Numerik-Vorlesungen

Teil 3

Eigenwerte, Ausgleichsprobleme, Minimierung

Peter Szyler, Horst Hollatz

Letzte Änderung: 20. Januar 2006

Inhaltsverzeichnis

9. Eigenwerte symmetrischer Matrizen **1**
- 9.1. Grundlegende Definitionen und Sätze 1
- 9.2. Störungstheorie 10
- 9.3. Das Jacobi-Verfahren 25
- 9.4. Die Vektor- und Teilraumiteration 34
- 9.5. Die Inverse Iteration nach Wielandt 47
- 9.6. Transformationen auf Tridiagonalform 51
- 9.7. Der Lanczos-Algorithmus 55
- 9.8. Der QR-Algorithmus 60
- 9.9. Der QR-Algorithmus für Tridiagonalmatrizen 74

10. Lineare Ausgleichsprobleme **87**
- 10.1. Problemstellung und klassische Lösung 87
- 10.2. Störungstheorie 94
- 10.3. Normalgleichungsverfahren 107
- 10.4. Orthogonalisierungsverfahren 110
- 10.5. Aufgaben 115

11. Freie Minimierung **119**
- 11.1. Einführung 119
  - 11.1.1. Aufgabenstellung und grundlegende Begriffe 119
  - 11.1.2. Differenzierbarkeit und Richtungsableitung 120
  - 11.1.3. Optimalitätskriterien 126
- 11.2. Ein Modellalgorithmus 135
  - 11.2.1. Schrittweiten bei glatter Zielfunktion 136
  - 11.2.2. Konvergenz des Modellalgorithmus 142
- 11.3. Quasi-Newton-Verfahren 150
  - 11.3.1. Gedämpftes und ungedämpftes Newton-Verfahren 150
  - 11.3.2. Verfahren der Oren-Luenberger-Klasse 154
  - 11.3.3. Algorithmen zur Aufdatierung von Zerlegungen 165
- 11.4. Trust-Region-Verfahren 167

Index **171**

Kapitel 9

Eigenwerte symmetrischer Matrizen

9.1. Grundlegende Definitionen und Sätze

Bekanntlich lassen sich für reelle und komplexe quadratische Matrizen Eigenwerte und Eigenvektoren definieren. Dabei sind die Eigenwerte und Eigenvektoren auch für eine reelle Matrix durchaus komplex.
Es sei $\boldsymbol{A} \in \mathbb{C}^{n\times n}$. Eine Zahl $\lambda \in \mathbb{C}$ heißt **Eigenwert** der Matrix $\boldsymbol{A}$, falls ein Vektor $\boldsymbol{x} \in \mathbb{C}^n$ mit $\boldsymbol{x} \neq \boldsymbol{o}$ existiert, so dass $\boldsymbol{Ax} = \lambda\boldsymbol{x}$ gilt. Jeder derartige Vektor heißt **Eigenvektor** zum Eigenwert λ.
Aus $\boldsymbol{Ax} = \lambda\boldsymbol{x}$ folgt $(\boldsymbol{A} - \lambda\boldsymbol{I})\boldsymbol{x} = \boldsymbol{o}$. Damit dieses Gleichungssystem nichttriviale Lösungen besitzt, muss der Rang der Matrix $\boldsymbol{A} - \lambda\boldsymbol{I}$ kleiner als n sein, d. h. es muss $\det(\boldsymbol{A} - \lambda\boldsymbol{I}) = 0$ gelten. Damit ergibt sich gleich die nächste Definition. Das Polynom $\varphi(\mu) = \det(\boldsymbol{A} - \mu\boldsymbol{I})$ heißt **charakteristisches Polynom**der Matrix $\boldsymbol{A}$. Offensichtlich gilt $\varphi \in \Pi_n$. Die Nullstellen des charakteristischen Polynoms sind gerade die Eigenwerte von $\boldsymbol{A}$. Das charakteristische Polynom der Matrix $\boldsymbol{A}$ habe die Darstellung:

$$\varphi(\mu) = (-1)^n(\mu - \lambda_1)^{\sigma_1}(\mu - \lambda_2)^{\sigma_2}\cdots(\mu - \lambda_k)^{\sigma_k}$$

mit paarweise verschiedenen $\lambda_1, \lambda_2, \ldots, \lambda_k$. Dann heißen die Zahlen

$$\sigma_i = \sigma(\lambda_i),\ i = 1, \ldots k,$$

algebraische Vielfachheiten der Eigenwerte λ_i, $i = 1, \ldots k$. Wegen $\text{Grad}(\varphi) = n$ folgt $\sigma_1 + \sigma_2 + \cdots + \sigma_k = n$. Die Eigenvektoren zu einem Eigenwert sind offensichtlich nicht eindeutig bestimmt. Sind $\boldsymbol{x}$ und $\boldsymbol{y}$ Eigenvektoren zum Eigenwert λ, so sind auch alle Linearkombinationen $\alpha\boldsymbol{x} + \beta\boldsymbol{y} \neq \boldsymbol{o}$ Eigenvektor zum Eigenwert λ. Die Menge aller Eigenvektoren zu einem festen Eigenwert bildet gemeinsam mit

dem Nullvektor einen Unterraum des $\mathbb{C}^n$.
Es sei $\boldsymbol{A} \in \mathbb{C}^{n\times n}$ und λ_i ein Eigenwert von $\boldsymbol{A}$.

$$\boldsymbol{\mathcal{L}}_i = \boldsymbol{\mathcal{L}}(\lambda_i) = \{\, \boldsymbol{x} \in \mathbb{C}^n \mid (\boldsymbol{A} - \lambda_i \boldsymbol{I})\boldsymbol{x} = \boldsymbol{o} \,\}$$

heißt **Eigenraum** des Eigenwertes λ_i. Die Zahlen

$$\varrho_i = \varrho(\lambda_i) = \dim(\boldsymbol{\mathcal{L}}_i) = n - rg(\boldsymbol{A} - \lambda_i \boldsymbol{I}), \quad i = 1, \ldots, k$$

heißen **geometrische Vielfachheiten** der Eigenwerte λ_i, $i = 1, \ldots, k$. Geometrische und algebraische Vielfachheit eines Eigenwertes brauchen nicht übereinzustimmen. Das zeigen schon die folgenden einfachen Beispiele.

9.1. Beispiel: Es sei $\boldsymbol{A} = \lambda \boldsymbol{I}$. Das charakteristische Polynom von $\boldsymbol{A}$ lautet

$$\varphi(\mu) = \det(\boldsymbol{A} - \mu \boldsymbol{I}) = (\mu - \lambda)^n.$$

λ ist der einzige Eigenwert von $\boldsymbol{A}$. Jeder Vektor $\boldsymbol{x} \in \mathbb{R}^n$ mit $\boldsymbol{x} \neq \boldsymbol{o}$ ist Eigenvektor zum Eigenwert λ. Geometrische und algebraische Vielfachheit von λ sind also jeweils gleich n. ♡

9.2. Beispiel: Es sei

$$\boldsymbol{B} = \begin{pmatrix} \lambda & 1 & & & & 0 \\ & \lambda & 1 & & & \\ & & \cdot & \cdot & & \\ & & & \cdot & \cdot & \\ & & & & \lambda & 1 \\ 0 & & & & & \lambda \end{pmatrix} \in \mathbb{R}^{n\times n}.$$

Wir erhalten als charakteristisches Polynom $\varphi(\mu) = \det(\boldsymbol{B} - \mu \boldsymbol{I}) = (\mu - \lambda)^n$. Die Zahl λ ist einziger Eigenwert von $\boldsymbol{A}$ mit der algebraischen Vielfachheit n. Für die geometrische Vielfachheit von λ ergibt sich aber

$$\varrho = n - rg(\boldsymbol{B} - \lambda \boldsymbol{I}) = n - (n-1) = 1 \quad .$$

λ hat die geometrische Vielfachheit $1 < n$. ♡

Die algebraische und geometrische Vielfachheit eines Eigenwertes unterscheiden sich daher. Wir werden später zeigen, dass die geometrische Vielfachheit eines Eigenwertes nicht größer ist als seine algebraische Vielfachheit. Eine Matrix, die mindestens einen Eigenwert λ mit $\varrho(\lambda) < \sigma(\lambda)$ besitzt, heißt defektiv. Wollen wir Eigenwerte und Eigenvektoren einer Matrix berechnen, so ist einerseits zu untersuchen,

wie diese Größen auf Störungen in der Matrix reagieren, und andererseits, wie die Eigenwerte und Eigenvektoren verschiedener Matrizen miteinander zusammenhängen. Der erste Gesichtspunkt wird Gegenstand des nächsten Abschnittes sein. Zum zweiten Aspekt notieren wir folgenden Satz.

9.3. Satz: *Ist $p(\mu) = \gamma_0 + \gamma_1\mu + \cdots + \gamma_m\mu^m$ ein beliebiges Polynom, und definiert man für eine Matrix $\boldsymbol{A} \in \mathbb{C}^{n\times n}$ die Matrix $p(\boldsymbol{A})$ durch*

$$p(\boldsymbol{A}) = \gamma_0\boldsymbol{I} + \gamma_1\boldsymbol{A} + \cdots + \gamma_m\boldsymbol{A}^m,$$

so besitzt die Matrix $p(\boldsymbol{A})$ den Eigenvektor $\boldsymbol{x}$ zum Eigenwert $p(\lambda)$, falls $\boldsymbol{x}$ Eigenvektor von $\boldsymbol{A}$ zum Eigenwert λ ist. Insbesondere besitzt $\alpha\boldsymbol{A}$ den Eigenwert $\alpha\lambda$ und $\boldsymbol{A} + \tau\boldsymbol{I}$ den Eigenwert $\lambda + \tau$.

Beweis: Aus $\boldsymbol{A}\boldsymbol{x} = \lambda\boldsymbol{x}$ folgt $\boldsymbol{A}^2\boldsymbol{x} = \boldsymbol{A}(\boldsymbol{A}\boldsymbol{x}) = \lambda\boldsymbol{A}\boldsymbol{x} = \lambda^2\boldsymbol{x}$ und allgemein

$$\boldsymbol{A}^i\boldsymbol{x} = \lambda^i\boldsymbol{x}.$$

Damit erhält man sofort $p(\boldsymbol{A})\boldsymbol{x} = p(\lambda)\boldsymbol{x}$. ❀

Weiterhin gilt

9.4. Satz: *Ist λ Eigenwert der Matrix $\boldsymbol{A}$, so ist λ auch Eigenwert der Matrix $\boldsymbol{A}^T$ und $\bar{\lambda}$ ist Eigenwert von $\boldsymbol{A}^H$.*

Beweis: Es gilt:

$$\det(\boldsymbol{A}^T - \lambda\boldsymbol{I}) = \det((\boldsymbol{A} - \lambda\boldsymbol{I})^T) = \det(\boldsymbol{A} - \lambda\boldsymbol{I})$$

$$\det(\boldsymbol{A}^H - \bar{\lambda}\boldsymbol{I}) = \det((\boldsymbol{A} - \lambda\boldsymbol{I})^H) = \overline{\det(\boldsymbol{A} - \lambda\boldsymbol{I})}.$$

❀

Wichtig sind Transformationen, die das Eigenwertspektrum einer Matrix unverändert lassen. Es sei $\boldsymbol{A} \in \mathbb{C}^{n\times n}$ eine beliebige Matrix und $\boldsymbol{T} \in \mathbb{C}^{n\times n}$ sei regulär. Dann bezeichnet man die Abbildung

$$\boldsymbol{A} \to \boldsymbol{B} = \boldsymbol{T}^{-1}\boldsymbol{A}\boldsymbol{T}$$

als **Ähnlichkeitstransformation**. Die Matrizen $\boldsymbol{A}$ und $\boldsymbol{B}$ heißen **ähnlich**.
Ist eine Matrix zu einer Diagonalmatrix ähnlich, so heißt sie **diagonalähnlich**. Die Eigenschaften einer Ähnlichkeitstransformation werden in folgendem Satz beschrieben.

9.5. Satz: *Die Matrix $\boldsymbol{A} \in \mathbb{C}^{n\times n}$ werde mittels einer regulären Matrix $\boldsymbol{T} \in \mathbb{C}^{n\times n}$ in $\boldsymbol{B} = \boldsymbol{T}^{-1}\boldsymbol{A}\boldsymbol{T}$ transformiert. Dann haben $\boldsymbol{A}$ und $\boldsymbol{B}$ dieselben Eigenwerte mit denselben algebraischen und geometrischen Vielfachheiten.*

Beweis: Aus

$$\boldsymbol{B} - \lambda\boldsymbol{I} = \boldsymbol{T}^{-1}\boldsymbol{A}\boldsymbol{T} - \lambda\boldsymbol{T}^{-1}\boldsymbol{T} = \boldsymbol{T}^{-1}(\boldsymbol{A} - \lambda\boldsymbol{I})\boldsymbol{T}$$

folgt

$$\det(\boldsymbol{B} - \lambda\boldsymbol{I}) = \det(\boldsymbol{T}^{-1})\det(\boldsymbol{A} - \lambda\boldsymbol{I})\det(\boldsymbol{T}) = \det(\boldsymbol{A} - \lambda\boldsymbol{I}).$$

Die charakteristischen Polynome von $\boldsymbol{A}$ und $\boldsymbol{B}$ stimmen daher überein. Damit haben $\boldsymbol{A}$ und $\boldsymbol{B}$ dieselben Eigenwerte mit denselben algebraischen Vielfachheiten. Für die geometrische Vielfachheit des Eigenwertes λ_i von $\boldsymbol{B}$ gilt wegen der Regularität von $\boldsymbol{T}$:

$$\varrho_i^{(B)} = n - rg(\boldsymbol{B} - \lambda_i\boldsymbol{I}) = n - rg(\boldsymbol{T}^{-1}(\boldsymbol{A} - \lambda_i\boldsymbol{I})\boldsymbol{T}) = n - rg(\boldsymbol{A} - \lambda_i\boldsymbol{I}) = \varrho_i^{(A)}.$$

❀

Viele Verfahren zur Eigenwertberechnung beruhen darauf, dass man versucht, die Matrix durch Ähnlichkeitstransformationen so umzuformen, dass Eigenwerte und eventuell Eigenvektoren leicht ablesbar sind. Eine Möglichkeit zeigt der folgende Satz.

9.6. Komplexe SCHUR**-Zerlegung:** *Zu jeder (n,n)-Matrix $\boldsymbol{A} \in \mathbb{C}^{n\times n}$ existiert eine unitäre Matrix $\boldsymbol{U} \in \mathbb{C}^{n\times n}$, so dass $\boldsymbol{U}^H\boldsymbol{A}\boldsymbol{U}$ eine obere Dreiecksmatrix ist, in deren Diagonale die Eigenwerte von $\boldsymbol{A}$ stehen.*

Beweis: Wir beweisen den Satz mittels Induktion über die Dimension n.
Für $n = 1$ ist die Aussage trivial.
Wir nehmen an, dass die Aussage für alle $(n-1)\times(n-1)$-Matrizen gilt. Es sei nun $\boldsymbol{A} \in \mathbb{C}^{n\times n}$ und λ_1 ein Eigenwert mit einem zugehörigen Eigenvektor $\boldsymbol{v}_1 \neq \boldsymbol{o}$. O.B.d.A. sei $\|\boldsymbol{v}_1\|_2 = 1$. Nun existieren weitere $n-1$ Vektoren $\boldsymbol{v}_2, \boldsymbol{v}_3, \ldots, \boldsymbol{v}_n$ finden, so dass $\boldsymbol{v}_1, \boldsymbol{v}_2, \ldots, \boldsymbol{v}_n$ eine Orthonormalbasis des $\mathbb{C}^n$ bilden. Die aus diesen Vektoren gebildete Matrix $\boldsymbol{V} = (\boldsymbol{v}_1, \boldsymbol{v}_2, \ldots, \boldsymbol{v}_n)$ ist unitär, und es gilt

$$\boldsymbol{V}^H\boldsymbol{A}\boldsymbol{V} = \begin{pmatrix} \lambda_1 & \boldsymbol{y}^T \\ \boldsymbol{o} & \bar{\boldsymbol{A}} \end{pmatrix}, \quad \bar{\boldsymbol{A}} \in \mathbb{C}^{(n-1)\times(n-1)}, \quad \boldsymbol{y} \in \mathbb{C}^{n-1}.$$

Nach Induktionsvoraussetzung existiert eine unitäre Matrix $\boldsymbol{W} \in \mathbb{C}^{(n-1)\times(n-1)}$ derart, dass $\boldsymbol{W}^H\bar{\boldsymbol{A}}\boldsymbol{W}$ eine obere Dreiecksmatrix ist. Die unitäre Matrix

$$\boldsymbol{U} = \boldsymbol{V}\begin{pmatrix} 1 & \boldsymbol{o}^T \\ \boldsymbol{o} & \boldsymbol{W} \end{pmatrix} \in \mathbb{C}^{n\times n}$$

transformiert dann die Matrix $\boldsymbol{A}$ orthogonal ähnlich auf obere Dreiecksform, denn

$$\boldsymbol{U}^H \boldsymbol{A} \boldsymbol{U} = \begin{pmatrix} 1 & \boldsymbol{o}^T \\ \boldsymbol{o} & \boldsymbol{W}^H \end{pmatrix} \begin{pmatrix} \lambda_1 & \boldsymbol{y}^T \overline{\boldsymbol{A}} \\ \boldsymbol{o} & \overline{\boldsymbol{A}} \end{pmatrix} \begin{pmatrix} 1 & \boldsymbol{o}^T \\ \boldsymbol{o} & \boldsymbol{W} \end{pmatrix} = \begin{pmatrix} \lambda_1 & \boldsymbol{y}^T \boldsymbol{W} \\ \boldsymbol{o} & \boldsymbol{W}^H \overline{\boldsymbol{A}} \boldsymbol{W} \end{pmatrix}.$$

Damit ist der Satz bewiesen. ❀

Ist die Matrix $\boldsymbol{A}$ reell und besitzt nur reelle Eigenwerte, so ist die Transformationsmatrix $\boldsymbol{U}$ orthogonal wählbar. Man erhält in diesem Falle eine reelle obere Dreiecksmatrix als Ergebnis. Besitzt eine reelle Matrix komplexe Eigenwerte, so ist sie nur mittels unitärer Ähnlichkeitstransformationen auf obere Dreiecksform transformierbar. Strebt man nur eine obere Blockdreiecksmatrix als Ergebnis an, so ist die ganze Transformation im Reellen durchführbar.

9.7. Reelle SCHUR**-Zerlegung:** *Zu jeder reellen (n,n)-Matrix $\boldsymbol{A}$ existiert eine orthogonale Matrix $\boldsymbol{Q}$ mit*

$$\boldsymbol{Q}^T \boldsymbol{A} \boldsymbol{Q} = \begin{pmatrix} \boldsymbol{R}_{11} & \boldsymbol{R}_{12} & \cdots & \boldsymbol{R}_{1m} \\ \boldsymbol{O} & \boldsymbol{R}_{22} & \cdots & \boldsymbol{R}_{2m} \\ \vdots & \ddots & \ddots & \vdots \\ \boldsymbol{O} & \cdots & \boldsymbol{O} & \boldsymbol{R}_{mm} \end{pmatrix},$$

wobei die Diagonalblöcke $\boldsymbol{R}_{ii}$ entweder 1×1-Matrizen oder 2×2-Matrizen mit einem Paar konjugiert komplexer Eigenwerte sind.

Beweis: Wir führen den Beweis mittels Induktion über die Anzahl k der Paare konjugiert komplexer Eigenwerte von $\boldsymbol{A}$. Für $k = 0$ besitzt die reelle Matrix $\boldsymbol{A}$ nur reelle Eigenwerte. Die Aussage folgt dann aus dem vorigen Satz. Wir nehmen an, dass eine derartige Zerlegung für alle Matrizen mit höchstens $k-1$ konjugiert komplexen Eigenwertpaaren gilt. Es sei nun $\boldsymbol{A}$ eine Matrix mit k konjugiert komplexen Eigenwertpaaren. Ein komplexer Eigenwert sei $\lambda = \alpha + i\beta$ mit $\beta \neq 0$. Der Vektor $\boldsymbol{y} + i\boldsymbol{z}$ mit $\boldsymbol{y}, \boldsymbol{z} \in \mathbb{R}^n$ sei ein zugehöriger Eigenvektor. Dann gilt $\boldsymbol{A}(\boldsymbol{y} + i\boldsymbol{z}) = (\alpha + i\beta)(\boldsymbol{y} + i\boldsymbol{z})$ und damit $\boldsymbol{A}\boldsymbol{y} = \alpha\boldsymbol{y} - \beta\boldsymbol{z}$ und $\boldsymbol{A}\boldsymbol{z} = \beta\boldsymbol{y} + \alpha\boldsymbol{z}$. In Matrixschreibweise lautet das

$$\boldsymbol{A}(\boldsymbol{y}, \quad \boldsymbol{z}) = (\boldsymbol{y}, \quad \boldsymbol{z}) \begin{pmatrix} \alpha & \beta \\ -\beta & \alpha \end{pmatrix}.$$

Die Vektoren $\boldsymbol{y}$ und $\boldsymbol{z}$ sind linear unabhängig, denn andernfalls wäre $\boldsymbol{z}$ ein reeller Eigenvektor zum komplexen Eigenwert λ, und das wäre ein Widerspruch zu $\boldsymbol{A} \in \mathbb{R}^{n \times n}$. Der Rang der Matrix $(\boldsymbol{y}, \quad \boldsymbol{z})$ ist somit 2, und es existiert eine orthogonale Matrix $\boldsymbol{V}$, so dass

$$(\boldsymbol{y}, \quad \boldsymbol{z}) = \boldsymbol{V} \begin{pmatrix} \boldsymbol{R}_1 \\ \boldsymbol{o} \end{pmatrix}$$

mit einer oberen Dreiecksmatrix $\boldsymbol{R}_1 \in \mathbb{R}^{2\times 2}$. Es folgt weiter

$$\begin{aligned}\boldsymbol{V}^T\boldsymbol{A}\boldsymbol{V}\begin{pmatrix}\boldsymbol{R}_1\\ \boldsymbol{o}\end{pmatrix} &= \boldsymbol{V}^T\boldsymbol{A}(\boldsymbol{y}, \quad \boldsymbol{z}) = \boldsymbol{V}^T(\boldsymbol{y}, \quad \boldsymbol{z})\begin{pmatrix}\alpha & \beta\\ -\beta & \alpha\end{pmatrix}\\ &= \begin{pmatrix}\boldsymbol{R}_1\\ \boldsymbol{o}\end{pmatrix}\begin{pmatrix}\alpha & \beta\\ -\beta & \alpha\end{pmatrix} = \begin{pmatrix}\bar{\boldsymbol{R}}_1\\ \boldsymbol{o}\end{pmatrix}.\end{aligned}$$

Partitioniert man $\boldsymbol{V}^T\boldsymbol{A}\boldsymbol{V}$ entsprechend

$$\boldsymbol{V}^T\boldsymbol{A}\boldsymbol{V} = \begin{pmatrix}\boldsymbol{B}_{11} & \boldsymbol{B}_{12}\\ \boldsymbol{B}_{21} & \boldsymbol{B}_{22}\end{pmatrix}, \quad \boldsymbol{B}_{11} \in \mathbb{R}^{2\times 2},$$

so folgt

$$\begin{pmatrix}\boldsymbol{B}_{11}\boldsymbol{R}_1\\ \boldsymbol{B}_{21}\boldsymbol{R}_1\end{pmatrix} = \begin{pmatrix}\bar{\boldsymbol{R}}_1\\ \boldsymbol{o}\end{pmatrix}.$$

Da $\boldsymbol{R}_1$ regulär ist, gilt

$$\boldsymbol{R}_1^{-1}\boldsymbol{B}_{11}\boldsymbol{R}_1 = \begin{pmatrix}\alpha & \beta\\ -\beta & \alpha\end{pmatrix}, \quad \boldsymbol{B}_{21} = \boldsymbol{o}.$$

Die Matrix $\boldsymbol{A}$ wird durch $\boldsymbol{V}$ orthogonal ähnlich auf

$$\boldsymbol{V}^T\boldsymbol{A}\boldsymbol{V} = \begin{pmatrix}\boldsymbol{B}_{11} & \boldsymbol{B}_{12}\\ \boldsymbol{o} & \boldsymbol{B}_{22}\end{pmatrix},$$

transformiert. Das Paar konjugiert komplexer Eigenwerte (λ und $\bar{\lambda}$) von $\boldsymbol{A}$ ist auch Eigenwertpaar von $\boldsymbol{B}_{11}$. Die restlichen Eigenwerte sind Eigenwerte von $\boldsymbol{B}_{22}$. Auf diese Matrix ist die Induktionsvoraussetzung anwendbar.
Damit folgt die Behauptung. ❀

Wir wollen nun noch den angekündigten Satz über den Zusammenhang zwischen den algebraischen und geometrischen Vielfachheiten von Eigenwerten einer Matrix beweisen.

9.8. Satz: *Es sei λ ein Eigenwert der Matrix $\boldsymbol{A} \in \mathbb{C}^{n\times n}$ mit der algebraischen Vielfachheit $\sigma(\lambda)$ und der geometrischen Vielfachheit $\varrho(\lambda)$. Dann gilt*

$$1 \leqq \varrho(\lambda) \leqq \sigma(\lambda) \leqq n.$$

Beweis: Die Gültigkeit der Ungleichungen $1 \leqq \varrho(\lambda)$ und $\sigma(\lambda) \leqq n$ ist offensichtlich. Es bleibt nur $\varrho(\lambda) \leqq \sigma(\lambda)$ zu zeigen. Es sei dazu $\varrho = \varrho(\lambda)$. Dann existieren ϱ linear unabhängige Eigenvektoren zum Eigenwert λ. Diese seien $\boldsymbol{v}_1, \ldots, \boldsymbol{v}_\varrho$. Wir ergänzen diese Vektoren durch Hinzunahme weiterer $n - \varrho$ Vektoren zu einer Basis des $\mathbb{C}^n$ und fassen alle Vektoren zu einer regulären Matrix $\boldsymbol{V} \in \mathbb{C}^{n\times n}$ zusammen. Wegen $\boldsymbol{V}\boldsymbol{e}_i = \boldsymbol{v}_i$ und $\boldsymbol{V}^{-1}\boldsymbol{v}_i = \boldsymbol{e}_i$ folgt für $i = 1, \ldots, \varrho$

$$\boldsymbol{V}^{-1}\boldsymbol{A}\boldsymbol{V}\boldsymbol{e}_i = \boldsymbol{V}^{-1}\boldsymbol{A}\boldsymbol{v}_i = \lambda\boldsymbol{V}^{-1}\boldsymbol{v}_i = \lambda\boldsymbol{e}_i.$$

Damit hat die Matrix $\boldsymbol{V}^{-1}\boldsymbol{A}\boldsymbol{V}$ die Gestalt

$$\boldsymbol{V}^{-1}\boldsymbol{A}\boldsymbol{V} = \begin{pmatrix} \lambda & & 0 & * & \cdots & * \\ & \ddots & & \vdots & & \vdots \\ 0 & & \lambda & * & \cdots & * \\ 0 & \cdots & 0 & * & \cdots & * \\ \vdots & & \vdots & \vdots & & \vdots \\ 0 & \cdots & 0 & * & \cdots & * \end{pmatrix} = \begin{pmatrix} \lambda\boldsymbol{I} & \boldsymbol{B} \\ \boldsymbol{o} & \boldsymbol{C} \end{pmatrix} \quad \lambda\boldsymbol{I} \in \mathbb{C}^{\varrho\times\varrho}.$$

Das charakteristische Polynom von $\boldsymbol{V}^{-1}\boldsymbol{A}\boldsymbol{V}$ und damit auch das von $\boldsymbol{A}$ lautet

$$\varphi(\mu) = \det(\boldsymbol{A} - \mu\boldsymbol{I}) = \det(\boldsymbol{V}^{-1}\boldsymbol{A}\boldsymbol{V} - \mu\boldsymbol{I}) = (\lambda - \mu)^\varrho \det(\boldsymbol{C} - \mu\boldsymbol{I}).$$

Die Größe λ ist daher mindestens ϱ-fache Nullstelle des charakteristischen Polynoms von $\boldsymbol{A}$. Die algebraische Vielfachheit $\sigma(\lambda)$ des Eigenwertes λ ist daher niemals kleiner als seine geometrische Vielfachheit $\varrho(\lambda)$. ❀

Im Falle hermitescher Matrizen folgt aus der komplexen SCHUR-Zerlegung:

9.9. Satz: *Es sei $\boldsymbol{A} \in \mathbb{C}^{n\times n}$ eine hermitesche Matrix ($\boldsymbol{A}^H = \boldsymbol{A}$). Dann gelten folgende Aussagen:*

(i) *$\boldsymbol{A}$ besitzt nur reelle Eigenwerte.*

(ii) *Eigenvektoren zu verschiedenen Eigenwerten sind orthogonal zueinander.*

(iii) *Algebraische und geometrische Vielfachheiten der Eigenwerte stimmen überein.*

Beweis: Nach Satz 9.6 existiert eine unitäre Matrix $\boldsymbol{U} \in \mathbb{C}^{n\times n}$, so dass $\Lambda = \boldsymbol{U}^H\boldsymbol{A}\boldsymbol{U}$ eine obere Dreiecksmatrix ist. Wegen $\boldsymbol{A}^H = \boldsymbol{A}$ folgt

$$\Lambda^H = (\boldsymbol{U}^H\boldsymbol{A}\boldsymbol{U})^H = \boldsymbol{U}^H\boldsymbol{A}^H\boldsymbol{U} = \Lambda.$$

Damit muss Λ reelle Diagonalmatrix sein, in deren Diagonale die reellen Eigenwerte von $\boldsymbol{A}$ stehen. Die Spalten von $\boldsymbol{U}$ sind die zugehörigen Eigenvektoren. Diese sind

paarweise orthogonal und bilden eine Basis des $\mathbb{C}^n$. Damit ist die Summe der geometrischen Vielfachheiten der Eigenwerte gleich n. Mit Satz 9.8 folgt dann, dass die geometrischen Vielfachheiten gleich den algebraischen Vielfachheiten sind. ❀

Im reellen Falle erhalten wir aus dem letzten Satz:

9.10. Satz: *Es sei $\boldsymbol{A}$ eine symmetrische Matrix ($\boldsymbol{A}^T = \boldsymbol{A}$). Dann gelten folgende Aussagen:*

(i) $\boldsymbol{A}$ *besitzt nur reelle Eigenwerte und die Eigenvektoren sind reell wählbar.*

(ii) Eigenvektoren zu verschiedenen Eigenwerten sind orthogonal zueinander.

(iii) Algebraische und geometrische Vielfachheiten der Eigenwerte stimmen überein.

Hermitesche Matrizen sind daher durch unitäre Ähnlichkeitstransformation auf Diagonalform transformierbar. Es gibt noch weitere Matrizen, die ebenfalls diagonalähnlich sind. Das sind gerade alle nichtdefektiven Matrizen: Algebraische und geometrische Vielfachheiten der Eigenwerte stimmen überein. Bevor wir die Diagonalähnlichkeit der nichtdefektiven Matrizen zeigen, benötigen wir noch

9.11. Satz: *Eigenvektoren zu paarweise verschiedenen Eigenwerten sind linear unabhängig.*

Beweis: Es seien $\lambda_1, \ldots, \lambda_s$ paarweise verschiedene Eigenwerte der Matrix $\boldsymbol{A} \in \mathbb{C}^{n\times n}$ mit den zugehörigen Eigenvektoren $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_s \in \mathbb{C}^n$. Wir nehmen an, dass die Vektoren $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_s$ linear abhängig sind. O.B.d.A. seien die Vektoren

$$\boldsymbol{x}_1, \ldots, \boldsymbol{x}_k \quad (k < s)$$

linear unabhängig. Die restlichen Vektoren lassen sich dann als Linearkombinationen dieser linear unabhängigen Vektoren darstellen. Es sei zum Beispiel

$$\boldsymbol{x}_s = \sum_{i=1}^{k} \alpha_i \boldsymbol{x}_i.$$

Es folgt

$$\boldsymbol{A}\boldsymbol{x}_s = \sum_{i=1}^{k} \alpha_i \boldsymbol{A}\boldsymbol{x}_i = \sum_{i=1}^{k} \alpha_i \lambda_i \boldsymbol{x}_i.$$

Andererseits gilt aber

$$\boldsymbol{A}\boldsymbol{x}_s = \lambda_s \boldsymbol{x}_s = \sum_{i=1}^{k} \alpha_i \lambda_s \boldsymbol{x}_i.$$

Daraus folgt

$$\sum_{i=1}^{k}(\lambda_i - \lambda_s)\alpha_i \boldsymbol{x}_i = 0.$$

Wegen der linearen Unabhängigkeit der Vektoren $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_k$ verschwinden die Koeffizienten $(\lambda_i - \lambda_s)\alpha_i$ für $i = 1, \ldots, k$ sämtlich. Da $\lambda_i \neq \lambda_s, \quad i = 1, \ldots, k,$ folgt $\alpha_i = 0, \quad i = 1, \ldots, k,$ und damit $\boldsymbol{x}_s = 0$ als Widerspruch. Damit ist die Annahme $k < s$ falsch. Die Vektoren $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_s$ sind inear unabhängig. ❀

Mit Hilfe dieses Satzes lässt sich nun die Diagonalähnlichkeit der nichtdefektiven Matrizen zeigen.

9.12. Satz: *Zu jeder nichtdefektiven komplexen (n, n)-Matrix $\boldsymbol{A}$ gibt es eine komplexe, reguläre (n, n)-Matrix $\boldsymbol{X}$ mit*

$$\boldsymbol{X}^{-1}\boldsymbol{A}\boldsymbol{X} = \Lambda = \text{diag}(\lambda_1, \ldots, \lambda_n) \in \mathbb{C}^{n\times n}.$$

Falls $\boldsymbol{A}$ reell ist mit nur reellen Eigenwerten, ist $\boldsymbol{X}$ reell wählbar.

Beweis: Da algebraische und geometrische Vielfachheiten der Eigenwerte übereinstimmen, existieren zu jedem Eigenwert $\lambda_i, \quad i = 1, \ldots, sgenau \quad \sigma_i$ linear unabhängige Eigenvektoren. Nach Satz 9.11 existieren damit $\sigma_1 + \sigma_2 + \cdots + \sigma_s = n$ linear unabhängige Eigenvektoren der Matrix $\boldsymbol{A}$. Diese fassen wir zur Matrix $\boldsymbol{X} \in \mathbb{C}^{n\times n}$ zusammen. $\boldsymbol{X}$ ist offensichtlich regulär und es gilt $\boldsymbol{A}\boldsymbol{X} = \boldsymbol{X}\Lambda$. ❀

Im Falle defektiver Matrizen ist nicht mehr zu erwarten, dass sie durch eine Ähnlichkeitstransformation auf Diagonalgestalt transformierbar sind. Hier gilt nur noch der folgende fundamentale Satz, den wir ohne Beweis angeben.

9.13. Satz: *Die Matrix $\boldsymbol{A} \in \mathbb{C}^{n\times n}$ habe die paarweise verschiedenen Eigenwerte*

$$\lambda_1, \ldots, \lambda_s$$

mit den algebraischen bzw. geometrischen Vielfachheiten

$$\sigma_i, \varrho_i, i = 1, \ldots, s.$$

Zu jedem Eigenwert λ_i gibt es dann ϱ_i natürliche Zahlen $\nu_j^{(i)}, j = 1, \ldots, \varrho_i$ mit

$$\sigma_i = \nu_1^{(i)} + \nu_2^{(i)} + \cdots + \nu_{\varrho_i}^{(i)}$$

und eine reguläre Matrix $\boldsymbol{T} \in \mathbb{C}^{n\times n}$, so dass $\boldsymbol{J} = \boldsymbol{T}^{-1}\boldsymbol{A}\boldsymbol{T}$ folgende Gestalt besitzt:

$$\boldsymbol{J} = \begin{pmatrix} \boldsymbol{C}_{\nu_1^{(1)}}(\lambda_1) & & & & & & 0 \\ & \ddots & & & & & \\ & & \boldsymbol{C}_{\nu_{\varrho_1}^{(1)}}(\lambda_1) & & & & \\ & & & \ddots & & & \\ & & & & \boldsymbol{C}_{\nu_1^{(s)}}(\lambda_s) & & \\ & & & & & \ddots & \\ 0 & & & & & & \boldsymbol{C}_{\nu_{\varrho s}^{(s)}}(\lambda_s) \end{pmatrix}.$$

Die Diagonalblöcke $\boldsymbol{C}_\nu(\lambda)$ haben dabei die Form

$$\boldsymbol{C}_\nu(\lambda) = \begin{pmatrix} \lambda & 1 & & 0 \\ & \ddots & \ddots & \\ & & \ddots & 1 \\ 0 & & & \lambda \end{pmatrix}.$$

Die Zahlen $\nu_j^{(i)}$ für $j = 1,\ldots,\varrho_i$ und $i = 1,\ldots,s$ sind bis auf die Reihenfolge eindeutig bestimmt. Die Matrix $\boldsymbol{J}$ heißt JORDANsche *Normalform der Matrix $\boldsymbol{A}$.*

9.2. Störungstheorie

Nun wenden wir uns dem Verhalten der Eigenwerte und Eigenvektoren einer Matrix unter dem Einfluss von Störungen zu und zeigen, dass die Eigenwerte einer Matrix stetige Funktionen der Matrixelemente sind. Da die Koeffizienten des charakteristischen Polynoms einer Matrix offensichtlich stetig von den Matrixelementen abhängen, genügt es zu zeigen, dass die Nullstellen eines Polynoms stetige Funktionen der Koeffizienten des Polynoms sind. Für den Beweis dieser Eigenschaft wenden wir einen Satz aus der Funktionentheorie an.

9.14. ROUCHÉ: *Es sei $D \in \mathbb{C}$ eine offene Kreisscheibe mit Rand γ. Die komplexwertigen Funktionen f und g seien analytisch in einer Umgebung der abgeschlossenen Kreisscheibe $\bar{D} = D \cup \gamma$ und es sei $|f(z) - g(z)| < |f(z)|$ für alle $z \in \gamma$. Dann haben f und g genau gleich viele Nullstellen in D.*

Den Beweis findet man in fast jedem Lehrbuch der Funktionentheorie. Wir zeigen nun, dass die Nullstellen eines Polynoms stetig von den Koeffizienten des Polynoms abhängen.

9.15. Satz: *Es sei*

$$p(z) = z^n + a_{n-1}z^{n-1} + \cdots + a_1 z + a_0$$

ein Polynom mit den Nullstellen $\lambda_1, \ldots, \lambda_n$. Dann gibt es zu jedem $\varepsilon > 0$ ein positives $\delta = \delta(\varepsilon)$ mit der Eigenschaft: Ist

$$q(z) = z^n + b_{n-1}z^{n-1} + \cdots + b_1 z + b_0$$

ein Polynom mit $|b_j - a_j| \leqq \delta$ für $j = 0, \ldots, n-1$, so sind die Nullstellen $\mu_1, \ldots, \mu_n$ von q so nummerierbar, dass $|\mu_i - \lambda_i| \leqq \varepsilon$ für $i = 1, \ldots, n$ gilt.

Beweis: Es seien $\hat{\lambda}_1, \ldots, \hat{\lambda}_m$ die paarweise verschiedenen Nullstellen von p.
O.B.d.A. sei

$$\varepsilon < \frac{1}{2}\min\left\{ \, |\hat{\lambda}_i - \hat{\lambda}_j| \;\middle|\; 1 \leqq i < j \leqq m \right\}.$$

Für jedes $i = 1, \ldots, m$ definieren wir die offene Kreisscheibe

$$D_i = \left\{ \, z \in \mathbb{C} \mid \quad |z - \hat{\lambda}_i| < \varepsilon \right\}$$

mit dem Rand γ_i. Weiterhin sei

$$m_i = \min\{ \, |p(z)| \mid z \in \gamma_i \, \}, \quad M_i = \max\left\{ \sum_{j=0}^{n-1} |z|^j \;\middle|\; z \in \gamma_i \right\}.$$

Wegen

$$\varepsilon < \frac{1}{2}\min\{|\hat{\lambda}_i - \hat{\lambda}_j| : 1 \leqq i < j \leqq m\}$$

besitzt p keine Nullstellen auf γ_i, so dass $m_i > 0$. Wählt man nun $\delta > 0$ so klein, dass $M_i\delta < m_i$ für $i = 1, \ldots, m$ ist, so gilt für ein Polynom

$$q(z) = z^n + b_{n-1}z^{n-1} + \cdots + b_1 z + b_0, \quad |b_j - a_j| \leqq \delta, \quad j = 0, \ldots, n-1$$

$$|q(z) - p(z)| \leqq \sum_{j=0}^{n-1} |b_j - a_j||z|^j \leqq \delta \sum_{j=0}^{n-1} |z|^j \leqq \delta M_i < m_i \leqq |p(z)|, \quad z \in \gamma_i.$$

Aus dem Satz von ROUCHÉ folgt, dass p und q gleiche Anzahl von Nullstellen in D_i haben, nämlich gerade die Vielfachheit von $\hat{\lambda}_i$. Damit sind diese so nummerierbar, dass für $i = 1, \ldots, n$ jeweils $|\mu_i - \lambda_i| \leqq \varepsilon$ gilt. ❀

Unmittelbar aus diesem Satz folgt:

9.16. Satz: *Ist $\boldsymbol{A} \in \mathbb{C}^{n\times n}$ eine Matrix mit den Eigenwerten $\lambda_1, \ldots, \lambda_n$ und $\|\circ\|$ eine Matrixgrenznorm, so gibt es zu jedem $\varepsilon > 0$ ein $\delta = \delta(\varepsilon) > 0$ mit der Eigenschaft: Ist $\boldsymbol{\delta A} \in \mathbb{C}^{n\times n}$ eine Matrix mit $\|\boldsymbol{\delta A}\| \leqq \delta$, so sind die Eigenwerte $\mu_1, \ldots, \mu_n$ von $\boldsymbol{A} + \boldsymbol{\delta A}$ so nummerierbar, dass $|\mu_j - \lambda_j| \leqq \varepsilon$,, $j = 1, \ldots, n$, gilt.*

Aus der stetigen Abhängigkeit der Eigenwerte von den Matrixelementen folgt der nächste Satz, der eine Lokalisierung der Eigenwerte einer Matrix ermöglicht.

9.17. GERSCHGORIN: *Es sei $\boldsymbol{A} = (a_{ij}) \in \mathbb{C}^{n\times n}$. Weiterhin seien durch*

$$G_i = \left\{ z \in \mathbb{C} \mid |z - a_{ii}| \leqq r_i \right\}, \quad r_i = \sum_{\substack{j=1 \\ j\neq i}}^{n} |a_{ij}|, \quad i = 1, \ldots, n$$

die GERSCHGORIN*-Kreise definiert. Dann gilt:*

(i) Ist λ ein Eigenwert von $\boldsymbol{A}$, so ist $\lambda \in \cup_{i=1}^{n} G_i$. Alle Eigenwerte von $\boldsymbol{A}$ sind also in der Vereinigung aller GERSCHGORIN*-Kreise enthalten.*

(ii) Hat die Vereinigung $\hat{G}$ von m Kreisen G_i einen leeren Durchschnitt mit den restlichen $n - m$ Kreisen, so enthält $\hat{G}$ genau m Eigenwerte von $\boldsymbol{A}$, wobei jeder Eigenwert entsprechend seiner algebraischen Vielfachheit zu zählen ist.

Beweis: Es sei λ ein Eigenwert von $\boldsymbol{A}$. Ist $\lambda = a_{ii}$ für ein $i \in \{1, \ldots, n\}$, so gilt die erste Behauptung offensichtlich.
Es sei $\lambda \neq a_{ii}, i = 1, \ldots, n$, $\boldsymbol{D} = \operatorname{diag}(a_{11}, a_{22}, \ldots, a_{nn})$; die Matrix $(\lambda \boldsymbol{I} - \boldsymbol{D})$ ist regulär. Die Matrix $(\lambda \boldsymbol{I} - \boldsymbol{D})^{-1}(\boldsymbol{A} - \boldsymbol{D})$ hat den Eigenwert 1, denn aus $\boldsymbol{A}\boldsymbol{x} = \lambda \boldsymbol{x}$ folgt $(\lambda \boldsymbol{I} - \boldsymbol{D})^{-1}(\boldsymbol{A} - \boldsymbol{D})\boldsymbol{x} = \boldsymbol{x}$. Für eine beliebige Matrixgrenznorm gilt somit:

$$1 \leqq \varrho((\lambda \boldsymbol{I} - \boldsymbol{D})^{-1}(\boldsymbol{A} - \boldsymbol{D})) \leqq \|(\lambda \boldsymbol{I} - \boldsymbol{D})^{-1}(\boldsymbol{A} - \boldsymbol{D})\|.$$

Wählt man als Matrixnorm die ∞-Norm, also die maximale Zeilenbetragssumme, so erhält man:

$$\begin{aligned} 1 \leqq \|(\lambda \boldsymbol{I} - \boldsymbol{D})^{-1}(\boldsymbol{A} - \boldsymbol{D})\|_\infty &= \max \left\{ \sum_{\substack{j=1 \\ j\neq i}}^{n} \frac{|a_{ij}|}{|\lambda - a_{ii}|} \;\middle|\; i = 1, \ldots, n \right\} \\ &= \max \left\{ \frac{r_i}{|\lambda - a_{ii}|} \;\middle|\; i = 1, \ldots, n \right\}. \end{aligned}$$

Daher existiert ein $i \in \{1,\ldots,n\}$ mit $\lambda \in G_i$. Zum Beweis des zweiten Teils des Satzes sei o.B.d.A.

$$\hat{G} = \bigcup_{i=1}^{m} G_i, \quad \tilde{G} = \bigcup_{i=m+1}^{n} G_i, \quad \hat{G} \cap \tilde{G} = \emptyset.$$

Für beliebiges $t \in [0,1]$ definieren wir die Matrix $\boldsymbol{A}(t) = \boldsymbol{D} + t(\boldsymbol{A} - \boldsymbol{D})$. Offensichtlich gilt $\boldsymbol{A}(0) = \boldsymbol{D}$ und $\boldsymbol{A}(1) = \boldsymbol{A}$. Die GERSCHGORIN-Kreise für die Matrix $\boldsymbol{A}(t)$ sind dann durch

$$G_i(t) = \left\{ z \in \mathbb{C} \mid |z - a_{ii}| \leqq t r_i| \right\} \subset G_i, \quad i = 1,\ldots,n$$

gegeben. Für $t \in [0,1]$ liegen daher alle Eigenwerte von $\boldsymbol{A}$(t) in

$$\hat{G} \cup \tilde{G} = \bigcup_{i=1}^{n} G_i.$$

Nun definieren wir

$$I = \left\{ t \in [0,1] \mid \text{ Genau m Eigenwerte von } \boldsymbol{A}(t) \text{ liegen in } \hat{G} \right\}.$$

Weiterhin sei

$$t_0 = \sup\{ t \mid t \in I \}, \quad \varepsilon = \frac{1}{2} \min\left\{ |\hat{z} - \tilde{z}| \;\middle|\; \hat{z} \in \hat{G}, \tilde{z} \in \tilde{G} \right\}.$$

Wegen $\boldsymbol{A}(0) = \boldsymbol{D}$ ist $0 \in I$; I ist daher nichtleer. Nun zeigen wir, dass auch $t_0 \in I$ gilt. Es seien dazu $\lambda_1(t_0),\ldots,\lambda_n(t_0)$ die Eigenwerte von $\boldsymbol{A}(t_0)$. Es gilt

$$\|\boldsymbol{A}(t) - \boldsymbol{A}(t_0)\| = (t - t_0)\|\boldsymbol{A} - \boldsymbol{D}\|.$$

Nach Satz 9.16 gibt es zu ε ein $\delta = \delta(\varepsilon) > 0$, so dass für alle $t \in [0,1]$ mit $|t - t_0| \leqq \delta$ eine Numerierung der Eigenwerte $\lambda_1(t),\ldots,\lambda_n(t)$ von $\boldsymbol{A}(t)$ mit

$$|\lambda_i(t) - \lambda_i(t_0)| \leqq \varepsilon, i = 1,\ldots,n,$$

existiert. Nach Definition von t_0 existiert ein $t \in [t_0 - \delta, t_0] \cap I$. Für dieses t gilt: Genau m Eigenwerte $\lambda_i(t)$ von $\boldsymbol{A}(t)$ liegen in $\hat{G}$, die anderen in $\tilde{G}$. Wegen

$$|\lambda_i(t) - \lambda_i(t_0)| \leqq \varepsilon = \frac{1}{2} \min\left\{ |\hat{z} - \tilde{z}| \;\middle|\; \hat{z} \in \hat{G}, \tilde{z} \in \tilde{G} \right\}$$

liegen die ersten m Eigenwerte von $\boldsymbol{A}(t_0)$ ebenfalls in $\hat{G}$ und die restlichen Eigenwerte in $\tilde{G}$. Damit gilt $t_0 \in I$. Nehmen wir nun an, dass $t_0 < 1$ gilt. Mit $\delta > 0$ und $t \in [t_0, t_0 + \delta] \cap I$ folgern wir genau wie eben, dass die Eigenwerte von $\boldsymbol{A}(t)$ so numerierbar sind, dass $|\lambda_i(t) - \lambda_i(t_0)| \leqq \varepsilon, \quad i = 1,\ldots,n$, gilt. Wegen $t_0 \in I$ ist dann auch $t \in I$ im Widerspruch zur Definition von t_0. Es muss daher $t_0 = 1$ gelten, und der Satz ist bewiesen. ❀

Eigentlich haben wir im ersten Teil des Beweises etwas mehr gezeigt, als zu beweisen war. Sind nämlich $\boldsymbol{A}, \boldsymbol{D} \in \mathbb{C}^{n\times n}$ und ist λ Eigenwert von $\boldsymbol{A}$, so ist entweder λ Eigenwert von $\boldsymbol{D}$ oder

$$1 \leqq \varrho((\lambda \boldsymbol{I} - \boldsymbol{D})^{-1}(\boldsymbol{A} - \boldsymbol{D})) \leqq \|(\lambda \boldsymbol{I} - \boldsymbol{D})^{-1}(\boldsymbol{A} - \boldsymbol{D})\|$$

und

$$1 \leqq \varrho((\boldsymbol{A} - \boldsymbol{D})(\lambda \boldsymbol{I} - \boldsymbol{D})^{-1}) \leqq \|(\boldsymbol{A} - \boldsymbol{D})(\lambda \boldsymbol{I} - \boldsymbol{D})^{-1}\|$$

für jede Matrixgrenznorm $\|\circ\|$. Je nach Wahl von $\boldsymbol{D}$ und der Matrixnorm erhält man so verschiedene Abschätzungen für die Lage der Eigenwerte. Wählt man wieder

$$\boldsymbol{D} = \text{diag}(a_{11}, a_{22}, \ldots, a_{nn}),$$

aber diesmal $\|\circ\| = \|\circ\|_1$, also die Spaltensummennorm, so erhält man:

$$1 \leqq \|(\boldsymbol{A} - \boldsymbol{D})(\lambda \boldsymbol{I} - \boldsymbol{D})^{-1})\|_1 = \max \left\{ \left. \frac{1}{|\lambda - a_{jj}|} \sum_{\substack{i=1 \\ i\neq j}}^{n} |a_{ij}| \right| j = 1, \ldots, n \right\}.$$

Damit gelten die Aussagen des Kreisesatzes von GERSCHGORIN auch für die durch

$$G_j = \left\{ z \in \mathbb{C} \mid |z - a_{jj}| \leqq q_j| \right\}, \quad q_j = \sum_{\substack{i=1 \\ i\neq j}}^{n} |a_{ij}|, \quad j = 1, \ldots, n$$

definierten Kreise.
Verwendet man die FROBENIUS-Norm, so existiert zu jedem Eigenwert ein $k \in \{1, \ldots, n\}$ mit

$$|\lambda - a_{kk}| \leqq \left(\sum_{\substack{i,j=1 \\ i\neq j}}^{n} |a_{ij}|^2 \right)^{1/2}.$$

Der Kreisesatz von GERSCHGORIN ist natürlich auch auf Matrizen anwendbar, die die gleichen Eigenwerte wie $\boldsymbol{A}$ besitzen, z. B. $\boldsymbol{A}^H$ oder jede durch Ähnlichkeitstransformation aus $\boldsymbol{A}$ hervorgegangenen Matrix.

9.18. Beispiel: Wir betrachten die Matrix

$$\boldsymbol{A} = \begin{pmatrix} 0.9 & 0.0 & 0.0 \\ 0.0 & 0.4 & 0.0 \\ 0.0 & 0.0 & 0.2 \end{pmatrix} + 10^{-5} \begin{pmatrix} 0.1 & 0.4 & -0.2 \\ -0.1 & 0.5 & 0.1 \\ 0.2 & 0.1 & 0.3 \end{pmatrix}.$$

Nach Satz 9.17 ergeben sich folgende Aussagen über die Lokalisierung der Eigenwerte:

$$|\lambda_1 - (0.9 + 0.1 \cdot 10^{-5})| \leqq 0.6 \cdot 10^{-5},$$
$$|\lambda_2 - (0.4 + 0.5 \cdot 10^{-5})| \leqq 0.2 \cdot 10^{-5},$$
$$|\lambda_3 - (0.2 + 0.3 \cdot 10^{-5})| \leqq 0.3 \cdot 10^{-5}.$$

Wendet man den Satz 9.17 auf die zu $\boldsymbol{A}$ ähnliche Matrix

$$\boldsymbol{P}^{-1}\boldsymbol{A}\boldsymbol{P}, \quad \boldsymbol{P} = \text{diag}(10^5, 1, 1)$$

an, so erhält man für den ersten Eigenwert folgende Abschätzung:

$$|\lambda_1 - (0.9 + 0.1 \cdot 10^{-5})| \leqq 0.6 \cdot 10^{-10}.$$

Die beiden anderen GERSCHGORIN-Kreise sind nicht mehr disjunkt. Eine verbesserte Abschätzung für den zweiten Eigenwert erhält man, falls man $\boldsymbol{P} = \text{diag}(1, 10^5, 1)$ verwendet:

$$|\lambda_2 - (0.4 + 0.5 \cdot 10^{-5})| \leqq 0.2 \cdot 10^{-10}.$$

Analog ergibt sich mit $\boldsymbol{P} = \text{diag}(1, 1, 10^5)$

$$|\lambda_3 - (0.2 + 0.3 \cdot 10^{-5})| \leqq 0.3 \cdot 10^{-10}.$$

♡

Der Kreisesatz von GERSCHGORIN diente dazu, die mögliche Lage der Eigenwerte einer gegebenen Matrix einzugrenzen. Wir wollen uns nun dem eigentlichen Thema dieses Abschnitts zuwenden, nämlich der Frage: Wie verhalten sich die Eigenwerte einer Matrix, falls die Matrixelemente gestört werden? Wir betrachten neben der Matrix $\boldsymbol{A} \in \mathbb{C}^{n\times n}$ eine Störung $\boldsymbol{\delta A} \in \mathbb{C}^{n\times n}$ und fragen nach dem Zusammenhang zwischen den Eigenwerten der Matrix $\boldsymbol{A}$ und der Matrix $\boldsymbol{A} + \boldsymbol{\delta A}$. Für beliebige Matrizen ist es wieder schwierig, genauere Aussagen zu treffen. Grenzt man die Klasse der zu betrachtenden Matrizen etwas ein (diagonalisierbare Matrizen, hermitesche Matrizen) so lässt sich das Verhalten der Eigenwerte dieser Matrizen gegenüber Störungen genauer beschreiben. Eine Matrix $\boldsymbol{A} \in \mathbb{C}^{n\times n}$ heißt **diagonalisierbar**, falls eine reguläre Matrix $\boldsymbol{X} \in \mathbb{C}^{n\times n}$ mit $\boldsymbol{X}^{-1}\boldsymbol{A}\boldsymbol{X} = \text{diag}(\lambda_1, \ldots, \lambda_n) = \boldsymbol{D}$ existiert.

9.19. BAUER-FIKE**:** *Es sei $\boldsymbol{A} \in \mathbb{C}^{n\times n}$ eine diagonalisierbare Matrix:*

$$\boldsymbol{X}^{-1}\boldsymbol{A}\boldsymbol{X} = \text{diag}(\lambda_1, \ldots, \lambda_n) = \boldsymbol{D}.$$

Es sei weiterhin $\boldsymbol{\delta A} \in \mathbb{C}^{n\times n}$ und λ ein Eigenwert von $\boldsymbol{A} + \boldsymbol{\delta A}$. Dann gilt bezüglich einer Matrixgrenznorm

$$\min\{\, |\lambda - \lambda_j| \mid j = 1, \ldots, n \,\} \leqq \text{cond}(\boldsymbol{X})\|\boldsymbol{\delta A}\|.$$

Beweis: Ist λ auch ein Eigenwert von $\boldsymbol{A}$, so ist die Behauptung offensichtlich wahr. Es sei nun $\lambda \neq \lambda_j$ für $j = 1, \ldots, n$. Der Vektor $\boldsymbol{x}$ sei ein zu λ gehörender Eigenvektor: $(\boldsymbol{A} + \boldsymbol{\delta A})\boldsymbol{x} = \lambda \boldsymbol{x}$. Dann folgt

$$\boldsymbol{\delta A x} = (\lambda \boldsymbol{I} - \boldsymbol{A})\boldsymbol{x} = (\lambda \boldsymbol{I} - \boldsymbol{X D X}^{-1})\boldsymbol{x} = \boldsymbol{X}(\lambda \boldsymbol{I} - \boldsymbol{D})\boldsymbol{X}^{-1}\boldsymbol{x}$$

und daraus

$$\boldsymbol{X}^{-1}\boldsymbol{x} = (\lambda \boldsymbol{I} - \boldsymbol{D})^{-1}(\boldsymbol{X}^{-1}\boldsymbol{\delta A X})\boldsymbol{X}^{-1}\boldsymbol{x}.$$

Wegen der Submultiplikativität der Matrixgrenznorm erhalten wir

$$\begin{aligned} \|\boldsymbol{X}^{-1}\boldsymbol{x}\| &\leqq \|(\lambda \boldsymbol{I} - \boldsymbol{D})^{-1}\| \, \|\boldsymbol{X}^{-1}\boldsymbol{\delta A X}\| \, \|\boldsymbol{X}^{-1}\boldsymbol{x}\| \\ &= \max\left\{ \frac{1}{|\lambda - \lambda_j|} \;\middle|\; j = 1, \ldots, n \right\} \|\boldsymbol{X}^{-1}\boldsymbol{\delta A X}\| \, \|\boldsymbol{X}^{-1}\boldsymbol{x}\| \end{aligned}$$

und weiter

$$\min\{\, |\lambda - \lambda_j| \mid j = 1, \ldots, n \,\} \leqq \operatorname{cond}(\boldsymbol{X}) \|\boldsymbol{\delta A}\|.$$

❀

Die Kondition der Matrix $\boldsymbol{X}$, die $\boldsymbol{A}$ diagonalisiert, ist daher ein Maß für die Störanfälligkeit der Eigenwerte von $\boldsymbol{A}$. Für hermitesche Matrizen folgt sofort:

9.20. Satz: *Es sei $\boldsymbol{A} \in \mathbb{C}^{n \times n}$ eine hermitesche Matrix und $\boldsymbol{\delta A} \in \mathbb{C}^{n \times n}$. $\lambda_1, \ldots, \lambda_n$ seien die (reellen) Eigenwerte von $\boldsymbol{A}$ und λ sei ein Eigenwert der gestörten Matrix $\boldsymbol{A} + \boldsymbol{\delta A}$. Dann gilt bezüglich der Spektralnorm*

$$\min\{\, |\lambda - \lambda_j| \mid j = 1, \ldots, n \,\} \leqq \|\boldsymbol{\delta A}\|_2.$$

Beweis: Für hermitesche Matrizen ist die Matrix $\boldsymbol{X}$, die $\boldsymbol{A}$ diagonalisiert, unitär wählbar. Damit ist $\operatorname{cond}_2(\boldsymbol{X}) = 1$ und mit Satz 9.19 folgt die Behauptung. ❀

Wie wir sehen, ist das Eigenwertproblem für hermitesche Matrizen (oder im Reellen für symmetrische Matrizen) besonders gut konditioniert. In diesem Falle lässt sich die Aussage des letzten Satzes sogar noch verschärfen. Vorher noch zwei wichtige Sätze über hermitesche Matrizen.

9.21. RAYLEIGH: *Für die hermitesche Matrix $\boldsymbol{A} \in \mathbb{C}^{n \times n}$ seien die (reellen) Eigenwerte gemäß $\lambda_1 \geqq \cdots \geqq \lambda_n$ geordnet. Durch $\{\boldsymbol{u}_1, \ldots, \boldsymbol{u}_n\}$ sei ein zugehöriges Orthonormalsystem von Eigenvektoren gegeben. Es gilt daher*

$$\boldsymbol{A}\boldsymbol{u}_i = \lambda_i \boldsymbol{u}_i, \quad \boldsymbol{u}_i^H \boldsymbol{u}_j = \delta_{ij}, \quad 1 \leqq i, j \leqq n.$$

Für $j = 1, \ldots, n$ definieren wir den $(n+1-j)$-dimensionalen Teilraum

$$\boldsymbol{\mathcal{M}}_j = \text{span}(\boldsymbol{u}_j, \ldots, \boldsymbol{u}_n).$$

Dann gilt

$$\lambda_j = \max \left\{ \frac{\boldsymbol{x}^H \boldsymbol{A} \boldsymbol{x}}{\boldsymbol{x}^H \boldsymbol{x}} \;\middle|\; \boldsymbol{x} \in \boldsymbol{\mathcal{M}}_j \setminus \{\boldsymbol{o}\} \right\}, \quad j = 1, \ldots, n.$$

Beweis: Für festes $j \in \{1, \ldots, n\}$ sei $\boldsymbol{x} \in \boldsymbol{\mathcal{M}}_j \setminus \{\boldsymbol{o}\}$ beliebig. Dann gilt

$$\boldsymbol{x} = \alpha_j \boldsymbol{u}_j + \cdots + \alpha_n \boldsymbol{u}_n$$

und

$$\boldsymbol{A}\boldsymbol{x} = \lambda_j \alpha_j \boldsymbol{u}_j + \cdots + \lambda_n \alpha_n \boldsymbol{u}_n.$$

Damit ergibt sich

$$\frac{\boldsymbol{x}^H \boldsymbol{A} \boldsymbol{x}}{\boldsymbol{x}^H \boldsymbol{x}} = \frac{\sum_{i=j}^n \lambda_i |\alpha_i|^2}{\sum_{i=j}^n |\alpha_i|^2} \leqq \lambda_j.$$

Es folgt daher

$$\sup \left\{ \frac{\boldsymbol{x}^H \boldsymbol{A} \boldsymbol{x}}{\boldsymbol{x}^H \boldsymbol{x}} \;\middle|\; \boldsymbol{x} \in \boldsymbol{\mathcal{M}}_j \setminus \{\boldsymbol{o}\} \right\} \leqq \lambda_j.$$

Andererseits ist $\boldsymbol{u}_j \in \boldsymbol{\mathcal{M}}_j$ und $\boldsymbol{u}_j^H \boldsymbol{A} \boldsymbol{u}_j / \boldsymbol{u}_j^H \boldsymbol{u}_j = \lambda_j$, so dass

$$\max \left\{ \frac{\boldsymbol{x}^H \boldsymbol{A} \boldsymbol{x}}{\boldsymbol{x}^H \boldsymbol{x}} \;\middle|\; \boldsymbol{x} \in \boldsymbol{\mathcal{M}}_j \setminus \{\boldsymbol{o}\} \right\} = \lambda_j.$$

❀

Bemerkungen: (i) Für eine hermitesche Matrix wird der Ausdruck

$$RQ(\boldsymbol{x}) = \boldsymbol{x}^H \boldsymbol{A} \boldsymbol{x} / \boldsymbol{x}^H \boldsymbol{x}$$

als RAYLEIGH-Quotient bezeichnet. Ist $\boldsymbol{x}$ ein Eigenvektor von $\boldsymbol{A}$ so ist $RQ(\boldsymbol{x})$ gleich dem zugehörigen Eigenwert. Ist $\boldsymbol{x}$ nur eine Näherung für einen Eigenvektor von $\boldsymbol{A}$, so erwartet man, dass durch $RQ(\boldsymbol{x})$ eine Näherung für den zugehörigen Eigenwert gegeben ist.
(ii) Ist $\boldsymbol{A}$ symmetrisch und reell, so gilt der Satz in analoger Weise im Reellen.

9.22. COURANT: *Für die hermitesche Matrix $\boldsymbol{A} \in \mathbb{C}^{n\times n}$ seien die (reellen) Eigenwerte gemäß $\lambda_1 \geqq \cdots \geqq \lambda_n$ geordnet. Für $j = 1,\ldots,n$ definieren wir die Mengen*

$$\mathcal{N}_j = \left\{ \boldsymbol{\mathcal{N}}_j \subseteqq \mathbb{C}^{n\times n} \;\middle|\; \boldsymbol{\mathcal{N}}_j \text{ ist linearer Teilraum der Dimension } n+1-j \right\}.$$

Dann ist

$$\lambda_j = \min\left\{ \max\left\{ \frac{\boldsymbol{x}^H\boldsymbol{A}\boldsymbol{x}}{\boldsymbol{x}^H\boldsymbol{x}} \;\middle|\; \boldsymbol{x} \in \boldsymbol{\mathcal{N}}_j \setminus \{\boldsymbol{o}\} \right\} \;\middle|\; \boldsymbol{\mathcal{N}}_j \in \mathcal{N}_j \right\}, \quad j = 1,\ldots,n.$$

Beweis: Es sei

$$\{\boldsymbol{u}_1,\ldots,\boldsymbol{u}_n\}$$

ein Orthonormalsystem von Eigenvektoren zu den Eigenwerten $\lambda_1,\ldots,\lambda_n$. Für festes $j \in \{1,\ldots,n\}$ definieren wir den linearen Teilraum

$$\boldsymbol{\mathcal{L}}_j = \operatorname{span}(\boldsymbol{u}_1,\ldots,\boldsymbol{u}_j).$$

Der Durchschnitt von $\boldsymbol{\mathcal{L}}_j$ mit einem beliebigen Teilraum $\boldsymbol{\mathcal{N}}_j \in \mathcal{N}_j$ enthält wegen

$$\dim(\boldsymbol{\mathcal{L}}_j \cap \boldsymbol{\mathcal{N}}_j) = \dim(\boldsymbol{\mathcal{L}}_j) + \dim(\boldsymbol{\mathcal{N}}_j) - \dim(\boldsymbol{\mathcal{L}}_j \oplus \boldsymbol{\mathcal{N}}_j) \geqq 1$$

ein $\boldsymbol{x} \neq \boldsymbol{o}$. Aus $\boldsymbol{x} \in \boldsymbol{\mathcal{L}}_j$ folgt $\boldsymbol{x}^H\boldsymbol{A}\boldsymbol{x}/\boldsymbol{x}^H\boldsymbol{x} \geqq \lambda_j$ und damit

$$\min\left\{ \max\left\{ \frac{\boldsymbol{x}^H\boldsymbol{A}\boldsymbol{x}}{\boldsymbol{x}^H\boldsymbol{x}} \;\middle|\; \boldsymbol{x} \in \boldsymbol{\mathcal{N}}_j \setminus \{\boldsymbol{o}\} \right\} \;\middle|\; \boldsymbol{\mathcal{N}}_j \in \mathcal{N}_j \right\} \geqq \lambda_j.$$

Wählt man nun

$$\boldsymbol{\mathcal{N}}_j = \boldsymbol{\mathcal{M}}_j = \left\{ \boldsymbol{x} \in \mathbb{C}^n \mid \boldsymbol{u}_i^H\boldsymbol{x} = 0, \quad i = 1,\ldots,j-1 \right\},$$

so folgt nach Satz 9.21

$$\max\left\{ \frac{\boldsymbol{x}^H\boldsymbol{A}\boldsymbol{x}}{\boldsymbol{x}^H\boldsymbol{x}} \;\middle|\; \boldsymbol{x} \in \boldsymbol{\mathcal{M}}_j \setminus \{\boldsymbol{o}\} \right\} = \lambda_j.$$

Damit ist der Satz bewiesen. ❀

Nun sind wir in der Lage, das Verhalten der Eigenwerte einer hermiteschen Matrix $\boldsymbol{A} \in \mathbb{C}^{n\times n}$ unter einer hermiteschen Störung $\boldsymbol{\delta A} \in \mathbb{C}^{n\times n}$ genauer anzugeben.

9.23. Satz: *Die hermitesche Matrix $\boldsymbol{A} \in \mathbb{C}^{n\times n}$ habe die (reellen) Eigenwerte*

$$\lambda_1 \geqq \cdots \geqq \lambda_n.$$

Für eine hermitesche Störung $\boldsymbol{\delta A} \in \mathbb{C}^{n\times n}$ betrachten wir die Matrix $\boldsymbol{A}+\boldsymbol{\delta A} \in \mathbb{C}^{n\times n}$ mit den (reellen) Eigenwerten $\mu_1 \geqq \cdots \geqq \mu_n$. Bezüglich jeder Matrixgrenznorm gilt dann die Abschätzung

$$|\lambda_j - \mu_j| \leqq \|\boldsymbol{\delta A}\|.$$

Beweis: Da $\boldsymbol{A}$ und $\boldsymbol{\delta A}$ hermitesch sind, gilt

$$\frac{\boldsymbol{x}^H(\boldsymbol{A}-(\boldsymbol{A}+\boldsymbol{\delta A}))\boldsymbol{x}}{\boldsymbol{x}^H\boldsymbol{x}} = \frac{\boldsymbol{x}^H(-\boldsymbol{\delta A})\boldsymbol{x}}{\boldsymbol{x}^H\boldsymbol{x}} \leqq \|\boldsymbol{\delta A}\|_2 = \varrho(\boldsymbol{\delta A}) \leqq \|\boldsymbol{\delta A}\|.$$

Daraus folgt

$$\frac{\boldsymbol{x}^H\boldsymbol{A}\boldsymbol{x}}{\boldsymbol{x}^H\boldsymbol{x}} \leqq \frac{\boldsymbol{x}^H(\boldsymbol{A}+\boldsymbol{\delta A})\boldsymbol{x}}{\boldsymbol{x}^H\boldsymbol{x}} + \|\boldsymbol{\delta A}\|.$$

Für festes $j \in \{1,\ldots,n\}$ folgt dann wie in Satz 9.22 für ein beliebiges $\boldsymbol{\mathcal{N}}_j \in \mathcal{N}_j$

$$\begin{aligned}
&\max\left\{ \frac{\boldsymbol{x}^H\boldsymbol{A}\boldsymbol{x}}{\boldsymbol{x}^H\boldsymbol{x}} \;\middle|\; \boldsymbol{x} \in \boldsymbol{\mathcal{N}}_j \setminus \{\boldsymbol{o}\} \right\} \\
&\quad \leqq \max\left\{ \frac{\boldsymbol{x}^H(\boldsymbol{A}+\boldsymbol{\delta A})\boldsymbol{x}}{\boldsymbol{x}^H\boldsymbol{x}} \;\middle|\; \boldsymbol{x} \in \boldsymbol{\mathcal{N}}_j \setminus \{\boldsymbol{o}\} \right\} + \|\boldsymbol{\delta A}\|
\end{aligned}$$

und weiter

$$\begin{aligned}
&\min\left\{ \max\left\{ \frac{\boldsymbol{x}^H\boldsymbol{A}\boldsymbol{x}}{\boldsymbol{x}^H\boldsymbol{x}} \;\middle|\; \boldsymbol{x} \in \boldsymbol{\mathcal{N}}_j \setminus \{\boldsymbol{o}\} \right\} \;\middle|\; \boldsymbol{\mathcal{N}}_j \in \mathcal{N}_j \right\} \\
&\quad \leqq \min\left\{ \max\left\{ \frac{\boldsymbol{x}^H(\boldsymbol{A}+\boldsymbol{\delta A})\boldsymbol{x}}{\boldsymbol{x}^H\boldsymbol{x}} \;\middle|\; \boldsymbol{x} \in \boldsymbol{\mathcal{N}}_j \setminus \{\boldsymbol{o}\} \right\} \;\middle|\; \boldsymbol{\mathcal{N}}_j \in \mathcal{N}_j \right\} + \|\boldsymbol{\delta A}\|.
\end{aligned}$$

Nach Satz 9.22 heißt das

$$\lambda_j \leqq \mu_j + \|\boldsymbol{\delta A}\|, \quad j = 1,\ldots,n.$$

Vertauscht man $\boldsymbol{A}$ und $\boldsymbol{A}+\boldsymbol{\delta A}$, so erhält man

$$\mu_j \leqq \lambda_j + \|\boldsymbol{\delta A}\|, \quad j = 1,\ldots,n,$$

insgesamt also

$$|\lambda_j - \mu_j| \leqq \|\boldsymbol{\delta A}\|, \quad j = 1,\ldots,n.$$

❀

Die letzten Sätze haben gezeigt, dass die Eigenwerte einer nichtdefektiven Matrix lokal lipschitzstetige Funktionen der Matrixelemente sind. Im Falle hermitescher Matrizen ist die Lipschitzkonstante gleich 1. Das Eigenwertproblem für hermitesche Matrizen (oder im Reellen für symmetrische Matrizen) ist besonders gut konditioniert.
Wir wollen nun noch untersuchen, wie sich die Eigenvektoren einer Matrix gegenüber Störungen verhalten.

