Numerische Mathematik

Skript zur Vorlesung
im
Wintersemester 2008/9
und
Sommersemester 2009

Helmut Harbrecht

Stand: 23. Juli 2009

Vorwort

Diese Mitschrift kann und soll nicht ganz den Wortlaut der Vorlesung wiedergeben. Sie soll das Nacharbeiten des Inhalts der Vorlesung erleichtern. Das erste Kapitel habe ich dem Skript hinzugefügt, um nochmals die wichtigen Grundlagen der Numerischen Mathematik bereitzustellen, die schon in der Vorlesung *Algorithmische Mathematik* erarbeitet wurden.

Literatur zur Vorlesung:

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

— R. Schaback und H. Wendland: *Numerische Mathematik*, Springer-Verlag

— J. Stoer und R. Bulirsch: *Numerische Mathematik I+II*, Springer-Verlag

— D. Braess: *Finite Elemente: Theorie, schnelle Löser und Anwendungen in der Elastizitätstheorie*, Springer-Verlag

— W. Hackbusch: *Theorie und Numerik elliptischer Differentialgleichungen*, Teubner-Verlag

Inhaltsverzeichnis

1 Grundlagen **7**
1.1 Gleitkommazahlen . . . 7
1.2 Rundung . . . 8
1.3 Vektor- und Matrixnormen . . . 8
1.4 Numerische Kondition . . . 14
1.5 LR-Zerlegung . . . 18
1.6 Cholesky-Zerlegung . . . 26

2 Lineare Ausgleichsprobleme **30**
2.1 Normalengleichungen . . . 30
2.2 QR-Zerlegung . . . 33
2.3 Methode der Orthogonalisierung . . . 40
2.4 Singulärwertzerlegung und Pseudoinverse . . . 42
2.5 CG- und CGLS-Verfahren . . . 47

3 Nichtlineare Ausgleichsprobleme **54**
3.1 Gradientenverfahren . . . 54
3.2 Gauß-Newton-Verfahren . . . 56
3.3 Levenberg-Marquardt-Verfahren . . . 59

4 Eigenwerte **65**
4.1 Eigenwerteinschließungen . . . 65
4.2 Kondition des Eigenwertproblems . . . 68
4.3 Potenzmethode . . . 72
4.4 QR-Verfahren . . . 77
4.5 Implementierung des QR-Verfahrens . . . 79
4.6 Lanczos-Verfahren . . . 83

5 Partielle Differentialgleichungen **88**
5.1 Beispiele . . . 88
5.2 Charakterisierung . . . 90
5.3 Maximum-Prinzip . . . 92
5.4 Poisson-Gleichung . . . 94
5.5 Beliebige Differentialoperatoren . . . 98
5.6 Diskretes Maximumprinzip . . . 99
5.7 Konvergenz . . . 101
5.8 Iterative Lösung . . . 103
5.9 Vorkonditionierung . . . 107

5.10 Mehrgitterverfahren 110

6 Orthogonalpolynome 115
6.1 Innenproduktträume, Orthogonalbasen und Gramsche Matrizen 115
6.2 Tschebyscheff-Polynome 122
6.3 Allgemeine Orthogonalpolynome 125
6.4 Nullstellen von Orthogonalpolynomen 130
6.5 Erneut das Lanczos-Verfahren 133
6.6 Bisektionsverfahren 135

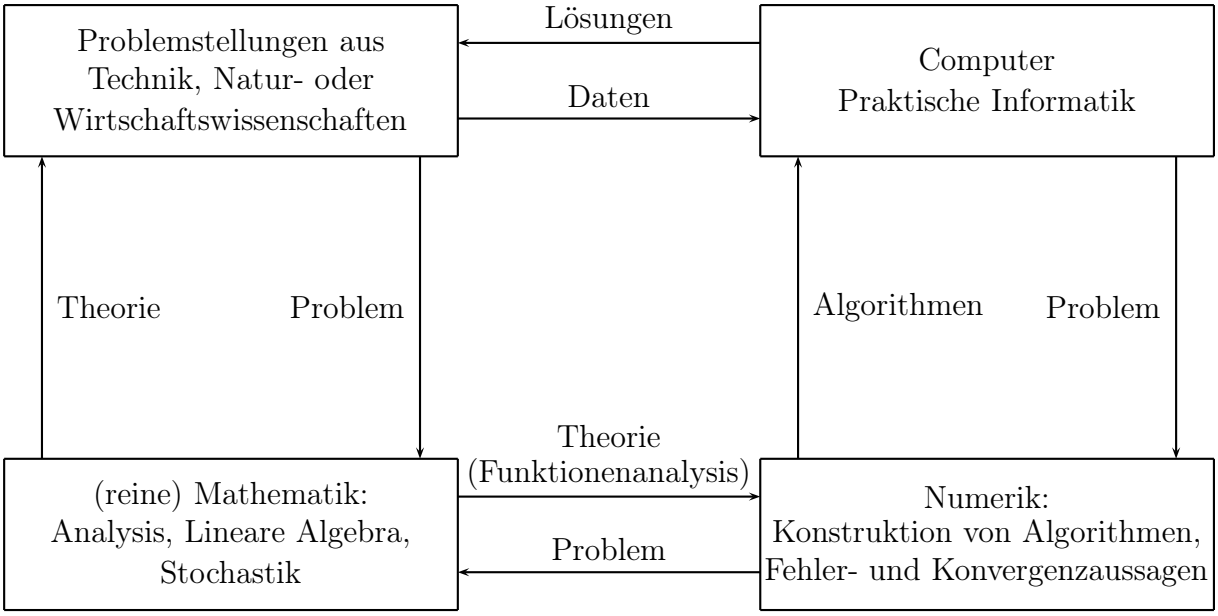
7 Numerische Quadratur 138
7.1 Einführung 138
7.2 Lagrange-Interpolation 141
7.3 Newton-Cotes-Formeln 143
7.4 Zusammengesetzte Quadraturformeln 144
7.5 Adaptive Quadratur 146
7.6 Euler-Maclaurinsche Summenformel 148
7.7 Romberg-Verfahren 150
7.8 Quadratur periodischer Funktionen 153
7.9 Gauß-Quadratur 154
7.10 *hp*-Quadratur 158
7.11 Tensorprodukt-Quadratur 162
7.12 Monte-Carlo-Quadratur 163
7.13 Dünngitter-Quadratur 166

8 Gewöhnliche Differentialgleichungen 170
8.1 Einleitung 170
8.2 Theoretische Grundlagen 172
8.3 Einschrittverfahren 175
8.4 Konsistenz 176
8.5 Konvergenz von Einschrittverfahren 178
8.6 Runge-Kutta-Verfahren 179
8.7 Mehrschrittverfahren 185
8.8 Adams-Bashforth-Verfahren 187
8.9 Adams-Moulton-Verfahren 188
8.10 Stabilität 188
8.11 Konvergenz von Mehrschrittverfahren 191
8.12 Steife Differentialgleichungen 195
8.13 Parabolische Differentialgleichungen 198

Einführung

Die Aufgabe der Numerik ist die Konstruktion und Analyse von Algorithmen zur Lösung mathematischer Aufgaben. Diese mathematischen Aufgaben stammen z.B. aus Technik, Naturwissenschaften, Wirtschaftswissenschaften und Sozialwissenschaften. Mathematische Methoden sind häufig auf ein spezielles Anwendungsgebiet zugeschnitten. Sobald Zahlenwerte erlaubt sind, treten jedoch überall ähnliche Probleme auf. Beispielsweise treten in 70% aller Anwendungen lineare Gleichungssysteme auf.

Beziehung der Numerik zu anderen Bereichen:



1. Grundlagen

1.1 Gleitkommazahlen

Bekanntlich lässt sich jede Zahl $x \in \mathbb{R}$ in normalisierter Dezimaldarstellung beschreiben:

$$x \pm \underbrace{a}_{\text{Mantisse}} \cdot 10 \overbrace{e}^{\text{Exponent}}, \qquad 0.1 \leq a < 1, \qquad e \in \mathbb{Z}.$$

Allgemein gibt es für eine Basis $b \in \mathbb{N}$, $b > 1$, eine eindeutige Schreibweise

$$x = \pm \left(\sum_{j=1}^{\infty} d_j \cdot b^{-j} \right) \cdot b^e, \qquad e = \pm \sum_{j=0}^{m} c_j \cdot b^j \tag{1.1}$$

mit $c_j, d_j \in \{0, 1, \ldots, b-1\}$ und $d_1 \neq 0$. Da Computer nur endliche Mantissenlängen m und Exponentenlängen n besitzen, das heißt

$$x = \pm \left(\sum_{j=1}^{m} d_j \cdot b^{-j} \right) \cdot b^e, \qquad e = \pm \sum_{j=0}^{n-1} c_j \cdot b^j \tag{1.2}$$

ist die Menge $\mathcal{A}$ der Maschinenzahlen *endlich*. Insbesondere gilt

$$x \in \mathcal{A} \quad \Longrightarrow \quad |x| \in [x_{\min}, x_{\max}]$$

mit

$$x_{\min} := b^{-1} \cdot b^{-\sum_{j=0}^{n-1}(b-1)\cdot b^j},$$
$$x_{\max} := \left(\sum_{j=1}^{m} (b-1) \cdot b^{-j} \right) \cdot b^{\sum_{j=0}^{n-1}(b-1)\cdot b^j}.$$

Beispiel 1.1 Für $x = 123.75$ gilt im sechsstelligen dezimalen Gleitkomma-Zahlensystem $(b = 10, m = 6)$

$$0.123\,750 \cdot 10^3$$

während sich im 12-stelligen binären Gleitkomma-Zahlensystem $(b = 2, m = 12)$

$$0.111\,101\,111\,000 \cdot 2^{111}$$

ergibt. △

1.2 Rundung

Problem: Approximation von $x \in \mathbb{R}$ mit $x \notin \mathcal{A}$ durch $y \in \mathcal{A}$.
Diese Approximation heißt *Rundung*. Die Rundung ist eine Abbildung

$$\mathrm{rd} : \mathbb{R} \to \mathcal{A} \qquad x \mapsto y = \mathrm{rd}(x),$$

die folgende Eigenschaften erfüllen soll:

- $\forall a \in \mathcal{A} : \mathrm{rd}(a) = a$
- $|x - \mathrm{rd}(x)| = \min_{a \in \mathcal{A}} |x - a| \quad \rightsquigarrow$ Rundung zur nächstgelegenen Maschinenzahl

Realisierung von rd: Sei $x \in \mathbb{R}$ mit Darstellung (1.1). Gilt $|x| \in [x_{\min}, x_{\max}]$, so erhalten wir die Maschinenzahl (1.2) vermittels

$$\mathrm{rd}(x) = \begin{cases} \pm \left(\sum_{j=1}^{m} d_j \cdot b^{-j} \right) \cdot b^e & \text{falls } d_{m+1} < \frac{b}{2} \\ \pm \left(\sum_{j=1}^{m} d_j \cdot b^{-j} + b^{-m} \right) \cdot b^e & \text{falls } d_{m+1} \geq \frac{b}{2} \end{cases}$$

Bemerkung: Abschneiden der Mantisse verletzt die zweite Eigenschaft!

Gilt $|x| < x_{\min}$ oder $|x| > x_{\max}$, so ergibt sich ein Underflow ($\mathrm{rd}(x) = 0$) bzw. Overflow ($\mathrm{rd}(x) = \pm \inf$).
Für den relativen Rundungsfehler gilt

$$\frac{|x - \mathrm{rd}(x)|}{|x|} \leq \frac{\frac{b}{2} \cdot b^{-(m+1)} \cdot b^e}{\underbrace{b^{-1} \cdot b^e}_{\text{untere Schranke für die Mantisse}}} \leq \frac{b}{2} \cdot b^{-m} =: \mathrm{eps}$$

Die Zahl eps heißt *Maschinengenauigkeit* (round-off unit). Bei heutigen Computern gilt i.a. $\mathrm{eps} \approx 2.2 \cdot 10^{-16}$.

Beispiel 1.2 Für die Zahl $x = 0.2$ ergibt sich im Binärsystem exakt $0.\overline{0011}$. Im Falle $m = 6$ ergibt sich gerundet $0.110\,011 \cdot 2^{-10}$. Rückkonvertiert ins Dezimalsystem bedeutet dies $\frac{51}{256} = 0.199\,218\,75$. △

1.3 Vektor- und Matrixnormen

Im folgenden bezeichnet $\mathbb{R}^n$ ($\mathbb{C}^n$) den Raum der reellwertigen (komplexwertigen) Vektoren

$$\mathbf{x} = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix}, \qquad x_i \in \mathbb{R} \quad (\mathbb{C})$$

und $\mathbb{R}^{m \times n}$ ($\mathbb{C}^{m \times n}$) den Raum der reellwertigen (komplexen) Matrizen

$$\mathbf{A} = \begin{bmatrix} a_{1,1} & a_{1,2} & \dots & a_{1,n} \\ a_{2,1} & a_{2,2} & & a_{2,n} \\ \vdots & & \ddots & \vdots \\ a_{m,1} & a_{m,2} & \dots & a_{m,n} \end{bmatrix}, \qquad a_{i,j} \in \mathbb{R} \quad (\mathbb{C}).$$

Meist ist es unerheblich, ob der zugrundeliegende Körper reell oder komplex ist. In diesem Fall schreiben wir $\mathbb{K}$ statt $\mathbb{R}$ oder $\mathbb{C}$.

Definition 1.3 Sei $X = \mathbb{K}^n$ oder $X = \mathbb{K}^{m\times n}$. Eine Abbildung

$$\|\cdot\| : X \to \mathbb{R}_{\geq 0}$$

heißt **Norm** auf X, wenn gilt

1. $\|\mathbf{x}\| > 0 \quad \forall \mathbf{x} \in X \setminus \{\mathbf{0}\}$
2. $\|\alpha\mathbf{x}\| = |\alpha| \cdot \|\mathbf{x}\| \quad \forall \mathbf{x} \in X,\ \alpha \in \mathbb{K}$
3. $\|\mathbf{x}+\mathbf{y}\| \leq \|\mathbf{x}\| + \|\mathbf{y}\| \quad \forall \mathbf{x}, \mathbf{y} \in X$

Bemerkung: Wegen $\mathbf{x} = \mathbf{x} - \mathbf{0}$ kann $\|\mathbf{x}\|$ als Abstand von $\mathbf{x}$ zum Nullpunkt in X interpretiert werden. In der Tat hat $\text{dist}(\mathbf{x}, \mathbf{y}) := \|\mathbf{x}-\mathbf{y}\|$ die Eigenschaften einer *Distanz* von zwei Elementen. Der Begriff Distanz ist allerdings allgemeiner, und nicht nur auf (normierte) Vektorräume beschränkt. Insofern liefern Normen spezielle Distanzbegriffe.

Häufig verwendete Normen:

1. $X = \mathbb{K}^n$:

Betragssummennorm: $$\|\mathbf{x}\|_1 := \sum_{i=1}^{n} |x_i|$$

Euklidnorm: $$\|\mathbf{x}\|_2 := \sqrt{\sum_{i=1}^{n} |x_i|^2} = \sqrt{\mathbf{x}^\star \cdot \mathbf{x}}, \quad \mathbf{x}^\star = [\overline{x}_1, \overline{x}_2, ..., \overline{x}_n]$$

($\mathbf{x}^\star$ ist der zu $\mathbf{x}$ konjugiert komplexe Vektor)

Maximumnorm: $$\|\mathbf{x}\|_\infty := \max_{1\leq i\leq n} |x_i|$$

2. $X = \mathbb{K}^{m\times n}$:

Spaltensummennorm: $$\|\mathbf{A}\|_1 := \max_{1\leq j\leq n} \sum_{i=1}^{m} |a_{i,j}|$$

Zeilensummennorm: $$\|\mathbf{A}\|_\infty := \max_{1\leq i\leq m} \sum_{j=1}^{n} |a_{i,j}|$$

Frobeniusnorm: $$\|\mathbf{A}\|_F := \sqrt{\sum_{i,j} |a_{i,j}|^2}$$

Beispiel 1.4 Für

$$\mathbf{A} = \begin{bmatrix} 2 & -3 \\ 1 & 1 \end{bmatrix}$$

gilt

$$\|\mathbf{A}\|_1 = 4, \qquad \|\mathbf{A}\|_\infty = 5, \qquad \|\mathbf{A}\|_F = \sqrt{15}.$$

△

Satz 1.5 Alle Normen auf $\mathbb{K}^n$ sind äquivalent, das heißt für zwei Normen $\|\cdot\|_a$ und $\|\cdot\|_b$ auf $\mathbb{K}^n$ gibt es positive Konstanten $c, C > 0$ mit

$$c\|\mathbf{x}\|_a \leq \|\mathbf{x}\|_b \leq C\|\mathbf{x}\|_a \quad \forall \mathbf{x} \in \mathbb{K}^n.$$

Beweis. Es genügt die Behauptung für $\|\cdot\|_a = \|\cdot\|_\infty$ zu zeigen. Dazu seien $\mathbf{x}, \mathbf{y} \in \mathbb{K}^n$ beliebig und $\|\cdot\|$ eine Norm im $\mathbb{K}^n$. Wegen

$$\mathbf{x} - \mathbf{y} = \sum_{i=1}^{m} (x_i - y_i)\mathbf{e}_i$$

folgt

$$\big|\|\mathbf{x}\| - \|\mathbf{y}\|\big| \leq \|\mathbf{x} - \mathbf{y}\| \leq \sum_{i=1}^{m} |x_i - y_i| \|\mathbf{e}_i\| \leq \|\mathbf{x} - \mathbf{y}\|_\infty \sum_{i=1}^{m} \|\mathbf{e}_i\|.$$

Folglich ist $\|\cdot\| : \mathbb{K}^n \to \mathbb{R}$ eine Lipschitz-stetige Funktion mit Lipschitz-Konstante $L := \sum_{i=1}^{m} \|\mathbf{e}_i\|$. Als solche nimmt $\|\cdot\|$ auf der kompakten Einheitssphäre $\{\mathbf{x} \in \mathbb{K}^n : \|\mathbf{x}\|_\infty = 1\}$ sowohl ihr Maximum C als auch ihr Minimum c an. Wegen der ersten Normeigenschaft aus Definition 1.3 ist insbesondere $c > 0$. Daher folgt für beliebiges $\mathbf{z} \in \mathbb{K}^n$, dass

$$c \leq \left\| \frac{\mathbf{z}}{\|\mathbf{z}\|_\infty} \right\| \leq C,$$

beziehungsweise

$$c\|\mathbf{z}\|_\infty \leq \|\mathbf{z}\| \leq C\|\mathbf{z}\|_\infty.$$

□

Beispiel 1.6 Für $\mathbf{x} \in \mathbb{K}^n$ folgt aus

$$\max_{1 \leq i \leq n} |x_i|^2 \leq \sum_{k=1}^{n} \underbrace{|x_k|^2}_{\leq \max_{i=1}^{n} |x_i|^2} \leq n \cdot \max_{1 \leq i \leq n} |x_i|^2$$

sofort die Ungleichung

$$\|\mathbf{x}\|_\infty \leq \|\mathbf{x}\|_2 \leq \sqrt{n} \cdot \|\mathbf{x}\|_\infty.$$

△

Bemerkung: Satz 1.5 gilt auch für Matrizen, das heißt, im Falle des $\mathbb{K}^{m \times n}$.

Der Vektorraum $\mathbb{K}^{n \times n}$ unterscheidet sich von allen anderen genannten Räumen dadurch, dass eine weitere Operation definiert ist, nämlich die Multiplikation $\mathbf{A} \cdot \mathbf{B}$ mit $\mathbf{A}, \mathbf{B} \in \mathbb{K}^{n \times n}$.

Definition 1.7 Eine Matrixnorm $\|\cdot\|_M$ auf $\mathbb{K}^{n\times n}$ heißt **submultiplikativ**, falls gilt

$$\|\mathbf{A}\cdot\mathbf{B}\|_M \le \|\mathbf{A}\|_M \cdot \|\mathbf{B}\|_M \quad \forall \mathbf{A},\mathbf{B}\in\mathbb{K}^{n\times n}.$$

Eine Matrixnorm $\|\cdot\|_M$ auf $\mathbb{K}^{n\times n}$ heißt **verträglich** mit einer Vektornorm $\|\cdot\|_V$ auf $\mathbb{K}^n$, wenn gilt

$$\|\mathbf{A}\cdot\mathbf{x}\|_V \le \|\mathbf{A}\|_M \cdot \|\mathbf{x}\|_V \quad \forall \mathbf{A}\in\mathbb{K}^{n\times n},\ \mathbf{x}\in\mathbb{K}^n.$$

Beispiel 1.8

1. $\|\mathbf{A}\| := \max_{1\le i,j\le n}|a_{i,j}|$ ist eine Matrixnorm auf $\mathbb{K}^{n\times n}$, aber nicht submultiplikativ:

$$\mathbf{A} = \begin{bmatrix}1 & 1\\ 1 & 1\end{bmatrix}, \quad \|\mathbf{A}\| = 1,$$
$$\mathbf{A}^2 = \begin{bmatrix}2 & 2\\ 2 & 2\end{bmatrix}, \quad \|\mathbf{A}^2\| = 2 \neq 1 = \|\mathbf{A}\|^2.$$

2. Die Frobeniusnorm ist mit der Euklidnorm verträglich:

$$\underbrace{[\mathbf{Ax}]_i^2}_{i\text{-te Komponente von } \mathbf{Ax}} = \left(\sum_{j=1}^n a_{i,j}\cdot x_j\right)^2 \overset{\text{CSU}}{\le} \left(\sum_{j=1}^n |a_{i,j}|^2\right)\cdot\left(\sum_{j=1}^n |x_j|^2\right)$$
$$= \left(\sum_{j=1}^n |a_{i,j}|^2\right)\cdot\|\mathbf{x}\|_2^2$$
$$\Longrightarrow \quad \|\mathbf{Ax}\|_2^2 = \sum_{i=1}^n [\mathbf{Ax}]_i^2 \le \underbrace{\sum_{i=1}^n\left(\sum_{j=1}^n |a_{i,j}|^2\right)}_{=\|\mathbf{A}\|_F^2}\cdot\|\mathbf{x}\|_2^2 = \|\mathbf{A}\|_F^2\cdot\|\mathbf{x}\|_2^2.$$

△

Definition 1.9 Sei $\|\cdot\|_V$ eine Vektornorm auf $\mathbb{K}^n$. Dann ist

$$|||\mathbf{A}||| := \sup_{\mathbf{x}\neq\mathbf{0}}\frac{\|\mathbf{Ax}\|_V}{\|\mathbf{x}\|_V} = \max_{\|\mathbf{x}\|_V=1}\|\mathbf{Ax}\|_V$$

eine Norm auf $\mathbb{K}^{n\times n}$, die sogenannte **induzierte Norm** von $\|\cdot\|_V$. (Die Normeigenschaften sind trivial nachgerechnet.)

Lemma 1.10 Die von $\|\cdot\|_V$ induzierte Norm $|||\cdot|||$ ist submultiplikativ und ist mit der Ausgangsnorm verträglich. Ist $\|\cdot\|_M$ eine mit $\|\cdot\|_V$ verträgliche Norm, dann gilt

$$|||\mathbf{A}||| \le \|\mathbf{A}\|_M \quad \forall \mathbf{A}\in\mathbb{K}^{n\times n}.$$

Beweis. 1. Sei $m = n$ und $B \neq 0$, dann gilt

$$\begin{aligned}\|\mathbf{AB}\| &= \sup_{\mathbf{x}\neq\mathbf{0}} \frac{\|\mathbf{ABx}\|_V}{\|\mathbf{x}\|_V} = \sup_{\mathbf{Bx}\neq\mathbf{0}} \frac{\|\mathbf{ABx}\|_V}{\|\mathbf{x}\|_V} = \sup_{\mathbf{Bx}\neq\mathbf{0}} \left(\frac{\|\mathbf{ABx}\|_V}{\|\mathbf{Bx}\|_V} \cdot \frac{\|\mathbf{Bx}\|_V}{\|\mathbf{x}\|_V} \right) \\ &\leq \sup_{\mathbf{Bx}\neq\mathbf{0}} \frac{\|\mathbf{ABx}\|_V}{\|\mathbf{Bx}\|_V} \cdot \sup_{\mathbf{Bx}\neq\mathbf{0}} \frac{\|\mathbf{Bx}\|_V}{\|\mathbf{x}\|_V} \leq \sup_{\mathbf{y}\neq\mathbf{0}} \frac{\|\mathbf{Ay}\|_V}{\|\mathbf{y}\|_V} \cdot \sup_{\mathbf{x}\neq\mathbf{0}} \frac{\|\mathbf{Bx}\|_V}{\|\mathbf{x}\|_V} \\ &= \|\mathbf{A}\| \cdot \|\mathbf{B}\|.\end{aligned}$$

2. Aus

$$\|\mathbf{A}\| = \sup_{\mathbf{x}\neq\mathbf{0}} \frac{\|\mathbf{Ax}\|_V}{\|\mathbf{x}\|_V} \geq \frac{\|\mathbf{Ax}\|_V}{\|\mathbf{x}\|_V} \quad \forall \mathbf{x} \in \mathbb{K}^n \backslash \{\mathbf{0}\}$$

folgt

$$\|\mathbf{Ax}\|_V \leq \|\mathbf{A}\| \cdot \|\mathbf{x}\|_V \quad \forall \mathbf{x} \in \mathbb{K}^n \backslash \{\mathbf{0}\}.$$

Im Falle $\mathbf{x} = \mathbf{0}$ folgt sofort

$$0 = \|\mathbf{Ax}\|_V \leq \|\mathbf{A}\| \cdot \underbrace{\|\mathbf{x}\|_V}_{=0} = 0.$$

3. Die Behauptung ergibt sich aus

$$\|\mathbf{A}\| = \sup_{\mathbf{x}\neq\mathbf{0}} \frac{\|\mathbf{Ax}\|_V}{\|\mathbf{x}\|_V} \leq \sup_{\mathbf{x}\neq\mathbf{0}} \frac{\|\mathbf{A}\|_M \cdot \|\mathbf{x}\|_V}{\|\mathbf{x}\|_V} = \|\mathbf{A}\|_M.$$

□

Bemerkung: Die Spaltensummennorm ist von der Betragssummennorm induziert, die Zeilensummennorm ist von der Maximumsnorm induziert.

Beachte: Verallgemeinerungen von Definitionen 1.7 und 1.9 auf Matrixnormen im $\mathbb{K}^{m\times n}$ gelten entsprechend; in diesem Fall müssen dann Vektornormen sowohl für den $\mathbb{K}^m$ als auch den $\mathbb{K}^n$ spezifiziert werden.

Frage: Welche Matrixnorm wird durch die Euklidnorm induziert?
Um diese Frage zu beantworten, betrachten wir

$$\|\mathbf{A}\|_2 := \max_{\|\mathbf{x}\|_2=1} \|\mathbf{Ax}\|_2 = \max_{\|\mathbf{x}\|_2=1} \sqrt{(\mathbf{Ax})^\star(\mathbf{Ax})} = \max_{\|\mathbf{x}\|_2=1} \sqrt{\mathbf{x}^\star\mathbf{A}^\star\mathbf{Ax}}.$$

Satz 1.11 Es gilt

$$\|\mathbf{A}\|_2 = \sqrt{\lambda_{\max}(\mathbf{A}^\star\mathbf{A})} = \sqrt{\max\{\lambda : \lambda \text{ ist Eigenwert von } \mathbf{A}^\star\mathbf{A}\}}.$$

Beweis. $\mathbf{A}^\star\mathbf{A}$ ist hermitesch (d.h. $\mathbf{A}^\star = \mathbf{A}$) und positiv semidefinit (d.h. alle Eigenwerte ≥ 0). Also hat $\mathbf{A}^\star\mathbf{A}$ n nichtnegative Eigenwerte $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_n \geq 0$ und zugehörige, paarweise orthonormale Eigenvektoren $\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_n$. Jeder Vektor $\mathbf{x} \in \mathbb{K}^n$, $\|\mathbf{x}\|_2 = 1$, lässt sich entwickeln

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

woraus folgt

$$1 = \|\mathbf{x}\|_2^2 = \mathbf{x}^\star\mathbf{x} = \left(\sum_{i=1}^n \overline{\xi}_i \mathbf{v}_i^\star\right) \cdot \left(\sum_{j=1}^n \xi_j \mathbf{v}_j\right) = \sum_{i,j=1}^n \overline{\xi}_i \xi_j \cdot \underbrace{\mathbf{v}_i^\star \mathbf{v}_j}_{=\delta_{i,j}} = \sum_{i=1}^n |\xi_i|^2.$$

Einerseits ergibt sich nun

$$\begin{aligned}\mathbf{x}^\star\mathbf{A}^\star\mathbf{A}\mathbf{x} &= \left(\sum_{i=1}^n \overline{\xi}_i \mathbf{v}_i^\star\right) \cdot \underbrace{\mathbf{A}^\star\mathbf{A} \cdot \left(\sum_{j=1}^n \xi_j \mathbf{v}_j\right)}_{=\sum_{j=1}^n \xi_j \mathbf{A}^\star\mathbf{A}\mathbf{v}_j} = \sum_{i,j=1}^n \overline{\xi}_i \xi_j \lambda_j \cdot \underbrace{\mathbf{v}_i^\star \mathbf{v}_j}_{=\delta_{i,j}} \\ &= \sum_{i=1}^n |\xi_i|^2 \cdot \lambda_i \leq \lambda_1 \cdot \underbrace{\sum_{i=1}^n |\xi_i|^2}_{=1} = \lambda_1.\end{aligned}$$

Andererseits gilt aber auch

$$\max_{\|\mathbf{x}\|_2=1} \mathbf{x}^\star\mathbf{A}^\star\mathbf{A}\mathbf{x} \geq \mathbf{v}_1^\star\mathbf{A}^\star\mathbf{A}\mathbf{v}_1 = \lambda_1 \cdot \mathbf{v}_1^\star\mathbf{v}_1 = \lambda_1.$$

Dies bedeutet aber

$$\max_{\|\mathbf{x}\|_2=1} \mathbf{x}^\star\mathbf{A}^\star\mathbf{A}\mathbf{x} = \lambda_1,$$

woraus die Behauptung folgt. □

Bemerkung: Wegen Satz 1.11 nennt man die $\|\cdot\|_2$-Matrixnorm auch *Spektralnorm*.

Beispiel 1.12 (Fortsetzung von Beipiel (1.4)) Wir wollen für

$$\mathbf{A} = \begin{bmatrix} 2 & -3 \\ 1 & 1 \end{bmatrix}$$

die Spektralnorm berechnen. Die Eigenwerte der Matrix

$$\mathbf{A}^\star\mathbf{A} = \begin{bmatrix} 5 & -5 \\ 5 & 10 \end{bmatrix}$$

kann man mit Hilfe der Regel von Sarrus über

$$\det(\mathbf{A}^\star\mathbf{A} - \lambda\mathbf{I}) = \begin{vmatrix} 5-\lambda & -5 \\ 5 & 10-\lambda \end{vmatrix} = \underbrace{(5-\lambda)(10-\lambda)}_{\lambda^2-15\lambda+50} + 25 \stackrel{!}{=} 0$$

bestimmen. Es folgt

$$\lambda_{1/2} = \frac{15 \pm 5\sqrt{5}}{2}$$

und damit

$$\|\mathbf{A}\|_2 = \sqrt{\frac{1}{2} \cdot (15 + 5\sqrt{5})}.$$

△

1.4 Numerische Kondition

Eine Fehlerquelle bei Implementierung eines numerischen Algorithmus sind Daten- und Rundungsfehler. Rundungsfehler je Rechenschritt können zwar in der Regel vernachlässigt werden, problematisch ist aber unter Umständen die Fehlerfortpflanzung über mehrere Rechenschritte. Man spricht in diesem Zusammenhang von der Stabilität von Algorithmen.
Die (etwas idealisierte) Annahme bezüglich Rundungsfehler ist jene, nachdem der Rechner bei einer Grundoperation anstelle des exakten Ergebnisses die nächstgelegene Maschinenzahl ausgibt.

$$a \boxdot b = \mathrm{rd}(a \circ b) = (a \circ b) \cdot (1 + \varepsilon), \qquad |\varepsilon| \leq \mathrm{eps}\,.$$

Dabei bezeichne $\boxdot$ die Rechnerverknüpfung und $\circ$ die exakte Verknüpfung.
Im weiteren beschäftigen wir uns mit der *Kondition* eines Algorithmus.

Gegeben: Eingangsdaten $\mathbf{x} \in \mathbb{K}^n$ und Rechenvorschrift

$$F : \mathbb{K}^n \to \mathbb{K}, \qquad F : \mathbf{x} \mapsto \mathbf{y} = F(\mathbf{x}) \in \mathbb{K}.$$

Die Eingangsdaten seien behaftet mit dem Datenfehler $\Delta\mathbf{x} \in \mathbb{K}^n$, also $\widetilde{\mathbf{x}} = \mathbf{x} + \Delta\mathbf{x}$.

Frage: Wie wirkt sich dies auf das Ergebnis $\mathbf{y} = F(\mathbf{x})$ aus?
Die Größe

$$\begin{aligned}\Delta\mathbf{y} :=& F(\mathbf{x} + \Delta\mathbf{x}) - F(\mathbf{x}) \\ =& F(x_1 + \Delta x_1, x_2 + \Delta x_2, \ldots, x_n + \Delta x_n) - F(x_1, x_2, ..., x_n)\end{aligned}$$

ist der *fortgepflanzte Datenfehler*. Falls $F \in C^1(\mathbb{K})$, gilt nach dem Mittelwertsatz

$$\Delta\mathbf{y} = \sum_{i=1}^{n} \frac{\partial F}{\partial x_i}(\boldsymbol{\xi}) \cdot \Delta x_i$$

mit

$$\boldsymbol{\xi} = \begin{bmatrix} \xi_1 \\ \xi_2 \\ \vdots \\ \xi_n \end{bmatrix} \in [\mathbf{x}, \mathbf{x} + \Delta\mathbf{x}] =: [x_1, x_1 + \Delta x_1] \times [x_2, x_2 + \Delta x_2] \times \cdots \times [x_n, x_n + \Delta x_n].$$

Ist die Ableitung Lipschitz-stetig, das heißt

$$|F(\mathbf{x}) - F(\mathbf{y})| \leq L\|\mathbf{x} - \mathbf{y}\| \quad \forall \mathbf{x}, \mathbf{y} \in \mathbb{K}^n,$$

dann gilt sogar

$$\Delta\mathbf{y} = \sum_{i=1}^{n} \frac{\partial F}{\partial x_i}(\mathbf{x}) \cdot \Delta x_i + \varepsilon, \qquad |\varepsilon| \leq C\|\Delta\mathbf{x}\|^2. \tag{1.3}$$

Für eine vereinfachende Stabilitätsaussage genügt es, den ε-Term zu vernachlässigen. Dabei ist der relative Fehler üblicherweise von größerer Bedeutung. Aus (1.3) folgt für diesen

$$\frac{\Delta\mathbf{y}}{\mathbf{y}} \approx \sum_{i=1}^{n} \frac{\partial F}{\partial x_i}(\mathbf{x}) \frac{\Delta x_i}{F(\mathbf{x})} = \sum_{i=1}^{n} \frac{\partial F}{\partial x_i}(\mathbf{x}) \cdot \frac{x_i}{F(\mathbf{x})} \cdot \frac{\Delta x_i}{x_i}.$$

Definition 1.13 Der Vektor

$$\boldsymbol{\kappa}_{\text{abs}}(\mathbf{x}) = \left[\left| \frac{\partial F}{\partial x_i}(\mathbf{x}) \right| \right]_{i=1}^{n} = \left| \nabla F(\mathbf{x}) \right|$$

gibt die **absoluten Konditionszahlen**, der Vektor

$$\boldsymbol{\kappa}_{\text{rel}}(\mathbf{x}) = \left[\left| \frac{\partial F}{\partial x_i}(\mathbf{x}) \cdot \frac{x_i}{F(\mathbf{x})} \right| \right]_{i=1}^{n}$$

gibt die **relativen Konditionszahlen** eines Algorithmus an. Ein Algorithmus heißt **schlecht konditioniert**, falls seine Konditionszahlen deutlich größer als 1 sind, ansonsten heißt er **gut konditioniert**.

Eine mögliche Verallgemeinerung auf mehrdimensionale Funktionen $F : \mathbb{K}^n \to \mathbb{K}^m$ ist offensichtlich, nämlich komponentenweise. Besser ist sogar "normweise", vgl. (1.4).

Beispiel 1.14 Addition $F(x_1, x_2) = x_1 + x_2$: Für $i = 1, 2$ gilt

$$\left[\boldsymbol{\kappa}_{\text{rel}}(\mathbf{x})\right]_i = \left| \frac{\partial F(\mathbf{x})}{\partial x_i} \frac{x_i}{F(\mathbf{x})} \right| = \left| \frac{x_i}{F(\mathbf{x})} \right|.$$

Die relativen Konditionszahlen sind sehr groß, falls $|x_1 + x_2| \ll |x_i|$, also wenn $|x_1| \approx |x_2|$ und $x_1 \cdot x_2 < 0$. Für

$$x_1 = 1.000\,001, \qquad x_2 = -1$$

und

$$\Delta x_1 = 0.001, \qquad \Delta x_2 = 0.001$$

ergibt sich etwa

$$x_1 + x_2 = 0.000\,001$$

und

$$(x_1 + \Delta x_1) + (x_2 + \Delta x_2) = 0.002\,001.$$

Der absolute Fehler wird zwar nur um den Faktor 2 verstärkt, der relative Fehler jedoch um den Faktor 10^6. Man spricht in diesem Fall von *Auslöschung*. △

Bemerkung: Ein typisches Beispiel für Auslöschung ist die numerische Differentiation

$$f'(x) \approx \frac{f(x+h) - f(x)}{h}$$

mit $h \approx 0$.

Beispiel 1.15 Multiplikation $F(x) = ax$: Hier gilt

$$F'(x) = a,$$

das heißt die absolute Kondition ist schlecht, falls $|a| \gg 1$. In diesem Fall ergibt sich eine starke Fehlerverstärkung. △

Man muß jedoch beachten, daß sowohl eine Aufgabenstellung schlecht konditioniert sein kann (etwa numerische Differentiation), oder lediglich ein spezieller Algorithmus für ein ansonsten gut konditioniertes Problem. Man spricht hier von *Stabilität*.

Wir betrachten folgendes Beispiel: Löse das lineare Gleichungssystem $\mathbf{Ax} = \mathbf{b}$, mit $\mathbf{A} \in \mathbb{K}^{n\times n}$ nichtsingulär. Offensichtlich ist bei Eingangsfehler $\Delta\mathbf{b}$

$$\mathbf{x} = \mathbf{A}^{-1}\mathbf{b}, \qquad \mathbf{x} + \Delta\mathbf{x} = \mathbf{A}^{-1}(\mathbf{b} + \Delta\mathbf{b}) = \mathbf{A}^{-1}\mathbf{b} + \mathbf{A}^{-1}\Delta\mathbf{b},$$

das heißt

$$\Delta\mathbf{x} = \mathbf{A}^{-1}\Delta\mathbf{b}.$$

Für ein verträgliches Matrix/Vektornormpaar ergibt sich somit

$$\frac{\|\Delta\mathbf{x}\|}{\|\mathbf{x}\|} = \frac{\|\mathbf{A}^{-1}\Delta\mathbf{b}\|}{\|\mathbf{x}\|} \le \|\mathbf{A}^{-1}\|\frac{\|\Delta\mathbf{b}\|}{\|\mathbf{b}\|}\frac{\|\mathbf{Ax}\|}{\|\mathbf{x}\|} \le \|\mathbf{A}^{-1}\|\|\mathbf{A}\|\frac{\|\Delta\mathbf{b}\|}{\|\mathbf{b}\|}. \tag{1.4}$$

Definition 1.16 Der Faktor

$$\operatorname{cond}_M \mathbf{A} := \|\mathbf{A}^{-1}\|_M\|\mathbf{A}\|_M$$

wird mit der **Kondition** der Matrix $\mathbf{A}$ bzgl. der Matrixnorm $\|\cdot\|_M$ bezeichnet.

Wieder beschreibt die Kondition die relative Fehlerverstärkung in diesem Problem, diesmal allerdings normweise für den schlimmstmöglichen Fall. Ist die Matrixnorm $\|\cdot\|_M$ durch eine Vektornorm induziert, so kann man einfache Beispiele für $\mathbf{b}$ und $\Delta\mathbf{b}$ konstruieren, für die diese Fehlerverstärkung exakt ist ("=").

Beispiel 1.17 Sei

$$\mathbf{A} = \begin{bmatrix} 10^{-3} & 1 \\ 1 & 2 \end{bmatrix}, \qquad \mathbf{b} = \begin{bmatrix} 1 \\ 3 \end{bmatrix}.$$

Die Matrix $\mathbf{A}$ ist gut konditioniert, denn mit

$$\mathbf{A}^{-1} \approx \begin{bmatrix} -2.004 & 1.002 \\ 1.002 & -0.001 \end{bmatrix}$$

folgt $3 = \|\mathbf{A}\|_\infty \approx \|\mathbf{A}^{-1}\|_\infty$ und daher ist

$$\operatorname{cond}_\infty \mathbf{A} \approx 9.$$

Die Lösung von $\mathbf{Ax} = \mathbf{b}$ ist also gut konditioniert. Sie lautet

$$\mathbf{x} \approx \begin{bmatrix} 1.002\ldots \\ 0.9989\ldots \end{bmatrix}.$$

Mit dem Gauß-Algorithmus und dreistelliger Gleitkommaarithmetik ergibt sich bei kleiner Datenstörung

$$\left[\begin{array}{cc|c} 0.001 & 1 & 1.01 \\ 1 & 2 & 3.01 \end{array}\right] \begin{matrix} \\ -1000 \end{matrix} \longrightarrow \left[\begin{array}{cc|c} 0.001 & 1 & 1.01 \\ 0 & -998 & -1010 \end{array}\right]$$

$$\Longrightarrow \quad x_2 = 1010 \boxslash 998 = 1.01$$
$$x_1 = 1000 \boxdot (1.01 \boxminus 1.01) = 0$$

△

Der Grund: Das kleine (1,1)-Element bewirkt einen großen Faktor (nämlich -1000) und damit starke Fehlerverstärkung, das heißt Instabilität.

Die Diagonalelemente, die bei der Gauß-Elimination im i-ten Schritt an Position (i,i) auftreten, werden *Pivotelemente* genannt. Zur Stabilisierung der Gauß-Elimination vertauscht man nun vor jedem Eliminierungsschritt die i-te und die k-te Zeile derart, dass das Pivotelement am betragsgrößten ist ("Spaltenpivotsuche" oder "partial pivoting"). Am exakten Resultat ändert das nichts!

Beispiel 1.18 (Fortsetzung von Beipiel 1.17) In unserem Fall würde man also die beiden Zeilen vertauschen, da $1 > 0.001$:

$$\left[\begin{array}{cc|c} 1 & 2 & 3.01 \\ 0.001 & 1 & 1.01 \end{array}\right] \begin{array}{c} \\ -\frac{1}{1000} \\ \hookleftarrow \end{array} \longrightarrow \left[\begin{array}{cc|c} 1 & 2 & 3.01 \\ 0 & 0.998 & 1.01 \end{array}\right]$$

$$\Longrightarrow \quad x_2 = 1.01 \boxslash 0.998 = 1.01$$
$$x_1 = 3.01 \boxminus (2 \boxdot 1.01) = 3.01 \boxminus 2.02 = 0.99$$

△

Die Auswahl des Pivotelements hängt stark von der Skalierung des linearen Gleichungssystems ab. Beispielsweise könnte man auch einfach die erste Gleichung mit 10^4 multiplizieren und das (1,1)-Element als Pivot behalten. Das Ergebnis wäre dann wieder so verheerend wie vorher. Man wählt daher im i-ten Teilschritt das Element welches am betragsgrößten ist ("totale Pivotisierung" oder "total pivoting"):

Im i-ten Teilschritt sei

$$\mathbf{A}^{(i)} = \begin{bmatrix} * & \cdots & * & * & \cdots & * \\ & \ddots & \vdots & \vdots & & \vdots \\ & & * & * & \cdots & * \\ & & & a_{i,i}^{(i)} & \cdots & a_{i,n}^{(i)} \\ & \mathbf{0} & & a_{i+1,i}^{(i)} & \cdots & a_{i+1,n}^{(i)} \\ & & & \vdots & & \\ & & & a_{n,i}^{(i)} & \cdots & a_{n,n}^{(i)} \end{bmatrix}.$$

Das entsprechende Element $a_{\ell,m}^{(i)}$ mit

$$|a_{\ell,m}^{(i)}| = \max_{i \le j,k \le n} |a_{j,k}^{(i)}|$$

kann dadurch in Position (i,i) gebracht werden, indem man wie zuvor die Zeilen i und ℓ und zusätzlich noch die Spalten i und m vertauscht. Dies ist die *stabilste* Variante des Gauß-Algorithmus.

Definition 1.19 Wir nennen einen Algorithmus **stabil**, falls für das berechnete Ergebnis $\boxed{F}$ und das exakte Ergebnis F gilt

$$\left| \frac{\boxed{F}(\mathbf{x}) - F(\mathbf{x})}{F(\mathbf{x})} \right| \le C \|\boldsymbol{\kappa}_{\text{rel}}(\mathbf{x})\| \text{ eps},$$

wobei C nicht zu groß und unabhängig von $\mathbf{x}$ sein soll.

1.5 LR-Zerlegung

Erinnerung: Im i-ten Teilschritt ($1 \leq i \leq n-1$) des Gauß-Algorithmus geht man wie folgt vor:

$$\underbrace{\left[\begin{array}{cccccc|c} \star & \cdots & \star & \star & \cdots & \star & \star \\ & \ddots & \vdots & \vdots & & \vdots & \vdots \\ & & \star & \star & \cdots & \star & \star \\ & & & a_{i,i}^{(i)} & \cdots & a_{i,n}^{(i)} & b_i^{(i)} \\ & \mathbf{0} & & a_{i+1,i}^{(i)} & \cdots & a_{i+1,n}^{(i)} & b_{i+1}^{(i)} \\ & & & \vdots & & \vdots & \vdots \\ & & & a_{n,i}^{(i)} & \cdots & a_{n,n}^{(i)} & b_n^{(i)} \end{array}\right]}_{=\mathbf{A}_i \qquad\qquad\qquad\qquad \underbrace{\qquad}_{=\mathbf{b}_i}} \begin{array}{ccc} -\tau_{i+1}^{(i)} & \cdots & -\tau_n^{(i)} \\ \hookleftarrow & & \\ & & \hookleftarrow \end{array}$$

mit

$$\tau_j^{(i)} = \frac{a_{j,i}^{(i)}}{a_{i,i}^{(i)}}, \qquad i < j \leq n.$$

Der gemeinsame Nenner $a_{i,i}^{(i)}$ der Faktoren $\tau_j^{(i)}$ wird *Pivotelement* genannt. Der obige Elimininationsschritt kann in Matrixnotation wie folgt geschrieben werden:

$$\underbrace{\begin{bmatrix} 1 & & & & & & \\ & \ddots & & & & \mathbf{0} & \\ & & 1 & & & & \\ & & & 1 & & & \\ & & & -\tau_{i+1}^{(i)} & 1 & & \\ & \mathbf{0} & & \vdots & & \ddots & \\ & & & -\tau_n^{(i)} & & & 1 \end{bmatrix}}_{=\mathbf{L}_i} \left[\begin{array}{c|c} \mathbf{A}_i & \mathbf{b}_i \end{array}\right] = \left[\begin{array}{c|c} \mathbf{A}_{i+1} & \mathbf{b}_{i+1} \end{array}\right],$$

wobei

$$\mathbf{L}_i = \mathbf{I} - \underbrace{\begin{bmatrix} 0 \\ \vdots \\ 0 \\ \tau_{i+1}^{(i)} \\ \vdots \\ \tau_n^{(i)} \end{bmatrix}}_{=:\,\boldsymbol{l}_i} \underbrace{\begin{bmatrix} 0 & \cdots & 0 & \overbrace{1}^{\text{Stelle } i} & 0 & \cdots & 0 \end{bmatrix}}_{=:\,\mathbf{e}_i^\star}.$$

Mit $\mathbf{A}_1 = \mathbf{A}$ und $\mathbf{b}_1 = \mathbf{b}$ ergibt sich durch Auflösen der Rekursion

$$\mathbf{L}_{n-1}\mathbf{L}_{n-2}\cdots\mathbf{L}_1 \left[\begin{array}{c|c} \mathbf{A} & \mathbf{b} \end{array}\right] = \left[\begin{array}{c|c} \mathbf{A}_n & \mathbf{b}_n \end{array}\right] = \left[\begin{array}{c|c} \mathbf{R} & \mathbf{c} \end{array}\right] = \left[\begin{array}{ccc|c} \star & \cdots & \star & \star \\ & \ddots & \vdots & \vdots \\ \mathbf{0} & & \star & \star \end{array}\right].$$

Insbesondere gilt

$$\mathbf{L}_{n-1}\mathbf{L}_{n-2}\cdots\mathbf{L}_1\mathbf{A} = \mathbf{R},$$

das heißt die Faktorisierung

$$\mathbf{A} = \underbrace{\mathbf{L}_1^{-1}\mathbf{L}_2^{-1}\cdots\mathbf{L}_{n-1}^{-1}}_{=:\mathbf{L}}\mathbf{R} = \mathbf{L}\mathbf{R}.$$

Die inversen Matrizen $\mathbf{L}_i^{-1}$ sowie $\mathbf{L}$ lassen sich explizit angeben:

Lemma 1.20

1. Die Inverse von $\mathbf{L}_i = \mathbf{I} - \boldsymbol{l}_i\mathbf{e}_i^\star$ berechnet sich gemäß

$$\mathbf{L}_i^{-1} = \mathbf{I} + \boldsymbol{l}_i\mathbf{e}_i^\star = \begin{bmatrix} 1 & & & & & \\ & \ddots & & & \mathbf{0} & \\ & & 1 & & & \\ & & & 1 & & \\ & & & \tau_{i+1}^{(i)} & 1 & \\ & \mathbf{0} & & \vdots & & \ddots \\ & & & \tau_n^{(i)} & & 1 \end{bmatrix}.$$

2. Die Matrix $\mathbf{L}$ erfüllt

$$\mathbf{L} = \mathbf{I} + \boldsymbol{l}_1\mathbf{e}_1^\star + \boldsymbol{l}_2\mathbf{e}_2^\star + \cdots + \boldsymbol{l}_{n-1}\mathbf{e}_{n-1}^\star = \begin{bmatrix} 1 & & & \\ \tau_2^{(1)} & 1 & & \mathbf{0} \\ \tau_3^{(1)} & \tau_3^{(2)} & \ddots & \\ \vdots & \vdots & & 1 \\ \tau_n^{(1)} & \tau_n^{(2)} & \cdots & \tau_n^{(n-1)} \quad 1 \end{bmatrix}. \tag{1.5}$$

Beweis. Aufgrund der Nulleinträge in $\boldsymbol{l}_i$ und $\mathbf{e}_i$ ist $\mathbf{e}_i^\star\boldsymbol{l}_j = 0$ für $i \le j$. Daraus folgt

$$(\underbrace{\mathbf{I} - \boldsymbol{l}_i\mathbf{e}_i^\star}_{=\mathbf{L}_i})(\mathbf{I} + \boldsymbol{l}_i\mathbf{e}_i^\star) = \mathbf{I} - \boldsymbol{l}_i\mathbf{e}_i^\star + \boldsymbol{l}_i\mathbf{e}_i^\star - \boldsymbol{l}_i\underbrace{\mathbf{e}_i^\star\boldsymbol{l}_i}_{=0}\mathbf{e}_i^\star = \mathbf{I}.$$

Weiter ergibt sich induktiv aus

$$\mathbf{L}_1^{-1}\mathbf{L}_2^{-1}\cdots\mathbf{L}_i^{-1} = \mathbf{I} + \boldsymbol{l}_1\mathbf{e}_1^\star + \boldsymbol{l}_2\mathbf{e}_2^\star + \cdots + \boldsymbol{l}_i\mathbf{e}_i^\star$$

dass

$$\begin{aligned} \mathbf{L}_1^{-1}\mathbf{L}_2^{-1}\cdots\mathbf{L}_{i+1}^{-1} &= (\mathbf{I} + \boldsymbol{l}_1\mathbf{e}_1^\star + \boldsymbol{l}_2\mathbf{e}_2^\star + \cdots + \boldsymbol{l}_i\mathbf{e}_i^\star)\mathbf{L}_{i+1}^{-1} \\ &= (\mathbf{I} + \boldsymbol{l}_1\mathbf{e}_1^\star + \boldsymbol{l}_2\mathbf{e}_2^\star + \cdots + \boldsymbol{l}_i\mathbf{e}_i^\star)(\mathbf{I} + \boldsymbol{l}_{i+1}\mathbf{e}_{i+1}^\star) \\ &= \mathbf{I} + \boldsymbol{l}_1\mathbf{e}_1^\star + \boldsymbol{l}_2\mathbf{e}_2^\star + \cdots + \boldsymbol{l}_i\mathbf{e}_i^\star + \boldsymbol{l}_{i+1}\mathbf{e}_{i+1}^\star + \sum_{j=1}^{i}\boldsymbol{l}_j\underbrace{\mathbf{e}_j^\star\boldsymbol{l}_{i+1}}_{=0}\mathbf{e}_{i+1}^\star. \end{aligned}$$

□

Wird im Verlauf des Gauß-Algorithmus ein Pivotelement $a_{i,i}^{(i)}$ Null, dann bricht das Verfahren in dieser Form zusammen. Sind hingegen alle Pivotelemente für $i = 1, 2, \ldots, n$ von Null verschieden, so haben wir das folgende Resultat bewiesen.

Satz 1.21 Falls kein Pivotelement Null wird, bestimmt der Gauß-Algorithmus neben der Lösung $\mathbf{x}$ von $\mathbf{Ax} = \mathbf{b}$ eine LR-Zerlegung $\mathbf{A} = \mathbf{L} \cdot \mathbf{R}$ in eine linke untere und eine rechte obere Dreiecksmatrix. Die Matrix $\mathbf{L}$ ist dabei durch (1.5) gegeben.

Beispiel 1.22

$$\mathbf{A} = \begin{bmatrix} 1 & 4 & 7 \\ 2 & 5 & 8 \\ 3 & 6 & 10 \end{bmatrix} \begin{matrix} -2 \\ \hookleftarrow \\ \end{matrix} \begin{matrix} -3 \\ \\ \hookleftarrow \end{matrix} \longrightarrow \begin{bmatrix} 1 & 4 & 7 \\ 0 & -3 & -6 \\ 0 & -6 & -11 \end{bmatrix} \begin{matrix} -2 \\ \hookleftarrow \\ \end{matrix} \longrightarrow \begin{bmatrix} 1 & 4 & 7 \\ 0 & -3 & -6 \\ 0 & 0 & 1 \end{bmatrix},$$

das heißt

$$\mathbf{R} = \begin{bmatrix} 1 & 4 & 7 \\ 0 & -3 & -6 \\ 0 & 0 & 1 \end{bmatrix}, \qquad \mathbf{L} = \begin{bmatrix} 1 & 0 & 0 \\ 2 & 1 & 0 \\ 3 & 2 & 1 \end{bmatrix}.$$

△

Bemerkung: Bei der Realisierung der LR-Zerlegung am Computer überschreibt man die ursprünglichen Einträge $a_{i,j}^{(1)} = a_{i,j}$ der Matrix $\mathbf{A}$ mit den jeweils aktuellen Einträgen $a_{i,j}^{(i)}$. Die Matrix $\mathbf{L}$ lässt sich sukzessive in die nicht mehr benötigte untere Hälfte von $\mathbf{A}$ schreiben. Damit wird kein zusätzlicher Speicherplatz für die LR-Zerlegung gebraucht.

Die Lösung des linearen Gleichungssystems $\mathbf{Ax} = \mathbf{b}$ wird mit Hilfe der LR-Zerlegung wie folgt berechnet:

① zerlege $\mathbf{A} = \mathbf{LR}$ mit dem Gauß-Algorithmus

② löse $\mathbf{Ax} = \mathbf{LRx} = \mathbf{b}$ in zwei Schritten:

- löse $\mathbf{Ly} = \mathbf{b}$ durch Vorwärtssubstitution
- löse $\mathbf{Rx} = \mathbf{y}$ durch Rückwärtssubstitution

Aufwand:

① Im i-ten Teilschritt werden

$$(n - i + 1)(n - i) = (n - i)^2 + n - i$$

Multiplikationen benötigt, das sind insgesamt

$$\sum_{i=1}^{n-1} \left\{(n - i)^2 + n - i\right\} \overset{j:=n-i}{=} \sum_{j=1}^{n-1} (j^2 + j) = \frac{1}{3}n^3 + \mathcal{O}(n^2)$$

Multiplikationen.

② Hier werden

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

Multiplikationen benötigt.

Demnach werden also insgesamt zum Lösen eines linearen Gleichungssystems mit Hilfe der LR-Zerlegung $n^3/3+\mathcal{O}(n^2)$ Multiplikationen (und, wie man leicht nachrechnet, nochmals ebensoviele Additionen) benötigt. Der Speicherplatzbedarf ist dabei allerdings nur von der Ordnung $\mathcal{O}(n^2)$.

Beispiel 1.23 (Fortsetzung von Beipiel 1.22) Für $\mathbf{b} = [1,1,1]^T$ wollen wir das lineare Gleichungssystem $\mathbf{Ax} = \mathbf{b}$ lösen:

1. bestimme $\mathbf{y}$ mit $\mathbf{Ly} = \mathbf{b}$ durch Vorwärtssubstitution:

$$\left[\begin{array}{c|c} \mathbf{L} & \mathbf{b} \end{array}\right] = \left[\begin{array}{ccc|c} 1 & 0 & 0 & 1 \\ 2 & 1 & 0 & 1 \\ 3 & 2 & 1 & 1 \end{array}\right] \quad \Longrightarrow \quad \begin{array}{l} y_1 = 1 \\ y_2 = 1 - 2 = -1 \\ y_3 = 1 - 3 + 2 = 0 \end{array}$$

2. bestimme $\mathbf{x}$ mit $\mathbf{Rx} = \mathbf{y}$ durch Rückwärtssubstitution:

$$\left[\begin{array}{c|c} \mathbf{R} & \mathbf{y} \end{array}\right] = \left[\begin{array}{ccc|c} 1 & 4 & 7 & 1 \\ 0 & -3 & -6 & -1 \\ 0 & 0 & 1 & 0 \end{array}\right] \quad \Longrightarrow \quad \begin{array}{l} x_3 = 0 \\ x_2 = (-1+0)/(-3) = 1/3 \\ x_1 = 1 - 4/3 - 0 = -1/3 \end{array}$$

△

Bemerkung: Eine weitere Möglichkeit, das lineare Gleichungssystem $\mathbf{Ax} = \mathbf{b}$ zu lösen, bietet bekanntlich die Cramersche Regel. Danach lautet die i-te Komponente x_i der Lösung

$$x_i = \frac{\det \mathbf{A}_i}{\det \mathbf{A}},$$

wobei hier $\mathbf{A}_i \in \mathbb{R}^{n\times n}$ diejenige Matrix ist, die aus $\mathbf{A}$ entsteht, wenn man die i-te Spalte durch $\mathbf{b}$ ersetzt. Berechnet man die Determinante nach dem Laplaceschen Entwicklungssatz, so benötigt man i.a. $n!$ Operationen. Bei Verwendung eines Rechners mit 10^8 Gleitkommaoperationen pro Sekunde (100 Megaflops) ergäben sich dann die folgenden Rechenzeiten:

n	10	12	14	16	18	20
Rechenzeit	0.4 s	1 min	3.6 h	41 Tage	38 Jahre	16 000 Jahre

Mit Spaltenpivotsuche wird die Matrizenformulierung des Gauß-Algorithmus schwieriger. Werden im i-ten Teilschritt die i-te und die k-te Zeile der Matrix $\mathbf{A}_i$ $(i < k \leq n)$ ver-

tauscht, dann kann dies durch die zugehörige Permutationsmatrix

$$\mathbf{P}_i = \begin{bmatrix} 1 & & & & & & & & \\ & \ddots & & & & & & & \\ & & 1 & & & & & & \\ & & & 0 & & & 1 & & \\ & & & & 1 & & & & \\ & & & & & \ddots & & & \\ & & & & & & 1 & & \\ & & & 1 & & & & 0 & \\ & & & & & & & & 1 \\ & & & & & & & & & \ddots \\ & & & & & & & & & & 1 \end{bmatrix} \begin{matrix} \\ \\ \\ \leftarrow i\text{-te Zeile} \\ \\ \\ \\ \leftarrow k\text{-te Zeile} \\ \\ \\ \\ \end{matrix} \tag{1.6}$$

$$\begin{matrix} & & \uparrow & & \uparrow & \\ & & i\text{-te Spalte} & & k\text{-te Spalte} & \end{matrix}$$

beschrieben werden. Es gelten nämlich die folgenden Rechenregeln:

Rechenregeln:

1. Multiplikation mit $\mathbf{P}_i$ von *links* $\longrightarrow$ vertausche *Zeilen* i und k

2. Multiplikation mit $\mathbf{P}_i$ von *rechts* $\longrightarrow$ vertausche *Spalten* i und k

3. Insbesondere gilt $\mathbf{P}_i^2 = \mathbf{I}$ (dies gilt für jede Permutationsmatrix)

Folglich wird aus der Matrix $\mathbf{A}_i$ durch Vertauschung die Matrix $\mathbf{P}_i\mathbf{A}_i$. Diese Matrix wird nun im Eliminationsschritt weiter reduziert. Der vollständige i-te Teilschritt des Gauß-Algorithmus mit Spaltenpivotsuche transformiert $\mathbf{A}_i$ also wie folgt:

$$\mathbf{A}_i \mapsto \mathbf{A}_{i+1} = \mathbf{L}_i\mathbf{P}_i\mathbf{A}_i. \tag{1.7}$$

Wir benötigen folgendes Lemma:

Lemma 1.24 Sei $j < i$ und $\mathbf{P}_i$ durch (1.6) gegeben. Dann ist $\mathbf{P}_i\mathbf{L}_j = \widetilde{\mathbf{L}}_j\mathbf{P}_i$, wobei $\widetilde{\mathbf{L}}_j$ wieder die gleiche Form hat wie $\mathbf{L}_j$, außer dass $\tau_k^{(j)}$ und $\tau_i^{(j)}$ vertauscht sind.

Beweis. Wegen $\mathbf{P}_i^2 = \mathbf{I}$ gilt

$$\mathbf{P}_i\mathbf{L}_j = \mathbf{P}_i\mathbf{L}_j\mathbf{P}_i^2 = (\mathbf{P}_i\mathbf{L}_j\mathbf{P}_i)\mathbf{P}_i,$$

dies bedeutet

$$\widetilde{\mathbf{L}}_j = \mathbf{P}_i\mathbf{L}_j\mathbf{P}_i.$$

Mit Hilfe der obigen Rechenregeln folgt

$$
\mathbf{P}_i\mathbf{L}_j\mathbf{P}_i = \begin{array}{l} \\ \\ \\ i\text{-te Zeile} \rightarrow \\ \\ \\ \\ k\text{-te Zeile} \rightarrow \\ \\ \\ \\

eine linke untere Dreiecksmatrix ist mit

$$\begin{aligned}
\widetilde{\mathbf{L}}_{n-1} &= \mathbf{L}_{n-1},\\
\widetilde{\mathbf{L}}_{n-2} &= \mathbf{P}_{n-1}\mathbf{L}_{n-2}\mathbf{P}_{n-1},\\
\widetilde{\mathbf{L}}_{n-3} &= \mathbf{P}_{n-1}\mathbf{P}_{n-2}\mathbf{L}_{n-3}\mathbf{P}_{n-2}\mathbf{P}_{n-1},\\
&\vdots\\
\widetilde{\mathbf{L}}_{1} &= \mathbf{P}_{n-1}\mathbf{P}_{n-2}\cdots\mathbf{P}_{2}\mathbf{L}_{1}\mathbf{P}_{2}\cdots\mathbf{P}_{n-2}\mathbf{P}_{n-1}.
\end{aligned}$$

Beweis. Nehmen wir zunächst an, dass der Gauß-Algorithmus mit Spaltenpivotsuche nicht zusammenbricht. Dann ergibt sich aus (1.7) durch sukzessive Anwendung von Lemma 1.24 dass

$$\begin{aligned}
\mathbf{R} = \mathbf{A}_n &= \mathbf{L}_{n-1}\mathbf{P}_{n-1}\mathbf{A}_{n-1}\\
&= \widetilde{\mathbf{L}}_{n-1}\mathbf{P}_{n-1}\mathbf{L}_{n-2}\mathbf{P}_{n-2}\mathbf{A}_{n-2}\\
&= \widetilde{\mathbf{L}}_{n-1}\widetilde{\mathbf{L}}_{n-2}\mathbf{P}_{n-1}\mathbf{P}_{n-2}\mathbf{L}_{n-3}\mathbf{P}_{n-3}\mathbf{A}_{n-3}\\
&\vdots\\
&= \widetilde{\mathbf{L}}_{n-1}\widetilde{\mathbf{L}}_{n-2}\cdots\widetilde{\mathbf{L}}_{1}\mathbf{P}_{n-1}\mathbf{P}_{n-2}\cdots\mathbf{P}_{1}\mathbf{A}.
\end{aligned}$$

Zu klären bleibt schließlich der Punkt, dass der Gauß-Algorithmus mit Spaltenpivotsuche nicht abbricht, also dass alle Pivotelemente nach der Spaltenpivotsuche von Null verschieden sind. Wäre das Pivotelement nach dem i-ten Teilschritt tatsächlich Null, dann gälte zwangsläufig

$$\mathbf{A}_i = \begin{array}{c} \lceil \\ \mathbf{B} \\ \lfloor \end{array} \left[\begin{array}{ccc|ccc} \star & \cdots & \star & & & \\ & \ddots & \vdots & & \star & \\ \mathbf{0} & & \star & & & \\ \hline & & & 0 & \star & \cdots & \star \\ & \mathbf{0} & & \vdots & \vdots & & \vdots \\ & & & 0 & \star & \cdots & \star \end{array}\right].$$

Daraus folgt jedoch

$$\det \mathbf{A}_i = \det \mathbf{B} \cdot \det \begin{bmatrix} 0 & \star & \cdots & \star \\ \vdots & \vdots & & \vdots \\ 0 & \star & \cdots & \star \end{bmatrix} = 0$$

und weiter

$$0 = \det \mathbf{A}_i = \det(\mathbf{L}_{i-1}\mathbf{P}_{i-1}\cdots\mathbf{L}_1\mathbf{P}_1\mathbf{A}) = \prod_{j=1}^{i-1} \underbrace{\det \mathbf{L}_j}_{=1} \cdot \prod_{j=1}^{i-1} \underbrace{\det \mathbf{P}_j}_{=\pm 1} \cdot \det \mathbf{A}.$$

Dies impliziert $\det \mathbf{A} = 0$ im Widerspruch zur Voraussetzung. □

Beispiel 1.26

$$\mathbf{A} = \begin{bmatrix} 1 & 1 & 0 & 2 \\ 1/2 & 1/2 & 2 & -1 \\ -1 & 0 & -1/8 & -5 \\ 1 & -7 & 9 & 10 \end{bmatrix} \begin{matrix} -1/2 \\ \hookleftarrow \\ \\ \end{matrix} \begin{matrix} +1 \\ \\ \hookleftarrow \\ \end{matrix} \begin{matrix} -1 \\ \\ \\ \hookleftarrow \end{matrix} \longrightarrow \begin{bmatrix} 1 & 1 & 0 & 2 \\ 0 & 0 & 2 & -2 \\ 0 & 1 & -1/8 & -3 \\ 0 & -8 & 9 & 8 \end{bmatrix} \begin{matrix} \hookleftarrow \\ \\ \\ \hookleftarrow \end{matrix}$$

$$\xrightarrow{\mathbf{P}_2} \begin{bmatrix} 1 & 1 & 0 & 2 \\ 0 & -8 & 9 & 8 \\ 0 & 1 & -1/8 & -3 \\ 0 & 0 & 2 & -2 \end{bmatrix} \begin{matrix} \\ \\ +1/8 \\ \hookleftarrow \end{matrix} \longrightarrow \begin{bmatrix} 1 & 1 & 0 & 2 \\ 0 & -8 & 9 & 8 \\ 0 & 0 & 1 & -2 \\ 0 & 0 & 2 & -2 \end{bmatrix} \begin{matrix} \\ \\ \hookleftarrow \\ \hookleftarrow \end{matrix}$$

$$\xrightarrow{\mathbf{P}_3} \begin{bmatrix} 1 & 1 & 0 & 2 \\ 0 & -8 & 9 & 8 \\ 0 & 0 & 2 & -2 \\ 0 & 0 & 1 & -2 \end{bmatrix} \begin{matrix} \\ \\ \\ -1/2 \\ \hookleftarrow \end{matrix} \longrightarrow \begin{bmatrix} 1 & 1 & 0 & 2 \\ 0 & -8 & 9 & 8 \\ 0 & 0 & 2 & -2 \\ 0 & 0 & 0 & -1 \end{bmatrix}.$$

Damit ergibt sich (beachte: $\mathbf{P}_1 = \mathbf{I}$)

$$\mathbf{R} = \begin{bmatrix} 1 & 1 & 0 & 2 \\ 0 & -8 & 9 & 8 \\ 0 & 0 & 2 & -2 \\ 0 & 0 & 0 & -1 \end{bmatrix}, \qquad \widetilde{\mathbf{L}} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 1/2 & 0 & 1 & 0 \\ -1 & -1/8 & 1/2 & 1 \end{bmatrix},$$

$$\mathbf{PA} = \begin{bmatrix} 1 & 1 & 0 & 2 \\ 1 & -7 & 9 & 10 \\ 1/2 & 1/2 & 2 & -1 \\ -1 & 0 & -1/8 & -5 \end{bmatrix}.$$

△

Faustregel: Um auf $\widetilde{\mathbf{L}}$ zu kommen, erstellt man zunächst eine Matrix $\mathbf{L}$ wie gewohnt, und führt dann in jeder Spalte *alle* Vertauschungen (in der Reihenfolge $\mathbf{P}_1, \mathbf{P}_2, \ldots, \mathbf{P}_{n-1}$) durch, bei denen nur Elemente unterhalb der Diagonalen betroffen sind.

Totale Pivotisierung: Im i-ten Teilschritt wird das Element $a_{k,\ell}^{(i)}$ $(i \le k, \ell \le n)$ als Pivotelement gewählt, das in der gesamten verbliebenen Restmatrix betragsmäßig am größten ist. Da man hierzu Zeilen und Spalten tauschen muss, benötigt man formal zwei Permutationsmatrizen $\mathbf{P}_i$ bzw. $\mathbf{\Pi}_i$:

$$\mathbf{A}_i \mapsto \mathbf{A}_{i+1} = \mathbf{L}_i \mathbf{P}_i \mathbf{A}_i \mathbf{\Pi}_i.$$

Man erhält so schließlich eine LR-Zerlegung der Matrix $\mathbf{PA\Pi}$ mit $\mathbf{\Pi} = \mathbf{\Pi}_1 \mathbf{\Pi}_2 \cdots \mathbf{\Pi}_{n-1}$.

Wird die Totalpivotsuche bei der Lösung eines linearen Gleichungssystems eingesetzt, dann entsprechen Spaltenvertauschungen Permutationen der Lösung $\mathbf{x}$. Der Ergebnisvektor ist also nicht mehr in der richtigen Reihenfolge.

Die totale Pivotisierung ist stabil, wird aber in der Praxis nur selten eingesetzt, da die Suche nach dem betragsgrößten Element im i-ten Schritt einem Aufwand $(n-i)^2$ entspricht. Der Gesamtaufwand

$$\sum_{i=1}^{n-1} (n-i)^2 \overset{j:=n-i}{=} \sum_{j=1}^{n-1} j^2 = \frac{1}{3} n^3$$

ist nicht mehr vernachlässigbar gegenüber der eigentlichen Rechnung.

1.6 Cholesky-Zerlegung

Wir betrachten zu gegebener Matrix $\mathbf{A} \in \mathbb{K}^{n\times n}$ die Blockpartitionierung

$$\mathbf{A} = \begin{bmatrix} \mathbf{A}_{1,1} & \mathbf{A}_{1,2} \\ \mathbf{A}_{2,1} & \mathbf{A}_{2,2} \end{bmatrix}$$

mit $\mathbf{A}_{1,1} \in \mathbb{K}^{p\times p}$ und $\mathbf{A}_{2,2} \in \mathbb{K}^{(n-p)\times(n-p)}$. Ist $\mathbf{A}_{1,1}$ nichtsingulär, so kann man das lineare Gleichungssystem

$$\mathbf{A}\begin{bmatrix}\mathbf{x}\\ \mathbf{y}\end{bmatrix} = \begin{bmatrix} \mathbf{A}_{1,1} & \mathbf{A}_{1,2} \\ \mathbf{A}_{2,1} & \mathbf{A}_{2,2} \end{bmatrix}\begin{bmatrix}\mathbf{x}\\ \mathbf{y}\end{bmatrix} = \begin{bmatrix}\mathbf{b}\\ \mathbf{c}\end{bmatrix} \tag{1.8}$$

vermittels Block-Gauß-Elimination lösen:

$$\mathbf{A} = \left[\begin{array}{cc|c} \mathbf{A}_{1,1} & \mathbf{A}_{1,2} & \mathbf{b} \\ \mathbf{A}_{2,1} & \mathbf{A}_{2,2} & \mathbf{c} \end{array}\right] \begin{array}{c} \\ -\mathbf{A}_{2,1}\mathbf{A}_{1,1}^{-1} \\ \longleftarrow \end{array} \longrightarrow \left[\begin{array}{cc|c} \mathbf{A}_{1,1} & \mathbf{A}_{1,2} & \mathbf{b} \\ 0 & \mathbf{A}_{2,2} - \mathbf{A}_{2,1}\mathbf{A}_{1,1}^{-1}\mathbf{A}_{1,2} & \mathbf{c} - \mathbf{A}_{2,1}\mathbf{A}_{1,1}^{-1}\mathbf{b} \end{array}\right]$$

Definition 1.27 Die Matrix

$$\mathbf{S} := \mathbf{A}_{2,2} - \mathbf{A}_{2,1}\mathbf{A}_{1,1}^{-1}\mathbf{A}_{1,2} \in \mathbb{K}^{(n-p)\times(n-p)} \tag{1.9}$$

heißt **Schurkomplement** von $\mathbf{A}$ bezüglich $\mathbf{A}_{1,1}$.

Für die Lösung von (1.8) folgt nun

$$\begin{aligned} \mathbf{y} &= \mathbf{S}^{-1}\big(\mathbf{c} - \mathbf{A}_{2,1}\mathbf{A}_{1,1}^{-1}\mathbf{b}\big), \\ \mathbf{x} &= \mathbf{A}_{1,1}^{-1}(\mathbf{b} - \mathbf{A}_{1,2}\mathbf{y}). \end{aligned}$$

Lemma 1.28 Sei $\mathbf{A} \in \mathbb{K}^{n\times n}$ hermitesch und positiv definit. Dann ist das Schurkomplement $\mathbf{S}$ wohldefiniert und sowohl $\mathbf{A}_{1,1}$ als auch $\mathbf{S}$ sind hermitesch und positiv definit.

Beweis. Sei $\left[\begin{smallmatrix}\mathbf{x}\\ \mathbf{y}\end{smallmatrix}\right]$ entsprechend zu $\mathbf{A}$ partitioniert. Wegen

$$\begin{bmatrix} \mathbf{A}_{1,1} & \mathbf{A}_{1,2} \\ \mathbf{A}_{2,1} & \mathbf{A}_{2,2} \end{bmatrix} = \mathbf{A} = \mathbf{A}^\star = \begin{bmatrix} \mathbf{A}_{1,1}^\star & \mathbf{A}_{2,1}^\star \\ \mathbf{A}_{1,2}^\star & \mathbf{A}_{2,2}^\star \end{bmatrix}$$

ergibt sich

$$\mathbf{A}_{1,1} = \mathbf{A}_{1,1}^\star, \qquad \mathbf{A}_{2,2} = \mathbf{A}_{2,2}^\star, \qquad \mathbf{A}_{1,2} = \mathbf{A}_{2,1}^\star.$$

Folglich ist $\mathbf{A}_{1,1}$ hermitesch und es gilt

$$0 \le \begin{bmatrix}\mathbf{x}\\ \mathbf{0}\end{bmatrix}^\star \mathbf{A} \begin{bmatrix}\mathbf{x}\\ \mathbf{0}\end{bmatrix} = \begin{bmatrix}\mathbf{x}\\ \mathbf{0}\end{bmatrix}^\star \begin{bmatrix}\mathbf{A}_{1,1}\mathbf{x}\\ \mathbf{A}_{2,1}\mathbf{x}\end{bmatrix} = \mathbf{x}^\star\mathbf{A}_{1,1}\mathbf{x},$$

wobei sich Gleichheit nur für $\mathbf{x} = \mathbf{0}$ ergibt. Also ist $\mathbf{A}_{1,1}$ positiv definit und $\mathbf{A}_{1,1}^{-1}$ existiert. $\mathbf{S}$ ist somit wohldefiniert und

$$\mathbf{S}^\star = \mathbf{A}_{2,2}^\star - \mathbf{A}_{1,2}^\star\mathbf{A}_{1,1}^{-\star}\mathbf{A}_{2,1}^\star = \mathbf{A}_{2,2} - \mathbf{A}_{2,1}\mathbf{A}_{1,1}^{-1}\mathbf{A}_{1,2} = \mathbf{S}.$$

Schließlich betrachten wir $\left[\begin{smallmatrix}\mathbf{x}\\ \mathbf{y}\end{smallmatrix}\right]$ mit $\mathbf{x} = -\mathbf{A}_{1,1}^{-1}\mathbf{A}_{1,2}\mathbf{y}$:

$$\begin{aligned} 0 \leq \begin{bmatrix}\mathbf{x}\\ \mathbf{y}\end{bmatrix}^\star \mathbf{A} \begin{bmatrix}\mathbf{x}\\ \mathbf{y}\end{bmatrix} &= \begin{bmatrix}\mathbf{x}\\ \mathbf{y}\end{bmatrix}^\star \begin{bmatrix}\mathbf{A}_{1,1}\mathbf{x} + \mathbf{A}_{1,2}\mathbf{y}\\ \mathbf{A}_{2,1}\mathbf{x} + \mathbf{A}_{2,2}\mathbf{y}\end{bmatrix} = \begin{bmatrix}\mathbf{x}\\ \mathbf{y}\end{bmatrix}^\star \begin{bmatrix}-\mathbf{A}_{1,2}\mathbf{y} + \mathbf{A}_{1,2}\mathbf{y}\\ -\mathbf{A}_{2,1}\mathbf{A}_{1,1}^{-1}\mathbf{A}_{1,2}\mathbf{y} + \mathbf{A}_{2,2}\mathbf{y}\end{bmatrix} \\ &= \begin{bmatrix}\mathbf{x}\\ \mathbf{y}\end{bmatrix}^\star \begin{bmatrix}\mathbf{0}\\ \mathbf{S}\mathbf{y}\end{bmatrix} = \mathbf{y}^\star\mathbf{S}\mathbf{y}. \end{aligned}$$

Da Gleichheit nur im Falle $\mathbf{x} = \mathbf{0}$ und $\mathbf{y} = \mathbf{0}$ gilt, ist $\mathbf{S}$ ebenfalls positiv definit. □

Im weiteren betrachten wir nur hermitesche Matrizen $\mathbf{A} = \mathbf{A}^\star \in \mathbb{K}^{n\times n}$.

