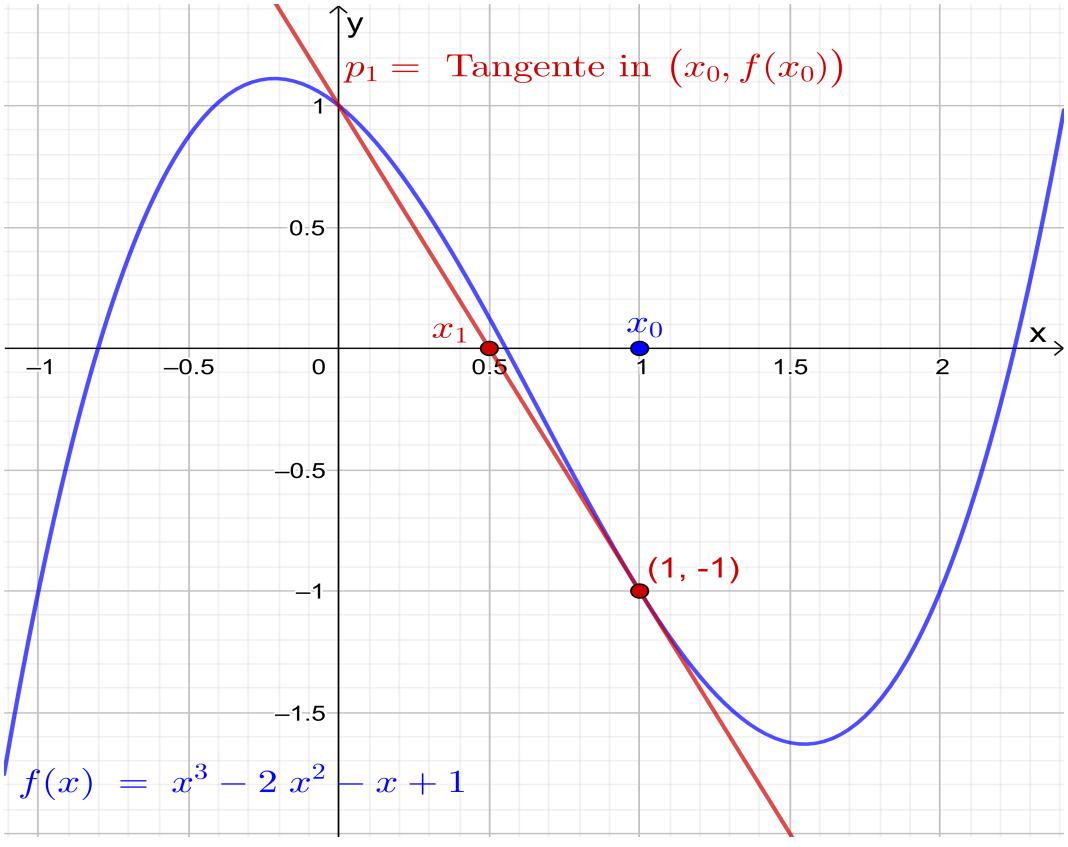
UNIVERSITÄT PADERBORN
Die Universität der Informationsgesellschaft

Mathematik 4 für Maschinenbau: Numerische Methoden

Kerstin Hesse

Universität Paderborn, Sommersemester 2020
(aktualisierte Version – 14. November 2020)



Erster Schritt im Newton-Verfahren zur Nullstellenbestimmung

Dieses Skript wurde in seiner ersten Version für das Sommersemester 2020 erstellt.

Bei der Aktualisierung des Skripts im November 2020 wurden die Inhalte einiger theorielastiger Aufgaben in das Skript integriert.

Paderborn, November 2020

Kerstin Hesse

Einleitung

Bei dem Thema **Numerik** oder **Numerische Methoden** geht es darum, wie man mathematische Probleme (angenähert) mit einem Computer löst. Hier sind zwei Beispiele, an denen deutlich wird, worum es geht und warum dieses interessant ist und für alle Anwender sehr wichtig ist:

Beispiel 1: Lösen linearer Gleichungssysteme. Angenommen, man hat ein lineares Gleichungssystem mit 100.000 Unbekannten. Sicher möchte man dieses nicht per Hand lösen, sondern ein geeignetes numerisches Verfahren (einen Algorithmus) auf einem Computer dazu nutzen. Dabei erhält man als Ergebnis oft nur eine Annäherung an die exakte Lösung, denn es treten einerseits Rundungsfehler auf und andererseits wird man zur Lösung oft kein direktes sondern ein iteratives Verfahren verwenden (dieses berechnet eine Folge von Näherungen der Lösung), welches man abbricht, wenn die Lösung hinreichend gut angenähert worden ist.

Beispiel 2: Funktion finden, die einen Datensatz beschreibt. Angenommen, die Temperatur y wurde an einem bestimmten Ort stündlich ein Jahr lang gemessen. Dann erhalten wir $365 \cdot 24 = 8.760$ Messdaten $(t_i; y_i)$, $1 = 1, 2, \ldots, 8.760$, wobei y_i die zum Zeitpunkt t_i gemessene Temperatur ist. Man möchte nun gerne eine Funktion $y(t)$ bestimmten, deren Graph (genau oder auch nur angenähert) durch alle Datenpunkte $(t_i; y_i)$, $i = 1, 2, \ldots, 8.760$, geht. Es stellt sich die Frage, wie gut diese Funktion die Temperatur $y(t)$ zu anderen Zeitpunkten t beschreibt.

Das erste Bespiel gehört in den Bereich der Numerischen Linearen Algebra, und das zweite Beispiel gehört in den Bereich der Numerischen Analysis. Beide Teilgebiete zusammen bilden die Numerik. In dieser Vorlesung werden wir einen kleinen **Querschnitt durch die wichtigsten Themen der Numerik** behandeln.

Dieses detailliert ausgearbeitete Skript können (und sollten) Sie wie ein Lehrbuch verwenden. Es folgt der Vorlesung ganz genau und enthält häufig zusätzliche Erklärungen und weitere Beispiele.

Ich freue mich auf Ihre Teilnahme an der „Mathematik 4 für Maschinenbau: Numerische Methoden“!

Kerstin Hesse April 2020

Literaturverzeichnis

Bei der Erstellung dieses Skripts wurde die unten aufgelistete Literatur verwendet:

[1] Kendall Atkinson, Weimin Han: Elementary Numerical Analysis (3. Auflage). John Wiley & Sons, Inc., 2004.

[2] Sören Bartels: Numerik 3x9: Drei Themengebiete in jeweils neun kurzen Kapiteln. Springer-Verlag, Berlin, Heidelberg, 2016.

[3] Günter Bärwolff: Numerik für Ingenieure, Physiker und Informatiker (2. Auflage). Springer-Verlag, Berlin, Heidelberg, 2016.

[4] Kerstin Hesse: Höhere Mathematik A für Elektrotechniker. Vorlesungsskript, Universität Paderborn, 2018.

[5] Kerstin Hesse: Höhere Mathematik B für Elektrotechniker. Vorlesungsskript, Universität Paderborn, 2019.

[6] Kerstin Hesse: Höhere Mathematik C für Elektrotechniker. Vorlesungsskript, Universität Paderborn, 2019.

[7] Kerstin Hesse: Numerical Linear Algebra. Vorlesungsskript, University of Sussex, 2010.

[8] Wolfgang Dahmen, Arnold Reusken: Numerik für Ingenieure und Naturwissenschaftler (2. Auflage). Springer-Verlag, Berlin, Heidelberg, 2008.

[9] Kshitij Kulshreshtha: Numerische Methoden für Maschinenbauer. Skript, Universität Padeborn, 2018.

[10] H. W. Lang: IEEE-Gleitkomma-Format. Online-Resource: https://www.inf.hs-flensburg.de/lang/informatik/ieee-format.htm

[11] Sebastian Peitz: Mathematik 4 für Maschinenbau – Numerische Methoden. Vorlesungsfolien, Universität Padeborn, 2019.

[12] Rainer Kress: Numerical Analysis. Springer-Verlag, New York, 1998.

[13] Hans Rudolf Schwarz: Numerische Mathematik (4. Auflage). B. G. Teubner, Stuttgart, 1997.

[14] J. Stoer, R. Bulirsch: Introduction to Numerical Analysis (3. Auflage). Springer-Verlag, New York, 2002.

Inhaltsverzeichnis

*Dieses Teilkapitel wird nicht behandelt und ist damit auch nicht klausurrelevant.

*Dieses Teilkapitel wird nicht behandelt und ist damit auch nicht klausurrelevant.

KAPITEL 1

Maschinenzahlen, Rundung, Fehler und Kondition

1.1 Gleitkommadarstellung von Zahlen

Im Dezimalsystem hat die Zahl $x = \frac{2}{3}$ die Darstellung

$$x = 0{,}\overline{6} = 0{,}6666\ldots = \sum_{j=1}^{\infty} 6 \cdot 10^{-j}$$

mit der von 0 verschiedenen Periode 6. Auch wenn man das Dualsystem (Basis 2), Oktalsystem (Basis 8) oder Hexadezimalsystem (Basis 16) verwendet, hat diese Zahl eine Reihendarstellung, die nicht abbricht.

Ist $b > 2$ eine beliebige natürliche Zahl, so kann man jede reelle Zahl x **bzgl. der Basis** b wie folgt als **Gleitkommazahl** darstellen:

$$\boxed{x = \underbrace{(-1)^{\nu}}_{=\sigma}\, b^N \underbrace{\sum_{j=1}^{\infty} c_j\, b^{-j}}_{=a},} \tag{1.1}$$

wobei $\nu \in \{0; 1\}$ (über $\sigma = (-1)^{\nu} \in \{-1; 1\}$) das **Vorzeichen** bestimmt und $N \in \mathbb{Z}$ der **Exponent** ist. Die Zahlen c_j, $j = 0, 1, 2, \ldots$, liegen in der Menge $\{0; 1; 2; \ldots; b-1\}$ und sind die **Ziffern** (oder **Koeffizienten**) von x (bzgl. der Basis b), und $a = \sum_{j=1}^{\infty} c_j\, b^{-j}$ wird als die **Mantisse** von x bezeichnet.

Die Zahl $x = \frac{2}{3}$ hat z.B. die Gleitkommadarstellungen bzgl. der Basis $b = 10$

$$x = (-1)^0 \, 10^0 \sum_{j=1}^{\infty} 6 \cdot 10^{-j} \qquad \text{und} \qquad x = (-1)^0 \, 10^1 \sum_{j=2}^{\infty} 6 \cdot 10^{-j},$$

und bzgl. der Basis $b = 2$ hat $x = \frac{2}{3}$ die Gleitkommadarstellungen

$$x = (-1)^0 \, 2^0 \sum_{\ell=1}^{\infty} 1 \cdot 2^{-(2\ell-1)} \qquad \text{und} \qquad x = (-1)^0 \, 2^1 \sum_{\ell=1}^{\infty} 1 \cdot 2^{-2\ell}.$$

(Im Dualsystem ist also jeder zweite Koeffizient gleich null.) Bzgl. beider Basen ist jeweils in der zweiten Darstellung der Koeffizient $c_1 = 0$, und $j = 1$ taucht daher nicht in der Reihe auf.

Wir sehen an den Beispielen, dass die Gleitkommadarstellung nicht eindeutig bestimmt ist. Daher trifft man die **Konvention, dass in** (1.1) **immer** $c_1 \neq 0$ **gelten soll**, und nennt dieses die **normalisierte Gleitkommadarstellung zur Basis** b. Dass $c_1 \neq 0$ ist, bedeutet, dass in der Mantisse a die erste Nachkommastelle ungleich null sein muss, und das somit $a \in [b^{-1}, 1[$ gilt.

Beispiel 1.1. (normalisierte Gleitkommadarstellung zur Basis b)
$x = \frac{2}{3}$ hat bzgl. der Basis $b = 10$ bzw. bzgl. der Basis $b = 2$ jeweils die folgende normalisierte Gleitkommadarstellung

$$x = (-1)^0 \, 10^0 \sum_{j=1}^{\infty} 6 \cdot 10^{-j} \qquad \text{bzw.} \qquad x = (-1)^0 \, 2^0 \sum_{\ell=1}^{\infty} 1 \cdot 2^{-2\ell-1}.$$

Die Zahlen $y = 5367{,}23$ und $z = -0{,}00123$ haben bzgl. der Basis $b = 10$ die folgenden normalisierten Gleitkommadarstellungen

$$y = (-1)^0 \, 10^4 \left(5 \cdot 10^{-1} + 3 \cdot 10^{-2} + 6 \cdot 10^{-3} + 7 \cdot 10^{-4} + 2 \cdot 10^{-5} + 3 \cdot 10^{-6}\right)$$

und $z = (-1)^1 \, 10^{-2} \left(1 \cdot 10^{-1} + 2 \cdot 10^{-2} + 3 \cdot 10^{-3}\right)$. ♠

1.2 Gleitkomma-Zahlenformat auf Computern

Im Computer kann man Zahlen wie $x = \frac{2}{3}$, die durch eine Gleitkommadarstellung (1.1) mit einer unendlichen Reihe gegeben sind, nicht exakt darstellen, denn im

Computer kann nur eine endliche Anzahl der Koeffizienten $c_1, c_2 \ldots$ in (1.1) abgespeichert werden. Der Computer stellt Zahlen **bzgl. der Basis** b **als Maschinenzahlen** dar, wobei Vorzeichen, Mantisse und Exponent abgespeichert werden und **der Exponent und die Mantisse jeweils eine endliche Länge** haben. Die Länge des Exponenten und der Mantisse sind durch das Maschinenzahlformat, genauer durch den in dem Maschinenzahlformat dafür vorgesehenen Speicherplatz, beschränkt.

Definition 1.2. (Maschinenzahlen/Gleitkomma-Zahlensystem)
Seien $b \in \mathbb{N}$ *mit* $b \geq 2$ *eine Basis,* $p \in \mathbb{N}$ *eine Mantissenlänge und* $N_{\min}, N_{\max} \in \mathbb{N}$ *mit* $N_{\min} < 0 < N_{\max}$ *Schranken für den Exponenten. Dann sind alle* ***Maschinenzahlen*** *oder* ***Gleitkommazahlen*** *ungleich* 0 *(bzgl. dieser Vorgaben) von der Form*

$$x = (-1)^{\nu}\, b^{N} \sum_{j=1}^{p} c_j\, b^{-j}$$

mit dem Vorzeichen $\sigma = (-1)^{\nu}$ *mit* $\nu \in \{0; 1\}$*, dem Exponenten* $N \in \mathbb{Z}$ *mit* $N_{\min} \leq N \leq N_{\max}$ *und den Ziffern (oder Koeffizienten)* $c_1, c_2, \ldots, c_p \in \{0; 1; \ldots; b-1\}$ *mit* $c_1 \neq 0$ *der Mantisse. Weiter gehört die Zahl* 0 *zu den Maschinenzahlen. Da* $c_1 \neq 0$ *verlangt wird, sprechen wir von einem* ***normalisierten Gleitkomma-Zahlensystem****.*

Wir können mit der Hilfe der geometrischen Summe die größte Zahl im normalisierten Gleitkomma-Zahlensystem berechnen:

$$\begin{aligned} x_{\max} &= (-1)^0\, b^{N_{\max}} \sum_{j=1}^{p} (b-1)\, b^{-j} = b^{N_{\max}} (b-1) \left(\sum_{j=0}^{p} b^{-j} - 1 \right) \\ &= b^{N_{\max}}\, (b-1) \left(\frac{1 - \left(\frac{1}{b}\right)^{p+1}}{1 - \frac{1}{b}} - 1 \right) = b^{N_{\max}}\, (b-1) \left(\frac{1 - \left(\frac{1}{b}\right)^{p+1}}{\frac{b-1}{b}} - 1 \right) \\ &= b^{N_{\max}} \left(b - \left(\frac{1}{b}\right)^{p} - (b-1) \right) = b^{N_{\max}} \left(1 - b^{-p}\right), \end{aligned} \tag{1.2}$$

wobei beim Übergang zur zweiten Zeile die geometrische Summe genutzt wurde. Analog finden wir

$$\text{kleinste Zahl:} \qquad x_{\min} = (-1)^1\, b^{N_{\max}} \sum_{j=1}^{p} (b-1)\, b^{-j} = -b^{N_{\max}} \left(1 - b^{-p}\right). \tag{1.3}$$

Die betraglich kleinsten Zahlen im normalisierten Gleitkomma-Zahlensystem sind:

$$\begin{aligned} &\text{betraglich kleinste positive Zahl:} \quad x_{\text{posmin}} = (-1)^0\, b^{N_{\min}} \cdot \left(1 \cdot b^{-1}\right) = b^{N_{\min}-1}, \\ &\text{betraglich kleinste negative Zahl:} \quad x_{\text{negmin}} = (-1)^1\, b^{N_{\min}} \cdot \left(1 \cdot b^{-1}\right) = -b^{N_{\min}-1}. \end{aligned} \tag{1.4}$$

Damit man auch betraglich noch kleinere Zahlen als $x_{\text{posmin}} = b^{N_{\min}-1}$ darstellen kann, wird die Menge der normalisierten Gleitkommazahlen oft durch die Zahlen

$$x = (-1)^\nu\, b^{N_{\min}} \sum_{j=1}^{p} c_j\, b^{-j} \qquad \text{mit} \qquad c_1, c_2, \ldots, c_p \in \{0; 1; \ldots; b-1\} \tag{1.5}$$

ergänzt. (Die Forderung $c_1 \neq 0$ ist in (1.5) weggefallen.) Durch Hinzunahme der Zahlen (1.5) erhält man ein **denormalisiertes Gleitkomma-Zahlensystem**. In ihm sind die betraglich kleinsten Zahlen:

$$\begin{aligned} &\text{betraglich kleinste positive Zahl:} \quad \widetilde{x}_{\text{posmin}} = (-1)^0\, b^{N_{\min}} \cdot \left(1 \cdot b^{-p}\right) = b^{N_{\min}-p}, \\ &\text{betraglich kleinste negative Zahl:} \quad \widetilde{x}_{\text{negmin}} = (-1)^1\, b^{N_{\min}} \cdot \left(1 \cdot b^{-p}\right) = -b^{N_{\min}-p}. \end{aligned}$$

Tritt in einem Gleitkomma-Zahlensystem bei Berechnungen eine Zahl x mit $|x| > x_{\max}$ auf, so bekommt man die Warnung **„arithmetical overflow“** und das Programm stürzt oft ab. Tritt in einem Gleitkomma-Zahlensystem bei Berechnungen eine Zahl x mit $|x| < \widetilde{x}_{\text{posmin}}$ auf, so bekommt man entweder die Warnung **„arithmetical underflow“** oder der Computer rechnet mit dem Wert $x = 0$ weiter.

1.3 Rundung und Abschneiden

Reelle Zahlen mit $|x| \leq x_{\max}$ kann man in einem Gleitkomma-Zahlensystem durch Abschneiden oder Runden angenähert darstellen.

Definition 1.3. (Runden und Abschneiden)
Gegeben sei ein normalisiertes Gleitkomma-Zahlensystem mit Basis b, Mantissenlänge p und Exponentenschranken $N_{\min} < 0 < N_{\max}$. Sei $x \in \mathbb{R} \setminus \{0\}$ mit $x_{\text{minpos}} \leq |x| \leq x_{\max}$ in normalisierter Gleitkommadarstellung gegeben als

$$x = (-1)^\nu\, b^N \underbrace{\sum_{j=1}^{\infty} c_j\, b^{-j}}_{=a} \qquad \textit{mit} \qquad c_1 \neq 0, \tag{1.6}$$

wobei (wegen $x_{\text{minpos}} \leq |x| \leq x_{\max}$*)* $N_{\min} \leq N \leq N_{\max}$ *gilt. Dann wird* x *eine Zahl* $\text{rd}(x)$ *aus dem normalisierten Gleitkomma-Zahlensystem durch* ***Abschneiden*** *bzw.* ***Runden*** *zugeordnet, die wie folgt definiert ist:*

$$\text{rd}(x) = (-1)^{\nu}\, b^{N}\, \text{rd}(a) \tag{1.7}$$

mit

$$\text{rd}(a) = \text{rd}\left(\sum_{j=1}^{\infty} c_j\, b^{-j}\right) = \begin{cases} c_1\, b^{-1} + c_2\, b^{-2} + \ldots + c_p\, b^{-p} & \textit{(Abschneiden)}, \\ c_1\, b^{-1} + c_2\, b^{-2} + \ldots + \widetilde{c}_p\, b^{-p} & \textit{(Runden)}, \end{cases}$$

wobei beim Runden für die letze Ziffer $\widetilde{c}_p$ *gilt*

$$\widetilde{c}_p := \begin{cases} c_p, & \textit{falls} \quad \displaystyle\sum_{j=p+1}^{\infty} c_j\, b^{p+1-j} < \frac{b}{2} \quad \textit{(Abrunden)}, \\ c_p + 1, & \textit{falls} \quad \displaystyle\sum_{j=p+1}^{\infty} c_j\, b^{p+1-j} \geq \frac{b}{2} \quad \textit{(Aufrunden)}. \end{cases}$$

Beim Aufrunden sind gegebenenfalls (einfache oder mehrfache) Überträge zu beachten. Ob gerundet oder abgeschnitten wird, hängt von der Software ab.

Beispiel 1.4. (Runden und Abschneiden)
Wir wollen durch Abschneiden bzw. Runden jeweils Darstellungen der nachfolgenden Zahlen in einem Gleitkomma-Zahlensystem mit Basis $b = 10$ und Mantissenlänge $p = 6$ bestimmen:

$$x = \frac{2}{3} = 0,\overline{6}, \qquad y = 4{,}7684\overline{9}, \qquad z = 12{,}34\overline{9}.$$

Wir finden

$$\text{rd}(x) = \text{rd}\left(0{,}666666\overline{6} \cdot 10^{0}\right) = \begin{cases} 0{,}666666 \cdot 10^{0} = 0{,}666666 & \text{(Abschneiden)}, \\ 0{,}666667 \cdot 10^{0} = 0{,}666667 & \text{(Runden)}, \end{cases}$$

$$\text{rd}(y) = \text{rd}\left(0{,}4768499\overline{9} \cdot 10^{1}\right) = \begin{cases} 0{,}476849 \cdot 10^{1} = 4{,}76849 & \text{(Abschneiden)}, \\ 0{,}476850 \cdot 10^{1} = 4{,}7685 & \text{(Runden)}, \end{cases}$$

$$\text{rd}(z) = \text{rd}\left(0{,}1234999\overline{9} \cdot 10^{2}\right) = \begin{cases} 0{,}123499 \cdot 10^{2} = 12{,}3499 & \text{(Abschneiden)}, \\ 0{,}123500 \cdot 10^{2} = 12{,}35 & \text{(Runden)}. \end{cases}$$

Im zweiten Beispiel findet beim Runden ein einfacher Übertrag statt, und im dritten Beispiel findet beim Runden ein zweifacher Übertrag statt. ♠

Welchen Fehler macht man man, wenn man (1.6) **rundet bzw. abschneidet?** Mit (1.6) und (1.7) erhält man sofort

$$x - \mathrm{rd}(x) = (-1)^\nu \, b^N \, a - (-1)^\nu \, b^N \, \mathrm{rd}(a) = (-1)^\nu \, b^N \, \big(a - \mathrm{rd}(a)\big),$$

und für $a - \mathrm{rd}(a)$ gilt

$$a - \mathrm{rd}(a) = \begin{cases} \displaystyle\sum_{j=p+1}^{\infty} c_j \, b^{-j} & \text{(Abschneiden)}, \\ (c_p - \widetilde{c}_p) \, b^{-p} + \displaystyle\sum_{j=p+1}^{\infty} c_j \, b^{-j} & \text{(Runden)}. \end{cases}$$

Damit ergibt sich für den sogenannten **absoluten Fehler** $|\mathrm{rd}(x) - x|$ **beim Abschneiden bzw. Runden** die Abschätzung

$$\begin{aligned} \big|\mathrm{rd}(x) - x\big| &= \big|x - \mathrm{rd}(x)\big| = \big|(-1)^\nu \, b^N \, \big(a - \mathrm{rd}(a)\big)\big| \\ &= b^N \, \big|a - \mathrm{rd}(a)\big| \le \begin{cases} b^N \cdot b^{-p} = b^{N-p} & \text{(Abschneiden)}, \\ b^N \cdot \dfrac{1}{2} \, b^{-p} = \dfrac{1}{2} \, b^{N-p} & \text{(Runden)}. \end{cases} \end{aligned} \tag{1.8}$$

Insbesondere erhalten wir für den Sonderfall $x = a$

$$\big|\mathrm{rd}(a) - a\big| \le \begin{cases} b^{-p} & \text{(Abschneiden)}, \\ \dfrac{1}{2} \, b^{-p} & \text{(Runden)}. \end{cases}$$

Für den sogenannten **relativen Fehler** $\frac{|\mathrm{rd}(x)-x|}{|x|}$ ergibt sich **beim Abschneiden bzw. beim Runden** mit Hilfe von (1.8) die Abschätzung

$$\frac{|\mathrm{rd}(x) - x|}{|x|} = \frac{|x - \mathrm{rd}(x)|}{|x|} = \frac{\big|(-1)^\nu \, b^N \, \big(a - \mathrm{rd}(a)\big)\big|}{\big|(-1)^\nu \, b^N \, a\big|} = \frac{b^N \, \big|a - \mathrm{rd}(a)\big|}{b^N \, a}$$

$$= \frac{\big|a - \mathrm{rd}(a)\big|}{a} \le \begin{cases} \dfrac{b^{-p}}{a} \le \dfrac{b^{-p}}{b^{-1}} = b^{1-p} & \text{(Abschneiden)}, \\ \dfrac{1}{2} \, \dfrac{b^{-p}}{a} \le \dfrac{1}{2} \, \dfrac{b^{-p}}{b^{-1}} = \dfrac{1}{2} \, b^{1-p} & \text{(Runden)}, \end{cases} \tag{1.9}$$

wobei wir im letzten Schritt $a \in [b^{-1}; 1[$ genutzt haben.

Die kleinste obere Schranke für den relativen Fehler beim Abschneiden bzw. Runden in einem Gleitkomma-Zahlensystem heißt die **Maschinengenauigkeit**. An (1.9) lesen wir die Maschinengenauigkeit ab:

$$\boxed{\text{Maschinengenauigkeit:} \qquad \tau := \begin{cases} b^{1-p} & \text{(Abschneiden)}, \\ \dfrac{1}{2} \, b^{1-p} & \text{(Runden)}. \end{cases}} \tag{1.10}$$

Aus (1.9) folgt

$$\left|\frac{x - \mathrm{rd}(x)}{x}\right| \leq \tau,$$

und es gilt

$$\frac{\mathrm{rd}(x) - x}{x} =: \tau_x \quad \Longleftrightarrow \quad \mathrm{rd}(x) - x = \tau_x\, x \quad \Longleftrightarrow \quad \mathrm{rd}(x) = (1 + \tau_x)\, x$$

mit dem **(relativen) Darstellungsfehler** τ_x mit $|\tau_x| \leq \tau$.

Ein weiterer wichtiger Begriff bei der Diskussion der Rundung und des Abschneidens sind die **signifikanten Ziffern**.

Definition 1.5. (signifikante Ziffern im Dezimalsystem)
*Eine Näherung $\widetilde{x} \in \mathbb{R}$ einer quantitativen reellen Größe $x \in \mathbb{R}$ im Dezimalsystem hat **mindestens** m **signifikante Ziffern**, wenn $|\widetilde{x} - x|$ kleiner oder gleich 5 Einheiten in der $(m+1)$-sten Ziffer von x ist.*

Beispiel 1.6. (signifikante Ziffern)

(a) Die Näherung $\widetilde{x} = 0{,}222$ von $x = \frac{2}{9} = 0{,}\overline{2}$ hat drei signifikante Ziffern, denn

$$0{,}00005 < |\widetilde{x} - x| = 0{,}000\overline{2} \leq 0{,}0005,$$

und die 5 in 0,0005 steht an der Stelle der $(3+1)$-sten Ziffer von $x = 0{,}\overline{2}$.

(b) Die Näherung $\widetilde{x} = 31{,}578$ von $x = 31{,}575$ hat vier signifikante Ziffern, denn

$$0{,}0005 < |\widetilde{x} - x| = 0{,}003 \leq 0{,}005,$$

und die 5 in 0,005 steht an der Stelle der $(4+1)$-sten Ziffer von $x = 31{,}578$.

(c) Die Näherung $\widetilde{x} = 0{,}02138$ von $x = 0{,}02144$ hat nur zwei signifikante Ziffern, denn

$$0{,}00005 < |\widetilde{x} - x| = 0{,}00006 \leq 0{,}0005,$$

und die Ziffer 5 in 0,0005 steht an der Stelle der $(2+1)$-sten Ziffer von $x = 0{,}02144$.

Nur signifikante Ziffern sind bei der Angabe eines Ergebnisses relevant. ♠

Ein großes Problem bei Berechnungen in einem Gleitkomma-Zahlensystem mit einer endlichen Mantissenlänge (also bei allen Berechnungen mit einem Computer oder einem Taschenrechner) ist der **mögliche Verlust signifikanter Ziffern** durch **Auslöschung**. Um dieses zu verstehen, betrachten wir ein Beispiel.

Beispiel 1.7. (Verlust signifikanter Ziffern durch Auslöschung)
Die Funktion

$$f : \mathbb{R} \to \mathbb{R}, \quad f(x) = x\left(\sqrt{x+1} - \sqrt{x}\right),$$

soll für $x = 10^k$ mit $k = 0, 1, 2, 3, 4, 5$ ausgewertet werden, wobei in einem Gleitkomma-Zahlensystem mit der Mantissenlänge 6 gerechnet wird. Dieses bedeutet, dass wir in jedem einzelnen Rechenschritt auf eine 6-stellige Mantisse runden müssen. Um die so erhaltenen genäherten Werte von $f(x)$ zu unterscheiden, bezeichnen wir diese mit $\widetilde{f}(x)$.

$$f(1) = 1\cdot\left(\sqrt{2} - \sqrt{1}\right) \doteq 1\cdot\big(\underbrace{1{,}41421 - 1}_{=0{,}414210}\big) = 0{,}41421 = \widetilde{f}(1)$$

$$f(10) = 10\cdot\left(\sqrt{11} - \sqrt{10}\right) \doteq 10\cdot\big(\underbrace{3{,}31662 - 3{,}16228}_{=0{,}154340}\big) = 1{,}5434 = \widetilde{f}(10)$$

$$f(10^2) = 10^2\cdot\left(\sqrt{101} - \sqrt{100}\right) \doteq 100\cdot\big(\underbrace{10{,}0499 - 10}_{=0{,}0499000}\big) = 4{,}99 = \widetilde{f}(10^2)$$

$$f(10^3) = 10^3\cdot\left(\sqrt{1001} - \sqrt{1000}\right) \doteq 1000\cdot\big(\underbrace{31{,}6386 - 31{,}6228}_{=0{,}0158000}\big) = 15{,}8 = \widetilde{f}(10^3)$$

$$f(10^4) = 10^4\cdot\left(\sqrt{10001} - \sqrt{10^4}\right) \doteq 10000\cdot\big(\underbrace{100{,}005 - 100}_{=0{,}00500000}\big) = 50 = \widetilde{f}(10^4)$$

$$f(10^5) = 10^5\cdot\left(\sqrt{100001} - \sqrt{10}\right) \doteq 100000\cdot\big(\underbrace{316{,}229 - 316{,}228}_{=0{,}0010000}\big) = 100 = \widetilde{f}(10^5)$$

Das Symbol $\doteq$ bedeutet, dass (hier auf eine 6-stellige Mantisse) gerundet wurde.

Dabei erhalten wir also die Ergebnisse in Tabelle 1.1, wobei der wahre Wert $f(x)$ jeweils auf eine Mantissenlänge von 6 Ziffern gerundet wurde.

Für $x = 10^5$ ist der absolute Fehler $|\widetilde{f}(10^5) - f(10^5)| \doteq 58{,}113$, und der relative Fehlers ist

$$\frac{|\widetilde{f}(10^5) - f(10^5)|}{|f(10^5)|} \doteq 0{,}368 \mathrel{\widehat{=}} 36{,}8\,\%.$$

Wie kommt dieses eklatant schlechte Ergebnis zustande? Für großes x sind $\sqrt{x+1}$ und $\sqrt{x}$ fast gleich. Daher führt das Bilden der Differenz $\sqrt{x+1} - \sqrt{x}$ nach der jeweiligen Rundung von $\sqrt{x+1}$ und $\sqrt{x}$ mit wachsendem x zur Auslöschung von immer mehr signifikanten Ziffern. Dieses kann man bei der konkreten Berechnung von $\widetilde{f}(10^4)$ oder $\widetilde{f}(10^5)$ sehr gut sehen.

x	berechnetes $\widetilde{f}(x)$	exakter Wert für $f(x)$	$\lvert\widetilde{f}(x) - f(x)\rvert$	$\dfrac{\lvert\widetilde{f}(x) - f(x)\rvert}{\lvert f(x)\rvert}$
$1 = 10^0$	0,414210	0,414214	0,000004	$9{,}66 \cdot 10^{-6}$
$10 = 10^1$	1,54340	1,54347	0,00007	$4{,}54 \cdot 10^{-5}$
$100 = 10^2$	4,99000	4,98756	0,00244	$4{,}89 \cdot 10^{-4}$
$1000 = 10^3$	15,8000	15,8074	0,0074	$4{,}68 \cdot 10^{-4}$
$10.000 = 10^4$	50,0000	49,9988	0,0012	$2{,}40 \cdot 10^{-5}$
$100.000 = 10^5$	100,000	158,113	58,113	$3{,}68 \cdot 10^{-1}$

Tabelle 1.1: Ergebnisse und deren absolute und relative Fehler für die Berechnung von $f(x) = x\left(\sqrt{x+1} - \sqrt{x}\right)$ für $x = 10^k$, $k = 0, 1, 2, 3, 4, 5$, in einem Gleitkomma-Zahlensystem mit der Mantissenlänge 6.

Man kann das Problem der Auslöschung signifikanter Ziffern in diesem Beispiel leicht umgehen, indem man $f(x)$ vorher geeignet umformt und dann mit der neuen Darstellung von $f(x)$ rechnet: Durch Erweitern mit $\sqrt{x+1} + \sqrt{x}$ erhalten wir mit der dritten binomischen Formel:

$$\begin{aligned} f(x) = x\left(\sqrt{x+1} - \sqrt{x}\right) &= x \cdot \frac{\left(\sqrt{x+1} - \sqrt{x}\right)\left(\sqrt{x+1} + \sqrt{x}\right)}{\sqrt{x+1} + \sqrt{x}} \\ &= x \cdot \frac{\left(\sqrt{x+1}\right)^2 - \left(\sqrt{x}\right)^2}{\sqrt{x+1} + \sqrt{x}} = x \cdot \frac{(x+1) - x}{\sqrt{x+1} + \sqrt{x}} = \frac{x}{\sqrt{x+1} + \sqrt{x}} \end{aligned} \tag{1.11}$$

Berechnet man $f(10^5)$ mit dieser neuen Formel in einem Gleitkomma-Zahlensystem mit der Mantissenlänge 6, so erhält man

$$\begin{aligned} f(10^5) &= \frac{100000}{\sqrt{100001} + \sqrt{100000}} \doteq \frac{100000}{316{,}229 + 316{,}228} \\ &= \frac{100000}{632{,}457} \doteq 158{,}114 = \widetilde{\widetilde{f}}(10^5), \end{aligned}$$

und wir haben nur in der letzten Ziffer von $\widetilde{\widetilde{f}}(10^5)$ eine Abweichung von exakten Wert $f(10^5)$. Anders ausgedrückt: $\widetilde{\widetilde{f}}(10^5) = 158{,}114$ hat 5 signifikante Ziffern. Der absolute Fehler von $\widetilde{\widetilde{f}}(10^5)$ (als Näherung von $f(10^5)$) ist nun $\lvert\widetilde{\widetilde{f}}(10^5) - f(10^5)\rvert \doteq 0{,}001$, und der relative Fehler von $\widetilde{\widetilde{f}}(10^5)$ (als Näherung von $f(10^5)$) ist nun

$$\frac{\lvert\widetilde{\widetilde{f}}(10^5) - f(10^5)\rvert}{\lvert f(10^5)\rvert} \doteq 6{,}32 \cdot 10^{-6} \mathrel{\widehat{=}} 0{,}000632\,\%.$$

Wir erhalten nun ein Ergebnis mit einer sehr guten Genauigkeit. ♠

1.4 IEEE 64 Bit Maschinenzahlformat

Wie werden Zahlen in einem modernen Computer im Gleitkomma-Zahlensystem gespeichert? Wir besprechen hier nur das gängige **IEEE-Format 64 Bit (= 8 Byte) Maschinenzahlformat** mit **doppelter Genauigkeit** mit der Basis $b = 2$, der Mantissenlänge $p = 53$ und den Exponentenschranken $N_{\min} = -1021$ und $N_{\max} = 1024$. Im Computer wird eine 64 Bit Maschinenzahl im **normalisierten Gleitkomma-Zahlensystem** mit 1 Bit für das Vorzeichen, 11 Bit für den Exponenten und 52 Bit für die Mantisse gespeichert. Wir benötigen nur 52 Bit für die Speicherung der Mantisse, da im Dualsystem wegen $c_1 \neq 0$ und $c_1 \in \{0; \ldots; 2-1\} = \{0; 1\}$ folgt, dass $c_1 = 1$ sein muss und somit nicht mehr abgespeichert werden muss.

Mit dem normalisierten 64 Bit Maschinenzahlformat

ν	$n_0\, n_2\, \ldots\, n_{10}$	$c_2\, c_2\, \ldots\, c_{53}$

wobei $\nu, n_0, n_1, \ldots, n_{10}, c_2, c_3 \ldots, c_{53} \in \{0; 1\}$ sind, wird also die normalisierte Gleitkommazahl

$$x = (-1)^{\nu}\, 2^{N} \left(2^{-1} + \sum_{j=2}^{53} c_j\, 2^{-j} \right) \quad \text{mit Exponenten} \quad -1021 \leq N \leq 1024$$

dargestellt. (Genauer gilt für den Exponenten

$$N = \sum_{i=0}^{10} n_i\, 2^i - 1022.$$

Die Verschiebungsdistanz -1022 im Exponenten N sorgt dafür, dass sowohl positive wie negative Exponenten dargestellt werden können. Die speziellen Werte

$$\sum_{i=0}^{10} 0 \cdot 2^i = 0 \qquad \text{und} \qquad \sum_{i=0}^{10} 1 \cdot 2^i = \frac{1 - 2^{11}}{1 - 2} = 2047$$

sind reserviert. Aus den verbleibenden Werten $1, 2, \ldots, 2046$ ergibt sich mit der Verschiebungsdistanz -1022 für die möglichen Exponenten $N_{\min} = 1 - 1022 =$

-1021 und $N_{\max} = 2046 - 1022 = 1024$.) Wir erhalten damit für das 64 Bit normalisierte Maschinenzahlformat (vgl. (1.2) und (1.3))

$$\text{größte Zahl:} \qquad x_{\max} = 2^{1024}\left(1 - 2^{-53}\right) \approx 2^{1024} \doteq 1{,}8 \cdot 10^{308},$$

$$\text{kleinste Zahl:} \qquad x_{\min} = -2^{1024}\left(1 - 2^{-53}\right) \approx -2^{1024} \doteq -1{,}8 \cdot 10^{308},$$

und die betraglich kleinsten Zahlen sind (vgl. (1.4))

$$\text{betraglich kleinste positive Zahl:} \qquad x_{\text{posmin}} = 2^{-1022} \doteq 2{,}2 \cdot 10^{-308},$$

$$\text{betraglich kleinste negative Zahl:} \qquad x_{\text{negmin}} = -2^{-1022} \doteq -2{,}2 \cdot 10^{-308}.$$

Bei dem 64 Bit Maschinenzahlformat treten folgende Sonderfälle auf:

Zahl Null (+0):	0	00 ... 0	00 ... 0
Zahl Null (−0):	1	00 ... 0	00 ... 0
„Zahl“ $+\infty$:	0	11 ... 1	00 ... 0
„Zahl“ $-\infty$:	1	11 ... 1	00 ... 0

Besteht der Exponent nur aus aus Einsen und enthält die Mantisse Einsen, so ist der Wert der Zahl NaN („not a number“). Ein Beispiel ist

NaN („not a number“):	0	11 ... 1	010110 ... 0

.

Die **Maschinengenauigkeit** im 64 Bit Maschinenzahlformat ist (vgl. (1.10))

$$\tau = \begin{cases} 2^{1-53} = 2^{-52} \doteq 2{,}2 \cdot 10^{-16} & \text{(Abschneiden)}, \\ \dfrac{1}{2} \cdot 2^{1-53} = 2^{-53} \doteq 1{,}1 \cdot 10^{-16} & \text{(Runden)}. \end{cases}$$

1.5 Fehlerfortpflanzung und Fehlerverstärkung

Rechnet man in einem Gleitkomma-Zahlensystem mit der Basis b mit **endlicher Mantissenlänge**, so entstehen unvermeidbare Fehler, denn reelle Zahlen mit einer nicht abbrechenden Reihendarstellung bzgl. der Basis b (und reelle Zahlen mit einer endlichen Mantisse bzgl. der Basis b, die länger ist als die Mantissenlänge des Gleitkomma-Zahlensystems) können nicht exakt dargestellt werden, sondern müssen durch eine Maschinenzahl angenähert werden.

Aber selbst, wenn man beispielsweise zwei Maschinenzahlen x und y mit $x+y \neq 0$ addiert, so muss das Ergebnis keine Maschinenzahl sein. Somit erhält man als Ergebnis

$$\text{rd}(x + y) = (1 + \tau_{x+y})\,(x + y)$$

mit dem (relativen) Darstellungsfehler $\tau_{x+y} = \frac{\mathrm{rd}(x+y)-(x+y)}{x+y}$.

Sind x und y keine Maschinenzahlen, so ist das Ergebnis der Addition

$$\begin{aligned}\mathrm{rd}\big(\mathrm{rd}(x)+\mathrm{rd}(y)\big) &= \big(1+\tau_{\mathrm{rd}(x)+\mathrm{rd}(y)}\big)\,\big(\mathrm{rd}(x)+\mathrm{rd}(y)\big)\\ &= \big(1+\tau_{\mathrm{rd}(x)+\mathrm{rd}(y)}\big)\,\big((1+\tau_x)\,x+(1+\tau_y)\,y\big)\\ &= x+y+\big(\tau_x+\tau_{\mathrm{rd}(x)+\mathrm{rd}(y)}+\tau_x\,\tau_{\mathrm{rd}(x)+\mathrm{rd}(y)}\big)\,x\\ &\qquad +\big(\tau_y+\tau_{\mathrm{rd}(x)+\mathrm{rd}(y)}+\tau_y\,\tau_{\mathrm{rd}(x)+\mathrm{rd}(y)}\big)\,y\end{aligned}$$

wobei für τ_x, τ_y und $\tau_{\mathrm{rd}(x)+\mathrm{rd}(y)}$ gilt $|\tau_x| \le \tau$, $|\tau_y| \le \tau$ und $|\tau_{\mathrm{rd}(x)+\mathrm{rd}(y)}| \le \tau$. (Dabei ist τ die Maschinengenauigkeit.) Jeder einzelne relative Fehler durch Rundung bzw. Abschneiden ist also betraglich durch die Maschinengenauigkeit τ begrenzt.

Insbesondere führt das Runden bzw. Abschneiden beim Rechnen dazu, dass **die Assoziativgesetze und das Distributivgesetz nicht mehr gelten**, sobald Zahlen auftreten, die keine Maschinenzahlen mehr sind.

Beispiel 1.8. (Rundungsfehler und verletztes Assoziativgesetz)
Wir Rechnen mit der 64 Bit Maschinenzahlformat (vgl. Teilkapitel 1.4). In dieser sind $x = 4$ und $y = 10^{-20}$ beides Maschinenzahlen, aber

$$x + y = 4 + 10^{-20} = 0{,}40000000000000000001 \cdot 10^1$$

ist keine Maschinenzahl mehr, denn wegen $\tau \approx 10^{-16}$ erhalten wir im Dezimalsystem nur eine Mantisse mit 16 Stellen, also

$$\mathrm{rd}(x+y) = \mathrm{rd}(4+10^{-20}) = 4.$$

Analog findet man für die Maschinenzahlen $x = 10^{-20}$, $y = 2$ und $z = -2$

$$\begin{aligned}\mathrm{rd}\big(\mathrm{rd}(x+y)+z\big) &= \mathrm{rd}\big(\mathrm{rd}(10^{-20}+2)+(-2)\big) = \mathrm{rd}\big(2+(-2)\big) = \mathrm{rd}(0) = 0\\ \mathrm{rd}\big(x+\mathrm{rd}(y+z)\big) &= \mathrm{rd}\big(10^{-20}+\mathrm{rd}(2+(-2))\big) = \mathrm{rd}\big(10^{-20}+\mathrm{rd}(0)\big)\\ &= \mathrm{rd}(10^{-20}+0) = 10^{-20} \neq 0.\end{aligned}$$

Also gilt hier

$$\mathrm{rd}\big(\mathrm{rd}(x+y)+z\big) \neq \mathrm{rd}\big(x+\mathrm{rd}(y+z)\big),$$

d.h. das Assoziativgesetz der Addition ist verletzt. ♠

Wir wollen nun die **Fehlerfortpflanzung** systematisch untersuchen. Dabei interessiert uns aktuell nicht, ob die fehlerhaften Eingangsdaten durch Rundungsfehler Messfehler oder eine andere Ursache zustande kommen. Wir werden sowohl den absoluten Fehler als auch den relativen Fehler betrachten.

Definition 1.9. (absoluter Fehler und relativer Fehler)

*(1) Sei $x \in \mathbb{R}$ eine **reellwertige** Größe und sei $\widetilde{x} \in \mathbb{R}$ ein fehlerbehafteter Näherungswert für x. Dann ist der **absolute Fehler** der Näherung $\widetilde{x}$*

$$\mathrm{Abs}(\widetilde{x}) := |\widetilde{x} - x|.$$

*Ist $x \neq 0$, so ist der **relative Fehler** der Näherung $\widetilde{x}$*

$$\mathrm{Rel}(\widetilde{x}) := \frac{|\widetilde{x} - x|}{|x|} = \frac{\mathrm{Abs}(\widetilde{x})}{|x|}.$$

*(2) Sei $\mathbb{R}^n$ mit einer Norm $\|\cdot\|$ (z.B. der euklidischen Norm) versehen. Sei $\mathbf{x} \in \mathbb{R}^n$ eine **vektorwertige** Größe, und sei $\widetilde{\mathbf{x}} \in \mathbb{R}^n$ eine fehlerbehaftete Näherung für $\mathbf{x}$. Dann ist der **absolute Fehler** der Näherung $\widetilde{\mathbf{x}}$*

$$\mathrm{Abs}(\widetilde{\mathbf{x}}) = \|\widetilde{\mathbf{x}} - \mathbf{x}\|.$$

*Ist $\mathbf{x} \neq \mathbf{0}$, so ist der **relative Fehler** der Näherung $\widetilde{\mathbf{x}}$*

$$\mathrm{Rel}(\widetilde{\mathbf{x}}) := \frac{\|\widetilde{\mathbf{x}} - \mathbf{x}\|}{\|\mathbf{x}\|} = \frac{\mathrm{Abs}(\widetilde{\mathbf{x}})}{\|\mathbf{x}\|}.$$

Wir haben den absoluten und den relativen Fehler einer reellwertigen Größe bereits bei der Analyse der Rundung und des Abschneidens und in Beispiel 1.7 verwendet.