9.24. Satz: *Die hermitesche Matrix $\boldsymbol{A} \in \mathbb{C}^{n\times n}$ habe die (reellen) Eigenwerte*

$$\lambda_1, \ldots, \lambda_n$$

und die zugehörigen orthonormierten Eigenvektoren

$$\boldsymbol{u}_1, \ldots, \boldsymbol{u}_n.$$

Weiterhin sei $\boldsymbol{\delta A} \in \mathbb{C}^{n\times n}$ und $(\mu, \boldsymbol{u})$ sei ein Eigenpaar von $\boldsymbol{A} + \boldsymbol{\delta A}$. Dann gelten folgende Aussagen:

(i) Falls für ein $j \in \{1, \ldots, n\}$ $\quad \boldsymbol{u}_j^H \boldsymbol{u} \neq 0$, so gilt

$$|\mu - \lambda_j| \leqq \frac{\|\boldsymbol{\delta A u}\|_2}{\boldsymbol{u}_j^H \boldsymbol{u}} \leqq \frac{\|\boldsymbol{\delta A}\|_2}{\boldsymbol{u}_j^H \boldsymbol{u}}. \tag{9.1}$$

(ii) Falls für ein $j \in \{1, \ldots, n\}$ $\quad \mu \neq \lambda_j$, so gilt

$$|\cos(\sphericalangle(\boldsymbol{u}_j, \boldsymbol{u}))| \leqq \frac{\|\boldsymbol{\delta A u}\|_2}{|\mu - \lambda_j|} \leqq \frac{\|\boldsymbol{\delta A}\|_2}{|\mu - \lambda_j|}. \tag{9.2}$$

(iii) Falls für ein $j \in \{1, \ldots, n\}$

$$\gamma_j = \min\{\, |\mu - \lambda_i| \mid i = 1, \ldots, n, i \neq j \,\} > 0$$

so gilt

$$\sin(\sphericalangle(\boldsymbol{u}_j, \boldsymbol{u})) \leqq \frac{\|\boldsymbol{\delta A u}\|_2}{\gamma_j} \leqq \frac{\|\boldsymbol{\delta A}\|_2}{\gamma_j}. \tag{9.3}$$

Beweis: Wir stellen den Vektor $\boldsymbol{u}$ als Linearkombination der Eigenvektoren der Matrix $\boldsymbol{A}$ dar:

$$\boldsymbol{u} = \sum_{i=1}^{n} \zeta_i \boldsymbol{u}_i = \boldsymbol{U}\boldsymbol{z}, \quad \boldsymbol{z} = (\zeta_1, \ldots, \zeta_n)^T, \quad \|\boldsymbol{z}\|_2 = 1.$$

Dann gilt $\boldsymbol{z} = \boldsymbol{U}^H \boldsymbol{u}$ und $\zeta_i = \boldsymbol{u}_i^H \boldsymbol{u} = \cos(\sphericalangle(\boldsymbol{u}_i, \boldsymbol{u}))$ für $i = 1, \ldots, n$. Es sei weiter

$$\begin{aligned}
\boldsymbol{y} &= (\mu \boldsymbol{I} - \Lambda)\boldsymbol{z} \\
&= (\mu \boldsymbol{U}^H \boldsymbol{U} - \boldsymbol{U}^H \boldsymbol{A}\boldsymbol{U})\boldsymbol{z} \\
&= \boldsymbol{U}^H(\mu \boldsymbol{I} - \boldsymbol{A})\boldsymbol{U}\boldsymbol{z} \\
&= \boldsymbol{U}^H(\mu \boldsymbol{I} - (\boldsymbol{A} + \boldsymbol{\delta A}) + \boldsymbol{\delta A})\boldsymbol{u} \\
&= \boldsymbol{U}^H(\mu \boldsymbol{I} - (\boldsymbol{A} + \boldsymbol{\delta A}))\boldsymbol{u} + \boldsymbol{U}^H \boldsymbol{\delta A}\boldsymbol{u} \\
&= \boldsymbol{U}^H \boldsymbol{\delta A}\boldsymbol{u}.
\end{aligned}$$

Für eine Komponente η_i des Vektors $\boldsymbol{y}$ gilt dann $\eta_i = (\mu - \lambda_i)\zeta_i = \boldsymbol{u}_i^H \boldsymbol{\delta A}\boldsymbol{u}$. Ist nun j ein Index, für den $\zeta_j = \boldsymbol{u}_j^H \boldsymbol{u} \neq 0$, so folgt sofort

$$|\mu - \lambda_j| = \frac{|\boldsymbol{u}_j^H \boldsymbol{\delta A}\boldsymbol{u}|}{\boldsymbol{u}_j^H \boldsymbol{u}} \leqq \frac{\|\boldsymbol{u}_j\|_2 \|\boldsymbol{\delta A}\boldsymbol{u}\|_2}{\boldsymbol{u}_j^H \boldsymbol{u}} = \frac{\|\boldsymbol{\delta A}\boldsymbol{u}\|_2}{\boldsymbol{u}_j^H \boldsymbol{u}} \leqq \frac{\|\boldsymbol{\delta A}\|_2}{\boldsymbol{u}_j^H \boldsymbol{u}},$$

also der erste Teil des Satzes.
Gilt für ein $j \in \{1, \ldots, n\}$ $\mu \neq \lambda_j$, so lässt sich abschätzen, wie stark $\boldsymbol{u}$ und $\boldsymbol{u}_j$ von der Orthogonalität abweichen:

$$|\cos(\sphericalangle(\boldsymbol{u}_j, \boldsymbol{u}))| = |\zeta_j| = \frac{|\boldsymbol{u}_j^H \boldsymbol{\delta A}\boldsymbol{u}|}{|\mu - \lambda_j|} \leqq \frac{\|\boldsymbol{\delta A}\boldsymbol{u}\|_2}{|\mu - \lambda_j|}.$$

Damit wäre die zweite Aussage des Satzes bewiesen. Für den Beweis der dritten Aussage beachte man, dass für

$$\gamma_j = \min\{|\mu - \lambda_j| : i = 1, \ldots, n, i \neq j\} > 0$$

$$(\sin(\sphericalangle(\boldsymbol{u}_j, \boldsymbol{u})))^2 = 1 - \zeta_j^2 = \sum_{\substack{i=1 \\ i \neq j}}^{n} \zeta_i^2 = \sum_{\substack{i=1 \\ i \neq j}}^{n} \frac{|\eta_i|^2}{|\mu - \lambda_j|^2} \leqq \frac{\|\boldsymbol{y}\|_2^2}{\gamma_j^2} = \frac{\|\boldsymbol{\delta A}\boldsymbol{u}\|_2^2}{\gamma_j^2}$$

folgt. Nach Wurzelziehen erhält man die letzte Aussage des Satzes. ❀

Bemerkungen: (i) Es seien

$$\lambda_1 \geqq \lambda_2 \geqq \cdots \geqq \lambda_n \text{ bzw. } \mu_1 \geqq \mu_2 \geqq \cdots \geqq \mu_n$$

die geordneten Eigenwerte der hermiteschen Matrizen $\boldsymbol{A}$ bzw. $\boldsymbol{A} + \boldsymbol{\delta A}$. Ferner seien

$$\{\boldsymbol{u}_1, \ldots, \boldsymbol{u}_n\} \text{ und } \{\boldsymbol{v}_1, \ldots, \boldsymbol{v}_n\}$$

zugehörige Orthonormalsysteme von Eigenvektoren. Setlzen wir nun

$$\{\mu, \boldsymbol{u}\} = \{\mu_i, \boldsymbol{v}_i\},$$

so lautet (9.2):

$$|\cos(\sphericalangle(\boldsymbol{u}_j, \boldsymbol{v}_i))| \leqq \frac{\|\boldsymbol{\delta A}\boldsymbol{v}_i\|_2}{|\mu_i - \lambda_j|} \leqq \frac{\|\boldsymbol{\delta A}\|_2}{|\mu_i - \lambda_j|}.$$

Sind die Eigenwerte μ_i und λ_j genügend stark voneinander getrennt, und ist die Störung $\boldsymbol{\delta A}$ hinreichend klein, so sind die Eigenvektoren $\boldsymbol{v}_i$ und $\boldsymbol{u}_j$ fast orthogonal. Nach Satz 9.23 gilt aber

$$|\mu_i - \lambda_j| = |\mu_i - \lambda_i + \lambda_i - \lambda_j| \geqq |\lambda_i - \lambda_j| - |\mu_i - \lambda_i| \geqq |\lambda_i - \lambda_j| - \|\boldsymbol{\delta A}\|_2.$$

Damit ist μ_i hinreichend stark von λ_j getrennt, falls λ_i hinreichend stark von λ_j getrennt ist und die Störung $\boldsymbol{\delta A}$ hinreichend klein ist. Die Orthogonalität zweier Eigenvektoren einer hermiteschen Matrix wird durch eine kleine Störung wenig beeinflusst, falls die zugehörigen Eigenwerte weit genug voneinander entfernt sind.
(ii) Aus (9.3) erhalten wir für $\{\mu, \boldsymbol{u}\} = \{\mu_j, \boldsymbol{v}_j\}$

$$\sin(\sphericalangle(\boldsymbol{u}_j, \boldsymbol{v}_j)) \leqq \frac{\|\boldsymbol{\delta A}\boldsymbol{v}_j\|_2}{\gamma_j} \leqq \frac{\|\boldsymbol{\delta A}\|_2}{\gamma_j}, \quad \gamma_j = \min\{\, |\mu_j - \lambda_i| \mid i = 1, \ldots, n, i \neq j \,\}.$$

Sind γ_j hinreichend groß und $\|\boldsymbol{\delta A}\|_2$ hinreichend klein, so weichen die Richtungen der Eigenvektoren $\boldsymbol{u}_j$ und $\boldsymbol{v}_j$ nur wenig voneinander ab. Wegen

$$\gamma_j \geqq \min\{\, |\lambda_j - \lambda_i| \mid i = 1, \ldots, n, i \neq j \,\} - \|\boldsymbol{\delta A}\|_2$$

ist dies der Fall, falls der Eigenwert λ_j zusätzlich genügend stark von allen anderen Eigenwerten λ_i getrennt ist. Für isolierte Eigenwerte einer hermiteschen Matrix haben daher kleine Störungen nur geringen Einfluss auf die Richtungen der zugehörigen Eigenvektoren.
(iii) In vielen Fällen gilt $\|\boldsymbol{\delta A}\boldsymbol{u}\|_2 \ll \|\boldsymbol{\delta A}\|_2\|\boldsymbol{u}\|_2$. Die Schranken mit $\|\boldsymbol{\delta A}\boldsymbol{u}\|_2$ sind dann wesentlich besser als die, die durch die Vergröberung $\|\boldsymbol{\delta A}\boldsymbol{u}\|_2 \leqq \|\boldsymbol{\delta A}\|_2$ entstehen. Da $\boldsymbol{\delta A}$ als hermitesch vorausgesetzt wurde, lässt sich wegen

$$|\boldsymbol{u}_i^H \boldsymbol{\delta A}\boldsymbol{u}| = |(\boldsymbol{\delta A}\boldsymbol{u})_i^H \boldsymbol{u}| \leqq \|\boldsymbol{\delta A}\boldsymbol{u}_i\|_2$$

der Term $\|\boldsymbol{\delta A}\boldsymbol{u}\|_2$ in (9.1) bzw. (9.2) durch $\|\boldsymbol{\delta A}\boldsymbol{u}_j\|_2$ bzw. $\|\boldsymbol{\delta A}\boldsymbol{u}_i\|_2$ ersetzen.
Als letztes wollen wir in diesem Abschnitt noch sogenannte Residualkriterien angeben, die es uns ermöglichen, von einem Paar $\{\mu, \boldsymbol{u}\}$ zu entscheiden, ob es im Rahmen einer vorgegebenen Genauigkeit als Eigenpaar einer Matrix $\boldsymbol{A}$ akzeptierbar ist.

9.25. Satz: *Es seien $\boldsymbol{A} \in \mathbb{C}^{n\times n}$, $\mu \in \mathbb{C}$ und $\boldsymbol{u} \in \mathbb{C}^n$ mit $\|\boldsymbol{u}\|_2 = 1$ gegeben. Dann sind folgende Aussagen äquivalent:*

1. *Es existiert ein $\boldsymbol{\delta A} \in \mathbb{C}^{n\times n}$ mit*

$$(\boldsymbol{A} + \boldsymbol{\delta A})\boldsymbol{u} = \mu\boldsymbol{u}, \quad \|\boldsymbol{\delta A}\|_2 \leqq \varepsilon.$$

2. *Es gilt*

$$\|\boldsymbol{r}\|_2 \leqq \varepsilon, \quad \boldsymbol{r} = \boldsymbol{A}\boldsymbol{u} - \mu\boldsymbol{u}.$$

Der Vektor $\boldsymbol{r} = \boldsymbol{A}\boldsymbol{u} - \mu\boldsymbol{u}$ heißt Residuum des Paares $\{\mu, \boldsymbol{u}\}$ in Bezug auf das Eigenwertproblem der Matrix $\boldsymbol{A}$.

Beweis: Aus $\boldsymbol{r} = \boldsymbol{A}\boldsymbol{u} - \mu\boldsymbol{u}$ und $(\boldsymbol{A} + \boldsymbol{\delta A})\boldsymbol{u} = \mu\boldsymbol{u}$ folgt

$$\boldsymbol{r} = \| - \boldsymbol{\delta A}\boldsymbol{u}\|_2 \leqq \|\boldsymbol{\delta A}\|_2 \leqq \varepsilon.$$

Es sei nun $\eta = \|\boldsymbol{r}\|_2 \leqq \varepsilon$. Dann existiert eine HOUSEHOLDER-Spiegelung $\boldsymbol{H} \in \mathbb{C}^{n\times n}$ mit $\|\boldsymbol{H}\|_2 = 1$, für die $\boldsymbol{r} = \eta\boldsymbol{H}\boldsymbol{u}$ gilt. Damit ist $\boldsymbol{A}\boldsymbol{u} - \mu\boldsymbol{u} = \eta\boldsymbol{H}\boldsymbol{u}$, also $\boldsymbol{\delta A} = -\eta\boldsymbol{H}$ und $\|\boldsymbol{\delta A}\|_2 = \|\eta\boldsymbol{H}\|_2 = \eta \leqq \varepsilon$. ❀

Wir akzeptieren daher ein Paar $\{\mu, \boldsymbol{u}\}$ als Eigenpaar einer benachbarten Matrix $\boldsymbol{A} + \boldsymbol{\delta A}$, falls sein Residuum bezüglich des Eigenwertproblems der Matrix $\boldsymbol{A}$ klein genug ist.
Die Algorithmen, die wir kennenlernen werden, liefern teilweise nur Näherungen für Eigenvektoren oder nur Näherungen für Eigenwerte. In diesen Fällen stellt sich die Aufgabe, zu einer gegebenen Näherung für einen Eigenvektor bzw. einen Eigenwert entsprechende gute Näherungen für einen zugehörigen Eigenwert bzw. Eigenvektor zu finden. Betrachten wir das erste Problem.

9.26. Satz: *Es seien die hermitesche Matrix $\boldsymbol{A} \in \mathbb{C}^{n\times n}$ und ein Vektor $\boldsymbol{u} \in \mathbb{C}^n$ mit $\|\boldsymbol{u}\|_2 = 1$ gegeben. Dann gilt:*

1. *Das lineare Ausgleichsproblem* $\min\{\, \|\boldsymbol{A}\boldsymbol{u} - \mu\boldsymbol{u}\|_2 \mid \mu \in \mathbb{R}\,\}$ *hat die eindeutige Lösung*

$$\mu^* = \boldsymbol{u}^H\boldsymbol{A}\boldsymbol{u} = RQ(\boldsymbol{u}).$$

   Das zugehörige Residuum $\boldsymbol{r} = \boldsymbol{A}\boldsymbol{u} - \mu^*\boldsymbol{u}$ ist orthogonal zu $\boldsymbol{u}$, es gilt $\boldsymbol{u}^H\boldsymbol{r} = 0$.

2. *Das Paar $\{\mu^*, \boldsymbol{u}\}$ ist Eigenpaar einer Matrix $\boldsymbol{A} + \boldsymbol{\delta A}$ mit*

$$\boldsymbol{\delta A} \in \mathbb{C}^{n\times n}, \quad \|\boldsymbol{\delta A}\|_2 \leqq \|\boldsymbol{r}\|_2.$$

3. Es gibt ein Eigenpaar $\{\lambda_j, \boldsymbol{u}_j\}$ *von* $\boldsymbol{A}$ *mit*

$$|\mu^* - \lambda_j| = \min\{\, |\mu^* - \lambda_i| \mid i = 1, \ldots, n \,\} \leqq \|\boldsymbol{r}\|_2.$$

Falls $\gamma_j = \min\{\, |\mu^* - \lambda_i| \mid i = 1, \ldots, n, i \neq j \,\} > \|\boldsymbol{r}\|_2$ *gilt*

$$\sin(\ \sphericalangle(\boldsymbol{u}, \boldsymbol{u}_j)) \leqq \|\boldsymbol{r}\|_2 / \gamma_j.$$

Beweis: 1. Die Normalengleichungen zu dem gegebenen Ausgleichsproblem lauten $\mu \boldsymbol{u}^H \boldsymbol{u} = \boldsymbol{u}^H \boldsymbol{A}\boldsymbol{u}$. Daraus folgt wegen $\boldsymbol{u}^H \boldsymbol{u} = 1$ sofort die erste Aussage. Die Orthogonalität zwischen Residuum und dem Vektor $\boldsymbol{u}$ erkennt man durch einfaches Nachrechnen.
2. Diese Aussage erhält man mit dem eben bewiesenen und Satz 9.25.
3. Die erste Aussage ergibt sich aus (ii) mit Hilfe von Satz 9.19. Für die Abschätzung des Winkels zwischen $\boldsymbol{u}$ und $\boldsymbol{u}_j$ wende man Satz 9.24 an. ❀

Die optimale Eigenwertnäherung zu einer Eigenvektornäherung $\boldsymbol{v}$besteht nach diesem Satz im schon erwähnten RAYLEIGH-Quotienten $RQ(\boldsymbol{v}) = \boldsymbol{v}^H \boldsymbol{A}\boldsymbol{v} / \boldsymbol{v}^H \boldsymbol{v}$. Zum Problem der Wahl einer Eigenvektornäherung bei gegebener Eigenwertnäherung erhalten wir den folgenden Satz.

9.27. Satz: *Die hermitesche Matrix* $\boldsymbol{A} \in \mathbb{C}^{n \times n}$ *besitze die (reellen) Eigenwerte*

$$\lambda_1, \ldots, \lambda_n.$$

Durch $\{\boldsymbol{u}_1, \ldots, \boldsymbol{u}_n\}$ *sei ein zugehöriges Orthonormalsystem von Eigenvektoren gegeben. Für festes* $\mu \in \mathbb{R}$ *hat die Aufgabe*

$$\min\{\, \|\boldsymbol{A}\boldsymbol{u} - \mu\boldsymbol{u}\|_2 \mid \boldsymbol{u} \in \mathbb{C}^n, \|\boldsymbol{u}\|_2 = 1 \,\}$$

die Lösung $\boldsymbol{u} = \boldsymbol{u}_j$*, wobei* j *durch*

$$\min\{\, |\lambda_i - \mu| \mid i = 1, \ldots, n \,\} = |\lambda_j - \mu|$$

definiert ist. Für das minimale Residuum gilt

$$\|\boldsymbol{A}\boldsymbol{u}_j - \mu\boldsymbol{u}_j\|_2 = |\lambda_j - \mu|.$$

Beweis: Mit der unitären Matrix $\boldsymbol{U} = (\boldsymbol{u}_1, \ldots, \boldsymbol{u}_n)$ und $\boldsymbol{u} = \boldsymbol{U}\boldsymbol{z}$ gilt

$$\|\boldsymbol{A}\boldsymbol{u} - \mu\boldsymbol{u}\|_2 = \|\boldsymbol{U}^H(\boldsymbol{A}\boldsymbol{u} - \mu\boldsymbol{u}\|_2 = \|\Lambda\boldsymbol{z} - \mu\boldsymbol{z}\|_2 = \sum_{i=1}^{n} (\lambda_i - \mu) z_i^2, \quad \sum_{i=1}^{n} z_i^2 = 1.$$

Damit erhält man für das äquivalente Minimierungsproblem

$$\min\{\, \|\Lambda\boldsymbol{z} - \mu\boldsymbol{z}\|_2 \mid \boldsymbol{z} \in \mathbb{C}^n, \|\boldsymbol{z}\|_2 = 1 \,\}$$

sofort die Lösung $\boldsymbol{z} = \boldsymbol{e}_j$ und damit $\boldsymbol{u} = \boldsymbol{u}_j$. ❀

Die optimale Eigenvektornäherung zu einer vorgegebenen Eigenwertnäherung besteht also in einem Eigenvektor zum nächstgelegenen exakten Eigenwert. Ein effektives Verfahren zum näherungsweisen Lösen dieser Aufgabe werden wir in Form der Inversen Iteration nach WIELANDT kennenlernen.

9.3. Das Jacobi-Verfahren

In diesem und den nächsten Abschnitten werden wir uns mit verschiedenen Verfahren zum Lösen von Eigenwertproblemen befassen. Dabei werden wir alle Verfahren im Reellen erläutern. Die Übertragung auf den komplexen Fall dürfte nicht schwierig sein. Prinzipiell könnte man zumindest im reellen symmetrischen Falle daran denken, die Eigenwerte als Nullstellen des charakteristischen Polynoms zu bestimmen. Dies führt im allgemeinen zu großen Fehlern. Denn während das symmetrische Eigenwertproblem gut konditioniert ist, reagieren die Nullstellen eines Polynoms i. a. äußerst empfindlich auf Änderungen in den Koeffizienten. Betrachten wir als Beispiel eine symmetrische Matrix mit den Eigenwerten $\lambda_1 = 1,\ \lambda_2 = 2, \ldots, \lambda_{20} = 20$. Das charakteristische Polynom dieser Matrix lautet

$$\varphi(\lambda) = (\lambda - 1)(\lambda - 2)\cdots(\lambda - 20).$$

Ändern wir ein Matrixelement um den Wert $\varepsilon = 2^{-23} \approx 10^{-7}$ so ändern sich die Eigenwerte dieser Matrix nach Satz 9.23 höchstens um den Wert ε. Stört man den Koeffizienten von λ^{19} im charakteristischen Polynom um ε, so erhält man ein Polynom mit folgenden Nullstellen:

1.000	000	000		10.095	266	145 ± 0.643	500	$904i$
2.000	000	000		11.793	633	881 ± 1.652	329	$728i$
3.000	000	000		13.992	358	137 ± 2.518	830	$070i$
4.000	000	000		16.730	737	466 ± 2.812	624	$894i$
4.999	999	928		19.502	439	400 ± 1.940	330	$347i$
6.000	006	944						
6.999	697	234						
8.007	267	603						
8.917	250	249						
20.846	908	101						

Die Nullstellen des gestörten Polynoms weichen erheblich von den exakten Nullstellen ab. Es treten sogar fünf komplexe Nullstellenpaare auf. Dies Beispiel zeigt schon, dass der Weg der Eigenwertberechnung über das charakteristische Polynom nicht gangbar ist.
Es gibt mit dem QR-Algorithmus ein leistungsfähiges Werkzeug zum Lösen des

vollständigen Eigenwertproblems; trotzdem haben andere Verfahren, die wir angeben wollen, ihre Berechtigung. Diese liegt darin, dass es die verschiedensten Aufgabenstellungen gibt. So z.B. ist in manchen Anwendungen nur der größte oder der kleinste Eigenwert einer Matrix gesucht. In anderen Anwendungen sucht man alle Eigenwerte oder alle Eigenwerte aus einem vorgegebenen Intervall und eventuell zugehörige Eigenvektoren.
Wie wir im letzten Abschnitt gesehen haben ist das Eigenwertproblem für hermitesche Matrizen besonders gut konditioniert. Da wir im folgenden nur reelle Matrizen betrachten wollen, spielen gerade die symmetrischen Matrizen eine besondere Rolle. Für sie ist außerdem garantiert, dass sie nur reelle Eigenwerte besitzen. Die zugehörigen Algorithmen laufen tatsächlich im Reellen ab. Zum Lösen des Eigenwertproblems für reelle Matrizen mit komplexen Eigenwerten (und damit auch komplexen Eigenvektoren) benötigt man natürlich spezielle reelle Algorithmen oder Algorithmen, die prinzipiell im Komplexen arbeiten. Soweit wir explizit nichts anderes erwähnen, beziehen sich alle folgenden Algorithmen auf den reellen symmetrischen Fall.
Das erste Verfahren zum Lösen des symmetrischen Eigenwertproblems, das wir behandeln werden, ist zugleich das älteste. Es handelt sich um das JACOBI-Verfahren, das aus dem Jahre 1846 stammt. Die Idee dieses Verfahrens besteht darin, die symmetrische (n,n)-Matrix $\boldsymbol{A}$ durch eine Folge orthogonaler Ähnlichkeitstransformationen näherungsweise auf Diagonalgestalt zu transformieren. Dazu werden wir sogenannte GIVENS-Rotationen verwenden. Diese sind von folgender Gestalt:

$$\boldsymbol{G}_{pq}(c,s) = \begin{pmatrix} 1 & & & & & & 0 \\ & \ddots & & & & & \\ & & c & & s & & \\ & & & \ddots & & & \\ & & -s & & c & & \\ & & & & & \ddots & \\ 0 & & & & & & 1 \end{pmatrix} \begin{matrix} \\ \\ \text{p-te Zeile} \\ \\ \text{q-te Zeile} \\ \\ \\ \end{matrix}, \quad c^2+s^2=1.$$

Sie unterscheiden sich von der Einheitsmatrix nur in der p-ten und q-ten Zeile bzw. Spalte. Wir erhalten folgenden Algorithmus:

$$\boldsymbol{A}^{(0)} = \boldsymbol{A}, \qquad \boldsymbol{A}^{(k+1)} = \boldsymbol{G}^{(k)^T}\boldsymbol{A}^{(k)}\boldsymbol{G}^{(k)}, \quad k=0,1,\ldots$$

Dabei sind die Transformationsmatrizen $\boldsymbol{G}^{(k)}$ in jedem Schritt in der Form

$$\boldsymbol{G}^{(k)} = \boldsymbol{G}_{pq}(c,s)$$

mit geeigneten Parametern p, q, c, s zu wählen. Die Matrizen $\boldsymbol{A}^{(k)}, k = 0, 1, \ldots,$ sind alle orthogonal ähnlich. Es gilt

$$\boldsymbol{A}^{(k)} = {\boldsymbol{V}^{(k)}}^T \boldsymbol{A} \boldsymbol{V}^{(k)}, \quad \boldsymbol{V}^{(k)} = \begin{cases} \boldsymbol{I} & \text{für} \quad k = 0 \\ \boldsymbol{G}^{(1)} \boldsymbol{G}^{(2)} \cdots \boldsymbol{G}^{(k-1)} & \text{für} \quad k \geqq 1 \end{cases}.$$

Das Ziel der Transformationen ist zumindest näherungsweise Diagonalgestalt der Matrix. Als Maß für die Abweichung der Iterationsmatrizen von der Diagonalgestalt bietet sich die Summe der Quadrate der Nichtdiagonalelemente an. Wegen der Symmetrie von $\boldsymbol{A}$ genügt es dabei, nur die Quadrate der Elemente oberhalb der Diagonalen aufzusummieren. Wir definieren darum:

$$\omega_k = \omega(\boldsymbol{A}^{(k)}) = \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} \left(a_{ij}^{(k)}\right)^2.$$

Im k-ten Schritt sind demnach die Parameter p, q, c, s so zu wählen, dass ω_{k+1} möglichst klein wird.
Es sei weiterhin

$$\boldsymbol{M}^{(k)} = \text{diag}(a_{11}^{(k)}, a_{22}^{(k)}, \ldots, a_{nn}^{(k)}), \quad \boldsymbol{R}^{(k)} = \boldsymbol{A}^{(k)} - \boldsymbol{M}^{(k)}.$$

Dann ergibt sich

$$\boldsymbol{M}^{(k)} = \boldsymbol{A}^{(k)} - \boldsymbol{R}^{(k)} = {\boldsymbol{V}^{(k)}}^T \left(\boldsymbol{A} + \boldsymbol{\delta A}^{(k)}\right) \boldsymbol{V}^{(k)}$$

mit der symmetrischen Störung

$$\boldsymbol{\delta A}^{(k)} = -\boldsymbol{V}^{(k)} \boldsymbol{R}^{(k)} {\boldsymbol{V}^{(k)}}^T, \quad \text{für die} \quad \|\boldsymbol{\delta A}^{(k)}\|_F = \|\boldsymbol{R}^{(k)}\|_F = \sqrt{2\omega_k} \quad \text{gilt.}$$

Die letzten Gleichungen besagen, dass die Diagonalelemente der Matrix $\boldsymbol{A}^{(k)}$ die exakten Eigenwerte der gestörten Matrix $\boldsymbol{A} + \boldsymbol{\delta A}^{(k)}$ mit den Spalten von $\boldsymbol{V}^{(k)}$ als zugehörigen Eigenvektoren sind. Nach Satz 9.23 lassen sich dann die Eigenwerte λ_i von $\boldsymbol{A}$ so ordnen, dass

$$|a_{ii}^{(k)} - \lambda_{l(i)}| \leqq \|\boldsymbol{\delta A}^{(k)}\|_2 \leqq \|\boldsymbol{\delta A}^{(k)}\|_F = \sqrt{2\omega_k} \quad \text{gilt.} \tag{9.4}$$

Für jedes k ist dabei $\{l(1), \ldots, l(n)\}$ eine Permutation der Zahlen $\{1, \ldots, n\}$ derart, dass die Eigenwerte von $\boldsymbol{A}$ in der gleichen Weise wie die Diagonalelemente von $\boldsymbol{A}^{(k)}$ geordnet sind. Die Diagonalelemente von $\boldsymbol{A}^{(k)}$ approximieren damit die Eigenwerte von $\boldsymbol{A}$ mit einem absoluten Fehler, der durch $\sqrt{2\omega_k}$ beschränkt ist.

Kommen wir nun zur Wahl der Parameter p, q, c, s. Ziel ist es in jedem Schritt, das Maß für die Nichtdiagonalität der Matrix so stark wie möglich zu verringern. Wir haben im k-ten Schritt die Matrix $\boldsymbol{G}^{(k)} = \boldsymbol{G}_{pq}(c, s)$ so zu wählen, dass

$$\boldsymbol{\Phi}_k(p, q, c, s) = \omega_k - \omega_{k+1}$$

maximiert wird. Betrachten wir einen Transformationsschritt

$$\bar{\boldsymbol{A}} = \boldsymbol{G}_{pq}(c, s)^T \boldsymbol{A} \boldsymbol{G}_{pq}(c, s).$$

Bei dieser Transformation werden nur die Elemente der p-ten und q-ten Zeile und Spalte von $\boldsymbol{A}$ geändert. Es gilt

$$\begin{aligned}
\bar{a}_{pp} &= c^2 a_{pp} - 2csa_{pq} + s^2 a_{qq}, \\
\bar{a}_{qq} &= s^2 a_{pp} + 2csa_{pq} + c^2 a_{qq}, \\
\bar{a}_{pq} &= \bar{a}_{qp} = cs(a_{pp} - a_{qq}) + (c^2 - s^2) a_{pq}, \\
\bar{a}_{ip} &= \bar{a}_{pi} = ca_{ip} - sa_{iq}, \quad i \neq p, q, \\
\bar{a}_{iq} &= \bar{a}_{qi} = sa_{ip} + ca_{iq}, \quad i \neq p, q.
\end{aligned}$$

Wegen $c^2 + s^2 = 1$ folgt

$$\bar{a}_{ip}^2 + \bar{a}_{iq}^2 = a_{ip}^2 + a_{iq}^2, \quad i \neq p, q.$$

Die Quadratsumme der Nichtdiagonalelemente wird nur durch die Elemente in den Positionen (p, q) und (q, p) verändert. Damit gilt

$$\boldsymbol{\Phi}(p, q, c, s) = \omega - \bar{\omega} = \omega(\boldsymbol{A}) - \omega(\bar{\boldsymbol{A}}) = a_{pq}^2 - \bar{a}_{pq}^2.$$

Hält man p und q einmal fest, so wird $\boldsymbol{\Phi}(p, q, c, s)$ offensichtlich maximal, falls c und s so gewählt werden, dass $\bar{a}_{pq} = 0$ gilt. Dann ist

$$\omega(\bar{\boldsymbol{A}}) = \omega(\boldsymbol{A}) - a_{pq}^2.$$

Als Bestimmungsgleichungen für c und s erhält man

$$\begin{aligned}
cs(a_{pp} - a_{qq}) + (c^2 - s^2) a_{pq} &= 0, \\
c^2 + s^2 &= 1.
\end{aligned}$$

Mit $t = s/c$ erhalten wir aus der ersten Gleichung

$$t^2 + 2\delta t - 1 = 0, \quad \delta = \frac{a_{qq} - a_{pp}}{2a_{pq}}.$$

Die betragskleinste Lösung dieser quadratischen Gleichung ist nach der Vorschrift

$$t = \begin{cases} 1/\left(|\delta| + \sqrt{1+\delta^2}\right) & \text{für} \quad \delta \geqq 0 \\ -1/\left(|\delta| + \sqrt{1+\delta^2}\right) & \text{für} \quad \delta < 0 \end{cases} \tag{9.5}$$

numerisch stabil berechenbar. Damit erhält man die gesuchten Parameter zu

$$c = 1/\sqrt{1+t^2}, \quad s = tc. \tag{9.6}$$

Mit der Hilfsgröße

$$\tau = s/(1+c) \tag{9.7}$$

ergeben sich jetzt folgende Transformationsformeln:

$$\begin{aligned} \overline{a}_{pp} &= a_{pp} - t a_{pq}, && (9.8)\\ \overline{a}_{qq} &= a_{qq} + t a_{pq}, && (9.9)\\ \overline{a}_{pq} &= \overline{a}_{qp} = 0, && (9.10)\\ \overline{a}_{ip} &= \overline{a}_{pi} = a_{ip} - s(a_{iq} + \tau a_{ip}), \quad i \neq p,q && (9.11)\\ \overline{a}_{iq} &= \overline{a}_{qi} = a_{iq} + \tau(a_{ip} + \overline{a}_{ip}), \quad i \neq p,q. && (9.12) \end{aligned}$$

lässt man die Wahl von p und q vorerst noch offen, so erhält man folgenden Basisalgorithmus:

9.28. JACOBI**-Basisverfahren:**

S0 *(Initialisierung) Setze* $\boldsymbol{A}^{(0)} = \boldsymbol{A}$, $k = 0$ *und berechne* $\omega_0 = \omega(\boldsymbol{A})$. *Wähle ein* $\varepsilon > 0$.

S1 *(Abbruchtest) Falls* $2\omega_k \leqq \varepsilon^2$ *STOPP.*

S2 *(Pivotwahl) Wähle Indizes* $p = p(k)$ *und* $q = q(k)$ *mit* $a_{pq}^{(k)} \neq 0$.

S3 *(Iterationsschritt)*

S4 *Berechne* $t = t_k, c = c_k, s = s_k$ *und* $\tau = \tau_k$ *gemäß (9.5)-(9.7).*

S5 *Berechne* $\boldsymbol{A}^{(k+1)}$ *aus* $\boldsymbol{A}^{(k)}$ *gemäß (9.8)-(9.12).*

S6 *Berechne* $\omega_{k+1} = \omega_k - \left(a_{pq}^{(k)}\right)^2$.

Setze $k = k+1$ *und gehe zu Schritt* **S1**.

Aufwand pro Schritt: $\sim 4n$ *Add./Sub. +* $\sim 3n$ *Mult./Div. + 2 Quadratwurzeln.*

Bemerkungen: (i) Das Verfahren darf auf dem Platz der Matrix, also auf dem oberen Dreieck, durchgeführt werden.
(ii) Ist der Abbruchtest in **S1** nach k Schritten erfüllt, so gilt nach (9.4)

$$|a_{ii}^{(k)} - \lambda_{l(i)}| \leqq \varepsilon.$$

Die Diagonalelemente von $\boldsymbol{A}^{(k)}$ approximieren die Eigenwerte von $\boldsymbol{A}$ mit der vorgegebenen Genauigkeit ε. Man sollte $\varepsilon \geqq \text{eps}\|\boldsymbol{A}\|_F$ wählen, da eine höhere Genauigkeit nicht erreichbar ist.
(iii) Falls ω_k klein wird (etwa $k\text{eps}\,\omega_1$), so sollte man statt der Aufdatierung in Schritt **S3** ω_k direkt aus $\boldsymbol{A}^{(k)}$ berechnen.
(iv) Im KonvergenzFalle gilt $\omega_k \to 0$. Dann werden die Nichtdiagonalelemente von $\boldsymbol{A}_k$ klein. Das δ aus der zu lösenden quadratischen Gleichung kann dann groß werden. Wegen

$$|\tau| \leqq \frac{|s|}{1+\sqrt{2}/2} \leqq |t| \leqq \frac{1}{1+|\delta|}$$

sind τ_k, s_k und t_k betragsmäßig klein, so dass die entsprechenden Korrekturformeln (9.8)-(9.12) gutartig sind.
Für eine optimale Wahl von p und q erkennt man sofort, dass a_{pq} das betragsgrößte Nichtdiagonalelement von $\boldsymbol{A}$ sein müßte. Mit dieser Strategie erhält man das klassische JACOBI-Verfahren. Zur Bestimmung dieses betragsgrößten Elementes ist in jedem Schritt ein großer Suchaufwand notwendig. Als einfachere Variante bietet es sich an, alle Nichtdiagonalelemente der Matrix zyklisch zu durchlaufen, so z.B. zeilenweise gemäß

$$(p,q) = (1,2),\ldots,(1,n),(2,3),\ldots,(2,n),\ldots,(n-2,n-1),(n-2,n),(n-1,n).$$

Ein derartiges Verfahren bezeichnet man als zyklisches JACOBI-Verfahren. Beim zyklischen JACOBI-Verfahren kommt es vor, dass zu einem Element a_{pq} eine GIVENS-Rotation durchgeführt wird, obwohl es klein ist. Die Abnahme von ω_k, die dadurch erreicht wird, ist also unwesentlich. Darum modifizieren wir das zyklische JACOBI-Verfahren so, dass wir nur dann eine GIVENS-Rotation $\boldsymbol{G}_{pq}$ durchführen, falls das Element a_{pq} größer als ein vorgegebener Schwellwert ist. Dadurch werden uneffiziente Transformationen vermieden. Dieses Verfahren heißt Schwellwert-JACOBI-Verfahren. Als Schwellwert sollte man das quadratische Mittel der Nichtdiagonalelemente $\sqrt{\omega_k/N}$, $N = n(n-1)/2$ verwenden. Man beachte, dass jede GIVENS-Rotation neue Nichtnullelemente in der Matrix erzeugt, d.h. dass auch nach Durchlaufen aller Nichtdiagonalelemente die Matrix nicht auf Diagonalgestalt transformiert ist. Das JACOBI-Verfahren stellt im allgemeinen einen unendlichen Algorithmus zur Eigenwertberechnung dar. Die Konvergenz des Algorithmus wollen wir nun untersuchen.

9.29. Satz: *Das* JACOBI*-Verfahren werde für eine symmetrische Matrix* $\boldsymbol{A} \in \mathbb{C}^{n\times n}$ *mit den Eigenwerten* $\lambda_1, \ldots, \lambda_n$ *in exakter Arithmetik durchgeführt. Es existiere ein* $\kappa < 1$*, so dass die Pivotwahl in Schritt* **S2** *so erfolgt, dass jeweils*

$$\omega_{k+1} \leqq \kappa\omega_k$$

erfüllt ist. Dann gilt für die Folge $\{\boldsymbol{A}^{(k)}\}$

$$\lim_{k\to\infty} \boldsymbol{A}^{(k)} = \boldsymbol{M} = \operatorname{diag}(\lambda_{m(1)}, \lambda_{m(2)}, \ldots, \lambda_{m(n)}) \tag{9.13}$$

mit einer geeigneten Permutation $\{m(1), \ldots, m(n)\}$ *der Zahlen* $\{1, \ldots, n\}$*. Weiterhin gilt:*

(i) Für $i \neq j$ *und* $k \geqq 0$ *ist*

$$|a_{ij}^{(k)}| \leqq \|\boldsymbol{R}^{(k)}\|_F \leqq \sqrt{2\omega_k} \leqq \left(\sqrt{\kappa}\right)^k \sqrt{2\omega_0}.$$

(ii) Im Falle $i = j$ *existiert ein* $k_0 > 0$*, so dass für alle* $k \geqq k_0$

$$|a_{ii}^{(k)} - \lambda_{m(i)}| \leqq \|\boldsymbol{M}^{(k)} - \boldsymbol{M}\| \leqq \sqrt{2\omega_k} \leqq \left(\sqrt{\kappa}\right)^k \sqrt{2\omega_0}.$$

Beweis: (i) Diese Aussage folgt sofort aus $\omega_{k+1} \leqq \kappa\omega_k$.
(ii) Für $i \neq p, q$ gilt

$$|a_{ii}^{(k+1)} - a_{ii}^{(k)}| = 0.$$

Für $i = p$ oder $i = q$ folgt aus den Transformationsformeln (9.8) und (9.12)

$$\begin{aligned}
|a_{pp}^{(k+1)} - a_{pp}^{(k)}| &= |a_{pp}^{(k)} - t_k a_{pq}^{(k)} - a_{pp}^{(k)}| \\
&= |t_k||a_{pq}^{(k)}| \\
|a_{qq}^{(k+1)} - a_{qq}^{(k)}| &= |a_{qq}^{(k)} + t_k a_{pq}^{(k)} - a_{qq}^{(k)}| \\
&= |t_k||a_{pq}^{(k)}|.
\end{aligned}$$

Damit erhält man

$$|a_{ii}^{(k+1)} - a_{ii}^{(k)}| \leqq |t_k||a_{pq}^{(k)}| \leqq 1 \cdot \left(\sqrt{\kappa}\right)^k \sqrt{2\omega_0},$$

und weiter für $r = 0, 1, \ldots$

$$\begin{aligned}
|a_{ii}^{(k+r+1)} - a_{ii}^{(k)}| &\leqq |a_{ii}^{(k+r+1)} - a_{ii}^{(k+r)}| + \cdots + |a_{ii}^{(k+1)} - a_{ii}^{(k)}| \\
&\leqq \left[\left(\sqrt{\kappa}\right)^{k+r} + \cdots + \left(\sqrt{\kappa}\right)^k\right] \sqrt{2\omega_0} \\
&= \left(\sqrt{\kappa}\right)^k \left[\left(\sqrt{\kappa}\right)^r + \cdots + 1\right] \sqrt{2\omega_0} \\
&\leqq \frac{\left(\sqrt{\kappa}\right)^k}{1-\sqrt{\kappa}} \sqrt{2\omega_0}.
\end{aligned}$$

Die Folge $\{a_{ii}^{(k)}\}$ ist damit CAUCHY-Folge und konvergiert gegen einen Grenzwert μ_i. Damit gilt (9.13) mit $\boldsymbol{M} = \text{diag}(\mu_1, \mu_2, \ldots, \mu_n)$. Da alle $\boldsymbol{A}^{(k)}$ orthogonal ähnlich zu $\boldsymbol{A}$ sind, müssen diese μ_i Eigenwerte von $\boldsymbol{A}$ sein. Es existiert also eine Permutation $\{m(1), \ldots, m(n)\}$ der Zahlen $\{1, \ldots, n\}$, so dass $\mu_i = \lambda_{m(i)}$, $i = 1, \ldots, n$. Die Abschätzung in (ii) folgt dann mit Satz 9.23. ❀

Bemerkungen: (i) Für das klassische JACOBI-Verfahren gilt

$$\omega_k \leqq \frac{n(n-1)}{2} \left(a_{pq}^{(k)}\right)^2 = N \left(a_{pq}^{(k)}\right)^2.$$

Daraus folgt

$$\begin{aligned} \omega_{k+1} &= \omega_k - \left(a_{pq}^{(k)}\right)^2 \\ &= \left(1 - \frac{\left(a_{pq}^{(k)}\right)^2}{\omega_k}\right) \omega_k \\ &\leqq \left(1 - \frac{1}{N}\right) \omega_k. \end{aligned}$$

Wir erhalten daher $\omega_{k+1} \leqq \kappa \omega_k$ mit $\kappa = 1 - \frac{1}{N} < 1$.
(ii) Für das Schwellwert-JACOBI-Verfahren sollte

$$|a_{pq}^{(k)}| \geqq \sqrt{\omega_k / N}$$

gelten. Damit folgt wieder

$$\begin{aligned} \omega_{k+1} &= \omega_k - \left(a_{pq}^{(k)}\right)^2 \\ &= \left(1 - \frac{\left(a_{pq}^{(k)}\right)^2}{\omega_k}\right) \omega_k \\ &\leqq \left(1 - \frac{1}{N}\right) \omega_k, \end{aligned}$$

daher $\omega_{k+1} \leqq \kappa \omega_k$ mit $\kappa = 1 - \frac{1}{N} < 1$.
(iii) Nach Satz 9.29 konvergieren die Elemente von $\boldsymbol{A}^{(k)}$ mindestens linear mit dem Faktor κ gegen Null bzw. gegen die Eigenwerte von $\boldsymbol{A}$. Da κ nahe bei 1 liegt, folgt

daraus langsame Konvergenz. Um eine vorgegebene Genauigkeit $\varepsilon = K\text{eps}\|\boldsymbol{A}\|_F$ zu erreichen würde dies eine Abschätzung

$$r = \frac{k}{N} \leqq 2\ln\left(\frac{1}{K\text{eps}}\right)$$

für die Anzahl der benötigten Zyklen ergeben. Praktische Erfahrungen zeigen aber, dass fast immer

$$r = \frac{k}{N} \leqq 4\ldots10$$

gilt. Grund für dieses Verhalten ist die asymptotisch quadratische Konvergenz des JACOBI-Verfahrens.
Als letztes wollen wir noch den Rundungsfehlereinfluss beim JACOBI-Verfahren abschätzen.

9.30. Rundungsfehleranalyse: *Das* JACOBI*-Verfahren werde in Computerarithmetik mit einer vorgegebenen Genauigkeit $\varepsilon = K\text{eps}\|\boldsymbol{A}\|_F$ und einer Pivotwahl, für die $\omega_{k+1} \leqq \kappa\omega_k$ mit $\kappa < 1$ gilt, durchgeführt. Dann bricht das Verfahren nach $k = Nr$ Schritten ab, wobei $r \leqq 4\ldots10$ gilt.*
Die berechneten Diagonalelemente $\mu_i = a_{ii}^{(k+1)}$ sind exakte Eigenwerte einer Matrix $\boldsymbol{A} + \boldsymbol{\delta A}$, wobei für die symmetrische Störung $\boldsymbol{\delta A}$

$$\|\boldsymbol{\delta A}\| \leqq \text{eps}(F+K)\|\boldsymbol{A}\|_F, \quad F \leqq 18n^{3/2}r$$

gilt. (Praktisch gilt fast immer $F \leqq k/n \approx nr/2$, oft sogar $F \leqq 10$.)
Für die berechnete Matrix der Eigenvektoren $\tilde{\boldsymbol{V}}$ gilt

$$\|\tilde{\boldsymbol{V}} - \boldsymbol{V}\| \leqq \text{eps}F_1, \quad F_1 \leqq 6n^2$$

wobei $\boldsymbol{V}$ die exakte Matrix der Eigenvektoren von $\boldsymbol{A} + \boldsymbol{\delta A}$ ist.

Der Aufwand des JACOBI-Verfahrens ist relativ hoch. Bricht man nach vier Zyklen ab, so benötigt man etwa $\sim 6n^3$ Additionen und Multiplikationen allein zum Berechnen der Eigenwerte und $\sim 12n^3$ zum Lösen des vollständigen Eigenwertproblems. Das QR-Verfahren, das wir später behandeln werden, löst dieselben Aufgaben mit einem Neuntel bzw. einem Drittel des Aufwands. Darum ist die Anwendung des JACOBI-Verfahrens auf Spezialfälle beschränkt, etwa das Berechnen der Eigenwerte mit geringer Genauigkeit oder das Berechnen der Eigenwerte mit voller Genauigkeit für Matrizen mit kleinen Nichtdiagonalelementen. In diesen Fällen kann man hoffen, mit ein bis zwei Zyklen ans Ziel zu kommen. Auch für Probleme kleiner Dimension ($n \leqq 10$) ist das JACOBI-Verfahren konkurrenzfähig.

9.4. Die Vektor- und Teilraumiteration

In diesem Abschnitt werden wir iterative Verfahren betrachten, die für eine reelle symmetrische Matrix Eigenvektoren approximieren, die zu den betragsgrößten Eigenwerten gehören. Die Algorithmen selbst haben keine besondere Bedeutung, sie dienen als Grundlage für die inverse Iteration oder den QR-Algorithmus, die leistungsfähige Algorithmen zur Behandlung von Eigenwertproblemen darstellen.
Es seien $\lambda_1, \lambda_2, \ldots, \lambda_n$ die gemäß

$$|\lambda_1| \geqq |\lambda_2| \geqq \cdots \geqq |\lambda_n|$$

geordneten Eigenwerte der symmetrischen Matrix $\boldsymbol{A}$. Ferner sei

$$\{\, \boldsymbol{u}_1, \boldsymbol{u}_2, \ldots, \boldsymbol{u}_n \,\}$$

ein zugehöriges Orthonormalsystem von Eigenvektoren. Dann heißen die Eigenwerte

$$\{\, \lambda_1, \lambda_2, \ldots, \lambda_p \,\}$$

dominant, falls $|\lambda_p| > |\lambda_{p+1}|$ gilt.
Der den dominanten Eigenwerten $\{\lambda_1, \lambda_2, \ldots, \lambda_p\}$ zugeordnete Eigenraum

$$\boldsymbol{\mathcal{S}}_p = \boldsymbol{\mathcal{S}}_p(\boldsymbol{A}) = \text{span}(\boldsymbol{u}_1, \boldsymbol{u}_2, \ldots, \boldsymbol{u}_p)$$

heißt p-ter dominanter Teilraum von $\boldsymbol{A}$. Die Eigenvektoren $\boldsymbol{u}_1, \boldsymbol{u}_2, \ldots, \boldsymbol{u}_p$ bilden eine orthonormale Basis des $\boldsymbol{\mathcal{S}}_p$ und es gilt $\dim(\boldsymbol{\mathcal{S}}_p) = p$. Weiterhin ist wegen $\boldsymbol{A}\boldsymbol{u}_j = \lambda_j \boldsymbol{u}_j \in \boldsymbol{\mathcal{S}}_p$ für $j = 1, 2, \ldots, p$

$$\boldsymbol{A}\boldsymbol{\mathcal{S}}_p = \{\boldsymbol{y} = \boldsymbol{A}\boldsymbol{x} : \boldsymbol{x} \in \boldsymbol{\mathcal{S}}_p\} \subseteqq \boldsymbol{\mathcal{S}}_p. \tag{9.14}$$

Der Eigenraum $\boldsymbol{\mathcal{S}}_p$ ist somit invariant unter der durch $\boldsymbol{A}$ vermittelten linearen Abbildung. Ein Teilraum $\boldsymbol{\mathcal{S}}_p$ des $\mathbb{R}^n$, für den (9.14) gilt, heißt auch $\boldsymbol{A}$-invarianter Teilraum des $\mathbb{R}^n$. Wir werden nun versuchen, $\boldsymbol{\mathcal{S}}_p$ zu approximieren. Dazu gehen wir von der Eigenwertzerlegung von $\boldsymbol{A}$ aus. Es sei

$$\begin{aligned} \boldsymbol{A} &= \boldsymbol{U}\boldsymbol{\Lambda}\boldsymbol{U}^T \\ &= \left[\lambda_1 \boldsymbol{u}_1 \boldsymbol{u}_1^T + \cdots + \lambda_p \boldsymbol{u}_p \boldsymbol{u}_p^T\right] + \left[\lambda_{p+1} \boldsymbol{u}_{p+1} \boldsymbol{u}_{p+1}^T + \cdots + \lambda_n \boldsymbol{u}_n \boldsymbol{u}_n^T\right]. \end{aligned}$$

Dann gilt

$$\begin{aligned}\boldsymbol{A}^k &= \boldsymbol{U}\Lambda^k\boldsymbol{U}^T \\ &= \left[\lambda_1^k\boldsymbol{u}_1\boldsymbol{u}_1^T+\cdots+\lambda_p^k\boldsymbol{u}_p\boldsymbol{u}_p^T\right]+\left[\lambda_{p+1}^k\boldsymbol{u}_{p+1}\boldsymbol{u}_{p+1}^T+\cdots+\lambda_n^k\boldsymbol{u}_n\boldsymbol{u}_n^T\right] \\ &= \lambda_p^k\left\{\left[\left(\frac{\lambda_1}{\lambda_p}\right)^k\boldsymbol{u}_1\boldsymbol{u}_1^T+\cdots+\boldsymbol{u}_p\boldsymbol{u}_p^T\right]\right. \\ &\qquad \left.+\left[\left(\frac{\lambda_{p+1}}{\lambda_p}\right)^k\boldsymbol{u}_{p+1}\boldsymbol{u}_{p+1}^T+\cdots+\left(\frac{\lambda_n}{\lambda_p}\right)^k\boldsymbol{u}_n\boldsymbol{u}_n^T\right]\right\}.\end{aligned}$$

Wegen $|\lambda_p| > |\lambda_{p+1}|$ wird die Dominanz der ersten p Terme mit zunehmenden k-Potenzen immer größer. Für hinreichend großes k wird $\boldsymbol{A}^k$ durch die ersten p Terme repräsentiert. Damit gilt für einen beliebigen Vektor $\boldsymbol{v} \in \mathbb{R}^n$

$$\boldsymbol{A}^k\boldsymbol{v} = \sum_{j=1}^{n}\lambda_j^k\left[\boldsymbol{u}_j^T\boldsymbol{v}\right]\boldsymbol{u}_j \approx \lambda_1^k\left[\boldsymbol{u}_1^T\boldsymbol{v}\right]\boldsymbol{u}_1+\cdots+\lambda_p^k\left[\boldsymbol{u}_p^T\boldsymbol{v}\right]\boldsymbol{u}_p$$

falls nicht alle Skalarprodukte $\boldsymbol{u}_j^T\boldsymbol{v}, j = 1,\ldots,p$, verschwinden. Für großes k liegt somit der Vektor $\boldsymbol{A}^k\boldsymbol{v}$ fast in $\boldsymbol{\mathcal{S}}_p$.
Wenn wir geeignete p linear unabhängige Vektoren $\boldsymbol{v}_1,\ldots,\boldsymbol{v}_p$ wählen, so werden die Vektoren $\boldsymbol{A}^k\boldsymbol{v}_1,\ldots,\boldsymbol{A}^k\boldsymbol{v}_p$ für genügend großes k einen Teilraum des $\mathbb{R}^n$ aufspannen, der $\boldsymbol{\mathcal{S}}_p$ gut approximiert. Das Berechnen von $\boldsymbol{A}^k\boldsymbol{v}$ erfolgt natürlich rekursiv gemäß $\boldsymbol{A}^{k+1}\boldsymbol{v} = \boldsymbol{A}\left(\boldsymbol{A}^k\boldsymbol{v}\right)$. Außerdem sind bei der praktischen Realisierung die Vektoren $\boldsymbol{A}^k\boldsymbol{v}$ zu normieren, da $\|\boldsymbol{A}^k\boldsymbol{v}\|$ im Falle $|\lambda_1| > 1$ unbeschränkt wächst und im Falle $|\lambda_1| < 1$ gegen Null geht.
Wir betrachten nun den einfachsten Fall $p = 1$. Das ist die sogenannte Vektoriteration.

9.31. Basisverfahren der Vektoriteration:

S0 *(Initialisierung) Wähle Vektor* $\boldsymbol{v}^{(0)} \in \mathbb{R}^n$ *mit* $\|\boldsymbol{v}^{(0)}\|_2 = 1$ *und setze* $k = 0$.

S1 *(Iteration) Berechne* $\boldsymbol{w}^{(k+1)} = \boldsymbol{A}\boldsymbol{v}^{(k)}$.

S2 *(Normierung) Setze* $\boldsymbol{v}^{(k+1)} = \boldsymbol{w}^{(k+1)}/\|\boldsymbol{w}^{(k+1)}\|_2$.

S3 *Setze* $k = k+1$ *und gehe zu Schritt* **S1**.

Aufwand pro Schritt:

- *Eine Auswertung* $Matrix \times Vektor$ **(S1)**
- $\sim n$ *Add./Sub. +* $\sim 2n$ *Mult./Div. + 1 Quadratwurzel (***S2***)*

Bemerkungen: (i) Die Matrix $\boldsymbol{A}$ wird bei der Iteration im Unterschied zu anderen Verfahren nicht verändert. Man benötigt eigentlich auch nur eine Prozedur zum

Berechnen von $\boldsymbol{Ax}$. Dies ist besonders bei schwach besetzten Matrizen hoher Dimension von Vorteil.
(ii) Die Vektoriteration wird auch als „VON-MISES-Iteration“ oder als Potenzmethode bezeichnet.
Der folgende Satz liefert uns Aussagen über das Konvergenzverhalten der Vektoriteration.

9.32. Satz: *Es sei λ_1 dominanter Eigenwert der symmetrischen Matrix $\boldsymbol{A}$, $\boldsymbol{u}_1$ bezeichne einen zugehörigen normierten Eigenvektor, der Startvektor $\boldsymbol{v}^{(0)}$ genüge der Bedingung*

$$\sigma = \boldsymbol{u}_1^T \boldsymbol{v}^{(0)} > 0. \tag{9.15}$$

Dann gilt für die Folge $\{\boldsymbol{v}^{(k)}\}$, die durch die Vektoriteration erzeugt wird,

$$0 \leqq \tan\varphi_k \leqq \kappa \tan\varphi_{k-1} \leqq \kappa^k \tan\varphi_0 = \kappa^k \sqrt{1-\sigma^2}/\sigma \tag{9.16}$$

mit $\kappa = |\lambda_1/\lambda_2|$ und $\varphi_k = \sphericalangle(\boldsymbol{u}_1, \vartheta_k \boldsymbol{v}^{(k)}) \in [0, \pi/2)$, $\vartheta_k = (\operatorname{sgn}(\lambda_1))^k \in \{1, -1\}$.

Beweis: Es sei $\boldsymbol{\mathcal{S}}_1 = \operatorname{span}(\boldsymbol{u}_1)$ und $\boldsymbol{\mathcal{Z}}_2 = \operatorname{span}(\boldsymbol{u}_2, \ldots, \boldsymbol{u}_n) = \boldsymbol{\mathcal{S}}_1^{\perp}$. Dann lässt sich der Startvektor $\boldsymbol{v}^{(0)}$ in Komponenten bezüglich $\boldsymbol{\mathcal{S}}_1$ und $\boldsymbol{\mathcal{Z}}_2$ zerlegen:

$$\boldsymbol{v}^{(0)} = \vartheta_0 \boldsymbol{v}^{(0)} = \boldsymbol{u}_1 \cos\varphi_0 + \boldsymbol{z}_0 \sin\varphi_0, \quad \boldsymbol{z}_0 \in \boldsymbol{\mathcal{Z}}_2, \quad \|\boldsymbol{z}_0\| = 1. \tag{9.17}$$

Wegen (9.15) gilt

$$\sigma = \cos\varphi_0 > 0,$$

und darum $0 \leqq \varphi_0 < \pi/2$ und $\tan\varphi_0 = \sqrt{1-\sigma^2}/\sigma \geqq 0$. Multiplizieren wir 9.17 von links mit $\operatorname{sgn}(\lambda_1)\,\boldsymbol{A}$, so erhalten wir

$$\begin{aligned} \lambda_1 \boldsymbol{A} \vartheta_0 \boldsymbol{v}^{(0)} &= \vartheta_1 \boldsymbol{A} \boldsymbol{v}^{(0)} = \vartheta_1 \boldsymbol{w}^{(1)} \\ &= \operatorname{sgn}(\lambda_1)\,(\lambda_1 \boldsymbol{u}_1 \cos\varphi_0 + \boldsymbol{A}\boldsymbol{z}_0 \sin\varphi_0), \quad \boldsymbol{A}\boldsymbol{z}_0 \in \boldsymbol{\mathcal{Z}}_2. \end{aligned}$$

Daraus folgt

$$\|\boldsymbol{w}^{(1)}\|^2 = \lambda_1^2 \cos^2\varphi_0 + \|\boldsymbol{A}\boldsymbol{z}_0\|^2 \sin^2\varphi_0 \geqq \lambda_1^2 \cos^2\varphi_0,$$

daher $\|\boldsymbol{w}^{(1)}\| \geqq |\lambda_1| \cos\varphi_0 > 0$. Der Schritt **S2** im Algorithmus ist durchführbar und es gilt

$$\vartheta_1 \boldsymbol{v}^{(1)} = \vartheta_1 \boldsymbol{w}^{(1)} / \|\boldsymbol{w}^{(1)}\|_2 = \boldsymbol{u}_1 \cos\varphi_1 + \boldsymbol{z}_1 \sin\varphi_1, \quad \boldsymbol{z}_1 \in \boldsymbol{\mathcal{Z}}_2, \quad \|\boldsymbol{z}_1\|_2 = 1,$$

wobei

$$\cos\varphi_1 = \frac{|\lambda_1|\cos\varphi_0}{\|\boldsymbol{w}^{(1)}\|}, \quad \sin\varphi_1 = \frac{\|\boldsymbol{A}\boldsymbol{z}_0\|\sin\varphi_0}{\|\boldsymbol{w}^{(1)}\|}$$

und

$$\boldsymbol{z}_1 = \begin{cases} \dfrac{\operatorname{sgn}(\lambda_1)\boldsymbol{A}\boldsymbol{z}_0}{\|\boldsymbol{A}\boldsymbol{z}_0\|} & \text{für} \quad \boldsymbol{A}\boldsymbol{z}_0 \neq \boldsymbol{o} \\ \\ \boldsymbol{u}_2 & \text{für} \quad \boldsymbol{A}\boldsymbol{z}_0 = \boldsymbol{o}. \end{cases}$$

Daraus folgt $0 \leqq \varphi_1 < \pi/2$ und

$$0 \leqq \tan\varphi_1 = \frac{\|\boldsymbol{A}\boldsymbol{z}_0\|}{|\lambda_1|}\tan\varphi_0 \leqq \frac{|\lambda_2|}{|\lambda_1|}\tan\varphi_0 = \kappa\tan\varphi_0.$$

(9.16) gilt daher für $k = 1$. Der Rest folgt durch vollständige Induktion. ❀

Bemerkungen: (i) Der Vorzeichenfaktor ϑ_k hat nur beweistechnische Bedeutung.
(ii) $|\tan(\sphericalangle(\boldsymbol{u},\boldsymbol{v}))|$ ist als „Entfernung“ zweier Richtungen interpretierbar. Orthogonale Richtungen sind unendlich weit voneinander entfernt, der Abstand der Richtungen paralleler Vektoren ist gleich Null.
(iii) Es gilt

$$\begin{aligned} \|\vartheta_k\boldsymbol{v}^{(k)} - \boldsymbol{u}_1\| &= \sqrt{(\vartheta_k\boldsymbol{v}^{(k)} - \boldsymbol{u}_1)^T(\vartheta_k\boldsymbol{v}^{(k)} - \boldsymbol{u}_1)} \\ &= \sqrt{\|\boldsymbol{v}^{(k)}\|^2 - 2\boldsymbol{u}_1^T\boldsymbol{v}^{(k)} + \|\boldsymbol{u}_1\|^2} \\ &= \sqrt{2(1-\cos\varphi_k)} \\ &= 2\sin(\varphi_k/2) \leqq \tan\varphi_k. \end{aligned}$$

Wegen $\varphi_k \to 0$ folgt somit $\vartheta_k\boldsymbol{v}^{(k)} \to \boldsymbol{u}_1$.
(iv) Normalerweise möchte man neben einer Eigenvektornäherung auch eine entsprechende Approximation für den zugehörigen Eigenwert haben. Die beste Eigenwertnäherung zu einem gegebenen Eigenvektor ist nach Satz 9.26 der RAYLEIGH-Quotient:

$$\varrho_k = RQ(\boldsymbol{v}^{(k)}) = \boldsymbol{v}^{(k)^T}\boldsymbol{A}\boldsymbol{v}^{(k)} = \boldsymbol{v}^{(k)^T}\boldsymbol{w}^{(k+1)}.$$

Stellen wir $\boldsymbol{v}^{(k)}$ in der Form

$$\boldsymbol{v}^{(k)} = \boldsymbol{u}_1\cos\varphi_k + \boldsymbol{z}_k\sin\varphi_k, \quad \boldsymbol{z}_k \in \boldsymbol{\mathcal{Z}}_2, \quad \|\boldsymbol{z}_k\| = 1$$

dar, so ergibt sich

$$\boldsymbol{A}\left(\vartheta_k \boldsymbol{v}^{(k)}\right) = \boldsymbol{u}_1 \lambda_1 \cos\varphi_k + \boldsymbol{A}\boldsymbol{z}_k \sin\varphi_k, \quad \boldsymbol{A}\boldsymbol{z}_k \in \boldsymbol{Z}_2$$

und weiter

$$\begin{aligned} \varrho_k = RQ(\vartheta_k \boldsymbol{v}^{(k)}) &= (\boldsymbol{u}_1 \cos\varphi_k + \boldsymbol{z}_k \sin\varphi_k)^T (\boldsymbol{u}_1 \lambda_1 \cos\varphi_k + \boldsymbol{A}\boldsymbol{z}_k \sin\varphi_k) \\ &= \lambda_1 \cos^2\varphi_k + \boldsymbol{z}_k^T \boldsymbol{A}\boldsymbol{z}_k \sin^2\varphi_k \\ &= \lambda_1 - (\lambda_1 - \boldsymbol{z}_k^T \boldsymbol{A}\boldsymbol{z}_k)\sin^2\varphi_k. \end{aligned}$$

Nun gilt

$$\begin{aligned} |\lambda_1 - \boldsymbol{z}_k^T \boldsymbol{A}\boldsymbol{z}_k| &\leqq |\lambda_1| + |\boldsymbol{z}_k^T \boldsymbol{A}\boldsymbol{z}_k| \\ &\leqq |\lambda_1| + \|\boldsymbol{z}_k\| \, \|\boldsymbol{A}\boldsymbol{z}_k\| \\ &\leqq |\lambda_1| + |\lambda_2| \\ &\leqq 2|\lambda_1| = 2\|\boldsymbol{A}\|. \end{aligned}$$

Damit folgt endlich

$$|\varrho_k - \lambda_1| \leqq 2\|\boldsymbol{A}\| \sin^2\varphi_k \leqq 2\|\boldsymbol{A}\| \tan^2\varphi_k \leqq 2\|\boldsymbol{A}\| \kappa^{2k} \tan^2\varphi_0.$$

Die Folge der RAYLEIGH-Quotienten konvergiert linear mit dem Faktor κ^2, also schneller als die Folge der Eigenvektorapproximationen.
Bei der praktischen Realisierung der Vektoriteration gilt wegen den auftretenden Rundungsfehlern nur noch der folgende Satz, den wir ohne Beweis angeben.

9.33. Satz: *Es sei λ_1 dominanter Eigenwert der symmetrischen Matrix $\boldsymbol{A}$:*

$$\kappa = |\lambda_2/\lambda_1| < 1,$$

$\boldsymbol{u}_1$ bezeichne einen zugehörigen normierten Eigenvektor. Die Vektoriteration werde in Computerarithmetik so durchgeführt, dass die berechneten Vektoren $\{\boldsymbol{w}^{(k+1)}\}$ der Beziehung

$$\boldsymbol{w}^{(k+1)} = (\boldsymbol{A} + \boldsymbol{\delta A}_k)\boldsymbol{v}^{(k)}, \quad \|\boldsymbol{\delta A}_k\| \leqq \mathrm{eps} F \|\boldsymbol{A}\|$$

genügen. Mit $\varepsilon = \mathrm{eps}(F + n/2)$ gelte

$$\kappa + 10 < 1, \quad |\boldsymbol{u}_1^T \boldsymbol{v}^{(0)}| > \bar{\varepsilon} = \varepsilon/(1 - \kappa - \varepsilon). \tag{9.18}$$

Dann gilt für die den berechneten Vektoren $\{\boldsymbol{v}^{(k)}\}$ zugeordneten Winkel $\{\varphi_k\}$ die Abschätzung

$$0 \leqq \tan\varphi_k \leqq \tau_k, \quad \tau_k \geqq \tau_{k+1}, \quad \lim_{k\to\infty} \tau_k = \bar{\varepsilon}.$$

Ist k_0 ein hinreichend großer Index mit $\tan\varphi_{k_0} < 1$, *so gilt für alle* $k \geqq k_0$

$$\tan\varphi_{k+1} \leqq \hat{\kappa}\tan\varphi_k + \hat{\varepsilon} \leqq \hat{\kappa}^{k-k_0+1}\tan\varphi_{k_0} + \frac{\hat{\varepsilon}}{1-\hat{\kappa}}$$

mit

$$\hat{\kappa} = \frac{\kappa+\varepsilon}{1-\sqrt{2}\varepsilon}, \quad \hat{\varepsilon} = \frac{\varepsilon}{1-\sqrt{2}\varepsilon}.$$

Für alle $k \geqq 1$ *gilt weiterhin*

$$\|\vartheta_k \boldsymbol{v}^{(k)} - \boldsymbol{u}_1\| \leqq \tan\varphi_k + \text{eps}\frac{n}{2},$$

und die gemäß

$$\varrho_k = {\boldsymbol{v}^{(k)}}^T \boldsymbol{w}^{(k+1)}$$

berechneten Eigenwertnäherungen genügen der Abschätzung

$$|\varrho_k - \lambda_1| \leqq \|\boldsymbol{A}\|[2\sin^2\varphi_k + \text{eps}(F+2n)].$$

Bemerkungen: (i) Durch die Bedingungen (9.18) wird einerseits gesichert, dass $|\lambda_1|$ hinreichend größer ist als $|\lambda_2|$, und andererseits wird verhindert, dass $\boldsymbol{u}_1$ und der Startvektor $\boldsymbol{v}^{(0)}$ fast orthogonal sind. Die zweite Bedingung ist nur von theoretischer Bedeutung. Praktisch konvergiert das Verfahren auch, falls $\boldsymbol{u}_1$ und $\boldsymbol{v}^{(0)}$ exakt orthogonal sind, da durch Rundungsfehler immer Komponenten in den Vektoren $\boldsymbol{v}^{(k)}, k > 1$, erzeugt werden, die parallel zu $\boldsymbol{u}_1$ sind.
(ii) Für vollbesetztes $\boldsymbol{A}$ wäre die Konstante $F = n^{3/2}$. Ist $\boldsymbol{A}$ jedoch wie in vielen Anwendungen schwach besetzt, so ergeben sich wesentlich kleinere Werte für F.
Wir wollen nun den Fall betrachten, bei dem mehrere Vektoren gleichzeitig iteriert werden, um einen dominanten Teilraum zu approximieren. Das ist die sogenannte Teilraumiteration.

9.34. Basisverfahren der Teilraumiteration:

S0 *Wähle einen Teilraum* $\boldsymbol{\mathcal{Y}}^{(0)} \subset \mathbb{R}^n$ *mit* $\dim\left(\boldsymbol{\mathcal{Y}}^{(0)}\right) = p$ *und setze* $k = 0$.

S1 *(Iteration) Definiere*

$$\boldsymbol{\mathcal{Y}}^{(k+1)} = \boldsymbol{A}\boldsymbol{\mathcal{Y}}^{(k)} = \left\{ \boldsymbol{A}\boldsymbol{y} \;\middle|\; \boldsymbol{y} \in \boldsymbol{\mathcal{Y}}^{(k)} \right\}.$$

S2 *Setze* $k = k+1$ *und gehe zu Schritt* **S1**.

Um die Konvergenz des Algorithmus zu beschreiben, benötigen wir ein Maß für die Entfernung zweier Teilräume. Es sei $\boldsymbol{\mathcal{S}} \subset \mathbb{R}^n$ ein Teilraum und $\boldsymbol{y} \in \mathbb{R}^n$ ein Vektor ungleich dem Nullvektor. Die Größe

$$\sphericalangle(\boldsymbol{y}, \boldsymbol{\mathcal{S}}) = \min\{\ \sphericalangle(\boldsymbol{y}, \boldsymbol{s}) \mid \boldsymbol{s} \in \boldsymbol{\mathcal{S}}, \boldsymbol{s} \neq \boldsymbol{o}\}$$

heißt Winkel zwischen dem Vektor $\boldsymbol{y}$ und dem Teilraum $\boldsymbol{\mathcal{S}}$.
Es sei $\boldsymbol{\mathcal{Y}} \subset \mathbb{R}^n$ ein weiterer Teilraum. Dann heißt die Größe

$$\sphericalangle(\boldsymbol{\mathcal{Y}}, \boldsymbol{\mathcal{S}}) = \max\{\ \sphericalangle(\boldsymbol{y}, \boldsymbol{\mathcal{S}}) \mid \boldsymbol{y} \in \boldsymbol{\mathcal{Y}}, \boldsymbol{y} \neq \boldsymbol{o}\}$$

Winkel zwischen den Teilräumen $\boldsymbol{\mathcal{Y}}$ und $\boldsymbol{\mathcal{S}}$. Es gilt offensichtlich stets

$$0 \leqq \ \sphericalangle(\boldsymbol{y}, \boldsymbol{\mathcal{S}}) \leqq \pi/2.$$

Man beachte aber, dass für den Winkel zwischen zwei Teilräumen im allgemeinen **nicht** $\sphericalangle(\boldsymbol{\mathcal{Y}}, \boldsymbol{\mathcal{S}}) = \sphericalangle(\boldsymbol{\mathcal{S}}, \boldsymbol{\mathcal{Y}})$ gilt. Um dies einzusehen, betrachte man nur eine Gerade und eine Ebene im $\mathbb{R}^3$. Im Falle $\dim(\boldsymbol{\mathcal{Y}}) = \dim(\boldsymbol{\mathcal{S}})$ gilt

$$\sphericalangle(\boldsymbol{\mathcal{Y}}, \boldsymbol{\mathcal{S}}) = \sphericalangle(\boldsymbol{\mathcal{S}}, \boldsymbol{\mathcal{Y}}) = \sphericalangle(\boldsymbol{\mathcal{Y}}^\perp, \boldsymbol{\mathcal{S}}^\perp).$$

Falls die Teilräume $\boldsymbol{\mathcal{Y}}$ und $\boldsymbol{\mathcal{S}}$ gleiche Dimension haben, ist $\tan(\ \sphericalangle(\boldsymbol{\mathcal{Y}}, \boldsymbol{\mathcal{S}}))$ zur Charakterisierung der Entfernung der beiden Teilräume verwendbar. Es gilt

$$\tan(\ \sphericalangle(\boldsymbol{\mathcal{Y}}, \boldsymbol{\mathcal{S}})) = 0$$

genau dann, wenn $\boldsymbol{\mathcal{Y}} = \boldsymbol{\mathcal{S}}$, und

$$\tan(\ \sphericalangle(\boldsymbol{\mathcal{Y}}, \boldsymbol{\mathcal{S}})) = \infty,$$

falls einer der Teilräume eine Richtung enthält, die orthogonal zum anderen ist. Nun formulieren wir einen Konvergenzsatz.

9.35. Satz: *Es seien $\{\lambda_1, \ldots, \lambda_p\}$ dominante Eigenwerte der symmetrischen Matrix $\boldsymbol{A}$. $\boldsymbol{\mathcal{S}}_p \subset \mathbb{R}^n$ sei der zugehörige p-te dominante Teilraum. Der Teilraum $\boldsymbol{\mathcal{Y}}^{(0)} \subset \mathbb{R}^n$ genüge den Bedingungen*

$$\dim\left(\boldsymbol{\mathcal{Y}}^{(0)}\right) = p, \quad \sigma = \cos\left(\ \sphericalangle(\boldsymbol{\mathcal{Y}}^{(0)}, \boldsymbol{\mathcal{S}}_p)\right) > 0.$$

Dann gilt für die durch die Teilraumiteration erzeugte Folge $\{\boldsymbol{\mathcal{Y}}^{(k)}\}$

1. $\dim\left(\boldsymbol{\mathcal{Y}}^{(k)}\right) = p, \quad k = 1, 2, \ldots,$

2.

$$\tan\varphi_k \leqq \kappa \tan\varphi_{k-1} \leqq \kappa^k \tan\varphi_0 = \kappa^k \sqrt{1-\sigma^2}/\sigma$$

mit

$$\kappa = |\lambda_{p+1}/\lambda_p| < 1, \quad \varphi_k = \sphericalangle\left(\boldsymbol{\mathcal{Y}}^{(k)}, \boldsymbol{\mathcal{S}}_p\right).$$

Beweis: Es ist $\boldsymbol{\mathcal{Y}}^{(k)}$ in Komponenten bezüglich $\boldsymbol{\mathcal{S}}_p$ und

$$\boldsymbol{\mathcal{Z}}_{p+1} = \text{span}(\boldsymbol{u}_{p+1}, \ldots, \boldsymbol{u}_n) = \boldsymbol{\mathcal{S}}_p^{\perp}$$

zu zerlegen. Der Rest verläuft analog zum Beweis des Satzes 9.18. ❀

Zur praktischen Realisierung der Teilraumiteration liegt es nahe, $\boldsymbol{\mathcal{Y}}^{(k)}$ durch p linear unabhängige Vektoren $\boldsymbol{v}_1^{(k)}, \ldots, \boldsymbol{v}_p^{(k)}$ festzulegen. Faßt man diese Vektoren zu einer Matrix

$$\boldsymbol{V}^{(k)} = \left(\boldsymbol{v}_1^{(k)}, \ldots, \boldsymbol{v}_p^{(k)}\right) \in \mathbb{R}^{n\times p}$$

zusammen, so besteht ein Iterationsschritt der Teilraumiteration in der Transformation

$$\boldsymbol{W}^{(k+1)} = \boldsymbol{A}\boldsymbol{V}^{(k)}.$$

Die Spalten der Matrix $\boldsymbol{W}^{(k+1)} \in \mathbb{R}^{n\times p}$ spannen dann den Teilraum $\boldsymbol{\mathcal{Y}}^{(k+1)}$ auf. Zur Vermeidung von Über- und Unterlauf ist $\boldsymbol{W}^{(k+1)}$ zu normieren:

$$\boldsymbol{V}^{(k+1)} = \boldsymbol{W}^{(k+1)} \text{diag}\left(\frac{1}{\|\boldsymbol{w}_1^{(k+1)}\|}, \ldots, \frac{1}{\|\boldsymbol{w}_p^{(k+1)}\|}\right).$$

Für großes k konvergieren alle Vektoren $\boldsymbol{v}_i^{(k)}$ gegen $\pm\boldsymbol{u}_1$ falls $\boldsymbol{u}_1^T\boldsymbol{v}_i^{(0)} \neq 0$ gilt. Die Vektoren $\boldsymbol{v}_1^{(k)}, \ldots, \boldsymbol{v}_p^{(k)}$ werden mit wachsendem k immer stärker linear abhängig. Darum sollte man mit paarweise orthogonalen Vektoren starten und diese immer wieder orthonormieren. Weiterhin ist es günstig, mit einem Teilraum $\boldsymbol{\mathcal{Y}}^{(0)}$ zu starten, dessen Dimension größer als p ist. Wir erhalten so den folgenden Algorithmus.

9.36. Teilraumiteration mit orthonormalen Basen:

S0 *(Initialisierung) Wähle ein m mit $p \leqq m \leqq n$ und eine spaltenorthonormale Matrix $\boldsymbol{V}^{(0)} \in \mathbb{R}^{n\times m}$. Setze $k = 0$.*

S1 *(Iteration) Berechne $\boldsymbol{W}^{(k+1)} = \boldsymbol{A}\boldsymbol{V}^{(k)}$.*

S2 *(Orthogonalisierung) Bestimme eine spaltenorthonormale Matrix*

$$\boldsymbol{V}^{(k+1)} \in \mathbb{R}^{n\times m}$$

und eine obere Dreiecksmatrix $\boldsymbol{R}^{(k+1)} \in \mathbb{R}^{m\times m}$ *mit*

$$\boldsymbol{W}^{(k+1)} = \boldsymbol{V}^{(k+1)}\boldsymbol{R}^{(k+1)}.$$

S3 *Setze* $k = k+1$ *und gehe zu Schritt* **S1**.