Definition 1.29 Eine Zerlegung $\mathbf{A} = \mathbf{L}\mathbf{L}^\star$ mit unterer Dreiecksmatrix $\mathbf{L}$ mit positiven Diagonaleinträgen heißt **Cholesky-Zerlegung** von $\mathbf{A}$.

Proposition 1.30 Hat $\mathbf{A}$ eine Cholesky-Zerlegung, dann ist $\mathbf{A}$ hermitesch und positiv definit.

Beweis. Aus $\mathbf{A} = \mathbf{L}\mathbf{L}^\star$ folgt

$$\mathbf{A}^\star = \left(\mathbf{L}^\star\right)^\star\mathbf{L}^\star = \mathbf{L}\mathbf{L}^\star = \mathbf{A},$$

das heißt $\mathbf{A}$ ist hermitsch. Wegen

$$\mathbf{x}^\star\mathbf{A}\mathbf{x} = \mathbf{x}^\star\mathbf{L}\mathbf{L}^\star\mathbf{x} = \left(\mathbf{L}^\star\mathbf{x}\right)^\star\mathbf{L}^\star\mathbf{x} = \|\mathbf{L}^\star\mathbf{x}\|_2^2 \geq 0$$

ist $\mathbf{A}$ auch positiv semidefinit. Da $\mathbf{L}$ nach Voraussetzung nichtsingulär ist, impliziert $\mathbf{L}^\star\mathbf{x} = \mathbf{0}$ auch $\mathbf{x} = \mathbf{0}$. Damit ist $\mathbf{A}$ sogar definit. □

Satz 1.31 Ist $\mathbf{A}$ hermitesch und positiv definit, dann existiert eine Cholesky-Zerlegung von $\mathbf{A}$.

Beweis. Induktion über n:

$\underline{n = 1}$: Da $\mathbf{A} = [a_{1,1}]$ positiv definit ist gilt $a_{1,1} > 0$. Wegen

$$\mathbf{A} = [a_{1,1}] \overset{!}{=} [\ell_{1,1}] \cdot [\overline{\ell_{1,1}}] = \mathbf{L} \cdot \mathbf{L}^\star$$

folgt damit $\ell_{1,1} = \sqrt{a_{1,1}} > 0$.

$\underline{n - 1 \mapsto n}$: Betrachte

$$\mathbf{A} = \left[\begin{array}{c|c} a_{1,1} & \mathbf{A}_{1,2} \\ \hline \mathbf{A}_{2,1} & \mathbf{A}_{2,2} \end{array}\right]$$

mit $\mathbf{A}_{2,1} = \mathbf{A}_{1,2}^\star$ und das Schurkomplement

$$\mathbf{S} = \mathbf{A}_{2,2} - \frac{1}{a_{1,1}}\mathbf{A}_{2,1}\mathbf{A}_{1,2} \tag{1.10}$$

von $\mathbf{A}$ bezüglich $a_{1,1}$. Nach Lemma 1.28 ist $a_{1,1} > 0$ und $\mathbf{S}$ hermitesch und positiv definit. Also ist $\ell_{1,1} = \sqrt{a_{1,1}} > 0$ und aufgrund der Induktionsannahme hat $\mathbf{S}$ eine Cholesky-Zerlegung

$$\mathbf{S} = \widetilde{\mathbf{L}}\widetilde{\mathbf{L}}^\star.$$

Definiere damit

$$\mathbf{L} = \left[\begin{array}{c|c} \ell_{1,1} & \mathbf{0} \\ \hline \frac{1}{\ell_{1,1}}\mathbf{A}_{2,1} & \widetilde{\mathbf{L}} \end{array}\right], \qquad \mathbf{L}^\star = \left[\begin{array}{c|c} \ell_{1,1} & \frac{1}{\ell_{1,1}}\mathbf{A}_{1,2} \\ \hline \mathbf{0} & \widetilde{\mathbf{L}}^\star \end{array}\right].$$

Es ergibt sich

$$\mathbf{L}\mathbf{L}^\star = \left[\begin{array}{c|c} \ell_{1,1} & \mathbf{0} \\ \hline \frac{1}{\ell_{1,1}}\mathbf{A}_{2,1} & \widetilde{\mathbf{L}} \end{array}\right] \cdot \left[\begin{array}{c|c} \ell_{1,1} & \frac{1}{\ell_{1,1}}\mathbf{A}_{1,2} \\ \hline \mathbf{0} & \widetilde{\mathbf{L}}^\star \end{array}\right] = \left[\begin{array}{c|c} a_{1,1} & \mathbf{A}_{1,2} \\ \hline \mathbf{A}_{2,1} & \frac{1}{a_{1,1}}\mathbf{A}_{2,1}\mathbf{A}_{1,2} + \widetilde{\mathbf{L}}\widetilde{\mathbf{L}}^\star \end{array}\right].$$

Wegen

$$\frac{1}{a_{1,1}}\mathbf{A}_{2,1}\mathbf{A}_{1,2} + \widetilde{\mathbf{L}}\widetilde{\mathbf{L}}^\star = \frac{1}{a_{1,1}}\mathbf{A}_{2,1}\mathbf{A}_{1,2} + \mathbf{S} \overset{(1.10)}{=} \mathbf{A}_{2,2}$$

folgt hieraus

$$\mathbf{L}\mathbf{L}^\star = \left[\begin{array}{c|c} a_{1,1} & \mathbf{A}_{1,2} \\ \hline \mathbf{A}_{2,1} & \mathbf{A}_{2,2} \end{array}\right] = \mathbf{A}.$$

□

Bemerkung: Durch Kombination von Proposition 1.30 und Satz 1.31 erhalten wir die Aussage, dass eine Cholesky-Zerlegung genau dann existiert, falls $\mathbf{A}$ hermitesch und positiv definit ist.

Die Berechnung von $\mathbf{L}$ ergibt sich durch Koeffizientenvergleich: Aus

$$\underbrace{\begin{bmatrix} a_{1,1} & a_{1,2} & \cdots & a_{1,n} \\ a_{2,1} & a_{2,2} & \cdots & a_{2,n} \\ \vdots & \vdots & & \vdots \\ a_{n,1} & a_{n,2} & \cdots & a_{n,n} \end{bmatrix}}_{=\mathbf{A}} = \underbrace{\begin{bmatrix} \ell_{1,1} & & & \mathbf{0} \\ \ell_{2,1} & \ell_{2,2} & & \\ \vdots & \vdots & \ddots & \\ \ell_{n,1} & \ell_{n,2} & \cdots & \ell_{n,n} \end{bmatrix}}_{=\mathbf{L}} \cdot \underbrace{\begin{bmatrix} \overline{\ell_{1,1}} & \overline{\ell_{2,1}} & \cdots & \overline{\ell_{n,1}} \\ & \overline{\ell_{2,2}} & & \overline{\ell_{n,2}} \\ & & \ddots & \vdots \\ \mathbf{0} & & & \overline{\ell_{n,n}} \end{bmatrix}}_{=\mathbf{L}^\star}$$

folgt

$$\begin{aligned}
a_{1,1} &= |\ell_{1,1}|^2 && \rightsquigarrow \ell_{1,1} = \sqrt{a_{1,1}} > 0 \\
a_{2,1} &= \ell_{2,1}\overline{\ell_{1,1}} && \rightsquigarrow \ell_{2,1} = a_{2,1}/\overline{\ell_{1,1}} \\
a_{3,1} &= \ell_{3,1}\overline{\ell_{1,1}} && \rightsquigarrow \ell_{3,1} = a_{3,1}/\overline{\ell_{1,1}} \\
&\vdots && \vdots \\
a_{n1} &= \ell_{n1}\overline{\ell_{1,1}} && \rightsquigarrow \ell_{n1} = a_{n1}/\overline{\ell_{1,1}} \\
a_{2,2} &= |\ell_{2,1}|^2 + |\ell_{2,2}|^2 && \rightsquigarrow \ell_{2,2} = \sqrt{a_{2,2} - |\ell_{2,1}|^2} > 0 \\
a_{3,2} &= \ell_{3,1}\overline{\ell_{2,1}} + \ell_{3,2}\overline{\ell_{2,2}} && \rightsquigarrow \ell_{3,2} = (a_{3,2} - \ell_{3,1}\overline{\ell_{2,1}})/\overline{\ell_{2,2}} \\
&\vdots && \vdots \\
a_{n,2} &= \ell_{n,1}\overline{\ell_{2,1}} + \ell_{n,2}\overline{\ell_{2,2}} && \rightsquigarrow \ell_{n,2} = (a_{n,2} - \ell_{n,1}\overline{\ell_{2,1}})/\overline{\ell_{2,2}}
\end{aligned}$$

und allgemein

$$\begin{aligned} \ell_{j,j} &= \sqrt{a_{j,j} - \sum_{k=1}^{j-1} |\ell_{j,k}|^2} > 0, && 1 \le j \le n, \\ \ell_{i,j} &= \frac{1}{\overline{\ell_{j,j}}} \left(a_{i,j} - \sum_{k=1}^{j-1} \ell_{i,k} \overline{\ell_{j,k}} \right), && j < i \le n. \end{aligned} \tag{1.11}$$

Die Berechenbarkeit ist durch den Existenzbeweis (Satz 1.31) gewährleistet (alle $\ell_{i,i} \neq 0$). Aus (1.11) ergibt sich sofort die folgende Aussage:

Korollar 1.32 Die Cholesky-Zerlegung von $\mathbf{A} = \mathbf{A}^\star$ positiv definit ist eindeutig.

Aufwand: Zur Berechnung von $\ell_{i,j}$ $(j \le i \le n)$ sind j Multiplikationen bzw. Wurzeln auszuführen. Damit werden insgesamt

$$\sum_{j=1}^{n} (n-j+1)j = \frac{n^2(n+1)}{2} - \frac{n(n+1)(2n+1)}{6} + \mathcal{O}(n^2) = \frac{1}{6}n^3 + \mathcal{O}(n^2)$$

Multiplikationen bzw. Wurzeln benötigt. Der Aufwand ist demnach nur halb so groß wie für die LR-Zerlegung.

Beispiel 1.33 Für

$$\mathbf{A} = \begin{bmatrix} 1 & 2 & 1 \\ 2 & 5 & 2 \\ 1 & 2 & 10 \end{bmatrix}$$

ergibt sich wegen

$$\begin{aligned} \ell_{1,1} &= \sqrt{1} = 1 & \qquad \ell_{3,1} &= 1/1 = 1 \\ \ell_{2,1} &= 2/1 = 2 & \qquad \ell_{3,2} &= (2-2)/1 = 0 \\ \ell_{2,2} &= \sqrt{5-4} = 1 & \qquad \ell_{3,3} &= \sqrt{10-1} = 3 \end{aligned}$$

die Cholesky-Zerlegung $\mathbf{L}\mathbf{L}^\star$ mit

$$\mathbf{L} = \begin{bmatrix} 1 & & \\ 2 & 1 & \\ 1 & 0 & 3 \end{bmatrix}, \qquad \mathbf{L}^\star = \begin{bmatrix} 1 & 2 & 1 \\ & 1 & 0 \\ & & 3 \end{bmatrix}.$$

△

Bemerkung: Im Gegensatz zur LR-Zerlegung ist die Cholesky-Zerlegung immer stabil.

2. Lineare Ausgleichsprobleme

2.1 Normalengleichungen

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

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

Da wir m Gleichungen für n Unbekannte haben, ist das lineare Gleichungssystem im allgemeinen nicht — oder nicht eindeutig — lösbar. Ist $m > n$, dann nennen wir das lineare Gleichungssystem *überbestimmt*, ist $m < n$, dann nennen wir es *unterbestimmt*.

Überbestimmte Probleme treten häufig in den Anwendungen auf, wenn es darum geht, Modellparameter an Messdaten anzupassen.

Beispiel 2.1 Unter Einfluss der Schwerkraft fliegen geworfene Körper auf Parabeln. Hat der Körper die Anfangsgeschwindigkeit $\mathbf{v} = (v_x, v_y)^\star$ zum Zeitpunkt $t = 0$ am Punkt $\mathbf{0}$ und fliegt er anschließend nur unter Einfluss der Schwerkraft, so ist er zum Zeitpunkt $t > 0$ am Ort

$$x = v_x t, \qquad y = v_y t - \frac{1}{2} g t^2,$$

wobei g die Erdbeschleunigung ist. Die Anfangsgeschwindigkeit v_y und die Erdbeschleunigung g seien unbekannt und sollen aus Messungen bestimmt werden. Hierzu wurde die Höhe über Grund des Körpers zu folgenden Zeiten gemessen:

i	1	2	3	4	5	6	7
t_i [s]	0.1	0.4	0.5	0.9	1.0	1.2	2.0
y_i [m]	0.96	3.26	3.82	5.11	5.2	5.05	0.58

Es ergeben sich damit sieben Gleichungen für die zwei unbekannten Parameter v_y und g:

$$y_i = t_i v_y - \frac{1}{2} t_i^2 g, \qquad i = 1, 2, \ldots, 7.$$

Führt man die Matrix $\mathbf{A} \in \mathbb{R}^{7\times 2}$ mit

$$a_{i,1} = t_i, \qquad a_{i,2} = -\frac{1}{2} t_i^2, \qquad i = 1, 2, \ldots, 7,$$

ein, so ergibt sich

$$\mathbf{A} \begin{bmatrix} v_y \\ g \end{bmatrix} = \begin{bmatrix} 0.1 & -0.005 \\ 0.4 & -0.08 \\ 0.5 & -0.125 \\ 0.9 & -0.405 \\ 1.0 & -0.5 \\ 1.2 & -0.72 \\ 2.0 & -2.0 \end{bmatrix} \cdot \begin{bmatrix} v_y \\ g \end{bmatrix} = \begin{bmatrix} 0.96 \\ 3.26 \\ 3.82 \\ 5.11 \\ 5.2 \\ 5.05 \\ 0.5 \end{bmatrix} = \mathbf{y}. \tag{2.1}$$

△

In Beispiel 2.1 erhalten wir also ein überbestimmtes Gleichungssystem mit $m = 7$ und $n = 2$, das keine klassische Lösung besitzt. Die Ausgleichsrechnung liefert nun eine Methode, sinnvolle Lösungen von überbestimmten Gleichungssystemen zu definieren.

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

$$\mathbf{r} = \mathbf{b} - \mathbf{A}\mathbf{x} \tag{2.2}$$

möglichst klein ist.

Definition 2.2 Für eine Matrix $\mathbf{A} \in \mathbb{K}^{m\times n}$ und ein $\mathbf{b} \in \mathbb{K}^m$ heißt das Problem

$$\|\mathbf{b} - \mathbf{A}\mathbf{x}\|_2 \to \min \tag{2.3}$$

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

Bemerkung: Der Lösungsbegriff in (2.3) ist eine Verallgemeinerung der klassischen Lösung. Ist nämlich $m = n$ und ist $\mathbf{x} \in \mathbb{K}^n$ eine klassische Lösung, das heißt, gilt $\mathbf{A}\mathbf{x} = \mathbf{b}$, dann ist offensichtlich $\mathbf{x}$ ebenfalls eine Lösung von (2.3).

Satz 2.3 Die Lösungen von (2.3) sind genau die Lösungen der **Gaußschen Normalengleichungen**

$$\mathbf{A}^\star\mathbf{A}\mathbf{x} = \mathbf{A}^\star\mathbf{b}, \tag{2.4}$$

insbesondere existiert eine Lösung $\mathbf{x}$. Ist $\mathbf{z}$ eine weitere Lösung, so gilt $\mathbf{A}\mathbf{x} = \mathbf{A}\mathbf{z}$. Das Residuum (2.2) ist eindeutig bestimmt und genügt der Gleichung $\mathbf{A}^\star\mathbf{r} = \mathbf{0}$.

Beweis. Das Bild der Matrix $\mathbf{A}$ ist der lineare Teilraum

$$\operatorname{img}(\mathbf{A}) = \{\mathbf{A}\mathbf{x} : \mathbf{x} \in \mathbb{K}^n\} \subset \mathbb{K}^m,$$

der von den Spalten von $\mathbf{A}$ aufgespannt wird. Wegen

$$\begin{aligned}\operatorname{img}(\mathbf{A})^\perp &= \{\mathbf{r} \in \mathbb{K}^m : \mathbf{r}^\star\mathbf{z} = 0 \text{ für alle } \mathbf{z} \in \operatorname{img}(\mathbf{A})\}\\ &= \{\mathbf{r} \in \mathbb{K}^m : \mathbf{r}^\star\mathbf{A} = \mathbf{0}\}\\ &= \operatorname{kern}(\mathbf{A}^\star)\end{aligned}$$

folgt

$$\mathbb{K}^m = \operatorname{img}(\mathbf{A}) \overset{\perp}{\oplus} \operatorname{kern}(\mathbf{A}^\star), \tag{2.5}$$

das heißt, der Vektor $\mathbf{b} \in \mathbb{K}^m$ lässt sich eindeutig schreiben als

$$\mathbf{b} = \mathbf{y} + \mathbf{r}, \quad \mathbf{y} \in \operatorname{img}(\mathbf{A}), \quad \mathbf{r} \in \operatorname{kern}(\mathbf{A}^\star).$$

Folglich gibt es mindestens ein $\mathbf{x} \in \mathbb{K}^n$ mit $\mathbf{A}\mathbf{x} = \mathbf{y}$ und es gilt

$$\mathbf{A}^\star\mathbf{b} = \mathbf{A}^\star\mathbf{y} + \underbrace{\mathbf{A}^\star\mathbf{r}}_{=\mathbf{0}} = \mathbf{A}^\star\mathbf{A}\mathbf{x},$$

das heißt, $\mathbf{x}$ löst die Normalengleichungen (2.4).
Um zu zeigen, dass $\mathbf{x}$ auch Ausgleichslösung ist, setzen wir für beliebiges $\mathbf{z} \in \mathbb{K}^n$

$$\mathbf{y} := \mathbf{A}\mathbf{z} - \mathbf{A}\mathbf{x}, \quad \mathbf{r} := \mathbf{b} - \mathbf{A}\mathbf{x}.$$

Wegen $\mathbf{A}^\star\mathbf{r} = \mathbf{0}$ folgt

$$\|\mathbf{b} - \mathbf{A}\mathbf{z}\|_2^2 = \|\mathbf{r} - \mathbf{y}\|_2^2 = \|\mathbf{r}\|_2^2 + \|\mathbf{y}\|_2^2 \geq \|\mathbf{r}\|_2^2 = \|\mathbf{b} - \mathbf{A}\mathbf{x}\|_2.$$

Der Vektor $\mathbf{x}$ minimiert also (2.3). Gleichheit ("=") gilt hierbei nur dann für ein $\mathbf{z} \in \mathbb{K}^n$, wenn

$$0 = \|\mathbf{y}\|_2^2 = \|\mathbf{A}(\mathbf{z} - \mathbf{x})\|_2^2$$

das heißt genau dann, wenn

$$\mathbf{z} - \mathbf{x} \in \operatorname{kern}(\mathbf{A}) \subseteq \operatorname{kern}(\mathbf{A}^\star\mathbf{A}),$$

also genau dann, wenn auch $\mathbf{z}$ (2.4) löst. □

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

Satz 2.4 Die Matrix $\mathbf{A}^\star\mathbf{A} \in \mathbb{K}^{n\times n}$ ist hermitesch und positiv semidefinit. Darüberhinaus ist $\mathbf{A}^\star\mathbf{A}$ genau dann positiv definit, wenn der Kern von $\mathbf{A}$ trivial ist, das heißt, wenn $\operatorname{kern}(\mathbf{A}) = \{\mathbf{0}\}$. Dies ist genau dann der Fall, wenn die Spalten von $\mathbf{A}$ linear unabhängig sind.

Beweis. Offensichtlich ist $\mathbf{A}^\star\mathbf{A}$ hermitesch und wegen

$$\mathbf{x}^\star\mathbf{A}^\star\mathbf{A}\mathbf{x} = \|\mathbf{A}\mathbf{x}\|_2^2 \geq 0 \qquad \forall \mathbf{x} \in \mathbb{K}^n$$

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

Beispiel 2.5 (Fortsetzung von Beipiel 2.1) Für das überbestimmte System (2.1) ist die gesuchte Normalengleichung

$$\mathbf{A}^\star\mathbf{A} = \begin{bmatrix} 0.1 & 0.4 & 0.5 & 0.9 & 1.0 & 1.2 & 2.0 \\ -0.005 & -0.08 & -0.125 & -0.405 & -0.5 & -0.72 & -2.0 \end{bmatrix} \begin{bmatrix} 0.1 & -0.005 \\ 0.4 & -0.08 \\ 0.5 & -0.125 \\ 0.9 & -0.405 \\ 1.0 & -0.5 \\ 1.2 & -0.72 \\ 2.0 & -2.0 \end{bmatrix}$$

$$= \begin{bmatrix} 7.6700 & -5.8235 \\ -5.8235 & 4.954475 \end{bmatrix}$$

und

$$\mathbf{A}^\star\mathbf{y} = \begin{bmatrix} 0.1 & 0.4 & 0.5 & 0.9 & 1.0 & 1.2 & 2.0 \\ -0.005 & -0.08 & -0.125 & -0.405 & -0.5 & -0.72 & -2.0 \end{bmatrix} \begin{bmatrix} 0.96 \\ 3.26 \\ 3.82 \\ 5.11 \\ 5.2 \\ 5.05 \\ 0.58 \end{bmatrix}$$

$$= \begin{bmatrix} 20.3290 \\ -10.20865 \end{bmatrix}.$$

Die gesuchte Ausgleichslösung (v_y, g) erfüllt also

$$\begin{bmatrix} 7.6700 & -5.8235 \\ -5.8235 & 4.954475 \end{bmatrix} \begin{bmatrix} v_y \\ g \end{bmatrix} = \begin{bmatrix} 20.3290 \\ -10.20865 \end{bmatrix}.$$

Die Matrix $\mathbf{A}^\star\mathbf{A}$ ist in der Tat symmetrisch und positiv definit, und die damit eindeutige Lösung (v_y, g) ist mit drei Stellen Genauigkeit

$$\begin{bmatrix} v_y \\ g \end{bmatrix} = \begin{bmatrix} 10.1 \\ 9.81 \end{bmatrix}.$$

△

Prinzipiell könnte man das lineare Ausgleichsproblem mit den Gaußschen Normalengleichungen auch numerisch behandeln, etwa mittels Cholesky-Zerlegung der Matrix $\mathbf{A}^\star\mathbf{A}$ in den Normalengleichungen. Dies ist jedoch ein typisches Beispiel für einen numerischen Algorithmus, der deutlich weniger stabil ist als das eigentlich zu lösende Problem. Man sieht das am einfachsten in dem Fall, dass $\mathbf{A} \in \mathbb{K}^{n\times n}$ invertierbar ist. Dann ist die Kondition des linearen Gleichungssystems $\mathbf{Ax} = \mathbf{b}$

$$\operatorname{cond}_2 \mathbf{A} = \|\mathbf{A}\|_2 \cdot \|\mathbf{A}^{-1}\|_2 = \sqrt{\frac{\lambda_{\max}(\mathbf{A}^\star\mathbf{A})}{\lambda_{\min}(\mathbf{A}^\star\mathbf{A})}},$$

aber die Kondition der Normalengleichung ist

$$\operatorname{cond}_2(\mathbf{A}^\star\mathbf{A}) = \|\mathbf{A}^\star\mathbf{A}\|_2 \cdot \|(\mathbf{A}^\star\mathbf{A})^{-1}\|_2 = \frac{\lambda_{\max}(\mathbf{A}^\star\mathbf{A})}{\lambda_{\min}(\mathbf{A}^\star\mathbf{A})} = (\operatorname{cond}_2 \mathbf{A})^2.$$

Wir wollen daher in den nachfolgenden Abschnitten einen Algorithmus herleiten, der die Normalengleichungen vermeidet.

2.2 QR-Zerlegung

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

Definition 2.6 Eine Matrix $\mathbf{Q} \in \mathbb{K}^{n\times n}$ heißt **unitär**, falls

$$\mathbf{Q}^\star\mathbf{Q} = \mathbf{I},$$

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

Eigenschaften unitärer Matrizen:

1. Wegen
$$\|\mathbf{Q}\mathbf{x}\|_2^2 = (\mathbf{Q}\mathbf{x})^\star\mathbf{Q}\mathbf{x} = \mathbf{x}^\star \underbrace{\mathbf{Q}^\star\mathbf{Q}}_{=\mathbf{I}}\mathbf{x} = \mathbf{x}^\star\mathbf{x} = \|\mathbf{x}\|_2^2$$
gilt
$$\|\mathbf{Q}\mathbf{x}\|_2 = \|\mathbf{x}\|_2 \quad \forall \mathbf{x} \in \mathbb{K}^n.$$

2. Es gilt
$$\operatorname{cond}_2 \mathbf{Q} = 1,$$
da
$$\|\mathbf{Q}\|_2 = \max_{\|\mathbf{x}\|_2=1} \|\mathbf{Q}\mathbf{x}\|_2 = 1$$
und
$$\|\mathbf{Q}^{-1}\|_2 = \|\mathbf{Q}^\star\|_2 = \|\mathbf{Q}\|_2 = 1.$$

3. Mit $\mathbf{P}, \mathbf{Q} \in \mathbb{K}^{n\times n}$ unitär ist auch $\mathbf{PQ}$ unitär, da
$$(\mathbf{PQ})^\star\mathbf{PQ} = \mathbf{Q}^\star \underbrace{\mathbf{P}^\star\mathbf{P}}_{=\mathbf{I}}\mathbf{Q} = \mathbf{Q}^\star\mathbf{Q} = \mathbf{I}.$$

Definition 2.7 Sei $\mathbf{v} \in \mathbb{K}^n \setminus \{\mathbf{0}\}$. Die Matrix

$$\mathbf{P} = \mathbf{I} - \frac{2}{\|\mathbf{v}\|_2^2}\mathbf{v}\mathbf{v}^\star \in \mathbb{K}^{n\times n}$$

heißt **Householder-Transformation**.

Lemma 2.8 $\mathbf{P}$ ist eine hermitesche unitäre Matrix mit

$$\mathbf{P}\mathbf{v} = -\mathbf{v}$$

und für alle $\mathbf{w} \in \mathbb{K}^n$ mit $\mathbf{w} \perp \mathbf{v}$ gilt

$$\mathbf{P}\mathbf{w} = \mathbf{w}.$$

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

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

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

$$\mathbf{P}\mathbf{v} = \mathbf{I}\mathbf{v} - \frac{2}{\|\mathbf{v}\|_2^2}\mathbf{v}\underbrace{\mathbf{v}^\star\mathbf{v}}_{=\|\mathbf{v}\|_2^2} = \mathbf{v} - 2\mathbf{v} = -\mathbf{v}$$

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

$$\mathbf{P}\mathbf{w} = \mathbf{I}\mathbf{w} - \frac{2}{\|\mathbf{v}\|_2^2}\mathbf{v}\underbrace{\mathbf{v}^\star\mathbf{w}}_{=0} = \mathbf{w}.$$

□

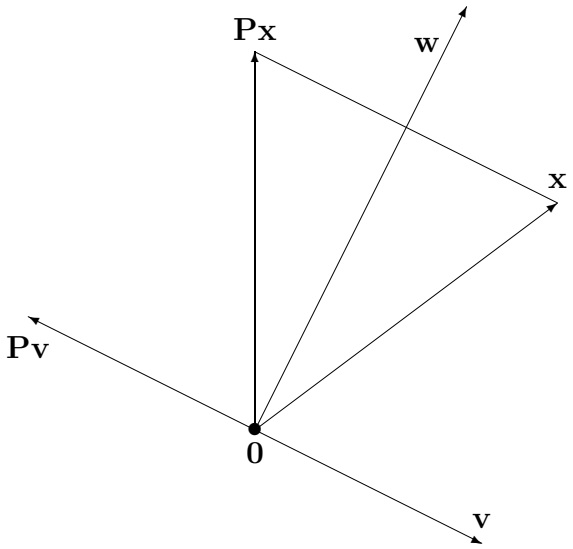


Abbildung 2.1: Householder-Transformationen sind Spiegelungen!

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

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

Lemma 2.9 Gegeben sei $\mathbf{x} \in \mathbb{K}^n \setminus \{\mathbf{0}\}$ mit $\mathbf{x} \notin \operatorname{span}\{\mathbf{e}_1\}$. Für

$$\mathbf{v} = \mathbf{x} + \sigma\mathbf{e}_1 \quad \text{mit} \quad \sigma = \pm \begin{cases} \frac{x_1}{|x_1|}\|\mathbf{x}\|_2, & \text{falls } x_1 \neq 0, \\ \|\mathbf{x}\|_2, & \text{falls } x_1 = 0, \end{cases} \tag{2.6}$$

gilt

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

Beweis. Wegen $\mathbf{x} \notin \operatorname{span}\{\mathbf{e}_1\}$ ist $\mathbf{v} \neq \mathbf{0}$. Weiter gilt

$$\|\mathbf{x} + \sigma\mathbf{e}_1\|_2^2 = \|\mathbf{x}\|_2^2 \underbrace{+\overline{\sigma}\mathbf{e}_1^\star\mathbf{x} + \sigma\mathbf{x}^\star\mathbf{e}_1}_{=2\overline{\sigma}\mathbf{e}_1^\star\mathbf{x}} + |\sigma|^2 = 2(\mathbf{x} + \sigma\mathbf{e}_1)^\star\mathbf{x}.$$

Daraus erhält man

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

was zusammen mit (2.6) die Darstellung

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

liefert. Dies implizert die Behauptung. □

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

$$\mathbf{v} = \mathbf{x} + \frac{x_1}{|x_1|}\|\mathbf{x}\|_2\mathbf{e}_1, \qquad \|\mathbf{v}\|_2^2 = 2\|\mathbf{x}\|_2^2 + 2|x_1|\|\mathbf{x}\|_2. \tag{2.7}$$

Satz 2.10 Sei $\mathbf{A} \in \mathbb{K}^{m\times n}$ mit $\operatorname{rang}(\mathbf{A}) = n$ (also $m \geq n$). Dann existiert eine unitäre Matrix $\mathbf{Q} \in \mathbb{K}^{m\times m}$ und eine obere Dreiecksmatrix $\mathbf{R} \in \mathbb{K}^{m\times n}$ mit

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

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

$$\mathbf{Q}_{n-1}\mathbf{Q}_{n-2}\cdots\mathbf{Q}_1\mathbf{A} = \mathbf{R}. \tag{2.8}$$

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

$$\mathbf{Q} = \mathbf{Q}_1 \mathbf{Q}_2 \cdots \mathbf{Q}_{n-1}.$$

Im ersten Schritt setzen wir $\mathbf{A}_1 := \mathbf{A}$ und $\mathbf{x} = \mathbf{a}_1$ (erste Spalte von $\mathbf{A}_1$) und bestimmen die Householder-Transformation $\mathbf{Q}_1 \in \mathbb{K}^{m \times m}$ gemäß (2.7). Es folgt

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

beziehungsweise

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

Im nächsten Schritt setzen wir $\mathbf{x} = \mathbf{a}_2 \in \mathbb{K}^{m-1}$ (erste Spalte von $\mathbf{A}_2$) und wählen wiederum die Householder-Matrix $\widetilde{\mathbf{Q}}_2 \in \mathbb{K}^{(m-1)\times(m-1)}$ gemäß (2.7). Wir erhalten

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

beziehungsweise

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

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

Auf diese Weise erhalten wir sukzessive die gewünschte Zerlegung (2.8). Man beachte, dass $|r_{i,i}| = \|\mathbf{a}_i\|$ $(1 \leq i \leq n)$ immer von Null verschieden ist, da ansonsten $\mathbf{A}_i$ und damit auch $\mathbf{A}$ einen Rangdefekt hätte. □

Bemerkungen:

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

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

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

2. Die während der QR-Zerlegung anfallenden Vektoren $\mathbf{v}_i = [0, \ldots, 0, 1, v_{i,i+1}, \ldots, v_{i,m}]^T$ lassen sich analog zur LR-Zerlegung wieder in der freiwerdenden linken unteren Dreiecksmatrix von $\mathbf{A}$ speichern. Die Matrix $\mathbf{Q}$ ist dann wie folgt gegeben

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

Algorithmus 2.11
Initialisierung: sei $\mathbf{A}_1 := \mathbf{A}$ und $\mathbf{a}_1$ die erste Spalte
Für $i = 1, 2, \ldots, n$:

① setze $\mathbf{x} = \mathbf{a}_i$ und bestimme gemäß (2.7)

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

② setze

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

vgl. (2.7).

③ berechne $\mathbf{w} = \beta \mathbf{A}_i^\star \mathbf{v}$

④ ersetze $\mathbf{A}_i$ durch $\mathbf{A}_i - \mathbf{v}\mathbf{w}^\star$

⑤ $\mathbf{A}_{i+1}$ bezeichne die rechte untere $(m-i+1) \times (n-i+1)$-Teilmatrix von $\mathbf{A}_i$, und $\mathbf{a}_{i+1}$ deren erste Spalte

Aufwand: Wir bilanzieren den Aufwand im i-ten Schritt:

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

Für den Gesamtaufwand ergibt sich daher

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

dies bedeutet dass der Aufwand etwa doppelt so hoch ist wie bei der LR-Zerlegung.
Die QR-Zerlegung kann wie die LR-Zerlegung zur Lösung eines nichtsingulären linearen Gleichungssystems $\mathbf{A}\mathbf{x} = \mathbf{b}$ (also $m = n$) verwendet werden. Dies geschieht in folgender Weise: Zerlege $\mathbf{A} = \mathbf{Q}\mathbf{R}$, und löse $\mathbf{Q}\mathbf{R}\mathbf{x} = \mathbf{b}$ durch Rückwärtssubstitution

$$\mathbf{R}\mathbf{x} = \underbrace{\mathbf{Q}^\star \mathbf{b}}_{\mathcal{O}(n^2) \text{ Operationen}}.$$

Bemerkung: Die QR-Zerlegung gehört zu den "stabilsten" Algorithmen in der numerischen linearen Algebra. Der Grund dafür ist, dass Orthogonaltransformationen keine

Fehlerverstärkung bringen, da $\operatorname{cond}_2 \mathbf{Q} = 1$. Die abschließende Rückwärtssubstitution hat die gleiche Kondition wie das Ausgangsproblem, da wegen $\|\mathbf{Qx}\|_2 = \|\mathbf{x}\|_2$ folgt

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

das heißt

$$\operatorname{cond}_2 \mathbf{R} = \operatorname{cond}_2 \mathbf{A}.$$

Beispiel 2.12 Gesucht ist die QR-Zerlegung von

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

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

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

Also ist

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

Somit folgt

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

woraus sich

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

ergibt. Dies bedeutet

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

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

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

Nun ist

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

Mit

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

und

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

folgt

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

Damit ergibt sich

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

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

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

Für $\mathbf{Q}$ erhalten wir schließlich

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

während $\mathbf{R}$ gegeben ist durch

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

△

Bemerkung: Algorithmus 2.11 bricht zusammen, wenn $\operatorname{rang}(\mathbf{A}) = p < n$. In diesem Fall muss man Spalten von $\mathbf{A}$ permutieren (ähnlich zur Pivotsuche) und erhält eine Faktorisierung der Art

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

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

2.3 Methode der Orthogonalisierung

Wir wollen nun die Methode der Orthogonalisierung herleiten, mit deren Hilfe das lineare Ausgleichsproblem ohne Verschlechterung der Kondition lösbar ist. Dazu betrachten wir

zunächst den Spezialfall, dass die Matrix $\mathbf{A} \in \mathbb{K}^{m\times n}$ mit $m \geq n = \operatorname{rang}(\mathbf{A})$ Rechtsdreiecksstruktur hat, das heißt

$$\mathbf{A} = \begin{bmatrix} \mathbf{R} \\ \mathbf{0} \end{bmatrix}$$

mit einer rechten oberen Dreiecksmatrix $\mathbf{R} \in \mathbb{K}^{n\times n}$. Der Vektor $\mathbf{b} \in \mathbb{K}^m$ sei analog zerlegt:

$$\mathbf{b} = \begin{bmatrix} \mathbf{c} \\ \mathbf{d} \end{bmatrix}, \qquad \mathbf{c} \in \mathbb{K}^n, \qquad \mathbf{d} \in \mathbb{K}^{m-n}.$$

Damit haben wir

$$\|\mathbf{b} - \mathbf{A}\mathbf{x}\|_2^2 = \left\| \begin{bmatrix} \mathbf{c} \\ \mathbf{d} \end{bmatrix} - \begin{bmatrix} \mathbf{R} \\ \mathbf{0} \end{bmatrix} \mathbf{x} \right\|_2^2 = \left\| \begin{bmatrix} \mathbf{c} - \mathbf{R}\mathbf{x} \\ \mathbf{d} - \mathbf{0}\mathbf{x} \end{bmatrix} \right\|_2^2 = \|\mathbf{c} - \mathbf{R}\mathbf{x}\|_2^2 + \|\mathbf{d}\|_2^2.$$

Ist die rechte obere Dreiecksmatrix $\mathbf{R}$ invertierbar, so ist die Lösung des Minimierungsproblems (2.3) offensichtlich gegeben durch $\mathbf{x} = \mathbf{R}^{-1}\mathbf{c} \in \mathbb{K}^n$. Die Größe des Residuums $\mathbf{r} = \mathbf{b} - \mathbf{A}\mathbf{x}$ können wir auch sofort ablesen, nämlich $\|\mathbf{r}\|_2 = \|\mathbf{d}\|_2$.

Satz 2.13 Sei $m \geq n$, $\mathbf{A} \in \mathbb{K}^{m\times n}$ eine Matrix mit linear unabhängigen Spalten und der QR-Zerlegung

$$\mathbf{A} = \mathbf{Q} \begin{bmatrix} \mathbf{R} \\ \mathbf{0} \end{bmatrix}, \qquad \mathbf{Q} \in \mathbb{K}^{m\times m}, \qquad \mathbf{R} \in \mathbb{K}^{n\times n}.$$

Für beliebiges $\mathbf{b} \in \mathbb{K}^m$ sei

$$\mathbf{Q}^\star \mathbf{b} = \begin{bmatrix} \mathbf{c} \\ \mathbf{d} \end{bmatrix}, \qquad \mathbf{c} \in \mathbb{K}^n, \qquad \mathbf{d} \in \mathbb{K}^{m-n}.$$

Dann ist die Lösung $\mathbf{x} \in \mathbb{K}^n$ des Ausgleichsproblems (2.3) eindeutig bestimmt durch

$$\mathbf{R}\mathbf{x} = \mathbf{c}.$$

Die Norm des Residuums $\mathbf{r} = \mathbf{b} - \mathbf{A}\mathbf{x}$ ist gegeben durch $\|\mathbf{r}\|_2 = \|\mathbf{d}\|_2$.

Beweis. Wegen $\|\mathbf{Q}\mathbf{z}\|_2 = \|\mathbf{z}\|_2$ für alle $\mathbf{z} \in \mathbb{K}^m$ folgt

$$\begin{aligned} \|\mathbf{b} - \mathbf{A}\mathbf{x}\|_2^2 = \|\mathbf{Q}(\mathbf{Q}^\star\mathbf{b} - \mathbf{Q}^\star\mathbf{A}\mathbf{x})\|_2^2 = \|\mathbf{Q}^\star\mathbf{b} - \mathbf{Q}^\star\mathbf{A}\mathbf{z}\|_2^2 &= \left\| \begin{bmatrix} \mathbf{c} \\ \mathbf{d} \end{bmatrix} - \begin{bmatrix} \mathbf{R} \\ \mathbf{0} \end{bmatrix} \mathbf{x} \right\|_2^2 \\ &= \|\mathbf{c} - \mathbf{R}\mathbf{x}\|_2^2 + \|\mathbf{d}\|_2^2. \end{aligned}$$

Da $\mathbf{A}$ vollen Rang besitzt, ist $\mathbf{R}$ nicht singulär. Damit wird dieser Ausdruck minimal für die eindeutig bestimmte Lösung $\mathbf{x} := \mathbf{R}^{-1}\mathbf{c} \in \mathbb{K}^n$. □

Beispiel 2.14 Wir führen das Vorgehen wieder mit den Zahlen aus Beispiel 2.1 vor. Die QR-Zerlegung der Matrix $\mathbf{A}$ aus (2.1) lautet $\mathbf{A} = \mathbf{Q}\mathbf{R}$ mit

$$\mathbf{Q} = \begin{bmatrix} -0.0361 & -0.0972 & -0.1684 & -0.3121 & -0.3493 & -0.4252 & -0.7488 \\ -0.1444 & -0.3064 & -0.3553 & -0.3943 & -0.3700 & -0.2804 & 0.6229 \\ -0.1805 & -0.3488 & 0.8855 & -0.1444 & -0.1433 & -0.1309 & 0.0562 \\ -0.3250 & -0.3813 & -0.1429 & 0.8069 & -0.1967 & -0.1934 & -0.0375 \\ -0.3611 & -0.3551 & -0.1412 & -0.1959 & 0.7984 & -0.2031 & -0.0802 \\ -0.4333 & -0.2618 & -0.1272 & -0.1905 & -0.2010 & 0.7844 & -0.1887 \\ -0.7222 & 0.6595 & 0.0693 & -0.0202 & -0.0628 & -0.1720 & 0.0690 \end{bmatrix}$$

und

$$\left[\frac{\mathbf{R}}{\mathbf{0}}\right] = \left[\begin{array}{cc} -2.7695 & 2.1027 \\ 0 & -0.73 \\ \hline 0 & 0 \\ 0 & 0 \\ 0 & 0 \\ 0 & 0 \\ 0 & 0 \end{array}\right].$$

Wir berechnen

$$\mathbf{Q}^\star\mathbf{y} = \mathbf{Q}^\star \begin{bmatrix} 0.96 \\ 3.26 \\ 3.82 \\ 5.11 \\ 5.2 \\ 5.05 \\ 0.58 \end{bmatrix} = \begin{bmatrix} -7.3404 \\ -7.1590 \\ -0.0037 \\ -0.0063 \\ 0.0058 \\ -0.0055 \\ 0.0046 \end{bmatrix}$$

Die gewünschte Partitionierung von $\mathbf{Q}^\star\mathbf{y}$ ist damit

$$\mathbf{Q}^\star\mathbf{y} = \begin{bmatrix} \mathbf{c} \\ \mathbf{d} \end{bmatrix}, \qquad \mathbf{c} = \begin{bmatrix} -7.3404 \\ -7.1590 \end{bmatrix}, \qquad \mathbf{d} = \begin{bmatrix} -0.0037 \\ -0.0063 \\ 0.0058 \\ -0.0055 \\ 0.0046 \end{bmatrix}.$$

Somit erhalten wir als das zu lösende Gleichungssystem

$$\mathbf{R}\mathbf{x} = \begin{bmatrix} -2.7695 & 2.1027 \\ 0 & -0.73 \end{bmatrix} \begin{bmatrix} v_y \\ g \end{bmatrix} = \begin{bmatrix} -7.3404 \\ -7.1590 \end{bmatrix} = \mathbf{c}.$$

Aufgelöst ergibt sich also

$$v_y = 10.1, \qquad g = 9.81.$$

Die Norm des Residuums ist übrigens

$$\|\mathbf{r}\|_2 = \|\mathbf{d}\|_2 = 0.118.$$

△

2.4 Singulärwertzerlegung und Pseudoinverse

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

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

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

$$\mathbf{u}_i := \frac{1}{\sqrt{\lambda_i}}\mathbf{A}\mathbf{v}_i, \qquad i = 1, 2 \ldots, p, \tag{2.9}$$

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

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

Die Vektoren $\{\mathbf{u}_i\}_{i=1}^p$ bilden folglich eine Orthonormalbasis von $\mathrm{img}(\mathbf{A})$ und können durch weitere $m - p$ Vektoren $\mathbf{u}_{p+1}, \ldots, \mathbf{u}_n$ zu einer von $\mathbb{K}^m$ ergänzt werden.
Es gilt wegen (2.9)

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

während

$$\mathbf{A}^\star \mathbf{u}_i = \mathbf{0}, \quad i = p+1, \ldots, m,$$

da

$$\{\mathbf{u}_{p+1}, \ldots, \mathbf{u}_m\} \subseteq \mathrm{img}(\mathbf{A})^\perp \overset{(2.5)}{=} \mathrm{kern}(\mathbf{A}^\star).$$

Wir fassen zusammen:

Satz 2.15 Zu jeder Matrix $\mathbf{A} \in \mathbb{K}^{m \times n}$ existiert eine **Singulärwertzerlegung** (SVD = singular value decomposition), das ist ein Tripel $(\{\sigma_i\}_{i=1}^p, \{\mathbf{u}_i\}_{i=1}^m, \{\mathbf{v}_i\}_{i=1}^n)$ mit

$$\begin{aligned} &\sigma_1 \ge \sigma_2 \ge \cdots \ge \sigma_p > 0, \\ &\mathbf{u}_i \in \mathbb{K}^m, \quad \mathbf{u}_i^\star \mathbf{u}_j = \delta_{i,j}, \quad i, j = 1, 2, \ldots, m, \\ &\mathbf{v}_i \in \mathbb{K}^n, \quad \mathbf{v}_i^\star \mathbf{v}_j = \delta_{i,j}, \quad i, j = 1, 2, \ldots, n, \end{aligned}$$

und

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

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

In Matrixnotation lässt sich Satz 2.15 kürzer schreiben. Wir setzen

$$\mathbf{U} := \big[\mathbf{u}_1, \mathbf{u}_2, \ldots, \mathbf{u}_m\big] \in \mathbb{K}^{m \times m}, \quad \mathbf{V} := \big[\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_n\big] \in \mathbb{K}^{n \times n},$$

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

und erhalten

$$\mathbf{A} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^\star, \quad \mathbf{A}^\star = \mathbf{V}\mathbf{\Sigma}^\star\mathbf{U}^\star. \tag{2.10}$$

Dabei sind die Matrizen $\mathbf{U}$ und $\mathbf{V}$ unitär.
Alternativ zu (2.10) gelten die Summendarstellungen

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

Definition 2.16 Sei $\mathbf{U}\mathbf{\Sigma}\mathbf{V}^\star$ die Singulärwertzerlegung von $\mathbf{A}$ und

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

Dann heißt die Matrix

$$\mathbf{A}^+ = \mathbf{V}\mathbf{\Sigma}^+\mathbf{U}^\star \in \mathbb{K}^{n\times m}$$

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

Auch für die Pseudoinverse gilt eine ensprechende Summendarstellung

$$\mathbf{A}^+ = \sum_{i=1}^{p} \frac{1}{\sigma_i}\mathbf{v}_i\mathbf{u}_i^\star, \tag{2.11}$$

aus der sofort folgt

$$\operatorname{kern}(\mathbf{A}^+) = \operatorname{kern}(\mathbf{A}^\star) \overset{(2.5)}{=} \operatorname{img}(\mathbf{A})^\perp, \quad \operatorname{img}(\mathbf{A}^+) = \operatorname{img}(\mathbf{A}^\star) \overset{(2.5)}{=} \operatorname{kern}(\mathbf{A})^\perp. \tag{2.12}$$

Beispiel 2.17 Für die Matrix

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

gilt

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

Hieraus folgt

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

und daher

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

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

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

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

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

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

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

ergibt sich $\alpha = 1/8$. △

Der Name "Pseudoinverse" beruht auf folgendem Resultat:

Satz 2.18 Die Pseudoinverse $\mathbf{A}^+$ von $\mathbf{A} \in \mathbb{K}^{m\times n}$ ist die eindeutige Lösung der vier Gleichungen

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

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

$$\mathbf{\Sigma\Sigma}^+ = \left[\begin{array}{ccc|c} \sigma_1 & & & \\ & \ddots & & \mathbf{0} \\ & & \sigma_p & \\ \hline & \mathbf{0} & & \mathbf{0} \end{array}\right] \left[\begin{array}{ccc|c} \sigma_1^{-1} & & & \\ & \ddots & & \mathbf{0} \\ & & \sigma_p^{-1} & \\ \hline & \mathbf{0} & & \mathbf{0} \end{array}\right] = \left[\begin{array}{c|c} \mathbf{I} & \mathbf{0} \\ \hline \mathbf{0} & \mathbf{0} \end{array}\right] \in \mathbb{K}^{m\times m} \tag{2.13}$$

folgen die ersten beiden Gleichungen

$$\begin{aligned} \mathbf{AA}^+\mathbf{A} &= \mathbf{U\Sigma V}^\star\mathbf{V\Sigma}^+\mathbf{U}^\star\mathbf{U\Sigma V}^\star = \mathbf{U\Sigma\Sigma}^+\mathbf{\Sigma V}^\star = \mathbf{U\Sigma V}^\star = \mathbf{A}, \\ \mathbf{A}^+\mathbf{AA}^+ &= \mathbf{V}^\star\mathbf{\Sigma}^+\mathbf{UU}^\star\mathbf{\Sigma V}^\star\mathbf{V\Sigma U}^\star = \mathbf{V}^\star\mathbf{\Sigma}^+\mathbf{\Sigma\Sigma}^+\mathbf{U}^\star = \mathbf{V\Sigma}^+\mathbf{U}^\star = \mathbf{A}^+. \end{aligned}$$

Weiter ist $\mathbf{AA}^+ = \mathbf{U\Sigma\Sigma}^+\mathbf{U}^\star$ und somit wegen (2.13) hermitesch. Entsprechend ist auch $\mathbf{A}^+\mathbf{A}$ hermitesch, womit auch die beiden letzten Gleichungen gezeigt sind.

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

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

Dies bedeutet, dass

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

Wegen $\mathbf{w}_i \in \operatorname{kern}(\mathbf{A}) \subseteq \operatorname{kern}(\mathbf{XA})$ und (iv) folgt für jedes $i = 1, 2, \ldots, p$

$$\begin{aligned} 0 = (\mathbf{XAw}_i)^\star\left(\frac{1}{\sigma_i}\mathbf{v}_i\right) = \mathbf{w}_i^\star\mathbf{XA}\left(\frac{1}{\sigma_i}\mathbf{v}_i\right) = \mathbf{w}_i^\star\mathbf{Xu}_i = \mathbf{w}_i^\star\left(\frac{1}{\sigma_i}\mathbf{v}_i + \mathbf{w}_i\right) \\ = \underbrace{\frac{1}{\sigma_i}\mathbf{w}_i^\star\mathbf{v}_i}_{\substack{=0,\text{ da} \\ \mathbf{v}_i\perp\operatorname{span}\{\mathbf{v}_{p+1},\ldots,\mathbf{v}_n\}=\operatorname{kern}(\mathbf{A})}} + \mathbf{w}_i^\star\mathbf{w}_i = \|\mathbf{w}_i\|^2. \end{aligned}$$

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

$$\mathbf{Xu}_i = \frac{1}{\sigma_i}\mathbf{v}_i, \quad i = 1, 2, \ldots, p. \tag{2.14}$$

Hieraus folgt die Inklusion

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

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

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

Aus (iii) folgt damit

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

Demnach ist

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

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

$$\mathbf{w}_i = \mathbf{Xu}_i = \mathbf{XAXu}_i = \mathbf{XAw}_i = \mathbf{0}, \quad i = p+1, \ldots, m. \tag{2.15}$$

Ein Vergleich von (2.11) mit (2.14) und (2.15) zeigt, dass $\mathbf{X}$ und $\mathbf{A}^+$ übereinstimmen, also $\mathbf{A}^+$ die einzige Lösung der Gleichungen (i)–(iv) ist. □

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

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

Satz 2.19 Der Vektor $\mathbf{A}^+\mathbf{b}$ ist die eindeutige Lösung des linearen Ausgleichsproblems (2.3) mit minimaler $\|\cdot\|_2$-Norm.

Beweis. Nach Satz 2.18 (iii) ist

$$\mathbf{AA}^+\mathbf{b} - \mathbf{b} \in \text{kern}(\mathbf{A}^+) \overset{(2.12)}{=} \text{img}(\mathbf{A})^\perp = \text{kern}(\mathbf{A}^\star).$$

Also erfüllt $\mathbf{A}^+\mathbf{b}$ die Normalengleichungen (2.4)

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

und ist daher eine Lösung des linearen Ausgleichsproblems.
Ist $\mathbf{z}$ eine zweite Lösung der Normalengleichungen, dann gilt gemäß Satz 2.3

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

Da $\mathbf{A}^+\mathbf{b} \in \text{img}(\mathbf{A}^+) \overset{(2.12)}{=} \text{kern}(\mathbf{A})^\perp$ haben wir $\mathbf{z} = \mathbf{A}^+\mathbf{b} - \mathbf{w}$ orthogonal zerlegt, und nach dem Satz des Pythagoras gilt

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

□

Korollar 2.20 Hat $\mathbf{A} \in \mathbb{K}^{m\times n}$ vollen Rang $\text{rang}(\mathbf{A}) = n \leq m$, dann gilt $\mathbf{A}^+ = (\mathbf{A}^\star\mathbf{A})^{-1}\mathbf{A}^\star$.

Beispiel 2.21 (Fortsetzung von Beispiel 2.17) Zu lösen sei das lineare Ausgleichsproblem $\|\mathbf{A}\mathbf{x}-\mathbf{b}\|_2 \to \min$ mit

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

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

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

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

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

△

2.5 CG- und CGLS-Verfahren

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

$$\phi(\mathbf{z}) = \frac{1}{2}(\mathbf{b}-\mathbf{A}\mathbf{z})^\star\mathbf{A}^{-1}(\mathbf{b}-\mathbf{A}\mathbf{z}) = \frac{1}{2}\mathbf{z}^\star\mathbf{A}\mathbf{z} - \operatorname{Re}(\mathbf{z}^\star\mathbf{b}) + \frac{1}{2}\mathbf{b}^\star\mathbf{A}^{-1}\mathbf{b} \geq 0$$

ist.

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

$$\phi(\mathbf{z}+\alpha\mathbf{d}) = \phi(\mathbf{z}) + \frac{|\alpha|^2}{2}\mathbf{d}^\star\mathbf{A}\mathbf{d} - \operatorname{Re}\left(\overline{\alpha}\mathbf{d}^\star(\mathbf{b}-\mathbf{A}\mathbf{z})\right) \to \min_{\alpha\in\mathbb{K}}.$$

Setzt man $\alpha = \xi + i\zeta$, so folgt

$$\phi(\mathbf{z}+\alpha\mathbf{d}) = \phi(\mathbf{z}) + \frac{\xi^2+\zeta^2}{2}\mathbf{d}^\star\mathbf{A}\mathbf{d} - \xi\operatorname{Re}\left(\mathbf{d}^\star(\mathbf{b}-\mathbf{A}\mathbf{z})\right) - \zeta\operatorname{Im}\left(\mathbf{d}^\star(\mathbf{b}-\mathbf{A}\mathbf{z})\right).$$

Aus

$$\frac{\partial\phi(z+\alpha\mathbf{d})}{\partial\xi} = \xi\mathbf{d}^\star\mathbf{A}\mathbf{d} - \operatorname{Re}\left(\mathbf{d}^\star(\mathbf{b}-\mathbf{A}\mathbf{z})\right) \stackrel{!}{=} 0,$$
$$\frac{\partial\phi(z+\alpha\mathbf{d})}{\partial\zeta} = \zeta\mathbf{d}^\star\mathbf{A}\mathbf{d} - \operatorname{Im}\left(\mathbf{d}^\star(\mathbf{b}-\mathbf{A}\mathbf{z})\right) \stackrel{!}{=} 0,$$

folgt daher

$$\alpha = \frac{\mathbf{d}^\star(\mathbf{b}-\mathbf{A}\mathbf{z})}{\mathbf{d}^\star\mathbf{A}\mathbf{d}}. \tag{2.16}$$

Lemma 2.22 Die Vektoren $\mathbf{d}_0, \mathbf{d}_1, \ldots, \mathbf{d}_k$ seien **A-konjugiert**, das heißt, es gelte $\mathbf{d}_i^\star \mathbf{A}\mathbf{d}_j = 0$ für alle $i \neq j$. Ist

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

und setzt man

$$\mathbf{x}_{k+1} = \mathbf{x}_k + \alpha_k \mathbf{d}_k, \qquad \alpha_k = \frac{\mathbf{d}_k^\star \mathbf{r}_k}{\mathbf{d}_k^\star \mathbf{A}\mathbf{d}_k}, \qquad \mathbf{r}_k = \mathbf{b} - \mathbf{A}\mathbf{x}_k, \tag{2.17}$$

so folgt

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

Beweis. Die $\mathbf{A}$-Konjugiertheit der Vektoren $\{\mathbf{d}_\ell\}$ impliziert $\mathbf{d}_k^\star \mathbf{A}(\mathbf{x}_{k-1} - \mathbf{x}_0) = 0$. Daher folgt

$$\begin{aligned}
\phi(\mathbf{x}_k + \alpha \mathbf{d}_k) &= \phi(\mathbf{x}_k) + \frac{|\alpha|^2}{2} \mathbf{d}_k^\star \mathbf{A}\mathbf{d}_k - \operatorname{Re}\left(\overline{\alpha} \mathbf{d}_k^\star (\mathbf{b} - \mathbf{A}\mathbf{x}_k)\right) \\
&= \phi(\mathbf{x}_k) + \frac{|\alpha|^2}{2} \mathbf{d}_k^\star \mathbf{A}\mathbf{d}_k - \operatorname{Re}\left(\overline{\alpha} \mathbf{d}_k^\star (\mathbf{b} - \mathbf{A}\mathbf{x}_0)\right) \\
&= \phi(\mathbf{x}_k) + \varphi(\alpha),
\end{aligned}$$

das heißt, das Minimierungsproblem entkoppelt. Da nach Voraussetzung $\mathbf{x}_k$ das Funktional ϕ über $\mathbf{x}_0 + \operatorname{span}\{\mathbf{d}_0, \mathbf{d}_1 \ldots, \mathbf{d}_{k-1}\}$ minimiert, wird das eindeutige Minimum angenommen, wenn $\varphi(\alpha)$ minimal ist. Dies ist aber nach (2.16) genau dann der Fall, wenn

$$\alpha_k = \frac{\mathbf{d}_k^\star (\mathbf{b} - \mathbf{A}\mathbf{x}_k)}{\mathbf{d}_k^\star \mathbf{A}\mathbf{d}_k} = \frac{\mathbf{d}_k^\star (\mathbf{b} - \mathbf{A}\mathbf{x}_0)}{\mathbf{d}_k^\star \mathbf{A}\mathbf{d}_k}. \tag{2.18}$$

□

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

$$\mathbf{r}_0 = \mathbf{b} - \mathbf{A}\mathbf{x}_0, \qquad \mathbf{r}_{k+1} = \mathbf{b} - \mathbf{A}\mathbf{x}_{k+1} \overset{(2.17)}{=} \mathbf{r}_k - \alpha_k \mathbf{A}\mathbf{d}_k, \quad k \geq 0, \tag{2.19}$$

erfüllt dann für alle $\ell < k$

$$\mathbf{d}_\ell^\star \mathbf{r}_k = \mathbf{d}_\ell^\star (\mathbf{b} - \mathbf{A}\mathbf{x}_k) = \mathbf{d}_\ell^\star \left(\mathbf{b} - \mathbf{A}\mathbf{x}_0 - \sum_{i=0}^{k-1} \alpha_i \mathbf{A}\mathbf{d}_i \right) = \mathbf{d}_\ell^\star (\mathbf{b} - \mathbf{A}\mathbf{x}_0) - \alpha_\ell \mathbf{d}_\ell^\star \mathbf{A}\mathbf{d}_\ell \overset{(2.18)}{=} 0. \tag{2.20}$$

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

Lemma 2.23 Für beliebiges $\mathbf{d}_0 = \mathbf{r}_0$ erzeugt die Rekursion

$$\mathbf{r}_{k+1} = \mathbf{r}_k - \alpha_k \mathbf{A}\mathbf{d}_k, \qquad \mathbf{d}_{k+1} = \mathbf{r}_{k+1} - \beta_k \mathbf{d}_k, \qquad \beta_k = \frac{\mathbf{d}_k^\star \mathbf{A}\mathbf{r}_{k+1}}{\mathbf{d}_k^\star \mathbf{A}\mathbf{d}_k} \tag{2.21}$$

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

Beweis. Sei

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

Wir zeigen zunächst induktiv, dass stets gilt

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

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

$$\mathbf{r}_k \overset{(2.21)}{=} \underbrace{\mathbf{r}_{k-1}}_{\in \mathcal{K}_k(\mathbf{A}, \mathbf{r}_0)} - \alpha_k \underbrace{\mathbf{A}\mathbf{d}_k}_{\in \mathcal{K}_{k+1}(\mathbf{A}, \mathbf{r}_0)} \in \mathcal{K}_{k+1}(\mathbf{A}, \mathbf{r}_0).$$

Gemäß (2.20) ist $\mathbf{r}_k \perp \operatorname{span}\{\mathbf{d}_0, \mathbf{d}_1, \ldots, \mathbf{d}_{k-1}\} = \mathcal{K}_k(\mathbf{A}, \mathbf{r}_0)$, dies bedeutet

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

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

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

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

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

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

$$\mathbf{d}_k^\star \mathbf{A}\mathbf{d}_{k+1} \overset{(2.21)}{=} \mathbf{d}_k^\star \mathbf{A}(\mathbf{r}_{k+1} - \beta_k \mathbf{d}_k) \overset{(2.21)}{=} \mathbf{d}_k^\star \mathbf{A}\mathbf{r}_{k+1} - \frac{\mathbf{d}_k^\star \mathbf{A}\mathbf{r}_{k+1}}{\mathbf{d}_k^\star \mathbf{A}\mathbf{d}_k} \mathbf{d}_k^\star \mathbf{A}\mathbf{d}_k = 0$$

und für alle $\ell < k$

$$\mathbf{d}_\ell^\star \mathbf{A}\mathbf{d}_{k+1} \overset{(2.21)}{=} \mathbf{d}_\ell^\star \mathbf{A}\mathbf{r}_{k+1} - \frac{\mathbf{d}_k^\star \mathbf{A}\mathbf{r}_{k+1}}{\mathbf{d}_k^\star \mathbf{A}\mathbf{d}_k} \underbrace{\mathbf{d}_\ell^\star \mathbf{A}\mathbf{d}_k}_{=0} = (\mathbf{A}\mathbf{d}_\ell)^\star \mathbf{r}_{k+1} = 0$$

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

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

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

$$\alpha_k \overset{(2.18)}{=} \frac{\mathbf{d}_k^\star \mathbf{r}_k}{\mathbf{d}_k^\star \mathbf{A}\mathbf{d}_k} \overset{(2.21)}{=} \frac{(\mathbf{r}_k - \beta_{k-1}\mathbf{d}_{k-1})^\star \mathbf{r}_k}{\mathbf{d}_k^\star \mathbf{A}\mathbf{d}_k} \overset{(2.20)}{=} \frac{\|\mathbf{r}_k\|_2^2}{\mathbf{d}_k^\star \mathbf{A}\mathbf{d}_k}. \tag{2.22}$$

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

$$\mathbf{d}_k^\star \mathbf{A}\mathbf{r}_{k+1} = (\mathbf{A}\mathbf{d}_k)^\star \mathbf{r}_{k+1} \overset{(2.21)}{=} \frac{1}{\overline{\alpha}_k}(\mathbf{r}_k - \mathbf{r}_{k+1})^\star \mathbf{r}_{k+1} \overset{(2.22)}{=} -\frac{\mathbf{d}_k^\star \mathbf{A}\mathbf{d}_k}{\|\mathbf{r}_k\|_2^2}\|\mathbf{r}_{k+1}\|_2^2$$

und damit

$$\beta_k = \frac{\mathbf{d}_k^\star \mathbf{A}\mathbf{r}_{k+1}}{\mathbf{d}_k^\star \mathbf{A}\mathbf{d}_k} = -\frac{\|\mathbf{r}_{k+1}\|_2^2}{\|\mathbf{r}_k\|_2^2}. \tag{2.23}$$

Kombination von (2.17) und (2.21)–(2.23) liefert schließlich:

Algorithmus 2.24 (CG-Verfahren)
input: Matrix $\mathbf{A} \in \mathbb{K}^{n\times n}$, rechte Seite $\mathbf{b} \in \mathbb{K}^n$ und Startnäherung $\mathbf{x}_0 \in \mathbb{K}^n$
output: Folge von Iterierten $\{\mathbf{x}_k\}_{k>0}$

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

② berechne

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

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

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

Satz 2.25 Die Iterierten $\{\mathbf{x}_k\}$ des CG-Verfahrens genügen bezüglich der **Energienorm**

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

der Fehlerabschätzung

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

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

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

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

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

und

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

Für ein beliebiges Polynom p folgt daher

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

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

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

und setzen

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

Wegen

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

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

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

mit

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

Zusammengefasst haben wir damit schließlich gezeigt, dass

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

□

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

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

$$\mathbf{A}^\star\mathbf{A}\mathbf{x} = \mathbf{A}^\star\mathbf{b}$$

zu lösen im Falle rechteckiger Matrizen $\mathbf{A} \in \mathbb{K}^{m\times n}$ mit $m \ge n = \operatorname{rang}(\mathbf{A})$. Allerdings soll das explizite Ausmultiplizieren der Matrix $\mathbf{A}^\star\mathbf{A}$ vermieden werden, da deren Kondition wesentlich schlechter ist als die von $\mathbf{A}$. Zudem ist mit der Matrix $\mathbf{A}$ nicht notwendig auch $\mathbf{A}^\star\mathbf{A}$ dünnbesetzt.
Es bezeichne $\mathbf{r}_k = \mathbf{b} - \mathbf{A}\mathbf{x}_k$ das Residuum und

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

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

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

Benutzt man ferner

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

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

Algorithmus 2.26 (CGLS-Verfahren)
input: Matrix $\mathbf{A} \in \mathbb{K}^{m\times n}$, rechte Seite $\mathbf{b} \in \mathbb{K}^m$ und Startnäherung $\mathbf{x}_0 \in \mathbb{K}^n$
output: Folge von Iterierten $\{\mathbf{x}_k\}_{k>0}$

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

② berechne

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

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

Proposition 2.27 Die k-te Iterierte $\mathbf{x}_k$ des CGLS-Verfahrens liegt im verschobenen Krylov-Raum

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

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

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

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

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

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

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

gilt für die Iterierten stets

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

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

3. Nichtlineare Ausgleichsprobleme

3.1 Gradientenverfahren

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

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

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

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

löst.

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

Zunächst wollen wir das *Gradientenverfahren* betrachten, das auch *Verfahren des steilsten Abstiegs* genannt wird. Die Idee dabei ist, die Iterierte $\mathbf{x}_k$ in Richtung des Antigradienten

$$-\nabla\phi(\mathbf{x}_k) = \big(\mathbf{f}'(\mathbf{x}_k)\big)^\star\big(\mathbf{y} - \mathbf{f}(\mathbf{x}_k)\big), \quad \mathbf{f}'(\mathbf{x}_k) = \begin{bmatrix} \frac{\partial f_1}{\partial z_1}(\mathbf{x}_k) & \frac{\partial f_1}{\partial z_2}(\mathbf{x}_k) & \ldots & \frac{\partial f_1}{\partial z_n}(\mathbf{x}_k) \\ \frac{\partial f_2}{\partial z_1}(\mathbf{x}_k) & \frac{\partial f_2}{\partial z_2}(\mathbf{x}_k) & \ldots & \frac{\partial f_2}{\partial z_n}(\mathbf{x}_k) \\ \vdots & \vdots & & \vdots \\ \frac{\partial f_m}{\partial z_1}(\mathbf{x}_k) & \frac{\partial f_m}{\partial z_2}(\mathbf{x}_k) & \ldots & \frac{\partial f_m}{\partial z_n}(\mathbf{x}_k) \end{bmatrix}$$

so aufzudatieren

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

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

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

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

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

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

Beweis. Die Funktion φ ist stetig differenzierbar und es gilt

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

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

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

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

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

③ solange

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

setze $\alpha_k := \alpha_k/2$

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

⑤ erhöhe $k := k+1$ und gehe nach ②

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

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

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

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

Satz 3.3 Es sei $D \subset \mathbb{K}^n$ eine offene Menge, in der $\mathbf{f}$ stetig differenzierbar und $\mathbf{f}'$ zudem Lipschitz-stetig ist. Ferner sei neben $\mathbf{x}_0$ auch die gesamte Niveaumenge $\{\mathbf{z} \in \mathbb{K}^n : \phi(\mathbf{z}) \le \phi(\mathbf{x}_0)\}$ in D enthalten. Dann gilt für die Iterierten $\{\mathbf{x}_k\}_{k\ge0}$ aus Algorithmus 3.2

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

Beweis. Da D die gesamte Niveaumenge enthält, ist sichergestellt, dass die Iterierten $\{\mathbf{x}_k\}_{k\ge0}$ alle in D enthalten ist. Nach Konstruktion ist dann die Folge $\{\phi(\mathbf{x}_k)\}_{k\ge0}$ monoton fallend und nach unten beschränkt. Daher folgt aus der Armijo-Goldstein-Bedingung (3.3), dass

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

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

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

Es verbleibt zu zeigen, dass $\alpha_k > \varepsilon$ für ein $\varepsilon > 0$.
Für festes k ist aufgrund von Algorithmus 3.2 $\alpha_k = 1$ oder die Armijo-Goldstein-Bedingung ist für $2\alpha_k$ verletzt:

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

mit dem Taylor-Restglied

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

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

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

dies bedeutet

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

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

Beachte: Satz 3.3 besagt nicht, dass die Folge $\{\mathbf{x}_k\}_{k\geq 0}$ selber konvergiert. Selbst wenn die Folge $\{\mathbf{x}_k\}_{k\geq 0}$ konvergiert, braucht der Grenzwert darüberhinaus kein Minimum von ϕ zu sein.

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

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

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

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

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

3.2 Gauß-Newton-Verfahren

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

$$\nabla\phi(\mathbf{z}) = -\big(\mathbf{f}'(\mathbf{z})\big)^\star\big(\mathbf{y} - \mathbf{f}(\mathbf{z})\big) \stackrel{!}{=} \mathbf{0}$$

lösen, was der Iteration

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

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

$$\phi''(\mathbf{z}) = \big(\mathbf{f}'(\mathbf{z})\big)^{\star}\mathbf{f}'(\mathbf{x}) - \big(\mathbf{y} - \mathbf{f}(\mathbf{z})\big)^{\star}\mathbf{f}''(\mathbf{z}).$$

In der Praxis will man die Berechnung des Tensors $\mathbf{f}''(\mathbf{z}) \in \mathbb{K}^{m\times(n\times n)}$ jedoch vermeiden. Daher vernachlässigt man den Term $\big(\mathbf{y} - \mathbf{f}(\mathbf{z})\big)^{\star}\mathbf{f}''(\mathbf{z})$ und erhält das *Gauß-Newton-Verfahren*.

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

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

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

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

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

$$\big(\mathbf{f}'(\mathbf{x}_k)\big)^{\star}\mathbf{f}'(\mathbf{x}_k)\mathbf{d}_k = \big(\mathbf{f}'(\mathbf{x}_k)\big)^{\star}\mathbf{r}_k$$

lösen. Dies führt auf folgenden Algorithmus:

Algorithmus 3.5 (Gauß-Newton–Verfahren)
input: Funktion $\mathbf{f} : \mathbb{K}^n \to \mathbb{K}^m$, Datenvektor $\mathbf{y} \in \mathbb{K}^m$ und Startnäherung $\mathbf{x}_0 \in \mathbb{K}^n$
output: Folge von Iterierten $\{\mathbf{x}_k\}_{k>0}$

① Initialisierung: setze $k = 0$

② löse die Normalengleichungen

$$\big(\mathbf{f}'(\mathbf{x}_k)\big)^{\star}\mathbf{f}'(\mathbf{x}_k)\mathbf{d}_k = \big(\mathbf{f}'(\mathbf{x}_k)\big)^{\star}\big(\mathbf{y} - \mathbf{f}(\mathbf{x}_k)\big) \tag{3.5}$$

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

④ erhöhe $k := k + 1$ und gehe nach ②

Satz 3.6 Sei $D \subset \mathbb{K}^n$ offen und $\mathbf{f} : D \to \mathbb{K}^m$ eine stetig differenzierbare Abbildung. Das Minimierungsproblem (3.1) habe eine Lösung $\mathbf{x} \in D$ mit $\operatorname{rang}\big(\mathbf{f}'(\mathbf{x})\big) = n \le m$. Sei $\lambda > 0$ der kleinste Eigenwert von $\big(\mathbf{f}'(\mathbf{x})\big)^{\star}\mathbf{f}'(\mathbf{x})$. Ferner gelte die Lipschitz-Bedingung

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

und

$$\big\|\big(\mathbf{f}'(\mathbf{z}) - \mathbf{f}'(\mathbf{x})\big)^{\star}\big(\mathbf{y} - \mathbf{f}(\mathbf{x})\big)\big\|_2 \le \beta\|\mathbf{z} - \mathbf{x}\|_2 \tag{3.7}$$

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

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

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

Wegen $\operatorname{rang}\big(\mathbf{f}'(\mathbf{x})\big) = n$ ist die Matrix $\big(\mathbf{f}'(\mathbf{x})\big)^\star \mathbf{f}'(\mathbf{x})$ regulär mit

$$\left\| \Big(\big(\mathbf{f}'(\mathbf{x})\big)^\star \mathbf{f}'(\mathbf{x}) \Big)^{-1} \right\|_2 = \frac{1}{\lambda}. \tag{3.8}$$

Daher gibt es zu beliebigem $\delta > 1$ ein $\varepsilon_2 > 0$ derart, dass $\big(\mathbf{f}'(\mathbf{z})\big)^\star \mathbf{f}'(\mathbf{z})$ regulär ist und

$$\left\| \Big(\big(\mathbf{f}'(\mathbf{z})\big)^\star \mathbf{f}'(\mathbf{z}) \Big)^{-1} \right\|_2 = \frac{\delta}{\lambda} \quad \text{für alle} \quad \mathbf{z} \in B_{\varepsilon_2}(\mathbf{x}). \tag{3.9}$$

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

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

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

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

Wir setzen nun

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

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

$$\begin{aligned}
\mathbf{x}_{k+1} - \mathbf{x} &= \mathbf{x}_k + \mathbf{d}_k - \mathbf{x} \\
&= \Big(\big(\mathbf{f}'(\mathbf{x}_k)\big)^\star \mathbf{f}'(\mathbf{x}_k) \Big)^{-1} \Big[\big(\mathbf{f}'(\mathbf{x}_k)\big)^\star \big(\mathbf{y} - \mathbf{f}(\mathbf{x}_k)\big) - \big(\mathbf{f}'(\mathbf{x}_k)\big)^\star \mathbf{f}'(\mathbf{x}_k)(\mathbf{x} - \mathbf{x}_k) \Big] \\
&= \Big(\big(\mathbf{f}'(\mathbf{x}_k)\big)^\star \mathbf{f}'(\mathbf{x}_k) \Big)^{-1} \Big[\big(\mathbf{f}'(\mathbf{x}_k)\big)^\star \big(\mathbf{y} - \mathbf{f}(\mathbf{x})\big) \\
&\qquad + \big(\mathbf{f}'(\mathbf{x}_k)\big)^\star \big(\mathbf{f}(\mathbf{x}) - \mathbf{f}(\mathbf{x}_k) - \mathbf{f}'(\mathbf{x}_k)(\mathbf{x} - \mathbf{x}_k)\big) \Big].
\end{aligned}$$

Hieraus folgt dann mit (3.8) und (3.9)

$$\begin{aligned}
\|\mathbf{x}_{k+1} - \mathbf{x}\|_2 &\leq \left\| \Big(\big(\mathbf{f}'(\mathbf{x}_k)\big)^\star \mathbf{f}'(\mathbf{x}_k) \Big)^{-1} \right\|_2 \Big[\left\| \big(\mathbf{f}'(\mathbf{x}_k)\big)^\star \big(\mathbf{y} - \mathbf{f}(\mathbf{x})\big) \right\|_2 \\
&\qquad + \|\mathbf{f}'(\mathbf{x}_k)\|_2 \|\mathbf{f}(\mathbf{x}) - \mathbf{f}(\mathbf{x}_k) - \mathbf{f}'(\mathbf{x}_k)(\mathbf{x} - \mathbf{x}_k)\|_2 \Big] \\
&\leq \frac{\delta}{\lambda} \left[\left\| \big(\mathbf{f}'(\mathbf{x}_k)\big)^\star \big(\mathbf{y} - \mathbf{f}(\mathbf{x})\big) \right\|_2 + \frac{\alpha\gamma}{2} \|\mathbf{x}_k - \mathbf{x}\|_2^2 \right].
\end{aligned} \tag{3.11}$$

Wegen $\mathbf{0} = \nabla\phi(\mathbf{x}) = -\big(\mathbf{f}'(\mathbf{x})\big)^\star\big(\mathbf{y} - \mathbf{f}(\mathbf{x})\big)$ ergibt sich mit (3.7)

$$\big\|\big(\mathbf{f}'(\mathbf{x}_k)\big)^\star\big(\mathbf{y} - \mathbf{f}(\mathbf{x})\big)\big\|_2 = \big\|\big(\mathbf{f}'(\mathbf{x}_k) - \mathbf{f}'(\mathbf{x})\big)^\star\big(\mathbf{y} - \mathbf{f}(\mathbf{x})\big)\big\|_2 \le \beta\|\mathbf{x}_k - \mathbf{x}\|_2,$$

dies bedeutet,

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

Wegen (3.10) ist die Konstante $(\lambda+\beta\delta)/(2\lambda) < 1$, das heißt, alle Iterierten $\{\mathbf{x}_k\}_{k\ge0}$ liegen in $B_\varepsilon(\mathbf{x})$ und konvergieren mindestens linear gegen $\mathbf{x}$. □

Bemerkung: Die Voraussetzung (3.7) besagt im Prinzip, dass das Residuum $\mathbf{r}(\mathbf{x}) = \mathbf{y} - \mathbf{f}(\mathbf{x})$ klein genug sein soll. Dies sieht man insbesondere, wenn man sie durch die stärkere Bedingung

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

ersetzt. Dann folgt nämlich (3.7):

$$\big\|\big(\mathbf{f}'(\mathbf{z}) - \mathbf{f}'(\mathbf{x})\big)^\star\big(\mathbf{y} - \mathbf{f}(\mathbf{x})\big)\big\|_2 \le \underbrace{\|\mathbf{f}'(\mathbf{z}) - \mathbf{f}'(\mathbf{x})\|_2}_{\le\alpha\|\mathbf{z}-\mathbf{x}\|_2}\underbrace{\|\mathbf{y} - \mathbf{f}(\mathbf{x})\|_2}_{<\lambda/\alpha} < \lambda\|\mathbf{z} - \mathbf{x}\|_2.$$

Korollar 3.7 Zusätzlich zu den Voraussetzungen aus Satz 3.6 gelte $\mathbf{f}(\mathbf{x}) = \mathbf{y}$. Dann existiert ein $\varepsilon > 0$, so dass für jeden Startvektor $\mathbf{x}_0 \in B_\varepsilon(\mathbf{x})$ die Folge der Iterierten $\{\mathbf{x}_k\}_{k\ge0}$ quadratisch gegen $\mathbf{x}$ konvergiert.

