.1 untersucht. Viel besser als das Gradientenverfahren, welches nur mit linearer Rate konvergiert (falls es überhaupt gegen ein Minimum konvergiert), ist das *Newton-Verfahren*, da dies im Fall der Konvergenz quadratisch konvergiert.

Beim Newton-Verfahren ist das Update $\mathbf{x}_{k+1} := \mathbf{x}_k + \mathbf{d}_k$ 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, $\left(\nabla^2 f(\mathbf{x}_k)\right)^{-1}$ 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*

$$\mathbf{H}_{k+1}\left\{\nabla f(\mathbf{x}_{k+1}) - \nabla f(\mathbf{x}_k)\right\} = \mathbf{x}_{k+1} - \mathbf{x}_k \tag{4.5}$$

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 $\left(\nabla^2 f(\mathbf{x}_k)\right)^{-1}$ 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 $q(\mathbf{x}) = \frac{1}{2}\mathbf{x}^\intercal \mathbf{A}\mathbf{x} + \mathbf{b}^\intercal \mathbf{x} + c$ mit positiv definiter Matrix $\mathbf{A}$ gilt (4.5) wegen $\nabla q(\mathbf{x}) = \mathbf{A}\mathbf{x} + \mathbf{b}$ sogar exakt. Ferner erscheint es sinnvoll, als $\mathbf{H}_k$ nur symmetrische und 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

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

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

$$\begin{aligned} \mathbf{H}_{k+1} &:= \Phi(\mathbf{H}_k, \mathbf{p}_k, \mathbf{q}_k, \gamma_k, \nu_k), \\ \Phi(\mathbf{H}, \mathbf{p}, \mathbf{q}, \gamma, \nu) &:= \gamma \mathbf{H} + \left(1 + \gamma\nu \frac{\mathbf{q}^\intercal \mathbf{H}\mathbf{q}}{\mathbf{p}^\intercal \mathbf{q}}\right) \frac{\mathbf{p}\mathbf{p}^\intercal}{\mathbf{p}^\intercal \mathbf{q}} \\ &\quad - \gamma \frac{1-\nu}{\mathbf{q}^\intercal \mathbf{H}\mathbf{q}} \mathbf{H}\mathbf{q}\mathbf{q}^\intercal \mathbf{H} - \frac{\gamma\nu}{\mathbf{p}^\intercal \mathbf{q}} (\mathbf{p}\mathbf{q}^\intercal \mathbf{H} + \mathbf{H}\mathbf{q}\mathbf{p}^\intercal). \end{aligned} \tag{4.6}$$

Die Update-Funktion Φ ist nur für $\mathbf{p}^\intercal \mathbf{q} \neq 0$ und $\mathbf{q}^\intercal \mathbf{H}\mathbf{q} \neq 0$ 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 (4.6) 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 Broyden, Fletcher, Goldfarb und Shanno (*BFGS-Verfahren*).
3. $\gamma_k \equiv 1$, $\nu_k = \mathbf{p}_k^\intercal \mathbf{q}_k / (\mathbf{p}_k^\intercal \mathbf{q}_k - \mathbf{p}_k^\intercal \mathbf{H}_k \mathbf{q}_k)$: *symmetrisches Rang-1-Verfahren von Broyden*.

Letzteres Verfahren ist nur für $\mathbf{p}_k^\mathsf{T}\mathbf{q}_k \neq \mathbf{p}_k^\mathsf{T}\mathbf{H}_k\mathbf{q}_k$ 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 4.11). Setzt man den gewählten Wert in (4.6) ein, erhält man für $\mathbf{H}_k$ eine Rekursionformel, die den Namen Rang-1-Verfahren erklärt:

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

Algorithmus 4.10 (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 (4.6)

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

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

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

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

Wir setzen für das folgende

$$\sigma_k < 1 \tag{4.7}$$

voraus. Falls $\nabla f(\mathbf{x}_k) \neq \mathbf{0}$ und $\mathbf{H}_k$ positiv definit ist, folgt aus (4.7) $\alpha_k > 0$ und deshalb

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

also auch $\mathbf{q}_k \neq \mathbf{0}$ und $\mathbf{q}_k^\mathsf{T}\mathbf{H}_k\mathbf{q}_k > 0$. Die Matrix $\mathbf{H}_{k+1}$ ist damit durch (4.6) wohldefiniert.

Die Forderung (4.7) kann nur dann nicht erfüllt werden, wenn

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

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

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

so dass $f(\mathbf{x}_k + \alpha\mathbf{d}_k)$ für $\alpha \to \infty$ nicht nach unten beschränkt ist. Die Forderung (4.7) bedeutet also keine wesentliche Einschränkung. Damit ist bereits der erste Teil des folgenden Satzes gezeigt, der besagt, dass das Quasi-Newton-Verfahren 4.10 unsere oben aufgestellten Forderungen erfüllt.

Satz 4.11 *Falls im Quasi-Newton-Verfahren 4.10 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. Seien $\mathbf{y} \in \mathbb{R}^n \setminus \{\mathbf{0}\}$ ein beliebiger Vektor und $\mathbf{H}_k = \mathbf{L}\mathbf{L}^\mathsf{T}$ die Cholesky-Zerlegung von $\mathbf{H}_k$. Mit Hilfe der Vektoren

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

lässt sich $\mathbf{y}^\mathsf{T}\mathbf{H}_{k+1}\mathbf{y}$ wegen (4.6) so schreiben:

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

Die Cauchy-Schwarzsche Ungleichung ergibt

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

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

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

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 (4.6). ♠

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 stets exakt ist. 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 4.12 *Sei*

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

eine quadratische Funktion mit einer positiv definiten Matrix $\mathbf{A} \in \mathbb{R}^{n\times n}$. *Wendet man das Quasi-Newton-Verfahren 4.10 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 $\mathbf{p}_k^\intercal \mathbf{q}_k > 0$ *und* $\mathbf{p}_k^\intercal \mathbf{q}_\ell = \mathbf{p}_k^\intercal \mathbf{A}\mathbf{p}_\ell = 0$ *für alle* $0 \leq k \neq \ell < m$. *Die Vektoren* $\mathbf{p}_k$ *sind demnach* $\mathbf{A}$**-konjugiert.**

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

(iv.) Es ist $\mathbf{H}_\ell \mathbf{q}_k = \gamma_{k,\ell}\mathbf{p}_k$ *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 $\nabla f(\mathbf{x}_m)^\intercal \mathbf{H}_m \nabla f(\mathbf{x}_m) > 0$. Weil exakt minimiert wird, ist α_m die Nullstelle von

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

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

$$\nabla f(\mathbf{x}_{m+1})^\intercal \mathbf{p}_m = \alpha_m \nabla f(\mathbf{x}_{m+1})^\intercal \mathbf{d}_m = 0. \tag{4.8}$$

Deshalb gilt

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

und folglich ist $\mathbf{H}_{m+1}$ nach Satz 4.11 positiv definit. Weiter ist für $k < m$ wegen $\mathbf{A}\mathbf{p}_k = \mathbf{q}_k$

$$\mathbf{p}_k^\intercal \mathbf{q}_m = \mathbf{p}_k^\intercal \mathbf{A}\mathbf{p}_m = \mathbf{q}_k^\intercal \mathbf{p}_m = -\alpha_m \mathbf{q}_k^\intercal \mathbf{H}_m \nabla f(\mathbf{x}_m) \overset{(iv.)}{=} -\alpha_m \gamma_{k,m} \mathbf{p}_k^\intercal \nabla f(\mathbf{x}_m) \overset{(iii.)}{=} 0. \quad (4.9)$$

Das ist der Induktionsschritt für Aussage (*ii.*).