Aufwand pro Schritt:

- m *Auswertungen* $Matrix \times Vektor$ *(***S1***)*
- $\sim K_1 nm^2$ *Add./Mult. (***S2***)*
  ($K_1 = 1$ *für* SCHMIDT*sches Orthogonalisierungsverfahren)*

Bemerkungen: (i) Falls die Voraussetzungen von Satz 9.35 mit m statt p erfüllt sind, gelten natürlich alle entsprechenden Aussagen. In exakter Arithmetik wird dann durch $\boldsymbol{W}^{(k+1)}$ der m-te dominante Teilraum approximiert. Für

$$\boldsymbol{\mathcal{Y}}^{(k+1)} = \text{span}\left(\boldsymbol{W}^{(k+1)}\right)$$

gilt

$$\dim\left(\boldsymbol{\mathcal{Y}}^{(k+1)}\right) = rg\left(\boldsymbol{W}^{(k+1)}\right) = m.$$

Der Schritt **S2** im obigen Algorithmus ist daher durchführbar.
(ii) Partitioniert man die Matrizen $\boldsymbol{W}^{(k)}$, $\boldsymbol{V}^{(k)}$ und $\boldsymbol{R}^{(k)}$ gemäß

$$\begin{aligned}\boldsymbol{W}^{(k)} &= \left(\boldsymbol{W}_1^{(k)}, \boldsymbol{W}_2^{(k)}\right), \quad \boldsymbol{W}_1^{(k)} \in \mathbb{R}^{n\times p}, \quad \boldsymbol{W}_2^{(k)} \in \mathbb{R}^{n\times(m-p)},\\ \boldsymbol{V}^{(k)} &= \left(\boldsymbol{V}_1^{(k)}, \boldsymbol{V}_2^{(k)}\right), \quad \boldsymbol{V}_1^{(k)} \in \mathbb{R}^{n\times p}, \quad \boldsymbol{V}_2^{(k)} \in \mathbb{R}^{n\times(m-p)}\end{aligned}$$

und

$$\boldsymbol{R}^{(k)} = \begin{pmatrix} \boldsymbol{R}_{11}^{(k)} & \boldsymbol{R}_{12}^{(k)} \\ \boldsymbol{o} & \boldsymbol{R}_{22}^{(k)} \end{pmatrix}$$

mit

$$\boldsymbol{R}_{11}^{(k)} \in \mathbb{R}^{p\times p}, \quad \boldsymbol{R}_{12}^{(k)} \in \mathbb{R}^{p\times(m-p)}, \quad \boldsymbol{R}_{22}^{(k)} \in \mathbb{R}^{(m-p)\times(m-p)},$$

dann gilt

$$\boldsymbol{W}_1^{(k+1)} = \boldsymbol{A}\boldsymbol{V}_1^{(k)}$$

und

$$\boldsymbol{W}_1^{(k+1)} = \boldsymbol{V}_1^{(k+1)} \boldsymbol{R}_{11}^{(k+1)}.$$

Die Spalten von $\boldsymbol{V}_1^{(k)}$ bilden eine orthonormale Basis von $\boldsymbol{\mathcal{Y}}_1^{(k)} = \text{span}(\boldsymbol{W}_1^{(k)})$, also desjenigen Teilraums, der durch Iteration aus $\boldsymbol{\mathcal{Y}}_1^{(0)} = \text{span}(\boldsymbol{V}_1^{(0)})$ entsteht. Sind die ersten p Eigenwerte dominant, so konvergiert damit die Folge $\{\boldsymbol{\mathcal{Y}}_1^{(k)}\}$ gegen den p-ten dominanten Teilraum $\boldsymbol{\mathcal{S}}_p$.
(iii) Für die Computerrealisierung lassen sich zu Satz 9.33 analoge Aussagen beweisen. Wesentlich für die Konvergenz des Verfahrens ist dabei, dass $\kappa = |\lambda_{p+1}/\lambda_p|$ hinreichend kleiner als 1 ist.
(iv) An das Orthogonalisierungsverfahren in Schritt **S2** brauchen keine besonders hohen Qualitätsanforderungen gestellt werden, da nicht die paarweise Orthogonalität der Spalten von $\boldsymbol{V}^{(k)}$ für die Konvergenz des Verfahrens notwendig ist, sondern nur ihre hinreichend starke lineare Unabhängigkeit.
Mit den Matrizen $\boldsymbol{V}^{(k)}$ hat man orthogonale Basen der Teilräume $\boldsymbol{\mathcal{Y}}^{(k)}$ berechnet, die Approximationen des p-ten dominanten Teilraums $\boldsymbol{\mathcal{S}}_p$ darstellen, sofern

$$\dim(\boldsymbol{V}^{(k)}) = p \text{ und } \kappa = |\lambda_{p+1}/\lambda_p| < 1$$

gilt. Die Spalten von $\boldsymbol{V}^{(k)}$ stellen nicht die besten Näherungen für die ersten p Eigenvektoren dar. Außerdem stellt sich natürlich die Frage nach den zugehörigen Eigenwertapproximationen. Wir haben die Aufgabe, zu einer gegebenen Näherung des p-ten dominanten Teilraums entsprechende Näherungen $(\mu_i, \boldsymbol{z}_i), i = 1, \ldots, p$, für die ersten p Eigenpaare der Matrix $\boldsymbol{A}$ zu konstruieren. Aus den Approximationen μ_i und $\boldsymbol{z}_i$ bilden wir die Matrizen

$$\boldsymbol{M} = \text{diag}(\mu_1, \ldots, \mu_p) \in \mathbb{R}^{p \times p}, \quad \boldsymbol{Z} = (\boldsymbol{z}_1, \ldots, \boldsymbol{z}_p) \in \mathbb{R}^{n \times p}.$$

Wären nun die Paare $(\mu_i, \boldsymbol{z}_i), i = 1, \ldots, p$, exakte Eigenpaare von $\boldsymbol{A}$, so würde $\boldsymbol{AZ} = \boldsymbol{ZM}$ gelten. Als Maß für die Güte der Approximation der Eigenpaare von $\boldsymbol{A}$ durch $\boldsymbol{M}$ und $\boldsymbol{Z}$ verwenden wir darum die Norm der Residuumsmatrix $\boldsymbol{R} = \boldsymbol{AZ} - \boldsymbol{ZM}$. Eine in diesem Sinne beste Approximation der ersten p Eigenpaare einer symmetrischen Matrix $\boldsymbol{A}$, bei einer durch die spaltenorthonormale Matrix $\boldsymbol{V} \in \mathbb{R}^{n \times p}$ gegebenen Approximation des p-ten dominanten Teilraums $\boldsymbol{\mathcal{S}}_p$, erhalten wir durch Lösen der Aufgabe

$$\min \{ \, \|\boldsymbol{AZ} - \boldsymbol{ZM}\|_F \mid \boldsymbol{M} \in \mathcal{D}_p, \boldsymbol{Z} \in \mathcal{OB}(\text{span}(\boldsymbol{V})) \, \}$$

wobei $\mathcal{D}_p$ die Menge aller p-dimensionalen Diagonalmatrizen ist und $\mathcal{OB}(\boldsymbol{\mathcal{Y}})$ für die Menge aller orthonormalen Matrizen $\boldsymbol{Z}$ steht, deren Spalten eine Orthonormalbasis des Teilraums $\boldsymbol{\mathcal{Y}}$ bilden. Eine Lösung dieses Problems liefert der folgende Satz.

9.37. Satz: *Gegeben seien eine symmetrische Matrix $\boldsymbol{A}$ und eine spaltenorthonormale (n,p)-Matrix $\boldsymbol{V}$. Weiterhin sei $\boldsymbol{\mathcal{Y}} = \text{span}(\boldsymbol{V})$ der Teilraum, der durch die Spalten von $\boldsymbol{V}$ aufgespannt wird. Aus $\boldsymbol{V}$ und $\boldsymbol{A}$ werde die symmetrische (p,p)-Matrix*

$$\boldsymbol{P} = \boldsymbol{V}^T \boldsymbol{A} \boldsymbol{V}$$

gebildet. Es sei

$$\boldsymbol{P} = \boldsymbol{X} \boldsymbol{M} \boldsymbol{X}^T$$

die zugehörige Eigenwertzerlegung von $\boldsymbol{P}$:

$$\boldsymbol{M} = \text{diag}(\mu_1, \ldots, \mu_p),$$

$$\boldsymbol{X} \in \mathbb{R}^{p \times p}, \quad \textit{orthogonal.}$$

Dann löst das Paar $(\boldsymbol{M}, \boldsymbol{Z})$ mit $\boldsymbol{Z} = \boldsymbol{V}\boldsymbol{X}$ die Aufgabe

$$\min \{ \, \| \boldsymbol{A}\boldsymbol{Z} - \boldsymbol{Z}\boldsymbol{M} \|_F \mid \boldsymbol{M} \in \mathcal{D}_p, \boldsymbol{Z} \in \mathcal{OB}(\text{span}(\boldsymbol{V})) \, \}.$$

Beweis: Aus $\dim(\boldsymbol{Z}) = p$, $\boldsymbol{\mathcal{Y}} = \text{span}(\boldsymbol{Z}) = \text{span}(\boldsymbol{V})$ und der Spaltenorthogonalität von $\boldsymbol{Z}$ folgt, dass $\boldsymbol{Z}$ die Darstellung $\boldsymbol{Z} = \boldsymbol{V}\tilde{\boldsymbol{X}}$ mit $\tilde{\boldsymbol{X}} \in \mathbb{R}^{p \times p}$ besitzt. Weiter folgt $\boldsymbol{I} = \boldsymbol{Z}^T\boldsymbol{Z} = \tilde{\boldsymbol{X}}^T\boldsymbol{V}^T\boldsymbol{V}\tilde{\boldsymbol{X}} = \tilde{\boldsymbol{X}}^T\tilde{\boldsymbol{X}}$, also die Orthogonalität von $\tilde{\boldsymbol{X}}$. Damit gilt

$$\begin{aligned}
\|\boldsymbol{R}\|_F &= \|\boldsymbol{A}\boldsymbol{Z} - \boldsymbol{Z}\boldsymbol{M}\|_F \\
&= \|(\boldsymbol{A}\boldsymbol{Z} - \boldsymbol{Z}\boldsymbol{M})\tilde{\boldsymbol{X}}^T\|_F \\
&= \|\boldsymbol{A}\boldsymbol{V}\tilde{\boldsymbol{X}}\tilde{\boldsymbol{X}}^T - \boldsymbol{V}\tilde{\boldsymbol{X}}\boldsymbol{M}\tilde{\boldsymbol{X}}^T\|_F \\
&= \|\boldsymbol{A}\boldsymbol{V} - \boldsymbol{V}\tilde{\boldsymbol{P}}\|_F, \quad \tilde{\boldsymbol{P}} = \tilde{\boldsymbol{X}}\boldsymbol{M}\tilde{\boldsymbol{X}}^T.
\end{aligned}$$

Wir ergänzen $\boldsymbol{V}$zu einer orthogonalen (n,n)-Matrix $\bar{\boldsymbol{V}} = (\boldsymbol{V}, \hat{\boldsymbol{V}})$. Es ergibt sich weiter

$$\begin{aligned}
\|\boldsymbol{R}\|_F &= \|\bar{\boldsymbol{V}}^T(\boldsymbol{A}\boldsymbol{V} - \boldsymbol{V}\tilde{\boldsymbol{P}})\|_F \\
&= \left\| \begin{pmatrix} \boldsymbol{V}^T \\ \hat{\boldsymbol{V}}^T \end{pmatrix} (\boldsymbol{A}\boldsymbol{V} - \boldsymbol{V}\tilde{\boldsymbol{P}}) \right\|_F \\
&= \left\| \begin{pmatrix} \boldsymbol{V}^T\boldsymbol{A}\boldsymbol{V} - \tilde{\boldsymbol{P}} \\ \hat{\boldsymbol{V}}^T\boldsymbol{A}\boldsymbol{V} \end{pmatrix} \right\|_F \quad \text{wegen} \quad \boldsymbol{V}^T\boldsymbol{V} = \boldsymbol{I}, \quad \hat{\boldsymbol{V}}^T\boldsymbol{V} = \boldsymbol{o}.
\end{aligned}$$

Damit erhalten wir

$$\|\boldsymbol{R}\|_F = \sqrt{\|\boldsymbol{V}^T\boldsymbol{A}\boldsymbol{V} - \tilde{\boldsymbol{P}})\|_F + \|\hat{\boldsymbol{V}}^T\boldsymbol{A}\boldsymbol{V}\|_F^2}$$

und es gilt

$$\begin{aligned}\min\{\,\|\boldsymbol{R}\|_F\,\} &= \min\{\,\|\boldsymbol{AZ}-\boldsymbol{ZM}\|_F \mid \boldsymbol{M}\in\mathcal{D}_p, \boldsymbol{Z}\in\mathcal{OB}(\text{span}(\boldsymbol{V}))\,\}\\ &\geqq \|\hat{\boldsymbol{V}}^T\boldsymbol{AV}\|_F\end{aligned}$$

und

$$\min\{\|\boldsymbol{R}\|_F\} = \|\hat{\boldsymbol{V}}^T\boldsymbol{AV}\|_F$$

genau dann, wenn

$$\tilde{\boldsymbol{P}} = \boldsymbol{V}^T\boldsymbol{AV}.$$

Ist $\tilde{\boldsymbol{P}} = \boldsymbol{XMX}^T$ die Eigenwertzerlegung von $\tilde{\boldsymbol{P}}$, so löst das Paar $(\boldsymbol{M}, \boldsymbol{VX})$ die vorgegebene Minimierungsaufgabe. ❀

Die in diesem Satz beschriebene Zuordnung

$$(\boldsymbol{A},\boldsymbol{V}) \to (\boldsymbol{M},\boldsymbol{Z})$$

wird RAYLEIGH-RITZ-Algorithmus (kurz RR-Algorithmus) genannt. Die Größen μ_i und $\boldsymbol{z}_i$ heißen RITZsche Eigenwerte bzw. Eigenvektoren der Matrix $\boldsymbol{A}$ bezüglich des Teilraums $\boldsymbol{\mathcal{Y}} = \text{span}(\boldsymbol{V})$. Die Matrix $\boldsymbol{P}$ mit den Elementen $\pi_{ij} = \boldsymbol{v}^{(i)^T}\boldsymbol{Av}^{(j)}$ stellt eine Verallgemeinerung des RAYLEIGH-Quotienten dar und wird Projektion von $\boldsymbol{A}$ auf den Teilraum $\boldsymbol{\mathcal{Y}} = \text{span}(\boldsymbol{V})$ genannt.

9.38. RAYLEIGH-RITZ**-Algorithmus:**
Es seien die symmetrische Matrix $\boldsymbol{A}\in\mathbb{R}^{n\times n}$ und die spaltenorthonormale Matrix $\boldsymbol{V}\in\mathbb{R}^{n\times p}$ gegeben.

S0
- *Berechne $\boldsymbol{B} = \boldsymbol{AV}\in\mathbb{R}^{n\times p}$.*
- *Berechne $\boldsymbol{P} = \boldsymbol{V}^T\boldsymbol{B}\in\mathbb{R}^{p\times p}$.*

S1 *Berechne die Eigenwertzerlegung $\boldsymbol{P} = \boldsymbol{XMX}^T$.*

S2 *Berechne $\boldsymbol{Z} = \boldsymbol{VX}$.*

Aufwand:

- *m Auswertungen $Matrix\times Vektor$ + $\sim np^2/2$ Add./Mult. (**S0**)*
- *$\sim K_2p^3$ Add./Mult. (**S1**) (K_2 hängt vom angewendeten Verfahren ab)*
- *$\sim np^3$ Add./Mult. (**S2**)*

Bemerkungen: (i) Ist $\boldsymbol{\mathcal{Y}}$ invarianter Teilraum von $\boldsymbol{A}$, so gilt

$$\text{span}(\boldsymbol{AV}) \subseteqq \text{span}(\boldsymbol{V}) = \boldsymbol{\mathcal{Y}}.$$

Es gibt daher eine Matrix $\boldsymbol{C} \in \mathbb{R}^{p\times p}$ mit $\boldsymbol{AV} = \boldsymbol{VC}$. Damit folgt

$$\boldsymbol{P} = \boldsymbol{V}^T\boldsymbol{AV} = \boldsymbol{V}^T\boldsymbol{VC} = \boldsymbol{C}.$$

Es gilt also $\boldsymbol{AV} = \boldsymbol{VP}$. (Diese Beziehung folgt nicht einfach aus $\boldsymbol{P} = \boldsymbol{V}^T\boldsymbol{AV}$, da im allgemeinen nicht $\boldsymbol{VV}^T = \boldsymbol{I}$ gilt.) Aus der letzten Beziehung folgt $\boldsymbol{AVX} = \boldsymbol{VPX}$, und weiter $\boldsymbol{AZ} - \boldsymbol{ZM} = \boldsymbol{o}$. Für einen invarianten Teilraum verschwindet die Residuumsmatrix $\boldsymbol{R} = \boldsymbol{AZ} - \boldsymbol{ZM}$. Die RITZschen Eigenpaare sind dann exakte Eigenpaare der Matrix $\boldsymbol{A}$.
(ii) Sind die Spalten von $\boldsymbol{V}$ gute Näherungen für die gesuchten Eigenvektoren, so ist $\boldsymbol{P}$ fast diagonal. Zum Berechnen der Eigenwertzerlegung im Schritt **S2** bietet sich damit das JACOBI-Verfahren an.
Für den Fall, dass $\boldsymbol{\mathcal{Y}}$ kein invarianter Teilraum ist, geben wir noch den folgenden Satz ohne Beweis an, der Aussagen über die Güte der RITZschen Näherungen in Abhängigkeit von der Residuumsmatrix $\boldsymbol{R}$ macht.

9.39. Satz: *Für eine symmetrische (n,n)-Matrix $\boldsymbol{A}$ und eine spaltenorthonormale (n,p)-Matrix $\boldsymbol{V}$ seien durch*

$$\boldsymbol{M} = \operatorname{diag}(\mu_1,\ldots,\mu_p), \quad \boldsymbol{Z} = (\boldsymbol{z}_1,\ldots,\boldsymbol{z}_p)$$

die in exakter Arithmetik berechneten RITZ*schen Eigenwerte und Eigenvektoren gegeben.*

$$\boldsymbol{R} = \boldsymbol{AZ} - \boldsymbol{ZM} = (\boldsymbol{r}_1,\ldots,\boldsymbol{r}_p)$$

sei die zugehörige Residuumsmatrix. Dann gilt

(i) Es gibt symmetrische Störungen $\boldsymbol{\delta A}_j$ mit

$$(\boldsymbol{A} + \boldsymbol{\delta A}_j)\boldsymbol{z}_j = \mu_j\boldsymbol{z}_j, \quad \|\boldsymbol{\delta A}_j\|_2 \leqq \|\boldsymbol{r}_j\|_2, \quad j = 1,\ldots,p.$$

(ii) Es gibt Eigenwerte $\lambda_{l(1)},\ldots,\lambda_{l(p)}$ von $\boldsymbol{A}$ mit

$$|\mu_j - \lambda_{l(j)}| \leqq \|\boldsymbol{R}\|_2, \quad j = 1,\ldots,p,$$

$$\sqrt{\sum_{j=1}^{p}\left(\mu_j - \lambda_{l(j)}\right)^2} \leqq \sqrt{2}\|\boldsymbol{R}\|_F.$$

9.5. Die Inverse Iteration nach Wielandt

Die Vektor- und die Teilraumiteration lieferten nur Approximationen für Eigenvektoren bzw. für Eigenräume, die zu dominanten Eigenwerten gehören. Die Konvergenz der Algorithmen ist im allgemeinen langsam. In vielen Anwendungen werden andere Eigenvektoren gesucht, so zum Beispiel die zu einer vorgegebenen Eigenwertnäherung gehörende beste Eigenvektornäherung oder ein Eigenvektor zum betragskleinsten Eigenwert. Diese Aufgaben lassen sich mit der Vektoriteration in der ursprünglichen Form nicht lösen. Wenden wir die Vektoriteration auf die Inverse $\boldsymbol{A}^{-1}$ der Matrix $\boldsymbol{A}$ an (vorausgesetzt $\boldsymbol{A}$ ist regulär), so approximieren wir auf diese Weise einen Eigenvektor zum betragskleinsten Eigenwert von $\boldsymbol{A}$, denn aus der Eigenwertzerlegung von $\boldsymbol{A}$

$$\boldsymbol{A} = \boldsymbol{U}\Lambda\boldsymbol{U}^T$$

ergibt sich die Eigenwertzerlegung der Inversen von $\boldsymbol{A}$

$$\boldsymbol{A}^{-1} = \boldsymbol{U}\Lambda^{-1}\boldsymbol{U}^T.$$

Sind $\lambda_1, \ldots, \lambda_n$ die Eigenwerte der regulären Matrix $\boldsymbol{A}$, so sind $\frac{1}{\lambda_1}, \ldots, \frac{1}{\lambda_n}$ die Eigenwerte von $\boldsymbol{A}^{-1}$. Durch Spektralverschiebung $\boldsymbol{A}_\mu = \boldsymbol{A} - \mu\boldsymbol{I} = \boldsymbol{U}(\Lambda - \mu\boldsymbol{I})\boldsymbol{U}^T$ lässt sich jeder Eigenwert von $\boldsymbol{A}$ zum betragskleinsten Eigenwert von $\boldsymbol{A}_\mu$ machen. Dies sind die grundlegenden Ideen der Inversen Iteration.

9.40. Basisverfahren der Inversen Iteration:

S0 *(Initialisierung) Wähle ein $\mu \neq \lambda_j, j = 1, \ldots, n$ und einen Startvektor $\boldsymbol{v}^{(0)} \in \mathbb{R}^n$ mit $\|\boldsymbol{v}^{(0)}\| = 1$. Setze $k = 0$.*

S1 *(Iteration) Berechne $\boldsymbol{w}^{(k+1)}$ aus*

$$\boldsymbol{A}_\mu \boldsymbol{w}^{(k+1)} = (\boldsymbol{A} - \mu\boldsymbol{I})\boldsymbol{w}^{(k+1)} = \boldsymbol{v}^{(k)}.$$

S2 *(Normierung) Setze $\boldsymbol{v}^{(k+1)} = \boldsymbol{w}^{(k+1)}/\|\boldsymbol{w}^{(k+1)}\|$.*

S3 *Setze $k = k + 1$ und gehe zu Schritt* **S1**.

Bemerkungen: (i) Die Matrix $(\boldsymbol{A} - \mu\boldsymbol{I})^{-1}$ muss nicht berechnet werden. Statt dessen werden im Schritt **S1** jeweils lineare Gleichungssysteme mit der Matrix $\boldsymbol{A}_\mu$ gelöst. Dazu bietet sich eine LDL^T-Zerlegung oder eine LU-Zerlegung an, die nur einmal berechnet wird. Das Lösen der Gleichungssysteme im Schritt **S1** benötigt dann jeweils $\sim n^2$ Add./Mult. Das ist derselbe Aufwand wie bei der einfachen Vektoriteration.
(ii) Wenn wir annehmen, dass die Eigenwerte von $\boldsymbol{A}$ so geordnet sind, dass

$$|\lambda_1 - \mu| < |\lambda_2 - \mu| \leqq |\lambda_3 - \mu| \leqq \cdots \leqq |\lambda_n - \mu|$$

gilt, so lässt sich der Konvergenzsatz 9.32 unmittelbar anwenden, falls

$$\kappa = \left| \frac{\lambda_1 - \mu}{\lambda_2 - \mu} \right|$$

gesetzt wird. Die Konvergenzgeschwindigkeit ist um so größer, je besser die Eigenwertnäherung μ ist. Mit besserer Eigenwertnäherung verschlechtert sich die Kondition der Matrix $\boldsymbol{A}_\mu$, so dass beim Lösen der linearen Gleichungssysteme in Schritt **S1** Probleme auftreten.
(iii) Für Schritt **S1** gilt im allgemeinen

$$(\boldsymbol{A}_\mu + \boldsymbol{\delta A}_k)\boldsymbol{w}^{(k+1)} = \boldsymbol{v}^{(k)}, \quad \|\boldsymbol{\delta A}_k\| \leqq \text{eps} F \|\boldsymbol{A}_\mu\|,$$

wobei F vom Verfahren zum Lösen des linearen Gleichungssystems abhängt. Ist $\boldsymbol{A}_\mu$ im Sinne von $\text{eps} F \text{cond}(\boldsymbol{A}_\mu) = \overline{\kappa} < 1$ genügend gut konditioniert (μ ist keine zu gute Näherung für λ_1), so ist $\boldsymbol{A}_\mu + \boldsymbol{\delta A}_k$ regulär, und es gilt

$$\boldsymbol{w}^{(k+1)} = (\boldsymbol{A}_\mu + \boldsymbol{\delta A}_k)^{-1}\boldsymbol{v}^{(k)} = (\hat{\boldsymbol{A}}_\mu + \boldsymbol{\delta}\hat{\boldsymbol{A}}_k)\boldsymbol{v}^{(k)}$$

mit

$$\hat{\boldsymbol{A}} = \boldsymbol{A}_\mu^{-1}, \quad \boldsymbol{\delta}\hat{\boldsymbol{A}}_k = (\boldsymbol{A}_\mu + \boldsymbol{\delta A}_k)^{-1} - \boldsymbol{A}_\mu^{-1}.$$

Für die Störung $\boldsymbol{\delta}\hat{\boldsymbol{A}}_k$ ergibt sich dann die Abschätzung

$$\|\boldsymbol{\delta}\hat{\boldsymbol{A}}_k\| \leqq \frac{\|\hat{\boldsymbol{A}}\|^2}{1-\overline{\kappa}} \|\boldsymbol{\delta A}_k\| \leqq \hat{F}\|\hat{\boldsymbol{A}}\|, \quad \hat{F} = \frac{\text{cond}(\boldsymbol{A}_\mu)}{1-\overline{\kappa}} F.$$

Die Voraussetzungen von Satz 9.33 sind mit $(\hat{\boldsymbol{A}}, \hat{F})$ statt $(\boldsymbol{A}, F)$ erfüllt, falls

$$\kappa + 10\text{eps}\left(\hat{F} + \frac{n}{2}\right) = \left| \frac{\lambda_1 - \mu}{\lambda_2 - \mu} \right| + 10\text{eps}\left(\hat{F} + \frac{n}{2}\right) < 1$$

gilt.
(iv) Falls die Eigenvektoren zu den p betragskleinsten Eigenwerten gesucht sind, lässt sich das Verfahren der Inversen Iteration natürlich auch im Sinne der Teilraumiteration verallgemeinern. Man wendet die Teilraumiteration mit $\boldsymbol{A}^{-1}$ statt $\boldsymbol{A}$ an. Das Verfahren konvergiert gegen den p-ten dominanten Teilraum von $\boldsymbol{A}^{-1}$, falls $0 < |\lambda_1| \leqq \cdots \leqq |\lambda_p| < |\lambda_{p+1}|$ gilt. Ist μ eine gute Näherung für λ_1, so ist $\boldsymbol{A}_\mu$ schlecht konditioniert. Die Bedingung $\overline{\kappa} < 1$ ist dann nicht mehr erfüllt. Wir dürfen daher den Satz 9.33 nicht anwenden. Trotzdem ist natürlich die Inverse Iteration anwendbar, um zu μ eine akzeptable Eigenvektorapproximation zu berechnen. Man sollte folgenden speziellen Algorithmus anwenden.

9.41. Inverse Iteration bei guter Eigenwertnäherung:
Es sei μ ein im Sinne von

$$|\mu - \lambda| \leqq \text{eps} F_0 \|\boldsymbol{A}\| = \varepsilon_0$$

gute Näherung für den Eigenwert λ von $\boldsymbol{A}$.

S0 *(Initialisierung)*

- *Berechne Zerlegung $\boldsymbol{A}_\mu = \boldsymbol{P}_\mu^T \boldsymbol{L}_\mu \boldsymbol{U}_\mu$.*
- *Berechne $\boldsymbol{w}^{(0)}$ aus $\boldsymbol{U}_\mu \boldsymbol{w}^{(0)} = \boldsymbol{e} = (1, \ldots, 1)^T$.*
- *Wähle ein $\varepsilon > 0$ und setze $k = 0$.*

S1 *(Iteration) Berechne $\boldsymbol{w}^{(k+1)}$ aus*

$$\boldsymbol{A}_\mu \boldsymbol{w}^{(k+1)} = \boldsymbol{P}_\mu^T \boldsymbol{L}_\mu \boldsymbol{U}_\mu \boldsymbol{w}^{(k+1)} = \boldsymbol{v}^{(k)}.$$

S2 *(Normierung) Setze*

$$\omega_{k+1} = \|\boldsymbol{w}^{(k+1)}\|, \quad \boldsymbol{v}^{(k+1)} = \boldsymbol{w}^{(k+1)} / \omega_{k+1}.$$

S3 *(Abbruchtest) Falls $\varepsilon \cdot \omega_{k+1} < 1$ so setze $k = k + 1$ und gehe zu Schritt* **S1**.
Aufwand:

- *$\sim n^3/3$ Add./Mult. (***S0***)*
- *$\sim n^2$ Add./Mult. pro Schritt für vollbesetztes $\boldsymbol{A}$ (***S1*** und* **S2***)*

Ohne Beweis geben wir folgenden Satz über die Konvergenz des Algorithmus an.

9.42. Satz: *Das obige Verfahren werde in Computerarithmetik durchgeführt. Dann gilt*

(i) Falls das Verfahren nach k Schritten mit $\varepsilon \cdot \omega_{k+1} \geqq 1$ abbricht, so gibt es eine symmetrische Störung $\boldsymbol{\delta A}$, so dass $\{\mu, \boldsymbol{v}\}$, mit $\boldsymbol{v} = \boldsymbol{v}^{(k+1)}$, der Beziehung

$$(\boldsymbol{A} + \boldsymbol{\delta A})\boldsymbol{v} = \mu \boldsymbol{v}, \quad \|\boldsymbol{\delta A}\| \sim \varepsilon \|\boldsymbol{A}\|$$

genügt; $\{\mu, \boldsymbol{v}\}$ ist daher ein akzeptables Eigenpaar von $\boldsymbol{A}$.

(ii) Das Abbruchkriterium

$$\varepsilon \cdot \omega_{k+1} \geqq 1$$

ist meist nach wenigen Schritten erfüllt ($k \leqq 3$).

Bemerkungen: (i) Für eine gute Eigenwertnäherung ist $|\mu - \lambda_1| \leqq \text{eps} F_0 \|\boldsymbol{A}\|$. Dann gilt für die Kondition der Matrix $\boldsymbol{A}_\mu$

$$\text{cond}\,(\boldsymbol{A}_\mu) = \left|\frac{\mu - \lambda_n}{\mu - \lambda_1}\right| \geqq \frac{|\mu - \lambda_n|}{\text{eps} F_0 \|\boldsymbol{A}\|}.$$

Für kleines F_0 liegt $\text{cond}\,(\boldsymbol{A}_\mu)$ in der Größenordnung $1/\text{eps}$. Die Durchführbarkeit der LU-Zerlegung mit regulärer oberer Dreiecksmatrix $\boldsymbol{U}_\mu$ ist nicht gesichert. Praktisch wird $\boldsymbol{U}_\mu$ fast immer regulär sein mit eventuell kleinen Diagonalelementen. Sollte im Algorithmus sogar $(\boldsymbol{U}_\mu)_{jj} = 0$ gelten, so setze man $(\boldsymbol{U}_\mu)_{jj} = \text{eps}\|\boldsymbol{A}\|$.
(ii) Man sollte in Schritt **S0** $\varepsilon = \text{eps}\sqrt{n}\|\boldsymbol{A}\|$ wählen.
(iii) Normalerweise ist im Falle $\text{cond}\,(\boldsymbol{A}_\mu) \approx 1/\text{eps}$ damit zu rechnen, dass die Lösung von $\boldsymbol{A}_\mu \boldsymbol{w} = \boldsymbol{v}$ einen großen Fehler aufweist. Für die Inverse Iteration ist nur die Richtung $\boldsymbol{w}/\|\boldsymbol{w}\|$ wesentlich, die auch bei schlechter Näherung für $\boldsymbol{w}$ noch gut approximiert wird, falls ein numerisch gutartiges Verfahren angewendet wird, und falls der Eigenwert λ_1 genügend stark von den anderen Eigenwerten getrennt ist. Nur für den unwahrscheinlichen Fall $\boldsymbol{\delta w} \approx -\boldsymbol{w}^* = -\boldsymbol{A}_\mu^{-1}\boldsymbol{v}$, also für kleines $\|\boldsymbol{w}\|$, weicht die berechnete Richtung stark von der Richtung des Eigenvektors $\boldsymbol{u}_1$ ab.
Ist μ keine gute Eigenwertnäherung, so lässt sich μ während der Iteration verbessern. Dazu bietet sich wieder der RAYLEIGH-Quotient an. Wir erhalten folgenden Algorithmus.

9.43. RAYLEIGH**-Quotienten-Iteration:**

S0 *(Initialisierung) Wähle* $\boldsymbol{v}^{(0)}$ *mit* $\|\boldsymbol{v}^{(0)}\|_2 = 1$ *und ein* $\varepsilon > 0$. *Setze* $k = 0$.

S1 *(Berechnen der Verschiebung) Berechne*

$$\varrho_k = RQ(\boldsymbol{v}^{(k)}) = \boldsymbol{v}^{(k)^T} \boldsymbol{A} \boldsymbol{v}^{(k)}.$$

S2 *(Inverse Iteration) Berechne* $\boldsymbol{w}^{(k+1)}$ *aus*

$$(\boldsymbol{A} - \varrho_k \boldsymbol{I})\boldsymbol{w}^{(k+1)} = \boldsymbol{v}^{(k)}.$$

S3 *(Normierung) Setze* $\omega_{k+1} = \|\boldsymbol{w}^{(k+1)}\|_2$ *und* $\boldsymbol{v}^{(k+1)} = \boldsymbol{w}^{(k+1)}/\omega_{k+1}$.

S4 *(Abbruchtest) Falls* $\varepsilon \cdot \omega_{k+1} < 1$ *so setze* $k = k + 1$ *und gehe zu Schritt* **S1**.

Aufwand pro Schritt:

- $\sim n^2$ *Add./Mult. (***S1***)*
- $\sim K_1 n^3$ *Add./Mult. für vollbesetztes* $\boldsymbol{A}$ *(***S2***)*
- $\sim Kn$ *Add./Mult. für tridiagonales* $\boldsymbol{A}$ *(***S2***)*
- $\sim n$ *Add. +* $\sim 2n$ *Mult. + 1 Quadratwurzel (***S3***)*

Es gilt der folgende Satz (ohne Beweis):

9.44. Satz: *Die* RAYLEIGH*-Quotienten-Iteration werde mit $\varepsilon = 0$ in exakter Arithmetik durchgeführt. Dann gilt:*

(i) Die Normen der Residuen $\boldsymbol{r}^{(k)} = (\boldsymbol{A} - \varrho_k \boldsymbol{I})\boldsymbol{v}^{(k)}$ fallen monoton, und die Folge $\{\varrho_k, \vartheta_k \boldsymbol{v}^{(k)}\}$ konvergiert fast immer gegen ein Eigenpaar $\{\lambda_j, \boldsymbol{u}_j\}$.
(Die Größen $\vartheta_k \in \{1, -1\}$ werden dabei so gewählt, dass $\boldsymbol{u}_j^T(\vartheta_k \boldsymbol{v}^{(k)}) \geqq 0$.)

(ii) Falls ϱ_k gegen λ_j konvergiert, so ist die Konvergenz kubisch im Sinne von

$$\tan \varphi_{k+1} \leqq \eta \, (\tan \varphi_k)^3 .$$

Dabei ist $\eta > 0$ und $\varphi_k = \sphericalangle(\boldsymbol{u}_j, \vartheta_k \boldsymbol{v}^{(k)})$.

Bemerkungen: (i) Falls $\boldsymbol{A} - \varrho_k \boldsymbol{I}$ bei der praktischen Realisierung numerisch singulär wird, so sollte man wie in Bemerkung (i) auf Seite 50 verfahren.
(ii) Das im Satz 9.44 beschriebene ideale Konvergenzverhalten wird bei der Computerrealisierung gestört. Praktisch tritt fast immer eine Verbesserung ein, so dass mit schneller Konvergenz zu rechnen ist.
(iii) Allgemein lässt sich nichts darüber ausgesagen, gegen welchen Eigenwert λ_j von $\boldsymbol{A}$ die RAYLEIGH-Quotienten ϱ_k konvergieren. Es braucht insbesondere nicht derjenige Eigenwert zu sein, dessen Abstand zu ϱ_1 am kleinsten ist.

9.6. Transformationen auf Tridiagonalform

Der Aufwand der meisten Algorithmen ist beträchtlich kleiner, falls sie auf Matrizen einfacherer Gestalt angewendet werden. Darum ist es oft günstig, die zu behandelnde Matrix in einem ersten Schritt durch Ähnlichkeitstransformation auf so eine Form zu bringen. Für symmetrische Matrizen arbeitet man dabei mit orthogonalen Transformationen. Als Ziel bietet sich Tridiagonalform an (Diagonalform wäre schon die Lösung des Problems). Wir suchen eine orthogonale (n,n)-Matrix $\boldsymbol{Q}$, die die symmetrische (n,n)-Matrix $\boldsymbol{A}$ gemäß

$$\boldsymbol{A} \to \boldsymbol{T} = \boldsymbol{Q}^T \boldsymbol{A} \boldsymbol{Q} = \begin{pmatrix} a_1 & b_2 & 0 & \cdots & 0 \\ b_2 & a_2 & b_3 & \ddots & \vdots \\ 0 & \ddots & \ddots & \ddots & 0 \\ \vdots & \ddots & b_{n-1} & a_{n-1} & b_n \\ 0 & \cdots & 0 & b_n & a_n \end{pmatrix} = \mathrm{trid}(a_1, \ldots, a_n, b_2, \ldots, b_n)$$

auf Tridiagonalgestalt transformiert. Wird $\boldsymbol{A}$ noch durch $\sim n^2/2$ Daten repräsentiert, so sind es bei $\boldsymbol{T}$ nur noch $\sim 2n$ Elemente. Durch diese Datenreduktion tritt

eine wesentliche Verringerung des Aufwands zum Lösen des Eigenwertproblems ein. Die orthogonale Matrix $\boldsymbol{Q}$ ist mit Hilfe von HOUSEHOLDER-Spiegelungen oder GIVENS-Drehungen konstruierbar. Wir wollen hier nur die Vorgehensweise bei der Anwendung von HOUSEHOLDER-Matrizen erläutern.

$$\begin{array}{lll} \text{Es sei} \quad \boldsymbol{A}^{(1)} & = & \boldsymbol{A} \\ & \boldsymbol{A}^{(k+1)} & = \boldsymbol{H}^{(k)}\boldsymbol{A}^{(k)}\boldsymbol{H}^{(k)}, \quad k = 1,\ldots,n-2 \\ & \boldsymbol{T} & = \boldsymbol{A}^{(n-1)} \end{array}$$

Die Matrizen $\boldsymbol{A}^{(k)}$ haben dabei folgende Struktur

$$\boldsymbol{A}^{(k)} = \left(\begin{array}{cc|c} \multicolumn{2}{c|}{\multirow{2}{*}{$\boldsymbol{T}^{(k)}$}} & \boldsymbol{o} \\ & & \boldsymbol{b}^{(k)T} \\ \hline \boldsymbol{o} & \boldsymbol{b}^{(k)} & \boldsymbol{B}^{(k)} \end{array}\right)$$

mit $\boldsymbol{T}^{(k)} = \text{trid}(a_1,\ldots,a_k,b_2,\ldots,b_k) \in \mathbb{R}^{k\times k}$, $\boldsymbol{b}^{(k)} \in \mathbb{R}^{n-k}$ und $\boldsymbol{B}^{(k)} \in \mathbb{R}^{(n-k)\times(n-k)}$. Die HOUSEHOLDER-Transformation $\boldsymbol{H}^{(k)}$ wird nun im k-ten Schritt so gewählt, dass durch sie $\boldsymbol{b}^{(k)}$ auf ein Vielfaches des entsprechenden Einheitsvektors transformiert wird:

$$\boldsymbol{H}^{(k)} = \boldsymbol{I} \quad \text{falls} \quad \boldsymbol{b}^{(k)} = \boldsymbol{o} \quad \text{bzw.} \tag{9.19}$$

$$\boldsymbol{H}^{(k)} = \begin{pmatrix} \boldsymbol{I}_k & \boldsymbol{o} \\ \boldsymbol{o} & \bar{\boldsymbol{H}}^{(k)} \end{pmatrix}, \quad \boldsymbol{I}_k \in \mathbb{R}^{k\times k}, \quad \bar{\boldsymbol{H}}^{(k)} \in \mathbb{R}^{(n-k)\times(n-k)} \tag{9.20}$$

Die Matrix $\bar{\boldsymbol{H}}^{(k)}$ wird gemäß

$$\bar{\boldsymbol{H}}^{(k)} = \boldsymbol{I} - \bar{\boldsymbol{v}}_k\bar{\boldsymbol{v}}_k^T/\gamma_k \tag{9.21}$$

mit

$$\bar{\boldsymbol{v}}_k = \boldsymbol{e}_1 - \boldsymbol{b}^{(k)}/\varrho_k, \tag{9.22}$$

(9.23)

$$\varrho_k = \begin{cases} \|\boldsymbol{b}^{(k)}\| & \text{für} \quad a_{k+1,k}^{(k)} \leq 0 \\ & \text{und} \\ -\|\boldsymbol{b}^{(k)}\| & \text{für} \quad a_{k+1,k}^{(k)} > 0 \end{cases} \tag{9.24}$$

(9.25)

$$\gamma_k = \bar{\boldsymbol{v}}_k^T\bar{\boldsymbol{v}}_k/2 \quad (\text{Damit gilt } \gamma_k = v_{k+1,k}.) \tag{9.26}$$

berechnet. Damit erhält man

$$\boldsymbol{A}^{(k+1)} = \boldsymbol{H}^{(k)}\boldsymbol{A}^{(k)}\boldsymbol{H}^{(k)} \tag{9.27}$$

$$= \left(\begin{array}{c|c|c} \boldsymbol{T}^{(k)} & & \boldsymbol{o} \\ & & \overline{\boldsymbol{b}}^{(k)^T} \\ \hline \boldsymbol{o} & \overline{\boldsymbol{b}}^{(k)} & \overline{\boldsymbol{B}}^{(k+1)} \end{array}\right) = \left(\begin{array}{c|c|c} \boldsymbol{T}^{(k+1)} & & \boldsymbol{o} \\ & & \boldsymbol{b}^{(k+1)^T} \\ \hline \boldsymbol{o} & \boldsymbol{b}^{(k+1)} & \boldsymbol{B}^{(k+1)} \end{array}\right) \tag{9.28}$$

mit

$$\overline{\boldsymbol{b}}^{(k)} = \overline{\boldsymbol{H}}^{(k)}\boldsymbol{b}^{(k)}, \tag{9.29}$$

$$\overline{\boldsymbol{B}}^{(k)} = \overline{\boldsymbol{H}}^{(k)}\boldsymbol{B}^{(k)}\overline{\boldsymbol{H}}^{(k)} = \begin{pmatrix} a_{k+1,k+1} & \boldsymbol{b}^{(k+1)^T} \\ \boldsymbol{b}^{(k+1)} & \boldsymbol{B}^{(k+1)} \end{pmatrix}. \tag{9.30}$$

Im Unterschied zum HOUSEHOLDER-Verfahren bei linearen Gleichungssystemen oder linearen Ausgleichsproblemen tritt hier die Spiegelung $\boldsymbol{H}^{(k)}$ auf beiden Seiten auf. Mit $\boldsymbol{B}$ ist auch $\overline{\boldsymbol{B}} = \boldsymbol{HBH}$ symmetrisch. Es ist daher nicht zweckmäßig, $\overline{\boldsymbol{B}}$ nach den Vorschriften $\overline{\boldsymbol{B}} = (\boldsymbol{HB})\boldsymbol{H}$ oder $\overline{\boldsymbol{B}} = \boldsymbol{H}(\boldsymbol{BH})$ zu berechnen. Hierbei würde durch den Einfluss von Rundungsfehlern die Symmetrie verloren gehen. Eine symmetrische Berechnungsvorschrift lässt sich folgendermaßen herleiten:
Es gilt $\boldsymbol{H} = \boldsymbol{I} - \boldsymbol{v}\boldsymbol{v}^T/\gamma$. Damit ergibt sich

$$\overline{\boldsymbol{B}} = \boldsymbol{HBH} \tag{9.31}$$

$$= (\boldsymbol{I} - \boldsymbol{v}\boldsymbol{v}^T/\gamma)\boldsymbol{B}(\boldsymbol{I} - \boldsymbol{v}\boldsymbol{v}^T/\gamma) \tag{9.32}$$

$$= \boldsymbol{B} - \boldsymbol{v}\boldsymbol{v}^T\boldsymbol{B}/\gamma - \boldsymbol{B}\boldsymbol{v}\boldsymbol{v}^T/\gamma + \boldsymbol{v}\boldsymbol{v}^T\boldsymbol{v}\boldsymbol{v}^T/\gamma^2 \tag{9.33}$$

$$= \boldsymbol{B} - \boldsymbol{w}\boldsymbol{v}^T - \boldsymbol{v}\boldsymbol{w}^T + \frac{\boldsymbol{v}^T\boldsymbol{w}}{\gamma}\boldsymbol{v}\boldsymbol{v}^T \tag{9.34}$$

mit

$$\boldsymbol{w} = \boldsymbol{B}\boldsymbol{v}/\gamma. \tag{9.35}$$

Das lässt sich noch weiter umformen zu

$$\overline{\boldsymbol{B}} = \boldsymbol{B} - (\boldsymbol{p}\boldsymbol{v}^T + \boldsymbol{v}\boldsymbol{p}^T) \tag{9.36}$$

mit

$$\boldsymbol{p} = \boldsymbol{w} - \frac{\boldsymbol{v}^T\boldsymbol{w}}{2\gamma}\boldsymbol{v}. \tag{9.37}$$

Hier ergibt sich $\bar{\boldsymbol{B}}$ aus $\boldsymbol{B}$durch eine symmetrische Rang-2-Modifikation. Nach $n-2$ Schritten erhält man durch diesen Algorithmus mit

$$\boldsymbol{A}^{(n-1)} = \left(\begin{array}{cc|c} \multicolumn{2}{c|}{\multirow{2}{*}{$\boldsymbol{T}^{(n-1)}$}} & \boldsymbol{o} \\ & & \overline{\boldsymbol{b}}^{(n-1)^T} \\ \hline \boldsymbol{o} & \overline{\boldsymbol{b}}^{(n-1)} & \overline{\boldsymbol{B}}^{(n-1)} \end{array}\right) = \boldsymbol{T} = \text{trid}(a_1, \ldots, a_n, b_2, \ldots, b_n)$$

die gesuchte Tridiagonalmatrix. Es gilt

$$\boldsymbol{T} = \boldsymbol{Q}^T \boldsymbol{A} \boldsymbol{Q}$$

mit

$$\boldsymbol{Q} = \boldsymbol{H}^{(1)} \boldsymbol{H}^{(2)} \cdots \boldsymbol{H}^{(n-2)}.$$

Insgesamt ergibt sich der folgende Algorithmus:

9.45. Tridiagonalisierung einer symmetrischen Matrix mittels HOUSEHOLDER**-Spiegelungen:**

S0 *(Initialisierung) Setze* $k = 1$ *und* $\boldsymbol{A}^{(1)} = \boldsymbol{A}$.

S1 *(Berechnen der* HOUSEHOLDER*-Spiegelung)*
Berechne $\boldsymbol{H}^{(k)}$ *nach (9.19)-(9.26).*
(Iteration) Berechne $\boldsymbol{A}^{(k+1)} = \boldsymbol{H}^{(k)} \boldsymbol{A}^{(k)} \boldsymbol{H}^{(k)}$ *nach (9.28)-(9.37).*

S2 *(Abbruch) Falls* $k < n-2$ *so setze* $k = k+1$ *und gehe zu Schritt* **S1**.

Aufwand: $\sim 2n^3/3$ *Add./Mult. und* $\sim n$ *Quadratwurzeln falls* $\boldsymbol{A}$ *voll besetzt ist.*

Bemerkungen:
(i) Der Algorithmus ist auf dem Platz von $\boldsymbol{A}$ durchführbar. Die Elemente $a_{ij}, i \geqq j+1$ werden mit den Vektoren $\boldsymbol{v}^{(k)}$ überspeichert. Die Subdiagonalelemente b_i sind in einem extra Feld zu speichern.
(ii) Der Algorithmus ist in dem Sinne gutartig, dass zur berechneten Matrix $\boldsymbol{T}$ eine orthogonale Matrix $\boldsymbol{P}$ und eine Störung $\boldsymbol{\delta A}$ existiert, so dass

$$\boldsymbol{A} + \boldsymbol{\delta A} = \boldsymbol{P} \boldsymbol{T} \boldsymbol{P}^T$$

mit $\|\boldsymbol{\delta A}\|_F \leqq \text{eps} 4.2 n^2 \|\boldsymbol{A}\|_F$ bzw. $\|\boldsymbol{\delta A}\|_2 \leqq \text{eps} 0.4 n^{5/2}(1 + 9.5/\sqrt{n})\|\boldsymbol{A}\|_2$ gilt.

9.7. Der Lanczos-Algorithmus

In diesem Abschnitt werden wir einen anderen Algorithmus zur Tridiagonalisierung einer Matrix kennenlernen. Durch diesen sogenannten LANCZOS-Algorithmus wird eine Folge von Tridiagonalmatrizen wachsender Dimension erzeugt, deren größte und kleinste Eigenwerte schnell gegen die entsprechenden Eigenwerte der zugrundeliegenden Matrix $\boldsymbol{A}$ konvergieren.
Es sei die symmetrische (n,n)-Matrix $\boldsymbol{A}$ gegeben. Für einen Vektor $\boldsymbol{q} \in \mathbb{R}^n$ heißt die Folge

$$\boldsymbol{q}, \boldsymbol{A}\boldsymbol{q}, \boldsymbol{A}^2\boldsymbol{q}, \ldots$$

KRYLOV-Sequenz. Die ersten i Vektoren einer KRYLOV-Sequenz spannen einen Teilraum des $\mathbb{R}^n$, den KRYLOV-Raum, auf:

$$\begin{aligned} \mathcal{K}_i(\boldsymbol{q}, \boldsymbol{A}) &= \text{span}(\boldsymbol{q}, \boldsymbol{A}\boldsymbol{q}, \boldsymbol{A}^2\boldsymbol{q}, \ldots, \boldsymbol{A}^{i-1}\boldsymbol{q}), \quad i \geqq 1, \\ \mathcal{K}_0(\boldsymbol{q}, \boldsymbol{A}) &= \{\boldsymbol{o}\}. \end{aligned}$$

Für einen festen Vektor $\boldsymbol{q} \in \mathbb{R}^n$ sei m die maximale Dimension der zugehörigen KRYLOV-Räume:

$$m = \max \left\{ \dim(\mathcal{K}_i(\boldsymbol{q}, \boldsymbol{A})) \mid 1 \leqq i \leqq n \right\}.$$

Für den KRYLOV-Raum $\mathcal{K}_m(\boldsymbol{q}, \boldsymbol{A})$ sind die Vektoren

$$\boldsymbol{q}, \boldsymbol{A}\boldsymbol{q}, \boldsymbol{A}^2\boldsymbol{q}, \ldots, \boldsymbol{A}^{m-1}\boldsymbol{q}$$

linear unabhängig, und für den Vektor $\boldsymbol{A}^m\boldsymbol{q}$ gilt

$$\boldsymbol{A}^m\boldsymbol{q} \in \mathcal{K}_m(\boldsymbol{q}, \boldsymbol{A}).$$

Daraus folgt

$$\boldsymbol{A}\mathcal{K}_m(\boldsymbol{q}, \boldsymbol{A}) \subseteqq \mathcal{K}_m(\boldsymbol{q}, \boldsymbol{A}).$$

Damit ist $\mathcal{K}_m(\boldsymbol{q}, \boldsymbol{A})$ ein $\boldsymbol{A}$-invarianter Teilraum des $\mathbb{R}^n$.
Die Idee des LANCZOS-Algorithmus besteht nun darin, die Abbildung

$$\Phi : \mathcal{K}_m(\boldsymbol{q}, \boldsymbol{A}) \to \mathcal{K}_m(\boldsymbol{q}, \boldsymbol{A}), \quad \Phi(\boldsymbol{x}) = \boldsymbol{A}\boldsymbol{x}$$

bezüglich einer speziellen orthonormalen Basis $\{\boldsymbol{q}_1, \ldots, \boldsymbol{q}_m\}$ von $\mathcal{K}_m(\boldsymbol{q}, \boldsymbol{A})$ zu beschreiben. Die Vektoren $\boldsymbol{q}_1, \ldots, \boldsymbol{q}_m$ werden so gewählt, dass $\{\boldsymbol{q}_1, \ldots, \boldsymbol{q}_i\}$ jeweils eine orthonormale Basis des $\mathcal{K}_i(\boldsymbol{q}, \boldsymbol{A})$ bildet. Diese Vektoren lassen sich leicht berechnen:

9.46. LANCZOS**-Algorithmus:**

S0 *(Initialisierung) Wähle Vektor* $\boldsymbol{q} \in \mathbb{R}^n$ *mit* $\|\boldsymbol{q}\|_2 = 1$. *Setze* $\boldsymbol{q}_1 = \boldsymbol{q}$ *und* $k = 1$, $b_1\boldsymbol{q}_0 = \boldsymbol{o}$.

S1 *(Iteration) Berechne*

$$\begin{aligned} a_k &= \boldsymbol{q}_k^T \boldsymbol{A}\boldsymbol{q}_k, \\ \boldsymbol{r}_k &= \boldsymbol{A}\boldsymbol{q}_k - a_k\boldsymbol{q}_k - b_k\boldsymbol{q}_{k-1}, \\ b_{k+1} &= \|\boldsymbol{r}_k\|_2. \end{aligned}$$

S2 *(Abbruchtest)*

- *Falls* $b_{k+1} \neq 0$, *so berechne* $\boldsymbol{q}_{k+1} = \boldsymbol{r}_k/b_{k+1}$, *setze* $k = k+1$ *und gehe zu Schritt* **S1**.
- *Falls* $b_{k+1} = 0$, *so setze* $m = k$. *STOPP*

Aufwand pro Schritt:

- *Eine Auswertung Matrix* $\times$ *Vektor*
- $\sim 4n$ *Add./Sub. +* $\sim 5n$ *Mult./Div. + 1 Quadratwurzel.*

Die durch den LANCZOS-Algorithmus erzeugten Vektoren bilden die gesuchte Orthonormalbasis des $\mathcal{K}_m(\boldsymbol{q}, \boldsymbol{A})$, wie der folgende Satz aussagt.

9.47. Satz: *Die nach dem* LANCZOS*-Algorithmus berechneten Vektoren sind paarweise orthogonal und es gilt*

$$\text{span}(\boldsymbol{q}_1, \ldots, \boldsymbol{q}_i) = \mathcal{K}_i(\boldsymbol{q}, \boldsymbol{A}), \quad i = 1, \ldots, m.$$

Für den Abbruchindex gilt

$$m = \max\left\{ \dim(\mathcal{K}_i(\boldsymbol{q}, \boldsymbol{A})) \mid 1 \leqq i \leqq n \right\}.$$

Beweis: Wir beweisen den Satz mittels vollständiger Induktion über i. Zunächst stellen wir fest, dass für $i = 1, \ldots, m-1$

$$\boldsymbol{A}\boldsymbol{q}_i = b_i\boldsymbol{q}_{i-1} + a_i\boldsymbol{q}_i + b_{i+1}\boldsymbol{q}_{i+1}$$

gilt. Für $i = 1$ ist $\boldsymbol{q}_1 = \boldsymbol{q}$ offensichtlich eine orthonormale Basis des $\mathcal{K}_1(\boldsymbol{q}, \boldsymbol{A})$. Nehmen wir nun an, dass für alle $i \leqq j$ die Vektoren $\boldsymbol{q}_1, \ldots, \boldsymbol{q}_i$ paarweise orthogonal sind, und dass

$$\text{span}(\boldsymbol{q}_1, \ldots, \boldsymbol{q}_i) = \mathcal{K}_i(\boldsymbol{q}, \boldsymbol{A})$$

gilt. Ist nun der berechnete Vektor $\boldsymbol{r}_j \neq \boldsymbol{o}$, so ist $b_{j+1} \neq 0$. Der Vektor $\boldsymbol{q}_{j+1}$ ist eindeutig definiert, und es gilt $\|\boldsymbol{q}_{j+1}\|_2 = 1$. Um die Orthogonalitätsbeziehungen zu beweisen, sind drei Fälle zu unterscheiden.

(i) Für $1 \leqq i \leqq j-2$ gilt

$$\begin{aligned} b_{j+1}\boldsymbol{q}_i^T\boldsymbol{q}_{j+1} &= \boldsymbol{q}_i^T\boldsymbol{r}_j \\ &= \boldsymbol{q}_i^T(\boldsymbol{A}\boldsymbol{q}_j - a_j\boldsymbol{q}_j - b_j\boldsymbol{q}_{j-1}) \\ &= \boldsymbol{q}_i^T\boldsymbol{A}\boldsymbol{q}_j \\ &= (\boldsymbol{A}\boldsymbol{q}_i)^T\boldsymbol{q}_j \\ &= (b_i\boldsymbol{q}_{i-1} + a_i\boldsymbol{q}_i + b_{i+1}\boldsymbol{q}_{i+1})^T\boldsymbol{q}_j \\ &= 0. \end{aligned}$$

(ii) Für $i = j-1$ gilt

$$\begin{aligned} b_{j+1}\boldsymbol{q}_{j-1}^T\boldsymbol{q}_{j+1} &= \boldsymbol{q}_{j-1}^T\boldsymbol{r}_j \\ &= \boldsymbol{q}_{j-1}^T(\boldsymbol{A}\boldsymbol{q}_j - a_j\boldsymbol{q}_j - b_j\boldsymbol{q}_{j-1}) \\ &= \boldsymbol{q}_{j-1}^T\boldsymbol{A}\boldsymbol{q}_j - b_j\boldsymbol{q}_{j-1}^T\boldsymbol{q}_{j-1} \\ &= (\boldsymbol{A}\boldsymbol{q}_{j-1})^T\boldsymbol{q}_j - b_j \\ &= (b_{j-1}\boldsymbol{q}_{j-2} + a_{j-1}\boldsymbol{q}_{j-1} + b_j\boldsymbol{q}_j)^T\boldsymbol{q}_j - b_j \\ &= b_j - b_j \\ &= 0. \end{aligned}$$

(iii) Für $i = j$ gilt

$$\begin{aligned} b_{j+1}\boldsymbol{q}_j^T\boldsymbol{q}_{j+1} &= \boldsymbol{q}_j^T\boldsymbol{r}_j \\ &= \boldsymbol{q}_j^T(\boldsymbol{A}\boldsymbol{q}_j - a_j\boldsymbol{q}_j - b_j\boldsymbol{q}_{j-1}) \\ &= \boldsymbol{q}_j^T\boldsymbol{A}\boldsymbol{q}_j - a_j\boldsymbol{q}_j^T\boldsymbol{q}_j \\ &= a_j - a_j \\ &= 0. \end{aligned}$$

Wegen

$$\text{span}(\boldsymbol{q}_1, \ldots, \boldsymbol{q}_i) = \boldsymbol{\mathcal{K}}_i(\boldsymbol{q}, \boldsymbol{A}) \subseteqq \boldsymbol{\mathcal{K}}_j(\boldsymbol{q}, \boldsymbol{A})$$

für $i \leqq j$ gilt außerdem $\boldsymbol{A}\boldsymbol{q}_j \in \boldsymbol{\mathcal{K}}_{j+1}(\boldsymbol{q}, \boldsymbol{A})$. Da

$$\boldsymbol{q}_{j+1} \in \text{span}(\boldsymbol{q}_{j-1}, \boldsymbol{q}_j, \boldsymbol{A}\boldsymbol{q}_j)$$

folgt $\boldsymbol{q}_{j+1} \in \boldsymbol{\mathcal{K}}_{j+1}(\boldsymbol{q}, \boldsymbol{A})$ und weiter

$$\text{span}(\boldsymbol{q}_1, \ldots, \boldsymbol{q}_{j+1}) \subseteqq \boldsymbol{\mathcal{K}}_{j+1}(\boldsymbol{q}, \boldsymbol{A}).$$

Da die Vektoren $\boldsymbol{q}_1, \ldots, \boldsymbol{q}_{j+1}$ wegen ihrer paarweisen Orthogonalität linear unabhängig sind, gilt

$$\text{span}(\boldsymbol{q}_1, \ldots, \boldsymbol{q}_{j+1}) = \boldsymbol{\mathcal{K}}_{j+1}(\boldsymbol{q}, \boldsymbol{A}).$$

Damit folgt sofort $j+1 \leqq \max\left\{ \dim(\boldsymbol{\mathcal{K}}_i(\boldsymbol{q}, \boldsymbol{A})) \mid 1 \leqq i \leqq n \right\}$ und für den Abbruchindex m ebenso

$$m \leqq \max\left\{ \dim(\boldsymbol{\mathcal{K}}_i(\boldsymbol{q}, \boldsymbol{A})) \mid 1 \leqq i \leqq n \right\}.$$

Andererseits gilt wegen der Abbruchbedingung

$$\boldsymbol{r}_m = \boldsymbol{A}\boldsymbol{q}_m - a_m \boldsymbol{q}_m - b_m \boldsymbol{q}_{m-1} = \boldsymbol{o},$$

daher

$$\boldsymbol{A}\boldsymbol{q}_m \in \text{span}(\boldsymbol{q}_{m-1}, \boldsymbol{q}_m) \subset \text{span}(\boldsymbol{q}_1, \ldots, \boldsymbol{q}_m) = \boldsymbol{\mathcal{K}}_m(\boldsymbol{q}, \boldsymbol{A}).$$

Für $i < m$ gilt

$$\boldsymbol{A}\boldsymbol{q}_i \in \boldsymbol{\mathcal{K}}_{i+1}(\boldsymbol{q}, \boldsymbol{A}) \subseteqq \boldsymbol{\mathcal{K}}_m(\boldsymbol{q}, \boldsymbol{A}).$$

Damit ist aber

$$\boldsymbol{A}\boldsymbol{\mathcal{K}}_m(\boldsymbol{q}, \boldsymbol{A}) \subseteqq \boldsymbol{\mathcal{K}}_m(\boldsymbol{q}, \boldsymbol{A}),$$

$\boldsymbol{\mathcal{K}}_m(\boldsymbol{q}, \boldsymbol{A})$ ist also $\boldsymbol{A}$-invarianter Teilraum; der kleinste unter ihnen ist jedoch durch die Bedingung

$$j = \max\left\{ \dim(\boldsymbol{\mathcal{K}}_i(\boldsymbol{q}, \boldsymbol{A})) \mid 1 \leqq i \leqq n \right\}$$

festgelegt. Damit gilt für den Abbruchindex

$$m \geqq \max\left\{ \dim(\boldsymbol{\mathcal{K}}_i(\boldsymbol{q}, \boldsymbol{A})) \mid 1 \leqq i \leqq n \right\}.$$

Zusammengefasst ergibt sich

$$m = \max\left\{ \dim(\boldsymbol{\mathcal{K}}_i(\boldsymbol{q}, \boldsymbol{A})) \mid 1 \leqq i \leqq n \right\}.$$

❀

Bemerkungen: (i) Die durch den LANCZOS-Algorithmus berechneten Vektoren genügen dem folgenden Gleichungssystem:

$$\begin{aligned}
\boldsymbol{A}\boldsymbol{q}_1 &= a_1\boldsymbol{q}_1 + b_2\boldsymbol{q}_2 \\
\boldsymbol{A}\boldsymbol{q}_2 &= b_2\boldsymbol{q}_1 + a_2\boldsymbol{q}_2 + b_3\boldsymbol{q}_3 \\
\vdots &\qquad\quad \vdots \\
\boldsymbol{A}\boldsymbol{q}_{m-1} &= b_{m-1}\boldsymbol{q}_{m-2} + a_{m-1}\boldsymbol{q}_{m-1} + b_m\boldsymbol{q}_m \\
\boldsymbol{A}\boldsymbol{q}_m &= b_m\boldsymbol{q}_{m-1} + a_m\boldsymbol{q}_m.
\end{aligned}$$

Mit der spaltenorthonormalen (n,m)-Matrix

$$\boldsymbol{Q} = (\boldsymbol{q}_1, \boldsymbol{q}_2, \ldots, \boldsymbol{q}_{m-1}, \boldsymbol{q}_m)$$

und der (m,m)-Tridiagonalmatrix

$$\boldsymbol{J} = \text{trid}(a_1, \ldots, a_m, b_2, \ldots, b_m) = \begin{pmatrix} a_1 & b_2 & 0 & \ldots & 0 \\ b_2 & a_2 & b_3 & \ddots & \vdots \\ 0 & \ddots & \ddots & \ddots & 0 \\ \vdots & \ddots & b_{m-1} & a_{m-1} & b_m \\ 0 & \ldots & 0 & b_m & a_m \end{pmatrix}$$

gilt dann

$$\boldsymbol{A}\boldsymbol{Q} = \boldsymbol{Q}\boldsymbol{J}.$$

Im Falle $m = n$ hat man mit diesem Verfahren die Matrix $\boldsymbol{A}$ auf symmetrische Tridiagonalform gebracht. Die Eigenwerte von $\boldsymbol{J}$ sind auch Eigenwerte von $\boldsymbol{A}$. Ist $m < n$, so lässt sich durch Wahl eines neuen Startvektors $\overline{\boldsymbol{q}}$ mit $\boldsymbol{Q}^T\overline{\boldsymbol{q}} = \boldsymbol{o}$ eine weitere spaltenorthonormale Matrix $\overline{\boldsymbol{Q}}$ und eine Tridiagonalmatrix $\overline{\boldsymbol{J}}$ berechnen, die dann weitere Eigenwerte von $\boldsymbol{A}$ liefern.
(ii) Der Vorteil des LANCZOS-Algorithmus besteht darin, dass die größten bzw. kleinsten Eigenwerte der Matrizen

$$\boldsymbol{J}_i = \text{trid}(a_1, \ldots, a_i, b_2, \ldots, b_i) = \begin{pmatrix} a_1 & b_2 & 0 & \ldots & 0 \\ b_2 & a_2 & b_3 & \ddots & \vdots \\ 0 & \ddots & \ddots & \ddots & 0 \\ \vdots & \ddots & b_{i-1} & a_{i-1} & b_i \\ 0 & \ldots & 0 & b_i & a_i \end{pmatrix}$$

schnell gegen die größten bzw. kleinsten Eigenwerte von $\boldsymbol{A}$ konvergieren. Diese lassen sich effektiv mit der Vektoriteration bzw. der Inversen Vektoriteration bestimmen. Benötigt man alle Eigenwerte, so wendet man auf die Tridiagonalmatrix den QR-Algorithmus an, der im nächsten Abschnitt beschrieben wird.
(iii) Dieser Algorithmus ist für große schwach besetzte Matrizen geeignet, da der wesentliche Aufwand in der Operation Matrix$\times$Vektor liegt.

9.8. Der QR-Algorithmus

Der QR-Algorithmus, der 1961 unabhängig voneinander von FRANCIS und KUBLANOVSKAJA entwickelt wurde, ist das zur Zeit effektivste Verfahren zum Lösen des vollständigen Eigenwertproblems einer symmetrischen Tridiagonalmatrix. In seiner Grundform ist er auch auf beliebige symmetrische Matrizen anwendbar. Wir setzen darum vorerst nur die Symmetrie der Matrix $\boldsymbol{A}$ voraus.

9.48. QR**-Algorithmus ohne Verschiebungen:**

S0 *(Initialisierung) Setze*

$$\boldsymbol{A}^{(0)} = \boldsymbol{A}, \quad \boldsymbol{V}^{(0)} = \boldsymbol{I}, \quad k = 0.$$

S1 *(QR-Zerlegung) Berechne eine orthogonale (n,n)-Matrix $\boldsymbol{Q}^{(k)}$ und eine obere Dreiecksmatrix $\boldsymbol{R}^{(k)}$ mit*

$$\boldsymbol{A}^{(k)} = \boldsymbol{Q}^{(k)}\boldsymbol{R}^{(k)}.$$

S2 *(Iteration) Berechne*

$$\boldsymbol{A}^{(k+1)} = \boldsymbol{R}^{(k)}\boldsymbol{Q}^{(k)}, \quad \boldsymbol{V}^{(k+1)} = \boldsymbol{V}^{(k)}\boldsymbol{Q}^{(k)}.$$

S3 *Setze $k = k+1$ und gehe zu Schritt* **S1.**

9.49. Satz: *Die vom obigen Algorithmus erzeugten Matrizen $\boldsymbol{A}^{(k)}, \boldsymbol{Q}^{(k)}, \boldsymbol{R}^{(k)}$ und $\boldsymbol{V}^{(k)}$ sowie die Matrix*

$$\boldsymbol{U}^{(k+1)} = \boldsymbol{R}^{(k)}\boldsymbol{R}^{(k-1)}\cdots\boldsymbol{R}^{(0)}$$

haben folgende Eigenschaften:

(i) $\boldsymbol{A}^{(k+1)}$ ist orthogonal ähnlich zu $\boldsymbol{A}^{(k)}$:

$$\boldsymbol{A}^{(k+1)} = {\boldsymbol{Q}^{(k)}}^T \boldsymbol{A}^{(k)}\boldsymbol{Q}^{(k)}, \quad k = 0,1,\ldots$$

(ii) Es gilt $\boldsymbol{A}^{(k)} = {\boldsymbol{V}^{(k)}}^T \boldsymbol{A}\boldsymbol{V}^{(k)}$ für $k = 0,1,\ldots$

(iii) Es gilt $\boldsymbol{A}^k = \boldsymbol{V}^{(k)}\boldsymbol{U}^{(k)}$ für $k = 1,2,\ldots$

Beweis: (i) Aus $\boldsymbol{A}^{(k)} = \boldsymbol{Q}^{(k)}\boldsymbol{R}^{(k)}$ folgt $\boldsymbol{R}^{(k)} = {\boldsymbol{Q}^{(k)}}^T \boldsymbol{A}^{(k)}$ und damit

$$\boldsymbol{A}^{(k+1)} = \boldsymbol{R}^{(k)}\boldsymbol{Q}^{(k)} = {\boldsymbol{Q}^{(k)}}^T \boldsymbol{A}^{(k)}\boldsymbol{Q}^{(k)}.$$

(ii) Die Aussage ergibt sich durch wiederholte Anwendung von (i).
(iii) Wir zeigen die Aussage mittels vollständiger Induktion. Zunächst gilt wegen $\boldsymbol{A}^{(k)} = \boldsymbol{V}^{(k)^T} \boldsymbol{A} \boldsymbol{V}^{(k)}$

$$\boldsymbol{V}^{(k)} \boldsymbol{A}^{(k)} = \boldsymbol{A} \boldsymbol{V}^{(k)}, \quad k = 0, 1, \ldots.$$

Offensichtlich gilt $\boldsymbol{A}^1 = \boldsymbol{A} = \boldsymbol{A}^{(0)} = \boldsymbol{Q}^{(0)} \boldsymbol{R}^{(0)} = \boldsymbol{V}^{(1)} \boldsymbol{U}^{(1)}$. Damit ist der Induktionsanfang gesichert. Wir nehmen nun an, dass für alle $j \leqq k$

$$\boldsymbol{A}^j = \boldsymbol{V}^{(j)} \boldsymbol{U}^{(j)}$$

gilt. Dann folgt

$$\begin{aligned} \boldsymbol{A}^{k+1} &= \boldsymbol{A} \boldsymbol{A}^k = \boldsymbol{A} \boldsymbol{V}^{(k)} \boldsymbol{U}^{(k)} = \boldsymbol{V}^{(k)} \boldsymbol{A}^{(k)} \boldsymbol{U}^{(k)} \\ &= \boldsymbol{V}^{(k)} \boldsymbol{Q}^{(k)} \boldsymbol{R}^{(k)} \boldsymbol{U}^{(k)} = \boldsymbol{V}^{(k+1)} \boldsymbol{U}^{(k+1)}. \end{aligned}$$

❀

Der QR-Algorithmus ist eng mit der Teilraumiteration verwandt. Um dies zu zeigen, benötigen wir noch folgenden Satz.

9.50. Satz: *Es sei $\boldsymbol{A}$ eine reguläre (n, n)-Matrix. Durch $\boldsymbol{A} = \boldsymbol{R}$ und $\boldsymbol{A} = \bar{\boldsymbol{Q}} \bar{\boldsymbol{R}}$ seien zwei Zerlegungen mit orthogonalen Matrizen $\boldsymbol{Q}$ und $\bar{\boldsymbol{Q}}$ und oberen Dreiecksmatrizen $\boldsymbol{R}$ und $\bar{\boldsymbol{R}}$ gegeben. Dann gilt*

$$\boldsymbol{Q} = \bar{\boldsymbol{Q}} \boldsymbol{D}, \quad \boldsymbol{R} = \boldsymbol{D} \bar{\boldsymbol{R}}$$

mit einer Diagonalmatrix $\boldsymbol{D}$, für die

$$|\boldsymbol{D}| = \boldsymbol{I}$$

gilt.

Beweis: Aus $\boldsymbol{R} = \bar{\boldsymbol{Q}} \bar{\boldsymbol{R}}$ folgt $\bar{\boldsymbol{Q}}^T \boldsymbol{Q} = \bar{\boldsymbol{R}} \boldsymbol{R}^{-1} = \boldsymbol{B}$. Die Matrix $\boldsymbol{B}$ ist damit gleichzeitig obere Dreiecksmatrix und Orthogonalmatrix; daher ist $\boldsymbol{B}$ eine Diagonalmatrix, deren Diagonalelemente sämtlich $+1$ oder -1 sind. ❀

Damit lässt sich jetzt der Zusammenhang zwischen dem QR-Algorithmus und der Teilraumiteration zeigen. Starten wir die Teilraumiteration mit der (n, n)-Matrix $\bar{\boldsymbol{V}}^{(0)}$, so hat man im k-ten Schritt $\bar{\boldsymbol{W}}^{(k+1)} = \boldsymbol{A} \bar{\boldsymbol{V}}^{(k)}$ zu berechnen und anschließend die QR-Zerlegung $\bar{\boldsymbol{W}}^{(k+1)} = \bar{\boldsymbol{V}}^{(k+1)} \bar{\boldsymbol{R}}^{(k+1)}$ durchzuführen. (Um Verwechslungen mit den Matrizen aus dem QR-Algorithmus zu vermeiden, sind die hier auftretenden

Matrizen mit einem oberen Querstrich gekennzeichnet.) Führen wir nun die Matrizen $\bar{\boldsymbol{A}}^{(k)}$ gemäß $\bar{\boldsymbol{A}}^{(k)} = \bar{\boldsymbol{V}}^{(k)^T} \boldsymbol{A} \bar{\boldsymbol{V}}^{(k)}$ ein, so gilt für diese

$$\begin{aligned} \bar{\boldsymbol{A}}^{(k)} &= \bar{\boldsymbol{V}}^{(k)^T} \bar{\boldsymbol{W}}^{(k+1)} \\ &= \bar{\boldsymbol{V}}^{(k)^T} \bar{\boldsymbol{V}}^{(k+1)} \bar{\boldsymbol{R}}^{(k+1)}. \end{aligned}$$

Durch die orthogonale Matrix $\bar{\boldsymbol{Q}}^{(k)} = \bar{\boldsymbol{V}}^{(k)^T} \bar{\boldsymbol{V}}^{(k+1)}$ und die obere Dreiecksmatrix $\bar{\boldsymbol{R}}^{(k+1)}$ ist eine QR-Zerlegung von $\bar{\boldsymbol{A}}^{(k)}$ gegeben. Für die Matrix $\bar{\boldsymbol{A}}^{(k+1)}$ erhält man

$$\begin{aligned} \bar{\boldsymbol{A}}^{(k+1)} &= \bar{\boldsymbol{V}}^{(k+1)^T} \boldsymbol{A} \bar{\boldsymbol{V}}^{(k+1)} \\ &= \bar{\boldsymbol{V}}^{(k+1)^T} \bar{\boldsymbol{V}}^{(k)} \bar{\boldsymbol{A}}^{(k)} \bar{\boldsymbol{V}}^{(k)^T} \bar{\boldsymbol{V}}^{(k+1)} \\ &= \bar{\boldsymbol{R}}^{(k+1)} \bar{\boldsymbol{V}}^{(k)^T} \bar{\boldsymbol{V}}^{(k+1)} \\ &= \bar{\boldsymbol{R}}^{(k+1)} \bar{\boldsymbol{Q}}^{(k)}. \end{aligned}$$

Die Matrix $\bar{\boldsymbol{A}}^{(k+1)}$ entsteht, indem die Faktoren der QR-Zerlegung von $\bar{\boldsymbol{A}}^{(k)}$ in umgekehrter Reihenfolge multipliziert werden. Das entspricht genau der Vorgehensweise des QR-Algorithmus. Die QR-Zerlegung einer Matrix ist bis auf die Vorzeichenverteilung auf die Spalten von $\boldsymbol{Q}$ und die Zeilen von $\boldsymbol{R}$ eindeutig. Es ist daher zu erwarten, dass die Folgen $\{\bar{\boldsymbol{V}}^{(k)}\}$, $\{\bar{\boldsymbol{R}}^{(k)}\}$, $\{\bar{\boldsymbol{A}}^{(k)}\}$ und $\{\bar{\boldsymbol{Q}}^{(k)}\}$, die von der Teilraumiteration erzeugt werden, eng mit den Folgen $\{\boldsymbol{V}^{(k)}\}$, $\{\boldsymbol{R}^{(k)}\}$, $\{\boldsymbol{A}^{(k)}\}$ und $\{\boldsymbol{Q}^{(k)}\}$, die der QR-Algorithmus liefert, zusammenhängen. Es gilt der folgende Satz.

9.51. Satz: *Für die reguläre symmetrische (n,n)-Matrix $\boldsymbol{A}$ seien die Folgen $\{\bar{\boldsymbol{V}}^{(k)}\}$ und $\{\bar{\boldsymbol{R}}^{(k)}\}$ durch die Teilraumiteration mit $\bar{\boldsymbol{V}}^{(0)} = \boldsymbol{I}$ erzeugt. Die Matrizen $\bar{\boldsymbol{A}}^{(k)}$ und $\bar{\boldsymbol{Q}}^{(k)}$ seien wie oben definiert. Der QR-Algorithmus liefere die Folgen $\left\{ \boldsymbol{V}^{(k)} \right\}$, $\left\{ \boldsymbol{R}^{(k)} \right\}$, $\left\{ \boldsymbol{A}^{(k)} \right\}$ und $\left\{ \boldsymbol{Q}^{(k)} \right\}$.*
Dann existieren (n,n)-Diagonalmatrizen $\boldsymbol{D}^{(k)}$ mit $\boldsymbol{D}^{(0)} = \boldsymbol{I}$ und $|\boldsymbol{D}^{(k)}| = \boldsymbol{I}$ für $k = 1, 2, \ldots$, so dass

$$\bar{\boldsymbol{V}}^{(k)} = \boldsymbol{V}^{(k)} \boldsymbol{D}^{(k)}$$

und

$$\begin{aligned} \bar{\boldsymbol{A}}^{(k)} &= \boldsymbol{D}^{(k)} \boldsymbol{A}^{(k)} \boldsymbol{D}^{(k)}, \quad \bar{\boldsymbol{Q}}^{(k)} = \boldsymbol{D}^{(k)} \boldsymbol{Q}^{(k)} \boldsymbol{D}^{(k+1)} \quad \textit{und} \quad \bar{\boldsymbol{R}}^{(k+1)} \\ &= \boldsymbol{D}^{(k+1)} \boldsymbol{R}^{(k)} \boldsymbol{D}^{(k)} \end{aligned}$$

gilt.