Beweis. Aus Satz 3.6 folgt die lineare Konvergenz der Iterierten $\{\mathbf{x}_k\}_{k\ge0}$ gegen $\mathbf{x}$. Zum Nachweis der quadratischen Konvergenz bemerken wir, dass aufgrund der Voraussetzung (3.7) mit $\beta = 0$ gilt. Daher folgt aus (3.11)

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

□

3.3 Levenberg-Marquardt-Verfahren

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

$$\min_{\mathbf{h}\in\mathbb{K}^n:\|\mathbf{h}\|_2\le\Delta}\psi(\mathbf{h}) = \min_{\mathbf{h}\in\mathbb{K}^n:\|\mathbf{h}\|_2\le\Delta}\|\mathbf{r}_k - \mathbf{f}'(\mathbf{x_k})\mathbf{h}\|_2^2, \tag{3.12}$$

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

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

$$\nabla\psi(\mathbf{d}_k) = \big(\mathbf{f}'(\mathbf{x}_k)\big)^\star\big(\mathbf{f}'(\mathbf{x}_k)\mathbf{d}_k - \mathbf{r}_k\big) = \mathbf{0}. \tag{3.13}$$

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

$$\nabla\psi(\mathbf{d}_k) = \big(\mathbf{f}'(\mathbf{x}_k)\big)^\star\big(\mathbf{f}'(\mathbf{x}_k)\mathbf{d}_k - \mathbf{r}_k\big) = -\lambda_k\mathbf{d}_k \quad \text{für ein} \quad \lambda_k > 0. \tag{3.14}$$

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

Lemma 3.8 Die Lösung $\mathbf{d}_k$ des restringierten Problems (3.12) genügt der Gleichung

$$\Big(\big(\mathbf{f}'(\mathbf{x}_k)\big)^\star\mathbf{f}'(\mathbf{x}_k) + \lambda_k\mathbf{I}\Big)\mathbf{d}_k = \big(\mathbf{f}'(\mathbf{x}_k)\big)^\star\mathbf{r}_k \tag{3.15}$$

für ein $\lambda_k \geq 0$. Dabei ist der Wert λ_k genau dann positiv, wenn $\|\mathbf{d}_k\|_2 = \Delta > 0$ gilt.

Der Vorteil des regularisierten Systems (3.15) liegt darin, dass es stets eindeutig lösbar ist, sofern $\lambda_k > 0$ ist. Die zugehörige Lösung

$$\mathbf{d}_k = \Big(\big(\mathbf{f}'(\mathbf{x}_k)\big)^\star\mathbf{f}'(\mathbf{x}_k) + \lambda_k\mathbf{I}\Big)^{-1}\big(\mathbf{f}'(\mathbf{x}_k)\big)^\star\mathbf{r}_k = -\Big(\big(\mathbf{f}'(\mathbf{x}_k)\big)^\star\mathbf{f}'(\mathbf{x}_k) + \lambda_k\mathbf{I}\Big)^{-1}\nabla\phi(\mathbf{x}_k)$$

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

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

Wir müssen uns noch ein Kriterium überlegen, wie wir Δ wählen. Eine neue Näherung $\mathbf{x}_{k+1} = \mathbf{x}_k + \mathbf{d}_k$ können wir anhand der Armijo-Goldstein-Bedingung (vergleiche (3.3)) bewerten:

$$\mu = \frac{\phi(\mathbf{x}_k + \mathbf{d}_k) - \phi(\mathbf{x}_k)}{\mathbf{d}_k^\star\nabla\phi(\mathbf{x}_k)} = \frac{1}{2}\frac{\|\mathbf{y} - \mathbf{f}(\mathbf{x}_k)\|_2^2 - \|\mathbf{y} - \mathbf{f}(\mathbf{x}_k + \mathbf{d}_k)\|_2^2}{\mathbf{d}_k^\star\big(\mathbf{f}'(\mathbf{x}_k)\big)^\star\big(\mathbf{y} - \mathbf{f}(\mathbf{x}_k)\big)}.$$

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

Algorithmus 3.9 (Levenberg-Marquardt-Verfahren)
input: Funktion $\mathbf{f} : \mathbb{K}^n \to \mathbb{K}^m$, Datenvektor $\mathbf{y} \in \mathbb{K}^m$ und Startnäherung $\mathbf{x}_0 \in \mathbb{K}^n$
output: Folge von Iterierten $\{\mathbf{x}_k\}_{k>0}$

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

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

③ berechne

$$\mu_k = \frac{1}{2}\frac{\|\mathbf{y}-\mathbf{f}(\mathbf{x}_k+\mathbf{d}_k)\|_2^2 - \|\mathbf{y}-\mathbf{f}(\mathbf{x}_k)\|_2^2}{\mathbf{d}_k^\star\big(\mathbf{f}'(\mathbf{x}_k)\big)^\star\big(\mathbf{y}-\mathbf{f}(\mathbf{x}_k)\big)} \tag{3.16}$$

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

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

⑥ erhöhe $k := k+1$ und gehe nach ②

Satz 3.10 Es sei $D \subset \mathbb{K}^n$ eine kompakte Menge, in der $\mathbf{f}$ stetig differenzierbar und $\mathbf{f}'$ zudem Lipschitz-stetig ist. Ferner sei neben $\mathbf{x}_0$ auch die gesamte Niveaumenge $\{\mathbf{z} \in \mathbb{K}^n : \phi(\mathbf{z}) \le \phi(\mathbf{x}_0)\}$ in D enthalten. Dann gilt für die Iterierten $\{\mathbf{x}_k\}_{k\ge 0}$ aus Algorithmus 3.9

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

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

$$\mathbf{d}_k^\star\Big(\big(\mathbf{f}'(\mathbf{x}_k)\big)^\star\mathbf{f}'(\mathbf{x}_k) + \lambda_k\mathbf{I}\Big)\mathbf{d}_k = \mathbf{d}_k^\star\big(\mathbf{f}'(\mathbf{x}_k)\big)^\star\mathbf{r}_k \le \|\mathbf{d}_k\|_2\big\|\big(\mathbf{f}'(\mathbf{x}_k)\big)^\star\mathbf{r}_k\big\|_2.$$

Da $\big(\mathbf{f}'(\mathbf{x}_k)\big)^\star\mathbf{f}'(\mathbf{x}_k)$ positiv semidefinit ist, kann die linke Seite nach unten durch λ_k abgeschätzt werden, so dass folgt

$$\lambda_k \le \frac{\big\|\big(\mathbf{f}'(\mathbf{x}_k)\big)^\star\mathbf{r}_k\big\|_2}{\Delta_k} = \frac{\|\nabla\phi(\mathbf{x}_k)\|_2}{\Delta_k}. \tag{3.17}$$

(*ii*) Als nächstes leiten wir eine untere Schranke für den Nenner ν_k in (3.16) her. Im Fall $\lambda_k > 0$ ist $\big(\mathbf{f}'(\mathbf{x}_k)\big)^\star\mathbf{f}'(\mathbf{x}_k) + \lambda_k\mathbf{I}$ positiv definit, weshalb eine Cholesky-Zerlegung $\mathbf{L}\mathbf{L}^\star$ existiert. Dabei gilt

$$\|\mathbf{L}^\star\mathbf{L}\|_2 = \|\mathbf{L}\mathbf{L}^\star\|_2 = \big\|\big(\mathbf{f}'(\mathbf{x}_k)\big)^\star\mathbf{f}'(\mathbf{x}_k) + \lambda_k\mathbf{I}\big\| = \underbrace{\|\mathbf{f}'(\mathbf{x}_k)\|_2^2}_{\le c \text{ für alle } \mathbf{z}\in D} + \lambda_k \overset{(3.17)}{\le} c + \frac{\|\nabla\phi(\mathbf{x}_k)\|_2}{\Delta_k}.$$

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

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

Im Fall $\lambda_k = 0$ folgt

$$\nu_k = \mathbf{r}_k^\star\mathbf{f}'(\mathbf{x}_k)\big(\mathbf{f}'(\mathbf{x}_k)\big)^+\mathbf{r}_k = \|\mathbf{P}_k\mathbf{r}_k\|_2^2,$$

wobei $\mathbf{P}_k = \mathbf{f}'(\mathbf{x}_k)\big(\mathbf{f}'(\mathbf{x}_k)\big)^+$ den Orthogonalprojektor auf $\operatorname{img}\big(\mathbf{f}'(\mathbf{x}_k)\big)$ bezeichnet. Wegen $\operatorname{img}\big(\mathbf{f}'(\mathbf{x}_k)\big) = \operatorname{kern}\Big(\big(\mathbf{f}'(\mathbf{x}_k)\big)^\star\Big)^\perp$ folgt daher

$$\|\nabla\phi(\mathbf{x}_k)\|_2^2 = \big\|\big(\mathbf{f}'(\mathbf{x}_k)\big)^\star\mathbf{r}_k\big\|_2^2 = \big\|\big(\mathbf{f}'(\mathbf{x}_k)\big)^\star\mathbf{P}_k\mathbf{r}_k\big\|_2^2 \le \big\|\big(\mathbf{f}'(\mathbf{x}_k)\big)^\star\big\|_2^2\|\mathbf{P}_k\mathbf{r}_k\|_2^2 \le c\nu_k,$$

das heißt, (3.18) ist auch im Fall $\lambda_k = 0$ gültig.
(*iii*) Wir beweisen nun, dass die rechte Seite von (3.18) gegen Null konvergiert. Bei erfolgreichem Iterationsschritt ist $\mu_k > \mu^-$ und aus (3.16) und (3.18) folgt

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

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

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

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

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

Aus (3.20) ergibt sich damit unmittelbar

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

Taylor-Entwicklung von μ_k liefert jedoch

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

Demnach existiert ein $M(\varepsilon) \geq K(\varepsilon)$, so dass $\mu_k > \mu^+$ für alle $k \geq M(\varepsilon)$. Ab dem $M(\varepsilon)$-ten Schritt wird folglich Δ_k in jedem Schritt von Algorithmus 3.9 verdoppelt, was jedoch im Widerspruch zu (3.21) steht.
(*v*) Wir beweisen nun die Aussage des Satzes. Dazu nehmen wir an, dass eine Teilfolge von $\{\|\nabla\phi(\mathbf{x}_k)\|_2\}_{k\geq 0}$ nicht gegen Null konvergiert. Nach Aussage (*iv*) existiert dann ein $\varepsilon > 0$ und zwei Indizes $\ell < m$, so dass

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

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

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

wobei $L > 1$ eine Lipschitz-Konstante von $\nabla\phi$ in D bezeichne. Wegen $\|\mathbf{x}_{k+1} - \mathbf{x}_k\|_2 \leq \Delta_k$ folgt aus (3.19) dass

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

Summation ergibt

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

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

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

und insgesamt

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

gilt. Dies ergibt

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

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

Wir kommen nun zur Implementierung. Um das restringierte Minimierungsproblem (3.12) zu lösen, berechnen wir zunächst die Lösung $\mathbf{d}_k$ bezüglich des unrestringierten Minimierungsproblems und akzeptieren den Schritt, falls $\|\mathbf{d}_k\|_2 \leq \Delta_k$. Ist hingegen $\|\mathbf{d}_k\|_2 > \Delta_k$, so wissen wir, dass das Minimum von (3.12) auf dem Rand liegt. Wir suchen dann dasjenige Tupel $(\lambda_k, \mathbf{d}_k)$, das (3.15) und $\|\mathbf{d}_k\|_2 = \Delta_k$ löst.

Es bezeichne $z_1 \geq \cdots \geq z_n \geq 0$ die Eigenwerte von $\big(\mathbf{f}'(\mathbf{x}_k)\big)^\star \mathbf{f}'(\mathbf{x}_k)$ und $\{\mathbf{v}_i\}_{i=1}^n$ die zugehörigen orthonormalen Eigenvektoren. Entwickeln wir die rechte Seite von (3.15) in diese Eigenbasis

$$\big(\mathbf{f}'(\mathbf{x}_k)\big)^\star \mathbf{r}_k = \sum_{i=1}^n \xi_i \mathbf{v}_i,$$

dann folgt

$$\mathbf{d}_k(\lambda_k) = \Big(\big(\mathbf{f}'(\mathbf{x}_k)\big)^\star \mathbf{f}'(\mathbf{x}_k) + \lambda_k \mathbf{I}\Big)^{-1} \sum_{i=1}^n \xi_i \mathbf{v}_i = \sum_{i=1}^n \frac{\xi_i}{z_i + \lambda_k} \mathbf{v}_i.$$

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

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

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

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

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

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

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

$$r'(\lambda) = -2\mathbf{r}_k^\star \mathbf{f}'(\mathbf{x}_k)\Big(\big(\mathbf{f}'(\mathbf{x}_k)\big)^\star \mathbf{f}'(\mathbf{x}_k) + \lambda\mathbf{I}\Big)^{-3}\big(\mathbf{f}'(\mathbf{x}_k)\big)^\star \mathbf{r}_k = -2\mathbf{d}_k(\lambda)^\star \mathbf{g}(\lambda)$$

$$\text{mit} \quad \Big(\big(\mathbf{f}'(\mathbf{x}_k)\big)^\star \mathbf{f}'(\mathbf{x}_k) + \lambda\mathbf{I}\Big)\mathbf{g}(\lambda) = \mathbf{d}_k(\lambda)$$

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

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

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

4. Eigenwerte

4.1 Eigenwerteinschließungen

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

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

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

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

Satz 4.1 (Gerschgorin) Sei $\mathbf{A} = [a_{i,j}] \in \mathbb{K}^{n\times n}$ und λ ein beliebiger Eigenwert von $\mathbf{A}$. Dann gilt

$$\lambda \in \bigcup_{i=1}^{n} K_i = \bigcup_{i=1}^{n} \left\{ z \in \mathbb{C} : |z - a_{i,i}| \leq \sum_{\substack{j=1 \\ j\neq i}}^{n} |a_{i,j}| \right\}. \tag{4.1}$$

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

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

und weiter

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

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

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

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

beziehungsweise

$$\lambda \in \bigcup_{i=1}^{n} \overline{K}_i := \bigcup_{i=1}^{n} \left\{ z \in \mathbb{C} : |z - a_{i,i}| \leq \sum_{\substack{j=1 \\ j \neq i}}^{n} |a_{j,i}| \right\}. \tag{4.2}$$

Dies ist der Satz von Gerschgorin angewendet auf $\mathbf{A}^\star$.
Ist $\mathbf{A}$ eine beliebige $(n \times n)$-Matrix, dann ist $(\mathbf{A} + \mathbf{A}^\star)/2$ hermitesch und $(\mathbf{A} - \mathbf{A}^\star)/2$ *schiefhermitesch*, dies bedeutet

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

Offensichtlich gilt

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

Definition 4.2 Unter dem **Wertebereich** einer Matrix $\mathbf{A} \in \mathbb{K}^{n \times n}$ versteht man die Menge aller **Rayleigh-Quotienten** $\mathbf{x}^\star \mathbf{A} \mathbf{x} / (\mathbf{x}^\star \mathbf{x})$ mit $\mathbf{x} \in \mathbb{C}^n \setminus \{\mathbf{0}\}$:

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

Lemma 4.3

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

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

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

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

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

eine stetige Kurve, die ξ_0 mit ξ_1 verbindet.

2. Dies haben wir bereits im Beweis von Satz 1.11 gesehen.

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

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

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

□

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

Satz 4.4 (Satz von Bendixson) Sei $\mathbf{A} \in \mathbb{K}^{n\times n}$ beliebig. Dann liegt das Spektrum von $\mathbf{A}$ in dem Rechteck

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

Beweis. Wir zeigen

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

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

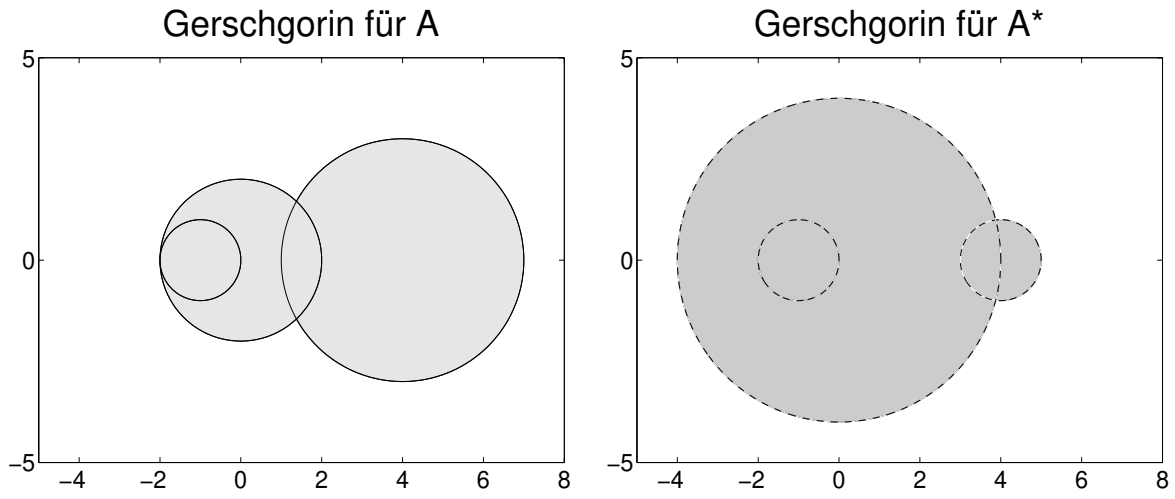
$$\begin{aligned}\frac{\mathbf{x}^\star\mathbf{A}\mathbf{x}}{\mathbf{x}^\star\mathbf{x}} &= \frac{\mathbf{x}^\star\left[\frac{1}{2}(\mathbf{A}+\mathbf{A}^\star)+\frac{1}{2}(\mathbf{A}-\mathbf{A}^\star)\right]\mathbf{x}}{\mathbf{x}^\star\mathbf{x}} \\ &= \frac{1}{2}\frac{\mathbf{x}^\star(\mathbf{A}+\mathbf{A}^\star)\mathbf{x}}{\mathbf{x}^\star\mathbf{x}}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

□

Beispiel 4.5 Für die Matrix

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

ergeben sich nach (4.1) beziehungsweise (4.2) die folgenden Einschließungen:



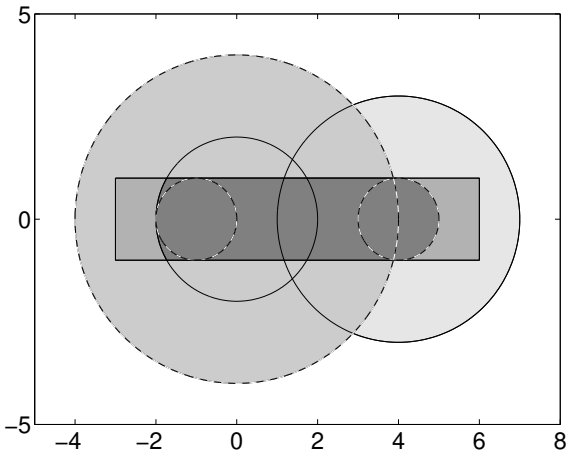
Der symmetrische und schiefsymmetrische Anteil von $\mathbf{A}$ ist

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

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

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

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



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

4.2 Kondition des Eigenwertproblems

Betrachte

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

Entwickelt man die Determinante nach der letzten Spalte, so gilt

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

Definition 4.6 Die Matrix $\mathbf{A}$ heißt **Frobenius-Begleitmatrix** von $p(\lambda)$. Die Eigenwerte von $\mathbf{A}$ sind die Nullstellen von p.

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

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

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

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

Allerdings haben die Eigenwerte der Begleitmatrizen den Abstand

$$|\triangle\lambda| = |\widehat{\lambda} - \lambda_k| = \varepsilon^{1/n}. \tag{4.3}$$

Für $a \neq 0$ folgt daher

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

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

Man kann jedoch zeigen, dass die Eigenwerte stetig von den Einträgen der Matrix abhängen. Der gefundene Exponent $1/n$ in (4.3) ist schlimmstmöglich.

Definition 4.7 Eine Matrix $\mathbf{A}$ heißt **diagonalisierbar**, falls es eine Basis aus Eigenvektoren gibt. Sind $(\lambda_1, \mathbf{v}_1), (\lambda_2, \mathbf{v}_2), \dots, (\lambda_n, \mathbf{v}_n)$ die Eigenpaare, dann gilt

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

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

Satz 4.8 (Bauer und Fike) Sei $\mathbf{A} = \mathbf{V}\mathbf{D}\mathbf{V}^{-1}$ diagonalisierbar und λ ein Eigenwert von $\mathbf{A} + \mathbf{E}$. Dann existiert ein Eigenwert λ_i von $\mathbf{A}$ mit

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

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

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

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

und daher

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

Folglich ist

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

□

Definition 4.9 Eine Matrix $\mathbf{A}$ heißt **normal**, falls gilt $\mathbf{A}\mathbf{A}^\star = \mathbf{A}^\star\mathbf{A}$. Insbesondere sind normale Matrizen diagonalisierbar mit $\mathbf{V}^{-1} = \mathbf{V}^\star$.

Bemerkung: Hermitesche Matrizen sind normal.

Korollar 4.10 Sei $\mathbf{E}$ eine beliebige Matrix. Ist $\mathbf{A}$ normal und λ ein Eigenwert von $\mathbf{A}+\mathbf{E}$, dann existiert ein $\lambda_i \in \sigma(\mathbf{A})$ mit

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

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

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

Lemma 4.11 Sei $\mathbf{A} \in \mathbb{K}^{n\times n}$ und λ_i ein einfacher Eigenwert zum Rechtseigenvektor $\mathbf{v}$, das heißt, $\mathbf{A}\mathbf{v}_i = \lambda_i\mathbf{v}_i$, und Linkseigenvektor $\mathbf{u}_i$, das heißt, $\mathbf{u}_i^\star\mathbf{A} = \mathbf{u}_i^\star\lambda_i$. Dann besitzt die Matrix $\mathbf{A}+\varepsilon\mathbf{C}$ für genügend kleines $|\varepsilon| > 0$ einen einfachen Eigenwert $\lambda(\varepsilon)$ und es gilt

$$\lambda(\varepsilon) = \lambda_i + \varepsilon\frac{\mathbf{u}_i^\star\mathbf{C}\mathbf{v}_i}{\mathbf{u}_i^\star\mathbf{v}_i} + \mathcal{O}(\varepsilon^2), \quad \varepsilon \to 0. \tag{4.4}$$

Beweis. Da der Eigenwert λ_i einfach ist, folgt aus der Jordan'schen Normalform

$$\mathbf{A} = \mathbf{T}\begin{bmatrix}\lambda_i & \mathbf{0}\\ \mathbf{0} & \mathbf{B}\end{bmatrix}\mathbf{T}^{-1} \quad \text{mit} \quad \mathbf{v}_i = \alpha\mathbf{T}\mathbf{e}_1, \quad \mathbf{u}_i = \beta\mathbf{T}^{-\star}\mathbf{e}_1, \quad \alpha,\beta \neq 0, \tag{4.5}$$

dies bedeutet,

$$\mathbf{u}_i^\star\mathbf{v}_i = \alpha\overline{\beta}\mathbf{e}_1^\star\mathbf{T}^{-1}\mathbf{T}\mathbf{e}_1 = \alpha\overline{\beta} \neq 0. \tag{4.6}$$

Wir betrachten nun die analytische Funktion

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

für die offensichtlich gilt

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

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

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

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

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

was wegen (4.6) $\lambda = 0$ impliziert und weiter $\mathbf{A}\mathbf{v} = \lambda_i \mathbf{v}$. Folglich ist $\mathbf{v}$ ein Vielfaches von $\mathbf{v}_i$ und wegen der zweiten Gleichung gilt $\mathbf{v} = \mathbf{0}$.

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

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

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

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

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

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

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

Korollar 4.12 Sei $\mathbf{A} \in \mathbb{K}^{n\times n}$ und λ_i ein einfacher Eigenwert zum Rechtseigenvektor $\mathbf{v}_i$ und Linkseigenvektor $\mathbf{u}_i$. Dann ist die Kondition der Berechnung des Eigenwerts

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

Beweis. Aus (4.4) ergibt sich für die absolute Kondition

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

woraus für die relative Kondition folgt

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

□

Bemerkung: Die Kondition wird demnach groß, wenn $\mathbf{u}_i^\star \mathbf{v}_i \approx 0$. Bei normalen Matrizen fallen Rechts- und Linkseigenvektor zusammen, weshalb sich hier $\kappa_{\text{abs}} = 1$ ergibt.

4.3 Potenzmethode

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

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

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

$$\mathbf{x} = \sum_{i=1}^{n} \xi_i \mathbf{v}_i, \tag{4.7}$$

und folglich ist

$$\mathbf{A}^k \mathbf{x} = \sum_{i=1}^{n} \lambda_i^k \xi_i \mathbf{v}_i. \tag{4.8}$$

Die Potenzmethode beruht nun auf der asymptotischen Identität

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

Algorithmus 4.13 (von Mises-Potenzmethode)
input: Matrix $\mathbf{A} \in \mathbb{K}^{n \times n}$ und Startvektor $\mathbf{x} \in \mathbb{K}^n$
output: Folge von Iterierten $\{\mathbf{z}_k\}_{k \geq 0}$

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

② berechne

$$\widetilde{\mathbf{z}}_{k+1} := \mathbf{A}\mathbf{z}_k, \qquad \mathbf{z}_{k+1} := \frac{\widetilde{\mathbf{z}}_{k+1}}{\|\widetilde{\mathbf{z}}_{k+1}\|} \tag{4.9}$$

③ erhöhe $k := k + 1$ und gehe nach ②

Die Potenzmethode besitzt die folgenden Eigenschaften:

Satz 4.14 Ist $\xi_1 \neq 0$ in (4.7), dann gilt mit $q := |\lambda_2/\lambda_1| < 1$:

1. Es ist
$$\|\widetilde{\mathbf{z}}_k\| = |\lambda_1| + \mathcal{O}(q^k), \quad k \to \infty.$$
2. Ist $\lambda_1 > 0$, dann gilt
$$\|\mathbf{z}_k - \operatorname{sign}(\xi_1)\mathbf{v}_1\| = \mathcal{O}(q^k), \quad k \to \infty.$$
3. Ist $\lambda_1 < 0$, dann gilt
$$\|(-1)^{k+1}\mathbf{z}_k - \operatorname{sign}(\xi_1)\mathbf{v}_1\| = \mathcal{O}(q^k), \quad k \to \infty.$$

Beweis. Aus (4.8) folgt

$$\mathbf{A}^k\mathbf{x} = \lambda_1^k\xi_1\Bigg(\mathbf{v}_1 + \underbrace{\sum_{i=2}^{n}\left[\frac{\lambda_i}{\lambda_1}\right]^k\frac{\xi_i}{\xi_1}\mathbf{v}_i}_{=:\mathbf{w}_k}\Bigg) \tag{4.10}$$

mit

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

Damit ist

$$\mathbf{z}_{k+1} = \frac{\mathbf{A}^k\mathbf{x}}{\|\mathbf{A}^k\mathbf{x}\|} = \operatorname{sign}(\lambda_1^k\xi_1)\,\frac{\mathbf{v}_1+\mathbf{w}_k}{\|\mathbf{v}_1+\mathbf{w}_k\|}. \tag{4.11}$$

Wegen

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

folgt also

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

und

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

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

Bemerkungen:

1. Die Normierung $\widetilde{\mathbf{z}}_k \mapsto \mathbf{z}_k$ in (4.9) ist notwendig, um overflow/underflow zu vermeiden. Die Wahl der Norm ist dabei unerheblich.
2. Aus Eigenschaft 1 ergibt sich $|\lambda_1|$, aus dem Vorzeichenverhalten von $\mathbf{z}_k$ das Vorzeichen von λ_1: alternieren die Vorzeichen von $\mathbf{z}_k$, dann folgt $\lambda_1 < 0$, ansonsten ist $\lambda_1 > 0$.
3. Die Voraussetzung $\xi_1 \neq 0$ kann natürlich nicht a priori überprüft werden. Wegen Rundungsfehlereinflüssen wird jedoch in der Regel eine Komponente von $\mathbf{z}_k$ längs $\mathbf{v}_1$ im Verlauf der Iteration eingeschleppt.

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

1. Gilt $\lambda_n \neq 0$, so kann man $\mathbf{A}^{-1}$ statt $\mathbf{A}$ in (4.9) verwenden. Dies ist dann die *inverse Iteration*. Da $\mathbf{A}^{-1}$ die Eigenwerte

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

   mit den gleichen Eigenvektoren besitzt, approximiert die inverse Iteration $|\lambda_n^{-1}|$ und den entsprechenden Eigenvektor $\mathbf{v}_n$.
2. Bei der *gebrochenen Iteration* von Wielandt verwendet man $(\mathbf{A}-\lambda\mathbf{I})^{-1}$ statt $\mathbf{A}$ in (4.9), $\lambda \notin \sigma(\mathbf{A})$. Die Matrix $(\mathbf{A}-\lambda\mathbf{I})^{-1}$ besitzt die Eigenwerte $(\lambda_j-\lambda)^{-1}$ und die gebrochene Iteration approximiert den Eigenvektor zu $\lambda_i \in \sigma(\mathbf{A})$, der am nächsten zu λ liegt.

Bemerkungen:

1. Bei beiden Varianten muss ein lineares Gleichungssystem pro Schritt gelöst werden!
2. Ist λ_i der nächste Eigenwert zu λ, dann konvergiert die gebrochene Iteration um so schneller, je näher λ ist, da der Konvergenzfaktor

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

dann am kleinsten ist.

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

$$\lambda_1 \approx \frac{\mathbf{z}_k^\star \mathbf{A} \mathbf{z}_k}{\underbrace{\mathbf{z}_k^\star \mathbf{z}_k}_{=1}} = \mathbf{z}_k^\star \mathbf{A} \mathbf{z}_k = \mathbf{z}_k^\star \widetilde{\mathbf{z}}_{k+1} \tag{4.12}$$

verwendet.

Satz 4.15 Ist $\mathbf{A} = \mathbf{A}^\star$, dann gilt

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

Beweis. Der Beweis verwendet, dass $\mathbf{w}_k$ aus (4.10) senkrecht auf $\mathbf{v}_1$ steht. Mit

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

folgt aus (4.11) nämlich

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

also mit $\mathbf{z}_k^\star = \gamma_k \mathbf{v}_1^\star + \gamma_k \mathbf{w}_{k-1}^\star$

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

Wegen

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

ergibt sich daher

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

□

Entsprechend kann man für "innere" Eigenwerte verfahren:

Algorithmus 4.16 (Rayleigh-Quotient-Iteration)

input: Matrix $\mathbf{A} \in \mathbb{K}^{n\times n}$ und Näherungseigenpaar $(\mu_0, \mathbf{z}_0)$ mit $\|\mathbf{z}_0\|_2 = 1$ (bestimmt etwa durch Potenzmethode)

output: Folge von Iterierten $\{(\mu_k, \mathbf{z}_k)\}_{k>0}$

① setze $k := 0$

② berechne

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

③ erhöhe $k := k + 1$ und gehe nach ②

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

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

Satz 4.17 Sei $\mathbf{A} = \mathbf{A}^\star$ und $(\mu_0, \mathbf{z}_0)$ eine hinreichend gute Näherung an ein Eigenpaar $(\lambda, \mathbf{v})$ von $\mathbf{A}$ ($\|\mathbf{v}\|_2 = \|\mathbf{z}_1\|_2 = 1$). Dann konvergiert μ_k aus der Rayleigh-Quotient-Iteration lokal kubisch gegen λ, das heißt, es existiert ein $C > 0$ mit

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

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

$$\mathbf{z}_{k-1} = \mathbf{x}_{k-1} + \mathbf{y}_{k-1} \quad \text{mit} \quad \mathbf{A}\mathbf{x}_{k-1} = \lambda\mathbf{x}_{k-1}, \quad \mathbf{x}_{k-1} \perp \mathbf{y}_{k-1} \tag{4.13}$$

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

$$\|\mathbf{y}_{k-1}\|_2 \le \frac{1}{2}\|\mathbf{z}_{k-1}\|_2 = \frac{1}{2}, \tag{4.14}$$

$$|\lambda - \mu_k| \le \frac{1}{3}|\lambda - \widehat{\lambda}| = \frac{1}{3}\min_{\widetilde{\lambda}\in\sigma(\mathbf{A})\setminus\{\lambda\}} |\lambda - \widetilde{\lambda}|. \tag{4.15}$$

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

$$|\mu_k - \widetilde{\lambda}| = |\mu_k - \lambda + \lambda - \widetilde{\lambda}| \ge |\lambda - \widetilde{\lambda}| - |\lambda - \mu_k| \ge \frac{2}{3}|\lambda - \widetilde{\lambda}|. \tag{4.16}$$

Es gilt

$$|\lambda - \mu_{k+1}| = |\lambda - \mathbf{z}_k^\star \mathbf{A}\mathbf{z}_k| = |\mathbf{z}_k^\star(\lambda\mathbf{I} - \mathbf{A})\mathbf{z}_k| = \frac{|\widetilde{\mathbf{z}}_k^\star(\lambda\mathbf{I} - \mathbf{A})\widetilde{\mathbf{z}}_k|}{\|\widetilde{\mathbf{z}}_k\|_2^2}, \tag{4.17}$$

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

$$|\lambda - \mu_{k+1}| = \left|\frac{\mathbf{z}_{k-1}^\star(\mu_k\mathbf{I} - \mathbf{A})^{-2}(\lambda\mathbf{I} - \mathbf{A})\mathbf{z}_{k-1}}{\mathbf{z}_{k-1}^\star(\mu_k\mathbf{I} - \mathbf{A})^{-2}\mathbf{z}_{k-1}}\right|. \tag{4.18}$$

Wegen (4.13) gilt auch $(\mu_k\mathbf{I}-\mathbf{A})^{-2}\mathbf{x}_{k-1} = (\mu_k-\lambda)^{-2}\mathbf{x}_{k-1} \perp \mathbf{y}_{k-1}$, und daher folgt

$$\begin{aligned}\mathbf{z}_{k-1}^\star(\mu_k\mathbf{I}-\mathbf{A})^{-2}\mathbf{z}_{k-1} &= \frac{1}{(\mu_k-\lambda)^2}\|\mathbf{x}_{k-1}\|_2^2 + \underbrace{\mathbf{y}_{k-1}^\star(\mu_k\mathbf{I}-\mathbf{A})^{-2}\mathbf{y}_{k-1}}_{\geq 0}\\ &\geq \frac{1}{(\mu_k-\lambda)^2}\|\mathbf{x}_{k-1}\|_2^2 \overset{(4.14)}{\geq} \frac{1}{4(\mu_k-\lambda)^2}.\end{aligned}$$

Andererseits ist

$$\begin{aligned}|\mathbf{z}_{k-1}^\star(\mu_k\mathbf{I}-\mathbf{A})^{-2}(\lambda\mathbf{I}-\mathbf{A})\mathbf{z}_{k-1}| &= |\mathbf{z}_{k-1}^\star(\mu_k\mathbf{I}-\mathbf{A})^{-2}(\lambda\mathbf{I}-\mathbf{A})\mathbf{y}_{k-1}|\\ &\overset{\text{Ortho-gonalität}}{=} |\mathbf{y}_{k-1}^\star(\mu_k\mathbf{I}-\mathbf{A})^{-2}(\lambda\mathbf{I}-\mathbf{A})\mathbf{y}_{k-1}|\\ &\leq \max_{\widetilde{\lambda}\in\sigma(\mathbf{A})\setminus\{\lambda\}} \frac{1}{(\mu_k-\widetilde{\lambda})^2} \underbrace{|\mathbf{y}_{k-1}^\star(\lambda\mathbf{I}-\mathbf{A})\mathbf{y}_{k-1}|}_{=|\mathbf{z}_{k-1}^\star(\lambda\mathbf{I}-\mathbf{A})\mathbf{z}_{k-1}|,\text{ da }(\lambda\mathbf{I}-\mathbf{A})\mathbf{x}_{k-1}=\mathbf{0}}\\ &\overset{(4.16)}{\leq} \left(\frac{3}{2}\right)^2 \frac{1}{(\lambda-\widehat{\lambda})^2}|\mathbf{z}_{k-1}^\star(\lambda\mathbf{I}-\mathbf{A})\mathbf{z}_{k-1}|\\ &\overset{(4.17)}{=} \frac{9}{4}\frac{1}{(\lambda-\widehat{\lambda})^2}|\lambda-\mu_k|.\end{aligned}$$

Mit mit (4.18) ergibt daher sich ingesamt

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

dies ist die kubische Konvergenzrate.
Wir müssen nur noch zeigen, dass die Iterierte $(\mu_{k+1}, \mathbf{z}_k)$ wieder den Abschätzungen (4.14) und (4.15) genügt. Die Bedingung (4.15) folgt aus

$$|\lambda-\mu_{k+1}| \leq \frac{9}{(\lambda-\widehat{\lambda})^2}\frac{(\lambda-\widehat{\lambda})^2}{9}|\lambda-\mu_k| = |\lambda-\mu_k| \overset{(4.15)}{\leq} \frac{1}{3}|\lambda-\widehat{\lambda}|.$$

Ferner folgt (4.14) aus

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

da wegen

$$\|\widetilde{\mathbf{y}}_k\|_2 = \|(\mu_k\mathbf{I}-\mathbf{A})^{-1}\mathbf{y}_{k-1}\|_2 \overset{(4.16)}{\leq} \frac{3}{2}\frac{1}{|\lambda-\widehat{\lambda}|}\|\mathbf{y}_{k-1}\|_2 \overset{(4.15)}{\leq} \frac{1}{2}\frac{1}{|\lambda-\mu_k|}\|\mathbf{y}_{k-1}\|_2$$

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

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

4.4 QR-Verfahren

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

Algorithmus 4.18 (QR-Verfahren)
input: Matrix $\mathbf{A} \in \mathbb{K}^{n\times n}$
output: Folge von Iterierten $\{\mathbf{A}_k\}_{k\geq 0}$

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

② berechne die QR-Zerlegung

$$\mathbf{A}_k = \mathbf{Q}_k\mathbf{R}_k \tag{4.19}$$

und setze

$$\mathbf{A}_{k+1} := \mathbf{R}_k\mathbf{Q}_k \tag{4.20}$$

③ erhöhe $k := k + 1$ und gehe nach ②

Lemma 4.19 Die Matrizen $\mathbf{A}_k$ aus Algorithmus 4.18 besitzten folgende Eigenschaften:

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

Beweis.

1. Mit (4.19) und (4.20) ergibt sich
$$\mathbf{A}_{k+1} = \mathbf{R}_k\mathbf{Q}_k = \mathbf{Q}_k^\star\mathbf{Q}_k\mathbf{R}_k\mathbf{Q}_k = \mathbf{Q}_k^\star(\mathbf{Q}_k\mathbf{R}_k)\mathbf{Q}_k = \mathbf{Q}_k^\star\mathbf{A}_k\mathbf{Q}_k.$$
2. Die Behauptung ergibt sich sofort aus der ersten Aussage wegen $\mathbf{A}_0 = \mathbf{A}$.
3. Für $k = 0$ entspricht die Aussage genau (4.19). Der Induktionsschritt $k \mapsto k + 1$ folgt nun aus
$$\mathbf{Q}_{k+1}\mathbf{R}_{k+1} = \mathbf{A}_{k+1} = (\mathbf{Q}_0\mathbf{Q}_1\cdots\mathbf{Q}_k)^\star\mathbf{A}(\mathbf{Q}_0\mathbf{Q}_1\cdots\mathbf{Q}_k)$$
zusammen mit der Induktionsannahme
$$(\mathbf{Q}_0\cdots\mathbf{Q}_k\underbrace{\mathbf{Q}_{k+1})(\mathbf{R}_{k+1}}_{=\mathbf{A}_{k+1}}\mathbf{R}_k\cdots\mathbf{R}_0) = \mathbf{A}\underbrace{(\mathbf{Q}_0\mathbf{Q}_1\cdots\mathbf{Q}_k)^\star(\mathbf{R}_k\mathbf{R}_{k-1}\cdots\mathbf{R}_0)}_{=\mathbf{A}^{k+1}} = \mathbf{A}^{k+2}.$$

□

Bemerkungen:

1. Wegen der ersten Aussage von Lemma 4.19 sind alle Matrizen $\mathbf{A}_k$ ähnlich zueinander und besitzten daher die selben Eigenwerte.
2. Anstelle der QR-Zerlegung kann man auch eine LR-Zerlegung verwenden: berechne in (4.19) die LR-Zerlegung $\mathbf{A}_k = \mathbf{L}_k\mathbf{R}_k$ und setze in (4.20) $\mathbf{A}_{k+1} := \mathbf{R}_k\mathbf{L}_k$. Dies ist das *$LR$-Verfahren*, das jedoch instabil ist.

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

Satz 4.20 Sei $\mathbf{A} = \mathbf{V}\mathbf{D}\mathbf{V}^{-1} \in \mathbb{K}^{n\times n}$ diagonalisierbar mit betragsmäßig verschiedenen Eigenwerten

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

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

Beweis. Die dritte Aussage aus Lemma 4.19 liefert die QR-Zerlegung von $\mathbf{A}^k$,

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

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

$$\mathbf{A}^k = (\mathbf{V}\mathbf{D}\mathbf{V}^{-1})^k = \mathbf{V}\mathbf{D}^k\mathbf{V}^{-1} = \mathbf{V}\mathbf{D}^k\mathbf{L}\mathbf{U} = (\underbrace{\mathbf{V}\mathbf{D}^k\mathbf{L}\mathbf{D}^{-k}}_{=:\mathbf{B}_k})\mathbf{D}^k\mathbf{U} = \mathbf{B}_k\mathbf{D}^k\mathbf{U}. \tag{4.21}$$

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

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

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

$$\mathbf{Q}_0\mathbf{Q}_1\cdots\mathbf{Q}_{k-1} = \mathbf{P}_k\mathbf{T}_k^\star, \qquad \mathbf{R}_{k-1}\mathbf{R}_{k-2}\cdots\mathbf{R}_0 = \mathbf{T}_k\mathbf{S}_k\mathbf{D}^k\mathbf{U} \tag{4.22}$$

für eine unitäre Diagonalmatrix $\mathbf{T}_k$. Wegen

$$\mathbf{Q}_k = (\mathbf{Q}_0\mathbf{Q}_1\cdots\mathbf{Q}_{k-1})^{-1}(\mathbf{Q}_0\mathbf{Q}_1\cdots\mathbf{Q}_k) \overset{(4.22)}{=} \mathbf{T}_k\mathbf{P}_k^\star\mathbf{P}_{k+1}\mathbf{T}_{k+1}^\star$$

und

$$\begin{aligned}\mathbf{R}_k &= (\mathbf{R}_k\mathbf{R}_{k-1}\cdots\mathbf{R}_0)(\mathbf{R}_{k-1}\mathbf{R}_{k-2}\cdots\mathbf{R}_0)^{-1}\\ &\overset{(4.22)}{=} \mathbf{T}_{k+1}\mathbf{S}_{k+1}\mathbf{D}^{k+1}\mathbf{U}\mathbf{U}^{-1}\mathbf{D}^{-k}\mathbf{S}_k^{-1}\mathbf{T}_k^\star\\ &= \mathbf{T}_{k+1}\mathbf{S}_{k+1}\mathbf{D}\mathbf{S}_k^{-1}\mathbf{T}_k^\star\end{aligned}$$

ergibt sich aus (4.19)

$$\begin{aligned}\mathbf{A}_k = \mathbf{Q}_k\mathbf{R}_k &= \mathbf{T}_k\mathbf{P}_k^\star\mathbf{P}_{k+1}\mathbf{T}_{k+1}^\star\mathbf{T}_{k+1}\mathbf{S}_{k+1}\mathbf{D}\mathbf{S}_k^{-1}\mathbf{T}_k^\star\\ &= \mathbf{T}_k\mathbf{S}_k\underbrace{\mathbf{S}_k^{-1}\mathbf{P}_k^\star}_{=\mathbf{B}_k^{-1}}\underbrace{\mathbf{P}_{k+1}\mathbf{S}_{k+1}}_{=\mathbf{B}_{k+1}}\mathbf{D}\mathbf{S}_k^{-1}\mathbf{T}_k^\star\\ &= \mathbf{T}_k\mathbf{S}_k\mathbf{B}_k^{-1}\mathbf{B}_{k+1}\mathbf{D}\mathbf{S}_k^{-1}\mathbf{T}_k^\star.\end{aligned} \tag{4.23}$$

Bezeichnen wir mir $\ell_{i,j}$ die Einträge von $\mathbf{L}$, dann ist insbesondere $\ell_{i,i} = 1$ und aus der Anordnung der Eigenwerte in $\mathbf{D}$ ergibt sich für ein $q \in (0,1)$

$$[\mathbf{D}^k\mathbf{L}\mathbf{D}^{-k}]_{i,j} = \lambda_i^k \ell_{i,j} \lambda_j^{-k} = \begin{cases} 0, & \text{falls } i < j, \\ 1, & \text{falls } i = j, \\ \mathcal{O}\Big(\big|\frac{\lambda_i}{\lambda_j}\big|^k\Big), & \text{falls } i > j. \end{cases} \tag{4.24}$$

Setzen wir $q := \min_{i>j}\{|\lambda_j/\lambda_i|\}$, so folgt (vgl. (4.21)) hieraus

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

Demzufolge erhalten wir

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

und eingesetzt in (4.23) ergibt sich

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

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

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

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

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

mindestens linear gegen eine obere Dreiecksmatrix und es gilt

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

□

4.5 Implementierung des QR-Verfahrens

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

Definition 4.21 Eine Matrix $\mathbf{H} = [h_{i,j}]_{i,j=1}^n$ besitzt **obere Hessenberg-Form**, wenn $h_{i,j} = 0$ für $j < i-1$, das heißt

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

Ziel ist es also zunächst, $\mathbf{A}$ durch Ähnlichkeitstransformationen auf obere Hessenberg-Form zu bringen. Dazu gehen wir wie in Abschnitt 2.2 bei der QR-Zerlegung vor: Wähle Householder-Spiegelungen $\mathbf{P}_1, \mathbf{P}_2, \ldots, \mathbf{P}_{n-2}$ und bilde

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

gemäß dem folgenden Schema:

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

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

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

Allgemein gilt im i-ten Schritt

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

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

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

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

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

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

Definition 4.22 Eine Matrix $\mathbf{G} = \mathbf{G}(i, j, \theta)$ mit

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

mit

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

heißt **Givens-Rotation**.

Givens-Rotationen sind orthogonale Matrizen (alle Zeilen sind paarweise zueinander senkrecht) und die Operation $\mathbf{G}(i,j,\theta)\mathbf{A}$ ersetzt die Zeilen i und j von $\mathbf{A}$ durch Linearkombinationen $c[a_{i,1}, a_{i,2}, \ldots, a_{i,n}] + s[a_{j,1}, a_{j,2}, \ldots, a_{j,n}]$ beziehungsweise $-s[a_{i,1}, a_{i,2}, \ldots, a_{i,n}] + c[a_{j,1}, a_{j,2}, \ldots, a_{j,n}]$, während $\mathbf{AG}(i,j,\theta)$ die Spalten i und j von $\mathbf{A}$ durch entsprechende Linearkombinationen ersetzt.

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

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

die Lösung

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

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

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

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

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

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

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

gemäß dem Schema:

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

$$\xrightarrow[i=3]{} \begin{bmatrix} \times & \times & \times & \times & \times \\ 0 & \times & \times & \times & \times \\ \hline 0 & 0 & + & + & + \\ 0 & 0 & 0 & + & + \\ \hline 0 & 0 & 0 & \times & \times \end{bmatrix} \xrightarrow[i=4]{} \begin{bmatrix} \times & \times & \times & \times & \times \\ 0 & \times & \times & \times & \times \\ 0 & 0 & \times & \times & \times \\ \hline 0 & 0 & 0 & + & + \\ 0 & 0 & 0 & 0 & + \end{bmatrix}$$

Im Teilschritt (4.20) muss das Produkt der Givens-Rotationen von rechts an $\mathbf{R}_k$ heranmultipliziert werden. Wegen

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

ergibt sich

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

Die Multiplikation von rechts mit $\mathbf{G}(i,i+1,\theta_i)$ verändert die Spalten i und $i+1$. Offensichtlich besitzt $\mathbf{A}_{k+1}$ dann wieder Hessenberg-Gestalt.
Als Aufwand für die beiden Teilschritte (4.19) und (4.20) ergeben sich

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

Multiplikationen.

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

Algorithmus 4.23 (QR-Verfahren mit Shift)
input: Matrix $\mathbf{A} \in \mathbb{K}^{n\times n}$
output: Folge von Iterierten $\{\mathbf{A}_k\}_{k\geq 0}$

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

② berechne die QR-Zerlegung

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

und setze

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

③ erhöhe $k := k+1$ und gehe nach ②

Für das QR-Verfahren mit Shift bleiben die ersten beiden Aussagen von Lemma 4.19 erhalten, während die dritte nun lautet

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

Ferner ist die für $i > j$ in (4.24) beschriebene Konvergenzgeschwindigkeit nun durch den Faktor

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

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

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

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

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

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

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

Eigenvektoren: Gemäß Lemma 4.19 liefert das QR-Verfahren asymptotisch die Faktorisierung

$$\mathbf{A} = \mathbf{Q}^\star \mathbf{R} \mathbf{Q}, \quad \mathbf{Q} \text{ unitär}, \quad \mathbf{R} \text{ obere Dreiecksmatrix}$$

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

4.6 Lanczos-Verfahren

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

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

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

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

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

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

Dann gilt

$$\lambda_{\max} \geq \mu_{\max}^{(k)} = \max_{\mathbf{x} \in \mathcal{K}_k(\mathbf{A}, \mathbf{z}_0)} \frac{\mathbf{x}^\star \mathbf{A} \mathbf{x}}{\mathbf{x}^\star \mathbf{x}} = \max_{\mathbf{y} \in \mathbb{K}^k \setminus \{\mathbf{0}\}} \frac{\mathbf{y}^\star \mathbf{W}_k^\star \mathbf{A} \mathbf{W}_k \mathbf{y}}{\mathbf{y}^\star \underbrace{\mathbf{W}_k^\star \mathbf{W}_k}_{=\mathbf{I}} \mathbf{y}} = \max_{\mathbf{y} \in \mathbb{K}^k \setminus \{\mathbf{0}\}} \frac{\mathbf{y}^\star \mathbf{W}_k^\star \mathbf{A} \mathbf{W}_k \mathbf{y}}{\mathbf{y}^\star \mathbf{y}}$$

und analog

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

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

Satz 4.24 Sei $\mathbf{A}\in\mathbb{K}^{n\times n}$ eine hermitesche Matrix mit absteigend sortierten Eigenwerten $\lambda_1\geq\lambda_2\geq\cdots\geq\lambda_n$ und $\mathbf{v}_1,\mathbf{v}_2,\ldots,\mathbf{v}_n$ die zugehörigen orthonormalen Eigenvektoren. Seien weiter $\mu_1^{(k)}\geq\mu_2^{(k)}\geq\cdots\geq\mu_k^{(k)}$ die Eigenwerte von $\mathbf{W}_k^\star\mathbf{A}\mathbf{W}_k$. Dann gilt

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

mit $\rho_1=(\lambda_1-\lambda_2)/(\lambda_2-\lambda_n)$ und $\cos(\phi_1)=|\mathbf{v}_1^\star\mathbf{z}|/\|\mathbf{z}\|_2$.

Beweis. Sei ohne Beschränkung der Allgemeinheit $\|\mathbf{z}\|_2=1$. Da $\mathcal{K}_k(\mathbf{A},\mathbf{z})=\operatorname{span}\{p(\mathbf{A})\mathbf{z}:p\in\Pi_{k-1}\}$ gilt

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

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

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

so folgt

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

Wir erhalten

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

Um eine möglichst scharfe Abschätzung zu erhalten, müssen wir ein Polynom $p\in\Pi_{k-1}$ einsetzen, das innerhalb des Intervalls $[\lambda_n,\lambda_1]$ möglichst klein ist. Wir wählen das transformierte Tschebyscheff-Polynom

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

mit der Eigenschaft $p(\lambda_i)\leq 1$ für $j=2,3,\ldots,n$. Damit gilt dann wegen $\sum_{i=1}^n|\xi_i|^2=\|\mathbf{z}\|_2^2=1$, dass

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

und die Behauptung folgt aus der Tatsache, dass

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

□

Ein analoges Resultat erhalten wir für den kleinsten Eigenwert, indem wir Satz 4.24 auf $-\mathbf{A}$ anwenden.

Korollar 4.25 Unter den Voraussetzungen von Satz 4.24 gilt

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

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

Bemerkungen:

1. Da der Krylov-Raum $\mathcal{K}_k(\mathbf{A}, \mathbf{z})$ den Vektor $\mathbf{A}^{k-1}\mathbf{z}$ enthält, ist durch den betragsgrößten Eigenwert von $\mathbf{W}_k^\star \mathbf{A}\mathbf{W}_k$ eine bessere Näherung an den betragsgrößten Eigenwert von $\mathbf{A}$ gegeben als durch den entsprechenden Rayleigh-Quotienten $\mathbf{A}^{k-1}\mathbf{z}/\|\mathbf{A}^{k-2}\mathbf{z}\|_2$ der Potenzmethode.
2. Auch die anderen Eigenwerte von $\mathbf{W}_k^\star \mathbf{A}\mathbf{W}_k$ können als Näherungen an die Eigenwerte von $\mathbf{A}$ herangezogen werden: Mit wachsendem k fällt der j-kleinste Eigenwert von $\mathbf{W}_k^\star \mathbf{A}\mathbf{W}_k$ monoton von oben gegen den j-kleinsten Eigenwert von $\mathbf{A}$, während der j-größte Eigenwert von $\mathbf{W}_k^\star \mathbf{A}\mathbf{W}_k$ monoton von unten gegen den j-größten Eigenwert von $\mathbf{A}$ wächst.
3. Ist $\mathbf{u}_1$ der Eigenvektor von $\mathbf{W}_k^\star \mathbf{A}\mathbf{W}_k$ zum Eigenwert $\mu_1^{(k)}$, dann ist gemäß Konstruktion $\mathbf{W}_k\mathbf{u}_1$ eine Näherung an den Eigenvektor zum größten Eigenwert λ_1 von $\mathbf{A}$. Entsprechendes gilt für die anderen Eigenpaare.

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

Algorithmus 4.26 (Lanczos-Prozess)
input: Matrix $\mathbf{A} \in \mathbb{K}^{n\times n}$ und Startvektor $\mathbf{z} \in \mathbb{K}^n$
output: Orthonormalbasis $\{\mathbf{w}_k\}_{k>0}$

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

② berechne

$$\beta_{k-1} := \|\mathbf{u}_k\|_2 \tag{4.25}$$
$$\mathbf{w}_k := \frac{\mathbf{u}_k}{\beta_{k-1}} \text{ falls } \beta_{k-1} \neq 0 \tag{4.26}$$
$$\alpha_k := \mathbf{w}_k^\star \mathbf{A}\mathbf{w}_k \tag{4.27}$$
$$\mathbf{u}_{k+1} := (\mathbf{A} - \alpha_k \mathbf{I})\mathbf{w}_k - \beta_{k-1}\mathbf{w}_{k-1} \tag{4.28}$$

③ erhöhe $k := k + 1$ und gehe nach ②

Satz 4.27 Die durch Algorithmus 4.26 gebildete Folge $\{\mathbf{w}_i\}_{i=1}^k$ ist eine Orthonormalbasis des Krylov-Raums $\mathcal{K}_k(\mathbf{A}, \mathbf{z})$.

Beweis. Wir beweisen die Aussage mittels vollständiger Induktion über k. Für $k = 1$ ist

die Aussage offensichtlich. Sei also $\{\mathbf{w}_i\}_{i=1}^k$ ist eine Orthonormalbasis des Krylov-Raums $\mathcal{K}_k(\mathbf{A},\mathbf{z})$. Dann folgt aus Algorithmus 4.26, dass

$$\mathbf{u}_{k+1} = \mathbf{A}\mathbf{w}_k + \mathbf{p}_k \quad \text{mit} \quad \mathbf{p}_k \in \mathcal{K}_k(\mathbf{A},\mathbf{z}) \quad \text{und} \quad \mathbf{A}\mathbf{w}_k \in \mathcal{K}_{k+1}(\mathbf{A},\mathbf{z}). \tag{4.29}$$

Folglich ist $\mathbf{w}_{k+1} \in \mathcal{K}_{k+1}(\mathbf{A},\mathbf{z})$ mit $\|\mathbf{w}_{k+1}\|_2 = 1$.
Wir zeigen nun, dass gilt $\mathbf{w}_{k+1} \perp \mathcal{K}_k(\mathbf{A},\mathbf{z})$. Für $1 \le i < k-1$ ergibt sich

$$\mathbf{w}_i^\star \mathbf{u}_{k+1} \overset{(4.28)}{=} \mathbf{w}_i^\star(\mathbf{A}-\alpha_k\mathbf{I})\mathbf{w}_k - \beta_{k-1}\underbrace{\mathbf{w}_i^\star\mathbf{w}_{k-1}}_{=0} = \big(\underbrace{(\mathbf{A}-\alpha_k\mathbf{I})\mathbf{w}_i}_{\in\mathcal{K}_i(\mathbf{A},\mathbf{z})=\operatorname{span}\{\mathbf{w}_1,\ldots,\mathbf{w}_{i+1}\}}\big)^\star\mathbf{w}_k = 0.$$

Ferner gilt

$$\begin{aligned}\mathbf{w}_{k-1}^\star\mathbf{u}_{k+1} &\overset{(4.28)}{=} \mathbf{w}_{k-1}^\star\mathbf{A}\mathbf{w}_k - \alpha_k\underbrace{\mathbf{w}_{k-1}^\star\mathbf{w}_k}_{=0} - \beta_{k-1}\underbrace{\mathbf{w}_{k-1}^\star\mathbf{w}_{k-1}}_{=1}\\ &\overset{(4.29)}{=} (\mathbf{u}_k - \underbrace{\mathbf{p}_{k-1}}_{\in\mathcal{K}_{k-1}(\mathbf{A},\mathbf{z})\perp\mathbf{w}_k})^\star\mathbf{w}_k - \beta_{k-1} = \mathbf{u}_k^\star\mathbf{w}_k - \beta_{k-1} \overset{(4.25)}{=} 0,\end{aligned}$$

sowie

$$\mathbf{w}_k^\star\mathbf{u}_{k+1} \overset{(4.28)}{=} \mathbf{w}_k^\star(\mathbf{A}-\alpha_k\mathbf{I})\mathbf{w}_k - \beta_{k-1}\underbrace{\mathbf{w}_k^\star\mathbf{w}_{k-1}}_{=0} = \mathbf{w}_k^\star\mathbf{A}\mathbf{w}_k - \alpha_k \overset{(4.27)}{=} 0.$$

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

Bemerkung: Der Lanczos-Prozess bricht zusammmen, falls sich in (4.26) $\beta_{k+1} = 0$ ergibt. Dies ist genau dann der Fall, wenn

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

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

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

Proposition 4.28 Es ist

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

mit α_i, β_i wie in Algorithmus 4.26.

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

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

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

$$[\mathbf{T}_k]_{i+1,i} = \mathbf{w}_{i+1}^\star \mathbf{A}\mathbf{w}_i \overset{(4.29)}{=} \mathbf{w}_{i+1}^\star(\mathbf{u}_{i+1} - \mathbf{p}_i) = \mathbf{w}_{i+1}^\star \mathbf{u}_{i+1} \overset{(4.26)}{=} \beta_i.$$

□

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

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

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

Auflösen nach dem letzten Term der Summe ergibt

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

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

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

Diese Gleichungen führen auf den Arnoldi-Prozess:

Algorithmus 4.29 (Arnoldi-Prozess)

input: Matrix $\mathbf{A} \in \mathbb{K}^{n\times n}$ und Startvektor $\mathbf{z} \in \mathbb{K}^n$

output: Orthonormalbasis $\{\mathbf{w}_k\}_{k\ge 0}$

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

② berechne

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

③ erhöhe $k := k+1$ und gehe nach ②

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

5. Partielle Differentialgleichungen

5.1 Beispiele

Potentialgleichung: Es sei $\Omega \subset \mathbb{R}^2$ ein *Gebiet*, das ist eine offene, zusammenhängende Menge, und $\Gamma := \partial\Omega$ der Rand. Der Graph der Funktion $g : \Gamma \to \mathbb{R}$ beschreibe eine Drahtschlinge, die eine Seifenhaut aufspannt. Diese Seifenhaut lässt sich als Funktion $u : \Omega \to \mathbb{R}$ beschreiben, deren Form minimale Oberfläche besitzt

$$\int_\Omega \sqrt{1 + u_x^2 + u_y^2}\,\mathrm{d}x\,\mathrm{d}y \to \min\,.$$

Wegen $\sqrt{1+z} = 1 + \frac{z}{2} + \mathcal{O}(z^2)$ kann man den Integranden für kleine Werte von u_x und u_y ersetzen durch

$$F(u) := \frac{1}{2}\int_\Omega u_x^2 + u_y^2\,\mathrm{d}x\,\mathrm{d}y \to \min\,.$$

Ist $u \in C^2(\Omega) \cap C(\overline{\Omega})$ mit $u|_\Gamma = g$ Lösung dieser Minimierungsaufgabe, dann folgt für beliebiges $v \in C^1(\Omega)$ mit $v|_\Gamma = 0$ dass

$$0 = \lim_{\varepsilon\to 0} \frac{F(u+\varepsilon v) - F(u)}{\varepsilon} = \int_\Omega u_x v_x + u_y v_y\,\mathrm{d}x\,\mathrm{d}y = \int_\Omega \nabla u \nabla v\,\mathrm{d}\mathbf{x}. \tag{5.1}$$

Für $\mathbf{f} := \nabla u v$ liefert der Gaußsche Integralsatz die Identität

$$\int_\Omega \Delta u v\,\mathrm{d}\mathbf{x} + \int_\Omega \nabla u \nabla v\,\mathrm{d}\mathbf{x} = \int_\Omega \operatorname{div}\mathbf{f}\,\mathrm{d}\mathbf{x} = \int_\Gamma \mathbf{f}\,\mathbf{n}\,\mathrm{d}\sigma = \int_\Gamma \underbrace{v}_{=0}\frac{\partial u}{\partial \mathbf{n}}\,\mathrm{d}\sigma = 0,$$

wobei $\Delta u = u_{xx} + u_{yy}$ den *Laplace-Operator* bezeichnet. Dies eingesetzt in (5.1) ergibt für u die Bedingung

$$0 = \int_\Omega \Delta u v\,\mathrm{d}\mathbf{x}$$

für alle $v \in C^1(\Omega)$ mit $v|_\Gamma = 0$. Daher muss die Funktion u der *Potential-* oder *Laplace-Gleichung*

$$\Delta u(\mathbf{x}) = 0, \quad \mathbf{x} \in \Omega \tag{5.2}$$

genügen.

Wärmeleitungsgleichung: In einem offenen, beschränkten Gebiet $\Omega \subset \mathbb{R}^d$ beschreibe die Funktion $u : \mathbb{R}_{\geq 0} \times \Omega \to \mathbb{R}$ die Temperaturverteilung. Zum Zeitpunkt $t = 0$ liege die Anfangsverteilung $u(0, \mathbf{x}) = u_0(\mathbf{x}) \in C(\Omega)$ vor. Zusätzlich seien im Gebiet Ω die

Wärmequelle $f \in C(\mathbb{R}_{\geq 0} \times \Omega)$ und an dessen Rand $\Gamma = \partial\Omega$ die Temperaturverteilung $g \in C\big(\mathbb{R}_{\geq 0} \times \Gamma\big)$ vorgegeben.
Aus dem Erhaltungssatz folgt dann für jedes Kontrollvolumen $V \subset \Omega$

$$\underbrace{\int_V \frac{\partial}{\partial t} u(t,\mathbf{x})\,\mathrm{d}\mathbf{x}}_{\text{Wärmegehalt in } V} = \underbrace{-\int_{\partial V} \mathbf{q}(t,\mathbf{x})\mathbf{n}(\mathbf{x})\,\mathrm{d}\sigma}_{\text{Wärmefluss von außen}} + \underbrace{\int_V f(t,\mathbf{x})\,\mathrm{d}\mathbf{x}}_{\text{Wärmequelle}}.$$

Gemäß dem Materialgesetz genügt der Wärmefluss der Beziehung

$$\mathbf{q}(t,\mathbf{x}) = -c(\mathbf{x})\nabla u(t,\mathbf{x})$$

mit der materialabhängigen Wärmeleitkonstante $c \geq c_0 > 0$. Eingesetzt in den Gaußschen Integralsatz folgt daher

$$\int_{\partial V} \mathbf{q}(t,\mathbf{x})\mathbf{n}(\mathbf{x})\,\mathrm{d}\sigma = \int_V \mathrm{div}\mathbf{q}(t,\mathbf{x})\,\mathrm{d}\mathbf{x} = \int_V \mathrm{div}\big(c(\mathbf{x})\nabla u(t,\mathbf{x})\big)\,\mathrm{d}\mathbf{x}.$$

Für die Temperaturverteilung folgt somit für alle Kontrollvolumen V die Gleichung

$$\int_V \left\{ \frac{\partial}{\partial t} u(t,\mathbf{x}) - \mathrm{div}\big(c(\mathbf{x})\nabla u(t,\mathbf{x})\big) \right\} \mathrm{d}\mathbf{x} = \int_V f(t,\mathbf{x})\,\mathrm{d}\mathbf{x},$$

dies bedeutet

$$\frac{\partial}{\partial t} u(t,\mathbf{x}) - \mathrm{div}\big(c(\mathbf{x})\nabla u(t,\mathbf{x})\big) = f(t,\mathbf{x}), \quad (t,\mathbf{x}) \in \mathbb{R}_{\geq 0} \times \Omega.$$

Ist c konstant, etwa $c = 1$, so genügt die Temperaturverteilung $u \in C^2(\mathbb{R}_{>0} \times \Omega) \cap C(\mathbb{R}_{\geq 0} \times \overline{\Omega})$ der Gleichung

$$\frac{\partial}{\partial t} u(t,\mathbf{x}) - \Delta u(t,\mathbf{x}) = f(t,\mathbf{x}), \quad (t,\mathbf{x}) \in \mathbb{R}_{\geq 0} \times \Omega \tag{5.3}$$

mit dem d-dimensionalen Laplace-Operator $\Delta = \frac{\partial^2}{\partial x_1^2} + \cdots + \frac{\partial^2}{\partial x_d^2}$.

Poisson-Gleichung: Sind die Daten f und g der Wärmeleitungsgleichung nicht zeitabhängig, dann stellt sich für $t \to \infty$ ein Gleichgewichtszustand ein. Dies bedeutet, es gilt $\partial u/\partial t = 0$ und (5.3) geht über in die *Poisson-Gleichung*

$$-\Delta u(\mathbf{x}) = f(\mathbf{x}), \quad \mathbf{x} \in \Omega. \tag{5.4}$$

Diese Gleichung hat in der Elektrostatik ebenfalls eine immense Bedeutung: ist in Ω die Ladungsdichte $f : \Omega \to \mathbb{R}$ bekannt, so genügt die Spannung u dort der Poisson-Gleichung.

Wellengleichung: Die Bewegung in einem idealen Gas wird durch drei Gesetze bestimmt. Wie üblich wird die Geschwindigkeit mit $\mathbf{v}$, die Dichte mit ρ und der Druck mit p bezeichnet.

1. *Kontinuitätsgleichung:*

$$\frac{\partial \rho}{\partial t} = -\rho_0 \,\mathrm{div}\,\mathbf{v}.$$

Wegen der Massenerhaltung ist die Änderung der Masse in einem Kontrollvolumen V gleich dem Fluss durch die Oberfläche, das ist $\int_{\partial V} \rho\mathbf{v}\mathbf{n}\,\mathrm{d}\sigma$. Aus dem Gaußschen Integralsatz folgt dann die obige Gleichung. Dabei wird ρ durch die feste Dichte ρ_0 approximiert.

2. *Newtonsches Gesetz:*
$$\rho_0 \frac{\partial \mathbf{v}}{\partial t} = -\nabla p.$$
Der Druckgradient induziert ein Kraftfeld, das die Beschleunigung der Teilchen bewirkt.
3. *Zustandsgleichung:*
$$p = c^2 \rho.$$
In idealen Gasen ist der Druck bei konstanter Temperatur proportional zur Dichte.

Aus den drei Gesetzen folgt
$$\frac{\partial^2 p}{\partial t^2} = c^2 \frac{\partial^2 \rho}{\partial t^2} = -c^2 \frac{\partial}{\partial t}(\rho_0 \operatorname{div} \mathbf{v}) = -c^2 \operatorname{div}\left(\rho_0 \frac{\partial \mathbf{v}}{\partial t}\right) = c^2 \operatorname{div}(\nabla p) = c^2 \Delta p.$$
Andere Beipsiele für die *Wellengleichung*
$$\frac{\partial^2 p}{\partial t^2}(t, \mathbf{x}) = c^2 \Delta p(t, \mathbf{x}), \quad (t, \mathbf{x}) \in \mathbb{R}_{\geq 0} \times \Omega \tag{5.5}$$
ergeben sich in zwei Raumdimensionen für eine schwingende Membran oder in einer Raumdimension für eine schwingende Saite.

5.2 Charakterisierung

Sei $\Omega \subset \mathbb{R}^d$ ein Gebiet und $\mathcal{L} : C^2(\Omega) \to C(\Omega)$ ein allgemeiner linearer Differentialoperator zweiter Ordnung
$$(\mathcal{L}u)(\mathbf{x}) = -\sum_{i,j=1}^{d} a_{i,j}(\mathbf{x}) \frac{\partial^2}{\partial x_i \partial x_j} u(\mathbf{x}) + \sum_{i=1}^{d} b_i(\mathbf{x}) \frac{\partial}{\partial x_i} u(\mathbf{x}) + c(\mathbf{x}) u(\mathbf{x}), \tag{5.6}$$
wobei $\mathbf{A} = [a_{i,j}]_{i,j=1}^d \in C(\Omega)$, $\mathbf{b} = [b_i]_{i=1}^d \in C(\Omega)$ und $c \in C(\Omega)$. Die zugehörige Differentialgleichung lautet dann
$$(\mathcal{L}u) = f(\mathbf{x}), \quad \mathbf{x} \in \Omega. \tag{5.7}$$
Da für $u \in C^2(\Omega)$ gilt $\partial^2 u / \partial x_i \partial x_j = \partial^2 u / \partial x_j \partial x_i$ kann ohne Beschränkung der Allgemeinheit $a_{i,j} = a_{j,i}$ angenommen werden. Demnach ist die Matrix $\mathbf{A}$ symmetrisch und besitzt nur reelle Eigenwerte. Der Differentialoperator $-\sum_{i,j=1}^d a_{i,j}(\mathbf{x}) \frac{\partial^2}{\partial x_i \partial x_j} u(\mathbf{x})$ ist der *Hauptteil* von $\mathcal{L}$.

Definition 5.1 Die Differentialoperator (5.6) heißt
- **elliptisch in $\mathbf{x}$**, falls die Eigenwerte von $\mathbf{A}(\mathbf{x})$ alle positiv sind,
- **parabolisch in $\mathbf{x}$**, falls $d-1$ Eigenwerte von $\mathbf{A}(\mathbf{x})$ alle positiv sind und ein Eigenwert verschwindet, aber $\operatorname{rang}\big([\mathbf{A}(\mathbf{x}), \mathbf{b}(\mathbf{x})]\big) = d$ ist.
- **hyperbolisch in $\mathbf{x}$**, falls $d-1$ Eigenwerte von $\mathbf{A}(\mathbf{x})$ alle positiv sind und ein Eigenwert negatives Vorzeichen besitzt.

Der Differentialoperator (5.7) heißt **elliptisch/parabolisch/hyperbolisch (in Ω)**, falls er elliptisch/parabolisch/hyperbolisch ist für alle $\mathbf{x} \in \Omega$. Entsprechend wird die Differentialgleichung (5.7) elliptisch/parabolisch/hyperbolisch genannt, wenn der zugehörige Differentialoperator diese Eigenschaft besitzt.

Beispiel 5.2 Die Potentialgleichung (5.2) und die Poisson-Gleichung (5.4) sind elliptisch, die Wärmeleitungsgleichung (5.3) ist parabolisch, während die Wellengleichung (5.5) hyperbolisch ist. △

Zusätzlich zur Differentialgleichung (5.7) müssen noch geeignete Anfangs- oder Randbedingungen gefordert werden, um eine sachgemäße Aufgabenstellung zu ergeben.

Definition 5.3 Ein Problem heißt **sachgemäß gestellt**, wenn eine Lösung existiert, diese eindeutig ist und stetig von den vorgegebenen Daten abhängt. Andernfalls heißt das Problem **schlecht gestellt**.

Die Unterscheidung partieller Differentialgleichungen in verschiedene Typen ergibt keinen Sinn, wenn nicht jeder Typ grundlegend andere Eigenschaften hätte.

1. *Elliptische Differentialgleichungen:* Bei elliptischen Problemen werden Randbedingungen vorgegeben: für gegebenes $f \in C(\Omega)$ und $g \in C(\Gamma)$ suche $u \in C^2(\Omega) \cap C(\overline{\Omega})$, so dass
$$\mathcal{L}u = f \text{ in } \Omega, \quad u = g \text{ auf } \Gamma. \tag{5.8}$$
Diese Randbedingungen heißen *Dirichlet-Randbedingungen*. In der Praxis treten oft auch *Neumann-Randbedingungen*, $\partial u/\partial \mathbf{n} = g$, auf. Lösungen elliptischer Differentialgleichungen erfüllen das Maximumprinzip (siehe nächster Abschnitt).

2. *Parabolische Differentialgleichungen:* Parabolische Differentialgleichungen beschreiben Diffusionsvorgänge. Die ausgezeichnete Koordinatenrichtung ist in der Regel die Zeit, so dass man oftmals die Differentialgleichung auf die Form $u_t + \mathcal{L}u = f$ bringen kann, wobei $\mathcal{L}$ ein elliptischer Differentialoperator ist. Zusätzlich werden Anfangsrandwerte vorgegeben: für gegebenes $f \in C(\mathbb{R}_{\geq 0} \times \Omega)$, $g \in C(\mathbb{R}_{\geq 0} \times \Gamma)$ und $u_0 \in C(\Omega)$ suche $u \in C^2(\mathbb{R}_{>0} \times \Omega) \cap C(\mathbb{R}_{\geq 0} \times \overline{\Omega})$, so dass
$$\begin{aligned} u_t + \mathcal{L}u &= f \quad \text{in } \mathbb{R}_{\geq 0} \times \Omega \\ u &= g \quad \text{auf } \mathbb{R}_{\geq 0} \times \Gamma \quad \text{(Randbedingung)} \\ u(0,\cdot) &= u_0 \text{ auf } \Omega \quad \text{(Anfangsbedingung)} \end{aligned}$$

3. *Hyperbolische Differentialgleichungen:* Hier ist ebenfalls eine Koordinate ausgezeichnet, die wieder als Zeit interpretiert werden kann, so dass man die Differentialgleichung oft auf die Form $u_{tt} + \mathcal{L}u = f$ bringen kann mit einem elliptischen Differentialoperator $\mathcal{L}$. Hyperbolische Gleichungen beschreiben physikalisch gesehen Schwingungsvorgänge. Sinnvolle Probleme erhält man mit Anfangsbedingungen: für eine gegebene Funktion $f \in C(\mathbb{R}_{>0} \times \Omega)$, $u_0, u_1 \in C(\overline{\Omega})$ suche $u \in C^2(\mathbb{R}_{>0} \times \Omega) \cap C(\mathbb{R}_{\geq 0} \times \overline{\Omega})$, so dass
$$u_{tt} + \mathcal{L}u = f \text{ in } \mathbb{R}_{\geq 0} \times \Omega, \quad u(0,\cdot) = u_0,\ u_t(0,\cdot) = u_1 \text{ auf } \overline{\Omega}.$$

Wenn Differentialgleichungen invariant gegenüber Bewegungen sind (also gegenüber Translation und Drehungen), dann hat der elliptische Differentialoperator $\mathcal{L}$ die Form
$$\mathcal{L}u = -a\Delta u + cu.$$

5.3 Maximum-Prinzip

Bei der Analyse von Differenzenverfahren spielt das diskrete Analogon des Maximumprinzips eine wichtige Rolle. Deshalb betrachten wir vorab eine einfache Fassung des Prinzips. Dazu seien $\Omega \subset \mathbb{R}^d$ stets ein beschränktes Gebiet und der elliptische Differentialoperator von der Form

$$(\mathcal{L}u)(\mathbf{x}) = -\sum_{i,j=1}^{d} a_{i,j}(\mathbf{x}) u_{x_i,x_j}(\mathbf{x}). \tag{5.9}$$

Satz 5.4 (Maximumprinzip) Die Funktion $u \in C^2(\Omega) \cap C(\overline{\Omega})$ genüge der elliptischen Differentialgleichung $\mathcal{L}u = f \leq 0$ in Ω. Dann nimmt u sein Maximum auf dem Rand Γ an.