Beispiel 1.10. (absoluter und relativer Fehler)

(a) Die Distanz zwischen zwei Städten sei exakt $x = 100$ km. Die Näherung dieser Entfernung (z.B. durch eine Messung) sei $\widetilde{x} = 101$ km. Dann gelten

$$\mathrm{Abs}(\widetilde{x}) = |\widetilde{x} - x| = |101 \text{ km} - 100 \text{ km}| = 1 \text{ km},$$
$$\mathrm{Rel}(\widetilde{x}) = \frac{|\widetilde{x} - x|}{|x|} = \frac{1 \text{ km}}{100 \text{ km}} = 0{,}01 \mathrel{\widehat{=}} 1\,\%.$$

(b) Die Distanz zwischen zwei Dörfern sei exakt $x = 2$ km. Die Näherung dieser Entfernung (z.B. durch eine Messung) sei $\widetilde{x} = 3$ km. Dann gelten

$$\mathrm{Abs}(\widetilde{x}) = |\widetilde{x} - x| = |3 \text{ km} - 2 \text{ km}| = 1 \text{ km},$$
$$\mathrm{Rel}(\widetilde{x}) = \frac{|\widetilde{x} - x|}{|x|} = \frac{1 \text{ km}}{2 \text{ km}} = 0{,}5 \mathrel{\widehat{=}} 50\,\%.$$

Der absolute Fehler ist in beiden Beispielen jeweils 1 km. Trotzdem ist es auch intuitiv klar, dass die Näherung für die Entfernung der beiden Städte „viel genauer“ ist als die Näherung für die Entfernung der beiden Dörfer. Dieses wird durch den relativen Fehler widergespiegelt. Dieser beträgt im Beispiel (a) nur 1 % der Entfernung, aber im Beispiel (b) dagegen 50 % der Entfernung. ♠

Der **relative Fehler** $\mathrm{Rel}(\widetilde{x})$ bzw. $\mathrm{Rel}(\widetilde{\mathbf{x}})$ einer Näherung $\widetilde{x}$ von $x \neq 0$ bzw. einer Näherung $\widetilde{\mathbf{x}}$ von $\mathbf{x} \neq \mathbf{0}$ **ist normalerweise aussagekräftiger**, da er den Fehler relativ zur „Größe“ von x bzw. $\mathbf{x}$ betrachtet. Genauer gibt

$$\mathrm{Rel}(\widetilde{x}) \cdot 100\% = \frac{|\widetilde{x} - x|}{|x|} \cdot 100\% \qquad \text{bzw.} \qquad \mathrm{Rel}(\widetilde{\mathbf{x}}) \cdot 100\% = \frac{\|\widetilde{\mathbf{x}} - \mathbf{x}\|}{\|\mathbf{x}\|} \cdot 100\%$$

an, wieviel Prozent des Wertes $|x|$ bzw. $\|\mathbf{x}\|$ der absolute Fehler $\mathrm{Abs}(\widetilde{x}) = |\widetilde{x} - x|$ bzw. $\mathrm{Abs}(\widetilde{\mathbf{x}}) = \|\widetilde{\mathbf{x}} - \mathbf{x}\|$ ausmacht.

Die Grundlage für eine Analyse der Fehlerfortpflanzung bildet die **Taylorsche Formel** (auch **Satz von Taylor** genannt) mit dem Taylorpolynom vom Grad 1: Seien $D \subseteq \mathbb{R}^n$ eine konvexe Menge und $\mathbf{x} = (x_1, x_2, \ldots, x_n) \in D$, und sei $f : D \to \mathbb{R}$ eine zweimal stetig differenzierbare Funktion. Dann gibt es zu jedem

$$\widetilde{\mathbf{x}} = \mathbf{x} + \Delta\mathbf{x} = (x_1 + \Delta x_1, x_2 + \Delta x_2, \ldots, x_n + \Delta x_n) \in D$$

ein $\mathbf{z}_{\Delta\mathbf{x}}$ auf der Verbindungsstrecke von $\widetilde{\mathbf{x}} = \mathbf{x} + \Delta\mathbf{x}$ und $\mathbf{x}$, so dass gilt

$$f(\mathbf{x} + \Delta\mathbf{x}) = \underbrace{f(\mathbf{x}) + \sum_{k=1}^{n} \frac{\partial f(\mathbf{x})}{\partial x_k}\, \Delta x_k}_{\text{Taylorpolynom vom Grad 1}} + \underbrace{\frac{1}{2} \sum_{j=1}^{n} \sum_{k=1}^{n} \frac{\partial^2 f(\mathbf{z}_{\Delta\mathbf{x}})}{\partial x_j \partial x_k}\, \Delta x_j\, \Delta x_k}_{\text{Restglied}}.$$

Wenn das Restglied hinreichend klein ist, dann gilt die Näherung

$$f(\mathbf{x} + \Delta\mathbf{x}) \approx f(\mathbf{x}) + \sum_{k=1}^{n} \frac{\partial f(\mathbf{x})}{\partial x_k}\, \Delta x_k$$

$$\iff \qquad f(\mathbf{x} + \Delta\mathbf{x}) - f(\mathbf{x}) \approx \sum_{k=1}^{n} \frac{\partial f(\mathbf{x})}{\partial x_k}\, \Delta x_k.$$

Also ist der **absolute Fehler** $\mathrm{Abs}\big(f(\mathbf{x}+\Delta x)\big) = |f(\mathbf{x}+\Delta\mathbf{x}) - f(\mathbf{x})|$ der Näherung $f(\widetilde{\mathbf{x}}) = f(\mathbf{x} + \Delta x)$ von $f(\mathbf{x})$ angenähert gegeben durch

$$\mathrm{Abs}\big(f(\mathbf{x} + \Delta x)\big) = |f(\mathbf{x} + \Delta\mathbf{x}) - f(\mathbf{x})| \approx \left| \sum_{k=1}^{n} \frac{\partial f(\mathbf{x})}{\partial x_k} \cdot \Delta x_k \right|. \tag{1.12}$$

Daraus folgt eine angenäherte Abschätzung von $\mathrm{Abs}\big(f(\mathbf{x}+\Delta x)\big)$ nach oben:

$$\mathrm{Abs}\big(f(\mathbf{x}+\Delta x)\big) = |f(\mathbf{x}+\Delta\mathbf{x}) - f(\mathbf{x})| \lessapprox \sum_{k=1}^{n} \left|\frac{\partial f(\mathbf{x})}{\partial x_k}\right| \cdot |\Delta x_k|. \tag{1.13}$$

Das Symbol $\lessapprox$ in (1.13) (und in (1.15) weiter unten) bedeutet, dass genähert und (nach rechts) nach oben abgeschätzt wurde.

Division in (1.12) durch $|f(\mathbf{x})| \neq 0$ liefert eine Näherung für den **relativen Fehler** $\mathrm{Rel}\big(f(\mathbf{x}+\Delta x)\big)$ der Näherung $f(\widetilde{\mathbf{x}}) = f(\mathbf{x}+\Delta\mathbf{x})$ von $f(\mathbf{x})$:

$$\begin{aligned}\mathrm{Rel}\big(f(\mathbf{x}+\Delta x)\big) &= \frac{|f(\mathbf{x}+\Delta\mathbf{x}) - f(\mathbf{x})|}{|f(\mathbf{x})|} \approx \frac{1}{|f(\mathbf{x})|} \cdot \left|\sum_{k=1}^{n} \frac{\partial f(\mathbf{x})}{\partial x_k} \cdot \Delta x_k\right| \\ &= \left|\sum_{k=1}^{n} \frac{1}{f(\mathbf{x})} \cdot \frac{\partial f(\mathbf{x})}{\partial x_k} \cdot \Delta x_k\right| = \left|\sum_{k=1}^{n} \left(\frac{x_k}{f(\mathbf{x})} \cdot \frac{\partial f(\mathbf{x})}{\partial x_k}\right) \cdot \frac{\Delta x_k}{x_k}\right| \end{aligned} \tag{1.14}$$

Aus (1.14) folgt die Abschätzung

$$\begin{aligned}&\mathrm{Rel}\big(f(\mathbf{x}+\Delta x)\big) = \frac{|f(\mathbf{x}+\Delta\mathbf{x}) - f(\mathbf{x})|}{|f(\mathbf{x})|} \\ &\lessapprox \sum_{k=1}^{n} \left|\frac{x_k}{f(\mathbf{x})} \cdot \frac{\partial f(\mathbf{x})}{\partial x_k}\right| \cdot \left|\frac{\Delta x_k}{x_k}\right| = \sum_{k=1}^{n} \left|\frac{x_k}{f(\mathbf{x})} \cdot \frac{\partial f(\mathbf{x})}{\partial x_k}\right| \cdot \mathrm{Rel}(\widetilde{x}_k). \end{aligned} \tag{1.15}$$

An (1.14) und (1.15) können wir ablesen, dass die relativen Fehler $\mathrm{Rel}(\widetilde{x}_k) = \left|\frac{\Delta x_k}{x_k}\right|$, $k = 1, 2, \ldots, n$, der einzelnen Komponenten $x_1, x_2, \ldots, x_n$ von $\mathbf{x}$ (also die Eingangsdaten der Funktion f) jeweils mit den Faktoren bzw. den Beträgen der Faktoren

$$\frac{x_k}{f(\mathbf{x})} \cdot \frac{\partial f(\mathbf{x})}{\partial x_k}, \qquad k = 1, 2, \ldots, n,$$

multipliziert werden. Die betragliche Größe dieser Faktoren bestimmt also im Wesentlichen, wie groß der relative Fehler der Näherung $f(\widetilde{\mathbf{x}}) = f(\mathbf{x}+\Delta\mathbf{x})$ von $f(\mathbf{x})$ schlimmstenfalls werden kann.

Betrachten wir zwei Beispiele.

Beispiel 1.11. (Fehlerfortpflanzung bei der Addition)

Sei $f : \mathbb{R}^2 \to \mathbb{R}$, $f(x, y) = x + y$, die Addition von x und y. Dann gelten

$$\frac{\partial f(x,y)}{\partial x} = 1, \qquad \frac{\partial f(x,y)}{\partial y} = 1, \qquad \frac{\partial^2(x,y)}{\partial x^2} = \frac{\partial^2(x,y)}{\partial y^2} = \frac{\partial^2(x,y)}{\partial x \partial y} = 0.$$

Da alle zweiten Ableitungen null sind, liefert uns die Taylorsche Formel hier exakt

$$f(x+\Delta x, y+\Delta y) = f(x,y) + 1\cdot\Delta x + 1\cdot\Delta y + 0 = f(x,y) + \Delta x + \Delta y$$
$$\iff \qquad f(x+\Delta x, y+\Delta y) - f(x,y) = \Delta x + \Delta y.$$

Also gilt für den absoluten Fehler der Näherung $f(x+\Delta x, y+\Delta y)$ von $f(x,y)$

$$\begin{aligned}\mathrm{Abs}\big(f(x+\Delta x, y+\Delta y)\big) &= \big|f(x+\Delta x, y+\Delta y) - f(x,y)\big| \\ &= |\Delta x + \Delta y| \le |\Delta x| + |\Delta y|.\end{aligned}$$

Für den relativen Fehler finden wir (mit den Annahmen $x+y\neq 0$, $x\neq 0$, $y\neq 0$)

$$\mathrm{Rel}\big(f(x+\Delta x, y+\Delta y)\big) = \frac{\big|f(x+\Delta x, y+\Delta y) - f(x,y)\big|}{|f(x,y)|} = \frac{|\Delta x+\Delta y|}{|x+y|}$$
$$= \left|\frac{\Delta x+\Delta y}{x+y}\right| = \left|\frac{\Delta x}{x+y} + \frac{\Delta y}{x+y}\right| = \left|\frac{x}{x+y}\cdot\frac{\Delta x}{x} + \frac{y}{x+y}\cdot\frac{\Delta y}{y}\right|. \qquad (1.16)$$

Also gilt für den relativen Fehler der Näherung $f(x+\Delta x, y+\Delta y)$ von $f(x,y)$

$$\begin{aligned}\mathrm{Rel}\big(f(x+\Delta x, y+\Delta y)\big) &\le \left|\frac{x}{x+y}\right|\cdot\left|\frac{\Delta x}{x}\right| + \left|\frac{y}{x+y}\right|\cdot\left|\frac{\Delta y}{y}\right| \\ &= \left|\frac{x}{x+y}\right|\cdot\mathrm{Rel}(\widetilde{x}) + \left|\frac{y}{x+y}\right|\cdot\mathrm{Rel}(\widetilde{y}), \qquad (1.17)\end{aligned}$$

wobei wir genutzt haben, dass der relative Fehler von $\widetilde{x} = x+\Delta x$ bzw. $\widetilde{y} = y+\Delta y$ wie folgt gegeben ist:

$$\mathrm{Rel}(\widetilde{x}) = \frac{|\Delta x|}{|x|} = \frac{|(x+\Delta x)-x|}{|x|} \quad \text{bzw.} \quad \mathrm{Rel}(\widetilde{y}) = \frac{|\Delta y|}{|y|} = \frac{|(y+\Delta y)-y|}{|y|}.$$

(a) Haben x und y das gleiche Vorzeichen, so gelten

$$0 \le \frac{x}{x+y} \le 1 \qquad \text{und} \qquad 0 \le \frac{y}{x+y} \le 1,$$

und es folgt aus (1.17)

$$\mathrm{Rel}\big(f(x+\Delta x, y+\Delta y)\big) \le \frac{x}{x+y}\cdot\mathrm{Rel}(\widetilde{x}) + \frac{y}{x+y}\cdot\mathrm{Rel}(\widetilde{y})$$
$$\le \frac{x}{x+y}\max\big\{\mathrm{Rel}(\widetilde{x}), \mathrm{Rel}(\widetilde{y})\big\} + \frac{y}{x+y}\max\big\{\mathrm{Rel}(\widetilde{x}), \mathrm{Rel}(\widetilde{y})\big\}$$
$$= \left(\frac{x}{x+y} + \frac{y}{x+y}\right)\max\big\{\mathrm{Rel}(\widetilde{x}), \mathrm{Rel}(\widetilde{y})\big\} = \max\big\{\mathrm{Rel}(\widetilde{x}), \mathrm{Rel}(\widetilde{y})\big\}.$$

Der relative Fehler von $f(x+\Delta x, y+\Delta y) = (x+\Delta x)+(y+\Delta y)$ als Näherung der Summe $f(x,y) = x+y$ ist bei gleichen Vorzeichen von x und y also höchstens so groß wie das Maximum der relativen Fehler

$$\mathrm{Rel}(\widetilde{x}) = \frac{|\Delta x|}{|x|} = \frac{|(x+\Delta x)-x|}{|x|} \quad \text{und} \quad \mathrm{Rel}(\widetilde{y}) = \frac{|\Delta y|}{|y|} = \frac{|(y+\Delta y)-y|}{|y|}$$

der Näherungen $\widetilde{x} = x+\Delta x$ und $\widetilde{y} = y+\Delta y$ für x bzw. y.

(b) Haben x und y dagegen unterschiedliche Vorzeichen, so können $\frac{x}{x+y}$ und $\frac{y}{x+y}$ durch Auslöschung im Nenner betraglich beliebig groß werden, so dass der relative Fehler in (1.16) und (1.17) beliebig groß werden kann. Dieses illustriert das folgende Beispiel: Seien $x = 10^6+1$ und $y = -10^6$, und seien $\widetilde{x} = x+\Delta x$ bzw. $\widetilde{y} = y+\Delta y$ Näherungswerte für x bzw. y mit absoluten Fehlern $|\Delta x| = |\widetilde{x}-x| \leq \frac{1}{2}$ und $|\Delta y| = |\widetilde{y}-y| \leq \frac{1}{2}$. Dann erfüllen die relativen Fehler der Näherungen $\widetilde{x}$ und $\widetilde{y}$ jeweils

$$\mathrm{Rel}(\widetilde{x}) = \frac{|\Delta x|}{|x|} \leq \frac{\frac{1}{2}}{10^6} = 0{,}5\cdot 10^{-6} \quad \text{bzw.} \quad \mathrm{Rel}(\widetilde{y}) = \frac{|\Delta y|}{|y|} \leq \frac{\frac{1}{2}}{10^6} = 0{,}5\cdot 10^{-6},$$

wobei wir genutzt haben, dass $|x| = |10^6+1| > 10^6$ und $|y| = |-10^6| = 10^6$ erfüllen. Mit (1.16) folgt für den relativen Fehler von $\widetilde{x}+\widetilde{y}$

$$\begin{aligned}\mathrm{Rel}(\widetilde{x}+\widetilde{y}) &= \left|\frac{10^6+1}{(10^6+1)+(-10^6)}\cdot\frac{\Delta x}{x} + \frac{-10^6}{(10^6+1)+(-10^6)}\cdot\frac{\Delta y}{y}\right| \\ &= \left|(10^6+1)\,\frac{\Delta x}{x} - 10^6\,\frac{\Delta y}{y}\right|. \qquad (1.18)\end{aligned}$$

Gelten beispielsweise $\Delta x = \frac{1}{2}$ und $\Delta y = \frac{1}{2}$ und damit

$$\frac{\Delta x}{x} = \frac{1}{2}\,(10^6+1)^{-1} \approx \frac{10^{-6}}{2} \qquad \text{und} \qquad \frac{\Delta y}{y} = -\,\frac{10^{-6}}{2}\,,$$

so folgt aus (1.18)

$$\begin{aligned}\mathrm{Rel}(\widetilde{x}+\widetilde{y}) &= \left|(10^6+1)\cdot\frac{(10^6+1)^{-1}}{2} - 10^6\cdot\left(-\frac{10^{-6}}{2}\right)\right| \\ &\approx \left|(10^6+1)\cdot\frac{10^{-6}}{2} + 10^6\cdot\frac{10^{-6}}{2}\right| = (2\cdot 10^6+1)\cdot\frac{10^{-6}}{2}\,,\end{aligned}$$

d.h. der relative Fehler der Summe $\widetilde{x}+\widetilde{y}$ ist das ungefähr $(2\cdot 10^6)$-fache der relativen Fehler von $\widetilde{x}$ und $\widetilde{y}$. Ein so stark fehlerbehaftetes Ergebnis ist natürlich völlig unbrauchbar.

Wir sehen also, dass bei der Addition von Zahlen mit unterschiedlichem Vorzeichen eine Fehlerverstärkung stattfinden kann, die zu völlig unbrauchbaren Ergebnissen führt! ♠

Beispiel 1.12. (Fehlerfortpflanzung bei Funktionsauswertung)
Sei $f : \mathbb{R} \to \mathbb{R}$, $f(x) = e^{3x^2}$. Nach der Taylorschen Formel (bzw. Satz von Taylor) gibt es zu jedem $\widetilde{x} = x + \Delta x$ ein $z_{\Delta x}$ zwischen x und $\widetilde{x} = x + \Delta x$, so dass gilt:

$$\begin{aligned} f(x+\Delta x) &= f(x) + f'(x)\big((x+\Delta x) - x\big) + \frac{1}{2} f''(z_{\Delta x})\big((x+\Delta x) - x\big)^2 \\ &= f(x) + f'(x)\,\Delta x + \frac{1}{2} f''(z_{\Delta x})\,(\Delta x)^2 \\ \iff \qquad f(x+\Delta x) - f(x) &= f'(x)\,\Delta x + \frac{1}{2} f''(z_{\Delta x})\,(\Delta x)^2 \end{aligned}$$

Also finden wir mit der ersten und zweiten Ableitung

$$f'(x) = 6\,x\,e^{3x^2} \qquad \text{und} \qquad f''(x) = 6\,e^{3x^2} + 36\,x^2\,e^{3x^2},$$

dass mit $z = z_{\Delta x}$ gilt

$$f(x+\Delta x) - f(x) = \underbrace{6\,x\,e^{3x^2}\,\Delta x}_{\substack{\text{Taylorpolynom} \\ \text{vom Grad 1}}} + \underbrace{\Big(3\,e^{3z^2} + 18\,z^2\,e^{3z^2}\Big)\,(\Delta x)^2}_{\text{Restglied}}.$$

Daran sehen wir, dass nur für x und $\widetilde{x} = x + \Delta x$ dicht bei 0, und damit auch z_x dicht bei 0, das Restglied vernachlässigbar ist. Dann gilt für den absoluten Fehler

$$\mathrm{Abs}\big(f(x+\Delta x)\big) = |f(x+\Delta x) - f(x)| \approx |6\,x\,e^{3x^2}\,\Delta x| = 6\,|x|\,e^{3x^2}\,|\Delta x|. \quad (1.19)$$

Betrachten wir nun den relativen Fehler für x und $\widetilde{x} = x + \Delta x$ dicht bei 0:

$$\begin{aligned} \mathrm{Rel}\big(f(x+\Delta x)\big) &= \frac{|f(x+\Delta x) - f(x)|}{|f(x)|} = \frac{6\,|x|\,e^{3x^2}\,|\Delta x|}{e^{3x^2}} \\ &= 6\,|x|\,|\Delta x| = 6\,|x|^2\,\frac{|\Delta x|}{|x|} = 6\,|x|^2\,\mathrm{Rel}(x+\Delta x) \qquad (1.20) \end{aligned}$$

Im relativen Fehler $\mathrm{Rel}\big(f(x+\Delta x)\big)$ von $f(x+\Delta x)$ als Näherung für $f(x)$ wird der relative Fehler $\mathrm{Rel}(\widetilde{x}) = \mathrm{Rel}(x+\Delta x) = \frac{|\Delta x|}{|x|}$ der genäherten Eingangsdaten $\widetilde{x} = x + \Delta x$ also durch den Faktor $6\,|x|^2$ verstärkt. Nur für x dicht bei 0 ist der relative Fehler $\mathrm{Rel}\big(f(x+\Delta x)\big)$ der Näherung $f(x+\Delta x) = e^{3(x+\Delta x)^2}$ für $f(x) = e^{3x^2}$ klein. Also ist $f(x+\Delta x) = e^{3(x+\Delta x)^2}$ auch nur für x dicht bei 0 und nur für kleines Δx eine brauchbare Näherung für $f(x) = e^{3x^2}$. ♠

Natürlich treten zusätzlich immer auch noch Fehler durch die Rundung bzw. das Abschneiden beim Rechnen in einem Gleitkomma-Zahlensystem mit endlicher Matissenlänge auf, sobald innerhalb des Rechenprozesses Zahlen auftreten, die keine Maschinenzahlen sind. Diese Rundungs- bzw. Abschneidefehler wurden in den vorigen beiden Beispielen bei der Analyse nicht berücksichtigt. Rundungs- und Abschneidefehler sind aber unvermeidbar und ihr Einfluss kann höchstens durch die Wahl eines geeigneten Algorithmus zur Lösung des Problems möglichst klein gehalten werden.

1.6 Kondition und Stabilität

Kondition und Stabilität sind zwei der zentralen Begriffe der Numerik.

Definition 1.13. (Kondition, schlecht bzw. gut konditioniert)
*Seien $D \subseteq \mathbb{R}^m$ und $Y \subseteq \mathbb{R}^n$. Sei $\mathbf{f} : D \to Y$ eine vektorwertige (und für $n = 1$ reellwertige) Funktion, die eine mathematische Aufgabe beschreibt. Sei $\mathbf{x} \in D$ und sei $\widetilde{\mathbf{x}} = \mathbf{x} + \Delta\mathbf{x}$ eine **Näherung oder Störung** von $\mathbf{x}$. Die **absolute Konditionszahl von** $\mathbf{f}(\mathbf{x})$ ist die **kleinste reelle Zahl** $\kappa_{\text{abs}}(\mathbf{x})$, für die gilt*

$$\|\mathbf{f}(\mathbf{x} + \Delta\mathbf{x}) - \mathbf{f}(\mathbf{x})\|_{(n)} \leq \kappa_{\text{abs}}(\mathbf{x}) \, \|\Delta\mathbf{x}\|_{(m)} \qquad \text{für alle } \Delta\mathbf{x} \text{ mit } \mathbf{x} + \Delta\mathbf{x} \in D. \tag{1.21}$$

*Sind $\mathbf{x} \neq \mathbf{0}$ und $\mathbf{f}(\mathbf{x}) \neq \mathbf{0}$, so heißt die **kleinste reelle Zahl** $\kappa(\mathbf{x})$, für die gilt*

$$\frac{\|\mathbf{f}(\mathbf{x} + \Delta\mathbf{x}) - \mathbf{f}(\mathbf{x})\|_{(n)}}{\|\mathbf{f}(\mathbf{x})\|_{(n)}} \leq \kappa(\mathbf{x}) \, \frac{\|\Delta\mathbf{x}\|_{(m)}}{\|\mathbf{x}\|_{(m)}} \qquad \text{für alle } \Delta\mathbf{x} \text{ mit } \mathbf{x} + \Delta\mathbf{x} \in D, \tag{1.22}$$

*die **relative Konditionszahl von** $\mathbf{f}(\mathbf{x})$. $\mathbf{f}(\mathbf{x})$ heißt **schlecht konditioniert (an der Stelle $\mathbf{x}$)**, wenn $\kappa(\mathbf{x}) \gg 1$ ist. („$\kappa(\mathbf{x}) \gg 1$“ bedeutet, dass $\kappa(\mathbf{x})$ sehr viel größer als 1 ist.). Ist $\mathbf{f}(\mathbf{x})$ **nicht schlecht konditioniert (an der Stelle $\mathbf{x}$)**, so heißt $\mathbf{f}(\mathbf{x})$ **gut konditioniert (an der Stelle $\mathbf{x}$)**. (In (1.21) und (1.22) sind $\|\cdot\|_{(n)}$ bzw. $\|\cdot\|_{(m)}$ jeweils der Absolutbetrag, falls $n = 1$ bzw. $m = 1$. Andernfalls handelt es sich jeweils um eine geeignete Norm für $\mathbb{R}^n$ bzw. $\mathbb{R}^m$.)*

Was bedeuten (1.21) und (1.22)? Die absolute Konditionszahl $\kappa_{\text{abs}}(\mathbf{x})$ beschränkt den Faktor bei der Fortpflanzung des absoluten Fehlers. Analog beschränkt die relative Konditionszahl $\kappa(\mathbf{x})$ den Faktor bei der Fortpflanzung des

relativen Fehlers. – Ist $\mathbf{f}(\mathbf{x})$ schlecht konditioniert, so bedeutet es, dass es ein $\widetilde{\mathbf{x}} = \mathbf{x} + \Delta\mathbf{x} \in D$ gibt, für dass gilt

$$\frac{\|\mathbf{f}(\mathbf{x}+\Delta\mathbf{x}) - \mathbf{f}(\mathbf{x})\|_{(n)}}{\|\mathbf{f}(\mathbf{x})\|_{(n)}} \gg \frac{\|\Delta\mathbf{x}\|_{(m)}}{\|\mathbf{x}\|_{(m)}}.$$

Es ist klar, dass wir nur brauchbare Ergebnisse erwarten können, wenn eine mathematische Aufgabe $f(x)$ gut konditioniert ist.

Betrachten wir einige Beispiele.

Beispiel 1.14. (schlecht bzw. gut konditioniert)

(a) In Beispiel 1.11 (a) haben wir die Addition $f(x,y) = x + y$ zweier Zahlen untersucht. Falls x und y das gleiche Vorzeichen haben und $x + y \neq 0$, $x \neq 0$ und $y \neq 0$ sind, fanden wir, dass der relative Fehler der Näherung $f(x+\Delta x, y+\Delta y) = f(\widetilde{x}, \widetilde{y}) = \widetilde{x}+\widetilde{y}$ für $f(x,y) = x+y$ höchstens so groß ist, wie der größere relative Fehler der Näherungen $\widetilde{x} = x+\Delta x$ und $\widetilde{y} = y+\Delta y$ für x bzw. y. Also sollte das Problem mit Sicherheit gut konditioniert sein. Wir zeigen dieses nun mit der obigen Definition:

Wir haben hier vektorwertige Eingangsdaten $(x,y) \in \mathbb{R}^2$ und nutzen auf $\mathbb{R}^2$ die 1-Norm $\|(x,y)\|_1 := |x| + |y|$. Mit den Näherungen $\widetilde{x} = x + \Delta x$ und $\widetilde{y} = y + \Delta y$ von x bzw. y gilt für den absoluten Fehler

$$\begin{aligned}\mathrm{Abs}\big(f(\widetilde{x},\widetilde{y})\big) &= |f(\widetilde{x},\widetilde{y}) - f(x,y)| = \big|\big((x+\Delta x)+(y+\Delta y)\big) - (x+y)\big| \\ &= |\Delta x + \Delta y| \leq |\Delta x| + |\Delta y| = \|(\Delta x, \Delta y)\|_1 \qquad \text{für alle } (\Delta x, \Delta y) \in \mathbb{R}^2,\end{aligned}$$

und wir lesen ab, dass die absolute Konditionszahl $\kappa_{\mathrm{abs}}(x,y) \leq 1$ ist. (Genauer gilt sogar $\kappa_{\mathrm{abs}}(x,y) = 1$.)

Wenn x und y mit $x + y \neq 0$ das gleiche Vorzeichen haben, gilt $|x+y| = |x| + |y|$. Division durch $|f(x,y)| = |x+y| = |x| + |y|$ liefert dann

$$\begin{aligned}\mathrm{Rel}\big(f(\widetilde{x},\widetilde{y})\big) &= \frac{|f(\widetilde{x},\widetilde{y}) - f(x,y)|}{|f(x,y)|} = \frac{|\Delta x + \Delta y|}{|x| + |y|} \\ &\leq \frac{|\Delta x| + |\Delta y|}{|x| + |y|} = \frac{\|(\Delta x, \Delta y)\|_1}{\|(x,y)\|_1} \qquad \text{für alle } (\Delta x, \Delta y) \in \mathbb{R}^2,\end{aligned}$$

und wir lesen ab, dass die relative Konditionszahl $\kappa(x) \leq 1$ erfüllt. Die Addition zweier Zahlen mit dem gleichen Vorzeichen ist also gut konditioniert.

(b) Betrachten wir nun noch einmal die Addition zweier Zahlen mit verschiedenen Vorzeichen, die betraglich ungefähr gleich groß sind. Für die Zahlen

aus Beispiel 1.11 (b) $x = 10^6 + 1$ und $y = -10^6$ mit den Näherungswerten $\widetilde{x} = 10^6 + \frac{3}{2}$ und $\widetilde{y} = -10^6 + \frac{1}{2}$ (also $\Delta x = \Delta y = \frac{1}{2}$) finden wir

$$\begin{aligned}
&\frac{|f(\widetilde{x},\widetilde{y}) - f(x,y)|}{|f(x,y)|} = \frac{|(\widetilde{x}+\widetilde{y}) - (x+y)|}{|x+y|} \\
&= \frac{\big|\big((10^6+\frac{3}{2}) + (-10^6+\frac{1}{2})\big) - \big((10^6+1)+(-10^6)\big)\big|}{|(10^6+1)+(-10^6)|} = \frac{1}{1} = 1 \\
&\gg \frac{\|(\widetilde{x}-x, \widetilde{y}-y)\|_1}{\|(x,y)\|_1} = \frac{\big\|(\frac{1}{2},\frac{1}{2})\big\|_1}{\|(10^6+1,-10^6)\|_1} = \frac{\frac{1}{2}+\frac{1}{2}}{10^6+1+10^6} \approx \frac{10^{-6}}{2}
\end{aligned}$$

wobei wir wieder die 1-Norm $\|(x,y)\|_1 := |x| + |y|$ verwendet haben. Offensichtlich ist die Addition von $x = 10^6 + 1$ und $y = -10^6$ im Sinne von Definition 1.13 schlecht konditioniert.

(c) In Beispiel 1.12 haben wir die Fehlerfortpflanzung bei der reellwertigen Funktion $f : \mathbb{R} \to \mathbb{R}$, $f(x) = e^{3x^2}$, untersucht. Für kleine x und eine Näherung/Störung $\widetilde{x} = x + \Delta x$ des Wertes x fanden wir (für hinreichend kleines x und Δx) in (1.19)

$$\mathrm{Abs}\big(f(\widetilde{x})\big) = |f(\widetilde{x}) - f(x)| = |f(x+\Delta x) - f(x)| \approx 6\,|x|\,e^{3x^2}\,|\Delta x|. \tag{1.23}$$

Daraus folgt (vgl. auch (1.20))

$$\mathrm{Rel}\big(f(\widetilde{x})\big) = \frac{|f(\widetilde{x}) - f(x)|}{|f(x)|} \approx \frac{|6\,x\,e^{3x^2}\,\Delta x|}{|e^{3x^2}|} = 6\,|x|^2\,\frac{|\Delta x|}{|x|}. \tag{1.24}$$

An dieser Formel sehen wir, dass nur für kleines $x \in \mathbb{R}$ (z.B. alle x mit $|x| < \frac{1}{2}$) die Funktionsauswertung $f(x)$ gut konditioniert ist. Für $x \in \mathbb{R}$ mit $|x| \geq 6$ gilt $6\,x^2 \geq 6^3 = 216$, und die Funktionsauswertung $f(x)$ ist sicherlich schlecht konditioniert. Hinzu kommt, dass die Näherungen in (1.23) und (1.24) für $|x| \geq 6$ nicht mehr passend (bzw. nur sehr grob) sind.

Das Konzept der Kondition wird uns noch öfter begegnen. ♠

Definition 1.15. (Stabilität eines Algorithmus)
*Seien $D \subseteq \mathbb{R}^m$ und $Y \subseteq \mathbb{R}^n$. Sei $\mathbf{f} : D \to Y$ eine vektorwertige (und für $n = 1$ reellwertige) Funktion, die eine mathematische Aufgabe beschreibt. Sei $\widetilde{\mathbf{f}}$ ein Algorithmus, der (näherungsweise), die mathematische Aufgabe $\mathbf{f}$ ausführt. Sei $\mathbf{x} \in D$ mit $\mathbf{f}(\mathbf{x}) \neq \mathbf{0}$, und sei $\widetilde{\mathbf{x}} = \mathbf{x} + \Delta\mathbf{x}$ immer eine **Näherung oder***

Störung von $\mathbf{x}$. *Der Algorithmus* $\widetilde{\mathbf{f}}(\mathbf{x})$ *heißt* ***instabil (an der Stelle*** $\mathbf{x}$***)****, wenn es eine Näherung oder Störung* $\widetilde{\mathbf{x}}$ *von* $\mathbf{x}$ *gibt, für die gilt*

$$\frac{\|\widetilde{\mathbf{f}}(\widetilde{\mathbf{x}}) - \mathbf{f}(\mathbf{x})\|}{\|\mathbf{f}(\mathbf{x})\|} \gg \frac{\|\mathbf{f}(\widetilde{\mathbf{x}}) - \mathbf{f}(\mathbf{x})\|}{\|\mathbf{f}(\mathbf{x})\|} \,. \tag{1.25}$$

Ist $\widetilde{\mathbf{f}}(\mathbf{x})$ ***nicht instabil (an der Stelle*** $\mathbf{x}$***)****, so heißt* $\widetilde{\mathbf{f}}(\mathbf{x})$ ***stabil (an der Stelle*** $\mathbf{x}$***)****. (In* (1.25) *ist* $\|\cdot\|$ *der Absolutbetrag, falls* $n = 1$ *ist, und ansonsten eine geeignete Norm für* $\mathbb{R}^n$*.)*

Was bedeutet (1.25)**?** Die Formel (1.25) bedeutet, dass für eine Näherung $\widetilde{\mathbf{x}}$ der durch Rundungsfehler und den Algorithmus $\widetilde{\mathbf{f}}$ verursachte relative Fehler $\frac{\|\widetilde{\mathbf{f}}(\widetilde{\mathbf{x}})-\mathbf{f}(\mathbf{x})\|}{\|\mathbf{f}(\mathbf{x})\|}$ sehr viel größer ist als der nur durch die Näherung $\widetilde{\mathbf{x}}$ verursachte relative Fehler $\frac{\|\mathbf{f}(\widetilde{\mathbf{x}})-\mathbf{f}(\mathbf{x})\|}{\|\mathbf{f}(\mathbf{x})\|}$ (bei der exakten Ausführung der Aufgabe).

Betrachten wir zwei Beispiele. In dem ersten Beispiel wollen wir uns den Unterschied zwischen der Aufgabe $\mathbf{f}$ und dem Algorithmus $\widetilde{\mathbf{f}}$ zur (angenäherten) Lösung dieser Aufgabe deutlich machen, damit klar wird, warum wir sowohl das Konzept Kondition als auch das Konzept Stabilität brauchen.

Beispiel 1.16. (Stabilität eines Algorithmus)

Gegeben sei eine invertierbare Matrix $\mathbf{A} \in \mathbb{R}^{n\times n}$. Die Aufgabe $\mathbf{f}$ ist das Lösen des linearen Gleichungssystem $\mathbf{A}\,\mathbf{x} = \mathbf{b}$ für eine gegebene rechte Seite $\vec{\mathbf{b}} \in \mathbb{R}^n$.

Da $\mathbf{A}$ invertierbar ist, kann man diese Aufgabe theoretisch exakt lösen, und die exakte Lösung ist durch die Formel $\mathbf{f}(\mathbf{b}) = \mathbf{A}^{-1}\,\mathbf{b}$ gegeben. Bei der Frage nach der Kondition von $\mathbf{f}(\mathbf{b})$ interessiert man sich also dafür, wie groß der relative Fehler für eine gestörte rechte Seite $\widetilde{\mathbf{b}} = \mathbf{b} + \Delta\mathbf{b}$

$$\frac{\|\mathbf{f}(\widetilde{\mathbf{b}}) - \mathbf{f}(\mathbf{b})\|}{\|\mathbf{f}(\mathbf{b})\|} = \frac{\|\mathbf{f}(\mathbf{b}+\Delta\mathbf{b}) - \mathbf{f}(\mathbf{b})\|}{\|\mathbf{f}(\mathbf{b})\|} = \frac{\|\mathbf{A}^{-1}\,(\mathbf{b}+\Delta\mathbf{b}) - \mathbf{A}^{-1}\,\mathbf{b}\|}{\|\mathbf{A}^{-1}\,\mathbf{b}\|} = \frac{\|\mathbf{A}^{-1}\,\Delta\mathbf{b}\|}{\|\mathbf{A}^{-1}\,\mathbf{b}\|}$$

verglichen mit dem relativen Fehler $\frac{\|\widetilde{\mathbf{b}}-\mathbf{b}\|}{\|\mathbf{b}\|} = \frac{\|\Delta\mathbf{b}\|}{\|\mathbf{b}\|}$ der gestörten rechten Seite $\widetilde{\mathbf{b}} = \mathbf{b} + \Delta\mathbf{b}$ ist. Dabei ist $\|\cdot\|$ beispielsweise die Euklidische Norm $\|\mathbf{y}\| = \sqrt{|y_1|^2 + \ldots + |y_n|^2}$. Wir werden im nächsten Kapitel sehen, dass

$$\frac{\|\mathbf{f}(\mathbf{b}+\Delta\mathbf{b}) - \mathbf{f}(\mathbf{b})\|}{\|\mathbf{f}(\mathbf{b})\|} = \frac{\|\mathbf{A}^{-1}\,\Delta\mathbf{b}\|}{\|\mathbf{A}^{-1}\,\mathbf{b}\|} \leq \kappa(\mathbf{b})\,\frac{\|\Delta\mathbf{b}\|}{\|\mathbf{b}\|}$$

mit $\kappa(\mathbf{b}) \leq \|\mathbf{A}\|\,\|\mathbf{A}^{-1}\|$ gilt, wobei $\|\mathbf{A}\|$ und $\|\mathbf{A}^{-1}\|$ jeweils die sogenannte induzierte Matrixnorm von $\mathbf{A}$ bzw. $\mathbf{A}^{-1}$ sind.

Berechnet man die Lösung von $\mathbf{A}\,\mathbf{x} = \mathbf{b}$ in der Praxis, so könnte man als Algorithmus das Gaußsche Eliminationsverfahren verwenden, welches nun unser $\widetilde{\mathbf{f}}$ ist. Dabei passieren aber durch das Gleitkomma-Zahlensystem im Computer bei der Berechnung Rundungsfehler, so dass das Ergebnis des Algorithmus $\widetilde{\mathbf{f}}(\mathbf{b})$ in der Regel ungleich $\mathbf{f}(\mathbf{b})$ ist. Bei der Frage der Stabilität interessiert man sich dafür, ob der relative Fehler

$$\frac{\|\widetilde{\mathbf{f}}(\widetilde{\mathbf{b}}) - \mathbf{f}(\mathbf{b})\|}{\|\mathbf{f}(\mathbf{b})\|}$$

durch den Algorithmus $\widetilde{\mathbf{f}}$ für irgendeine Störung $\widetilde{\mathbf{b}}$ von $\mathbf{b}$ deutlich größer ist als der relative Fehler

$$\frac{\|\mathbf{f}(\widetilde{\mathbf{b}}) - \mathbf{f}(\mathbf{b})\|}{\|\mathbf{f}(\mathbf{b})\|}$$

der (exakt gelösten) Aufgabe $\mathbf{f}$ für diese Störung $\widetilde{\mathbf{b}}$ ♠

Es ist klar, dass man auch bei einem gut konditionierten Problem $\mathbf{f}$ nur dann eine brauchbare Lösung erwarten kann, wenn der Algorithmus $\widetilde{\mathbf{f}}$ stabil ist.

Beispiel 1.17. (stabiler und instabiler Algorithmus)

In Beispiel 1.7 wurde für $x = 10^5$

$$f(x) = x\left(\sqrt{x+1} - \sqrt{x}\right) \tag{1.26}$$

in einem Gleitkomma-Zahlensystem mit der Mantissenlänge 6 auf zwei verschiedene Arten berechnet. Einmal wurde die gegebene Formel (1.26) für f verwendet und einmal die Umformung (vgl. (1.11) in Beispiel 1.7)

$$f(x) = \frac{x}{\sqrt{x+1} + \sqrt{x}} \tag{1.27}$$

verwendet. Durch die Rundung im Gleitkomma-Zahlensystem mit der Mantissenlänge 6 liefern uns (1.26) und (1.27) jeweils einen Algorithmus den wir mit $\widetilde{f}_1$ (für (1.26)) bzw. mit $\widetilde{f}_2$ (für (1.27)) bezeichnen wollen.

Wir betrachten nun $x = 10^5$ und die Näherung (oder Störung) $\widetilde{x} = 10^5 + 1$ von x. Dann sind die exakten Werte (mit Rundung auf eine Mantisse der Länge 6)

$$f(x) \doteq 158{,}113 \qquad \text{und} \qquad f(\widetilde{x}) \doteq 158{,}114.$$

Wie berechnen nun $\widetilde{f}_1(\widetilde{x})$ und $\widetilde{f}_2(\widetilde{x})$:

$$\widetilde{f}_1(\widetilde{x}) = \widetilde{f}_1(10^5 + 1) = 100001 \cdot \left(\sqrt{100002} - \sqrt{100001}\right)$$

$$= 100001 \cdot \big(\underbrace{316{,}231 - 316{,}229}_{= 0{,}002} \big) = 100001 \cdot 0{,}002 = 200{,}002,$$

$$\widetilde{f}_2(\widetilde{x}) = \widetilde{f}_2(10^5 + 1) = \frac{100001}{\sqrt{100002} + \sqrt{100001}}$$
$$= \frac{100001}{316{,}231 + 316{,}229} = \frac{100001}{632{,}46} = 158{,}114.$$

Wir berechnen nun jeweils beide Seiten in (1.25):

$$\frac{|\widetilde{f}_1(\widetilde{x}) - f(x)|}{|f(x)|} = \frac{|200{,}002 - 158{,}113|}{|158{,}113|} \doteq 0{,}2649,$$
$$\frac{|\widetilde{f}_2(\widetilde{x}) - f(x)|}{|f(x)|} = \frac{|158{,}114 - 158{,}113|}{|158{,}113|} \doteq 0{,}6325 \cdot 10^{-5},$$
$$\frac{|f(\widetilde{x}) - f(x)|}{|f(x)|} = \frac{|158{,}114 - 158{,}113|}{|158{,}113|} \doteq 0{,}6325 \cdot 10^{-5}.$$

Der Algorithmus $\widetilde{f}_1(x)$ ist an der Stelle $x = 10^5$ offenbar instabil, denn es gilt

$$\frac{|\widetilde{f}_1(\widetilde{x}) - f(x)|}{|f(x)|} \doteq 0{,}2649 \gg \frac{|f(\widetilde{x}) - f(x)|}{|f(x)|} \doteq 0{,}6325 \cdot 10^{-5}.$$

Der Algorithmus $\widetilde{f}_2(x)$ ist an der Stelle $x = 10^5$ offenbar stabil, denn es gilt

$$\frac{|\widetilde{f}_1(\widetilde{x}) - f(x)|}{|f(x)|} \doteq 0{,}6325 \cdot 10^{-5} = \frac{|f(\widetilde{x}) - f(x)|}{|f(x)|} \doteq 0{,}6325 \cdot 10^{-5}.$$

Wir sind nicht darüber überrascht, denn wir hatten bereits in Beispiel 1.7 gesehen, wie schlecht das Ergebnis der Berechnung von $f(10^5)$ mit (1.26) in Gleitkomma-Zahlensystem mit der Mantissenlänge 6 war. ♠

Sei die Notation wie in Definition 1.15. Mit der Dreiecksungleichung folgt

$$\big\|\widetilde{\mathbf{f}}(\widetilde{\mathbf{x}}) - \mathbf{f}(\mathbf{x})\big\| = \big\| \big(\widetilde{\mathbf{f}}(\widetilde{\mathbf{x}}) - \mathbf{f}(\widetilde{\mathbf{x}})\big) + \big(\mathbf{f}(\widetilde{\mathbf{x}}) - \mathbf{f}(\mathbf{x})\big) \big\| \leq \big\|\widetilde{\mathbf{f}}(\widetilde{\mathbf{x}}) - \mathbf{f}(\widetilde{\mathbf{x}})\big\| + \big\|\mathbf{f}(\widetilde{\mathbf{x}}) - \mathbf{f}(\mathbf{x})\big\|.$$

Division durch $\|\mathbf{f}(\mathbf{x})\|$ liefert:

$$\boxed{\frac{\big\|\widetilde{\mathbf{f}}(\widetilde{\mathbf{x}}) - \mathbf{f}(\mathbf{x})\big\|}{\|\mathbf{f}(\mathbf{x})\|} \leq \frac{\big\|\widetilde{\mathbf{f}}(\widetilde{\mathbf{x}}) - \mathbf{f}(\widetilde{\mathbf{x}})\big\|}{\|\mathbf{f}(\mathbf{x})\|} + \frac{\big\|\mathbf{f}(\widetilde{\mathbf{x}}) - \mathbf{f}(\mathbf{x})\big\|}{\|\mathbf{f}(\mathbf{x})\|}} \tag{1.28}$$

Der zweite Term auf der rechten Seite von (1.28) wird bei der Bewertung der **Kondition** von $\mathbf{f}(\mathbf{x})$ mit $\frac{\|\widetilde{\mathbf{x}} - \mathbf{x}\|_{(m)}}{\|\mathbf{x}\|_{(m)}}$ verglichen, wobei $\|\cdot\|_{(m)}$ eine Norm für $\mathbb{R}^m$

ist. Nur wenn $\frac{\|\mathbf{f}(\widetilde{\mathbf{x}})-\mathbf{f}(\mathbf{x})\|}{\|\mathbf{f}(\mathbf{x})\|}$ eine ähnliche Größenordnung wie $\frac{\|\widetilde{\mathbf{x}}-\mathbf{x}\|_{(m)}}{\|\mathbf{x}\|_{(m)}}$ hat, ist $\mathbf{f}(\mathbf{x})$ gut konditioniert, und es macht überhaupt Sinn, den Algorithmus $\widetilde{\mathbf{f}}$ zur Lösung der mathematischen Aufgabe $\mathbf{f}$ weiter zu untersuchen.

Wenn die mathematische Aufgabe $\mathbf{f}(\mathbf{x})$ gut konditioniert ist, dann gilt $\mathbf{f}(\widetilde{\mathbf{x}}) \approx \mathbf{f}(\mathbf{x})$ und somit gilt für den ersten Term auf der rechten Seite von (1.28)

$$\frac{\|\widetilde{\mathbf{f}}(\widetilde{\mathbf{x}}) - \mathbf{f}(\widetilde{\mathbf{x}})\|}{\|\mathbf{f}(\mathbf{x})\|} \approx \frac{\|\widetilde{\mathbf{f}}(\widetilde{\mathbf{x}}) - \mathbf{f}(\widetilde{\mathbf{x}})\|}{\|\mathbf{f}(\widetilde{\mathbf{x}})\|}\,. \tag{1.29}$$

Der Term auf der rechten Seite von (1.29) beschreibt den relativen Fehler der Näherung $\widetilde{\mathbf{f}}(\widetilde{\mathbf{x}})$ für $\mathbf{f}(\widetilde{\mathbf{x}})$. Da in diesem Term überall $\widetilde{\mathbf{x}}$ steht, wird hier nur die Qualität der Näherung der mathematischen Aufgabe $\mathbf{f}$ durch den Algorithmus $\widetilde{\mathbf{f}}$ bewertet.

Alternativ folgt für eine gut konditionierte mathematische Aufgabe $\mathbf{f}(\mathbf{x})$ aus $\mathbf{f}(\widetilde{\mathbf{x}}) \approx \mathbf{f}(\mathbf{x})$ für den ersten Term auf der rechten Seite von (1.28) auch

$$\frac{\|\widetilde{\mathbf{f}}(\widetilde{\mathbf{x}}) - \mathbf{f}(\widetilde{\mathbf{x}})\|}{\|\mathbf{f}(\mathbf{x})\|} \approx \frac{\|\widetilde{\mathbf{f}}(\widetilde{\mathbf{x}}) - \mathbf{f}(\mathbf{x})\|}{\|\mathbf{f}(\mathbf{x})\|}\,,$$

und wir erhalten den Term auf der linken Seite von (1.25), mit dem die Stabilität von $\widetilde{\mathbf{f}}(\widetilde{\mathbf{x}})$ bewertet wird.