Weiter gilt für $k < m$

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

nach dem eben bewiesenen und Aussage (*iii.*). Zusammen mit (4.8) ergibt dies Aussage (*iii.*) für $m+1$.

Den Induktionsschritt für Aussage (*iv.*) sieht man wie folgt ein. Anhand von (4.6) 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 $k < m$

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

so dass für $k < m$ aus (4.6) 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 $\mathbf{p}_0, \mathbf{p}_1, \dots, \mathbf{p}_{m-1}$ linear unabhängig sind. Aus $\mathbf{0} = \sum_{\ell=0}^{m-1} \lambda_\ell \mathbf{p}_\ell$ folgt nämlich durch Multiplikation mit $\mathbf{p}_k^\intercal \mathbf{A}$, $k = 0, 1, \dots, m-1$, wegen Aussage (*ii.*) $\lambda_k \mathbf{p}_k^\intercal \mathbf{A}\mathbf{p}_k = 0$, 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 $m \leq n$ 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 $\mathbf{H}_n\mathbf{Q} = \mathbf{PD}$ für die Matrizen

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

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

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

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 4.12 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 Vorraussetzungen 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*. ♦

4.5 Nichtlineares CG-Verfahren

In Anlehnung an das CG-Verfahren aus Abschnitt 2.3 ist das *nichtlineare CG-Verfahren* zur Lösung von nichtlinearen Optimierungsproblemen $f(\mathbf{x}) \to \min$ definiert.

Algorithmus 4.13 (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

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

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

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

④ erhöhe $k := k + 1$ und gehe nach ②

Bemerkung Ist $f(\mathbf{x}) = \frac{1}{2}\mathbf{x}^\mathsf{T}\mathbf{A}\mathbf{x} - \mathbf{b}^\mathsf{T}\mathbf{x} + c$ 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 $\nabla f(\mathbf{x}_k)^\mathsf{T}\nabla f(\mathbf{x}_{k+1}) = \mathbf{r}_k^\mathsf{T}\mathbf{r}_{k+1} = 0$. ♦

Lemma 4.14 *Die Funktion $f : D \subset \mathbb{R}^n \to \mathbb{R}$ sei gleichmäßig konvex. Weiter sei f differenzierbar mit Lipschitz-stetigem Gradienten:*

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

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

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

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

$$0 = \nabla f(\mathbf{x}_\ell + \alpha_\ell \mathbf{d}_\ell)^\mathsf{T} \mathbf{d}_\ell = \nabla f(\mathbf{x}_{\ell+1})^\mathsf{T} \mathbf{d}_\ell \quad \text{für alle } \ell \leq k. \tag{4.10}$$

Daher können wir den Nenner in

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

folgendemaßen umformen:

$$\begin{aligned} \|\nabla f(\mathbf{x}_k)\|_2^2 &= (\beta_{k-1} \mathbf{d}_{k-1} - \mathbf{d}_k)^\mathsf{T} \nabla f(\mathbf{x}_k) \\ &\overset{(4.10)}{=} -\mathbf{d}_k^\mathsf{T} \nabla f(\mathbf{x}_k) \\ &= -\frac{1}{\alpha_k} (\mathbf{x}_{k+1} - \mathbf{x}_k)^\mathsf{T} \nabla f(\mathbf{x}_k). \end{aligned}$$

Die gleichmäßigen Konvexität impliziert

$$-(\mathbf{x}_{k+1} - \mathbf{x}_k)^\mathsf{T} \nabla f(\mathbf{x}_k) \geq \mu \|\mathbf{x}_{k+1} - \mathbf{x}_k\|_2^2 + \underbrace{f(\mathbf{x}_k) - f(\mathbf{x}_{k+1})}_{\geq 0} \geq \mu \|\mathbf{x}_{k+1} - \mathbf{x}_k\|_2^2,$$

das heißt

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

Mit Hilfe der Lipschitz-Bedingung erhalten wir daraus

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

Dies führt auf

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

woraus dann die Behauptung folgt

$$\begin{aligned} -\frac{\mathbf{d}_{k+1}^{\mathsf{T}}\nabla f(\mathbf{x}_{k+1})}{\|\mathbf{d}_{k+1}\|_2\|\nabla f(\mathbf{x}_{k+1})\|_2} &= \frac{\{\nabla f(\mathbf{x}_{k+1}) - \beta_k\mathbf{d}_k\}^{\mathsf{T}}\nabla f(\mathbf{x}_{k+1})}{\|\mathbf{d}_{k+1}\|_2\|\nabla f(\mathbf{x}_{k+1})\|_2} \\ &\overset{(4.10)}{=} \frac{\|\nabla f(\mathbf{x}_{k+1})\|_2^2}{\|\mathbf{d}_{k+1}\|_2\|\nabla f(\mathbf{x}_{k+1})\|_2} \\ &\ge \frac{\mu}{\mu + L}. \end{aligned}$$

♠

Bemerkung Die geometrische Interpretation von Lemma 4.14 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 4.15 *Die Funktion $f : D \subset \mathbb{R}^n \to \mathbb{R}$ sei gleichmäßig konvex. Weiter sei f differenzierbar mit Lipschitz-stetigem Gradienten:*

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

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

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

Beweis. Aufgrund der Minimierungsbedingung gilt für ein $\gamma > 0$

$$\begin{aligned} f(\mathbf{x}_{k+1}) &\le f(\mathbf{x}_k + \gamma\mathbf{d}_k) \\ &= f(\mathbf{x}_k) + \gamma\int_0^1 \nabla f(\mathbf{x}_k + \gamma t\mathbf{d}_k)^{\mathsf{T}}\mathbf{d}_k\,\mathrm{d}t \\ &= f(\mathbf{x}_k) + \gamma\nabla f(\mathbf{x}_k)^{\mathsf{T}}\mathbf{d}_k + \gamma\int_0^1 \left\{\nabla f(\mathbf{x}_k + \gamma t\mathbf{d}_k) - \nabla f(\mathbf{x}_k)\right\}^{\mathsf{T}}\mathbf{d}_k\,\mathrm{d}t \end{aligned}$$

und daher

$$f(\mathbf{x}_{k+1}) \le f(\mathbf{x}_k) + \gamma \nabla f(\mathbf{x}_k)^\mathsf{T} \mathbf{d}_k + \gamma \int_0^1 \underbrace{\|\nabla f(\mathbf{x}_k + \gamma t \mathbf{d}_k) + \nabla f(\mathbf{x}_k)\|_2}_{\le t\gamma L \|\mathbf{d}_k\|_2} \|\mathbf{d}_k\|_2 \, \mathrm{d}t$$

$$\le f(\mathbf{x}_k) + \gamma \nabla f(\mathbf{x}_k)^\mathsf{T} \mathbf{d}_k + \gamma^2 \frac{L}{2} \|\mathbf{d}_k\|_2^2.$$

Für die Wahl

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

folgern wir mit Lemma 4.14

$$f(\mathbf{x}_{k+1}) - f(\mathbf{x}^\star) \le f(\mathbf{x}_k) - f(\mathbf{x}^\star) - \frac{(\nabla f(\mathbf{x}_k)^\mathsf{T} \mathbf{d}_k)^2}{2L \|\mathbf{d}_k\|_2^2}$$

$$\le f(\mathbf{x}_k) - f(\mathbf{x}^\star) - \frac{\mu^2}{2L(\mu + L)^2} \|\nabla f(\mathbf{x}_k) - \underbrace{\nabla f(\mathbf{x}^\star)}_{=\mathbf{0}}\|_2^2.$$

Aus der gleichmäßigen Konvexität ergibt sich

$$\mu \|\mathbf{x}_k - \mathbf{x}^\star\|_2^2 \le \left\{\nabla f(\mathbf{x}_k) - \nabla f(\mathbf{x}^\star)\right\}^\mathsf{T} (\mathbf{x}_k - \mathbf{x}^\star) \le \|\nabla f(\mathbf{x}_k) - \nabla f(\mathbf{x}^\star)\|_2 \|\mathbf{x}_k - \mathbf{x}^\star\|_2,$$

während die Lipschitz-Stetigkeit impliziert

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

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

Bemerkungen

1. Das Verfahren von Polak und Ribière konvergiert im allgemeinen schneller als das Verfahren von Fletcher und Reeves.
2. In der Praxis verwendet man Restarts: Wird der Winkel zwischen dem Antigradienten und der Suchrichtung zu groß, etwa

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

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

4.6 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 4.13 durchgeführt wird. Eine direkt zu implementierende Schrittweitensteuerung stellen wir im folgenden modifizierten Verfahren von Polak und Ribière vor.