Beweis. (i) Wir führen den Beweis zunächst unter der stärkeren Voraussetzung $f < 0$. Angenommen, es sei $\mathbf{y} \in \Omega$ mit

$$u(\mathbf{y}) = \sup_{\mathbf{x}\in\Omega} u(\mathbf{x}) > \max_{\mathbf{x}\in\Gamma} u(\mathbf{x}).$$

Bei einer linearen Koordinatentransformation $\mathbf{x} \mapsto \boldsymbol{\xi} = \mathbf{U}\mathbf{x}$ lautet der Differentialoperator in den neuen Koordinaten

$$\mathcal{L}u = -\sum_{i,j=1}^{d} \left[\mathbf{U}^\star \mathbf{A}(\mathbf{x})\mathbf{U}\right]_{i,j} u_{\xi_i,\xi_j},$$

wobei $\mathbf{A}(\mathbf{x}) = [a_{i,j}(\mathbf{x})]_{i,j=1}^d$ die Koeffizientenmatrix ist. Wegen der Symmetrie von $\mathbf{A}(\mathbf{x})$ können wir eine orthogonale Matrix $\mathbf{U}$ wählen, mit der $\mathbf{U}^T\mathbf{A}(\mathbf{y})\mathbf{U}$ diagonal wird. Aus der positiven Definitheit schließen wir, dass die Diagonalelemente positiv sind. Weil $\mathbf{y}$ Extremalpunkt ist, gilt

$$\nabla u(\mathbf{y}) = \mathbf{0}, \quad u_{\xi_i,\xi_i}(\mathbf{y}) \leq 0.$$

Dies bedeutet

$$\mathcal{L}u = -\sum_{i,j=1}^{d} \left[\mathbf{U}^T \mathbf{A}(\mathbf{x})\mathbf{U}\right]_{i,j} u_{\xi_i,\xi_j} \geq 0$$

im Widerspruch zu $(\mathcal{L}u)(\mathbf{y}) = f(\mathbf{y}) < 0$.

(ii) Sei nun $f(\mathbf{x}) \leq 0$ angenommen und es gebe ein $\mathbf{x} = \mathbf{y} \in \Omega$ mit $u(\mathbf{y}) > \max_{\mathbf{x}\in\Gamma} u(\mathbf{x})$. Die Hilfsfunktion

$$h(\mathbf{x}) := (x_1 - y_1)^2 + (x_2 - y_2)^2 + \cdots + (x_d - y_d)^2$$

ist auf Γ beschränkt. Wenn $\delta > 0$ hinreichend klein gewählt wird, nimmt also auch die Funktion

$$w = u + \delta h$$

ihr Maximum in einem Punkt $\mathbf{z}$ im Innern an. Wegen $h_{x_i,x_j} = 2\delta_{i,j}$ ist

$$\mathcal{L}w(\mathbf{z}) = \mathcal{L}u(\mathbf{z}) + \delta\mathcal{L}h(\mathbf{z}) = f(\mathbf{z}) - 2\delta \sum_{i=1}^{d} a_{i,i}(\mathbf{z}) < 0.$$

Wie im ersten Teil des Beweises ergibt sich ein Widerspruch. □

Folgerungen:

1. *Minimumprinzip:* Ist $\mathcal{L}u = f \geq 0$ in Ω, so nimmt u sein Minimum auf dem Rand Γ an.

   Beweis. Man wende auf $v := -u$ das Maximumprinzip an. □

2. *Vergleichsprinzip:* Wenn für $u, v \in C^2(\Omega) \cap C(\overline{\Omega})$ gilt

   $$\mathcal{L}u \leq \mathcal{L}v \text{ in } \Omega, \quad u \leq v \text{ auf } \Gamma,$$

   so folgt $u \leq v$ in Ω.

   Beweis. Für $w := v - u$ ist nach Konstruktion $\mathcal{L}w = \mathcal{L}v - \mathcal{L}u \geq 0$ und auf Γ auch $w \geq 0$. Nach dem Minimumprinzip folgt $\inf_{\mathbf{x}\in\Omega} w(\mathbf{x}) \geq 0$ und folglich $v(\mathbf{x}) \geq u(\mathbf{x})$ für alle $\mathbf{x} \in \Omega$. □

3. *Eindeutigkeit der Lösung:* Die Lösung des Dirichlet-Problems (5.8) ist eindeutig.

   Beweis. Seien u_1 und u_2 zwei Lösungen von (5.8), dann erfüllt $v = u_1 - u_2$ die Gleichung

   $$\mathcal{L}v = 0 \text{ in } \Omega, \quad v = 0 \text{ auf } \Gamma.$$

   Minimum- und Maximumprinzip implizieren

   $$0 = \inf_{\mathbf{z}\in\Omega} v(\mathbf{z}) \leq v(\mathbf{x}) \leq \sup_{\mathbf{z}\in\Omega} v(\mathbf{z}) = 0, \quad \mathbf{x} \in \Omega.$$

   □

4. *Stetige Abhängigkeit von den Randdaten:* Die Lösung des Dirichlet-Problems (5.8) hängt stetig von den Randdaten ab. Sind u_1 und u_2 Lösungen zu verschiedenen-Randwerten, so ist

   $$\max_{\mathbf{x}\in\overline{\Omega}} |u_1(\mathbf{x}) - u_2(\mathbf{x})| = \max_{\mathbf{x}\in\Gamma} |u_1(\mathbf{x}) - u_2(\mathbf{x})|.$$

   Beweis. Für $v = u_1 - u_2$ ist $\mathcal{L}v = 0$. Aus dem Maximumprinzip folgt

   $$v(\mathbf{x}) \leq \max_{\mathbf{z}\in\Gamma} v(\mathbf{z}) \leq \max_{\mathbf{z}\in\Gamma} |v(\mathbf{z})|, \quad \mathbf{x} \in \Omega.$$

   Ebenso liefert das Minimumprinzip die Aussage

   $$v(\mathbf{x}) \geq \min_{\mathbf{z}\in\Gamma} v(\mathbf{z}) \geq -\max_{\mathbf{z}\in\Gamma} |v(\mathbf{z})|, \quad \mathbf{x} \in \Omega.$$

   □

Definition 5.5 Ein Differentialoperator $\mathcal{L}$ der Form (5.9) heißt **gleichmäßig elliptisch**, wenn ein $\alpha > 0$ existiert, so dass die Koeffizientenmatrix $\mathbf{A}(\mathbf{x}) = [a_{i,j}(\mathbf{x})]_{i,j=1}^d$ der Abschätzung

$$\boldsymbol{\xi}^T \mathbf{A}(\mathbf{x}) \boldsymbol{\xi} \geq \alpha \|\boldsymbol{\xi}\|_2^2$$

für alle $\boldsymbol{\xi} \in \mathbb{R}^d$ und $\mathbf{x} \in \Omega$ genügt. Die Zahl α wird als **Elliptizitätskonstante** bezeichnet.

5. *Stetige Abhängigkeit von der rechten Seite:* Sei $\mathcal{L}$ gleichmäßig elliptisch in Ω. Dann gibt es eine nur von Ω und der Elliptizitätskonstante α anhängige Zahl c, so dass für jedes $u \in C^2(\Omega) \cap C(\overline{\Omega})$ gilt

$$|u(\mathbf{x})| \leq \max_{\mathbf{z}\in\Gamma} |u(\mathbf{z})| + c \sup_{\mathbf{z}\in\Omega} |(\mathcal{L}u)(\mathbf{z})|, \quad \mathbf{x} \in \Omega.$$

Beweis. Sei $\Omega \subset \{\mathbf{x} \in \mathbb{R}^d : \|\mathbf{x}\|_2 < R\}$ und setze

$$w(\mathbf{x}) = R^2 - \sum_{i=1}^{d} x_i^2.$$

Im Hinblick auf $w_{x_i,x_j} = -2\delta_{i,j}$ ist offensichtlich

$$\mathcal{L}w \geq 2\alpha \quad \text{und} \quad 0 \leq w \leq R^2 \text{ in } \Omega,$$

wobei α die Elliptizitätskonstante ist. Für

$$v(\mathbf{x}) := \max_{\mathbf{z}\in\Gamma} |u(\mathbf{z})| + w(\mathbf{x}) \frac{1}{2\alpha} \sup_{\mathbf{z}\in\Omega} |(\mathcal{L}u)(\mathbf{z})|$$

ist nach Konstruktion $\mathcal{L}v \geq |\mathcal{L}u|$ in Ω und $v \geq |u|$ auf Γ. Das Vergleichsprinzip liefert $-v(\mathbf{x}) \leq u(\mathbf{x}) \leq v(\mathbf{x})$ für alle $\mathbf{x} \in \Omega$ und wegen $w \leq R^2$ erhalten wir die gewünschte Abschätzung mit $c = R^2/(2\alpha)$. □

6. *Elliptische Operatoren mit Termen der Ordnung* 0*:* Für den allgemeineren Differentialoperator

$$(\mathcal{L}u)(\mathbf{x}) = c(\mathbf{x})u(\mathbf{x}) - \sum_{i,j=1}^{d} a_{i,j}(\mathbf{x}) u_{x_i,x_j}(\mathbf{x}) \quad \text{mit} \quad c(\mathbf{x}) \geq 0$$

gilt ein abgeschwächtes Maximumprinzip. Aus $\mathcal{L}u \leq 0$ folgt

$$\max_{\mathbf{x}\in\overline{\Omega}} u(\mathbf{x}) \leq \max\{0, \max_{\mathbf{x}\in\Gamma} u(\mathbf{x})\}.$$

Beweis. Ein Beweis ist nur für $\mathbf{y} \in \Omega$ und $u(\mathbf{y}) = \sup_{\mathbf{x}\in\Omega} u(\mathbf{x}) > 0$ erforderlich. Dann ist $(\mathcal{L}u)(\mathbf{y}) - c(\mathbf{y})u(\mathbf{y}) \leq (\mathcal{L}u)(\mathbf{y}) \leq 0$. Außerdem ist durch den Hauptteil $\mathcal{L}u - cu$ ein elliptischer Operator der Form (5.9) definiert. Deshalb kann der Beweis wie für Satz 5.4 vollzogen werden. □

5.4 Poisson-Gleichung

Im folgenden wollen wir uns auf die Poisson-Gleichung beschränken. Sei dazu $\Omega \subset \mathbb{R}^d$ ein Gebiet und $f \in C(\Omega)$ und $g \in C(\Gamma)$. Gesucht ist $u \in C^2(\Omega) \cap C(\overline{\Omega})$, so dass gilt

$$-\Delta u = f \text{ in } \Omega, \quad u = g \text{ auf } \Gamma.$$

Definition 5.6 Eine Lösung $u \in C^2(\Omega) \cap C(\overline{\Omega})$ der Poisson-Gleichung ist eine **klassische Lösung**. Gilt speziell $f = 0$, das heißt, ist $\Delta u = 0$ in Ω, so ist u **harmonisch**.

Wir werden mit Lösung stets die klassische Lösung meinen. Um diese zu berechnen, benötigen wir finite Differenzen:

Definition 5.7 Für $u \in C(\mathbb{R}^d)$ und eine Richtung $1 \leq j \leq d$ definieren wir die **Vorwärts-** oder **rechtsseitige Differenz** durch

$$(\partial_j^{+h} u)(\mathbf{x}) := \frac{u(\mathbf{x} + h\mathbf{e}_j) - u(\mathbf{x})}{h},$$

die **Rückwärts-** oder **linksseitige Differenz** durch

$$(\partial_j^{-h} u)(\mathbf{x}) := \frac{u(\mathbf{x}) - u(\mathbf{x} - h\mathbf{e}_j)}{h}$$

und die **symmetrische** oder **zentrale Differenz** durch

$$(\partial_j^{h} u)(\mathbf{x}) := \frac{u(\mathbf{x} + h\mathbf{e}_j) - u(\mathbf{x} - h\mathbf{e}_j)}{2h}.$$

Lemma 5.8 Ist $\{\mathbf{x} + th\mathbf{e}_j : |t| \leq 1\} \subset \overline{\Omega}$ und $u \in C^4(\overline{\Omega})$, dann gilt

$$\begin{aligned} \frac{\partial u}{\partial \mathbf{e}_j}(\mathbf{x}) &= (\partial_j^{\pm h} u)(\mathbf{x}) + R_1^{\pm}, \quad |R_1^{\pm}| \leq \frac{h}{2} \|u\|_{C^2(\overline{\Omega})}, \\ \frac{\partial u}{\partial \mathbf{e}_j}(\mathbf{x}) &= (\partial_j^{h} u)(\mathbf{x}) + R_2, \qquad |R_2| \leq \frac{h^2}{6} \|u\|_{C^3(\overline{\Omega})}, \end{aligned}$$

und

$$\begin{aligned} \frac{\partial^2 u}{\partial \mathbf{e}_j^2}(\mathbf{x}) &= (\partial_j^{-h} \partial_j^{+h} u)(\mathbf{x}) + R_3 \\ &= \frac{u(\mathbf{x} + h\mathbf{e}_j) - 2u(\mathbf{x}) + u(\mathbf{x} - h\mathbf{e}_j)}{h^2} + R_3, \quad |R_3| \leq \frac{h^2}{12} \|u\|_{C^4(\overline{\Omega})}. \end{aligned}$$

Beweis. Es genügt, die Behauptung im Eindimensionalen zu beweisen. Taylor-Entwicklung von u liefert

$$u(x \pm h) = u(x) \pm hu'(x) + \frac{h^2}{2} u''(\xi), \quad \xi \in (x, x \pm h),$$

woraus sofort die erste Aussage folgt. Subtrahieren wir ferner

$$\begin{aligned} u(x - h) &= u(x) - hu'(x) + \frac{h^2}{2} u''(x) - \frac{h^3}{6} u'''(\xi_1), \quad \xi_1 \in (x - h, x), \\ u(x + h) &= u(x) + hu'(x) + \frac{h^2}{2} u''(x) + \frac{h^3}{6} u'''(\xi_2), \quad \xi_2 \in (x, x + h), \end{aligned}$$

so folgt die zweite Aussage

$$u(x+h) - u(x-h) = 2hu'(x) + \frac{h^3}{6}\big(u'''(\xi_2) - u'''(\xi_1)\big).$$

Schließlich folgt aus Addition der drei Gleichungen

$$\begin{aligned}
u(x-h) &= u(x) - hu'(x) + \frac{h^2}{2}u''(x) - \frac{h^3}{6}u'''(x) + \frac{h^4}{24}u^{(4)}(\xi_1), \quad \xi_1 \in (x-h, x),\\
-2u(x) &= -2u(x),\\
u(x+h) &= u(x) + hu'(x) + \frac{h^2}{2}u''(x) + \frac{h^3}{6}u'''(x) + \frac{h^4}{24}u^{(4)}(\xi_2), \quad \xi_2 \in (x, x+h),
\end{aligned}$$

dass

$$\frac{u(x+h) - 2u(x) + u(x-h)}{h^2} = u''(x) + \frac{h^2}{24}\big(u^{(4)}(\xi_1) - u^{(4)}(\xi_2)\big).$$

□

Zur Diskretisierung wird über das Gebiet Ω ein *Gitter* mit Maschenweite h gelegt

$$\begin{aligned}
\Omega_h &:= \big\{\mathbf{x} \in \Omega : \mathbf{x} = h\mathbf{k} \text{ mit } \mathbf{k} \in \mathbb{Z}^d\big\},\\
\Gamma_h &:= \{\mathbf{x} \in \Gamma : \exists 1 \le i \le d \text{ mit } x_i = hk,\ k \in \mathbb{Z}\}.
\end{aligned}$$

In Anlehnung an $\overline{\Omega} = \Omega \cup \Gamma$ setzen wir $\overline{\Omega}_h := \Omega_h \cup \Gamma_h$. Punkte aus Γ_h werden *Randpunkte* genannt. Ein Gitterpunkt $\mathbf{x} \in \Omega_h$, der einen Nachbarn aus Γ_h besitzt, heißt *randnah*. Alle anderen Punkte aus Ω_h sind *randfern*. Ist $\overline{\Omega}$ die Vereinigung von Würfeln der Kantenlänge h, so sprechen wir von einem *Würfelgebiet*. In diesem Fall besitzen dann auch alle Rand- und randnahen Punkte immer den Abstand h zu ihren Nachbarn.

In den Randpunkten $\mathbf{x}$ aus Γ_h ist $u(\mathbf{x})$ durch die Randwerte $g(\mathbf{x})$ vorgegeben. Hingegen erhält man für jeden Punkt $\mathbf{x}$ aus Ω_h eine Gleichung für $u(\mathbf{x})$, indem man die Poisson-Gleichung durch Differenzenquotienten approximiert. In jedem randfernen Gitterpunkt $\mathbf{x}$ diskretisieren wir den Laplace-Operator durch $(\Delta_h u)(\mathbf{x}) := \sum_{i,j=1}^{d} (\partial_i^{-h}\partial_j^{+h}u)(\mathbf{x})$, wobei sich

$$(\Delta u)(\mathbf{x}) = \sum_{i,j=1}^{d} (\partial_i^{-h}\partial_j^{+h}u)(\mathbf{x}) + \mathcal{O}(h^2)$$

ergibt. Für $d = 2$ erhält man durch den sogenannten *5-Punkte-Differenzenstern*

$$\begin{bmatrix} \alpha_{NW} & \alpha_N & \alpha_{NO} \\ \alpha_W & \alpha_Z & \alpha_O \\ \alpha_{SW} & \alpha_S & \alpha_{SO} \end{bmatrix}_\star = \frac{1}{h^2}\begin{bmatrix} & -1 & \\ -1 & 4 & -1 \\ & -1 & \end{bmatrix}_\star.$$

Dieser ist für Würfelgebiete ausreichend.

In beliebigen Gebieten muss der Differenzenquotient für randnahe Punkte entsprechend modifiziert werden. Für $u \in C^3(\overline{\Omega})$ erhält man durch Taylor-Entwicklung in einer Dimension

$$u_{xx} = \frac{2}{h_O(h_O + h_W)}u_O - \frac{2}{h_O h_W}u_Z + \frac{2}{h_W(h_O + h_W)}u_W + \mathcal{O}(h)$$

und in zwei Dimensionen

$$\Delta_h u = \frac{2}{h_O(h_O+h_W)}u_O + \frac{2}{h_W(h_O+h_W)}u_W + \frac{2}{h_S(h_S+h_N)}u_S + \frac{2}{h_N(h_S+h_N)}u_N - \left(\frac{2}{h_O h_W} + \frac{2}{h_S h_N}\right)u_Z + \mathcal{O}(h).$$

Dies ist die *Shortley-Weller-Approximation*.

Beispiel 5.9 *Eindimensionaler Fall:* Sei

$$-u_{xx} = f \text{ in } (a,b), \quad u(a) = \alpha, \quad u(b) = \beta.$$

Für $h = (b-a)/n$ und $x_i = a + hi$, $i = 1, \ldots, n-1$, setzen wir $u_i = u(x_i)$ und $f_i = f(x_i)$. Dann erhalten wir das lineare Gleichungssystem

$$\frac{1}{h^2}\begin{bmatrix} 2 & -1 & & & \\ -1 & 2 & -1 & & \\ & \ddots & \ddots & \ddots & \\ & & -1 & 2 & -1 \\ & & & -1 & 2 \end{bmatrix}\begin{bmatrix} u_1 \\ u_2 \\ \vdots \\ u_{n-2} \\ u_{n-1} \end{bmatrix} = \begin{bmatrix} f_1 + \alpha/h^2 \\ f_2 \\ \vdots \\ f_{n-2} \\ f_{n-1} + \beta/h^2 \end{bmatrix}$$

Zweidimensionales Würfelgebiet: Zur Lösung der Poisson-Gleichung im Einheitsquadrat

$$-\Delta u = f \text{ in } \Omega, \quad u = 0 \text{ auf } \Gamma$$

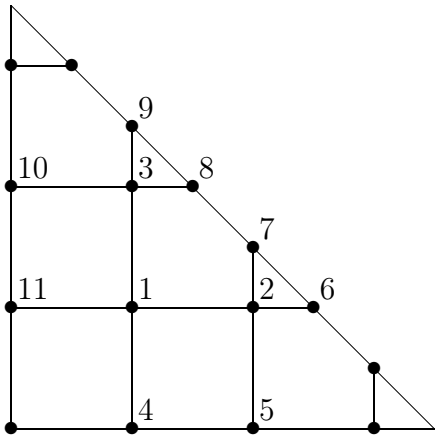
überziehen wir Ω mit einem Gitter der Maschenweite $h = 1/n$. Das entstehende Gleichungssystem wird übersichtlicher bei Benutzung von Doppelindizes $u_{i,j} = u(ih, jh)$, $1 \le i, j < n$. Es ergibt sich das Gleichungssystem

$$4u_{i,j} - u_{i-1,j} - u_{i+1,j} - u_{i,j-1} - u_{i,j+1} = f_{i,j}, \quad 1 \le i, j < n$$

mit $f_{i,j} = h^2 f(ih, jh)$. Die Terme mit Indizes 0 oder n gelten als nicht geschrieben.
Im Fall von Würfelgebieten ist die Systemmatrix immer symmetrisch. Ordnet man die Indizes schachbrettartig an (zum Beispiel durch abwechselndes rotes und scharzes Einfärben), so erhält man ein Gleichungssystem der Form

$$\left[\begin{array}{ccc|ccc} 4 & & & & & \\ & \ddots & & & \mathbf{A} & \\ & & 4 & & & \\ \hline & & & 4 & & \\ & \mathbf{A}^T & & & \ddots & \\ & & & & & 4 \end{array}\right] \mathbf{u} = \mathbf{f}.$$

Beliebiges zweidimensionales Gebiet: Sei Ω ein rechtwinkliges gleichschenkliges Dreieck mit Katheten der Länge 7:



Zu lösen sei die Laplace-Gleichung mit Dirichlet-Randbedingungen. Für $h = 2$ enthält Ω_h drei Punkte. Es entsteht ein lineares Gleichungssystem für u_1, u_2 und u_3:

$$\begin{aligned} u_1 - \frac{u_2}{4} - \frac{u_3}{4} &= \frac{u_4}{4} + \frac{u_{11}}{4} \\ -\frac{u_1}{6} + u_2 &= \frac{u_5}{6} + \frac{u_6}{3} + \frac{u_7}{3} \\ -\frac{u_1}{6} + u_3 &= \frac{u_8}{3} + \frac{u_9}{3} + \frac{u_{10}}{6}. \end{aligned}$$

Man beachte, dass das System unsymmetrisch ist! △

Zusammengefasst erhalten wir demnach ein *Differenzenverfahren* für die näherungsweise Lösung des Poisson-Problems: suche eine *Gitterfunktion* $u_h : \overline{\Omega}_h \to \mathbb{R}$, so dass

$$\begin{aligned} -(\Delta_h u_h)(\mathbf{x}) &= f(\mathbf{x}) \qquad \text{für alle } \mathbf{x} \in \Omega_h, \\ u_h(\mathbf{x}) &= g(\mathbf{x}) \qquad \text{für alle } \mathbf{x} \in \Gamma_h. \end{aligned} \tag{5.10}$$

Sammelt man alle Unbekannten im Vektor $\mathbf{u}_h$, so führt (5.10) auf ein Gleichungsystem $\mathbf{A}_h \mathbf{u}_h = \mathbf{f}_h$.

5.5 Beliebige Differentialoperatoren

Der allgemeine elliptische Differentialoperator

$$(\mathcal{L}u)(\mathbf{x}) = -\sum_{i,j=1}^{d} a_{i,j}(\mathbf{x}) \frac{\partial^2}{\partial x_i \partial x_j} u(\mathbf{x}) + \sum_{i=1}^{d} b_i(\mathbf{x}) \frac{\partial}{\partial x_i} u(\mathbf{x}) + c(\mathbf{x}) u(\mathbf{x})$$

wird diskretisiert durch

$$(\mathcal{L}_h u)(\mathbf{x}) = \left[-\sum_{i=1}^{d} a_{i,i}(\mathbf{x}) \partial_i^{-h} \partial_i^{+h} - \sum_{\substack{i,j=1 \\ i \neq j}}^{d} a_{i,j}(\mathbf{x}) \partial_i^h \partial_j^h u + \sum_{i=1}^{d} b_i(\mathbf{x}) \partial_i^h + c(\mathbf{x}) \right] u(\mathbf{x}).$$

Falls $u \in C^4(\overline{\Omega})$ gilt dann $|(\mathcal{L}u)(\mathbf{x}) - (\mathcal{L}_h u)(\mathbf{x})| = \mathcal{O}(h^2)$.

Beispiel 5.10 Im Zweidimensionalen ergibt sich

$$\begin{aligned}(\mathcal{L}_h u)(\mathbf{x}) &= \big[-a_{1,1}(\mathbf{x})\partial_1^{-h}\partial_1^{+h} - 2a_{1,2}(\mathbf{x})\partial_1^h\partial_2^h - a_{2,2}(\mathbf{x})\partial_2^{-h}\partial_2^{+h}\big]u(\mathbf{x}) \\ &\qquad + \big[b_1(\mathbf{x})\partial_1^h + b_2(\mathbf{x})\partial_2^h\big]u(\mathbf{x}) + c(\mathbf{x})u(\mathbf{x}) \\ &= \frac{1}{2h^2}\begin{bmatrix} a_{1,2}(\mathbf{x}) & -2a_{2,2}(\mathbf{x}) & -a_{1,2}(\mathbf{x}) \\ -2a_{1,1}(\mathbf{x}) & 4[a_{1,1}(\mathbf{x}) + a_{2,2}(\mathbf{x})] & -2a_{1,1}(\mathbf{x}) \\ -a_{1,2}(\mathbf{x}) & -2a_{2,2}(\mathbf{x}) & a_{1,2}(\mathbf{x}) \end{bmatrix}_\star u(\mathbf{x}) \\ &\qquad + \frac{1}{2h}\begin{bmatrix} 0 & b_2(\mathbf{x}) & 0 \\ -b_1(\mathbf{x}) & 0 & b_1(\mathbf{x}) \\ 0 & -b_2(\mathbf{x}) & 0 \end{bmatrix}_\star u(\mathbf{x}) + \begin{bmatrix} 0 & 0 & 0 \\ 0 & c(\mathbf{x}) & 0 \\ 0 & 0 & 0 \end{bmatrix}_\star u(\mathbf{x}).\end{aligned}$$

△

Ist Ω ein Würfelgebiet, so führt das Randwertproblem

$$\begin{aligned}(\mathcal{L}u)(\mathbf{x}) &= f(\mathbf{x}) \quad \text{für alle } \mathbf{x} \in \Omega, \\ u(\mathbf{x}) &= g(\mathbf{x}) \quad \text{für alle } \mathbf{x} \in \Gamma,\end{aligned}$$

auf das Differenzenverfahren

$$\begin{aligned}(\mathcal{L}_h u_h)(\mathbf{x}) &= f(\mathbf{x}) \quad \text{für alle } \mathbf{x} \in \Omega_h, \\ u_h(\mathbf{x}) &= g(\mathbf{x}) \quad \text{für alle } \mathbf{x} \in \Gamma_h.\end{aligned} \tag{5.11}$$

Dies ist gleichbedeutend mit einem linearen Gleichungssystem $\mathbf{A}_h\mathbf{u}_h = \mathbf{f}_h$ für die unbekannten Gitterwerte $\mathbf{u}_h$. Die Systemmatrix $\mathbf{A}_h$ ist symmetrisch, falls $b_1 = b_2 = 0$ gilt.

5.6 Diskretes Maximumprinzip

Alle verwendeten Differenzensternen entsprechen einer gewichteten Mittelung von Nachbarwerten. Daher kann offensichtlich kein Wert größer sein als das Maximum über alle Nachbarwerte. Dies ist der Spezialfall der Theorie der Differenzensterne, deren Koeffizienten ein bestimmtes Vorzeichenverhalten aufweisen.

Lemma 5.11 (Sternlemma) Sei $k > 1$. Für die Zahlen α_ℓ und p_ℓ, $0 \le \ell \le k$, gelte $\alpha_\ell < 0$ für alle $\ell = 1, 2, \ldots, k$ und

$$\sum_{\ell=0}^{k} \alpha_\ell \ge 0, \qquad \sum_{\ell=0}^{k} \alpha_\ell p_\ell \le 0.$$

Ferner sei $p_0 \ge 0$ oder $\sum_{\ell=0}^{k} \alpha_\ell = 0$. Dann folgt aus $p_0 \ge \max_{1\le\ell\le k}\{p_\ell\}$ die Gleichheit

$$p_0 = p_1 = \cdots = p_k.$$

Beweis. Aus den Voraussetzungen folgt

$$\sum_{\ell=1}^{k} \alpha_\ell(p_\ell - p_0) = \sum_{\ell=0}^{k} \alpha_\ell(p_\ell - p_0) = \sum_{\ell=0}^{k} \alpha_\ell p_\ell - p_0 \sum_{\ell=0}^{k} \alpha_\ell \le 0.$$

In der links stehenden Summe sind alle Summanden wegen $\alpha_\ell < 0$ und $p_\ell - p_0 \leq 0$ nicht negativ. Also hat jeder Summand den Wert 0. Aus $\alpha_\ell \neq 0$ folgt die Behauptung. □

Definition 5.12 Das Gebiet Ω_h heißt **(diskret) zusammenhängend**, , wenn zu jedem Punktepaar $\mathbf{x}, \mathbf{y} \in \Omega_h$ auch ein Verbindungsweg existiert, der entlang den Gitterlinien und ganz in Ω_h verläuft.

Bemerkung: Für genügend kleines h ist das Gebiet Ω_h diskret zusammenhängend.

Satz 5.13 (Diskretes Maximumprinzip) Sei u_h die Lösung der diskreten Differentialgleichung

$$(\mathcal{L}_h u_h)(\mathbf{x}) = f(\mathbf{x}) \leq 0 \quad \text{für alle } \mathbf{x} \in \Omega_h,$$

die von der Diskretisierung der elliptischen Differentialgleichung $\mathcal{L}u = f \leq 0$ in Ω herrührt. Der Differenzenstern zu jedem Gitterpunkt in Ω_h genüge folgenden drei Bedingungnen:

1. Alle Koeffizienten, abgesehen vom Zentrum, sind nicht positiv.
2. Der Koeffizient in Ostrichtung sei negativ: $\alpha_O < 0$.
3. Die Summe aller Koeffizienten ist nicht negativ.

Dann ist

$$\max_{\mathbf{x} \in \overline{\Omega}_h} u_h(\mathbf{x}) \leq \max\{\max_{\mathbf{x} \in \Gamma_h} u_h, 0\}.$$

Wenn das Maximum im Innern angenommen wird, die Koeffizienten in den Hauptrichtungen (also in zwei Dimensionen $\alpha_O, \alpha_W, \alpha_S, \alpha_N$) negativ sind und Ω_h zusammenhängend ist, dann ist u_h konstant.

Beweis. Wenn das Maximum im Punkt $\mathbf{z} \in \Omega_h$, also im Innern, angenommen wird, dann setzte $p_0 = u_h(\mathbf{z}) > 0$ und identifiziere $p_1, p_2, \ldots$ mit den Werten in allen Nachbarpunkten, die im Differenzenstern auftreten. Wegen $\sum_{\ell=0}^{k} \alpha_\ell p_\ell = f(\mathbf{z}) \leq 0$ impliziert das Sternlemma $p_0 = p_1 = \cdots = p_k$, das heißt, $u_h(\mathbf{z})$ stimmt mit allen Nachbarn überein.

Nun marschieren wir zum Rand: wir wiederholen dieses Argument solange im jeweils östlichen Nachbarn, bis wir am Rand angekommen sind.

Wenn Ω_h zusammenhängt, können wir gemäß der Voraussetzung das obige Argument solange in alle Hauptrichtungen anwenden, bis alle Punkte von $\overline{\Omega}_h$ erreicht sind. □

Bemerkung: Wenn man als dritte Bedingung sogar verlangt, dass die Summe aller Koeffizienten des Sterns 0 ergibt, dann folgt das strenge Maximumprinzip $\max_{\mathbf{x} \in \overline{\Omega}_h} u_h(\mathbf{x}) \leq \max_{\mathbf{x} \in \Gamma_h} u_h$.

Aus dem diskreten Maximumprinzip kann man genau die selben Folgerungen schließen wie aus dem kontinuierlichen Maximumprinzip. Insbesondere sei auf das Vergleichsprinzip und die stetige Abhängigkeit von den Daten f und g hingewiesen. Eine weitere wollen wir explizit benennen:

Proposition 5.14 Wenn das diskrete Maximumprinzip gilt, ist das Gleichungssystem $\mathbf{A}_h\mathbf{u}_h = \mathbf{f}_h$ eindeutig lösbar.

5.7 Konvergenz

Auf Ω_h und $\overline{\Omega}_h$ definieren wir die Maximumnorm durch

$$\|v_h\|_{\Omega_h} := \max_{\mathbf{x}_i \in \Omega_h} |v_h(\mathbf{x}_i)|, \qquad \|v_h\|_{\overline{\Omega}_h} := \max_{\mathbf{x}_i \in \overline{\Omega}_h} |v_h(\mathbf{x}_i)|.$$

Definition 5.15 Das Differenzenverfahren (5.11) heißt

- **konvergent** mit der Ordnung p, wenn
$$\|u - u_h\|_{\overline{\Omega}_h} = \mathcal{O}(h^p),$$
- **konsistent** mit der Ordnung p, wenn
$$\|\mathcal{L}_h u - \mathcal{L}u\|_{\Omega_h} = \mathcal{O}(h^p),$$
- **stabil** (bzgl. der rechten Seite), wenn eine Konstante $C_s > 0$ existiert, so dass für alle Gitterfunktionen v_h mit $v_h = 0$ am Rand gilt
$$\|v_h\|_{\overline{\Omega}_h} \leq C_s \|\mathcal{L}_h v_h\|_{\Omega_h}.$$

Satz 5.16 Ist ein Differenzenverfahren stabil und konsistent mit der Ordnung p, dann ist es auch konvergent mit der Ordnung p.

Beweis. Es gilt

$$\|u - u_h\|_{\overline{\Omega}_h} \leq C_s \|\mathcal{L}_h(u - u_h)\|_{\Omega_h} = C_s \|\mathcal{L}_h u - \mathcal{L}_h u_h\|_{\Omega_h}.$$

Wegen $(\mathcal{L}_h u_h)(\mathbf{x}) = f(\mathbf{x}) = (\mathcal{L}u)(\mathbf{x})$ für alle $\mathbf{x} \in \Omega_h$, folgt

$$\underbrace{\|u - u_h\|_{\overline{\Omega}_h}}_{\text{Diskretisierungsfehler}} \leq C_s \underbrace{\|\mathcal{L}_h u - \mathcal{L}u\|_{\Omega_h}}_{\text{Konsistenzfehler}} = \mathcal{O}(h^p).$$

□

Beispiel 5.17 Auf Würfelgebieten ist wegen $\|\mathcal{L}_h u - \mathcal{L}u\|_{\Omega_h} = \mathcal{O}(h^2)$ sind die Differenzenverfahren (5.10) und (5.11) konsistent mit der Ordnung 2. Da die Shortley-Weller-Approximation nur von der Ordnung 1 ist, ist hingegen das Verfahren (5.10) für allgemeine Gebiete jedoch nur konsistent mit der Ordnung 1. △

Bemerkung: Stabilität bedeutet nichts anderes, als dass $\|\mathbf{A}_h^{-1}\|_\infty \leq C_s$ unabhängig von der Maschenweite h ist. Das sieht man wie folgt: Bezeichnen $\mathbf{v}_h$ und $\mathbf{w}_h$ die Vektoren der

Werte der Gitterfunktion $v_h|_{\Omega_h}$ und $\mathcal{L}_h v_h$, dann folgt $\mathbf{w}_h = \mathbf{A}_h \mathbf{v}_h$. Die Stabilitätsbedingung kann nun übersetzt werden gemäß

$$\|v_h\|_{\overline{\Omega}_h} = \|\mathbf{v}_h\|_\infty = \|\mathbf{A}_h^{-1}\mathbf{w}_h\|_\infty \leq C_s \|\mathbf{w}_h\|_\infty = C_s \|\mathbf{A}_h \mathbf{v}_h\|_\infty = \|\mathcal{L}_h v_h\|_{\Omega_h}.$$

Hieraus folgt das Behauptete, da diese Ungleichung für beliebe Vektoren $\mathbf{w}_h$ gilt.

Lemma 5.18 Sei $\Omega \subset \{\mathbf{x} \in \mathbb{R}^d : \|\mathbf{x}\|_2 < R\}$ enthalten. Die Gitterfunktion v_h sei Lösung der Gleichung

$$\begin{aligned} -(\Delta_h v_h)(\mathbf{x}) &= 1 \quad \text{für alle } \mathbf{x} \in \Omega_h, \\ v_h(\mathbf{x}) &= 0 \quad \text{für alle } \mathbf{x} \in \Gamma_h. \end{aligned}$$

Dann gilt

$$0 \leq v_h(\mathbf{x}) \leq \frac{1}{2d}(R^2 - \|\mathbf{x}\|_2^2), \quad \mathbf{x} \in \overline{\Omega}_h.$$

Beweis. Man betrachte die Funktion $w(\mathbf{x}) = \frac{1}{2d}(R^2 - \|\mathbf{x}\|_2^2)$. Da w ein Polynom zweiten Grades ist, verschwinden die bei der Bildung des Differenzensterns vernachlässigten Ableitungen, das heißt, es gilt $-\Delta_h w = -\Delta w = 1$ in Ω_h. Außerdem ist $w \geq 0$ auf Γ_h. Aus dem diskreten Vergleichsprinzip folgt daher $v_h(\mathbf{x}) \leq w(\mathbf{x})$ für alle $\mathbf{x} \in \overline{\Omega}_h$. □

Dieses Lemma impliziert die Stabilität: Sei w_h mit $w_h = 0$ auf Γ_h beliebig, dann folgt

$$-\frac{(\Delta_h w_h)(\mathbf{x})}{\|\Delta_h w_h\|_{\Omega_h}} \leq 1 = -(\Delta_h v_h)(\mathbf{x}), \quad \mathbf{x} \in \Omega_h.$$

Das diskrete Vergleichsprinzip liefert sofort

$$\frac{w_h(\mathbf{x})}{\|\Delta_h w_h\|_{\Omega_h}} \leq v_h(\mathbf{x}) \leq \frac{1}{2d}(R^2 - \|\mathbf{x}\|_2^2), \quad \mathbf{x} \in \overline{\Omega}_h.$$

dies bedeutet

$$\|w_h\|_{\overline{\Omega}_h} \leq \frac{R^2}{2d} \|\Delta_h w_h\|_{\Omega_h}.$$

Korollar 5.19 Die Lösung der Poisson-Gleichung efülle $u \in C^4(\overline{\Omega})$. Dann konvergiert das Differenzenverfahren (5.10) und es gilt

$$\|u - u_h\|_{\overline{\Omega}_h} \leq \mathcal{O}(h^p).$$

mit $p = 2$ im Falle von Würfelgebieten und $p = 1$ im Falle von allgemeinen Gebieten.

Bemerkung: Ist der Differentialoperator

$$(\mathcal{L}u)(\mathbf{x}) = \sum_{i,j=1}^{d} a_{i,j}(\mathbf{x}) \frac{\partial^2}{\partial x_i \partial x_j} u(\mathbf{x}) + c(\mathbf{x}) u(\mathbf{x}), \quad c(\mathbf{x}) \geq 0,$$

gleichmäßig elliptisch mit Elliptizitätskonstante $\alpha > 0$, dann folgt für $w(\mathbf{x}) = R^2 - \|\mathbf{x}\|_2^2$ dass $(\mathcal{L}_h w)(\mathbf{x}) = (\mathcal{L}w)(\mathbf{x}) \geq 2\alpha$ für alle $\mathbf{x} \in \Omega_h$ (vergleiche fünfte Folgerung des

kontinuierlichen Maximumprinzips). Erfüllt der zugehörige Differenzenstern das diskrete Maximumprinzip (und damit das Vergleichsprinzip), so gilt daher Lemma 5.18 mit

$$0 \leq v_h(\mathbf{x}) \leq \frac{1}{4d\alpha}(R^2 - \|\mathbf{x}\|_2^2), \quad \mathbf{x} \in \overline{\Omega}_h.$$

Enthält hingegen $\mathcal{L}$ zusätzlich Terme erster Ordnung, dann gilt das diskrete Maximumprinzip nur für genügend kleine Maschenweiten h. Auf Würfelgebieten erhalten wir folglich für beliebige Differentialoperatoren mit $c(\mathbf{x}) \geq 0$ ebenfalls eine quadratische Konvergenzordnung.

5.8 Iterative Lösung

Die Diskretisierung von Differentialgleichungen führt auf sehr große, dünnbesetzte Gleichungssysteme $\mathbf{Ax} = \mathbf{b}$. Die effizientesten Löser für dünnbesetzte Gleichungssysteme sind Krylov-Raum-Verfahren.
Besitzt der Differenzenstern die Eigenschaft

$$\sum_{\ell=0}^{k} \alpha_\ell \geq 0 \quad \text{und} \quad a_\ell \leq 0 \quad \text{für alle } 1 \leq \ell \leq k,$$

dann folgt aus dem Satz von Gerschgorin, dass alle Eigenwerte > 0 sind, denn

$$0 \neq \lambda \in K = \left\{ z \in \mathbb{C} : |z - \alpha_0| \leq \underbrace{\sum_{\ell=1}^{k} |\alpha_\ell|}_{\leq \alpha_0} \right\}.$$

Ist die Systemmatrix $\mathbf{A}_h$ zusätzlich symmetrisch, dann kann das CG-Verfahren zur iterativen Lösung benutzt werden. Dies tritt beispielsweise auf Würfelgebieten ein bei der Diskretisierung des Differentialoperators

$$(\mathcal{L}u)(\mathbf{x}) = c(\mathbf{x})u(\mathbf{x}) - \sum_{i,j=1}^{d} a_{i,j}(\mathbf{x}) \frac{\partial^2}{\partial x_i \partial x_j} u(\mathbf{x}), \qquad c(\mathbf{x}) \geq 0 \tag{5.12}$$

gemäß Abschnitt 5.5.

GMRES-Verfahren: Das *GMRES-Verfahren* (kurz für *Generalized Minimal Residual-Verfahren*) ist ein sehr schnell konvergierendes Verfahren für nichtsymmetrische Gleichungssysteme, das auf dem Arnoldi-Verfahren beruht. Die Systemmatrix ist unsymmetrisch bei Differentialoperatoren mit Termen erster Ordnung, aber auch bei der Shortley-Weller-Approximation.
Im Arnoldi-Verfahren erhalten wir nach k Schritten die Faktorisierung

$$\mathbf{AW}_k = \mathbf{W}_{k+1}\widetilde{\mathbf{H}}_k,$$

wobei die Spalten von $\mathbf{W}_{k+1} = [\mathbf{w}_1, \mathbf{w}_2, \ldots, \mathbf{w}_k, \mathbf{w}_{k+1}]$ gerade die Orthonomalbasis des Krylov-Raums $\mathcal{K}_{k+1}(\mathbf{A}, \mathbf{z})$ ist und

$$\widetilde{\mathbf{H}}_k = \begin{bmatrix} h_{1,1} & h_{1,2} & h_{1,3} & \ldots & h_{1,k} \\ h_{2,1} & h_{2,2} & h_{2,3} & \ldots & h_{2,k} \\ 0 & h_{3,2} & h_{3,3} & \ldots & h_{3,k} \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ 0 & \ldots & 0 & h_{k,k-1} & h_{k,k} \\ 0 & \ldots & 0 & 0 & h_{k+1,k} \end{bmatrix} \in \mathbb{K}^{(k+1)\times k}$$

eine obere Hessenberg-Matrix ist.

Sei $\mathbf{x}_0$ eine Startnäherung und $\mathbf{z} := \mathbf{r}_0 = \mathbf{b} - \mathbf{A}\mathbf{x}_0$. Wegen $\mathbf{w}_1 = \mathbf{r}_0/\|\mathbf{r}_0\|_2$ folgt dann

$$\begin{aligned} \min_{\mathbf{x}\in\mathbf{x}_0+\mathcal{K}_k(\mathbf{A},\mathbf{r}_0)} \|\mathbf{b} - \mathbf{A}\mathbf{x}\|_2 &= \min_{\mathbf{y}\in\mathbb{K}^k} \|\mathbf{b} - \mathbf{A}(\mathbf{x}_0 + \mathbf{W}_k\mathbf{y})\|_2 \\ &= \min_{\mathbf{y}\in\mathbb{K}^k} \|\mathbf{r}_0 - \mathbf{W}_{k+1}\widetilde{\mathbf{H}}_k\mathbf{y}\|_2 \\ &= \min_{\mathbf{y}\in\mathbb{K}^k} \left\|\mathbf{e}_1\|\mathbf{r}_0\|_2 - \widetilde{\mathbf{H}}_k\mathbf{y}\right\|_2. \end{aligned}$$

Dies bedeutet, die Lösung $\mathbf{y}$ eines $\big((k+1)\times k\big)$-Ausgleichsproblems liefert die neue Iterierte gemäß $\mathbf{x}_k = \mathbf{x}_0 + \mathbf{W}_k\mathbf{y}$. Das Ausgleichsproblem lässt sich schnell duch die k Givens-Rotationen $\mathbf{G}(2,1,\theta_1), \mathbf{G}(3,2,\theta_2), \ldots, \mathbf{G}(k+1,k,\theta_k)$ lösen, wobei das Residuum offensichtlich gerade $\|\mathbf{b} - \mathbf{A}\mathbf{x}_k\|_2 = \|\mathbf{r}_k\|_2 = h_{k+1,k}$ ist.

Algorithmus 5.20 (GMRES-Verfahren)

input: Matrix $\mathbf{A} \in \mathbb{K}^{n\times n}$, rechte Seite $\mathbf{b} \in \mathbb{K}^n$ und Startnäherung $\mathbf{x}_0 \in \mathbb{K}^n$

output: Näherungslösung $\mathbf{x}_k$ mit $\|\mathbf{b} - \mathbf{A}\mathbf{x}_k\|_2 \le \varepsilon$

① Initialisierung: setze $\mathbf{r}_0 := \mathbf{b} - \mathbf{A}\mathbf{x}_0$, $h_{1,0} := \|\mathbf{r}_0\|_2$ und $k := 0$

② solange $(h_{k+1,k} > \varepsilon)$ berechne

$$\begin{gathered} k := k + 1 \\ \mathbf{w}_k := \mathbf{r}_{k-1}/h_{k,k-1} \\ \mathbf{r}_k := \mathbf{A}\mathbf{w}_k \\ h_{1,k} := \mathbf{w}_1^\star\mathbf{r}_k, \ldots, h_{k,k} := \mathbf{w}_k^\star\mathbf{r}_k \\ \mathbf{r}_k := \mathbf{r}_k - \sum_{i=1}^{k} h_{i,k}\mathbf{w}_i \\ h_{k+1,k} := \|\mathbf{r}_k\|_2 \end{gathered}$$

③ löse

$$\mathbf{z} = \arg\min_{\mathbf{y}\in\mathbb{K}^k} \|\mathbf{e}_1 h_{1,0} - \widetilde{\mathbf{H}}_k\mathbf{y}\|_2$$

und setze $\mathbf{x}_k := \mathbf{x}_0 + \mathbf{W}_k\mathbf{z}$

MINRES-Verfahren: Ist $c(\mathbf{x})$ in (5.12) beliebig, so kann die Matrix indefinit symmetrisch sein. Wendet man dann das GMRES-Verfahren auf das Gleichungssystem an, dann

ist die obere Hessenberg-Matrix $\widetilde{\mathbf{H}}_k$ aus Symmetriegründen eine symmetrische Tridiagonalmarix und der Arnoldi-Prozess wird zum Lanzcos-Prozess. Dies führt auf das *MINRES-Verfahren* (kurz für *Minimal Residual-Verfahren*). Wir stellen hier jedoch eine Variante vor, bei der direkt die Iterierte und das Residuum aufdatiert wird.

Lemma 5.21 Seien $\mathbf{d}_0, \mathbf{d}_1, \ldots, \mathbf{d}_{n-1}$ bezüglich $\mathbf{A}^2$ konjugierte Richtungen. Für jedes $\mathbf{x}_0 \in \mathbb{K}^n$ liefert die durch

$$\mathbf{x}_{k+1} = \mathbf{x}_k + \alpha_k \mathbf{d}_k, \qquad k = 0, 1, \ldots, n-1,$$

mit

$$\alpha_k = \frac{\mathbf{d}_k^\star \mathbf{A} \mathbf{r}_k}{\mathbf{d}_k^\star \mathbf{A}^2 \mathbf{d}_k}, \qquad \mathbf{r}_k = \mathbf{b}_k - \mathbf{A}\mathbf{x}_k$$

erzeugte Folge nach (höchstens) n Schritten die Lösung $\mathbf{x}_n = \mathbf{A}^{-1}\mathbf{b}$.

Beweis. Mit dem Ansatz

$$\mathbf{x} - \mathbf{x}_0 = \sum_{\ell=0}^{n-1} \alpha_\ell \mathbf{d}_\ell$$

erhalten wir wegen den Orthogonalitätsrelationen

$$\mathbf{d}_k^\star \mathbf{A}^2 (\mathbf{x} - \mathbf{x}_0) = \mathbf{d}_k^\star \mathbf{A}^2 \left(\sum_{\ell=0}^{n-1} \alpha_\ell \mathbf{d}_\ell \right) = \sum_{\ell=0}^{n-1} \alpha_\ell \underbrace{\mathbf{d}_k^\star \mathbf{A}^2 \mathbf{d}_\ell}_{=0 \text{ falls } k \neq \ell} = \alpha_k \mathbf{d}_k^\star \mathbf{A}^2 \mathbf{d}_k$$

die Beziehung

$$\alpha_k = \frac{\mathbf{d}_k^\star \mathbf{A}^2 (\mathbf{x} - \mathbf{x}_0)}{\mathbf{d}_k^\star \mathbf{A}^2 \mathbf{d}_k}.$$

Weil $\mathbf{d}_k$ zu den anderen Richtungen konjugiert ist, gilt

$$\mathbf{d}_k \mathbf{A}^2 (\mathbf{x}_k - \mathbf{x}_0) = \mathbf{d}_k^\star \mathbf{A}^2 \left(\sum_{\ell=0}^{k-1} \alpha_\ell \mathbf{d}_\ell \right) = \sum_{\ell=0}^{k-1} \alpha_\ell \underbrace{\mathbf{d}_k^\star \mathbf{A}^2 \mathbf{d}_\ell}_{=0} = 0.$$

Deshalb ist

$$\alpha_k = \frac{\mathbf{d}_k^\star \mathbf{A}(\overbrace{\mathbf{A}\mathbf{x}}^{=\mathbf{b}} - \mathbf{A}\mathbf{x}_k + \mathbf{A}\mathbf{x}_k - \mathbf{A}\mathbf{x}_0)}{\mathbf{d}_k^\star \mathbf{A}^2 \mathbf{d}_k} = \frac{\mathbf{d}_k^\star \mathbf{A}(\mathbf{b} - \mathbf{A}\mathbf{x}_k)}{\mathbf{d}_k^\star \mathbf{A}^2 \mathbf{d}_k} + \underbrace{\frac{\mathbf{d}_k^\star \mathbf{A}^2 (\mathbf{x}_k - \mathbf{x}_0)}{\mathbf{d}_k^\star \mathbf{A}^2 \mathbf{d}_k}}_{=0}$$

$$= \frac{\mathbf{d}_k^\star \mathbf{A} \mathbf{r}_k}{\mathbf{d}_k^\star \mathbf{A}^2 \mathbf{d}_k}.$$

□

Beim MINRES-Verfahren werden die Richtungen $\mathbf{d}_0, \mathbf{d}_1, \ldots, \mathbf{d}_{n-1}$ nicht von vornherein gewählt, sondern sukzessive vermittels einer Dreitermrekursion ermittelt. Das vollständige Verfahren lautet:

Algorithmus 5.22 (MINRES-Verfahren)
input: Matrix $\mathbf{A} \in \mathbb{K}^{n\times n}$, rechte Seite $\mathbf{b} \in \mathbb{K}^n$ und Startnäherung $\mathbf{x}_0 \in \mathbb{K}^n$
output: Folge von Iterierten $\{\mathbf{x}_k\}_{k>0}$

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

② berechne

$$\alpha_k := \frac{\mathbf{d}_k^\star \mathbf{A}\mathbf{r}_k}{\mathbf{d}_k^\star \mathbf{A}^2 \mathbf{d}_k}$$
$$\mathbf{x}_{k+1} := \mathbf{x}_k + \alpha_k \mathbf{d}_k$$
$$\mathbf{r}_{k+1} := \mathbf{r}_k - \alpha_k \mathbf{A}\mathbf{d}_k$$
$$\beta_k := \frac{\mathbf{d}_k^\star \mathbf{A}^3 \mathbf{d}_k}{\mathbf{d}_k^\star \mathbf{A}^2 \mathbf{d}_k} \tag{5.13}$$
$$\gamma_k := \frac{\mathbf{d}_k^\star \mathbf{A}^3 \mathbf{d}_{k-1}}{\mathbf{d}_{k-1}^\star \mathbf{A}^2 \mathbf{d}_{k-1}} \tag{5.14}$$
$$\mathbf{d}_{k+1} := \mathbf{A}\mathbf{d}_k - \beta_k \mathbf{d}_k - \gamma_k \mathbf{d}_{k-1} \tag{5.15}$$

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

Bemerkung: Mithilfe einer zusätzlichen Variablen $\mathbf{s}_k = \mathbf{A}\mathbf{d}_k$ lässt sich das Verfahren so umformulieren, dass nur eine Matrix-Vektor-Multiplikation pro Iteration benötigt wird.

Satz 5.23 Solange $\mathbf{r}_k \neq \mathbf{0}$ ist, produziert das MINRES-Verfahren von Null verschiedene, paarweise $\mathbf{A}^2$ konjugierte Vektoren $\mathbf{d}_0, \mathbf{d}_1, \ldots, \mathbf{d}_k$ und es gilt $\mathcal{K}_{k+1}(\mathbf{A}, \mathbf{r}_0) = \operatorname{span}\{\mathbf{d}_0, \mathbf{d}_1, \ldots, \mathbf{d}_k\}$.

Beweis. Zunächst bemerken wir, dass aus $\mathcal{K}_{k+1}(\mathbf{A}, \mathbf{r}_0) = \mathcal{K}_k(\mathbf{A}, \mathbf{r}_0)$ die Existenz von Zahlen γ_ℓ folgt mit

$$\mathbf{A}(\mathbf{x} - \mathbf{x}_0) = \mathbf{r}_0 = \mathbf{A} \sum_{\ell=0}^{k-1} \gamma_\ell \mathbf{A}^\ell \mathbf{r}_0.$$

Dies bedeutet, dass dann $\mathbf{x} \in \mathbf{x}_0 + \mathcal{K}_k(\mathbf{A}, \mathbf{r}_0)$ und folglich $\mathbf{r}_k = \mathbf{0}$ gelten würden. Daher besitzt $\mathcal{K}_{k+1}(\mathbf{A}, \mathbf{r}_0)$ stets die Dimension $k+1$.
Wir wollen nun den Beweis vermittels Induktion führen. Wegen

$$\mathbf{d}_1^\star \mathbf{A}^2 \mathbf{d}_0 \overset{(5.15)}{=} \mathbf{d}_0^\star \mathbf{A}^3 \mathbf{d}_0 - \beta_0 \mathbf{d}_0^\star \mathbf{A}^2 \mathbf{d}_0 \overset{(5.13)}{=} 0$$

ist die Behauptung für $k = 1$ klar.
Für den Induktionsschritt $k \mapsto k+1$ nehmen wir an, dass alle Aussagen für $k \geq 0$ gelten. Dann folgt aus (5.15)

$$\mathbf{0} \neq \mathbf{d}_{k+1} = \underbrace{\mathbf{A}\mathbf{d}_k}_{\notin \mathcal{K}_{k+1}(\mathbf{A},\mathbf{r}_0)} \underbrace{- \beta_k \mathbf{d}_k - \gamma_k \mathbf{d}_{k-1}}_{\in \mathcal{K}_{k+1}(\mathbf{A},\mathbf{r}_0)} \in \mathcal{K}_{k+2}(\mathbf{A}, \mathbf{r}_0),$$

dies bedeutet

$$\operatorname{span}\{\mathbf{d}_0, \mathbf{d}_1, \ldots, \mathbf{d}_{k+1}\} = \mathcal{K}_{k+2}(\mathbf{A}, \mathbf{r}_0).$$

Für alle $\ell \leq k$ folgt außerdem sofort

$$\mathbf{d}_{k+1}^\star \mathbf{A}^2 \mathbf{d}_\ell \overset{(5.15)}{=} \mathbf{d}_k^\star \mathbf{A}^3 \mathbf{d}_\ell - \underbrace{\beta_k}_{\overset{(5.13)}{=} \frac{\mathbf{d}_k^\star \mathbf{A}^3 \mathbf{d}_k}{\mathbf{d}_k^\star \mathbf{A}^2 \mathbf{d}_k}} \mathbf{d}_k^\star \mathbf{A}^2 \mathbf{d}_\ell - \underbrace{\gamma_k}_{\overset{(5.14)}{=} \frac{\mathbf{d}_k^\star \mathbf{A}^3 \mathbf{d}_{k-1}}{\mathbf{d}_{k-1}^\star \mathbf{A}^2 \mathbf{d}_{k-1}}} \mathbf{d}_{k-1}^\star \mathbf{A}^2 \mathbf{d}_\ell,$$

was für $\ell = k-1, k$ offensichtlich gleich Null ist. Für $\ell < k-1$ sind alle Terme Null aufgrund der Induktionsannahme, wobei man dies im Fall $\mathbf{d}_k^\star \mathbf{A}^3 \mathbf{d}_\ell = 0$ wie folgt einsieht:

$$\mathbf{d}_k^\star \perp_{\mathbf{A}^2} \mathcal{K}_{k-1}(\mathbf{A}, \mathbf{r}_0) \supset \mathbf{A}\mathcal{K}_{k-2}(\mathbf{A}, \mathbf{r}_0) \ni \mathbf{A}\mathbf{d}_\ell.$$

□

Lemma 5.24 Die Iterierten $\mathbf{x}_k$ aus dem MINRES-Verfahren erfüllen

$$\min_{\mathbf{y} \in \mathbf{x}_0 + \mathcal{K}_k(\mathbf{A}, \mathbf{r}_0)} \|\mathbf{b} - \mathbf{A}\mathbf{x}_k\|_2 = \|\mathbf{b} - \mathbf{A}\mathbf{x}_k\|_2.$$

Beweis. Für ein beliebiges $\mathbf{y} \in \mathbf{x}_0 + \mathcal{K}_k(\mathbf{A}, \mathbf{r}_0)$ folgt

$$\|\mathbf{b} - \mathbf{A}\mathbf{y}\|_2 = \|\mathbf{x} - \mathbf{y}\|_{\mathbf{A}^2}^2 = \left\| \sum_{\ell=0}^{n-1} \alpha_\ell \mathbf{d}_\ell - \mathbf{y} \right\|_{\mathbf{A}^2}^2 \geq \left\| \sum_{\ell=k}^{n-1} \alpha_\ell \mathbf{d}_\ell \right\|_{\mathbf{A}^2}^2 = \sum_{\ell=k}^{n-1} |\alpha_\ell|^2 \|\mathbf{d}_\ell\|_{\mathbf{A}^2}^2,$$

wobei nach Konstruktion das Minimum für $\mathbf{x}_k$ angenommen wird. □

Bemerkung: Man kann (5.13)–(5.15) auch ersetzen durch

$$\beta_k := -\frac{\mathbf{d}_k^\star \mathbf{A}\mathbf{r}_{k+1}}{\mathbf{r}_k^\star \mathbf{A}\mathbf{r}_k}, \qquad \mathbf{d}_{k+1} := \mathbf{r}_{k+1} + \beta_k \mathbf{d}_k.$$

In diesem Fall kann jedoch die Suchrichtung $\mathbf{d}_{k+1} = \mathbf{0}$ werden ohne dass $\mathbf{r}_{k+1} = \mathbf{0}$ gilt. Dies bedeutet, der Algorithmus kann zusammenbrechen, bevor die exakte Lösung gefunden ist.

5.9 Vorkonditionierung

Die Stabiltätsanalyse liefert $\|\mathbf{A}^{-1}\|_\infty \leq C_s$, während wir aus der Definition des Differenzenstern schließen, dass $\|\mathbf{A}\|_\infty \leq C_c/h^2$. Hieraus erhalten wir für die Kondition der Systemmatrix $\mathbf{A}$ die Abschätzung

$$\mathrm{cond}_\infty(\mathbf{A}) = \|\mathbf{A}\|_\infty \left\|\mathbf{A}^{-1}\right\|_\infty \leq \frac{C_s C_c}{h^2} \overset{h \to 0}{\longrightarrow} \infty.$$

Dass diese Abschätzung scharf ist, zeigt das folgende Beispiel.

Beispiel 5.25 Wie man leicht nachrechnet, besitzt die bei der Diskretisierung der eindimensionalen Poisson-Gleichung auftretende Systemmatrix $\mathbf{A} \in \mathbb{R}^{(n-1)\times(n-1)}$ (vergleiche Beispiel 5.9) die Eigenwerte

$$\lambda_k = \frac{2}{h^2}\left[1 - \cos\left(\frac{k\pi}{n}\right)\right], \quad k = 1, 2, \ldots, n-1$$

und die Eigenvektoren $\mathbf{v}_k$ mit

$$[\mathbf{v}_k]_\ell = \sin\left(\frac{k\ell\pi}{n}\right), \quad \ell = 1, 2, \ldots, n-1.$$

Wegen

$$\lambda_{\min} = \frac{2}{h^2}\left[1 - \left(1 - \frac{h^2\pi^2}{2} + \mathcal{O}(h^4)\right)\right] = \pi^2 + \mathcal{O}(h^2),$$
$$\lambda_{\max} = \frac{2}{h^2}\left[1 - \left(-1 - \frac{h^2\pi^2}{2} + \mathcal{O}(h^4)\right)\right] = \frac{1}{h^2} + \mathcal{O}(h^2),$$

folgt schließlich

$$\operatorname{cond}_2(\mathbf{A}) = \frac{\lambda_{\max}}{\lambda_{\min}} = \frac{1}{h^2\pi^2} + \mathcal{O}(h^2).$$

△

Um eine Genauigkeit $\varepsilon > 0$ mit dem CG-Verfahren zu erzielen, benötigt man gemäß Satz 2.25

$$K \geq \frac{\log(\varepsilon/2)}{\log q}$$

Iterationen, wobei

$$q = \frac{\sqrt{\operatorname{cond}_2 \mathbf{A}} - 1}{\sqrt{\operatorname{cond}_2 \mathbf{A}} + 1} = \frac{1 - h\pi}{1 + h\pi} = 1 - \frac{2h\pi}{1 + h\pi}.$$

Wegen

$$\log q = \log\left(1 - \frac{2h\pi}{1 + h\pi}\right) \approx -h$$

folgt also, dass man $K = \mathcal{O}(1/h)$ Iterationen benötigt.

Bemerkung: Bezeichnet man die Anzahl der Unbekannten mit n, dann folgt für eine Differentialgleichung auf dem d-dimensionalen Würfel der Zusammenhang $h \approx n^{1/d}$. Folglich besitzt das Berechnen einer Lösung die Gesamtkomplexität $\mathcal{O}(n^{(d+1)/d})$.

Folglich versucht man, durch eine geeignete *Vorkonditionierung* die Konvergenz zu beschleunigen. Dazu wählt man eine hermitesche und positiv definite Matrix $\mathbf{M}$, so dass $\operatorname{cond}_2\left(\mathbf{M}^{-1}\mathbf{A}\right) = \operatorname{cond}_2\left(\mathbf{M}^{-1/2}\mathbf{A}\mathbf{M}^{-1/2}\right)$ klein wird. Das CG-Verfahren wird nun auf das zu $\mathbf{A}\mathbf{u} = \mathbf{f}$ äquivalente System

$$\mathbf{M}^{-1}\mathbf{A}\mathbf{x} = \mathbf{M}^{-1}\mathbf{f}$$

angewendet. Da $\mathbf{M}^{-1}\mathbf{A}$ im allgmeinen nicht mehr hermitesch ist, führt man das CG-Verfahren im modifiziertem Innenprodukt $\langle \mathbf{x}, \mathbf{y}\rangle_{\mathbf{M}} := \mathbf{x}^\star\mathbf{M}\mathbf{y}$ durch, denn hier gilt

$$\langle \mathbf{M}^{-1}\mathbf{A}\mathbf{x}, \mathbf{y}\rangle_{\mathbf{M}} = \mathbf{x}^\star(\mathbf{M}^{-1}\mathbf{A})^\star\mathbf{M}\mathbf{y} = \mathbf{x}^\star\mathbf{M}(\mathbf{M}^{-1}\mathbf{A})\mathbf{y} = \langle \mathbf{x}, \mathbf{M}^{-1}\mathbf{A}\mathbf{y}\rangle_{\mathbf{M}}.$$

Man beachte, dass wegen

$$\left\|\mathbf{M}^{-1}\mathbf{A}\right\|_{\mathbf{M}}^2 = \sup_{\mathbf{x}\neq\mathbf{0}} \frac{\mathbf{x}^\star\mathbf{A}\mathbf{M}^{-1}\mathbf{A}\mathbf{x}}{\mathbf{x}^\star\mathbf{M}\mathbf{x}} \overset{\mathbf{y}:=\mathbf{M}^{1/2}\mathbf{x}}{=} \sup_{\mathbf{y}\neq\mathbf{0}} \frac{\mathbf{x}^\star\mathbf{M}^{-1/2}\mathbf{A}\mathbf{M}^{-1}\mathbf{A}\mathbf{M}^{-1/2}\mathbf{y}}{\mathbf{y}^\star\mathbf{y}}$$
$$= \left\|\mathbf{M}^{-1/2}\mathbf{A}\mathbf{M}^{-1/2}\right\|_2^2,$$

und analog für die Inverse, in der Fehlerabschätzung des vorkonditionierten CG-Verfahrens tatsächlich die Konditionszahl $\operatorname{cond}_2\left(\mathbf{M}^{-1}\mathbf{A}\right)$ auftritt. Das fertige Verfahren nennt man *PCG-Verfahren* (kurz für *preconditioned conjugate gradients*).

Algorithmus 5.26 (PCG-Verfahren)

input: Matrix $\mathbf{A} \in \mathbb{K}^{n\times n}$, Vorkonditionierer $\mathbf{M} \in \mathbb{K}^{n\times n}$, rechte Seite $\mathbf{b} \in \mathbb{K}^n$ und Startnäherung $\mathbf{x}_0 \in \mathbb{K}^n$

output: Folge von Iterierten $\{\mathbf{x}_k\}_{k>0}$

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

② berechne

$$\begin{aligned}
\alpha_k &:= \frac{\mathbf{s}_k^\star \mathbf{r}_k}{\mathbf{d}_k^\star \mathbf{A}\mathbf{d}_k} \\
\mathbf{x}_{k+1} &:= \mathbf{x}_k + \alpha_k \mathbf{d}_k \\
\mathbf{s}_{k+1} &:= \mathbf{s}_k - \alpha_k \mathbf{A}\mathbf{d}_k \\
\mathbf{r}_{k+1} &:= \mathbf{M}^{-1}\mathbf{s}_{k+1} \\
\beta_k &:= \frac{\mathbf{s}_{k+1}^\star \mathbf{r}_{k+1}}{\mathbf{s}_k^\star \mathbf{r}_k} \\
\mathbf{d}_{k+1} &:= \mathbf{r}_{k+1} + \beta_k \mathbf{d}_k
\end{aligned}$$

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

Bemerkung: Optimal ist eine Vorkonditionierung mit einer Konditionszahl $\operatorname{cond}_2\left(\mathbf{M}^{-1}\mathbf{A}\right)$, welche klein und möglichst unabhängig von der Maschenweite h ist. Dies bedeutet, dass $\mathbf{M}$ möglichst "nahe" an $\mathbf{A}$ gewählt werden sollte. Zusätzlich muss $\mathbf{M}^{-1}$ jedoch auch billig anzuwenden sein.

Eine einfache Vorkonditionierung bietet die ILU-Zerlegung (kurz für *inclomplete LU-decomposition*). Da bei einer gewöhnlichen Cholesky-Zerlegung $\mathbf{A} = \mathbf{L}\mathbf{L}^\star$ die Dreicksmatrix $\mathbf{L}$ weit mehr Nichtnullelemente besitzt als $\mathbf{A}$, rechnet man die Cholesky-Zerlegung nur auf einer Untermenge $E \subseteq \{(i,j) : 1 \le i,j \le n\}$ aus. Geignete Mengen E kann man über Sternmuster definieren, zum Beispiel

$$\begin{bmatrix} 0 & \star & 0 \\ \star & \star & \star \\ 0 & \star & 0 \end{bmatrix}_\star, \qquad \begin{bmatrix} \star & \star & 0 \\ \star & \star & \star \\ 0 & \star & \star \end{bmatrix}_\star, \qquad \begin{bmatrix} \star & \star & \star \\ \star & \star & \star \\ \star & \star & \star \end{bmatrix}_\star.$$

Gemäß (1.11) sind die Diagonalelemente dann gegeben durch

$$\ell_{j,j} = \sqrt{a_{j,j} - \sum_{\substack{k<j \\ (j,k)\in E}} \ell_{j,k}^2} > 0, \quad 1 \le j \le n,$$

und alle Nebendiagonalelemente aus E durch

$$\ell_{i,j} = \frac{1}{\ell_{j,j}}\left(a_{i,j} - \sum_{\substack{k<j \\ (i,k),(j,k)\in E}} \ell_{i,k}\ell_{j,k}\right), \quad j < i \le n \text{ mit } (i,j) \in E.$$

Insgesamt erhält man also eine Zerlegung $\mathbf{A} = \widetilde{\mathbf{L}}\widetilde{\mathbf{L}}^\star + \mathbf{R}$ mit einer Restmatrix $\mathbf{R} \neq \mathbf{0}$.

5.10 Mehrgitterverfahren

Definition 5.27 Ist $\mathbf{A} \in \mathbb{K}^{m\times n}$ und $\mathbf{B} \in \mathbb{K}^{p\times q}$, so ist das **Tensorprodukt** oder **Kronecker-Produkt** $\mathbf{C} = \mathbf{A} \otimes \mathbf{B} \in \mathbb{K}^{mp\times nq}$ definiert durch

$$\mathbf{C} = \big[a_{i,j}\mathbf{B}\big]_{i=1,\dots,n,j=1,\dots,m} = \begin{bmatrix} a_{1,1}\mathbf{B} & \dots & a_{1,m}\mathbf{B} \\ \vdots & & \vdots \\ a_{n,1}\mathbf{B} & \dots & a_{n,m}\mathbf{B} \end{bmatrix}.$$

Lemma 5.28 Für $\mathbf{A} \in \mathbb{K}^{m\times n}, \mathbf{B} \in \mathbb{K}^{n\times p}, \mathbf{C} \in \mathbb{K}^{q\times r}, \mathbf{D} \in \mathbb{K}^{r\times s}$ gilt

$$(\mathbf{A} \otimes \mathbf{C}) \cdot (\mathbf{B} \otimes \mathbf{D}) = \mathbf{AB} \otimes \mathbf{CD}.$$

Beweis. Wegen

$$\mathbf{A} \otimes \mathbf{C} = [a_{i,j}\mathbf{C}]_{i,j}, \qquad \mathbf{B} \otimes \mathbf{D} = [b_{i,j}\mathbf{D}]_{i,j}$$

besteht $\mathbf{F} = (\mathbf{A} \otimes \mathbf{C}) \cdot (\mathbf{B} \otimes \mathbf{D})$ aus den Blockmatrizen

$$\mathbf{F}_{i,j} := \sum_{k=1}^{n} (a_{i,k}\mathbf{C})(b_{k,j}\mathbf{D}) = \sum_{k=1}^{n} a_{i,k}b_{k,j} \cdot \mathbf{C} \cdot \mathbf{D} = \left(\sum_{k=1}^{n} a_{i,k}b_{k,j}\right) \cdot \mathbf{CD},$$
$$1 \le i \le m, \quad 1 \le j \le p.$$

Andererseits besteht $\mathbf{G} = \mathbf{AB} \otimes \mathbf{CD}$ aus den Blockmatrizen

$$\mathbf{G}_{i,j} := g_{i,j} \cdot \mathbf{CD} = \left(\sum_{k=1}^{n} a_{i,k}b_{k,j}\right) \cdot \mathbf{CD}, \quad 1 \le i \le m, \quad 1 \le j \le p,$$

woraus sich die Behauptung ergibt. □

Folgerungen:

1. Es gilt $(\mathbf{A} \otimes \mathbf{B})^{-1} = \mathbf{A}^{-1} \otimes \mathbf{B}^{-1}$.
2. Sind $(\lambda, \mathbf{v})$ und $(\mu, \mathbf{w})$ Eigenpaare von $\mathbf{A}$ bzw. $\mathbf{B}$, dann ist $(\lambda\mu, \mathbf{v} \otimes \mathbf{w})$ ein Eigenpaar von $\mathbf{A} \otimes \mathbf{B}$.
3. Seien $\mathbf{A} \in \mathbb{K}^{n\times n}$ bzw. $\mathbf{B} \in \mathbb{K}^{m\times m}$ und $(\lambda, \mathbf{v})$ bzw. $(\mu, \mathbf{w})$ zugehörige Eigenpaare. Dann ist $(\mu + \lambda, \mathbf{v} \otimes \mathbf{w})$ ein Eigenpaar von $\mathbf{A} \otimes \mathbf{I}^{(m\times m)} + \mathbf{I}^{(n\times n)} \otimes \mathbf{B}$.

Wir betrachten die Diskretisierung der Poisson-Gleichnung auf dem Einheitsquadrat durch den 5-Punkte-Stern zur Schrittweite $h = 1/n$ (vergleiche Beispiel 5.9). Bezeichnet

$$\mathbf{L} = \begin{bmatrix} 2 & -1 & & 0 \\ -1 & 2 & \ddots & \\ & \ddots & \ddots & -1 \\ 0 & & -1 & 2 \end{bmatrix} \in \mathbb{R}^{(n-1)\times(n-1)}$$

die Diskretisierung des eindimensionalen Laplace-Operators und $\mathbf{I}$ die Einheitmatrix im $\mathbb{R}^{n-1}$, dann folgt bei zeilenweiser Numerierung der Unbekannten das lineare Gleichungssystem

$$\mathbf{A}\mathbf{u} = (\mathbf{L} \otimes \mathbf{I} + \mathbf{I} \otimes \mathbf{L})\mathbf{u} = \mathbf{f}.$$

Wir wollen dieses Gleichungssystem mit der *Richardson-Iteration*

$$\mathbf{x}_{k+1} = \mathbf{x}_k + \omega(\mathbf{f} - \mathbf{A}\mathbf{x}_k), \quad k = 0, 1, 2, \ldots \tag{5.16}$$

lösen.

Bemerkung: Die Richardson-Iteration ist das Gradientenverfahren zur Minimierung von

$$\Phi(\mathbf{x}) = \frac{1}{2}\mathbf{x}^\star\mathbf{A}\mathbf{x} - \mathbf{x}^\star\mathbf{f} \to \min$$

mit fester Schrittweite ω. Für das Modellproblem und $\omega = 1/4$ stimmt das Richardson-Iteration mit dem *Gesamtschrittverfahren*

$$\mathbf{x}_{k+1} = \mathbf{D}^{-1}[\mathbf{f} + (\mathbf{L} + \mathbf{R})\mathbf{x}_k], \quad k = 0, 1, 2, \ldots$$

überein, wobei $\mathbf{A} = \mathbf{D} - \mathbf{L} - \mathbf{R}$.