Beweis: Wir beweisen den Satz mittels vollständiger Induktion. Der Induktionsanfang für $k = 0$ ist offensichtlich. Aus der Annahme, dass die Aussagen bis zu einem Index k gelten, folgt dann

$$\begin{aligned}\bar{\boldsymbol{V}}^{(k+1)} &= \bar{\boldsymbol{V}}^{(k)}\bar{\boldsymbol{Q}}^{(k)} \\ &= \left(\boldsymbol{V}^{(k)}\boldsymbol{D}^{(k)}\right)\left(\boldsymbol{D}^{(k)}\boldsymbol{Q}^{(k)}\boldsymbol{D}^{(k+1)}\right) \\ &= \boldsymbol{V}^{(k)}\boldsymbol{Q}^{(k)}\boldsymbol{D}^{(k+1)} \\ &= \boldsymbol{V}^{(k+1)}\boldsymbol{D}^{(k+1)}.\end{aligned}$$

Weiter ergibt sich

$$\begin{aligned}\bar{\boldsymbol{A}}^{(k+1)} &= \bar{\boldsymbol{Q}}^{(k)^T}\bar{\boldsymbol{A}}^{(k)}\bar{\boldsymbol{Q}}^{(k)} \\ &= \left(\boldsymbol{D}^{(k+1)}\boldsymbol{Q}^{(k)^T}\boldsymbol{D}^{(k)}\right)\left(\boldsymbol{D}^{(k)}\boldsymbol{A}^{(k)}\boldsymbol{D}^{(k)}\right)\left(\boldsymbol{D}^{(k)}\boldsymbol{Q}^{(k)}\boldsymbol{D}^{(k+1)}\right) \\ &= \boldsymbol{D}^{(k+1)}\boldsymbol{A}^{(k+1)}\boldsymbol{D}^{(k+1)}\end{aligned}$$

und

$$\begin{aligned}\bar{\boldsymbol{Q}}^{(k+1)}\bar{\boldsymbol{R}}^{(k+2)} &= \bar{\boldsymbol{A}}^{(k+1)} \\ &= \boldsymbol{D}^{(k+1)}\boldsymbol{A}^{(k+1)}\boldsymbol{D}^{(k+1)} \\ &= \left(\boldsymbol{D}^{(k+1)}\boldsymbol{Q}^{(k+1)}\right)\left(\boldsymbol{R}^{(k+1)}\boldsymbol{D}^{(k+1)}\right).\end{aligned}$$

Da die QR-Zerlegung einer regulären Matrix bis auf die Verteilung der Vorzeichen auf die Spalten von $\boldsymbol{Q}$ und die Zeilen von $\boldsymbol{R}$ eindeutig ist, existiert eine Diagonalmatrix $\boldsymbol{D}^{(k+2)}$ mit $|\boldsymbol{D}^{(k+2)}| = \boldsymbol{I}$, so dass

$$\bar{\boldsymbol{Q}}^{(k+1)} = \left(\boldsymbol{D}^{(k+1)}\boldsymbol{Q}^{(k+1)}\right)\boldsymbol{D}^{(k+2)}, \quad \bar{\boldsymbol{R}}^{(k+2)} = \boldsymbol{D}^{(k+2)}\left(\boldsymbol{R}^{(k+1)}\boldsymbol{D}^{(k+1)}\right).$$

Damit ist der Satz bewiesen. ❀

Bemerkung: Wird in den Algorithmen bei der QR-Zerlegung jeweils $(\boldsymbol{R})_{jj} > 0$ für $j = 1, \ldots, n$ gefordert, so gilt immer $\boldsymbol{D}^{(k)} = \boldsymbol{I}$. Dann stimmen die Matrizenfolgen beider Algorithmen exakt überein.
Da sich die Matrizen $\boldsymbol{V}^{(k)}$ und $\bar{\boldsymbol{V}}^{(k)}$ nur in den Vorzeichen der Spalten unterscheiden, gilt für eine Partitionierung

$$\boldsymbol{V}^{(k)} = \left(\boldsymbol{V}_1^{(k)}, \boldsymbol{V}_2^{(k)}\right), \quad \boldsymbol{V}_1^{(k)} \in \mathbb{R}^{n\times m}$$

und eine entsprechende Partitionierung

$$\bar{\boldsymbol{V}}^{(k)} = \left(\bar{\boldsymbol{V}}_1^{(k)}, \bar{\boldsymbol{V}}_2^{(k)}\right), \quad \bar{\boldsymbol{V}}_1^{(k)} \in \mathbb{R}^{n\times m}$$

$$\boldsymbol{\mathcal{Y}}_1 = \operatorname{span}\left(\boldsymbol{V}_1^{(k)}\right) = \bar{\boldsymbol{\mathcal{Y}}}_1 = \operatorname{span}\left(\bar{\boldsymbol{V}}_1^{(k)}\right).$$

Die Algorithmen erzeugen die gleichen Teilraumfolgen. Damit lassen sich alle Konvergenzaussagen der Teilraumiteration aus Satz 9.35 auf den QR-Algorithmus übertragen. Es gilt der folgende Satz.

9.52. Satz: *Für die symmetrische (n,n)-Matrix $\boldsymbol{A}$ mit den Eigenwerten*

$$0 < |\lambda_n| \leqq |\lambda_{n-1}| \leqq \cdots \leqq |\lambda_{p+1}| < |\lambda_p| \leqq \cdots \leqq |\lambda_2| \leqq |\lambda_1|,$$

und zugehörigen orthonormierten Eigenvektoren

$$\boldsymbol{u}_1, \boldsymbol{u}_2, \ldots, \boldsymbol{u}_n$$

werde der QR-Algorithmus in exakter Arithmetik durchgeführt. Mit

$$\boldsymbol{\mathcal{Y}}^{(0)} = \operatorname{span}(\boldsymbol{e}_1, \ldots, \boldsymbol{e}_p), \quad \boldsymbol{\mathcal{S}}_p = \operatorname{span}(\boldsymbol{u}_1, \ldots, \boldsymbol{u}_p)$$

gelte

$$\sigma = \cos\left(\ \sphericalangle\left(\boldsymbol{\mathcal{Y}}^{(0)}, \boldsymbol{\mathcal{S}}_p\right)\right) > 0.$$

Die durch den QR-Algorithmus ohne Verschiebungen erzeugten Matrizen seien wie folgt partitioniert:

$$\boldsymbol{V}^{(k)} = \left(\boldsymbol{V}_1^{(k)}, \boldsymbol{V}_2^{(k)}\right), \quad \boldsymbol{V}_1^{(k)} \in \mathbb{R}^{n\times p}$$

und

$$\boldsymbol{A}^{(k)} = \begin{pmatrix} \boldsymbol{A}_{11}^{(k)} & \boldsymbol{A}_{12}^{(k)} \\ \boldsymbol{A}_{21}^{(k)} & \boldsymbol{A}_{22}^{(k)} \end{pmatrix}, \quad \boldsymbol{Q}^{(k)} = \begin{pmatrix} \boldsymbol{Q}_{11}^{(k)} & \boldsymbol{Q}_{12}^{(k)} \\ \boldsymbol{Q}_{21}^{(k)} & \boldsymbol{Q}_{22}^{(k)} \end{pmatrix}, \quad \boldsymbol{R}^{(k)} = \begin{pmatrix} \boldsymbol{R}_{11}^{(k)} & \boldsymbol{R}_{12}^{(k)} \\ \boldsymbol{o} & \boldsymbol{R}_{22}^{(k)} \end{pmatrix}$$

mit

$$\boldsymbol{A}_{11}^{(k)} \in \mathbb{R}^{p\times p}, \quad \boldsymbol{Q}_{11}^{(k)} \in \mathbb{R}^{p\times p}, \quad \boldsymbol{R}_{11}^{(k)} \in \mathbb{R}^{p\times p}.$$

$$\boldsymbol{\mathcal{Y}}^{(k)} = \operatorname{span}\left(\boldsymbol{V}_1^{(k)}\right).$$

Dann gilt

$$\tan\varphi_{k+1} \leqq \kappa \tan\varphi_k \leqq \kappa^{k+1}\tan\varphi_0 = \kappa^{k+1}\frac{\sqrt{1-\sigma^2}}{\sigma} = \varepsilon_{k+1} \tag{9.38}$$

mit

$$\kappa = \left|\frac{\lambda_{p+1}}{\lambda_p}\right|, \quad \varphi_k = \sphericalangle\left(\boldsymbol{\mathcal{Y}}^{(k)}, \boldsymbol{\mathcal{S}}_p\right) = \sphericalangle\left(\boldsymbol{\mathcal{Y}}^{(k)^\perp}, \boldsymbol{\mathcal{Z}}_{p+1}\right).$$

Dabei ist

$$\boldsymbol{\mathcal{Y}}^{(k)^\perp} = \operatorname{span}\left(\boldsymbol{V}_2^{(k)}\right), \quad \boldsymbol{\mathcal{Z}}_{p+1} = \operatorname{span}\left(\boldsymbol{u}_{p+1},\ldots,\boldsymbol{u}_n\right) = \boldsymbol{\mathcal{S}}_p^\perp.$$

Weiterhin gelten die Abschätzungen

$$\|\boldsymbol{A}_{12}^{(k)}\|_2 = \|\boldsymbol{A}_{21}^{(k)}\|_2 \leqq (1+\kappa)\varepsilon_k\|\boldsymbol{A}\|_2 \leqq 2\varepsilon_k\|\boldsymbol{A}\|_2, \tag{9.39}$$

$$\|\boldsymbol{Q}_{12}^{(k)}\|_2, \|\boldsymbol{Q}_{21}^{(k)}\|_2 \leqq (1+\kappa)\varepsilon_k \leqq 2\varepsilon_k, \tag{9.40}$$

$$\|\boldsymbol{R}_{12}^{(k)}\|_2 \leqq (1+\kappa)^2\varepsilon_k\|\boldsymbol{A}\|_2 \leqq 4\varepsilon_k\|\boldsymbol{A}\|_2. \tag{9.41}$$

Beweis: Die Gültigkeit von (9.38) folgt sofort aus Satz 9.35 falls man berücksichtigt, dass im Falle $\dim(\boldsymbol{\mathcal{Y}}) = \dim(\boldsymbol{\mathcal{S}})$ die Beziehung

$$\sphericalangle(\boldsymbol{\mathcal{Y}},\boldsymbol{\mathcal{S}}) = \sphericalangle(\boldsymbol{\mathcal{S}},\boldsymbol{\mathcal{Y}}) = \sphericalangle(\boldsymbol{\mathcal{Y}}^\perp,\boldsymbol{\mathcal{S}}^\perp)$$

gilt. Um (9.40) zu beweisen gehen wir von

$$\boldsymbol{Q}^{(k)} = \boldsymbol{V}^{(k)^T}\boldsymbol{V}^{(k+1)}$$

aus. Es gilt dann

$$\boldsymbol{Q}_{21}^{(k)} = \boldsymbol{V}_2^{(k)^T}\boldsymbol{V}_1^{(k+1)} = \left[\boldsymbol{U}^T\boldsymbol{V}_2^{(k)}\right]^T\left[\boldsymbol{U}^T\boldsymbol{V}_1^{(k+1)}\right]$$

mit $\boldsymbol{U} = \left(\boldsymbol{u}_1,\ldots,\boldsymbol{u}_p,\boldsymbol{u}_{p+1},\ldots,\boldsymbol{u}_n\right) = (\boldsymbol{U}_1,\boldsymbol{U}_2)$. Damit erhalten wir

$$\boldsymbol{U}^T\boldsymbol{V}_1^{(k+1)} = \begin{pmatrix} \boldsymbol{U}_1^T\boldsymbol{V}_1^{(k+1)} \\ \\ \boldsymbol{U}_2^T\boldsymbol{V}_1^{(k+1)} \end{pmatrix} = \begin{pmatrix} \boldsymbol{F} \\ \boldsymbol{G} \end{pmatrix}.$$

Wegen der Orthogonalität der Matrizen $\boldsymbol{U}$und $\boldsymbol{V}^{(k+1)}$ gilt

$$\|\boldsymbol{U}^T\boldsymbol{V}_1^{(k+1)}\|_2 = \|\boldsymbol{V}_1^{(k+1)}\|_2 = 1$$

und damit

$$\|\boldsymbol{F}\|_2 \leqq 1.$$

Es lässt sich zeigen, dass $\|\boldsymbol{G}\|_2 = \sin\varphi_{k+1}$. Damit ergibt sich $\|\boldsymbol{G}\|_2 \leqq \varepsilon_{k+1}$. Für

$$\boldsymbol{U}^T\boldsymbol{V}_2^{(k+1)} = \begin{pmatrix} \boldsymbol{U}_1^T\boldsymbol{V}_2^{(k+1)} \\ \\ \boldsymbol{U}_2^T\boldsymbol{V}_2^{(k+1)} \end{pmatrix} = \begin{pmatrix} \boldsymbol{G}' \\ \boldsymbol{F}' \end{pmatrix}$$

erhält man analog $\|\boldsymbol{F}'\|_2 \leqq 1$ und $\|\boldsymbol{G}'\|_2 \leqq \varepsilon_k$. Das liefert insgesamt

$$\begin{aligned} \|\boldsymbol{Q}_{21}^{(k)}\|_2 &= \|\boldsymbol{G}'^T\boldsymbol{F} + \boldsymbol{F}'^T\boldsymbol{G}\|_2 \\ &\leqq \|\boldsymbol{G}'\|_2\,\|\boldsymbol{F}\|_2 + \|\boldsymbol{F}'\|_2\,\|\boldsymbol{G}\|_2 \\ &\leqq \varepsilon_k + \varepsilon_{k+1} \\ &= (1+\kappa)\varepsilon_k. \end{aligned}$$

Dieselbe Abschätzung erhält man für $\boldsymbol{Q}_{12}$. Weiterhin gilt $\boldsymbol{A}^{(k)} = \boldsymbol{Q}^{(k)}\boldsymbol{R}^{(k)}$, und damit $\boldsymbol{A}_{21}^{(k)} = \boldsymbol{Q}_{21}^{(k)}\boldsymbol{R}_{11}^{(k)}$. Mit

$$\|\boldsymbol{R}_{11}^{(k)}\|_2 \leqq \|\boldsymbol{R}^{(k)}\|_2 = \|\boldsymbol{A}^{(k)}\|_2 = \|\boldsymbol{A}\|_2$$

folgt dann

$$\|\boldsymbol{A}_{21}^{(k)}\|_2 \leqq \|\boldsymbol{Q}_{21}^{(k)}\|_2\,\|\boldsymbol{R}_{11}^{(k)}\|_2 \leqq (1+\kappa)\varepsilon_k\|\boldsymbol{A}\|_2 \leqq 2\varepsilon_k\|\boldsymbol{A}\|_2.$$

Die letzte Abschätzung erhält man aus der Darstellung $\boldsymbol{R}^{(k)} = \boldsymbol{A}^{(k+1)}{\boldsymbol{Q}^{(k)}}^T$. Hieraus ergibt sich

$$\boldsymbol{R}_{12}^{(k)} = \boldsymbol{A}_{11}^{(k+1)}{\boldsymbol{Q}_{21}^{(k)}}^T + \boldsymbol{A}_{12}^{(k+1)}{\boldsymbol{Q}_{22}^{(k)}}^T.$$

Wegen $\|\boldsymbol{A}_{11}^{(k+1)}\|_2 \leqq \|\boldsymbol{A}^{(k+1)}\|_2 \leqq \|\boldsymbol{A}\|_2$ und $\|\boldsymbol{Q}_{22}^{(k)}\|_2 \leqq 1$ folgt

$$\begin{aligned} \|\boldsymbol{R}_{12}^{(k)}\|_2 &\leqq \|\boldsymbol{A}_{11}^{(k+1)}\|_2\,\|\boldsymbol{Q}_{21}^{(k)}\|_2 + \|\boldsymbol{A}_{12}^{(k+1)}\|_2\,\|\boldsymbol{Q}_{22}^{(k)}\|_2 \\ &\leqq \|\boldsymbol{A}\|_2(1+\kappa)\varepsilon_k + (1+\kappa)\varepsilon_{k+1}\|\boldsymbol{A}\|_2 \cdot 1 \\ &= (1+\kappa)(\varepsilon_k + \varepsilon_{k+1})\|\boldsymbol{A}\|_2 \\ &= (1+\kappa)^2\varepsilon_k\|\boldsymbol{A}\|_2 \\ &\leqq 4\varepsilon_k\|\boldsymbol{A}\|_2. \end{aligned}$$

❀

Bemerkungen: (i) Nach Ungleichung (9.38) konvergieren die Teilräume $\boldsymbol{\mathcal{Y}}^{(k)}$ bzw. $\boldsymbol{\mathcal{Y}}^{(k)^\perp}$ linear mit dem Konvergenzfaktor $\kappa < 1$ gegen $\boldsymbol{\mathcal{S}}_p$ bzw. $\boldsymbol{\mathcal{Z}}_{p+1}$. Im Sinne von $\tan\varphi_{k+1}/\tan\varphi_k < \kappa$ ist diese Konvergenz monoton. Nach (9.39)-(9.41) konvergieren die Nichtdiagonalblöcke

$$\boldsymbol{A}_{12}^{(k)}, \boldsymbol{A}_{21}^{(k)}, \boldsymbol{Q}_{12}^{(k)}, \boldsymbol{Q}_{21}^{(k)}$$

und

$$\boldsymbol{R}_{12}^{(k)} \text{ für } k \to \infty \text{ wie } \varepsilon_k$$

gegen $\boldsymbol{o}$.
(ii) In den meisten Fällen werden mehrere dominante Eigenwerte existieren. Es gilt dann zum Beispiel für zwei Indizes p und p'

$$0 < |\lambda_n| \leqq |\lambda_{n-1}| \leqq \cdots \leqq |\lambda_{p'+1}| < |\lambda_{p'}| \leqq \cdots \leqq |\lambda_{p+1}| < |\lambda_p| \leqq \cdots \leqq |\lambda_2| \leqq |\lambda_1|.$$

Die Aussagen des Satzes sind dann sowohl für p als auch für p' gültig falls

$$\cos\left(\measuredangle\left(\boldsymbol{\mathcal{Y}}^{(0)}, \boldsymbol{\mathcal{S}}_p\right)\right) > 0$$

und

$$\cos\left(\measuredangle\left(\boldsymbol{\mathcal{Y}}^{(0)}, \boldsymbol{\mathcal{S}}_{p'}\right)\right) > 0$$

gilt. Partitioniert man die Matrizen $\boldsymbol{A}^{(k)}$ gemäß

$$\boldsymbol{A}^{(k)} = \begin{pmatrix} \boldsymbol{A}_{11}^{(k)} & \boldsymbol{A}_{12}^{(k)} & \boldsymbol{A}_{13}^{(k)} \\ \boldsymbol{A}_{21}^{(k)} & \boldsymbol{A}_{22}^{(k)} & \boldsymbol{A}_{23}^{(k)} \\ \boldsymbol{A}_{31}^{(k)} & \boldsymbol{A}_{32}^{(k)} & \boldsymbol{A}_{33}^{(k)} \end{pmatrix} \text{ mit } \boldsymbol{A}_{11}^{(k)} \in \mathbb{R}^{p\times p}, \boldsymbol{A}_{22}^{(k)} \in \mathbb{R}^{(p'-p)\times(n-p')},$$

so folgt aus Satz 9.52

$$\|\boldsymbol{A}_{21}^{(k)}\|_2, \|\boldsymbol{A}_{31}^{(k)}\|_2 \leqq \left\| \begin{pmatrix} \boldsymbol{A}_{21}^{(k)} \\ \\ \boldsymbol{A}_{31}^{(k)} \end{pmatrix} \right\|_2 \leqq 2\varepsilon_k \|\boldsymbol{A}\|_2 \quad \text{und}$$

$$\|\boldsymbol{A}_{31}^{(k)}\|_2, \|\boldsymbol{A}_{32}^{(k)}\|_2 \leqq \left\| \left(\boldsymbol{A}_{31}^{(k)}, \boldsymbol{A}_{32}^{(k)}\right) \right\|_2 \leqq 2\varepsilon'_k \|\boldsymbol{A}\|_2.$$

Insgesamt gilt dann $\|\boldsymbol{A}_{31}^{(k)}\|_2 \leqq 2\min\{\varepsilon_k, \varepsilon_k'\}\|\boldsymbol{A}\|_2$.
(iii) Sind alle Eigenwerte dem Betrag nach paarweise verschieden, also

$$0 < |\lambda_n| < |\lambda_{n-1}| < \cdots < |\lambda_2| < |\lambda_1|,$$

und gilt für alle $p \in \{1, \ldots, n\}$ $\cos\left(\sphericalangle\left(\boldsymbol{\mathcal{Y}}^{(0)}, \boldsymbol{\mathcal{S}}_p\right)\right) > 0$, so konvergieren alle Nichtdiagonalelemente gegen Null. Die Diagonalelemente konvergieren dann gegen die Eigenwerte von $\boldsymbol{A}$.
Satz 9.52 zeigt, dass nach hinreichend vielen Schritten die Nichtdiagonalblöcke $\boldsymbol{A}_{12}^{(k)}$ und $\boldsymbol{A}_{21}^{(k)}$ klein werden. Für genügend großes k gilt also

$$\|\boldsymbol{A}_{12}^{(k)}\|_2 = \|\boldsymbol{A}_{21}^{(k)}\|_2 \leqq \varepsilon = \text{eps}F\|\boldsymbol{A}\|_2$$

mit einer Konstanten F, die den Rundungsfehlereinfluss im Algorithmus beschreibt. Vernachlässigt man unter diesen Voraussetzungen die Nichtdiagonalblöcke, so entspricht dies einem Übergang zur Matrix

$$\hat{\boldsymbol{A}}^{(k)} = \begin{pmatrix} \boldsymbol{A}_{11}^{(k)} & \boldsymbol{o} \\ \\ \boldsymbol{o} & \boldsymbol{A}_{22}^{(k)} \end{pmatrix}.$$

Es gilt dann

$$\hat{\boldsymbol{A}}^{(k)} = \boldsymbol{A}^{(k)} + \boldsymbol{\delta}\hat{\boldsymbol{A}}^{(k)}$$

mit der Störung

$$\boldsymbol{\delta}\hat{\boldsymbol{A}}^{(k)} = -\begin{pmatrix} \boldsymbol{o} & \boldsymbol{A}_{12}^{(k)} \\ \\ \boldsymbol{A}_{21}^{(k)} & \boldsymbol{o} \end{pmatrix}.$$

Für die Störung $\boldsymbol{\delta}\hat{\boldsymbol{A}}^{(k)}$ erhalten wir die Abschätzung

$$\|\boldsymbol{\delta}\hat{\boldsymbol{A}}^{(k)}\|_2 \leqq \varepsilon.$$

Die Eigenwerte der Matrix $\hat{\boldsymbol{A}}^{(k)}$ weichen höchstens um den Betrag ε von den Eigenwerten der Matrix $\boldsymbol{A}^{(k)}$ und damit von den Eigenwerten von $\boldsymbol{A}$ ab. Das Eigenwertproblem für $\hat{\boldsymbol{A}}^{(k)}$ zerfällt in zwei Eigenwertprobleme geringerer Dimension, nämlich die für $\boldsymbol{A}_{11}^{(k)} \in \mathbb{R}^{p\times p}$ und für $\boldsymbol{A}_{22}^{(k)} \in \mathbb{R}^{(n-p)\times(n-p)}$. Der Übergang von $\boldsymbol{A}^{(k)}$ zu $\hat{\boldsymbol{A}}^{(k)}$

ist damit mit einer wesentlichen Aufwandsverminderung verbunden. Dieses Vorgehen bezeichnet man als Deflation. Sind nach weiteren Schritten die Nichtdiagonalblöcke für die Matrizen $\boldsymbol{A}_{11}^{(k)}$ oder $\boldsymbol{A}_{22}^{(k)}$ wieder hinreichend klein, so zerfällt das Eigenwertproblem in noch kleinere Teilprobleme. So wird sich durch fortwährende Deflation der Gesamtaufwand erheblich verringern. Der Satz 9.52 zeigte, dass die Konvergenzgeschwindigkeit, mit der die Nichtdiagonalblöcke gegen Null streben, durch den Faktor $\kappa = |\lambda_{p+1}/\lambda_p|$ bestimmt wird. Wie bei der Inversen Iteration kann man nun wieder versuchen, diesen Konvergenzfaktor durch eine geeignete Spektralverschiebung zu verkleinern. Für ein vorgegebenes $\mu \in \mathbb{R}$ definieren wir die Matrix

$$\bar{\boldsymbol{A}} = \boldsymbol{A} - \mu \boldsymbol{I}.$$

Weiter nehmen wir an, dass die Eigenwerte $\lambda_i - \mu$ von $\bar{\boldsymbol{A}}$ so geordnet sind, dass

$$0 < |\lambda_1 - \mu| \ll |\lambda_2 - \mu| \leqq \cdots \leqq |\lambda_n - \mu|$$

gilt; μ ist eine gute Näherung für einen einfachen Eigenwert von $\boldsymbol{A}$, stimmt aber mit diesem nicht überein. Wir bezeichnen die Eigenwerte der Matrix $\bar{\boldsymbol{A}}$ mit

$$\lambda_n[\bar{\boldsymbol{A}}] = \bar{\lambda}_n = \lambda_1 - \mu, \quad \lambda_{n-1}[\bar{\boldsymbol{A}}] = \bar{\lambda}_{n-1} = \lambda_2 - \mu, \ldots, \lambda_1[\bar{\boldsymbol{A}}] = \bar{\lambda}_1 = \lambda_n - \mu.$$

Führt man den QR-Algorithmus mit der Startmatrix $\bar{\boldsymbol{A}}^{(0)} = \bar{\boldsymbol{A}} = \boldsymbol{A} - \mu \boldsymbol{I}$ durch, so ergibt sich für die Approximation des zum Eigenwert $\bar{\lambda}_n = \lambda_1 - \mu$ gehörenden Teilraums der Konvergenzfaktor

$$\bar{\kappa} = \left| \frac{\bar{\lambda}_n}{\lambda_{n-1}} \right| = \left| \frac{\lambda_1 - \mu}{\lambda_2 - \mu} \right| \ll 1.$$

Nach k Schritten erhalten wir die Matrix

$$\bar{\boldsymbol{A}}^{(k)} = \begin{pmatrix} \bar{\boldsymbol{A}}_{11}^{(k)} & \boldsymbol{a}^{(k)} \\ \\ \boldsymbol{a}^{(k)T} & \bar{a}_{nn}^{(k)} \end{pmatrix}, \quad \bar{\boldsymbol{A}}_{11}^{(k)} \in \mathbb{R}^{(n-1)\times(n-1)}, \quad \boldsymbol{a}^{(k)} \in \mathbb{R}^{n-1}.$$

Mit Satz 9.52 folgt, dass unter der Voraussetzung $\cos\bar{\varphi}_1 > 0$ mit

$$\bar{\varphi}_1 = \sphericalangle(\text{span}\,(\boldsymbol{e}_1, \ldots, \boldsymbol{e}_{n-1}), \boldsymbol{\mathcal{S}}_{n-1})$$

$$\|\boldsymbol{a}^{(k)}\|_2 \leqq (1+\bar{\kappa})\bar{\varepsilon}_k \|\bar{\boldsymbol{A}}\|_2, \quad \bar{\varepsilon}_{k+1} = \bar{\kappa}\bar{\varepsilon}_k = \bar{\kappa}^k \tan\bar{\varphi}_1$$

gilt. Folglich wird $\|\boldsymbol{a}^{(k)}\|_2$ schnell klein, falls $\tan\bar{\varphi}_1$ nicht zu groß ist. Nach wenigen Schritten ist daher $\boldsymbol{a}^{(k)}$ vernachlässigbar, und $\bar{a}_{nn}^{(k)}$ ist dann eine akzeptable Näherung

für $\bar{\lambda}_n = \lambda_1 - \mu$.
Die Matrizen $\bar{\boldsymbol{A}}^{(k)}$ sind nach Satz 9.49 orthogonal ähnlich zur Matrix $\bar{\boldsymbol{A}} = \boldsymbol{A} - \mu\boldsymbol{I}$. Definieren wir nun über

$$\boldsymbol{A}^{(k)} = \bar{\boldsymbol{A}}^{(k)} + \mu\boldsymbol{I}$$

eine weitere Folge von Matrizen, so gilt wegen

$$\bar{\boldsymbol{A}}^{(k)} = \bar{\boldsymbol{Q}}^{(k)}\bar{\boldsymbol{R}}^{(k)}, \quad \bar{\boldsymbol{A}}^{(k+1)} = \bar{\boldsymbol{R}}^{(k)}\bar{\boldsymbol{Q}}^{(k)}$$

$$\begin{aligned}
\boldsymbol{A}^{(k+1)} &= \bar{\boldsymbol{A}}^{(k+1)} + \mu\boldsymbol{I} \\
&= \bar{\boldsymbol{R}}^{(k)}\bar{\boldsymbol{Q}}^{(k)} + \mu\boldsymbol{I} \\
&= \bar{\boldsymbol{Q}}^{(k)^T}\bar{\boldsymbol{Q}}^{(k)}\bar{\boldsymbol{R}}^{(k)}\bar{\boldsymbol{Q}}^{(k)} + \mu\boldsymbol{I} \\
&= \bar{\boldsymbol{Q}}^{(k)^T}\bar{\boldsymbol{A}}^{(k)}\bar{\boldsymbol{Q}}^{(k)} + \mu\boldsymbol{I} \\
&= \bar{\boldsymbol{Q}}^{(k)^T}\left(\boldsymbol{A}^{(k)} - \mu\boldsymbol{I}\right)\bar{\boldsymbol{Q}}^{(k)} + \mu\boldsymbol{I} \\
&= \bar{\boldsymbol{Q}}^{(k)^T}\boldsymbol{A}^{(k)}\bar{\boldsymbol{Q}}^{(k)}.
\end{aligned}$$

Die so definierten Matrizen $\boldsymbol{A}^{(k)}$ sind daher orthogonal ähnlich zu $\boldsymbol{A}^{(0)} = \boldsymbol{A}$. Partitionieren wir die Matrix $\boldsymbol{A}^{(k)}$ genau wie $\bar{\boldsymbol{A}}^{(k)}$, so gilt

$$\boldsymbol{A}^{(k)} = \begin{pmatrix} \boldsymbol{A}_{11}^{(k)} & \boldsymbol{a}^{(k)} \\ \\ \boldsymbol{a}^{(k)^T} & a_{nn}^{(k)} \end{pmatrix} = \begin{pmatrix} \bar{\boldsymbol{A}}_{11}^{(k)} + \mu\boldsymbol{I} & \boldsymbol{a}^{(k)} \\ \\ \boldsymbol{a}^{(k)^T} & \bar{\boldsymbol{a}}_{nn}^{(k)} + \mu \end{pmatrix}.$$

Die Nichtdiagonalblöcke von $\boldsymbol{A}^{(k)}$ sind dieselben wie die von $\bar{\boldsymbol{A}}^{(k)}$. Sie konvergieren gegen Null. Gilt für genügend großes k

$$\|\boldsymbol{a}^{(k)}\|_2 \leqq \varepsilon,$$

so lässt sich zur Matrix

$$\hat{\boldsymbol{A}}^{(k)} = \begin{pmatrix} \boldsymbol{A}_{11}^{(k)} & \boldsymbol{o} \\ \\ \boldsymbol{o}^T & a_{nn}^{(k)} \end{pmatrix}$$

übergehen. Das Diagonalelement $a_{nn}^{(k)}$ ist in diesem Falle eine gute Näherung für den Eigenwert λ_1 von $\boldsymbol{A}$. Der Algorithmus ist dann mit der symmetrischen Matrix

$\boldsymbol{A}_{11}^{(k)} \in \mathbb{R}^{(n-1)\times(n-1)}$ fortsetzbar. Die Konvergenzgeschwindigkeit des Algorithmus erhöht sich, falls die Verschiebung $\mu = \mu_k$ in jedem Schritt so festgelegt wird, dass sie eine möglichst gute Eigenwertapproximation darstellt. Wir erhalten damit folgenden Algorithmus:

9.53. QR**-Algorithmus mit Verschiebungen:**

S0 *(Initialisierung) Setze* $\boldsymbol{A}^{(0)} = \boldsymbol{A}, \boldsymbol{V}^{(0)} = \boldsymbol{I}$ *und* $k = 0$.

S1 *(Spektralverschiebung) Wähle Verschiebungsparameter* μ_k.

S2 *(*QR*-Zerlegung) Berechne eine orthogonale Matrix* $\boldsymbol{Q}^{(k)}$ *und eine obere Dreiecksmatrix* $\boldsymbol{R}^{(k)}$ *mit*

$$\boldsymbol{A}^{(k)} - \mu_k \boldsymbol{I} = \boldsymbol{Q}^{(k)} \boldsymbol{R}^{(k)}.$$

S3 *(Iteration) Berechne*

$$\boldsymbol{A}^{(k+1)} = \boldsymbol{R}^{(k)} \boldsymbol{Q}^{(k)} + \mu_k \boldsymbol{I}, \quad \boldsymbol{V}^{(k+1)} = \boldsymbol{V}^{(k)} \boldsymbol{Q}^{(k)}.$$

S3 *Setze* $k = k + 1$ *und gehe zu Schritt* **S1**.

Bemerkungen: (i) Auch für die von diesem Algorithmus erzeugten Matrizen gilt

$$\boldsymbol{A}^{(k)} = {\boldsymbol{V}^{(k)}}^T \boldsymbol{A} \boldsymbol{V}^{(k)}, \quad k = 0, 1, \ldots.$$

Es wird wieder eine Folge von orthogonal ähnlichen Matrizen erzeugt.
(ii) Falls μ_k kein Eigenwert der Matrix $\boldsymbol{A}$ ist, ist die Matrix $\boldsymbol{A}^{(k)} - \mu_k \boldsymbol{I}$ regulär. Bis auf die Vorzeichen der Spalten bzw. Zeilen sind dann die Faktoren $\boldsymbol{Q}^{(k)}$ und $\boldsymbol{R}^{(k)}$ eindeutig festgelegt. Der Algorithmus ist auch für den Fall, dass μ_k Eigenwert der Matrix $\boldsymbol{A}$ ist, durchführbar. In diesem Falle treten bei der QR-Zerlegung Nullen in der Diagonalen von $\boldsymbol{R}^{(k)}$ auf, ein Eigenwert lässt sich abgespalten, und die Dimension des Problems verringert sich.
Es blieb im Algorithmus noch die Frage nach der Wahl der Verschiebungsparameter μ_k offen. Als einfachste Wahl bietet es sich an, $\mu_k = a_{nn}^{(k)}$ zu wählen.

9.54. QR**-Algorithmus mit** RAYLEIGH**-Quotienten-Verschiebungen:**

S0 *(Initialisierung) Setze*

$$\boldsymbol{A}^{(0)} = \boldsymbol{A}, \quad \boldsymbol{V}^{(0)} = \boldsymbol{I}, \quad k = 0.$$

S1 *(Spektralverschiebung) Wähle Verschiebungsparameter* $\mu_k = a_{nn}^{(k)}$.
(QR*-Zerlegung) Berechne eine orthogonale Matrix* $\boldsymbol{Q}^{(k)}$ *und eine obere Dreiecksmatrix* $\boldsymbol{R}^{(k)}$ *mit*

$$\boldsymbol{A}^{(k)} - \mu_k \boldsymbol{I} = \boldsymbol{Q}^{(k)} \boldsymbol{R}^{(k)}.$$

S2 *(Iteration) Berechne*

$$\boldsymbol{A}^{(k+1)} = \boldsymbol{R}^{(k)}\boldsymbol{Q}^{(k)} + \mu_k \boldsymbol{I}, \quad \boldsymbol{V}^{(k+1)} = \boldsymbol{V}^{(k)}\boldsymbol{Q}^{(k)}.$$

S3 *Setze $k = k+1$ und gehe zu Schritt* **S1.**

Es gibt einen Zusammenhang mit der RAYLEIGH-Quotienten-Iteration, wie der folgende Satz zeigt.

9.55. Satz: *Wenn der Algorithmus 9.54 exakt durchgeführt wird und in jedem Schritt μ_k kein Eigenwert der Matrix $\boldsymbol{A}$ ist, stimmen die n-ten Spalten $\boldsymbol{v}_n^{(k)} = \boldsymbol{V}^{(k)}\boldsymbol{e}_n$ der Matrizen $\boldsymbol{V}^{(k)}$ bis auf das Vorzeichen mit den Iterationsvektoren $\boldsymbol{v}^{(k)}$, die sich nach der* RAYLEIGH*-Quotienten-Iteration mit dem Startvektor $\boldsymbol{v}^{(0)} = \boldsymbol{e}_n$ ergeben, überein.*

Beweis: Es gilt

$$\mu_k = a_{nn}^{(k)} = \boldsymbol{e}_n^T \boldsymbol{A}^{(k)} \boldsymbol{e}_n = \boldsymbol{e}_n^T {\boldsymbol{V}^{(k)}}^T \boldsymbol{A} \boldsymbol{V}^{(k)} \boldsymbol{e}_n = {\boldsymbol{v}_n^{(k)}}^T \boldsymbol{A} \boldsymbol{v}_n^{(k)} = \varrho\left(\boldsymbol{v}_n^{(k)}\right).$$

Die Größe μ_k ist der zu $\boldsymbol{v}_n^{(k)}$ gehörende RAYLEIGH-Quotient. Es folgt weiter

$$\begin{aligned}
(\boldsymbol{A} - \mu_k \boldsymbol{I})\, \boldsymbol{v}_n^{(k+1)} &= \boldsymbol{V}^{(k)} \left(\boldsymbol{A}^{(k)} - \mu_k \boldsymbol{I}\right)^T {\boldsymbol{V}^{(k)}}^T \boldsymbol{v}_n^{(k+1)} \\
&= \boldsymbol{V}^{(k)} \left(\boldsymbol{Q}^{(k)} \boldsymbol{R}^{(k)}\right)^T {\boldsymbol{V}^{(k)}}^T \boldsymbol{v}_n^{(k+1)} \\
&= \boldsymbol{V}^{(k)} {\boldsymbol{R}^{(k)}}^T {\boldsymbol{Q}^{(k)}}^T {\boldsymbol{V}^{(k)}}^T \boldsymbol{v}_n^{(k+1)} \\
&= \boldsymbol{V}^{(k)} {\boldsymbol{R}^{(k)}}^T {\boldsymbol{V}^{(k+1)}}^T \boldsymbol{v}_n^{(k+1)} \\
&= \boldsymbol{V}^{(k)} {\boldsymbol{R}^{(k)}}^T \boldsymbol{e}_n \\
&= r_{nn}^{(k)} \boldsymbol{V}^{(k)} \boldsymbol{e}_n \\
&= r_{nn}^{(k)} \boldsymbol{v}_n^{(k)}.
\end{aligned}$$

Da μ_k kein Eigenwert von $\boldsymbol{A}$ ist, muss $r_{nn}^{(k)} \neq 0$ sein, so dass der Vektor

$$\boldsymbol{w}_n^{(k+1)} = \boldsymbol{v}_n^{(k+1)} / r_{nn}^{(k)}$$

der Gleichung

$$(\boldsymbol{A} - \mu_k \boldsymbol{I})\, \boldsymbol{w}_n^{(k+1)} = \boldsymbol{v}_n^{(k)}$$

genügt. Das ist genau die Iterationsvorschrift der RAYLEIGH-Quotienten-Iteration. Damit folgt die Behauptung. ❀

Aus Satz 9.55 folgt sofort, dass die Konvergenzaussagen aus Satz 9.44 übernehmbar sind. Insbesondere ist die Konvergenz asymptotisch kubisch. Falls für hinreichen großes k $r_{nn}^{(k)} \leqq \varepsilon$ gilt, so ist μ_k als Eigenwert von $\boldsymbol{A}$ akzeptierbar. Es gilt dann

$$\boldsymbol{A}^{(k+1)} = \begin{pmatrix} \bar{\boldsymbol{A}}^{(k+1)} & \boldsymbol{o} \\ \boldsymbol{o}^T & \mu_k \end{pmatrix} + \boldsymbol{\delta A}^{(k+1)}, \quad \|\boldsymbol{\delta A}^{(k+1)}\|_2 \leqq \varepsilon.$$

Der Algorithmus ist dann mit $\bar{\boldsymbol{A}}^{(k+1)}$ statt $\boldsymbol{A}^{(k+1)}$ fortsetzbar.
Eine etwas feinere Verschiebungsstrategie ist in folgendem Algorithmus enthalten.

9.56. QR**-Algorithmus mit**WILKINSON**-Verschiebungen:**

S0 *(Initialisierung) Setze*

$$\boldsymbol{A}^{(0)} = \boldsymbol{A}, \quad \boldsymbol{V}^{(0)} = \boldsymbol{I}, \quad k = 0.$$

S1 *(Spektralverschiebung)*

- *Berechne die Eigenwerte μ' und μ'' der symmetrischen Matrix*

$$\boldsymbol{P}^{(k)} = \begin{pmatrix} a_{n-1,n-1}^{(k)} & a_{n-1,n}^{(k)} \\ \\ a_{n,n-1}^{(k)} & a_{n,n}^{(k)} \end{pmatrix} \in \mathbb{R}^{2\times 2}.$$

- *Ordne μ' und μ'' so, dass*

$$|\mu' - a_{n,n}^{(k)}| < |\mu'' - a_{n,n}^{(k)}|$$

bzw.

$$|\mu'| < |\mu''|$$

im Falle

$$|\mu' - a_{n,n}^{(k)}| = |\mu'' - a_{n,n}^{(k)}|$$

gilt.

- *Setze $\mu_k = \mu'$.*

S2 *(QR-Zerlegung) Berechne eine orthogonale (n,n)-Matrix $\boldsymbol{Q}^{(k)}$ und eine obere (n,n)-Dreiecksmatrix $\boldsymbol{R}^{(k)}$ mit*

$$\boldsymbol{A}^{(k)} - \mu_k \boldsymbol{I} = \boldsymbol{Q}^{(k)} \boldsymbol{R}^{(k)}.$$

S3 *(Iteration) Berechne*

$$\boldsymbol{A}^{(k+1)} = \boldsymbol{R}^{(k)}\boldsymbol{Q}^{(k)} + \mu_k \boldsymbol{I}, \quad \boldsymbol{V}^{(k+1)} = \boldsymbol{V}^{(k)}\boldsymbol{Q}^{(k)}.$$

S4 *Setze* $k = k+1$ *und gehe zu Schritt* **S1**.

Bemerkung: Die Matrix $\boldsymbol{P}^{(k)}$ ist gerade die zur Matrix $\boldsymbol{Q} = (\boldsymbol{v}_{n-1}^{(k)}, \boldsymbol{v}_n^{(k)})$ gehörende Matrix aus dem RAYLEIGH-RITZ-Algorithmus. Durch μ' und μ'' sind gerade die RITZschen Eigenwerte bezüglich des Unterraums span$(\boldsymbol{v}_{n-1}^{(k)}, \boldsymbol{v}_n^{(k)})$ gegeben. Damit ist der Algorithmus 9.56 eine Erweiterung des Algorithmus 9.54 auf den zweidimensionalen Fall.

9.9. Der QR-Algorithmus für Tridiagonalmatrizen

Obwohl die Konvergenz des im vorigen Abschnitt angegebenen QR-Algorithmus asymptotisch kubisch ist, ist er gegenüber dem JACOBI-Verfahren noch nicht konkurrenzfähig. Normalerweise sind immer mehrere QR-Schritte zur Abspaltung eines Eigenwertes notwendig. Selbst wenn nach jedem QR-Schritt ein Eigenwert abgespalten würde, und falls man außerdem die Dimensionsreduktion in jedem Schritt berücksichtigt, so würde das Berechnen aller Eigenwerte rund $n^4/3$ Additionen und Multiplikationen kosten. Der Aufwand beim JACOBI-Verfahren liegt dagegen bei $6n^3$ Operationen, also um eine n-Potenz niedriger. Die Situation ändert sich grundlegend, falls man die Matrix mit den Methoden aus Abschnitt 9.6. orthogonal ähnlich auf Tridiagonalform transformiert, und den QR-Algorithmus dann auf diese Tridiagonalmatrix anwendet.
Bei der QR-Zerlegung einer Tridiagonalmatrix $\boldsymbol{T}$ bietet es sich wegen der speziellen Struktur an, mit GIVENS-Drehungen zu arbeiten. Mit Hilfe von $n-1$ GIVENS-Transformationen lassen sich so nacheinander die Subdiagonalelemente an den Positionen $(2,1),(3,2),(4,3),\ldots,(n,n-1)$ zu Null machen. Man erhält

$$\boldsymbol{G}_{n-1,n}\boldsymbol{G}_{n-2,n-1}\cdots\boldsymbol{G}_{2,3}\boldsymbol{G}_{1,2}\boldsymbol{T} = \boldsymbol{R}.$$

Mit

$$\boldsymbol{Q} = \boldsymbol{G}_{1,2}^T\boldsymbol{G}_{2,3}^T\cdots\boldsymbol{G}_{n-2,n-1}^T\boldsymbol{G}_{n-1,n}^T$$

gilt dann

$$\boldsymbol{T} = \boldsymbol{Q}\boldsymbol{R}.$$

Da die GIVENS-Drehungen jeweils nur auf zwei Zeilen wirken, hat die Matrix $\boldsymbol{R}$ die spezielle Struktur

$$\boldsymbol{R} = \begin{pmatrix} \times & \times & \times & & & \\ & \times & \times & \times & & \\ & & \ddots & \ddots & \ddots & \\ & & & \times & \times & \times \\ & & & & \times & \times \\ & & & & & \times \end{pmatrix}.$$

Bildet man nun die Matrix

$$\bar{\boldsymbol{T}} = \boldsymbol{RQ} = \boldsymbol{R}\boldsymbol{G}_{1,2}^T\boldsymbol{G}_{2,3}^T\cdots\boldsymbol{G}_{n-2,n-1}^T\boldsymbol{G}_{n-1,n}^T,$$

so wird durch jede GIVENS-Transformation $\boldsymbol{G}_{j,j+1}^T$ ein Subdiagonalelement in der Spalte j erzeugt. Die Matrix $\bar{\boldsymbol{T}}$ ist damit eine obere HESSENBERG-Matrix:

$$\bar{\boldsymbol{T}} = \begin{pmatrix} \times & \times & \times & & & \\ \times & \times & \times & \times & & \\ & \ddots & \ddots & \ddots & \ddots & \\ & & \times & \times & \times & \times \\ & & & \times & \times & \times \\ & & & & \times & \times \end{pmatrix}.$$

Andererseits wissen wir, dass im QR-Algorithmus die Symmetrie der Matrizen erhalten bleibt. Damit folgt aber, dass $\bar{\boldsymbol{T}}$ tridiagonal ist. Da die QR-Zerlegung einer Matrix bis auf die Vorzeichen der Spalten von $\boldsymbol{Q}$ bzw. der Zeilen von $\boldsymbol{R}$ eindeutig ist, gilt diese Aussage natürlich auch bei Anwendung eines anderen Verfahrens zur QR-Zerlegung. Starten wir den QR-Algorithmus mit einer Tridiagonalmatrix, so wird eine Folge von Tridiagonalmatrizen erzeugt. Der Aufwand für die Durchführung eines QR-Schritts verringert sich damit von $\sim 4n^3/3$ Additionen und Multiplikationen im Falle einer vollbesetzten Matrix auf $\sim K_1 n$ Additionen, $\sim K_2 n$ Multiplikationen und n Quadratwurzeln, falls mit Tridiagonalmatrizen gearbeitet wird. Die Konstanten K_1 und K_2 hängen von der konkreten Implementierung ab. Die einmalige Transformation der Ausgangsmatrix auf Tridiagonalform benötigt weitere $\sim 2n^3/3$ Operationen. Das ist etwa der halbe Aufwand eines QR-Schritts für eine vollbesetzte Matrix. Man erkennt, dass sich der Gesamtaufwand des QR-Algorithmus entscheidend verringert. Der Übergang zu Tridiagonalmatrizen hat noch einen weiteren Vorteil. Eine symme-

trische Tridiagonalmatrix

$$\boldsymbol{T} = \text{trid}(a_1,\ldots,a_n,b_2,\ldots,b_n) = \begin{pmatrix} a_1 & b_2 & & & & & \\ b_2 & a_2 & b_3 & & & & \\ & b_3 & a_3 & b_4 & & & \\ & & \ddots & \ddots & \ddots & & \\ & & & b_{n-2} & a_{n-2} & b_{n-1} & \\ & & & & b_{n-1} & a_{n-1} & b_n \\ & & & & & b_n & a_n \end{pmatrix}$$

heißt nichtzerfallend, falls alle Nebendiagonalelemente von Null verschieden sind. Existiert ein Index k mit $b_k = 0$, so heißt die Matrix $\boldsymbol{T}$ zerfallend. Es gilt dann

$$\boldsymbol{T} = \left(\begin{array}{ccccc|ccccc} a_1 & b_2 & & & & & & & & \\ b_2 & a_2 & b_3 & & & & & & & \\ & \ddots & \ddots & \ddots & & & & & & \\ & & b_{k-2} & a_{k-2} & b_{k-1} & & & & & \\ & & & b_{k-1} & a_{k-1} & & & & & \\ \hline & & & & & a_k & b_{k+1} & & & \\ & & & & & b_{k+1} & a_{k+1} & b_{k+2} & & \\ & & & & & & \ddots & \ddots & \ddots & \\ & & & & & & & b_{n-1} & a_{n-1} & b_n \\ & & & & & & & & b_n & a_n \end{array}\right)$$

$$= \begin{pmatrix} \boldsymbol{T}_1 & \\ & \boldsymbol{T}_2 \end{pmatrix}$$

mit den Tridiagonalmatrizen $\boldsymbol{T}_1 \in \mathbb{R}^{(k-1)\times(k-1)}$ und $\boldsymbol{T}_2 \in \mathbb{R}^{(n-k+1)\times(n-k+1)}$.
Ist die Matrix $\boldsymbol{T}$ zerfallend, so zerfällt das zugehörige Eigenwertproblem offensichtlich auch in zwei Eigenwertprobleme geringerer Dimension. Damit ist fast immer eine beträchtliche Verringerung des Aufwands verbunden. Darum ist es günstig, auch hinreichend kleine nichtverschwindende Nebendiagonalelemente als Null anzusehen. Als Kriterium für die Kleinheit von Nichtdiagonalelementen ist

$$|b_k| \leqq \text{eps}\|\boldsymbol{T}\|_2 = \varepsilon$$

anwendbar. Etwas feiner ist das Kriterium

$$|b_k| \leqq \text{eps}(|a_{k-1}| + |a_k|) = \varepsilon.$$

Das Nullsetzen von b_k entspricht in beiden Fällen einer Störung $\boldsymbol{\delta T}$ der Matrix $\boldsymbol{T}$ und damit einer Störung $\boldsymbol{\delta A}$ der Matrix $\boldsymbol{A}$ vom Niveau

$$\|\boldsymbol{\delta T}\|_2 = \|\boldsymbol{\delta A}\|_2 \leqq \varepsilon.$$

Eine wichtige Eigenschaft von Tridiagonalmatrizen wird in folgendem Satz beschrieben.

9.57. Satz: *Alle Eigenwerte einer nichtzerfallenden symmetrischen Tridiagonalmatrix sind einfach. Es existiert eine Nummerierung mit*

$$\lambda_1 < \lambda_2 < \cdots < \lambda_{n-1} < \lambda_n.$$

Beweis: Da $\boldsymbol{T}$ als nichtzerfallend vorausgesetzt wurde, sind für jedes λ die ersten $n-1$ Spalten der Matrix $\boldsymbol{T}-\lambda\boldsymbol{I}$ linear unabhängig. Es gilt also $rg(\boldsymbol{T}-\lambda\boldsymbol{I}) \geqq n-1$. Ist λ ein Eigenwert von $\boldsymbol{T}$, so ist $\boldsymbol{T}-\lambda\boldsymbol{I}$ singulär. Dann gilt aber $rg(\boldsymbol{T}-\lambda\boldsymbol{I}) \leqq n-1$. Insgesamt folgt dann $rg(\boldsymbol{T}-\lambda\boldsymbol{I}) = n-1$. Die geometrische Vielfachheit des Eigenwertes λ ist 1, λ ist einfacher Eigenwert. ❀

Für eine Matrix mit mehrfachen Eigenwerten wird bei exakter orthogonaler Transformation die zugehörige Tridiagonalmatrix zerfallend sein. Es ist dann wieder zu erwarten, dass der Gesamtaufwand sich deutlich verringert. Sind einige Eigenwerte nur dicht benachbart, so muss nach Transformation auf Tridiagonalform keine Matrix entstehen, die „fast zerfallend“ in dem Sinne ist, dass einige Nebendiagonalelemente klein werden. Ein bekanntes Beispiel dafür ist die WILKINSON-Matrix, die bei deutlich von Null verschiedenen Nichtdiagonalelementen eng benachbarte Eigenwerte besitzt.

9.58. Beispiel: Die WILKINSON-Matrix:

$$\boldsymbol{W} = \text{trid}(10,9,\ldots,1,0,1,\ldots,9,10;1,\ldots,1)$$

hat die Eigenwerte

$$\lambda_{20} = 10.74619418290339\ldots,$$

$$\lambda_{21} = 10.74619418290332\ldots$$

♡

Wendet man den QR-Algorithmus mit Verschiebungen auf eine Tridiagonalmatrix an, so ist die Tridiagonalform unter der Transformation

$$\boldsymbol{A}^{(k)} \to \boldsymbol{A}^{(k+1)} = \boldsymbol{Q}^{(k)^T}\boldsymbol{A}^{(k)}\boldsymbol{Q}^{(k)} = \boldsymbol{R}^{(k)}\boldsymbol{Q}^{(k)} + \mu_k\boldsymbol{I}$$

ebenfalls invariant solange μ_k kein Eigenwert von $\boldsymbol{A}$ ist. Es lässt sich nun zeigen, dass falls $\boldsymbol{A}^{(k)}$ nicht zerfallend ist und μ_k Eigenwert von $\boldsymbol{A}$ ist, die letzte Zeile der

Matrix $\boldsymbol{A}^{(k+1)}$ bei exakter Rechnung die Gestalt $(0,\ldots,0,\mu_k)$ hat. Das Subdiagonalelement $a_{n,n-1}^{(k+1)}$ verschwindet und μ_k wird damit als Eigenwert erkannt und abgespalten. Wendet man die RAYLEIGH-Quotienten-Verschiebung an, so konvergiert $|a_{n,n-1}^{(k)}| = |b_n^{(k)}|$ in exakter Arithmetik fast immer monoton und asymptotisch kubisch gegen 0. Bei Verwendung von WILKINSON-Verschiebungen liegt sogar globale Konvergenz vor. Die Konvergenz ist dabei mindestens quadratisch und meistens besser als kubisch.
Arbeitet man bei der numerischen Realisierung mit GIVENS-Drehungen, so zeigt es sich, dass die QR-Zerlegung und das anschließende Berechnen von $\boldsymbol{RQ}$ in gewisser Weise gleichzeitig durchführbar ist. Hat man nämlich mit j GIVENS-Drehungen die Subdiagonalelemente in den Spalten $1,\ldots,j$ von $\boldsymbol{A}^{(k)} - \mu_k \boldsymbol{I}$ eliminiert, so bleiben diese Spalten von den weiteren Transformationen unberührt. Sie werden weder benötigt noch geändert. Diese Spalten sind schon die entsprechenden Spalten der Matrix $\boldsymbol{R}^{(k)}$, und auf sie sind schon die ersten $j-1$ Transformationen von rechts anwendbar. Ein Schritt des QR-Algorithmus ist somit im wesentlichen auf dem Platz der Matrix $\boldsymbol{A}$ durchführbar.

9.59. Expliziter QR-Schritt:

S0 *(Initialisierung) Wähle Verschiebungsparameter μ_k und bilde*

$$\bar{\boldsymbol{A}} = \boldsymbol{A}^{(k)} - \mu_k \boldsymbol{I}.$$

Setze $\boldsymbol{G}_{01} = \boldsymbol{I}$ und $\bar{\boldsymbol{V}} = \boldsymbol{V}^{(k)}$.

S1 *(Transformation) Für $j = 1,\ldots,n-1$ führe aus:*

- *Lege $\boldsymbol{G}_{j,j+1}$ so fest, dass $(\boldsymbol{G}_{j,j+1}\bar{\boldsymbol{A}})_{j+1,j} = 0$ gilt.*
- *Bilde $\bar{\boldsymbol{A}} = \boldsymbol{G}_{j,j+1}\bar{\boldsymbol{A}}$, $\bar{\boldsymbol{A}} = \bar{\boldsymbol{A}}\boldsymbol{G}_{j-1,j}$ und $\bar{\boldsymbol{V}} = \bar{\boldsymbol{V}}\boldsymbol{G}_{j,j+1}^T$.*

S2 *Berechne*

$$\bar{\boldsymbol{A}} = \bar{\boldsymbol{A}}\boldsymbol{G}_{n-1,n}^T, \quad \boldsymbol{A}^{(k+1)} = \bar{\boldsymbol{A}} + \mu_k \boldsymbol{I}$$

und setze

$$\boldsymbol{V}^{(k+1)} = \bar{\boldsymbol{V}}.$$

Aufwand:

- *$\sim 4n$ Add./Sub., $\sim 13n$ Mult./Div. und n Quadratwurzeln zum Berechnen der Matrix $\boldsymbol{A}^{(k+1)}$.*
- *$\sim 3n^2$ Add./Sub. und $\sim 3n^2$ Mult./Div. zum Berechnen von $\boldsymbol{V}^{(k+1)}$.*

Bemerkung: Beim expliziten QR-Schritt ist es günstig, sofort

$$\boldsymbol{A}^{(k+1)} = \boldsymbol{R}^{(k)}\boldsymbol{Q}^{(k)} = {\boldsymbol{Q}^{(k)}}^T(\boldsymbol{A}^{(k)} - \mu_k\boldsymbol{I})\boldsymbol{Q}^{(k)}$$

zu setzen. Es gilt dann

$$\boldsymbol{A}^{(k+1)} = {\boldsymbol{V}^{(k)}}^T\left[\boldsymbol{A} - (\mu_0 + \cdots + \mu_k)\boldsymbol{I}\right]\boldsymbol{V}^{(k)}.$$

Die Verschiebungen sind dann gemäß $\sigma_0 = 0$, $\sigma_{k+1} = \sigma_k + \mu_k$ für $k = 0, 1, \ldots$ gesondert aufzusummieren.
Im obigen Expliziten QR-Schritt wirkt es noch etwas störend, dass die GIVENS-Rotationen nicht simultan von links und von rechts anwendbar sind. Mit dem folgenden Satzes lässt sich der Algorithmus so modifizieren, dass dieser Mangel beseitigt wird.

9.60. Satz: *Für die symmetrische (n,n)-Matrix $\boldsymbol{A}$ gelte*

$$\boldsymbol{A} = \boldsymbol{Q}\boldsymbol{T}\boldsymbol{Q}^T = \bar{\boldsymbol{Q}}\bar{\boldsymbol{T}}\bar{\boldsymbol{Q}}^T$$

mit Tridiagonalmatrizen

$$\boldsymbol{T} = \text{trid}(a_1, \ldots, a_n, b_2, \ldots, b_n), \quad \bar{\boldsymbol{T}} = \text{trid}(\bar{a}_1, \ldots, \bar{a}_n, \bar{b}_2, \ldots, \bar{b}_n)$$

und orthogonalen Matrizen

$$\boldsymbol{Q} = (\boldsymbol{q}_1, \ldots, \boldsymbol{q}_n), \quad \bar{\boldsymbol{Q}} = (\bar{\boldsymbol{q}}_1, \ldots, \bar{\boldsymbol{q}}_n).$$

Ist die Matrix $\boldsymbol{T}$ nichtzerfallend, also $b_i \neq 0$ für $i = 2, \ldots, n$, und gilt

$$\boldsymbol{q}_1 = \boldsymbol{Q}\boldsymbol{e}_1 = \bar{\boldsymbol{Q}}\boldsymbol{e}_1 = \bar{\boldsymbol{q}}_1,$$

so ist

$$\bar{\boldsymbol{Q}} = \boldsymbol{Q}\boldsymbol{D}, \quad \bar{\boldsymbol{T}} = \boldsymbol{D}\boldsymbol{T}\boldsymbol{D}$$

mit

$$\boldsymbol{D} = \text{diag}(d_1, \ldots, d_n), \quad d_1 = 1, \quad |d_i| = 1, \quad i = 2, \ldots, n.$$

Beweis: Aus $\boldsymbol{Q}\boldsymbol{T}\boldsymbol{Q}^T = \bar{\boldsymbol{Q}}\bar{\boldsymbol{T}}\bar{\boldsymbol{Q}}^T$ folgt $\bar{\boldsymbol{T}}\boldsymbol{P} = \boldsymbol{P}\boldsymbol{T}$ mit der orthogonalen Matrix

$$\boldsymbol{P} = (\boldsymbol{p}_1, \ldots, \boldsymbol{p}_n) = \bar{\boldsymbol{Q}}^T\boldsymbol{Q}.$$

Es gilt

$$\boldsymbol{p}_1 = \boldsymbol{P}\boldsymbol{e}_1 = \bar{\boldsymbol{Q}}^T\boldsymbol{Q}\boldsymbol{e}_1 = \bar{\boldsymbol{Q}}^T\boldsymbol{q}_1 = \bar{\boldsymbol{Q}}^T\bar{\boldsymbol{q}}_1 = \boldsymbol{e}_1.$$

Damit folgt

$$\bar{\boldsymbol{T}}\boldsymbol{P}\boldsymbol{e}_1 = \bar{\boldsymbol{T}}\boldsymbol{e}_1 = \bar{a}_1\boldsymbol{e}_1 + \bar{b}_2\boldsymbol{e}_2 = \boldsymbol{P}\boldsymbol{T}\boldsymbol{e}_1 = a_1\boldsymbol{p}_1 + b_2\boldsymbol{p}_2.$$

Linksmultiplikation mit $\boldsymbol{e}_1^T$ liefert $\bar{a}_1 = a_1$ und damit $\bar{b}_2\boldsymbol{e}_2 = b_2\boldsymbol{p}_2$, also $|\bar{b}_2| = |b_2|$ und $\boldsymbol{p}_2 = d_2\boldsymbol{e}_2$ mit $d_2 = b_2/b_2$ (da $\boldsymbol{T}$als nichtzerfallend vorausgesetzt wurde gilt $b_2 \neq 0$). Damit gilt $|d_2| = 1$. In analoger Weise folgt

$$\begin{aligned}\bar{\boldsymbol{T}}\boldsymbol{P}\boldsymbol{e}_2 &= \bar{\boldsymbol{T}}\boldsymbol{p}_2 = d_2\bar{\boldsymbol{T}}\boldsymbol{e}_2 = d_2(\bar{b}_2\boldsymbol{e}_1 + \bar{a}_2\boldsymbol{e}_2 + \bar{b}_3\boldsymbol{e}_3)\\ &= \boldsymbol{P}\boldsymbol{T}\boldsymbol{e}_2 = b_2\boldsymbol{p}_1 + a_2\boldsymbol{p}_2 + b_3\boldsymbol{p}_3 = b_2\boldsymbol{e}_1 + d_2a_2\boldsymbol{e}_2 + b_3\boldsymbol{p}_3.\end{aligned}$$

Wegen $d_2\bar{b}_2 = \bar{b}_2^2/b_2 = b_2^2/b_2 = b_2$ folgt

$$d_2(\bar{a}_2 - a_2\boldsymbol{e}_2) + \bar{b}_3\boldsymbol{e}_3 - b_3\boldsymbol{p}_3 = 0.$$

Multipliziert man diese Gleichung von links mit $\boldsymbol{e}_2^T$ und beachtet dabei, dass aus der Orthogonalität von $\boldsymbol{P}$ und $\boldsymbol{p}_2 = d_2\boldsymbol{e}_2$ $\boldsymbol{e}_2^T\boldsymbol{p}_3 = 0$ folgt, so ergibt sich $\bar{a}_2 = a_2$ und $\boldsymbol{p}_3 = d_3\boldsymbol{e}_3$ mit $d_3 = d_2\bar{b}_3/b_3$, daher $|d_3| = 1$. Durch Induktion folgt dann weiter $\boldsymbol{p}_i = d_i\boldsymbol{e}_i$ mit $d_i = d_{i-1}\bar{b}_i/b_i$ (also $|d_i| = 1$) für $i = 4, \ldots, n$. ❀

Bemerkung: Die Bedingung $\boldsymbol{q}_1 = \bar{\boldsymbol{q}}_1$ ist durch $\boldsymbol{q}_n = \bar{\boldsymbol{q}}_n$ ersetzbar. Der Beweis ist dann in umgekehrter Reihenfolge, mit Spalte n beginnend, durchzuführen.
Satz 9.60 sagt aus, dass in einem QR-Schritt die Matrix $\boldsymbol{Q}^{(k)}$ und damit die Matrizen $\boldsymbol{A}^{(k+1)}$ und $\boldsymbol{V}^{(k+1)}$ im wesentlichen festgelegt sind, falls $\boldsymbol{A}^{(k)}$ nichtzerfallend ist und die erste Spalte von $\boldsymbol{Q}^{(k)}$ festgelegt ist. Die Matrix $\boldsymbol{Q}^{(k)}$ wird im Algorithmus durch ein Produkt von $n-1$ GIVENS-Transformationen dargestellt. Damit ergibt sich

$$\boldsymbol{Q}^{(k)}\boldsymbol{e}_1 = \boldsymbol{G}_{12}^T\boldsymbol{G}_{23}^T\cdots\boldsymbol{G}_{n-1,n}^T\boldsymbol{e}_1 = \boldsymbol{G}_{12}^T\boldsymbol{e}_1.$$

Die Matrix $\boldsymbol{G}_{12}$ legt im wesentlichen den gesamten Transformationsschritt fest. Bestimmen wir nun weitere Matrizen $\bar{\boldsymbol{G}}_{23}, \ldots, \bar{\boldsymbol{G}}_{n-1,n}$ und damit eine Matrix

$$\bar{\boldsymbol{Q}}^{(k)} = \boldsymbol{G}_{12}\bar{\boldsymbol{G}}_{23}\cdots\bar{\boldsymbol{G}}_{n-1,n}$$

so, dass

$$\bar{\boldsymbol{A}}^{(k+1)} = \bar{\boldsymbol{Q}}^{(k)^T}\boldsymbol{A}^{(k)}\bar{\boldsymbol{Q}}^{(k)} = \bar{\boldsymbol{G}}_{n-1,n}\cdots\bar{\boldsymbol{G}}_{23}\boldsymbol{G}_{12}\boldsymbol{A}^{(k)}\boldsymbol{G}_{12}^T\bar{\boldsymbol{G}}_{23}^T\cdots\bar{\boldsymbol{G}}_{n-1,n}^T$$

tridiagonal ist, so stimmen die Matrizen $\bar{\boldsymbol{A}}^{(k+1)}$ und $\bar{\boldsymbol{Q}}^{(k)}$ im Sinne von Satz 9.60 im wesentlichen mit den Matrizen $\boldsymbol{A}^{(k+1)}$ und $\boldsymbol{Q}^{(k)}$ überein. Betrachtet man nun die Matrix $\boldsymbol{M}_2 = \boldsymbol{G}_{12}\boldsymbol{A}^{(k)}\boldsymbol{G}_{12}^T$, so erkennt man, dass sie sich von einer Tridiagonalmatrix nur in den Nichtnullelementen an den Positionen $(1,3)$ und $(3,1)$ unterscheidet.

$$\boldsymbol{M}_2 = \begin{pmatrix} \times & \times & \otimes & & & & \\ \times & \times & \times & & & & \\ \otimes & \times & \times & \times & & & \\ & & \times & \times & \times & & \\ & & & \ddots & \ddots & \ddots & \\ & & & & \times & \times & \times \\ & & & & & \times & \times \end{pmatrix}$$

Die nächste GIVENS-Transformation $\bar{\boldsymbol{G}}_{23}$ wird nun so festgelegt, dass durch sie die störenden Nichtnullelemente eliminiert werden. Dabei werden neue Nichtnullelemente an den Positionen $(2,4)$ und $(4,2)$ erzeugt, so dass die Matrix $\boldsymbol{M}_3 = \bar{\boldsymbol{G}}_{23}\boldsymbol{M}_2\bar{\boldsymbol{G}}_{23}^T = \bar{\boldsymbol{G}}_{23}\boldsymbol{G}_{12}\boldsymbol{A}^{(k)}\boldsymbol{G}_{12}^T\bar{\boldsymbol{G}}_{23}^T$ folgende Struktur besitzt:

$$\boldsymbol{M}_3 = \begin{pmatrix} \times & \times & & & & & \\ \times & \times & \times & \otimes & & & \\ & \times & \times & \times & & & \\ & \otimes & \times & \times & \times & & \\ & & & \ddots & \ddots & \ddots & \\ & & & & \times & \times & \times \\ & & & & & \times & \times \end{pmatrix}$$

Die in diesem Schritt entstandenen neuen Nichtnullelemente werden mit der nächsten GIVENS-Transformation in analoger Weise eliminiert. Nach $n-3$ GIVENS-Drehungen

$$\bar{\boldsymbol{G}}_{23}, \bar{\boldsymbol{G}}_{34}, \ldots, \bar{\boldsymbol{G}}_{n-2,n-1}$$

ist dann die Matrix $\boldsymbol{M}_{n-2}$ entstanden, die sich von einer Tridiagonalmatrix nur durch die zusätzlichen Elemente an den Positionen $(n-2,n)$ und $(n,n-2)$ unterscheidet. Diese beiden Nichtnullelemente lassen sich durch eine letzte GIVENS-Drehung $\bar{\boldsymbol{G}}_{n-1,n}$ beseitigen, so dass dann $\boldsymbol{M}_{n-1} = \boldsymbol{A}^{(k+1)}$ wieder tridiagonal ist. Bei dieser Variante des QR-Algorithmus muss die Verschiebung μ_k nicht mehr explizit ausgeführt werden. Sie wird nur im ersten Schritt zur Festlegung der Matrix $\boldsymbol{G}_{12}$ benötigt. Darum bezeichnet man dieses Vorgehen auch als Impliziten QR-Schritt.

9.61. Impliziter QR-Schritt:

S0 *(Initialisierung) Wähle Verschiebungsparameter μ_k.*
Setze $\bar{\boldsymbol{A}} = \boldsymbol{A}^{(k)}$ und $\bar{\boldsymbol{V}} = \boldsymbol{V}^{(k)}$.

S1 *(Start-Transformation)*

- *Lege $\boldsymbol{G}_{12}$ so fest, dass $[\boldsymbol{G}_{12}(\bar{\boldsymbol{A}} - \mu_k \boldsymbol{I})]_{21} = 0$ gilt.*
- *Bilde $\bar{\boldsymbol{A}} = \boldsymbol{G}_{12}\bar{\boldsymbol{A}}\boldsymbol{G}_{12}^T$ und $\bar{\boldsymbol{V}} = \bar{\boldsymbol{V}}\boldsymbol{G}_{12}^T$.*

S2 *(Rest-Transformation) Für $j = 2, \ldots, n-1$ führe aus:*

- *Lege $\boldsymbol{G}_{j,j+1}$ so fest, dass $(\boldsymbol{G}_{j,j+1}\bar{\boldsymbol{A}})_{j+1,j-1} = 0$ gilt.*
- *Bilde $\bar{\boldsymbol{A}} = \boldsymbol{G}_{j,j+1}\bar{\boldsymbol{A}}\boldsymbol{G}_{j,j+1}^T$ und $\bar{\boldsymbol{V}} = \bar{\boldsymbol{V}}\boldsymbol{G}_{j,j+1}^T$.*

S3 *Setze $\boldsymbol{A}^{(k+1)} = \bar{\boldsymbol{A}}$ und $\boldsymbol{V}^{(k+1)} = \bar{\boldsymbol{V}}$.*

Aufwand:

- *$\sim 6n$ Add./Sub. + $\sim 11n$ Mult./Div. + n Quadratwurzeln zum Berechnen von $\boldsymbol{A}^{(k+1)}$.*
- *$\sim 3n^2$ Add./Sub. + $\sim 3n^2$ Mult./Div. zum Berechnen von $\boldsymbol{V}^{(k+1)}$.*

Bemerkungen: (i) Der wesentliche Vorteil des Impliziten QR-Schritts liegt darin, dass $\boldsymbol{A}^{(k)} - \mu_k \boldsymbol{I}$ nicht explizit berechnet werden muss. Dadurch kann es nicht zu einem Informationsverlust im Falle $|\mu_k| \gg |a_{jj}^{(k)}|$ kommen. Die Verschiebung μ_k wird nur bei der Festlegung der ersten GIVENS-Drehung $\boldsymbol{G}_{12}$ benötigt.
(ii) Nach jedem Schritt sollten alle Nichtdiagonalelemente auf Kleinheit überprüft werden. Hinreichend kleine Elemente sollten Null gesetzt werden. Das führt nicht nur zu einer deutlichen Reduktion des Aufwands, sondern auch zu einer Verbesserung des Konvergenzverhaltens, da durch die Aufspaltung in Teilprobleme geringerer Dimension für diese Probleme besser angepasste Verschiebungsparameter wählbar sind.
(iii) Ein mögliches Kriterium zum Nullsetzen von Subdiagonalelementen wäre

$$|b_i^{(k)}| \leqq \text{eps}(|a_{i-1}^{(k)}| + |a_i^{(k)}|)$$

für $i \in \{2, \ldots, n\}$, wobei

$$\boldsymbol{A}^{(k)} = \text{trid}(a_1^{(k)}, \ldots, a_n^{(k)}, b_2^{(k)}, \ldots, b_n^{(k)})$$

gilt. Ein aufwendigeres, aber besseres Kriterium ist

$$|b_i^{(k)}| \leqq \text{eps}(\beta_i^{(k)} + \gamma_i^{(k)})(1 + \alpha_i^{(k)}/\beta_i^{(k)}), \quad i \in \{2, \ldots, n\}$$

mit $\alpha_i^{(k)} = |a_{i-1}^{(k)}| + |a_i^{(k)}|$ und

$$\gamma_i^{(k)} = |a_{i-1}^{(k)} - a_i^{(k)}|, \quad i \in \{2, \ldots, n\}$$

sowie

$$\beta_i^{(k)} = |b_{i-1}^{(k)}| + |b_i^{(k)}| + |b_{i+1}^{(k)}|, \quad i \in \{2, \ldots, n-1\}$$

und

$$\beta_n^{(k)} = |b_{n-1}^{(k)}| + |b_n^{(k)}|.$$

Dieser Test sollte nur ausgeführt werden, falls $|b_i^{(k)}| \leq \sqrt{\text{eps}}\,\|\boldsymbol{A}\|_\infty$ gilt.
Nach hinreichend vielen Schritten sind die Nullsetzungskriterien für alle Nichtdiagonalelemente erfüllt. Dann ist $\boldsymbol{A}^{(k)}$ diagonal, und die Diagonalelemente approximieren die Eigenwerte von $\boldsymbol{A}$. Die Spalten von $\boldsymbol{V}^{(k)}$ sind die zugehörigen Eigenvektornäherungen zur Startmatrix $\boldsymbol{A}$. Wurde die Matrix $\boldsymbol{A}$ durch orthogonale Tridiagonalisierung aus einer Matrix $\bar{\boldsymbol{A}}$ gemäß $\bar{\boldsymbol{A}} = \boldsymbol{Q}\boldsymbol{A}\boldsymbol{Q}^T$ gewonnen, so sind die Eigenvektornäherungen für $\bar{\boldsymbol{A}}$ durch $\boldsymbol{Q}\boldsymbol{V}^{(k)}$ gegeben. Es bietet sich daher an, in diesem Falle den Algorithmus mit $\boldsymbol{V}^{(0)} = \boldsymbol{Q}$ statt mit $\boldsymbol{V}^{(0)} = \boldsymbol{I}$ zu starten.
Insgesamt ergibt sich folgendes Grobschema zur Realisierung des QR-Algorithmus für eine Tridiagonalmatrix $\boldsymbol{T}$. Die Indizes $1 \leq p \leq q \leq n$ geben dabei die Position des gerade zu bearbeitenden untersten nichtzerfallenden Diagonalblockes von $\boldsymbol{T}$ an.

9.62. QR**-Algorithmus für Tridiagonalmatrizen:**
Gegeben seien die (n,n)*-Tridiagonalmatrix*

$$\boldsymbol{T} = \text{trid}(a_1, \ldots, a_n, b_2, \ldots, b_n)$$

und die orthogonale (n,n)*-Matrix* $\boldsymbol{V}$*.*

S0 *(Initialisierung) Setze k=1, q=n.*

S1 *(Abbruchtest) Falls* $q = 1$*: STOPP.*

S2 *(Nullsetzen kleiner Nichtdiagonalelemente und Festlegen des zu aktuellen Diagonalblocks*

$$\boldsymbol{T}^{(p,q)} = \text{trid}(a_p, \ldots, a_q, b_{p+1}, \ldots, b_q)$$

der Dimension $m = q - p + 1$ *von* $\boldsymbol{T}$*)*
Für p=q, q-1, . . . ,2 führe aus:

- *Ist* $|b_p|$ *hinreichend klein, so*
  - *setze* $b_p = 0$*.*
  - *Ist* $p = q$*, so setze* $q = q - 1$ *und gehe zu Schritt* **S1***.*
  - *Ist* $p < q$*, so gehe zu Schritt* **S3***.*
- *Setze* $p = 1$*.*

S3 *(k-ter QR-Schritt) Führe QR-Schritt mit* WILKINSON*-Verschiebung für*

$$\boldsymbol{A}^{(k)} = \boldsymbol{T}^{(p,q)} \in \mathbb{R}^{m\times m} \quad , \quad \boldsymbol{V}^{(k)} = \boldsymbol{V}^{(p,q)} \in \mathbb{R}^{n\times m}$$

auf dem Platz von $\boldsymbol{T}$ bzw. $\boldsymbol{V}$ durch. Dabei bezeichnet $\boldsymbol{V}^{(p,q)}$ die aus den Spalten p bis q von $\boldsymbol{V}$ gebildete spaltenorthonormale Matrix.

S4 *Setze $k = k+1$ und gehe zu Schritt* **S1**.