1.7 Rechenaufwand eines numerischen Verfahrens

Definition 1.18. (Aufwand)
*Sei $\mathbf{f}$ eine mathematische Aufgabe, und sei $\widetilde{\mathbf{f}}$ ein Algorithmus zur Lösung dieser Aufgabe. Unter dem **Aufwand** von $\widetilde{\mathbf{f}}$ versteht man die **Anzahl der benötigten elementaren Rechenoperationen** (Additionen/Subtraktionen und Multiplikationen/Divisionen).*

Der Aufwand wird **nicht exakt** berechnet, sondern man gibt **nur seine Größenordnung** an. Betrachten wir dazu ein Beispiel.

Beispiel 1.19. (Aufwand)
Sei $\mathbf{A} \in \mathbb{R}^{n\times n}$ eine $n \times n$-Matrix, und sei $\mathbf{f} : \mathbb{R}^n \to \mathbb{R}^n$, $\mathbf{f}(\mathbf{x}) = \mathbf{A}\,\mathbf{x}$, die Matrix-Vektor-Multiplikation der Matrix $\mathbf{A}$ mit dem Vektor $\mathbf{x} \in \mathbb{R}^n$. Mit $\widetilde{\mathbf{f}}$ bezeichnen wir die numerische Berechnung von $\mathbf{f}$ (einschließlich der durch Rundung verursachten

Fehler). Die numerische Berechnung der k-ten Komponente von $\mathbf{A}\,\mathbf{x}$ erfordert n Multiplikationen und $n-1$ Additionen, also insgesamt $2n-1$ elementare Rechenoperationen. Also benötigt die numerische Berechnung $\widetilde{\mathbf{f}}(\mathbf{x})$ von $\mathbf{f}(\mathbf{x}) = \mathbf{A}\,\mathbf{x}$ insgesamt $n \cdot (2n-1) \approx 2\,n^2$ elementare Rechenoperationen. Der Aufwand des Algorithmus $\widetilde{\mathbf{f}}$ hat also die Größenordnung n^2. ♠

Damit wir beim Aufwand nicht umständlich „von der Größenordnung ...“ reden müssen, führen wir das **Landau Symbol** $\mathcal{O}$ ein: Seien $f : \mathbb{N} \to \mathbb{R}$ und $g : \mathbb{N} \to \mathbb{R}$ zwei Funktionen. Dann gilt $f = \mathcal{O}(g)$ per Definition genau dann, wenn es eine Konstante $C > 0$ und ein $N \in \mathbb{N}$ gibt, so dass gilt

$$|f(n)| \leq C\,|g(n)| \qquad \text{für alle } n \in \mathbb{N} \text{ mit } n \geq N.$$

Beispiel 1.20. (Landau Symbol)
Sei $\mathbf{A} \in \mathbb{R}^{n\times n}$ eine $n \times n$-Matrix, und sei $\mathbf{f} : \mathbb{R}^n \to \mathbb{R}^n$, $\mathbf{f}(\mathbf{x}) = \mathbf{A}\,\mathbf{x}$, die Matrix-Vektor-Multiplikation der Matrix $\mathbf{A}$ mit dem Vektor $\mathbf{x} \in \mathbb{R}^n$. Mit $\widetilde{\mathbf{f}}$ bezeichnen wir die numerische Berechnung von $\mathbf{f}$ (einschließlich der durch Rundung verursachten Fehler). In Beispiel 1.19 hatten wir uns überlegt, dass die numerische Berechnung $\widetilde{\mathbf{f}}(\mathbf{x})$ von $\mathbf{f}(\mathbf{x}) = \mathbf{A}\,\mathbf{x}$ insgesamt $n \cdot (2n-1)$ elementare Rechenoperationen benötigt. Mit dem Landau Symbol, gilt nun, dass die Berechnung $\mathcal{O}(n^2)$ elementare Rechenoperationen benötigt, denn es gilt

$$|n\,(2n-1)| = n\,(2n-1) \leq n\,2\,n = 2\,n^2 \leq C\,n^2 = C\,|n^2| \qquad \text{für alle } n \in \mathbb{N}$$

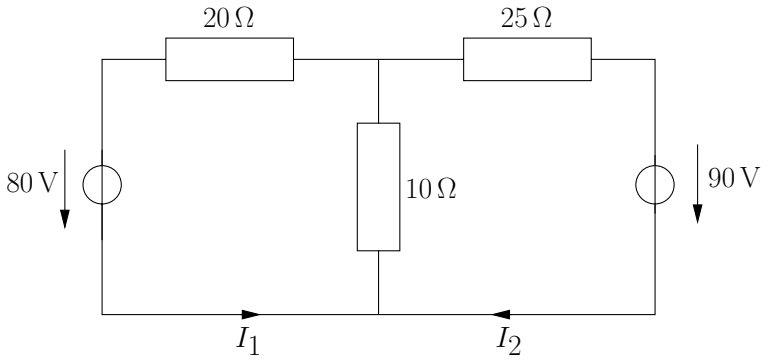
mit der Konstante $C = 2$. ♠

KAPITEL 2

Direkte Lösungsverfahren für lineare Gleichungssysteme

Betrachten wir als Einstig ein Anwendungsbeispiel aus der Elektrotechnik.

Das Bild rechts zeigt einen Schaltkreis. Gesucht sind die Stromstärken I_1 und I_2 in A (= Ampere).



Ohmsches Gesetz:

$$R = \frac{U}{I} \quad \Longleftrightarrow \quad U = R\,I$$

wobei U = Spannung, I = Stromstärke, R = Widerstand.

Erstes Kirchhoffsches Gesetz: In jeder Masche ist die Summe der Teilspannungen an den Leitern (Widerständen, Verbrauchern) gleich der Summe der Spannungen der eingeschalteten Stromquellen.

Das erste Kirchhoffsche Gesetz (Maschensatz) liefert:

$$\left.\begin{array}{rcl} 20 \cdot I_1 + 10 \cdot (I_1 + I_2) & = & 80 \\ 25 \cdot I_2 + 10 \cdot (I_1 + I_2) & = & 90 \end{array}\right\} \quad \Longleftrightarrow \quad \left\{\begin{array}{l} 3\,I_1 + \quad I_2 = 8 \\ I_1 + 3{,}5\,I_2 = 9 \end{array}\right.$$

Wir erhalten also ein lineares Gleichungssystem, dessen Lösungen die gesuchten Stromstärken I_1 und I_2 sind.

Bevor wir uns dem Lösen linearer Gleichungssysteme widmen können, müssen wir im nächsten Teilkapitel einige Vorbereitungen treffen.

2.1 Normen

Als Vorbereitung für die direkten Lösungsverfahren linearer Gleichungssysteme benötigen wir den Begriff einer Norm eines $\mathbb{R}$-Vektorraums.

Definition 2.1. (Norm eines Vektorraums)
*Sei V ein $\mathbb{R}$-Vektorraum mit dem Nullvektor 0_V. Eine **Norm** auf V ist eine Abbildung $\|\cdot\| : V \to [0;\infty[$, welche die folgenden Bedingungen erfüllt:*

(1) Für $x \in V$ gilt: $\|x\| = 0 \implies x = 0_V$

(2) $\|\lambda\, x\| = |\lambda|\, \|x\|$ für alle $x \in V$ und alle Skalare $\lambda \in \mathbb{R}$.

(3) $\|x + y\| \leq \|x\| + \|y\|$ für alle $x, y \in V$ (Dreiecksungleichung).

Beachten Sie, dass für eine Norm $\|\cdot\|$ immer gilt

$$\|x\| \geq 0 \qquad \text{für alle } x \in V, \tag{2.1}$$

denn die Zielmenge der Abbildung/Funktion $\|\cdot\|$ ist $[0;\infty[$. Die **„Nichtnegativität" der Norm** (2.1) ist mit zu überprüfen, wenn man nachweist, dass eine gegebene Abbildung/Funktion $\|\cdot\|$ auf einem Vektorraum eine Norm ist.

Betrachten wir einige Beispiele für Normen.

Beispiel 2.2. (Normen auf einem Vektorraum)

(a) Der $\mathbb{R}$-Vektorraum $\mathbb{R}^n$ kann mit den **p-Normen** (oder **ℓ_p-Normen**) $\|\cdot\|_p$, wobei $1 \leq p \leq \infty$, versehen werden:

$$\|\mathbf{x}\|_p := \begin{cases} \left(\sum_{j=1}^{n} |x_j|^p \right)^{1/p} & \text{für } 1 \leq p < \infty, \\ \max\limits_{j=1,\ldots,n} |x_j| & \text{für } p = \infty. \end{cases}$$

Die **2-Norm** kennen Sie schon aus Ihrer Mathematik-Vorlesung; sie heißt die **Euklidische Norm** und ist die „Standardnorm" für $\mathbb{R}^n$. Sehr wichtig für uns sind auch noch die **1-Norm** und die **∞-Norm** (welche auch **Maximumsnorm** genannt wird). – *Zahlenbeispiel:* Für

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

berechnen sich die 2-Norm, die 1-Norm und die ∞-Norm wie folgt:

$$\|\mathbf{x}\|_2 = \sqrt{(-2)^2 + 1^2 + 3^2} = \sqrt{14},$$
$$\|\mathbf{x}\|_1 = |-2| + |1| + |3| = 6,$$
$$\|\mathbf{x}\|_\infty = \max\big\{|-2|; |1|; |3|\big\} = 3.$$

(b) Der $\mathbb{R}$-Vektorraum aller Polynome $\mathbb{P}_n([-1;1])$ vom Grad $\leq n$ auf dem Intervall $[-1;1]$ kann beispielsweise mit der **Supremumsnorm** (auch L_∞-Norm genannt)

$$\|p\|_\infty := \max_{x\in[-1,1]} |p(x)|$$

oder mit der **L_2-Norm**

$$\|p\|_2 := \left(\int_{-1}^{1} |p(x)|^2 \,\mathrm{d}x\right)^{1/2}$$

versehen werden.

Wir lernen später noch weitere Beispiele kennen. ♠

Definition 2.3. (äquivalente Normen)
*Sei V ein $\mathbb{R}$-Vektorraum, und seien $\|\cdot\|$ und $|||\cdot|||$ zwei Normen auf V. Die beiden Normen $\|\cdot\|$ und $|||\cdot|||$ heißen **äquivalent**, wenn es eine Konstante $c > 0$ gibt, so dass gilt*

$$c^{-1}\,\|x\| \leq |||x||| \leq c\,\|x\| \qquad \text{für alle } x \in V. \tag{2.2}$$

Wir bemerken, dass aus (2.2) durch Umformen auch folgt

$$c^{-1}\,|||x||| \leq \|x\| \leq c\,|||x||| \qquad \text{für alle } x \in V,$$

so dass die beiden Normen in (2.2) vertauschbar sind.

Das folgende Resultat ermöglicht es uns, eine „geeignete“ Norm auf $\mathbb{R}^n$ zu wählen.

Hilfssatz 2.4. (Äquivalenz aller Normen auf $\mathbb{R}^n$)
*Auf $\mathbb{R}^n$ sind alle Normen **äquivalent**. Insbesondere sind alle p-Normen **äquivalent**, d.h. zu $1 \leq p, q \leq \infty$ es gibt eine Konstante $c_{p,q} > 0$, so dass gilt*

$$c_{p,q}^{-1}\,\|\mathbf{x}\|_p \leq \|\mathbf{x}\|_q \leq c_{p,q}\,\|\mathbf{x}\|_p \qquad \text{für alle } \mathbf{x} \in \mathbb{R}^n.$$

Wir führen nun mit Hilfe von Normen für $\mathbb{R}^n$ und $\mathbb{R}^m$ eine „induzierte“ Matrixnorm auf dem $\mathbb{R}$-Vektorraum $\mathbb{R}^{m\times n}$ aller $m \times n$-Matrizen mit reellen Einträgen/Koeffizienten ein.

Definition 2.5. (induzierte Matrixnorm und Spektralradius)

*(1) Seien $\|\cdot\|_{(m)}$ eine Norm auf $\mathbb{R}^m$ und $\|\cdot\|_{(n)}$ eine Norm auf $\mathbb{R}^n$. Dann ist die (zugehörige) **induzierte Matrixnorm** auf $\mathbb{R}^{m\times n}$ definiert durch*

$$\|\mathbf{A}\|_{(m,n)} := \max_{\substack{\mathbf{x}\in\mathbb{R}^n,\\ \|\mathbf{x}\|_{(n)}=1}} \|\mathbf{A}\,\mathbf{x}\|_{(m)} \qquad \text{für } \mathbf{A} \in \mathbb{R}^{m\times n}.$$

*(2) Sei $\mathbf{B} \in \mathbb{R}^{n\times n}$ eine **quadratische Matrix** mit den n nicht notwendigerweise verschiedenen (reellen oder komplexen) Eigenwerten $\lambda_1, \lambda_2, \ldots, \lambda_n \in \mathbb{C}$. Der **Spektralradius von $\mathbf{B}$** ist definiert durch*

$$\varrho(\mathbf{B}) := \max_{j=1,\ldots,n} |\lambda_j|.$$

(Zur Erinnerung: $\lambda \in \mathbb{C}$ ist ein Eigenwert von $\mathbf{B} \in \mathbb{R}^{n\times n}$, wenn es einen Vektor $\mathbf{x} \in \mathbb{C}^n \setminus \{\mathbf{0}\}$ gibt mit $\mathbf{B}\,\mathbf{x} = \lambda\,\mathbf{x}$.)

Wir können die induzierte Matrixnorm auch mittels

$$\|\mathbf{A}\| = \max_{\mathbf{x}\in\mathbb{R}^n\setminus\{\mathbf{0}\}} \frac{\|\mathbf{A}\,\mathbf{x}\|_{(m)}}{\|\mathbf{x}\|_{(n)}} \tag{2.3}$$

berechnen, denn

$$\frac{\|\mathbf{A}\,\mathbf{x}\|_{(m)}}{\|\mathbf{x}\|_{(n)}} = \left\|\mathbf{A}\,\frac{\mathbf{x}}{\|\mathbf{x}\|_{(n)}}\right\|_{(m)} = \|\mathbf{A}\,\mathbf{y}\|_{(m)} \qquad \text{mit} \qquad \mathbf{y} := \frac{\mathbf{x}}{\|\mathbf{x}\|_{(n)}}$$

und

$$\|\mathbf{y}\|_{(n)} = \left\|\frac{\mathbf{x}}{\|\mathbf{x}\|_{(n)}}\right\|_{(n)} = \frac{\|\mathbf{x}\|_{(n)}}{\|\mathbf{x}\|_{(n)}} = 1.$$

Im nächsten Hilfssatz lernen wir einige wichtige Beispiele für induzierte Matrixnormen kennen.

Hilfssatz 2.6. (wichtige induzierte Matrixnormen)

*(1) Sind $\mathbb{R}^m$ und $\mathbb{R}^n$ mit der **1-norm** $\|\mathbf{x}\|_1 = |x_1| + |x_2| + \ldots + |x_\ell|$ mit $\ell = m$ bzw. $\ell = n$ versehen, so ist die induzierte Matrixnorm für $\mathbb{R}^{m\times n}$*

die sogenannte ***Spaltensummennorm***

$$\|\mathbf{A}\|_1 = \max_{j=1,\ldots,n} \left(\sum_{i=1}^{m} |a_{i,j}| \right), \qquad \mathbf{A} = [a_{i,j}] \in \mathbb{R}^{m\times n}.$$

(2) Sind $\mathbb{R}^m$ *und* $\mathbb{R}^n$ *mit der* $\boldsymbol{\infty}$***-norm*** $\|\mathbf{x}\|_\infty = \max\{|x_1|, |x_2|, \ldots, |x_\ell|\}$ *mit* $\ell = m$ *bzw.* $\ell = n$ *versehen, so ist die induzierte Matrixnorm für* $\mathbb{R}^{m\times n}$ *die sogenannte* ***Zeilensummennorm***

$$\|\mathbf{A}\|_\infty = \max_{i=1,\ldots,m} \left(\sum_{j=1}^{n} |a_{i,j}| \right), \qquad \mathbf{A} = [a_{i,j}] \in \mathbb{R}^{m\times n}.$$

(3) Sind $\mathbb{R}^m$ *und* $\mathbb{R}^n$ *mit der* $\mathbf{2}$***-norm*** $\|\mathbf{x}\|_2 = \sqrt{|x_1|^2 + |x_2|^2 + \ldots + |x_\ell|^2}$ *mit* $\ell = m$ *bzw.* $\ell = n$ *versehen, so ist die induzierte Matrixnorm für* $\mathbb{R}^{m\times n}$ *die sogenannte* ***Spektralnorm***

$$\|\mathbf{A}\|_2 = \sqrt{\varrho\big(\mathbf{A}^T\mathbf{A}\big)}, \qquad \mathbf{A} \in \mathbb{R}^{m\times n}.$$

(Dabei ist $\mathbf{A}^T$ *die transponierte Matrix von* $\mathbf{A}$*, und* ϱ *ist der in Definition 2.5 (2) eingeführte Spektralradius.)*

Beweisidee von Hilfssatz 2.6: Die grundsätzliche Vorgehensweise ist in allen drei Fällen wie folgt: $\|\cdot\|_p$ bezeichne jeweils die in Hilfssatz 2.6 (1), (2) und (3) gegebene Formel mit $p = 1$, $p = \infty$ bzw. $p = 2$. Durch geeignetes Abschätzen zeigt man, dass für jeden Vektor $\mathbf{x} \in \mathbb{R}^n$ mit $\|\mathbf{x}\|_p = 1$ gilt

$$\|\mathbf{A}\,\mathbf{x}\|_p \leq \|\mathbf{A}\|_p.$$

Daraus folgt dann, dass

$$\max_{\substack{\mathbf{x}\in\mathbb{R}^n,\\ \|\mathbf{x}\|_p=1}} \|\mathbf{A}\,\mathbf{x}\|_p \leq \|\mathbf{A}\|_p. \tag{2.4}$$

gilt. Nun muss man ein passendes $\mathbf{z} \in \mathbb{R}^n$ mit $\|\mathbf{z}\|_p = 1$ konstruieren, für das gilt

$$\|\mathbf{A}\,\mathbf{z}\|_p = \|\mathbf{A}\|_p. \tag{2.5}$$

Aus (2.4) und (2.5) folgt dann, dass $\|\mathbf{A}\|_p$ in der Tat die von der p-Norm $\|\cdot\|_p$ induzierte Matrixnorm ist. Der Beweis kann beispielsweise in [7, Theorem 2.37

und sein Beweis] nachgelesen werden. □

Betrachten wir einige Beispiele.

Beispiel 2.7. (induzierte Matrixnormen)
Gegeben seien die folgenden reellen Matrizen:

$$\mathbf{A} = \begin{bmatrix} -1 & 3 & 2 \\ 2 & -2 & 4 \\ 0 & 4 & -1 \\ 1 & -1 & -2 \end{bmatrix}, \qquad \mathbf{B} = \begin{bmatrix} 1 & -2 & 3 & -4 \\ -2 & 3 & -4 & 5 \\ 3 & -4 & 5 & -6 \end{bmatrix}, \qquad \mathbf{C} = \begin{bmatrix} 1 & 0 \\ -1 & -1 \\ 1 & 1 \\ -1 & 0 \end{bmatrix}.$$

(a) Die Spaltensummennorm von $\mathbf{A}$ bzw. $\mathbf{B}$ bzw. $\mathbf{C}$ ist jeweils

$$\|\mathbf{A}\|_1 = \max \left\{ \begin{array}{c} |-1|+|2|+|0|+|1|; \\ |3|+|-2|+|4|+|-1|; \\ |2|+|4|+|-1|+|-2| \end{array} \right\} = \max\{4;10;9\} = 10,$$

$$\|\mathbf{B}\|_1 = \max \left\{ \begin{array}{c} |1|+|-2|+|3|; \\ |-2|+|3|+|-4|; \\ |3|+|-4|+|5|; \\ |-4|+|5|+|-6| \end{array} \right\} = \max\{6;9;12;15\} = 15,$$

$$\|\mathbf{C}\|_1 = \max \left\{ \begin{array}{c} |1|+|-1|+|1|+|-1|; \\ |0|+|-1|+|1|+|0| \end{array} \right\} = \max\{4;2\} = 4.$$

(b) Die Zeilensummennorm von $\mathbf{A}$ bzw. $\mathbf{B}$ bzw. $\mathbf{C}$ ist jeweils

$$\|\mathbf{A}\|_\infty = \max \left\{ \begin{array}{c} |-1|+|3|+|2|; \\ |2|+|-2|+|4|; \\ |0|+|4|+|-1|; \\ |1|+|-1|+|-2| \end{array} \right\} = \max\{6;8;5;4\} = 8,$$

$$\|\mathbf{B}\|_\infty = \max \left\{ \begin{array}{c} |1|+|-2|+|3|+|-4|; \\ |-2|+|3|+|-4|+|5|; \\ |3|+|-4|+|5|+|-6| \end{array} \right\} = \max\{10;14;18\} = 18,$$

$$\|\mathbf{C}\|_\infty = \max \left\{ \begin{array}{c} |1|+|0|; \\ |-1|+|-1|; \\ |1|+|1|; \\ |-1|+|0| \end{array} \right\} = \max\{1;2;2;1\} = 2.$$

(c) Da die Berechnung der Spektralnorm aufwendiger ist, wollen wir diese nur für $\mathbf{C}$ berechnen. Wegen $\|\mathbf{C}\|_2 = \sqrt{\varrho\big(\mathbf{C}^T\mathbf{C}\big)}$ benötigen zunächst $\mathbf{C}^T\mathbf{C}$:

$$\mathbf{C}^T\mathbf{C} = \begin{bmatrix} 1 & -1 & 1 & -1 \\ 0 & -1 & 1 & 0 \end{bmatrix} \begin{bmatrix} 1 & 0 \\ -1 & -1 \\ 1 & 1 \\ -1 & 0 \end{bmatrix} = \begin{bmatrix} 4 & 2 \\ 2 & 2 \end{bmatrix}$$

Die Eigenwerte von $\mathbf{C}^T\mathbf{C}$ finden wir, indem wir die Nullstellen des charakteristischen Polynoms $p_{\mathbf{C}^T\mathbf{C}}(\lambda) := \det\big(\mathbf{C}^T\mathbf{C} - \lambda\,\mathbf{E}_2\big)$ von $\mathbf{C}^T\mathbf{C}$ bestimmen:

$$\begin{aligned} 0 &= \det\big(\mathbf{C}^T\mathbf{C} - \lambda\,\mathbf{E}_2\big) = \det\left(= \begin{bmatrix} 4-\lambda & 2 \\ 2 & 2-\lambda \end{bmatrix}\right) \\ &= (4-\lambda)\,(2-\lambda) - 4 = \lambda^2 - 6\,\lambda + 4 = (\lambda^2 - 6\lambda + 9) - 5 \\ &= (\lambda-3)^2 - \big(\sqrt{5}\big)^2 = (\lambda - 3 - \sqrt{5})\,(\lambda - 3 + \sqrt{5}) \end{aligned}$$

Also sind die Eigenwerte von $\mathbf{C}^T\mathbf{C}$ durch $\lambda_1 = 3 + \sqrt{5}$ und $\lambda_2 = 3 - \sqrt{5}$ gegeben. Damit ist der Spektralradius von $\mathbf{C}^T\mathbf{C}$

$$\varrho\big(\mathbf{C}^T\mathbf{C}\big) = \max\big\{|3+\sqrt{5}|; |3-\sqrt{5}|\big\} = \max\big\{3+\sqrt{5}; 3-\sqrt{5}\big\} = 3+\sqrt{5},$$

und die Spektralnorm von $\mathbf{C}$ ist

$$\|\mathbf{C}\|_2 = \sqrt{\varrho\big(\mathbf{C}^T\mathbf{C}\big)} = \sqrt{3+\sqrt{5}}.$$

Wir sehen an den Beispielen, dass die verschiedenen induzierten Matrixnormen für die gleiche Matrix in der Regel unterschiedliche Werte annehmen. ♠

Wir lernen zunächst die Eigenschaften der induzierten Matrixnorm kennen.

Hilfssatz 2.8. (Eigenschaften der induzierten Matrixnorm)
Seien $\|\cdot\|_{(m)}$ eine Norm auf $\mathbb{R}^m$ und $\|\cdot\|_{(n)}$ eine Norm auf $\mathbb{R}^n$. Dann hat die induzierte Matrixnorm $\|\cdot\|_{(m,n)}$ auf $\mathbb{R}^{m\times n}$ die folgenden Eigenschaften:

*(1) Die induzierte Matrixnorm $\|\cdot\|_{(m,n)}$ ist eine **Norm** für den Vektorraum $\mathbb{R}^{m\times n}$ der $m \times n$-Matrizen im Sinne von Definition 2.1.*

(2) Es gilt $\|\mathbf{A}\,\mathbf{x}\|_{(n)} \leq \|\mathbf{A}\|_{(m,n)}\,\|\mathbf{x}\|_{(n)}$ für alle $\mathbf{x} \in \mathbb{R}^n$ und alle $\mathbf{A} \in \mathbb{R}^{m\times n}$.

(3) Ist $m = n$ und $\|\cdot\|_{(m)} = \|\cdot\|_{(n)}$, so gilt für die Einheitsmatrix $\mathbf{E}_n \in \mathbb{R}^{n\times n}$ (mit Einsen auf der Diagonale und sonst Nullen) stets $\|\mathbf{E}_n\|_{(n,n)} = 1$.

(4) Ist $\|\cdot\|_{(\ell)}$ zusätzlich eine Norm auf $\mathbb{R}^\ell$. Dann gilt

$$\|\mathbf{A}\,\mathbf{B}\|_{(\ell,n)} \leq \|\mathbf{A}\|_{(\ell,m)}\,\|\mathbf{B}\|_{(m,n)} \qquad \text{für alle } \mathbf{A} \in \mathbb{R}^{\ell\times m},\ \mathbf{B} \in \mathbb{R}^{m\times n}.$$

(5) Ist $\mathbf{A} \in \mathbb{R}^{n\times n}$ symmetrisch, d.h. $\mathbf{A}^T = \mathbf{A}$, dann gilt für die Spektralnorm

$$\|\mathbf{A}\|_2 = \varrho(\mathbf{A}) = \max\big\{|\lambda| \ :\ \lambda \in \mathbb{C} \text{ ist Eigenwert von } \mathbf{A}\big\}.$$

Den Nachweis von Hilfssatz 2.8 wird am Ende dieses Teilkapitels gezeigt.

Wir bemerken abschließend, dass $\mathbb{R}^{m\times n}$ auch als $\mathbb{R}^\ell$ mit $\ell = m \cdot n$ aufgefasst werden kann, wobei natürlich die Sortierung der Matrixeinträge in $\mathbb{K}^\ell$ angegeben werden muss. Verwendet man nun für $\mathbb{R}^\ell$ mit $\ell = m \cdot n$ die Euklidische Norm, so erhält man die sogenannte **Frobenius-Norm**

$$\|\mathbf{A}\|_F := \left(\sum_{i=1}^{m}\sum_{j=1}^{n} |a_{i,j}|^2\right)^{1/2}, \qquad \mathbf{A} \in \mathbb{R}^{m\times n}. \tag{2.6}$$

Die Frobenius-Norm ist **keine** induzierte Matrixnorm.

Fassen wir $\mathbb{R}^{m\times n}$ als $\mathbb{R}^\ell$ mit $\ell = m \cdot n$ auf, so folgt mit Hilfssatz 2.4, dass die Matrixnormen in Hilfssatz 2.6 und die Frobenius-Norm (2.6) alle äquivalent sind.

Beweis von Hilfssatz 2.8: Die Formel für die induzierte Matrixnorm auf $\mathbb{R}^{m\times n}$ lautet nach Definition 2.5 (1) und (2.3)

$$\|\mathbf{A}\|_{(m,n)} := \max_{\substack{\mathbf{x}\in\mathbb{R}^n,\\ \|\mathbf{x}\|_{(n)}=1}} \|\mathbf{A}\,\mathbf{x}\|_{(m)} = \max_{\mathbf{x}\in\mathbb{R}^n\setminus\{\mathbf{0}\}} \frac{\|\mathbf{A}\,\mathbf{x}\|_{(m)}}{\|\mathbf{x}\|_{(n)}}. \tag{2.7}$$

(1) Wegen $\|\mathbf{A}\,\mathbf{x}\|_{(m)} \geq 0$ für alle $\mathbf{x} \in \mathbb{R}^n$, folgt

$$\|\mathbf{A}\|_{(m,n)} = \max_{\substack{\mathbf{x}\in\mathbb{R}^n,\\ \|\mathbf{x}\|_{(n)}=1}} \underbrace{\|\mathbf{A}\,\mathbf{x}\|_{(m)}}_{\geq 0} \geq 0.$$

Also ist die induzierte Matrixnorm $\|\cdot\|_{(m,n)}$ eine Abbildung von $\mathbb{R}^{m\times n}$ nach $[0;\infty[$. Wir überprüfen nun die drei Normeigenschaften:

(1) Es sei $\|\mathbf{A}\|_{(m,n)} = 0$, also

$$0 = \|\mathbf{A}\|_{(m,n)} = \max_{\mathbf{x}\in\mathbb{R}^n\setminus\{\mathbf{0}\}} \frac{\|\mathbf{A}\,\mathbf{x}\|_{(m)}}{\|\mathbf{x}\|_{(n)}}. \tag{2.8}$$

Wegen $\|\mathbf{A}\,\mathbf{x}\|_{(m)} \geq 0$ und $\|\mathbf{x}\| > 0$ für $\mathbf{x} \in \mathbb{R}^n \setminus \{\mathbf{0}\}$ folgt aus (2.8):

$$\begin{aligned} &\frac{\|\mathbf{A}\,\mathbf{x}\|_{(m)}}{\|\mathbf{x}\|_{(n)}} = 0 \quad \text{für alle } \mathbf{x} \in \mathbb{R}^n \setminus \{\mathbf{0}\} \\ \iff \quad &\|\mathbf{A}\,\mathbf{x}\|_{(m)} = 0 \quad \text{für alle } \mathbf{x} \in \mathbb{R}^n \setminus \{\mathbf{0}\} \\ \iff \quad &\mathbf{A}\,\mathbf{x} = \mathbf{0} \quad \text{für alle } \mathbf{x} \in \mathbb{R}^n \setminus \{\mathbf{0}\}, \end{aligned}$$

wobei der letzte Schritt aus der Eigenschaft (1) der Norm $\|\cdot\|_{(m)}$ folgt. Da auch $\mathbf{A}\,\mathbf{0} = \mathbf{0}$ gilt, folgt

$$\mathbf{A}\,\mathbf{x} = \mathbf{0} \qquad \text{für alle } \mathbf{x} \in \mathbb{R}^n. \tag{2.9}$$

Setzt man in (2.9) die Standardbasisvektoren $\mathbf{e}_1, \mathbf{e}_2, \ldots, \mathbf{e}_n$ von $\mathbb{R}^n$ (in $\mathbf{e}_k$ ist der k-te Eintrag 1 und alle anderen Einträge sind 0) ein, so folgt $\mathbf{A}\,\mathbf{e}_k = \mathbf{0}$ für alle $k = 1, 2, \ldots, n$. Der Vektor $\mathbf{A}\,\mathbf{e}_k$ ist aber der k-te Spaltenvektor von $\mathbf{A}$. Daher sind alle Spaltenvektoren von $\mathbf{A}$ der Nullvektor $\mathbf{0} \in \mathbb{R}^m$, d.h. $\mathbf{A}$ ist die Nullmatrix $\mathbf{O}_{m\times n}$ von $\mathbb{R}^{m\times n}$.

(2) Seien $\mathbf{A} \in \mathbb{R}^{m\times n}$ und $\lambda \in \mathbb{R}$ beliebig. Dann gilt wegen der Eigenschaft (2) der Norm $\|\cdot\|_{(m)}$

$$\|(\lambda\,\mathbf{A})\,\mathbf{x}\|_{(m)} = \|\lambda\,\mathbf{A}\,\mathbf{x}\|_{(m)} = |\lambda|\,\|\mathbf{A}\,\mathbf{x}\|_{(m)} \qquad \text{für alle } \mathbf{x} \in \mathbb{R}^n.$$

Damit folgt

$$\begin{aligned} \|\lambda\,\mathbf{A}\|_{(m,n)} &= \max_{\substack{\mathbf{x}\in\mathbb{R}^n,\\ \|\mathbf{x}\|_{(n)}=1}} \|(\lambda\,\mathbf{A})\,\mathbf{x}\|_{(m)} = \max_{\substack{\mathbf{x}\in\mathbb{R}^n,\\ \|\mathbf{x}\|_{(n)}=1}} \big(|\lambda|\,\|\mathbf{A}\,\mathbf{x}\|_{(m)}\big) \\ &= |\lambda| \max_{\substack{\mathbf{x}\in\mathbb{R}^n,\\ \|\mathbf{x}\|_{(n)}=1}} \|\mathbf{A}\,\mathbf{x}\|_{(m)} = |\lambda|\,\|\mathbf{A}\|_{(m,n)}. \end{aligned}$$

(3) Seien $\mathbf{A}, \mathbf{B} \in \mathbb{R}^{m\times n}$ beliebig. Dann gilt nach der Dreiecksungleichung für $\|\cdot\|_{(m)}$

$$\begin{aligned} \|(\mathbf{A}+\mathbf{B})\,\mathbf{x}\|_{(m)} &= \|\mathbf{A}\,\mathbf{x} + \mathbf{B}\,\mathbf{x}\|_{(m)} \\ &\leq \|\mathbf{A}\,\mathbf{x}\|_{(m)} + \|\mathbf{B}\,\mathbf{x}\|_{(m)} \qquad \text{für alle } \mathbf{x} \in \mathbb{R}^n. \end{aligned}$$

Anwenden dieser Abschätzung in der Formel für die induzierte Matrixnorm liefert

$$\begin{aligned} \|\mathbf{A}+\mathbf{B}\|_{(m,n)} &= \max_{\substack{\mathbf{x}\in\mathbb{R}^n,\\ \|\mathbf{x}\|_{(n)}=1}} \|(\mathbf{A}+\mathbf{B})\,\mathbf{x}\|_{(m)} \\ &\leq \max_{\substack{\mathbf{x}\in\mathbb{R}^n,\\ \|\mathbf{x}\|_{(n)}=1}} \big(\|\mathbf{A}\,\mathbf{x}\|_{(m)} + \|\mathbf{B}\,\mathbf{x}\|_{(m)}\big) \\ &\leq \max_{\substack{\mathbf{x}\in\mathbb{R}^n,\\ \|\mathbf{x}\|_{(n)}=1}} \|\mathbf{A}\,\mathbf{x}\|_{(m)} + \max_{\substack{\mathbf{x}\in\mathbb{R}^n,\\ \|\mathbf{x}\|_{(n)}=1}} \|\mathbf{B}\,\mathbf{x}\|_{(m)} = \|\mathbf{A}\|_{(m,n)} + \|\mathbf{B}\|_{(m,n)}. \end{aligned}$$

Da alle Normeigenschaften erfüllt sind, ist $\|\cdot\|_{(m,n)}$ eine Norm auf $\mathbb{R}^{m\times n}$.

(2) Aus der zweiten Darstellung von $\|\mathbf{A}\|_{(m,n)}$ in (2.7) folgt für alle $\mathbf{x}\in\mathbb{R}^n\backslash\{\mathbf{0}\}$:

$$\frac{\|\mathbf{A}\,\mathbf{x}\|_{(m)}}{\|\mathbf{x}\|_{(n)}} \leq \max_{\mathbf{x}\in\mathbb{R}^n\backslash\{\mathbf{0}\}} \frac{\|\mathbf{A}\,\mathbf{x}\|_{(m)}}{\|\mathbf{x}\|_{(n)}} = \|\mathbf{A}\|_{(m,n)} \quad\Longrightarrow\quad \frac{\|\mathbf{A}\,\mathbf{x}\|_{(m)}}{\|\mathbf{x}\|_{(n)}} \leq \|\mathbf{A}\|_{(m,n)}$$

Multiplizieren mit $\|\mathbf{x}\|_{(n)}$ liefert

$$\|\mathbf{A}\,\mathbf{x}\|_{(m)} \leq \|\mathbf{A}\|_{(m,n)}\,\|\mathbf{x}\|_{(n)} \qquad \text{für alle } \mathbf{x}\in\mathbb{R}^n\setminus\{\mathbf{0}\}.$$

Diese Ungleichung ist auch wahr, wenn $\mathbf{x}=\mathbf{0}$ ist, denn dann sind beide Seiten null.

(3) Für die Einheitsmatrix $\mathbf{E}_n$ gilt $\mathbf{E}_n\,\mathbf{x}=\mathbf{x}$ für alle $\mathbf{x}\in\mathbb{R}^n$. Daraus folgt

$$\|\mathbf{E}_n\|_{(n,n)} = \max_{\substack{\mathbf{x}\in\mathbb{R}^n,\\ \|\mathbf{x}\|_{(n)}=1}} \|\mathbf{E}_n\,\mathbf{x}\|_{(n)} = \max_{\substack{\mathbf{x}\in\mathbb{R}^n,\\ \|\mathbf{x}\|_{(n)}=1}} \underbrace{\|\mathbf{x}\|_{(n)}}_{=1} = 1.$$

(4) Mit der ersten Darstellung von $\|\mathbf{A}\,\mathbf{B}\|_{(\ell,n)}$ in (2.7) gilt

$$\|\mathbf{A}\,\mathbf{B}\|_{(\ell,n)} = \max_{\substack{\mathbf{x}\in\mathbb{R}^n,\\ \|\mathbf{x}\|_{(n)}=1}} \|\mathbf{A}\,\mathbf{B}\,\mathbf{x}\|_{(\ell)}. \tag{2.10}$$

Durch zweimalige Anwendung von Hilfssatz 2.8 (2), den wir schon bewiesen haben, erhalten wir

$$\|\mathbf{A}\,\mathbf{B}\,\mathbf{x}\|_{(\ell)} = \big\|\mathbf{A}\,(\mathbf{B}\,\mathbf{x})\big\|_{(\ell)} \leq \|\mathbf{A}\|_{(\ell,m)}\,\|\mathbf{B}\,\mathbf{x}\|_{(m)} \leq \|\mathbf{A}\|_{(\ell,m)}\,\|\mathbf{B}\|_{(m,n)}\,\|\mathbf{x}\|_{(n)}.$$

Anwenden dieser Abschätzung in (2.10) liefert

$$\begin{aligned}\|\mathbf{A}\,\mathbf{B}\|_{(\ell,n)} &= \max_{\substack{\mathbf{x}\in\mathbb{R}^n,\\ \|\mathbf{x}\|_{(n)}=1}} \|\mathbf{A}\,\mathbf{B}\,\mathbf{x}\|_{(\ell)} \leq \max_{\substack{\mathbf{x}\in\mathbb{R}^n,\\ \|\mathbf{x}\|_{(n)}=1}} \big(\|\mathbf{A}\|_{(\ell,m)}\,\|\mathbf{B}\|_{(m,n)}\,\|\mathbf{x}\|_{(n)}\big)\\ &= \|\mathbf{A}\|_{(\ell,m)}\,\|\mathbf{B}\|_{(m,n)} \underbrace{\max_{\substack{\mathbf{x}\in\mathbb{R}^n,\\ \|\mathbf{x}\|_{(n)}=1}} \|\mathbf{x}\|_{(n)}}_{=1} = \|\mathbf{A}\|_{(\ell,m)}\,\|\mathbf{B}\|_{(m,n)}.\end{aligned}$$

(5) Die Formel für die Spektralnorm ist

$$\|\mathbf{A}\|_2 = \sqrt{\varrho(\mathbf{A}^T\,\mathbf{A})}, \qquad \text{wobei}$$
$$\varrho(\mathbf{A}^T\,\mathbf{A}) := \max\big\{|\lambda| \;:\; \lambda\in\mathbb{C} \text{ ist Eigenwert von } \mathbf{A}^T\,\mathbf{A}\big\}.$$

Sei $\mathbf{A}$ nun symmetrisch, d.h. $\mathbf{A}^T=\mathbf{A}$. Dann folgt $\mathbf{A}^T\,\mathbf{A}=\mathbf{A}\,\mathbf{A}$.

Sei nun λ ein Eigenwert von $\mathbf{A}$, und sei $\mathbf{x}$ ein zugehöriger Eigenvektor. Dann gilt

$$(\mathbf{A}^T\,\mathbf{A})\,\mathbf{x} = (\mathbf{A}\,\mathbf{A})\,\mathbf{x} = \mathbf{A}\,\big(\mathbf{A}\,\mathbf{x}\big) = \mathbf{A}\,(\lambda\,\mathbf{x}) = \lambda\,(\mathbf{A}\,\mathbf{x}) = \lambda\,(\lambda\,\mathbf{x}) = \lambda^2\,\mathbf{x}.$$

(Man kann auch zeigen, dass aus $\mathbf{A} = \mathbf{A}^T$ folgt, dass jeder Eigenwert von $\mathbf{A}$ reell ist, aber dieses benötigen wir hier nicht.) Also sind die Eigenwerte von $\mathbf{A}^T\,\mathbf{A} = \mathbf{A}\,\mathbf{A}$ die Quadrate der Eigenwerte von $\mathbf{A}$:

$$\varrho(\mathbf{A}^T\,\mathbf{A}) = \max\big\{\underbrace{|\lambda^2|}_{=|\lambda|^2} \;:\; \lambda \in \mathbb{C} \text{ ist Eigenwert von } \mathbf{A}\big\}.$$

Damit folgt

$$\begin{aligned}\|\mathbf{A}\|_2 &= \sqrt{\varrho(\mathbf{A}^T\,\mathbf{A})} = \sqrt{\max\big\{|\lambda|^2 \;:\; \lambda \in \mathbb{C} \text{ ist Eigenwert von } \mathbf{A}\big\}} \\ &= \max\big\{|\lambda| \;:\; \lambda \in \mathbb{C} \text{ ist Eigenwert von } \mathbf{A}\big\}.\end{aligned}$$

Damit sind alle fünf Ausagen von Hilfssatz 2.8 nachgewiesen. □

2.2 Störungen und Kondition einer Matrix

In diesem Teilkapitel sei $\mathbf{A} \in \mathbb{R}^{n\times n}$ eine **quadratische invertierbare** Matrix. Die inverse Matrix zu $\mathbf{A}$ werde wie üblich mit $\mathbf{A}^{-1}$ bezeichnet.

Wir erinnern kurz daran, dass die folgenden Eigenschaften einer **quadratischen Matrix $\mathbf{A} \in \mathbb{R}^{n\times n}$ äquivalent** sind:

(i) $\mathbf{A}$ ist invertierbar (d.h. es gibt eine Matrix $\mathbf{A}^{-1} \in \mathbb{R}^{n\times n}$, genannt die Inverse von $\mathbf{A}$, mit der Eigenschaft $\mathbf{A}\,\mathbf{A}^{-1} = \mathbf{A}^{-1}\,\mathbf{A} = \mathbf{E}_n$).

(ii) $\mathbf{A}$ hat den Rang n (d.h. $\mathbf{A}$ hat genau n linear unabhängige Zeilenvektoren und genau n linear unabhängige Spaltenvektoren).

(iii) $\det(\mathbf{A}) \neq 0$

(iv) Das LGS $\mathbf{A}\,\mathbf{x} = \mathbf{b}$ ist für jedes $\mathbf{b} \in \mathbb{R}^n$ eindeutig lösbar.

(v) Das LGS $\mathbf{A}\,\mathbf{x} = \mathbf{0}$ hat nur die Lösung $\mathbf{x} = \mathbf{0}$.

(vi) $\mathbf{A}$ ist regulär.

Ist eine quadratische Matrix $\mathbf{A} \in \mathbb{R}^{n\times n}$ **nicht regulär** (also nicht invertierbar), dann nennen wir die Matrix **singulär**.

Wir wollen nun untersuchen, wie sich eine Störung der rechten Seite des linearen Gleichungssystems (LGS) $\mathbf{A}\,\mathbf{x} = \mathbf{b}$ auf die Lösung auswirkt.

Satz 2.9. (Störung der rechten Seite eines LGS und Konditionszahl)
*Sei $\|\cdot\|$ eine Norm auf $\mathbb{R}^n$, und sei $\|\cdot\|$ die induzierte Matrixnorm auf $\mathbb{R}^{n\times n}$. Sei $\mathbf{A} \in \mathbb{R}^{n\times n}$ **invertierbar**. Es seien eine rechte Seite $\mathbf{b} \in \mathbb{R}^n \setminus \{\mathbf{0}\}$ und eine **gestörte rechte Seite** $\widetilde{\mathbf{b}} = \mathbf{b} + \Delta\mathbf{b} \in \mathbb{R}^n$ ($\Delta\mathbf{b}$ ist also die Störung von $\mathbf{b}$) gegeben, und die zugehörigen Lösungen seien $\mathbf{x} \in \mathbb{R}^n$ bzw. $\widetilde{\mathbf{x}} \in \mathbb{R}^n$, d.h.*

$$\mathbf{A}\,\mathbf{x} = \mathbf{b} \qquad \text{und} \qquad \mathbf{A}\,\widetilde{\mathbf{x}} = \widetilde{\mathbf{b}}.$$

*Dann erfüllt der **relative Fehler** der (gestörten) Lösung $\widetilde{\mathbf{x}}$ die Abschätzung*

$$\mathrm{Rel}(\widetilde{\mathbf{x}}) = \frac{\|\widetilde{\mathbf{x}} - \mathbf{x}\|}{\|\mathbf{x}\|} \leq \|\mathbf{A}\|\,\|\mathbf{A}^{-1}\|\,\frac{\|\widetilde{\mathbf{b}} - \mathbf{b}\|}{\|\mathbf{b}\|} = \|\mathbf{A}\|\,\|\mathbf{A}^{-1}\|\,\mathrm{Rel}(\widetilde{\mathbf{b}})\,. \qquad (2.11)$$

Da die Zahl $\|\mathbf{A}\|\,\|\mathbf{A}^{-1}\|$ in (2.11) eine Abschätzung nach oben dafür angibt, wie stark der relative Fehler der rechten Seite $\widetilde{\mathbf{b}}$ höchstens verstärkt wird, heißt

$$\mathrm{cond}_{\|\cdot\|}(\mathbf{A}) := \|\mathbf{A}\|\,\|\mathbf{A}^{-1}\|$$

*die **Konditionszahl** der Matrix $\mathbf{A}$ (bzgl. der induzierten Matrixnorm $\|\cdot\|$).*

Bei den p-Normen $\|\cdot\|_p$ für $\mathbb{R}^n$ schreiben wir auch $\mathrm{cond}_p(\mathbf{A})$ statt $\mathrm{cond}_{\|\cdot\|_p}(\mathbf{A})$, also $\mathrm{cond}_p(\mathbf{A}) = \|\mathbf{A}\|_p\,\|\mathbf{A}^{-1}\|_p$.

In Verallgemeinerung von Satz 2.9 kann man auch den Fall betrachten, dass die Matrix und die rechte Seite gestört sind, also das lineare Gleichungssystem

$$(\mathbf{A} + \Delta\mathbf{A})\,(\mathbf{x} + \Delta\mathbf{x}) = \mathbf{b} + \Delta\mathbf{b}.$$

Dieses soll hier aber nicht weiter untersucht werden.

Betrachten wir ein Beispiel.