Die Eigenpaare von $\mathbf{L}$ sind gemäß Beispiel 5.25

$$\lambda_i = 2 - 2\cos\left(\frac{i\pi}{n}\right), \qquad \mathbf{v}_i = \left[\sin\left(\frac{ik\pi}{n}\right)\right]_{k=1}^{n-1}, \qquad i = 1, 2, \ldots, n-1.$$

Folglich besitzt die Iterationsmatrix $\mathbf{I} - \omega\mathbf{A}$ die Eigenwerte

$$\begin{aligned}\mu_{i,j} &= 1 - \omega\Bigg[\underbrace{2 - 2\cos\left(\frac{i\pi}{n}\right)}_{=4\sin^2(i\pi/(2n))} + \underbrace{2 - 2\cos\left(\frac{j\pi}{n}\right)}_{=4\sin^2(j\pi/(2n))}\Bigg] \\ &= 1 - 4\omega\left[\sin^2\left(\frac{i\pi}{2n}\right) + \sin^2\left(\frac{j\pi}{2n}\right)\right]\end{aligned}$$

mit zugehörigen Eigenvektoren $\mathbf{w}_{i,j} := \mathbf{v}_i \otimes \mathbf{v}_j$. Für $0 < \omega \leq 1/4$ sind alle Eigenwerte betragsmäßig kleiner als Eins und die Richardson-Iteration konvergiert. Aber wegen $\|\mathbf{I} - \omega\mathbf{A}\|_2 = \mu_{1,1} = 1 - \mathcal{O}(h^2)$ wird die Konvergenz für $h \to 0$ immer langsamer. Hingegen ist die Konvergenz für alle Eigenwerte mit hohem Frequenzanteil in mindestens in einer Richtung unabhängig von h gegeben durch

$$\max\{|\mu_{i,j}| : i \geq n/2 \text{ oder } j \geq n/2\} \leq 1 - 4\omega\left[\sin^2\left(\frac{\pi}{4}\right) + \sin^2(0)\right] = 1 - 2\omega < 1.$$

Somit haben wir folgendes Ergebnis gezeigt:

Lemma 5.29 Sei

$$V_{\text{osc}} := \operatorname{span}\left\{\mathbf{w}_{i,j} : 1 \leq i, j < n,\ \max\{i, j\} > n/2\right\}$$

der Unterraum der hohen Frequenzen. Für den Startwert $\mathbf{x}_0$ der Richardson-Iteration (5.16) gelte $\mathbf{e}_0 := \mathbf{x} - \mathbf{x}_0 \in V_{\text{osc}}$. Dann liegen alle weiteren Fehler $\mathbf{e}_k := \mathbf{x} - \mathbf{x}_k$ ebenfalls in V_{osc} und es gilt

$$\|\mathbf{e}_{k+1}\|_2 \leq (1 - 2\omega)\|\mathbf{e}_k\|_2.$$

Korollar 5.30 Der Startfehler $\mathbf{e}_0$ sei aufgeteilt in

$$\mathbf{e}_0 = \mathbf{e}_{0,\text{osc}} + \mathbf{e}_{0,\text{glatt}}, \qquad \mathbf{e}_{0,\text{osc}} \in V_{\text{osc}}, \qquad \mathbf{e}_{0,\text{glatt}} \in V_{\text{glatt}} := V_{\text{osc}}^{\perp}.$$

Dann gilt

$$\mathbf{e}_k = \mathbf{e}_{k,\text{osc}} + \mathbf{e}_{k,\text{glatt}}$$

mit

$$\mathbf{e}_{k,\text{osc}} = (\mathbf{I} - \omega\mathbf{A})^k \mathbf{e}_{0,\text{osc}} \in V_{\text{osc}}, \qquad \mathbf{e}_{k,\text{glatt}} = (\mathbf{I} - \omega\mathbf{A})^k \mathbf{e}_{k,\text{glatt}} \in V_{\text{glatt}},$$

und

$$\|\mathbf{e}_{k,\text{osc}}\|_2 \leq (1 - 2\omega)^k \|\mathbf{e}_{0,\text{osc}}\|_2.$$

Folglich werden die hochfrequenten Fehleranteile $\mathbf{e}_{0,\text{osc}}$ mit der Richardson-Iteration schnell gedämpft, während die tieffrequenten Fehleranteile $\mathbf{e}_{0,\text{glatt}}$ sehr langsam reduziert werden. Die Idee des Mehrgitterverfahrens ist es nun, diese tieffrequenten Lösungsanteile auf einem gröberen Gitter zu approximieren.

Zutaten eines Mehrgitterverfahrens:

1. Wir benötigen eine Hierarchie von Gittern

   $$\Omega_0 \subset \Omega_1 \subset \Omega_2 \subset \cdots, \qquad \Omega_j := \Omega_{h_j},$$

   wobei $h_j = 2^{-j} h_0$. Auf jedem Gitter erhalten wir ein Gleichungssystem

   $$\mathbf{A}_j \mathbf{u}_j = \mathbf{f}_j, \quad \mathbf{u}_j \in V_j := \mathbb{R}^{|\Omega_j|}.$$

2. Für den Gittertransfer $\Omega_{j-1} \mapsto \Omega_j$ benötigen wir eine *Prolongation*

   $$\mathbf{P}_j : V_{j-1} \to V_j,$$

   welche eine Gitterfunktion $\mathbf{u}_{j-1} \in V_{j-1}$ auf das feinere Gitter Ω_j transferiert. Naheliegend ist die 9-Punkt-Prolongation

   $$\begin{bmatrix} 1/4 & 1/2 & 1/4 \\ 1/2 & 1 & 1/2 \\ 1/4 & 1/2 & 1/4 \end{bmatrix}_\star,$$

   die einer linearen Interpolation von $\mathbf{u}_{j-1}$ in allen Gitterpunkten von Ω_j entspricht.

3. Umgekehrt benötigen wir für den Gittertransfer $\Omega_j \mapsto \Omega_{j-1}$ eine *Restriktion*

   $$\mathbf{R}_j : V_j \to V_{j-1},$$

   welche einer Gitterfunktion $\mathbf{u}_j \in V_j$ auf das gröbere Gitter Ω_{j-1} einschränkt. Viel besser als die triviale Restriktion

   $$\begin{bmatrix} 0 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 0 \end{bmatrix}_\star$$

ist die 9-Punkt-Restriktion

$$\frac{1}{4}\begin{bmatrix} 1/4 & 1/2 & 1/4 \\ 1/2 & 1 & 1/2 \\ 1/4 & 1/2 & 1/4 \end{bmatrix}_\star ,$$

da diese wegen $4\mathbf{R}_j^\star = \mathbf{P}_j$ die Symmetrie bewahrt. Insbesondere gilt dann $\mathbf{A}_{j-1} = \mathbf{R}_{j-1}\mathbf{A}_j\mathbf{P}_{j-1}$.

4. Als *Glätter* verwendet man ein Iterationsverfahren wie die Richardson-Iteration, das Gesamtschrittverfahren oder das *Einzelschrittverfahren*

$$\mathbf{x}_{k+1} = (\mathbf{D} - \mathbf{L})^{-1}(\mathbf{b} + \mathbf{R}\mathbf{x}_k), \quad k = 0, 1, 2, \ldots .$$

5. Auf dem gröbsten Gitter Ω_0 wird ein direkter *Grobgitterlöser* für das Gleichungssystem $\mathbf{A}_0\mathbf{x}_0 = \mathbf{f}_0$ eingesetzt.

Bemerkung: Obwohl Restriktion und Prolongation mit dem selben Stern beschrieben werden, ist ihre Bedeutung komplett unterschiedlich: bei der Restriktion werden Daten gemittelt, währen sie bei der Prolongation verteilt werden.

Algorithmus 5.31 (Zweigitterverfahren)

input: Diskretisierungslevel j, Vorglättungsschritte ν_1, Nachglättungsschritte ν_2
Startnäherung $\mathbf{x}_{\text{in}} \in \mathbb{K}^n$, Systemmatrizen $\mathbf{A}_{j-1}, \mathbf{A}_j$, rechte Seite $\mathbf{f}_j \in \mathbb{K}^n$

output: approximative Lösung $\mathbf{x}_{\text{out}}$

① wende ν_1-mal den Glätter zur Lösung von $\mathbf{A}_j\mathbf{u}_j = \mathbf{f}_j$ mit Startwert $\mathbf{x}_{\text{in}}$ an und erhalte $\mathbf{x}_j$

② bilde Residuum $\mathbf{r}_j := \mathbf{f}_j - \mathbf{A}_j\mathbf{x}_j$

③ bilde $\mathbf{f}_{j-1} := \mathbf{R}_j\mathbf{r}_j$

④ löse $\mathbf{A}_{j-1}\mathbf{x}_{j-1} = \mathbf{f}_{j-1}$

⑤ bilde $\mathbf{x}_j := \mathbf{x}_j + \mathbf{P}_{j-1}\mathbf{x}_{j-1}$

⑥ wende ν_2-mal den Glätter zur Lösung von $\mathbf{A}_j\mathbf{u}_j = \mathbf{f}_j$ mit Startwert $\mathbf{x}_j$ an und erhalte $\mathbf{x}_{\text{out}}$

Benutzen wir das Richardson-Verfahren als Glätter und setzen wir $\mathbf{e}_{\text{in}} := \mathbf{A}_j^{-1}\mathbf{f}_j - \mathbf{x}_{\text{in}}$, dann ist der Fehler nach Schritt ① genau $\mathbf{e}_{①} = (\mathbf{I} - \omega\mathbf{A}_j)^{\nu_1}\mathbf{e}_{\text{in}}$. Die *Grobgitterkorrektur*, die in den Schritten ②–⑤ berechnet wird, ist gerade $\mathbf{d} = \mathbf{P}_{j-1}\mathbf{A}_{j-1}^{-1}\mathbf{R}_{j-1}\mathbf{A}_j\mathbf{e}_{①}$. Beachtet man noch das Nachglätten, dann folgt

$$\mathbf{e}_{\text{out}} := \mathbf{A}_j^{-1}\mathbf{f}_j - \mathbf{x}_{\text{out}} = (\mathbf{I} - \omega\mathbf{A}_j)^{\nu_2}(\mathbf{I} - \mathbf{P}_j\mathbf{A}_{j-1}^{-1}\mathbf{R}_j\mathbf{A}_j)(\mathbf{I} - \omega\mathbf{A}_j)^{\nu_1}\mathbf{e}_{\text{in}}.$$

Rekursive Anwendung des obigen Algorithmus liefert das Mehrgitterverfahren:

Algorithmus 5.32 (Mehrgitterverfahren)

input: Diskretisierungslevel j, Rekursionparameter γ,
Vorglättungsschritte ν_1, Nachglättungsschritte ν_2
Startnäherung $\mathbf{x}_{\text{in}} \in \mathbb{K}^n$, Systemmatrizen $\mathbf{A}_0, \ldots, \mathbf{A}_j$, rechte Seite $\mathbf{f}_j \in \mathbb{K}^n$

output: approximative Lösung $\mathbf{x}_{\text{out}}$

① falls $j = 0$, dann $\mathbf{x}_{\text{out}} = \mathbf{A}_j^{-1}\mathbf{f}_j$ und stop

② wende ν_1-mal den Glätter zur Lösung von $\mathbf{A}_j\mathbf{u}_j = \mathbf{f}_j$ mit Startwert $\mathbf{x}_{\text{in}}$ an und erhalte $\mathbf{x}_j$

③ bilde Residuum $\mathbf{r}_j := \mathbf{f}_j - \mathbf{A}_j\mathbf{x}_j$

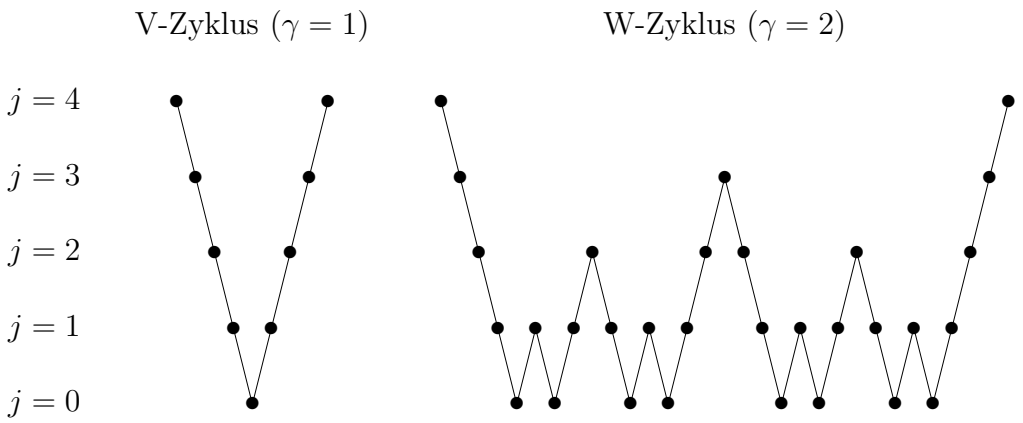
④ bilde $\mathbf{f}_{j-1} := \mathbf{R}_j\mathbf{r}_j$

⑤ wende Mehrgitterverfahren mit Startwert $\mathbf{0}$ zur Lösung von $\mathbf{A}_{j-1}\mathbf{u}_{j-1} = \mathbf{f}_{j-1}$ an und erhalte $\mathbf{x}_{j-1}$

⑥ bilde $\mathbf{x}_j := \mathbf{x}_j + \mathbf{P}_{j-1}\mathbf{x}_{j-1}$

⑦ wende ν_2-mal den Glätter zur Lösung von $\mathbf{A}_j\mathbf{u}_j = \mathbf{f}_j$ mit Startwert $\mathbf{x}_j$ an und erhalte $\mathbf{x}_{\text{out}}$

Je nach Wahl des Rekursionsparamters γ erhalten wir folgendes Schema:



Lemma 5.33 Für $\gamma < 2^d$ ist der Aufwand des Mehrgitterverfahrens 5.32 linear.

Beweis. Für die gewählten Parameter (ν_1, ν_2, γ) bezeichne $T(j)$ den Aufwand des Mehrgitterverfahrens zum Level j. Die Zahl der Unbekannten n_j im Level j skaliert wegen $h_j = 2^{-j}h_0$ wie 2^{dj}. Da das Matrix-Vektor-Produkt linear skaliert, besitzt ein Glättungsschritt die Komplexität $C_G n_j$. Ebenso skalieren die Prolongation und die Restriktion linear, $c_R n_j$ bzw. $c_P n_j$. Damit erhalten wir die Rekursion

$$T(j) \le \underbrace{\nu_1 c_G n_j}_{\text{vorglätten}} + \underbrace{c_R n_j}_{\text{Restriktion}} + \gamma T(j-1) + \underbrace{c_P n_j}_{\text{Prolongation}} + \underbrace{\nu_2 c_G n_j}_{\text{nachglätten}} \le 2^{jd}C + \gamma T(j-1),$$

wobei $T(0) = C_L n_0^3$ gilt. Rekursiv erhalten wir

$$T(j) \le \gamma^j T(0) + \sum_{k=1}^{j} 2^{kd} C \gamma^{j-k} \le 2^{jd} \max\{C, T(0)\} \sum_{k=0}^{j} \Bigg(\underbrace{\frac{\gamma}{2^d}}_{<1} \Bigg)^k = \mathcal{O}(n_j).$$

□

6. Orthogonalpolynome

6.1 Innenprodukträume, Orthogonalbasen und Gramsche Matrizen

Orthogonalität ist ein zentrales Hilfsmittel in der Mathematik, insbesondere in der Numerik. Der Begiff ist vertraut und anschaulich im $\mathbb{R}^n$: Zwei Vektoren $\mathbf{x}$ und $\mathbf{y}$ heißen zueinander orthogonal, kurz $\mathbf{x} \perp \mathbf{y}$, falls gilt

$$\mathbf{x}^\star \mathbf{y} = \sum_{i=1}^{n} x_i y_i = 0.$$

Geometrisch bedeutet dies, dass $\mathbf{x}$ senkrecht auf $\mathbf{y}$ fußt:

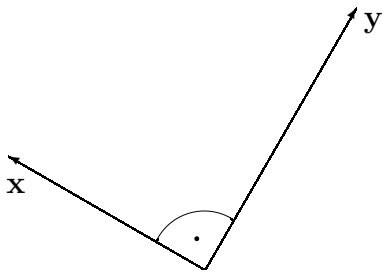


Abbildung 6.1: $\mathbf{x}$ steht senkrecht auf $\mathbf{y}$.

Wir interessieren uns für Funktionenräume und zugehörige Basen, bezüglich deren wir eine gegebene Funktion f darstellen können. Dazu werden wir den Orthogonalitätsbegriff in solchen Räumen einführen.

Definition 6.1 Eine Abbildung $\langle \cdot, \cdot \rangle : X \times X \to \mathbb{R}$ in einem Vektorraum X über $\mathbb{R}$ heißt **Innenprodukt** oder **Skalarprodukt**, falls

1. $\langle f, g \rangle = \langle g, f \rangle \quad \forall f, g \in X$ (Symmetrie)
2. $\langle \alpha f + \beta g, h \rangle = \alpha \langle f, h \rangle + \beta \langle g, h \rangle \quad \forall f, g, h \in X, \forall \alpha, \beta \in \mathbb{R}$ (Linearität)
3. $\langle f, f \rangle > 0 \quad \forall f \in X \setminus \{0\}$ (Definitheit)

Bemerkung: Wegen der Symmetrie ist $\langle \cdot, \cdot \rangle$ auch im zweiten Argument linear. Man nennt eine Abbildung $\langle \cdot, \cdot \rangle$, die die ersten beiden Aussagen erfüllt, daher auch *symmetrische Bilinearform*.

Proposition 6.2 Es gilt die Cauchy-Schwarzsche Ungleichung

$$\langle f, g\rangle^2 \le \langle f, f\rangle \cdot \langle g, g\rangle.$$

Beweis. Im Fall $g = 0$ ist nichts zu zeigen. Sei also $g \neq 0$, dann gilt für jedes $\alpha \in \mathbb{R}$

$$0 \le \langle f - \alpha g, f - \alpha g\rangle = \langle f, f\rangle - 2\alpha\langle f, g\rangle + \alpha^2\langle g, g\rangle.$$

Setzt man speziell $\alpha = \langle f, g\rangle/\langle g, g\rangle$ ein, dann folgt

$$0 \le \langle f, f\rangle - \frac{\langle f, g\rangle^2}{\langle g, g\rangle},$$

woraus die Behauptung folgt. □

Satz 6.3 Ein Innenprodukt $\langle\cdot,\cdot\rangle : X \times X \to \mathbb{R}$ induziert eine Norm in X, die wegen der dritten Aussage aus Definition 6.1 wohldefiniert ist:

$$\|f\| := \sqrt{\langle f, f\rangle}.$$

Beweis. Der einfache Beweis verbleibt dem Leser zur Übung. □

Im $\mathbb{R}^n$ ist die induzierte Norm die übliche Euklidnorm $\|\cdot\|_2$.

Beispiele 6.4

1. Es sei $I = [a, b] \subset \mathbb{R}$ ein Intervall. Der Raum

$$X = \left\{\text{alle Funktionen } f : I \to \mathbb{R} : \int_I f(t)^2\,\mathrm{d}t < \infty\right\},$$

versehen mit dem Innenprodukt

$$\langle f, g\rangle_{L^2(I)} := \int_I f(t)g(t)\,\mathrm{d}t, \qquad \|f\|_{L^2(I)} = \left(\int_I f(t)^2\,\mathrm{d}t\right)^{1/2},$$

heißt $L^2(I)$. Entsprechend definieren wir für eine beschränkte positive Funktion w den Raum $L^2(I; w)$ mit

$$\langle f, g\rangle := \int_I f(t)g(t)w(t)\,\mathrm{d}t.$$

Die Funktion w heißt *Gewichtsfunktion*.

2. Sei

$$X = \left\{F \in L^2(I) : \exists f \in L^2(I) \text{ mit } F(x) = F(a) + \int_a^x f(t)\,\mathrm{d}t\right\}$$

und $a \le x < y \le b$. Wegen

$$|F(y) - F(x)| = \left|\int_x^y f(t)\,\mathrm{d}t\right| \le \int_x^y 1 \cdot |f(t)|\,\mathrm{d}t \underset{\text{CSU}}{\le} \left(\int_x^y 1\,\mathrm{d}t \cdot \int_x^y f(t)^2\,\mathrm{d}t\right)^{1/2}$$
$$\le (y - x)^{1/2}\|f\|_{L^2(I)}$$

ist $F \in X$ immer stetig, sogar Hölder-stetig mit Exponent $1/2$. Das Innenprodukt

$$\langle F, G\rangle_{H^1(I)} := \int_a^b F(t)G(t)\,\mathrm{d}t + \int_a^b f(t)g(t)\,\mathrm{d}t,$$

wobei g zu G gehöre wie f zu F, induziert die Norm

$$\|F\|_{H^1(I)} = \Big(\int_I F(t)^2\,\mathrm{d}t + \int_I f(t)^2\,\mathrm{d}t\Big)^{1/2} = \Big(\|F\|^2_{L^2(I)} + \|f\|^2_{L^2(I)}\Big)^{1/2}.$$

Dieser Raum wird als *Sobolev-Raum* $H^1(I)$ bezeichnet. Die Funktion f wird *schwache Ableitung* von $F \in H^1(I)$ genannt.

Enstprechend definiert man für beliebiges $s \in \mathbb{N}$ die Sobelev-Räume $H^s(I)$ mit Hilfe höherer Ableitungen:

$$H^s(I) = \left\{F \in L^2(I) : \exists f \in H^{s-1}(I) \text{ mit } F(x) = F(a) + \int_a^x f(t)\,\mathrm{d}t\right\} \subset H^{s-1}(I).$$

Man beachte, dass

$$\langle f, g\rangle := \int_I f'(t)g'(t)\,\mathrm{d}t$$

kein Innenprodukt definiert (etwa im $C^1(I)$).

3. Es seien $\{t_n\}_{n=1}^\infty \subset I$ abzählbar viele Punkte und $\{w_n\}_{n=1}^\infty \subset \mathbb{R}_{>0}$ positive Gewichte mit $\sum_{n=1}^\infty w_n < \infty$. Dann definiert

$$\langle f, g\rangle := \sum_{n=1}^\infty w_n f(t_n)g(t_n)$$

ein diskretes Innenprodukt im Raum

$$X = \big\{\text{Polynome über } I \subset \mathbb{R}\big\}.$$

△

Definition 6.5 Eine Basis $\{\phi_i\}_{i=1}^n$ des endlichdimensionalen Teilraums $X_n \subset X$ heißt **Orthogonalbasis**, falls

$$\langle \phi_i, \phi_j\rangle = 0, \quad i \neq j.$$

Gilt zudem $\|\phi_i\|^2 = \langle \phi_i, \phi_i\rangle = 1$ für alle $i = 1, 2, \ldots, n$, dann heißt $\{\phi_i\}$ **Orthonormalbasis**.

Beispiel 6.6 Die Funktionen

$$\left\{\frac{1}{\sqrt{2\pi}}, \frac{\sin x}{\sqrt{\pi}}, \frac{\cos x}{\sqrt{\pi}}, \ldots, \frac{\sin nx}{\sqrt{\pi}}, \frac{\cos nx}{\sqrt{\pi}}\right\}$$

bilden eine Orthonormalbasis der trigonometrischen Polynome vom Grad n bezüglich $L^2([-\pi, \pi])$. △

Für eine Orthonormalbasis von X_n gelten angenehme Darstellungsformeln:

Satz 6.7 Sei $\{\phi_i\}_{i=1}^n$ eine Orthonormalbasis von $X_n \subset X$. Dann gelten die folgenden Aussagen:

1. Für $f \in X_n$ gilt die Entwicklungsformel
$$f = \sum_{i=1}^{n} \langle f, \phi_i \rangle \phi_i.$$
2. Für $f \in X_n$ gilt der Satz von Pythagoras
$$\|f\|^2 = \sum_{i=1}^{n} \langle f, \phi_i \rangle^2.$$
3. Die Bestapproximation von $f \in X$ in X_n ist
$$f_n = \sum_{i=1}^{n} \langle f, \phi_i \rangle \phi_i,$$
dies bedeutet,
$$\|f - f_n\| \leq \|f - g\| \quad \text{für alle } g \in X_n.$$
4. Für alle $f \in X$ gilt die Besselsche Ungleichung
$$\sum_{i=1}^{n} \langle f, \phi_i \rangle^2 \leq \|f\|^2.$$

Beweis. 1. Wegen $f \in X_n$ gilt $f = \sum_{i=1}^n \alpha_i \phi_i$. Daher folgt

$$\langle f, \phi_j \rangle = \Big\langle \sum_{i=1}^{n} \alpha_i \phi_i, \phi_j \Big\rangle = \sum_{i=1}^{n} \alpha_i \langle \phi_i, \phi_j \rangle = \alpha_j \cdot 1.$$

2. Aus der ersten Aussage folgt

$$\|f\|^2 = \langle f, f \rangle = \Big\langle \sum_{i=1}^{n} \alpha_i \phi_i, \sum_{j=1}^{n} \alpha_j \phi_j \Big\rangle = \sum_{i=1}^{n} \sum_{j=1}^{n} \alpha_i \alpha_j \underbrace{\langle \phi_i, \phi_j \rangle}_{=\delta_{i,j}}$$
$$= \sum_{i=1}^{n} \alpha_i^2 = \sum_{i=1}^{n} \langle f, \phi_i \rangle^2$$

3. Für eine beliebige Funktion $g = \sum_{i=1}^{n} \beta_i \phi_i \in X_n$ gilt

$$\begin{aligned}\|f-g\|^2 &= \left\| f - \sum_{i=1}^{n} \beta_i \phi_i \right\|^2 = \left\langle f - \sum_{i=1}^{n} \beta_i \phi_i, f - \sum_{j=1}^{n} \beta_j \phi_j \right\rangle \\ &= \|f\|^2 - 2\sum_{i=1}^{n} \beta_i \underbrace{\langle f, \phi_i \rangle}_{=:\alpha_i} + \sum_{i=1}^{n}\sum_{j=1}^{n} \beta_i \beta_j \langle \phi_i, \phi_j \rangle \\ &= \|f\|^2 - 2\sum_{i=1}^{n} \alpha_i \beta_i + \sum_{i=1}^{n} \beta_i^2 \\ &= \|f\|^2 - \sum_{i=1}^{n} \alpha_i^2 + \underbrace{\sum_{i=1}^{n} (\alpha_i - \beta_i)^2}_{\geq 0}. \end{aligned} \tag{6.1}$$

   Das Minimum wird offensichtlich genau dann angenommen, wenn gilt $\alpha_i = \beta_i$ für alle i.

4. Für $f \in X$ und $\alpha_i := \langle f, \phi_i \rangle$ ergibt sich analog zu (6.1)

$$\begin{aligned} 0 \leq \left\| f - \sum_{i=1}^{n} \alpha_i \phi_i \right\|^2 &= \|f\|^2 - 2\sum_{i=1}^{n} \alpha_i^2 + \sum_{i=1}^{n} \alpha_i^2 = \|f\|^2 - \sum_{i=1}^{n} \alpha_i^2 \\ &= \|f\|^2 - \sum_{i=1}^{n} \langle f, \phi_i \rangle^2. \end{aligned}$$

□

Dieser Satz macht klar, warum Orthogonalität in der Numerik so wichtig ist: Man hat

- unmittelbar die Bestapproximation und
- einfache Basisdarstellungen.

Es bleibt die Frage: Wie erhält man Orthonormalbasen?

Definition 6.8 Sei $\{\phi_i\}_{i=1}^n$ ein beliebiges Funktionensystem von X_n. Dann heißt

$$\mathbf{G} := \left[\langle \phi_i, \phi_j \rangle\right]_{i,j=1}^{n}$$

Gramsche Matrix bezüglich $\{\phi_i\}$.

Proposition 6.9 Die Gramsche Matrix $\mathbf{G}$ ist symmetrisch und positiv semidefinit. Sie ist genau dann positiv definit, wenn die $\{\phi_i\}$ linear unabhängig sind.

Beweis. Die Symmetrie $\mathbf{G} = \mathbf{G}^\star$ folgt sofort aus der Symmetrie des Innenprodukts. Für

$\mathbf{x} = [x_1, x_2, \ldots, x_n]^\star \in \mathbb{R}^n$ folgt

$$\mathbf{x}^\star \mathbf{G} \mathbf{x} = \Big[x_1, x_2, \ldots, x_n\Big] \mathbf{G} \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix} = \Big[x_1, x_2, \ldots, x_n\Big] \begin{bmatrix} \big\langle \phi_1, \sum_{j=1}^n x_j \phi_j \big\rangle \\ \big\langle \phi_2, \sum_{j=1}^n x_j \phi_j \big\rangle \\ \vdots \\ \big\langle \phi_n, \sum_{j=1}^n x_j \phi_j \big\rangle \end{bmatrix}$$
$$= \Big\langle \sum_{i=1}^n x_i \phi_i, \sum_{j=1}^n x_j \phi_j \Big\rangle = \Big\| \sum_{i=1}^n x_i \phi_i \Big\|^2. \tag{6.2}$$

Folglich ist $\mathbf{x}^\star \mathbf{G} \mathbf{x} \geq 0$ für alle $\mathbf{x} \in \mathbb{R}^n$, während "=0" genau dann gilt, wenn $\sum_{i=1}^n x_i \phi_i = 0$. Ist $\{\phi_i\}$ eine Basis des X_n, so bedeutet letzteres gerade $x_i = 0$ für alle $i = 1, 2, \ldots, n$. Ist umgekehrt $\mathbf{x}^\star \mathbf{G} \mathbf{x} > 0$ für alle $\mathbf{x} \neq \mathbf{0}$, dann sind die $\{\phi_i\}$ linear unabhängig. □

Man sieht sofort, dass (6.2) verallgemeinert werden kann:

$$\mathbf{x}^\star \mathbf{G} \mathbf{y} = \Big\langle \sum_{i=1}^n x_i \phi_i, \sum_{j=1}^n y_j \phi_j \Big\rangle. \tag{6.3}$$

Bemerkung: Auf dem Rechner werden Funktion f und g vermittels deren Entwicklungskoeffizienten $\{x_i\}_{i=1}^n$ und $\{y_i\}_{i=1}^n$ bezüglich einer geeigneten Basis $\{\phi_i\}_{i=1}^n$ dargestellt. Dann gilt also gemäß (6.3)

$$\langle f, g \rangle = \mathbf{x}^\star \mathbf{G} \mathbf{y}.$$

Die Gramsche Matrix kann man verwenden, um eine Orthonormalbasis zu konstruieren:

Satz 6.10 Es sei $\{\phi_i\}_{i=1}^n$ eine Basis von X_n und $\mathbf{G}$ die zugehörige Gramsche Matrix. Ist $\mathbf{G} = \mathbf{L}\mathbf{L}^\star$ die Cholesky-Zerlegung von $\mathbf{G}$ und $\mathbf{L}^{-1} = [\ell_{i,j}]_{i,j=1}^n$, dann ist

$$\psi_i = \sum_{j=1}^i \ell_{i,j} \phi_j, \quad i = 1, 2, \ldots, n,$$

eine Orthonormalbasis von X_n. Formal können wir schreiben

$$\begin{bmatrix} \psi_1 \\ \psi_2 \\ \vdots \\ \psi_n \end{bmatrix} = \mathbf{L}^{-1} \begin{bmatrix} \phi_1 \\ \phi_2 \\ \vdots \\ \phi_n \end{bmatrix}.$$

Beweis. Sei $\boldsymbol{\ell}_i$ die i-te Zeile von $\mathbf{L}^{-1}$. Dann folgt mit (6.3)

$$\langle \psi_i, \psi_j \rangle = \Big\langle \sum_{\nu=1}^{i} \ell_{i,\nu}\phi_\nu, \sum_{\mu=1}^{j} \ell_{j,\mu}\phi_\mu \Big\rangle = \Big[\ell_{i,1}, \ell_{i,2}, \ldots, \ell_{i,i}, 0, \ldots, 0\Big]\mathbf{G} \begin{bmatrix} \ell_{j,1} \\ \ell_{j,2} \\ \vdots \\ \ell_{j,j} \\ 0 \\ \vdots \\ 0 \end{bmatrix}$$

$$= \boldsymbol{\ell}_i \mathbf{L}\mathbf{L}^\star \boldsymbol{\ell}_j^\star = \delta_{i,j}.$$

□

Bemerkung: Hinter Satz 6.10 steckt im wesentlichen das Gram-Schmidtsche Orthogonalisierungsverfahren: für jedes i orthogonalisiere ϕ_i gegen alle ϕ_j mit $j < i$ und erhalte ψ_i.

Hat man die Gramsche Matrix $\mathbf{G}$ für eine Basis $\{\phi_i\}_{i=1}^n$ von $X_n \subset X$ zur Hand, dann lässt sich die Bestapproximation $f_n \in X_n$ an eine Funktion $f \in X$ auch direkt berechnen, also ohne Konstruktion einer Orthonormalbasis:

Satz 6.11 Ist $f \in X$, $\{\phi_i\}_{i=1}^n$ eine Basis von X_n und $\mathbf{G} = \big[\langle \phi_i, \phi_j \rangle\big]_{i,j=1}^n$ deren Gramsche Matrix. Die Funktion $f_n = \sum_{i=1}^n x_i \phi_i$ ist genau dann die Bestapproximation an f aus X_n, wenn

$$\mathbf{G}\mathbf{x} = \mathbf{b} \quad \text{für} \quad \mathbf{x} = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix}, \quad \mathbf{b} = \begin{bmatrix} \langle f, \phi_1 \rangle \\ \langle f, \phi_2 \rangle \\ \vdots \\ \langle f, \phi_n \rangle \end{bmatrix}.$$

Beweis. Sei $\mathbf{G} = \mathbf{L}\mathbf{L}^\star$ die Cholesky-Zerlegung von $\mathbf{G}$ und

$$\begin{bmatrix} \psi_1 \\ \psi_2 \\ \vdots \\ \psi_n \end{bmatrix} = \mathbf{L}^{-1} \begin{bmatrix} \phi_1 \\ \phi_2 \\ \vdots \\ \phi_n \end{bmatrix}$$

eine Orthonormalbasis von X_n. Daraus folgt

$$\begin{bmatrix} \langle f, \psi_1 \rangle \\ \langle f, \psi_2 \rangle \\ \vdots \\ \langle f, \psi_n \rangle \end{bmatrix} = \mathbf{L}^{-1} \begin{bmatrix} \langle f, \phi_1 \rangle \\ \langle f, \phi_2 \rangle \\ \vdots \\ \langle f, \phi_n \rangle \end{bmatrix} = \mathbf{L}^{-1}\mathbf{b}.$$

Gemäß der dritten Aussage aus Satz 6.7 gilt

$$f_n = \sum_{i=1}^n \langle f, \psi_i \rangle \psi_i = \Big[\langle f, \psi_1 \rangle, \langle f, \psi_2 \rangle, \ldots, \langle f, \psi_n \rangle\Big] \begin{bmatrix} \psi_1 \\ \psi_2 \\ \vdots \\ \psi_n \end{bmatrix} = \big(\mathbf{L}^{-1}\mathbf{b}\big)^\star \begin{bmatrix} \psi_1 \\ \psi_2 \\ \vdots \\ \psi_n \end{bmatrix}.$$

Andererseits ist

$$f_n = \sum_{i=1}^{n} x_i \phi_i = \begin{bmatrix} x_1, x_2, \dots, x_n \end{bmatrix} \begin{bmatrix} \phi_1 \\ \phi_2 \\ \vdots \\ \phi_n \end{bmatrix} = \begin{bmatrix} x_1, x_2, \dots, x_n \end{bmatrix} \mathbf{L} \begin{bmatrix} \psi_1 \\ \psi_2 \\ \vdots \\ \psi_n \end{bmatrix} = \left(\mathbf{L}^\star \mathbf{x}\right)^\star \begin{bmatrix} \psi_1 \\ \psi_2 \\ \vdots \\ \psi_n \end{bmatrix}.$$

Daraus folgt

$$\mathbf{L}^\star \mathbf{x} = \mathbf{L}^{-1} \mathbf{b},$$

dies bedeutet,

$$\mathbf{L}\mathbf{L}^\star \mathbf{x} = \mathbf{b}.$$

□

6.2 Tschebyscheff-Polynome

Die Tschebyscheff-Polynome sind auf $[-1, 1]$ definiert durch die Gleichung

$$T_n(x) = \cos(n \arccos x), \quad -1 \le x \le 1. \tag{6.4}$$

Obwohl T_n auf den ersten Blick nicht wie ein Polynom aussieht, überzeugt man sich leicht, dass

$$T_0(x) = 1, \quad T_1(x) = x,$$

und — mittels trigonometrischer Identitäten und der Substitution $\xi = \arccos x$ —

$$\begin{aligned} T_{n-1}(x) + T_{n+1}(x) &= \cos\big((n-1)\xi\big) + \cos(\big(n+1)\xi\big) \\ &= \cos n\xi \cos \xi - \sin n\xi \sin(-\xi) + \cos n\xi \cos \xi - \sin n\xi \sin \xi \\ &= 2\cos(\arccos x)\cos(n \arccos x) \\ &= 2xT_n(x). \end{aligned}$$

Mit anderen Worten, es gilt die Dreitermrekursion

$$T_{n+1}(x) = 2xT_n(x) - T_{n-1}(x), \quad n = 1, 2, \dots, \tag{6.5}$$

und man erkennt, dass $T_n(x)$ ein Polynom vom Grad n ist.
Wir werden nun zeigen, dass $\{T_i\}_{i=0}^n$ eine Orthogonalbasis der Polynome vom Grad $\le n$ bezüglich einem speziell gewichteten L^2-Innenprodukt ist.

Satz 6.12 Es gilt

$$\langle T_i, T_j \rangle := \int_{-1}^{1} T_i(x) T_j(x) \frac{1}{\sqrt{1-x^2}} \, \mathrm{d}x = \begin{cases} 0, & \text{falls } i \neq j, \\ \pi, & \text{falls } i = j = 0, \\ \pi/2, & \text{falls } i = j \neq 0. \end{cases}$$

Beweis. Nach Definition ist

$$\langle T_i, T_j \rangle := \int_{-1}^{1} \cos(i \arccos x) \cos(j \arccos x) \frac{1}{\sqrt{1-x^2}} \, \mathrm{d}x.$$

Substitutieren wir

$$x = \cos \xi, \quad \mathrm{d}x = -\sin \xi \, \mathrm{d}\xi = -\sqrt{1-x^2} \, \mathrm{d}\xi,$$

so ergibt sich

$$\begin{aligned} \langle T_i, T_j \rangle &= -\int_{\pi}^{0} \cos(i\xi) \cos(j\xi) \, \mathrm{d}\xi = \int_{0}^{\pi} \cos(i\xi) \cos(j\xi) \, \mathrm{d}\xi \\ &\overset{\text{Symmetrie}}{=} \frac{1}{2} \int_{-\pi}^{\pi} \cos(i\xi) \cos(j\xi) \, \mathrm{d}\xi \overset{\text{Beispiel 6.6}}{=} \begin{cases} 0, & \text{falls } i \neq j, \\ \pi, & \text{falls } i = j = 0, \\ \pi/2, & \text{falls } i = j \neq 0. \end{cases} \end{aligned}$$

□

Aus (6.4) ergibt sich

$$|T_n(x)| \leq 1, \quad -1 \leq x \leq 1,$$

während aus (6.5) folgt

$$T_n(x) = 2^{n-1} x^n + \ldots.$$

Die Tschebyscheff-Polynome haben eine Reihe von extremalen Eigenschaften, die sie für viele Anwendungen in der Numerik interessant machen. Zwei dieser Eigenschaften sollen im folgenden vorgeführt werden.

Satz 6.13

1. Unter allen monischen Polynomen $p_n(x) = x^n + \ldots$ minimiert $2^{1-n} T_n$ die Maximumsnorm $\|p_n\|_{C([-1,1])}$ über $[-1, 1]$.
2. Sei $\xi \notin [-1, 1]$: Unter allen Polynomen p_n vom Grad n mit $p_n(\xi) = 1$ minimiert $p_n = T_n / T_n(\xi)$ die Maximumsnorm $\|p_n\|_{C([-1,1])}$ über $[-1, 1]$.

Beweis. 1. Sei $s_n := 2^{n-1} T_n$ und p_n ein weiteres monisches Polynom vom Grad n. Dann hat $s_n - p_n$ höchstens den Grad $n - 1$. Wir nehmen nun an, dass gilt

$$\|p_n\|_{C([-1,1])} < 2^{1-n} = \|s_n\|_{C([-1,1])}.$$

Da s_n nach (6.4) an den $n + 1$ Stellen $x_k = \cos(k\pi/n)$, $0 \leq k \leq n$, alternierend die Extremalwerte $\pm 2^{1-n}$ annimmt, hat $s_n - p_n$ mindestens n Vorzeichenwechsel, also n Nullstellen. Wegen $\deg(s_n - p_n) \leq n - 1$ muss daher $s_n - p_n \equiv 0$ gelten im Widerspruch zur gemachten Annahme.

2. Wieder nehmen wir an, dass $t_n := T_n / T_n(\xi)$ eine größere Maximumsnorm $\|t_n\|_{C([-1,1])}$ hat als ein anderes Polynom $p_n \in \Pi_n$ mit $p_n(\xi) = 1$. Wieder ergeben sich mindestens n Nullstellen von $t_n - p_n$ in $[-1, 1]$, sowie eine $(n+1)$-te Nullstelle in ξ. Wegen $\deg(t_n - p_n) \leq n$ folgt $t_n \equiv p_n$ im Widerspruch zur Annahme.

□

Bemerkung: Die zweite Aussage aus Satz 6.13 impliziert, dass T_n unter allen Polynomen $p_n \in \Pi_n$ mit $\|p_n\|_{C([-1,1])} = 1$ am schnellsten außerhalb von $[-1,1]$ wächst. Dies folgt punktweise für jedes $\xi \notin [-1,1]$, indem man $T_n/T_n(\xi)$ und $p_n/p_n(\xi)$ mit Hilfe von Satz 6.13 vergleicht:

$$\frac{1}{|p_n(\xi)|} = \left\| \frac{p_n}{p_n(\xi)} \right\|_{C([-1,1])} \geq \left\| \frac{T_n}{T_n(\xi)} \right\|_{C([-1,1])} = \frac{1}{|T_n(\xi)|}.$$

Die Orthogonalität der Tschebyscheff-Polynome und ihre Beschränkung durch 1 in $[-1,1]$ machen sie zu einer beliebten Basis des Polynomraumes. Entwickelt man gemäß Satz 6.7 ein beliebiges Polynom $p_n \in \Pi_n$ in Tschebyscheff-Polynome, so erhält man

$$p_n(x) = \sum_{i=0}^{n} \alpha_i T_i(x). \tag{6.6}$$

Dabei gilt für alle $1 \leq i \leq n$, dass

$$\begin{aligned} |\alpha_i| &= \left| \int_{-1}^{1} p_n(x) \frac{2}{\pi} \cos(i \arccos x) \frac{1}{\sqrt{1-x^2}} \, \mathrm{d}x \right| \\ &= \left| \int_{0}^{\pi} \frac{2}{\pi} p_n(\cos t) \cos(it) \, \mathrm{d}t \right| \\ &\leq \frac{4}{\pi} \|p_n\|_{C([-1,1])} \end{aligned}$$

und entsprechend $|\alpha_0| \leq 1 \cdot \|p_n\|_{C([-1,1])}$. Folglich sind alle Entwicklungskoeffizienten α_i gleichmäßig beschränkt, weshalb die Darstellung (6.6) sehr stabil ist.

Eine Entwicklung der Form (6.6) lässt sich sehr effizient mit Hilfe der Dreitermrekursion (6.5) auswerten:

$$\begin{aligned} p_n(x) &= \underbrace{\alpha_n}_{=:\beta_n} \underbrace{T_n(x)}_{=2xT_{n-1}(x)-T_{n-2}(x)} + \sum_{i=0}^{n-1} \alpha_i T_i(x) \\ &= (\underbrace{2x\beta_n + \alpha_{n-1}}_{=:\beta_{n-1}}) \underbrace{T_{n-1}(x)}_{=2xT_{n-2}(x)-T_{n-3}(x)} + (\alpha_{n-2} - \beta_n) T_{n-2}(x) + \sum_{i=0}^{n-3} \alpha_i T_i(x) \\ &= (\underbrace{2x\beta_{n-1} + \alpha_{n-2} - \beta_n}_{=:\beta_{n-2}}) T_{n-2}(x) + (\alpha_{n-3} - \beta_{n-1}) T_{n-3}(x) + \sum_{i=0}^{n-4} \alpha_i T_i(x) \\ &= \cdots = \beta_1 \underbrace{T_1(x)}_{=x} + (\alpha_0 - \beta_2) \underbrace{T_0(x)}_{=1}. \end{aligned}$$

Dies führt auf den folgenden Algorithmus, der die Darstellung (6.6) in einem Punkt $x \in \mathbb{R}$ in $n+1$ Multiplikationen auswertet:

Algorithmus 6.14 (Clenshaw-Algorithmus)
input: Entwicklungskoeffizienten $\{\alpha_i\}_{i=0}^n$ und Auswertungspunkt $x \in \mathbb{R}$
output: Auswertung $p_n(x) = \sum_{i=0}^n \alpha_i T_i(x)$

① Initialisierung: setze $y := 2x$, $\beta_n := \alpha_n$ und $\beta_{n-1} := y\beta_n + \alpha_{n-1}$

② für alle $i = n-2, n-3, \ldots, 1$ setze $\beta_i := y\beta_{i+1} + \alpha_i - \beta_{i+2}$

③ setze $p_n(x) := x\beta_1 + \alpha_0 - \beta_2$

6.3 Allgemeine Orthogonalpolynome

Im weiteren sei $w(x)$ eine positive integrierbare Funktion über dem Intervall $I \subset \mathbb{R}$ mit

$$\int_I w(x)\,\mathrm{d}x < \infty.$$

Eine solche Funktion w bezeichnen wir als *Gewichtsfunktion* und definieren das Innenprodukt

$$\langle f, g\rangle := \int_I f(x)g(x)w(x)\,\mathrm{d}x \tag{6.7}$$

über dem Raum Π der reellwertigen Polynome. $\Pi_n \subset \Pi$ bezeichne die Teilmenge der Polynome vom Grad $\leq n$.

Satz 6.15 Zu jeder Gewichtsfunktion w und zugehörigem Innenprodukt (6.7) existiert eine eindeutig bestimmte Folge von Polynomen $\{u_n\}_{n=0}^{\infty}$ mit $u_n \in \Pi_n$, für die gilt

$$\langle u_n, u_m\rangle = \delta_{m,n}, \quad u_n(x) = \gamma_n x^n + \cdots \text{ für ein } \gamma_n > 0. \tag{6.8}$$

Die Folge $\{u_n\}$ genügt zudem der Dreitermrekursion

$$a_{n+1}u_{n+1}(x) = (x - b_n)u_n(x) - a_n u_{n-1}(x), \quad n \geq 0, \tag{6.9}$$

wobei $a_n = \gamma_{n-1}/\gamma_n > 0$, $b_n = \langle xu_n, u_n\rangle$, $u_{-1} \equiv 0$ und

$$u_0 = \frac{1}{\sqrt{\langle 1, 1\rangle}} = \frac{1}{\sqrt{\int_I w(x)\,\mathrm{d}x}}.$$

Beweis. Wir führen den Beweis induktiv. Sei $\{u_0, u_1, \ldots, u_n\}$ eine Orthonormalbasis von Π_n bezüglich $\langle\cdot,\cdot\rangle$ aus (6.7), die alle im Satz genannten Eigenschaften erfüllt. Ist

$$p_{n+1}(x) = xu_n(x) - b_n u_n(x) - a_n u_{n-1}(x),$$

dann gilt

$$\langle p_{n+1}, u_n\rangle = \langle xu_n, u_n\rangle - b_n \underbrace{\langle u_n, u_n\rangle}_{=1} - a_n \underbrace{\langle u_{n-1}, u_n\rangle}_{=0} = \underbrace{\langle xu_n, u_n\rangle}_{=b_n} - b_n = 0.$$

Offensichtlich gilt zudem

$$\begin{aligned}
\langle p_{n+1}, u_{n-1}\rangle &= \langle xu_n, u_{n-1}\rangle - b_n \underbrace{\langle u_n, u_{n-1}\rangle}_{=0} - a_n \underbrace{\langle u_{n-1}, u_{n-1}\rangle}_{=1} \\
&= \langle xu_n, u_{n-1}\rangle - a_n \\
&\overset{(6.7)}{=} \langle u_n, xu_{n-1}\rangle - a_n \\
&\overset{(6.9)}{=} \langle u_n, a_n u_n + b_{n-1}u_{n-1} + a_{n-1}u_{n-2}\rangle - a_n \\
&= a_n \underbrace{\langle u_n, u_n\rangle}_{=1} + b_{n-1} \underbrace{\langle u_n, u_{n-1}\rangle}_{=0} + a_{n-1} \underbrace{\langle u_n, u_{n-2}\rangle}_{=0} - a_n \\
&= 0
\end{aligned}$$

und für jedes $m < n-1$

$$\begin{aligned}\langle p_{n+1}, u_m\rangle &= \langle xu_n, u_m\rangle - b_n \underbrace{\langle u_n, u_m\rangle}_{=0} - a_n \underbrace{\langle u_{n-1}, u_m\rangle}_{=0} \\ &= \langle u_n, xu_m\rangle \\ &= \langle u_n, a_{m+1}u_{m+1} + b_m u_m + a_m u_{m-1}\rangle \\ &= 0.\end{aligned}$$

Folglich ist p_{n+1} für alle $0 \le m \le n$ orthogonal zu u_m, und nach Konstruktion gilt

$$p_{n+1}(x) = \gamma_n x^{n+1} + \cdots .$$

Um das normalisierte Polynom u_{n+1} zu finden, muss p_{n+1} geeignet umskaliert werden. Gesucht ist u_{n+1} derart, dass der Höchstkoeffizent γ_{n+1} positiv ist. Also gilt

$$u_{n+1} = \gamma_{n+1}x^{n+1} + \cdots = \frac{\gamma_{n+1}}{\gamma_n} p_{n+1},$$

beziehungsweise $p_{n+1} = a_{n+1}u_{n+1}$.
Ist q_{n+1} ein zweites Polynom, das die Bedingung (6.8) erfüllt, dann gilt nach Satz 6.7

$$q_{n+1} = \sum_{m=0}^{n+1} \langle q_{n+1}, u_m\rangle u_m = \underbrace{\langle q_{n+1}, u_{n+1}\rangle}_{=:c} u_{n+1},$$

das heißt, q_{n+1} ist ein skalares Vielfaches von u_{n+1}. Wegen

$$\langle q_{n+1}, q_{n+1}\rangle = c^2 \langle u_{n+1}, u_{n+1}\rangle$$

ergibt sich $c^2 = 1$, also $c = \pm 1$. Da $c = -1$ auf einen negativen Höchstkoeffizienten führt, folgt schließlich $c = 1$ und $q_{n+1} = u_{n+1}$. □

Beispiele 6.16 Die gängisten Orthogonalpolynome sind:

- Tschebyscheff-Polynome: $w(x) = 1/\sqrt{1-x^2}$ und $I = (-1, 1)$.
- Legendre-Polynome: $w(x) = 1$ und $I = (-1, 1)$
- Jacobi-Polynome: $w(x) = (1-x)^\alpha (1+x)^\beta$ für $\alpha, \beta > -1$ und $I = (-1, 1)$
- Hermite-Polynome: $w(x) = e^{-x^2}$ und $I = \mathbb{R}$
- Laguerre-Polynome: $w(x) = x^\alpha e^{-x}$ für $\alpha > -1$ und $I = \mathbb{R}_{>0}$

△

Der Fall $w(x) = 1$ ist dabei für Anwendungen besonders interessant:

Beispiel 6.17 (Legendre Polynome) Wir betrachten das Intervall $I = [-1, 1]$, ausgestattet mit dem Standard-Innenprodukt, das heißt, $w(x) \equiv 1$. Das *Legendre-Polynom* vom Grad n ist definiert durch

$$P_n(x) := \frac{1}{2^n n!} \frac{d^n}{dx^n}[(x^2-1)^n] \in \Pi_n. \tag{6.10}$$

Es gilt $\langle P_n, P_m\rangle = 0$ für $n \neq m$, denn für $m < n$ folgt durch partielle Integration

$$\begin{aligned}
2^n n!\, 2^m m!\, \langle P_n, P_m\rangle &= \int_{-1}^{1} \frac{d^n}{dx^n}[(x^2-1)^n]\frac{d^m}{dx^m}[(x^2-1)^m]\,\mathrm{d}x \\
&= \frac{d^{n-1}}{dx^{n-1}}[\underbrace{(x^2-1)^n}_{n\text{-fache Nst. in } \pm 1}]\frac{d^m}{dx^m}[(x^2-1)^m]\Big|_{-1}^{1} \\
&\qquad - \int_{-1}^{1} \frac{d^{n-1}}{dx^{n-1}}[(x^2-1)^n]\frac{d^{m+1}}{dx^{m+1}}[(x^2-1)^m]\,\mathrm{d}x \\
&= -\int_{-1}^{1} \frac{d^{n-1}}{dx^{n-1}}[(x^2-1)^n]\frac{d^{m+1}}{dx^{m+1}}[(x^2-1)^m]\,\mathrm{d}x \\
&\;\;\vdots \\
&= (-1)^n \int_{-1}^{1} (x^2-1)^n \underbrace{\frac{d^{m+n}}{dx^{m+n}}[(x^2-1)^m]}_{\equiv 0,\ \text{da } m+n > 2m = \deg[(x^2-1)^m]}\,\mathrm{d}x = 0.
\end{aligned}$$

Wiederum mit Hilfe partieller Integration folgt

$$\begin{aligned}
2^n n!\, 2^n n!\, \langle P_n, P_n\rangle &= (-1)^n \int_{-1}^{1} \underbrace{\frac{d^{2n}}{dx^{2n}}[(x^2-1)^n]}_{=(2n)!}(x^2-1)^n\,\mathrm{d}x \\
&= (-1)^n (2n!) \int_{-1}^{1} (x-1)^n (x+1)^n\,\mathrm{d}x \\
&= (n!)^2 \frac{2^{2n+1}}{2n+1}.
\end{aligned}$$

Also ergibt sich $\langle P_n, P_n\rangle = 2/(2n+1)$ und daher ist

$$u_n = \sqrt{\frac{2n+1}{2}} P_n = \sqrt{\frac{2n+1}{2}} \frac{(2n)!}{2^n (n!)^2}\, x^n + \cdots = \gamma_n x^n + \cdots \tag{6.11}$$

das zugehörige orthonormale Polynom.
Für die Dreitermrekursion (6.9) ergibt sich

$$\begin{aligned}
a_n = \frac{\gamma_{n-1}}{\gamma_n} &= \frac{2^n (n!)^2}{(2n)!} \sqrt{\frac{2}{2n+1}} \sqrt{\frac{2n-1}{2}} \frac{(2n-2)!}{2^{n-1}\big((n-1)!\big)^2} \\
&= \sqrt{\frac{2n-1}{2n+1}} \frac{2n^2}{2n(2n-1)} \\
&= \frac{n}{\sqrt{4n^2-1}}.
\end{aligned}$$

Wegen (6.10) ist P_n ein gerades Polynom, falls n gerade, und ein ungerades Polynom, falls n ungerade ist. Daher muss b_n in (6.9) verschwinden. Dies bedeutet,

$$\frac{n+1}{\sqrt{4(n+1)^2-1}} u_{n+1}(x) = x u_n(x) - \frac{n}{\sqrt{4n^2-1}} u_{n-1}(x).$$

△

Bemerkung: Satz 6.15 (inklusive Beweis), sowie alle nachfolgenden Resultate, gelten auch für diskrete Innenprodukte

$$\langle f, g\rangle = \sum_{i=1}^{N} w_i f(\lambda_i) g(\lambda_i), \quad w_i > 0, \quad N \leq \infty, \tag{6.12}$$

vorausgesetzt im Falle $N < \infty$ gilt $n < N$. In Π_N ist nämlich $\langle \cdot, \cdot \rangle$ kein Innenprodukt mehr, da $\langle p, p\rangle = 0$ für $p = \prod_{i=1}^{N}(x - \lambda_i)$. Aus dem Beweis von Satz 6.15 erkennt man zudem, dass für $n = N-1$ die Dreitermrekursion (6.9) ein Vielfaches u_N von eben diesem p liefert, da $\langle u_N, q\rangle = 0$ ist für alle $q \in \Pi_{N-1}$. Lediglich a_N ist nicht festgelegt.

Folgende Optimalitätseigenschaft ähnelt in gewisser Weise der ersten Aussage aus Satz 6.13.

Satz 6.18 Sei $u_n = \gamma_n x^n + \cdots$ das Orthonormalpolynom aus (6.8) bezüglich $\langle \cdot, \cdot \rangle$ aus (6.7). Dann minimiert $p_n = u_n/\gamma_n$ unter allen Polynomen $p_n(x) = x^n + \ldots$ die gewichtete L^2-Norm

$$\|p_n\|^2_{L^2(I;w)} = \int_I p_n^2(x) w(x)\,\mathrm{d}x.$$

Beweis. Die gestellte Minimierungsaufgabe $\|p_n\|_{L^2(I;w)} \to \min$ lässt sich auch wie folgt formulieren: schreibe $p_n = x^n - q_{n-1} \in \Pi_n$ und bestimme $q_{n-1} \in \Pi_{n-1}$ derart, dass

$$\|x^n - q_{n-1}\|_{L^2(I;w)} \to \min_{q_{n-1} \in \Pi_{n-1}}.$$

Die dritte Aussage aus Satz 6.7 liefert nun

$$q_{n-1} = \sum_{i=0}^{n-1} \langle x^n, u_i\rangle u_i,$$

dies bedeutet,

$$\langle x^n - q_{n-1}, u_i\rangle = \langle x^n, u_i\rangle - \langle q_{n-1}, u_i\rangle = 0, \quad 0 \leq i \leq n-1.$$

Also stimmen $p_n = x^n - q_{n-1}$ und u_n bis auf einen konstanten Faktor überein. Dieser Faktor ist γ_n, wie sich durch Vergleich der Höchstkoeffizienten ergibt. □

Korollar 6.19 Die Tschebyscheff-Polynome $p_n = 2^{1-n}T_n$ erfüllen neben der Optimalitätseigenschaft aus Satz 6.13 auch

$$\int_{-1}^{1} \frac{p_n^2(x)}{\sqrt{1-x^2}}\,\mathrm{d}x \to \min_{p_n \in \Pi_n \text{ monisch}}.$$

Beispiel 6.20 Gemäß der ersten Aussage aus Satz 6.13 gilt hinsichtlich der Maximumsnorm

$$\min_{p_n \in \Pi_n \text{ monisch}} \|p_n\|_{C([-1,1])} = \|2^{1-n}T_n\|_{C([-1,1])} = 2^{1-n}.$$

Bezüglich der L^2-Norm ergibt sich nach Satz 6.18

$$\min_{p_n\in\Pi_n \text{ monisch}} \|p_n\|_{L^2([-1,1])} = \left\|\frac{u_n}{\gamma_n}\right\|_{L^2([-1,1])} = \frac{1}{\gamma_n} = \sqrt{\frac{2}{2n+1}}\frac{2^n(n!)^2}{(2n)!},$$

denn u_n und γ_n sind die entsprechenden Größen der Legendre-Polynome aus Beispiel 6.17. Mit Hilfe der *Stirlingschen Formel*

$$n! = \sqrt{2\pi n}\left(\frac{n}{e}\right)^n \underbrace{e^{\lambda(n)}}_{\approx 1} \quad \text{mit} \quad \frac{1}{12n+1} < \lambda(n) < \frac{1}{12n}$$

ergibt sich

$$\min_{p_n\in\Pi_n \text{ monisch}} \|p_n\|_{L^2([-1,1])} \approx \frac{\sqrt{\pi}}{2}2^{1-n} \approx 0.8862 \cdot 2^{1-n}.$$

△

Definition 6.21 1. Sei $\xi \in \mathbb{R}$ beliebig aber fest. Die Polynome

$$K_n(x,\xi) = \sum_{m=0}^{n} u_m(\xi)u_m(x) \in \Pi_n$$

heißen **Kernpolynome** bezüglich (6.7).

2. Die rationalen Funktionen

$$\Lambda_n(x) = \frac{1}{K_n(x,x)} = \frac{1}{\sum_{m=0}^{n} u_m^2(x)}$$

heißen **Christoffel-Funktionen** bezüglich (6.7).

Ist $\xi \notin I$, dann bilden die Kernpolynome selbst ein orthogonales System, und zwar bezüglich der Gewichtsfunktion $\widetilde{w}(x) = |x-\xi|w(x)$, $x \in I$. Dies beweisen wir allerdings erst in Satz 6.26.

Zum Schluss noch ein für die Numerik sehr bedeutsames Resultat über Orthogonalpolynome (vergleiche Aussage 2 aus Satz 6.13):

Satz 6.22 Sei $\xi \in \mathbb{R}$. Unter allen Polynomen p_n vom Grad n mit $p_n(\xi) = 1$ minimiert $p_n(x) = K_n(x,\xi)/K_n(\xi,\xi)$ die gewichtete L^2-Norm $\|p_n\|_{L^2(I;w)}$, und das Minimum ist gerade $\sqrt{\Lambda_n(\xi)}$.

Beweis. Sei $p_n \in \Pi_n$ mit $p_n(\xi) = 1$. Wir schreiben

$$p_n(x) = \sum_{m=0}^{n} \langle p_n, u_m\rangle u_m(x), \quad x \in \mathbb{R},$$

und erhalten aus der Cauchy-Schwarzschen Ungleichung im $\mathbb{R}^{n+1}$

$$1 = p_n^2(\xi) \le \left(\sum_{m=0}^{n}\langle p_n, u_m\rangle^2\right)\underbrace{\left(\sum_{m=0}^{n} u_m^2(\xi)\right)}_{=1/\Lambda_n(\xi)} \underset{\text{Satz 6.7}}{=} \frac{\|p_n\|^2_{L^2(I;w)}}{\Lambda_n(\xi)}.$$

Dabei gilt Gleichheit genau dann, wenn die Vektoren $[\langle p_n, u_0\rangle, \langle p_n, u_1\rangle, \ldots, \langle p_n, u_n\rangle]$ und $[u_0(\xi), u_1(\xi), \ldots, u_n(\xi)]$ linear abhängig sind, also wenn

$$p_n(x) = c\sum_{m=0}^{n} u_m(\xi)u_m(x) = cK_n(x,\xi), \quad c = \frac{1}{K_n(\xi,\xi)}.$$

Damit ist gezeigt, dass

$$\|p_n\|^2_{L^2(I;w)} \geq \Lambda_n(\xi),$$

und das Minimum wird angenommen für $p_n = K_n(\cdot,\xi)/K_n(\xi,\xi)$. □

6.4 Nullstellen von Orthogonalpolynomen

Die Nullstellen von Orthogonalpolynomen $\{u_n\}$ bezüglich einer Gewichtsfunktion w über I haben einige wichtige Eigenschaften. Entspechende Eigenschaften gelten auch für diskrete Innenprodukte der Form (6.12), indem man

$$I := \operatorname{conv}\{\lambda_i, 1 \leq i \leq N\}$$

setzt. Dabei bezeichnet $\operatorname{conv}(M)$ die konvexe Hülle eine Menge $M \subset \mathbb{R}$.

Satz 6.23 Die Nullstellen der Orthogonalpolynome $\{u_n\}$ sind alle einfach und liegen im Inneren von I.

Beweis. Wir nehmen an, die Aussage sei falsch. Hat u_n etwa eine Nullstelle z auf dem Rand von I oder in $\mathbb{R} \setminus I$, dann ist

$$p_{n-1}(x) := \frac{u_n(x)}{x-z} \in \Pi_{n-1}$$

und daher

$$0 = \langle p_{n-1}, u_n\rangle = \int_I \frac{u_n^2(x)}{x-z} w(x)\,\mathrm{d}x.$$

Da jedoch $w(x)/(x-z)$ in I keinen Vorzeichenwechsel hat und $0 \not\equiv u_n^2(x) \geq 0$ ist, ergibt sich ein Widerspruch.
Ist hingegen $z \in I$ eine mehrfache Nullstelle von u_n oder liegt $z \notin \mathbb{R}$, dann wenden wir das entsprechende Argument auf

$$p_{n-2}(x) := \frac{u_n(x)}{(x-z)(x-\overline{z})} = \frac{u_n(x)}{|x-z|^2} \in \Pi_{n-2}$$

an. Dazu beachte man, dass gemäß Satz 6.15 das Polynom u_n reell, also mit $z \notin \mathbb{R}$ auch $\overline{z}$ eine Nullstelle von u_n ist. □

Bevor wir weitere Eigenschaften der Nullstellen von $\{u_n\}$ herleiten, wenden wir uns noch einmal den Kernpolynomen zu.

Lemma 6.24 (Christoffel-Darboux-Identität) Für alle $x, \xi \in \mathbb{R}$ mit $x \neq \xi$ gilt

$$K_n(x,\xi) = a_{n+1}\frac{u_{n+1}(x)u_n(\xi) - u_n(x)u_{n+1}(\xi)}{x-\xi}, \tag{6.13}$$

wobei $a_{n+1} = \gamma_n/\gamma_{n+1}$ aus der entsprechenden Dreitermrekursion (6.9) herrührt.

Beweis. Durch Anwendung der Dreitermrekursion (6.9) auf $u_{n+1}(x)$ und $u_{n+1}(\xi)$ ergibt sich

$$\begin{aligned}
&a_{n+1}\Big(u_{n+1}(x)u_n(\xi) - u_n(x)u_{n+1}(\xi)\Big)\\
&\quad= u_n(\xi)\Big(xu_n(x) - b_nu_n(x) - a_nu_{n-1}(x)\Big) - u_n(x)\Big(\xi u_n(\xi) - b_nu_n(\xi) - a_nu_{n-1}(\xi)\Big)\\
&\quad= (x-\xi)u_n(x)u_n(\xi) + a_n\Big(u_n(x)u_{n-1}(\xi) - u_n(\xi)u_{n-1}(x)\Big)\\
&\quad\vdots\\
&\quad= \sum_{i=0}^{n}(x-\xi)u_i(x)u_i(\xi) + a_0\Big(u_0(x)\underbrace{u_{-1}(\xi)}_{=0} - u_0(\xi)\underbrace{u_{-1}(x)}_{=0}\Big)\\
&\quad= (x-\xi)K_n(x,\xi).
\end{aligned}$$

□

Korollar 6.25 Es gilt für alle $x \in \mathbb{R}$

$$K_n(x,x) = a_{n+1}\Big(u'_{n+1}(x)u_n(x) - u_{n+1}(x)u'_n(x)\Big). \tag{6.14}$$

Beweis. Ersetze in (6.13) den Zähler durch

$$u_{n+1}(x)u_n(\xi) - u_n(x)u_{n+1}(\xi) = \Big(u_{n+1}(x) - u_{n+1}(\xi)\Big)u_n(\xi) - u_{n+1}(\xi)\Big(u_n(x) - u_n(\xi)\Big)$$

und lasse $\xi \to x$ gehen. □

Damit können wir die bereits in Abschnitt 6.3 angekündigte Orthogonalität der Kernpolynome beweisen:

Satz 6.26 Die Kernpolynome $K_n(x,\xi)$ sind orthogonal bezüglich $|x-\xi|w(x)$ falls $\xi \notin I$, das heißt,

$$\int_I K_i(x,\xi)K_j(x,\xi)|x-\xi|w(x)\,\mathrm{d}x = 0, \quad i \neq j.$$

Beweis. Ohne Einschränkung der Allgemeinheit sei $i > j$. Aus der Christoffel-Darboux-

Identität (6.13) folgt

$$\begin{aligned}
&\int_I K_i(x,\xi)K_j(x,\xi)|x-\xi|w(x)\,\mathrm{d}x \\
&\quad = \pm a_{i+1}\int_I \Big(u_{i+1}(x)u_i(\xi) - u_i(x)u_{i+1}(\xi)\Big)K_j(x,\xi)w(x)\,\mathrm{d}x \\
&\quad = \pm a_{i+1}\Big(u_i(\xi)\underbrace{\langle u_{i+1}, K_j(\cdot,\xi)\rangle}_{=0} - u_{i+1}(\xi)\underbrace{\langle u_i, K_j(\cdot,\xi)\rangle}_{=0}\Big) \\
&\quad = 0.
\end{aligned}$$

Dabei gilt "+", falls $\xi \geq x$ für $x \in I$, beziehungsweise "−", falls $\xi \leq x$ für $x \in I$. □

Nun zurück zu den Nullstellen von $\{u_n\}$. Es gilt folgende Trennungseigenschaft:

Satz 6.27 Zwischen je zwei Nullstellen von u_{n+1} befindet sich genau eine Nullstelle von u_n.

Beweis. Da $K_n(x,x) = \sum_{m=0}^n u_m^2(x) \geq u_0^2(x) > 0$ und $a_{n+1} > 0$ folgt aus (6.14), dass

$$u'_{n+1}(x_i)u_n(x_i) > 0$$

für jede Nullstelle x_i von u_{n+1}. Da u'_{n+1} das Vorzeichen zwischen zwei aufeinanderfolgenden Nullstellen x_i und x_{i+1} genau einmal wechselt, hat auch u_n mindestens einen Vorzeichenwechsel, also eine Nullstelle, im Interval (x_i, x_{i+1}). Auf diese Weise findet man n Nullstellen von u_n, so dass in jedem Intervall (x_i, x_{i+1}) auch nur eine Nullstelle von u_n liegen kann. □

Zum Abschluss beweisen wir noch ein Resultat für diskrete Innenprodukte der Form (6.12), vergleiche auch Bemerkung 6.3.

Satz 6.28 Seien $\{u_m\}_{m=0}^N$ die Orthonormalpolynome bezüglich des Innenprodukts (6.12). Für jedes $n < N$ besitzt u_n zwischen je zwei aufeinanderfolgenden Stützstellen λ_k und λ_ℓ höchstens eine Nullstelle.

Beweis. Ohne Einschränkung der Allgemeinheit sei $\lambda_k < \lambda_\ell$. Wir nehmen an, dass gilt $u_n(\xi_k) = u_n(\xi_\ell) = 0$ für zwei Punkte $\lambda_k < \xi_k < \xi_\ell < \lambda_\ell$. Dann ist

$$p_{n-2}(x) = \frac{u_n(x)}{(x-\xi_k)(x-\xi_\ell)} \in \Pi_{n-2}$$

und für alle $x \in \{\lambda_i : 1 \leq i \leq N\}$ gilt

$$\frac{u_n^2(x)}{(x-\xi_k)(x-\xi_\ell)} \geq 0.$$

Dabei gilt mindestens für ein $x \in \{\lambda_i : 1 \leq i \leq N\}$ die strenge Ungleichung, da u_n nur $n-2 < N$ weitere Nullstellen neben ξ_k und ξ_ℓ besitzt. Hieraus ergibt sich aber ein Widerspruch:

$$0 = \langle p_{n-2}, u_n\rangle = \sum_{i=1}^N w_i \frac{u_n^2(\lambda_i)}{(\lambda_i-\xi_k)(\lambda_i-\xi_\ell)} > 0.$$

□

Beachte: Nach Bemerkung 6.3 sind insbesondere die Nullstellen von u_N gerade die Stützstellen $\{\lambda_i\}$. Nach Satz 6.27 hat u_{N-1} demnach je eine Nullstelle zwischen zwei aufeinanderfolgenden Stützstellen.

6.5 Erneut das Lanczos-Verfahren

In Abschnitt 4.6 wurde das Lanczos-Verfahren zur Bestimmung von Eigenwerten einer hermiteschen Matrix $\mathbf{A} = \mathbf{A}^\star \in \mathbb{K}^{N\times N}$ eingeführt. Beim Lanczos-Prozess (Algorithmus 4.26) wird mittels der Iteration

$$\mathbf{u}_{i+1} = (\mathbf{A} - \alpha_i \mathbf{I})\mathbf{w}_i - \beta_{i-1}\mathbf{w}_{i-1}, \quad \mathbf{w}_{i+1} = \frac{\mathbf{u}_{i+1}}{\|\mathbf{u}_{i+1}\|_2}, \quad i = 1, 2, \ldots, n-1, \tag{6.15}$$

eine Orthonormalbasis $\{\mathbf{w}_1, \mathbf{w}_2, \ldots, \mathbf{w}_n\}$ des Krylov-Raums

$$\mathcal{K}_n(\mathbf{A}, \mathbf{z}) = \operatorname{span}\{\mathbf{z}, \mathbf{A}\mathbf{z}, \ldots, \mathbf{A}^{n-1}\mathbf{z}\}$$

erzeugt, wobei $\mathbf{w}_0 = \mathbf{0}$, $\mathbf{u}_1 = \mathbf{z}$, $\mathbf{w}_1 = \mathbf{z}/\|\mathbf{z}\|_2$, und

$$\alpha_i = \mathbf{w}_i^\star \mathbf{A}\mathbf{w}_i, \quad \beta_{i-1} = \|\mathbf{u}_i\|_2, \quad i = 1, 2, \ldots, n.$$

Jedes Element $\mathbf{x} \in \mathcal{K}_n(\mathbf{A}, \mathbf{z})$ lässt sich schreiben als

$$\mathbf{x} = \sum_{i=0}^{n-1} \zeta_i \mathbf{A}^i \mathbf{z} = \left(\sum_{i=0}^{n-1} \zeta_i \mathbf{A}^i \right) \mathbf{z},$$

dies bedeutet,

$$\mathbf{x} = p(\mathbf{A})\mathbf{z}, \quad \text{wobei} \quad p(\lambda) = \sum_{i=0}^{n-1} \zeta_i \lambda^i \in \Pi_{n-1}. \tag{6.16}$$

Durch (6.16) wird jedem Element aus $\mathcal{K}_n(\mathbf{A}, \mathbf{z})$ genau ein Polynom zugeordnet und umgekehrt, jedoch nur solange wie $\dim \mathcal{K}_n(\mathbf{A}, \mathbf{z}) = n$.
Wir definieren nun für Polynome $p, q \in \Pi_n$ die Bilinearform

$$\langle p, q \rangle := \big(p(\mathbf{A})\mathbf{z}\big)^\star \big(q(\mathbf{A})\mathbf{z}\big). \tag{6.17}$$

Offensichtlich ist (6.17) ein Innenprodukt, sofern für $p \not\equiv 0$ gilt

$$\|p\|^2 = \big(p(\mathbf{A})\mathbf{z}\big)^\star \big(p(\mathbf{A})\mathbf{z}\big) = \|p(\mathbf{A})\mathbf{z}\|_2^2 > 0. \tag{6.18}$$

Dies ist genau dann der Fall, wenn $\dim \mathcal{K}_n(\mathbf{A}, \mathbf{z}) = n$.
Seien $(\lambda_1, \mathbf{v}_1), (\lambda_2, \mathbf{v}_2), \ldots, (\lambda_N, \mathbf{v}_N)$ die Eigenpaare von $\mathbf{A}$, wobei wir der Einfachheit halber annehmen, dass alle Eigenwerte $\lambda_1, \lambda_2, \ldots, \lambda_N$ paarweise verschieden sind und dass $\|\mathbf{v}_i\|_2 = 1$. Enwickeln wir den Startvektor in die Eigenbasis

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

so folgt aus (6.17)

$$\begin{aligned}\langle p, q\rangle &= \Big(p(\mathbf{A})\sum_{i=1}^{N}\xi_i\mathbf{v}_i\Big)^\star\Big(q(\mathbf{A})\sum_{j=1}^{N}\xi_j\mathbf{v}_j\Big)\\ &= \Big(\sum_{i=1}^{N}p(\lambda_i)\xi_i\mathbf{v}_i\Big)^\star\Big(\sum_{j=1}^{N}q(\lambda_j)\xi_j\mathbf{v}_j\Big)\\ &= \sum_{i,j=1}^{N}p(\lambda_i)q(\lambda_j)\xi_i\xi_j\underbrace{\mathbf{v}_i^\star\mathbf{v}_j}_{=\delta_{i,j}}\\ &= \sum_{i=1}^{N}p(\lambda_i)q(\lambda_i)|\xi_i|^2.\end{aligned}$$

Dies bedeutet, das Innenprodukt (6.17) ist ein diskretes Innenprodukt der Form (6.12). Dabei sind die Abszissen λ_i gerade die Eigenwerte von $\mathbf{A}$ und die Gewichte $w_i = |\xi_i|^2$ die Eigenanteile im Startvektor. Insbesondere folgt, dass (6.18) positiv ist für $p \not\equiv 0$, solange $n < N$ gilt und $\xi_i \neq 0$ für alle i.

Die Vektoren $\mathbf{w}_i$ aus (6.15) können gemäß (6.16) mit Polynomen $u_{i-1} \in \Pi_{i-1}$ identifiziert werden, wobei die Orthonormalität der $\mathbf{w}_i$ impliziert, dass

$$\delta_{i,j} = \mathbf{w}_i^\star\mathbf{w}_j = \big(u_{i-1}(\mathbf{A})\mathbf{z}\big)^\star\big(u_{j-1}(\mathbf{A})\mathbf{z}\big) = \langle u_{i-1}, u_{j-1}\rangle.$$

Folglich sind die Polynome $\{u_i\}_{i=0}^{n-1}$ gerade die Orthonormalpolynome bezüglich des Innenprodukts (6.17). Nach (6.15) genügen sie der Dreitermrekursion

$$\beta_i u_i(\mathbf{A})\mathbf{z} = (\mathbf{A} - \alpha_i\mathbf{I})u_{i-1}(\mathbf{A})\mathbf{z} - \beta_{i-1}u_{i-2}(\mathbf{A})\mathbf{z},$$

also

$$\beta_i u_i(x) = (x - \alpha_i)u_{i-1}(x) - \beta_{i-1}u_{i-2}(x). \tag{6.19}$$

Diese Rekursion stimmt bis auf Bezeichnungen mit (6.9) überein.