Algorithmus 4.16 (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

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

③ berechne

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

④ ist eine der Bedingungen

$$f(\mathbf{x}_{k+1}) \le f(\mathbf{x}_k) - \sigma \alpha_k^2 \|\mathbf{d}_k\|_2^2 \tag{4.11}$$

$$-\overline{\gamma}\|\nabla f(\mathbf{x}_{k+1})\|_2^2 \le \nabla f(\mathbf{x}_{k+1})^\intercal \mathbf{d}_{k+1} \le -\underline{\gamma}\|\nabla f(\mathbf{x}_{k+1})\|_2^2 \tag{4.12}$$

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 4.17 *Ist $f : \mathbb{R}^n \to \mathbb{R}$ stetig differenzierbar, so ist Algorithmus 4.16 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 ① im Fall $k = 0$ beziehungsweise aus (4.12) 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 $\nabla f(\mathbf{x}_k)^\mathsf{T}\mathbf{d}_k < 0$ ist. Als erstes stellen wir fest, dass die Bedingung (4.11) wegen

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

nach endlich vielen erfolglosen Schritten der Liniensuche immer erfüllt ist.

Als nächstes zeigen wir, dass die Bedingung (4.12) 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

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

zumindest eine der beiden Bedingungen

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

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

Damit ist gezeigt, dass Algorithmus 4.16 wohldefiniert ist, sofern die Abstiegsbedingung $\nabla f(\mathbf{x}_k)^\mathsf{T}\mathbf{d}_k < 0$ 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 (4.12). ♠

Lemma 4.18 *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 $N := \{\mathbf{x} \in \mathbb{R}^n : f(\mathbf{x}) \leq f(\mathbf{x}_0)\}$ 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*

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

für alle $k \in \mathbb{N}_0$.

Beweis.

(i.) Diese Aussage ergibt sich unmittelbar aus der Bedingung (4.11).

(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 (4.11) 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 4.16 folgt

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

für alle $k \in \mathbb{N}_0$.

(v.) Zum Nachweis dieser Aussage führen wir eine Fallunterscheidung durch.

Fall 1: $\alpha_k = |\nabla f(\mathbf{x}_k)^\mathsf{T} \mathbf{d}_k| / \|\mathbf{d}_k\|_2^2$.

Dann ist offensichtlich

$$\alpha_k \geq \frac{|\nabla f(\mathbf{x}_k)^\mathsf{T} \mathbf{d}_k|}{\|\mathbf{d}_k\|_2^2}. \tag{4.13}$$

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

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 (4.11) oder (4.12) 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 (4.11).

Dann gilt

$$f(\mathbf{z}_k) > f(\mathbf{x}_k) - \sigma(2\alpha_k)^2\|\mathbf{d}_k\|_2^2. \tag{4.14}$$

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

$$f(\mathbf{z}_k) = f(\mathbf{x}_k) + \nabla f(\boldsymbol{\xi}_k)^\mathsf{T}(\mathbf{z}_k - \mathbf{x}_k) = f(\mathbf{x}_k) + 2\alpha_k\nabla f(\boldsymbol{\xi}_k)^\mathsf{T}\mathbf{d}_k. \tag{4.15}$$

Aus (4.14) und (4.15) folgt daher

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

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

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

Dies liefert unmittelbar

$$\alpha_k \geq \frac{1}{2(L+\sigma)}\frac{|\nabla f(\mathbf{x}_k)^\mathsf{T}\mathbf{d}_k|}{\|\mathbf{d}_k\|_2^2}. \tag{4.16}$$

Fall 2B: Der Punkt $\mathbf{z}_k \in D$ verletzt die linke Ungleichung in (4.12).

Wegen

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

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 (4.12) genügt, folgt daher nach kurzer Rechnung

$$\alpha_k \geq \frac{(\overline{\gamma}-1)\|\nabla f(\mathbf{x}_k)\|_2^2}{2L\|\mathbf{d}_k\|_2^2} \geq \frac{\overline{\gamma}-1}{2\overline{\gamma}L}\frac{|\nabla f(\mathbf{x}_k)^\mathsf{T}\mathbf{d}_k|}{\|\mathbf{d}_k\|_2^2}. \tag{4.17}$$

Fall 2C: Der Punkt $\mathbf{z}_k \in D$ verletzt die linke Ungleichung in (4.12). Es ist

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

Analog zum *Fall 2B* erhält man hieraus

$$\alpha_k \geq \frac{1-\underline{\gamma}}{2\overline{\gamma}L}\frac{|\nabla f(\mathbf{x}_k)^\mathsf{T}\mathbf{d}_k|}{\|\mathbf{d}_k\|_2^2}. \tag{4.18}$$

Wegen (4.13), (4.16), (4.17), (4.18) 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 4.19 *Unter den Voraussetzungen von Lemma 4.18 gilt für die Iterierten $\{\mathbf{x}_k\}_{k\geq 0}$ des modifizierten Verfahren von Polak und Ribière*

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

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 4.18 (*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 4.18 (*iii.*) ergibt sich hieraus

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

und weiter aus Lemma 4.18 (*v.*)

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

Die rechte Ungleichung in (4.12) liefert daher

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

Weil nach Lemma 4.18 (*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. ♠

4.7 Projiziertes Gradientenverfahren

Bislang haben wir uns mit gradientenbasierten Verfahren für Optimierungsprobleme ohne Nebenbedingung beschäftigt. In diesem Abschnitt wollen wir nun die Situation einer vorgegebenen Nebenbedingung in Betracht ziehen. Dazu seien $f : \mathbb{R}^n \to \mathbb{R}$ eine stetig differenzierbare Funktion und $K \subset \mathbb{R}^n$ eine abgeschlossene und konvexe Menge. Wir betrachten folgendes Optimierungsproblem unter Nebenbedingungen

$$\text{minimiere } f(\mathbf{x}) \text{ unter der Nebenbedingung } \mathbf{x} \in K. \tag{4.19}$$

Da das Minimum auch auf dem Rand von K liegen kann, lautet im Fall von (4.19) die notwendige Optimalitätsbedingung für ein Minimum $\mathbf{x}^\star \in K$

$$\nabla f(\mathbf{x}^\star)^\intercal (\mathbf{x} - \mathbf{x}^\star) \geq 0 \text{ für alle } \mathbf{x} \in K. \tag{4.20}$$

Grundlage des *projizierten Gradientenverfahrens* ist die orthogonale Projektion auf die zulässige Menge.