Bemerkungen: (i) Praktische Erfahrungen zeigen, dass man im Mittel etwa $k \approx 1.7n$ QR-Schritte zur Diagonalisierung einer Tridiagonalmatrix $\boldsymbol{T}$ benötigt. Das Berechnen der ersten Eigenwerte beansprucht dabei etwas mehr Schritte (etwa 4 bis 6), die letzten Eigenwerte werden nach höchstens ein bis zwei Schritten abgespalten. Damit beträgt der Gesamtaufwand zur Eigenwertberechnung einer Tridiagonalmatrix $\sim 10n^2$ Additionen und Multiplikationen und $\sim n^2$ Quadratwurzeln. Meistens ist der Aufwand sogar noch geringer. Will man auch alle Eigenvektoren berechnen, so kostet die Aufdatierung von $\boldsymbol{V}$ $\sim 3n^3$ Additionen und Multiplikationen, ist daher wesentlich aufwendiger.
(ii) Für eine vollbesetzte Matrix ist natürlich noch der Aufwand für die Tridiagonalisierung zu berücksichtigen. Hierbei kostet das Berechnen der Transformation $\boldsymbol{T} = \boldsymbol{Q}^T\boldsymbol{A}\boldsymbol{Q}$ ungefähr $2n^3/3$ Additionen und Multiplikationen. Die Bildung der Matrix $\boldsymbol{Q}$, die man für die Ermittlung der Eigenvektoren benötigt, kostet weitere $\sim 2n^3/3$ Additionen und Multiplikationen. Damit beträgt der Gesamtaufwand für das Berechnen der Eigenwerte einer vollbesetzten Matrix mit Hilfe des QR-Algorithmus mehr als $2n^3/3$ Additionen und Multiplikationen und $\sim n^2$ Quadratwurzeln. Das Lösen des vollständigen Eigenwertproblems für eine vollbesetzte symmetrische Matrix benötigt dagegen mindestens $11n^3/3$ Additionen und Multiplikationen, ist also wesentlich teurer.
(iii) Sind nur wenige Eigenvektoren zu berechnen, so erhält man diese billiger mittels Inverser Iteration. Diese benötigt pro Eigenvektor einer Tridiagonalmatrix $\sim Kn$ Additionen und Multiplikationen. Verwendet man für $\boldsymbol{Q}$ die Darstellung als Produkt von GIVENS-Drehungen, so benötigt das Berechnnen eines Eigenvektors der Ausgangsmatrix $\boldsymbol{A}$ aus einem Eigenvektor der Tridiagonalmatrix $\boldsymbol{T}$ $\sim n^2$ Additionen und Multiplikationen. Beim Berechnen von Eigenvektoren zu dicht benachbarten Eigenwerten ist dabei eventuell eine Reorthogonalisierung notwendig.
(iv) Für eine Bandmatrix $\boldsymbol{A}$ wird der QR-Algorithmus direkt mit $\boldsymbol{A}$ ohne vorherige Tridiagonalisierung durchgeführt, da die Bandgestalt invariant unter den QR-Transformationen ist.
(v) Man kann zeigen, dass die bei Abbruch nach k Schritten vorliegende Matrix $\boldsymbol{M} = \boldsymbol{T}^{(k)}$ die exakten Eigenwerte einer gestörten Matrix $\boldsymbol{A} + \boldsymbol{\delta A}$ enthält. Es gilt

$$\boldsymbol{A} + \boldsymbol{\delta A} = \boldsymbol{V}\boldsymbol{M}\boldsymbol{V}^T$$

mit der symmetrischen (n,n)-Störung $\boldsymbol{\delta A}$, die der Abschätzung

$$\|\boldsymbol{\delta A}\|_2 \leqq \text{eps}F\|\boldsymbol{A}\|_2$$

genügt. Dabei ist $\boldsymbol{V}$ eine exakt orthogonale Matrix. Das Berechnen der Eigenwerte der Matrix $\boldsymbol{A}$ mit dem QR-Algorithmus ist demnach ein numerisch gutartiger Prozess. Die berechnete Matrix $\tilde{\boldsymbol{V}}$ der Eigenvektoren approximiert dabei die exakte Matrix $\boldsymbol{V}$ im Sinne von

$$\|\boldsymbol{V} - \tilde{\boldsymbol{V}}\|_2 \leqq \text{eps}F_1$$

hinreichend genau.
(vi) Es besteht auch die Möglichkeit, den Algorithmus in inverser Reihenfolge durchzuführen. Dabei werden die Verschiebungen jeweils durch die links oben stehende 2×2-Matrix festgelegt. Statt der QR-Zerlegung der Matrix $\boldsymbol{A}^{(k)} - \mu_k \boldsymbol{I}$ wird dann eine QL-Faktorisierung $\boldsymbol{A}^{(k)} - \mu_k \boldsymbol{I} = \boldsymbol{Q}^{(k)}\boldsymbol{L}^{(k)}$ mit einer unteren Dreiecksmatrix $\boldsymbol{L}^{(k)}$ durchgeführt. Diese Variante wird auch QL-Algorithmus genannt und heute in Programmpaketen häufig angewendet.

Kapitel 10

Lineare Ausgleichsprobleme

10.1. Problemstellung und klassische Lösung

Viele Anwendungen in der Physik, Technik oder Ökonomie führen auf folgende Problemstellung:
Zwischen zwei Größen y und z wird ein funktionaler Zusammenhang der Art

$$y = f(z; x_1, \ldots, x_n)$$

angenommen. Die Variablen $x_1, \ldots, x_n$ sind unbekannte Parameter, die zu bestimmen sind. Dazu führt man m „Messungen“ durch. Man erhält Wertepaare (z_i, y_i) $(i = 1, \ldots, m)$ und m Gleichungen

$$y_i = f(z_i; x_1, \ldots, x_n), \quad i = 1, \ldots, m,$$

um aus diesen die unbekannten Parameter zu ermitteln. Für ein exaktes Bestimmen der Parameter sind mindestens n Messungen nötig. Da aber die Messungen meist fehlerbehaftet sind, wird man mehr als n Messungen durchführen, um eine größere statistische Sicherheit zu bekommen. Damit ergibt sich ein überbestimmtes Gleichungssystem, das im allgemeinen keine Lösung besitzt. Man wird darum das Gleichungssystem „so gut wie möglich“ lösen wollen. Das führt auf einen neuen Lösungsbegriff.
Es seien $\boldsymbol{y} = (y_1, \ldots, y_m)^T \in \mathbb{R}^m$, $\boldsymbol{z} = (z_1, \ldots, z_m)^T \in \mathbb{R}^m$, $\boldsymbol{x} = (x_1, \ldots, x_n)^T \in \mathbb{R}^n$ und

$$\boldsymbol{F}(\boldsymbol{z}; \boldsymbol{x}) = \begin{pmatrix} f(z_1; x_1, \ldots, x_n) \\ f(z_2; x_1, \ldots, x_n) \\ \vdots \\ f(z_m; x_1, \ldots, x_n) \end{pmatrix} = \begin{pmatrix} f_1(\boldsymbol{x}) \\ f_2(\boldsymbol{x}) \\ \vdots \\ f_m(\boldsymbol{x}) \end{pmatrix}.$$

Das überbestimmte Gleichungssystem lautet dann

$$\boldsymbol{y} = \boldsymbol{F}(\boldsymbol{z}; \boldsymbol{x}).$$

Es liegt nahe, den Vektor $\boldsymbol{x}$ so zu bestimmen, dass das Residuum

$$\boldsymbol{r}(\boldsymbol{x}) = \boldsymbol{y} - \boldsymbol{F}(\boldsymbol{z};\boldsymbol{x})$$

möglichst klein wird. Wir lösen daher folgende Ersatzaufgabe

$$\boldsymbol{x}^* = \arg\min_{\boldsymbol{x}\in\mathbb{R}^n} \|\boldsymbol{r}(\boldsymbol{x})\|,$$

wobei $\|\circ\|$ eine beliebige Vektornorm bezeichnen soll. Dieser neue Lösungsbegriff für das Gleichungssystem hat den Vorteil, dass er den alten Lösungsbegriff einschließt. Im Falle exakter Lösbarkeit stimmt $\boldsymbol{x}^*$ mit der exakten Lösung überein.
Je nach Wahl der Norm ergeben sich verschiedene Aufgabenstellungen.
Verwendet man die Maximumnorm, so ergibt sich

$$\|\boldsymbol{r}(\boldsymbol{x})\|_\infty = \max_{k=1,\ldots,n} |y_k - f(z_k;x_1,\ldots,x_n)| = \max_{k=1,\ldots,n} |y_k - f_k(\boldsymbol{x})|$$

und

$$\boldsymbol{x}^* = \arg\min_{\boldsymbol{x}\in\mathbb{R}^n} \max_{k=1,\ldots,n} |y_k - f_k(\boldsymbol{x})|\,.$$

Diese Aufgabe bezeichnet man als **Diskretes** TSCHEBYSCHEFF**-Problem**.
Wir werden jedoch die euklidische Vektornorm

$$\|\boldsymbol{r}(\boldsymbol{x})\|_2^2 = \sum_{k=1}^{n} [y_k - f(z_k;x_1,\ldots,x_n)]^2 = \sum_{k=1}^{n} [y_k - f_k(\boldsymbol{x})]^2$$

verwenden. Hier gilt

$$\boldsymbol{x}^* = \arg\min_{\boldsymbol{x}\in\mathbb{R}^n} \sum_{k=1}^{n} [y_k - f_k(\boldsymbol{x})]^2\,.$$

Dieses Problem bezeichnet man als **Ausgleichsproblem im engeren Sinne**. Es wurde schon von GAUSS als Methode der kleinsten Quadrate untersucht.
Besitzen die Funktionen $f_k(\boldsymbol{x}) = f(z_k;x_1,\ldots,x_n)$ stetige partielle Ableitungen nach den x_i, so lassen sich sofort notwendige Bedingungen für die Existenz eines Minimums angeben:

$$\frac{\partial}{\partial x_i} \sum_{k=1}^{n} [y_k - f_k(x_1,\ldots,x_n)]^2 = 2\sum_{k=1}^{n} \frac{\partial}{\partial x_i} f_k(x_1,\ldots,x_n)\,[y_k - f_k(x_1,\ldots,x_n)] = 0$$

für $i = 1,\ldots,n$. Diese Gleichungen heißen **Normalgleichungen**. Besonders einfach wird die Situation, falls die Funktion f linear von den Parametern x_i abhängt. Wir stellen dann f in der Form

$$f(z;\boldsymbol{x}) = x_1\varphi_1(z) + x_2\varphi_2(z) + \cdots + x_n\varphi_n(z)$$

und die f_k für $k = 1, \ldots, m$ in der Form

$$f_k(\boldsymbol{x}) = x_1\varphi_1(z_k) + x_2\varphi_2(z_k) + \cdots + x_n\varphi_n(z_k)$$

dar. Mit den Bezeichnungen $a_{ki} = \varphi_i(z_k)$ für $i = 1, \ldots, n$ und $k = 1, \ldots, m$ gilt dann

$$\boldsymbol{F}(\boldsymbol{z}; \boldsymbol{x}) = \boldsymbol{A}\boldsymbol{x}$$

mit

$$\boldsymbol{A} = \begin{pmatrix} a_{11} & \ldots & a_{1n} \\ \vdots & \ddots & \vdots \\ a_{m1} & \ldots & a_{mn} \end{pmatrix}.$$

Wir erhalten das **lineare Ausgleichsproblem**

$$\boldsymbol{x}^* = \arg\min_{\boldsymbol{x}\in\mathbb{R}^n} \|\boldsymbol{y} - \boldsymbol{A}\boldsymbol{x}\|_2^2$$

mit den Normalgleichungen

$$\boldsymbol{A}^T(\boldsymbol{y} - \boldsymbol{A}\boldsymbol{x}) = \boldsymbol{o}$$

beziehungsweise

$$\boldsymbol{A}^T\boldsymbol{A}\boldsymbol{x} = \boldsymbol{A}^T\boldsymbol{y}.$$

Dieses Gleichungssystem ist stets lösbar (Übungsaufgabe 1).
Im Falle $\mathrm{rg}(\boldsymbol{A}) = n$ ($\boldsymbol{A}$ ist **spaltenregulär**) ist die Matrix $\boldsymbol{A}^T\boldsymbol{A}$ regulär und darüber hinaus sogar positiv definit (Übungsaufgabe 2). In diesem Falle ist das Normalgleichungssystem eindeutig lösbar. Wir wenden zum Lösen das CHOLESKY-Verfahren an. Spätere Überlegungen werden jedoch zeigen, dass dies nicht der günstigste Weg ist.
Zunächst wollen wir jedoch den Zusammenhang zwischen dem eigentlichen Ausgleichsproblem und den Normalgleichungen näher beleuchten.

10.1. Satz: *Das lineare Ausgleichsproblem*

$$\min_{\boldsymbol{x}\in\mathbb{R}^n} \|\boldsymbol{y} - \boldsymbol{A}\boldsymbol{x}\|_2$$

besitzt mindestens eine Lösung $\boldsymbol{x}^*$. Ist $\bar{\boldsymbol{x}}$ eine weitere Lösung, so gilt $\boldsymbol{A}\boldsymbol{x}^* = \boldsymbol{A}\bar{\boldsymbol{x}}$. Das Residuum $\boldsymbol{r} = \boldsymbol{y} - \boldsymbol{A}\boldsymbol{x}^*$ genügt der Gleichung $\boldsymbol{A}^T\boldsymbol{r} = \boldsymbol{o}$. Jede Lösung $\boldsymbol{x}^*$ des Ausgleichsproblems ist auch Lösung der Normalgleichungen und umgekehrt.

Beweis: Es sei

$$\boldsymbol{\mathcal{V}} = \text{span}(\boldsymbol{A}) = \{\, \boldsymbol{Ax} \mid \boldsymbol{x} \in \mathbb{R}^n \,\} \subseteq \mathbb{R}^m$$

der Teilraum, der von den Spalten von $\boldsymbol{A}$ aufgespannt wird. Weiterhin bezeichne

$$\boldsymbol{\mathcal{V}}^\perp = \left\{ \boldsymbol{r} \in \mathbb{R}^m \mid \boldsymbol{r}^T\boldsymbol{z} = 0 \,\forall \boldsymbol{z} \in \boldsymbol{\mathcal{V}} \right\} = \left\{ \boldsymbol{r} \in \mathbb{R}^m \mid \boldsymbol{r}^T\boldsymbol{A} = \boldsymbol{o} \right\}$$

den zu $\boldsymbol{\mathcal{V}}$ gehörenden Orthogonalraum. Wegen $\mathbb{R}^m = \boldsymbol{\mathcal{V}} \oplus \boldsymbol{\mathcal{V}}^\perp$ lässt sich der Vektor $\boldsymbol{y} \in \mathbb{R}^m$ eindeutig in der Form $\boldsymbol{y} = \boldsymbol{s} + \boldsymbol{r}$ mit $\boldsymbol{s} \in \boldsymbol{\mathcal{V}}$ und $\boldsymbol{r} \in \boldsymbol{\mathcal{V}}^\perp$ zerlegen. Es gibt mindestens ein $\boldsymbol{x}^* \in \mathbb{R}^n$ mit $\boldsymbol{s} = \boldsymbol{Ax}^*$. Wegen $\boldsymbol{A}^T\boldsymbol{r} = \boldsymbol{o}$ gilt

$$\boldsymbol{A}^T\boldsymbol{y} = \boldsymbol{A}^T\boldsymbol{s} = \boldsymbol{A}^T\boldsymbol{Ax}^*.$$

$\boldsymbol{x}^*$ ist somit Lösung der Normalgleichungen. Umgekehrt entspricht jeder Lösung $\bar{\boldsymbol{x}}$ der Normalgleichungen die Zerlegung

$$\boldsymbol{y} = \bar{\boldsymbol{s}} + \bar{\boldsymbol{r}}, \bar{\boldsymbol{s}} = \boldsymbol{A}\bar{\boldsymbol{x}} \in \boldsymbol{\mathcal{V}}, \quad \bar{\boldsymbol{r}} = \boldsymbol{y} - \boldsymbol{A}\bar{\boldsymbol{x}} \in \boldsymbol{\mathcal{V}}^\perp.$$

Wegen der Eindeutigkeit der Zerlegung muss aber $\boldsymbol{s} = \bar{\boldsymbol{s}}$ und damit $\boldsymbol{Ax}^* = \boldsymbol{A}\bar{\boldsymbol{x}}$ gelten. Weiterhin ist jede Lösung $\boldsymbol{x}^*$ der Normalgleichungen Optimallösung des Ausgleichsproblems. Ist nämlich $\boldsymbol{x}$ beliebig und $\boldsymbol{z} = \boldsymbol{Ax} - \boldsymbol{Ax}^* \in \boldsymbol{\mathcal{V}}$ und $\boldsymbol{r} = \boldsymbol{y} - \boldsymbol{Ax}^*$, so gilt wegen $\boldsymbol{z}^T\boldsymbol{r} = 0$

$$\|\boldsymbol{y} - \boldsymbol{Ax}\|_2^2 = \|\boldsymbol{r} - \boldsymbol{z}\|_2^2 = \|\boldsymbol{r}\|_2^2 + \|\boldsymbol{z}\|_2^2 \geqq \|\boldsymbol{r}\|_2^2 = \|\boldsymbol{y} - \boldsymbol{Ax}^*\|_2^2.$$

Damit ist die Existenz einer Optimallösung und die Eindeutigkeit des zugehörigen Residuums bewiesen. ❀

Im spaltenregulären Falle ist das Ausgleichsproblem mittels einer CHOLESKYZerlegung der Matrix $\boldsymbol{A}^T\boldsymbol{A}$ lösbar. Eine andere (mehr theoretische) Möglichkeit ergibt sich aus der Singulärwertzerlegung der Matrix $\boldsymbol{A}$. Es sei $\boldsymbol{A} = \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^T$ mit $\boldsymbol{\Sigma} = \text{diag}(\sigma_1, \ldots, \sigma_l)$, $l = \min\{m, n\}$, die Singulärwertzerlegung von $\boldsymbol{A}$. Für die Singulärwerte gelte

$$\sigma_1 \geqq \cdots \geqq \sigma_r > \sigma_{r+1} = \cdots = \sigma_l = 0$$

mit $r = \text{rg}(\boldsymbol{A})$. Wegen der Invarianz der euklidischen Vektornorm gegenüber orthogonalen Transformationen gilt

$$\|\boldsymbol{y} - \boldsymbol{Ax}\|_2^2 = \left\|\boldsymbol{y} - \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^T\boldsymbol{x}\right\|_2^2 = \left\|\boldsymbol{U}^T\left(\boldsymbol{y} - \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^T\boldsymbol{x}\right)\right\|_2^2 = \left\|\boldsymbol{U}^T\boldsymbol{y} - \boldsymbol{\Sigma}\boldsymbol{V}^T\boldsymbol{x}\right\|_2^2.$$

Mit $\boldsymbol{\eta} = \boldsymbol{U}^T\boldsymbol{y}$ und $\boldsymbol{\xi} = \boldsymbol{V}^T\boldsymbol{x}$ erhalten wir das äquivalente Ausgleichsproblem

$$\min_{\boldsymbol{\xi}\in\mathbb{R}^n} \|\boldsymbol{\eta} - \boldsymbol{\Sigma}\boldsymbol{\xi}\|_2.$$

Hier gilt aber

$$\|\boldsymbol{\eta} - \boldsymbol{\Sigma}\boldsymbol{\xi}\|_2^2 = \sum_{i=1}^{r}(\eta_i - \sigma_i\xi_i)^2 + \sum_{i=r+1}^{m}\eta_i^2.$$

Der zweite Summand hängt nicht von $\boldsymbol{\xi}$ ab; er spielt daher bei der Minimierung keine Rolle. Für den ersten Summanden gilt

$$\sum_{i=1}^{r}(\eta_i - \sigma_i\xi_i)^2 \geqq 0$$

und

$$\sum_{i=1}^{r}(\eta_i - \sigma_i\xi_i)^2 = 0 \quad \Longleftrightarrow \quad \xi_i = \frac{\eta_i}{\sigma_i}, \quad i = 1,\ldots,r.$$

Wir erhalten die Lösung des Ausgleichsproblems in der Form

$$\boldsymbol{x}^* = \boldsymbol{V}\boldsymbol{\xi}^*$$

mit

$$\boldsymbol{\xi}^* = \left(\frac{\eta_1}{\sigma_1},\ldots,\frac{\eta_r}{\sigma_r},\xi_{r+1}^*,\ldots,\xi_n^*\right)^T.$$

Im spaltenregulären Falle ($r = n$) ist die Lösung eindeutig bestimmt. Für $r < n$ sind die $\xi_{r+1}^*,\ldots,\xi_n^*$ frei wählbar. Setzt man $\xi_{r+1}^* = \cdots = \xi_n^* = 0$, so erhält man die normkleinste Lösung. Für das Residuum gilt

$$\boldsymbol{r}^* = \boldsymbol{y} - \boldsymbol{A}\boldsymbol{x}^* = \boldsymbol{U}\boldsymbol{\eta} - \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^T\boldsymbol{V}\boldsymbol{\xi}^* = \boldsymbol{U}(\boldsymbol{\eta} - \boldsymbol{\Sigma}\boldsymbol{\xi}^*) = \boldsymbol{U}\begin{pmatrix} 0 \\ \vdots \\ 0 \\ \eta_{r+1} \\ \vdots \\ \eta_m \end{pmatrix}.$$

Die normkleinste Lösung des Ausgleichsproblems lässt sich formal als

$$\boldsymbol{x}^* = \boldsymbol{A}^+\boldsymbol{y}$$

schreiben. Die Matrix $\boldsymbol{A}^+$ wird als **Pseudoinverse** bezeichnet. Sie ist für eine beliebige (m,n)-Matrix $\boldsymbol{A}$ mit der Singulärwertzerlegung $\boldsymbol{A} = \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^T$ durch

$$\boldsymbol{A}^+ = \boldsymbol{V}\boldsymbol{\Sigma}^+\boldsymbol{U}^T$$

definiert. Dabei gilt für

$$\boldsymbol{\Sigma} = \operatorname{diag}(\sigma_1, \ldots, \sigma_r, 0, \ldots, 0)$$

$$\boldsymbol{\Sigma}^+ = \operatorname{diag}(1/\sigma_1, \ldots, 1/\sigma_r, 0, \ldots, 0) \in \mathbb{R}^{n\times m}$$

mit $r = \operatorname{rg}(\boldsymbol{A})$. Somit für

$$\boldsymbol{\Sigma} = \left(\begin{array}{ccc|c} \sigma_1 & & & \\ & \ddots & & \boldsymbol{O} \\ & & \sigma_r & \\ \hline & \boldsymbol{O} & & \boldsymbol{O} \end{array}\right) \in \mathbb{R}^{m\times n}$$

ergibt sich

$$\boldsymbol{\Sigma}^+ = \left(\begin{array}{ccc|c} \frac{1}{\sigma_1} & & & \\ & \ddots & & \boldsymbol{O} \\ & & \frac{1}{\sigma_r} & \\ \hline & \boldsymbol{O} & & \boldsymbol{O} \end{array}\right) \in \mathbb{R}^{n\times m}.$$

Für $m = n = r$ gilt $\boldsymbol{A}^+ = \boldsymbol{A}^{-1}$, so dass die Pseudoinverse als Verallgemeinerung der gewöhnlichen Inversen einer Matrix anzusehen ist. Eine etwas andere Charakterisierung der Pseudoinversen ist im folgenden Satz angegeben.

10.2. Satz:

1. *Zu jeder (m,n)-Matrix $\boldsymbol{A}$ gibt es genau eine (n,m)-Matrix $\boldsymbol{A}^+$, so dass*

   $$\boldsymbol{x}^* = \boldsymbol{A}^+\boldsymbol{y}$$

   die normkleinste Lösung (Normallösung) des linearen Ausgleichsproblems

   $$\min_{\boldsymbol{x}\in\mathbb{R}^n} \|\boldsymbol{y} - \boldsymbol{A}\boldsymbol{x}\|_2$$

   ist. Ist die Singulärwertzerlegung von $\boldsymbol{A}$ durch $\boldsymbol{A} = \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^T$ mit

   $$\boldsymbol{\Sigma} = \operatorname{diag}(\sigma_1, \ldots, \sigma_r, 0, \ldots, 0)$$

und

$$\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_r > 0$$

gegeben, so gilt

$$\boldsymbol{A}^+ = \boldsymbol{V}\boldsymbol{\Sigma}^+\boldsymbol{U}^T$$

mit

$$\boldsymbol{\Sigma}^+ = \text{diag}(1/\sigma_1, \ldots, 1/\sigma_r, 0, \ldots, 0).$$

2. $\boldsymbol{A}^+$ *ist durch die* PENROSE*-Bedingungen*
   (a) $\boldsymbol{A}\boldsymbol{A}^+\boldsymbol{A} = \boldsymbol{A}$,
   (b) $\boldsymbol{A}^+\boldsymbol{A}\boldsymbol{A}^+ = \boldsymbol{A}^+$,
   (c) $(\boldsymbol{A}\boldsymbol{A}^+)^T = \boldsymbol{A}\boldsymbol{A}^+$,
   (d) $(\boldsymbol{A}^+\boldsymbol{A})^T = \boldsymbol{A}^+\boldsymbol{A}$
   eindeutig festgelegt.

Beweis:

1. Wurde bereits gezeigt.
2. Die Gültigkeit der PENROSE-Bedingungen für

   $$\boldsymbol{A} = \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^T, \quad \boldsymbol{A}^+ = \boldsymbol{V}\boldsymbol{\Sigma}^+\boldsymbol{U}^T$$

   zeigt man durch einfaches Nachrechnen. Es bleibt noch zu beweisen, dass die Pseudoinverse durch die PENROSE-Bedingungen eindeutig festgelegt ist. Dazu nehmen wir an, dass die Matrix $\boldsymbol{B}$ ebenfalls die PENROSE-Bedingungen erfüllt. Dann gilt

$$\begin{aligned}
\boldsymbol{B} &= \boldsymbol{B}\boldsymbol{A}\boldsymbol{B} && \text{(nach Bedingung } b) \\
&= \boldsymbol{B}\boldsymbol{A}\boldsymbol{A}^+\boldsymbol{A}\boldsymbol{B} && \text{(nach Bedingung } a) \\
&= \boldsymbol{A}^T\boldsymbol{B}^T\boldsymbol{A}^+\boldsymbol{A}\boldsymbol{B} && \text{(nach Bedingung } d) \\
&= \boldsymbol{A}^T\boldsymbol{B}^T\boldsymbol{A}^T\boldsymbol{A}^{+T}\boldsymbol{B} = (\boldsymbol{A}\boldsymbol{B}\boldsymbol{A})^T\boldsymbol{A}^{+T}\boldsymbol{B} && \text{(nach Bedingung } d) \\
&= \boldsymbol{A}^T\boldsymbol{A}^{+T}\boldsymbol{B} && \text{(nach Bedingung } a) \\
&= \boldsymbol{A}^+\boldsymbol{A}\boldsymbol{B} && \text{(nach Bedingung } d) \\
&= \boldsymbol{A}^+\boldsymbol{B}^T\boldsymbol{A}^T && \text{(nach Bedingung } c) \\
&= \boldsymbol{A}^+\boldsymbol{A}\boldsymbol{A}^+\boldsymbol{B}^T\boldsymbol{A}^T && \text{(nach Bedingung } b) \\
&= \boldsymbol{A}^+\boldsymbol{A}^{+T}\boldsymbol{A}^T\boldsymbol{B}^T\boldsymbol{A}^T = \boldsymbol{A}^+\boldsymbol{A}^{+T}(\boldsymbol{A}\boldsymbol{B}\boldsymbol{A})^T && \text{(nach Bedingung } c) \\
&= \boldsymbol{A}^+\boldsymbol{A}^{+T}\boldsymbol{A}^T && \text{(nach Bedingung } a) \\
&= \boldsymbol{A}^+\boldsymbol{A}\boldsymbol{A}^+ && \text{(nach Bedingung } c) \\
&= \boldsymbol{A}^+ && \text{(nach Bedingung } b).
\end{aligned}$$

❀

Für einige spezielle Matrizen lässt sich die Pseudoinverse sofort angeben. Es gilt:

- $\boldsymbol{O}^+ = \boldsymbol{O}$,
- $(\boldsymbol{A}^T)^+ = (\boldsymbol{A}^+)^T$,
- $(\lambda \boldsymbol{A})^+ = \frac{1}{\lambda}\boldsymbol{A}^+, \quad \lambda \neq 0$,
- $\boldsymbol{A}^+ = (\boldsymbol{A}^T\boldsymbol{A})^{-1}\boldsymbol{A}^T, \quad \text{rg}(\boldsymbol{A}) = n \leqq m$,
- $\boldsymbol{A}^+ = \boldsymbol{A}^T(\boldsymbol{A}\boldsymbol{A}^T)^{-1}, \quad \text{rg}(\boldsymbol{A}) = m \leqq n$,
- $\boldsymbol{A}^+ = \boldsymbol{A}^{-1}, \quad \text{rg}(\boldsymbol{A}) = m = n$,
- $(\boldsymbol{A}\boldsymbol{B})^+ = \boldsymbol{B}^+\boldsymbol{A}^+, \quad \boldsymbol{A} \in \mathbb{R}^{m\times r}, \boldsymbol{B} \in \mathbb{R}^{r\times n}, \quad \text{rg}(\boldsymbol{A}) = \text{rg}(\boldsymbol{B}) = r$,
- $(\boldsymbol{U}\boldsymbol{A}\boldsymbol{V})^+ = \boldsymbol{V}^T\boldsymbol{A}^+\boldsymbol{U}^T$ für orthogonale Matrizen $\boldsymbol{U}, \quad \boldsymbol{V}$,
- $\begin{pmatrix} \boldsymbol{A} \\ \boldsymbol{O} \end{pmatrix}^+ = (\boldsymbol{A}^+, \boldsymbol{O})$,
- $(\boldsymbol{a}\boldsymbol{b}^T)^+ = \frac{\boldsymbol{b}\boldsymbol{a}^T}{(\boldsymbol{a}^T\boldsymbol{a})(\boldsymbol{b}^T\boldsymbol{b})}$ für beliebige Vektoren $\boldsymbol{a} \in \mathbb{R}^m$ und $\boldsymbol{b} \in \mathbb{R}^n$ mit $\boldsymbol{a}, \boldsymbol{b} \neq \boldsymbol{o}$.

10.2. Störungstheorie

Wir wollen wieder untersuchen, welchen Einfluss Störungen in der Matrix $\boldsymbol{A}$ und im Vektor $\boldsymbol{y}$ auf die Lösung haben. Dazu betrachten wir neben dem Ausgleichsproblem

$$\min_{\boldsymbol{x}\in\mathbb{R}^n} \|\boldsymbol{y} - \boldsymbol{A}\boldsymbol{x}\|_2$$

mit der Normallösung $\boldsymbol{x} = \boldsymbol{A}^+\boldsymbol{y}$ ein gestörtes Problem

$$\min_{\boldsymbol{x}\in\mathbb{R}^n} \|(\boldsymbol{y} + \boldsymbol{\delta y}) - (\boldsymbol{A} + \boldsymbol{\delta A})\boldsymbol{x}\|_2$$

mit der Normallösung $\boldsymbol{x} + \boldsymbol{\delta x} = (\boldsymbol{A} + \boldsymbol{\delta A})^+(\boldsymbol{y} + \boldsymbol{\delta y})$. Durch die Störungen $\boldsymbol{\delta A}$ und $\boldsymbol{\delta y}$ wird der Fehler

$$\boldsymbol{\delta x} = (\boldsymbol{A} + \boldsymbol{\delta A})^+(\boldsymbol{y} + \boldsymbol{\delta y}) - \boldsymbol{A}^+\boldsymbol{y} = \left[(\boldsymbol{A} + \boldsymbol{\delta A})^+ - \boldsymbol{A}^+\right]\boldsymbol{y} + (\boldsymbol{A} + \boldsymbol{\delta A})^+\, dy$$

erzeugt. Zunächst untersuchen wir, wie sich die Pseudoinverse einer Matrix gegenüber Störungen verhält. Dazu zwei Beispiele.

10.3. Beispiel: Es sei $\boldsymbol{A} \neq \boldsymbol{O}$ durch die Singulärwertzerlegung $\boldsymbol{A} = \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^T$ mit

$$\boldsymbol{\Sigma} = \text{diag}(\sigma_1, \ldots, \sigma_r, 0, \ldots, 0), \quad \sigma_1 \geqq \cdots \geqq \sigma_r > 0$$

gegeben. Es gilt $\text{rg}(\boldsymbol{A}) = r$. Wir betrachten die spezielle Störung $\boldsymbol{\delta A} = \boldsymbol{U}\boldsymbol{\delta\Sigma}\boldsymbol{V}^T$ mit

$$(\boldsymbol{\delta\Sigma})_{ij} = \begin{cases} \varepsilon & \text{für} \quad i = j = r \\ 0 & \text{sonst} \end{cases} .$$

Dann folgt $\boldsymbol{A} + \boldsymbol{\delta A} = \boldsymbol{U}(\boldsymbol{\Sigma} + \boldsymbol{\delta\Sigma})\boldsymbol{V}^T$ mit

$$\boldsymbol{\Sigma} + \boldsymbol{\delta\Sigma} = \text{diag}(\sigma_1, \ldots, \sigma_{r-1}, \sigma_r + \varepsilon, 0, \ldots, 0).$$

Weiterhin gilt $\|\boldsymbol{A}^+\|_2 = \|\boldsymbol{\Sigma}^+\|_2 = 1/\sigma_r$ und $\|\boldsymbol{\delta A}\|_2 = \|dSigma\|_2 = |\varepsilon|$. Falls außerdem

$$\kappa = \|\boldsymbol{A}^+\|_2 \|\boldsymbol{\delta A}\|_2 = \frac{|\varepsilon|}{\sigma_r} < 1$$

gilt, folgt

$$\sigma_r + \varepsilon \geqq \sigma_r - |\varepsilon| = \sigma_r(1 - \kappa) > 0.$$

Damit gilt

$$\text{rg}(\boldsymbol{A} + \boldsymbol{\delta A}) = \text{rg}(\boldsymbol{\Sigma} + \boldsymbol{\delta\Sigma}) = \text{rg}(\boldsymbol{\Sigma}) = \text{rg}(\boldsymbol{A}) = r.$$

Für die Pseudoinverse der gestörten Matrix gilt dann

$$(\boldsymbol{A} + \boldsymbol{\delta A})^+ = \boldsymbol{V}(\boldsymbol{\Sigma} + \boldsymbol{\delta\Sigma})^+ \boldsymbol{U}^T$$

mit

$$(\boldsymbol{\Sigma} + \boldsymbol{\delta\Sigma})^+ = \text{diag}(1/\sigma_1, \ldots, 1/\sigma_{r-1}, 1/(\sigma_r + \varepsilon), 0, \ldots, 0).$$

Für die Norm der Pseudoinversen der gestörten Matrix erhält man

$$\begin{aligned} \left\|(\boldsymbol{A} + \boldsymbol{\delta A})^+\right\|_2 &= \left\|(\boldsymbol{\Sigma} + \boldsymbol{\delta\Sigma})^+\right\|_2 \\ &= \max\left\{\frac{1}{\sigma_{r-1}}, \frac{1}{\sigma_r + \varepsilon}\right\} \\ &\leqq \max\left\{\frac{1}{\sigma_r}, \frac{1}{\sigma_r + \varepsilon}\right\} \\ &\leqq \frac{1}{\sigma_r - |\varepsilon|} = \frac{1}{\sigma_r(1 - \kappa)}. \end{aligned}$$

Damit gilt

$$\left\|(\boldsymbol{A}+\boldsymbol{\delta A})^{+}\right\|_2 \leqq \frac{\left\|\boldsymbol{A}^{+}\right\|_2}{1-\kappa}.$$

Wir wollen noch abschätzen, wie stark sich die Pseudoinverse der gestörten Matrix $\boldsymbol{A}+\boldsymbol{\delta A}$ von der Pseudoinversen der Matrix $\boldsymbol{A}$ unterscheidet. Es gilt

$$(\boldsymbol{A}+\boldsymbol{\delta A})^{+}-\boldsymbol{A}^{+}=\boldsymbol{V}(\Sigma+\boldsymbol{\delta\Sigma})^{+}\boldsymbol{U}^{T}-\boldsymbol{V}\Sigma^{+}\boldsymbol{U}^{T}=\boldsymbol{V}\left[(\Sigma+\boldsymbol{\delta\Sigma})^{+}-\Sigma^{+}\right]\boldsymbol{U}^{T}$$

mit

$$\left((\Sigma+\boldsymbol{\delta\Sigma})^{+}-\Sigma^{+}\right)_{ij}=\begin{cases}\frac{1}{\sigma_r+\varepsilon}-\frac{1}{\sigma_r} & \text{für} \quad i=j=r \\ 0 & \text{sonst}\end{cases}.$$

Damit folgt

$$\left\|(\Sigma+\boldsymbol{\delta\Sigma})^{+}-\Sigma^{+}\right\|_2=\left|\frac{1}{\sigma_r+\varepsilon}-\frac{1}{\sigma_r}\right|=\frac{|\varepsilon|}{\sigma_r|\sigma_r+\varepsilon|}\leqq\frac{|\varepsilon|}{\sigma_r^2(1-\kappa)}$$

und

$$\left\|(\boldsymbol{A}+\boldsymbol{\delta A})^{+}-\boldsymbol{A}^{+}\right\|_2=\left\|(\Sigma+\boldsymbol{\delta\Sigma})^{+}-\Sigma^{+}\right\|_2\leqq\frac{\left\|\boldsymbol{A}^{+}\right\|_2}{1-\kappa}\left\|\boldsymbol{\delta A}\right\|_2.$$

Diese Abschätzungen entsprechen genau den Abschätzungen, die wir bei der Behandlung der Störungen einer regulären Matrix für die Inverse erhalten hatten. ♡

10.4. Beispiel: Es sei $\boldsymbol{A}\neq\boldsymbol{O}$ durch die Singulärwertzerlegung $\boldsymbol{A}=\boldsymbol{U}\Sigma\boldsymbol{V}^{T}$ mit

$$\Sigma=\operatorname{diag}(\sigma_1,\ldots,\sigma_r,0,\ldots,0),\quad \sigma_1\geqq\cdots\geqq\sigma_r>0$$

gegeben. Es gelte $\operatorname{rg}(\boldsymbol{A})=r<\min\{m,n\}$. Wir betrachten die spezielle Störung $\boldsymbol{\delta A}=\boldsymbol{U}\boldsymbol{\delta\Sigma}\boldsymbol{V}^{T}$ mit

$$(\boldsymbol{\delta\Sigma})_{ij}=\begin{cases}\varepsilon & \text{für} \quad i=j=r+1 \\ 0 & \text{sonst}\end{cases}.$$

Dann folgt $\boldsymbol{A}+\boldsymbol{\delta A}=\boldsymbol{U}(\Sigma+\boldsymbol{\delta\Sigma})\boldsymbol{V}^{T}$ mit $\Sigma+\boldsymbol{\delta\Sigma}=\operatorname{diag}(\sigma_1,\ldots,\sigma_r,\varepsilon,0,\ldots,0)$ und

$$(\Sigma+\boldsymbol{\delta\Sigma})^{+}=\operatorname{diag}\left(\frac{1}{\sigma_1},\ldots,\frac{1}{\sigma_r},\frac{1}{\varepsilon},0,\ldots,0\right).$$

Weiterhin gilt $\operatorname{rg}(\boldsymbol{A}+\boldsymbol{\delta A})=\operatorname{rg}(\Sigma+\boldsymbol{\delta\Sigma})=r+1>r=\operatorname{rg}(\Sigma)=\operatorname{rg}(\boldsymbol{A})$ und

$$\left\|(\boldsymbol{A}+\boldsymbol{\delta A})^{+}\right\|_2=\left\|(\Sigma+\boldsymbol{\delta\Sigma})^{+}\right\|_2=\max\left\{\frac{1}{\sigma_r},\frac{1}{|\varepsilon|}\right\}\geqq\frac{1}{|\varepsilon|}=\frac{1}{\|\boldsymbol{\delta A}\|_2},$$

sowie

$$\left\|(\boldsymbol{A}+\boldsymbol{\delta A})^{+}-\boldsymbol{A}^{+}\right\|_2=\left\|(\boldsymbol{\Sigma}+\boldsymbol{\delta\Sigma})^{+}-\boldsymbol{\Sigma}^{+}\right\|_2=\frac{1}{|\varepsilon|}=\frac{1}{\|\boldsymbol{\delta A}\|_2}.$$

Diese rangerhöhende Störung hat einen katastrophalen Einfluss auf die Pseudoinverse. Je kleiner die Störung ist, desto mehr weicht die Pseudoinverse der gestörten Matrix von der Pseudoinversen der exakten Matrix ab. Die Pseudoinverse einer Matrix ist bezüglich dieser rangerhöhenden Störung unstetig und unbeschränkt. Es liegt ein extrem schlechtes Unstetigkeitsverhalten vor. ♡

Im allgemeinen Falle gibt es genau die zwei Klassen von Störungen, die in den beiden Beispielen zum Ausdruck kamen. Zum Beweis der entsprechende Störungssätze benötigen wir den folgenden Hilfssatz.

10.5. Satz: *Es seien $\boldsymbol{A}$ und $\boldsymbol{\delta A}$ (m,n)-Matrizen. Die Größen σ_i und $\delta\sigma_i$ seien die gemäß*

$$\sigma_1 \geqq \sigma_2 \geqq \cdots \geqq \sigma_l$$

bzw.

$$\delta\sigma_1 \geqq \delta\sigma_2 \geqq \cdots \geqq \delta\sigma_l$$

geordneten Singulärwerte der Matrizen $\boldsymbol{A}$ bzw. $\boldsymbol{A}+\boldsymbol{\delta A}$, wobei $l=\min\{m,n\}$. Dann gelten die Abschätzungen

$$|\delta\sigma_i| \leqq \|\boldsymbol{\delta A}\|_2$$

für $i=1,\ldots,l$.

Beweis: Wir definieren die Matrizen

$$\boldsymbol{B}=\left(\begin{array}{c|c}\boldsymbol{O} & \boldsymbol{A}\\ \hline \boldsymbol{A}^T & \boldsymbol{O}\end{array}\right)\in\mathbb{R}^{(m+n)\times(m+n)},\quad \boldsymbol{\delta B}=\left(\begin{array}{c|c}\boldsymbol{O} & \boldsymbol{\delta A}\\ \hline \boldsymbol{\delta A}^T & \boldsymbol{O}\end{array}\right)\in\mathbb{R}^{(m+n)\times(m+n)}.$$

Aus der Singulärwertzerlegung $\boldsymbol{A}=\boldsymbol{U\Sigma V}^T$ folgt

$$\boldsymbol{B}=\left(\begin{array}{c|c}\boldsymbol{O} & \boldsymbol{U\Sigma V}^T\\ \hline \boldsymbol{V\Sigma}^T\boldsymbol{U}^T & \boldsymbol{O}\end{array}\right)=\boldsymbol{SCS}^T,\quad \boldsymbol{S}=\left(\begin{array}{c|c}\boldsymbol{U} & \boldsymbol{O}\\ \hline \boldsymbol{O} & \boldsymbol{V}\end{array}\right),\quad \boldsymbol{C}=\left(\begin{array}{c|c}\boldsymbol{O} & \boldsymbol{\Sigma}\\ \hline \boldsymbol{\Sigma}^T & \boldsymbol{O}\end{array}\right).$$

Wegen der Orthogonalität der Matrix $\boldsymbol{S}$ sind die Matrizen $\boldsymbol{B}$ und $\boldsymbol{C}$ ähnlich, sie besitzen daher die gleichen Eigenwerte. Die Eigenwertgleichung $\boldsymbol{Cz}=\lambda\boldsymbol{z}$ lässt sich mit

$$\boldsymbol{z}=\begin{pmatrix}\boldsymbol{y}\\ \boldsymbol{x}\end{pmatrix}\quad\text{als}\quad\begin{pmatrix}\boldsymbol{\Sigma x}\\ \boldsymbol{\Sigma}^T\boldsymbol{x}\end{pmatrix}=\lambda\begin{pmatrix}\boldsymbol{y}\\ \boldsymbol{x}\end{pmatrix}\tag{10.1}$$

schreiben. Für $m \geqq n$ hat $\boldsymbol{\Sigma}$ die Form

$$\boldsymbol{\Sigma} = \begin{pmatrix} \boldsymbol{D} \\ \boldsymbol{O} \end{pmatrix}, \quad \boldsymbol{D} = \begin{pmatrix} \sigma_1 & & \\ & \ddots & \\ & & \sigma_n \end{pmatrix}.$$

Dann folgt aus Gleichung 10.1 mit $\boldsymbol{y} = (\bar{\boldsymbol{y}}^T, \tilde{\boldsymbol{y}}^T)^T$

$$\begin{aligned} \boldsymbol{D}\boldsymbol{x} &= \lambda \bar{\boldsymbol{y}}, \\ \boldsymbol{o} &= \lambda \tilde{\boldsymbol{y}}, \\ \boldsymbol{D}\bar{\boldsymbol{y}} &= \lambda \boldsymbol{x}. \end{aligned}$$

Hieraus ergeben sich die Eigenwerte

$$\begin{aligned} &\lambda_i = \sigma_i, \ \ i = 1, \ldots, n, \\ &\lambda_i = -\sigma_{i-n}, \ \ i = n+1, \ldots, 2n, \\ &\lambda_i = 0, \ \ i = 2n+1, \ldots, m+n. \end{aligned}$$

Für $m \leqq n$ ergibt sich entsprechend

$$\begin{aligned} &\lambda_i = \sigma_i, \ \ i = 1, \ldots, m, \\ &\lambda_i = -\sigma_{i-m}, \ \ i = m+1, \ldots, 2m, \lambda_i \qquad\qquad = 0, \ \ i = 2m+1, \ldots, m+n. \end{aligned}$$

Die Matrizen $\boldsymbol{C}$ und $\boldsymbol{B}$ besitzen somit in jedem Falle die Eigenwerte $\pm\sigma_i$, $i = 1, \ldots, l$. Genauso zeigt man, dass die Matrix $\boldsymbol{A} + \boldsymbol{\delta A}$ die Eigenwerte $\pm(\sigma_i + \delta\sigma_i)$, $i = 1, \ldots, l$, besitzt. Mit Satz 9.23 folgt dann $|\delta\sigma_i| \leqq \|\boldsymbol{\delta A}\|_2$ für $i = 1, \ldots, l$. ❀

Nun untersuchen wir, wie sich rangerhöhende Störungen auswirken. Es gilt der folgende Satz.

10.6. Satz: *Es sei $\boldsymbol{A}$ eine (m,n)-Matrix mit* $\operatorname{rg}(\boldsymbol{A}) = r < l = \min\{m,n\}$, *und $\boldsymbol{\delta A}$ sei eine (m,n)-Störung mit* $\operatorname{rg}(\boldsymbol{A} + \boldsymbol{\delta A}) > \operatorname{rg}(\boldsymbol{A})$. *Dann gilt*

$$\left\|(\boldsymbol{A} + \boldsymbol{\delta A})^+\right\|_2 \geqq \frac{1}{\|\boldsymbol{\delta A}\|_2}.$$

Beweis: Für die geordneten Singulärwerte σ_i bzw. $\tilde{\sigma}_i$, $i = 1, \ldots, l$, der Matrizen $\boldsymbol{A}$ bzw. $\tilde{\boldsymbol{A}} = \boldsymbol{A} + \boldsymbol{\delta A}$ gilt nach Satz 10.5

$$|\sigma_i - \tilde{\sigma}_i| \leqq \|\boldsymbol{\delta A}\|_2, \quad i = 1, \ldots, l.$$

Speziell gilt

$$|\sigma_{r+1} - \tilde{\sigma}_{r+1}| = \tilde{\sigma}_{r+1} \leqq \|\boldsymbol{\delta A}\|_2.$$

Damit folgt

$$\left\|(\boldsymbol{A}+\boldsymbol{\delta A})^+\right\|_2 = \frac{1}{\tilde{\sigma}_{r+1}} \geqq \frac{1}{\|\boldsymbol{\delta A}\|_2}.$$

❀

Bemerkung: Wegen

$$\left\|(\boldsymbol{A}+\boldsymbol{\delta A})^+ - \boldsymbol{A}^+\right\|_2 \geqq \left\|(\boldsymbol{A}+\boldsymbol{\delta A})^+\right\|_2 - \left\|\boldsymbol{A}^+\right\|_2 \geqq \frac{1}{\|\boldsymbol{\delta A}\|_2} - \left\|\boldsymbol{A}^+\right\|_2$$

kann die Pseudoinverse der gestörten Matrix beliebig stark von der Pseudoinversen der exakten Matrix abweichen.
Für Störungen, die den Rang der Matrix nicht erhöhen gilt:

10.7. Satz: *Für die (m,n)-Matrix $\boldsymbol{A}$ und die (m,n)-Störung $\boldsymbol{\delta A}$ sei*

$$\mathrm{rg}(\boldsymbol{A}+\boldsymbol{\delta A}) \leqq \mathrm{rg}(\boldsymbol{A}), \quad \kappa = \|\boldsymbol{A}^+\|_2\|\boldsymbol{\delta A}\|_2 < 1.$$

Dann gilt

1.

$$\mathrm{rg}(\boldsymbol{A}+\boldsymbol{\delta A}) = \mathrm{rg}(\boldsymbol{A}),$$

2.

$$\left\|(\boldsymbol{A}+\boldsymbol{\delta A})^+\right\|_2 \leqq \frac{\left\|\boldsymbol{A}^+\right\|_2}{1-\kappa},$$

3.

$$(\boldsymbol{A}+\boldsymbol{\delta A})^+ - \boldsymbol{A}^+ = -\boldsymbol{A}^+\boldsymbol{\delta A}\boldsymbol{A}^+ + \boldsymbol{A}^+\boldsymbol{A}^{+T}\boldsymbol{\delta A}^T(\boldsymbol{I}-\boldsymbol{A}\boldsymbol{A}^+) + \\ +(\boldsymbol{I}-\boldsymbol{A}^+\boldsymbol{A})\boldsymbol{\delta A}^T\boldsymbol{A}^{+T}\boldsymbol{A}^+ + O(\|\boldsymbol{\delta A}\|_2^2),$$

4.

$$\left\|(\boldsymbol{A}+\boldsymbol{\delta A})^+ - \boldsymbol{A}^+\right\|_2 \leqq \mu\frac{\left\|\boldsymbol{A}^+\right\|_2^2}{1-\kappa}\|\boldsymbol{\delta A}\|_2$$

mit

$$\mu = \begin{cases} \frac{1+\sqrt{5}}{2} & fr \quad \text{rg}(\boldsymbol{A}) < \min\{m,n\} \\ \sqrt{2} & fr \quad \text{rg}(\boldsymbol{A}) = \min\{m,n\} < \max\{m,n\} \\ 1 & fr \quad \text{rg}(\boldsymbol{A}) = m = n \end{cases} .$$

Beweis: Es sei $r = \text{rg}(\boldsymbol{A})$. Mit σ_i, $i = 1,\ldots,r$, bezeichnen wir wieder die geordneten Singulärwerte von der matrix $\boldsymbol{A}$ und mit $\tilde{\sigma}_i = \sigma_i + \delta\sigma_i$, $i = 1,\ldots,l$ die geordneten Singulärwerte der matrix $\tilde{\boldsymbol{A}} = \boldsymbol{A} + \boldsymbol{\delta A}$. Dann gilt

$$\tilde{\sigma}_i = \sigma_i + \delta\sigma_i \geqq \sigma_i - |\delta\sigma_i|,$$

und wegen $|\delta\sigma_i| \leqq \|\boldsymbol{\delta A}\|_2$ folgt

$$\tilde{\sigma}_i \geqq \sigma_i - \|\boldsymbol{\delta A}\|_2 \geqq \sigma_r - \|\boldsymbol{\delta A}\|_2 = \frac{1}{\left\|\boldsymbol{A}^+\right\|_2} - \|\boldsymbol{\delta A}\|_2 = \frac{1-\kappa}{\left\|\boldsymbol{A}^+\right\|_2} > 0$$

für $i = 1,\ldots,r$. Folglich gilt $\text{rg}(\boldsymbol{A} + \boldsymbol{\delta A}) \geqq r$ und mit der Bedingung $\text{rg}(\boldsymbol{A} + \boldsymbol{\delta A}) \leqq r$ folgt $\text{rg}(\boldsymbol{A} + \boldsymbol{\delta A}) = r$. Damit ist $\tilde{\sigma}_{r+1} = \cdots = \tilde{\sigma}_l = 0$ und

$$\left\|(\boldsymbol{A} + \boldsymbol{\delta A})^+\right\|_2 = \frac{1}{\tilde{\sigma}_r} \leqq \frac{\left\|\boldsymbol{A}^+\right\|_2}{1-\kappa}.$$

Damit sind die ersten beiden Aussagen bewiesen.
Es sei nun

$$\tilde{\boldsymbol{A}} = \boldsymbol{A} + \boldsymbol{\delta A}, \quad \boldsymbol{G} = \tilde{\boldsymbol{A}}^+ - \boldsymbol{A}^+$$

sowie

$$\boldsymbol{P} = \boldsymbol{A}\boldsymbol{A}^+, \quad \boldsymbol{S} = \boldsymbol{A}^+\boldsymbol{A}, \quad \tilde{\boldsymbol{P}} = \tilde{\boldsymbol{A}}\tilde{\boldsymbol{A}}^+, \quad \tilde{\boldsymbol{S}} = \tilde{\boldsymbol{A}}^+\tilde{\boldsymbol{A}}.$$

Damit folgt

$$\begin{aligned} \boldsymbol{G} &= [\tilde{\boldsymbol{S}} + (\boldsymbol{I} - \tilde{\boldsymbol{S}})](\tilde{\boldsymbol{A}}^+ - \boldsymbol{A}^+)[\boldsymbol{P} + (\boldsymbol{I} - \boldsymbol{P})] \\ &= \tilde{\boldsymbol{S}}\tilde{\boldsymbol{A}}^+\boldsymbol{P} + \tilde{\boldsymbol{S}}\tilde{\boldsymbol{A}}^+(\boldsymbol{I} - \boldsymbol{P}) - \tilde{\boldsymbol{S}}\boldsymbol{A}^+\boldsymbol{P} - \tilde{\boldsymbol{S}}\boldsymbol{A}^+(\boldsymbol{I} - \boldsymbol{P}) + \\ &\quad + (\boldsymbol{I} - \tilde{\boldsymbol{S}})\tilde{\boldsymbol{A}}^+\boldsymbol{P} + (\boldsymbol{I} - \tilde{\boldsymbol{S}})\tilde{\boldsymbol{A}}^+(\boldsymbol{I} - \boldsymbol{P}) \\ &\quad - (\boldsymbol{I} - \tilde{\boldsymbol{S}})\boldsymbol{A}^+\boldsymbol{P} - (\boldsymbol{I} - \tilde{\boldsymbol{S}})\boldsymbol{A}^+(\boldsymbol{I} - \boldsymbol{P}). \end{aligned}$$

Nun gilt

$$\tilde{\boldsymbol{S}}\tilde{\boldsymbol{A}}^+ = \tilde{\boldsymbol{A}}^+\tilde{\boldsymbol{A}}\tilde{\boldsymbol{A}}^+ = \tilde{\boldsymbol{A}}^+, \quad \tilde{\boldsymbol{A}}^+\boldsymbol{P} = \tilde{\boldsymbol{A}}^+\boldsymbol{A}\boldsymbol{A}^+ = \boldsymbol{A}^+,$$

und damit

$$(\boldsymbol{I}-\tilde{\boldsymbol{S}})\tilde{\boldsymbol{A}}^{+}=\boldsymbol{O},\quad \boldsymbol{A}^{+}(\boldsymbol{I}-\boldsymbol{P})=\boldsymbol{O}.$$

Somit ergibt sich

$$\begin{array}{rcccccc}\boldsymbol{G} &=& \underbrace{(\tilde{\boldsymbol{A}}^{+}\boldsymbol{P}-\tilde{\boldsymbol{S}}\boldsymbol{A}^{+})} &+& \underbrace{\tilde{\boldsymbol{A}}^{+}(\boldsymbol{I}-\boldsymbol{P})} &+& \underbrace{(\boldsymbol{I}-\tilde{\boldsymbol{S}})(-\boldsymbol{A}^{+})} \\ &=& \boldsymbol{G}_1 &+& \boldsymbol{G}_2 &+& \boldsymbol{G}_3.\end{array}$$

Für die einzelnen Summanden erhalten wir weiter:

$$\begin{aligned}\boldsymbol{G}_1 &= \tilde{\boldsymbol{A}}^{+}\boldsymbol{P}-\tilde{\boldsymbol{S}}\boldsymbol{A}^{+},\\ &= \tilde{\boldsymbol{A}}^{+}\boldsymbol{A}\boldsymbol{A}^{+}-\tilde{\boldsymbol{A}}^{+}\tilde{\boldsymbol{A}}\boldsymbol{A}^{+},\\ &= \tilde{\boldsymbol{A}}^{+}(\boldsymbol{A}-\tilde{\boldsymbol{A}})\boldsymbol{A}^{+},\\ &= -\tilde{\boldsymbol{A}}^{+}\boldsymbol{\delta}\boldsymbol{A}\boldsymbol{A}^{+}.\\ \boldsymbol{G}_2 &= \tilde{\boldsymbol{A}}^{+}(\boldsymbol{I}-\boldsymbol{P}),\\ &= \tilde{\boldsymbol{A}}^{+}\tilde{\boldsymbol{A}}\tilde{\boldsymbol{A}}^{+}(\boldsymbol{I}-\boldsymbol{P}),\\ &= \tilde{\boldsymbol{A}}^{+}\tilde{\boldsymbol{A}}^{+^T}\tilde{\boldsymbol{A}}^{T}(\boldsymbol{I}-\boldsymbol{P}).\end{aligned}$$

Berücksichtigt man, dass

$$\boldsymbol{A}^{T}(\boldsymbol{I}-\boldsymbol{P})=\boldsymbol{A}^{T}-\boldsymbol{A}^{T}\boldsymbol{A}\boldsymbol{A}^{+}=\boldsymbol{A}^{T}-\boldsymbol{A}^{T}\boldsymbol{A}^{+^T}\boldsymbol{A}^{T}=\boldsymbol{A}^{T}-\boldsymbol{A}^{T}=\boldsymbol{O}$$

gilt, so folgt

$$\begin{aligned}\boldsymbol{G}_2 &= \tilde{\boldsymbol{A}}^{+}\tilde{\boldsymbol{A}}^{+^T}(\tilde{\boldsymbol{A}}^{T}-\boldsymbol{A}^{T})(\boldsymbol{I}-\boldsymbol{P}),\\ &= \tilde{\boldsymbol{A}}^{+}\tilde{\boldsymbol{A}}^{+^T}\boldsymbol{\delta}\boldsymbol{A}^{T}(\boldsymbol{I}-\boldsymbol{P}).\end{aligned}$$

Für $\boldsymbol{G}_3$ folgt mit

$$(\boldsymbol{I}-\tilde{\boldsymbol{S}})\tilde{\boldsymbol{A}}^{T}=\tilde{\boldsymbol{A}}^{T}-\tilde{\boldsymbol{A}}^{+}\tilde{\boldsymbol{A}}\tilde{\boldsymbol{A}}^{T}=\tilde{\boldsymbol{A}}^{T}-\tilde{\boldsymbol{A}}^{T}\tilde{\boldsymbol{A}}^{+^T}\tilde{\boldsymbol{A}}^{T}=\tilde{\boldsymbol{A}}^{T}-\tilde{\boldsymbol{A}}^{T}=\boldsymbol{O}$$

$$\begin{aligned}\boldsymbol{G}_3 &= -(\boldsymbol{I}-\tilde{\boldsymbol{S}})\boldsymbol{A}^{+},\\ &= -(\boldsymbol{I}-\tilde{\boldsymbol{S}})\boldsymbol{A}^{+}\boldsymbol{A}\boldsymbol{A}^{+},\\ &= -(\boldsymbol{I}-\tilde{\boldsymbol{S}})\boldsymbol{A}^{T}\boldsymbol{A}^{+^T}\boldsymbol{A}^{+},\\ &= -(\boldsymbol{I}-\tilde{\boldsymbol{S}})(\boldsymbol{A}^{T}-\tilde{\boldsymbol{A}}^{T})\boldsymbol{A}^{+^T}\boldsymbol{A}^{+},\\ &= -(\boldsymbol{I}-\tilde{\boldsymbol{S}})\boldsymbol{\delta}\boldsymbol{A}^{T}\boldsymbol{A}^{+^T}\boldsymbol{A}^{+}.\end{aligned}$$

Mit Hilfe der Abschätzung

$$\|\tilde{\boldsymbol{A}}^+\|_2 \leqq \frac{\|\boldsymbol{A}^+\|_2}{1-\kappa}$$

erhalten wir weiter

$$\|\boldsymbol{G}_1\|_2 \leqq \|\tilde{\boldsymbol{A}}^+\|_2\|\boldsymbol{A}^+\|_2\|\boldsymbol{\delta A}\|_2 \leqq \frac{\|\boldsymbol{A}^+\|_2}{1-\kappa}\|\boldsymbol{A}^+\|_2\|\boldsymbol{\delta A}\|_2 = \alpha$$

und

$$\|\boldsymbol{G}_3\|_2 \leqq \|\boldsymbol{I}-\tilde{\boldsymbol{S}}\|_2\|\boldsymbol{A}^+\|_2^2\|\boldsymbol{\delta A}\|_2 = \|\boldsymbol{A}^+\|_2^2\|\boldsymbol{\delta A}\|_2 \leqq \alpha.$$

(Man beachte, dass wegen $(\boldsymbol{I}-\tilde{\boldsymbol{S}})^2 = \boldsymbol{I}-\tilde{\boldsymbol{S}}$ gilt: $\|\boldsymbol{I}-\tilde{\boldsymbol{S}}\|_2 = 1$.)
Bei der Abschätzung von $\boldsymbol{G}_2$ würde der Faktor $\|\tilde{\boldsymbol{A}}^+\|_2^2$ auftauchen. Um dies zu vermeiden, formen wir $\boldsymbol{G}_2$ weiter um. Es gilt

$$\begin{aligned}
\|\boldsymbol{G}_2\|_2 &= \|\tilde{\boldsymbol{A}}^+(\boldsymbol{I}-\boldsymbol{P})\|_2, \\
&= \|\tilde{\boldsymbol{A}}^+\tilde{\boldsymbol{P}}(\boldsymbol{I}-\boldsymbol{P})\|_2, \\
&\leqq \|\tilde{\boldsymbol{A}}^+\|_2\|\tilde{\boldsymbol{P}}(\boldsymbol{I}-\boldsymbol{P})\|_2.
\end{aligned}$$

In Übungsaufgabe 4 ist zu zeigen, dass $\|\tilde{\boldsymbol{P}}(\boldsymbol{I}-\boldsymbol{P})\|_2 = \|\boldsymbol{P}(\boldsymbol{I}-\tilde{\boldsymbol{P}})\|_2$ gilt. Damit folgt

$$\begin{aligned}
\|\boldsymbol{G}_2\|_2 &\leqq \|\tilde{\boldsymbol{A}}^+\|_2\|\boldsymbol{P}(\boldsymbol{I}-\tilde{\boldsymbol{P}})\|_2, \\
&= \|\tilde{\boldsymbol{A}}^+\|_2\|\boldsymbol{A}\boldsymbol{A}^+(\boldsymbol{I}-\tilde{\boldsymbol{P}})\|_2, \\
&= \|\tilde{\boldsymbol{A}}^+\|_2\|\boldsymbol{A}^{+T}\boldsymbol{A}^T(\boldsymbol{I}-\tilde{\boldsymbol{P}})\|_2, \\
&= \|\tilde{\boldsymbol{A}}^+\|_2\|\boldsymbol{A}^{+T}(\boldsymbol{A}^T-\tilde{\boldsymbol{A}}^T)(\boldsymbol{I}-\tilde{\boldsymbol{P}})\|_2, \\
&= \|\tilde{\boldsymbol{A}}^+\|_2\|\boldsymbol{A}^{+T}\boldsymbol{\delta A}^T(\boldsymbol{I}-\tilde{\boldsymbol{P}})\|_2, \\
&\leqq \frac{\|\boldsymbol{A}^+\|_2}{1-\kappa}\|\boldsymbol{A}^+\|_2\|\boldsymbol{\delta A}\|_2\|\boldsymbol{I}-\tilde{\boldsymbol{P}}\|_2, \\
&= \frac{\|\boldsymbol{A}^+\|_2}{1-\kappa}\|\boldsymbol{A}^+\|_2\|\boldsymbol{\delta A}\|_2 = \alpha.
\end{aligned}$$

Dabei wurde wieder ausgenutzt, dass $\tilde{\boldsymbol{A}}^T(\boldsymbol{I}-\tilde{\boldsymbol{P}}) = \boldsymbol{O}$ und $\|\boldsymbol{I}-\tilde{\boldsymbol{P}}\|_2 = 1$ gilt. Insgesamt erhalten wir die Abschätzung

$$\|\boldsymbol{G}\|_2 \leqq 3\alpha.$$

Die verbesserten Werte sind in Übungsaufgabe 5 zu zeigen. Aus

$$\left\|(\boldsymbol{A}+\boldsymbol{\delta A})^{+}-\boldsymbol{A}^{+}\right\|_2 \leqq 3\frac{\|\boldsymbol{A}^{+}\|_2}{1-\kappa}\|\boldsymbol{A}^{+}\|_2\|\boldsymbol{\delta A}\|_2$$

folgt

$$(\boldsymbol{A}+\boldsymbol{\delta A})^{+}=\boldsymbol{A}^{+}+O(\|\boldsymbol{\delta A}\|_2).$$

Setzt man dies in die Darstellung

$$\boldsymbol{G}=(\boldsymbol{A}+\boldsymbol{\delta A})^{+}-\boldsymbol{A}^{+}=\boldsymbol{G}_1+\boldsymbol{G}_2+\boldsymbol{G}_3$$

ein, so erhält man die linearisierte Formel

$$\begin{aligned}(\boldsymbol{A}+\boldsymbol{\delta A})^{+}-\boldsymbol{A}^{+} \;=\; & -\boldsymbol{A}^{+}\boldsymbol{\delta A}\boldsymbol{A}^{+}+\boldsymbol{A}^{+}\boldsymbol{A}^{+^T}\boldsymbol{\delta A}^{T}(\boldsymbol{I}-\boldsymbol{A}\boldsymbol{A}^{+})+\\ & (\boldsymbol{I}-\boldsymbol{A}^{+}\boldsymbol{A})\boldsymbol{\delta A}^{T}\boldsymbol{A}^{+^T}\boldsymbol{A}^{+}+O(\|\boldsymbol{\delta A}\|_2^2).\end{aligned}$$

❀

Bemerkungen: (i) Die Bedingung $\mathrm{rg}(\boldsymbol{A}+\boldsymbol{\delta A}) \leqq \mathrm{rg}(\boldsymbol{A})$ ist insbesondere erfüllt, falls die Matrix $\boldsymbol{A}$ den maximalen Rang $r=l=\min\{m,n\}$ hat. In diesem Falle erhöht keine Störung den Rang der Matrix. Die Bestimmung der Pseudoinversen ist dann eine lokal lipschitzstetige Aufgabe, daher ein korrekt gestelltes Problem. Das folgt auch schon aus der expliziten Darstellung von $\boldsymbol{A}^{+}$ für diese Fälle:

$$\begin{aligned}\boldsymbol{A}^{+} &= (\boldsymbol{A}^{T}\boldsymbol{A})^{-1}\boldsymbol{A}^{T}, \quad r=n,\\ \boldsymbol{A}^{+} &= \boldsymbol{A}^{T}(\boldsymbol{A}\boldsymbol{A}^{T})^{-1}, \quad r=m.\end{aligned}$$

(ii) Bei rangdefizienten Problemen, also $\mathrm{rg}(\boldsymbol{A})<l=\min\{m,n\}$, hat man immer mit Störungen zu rechnen, die den Rang der Matrix erhöhen. In diesem Falle ist die Norm der Pseudoinversen unbeschränkt und $(\boldsymbol{A}+\boldsymbol{\delta A})^{+}$ weicht beliebig stark von $\boldsymbol{A}^{+}$ ab. Das Bestimmen der Pseudoinversen für rangdefiziente Probleme ist daher ein inkorrekt gestelltes Problem. Wir wollen nun noch das Störungsverhalten des gesamten linearen Ausgleichsproblems betrachten.

10.8. Satz: *Gegeben seien die linearen Ausgleichsprobleme*

$$\min_{\boldsymbol{x}\in\mathbb{R}^n}\|\boldsymbol{y}-\boldsymbol{A}\boldsymbol{x}\|_2$$

und

$$\min_{\boldsymbol{x}+\boldsymbol{\delta x}\in\mathbb{R}^n}\|\boldsymbol{y}+\boldsymbol{\delta y}-(\boldsymbol{A}+\boldsymbol{\delta A})(\boldsymbol{x}+\boldsymbol{\delta x})\|_2$$

mit den Normallösungen

$$\boldsymbol{x} = \boldsymbol{A}^+\boldsymbol{y}, \quad \boldsymbol{x}+\boldsymbol{\delta x} = (\boldsymbol{A}+\boldsymbol{\delta A})^+(\boldsymbol{y}+\boldsymbol{\delta y}).$$

Die Störung $\boldsymbol{\delta A}$ genüge den Bedingungen

$$\mathrm{rg}(\boldsymbol{A}+\boldsymbol{\delta A}) \leqq \mathrm{rg}(\boldsymbol{A}), \quad \kappa = \|\boldsymbol{A}^+\|_2\|\boldsymbol{\delta A}\|_2 < 1.$$

Dann gilt

1. *Der Fehler $\boldsymbol{\delta x}$ besitzt die Darstellung*

   $$\boldsymbol{\delta x} = \boldsymbol{\delta x}' + O\left(\|\boldsymbol{\delta A}\|_2(\|\boldsymbol{\delta A}\|_2 + \|\boldsymbol{\delta y}\|_2)\right)$$

   mit dem bezüglich $\boldsymbol{\delta A}$ und $\boldsymbol{\delta y}$ linearen Teil

   $$\boldsymbol{\delta x}' = \boldsymbol{A}^+(-\boldsymbol{\delta A}\boldsymbol{x}+\boldsymbol{\delta y}) + \boldsymbol{A}^+\boldsymbol{A}^{+T}\boldsymbol{\delta A}^T\boldsymbol{r} + (\boldsymbol{I}-\boldsymbol{A}^+\boldsymbol{A})\boldsymbol{\delta A}^T\boldsymbol{A}^{+T}\boldsymbol{x}.$$

   Dabei bezeichnet $\boldsymbol{r} = \boldsymbol{y}-\boldsymbol{A}\boldsymbol{x} = (\boldsymbol{I}-\boldsymbol{A}\boldsymbol{A}^+)\boldsymbol{x}$ wie üblich das Residuum von $\boldsymbol{x}$.
2. *Der Fehler $\boldsymbol{\delta x}$ genügt der Abschätzung*

   $$\|\boldsymbol{\delta x}\|_2 \leqq \frac{\|\boldsymbol{A}^+\|_2}{1-\kappa}\left[\|\boldsymbol{\delta A}\|_2\left(\omega\|\boldsymbol{x}\|_2 + \|\boldsymbol{A}^+\|_2\|\boldsymbol{r}\|_2\right) + \|\boldsymbol{y}\|_2\right]$$

   und für $\boldsymbol{x} \neq \boldsymbol{o}$

   $$\frac{\|\boldsymbol{\delta x}\|_2}{\|\boldsymbol{x}\|_2} \leqq \frac{1}{1-\kappa}\left[\left(\omega\mathrm{cond}(\boldsymbol{A}) + \frac{\|\boldsymbol{r}\|_2}{\|\boldsymbol{A}\|_2\|\boldsymbol{x}\|_2}\mathrm{cond}^2(\boldsymbol{A})\right)\frac{\|\boldsymbol{\delta A}\|_2}{\|\boldsymbol{A}\|_2} + \frac{\|\boldsymbol{A}^+\|_2\|\boldsymbol{y}\|_2}{\|\boldsymbol{x}\|_2}\frac{\|\boldsymbol{\delta y}\|_2}{\|\boldsymbol{y}\|_2}\right]$$

   mit $\mathrm{cond}(\boldsymbol{A}) = \|\boldsymbol{A}\|_2\|\boldsymbol{A}^+\|_2$ *und*

   $$\omega = \begin{cases} \sqrt{2} & fr \quad \mathrm{rg}(\boldsymbol{A}) < n \\ 1 & fr \quad \mathrm{rg}(\boldsymbol{A}) = n \end{cases}.$$

Beweis: Mit den Bezeichnungen aus dem Beweis von Satz 10.7 gilt

$$\begin{aligned} \boldsymbol{\delta x} &= (\boldsymbol{A}+\boldsymbol{\delta A})^+(\boldsymbol{y}+\boldsymbol{\delta y}) - \boldsymbol{A}^+\boldsymbol{y}, \\ &= \left[(\boldsymbol{A}+\boldsymbol{\delta A})^+ - \boldsymbol{A}^+\right]\boldsymbol{y} + (\boldsymbol{A}+\boldsymbol{\delta A})^+\boldsymbol{\delta y}, \\ &= \boldsymbol{G}\boldsymbol{y} + \tilde{\boldsymbol{A}}^+\boldsymbol{\delta y}, \\ &= \boldsymbol{G}_1\boldsymbol{y} + \boldsymbol{G}_2\boldsymbol{y} + \boldsymbol{G}_3\boldsymbol{y} + \tilde{\boldsymbol{A}}^+\boldsymbol{\delta y}. \end{aligned}$$

Nun gilt

$$\begin{aligned}
\|\boldsymbol{G}_1\boldsymbol{y}\|_2 &= \|\tilde{\boldsymbol{A}}^+\boldsymbol{\delta A}\boldsymbol{A}^+\boldsymbol{y}\|_2, \\
&= \|\tilde{\boldsymbol{A}}^+\boldsymbol{\delta A}\boldsymbol{x}\|_2, \\
&\leqq \|\tilde{\boldsymbol{A}}^+\|_2\|\boldsymbol{\delta A}\|_2\|\boldsymbol{x}\|_2, \\
&\leqq \frac{\|\boldsymbol{A}^+\|_2}{1-\kappa}\|\boldsymbol{\delta A}\|_2\|\boldsymbol{x}\|_2.
\end{aligned}$$

und

$$\begin{aligned}
\|\boldsymbol{G}_3\boldsymbol{y}\|_2 &= \|(\boldsymbol{I}-\tilde{\boldsymbol{S}})\boldsymbol{\delta A}^T\boldsymbol{A}^{+T}\boldsymbol{A}^+\boldsymbol{y}\|_2, \\
&= \|\boldsymbol{\delta A}\|_2\|\boldsymbol{A}^+\|_2\|\boldsymbol{x}\|_2.
\end{aligned}$$

Weiter gilt

$$\|\boldsymbol{G}_1\boldsymbol{y}+\boldsymbol{G}_3\boldsymbol{y}\|_2^2 = \|\boldsymbol{G}_1\boldsymbol{y}\|_2^2 + 2\boldsymbol{y}^T\boldsymbol{G}_1^T\boldsymbol{G}_3\boldsymbol{y} + \|\boldsymbol{G}_3\boldsymbol{y}\|_2^2.$$

Wegen

$$\boldsymbol{G}_1^T\boldsymbol{G}_3 = \boldsymbol{A}^{+T}\boldsymbol{\delta A}^T\underbrace{\tilde{\boldsymbol{A}}^{+T}(\boldsymbol{I}-\tilde{\boldsymbol{S}})}_{=\boldsymbol{O}}\boldsymbol{\delta A}^T\boldsymbol{A}^{+T}\boldsymbol{A}^+ = \boldsymbol{O}$$

folgt

$$\begin{aligned}
\|\boldsymbol{G}_1\boldsymbol{y}+\boldsymbol{G}_3\boldsymbol{y}\|_2^2 &= \|\boldsymbol{G}_1\boldsymbol{y}\|_2^2 + \|\boldsymbol{G}_3\boldsymbol{y}\|_2^2, \\
&\leqq \left(\frac{\|\boldsymbol{A}^+\|_2\|\boldsymbol{\delta A}\|_2\|\boldsymbol{x}\|_2}{1-\kappa}\right)^2 + \left(\|\boldsymbol{A}^+\|_2\|\boldsymbol{\delta A}\|_2\|\boldsymbol{x}\|_2\right)^2, \\
&\leqq 2\left(\frac{\|\boldsymbol{A}^+\|_2\|\boldsymbol{\delta A}\|_2\|\boldsymbol{x}\|_2}{1-\kappa}\right)^2, \\
\|\boldsymbol{G}_1\boldsymbol{y}+\boldsymbol{G}_3\boldsymbol{y}\|_2 &\leqq \sqrt{2}\frac{\|\boldsymbol{A}^+\|_2}{1-\kappa}\|\boldsymbol{\delta A}\|_2\|\boldsymbol{x}\|_2.
\end{aligned}$$

Weiter gilt

$$\begin{aligned}
\|\boldsymbol{G}_2\boldsymbol{y}\|_2 &= \|\tilde{\boldsymbol{A}}^+(\boldsymbol{I}-\boldsymbol{P})\boldsymbol{y}\|_2, \\
&= \|\tilde{\boldsymbol{A}}^+(\boldsymbol{I}-\boldsymbol{A}\boldsymbol{A}^+)\boldsymbol{y}\|_2, \\
&= \|\tilde{\boldsymbol{A}}^+\boldsymbol{r}\|_2, \\
&= \|\tilde{\boldsymbol{A}}^+(\boldsymbol{I}-\boldsymbol{A}\boldsymbol{A}^+)\boldsymbol{r}\|_2, \\
&= \|\boldsymbol{G}_2\boldsymbol{r}\|_2, \\
&\leqq \|\boldsymbol{G}_2\|_2\|\boldsymbol{r}\|_2, \\
&\leqq \frac{\|\boldsymbol{A}^+\|_2^2}{1-\kappa}\|\boldsymbol{\delta A}\|_2\|\boldsymbol{r}\|_2
\end{aligned}$$

und

$$\begin{aligned}\|\tilde{\boldsymbol{A}}^+\boldsymbol{\delta y}\|_2 &\leqq \|\tilde{\boldsymbol{A}}^+\|_2\|\boldsymbol{\delta y}\|_2,\\ &\leqq \frac{\|\boldsymbol{A}^+\|_2}{1-\kappa}\|\boldsymbol{\delta y}\|_2.\end{aligned}$$

Insgesamt erhalten wir

$$\begin{aligned}\|\boldsymbol{\delta x}\|_2 &\leqq \|\boldsymbol{G}_1\boldsymbol{y}+\boldsymbol{G}_3\boldsymbol{y}\|_2+\|\boldsymbol{G}_2\boldsymbol{y}\|_2+\|\tilde{\boldsymbol{A}}^+\boldsymbol{\delta y}\|_2,\\ &\leqq \sqrt{2}\frac{\|\boldsymbol{A}^+\|_2}{1-\kappa}\|\boldsymbol{\delta A}\|_2\|\boldsymbol{x}\|_2+\frac{\|\boldsymbol{A}^+\|_2^2}{1-\kappa}\|\boldsymbol{\delta A}\|_2\|\boldsymbol{r}\|_2+\frac{\|\boldsymbol{A}^+\|_2}{1-\kappa}\|\boldsymbol{\delta y}\|_2.\end{aligned}$$

Damit ist die Abschätzung für den absoluten Fehler bewiesen. Für $\mathrm{rg}(\boldsymbol{A})=n$ gilt

$$\tilde{\boldsymbol{A}}^+=\left(\tilde{\boldsymbol{A}}^T\tilde{\boldsymbol{A}}\right)^{-1}\tilde{\boldsymbol{A}}^T$$

und damit

$$\boldsymbol{I}-\tilde{\boldsymbol{S}}=\boldsymbol{I}-\left(\tilde{\boldsymbol{A}}^T\tilde{\boldsymbol{A}}\right)^{-1}\tilde{\boldsymbol{A}}^T\tilde{\boldsymbol{A}}=\boldsymbol{O}.$$

Dann ist $\boldsymbol{G}_3=\boldsymbol{O}$ und der Faktor $\sqrt{2}$ in der Abschätzung ist durch 1 zu ersetzen. Die Abschätzung des relativen Fehlers folgt sofort aus der Abschätzung für den absoluten Fehler nach Division durch $\|\boldsymbol{x}\|_2$. Die linearisierte Darstellung von $\boldsymbol{\delta x}$ ergibt sich aus

$$\begin{aligned}\boldsymbol{\delta x} &= \boldsymbol{G}_1\boldsymbol{y}+\boldsymbol{G}_2\boldsymbol{y}+\boldsymbol{G}_3\boldsymbol{y}+\tilde{\boldsymbol{A}}^+\boldsymbol{\delta y},\\ &= -\tilde{\boldsymbol{A}}^+\boldsymbol{\delta A}\boldsymbol{A}^+\boldsymbol{y}+\boldsymbol{G}_2\boldsymbol{r}+(\boldsymbol{I}-\tilde{\boldsymbol{S}})\boldsymbol{\delta A}^T\boldsymbol{A}^{+T}\boldsymbol{A}^+\boldsymbol{y}+\tilde{\boldsymbol{A}}^+\boldsymbol{\delta y},\\ &= -\tilde{\boldsymbol{A}}^+\boldsymbol{\delta A}\boldsymbol{x}+\tilde{\boldsymbol{A}}^+\tilde{\boldsymbol{A}}^{+T}\boldsymbol{\delta A}^T\boldsymbol{r}+(\boldsymbol{I}-\tilde{\boldsymbol{S}})\boldsymbol{\delta A}^T\boldsymbol{A}^{+T}\boldsymbol{x}+\boldsymbol{\delta y}.\end{aligned}$$

Mit $\tilde{\boldsymbol{A}}^+=\boldsymbol{A}^++O(\|\boldsymbol{\delta A}\|_2)$ folgt

$$\begin{aligned}\boldsymbol{\delta x}=\boldsymbol{A}^+(-\boldsymbol{\delta A}\boldsymbol{x}+\boldsymbol{\delta y})+\boldsymbol{A}^+\boldsymbol{A}^{+T}\boldsymbol{\delta A}^T\boldsymbol{r}+(\boldsymbol{I}-\boldsymbol{A}^+\boldsymbol{A})\boldsymbol{\delta A}^T\boldsymbol{A}^{+T}\boldsymbol{x}\\ +O\left(\|\boldsymbol{\delta A}\|_2(\|\boldsymbol{\delta A}\|_2+\|\boldsymbol{\delta y}\|_2)\right).\end{aligned}$$

❀

Bemerkungen: (i) Für konsistente Probleme ($\boldsymbol{r}=\boldsymbol{o}$) erhält man die Abschätzung

$$\frac{\|\boldsymbol{\delta x}\|_2}{\|\boldsymbol{x}\|_2}\leqq\frac{1}{1-\kappa}\left[\omega\mathrm{cond}(\boldsymbol{A})\frac{\|\boldsymbol{\delta A}\|_2}{\|\boldsymbol{A}\|_2}+\frac{\|\boldsymbol{A}^+\|_2\|\boldsymbol{y}\|_2}{\|\boldsymbol{x}\|_2}\frac{\|\boldsymbol{\delta y}\|_2}{\|\boldsymbol{y}\|_2}\right].$$

Bis auf den Faktor ω entspricht das genau der Fehlerabschätzung bei linearen Gleichungssystemen.
(ii) Für inkonsistente Probleme ($\boldsymbol{r} \neq \boldsymbol{o}$) tritt zusätzlich der Term

$$\frac{\|\boldsymbol{r}\|_2}{\|\boldsymbol{A}\|_2\|\boldsymbol{x}\|_2}\text{cond}^2(\boldsymbol{A})$$

auf. Für hinreichend kleines $\|\boldsymbol{r}\|_2$ spielt er keine Rolle. Das Fehlerverhalten ist proportional zu $\text{cond}(\boldsymbol{A})$ wie bei konsistenten Problemen. Für stark inkonsistente Probleme wird dieser Term aber bestimmend. Das Fehlerverhalten ist dann proportional zu $\text{cond}^2(\boldsymbol{A})$ und somit bedeutend schlechter.
(iii) Der Fehlereinfluss von Störungen der Matrix hängt bei Ausgleichsproblemen über das Residuum auch von der rechten Seite $\boldsymbol{y}$ ab.