Beispiel 2.10. (Konditionszahl einer Matrix)
Wir betrachten die Matrix

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

welche für jedes $\alpha \neq 0$ regulär ist. Für $\alpha = 0$ ist die Matrix allerdings singulär, d.h. nicht invertierbar (denn nur für $\alpha = 0$ sind die beiden Zeilenvektoren linear abhängig). Für $\alpha \neq 0$ können wir die Konditionszahl bzgl. der Spaltensummen-

norm und bzgl. der Zeilensummennorm bequem berechnen. Es gilt (mit der aus der linearen Algebra bekannten Formel für die Inverse einer 2×2-Matrix)

$$\mathbf{A}^{-1} = \frac{1}{\det(\mathbf{A})} \begin{bmatrix} 1+\alpha & -1 \\ -1 & 1 \end{bmatrix} = \frac{1}{\alpha} \begin{bmatrix} 1+\alpha & -1 \\ -1 & 1 \end{bmatrix} = \begin{bmatrix} 1+\frac{1}{\alpha} & -\frac{1}{\alpha} \\ -\frac{1}{\alpha} & \frac{1}{\alpha} \end{bmatrix},$$

wobei wir $\det(\mathbf{A}) = 1\,(1+\alpha) - 1 = \alpha$ genutzt haben. Damit finden wir

$$\mathrm{cond}_1(\mathbf{A}) = \|\mathbf{A}\|_1 \, \|\mathbf{A}^{-1}\|_1 = \max\left\{2, 1+|1+\alpha|\right\} \max\left\{\left|1+\frac{1}{\alpha}\right| + \frac{1}{|\alpha|}, \frac{2}{|\alpha|}\right\},$$

$$\mathrm{cond}_\infty(\mathbf{A}) = \|\mathbf{A}\|_\infty \, \|\mathbf{A}^{-1}\|_\infty = \max\left\{2, 1+|1+\alpha|\right\} \max\left\{\left|1+\frac{1}{\alpha}\right| + \frac{1}{|\alpha|}, \frac{2}{|\alpha|}\right\}.$$

Dass wir für die Spaltensummennorm und für die Zeilensummennorm die gleiche Konditionszahl erhalten liegt daran, dass $\mathbf{A}$ und damit auch ihre Inverse $\mathbf{A}^{-1}$ symmetrisch sind (also $\mathbf{A}^T = \mathbf{A}$ bzw. $(\mathbf{A}^{-1})^T = \mathbf{A}^{-1}$ erfüllen).

Die Matrix $\mathbf{A}$ wird singulär wenn $\alpha = 0$ ist. Betrachten wir daher, was passiert, wenn $|\alpha|$ sehr dicht bei 0 liegt d.h. wenn $\mathbf{A}$ „fast singulär“ ist. Dann gilt genähert

$$\mathrm{cond}_1(\mathbf{A}) = \mathrm{cond}_\infty(\mathbf{A}) \approx 2 \cdot \frac{2}{|\alpha|} = \frac{4}{|\alpha|}, \tag{2.12}$$

denn $|1+\alpha| \approx 1$ und $\left|1+\frac{1}{\alpha}\right| \approx \frac{1}{|\alpha|}$ für $|\alpha|$ sehr dicht bei 0. Wir sehen in (2.12), dass die Konditionszahl für $\alpha \to 0$ gegen ∞ strebt. Relative Fehler der rechten Seite können in einem linearen Gleichungssystem mit einer „fast singulären“ Matrix also beliebig verstärkt werden. ♠

Wir halten einige Eigenschaften der Konditionszahl fest. Den Nachweis dieser Eigenschaften geben wir am Ende dieses Teilkapitels.

Hilfssatz 2.11. (Eigenschaften der Konditionszahl)

Seien die Voraussetzungen und die Notation wie in Satz 2.9. Dann gelten:

(1) Es gilt stets $\mathrm{cond}_{\|\cdot\|}(\mathbf{A}) \geq 1$.

(2) Ist auf $\mathbb{R}^n$ *die* 2***-Norm*** *(oder* ***Euklidische Norm****)* $\|\cdot\|_2$ *gegeben und ist* $\mathbf{A} \in \mathbb{R}^{n\times n}$ ***symmetrisch*** *(d.h.* $\mathbf{A} = \mathbf{A}^T$*), so gilt*

$$\mathrm{cond}_2(\mathbf{A}) = \|\mathbf{A}\|_2 \, \|\mathbf{A}^{-1}\|_2 = \frac{\max\limits_{j=1,\ldots,n} |\lambda_j|}{\min\limits_{k=1,\ldots,n} |\lambda_k|},$$

wobei $\lambda_1, \lambda_2, \ldots, \lambda_n$ *die* n *(nicht notwendigerweise verschiedenen) reellen Eigenwerte von* $\mathbf{A}$ *sind.*

Betrachten wir zunächst ein Beispiel.

Beispiel 2.12. (Konditionszahl einer Matrix)
Wir betrachten die Matrix

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

welche für jedes $\alpha \neq 0$ regulär (d.h. invertierbar) ist. Für $\alpha = 0$ ist die Matrix allerdings singulär, d.h. nicht invertierbar. Wir wollen nun die Konditionszahl bzgl. der Spektralnorm mit Hilfe von Hilfssatz 2.11 (2) berechnen. Wir faktorisieren das charakteristische Polynom von $\mathbf{A}$, um die Eigenwerte von $\mathbf{A}$ zu bestimmen. Mit quadratischer Ergänzung, der zweiten binomischen Formel und dann der dritten binomischen Formel finden wir

$$\begin{aligned} p_{\mathbf{A}}(\lambda) &= \det\left(\mathbf{A} - \lambda\,\mathbf{E}_2\right) = \det\left(\begin{bmatrix} 1-\lambda & 1 \\ 1 & 1+\alpha-\lambda \end{bmatrix}\right) = (1-\lambda)\,(1+\alpha-\lambda) - 1 \\ &= \lambda^2 - (2+\alpha)\,\lambda + \alpha = \left(\lambda^2 - 2\left(1+\frac{\alpha}{2}\right)\lambda + \left(1+\frac{\alpha}{2}\right)^2\right) - \left(1+\frac{\alpha}{2}\right)^2 + \alpha \\ &= \left(\lambda - 1 - \frac{\alpha}{2}\right)^2 - \left(1 + \alpha + \frac{\alpha^2}{4} - \alpha\right) = \left(\lambda - 1 - \frac{\alpha}{2}\right)^2 - \left(1 + \frac{\alpha^2}{4}\right) \\ &= \left(\lambda - 1 - \frac{\alpha}{2} - \sqrt{1+\frac{\alpha^2}{4}}\right)\left(\lambda - 1 - \frac{\alpha}{2} + \sqrt{1+\frac{\alpha^2}{4}}\right). \end{aligned}$$

Also sind die Eigenwerte von $\mathbf{A}$ durch $\lambda = 1 + \frac{\alpha}{2} \pm \sqrt{1+\frac{\alpha^2}{4}}$ gegeben. Der Einfachheit halber betrachten wir nun nur noch den Fall $\alpha > 0$. Für $\alpha > 0$ gilt

$$0 \leq 1 + \frac{\alpha}{2} - \sqrt{1+\frac{\alpha^2}{4}} \leq 1 + \frac{\alpha}{2} + \sqrt{1+\frac{\alpha^2}{4}}\,.$$

Damit finden wir für $\alpha > 0$ mit Hilfssatz 2.11 (2) die Konditionszahl bzgl. von der 2-Norm induzierten Matrixnorm

$$\mathrm{cond}_2(\mathbf{A}) = \frac{1 + \frac{\alpha}{2} + \sqrt{1+\frac{\alpha^2}{4}}}{1 + \frac{\alpha}{2} - \sqrt{1+\frac{\alpha^2}{4}}}\,.$$

Für $\alpha > 0$ dicht bei 0, also wenn die Matrix „fast singulär“ ist, finden wir mit Umformen mit der dritten binomischen Formel

$$\mathrm{cond}_2(\mathbf{A}) = \frac{1 + \frac{\alpha}{2} + \sqrt{1+\frac{\alpha^2}{4}}}{1 + \frac{\alpha}{2} - \sqrt{1+\frac{\alpha^2}{4}}} = \frac{\left(1 + \frac{\alpha}{2} + \sqrt{1+\frac{\alpha^2}{4}}\right)^2}{\left(1 + \frac{\alpha}{2} - \sqrt{1+\frac{\alpha^2}{4}}\right)\left(1 + \frac{\alpha}{2} + \sqrt{1+\frac{\alpha^2}{4}}\right)}$$

$$= \frac{\left(1+\frac{\alpha}{2}+\sqrt{1+\frac{\alpha^2}{4}}\right)^2}{\left(1+\frac{\alpha}{2}\right)^2-\left(\sqrt{1+\frac{\alpha^2}{4}}\right)^2} = \frac{\left(1+\frac{\alpha}{2}+\sqrt{1+\frac{\alpha^2}{4}}\right)^2}{\left(1+2\,\frac{\alpha}{2}+\frac{\alpha^2}{4}\right)-\left(1+\frac{\alpha^2}{4}\right)}$$

$$= \frac{\left(1+\frac{\alpha}{2}+\sqrt{1+\frac{\alpha^2}{4}}\right)^2}{\alpha} \approx \frac{4}{\alpha} \quad \text{für } \alpha > 0 \text{ dicht bei } 0.$$

Wir beobachten, dass auch die Konditionszahl bzgl. der Spektralnorm beliebig groß wird, wenn $\alpha > 0$ gegen 0 strebt und die Matrix dabei immer stärker „fast singulär“ wird. ♠

Nun zeigen wir zum Abschluss die Beweise von Satz 2.9 und Hilfssatz 2.11.

Beweis von Satz 2.9: Wegen $\mathbf{A}\,\mathbf{x} = \mathbf{b} \iff \mathbf{x} = \mathbf{A}^{-1}\,\mathbf{b}$ und $\mathbf{A}\,\widetilde{\mathbf{x}} = \widetilde{\mathbf{b}} \iff \widetilde{\mathbf{x}} = \mathbf{A}^{-1}\,\widetilde{\mathbf{b}}$ (jeweils durch Multiplizieren von links mit $\mathbf{A}^{-1}$) folgt

$$\widetilde{\mathbf{x}} - \mathbf{x} = \mathbf{A}^{-1}\,\widetilde{\mathbf{b}} - \mathbf{A}^{-1}\,\mathbf{b} = \mathbf{A}^{-1}\,(\widetilde{\mathbf{b}} - \mathbf{b}),$$

und somit gilt (mit Hilfssatz 2.8 (2))

$$\|\widetilde{\mathbf{x}} - \mathbf{x}\| = \|\mathbf{A}^{-1}\,(\widetilde{\mathbf{b}} - \mathbf{b})\| \leq \|\mathbf{A}^{-1}\|\,\|\widetilde{\mathbf{b}} - \mathbf{b}\|. \tag{2.13}$$

Weiter gilt (mit Hilfssatz 2.8 (2))

$$\|\mathbf{b}\| = \|\mathbf{A}\,\mathbf{x}\| \leq \|\mathbf{A}\|\,\|\mathbf{x}\| \qquad \Longleftrightarrow \qquad \|\mathbf{x}\| \geq \|\mathbf{A}\|^{-1}\,\|\mathbf{b}\|. \tag{2.14}$$

Damit folgt für den relativen Fehler der gestörten Lösung

$$\frac{\|\widetilde{\mathbf{x}} - \mathbf{x}\|}{\|\mathbf{x}\|} \leq \frac{\|\mathbf{A}^{-1}\|\,\|\widetilde{\mathbf{b}} - \mathbf{b}\|}{\|\mathbf{x}\|} \leq \frac{\|\mathbf{A}^{-1}\|\,\|\widetilde{\mathbf{b}} - \mathbf{b}\|}{\|\mathbf{A}\|^{-1}\,\|\mathbf{b}\|} = \|\mathbf{A}\|\,\|\mathbf{A}^{-1}\|\,\frac{\|\widetilde{\mathbf{b}} - \mathbf{b}\|}{\|\mathbf{b}\|},$$

wobei wir im ersten Schritt (2.13) durch $\|\mathbf{x}\|$ geteilt haben und danach im Nenner die Abschätzung in (2.14) genutzt haben. □

Beweis von Hilfssatz 2.11:

(1) Sei $\mathbf{A} \in \mathbb{R}^{n\times n}$ invertierbar, aber beliebig. Per Definition gilt

$$\mathrm{cond}_{\|\cdot\|}(\mathbf{A}) = \|\mathbf{A}\|\,\|\mathbf{A}^{-1}\|.$$

Nach Hilfssatz 2.8 (4) gilt $\|\mathbf{A}\,\mathbf{A}^{-1}\| \leq \|\mathbf{A}\|\,\|\mathbf{A}^{-1}\|$. Also folgt

$$\mathrm{cond}_{\|\cdot\|}(\mathbf{A}) = \|\mathbf{A}\|\,\|\mathbf{A}^{-1}\| \geq \|\mathbf{A}\,\mathbf{A}^{-1}\| = \|\mathbf{E}_n\| = 1,$$

wobei wir im letzten Schritt Hilfssatz 2.8 (3) genutzt haben.

(2) Da $\mathbf{A}$ symmetrisch ist, gilt nach Hilfssatz 2.8 (5)

$$\|\mathbf{A}\|_2 = \max\big\{|\lambda_1|; |\lambda_2|; \ldots; |\lambda_n|\big\}, \tag{2.15}$$

wobei $\lambda_1, \lambda_2, \ldots, \lambda_n$ die n nicht notwendigerweise verschiedenen Eigenwerte von $\mathbf{A}$ sind. Da $\mathbf{A}$ invertierbar ist, kann 0 kein Eigenwert sein. (Ansonsten gäbe es ein $\mathbf{x} \neq \mathbf{0}$ mit $\mathbf{A}\,\mathbf{x} = 0\,\mathbf{x} = \mathbf{0}$, aber dieses ist ein Widerspruch zur Invertierbarkeit von $\mathbf{A}$.)

Die inverse Matrix $\mathbf{A}^{-1}$ von $\mathbf{A}$ ist auch symmetrisch, denn $\big(\mathbf{A}^{-1}\big)^T = \big(\mathbf{A}^T\big)^{-1} = \mathbf{A}^{-1}$. Da $\mathbf{A}^{-1}$ auch symmetrisch ist, können wir auch für $\mathbf{A}^{-1}$ Hilfssatz 2.8 (5) zur Berechnung von $\|\mathbf{A}^{-1}\|_2$ nutzen.

Wie sehen die Eigenwerte von $\mathbf{A}^{-1}$ aus? Sei $\lambda \neq 0$ ein Eigenwert von $\mathbf{A}$ und $\mathbf{x} \neq \mathbf{0}$ ein zugehöriger Eigenvektor. Dann gilt,

$$\begin{aligned}\mathbf{A}\,\mathbf{x} = \lambda\,\mathbf{x} \quad &\Longleftrightarrow \quad \underbrace{\mathbf{A}^{-1}\,\mathbf{A}}_{=\mathbf{E}_n}\,\mathbf{x} = \mathbf{A}^{-1}\,\lambda\,\mathbf{x} \quad \Longleftrightarrow \quad \underbrace{\mathbf{E}_n\,\mathbf{x}}_{=\mathbf{x}} = \lambda\,\mathbf{A}^{-1}\,\mathbf{x} \\ &\Longleftrightarrow \quad \lambda^{-1}\,\mathbf{x} = \mathbf{A}^{-1}\,\mathbf{x} \quad \Longleftrightarrow \quad \mathbf{A}^{-1}\,\mathbf{x} = \lambda^{-1}\,\mathbf{x},\end{aligned}$$

d.h. $\mathbf{x}$ ist ein Eigenvektor von $\mathbf{A}^{-1}$ zum Eigenwert λ^{-1}. Also hat $\mathbf{A}^{-1}$ die Eigenwerte $\lambda_1^{-1}, \lambda_1^{-1}, \ldots, \lambda_n^{-1}$, wobei $\lambda_1, \lambda_2, \ldots, \lambda_n$ die n nicht notwendigerweise verschiedenen Eigenwerte von $\mathbf{A}$ sind.

Es folgt also

$$\begin{aligned}\|\mathbf{A}^{-1}\|_2 &= \max\big\{|\lambda_1^{-1}|; |\lambda_2^{-1}|; \ldots; |\lambda_n^{-1}|\big\} \\ &= \max\big\{|\lambda_1|^{-1}; |\lambda_2|^{-1}; \ldots; |\lambda_n|^{-1}\big\} = \frac{1}{\min\big\{|\lambda_1|; |\lambda_2|; \ldots; |\lambda_n|\big\}}\,. \end{aligned} \tag{2.16}$$

Damit folgt aus (2.15) und (2.16)

$$\mathrm{cond}_2(\mathbf{A}) = \|\mathbf{A}\|_2\,\|\mathbf{A}^{-1}\|_2 = \frac{\max\big\{|\lambda_1|; |\lambda_2|; \ldots; |\lambda_n|\big\}}{\min\big\{|\lambda_1|; |\lambda_2|; \ldots; |\lambda_n|\big\}}\,.$$

Somit sind beide Eigenschaften aus Hilfssatz 2.8 bewiesen. □

2.3 Gaußsches Eliminationsverfahren mit Pivot-Strategie und LR-Zerlegung

Wir wollen im Folgenden lineare Gleichungssysteme (LGS) $\mathbf{A}\,\mathbf{x} = \mathbf{b}$ mit der Matrix $\mathbf{A} \in \mathbb{R}^{m\times n}$ und der rechten Seite $\mathbf{b} \in \mathbb{R}^m$ mit dem Gaußschen Eliminationsverfahren lösen. Zum Einstieg wiederholen wir einige Grundlagen.

Das **lineare Gleichungssystem** $\mathbf{A}\,\mathbf{x} = \mathbf{b}$ mit der Matrix $\mathbf{A} = [a_{i,j}] \in \mathbb{R}^{m\times n}$ und der rechten Seite $\mathbf{b} \in \mathbb{R}^n$, also

$$\mathbf{A}\,\mathbf{x} = \mathbf{b} \qquad \Longleftrightarrow \qquad \begin{bmatrix} a_{1,1} & a_{1,2} & \cdots & a_{1,n} \\ a_{2,1} & a_{2,2} & \cdots & a_{2,n} \\ \vdots & \vdots & & \vdots \\ a_{m,1} & a_{m,2} & \cdots & a_{m,n} \end{bmatrix} \cdot \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix} = \begin{bmatrix} b_1 \\ b_2 \\ \vdots \\ b_m \end{bmatrix}, \tag{2.17}$$

hat m **lineare Gleichungen** und n **Unbekannte** $x_1, x_2, \ldots, x_n$:

$$\mathbf{A}\,\mathbf{x} = \mathbf{b} \qquad \Longleftrightarrow \qquad \begin{cases} a_{1,1}\,x_1 + a_{1,2}\,x_2 + \ldots + a_{1,n}\,x_n = b_1 \\ a_{2,1}\,x_1 + a_{2,2}\,x_2 + \ldots + a_{2,n}\,x_n = b_2 \\ \qquad\qquad\qquad\qquad\qquad\qquad\quad \vdots \\ a_{m,1}\,x_1 + a_{m,2}\,x_2 + \ldots + a_{m,n}\,x_n = b_m \end{cases}$$

Es ist üblich, das lineare Gleichungssystem (2.17) kompakter mit der **erweiterten Koeffizientenmatrix** zu schreiben:

$$\big[\,\mathbf{A}\big|\mathbf{b}\,\big] := \left[\begin{array}{cccc|c} a_{1,1} & a_{1,2} & \cdots & a_{1,n} & b_1 \\ a_{2,1} & a_{2,2} & \cdots & a_{2,n} & b_2 \\ \vdots & \vdots & & \vdots & \vdots \\ a_{m,1} & a_{m,2} & \cdots & a_{m,n} & b_m \end{array}\right]$$

Bezeichnen $\mathbf{a}_1, \mathbf{a}_2, \ldots, \mathbf{a}_n \in \mathbb{R}^m$ die Spaltenvektoren von $\mathbf{A}$, so können wir das lineare Gleichungssystem (2.17) auch als

$$x_1\,\mathbf{a}_1 + x_2\,\mathbf{a}_2 + \ldots + x_n\,\mathbf{a}_n = \mathbf{b} \qquad \Longleftrightarrow \qquad \sum_{j=1}^{n} x_j\,\mathbf{a}_j = \mathbf{b} \tag{2.18}$$

schreiben. An (2.18) sehen wir, dass das lineare Gleichungssystem $\mathbf{A}\,\mathbf{x} = \mathbf{b}$ **genau dann lösbar ist, wenn die rechte Seite eine Linearkombination der Spaltenvektoren von A ist**. (Falls Sie den Begriff des Rangs einer Matrix kennen, so kann man die Lösbarkeit von $\mathbf{A}\,\mathbf{x} = \mathbf{b}$ auch wie folgt charakterisieren: $\mathbf{A}\,\mathbf{x} = \mathbf{b}$ ist genau dann lösbar, wenn die Matrix $\mathbf{A}$ und die erweiterte Koeffizientenmatrix $[\mathbf{A}|\mathbf{b}]$ den gleichen Rang haben.)

Beispiel 2.13. (lineares Gleichungssystem)
Das lineare Gleichungssystem

$$\begin{aligned} x_1 + x_2 - 3\,x_3 + x_4 &= 1 \\ 2\,x_1 + x_2 + x_3 - x_4 &= 0 \\ 2\,x_2 - 13\,x_3 + x_4 &= -1 \end{aligned}$$

lautet in Matrix-Vektor Schreibweise

$$\begin{bmatrix} 1 & 1 & -3 & 1 \\ 2 & 1 & 1 & -1 \\ 0 & 2 & -13 & 1 \end{bmatrix} \cdot \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{bmatrix} = \begin{bmatrix} 1 \\ 0 \\ -1 \end{bmatrix}$$

und hat die erweiterte Koeffizientenmatrix $\left[\begin{array}{cccc|c} 1 & 1 & -3 & 1 & 1 \\ 2 & 1 & 1 & -1 & 0 \\ 0 & 2 & -13 & 1 & -1 \end{array}\right]$. ♠

Wir erinnern uns nun, welche Fälle bei der Lösungsmenge eines linearen Gleichungssystems austreten können.

Hilfssatz 2.14. (Lösbarkeit eines linearen Gleichungssystems)
Seien $\mathbf{A} \in \mathbb{R}^{m \times n}$ *und* $\mathbf{b} \in \mathbb{R}^m$. *Für die Lösungsmenge* $\mathbb{L}$ *des linearen Gleichungssystems* $\mathbf{A}\,\mathbf{x} = \mathbf{b}$ *tritt* ***immer genau einer*** *der folgenden Fälle auf:*

(1) $\mathbf{A}\,\mathbf{x} = \mathbf{b}$ *hat* ***genau eine*** *Lösung* $\mathbf{x} = \mathbf{z}$, *d.h.* $\mathbb{L} = \{\mathbf{z}\}$.

(2) $\mathbf{A}\,\mathbf{x} = \mathbf{b}$ *hat* ***unendlich viele*** *Lösungen. Ist* $\mathbf{x} = \mathbf{z}$ *eine Lösung von* $\mathbf{A}\,\mathbf{x} = \mathbf{b}$, *so ist die Lösungsmenge*

$$\mathbb{L} = \big\{\mathbf{z} + \mathbf{x} \;:\; \mathbf{A}\,\mathbf{x} = \mathbf{0}\big\}.$$

(3) $\mathbf{A}\,\mathbf{x} = \mathbf{b}$ *hat* ***keine*** *Lösung, d.h.* $\mathbb{L} = \{\} = \emptyset$.

Die Grundlage für das Gaußsche Eliminationsverfahren bildet die Information, dass **die Lösungsmenge $\mathbb{L}$ des linearen Gleichungssystems $\mathbf{A}\,\mathbf{x} = \mathbf{b}$ sich unter den folgenden elementaren Zeilenoperationen nicht ändert**:

(E1) Multiplikation einer Zeile mit $\lambda \in \mathbb{R} \setminus \{0\}$ (Notation: $Z_i \to \lambda\, Z_i$).

(E2) Ersetzen einer Zeile durch die Summe aus dieser Zeile und dem μ-fachen einer anderen Zeile ($\mu \in \mathbb{R}$) (Notation: $Z_i \to Z_i + \mu\, Z_j$, wobei $i \neq j$).

(E3) Vertauschen zweier Zeilen (Notation: $Z_i \leftrightarrow Z_j$).

(In der Notation bezeichnet Z_i bzw. Z_j die i-te bzw. j-te Zeile der erweiterten Koeffizientenmatrix.) Durch ein systematisches Anwenden dieser elementaren Zeilenoperationen auf die erweiterte Koeffizientenmatrix bringt man das lineare Gleichungssystem in **Stufenform**. Wir demonstrieren dieses an einem Beispiel.

Beispiel 2.15. (Gaußsches Eliminationsverfahren für LGS)
Das lineare Gleichungssystem

$$\begin{aligned} x_1 + 2\,x_2 + 3\,x_3 &= 1 \\ -\,x_1 + x_2 &= 2 \\ 2\,x_1 - 2\,x_2 + x_3 &= -2 \end{aligned}$$

hat die folgende erweiterte Koeffizientenmatrix:

$$\left[\begin{array}{ccc|c} 1 & 2 & 3 & 1 \\ -1 & 1 & 0 & 2 \\ 2 & -2 & 1 & -2 \end{array}\right].$$

Wir bringen diese nun mit elementaren Zeilenoperationen in Stufenform:

$$\left[\begin{array}{ccc|c} 1 & 2 & 3 & 1 \\ -1 & 1 & 0 & 2 \\ 2 & -2 & 1 & -2 \end{array}\right] \overset{Z_2 \to Z_2 + Z_1}{\Longleftrightarrow} \left[\begin{array}{ccc|c} 1 & 2 & 3 & 1 \\ 0 & 3 & 3 & 3 \\ 2 & -2 & 1 & -2 \end{array}\right]$$

$$\overset{Z_3 \to Z_3 - 2\,Z_1}{\Longleftrightarrow} \left[\begin{array}{ccc|c} 1 & 2 & 3 & 1 \\ 0 & 3 & 3 & 3 \\ 0 & -6 & -5 & -4 \end{array}\right] \overset{Z_3 \to Z_3 + 2\,Z_2}{\Longleftrightarrow} \left[\begin{array}{ccc|c} 1 & 2 & 3 & 1 \\ 0 & 3 & 3 & 3 \\ 0 & 0 & 1 & 2 \end{array}\right] \tag{2.19}$$

Das lineare Gleichungssystem rechts in (2.19) befindet sich nun in Stufenform.

Das lineare Gleichungssystem ganz rechts in (2.19) lautet nun:

$$\begin{aligned} x_1 + 2\,x_2 + 3\,x_3 &= 1 \quad (\mathrm{I}) \\ 3\,x_2 + 3\,x_3 &= 3 \quad (\mathrm{II}) \\ x_3 &= 2 \quad (\mathrm{III}) \end{aligned}$$

Mit Rückwärtsrechnen finden wir also:

$$\begin{aligned} &\text{Aus (III) :} && x_3 = 2 \\ &\text{In (II) einsetzen:} && x_2 = \frac{1}{3}\left(3 - 3\,x_3\right) = \frac{1}{3}\left(3 - 6\right) = -1 \end{aligned}$$

In (I) einsetzen: $x_1 = 1 - 2\,x_2 - 3\,x_3 = 1 - 2 \cdot (-1) - 3 \cdot 2 = -3$

Also ist die Lösungsmenge des LGS $\mathbb{L} = \left\{ \begin{bmatrix} -3 \\ -1 \\ 2 \end{bmatrix} \right\}$. ♠

Wir halten kurz fest, was beim Gaußschen Eliminationsverfahren und gegebenenfalls anschließendem Rückwärtsrechnen passiert und was wir dabei über die Lösbarkeit des linearen Gleichungssystems ablesen können.

Methode 2.16. (Gaußsches Eliminationsverfahren)
Wir betrachten ein lineares Gleichungssystem (LGS) $\mathbf{A}\,\mathbf{x} = \mathbf{b}$ *mit der Matrix* $\mathbf{A} \in \mathbb{R}^{m \times n}$ *und der rechten Seite* $\mathbf{b} \in \mathbb{R}^m$.

(1) Durch elementare Zeilenoperationen lässt sich jede erweiterte Koeffizientenmatrix $\left[\,\mathbf{A}\middle|\mathbf{b}\,\right]$ *in die sogenannte* ***Stufenform*** *bringen:*

$$\underbrace{\left[\begin{array}{cccccccccccccc|c}
0 & \cdots & 0 & \# & * & \cdots & * & * & * & \cdots & * & * & * & \cdots & * & * \\
\vdots & & \vdots & 0 & \cdots & \cdots & 0 & \# & * & \cdots & * & \vdots & \vdots & & \vdots & \vdots \\
\vdots & & \vdots & \vdots & & & \vdots & 0 & \cdots & \cdots & 0 & * & \vdots & & \vdots & \vdots \\
\vdots & & \vdots & \vdots & & & \vdots & \vdots & & & \vdots & \cdots & \# & * & \cdots & * & * \\
\vdots & & \vdots & \vdots & & & \vdots & \vdots & & & \vdots & & 0 & \cdots & \cdots & 0 & d_{r+1} \\
\vdots & & \vdots & \vdots & & & \vdots & \vdots & & & \vdots & & \vdots & & & \vdots & \vdots \\
0 & \cdots & 0 & 0 & \cdots & \cdots & 0 & 0 & \cdots & \cdots & 0 & & 0 & \cdots & \cdots & 0 & d_m
\end{array}\right]}_{n} \begin{matrix} \left.\begin{matrix} \\ \\ \\ \\ \\ \\ \\ \end{matrix}\right\} r \\ \left.\begin{matrix} \\ \\ \\ \\ \\ \\ \\ \end{matrix}\right\} m \end{matrix}$$

Die ***Stufenform*** *hat die folgenden Eigenschaften:*

- *Die Zeilen, in denen alle Koeffizienten null sind, sind in den unteren Zeilen der erweiterten Koeffizientenmatrix angeordnet.*
- *Jede andere Zeile der erweiterten Matrix ist von der Form*

$$\left[\; 0 \;\cdots\; 0 \;\; \# \;\; * \;\cdots\; * \,\middle|\, * \;\right],$$

  wobei das #-Symbol jeweils für eine beliebige Zahl in $\mathbb{R} \setminus \{0\}$ *steht und die* $*$*-Symbole jeweils für beliebige reelle Zahlen stehen.*
- *Wandert man durch die Zeilen der Matrix von oben nach unten, so muss die (erste) Zahl # in einer Zeile immer weiter rechts als in der vorhergehenden Zeile auftreten.*

(2) Es gilt immer $r \leq \min\{m, n\}$.

(3) ***Lösbarkeit:***

Fall 1: $r = m$ *oder* $(r < m$ *und* $d_{r+1} = \ldots = d_m = 0)$

$\Longrightarrow$ *Das LGS ist lösbar, d.h.* $\mathbb{L} \neq \emptyset$.

Falls $r = n$ *ist, hat das LGS genau eine Lösung.*

Falls $r < n$ *ist, hat das LGS unendlich viele Lösungen.*

Die Anzahl der frei wählbaren Parameter ist dann $n - r$.

Fall 2: $r < m$ *und* $d_i \neq 0$ *für mindestens ein* $i > r$

$\Longrightarrow$ *Das LGS ist unlösbar, d.h.* $\mathbb{L} = \emptyset$.

(4) In Fall 1 erhält man die Lösungsmenge aus der Stufenform durch ***Rückwärtsrechnen****.*

Als eine elementare Zeilenoperation ist es möglich Zeilen zu tauschen, und manchmal ist dieses auch erforderlich, weil wir das lineare Gleichungssystem sonst nicht in Stufenform bringen können. Das nächste Beispiel zeigt, dass es auch aus anderen Gründen sinnvoll sein kann, einen Zeilentausch vorzunehmen.

Beispiel 2.17. (Gaußsches Eliminationsverfahren)
Wir wollen das folgende lineare Gleichungssystem in einem Gleitkomma-Zahlensystem mit der Mantissenlänge 4 lösen. Es wird also in jedem einzelnen Rechenschritt auf eine 4-stellige Mantisse gerundet.

$$\begin{aligned} 0{,}729\,x_1 + 0{,}81\,x_2 + 0{,}9\,x_3 &= 0{,}6867 \\ x_1 + \quad x_2 + \quad x_3 &= 0{,}8338 \\ 1{,}331\,x_1 + 1{,}21\,x_2 + 1{,}1\,x_3 &= 1{,}000 \end{aligned} \tag{2.20}$$

Gerundet auf eine Mantisse der Länge 4 ist die exakte Lösung

$$x_1 \doteq 0{,}2245, \qquad x_2 \doteq 0{,}2814, \qquad x_3 \doteq 0{,}3279, \tag{2.21}$$

wie man leicht durch Einsetzen in (2.20) überprüft.

Wir lösen das LGS nun mit dem Gaußschen Eliminationsverfahren systematisch nach Schema (d.h. ohne gegebenenfalls „geschickte“ Zeilenumformungen außerhalb der üblichen Vorgehensweise vorzunehmen) und **ohne Zeilentausch**, was in diesem konkreten Beispiel geht.

$$\left[\begin{array}{ccc|c} 0{,}7290 & 0{,}8100 & 0{,}9000 & 0{,}6867 \\ 1{,}000 & 1{,}000 & 1{,}000 & 0{,}8338 \\ 1{,}331 & 1{,}210 & 1{,}100 & 1{,}000 \end{array}\right]$$

$$\begin{array}{c} Z_2 \to Z_2 - 1{,}372\, Z_1 \\ Z_3 \to Z_3 - 1{,}826\, Z_1 \\ \downarrow \\ \Longleftrightarrow \end{array} \quad \left[\begin{array}{rrr|r} 0{,}7290 & 0{,}8100 & 0{,}9000 & 0{,}6867 \\ 0{,}0 & -0{,}1110 & -0{,}2350 & -0{,}1084 \\ 0{,}0 & -0{,}2690 & -0{,}5430 & -0{,}2540 \end{array}\right]$$

$$\begin{array}{c} Z_3 \to Z_3 - 2{,}423\, Z_2 \\ \downarrow \\ \Longleftrightarrow \end{array} \quad \left[\begin{array}{rrr|r} 0{,}7290 & 0{,}8100 & 0{,}9000 & 0{,}6867 \\ 0{,}0 & -0{,}1110 & -0{,}2350 & -0{,}1084 \\ 0{,}0 & 0{,}0 & 0{,}02640 & 0{,}008700 \end{array}\right]$$

Dabei haben wir im ersten Schritt bzw. im zweiten Schritt genutzt, dass die Faktoren, mit denen die erste bzw. zweite Zeile multipliziert wird, um dann von der entsprechenden unteren Zeile subtrahiert zu werden, (mit Rundung auf die Mantissenlänge 4) wie folgt lauten:

$$\frac{1}{0{,}7290} \doteq 1{,}372, \qquad \frac{1{,}331}{0{,}7290} \doteq 1{,}826, \qquad \frac{-0{,}2690}{-0{,}1110} \doteq 2{,}423.$$

Rückwärtsrechnen liefert (mit in jedem Schritt Rundung auf Mantissenlänge 4):

$$\begin{aligned} x_3 &= \frac{0{,}008700}{0{,}02640} \doteq 0{,}3295 \\ x_2 &= \frac{-0{,}1084 + 0{,}2350 \cdot 0{,}3295}{-0{,}1110} \doteq 0{,}2790, \\ x_1 &= \frac{0{,}6867 - 0{,}8100 \cdot 0{,}2790 - 0{,}9000 \cdot 0{,}3295}{0{,}7290} \doteq 0{,}2251. \end{aligned} \tag{2.22}$$

Vergleicht man (2.22) mit (2.21), so sieht man, dass x_2 nur eine signifikante Ziffer und x_1 und x_3 nur zwei signifikante Ziffern haben. Die Genauigkeit des berechneten Ergebnisses lässt deutlich zu wünschen übrig!

Woran liegt die schlechte Qualität des Ergebnisses? Wir haben bei der Berechnung der Faktoren, mit denen die erste bzw. zweite Zeile multipliziert wird, um dann von der entsprechenden unteren Zeile subtrahiert zu werden, immer eine betraglich größere Zahl durch eine betraglich kleinere Zahl geteilt, und dabei werden die absoluten Fehler, also auch die Rundungsfehler, verstärkt (siehe Übungsaufgabe). Man kann dieses Problem umgehen, indem man in jedem Schritt des Gaußschen Eliminationsverfahrens einen geeigneten Zeilentausch vornimmt, so dass anschließend bei der Berechnung der Faktoren $a_{i,j}/a_{j,j}$, $i = j+1, \ldots, n$, (mit denen die j-te Zeile im j-ten Schritt multipliziert und dann jeweils zur i-ten Zeile addiert wird) höchstens durch eine betraglich gleich große Zahl geteilt wird. (Die Einträge $a_{i,j}$ und $a_{j,j}$ sind hier jeweils die Einträge der modifizierten Matrix nach $j-1$ Schritten.) Wir illustrieren dieses am Beispiel; danach formulieren wir die

Vorgehensweise allgemein.

$$\left[\begin{array}{cccc} 0{,}7290 & 0{,}8100 & 0{,}9000 & 0{,}6867 \\ 1{,}000 & 1{,}000 & 1{,}000 & 0{,}8338 \\ 1{,}331 & 1{,}210 & 1{,}100 & 1{,}000 \end{array}\right]$$

$$\begin{array}{c} Z_1 \leftrightarrow Z_3 \\ \downarrow \\ \Longleftrightarrow \end{array} \quad \left[\begin{array}{cccc} 1{,}331 & 1{,}210 & 1{,}100 & 1{,}000 \\ 1{,}000 & 1{,}000 & 1{,}000 & 0{,}8338 \\ 0{,}7290 & 0{,}8100 & 0{,}9000 & 0{,}6867 \end{array}\right]$$

$$\begin{array}{c} Z_2 \to Z_2 - 0{,}7513\, Z_1 \\ Z_3 \to Z_3 - 0{,}5477\, Z_1 \\ \downarrow \\ \Longleftrightarrow \end{array} \quad \left[\begin{array}{cccc} 1{,}331 & 1{,}210 & 1{,}100 & 1{,}000 \\ 0{,}0 & 0{,}09090 & 0{,}1736 & 0{,}08250 \\ 0{,}0 & 0{,}1473 & 0{,}2975 & 0{,}1390 \end{array}\right]$$

$$\begin{array}{c} Z_2 \leftrightarrow Z_3 \\ \downarrow \\ \Longleftrightarrow \end{array} \quad \left[\begin{array}{cccc} 1{,}331 & 1{,}210 & 1{,}100 & 1{,}000 \\ 0{,}0 & 0{,}1473 & 0{,}2975 & 0{,}1390 \\ 0{,}0 & 0{,}09090 & 0{,}1736 & 0{,}08250 \end{array}\right]$$

$$\begin{array}{c} Z_3 \to Z_3 - 0{,}6171\, Z_2 \\ \downarrow \\ \Longleftrightarrow \end{array} \quad \left[\begin{array}{cccc} 1{,}331 & 1{,}210 & 1{,}100 & 1{,}000 \\ 0{,}0 & 0{,}1473 & 0{,}2975 & 0{,}1390 \\ 0{,}0 & 0{,}0 & -0{,}01000 & -0{,}003280 \end{array}\right] \tag{2.23}$$

Dabei haben wir im ersten Schritt bzw. im zweiten Schritt genutzt, dass die Faktoren, mit denen die erste bzw. zweite Zeile multipliziert wird, um dann von der entsprechenden unteren Zeile subtrahiert zu werden, (mit Rundung auf Mantissenlänge 4) wie folgt lauten:

$$\frac{1{,}000}{1{,}331} \doteq 0{,}7513, \qquad \frac{0{,}7290}{1{,}331} \doteq 0{,}5477, \qquad \frac{0{,}09090}{0{,}1473} \doteq 0{,}6171.$$

Wir haben also, bevor wir im ersten Schritt des Gaußschen Eliminationsverfahrens die erste Spalte bearbeitet haben, zuerst die Zeile mit dem betraglich größten Eintrag in der ersten Spalte mit der ersten Zeile getauscht. Erst danach haben haben wir die Eliminationsschritte durchgeführt. Durch den Zeilentausch wird es vermieden, dass wir eine betraglich größere Zahl durch eine betraglich kleinere Zahl teilen. Analog wird in zweiten Schritt des Gaußschen Eliminationsverfahrens vor der Bearbeitung der zweiten Spalte vorgegangen. Hier wird natürlich für den Zeilentausch nur nachgeschaut, wo in der zweiten Spalte von der zweiten Zeile abwärts das betraglich größte Element auftritt.

Berechnen wir die Lösung von (2.23) durch Rückwärtsrechnen und schauen, ob sich die Genauigkeit verbessert hat. Rückwärtsrechnen liefert (ebenfalls mit in

jedem Schritt Rundung auf die Mantissenlänge 4):

$$\begin{aligned} x_3 &= \frac{-0{,}003280}{-0{,}01000} \doteq 0{,}3280 \\ x_2 &= \frac{0{,}1390 - 0{,}2975 \cdot 0{,}3280}{0{,}1473} \doteq 0{,}2812, \\ x_1 &= \frac{1{,}000 - 1{,}210 \cdot 0{,}2812 - 1{,}100 \cdot 0{,}3280}{1{,}331} \doteq 0{,}2246. \end{aligned} \tag{2.24}$$

Vergleicht man (2.24) mit (2.21), so sieht man, dass x_3 nun zwei signifikante Ziffern hat und x_1 und x_2 jeweils drei signifikante Ziffern haben. ♠

Die im vorherigen Beispiel angewendete Vorgehensweise des systematischen Zeilentauschs nennt sich eine **(partielle) Pivotstrategie** für das Gaußsche Eliminationsverfahren, und wir wollen diese Vorgehensweise nun allgemein beschreiben.

Verfahren 2.18. (Gaußsches Eliminationsverf. mit Pivotstrategie)
*Sei $\mathbf{A} \in \mathbb{R}^{m \times n}$ und sei $\mathbf{b} \in \mathbb{R}^m$. Die Lösungsmenge des linearen Gleichungssystems $\mathbf{A}\,\mathbf{x} = \mathbf{b}$ lässt sich mit den **Gaußschen Eliminationsverfahren mit (partieller) Pivotstrategie** und mit anschließendem Rückwärtsrechnen wie folgt berechnen:*

Initialisierung: $\mathbf{A}^{(1)} := \mathbf{A}$, $\mathbf{b}^{(1)} := \mathbf{b}$

Für $j = 1, 2, \ldots, m-1$ (Zeilenzähler) führe die folgenden Schritte durch:

(1) Suche nach dem Index q der ersten Spalte mit $\begin{bmatrix} a_{j,q}^{(j)} \\ \vdots \\ a_{m,q}^{(j)} \end{bmatrix} \neq \mathbf{0}$.

(Falls sich kein q mit dieser Eigenschaft findet, bricht das Verfahren ab und wir setzen $\mathbf{A}^{(m)} := \mathbf{A}^{(j)}$ und $\mathbf{b}^{(m)} := \mathbf{b}^{(j)}$.)

In $\begin{bmatrix} a_{j,q}^{(j)} \\ \vdots \\ a_{m,q}^{(j)} \end{bmatrix}$ bestimme den Index $p \in \{j, \ldots, m\}$, für den $|a_{p,q}^{(j)}|$ am

größten ist. Tausche nun die j-te und die p-te Zeile von $\left[\mathbf{A}^{(j)} | \mathbf{b}^{(j)}\right]$.

*(Die Zahl $|a_{p,q}^{(j)}|$ heißt das **Pivotelement** der q-ten Spalte.)*

Die modifizierte Matrix und die modifizierte rechte Seite nach dem Zeilentausch seien mit $\widetilde{\mathbf{A}}^{(j)}$ und $\widetilde{\mathbf{b}}^{(j)}$ bezeichnet.

(2) Eliminationsschritt: Sei $\mathbf{A}^{(j+1)} := \widetilde{\mathbf{A}}^{(j)}$ *und* $\mathbf{b}^{(j+1)} := \widetilde{\mathbf{b}}^{(j)}$. *Für* $i = j+1, \ldots, m$ *überschreiben wir in* $\mathbf{A}^{(j+1)}$ *und* $\mathbf{b}^{(j+1)}$ *nun folgende Einträge:*

$$a_{i,k}^{(j+1)} := \widetilde{a}_{i,k}^{(j)} - \frac{\widetilde{a}_{i,q}^{(j)}}{\widetilde{a}_{j,q}^{(j)}} \cdot \widetilde{a}_{j,k}^{(j)}, \quad k = q, \ldots, n; \qquad b_i^{(j+1)} := \widetilde{b}_i^{(j)} - \frac{\widetilde{a}_{i,q}^{(j)}}{\widetilde{a}_{j,q}^{(j)}} \cdot \widetilde{b}_j^{(j)}$$

Nach höchstens m *Schritten bricht der Algorithmus ab: Wir erhalten dann eine Matrix* $\mathbf{A}^{(m)}$ *in Stufenform und eine rechte Seite* $\mathbf{b}^{(m)}$, *und das lineare Gleichungssystem* $\mathbf{A}^{(m)}\,\mathbf{x} = \mathbf{b}^{(m)}$ *kann durch Rückwärtsrechnen gelöst werden.*

Wie viele elementare Rechenoperationen benötigt das Gaußsche Eliminationsverfahren höchstens? Beim Rückwärtsrechnen benötigen wir zur Berechnung von x_j, $j = n, n-1, \ldots, 1$, höchstens $2(n-j)+1$ elementare Rechenoperationen, also insgesamt

$$\sum_{j=1}^{n} \big(2\,(n-j)+1\big) \overset{k=n-j}{\underset{\downarrow}{=}} \sum_{k=0}^{n-1} (2\,k+1) \leq \int_0^n (2\,k+1)\,\mathrm{d}k = n^2 + n = \mathcal{O}(n^2)$$

elementare Rechenoperationen. Bei dem Umformen in Stufenform benötigt man im j-ten Schritt höchstens $(n-j)\big(2\,n+3-2\,(j-1)\big) = (n-j)\big(2\,(n-j)+5\big)$ elementare Rechenoperationen, also insgesamt

$$\sum_{j=1}^{n-1} (n-j)\big(2\,(n-j)+5\big) \overset{k=n-j}{\underset{\downarrow}{=}} \sum_{k=1}^{n-1} (2\,k^2+5\,k) \leq \int_1^n (2\,k^2+5\,k)\,\mathrm{d}k$$

$$= \left[\frac{2}{3}\,k^3 + \frac{5}{2}\,k^2\right]_{k=1}^{k=n} = \frac{2}{3}\,n^3 + \frac{5}{2}\,n^2 - \left(\frac{2}{3}+\frac{5}{2}\right) = \mathcal{O}(n^3)$$

elementare Rechenoperationen. Also benötigt das **Gaußsche Eliminationsverfahren einschließlich des anschließenden Rückwärtsrechnens** insgesamt $\mathcal{O}(n^3) + \mathcal{O}(n^2) = \mathcal{O}(n^3)$ **elementare Rechenoperationen**.

Wir wollen nun noch den Sonderfall betrachten, dass die Matrix $\mathbf{A} \in \mathbb{R}^{n\times n}$ **quadratisch und regulär** ist. In diesem Fall liefert uns das Gaußsche Eliminationsverfahren mit Pivotstrategie eine sogenannte **LR-Zerlegung** von $\mathbf{A}$:

$$\mathbf{P}\,\mathbf{A} = \mathbf{L}\,\mathbf{R},$$

wobei $\mathbf{P}$ eine **Permutationsmatrix** ist, welche die im Gaußschen Eliminationsverfahren mit Pivotstrategie vorgenommenen Zeilenvertauschungen widerspiegelt, $\mathbf{L}$ eine **untere Dreiecksmatrix mit Einsen auf der Diagonalen** ist und $\mathbf{R}$

die aus dem Eliminationsverfahren mit Pivotstrategie erhaltene Stufenform ist. Für reguläres $\mathbf{A} \in \mathbb{R}^{n\times n}$ ist $\mathbf{R}$ eine **obere Dreiecksmatrix mit allen Diagonaleinträgen ungleich null**. Das nachfolgende Verfahren gibt an, wie man diese Matrizen genau berechnet.

Verfahren 2.19. (LR-Zerlegung)
*Sei $\mathbf{A} \in \mathbb{R}^{n\times n}$ **regulär**. Die **LR-Zerlegung** von $\mathbf{A}$ berechnet sich wie folgt:*

Initialisierung: $\mathbf{A}^{(1)} := \mathbf{A}$, $\mathbf{P}^{(1)} := \mathbf{E}_n$, $\mathbf{L}^{(1)} := \mathbf{E}_n$

Für $j = 1, 2, \ldots, n-1$ führe die folgenden Schritte durch:

(1) In $\begin{bmatrix} a_{j,j}^{(j)} \\ \vdots \\ a_{n,j}^{(j)} \end{bmatrix}$ *bestimme den Index $p \in \{j, \ldots, n\}$, für den $|a_{p,j}^{(j)}|$ am größten ist. Tausche nun jeweils die j-te und die p-te Zeile von $\mathbf{A}^{(j)}$ und $\mathbf{P}^{(j)}$. Setze weiter $\mathbf{L}^{(j+1)} := \mathbf{L}^{(j)}$ und überschreibe danach*

$$\ell_{j,k}^{(j+1)} := \ell_{p,k}^{(j)}, \qquad k = 1, \ldots, j-1,$$
$$\ell_{p,k}^{(j+1)} := \ell_{j,k}^{(j)}, \qquad k = 1, \ldots, j-1.$$

Die modifizierte Matrix $\mathbf{A}^{(j)}$ und die modifizierte Matrix $\mathbf{P}^{(j)}$ nach dem Zeilentausch seien mit $\widetilde{\mathbf{A}}^{(j)}$ bzw. $\mathbf{P}^{(j+1)}$ bezeichnet.