Definition 4.20 Es sei $K \subset \mathbb{R}^n$ eine abgeschlossene, konvexe Menge. Dann ist die **orthogonale Projektion** $\mathbf{P}_K : \mathbb{R}^n \to K$ definiert durch die Bedingung

$$\|\mathbf{P}_K(\mathbf{x}) - \mathbf{x}\|_2 = \min_{\mathbf{y} \in K} \|\mathbf{y} - \mathbf{x}\|_2.$$

Der Punkt $\mathbf{P}_K(\mathbf{x}) \in K$ besitzt also die Eigenschaft, den kürzesten Abstand zu einem gegebenen Punkt $\mathbf{x} \in \mathbb{R}^n$ zu besitzen.

Die Grundversion des projizierten Gradientenverfahren ist im folgenden Algorithmus beschrieben:

Algorithmus 4.21 (projiziertes Gradientenverfahren)
input: Funktion $f : \mathbb{R}^n \to \mathbb{R}$, konvexe zulässige Menge $K \subset \mathbb{R}^n$ und Startnäherung $\mathbf{x}_0 \in K$
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

$$f\big(\mathbf{P}_K(\mathbf{x}_k + \alpha_k \mathbf{d}_k)\big) > f(\mathbf{x}_k) - \sigma \mathbf{d}_k^{\mathsf{T}} \big(\mathbf{P}_K(\mathbf{x}_k + \alpha_k \mathbf{d}_k) - \mathbf{x}_k\big) \tag{4.21}$$

setze $\alpha_k := \alpha_k / 2$

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

⑤ erhöhe $k := k + 1$ und gehe nach ②

Für den Fall der Minimierung ohne Nebenbedingungen, das heißt $K = \mathbb{R}^n$, stellt obiger Algorithmus das klassische Gradientverfahren dar. Insbesondere geht die Bedingung an die Reduktion des Funktionals über in die Armijo-Goldstein-Bedingung

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

Man kann zeigen, dass für das projizierte Gradientenverfahren ein $\alpha_k > 0$ existiert, für das die Reduktionsbedingung erfüllt ist.

Lemma 4.22 *Die orthogonale Projektion $\mathbf{P}_K$ besitzt die folgenden Eigenschaften:*

(i.) Es gilt $\big(\mathbf{P}_K(\mathbf{x})-\mathbf{x}\big)^\mathsf{T}\big(\mathbf{P}_K(\mathbf{x})-\mathbf{y}\big)\le 0$ für alle $\mathbf{x}\in\mathbb{R}^n$, $\mathbf{y}\in K$.

(ii.) Es gilt $\big(\mathbf{P}_K(\mathbf{y})-\mathbf{P}_K(\mathbf{x})\big)^\mathsf{T}(\mathbf{y}-\mathbf{x})\ge\|\mathbf{P}_K(\mathbf{y})-\mathbf{P}_K(\mathbf{x})\|_2^2\ge 0$ für alle $\mathbf{x},\mathbf{y}\in\mathbb{R}^n$, das heißt, $\mathbf{P}_K$ ist monoton.

(iii.) Es gilt $\|\mathbf{P}_K(\mathbf{y})-\mathbf{P}_K(\mathbf{x})\|_2\le\|\mathbf{y}-\mathbf{x}\|_2$ für alle $\mathbf{x},\mathbf{y}\in\mathbb{R}^n$, das heißt, $\mathbf{P}_K$ ist nicht expandierend.

Beweis. (*i.*) Wegen der Konvexität folgt aus $\mathbf{y}\in K$ auch $\widehat{\mathbf{y}}:=(1-t)\mathbf{P}_K(\mathbf{x})+t\mathbf{y}\in K$ für alle $t\in[0,1]$. Aus

$$\begin{aligned}\|\widehat{\mathbf{y}}-\mathbf{x}\|_2^2&=\|\widehat{\mathbf{y}}-\mathbf{P}_K(\mathbf{x})+\mathbf{P}_K(\mathbf{x})-\mathbf{x}\|_2^2\\&=\|\widehat{\mathbf{y}}-\mathbf{P}_K(\mathbf{x})\|_2^2+\|\mathbf{P}_K(\mathbf{x})-\mathbf{x}\|_2^2-2\big(\mathbf{P}_K(\mathbf{x})-\mathbf{x}\big)^\mathsf{T}\big(\mathbf{P}_K(\mathbf{x})-\widehat{\mathbf{y}}\big)\end{aligned}$$

folgt aufgrund der Minimierungseigenschaft von $\mathbf{P}_K$, dass

$$\|\widehat{\mathbf{y}}-\mathbf{P}_K(\mathbf{x})\|_2^2-2\big(\mathbf{P}_K(\mathbf{x})-\mathbf{x}\big)^\mathsf{T}\big(\mathbf{P}_K(\mathbf{x})-\widehat{\mathbf{y}}\big)=\|\widehat{\mathbf{y}}-\mathbf{x}\|_2^2-\|\mathbf{P}_K(\mathbf{x})-\mathbf{x}\|_2^2\ge 0.$$

Einsetzen von $\mathbf{P}_K(\mathbf{x})-\widehat{\mathbf{y}}=t\big(\mathbf{P}_K(\mathbf{x})-\mathbf{y}\big)$ führt auf

$$t^2\|\mathbf{y}-\mathbf{P}_K(\mathbf{x})\|_2^2-2t\big(\mathbf{P}_K(\mathbf{x})-\mathbf{x}\big)^\mathsf{T}\big(\mathbf{P}_K(\mathbf{x})-\mathbf{y}\big)\ge 0,$$

was für $t\to 0$ die gewünschte Aussage liefert.

(*ii.*) Die bereits bewiesene Aussage (*i.*) impliziert

$$\begin{aligned}\big(\mathbf{P}_K(\mathbf{x})-\mathbf{x}\big)^\mathsf{T}\big(\mathbf{P}_K(\mathbf{x})-\mathbf{P}_K(\mathbf{y})\big)&\le 0,\\\big(\mathbf{P}_K(\mathbf{y})-\mathbf{y}\big)^\mathsf{T}\big(\mathbf{P}_K(\mathbf{y})-\mathbf{P}_K(\mathbf{x})\big)&\le 0.\end{aligned}$$

Zusammen führt dies auf

$$\big(\mathbf{P}_K(\mathbf{y})-\mathbf{y}+\mathbf{x}-\mathbf{P}_K(\mathbf{x})\big)^\mathsf{T}\big(\mathbf{P}_K(\mathbf{y})-\mathbf{P}_K(\mathbf{x})\big)\le 0,$$

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 $\mathbf{x}_k=\mathbf{P}_K(\mathbf{x}_k)$ für die neue Iterierte $\mathbf{x}_{k+1}=\mathbf{P}_K\big(\mathbf{x}_k-\alpha_k\nabla f(\mathbf{x}_k)\big)$ des projizierten Gradientenverfahrens, dass

$$\|\mathbf{x}_{k+1}-\mathbf{x}_k\|_2^2\le(\mathbf{x}_{k+1}-\mathbf{x}_k)^\mathsf{T}\big(\mathbf{x}_k-\alpha_k\nabla f(\mathbf{x}_k)-\mathbf{x}_k\big).$$

Wir erhalten daher

$$-\nabla f(\mathbf{x}_k)^\mathsf{T}(\mathbf{x}_{k+1} - \mathbf{x}_k) \geq \frac{1}{\alpha_k}\|\mathbf{x}_{k+1} - \mathbf{x}_k\|_2^2. \tag{4.22}$$

Dies bedeutet, dass durch die Abstiegsbedingung (4.21) des Algorithmus 4.21 tatsächlich $f(\mathbf{x}_{k+1}) < f(\mathbf{x}_k)$ erreicht wird. ♦

Lemma 4.23 *Für beliebige $\mathbf{x} \in \mathbb{R}^n$ und $\mathbf{d} \in \mathbb{R}^n$ ist die Funktion*

$$\varphi(\alpha) := \frac{\|\mathbf{P}_K(\mathbf{x} + \alpha\mathbf{d}) - \mathbf{x}\|_2}{\alpha}$$

für alle $\alpha > 0$ monoton fallend.