10.3. Normalgleichungsverfahren

Wie wir im Abschnitt 10.1. gesehen haben, ist das Lösen des linearen Ausgleichsproblems

$$\min_{\boldsymbol{x}\in\mathbb{R}^n} \|\boldsymbol{y} - \boldsymbol{A}\boldsymbol{x}\|_2$$

dem Lösen des Normalgleichungssystems

$$\boldsymbol{A}^T\boldsymbol{A}\boldsymbol{x} = \boldsymbol{A}^T\boldsymbol{y}$$

äquivalent. Im spaltenregulären Falle ($\text{rg}(\boldsymbol{A}) = n$) ist $\boldsymbol{A}^T\boldsymbol{A}$ regulär und positiv definit. Das Normalgleichungssystem ist dann eindeutig lösbar. Für $\text{rg}(\boldsymbol{A}) < n$ ist die Matrix $\boldsymbol{A}^T\boldsymbol{A}$ singulär. Das Normalgleichungssystem ist dann aber immer noch konsistent, hat aber unendlich viele Lösungen. Wir werden uns darum auf den spaltenregulären Fall beschränken. Hier bietet sich zum Lösen das CHOLESKY-Verfahren an.

10.9. Lösen des Normalgleichungssystems mit Hilfe einer CHOLESKY-Zerlegung:
Es ist das lineare Ausgleichsproblem

$$\min_{\boldsymbol{x}\in\mathbb{R}^n} \|\boldsymbol{y} - \boldsymbol{A}\boldsymbol{x}\|_2$$

mit der spaltenregulären Matrix (m,n)-$\boldsymbol{A}$ zu lösen.
S0 *Berechne die (n,n)-Matrix $\boldsymbol{M} = \boldsymbol{A}^T\boldsymbol{A}$.*

S1 *Bestimme eine* CHOLESKY*-Zerlegung* $\boldsymbol{M} = \boldsymbol{L}\boldsymbol{L}^T$ *mit einer unteren Dreiecksmatrix* $\boldsymbol{L}$.

S2 *Berechne* $\boldsymbol{z} = \boldsymbol{A}^T\boldsymbol{y}$.

S3 *Löse die Dreieckssysteme* $\boldsymbol{L}\boldsymbol{w} = \boldsymbol{z}$ *und* $\boldsymbol{L}^T\boldsymbol{x} = \boldsymbol{w}$.

Aufwand:

S0 $\sim mn^2/2$ *Additionen/Multiplikationen,*

S1 $\sim n^3/6$ *Additionen/Multiplikationen+* n *Quadratwurzeln,*

S2 $\sim mn$ *Additionen/Multiplikationen,*

S3 $\sim n^2$ *Additionen/Multiplikationen*

Bemerkungen:
(i) In praktischen Fällen ist meist m größer als n. Dann liegt der Hauptaufwand im Berechnen der Matrix $\boldsymbol{M}$.
(ii) Die Matrix $\boldsymbol{M}$ sollte als unteres Dreieck gesondert gespeichert werden, damit die Matrix $\boldsymbol{A}$ für eine Nachiteration weiter zur Verfügung steht.
Für das Rundungsfehlerverhalten dieses Verfahrens geben wir den folgenden Satz ohne Beweis an.

10.10. Satz: *Für eine spaltenreguläre* (m,n)*-Matrix* $\boldsymbol{A}$ *ist der Algorithmus durchführbar, falls* $\boldsymbol{A}$ *im Sinne von*

$$\hat{\kappa} = \text{eps} N_1 \text{cond}^2(\boldsymbol{A}) \leqq \frac{1}{2}$$

mit

$$\begin{aligned} N_1 &= mn + F_1 \approx mn, \\ F_1 &= n^{3/2} + n + n^{1/2}F \approx 2n^{3/2}, \\ F &= n + 1 + \frac{1}{2}\ln n \approx n \end{aligned}$$

nicht zu schlecht konditioniert ist. Die berechneten Größen $\hat{\boldsymbol{M}}$, $\hat{\boldsymbol{L}}$, $\hat{\boldsymbol{z}}$ *und* $\hat{\boldsymbol{x}}$ *genügen den Beziehungen*

$$\begin{aligned} \hat{\boldsymbol{M}} &= \boldsymbol{A}^T\boldsymbol{A} + \boldsymbol{\delta}_0\boldsymbol{M}, \qquad \|\boldsymbol{\delta}_0\boldsymbol{M}\|_2 \leqq \text{eps} mn \|\boldsymbol{A}\|_2^2, \\ \hat{\boldsymbol{z}} &= (\boldsymbol{A} + \boldsymbol{\delta}_0\boldsymbol{A})^T\boldsymbol{y}, \qquad \|\boldsymbol{\delta}_0\boldsymbol{A}\|_2 \leqq \text{eps} m\sqrt{n}\|\boldsymbol{A}\|_2, \\ \hat{\boldsymbol{L}}\hat{\boldsymbol{L}}^T &= \boldsymbol{M} + \boldsymbol{\delta}_1\boldsymbol{M}, \qquad \|\boldsymbol{\delta}_1\boldsymbol{M}\|_2 \leqq \text{eps} nF \|\boldsymbol{M}\|_2, \\ (\boldsymbol{M} + \boldsymbol{\delta}_2\boldsymbol{M})\hat{\boldsymbol{x}} &= \hat{\boldsymbol{x}}, \qquad \|\boldsymbol{\delta}_2\boldsymbol{M}\|_2 \leqq \text{eps} F_1 \|\boldsymbol{M}\|_2. \end{aligned}$$

Bemerkung: Die Bedingung $\hat{\kappa} \leqq 0.5$ schränkt die Klasse der mit diesem Algorithmus lösbaren Probleme wesentlich ein. Die Klasse der Probleme, für die das lineare Ausgleichsproblem korrekt gestellt ist, ist nach Satz 10.8 durch

$$\kappa = \|\boldsymbol{A}^+\|_2 \|\boldsymbol{\delta A}\|_2 < 1$$

charakterisiert. Nimmt man für $\boldsymbol{\delta A}$ nur den Darstellungsfehler auf dem Rechner an, so gilt $\|\boldsymbol{\delta A}\|_2 \leqq \text{eps}\sqrt{n}\|\boldsymbol{A}\|_2$ und daher

$$\kappa = \text{eps}\sqrt{n}\text{cond}(\boldsymbol{A}) < 1.$$

Damit ist für alle Matrizen mit

$$\text{cond}(\boldsymbol{A}) < \frac{1}{\text{eps}\sqrt{n}}$$

das lineare Ausgleichsproblem korrekt gestellt. Aber nur für Matrizen mit

$$\text{cond}(\boldsymbol{A}) < \frac{1}{\sqrt{2\text{eps}mn}}$$

ist das Normalgleichungsverfahren anwendbar.
Mit Hilfe von Satz 10.10 lässt sich auch der erzeugte Rundungsfehler abschätzen.

10.11. Satz: *Unter den Voraussetzungen von Satz 10.10 gilt für den erzeugten Rundungsfehler $\boldsymbol{\delta x} = \hat{\boldsymbol{x}} - \boldsymbol{x}$, $\boldsymbol{x} = \boldsymbol{A}^+\boldsymbol{y} = (\boldsymbol{A}^T\boldsymbol{A})^{-1}\boldsymbol{A}^T\boldsymbol{y}$, die Abschätzung*

$$\|\boldsymbol{\delta x}\|_2 \leqq \text{eps}\frac{\text{cond}^2(\boldsymbol{A})}{1-\hat{\kappa}}\left[N_1\|\boldsymbol{x}\|_2 + m\sqrt{n}\frac{\|\boldsymbol{y}\|_2}{\|\boldsymbol{A}\|_2}\right].$$

Eine weitere Analyse zeigt, dass der erzeugte Rundungsfehler den unvermeidbaren Fehler beliebig stark übersteigen kann. Dieser Fall liegt besonders vor, falls die Norm des Residuums klein ist und $\|\boldsymbol{A}^+\boldsymbol{y}\|_2$ in der Größenordnung von $\|\boldsymbol{A}^+\|_2\|\boldsymbol{y}\|_2$ liegt. Andererseits ist das Normalgleichungsverfahren für stark inkonsistente Probleme ($\|\boldsymbol{r}\|_2 \gg 1$) mit kleinen Lösungen ($\|\boldsymbol{x}\|_2 \ll 1$)stabil und gutartig. Das sind gerade die Ausgleichsprobleme, bei denen die Fehlerverstärkung durch $\text{cond}^2(\boldsymbol{A})$ bestimmt wird.
Die numerische Instabilität des Normalgleichungsverfahrens lässt sich in ihrer Wirkung durch eine Nachiteration mildern.

10.12. Normalgleichungsverfahren mit Nachiteration:
Es ist das lineare Ausgleichsproblem

$$\min_{\boldsymbol{x}\in\mathbb{R}^n} \|\boldsymbol{y} - \boldsymbol{A}\boldsymbol{x}\|_2$$

mit der spaltenregulären (m,n)-Matrix $\boldsymbol{A}$ zu lösen.

S0 *Berechne die* (n,n)*-Matrix* $\boldsymbol{M} = \boldsymbol{A}^T\boldsymbol{A}$*.*
Bestimme eine CHOLESKY*-Zerlegung* $\boldsymbol{M} = \boldsymbol{L}\boldsymbol{L}^T$ *mit einer unteren Dreiecksmatrix* $\boldsymbol{L}$*.*
Setze $\boldsymbol{x}^{(0)} = \boldsymbol{o}$ *und* $k = 0$*.*

S1 *Berechne* $\boldsymbol{r}^{(k)} = \boldsymbol{y} - \boldsymbol{A}\boldsymbol{x}^{(k)}$ *und* $\boldsymbol{g}^{(k)} = \boldsymbol{A}^T\boldsymbol{r}^{(k)}$*.*

S2 *Löse die Dreieckssysteme*

$$\boldsymbol{L}\boldsymbol{w} = \boldsymbol{g}^{(k)}, \quad \boldsymbol{L}^T\boldsymbol{h}^{(k)} = \boldsymbol{w}.$$

S3 *Setze* $\boldsymbol{x}^{(k+1)} = \boldsymbol{x}^{(k)} + \boldsymbol{h}^{(k)}$*,* $k = k+1$ *und gehe zu Schritt* **S1**

Im ersten Schritt ergibt sich mit $\boldsymbol{x}^{(1)} = \boldsymbol{h}^{(0)}$ wieder die Lösung des üblichen Normalgleichungsverfahrens.
Nach hinreichend vielen Iterationen liefert dieses Verfahren eine Lösung $\boldsymbol{x}^{(k)}$, die als exakte Lösung eines benachbarten Ausgleichsproblems

$$\min_{\boldsymbol{x}\in\mathbb{R}^n} \|\boldsymbol{y} - (\boldsymbol{A} + \boldsymbol{\delta A})\boldsymbol{x}\|_2$$

interpretierbar ist, falls die spaltenreguläre Matrix $\boldsymbol{A}$ der Bedingung

$$\overline{\kappa} = \text{eps}\overline{N}_1\text{cond}^2(\boldsymbol{A}) \leqq 0.23$$

mit $\overline{N}_1 = (m+1)(n+1) + 2n^2$ genügt. Unter diesen Voraussetzungen liegt numerische Gutartigkeit vor.
Man beachte:
Hier spielt die Nachiteration eine ganz andere Rolle als bei linearen Gleichungssystemen. Bei linearen Gleichungssystemen diente sie zur Genauigkeitsverbesserung der Lösung eines ohnehin gutartigen Verfahrens. Hier wird dagegen erst durch die Nachiteration die Gutartigkeit des Verfahrens erreicht.
Die Konvergenzgeschwindigkeit des Verfahrens hängt entscheidend von $\overline{\kappa}$ ab. Die Klasse der lösbaren Probleme ist durch die Bedingung $\overline{\kappa} \leqq 0.23$ wegen $\overline{\kappa} \sim \text{cond}^2(\boldsymbol{A})$ immer noch wesentlich eingeschränkt!

10.4. Orthogonalisierungsverfahren

Das in Abschnitt 8.2.5 beschriebene HOUSEHOLDER-Verfahren zum Lösen linearer Gleichungssysteme lässt sich unmittelbar auf das Lösen linearer Ausgleichsprobleme übertragen. Wegen der Invarianz der euklidischen Vektornorm gegenüber orthogonalen Transformationen sind nämlich die Probleme

$$\min_{\boldsymbol{x}\in\mathbb{R}^n} \|\boldsymbol{y} - \boldsymbol{A}\boldsymbol{x}\|_2 \tag{10.2}$$

und

$$\min_{\boldsymbol{x}\in\mathbb{R}^n} \|\boldsymbol{h}-\boldsymbol{S}\boldsymbol{x}\|_2, \tag{10.3}$$

wobei $\boldsymbol{h}=\boldsymbol{Q}^T\boldsymbol{y}$ und $\boldsymbol{S}=\boldsymbol{Q}^T\boldsymbol{A}$ mit einer orthogonalen (m,m)-Matrix $\boldsymbol{Q}$ gilt, äquivalent. Es kommt nun darauf an, die orthogonale Matrix $\boldsymbol{Q}$ so zu wählen, dass das Problem 10.3 möglichst einfach wird. Das ist zum Beispiel der Fall, wenn die Matrix $\boldsymbol{S}$ die Struktur $\boldsymbol{S}=\begin{pmatrix}\boldsymbol{R}\\ \boldsymbol{O}\end{pmatrix}$ mit einer oberen (n,n)-Dreiecksmatrix $\boldsymbol{R}$ besitzt. Mit der Partitionierung $\boldsymbol{h}=\begin{pmatrix}\boldsymbol{h}_1\\ \boldsymbol{h}_2\end{pmatrix}$, $\boldsymbol{h}_1\in\mathbb{R}^n$ und $\boldsymbol{h}_2\in\mathbb{R}^{m-n}$, gilt dann

$$\begin{aligned}\|\boldsymbol{y}-\boldsymbol{A}\boldsymbol{x}\|_2^2 &= \|\boldsymbol{h}-\boldsymbol{S}\boldsymbol{x}\|_2^2\\ &= \left\|\begin{pmatrix}\boldsymbol{h}_1\\ \boldsymbol{h}_2\end{pmatrix}-\begin{pmatrix}\boldsymbol{R}\\ \boldsymbol{O}\end{pmatrix}\boldsymbol{x}\right\|_2^2\\ &= \left\|\begin{pmatrix}\boldsymbol{h}_1-\boldsymbol{R}\boldsymbol{x}\\ \boldsymbol{h}_2\end{pmatrix}\right\|_2^2\\ &= \|\boldsymbol{h}_1-\boldsymbol{R}\boldsymbol{x}\|_2^2+\|\boldsymbol{h}_2\|_2^2.\end{aligned}$$

Der zweite Summand in der letzten Gleichung ist konstant. Er spielt daher bei der Minimierung keine Rolle. Für den ersten Summanden gilt

$$\|\boldsymbol{h}_1-\boldsymbol{R}\boldsymbol{x}\|_2^2 \geqq 0, \quad \|\boldsymbol{h}_1-\boldsymbol{R}\boldsymbol{x}\|_2^2=0 \Longleftrightarrow \boldsymbol{R}\boldsymbol{x}=\boldsymbol{h}_1.$$

Die Lösung des linearen Ausgleichsproblems 10.2 ergibt sich somit nach der Transformation auf das Problem 10.3 aus dem linearen Gleichungssystem $\boldsymbol{R}\boldsymbol{x}=\boldsymbol{h}_1$. Der Algorithmus zur QR-Zerlegung aus Abschnitt 8.2.5 lässt sich direkt auf das Problem übertragen.

10.13. Lösen eines linearen Ausgleichsproblems mit HOUSEHOLDER-Transformationen:
Es ist das lineare Ausgleichsproblem

$$\min_{\boldsymbol{x}\in\mathbb{R}^n} \|\boldsymbol{y}-\boldsymbol{A}\boldsymbol{x}\|_2$$

mit der spaltenregulären (m,n)-Matrix $\boldsymbol{A}$ zu lösen.

{Initialisierung}
Wähle eine Genauigkeitsschranke $\varepsilon>0$.
{QR-Zerlegung}
for $k=1$ ***to*** n ***do***

{Berechnen der Transformationsmatrix $\boldsymbol{H}^{(k)}$}
$\varrho_k = \sqrt{a_{kk}^2 + a_{k+1,k}^2 + \cdots + a_{mk}^2}$
if $\varrho_k \leq \varepsilon$ ***then***
STOPP
endif
if $a_{kk} > 0$ ***then***
$\varrho_k = -\varrho_k$
endif
$a_{kk} = a_{kk} - \varrho_k$
$\gamma_k = -\varrho_k \cdot a_{kk}$
{Transformation der Restmatrix}
for $j = k+1$ ***to*** n ***do***
$\beta = 0$
for $i = k$ ***to*** m ***do***
$\beta = \beta + a_{ik} \cdot a_{ij}$
endfor
$\beta = \beta / \gamma_k$
for $i = k$ ***to*** m ***do***
$a_{ij} = a_{ij} - \beta \cdot a_{ik}$
endfor
{Transformation des Vektors $\boldsymbol{y}$}
$\beta = 0$
for $i = k$ ***to*** m ***do***
$\beta = \beta + a_{ik} \cdot y_i j$
endfor
$\beta = \beta / \gamma_k$
for $i = k$ ***to*** m ***do***
$y_i = y_i - \beta \cdot a_{ik}$
endfor
endfor
endfor
$\varrho_n = a_{nn}$
{Lösen von $\boldsymbol{Rx} = \boldsymbol{c}$}
for $k = n$ ***to*** 1 ***step*** -1 ***do***
for $i = k+1$ ***to*** n ***do***
$y_k = y_k - y_i \cdot a_{ki}$
endfor
$y_k = y_k / \varrho_k$
endfor

Aufwand:
Transf.: $\sim(m-n/3)n^2$ Additionen/Multiplikationen, n Quadratwurzeln,

Lösen: $\sim n^2/2$ Additionen/Multiplikationen.

Für $m \gg n$ ist der Aufwand damit etwa doppelt so groß wie der des Normalgleichungsverfahrens. Dafür zeigt aber eine Rundungsfehleranalyse, dass der Algorithmus gutartig ist und für eine große Klasse von Matrizen anwendbar ist. Wir geben den entsprechenden Satz ohne Beweis an.

10.14. Satz: *Für eine spaltenreguläre (m,n)-Matrix $\boldsymbol{A}$ ist der Algorithmus mit $r_{kk} \neq 0$ für $k = 1,\ldots,n$ durchführbar, falls $\boldsymbol{A}$ der Bedingung*

$$\kappa = \text{eps}F\text{cond}(\boldsymbol{A}) < 1$$

mit

$$F = Kmn^{3/2}, \quad K = 3.14(1+3/m) \approx 4$$

genügt. Für einen beliebigen Vektor $\boldsymbol{y} \in \mathbb{R}^m$ liefert der Algorithmus eine Lösung $\hat{\boldsymbol{x}} \in \mathbb{R}^n$, die als exakte Lösung eines benachbarten Problems

$$\min_{\boldsymbol{x}\in\mathbb{R}^n} \|(\boldsymbol{y}+\boldsymbol{\delta y}) - (\boldsymbol{A}+\boldsymbol{\delta A})\boldsymbol{x}\|_2$$

interpretierbar ist. Die Störungen genügen den Abschätzungen

$$\begin{aligned} \|\boldsymbol{\delta A}\|_2 &\leqq \text{eps}F_1\|\boldsymbol{A}\|_2, \quad F_1 = F + n^{3/2} \approx F, \\ \|\boldsymbol{\delta y}\|_2 &\leqq \text{eps}\frac{F}{\sqrt{n}}\|\boldsymbol{y}\|_2. \end{aligned}$$

Für das berechnete Residuum $\hat{\boldsymbol{r}} = \boldsymbol{r} + \boldsymbol{\delta r}$ gilt

$$\boldsymbol{r} + \boldsymbol{\delta r} = \boldsymbol{y} + \boldsymbol{\delta y} - (\boldsymbol{A} + \boldsymbol{\delta A})\hat{\boldsymbol{x}}$$

mit

$$\|\boldsymbol{\delta r}\|_2 \leqq \text{eps}\frac{F}{\sqrt{n}}\|\boldsymbol{r}\|_2.$$

Bemerkungen (i) Das HOUSEHOLDER-Verfahren zum Lösen von linearen Ausgleichsproblemen ist ein gutartiger Prozess. Die Bedingung $\kappa = \text{eps}F\text{cond}(\boldsymbol{A}) < 1$ für die Anwendbarkeit des Verfahrens ist wesentlich schwächer als die Bedingung $\hat{\kappa} = \text{eps}N_1\text{cond}^2(\boldsymbol{A}) \leqq 0.5$ für die Anwendbarkeit des Normalgleichungsverfahrens.

Das HOUSEHOLDER-Verfahren ist trotz des etwa doppelt so großen Aufwands dem Normalgleichungsverfahren vorzuziehen.
(ii) Der Algorithmus zur QR-Zerlegung der rechteckigen (m,n)-Matrix $\boldsymbol{A}$ stimmt fast mit dem entsprechenden Algorithmus zur QR-Zerlegung einer quadratischen Matrix $\boldsymbol{A}$ überein. Es ist ein zusätzlicher n-ter Schritt auszuführen. Insbesondere darf die gleiche Speicherungsform gewählt werden.
(iii) In speziellen Anwendungen ist statt der Lösung $\boldsymbol{x}$ der Residuumsvektor $\boldsymbol{r} = \boldsymbol{y} - \boldsymbol{A}\boldsymbol{x}$ gesucht; dieser lässt sich ohne Kenntnis von $\boldsymbol{x}$ berechnen. Es gilt

$$\begin{aligned}
\boldsymbol{r} &= \boldsymbol{y} - \boldsymbol{A}\boldsymbol{x}, \\
&= \boldsymbol{Q}\boldsymbol{h} - \boldsymbol{Q}\boldsymbol{S}\boldsymbol{x}, \\
&= \boldsymbol{Q}\begin{pmatrix} \boldsymbol{h}_1 \\ \boldsymbol{h}_2 \end{pmatrix} - \boldsymbol{Q}\begin{pmatrix} \boldsymbol{R} \\ \boldsymbol{O} \end{pmatrix}\boldsymbol{R}^{-1}\boldsymbol{h}_1, \\
&= \boldsymbol{Q}\begin{pmatrix} \boldsymbol{h}_1 \\ \boldsymbol{h}_2 \end{pmatrix} - \boldsymbol{Q}\begin{pmatrix} \boldsymbol{h}_1 \\ \boldsymbol{o} \end{pmatrix}, \\
&= \boldsymbol{Q}\begin{pmatrix} \boldsymbol{o} \\ \boldsymbol{h}_2 \end{pmatrix}.
\end{aligned}$$

Auch bei Orthogonalisierungsverfahren ist eine Nachiteration möglich. Hier dient sie aber wieder nur einer Genauigkeitsverbesserung. Für spaltenreguläres $\boldsymbol{A}$ sei $\boldsymbol{x}^*$ die eindeutige Lösung des linearen Ausgleichsproblems. $\boldsymbol{r}^*$ sei das zugehörige minimale Residuum. Dann gilt $\boldsymbol{r}^* + \boldsymbol{A}\boldsymbol{x}^* = \boldsymbol{y}$ und $\boldsymbol{A}^T\boldsymbol{r}^* = \boldsymbol{o}$. Damit ist der Vektor $\begin{pmatrix} \boldsymbol{r}^* \\ \boldsymbol{x}^* \end{pmatrix}$ Lösung des linearen Gleichungssystems

$$\begin{pmatrix} \boldsymbol{I} & \boldsymbol{A} \\ \boldsymbol{A}^T & \boldsymbol{O} \end{pmatrix}\begin{pmatrix} \boldsymbol{r} \\ \boldsymbol{x} \end{pmatrix} = \begin{pmatrix} \boldsymbol{y} \\ \boldsymbol{o} \end{pmatrix}.$$

Es sei nun

$$\boldsymbol{A} + \boldsymbol{\delta A} = \boldsymbol{\mathcal{Q}}\begin{pmatrix} \boldsymbol{\mathcal{R}} \\ \boldsymbol{O} \end{pmatrix}$$

die näherungsweise QR-Zerlegung von $\boldsymbol{A}$. Wendet man auf das obige Gleichungssystem ein übliches Iterationsverfahren mit

$$\boldsymbol{B} = \begin{pmatrix} \boldsymbol{I} & \boldsymbol{\mathcal{Q}}\begin{pmatrix} \boldsymbol{\mathcal{R}} \\ \boldsymbol{O} \end{pmatrix} \\ (\boldsymbol{\mathcal{R}}^T, \boldsymbol{O})\,\boldsymbol{\mathcal{Q}}^T & \boldsymbol{O} \end{pmatrix}$$

an, so erhält man:

10.15. Nachiteration für QR-Zerlegung bei linearen Ausgleichsproblemen:

Es ist das lineare Ausgleichsproblem

$$\min_{\boldsymbol{x}\in\mathbb{R}^n} \|\boldsymbol{y}-\boldsymbol{A}\boldsymbol{x}\|_2$$

mit der spaltenregulären (m,n)-Matrix $\boldsymbol{A}$ zu lösen.

S0 *Berechne die QR-Zerlegung $\boldsymbol{A}=\boldsymbol{Q}\begin{pmatrix}\boldsymbol{R}\\ \boldsymbol{O}\end{pmatrix}$ mit einer orthogonalen (m,m)-Matrix $\boldsymbol{Q}$ und einer oberen (n,n)-Dreiecksmatrix $\boldsymbol{R}$.*
Setze $\boldsymbol{x}^{(0)}=\boldsymbol{o}$, $\boldsymbol{r}^{(0)}=\boldsymbol{y}$ und $k=0$.

S1 *Berechne $\boldsymbol{e}=\boldsymbol{y}-\boldsymbol{r}^{(k)}-\boldsymbol{A}\boldsymbol{x}^{(k)}$ und $\boldsymbol{f}=-\boldsymbol{A}^T\boldsymbol{r}^{(k)}$.*
Berechne

$$\boldsymbol{Q}^T\boldsymbol{e}=\begin{pmatrix}\mu\\ \boldsymbol{v}\end{pmatrix},\quad \mu\in\mathbb{R}^n,\quad \boldsymbol{v}\in\mathbb{R}^{m-n}.$$

S2 *Löse das Dreieckssystem $\boldsymbol{R}^T\boldsymbol{w}=\boldsymbol{f}$.*
Löse das Dreieckssystem $\boldsymbol{R}\boldsymbol{h}=\mu-\boldsymbol{w}$.

S3 *Berechne*

$$\boldsymbol{s}=\boldsymbol{Q}\begin{pmatrix}\boldsymbol{w}\\ \boldsymbol{v}\end{pmatrix}.$$

Setze $\boldsymbol{x}^{(k+1)}=\boldsymbol{x}^{(k)}+\boldsymbol{h}$, $\boldsymbol{r}^{(k+1)}=\boldsymbol{r}^{(k)}+\boldsymbol{s}$, $k=k+1$ und gehe zu Schritt **S1**

10.5. Aufgaben

1. Man zeige: Für eine beliebige (m,n)-Matrix $\boldsymbol{A}$ und einen beliebigen Vektor $\boldsymbol{y}\in\mathbb{R}^m$ ist das Gleichungssystem

   $$\boldsymbol{A}^T\boldsymbol{A}\boldsymbol{x}=\boldsymbol{A}^T\boldsymbol{y}$$

   stets lösbar.

2. Man zeige: Gilt für die (m,n)-Matrix $\boldsymbol{A}$ $\operatorname{rg}(\boldsymbol{A})=n$, so ist die Matrix $\boldsymbol{M}=\boldsymbol{A}^T\boldsymbol{A}$ regulär und positiv definit.

3. Es liege das folgende mathematische Gesetz mit zwei unbekannten Parametern α und β vor:

   $$z=\alpha y+\beta$$

   Weiterhin sei ein Satz von Messdaten $(y_i,z_i)\quad i=1,\ldots,m$ mit $y_i=i$ gegeben. Man bestimme mit Hilfe der linearen Ausgleichsrechnung die Parameter α und β!

(a) Wie lauten die Normalgleichungen?

(b) Wie lautet die CHOLESKY-Zerlegung der Matrix der Normalgleichungen? ($\boldsymbol{A}^T\boldsymbol{A} = \boldsymbol{L}\boldsymbol{L}^T$)

(c) Man gebe eine Abschätzung von $\text{cond}_2(\boldsymbol{L})$ an.

(d) Wie ändert sich die Kondition mit der Anzahl der Messdaten?

4. Für $\boldsymbol{A}, \tilde{\boldsymbol{A}}$ gelte rg($\boldsymbol{A}$) = rg($\tilde{\boldsymbol{A}}$) = r. Man zeige:

$$\|\tilde{\boldsymbol{P}}(\boldsymbol{I}-\boldsymbol{P})\|_2 = \|\boldsymbol{P}(\boldsymbol{I}-\tilde{\boldsymbol{P}})\|_2.$$

Hinweis: Unter Verwendung der Singulärwertzerlegung stelle man $\boldsymbol{P}$ und $\tilde{\boldsymbol{P}}$ gemäß
$\boldsymbol{P} = \boldsymbol{U}\Pi\boldsymbol{U}^T$ und $\tilde{\boldsymbol{P}} = \tilde{\boldsymbol{U}}\Pi\tilde{\boldsymbol{U}}^T$ mit

$$\Pi = \begin{pmatrix} \boldsymbol{I}^{(r)} & \boldsymbol{O} \\ \boldsymbol{O} & \boldsymbol{O} \end{pmatrix}$$

dar und partitioniere $\boldsymbol{W} = \tilde{\boldsymbol{U}}^T\boldsymbol{U}$ gemäß

$$\boldsymbol{W} = \begin{pmatrix} \boldsymbol{W}_{11} & \boldsymbol{W}_{12} \\ \boldsymbol{W}_{21} & \boldsymbol{W}_{22} \end{pmatrix}.$$

Man beweise:

$$\|\tilde{\boldsymbol{P}}(\boldsymbol{I}-\boldsymbol{P})\| = \|\boldsymbol{W}_{12}\| \text{ und } \|\boldsymbol{P}(\boldsymbol{I}-\tilde{\boldsymbol{P}})\| = \|\boldsymbol{W}_{21}\|$$

und folgere

$$\|\boldsymbol{W}_{12}\| = \|\boldsymbol{W}_{21}\|$$

aus der Orthogonalität von $\boldsymbol{W}$.

5. Für

$$\boldsymbol{y} \in \mathbb{R}^m \text{ und } \boldsymbol{G}_1 = \tilde{\boldsymbol{A}}^+\boldsymbol{P} - \tilde{\boldsymbol{S}}\boldsymbol{A}^+, \boldsymbol{G}_2 = \tilde{\boldsymbol{A}}^+(\boldsymbol{I}-\boldsymbol{P}), \boldsymbol{G}_3 = -(\boldsymbol{I}-\tilde{\boldsymbol{S}})\boldsymbol{A}^+$$

sowie

$$\boldsymbol{P} = \boldsymbol{A}\boldsymbol{A}^+, \tilde{\boldsymbol{P}} = \tilde{\boldsymbol{A}}\tilde{\boldsymbol{A}}^+, \tilde{\boldsymbol{S}} = \tilde{\boldsymbol{A}}^+\tilde{\boldsymbol{A}}, \boldsymbol{S} = \boldsymbol{A}^+\boldsymbol{A} \text{ mit } \|\boldsymbol{G}_i\| \leqq \alpha$$

werde

$$\boldsymbol{z} = \boldsymbol{G}_1\boldsymbol{y} + \boldsymbol{G}_3\boldsymbol{y} = \boldsymbol{z}_1 + \boldsymbol{z}_2 + \boldsymbol{z}_3$$

gebildet. Man zeige:

(a) $\boldsymbol{z}_1 \perp \boldsymbol{z}_3, \boldsymbol{z}_2 \perp \boldsymbol{z}_3$ folglich $\|\boldsymbol{z}\|^2 = \|\boldsymbol{z}_1 + \boldsymbol{z}_2\|^2 + \|\boldsymbol{z}_3\|^2$

(b) mit $\boldsymbol{u} = \boldsymbol{P}\boldsymbol{y}, \boldsymbol{v} = (\boldsymbol{I} - \boldsymbol{P})\boldsymbol{y}$ gilt:

$$\|\boldsymbol{u}\|^2 + \|\boldsymbol{v}\|^2 = \|\boldsymbol{y}\|^2 \text{ und } \boldsymbol{G}_1\boldsymbol{y} = \boldsymbol{G}_1\boldsymbol{u}, \boldsymbol{G}_2\boldsymbol{y} = \boldsymbol{G}_2\boldsymbol{v}, \boldsymbol{G}_3\boldsymbol{y} = \boldsymbol{G}_3\boldsymbol{u}.$$

Daraus folgere man weiter

$$\|\boldsymbol{z}\|^2 \leqq \alpha^2[(\|\boldsymbol{u}\| + \|\boldsymbol{v}\|)^2 + \|\boldsymbol{u}\|^2] = \alpha^2\|\boldsymbol{y}\|^2 N$$

mit

$$N = (\cos\varphi + \sin\varphi)^2 + \cos^2\varphi \leqq [\frac{1+\sqrt{5}}{2}]^2$$

(Dabei wurde $\|\boldsymbol{u}\| = \|\boldsymbol{y}\|\cos\varphi$ und $\|\boldsymbol{v}\| = \|\boldsymbol{y}\|\sin\varphi$ gesetzt.)

(c) Im Falle $\text{rg}(\boldsymbol{A}) = n$ ist $\boldsymbol{z}_3 = \boldsymbol{o}$, also

$$\|\boldsymbol{z}\|^2 = \|\boldsymbol{z}_1 + \boldsymbol{z}_2\|^2 \leqq 2\alpha^2\|\boldsymbol{y}\|^2.$$

Im Falle $\text{rg}(\boldsymbol{A}) = m$ ist $\boldsymbol{z}_2 = \boldsymbol{o}$, also

$$\|\boldsymbol{z}\|^2 = \|\boldsymbol{z}_1\|^2 + \|\boldsymbol{z}_3\|^2 \leqq 2\alpha^2\|\boldsymbol{y}\|^2.$$

Im Falle $\text{rg}(\boldsymbol{A}) = m = n$ ist $\|\boldsymbol{z}\| = \|\boldsymbol{z}_1\| \leqq \alpha\|\boldsymbol{y}\|$.

6. Es sei eine beliebige (m,n)-Matrix $\boldsymbol{A}$ gegeben. Man zeige

$$\lim_{\varrho \to 0}[\varrho\boldsymbol{I} + \boldsymbol{A}^T\boldsymbol{A}]^{-1}\boldsymbol{A}^T = \boldsymbol{A}^+.$$

Hinweis: Singulärwertzerlegung von $\boldsymbol{A}$ anwenden.

7. Für eine (m,n)-Matrix $\boldsymbol{A}$ mit $m \geqq n$ sei

$$\boldsymbol{M}(\boldsymbol{A}) = \begin{pmatrix} \boldsymbol{I} & \boldsymbol{A} \\ \boldsymbol{A}^T & \boldsymbol{O} \end{pmatrix}.$$

(a) Man zeige, dass $\boldsymbol{M}(\boldsymbol{A})$ genau dann regulär ist, wenn $\text{rg}(\boldsymbol{A}) = n$ gilt.

(b)

$$\boldsymbol{M}(\boldsymbol{A})^+ = \begin{pmatrix} \boldsymbol{I} - \boldsymbol{A}\boldsymbol{A}^+ & \boldsymbol{A}^{+T} \\ \boldsymbol{A}^+ & -\boldsymbol{A}^+\boldsymbol{A}^{+T} \end{pmatrix}.$$

Kapitel 11

Freie Minimierung

11.1. Einführung

11.1.1. Aufgabenstellung und grundlegende Begriffe

In diesem Kapitel wollen wir folgendes Problem behandeln:
Es sei eine Funktion $f : \mathbb{R}^n \longrightarrow \mathbb{R}$ gegeben. Zu bestimmen ist ein Punkt $\boldsymbol{x}^* \in \mathbb{R}^n$, der die Funktion f minimiert. Probleme dieser Art heißen freie Minimumprobleme bzw. Minimumprobleme ohne Nebenbedingungen (Restriktionen).
Wir unterscheiden zwischen globalen und lokalen Lösungen.
Ein Punkt $\boldsymbol{x}^* \in \mathbb{R}^n$, für den es eine Umgebung $U(\boldsymbol{x}^*)$ gibt, so dass $f(\boldsymbol{x}^*) \leqq f(\boldsymbol{x})$ für alle $\boldsymbol{x} \in U(\boldsymbol{x}^*)$ gilt, heißt **lokale Lösung** des Minimumproblems

$$\min_{\boldsymbol{x} \in \mathbb{R}^n} f(\boldsymbol{x}).$$

Im Falle $U(\boldsymbol{x}^*) = \mathbb{R}^n$ heißt der Punkt **globale Lösung** Alle globalen Lösungen eines beliebigen Minimumproblems zu finden, ist i. a. eine schwierige Aufgabe. Wir werden uns damit begnügen, nach lokalen Lösungen oder nach Punkten zu suchen, die als lokale Lösungen in Frage kommen.
Die meisten Verfahren zum Lösen von Minimumproblemen erzeugen ausgehend von einem Startpunkt $\boldsymbol{x}^{(0)} \in \mathbb{R}^n$ eine Folge $\{\boldsymbol{x}^{(k)}\}_{k \in \mathbb{N}}$ deren zugehörige Folge von Funktionswerten $\{f(\boldsymbol{x}^{(k)})\}_{k \in \mathbb{N}}$ monoton fallend ist.
Als Beispiel stellen wir uns folgende Situation vor. Wir befinden uns in einem Gebirge und werden von einem starken Gewitter überrascht. Natürlich wollen wir schnell talwärts in eine Hütte. Da wir nur die nahe Umgebung überblicken, werden wir eine Richtung wählen, in der es von unserem Standpunkt aus abwärts geht. In dieser Richtung gehen wir ein Stück voran und bestimmen dann wieder eine neue Richtung, usw.
Mathematisch gesprochen:

Wir sind in einem Punkt $\boldsymbol{x} \in \mathbb{R}^n$ und suchen eine Richtung $\boldsymbol{p} \in \mathbb{R}^n$, für die es ein $\alpha_0 > 0$ gibt, so dass

$$f(\boldsymbol{x} + \alpha \boldsymbol{p}) < f(\boldsymbol{x})$$

für alle $\alpha \in (0, \alpha_0)$ gilt. Gibt es eine solche **Abstiegsrichtung** $\boldsymbol{p}$, so bestimmen wir anschließend eine **Schrittweite** α mit $f(\boldsymbol{x} + \alpha \boldsymbol{p}) < f(\boldsymbol{x})$. Mit dem Punkt $\bar{\boldsymbol{x}} = \boldsymbol{x} + \alpha \boldsymbol{p}$ wird neu gestartet. Sind wir in einem Punkte $\boldsymbol{x}^*$ angelangt, zu dem es keine Abstiegsrichtung gibt, also

$$f(\boldsymbol{x}^* + \alpha \boldsymbol{p}) \geqq f(\boldsymbol{x}^*)$$

für alle Richtungen $\boldsymbol{p} \in \mathbb{R}^n$ und alle $\alpha \in (0, \alpha_0)$ gilt, so sind wir in einem Tal angelangt. Es ist nicht klar, ob es der tiefste Punkt in unserem Gebirge ist.
Wissen wir dagegen, dass eine Funktion f konvex ist, das heißt

$$f(\alpha \boldsymbol{x} + (1-\alpha)\boldsymbol{y}) \leqq \alpha f(\boldsymbol{x}) + (1-\alpha) f(\boldsymbol{y})$$

für alle $\boldsymbol{x} \in \mathbb{R}^n$ und alle $\alpha \in [0,1]$, so ist ein lokales Minimum auch globales Minimum. Das sieht man folgendermaßen ein. Für beliebige $\boldsymbol{x} \in \mathbb{R}^n$ gilt wegen der Konvexität

$$f(\boldsymbol{x}^* + \alpha(\boldsymbol{x} - \boldsymbol{x}^*)) \leqq (1-\alpha) f(\boldsymbol{x}^*) + \alpha f(\boldsymbol{x}) \quad \forall \alpha \in [0,1].$$

Daraus folgt

$$f(\boldsymbol{x}^* + \alpha(\boldsymbol{x} - \boldsymbol{x}^*)) - f(\boldsymbol{x}^*) \leqq \alpha (f(\boldsymbol{x}) - f(\boldsymbol{x}^*)) \quad \forall \alpha \in [0,1].$$

Da $\boldsymbol{x}^*$ lokales Minimum ist, gilt für $0 < \alpha < \min\{\alpha_0, 1\}$

$$0 \leqq \alpha (f(\boldsymbol{x}) - f(\boldsymbol{x}^*)).$$

Nach Division durch α folgt

$$f(\boldsymbol{x}^*) \leqq f(\boldsymbol{x}).$$

$\boldsymbol{x}^*$ ist somit globales Minimum.
Damit haben wir einige grundlegende Begriffe definiert.

11.1.2. Differenzierbarkeit und Richtungsableitung

Ist die Funktion $\boldsymbol{F} : \mathbb{R}^n \longrightarrow \mathbb{R}^m$ im Punkt $\boldsymbol{x}$ stetig partiell differenzierbar, das heißt, alle partiellen Ableitungen $\frac{\partial F_i}{\partial x_j}$, $i = 1, \ldots, m$ und $j = 1, \ldots, n$, existieren im Punkt $\boldsymbol{x}$ und sind stetig, so heißt $\boldsymbol{F}$ im Punkt $\boldsymbol{x}$ **stetig differenzierbar**. Die (m,n)-Matrix

$$\boldsymbol{F}'(\boldsymbol{x}) = \left(\frac{\partial F_i(\boldsymbol{x})}{\partial x_j} \right)_{\substack{i=1,\ldots,m \\ j=1,\ldots,n}}$$

heißt **Funktionalmatrix** von $\boldsymbol{F}$ in $\boldsymbol{x}$. Ist $f : \mathbb{R}^n \longrightarrow \mathbb{R}$, so heißt

$$f'(\boldsymbol{x}) = \boldsymbol{\nabla} f(\boldsymbol{x}) = \begin{pmatrix} \frac{\partial f(\boldsymbol{x})}{\partial x_1} \\ \vdots \\ \frac{\partial f(\boldsymbol{x})}{\partial x_n} \end{pmatrix}$$

Gradient von f in $\boldsymbol{x}$.
$\boldsymbol{F}$ heißt stetig differenzierbar auf einer offenen Menge $D \subset \mathbb{R}^n$ ($\boldsymbol{F} \in C^1(D)$), falls $\boldsymbol{F}$ für jedes $\boldsymbol{x} \in D$ stetig differenzierbar ist.

11.1. Beispiel: Wir betrachten das nichtlineare Ausgleichsproblem

$$\min_{\boldsymbol{x} \in \mathbb{R}^n} f(\boldsymbol{x}), \quad f(\boldsymbol{x}) = \frac{1}{2} \sum_{i=1}^{m} \left(F_i(\boldsymbol{x})\right)^2 .$$

Die Funktionen $F_i : \mathbb{R}^n \longrightarrow \mathbb{R}$, $i = 1, \ldots, m$, seien dabei stetig differenzierbar. Es gilt

$$\frac{\partial f(\boldsymbol{x})}{\partial x_j} = \sum_{i=1}^{m} \frac{\partial F_i}{\partial x_j}(\boldsymbol{x}) \cdot F_i(\boldsymbol{x}), \quad j = 1, \ldots, n.$$

Fasst man die F_i als Komponenten einer Funktion

$$\boldsymbol{F} : \mathbb{R}^n \longrightarrow \mathbb{R}^m$$

auf, so gilt

$$f(\boldsymbol{x}) = \frac{1}{2} \boldsymbol{F}(\boldsymbol{x})^T \boldsymbol{F}(\boldsymbol{x})$$

und

$$\boldsymbol{\nabla} f(\boldsymbol{x}) = \boldsymbol{F}'(\boldsymbol{x})^T \boldsymbol{F}(\boldsymbol{x}).$$

♡

11.2. Beispiel: Wir betrachten das nichtlineare Ausgleichsproblem bezüglich einer anderen Norm:

$$\min_{\boldsymbol{x} \in \mathbb{R}^n} f(\boldsymbol{x}), \quad f(\boldsymbol{x}) = \|\boldsymbol{F}(\boldsymbol{x})\|_\infty = \max_{i=1,\ldots,m} |F_i(\boldsymbol{x})|$$

oder

$$\min_{\boldsymbol{x} \in \mathbb{R}^n} f(\boldsymbol{x}), \quad f(\boldsymbol{x}) = \|\boldsymbol{F}(\boldsymbol{x})\|_1 = \sum_{i=1}^{m} |F_i(\boldsymbol{x})| .$$

Hier sollte man auch für stetig differenzierbares $\boldsymbol{F}$ nicht mehr erwarten, dass die Funktion f differenzierbar ist. ♡

Für die Fälle aus dem letzten Beispiel benötigen wir einen schwächeren Ableitungsbegriff. Eine Funktion

$$\boldsymbol{F} : \mathbb{R}^n \longrightarrow \mathbb{R}^m$$

heißt im Punkte $\boldsymbol{x} \in \mathbb{R}^n$ **in Richtung** $\boldsymbol{p} \in \mathbb{R}^n$ **richtungsdifferenzierbar**, falls der Grenzwert

$$\boldsymbol{F}'(\boldsymbol{x};\boldsymbol{p}) = \lim_{\alpha \to +0} \frac{\boldsymbol{F}(\boldsymbol{x}+\alpha\boldsymbol{p}) - \boldsymbol{F}(\boldsymbol{x})}{\alpha}$$

existiert. Die Funktion $\boldsymbol{F}'(\boldsymbol{x};\boldsymbol{p})$ heißt dann **Richtungsableitung von $\boldsymbol{F}$ im Punkt $\boldsymbol{x}$ in Richtung $\boldsymbol{p}$**.
Die Funktion $\boldsymbol{F}$ heißt im Punkt $\boldsymbol{x} \in \mathbb{R}^n$ **richtungsdifferenzierbar**, falls $\boldsymbol{F}$ in $\boldsymbol{x}$ in jede Richtung $\boldsymbol{p} \in \mathbb{R}^n$ richtungsdifferenzierbar ist. In diesem Falle nennt man die Abbildung $\boldsymbol{F}'(\boldsymbol{x};\circ) : \mathbb{R}^n \longrightarrow \mathbb{R}^m$ GATEAUX-**Ableitung** von $\boldsymbol{F}$ in $\boldsymbol{x}$. Für den Zusammenhang der verschiedenen Ableitungsbegriffe gilt der folgende Satz.

11.3. Satz: *Ist die Funktion $\boldsymbol{F} : \mathbb{R}^n \longrightarrow \mathbb{R}^m$ im Punkt $\boldsymbol{x} \in \mathbb{R}^n$ stetig differenzierbar, so ist $\boldsymbol{F}$ in $\boldsymbol{x}$ richtungsdifferenzierbar. Die* GATEAUX*-Ableitung ist durch*

$$\boldsymbol{F}'(\boldsymbol{x};\boldsymbol{p}) = \boldsymbol{F}'(\boldsymbol{x})\boldsymbol{p}$$

gegeben.

Im Falle konvexer Funktionen ist jede stationäre Lösung auch globale Lösung des Minimumproblems. Konvexe Funktionen haben noch andere nützliche Eigenschaften.

11.4. Satz: *Es sei $f : \mathbb{R}^n \longrightarrow \mathbb{R}$ auf dem $\mathbb{R}^n$ konvex. Dann ist f in jedem Punkt $\boldsymbol{x} \in \mathbb{R}^n$ richtungsdifferenzierbar und es gilt*

$$f'(\boldsymbol{x};\boldsymbol{p}) \leqq f(\boldsymbol{x}+\boldsymbol{p}) - f(\boldsymbol{x}).$$

Die GATEAUX*-Ableitung $f'(\boldsymbol{x};\circ) : \mathbb{R}^n \longrightarrow \mathbb{R}$ von f in $\boldsymbol{x}$ hat folgende Eigenschaften:*

1. *$f'(\boldsymbol{x};\circ)$ ist nichtnegativ homogen:*

$$f'(\boldsymbol{x};\alpha\boldsymbol{p}) = \alpha f'(\boldsymbol{x};\boldsymbol{p}) \quad \forall \boldsymbol{p} \in \mathbb{R}^n \forall \alpha > 0,$$

2. *$f'(\boldsymbol{x};\circ)$ ist subadditiv:*

$$f'(\boldsymbol{x};\boldsymbol{p}+\boldsymbol{q}) \leqq f'(\boldsymbol{x};\boldsymbol{p}) + f'(\boldsymbol{x};\boldsymbol{q}) \quad \forall \boldsymbol{p},\boldsymbol{q} \in \mathbb{R}^n,$$

3. $f'(\boldsymbol{x};\circ)$ *ist konvex:*

$$f'(\boldsymbol{x};\alpha\boldsymbol{p}+(1-\alpha)\boldsymbol{q}) \leqq \alpha f'(\boldsymbol{x};\boldsymbol{p})+(1-\alpha)f'(\boldsymbol{x};\boldsymbol{q})\ \forall \boldsymbol{p},\boldsymbol{q}\in\mathbb{R}^n \forall \alpha\in[0,1].$$

Beweis: Für beliebige aber feste $\boldsymbol{x},\boldsymbol{p}\in\mathbb{R}^n$ definieren wir die Funktion

$$\boldsymbol{\Phi} : (0,1] \longrightarrow \mathbb{R}$$

durch

$$\boldsymbol{\Phi}(\alpha) = \frac{f(\boldsymbol{x}+\alpha\boldsymbol{p})-f(\boldsymbol{x})}{\alpha}.$$

Wir zeigen als erstes die Existenz des Grenzwertes

$$\lim_{\alpha\to+0} \boldsymbol{\Phi}(\alpha).$$

Wegen der Konvexität von f gilt

$$\begin{aligned} f(\boldsymbol{x}) &= f\left(\frac{1}{1+\alpha}(\boldsymbol{x}+\alpha\boldsymbol{p})+\frac{\alpha}{1+\alpha}(\boldsymbol{x}-\boldsymbol{p})\right) \\ &\leqq \frac{1}{1+\alpha}f(\boldsymbol{x}+\alpha\boldsymbol{p})+\frac{\alpha}{1+\alpha}f(\boldsymbol{x}-\boldsymbol{p}) \end{aligned}$$

für alle $\alpha\in(0,1]$. Daraus folgt weiter

$$f(\boldsymbol{x})-\frac{\alpha}{1+\alpha}f(\boldsymbol{x}-\boldsymbol{p}) \leqq \frac{1}{1+\alpha}f(\boldsymbol{x}+\alpha\boldsymbol{p})$$

und

$$\begin{aligned} (1+\alpha)f(\boldsymbol{x})-\alpha f(\boldsymbol{x}-\boldsymbol{p}) &\leqq f(\boldsymbol{x}+\alpha\boldsymbol{p}), \\ f(\boldsymbol{x})-f(\boldsymbol{x}-\boldsymbol{p}) &\leqq \frac{f(\boldsymbol{x}+\alpha\boldsymbol{p})-f(\boldsymbol{x})}{\alpha} = \boldsymbol{\Phi}(\alpha). \end{aligned}$$

Damit ist $\boldsymbol{\Phi}(\alpha)$ nach unten beschränkt. Für ein β mit $0<\beta\leqq\alpha\leqq 1$ gilt

$$\begin{aligned} \boldsymbol{\Phi}(\beta) &= \frac{f(\boldsymbol{x}+\beta\boldsymbol{p})-f(\boldsymbol{x})}{\beta} \\ &= \frac{f\left(\frac{\beta}{\alpha}(\boldsymbol{x}+\alpha\boldsymbol{p})+\frac{\alpha-\beta}{\alpha}\boldsymbol{x}\right)-f(\boldsymbol{x})}{\beta} \\ &\leqq \frac{\frac{\beta}{\alpha}f(\boldsymbol{x}+\alpha\boldsymbol{p})+\frac{\alpha-\beta}{\alpha}f(\boldsymbol{x})-f(\boldsymbol{x})}{\beta} \\ &= \frac{\beta f(\boldsymbol{x}+\alpha\boldsymbol{p})+(\alpha-\beta)f(\boldsymbol{x})-\alpha f(\boldsymbol{x})}{\alpha\beta} \\ &= \boldsymbol{\Phi}(\alpha). \end{aligned}$$

Die Funktion Φ ist somit im Intervall $(0,1]$ monoton fallend. Mit der oben bewiesenen Beschränktheit folgt die Existenz des Grenzwertes.
Weiter gilt:

$$\begin{aligned} f'(\boldsymbol{x};\alpha\boldsymbol{p}) &= \lim_{\beta\to+0}\frac{f(\boldsymbol{x}+\beta\alpha\boldsymbol{p})-f(\boldsymbol{x})}{\beta} \\ &= \lim_{\beta\to+0}\alpha\frac{f(\boldsymbol{x}+\beta\alpha\boldsymbol{p})-f(\boldsymbol{x})}{\alpha\beta} \\ &= \alpha\lim_{\gamma\to+0}\frac{f(\boldsymbol{x}+\gamma\boldsymbol{p})-f(\boldsymbol{x})}{\gamma} \\ &= \alpha f'(\boldsymbol{x};\boldsymbol{p}), \end{aligned}$$

$$\begin{aligned} f'(\boldsymbol{x};\boldsymbol{p}+\boldsymbol{q}) &= \lim_{\alpha\to+0}\frac{f(\boldsymbol{x}+\alpha(\boldsymbol{p}+\boldsymbol{q}))-f(\boldsymbol{x})}{\alpha} \\ &= \lim_{\alpha\to+0}\frac{f\left(\frac{1}{2}(\boldsymbol{x}+2\alpha\boldsymbol{p})+\frac{1}{2}(\boldsymbol{x}+2\alpha\boldsymbol{q})\right)-f(\boldsymbol{x})}{\alpha} \\ &\leqq \lim_{\alpha\to+0}\frac{\frac{1}{2}f(\boldsymbol{x}+2\alpha\boldsymbol{p})+\frac{1}{2}f(\boldsymbol{x}+2\alpha\boldsymbol{q})-f(\boldsymbol{x})}{\alpha} \\ &= \lim_{\alpha\to+0}\frac{f(\boldsymbol{x}+2\alpha\boldsymbol{p})-f(\boldsymbol{x})+f(\boldsymbol{x}+2\alpha\boldsymbol{q})-f(\boldsymbol{x})}{2\alpha} \\ &= f'(\boldsymbol{x};\boldsymbol{p})+f'(\boldsymbol{x};\boldsymbol{q}). \end{aligned}$$

Die Konvexität der GATEAUX-Ableitung folgt sofort aus der nichtnegativen Homogenität und der Subadditivität. ❀

Typische Beispiele konvexer Funktionen sind Vektornormen. Sie besitzen nach Satz 11.4 stets eine GATEAUX-Ableitung. Im Einzelfall ist es oft schwierig, diese zu berechnen. Es gilt

- für $f(\boldsymbol{x}) = \|\boldsymbol{x}\|_\infty = \max\{|x_i| : i = 1,\ldots,n\}$

$$f'(\boldsymbol{x};\boldsymbol{p}) = \begin{cases} \|\boldsymbol{p}\|_\infty & \text{für} \quad \boldsymbol{x} = \boldsymbol{o} \\ \max\{\ \text{sign}(x_j)p_j \mid j \in J(\boldsymbol{x})\ \} & \text{für} \quad \boldsymbol{x} \neq \boldsymbol{o} \end{cases}$$

mit $J(\boldsymbol{x}) = \{\, j \in \{1,\ldots,n\} \mid |x_j| = \|\boldsymbol{x}\|_\infty \,\}$,

- für $f(\boldsymbol{x}) = \|\boldsymbol{x}\|_1 = \sum_{i=1}^n |x_i|$

$$f'(\boldsymbol{x};\boldsymbol{p}) = \sum_{j\in J(\boldsymbol{x})} |p_j| + \sum_{j\notin J(\boldsymbol{x})} \text{sign}(x_j)p_j$$

mit $J(\boldsymbol{x}) = \{\, j \in \{1,\ldots,n\} \mid x_j = 0 \,\}$,

- für $f(\boldsymbol{x}) = \|\boldsymbol{x}\|_2^2 = \sum_{i=1}^n x_i^2$

$$f'(\boldsymbol{x};\boldsymbol{p}) = 2\sum_{i=1}^n x_i p_i.$$

Im letzten Falle ist f sogar stetig differenzierbar und es gilt $f'(\boldsymbol{x};\boldsymbol{p}) = f'(\boldsymbol{x})^T\boldsymbol{p}$. Typische freie Minimumprobleme sind Probleme der Form

$$\min_{\boldsymbol{x}\in\mathbb{R}^n} f(\boldsymbol{x}), \quad f(\boldsymbol{x}) = \|\boldsymbol{F}(\boldsymbol{x})\|,$$

wobei $\boldsymbol{F} : \mathbb{R}^n \longrightarrow \mathbb{R}^m$ eine im allgemeinen nichtlineare Funktion und $\|\circ\|$ eine Vektornorm auf dem $\mathbb{R}^m$ ist. Aufgaben diesen Typs heißen diskrete, nichtlineare Approximationsaufgaben. Man erhält im einzelnen:

- $\|\circ\| = \|\circ\|_2$: nichtlineares Ausgleichsproblem,
- $\|\circ\| = \|\circ\|_\infty$: diskretes, nichtlineares TSCHEBYSCHEFF-Problem,
- $\|\circ\| = \|\circ\|_1$: diskrete L_1-Approximationsaufgabe.

Sind die Komponenten F_i der Funktion $\boldsymbol{F}$ hinreichend glatt, so darf man annehmen, dass f noch richtungsdifferenzierbar ist. So gilt zum Beispiel für die diskrete TSCHEBYSCHEFF-Approximation:

11.5. Satz: *Mit der Abbildung $\boldsymbol{F} : \mathbb{R}^n \longrightarrow \mathbb{R}^m$ sei $f : \mathbb{R}^n \longrightarrow \mathbb{R}$ durch*

$$f(\boldsymbol{x}) = \|\boldsymbol{F}(\boldsymbol{x})\|_\infty = \max_{i=1,\ldots,m} |F_i(\boldsymbol{x})|$$

definiert. Dann gilt:
Ist $\boldsymbol{F}$ in $\boldsymbol{x}^* \in \mathbb{R}^n$ stetig differenzierbar, so ist die Funktion f in $\boldsymbol{x}^*$ richtungsdifferenzierbar mit der GATEAUX*-Ableitung*

$$f'(\boldsymbol{x}^*;\boldsymbol{p}) = \begin{cases} \max\{|F_i'(\boldsymbol{x})\boldsymbol{p}|, i = 1,\ldots,m\} & \textit{für} \quad \boldsymbol{F}(\boldsymbol{x}^*) = \boldsymbol{o}, \\ \max\{\text{sign}(F_i(\boldsymbol{x}^*))F_i'(\boldsymbol{x})\boldsymbol{p}, i \in I(\boldsymbol{x}^*)\} & \textit{für} \quad \boldsymbol{F}(\boldsymbol{x}^*) \neq \boldsymbol{o} \end{cases}$$

mit $I(\boldsymbol{x}^*) = \{\, i \in \{1,\ldots,m\} \mid |F_i(\boldsymbol{x}^*)| = \|\boldsymbol{F}(\boldsymbol{x}^*)\|_\infty \,\}$.

11.1.3. Optimalitätskriterien

Nun sind wir in der Lage, für das Minimumproblem eine erste notwendige Bedingung zu formulieren.

11.6. Satz: *Es sei $\boldsymbol{x}^* \in \mathbb{R}^n$ eine lokale Lösung des Minimumproblems*

$$\min\{\, f(\boldsymbol{x}) \mid \boldsymbol{x} \in \mathbb{R}^n \,\}.$$

Ist die Zielfunktion f in $\boldsymbol{x}^*$ richtungsdifferenzierbar, so gilt

$$f'(\boldsymbol{x}^*; \boldsymbol{p}) \geqq 0$$

für jede Richtung $\boldsymbol{p} \in \mathbb{R}^n$. Ist f in $\boldsymbol{x}^*$ stetig differenzierbar, so gilt $f'(\boldsymbol{x}^*) = 0$.

Beweis: Es existiert eine Umgebung $U(\boldsymbol{x}^*)$ mit $f(\boldsymbol{x}^*) \leqq f(\boldsymbol{x})$ für alle $\boldsymbol{x} \in U(\boldsymbol{x}^*)$. Für ein beliebiges $\boldsymbol{p} \in \mathbb{R}^n$ existiert ein $\alpha_0 > 0$, so dass $\boldsymbol{x}^* + \alpha\boldsymbol{p} \in U(\boldsymbol{x}^*)$ und damit $f(\boldsymbol{x}^*) \leqq f(\boldsymbol{x}^* + \alpha\boldsymbol{p})$ für alle $\alpha \in (0, \alpha_0)$ gilt. Daraus folgt

$$f'(\boldsymbol{x}^*; \boldsymbol{p}) = \lim_{\alpha \to +0} \frac{f(\boldsymbol{x}^* + \alpha\boldsymbol{p}) - f(\boldsymbol{x}^*)}{\alpha} \geqq 0.$$

Ist f in $\boldsymbol{x}^*$ stetig differenzierbar, so gilt $f'(\boldsymbol{x}^*; \boldsymbol{p}) = f'(\boldsymbol{x}^*)^T \boldsymbol{p} \geqq 0$ für alle $\boldsymbol{p} \in \mathbb{R}^n$. Wir wählen speziell $\boldsymbol{p} = -f'(\boldsymbol{x}^*)$ und erhalten

$$-f'(\boldsymbol{x}^*)^T f'(\boldsymbol{x}^*) = -\|f'(\boldsymbol{x}^*)\|^2 \geqq 0.$$

Andererseits gilt aber $\|f'(\boldsymbol{x}^*)\|^2 \geqq 0$ und daher $\|f'(\boldsymbol{x}^*)\|^2 = 0$. ❀

Punkte, die diese notwendige Bedingung erfüllen, haben eine besondere Bedeutung bei der Konstruktion von Minimierungsverfahren. Ist die Funktion $f : \mathbb{R}^n \longrightarrow \mathbb{R}$ im Punkt $\boldsymbol{x}^* \in \mathbb{R}^n$ richtungsdifferenzierbar, so heißt ein Punkt $\boldsymbol{x}^*$ **stationär** von f oder eine **stationäre Lösung** des Minimumproblems $\min\{\, f(\boldsymbol{x}) \mid \boldsymbol{x} \in \mathbb{R}^n \,\}$, falls $f'(\boldsymbol{x}^*; \boldsymbol{p}) \geqq 0$ für alle Richtungen $\boldsymbol{p} \in \mathbb{R}^n$ gilt. Mit den Sätzen 11.4 und 11.5 erhält man für das diskrete TSCHEBYSCHEFF-Problem ein Optimalitätskriterium erster Ordnung[1].

11.7. Satz: *Gegeben sei die diskrete* TSCHEBYSCHEFF*sche Approximationsaufgabe*

$$\min_{\boldsymbol{x} \in \mathbb{R}^n} f(\boldsymbol{x}), \quad f(\boldsymbol{x}) = \|\boldsymbol{F}(\boldsymbol{x})\|_\infty = \max_{i=1,\ldots,m} |F_i(\boldsymbol{x})|.$$

Die Funktionen $F_i : \mathbb{R}^n \longrightarrow \mathbb{R}$, $i = 1, \ldots, m$, seien in $\boldsymbol{x}^*$ stetig differenzierbar und es sei $\boldsymbol{F}(\boldsymbol{x}^*) \neq \boldsymbol{o}$. Dann gilt:
Ist $\boldsymbol{x}^*$ eine lokale Lösung des Minimumproblems, so ist $\boldsymbol{x}^*$ auch eine stationäre Lösung und daher

[1] Optimalitätskriterium erster Ordnung bedeutet, dass nur Ableitungen erster Ordnung verwendet werden

1. $f'(\boldsymbol{x}^*;\boldsymbol{p}) = \max\{\text{sign}(F_i(\boldsymbol{x}^*))F_i'(\boldsymbol{x}^*)\boldsymbol{p}, i \in I(\boldsymbol{x}^*)\} \geqq 0 \; \forall \boldsymbol{p} \in \mathbb{R}^n$ *mit* $I(\boldsymbol{x}^*) = \{\, i \in \{1,\ldots,m \mid |F_i(\boldsymbol{x}^*)| = \|\boldsymbol{F}(\boldsymbol{x}^*)\|_\infty \,\}$.
2. *Die erste Bedingung ist äquivalent zur Existenz von nichtnegativen reellen Zahlen* λ_i^*, $i \in I(\boldsymbol{x}^*)$, *mit*

$$\sum_{i \in I(\boldsymbol{x}^*)} \lambda_i^* = 1, \qquad \sum_{i \in I(\boldsymbol{x}^*)} \lambda_i^* \text{sign}(F_i(\boldsymbol{x}^*))F_i'(\boldsymbol{x}^*) = 0.$$

Im stetig differenzierbaren Falle werden die notwendigen Optimalitätskriterien natürlich einfacher. Für das nichtlineare Ausgleichsproblem

$$\min_{\boldsymbol{x} \in \mathbb{R}^n} f(\boldsymbol{x}), \quad f(\boldsymbol{x}) = \|\boldsymbol{F}(\boldsymbol{x})\|_2^2 = \sum_{i=1}^{m} \left(F_i(\boldsymbol{x})\right)^2$$

mit einer stetig differenzierbaren Funktion $\boldsymbol{F} : \mathbb{R}^n \longrightarrow \mathbb{R}^m$ ergibt sich

$$f'(\boldsymbol{x}^*) = 2\boldsymbol{F}'(\boldsymbol{x}^*)\boldsymbol{F}(\boldsymbol{x}^*) = 0$$

als notwendige Bedingung für ein lokales Minimum.
Neben den notwendigen Optimalitätskriterien erster Ordnung interessieren auch notwendige und hinreichende Optimalitätskriterien zweiter Ordnung. Dazu definieren wir:
Ist die Funktion $f : \mathbb{R}^n \longrightarrow \mathbb{R}$ in einem Punkt $\boldsymbol{x} \in \mathbb{R}^n$ zweimal stetig partiell differenzierbar (alle partiellen Ableitungen zweiter Ordnung existieren in einer Umgebung von $\boldsymbol{x}$ und diese sind in $\boldsymbol{x}$ stetig), so heißt f in $\boldsymbol{x}$ zweimal stetig differenzierbar. Die symmetrische (n,n)-Matrix

$$f''(\boldsymbol{x}) = \left(\frac{\partial^2}{\partial x_i \partial x_j} f(\boldsymbol{x})\right)_{1 \leqq i,j \leqq n}$$

heißt HESSE-**Matrix** von f im Punkt $\boldsymbol{x}$.
Eine Funktion f heißt zweimal stetig differenzierbar auf einer offenen Menge $D \in \mathbb{R}^n$, falls f für jedes $\boldsymbol{x} \in D$ zweimal stetig differenzierbar ist. Für zweimal stetig differenzierbare Funktionen kennen wir aus der Analysis notwendige und hinreichende Optimalitätskriterien.

11.8. Satz: *Die Funktion f sei auf einer offenen Umgebung des Punktes $\boldsymbol{x}^* \in \mathbb{R}^n$ zweimal stetig differenzierbar. Dann gilt:*

1. *Ist $\boldsymbol{x}^*$ eine lokale Lösung des Minimumproblems*

$$\min_{\boldsymbol{x} \in \mathbb{R}^n} f(\boldsymbol{x}),$$

so ist $f'(\boldsymbol{x}^*) = \boldsymbol{o}$ und $f''(\boldsymbol{x}^*)$ ist positiv semidefinit.

2. *Ist $f'(\boldsymbol{x}^*) = \boldsymbol{o}$ und $f''(\boldsymbol{x}^*)$ positiv definit, so ist $\boldsymbol{x}^*$ eine isolierte lokale Lösung des Minimumproblems: Es gibt eine Umgebung $U(\boldsymbol{x}^*)$ mit*

$$f(\boldsymbol{x}^*) < f(\boldsymbol{x}) \quad \forall \boldsymbol{x} \in U(\boldsymbol{x}^*).$$

Im nichtglatten Falle ist die Situation wieder komplizierter. Es lassen sich aber für spezielle Probleme auch hinreichende Bedingungen formulieren.

11.9. Satz: *Gegeben sei die diskrete* TSCHEBYSCHEFF*sche Approximationsaufgabe*

$$\min_{\boldsymbol{x}\in\mathbb{R}^n} f(\boldsymbol{x}), \quad f(\boldsymbol{x}) = \|\boldsymbol{F}(\boldsymbol{x})\|_\infty = \max_{i=1,\ldots,m} |F_i(\boldsymbol{x})|.$$

Die Funktionen

$$F_i : \mathbb{R}^n \longrightarrow \mathbb{R}, \quad i = 1, \ldots, m,$$

seien auf einer offenen Umgebung von $\boldsymbol{x}^* \in \mathbb{R}^n$ zweimal stetig differenzierbar und es sei $\boldsymbol{F}(\boldsymbol{x}^*) \neq \boldsymbol{o}$. Wir definieren die Menge

$$I(\boldsymbol{x}^*) = \{\, i \in \{1, \ldots, m\} \mid |F_i(\boldsymbol{x}^*)| = \|\boldsymbol{F}(\boldsymbol{x}^*)\|_\infty \,\}$$

und nehmen an, dass nichtnegative reelle Zahlen λ_i^*, $i \in I(\boldsymbol{x}^*)$, existieren mit

$$\sum_{i\in I(\boldsymbol{x}^*)} \lambda_i^* = 1, \qquad \sum_{i\in I(\boldsymbol{x}^*)} \lambda_i^* \operatorname{sign}(F_i(\boldsymbol{x}^*)) F_i'(\boldsymbol{x}^*) = \boldsymbol{o}.$$

Weiterhin sei

$$T^* = \left\{\, \boldsymbol{p} \in \mathbb{R}^n \mid F_i'(\boldsymbol{x}^*)\boldsymbol{p} = 0 \,\right\} \quad \forall i \in I(\boldsymbol{x}^*) \wedge \lambda_i^* > 0.$$

Gilt nun

$$\boldsymbol{p}^T \left\{ \sum_{i\in I(\boldsymbol{x}^*)} \lambda_i^* \operatorname{sign}(F_i(\boldsymbol{x}^*)) F_i''(\boldsymbol{x}^*) \right\} \boldsymbol{p} > 0 \quad \forall \boldsymbol{p} \in T^* \setminus \{\boldsymbol{o}\},$$

so ist $\boldsymbol{x}^*$ eine isolierte Lösung des Minimumproblems.