(2) Setze $\mathbf{A}^{(j+1)} := \widetilde{\mathbf{A}}^{(j)}$. Für $i = j+1, \ldots, n$ überschreiben wir in $\mathbf{L}^{(j+1)}$ und $\mathbf{A}^{(j+1)}$ nun folgende Einträge:

$$\ell_{i,j}^{(j+1)} := \frac{\widetilde{a}_{i,j}^{(j)}}{\widetilde{a}_{j,j}^{(j)}}; \qquad a_{i,k}^{(j+1)} := \widetilde{a}_{i,k}^{(j)} - \ell_{i,j}^{(j+1)} \cdot \widetilde{a}_{j,k}^{(j)}, \quad k = j, \ldots, n$$

Setze nun $\mathbf{P} := \mathbf{P}^{(n)}$ und $\mathbf{R} := \mathbf{A}^{(n)}$ und $\mathbf{L} := \mathbf{L}^{(n)}$. Dann gilt

$$\mathbf{P}\,\mathbf{A} = \mathbf{L}\,\mathbf{R}.$$

*$\mathbf{L} \in \mathbb{R}^{n\times n}$ ist eine **untere Dreiecksmatrix mit Einsen auf der Diagonalen**, und $\mathbf{R} \in \mathbb{R}^{n\times n}$ ist eine **obere Dreiecksmatrix mit allen Diagonaleinträgen ungleich null**. Die **Permutationsmatrix** $\mathbf{P} \in \mathbb{R}^{n\times n}$ geht aus der Einheitsmatrix durch wiederholte Zeilenvertauschungen hervor und spiegelt die Zeilenvertauschungen in der Pivotstrategie wider.*

Beweis: Den Beweis bzw. die Herleitung der LR-Zerlegung kann man beispielsweise in [2, Teilkapitel 4.2] oder in [7, Teilkapitel 4.3] nachlesen. □

Berechnen wir die LR-Zerlegung einer Matrix per Hand, um uns die Vorgehensweise in Verfahren 2.19 klar zu machen.

Beispiel 2.20. (LR-Zerlegung)
Gesucht ist die LR-Zerlegung (mittels Verfahren 2.19) der regulären Matrix

$$\mathbf{A} := \begin{bmatrix} 1 & 2 & 3 \\ -1 & 2 & 0 \\ 2 & -2 & 1 \end{bmatrix}.$$

(Die Matrix regulär ist, denn $\det(\mathbf{A}) = 2 + 0 + 6 - 12 - 0 - (-2) = -2 \neq 0$.)

Berechnung der LR-Zerlegung mit Verfahren 2.19:

Initialisierung: $\mathbf{A}^{(1)} := \begin{bmatrix} 1 & 2 & 3 \\ -1 & 2 & 0 \\ 2 & -2 & 1 \end{bmatrix}$, $\mathbf{P}^{(1)} := \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}$, $\mathbf{L}^{(1)} := \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}$

Schritt $j = 1$**:** Da $a_{3,1}^{(1)} = 2$ der betraglich größte Eintrag in der ersten Spalte von $\mathbf{A}^{(1)}$ ist, müssen wir in $\mathbf{A}^{(1)}$ (und dann in $\mathbf{P}^{(1)}$) die erste und dritte Zeile tauschen:

$$\mathbf{A}^{(1)} = \begin{bmatrix} 1 & 2 & 3 \\ -1 & 2 & 0 \\ 2 & -2 & 1 \end{bmatrix} \quad \begin{matrix} Z_1 \leftrightarrow Z_3 \\ \downarrow \\ \Longleftrightarrow \end{matrix} \quad \begin{bmatrix} 2 & -2 & 1 \\ -1 & 2 & 0 \\ 1 & 2 & 3 \end{bmatrix} =: \widetilde{\mathbf{A}}^{(1)}$$

Damit bekommen wir die folgende neue Permutationsmatrix:

$$\mathbf{P}^{(1)} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \quad \begin{matrix} Z_1 \leftrightarrow Z_3 \\ \downarrow \\ \Longleftrightarrow \end{matrix} \quad \begin{bmatrix} 0 & 0 & 1 \\ 0 & 1 & 0 \\ 1 & 0 & 0 \end{bmatrix} =: \mathbf{P}^{(2)}$$

Weiter gilt vorläufig $\mathbf{L}^{(2)} := \mathbf{L}^{(1)}$.

Nun erhalten wir $\ell_{2,1}^{(2)} := \dfrac{\widetilde{a}_{2,1}^{(1)}}{\widetilde{a}_{1,1}^{(1)}} = -\dfrac{1}{2}$ und $\ell_{3,1}^{(2)} := \dfrac{\widetilde{a}_{3,1}^{(1)}}{\widetilde{a}_{1,1}^{(1)}} = \dfrac{1}{2}$ und

$$\widetilde{\mathbf{A}}^{(1)} = \begin{bmatrix} 2 & -2 & 1 \\ -1 & 2 & 0 \\ 1 & 2 & 3 \end{bmatrix} \quad \begin{matrix} Z_2 \to Z_2 - \left(-\frac{1}{2}\right) Z_1 \\ Z_3 \to Z_3 - \frac{1}{2} Z_1 \\ \downarrow \\ \Longleftrightarrow \end{matrix} \quad \begin{bmatrix} 2 & -2 & 1 \\ 0 & 1 & \frac{1}{2} \\ 0 & 3 & \frac{5}{2} \end{bmatrix} =: \mathbf{A}^{(2)},$$

sowie die endgültige Matrix $\mathbf{L}^{(2)}$ als

$$\mathbf{L}^{(2)} := \begin{bmatrix} 1 & 0 & 0 \\ -\frac{1}{2} & 1 & 0 \\ \frac{1}{2} & 0 & 1 \end{bmatrix}.$$

Schritt $j = 2$**:** Da $a_{3,2}^{(2)} = 3$ der betraglich größte Eintrag in der zweiten Spalte von der zweiten Zeile abwärts in $\mathbf{A}^{(2)}$ ist, müssen wir in $\mathbf{A}^{(2)}$ (und dann in $\mathbf{P}^{(2)}$) die zweite und die dritte Zeile tauschen:

$$\mathbf{A}^{(2)} = \begin{bmatrix} 2 & -2 & 1 \\ 0 & 1 & \frac{1}{2} \\ 0 & 3 & \frac{5}{2} \end{bmatrix} \quad \begin{array}{c} Z_2 \leftrightarrow Z_3 \\ \downarrow \\ \Longleftrightarrow \end{array} \quad \begin{bmatrix} 2 & -2 & 1 \\ 0 & 3 & \frac{5}{2} \\ 0 & 1 & \frac{1}{2} \end{bmatrix} =: \widetilde{\mathbf{A}}^{(2)}$$

Damit bekommen wir die folgende neue Permutationsmatrix:

$$\mathbf{P}^{(2)} = \begin{bmatrix} 0 & 0 & 1 \\ 0 & 1 & 0 \\ 1 & 0 & 0 \end{bmatrix} \quad \begin{array}{c} Z_2 \leftrightarrow Z_3 \\ \downarrow \\ \Longleftrightarrow \end{array} \quad \begin{bmatrix} 0 & 0 & 1 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \end{bmatrix} =: \mathbf{P}^{(3)}$$

Wir setzen nun $\mathbf{L}^{(3)} := \mathbf{L}^{(2)}$.

Nun müssen wir (wegen des Zeilentauschs) im $\mathbf{L}^{(3)} = \mathbf{L}^{(2)}$ folgende Änderungen vornehmen, um eine vorläufige Version von $\mathbf{L}^{(3)}$ zu bekommen:

$$\ell_{2,1}^{(3)} := \ell_{3,1}^{(2)} = \frac{1}{2} \qquad \text{und} \qquad \ell_{3,1}^{(3)} := \ell_{2,1}^{(2)} = -\frac{1}{2}$$

Also erhalten wir als vorläufige Version von $\mathbf{L}^{(3)}$

$$\mathbf{L}^{(3)} := \begin{bmatrix} 1 & 0 & 0 \\ \frac{1}{2} & 1 & 0 \\ -\frac{1}{2} & 0 & 1 \end{bmatrix}.$$

Nun erhalten wir $\ell_{3,2}^{(3)} := \dfrac{\widetilde{a}_{3,2}^{(2)}}{\widetilde{a}_{2,2}^{(2)}} = \dfrac{1}{3}$ und

$$\widetilde{\mathbf{A}}^{(2)} = \begin{bmatrix} 2 & -2 & 1 \\ 0 & 3 & \frac{5}{2} \\ 0 & 1 & \frac{1}{2} \end{bmatrix} \quad \begin{array}{c} Z_3 \to Z_3 - \frac{1}{3} Z_2 \\ \downarrow \\ \Longleftrightarrow \end{array} \quad \begin{bmatrix} 2 & -2 & 1 \\ 0 & 3 & \frac{5}{2} \\ 0 & 0 & -\frac{1}{3} \end{bmatrix} =: \mathbf{A}^{(3)},$$

sowie die finale Version der Matrix $\mathbf{L}^{(3)}$ als

$$\mathbf{L}^{(3)} := \begin{bmatrix} 1 & 0 & 0 \\ \frac{1}{2} & 1 & 0 \\ -\frac{1}{2} & \frac{1}{3} & 1 \end{bmatrix}.$$

Mit den Matrizen

$$\mathbf{L} = \mathbf{L}^{(3)} = \begin{bmatrix} 1 & 0 & 0 \\ \frac{1}{2} & 1 & 0 \\ -\frac{1}{2} & \frac{1}{3} & 1 \end{bmatrix}, \quad \mathbf{R} := \mathbf{A}^{(3)} = \begin{bmatrix} 2 & -2 & 1 \\ 0 & 3 & \frac{5}{2} \\ 0 & 0 & -\frac{1}{3} \end{bmatrix}, \quad \mathbf{P} := \mathbf{P}^{(3)} = \begin{bmatrix} 0 & 0 & 1 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \end{bmatrix}$$

gilt dann $\mathbf{L}\,\mathbf{R} = \mathbf{P}\,\mathbf{A}$. In der Tat finden wir

$$\mathbf{L}\,\mathbf{R} = \begin{bmatrix} 1 & 0 & 0 \\ \frac{1}{2} & 1 & 0 \\ -\frac{1}{2} & \frac{1}{3} & 1 \end{bmatrix} \cdot \begin{bmatrix} 2 & -2 & 1 \\ 0 & 3 & \frac{5}{2} \\ 0 & 0 & -\frac{1}{3} \end{bmatrix} = \begin{bmatrix} 2 & -2 & 1 \\ 1 & 2 & 3 \\ -1 & 2 & 0 \end{bmatrix} = \mathbf{P}\,\mathbf{A}$$

wie erwartet. ♠

Bemerkung 2.21. (LGS mittels LR-Zerlegung lösen)
Sei $\mathbf{A} \in \mathbb{R}^{n\times n}$ **regulär** (also invertierbar), so dass uns Verfahren 2.19 eine **LR-Zerlegung** von $\mathbf{A}$

$$\mathbf{L}\,\mathbf{R} = \mathbf{P}\,\mathbf{A}$$

liefert. Dann kann man diese LR-Zerlegung (sofern diese bereits vorliegt) nutzen, um das lineare Gleichungssystem $\mathbf{A}\,\mathbf{x} = \mathbf{b}$ effizient zu lösen:

$$\mathbf{A}\,\mathbf{x} = \mathbf{b} \quad \Longleftrightarrow \quad \mathbf{P}\,\mathbf{A}\,\mathbf{x} = \mathbf{P}\,\mathbf{b} \quad \Longleftrightarrow \quad \mathbf{L}\,\mathbf{R}\,\mathbf{x} = \mathbf{P}\,\mathbf{b}$$

Also können wir erst mit **Vorwärtsrechnen**

$$\mathbf{L}\,\mathbf{y} = \mathbf{P}\,\mathbf{b} \tag{2.25}$$

lösen und dann mit **Rückwärtsrechnen**

$$\mathbf{R}\,\mathbf{x} = \mathbf{y} \tag{2.26}$$

lösen. Da $\mathbf{L}$ eine untere Dreiecksmatrix ist, kann (2.25) mit Vorwärtsrechnen in $\mathcal{O}(n^2)$ Operationen gelöst werden. Da $\mathbf{R}$ eine obere Dreiecksmatrix ist, kann (2.26) mit Rückwärtsrechnen in $\mathcal{O}(n^2)$ elementaren Rechenoperationen gelöst werden. Liegt die LR-Zerlegung von $\mathbf{A}$ also bereits vor, so kann man Hilfe

von dieser das lineare Gleichungssystem $\mathbf{A}\,\mathbf{x} = \mathbf{b}$ mit $\mathcal{O}(n^2)$ **elementaren Rechenoperationen** lösen.

Die **Berechnung der LR-Zerlegung** selber erfordert allerdings wie das Gaußsche Eliminationsverfahren $\mathcal{O}(n^3)$ **elementare Rechenoperationen**. Also ist es vor allem dann interessant, erst die LR-Zerlegung von $\mathbf{A}$ zu berechnen, wenn man viele lineare Gleichungssysteme mit der gleichen regulären Matrix $\mathbf{A}$ lösen muss.

Betrachten wir ein Beispiel, in dem ein lineares Gleichungssystem mit einer invertierbaren Matrix mit Hilfe von deren LR-Zerlegung gelöst wird.

Beispiel 2.22. (LGS mit LR-Zerlegung lösen)
Gesucht ist die Lösung des LGS $\mathbf{A}\,\mathbf{x} = \mathbf{b}$ mit

$$\mathbf{A} := \begin{bmatrix} 1 & 2 & 3 \\ -1 & 2 & 0 \\ 2 & -2 & 1 \end{bmatrix}, \qquad \mathbf{b} = \begin{bmatrix} 5 \\ -3 \\ 6 \end{bmatrix}$$

mit Hilfe der LR-Zerlegung von $\mathbf{A}$. In Beispiel 2.20 haben wir die LR-Zerlegung $\mathbf{L}\,\mathbf{R} = \mathbf{P}\,\mathbf{A}$ von $\mathbf{A}$ berechnet und fanden dabei

$$\mathbf{L} = \begin{bmatrix} 1 & 0 & 0 \\ \frac{1}{2} & 1 & 0 \\ -\frac{1}{2} & \frac{1}{3} & 1 \end{bmatrix}, \quad \mathbf{R} = \begin{bmatrix} 2 & -2 & 1 \\ 0 & 3 & \frac{5}{2} \\ 0 & 0 & -\frac{1}{3} \end{bmatrix}, \quad \mathbf{P} = \begin{bmatrix} 0 & 0 & 1 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \end{bmatrix}.$$

Wegen

$$\mathbf{A}\,\mathbf{x} = \mathbf{b} \qquad \Longleftrightarrow \qquad \mathbf{P}\,\mathbf{A}\,\mathbf{x} = \mathbf{P}\,\mathbf{b} \qquad \Longleftrightarrow \qquad \mathbf{L}\,\underbrace{\mathbf{R}\,\mathbf{x}}_{=\mathbf{y}} = \mathbf{P}\,\mathbf{b}$$

lösen wir zuerst

$$\mathbf{L}\,\mathbf{y} = \mathbf{P}\,\mathbf{b} \qquad \Longleftrightarrow \qquad \begin{bmatrix} 1 & 0 & 0 \\ \frac{1}{2} & 1 & 0 \\ -\frac{1}{2} & \frac{1}{3} & 1 \end{bmatrix} \cdot \begin{bmatrix} y_1 \\ y_2 \\ y_3 \end{bmatrix} = \begin{bmatrix} 0 & 0 & 1 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \end{bmatrix} \cdot \begin{bmatrix} 5 \\ -3 \\ 6 \end{bmatrix} = \begin{bmatrix} 6 \\ 5 \\ -3 \end{bmatrix}.$$

Vorwärtsrechnen liefert:

$$\begin{aligned} y_1 &= 6, \\ y_2 &= 5 - \frac{1}{2}\,y_1 = 5 - \frac{1}{2} \cdot 6 = 2, \end{aligned}$$

$$y_3 = -3 + \frac{1}{2}\,y_1 - \frac{1}{3}\,y_2 = -3 + \frac{1}{2}\cdot 6 - \frac{1}{3}\cdot 2 = -\frac{2}{3}\,,$$

also $\mathbf{y} = \begin{bmatrix} 6 \\ 2 \\ -\frac{2}{3} \end{bmatrix}$. Nun müssen wir noch

$$\mathbf{R}\,\mathbf{x} = \mathbf{y} \qquad \Longleftrightarrow \qquad \begin{bmatrix} 2 & -2 & 1 \\ 0 & 3 & \frac{5}{2} \\ 0 & 0 & -\frac{1}{3} \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} 6 \\ 2 \\ -\frac{2}{3} \end{bmatrix}$$

lösen. Rückwärtsrechnen liefert:

$$\begin{aligned} x_3 &= (-3)\cdot\left(-\frac{2}{3}\right) = 2, \\ x_2 &= \frac{1}{3}\left(2 - \frac{5}{2}\,x_3\right) = \frac{1}{3}\left(2 - \frac{5}{2}\cdot 2\right) = -1, \\ x_1 &= \frac{1}{2}\left(6 + 2\,x_2 - x_3\right) = \frac{1}{2}\left(6 + 2\cdot(-1) - 2\right) = 1. \end{aligned}$$

Also ist der Lösungsvektor von $\mathbf{A}\,\mathbf{x} = \mathbf{b}$ durch $\mathbf{x} = \begin{bmatrix} 1 \\ -1 \\ 2 \end{bmatrix}$ gegeben. ♠

2.4 QR-Zerlegung

In diesem Teilkapitel betrachten wir nur **quadratische Matrizen**.

Bei der **QR-Zerlegung** einer regulären Matrix $\mathbf{A} \in \mathbb{R}^{n\times n}$ geht es um eine Zerlegung $\mathbf{A} = \mathbf{Q}\,\mathbf{R}$, bei der $\mathbf{R} \in \mathbb{R}^{n\times n}$ ebenfalls eine **obere Dreiecksmatrix** mit Diagonaleinträgen ungleich null ist, aber $\mathbf{Q} \in \mathbb{R}^{n\times n}$ eine sogenannte **orthogonale Matrix** ist. Die QR-Zerlegung (wenn sie bereits vorliegt) erlaubt es uns, in Analogie zur LR-Zerlegung lineare Gleichungssysteme $\mathbf{A}\,\mathbf{x} = \mathbf{b}$ mit einer invertierbaren Matrix $\mathbf{A} \in \mathbb{R}^{n\times n}$ effizient mit $\mathcal{O}(n^2)$ elementaren Operationen zu lösen. Darüber hinaus ist die QR-Zerlegung ein wichtiges Hilfsmittel, um das lineare Ausgleichsproblem (siehe Teilkapitel 2.5) zu lösen, und sie wird in Teilkapitel 5.4 ein zentrale Rolle im QR-Verfahren zur Eigenwertberechnung spielen.

Definition 2.23. (orthogonale Matrix)
Eine invertierbare Matrix $\mathbf{Q} \in \mathbb{R}^{n\times n}$ *heißt* ***orthogonal****, wenn gilt* $\mathbf{Q}^T = \mathbf{Q}^{-1}$.

Die Aussage $\mathbf{Q}^T = \mathbf{Q}^{-1}$ für eine invertierbare Matrix $\mathbf{Q} \in \mathbb{R}^{n\times n}$ ist äquivalent zu

$$\mathbf{Q}^T\mathbf{Q} = \mathbf{E}_n \qquad \Longleftrightarrow \qquad \mathbf{Q}\,\mathbf{Q}^T = \mathbf{E}_n. \tag{2.27}$$

Seien nun $\mathbf{q}_1, \mathbf{q}_2, \ldots, \mathbf{q}_n \in \mathbb{R}^n$ die Spaltenvektoren von $\mathbf{Q}$. Die erste Gleichung in (2.27) können wir dann auch wie folgt schreiben:

$$\mathbf{q}_j^T\mathbf{q}_k = \delta_{j,k}, \qquad j,k = 1,2,\ldots,n, \tag{2.28}$$

wobei $\delta_{j,k}$ das Kronecker-Delta ist mit $\delta_{j,k} = 1$ für $j = k$ und $\delta_{j,k} = 0$ für $j \neq k$. An (2.28) sehen wir, dass die Spaltenvektoren $\mathbf{q}_1, \mathbf{q}_2, \ldots, \mathbf{q}_n$ von $\mathbf{Q}$ eine **Orthonormalbasis von** $\mathbb{R}^n$ mit dem Euklidischen Skalarprodukt $\mathbf{x}\cdot\mathbf{y} := \mathbf{x}^T\mathbf{y}$ bilden. (Eine Orthonormalbasis von $\mathbb{R}^n$ mit dem Euklidischen Skalarprodukt ist eine Basis aus paarweise orthogonalen Vektoren, die jeweils die Länge 1 haben, also normiert sind.) Weiter folgt in der 2-Norm (oder Euklidischen Norm) $\|\cdot\|_2$ für alle $\mathbf{x} \in \mathbb{R}^n$

$$\|\mathbf{Q}\,\mathbf{x}\|_2^2 = (\mathbf{Q}\,\mathbf{x})^T(\mathbf{Q}\,\mathbf{x}) = \mathbf{x}^T\underbrace{\mathbf{Q}^T\mathbf{Q}}_{=\mathbf{E}_n}\mathbf{x} = \mathbf{x}^T\mathbf{E}_n\,\mathbf{x} = \mathbf{x}^T\mathbf{x} = \|\mathbf{x}\|_2^2, \tag{2.29}$$

d.h. die **Multiplikation eines Vektors** $\mathbf{x} \in \mathbb{R}^n$ **mit einer orthogonalen Matrix Q ändert die 2-Norm des Vektors x (und damit die geometrische Länge von x) nicht**.

Für eine orthogonalen Matrix $\mathbf{Q}$ folgt wegen $\mathbf{Q}^T = \mathbf{Q}^{-1}$, dass

$$(\mathbf{Q}^T)^T = \mathbf{Q} = (\mathbf{Q}^{-1})^{-1} = (\mathbf{Q}^T)^{-1},$$

also $(\mathbf{Q}^T)^T = (\mathbf{Q}^T)^{-1}$, d.h. auch $\mathbf{Q}^T$ **ist orthogonal**.

Weiter ist es wichtig zu wissen, dass das **Produkt orthogonaler Matrizen wieder eine orthogonale Matrix ist**. Wir zeigen dieses hier nur für das Produkt von zwei orthogonalen Matrizen, denn der Beweis für mehr als zwei Matrizen ist analog: Seien also $\mathbf{Q}_1, \mathbf{Q}_2 \in \mathbb{R}^{n\times n}$ zwei orthogonale Matrizen. Dann gelten $\mathbf{Q}_1^T = \mathbf{Q}_1^{-1}$ und $\mathbf{Q}_2^T = \mathbf{Q}_2^{-1}$. Mit $(\mathbf{A}\,\mathbf{B})^T = \mathbf{B}^T\mathbf{A}^T$ für $\mathbf{A}, \mathbf{B} \in \mathbb{R}^{n\times n}$ folgt

$$\big(\mathbf{Q}_1\,\mathbf{Q}_2\big)^T = \mathbf{Q}_2^T\mathbf{Q}_1^T = \mathbf{Q}_2^{-1}\mathbf{Q}_1^{-1} = \big(\mathbf{Q}_1\,\mathbf{Q}_2\big)^{-1},$$

wobei der letzte Schritt aus der nachfolgenden Rechnung klar wird:

$$\big(\mathbf{Q}_2^{-1}\mathbf{Q}_1^{-1}\big)\big(\mathbf{Q}_1\,\mathbf{Q}_2\big) = \mathbf{Q}_2^{-1}\underbrace{\mathbf{Q}_1^{-1}\mathbf{Q}_1}_{=\mathbf{E}_n}\mathbf{Q}_2 = \mathbf{Q}_2^{-1}\underbrace{\mathbf{E}_n\,\mathbf{Q}_2}_{=\mathbf{Q}_2} = \mathbf{Q}_2^{-1}\mathbf{Q}_2 = \mathbf{E}_n,$$

d.h. $\mathbf{Q}_2^{-1}\mathbf{Q}_1^{-1}$ ist die inverse Matrix von $\mathbf{Q}_1\,\mathbf{Q}_2$, also $\mathbf{Q}_2^{-1}\mathbf{Q}_1^{-1} = \big(\mathbf{Q}_1\,\mathbf{Q}_2\big)^{-1}$.

Betrachten wir zunächst zwei Beispiele für orthogonale Matrizen.

Beispiel 2.24. (orthogonale Matrizen und orthogonale Matrizen)
Gegeben seien die folgenden quadratischen Matrizen:

$$\mathbf{Q}_1 = \begin{bmatrix} \cos(\alpha) & \sin(\alpha) \\ \sin(\alpha) & -\cos(\alpha) \end{bmatrix} \quad \text{mit } \alpha \in \mathbb{R} \qquad \text{und} \qquad \mathbf{Q}_2 = \begin{bmatrix} \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} & 0 \\ 0 & 0 & 1 \\ -\frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} & 0 \end{bmatrix}.$$

Die Matrix $\mathbf{Q}_1 \in \mathbb{R}^{2\times 2}$ ist orthogonal, denn wegen

$$\mathbf{Q}_1^T\mathbf{Q}_1 = \begin{bmatrix} \cos(\alpha) & \sin(\alpha) \\ \sin(\alpha) & -\cos(\alpha) \end{bmatrix} \begin{bmatrix} \cos(\alpha) & \sin(\alpha) \\ \sin(\alpha) & -\cos(\alpha) \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} = \mathbf{E}_2$$

gilt $\mathbf{Q}_1^T = \mathbf{Q}_1^{-1}$. Dabei haben wir $\cos^2(\alpha) + \sin^2(\alpha) = 1$ für alle $\alpha \in \mathbb{R}$ genutzt.

Die Matrix $\mathbf{Q}_2 \in \mathbb{C}^{2\times 2}$ ist orthogonal, denn aus

$$\mathbf{Q}_2^T\mathbf{Q}_2 = \begin{bmatrix} \frac{1}{\sqrt{2}} & 0 & -\frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} & 0 & \frac{1}{\sqrt{2}} \\ 0 & 1 & 0 \end{bmatrix} \begin{bmatrix} \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} & 0 \\ 0 & 0 & 1 \\ -\frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} & 0 \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} = \mathbf{E}_3$$

folgt $\mathbf{Q}_2^T = \mathbf{Q}_2^{-1}$.

Wir lernen in Definition 2.25 noch eine für uns ganz wichtige Klasse orthogonaler Matrizen kennen. ♠

Da eine orthogonale Matrix $\mathbf{Q} \in \mathbb{R}^{n\times n}$ nach (2.29) die 2-Norm einer Vektors nicht ändert, finden wir für eine quadratische invertierbare Matrix $\mathbf{A} \in \mathbb{R}^{n\times n}$

$$\begin{aligned} &\|(\mathbf{Q}\,\mathbf{A})\,\mathbf{x}\|_2 = \|\mathbf{Q}\,(\mathbf{A}\,\mathbf{x})\| = \|\mathbf{A}\,\mathbf{x}\|_2 \qquad \text{für alle } \mathbf{x} \in \mathbb{R}^n, \\ &\|(\mathbf{Q}\,\mathbf{A})^{-1}\,\mathbf{x}\|_2 = \|\mathbf{A}^{-1}\,\mathbf{Q}^{-1}\,\mathbf{x}\| = \|\mathbf{A}^{-1}\,\mathbf{Q}^T\mathbf{x}\| \qquad \text{für alle } \mathbf{x} \in \mathbb{R}^n, \end{aligned}$$

wobei wir in der zweiten Zeile $(\mathbf{C}\,\mathbf{B})^{-1} = \mathbf{B}^{-1}\,\mathbf{C}^{-1}$ für invertierbare Matrizen $\mathbf{B}, \mathbf{C} \in \mathbb{R}^{n\times n}$ und $\mathbf{Q}^{-1} = \mathbf{Q}^T$ genutzt haben. Daraus folgen direkt

$$\|\mathbf{Q}\,\mathbf{A}\|_2 = \max_{\substack{\mathbf{x}\in\mathbb{R}^n, \\ \|\mathbf{x}\|_2=1}} \|(\mathbf{Q}\,\mathbf{A})\,\mathbf{x}\|_2 = \max_{\substack{\mathbf{x}\in\mathbb{R}^n, \\ \|\mathbf{x}\|_2=1}} \|\mathbf{A}\,\mathbf{x}\|_2 = \|\mathbf{A}\|_2, \tag{2.30}$$

$$\|(\mathbf{Q}\,\mathbf{A})^{-1}\|_2 = \max_{\substack{\mathbf{x}\in\mathbb{R}^n, \\ \|\mathbf{x}\|_2=1}} \|(\mathbf{Q}\,\mathbf{A})^{-1}\,\mathbf{x}\|_2 = \max_{\substack{\mathbf{x}\in\mathbb{R}^n, \\ \|\mathbf{x}\|_2=1}} \|\mathbf{A}^{-1}\mathbf{Q}^T\mathbf{x}\|_2$$

$$= \max_{\substack{\mathbf{y}\in\mathbb{R}^n,\\ \|\mathbf{Q}\mathbf{y}\|_2=1}} \|\mathbf{A}^{-1}\underbrace{\mathbf{Q}^T\mathbf{Q}}_{=\mathbf{E}_n}\mathbf{y}\|_2 = \max_{\substack{\mathbf{y}\in\mathbb{R}^n,\\ \|\mathbf{y}\|_2=1}} \|\mathbf{A}^{-1}\mathbf{y}\|_2 = \|\mathbf{A}^{-1}\|_2. \qquad (2.31)$$

Dabei haben wir in der dritten Zeile genutzt, dass die orthogonale Matrix eine Bijektion ist und somit $\mathbf{Q}\,\mathbf{y}$ mit $\mathbf{y} \in \mathbb{R}^n$ alle Vektoren in $\mathbb{R}^n$ liefert. Danach wurde lediglich noch $\|\mathbf{Q}\,\mathbf{y}\|_2 = \|\mathbf{y}\|_2$ für alle $\mathbf{y} \in \mathbb{R}^n$ genutzt.

Also folgt aus (2.30) und (2.31), dass

$$\mathrm{cond}_2(\mathbf{Q}\,\mathbf{A}) = \|\mathbf{Q}\,\mathbf{A}\|_2\,\|(\mathbf{Q}\,\mathbf{A})^{-1}\|_2 = \|\mathbf{A}\|_2\,\|\mathbf{A}^{-1}\|_2 = \mathrm{cond}(\mathbf{A}), \qquad (2.32)$$

d.h. die **Konditionszahl bzgl. der Spektralnorm $\|\cdot\|_2$ ändert sich nicht, wenn man eine reguläre Matrix $\mathbf{A} \in \mathbb{R}^{n\times n}$ mit einer orthogonalen Matrix multipliziert**. Unser Ziel ist es daher, eine Abfolge von orthogonalen Matrizen $\mathbf{Q}_1, \mathbf{Q}_2, \ldots, \mathbf{Q}_n \in \mathbb{R}^{n\times n}$ zu finden, so dass

$$\mathbf{Q}_n \cdots \mathbf{Q}_2\,\mathbf{Q}_1\,\mathbf{A} =: \mathbf{R} \qquad (2.33)$$

eine obere Dreiecksmatrix ist. Durch wiederholte Anwendung von (2.32) folgt, dass die obere Dreiecksmatrix $\mathbf{R}$ dann bzgl. der Spektralnorm $\|\cdot\|_2$ die **gleiche Konditionszahl hat wie invertierbare Matrix $\mathbf{A}$**. Nutzt man (2.33) zum Lösen eines linearen Gleichungssystems, so verschlechtert sich die Kondition nicht.

Definition 2.25. (Householder-Matrix)
*Eine **Householder-Matrix** ist jede Matrix in $\mathbb{R}^{n\times n}$ der folgenden Form*

$$\mathbf{H}(\mathbf{w}) := \mathbf{E}_n - 2\,\mathbf{w}\,\mathbf{w}^T, \qquad (2.34)$$

wobei $\mathbf{w} \in \mathbb{R}^n$ die Bedingung $\mathbf{w}^T\mathbf{w} = \|\mathbf{w}\|_2^2 = 1$ erfüllt oder $\mathbf{w} = \mathbf{0}$ ist.

Abbildung 2.1 illustriert, dass die Householder-Matrix $\mathbf{H}(\mathbf{w})$ mit $\mathbf{w} \neq \mathbf{0}$ eine **Spiegelung an der Hyperebene**

$$S_{\mathbf{w}} = \left\{\mathbf{z} \in \mathbb{R}^n \,:\, \mathbf{w}^T\mathbf{z} = 0\right\}$$

durch der Ursprung ist, die senkrecht (also orthogonal) zu w ist. Dieses zeigt man wie folgt: Sei $\mathbf{a} \in \mathbb{R}^n$. Wir zerlegen $\mathbf{a}$ additiv in die Komponente $(\mathbf{w}^T\mathbf{a})\,\mathbf{w}$ von $\mathbf{a}$ in Richtung von $\mathbf{w}$ und die dazu orthogonale Komponente $\mathbf{a} - (\mathbf{w}^T\mathbf{a})\,\mathbf{w}$ (welche in $S_{\mathbf{w}}$ liegt):

$$\mathbf{a} = (\mathbf{w}^T\mathbf{a})\,\mathbf{w} + \big(\mathbf{a} - (\mathbf{w}^T\mathbf{a})\,\mathbf{w}\big). \qquad (2.35)$$

(In der Tat gilt mit $\mathbf{w}^T\mathbf{w} = 1$ dann $\mathbf{w}^T\big(\mathbf{a} - (\mathbf{w}^T\mathbf{a})\,\mathbf{w}\big) = \mathbf{w}^T\mathbf{a} - \mathbf{w}^T\mathbf{a}\,(\mathbf{w}^T\mathbf{w}\big) = 0$, d.h. $\mathbf{a} - (\mathbf{w}^T\mathbf{a})\,\mathbf{w}$ liegt in $S_{\mathbf{w}}$.)

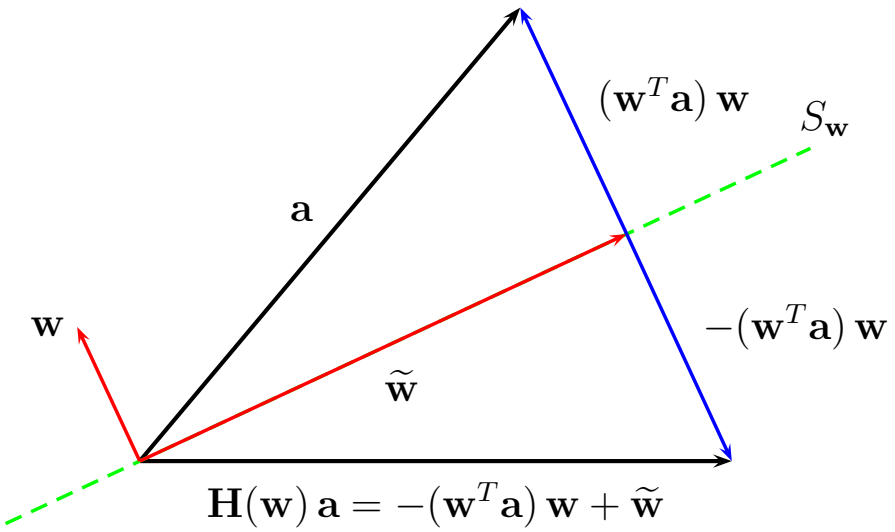


Abb. 2.1: Spiegelung mit Householder-Matrix: Für $\mathbf{a} \in \mathbb{R}^n$ gilt $\mathbf{a} = (\mathbf{w}^T\mathbf{a})\,\mathbf{w} + \widetilde{\mathbf{w}}$ mit $\widetilde{\mathbf{w}} = \mathbf{a} - (\mathbf{w}^T\mathbf{a})\,\mathbf{w}$. Der Vektor $(\mathbf{w}^T\mathbf{a})\,\mathbf{w}$ ist die zu $S_{\mathbf{w}}$ senkrechte Komponente von $\mathbf{a}$, und $\widetilde{\mathbf{w}} = \mathbf{a} - (\mathbf{w}^T\mathbf{a})\,\mathbf{w}$ ist die Komponente von $\mathbf{a}$ in $S_{\mathbf{w}}$. Nach (2.36) gilt dann $\mathbf{H}(\mathbf{w})\,\mathbf{a} = -(\mathbf{w}^T\mathbf{a})\,\mathbf{w} + \widetilde{\mathbf{w}}$.

Wenn wir den Vektor $\mathbf{a}$ mit $\mathbf{H}(\mathbf{w})$ multiplizieren, so folgt mit (2.35) und $\mathbf{w}^T\mathbf{w} = 1$

$$\begin{aligned}\mathbf{H}(\mathbf{w})\,\mathbf{a} &= (\mathbf{E}_n - 2\,\mathbf{w}\,\mathbf{w}^T)\,\mathbf{a} = \mathbf{a} - 2\,\mathbf{w}\,\mathbf{w}^T\mathbf{a} \\ &= \underbrace{(\mathbf{w}^T\mathbf{a})\,\mathbf{w} + \big(\mathbf{a} - (\mathbf{w}^T\mathbf{a})\,\mathbf{w}\big)}_{=\,\mathbf{a} \text{ nach } (2.35)} - 2\,\mathbf{w}\,(\mathbf{w}^T\mathbf{a}) \\ &= -(\mathbf{w}^T\mathbf{a})\,\mathbf{w} + \big(\mathbf{a} - (\mathbf{w}^T\mathbf{a})\,\mathbf{w}\big). \qquad (2.36)\end{aligned}$$

An der Darstellung in der letzten Zeile sieht man, dass $\mathbf{H}(\mathbf{w})\,\mathbf{a}$ tatsächlich die Spiegelung von $\mathbf{a}$ an der Hyperebene $S_{\mathbf{w}}$ ist, denn nur die zu $S_{\mathbf{w}}$ senkrechte Komponente hat ihre Vorzeichen geändert (vgl. Abbildung 2.1).

Beispiel 2.26. (Householder-Matrix)
Sei $\mathbf{w} = \big[0; -\frac{3}{5}; \frac{4}{5}\big]^T$. Dann gilt $\|\mathbf{w}\|_2 = 1$, und die Matrix

$$\mathbf{H}(\mathbf{w}) = \mathbf{E}_n - 2\begin{bmatrix} 0 \\ -\frac{3}{5} \\ \frac{4}{5} \end{bmatrix}\Big[0; -\frac{3}{5}; \frac{4}{5}\Big] = \mathbf{E}_n - 2\begin{bmatrix} 0 & 0 & 0 \\ 0 & \frac{9}{25} & -\frac{12}{25} \\ 0 & -\frac{12}{25} & \frac{16}{25} \end{bmatrix}$$

$$= \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} - \begin{bmatrix} 0 & 0 & 0 \\ 0 & \frac{18}{25} & -\frac{24}{25} \\ 0 & -\frac{24}{25} & \frac{32}{25} \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & \frac{7}{25} & \frac{24}{25} \\ 0 & \frac{24}{25} & -\frac{7}{25} \end{bmatrix}$$

ist eine 3×3 Householder-Matrix. ♠

Der nächste Satz stellt wichtige Eigenschaften von Householder-Matrizen zusammen, deren Beweis wir auf einem Übungsblatt herleiten.

Hilfssatz 2.27. (Eigenschaften von Householder-Matrizen)
Eine Householder-Matrix $\mathbf{H}(\mathbf{w})$ *mit* $\mathbf{w} \in \mathbb{R}^n$ *hat die folgenden Eigenschaften:*

(1) $\mathbf{H}(\mathbf{w})$ *ist* ***symmetrisch****, d.h.* $\big(\mathbf{H}(\mathbf{w})\big)^T = \mathbf{H}(\mathbf{w})$.

(2) $\mathbf{H}(\mathbf{w})$ *ist* ***invertierbar*** *(d.h.* ***regulär****).*

(3) $\det\big(\mathbf{H}(\mathbf{w})\big) = -1$ *für jeden Vektor* $\mathbf{w} \neq \mathbf{0}$.

(4) $\mathbf{H}(\mathbf{w})$ *ist* ***orthogonal****, d.h.* $\big(\mathbf{H}(\mathbf{w})\big)^T = \big(\mathbf{H}(\mathbf{w})\big)^{-1}$.

Wir bemerken noch, dass die Abspeicherung von $\mathbf{H}(\mathbf{w})$ nur die Abspeicherung der n Elemente von $\mathbf{w} \in \mathbb{R}^n$ erfordert.

Beweis von Hilfssatz 2.27:

(1) Unter Ausnutzung der Rechenregeln $(\mathbf{A}+\mathbf{B})^T = \mathbf{A}^T + \mathbf{B}^T$, $(\mathbf{A}\,\mathbf{B})^T = \mathbf{B}^T\mathbf{A}^T$ und $(\mathbf{A}^T)^T = \mathbf{A}$ für transponierten Matrizen finden wir

$$\begin{aligned}\big(\mathbf{H}(\mathbf{w})\big)^T &= \big(\mathbf{E}_n - 2\,\mathbf{w}\,\mathbf{w}^T\big)^T = \mathbf{E}_n^T - 2\,(\mathbf{w}\,\mathbf{w}^T)^T \\ &= \mathbf{E}_n - 2\,(\mathbf{w}^T)^T\,\mathbf{w}^T = \mathbf{E}_n - 2\,\mathbf{w}\,\mathbf{w}^T = \mathbf{H}(\mathbf{w}).\end{aligned}$$

Also ist $\mathbf{H}(\mathbf{w})$ symmetrisch.

(4) Um zu zeigen, dass $\mathbf{H}(\mathbf{w})$ orthogonal ist, müssen wir

$$\big(\mathbf{H}(\mathbf{w})\big)^T\,\mathbf{H}(\mathbf{w}) = \mathbf{E}_n$$

nachweisen. Da nach (1) $\mathbf{H}(\mathbf{w})^T = \mathbf{H}(\mathbf{w})$ ist folgt mit $\mathbf{H}(\mathbf{w}) = \mathbf{E}_n - 2\,\mathbf{w}\,\mathbf{w}^T$

$$\begin{aligned}\mathbf{H}(\mathbf{w})^T\,\mathbf{H}(\mathbf{w}) &= \mathbf{H}(\mathbf{w})\,\mathbf{H}(\mathbf{w}) = \big(\mathbf{E}_n - 2\,\mathbf{w}\,\mathbf{w}^T\big)\,\big(\mathbf{E}_n - 2\,\mathbf{w}\,\mathbf{w}^T\big) \\ &= \underbrace{\mathbf{E}_n\,\mathbf{E}_n}_{=\mathbf{E}_n} - 2\,\mathbf{w}\,\underbrace{\mathbf{w}^T\,\mathbf{E}_n}_{=\mathbf{w}^T} - 2\,\underbrace{\mathbf{E}_n\,\mathbf{w}}_{=\mathbf{w}}\,\mathbf{w}^T + 4\,(\mathbf{w}\,\mathbf{w}^T)\,(\mathbf{w}\,\mathbf{w}^T) \\ &= \mathbf{E}_n - 4\,\mathbf{w}\,\mathbf{w}^T + 4\,(\mathbf{w}\,\mathbf{w}^T)\,(\mathbf{w}\,\mathbf{w}^T) \\ &= \mathbf{E}_n - 4\,\mathbf{w}\,\mathbf{w}^T + 4\,\mathbf{w}\,\underbrace{(\mathbf{w}^T\,\mathbf{w})}_{=1}\,\mathbf{w}^T \\ &= \mathbf{E}_n - 4\,\mathbf{w}\,\mathbf{w}^T + 4\,\mathbf{w}\,\mathbf{w}^T = \mathbf{E}_n.\end{aligned}$$

Also ist $\mathbf{H}(\mathbf{w})$ orthogonal.

Nachweis von (2) für $\mathbf{w} = \mathbf{0}$: Ist $\mathbf{w} = \mathbf{0}$, so folgt $\mathbf{H}(\mathbf{w}) = \mathbf{E}_n$, und $\mathbf{E}_n$ ist invertierbar.

Vorbereitung für den Nachweis von (2) und (3) für $\mathbf{w} \neq \mathbf{0}$: Sei $\mathbf{w} \neq \mathbf{0}$. Dann gilt

$$\mathbf{H}(\mathbf{w})\,\mathbf{w} = \big(\mathbf{E}_n - 2\,\mathbf{w}\,\mathbf{w}^T\big)\,\mathbf{w} = \underbrace{\mathbf{E}_n\,\mathbf{w}}_{=\mathbf{w}} - 2\,\mathbf{w}\,\underbrace{(\mathbf{w}^T\,\mathbf{w})}_{=1} = \mathbf{w} - 2\,\mathbf{w} = -\mathbf{w},$$

d.h. $\mathbf{w}$ ist ein Eigenvektor zum Eigenwert $\lambda_1 = -1$. Sei $\mathbf{b}_1 := \mathbf{w}$.

Sei nun $\mathbf{z}$ ein Vektor aus der Hyperebene $S_{\mathbf{w}} = \{\mathbf{z} \in \mathbb{R}^n \ : \ \mathbf{w}^T\,\mathbf{z} = 0\}$, welche auf $\mathbf{w}$ senkrecht steht und welche ein $(n-1)$-dimensionaler Unterraum von $\mathbb{R}^n$ ist. Für $\mathbf{z} \in S_{\mathbf{w}}$ gilt

$$\mathbf{H}(\mathbf{w})\,\mathbf{z} = \big(\mathbf{E}_n - 2\,\mathbf{w}\,\mathbf{w}^T\big)\,\mathbf{z} = \mathbf{z} - 2\,\mathbf{w}\,\underbrace{(\mathbf{w}^T\,\mathbf{z})}_{=0} = \mathbf{z},$$

d.h. jedes $\mathbf{z} \in S_{\mathbf{w}}$ ist ein Eigenvektor zum Eigenwert $\lambda_2 = 1$. Da $S_{\mathbf{w}}$ ein $(n-1)$-dimensionaler Unterraum von $\mathbb{R}^n$ ist, können wir eine Orthonormalbasis $\mathbf{b}_2, \ldots, \mathbf{b}_n$ von $S_{\mathbf{w}}$ wählen. Diese besteht automatisch aus Eigenvektoren von $\mathbf{H}(\mathbf{w})$ zum Eigenwert $\lambda_2 = 1$.

Damit ist $\mathbf{b}_1, \mathbf{b}_2, \ldots, \mathbf{b}_n$ eine Orthonormalbasis von $\mathbb{R}^n$, und es gilt somit für die $n \times n$ Matrix $\mathbf{B} = \big[\mathbf{b}_1, \mathbf{b}_2, \ldots \mathbf{b}_n\big]$ mit den Spaltenvektoren $\mathbf{b}_1, \mathbf{b}_2, \ldots \mathbf{b}_n$, dass $\mathbf{B}$ orthogonal ist, da wegen $(\mathbf{B}^T\,\mathbf{B})_{j,k} = \mathbf{b}_j^T\,\mathbf{b}_k = \delta_{j,k}$ für $j, k = 1, 2, \ldots, n$ die Gleichheit $\mathbf{B}^T\,\mathbf{B} = \mathbf{E}_n$ gilt. Weiter folgt

$$\mathbf{B}^T\,\mathbf{H}(\mathbf{w})\,\mathbf{B} = \begin{bmatrix} \mathbf{b}_1^T \\ \mathbf{b}_2^T \\ \vdots \\ \mathbf{b}_n^T \end{bmatrix} \cdot \big[-\mathbf{b}_1,\ \mathbf{b}_2,\ \ldots,\ \mathbf{b}_n\big] = \begin{bmatrix} -1 & 0 & \cdots & \cdots & 0 \\ 0 & 1 & 0 & \cdots & 0 \\ \vdots & 0 & 1 & \ddots & \vdots \\ \vdots & \vdots & \ddots & \ddots & 0 \\ 0 & 0 & \cdots & 0 & 1 \end{bmatrix}, \quad (2.37)$$

wobei wir zunächst $\mathbf{H}(\mathbf{w})\,\mathbf{b}_1 = -\mathbf{b}_1$ und $\mathbf{H}(\mathbf{w})\,\mathbf{b}_j = \mathbf{b}_j$ für alle $j = 2, \ldots, n$ und danach $\mathbf{b}_j^T\,\mathbf{b}_k = \delta_{j,k}$, $j, k = 1, 2, \ldots, n$ (wegen der Tatsache, dass $\mathbf{b}_1, \mathbf{b}_2, \ldots, \mathbf{b}_n$ eine Orthonormalbasis von $\mathbb{R}^n$ ist) ausgenutzt haben.