Beweis. (*i.*) Für $0 < \alpha < \beta$ setzen wir

$$\mathbf{u} := \mathbf{P}_K(\mathbf{x} + \alpha\mathbf{d}) - \mathbf{x}, \quad \mathbf{v} := \mathbf{P}_K(\mathbf{x} + \beta\mathbf{d}) - \mathbf{x}$$

und erhalten unter Verwendung von Lemma 4.22 (*i.*)

$$\begin{aligned}\mathbf{u}^\mathsf{T}(\mathbf{u} - \mathbf{v}) &= \{\mathbf{P}_K(\mathbf{x} + \alpha\mathbf{d}) - (\mathbf{x} + \alpha\mathbf{d}) + \alpha\mathbf{d}\}^\mathsf{T}\{\mathbf{P}_K(\mathbf{x} + \alpha\mathbf{d}) - \mathbf{P}_K(\mathbf{x} + \beta\mathbf{d})\} \\ &\leq \alpha\mathbf{d}^\mathsf{T}\{\mathbf{P}_K(\mathbf{x} + \alpha\mathbf{d}) - \mathbf{P}_K(\mathbf{x} + \beta\mathbf{d})\}\end{aligned}$$

und analog

$$\mathbf{v}^\mathsf{T}(\mathbf{v} - \mathbf{u}) \leq \beta\mathbf{d}^\mathsf{T}\{\mathbf{P}_K(\mathbf{x} + \beta\mathbf{d}) - \mathbf{P}_K(\mathbf{x} + \alpha\mathbf{d})\}.$$

Zusammen ergibt dies

$$\frac{\mathbf{u}^\mathsf{T}(\mathbf{u} - \mathbf{v})}{\alpha} \leq \frac{\mathbf{v}^\mathsf{T}(\mathbf{u} - \mathbf{v})}{\beta}. \tag{4.23}$$

(*ii.*) Weiter erhalten wir mit Lemma 4.22 (*ii.*)

$$\begin{aligned}\mathbf{u}^\mathsf{T}(\mathbf{u} - \mathbf{v}) &\leq \alpha\mathbf{d}^\mathsf{T}\{\mathbf{P}_K(\mathbf{x} + \alpha\mathbf{d}) - \mathbf{P}_K(\mathbf{x} + \beta\mathbf{d})\} \\ &= -\frac{\alpha}{\beta - \alpha}(\alpha\mathbf{d} - \beta\mathbf{d})^\mathsf{T}\{\mathbf{P}_K(\mathbf{x} + \beta\mathbf{d}) - \mathbf{P}_K(\mathbf{x} + \alpha\mathbf{d})\} \\ &\leq -\frac{\alpha}{\beta - \alpha}\|\mathbf{P}_K(\mathbf{x} + \beta\mathbf{d}) - \mathbf{P}_K(\mathbf{x} + \alpha\mathbf{d})\|_2^2 \\ &\leq 0.\end{aligned}$$

Aus der Cauchy-Schwarzschen Ungleichung ergibt sich

$$\mathbf{u}^\mathsf{T}\mathbf{v}(\|\mathbf{u}\|_2 + \|\mathbf{v}\|_2) \leq \|\mathbf{u}\|_2\|\mathbf{v}\|_2(\|\mathbf{u}\|_2 + \|\mathbf{v}\|_2),$$

woraus

$$\|\mathbf{u}\|_2\mathbf{v}^\mathsf{T}(\mathbf{u}-\mathbf{v}) = \|\mathbf{u}\|_2(\mathbf{u}^\mathsf{T}\mathbf{v} - \|\mathbf{v}\|_2^2) \leq \|\mathbf{v}\|_2(\|\mathbf{u}\|_2^2 - \mathbf{u}^\mathsf{T}\mathbf{v}) = \|\mathbf{v}\|_2\mathbf{u}^\mathsf{T}(\mathbf{u}-\mathbf{v}) \quad (4.24)$$

folgt.

(*iii.*) Wir unterscheiden nun zwei Fälle: Für $\mathbf{u}^\mathsf{T}(\mathbf{u}-\mathbf{v}) = 0$ gilt $\mathbf{P}_K(\mathbf{x}+\alpha\mathbf{d}) = \mathbf{P}_K(\mathbf{x}+\beta\mathbf{d})$ 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 $\mathbf{u}^\mathsf{T}(\mathbf{u}-\mathbf{v}) < 0$ folgt aus (4.23)

$$\frac{\beta}{\alpha} \geq \frac{\mathbf{v}^\mathsf{T}(\mathbf{u}-\mathbf{v})}{\mathbf{u}^\mathsf{T}(\mathbf{u}-\mathbf{v})}$$

und aus (4.24)

$$\|\mathbf{v}\|_2 \leq \|\mathbf{u}\|_2\frac{\mathbf{v}^\mathsf{T}(\mathbf{u}-\mathbf{v})}{\mathbf{u}^\mathsf{T}(\mathbf{u}-\mathbf{v})}.$$

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 4.24 *Die Funktion $f : \mathbb{R}^n \to \mathbb{R}$ sei auf K stetig differenzierbar und nach unten beschränkt. Weiter sei ∇f auf K gleichmäßig stetig. Dann gilt für die Iterierten $\{\mathbf{x}_k\}_{k\in\mathbb{N}}$ des projizierten Gradientenverfahrens*

$$\lim_{k\to\infty} \frac{\|\mathbf{x}_{k+1} - \mathbf{x}_k\|_2}{\alpha_k} = 0.$$

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

$$\frac{\|\mathbf{x}_{k_\ell+1} - \mathbf{x}_{k_\ell}\|_2^2}{\alpha_{k_\ell}} \geq \varepsilon \max\{\varepsilon\alpha_{k_\ell}, \|\mathbf{x}_{k_\ell+1} - \mathbf{x}_{k_\ell}\|_2\}. \tag{4.25}$$

Da die Folge $\{f(\mathbf{x}_{k_\ell})\}_{\ell\in\mathbb{N}}$ monoton fallend und nach unten beschränkt ist, folgt aus der Abstiegsbedingung (4.21) des projizierten Gradientenverfahrens

$$\lim_{\ell\to\infty} \nabla f(\mathbf{x}_{k_\ell})^\mathsf{T}(\mathbf{x}_{k_\ell+1} - \mathbf{x}_{k_\ell}) = 0,$$

was wiederum gemäß (4.22)

$$\lim_{\ell\to\infty} \frac{\|\mathbf{x}_{k_\ell+1} - \mathbf{x}_{k_\ell}\|_2^2}{\alpha_{k_\ell}} = 0 \tag{4.26}$$

nach sich zieht. Aufgrund von (4.25) erhalten wir hieraus

$$\lim_{\ell\to\infty} \alpha_{k_\ell} = 0 \quad \text{und} \quad \lim_{\ell\to\infty} \|\mathbf{x}_{k_\ell+1} - \mathbf{x}_{k_\ell}\|_2 = 0.$$

Für $\mathbf{y}_{k_\ell+1} := \mathbf{P}_K(\mathbf{x}_{k_\ell} + 2\alpha_{k_\ell}\mathbf{d}_{k_\ell})$ gilt aufgrund der algorithmischen Umsetzung des projizierten Gradientenverfahrens

$$f(\mathbf{y}_{k_\ell+1}) > f(\mathbf{x}_{k_\ell}) + \sigma\nabla f(\mathbf{x}_{k_\ell})^\mathsf{T}(\mathbf{y}_{k_\ell+1} - \mathbf{x}_{k_\ell}),$$

also auch

$$f(\mathbf{x}_{k_\ell}) - f(\mathbf{y}_{k_\ell+1}) < \sigma\nabla f(\mathbf{x}_{k_\ell})^\mathsf{T}(\mathbf{x}_{k_\ell} - \mathbf{y}_{k_\ell+1}). \tag{4.27}$$