Die Näherungseigenwerte, die das Lanczos-Verfahren berechnet, sind die exakten Eigenwerte der *Jacobi-Matrix*

$$\mathbf{T}_n = \begin{bmatrix}\alpha_1 & \beta_1 & & 0\\ \beta_1 & \alpha_2 & \ddots & \\ & \ddots & \ddots & \beta_{n-1}\\ 0 & & \beta_{n-1} & \alpha_n\end{bmatrix}.$$

Ist λ ein Eigenwert von $\mathbf{T}_n$ und $[t_0, t_1, \ldots, t_{n-1}]^T$ ein zugehöriger Eigenvektor, dann folgt

$$\begin{aligned}\lambda t_0 &= \alpha_1 t_0 + \beta_1 t_1,\\ \lambda t_i &= \beta_i t_{i-1} + \alpha_{i+1}t_i + \beta_{i+1}t_{i+1}, \qquad i = 1, 2, \ldots, n-2,\end{aligned}$$

und

$$\lambda t_{n-1} = \beta_{n-1}t_{n-2} + \alpha_n t_{n-1}. \tag{6.20}$$

Setzen wir $t_{-1} = \beta_0 := 0$, so bedeutet dies gerade

$$\beta_i t_i = (\lambda - \alpha_i)t_{i-1} - \beta_{i-1}t_{i-2}, \quad i = 1, 2, \ldots, n-1. \tag{6.21}$$

Offensichtlich muss $t_0 \neq 0$ sein, sonst sind zwangsläufig alle $t_i = 0$. Wählen wir $t_0 = u_0$, dann liefert ein Vergleich zwischen (6.19) und (6.21) dass

$$t_i = u_i(\lambda), \quad i = 1, 2, \ldots, n-1.$$

Daher erhalten wir

$$\beta_n u_n(\lambda) \overset{(6.19)}{=} (\lambda - \alpha_n) u_{n-1}(\lambda) - \beta_{n-1} u_{n-2}(\lambda) \overset{(6.20)}{=} 0.$$

Folglich sind die Eigenwerte von $\mathbf{T}_n$ gerade die Nullstellen des Orthogonalpolynoms u_n.
Nach Abschnitt 6.4 liegen alle Nullstellen von u_n, das sind die approximierten Eigenwerte von $\mathbf{A}$, in der konvexen Hülle der λ_i, also des Spektrums von $\mathbf{A}$, und die Approximationen nach n und $n+1$ Schritten trennen sich (Satz 6.27). Darüberhinaus lässt sich aber noch mehr sagen:

- alle approximierenden Eigenwerte von $\mathbf{T}_n$ sind einfach (Satz 6.23) und
- zwischen je zwei Eigenwerten von $\mathbf{A}$ liegt höchstens eine Näherung (Satz 6.28).

6.6 Bisektionsverfahren

Sei

$$\mathbf{T} = \begin{bmatrix} \alpha_1 & \beta_1 & & 0 \\ \beta_1 & \alpha_2 & \ddots & \\ & \ddots & \ddots & \beta_{N-1} \\ 0 & & \beta_{N-1} & \alpha_N \end{bmatrix} \in \mathbb{R}^{N \times N} \tag{6.22}$$

eine symmetrische Tridiagonalmatrix mit $\beta_i \neq 0$ für alle $1 \leq i < N$. Im letzten Abschnitt haben wir gesehen, dass eine solche Jacobi-Matrix aus den Rekursionskoeffizienten orthogonaler Polynome entsteht. Wir betrachten nun das umgekehrte Problem, das heißt, wir suchen die Orthogonalpolynome, deren Jacobi-Matrix (6.22) ergibt.
Dazu verwenden wir die Spektralzerlegung $\mathbf{WDW}^\star$ von $\mathbf{T}$ mit $\mathbf{W}^\star\mathbf{W} = \mathbf{I}$ und definieren

$$\langle p, q \rangle := \sum_{i=1}^{N} w_{1,i}^2 p(\lambda_i) q(\lambda_i), \tag{6.23}$$

wobei $\mathbf{D} = \operatorname{diag}(\lambda_1, \ldots, \lambda_N)$ und $\mathbf{W} = [w_{i,j}]_{i,j=1}^N$. Wir werden gleich sehen, dass (6.23) ein Innenprodukt in Π_{N-1} definiert.

Satz 6.29 Sei $\mathbf{T}$ durch (6.22) gegeben und die Polynome u_i seien wie folgt definiert (vergleiche (6.19))

$$\beta_i u_i(x) = (x - \alpha_i) u_{i-1}(x) - \beta_{i-1} u_{i-2}(x), \quad i = 1, 2, \ldots, N, \tag{6.24}$$

mit $u_{-1} \equiv 0$, $u_0 \equiv 1$, und $\beta_N \neq 0$. Dann bilden die $\{u_i\}_{i=0}^N$ eine Orthonormalbasis bezüglich dem Innenprodukt (6.23), wobei die Abszissen $\{\lambda_i\}$ paarweise verschieden und $w_{1,i} \neq 0$ für alle $1 \leq i \leq N$ sind.

Beweis. Wie wir bereits im vorigen Abschnitt gesehen haben, muss $w_{1,i} \neq 0$ gelten. Wir zeigen nun, dass gilt

$$w_{i,j} = w_{1,j} u_{i-1}(\lambda_j), \quad i = 1, 2, \ldots, N. \tag{6.25}$$

Für $i = 1$ ist (6.25) offensichtlich richtig, da $u_0 \equiv 1$. Für $i = 0$ gilt dies ebenfalls, falls $w_{0,j} = 0$ gesetzt wird. Wegen $\mathbf{T} = \mathbf{W}\mathbf{D}\mathbf{W}^\star$ folgt

$$\mathbf{T} \begin{bmatrix} w_{1,j} \\ w_{2,j} \\ \vdots \\ w_{N,j} \end{bmatrix} = \lambda_j \begin{bmatrix} w_{1,j} \\ w_{2,j} \\ \vdots \\ w_{N,j} \end{bmatrix},$$

dies bedeutet,

$$\begin{aligned} \lambda_1 w_{1,j} &= \underbrace{\beta_0 w_{0,j}}_{=0} + \alpha_1 w_{1,j} + \beta_1 w_{2,j}, \\ \lambda_j w_{i,j} &= \beta_{i-1} w_{i-1,j} + \alpha_i w_{i,j} + \beta_i w_{i+1,j}, \quad 2 \leq i \leq N-1, \\ \lambda_j w_{N,j} &= \beta_{N-1} w_{N-1,j} + \alpha_N w_{N,j}. \end{aligned} \tag{6.26}$$

Durch einen Vergleich von (6.26) mit (6.24) folgt induktiv (6.25) für alle $i = 1, 2, \ldots, N$. Hieraus ergibt sich dann sofort die behauptete Orthogonalitätsbeziehung aus der Orthogonalität der Zeilen von $\mathbf{W}$:

$$\delta_{k,\ell} = [w_{k,1}, w_{k,2}, \ldots, w_{k,N}] \begin{bmatrix} w_{\ell,1} \\ w_{\ell,2} \\ \vdots \\ w_{\ell,n} \end{bmatrix} \overset{(6.25)}{=} \sum_{j=1}^{N} w_{1,j}^2 u_{k-1}(\lambda_j) u_{\ell-1}(\lambda_j) \overset{(6.23)}{=} \langle u_{k-1}, u_{\ell-1} \rangle.$$

Insbesondere ist (6.23) ein definites Innenprodukt in Π_{N-1}, das heißt, keine zwei Eigenwerte von $\mathbf{T}$ können gleich sein.
Es verbleibt zu zeigen, dass $u_N(\lambda_j) = 0$ gilt für alle $j = 1, 2, \ldots, N$. Dies folgt aber aus

$$\begin{aligned} \beta_N u_N(\lambda_j) &\overset{(6.24)}{=} (\lambda_j - \alpha_N) u_{N-1}(\lambda_j) - \beta_{N-1} u_{N-2}(\lambda_j) \\ &\overset{(6.25)}{=} \frac{1}{w_{1,j}} \big[(\lambda_j - \alpha_N) w_{N,j} - \beta_{N-1} w_{N-1,j} \big] \\ &\overset{(6.26)}{=} 0. \end{aligned}$$

□

Beachte: Da ein β_i negativ sein kann, ist im allgemeinen die Positivität aus (6.8) des Höchstkoeffizienten γ_n an nicht erfüllt.

Wir können nun unsere Kenntnisse über die Nullstellen orthogonaler Polynome ausnutzen, um die Eigenwerte von $\mathbf{T}$ einzuschließen. Ist nämlich λ eine beliebige reelle Zahl und hat u_i k Nullstellen, die größer als λ sind, dann ist

$$\operatorname{sign}\big(u_i(\lambda)\big) = (-1)^k \operatorname{sign}\Bigg(\underbrace{\frac{1}{\beta_1 \beta_2 \cdots \beta_i}}_{\text{Höchstkoeffizent } \gamma_i} \Bigg) = (-1)^k \operatorname{sign}(\beta_1 \beta_2 \cdots \beta_i).$$

Entsprechend ist

$$\operatorname{sign}\big(u_{i+1}(\lambda)\big) = (-1)^k \operatorname{sign}(\beta_1\beta_2\cdots\beta_{i+1}),$$

falls auch u_{i+1} nur k Nullstellen größer λ besitzt, oder

$$\operatorname{sign}\big(u_{i+1}(\lambda)\big) = (-1)^{k+1} \operatorname{sign}(\beta_1\beta_2\cdots\beta_{i+1}),$$

falls u_{i+1} nun $k+1$ Nullstellen größer λ besitzt. Andere Fälle sind nicht möglich. Im ersten Fall ist

$$\operatorname{sign}\big(\beta_1\beta_2\cdots\beta_{i+1}u_{i+1}(\lambda)\big) = \operatorname{sign}\big(\beta_1\beta_2\cdots\beta_i u_i(\lambda)\big),$$

im letzten Fall nicht!

Da $u_0 \equiv 1$ keine Nullstelle hat und nach Bemerkung 6.3 die Nullstellen von u_N gerade die Eigenwerte von T sind, erhalten wir:

Satz 6.30 (Sturm) Sei $\lambda \in \mathbb{R}$ beliebig. Die Anzahl der Vorzeichenwechsel in der Folge $\{\beta_1\beta_2\cdots\beta_i u_i(\lambda)\}_{i=0}^N$ entspricht genau der Anzahl der Eigenwerte von $\mathbf{T}$, die größer sind als λ.

Auf diesem Resultat beruht das nachfolgende Bisektionsverfahren zur Bestimmung des k-größten Eigenwerts λ_k von $\mathbf{T}$:

Algorithmus 6.31 (Bisektionsverfahren)

input: Matrix $\mathbf{T}$, Toleranz ε und Schätzwerte $\underline{x} < \lambda_k < \overline{x}$

output: Schätzwert x an λ_k mit $|x - \lambda_k| < \varepsilon$

① setze $x := (\underline{x} + \overline{x})/2$

② werte $\{u_i(x)\}_{i=0}^N$ mittels Dreitermrekursion (6.24) aus

③ bestimme die Anzahl j der Vorzeichenwechsel in der Folge $\{\beta_1\beta_2\cdots\beta_i u_i(x)\}_{i=0}^N$

④ falls $j \geq k$ setze $\underline{x} := x$, ansonsten setze $\overline{x} := x$

⑤ falls $\overline{x} - \underline{x} \geq \varepsilon$ gehe nach ①

Die Konvergenzrate ist linear, da die Näherung $x^{(j)}$ im j-ten Schritt

$$\big|x^{(j)} - \lambda_k\big| \leq \frac{1}{2}\big|x^{(j-1)} - \lambda_k\big|,$$

erfüllt. Dabei kann das Bisektionsverfahren so implementiert werden, dass jeder Iterationsschritt nur $2N$ Multiplikationen kostet.

7. Numerische Quadratur

7.1 Einführung

Wir wollen bestimmte Integrale

$$I[f] = \int_a^b f(x)\,\mathrm{d}x$$

numerisch approximieren, die nicht in geschlossener Form durch Angabe einer Stammfunktion integriert werden können.

Definition 7.1 Wir sprechen von einer **Quadraturformel**

$$Q[f] = \sum_{i=1}^{n} w_i f(x_i) \tag{7.1}$$

mit **Quadraturgewichten** (oder einfach **Gewichte**) w_i und **Quadraturpunkte** oder **-knoten** x_i, wenn die Wahl von n, $\{x_i\}$ und $\{w_i\}$ fix ist. Sind die Integrationsgrenzen der Quadraturformel von Bedeutung, so schreiben wir Q_a^b.

Satz 7.2 Sei Q eine Quadraturformel von der Form (7.1). Dann ist $Q : C([a,b]) \to \mathbb{R}$ linear, das heißt

$$Q[\alpha f + \beta g] = \alpha Q[f] + \beta Q[g]$$

für alle $f, g \in C([a,b])$ und $\alpha, \beta \in \mathbb{R}$. Ferner gilt

$$\big|Q[f]\big| \le \left(\sum_{i=1}^{n} |w_i|\right) \|f\|_{C([a,b])}.$$

Beweis. Wir zeigen nur die letzte Eigenschaft:

$$\big|Q[f]\big| \le \sum_{i=1}^{n} |w_i||f(x_i)| \le \sum_{i=1}^{n} |w_i| \|f\|_{C([a,b])}.$$

□

Bemerkung: Die Zahl $\sum_{i=1}^{n} |w_i|$ misst die *Stabilität* der Quadraturformel, denn für Funktionen f, $\Delta f \in C([a,b])$ gilt

$$\big|Q[f+\Delta f] - Q[f]\big| = \big|Q[\Delta f]\big| \leq \sum_{i=1}^{n} |w_i| \|\Delta f\|_{C([a,b])},$$

das heißt, sie gibt an, wie Störungen verstärkt werden. Um Auslöschung zu vermeiden, sind insbesondere Quadraturformeln vorzuziehen, bei denen die Gewichte positiv sind.

Definition 7.3 Eine Quadraturformel $Q[f]$ besitzt den **Exaktheitsgrad** m, falls gilt

$$Q[p] = I[p] \quad \text{für alle } p \in \Pi_m.$$

Besitzt eine Quadraturformel den Exaktheitsgrad m, so kann man sofort eine Fehlerabschätzung angeben. Wir formulieren dieses Resultat im Fall positiver Gewichte.

Satz 7.4 Seien $I[f] = \int_a^b f(x)\,\mathrm{d}x$ und $Q[f] = \sum_{i=1}^{n} w_i f(x_i)$ eine zugehörige Quadraturformel mit positiven Gewichten und Exaktheitsgrad $m \in \mathbb{N}_0$. Dann ist

$$\sum_{i=1}^{n} w_i = b - a$$

und für jedes $f \in C([a,b])$ gilt

$$\left| \int_a^b f(x)\,\mathrm{d}x - Q[f] \right| \leq 2(b-a) \min_{p \in \Pi_m} \|f - p\|_{C([a,b])}.$$

Beweis. Wegen $m \geq 0$ ist die Quadraturformel exakt für konstante Funktionen. Damit gilt für die Gewichte w_i

$$\sum_{i=1}^{n} w_i = Q[1] = \int_a^b 1\,\mathrm{d}x = b - a.$$

Sei nun $f \in C([a,b])$ und $p \in \Pi_m$ beliebig. Wegen $Q[p] = I[p]$ folgt

$$\begin{aligned}
\left| \int_a^b f(x)\,\mathrm{d}x - Q[f] \right| &= \left| \int_a^b f(x) - p(x)\,\mathrm{d}x - Q[f-p] \right| \\
&\leq (b-a)\|f-p\|_{C([a,b])} + \sum_{i=1}^{n} w_i \underbrace{|f(x_i) - p(x_i)|}_{\leq \|f-p\|_{C([a,b])}} \\
&\leq 2(b-a)\|f-p\|_{C([a,b])}.
\end{aligned}$$

Da $p \in \Pi_m$ beliebig war, ergibt sich die Behauptung. □

Oftmals ist das folgende Resultat hilfreich:

Lemma 7.5 Sei $Q[f] = \sum_{i=1}^{n} w_i f(x_i)$ eine Quadraturformel für $\int_a^b f(x)\,\mathrm{d}x$ mit zu $(a+b)/2$ symmetrischen Knoten und Gewichten. Ist $Q[p] = I[p]$ für alle Polynome $p \in \Pi_{2q}$, dann hat $Q[\cdot]$ mindestens den Exaktheitsgrad $2q+1$.

Beweis. Betrachte die Basis

$$\left\{x^0, x^1, \ldots, x^{2q}, \left(x - \frac{a+b}{2}\right)^{2q+1}\right\}$$

von Π_{2q+1}. Nach Voraussetzung ist $Q[x^i] = I[x^i]$ für alle $i = 0, 1, \ldots, 2q$. Ferner ist

$$Q\big[\big(x - \tfrac{a+b}{2}\big)^{2q+1}\big] = 0,$$

da die Knoten und Gewichte symmetrisch zu $(a+b)/2$ liegen und $\big(x - (a+b)/2\big)^{2q+1}$ punktsymmetrisch zum Punkt $\big((a+b)/2, 0\big)$ ist. Andererseits ist auch

$$\begin{aligned} I\big[\big(x - \tfrac{a+b}{2}\big)^{2q+1}\big] &= \frac{1}{2q+2}\left(x - \frac{a+b}{2}\right)^{2q+2}\Bigg|_a^b \\ &= \frac{1}{2q+2}\left\{\left(\frac{b-a}{2}\right)^{2q+2} - \left(\frac{a-b}{2}\right)^{2q+2}\right\} \\ &= 0. \end{aligned}$$

Folglich ist $Q[p] = I[p]$ für alle $p \in \Pi_{2q+1}$. □

Mithilfe der Substitutionsregel für Integrale kann man Quadraturformeln für neue Intervalle aus Quadraturformeln für bekannte Intervalle erzeugen. Sei Q_a^b eine Quadraturformel auf dem Intervall $[a, b]$ gegeben durch $Q_a^b[f] = \sum_{i=1}^{n} w_i f(x_i)$. Dann kann eine Formel für das Intervall $[a', b']$ definiert werden durch

$$Q_{a'}^{b'}[f] := \sum_{i=1}^{n} w_i' f(x_i'), \qquad w_i' := \frac{b'-a'}{b-a} w_i, \qquad x_i' := a' + \frac{b'-a'}{b-a}(x_i - a).$$

Diese Wahl ist dadurch motiviert, dass man mit der linearen Substitution

$$\begin{aligned} L : [a, b] &\to [a', b'] \\ x &\mapsto L(x) := a' + \frac{b'-a'}{b-a}(x-a) \end{aligned}$$

schreiben kann:

$$\int_{a'}^{b'} f(x)\,\mathrm{d}x = \int_a^b f\big(L(x)\big)\,L'(x)\,\mathrm{d}x = \frac{b'-a'}{b-a}\int_a^b f\left(a' + \frac{b'-a'}{b-a}(x-a)\right)\mathrm{d}x.$$

Wendet man nun die bekannte Quadraturformel Q_a^b an, dann erhält man die obige Formel für $Q_{a'}^{b'}$.

Aufgrund der Tatsache, dass Quadraturformeln mithilfe linearer Substitution aus einer bekannten erzeugt werden können, werden Quadraturformeln normalerweise nur auf *Referenzintervallen* angegeben. Das Referenzintervall ist üblicherweise $I = [0, 1]$ oder $I = [-1, 1]$.

7.2 Lagrange-Interpolation

Gegeben seien $n+1$ paarweise verschiedene Knoten $x_0 < x_1 < \cdots < x_n$ sowie $n+1$ Werte $y_0, y_1, \ldots, y_n$, kurz $n+1$ *Stützstellen* $(x_0, y_0), (x_1, y_1), \ldots, (x_n, y_n)$. Bei der Lagrangeschen Interpolationsaufgabe sucht man ein Polynom $p \in \Pi_n$ mit

$$p(x_i) \stackrel{!}{=} y_i, \qquad i = 0, 1, \ldots, n. \tag{7.2}$$

Definition 7.6 Wir bezeichnen als **Knotenpolynom**

$$w(x) = \prod_{j=0}^{n} (x - x_j) \in \Pi_{n+1}$$

und als **Lagrange-Grundpolynome** die Polynome

$$L_i(x) = \frac{w(x)}{(x - x_i) w'(x_i)} = \prod_{j=0,\, j \neq i}^{n} \frac{x - x_j}{x_i - x_j} \in \Pi_n.$$

Beachte: Es gilt offensichtlich

$$L_i(x_j) = \delta_{i,j}. \tag{7.3}$$

Satz 7.7 Die Interpolationsaufgabe (7.2) hat genau eine Lösung

$$p = \sum_{i=0}^{n} y_i L_i.$$

Beweis. Wegen (7.3) gilt

$$p(x_j) = \sum_{i=0}^{n} y_i L_i(x_j) = y_j,$$

also (7.2). Damit ist die Existenz einer Lösung gesichert. Seien $p, q \in \Pi_n$ zwei Lösungen der Interpolationsaufgabe (7.2). Dann folgt aus

$$(p - q)(x_i) = 0, \qquad i = 0, 1, \ldots, n,$$

dass das Polynom $p - q \in \Pi_n$ $n + 1$ Nullstellen besitzt. Daraus folgt jedoch $p \equiv q$. □

Satz 7.8 Sei $f \in C^{n+1}([a, b])$ und $p \in \Pi_n$ das Interpolationspolynom zu den Knoten $\{x_i\}_{i=0}^n$ und Werten $\{y_i = f(x_i)\}_{i=0}^n$. Dann existiert zu jedem x ein $\xi \in I := \operatorname{conv}\{x, x_0, x_1, \ldots, x_n\}$ mit

$$f(x) - p(x) = \frac{f^{(n+1)}(\xi)}{(n+1)!} w(x). \tag{7.4}$$

Beweis. Gilt $x = x_i$ für ein $0 \le i \le n$, so ist (7.4) gültig mit $\xi = x$. Sei also $x \notin \{x_i\}_{i=0}^m$. Dann hat

$$h(t) := f(t) - p(t) - \frac{w(t)}{w(x)}\big(f(x) - p(x)\big)$$

Nullstellen in $x_0, x_1, \ldots, x_n$ sowie für $t = x$. Damit enthält jedes der $n+1$ Teilintervalle zwischen diesen Nullstellen nach dem Satz von Rolle eine Nullstelle $\tau_i^{(1)}$ von h', $i = 1, 2, \ldots, n+1$. So fortfahrend erhält man $n, n-1, \ldots$ Nullstellen $\tau_i^{(k)}$ von $h^{(k)}$, $k = 2, 3, \ldots$ im Intervall I. Insbesondere ergibt sich eine Nullstelle $\xi = \tau_1^{(n+1)}$ von $h^{(n+1)}$ in I. Wegen $p \in \Pi_n$ ist $p^{(n+1)} \equiv 0$, und wegen

$$w^{(n+1)}(t) = \left(\frac{d}{dt}\right)^{n+1} (t^{n+1} + \ldots) \equiv (n+1)!,$$

folgt hieraus

$$0 = h^{(n+1)}(\xi) = f^{(n+1)}(\xi) - \frac{(n+1)!}{w(x)}\big(f(x) - p(x)\big),$$

beziehungsweise

$$f(x) - p(x) = \frac{f^{(n+1)}(\xi)}{(n+1)!} w(x).$$

□

Korollar 7.9 Sei $f \in C^{n+1}([a,b])$, dann gilt

$$\inf_{p \in \Pi_n} \|f - p\|_{C([a,b])} \le \frac{(b-a)^{n+1}}{2^{2n+1}} \frac{\|f^{(n+1)}\|_{C([a,b])}}{(n+1)!}.$$

Beweis. Wir wählen die Nullstellen des transformierten Tschebyscheff-Polynoms $T_{n+1}\big((2x - b - a)/(b-a)\big)$ als Stützstellen $x_0 < x_1 < \cdots < x_n$ der Polynominterpolation auf dem Intervall $[a,b]$. Damit ist das Knotenpolynom gerade

$$w(x) = \prod_{j=0}^{n} (x - x_j) = \left(\frac{b-a}{2}\right)^{n+1} 2^{-n} T_{n+1}\left(\frac{2x - b - a}{b-a}\right),$$

dies bedeutet

$$\|w(x)\|_{C([a,b])} \le \frac{(b-a)^{n+1}}{2^{2n+1}} \underbrace{\left\| T_{n+1}\left(\frac{2x - b - a}{b-a}\right)\right\|_{C([a,b])}}_{=1}.$$

Somit genügt das Interpolationspolynom $p \in \Pi_n$ der Fehlerabschätzung

$$\|f - p\|_{C([a,b])} \le \|w(x)\|_{C([a,b])} \frac{\|f^{(n+1)}\|_{C([a,b])}}{(n+1)!} \le \frac{(b-a)^{n+1}}{2^{2n+1}} \frac{\|f^{(n+1)}\|_{C([a,b])}}{(n+1)!}.$$

□

7.3 Newton-Cotes-Formeln

Mit Hilfe der Polynominterpolation lassen sich leicht Quadraturformeln für $I[f]$ mit beliebigen Exaktheitsgrad m angeben. Seien $x_0 < x_1 < \cdots < x_m$ $m+1$ vorgegebene Knoten in $[a, b]$ und

$$w_i := \int_a^b L_i(x)\,\mathrm{d}x \tag{7.5}$$

das Integral des i-ten zugehörigen Lagrange-Grundpolynoms. Dann gilt:

Proposition 7.10 Die Quadraturformel Q mit Knoten $\{x_i\}$ und Gewichten $\{w_i\}$ gemäß (7.5) hat den Exaktheitsgrad m.

Beweis. Sei $p \in \Pi_m$. Offensichtlich interpoliert p sich selbst. Wegen der Eindeutigkeit des Interpolationspolynoms gilt daher

$$p(x) = \sum_{i=0}^{m} p(x_i) L_i(x)$$

(vgl. Satz 7.7). Daraus folgt

$$\begin{aligned} I[p] = \int_a^b p(x)\,\mathrm{d}x &= \int_a^b \sum_{i=0}^{m} p(x_i) L_i(x)\,\mathrm{d}x \\ &= \sum_{i=0}^{m} p(x_i) \int_a^b L_i(x)\,\mathrm{d}x \overset{(7.5)}{=} \sum_{i=0}^{m} w_i p(x_i) = Q[p], \end{aligned}$$

was zu zeigen war. □

Beschränken wir uns auf den Fall äquidistanter Knoten $a = x_0 < x_1 < \cdots < x_{m-1} < x_m = b$, dann erhalten wir die *Newton-Cotes-Formeln.*

Beispiel 7.11 Im Fall $m = 1$ erhält man die *Trapezregel*

$$\int_a^b f(x)\,\mathrm{d}x \approx T[f] := \frac{b-a}{2}\big(f(a) + f(b)\big).$$

Für $m = 2$ ergibt sich die *Simpson-Regel*

$$\int_a^b f(x)\,\mathrm{d}x \approx S[f] := \frac{b-a}{6}\left(f(a) + 4f\left(\frac{a+b}{2}\right) + f(b)\right).$$

△

Leider treten aber ab $m = 7$ negative Gewichte und dadurch unter Umständen Stabilitätsverluste selbst bei positiven Integranden auf. Die Gewichte der Newton-Cotes-Formeln für $m \leq 6$ sind in nachfolgender Tabelle zusammengefasst:

m	Gewichte	$\left\lvert \int_a^b f(x)\,\mathrm{d}x - Q_a^b[f] \right\rvert$	Name
1	$\frac{1}{2}\quad\frac{1}{2}$	$\frac{1}{12}(b-a)^3\|f''\|_{C([a,b])}$	Trapezregel
2	$\frac{1}{6}\quad\frac{4}{6}\quad\frac{1}{6}$	$\frac{1}{90}(\frac{b-a}{2})^5\|f^{(4)}\|_{C([a,b])}$	Simpson-Regel
3	$\frac{1}{8}\quad\frac{3}{8}\quad\frac{3}{8}\quad\frac{1}{8}$	$\frac{3}{80}(\frac{b-a}{3})^5\|f^{(4)}\|_{C([a,b])}$	3/8-Regel
4	$\frac{7}{90}\quad\frac{32}{90}\quad\frac{12}{90}\quad\frac{32}{90}\quad\frac{7}{90}$	$\frac{8}{945}(\frac{b-a}{4})^7\|f^{(6)}\|_{C([a,b])}$	Milne-Regel
5	$\frac{19}{288}\quad\frac{75}{288}\quad\frac{50}{288}\quad\frac{50}{288}\quad\frac{75}{288}\quad\frac{19}{288}$	$\frac{275}{12096}(\frac{b-a}{5})^7\|f^{(6)}\|_{C([a,b])}$	—
6	$\frac{41}{840}\quad\frac{216}{840}\quad\frac{27}{840}\quad\frac{272}{840}\quad\frac{27}{840}\quad\frac{216}{840}\quad\frac{41}{840}$	$\frac{9}{1400}(\frac{b-a}{6})^9\|f^{(8)}\|_{C([a,b])}$	Weddle-Regel

Die Newton-Cotes-Formeln sind für $m = 1, 3, 5$ exakt von der Ordnung m. Da die Formeln jedoch symmetrisch sind, sind sie für $m = 2, 4, 6$ gemäß Lemma 7.5 sogar exakt von der Ordnung $m + 1$. Die Simpson-Regel besitzt also Exaktheitsgrad 3 und nicht 2! Fehlerabschätzungen ergeben sich durch Kombination von Satz 7.4 und Korollar 7.9. Die in obiger Tabelle angegebenen Fehlerschranken ergeben sich aus einer verbesserten Analyse.

7.4 Zusammengesetzte Quadraturformeln

Um die Genauigkeit einer festen Quadraturformel zu erhöhen, zerlegt man das Intervall $[a, b]$ in N gleichgroße Teilintervalle $[a_i, a_{i+1}]$

$$a = a_0 < a_1 < a_2 < \cdots < a_N = b, \qquad a_i = a + i\frac{b-a}{N}.$$

Wegen

$$\int_a^b f(x)\,\mathrm{d}x = \sum_{i=0}^{N-1} \int_{a_i}^{a_{i+1}} f(x)\,\mathrm{d}x.$$

kann die Quadraturformel dann für jedes der Teilintervalle verwendet werden

$$Q_N[f] = \sum_{i=0}^{N-1} Q_{a_i}^{a_{i+1}}[f].$$

Definition 7.12 Unter der dazugehörigen **zusammengesetzten Quadraturformel** $Q_N[f]$ verstehen wir dann die Unterteilung von $[a, b]$ in N gleich große Teilintervalle, in denen jeweils die Quadraturformel angewandt wird.

Beispiel 7.13 Basierend auf den Newton-Cotes-Formeln erhält man die *zusammengesetzten Newton-Cotes-Formeln*. Im Fall $m = 1$ erhält man so die *zusammengesetzte Trapezregel*

$$\int_a^b f(x)\,\mathrm{d}x \approx T_N[f] := \frac{h}{2}\big\{f(a) + 2f(x_1) + 2f(x_2) + \cdots + 2f(x_{N-1}) + f(b)\big\},$$
$$h = \frac{b-a}{N}, \qquad x_i = a + ih, \qquad i = 1, 2, \ldots, N-1.$$

Im Fall $m = 2$ ergibt sich die *zusammengesetzte Simpson-Regel*

$$\int_a^b f(x)\,\mathrm{d}x \approx S_N[f] := \frac{h}{3}\big\{f(a) + 4f(x_1) + 2f(x_2) + 4f(x_3) + \cdots + 2f(x_{2N-2}) + 4f(x_{2N-1}) + f(b)\big\},$$
$$h = \frac{b-a}{2N}, \qquad x_i = a + ih, \qquad i = 1, 2, \ldots, 2N-1.$$

△

Im Zusammenhang mit zusammengesetzten Quadraturformeln steht die Asymptotik $N \to \infty$ im Vordergrund.

Definition 7.14 Eine zusammengesetzte Quadraturformel konvergiert gegen $I[f]$ mit der **Ordnung** s, falls

$$\big|Q_N[f] - I[f]\big| = \mathcal{O}(N^{-s}), \quad N \to \infty.$$

Satz 7.15 Besitzt die Quadraturformel Q den Exaktheitsgrad m, dann konvergiert die entsprechende zusammengesetzte Formel Q_N mit der Ordnung $m + 1$.

Beweis. Wir beweisen die Aussage auf dem Referenzintervall $[0, 1]$. Nach Satz 7.4 und Korollar 7.9 genügt Q_0^1 der Fehlerabschätzung

$$\left|\int_0^1 f(x)\,\mathrm{d}x - Q_0^1[f]\right| \le C\big\|f^{(m+1)}\big\|_{C([0,1])}.$$

Folglich erfüllt für alle c und $h > 0$ die transformierte Formel Q_c^{c+h}

$$\left|\int_c^{c+h} f(x)\,\mathrm{d}x - Q_c^{c+h}[f]\right| = h\left|\int_0^1 f(hx + c)\,\mathrm{d}x - Q_0^1[f(h\,\cdot\, + c)]\right| \le Ch^{m+2}\big\|f^{(m+1)}\big\|_{C([c,c+h])}.$$

Mit $h = 1/N$ gilt daher

$$\begin{aligned}\left|\int_0^1 f(x)\,\mathrm{d}x - \sum_{i=1}^N Q_{h(i-1)}^{hi}[f]\right| &\le \sum_{i=1}^N \left|\int_{h(i-1)}^{hi} f(x)\,\mathrm{d}x - Q_{h(i-1)}^{hi}[f]\right| \\ &\le \sum_{i=1}^N Ch^{m+2}\big\|f^{(m+1)}\big\|_{C([h(i-1),hi])} \\ &\le CNh^{m+2}\big\|f^{(m+1)}\big\|_{C([0,1])} \\ &= Ch^{m+1}\big\|f^{(m+1)}\big\|_{C([0,1])}.\end{aligned}$$

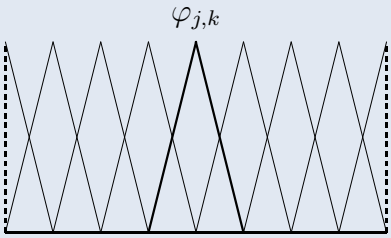
□

7.5 Adaptive Quadratur

Wir wollen adaptive Quadraturstrategien anhand der Trapezsumme vorstellen. Dazu führen wir lineare Splinefunktionen ein.

Definition 7.16 Sei

$$\varphi(x) = \begin{cases} 1 - |x|, & -1 \le x \le 1 \\ 0, & \text{sonst} \end{cases}$$



die **Hutfunktion**. Für jedes $j \in \mathbb{N}_0$ bilden die Funktionen

$$\varphi_{j,k}(x) = \varphi(2^j x - k)\big|_{[0,1]}, \quad k \in \Delta_j := \{0, 1, \ldots, 2^j\},$$

die **nodale Basis** im **Raum der stückweise linearen Funktionen**

$$V_j = \Big\{ f \in C([0,1]) : f\big|_{[2^{-j}k, 2^{-j}(k+1)]} \in \Pi_1 \text{ für alle } k = 0, 1, \ldots, 2^j - 1 \Big\}.$$

Eine Funktion $f \in C([0,1])$ kann mit Hilfe der nodalen Basis in V_j approximiert werden gemäß

$$f(x) \approx f_j(x) := \sum_{k \in \Delta_j} \alpha_{j,k} \varphi_{j,k}(x) \quad \text{mit} \quad \alpha_{j,k} = f(2^{-j}k).$$

Folglich gilt

$$I[f] = \int_0^1 f(x)\,\mathrm{d}x \approx I[f_j] = \sum_{k \in \Delta_j} \alpha_{j,k} \underbrace{\int_0^1 \varphi_{j,k}(x)\,\mathrm{d}x}_{= \begin{cases} 2^{-j}, & \text{falls } 0<k<2^j \\ 2^{-(j+1)}, & \text{sonst} \end{cases}}$$

Offensichtlich ist $I[f_j]$ gerade die Trapezsumme

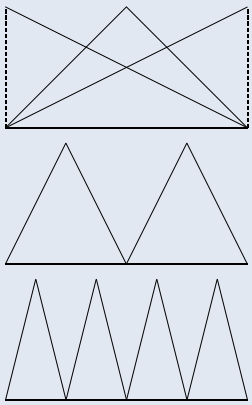
$$T_j[f] = 2^{-(j+1)} \Big\{ f(0) + 2 \sum_{k=1}^{2^j - 1} f(2^{-j}k) + f(1) \Big\}.$$

Definition 7.17 Setzen wir

$$\nabla_\ell := \begin{cases} \{0, 1\}, & \text{falls } \ell = 0, \\ \{1, 3, 5, \ldots, 2^\ell - 1\}, & \text{falls } \ell \in \mathbb{N}. \end{cases}$$

Dann ist die **hierarchische Basis** in V_j definiert durch

$$\{\varphi_{\ell,k}\}_{k \in \nabla_\ell, 0 \le \ell \le j}.$$



In der hierarchischen Basis besitzt f_j die Darstellung

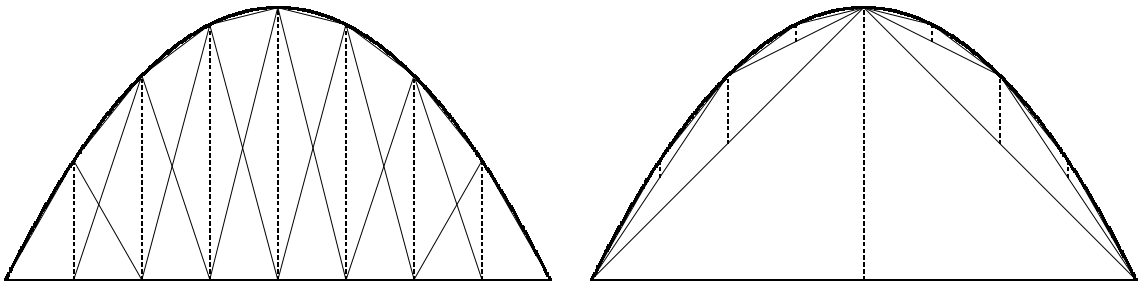
$$f_j(x) = \sum_{\ell=0}^{j} \sum_{k \in \nabla_\ell} \beta_{\ell,k} \varphi_{\ell,k}(x)$$

$$\text{mit} \quad \beta_{\ell,k} = \begin{cases} f(2^{-\ell}k), & \text{falls } \ell = 0, \\ f(2^{-\ell}k) - \frac{1}{2}\big\{f\big(2^{-\ell}(k-1)\big) + f\big(2^{-\ell}(k+1)\big)\big\}, & \text{falls } \ell \in \mathbb{N}. \end{cases}$$

Die Funktion f_j heißt dann *hierarchischer Interpoland von* f, die Koeffizienten $\beta_{\ell,k}$ heißen *hierarchische Überschüsse*. Ist $f \in C^2([0,1])$, dann liefert eine Taylor-Entwicklung von $f(x \pm h)$ in x für $\ell > 0$ die Abschätzung

$$\beta_{\ell,k} \le 2^{-(2\ell+1)} \|f''\|_{C([2^{-\ell}(k-1), 2^{-\ell}(k+1)])} \le 2^{-(2\ell+1)} \|f''\|_{C([0,1])}. \tag{7.6}$$

Beispiel 7.18 Im Fall einer Parabel ergeben sich folgende Darstellungen auf Level $j = 3$:



nodale Basis hierarchische Basis

△

Mit Hilfe des hierarchischen Interpolanden erhalten wir die Quadraturformel

$$I[f_j] = \sum_{\ell=0}^{j} \sum_{k \in \nabla_\ell} \beta_{\ell,k} \int_0^1 \varphi_{\ell,k}(x)\,\mathrm{d}x = \frac{f(0)}{2} + \frac{f(1)}{2} + \sum_{\ell=1}^{j} \sum_{k \in \nabla_\ell} 2^{-\ell} \beta_{\ell,k}.$$

Ist die Funktion glatt, dann werden die hierarchischen Überschüsse gemäß (7.6) sehr schnell sehr klein. Sind sie hingegen groß, ist dies ein Indikator dafür, dass die Funktion an dieser Stelle nicht glatt ist. Daher ist folgender adaptive Quadraturalgorithmus naheliegend:

Algorithmus 7.19 (adaptive Quadratur)
input: Funktion $f \in C([0,1])$ und Fehlertoleranz ε
output: Approximation $Q \approx I[f]$

① Initialisierung: setze $Q := \big(f(0) + f(1)\big)/2$ und $(x,h) := (1/2, 1/2)$

② berechne den hierarchischen Überschuss

$$E(x,h) = f(x) - \frac{1}{2}\big(f(x-h) + f(x+h)\big)$$

③ falls $|E(x,h)| > \varepsilon$, dann datiere auf $Q := Q + hE(x,h)$ und gehe nach ② mit $(x,h) := (x - h/2, h/2)$ und $(x,h) := (x + h/2, h/2)$

Bemerkung: Der Algorithmus kann zu früh terminieren, insbesondere wenn Nullstellen und Stützstellen zusammenfallen, wie beispielsweise im Fall der Funktion $\sin(2^\ell \pi x)$.

7.6 Euler-Maclaurinsche Summenformel

Definition 7.20 Die **Bernoulli-Polynome** B_n vom Grad n sind rekursiv definiert durch $B_0(x) := 1$ und

$$B_n' := B_{n-1}, \quad \int_0^1 B_n(x)\,\mathrm{d}x = 0, \quad n = 1, 2, \ldots. \tag{7.7}$$

Die rationalen Zahlen

$$b_n = n! B_n(0), \quad n = 1, 2, \ldots$$

heißen **Bernoulli-Zahlen**.

Die ersten Bernoulli-Polynome sind gegeben durch

$$B_0(x) = 1, \quad B_1(x) = x - \frac{1}{2}, \quad B_2(x) = \frac{1}{2}x^2 - \frac{1}{2}x + \frac{1}{12}.$$

Wegen $\int_0^1 B_{n-1}\,\mathrm{d}x = B_n(1) - B_n(0)$, ist die Normalisierungsbedingung äquivalent zu

$$B_n(0) = B_n(1), \quad n = 2, 3, \ldots. \tag{7.8}$$

Lemma 7.21 Die Bernoulli-Polynome erfüllen die Symmetrieeigenschaft

$$B_n(x) = (-1)^n B_n(1 - x), \quad x \in \mathbb{R}, \quad n = 0, 1, 2, \ldots. \tag{7.9}$$

Beweis. Offensichtlich gilt (7.9) für $n = 0$. Angenommen, (7.9) gilt für ein $n \geq 0$. Integration von (7.9) liefert

$$B_{n+1}(x) = (-1)^{n+1} B_{n+1}(1 - x) + \beta_{n+1}, \quad \beta_{n+1} \in \mathbb{R}.$$

Die Normalisierungsbedingung aus (7.7) liefert $\beta_{n+1} = 0$ und somit die Behauptung im Fall $n + 1$. □

Lemma 7.22 Die Bernoulli-Polynome B_{2m+1}, $m = 1, 2, \ldots$, ungeraden Grades besitzen exakt 3 Nullstellen in $[0, 1]$, und diese sind in den Punkten $0, 1/2$ und 1. Die Bernoulli-Polynome B_{2m}, $m = 1, 2, \ldots$, geraden Grades erfüllen $B_{2m}(0) \neq 0$.

Beweis. Aus (7.8) und (7.9) folgt, dass B_{2m+1} verschwindet in den Punkten $0, 1/2$ und 1. Wir beweisen durch Induktion, dass dies die einzigen Nullstellen von B_{2m+1} in $[0, 1]$ sind. Für $m = 1$ ist die Aussage richtig, da $B_3 \in \Pi_3$ nur drei Nullstellen besitzt. Angenommen, dass die Aussage ist richtig für B_{2m+1} und dass B_{2m+3} noch eine weitere Nullstelle $\alpha \in [0, 1]$ besitzt. Wegen der Symmetrie (7.9) dürfen wir annehmen, dass $\alpha \in (0, 1/2)$ gilt. Aus dem Satz von Rolle folgt dann aber, dass $B_{2m+3}' = B_{2m+2}$ mindestens je eine Nullstelle in $(0, \alpha)$

und $(\alpha, 1/2)$ besitzt. Damit besitzt aber auf $B'_{2m+2} = B_{2m+1}$ eine Nullstelle in $(0, 1/2)$, was im Widerspruch zu unserer Annahme steht.
Aufgrund der Nullstellen von B_{2m+1} impliziert der Satz von Rolle, dass $B'_{2m+1} = B_{2m}$ eine Nullstelle in $(0, 1/2)$ besitzt. Gilt nun $B_{2m}(0) = 0$, dann besitzt $B'_{2m} = B_{2m-1}$ eine Nullstelle in $(0, 1/2)$, was dem oben schon bewiesenen widerspricht. □

Für $N \in \mathbb{N}$ sei $h = (b-a)/N$ und $x_i = a + ih$, $i = 0, 1, \ldots, N$. Die zusammengesetzte Trapezregel ist definiert durch

$$T_N[f] := \frac{h}{2}\big\{f(a) + 2f(x_1) + 2f(x_2) + \cdots + 2f(x_{N-1}) + f(b)\big\}.$$

Wir beweisen nun die *Euler-Maclaurinsche Summenformel*.

Satz 7.23 Sei $f \in C^m([a,b])$ für ein $m \geq 2$. Dann gilt die Euler-Maclaurinsche Summenformel

$$\int_a^b f(x)\,\mathrm{d}x = T_N(f) - \sum_{j=1}^{\lfloor m/2 \rfloor} \frac{b_{2j} h^{2j}}{(2j)!}\big(f^{(2j-1)}(b) - f^{(2j-1)}(a)\big) + (-1)^m h^m \int_a^b \widetilde{B}_m\left(\frac{x-a}{h}\right) f^{(m)}(x)\,\mathrm{d}x, \tag{7.10}$$

wobei $\widetilde{B}_m$ das periodisierte Bernoulli-Polynom B_m bezeichnet

$$\widetilde{B}_m(x) := B_m(x - \lfloor x \rfloor).$$

Beweis. Sei $g \in C^m([0,1])$, dann ergibt $(m-1)$-malige partielle Integration zusammen mit (7.8)

$$\int_0^1 B_1(z) g'(z)\,\mathrm{d}x = \sum_{j=2}^{m} (-1)^j B_j(0)\big(g^{(j-1)}(1) - g^{(j-1)}(0)\big) - (-1)^m \int_0^1 B_m(z) g^{(m)}(z)\,\mathrm{d}z.$$

Einsetzen von

$$\int_0^1 B_1(z) g'(z)\,\mathrm{d}z = \frac{1}{2}\big(g(1) + g(0)\big) - \int_0^1 g(z)\,\mathrm{d}z$$

liefert unter Beachtung, dass alle ungeraden Bernoulli-Zahlen verschwinden, die Gleichung

$$\int_0^1 g(z)\,\mathrm{d}z = \frac{h}{2}\big(g(1) + g(0)\big) - \sum_{j=1}^{\lfloor m/2 \rfloor} \frac{b_{2j}}{(2j)!}\big(f^{(2j-1)}(b) - f^{(2j-1)}(a)\big) + (-1)^m \int_0^1 B_m(x) g^{(m)}(z)\,\mathrm{d}z.$$

Substituieren wir $x = x_k + hz$ und $g(z) = f(x_k + hz)$, dann erhalten wir

$$\int_{x_k}^{x_{k+1}} f(x)\,\mathrm{d}x = \frac{1}{2}\big(g(x_k) - g(x_{k+1})\big) - \sum_{j=1}^{\lfloor m/2 \rfloor} \frac{b_{2j} h^{2j}}{(2j)!}\big(f^{(2j-1)}(x_{k+1}) - f^{(2j-1)}(x_k)\big) + (-1)^m h^m \int_{x_k}^{x_{k+1}} B_m\left(\frac{x-x_k}{h}\right) f^{(m)}(x)\,\mathrm{d}x.$$

Aufsummieren über alle $k = 0, 1, \ldots, N-1$ liefert schließlich (7.10). □

7.7 Romberg-Verfahren

Während die Erhöhung der Knotenzahl N bei Trapez- oder der Simpson-Regel lediglich eine Verbesserung der Form N^{-p} für ein festes p ergibt, wollen wir hier eine Methode vorstellen, bei der N und p gleichzeitig erhöht werden.

Sei

$$T_N^1[f] := \frac{h}{2}\left\{ f(a) + 2\sum_{n=1}^{N-1} f(a+nh) + f(b) \right\}$$

die zusammengesetzte Trapezregel zur Schrittweite $h = (b-a)/N$. Vorausgesetzt $f \in C^4([a,b])$, dann liefert die Euler-Maclaurinsche Summenformel (7.10) die Fehlerdarstellung

$$\int_a^b f(x)\,\mathrm{d}x = T_N^1[f] + \gamma_1 h^2 + \mathcal{O}(h^4)$$

für eine Konstante γ_1, die von f abhängt, aber nicht von h. Daher folgt für die halbe Schrittweite

$$\int_a^b f(x)\,\mathrm{d}x = T_{2N}^1[f] + \gamma_1 \frac{h^2}{4} + \mathcal{O}(h^4).$$

Aus diesen beiden Gleichungen können wir nun den Term $\gamma_1 h^2$ eliminieren:

$$\int_a^b f(x)\,\mathrm{d}x = \frac{1}{3}\Big(4T_{2N}^1[f] - T_N^1[f]\Big) + \mathcal{O}(h^4).$$

Die Quadraturformel

$$T_N^2[f] := \frac{1}{3}\Big(4T_{2N}^1[f] - T_N^1[f]\Big)$$

besitzt demnach die verbesserte Fehlerordnung $\mathcal{O}(h^4)$. Wie man leicht nachrechnet, entspricht $T_N^2[f]$ gerade der zusammengesetzten Simpson-Regel $S_N[f]$.

Ist $f \in C^6([a,b])$, dann folgt aus der entsprechenden Linearkombination der Euler-Maclaurinschen Summenformeln (7.10) zu den Schrittweiten h und $h/2$, dass

$$\int_a^b f(x)\,\mathrm{d}x = T_N^2[f] + \gamma_2 h^4 + \mathcal{O}(h^6)$$

mit einer nur von f abhängigen Konstante γ_2. Geeignete Linearkombination mit

$$\int_a^b f(x)\,\mathrm{d}x = T_{2N}^2[f] + \gamma_2 \frac{h^4}{16} + \mathcal{O}(h^6)$$

liefert die Quadraturformel (es ist gerade die zusammengesetzte Milne-Regel)

$$T_N^3[f] := \frac{1}{15}\Big(16T_{2N}^2 - T_N^2[f]\Big)$$

von der Fehlerordnung $\mathcal{O}(h^6)$. Diese benötigt die Auswertungen der Trapezsummen zu den Schrittweiten $h, h/2$ und $h/4$.

Offensichtlich kann dieses Vorgehen fortgesetzt werden, was auf das *Romberg-Verfahren* führt:

Algorithmus 7.24 (Romberg-Verfahren)
input: Funktion f, Ausgangsschrittweite $h = 1/N$, Rekursionstiefe M
output: Approximation t_0^{M+1} an $I[f]$

① berechne für alle $k = 0, 1, 2, \ldots, M$ die Trapezsummen $t_k^1 := T_{2^k N}^1[f]$

② für alle $m = 1, 2, \ldots, M$ berechne

$$t_k^{m+1} := \frac{1}{4^m - 1}\big(4^m t_{k+1}^m - t_k^m\big), \quad k = 0, 1, 2, \ldots, M - m \tag{7.11}$$

Der Aufwand von Romberg-Verfahrens ist dabei

$$\mathcal{O}\Big(\sum_{k=0}^{M} 2^k N\Big) = \mathcal{O}(2^M N).$$

Der Quadraturfehler wird in folgendem Satz abgeschätzt.

Satz 7.25 Sei $f \in C^m([a, b])$, dann gilt

$$\left| \int_a^b f(x)\,\mathrm{d}x - t_k^m \right| \le C_m \big\| f^{(2m)} \big\|_{C([a,b])} \left(\frac{h}{2^k}\right)^{2m}, \quad k = 0, 1, \ldots, \tag{7.12}$$

mit einer von m abhängigen Konstante C_m.

Beweis. Durch Induktion zeigen wir, dass Konstanten $\gamma_{j,i}$ existieren, so dass

$$\left| \int_a^b f(x)\,\mathrm{d}x - t_k^i - \sum_{j=i}^{m-1} \gamma_{j,i}\big(f^{(2j-1)}(b) - f^{(2j-1)}(a)\big)\left(\frac{h}{2^k}\right)^{2j} \right| \le \gamma_{m,i} \big\| f^{(2m)} \big\|_{C([a,b])} \left(\frac{h}{2^k}\right)^{2m} \tag{7.13}$$

für alle $i = 1, 2, \ldots, m$ und $m = 0, 1, \ldots$. Dabei ist die Summe gleich Null, falls $i = m$ ist. Aufgrund der Euler-Maclaurinschen Entwicklung (7.10) ist (7.13) im Falle $i = 1$ wahr, falls $\gamma_{j,1} = b_{2j}/(2j)!$ für alle $j = 1, 2, \ldots, m - 1$ und

$$\gamma_{m,1} = (b - a)\|B_{2m}\|_{C([0,1])}$$

gesetzt wird.
Wir setzen

$$F_j := f^{(2j-1)}(b) - f^{(2j-1)}(a), \quad j = 1, 2, \ldots, m - 1.$$

Angenommen, (7.13) gilt für ein $1 \le i < m$. Dann folgt aus (7.11), dass

$$\begin{aligned}\frac{4^i}{4^i-1}\Big[\int_a^b f(x)\,\mathrm{d}x - t_{k+1}^i - \sum_{j=i}^{m-1}\left(\frac{h}{2^{k+1}}\right)^{2j}\gamma_{j,i}F_j\Big] \\ - \frac{1}{4^i-1}\Big[\int_a^b f(x)\,\mathrm{d}x - t_k^i - \sum_{j=i}^{m-1}\left(\frac{h}{2^k}\right)^{2j}\gamma_{j,i}F_j\Big] \\ = \int_a^b f(x)\,\mathrm{d}x - t_k^{i+1} - \sum_{j=i+1}^{m-1}\left(\frac{h}{2^k}\right)^{2j}\gamma_{j,i+1}F_j,\end{aligned}$$

wobei

$$\gamma_{j,i+1} = \frac{4^{i-j}-1}{4^i-1}\gamma_{j,i}, \quad j = i+1, i+2, \ldots, m-1.$$

Mit Hilfe der Induktionsannahme und

$$\gamma_{m,i+1} := \frac{4^{i-m}+1}{4^i-1}$$

ergibt sich

$$\left|\int_a^b f(x)\,\mathrm{d}x - t_k^{i+1} - \sum_{j=i+1}^{m-1}\left(\frac{h}{2^k}\right)^{2j}\gamma_{j,i+1}F_j\right| \le \gamma_{m,i+1}\left\|f^{(2m)}\right\|_{C([a,b])}\left(\frac{h}{2^k}\right)^{2m}.$$

□

Beispiel 7.26 Zu $f(x) = \sin^4 x$ berechnen wir $I[f] = \int_0^\pi \sin^4 x\,\mathrm{d}x$ mit dem Romberg-Verfahren. Ausgehend von

$$\begin{aligned}T_1[f] &= \frac{\pi}{2}(\sin^4 0 + \sin^4 \pi) = 0,\\ T_2[f] &= \frac{\pi}{4}(\sin^4 0 + 2\sin^4\frac{\pi}{2} + \sin^4\pi) = \frac{\pi}{2},\\ T_4[f] &= \frac{\pi}{8}(\sin^4 0 + 2\sin^4\frac{\pi}{4} + 2\sin^4\frac{\pi}{2} + 2\sin^4\frac{3\pi}{4} + \sin^4\pi) = \frac{3\pi}{8},\end{aligned}$$

erhalten wir die Näherung

$$\begin{array}{ccccc} t_0^1 = T_1[f] = 0 & & & & \\ & \searrow^{-\frac{1}{3}} & & & \\ t_1^1 = T_2[f] = \dfrac{\pi}{2} & \xrightarrow{\frac{4}{3}} & t_0^2 = \dfrac{2\pi}{3} & \searrow^{-\frac{1}{15}} & \\ & \searrow^{-\frac{1}{3}} & & & \\ t_2^1 = T_4[f] = \dfrac{3\pi}{8} & \xrightarrow{\frac{4}{3}} & t_1^2 = \dfrac{\pi}{3} & \xrightarrow{\frac{16}{15}} & t_0^2 = \dfrac{14}{45}\pi \end{array}$$

△

7.8 Quadratur periodischer Funktionen

Im Fall 2π-periodischer Funktionen $f : \mathbb{R} \to \mathbb{R}$ stimmt die Trapezregel mit der *Mittelpunktsregel* überein

$$I[f] = \int_0^{2\pi} f(x)\,\mathrm{d}x \approx Q_N[f] = \frac{2\pi}{N} \sum_{k=1}^{N} f\left(\frac{2\pi k}{N}\right).$$

Für den Quadraturfehler liefert die Euler-Maclaurinsche Summenformel (7.10) das folgende Korollar.

Korollar 7.27 Sei $f \in C^{2m+1}_{\mathrm{per}}([0, 2\pi])$ und sei $N \in \mathbb{N}$. Dann ist der Fehler der Mittelpunktsregel beschränkt durch

$$\left|I[f] - Q_N[f]\right| \leq \frac{C}{N^{2m+1}} \int_0^{2\pi} \left|f^{(2m+1)}(x)\right| \mathrm{d}x$$

mit

$$C := 2 \sum_{k=1}^{\infty} \frac{1}{k^{2m+1}}.$$

Beweis. Wegen der Periodizität gilt $f^{(j)}(0) = f^{(j)}(2\pi)$, $j = 0, 1, \ldots, 2m+1$. Daher reduziert sich die Euler-Maclaurinsche Summenformel (7.10) auf

$$\left|I[f] - Q_N[f]\right| = -\left(\frac{2\pi}{N}\right)^{2m+1} \int_0^{2\pi} \widetilde{B}_{2m+1}\left(\frac{2\pi x}{N}\right) f^{(2m+1)}(x)\,\mathrm{d}x.$$

Die periodisierten Bernoulli-Polynome genügen der Fourier-Entwicklung

$$\widetilde{B}_{2m}(x) = 2(-1)^{m-1} \sum_{k=1}^{\infty} \frac{\cos(2\pi k x)}{(2\pi k)^{2m}}, \quad \widetilde{B}_{2m+1}(x) = 2(-1)^{m-1} \sum_{k=1}^{\infty} \frac{\sin(2\pi k x)}{(2\pi k)^{2m+1}}$$

für alle $m = 0, 1, 2, \ldots$. Hieraus folgt

$$\left|\widetilde{B}_{2m+1}(x)\right| \leq 2 \sum_{k=1}^{\infty} \frac{1}{(2\pi k)^{2m+1}}, \quad x \in \mathbb{R},$$

und damit die Behauptung. □

Ist f eine analytische Funktion, dann kann f in jedem Punkt $x_0 \in [0, 2\pi]$ in eine Taylor-Reihe entwickelt werden

$$f(x) = \sum_{k=0}^{\infty} \frac{f^{(k)}(x_0)}{k!}(x - x_0)^k, \quad |x - x_0| < r(x_0)$$

mit Konvergenzradius

$$\frac{1}{r(x_0)} = \limsup_{k\to\infty} \sqrt[k]{\frac{f^{(k)}(x_0)}{k!}} > 0.$$

Folglich existiert eine Konstante $c_f > 0$, so dass

$$\left\|f^{(k)}\right\|_{C([0,2\pi])} \leq c_f \frac{k!}{R^k}, \quad R := \min_{x\in[0,2\pi]} r(x) > 0.$$

Wir erhalten demnach im Falle einer *analytischen* Funktion f die Fehlerabschätzung

$$\left|I[f] - Q_N[f]\right| \leq \frac{2\pi c_f\, C\,(2m+1)!}{(RN)^{2m+1}}$$

für alle $m \in \mathbb{N}$, unabhängig von N. Wählen wir nun $m \sim (RN-1)/2$, dann folgt

$$\frac{(2m+1)!}{(RN)^{2m+1}} \sim \frac{(RN)!}{(RN)^{RN}} \leq \frac{(RN/2)^{RN/2}(RN)^{RN/2}}{(RN)^{RN}} \leq 2^{-RN/2},$$

das heißt, die Mittelpunktsregel konvergiert *exponentiell* in N,

$$\left|I[f] - Q_N[f]\right| \leq a\exp(-bN), \quad a, b > 0.$$

Exponentielle Konvergenz bedeutet, dass eine Verdoppelung der Anzahl N der Funktionsauswertungen die doppelte Anzahl richtiger Ziffern ergibt.

7.9 Gauß-Quadratur

Wir beantworten zunächst folgende Frage: Gegeben seien m Knoten $x_1, x_2, \ldots, x_m$ und m Gewichte $w_1, w_2, \ldots, w_m$. Wie groß ist maximal der Exaktheitsgrad der Quadraturformel

$$Q[f] = \sum_{i=1}^{m} w_i f(x_i) \approx I_w[f] = \int_a^b f(x)w(x)\,\mathrm{d}x? \tag{7.14}$$

Proposition 7.28 Sei $w > 0$ in (a, b). Dann ist der Exaktheitsgrad der Quadraturformel $Q[\cdot]$ aus (7.14) maximal $q = 2m - 1$.

Beweis. Für

$$p(x) = \prod_{i=1}^{m}(x - x_i)^2 \in \Pi_{2m}$$

ist $Q[p] = 0$, während gilt

$$I_w[p] = \int_a^b \underbrace{\prod_{i=1}^{m}(x - x_i)^2 w(x)}_{>0 \text{ für } x\neq a,x_1,x_2,\ldots,x_m,b}\,\mathrm{d}x > 0.$$

□

Wir nehmen nun an, Q besitze den Exaktheitsgrad $q = 2m - 1$. Definieren wir

$$p(x) = \prod_{i=1}^{m}(x - x_i) \in \Pi_m,$$

dann gilt für jedes beliebige $r \in \Pi_{m-1}$ dass $rp \in \Pi_{2m-1}$ und $(rp)(x_i) = 0$, $i = 1, 2, \ldots, m$. Aufgrund unserer Annahme folgt also

$$0 = Q[rp] = I_w[rp] = \int_a^b r(x)p(x)w(x)\,\mathrm{d}x. \tag{7.15}$$

Dies bedeutet, r und p sind zueinander orthogonal bezüglich des Innenprodukts

$$\langle f, g\rangle = \int_a^b f(x)g(x)w(x)\,\mathrm{d}x. \tag{7.16}$$

Da dies für alle $r \in \Pi_{m-1}$ gilt, muss p ein Vielfaches des m-ten Orthonormalpolynoms bezüglich (7.16) sein.

Seien umgekehrt p das m-te Orthonormalpolynom bezüglich (7.16) und x_i dessen Nullstellen. Wählen wir die Gewichte w_i wie in Abschnitt 7.3 über die Lagrange-Grundpolynome bezüglich $\{x_i\}$,

$$w_i = \int_a^b L_i(x)w(x)\,\mathrm{d}x, \tag{7.17}$$

dann ist Q für $p \in \Pi_m$ exakt gemäß Proposition 7.10. Eine vollständige Basis von Π_{2m-1} erhält man durch Hinzunahme der Funktionen $\{px, px^2, \ldots, px^{m-1}\}$, für die Q, wie in (7.15) gezeigt, ebenfalls exakt ist. Dies ist die *m-stufige Gauß-Formel* G_w bezüglich w.

Satz 7.29 Wählt man die Knoten $\{x_i\}_{i=0}^m$ der Quadraturformel G_w als Nullstellen des m-ten Orthonormalpolynoms bezüglich (7.16) und die Gewichte w_i durch (7.17), dann besitzt G_w den Exaktheitsgrad $2m-1$. Die Gewichte w_i sind dabei allesamt positiv und durch die Christoffel-Funktion aus Definition 6.21 gegeben

$$w_i = \Lambda_{m-1}(x_i). \tag{7.18}$$

Beweis. Es verbleibt die Eigenschaft der Gewichte nachzuweisen. Setzen wir

$$0 \le p(x) = \prod_{\substack{j=1\\ j\neq i}}^{m} \frac{(x-x_j)^2}{(x_i-x_j)^2} \in \Pi_{2m-2}, \tag{7.19}$$

dann ergibt sich

$$w_i = w_i p(x_i) = G_w[p] = I_w[p] = \int_a^b p(x)w(x)\,\mathrm{d}x > 0. \tag{7.20}$$

Also ist $w_i > 0$ für alle $i = 1, \ldots, m$. Damit folgt jedoch für beliebiges $q \in \Pi_{m-1}$ mit $q(x_i) = 1$ dass

$$w_i = w_i q^2(x_i) \le G_w[q^2] = I_w[q^2] = \|q\|^2_{L^2([a,b];w)}.$$

Also ist $w_i \le \|q\|^2_{L^2([a,b];w)}$, wobei gemäß (7.20) Gleichheit gilt, falls q^2 mit p aus (7.19) übereinstimmt. Nach Satz 6.22 ist jedoch das Minimum von $\|q\|^2_{L^2([a,b];w)}$ gerade $\Lambda_{m-1}(x_i)$. □

Beispiel 7.30 Im in der Praxis wichtigsten Fall $w \equiv 1$ sind die Orthonormalpolynome gerade die Legendre-Polynome aus Beispiel 6.17. Bis $m = 5$ lauten die Gewichte und Knoten:

m	x_i	w_i
1	0	2
2	$-\frac{1}{\sqrt{3}}$, $\frac{1}{\sqrt{3}}$	1, 1
3	$-\frac{\sqrt{15}}{5}$, 0, $\frac{\sqrt{15}}{5}$	$\frac{5}{9}$, $\frac{8}{9}$, $\frac{5}{9}$
4	$\pm\frac{1}{35}\sqrt{525-70\sqrt{30}}$ $\pm\frac{1}{35}\sqrt{525+70\sqrt{30}}$	$\frac{1}{36}(18+\sqrt{30})$ $\frac{1}{36}(18-\sqrt{30})$
5	0 $\pm\frac{1}{21}\sqrt{245-14\sqrt{70}}$ $\pm\frac{1}{21}\sqrt{245+14\sqrt{70}}$	$\frac{128}{225}$ $\frac{1}{900}(322+13\sqrt{70})$ $\frac{1}{900}(322-13\sqrt{70})$

△

Beispiel 7.31 Von großer Bedeutung ist auch die Gewichtsfunktion $w(x) = 1/\sqrt{1-x^2}$, vor allem, da die zugehörenden Orthonormalpolynome, die Tschebyscheff-Polynome, gut bekannt sind (vergleiche Anschnitt 6.2). Die Nullstellen $\{x_i\}$ des m-ten Polynoms $T_m(x) = \cos(m \arccos x)$ sind gerade

$$x_i = \cos\left(\frac{2i-1}{2m}\pi\right), \quad i = 1, 2, \ldots, m.$$

Die Gewichte berechnen sich über (7.18) und Definition 6.21. Bezeichnet nämlich $u_0 \equiv 1/\sqrt{\pi}$ und $u_k(x) = \sqrt{2/\pi}\cos(k \arccos x)$, $k > 0$, die Orthonormalpolynome bezüglich $w(x)$, dann ist

$$\begin{aligned}
\frac{1}{w_i} = \frac{1}{\Lambda_{m-1}(x_i)} &= \sum_{k=0}^{m-1} u_k^2(x_i) \\
&= \frac{1}{\pi} + \frac{2}{\pi}\sum_{k=1}^{m-1} \cos^2\left(k\left(i-\frac{1}{2}\right)\frac{\pi}{m}\right) \\
&= \frac{1}{\pi} + \frac{1}{\pi}\sum_{k=1}^{m-1}\left[1+\cos\left(k(2i-1)\frac{\pi}{m}\right)\right] \\
&= \frac{m}{\pi} + \frac{1}{\pi}\sum_{k=1}^{m-1}\cos\left(k(2i-1)\frac{\pi}{m}\right).
\end{aligned}$$

In der verbliebenen Summe tritt für $k = \ell$ und $k = m - \ell$ der gleiche Summand mit verschiedenen Vorzeichen auf. Also verschwindet die Summe, dies bedeutet $w_i = \pi/m$, und die Quadraturformel lautet

$$\int_{-1}^{1} \frac{f(x)}{\sqrt{1-x^2}}\,\mathrm{d}x \approx \frac{\pi}{m}\sum_{i=1}^{m} f\left(\cos\left(\frac{2i-1}{2m}\pi\right)\right). \tag{7.21}$$

△

Für den Fehler bei der Gauß-Quadratur gilt folgendes Resultat:

Satz 7.32 Sei $f \in C^{2m}([a,b])$ und G_w die m-stufige Gauß-Formel für I_w. Dann gilt

$$\big|I_w[f] - G_w[f]\big| \le \frac{\|f^{(2m)}\|_{C([a,b])}}{(2m)!\gamma_m^2},$$

wobei γ_m wie in (6.8) den Höchstkoeffizenten des Orthonormalpolynoms $u_m(x)$ bezeichnet.

Beweis. Es sei $p \in \Pi_{2m-1}$ dasjenige Polynom, für das gilt

$$f(x_i) = p(x_i), \quad f'(x_i) = p'(x_i), \quad i = 1, 2, \dots, m.$$

Ferner bezeichne $s(x) := \prod_{i=1}^m (x - x_i)^2 \in \Pi_{2m}$ das zugehörige Knotenpolynom. Wie im Beweis von Satz 7.8 betrachten wir die Funktion

$$h(t) := f(t) - p(t) - \frac{s(t)}{s(x)}\big(f(x) - p(x)\big), \quad x \in [a,b].$$

Da sie m doppelte Nullstellen in den Punkten $x_1, x_2, \dots, x_m$, sowie eine einfache Nullstelle in $t = x$ besitzt, folgt, dass h' $2m$ Nullstellen $\tau_1^{(1)} < \tau_2^{(1)} < \cdots < \tau_{2m}^{(1)}$ in $[a,b]$ besitzt, h'' $2m-1$ Nullstellen $a < \tau_1^{(2)} < \tau_2^{(2)} < \cdots < \tau_{2m-1}^{(2)} < b$ besitzt, und so weiter. Schließlich besitzt $h^{(2m)}$ eine Nullstelle $\xi \in (a,b)$. Da $p^{(2m)} \equiv 0$ gilt, folgt

$$0 = h^{(2m)}(\xi) := f^{(2m)}(\xi) - \frac{(2m)!}{s(x)}\big(f(x) - p(x)\big),$$

beziehungsweise

$$f(x) - p(x) = \frac{f^{(2m)}(\xi)}{(2m)!}\prod_{i=1}^m (x - x_i)^2.$$

Da die x_i gerade die Nullstellen des $(m+1)$-ten Orthonormalpolynoms sind, gilt

$$f(x) - p(x) = \frac{f^{(2m)}(\xi)}{(2m)!\gamma_m^2} u_m^2(x).$$

Es folgt

$$\begin{aligned}\big|I_w[f] - G_w[f]\big| &= \big|I_w[f] - G_w[p]\big| = \big|I_w[f-p]\big| \\ &\le \frac{\|f^{(2m)}\|_{C([a,b])}}{(2m)!\gamma_m^2} \underbrace{\int_a^b u_m^2(x) w(x)\,\mathrm{d}x}_{=1} = \frac{\|f^{(2m)}\|_{C([a,b])}}{(2m)!\gamma_m^2}.\end{aligned}$$

□

Bemerkung: Für die Gauß-Legendre-Formel ($w \equiv 1$) ist

$$\gamma_m^2 = \frac{2m+1}{2} \cdot \frac{(2m)!^2}{2^{2m} m!^4} \approx \frac{1}{\pi} 4^m,$$

vergleiche Beispiel 6.20. Hingegen gilt für die Gauß-Tschebyscheff-Formel (7.21) gemäß Abschnitt 6.2

$$\gamma_m^2 = \begin{cases} 1/\pi, & \text{falls } m = 0, \\ 4^m/(2\pi), & \text{falls } m \ge 1. \end{cases}$$

Abschließend überlegen wir uns noch, wie man die Knoten x_i und die Gewichte w_i numerisch bestimmt. Dazu betrachten wir die aus den Rekursionskoeffizenten der zugehörigen Orthonormalpolynomen generierte Jacobi-Matrix

$$\mathbf{T}_m = \begin{bmatrix} \alpha_1 & \beta_1 & & 0 \\ \beta_1 & \alpha_2 & \ddots & \\ & \ddots & \ddots & \beta_{m-1} \\ 0 & & \beta_{m-1} & \alpha_m \end{bmatrix}.$$

Die Eigenwerte von $\mathbf{T}_m$ sind gerade die Nullstellen von u_m, also die Knoten x_i der Gauß-Quadraturformel. Die Gewichte bestimmen wir mit Hilfe der Resultate aus Abschnitt 6.6. Definiere

$$R[f] = \sum_{i=1}^{m} w_{1,i}^2 f(x_i), \tag{7.22}$$

wobei $w_{1,i}$ die erste Komponente des zu x_i gehörigen Eigenvektors von $\mathbf{T}_m$ mit Euklidnorm 1 bezeichnet. Bezeichnet $\langle \cdot, \cdot \rangle$ das diskrete Innenprodukt (6.23), dann gilt gerade $R[f] = \langle f, 1 \rangle$.
Sei nun

$$\langle\!\langle f, g \rangle\!\rangle = \int_a^b f(x) g(x) w(x) \, \mathrm{d}x$$

das zur Gauß-Formel gehörende Innenprodukt. Nach Proposition 6.29 gilt, dass

$$\langle\!\langle u_k, u_\ell \rangle\!\rangle = \langle u_k, u_\ell \rangle = \delta_{k,\ell}, \quad k, \ell = 0, 1, \ldots, m.$$

Wie wir zu Beginn dieses Abschnitts gesehen haben, müssen die Gewichte der Gauß-Formel lediglich so bestimmt werden, dass die Quadraturformel für alle $p \in \Pi_m$ exakt ist, also etwa $p \in \{u_0, u_1, \ldots, u_m\}$. Nun gilt aber

$$I_w[u_i] = \langle\!\langle u_i, 1 \rangle\!\rangle = \langle u_i, 1 \rangle = R[u_i], \quad i = 0, 1, \ldots, m.$$

Folglich ist R aus (7.22) gerade die gesuchte Gauß-Formel.

Algorithmus 7.33 (Golub/Welsch) Bestimme mit dem QR-Verfahren die Eigenwerte x_i der Jacobi-Matrix $\mathbf{T}_m$ sowie die erste Komponente $w_{1,i}$ der zugehörigen Eigenvektoren. Dann ist (7.22) die Gauß-Quadraturformel.

Beispiel 7.34 Für $w \equiv 1$ ist $\alpha_i = 0$ und $\beta_i = i/\sqrt{4i^2 - 1}$. △

7.10 *hp*-Quadratur

In der Praxis treten häufig singuläre Integrale der Form

$$I_0^1[g] = \int_0^1 g(x) \, \mathrm{d}x, \quad \text{wobei} \quad g(x) = \frac{f(x)}{x^\alpha} \tag{7.23}$$

mit analytischem $f : [0, 1] \to \mathbb{R}$ und nicht ganzzahligem $\alpha \in (0, 1)$ auf. Daher wollen wir im folgenden deren effiziente Behandlung diskutieren.

Beachte: Da der Integrand und alle seine Ableitungen in $x = 0$ singulär sind, versagen hier alle bisher bekannten Fehlerabschätzungen.

Lemma 7.35 Es sei $f : [0,1] \to \mathbb{R}$ eine analytische Funktion und $\alpha \in (0,1)$. Dann existieren Konstanten $R > 0$ und $c_f > 0$, die nur von f abhängen, so dass gilt

$$\left| \left(\frac{f(x)}{x^\alpha} \right)^{(n)} \right| \le \frac{c_f n!}{R^n x^{\alpha+n}} \tag{7.24}$$

für alle $x \in (0,1]$ und $n \in \mathbb{N}_0$.

Beweis. Wie wir im Abschnitt 7.8 gesehen haben, existiert ein $R > 0$, so dass gilt

$$\frac{|f^{(k)}(x)|}{k!} \le \frac{\widetilde{c}_f}{R^k}, \quad k \in \mathbb{N}_0.$$

Ferner haben wir

$$\left| \left(x^{-\alpha}\right)^{(k)} \right| = \alpha(\alpha+1)\cdots(\alpha+k-1)\, x^{-(\alpha+k)} \le k!\, x^{-(\alpha+k)}, \quad k \in \mathbb{N}_0.$$

Die Leibnizsche Regel

$$(uv)^n = \sum_{k=0}^{n} \binom{n}{k} u^{(k)} v^{(n-k)}$$

liefert dann

$$\left| \left(\frac{f(x)}{x^\alpha} \right)^{(n)} \right| \le \widetilde{c}_f \sum_{k=0}^{n} \binom{n}{k} \frac{k!(n-k)!}{R^k x^{\alpha+n-k}} = \frac{\widetilde{c}_f n!}{R^n x^{\alpha+n}} \sum_{k=0}^{n} R^{n-k} x^k.$$

Die Summe lässt sich durch ein Integral abschätzen:

$$\sum_{k=0}^{n} R^{n-k} x^k \sim R^n \int_0^n \left(\frac{x}{R} \right)^t \mathrm{d}t = \frac{x^n - R^n}{\log(x) - \log(R)} \le c_R.$$

Hieraus folgt die Behauptung. □

Wir wollen Gauß-Legendre-Quadraturformeln von variabler Knotenzahl m benutzen, die gemäß Abschnitt 7.9 der Fehlerabschätzung

$$\left| I_a^b[g] - Q_{a,m}^b[g] \right| \le c_Q 4^{-2m} (b-a)^{2m+1} \frac{\|g^{(2m)}\|_{C([a,b])}}{(2m)!}$$

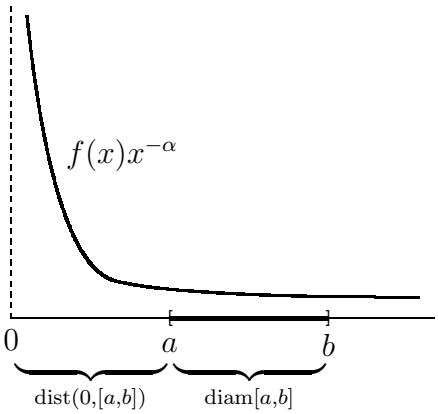
mit einer von m unabhängigen Konstanten c_Q genügen. Auf dem Intervall $[a,b] \subset (0,1]$ ergibt sich daher für die Funktion $g(x) = f(x)x^{-\alpha}$ wegen (7.24) die Fehlerschranke

$$\left| I_a^b[g] - Q_{a,m}^b[g] \right| \le \widetilde{C}_f \frac{b-a}{a^\alpha} \left(\frac{b-a}{4Ra} \right)^{2m}.$$

Ein Erhöhen der Knotenzahl m implizert also nur dann Konvergenz, falls $(b-a)/(4Ra) < 1$ ist, das heißt, falls gilt

$$a = \operatorname{dist}(0,[a,b]) \ge c_a \operatorname{diam}[a,b] = c_a(b-a) \quad \text{für ein } c_a > \frac{1}{4R}. \tag{7.25}$$

Gilt (7.25), so nennen wir das Intervall $[a,b]$ *zulässig*. Ist ein Intervall nicht zulässig, so werden wir es unterteilen.