(3) Sei $\mathbf{w} \neq \mathbf{0}$. Aus (2.37) folgt, weil $\mathbf{B}$ orthogonal ist und somit $\mathbf{B}^T = \mathbf{B}^{-1}$ gilt, dass

$$\mathbf{B}^{-1}\,\mathbf{H}(\mathbf{w})\,\mathbf{B} = \mathbf{B}^T\,\mathbf{H}(\mathbf{w})\,\mathbf{B} = \begin{bmatrix} -1 & 0 & \cdots & \cdots & 0 \\ 0 & 1 & 0 & \cdots & 0 \\ \vdots & 0 & 1 & \ddots & \vdots \\ \vdots & \vdots & \ddots & \ddots & 0 \\ 0 & 0 & \cdots & 0 & 1 \end{bmatrix}. \quad (2.38)$$

Mit Hilfe von $\det(\mathbf{A}\,\mathbf{B}) = \det(\mathbf{A})\,\det(\mathbf{B})$ für alle $\mathbf{A}, \mathbf{B} \in \mathbb{R}^{n\times n}$ folgt

$$\det\big(\mathbf{B}^{-1}\,\mathbf{H}(\mathbf{w})\,\mathbf{B}\big) = \underbrace{\det(\mathbf{B}^{-1})}_{=(\det(\mathbf{B}))^{-1}}\,\det\big(\mathbf{H}(\mathbf{w})\big)\,\det(\mathbf{B}) = \det\big(\mathbf{H}(\mathbf{w})\big),$$

wobei wir $\det(\mathbf{B}^{-1}) = \big(\det(\mathbf{B})\big)^{-1}$ genutzt haben. Andererseits erhalten wir durch Berechnen der Determinante der Diagonalmatrix auf der rechten Seite von (2.38)

$$\det\big(\mathbf{H}(\mathbf{w})\big) = \det\big(\mathbf{B}^{-1}\,\mathbf{H}(\mathbf{w})\,\mathbf{B}\big) = (-1)\cdot 1^{n-1} = -1.$$

(2) Ist $\mathbf{w} \neq \mathbf{0}$, so folgt aus $\det\big(\mathbf{H}(\mathbf{w})\big) = -1 \neq 0$ direkt, dass $\mathbf{H}(\mathbf{w})$ invertierbar ist. □

Hilfssatz 2.28. (Konstruktion von Householder-Matrizen)
Seien $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$ *Vektoren mit* $\mathbf{x} \neq \mathbf{y}$ *und* $\|\mathbf{x}\|_2 = \|\mathbf{y}\|_2$. *Dann bildet die Householder-Matrix* $\mathbf{H}(\mathbf{w}) = \mathbf{E}_n - 2\,\mathbf{w}\,\mathbf{w}^T$ *mit* $\mathbf{w} = \frac{\mathbf{x}-\mathbf{y}}{\|\mathbf{x}-\mathbf{y}\|_2}$ *die Vektoren* $\mathbf{x}$ *und* $\mathbf{y}$ *aufeinander ab, d.h. es gilt* $\mathbf{H}(\mathbf{w})\,\mathbf{x} = \mathbf{y}$ *und* $\mathbf{H}(\mathbf{w})\,\mathbf{y} = \mathbf{x}$.

Beweis von Hilfssatz 2.28: Die Matrix $\mathbf{H}(\mathbf{w}) = \mathbf{E}_n - 2\,\mathbf{w}\,\mathbf{w}^T$ ist in der Tat eine Householder-Matrix, da der Vektor

$$\mathbf{w} = \frac{\mathbf{x}-\mathbf{y}}{\|\mathbf{x}-\mathbf{y}\|_2} \tag{2.39}$$

per Definition $\|\mathbf{w}\|_2 = \mathbf{w}^T\mathbf{w} = 1$ erfüllt. Mit (2.39) folgt dann

$$\mathbf{H}(\mathbf{w})\,\mathbf{x} = \big(\mathbf{E}_n - 2\,\mathbf{w}\,\mathbf{w}^T\big)\,\mathbf{x} = \mathbf{x} - (\mathbf{x}-\mathbf{y})\,\frac{2\,(\mathbf{x}-\mathbf{y})^T\,\mathbf{x}}{\|\mathbf{x}-\mathbf{y}\|_2^2}\,. \tag{2.40}$$

Mit $\mathbf{x}^T\mathbf{x} = \|\mathbf{x}\|_2^2 = \|\mathbf{y}\|_2^2 = \mathbf{y}^T\mathbf{y}$ und $\mathbf{x}^T\mathbf{y} = \mathbf{y}^T\mathbf{x}$ (weil $\mathbf{x}^T\mathbf{y}$ reell ist und somit $\mathbf{x}^T\mathbf{y} = (\mathbf{x}^T\mathbf{y})^T = \mathbf{y}^T\mathbf{x}$ gilt) folgt für den Nenner des Bruchs in (2.40)

$$\begin{aligned} 2\,(\mathbf{x}-\mathbf{y})^T\,\mathbf{x} &= 2\,\mathbf{x}^T\mathbf{x} - 2\,\mathbf{y}^T\mathbf{x} = \big(\mathbf{x}^T\mathbf{x} + \mathbf{y}^T\mathbf{y}\big) - \big(\mathbf{y}^T\mathbf{x} + \mathbf{x}^T\mathbf{y}\big) \\ &= \mathbf{x}^T\mathbf{x} - \mathbf{x}^T\mathbf{y} + \mathbf{y}^T\mathbf{y} - \mathbf{y}^T\mathbf{x} = \mathbf{x}^T(\mathbf{x}-\mathbf{y}) + \mathbf{y}^T(\mathbf{y}-\mathbf{x}) \\ &= \mathbf{x}^T(\mathbf{x}-\mathbf{y}) - \mathbf{y}^T(\mathbf{x}-\mathbf{y}) = (\mathbf{x}-\mathbf{y})^T\,(\mathbf{x}-\mathbf{y}) = \|\mathbf{x}-\mathbf{y}\|_2^2. \end{aligned} \tag{2.41}$$

Einsetzen von (2.41) in (2.40) und Vereinfachen liefert $\mathbf{H}(\mathbf{w})\,\mathbf{x} = \mathbf{y}$, denn

$$\mathbf{H}(\mathbf{w})\,\mathbf{x} = \mathbf{x} - (\mathbf{x}-\mathbf{y})\,\frac{\|\mathbf{x}-\mathbf{y}\|_2^2}{\|\mathbf{x}-\mathbf{y}\|_2^2} = \mathbf{x} - (\mathbf{x}-\mathbf{y})\cdot 1 = \mathbf{x} - \mathbf{x} + \mathbf{y} = \mathbf{y}.$$

Mit $\mathbf{H}(\mathbf{w}) = (\mathbf{H}(\mathbf{w}))^T = (\mathbf{H}(\mathbf{w}))^{-1}$ (siehe (1) und (4) in Hilfssatz 2.27) und $\mathbf{H}(\mathbf{w})\,\mathbf{x} = \mathbf{y}$ folgt nun

$$\mathbf{H}(\mathbf{w})\,\mathbf{y} = \mathbf{H}(\mathbf{w})\,\big(\mathbf{H}(\mathbf{w})\,\mathbf{x}\big) = \mathbf{H}(\mathbf{w})\,\mathbf{H}(\mathbf{w})\,\mathbf{x} = \big(\mathbf{H}(\mathbf{w})\big)^{-1}\,\mathbf{H}(\mathbf{w})\,\mathbf{x} = \mathbf{E}_n\,\mathbf{x} = \mathbf{x}.$$

Damit ist der Hilfssatz bewiesen. □

Im Folgenden werden wir Hilfssatz 2.28 nutzen, um einen Vektor $\mathbf{x} = [x_1; \ldots; x_n]^T \in \mathbb{R}^n \setminus \{\mathbf{0}\}$ auf ein skalares Vielfaches von $\mathbf{e}_1 = [1; 0; \ldots; 0]^T$ abzubilden. Wir suchen also eine Householder-Matrix $\mathbf{H}(\mathbf{w})$ mit $\mathbf{H}(\mathbf{w})\,\mathbf{x} = c\,\mathbf{e}_1$ mit einem geeigneten $c \in \mathbb{R} \setminus \{0\}$. Die Bedingung $\|\mathbf{x}\|_2 = \|c\,\mathbf{e}_1\| = |c|\,\|\mathbf{e}_1\|_2 = |c|$ in Hilfssatz 2.28 liefert $c = \|\mathbf{x}\|_2$ oder $c = -\|\mathbf{x}\|_2$. Wir treffen die erste Wahl $c = \|\mathbf{x}\|_2$ und bekommen als Vektor $\mathbf{w} \in \mathbb{R}^n$ der Householder-Matrix nach Hilfssatz 2.28 (mit $\mathbf{x} = \mathbf{x}$ und $\mathbf{y} = \|\mathbf{x}\|_2\,\mathbf{e}_1$)

$$\mathbf{w} := \frac{\mathbf{x} - \|\mathbf{x}\|_2\,\mathbf{e}_1}{\big\|\mathbf{x} - \|\mathbf{x}\|_2\,\mathbf{e}_1\big\|_2}\,. \tag{2.42}$$

Die Householder-Matrix $\mathbf{H}(\mathbf{w})$ mit dem in (2.42) gegebenen Vektor $\mathbf{w}$ erfüllt dann $\mathbf{H}(\mathbf{w})\,\mathbf{x} = \|\mathbf{x}\|_2\,\mathbf{e}_1$ (und $\mathbf{H}(\mathbf{w})\,(\|\mathbf{x}\|_2\,\mathbf{e}_1) = \mathbf{x}$).

Wir werden nun **Householder-Matrizen nutzen, um eine invertierbare Matriz $\mathbf{A} \in \mathbb{R}^{n\times n}$ in eine obere Dreiecksmatrix zu transformieren**, ohne dass sich die Konditionszahl ändert.

Sei als $\mathbf{A} \in \mathbb{R}^{n\times n}$ eine invertierbare Matrix. Dann ist der erste Spaltenvektor $\mathbf{x}$ von $\mathbf{A}$ ungleich dem Nullvektor.

Wir wählen nun eine Householder-Matrix $\mathbf{H}_1 = \mathbf{H}(\mathbf{w}_1) \in \mathbb{R}^{n\times n}$, die $\mathbf{x}$ auf $\|\mathbf{x}\|_2\,\mathbf{e}_1$ mit $\mathbf{e}_1 = [1; 0; \ldots; 0]^T \in \mathbb{R}^n$ abbildet. Dann gilt $\mathbf{H}(\mathbf{w}_1)\,\mathbf{x} = \|\mathbf{x}\|_2\,\mathbf{e}_1$, und es folgt

$$\mathbf{H}_1\,\mathbf{A} = \begin{bmatrix} \alpha_1 & * & \cdots & * \\ 0 & & & \\ \vdots & & \mathbf{A}_1 & \\ 0 & & & \end{bmatrix} \qquad \text{mit einer Matrix } \mathbf{A}_1 \in \mathbb{R}^{(n-1)\times(n-1)}.$$

Da die Matrix $\mathbf{H}_1\,\mathbf{A}$ (wegen der Invertierbarkeit von $\mathbf{A}$ und $\mathbf{H}_1$) ebenfalls invertierbar ist, kann der erste Spaltenvektor $\widetilde{\mathbf{x}}$ von $\mathbf{A}_1$ nicht gleich dem Nullvektor sein. Wir wählen nun eine Housholder-Matrix $\widetilde{\mathbf{H}}_2 = \mathbf{H}(\mathbf{w}_2) \in \mathbb{R}^{(n-1)\times(n-1)}$ (mit $\mathbf{w}_2 \in \mathbb{R}^{n-1}$) mit $\widetilde{\mathbf{H}}_2\,\widetilde{\mathbf{x}} = \|\widetilde{\mathbf{x}}\|_2\,\mathbf{e}_1$, wobei nun $\mathbf{e}_1 = [1; 0; \ldots; 0]^T \in \mathbb{R}^{n-1}$. Dann gilt

$$\widetilde{\mathbf{H}}_2\,\mathbf{A}_1 = \begin{bmatrix} \alpha_2 & * & \cdots & * \\ 0 & & & \\ \vdots & & \mathbf{A}_2 & \\ 0 & & & \end{bmatrix} \qquad \text{mit einer Matrix } \mathbf{A}_2 \in \mathbb{R}^{(n-2)\times(n-2)}.$$

Wir betten $\widetilde{\mathbf{H}}_2$ nun passend in eine $n \times n$-Matrix ein: Die Matrix

$$\mathbf{H}_2 = \begin{bmatrix} 1 & 0 & \cdots & 0 \\ 0 & & & \\ \vdots & & \widetilde{\mathbf{H}}_2 & \\ 0 & & & \end{bmatrix} \in \mathbb{R}^{n \times n}$$

ist dann auch symmetrisch und orthogonal, da $\mathbf{H}_2^T \, \mathbf{H}_2 = \mathbf{E}_n$ aus $\widetilde{\mathbf{H}}_2^T \, \widetilde{\mathbf{H}}_2 = \mathbf{E}_{n-1}$ folgt. (Man kann zeigen, dass $\mathbf{H}_2$ eine $n \times n$ Householder-Matrix ist.) Es gilt dann

$$\mathbf{H}_2 \, \mathbf{H}_1 \, \mathbf{A} = \mathbf{H}_2 \begin{bmatrix} \alpha_1 & * & \cdots & * \\ 0 & & & \\ \vdots & & \mathbf{A}_1 & \\ 0 & & & \end{bmatrix} = \begin{bmatrix} \alpha_1 & * & * & \cdots & * \\ 0 & \alpha_2 & * & \cdots & * \\ 0 & 0 & & & \\ \vdots & \vdots & & \mathbf{A}_2 & \\ 0 & 0 & & & \end{bmatrix}.$$

Wir können diesen Prozess fortsetzen und erhalten nach $n-1$ Schritten

$$\mathbf{H}_{n-1} \, \mathbf{H}_{n-2} \cdots \mathbf{H}_2 \, \mathbf{H}_1 \, \mathbf{A} = \mathbf{R}, \tag{2.43}$$

wobei $\mathbf{R}$ eine obere Dreiecksmatrix ist, deren Diagonaleinträge alle ungleich null sind. Nach Konstruktion sind die Matrizen $\mathbf{H}_1, \mathbf{H}_2, \ldots, \mathbf{H}_{n-1}$ alle orthogonal. Da ein Produkt orthogonaler Matrizen wieder orthogonal ist (siehe Übungszettel) folgt, dass $\widetilde{\mathbf{Q}} := \mathbf{H}_{n-1} \, \mathbf{H}_{n-2} \cdots \mathbf{H}_2 \, \mathbf{H}_1$ orthogonal und damit insbesondere invertierbar ist. Die Inverse $\widetilde{\mathbf{Q}}^{-1}$ ist ebenfalls orthogonal. Also erhalten wir aus (2.43)

$$\widetilde{\mathbf{Q}} \, \mathbf{A} = \mathbf{R} \quad \Longleftrightarrow \quad \mathbf{A} = \widetilde{\mathbf{Q}}^{-1} \, \mathbf{R} \quad \text{mit } \widetilde{\mathbf{Q}}^{-1} = \mathbf{H}_1^{-1} \, \mathbf{H}_2^{-1} \cdots \mathbf{H}_{n-1}^{-1}, \tag{2.44}$$

wobei wir $\widetilde{\mathbf{Q}}^{-1} = \big(\mathbf{H}_{n-1} \, \mathbf{H}_{n-2} \cdots \mathbf{H}_2 \, \mathbf{H}_1\big)^{-1} = \mathbf{H}_1^{-1} \, \mathbf{H}_2^{-1} \cdots \mathbf{H}_{n-1}^{-1}$ wegen der Rechenregel $(\mathbf{B} \, \mathbf{C})^{-1} = \mathbf{C}^{-1} \mathbf{B}^{-1}$ für invertierbare Matrizen $\mathbf{B}, \mathbf{C} \in \mathbb{R}^{n \times n}$ bekommen. Da die Householder-Matrizen $\mathbf{H}_1, \mathbf{H}_2, \ldots, \mathbf{H}_{n-1}$ alle orthogonal und symmetrisch sind, gilt $\mathbf{H}_k^{-1} = \mathbf{H}_k^T = \mathbf{H}_k$ für alle $k = 1, 2, \ldots, n-1$. Damit vereinfacht sich $\widetilde{\mathbf{Q}}^{-1}$ zu $\widetilde{\mathbf{Q}}^{-1} = \mathbf{H}_1 \, \mathbf{H}_2 \cdots \mathbf{H}_{n-1}$. Also folgt aus (2.44)

$$\boxed{\mathbf{A} = \mathbf{Q} \, \mathbf{R} \qquad \text{mit} \qquad \mathbf{Q} := \widetilde{\mathbf{Q}}^{-1} = \mathbf{H}_1 \, \mathbf{H}_2 \cdots \mathbf{H}_{n-1},} \tag{2.45}$$

wobei $\mathbf{Q}$ orthogonal und $\mathbf{R}$ eine obere Dreiecksmatrix ist, deren Diagonaleinträge alle ungleich null sind. Die „Zerlegung“ $\mathbf{A} = \mathbf{Q} \, \mathbf{R}$ in (2.45) nennt sich die **QR-Zerlegung von A**. Aufgrund unsere Überlegungen in (2.32) ist klar, dass $\mathbf{R}$ die gleiche Kondition wir $\mathbf{A}$ hat, also $\text{cond}_2(\mathbf{R}) = \text{cond}_2(\mathbf{A})$.

Wir halten das hergeleitete Resultat als Satz fest.

Satz 2.29. (QR-Zerlegung)
Jede reguläre Matrix $\mathbf{A} \in \mathbb{R}^{n \times n}$ *hat eine* ***QR-Zerlegung*** $\mathbf{A} = \mathbf{Q}\,\mathbf{R}$ *mit einer* ***orthogonalen Matrix*** $\mathbf{Q}$ *und einer* ***oberen Dreiecksmatrix*** $\mathbf{R}$.

Die Multiplikation einer Householder-Matrix $\mathbf{H}(\mathbf{w}) = \mathbf{E}_n - 2\,\mathbf{w}\,\mathbf{w}^T$ mit einem Vektor benötigt aufgrund der speziellen Form von $\mathbf{H}(\mathbf{w})$ nur $\mathcal{O}(n)$ elementare Rechenoperationen. Daher erfordert jeder Schritt bei der Herleitung von $\mathbf{R}$ (also jede Matrix-Matrix-Multiplikation mit einer Matrix $\mathbf{H}_k$ in (2.43)) $\mathcal{O}(n^2)$ elementare Operationen. Da wir $n-1$ Schritte haben, werden bei der Berechnung der QR-Zerlegung insgesamt $\mathcal{O}(n^3)$ **elementare Operationen** benötigt.

Bemerkung 2.30. (Lösen eines LGS mit regulärer Matrix mit Hilfe der QR-Zerlegung)
Analog zur LR-Zerlegung können wir nun auch die QR-Zerlegung nutzen, um ein lineares Gleichungssystem $\mathbf{A}\,\mathbf{x} = \mathbf{b}$ mit einer invertierbaren Matrix $\mathbf{A} \in \mathbb{R}^{n \times n}$ zu lösen. Ist $\mathbf{A} = \mathbf{Q}\,\mathbf{R}$ die QR-Zerlegung von $\mathbf{A}$, so gilt (weil $\mathbf{Q}^{-1} = \mathbf{Q}^T$, da $\mathbf{Q}$ orthogonal ist)

$$\mathbf{A}\,\mathbf{x} = \mathbf{b} \qquad \Longleftrightarrow \qquad \mathbf{Q}\,\mathbf{R}\,\mathbf{x} = \mathbf{b} \qquad \Longleftrightarrow \qquad \mathbf{R}\,\mathbf{x} = \mathbf{Q}^T\mathbf{b},$$

und $\mathbf{R}\,\mathbf{x} = \mathbf{Q}^T\mathbf{b}$ kann mit Rückwärtsrechnen mit $\mathcal{O}(n^2)$ elementaren Rechenoperationen gelöst werden.

Betrachten wir ein Beispiel.

Beispiel 2.31. (QR-Zerlegung)
Gesucht ist die QR-Zerlegung der Matrix $\mathbf{A} = \begin{bmatrix} 2 & -3 & 3 \\ -2 & 6 & 6 \\ 1 & 0 & 3 \end{bmatrix}$.

Da für den ersten Spaltenvektor $\mathbf{x} = \begin{bmatrix} 2 \\ -2 \\ 1 \end{bmatrix}$ of $\mathbf{A}$ gilt $\|\mathbf{x}\|_2 = 3$, wählen wir

$$\mathbf{w}_1 := \frac{\mathbf{z}}{\|\mathbf{z}\|_2} \qquad \text{mit} \qquad \mathbf{z} := \begin{bmatrix} 2 \\ -2 \\ 1 \end{bmatrix} - 3 \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix} = \begin{bmatrix} -1 \\ -2 \\ 1 \end{bmatrix}, \qquad \|\mathbf{z}\|_2 = \sqrt{6}.$$

Also gilt

$$\mathbf{w}_1 = \frac{1}{\sqrt{6}} \begin{bmatrix} -1 \\ -2 \\ 1 \end{bmatrix},$$

und die erste Householder-Matrix $\mathbf{H}_1$ (welche $\mathbf{H}_1\,\mathbf{x} = 3\,\mathbf{e}_1$ erfüllt) ist

$$\mathbf{H}_1 := \mathbf{H}(\mathbf{w}_1) = \mathbf{E}_3 - 2\,\mathbf{w}_1\,\mathbf{w}_1^T = \mathbf{E}_3 - \frac{2}{\left(\sqrt{6}\right)^2}\begin{bmatrix}-1\\-2\\1\end{bmatrix}\,\big[-1;-2;1\big]$$

$$= \mathbf{E}_3 - \frac{1}{3}\begin{bmatrix}1 & 2 & -1\\2 & 4 & -2\\-1 & -2 & 1\end{bmatrix} = \begin{bmatrix}1 & 0 & 0\\0 & 1 & 0\\0 & 0 & 1\end{bmatrix} - \begin{bmatrix}\frac{1}{3} & \frac{2}{3} & -\frac{1}{3}\\\frac{2}{3} & \frac{4}{3} & -\frac{2}{3}\\-\frac{1}{3} & -\frac{2}{3} & \frac{1}{3}\end{bmatrix} = \begin{bmatrix}\frac{2}{3} & -\frac{2}{3} & \frac{1}{3}\\-\frac{2}{3} & -\frac{1}{3} & \frac{2}{3}\\\frac{1}{3} & \frac{2}{3} & \frac{2}{3}\end{bmatrix}.$$

Damit erhalten wir

$$\mathbf{H}_1\,\mathbf{A} = \begin{bmatrix}\frac{2}{3} & -\frac{2}{3} & \frac{1}{3}\\-\frac{2}{3} & -\frac{1}{3} & \frac{2}{3}\\\frac{1}{3} & \frac{2}{3} & \frac{2}{3}\end{bmatrix}\cdot\begin{bmatrix}2 & -3 & 3\\-2 & 6 & 6\\1 & 0 & 3\end{bmatrix} = \begin{bmatrix}3 & -6 & -1\\0 & 0 & -2\\0 & 3 & 7\end{bmatrix}.$$

Im nächsten Schritt betrachten wir die Teilmatrix

$$\mathbf{A}_1 = \begin{bmatrix}0 & -2\\3 & 7\end{bmatrix},$$

und wählen für die Teilmatrix

$$\mathbf{w}_2 = \frac{\mathbf{v}}{\|\mathbf{v}\|_2} \qquad \text{mit} \qquad \mathbf{v} = \begin{bmatrix}0\\3\end{bmatrix} - 3\,\begin{bmatrix}1\\0\end{bmatrix} = \begin{bmatrix}-3\\3\end{bmatrix}, \qquad \|\mathbf{v}\|_2 = 3\,\sqrt{2}.$$

Also gilt

$$\mathbf{w}_2 = \frac{1}{\sqrt{2}}\begin{bmatrix}-1\\1\end{bmatrix},$$

und die 2×2 Householder-Matrix $\mathbf{H}(\mathbf{w}_2)$ (welche $\mathbf{H}(\mathbf{w}_2)\begin{bmatrix}0\\3\end{bmatrix} = 3\,\mathbf{e}_1$ erfüllt) ist

$$\mathbf{H}(\mathbf{w}_2) = \mathbf{E}_2 - \frac{2}{\left(\sqrt{2}\right)^2}\begin{bmatrix}-1\\1\end{bmatrix}\big[-1;1\big] = \mathbf{E}_2 - \begin{bmatrix}1 & -1\\-1 & 1\end{bmatrix}$$

$$= \begin{bmatrix}1 & 0\\0 & 1\end{bmatrix} - \begin{bmatrix}1 & -1\\-1 & 1\end{bmatrix} = \begin{bmatrix}0 & 1\\1 & 0\end{bmatrix}.$$

Also ist die orthogonale Matrix $\mathbf{H}_2 \in \mathbb{R}^{3\times 3}$

$$\mathbf{H}_2 = \begin{bmatrix}1 & 0 & 0\\0 & 0 & 1\\0 & 1 & 0\end{bmatrix},$$

und wir finden

$$\mathbf{R} := \mathbf{H}_2\,\mathbf{H}_1\,\mathbf{A} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & 1 & 0 \end{bmatrix} \cdot \begin{bmatrix} 3 & -6 & -1 \\ 0 & 0 & -2 \\ 0 & 3 & 7 \end{bmatrix} = \begin{bmatrix} 3 & -6 & -1 \\ 0 & 3 & 7 \\ 0 & 0 & -2 \end{bmatrix}.$$

Die QR-Zerlegung ist dann $\mathbf{A} = \mathbf{Q}\,\mathbf{R}$ mit der orthogonalen Matrix $\mathbf{Q}$, die durch

$$\begin{aligned} \mathbf{Q} &= \big(\mathbf{H}_2\,\mathbf{H}_1\big)^{-1} = \mathbf{H}_1^{-1}\,\mathbf{H}_2^{-1} = \mathbf{H}_1^T\,\mathbf{H}_2^T = \mathbf{H}_1\,\mathbf{H}_2 \\ &= \begin{bmatrix} \frac{2}{3} & -\frac{2}{3} & \frac{1}{3} \\ -\frac{2}{3} & -\frac{1}{3} & \frac{2}{3} \\ \frac{1}{3} & \frac{2}{3} & \frac{2}{3} \end{bmatrix} \cdot \begin{bmatrix} 1 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & 1 & 0 \end{bmatrix} = \begin{bmatrix} \frac{2}{3} & \frac{1}{3} & -\frac{2}{3} \\ -\frac{2}{3} & \frac{2}{3} & -\frac{1}{3} \\ \frac{1}{3} & \frac{2}{3} & \frac{2}{3} \end{bmatrix} \end{aligned}$$

definiert ist. ♠

2.5 Lineares Ausgleichsproblem bei überbestimmtem LGS: Methode der kleinsten Quadrate

Wir betrachten nun die Situation, in der bei einem linearen Gleichungssystem **mehr Gleichungen als Unbekannte** vorliegen. Ein solches **„lineares Ausgleichsproblem“** tritt beispielsweise auf, wenn man eine physikalische Größe f durch ein Polynom vom Grad $\leq n-1$, approximieren (d.h. angenähert darstellen) will und $m > n$ Messdaten $\big(t_j, y_j\big)$, $j = 1, 2, \ldots, m$ vorliegen (allgemeiner mehr Messdaten als die Anzahl der in dem Ansatz zu bestimmenden Parameter). Beispielsweise könnte $f(t)$ eine Funktion der Zeit sein, welche die Temperatur an einem festen Ort zur Zeit t beschreibt. Für das Polynom von Grad $\leq n-1$ macht man den folgenden Ansatz

$$P_{n-1}(t) = c_0 + c_1\,t + c_2\,t^2 + \ldots + c_{n-1}\,t^{n-1}$$

mit den n zu bestimmenden Koeffizienten $c_0, c_1, \ldots, c_{n-1} \in \mathbb{R}$. Dann bekommt man aus dem Datensatz $m > n$ Gleichungen

$$\underbrace{1}_{=a_{j,1}}\,c_0 + c_1\,\underbrace{t_j}_{=a_{j,2}} + c_2\,\underbrace{t_j^2}_{=a_{j,3}} + \ldots + c_{n-1}\,\underbrace{t_j^{n-1}}_{=a_{j,n}} = y_j, \qquad j = 1, 2, \ldots, m,$$

also ein lineares Gleichungssystem $\mathbf{A}\,\mathbf{c} = \mathbf{y}$ mit $\mathbf{A} = [a_{j,k}] \in \mathbb{R}^{m\times n}$ und $\mathbf{y} = [y_1; y_2, \ldots; y_m]^T$, also mit $m > n$ Gleichungen zur Bestimmung des Vektors $\mathbf{c} \in \mathbb{R}^n$ der unbekannten Koeffizienten $c_0, c_1, c_2, \ldots, c_{n-1}$ des Polynoms P_{n-1}. Ist $m \gg n$, so ist dieses lineare Gleichungssystem in der Regel **nicht lösbar**.

Wir betrachten also lineare Gleichungssysteme $\mathbf{A}\,\mathbf{x} = \mathbf{b}$ mit $\mathbf{A} = [a_{j,k}] \in \mathbb{R}^{m\times n}$ und einer rechten Seite $\mathbf{b} = [b_j] \in \mathbb{R}^m$, **bei denen** $m > n$ **gilt**. Wir haben also **mehr Gleichungen als Unbekannte**. Da das lineare Gleichungssystem $\mathbf{A}\,\mathbf{x} = \mathbf{b}$ dann im Allgemeinen nicht lösbar sein wird, betrachtet man statt dessen das folgende **Minimierungsproblem:** Bestimme $\mathbf{x} \in \mathbb{R}^n$, welches das Funktional

$$\mu(\mathbf{x}) = \|\mathbf{A}\,\mathbf{x} - \mathbf{b}\|_2^2 \tag{2.46}$$

minimiert. Weil in (2.46) der **quadrierte Euklidische Abstand**

$$\|\mathbf{A}\,\mathbf{x} - \mathbf{b}\|_2^2 = \sum_{j=1}^{n} \left(\sum_{k=1}^{n} a_{j,k}\, x_k - b_j \right)^2$$

minimiert wird, ist dieses Problem auch als **Methode der kleinsten Quadrate** (englisch: „**least-squares approximation**“) bekannt geworden. Man spricht von einem **linearen Ausgleichsproblem**.

Im Folgenden setzen wir voraus, dass $\mathbf{A} \in \mathbb{R}^{m\times n}$ **mit** $m > n$ **den Rang** n **hat (A hat vollen Rang)**, was bedeutet, dass **die** n **Spaltenvektoren von A linear unabhängig sind**. Dieses garantiert (wie wir gleich sehen werden), dass es einen eindeutig bestimmten Vektor $\mathbf{x} \in \mathbb{R}^n$ gibt, der das Funktional μ in (2.46) minimiert.

Wir schreiben der Funktional (2.46) zunächst wie folgt um:

$$\begin{aligned}
\mu(\mathbf{x}) = \|\mathbf{A}\,\mathbf{x} - \mathbf{b}\|_2^2 &= (\mathbf{A}\,\mathbf{x} - \mathbf{b})^T\,(\mathbf{A}\,\mathbf{x} - \mathbf{b}) \\
&= \left(\mathbf{x}^T\mathbf{A}^T - \mathbf{b}^T\right)(\mathbf{A}\,\mathbf{x} - \mathbf{b}) \\
&= \mathbf{x}^T\mathbf{A}^T\mathbf{A}\,\mathbf{x} - \mathbf{x}^T\mathbf{A}^T\,\mathbf{b} - \mathbf{b}^T\mathbf{A}\,\mathbf{x} + \mathbf{b}^T\mathbf{b} \\
&= \mathbf{x}^T\mathbf{A}^T\mathbf{A}\,\mathbf{x} - 2\,\mathbf{x}^T\mathbf{A}^T\,\mathbf{b} + \mathbf{b}^T\mathbf{b},
\end{aligned} \tag{2.47}$$

wobei wir für das skalare (also reellwertige) $\mathbf{b}^T\mathbf{A}\,\mathbf{x}$ genutzt haben, dass $\mathbf{b}^T\mathbf{A}\,\mathbf{x} = \left(\mathbf{b}^T\mathbf{A}\,\mathbf{x}\right)^T = \mathbf{x}^T\mathbf{A}^T\mathbf{b}$ gilt (weil $\mathbf{b}^T\mathbf{A}\,\mathbf{x}$ eine reelle Zahl ist).

Um Kandidaten für die Minimalstelle(n) zu finden, bestimmen wir den Gradienten von μ, wobei wir die Darstellung in der letzten Zeile von (2.47) verwenden:

$$(\nabla\mu)(\mathbf{x}) = 2\,\mathbf{A}^T\mathbf{A}\,\mathbf{x} - 2\,\mathbf{A}^T\,\mathbf{b} = 2\left(\mathbf{A}^T\mathbf{A}\,\mathbf{x} - \mathbf{A}^T\,\mathbf{b}\right) \tag{2.48}$$

Durch Nullsetzen des Gradienten folgt aus (2.48)

$$\mathbf{A}^T\mathbf{A}\,\mathbf{x} - \mathbf{A}^T\,\mathbf{b} = \mathbf{0} \qquad \Longleftrightarrow \qquad \boxed{\mathbf{A}^T\mathbf{A}\,\mathbf{x} = \mathbf{A}^T\,\mathbf{b}.} \tag{2.49}$$

Die Gleichungen in (2.49) werden als die **Normalengleichungen** bezeichnet. Das Berechnen der Hesse-Matrix von μ liefert $(\mathbf{H}\mu)(\mathbf{x}) = 2\,\mathbf{A}^T\mathbf{A}$, und man kann

zeigen, dass diese Matrix positiv definiert ist, so dass bei jedem $\mathbf{x}$, welches (2.49) löst, ein lokales Minimum vorliegt. Dieses ist auch das globale Minimum. Wir zeigen dieses am Ende dieses Teilkapitels.

Da wir vorausgesetzt haben, dass $\mathbf{A} \in \mathbb{R}^{m \times n}$ den Rang n und somit n linear unabhängige Spaltenvektoren hat, ist die $n \times n$ Matrix $\mathbf{A}^T \mathbf{A}$ invertierbar. (Wir erklären dieses am Ende dieses Teilkapitels.) Also sind die Normalengleichungen (2.49) **eindeutig lösbar**. Die Normalengleichungen (2.49) sind aber oft schlecht konditioniert, so dass das direkte Lösen von $\mathbf{A}^T \mathbf{A}\, \mathbf{x} = \mathbf{A}^T \mathbf{b}$ (z.B. mit dem Gaußschen Eliminationsverfahren) oft keine Option ist. Statt dessen wollen wir die **QR-Zerlegung von A nutzen, um die Normalengleichungen zu lösen**.

Zunächst machen wir uns klar, warum man überhaupt eine **QR-Zerlegung der nicht-quadratischen Matrix** $\mathbf{A} \in \mathbb{R}^{m \times n}$ **mit** $m > n$ **und** $\mathrm{Rang}(\mathbf{A}) = n$ bestimmen kann und wie diese aussieht: Da $\mathrm{Rang}(\mathbf{A}) = n$ gilt, sind die n Spaltenvektoren von $\mathbf{A}$ linear unabhängig. Wir können daher zunächst eine $m \times m$ Householder-Matrix $\mathbf{H}_1$ finden, so dass der erste Spaltenvektor von $\mathbf{A}$ auf den Vektor $\alpha_1 \, \mathbf{e}_1 \in \mathbb{R}^m$ (mit einem passend gewählten α_1) abgebildet wird. Dann gilt

$$\mathbf{H}_1 \, \mathbf{A} = \begin{bmatrix} \alpha_1 & * & \cdots & * \\ 0 & & & \\ \vdots & & \mathbf{A}_1 & \\ 0 & & & \end{bmatrix} \qquad \text{mit einer Matrix } \mathbf{A}_1 \in \mathbb{R}^{(m-1) \times (n-1)}.$$

Nun bestimmen wir eine $(m-1) \times (m-1)$ Householder-Matrix $\widetilde{\mathbf{H}}_2$, so dass der erste Spaltenvektor von $\mathbf{A}_1$ auf $\alpha_1 \, \mathbf{e}_1 \in \mathbb{R}^{m-1}$ abgebildet wird. Dann gilt

$$\widetilde{\mathbf{H}}_2 \, \mathbf{A}_1 = \begin{bmatrix} \alpha_2 & * & \cdots & * \\ 0 & & & \\ \vdots & & \mathbf{A}_2 & \\ 0 & & & \end{bmatrix} \qquad \text{mit einer Matrix } \mathbf{A}_2 \in \mathbb{R}^{(m-2) \times (n-2)}.$$

Wir betten $\widetilde{\mathbf{H}}_2$ nun passend in eine $m \times m$-Matrix ein: Die Matrix

$$\mathbf{H}_2 = \begin{bmatrix} 1 & 0 & \cdots & 0 \\ 0 & & & \\ \vdots & & \widetilde{\mathbf{H}}_2 & \\ 0 & & & \end{bmatrix} \in \mathbb{R}^{m \times m}$$

ist dann ebenfalls symmetrisch und orthogonal, da $\mathbf{H}_2^T \, \mathbf{H}_2 = \mathbf{E}_n$ aus $\widetilde{\mathbf{H}}_2^T \, \widetilde{\mathbf{H}}_2 =$

$\mathbf{E}_{n-1}$ folgt. Die Matrix $\mathbf{H}_2$ ist sogar eine Householder-Matrix, und es gilt

$$\mathbf{H}_2\,\mathbf{H}_1\,\mathbf{A} = \mathbf{H}_2 \begin{bmatrix} \alpha_1 & * & \cdots & * \\ 0 & & & \\ \vdots & & \mathbf{A}_1 & \\ 0 & & & \end{bmatrix} = \begin{bmatrix} \alpha_1 & * & * & \cdots & * \\ 0 & \alpha_2 & * & \cdots & * \\ 0 & 0 & & & \\ \vdots & \vdots & & \mathbf{A}_2 & \\ 0 & 0 & & & \end{bmatrix}.$$

Wir setzen diesen Prozess fort und erhalten **nach n Schritten**

$$\mathbf{H}_n \cdots \mathbf{H}_2\,\mathbf{H}_1\,\mathbf{A} = \begin{bmatrix} \alpha_1 & * & * & \cdots & * \\ 0 & \alpha_2 & * & \cdots & * \\ 0 & 0 & \alpha_3 & & \vdots \\ 0 & 0 & 0 & \ddots & * \\ \vdots & \vdots & \vdots & \ddots & \alpha_n \\ 0 & 0 & 0 & \dots & 0 \\ \vdots & \vdots & \vdots & & \vdots \\ 0 & 0 & 0 & \dots & 0 \end{bmatrix} =: \mathbf{R},$$

wobei die Zahlen $\alpha_1, \alpha_2, \dots, \alpha_n$ alle ungleich null sind und wir unten entsprechend $m-n$ Nullzeilen vorfinden. Auflösen nach $\mathbf{A}$ liefert

$$\mathbf{A} = \mathbf{H}_1^{-1}\,\mathbf{H}_2^{-1} \cdots \mathbf{H}_n^{-1}\,\mathbf{R}, \tag{2.50}$$

und weil die Matrizen $\mathbf{H}_1, \mathbf{H}_2, \dots, \mathbf{H}_n$ alle orthogonal und symmetrisch sind gilt $\mathbf{H}_j^{-1} = \mathbf{H}_j^T = \mathbf{H}_j$, $j = 1, 2, \dots, n$. Damit vereinfacht sich (2.50) zu

$$\mathbf{A} = \mathbf{H}_1\,\mathbf{H}_2 \cdots \mathbf{H}_n\,\mathbf{R},$$

und wir erhalten eine **QR-Zerlegung von $\mathbf{A}$**

$$\mathbf{A} = \mathbf{Q}\,\mathbf{R} \qquad \text{mit} \qquad \mathbf{Q} := \mathbf{H}_1\,\mathbf{H}_2 \cdots \mathbf{H}_n, \qquad \mathbf{R} = \begin{bmatrix} * & \cdots & * \\ 0 & \ddots & \vdots \\ \vdots & \ddots & * \\ 0 & \dots & 0 \\ \vdots & & \vdots \\ 0 & \dots & 0 \end{bmatrix} = \begin{bmatrix} \mathbf{R}_1 \\ \mathbf{0} \end{bmatrix},$$

wobei **$\mathbf{Q}$ orthogonal** und $\mathbf{0} = \mathbf{0}_{(m-n)\times n}$ die $(m-n)\times n$ Nullmatrix ist und wobei die **obere Dreiecksmatrix $\mathbf{R}_1 \in \mathbb{R}^{n\times n}$ auf der Diagonalen nur Einträge ungleich null hat** und somit regulär, also invertierbar, ist.

Wir setzen nun die hergeleitete QR-Zerlegung $\mathbf{A} = \mathbf{Q}\,\mathbf{R}$ in das zu minimierende Funktional μ (2.46) ein und formen geeignet um:

$$\mu(\mathbf{x}) = \|\mathbf{A}\,\mathbf{x} - \mathbf{b}\|_2^2 = \big\|\mathbf{Q}\,\mathbf{R}\,\mathbf{x} - \mathbf{b}\big\|_2^2 = \big\|\mathbf{Q}\,\mathbf{R}\,\mathbf{x} - \underbrace{\mathbf{Q}\,\mathbf{Q}^T}_{=\mathbf{E}_n}\mathbf{b}\big\|_2^2$$

$$= \left\| \mathbf{Q} \left(\mathbf{R}\,\mathbf{x} - \mathbf{Q}^T\mathbf{b} \right) \right\|_2^2 = \left\| \mathbf{R}\,\mathbf{x} - \mathbf{Q}^T\mathbf{b} \right\|_2^2$$
$$= \left\| \begin{bmatrix} \mathbf{R}_1\,\mathbf{x} \\ \mathbf{0} \end{bmatrix} - \begin{bmatrix} \mathbf{c} \\ \mathbf{d} \end{bmatrix} \right\|_2^2 = \left\| \mathbf{R}_1\,\mathbf{x} - \mathbf{c} \right\|_2^2 + \left\| \mathbf{d} \right\|_2^2, \tag{2.51}$$

wobei der Vektor $\mathbf{Q}^T\mathbf{b}$ in $\mathbf{Q}^T\mathbf{b} =: \begin{bmatrix} \mathbf{c} \\ \mathbf{d} \end{bmatrix}$ mit $\mathbf{c} \in \mathbb{R}^n$ und $\mathbf{d} \in \mathbb{R}^{m-n}$ zerlegt wurde. In der zweiten Zeile haben wir genutzt, dass $\mathbf{Q}$ orthogonal ist und dass daher $\|\mathbf{Q}\,\mathbf{y}\|_2 = \|\mathbf{y}\|_2$ für alle $\mathbf{y} \in \mathbb{R}^m$ gilt. An der Darstellung von $\mu(\mathbf{x})$ in der letzten Zeile von (2.51) sieht man, dass wir den unvermeidbaren Fehler $\|\mathbf{d}\|_2^2$ haben. Der erste Term $\|\mathbf{R}_1\,\mathbf{x} - \mathbf{c}\|_2^2$ wird genau dann minimal $\mathbf{R}_1\,\mathbf{x} = \mathbf{c}$ gilt. Da $\mathbf{R}_1$ eine reguläre obere Dreiecksmatrix ist, können wir $\mathbf{R}_1\,\mathbf{x} = \mathbf{c}$ mit Rückwärtsrechnen eindeutig lösen.

Betrachten wir die Vorgehensweise an einem Beispiel.

Beispiel 2.32. (Methode der kleinsten Quadrate mit QR-Zerlegung)
Gesucht ist die Lösung des linearen Ausgleichsproblems $\mathbf{A}\,\mathbf{x} = \mathbf{b}$ mit

$$\mathbf{A} = \begin{bmatrix} 8 & -3 & -1 \\ -8 & -3 & -11 \\ 0 & 3 & 3 \\ -4 & 0 & 2 \\ 0 & -3 & -9 \end{bmatrix} \qquad \text{und} \qquad \mathbf{b} = \begin{bmatrix} 18 \\ -9 \\ 21 \\ 0 \\ 0 \end{bmatrix}.$$

$\mathbf{A}\,\mathbf{x} = \mathbf{b}$ hat also 5 Gleichungen, aber nur 3 Unbekannte.

Die Lösung des linearen Ausgleichsproblems soll wie oben beschrieben mit Hilfe der QR-Zerlegung von $\mathbf{A}$ berechnet werden.

Da für den ersten Spaltenvektor $\mathbf{x} = \begin{bmatrix} 8 \\ -8 \\ 0 \\ -4 \\ 0 \end{bmatrix}$ of $\mathbf{A}$ gilt $\|\mathbf{x}\|_2 = \sqrt{144} = 12$, wählen wir den Vektor für die erste Householder-Matrix als

$$\mathbf{w}_1 := \frac{\mathbf{z}}{\|\mathbf{z}\|_2} \qquad \text{mit} \qquad \mathbf{z} := \begin{bmatrix} 8 \\ -8 \\ 0 \\ -4 \\ 0 \end{bmatrix} - 12 \begin{bmatrix} 1 \\ 0 \\ 0 \\ 0 \\ 0 \end{bmatrix} = \begin{bmatrix} -4 \\ -8 \\ 0 \\ -4 \\ 0 \end{bmatrix}, \qquad \|\mathbf{z}\|_2 = \sqrt{96} = 4\sqrt{6}.$$

Also gilt

$$\mathbf{w}_1 = \frac{1}{4\sqrt{6}} \begin{bmatrix} -4 \\ -8 \\ 0 \\ -4 \\ 0 \end{bmatrix} = \frac{1}{\sqrt{6}} \begin{bmatrix} -1 \\ -2 \\ 0 \\ -1 \\ 0 \end{bmatrix},$$

und die erste Householder-Matrix ist

$$\mathbf{H}_1 := \mathbf{H}(\mathbf{w}_1) = \mathbf{E}_5 - 2\,\mathbf{w}_1\,\mathbf{w}_1^T = \mathbf{E}_5 - \frac{2}{\left(\sqrt{6}\right)^2} \begin{bmatrix} -1 \\ -2 \\ 0 \\ -1 \\ 0 \end{bmatrix} \begin{bmatrix} -1; -2; 0; -1; 0 \end{bmatrix}$$

$$= \frac{1}{3} \cdot 3\,\mathbf{E}_5 - \frac{1}{3} \begin{bmatrix} 1 & 2 & 0 & 1 & 0 \\ 2 & 4 & 0 & 2 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 1 & 2 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{bmatrix} = \frac{1}{3} \begin{bmatrix} 2 & -2 & 0 & -1 & 0 \\ -2 & -1 & 0 & -2 & 0 \\ 0 & 0 & 3 & 0 & 0 \\ -1 & -2 & 0 & 2 & 0 \\ 0 & 0 & 0 & 0 & 3 \end{bmatrix}.$$

Damit erhalten wir

$$\mathbf{H}_1\,\mathbf{A} = \frac{1}{3} \begin{bmatrix} 2 & -2 & 0 & -1 & 0 \\ -2 & -1 & 0 & -2 & 0 \\ 0 & 0 & 3 & 0 & 0 \\ -1 & -2 & 0 & 2 & 0 \\ 0 & 0 & 0 & 0 & 3 \end{bmatrix} \cdot \begin{bmatrix} 8 & -3 & -1 \\ -8 & -3 & -11 \\ 0 & 3 & 3 \\ -4 & 0 & 2 \\ 0 & -3 & -9 \end{bmatrix}$$

$$= \frac{1}{3} \begin{bmatrix} 36 & 0 & 18 \\ 0 & 9 & 9 \\ 0 & 9 & 9 \\ 0 & 9 & 27 \\ 0 & -9 & -27 \end{bmatrix} = \begin{bmatrix} 12 & 0 & 6 \\ 0 & 3 & 3 \\ 0 & 3 & 3 \\ 0 & 3 & 9 \\ 0 & -3 & -9 \end{bmatrix}.$$

Im nächsten Schritt betrachten wir die Teilmatrix

$$\mathbf{A}_1 = \begin{bmatrix} 3 & 3 \\ 3 & 3 \\ 3 & 9 \\ -3 & -9 \end{bmatrix},$$

Für den ersten Spaltenvektor gilt $\widetilde{\mathbf{x}} = \begin{bmatrix} 3 \\ 3 \\ 3 \\ -3 \end{bmatrix}$ gilt $\|\widetilde{\mathbf{x}}\|_2 = \sqrt{36} = 6$.