Beweis: Wir nehmen an, die Behauptung sei falsch. Es sei $\boldsymbol{x}^*$ keine isolierte lokale Lösung. Dann gibt es eine gegen $\boldsymbol{x}^*$ konvergierende Folge $\{\boldsymbol{x}_k\}_{k\in\mathbb{N}}$ mit $\boldsymbol{x}_k \neq \boldsymbol{x}^*$ für alle k, für die $f(\boldsymbol{x}_k) \leqq f(\boldsymbol{x}^*)$ gilt. Wir stellen die $\boldsymbol{x}_k$ in der Form $\boldsymbol{x}_k = \boldsymbol{x}^* + \alpha_k \boldsymbol{p}_k$ mit $\alpha_k > 0$ und $\|\boldsymbol{p}_k\| = 1$ dar. Wegen

$$\lim_{k\to\infty} \boldsymbol{x}_k = \boldsymbol{x}^* \quad \text{gilt} \quad \lim_{k\to\infty} \alpha_k = 0.$$

Aus $\{\boldsymbol{p}_k\}_{k\in\mathbb{N}}$ lässt sich eine konvergente Teilfolge auswählen. O.B.d.A. sei darum

$$\boldsymbol{x}_k = \boldsymbol{x}^* + \alpha_k \boldsymbol{p}_k, \quad \forall k : \alpha_k \geqq 0, \quad \lim_{k\to\infty} \boldsymbol{p}_k = \boldsymbol{p} \neq \boldsymbol{o}.$$

Es sei weiterhin $\boldsymbol{r}_k = \alpha_k(\boldsymbol{p}_k - \boldsymbol{p})$. Dann gilt

$$\lim_{k\to\infty} \frac{\boldsymbol{r}_k}{\alpha_k} = \boldsymbol{o}.$$

Damit folgt

$$\begin{aligned}
&\lim_{k\to\infty} \frac{f(\boldsymbol{x}^* + \alpha_k \boldsymbol{p}_k) - f(\boldsymbol{x}^*)}{\alpha_k} = \\
&= \lim_{k\to\infty} \frac{f(\boldsymbol{x}^* + \alpha_k \boldsymbol{p} + \boldsymbol{r}_k) - f(\boldsymbol{x}^*)}{\alpha_k} \\
&= \lim_{k\to\infty} \frac{f(\boldsymbol{x}^* + \alpha_k \boldsymbol{p}) - f(\boldsymbol{x}^*) + f(\boldsymbol{x}^* + \alpha_k \boldsymbol{p} + \boldsymbol{r}_k) - f(\boldsymbol{x}^* + \alpha_k \boldsymbol{p})}{\alpha_k} \\
&= \lim_{k\to\infty} \frac{f(\boldsymbol{x}^* + \alpha_k \boldsymbol{p}) - f(\boldsymbol{x}^*)}{\alpha_k} + \lim_{k\to\infty} \frac{f(\boldsymbol{x}^* + \alpha_k \boldsymbol{p} + \boldsymbol{r}_k) - f(\boldsymbol{x}^* + \alpha_k \boldsymbol{p})}{\alpha_k} \\
&= f'(\boldsymbol{x}^*, \boldsymbol{p}) + \lim_{k\to\infty} \frac{\|\boldsymbol{F}(\boldsymbol{x}^* + \alpha_k \boldsymbol{p} + \boldsymbol{r}_k)\|_\infty - \|\boldsymbol{F}(\boldsymbol{x}^* + \alpha_k \boldsymbol{p})\|_\infty}{\alpha_k}.
\end{aligned}$$

Nun gilt

$$\begin{aligned}
0 &\leqq \lim_{k\to\infty} \frac{|\|\boldsymbol{F}(\boldsymbol{x}^* + \alpha_k \boldsymbol{p} + \boldsymbol{r}_k)\|_\infty - \|\boldsymbol{F}(\boldsymbol{x}^* + \alpha_k \boldsymbol{p})\|_\infty|}{\alpha_k} \\
&\leqq \lim_{k\to\infty} \frac{\|\boldsymbol{F}(\boldsymbol{x}^* + \alpha_k \boldsymbol{p} + \boldsymbol{r}_k) - \boldsymbol{F}(\boldsymbol{x}^* + \alpha_k \boldsymbol{p})\|_\infty}{\alpha_k} \\
&\leqq \lim_{k\to\infty} \frac{\|\boldsymbol{F}(\boldsymbol{x}^* + \alpha_k \boldsymbol{p}) + \boldsymbol{F}'(\boldsymbol{x}^* + \alpha_k \boldsymbol{p} + \vartheta \boldsymbol{r}_k)\boldsymbol{r}_k - \boldsymbol{F}(\boldsymbol{x}^* + \alpha_k \boldsymbol{p})\|_\infty}{\alpha_k} \\
&\leqq \lim_{k\to\infty} \frac{\|\boldsymbol{F}'(\boldsymbol{x}^* + \alpha_k \boldsymbol{p} + \vartheta_k \boldsymbol{r}_k)\|_\infty \|\boldsymbol{r}_k\|_\infty}{\alpha_k} \quad (\vartheta_k \in (0,1)) \\
&= \lim_{k\to\infty} \|\boldsymbol{F}'(\boldsymbol{x}^* + \alpha_k \boldsymbol{p} + \vartheta_k \boldsymbol{r}_k)\|_\infty \lim_{k\to\infty} \frac{\|\boldsymbol{r}_k\|_\infty}{\alpha_k} \\
&= 0.
\end{aligned}$$

Insgesamt erhalten wir

$$\lim_{k\to\infty} \frac{f(\boldsymbol{x}^* + \alpha_k \boldsymbol{p}_k) - f(\boldsymbol{x}^*)}{\alpha_k} = f'(\boldsymbol{x}^*, \boldsymbol{p}).$$

Wegen $f(\boldsymbol{x}^* + \alpha_k \boldsymbol{p}_k) \leqq f(\boldsymbol{x}^*)$ und $\alpha_k > 0$ folgt $f'(\boldsymbol{x}^*, \boldsymbol{p}) \leqq \boldsymbol{o}$. Mit Satz 11.5 ergibt sich

$$f'(\boldsymbol{x}^*, \boldsymbol{p}) = \max_{i \in I(\boldsymbol{x}^*)} \text{sign}(F_i(\boldsymbol{x}^*)) F_i'(\boldsymbol{x}^*) \boldsymbol{p} \leqq 0.$$

Nach der ersten Voraussetzung gilt

$$\sum_{i \in I(\boldsymbol{x}^*)} \lambda_i^* \text{sign}(F_i(\boldsymbol{x}^*)) F_i'(\boldsymbol{x}^*) \boldsymbol{p} = 0.$$

Wegen $\lambda_i^* \geqq 0$ und $\text{sign}(F_i(\boldsymbol{x}^*)) F_i'(\boldsymbol{x}^*) \boldsymbol{p} \leqq 0$ für $i \in I(\boldsymbol{x}^*)$ folgt daraus $F_i'(\boldsymbol{x}^*) \boldsymbol{p} = \boldsymbol{o}$ für $i \in I(\boldsymbol{x}^*)$. Es gilt daher $\boldsymbol{p} \in T^*$. Für $i \in I(\boldsymbol{x}^*)$ ist weiterhin

$$|F_i(\boldsymbol{x}_k)| \leqq \|\boldsymbol{F}(\boldsymbol{x}_k)\|_\infty \leqq \|\boldsymbol{F}(\boldsymbol{x}^*)\|_\infty = |F_i(\boldsymbol{x}^*)|,$$

und für hinreichend großes k ist $\text{sign}(F_i(\boldsymbol{x}_k)) = \text{sign}(F_i(\boldsymbol{x}^*))$. Damit folgt für alle $i \in I(\boldsymbol{x}^*)$ und für hinreichend großes k

$$\begin{aligned}
0 &\geqq |F_i(\boldsymbol{x}_k)| - |F_i(\boldsymbol{x}^*)| \\
&= \text{sign}(F_i(\boldsymbol{x}^*)) \left[F_i(\boldsymbol{x}_k) - F_i(\boldsymbol{x}^*)\right] \\
&= \text{sign}(F_i(\boldsymbol{x}^*)) \left[F_i(\boldsymbol{x}^* + \alpha_k \boldsymbol{p}_k) - F_i(\boldsymbol{x}^*)\right] \\
&\geqq \alpha_k \text{sign}(F_i(\boldsymbol{x}^*)) F_i'(\boldsymbol{x}^*) \boldsymbol{p}_k + \frac{1}{2} \alpha_k^2 \text{sign}(F_i(\boldsymbol{x}^*)) \boldsymbol{p}_k^T F_i''(\boldsymbol{x}^* + \vartheta_{ik} \alpha_k \boldsymbol{p}_k) \boldsymbol{p}_k
\end{aligned}$$

mit $\vartheta_{ik} \in (0,1)$. Multiplizieren wir die letzte Ungleichung mit $\lambda_i^* \neq 0$ und summieren über alle $i \in I(\boldsymbol{x}^*)$, so erhalten wir

$$\begin{aligned}
0 &\geqq \alpha_k \sum_{i \in I(\boldsymbol{x}^*)} \lambda_i^* \text{sign}(F_i(\boldsymbol{x}^*)) F_i'(\boldsymbol{x}^*) \boldsymbol{p}_k \\
&\quad + \frac{\alpha_k^2}{2} \sum_{i \in I(\boldsymbol{x}^*)} \lambda_i^* \text{sign}(F_i(\boldsymbol{x}^*)) \boldsymbol{p}_k^T F_i''(\boldsymbol{x}^* + \vartheta_{ik} \alpha_k \boldsymbol{p}_k) \boldsymbol{p}_k \\
&= \frac{\alpha_k^2}{2} \boldsymbol{p}_k^T \left\{ \sum_{i \in I(\boldsymbol{x}^*)} \lambda_i^* \text{sign}(F_i(\boldsymbol{x}^*)) F_i''(\boldsymbol{x}^* + \vartheta_{ik} \alpha_k \boldsymbol{p}_k) \right\} \boldsymbol{p}_k.
\end{aligned}$$

Der Grenzübergang $k \longrightarrow \infty$ ($\boldsymbol{p}_k \longrightarrow \boldsymbol{p}$ und $\vartheta_{ik} \longrightarrow 0$) liefert

$$0 \geqq \boldsymbol{p} \left\{ \sum_{i \in I(\boldsymbol{x}^*)} \lambda_i^* \text{sign}(F_i(\boldsymbol{x}^*)) F_i''(\boldsymbol{x}^*) \right\} \boldsymbol{p}$$

als Widerspruch zur Voraussetzung

$$0 < \boldsymbol{p} \left\{ \sum_{i \in I(\boldsymbol{x}^*)} \lambda_i^* \text{sign}(F_i(\boldsymbol{x}^*)) F_i''(\boldsymbol{x}^*) \right\} \boldsymbol{p}.$$

Damit muss die Annahme falsch sein; $\boldsymbol{x}^*$ ist isolierte lokale Lösung. ❀

Für konvexe Funktionen lassen sich allgemeine globale Konvergenzaussagen machen. Für hinreichend glatte Funktionen wollen wir die Konvexität durch geeignete Bedingungen charakterisieren.
Eine Funktion $f : \mathbb{R}^n \longrightarrow \mathbb{R}$ heißt **gleichmäßig konvex** auf der konvexen Menge $D \subseteq \mathbb{R}^n$, falls eine Konstante $\gamma > 0$ existiert, so dass

$$(1-\alpha) f(\boldsymbol{x}) + \alpha f(\boldsymbol{y}) - f((1-\alpha)\boldsymbol{x} + \alpha \boldsymbol{y}) \geqq \frac{\gamma}{2} \alpha (1-\alpha) \|\boldsymbol{x} - \boldsymbol{y}\|_2^2$$

für alle $\boldsymbol{x}, \boldsymbol{y} \in D$ und alle $\alpha \in [0,1]$ gilt.
Bemerkung: Eine Funktion f ist konvex auf D, falls die obige Bedingung mit $\gamma = 0$ gilt.

11.10. Beispiel: Wir betrachten für die (n,n)-symmetrische Matrix $\boldsymbol{Q}$ die quadratische Funktion

$$f(\boldsymbol{x}) = \boldsymbol{c}^T \boldsymbol{x} + \frac{1}{2} \boldsymbol{x}^T \boldsymbol{Q} \boldsymbol{x}$$

für beliebige $\boldsymbol{x} \in \mathbb{R}^n$. Es gilt

$$\begin{aligned}
&(1-\alpha) f(\boldsymbol{x}) + \alpha f(\boldsymbol{y}) - f((1-\alpha)\boldsymbol{x} + \alpha \boldsymbol{y}) = \\
&= (1-\alpha)\boldsymbol{c}^T \boldsymbol{x} + \frac{1-\alpha}{2} \boldsymbol{x}^T \boldsymbol{Q} \boldsymbol{x} + \alpha \boldsymbol{c}^T \boldsymbol{y} + \frac{\alpha}{2} \boldsymbol{y}^T \boldsymbol{Q} \boldsymbol{y} - \\
&\quad - \boldsymbol{c}^T [(1-\alpha)\boldsymbol{x} + \alpha \boldsymbol{y}] - \frac{1}{2} [(1-\alpha)\boldsymbol{x} + \alpha \boldsymbol{y}]^T \boldsymbol{Q} [(1-\alpha)\boldsymbol{x} + \alpha \boldsymbol{y}] \\
&= \frac{1-\alpha}{2} \boldsymbol{x}^T \boldsymbol{Q} \boldsymbol{x} + \frac{\alpha}{2} \boldsymbol{y}^T \boldsymbol{Q} \boldsymbol{y} - \frac{(1-\alpha)^2}{2} \boldsymbol{x}^T \boldsymbol{Q} \boldsymbol{x} - \\
&\quad - \frac{\alpha(1-\alpha)}{2} \boldsymbol{x}^T \boldsymbol{Q} \boldsymbol{y} - \frac{\alpha(1-\alpha)}{2} \boldsymbol{y}^T \boldsymbol{Q} \boldsymbol{x} - \frac{\alpha^2}{2} \boldsymbol{y}^T \boldsymbol{Q} \boldsymbol{y} \\
&= \frac{\alpha(1-\alpha)}{2} \left[\boldsymbol{x}^T \boldsymbol{Q} \boldsymbol{x} - \boldsymbol{x}^T \boldsymbol{Q} \boldsymbol{y} - \boldsymbol{y}^T \boldsymbol{Q} \boldsymbol{x} + \boldsymbol{y}^T \boldsymbol{Q} \boldsymbol{y} \right] \\
&= \frac{\alpha(1-\alpha)}{2} (\boldsymbol{x} - \boldsymbol{y})^T \boldsymbol{Q} (\boldsymbol{x} - \boldsymbol{y}).
\end{aligned}$$

Ist $\boldsymbol{Q}$ positiv semidefinit, so gilt

$$(\boldsymbol{x}-\boldsymbol{y})^T\boldsymbol{Q}(\boldsymbol{x}-\boldsymbol{y}) \geqq 0$$

für beliebige $\boldsymbol{x}$ und $\boldsymbol{y}$. f ist in diesem Falle konvex. Ist $\boldsymbol{Q}$ sogar positiv definit, so gilt

$$(\boldsymbol{x}-\boldsymbol{y})^T\boldsymbol{Q}(\boldsymbol{x}-\boldsymbol{y}) \geqq \lambda_{min}\|\boldsymbol{x}-\boldsymbol{y}\|_2^2$$

mit $\lambda_{min} > 0$ als kleinstem Eigenwert von $\boldsymbol{Q}$. In diesem Falle ist f gleichmäßig konvex. ♡

Mit Hilfe von ersten und zweiten Ableitungen lassen sich Bedingungen für die Konvexität und gleichmäßige Konvexität angeben.

11.11. Satz: *Die Menge $D \subseteq \mathbb{R}^n$ sei konvex und die Funktion $f : \mathbb{R}^n \longrightarrow \mathbb{R}$ sei auf einer offenen Obermenge von D stetig differenzierbar. Dann gilt*

1. *f ist genau dann auf D konvex, wenn*

   $$f'(\boldsymbol{x})(\boldsymbol{y}-\boldsymbol{x}) \leqq f(\boldsymbol{y}) - f(\boldsymbol{x})$$

   für alle $\boldsymbol{x}, \boldsymbol{y} \in D$ gilt.
2. *f ist genau dann auf D gleichmäßig konvex mit einer Konstanten $\gamma > 0$, wenn*

   $$\frac{\gamma}{2}\|\boldsymbol{x}-\boldsymbol{y}\|_2^2 + f'(\boldsymbol{x})(\boldsymbol{y}-\boldsymbol{x}) \leqq f(\boldsymbol{y}) - f(\boldsymbol{x})$$

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

Beweis: Wir nehmen zunächst an, die Funktion f sei konvex bzw. gleichmäßig konvex. Dann gilt

$$(1-\alpha)f(\boldsymbol{x}) + \alpha f(\boldsymbol{y}) - f((1-\alpha)\boldsymbol{x} + \alpha\boldsymbol{y}) \geqq \frac{\gamma}{2}\alpha(1-\alpha)\|\boldsymbol{x}-\boldsymbol{y}\|_2^2$$

für alle $\boldsymbol{x}, \boldsymbol{y} \in D$ und alle $\alpha \in [0,1]$ mit $\gamma = 0$ für konvexes f bzw. $\gamma > 0$ für gleichmäßig konvexes f. Es folgt

$$\begin{aligned}
\alpha\,(f(\boldsymbol{y}) - f(\boldsymbol{x})) &\geqq f(\boldsymbol{x} + \alpha(\boldsymbol{y}-\boldsymbol{x})) - f(\boldsymbol{x}) + \frac{\gamma}{2}\alpha(1-\alpha)\|\boldsymbol{x}-\boldsymbol{y}\|_2^2, \\
f(\boldsymbol{y}) - f(\boldsymbol{x}) &\geqq \frac{f(\boldsymbol{x} + \alpha(\boldsymbol{y}-\boldsymbol{x})) - f(\boldsymbol{x})}{\alpha} + \frac{\gamma}{2}(1-\alpha)\|\boldsymbol{x}-\boldsymbol{y}\|_2^2.
\end{aligned}$$

Für $\alpha \to +0$ folgt daraus

$$f(\boldsymbol{y}) - f(\boldsymbol{x}) \geqq f'(\boldsymbol{x})(\boldsymbol{y}-\boldsymbol{x}) + \frac{\gamma}{2}\|\boldsymbol{x}-\boldsymbol{y}\|_2^2.$$

Nehmen wir nun umgekehrt an, dass eine Konstante $\gamma \geqq 0$ existiert, so dass für alle $\boldsymbol{x}, \boldsymbol{y} \in D$

$$f(\boldsymbol{y}) - f(\boldsymbol{x}) \geqq f'(\boldsymbol{x})(\boldsymbol{y}-\boldsymbol{x}) + \frac{\gamma}{2}\|\boldsymbol{x}-\boldsymbol{y}\|_2^2$$

gilt. Für ein $\alpha \in [0,1]$ ist $\boldsymbol{z} = (1-\alpha)\boldsymbol{x} + \alpha\boldsymbol{y} \in D$. Dann gilt

$$\begin{aligned} f(\boldsymbol{x}) - f(\boldsymbol{z}) &\geqq f'(\boldsymbol{z})(\boldsymbol{x}-\boldsymbol{z}) + \frac{\gamma}{2}\|\boldsymbol{x}-\boldsymbol{z}\|_2^2, \\ f(\boldsymbol{y}) - f(\boldsymbol{z}) &\geqq f'(\boldsymbol{z})(\boldsymbol{y}-\boldsymbol{z}) + \frac{\gamma}{2}\|\boldsymbol{y}-\boldsymbol{z}\|_2^2. \end{aligned}$$

Wir multiplizieren die erste Ungleichung mit $1-\alpha$ und die zweite Ungleichung mit α und addieren beide. Es ergibt sich

$$\begin{aligned} (1-\alpha)f(\boldsymbol{x}) + \alpha f(\boldsymbol{y}) - f(\boldsymbol{z}) \geqq f'(\boldsymbol{z})\left[(1-\alpha)\boldsymbol{x} + \alpha\boldsymbol{y} - \boldsymbol{z}\right] + \\ \frac{\gamma}{2}\left[(1-\alpha)\|\boldsymbol{x}-\boldsymbol{z}\|_2^2 + \alpha\|\boldsymbol{y}-\boldsymbol{z}\|_2\right]. \end{aligned}$$

Mit $\boldsymbol{z} = (1-\alpha)\boldsymbol{x} + \alpha\boldsymbol{y}$ folgt daraus

$$\begin{aligned} &(1-\alpha)f(\boldsymbol{x}) + \alpha f(\boldsymbol{y}) - f((1-\alpha)\boldsymbol{x} + \alpha\boldsymbol{y}) \geqq \\ &\quad\geqq \frac{\gamma}{2}\left\{(1-\alpha)(\boldsymbol{x}^T\boldsymbol{x} - 2\boldsymbol{x}^T\boldsymbol{z} + \boldsymbol{z}^T\boldsymbol{z}) + \alpha(\boldsymbol{y}^T\boldsymbol{y} - 2\boldsymbol{y}^T\boldsymbol{z} + \boldsymbol{z}^T\boldsymbol{z})\right\} \\ &\quad= \frac{\gamma}{2}\left\{(1-\alpha)\boldsymbol{x}^T\boldsymbol{x} - 2\left[(1-\alpha)\boldsymbol{x} + \alpha\boldsymbol{y}\right]^T\boldsymbol{z} + (1-\alpha)\boldsymbol{z}^T\boldsymbol{z} + \alpha\boldsymbol{y}^T\boldsymbol{y} + \alpha\boldsymbol{z}^T\boldsymbol{z}\right\} \\ &\quad= \frac{\gamma}{2}\left\{(1-\alpha)\boldsymbol{x}^T\boldsymbol{x} + \alpha\boldsymbol{y}^T\boldsymbol{y} - \boldsymbol{z}^T\boldsymbol{z}\right\} \\ &\quad= \frac{\gamma}{2}\left\{(1-\alpha)\boldsymbol{x}^T\boldsymbol{x} + \alpha\boldsymbol{y}^T\boldsymbol{y} - (1-\alpha)^2\boldsymbol{x}^T\boldsymbol{x} - 2(1-\alpha)\alpha\boldsymbol{y}^T\boldsymbol{x} - \alpha^2\boldsymbol{y}^T\boldsymbol{y}\right\} \\ &\quad= \frac{\gamma}{2}\alpha(1-\alpha)\|\boldsymbol{x}-\boldsymbol{y}\|_2^2. \end{aligned}$$

Für $\gamma = 0$ ist f damit konvex und für $\gamma > 0$ ist f gleichmäßig konvex. ❀

11.12. Satz: *Es sei $D \subseteqq \mathbb{R}^n$ konvex und die Funktion $f : \mathbb{R}^n \longrightarrow \mathbb{R}$ sei auf einer offenen Obermenge von D zweimal stetig differenzierbar. Dann gilt:*

1. *Ist $f''(\boldsymbol{x})$ positiv definit für alle $\boldsymbol{x} \in D$, so ist f auf D konvex.*

2. *Existiert eine Konstante $\gamma > 0$, so dass für alle $\boldsymbol{p} \in \mathbb{R}^n$ und alle $\boldsymbol{x} \in D$*

$$\gamma \|\boldsymbol{p}\|_2^2 \leqq \boldsymbol{p}^T f''(\boldsymbol{x}) \boldsymbol{p}$$

gilt, so ist f auf D gleichmäßig konvex.

3. *Ist D offen, so gelten in 1 und 2 die Umkehrungen.*

Beweis: Für beliebige, aber feste $\boldsymbol{x}, \boldsymbol{y} \in D$ definieren wir die zweimal stetig differenzierbare Funktion

$$\Phi : [0,1] \longrightarrow \mathbb{R}, \quad \Phi(\alpha) = f\left(\boldsymbol{x} + \alpha(\boldsymbol{y} - \boldsymbol{x})\right).$$

Es gilt

$$\Phi(1) - \Phi(0) - \Phi'(0) = \frac{1}{2}\Phi''(\alpha_0) \text{ mit } \alpha_0 \in (0,1).$$

Damit folgt

$$f(\boldsymbol{y}) - f(\boldsymbol{x}) - f'(\boldsymbol{x})(\boldsymbol{y} - \boldsymbol{x}) = \frac{1}{2}(\boldsymbol{y} - \boldsymbol{x})^T f''\left(\boldsymbol{x} + \alpha_0(\boldsymbol{y} - \boldsymbol{x})\right)(\boldsymbol{y} - \boldsymbol{x}).$$

Wegen $\boldsymbol{x} + \alpha(\boldsymbol{y} - \boldsymbol{x}) \in D$ gilt $(\boldsymbol{y} - \boldsymbol{x})^T f''\left(\boldsymbol{x} + \alpha_0(\boldsymbol{y} - \boldsymbol{x})\right)(\boldsymbol{y} - \boldsymbol{x}) \geqq \gamma \|\boldsymbol{y} - \boldsymbol{x}\|_2^2$. Damit folgt

$$f(\boldsymbol{y}) - f(\boldsymbol{x}) - f'(\boldsymbol{x})(\boldsymbol{y} - \boldsymbol{x}) \geqq \frac{\gamma}{2}\|\boldsymbol{y} - \boldsymbol{x}\|_2^2.$$

Nach Satz 11.11 ist f damit konvex und für positiv definites f'' ($\gamma > 0$) gleichmäßig konvex.
Es sei nun D offen und f auf D konvex ($\gamma = 0$) bzw. gleichmäßig konvex($\gamma > 0$). Für beliebiges $\boldsymbol{x} \in D$ und eine beliebige Richtung $\boldsymbol{p} \in \mathbb{R}^n$ existiert ein $\alpha_0 > 0$, so dass $\boldsymbol{x} + \alpha\boldsymbol{p} \in D$ für alle $\alpha \in [0, \alpha_0]$. Nach Satz 11.11 gilt dann

$$f(\boldsymbol{x} + \alpha\boldsymbol{p}) - f(\boldsymbol{x}) \geqq f'(\boldsymbol{x})(\alpha\boldsymbol{p}) + \frac{\gamma}{2}\alpha^2\|\boldsymbol{p}\|_2^2$$

und

$$f(\boldsymbol{x}) - f(\boldsymbol{x} + \alpha\boldsymbol{p})) \geqq -f'(\boldsymbol{x} + \alpha\boldsymbol{p})(\alpha\boldsymbol{p}) + \frac{\gamma}{2}\alpha^2\|\boldsymbol{p}\|_2^2.$$

Addition der beiden Ungleichungen liefert

$$\left[f'(\boldsymbol{x} + \alpha\boldsymbol{p}) - f'(\boldsymbol{x})\right](\alpha\boldsymbol{p}) \geqq \frac{\gamma}{2}\alpha^2\|\boldsymbol{p}\|_2^2.$$

Daraus erhalten wir

$$\begin{aligned}\boldsymbol{p}^T f''(\boldsymbol{x})\boldsymbol{p} &= \lim_{\alpha\to+0} \frac{[f'(\boldsymbol{x}+\alpha\boldsymbol{p}) - f'(\boldsymbol{x})]\,\boldsymbol{p}}{\alpha} \\ &= \lim_{\alpha\to+0} \frac{[f'(\boldsymbol{x}+\alpha\boldsymbol{p}) - f'(\boldsymbol{x})]\,\alpha\boldsymbol{p}}{\alpha^2} \\ &\geqq \frac{\gamma}{2}\alpha^2\|\boldsymbol{p}\|_2^2.\end{aligned}$$

♣

Für konvexe Funktionen lassen sich so globale Konvergenzaussagen machen. Ist andererseits für eine lokale Lösung $\boldsymbol{x}^*$ die hinreichende Bedingung zweiter Ordnung erfüllt, also f in einer Umgebung $U(\boldsymbol{x}^*)$ zweimal stetig differenzierbar und $f''(\boldsymbol{x}^*)$ positiv definit, so existiert eine Umgebung von $\boldsymbol{x}^*$, in der f'' positiv definit und damit f gleichmäßig konvex ist. Die Konvexitätseigenschaften der Funktion f spielen damit auch für lokale Konvergenzaussagen und die Abschätzung von Konvergenzgeschwindigkeiten eine große Rolle.

11.2. Ein Modellalgorithmus

Die Funktion $f : \mathbb{R}^n \longrightarrow \mathbb{R}$ sei in jedem Punkte $\boldsymbol{x} \in \mathbb{R}^n$ richtungsdifferenzierbar. Zum Lösen des Problems

$$\min_{\boldsymbol{x}\in\mathbb{R}^n} f(\boldsymbol{x})$$

wollen wir einen Algorithmus folgender Form anwenden.

11.13. Modellalgorithmus zum Lösen freier Minimumprobleme:
(Initialisierung) Wähle einen Startpunkt $\boldsymbol{x}^{(0)} \in \mathbb{R}^n$ und setze $k = 0$.

S0 *(Abbruchkriterium)*
Ist $f'(\boldsymbol{x}^{(k)};\boldsymbol{p}) \geqq 0$ für alle Richtungen $\boldsymbol{p} \in \mathbb{R}^n$, dann STOPP; $\boldsymbol{x}^{(k)}$ ist stationäre Lösung.

S1 *(Richtungswahl) Wähle Abstiegsrichtung $\boldsymbol{p}^{(k)}$ mit*

$$f'(\boldsymbol{x}^{(k)};\boldsymbol{p}^{(k)}) < 0.$$

S2 *(Schrittweitenbestimmung) Bestimme Schrittweite α_k, so dass*

$$f(\boldsymbol{x}^{(k)} + \alpha_k\boldsymbol{p}^{(k)}) < f(\boldsymbol{x}^{(k)}).$$

S3 *(Iterationsschritt) Setze*

$$\boldsymbol{x}^{(k+1)} = \boldsymbol{x}^{(k)} + \alpha_k \boldsymbol{p}^{(k)}, \quad k = k+1$$

und gehe zu Schritt **S1**.

Legt man nun Strategien fest, nach denen Abstiegsrichtung und Schrittweite zu bestimmen sind, erhält man konkrete Verfahren.

11.2.1. Schrittweiten bei glatter Zielfunktion

Wir betrachten wieder das freie Minimumproblem

$$\min_{\boldsymbol{x}\in\mathbb{R}^n} f(\boldsymbol{x}).$$

Die Funktion $f : \mathbb{R}^n \longrightarrow \mathbb{R}$ erfülle folgende Voraussetzungen:

V1 Für ein gegebenes $\boldsymbol{x}^{(0)} \in \mathbb{R}^n$ (i.a. der Startpunkt des Verfahrens) ist die Niveaumenge
$L_0 = \left\{ \boldsymbol{x} \in \mathbb{R}^n \mid f(\boldsymbol{x}) \leqq f(\boldsymbol{x}^{(0)}) \right\}$ beschränkt.

V2 Die Zielfunktion f ist auf einer offenen Menge $D \supseteq L_0$ stetig differenzierbar.

V3 Der Gradient $\boldsymbol{\nabla} f(\boldsymbol{x}) = f'(\boldsymbol{x})^T$ ist auf L_0 lipschitzstetig: Es existiert eine Konstante $L > 0$ mit

$$\|\boldsymbol{\nabla} f(\boldsymbol{x}) - \boldsymbol{\nabla} f(\boldsymbol{y})\|_2 \leqq L\|\boldsymbol{x} - \boldsymbol{y}\|_2$$

für alle $\boldsymbol{x}, \boldsymbol{y} \in L_0$.

Bei vorgegebener Abstiegsrichtung $\boldsymbol{p} \in \mathbb{R}^n$ ist eine solche Schrittweite $\alpha > 0$ zu bestimmen, die einen hinreichend starken Abstieg sichert. Die Forderung $f(\boldsymbol{x}^{(k+1)}) < f(\boldsymbol{x}^{(k)})$ ist nicht hinreichend dafür, dass die Folge $\left\{\boldsymbol{x}^{(k)}\right\}_{k\in\mathbb{N}}$ gegen eine lokale Lösung des Minumumproblems konvergiert.

11.14. Beispiel: Es sei $f(x) = x^2$ und $x_k = (-1)^k(\frac{1}{2} + 2^{-k})$. Dann gilt

$$f(x_{k+1}) = x_{k+1}^2 = \frac{1}{4} + 2^{-k-1} + 2^{-2k-2} < \frac{1}{4} + 2^{-k} + 2^{-2k} = x_k^2 = f(x_k).$$

Die Folge $\{x_k\}_{k\in\mathbb{N}}$ ist daher streng monoton fallend. Andererseits besitzt sie nur die beiden Häufungspunkte $-\frac{1}{2}$ und $\frac{1}{2}$, die beide keine lokalen Minima von f sind. ♡

Das folgende Lemma benötigen wir, um für konkrete Verfahren den erzielten Abstieg $f(\boldsymbol{x}) - f(\boldsymbol{x}+\alpha\boldsymbol{p})$ nach unten abzuschätzen.

11.15. Satz: *Die Zielfunktion f genüge den Voraussetzungen* **V1**, **V2** *und* **V3**. *Es seien $\boldsymbol{x} \in L_0$, $\boldsymbol{p} \in \mathbb{R}^n$ eine Abstiegsrichtung für f in $\boldsymbol{x}$ und $\hat{\alpha} = \hat{\alpha}(\boldsymbol{x};\boldsymbol{p})$ die erste positive Nullstelle der Funktion $\psi(\alpha) = f(\boldsymbol{x}) - f(\boldsymbol{x}+\alpha\boldsymbol{p})$. Dann gilt*

$$f(\boldsymbol{x}+\alpha\boldsymbol{p}) \leqq f(\boldsymbol{x}) + \alpha\boldsymbol{\nabla} f(\boldsymbol{x})^T\boldsymbol{p} + \alpha^2\frac{L}{2}\|\boldsymbol{p}\|_2^2$$

für alle $\alpha \in [0,\hat{\alpha}]$.

Beweis: Wegen Voraussetzung **V1** existiert $\hat{\alpha}$ und es ist

$$\boldsymbol{x}+\alpha\boldsymbol{p} \in L_0 \quad \forall \alpha \in [0,\hat{\alpha}].$$

Für ein $\alpha \in [0,\hat{\alpha}]$ erhält man

$$f(\boldsymbol{x}+\alpha\boldsymbol{p}) = f(\boldsymbol{x}) + \alpha\boldsymbol{\nabla} f(\boldsymbol{x})^T\boldsymbol{p} + \int_0^\alpha [\boldsymbol{\nabla} f(\underbrace{\boldsymbol{x}+\lambda\boldsymbol{p}}_{\in L_0}) - \boldsymbol{\nabla} f(\boldsymbol{x})]^T\boldsymbol{p}\, d\lambda.$$

Nach der CAUCHY-SCHWARZschen Ungleichung gilt

$$[\boldsymbol{\nabla} f(\boldsymbol{x}+\lambda\boldsymbol{p}) - \boldsymbol{\nabla} f(\boldsymbol{x})]^T\boldsymbol{p} \leqq \|\boldsymbol{\nabla} f(\boldsymbol{x}+\lambda\boldsymbol{p}) - \boldsymbol{\nabla} f(\boldsymbol{x})\|_2\|\boldsymbol{p}\|_2,$$

und mit Voraussetzung **V3** folgt

$$[\boldsymbol{\nabla} f(\boldsymbol{x}+\lambda\boldsymbol{p}) - \boldsymbol{\nabla} f(\boldsymbol{x})]^T\boldsymbol{p} \leqq \lambda L\|\boldsymbol{p}\|_2^2.$$

Das liefert

$$\begin{aligned} f(\boldsymbol{x}+\alpha\boldsymbol{p}) &= f(\boldsymbol{x}) + \alpha\boldsymbol{\nabla} f(\boldsymbol{x})^T\boldsymbol{p} + L\|\boldsymbol{p}\|_2^2\int_0^\alpha \lambda\, d\lambda \\ &= f(\boldsymbol{x}) + \alpha\boldsymbol{\nabla} f(\boldsymbol{x})^T\boldsymbol{p} + \alpha^2\frac{L}{2}\|\boldsymbol{p}\|_2^2. \end{aligned}$$

❀

Bemerkung: Für $\alpha = \hat{\alpha}$ erhält man wegen $f(\boldsymbol{x}+\hat{\alpha}\boldsymbol{p}) = f(\boldsymbol{x})$

$$-\frac{2\boldsymbol{\nabla} f(\boldsymbol{x})^T\boldsymbol{p}}{L\|\boldsymbol{p}\|_2^2} \leqq \hat{\alpha}.$$

Es liegt nun nahe, die Schrittweite α^* als globale oder erste lokale Lösung des eindimensionalen Minumumproblems

$$\min_{\alpha \in \mathbb{R}_+} f(\boldsymbol{x} + \alpha \boldsymbol{p})$$

zu bestimmen. Im ersten Falle gilt

$$f(\boldsymbol{x} + \alpha^* \boldsymbol{p}) = \min_{\alpha \in \mathbb{R}_+} f(\boldsymbol{x} + \alpha \boldsymbol{p})$$

und im zweiten Fall

$$\boldsymbol{\nabla} f(\boldsymbol{x} + \alpha^* \boldsymbol{p})^T \boldsymbol{p} = 0, \quad \forall \alpha \in [0, \alpha^*) : \boldsymbol{\nabla} f(\boldsymbol{x} + \alpha^* \boldsymbol{p})^T \boldsymbol{p} < 0.$$

In diesen Fällen spricht man von einer **exakten Schrittweite**. Im nächsten Satz wollen wir abschätzen, wie groß der erreichbare Abstieg für eine exakte Schrittweite ist.

11.16. Satz: *Die Zielfunktion f genüge den Voraussetzungen* **V1**, **V2** *und* **V3**. *Der Punkt $\boldsymbol{x} \in L_0$ sei keine stationäre Lösung und $\boldsymbol{p} \in \mathbb{R}^n$ sei eine Abstiegsrichtung für f in $\boldsymbol{x}$. Weiterhin sei $\varphi(\alpha) = f(\boldsymbol{x} + \alpha \boldsymbol{p})$. Ist α^* die erste positive Nullstelle von φ', so ist*

1.
$$-\frac{\boldsymbol{\nabla} f(\boldsymbol{x})^T \boldsymbol{p}}{L \|\boldsymbol{p}\|_2^2} \leqq \alpha^*,$$

2.
$$f(\boldsymbol{x}) - f(\boldsymbol{x} + \alpha^* \boldsymbol{p}) \geqq \frac{1}{2L} \left(\frac{\boldsymbol{\nabla} f(\boldsymbol{x})^T \boldsymbol{p}}{\|\boldsymbol{p}\|_2} \right)^2.$$

Beweis: φ ist monoton fallend auf $[0, \alpha^*]$. Weiterhin ist $\alpha^* \leqq \hat{\alpha}$, wobei $\hat{\alpha}$ die erste positive Nullstelle von $\varphi(0) - \varphi(\alpha)$ bezeichnet. Wegen der Lipschitzstetigkeit von $\boldsymbol{\nabla} f$ auf L_0 gilt

$$\begin{aligned} 0 &= \boldsymbol{\nabla} f(\boldsymbol{x} + \alpha^* \boldsymbol{p})^T \boldsymbol{p} + \boldsymbol{\nabla} f(\boldsymbol{x})^T \boldsymbol{p} - \boldsymbol{\nabla} f(\boldsymbol{x})^T \boldsymbol{p} \\ &= \boldsymbol{\nabla} f(\boldsymbol{x})^T \boldsymbol{p} + [\boldsymbol{\nabla} f(\boldsymbol{x} + \alpha^* \boldsymbol{p}) - \boldsymbol{\nabla} f(\boldsymbol{x})]^T \boldsymbol{p} \\ &\leqq \boldsymbol{\nabla} f(\boldsymbol{x})^T \boldsymbol{p} + \alpha^* L \|\boldsymbol{p}\|_2^2. \end{aligned}$$

Daraus folgt

$$\alpha^* \geqq -\frac{\boldsymbol{\nabla} f(\boldsymbol{x})^T \boldsymbol{p}}{L \|\boldsymbol{p}\|_2^2} = \tilde{\alpha}.$$

Wegen der Monotonie von φ gilt

$$\varphi(\alpha^*) \leqq \varphi(\tilde{\alpha})$$

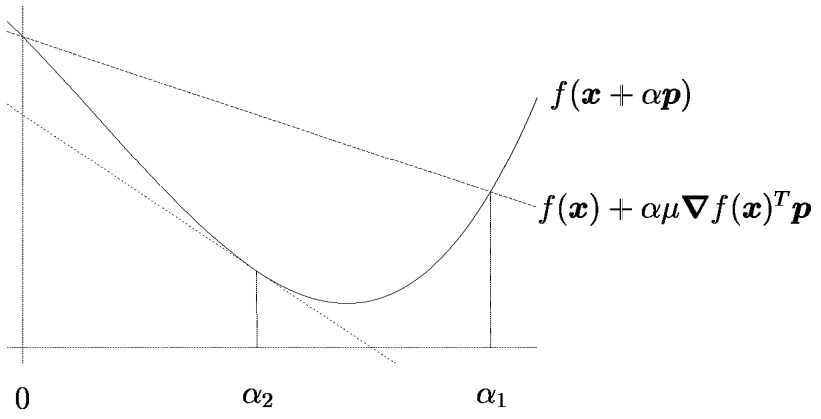
und damit

$$\begin{aligned} f(\boldsymbol{x}+\alpha^*\boldsymbol{p}) &\leqq f(\boldsymbol{x}+\tilde{\alpha}\boldsymbol{p}) \\ &\leqq f(\boldsymbol{x})+\tilde{\alpha}\boldsymbol{\nabla} f(\boldsymbol{x})^T\boldsymbol{p}+\tilde{\alpha}^2\frac{L}{2}\|\boldsymbol{p}\|_2^2 \\ &= f(\boldsymbol{x})+\left(-\frac{\boldsymbol{\nabla} f(\boldsymbol{x})^T\boldsymbol{p}}{L\|\boldsymbol{p}\|_2^2}\right)\boldsymbol{\nabla} f(\boldsymbol{x})^T\boldsymbol{p}+\left(-\frac{\boldsymbol{\nabla} f(\boldsymbol{x})^T\boldsymbol{p}}{L\|\boldsymbol{p}\|_2^2}\right)^2\frac{L}{2}\|\boldsymbol{p}\|_2^2 \\ &= f(\boldsymbol{x})-\frac{1}{2L}\left(\frac{\boldsymbol{\nabla} f(\boldsymbol{x})^T\boldsymbol{p}}{\|\boldsymbol{p}\|_2}\right)^2. \end{aligned}$$

❀

Leider lässt sich eine exakte Schrittweite i.a. nicht in endlich vielen Schritten berechnen. Darum verwendet man inexakte Schrittweiten.
Eine Variante ist die POWELL-**Schrittweite**:

> Für ein $\mu \in (0, 1/2)$ und ein $\nu \in (\mu, 1)$ wird zu einem Punkt $\boldsymbol{x}$ und einer Abstiegsrichtung $\boldsymbol{p}$ eine Schrittweite α^* so bestimmt, dass
>
> - $$f(\boldsymbol{x}+\alpha^*\boldsymbol{p}) \leqq f(\boldsymbol{x})+\alpha^*\mu\boldsymbol{\nabla} f(\boldsymbol{x})^T\boldsymbol{p} \quad \text{und}$$
> - $$\boldsymbol{\nabla} f(\boldsymbol{x}+\alpha^*\boldsymbol{p})^T\boldsymbol{p} \geqq \nu\boldsymbol{\nabla} f(\boldsymbol{x})^T\boldsymbol{p}$$
>
> gilt.



Für $0 \leqq \alpha^* \leqq \alpha_1$ ist die erste Bedingung erfüllt. Sie sichert einen hinreichend starken Abstieg.
Für $\alpha^* \geqq \alpha_2$ ist die zweite Bedingung erfüllt. Sie verhindert, dass zu kleine Schrittweiten gewählt werden.

Wir wollen nun zeigen, dass stets eine POWELL-Schrittweite existiert; Außerdem ist der erreichbare Abstieg abzuschätzen.

11.17. Satz: *Die Zielfunktion f genüge den Voraussetzungen* **V1**, **V2** *und* **V3**. $\boldsymbol{x} \in L_0$ *sei keine stationäre Lösung und $\boldsymbol{p} \in \mathbb{R}^n$ sei eine Abstiegsrichtung für f in $\boldsymbol{x}$. Für gegebenes $\mu \in (0, 1/2)$ und $\nu \in (\mu, 1)$ bezeichne*

$$T_P(\boldsymbol{x};\boldsymbol{p}) = \left\{ \alpha > 0 \;\middle|\; \begin{array}{l} f(\boldsymbol{x}+\alpha\boldsymbol{p}) \leqq f(\boldsymbol{x}) + \alpha\mu\boldsymbol{\nabla} f(\boldsymbol{x})^T, \\ \boldsymbol{\nabla} f(\boldsymbol{x}+\alpha\boldsymbol{p})^T\boldsymbol{p} \geqq \nu\boldsymbol{\nabla} f(\boldsymbol{x})^T\boldsymbol{p} \end{array} \right\}$$

die Menge der POWELL*-Schrittweiten im Punkt $\boldsymbol{x}$ in Richtung $\boldsymbol{p}$. Dann gilt*

1. *Die Menge $T_P(\boldsymbol{x};\boldsymbol{p})$ ist nicht leer.*
2. *Es existiert eine nur von μ, ν und der* LIPSCHITZ*-Konstanten L anhängende Konstante $\Theta > 0$ mit*

$$f(\boldsymbol{x}) - f(\boldsymbol{x}+\alpha\boldsymbol{p}) \geqq \Theta \left(\frac{\boldsymbol{\nabla} f(\boldsymbol{x})^T\boldsymbol{p}}{\|\boldsymbol{p}\|_2} \right)^2$$

für alle $\alpha \in T_P(\boldsymbol{x};\boldsymbol{p})$.

Beweis: Es sei

$$\boldsymbol{\Phi}(\alpha) = f(\boldsymbol{x}) - f(\boldsymbol{x}+\alpha\boldsymbol{p}) + \alpha\mu\boldsymbol{\nabla} f(\boldsymbol{x})^T \mp$$

und α^* die erste positive Nullstelle von $\boldsymbol{\nabla} f(\boldsymbol{x}+\alpha\boldsymbol{p})^T\boldsymbol{p}$. Dann gilt

$$\begin{aligned} \boldsymbol{\Phi}'(\alpha) &= -\boldsymbol{\nabla} f(\boldsymbol{x}+\alpha\boldsymbol{p})^T\boldsymbol{p} + \mu\boldsymbol{\nabla} f(\boldsymbol{x})^T\boldsymbol{p}, \\ \boldsymbol{\Phi}'(0) &= -\underbrace{(1-\mu)}_{>0}\underbrace{\boldsymbol{\nabla} f(\boldsymbol{x})^T\boldsymbol{p}}_{<0} > 0, \\ \boldsymbol{\Phi}'(\alpha^*) &= \mu\boldsymbol{\nabla} f(\boldsymbol{x})^T\boldsymbol{p} < 0. \end{aligned}$$

Wegen $\Phi(0) = 0$ existiert dann ein $\bar{\alpha} \in (0, \alpha^*)$ mit $\Phi(\bar{\alpha}) > 0$ und $\Phi'(\bar{\alpha}) = 0$. Daraus folgt

$$\begin{aligned} f(\boldsymbol{x}+\bar{\alpha}\boldsymbol{p}) &> f(\boldsymbol{x}) + \bar{\alpha}\mu\boldsymbol{\nabla} f(\boldsymbol{x})^T\boldsymbol{p}, \\ \boldsymbol{\nabla} f(\boldsymbol{x}+\bar{\alpha}\boldsymbol{p})^T\boldsymbol{p} &= \mu\boldsymbol{\nabla} f(\boldsymbol{x})^T\boldsymbol{p} > \nu\boldsymbol{\nabla} f(\boldsymbol{x})^T\boldsymbol{p}, \end{aligned}$$

daher $\bar{\alpha} \in T_P(\boldsymbol{x};\boldsymbol{p})$ und $T_P(\boldsymbol{x};\boldsymbol{p}) \neq \emptyset$. Weiter gilt

$$\begin{aligned} -(1-\nu)\boldsymbol{\nabla} f(\boldsymbol{x})^T\boldsymbol{p} &= \nu\boldsymbol{\nabla} f(\boldsymbol{x})^T\boldsymbol{p} - \boldsymbol{\nabla} f(\boldsymbol{x})^T\boldsymbol{p} \\ &\leqq [\boldsymbol{\nabla} f(\boldsymbol{x}+\alpha\boldsymbol{p}) - \boldsymbol{\nabla} f(\boldsymbol{x})]^T\boldsymbol{p} \\ &\leqq \|\boldsymbol{\nabla} f(\boldsymbol{x}+\alpha\boldsymbol{p}) - \boldsymbol{\nabla} f(\boldsymbol{x})\|_2 \, \|\boldsymbol{p}\|_2 \\ &\leqq L\|\boldsymbol{x}+\alpha\boldsymbol{p}-\boldsymbol{x}\|_2\|\boldsymbol{p}\|_2 \\ &= L\alpha\|\boldsymbol{p}\|_2^2. \end{aligned}$$

Daraus folgt

$$\alpha \geqq -\frac{1-\nu}{L}\frac{\boldsymbol{\nabla} f(\boldsymbol{x})^T\boldsymbol{p}}{\|\boldsymbol{p}\|_2^2}$$

und weiter

$$f(\boldsymbol{x}+\alpha\boldsymbol{p}) \leqq f(\boldsymbol{x}) + \alpha\mu\boldsymbol{\nabla} f(\boldsymbol{x})^T\boldsymbol{p} \leqq f(\boldsymbol{x}) - \frac{\mu(1-\nu)}{L}\left(\frac{\boldsymbol{\nabla} f(\boldsymbol{x})^T\boldsymbol{p}}{\|\boldsymbol{p}\|_2}\right)^2.$$

Es gilt daher die Ungleichung mit $\Theta = \frac{\mu(1-\nu)}{L}$. ❀

Eine andere Variante der Schrittweitenbestimmung ist das ARMIJO-Prinzip:
Man wählt wieder ein $\mu \in (0, 1/2)$ und testet zunächst mit der Schrittweite $\alpha = 1$ die Gültigkeit der Ungleichung

$$f(\boldsymbol{x}+\alpha\boldsymbol{p}) \leqq f(\boldsymbol{x}) + \alpha\mu\boldsymbol{\nabla} f(\boldsymbol{x})^T\boldsymbol{p}.$$

Ist die Ungleichung erfüllt, so wird α als Schrittweite akzeptiert, Anderenfalls wird α kontrolliert verkleinert. Das könnte wie folgt geschehen:

S0. Wähle Zahlen $\mu \in (0, 1/2)$ und $0 < \varrho \leqq \sigma < 1$. Setze $\alpha_0 = 1$ und $k = 0$.

S1. Falls $f(\boldsymbol{x}+\alpha_k\boldsymbol{p}) \leqq f(\boldsymbol{x}) + \alpha_k\mu\boldsymbol{\nabla} f(\boldsymbol{x})^T\boldsymbol{p}$, so STOPP. α_k ist die gesuchte Schrittweite.

S2. Wähle $\alpha_{k+1} \in [\varrho\alpha_k, \sigma\alpha_k]$.

S3. Setze $k = k+1$ und gehe zu Schritt **S1**.

Für $\varrho = \sigma$ ist die ARMIJO-Schrittweite durch $\alpha = \varrho^j$ gegeben, wobei j die kleinste ganze Zahl ist, für die die Ungleichung

$$f(\boldsymbol{x}+\varrho^j\boldsymbol{p}) \leqq f(\boldsymbol{x}) + \varrho^j\mu\boldsymbol{\nabla} f(\boldsymbol{x})^T\boldsymbol{p}$$

erfüllt ist.
Eine andere Variante stammt von HAN (1981). Er ersetzt die Funktion

$$\varphi(\alpha) = f(\boldsymbol{x}+\alpha\boldsymbol{p})$$

durch eine quadratische Funktion $q(\alpha)$ mit

- $q(0) = \varphi(0) = f(\boldsymbol{x})$,
- $q(\alpha_k) = \varphi(\alpha_k) = f(\boldsymbol{x} + \alpha_k \boldsymbol{p})$,
- $q'(0) = \varphi'(0) = \boldsymbol{\nabla} f(\boldsymbol{x})^T \boldsymbol{p}$.

Als Minimalpunkt dieser Ersatzfunktion ergibt sich

$$\alpha_k^* = -\frac{1}{2} \frac{\alpha_k^2 \boldsymbol{\nabla} f(\boldsymbol{x})^T \boldsymbol{p}}{f(\boldsymbol{x} + \alpha_k \boldsymbol{p}) - f(\boldsymbol{x}) - \alpha_k \boldsymbol{\nabla} f(\boldsymbol{x})^T \boldsymbol{p}}.$$

Nun wird die nächste Schrittweite durch

$$\alpha_{k+1} = \max\{0.1\alpha_k, \alpha_k^*\}$$

bestimmt.
Ist die Ungleichung $f(\boldsymbol{x} + \alpha_k \boldsymbol{p}) \leqq f(\boldsymbol{x}) + \alpha_k \mu \boldsymbol{\nabla} f(\boldsymbol{x})^T \boldsymbol{p}$ im k-ten Schritt nicht erfüllt, so folgt aus $\alpha_{k+1} \leqq \alpha_k^*$

$$\alpha_{k+1} \leqq \frac{1}{2(1-\mu)} \alpha_k.$$

Andererseits gilt immer $\alpha_{k+1} \geqq 0.1\alpha_k$. Damit entsprechen die gemäß HAN berechneten Schrittweiten dem ARMIJO-Konzept mit $\varrho = 0.1$ und $\sigma = \frac{1}{2(1-\mu)}$.
Auch für ARMIJO-Schrittweiten lässt sich eine zu Satz 11.17 analoge Existenzaussage machen, die wir hier ohne Beweis angeben.

11.18. Satz: *Die Zielfunktion f genüge den Voraussetzungen* **V1**, **V2** *und* **V3**. *$\boldsymbol{x} \in L_0$ sei keine stationäre Lösung und $\boldsymbol{p} \in \mathbb{R}^n$ sei eine Abstiegsrichtung für f in $\boldsymbol{x}$. Für gegebenes $\mu \in (0, 1/2)$ und $0 < \varrho \leqq \sigma < 1$ sei α eine zugehörige* ARMIJO*-Schrittweite. Dann existiert eine Konstante $\Theta > 0$, die nur von μ, ϱ, σ und der* LIPSCHITZ*-Konstanten L abhängt, mit*

$$f(\boldsymbol{x}) - f(\boldsymbol{x} + \alpha \boldsymbol{p}) \geqq \Theta \min\left\{ \left(\frac{\boldsymbol{\nabla} f(\boldsymbol{x})^T \boldsymbol{p}}{\|\boldsymbol{p}\|_2} \right)^2 ; -\boldsymbol{\nabla} f(\boldsymbol{x})^T \boldsymbol{p} \right\}.$$

11.2.2. Konvergenz des Modellalgorithmus

11.19. Satz: *Die Zielfunktion f genüge den Voraussetzungen* **V1**, **V2** *und* **V3**. *Als Schrittweite im Modellalgorithmus werde*

- *die exakte Schrittweite $\alpha_k = \alpha^*\left(\boldsymbol{x}^{(k)}; \boldsymbol{p}^{(k)}\right)$ oder*

- *die* POWELL*-Schrittweite* $\alpha_k = \alpha_P\left(\boldsymbol{x}^{(k)};\boldsymbol{p}^{(k)}\right)$ *oder*
- *die* ARMIJO*-Schrittweite* $\alpha_k = \alpha_A\left(\boldsymbol{x}^{(k)};\boldsymbol{p}^{(k)}\right)$

verwendet. Weiterhin existiere eine Konstante $\gamma > 0$ *mit*

$$-\boldsymbol{\nabla} f(\boldsymbol{x}^{(k)})^T \boldsymbol{p}^{(k)} \geqq \gamma \|\boldsymbol{\nabla} f(\boldsymbol{x}^{(k)})\| \, \|\boldsymbol{p}^{(k)}\|$$

und eine Konstante $\tau > 0$ *mit*

$$\|\boldsymbol{p}^{(k)}\| \geqq \tau \|\boldsymbol{\nabla} f(\boldsymbol{x}^{(k)})\|, \quad k = 0, 1, \ldots.$$

Dann gilt:

1. *Jeder Häufungspunkt der erzeugten Folge* $\left\{\boldsymbol{x}^{(k)}\right\}_{k\in\mathbb{N}}$ *ist eine stationäre Lösung des Minumumproblems*

$$\min_{\boldsymbol{x}\in\mathbb{R}^n} f(\boldsymbol{x}).$$

2. *Besitzt das Minumumproblem genau eine stationäre Lösung* $\boldsymbol{x}^*$ *in der Niveaumenge* L_0*, so konvergiert die gesamte Folge* $\left\{\boldsymbol{x}^{(k)}\right\}_{k\in\mathbb{N}}$ *gegen* $\boldsymbol{x}^*$*.*

Beweis: Wegen Satz 11.16, Satz 11.17 und Satz 11.18 existiert eine Konstante $\Theta > 0$ mit

$$f(\boldsymbol{x}) - f(\boldsymbol{x} + \alpha\boldsymbol{p}) \geqq \Theta \min\left\{\left(\frac{\boldsymbol{\nabla} f(\boldsymbol{x})^T \boldsymbol{p}}{\|\boldsymbol{p}\|_2}\right)^2; -\boldsymbol{\nabla} f(\boldsymbol{x})^T \boldsymbol{p}\right\}.$$

Mit den beiden Voraussetzungen folgt dann

$$f(\boldsymbol{x}) - f(\boldsymbol{x} + \alpha\boldsymbol{p}) \geqq \Theta \min\{\gamma\tau, \gamma^2\} \, \|\boldsymbol{\nabla} f(\boldsymbol{x}^{(k)})\|^2.$$

Da $\left\{f(\boldsymbol{x}^{(k)})\right\}_{k\in\mathbb{N}}$ eine monoton fallende, nach unten beschränkte Folge ist, muss die Folge $\left\{\boldsymbol{\nabla} f(\boldsymbol{x}^{(k)})\right\}_{k\in\mathbb{N}}$ gegen den Nullvektor konvergieren. Ist $\boldsymbol{x}^*$ ein Häufungspunkt von $\left\{\boldsymbol{x}^{(k)}\right\}_{k\in\mathbb{N}}$, so existiert eine Teilfolge $\left\{\boldsymbol{x}^{(k_j)}\right\}_{j\in\mathbb{N}} \subseteq \left\{\boldsymbol{x}^{(k)}\right\}_{k\in\mathbb{N}}$ mit

$$\lim_{j\to\infty} \boldsymbol{x}^{(k_j)} = \boldsymbol{x}^*.$$

Wegen

$$\lim_{k\to\infty} \boldsymbol{\nabla} f(\boldsymbol{x}^{(k)}) = \boldsymbol{o}$$

gilt dann aber $\nabla f(\boldsymbol{x}^*) = \boldsymbol{o}$. $\boldsymbol{x}^*$ ist daher stationäre Lösung. Damit ist die erste Aussage bewiesen.
Wir nehmen nun an, das Minumumproblem besitzt genau eine stationäre Lösung $\boldsymbol{x}^*$. Falls $\left\{\boldsymbol{x}^{(k)}\right\}_{k\in\mathbb{N}}$ nicht gegen $\boldsymbol{x}^*$ konvergiert, existieren ein $\varepsilon > 0$ und eine Teilfolge $\left\{\boldsymbol{x}^{(k_j)}\right\}_{j\in\mathbb{N}} \subseteq \left\{\boldsymbol{x}^{(k)}\right\}_{k\in\mathbb{N}}$ mit

$$\|\boldsymbol{x}^{(k_j)} - \boldsymbol{x}^*\| \geqq \varepsilon \quad j = 1, 2, \ldots$$

Da L_0 als kompakt vorausgesetzt wurde, besitzt die Folge $\left\{\boldsymbol{x}^{(k_j)}\right\}_{j\in\mathbb{N}}$ einen Häufungspunkt $\hat{\boldsymbol{x}}$. Nach dem gerade bewiesenen ersten Teil gilt dann aber $\nabla f(\hat{\boldsymbol{x}}) = \boldsymbol{o}$, d. h. $\hat{\boldsymbol{x}}$ ist stationäre Lösung. Andererseits folgt aber aus $\|\hat{\boldsymbol{x}} - \boldsymbol{x}^*\| \geqq \varepsilon > 0$, dass $\hat{\boldsymbol{x}} \neq \boldsymbol{x}^*$ sein muss, im Widerspruch zur Voraussetzung „ $\boldsymbol{x}^*$ ist einzige stationäre Lösung“. Damit muss die Folge $\left\{\boldsymbol{x}^{(k)}\right\}_{k\in\mathbb{N}}$ gegen $\boldsymbol{x}^*$ konvergieren. ❀

Bemerkungen: (i) Die konkrete Art der Schrittweitenberechnung spielt in diesem Konvergenzsatz keine Rolle. Wichtig ist nur die Sicherung eines im Sinne von

$$f(\boldsymbol{x}) - f(\boldsymbol{x} + \alpha\boldsymbol{p}) \geqq \Theta \min\left\{\left(\frac{\nabla f(\boldsymbol{x})^T\boldsymbol{p}}{\|\boldsymbol{p}\|_2}\right)^2; -\nabla f(\boldsymbol{x})^T\boldsymbol{p}\right\}$$

hinreichend starken Abstiegs pro Schritt.
(ii) Für exakte und POWELL-Schrittweiten darf auf die Voraussetzung

$$\|\boldsymbol{p}^{(k)}\| \geqq \tau\|\nabla f(\boldsymbol{x}^{(k)})\|$$

mit $\tau > 0$ verzichtet werden. Nun wollen wir wieder eine glatte, gleichmäßig konvexe Zielfunktion betrachten. Zur Vorbereitung beweisen wir den folgenden Hilfssatz.

11.20. Satz: *Gegeben sei das freie Minimumproblem*

$$\min_{\boldsymbol{x}\in\mathbb{R}^n} f(\boldsymbol{x}).$$

Die Zielfunktion $f : \mathbb{R}^n \longrightarrow \mathbb{R}$ erfülle folgende Voraussetzungen:

K1 *Für ein gegebenes $\boldsymbol{x}^{(0)} \in \mathbb{R}^n$ (i.a. der Startpunkt des Verfahrens) ist die Niveaumenge*
$L_0 = \left\{\, \boldsymbol{x} \in \mathbb{R}^n \mid f(\boldsymbol{x}) \leqq f(\boldsymbol{x}^{(0)}) \,\right\}$ *konvex.*

K2 *Die Zielfunktion f ist auf einer offenen Menge $D \supseteq L_0$ stetig differenzierbar und auf L_0 gleichmäßig konvex: Es existiert eine Konstante $\gamma > 0$ mit*

$$\frac{\gamma}{2}\|\boldsymbol{y}-\boldsymbol{x}\|_2^2 + \boldsymbol{\nabla} f(\boldsymbol{x})^T(\boldsymbol{y}-\boldsymbol{x}) \leqq f(\boldsymbol{y}) - f(\boldsymbol{x})$$

für alle $\boldsymbol{x}, \boldsymbol{y} \in L_0$.

K3 *Der Gradient $\boldsymbol{\nabla} f(\boldsymbol{x}) = f'(\boldsymbol{x})^T$ ist auf L_0 lipschitzstetig: Es existiert eine Konstante $L > 0$ mit*

$$\|\boldsymbol{\nabla} f(\boldsymbol{x}) - \boldsymbol{\nabla} f(\boldsymbol{y})\|_2 \leqq L\|\boldsymbol{x}-\boldsymbol{y}\|_2$$

für alle $\boldsymbol{x}, \boldsymbol{y} \in L_0$.

Dann ist die Niveaumenge L_0 kompakt und das Minumumproblem besitzt genau eine globale Lösung $\boldsymbol{x}^* \in L_0$. $\boldsymbol{x}^*$ ist auch die einzige stationäre Lösung auf L_0. Es gilt die Fehlerabschätzung

$$\frac{\gamma}{2}\|\boldsymbol{x}-\boldsymbol{x}^*\|_2^2 \leqq f(\boldsymbol{x}) - f(\boldsymbol{x}^*) \leqq \frac{1}{2\gamma}\|\boldsymbol{\nabla} f(\boldsymbol{x})\|_2^2 \tag{11.1}$$

für alle $\boldsymbol{x} \in L_0$.

Beweis: Die Niveaumenge L_0 ist offensichtlich abgeschlossen. Aus der gleichmäßigen Konvexität von f auf L_0 folgt

$$\frac{\gamma}{2}\|\boldsymbol{x}-\boldsymbol{x}^{(0)}\|_2^2 + \boldsymbol{\nabla} f(\boldsymbol{x}^{(0)})^T(\boldsymbol{x}-\boldsymbol{x}^{(0)}) \leqq f(\boldsymbol{x}) - f(\boldsymbol{x}^{(0)}) \leqq 0.$$

Mit Hilfe der CAUCHY-SCHWARZschen Ungleichung erhält man daraus

$$\frac{\gamma}{2}\|\boldsymbol{x}-\boldsymbol{x}^{(0)}\|_2^2 \leqq -\boldsymbol{\nabla} f(\boldsymbol{x}^{(0)})^T(\boldsymbol{x}-\boldsymbol{x}^{(0)}) \leqq \|\boldsymbol{\nabla} f(\boldsymbol{x}^{(0)})\|_2\|\boldsymbol{x}-\boldsymbol{x}^{(0)}\|_2,$$

und weiter

$$\|\boldsymbol{x}-\boldsymbol{x}^{(0)}\|_2 \leqq \frac{2}{\gamma}\|\boldsymbol{\nabla} f(\boldsymbol{x}^{(0)})\|_2.$$

Damit ist L_0 auch beschränkt und daher kompakt. f nimmt dann sein globales Minimum auf L_0 an. Dieses globale Minimum ist offensichtlich auch globale Lösung des Minumumproblems. Damit ist die Existenz einer globalen Lösung $\boldsymbol{x}^*$ mit $\boldsymbol{x}^* \in L_0$ bewiesen.
Wir zeigen nun die Gültigkeit der Abschätzung 11.1, aus der auch die Eindeutigkeit

von $\boldsymbol{x}^*$ folgt. Aus der gleichmäßigen Konvexität von f auf L_0 **(K2)** erhält man mit $\boldsymbol{y} = \boldsymbol{x}$, $\boldsymbol{x} = \boldsymbol{x}^*$ und $\boldsymbol{\nabla} f(\boldsymbol{x}^*) = \boldsymbol{o}$

$$\frac{\gamma}{2}\|\boldsymbol{x} - \boldsymbol{x}^*\| \leqq f(\boldsymbol{x}) - f(\boldsymbol{x}^*).$$

Das ist die erste zu beweisende Ungleichung. Zum Beweis der zweiten Ungleichung betrachten wir das Minumumproblem

$$\min_{\boldsymbol{p}\in\mathbb{R}^n} \frac{\gamma}{2}\|\boldsymbol{p}\|_2^2 + \boldsymbol{\nabla} f(\boldsymbol{x})^T \boldsymbol{p}$$

für ein festes $\boldsymbol{x} \in L_0$. Die Lösung ist offensichtlich durch

$$\boldsymbol{p} = -\frac{1}{\gamma}\boldsymbol{\nabla} f(\boldsymbol{x})$$

gegeben. Dann gilt für beliebige $\boldsymbol{p} \in \mathbb{R}^n$

$$\frac{\gamma}{2}\|\boldsymbol{p}\|_2^2 + \boldsymbol{\nabla} f(\boldsymbol{x})^T \boldsymbol{p} \geqq \frac{\gamma}{2}\left\|\frac{1}{\gamma}\boldsymbol{\nabla} f(\boldsymbol{x})\right\|_2^2 - \frac{1}{\gamma}\boldsymbol{\nabla} f(\boldsymbol{x})^T \boldsymbol{\nabla} f(\boldsymbol{x}) = -\frac{1}{2\gamma}\|\boldsymbol{\nabla} f(\boldsymbol{x})\|_2^2.$$

Speziell für $\boldsymbol{p} = \boldsymbol{x}^* - \boldsymbol{x}$ gilt

$$\frac{\gamma}{2}\|\boldsymbol{x}^* - \boldsymbol{x}\|_2^2 + \boldsymbol{\nabla} f(\boldsymbol{x})^T (\boldsymbol{x}^* - \boldsymbol{x}) \geqq -\frac{1}{2\gamma}\|\boldsymbol{\nabla} f(\boldsymbol{x})\|_2^2.$$

Wegen der gleichmäßigen Konvexität gilt

$$f(\boldsymbol{x}^*) - f(\boldsymbol{x}) \geqq \frac{\gamma}{2}\|\boldsymbol{x}^* - \boldsymbol{x}\|_2^2 + \boldsymbol{\nabla} f(\boldsymbol{x})^T (\boldsymbol{x}^* - \boldsymbol{x}),$$

also insgesamt

$$f(\boldsymbol{x}^*) - f(\boldsymbol{x}) \geqq -\frac{1}{2\gamma}\|\boldsymbol{\nabla} f(\boldsymbol{x})\|_2^2$$

beziehungsweise

$$f(\boldsymbol{x}) - f(\boldsymbol{x}^*) \leqq \frac{1}{2\gamma}\|\boldsymbol{\nabla} f(\boldsymbol{x})\|_2^2.$$

Wäre nun $\bar{\boldsymbol{x}}$ eine weitere stationäre Lösung in L_0, so würde aus der Abschätzung 11.1

$$\|\bar{\boldsymbol{x}} - \boldsymbol{x}^*\| \leqq \frac{1}{\gamma^2}\|\boldsymbol{\nabla} f(\bar{\boldsymbol{x}})\|_2^2 = 0$$

folgen, d. h. $\bar{\boldsymbol{x}} = \boldsymbol{x}^*$. ❀

Für gleichmäßig konvexe Zielfunktionen ist nun eine hinreichende Bedingung für die Konvergenz des Modellalgorithmus angebbar.

11.21. Satz: *Die Zielfunktion f genüge den Voraussetzungen* **K1**, **K2** *und* **K3** *aus Satz 11.20. Als Schrittweite im Modellalgorithmus werde*

- *die exakte Schrittweite* $\alpha_k = \alpha^*\left(\boldsymbol{x}^{(k)};\boldsymbol{p}^{(k)}\right)$ *oder*
- *die* POWELL*-Schrittweite* $\alpha_k = \alpha_P\left(\boldsymbol{x}^{(k)};\boldsymbol{p}^{(k)}\right)$ *oder*
- *die* ARMIJO*-Schrittweite* $\alpha_k = \alpha_A\left(\boldsymbol{x}^{(k)};\boldsymbol{p}^{(k)}\right)$

verwendet. Weiterhin sei

$$\delta_k = \min\left\{-\frac{\boldsymbol{\nabla} f(\boldsymbol{x}^{(k)})^T\boldsymbol{p}^{(k)}}{\|\boldsymbol{\nabla} f(\boldsymbol{x}^{(k)})\|^2};\left(\frac{\boldsymbol{\nabla} f(\boldsymbol{x}^{(k)})^T\boldsymbol{p}^{(k)}}{\|\boldsymbol{\nabla} f(\boldsymbol{x}^{(k)})\|\,\|\boldsymbol{p}^{(k)}\|}\right)^2\right\}$$

für ARMIJO*-Schrittweiten, bzw.*

$$\delta_k = \left(\frac{\boldsymbol{\nabla} f(\boldsymbol{x}^{(k)})^T\boldsymbol{p}^{(k)}}{\|\boldsymbol{\nabla} f(\boldsymbol{x}^{(k)})\|\,\|\boldsymbol{p}^{(k)}\|}\right)^2$$

für exakte oder POWELL*-Schrittweiten. Dann gilt:*

1. *Ist*

   $$\sum_{j=0}^{\infty}\delta_j = \infty,$$

   so konvergiert die durch den Modellalgorithmus erzeugte Folge $\left\{\boldsymbol{x}^{(k)}\right\}_{k\in\mathbb{N}}$ *gegen die eindeutige globale Lösung* $\boldsymbol{x}^*$.
2. *Existiert ein* $\delta > 0$ *mit*

   $$\frac{1}{k+1}\sum_{j=0}^{k}\delta_j > \delta$$

   für $k = 0,1,\ldots$, *so existieren Konstanten* $C > 0$ *und* $q \in (0,1)$ *mit*

   $$\|\boldsymbol{x}^{(k)} - \boldsymbol{x}^*\| \leqq Cq^k$$

   für $k = 0,1,\ldots$

Beweis: Nach Satz 11.16 und Satz 11.17 existiert sowohl für exakte als auch für und POWELL-Schrittweiten eine Konstante $\Theta > 0$ mit

$$f(\boldsymbol{x}^{(k)}) - f(\boldsymbol{x}^{(k+1)}) = f(\boldsymbol{x}^{(k)}) - f(\boldsymbol{x}^{(k)} + \alpha_k \boldsymbol{p}^{(k)}) \geqq \Theta \left(\frac{\boldsymbol{\nabla} f(\boldsymbol{x}^{(k)})^T \boldsymbol{p}^{(k)}}{\|\boldsymbol{p}^{(k)}\|_2} \right)^2.$$

Nach Satz 11.18 existiert für ARMIJO-Schrittweiten eine Konstante $\Theta > 0$ mit

$$\begin{aligned} f(\boldsymbol{x}^{(k)}) - f(\boldsymbol{x}^{(k+1)}) &= f(\boldsymbol{x}^{(k)}) - f(\boldsymbol{x}^{(k)} + \alpha_k \boldsymbol{p}^{(k)}) \\ &\geqq \Theta \min \left\{ \left(\frac{\boldsymbol{\nabla} f(\boldsymbol{x}^{(k)})^T \boldsymbol{p}^{(k)}}{\|\boldsymbol{p}^{(k)}\|_2} \right)^2 ; -\boldsymbol{\nabla} f(\boldsymbol{x}^{(k)})^T \boldsymbol{p}^{(k)} \right\}. \end{aligned}$$

Mit der Definition von δ_k folgt dann

$$f(\boldsymbol{x}^{(k)}) - f(\boldsymbol{x}^{(k+1)}) \geqq \Theta \delta_k \|\boldsymbol{\nabla} f(\boldsymbol{x}^{(k)})\|^2.$$

Wendet man die Fehlerabschätzung aus Satz 11.20 an, so erhält man

$$f(\boldsymbol{x}^{(k)}) - f(\boldsymbol{x}^{(k+1)}) \geqq 2\gamma\Theta\delta_k \left[f(\boldsymbol{x}^{(k)}) - f(\boldsymbol{x}^*) \right].$$

Damit folgt

$$\begin{aligned} 0 &\leqq f(\boldsymbol{x}^{(k+1)}) - f(\boldsymbol{x}^*) \\ &= f(\boldsymbol{x}^{(k+1)}) - f(\boldsymbol{x}^{(k)}) + f(\boldsymbol{x}^{(k)}) - f(\boldsymbol{x}^*) \\ &\leqq (1 - 2\gamma\Theta\delta_k) \left[f(\boldsymbol{x}^{(k)}) - f(\boldsymbol{x}^*) \right] \\ &\leqq \prod_{j=0}^{k} (1 - 2\gamma\Theta\delta_j) \cdot \left[f(\boldsymbol{x}^{(0)}) - f(\boldsymbol{x}^*) \right] \\ &\leqq \exp\left(-2\gamma\Theta \sum_{j=0}^{k} \delta_j \right) \cdot \left[f(\boldsymbol{x}^{(0)}) - f(\boldsymbol{x}^*) \right]. \end{aligned}$$

Wegen

$$\lim_{k\to\infty} \sum_{j=0}^{k} \delta_j = \infty$$

gilt dann

$$\lim_{k\to\infty} \exp\left(-2\gamma\Theta \sum_{j=0}^{k} \delta_j \right) = 0$$

und

$$0 \leqq \lim_{k\to\infty} f(\boldsymbol{x}^{(k+1)}) - f(\boldsymbol{x}^*) \leqq 0,$$

also

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

Nach Satz 11.20 gilt weiterhin für $\boldsymbol{x}^{(k)} \in L_0$

$$\begin{aligned}\|\boldsymbol{x}^{(k)} - \boldsymbol{x}^*\| &\leqq \sqrt{\frac{2}{\gamma}\left[f(\boldsymbol{x}^{(k)}) - f(\boldsymbol{x}^*)\right]} \\ &\leqq \sqrt{\frac{2}{\gamma}\exp\left(-2\gamma\Theta\sum_{j=0}^{k-1}\delta_j\right)\left[f(\boldsymbol{x}^{(0)}) - f(\boldsymbol{x}^*)\right]}.\end{aligned}$$

Daraus ergibt sich sofort

$$\lim_{k\to\infty} \boldsymbol{x}^{(k)} = \boldsymbol{x}^*.$$

Falls ein δ mit

$$\sum_{j=0}^{k-1} \delta_j \geqq k\delta$$

für $k = 0, 1, \ldots$ existiert, so folgt

$$\|\boldsymbol{x}^{(k)} - \boldsymbol{x}^*\| \leqq \mathrm{e}^{-k\gamma\Theta\delta}\sqrt{\frac{2}{\gamma}\left[f(\boldsymbol{x}^{(0)}) - f(\boldsymbol{x}^*)\right]} = Cq^k$$

mit

$$C = \sqrt{\frac{2}{\gamma}\left[f(\boldsymbol{x}^{(0)}) - f(\boldsymbol{x}^*)\right]}, \quad q = \mathrm{e}^{-\gamma\Theta\delta} \in (0,1).$$

❀

Bemerkungen: (i) Die konkrete Schrittweitenwahl spielt für die Konvergenzaussage wieder keine Rolle. Wesentlich ist nur die Sicherung eines hinreichend starken

Abstiegs.
(ii) Für exakte und POWELL-Schrittweiten ist

$$\delta_k = \left(\frac{\nabla f(\boldsymbol{x}^{(k)})^T \boldsymbol{p}^{(k)}}{\|\nabla f(\boldsymbol{x}^{(k)})\| \, \|\boldsymbol{p}^{(k)}\|} \right)^2$$

ein Maß für die Größe des Winkels zwischen der Richtung $\boldsymbol{p}^{(k)}$ und dem negativen Gradienten von f im Punkt $\boldsymbol{x}^{(k)}$. Die Bedingung

$$\sum_{j=0}^{\infty} \delta_j = \infty$$

besagt dann, dass diese Winkel nicht zu schnell gegen $\pi/2$ gehen dürfen.