Zuletzt müssen wir noch das Integral $[0, b]$ für kleines b behandeln. Wegen

$$\left| \int_0^b \frac{f(x)}{x^\alpha} \, \mathrm{d}x \right| \leq \|f\|_{C([0,b])} \int_0^b x^{-\alpha} \, \mathrm{d}x = \|f\|_{C([0,b])} \frac{b^{1-\alpha}}{1-\alpha} \xrightarrow{b \to 0} 0$$

lassen wir es für b klein genug einfach weg.
Wir können unsere Überlegungen wie folgt zusammenfassen.

Lemma 7.36 Ist $[a, b]$ zulässig und wählt man

$$m = \left\lceil \frac{\log\left(\frac{\varepsilon a^\alpha}{b-a}\right)}{2\log\left(\frac{b-a}{4Ra}\right)} \right\rceil, \tag{7.26}$$

dann gilt

$$\left| I_a^b[g] - Q_{a,m}^b[g] \right| = \mathcal{O}(\varepsilon).$$

Ist hingegen $a = 0$ und

$$b \leq B(\varepsilon) := \big((1-\alpha)\varepsilon\big)^{1/(1-\alpha)}, \tag{7.27}$$

so folgt

$$\left| I_0^b[g] \right| = \mathcal{O}(\varepsilon).$$

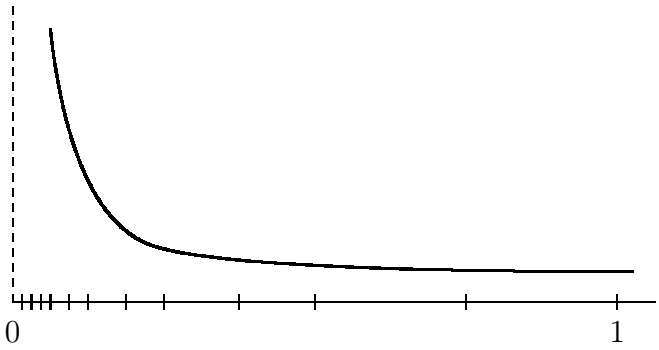
Wir erhalten damit einen rekursiven Quadraturalgorithmus:

Algorithmus 7.37 (*hp*-Quadratur)
input: Integrand g vom Typ (7.23) und Fehlertoleranz ε
output: Approximation $Q \approx I_0^1[g]$ mit $\left| Q - I_0^1[g] \right| = \mathcal{O}(\varepsilon)$

① Initialisierung: setze $Q := 0$ und $[a, b] := [0, 1]$

② Ist $[a, b]$ zulässig, dann setze $Q := Q + Q_{a,m}^b[g]$ mit m gemäß (7.26). Ansonsten, falls $[a, b] \not\subset [0, B(\varepsilon)]$, gehe nach ② um $I_a^{(a+b)/2}[g]$ und $I_{(a+b)/2}^b[g]$ mit der Genauigkeit $\varepsilon/2$ zu approximieren.

Beispiel 7.38 Im Fall der Zulässigkeitskonstante $c_a = 2$ liefert der Algorithmus folgende Unterteilung des Intervalls $[0, 1]$:



△

Die zulässigen Intervalle $[a,b]$ erfüllen stets $b-a=2^{-\ell}$, wobei gilt

$$1 \le \ell \le L := \lceil -\log_2\big(B(\varepsilon)\big)\rceil.$$

Speziell gibt es für jedes $1 \le \ell \le L$ nur $\mathcal{O}(1)$ zulässige Intervalle. Daher produziert Algorithmus 7.37 in der Tat die gewünschte Fehlerschranke

$$\big|Q - I_0^1[g]\big| \le C \sum_{\ell=0}^{L} 2^{-\ell}\varepsilon = \mathcal{O}(\varepsilon).$$

Satz 7.39 Algorithmus 7.37 konvergiert exponentiell, genauer, um eine Genauigkeit ε zu erzielen werden nur $\mathcal{O}\big(|\log(\varepsilon)|\log|\log(\varepsilon)|\big)$ Funktionsauswertungen benötigt.

Beweis. Auf dem Level $1 \le \ell \le L$ wird das Integral $I_a^b[g]$ wird mit der Genauigkeit $2^{-\ell}\varepsilon = (b-a)\varepsilon$ gebildet. Dabei ist die Knotenzahl m wegen $a \ge c_a(b-a) = 2^{-\ell}c_a$ beschränkt durch

$$\begin{aligned} m \le M(\ell) &= \left\lceil -\frac{\log(2^{-\alpha\ell}\varepsilon c_a^\alpha)}{2\log(2^{\ell+2}c_a R)} \right\rceil = \left\lceil \frac{\alpha\ell\log(2) - \alpha\log(c_a) - \log(\varepsilon)}{2\ell\log(2) + 2\underbrace{\log(4c_aR)}_{>0}} \right\rceil \\ &\le \left\lceil \frac{\alpha\ell\log(2) - \alpha\log(c_a) - \log(\varepsilon)}{2\ell\log(2)} \right\rceil \le \frac{|\log(\varepsilon)|}{\ell} + C_1. \end{aligned}$$

Da nur $\mathcal{O}(1)$ Integrale je Level gebildet werden, werden demnach insgesamt

$$\sum_{\ell=1}^{L} \frac{|\log(\varepsilon)|}{\ell} + C_1 \le |\log(\varepsilon)|\log(L) + C_1 L + 1$$

Funktionsauswertungen durchgeführt. Dabei wurde die harmonische Reihe durch $\sum_{\ell=1}^{L} 1/\ell \le \log(L)+1$ abgeschätzt.

Wir benötigen nun nur noch eine Schranke für das maximale Level L. Einsetzen der Definition von $B(\varepsilon)$ ergibt

$$L = \lceil -\log_2\big(B(\varepsilon)\big)\rceil = \lceil \log_2\big((1-\alpha)\varepsilon\big)/(\alpha-1)\rceil \le C_2|\log(\varepsilon)| + C_3.$$

Hieraus folgt die Behauptung. □

7.11 Tensorprodukt-Quadratur

Gegeben sei eine auf $\square = [0,1] \times [0,1]$ stetige Funktion $f : \square \to \mathbb{R}$. Auszuwerten sei das Integral

$$(I \otimes I)[f] = \int_0^1 \int_0^1 f(x,y)\,\mathrm{d}y\,\mathrm{d}x.$$

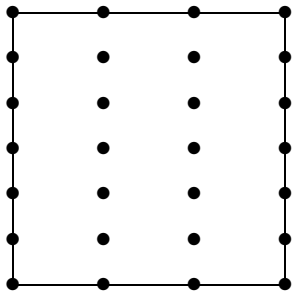
Um das Integral auf $\square$ zu approximieren, gehen wir von zwei eindimensionalen Quadraturformeln

$$Q_x[g] = \sum_{i=0}^{m} w_{x,i} g(x_{x,i}), \qquad Q_y[g] = \sum_{j=0}^{n} w_{y,j} g(x_{y,j})$$

aus und bilden deren Tensorprodukt

$$(Q_x \otimes Q_y)[f] = \sum_{i=0}^{m} \sum_{j=0}^{n} w_{x,i} w_{y,j} f(x_{x,i}, x_{y,j}).$$

Diese Quadraturformel heißt dann *Tensorprodukt-Quadraturformel*. Dies bedeutet, wir werten f auf einem regelmäßigen $(m \times n)$-Gitter aus:



Satz 7.40 Die Tensorprodukt-Quadraturformel $Q_x \otimes Q_y$ genügt der Fehlerabschätzung

$$\big|(I \otimes I)[f] - (Q_x \otimes Q_y)[f]\big| \leq \sup_{y \in [0,1]} E_x[f(\cdot, y)] + \sup_{x \in [0,1]} E_y[f(x, \cdot)],$$

wobei $E_x[g] := |I[g] - Q_x[g]|$ und $E_y[g] := |I[g] - Q_y[g]|$ die eindimensionalen Quadraturfehler bezeichnen.

Beweis. Wir benutzen ein Standard-Tensorproduktargument:

$$\begin{aligned}
&(I \otimes I)[f] - (Q_x \otimes Q_y)[f] \\
&\quad = (I \otimes I)[f] - (I \otimes Q_y)[f] + (I \otimes Q_y)[f] - (Q_x \otimes Q_y)[f] \\
&\quad = \big(I \otimes (I - Q_y)\big)[f] + \big((I - Q_x) \otimes Q_y\big)[f] \\
&\quad = \int_0^1 \left(\int_0^1 f(x,y)\,\mathrm{d}y - \sum_{j=0}^{n} w_{y,j} f(x, x_{y,j}) \right) \mathrm{d}x \\
&\qquad\qquad + \sum_{j=0}^{n} w_{y,j} \left(\int_0^1 f(x, x_{y,i})\,\mathrm{d}x - \sum_{i=0}^{m} w_{x,i} f(x_{x,i}, x_{y,j}) \right).
\end{aligned}$$

Nun gilt

$$\left| \int_0^1 \Bigg(\underbrace{\int_0^1 f(x,y)\,\mathrm{d}y - \sum_{j=0}^{n} w_{y,j} f(x,x_{y,j})}_{|\cdot| \le E_y[f(x,\cdot)]} \Bigg) \mathrm{d}x \right| \le \int_0^1 E_y[f(x,\cdot)]\,\mathrm{d}x \le \sup_{x\in[0,1]} E_y[f(x,\cdot)].$$

Andererseits ist

$$\left| \sum_{j=0}^{n} w_{y,j} \underbrace{\int_0^1 f(x,x_{y,i})\,\mathrm{d}x - \sum_{i=0}^{m} w_{x,i} f(x_{x,i},x_{y,j})}_{|\cdot| \le E_x[f(\cdot,x)]} \right| \le \sum_{j=0}^{n} w_{y,j} E_x[f(\cdot,x_{y,j})] \le \sup_{y\in[0,1]} E_y[f(\cdot,y)].$$

Zusammen ergibt dies die Behauptung. □

Beispiel 7.41 Ist $Q_x = Q_y = T_N$ die Trapezsumme zur Schrittweite $h = 1/N$, dann gilt gerade

$$E_x[g] = E_y[g] = \frac{h^2}{12} \|g''\|_{C([0,1])}.$$

Die entsprechende Tensorprodukt-Trapezsumme genügt somit der Fehlerabschätzung

$$\begin{aligned} \big|(I\otimes I)[f] - (T_N \otimes T_N)[f]\big| &\le \sup_{y\in[0,1]} \frac{h^2}{12} \left\| \frac{\partial^2 f}{\partial x^2}(\cdot,y) \right\|_{C([0,1])} + \sup_{x\in[0,1]} \frac{h^2}{12} \left\| \frac{\partial^2 f}{\partial y^2}(x,\cdot) \right\|_{C([0,1])} \\ &\le \frac{h^2}{6} \max\left\{ \left\| \frac{\partial^2 f}{\partial x^2} \right\|_{C(\square)}, \left\| \frac{\partial^2 f}{\partial y^2} \right\|_{C(\square)} \right\}. \end{aligned}$$

Die Fehlerordnung ist also wie im Eindimensionalen $\mathcal{O}(h^2)$. Die Tensorprodukt-Trapezsumme benötigt allerdings $(N+1)^2$ Funktionsauswertungen, im Gegensatz zu $N+1$ Funktionsauswertungen im Fall der eindimensionalen Trapezsumme. △

Bemerkung: Die Tensorprodukt-Konstruktion kann leicht für Integrale über dem d-dimensionalen Einheitswürfel $[0,1]^d$ verallgemeinert werden: die Tensorprodukt-Formel lautet $Q_{x_1} \otimes Q_{x_2} \otimes \cdots \otimes Q_{x_d}$. Satz 7.40 besagt nun, dass sich der Quadraturfehler durch die Summe der eindimensionalen Quadraturfehler abschätzen lässt, dies bedeutet

$$\begin{aligned} &\big|(I\otimes I\otimes\cdots\otimes I)[f] - (Q_{x_1}\otimes Q_{x_2}\otimes\cdots\otimes Q_{x_d})[f]\big| \\ &\quad\le \sum_{i=1}^{d} \sup_{x_1,\ldots,x_{i-1},x_{i+1},\ldots,x_d\in[0,1]} E_{x_i}[f(x_1,\ldots,x_{i-1},\cdot,x_{i+1},\ldots,x_d)]. \end{aligned}$$

Beachte: Die d-dimensionale Tensorprodukt-Trapezsumme besitzt die feste Fehlerordnung $\mathcal{O}(h^2)$. Der Aufwand wächst allerdings *exponentiell* mit der Dimension: $\mathcal{O}(N^d)$. Daher spricht man auch vom *Fluch der Dimension*.

7.12 Monte-Carlo-Quadratur

Im folgenden stehe uns ein Zufallsgenerator zur Verfügung, der eine Folge

$$x_1, x_2, x_3, \ldots$$

auf $[0,1]$ gleichverteilter und stochastisch unabhängiger Zufallsvariablen liefert.
Nun gilt für eine gleichverteilte Zufallsgröße X die Beziehung

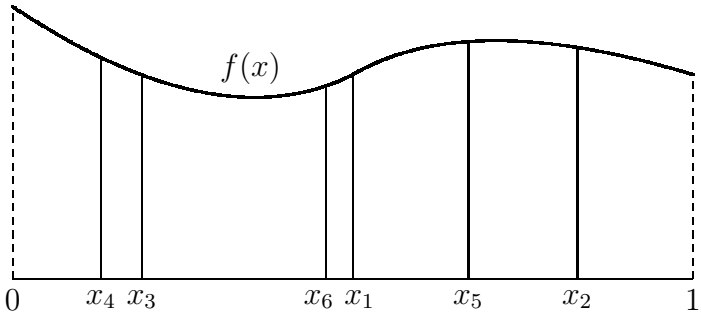
$$\mathbb{E}\big(f(X)\big) = \int_0^1 f(x)\,\mathrm{d}x =: I[f].$$

Daher kann man das Integral $I[f]$ wie folgt approximieren: wähle zufällig N Stützstellen $x_1, x_2, \ldots, x_N$ aus und bestimme den Mittelwert

$$Q_N := \frac{1}{N}\sum_{i=1}^{N} f(x_i).$$

Im statistischen Mittel liefert Q_N das richtige Ergebnis, denn es gilt

$$\mathbb{E}(Q_N) = \frac{1}{N}\sum_{i=1}^{N}\mathbb{E}\big(f(X_i)\big) = \frac{1}{N}\sum_{i=1}^{N}\int_0^1 f(x)\,\mathrm{d}x = \int_0^1 f(x)\,\mathrm{d}x.$$



Natürlich kann der Fehler beliebig groß sein. Jedoch wird mit wachsendem N die mittlere Abweichung $\mathbb{E}\big(|I[f]-Q_N|\big)$ klein.

Satz 7.42 Sei $g(x) := I[f] - f(x)$ und

$$v_f := \int_0^1 g^2(x)\,\mathrm{d}x.$$

Dann ist die mittlere Abweichung beschränkt durch

$$\mathbb{E}\big(|I[f]-Q_N|\big) \le \sqrt{\frac{v_f}{N}}.$$

Beweis. Wir bestimmen zunächst die mittlere quadratische Abweichung

$$\mathbb{V}(Q_N) = \mathbb{E}\big((I[f]-Q_N)^2\big) = \mathbb{E}(Q_N^2) - I[f]^2.$$

Es gilt

$$\begin{aligned}
\mathbb{E}\big((I[f]-Q_N)^2\big) &= \mathbb{E}\left[\left(I[f]-\frac{1}{N}\sum_{i=1}^{N} f(X_i)\right)^2\right] \\
&= \mathbb{E}\left[\left(\frac{1}{N}\sum_{i=1}^{N}\big(\underbrace{I[f]-f(X_i)}_{=g(X_i)}\big)\right)^2\right] \\
&= \frac{1}{N^2}\int_0^1\int_0^1\cdots\int_0^1\left(\sum_{i=1}^{N} g(x_i)\right)^2 \mathrm{d}x_N\cdots\mathrm{d}x_2\,\mathrm{d}x_1 \\
&= \frac{1}{N^2}\sum_{i=1}^{N}\int_0^1\int_0^1\cdots\int_0^1 g^2(x_i)\,\mathrm{d}x_N\cdots\mathrm{d}x_2\,\mathrm{d}x_1 \\
&\qquad + \frac{1}{N^2}\sum_{i\neq j}\int_0^1\int_0^1\cdots\int_0^1 g(x_i)g(x_j)\,\mathrm{d}x_N\cdots\mathrm{d}x_2\,\mathrm{d}x_1 \\
&= \frac{1}{N^2}\sum_{i=1}^{N}\underbrace{\int_0^1 g^2(x_i)\,\mathrm{d}x_i}_{=v_f} + \frac{1}{N^2}\sum_{i\neq j}\left(\underbrace{\int_0^1 g^2(x_i)\,\mathrm{d}x_i}_{=0}\right)\left(\underbrace{\int_0^1 g^2(x_j)\,\mathrm{d}x_j}_{=0}\right) \\
&= \frac{v_f}{N}.
\end{aligned}$$

Aus

$$\mathbb{E}^2\big(|I[f]-Q_N|\big) \le \mathbb{E}\big((I[f]-Q_N)^2\big)$$

folgt schließlich die Behauptung. □

Da $\sqrt{v_f}$ eine Konstante ist, die nur von f abhängt, ist $\sqrt{N}$ die Konvergenzrate, das heißt, der Erwartungswert des Fehlers konvergiert wie $\mathcal{O}(\sqrt{N})$. Der Aufwand der Monte-Carlo-Quadratur besteht in $\mathcal{O}(N)$ arithmetischen Operationen.
Zwar ist die Konvergenzrate $\mathcal{O}(\sqrt{N})$ nicht sehr hoch, allerdings bleibt sie auch im höherdimensionalen Fall erhalten, ohne dass der Aufwand mit der Dimension d exponentiell wächst. Die Monte-Carlo-Quadratur zur Berechnung von

$$I[f] = \int_{[0,1]^d} f(\mathbf{x})\,\mathrm{d}\mathbf{x}$$

besitzt den Aufwand $\mathcal{O}(N\cdot d)$:

Algorithmus 7.43 (Monte-Carlo-Quadratur)
input: Funktion $f\in C([0,1]^d)$ und Parameter N
output: Approximation $Q_N \approx I[f]$
① Initialisierung: setze $Q_N := 0$
② für alle $i = 1,2,\ldots,N$ erzeuge den Zufallsvektor $\mathbf{x}_i\in[0,1]^d$ durch d-maliges Auswerten des Zufallsgenerators und datiere auf $Q_N := Q_N + f(\mathbf{x}_i)/N$

7.13 Dünngitter-Quadratur

Abschließend wollen wir eine deterministische Methode betrachten, mit der Fluch der Dimension besiegt werden kann.

Hierzu interpretieren wir wie in Abschnitt 7.5 die Tensorprodukt-Trapezsumme $(T_j \otimes T_j)[f]$ zur Schrittweite $h_x = h_y = 2^{-j}$ wieder als exakte Quadratur $(I \otimes I)[f_j]$ des stückweise bilinearen Interpolanden

$$f_j(x,y) = \sum_{k_x \in \triangle_j} \sum_{k_y \in \triangle_j} \alpha_{j,k_x,k_y} \varphi_{j,k_x}(x) \varphi_{j,k_y}(y), \quad \alpha_{j,k_x,k_y} = f(2^{-j}k_x, 2^{-j}k_y).$$

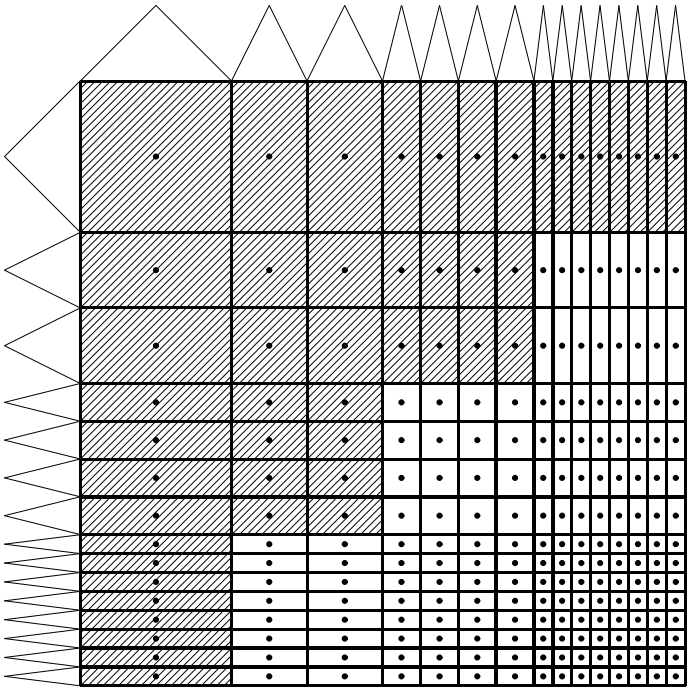
Ersetzen wir die nodale Basis durch die hierarchische Basis, dann folgt

$$f_j(x,y) = \sum_{\ell_x,\ell_y \le j} \sum_{k_x \in \nabla_{\ell_x}} \sum_{k_y \in \nabla_{\ell_y}} \beta_{(\ell_x,k_x),(\ell_y,k_y)} \varphi_{\ell_x,k_x}(x) \varphi_{\ell_y,k_y}(y),$$

wobei die hierarchischen Überschüsse $\beta_{(\ell_x,k_x),(\ell_y,k_y)}$ durch Anwenden des Differenzensterns

$$\mathcal{S}_{(h_x,h_y)} := \begin{bmatrix} 1/4 & -1/2 & 1/4 \\ -1/2 & 1 & -1/2 \\ 1/4 & -1/2 & 1/4 \end{bmatrix}_\star$$

bezüglich des Gitters zu den Schrittweiten $(h_x, h_y) = (2^{-\ell_x}, 2^{-\ell_y})$ auf f entstehen.



Wir wollen nun zeigen, dass wir für $\ell_x + \ell_y > j$ alle Koeffizienten vernachlässigen dürfen, falls f genügend glatt ist.

Lemma 7.44 Existiert die Ableitung $\partial^4 f(x,y)/(\partial x^2 \partial y^2)$ für alle $x, y \in \square$, dann gilt für alle $\ell_x, \ell_y > 0$ die Abschätzung

$$\left|\beta_{(\ell_x,k_x),(\ell_y,k_y)}\right| = \left|\mathcal{S}_{(2^{-\ell_x},2^{-\ell_y})} f(2^{-\ell_x}k_x, 2^{-\ell_y}k_y)\right| \leq 4^{-(\ell_x+\ell_y-1)} \left\| \frac{\partial^4 f}{\partial x^2 \partial y^2} \right\|_{C(\square)}.$$

Beweis. Taylor-Entwicklung zunächst in der ersten und danach in der zweiten Variable liefert

$$\begin{aligned}
&f(x+h, y \pm k) + f(x-h, y \pm k) \\
&\quad = 2f(x, y \pm k) + \int_0^h s\big\{f_{xx}(x+s, y \pm k) - f_{xx}(x-s, y \pm k)\big\}\,\mathrm{d}s \\
&\quad = 2f(x, y \pm k) + \int_0^h s\big\{f_{xx}(x+s, y) - f_{xx}(x-s, y)\big\}\,\mathrm{d}s \\
&\qquad \pm \int_0^h sk\big\{f_{xxy}(x+s, y) - f_{xxy}(x-s, y)\big\}\,\mathrm{d}s \\
&\qquad \pm \int_0^h \int_0^k st\big\{f_{xxyy}(x+s, y \pm t) - f_{xxyy}(x-s, y \pm t)\big\}\,\mathrm{d}t\,\mathrm{d}s.
\end{aligned}$$

Mit Hilfe der Gleichung

$$f(x+h, y) + f(x-h, y) = 2f(x, y) + \int_0^h s\big\{f_{xx}(x+s, y) - f_{xx}(x-s, y)\big\}\,\mathrm{d}s$$

folgt

$$\begin{aligned}
&f(x+h, y+k) + f(x-h, y+k) + f(x+h, y-k) + f(x-h, y-k) \\
&\quad = 2f(x, y+k) + 2f(x, y-k) + 2f(x+h, y) + 2f(x-h, y) - 4f(x, y) \\
&\qquad + \int_0^h \int_0^k st\big\{f_{xxyy}(x+s, y+t) + f_{xxyy}(x-s, y-t) \\
&\qquad\qquad - f_{xxyy}(x+s, y-t) - f_{xxyy}(x-s, y+t)\big\}\,\mathrm{d}t\,\mathrm{d}s.
\end{aligned}$$

Hieraus ergibt sich die Behauptung. □

Satz 7.45 Es sei

$$\widehat{f}_j(x,y) := \sum_{\ell_x+\ell_y \leq j} \sum_{k_x \in \nabla_{\ell_x}} \sum_{k_y \in \nabla_{\ell_y}} \beta_{(\ell_x,k_x),(\ell_y,k_y)} \varphi_{\ell_x,k_x}(x) \varphi_{\ell_y,k_y}(y).$$

Existiert die Ableitung $\partial^4 f(x,y)/(\partial x^2 \partial y^2)$ für alle $x, y \in \square$, dann gilt die Abschätzung

$$\left|(I \otimes I)[f_j - \widehat{f}_j]\right| \leq 4^{-j} \frac{j}{8} \left\| \frac{\partial^4 f}{\partial x^2 \partial y^2} \right\|_{C(\square)}.$$

Beweis. Aus Lemma 7.44 erhalten wir die Abschätzung

$$\begin{aligned}\big|(I\otimes I)[f_j-\widehat{f}_j]\big| &= \left|\sum_{\substack{\ell_x,\ell_y\le j\\ \ell_x+\ell_y>j}}\sum_{k_x\in\nabla_{\ell_x}}\sum_{k_y\in\nabla_{\ell_y}}\beta_{(\ell_x,k_x),(\ell_y,k_y)}\int_0^1\int_0^1\varphi_{\ell_x,k_x}(x)\varphi_{\ell_y,k_y}(y)\,\mathrm{d}y\,\mathrm{d}x\right|\\ &\le \sum_{\substack{\ell_x,\ell_y\le j\\ \ell_x+\ell_y>j}}\sum_{k_x\in\nabla_{\ell_x}}\sum_{k_y\in\nabla_{\ell_y}}2^{-(2\ell_x+1)}2^{-(2\ell_y+1)}2^{1-\ell_x}2^{1-\ell_y}\left\|\frac{\partial^4 f}{\partial x^2\partial y^2}\right\|_{C(\square)}.\end{aligned}$$

Nun gilt $|\nabla_\ell| = 2^{\ell-1}$ und daher

$$\big|(I\otimes I)[f_j-\widehat{f}_j]\big| \le \sum_{\substack{\ell_x,\ell_y\le j\\ \ell_x+\ell_y>j}} 4^{-(\ell_x+\ell_y+1)}\left\|\frac{\partial^4 f}{\partial x^2\partial y^2}\right\|_{C(\square)}.$$

Setzen wir $j<\ell:=\ell_x+\ell_y\le 2j$ und beachten, dass es stets $2j-\ell$ verschiedene Kombinationen mit $\ell=\ell_x+\ell_y$ gibt, dann folgt schließlich

$$\big|(I\otimes I)[f_j-\widehat{f}_j]\big| \le \sum_{\ell=j+1}^{2j} 4^{-(\ell+1)}(2j-\ell)\left\|\frac{\partial^4 f}{\partial x^2\partial y^2}\right\|_{C(\square)} \le 4^{-(j+1)}\frac{j}{2}\left\|\frac{\partial^4 f}{\partial x^2\partial y^2}\right\|_{C(\square)}.$$

□

Proposition 7.46 Existiert die Ableitung $\partial^4 f(x,y)/(\partial x^2\partial y^2)$ für alle $x,y\in\square$, dann genügt die Dünngitter-Quadratur der Abschätzung

$$\big|(I\otimes I)[f-\widehat{f}_j]\big| \le \frac{4^{-j}j}{6}\max\left\{\left\|\frac{\partial^2 f}{\partial x^2}\right\|_{C(\square)},\left\|\frac{\partial^2 f}{\partial y^2}\right\|_{C(\square)},\left\|\frac{\partial^4 f}{\partial x^2\partial y^2}\right\|_{C(\square)}\right\}.$$

Beweis. Gemäß Beispiel 7.41 gilt

$$\big|(I\otimes I)[f-f_j]\big| = \big|(I\otimes I)[f]-(T_j\otimes T_j)[f]\big| \le \frac{4^{-j}}{3}\max\left\{\left\|\frac{\partial^2 f}{\partial x^2}\right\|_{C(\square)},\left\|\frac{\partial^2 f}{\partial y^2}\right\|_{C(\square)}\right\}.$$

Diese Abschätzung muss nun nur noch mit Satz 7.45 kombiniert werden. □

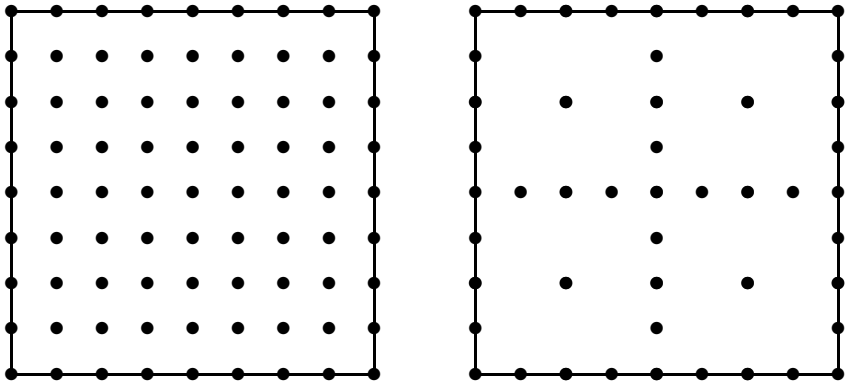
Obwohl die Dünngitter-Quadraturformel fast dieselbe Genauigkeit besitzt wie die Tensorprodukt-Trapezsumme, werden erheblich weniger Quadraturknoten benötigt:

Satz 7.47 Bei der Berechnung von $(I\otimes I)[\widehat{f}_j]$ werden nur $\mathcal{O}(2^j j)$ Funktionsauswertungen benötigt.

Beweis. Pro Koeffizient werden höchstens 9 Funktionsauswertungen ausgeführt. Aus

$$\sum_{\ell_x+\ell_y\le j}\sum_{k_x\in\nabla_{\ell_x}}\sum_{k_y\in\nabla_{\ell_y}}1 \le \sum_{\ell_x+\ell_y\le j}2^{\ell_x}2^{\ell_y} \le \sum_{\ell=1}^{j}2^\ell j = 2^{j+1}j$$

folgt das Behauptete. □



volles Gitter dünnes Gitter

Bemerkung: Die Dünne-Gitter-Konstruktion kann analog auf den d-dimensionalen Würfel $[0,1]^d$ verallgemeinert werden. Die Genauigkeit der Dünngitter-Quadratur ist $\mathcal{O}(4^{-j} j^{d-1})$, während der Aufwand im wesentlichen dem der Trapezsumme entspricht: $\mathcal{O}(2^j j^{d-1})$.

8. Gewöhnliche Differentialgleichungen

8.1 Einleitung

Viele Aufgabenstellungen in den Naturwissenschaften führen auf *gewöhnliche Differentialgleichungen*. Im einfachsten Fall ist dabei eine differenzierbare Funktion $y = y(x)$ gesucht, deren Ableitung $y'(x)$ einer Gleichung der Form $y'(x) = f\big(x, y(x)\big)$ genügt, kurz

$$y' = f(x, y). \tag{8.1}$$

Im allgemeinen besitzt (8.1) unendlich viele Lösungen y, weshalb *Anfangsbedingungen* vorgegeben werden:

$$y(x_0) = y_0. \tag{8.2}$$

Beispiel 8.1 Für jedes $c \in \mathbb{R}$ ist die Funktion $y(x) = c\exp(x)$ eine Lösung der gewöhnlichen Differentialgleichung $y' = y$. Die Anfangsbedingung $y(0) = c_0$ sichert die Eindeutigkeit. △

Allgemeiner betrachtet man Systeme von n gewöhnlichen Differentialgleichungen

$$\begin{aligned}
y_1'(x) &= f_1\big(x, y_1(x), y_2(x), \ldots, y_n(x)\big)\\
y_2'(x) &= f_2\big(x, y_1(x), y_2(x), \ldots, y_n(x)\big)\\
&\vdots\\
y_n'(x) &= f_2\big(x, y_1(x), y_2(x), \ldots, y_n(x)\big)
\end{aligned}$$

für n gesuchte Funktionen $y_i : \mathbb{R} \to \mathbb{R}$, $i = 1, 2, \ldots, n$. Solche Systeme schreibt man analog zu (8.1) vektoriell als

$$\mathbf{y}' = \mathbf{f}(x, \mathbf{y}), \quad \mathbf{y}' := \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}, \quad f(x, \mathbf{y}) := \begin{bmatrix} f_1\big(x, y_1, y_2, \ldots, y_n\big) \\ f_2\big(x, y_1, y_2, \ldots, y_n\big) \\ \vdots \\ f_n\big(x, y_1, y_2, \ldots, y_n\big) \end{bmatrix}. \tag{8.3}$$

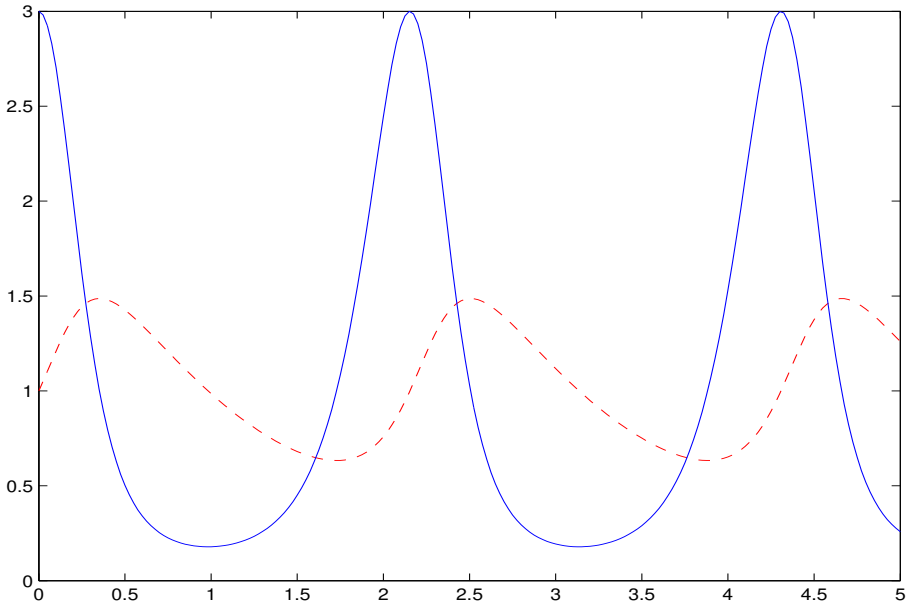
Die Anfangsbedingung (8.2) lautet nun

$$\mathbf{y}(x_0) = \mathbf{y}_0 = \begin{bmatrix} y_{0,1} \\ y_{0,2} \\ \vdots \\ y_{0,n} \end{bmatrix}.$$

Beispiel 8.2 (Räuber-Beute-Modell von Lotka-Volterra) Seien $y_1(t)$ bzw. $y_2(t)$ die Population von Beutetieren bzw. Raubtieren zum Zeitpunkt t, wobei sich die Raubtiere ausschließlich von den Beutetieren ernähren mögen. Ausgehend vom Gleichgewichtszustand $y_1(t) \equiv y_2(t) \equiv 1$ erhalten wir die Gleichungen

$$\begin{aligned} y_1'(x) &= \alpha y_1(t)\big(1 - y_2(t)\big), \\ y_2'(x) &= y_2(t)\big(y_1(t) - 1\big). \end{aligned}$$

Gegeben sei die Population zum Zeitpunkt $t = 0$. Das nachfolgende Bild gibt den Verlauf der beiden Funktionen $y_1(t)$ und $y_2(t)$ wieder:



△

Neben *gewöhnlichen Differentialgleichungen 1. Ordnung* (8.1) und (8.3), in denen nur Ableitungen erster Ordnung der unbekannten Funktion vorkommen, gibt es *gewöhnliche Differentialgleichungen m-ter Ordnung* der Form

$$y^{(m)}(x) = f\big(x, y(x), y'(x), \ldots, y^{(m-1)}(x)\big). \tag{8.4}$$

Man kann diese jedoch mit Hilfe von Hilfsfunktionen

$$\begin{aligned} z_1(x) &:= y(x) \\ z_2(x) &:= y'(x) \\ &\vdots \\ z_m(x) &:= y^{(m-1)}(x) \end{aligned}$$

stets in ein äquivalentes System gewöhnlicher Differentialgleichungen 1. Ordnung transformieren:

$$\mathbf{z}' := \begin{bmatrix} z_1' \\ z_2' \\ \vdots \\ z_{m-1}' \\ z_m' \end{bmatrix} = \begin{bmatrix} z_2 \\ z_3 \\ \vdots \\ z_m \\ f(x, z_1, z_2, \ldots, z_m) \end{bmatrix}.$$

Unter Anfangswertproblem für die Differentialgleichungen m-ter Ordnung versteht man die Aufgabe, eine m-mal stetig differenzierbare Funktion $y(x)$ zu finden, so dass (8.4) und

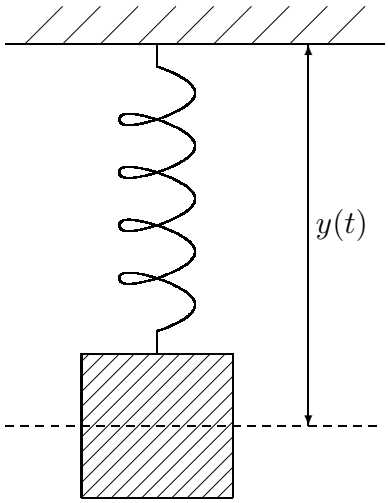
$$y^{(i)}(x_0) = y_{0,i}, \quad i = 0, 1, \ldots, m-1$$

erfüllt ist.

Beispiel 8.3 (elastische Schwingung) Eine Federpendel sei am oberen Ende fest eingespannt, während sich am unteren Ende ein Körper der Masse m befestigt ist. Bezüglich der Auslenkung $y(t)$ führt die Aufstellung des Kräftegleichgewichts auf die Gleichung

$$\underbrace{my''(t)}_{\text{Trägheitskraft}} + \underbrace{ry'(t)}_{\text{Reibungskraft}} + \underbrace{D\big(y(t)-\ell\big)}_{\text{Dehnungskraft}} = \underbrace{g(t)}_{\text{äußere Kraft}}.$$

Mit vorgegebenen Werten $y(0)$ und $y'(0)$, die der Lage und der Geschwindigkeit im Zeitpunkt $t = 0$ entsprechen, gelangen wir zu einem Anfangswertproblem.



△

8.2 Theoretische Grundlagen

Existenz und Eindeutigkeit der Lösung von Anfangswertproblemen sichert der folgende Satz:

Satz 8.4 (Picard-Lindelöf) Auf dem Streifen $S := \{(x, \mathbf{y}) : a \le x \le b, \mathbf{y} \in \mathbb{R}^n\}$ sei die Funktion $\mathbf{f} : S \to \mathbb{R}^n$ stetig und genüge der Lipschitz-Bedingung

$$\|\mathbf{f}(x, \mathbf{y}) - \mathbf{f}(x, \mathbf{z})\| \le L\|\mathbf{y} - \mathbf{z}\|$$

für alle $(x, \mathbf{y}), (x, \mathbf{z}) \in S$. Dann ist für jedes $x_0 \in [a, b]$ das Anfangswertproblem

$$\mathbf{y}' = \mathbf{f}(x, \mathbf{y}), \quad \mathbf{y}(x_0) = \mathbf{y}_0$$

eindeutig lösbar.

Beweis. Für $\mathbf{u} \in C([a,b],\mathbb{R}^n)$ sei

$$\Phi\big(\mathbf{u}(x)\big) := \mathbf{y}_0 + \int_{x_0}^{x} \mathbf{f}\big(t,\mathbf{u}(t)\big)\,\mathrm{d}t, \quad x \in [a,b].$$

Jeder Fixpunkt $\mathbf{y}(x) = \Phi(\mathbf{y}(x))$ löst das Anfangswertproblem wegen

$$\mathbf{y}'(x) = \Phi'\big(\mathbf{y}(x)\big) = \mathbf{f}\big(x,\mathbf{y}(x)\big), \qquad \mathbf{y}(x_0) = \Phi\big(\mathbf{y}(x_0)\big) = \mathbf{y}_0.$$

Umgekehrt ist jede Lösung $\mathbf{y}$ des Anfangswertproblems auch ein Fixpunkt von Φ, denn es gilt

$$\mathbf{y}(x) = \mathbf{y}(x_0) + \int_{x_0}^{x} \mathbf{y}'(x)\,\mathrm{d}x = \mathbf{y}_0 + \int_{x_0}^{x} \mathbf{f}\big(t,\mathbf{y}(t)\big)\,\mathrm{d}t = \Phi(\mathbf{y}(x)).$$

Um den Banachschen Fixpunktssatz anwenden zu können, führen wir auf $C([a,b],\mathbb{R}^n)$ eine geeignete Norm

$$|\!|\!|\mathbf{u}|\!|\!| := \sup_{x\in[a,b]} \big\{e^{-2Lx}\|\mathbf{u}(x)\|\big\}$$

ein. Für diese gilt

$$\begin{aligned}
e^{-2Lx}\big\|\Phi\big(\mathbf{u}(x)\big) - \Phi\big(\mathbf{v}(x)\big)\big\| &= e^{-2Lx}\left\| \int_{x_0}^{x} \mathbf{f}\big(t,\mathbf{u}(t)\big) - \mathbf{f}\big(t,\mathbf{v}(t)\big)\,\mathrm{d}t\right\| \\
&\le e^{-2Lx} \int_{x_0}^{x} \big\|\mathbf{f}\big(t,\mathbf{u}(t)\big) - \mathbf{f}\big(t,\mathbf{v}(t)\big)\big\|\,\mathrm{d}t \\
&\le Le^{-2Lx} \int_{x_0}^{x} \|\mathbf{u}(t) - \mathbf{v}(t)\|\,\mathrm{d}t \\
&= Le^{-2Lx} \int_{x_0}^{x} e^{2Lt} \underbrace{e^{-2Lt}\|\mathbf{u}(t) - \mathbf{v}(t)\|}_{\le |\!|\!|\mathbf{u}-\mathbf{v}|\!|\!|}\,\mathrm{d}t \\
&\le Le^{-2Lx}|\!|\!|\mathbf{u} - \mathbf{v}|\!|\!| \underbrace{\int_{x_0}^{x} e^{2Lt}\,\mathrm{d}t}_{=(e^{2Lx}-e^{2Lx_0})/(2L)} \\
&\le \frac{1}{2}|\!|\!|\mathbf{u} - \mathbf{v}|\!|\!|,
\end{aligned}$$

das heißt, Φ ist eine Kontraktion:

$$|\!|\!|\Phi(\mathbf{u}) - \Phi(\mathbf{v})|\!|\!| \le \frac{1}{2}|\!|\!|\mathbf{u} - \mathbf{v}|\!|\!|.$$

Der Banachsche Fixpunktsatz liefert uns die Existenz und Eindeutigkeit eines Fixpunktes von Φ und damit vom Anfangswertproblem. □

Für die Stabilitätsuntersuchung numerischer Verfahren wesentlich ist das nachfolgende Resultat, das zeigt, dass die Lösung stetig abhängt von den Anfangsdaten.

Satz 8.5 (Grönwall) Auf dem Streifen $S := \{(x,\mathbf{y}) : a \le x \le b,\ \mathbf{y} \in \mathbb{R}^n\}$ sei die Funktion $\mathbf{f} : S \to \mathbb{R}^n$ stetig und genüge der Lipschitz-Bedingung

$$\|\mathbf{f}(x,\mathbf{y}) - \mathbf{f}(x,\mathbf{z})\| \le L\|\mathbf{y} - \mathbf{z}\|$$

für alle $(x, \mathbf{y}), (x, \mathbf{z}) \in S$. Dann erfüllt für jedes $x_0 \in [a, b]$ die Lösung des Anfangswertproblems

$$\mathbf{y}' = \mathbf{f}(x, \mathbf{y}), \quad \mathbf{y}(x_0, \mathbf{u}) = \mathbf{u}$$

die Abschätzung

$$\|\mathbf{y}(\mathbf{u}) - \mathbf{y}(\mathbf{v})\| \le e^{L|x-x_0|}\|\mathbf{u} - \mathbf{v}\|.$$

Beweis. Aus der Integraldarstellung

$$\mathbf{y}(x, \mathbf{u}) = \mathbf{u} + \int_{x_0}^{x} \mathbf{f}\big(t, \mathbf{y}(t, \mathbf{u})\big)\,\mathrm{d}t, \quad x \in [a, b]$$

folgt

$$\begin{aligned}\|\mathbf{y}(x, \mathbf{u}) - \mathbf{y}(x, \mathbf{v})\| &= \left\|\mathbf{u} - \mathbf{v} + \int_{x_0}^{x} \mathbf{f}\big(t, \mathbf{y}(t, \mathbf{u})\big) - \mathbf{f}\big(t, \mathbf{y}(t, \mathbf{v})\big)\,\mathrm{d}t\right\| \\ &\le \|\mathbf{u} - \mathbf{v}\| + \int_{x_0}^{x} \big\|\mathbf{f}\big(t, \mathbf{y}(t, \mathbf{u})\big) - \mathbf{f}\big(t, \mathbf{y}(t, \mathbf{v})\big)\big\|\,\mathrm{d}t \\ &\le \|\mathbf{u} - \mathbf{v}\| + L\int_{x_0}^{x} \|\mathbf{y}(t, \mathbf{u}) - \mathbf{y}(t, \mathbf{v})\|\,\mathrm{d}t.\end{aligned}$$

Setzen wir

$$\Phi(x) := \int_{x_0}^{x} \|\mathbf{y}(t, \mathbf{u}) - \mathbf{y}(t, \mathbf{v})\|\,\mathrm{d}t,$$

dann ergibt sich

$$\Phi'(x) = \|\mathbf{y}(x, \mathbf{u}) - \mathbf{y}(x, \mathbf{v})\|$$

und daher für $x \ge x_0$

$$\Psi(x) := \Phi'(x) - L\Phi(x) \le \|\mathbf{u} - \mathbf{v}\|.$$

Das Anfangswertproblem

$$\Phi'(x) = \Psi(x) + L\Phi(x), \quad \Phi(x_0) = 0$$

besitzt für $x \ge x_0$ die Lösung

$$\Phi(x) = e^{L(x-x_0)} \int_{x_0}^{x} \Psi(t) e^{-L(t-x_0)}\,\mathrm{d}t.$$

Wegen $\Psi(x) \le \|\mathbf{u} - \mathbf{v}\|$ folgt so die Abschätzung

$$0 \le \Phi(x) \le e^{L(x-x_0)}\|\mathbf{u} - \mathbf{v}\| \int_{x_0}^{x} e^{-L(t-x_0)}\,\mathrm{d}t = \frac{1}{L}\|\mathbf{u} - \mathbf{v}\|\big(e^{L(x-x_0)} - 1\big).$$

Es ergibt sich schließlich das verlangte Resultat für $x \ge x_0$:

$$\|\mathbf{y}(x, \mathbf{u}) - \mathbf{y}(x, \mathbf{v})\| = \Phi'(x) = \Psi(x) + L\Phi(x) \le \|\mathbf{u} - \mathbf{v}\| e^{L(x-x_0)}.$$

Ähnlich geht man im Fall $x < x_0$ vor. □

8.3 Einschrittverfahren

Zu einer ersten numerischen Methode zur Lösung des Anfangswertproblems

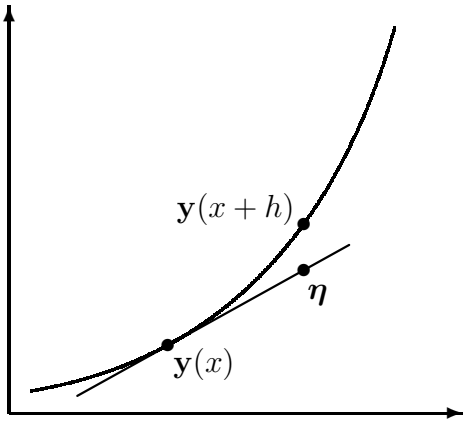
$$\mathbf{y}' = \mathbf{f}(x, \mathbf{y}), \quad \mathbf{y}(x_0) = \mathbf{y}_0 \tag{8.5}$$

kommt man durch eine einfache Überlegung: Da $\mathbf{f}\big(x, \mathbf{y}(x)\big)$ gerade die Steigung $\mathbf{y}'(x)$ der exakten Lösung $\mathbf{y}(x)$ von (8.5) ist, gilt näherungsweise für $h > 0$

$$\frac{\mathbf{y}(x+h) - \mathbf{y}(x)}{h} \approx \mathbf{y}'(x) = \mathbf{f}(x, \mathbf{y}),$$

das heißt,

$$\mathbf{y}(x+h) \approx \mathbf{y}(x) + h\mathbf{f}(x, \mathbf{y}).$$



Nach Wahl einer Schrittweite $h > 0$ erhält man ausgehend von den gegebenen Anfangswerten x_0, $\mathbf{y}_0 = \mathbf{y}(x_0)$ an den äquidistanen Stützstellen $x_i := x_0 + hi$, $i = 1, 2, \ldots$, Näherungswerte $\boldsymbol{\eta}_i$ für die Werte $\mathbf{y}_i := \mathbf{y}(x_i)$ der exakten Lösung $\mathbf{y}(x)$:

Algorithmus 8.6 (Eulersches Polygonzug-Verfahren)
input: Funktion $\mathbf{f} \in C([a, b] \times \mathbb{R}^n)$, Anfangswerte x_0, $\mathbf{y}_0$ und Schrittweite h
output: Approximation $\{\boldsymbol{\eta}_i\}$
① Initialisierung: setze $\boldsymbol{\eta}_0 := \mathbf{y}_0$
② für alle $i = 0, 1, \ldots$ berechne

$$\boldsymbol{\eta}_{i+1} := \boldsymbol{\eta}_i + h\mathbf{f}(x_i, \boldsymbol{\eta}_i), \qquad x_{i+1} := x_i + h$$

Definition 8.7 Ein Verfahren der Form

$$\boldsymbol{\eta}_{i+1} := \boldsymbol{\eta}_i + h\Phi(x_i, \boldsymbol{\eta}_i, \boldsymbol{\eta}_{i+1}, h), \qquad x_{i+1} := x_i + h \tag{8.6}$$

heißt **Einschrittverfahren**. Die Funktion Φ heißt dabei **Inkrementfunktion**. Falls Φ nicht von $\boldsymbol{\eta}_{i+1}$ abhängt, so sprechen wir von einem **expliziten**, andernfalls von einem **impliziten** Verfahren.

Beispiel 8.8 Das Eulersche Polygonzug-Verfahren ist ein explizites Verfahren, weshalb es auch *explizites Euler-Verfahren* genannt wird. Das *implizite Euler-Verfahren* erhält man, wenn man nicht wie beim expliziten Euler-Verfahren den rechtsseitigen Differenzenquotienten nimmt, sondern den linksseitigen:

$$\frac{\mathbf{y}(x+h)-\mathbf{y}(x)}{h} \approx \mathbf{y}'(x+h) = \mathbf{f}(x+h,\mathbf{y}),$$

dies bedeutet

$$\boldsymbol{\eta}_{i+1} = \boldsymbol{\eta}_i + h\mathbf{f}(x_{i+1},\boldsymbol{\eta}_{i+1}).$$

△

Bemerkung: Bei expliziten Verfahren lässt sich $\boldsymbol{\eta}_{i+1}$ direkt durch Einsetzen vorher berechneter Größen ermitteln. Bei impliziten Verfahren muss hingegen ein im allgemeinen nichtlineares Gleichungssystem gelöst werden.

8.4 Konsistenz

Um die Genauigkeit von Einschrittverfahren zu beurteilen, vergleichen wir die approximierte Lösung $\boldsymbol{\eta}_{i+1}$ aus (8.6) mit der lokal exakten Lösung $\mathbf{z}(x_{i+1})$ aus

$$\mathbf{z}' = \mathbf{f}(x,\mathbf{z}), \quad \mathbf{z}(x_i) = \boldsymbol{\eta}_i. \tag{8.7}$$

Die *Abweichung* von der lokal exakten Lösung wird beschrieben durch

$$h\boldsymbol{\tau}(x_i,\boldsymbol{\eta}_i,h) = \mathbf{z}(x_{i+1}) - \boldsymbol{\eta}_{i+1} = \mathbf{z}(x_{i+1}) - \boldsymbol{\eta}_i - h\Phi(x_i,\boldsymbol{\eta}_i,\boldsymbol{\eta}_{i+1},h).$$

Definition 8.9 Der **lokale Diskretisierungsfehler** des Einschrittverfahrens zur Inkrementfunktion Φ ist definiert als

$$\boldsymbol{\tau}(x_i,\boldsymbol{\eta}_i,h) = \frac{\mathbf{z}(x_{i+1})-\boldsymbol{\eta}_i}{h} - \Phi(x_i,\boldsymbol{\eta}_i,\boldsymbol{\eta}_{i+1},h),$$

wobei $\mathbf{z}$ die lokal exakte Lösung (8.7) bezeichnet. Das Einschrittverfahren heißt **konsistent** von der Ordnung p, falls $\|\boldsymbol{\tau}(x_i,\boldsymbol{\eta}_i,h)\| = \mathcal{O}(h^p)$.

Bemerkung: Ist die Inkrementfunktion $\Phi(x,\mathbf{y},\mathbf{z},h)$ eines impliziten Verfahren Lipschitzstetig in $\mathbf{z}$ mit Lipschitz-Konstante L, dann genügt es, die exakte Lösung in das Verfahren einzusetzen, um die Konsistenzordnung zu untersuchen. Denn gilt

$$\left\|\widetilde{\boldsymbol{\tau}}(x_i,\boldsymbol{\eta}_i,h)\right\| = \left\|\frac{\mathbf{z}(x_{i+1})-\boldsymbol{\eta}_i}{h} - \Phi\big(x_i,\boldsymbol{\eta}_i,\mathbf{z}(x_{i+1}),h\big)\right\| = \mathcal{O}(h^p),$$

so folgt aufgrund der Lipschitz-Stetigkeit

$$\begin{aligned}\left\|\boldsymbol{\tau}(x_i,\boldsymbol{\eta}_i,h) - \widetilde{\boldsymbol{\tau}}(x_i,\boldsymbol{\eta}_i,h)\right\| &= \left\|\Phi(x_i,\boldsymbol{\eta}_i,\boldsymbol{\eta}_{i+1},h) - \Phi\big(x_i,\boldsymbol{\eta}_i,\mathbf{z}(x_{i+1}),h\big)\right\| \\ &\le L\|\boldsymbol{\eta}_{i+1}-\mathbf{z}(x_{i+1})\| = Lh\|\boldsymbol{\tau}(x_i,\boldsymbol{\eta}_i,h)\|.\end{aligned}$$

Dies impliziert aber

$$\|\boldsymbol{\tau}(x_i,\boldsymbol{\eta}_i,h)\|(1-Lh) = \left\|\widetilde{\boldsymbol{\tau}}(x_i,\boldsymbol{\eta}_i,h)\right\|$$

also auch $\|\boldsymbol{\tau}(x_i,\boldsymbol{\eta}_i,h)\| = \mathcal{O}(h^p)$. Oftmals wird daher der lokale Diskretisierungsfehler auch einfach als $\widetilde{\boldsymbol{\tau}}(x_i,\boldsymbol{\eta}_i,h)$ definiert.

Beispiel 8.10 Die Taylor-Entwicklung liefert für das explizite Euler-Verfahren

$$\frac{\mathbf{z}(x_{i+1}) - \boldsymbol{\eta}_i}{h} = \underbrace{\mathbf{z}'(x_i)}_{=\mathbf{f}(x_i,\mathbf{z}(x_i))} + \underbrace{\boldsymbol{\tau}\big(x_i, \mathbf{z}(x_i), h\big)}_{\|\cdot\|=\mathcal{O}(h)} = \mathbf{f}(x_i, \boldsymbol{\eta}_i) + \boldsymbol{\tau}(x_i, \boldsymbol{\eta}_i, h)$$

mit $\|\boldsymbol{\tau}(x_i, \boldsymbol{\eta}_i, h)\| = \mathcal{O}(h)$. Gleiches gilt auch für das implizite Euler-Verfahren. Folglich ist das Euler-Verfahren konsistent von erster Ordnung. △

Um Verfahren höherer Ordnung zu konstruieren ist es naheliegend, die Taylor-Entwicklung weiterzutreiben. Es gilt

$$\begin{aligned}
\frac{\mathbf{z}(x_{i+1}) - \boldsymbol{\eta}_i}{h} &= \mathbf{z}'(x_i) + \frac{h}{2}\mathbf{z}''(x_i) + \underbrace{\boldsymbol{\tau}\big(x_i, \mathbf{z}(x_i), h\big)}_{\|\cdot\|=\mathcal{O}(h^2)} \\
&= \mathbf{f}\big(x_i, \mathbf{z}(x_i)\big) + \frac{h}{2}\Big\{\mathbf{f}_x\big(x_i, \mathbf{z}(x_i)\big) + \mathbf{f}_{\mathbf{y}}\big(x_i, \mathbf{z}(x_i)\big)\mathbf{z}'(x_i)\Big\} + \boldsymbol{\tau}\big(x_i, \mathbf{z}(x_i), h\big) \\
&= \mathbf{f}(x_i, \boldsymbol{\eta}_i) + \frac{h}{2}\Big\{\mathbf{f}_x(x_i, \boldsymbol{\eta}_i) + \mathbf{f}_{\mathbf{y}}(x_i, \boldsymbol{\eta}_i)\mathbf{f}(x_i, \boldsymbol{\eta}_i)\Big\} + \boldsymbol{\tau}\big(x_i, \boldsymbol{\eta}_i, h\big) \qquad (8.8)
\end{aligned}$$

mit $\|\boldsymbol{\tau}(x_i, \boldsymbol{\eta}_i, h)\| = \mathcal{O}(h^2)$. Die Inkrementfunktion

$$\Phi(x, \mathbf{y}, h) := \mathbf{f}(x, \mathbf{y}) + \frac{h}{2}\Big\{\mathbf{f}_x(x, \mathbf{y}) + \mathbf{f}_{\mathbf{y}}(x, \mathbf{y})\mathbf{f}(x, \mathbf{y})\Big\}$$

definiert demnach ein explizites Verfahren zweiter Ordnung. In der Praxis kann man die auf diese Art gewonnen Verfahren jedoch meist nicht brauchen, weil die Ableitungen von $\mathbf{f}$ explizit eingehen.

Einfachere Verfahren höherer Ordnung erhält man über den Ansatz

$$\Phi(x, \mathbf{y}, h) := a\mathbf{f}(x, \mathbf{y}) + b\mathbf{f}\big(x + hc, \mathbf{y} + hd\mathbf{f}(x, \mathbf{y})\big),$$

indem man die Konstanten a, b, c, d so bestimmt, dass die Taylor-Entwicklung des Fehlers $\boldsymbol{\tau}(x, \mathbf{y}, h)$ mit möglichst hoher Ordnung anfängt. Taylor-Entwicklung liefert

$$\begin{aligned}
\boldsymbol{\tau}(x, \mathbf{y}, h) &= \frac{\mathbf{y}(x + h) - \mathbf{y}(x)}{h} - \Phi(x, \mathbf{y}, h) \\
&= \underbrace{\mathbf{y}'(x_i)}_{=\mathbf{f}(x,\mathbf{y})} + \frac{h}{2}\underbrace{\mathbf{y}''(x_i)}_{=\mathbf{f}_x(x,\mathbf{y})+\mathbf{f}_{\mathbf{y}}(x,\mathbf{y})\mathbf{f}(x,\mathbf{y})} - (a + b)\mathbf{f}(x, \mathbf{y}) \\
&\qquad - hb\big\{c\mathbf{f}_x(x, \mathbf{y}) + d\mathbf{f}_{\mathbf{y}}(x, \mathbf{y})\mathbf{f}(x, \mathbf{y})\big\} + \underbrace{\boldsymbol{\varepsilon}(h)}_{\|\cdot\|=\mathcal{O}(h^2)}.
\end{aligned}$$

Die Fehlerordnung $\|\boldsymbol{\varepsilon}(h)\| = \mathcal{O}(h^2)$ erhält man nun, wenn

$$a + b = 1, \quad bc = \frac{1}{2}, \quad bd = \frac{1}{2}$$

erfüllt ist. Die Lösung $a = b = 1/2$, $c = d = 1$ liefert das *Verfahren von Heun*

$$\Phi(x, \mathbf{y}, h) := \frac{1}{2}\Big\{\mathbf{f}(x, \mathbf{y}) + \mathbf{f}\big(x + h, \mathbf{y} + h\mathbf{f}(x, \mathbf{y})\big)\Big\},$$

die Lösung $a = 0$, $b = 1$, $c = d = 1/2$ ist das *Euler-CollatzVerfahren*

$$\Phi(x, \mathbf{y}, h) := \mathbf{f}\left(x + \frac{h}{2}, \mathbf{y} + \frac{h}{2}\mathbf{f}(x, \mathbf{y})\right).$$

In Abschnitt 8.6 werden wir dieses Konstruktionsprinzip weiter beleuchten.

8.5 Konvergenz von Einschrittverfahren

Wir wollen nun die Konvergenz der Näherungslösung $\boldsymbol{\eta}(x,h)$ für $h \to 0$ untersuchen, wobei gilt $\boldsymbol{\eta}(x_i,h) = \boldsymbol{\eta}_i$. Wir interessieren uns für den *globalen Diskretisierungsfehler*

$$e(x,h) := \|\boldsymbol{\eta}(x,h) - \mathbf{y}(x)\|$$

bei festem x für

$$H_x := \left\{ \frac{x - x_0}{n} : n \in \mathbb{N}_0 \right\} \ni h \to 0.$$

Da $e(x,h)$ wie $\boldsymbol{\eta}(x,h)$ nur für $h \in H_x$ definiert ist, bedeutet dies die Untersuchung der Konvergenz von

$$e(x,h_n), \quad h_n := \frac{x - x_0}{n}, \quad n \to \infty.$$

Definition 8.11 Das Einschrittverfahrens zur Inkrementfunktion Φ heißt **konvergent** von der Ordnung p, falls gilt

$$e(x,h_n) = \mathcal{O}(h_n^p).$$

Satz 8.12 Gegeben sei für $x_0 \in [a,b]$ das Anfangswertproblem

$$\mathbf{y}' = \mathbf{f}(x,\mathbf{y}), \quad \mathbf{y}(x_0) = \mathbf{y}_0$$

mit der exakten Lösung $\mathbf{y}(x)$. Das explizite Einschrittverfahrens zur Inkrementfunktion Φ sei konsistent von der Ordnung p. Ferner sei Φ stetig auf

$$G := \{(x,\mathbf{y},h) : a \le x \le b,\ \mathbf{y} \in \mathbb{R}^n, |h| \le h_0\}$$

und genüge der Lipschitz-Bedingung

$$\|\Phi(x,\mathbf{u},h) - \Phi(x,\mathbf{v},h)\| \le M\|\mathbf{u} - \mathbf{v}\|$$

für alle $(x,\mathbf{u},h), (x,\mathbf{v},h) \in G$. Dann gilt für den globalen Diskretisierungsfehler

$$e(x,h_n) := \|\boldsymbol{\eta}(x,h_n) - \mathbf{y}(x)\| \le \left\{ e(x_0,h_n) + |x - x_0| \sup_{(x,\mathbf{y},h)\in G} \|\boldsymbol{\tau}(x,\mathbf{y},h)\| \right\} e^{M|x-x_0|}$$

für alle $x \in [a,b]$ und $h_n = (x - x_0)/n$, $n \in \mathbb{N}$, mit $|h_n| \le h_0$.

Beweis. Sei $\mathbf{y}_i = \mathbf{y}(x_i)$. Das durch Φ erzeugte Einschrittverfahren liefert Näherungswerte $\boldsymbol{\eta}_i = \boldsymbol{\eta}(x_i,h)$ mit

$$\mathbf{y}_{i+1} - \boldsymbol{\eta}_{i+1} = \mathbf{y}_i - \boldsymbol{\eta}_i + h\big\{\Phi(x,\mathbf{y}_i,h) - \Phi(x,\boldsymbol{\eta}_i,h)\big\} + h\boldsymbol{\tau}(x,\mathbf{y}_i,h).$$

Setzen wir $e_i := \|\mathbf{y}_i - \boldsymbol{\eta}_i\|$ und $\boldsymbol{\tau}_i = \boldsymbol{\tau}(x,\mathbf{y}_i,h)$, dann folgt hieraus aufgrund der Lipschitz-

Bedingung

$$\begin{aligned} e_{i+1} &\le e_i + |h| \underbrace{\|\Phi(x,\mathbf{y}_i,h) - \Phi(x,\boldsymbol{\eta}_i,h)\|}_{\le M\|\mathbf{y}_i-\boldsymbol{\eta}_i\|} + |h|\|\boldsymbol{\tau}_i\| \\ &\le (1+M|h|)e_i + N|h| \quad \text{mit} \quad N := \sup_{(x,\mathbf{y},h)\in G} \|\boldsymbol{\tau}(x,\mathbf{y},h)\|. \end{aligned}$$

Rekursiv erhalten wir demnach

$$\begin{aligned} e_i &\le (1+M|h|)e_{i-1} + N|h| \\ &\le (1+M|h|)^2 e_{i-2} + (1+M|h|)N|h| + N|h| \\ &\;\;\vdots \\ &\le (1+M|h|)^i e_0 + \big\{(1+M|h|)^{i-1} + \cdots + (1+M|h|) + 1\big\}N|h| \\ &= (1+M|h|)^i e_0 + \frac{(1+M|h|)^i - 1}{M|h|} N|h| \\ &\le e^{iM|h|}e_0 + \frac{e^{iM|h|}-1}{M}N \end{aligned}$$

wegen $0 < 1 + M|h| \le e^{M|h|}$ für $M|h| > -1$. Da $iM|h| \ge 0$ ist, gilt die Ungleichung

$$\frac{e^{iM|h|}-1}{M}N \le e^{iM|h|} iN|h|.$$

Diese oben eingesetzt liefert

$$e_i \le e^{iM|h|}\big\{e_0 + iN|h|\big\}, \quad i \in \mathbb{N}_0.$$

Sei nun $x_0 \ne x \in [a,b]$ fest gewählt und $h := h_n = (x - x_0)/n$, $n \in \mathbb{N}$. Dann ist $x_n = x_0 + nh = x$ und es folgt

$$e(x,h_n) = e_n \le e^{M|x-x_0|}\big\{e_0 + |x-x_0|N\big\},$$

das ist die Behauptung. □

Bemerkung: Aus diesem Satz folgt, dass ein konsistentes Verfahren der Ordnung p auch mit der Ordnung p konvergiert, vorausgesetzt die Funktion $\mathbf{f}$ ist hinreichend glatt.

8.6 Runge-Kutta-Verfahren

Runge-Kutta-Verfahren basieren auf der Integraldarstellung der Lösung

$$\mathbf{y}(x_{i+1}) = \mathbf{y}(x_i) + \int_{x_i}^{x_{i+1}} \mathbf{f}\big(t,\mathbf{y}(t)\big)\,\mathrm{d}t.$$

Ein diskretes Verfahren erhalten wir, wenn wir eine Quadraturformel einsetzen

$$\mathbf{y}(x_{i+1}) \approx \mathbf{y}(x_i) + h\sum_{\ell=1}^{m} \gamma_\ell \mathbf{f}\big(x_i + h\xi_\ell, \mathbf{y}(x_i + h\xi_\ell)\big).$$

Da $\mathbf{y}(x_i + h\xi_\ell)$ auch nicht bekannt ist, approximiert man weiter

$$\mathbf{f}\big(x_i + h\xi_\ell, \mathbf{y}(x_i + h\xi_\ell)\big) \approx \mathbf{k}_\ell\big(x_i, \mathbf{y}(x_i), h\big).$$

Definition 8.13 Gegeben seien reelle Gewichte $\alpha_\ell, \beta_{j,\ell}, \gamma_\ell$, $j, \ell = 1, 2, \ldots, m$, wobei $\alpha_1 = 0$. Ein allgemeines ***m*-stufiges Runge-Kutta-Verfahren** besitzt dann die Form: setze $\boldsymbol{\eta}_0 := \mathbf{y}_0$ und berechne für alle $i = 0, 1, 2, \ldots$

$$\begin{aligned} x_{i+1} &:= x_i + h \\ \mathbf{k}_j(x_i, \boldsymbol{\eta}_i, h) &:= \mathbf{f}\left(x_i + h\alpha_j, \boldsymbol{\eta}_i + h \sum_{\ell=1}^{m} \beta_{j,\ell} \mathbf{k}_\ell(x_i, \boldsymbol{\eta}_i, h) \right), \quad j = 1, 2, \ldots, m \\ \boldsymbol{\eta}_{i+1} &:= \boldsymbol{\eta}_i + h \underbrace{\sum_{\ell=1}^{m} \gamma_\ell \mathbf{k}_\ell(x_i, \boldsymbol{\eta}_i, h)}_{=:\Phi(x_i, \boldsymbol{\eta}_i, \boldsymbol{\eta}_{i+1}, h)} . \end{aligned} \tag{8.9}$$

Die Gewichte sind dabei so zu wählen, dass eine möglichst hohe Genauigkeit erreicht wird. Der Übersichtlichkeit halber werden sie üblicherweise im *Butcher-Tableau* erfasst:

$$\begin{array}{c|ccc} \alpha_1 & \beta_{1,1} & \cdots & \beta_{1,m} \\ \alpha_2 & \beta_{2,1} & \cdots & \beta_{2,m} \\ \vdots & \vdots & & \vdots \\ \alpha_m & \beta_{m,1} & \cdots & \beta_{m,m} \\ \hline & \gamma_1 & \cdots & \gamma_m \end{array}$$

Beispiel 8.14 (Butcher-Tableaus)

$$\begin{array}{c|c} 0 & 0 \\ \hline & 1 \end{array}$$

Eulersches Polygonzug-Verfahren

$$\begin{array}{c|cc} 0 & 0 & 0 \\ \frac{1}{2} & \frac{1}{2} & 0 \\ \hline & 0 & 1 \end{array}$$

Euler-Collatz-Verfahren

$$\begin{array}{c|cc} 0 & 0 & 0 \\ 1 & 1 & 0 \\ \hline & \frac{1}{2} & \frac{1}{2} \end{array}$$

Verfahren von Heun

△

Zunächst werden wir explizite Runge-Kutta-Verfahren betrachten. Dann ist $\beta_{j,\ell} = 0$ für alle $\ell \geq i$, das heißt, (8.9) lautet

$$\begin{aligned} \mathbf{k}_1(x, \mathbf{y}, h) &= \mathbf{f}(x, \mathbf{y}) \\ \mathbf{k}_2(x, \mathbf{y}, h) &= \mathbf{f}\big(x + h\alpha_2, \mathbf{y} + h\beta_{2,1}\mathbf{k}_1(x, \mathbf{y}, h)\big) \\ \mathbf{k}_3(x, \mathbf{y}, h) &= \mathbf{f}\big(x + h\alpha_3, \mathbf{y} + h\beta_{3,1}\mathbf{k}_1(x, \mathbf{y}, h) + h\beta_{3,2}\mathbf{k}_2(x, \mathbf{y}, h)\big) \\ &\vdots \\ \mathbf{k}_m(x, \mathbf{y}, h) &= \mathbf{f}\big(x + h\alpha_m, \mathbf{y} + h\beta_{m,1}\mathbf{k}_1(x, \mathbf{y}, h) + \cdots + h\beta_{m,m-1}\mathbf{k}_{m-1}(x, \mathbf{y}, h)\big). \end{aligned}$$

Eine Taylor-Entwicklung der Inkrementfunktion liefert

$$\begin{aligned}\Phi(x,\mathbf{y},h) &= \sum_{i=1}^{m}\gamma_i\mathbf{k}_i(x,\mathbf{y},h)\\ &= \sum_{i=1}^{m}\gamma_i\mathbf{f}\Big(x+h\alpha_i,\mathbf{y}+h\sum_{j=1}^{i-1}\beta_{i,j}\mathbf{k}_j(x,\mathbf{y},h)\Big)\\ &= \sum_{i=1}^{m}\gamma_i\Big(\mathbf{f}(x,\mathbf{y})+h\alpha_i\mathbf{f}_x(x,\mathbf{y})+h\sum_{j=1}^{i-1}\beta_{i,j}\mathbf{f}_{\mathbf{y}}(x,\mathbf{y})\mathbf{f}(x,\mathbf{y})+\cdots\Big).\end{aligned}$$

Vergleicht man dies mit (8.8), so folgt, dass

$$\sum_{i=1}^{m}\gamma_i=1,\qquad \sum_{i=1}^{m}\gamma_i\alpha_i=\frac{1}{2},\qquad \alpha_i=\sum_{j=1}^{i-1}\beta_{i,i},\quad i=1,2,\ldots,m$$

notwending ist, um ein Verfahren von zweiter Ordnung oder höher zu konstruieren. Treibt man die Taylor-Entwicklung weiter, so erhält man zum Beispiel folgende Runge-Kutta-Verfahren vierter Ordnung:

Beispiel 8.15 (explizite Runge-Kutta-Verfahren vierter Ordnung)

$$\begin{array}{c|cccc}0&&&&\\ \frac12&\frac12&&&\\ \frac12&0&\frac12&&\\ 1&0&0&1&\\ \hline &\frac16&\frac13&\frac13&\frac16\end{array}$$

klassisches Runge-Kutta-Verfahren

$$\begin{array}{c|cccc}0&&&&\\ \frac12&\frac12&&&\\ \frac12&\frac{\sqrt{2}-1}{2}&\frac{1-\sqrt{2}}{2}&&\\ 1&0&-\frac{\sqrt{2}}{2}&\frac{1+\sqrt{2}}{2}&\\ \hline &\frac16&\frac{2-\sqrt{2}}{6}&\frac{2+\sqrt{2}}{6}&\frac16\end{array}$$

Gills-Formel

$$\begin{array}{c|cccc}0&&&&\\ \frac13&\frac13&&&\\ \frac23&-\frac13&1&&\\ 1&1&-1&1&\\ \hline &\frac18&\frac38&\frac38&\frac18\end{array}$$

3/8-Regel

△

Oftmals sind implizite Runge-Kutta-Verfahren den expliziten vorzuziehen. Das Butcher-Tableau besitzt in diesem Fall nichttriviale Koeffzienten $\beta_{j,\ell}$ für $\ell \geq j$. Daher muss für jedes j das nichtlineare Gleichungssystem

$$\begin{aligned}\mathbf{k}_1(x,\mathbf{y},h)&=\mathbf{f}\big(x+h\alpha_1,\mathbf{y}+h\beta_{1,1}\mathbf{k}_1(x,\mathbf{y},h)+\cdots+h\beta_{1,m}\mathbf{k}_m(x,\mathbf{y},h)\big)\\ \mathbf{k}_2(x,\mathbf{y},h)&=\mathbf{f}\big(x+h\alpha_2,\mathbf{y}+h\beta_{2,1}\mathbf{k}_1(x,\mathbf{y},h)+\cdots+h\beta_{2,m}\mathbf{k}_m(x,\mathbf{y},h)\big)\\ &\vdots\\ \mathbf{k}_m(x,\mathbf{y},h)&=\mathbf{f}\big(x+h\alpha_m,\mathbf{y}+h\beta_{m,1}\mathbf{k}_1(x,\mathbf{y},h)+\cdots+h\beta_{m,m}\mathbf{k}_m(x,\mathbf{y},h)\big)\end{aligned}\tag{8.10}$$

in mn Unbekannten $\mathbf{k}_i=\mathbf{k}_i(x,\mathbf{y},h)$, $i=1,2,\ldots,m$, gelöst werden.

Satz 8.16 Auf dem Streifen $S := \{(x,\mathbf{y}) : a \leq x \leq b, \mathbf{y} \in \mathbb{R}^n\}$ sei die Funktion $\mathbf{f} : S \to \mathbb{R}^n$ Lipschitz-stetig bezüglich $\mathbf{y}$ mit Lipschitz-Konstante L. Die Schrittweite h sei klein genug, so dass

$$q := Lh\max_{i=1}^{m}\sum_{j=1}^{m}|\beta_{i,j}| < 1$$

ist. Dann existieren für alle $(x, \mathbf{y}) \in S$ eindeutige Lösungsvektoren $\mathbf{k}_i = \mathbf{k}_i(x, \mathbf{y}, h)$, $i = 1, 2, \ldots, m$, von (8.10).