Als Vektor für die 4×4 Householder-Matrix wählen wir

$$\mathbf{w}_2 := \frac{\mathbf{v}}{\|\mathbf{v}\|_2} \qquad \text{mit} \qquad \mathbf{v} := \begin{bmatrix} 3 \\ 3 \\ 3 \\ -3 \end{bmatrix} - 6 \begin{bmatrix} 1 \\ 0 \\ 0 \\ 0 \end{bmatrix} = \begin{bmatrix} -3 \\ 3 \\ 3 \\ -3 \end{bmatrix}, \qquad \|\mathbf{v}\|_2 = \sqrt{36} = 6.$$

Also gilt

$$\mathbf{w}_2 = \frac{1}{6} \begin{bmatrix} -3 \\ 3 \\ 3 \\ -3 \end{bmatrix} = \frac{1}{2} \begin{bmatrix} -1 \\ 1 \\ 1 \\ -1 \end{bmatrix},$$

und die 4×4 Householder-Matrix ist

$$\mathbf{H}(\mathbf{w}_2) = \mathbf{E}_4 - 2\,\mathbf{w}_2\,\mathbf{w}_2^T = \mathbf{E}_4 - \frac{2}{2^2} \begin{bmatrix} -1 \\ 1 \\ 1 \\ -1 \end{bmatrix} \big[-1; 1; 1; -1\big]$$

$$= \frac{1}{2} \cdot 2\,\mathbf{E}_4 - \frac{1}{2} \begin{bmatrix} 1 & -1 & -1 & 1 \\ -1 & 1 & 1 & -1 \\ -1 & 1 & 1 & -1 \\ 1 & -1 & -1 & 1 \end{bmatrix} = \frac{1}{2} \begin{bmatrix} 1 & 1 & 1 & -1 \\ 1 & 1 & -1 & 1 \\ 1 & -1 & 1 & 1 \\ -1 & 1 & 1 & 1 \end{bmatrix}.$$

Also ist die orthogonale Matrix $\mathbf{H}_2$

$$\mathbf{H}_2 = \frac{1}{2} \begin{bmatrix} 2 & 0 & 0 & 0 & 0 \\ 0 & 1 & 1 & 1 & -1 \\ 0 & 1 & 1 & -1 & 1 \\ 0 & 1 & -1 & 1 & 1 \\ 0 & -1 & 1 & 1 & 1 \end{bmatrix},$$

und wir finden

$$\mathbf{R} := \mathbf{H}_2\,\mathbf{H}_1\,\mathbf{A} = \frac{1}{2} \begin{bmatrix} 2 & 0 & 0 & 0 & 0 \\ 0 & 1 & 1 & 1 & -1 \\ 0 & 1 & 1 & -1 & 1 \\ 0 & 1 & -1 & 1 & 1 \\ 0 & -1 & 1 & 1 & 1 \end{bmatrix} \cdot \begin{bmatrix} 12 & 0 & 6 \\ 0 & 3 & 3 \\ 0 & 3 & 3 \\ 0 & 3 & 9 \\ 0 & -3 & -9 \end{bmatrix}$$

$$= \frac{1}{2} \begin{bmatrix} 24 & 0 & 12 \\ 0 & 12 & 24 \\ 0 & 0 & -12 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix} = \begin{bmatrix} 12 & 0 & 6 \\ 0 & 6 & 12 \\ 0 & 0 & -6 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix} \qquad \text{mit} \qquad \mathbf{R}_1 = \begin{bmatrix} 12 & 0 & 6 \\ 0 & 6 & 12 \\ 0 & 0 & -6 \end{bmatrix}.$$

Achtung: Normalerweise ist noch ein dritter Schritt erforderlich, aber hier erhalten wir schon nach zwei Schritten die Stufenform.

Die QR-Zerlegung ist dann $\mathbf{A} = \mathbf{Q}\,\mathbf{R}$ mit der orthogonalen Matrix $\mathbf{Q}$, die durch

$$\mathbf{Q} = \big(\mathbf{H}_2\,\mathbf{H}_1\big)^{-1} = \mathbf{H}_1^{-1}\,\mathbf{H}_2^{-1} = \mathbf{H}_1^T\,\mathbf{H}_2^T = \mathbf{H}_1\,\mathbf{H}_2$$

$$= \frac{1}{3}\begin{bmatrix} 2 & -2 & 0 & -1 & 0 \\ -2 & -1 & 0 & -2 & 0 \\ 0 & 0 & 3 & 0 & 0 \\ -1 & -2 & 0 & 2 & 0 \\ 0 & 0 & 0 & 0 & 3 \end{bmatrix} \cdot \frac{1}{2}\begin{bmatrix} 2 & 0 & 0 & 0 & 0 \\ 0 & 1 & 1 & 1 & -1 \\ 0 & 1 & 1 & -1 & 1 \\ 0 & 1 & -1 & 1 & 1 \\ 0 & -1 & 1 & 1 & 1 \end{bmatrix}$$

$$= \frac{1}{6}\begin{bmatrix} 4 & -3 & -1 & -3 & 1 \\ -4 & -3 & 1 & -3 & -1 \\ 0 & 3 & 3 & -3 & 3 \\ -2 & 0 & -4 & 0 & 4 \\ 0 & -3 & 3 & 3 & 3 \end{bmatrix}$$

gegeben ist.

Zur Lösung des linearen Ausgleichsproblems benötigen wir noch $\mathbf{Q}^T\mathbf{b}$, also

$$\mathbf{Q}^T\mathbf{b} = \frac{1}{6}\begin{bmatrix} 4 & -4 & 0 & -2 & 0 \\ -3 & -3 & 3 & 0 & -3 \\ -1 & 1 & 3 & -4 & 3 \\ -3 & -3 & -3 & 0 & 3 \\ 1 & -1 & 3 & 4 & 3 \end{bmatrix}\begin{bmatrix} 18 \\ -9 \\ 21 \\ 0 \\ 0 \end{bmatrix} = \frac{1}{6}\begin{bmatrix} 108 \\ 36 \\ 36 \\ -90 \\ 90 \end{bmatrix} = \begin{bmatrix} 18 \\ 6 \\ 6 \\ -15 \\ 15 \end{bmatrix}$$

Wir lesen ab: $\mathbf{c} = \begin{bmatrix} 18 \\ 6 \\ 6 \end{bmatrix}$.

Die Lösung des linearen Ausgleichsproblems kann nun durch Rückwärtsrechnen aus $\mathbf{R}_1\,\mathbf{x} = \mathbf{c}$ berechnet werden:

$$\begin{bmatrix} 12 & 0 & 6 \\ 0 & 6 & 12 \\ 0 & 0 & -6 \end{bmatrix}\begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} 18 \\ 6 \\ 6 \end{bmatrix} \quad \Longrightarrow \quad \begin{cases} x_3 = -1, \\ x_2 = \dfrac{1}{6}\,(6 - 12\,x_3) = \dfrac{1}{6}\,(6 + 12) = 3, \\ x_1 = \dfrac{1}{12}\,(18 - 6\,x_3) = \dfrac{1}{12}\,(18 + 6) = 2. \end{cases}$$

Also ist die Lösung des linearen Ausgleichsproblems $\mathbf{x} = \begin{bmatrix} 2 \\ 3 \\ -1 \end{bmatrix}$. ♠

Wir zeigen nun, dass die Normalengleichungen $\mathbf{A}^T\mathbf{A}\,\mathbf{x} = \mathbf{A}^T\,\mathbf{b}$ für jedes $\mathbf{b} \in \mathbb{R}^m$ eindeutig lösbar sind, wenn die Matrix $\mathbf{A} \in \mathbb{R}^{m\times n}$ eine $m \times n$-Matrix mit $m > n$ und mit Rang n ist (d.h. wenn $\mathbf{A}$ mehr Zeilen als Spalten hat und wenn die n Spaltenvektoren von $\mathbf{A}$ linear unabhängig sind).

Nachweis: Die Matrix $\mathbf{A}^T\mathbf{A}$ ist eine $n \times n$ Matrix. Sie ist genau dann invertierbar, wenn das lineare Gleichungssystem $\mathbf{A}^T\mathbf{A}\,\mathbf{x} = \mathbf{0}$ nur die einzige Lösung $\mathbf{x} = \mathbf{0}$ hat.

$$\mathbf{A}^T\mathbf{A}\,\mathbf{x} = \mathbf{0} \quad \Longrightarrow \quad \underbrace{\mathbf{x}^T\mathbf{A}^T\mathbf{A}\,\mathbf{x}}_{=(\mathbf{A}\mathbf{x})^T(\mathbf{A}\mathbf{x})=\|\mathbf{A}\,\mathbf{x}\|_2^2} = \underbrace{\mathbf{x}^T\mathbf{0}}_{=0}$$

$$\Longleftrightarrow \quad \|\mathbf{A}\,\mathbf{x}\|_2^2 = 0 \quad \Longleftrightarrow \quad \mathbf{A}\,\mathbf{x} = \mathbf{0},$$

wobei der letzte Schritt aus den Eigenschaften einer Norm folgt. Da $\mathbf{A}$ den Rang n hat, sind die n Spaltenvektoren $\mathbf{a}_1, \mathbf{a}_2, \ldots, \mathbf{a}_n$ von $\mathbf{A}$ linear unabhängig. Der Vektor $\mathbf{A}\,\mathbf{x}$ ist aber gerade eine Linearkombination der Spaltenvektoren $\mathbf{a}_1, \mathbf{a}_2, \ldots, \mathbf{a}_n$ von $\mathbf{A}$, wobei $x_1, x_2, \ldots, x_n$ die Koeffizienten der Spaltenvektoren sind. Also gilt:

$$\mathbf{A}\,\mathbf{x} = \mathbf{0} \quad \Longleftrightarrow \quad x_1\,\mathbf{a}_1 + x_2\,\mathbf{a}_2 + \ldots + x_n\,\mathbf{a}_n = \mathbf{0}.$$

Eine Linearkombination der linear unabhängigen Vektoren $\mathbf{a}_1, \mathbf{a}_2, \ldots, \mathbf{a}_n$ ist aber nur dann der Nullvektor, wenn alle Koeffizienten null sind. Also folgt für die Koeffizienten $x_1 = x_2 = \ldots = x_n = 0$, d.h. $\mathbf{x} = \mathbf{0}$.

Also folgt aus $\mathbf{A}^T\mathbf{A}\,\mathbf{x} = \mathbf{0}$, dass $\mathbf{x} = \mathbf{0}$ ist. Folglich ist $\mathbf{A}^T\mathbf{A}$ invertierbar. Die Normalengleichungen $\mathbf{A}^T\mathbf{A}\,\mathbf{x} = \mathbf{A}^T\,\mathbf{b}$ haben damit für jedes $\mathbf{b} \in \mathbb{R}^n$ eine eindeutige Lösung. □

Seien $m > n$ und $\mathbf{A} \in \mathbb{R}^{m\times n}$ eine $m \times n$-Matrix mit Rang n, d.h. die n Spaltenvektoren von $\mathbf{A}$ sind linear unabhängig. Wir zeigen nun, dass die eindeutige Lösung $\widehat{\mathbf{x}}$ der Normalengleichungen $\mathbf{A}^T\mathbf{A}\,\mathbf{x} = \mathbf{A}^T\,\mathbf{b}$ die eindeutig bestimmte globale Minimalstelle des Funktionals (2.46) ist.

Nachweis: Sei $\widehat{\mathbf{x}}$ die eindeutige Lösung der Normalengleichungen $\mathbf{A}^T\mathbf{A}\,\mathbf{x} = \mathbf{A}^T\,\mathbf{b}$, d.h. es gilt $\mathbf{A}^T\mathbf{A}\,\widehat{\mathbf{x}} = \mathbf{A}^T\,\mathbf{b}$. Wir betrachten nun $\mu(\widehat{\mathbf{x}} + \mathbf{x})$ mit $\mathbf{x} \in \mathbb{R}^n$ beliebig:

$$\begin{aligned}
\mu(\widehat{\mathbf{x}} + \mathbf{x}) &= \big\|\mathbf{A}\,(\widehat{\mathbf{x}} + \mathbf{x}) - \mathbf{b}\big\|_2^2 = \big\|(\mathbf{A}\,\widehat{\mathbf{x}} - \mathbf{b}) + \mathbf{A}\,\mathbf{x}\big\|_2^2 \\
&= \big((\mathbf{A}\,\widehat{\mathbf{x}} - \mathbf{b}) + \mathbf{A}\,\mathbf{x}\big)^T\big((\mathbf{A}\,\widehat{\mathbf{x}} - \mathbf{b}) + \mathbf{A}\,\mathbf{x}\big) \\
&= \big((\mathbf{A}\,\widehat{\mathbf{x}} - \mathbf{b})^T + (\mathbf{A}\,\mathbf{x})^T\big)\big((\mathbf{A}\,\widehat{\mathbf{x}} - \mathbf{b}) + \mathbf{A}\,\mathbf{x}\big) \\
&= (\mathbf{A}\,\widehat{\mathbf{x}} - \mathbf{b})^T(\mathbf{A}\,\widehat{\mathbf{x}} - \mathbf{b}) + (\mathbf{A}\,\mathbf{x})^T(\mathbf{A}\,\widehat{\mathbf{x}} - \mathbf{b}) \qquad (2.52) \\
&\quad + \underbrace{(\mathbf{A}\,\widehat{\mathbf{x}} - \mathbf{b})^T\mathbf{A}\,\mathbf{x}}_{=(\mathbf{A}\,\mathbf{x})^T(\mathbf{A}\,\widehat{\mathbf{x}}-\mathbf{b})} + (\mathbf{A}\,\mathbf{x})^T\mathbf{A}\,\mathbf{x}
\end{aligned}$$

$$
\begin{aligned}
&= \|\mathbf{A}\,\widehat{\mathbf{x}} - \mathbf{b}\|_2^2 + 2\,(\mathbf{A}\,\mathbf{x})^T(\mathbf{A}\,\widehat{\mathbf{x}} - \mathbf{b}) + \|\mathbf{A}\,\mathbf{x}\|_2^2 \\
&= \|\mathbf{A}\,\widehat{\mathbf{x}} - \mathbf{b}\|_2^2 + 2\,\mathbf{x}^T\big(\underbrace{\mathbf{A}^T\mathbf{A}\,\widehat{\mathbf{x}} - \mathbf{A}^T\,\mathbf{b}}_{=\mathbf{0}}\big) + \|\mathbf{A}\,\mathbf{x}\|_2^2 \\
&= \|A\,\widehat{\mathbf{x}} - \mathbf{b}\|_2^2 + \|\mathbf{A}\,\mathbf{x}\|_2^2,
\end{aligned}
$$

wobei wir beim Wechsel von der vierte in die fünfte Zeile $(\mathbf{A}\,\widehat{\mathbf{x}} - \mathbf{b})^T\mathbf{A}\,\mathbf{x} \in \mathbb{R}$ und somit

$$(\mathbf{A}\,\widehat{\mathbf{x}} - \mathbf{b})^T\mathbf{A}\,\mathbf{x} = \big((\mathbf{A}\,\widehat{\mathbf{x}} - \mathbf{b})^T\mathbf{A}\,\mathbf{x}\big)^T = (\mathbf{A}\,\mathbf{x})^T(\mathbf{A}\,\widehat{\mathbf{x}} - \mathbf{b})$$

genutzt haben. Also gilt

$$\mu(\mathbf{x} + \widehat{\mathbf{x}}) = \|\mathbf{A}\,\widehat{\mathbf{x}} - \mathbf{b}\|_2^2 + \|\mathbf{A}\,\mathbf{x}\|_2^2 \qquad \text{für alle } \mathbf{x} \in \mathbb{R}^n,$$

oder äquivalent dazu, indem man $\mathbf{y} = \mathbf{x} + \widehat{\mathbf{x}} \iff \mathbf{x} = \mathbf{y} - \widehat{\mathbf{x}}$ ersetzt,

$$
\begin{aligned}
\mu(\mathbf{y}) &= \|\mathbf{A}\,\widehat{\mathbf{x}} - \mathbf{b}\|_2^2 + \|\mathbf{A}\,(\mathbf{y} - \widehat{\mathbf{x}})\|_2^2 \\
&= \mu(\widehat{\mathbf{x}}) + \underbrace{\|\mathbf{A}\,(\mathbf{y} - \widehat{\mathbf{x}})\|_2^2}_{\geq 0} \qquad \text{für alle } \mathbf{y} \in \mathbb{R}^n. \qquad (2.53)
\end{aligned}
$$

An (2.53) sieht man, dass für jedes $\mathbf{y} \in \mathbb{R}^n$ gilt $\mu(\mathbf{y}) \geq \mu(\widehat{\mathbf{x}})$, und Gleichheit tritt nur auf, wenn $\|\mathbf{A}\,(\mathbf{y} - \widehat{\mathbf{x}})\|_2 = 0$ ist. Die Bedingung $\|\mathbf{A}\,(\mathbf{y} - \widehat{\mathbf{x}})\|_2 = 0$ ist äquivalent zu $\mathbf{A}\,(\mathbf{y} - \widehat{\mathbf{x}}) = \mathbf{0}$. Aus $\mathbf{A}\,(\mathbf{y} - \widehat{\mathbf{x}}) = \mathbf{0}$ folgt analog zu den Überlegungen in Teil (a), dass $\mathbf{y} - \widehat{\mathbf{x}} = \mathbf{0} \iff \mathbf{y} = \widehat{\mathbf{x}}$ gelten muss. Also gilt $\mu(\mathbf{y}) > \mu(\widehat{\mathbf{x}})$ für alle $\mathbf{y} \in \mathbb{R}^n \setminus \{\widehat{\mathbf{x}}\}$, d.h. $\widehat{\mathbf{x}}$ ist die eindeutig bestimmte globale Minimalstelle des Funktionals μ. □

KAPITEL 3

Iterative Lösungsverfahren für lineare Gleichungssysteme

In diesem Kapitel lernen wir iterative Lösungsverfahren für lineare Gleichungssysteme kennen. Im Gegensatz zu direkten Lösungsverfahren für lineare Gleichungssysteme, welche Sie aus dem vorigen Kapitel kennen, wird bei einem **iterativen Lösungsverfahren** (oder **Iterationsverfahren**) eine **Folge von Näherungen der Lösung** berechnet, die unter geeigneten Voraussetzungen **gegen die Lösung des linearen Gleichungssystems konvergiert**.

3.1 Fixpunktiteration

Die Grundlage für viele Iterationsverfahren bildet die Umschreibung der zu lösenden (linearen oder nichtlinearen) Gleichungen in eine Fixpunktgleichung.

Definition 3.1. (Fixpunkt und Fixpunktgleichung)
*Seien $D \subseteq \mathbb{R}^n$ und $\mathbf{f} : D \to \mathbb{R}^n$ eine Funktion. Die Gleichung $\mathbf{f}(\mathbf{x}) = \mathbf{x}$ heißt eine **Fixpunktgleichung**. Ein $\widehat{\mathbf{x}} \in D$ heißt ein **Fixpunkt von** $\mathbf{f}$, wenn $\widehat{\mathbf{x}}$ eine Lösung der Fixpunktgleichung ist, also wenn gilt $\mathbf{f}(\widehat{\mathbf{x}}) = \widehat{\mathbf{x}}$.*

Beispiel 3.2. (Fixpunkte und Fixpunktgleichungen)

(a) Die Funktion $h : \mathbb{R} \to \mathbb{R}$, $h(x) = x^3$, hat genau die drei Fixpunkte $x_1 = -1$,

$x_2 = 0$ und $x_3 = 1$. Dieses folgt (mit der dritten binomischen Formel) aus

$$x^3 = x \qquad \Longleftrightarrow \qquad 0 = x^3 - x = x\,(x^2 - 1) = x\,(x-1)\,(x+1).$$

(b) Die Funktion $f : \mathbb{R} \to \mathbb{R}$, $f(x) = \frac{\pi}{2}\sin(x)$, hat genau drei Fixpunkte, nämlich $x_1 = -\frac{\pi}{2}$, $x_2 = 0$ und $x_3 = \frac{\pi}{2}$. (Dass es sich um Fixpunkte handelt überprüft man leicht durch Nachrechnen, und, dass es keine weiteren Fixpunkte geben kann, macht man sich mit Hilfe des Graphen klar. Man kann dieses auch rechnerisch nachweisen.)

(c) Für $a > 0$ lässt sich $\sqrt{a}$ als die positive Lösung der Gleichung $x^2 = a$ eindeutig bestimmen. Wir können die Gleichung $x^2 = a$ wie folgt in eine Fixpunktgleichung umschreiben:

$$x^2 = a \qquad \overset{x\neq 0}{\Longleftrightarrow} \qquad x = \frac{a}{x} \qquad \Longleftrightarrow \qquad \frac{1}{2}\,x = \frac{1}{2}\,\frac{a}{x}$$

$$\Longleftrightarrow \qquad x = \frac{1}{2}\,x + \frac{1}{2}\,\frac{a}{x} \qquad \Longleftrightarrow \qquad x = \frac{1}{2}\left(x + \frac{a}{x}\right)$$

Also ist $\sqrt{a}$ ein Fixpunkt (und auch der einzige positive Fixpunkt) der Funktion $g : \mathbb{R} \setminus \{0\} \to \mathbb{R}$, $g(x) = \frac{1}{2}\left(x + \frac{a}{x}\right)$. In der Formel $\frac{1}{2}\left(x + \frac{a}{x}\right)$ wird das arithmetische Mittel von x und $\frac{a}{x}$ gebildet. Für x und $\frac{a}{x}$ gilt $x \cdot \frac{a}{x} = a$; also ist ihr Produkt gleich a. Damit muss eine der beiden positiven Zahlen x und $\frac{a}{x}$ aber $\leq \sqrt{a}$ und eine $\geq \sqrt{a}$ sein. Das arithmetische Mittel $\frac{1}{2}\left(x + \frac{a}{x}\right)$ ist auf jeden Fall eine bessere Näherung als für $\sqrt{a}$ als die schlechtere der beiden Näherungen x und $\frac{a}{x}$.

Wir werden in diesem Kapitel noch weitere Beispiele für Fixpunktgleichungen kennenlernen, bei denen es um das Lösen linearer Gleichungssysteme geht. ♠

Wie betrachten nun ein **lineares Gleichungssystem $\mathbf{A}\,\mathbf{x} = \mathbf{b}$** mit einer quadratischen **regulären** (also **invertierbaren**) Matrix $\mathbf{A} \in \mathbb{R}^{n\times n}$, welches damit eine eindeutig bestimmte Lösung hat. Wir bezeichnen diese **eindeutige Lösung** mit $\widehat{\mathbf{x}}$, also $\mathbf{A}\,\widehat{\mathbf{x}} = \mathbf{b}$.

Wir wollen das lineare Gleichungssystem $\mathbf{A}\,\mathbf{x} = \mathbf{b}$ nun in eine geeignete Fixpunktgleichung umschreiben. Dazu sei $\mathbf{B} \in \mathbb{R}^{n\times n}$ eine reguläre (also invertierbare) Matrix, die in der Regel eine **einfache Näherung der Matrix A ist und sich leicht invertieren lässt**. Dann erhalten wir durch Umformen:

$$\mathbf{A}\,\mathbf{x} = \mathbf{b} \quad \Longleftrightarrow \quad (\mathbf{A}-\mathbf{B})\,\mathbf{x} + \mathbf{B}\,\mathbf{x} = \mathbf{b} \quad \Longleftrightarrow \quad \mathbf{B}\,\mathbf{x} = -(\mathbf{A}-\mathbf{B})\,\mathbf{x} + \mathbf{b}$$

$$\Longleftrightarrow \quad \mathbf{x} = \mathbf{B}^{-1}\left[-(\mathbf{A}-\mathbf{B})\,\mathbf{x} + \mathbf{b}\right] \quad \Longleftrightarrow \quad \mathbf{x} = -\mathbf{B}^{-1}\,(\mathbf{A}-\mathbf{B})\,\mathbf{x} + \mathbf{B}^{-1}\,\mathbf{b}$$

Wir halten also die Fixpunktgleichung

$$\boxed{\mathbf{x} = \mathbf{C}\,\mathbf{x} + \mathbf{c} \qquad \text{mit} \qquad \mathbf{C} := -\mathbf{B}^{-1}\,(\mathbf{A}-\mathbf{B}), \quad \mathbf{c} := \mathbf{B}^{-1}\,\mathbf{b},} \tag{3.1}$$

deren einzige Lösung (immer noch) die eindeutige Lösung $\widehat{\mathbf{x}}$ von $\mathbf{A}\,\mathbf{x} = \mathbf{b}$ ist. Die eindeutige Lösung $\widehat{\mathbf{x}}$ von $\mathbf{A}\,\mathbf{x} = \mathbf{b}$ ist also der einzige Fixpunkt der Funktion

$$\boxed{\mathbf{f} : \mathbb{R}^n \to \mathbb{R}^n, \quad \mathbf{f}(\mathbf{x}) := \mathbf{C}\,\mathbf{x} + \mathbf{c} \quad \text{mit} \quad \mathbf{C} := -\mathbf{B}^{-1}\,(\mathbf{A} - \mathbf{B}), \quad \mathbf{c} := \mathbf{B}^{-1}\,\mathbf{b}.} \tag{3.2}$$

Bei einer **Fixpunktiteration** werden nun ausgehend von einem geeigneten Startvektor $\mathbf{x}^{(0)} \in \mathbb{R}^n$ nacheinander Iterierte

$$\boxed{\mathbf{x}^{(j+1)} = \mathbf{f}(\mathbf{x}^{(j)}) = \mathbf{C}\,\mathbf{x}^{(j)} + \mathbf{c}, \qquad j = 0, 1, 2, \ldots,} \tag{3.3}$$

berechnet. Unter geeigneten Voraussetzungen konvergiert die Folge $\left(\mathbf{x}^{(j)}\right)_{j \in \mathbb{N}_0}$ der Iterierten $\mathbf{x}^{(j)}$ gegen die eindeutige Lösung von $\widehat{\mathbf{x}}$ von $\mathbf{A}\,\mathbf{x} = \mathbf{b}$.

Bevor wir uns für die Konvergenz einer Fixpunktiteration (3.3) interessieren, wollen wir noch einmal (3.3) mit der Wahl $\mathbf{C} := -\mathbf{B}^{-1}\,(\mathbf{A} - \mathbf{B})$ und $\mathbf{c} := \mathbf{B}^{-1}\,\mathbf{b}$ aus (3.1) betrachten, um ein besseres **Verständnis für die Fixpunktiterationen zur Lösung eines eindeutig lösbaren linearen Gleichungssystems** $\mathbf{A}\,\mathbf{x} = \mathbf{b}$ zu gewinnen. Sei also

$$\mathbf{x}^{(j+1)} = -\mathbf{B}^{-1}\,(\mathbf{A} - \mathbf{B})\,\mathbf{x}^{(j)} + \mathbf{B}^{-1}\,\mathbf{b}.$$

Umformen liefert:

$$\begin{aligned}\mathbf{x}^{(j+1)} &= -\mathbf{B}^{-1}\,(\mathbf{A} - \mathbf{B})\,\mathbf{x}^{(j)} + \mathbf{B}^{-1}\,\mathbf{b} = -\mathbf{B}^{-1}\,\mathbf{A}\,\mathbf{x}^{(j)} + \underbrace{\mathbf{B}^{-1}\,\mathbf{B}}_{=\mathbf{E}_n}\,\mathbf{x}^{(j)} + \mathbf{B}^{-1}\,\mathbf{b} \\ &= -\mathbf{B}^{-1}\,\mathbf{A}\,\mathbf{x}^{(j)} + \underbrace{\mathbf{E}_n\,\mathbf{x}^{(j)}}_{=\mathbf{x}^{(j)}} + \mathbf{B}^{-1}\,\mathbf{b} = \mathbf{x}^{(j)} + \mathbf{B}^{-1}\left(\mathbf{b} - \mathbf{A}\,\mathbf{x}^{(j)}\right),\end{aligned}$$

also

$$\boxed{\mathbf{x}^{(j+1)} = \mathbf{x}^{(j)} + \mathbf{B}^{-1}\left(\mathbf{b} - \mathbf{A}\,\mathbf{x}^{(j)}\right).} \tag{3.4}$$

Mit $\mathbf{A}\,\widehat{\mathbf{x}} = \mathbf{b}$ ist der Term

$$\boxed{\mathbf{r}^{(j)} := \mathbf{b} - \mathbf{A}\,\mathbf{x}^{(j)} = \mathbf{A}\,\widehat{\mathbf{x}} - \mathbf{A}\,\mathbf{x}^{(j)} = \mathbf{A}\,(\widehat{\mathbf{x}} - \mathbf{x}^{(j)})}$$

das sogenannte „**Residuum**“, also die **Abweichung von $\widehat{\mathbf{x}} - \mathbf{x}^{(j)}$ abgebildet in die Bildmenge von $\mathbf{g} : \mathbb{R}^n \to \mathbb{R}^n$, $\mathbf{g}(\mathbf{x}) := \mathbf{A}\,\mathbf{x}$.**

Wir haben zu Beginn unserer Überlegungen erklärt, dass $\mathbf{B}$ eine Näherung von $\mathbf{A}$ sein soll, deren Inverse leicht zu berechnen ist. In diesem Sinn ist $\mathbf{B}^{-1}$ dann eine Näherung von $\mathbf{A}^{-1}$. Ersetzen wir in (3.4) $\mathbf{B}^{-1}$ durch $\mathbf{A}^{-1}$, so erhalten wir:

$$\mathbf{x}^{(j+1)} = \mathbf{x}^{(j)} + \mathbf{B}^{-1}\Big(\underbrace{\mathbf{b} - \mathbf{A}\,\mathbf{x}^{(j)}}_{=\mathbf{r}^{(j)}}\Big) \approx \mathbf{x}^{(j)} + \mathbf{A}^{-1}\Big(\underbrace{\mathbf{b} - \mathbf{A}\,\mathbf{x}^{(j)}}_{=\mathbf{r}^{(j)}}\Big)$$

$$= \mathbf{x}^{(j)} + \mathbf{A}^{-1}\,\mathbf{b} - \underbrace{\mathbf{A}^{-1}\mathbf{A}}_{=\mathbf{E}_n}\,\mathbf{x}^{(j)} = \mathbf{x}^{(j)} + \mathbf{A}^{-1}\,\mathbf{b} - \underbrace{\mathbf{E}_n\,\mathbf{x}^{(j)}}_{=\mathbf{x}^{(j)}} = \mathbf{A}^{-1}\,\mathbf{b}$$

Ersetzt man in (3.4) also $\mathbf{B}^{-1}$ durch $\mathbf{A}^{-1}$ so erhält man also die Lösung der LGS $\mathbf{A}\,\mathbf{x} = \mathbf{b}$. Also wird die Iterierte $\mathbf{x}^{(j)}$ in (3.4) mit Hilfe der angenäherten Lösung $\mathbf{B}^{-1}\left(\mathbf{b} - \mathbf{A}\,\mathbf{x}^{(j)}\right) = \mathbf{B}^{-1}\,\mathbf{r}^{(j)}$ der Gleichung $\mathbf{A}\,\mathbf{y} = \mathbf{b} - \mathbf{A}\,\mathbf{x}^{(j)} = \mathbf{r}^{(j)}$, deren rechte Seite das Residuum $\mathbf{r}^{(j)} = \mathbf{b} - \mathbf{A}\,\mathbf{x}^{(j)}$ ist, „verbessert“.

In dem nachfolgenden Teilkapitel werden wir zwei verschiedene Wahlen der Näherung $\mathbf{B}$ von $\mathbf{A}$ (und damit der Näherung $\mathbf{B}^{-1}$ von $\mathbf{A}^{-1}$) genauer untersuchen.

Die Voraussetzungen für die mögliche Konvergenz eine Fixpunktiteration (3.3) liefert der nachfolgende Banachsche Fixpunktsatz, den wir auch noch in einem späteren Kapitel benutzen werden.

Satz 3.3. (Banachscher Fixpunktsatz)
*Sei $\mathbb{R}^n$ mit einer Norm $\|\cdot\|$ gegeben, und sei $D \subseteq \mathbb{R}^n$ abgeschlossen. Sei die Funktion $\mathbf{f} : D \to D$ eine **Kontraktion (bzgl. $\|\cdot\|$ auf D)**, d.h. es gibt eine Konstante q mit $0 < q < 1$, so dass gilt*

$$\|\mathbf{f}(\mathbf{x}) - \mathbf{f}(\mathbf{y})\| \leq q\,\|\mathbf{x} - \mathbf{y}\| \qquad \text{für alle } \mathbf{x}, \mathbf{y} \in D. \tag{3.5}$$

Dann gelten:

*(1) $\mathbf{f}$ hat **genau einen einzigen** Fixpunkt $\widehat{\mathbf{x}}$ in D.*

*(2) Die rekursiv durch $\mathbf{x}^{(j+1)} := \mathbf{f}\left(\mathbf{x}^{(j)}\right)$ definierte Folge $\left(\mathbf{x}^{(j)}\right)_{j\in\mathbb{N}_0}$ **konvergiert für jeden Startvektor** $\mathbf{x}^{(0)} \in D$ gegen den Fixpunkt $\widehat{\mathbf{x}}$.*

*(3) Für $\mathbf{x}^{(j)}$ gilt die folgende **a posteriori Fehlerabschätzung**:*

$$\|\mathbf{x}^{(j)} - \widehat{\mathbf{x}}\| \leq \frac{q}{1-q}\,\|\mathbf{x}^{(j)} - \mathbf{x}^{(j-1)}\|, \qquad j = 1, 2, \ldots. \tag{3.6}$$

*(4) Für $\mathbf{x}^{(j)}$ gilt die folgende **a priori Fehlerabschätzung**:*

$$\|\mathbf{x}^{(j)} - \widehat{\mathbf{x}}\| \leq \frac{q^j}{1-q}\,\|\mathbf{x}^{(1)} - \mathbf{x}^{(0)}\|, \qquad j = 1, 2, \ldots. \tag{3.7}$$

Was bedeutet (3.5)**?** Auf der rechten Seite von (3.5) steht der Abstand $\|\mathbf{x}-\mathbf{y}\|$ von $\mathbf{x}$ und $\mathbf{y}$ (gemessen in der Norm $\|\cdot\|$). Auf der linken Seite von (3.5) steht der Abstand $\|\mathbf{f}(\mathbf{x}) - \mathbf{f}(\mathbf{y})\|$ von $\mathbf{f}(\mathbf{x})$ und $\mathbf{f}(\mathbf{y})$ (gemessen in der Norm $\|\cdot\|$). Da die Konstante q in $]0, 1[$ ist, besagt (3.5), dass $\mathbf{f}(\mathbf{x})$ und $\mathbf{f}(\mathbf{y})$ einen kleineren Abstand

haben als $\mathbf{x}$ und $\mathbf{y}$. Die Funktion **f kontrahiert** also (d.h. zieht also zusammen).

Die beiden Abschätzungen (3.6) und (3.7) liefern jeweils eine **Fehlerabschätzung für den absoluten Fehler** $\|\mathbf{x}^{(j)} - \widehat{\mathbf{x}}\|$ der Näherung $\mathbf{x}^{(j)}$ von $\widehat{\mathbf{x}}$.

Dabei scheint die Fehlerabschätzung (3.7) auf den ersten Blick vielleicht „vorteilhafter“, da in ihr der absolute Fehler durch eine Konstante $\frac{q^j}{1-q}$, die (wegen $0 < q < 1$) für $j \to \infty$ gegen null strebt, mal den Abstand $\|\mathbf{x}^{(1)} - \mathbf{x}^{(0)}\|$ der beiden ersten Iterierten ausgedrückt wird. Man nennt dieses eine **a priori Fehlerabschätzung** („a priori” bedeutet „im Voraus“), weil die Angabe von einem Index j, bei dem mindestens die gewünschte Genauigkeit erreicht wird, keine Kenntnis von $\mathbf{x}^{(j)}$ und den vorherigen Iterierten $\mathbf{x}^{(j-1)}$ erfordert. Alle Größen auf der rechten Seite von(3.7) lassen sich mit der Kenntnis von $\mathbf{x}^{(0)}$, $\mathbf{x}^{(1)}$ und q berechnen.

Tatsächlich ist solch eine a priori Abschätzung aber in der Regel nicht so gut wie eine a posteriori Fehlerabschätzung, bei der der absolute Fehler $\|\mathbf{x}^{(j)} - \widehat{\mathbf{x}}\|$ mit einer Konstante (hier mit $\frac{q}{1-q}$) mal dem Abstand $\|\mathbf{x}^{(j)} - \mathbf{x}^{(j-1)}\|$ der aktuellen und der vorherigen Iterierten nach oben abgeschätzt wird. Man nennt dieses eine **a posteriori Fehlerabschätzung** („a posteriori“ bedeutet „im Nachhinein“), da man die obere Schranke für den absoluten Fehler $\|\mathbf{x}^{(j)} - \widehat{\mathbf{x}}\|$ erst ausrechnen kann, wenn man die Iterierten $\mathbf{x}^{(i)}$, $i = 0, 1, 2, \ldots, j$, berechnet hat. Anders ausgedrückt: Die obere Schranke in der a posteriori Fehlerabschätzung für den Fehler $\|\mathbf{x}^{(j)} - \widehat{\mathbf{x}}\|$ wird mit Hilfe der Iterierten $\mathbf{x}^{(j)}$ und $\mathbf{x}^{(j-1)}$ berechnet und liefert daher in der Regel eine **bessere Genauigkeit als die a priori Fehlerabschätzung**.

Was bedeutet die Kontraktionsbedingung in Fall einer Fixpunktiteration der Form (3.3)**?** Für $\mathbb{R}^n$ mit der Norm $\|\cdot\|$ und für die Funktion

$$\mathbf{f} : \mathbb{R}^n \to \mathbb{R}^n, \qquad \mathbf{f}(\mathbf{x}) = \mathbf{C}\,\mathbf{x} + \mathbf{c}, \tag{3.8}$$

finden wir

$$\begin{aligned} \|\mathbf{f}(\mathbf{x}) - \mathbf{f}(\mathbf{y})\| &= \big\|\big(\mathbf{C}\,\mathbf{x} + \mathbf{c}\big) - \big(\mathbf{C}\,\mathbf{y} + \mathbf{c}\big)\big\| \\ &= \|\mathbf{C}\,(\mathbf{x} - \mathbf{y})\| \le \|\mathbf{C}\|\,\|\mathbf{x} - \mathbf{y}\|, \end{aligned} \tag{3.9}$$

wobei $\|\mathbf{C}\|$ die von $\|\cdot\|$ induzierte Matrixnorm ist. Die durch (3.8) gegebene Funktion $\mathbf{f}$ erfüllt also die Kontraktionsbedingung (3.5), wenn es eine Norm $\|\cdot\|$ für $\mathbb{R}^n$ gibt, so dass in der induzierten Matrixnorm gilt $\|\mathbf{C}\| < 1$. (Da auf $\mathbb{R}^n$ alle Normen äquivalent sind (vgl. Hilfssatz 2.4), reicht es, die Kontraktionsbedingung in einer geeigneten Norm nachzuweisen.) Dieses ist im Wesentlichen die Aussage des nächsten Satzes, wobei die hinreichende Bedingung für die Kontraktionseigenschaft etwas allgemeiner angegeben werden kann.

Satz 3.4. (Konvergenz der Fixpunktiteration für $\mathbf{f}(\mathbf{x}) = \mathbf{C}\,\mathbf{x} + \mathbf{c}$)
Sei $\mathbf{f} : \mathbb{R}^n \to \mathbb{R}^n$, $\mathbf{f}(\mathbf{x}) = \mathbf{C}\,\mathbf{x} + \mathbf{c}$, *mit einer Matrix* $\mathbf{C} \in \mathbb{R}^{n\times n}$ *und einem Vektor* $\mathbf{c} \in \mathbb{R}^n$. ***Wenn*** $\varrho(\mathbf{C}) < 1$ ***ist, konvergiert die Fixpunktiteration***

$$\mathbf{x}^{(j)} = \mathbf{f}\big(\mathbf{x}^{(j-1)}\big) = \mathbf{C}\,\mathbf{x}^{(j-1)} + \mathbf{c}, \qquad j = 1, 2, \ldots,$$

für jeden Startvektor $\mathbf{x}^{(0)} \in \mathbb{R}^n$ ***gegen denselben Vektor*** $\widehat{\mathbf{x}}$. *Ist* $\varrho(\mathbf{C}) < 1$, *so ist der Grenzwert von* $\big(\mathbf{x}^{(j)}\big)_{j\in\mathbb{N}_0}$ *der* ***einzige Fixpunkt*** $\widehat{\mathbf{x}}$ ***von*** $\mathbf{f}$.

Wir erinnern uns, dass der **Spektralradius** $\varrho(\mathbf{C})$ einer quadratischen Matrix $\mathbf{C} \in \mathbb{R}^{n\times n}$ über den betraglich größten Eigenwert durch

$$\varrho(\mathbf{C}) = \max\big\{|\lambda| \;:\; \lambda \in \mathbb{C} \text{ mit } \mathbf{C}\,\mathbf{y} = \lambda\,\mathbf{y} \text{ für ein } \mathbf{y} \in \mathbb{C}^n \setminus \{\mathbf{0}\}\big\}$$

definiert ist.

Im nachfolgenden Teilkapitel werden wir uns für zwei verschiedene Wahlen von $\mathbf{C}$ in der Fixpunktiteration (siehe (3.3) und (3.2)) zur Lösung von $\mathbf{A}\,\mathbf{x} = \mathbf{b}$ interessieren und untersuchen, unter welchen Bedingungen uns Satz 3.4 die Konvergenz der Fixpunktiteration liefert. Genauer werden wir uns für verschiedene Wahlen der invertierbaren Matrix $\mathbf{B} \in \mathbb{R}^{n\times n}$ in (3.1) interessieren.

Betrachten wir zunächst ein Beispiel für die Anwendung von Satz 3.4.

Beispiel 3.5. (Fixpunktiteration für $\mathbf{f}(\mathbf{x}) = \mathbf{C}\,\mathbf{x} + \mathbf{c}$)
Gegeben sei die affin lineare Function

$$\mathbf{f} : \mathbb{R}^2 \to \mathbb{R}^2, \qquad \mathbf{f}(\mathbf{x}) = \mathbf{C}\,\mathbf{x} + \mathbf{c} = \underbrace{\begin{bmatrix} \frac{1}{4} & 1 \\ 0 & -\frac{1}{2} \end{bmatrix}}_{=\mathbf{C}} \mathbf{x} + \underbrace{\begin{bmatrix} -\frac{7}{4} \\ \frac{3}{2} \end{bmatrix}}_{=\mathbf{c}}.$$

Wir berechnen die Eigenwerte von $\mathbf{C}$, um den Spektralradius $\varrho(\mathbf{C})$ zu bestimmen: Das charakteristische Polynom ist

$$p_{\mathbf{C}}(\lambda) = \det\big(\mathbf{C} - \lambda\,\mathbf{E}_2\big) = \det\left(\begin{bmatrix} \frac{1}{4} - \lambda & 1 \\ 0 & -\frac{1}{2} - \lambda \end{bmatrix}\right) = \left(\frac{1}{4} - \lambda\right)\left(-\frac{1}{2} - \lambda\right).$$

Die Eigenwerte von $\mathbf{C}$ sind also $\lambda_1 = \frac{1}{4}$ und $\lambda_2 = -\frac{1}{2}$. Damit folgt

$$\varrho(\mathbf{C}) = \max\left\{\left|\frac{1}{4}\right|; \left|-\frac{1}{2}\right|\right\} = \frac{1}{2} < 1,$$

und nach Satz 3.4 konvergiert die Fixpunktiteration $\big(\mathbf{x}^{(j)}\big)_{j\in\mathbb{N}_0}$ mit $\mathbf{x}^{(j)} = \mathbf{f}\big(\mathbf{x}^{(j-1)}\big)$ für jeden Startvektor $\mathbf{x}^{(0)}$ gegen den einzigen Fixpunkt $\widehat{\mathbf{x}}$ von $\mathbf{f}$.

Wir berechnen nun den einzigen Fixpunkt $\widehat{\mathbf{x}}$ direkt: Aus $\mathbf{f}(\widehat{\mathbf{x}}) = \widehat{\mathbf{x}}$ folgt

$$\mathbf{C}\,\widehat{\mathbf{x}} + \mathbf{c} = \widehat{\mathbf{x}} \qquad \Longleftrightarrow \qquad \mathbf{C}\,\widehat{\mathbf{x}} - \widehat{\mathbf{x}} = -\mathbf{c} \qquad \Longleftrightarrow \qquad \big(\mathbf{C} - \mathbf{E}_2\big)\,\widehat{\mathbf{x}} = -\mathbf{c}.$$

Wir lösen also das lineare Gleichungssystem $\big(\mathbf{C} - \mathbf{E}_2\big)\,\widehat{\mathbf{x}} = -\mathbf{c}$, um den einzigen Fixpunkt $\widehat{\mathbf{x}}$ von $\mathbf{f}$ zu bestimmen:

$$\big[\mathbf{C} - \mathbf{E}_2 \big| -\mathbf{c}\big] \quad \Longleftrightarrow \quad \left[\begin{array}{cc|c} \frac{1}{4}-1 & 1 & \frac{7}{4} \\ 0 & -\frac{1}{2}-1 & -\frac{3}{2} \end{array}\right] \quad \Longleftrightarrow \quad \left[\begin{array}{cc|c} -\frac{3}{4} & 1 & \frac{7}{4} \\ 0 & -\frac{3}{2} & -\frac{3}{2} \end{array}\right]$$

$$\overset{\substack{Z_1\to -\frac{4}{3}Z_1 \\ Z_2\to -\frac{2}{3}Z_2 \\ \downarrow \\ \Longleftrightarrow}}{} \left[\begin{array}{cc|c} 1 & -\frac{4}{3} & -\frac{7}{3} \\ 0 & 1 & 1 \end{array}\right] \quad \overset{\substack{Z_1\to Z_1+\frac{4}{3}Z_2 \\ \downarrow \\ \Longleftrightarrow}}{} \left[\begin{array}{cc|c} 1 & 0 & -1 \\ 0 & 1 & 1 \end{array}\right]$$

Der einzige Fixpunkt ist also $\widehat{\mathbf{x}} = \begin{bmatrix} -1 \\ 1 \end{bmatrix}$.