Aus Lemma 4.23 folgt

$$\begin{aligned}\frac{\|\mathbf{x}_{k_\ell+1} - \mathbf{x}_{k_\ell}\|_2^2}{\alpha_{k_\ell}} &\geq \|\mathbf{x}_{k_\ell+1} - \mathbf{x}_{k_\ell}\|_2 \frac{\|\mathbf{y}_{k_\ell+1} - \mathbf{x}_{k_\ell}\|_2}{2\alpha_{k_\ell}} \\ &\geq \varepsilon\alpha_{k_\ell} \frac{\|\mathbf{y}_{k_\ell+1} - \mathbf{x}_{k_\ell}\|_2}{2\alpha_{k_\ell}} = \frac{\varepsilon}{2}\|\mathbf{y}_{k_\ell+1} - \mathbf{x}_{k_\ell}\|_2.\end{aligned}$$

Weiter ergibt sich mit Lemma 4.22 (*ii.*)

$$(\mathbf{x}_{k_\ell+1} - \mathbf{x}_{k_\ell})^\mathsf{T}\Big\{ \underbrace{\mathbf{x}_{k_\ell} - \alpha_{k_\ell}\nabla f(\mathbf{x}_{k_\ell}) - \mathbf{x}_{k_\ell}}_{=-\alpha_{k_\ell}\nabla f(\mathbf{x}_{k_\ell})} \Big\} \geq \|\mathbf{x}_{k_\ell+1} - \mathbf{x}_{k_\ell}\|_2^2,$$

$$(\mathbf{y}_{k_\ell+1} - \mathbf{x}_{k_\ell+1})^\mathsf{T}\Big\{ \underbrace{\mathbf{x}_{k_\ell} - 2\alpha_{k_\ell}\nabla f(\mathbf{x}_{k_\ell}) - \big(\mathbf{x}_{k_\ell} - \alpha_{k_\ell}\nabla f(\mathbf{x}_{k_\ell})\big)}_{=-\alpha_{k_\ell}\nabla f(\mathbf{x}_{k_\ell})} \Big\} \geq 0.$$

Zusammen führt dies auf

$$\begin{aligned}(\mathbf{x}_{k_\ell} - \mathbf{y}_{k_\ell+1})^\intercal \nabla f(\mathbf{x}_{k_\ell}) &\geq (\mathbf{x}_{k_\ell} - \mathbf{x}_{k_\ell+1})^\intercal \nabla f(\mathbf{x}_{k_\ell}) \\ &\geq \frac{\|\mathbf{x}_{k_\ell+1} - \mathbf{x}_{k_\ell}\|_2^2}{\alpha_{k_\ell}} \\ &\geq \frac{\varepsilon}{2}\|\mathbf{y}_{k_\ell+1} - \mathbf{x}_{k_\ell}\|_2.\end{aligned}$$

Speziell ergibt sich wegen (4.26) auch $\|\mathbf{y}_{k_\ell+1} - \mathbf{x}_{k_\ell}\|_2 \to 0$ für $\ell \to \infty$. Die gleichmäßige Stetigkeit von ∇f impliziert daher

$$\begin{aligned}\left|1 - \frac{f(\mathbf{x}_{k_\ell}) - f(\mathbf{y}_{k_\ell+1})}{(\mathbf{x}_{k_\ell} - \mathbf{y}_{k_\ell+1})^\intercal \nabla f(\mathbf{x}_{k_\ell})}\right| &= \frac{\mathcal{O}(\|\mathbf{y}_{k_\ell+1} - \mathbf{x}_{k_\ell}\|_2)}{(\mathbf{x}_{k_\ell} - \mathbf{y}_{k_\ell+1})^\intercal \nabla f(\mathbf{x}_{k_\ell})} \\ &\leq \frac{2}{\varepsilon}\frac{\mathcal{O}(\|\mathbf{y}_{k_\ell+1} - \mathbf{x}_{k_\ell}\|_2)}{\|\mathbf{x}_{k_\ell} - \mathbf{y}_{k_\ell+1}\|_2} \\ &\xrightarrow{\ell\to\infty} 0.\end{aligned}$$

Dies steht jedoch im Widerspruch zu der aus (4.27) folgenden Abschätzung

$$\frac{f(\mathbf{x}_{k_\ell}) - f(\mathbf{y}_{k_\ell+1})}{(\mathbf{x}_{k_\ell} - \mathbf{y}_{k_\ell+1})^\intercal \nabla f(\mathbf{x}_{k_\ell})} < \sigma < 1. \qquad \spadesuit$$

Bemerkung Da $\alpha_k \leq 1$ ist, folgt aus Satz 4.24 $\|\mathbf{x}_{k+1} - \mathbf{x}_k\|_2 \to 0$ für $k \to \infty$. ♦

Definition 4.25 Eine Menge $C \subset \mathbb{R}^n$ heißt **Kegel**, wenn aus $\mathbf{x} \in C$ auch $\lambda\mathbf{x} \in C$ folgt für alle $\lambda \geq 0$. Der **Tangentialkegel** $T_D(\mathbf{x})$ von der Menge $D \subset \mathbb{R}^n$ an einen Punkt $\mathbf{x} \in D$ ist der kleinste abgeschlossene Kegel, der die Menge

$$M := \{\mathbf{d} = \mathbf{y} - \mathbf{x} : \mathbf{y} \in D\}$$

enthält.

Bemerkung Es sei $\mathbf{x} \in K$ und $\{\mathbf{y}_k\}_{k\in\mathbb{N}} \subset K \setminus \{\mathbf{x}\}$ eine Folge mit $\lim_{k\to\infty} \mathbf{y}_k = \mathbf{x}$. Dann ist

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

offenbar im Tangentialkegel $T_K(\mathbf{x})$ enthalten. Die Richtung $\mathbf{d}$ wird *Grenzrichtung* der Folge genannt. Umgekehrt gibt es zu jedem $\mathbf{d} \in T_K(\mathbf{x})$ mit $\|\mathbf{d}\|_2 = 1$ eine Folge

$\{\mathbf{y}_k\}_{k\in\mathbb{N}} \subset K$ derart, dass

$$\mathbf{d} = \lim_{k\to\infty} \frac{\mathbf{y}_k - \mathbf{x}}{\|\mathbf{y}_k - \mathbf{x}\|_2} \quad \text{und} \quad \lim_{k\to\infty} \mathbf{y}_k = \mathbf{x}. \tag{4.28}$$

Der Tangentialkegel enthält also gerade die Grenzrichtungen von allen Folgen

$$\{\mathbf{y}_k\}_{k\in\mathbb{N}} \subset K \setminus \{\mathbf{x}\} \quad \text{mit} \quad \lim_{k\to\infty} \mathbf{y}_k = \mathbf{x}.$$

Insbesondere ist der Tangentialkegel konvex, weil K konvex ist. ♦

Lemma 4.26 *Für jeden Punkt $\mathbf{x} \in K$ erfüllt die orthogonale Projektion $\mathbf{P}_{T_K(\mathbf{x})}\big(-\nabla f(\mathbf{x})\big)$ der Richtung des steilsten Abstiegs auf den Tangentialkegel $T_K(\mathbf{x})$ die folgenden Eigenschaften:*

(i.) Es gilt

$$\nabla f(\mathbf{x})^\mathsf{T} \mathbf{P}_{T_K(\mathbf{x})}\big(-\nabla f(\mathbf{x})\big) = -\big\|\mathbf{P}_{T_K(\mathbf{x})}\big(-\nabla f(\mathbf{x})\big)\big\|_2^2.$$

(ii.) Es ist

$$\min\{\nabla f(\mathbf{x})^\mathsf{T}\mathbf{d} \,:\, \mathbf{d} \in T_K(\mathbf{x}),\, \|\mathbf{d}\|_2 \le 1\} = -\big\|\mathbf{P}_{T_K(\mathbf{x})}\big(-\nabla f(\mathbf{x})\big)\big\|_2.$$

(iii.) Der Punkt $\mathbf{x}$ ist genau dann ein stationärer Punkt des Minimierungsproblems mit Nebenbedingungen (4.19), wenn $\mathbf{P}_{T_K(\mathbf{x})}\big(-\nabla f(\mathbf{x})\big) = \mathbf{0}$.