Beweis. Setze

$$\mathbf{K} = \big[\mathbf{k}_1, \mathbf{k}_2, \ldots, \mathbf{k}_m\big] := \big[\mathbf{k}_1(x, \mathbf{y}, h), \mathbf{k}_2(x, \mathbf{y}, h), \ldots, \mathbf{k}_m(x, \mathbf{y}, h)\big] \in \mathbb{R}^{n\times m}$$

und

$$\Psi : \mathbb{R}^{n\times m} \to \mathbb{R}^{n\times m}, \quad \Psi(\mathbf{K}) = \big[\boldsymbol{\psi}_1(\mathbf{K}), \boldsymbol{\psi}_2(\mathbf{K}), \ldots, \boldsymbol{\psi}_m(\mathbf{K})\big]$$

mit

$$\boldsymbol{\psi}_j(\mathbf{K}) := \mathbf{f}\Big(x + h\alpha_j, \mathbf{y} + h\sum_{\ell=1}^{m} \beta_{j,\ell}\mathbf{k}_\ell\Big).$$

Jeder Fixpunkt der Iteration

$$\mathbf{K}^{(r+1)} = \Psi\big(\mathbf{K}^{(r)}\big), \quad r = 0, 1, 2, \ldots$$

löst dann das nichtlineare Gleichungssystem (8.10). Wir zeigen nun, dass Ψ eine Kontraktion bezüglich der Norm $|\!|\!|\mathbf{K}|\!|\!| := \max_{i=1}^{m} \|\mathbf{k}_i\|$ ist:

$$\begin{aligned}
|\!|\!|\Psi(\mathbf{K}) - \Psi(\mathbf{L})|\!|\!| &= \max_{i=1}^{m} \|\boldsymbol{\psi}_i(\mathbf{K}) - \boldsymbol{\psi}_i(\mathbf{L})\| \\
&\le L \max_{i=1}^{m} \left\| \mathbf{y} + h\sum_{j=1}^{m} \beta_{i,j}\mathbf{k}_j - \mathbf{y} - h\sum_{j=1}^{m} \beta_{i,j}\boldsymbol{\ell}_j \right\| \\
&\le Lh \max_{i=1}^{m} \left(\sum_{j=1}^{m} |\beta_{i,j}| \|\mathbf{k}_j - \boldsymbol{\ell}_j\| \right) \\
&\le Lh \left(\max_{i=1}^{m} \sum_{j=1}^{m} |\beta_{i,j}| \right) \left(\max_{p=1}^{m} \|\mathbf{k}_p - \boldsymbol{\ell}_p\| \right) \\
&= \underbrace{Lh \left(\max_{i=1}^{m} \sum_{j=1}^{m} |\beta_{i,j}| \right)}_{=q<1} |\!|\!|\mathbf{K} - \mathbf{L}|\!|\!|.
\end{aligned}$$

Damit folgen Existenz und Eindeutigkeit der Lösung von (8.10) aus dem Banachschen Fixpunktsatz. □

Bemerkung: Am schnellsten löst man das nichtlineare Gleichungsystem (8.10) mit dem Newton-Verfahren, wobei dann wieder erste Ableitungen von $\mathbf{f}$ eingehen.

Beispiel 8.17 (implizite Runge-Kutta-Verfahren)

$$\begin{array}{c|c} \frac{1}{2} & \frac{1}{2} \\ \hline & 1 \end{array}$$

implizite Mittelpunktsregel (Ordnung 2)

$$\begin{array}{c|cc} \frac{2-\sqrt{2}}{2} & \frac{2-\sqrt{2}}{2} & \\ 1 & \frac{\sqrt{2}}{2} & \frac{2-\sqrt{2}}{2} \\ \hline & \frac{\sqrt{2}}{2} & \frac{2-\sqrt{2}}{2} \end{array}$$

SDIRK2-Verfahren (single-diagonal implizites Runge-Kutta-Verfahren, Ordnung 2)

$$\begin{array}{c|cc} \frac{1}{3} & \frac{5}{12} & -\frac{1}{12} \\ 1 & \frac{3}{4} & \frac{1}{4} \\ \hline & \frac{3}{4} & \frac{1}{4} \end{array}$$

Radau3-Verfahren (basiert auf der Radau-Quadraturformel, Ordnung 3)

△

Satz 8.18 Die Funktion $\mathbf{f} : S \to \mathbb{R}^n$ sei auf dem Streifen $S := \{(x, \mathbf{y}) : a \le x \le b, \mathbf{y} \in \mathbb{R}^n\}$ Lipschitz-stetig bezüglich $\mathbf{y}$ mit Lipschitz-Konstante L. Wählt man zu paarweise verschiedenen $\alpha_i \in [0,1]$, $i = 1, 2, \ldots, m$, Parameter γ_i, $\beta_{i,j}$, $i, j = 1, 2, \ldots, m$ derart, dass für ein $r \in \mathbb{N}$ gilt

$$\sum_{j=1}^{m} \beta_{i,j} \alpha_j^{\ell} = \frac{\alpha_i^{\ell+1}}{\ell+1}, \quad i = 1, 2, \ldots, m, \quad \ell = 0, 1, 2, \ldots, r-1, \tag{8.11}$$

und

$$\sum_{j=1}^{m} \gamma_j \alpha_j^{\ell} = \frac{1}{\ell+1}, \quad \ell = 0, 1, 2, \ldots, r-1, \tag{8.12}$$

gilt, dann ist die entsprechende Runge-Kutta-Formel konsistent von der Ordnung r.

Beweis. Gemäß (8.11) und (8.12) folgt

$$\frac{\alpha_i^{\ell}}{\ell+1} = \int_0^{\alpha_i} t^{\ell} \, \mathrm{d}t = \sum_{j=1}^{m} \beta_{i,j} \alpha_j^{\ell}, \quad i = 1, 2, \ldots, m, \quad \ell = 0, 1, 2, \ldots, r-1,$$

und

$$\frac{1}{\ell+1} = \int_0^{1} t^{\ell} \, \mathrm{d}t = \sum_{j=1}^{m} \gamma_j \alpha_j^{\ell}, \quad \ell = 0, 1, 2, \ldots, r-1,$$

dies bedeutet, die Quadraturformeln mit den Knoten $\{\alpha_j\}_{j=1}^m$ und Gewichten $\{\beta_{i,j}\}_{j=1}^m$ sind auf dem Intervall $[0, \alpha_i]$ exakt von der Ordnung r, ebenso die Quadraturformel mit den Gewichten $\{\gamma_j\}_{j=1}^m$ auf $[0,1]$. Daher folgt

$$\begin{aligned}
\mathbf{y}(x+h) - \mathbf{y}(x) &= \int_x^{x+h} \mathbf{f}\big(t, \mathbf{y}(t)\big) \, \mathrm{d}t \\
&= h \int_0^1 \mathbf{f}\big(x + hs, \mathbf{y}(x+hs)\big) \, \mathrm{d}s \\
&= h \sum_{i=1}^{m} \gamma_i \mathbf{f}\big(x + h\alpha_i, \mathbf{y}(x+h\alpha_i)\big) + \mathcal{O}(h^{r+1}).
\end{aligned}$$

Zusammen mit (8.10) ergibt sich daher

$$\begin{aligned}
\big\|\widetilde{\boldsymbol{\tau}}(x, \mathbf{y}, h)\big\| &= \left\| \frac{\mathbf{y}(x+h) - \mathbf{y}(x)}{h} - h \sum_{i=1}^{m} \gamma_i \mathbf{k}_i(x, \mathbf{y}, h) \right\| \\
&= \left\| \sum_{i=1}^{m} \gamma_i \mathbf{f}\big(x + h\alpha_i, \mathbf{y}(x+h\alpha_i)\big) \right. \\
&\qquad \left. - \sum_{i=1}^{m} \gamma_i \mathbf{f}\left(x + h\alpha_i, \mathbf{y}(x) + h \sum_{j=1}^{m} \beta_{i,j} \mathbf{k}_j(x, \mathbf{y}, h)\right) \right\| + \mathcal{O}(h^r) \\
&\le L \sum_{i=1}^{m} |\gamma_i| \left\| \mathbf{y}(x + h\alpha_i) - \mathbf{y}(x) - h \sum_{j=1}^{m} \beta_{i,j} \mathbf{k}_j(x, \mathbf{y}, h) \right\| + \mathcal{O}(h^r).
\end{aligned}$$

Auf die gleiche Weise folgt unter Ausnutzung der Quadraturformeln auf $[0, \alpha_i]$, dass

$$\begin{aligned}
A_i &:= \left\| \mathbf{y}(x+h\alpha_i) - \mathbf{y}(x) - h\sum_{j=1}^{m} \beta_{i,j}\mathbf{k}_j(x,\mathbf{y},h) \right\| \\
&= \left\| h\int_0^{\alpha_i} \mathbf{f}\big(x+hs, \mathbf{y}(x+hs)\big)\,\mathrm{d}s - h\sum_{j=1}^{m} \beta_{i,j}\mathbf{k}_j(x,\mathbf{y},h) \right\| \\
&= h\Bigg\| \sum_{j=1}^{m} \beta_{i,j}\mathbf{f}\big(x+h\alpha_j, \mathbf{y}(x+h\alpha_j)\big) \\
&\qquad - \sum_{j=1}^{m} \beta_{i,j}\mathbf{f}\Bigg(x+h\alpha_j, \mathbf{y}(x) + h\sum_{\ell=1}^{m} \beta_{j,\ell}\mathbf{k}_\ell(x,\mathbf{y},h)\Bigg)\Bigg\| + \mathcal{O}(h^{r+1}) \\
&\le Lh\sum_{j=1}^{m} |\beta_{i,j}| \underbrace{\left\| \mathbf{y}(x+h\alpha_j) - \mathbf{y}(x) - h\sum_{\ell=1}^{m} \beta_{j,\ell}\mathbf{k}_\ell(x,\mathbf{y},h) \right\|}_{=A_j} + \mathcal{O}(h^{r+1}) \\
&\le Lh\Bigg(\underbrace{\max_{j=1}^{m} A_j}_{=:A}\Bigg)\Bigg(\underbrace{\max_{j=1}^{m} \sum_{\ell=1}^{m} |\beta_{j,\ell}|}_{=:B}\Bigg) + \mathcal{O}(h^{r+1}).
\end{aligned}$$

Hieraus folgt

$$A(1 - LhB) = \mathcal{O}(h^{r+1}),$$

dies bedeutet, $A = \mathcal{O}(h^{r+1})$. Schließlich folgt die Behauptung aus

$$\big\|\widetilde{\boldsymbol{\tau}}(x,\mathbf{y},h)\big\| \le L\sum_{i=1}^{m} |\gamma_i| A_i + \mathcal{O}(h^r) \le LA\sum_{i=1}^{m} |\gamma_i| + \mathcal{O}(h^r) = \mathcal{O}(h^r).$$

□

Bemerkung: Setzt man

$$\mathbf{A} = \begin{bmatrix} 1 & 1 & \dots & 1 \\ \alpha_1 & \alpha_2 & \dots & \alpha_m \\ \alpha_1^2 & \alpha_2^2 & \dots & \alpha_m^2 \\ \vdots & \vdots & & \vdots \\ \alpha_1^{r-1} & \alpha_2^{r-1} & \dots & \alpha_m^{r-1} \end{bmatrix} \in \mathbb{R}^{r\times m}, \qquad \mathbf{B} = \begin{bmatrix} \beta_{1,1} & \beta_{1,2} & \dots & \beta_{1,m} \\ \beta_{2,1} & \beta_{2,2} & \dots & \beta_{2,m} \\ \vdots & \vdots & & \vdots \\ \beta_{m,1} & \beta_{m,2} & \dots & \beta_{m,m} \end{bmatrix} \in \mathbb{R}^{m\times m},$$

und

$$\mathbf{C} = \begin{bmatrix} \alpha_1 & \alpha_2 & \dots & \alpha_m \\ \alpha_1^2/2 & \alpha_2^2/2 & \dots & \alpha_m^2/2 \\ \vdots & \vdots & & \vdots \\ \alpha_1^r/r & \alpha_2^r/r & \dots & \alpha_m^r/r \end{bmatrix} \in \mathbb{R}^{r\times m}, \qquad \boldsymbol{\gamma} = \begin{bmatrix} \gamma_1 \\ \gamma_2 \\ \vdots \\ \gamma_m \end{bmatrix} \in \mathbb{R}^m, \qquad \mathbf{c} = \begin{bmatrix} 1 \\ 1/2 \\ \vdots \\ 1/r \end{bmatrix} \in \mathbb{R}^{r+1},$$

dann lassen sich (8.11) und (8.12) kurz schreiben als

$$\mathbf{A}\mathbf{B}^\star = \mathbf{C}, \qquad \mathbf{A}\boldsymbol{\gamma} = \mathbf{c}.$$

Da $\mathbf{A}$ eine Vandermonde-Matrix ist, besitzt sie den vollen Rang r, und beide Gleichungssysteme sind lösbar, falls $r < m$. Insbesondere folgt die Eindeutigkeit im Fall $r = m$. Damit eine Exaktheit $r > m$ erreicht werden kann, müssen die Knoten $\{\alpha_j\}_{j=1}^m$ speziell gewählt werden, beispielsweise als Gauß-Quadraturpunkte.

8.7 Mehrschrittverfahren

Genereller Nachteil von Einschrittverfahren ist die hohe Anzahl von Funktionsauswertungen im Fall von Verfahren höherer Ordnung. Ursache dafür ist, dass jeder Schritt so behandelt wird, als wäre es der erste.

Definition 8.19 Sei $x_i := x_0 + hi$ für $i \in \mathbb{N}$. Ein Verfahren der Form

$$\boldsymbol{\eta}_{i+k} := \sum_{\ell=0}^{k-1} \alpha_\ell \boldsymbol{\eta}_{i+\ell} + h\Phi(x_i, \boldsymbol{\eta}_i, \boldsymbol{\eta}_{i+1}, \ldots, \boldsymbol{\eta}_{i+k}, h) \tag{8.13}$$

mit festen Koeffizienten α_ℓ heißt ***k*-Schrittverfahren**. Hängt die Inkrementfunktion Φ nicht von $\boldsymbol{\eta}_{i+k}$ ab, dann ist das Mehrschrittverfahren **explizit**, andernfalls **implizit**. Das Mehrschrittverfahren heißt **linear**, falls es von der Form

$$\boldsymbol{\eta}_{i+k} := \sum_{\ell=0}^{k-1} \alpha_\ell \boldsymbol{\eta}_{i+\ell} + h \sum_{\ell=0}^{k} \beta_\ell \mathbf{f}(x_{i+\ell}, \boldsymbol{\eta}_{i+\ell}) \tag{8.14}$$

mit festen Koeffizienten α_ℓ, β_ℓ ist.

Bemerkungen:

1. Ein lineares Mehrschrittverfahren ist genau dann implizit, falls $\beta_k \neq 0$ gilt.
2. Ein k-Schrittverfahren benötigt immer k Anlaufwerte

   $$\boldsymbol{\eta}_0 = \mathbf{y}_0, \boldsymbol{\eta}_1 \approx \mathbf{y}(x_1), \boldsymbol{\eta}_2 \approx \mathbf{y}(x_2), \ldots, \boldsymbol{\eta}_{k-1} \approx \mathbf{y}(x_{k-1}).$$

   In der Praxis besorgt man sich diese aus einem hinreichend genauen Einschrittverfahren.

Algorithmus 8.20 (lineares Mehrschrittverfahren)

input: Funktion $\mathbf{f} \in C([a,b] \times \mathbb{R}^n)$, Anfangswerte x_0, $\mathbf{y}_0, \mathbf{y}_1, \ldots, \mathbf{y}_{k-1}$ und Schrittweite h

output: Approximation $\{\boldsymbol{\eta}_i\}$

① Initialisierung: setze $x_i := x_0 + ih$ und $\boldsymbol{\eta}_i := \mathbf{y}_i$ für alle $i = 0, 1, \ldots, k-1$

② für alle $i = 0, 1, \ldots$ berechne

$$\boldsymbol{\eta}_{i+k} := \sum_{\ell=0}^{k-1} \alpha_\ell \boldsymbol{\eta}_{i+\ell} + h \sum_{\ell=0}^{k} \beta_\ell \mathbf{f}\big(x_{i+\ell}, \boldsymbol{\eta}_{i+\ell}\big), \qquad x_{i+k} := x_{i+k-1} + h$$

Beispiel 8.21 Folgende Spezialfälle linearer Mehrschrittverfahren sind von Bedeutung:

- Beim *Adams-Bashforth-Verfahren* gilt

  $$\alpha_0 = \alpha_1 = \cdots = \alpha_{k-2} = 0, \quad \alpha_{k-1} = 1, \quad \beta_k = 0.$$

- Beim *Adams-Moulton-Verfahren* ist

$$\alpha_0 = \alpha_1 = \cdots = \alpha_{k-2}, \quad \alpha_{k-1} = 1, \quad \beta_k \neq 0.$$

△

Definition 8.22 Der **lokale Diskretisierungsfehler** des Mehrschrittverfahrens ist

$$\boldsymbol{\tau}(x_i, \mathbf{y}, h) = \frac{1}{h}\left(\mathbf{y}(x_{i+k}) - \sum_{\ell=0}^{k-1} \alpha_\ell \mathbf{y}(x_{i+\ell})\right) - \Phi\big(x_i, \mathbf{y}(x_i), \ldots, \mathbf{y}(x_{i+k}), h\big).$$

Das Mehrschrittverfahren heißt **konsistent** von der Ordnung p, falls gilt

$$\|\boldsymbol{\tau}(x, \mathbf{y}, h)\| = \mathcal{O}(h^p).$$

Im Gegensatz zu Runge-Kutta-Verfahren kann man für lineare Mehrschrittverfahren sehr einfache Konsistenzbedingungen formulieren.

Satz 8.23 Sei $\mathbf{f}$ hinreichend glatt. Dann ist ein lineares Mehrschrittverfahren der Form

$$\sum_{\ell=0}^{k} \alpha_\ell \boldsymbol{\eta}_{i+\ell} - h \sum_{\ell=0}^{k} \beta_\ell \mathbf{f}(x_{i+\ell}, \boldsymbol{\eta}_{i+\ell}), \quad i = 0, 1, 2, \ldots$$

konsistent von der Ordnung p genau dann, wenn die folgenden $p+1$ Bedingungen erfüllt sind:

$$\sum_{\ell=0}^{k} \alpha_\ell = 0, \qquad \sum_{\ell=0}^{k} \big(\ell^q \alpha_\ell - q\ell^{q-1}\beta_\ell\big) = 0, \quad q = 1, 2, \ldots, p. \tag{8.15}$$

Beweis. Taylor-Entwicklung von $\mathbf{y}$ liefert

$$\begin{aligned}
\alpha_k \boldsymbol{\tau}(x, \mathbf{y}, h) &= \frac{1}{h}\sum_{\ell=0}^{k} \alpha_\ell \mathbf{y}(x + h\ell) - \sum_{\ell=0}^{k} \beta_\ell \mathbf{y}'(x + h\ell) \\
&= \frac{1}{h}\sum_{\ell=0}^{k} \alpha_\ell \left(\sum_{q=0}^{p} \frac{(h\ell)^q}{q!} \mathbf{y}^{(q)}(x)\right) - \sum_{\ell=0}^{k} \beta_\ell \left(\sum_{q=0}^{p-1} \frac{(h\ell)^q}{q!} \mathbf{y}^{(q+1)}(x)\right) + \mathcal{O}(h^p) \\
&= \frac{1}{h}\left(\sum_{\ell=0}^{k} \alpha_\ell \mathbf{y}(x)\right) + \left(\sum_{q=1}^{p} \frac{h^{q-1}}{q!} \mathbf{y}^{(q)}(x) \sum_{\ell=0}^{k} \ell^q \alpha_\ell\right) \\
&\qquad - \left(\sum_{q=1}^{p} \frac{h^{q-1}}{q!} \mathbf{y}^{(q)}(x) \sum_{\ell=0}^{k} q\ell^{q-1}\beta_\ell\right) + \mathcal{O}(h^p) \\
&= \frac{1}{h}\left(\sum_{\ell=0}^{k} \alpha_\ell \mathbf{y}(x)\right) + \sum_{q=1}^{p} \frac{h^{q-1}}{q!} \mathbf{y}^{(q)}(x) \sum_{\ell=0}^{k} \big(\ell^q \alpha_\ell - q\ell^{q-1}\beta_\ell\big) + \mathcal{O}(h^p).
\end{aligned}$$

Gelten die Bedingungen (8.15), dann ist das Verfahren von der Konsistenzordnung p.

Damit umgekehrt das Verfahren die Konsistenzordnung p besitzt, müssen die Koeffizienten von h^q verschwinden. Da es Fälle mit $\mathbf{y}^{(q)}(x) \neq 0$ gibt, müssen die Bedingungen (8.15) gelten. □

8.8 Adams-Bashforth-Verfahren

Diese Klasse von Mehrschrittverfahren beruht auf der Diskretisierung der Integralgleichung

$$\mathbf{y}(x_{i+k}) = \mathbf{y}(x_{i+k-1}) + \int_{x_{i+k-1}}^{x_{i+k}} \mathbf{f}\big(t, \mathbf{y}(t)\big)\,\mathrm{d}t \tag{8.16}$$

an den Stüzstellen $x_i, x_{i+1}, \ldots, x_{i+k-1}$ mit Hilfe von Newton-Cotes-Formeln. Da die Stützstellen außerhalb des Integrationsintervalls $[x_{i+k-1}, x_{i+k}]$ liegen, liefert (7.5) folgende, nicht den üblichen aus Abschnitt 7.3 entsprechenden Gewichte:

k	b_0	b_1	b_2	b_3	b_4	Konsistenzordnung
1	1					1
2	$-1/2$	$3/2$				2
3	$5/12$	$-4/3$	$23/12$			3
4	$-3/8$	$37/24$	$-59/24$	$55/24$		4
5	$251/720$	$-637/360$	$327/90$	$-1387/360$	$1901/720$	5

Da die entsprechenden Quadraturformeln gemäß Proposition 7.10 exakt von der Ordnung k sind, folgt die Abschätzung

$$\int_{x_{i+k-1}}^{x_{i+k}} \mathbf{f}\big(t, \mathbf{y}(t)\big)\,\mathrm{d}t = h \sum_{\ell=0}^{k-1} \beta_\ell \underbrace{\mathbf{f}\big(x_{i+\ell}, \mathbf{y}(x_{i+\ell})\big)}_{=\mathbf{y}'(x_{i+\ell})} + \mathcal{O}(h^{k+1}).$$

Dies bedeutet, das so konstruierte k-Schritt-Verfahren besitzt die Konsistenzordnung k.

Algorithmus 8.24 (Adams-Bashforth-Verfahren)

input: Funktion $\mathbf{f} \in C([a,b] \times \mathbb{R}^n)$, Anfangswerte x_0, $\mathbf{y}_0, \mathbf{y}_1, \ldots, \mathbf{y}_{k-1}$ und Schrittweite h

output: Approximation $\{\boldsymbol{\eta}_i\}$

① Initialisierung: setze $x_i := x_0 + ih$ und $\boldsymbol{\eta}_i := \mathbf{y}_i$ für alle $i = 0, 1, \ldots, k-1$

② für alle $i = 0, 1, \ldots$ berechne

$$\boldsymbol{\eta}_{i+k} := \boldsymbol{\eta}_{i+k-1} + h \sum_{\ell=0}^{k-1} \beta_\ell \mathbf{f}\big(x_{i+\ell}, \boldsymbol{\eta}_{i+\ell}\big), \qquad x_{i+k} := x_{i+k-1} + h$$

Bemerkung: Beachte, dass pro Iterationsschritt nur eine Funktionsauswertung von $\mathbf{f}$ benötigt wird.

8.9 Adams-Moulton-Verfahren

Das Integral in (8.16) wird nun durch Newton-Cotes-Formeln in den Knoten $x_i, x_{i+1}, \ldots, x_{i+k-1}$ und zusätzlich x_{i+k} approximiert. Dies führt auf:

k	b_0	b_1	b_2	b_3	b_4	Konsistenzordnung
1	$1/2$	$1/2$				2
2	$-1/12$	$2/3$	$5/12$			3
3	$1/24$	$-5/24$	$19/24$	$3/8$		4
4	$-19/720$	$53/360$	$-11/30$	$323/360$	$251/720$	5

Da die Newton-Cotes-Quadraturformeln gemäß Proposition 7.10 exakt von der Ordnung $k+1$ sind, ergibt sich die Abschätzung

$$\int_{x_{i+k-1}}^{x_{i+k}} \mathbf{f}\big(t, \mathbf{y}(t)\big)\,\mathrm{d}t = h \sum_{\ell=0}^{k} \beta_\ell \underbrace{\mathbf{f}\big(x_{i+\ell}, \mathbf{y}(x_{i+\ell})\big)}_{=\mathbf{y}'(x_{i+\ell})} + \mathcal{O}(h^{k+2}).$$

Hieraus folgt, dass das Adams-Moulton-k-Schrittverfahren die Konsistenzordnung $k+1$ besitzt.

Pro Schritt muss das nichtlineare Gleichungssystem

$$\boldsymbol{\eta}_{i+k} - h\beta_k \mathbf{f}(x_{i+k}, \boldsymbol{\eta}_{i+k}) = \boldsymbol{\eta}_{i+k-1} + h \sum_{\ell=0}^{k-1} \beta_\ell \mathbf{f}(x_{i+\ell}, \boldsymbol{\eta}_{i+\ell})$$

gelöst werden. Wie schon bei den impliziten Einschrittverfahren kann nachgewiesen werden, dass diese Gleichung für hinreichend kleine Schrittweiten lösbar ist, falls $\mathbf{f}$ einer Lipschitz-Bedingung in der $\mathbf{y}$-Variablen genügt. Setzen wir

$$\Phi(\mathbf{u}) := h\beta_k \mathbf{f}(x_{i+k}, \mathbf{u}) + \boldsymbol{\eta}_{i+k-1} + h \sum_{\ell=0}^{k-1} \beta_\ell \mathbf{f}(x_{i+\ell}, \boldsymbol{\eta}_{i+\ell}),$$

dann folgt

$$\|\Phi(\mathbf{u}) - \Phi(\mathbf{v})\| = h\beta_k \|\mathbf{f}(x_{i+k}, \mathbf{u}) - \mathbf{f}(x_{i+k}, \mathbf{v})\| \le h\beta_k L \|\mathbf{u} - \mathbf{v}\|.$$

Ist h so klein, dass $h\beta_k L < 1$ gilt, dann liefert der Banachsche Fixpunktsatz Existenz und Eindeutigkeit der Lösung.

Bemerkung: Eine gute Startnäherung $\widetilde{\boldsymbol{\eta}}_{i+k}$ des nichtlinearen Lösers berechnet man üblicherweise durch ein explizites Verfahren — dies ist der Prädiktorschritt. Mit Hilfe einer Fixpunktiteration oder einem Newton-Verfahren verbessert man die Näherung $\widetilde{\boldsymbol{\eta}}_{i+k}$ und erhält die gesuchte Lösung $\boldsymbol{\eta}_{i+k}$ — dies ist der Korrektorschritt. Man spricht hier von *Prädiktor-Korrektor-Verfahren.*

8.10 Stabilität

Einschrittverfahren sind mit $k = 1$ natürlich formal in der Klasse der Mehrschrittverfahren enthalten. Allerdings impliziert im Gegensatz zu Einschrittverfahren die Konsistenz nicht die Konvergenz von Mehrschrittverfahren im Fall einer Schrittzahl $k > 1$.

Beispiel 8.25 Wir betrachten das Anfangswertproblem

$$y'(x) = y(x) - 2\sin x, \quad x \in [0,4], \qquad y(0) = 1,$$

mit der Lösung $y(x) = \sin(x) + \cos(x)$. Zur numerischen Berechnung dieser Lösung verwenden wir das explizite Zweischrittverfahren

$$\eta_{i+2} = -4\eta_{i+1} + 5\eta_i + h\big(2f(x_i,\eta_i) + 4f(x_{i+1},\eta_{i+1})\big), \quad i = 0,1,2,\ldots, \tag{8.17}$$

welche die Konsistenzordnung 3 besitzt. Führen wir ausgehend von $\eta_0 = 1$, $\eta_1 = \sin h + \cos h$ für verschiedene Schrittweiten das Verfahren durch, erhalten wir:

h	$\lvert\eta_{1/h} - y(1)\rvert$	$\lvert\eta_{2/h} - y(2)\rvert$	$\lvert\eta_{4/h} - y(4)\rvert$
2^{-2}	0.0094	2.87	$3.3 \cdot 10^5$
2^{-3}	0.280	$6.1 \cdot 10^4$	$2.7 \cdot 10^{15}$
2^{-4}	$6.4 \cdot 10^3$	$5.4 \cdot 10^{14}$	$3.7 \cdot 10^{36}$

Die Resultate deuten an, dass keine Konvergenz vorliegt. Man beobachtet sogar ein starkes Anwachsen des Fehlers, wenn die Schrittweite h kleiner wird.

Die im Verfahren (8.17) auftretende Instabiliät lässt sich wie folgt erklären. Wir betrachten die Aufgabe

$$y'(x) = 0, \quad x \in [0,b], \qquad y(0) = 1,$$

mit der Lösung $y(x) = 1$. Das Zweischrittverfahren hierauf angewandt liefert die *lineare homogene Differenzengleichung*

$$\eta_{i+2} + 4\eta_{i+1} - 5\eta_i = 0, \quad i = 0,1,2,\ldots. \tag{8.18}$$

Der Ansatz $\eta_i = z^i$ führt auf $z^i(z^2 + 4z - 5) = 0$ für alle i, also $z = 0$, $z = 1$ oder $z = -5$. Folglich ist $\eta_i = 1^i \cdot A + (-5)^i \cdot B$ für jedes $A, B \in \mathbb{R}$ eine Lösung von (8.18). Sind nun gestörte Startwerte $\eta_0 = 1$ und $\eta_1 = 1 + \delta$ vorgegeben, so sind $A = 1 + \delta/6$ und $B = -\delta/6$ eindeutig bestimmt und es folgt

$$\eta_i = 1 + \frac{\delta}{6} - \frac{\delta}{6}(-5)^i, \quad i = 0,1,2,\ldots.$$

Man sieht, dass eine kleine Störung $\delta \neq 0$ im Startwert η_1 exponentiell wächst. △

Definition 8.26 Eine Gleichung der Form

$$\sum_{\ell=0}^{k} \alpha_\ell \eta_{i+\ell} = 0, \quad i = 0,1,2,\ldots. \tag{8.19}$$

mit $\alpha_k \neq 0$ heißt **lineare homogene Differenzengleichung *k*-ter Ordnung**.

Zu jedem k-Schrittverfahren der Form

$$\sum_{\ell=0}^{k} \alpha_\ell \boldsymbol{\eta}_{i+\ell} = h\Phi(x_i, \boldsymbol{\eta}_i, \boldsymbol{\eta}_{i+1}, \ldots, \boldsymbol{\eta}_{i+k}, h), \qquad i = 0,1,2,\ldots, \tag{8.20}$$

gehört eine lineare homogene Differenzengleichung der Ordnung k (8.19), welche offensichtlich k linear unabhängige Lösungen hat. Hingegen besitzt die approximierte Differentialgleichung nur eine Lösung. Dieser einen Lösung entspricht aber nur eine Lösung der Differenzengleichung. Daher kann man im Fall $k > 1$ nur dann Konvergenz einer Mehrschrittmethode erwarten, wenn alle Lösungen der zugehörigen homogenen Differenzengleichung beschränkt bleiben, da auch diese durch Rundungsfehler eingeschleppt werden. Aus diesem Grund untersuchen wir zunächst die Stabilität der Differenzengleichung.

Definition 8.27 Die lineare homogene Differenzengleichung (8.19) wird **stabil** genannt, wenn alle ihre Lösungen beschränkt sind.

Satz 8.28 Die lineare homogene Differenzengleichung (8.19) ist stabil genau dann, falls sie die Dahlquistsche Wurzelbedingung erfüllt, das heißt, falls alle Nullstellen λ des charakteristischen Polynoms

$$p(\lambda) = \sum_{\ell=0}^{k} \alpha_\ell \lambda^\ell \tag{8.21}$$

alle $|\lambda| \leq 1$ erfüllen, und im Falle $|\lambda| = 1$ zusätzlich einfach sind.

Beweis. Zunächst sei angemerkt, dass jede Lösung der linearen Differenzengleichung (8.19) eindeutig bestimmt ist durch die k Startwerte $\eta_0, \eta_1, \ldots, \eta_{k-1}$. Die nachfolgenden Terme $\eta_k, \eta_{k+1}, \ldots$ sind dann rekursiv durch (8.19) bestimmt.
Sei A der Differentialoperator definiert durch $(Af)(\lambda) = \lambda f'(\lambda)$. Dann erfüllt die Folge

$$\eta_i = i^n \lambda^i, \quad i = 0, 1, 2, \ldots, \tag{8.22}$$

für jedes i die Beziehung

$$\begin{aligned}
\sum_{\ell=0}^{k} \alpha_\ell \eta_{i+\ell} &= \sum_{\ell=0}^{k} \alpha_\ell \underbrace{(i+\ell)^n}_{=\sum_{m=0}^{n} \binom{n}{m} i^m \ell^{n-m}} \lambda^{i+\ell} \\
&= \lambda^i \sum_{m=0}^{n} \binom{n}{m} i^m \sum_{\ell=0}^{k} \alpha_\ell \ell^{n-m} \lambda^\ell \\
&= \lambda^i \sum_{m=0}^{n} \binom{n}{m} i^m (A^{n-m} p)(\lambda).
\end{aligned}$$

Falls λ eine Nullstelle des charaktersistischen Polynoms mit der Vielfachheit s ist, dann löst die Folge (8.22) die Differenzengleichung (8.19).

Nun nehmen wir an, dass $\lambda_1, \lambda_2, \ldots, \lambda_n$ die Nullstellen des charakteristischen Polynoms sind und die Vielfachheiten $s_1, s_2, \ldots, s_n$ besitzen, dies bedeutet

$$p(\lambda) = \prod_{m=0}^{n} (\lambda - \lambda_m)^{s_m}.$$

Dann besitzen alle Lösungen der linearen homogenen Differentialgleichung (8.19) die Form

$$\eta_i = \sum_{m=0}^{n} \sum_{s=0}^{s_m - 1} \gamma_{m,s} i^s \lambda_m^i \tag{8.23}$$

mit beliebigen Koeffizienten $\gamma_{m,s} \in \mathbb{R}$. Um dies zu zeigen, müssen wir zeigen, dass die Koeffizienten $\gamma_{m,s}$ so gewählt werden können, dass beliebig vorgegebene Anfangsbedingungen

$$\sum_{m=0}^{n} \sum_{s=0}^{s_m-1} \gamma_{m,s} i^s \lambda_m^i = \eta_i, \quad i = 0, 1, \ldots, k-1, \tag{8.24}$$

erfüllt sind. Das homogene adjungierte System zu (8.24) lautet

$$\sum_{i=0}^{k-1} \rho_i i^s \lambda_m^i = 0, \quad s = 0, 1, \ldots, s_m - 1, \quad m = 0, 1, \ldots, n.$$

Angenommen, ρ_i, $i = 0, 1, \ldots, k-1$ ist eine Lösung, dann besitzt das Polynom

$$q(\lambda) := \sum_{i=0}^{k-1} \rho_i \lambda^i \in \Pi_{k-1}$$

die Nullstellen $\lambda_1, \lambda_2, \ldots, \lambda_n$ mit Vielfachheiten $s_1, s_2, \ldots, s_n$. Da dies k Nullstellen sind, muss $\rho_0 = \rho_1 = \cdots = \rho_{k-1} = 0$ gelten, dies bedeutet, das Gleichungssytem (8.24) ist eindeutig lösbar.

Anhand der Darstellung (8.23) folgt schließlich die Äquivalenz von Wurzelbedingung und Stabilität. □

Definition 8.29 Das Mehrschrittverfahren (8.20) ist **nullstabil**, falls die Dahlquistsche Wurzelbedingung erfüllt ist, wenn also für alle Nullstellen λ des charakteristischen Polnoms (8.21) gilt $|\lambda| \leq 1$ und darüberhinaus, falls diese Nullstelle sogar mehrfach ist, $|\lambda| < 1$.

Bemerkung: Bei einem linearen Einschrittverfahren gilt immer $p(\lambda) = \lambda - 1$, das heißt, $\lambda = 1$ ist einzige Nullstelle von p. Damit ist die Wurzelbedingung erfüllt und das Einschrittverfahren immer nullstabil.

Beim Adams-Bashforth- und Adams-Moulton-Verfahren hat man $p(\lambda) = \lambda^{k-1}(\lambda - 1)$, also eine einfache Nullstelle bei $\lambda = 1$ und eine $(k-1)$-fache Nullstelle bei $\lambda = 0$. Beide Verfahren sind folglich nullstabil.

8.11 Konvergenz von Mehrschrittverfahren

Zur Konvergenzuntersuchung müssen wir den Einfluss von Rundungsfehlern beachten. Daher betrachten wir zur näherungsweisen Lösung der Anfangswertaufgabe (8.5) das *gestörte* Mehrschrittverfahren

$$\begin{gathered}\sum_{\ell=0}^{k} \alpha_\ell \boldsymbol{\eta}_{i+\ell} = h\Phi(x_i, \boldsymbol{\eta}_i, \boldsymbol{\eta}_{i+1}, \ldots, \boldsymbol{\eta}_{i+k}, h) + h\boldsymbol{\varepsilon}_{i+k}(h), \quad x_i = x_0 + hi, \quad i = 0, 1, \ldots \\ \boldsymbol{\eta}_i = \mathbf{y}(x_i) + \boldsymbol{\varepsilon}_i(h), \quad i = 0, 1, \ldots, k-1.\end{gathered} \tag{8.25}$$

Im folgenden wollen wir stets dabei annehmen, dass $\alpha_k = 1$ ist.

Definition 8.30 Für alle Startwerte und Störungen gelte $\|\boldsymbol{\eta}_i - \mathbf{y}(x_i)\| = \mathcal{O}(h^p)$ und $\|\boldsymbol{\varepsilon}_i(h)\| = \mathcal{O}(h^p)$. Dann heißt das Mehrschrittverfahren (8.25) **konvergent** von der Ordnung p, falls der globale Diskretisierungsfehler zur Näherungslösung $\boldsymbol{\eta}(x_i, h_n) = \boldsymbol{\eta}_i$ für jedes feste x der Abschätzung

$$e(x, h_n) = \|\boldsymbol{\eta}(x, h_n) - \mathbf{y}(x)\| = \mathcal{O}(h_n^p), \quad h_n = \frac{x - x_0}{n} \to 0$$

genügt.

Satz 8.31 Es gelte die Bedingung

$$\mathbf{f} \equiv 0 \implies \Phi(x, \boldsymbol{\eta}_0, \boldsymbol{\eta}_1, \ldots, \boldsymbol{\eta}_k, h) \equiv 0 \tag{8.26}$$

für alle $x \in [a, b]$, h und $\boldsymbol{\eta}_i$. Dann folgt aus der Konvergenz des Mehrschrittverfahrens (8.25) stets seine Nullstabilität.

Beweis. Wählt man die spezielle skalare Anfangswertaufgabe

$$y'(x) = 0, \quad x \in [0, b], \qquad y(0) = 1,$$

dann folgt wegen (8.26), dass alle Näherungslösungen im Fall $\varepsilon_i(h) = 0$, $i \geq k$, die homogene Differenzengleichung (8.19) erfüllen. Aus der Konvergenz des Mehrschrittverfahrens folgt, dass die Lösungen für alle Startwerte beschränkt sind. Daraus ergibt sich mit Satz 8.28 das Behauptete. □

Lemma 8.32 Das Polynom

$$p(\lambda) = \lambda^k + \alpha_{k-1}\lambda^{k-1} + \cdots \alpha_1\lambda + \alpha_0$$

erfülle die Dahlquistsche Wurzelbedingung und sei $\mathbf{A}$ die zugehörige transponierte Frobenius-Begleitmatrix

$$\mathbf{A} = \begin{bmatrix} 0 & 1 & & \\ & \ddots & \ddots & \\ & & 0 & 1 \\ -\alpha_0 & \cdots & -\alpha_{k-2} & -\alpha_{k-1} \end{bmatrix} \in \mathbb{C}^{k\times k}.$$

Dann existiert eine reguläre Matrix $\mathbf{R} \in \mathbb{C}^{k\times k}$, so dass für die von der Vektornorm $\|\mathbf{x}\|_{\mathbf{R}} := \|\mathbf{R}\mathbf{x}\|_\infty$ induzierte Matrixnorm gilt

$$\|\mathbf{A}\|_{\mathbf{R}} = \max_{\|\mathbf{x}\|_{\mathbf{R}}=1} \|\mathbf{A}\mathbf{x}\|_{\mathbf{R}} = \max_{\|\mathbf{R}\mathbf{x}\|_\infty=1} \|\mathbf{R}\mathbf{A}\mathbf{x}\|_\infty \leq 1.$$

Beweis. Seien $\lambda_1, \lambda_2 \ldots, \lambda_n$ Nullstellen von p mit den Vielfachheiten $s_1, s_2, \ldots, s_n$. Da diese wegen $\det(\lambda\mathbf{I} - \mathbf{A}) = p(\lambda)$ mit den Eigenwerten von $\mathbf{A}$ zusammenfallen, existiert

eine reguläre Matrix $\mathbf{T} \in \mathbb{C}^{k\times k}$, die $\mathbf{A}$ in die Jordan-Normalform überführt

$$\mathbf{T}^{-1}\mathbf{A}\mathbf{T} = \begin{bmatrix} \mathbf{J}_1 & & & \\ & \mathbf{J}_2 & & \\ & & \ddots & \\ & & & \mathbf{J}_n \end{bmatrix}, \qquad \mathbf{J}_i = \begin{bmatrix} \lambda_i & 1 & & \\ & \ddots & \ddots & \\ & & \lambda_i & 1 \\ & & & \lambda_i \end{bmatrix} \in \mathbb{C}^{s_i\times s_i}.$$

Wir setzen

$$\varepsilon := \max_{i=1}^{n}\{1 - |\lambda_i|\} \geq 0$$

und definieren

$$\mathbf{D} = \begin{bmatrix} 1 & & & \\ & \varepsilon & & \\ & & \ddots & \\ & & & \varepsilon^{k-1} \end{bmatrix} \in \mathbb{R}^{k\times k},$$

falls $\varepsilon > 0$, ansonsten sei $\mathbf{D} = \mathbf{I}$. Hiermit folgt

$$\mathbf{D}^{-1}\mathbf{T}^{-1}\mathbf{A}\mathbf{T}\mathbf{D} = \begin{bmatrix} \widetilde{\mathbf{J}}_1 & & & \\ & \widetilde{\mathbf{J}}_2 & & \\ & & \ddots & \\ & & & \widetilde{\mathbf{J}}_n \end{bmatrix}, \qquad \widetilde{\mathbf{J}}_i = \begin{bmatrix} \lambda_i & \varepsilon & & \\ & \ddots & \ddots & \\ & & \lambda_i & \varepsilon \\ & & & \lambda_i \end{bmatrix} \in \mathbb{C}^{s_i\times s_i},$$

wobei nach Konstruktion gilt

$$\left\|\widetilde{\mathbf{J}}_i\right\|_\infty = \begin{cases} |\lambda_i| + \varepsilon \leq 1, & \text{falls } |\lambda_i| < 1, \\ 1, & \text{falls } |\lambda_i| = 1. \end{cases}$$

Letzteres gilt, da der Jordan-Block im Fall $|\lambda_i| = 1$ trivial ist. Es ergibt sich $\|\mathbf{D}^{-1}\mathbf{T}^{-1}\mathbf{A}\mathbf{T}\mathbf{D}\|_\infty = 1$ und weiter mit $\mathbf{R} = \mathbf{D}^{-1}\mathbf{T}^{-1}$ schließlich

$$\|\mathbf{A}\|_{\mathbf{R}} = \max_{\|\mathbf{D}^{-1}\mathbf{T}^{-1}\mathbf{x}\|_\infty = 1} \|\mathbf{D}^{-1}\mathbf{T}^{-1}\mathbf{A}\mathbf{x}\|_\infty = \max_{\|\mathbf{y}\|_\infty = 1} \|\mathbf{D}^{-1}\mathbf{T}^{-1}\mathbf{A}\mathbf{T}\mathbf{D}\mathbf{y}\|_\infty \leq 1.$$

□

Satz 8.33 Das Mehrschrittverfahren (8.25) sei konsistent von der Ordnung p und für alle Startwerte und Störungen gelte $\|\boldsymbol{\eta}_i - \mathbf{y}(x_i)\| = \mathcal{O}(h^p)$ und $\|\boldsymbol{\varepsilon}_i(h)\| = \mathcal{O}(h^p)$. Im Streifen $S := \{(x, \mathbf{y}) : a \leq x \leq b, \mathbf{y} \in \mathbb{R}^n\}$ sei die Funktion $\mathbf{f}$ stetig, und die Inkrementfunktion erfülle (8.26) und genüge der Lipschitz-Bedingung

$$\|\Phi(x, \mathbf{u}_0, \ldots, \mathbf{u}_k, h) - \Phi(x, \mathbf{v}_0, \ldots, \mathbf{v}_k, h)\| \leq L \sum_{\ell=0}^{k} \|\mathbf{u}_\ell - \mathbf{v}_\ell\|$$

für alle $x \in [a, b]$, h und $\mathbf{u}_i, \mathbf{v}_i \in \mathbb{R}^n$. Dann ist das Mehrschrittverfahren (8.25) genau dann konvergent, wenn es nullstabil ist.

Beweis. Die Notwendigkeit der Nullstabilität wurde schon in Satz 8.31 gezeigt.

Zum Beweis der Hinlänglichkeit betrachte den Fehler $\mathbf{e}_i = \boldsymbol{\eta}_i - \mathbf{y}(x_i)$. Nach Voraussetzung gilt

$$\|\mathbf{e}_i\| = \mathcal{O}(h^p), \quad i = 0, 1, \ldots, k-1$$

und für alle $i \geq k$

$$\sum_{\ell=0}^{k} \alpha_\ell \mathbf{e}_{i+\ell} = h\Big(\Phi(x, \boldsymbol{\eta}_i, \boldsymbol{\eta}_{i+1}, \ldots, \boldsymbol{\eta}_{i+k}, h) - \Phi\big(x, \mathbf{y}(x_i), \mathbf{y}(x_{i+1}), \ldots, \mathbf{y}(x_k), h\big) + \boldsymbol{\varepsilon}_{i+k}(h) - \boldsymbol{\tau}(x, \mathbf{y}, h)\Big) =: \mathbf{g}_i, \quad (8.27)$$

wobei aus der Lipschitz-Bedingung folgt

$$\|\mathbf{g}_i\| \leq h\bigg(L \sum_{\ell=0}^{k} \|\mathbf{e}_{i+\ell}\| + \|\boldsymbol{\varepsilon}_{i+k}(h)\| + \|\boldsymbol{\tau}(x, \mathbf{y}, h)\|\bigg). \quad (8.28)$$

Mit Hilfe der Matrizen

$$\mathbf{E}_i^\star = \left[\begin{array}{c|c|c|c} \mathbf{e}_i & \mathbf{e}_{i+1} & \cdots & \mathbf{e}_{i+k-1} \end{array}\right] \in \mathbb{R}^{k\times n}, \quad \mathbf{G}_i^\star = \left[\begin{array}{c|c|c|c} \mathbf{0} & \cdots & \mathbf{0} & \mathbf{g}_i \end{array}\right] \in \mathbb{R}^{k\times n}$$

und

$$\mathbf{A} = \begin{bmatrix} 0 & 1 & & \\ & \ddots & \ddots & \\ & & 0 & 1 \\ -\alpha_0 & \cdots & -\alpha_{k-2} & -\alpha_{k-1} \end{bmatrix} \in \mathbb{R}^{k\times k}$$

können wir (8.27) schreiben als

$$\mathbf{E}_{i+1} = \mathbf{A}\mathbf{E}_i + \mathbf{G}_i, \quad i = 0, 1, \ldots.$$

Für Matrizen $\mathbf{B} = [\mathbf{b}_1|\mathbf{b}_2|\cdots|\mathbf{b}_n] \in \mathbb{R}^{k\times n}$ sei die Norm $|\!|\!|\mathbf{B}|\!|\!|$ gegeben durch

$$|\!|\!|\mathbf{B}|\!|\!| := \max_{i=1}^{n} \|\mathbf{b}_i\|_{\mathbf{R}},$$

wobei $\|\cdot\|_{\mathbf{R}}$ die in Lemma 8.32 eingeführte Norm bezeichnet. Es folgt

$$|\!|\!|\mathbf{E}_{i+1}|\!|\!| \leq |\!|\!|\mathbf{A}\mathbf{E}_i|\!|\!| + |\!|\!|\mathbf{G}_i|\!|\!| \leq \|\mathbf{A}\|_{\mathbf{R}} |\!|\!|\mathbf{E}_i|\!|\!| + |\!|\!|\mathbf{G}_i|\!|\!|.$$

Aus (8.28) folgt wegen der Äquivalenz von Normen in endlichdimensionalen Räumen

$$|\!|\!|\mathbf{G}_i|\!|\!| \leq \widetilde{c}Lh\bigg(\sum_{\ell=0}^{k} \|\mathbf{e}_{i+\ell}\| + \underbrace{\|\boldsymbol{\varepsilon}_{i+k}(h)\| + \|\boldsymbol{\tau}(x, \mathbf{y}, h)\|}_{=\mathcal{O}(h^p)}\bigg) \leq ch\big(|\!|\!|\mathbf{E}_i|\!|\!| + |\!|\!|\mathbf{E}_{i+1}|\!|\!| + h^p\big),$$

das heißt, für genügend kleines h gilt $1 - ch > 0$ und weiter

$$|\!|\!|\mathbf{E}_{i+1}|\!|\!| \leq \frac{1+ch}{1-ch} |\!|\!|\mathbf{E}_i|\!|\!| + \frac{ch^{p+1}}{1-ch}, \quad i = 0, 1, \ldots.$$

Wegen

$$\frac{1+x}{1-x} = (1+x)\sum_{i=0}^{\infty} x^i = 1 + 2\sum_{i=1}^{\infty} x^i = 1 + \frac{2x}{1-x} \leq 1 + 4x, \quad x \leq \frac{1}{2}$$

folgt hieraus

$$\begin{aligned}\|\mathbf{E}_{i+1}\| &\le \underbrace{(1+4ch)}_{=:M}\|\mathbf{E}_i\| + \underbrace{ch^{p+1}(1+4ch)}_{=:N}\\ &\le M^2\|\mathbf{E}_{i-1}\| + MN + N\\ &\vdots\\ &\le M^{i+1}\|\mathbf{E}_0\| + \{M^i + \cdots + M + 1\}N\\ &= M^{i+1}\|\mathbf{E}_0\| + \frac{M^{i+1}-1}{M-1}N.\end{aligned}$$

Mit $0 < 1 + x \le e^x$ für $x > -1$ ergibt sich schließlich

$$\|\mathbf{E}_i\| \le e^{4cih}\|\mathbf{E}_0\| + \frac{e^{4cih}-1}{4ch}ch^{p+1}(1+4ch) \le e^{4cih}\left(\|\mathbf{E}_0\| + \frac{1+4ch}{4}h^p\right),$$

was für festes $x \ge x_0$ bedeutet

$$\|\mathbf{E}_{x/h}\| \le e^{4c|x-x_0|}\left(\underbrace{\|\mathbf{E}_0\|}_{=\mathcal{O}(h^p)} + \frac{1+4ch}{4}h^p\right) = \mathcal{O}(h^p).$$

□

8.12 Steife Differentialgleichungen

Beispiel 8.34 Gegeben sei das System

$$\left.\begin{aligned} y_1' &= \frac{\lambda_1+\lambda_2}{2}y_1 + \frac{\lambda_1-\lambda_2}{2}y_2\\ y_2' &= \frac{\lambda_1-\lambda_2}{2}y_1 + \frac{\lambda_1+\lambda_2}{2}y_2\end{aligned}\right\} \quad \lambda_1, \lambda_2 < 0. \tag{8.29}$$

Die allgmeine Lösung für $x \ge 0$ lautet

$$\begin{aligned} y_1(x) &= C_1e^{\lambda_1 x} + C_2e^{\lambda_2 x}\\ y_2(x) &= C_1e^{\lambda_1 x} - C_2e^{\lambda_2 x}\end{aligned} \tag{8.30}$$

mit $C_1, C_2 \in \mathbb{R}$.

Berechnet man (8.29) mit dem Eulerschen Polygonzug-Verfahren, so lässt sich die numerische Näherung geschlossen darstellen

$$\begin{aligned}\eta_{1,i} &= C_1(1+h\lambda_1)^i + C_2(1+h\lambda_2)^i\\ \eta_{2,i} &= C_1(1+h\lambda_1)^i - C_2(1+h\lambda_2)^i.\end{aligned}$$

Offensichtlich konvergieren die Lösungen nur dann, falls die Schrittweite h so klein gewählt wird, dass

$$|1+h\lambda_1| < 1, \quad |1+h\lambda_2| < 1. \tag{8.31}$$

Es sei nun $|\lambda_2|$ groß gegen $|\lambda_1|$. Wegen $\lambda_2 < 0$ ist dann in (8.30) der Einfluss der Komponente $e^{\lambda_2 x}$ gegenüber $e^{\lambda_1 x}$ vernachlässigbar klein. Leider gilt das nicht für das numerische Verfahren. Wegen (8.31) muss nämlich die Schrittweite so klein gewählt werden, dass

$$h < \frac{2}{|\lambda_2|}.$$

Für den Fall $\lambda_1 = -1$ und $\lambda_2 = -1000$ bedeutet dies $h \leq 0.002$. Obwohl also e^{-1000x} zur Lösung praktisch nichts beiträgt, bestimmt der Faktor 1000 im Exponenten die Schrittweite. Dieses Verhalten bezeichnet man als *steif*. △

Definition 8.35 Die lineare Differentialgleichung

$$\mathbf{y}' = \mathbf{A}\mathbf{y} + \mathbf{b}(x), \quad \mathbf{A} \in \mathbb{R}^{n\times n}, \quad \mathbf{b}(x) \in \mathbf{R}^n$$

heißt **steif**, wenn $\mathrm{Re}(\lambda_i) < 0$ für alle Eigenwerte λ_i von $\mathbf{A}$ und $\min_{i=1}^n\{\mathrm{Re}(\lambda_i)\} \ll \max_{i=1}^n\{\mathrm{Re}(\lambda_i)\}$. Der Quotient

$$q := \frac{\max_{i=1}^n\{\mathrm{Re}(\lambda_i)\}}{\min_{i=1}^n\{\mathrm{Re}(\lambda_i)\}}$$

heißt **Steifigkeitsquotient**.

Das Anwenden eines Einschrittverfahrens auf das Problem

$$y' = \lambda y \quad \text{mit} \quad \lambda < 0$$

führt auf eine Rekursion

$$\eta_{i+1} = g(h\lambda)\eta_i, \quad i = 0, 1, 2, \ldots,$$

wobei die *Stabilitätsfunktion* g vom jeweiligen Verfahren abhängt.

Beispiel 8.36 Beim Eulerschen Polygonzug-Verfahren folgt gemäß Beispiel 8.34 $g(z) = 1 + z$. Beim klassischen Runge-Kutta-Verfahren gilt

$$\begin{aligned} k_1 &= \eta_i z \\ k_2 &= \eta_i(z + z^2/2) \\ k_3 &= \eta_i(z + z^2/2 + z^3/4) \\ k_4 &= \eta_i(z + z^2 + z^3/2 + z^4/4) \\ \eta_{i+1} &= \eta_i + k_1/6 + k_2/3 + k_3/3 + k_4/6, \end{aligned}$$

dies bedeutet,

$$g(z) = 1 + z + \frac{z^2}{2} + \frac{z^3}{6} + \frac{z^4}{24}.$$

Beim impliziten Euler-Vefahren erhält man aus

$$\eta_{i+1} = \eta_i + hf(x_{i+1}, \eta_{i+1}) = \eta_i + z\eta_{i+1}$$

dass

$$g(z) = \frac{1}{1-z}.$$

Die Trapez-Methode

$$\eta_{i+1} = \eta_i + \frac{h}{2}\big(f(x_i, \eta_i + f(x_{i+1}, \eta_{i+1})\big) = \eta_i\left(1 + \frac{z}{2}\right) + \frac{z}{2}\eta_{i+1}$$

induziert schließlich

$$g(z) = \frac{1 + z/2}{1 - z/2} = \frac{2+z}{2-z}.$$

△

Für $\lambda < 0$ folgt nun $e^{\lambda x} \overset{x\to\infty}{\longrightarrow} 0$, während $\eta_i \to 0$ genau dann gilt, wenn $|g(z)| < 1$. Da bei den beiden expliziten Verfahren $g(z)$ ein Polynom ist, gilt

$$|g(z)| \to \infty \quad \text{für} \quad z \to -\infty.$$

Daher sind diese Verfahren nur stabil, falls h klein genug ist.
Demgegenüber ist bei den beiden impliziten Verfahren $g(z)$ eine rationale Funktion mit der Eigenschaft

$$|g(z)| < 1 \quad \text{für} \quad z < 0.$$

Allerdings sieht man, dass die Dämpfung beim impliziten Euler-Verfahren für betragsgroße Werte $z = h\lambda$ sehr viel stärker als bei der Trapez-Regel ist. Bei letzterer nähert man sich der Stabilitätsgrenze $|g(z)| \overset{z\to-\infty}{\longrightarrow} 1$.

Definition 8.37 Der Bereich

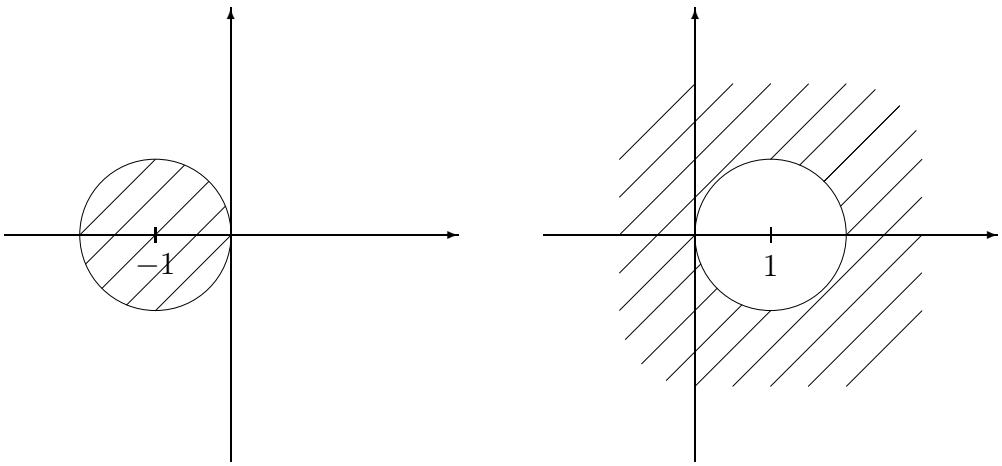
$$\mathcal{M} := \{z \in \mathbb{C} : |g(z)| < 1\}$$

heißt **Stabilitätsgebiet**. Ein Verfahren heißt **absolut stabil** oder ***A*-stabil**, falls

$$\{z \in \mathbb{C} : \mathrm{Re}(z) < 0\} \subset \mathcal{M},$$

wenn also $|g(z)| < 1$ gilt für alle $\mathrm{Re}(z) < 0$.

Beispiel 8.38



Stabilitätsgebiet des expliziten Euler-Vefahren

Stabilitätsgebiet des impliziten Euler-Vefahren

△

Das A-stabile implizite Euler-Verfahren besitzt uneingeschränkte Dämpfung für $z = h\lambda \to -\infty$, während die Stabilitätsfunktion der A-stabilen Trapez-Methode betragsmäßig gegen 1 strebt, also letzlich die Dämpfung verliert, was sich mit Rundungseffekten durchaus stark auswirken kann.

Definition 8.39 Ein Verfahren heißt ***L*-stabil**, falls es A-stabil ist und zusätzlich gilt

$$g(-\infty) = 0.$$

Bemerkungen:

1. Explizite Ein- oder Mehrschrittverfahren sind niemals A-stabil.
2. Die Trapez-Methode ist A-stabil, das implizite Euler-Verfahren sogar L-stabil.
3. Das SDIRK2- und Radau3-Verfahren aus Beispiel 8.17 sind beide L-stabil.
4. Ein A-stabiles Mehrschrittverfahren besitzt höchstens die Konsistenzordnung $p = 2$.

8.13 Parabolische Differentialgleichungen

Wir betrachten die Wärmeleitungsgleichung

$$\frac{\partial}{\partial t}u(t,\mathbf{x}) - \Delta u(t,\mathbf{x}) = f(t,\mathbf{x}), \quad (t,\mathbf{x}) \in [0,T] \times \Omega$$

für ein Gebiet $\Omega \subset \mathbb{R}^d$ und einen Endzeitpunkt T. Der Einfachheit halber geben wir uns homogene Dirichlet-Randwerte vor

$$u(\cdot,\mathbf{x}) = 0 \quad \text{auf } \Gamma = \partial\Omega$$

Um diese Gleichung numerisch zu lösen, führen wir zunächst eine Semidiskretisierung im Raum durch, indem wir wie in Kapitel 5 vorgehen. Dazu sei $\Omega_h \cup \Gamma_h$ ein Gitter von $\overline{\Omega}$ und

$$u_h(t,\mathbf{x}) : [0,T] \times \overline{\Omega}_h \to \mathbb{R}$$

eine zugehörige Gitterfunktion. Die semidiskrete Gleichung

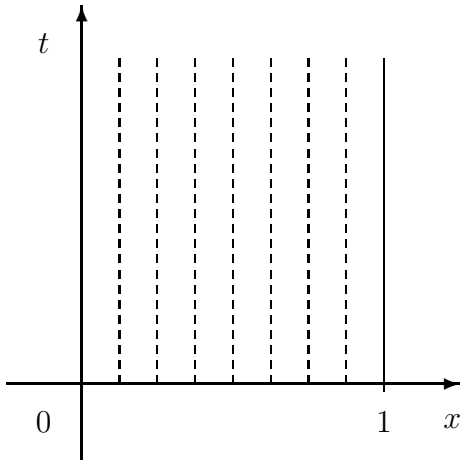
$$\begin{aligned} \frac{\partial}{\partial t}u_h(t,\mathbf{x}) + \big(\Delta_h u_h\big)(t,\mathbf{x})\big) &= f(t,\mathbf{x}) \quad \text{für alle } \mathbf{x} \in \Omega_h, \\ u_h(t,\mathbf{x}) &= 0 \qquad\quad \text{für alle } \mathbf{x} \in \Gamma_h, \end{aligned}$$

mit der Anfangsbedingung $u_h(0,\mathbf{x}) = g(\mathbf{x})$ für alle $\mathbf{x} \subset \Omega_h$ entspricht einem System gewöhnlicher Differentialgleichungen

$$\frac{\partial}{\partial t}\mathbf{u}(t) + \mathbf{A}\mathbf{u}(t) = \mathbf{f}(t), \quad \mathbf{u}(0) = \mathbf{g}. \tag{8.32}$$

Dabei ist die Matrix $\mathbf{A}$ unabhängig von der Zeit und die Elliptizität des Laplace-Operators impliziert, dass alle Eigenwerte von $\mathbf{A}$ einen positiven Realteil besitzten. Folglich ist die Differentialgleichung (8.32) steif.

Bemerkung: Aus anschaulichen Gründen wird die Semidiskretisierung (8.32) *Linienmethode* genannt. Für alle $t \geq 0$ enthält nämlich die vektorwertige Funktion $\mathbf{u}(t)$ die Funktionswerte der Approximation $u_h(t,\mathbf{x})$ in den Gitterpunkten.



Linienmethode für $\Omega = (0, 1)$

Wir benötigen noch eine geeignete Zeitdiskretisierung. Das explizite Euler-Verfahren führt auf

$$\frac{\mathbf{u}_{i+1} - \mathbf{u}_i}{\Delta t} + \mathbf{A}\mathbf{u}_i = \mathbf{f}(t_i) \quad \text{bzw.} \quad \Delta t\, \mathbf{u}_{i+1} = (\mathbf{I} - \Delta t\, \mathbf{A})\mathbf{u}_i + \Delta t\, \mathbf{f}(t_i). \tag{8.33}$$

Hingegen liefert das implizite Euler-Verfahren

$$\frac{\mathbf{u}_{i+1} - \mathbf{u}_i}{\Delta t} + \mathbf{A}\mathbf{u}_{i+1} = \mathbf{f}(t_{i+1}) \quad \text{bzw.} \quad (\mathbf{I} + \Delta t\, \mathbf{A})\mathbf{u}_{i+1} = \mathbf{u}_i + \Delta t\, \mathbf{f}(t_{i+1}). \tag{8.34}$$

Der Startwert lautet in beiden Fällen $\mathbf{u}_0 = \mathbf{g}$. Kombinieren wir (8.33) und (8.34), so erhalten wir das *θ-Schema*

$$\frac{\mathbf{u}_{i+1} - \mathbf{u}_i}{\Delta t} + (1-\theta)\mathbf{A}\mathbf{u}_i + \theta\mathbf{A}\mathbf{u}_{i+1} = (1-\theta)\mathbf{f}(t_i) + \theta\mathbf{f}(t_{i+1}),$$

beziehungsweise

$$(\mathbf{I} + \Delta t\, \theta\, \mathbf{A})\mathbf{u}_{i+1} = \big(\mathbf{I} - \Delta t\, (1-\theta)\, \mathbf{A}\big)\mathbf{u}_i + (1-\theta)\mathbf{f}(t_i) + \theta\mathbf{f}(t_{i+1}),$$

Dabei gilt

$$\theta = \begin{cases} 0, & \text{explizites Euler-Verfahren} \\ 1/2, & \text{Trapez-Methode} \\ 1, & \text{implizites Euler-Verfahren} \end{cases}$$

wobei die Trapez-Methode im Zusammenhang mit parabolischen Differentialgleichungen auch *Crank-Nicolson-Verfahren* genannt wird. Das θ-Schema ist konsistent von erster Ordnung, im Falle $\theta = 0.5$ sogar von zweiter Ordnung. Es ist A-stabil für alle $\theta \in [0.5, 1]$ und L-stabil für alle $\theta \in (0.5, 1]$.

Bemerkung: Da das Crank-Nicolson-Verfahren das einfachste Verfahren von zweiter Ordnung ist, ist es unheimlich populär. Aufgrund der fehlenden L-Stabilität produziert es aber unphysikalische Oszillationen und sollte daher gemieden werden. Am besten wählt man $\theta = 1/2 + \xi$ mit einem kleinen ξ.

Ein geeignetes, weil L-stabiles, Verfahren höherer Ordnung ist das Radau3-Verfahren aus Beispiel 8.17. Mit $t_{i+1/2} := t_i + \Delta t/3$ lautet es

$$\begin{bmatrix} \mathbf{I} + \frac{5\Delta t}{12}\mathbf{A} & -\frac{\Delta t}{12}\mathbf{A} \\ \frac{3\Delta t}{4}\mathbf{A} & \mathbf{I} + \frac{\Delta t}{4}\mathbf{A} \end{bmatrix} \begin{bmatrix} \mathbf{k}_1 \\ \mathbf{k}_2 \end{bmatrix} = \begin{bmatrix} \mathbf{f}(t_{i+1/2}) - \mathbf{A}\mathbf{u}_i \\ \mathbf{f}(t_{i+1}) - \mathbf{A}\mathbf{u}_i \end{bmatrix}, \quad \mathbf{u}_{i+1} = \mathbf{u}_i + \frac{3\Delta t}{4}\mathbf{k}_1 + \frac{\Delta t}{4}\mathbf{k}_2.$$

Jedoch muss hier in jedem Schritt ein Gleichungssystem gelöst werden, das doppelt so viele Unbekannte wie Ortsunbekannte enthält.

Letzteres ist der Grund für die Klasse der SDIRK-Verfahren, vergleiche Beispiel 8.17. Das SDIRK2-Verfahren führt auf

$$\begin{aligned}(\mathbf{I}+\gamma\Delta t\mathbf{A})\mathbf{k}_1 &= \mathbf{f}(t_{i+1/2}) - \mathbf{A}\mathbf{u}_i\\ (\mathbf{I}+\gamma\Delta t\mathbf{A})\mathbf{k}_2 &= \mathbf{f}(t_{i+1}) - \mathbf{A}\mathbf{u}_i - (1-\gamma)\Delta t\mathbf{k}_1\\ \mathbf{u}_{i+1} &= \mathbf{u}_i + (1-\gamma)\Delta t\mathbf{k}_1 + \gamma\Delta t\mathbf{k}_2,\end{aligned}$$

wobei $t_{i+1/2} := t_i + \gamma\Delta$ und $\gamma := (2-\sqrt{2})/2$ ist. Es müssen nun in jedem Schritt zwei Gleichungssysteme gelöst werden, die allerdings nur soviel Unbekannte wie Ortsunbekannte besitzen. Speziell ist die Systemmatrix immer dieselbe, weshalb beispielsweise nur eine einzige Cholesky-Zerlegung berechnet werden muss.

Index

3/8-Regel, 144, 181
LR-Verfahren, 77
LR-Zerlegung, 20
QR-Verfahren, 77
    mit Shift, 82
θ-Schema, 199
Überschuss
    hierarchischer, 147

Abstand, 9
Abstiegsbedingung, 60
Adams-Bashforth-Verfahren, 185, 187
Adams-Moulton-Verfahren, 186, 188
Algorithmus
    QR-Verfahren, 77
    QR-Verfahren mit Shift, 82
    QR-Zerlegung, 38
    hp-Quadratur, 160
    Adams-Bashforth-Verfahren, 187
    adaptive Quadratur, 147
    Arnoldi-Prozess, 87
    Bisektionsverfahren, 137
    CG-Verfahren, 50
    CGLS-Verfahren, 52
    Eulersches Polygonzug-Verfahren, 175
    Gauß-Newton–Verfahren, 57
    GMRES-Verfahren, 104
    Gradientenverfahren, 55
    Lanczos-Prozess, 85
    Levenberg-Marquardt-Verfahren, 60
    lineares Mehrschrittverfahren, 185
    Mehrgitterverfahren, 113
    MINRES-Verfahren, 106
    Monte-Carlo-Quadratur, 165
    PCG-Verfahren, 109
    Rayleigh-Quotient-Iteration, 75
    von Clenshaw, 124
    von Golub und Welsch, 158
    von Mises-Potenzmethode, 72
    Zweigitterverfahren, 113
Anfangsbedingung, 170
Armijo-Goldstein-Bedingung, 55
Arnoldi-Prozess, 87
Ausgleichslösung, 31
Ausgleichsproblem, 31
Auslöschung, 15

Basis
    hierarchische, 146
    nodale, 146
Bernoulli-Polynom, 148
Bernoulli-Zahlen, 148
Bilinearform
    symmetrische, 115
Bisektionsverfahren, 137
Butcher-Tableau, 180

CG-Verfahren, 50
CGLS-Verfahren, 52
Choleski-Zerlegung, 27
Christoffel-Darboux-Identität, 131
Christoffel-Funktion, 129
Clenshaw-Algorithmus, 124
Crank-Nicolson-Verfahren, 199

Dahlquistsche Wurzelbedingung, 190
Deflation, 83
Differentialgleichung
    elliptische, 90
    gewöhnliche, 170
    hyperbolische, 90
    parabolische, 90
    steife, 196
Differentialoperator, 90
    elliptischer, 90
    hyperbolischer, 90
    parabolischer, 90
Differenz
    linksseitige, 95
    rechtsseitige, 95

zentrale, 95
Differenzengleichung
lineare, 189
Differenzenstern, 96
5-Punkte-Stern, 96
für beliebigen Differentialoperator, 99
Differenzenverfahren, 98, 99
Diskretisierungsfehler
globaler, 178
lokaler, 176, 186
Distanz, 9

Einschrittverfahren, 175
explizites, 175
implizites, 175
Einzelschrittverfahren, 113
Energienorm, 50
eps, 8
Euler-Collatz-Verfahren, 177
Euler-Maclaurinsche Summenformel, 149
Euler-Verfahren
explizites, 176
implizites, 176
Eulersches Polygonzug-Verfahren, 175
Exaktheitsgrad, 139
Exponent, 7

Fluch der Dimension, 163
Frobenius-Begleitmatrix, 68
Funktion
harmonische, 95

Gauß-Newton-Verfahren, 57
Gauß-Quadratur, 155
Gebiet, 88
diskret zusammenhängend, 100
gebrochene Iteration, 73
Gesamtpivotsuche, 17
Gesamtschrittverfahren, 111
Gewichtsfunktion, 116, 125
Gills-Formel, 181
Gitter, 96
Gitterfunktion, 98
Gitterpunkt, 96
randferner, 96
randnaher, 96
Givens-Rotation, 81
Gleichung
Laplace-, 88
Poisson-, 89
Potential-, 88
Wärmeleitungs-, 89, 198
Wellen-, 90
Gleichungssystem
überbestimmt, 30
unterbestimmt, 30
GMRES-Verfahren, 103
Gradientenverfahren, 54
Grobgitter
-korrektur, 113
-löser, 113

Hebden-Verfahren, 63
Hessenberg-Form
obere, 79
hierarchische Basis, 146
Householder-Transformation, 34
Hutfunktion, 146

Inkrementfunktion, 175
Innenprodukt, 115
Interpoland
hierarchischer, 147
Interpolation
Lagrange-, 141
inverse Iteration, 73
Iteration
gebrochene, 73
inverse, 73

Kernpolynom, 129
kleinste-Quadrate-Lösung, 31
Knoten, 141
-polynom, 141
Kondition, 15
des Eigenwertproblems, 68
einer Matrix, 16
Konditionszahl
absolute, 15
relative, 15
Konistenz, 101
Konsistenz, 176, 186
Konvergenz, 101, 178, 192
exponentielle, 154
Kronecker-Produkt, 110
Krylov-Raum, 49

Lösung
klassische, 95

Lagrange
-Grundpolynome, 141
Lanczos-Prozess, 85
Laplace-Gleichung, 88
Laplace-Operator, 88, 89
Legendre-Polynom, 126
Linienmethode, 198
Linkseigenvektor, 70

Mantisse, 7
Maschinengenauigkeit, 8
Maschinenzahl, 7
Matrix
diagonalisierbare, 69
Frobenius-Begleitmatrix, 68
Gramsche, 119
normale, 70
Permutations-, 22
unitär, 34
Matrixnorm, 9
Frobeniusnorm, 9
induzierte, 11
Spaltensummennorm, 9
Spektralnorm, 13
submultiplikative, 11
verträgliche, 11
Zeilensummennorm, 9
Maximumprinzip, 92
diskretes, 100
Mehrgitterverfahren, 113
Mehrschrittverfahren, 185
explizites, 185
implizites, 185
lineares, 185
Milne-Regel, 144
Minimumprinzip, 93
MINRES-Verfahren, 105
Mittelpunktsregel, 153, 182
Moore-Penrose-Inverse, 44

Newton-Cotes-Formel, 143
zusammengesetzte, 144
nodale Basis, 146
Norm, 9
Energie-, 50
induzierte, 11
Normalengleichungen, 31

Operator
Laplace-, 88, 89
Ordnung
der Quadratur, 145
Orthogonalbasis, 117
Orthonormalbasis, 117
Overflow, 8

PCG-Verfahren, 109
Permutationsmatrix, 22
Pivotelement, 17, 18
Pivotisierung
partielle, 17
totale, 17
Poisson-Gleichung, 89
Polynom
Bernoulli-, 148
Legendre-, 126
Tschebyscheff-, 51, 122
Potentialgleichung, 88
Potenzmethode, 72
Prädiktor-Korrektor-Verfahren, 188
Problem
sachgemäß gestellt, 91
schlecht gestellt, 91
Prolongation, 112
Pseudoinverse, 44

Quadratur
3/8-Regel, 144
hp-, 160
-formel, 138
-gewicht, 138
-knoten, 138
-ordnung, 145
-punkt, 138
adaptive, 147
Fehlerabschätzung, 139
Gauß-Formel, 155
Milne-Regel, 144
Mittelpunktsregel, 153
Monte-Carlo-Quadratur, 165
Simpson-Regel, 143
Stabilität, 138
Tensorprodukt-Formel, 162
Weddle-Regel, 144
zusammengesetzte Fomel, 144

Radau3-Verfahren, 182, 199
Randbedingung

Dirichlet-, 91
Neumann-, 91
Randpunkt, 96
Rayleigh-Quotient, 66
Rayleigh-Quotient-Iteration, 75
Rechtseigenvektor, 70
Referenzintervallen, 140
Regel
3/8-, 144, 181
Milne-, 144
Simpson-, 143
Weddle-, 144
Residuum, 31
Restriktion, 112
Richardson-Iteration, 111
Romberg-Verfahren, 151
Rundung, 8
Runge-Kutta-Verfahren, 180
explizites, 180
implizites, 181
klassisches, 181

Satz
von Bauer und Fike, 69
von Bendixson, 67
von Gerschgorin, 65
von Sturm, 137
Schema
θ-, 199
schiefhermitesch, 66
Schurkomplement, 26
SDIRK2-Verfahren, 182
Shift-Strategie, 82
Shortley-Weller-Approximation, 97
Simpson-Regel, 143
zusammengesetzte, 145
Skalarprodukt, 115
Sobolev-Raum, 117
Spaltenpivotsuche, 17
Spektrum, 65
Stützstelle, 141
Stabilität, 17, 101
A-, 197
L-, 198
Stabilitätsgebiet, 197
Steifigkeitsquotient, 196
Stirlingsche Formel, 129

Tensorprodukt, 110
-Quadraturformel, 162
Transformation
Householder, 34
Trapezregel, 143
zusammengesetzte, 144
Trust-Region-Verfahren, 59
Tschebyscheff-Polynom, 122
Tschebyscheff-Polynome, 51

Underflow, 8

V-Zyklus, 114
Vektoren
konjugierte, 48
Vektornorm, 9
Betragssummennorm, 9
Euklidnorm, 9
Maximumnorm, 9
Verfahren
QR-, 77
QR- mit Shift, 82
Adams-Bashforth-, 185, 187
Adams-Moulton-, 186, 188
Bisektions-, 137
CGLS-, 52
Crank-Nicolson-, 199
der konjugierten Gradienten, 50
der konjugierten Gradienten mit Vorkonditionierung, 109
des steilsten Abstiegs, 54
Differenzen-, 98, 99
Einschritt-, 175
Einzelschritt-, 113
Euler-
explizites, 176
implizites, 176
Euler-Collatz-, 177
Eulersches Polygonzug-, 175
Gauß-Newton-, 57
Gesamtschritt-, 111
GMRES-, 103
Hebden-, 63
Levenberg-Marquardt-, 60
lineares Mehrschritt-, 185
Mehrgitter-, 113
Mehrschritt-, 185
MINRES-, 105
Prädiktor-Korrektor-, 188
Radau3-, 182, 199

Romberg-, 151
Runge-Kutta-, 180
explizites, 180
implizites, 181
klassisches, 181
SDIRK2-, 182
Trust-Region-, 59
von Heun, 177
von Mises-, 72
Zweigitter-, 113
Vergleichsprinzip, 93
von Mises-Potenzmethode, 72
von Mises-Verfahren, 72
Vorkonditionierung, 108

W-Zyklus, 114
Wärmeleitungsgleichung, 88, 89, 198
Weddle-Regel, 144
Wellengleichung, 90
Wertebereich, 66
Wurzelbedingung
von Dahlquist, 190

Zweigitterverfahren, 113