11.3. Quasi-Newton-Verfahren

11.3.1. Gedämpftes und ungedämpftes Newton-Verfahren

Wir betrachten wieder das freie Minimumproblem

$$\min_{\boldsymbol{x} \in \mathbb{R}^n} f(\boldsymbol{x})$$

mit der auf dem gesamten $\mathbb{R}^n$ zweimal stetig differenzierbaren Zielfunktion

$$f : \mathbb{R}^n \longrightarrow \mathbb{R}.$$

Stationäre Lösungen dieses Problems sind Lösungen des im allgemeinen nichtlinearen Gleichungssystems

$$\nabla f(\boldsymbol{x}) = \boldsymbol{o}.$$

Zum Lösen dieses Systems könnte man das NEWTON-Verfahren anwenden. Hier wird ausgehend von einem Startpunkt $\boldsymbol{x}^{(0)} \in \mathbb{R}^n$ mit der Rekursionsformel

$$\boldsymbol{x}^{(k+1)} = \boldsymbol{x}^{(k)} - \left(\nabla^2 f(\boldsymbol{x}^{(k)}) \right)^{-1} \nabla f(\boldsymbol{x}^{(k)})$$

eine Folge $\{\boldsymbol{x}^{(k)}\}_{k \in \mathbb{N}}$ berechnet. Der Konvergenzsatz für das NEWTON-Verfahren, angewendet auf dieses spezielle Problem, lautet:

11.22. Satz: *Die Funktion $f : \mathbb{R}^n \longrightarrow \mathbb{R}$ sei auf einer offenen Umgebung des Punktes $\boldsymbol{x}^* \in \mathbb{R}^n$ zweimal stetig differenzierbar. Es sei $\boldsymbol{\nabla} f(\boldsymbol{x}^*) = \boldsymbol{o}$ und $\boldsymbol{\nabla}^2 f(\boldsymbol{x}^*)$ sei regulär. Dann existiert ein $\delta > 0$, so dass für jedes*

$$\boldsymbol{x}^{(0)} \in U_\delta(\boldsymbol{x}^*) = \{\, \boldsymbol{x} \in \mathbb{R}^n \mid \|\boldsymbol{x} - \boldsymbol{x}^*\| < \delta \,\}$$

die durch das NEWTON*-Verfahren*

$$\boldsymbol{x}^{(k+1)} = \boldsymbol{x}^{(k)} - \left(\boldsymbol{\nabla}^2 f(\boldsymbol{x}^{(k)})\right)^{-1} \boldsymbol{\nabla} f(\boldsymbol{x}^{(k)})$$

erzeugte Folge $\{\boldsymbol{x}^{(k)}\}_{k\in\mathbb{N}}$ definiert ist und superlinear gegen $\boldsymbol{x}^*$ konvergiert:

$$\lim_{k\to\infty} \boldsymbol{x}^{(k)} = \boldsymbol{x}^*, \quad \lim_{k\to\infty} \frac{\|\boldsymbol{x}^{(k+1)} - \boldsymbol{x}^*\|}{\|\boldsymbol{x}^{(k)} - \boldsymbol{x}^*\|} = 0.$$

Ist außerdem $\boldsymbol{\nabla}^2 f(\boldsymbol{x})$ auf einer hinreichend kleinen Kugel

$$U_\eta(\boldsymbol{x}^*) = \{\, \boldsymbol{x} \in \mathbb{R}^n \mid \|\boldsymbol{x} - \boldsymbol{x}^*\| < \eta \,\}$$

lipschitzstetig mit der LIPSCHITZ*-Konstanten L:*

$$\forall \boldsymbol{x} \in U_\eta : \quad \|\boldsymbol{\nabla}^2 f(\boldsymbol{x}) - \boldsymbol{\nabla}^2 f(\boldsymbol{x}^*)\| \leqq L\|\boldsymbol{x} - \boldsymbol{x}^*\|,$$

so konvergiert die Folge $\{\boldsymbol{x}^{(k)}\}_{k\in\mathbb{N}}$ lokal quadratisch gegen $\boldsymbol{x}^*$.

Die Konvergenzeigenschaften dieses Verfahrens hängen stark von der Größe der Umgebungen $U_\delta(\boldsymbol{x}^*)$ und $U_\eta(\boldsymbol{x}^*)$ ab. Man sollte nun versuchen, durch Einführung von Schrittweiten globale Konvergenzeigenschaften zu erhalten. Damit kommt man zum gedämpften NEWTON-Verfahren:

$$\boldsymbol{x}^{(k+1)} = \boldsymbol{x}^{(k)} - \alpha_k \left(\boldsymbol{\nabla}^2 f(\boldsymbol{x}^{(k)})\right)^{-1} \boldsymbol{\nabla} f(\boldsymbol{x}^{(k)}).$$

Unter entsprechenden Konvexitätsvoraussetzungen an f lässt sich die Regularität und sogar die positive Definitheit von $\boldsymbol{\nabla}^2 f(\boldsymbol{x}^{(k)})$ sichern. Dann ist die NEWTON-Richtung

$$\boldsymbol{p}^{(k)} = \left(\boldsymbol{\nabla}^2 f(\boldsymbol{x}^{(k)})\right)^{-1} \boldsymbol{\nabla} f(\boldsymbol{x}^{(k)})$$

eine Abstiegsrichtung, und man wird die Konvergenz des gedämpften NEWTON-Verfahrens erwarten. Die Schrittweiten sollten so gewählt werden, dass das Verfahren bei hinreichender Annäherung an $\boldsymbol{x}^*$ automatisch in das ungedämpfte NEWTON-Verfahren (Schrittweite 1) übergeht. Wir geben den entsprechenden Konvergenzsatz ohne Beweis an.

11.23. Satz: *Die Zielfunktion f des freien Minimumproblems*

$$\min_{\boldsymbol{x}\in\mathbb{R}^n} f(\boldsymbol{x})$$

genüge den Voraussetzungen:

1. *Mit einem $\boldsymbol{x}^{(0)} \in \mathbb{R}^n$ ist die Niveaumenge*

   $$L_0 = \left\{ \boldsymbol{x} \in \mathbb{R}^n \mid f(\boldsymbol{x}) \leqq f(\boldsymbol{x}^{(0)}) \right\}$$

   konvex.

2. *Die Funktion f ist auf einer Umgebung von L_0 zweimal stetig differenzierbar und es existieren Konstanten $0 < m \leqq M < \infty$, so dass*

   $$m\|\boldsymbol{p}\|_2^2 \leqq \boldsymbol{p}^T \boldsymbol{\nabla}^2 f(\boldsymbol{x})\boldsymbol{p} \leqq M\|\boldsymbol{p}\|_2^2$$

   für alle $\boldsymbol{x} \in L_0$ und alle $\boldsymbol{p} \in \mathbb{R}^n$ gilt.

Für das gedämpfte NEWTON*-Verfahren*

$$\boldsymbol{x}^{(k+1)} = \boldsymbol{x}^{(k)} + \alpha_k \boldsymbol{p}^{(k)}, \quad \boldsymbol{p}^{(k)} = -\left(\boldsymbol{\nabla}^2 f(\boldsymbol{x}^{(k)})\right)^{-1} \boldsymbol{\nabla} f(\boldsymbol{x}^{(k)}),$$

wobei zur Schrittweitenbestimmung das ARMIJO*-Prinzip mit $\mu \in (0, 1/2)$ angewendet wird, gilt:*

- *Bricht das Verfahren nicht vorzeitig mit der Lösung $\boldsymbol{x}^*$ ab, so konvergiert die vom Verfahren erzeugte Folge $\{\boldsymbol{x}^{(k)}\}_{k\in\mathbb{N}}$ gegen $\boldsymbol{x}^*$.*
- *Für hinreichend großes k geht das Verfahren in das ungedämpfte* NEWTON*-Verfahren über.*

Bemerkung: Falls f zweimal stetig differenzierbar und $\boldsymbol{\nabla}^2 f(\boldsymbol{x}^*)$ positiv definit ist, ist die zweite Voraussetzung immer erfüllt. Die entsprechende Umgebung ist oft klein, wie das folgende Beispiel zeigt.

11.24. Beispiel: Wir betrachten die ROSENBROCK-Funktion

$$f(\boldsymbol{x}) = 100(x_2 - x_1^2)^2 + (1 - x_1)^2.$$

Es gilt

$$\boldsymbol{\nabla} f(\boldsymbol{x}) = \begin{pmatrix} 400x_1^3 - 400x_1x_2 + 2x_1 - 2 \\ -200x_1^2 + 200x_2 \end{pmatrix}$$

und

$$\nabla^2 f(\boldsymbol{x}) = \begin{pmatrix} 1200x_1^2 - 400x_2 + 2 & -400x_1 \\ -400x_1 & 200 \end{pmatrix}.$$

f besitzt das einzige globale Minimum $\boldsymbol{x}^* = (1,1)^T$ mit

$$\nabla^2 f(\boldsymbol{x}^*) = \begin{pmatrix} 802 & -400 \\ -400 & 200 \end{pmatrix}.$$

Die HESSE-Matrix $\nabla^2 f$ besitzt im Punkt $\boldsymbol{x}^*$ die Eigenwerte

$$\lambda_1 = 501 + \sqrt{250601} \approx 1001.6, \quad \lambda_2 = 501 - \sqrt{250601} \approx 0.4.$$

Sie ist daher positiv definit. Der nächste Punkt mit nur positiv semidefiniter Matrix (ein Eigenwert ist 0) ist

$$\tilde{\boldsymbol{x}} = (0.9979\ldots, 1.0010\ldots).$$

Der Abstand zwischen $\boldsymbol{x}^*$ und $\tilde{\boldsymbol{x}}$ ist klein:

$$\|\boldsymbol{x}^* - \tilde{\boldsymbol{x}}\|_2 = 0.0022\ldots.$$

Für die ROSENBROCK-Funktion würde damit das gedämpfte NEWTON-Verfahren erst in einer Umgebung des globalen Minimums vom Radius $\delta \approx 0.0022$ konvergieren. ♡

Das gedämpfte NEWTON-Verfahren ist ein typisches lokales Verfahren. Weitab von einem lokalen Minimum wird die HESSE-Matrix im allgemeinen nicht positiv definit sein. Dann ist aber durch die NEWTON-Richtung

$$\boldsymbol{p}^{(k)} = -\left(\nabla^2 f(\boldsymbol{x}^{(k)})\right)^{-1} \nabla f(\boldsymbol{x}^{(k)})$$

keine Abstiegsrichtung gegeben. Doch selbst dann, wenn man so nahe an einer lokalen Lösung ist, dass die positive Definitheit der HESSE-Matrix gesichert ist, hat das gedämpfte NEWTON-Verfahren noch einen großen Nachteil. In jedem Schritt ist nicht nur der Gradient von f zu berechnen, sondern auch noch die HESSE-Matrix. Oft ist aber die Funktion f nicht durch einen analytischen Ausdruck gegeben, so dass man schon bei der Gradientenberechnung Näherungsformeln anzuwenden hat. Dann ist die Approximation der HESSE-Matrix erst recht problematisch. Außerdem ist in jedem Schritt ein lineares Gleichungssystem der Dimension n zu lösen. Da es im allgemeinen nicht möglich ist, Informationen aus dem vorigen Schritt zu verwenden, liegt der Aufwand pro Schritt bestenfalls in der Größenordnung von $n^3/6$ Rechenoperationen.

11.3.2. Verfahren der Oren-Luenberger-Klasse

Das NEWTON-Verfahren und das gedämpfte NEWTON-Verfahren sind nur lokal konvergent und damit praktisch nicht von großem Nutzen. Sie geben aber Hinweise zur Konstruktion global konvergenter Verfahren. Wir haben gesehen, dass die NEWTON-Richtung

$$\boldsymbol{p}^{(k)} = -\left(\boldsymbol{\nabla}^2 f(\boldsymbol{x}^{(k)})\right)^{-1} \boldsymbol{\nabla} f(\boldsymbol{x}^{(k)})$$

in der Nähe einer isolierten lokalen Lösung zu schnell konvergenten Verfahren führt. Wählt man nun positiv definite Matrizen $\boldsymbol{H}_k$ und damit Richtungen der Form

$$\boldsymbol{p}^{(k)} = -\boldsymbol{H}_k \boldsymbol{\nabla} f(\boldsymbol{x}^{(k)}),$$

so sind diese wegen

$$\boldsymbol{p}^{(k)^T} \boldsymbol{\nabla} f(\boldsymbol{x}^{(k)}) = -\boldsymbol{\nabla} f(\boldsymbol{x}^{(k)})^T \boldsymbol{H}_k \boldsymbol{\nabla} f(\boldsymbol{x}^{(k)}) < 0$$

Abstiegsrichtungen. Soll mit so definierten Abstiegsrichtungen ein im Sinne von Satz 11.19 konvergentes Verfahren erzeugt werden, so müssen Konstanten $\gamma > 0$ und $\tau > 0$ existieren, so dass für alle $k \geqq 0$ die Ungleichungen

$$-\boldsymbol{\nabla} f(\boldsymbol{x}^{(k)})^T \boldsymbol{p}^{(k)} \geqq \gamma \|\boldsymbol{\nabla} f(\boldsymbol{x}^{(k)})\| \, \|\boldsymbol{p}^{(k)}\|$$

und

$$\|\boldsymbol{p}^{(k)}\| \geqq \tau \|\boldsymbol{\nabla} f(\boldsymbol{x}^{(k)})\|$$

erfüllt sind. Bezüglich der euklidischen Norm heißt das

$$\begin{aligned}
\frac{\boldsymbol{\nabla} f(\boldsymbol{x}^{(k)})^T \boldsymbol{H}_k \boldsymbol{\nabla} f(\boldsymbol{x}^{(k)})}{\|\boldsymbol{\nabla} f(\boldsymbol{x}^{(k)})\|_2 \|\boldsymbol{H}_k \boldsymbol{\nabla} f(\boldsymbol{x}^{(k)})\|_2} &= \frac{\left(\boldsymbol{H}_k^{-1/2}\boldsymbol{y}^{(k)}\right)^T \boldsymbol{H}_k \left(\boldsymbol{H}_k^{-1/2}\boldsymbol{y}^{(k)}\right)}{\left\|\boldsymbol{H}_k^{-1/2}\boldsymbol{y}^{(k)}\right\|_2 \left\|\boldsymbol{H}_k \boldsymbol{H}_k^{-1/2}\boldsymbol{y}^{(k)}\right\|_2} \\
&= \frac{\boldsymbol{y}^{(k)^T}\boldsymbol{y}^{(k)}}{\left\|\boldsymbol{H}_k^{-1/2}\boldsymbol{y}^{(k)}\right\|_2 \left\|\boldsymbol{H}_k^{1/2}\boldsymbol{y}^{(k)}\right\|_2} \\
&= \frac{1}{\frac{\left\|\boldsymbol{H}_k^{-1/2}\boldsymbol{y}^{(k)}\right\|_2}{\|\boldsymbol{y}^{(k)}\|_2} \frac{\left\|\boldsymbol{H}_k^{1/2}\boldsymbol{y}^{(k)}\right\|_2}{\|\boldsymbol{y}^{(k)}\|_2}} \\
&\geqq \frac{1}{\|\boldsymbol{H}_k^{-1/2}\|_2 \|\boldsymbol{H}_k^{1/2}\|_2} \\
&= \frac{1}{\text{cond}_2(\boldsymbol{H}_k^{1/2})} \\
&= \frac{1}{\sqrt{\text{cond}_2(\boldsymbol{H}_k)}}.
\end{aligned}$$

Die erste Bedingung ist erfüllt, falls für alle Matrizen $\boldsymbol{H}_k$

$$\text{cond}_2(\boldsymbol{H}_k) = \frac{\lambda_{max}^{(k)}}{\lambda_{min}^{(k)}} \leqq \frac{1}{\gamma^2}$$

gilt. Weiterhin erhalten wir

$$\frac{\|\boldsymbol{H}_k \boldsymbol{\nabla} f(\boldsymbol{x}^{(k)})\|_2}{\|\boldsymbol{\nabla} f(\boldsymbol{x}^{(k)})\|_2} = \frac{\|\boldsymbol{H}_k \boldsymbol{H}_k^{-1} \boldsymbol{z}^{(k)}\|_2}{\|\boldsymbol{H}_k^{-1} \boldsymbol{z}^{(k)}\|_2} = \frac{\|\boldsymbol{z}^{(k)}\|_2}{\|\boldsymbol{H}_k^{-1} \boldsymbol{z}^{(k)}\|_2} \geqq \frac{1}{\|\boldsymbol{H}_k^{-1}\|_2}.$$

Die zweite Bedingung ist erfüllt, falls für alle Matrizen $\boldsymbol{H}_k$

$$\frac{1}{\|\boldsymbol{H}_k^{-1}\|_2} = \lambda_{min}^{(k)} \geqq \tau$$

gilt. (Dabei bezeichnen $\lambda_{max}^{(k)}$ und $\lambda_{min}^{(k)}$ den größten bzw. kleinsten Eigenwert der symmetrischen, positiv definiten Matrix $\boldsymbol{H}_k$.) Wir sehen, dass für Abstiegsrichtungen der Form $\boldsymbol{p}^{(k)} = -\boldsymbol{H}_k \boldsymbol{\nabla} f(\boldsymbol{x}^{(k)})$ die Voraussetzungen von Satz 11.19 erfüllt sind, falls die Folge $\{\boldsymbol{H}_k\}_{k\in\mathbb{N}}$ eine Folge symmetrischer, gleichmäßig positiv definiter und gleichmäßig beschränkter Matrizen ist. Unter diesen Bedingungen existieren Konstanten $0 < m \leqq M < \infty$, so dass für alle $k = 0, 1, \ldots$ und alle Vektoren $\boldsymbol{z} \in \mathbb{R}^n$

$$m\|\boldsymbol{z}\|_2^2 \leqq \boldsymbol{z}^T \boldsymbol{H}_k \boldsymbol{z} \leqq M\|\boldsymbol{z}\|_2^2$$

gilt.
Als einfachste Wahl bietet sich $\boldsymbol{H}_k = \boldsymbol{I}$ für alle $k = 0, 1, \ldots$ an. Hier ist $m = M = 1$. Das zugehörige Verfahren heißt **Gradientenverfahren**. Es ist im Sinne von Satz 11.19 global konvergent. Bis zum Jahre 1959 war dieses Verfahren auch das gebräuchlichste zum Lösen von freien Minimumproblemen. Vom Gradientenverfahren ist lokal höchstens lineare Konvergenz zu erwarten. Darum liegt es nahe, andere Folgen $\{\boldsymbol{H}_k\}_{k\in\mathbb{N}}$ gleichmäßig positiv definiter und gleichmäßig beschränkter Matrizen zu verwenden, die sich in der Nähe einer lokalen Lösung ähnlich wie die HESSE-Matrix verhalten. Entwickelt man den Gradienten an einer Stelle $\boldsymbol{x}$, so ergibt sich

$$\boldsymbol{\nabla} f(\boldsymbol{x}+\boldsymbol{s}) = \boldsymbol{\nabla} f(\boldsymbol{x}) + \boldsymbol{\nabla}^2 f(\boldsymbol{x})\boldsymbol{s} + O(\|\boldsymbol{s}\|^2).$$

Wir lassen nun $O(\|\boldsymbol{s}\|^2)$ weg, ersetzen $\boldsymbol{\nabla}^2 f(\boldsymbol{x})$ durch $\boldsymbol{H}^{-1}$ und erhalten

$$\boldsymbol{\nabla} f(\boldsymbol{x}+\boldsymbol{s}) = \boldsymbol{\nabla} f(\boldsymbol{x}) + \boldsymbol{H}^{-1} f(\boldsymbol{x})\boldsymbol{s}$$

oder

$$\boldsymbol{H}\left(\boldsymbol{\nabla} f(\boldsymbol{x}+\boldsymbol{s}) - \boldsymbol{\nabla} f(\boldsymbol{x})\right) = \boldsymbol{s}.$$

Ist nun

$$\boldsymbol{x} = \boldsymbol{x}^{(k)}, \quad \boldsymbol{x}+\boldsymbol{s} = \boldsymbol{x}^{(k+1)}, \quad \boldsymbol{H} = \boldsymbol{H}_k,$$

so erhalten wir die **Quasi-**NEWTON**-Gleichung**

$$\boldsymbol{H}_k\left(\boldsymbol{\nabla} f(\boldsymbol{x}^{(k+1)}) - \boldsymbol{\nabla} f(\boldsymbol{x}^{(k)})\right) = \boldsymbol{x}^{(k+1)} - \boldsymbol{x}^{(k)}.$$

Bestimmt man nun die Matrizen $\boldsymbol{H}_k$ so, dass sie nicht nur gleichmäßig positiv definit und gleichmäßig beschränkt sind, sondern dass sie auch in jedem Schritt die Quasi-NEWTON-Gleichung erfüllen, so ist neben der globalen Konvergenz auch lokal ein gutes Konvergenzverhalten zu erwarten. Diese Erwartung bestätigt sich mit Satz 11.27. Alle so erzeugten Verfahren bezeichnet man als **Quasi-**NEWTON**-Verfahren**. Die bekanntesten Vertreter sind das BFGS-Verfahren[2] und das DFP-Verfahren[3]. Aus verschiedenen Quasi-NEWTON-Verfahren wurde 1974 von OREN und LUENBERGER die folgende Klasse von Algorithmen definiert.

11.25. Verfahren der OREN-LUENBERGER**-Klasse:**

S0 *(Initialisierung) Wähle einen Startvektor* $\boldsymbol{x}^{(0)} \in \mathbb{R}^n$
und eine positiv definite Startmatrix $\boldsymbol{H}_0 \in \mathbb{R}^{n\times n}$ *(z.B.* $\boldsymbol{H}_0 = \boldsymbol{I}$*). Setze* $k = 0$.

[2] nach **B**ROYDEN, **F**LETCHER, **G**OLDFARB und **S**HANNO, 1970
[3] nach **D**AVIDON, **F**LETCHER und **P**OWELL, 1959 bzw. 1963

S1 *(Abbruchbedingung) Falls $\nabla f(\boldsymbol{x}^{(k)}) = \boldsymbol{o}$, so STOPP; $\boldsymbol{x}^{(k)}$ ist stationäre Lösung.*

S2 *(Richtungswahl) Berechne eine Abstiegsrichtung*

$$\boldsymbol{p}^{(k)} = -\boldsymbol{H}_k \nabla f(\boldsymbol{x}^{(k)}).$$

S3 *(Schrittweitenbestimmung) Bestimme eine Schrittweite α_k als exakte,* POWELL- *oder* ARMIJO-*Schrittweite.*

S4 *(Iterationsschritt) Setze*

$$\begin{aligned} \boldsymbol{x}^{(k+1)} &= \boldsymbol{x}^{(k)} + \alpha_k \boldsymbol{p}^{(k)}, \\ \boldsymbol{s}^{(k)} &= \alpha_k \boldsymbol{p}^{(k)} = \boldsymbol{x}^{(k+1)} - \boldsymbol{x}^{(k)}, \\ \boldsymbol{q}^{(k)} &= \nabla f(\boldsymbol{x}^{(k+1)}) - \nabla f(\boldsymbol{x}^{(k)}). \end{aligned}$$

Wähle Konstanten $\gamma_k > 0$ und $\Theta_k \geqq 0$ und berechne

$$\boldsymbol{H}_{k+1} = \Psi(\boldsymbol{H}_k, \boldsymbol{s}^{(k)}, \boldsymbol{q}^{(k)}; \gamma_k, \Theta_k)$$

wobei

$$\begin{aligned} \Psi(\boldsymbol{H}, \boldsymbol{s}, \boldsymbol{q}; \gamma, \Theta) = \gamma \boldsymbol{H} \;&+\; \left(1 + \gamma\Theta \frac{\boldsymbol{q}^T \boldsymbol{H} \boldsymbol{q}}{\boldsymbol{s}^T \boldsymbol{q}}\right) \frac{\boldsymbol{s}\boldsymbol{s}^T}{\boldsymbol{s}^T \boldsymbol{q}} \\ &-\; \gamma \frac{1-\Theta}{\boldsymbol{q}^T \boldsymbol{H} \boldsymbol{q}} \boldsymbol{H}\boldsymbol{q} \cdot \boldsymbol{q}^T \boldsymbol{H} \\ &-\; \frac{\gamma\Theta}{\boldsymbol{s}^T \boldsymbol{q}} (\boldsymbol{s} \cdot \boldsymbol{q}^T \boldsymbol{H} + \boldsymbol{H}\boldsymbol{q} \cdot \boldsymbol{s}^T). \end{aligned}$$

Eine Formel, wie sie in Schritt **S4** verwendet wird, bezeichnet man als Update-Formel. Die neue Matrix $\Psi(\boldsymbol{H})$ entsteht hier aus der alten Matrix $\boldsymbol{H}$ durch eine Rang-2-Modifikation. Es gilt $\operatorname{rg}(\Psi(\boldsymbol{H}) - \gamma \boldsymbol{H}) \leqq 2$.
Je nach Wahl der Parameter γ_k und Θ_k erhält man verschiedene Verfahren:

- $\gamma_k \equiv 1$ und $\Theta_k \equiv 0$. Wir erhalten das DFP-Verfahren

$$\Psi(\boldsymbol{H}, \boldsymbol{s}, \boldsymbol{q}) = \boldsymbol{H} + \frac{\boldsymbol{s}\boldsymbol{s}^T}{\boldsymbol{s}^T \boldsymbol{q}} - \frac{\boldsymbol{H}\boldsymbol{q} \cdot \boldsymbol{q}^T \boldsymbol{H}}{\boldsymbol{q}^T \boldsymbol{H} \boldsymbol{q}}.$$

- $\gamma_k \equiv 1$ und $\Theta_k \equiv 1$. Wir erhalten das BFGS-Verfahren

$$\Psi(\boldsymbol{H}, \boldsymbol{s}, \boldsymbol{q}) = \boldsymbol{H} + \left(1 + \frac{\boldsymbol{q}^T \boldsymbol{H} \boldsymbol{q}}{\boldsymbol{s}^T \boldsymbol{q}}\right) \frac{\boldsymbol{s}\boldsymbol{s}^T}{\boldsymbol{s}^T \boldsymbol{q}} - \frac{\boldsymbol{s} \cdot \boldsymbol{q}^T \boldsymbol{H} + \boldsymbol{H}\boldsymbol{q} \cdot \boldsymbol{s}^T}{\boldsymbol{s}^T \boldsymbol{q}}.$$

- $\gamma_k \equiv 1$ und

$$\Theta_k = \frac{\boldsymbol{s}^{(k)^T}\boldsymbol{q}^{(k)}}{\boldsymbol{s}^{(k)^T}\boldsymbol{q}^{(k)} - \boldsymbol{q}^{(k)^T}\boldsymbol{H}_k\boldsymbol{q}^{(k)}}.$$

Wir erhalten das symmetrische Rang-1-Verfahren von BROYDEN

$$\boldsymbol{H}_{k+1} = \boldsymbol{H}_k + \frac{\left(\boldsymbol{s}^{(k)} - \boldsymbol{H}_k\boldsymbol{q}^{(k)}\right)\left(\boldsymbol{s}^{(k)} - \boldsymbol{H}_k\boldsymbol{q}^{(k)}\right)^T}{\boldsymbol{q}^{(k)^T}\left(\boldsymbol{s}^{(k)} - \boldsymbol{H}_k\boldsymbol{q}^{(k)}\right)}.$$

(Hier ist $\Theta_k < 0$ möglich. $\boldsymbol{H}_{k+1}$ kann dann indefinit werden.)

Wir wollen nun zeigen, unter welchen Bedingungen durch die Update-Formel der OREN-LUENBERGER-Klasse Folgen von positiv definiten Matrizen erzeugt werden.

11.26. Satz: *Es sei $\boldsymbol{H}$ eine positiv definite Matrix. Für die Vektoren $\boldsymbol{q} \in \mathbb{R}^n$ und $\boldsymbol{s} \in \mathbb{R}^n$ gelte $\boldsymbol{s}^T\boldsymbol{q} > 0$. Dann ist die Matrix*

$$\begin{aligned}\boldsymbol{H}_+ &= \Psi(\boldsymbol{H}, \boldsymbol{s}, \boldsymbol{q}; \gamma, \Theta) \\ &= \gamma\boldsymbol{H} + \left(1 + \gamma\Theta\frac{\boldsymbol{q}^T\boldsymbol{H}\boldsymbol{q}}{\boldsymbol{s}^T\boldsymbol{q}}\right)\frac{\boldsymbol{s}\boldsymbol{s}^T}{\boldsymbol{s}^T\boldsymbol{q}} - \gamma\frac{1-\Theta}{\boldsymbol{q}^T\boldsymbol{H}\boldsymbol{q}}\boldsymbol{H}\boldsymbol{q}\cdot\boldsymbol{q}^T\boldsymbol{H} \\ &\quad - \frac{\gamma\Theta}{\boldsymbol{s}^T\boldsymbol{q}}\left(\boldsymbol{s}\cdot\boldsymbol{q}^T\boldsymbol{H} + \boldsymbol{H}\boldsymbol{q}\cdot\boldsymbol{s}^T\right)\end{aligned}$$

*für beliebige $\gamma > 0$ und $\Theta \geqq 0$ ebenfalls positiv definit. Außerdem erfüllt $\boldsymbol{H}_+$ die Quasi-*NEWTON*-Gleichung*

$$\boldsymbol{H}_+\boldsymbol{q} = \boldsymbol{s}.$$

Beweis: Da $\boldsymbol{H}$ positiv definit ist existiert eine Zerlegung $\boldsymbol{H} = \boldsymbol{L}\boldsymbol{L}^T$ mit einer unteren Dreiecksmatrix $\boldsymbol{L}$. Es sei $\boldsymbol{u} = \boldsymbol{L}^T\boldsymbol{y}$ mit einem beliebigen Vektor $\boldsymbol{y} \in \mathbb{R}^n$ und $\boldsymbol{y} \neq \boldsymbol{o}$.

Weiterhin sei $\boldsymbol{v} = \boldsymbol{L}^T\boldsymbol{q}$. Dann gilt

$$\begin{aligned}
\boldsymbol{y}^T\boldsymbol{H}_+\boldsymbol{y} &= \gamma\boldsymbol{y}^T\boldsymbol{H}\boldsymbol{y} + \left(1+\gamma\Theta\frac{\boldsymbol{q}^T\boldsymbol{H}\boldsymbol{q}}{\boldsymbol{s}^T\boldsymbol{q}}\right)\frac{\boldsymbol{y}^T\boldsymbol{s}\boldsymbol{s}^T\boldsymbol{y}}{\boldsymbol{s}^T\boldsymbol{q}} - \\
&\quad -\gamma\frac{1-\Theta}{\boldsymbol{q}^T\boldsymbol{H}\boldsymbol{q}}\boldsymbol{y}^T\boldsymbol{H}\boldsymbol{q}\cdot\boldsymbol{q}^T\boldsymbol{H}\boldsymbol{y} - \frac{\gamma\Theta}{\boldsymbol{s}^T\boldsymbol{q}}\left(\boldsymbol{y}^T\boldsymbol{s}\cdot\boldsymbol{q}^T\boldsymbol{H}\boldsymbol{y} + \boldsymbol{y}^T\boldsymbol{H}\boldsymbol{q}\cdot\boldsymbol{s}^T\boldsymbol{y}\right) \\
&= \gamma\boldsymbol{u}^T\boldsymbol{u} + \left(1+\gamma\Theta\frac{\boldsymbol{v}^T\boldsymbol{v}}{\boldsymbol{s}^T\boldsymbol{q}}\right)\frac{(\boldsymbol{s}^T\boldsymbol{y})^2}{\boldsymbol{s}^T\boldsymbol{q}} - \gamma\frac{1-\Theta}{\boldsymbol{v}^T\boldsymbol{v}}(\boldsymbol{u}^T\boldsymbol{v})^2 \\
&\quad -\frac{2\gamma\Theta}{\boldsymbol{s}^T\boldsymbol{q}}(\boldsymbol{y}^T\boldsymbol{s})(\boldsymbol{u}^T\boldsymbol{v}) \\
&= \gamma\left[\boldsymbol{u}^T\boldsymbol{u} - \frac{(\boldsymbol{u}^T\boldsymbol{v})^2}{\boldsymbol{v}^T\boldsymbol{v}}\right] + \frac{(\boldsymbol{s}^T\boldsymbol{y})^2}{\boldsymbol{s}^T\boldsymbol{q}} \\
&\quad +\gamma\Theta\left[\frac{(\boldsymbol{v}^T\boldsymbol{v})(\boldsymbol{s}^T\boldsymbol{y})^2}{(\boldsymbol{s}^T\boldsymbol{q})^2} + \frac{(\boldsymbol{u}^T\boldsymbol{v})^2}{\boldsymbol{v}^T\boldsymbol{v}} - 2\frac{(\boldsymbol{y}^T\boldsymbol{s})(\boldsymbol{u}^T\boldsymbol{v})}{\boldsymbol{s}^T\boldsymbol{q}}\right] \\
&= \gamma\frac{(\boldsymbol{u}^T\boldsymbol{u})(\boldsymbol{v}^T\boldsymbol{v}) - (\boldsymbol{u}^T\boldsymbol{v})^2}{\boldsymbol{v}^T\boldsymbol{v}} + \frac{(\boldsymbol{s}^T\boldsymbol{y})^2}{\boldsymbol{s}^T\boldsymbol{q}} + \gamma\Theta(\boldsymbol{v}^T\boldsymbol{v})\left[\frac{\boldsymbol{s}^T\boldsymbol{y}}{\boldsymbol{s}^T\boldsymbol{q}} - \frac{\boldsymbol{u}^T\boldsymbol{v}}{\boldsymbol{v}^T\boldsymbol{v}}\right]^2.
\end{aligned}$$

Alle drei Terme sind nichtnegativ. Damit gilt zumindest

$$\boldsymbol{y}^T\boldsymbol{H}_+\boldsymbol{y} \geqq 0.$$

Der erste Term ist nach der CAUCHY-SCHWARZschen Ungleichung genau dann Null, wenn die Vektoren $\boldsymbol{u}$ und $\boldsymbol{v}$ linear abhängig sind: $\boldsymbol{u} = \beta\boldsymbol{v}$. Dann ist aber wegen der Regularität von $\boldsymbol{L}$ auch $\boldsymbol{y} = \beta\boldsymbol{q}$. In diesem Falle folgt

$$\boldsymbol{y}^T\boldsymbol{H}_+\boldsymbol{y} = \beta^2(\boldsymbol{s}^T\boldsymbol{q}) + \gamma\Theta(\boldsymbol{v}^T\boldsymbol{v})[\beta-\beta] = \beta^2(\boldsymbol{s}^T\boldsymbol{q}) > 0.$$

Damit ist $\boldsymbol{H}_+$ positiv definit.
Die Gültigkeit der Quasi-NEWTON-Gleichung folgt durch einfaches Ausrechnen.

$$\begin{aligned}
\boldsymbol{H}_+\boldsymbol{q} &= \gamma\boldsymbol{H}\boldsymbol{q} + \left(1+\gamma\Theta\frac{\boldsymbol{q}^T\boldsymbol{H}\boldsymbol{q}}{\boldsymbol{s}^T\boldsymbol{q}}\right)\frac{\boldsymbol{s}\boldsymbol{s}^T\boldsymbol{q}}{\boldsymbol{s}^T\boldsymbol{q}} - \gamma\frac{1-\Theta}{\boldsymbol{q}^T\boldsymbol{H}\boldsymbol{q}}\boldsymbol{H}\boldsymbol{q}\cdot\boldsymbol{q}^T\boldsymbol{H}\boldsymbol{q} - \\
&\quad -\frac{\gamma\Theta}{\boldsymbol{s}^T\boldsymbol{q}}\left(\boldsymbol{s}\cdot\boldsymbol{q}^T\boldsymbol{H}\boldsymbol{q} + \boldsymbol{H}\boldsymbol{q}\cdot\boldsymbol{s}^T\boldsymbol{q}\right) \\
&= \boldsymbol{H}\boldsymbol{q}[\gamma - \gamma(1-\Theta) - \gamma\Theta] + \boldsymbol{s}\left[1+\gamma\Theta\frac{\boldsymbol{q}^T\boldsymbol{H}\boldsymbol{q}}{\boldsymbol{s}^T\boldsymbol{q}} - \gamma\Theta\frac{\boldsymbol{q}^T\boldsymbol{H}\boldsymbol{q}}{\boldsymbol{s}^T\boldsymbol{q}}\right] \\
&= \boldsymbol{s}
\end{aligned}$$

❀

Bemerkung: Die Bedingung $\boldsymbol{s}^T\boldsymbol{q} > 0$ ist eine Forderung an den Schrittweitenalgorithmus. Aus

$$\boldsymbol{s}^{(k)^T}\boldsymbol{q}^{(k)} = \alpha_k \boldsymbol{p}^{(k)^T}\left[\boldsymbol{\nabla} f(\boldsymbol{x}^{(k+1)}) - \boldsymbol{\nabla} f(\boldsymbol{x}^{(k)})\right] > 0$$

folgt

$$\boldsymbol{p}^{(k)^T}\boldsymbol{\nabla} f(\boldsymbol{x}^{(k+1)}) > \boldsymbol{p}^{(k)^T}\boldsymbol{\nabla} f(\boldsymbol{x}^{(k)}).$$

Für exakte Schrittweiten ist dann wegen

$$\boldsymbol{p}^{(k)^T}\boldsymbol{\nabla} f(\boldsymbol{x}^{(k+1)}) = 0, \quad \boldsymbol{p}^{(k)^T}\boldsymbol{\nabla} f(\boldsymbol{x}^{(k)}) < 0$$

diese Ungleichung immer erfüllt.
Für POWELL-Schrittweiten gilt wegen $\boldsymbol{p}^{(k)^T}\boldsymbol{\nabla} f(\boldsymbol{x}^{(k)}) < 0$ und $\nu < 1$ ebenfalls

$$\boldsymbol{p}^{(k)^T}\boldsymbol{\nabla} f(\boldsymbol{x}^{(k+1)}) \geqq \nu \boldsymbol{p}^{(k)^T}\boldsymbol{\nabla} f(\boldsymbol{x}^{(k)}) > \boldsymbol{p}^{(k)^T}\boldsymbol{\nabla} f(\boldsymbol{x}^{(k)}).$$

Für ARMIJO-Schrittweiten kann die Bedingung $\boldsymbol{s}^{(k)^T}\boldsymbol{q}^{(k)} > 0$ verletzt sein. Es können indefinite Matrizen entstehen.
Das Konvergenzverhalten von Verfahren der OREN-LUENBERGER-Klasse klärt der folgende Satz.

11.27. Satz: *Mit der positiv definiten Matrix $\boldsymbol{A}$, dem Vektor $\boldsymbol{b} \in \mathbb{R}^n$ und einem $c \in \mathbb{R}$ definieren wir die quadratische Funktion $h : \mathbb{R}^n \longrightarrow \mathbb{R}$ gemäß*

$$h(\boldsymbol{x}) = \frac{1}{2}\boldsymbol{x}^T\boldsymbol{A}\boldsymbol{x} + \boldsymbol{b}^T\boldsymbol{x} + c.$$

Wendet man ein Verfahren der OREN-LUENBERGER*-Klasse zum Lösen des Minimumproblems*

$$\min_{\boldsymbol{x}\in\mathbb{R}^n} h(\boldsymbol{x})$$

an, so gilt für einen beliebigen Startpunkt $\boldsymbol{x}^{(0)} \in \mathbb{R}^n$ und eine beliebige positiv definite Startmatrix $\boldsymbol{H}_0$ bei Anwendung exakter Schrittweiten:

1. *Es existiert ein kleinstes $m \leqq n$, so dass $\boldsymbol{x}^{(m)} = \bar{\boldsymbol{x}} = -\boldsymbol{A}^{-1}\boldsymbol{b}$ die exakte Minimumstelle von h mit $\boldsymbol{\nabla} h(\bar{\boldsymbol{x}}) = \boldsymbol{o}$ ist.*
2. *Für $0 \leqq i \neq k \leqq m-1$ ist*

   $$\boldsymbol{s}^{(i)^T}\boldsymbol{q}^{(k)} = \boldsymbol{s}^{(i)^T}\boldsymbol{A}\boldsymbol{s}^{(k)}$$

   und für $0 \leqq i \leqq m-1$ ist

   $$\boldsymbol{s}^{(i)^T}\boldsymbol{q}^{(i)} > 0.$$

3. *Für* $0 \leqq i < k \leqq m-1$ *ist*

$$\boldsymbol{s}^{(i)^T} \boldsymbol{\nabla} h(\boldsymbol{x}^{(k)}) = \boldsymbol{o}.$$

4. *Für* $0 \leqq i < k \leqq m-1$ *ist*

$$\boldsymbol{H}_k \boldsymbol{q}^{(i)} = \gamma_{ik} \boldsymbol{s}^{(i)}$$

mit

$$\gamma_{ik} = \begin{cases} \gamma_{i+1}\gamma_{i+2}\cdots\gamma_{k-1} & \textit{für} \quad i < k-1 \\ 1 & \textit{für} \quad i = k-1 \end{cases}.$$

5. *Ist* $m = n$*, so gilt zusätzlich*

$$\boldsymbol{H}_m = \boldsymbol{H}_n = \boldsymbol{S}\boldsymbol{D}\boldsymbol{S}^{-1}\boldsymbol{A}^{-1}$$

mit

$$\boldsymbol{D} = \operatorname{diag}(\gamma_{0n}, \ldots, \gamma_{n-1,n}), \quad \boldsymbol{S} = \left(\boldsymbol{s}^{(0)}, \ldots, \boldsymbol{s}^{(n-1)}\right).$$

Für $\gamma_i \equiv 1$ *folgt* $\boldsymbol{H}_n = \boldsymbol{A}^{-1}$.

Einen Beweis findet man in STOER „Numerische Mathematik“ Bd. 1.
Da sich jede hinreichend oft differenzierbare Funktion in der Nähe eines lokalen Minimums beliebig genau durch eine quadratische Funktion approximieren lässt, ist zu erwarten, dass die Quasi-NEWTON-Verfahren der OREN-LUENBERGER-Klasse auch bei der Anwendung auf beliebige Funktionen lokal ein gutes Konvergenzverhalten zeigen.
Es bleibt nun noch die Frage zu klären, wie die Parameter γ_k und Θ_k in jedem Schritt zu wählen sind. Praktische Erfahrungen zeigen, dass man mit dem BFGS-Verfahren ($\gamma_k \equiv 1$ und $\Theta_k \equiv 1$) gute Resultate erzielt. Andererseits folgte aus Satz 11.19, dass die Matrizen $\boldsymbol{H}_k$ gleichmäßig positiv definit und gleichmäßig beschränkt sind, damit globale Konvergenz eintritt. Man kann nun versuchen, über die Parameter γ_k und Θ_k günstige Eigenschaften der Matrizen $\boldsymbol{H}_k$ einzustellen.
Von OREN und SPEDICATO (1974) stammt das folgende Ergebnis. Es sei

$$\begin{aligned} \sigma_k &= \boldsymbol{s}^{(k)^T} \boldsymbol{H}_k^{-1} \boldsymbol{s}^{(k)}, \\ \varrho_k &= \boldsymbol{q}^{(k)^T} \boldsymbol{H}_k \boldsymbol{q}^{(k)}, \\ \tau_k &= \boldsymbol{s}^{(k)^T} \boldsymbol{q}^{(k)}. \end{aligned}$$

Man setze

$$\gamma_k = \frac{\sigma_k}{\tau_k}, \quad \Theta_k = 0 \quad \text{falls} \quad \frac{\sigma_k}{\tau_k} \leqq 1,$$

$$\gamma_k = \frac{\tau_k}{\varrho_k}, \quad \Theta_k = 1 \quad \text{falls} \quad \frac{\tau_k}{\varrho_k} \geqq 1$$

beziehungsweise

$$\gamma_k = 1, \quad \Theta_k = \tau_k(\sigma_k - \tau_k)(\sigma_k\varrho_k - \tau_k^2) \quad \text{falls} \quad \frac{\tau_k}{\varrho_k} \leqq 1 \leqq \frac{\sigma_k}{\tau_k}.$$

Durch diese Wahl wird eine obere Schranke für $\text{cond}(\boldsymbol{H}_{k+1})/\text{cond}(\boldsymbol{H}_k)$ in jedem Schritt minimiert.
Eine andere Variante stammt von DAVIDON (1975). Er betrachtete Verfahren mit $\gamma_k \equiv 1$ und wählt

$$\Theta_k = \tau_k(\sigma_k - \tau_k)(\sigma_k\varrho_k - \tau_k^2) \quad \text{falls} \quad \tau_k \leqq 2\frac{\sigma_k\varrho_k}{\sigma_k + \varrho_k}$$

beziehungsweise

$$\Theta_k = \frac{\tau_k}{\tau_k - \varrho_k} \quad \text{sonst.}$$

Durch diese Wahl wird der Quotient $\lambda_{max}/\lambda_{min}$ bezüglich des allgemeinen Eigenwertproblems

$$\text{Bestimme ein } \lambda \in \mathbb{C} \text{ und ein } \boldsymbol{y} \in \mathbb{C}^n \text{ mit } \boldsymbol{y} \neq \boldsymbol{o} \text{ und } \boldsymbol{H}_{k+1}\boldsymbol{y} = \lambda \boldsymbol{H}_k\boldsymbol{y}.$$

minimiert.
Wie wir zu Satz 11.26 bemerkten, kann bei Verwendung von ARMIJO-Schrittweiten aus einer positiv definiten Matrix $\boldsymbol{H}_k$ durch die update-Formel von OREN und LUENBERGER eine indefinite Matrix $\boldsymbol{H}_{k+1}$ entstehen. Die Sicherung der Bedingung

$$\boldsymbol{\nabla} f(\boldsymbol{x}^{(k+1)})^T \boldsymbol{p}^{(k)} \geqq \nu \boldsymbol{\nabla} f(\boldsymbol{x}^{(k)})^T \boldsymbol{p}^{(k)},$$

durch die man hinreichend genaue Schrittweiten bekommen würde, ist in der Praxis aber nur mit einem unvertretbar großen Aufwand möglich. Wir wollen nun eine andere Variante betrachten, mit der man automatisch die gleichmäßige Beschränktheit und gleichmäßige positive Definitheit der Matrizen $\boldsymbol{H}_k$ sichert. Die Abstiegsrichtungen werden in den bisher behandelten Verfahren in der Form

$$\text{(positiv definite Matrix } \boldsymbol{H}_k) \times \text{(negativer Gradient } -\boldsymbol{\nabla} f(\boldsymbol{x}^{(k)}))$$

bestimmt. Hätte man statt der Matrix $\boldsymbol{H}_k$ nur die Inverse $\boldsymbol{B}_k = \boldsymbol{H}_k^{-1}$, so ergäbe sich die Abstiegsrichtung $\boldsymbol{p}^{(k)}$ als Lösung des linearen Gleichungssystems

$$\boldsymbol{B}_k \boldsymbol{p}^{(k)} = -\boldsymbol{\nabla} f(\boldsymbol{x}^{(k)}).$$

Die Matrix $\boldsymbol{B}_k$ erfüllt die Quasi-NEWTON-Gleichung

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

Für die Matrizen $\boldsymbol{B}_k$ lässt sich ebenfalls eine zweiparametrige Update-Formel angeben. Man braucht dazu nur die Update-Formel für die $\boldsymbol{H}_k$ durch mehrmalige Anwendung der SHERMAN-MORRISON-Formel zu invertieren. Es ergibt sich

$$\boldsymbol{B}_{k+1} = \Psi(\boldsymbol{B}_k, \boldsymbol{q}_k, \boldsymbol{s}_k; \tilde{\gamma}_k, \tilde{\Theta}_k)$$

mit

$$\begin{aligned} \Psi(\boldsymbol{B}, \boldsymbol{q}, \boldsymbol{s}; \tilde{\gamma}, \tilde{\Theta}) = \tilde{\gamma}\boldsymbol{B} \;&+\; \left(1 + \tilde{\gamma}\tilde{\Theta}\frac{\boldsymbol{s}^T \boldsymbol{B} \boldsymbol{s}}{\boldsymbol{q}^T \boldsymbol{s}}\right) \frac{\boldsymbol{q}\boldsymbol{q}^T}{\boldsymbol{q}^T \boldsymbol{s}} \\ &-\; \tilde{\gamma}\frac{1-\tilde{\Theta}}{\boldsymbol{s}^T \boldsymbol{B} \boldsymbol{s}} \boldsymbol{B}\boldsymbol{s} \cdot \boldsymbol{s}^T \boldsymbol{B} \\ &-\; \frac{\tilde{\gamma}\tilde{\Theta}}{\boldsymbol{q}^T \boldsymbol{s}} (\boldsymbol{q} \cdot \boldsymbol{s}^T \boldsymbol{B} + \boldsymbol{B}\boldsymbol{s} \cdot \boldsymbol{q}^T). \end{aligned}$$

Der Zusammenhang zu den entsprechenden Update-Formeln für $\boldsymbol{H}$ ist durch

$$\tilde{\gamma} = \frac{1}{\gamma}, \quad \tilde{\Theta} = \frac{(1-\Theta)(\boldsymbol{s}^T \boldsymbol{q})^2}{(1-\Theta)(\boldsymbol{s}^T \boldsymbol{q})^2 + \Theta \boldsymbol{s}^T \boldsymbol{B} \boldsymbol{s} \cdot \boldsymbol{q}^T \boldsymbol{H} \boldsymbol{q}}$$

gegeben.
Wir erhalten

- aus dem DFP-Verfahren mit $\gamma_k \equiv 1$ und $\Theta_k \equiv 0$ $\tilde{\gamma}_k \equiv 1$ und $\tilde{\Theta}_k \equiv 1$, also

$$\boldsymbol{B}_{k+1} = \boldsymbol{B}_k + \left(1 + \frac{\boldsymbol{s}^{(k)^T} \boldsymbol{B}_k \boldsymbol{s}^{(k)}}{\boldsymbol{q}^{(k)^T} \boldsymbol{s}^{(k)}}\right) \frac{\boldsymbol{q}^{(k)} \boldsymbol{q}^{(k)^T}}{\boldsymbol{q}^{(k)^T} \boldsymbol{s}^{(k)}} - \frac{\boldsymbol{q}^{(k)} \cdot \boldsymbol{s}^{(k)^T} \boldsymbol{B}_k + \boldsymbol{B}_k \boldsymbol{s}^{(k)} \cdot \boldsymbol{q}^{(k)^T}}{\boldsymbol{q}^{(k)^T} \boldsymbol{s}^{(k)}},$$

- aus dem BFGS-Verfahren mit $\gamma_k \equiv 1$ und $\Theta_k \equiv 1$ $\tilde{\gamma}_k \equiv 1$ und $\tilde{\Theta}_k \equiv 0$, also

$$\boldsymbol{B}_{k+1} = \boldsymbol{B}_k + \frac{\boldsymbol{q}^{(k)} \boldsymbol{q}^{(k)^T}}{\boldsymbol{q}^{(k)^T} \boldsymbol{s}^{(k)}} - \frac{\boldsymbol{B}_k \boldsymbol{s}^{(k)} \cdot \boldsymbol{s}^{(k)^T} \boldsymbol{B}_k}{\boldsymbol{s}^{(k)^T} \boldsymbol{B}_k \boldsymbol{s}^{(k)}},$$

- aus dem symmetrischen Rang-1-Verfahren von BROYDEN mit $\gamma_k \equiv 1$ und

  $$\Theta_k = \frac{\boldsymbol{s}^{(k)^T}\boldsymbol{q}^{(k)}}{\boldsymbol{s}^{(k)^T}\boldsymbol{q}^{(k)} - \boldsymbol{q}^{(k)^T}\boldsymbol{H}_k\boldsymbol{q}^{(k)}}$$

  $\tilde{\gamma}_k \equiv 1$ und

  $$\tilde{\Theta}_k = \frac{\boldsymbol{s}^{(k)^T}\boldsymbol{q}^{(k)}}{\boldsymbol{s}^{(k)^T}\boldsymbol{q}^{(k)} - \boldsymbol{q}^{(k)^T}\boldsymbol{B}_k\boldsymbol{q}^{(k)}},$$

  also

  $$\boldsymbol{B}_{k+1} = \boldsymbol{B}_k + \frac{(\boldsymbol{q}^{(k)} - \boldsymbol{B}_k\boldsymbol{s}^{(k)})(\boldsymbol{q}^{(k)} - \boldsymbol{B}_k\boldsymbol{s}^{(k)})^T}{\boldsymbol{s}^{(k)^T}\boldsymbol{q}^{(k)} - \boldsymbol{q}^{(k)^T}\boldsymbol{B}_k\boldsymbol{q}^{(k)}}.$$

Die Formeln ähneln den Update-Formeln für $\boldsymbol{H}_k$. Auf den ersten Blick scheint es aber so, als ob durch diese Vorgehensweise der Rechenaufwand stark ansteigt. Für das Berechnen von

$$\boldsymbol{p}^{(k)} = -\boldsymbol{H}_k\boldsymbol{\nabla} f(\boldsymbol{x}^{(k)})$$

(Matrix$\times$Vektor) benötigt man rund n^2 Rechenoperationen, für das Lösen von

$$\boldsymbol{B}_k\boldsymbol{p}^{(k)} = -\boldsymbol{\nabla} f(\boldsymbol{x}^{(k)})$$

(lineares Gleichungssystem) dagegen rund $n^3/6$ Rechenoperationen.
Die Matrix $\boldsymbol{B}_{k+1}$ entsteht durch eine symmetrische Rang-2-Modifikation aus der Matrix $\tilde{\gamma}_k\boldsymbol{B}_k$, denn es gilt

$$\operatorname{rg}(\boldsymbol{B}_{k+1} - \tilde{\gamma}_k\boldsymbol{B}_k) \leqq 2.$$

Daher lässt sich der Aufwand zum Lösen der Gleichungssysteme beträchtlich verringern. Kennt man nämlich eine Zerlegung $\boldsymbol{B}_k = \boldsymbol{L}_k\boldsymbol{D}_k\boldsymbol{L}_k^T$, so lässt sich die entsprechende Zerlegung von $\boldsymbol{B}_{k+1}$ mit einem Aufwand der Größenordnung n^2 berechnen. Dann ist der Aufwand zum Berechnen einer Abstiegsrichtung mittels $\boldsymbol{B}_k$ von gleicher Größenordnung wie der Aufwand zum Berechnen einer Abstiegsrichtung mittels $\boldsymbol{H}_k$. Der große Vorteil der Verwendung einer LDL^T-Zerlegung der Matrix $\boldsymbol{B}_k$ besteht darin, dass aus der Zerlegung immer die positive Definitheit von $\boldsymbol{B}_k$ abzulesen ist. Denn $\boldsymbol{B} = \boldsymbol{L}\boldsymbol{D}\boldsymbol{L}^T$ ist genau dann positiv definit, wenn für die Diagonalmatrix $\boldsymbol{D} = \operatorname{diag}(d_1, \ldots, d_n)$ $d_i > 0$ für $i = 1, \ldots, n$ gilt. Ist diese Bedingung irgendwann verletzt, so lässt sich die positive Definitheit der Matrix $\boldsymbol{B}_{k+1}$ erzwingen, indem man die entsprechenden Elemente von $\boldsymbol{D}_{k+1}$ durch positive Zahlen ersetzt. Dann ist zwar die Quasi-NEWTON-Gleichung nicht mehr erfüllt; dies lässt sich

aber als Neustart des Verfahrens mit dem aktuellen Iterationspunkt als $\boldsymbol{x}^{(0)}$ und der korrigierten Matrix $\boldsymbol{B}_{k+1}$ als positiv definiter Startmatrix $\boldsymbol{B}_0$ interpretieren. Andererseits wird man von dem Verfahren verlangen, dass für hinreichend große k sich Schrittweiten $\alpha_k = 1$ einstellen. Dann lässt sich aber auch annehmen, dass die zweite Bedingung aus Satz 11.19 erfüllt und damit die positive Definitheit der Matrix $\boldsymbol{B}_{k+1}$ gesichert ist. Die Eigenschaft $\alpha_k = 1$ für alle $k \geqq k_0$ ist für jeden konkreten Algorithmus gesondert nachzuweisen.

11.3.3. Algorithmen zur Aufdatierung von Zerlegungen

Wir unterscheiden die Fälle $\bar{\boldsymbol{B}} = \boldsymbol{B} + \boldsymbol{v}\boldsymbol{v}^T$ oder $\bar{\boldsymbol{B}} = \boldsymbol{B} - \boldsymbol{v}\boldsymbol{v}^T$ und erhalten die folgenden beiden Algorithmen.

11.28. Aufdatierung einer LDL^T-Zerlegung $\bar{\boldsymbol{B}} = \boldsymbol{B} + \boldsymbol{v}\boldsymbol{v}^T$:
Von der positiv definiten (n,n)-Matrix $\boldsymbol{B}$ sei eine Zerlegung

$$\boldsymbol{B} = \boldsymbol{L}\boldsymbol{D}\boldsymbol{L}^T$$

mit einer unteren (n,n)-Einsdreiecksmatrix $\boldsymbol{L}$ und einer (n,n)-Diagonalmatrix $\boldsymbol{D}$ bekannt.
Es sind eine untere Einsdreiecksmatrix $\bar{\boldsymbol{L}}$ und eine Diagonalmatrix $\bar{\boldsymbol{D}}$ derart zu berechnen, dass

$$\bar{\boldsymbol{B}} = \boldsymbol{B} + \boldsymbol{v}\boldsymbol{v}^T = \bar{\boldsymbol{L}}\bar{\boldsymbol{D}}\bar{\boldsymbol{L}}^T$$

gilt.

$t_0 = 1$; $\boldsymbol{v}^{(1)} = \boldsymbol{v}$
for $j = 1$ ***to*** n ***do***
  $u_j = v_j^{(j)}$
  $t_j = t_{j-1} + u_j^2/d_j$
  $\bar{d}_j = d_j t_j / t_{j-1}$
  $\beta_j = u_j/(d_j t_j)$
  for $k = j+1$ ***to*** n ***do***
    $v_k^{(j+1)} = v_k^{(j)} - u_j l_{kj}$
    $\bar{l}_{kj} = l_{kj} + \beta_j v_k^{(j+1)}$
  endfor
endfor

11.29. Aufdatierung einer LDL^T-Zerlegung $\bar{\boldsymbol{B}} = \boldsymbol{B} - \boldsymbol{v}\boldsymbol{v}^T$:

Von der positiv definiten (n,n)-Matrix $\boldsymbol{B}$ sei eine Zerlegung $\boldsymbol{B} = \boldsymbol{L}\boldsymbol{D}\boldsymbol{L}^T$ mit einer unteren Einsdreiecksmatrix $\boldsymbol{L}$ und einer Diagonalmatrix $\boldsymbol{D}$ bekannt.
Es sind eine untere Einsdreiecksmatrix $\bar{\boldsymbol{L}}$ und eine Diagonalmatrix $\bar{\boldsymbol{D}}$ derart zu berechnen, dass

$$\bar{\boldsymbol{B}} = \boldsymbol{B} - \boldsymbol{v}\boldsymbol{v}^T = \bar{\boldsymbol{L}}\bar{\boldsymbol{D}}\bar{\boldsymbol{L}}^T$$

gilt.

Wähle ein $\delta > 0$
Löse das Gleichungssystem $\boldsymbol{L}\boldsymbol{u} = \boldsymbol{v}$
$t_{n+1} = 1 - \boldsymbol{u}^T\boldsymbol{D}^{-1}\boldsymbol{u}$
if $t_{n+1} < \delta$ ***then***
  $t_{n+1} = \delta$
endif
for $j = n$ ***to*** 1 ***step*** -1 ***do***
  $\bar{t_j} = t_{j+1} + u_j^2/d_j$
  $\bar{d_j} = d_j t_{j+1}/t_j$
  $\beta_j = -u_j/(d_j t_{j+1})$
  $v_j^{(j)} = u_j$
  for $k = j+1$ ***to*** n ***do***
    $v_k^{(j+1)} = v_k^{(j)} - u_j l_{kj}$
    $\bar{l}_{kj} = l_{kj} + \beta_j v_k^{(j+1)}$
  endfor
endfor

Beim ersten Algorithmus gilt wegen $t_j \geqq t_{j-1}$ $\bar{d_j} \geqq d_j$. Die positive Definitheit der Matrix $\bar{\boldsymbol{B}}$ wird gegenüber der Matrix $\boldsymbol{B}$ verbessert. Beim zweiten Algorithmus gilt wegen $\bar{t_j} \leqq t_{j-1}$ $\bar{d_j} \leqq d_j$. Da aber immer $\bar{t_j} \geqq \delta > 0$ ist die positive Definitheit der Matrix $\bar{\boldsymbol{B}}$ gesichert.
Falls man im Algorithmus die positive Definitheit zu erzwingen hat, ist die berechnete Zerlegung nur noch Zerlegung einer benachbarten Matrix $\tilde{\boldsymbol{B}} = \bar{\boldsymbol{B}} + \boldsymbol{\delta B}$. Im Algorithmus erkennt man, dass in die Berechnung der Elemente von $\bar{\boldsymbol{D}}$ und $\bar{\boldsymbol{L}}$ nur die Elemente von $\boldsymbol{D}$, $\boldsymbol{L}$ und $\boldsymbol{v}$ eingehen. Ändert man irgendwelche Elemente, speziell die Elemente von $\bar{\boldsymbol{D}}$, ab, so hat das keinen Einfluss auf die Berechnung der restlichen Elemente. Damit ist es unproblematisch, Beziehungen der Form

$$0 < \delta \leqq \min_{i=1,\ldots,n} d_i \leqq \max_{i=1,\ldots,n} d_i \leqq \Delta < \infty$$

und

$$\max_{i,j=1,\ldots,n} |l_{ij}| \leqq \Lambda < \infty$$

zu sichern. Hiermit sichert man aber auch die gleichmäßige Beschränktheit und gleichmäßige positive Definitheit der Matrizen $\boldsymbol{B}_k$.

11.4. Trust-Region-Verfahren

Dieser Verfahrensgruppe liegt eine andere Idee zugrunde als jene, die zu den Quasi-NEWTON-Verfahren und NEWTON-Verfahren führte. Wir ersetzen das Minimumproblem

$$\min_{\boldsymbol{x}\in\mathbb{R}^n} f(\boldsymbol{x})$$

lokal, d. h. in einer Kugelumgebung des aktuellen Iterationspunktes $\boldsymbol{x}^{(k)}$, durch ein einfacheres, leicht zu lösendes Modell. Wird mit der Lösung dieses vereinfachten Problems eine hinreichende Verminderung der Zielfunktion erreicht, so wird diese als neuer Iterationspunkt $\boldsymbol{x}^{(k+1)}$ akzeptiert. Erreicht man mit der Lösung des Ersatzproblems keine hinreichende Verminderung der Zielfunktion, so bedeutet das, dass man dem Modell auf einer zu großen Umgebung von $\boldsymbol{x}^{(k)}$ vertraut hat.[4] Man wird die Umgebung verkleinern und das Ersatzproblem erneut lösen. Je nach Wahl des Hilfsproblems und der Norm zur Definition der Kugelumgebungen von $\boldsymbol{x}^{(k)}$ erhält man die verschiedensten Trust-Region-Verfahren. Als vereinfachte Probleme bieten sich die folgenden an.

- $f \in C^1(\mathbb{R})$:

  $$f_{\boldsymbol{x}}(\boldsymbol{p}) = f(\boldsymbol{x}) + \boldsymbol{\nabla} f(\boldsymbol{x})^T \boldsymbol{p}.$$

- $f \in C^2(\mathbb{R})$:

  $$f_{\boldsymbol{x}}(\boldsymbol{p}) = f(\boldsymbol{x}) + \boldsymbol{\nabla} f(\boldsymbol{x})^T \boldsymbol{p} + \frac{1}{2}\boldsymbol{p}^T \boldsymbol{\nabla}^2 f(\boldsymbol{x})\boldsymbol{p}$$

  oder

  $$f_{\boldsymbol{x}}(\boldsymbol{p}) = f(\boldsymbol{x}) + \boldsymbol{\nabla} f(\boldsymbol{x})^T \boldsymbol{p} + \frac{1}{2}\boldsymbol{p}^T \boldsymbol{B}\boldsymbol{p}$$

  mit einer symmetrischen Matrix $\boldsymbol{B}$.

- $f(\boldsymbol{x}) = \|F(\boldsymbol{x})\|$ mit $F : \mathbb{R}^n \longrightarrow \mathbb{R}^m$:

  $$f_{\boldsymbol{x}}(\boldsymbol{p}) = \|F(\boldsymbol{x}) + F'(\boldsymbol{x})\boldsymbol{p}\|.$$

[4] Daher stammt auch der Name der Verfahren. Den englischen Begriff „trust region“ würde man mit „Vertrauensbereich“ übersetzen.

Es sei nun $f_{\boldsymbol{x}}(\boldsymbol{p})$ für jedes $\boldsymbol{x} \in \mathbb{R}^n$ gegeben. Eine einfache Version eines Trust-Region-Verfahrens könnte folgendermaßen aussehen.

11.30. Trust-region-Verfahren:

S0 *Wähle Konstanten* $0 < \varrho_1 < \varrho_2 < 1$, $0 < \sigma_1 < 1 < \sigma_2$ *und* $\Delta_0 > 0$.
Wähle einen Startpunkt $\boldsymbol{x}^{(0)} \in \mathbb{R}^n$ *und setze* $k = 0$.

S1 *Berechne die Lösung* $\boldsymbol{p}^*$ *der Aufgabe*

$$\min_{\|\boldsymbol{p}\| \leqq \Delta_k} f_{\boldsymbol{x}^{(k)}}(\boldsymbol{p}).$$

S2 *Falls* $f(\boldsymbol{x}^{(k)}) = f_{\boldsymbol{x}^{(k)}}(\boldsymbol{p}^*)$ *STOPP*
(Bei vernünftiger Wahl von $f_{\boldsymbol{x}}$ *ist dann* $\boldsymbol{x}^{(k)}$ *zumindest stationäre Lösung es ursprünglichen Minimumproblems.)*

S3 *Berechne*

$$r = \frac{f(\boldsymbol{x}^{(k)}) - f(\boldsymbol{x}^{(k)} + \boldsymbol{p}^*)}{f(\boldsymbol{x}^{(k)}) - f_{\boldsymbol{x}^{(k)}}(\boldsymbol{p}^*)}.$$

Falls $r \geqq \varrho_1$, *so* $\boldsymbol{x}^{(k+1)} = \boldsymbol{x}^{(k)} + \boldsymbol{p}^*$.
Falls $r < \varrho_1$, *wähle ein* $\Delta_{k+1} \in (0, \sigma_1 \Delta_k]$.
Falls $\varrho_1 \leqq r < \varrho_2$, *wähle ein* $\Delta_{k+1} \in [\sigma_1 \Delta_k, \Delta_k]$
Falls $\varrho_2 \leqq r$, *wähle ein* $\Delta_{k+1} \in [\Delta_k, \sigma_2 \Delta_k]$

S4 *Setze* $k = k + 1$ *und gehe zu Schritt* **S1**.

Bemerkung: Wird im Schritt **S2** nicht abgebrochen, so ist $f_{\boldsymbol{x}^{(k)}}(\boldsymbol{p}^*) < f(\boldsymbol{x}^{(k)})$. Die Größe r drückt damit aus, wie gut das Modell mit der tatsächlichen Funktion übereinstimmt. Je näher r an 1 liegt, desto genauer ist das Modell. Ist nun $r \geqq \varrho_1$, so stimmen die tatsächliche Veränderung der Zielfunktion und die durch das Modell vorausgesagte Verminderung der Zielfunktion hinreichend gut überein. Wir können $\boldsymbol{x}^{(k)} + \boldsymbol{p}^*$ als neuen Iterationspunkt akzeptieren. Ist sogar $r \geqq \varrho_2$, so stimmt das Modell so gut, dass wir im nächsten Schritt den Vertrauensbereich vergrößern dürfen. Ist dagegen $r < \varrho_1$, so war das Modell zu ungenau oder der Bereich, auf dem wir es angewendet haben, zu groß. Wir müssen den Vertrauensbereich verkleinern und das Modellproblem auf dem verkleinerten Bereich noch einmal lösen.

11.31. Beispiel: Für $f \in C^1(\mathbb{R})$ ergibt sich das einfachste Trust-region-Verfahren, falls man die Modellfunktion $f_{\boldsymbol{x}}(\boldsymbol{p}) = f(\boldsymbol{x}) + \boldsymbol{\nabla} f(\boldsymbol{x})^T \boldsymbol{p}$ und zur Festlegung der Kugelumgebung die euklidische Norm verwendet. Wir lösen damit in jedem Schritt das Hilfsproblem

$$\min_{\|\boldsymbol{p}\|_2 \leqq \Delta} f(\boldsymbol{x}) + \boldsymbol{\nabla} f(\boldsymbol{x})^T \boldsymbol{p}.$$

Ist $\boldsymbol{p}^*$ eine Lösung dieses Problems und gilt $f_{\boldsymbol{x}}(\boldsymbol{p}^*) = f(\boldsymbol{x})$, so ist $\boldsymbol{\nabla} f(\boldsymbol{x})^T \boldsymbol{p} \geqq 0$ für alle $\boldsymbol{p}$ mit $\|\boldsymbol{p}\|_2 \leqq \Delta$. Wählt man speziell $\boldsymbol{p} = -\alpha \boldsymbol{\nabla} f(\boldsymbol{x})$ mit einem $\alpha > 0$, so dass $\|\boldsymbol{p}\|_2 \leqq \Delta$, so folgt $\|\boldsymbol{\nabla} f(\boldsymbol{x})\|_2 \leqq 0$ und weiter $\boldsymbol{\nabla} f(\boldsymbol{x}) = \boldsymbol{o}$. Ist also für die Modellfunktion der Test im Schritt **S2** des Verfahrens erfüllt, so ist $\boldsymbol{x}$ stationäre Lösung des Minimumproblems. Für $\boldsymbol{\nabla} f(\boldsymbol{x}) \neq \boldsymbol{o}$ ergibt sich die Lösung des Hilfsproblems zu

$$\boldsymbol{p}^* = -\Delta \frac{\boldsymbol{\nabla} f(\boldsymbol{x})}{\|\boldsymbol{\nabla} f(\boldsymbol{x})\|_2}.$$

♡

Betrachten wir nun glatte Zielfunktionen $f \in C^2(\mathbb{R})$. Für derartige Probleme hatten wir das NEWTON-Verfahren angewendet. Das Verfahren hat aber den Nachteil, dass die NEWTON-Richtung

$$\boldsymbol{p} = -\left(\boldsymbol{\nabla}^2 f(\boldsymbol{x})\right)^{-1} \boldsymbol{\nabla} f(\boldsymbol{x})$$

nicht immer eine Abstiegsrichtung ist. Das Trust-Region-Verfahren ist auch im Falle einer indefiniten HESSE-Matrix noch durchführbar. Wir verwenden dazu die Ersatzfunktion

$$f_{\boldsymbol{x}}(\boldsymbol{p}) = f(\boldsymbol{x}) + \boldsymbol{\nabla} f(\boldsymbol{x})^T \boldsymbol{p} + \frac{1}{2} \boldsymbol{p}^T \boldsymbol{\nabla}^2 f(\boldsymbol{x}) \boldsymbol{p}.$$

Das Hilfsproblem

$$(P_{\boldsymbol{x},\Delta}) \qquad \min_{\|\boldsymbol{p}\|_2 \leqq \Delta} f_{\boldsymbol{x}}(\boldsymbol{p}) = \min_{\|\boldsymbol{p}\|_2 \leqq \Delta} \left(f(\boldsymbol{x}) + \boldsymbol{\nabla} f(\boldsymbol{x})^T \boldsymbol{p} + \frac{1}{2} \boldsymbol{p}^T \boldsymbol{\nabla}^2 f(\boldsymbol{x}) \boldsymbol{p} \right)$$

besitzt, da $f_{\boldsymbol{x}}$ stetig ist, stets eine globale Lösung $\boldsymbol{p}^*$. Falls $\boldsymbol{\nabla}^2 f(\boldsymbol{x})$ indefinit ist, ist $f_{\boldsymbol{x}}$ nicht mehr gleichmäßig konvex. In diesem Falle können auch zusätzliche lokale Lösungen auftreten.
Der folgende Satz gibt notwendige und hinreichende Bedingungen für ein globales Minimum des Hilfsproblems $(P_{\boldsymbol{x},\Delta})$ an.

11.32. Satz: *Wir betrachten die Aufgabe*

$$(P_\Delta) \qquad \min_{\|\boldsymbol{p}\|_2 \leqq \Delta} \varphi(\boldsymbol{p}), \quad \varphi(\boldsymbol{p}) = f + \boldsymbol{g}^T \boldsymbol{p} + \frac{1}{2} \boldsymbol{p}^T \boldsymbol{B} \boldsymbol{p},$$

wobei $\Delta > 0$, $f \in \mathbb{R}$, $\boldsymbol{g} \in \mathbb{R}^n$ und $\boldsymbol{B} \in \mathbb{R}^{n \times n}$ gelte. Die Matrix $\boldsymbol{B}$ sei darüber hinaus symmetrisch. Dann ist $\boldsymbol{p}^* \in \mathbb{R}^n$ mit $\|\boldsymbol{p}^*\|_2 \leqq \Delta$ genau dann eine globale Lösung von (P_Δ), wenn ein $\lambda^* \geqq 0$ existiert, so dass

1. $(\boldsymbol{B}+\lambda^*\boldsymbol{I})\boldsymbol{p}^* = -\boldsymbol{g}$,
2. $\lambda^*(\|\boldsymbol{p}^*\|_2 - \Delta) = 0$,
3. $\boldsymbol{B}+\lambda^*\boldsymbol{I}$ *ist positiv semidefinit*

gilt. Ist $\boldsymbol{B}+\lambda^*\boldsymbol{I}$ *sogar positiv definit, so ist* $\boldsymbol{p}^*$ *eindeutige globale Lösung von* (P_Δ). *Weiterhin ist* $\varphi(\boldsymbol{p}^*) = f$ *genau dann, wenn* $\boldsymbol{g} = \boldsymbol{o}$ *und* $\boldsymbol{B}$ *positiv semidefinit ist. Es gilt die Abschätzung*

$$f - \varphi(\boldsymbol{p}^*) \geqq \frac{1}{2}\|\boldsymbol{g}\|_2 \min\left\{\Delta, \frac{\|\boldsymbol{g}\|_2}{\|\boldsymbol{B}\|_2}\right\}.$$

Ein Trust-Region-Verfahren mit dem Hilfsproblem $(P_{\boldsymbol{x},\Delta})$ bricht nach Satz 11.32 genau dann in Schritt **S2** ab, wenn $\boldsymbol{\nabla} f(\boldsymbol{x}^{(k)}) = 0$ und $\boldsymbol{\nabla}^2 f(\boldsymbol{x}^{(k)})$ positiv semidefinit ist. Damit sind aber im Punkt $\boldsymbol{x}^{(k)}$ die Optimalitätsbedingungen zweiter Ordnung aus Satz 11.8 erfüllt.
Zum Abschluss wollen wir noch einen Konvergenzsatz für die Trust-Region-Variante des NEWTON-Verfahrens angeben.

11.33. Satz: *Gegeben sei das freie Minimumproblem*

$$(P) \quad \min_{\boldsymbol{x}\in\mathbb{R}^n} f(\boldsymbol{x})$$

und ein Startpunkt $\boldsymbol{x}^{(0)} \in \mathbb{R}^n$. *Die Niveaumenge*

$$L_0 = \left\{ \boldsymbol{x} \in \mathbb{R}^n \mid f(\boldsymbol{x}) \leqq f(\boldsymbol{x}^{(0)}) \right\}$$

sei kompakt. Die Zielfunktion sei auf einer offenen Obermenge von L_0 *zweimal stetig differenzierbar und der Gradient von* f *sei auf* L_0 *lipschitzstetig. Wir betrachten ein Trust-Region-Verfahren mit dem Hilfsproblem*

$$(P_{\boldsymbol{x},\Delta}) \quad \min_{\|\boldsymbol{p}\|_2 \leqq \Delta} f_{\boldsymbol{x}}(\boldsymbol{p}) = \min_{\|\boldsymbol{p}\|_2 \leqq \Delta} \left(f(\boldsymbol{x}) + \boldsymbol{\nabla} f(\boldsymbol{x})^T\boldsymbol{p} + \frac{1}{2}\boldsymbol{p}^T\boldsymbol{\nabla}^2 f(\boldsymbol{x})\boldsymbol{p} \right).$$

Falls das Verfahren nicht vorzeitig abbricht, liefert es eine Folge $\{\boldsymbol{x}^{(k)}\}_{k\in\mathbb{N}}$ *mit*

1. *Jeder Häufungspunkt von* $\{\boldsymbol{x}^{(k)}\}_{k\in\mathbb{N}}$ *ist stationäre Lösung von* (P).
2. *Die Folge* $\{\boldsymbol{x}^{(k)}\}_{k\in\mathbb{N}}$ *besitzt mindestens einen Häufungspunkt in dem die Optimalitätsbedingungen zweiter Ordnung erfüllt sind.*
3. *Ist der Vektor* $\boldsymbol{x}^*$ *Häufungspunkt von* $\{\boldsymbol{x}^{(k)}\}_{k\in\mathbb{N}}$ *mit positiv definiter* HESSE-*Matrix* $\boldsymbol{\nabla}^2 f(\boldsymbol{x}^*)$, *so konvergiert die Folge* $\{\boldsymbol{x}^{(k)}\}_{k\in\mathbb{N}}$ *gegen* $\boldsymbol{x}^*$. *Ist darüber hinaus* $\boldsymbol{\nabla}^2 f(\circ)$ *auf einer Kugelumgebung von* $\boldsymbol{x}^*$ *lipschitzstetig, so konvergiert die Folge* $\{\boldsymbol{x}^{(k)}\}_{k\in\mathbb{N}}$ *mindestens quadratisch.*

Index

Ähnlichkeitstransformation, 3
GERSCHGORIN-Kreis, 12
GIVENS-Rotation, 26

Abstiegsrichtung, 120
Ausgleichsproblem, 88
    lineares, 89

Eigenraum, 2
Eigenvektor, 1
Eigenwert, 1
Eigenwerte
    dominante, 34

Funktion
    gleichmäßig konvexe, 131
    richtungsdifferenzierbare, 122
    stetig differenzierbare, 120
Funktionalmatrix, 121

GATEAUX-Ableitung, 122
Gradient, 121
Gradientenverfahren, 156

HESSE-Matrix, 127

Lösung
    globale, 119
    lokale, 119
    stationäre, 126

Matrix
    ähnliche, 3
    defektive, 2
    diagonalähnliche, 3
    diagonalisierbar, 15
    pseudoinverse, 92
    spaltenreguläre, 89

Normalform
    JORDANsche, 10
Normalgleichung, 88

Polynom
    charakteristisches, 1
POWELL-Schrittweite, 139
Punkt
    stationärer, 126

Quasi-NEWTON-Gleichung, 156

Richtungsableitung, 122

Schrittweite, 120
    exakte, 138

Teilraum
    dominanter, 34
TSCHEBYSCHEFF-Problem
    diskretes, 88

Vielfachheit
    algebraische, 1
    geometrische, 2

Winkel, 40