Beweis. (*i.*) Nach Definition der Orthogonalprojektion besitzt die Funktion

$$g(\lambda) := \frac{1}{2}\big\|\lambda \mathbf{P}_{T_K(\mathbf{x})}\big(-\nabla f(\mathbf{x})\big) + \nabla f(\mathbf{x})\big\|_2^2$$

ein Minimum bei $\lambda = 1$. Daher gilt

$$g'(1) := \big\|\mathbf{P}_{T_K(\mathbf{x})}\big(-\nabla f(\mathbf{x})\big)\big\|_2^2 + \nabla f(\mathbf{x})^\mathsf{T} \mathbf{P}_{T_K(\mathbf{x})}\big(-\nabla f(\mathbf{x})\big) = 0.$$

(*ii.*) Wegen Aussage (*i.*) gilt

$$\big\|\mathbf{P}_{T_K(\mathbf{x})}\big(-\nabla f(\mathbf{x})\big) + \nabla f(\mathbf{x})\big\|_2^2 = \|\nabla f(\mathbf{x})\|_2^2 - \big\|\mathbf{P}_{T_K(\mathbf{x})}\big(-\nabla f(\mathbf{x})\big)\big\|_2^2.$$

Für alle $\mathbf{d} \in T_K(\mathbf{x})$ mit $\|\mathbf{d}\|_2 \le \big\|\mathbf{P}_{T_K(\mathbf{x})}\big(-\nabla f(\mathbf{x})\big)\big\|_2$ gilt nach Definition der orthogonalen Projektion

$$\begin{aligned}\big\|\mathbf{P}_{T_K(\mathbf{x})}\big(-\nabla f(\mathbf{x})\big) + \nabla f(\mathbf{x})\big\|_2^2 &\le \|\mathbf{d} + \nabla f(\mathbf{x})\|_2^2 \\ &\le \big\|\mathbf{P}_{T_K(\mathbf{x})}\big(-\nabla f(\mathbf{x})\big)\big\|_2^2 + 2\nabla f(\mathbf{x})^\mathsf{T}\mathbf{d} + \|\nabla f(\mathbf{x})\|_2^2.\end{aligned}$$

Zusammen ergibt dies

$$\nabla f(\mathbf{x})^\mathsf{T}\mathbf{d} \geq -\big\|\mathbf{P}_{T_K(\mathbf{x})}\big(-\nabla f(\mathbf{x})\big)\big\|_2^2.$$

Das Behauptete erhält man, indem man $\widehat{\mathbf{d}} = \mathbf{d}/\big\|\mathbf{P}_{T_K(\mathbf{x})}\big(-\nabla f(\mathbf{x})\big)\big\|_2$ setzt.

(*iii.*) Definitionsgemäß ist $\mathbf{x} \in K$ genau dann ein stationärer Punkt, wenn

$$\nabla f(\mathbf{x})^\mathsf{T}(\mathbf{y} - \mathbf{x}) \geq 0$$

für alle $\mathbf{y} \in K$ ist. Dies ist gleichbedeutend damit, dass $\nabla f(\mathbf{x})^\mathsf{T}\mathbf{d} \geq 0$ für alle $\mathbf{d} \in T_K(\mathbf{x})$ ist. Aussage (*ii.*) impliziert, dass dies genau dann der Fall ist, wenn $\mathbf{P}_{T_K(\mathbf{x})}\big(-\nabla f(\mathbf{x})\big) = \mathbf{0}$ erfüllt ist. ♠

Bemerkung Ist $\mathbf{P}_{T_K(\mathbf{x})}\big(-\nabla f(\mathbf{x})\big) \neq \mathbf{0}$, so kann Aussage (*ii.*) des obigen Lemmas auch als

$$\min\{\nabla f(\mathbf{x})^\mathsf{T}\mathbf{d} \,:\, \mathbf{d} \in T_K(\mathbf{x}),\ \|\mathbf{d}\|_2 = 1\} = -\big\|\mathbf{P}_{T_K(\mathbf{x})}\big(-\nabla f(\mathbf{x})\big)\big\|_2 \tag{4.29}$$

geschrieben werden, denn das Minimum wird für $\|\mathbf{d}\|_2 = 1$ angenommen. ♦

Satz 4.27 *Die Funktion $f : \mathbb{R}^n \to \mathbb{R}$ sei auf K stetig differenzierbar und nach unten beschränkt. Weiter sei ∇f auf K gleichmäßig stetig. Dann gilt für die Iterierten $\{\mathbf{x}_k\}_{k\in\mathbb{N}}$ des projizierten Gradientenverfahrens*

$$\lim_{k\to\infty} \mathbf{P}_{T_K(\mathbf{x}_k)}\big(-\nabla f(\mathbf{x}_k)\big) = \mathbf{0}.$$

Beweis. Zu beliebigen $\varepsilon > 0$ gibt es nach Lemma 4.26 (*ii.*) zu jeder Iterierten $\mathbf{x}_k$ ein $\mathbf{d}_k \in T_K(\mathbf{x}_k)$ mit $\|\mathbf{d}_k\|_2 = 1$, so dass

$$\nabla f(\mathbf{x}_k)^\mathsf{T}\mathbf{d}_k \leq -\big\|\mathbf{P}_{T_K(\mathbf{x}_k)}\big(-\nabla f(\mathbf{x}_k)\big)\big\|_2 + \varepsilon \tag{4.30}$$

gilt. Da $\mathbf{d}_k$ Grenzrichtung einer zulässigen Folge ist, gibt es ein $\mathbf{y}_k \in K$ mit

$$\left\| \frac{\mathbf{y}_k - \mathbf{x}_k}{\|\mathbf{y}_k - \mathbf{x}_k\|_2} - \mathbf{d}_k \right\|_2 \leq \varepsilon.$$

Aus Lemma 4.22 (*i.*) folgt

$$\begin{aligned}\big\{\mathbf{x}_{k+1} - \big(\mathbf{x}_k - \alpha_k\nabla f(\mathbf{x}_k)\big)\big\}^\mathsf{T}&(\mathbf{x}_{k+1} - \mathbf{y}_{k+1}) \\ &= \big\{\mathbf{P}_K\big(\mathbf{x}_k - \alpha_k\nabla f(\mathbf{x}_k)\big) - \big(\mathbf{x}_k - \alpha_k\nabla f(\mathbf{x}_k)\big)\big\}^\mathsf{T} \\ &\qquad \big\{\mathbf{P}_K\big(\mathbf{x}_k - \alpha_k\nabla f(\mathbf{x}_k)\big) - \mathbf{y}_{k+1}\big\} \leq 0,\end{aligned}$$

was auf

$$\alpha_k \nabla f(\mathbf{x}_k)^\mathsf{T}(\mathbf{x}_{k+1} - \mathbf{y}_{k+1}) \leq \|\mathbf{x}_{k+1} - \mathbf{x}_k\|_2 \|\mathbf{x}_{k+1} - \mathbf{y}_{k+1}\|_2$$

beziehungsweise

$$-\frac{\nabla f(\mathbf{x}_k)^\mathsf{T}(\mathbf{y}_{k+1} - \mathbf{x}_{k+1})}{\|\mathbf{x}_{k+1} - \mathbf{y}_{k+1}\|_2} \leq \frac{\|\mathbf{x}_{k+1} - \mathbf{x}_k\|_2}{\alpha_k}$$

führt. Insgesamt erhalten wir deshalb

$$\begin{aligned} -\nabla f(\mathbf{x}_k)^\mathsf{T}\mathbf{d}_{k+1} &\leq \|\nabla f(\mathbf{x}_k)\|_2 \left\| \frac{\mathbf{y}_{k+1} - \mathbf{x}_{k+1}}{\|\mathbf{y}_{k+1} - \mathbf{x}_{k+1}\|_2} - \mathbf{d}_{k+1} \right\|_2 - \frac{\nabla f(\mathbf{x}_k)^\mathsf{T}(\mathbf{y}_{k+1} - \mathbf{x}_{k+1})}{\|\mathbf{y}_{k+1} - \mathbf{x}_{k+1}\|_2} \\ &\leq \varepsilon \|\nabla f(\mathbf{x}_k)\|_2 + \frac{\|\mathbf{x}_{k+1} - \mathbf{x}_k\|_2}{\alpha_k}. \end{aligned}$$

Die Kombination mit (4.30) ergibt

$$\begin{aligned} &\left\| \mathbf{P}_{T_k(\mathbf{x}_{k+1})}\big(-\nabla f(\mathbf{x}_{k+1})\big) \right\|_2 \\ &\quad \leq -\nabla f(\mathbf{x}_{k+1})^\mathsf{T}\mathbf{d}_{k+1} + \varepsilon \\ &\quad \leq -\nabla f(\mathbf{x}_k)^\mathsf{T}\mathbf{d}_{k+1} + \|\nabla f(\mathbf{x}_{k+1}) - \nabla f(\mathbf{x}_k)\|_2 \underbrace{\|\mathbf{d}_{k+1}\|_2}_{=1} + \varepsilon \\ &\quad \leq \varepsilon \|\nabla f(\mathbf{x}_k)\|_2 + \frac{\|\mathbf{x}_{k+1} - \mathbf{x}_k\|_2}{\alpha_k} + \|\nabla f(\mathbf{x}_{k+1}) - \nabla f(\mathbf{x}_k)\|_2 + \varepsilon. \end{aligned}$$

Weil $\varepsilon > 0$ beliebig war, folgt hieraus schließlich

$$\begin{aligned} &\lim_{k\to\infty} \left\| \mathbf{P}_{T_k(\mathbf{x}_{k+1})}\big(-\nabla f(\mathbf{x}_{k+1})\big) \right\|_2 \\ &\quad \leq \lim_{k\to\infty} \frac{\|\mathbf{x}_{k+1} - \mathbf{x}_k\|_2}{\alpha_k} + \lim_{k\to\infty} \|\nabla f(\mathbf{x}_{k+1}) - \nabla f(\mathbf{x}_k)\|_2 = 0 \end{aligned}$$

wobei Satz 4.24 und die gleichmäßige Stetigkeit von ∇f zur Anwendung kommt. ♠

In der Regel folgt aus der Stetigkeit von ∇f nicht, dass auch $\mathbf{P}_{T_k(\mathbf{x})}\big(-\nabla f(\mathbf{x})\big)$ stetig ist. Um sicherzustellen, dass die Iterierten des projizierten Gradientenverfahrens tatsächlich gegen einen stationären Punkt konvergieren, benötigen wir daher das folgende Resultat.

Satz 4.28 *Die Funktion $f : \mathbb{R}^n \to \mathbb{R}$ sei auf K stetig differenzierbar. Dann folgt für jede Folge $\{\mathbf{x}_k\}_{k\in\mathbb{N}} \subset K$ mit $\mathbf{x}_k \to \mathbf{x}^\star \in K$*

$$\left\|\mathbf{P}_{T_K(\mathbf{x}^\star)}\big(-\nabla f(\mathbf{x}^\star)\big)\right\|_2 \le \liminf_{k\to\infty} \left\|\mathbf{P}_{T_K(\mathbf{x}_k)}\big(-\nabla f(\mathbf{x}_k)\big)\right\|_2.$$

Beweis. Aus Lemma 4.26 (*ii.*) folgt für jedes $\mathbf{y} \in K$

$$-\nabla f(\mathbf{x}_k)^\mathsf{T}(\mathbf{y}-\mathbf{x}_k) \le \left\|\mathbf{P}_{T_K(\mathbf{x}_k)}\big(-\nabla f(\mathbf{x}_k)\big)\right\|_2 \|\mathbf{y}-\mathbf{x}_k\|_2,$$

woraus sich für $k \to \infty$ die Ungleichung

$$-\nabla f(\mathbf{x}^\star)^\mathsf{T}(\mathbf{y}-\mathbf{x}^\star) \le \liminf_{k\to\infty} \left\|\mathbf{P}_{T_K(\mathbf{x}_k)}\big(-\nabla f(\mathbf{x}_k)\big)\right\|_2 \|\mathbf{y}-\mathbf{x}^\star\|_2$$

ergibt. Zu jedem $\mathbf{d} \in T_K(\mathbf{x}^\star)$ mit $\|\mathbf{d}\|_2 = 1$ lässt sich eine Folge $\{\mathbf{y}_k\}_{k\in\mathbb{N}}$ aus K derart finden, dass

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

Somit erhalten wir

$$-\nabla f(\mathbf{x}^\star)^\mathsf{T}\mathbf{d} \le \liminf_{k\to\infty} \left\|\mathbf{P}_{T_K(\mathbf{x}_k)}\big(-\nabla f(\mathbf{x}_k)\big)\right\|_2$$

und daraus wegen (4.29) die Behauptung:

$$\begin{aligned}\left\|\mathbf{P}_{T_K(\mathbf{x}^\star)}\big(-\nabla f(\mathbf{x}^\star)\big)\right\|_2 &= \max\{-\nabla f(\mathbf{x}^\star)^\mathsf{T}\mathbf{d} : \mathbf{d} \in T_K(\mathbf{x}^\star), \|\mathbf{d}\|_2 = 1\} \\ &\le \liminf_{k\to\infty} \left\|\mathbf{P}_{T_K(\mathbf{x}_k)}\big(-\nabla f(\mathbf{x}_k)\big)\right\|_2.\end{aligned}$$

♠

Bemerkung Die Kombination der Sätze 4.24, 4.27 und 4.28 liefert die folgende Aussage: Ist die zulässige Menge $K \subset \mathbb{R}^n$ konvex und abgeschlossen und ist die Funktion $f : \mathbb{R}^n \to \mathbb{R}$ auf K stetig differenzierbar mit gleichmäßig stetigem Gradienten und nach unten beschränkt, dann gilt für jeden Häufungspunkt $\mathbf{x}^\star \in K$ der Iterierten $\{\mathbf{x}_k\}_{k\in\mathbb{N}}$ des projizierten Gradientenverfahrens 4.21

$$\mathbf{P}_{T_K(\mathbf{x}^\star)}\big(-\nabla f(\mathbf{x}^\star)\big) = \mathbf{0}.$$

Gemäß Lemma 4.26 (*iii.*) bedeutet dies, dass $\mathbf{x}^\star$ ein stationärer Punkt ist. ♦

INDEX