Für den Startvektor $\mathbf{x}^{(0)} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$ berechnen wir die ersten vier Iterierten:

$$\mathbf{x}^{(1)} = \mathbf{f}\big(\mathbf{x}^{(0)}\big) = \begin{bmatrix} \frac{1}{4} & 1 \\ 0 & -\frac{1}{2} \end{bmatrix} \begin{bmatrix} 0 \\ 0 \end{bmatrix} + \begin{bmatrix} -\frac{7}{4} \\ \frac{3}{2} \end{bmatrix} = \begin{bmatrix} -\frac{7}{4} \\ \frac{3}{2} \end{bmatrix},$$

$$\mathbf{x}^{(2)} = \mathbf{f}\big(\mathbf{x}^{(1)}\big) = \begin{bmatrix} \frac{1}{4} & 1 \\ 0 & -\frac{1}{2} \end{bmatrix} \begin{bmatrix} -\frac{7}{4} \\ \frac{3}{2} \end{bmatrix} + \begin{bmatrix} -\frac{7}{4} \\ \frac{3}{2} \end{bmatrix} = \begin{bmatrix} \frac{1}{4}\cdot\big(-\frac{7}{4}\big) + 1\cdot\frac{3}{2} - \frac{7}{4} \\ 0 + \big(-\frac{1}{2}\big)\cdot\frac{3}{2} + \frac{3}{2} \end{bmatrix} = \begin{bmatrix} -\frac{11}{16} \\ \frac{3}{4} \end{bmatrix},$$

$$\mathbf{x}^{(3)} = \mathbf{f}\big(\mathbf{x}^{(2)}\big) = \begin{bmatrix} \frac{1}{4} & 1 \\ 0 & -\frac{1}{2} \end{bmatrix} \begin{bmatrix} -\frac{11}{16} \\ \frac{3}{4} \end{bmatrix} + \begin{bmatrix} -\frac{7}{4} \\ \frac{3}{2} \end{bmatrix} = \begin{bmatrix} \frac{1}{4}\cdot\big(-\frac{11}{16}\big) + 1\cdot\frac{3}{4} - \frac{7}{4} \\ 0 + \big(-\frac{1}{2}\big)\cdot\frac{3}{4} + \frac{3}{2} \end{bmatrix} = \begin{bmatrix} -\frac{75}{64} \\ \frac{9}{8} \end{bmatrix},$$

$$\mathbf{x}^{(4)} = \mathbf{f}\big(\mathbf{x}^{(3)}\big) = \begin{bmatrix} \frac{1}{4} & 1 \\ 0 & -\frac{1}{2} \end{bmatrix} \begin{bmatrix} -\frac{75}{64} \\ \frac{9}{8} \end{bmatrix} + \begin{bmatrix} -\frac{7}{4} \\ \frac{3}{2} \end{bmatrix} = \begin{bmatrix} \frac{1}{4}\cdot\big(-\frac{75}{64}\big) + 1\cdot\frac{9}{8} - \frac{7}{4} \\ 0 + \big(-\frac{1}{2}\big)\cdot\frac{9}{8} + \frac{3}{2} \end{bmatrix} = \begin{bmatrix} -\frac{235}{256} \\ \frac{15}{16} \end{bmatrix}.$$

Nach vier Iterationsschritten erhalten wir also die Näherung

$$\mathbf{x}^{(4)} = \begin{bmatrix} -\frac{235}{256} \\ \frac{15}{16} \end{bmatrix} \doteq \begin{bmatrix} -0{,}91797 \\ 0{,}93750 \end{bmatrix}$$

für den Fixpunkt $\widehat{\mathbf{x}} = \begin{bmatrix} -1 \\ 1 \end{bmatrix}$. ♠

Beweisskizze von Satz 3.4: Sei $\varrho(\mathbf{C}) < 1$. Man kann zeigen (dieses ist nicht offensichtlich!), dass es zu jedem Abstand $\epsilon > 0$ eine Norm auf $\mathbb{R}^n$ gibt, so dass für die induzierte Matrixnorm auf $\mathbb{R}^{n\times n}$ gilt

$$\|\mathbf{C}\| \leq \varrho(\mathbf{C}) + \epsilon \qquad \text{für alle } \mathbf{C} \in \mathbb{R}^{n\times n}.$$

Da $\varrho(\mathbf{C}) < 1$ ist, kann man das $\epsilon > 0$ so klein wählen, dass auch $\varrho(\mathbf{C}) + \epsilon < 1$ ist. Dann folgt also $\|\mathbf{C}\| \leq \varrho(\mathbf{C}) + \epsilon < 1$, und nach (3.9) ist $\mathbf{f}$ ein Kontraktion. Satz 3.3 liefert, dann dass $\mathbf{f}$ genau einen Fixpunkt $\widehat{\mathbf{x}}$ hat und dass die Fixpunktiteration $\left(\mathbf{x}^{(j)}\right)_{j\in\mathbb{N}_0}$ für jeden Startvektor $\mathbf{x}^{(0)}$ gegen den einzigen Fixpunkt $\widehat{\mathbf{x}}$ konvergiert. □

Zum Abschluss skizzieren wir kurz den Beweis des Banachschen Fixpunktsatzes. Dieser Beweis wird nicht in der Vorlesung besprochen und ist für mathematisch Interessierte ins Skript aufgenommen.

Beweisskizze von Satz 3.3: Das Kernstück des Beweises ist die Kontraktionseigenschaft (3.5). Insbesondere folgt aus (3.5), dass $\mathbf{f}$ eine stetige Funktion ist.

Zunächst bemerken wir, dass für jeden Startvektor $\mathbf{x}^{(0)} \in D$ folgt, dass alle Ierierten $\mathbf{x}^{(j)}$ ebenfalls in D liegen, denn $\mathbf{x}^{(j)}$ erhält man aus $\mathbf{x}^{(0)}$ durch j-malige Anwendung von $\mathbf{f}$, und das Bild von $\mathbf{f}$ liegt per Definition in D. Mit Hilfe der Kontraktionseigenschaft (3.5) kann man zeigen, dass für alle $j \in \mathbb{N}$ gilt:

$$\|\mathbf{x}^{(j+k)} - \mathbf{x}^{(j)}\| \leq \frac{q^j}{1-q} \|\mathbf{x}^{(1)} - \mathbf{x}^{(0)}\| \qquad \text{für alle } k \in \mathbb{N}. \tag{3.10}$$

Weil $q \in]0,1[$ ist und somit q^j für j groß genug beliebig dicht bei null liegt, folgt aus (3.10), dass $\left(\mathbf{x}^{(j)}\right)_{j\in\mathbb{N}_0}$ eine sogenannte Cauchyfolge ist. Da $D \subseteq \mathbb{R}^n$ abgeschlossen ist, folgt, dass die Cauchyfolge $\left(\mathbf{x}^{(j)}\right)_{j\in\mathbb{N}_0}$ gegen einen Grenzwert $\widehat{\mathbf{x}} \in D$ konvergiert. Dieser Grenzwert $\widehat{\mathbf{x}}$ ist ein Fixpunkt von $\mathbf{f}$, denn aus der rekursiven Beziehung $\mathbf{x}^{(j+1)} = \mathbf{f}\left(\mathbf{x}^{(j)}\right)$ folgt durch Grenzwertbildung

$$\widehat{\mathbf{x}} = \lim_{j\to\infty} \mathbf{x}^{(j+1)} = \lim_{j\to\infty} \mathbf{f}\left(\mathbf{x}^{(j)}\right) = \mathbf{f}\left(\lim_{j\to\infty} \mathbf{x}^{(j)}\right) = \mathbf{f}(\widehat{\mathbf{x}}),$$

wobei wir im vorletzten Schritt die Stetigkeit von $\mathbf{f}$ genutzt haben.

(1) Wir wissen aus den obigen Überlegungen bereits, dass die Folge $\left(\mathbf{x}^{(j)}\right)_{j\in\mathbb{N}_0}$ für jeden festen Startwert $\mathbf{x}^{(0)}$ konvergiert und dass ihr Grenzwert $\widehat{\mathbf{x}}$ ein Fixpunkt von $\mathbf{f}$ ist. Also hat $\mathbf{f}$ mindestens einen Fixpunkt. Seien nun $\widehat{\mathbf{x}}, \widehat{\mathbf{y}} \in D$ zwei Fixpunkte von $\mathbf{f}$. Dann folgt mit $\widehat{\mathbf{x}} = \mathbf{f}(\widehat{\mathbf{x}})$ und $\widehat{\mathbf{y}} = \mathbf{f}(\widehat{\mathbf{y}})$ aus der Kontraktionseigenschaft

$$\|\widehat{\mathbf{x}} - \widehat{\mathbf{y}}\| = \|\mathbf{f}(\widehat{\mathbf{x}}) - \mathbf{f}(\widehat{\mathbf{y}})\| \leq q \|\widehat{\mathbf{x}} - \widehat{\mathbf{y}}\|.$$

Da $q \in]0,1[$ ist, kann diese Gleichung nur dann gelten, wenn $\|\widehat{\mathbf{x}} - \widehat{\mathbf{y}}\| = 0$ ist, also wenn $\widehat{\mathbf{x}} = \widehat{\mathbf{y}}$ gilt. Also hat $\mathbf{f}$ genau einen Fixpunkt, den wir im Folgenden immer mit $\widehat{\mathbf{x}}$ bezeichnen.

(2) Aus unseren Vorüberlegungen wissen wir bereits, die Folge $\big(\mathbf{x}^{(j)}\big)_{j\in\mathbb{N}_0}$ für jeden festen Startwert $\mathbf{x}^{(0)}$ konvergiert und dass ihr Grenzwert $\widehat{\mathbf{x}} = \lim_{j\to\infty} \mathbf{x}^{(j)}$ ein Fixpunkt von $\mathbf{f}$ ist. Da $\mathbf{f}$ nach (1) nur einen Fixpunkt $\widehat{\mathbf{x}}$ hat, muss die Folge $\big(\mathbf{x}^{(j)}\big)_{j\in\mathbb{N}_0}$ für jeden festen Startwert $\mathbf{x}^{(0)}$ gegen den einzigen Fixpunkt $\widehat{\mathbf{x}}$ konvergieren.

(3) Nachweis der a posteriori Fehlerabschätzung: Wegen der Kontraktionseigenschaft (3.5) gilt (mit $\mathbf{x}^{(j)} = \mathbf{f}(\mathbf{x}^{(j-1)})$ und $\mathbf{f}(\widehat{\mathbf{x}}) = \widehat{\mathbf{x}}$)

$$\begin{aligned}\|\mathbf{x}^{(j)} - \widehat{\mathbf{x}}\| = \|\mathbf{f}\big(\mathbf{x}^{(j-1)}\big) - \mathbf{f}(\widehat{\mathbf{x}})\| &\le q\,\|\mathbf{x}^{(j-1)} - \widehat{\mathbf{x}}\| \\ &\le q\,\|(\mathbf{x}^{(j-1)} - \mathbf{x}^{(j)}) + (\mathbf{x}^{(j)} - \widehat{\mathbf{x}})\| \\ &\le q\,\big(\|\mathbf{x}^{(j-1)} - \mathbf{x}^{(j)}\| + \|\mathbf{x}^{(j)} - \widehat{\mathbf{x}}\|\big),\end{aligned}$$

wobei wir im letzten Schritt die Dreiecksungleichung genutzt haben. Subtrahieren von $q\,\|\mathbf{x}^{(j)} - \widehat{\mathbf{x}}\|$ auf beiden Seiten liefert

$$\begin{aligned}&(1-q)\,\|\mathbf{x}^{(j)} - \widehat{\mathbf{x}}\| \le q\,\|\mathbf{x}^{(j-1)} - \mathbf{x}^{(j)}\| = q\,\|\mathbf{x}^{(j)} - \mathbf{x}^{(j-1)}\| \\ \iff \qquad &\|\mathbf{x}^{(j)} - \widehat{\mathbf{x}}\| \le \frac{q}{1-q}\,\|\mathbf{x}^{(j)} - \mathbf{x}^{(j-1)}\|, \qquad (3.11)\end{aligned}$$

und wir haben die a posteriori Fehlerabschätzung bewiesen.

(4) Nachweis der a priori Fehlerabschätzung: Wir beginnen unsere Argumentation mit der a posteriori Fehlerabschätzung (3.11) und wenden die Kontraktionseigenschaft (3.5) wiederholt an: Mit $\mathbf{x}^{(k)} = \mathbf{f}(\mathbf{x}^{(k-1)})$, $k = 1, 2, \ldots, j$, folgt dann

$$\begin{aligned}\|\mathbf{x}^{(j)} - \widehat{\mathbf{x}}\| &\le \frac{q}{1-q}\,\|\mathbf{x}^{(j)} - \mathbf{x}^{(j-1)}\| = \frac{q}{1-q}\,\|\mathbf{f}\big(\mathbf{x}^{(j-1)}\big) - \mathbf{f}\big(\mathbf{x}^{(j-2)}\big)\| \\ &\le \frac{q^2}{1-q}\,\|\mathbf{x}^{(j-1)} - \mathbf{x}^{(j-2)}\| = \frac{q^2}{1-q}\,\|\mathbf{f}\big(\mathbf{x}^{(j-2)}\big) - \mathbf{f}\big(\mathbf{x}^{(j-3)}\big)\| \\ &\le \frac{q^3}{1-q}\,\|\mathbf{x}^{(j-2)} - \mathbf{x}^{(j-3)}\| \le \ldots \le \frac{q^j}{1-q}\,\|\mathbf{x}^{(1)} - \mathbf{x}^{(0)}\|,\end{aligned}$$

womit die a priori Fehlerabschätzung bewiesen ist.

Damit ist der Banachsche Fixpunktsatz bewiesen. □

3.2 Jacobi-Verfahren und Gauß-Seidel-Verfahren

Wir kommen nun zu der Ausgangsfrage des vorigen Kapitels zurück: Gelöst werden soll das lineare Gleichungssystem $\mathbf{A}\,\mathbf{x} = \mathbf{b}$, wobei $\mathbf{A} \in \mathbb{R}^{n\times n}$ **regulär** (also **invertierbar**) ist und $\mathbf{b} \in \mathbb{R}^n$ ist. Das lineare Gleichungssystem $\mathbf{A}\,\mathbf{x} = \mathbf{b}$ hat also n Gleichungen und n Unbekannte und besitzt eine **eindeutige Lösung** $\widehat{\mathbf{x}} \in \mathbb{R}^n$. Durch geeignetes Umformen (siehe (3.1)) hatten wir das lineare Gleichungssystem $\mathbf{A}\,\mathbf{x} = \mathbf{b}$ in die Fixpunktgleichung

$$\boxed{\mathbf{x} = \underbrace{-\mathbf{B}^{-1}\,(\mathbf{A}-\mathbf{B})}_{=\mathbf{C}}\,\mathbf{x} + \underbrace{\mathbf{B}^{-1}\,\mathbf{b}}_{=\mathbf{c}}} \tag{3.12}$$

umgewandelt, wobei $\mathbf{B} \in \mathbb{R}^{n\times n}$ eine **reguläre** (also invertierbare) **Matrix** ist, die eine **(leichter zu invertierende) Näherung für A** darstellt. Die zu (3.12) gehörende Fixpunktiteration lautet also:

$$\boxed{\mathbf{x}^{(j+1)} = \mathbf{f}\big(\mathbf{x}^{(j)}\big) = \underbrace{-\mathbf{B}^{-1}\,(\mathbf{A}-\mathbf{B})}_{=\mathbf{C}}\,\mathbf{x}^{(j)} + \underbrace{\mathbf{B}^{-1}\,\mathbf{b}}_{=\mathbf{c}}} \tag{3.13}$$

mit $\mathbf{f} : \mathbb{R}^n \to \mathbb{R}^n$, $\mathbf{f}(\mathbf{x}) := -\mathbf{B}^{-1}\,(\mathbf{A}-\mathbf{B}) + \mathbf{B}^{-1}\,\mathbf{b}$.

Von nun an nehmen wir an, dass **alle Diagonaleinträge der regulären Matrix** $\mathbf{A} = [a_{j,k}] \in \mathbb{R}^{n\times n}$ **von null verschieden sind**, also $a_{j,j} \neq 0$ für alle $j = 1, 2, \ldots, n$. Erfüllt eine reguläre Matrix $\mathbf{A}$ diese Bedingung nicht, so kann man durch passende Zeilentauschungen und/oder Spaltentauschungen immer erreichen, dass die Diagonaleinträge ungleich null sind, denn $\mathbf{A}$ hat n linear unabhängige Spaltenvektoren.

Wir zerlegen $\mathbf{A}$ nun in die **Diagonalmatrix D** (den diagonalen Anteil von $\mathbf{A}$), die **streng untere Linksdreiecksmatrix L** (den Anteil von von $\mathbf{A}$ links unterhalb der Diagonalen) und die **streng obere Rechtsdreiecksmatrix R** (den Anteil von $\mathbf{A}$ rechts oberhalb der Diagonalen):

$$\mathbf{A} = \begin{bmatrix} a_{1,1} & a_{1,2} & \cdots & \cdots & a_{1,n} \\ a_{2,1} & a_{2,2} & a_{2,3} & \cdots & a_{2,n} \\ a_{3,1} & a_{3,2} & a_{3,3} & \ddots & \vdots \\ \vdots & & \ddots & \ddots & a_{n-1,n} \\ a_{n,1} & \cdots & \cdots & a_{n,n-1} & a_{n,n} \end{bmatrix} = \mathbf{L} + \mathbf{D} + \mathbf{R} \qquad \text{mit} \tag{3.14}$$

$$\mathbf{L} = \begin{bmatrix} 0 & 0 & \cdots & 0 \\ a_{2,1} & 0 & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ a_{n,1} & \cdots & a_{n,n-1} & 0 \end{bmatrix}, \quad \mathbf{D} = \begin{bmatrix} a_{1,1} & 0 & \cdots & 0 \\ 0 & a_{2,2} & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ 0 & \cdots & 0 & a_{n,n} \end{bmatrix}, \quad \mathbf{R} = \begin{bmatrix} 0 & a_{1,2} & \cdots & a_{1,n} \\ 0 & 0 & \ddots & \vdots \\ \vdots & \ddots & \ddots & a_{n-1,n} \\ 0 & \cdots & 0 & 0 \end{bmatrix}$$

Da wir vorausgesetzt haben, dass $\mathbf{A}$ nur Diagonaleinträge ungleich null hat, ist die **Matrix $\mathbf{D}$ invertierbar**. Die Wahl $\mathbf{B} = \mathbf{D}$ als (leicht invertierbare) Näherung von $\mathbf{A}$ stellt die einfachste Möglichkeit da, um eine Fixpunktgleichung (3.12) zu erhalten. Dann gilt mit $\mathbf{B} = \mathbf{D}$ nach (3.12) und (3.14) für die Matrix $\mathbf{C}$

$$\mathbf{C} = -\mathbf{D}^{-1}\,(\mathbf{A} - \mathbf{D}) = -\mathbf{D}^{-1}\,\big((\mathbf{L} + \mathbf{D} + \mathbf{R}) - \mathbf{D}\big) = -\mathbf{D}^{-1}\,(\mathbf{L} + \mathbf{R}),$$

und die zugehörige Fixpunktiteration (3.13) lautet

$$\boxed{\mathbf{x}^{(j+1)} = -\mathbf{D}^{-1}\,(\mathbf{L} + \mathbf{R})\,\mathbf{x}^{(j)} + \mathbf{D}^{-1}\,\mathbf{b} = \mathbf{D}^{-1}\,\big(\mathbf{b} - (\mathbf{L} + \mathbf{R})\,\mathbf{x}^{(j)}\big).} \tag{3.15}$$

Da die Matrix $\mathbf{D}$ eine reguläre Diagonalmatrix ist, kann man ihre Inverse $\mathbf{D}^{-1}$ direkt durch bilden des Kehrwerts der Diagonaleinträge bestimmen, also

$$\mathbf{D} = \begin{bmatrix} a_{1,1} & 0 & \cdots & 0 \\ 0 & a_{2,2} & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ 0 & \cdots & 0 & a_{n,n} \end{bmatrix} \quad \Longrightarrow \quad \mathbf{D}^{-1} = \begin{bmatrix} a_{1,1}^{-1} & 0 & \cdots & 0 \\ 0 & a_{2,2}^{-1} & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ 0 & \cdots & 0 & a_{n,n}^{-1} \end{bmatrix}. \tag{3.16}$$

Nach (3.15) berechnen sich die Komponenten von von $\mathbf{x}^{(j+1)}$ damit wie folgt:

$$\boxed{x_i^{(j+1)} = \frac{1}{a_{i,i}} \left(b_i - \sum_{\substack{k=1,\\ k \neq i}}^{n} a_{i,k}\, x_k^{(j)} \right), \qquad i = 1, 2, \ldots, n.}$$

Wir halten die hergeleitete Fixpunktiteration zur Lösung von $\mathbf{A}\,\mathbf{x} = \mathbf{b}$ als Iterationsverfahren fest.

Verfahren 3.6. (Jacobi-Verfahren)
*Sei $\mathbf{A} = [a_{i,j}] \in \mathbb{R}^{n \times n}$ eine **reguläre** (d.h. invertierbare) Matrix, deren **Diagonalelemente alle von null verschieden** sind (d.h. $a_{j,j} \neq 0$ für $j = 1, 2, \ldots, n$). Sei $\mathbf{b} \in \mathbb{R}^n$, und sei $\mathbf{x}^{(0)} \in \mathbb{R}^n$ ein Startvektor. Die Fixpunktiteration $\big(\mathbf{x}^{(j)}\big)_{j \in \mathbb{N}_0}$ mit $\mathbf{x}^{(j+1)} = \mathbf{D}^{-1}\,\big(\mathbf{b} - (\mathbf{L} + \mathbf{R})\,\mathbf{x}^{(j)}\big)$, also*

$$x_i^{(j+1)} = \frac{1}{a_{i,i}} \left(b_i - \sum_{\substack{k=1,\\ k \neq i}}^{n} a_{i,k}\, x_k^{(j)} \right), \qquad i = 1, 2, \ldots, n, \tag{3.17}$$

*heißt das **Jacobi-Verfahren** oder **Gesamtschrittverfahren**.*

Aus Satz 3.4 wissen wir bereits, dass die Fixpunktiteration genau dann konvergiert, wenn $\varrho\big(-\mathbf{D}^{-1}(\mathbf{L}+\mathbf{R})\big) < 1$ ist. Sicherlich kann man dieses nur erwarten, wenn $\mathbf{D}$ eine hinreichend gute Näherung von $\mathbf{A}$ ist, und dieses ist nur der Fall wenn die Diagonaleinträge von $\mathbf{A}$ die Matrix $\mathbf{A}$ in einem gewissen Sinne dominieren. Die nächste Definition präzisiert dieses.

Definition 3.7. (streng diagonal dominant)
Eine Matrix $\mathbf{A} = [a_{j,k}] \in \mathbb{R}^{n\times n}$ *heißt* ***streng diagonal dominant****, wenn gilt*

$$\sum_{\substack{k=1,\\ k\neq j}}^{n} |a_{j,k}| < |a_{j,j}| \qquad \textit{für alle } j = 1, 2, \ldots, n. \tag{3.18}$$

(In Worten: Der Absolutbetrag jedes Diagonaleintrags von $\mathbf{A}$ *ist größer als die Summe der Absolutbeträge aller übrigen Einträge in der gleichen Zeile.)*

Beispiel 3.8. (streng diagonal dominante Matrix)
Die Matrix

$$\mathbf{A} = [a_{j,k}] = \begin{bmatrix} 1 & \frac{1}{2} & \frac{1}{3} \\ -1 & 2 & \frac{1}{2} \\ \frac{1}{2} & \frac{1}{4} & \frac{5}{6} \end{bmatrix}$$

ist streng diagonal dominant, da

$$|a_{1,2}| + |a_{1,3}| = \frac{1}{2} + \frac{1}{3} = \frac{5}{6} < 1 = |a_{1,1}|,$$
$$|a_{2,1}| + |a_{2,3}| = |-1| + \frac{1}{2} = \frac{3}{2} < 2 = |a_{2,2}|,$$
$$|a_{3,1}| + |a_{3,2}| = \frac{1}{2} + \frac{1}{4} = \frac{3}{4} = \frac{9}{12} < \frac{10}{12} = \frac{5}{6} = |a_{3,3}|$$

gelten. ♠

Der nächste Satz formuliert eine hinreichende Bedingung für die Konvergenz des Jacobi-Verfahrens.

Satz 3.9. (Konvergenz des Jacobi-Verfahrens)
Seien die Notation und die Voraussetzungen wie in Verfahren 3.6. Wenn die reguläre Matrix $\mathbf{A} \in \mathbb{R}^{n\times n}$ ***streng diagonal dominant*** *ist, dann* ***konvergiert***

das Jacobi-Verfahren (3.17) ***für jeden Startvektor*** $\mathbf{x}^{(0)} \in \mathbb{R}^n$ *gegen die eindeutige Lösung* $\widehat{\mathbf{x}}$ *von* $\mathbf{A}\,\mathbf{x} = \mathbf{b}$.

Beweis von Satz 3.9: Sei $\mathbf{c} := \mathbf{D}^{-1}\,\mathbf{b}$, und sei $\mathbf{C} := -\mathbf{D}^{-1}\,(\mathbf{L}+\mathbf{R})$ die Matrix der Fixpunktiteration $\mathbf{x}^{(j+1)} = \mathbf{C}\,\mathbf{x}^{(j)} + \mathbf{c}$, $j = 0, 1, 2, \ldots$, des Jacobi-Verfahrens. Dann können wir $\mathbf{C}$ unter Nutzung von (3.16) explizit angeben:

$$\mathbf{C} = -\mathbf{D}^{-1}\,(\mathbf{L}+\mathbf{R}) = -\mathbf{D}^{-1}\,(\mathbf{A}-\mathbf{D})$$

$$= -\begin{bmatrix} 0 & a_{1,1}^{-1}\,a_{1,2} & \cdots & \cdots & a_{1,1}^{-1}\,a_{1,n} \\ a_{2,2}^{-1}\,a_{2,1} & 0 & a_{2,2}^{-1}\,a_{2,3} & \cdots & a_{2,2}^{-1}\,a_{2,n} \\ \vdots & a_{3,3}^{-1}\,a_{3,2} & 0 & \ddots & \vdots \\ \vdots & & \ddots & \ddots & a_{n-1,n-1}^{-1}\,a_{n-1,n} \\ a_{n,n}^{-1}\,a_{n,1} & \cdots & \cdots & a_{n,n}^{-1}\,a_{n,n-1} & 0 \end{bmatrix}.$$

Es gilt $\mathbf{x}^{(j+1)} = \mathbf{f}(\mathbf{x}^{(j)})$ mit $\mathbf{f} : \mathbb{R}^n \to \mathbb{R}^n$, $\mathbf{f}(\mathbf{x}) := \mathbf{C}\,\mathbf{x}+\mathbf{c}$, und nach Satz 3.3 reicht es für die Konvergenz der Fixpunktiteration zu zeigen, dass $\mathbf{f}$ eine Kontraktion ist. Ist $\|\cdot\|$ eine Norm für $\mathbb{R}^n$ und bezeichnet $\|\cdot\|$ ebenfalls die zugehörige induzierte Matrixnorm, so gilt für alle $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$

$$\begin{aligned} \|\mathbf{f}(\mathbf{x}) - \mathbf{f}(\mathbf{y})\| &= \big\|\big(\mathbf{C}\,\mathbf{x}+\mathbf{c}\big) - \big(\mathbf{C}\,\mathbf{y}+\mathbf{c}\big)\big\| \\ &= \|\mathbf{C}\,\big(\mathbf{x}-\mathbf{y}\big)\| \le \|\mathbf{C}\|\,\|\mathbf{x}-\mathbf{y}\|. \end{aligned} \tag{3.19}$$

Da alle Normen auf $\mathbb{R}^n$ nach Hilfssatz 2.4 äquivalent sind reicht es die Konvergenz der Fixpunktiteration in einer Norm unserer Wahl zu zeigen. Es reicht daher, wenn wir für eine Norm unserer Wahl zeigen, dass $\mathbf{f}$ eine Kontraktion ist, indem wir für die zugehörige induzierte Matrixnorm in (3.19) die Bedingung $\|\mathbf{C}\| < 1$ nachweisen. Wir nutzen nun die von der ∞-Norm induzierte Zeilensummennorm

$$\|\mathbf{C}\|_\infty = \max_{j=1,\ldots,n} \left(\sum_{k=1}^{n} |c_{j,k}| \right) = \max_{j=1,\ldots,n} \left(\sum_{\substack{k=1,\\ k\neq j}}^{n} \left| \frac{-a_{j,k}}{a_{j,j}} \right| \right)$$

$$= \max_{j=1,\ldots,n} \left(\sum_{\substack{k=1,\\ k\neq j}}^{n} \frac{|a_{j,k}|}{|a_{j,j}|} \right) = \max_{j=1,\ldots,n} \left(\frac{1}{|a_{j,j}|} \sum_{\substack{k=1,\\ k\neq j}}^{n} |a_{j,k}| \right).$$

Falls also gilt

$$\|\mathbf{C}\|_\infty = \max_{j=1,\ldots,n} \left(\frac{1}{|a_{j,j}|} \sum_{\substack{k=1,\\ k\neq j}}^{n} |a_{j,k}| \right) < 1$$

$$\iff \qquad \frac{1}{|a_{j,j}|} \sum_{\substack{k=1,\\ k\neq j}}^{n} |a_{j,k}| < 1 \quad \text{für alle } j = 1, \ldots, n$$

$$\iff \qquad \sum_{\substack{k=1,\\ k\neq j}}^{n} |a_{j,k}| < |a_{j,j}| \quad \text{für alle } j = 1, \ldots, n, \tag{3.20}$$

so ist $\mathbf{f}$ eine Kontraktion und die Fixpunktiteration, also das Jacobi-Verfahren, konvergiert. Die letzte Aussage in (3.20) besagt aber gerade, dass $\mathbf{A}$ streng diagonal dominant ist. Also konvergiert das Jacobi-Verfahren, wenn $\mathbf{A}$ streng diagonal dominant ist. □

Betrachten wir ein Beispiel.

Beispiel 3.10. (Jacobi-Verfahren)
Gegeben Sei das lineare Gleichungssystem

$$\mathbf{A}\,\mathbf{x} = \mathbf{b}, \qquad \text{mit} \qquad \mathbf{A} = \begin{bmatrix} 2 & \frac{1}{2} & \frac{1}{2} \\ 1 & 3 & 1 \\ 2 & 0 & 3 \end{bmatrix}, \qquad \mathbf{b} = \begin{bmatrix} \frac{3}{2} \\ -2 \\ 2 \end{bmatrix}.$$

Durch Berechnung der Determinante überprüft man leicht, dass $\mathbf{A}$ regulär ist:

$$\det(\mathbf{A}) = 12 + 1 + 0 - 3 - 0 - \frac{3}{2} = \frac{17}{2} \neq 0.$$

Die eindeutige Lösung des gegebenen LGS ist $\widehat{\mathbf{x}} = \begin{bmatrix} 1 \\ -1 \\ 0 \end{bmatrix}$, wie man leicht durch Berechnen von $\mathbf{A}\,\widehat{\mathbf{x}}$ überprüft. Die Matrix $\mathbf{A}$ ist streng diagonal dominant, denn

$$\frac{1}{2} + \frac{1}{2} = 1 < 2 = |a_{1,1}|, \quad 1 + 1 = 2 < 3 = |a_{2,2}| \quad \text{und} \quad 2 + 0 = 2 < 3 = |a_{3,3}|.$$

Also wird das Jacobi-Verfahren nach Satz 3.9 für jeden Startvektor $\mathbf{x}^{(0)}$ gegen die eindeutig bestimmte Lösung $\widehat{\mathbf{x}}$ von $\mathbf{A}\,\mathbf{x} = \mathbf{0}$ konvergieren, und wir können das Jacobi-Verfahren zur Berechnung einer (Näherungs-)Lösung von $\mathbf{A}\,\mathbf{x} = \mathbf{b}$ anwenden. Wir berechnen mit der Formel (3.17) die ersten vier Iterierten des Jacobi-Verfahrens mit dem Startvektor $\mathbf{x}^{(0)} = \mathbf{0}$:

$$\left.\begin{aligned} x_1^{(1)} &= \tfrac{1}{2}\left(\tfrac{3}{2} - \tfrac{1}{2}\cdot 0 - \tfrac{1}{2}\cdot 0\right) = \tfrac{3}{4}, \\ x_2^{(1)} &= \tfrac{1}{3}\left(-2 - 1\cdot 0 - 1\cdot 0\right) = -\tfrac{2}{3}, \\ x_3^{(1)} &= \tfrac{1}{3}\left(2 - 2\cdot 0 - 0\cdot 0\right) = \tfrac{2}{3} \end{aligned}\right\} \quad \Longrightarrow \quad \mathbf{x}^{(1)} = \begin{bmatrix} \frac{3}{4} \\ -\frac{2}{3} \\ \frac{2}{3} \end{bmatrix},$$

$$\left.\begin{aligned} x_1^{(2)} &= \tfrac{1}{2}\left(\tfrac{3}{2} - \tfrac{1}{2}\cdot\left(-\tfrac{2}{3}\right) - \tfrac{1}{2}\cdot\tfrac{2}{3}\right) = \tfrac{3}{4}, \\ x_2^{(2)} &= \tfrac{1}{3}\left(-2 - 1\cdot\tfrac{3}{4} - 1\cdot\tfrac{2}{3}\right) = -\tfrac{41}{36}, \\ x_3^{(2)} &= \tfrac{1}{3}\left(2 - 2\cdot\tfrac{3}{4} - 0\cdot\left(-\tfrac{2}{3}\right)\right) = \tfrac{1}{6} \end{aligned}\right\} \quad\Longrightarrow\quad \mathbf{x}^{(2)} = \begin{bmatrix} \frac{3}{4} \\ -\frac{41}{36} \\ \frac{1}{6} \end{bmatrix},$$

$$\left.\begin{aligned} x_1^{(3)} &= \tfrac{1}{2}\left(\tfrac{3}{2} - \tfrac{1}{2}\cdot\left(-\tfrac{41}{36}\right) - \tfrac{1}{2}\cdot\tfrac{1}{6}\right) = \tfrac{143}{144}, \\ x_2^{(3)} &= \tfrac{1}{3}\left(-2 - 1\cdot\tfrac{3}{4} - 1\cdot\tfrac{1}{6}\right) = -\tfrac{35}{36}, \\ x_3^{(3)} &= \tfrac{1}{3}\left(2 - 2\cdot\tfrac{3}{4} - 0\cdot\left(-\tfrac{41}{36}\right)\right) = \tfrac{1}{6} \end{aligned}\right\} \quad\Longrightarrow\quad \mathbf{x}^{(3)} = \begin{bmatrix} \frac{143}{144} \\ -\frac{35}{36} \\ \frac{1}{6} \end{bmatrix},$$

$$\left.\begin{aligned} x_1^{(4)} &= \tfrac{1}{2}\left(\tfrac{3}{2} - \tfrac{1}{2}\cdot\left(-\tfrac{35}{36}\right) - \tfrac{1}{2}\cdot\tfrac{1}{6}\right) = \tfrac{137}{144}, \\ x_2^{(4)} &= \tfrac{1}{3}\left(-2 - 1\cdot\tfrac{143}{144} - 1\cdot\tfrac{1}{6}\right) = -\tfrac{455}{432}, \\ x_3^{(4)} &= \tfrac{1}{3}\left(2 - 2\cdot\tfrac{143}{144} - 0\cdot\left(-\tfrac{35}{36}\right)\right) = \tfrac{1}{216} \end{aligned}\right\} \quad\Longrightarrow\quad \mathbf{x}^{(4)} = \begin{bmatrix} \frac{137}{144} \\ -\frac{455}{432} \\ \frac{1}{216} \end{bmatrix}.$$

Bereits nach vier Iterationsschritten, kann man mit $\mathbf{x}^{(4)} \doteq \begin{bmatrix} 0{,}95139 \\ -1{,}0532 \\ 0{,}0046296 \end{bmatrix}$ eine deutliche Annäherung an den Lösungsvektor $\widehat{\mathbf{x}} = \begin{bmatrix} 1 \\ -1 \\ 0 \end{bmatrix}$ erkennen. ♠

Inspiziert man die Formel (3.17) des Jacobi-Verfahrens, so sieht man, dass man das Verfahren **parallelisieren** kann, denn die Komponenten von $\mathbf{x}^{(j+1)}$ lassen sich alle unabhängig von einander berechnen und können somit parallel unabhängig voneinander auf verschiedenen untereinander vernetzten Computern berechnet werden.

Man beobachtet weiter, dass man das Jacobi-Verfahren **vermutlich verbessern kann**, indem man bei der Berechnung von $x_i^{(j+1)}$ die $x_k^{(j)}$ mit $k = 1, 2, \ldots, i-1$ durch die bereits berechneten Komponenten $x_k^{(j+1)}$ mit $k = 1, 2, \ldots, i-1$ von $\mathbf{x}^{(j+1)}$ ersetzt. (Dann ist eine Parallelisierung natürlich nicht mehr möglich.) das so erhaltene Verfahren nennt man das **Gauß-Seidel-Verfahren**.

Verfahren 3.11. (Gauß-Seidel-Verfahren)
Sei $\mathbf{A} = [a_{i,j}] \in \mathbb{R}^{n\times n}$ *eine* ***reguläre*** *(d.h. invertierbare) Matrix, deren* ***Diagonalelemente alle von null verschieden*** *sind (d.h.* $a_{j,j} \neq 0$ *für* $j = 1, 2, \ldots, n$*). Sei* $\mathbf{b} \in \mathbb{R}^n$*, und sei* $\mathbf{x}^{(0)} \in \mathbb{R}^n$ *ein Startvektor. Die Fix-*

punktiteration $\left(\mathbf{x}^{(j)}\right)_{j\in\mathbb{N}_0}$ *mit* $\mathbf{x}^{(j+1)} = \mathbf{D}^{-1}\left(\mathbf{b} - \mathbf{L}\,\mathbf{x}^{(j+1)} - \mathbf{R}\,\mathbf{x}^{(j)}\right)$, *also*

$$x_i^{(j+1)} = \frac{1}{a_{i,i}}\left(b_i - \sum_{k=1}^{i-1} a_{i,k}\,x_k^{(j+1)} - \sum_{k=i+1}^{n} a_{i,k}\,x_k^{(j)}\right), \qquad i = 1, 2, \ldots, n, , \tag{3.21}$$

heißt das ***Gauß-Seidel-Verfahren*** *oder* ***Einzelschrittverfahren****.*

Wie sieht es mit der **Konvergenz des Gauß-Seidel-Verfahrens** aus? Der nächste Satz liefert eine hinreichende Bedingung.

Satz 3.12. (Konvergenz des Gauß-Seidel-Verfahrens)
Seien die Notation und die Voraussetzungen wir in Verfahren 3.11. Wenn die reguläre Matrix $\mathbf{A} \in \mathbb{R}^{n\times n}$ ***streng diagonal dominant*** *ist,* ***konvergiert das Gauß-Seidel-Verfahren*** (3.21) ***für jeden Startvektor*** $\mathbf{x}^{(0)} \in \mathbb{R}^n$ *gegen die eindeutige Lösung* $\widehat{\mathbf{x}}$ *von* $\mathbf{A}\,\mathbf{x} = \mathbf{b}$.

Betrachten wir ein Beispiel.

Beispiel 3.13. (Gauß-Seidel-Verfahren)
Das lineare Gleichungssystem

$$\mathbf{A}\,\mathbf{x} = \mathbf{b} \qquad \text{mit} \qquad \mathbf{A} = \begin{bmatrix} 2 & \frac{1}{2} & \frac{1}{2} \\ 1 & 3 & 1 \\ 2 & 0 & 3 \end{bmatrix}, \qquad \mathbf{b} = \begin{bmatrix} \frac{3}{2} \\ -2 \\ 2 \end{bmatrix},$$

hat nach Beispiel 3.10 eine reguläre Matrix und die eindeutige Lösung $\widehat{\mathbf{x}} = \begin{bmatrix} 1 \\ -1 \\ 0 \end{bmatrix}$.

In Beispiel 3.10 hatten wir bereits geprüft, dass $\mathbf{A}$ streng diagonal dominant ist, so dass das Gauß-Seidel-Verfahren nach Satz 3.12 für jeden Startvektor $\mathbf{x}^{(0)}$ gegen $\widehat{\mathbf{x}}$ konvergieren wird. Wir berechnen nun die ersten drei Iterierten des Gauß-Seidel-Verfahrens mit dem Startvektor $\mathbf{x}^{(0)} = \mathbf{0}$:

$$\left.\begin{aligned} x_1^{(1)} &= \tfrac{1}{2}\left(\tfrac{3}{2} - \tfrac{1}{2}\cdot 0 - \tfrac{1}{2}\cdot 0\right) = \tfrac{3}{4}, \\ x_2^{(1)} &= \tfrac{1}{3}\left(-2 - 1\cdot\tfrac{3}{4} - 1\cdot 0\right) = -\tfrac{11}{12}, \\ x_3^{(1)} &= \tfrac{1}{3}\left(2 - 2\cdot\tfrac{3}{4} - 0\cdot\left(-\tfrac{11}{12}\right)\right) = \tfrac{1}{6} \end{aligned}\right\} \quad\Longrightarrow\quad \mathbf{x}^{(1)} = \begin{bmatrix} \frac{3}{4} \\ -\frac{11}{12} \\ \frac{1}{6} \end{bmatrix},$$

$$\left.\begin{aligned} x_1^{(2)} &= \tfrac{1}{2}\left(\tfrac{3}{2} - \tfrac{1}{2}\cdot\left(-\tfrac{11}{12}\right) - \tfrac{1}{2}\cdot\tfrac{1}{6}\right) = \tfrac{45}{48} = \tfrac{15}{16}, \\ x_2^{(2)} &= \tfrac{1}{3}\left(-2 - 1\cdot\tfrac{15}{16} - 1\cdot\tfrac{1}{6}\right) = -\tfrac{149}{144}, \\ x_3^{(2)} &= \tfrac{1}{3}\left(2 - 2\cdot\tfrac{15}{16} - 0\cdot\left(-\tfrac{149}{144}\right)\right) = \tfrac{1}{24} \end{aligned}\right\} \quad\Longrightarrow\quad \mathbf{x}^{(2)} = \begin{bmatrix} \frac{15}{16} \\ -\frac{149}{144} \\ \frac{1}{24} \end{bmatrix},$$

$$\left.\begin{aligned} x_1^{(3)} &= \tfrac{1}{2}\left(\tfrac{3}{2} - \tfrac{1}{2}\cdot\left(-\tfrac{149}{144}\right) - \tfrac{1}{2}\cdot\tfrac{1}{24}\right) = \tfrac{575}{576}, \\ x_2^{(3)} &= \tfrac{1}{3}\left(-2 - 1\cdot\tfrac{575}{576} - 1\cdot\tfrac{1}{24}\right) = -\tfrac{1751}{1728}, \\ x_3^{(3)} &= \tfrac{1}{3}\left(2 - 2\cdot\tfrac{575}{576} - 0\cdot\left(-\tfrac{1751}{1728}\right)\right) = \tfrac{1}{864} \end{aligned}\right\} \quad\Longrightarrow\quad \mathbf{x}^{(3)} = \begin{bmatrix} \frac{575}{576} \\ -\frac{1751}{1728} \\ \frac{1}{864} \end{bmatrix}.$$

Drei Iterationsschritte liefern die Näherung $\mathbf{x}^{(3)} \doteq \begin{bmatrix} 0{,}99826 \\ -1{,}0133 \\ 0{,}0011574 \end{bmatrix}$ von $\widehat{\mathbf{x}} = \begin{bmatrix} 1 \\ -1 \\ 0 \end{bmatrix}$.

Der Vergleich mit Beispiel 3.10 zeigt, dass wir nach drei Iterationsschritten eine bessere Näherung von $\widehat{\mathbf{x}}$ erhalten als mit dem Jacobi-Verfahren nach vier Iterationsschritten. ♠

Das Gauß-Seidel-Verfahren (vgl. Verfahren 3.11) zur iterativen Berechnung einer (Näherungs-)Lösung des linearen Gleichungssystems $\mathbf{A}\,\mathbf{x} = \mathbf{b}$ (mit einer regulären Matrix $\mathbf{A} \in \mathbb{R}^{n\times n}$) hat die Iterationsvorschrift

$$\mathbf{x}^{(j+1)} = \mathbf{D}^{-1}\left(\mathbf{b} - \mathbf{L}\,\mathbf{x}^{(j+1)} - \mathbf{R}\,\mathbf{x}^{(j)}\right), \qquad j = 0, 1, 2, \ldots, \tag{3.22}$$

wobei die reguläre Matrix $\mathbf{A}$ wie folgt eindeutig zerlegt wurde: $\mathbf{A} = \mathbf{L} + \mathbf{D} + \mathbf{R}$ mit der Diagonalmatrix $\mathbf{D}$, der streng unteren Linksdreiecksmatrix $\mathbf{L}$ und der streng oberen Rechtsdreiecksmatrix $\mathbf{R}$. Per Annahme hat $\mathbf{A}$ dabei nur Einträge ungleich Null auf der Diagonalen. – Die Darstellung (3.22) der Iterierten des Gauß-Seidel-Verfahrens ist nicht in der „üblichen“ Form einer Fixpunktiteration $\mathbf{x}^{(j+1)} = \mathbf{C}\,\mathbf{x}^{(j)} + \mathbf{c}$, $j \in \mathbb{N}_0$, da auf der rechten Seite in (3.22) auch $\mathbf{x}^{(j+1)}$ auftaucht. Wir wollen (3.22) nun mit Umformungen in die „übliche“ Form einer Fixpunktiteration $\mathbf{x}^{(j+1)} = \mathbf{C}\,\mathbf{x}^{(j)} + \mathbf{c}$ bringen.

Wir lösen (3.22) nach $\mathbf{x}^{(j+1)}$ auf:

$$\begin{aligned} & \mathbf{x}^{(j+1)} = \mathbf{D}^{-1}\left(\mathbf{b} - \mathbf{L}\,\mathbf{x}^{(j+1)} - \mathbf{R}\,\mathbf{x}^{(j)}\right) \\ \Longleftrightarrow\quad & \mathbf{D}\,\mathbf{x}^{(j+1)} = \mathbf{b} - \mathbf{L}\,\mathbf{x}^{(j+1)} - \mathbf{R}\,\mathbf{x}^{(j)} \\ \Longleftrightarrow\quad & \mathbf{D}\,\mathbf{x}^{(j+1)} + \mathbf{L}\,\mathbf{x}^{(j+1)} = \mathbf{b} - \mathbf{R}\,\mathbf{x}^{(j)} \\ \Longleftrightarrow\quad & (\mathbf{L} + \mathbf{D})\,\mathbf{x}^{(j+1)} = \mathbf{b} - \mathbf{R}\,\mathbf{x}^{(j)} \\ \Longleftrightarrow\quad & \mathbf{x}^{(j+1)} = (\mathbf{L} + \mathbf{D})^{-1}\,\mathbf{b} - (\mathbf{L} + \mathbf{D})^{-1}\,\mathbf{R}\,\mathbf{x}^{(j)}, \end{aligned}$$

wobei wir im letzten Schritt genutzt haben, dass $\mathbf{L}+\mathbf{D}$ invertierbar ist, da diese untere Dreiecksmatrix auf der Diagonale nur Einträge ungleich null hat. Es gilt also

$$\begin{aligned}\mathbf{x}^{(j+1)} &= \underbrace{-(\mathbf{L}+\mathbf{D})^{-1}\,\mathbf{R}}_{=\mathbf{C}}\,\mathbf{x}^{(j)} + \underbrace{(\mathbf{L}+\mathbf{D})^{-1}\,\mathbf{b}}_{=\mathbf{c}} \\ &= (\mathbf{L}+\mathbf{D})^{-1}\left(\mathbf{b}-\mathbf{R}\,\mathbf{x}^{(j)}\right), \qquad j\in\mathbb{N}_0. \end{aligned} \tag{3.23}$$

Wir lesen ab, dass die Matrix der Fixpunktiteration des Gauß-Seidel-Verfahrens

$$\mathbf{C} = -(\mathbf{L}+\mathbf{D})^{-1}\,\mathbf{R} \tag{3.24}$$

ist. Weiteres Umformen von (3.24) mit $\mathbf{A}=\mathbf{L}+\mathbf{D}+\mathbf{R}$ liefert

$$\mathbf{C} = -(\mathbf{L}+\mathbf{D})^{-1}\,\mathbf{R} = -(\mathbf{L}+\mathbf{D})^{-1}\left(\mathbf{A}-(\mathbf{L}+\mathbf{D})\right) = -(\mathbf{L}+\mathbf{D})^{-1}\mathbf{A}+\mathbf{E}_n.$$

Daraus folgt in (3.23)

$$\begin{aligned}\mathbf{x}^{(j+1)} &= -(\mathbf{L}+\mathbf{D})^{-1}\left(\mathbf{A}-(\mathbf{L}+\mathbf{D})\right)\mathbf{x}^{(j)} + (\mathbf{L}+\mathbf{D})^{-1}\,\mathbf{b} && (3.25)\\ &= \left(-(\mathbf{L}+\mathbf{D})^{-1}\mathbf{A}+\mathbf{E}_n\right)\mathbf{x}^{(j)} + (\mathbf{L}+\mathbf{D})^{-1}\,\mathbf{b} \\ &= \mathbf{x}^{(j)} + (\mathbf{L}+\mathbf{D})^{-1}\,\underbrace{\left(\mathbf{b}-\mathbf{A}\,\mathbf{x}^{(j)}\right)}_{=\mathbf{r}^{(j)}}, \qquad j\in\mathbb{N}_0. && (3.26)\end{aligned}$$

An (3.25) sehen wir, dass hier $\mathbf{B}=\mathbf{L}+\mathbf{D}$ als Näherung für $\mathbf{A}$, und damit $\mathbf{B}^{-1}=(\mathbf{L}+\mathbf{D})^{-1}$ als Näherung von $\mathbf{A}^{-1}$, verwendet wurde.

An (3.26) sehen wir, dass $\mathbf{x}^{(j)}$ durch $(\mathbf{L}+\mathbf{D})^{-1}\left(\mathbf{b}-\mathbf{A}\,\mathbf{x}^{(j)}\right) = (\mathbf{L}+\mathbf{D})^{-1}\,\mathbf{r}^{(j)}$ (eine Näherung für die Lösung von $\mathbf{A}\,\mathbf{y}=\mathbf{r}^{(j)}$) „verbessert“ wird.

3.3 Methode der konjugierten Gradienten (CG-Verfahren)

Bei der **Methode der konjugierten Gradienten** oder dem **CG-Verfahren** (CG steht für „conjugate gradient“) zur Lösung eines linearen Gleichungssystems $\mathbf{A}\,\mathbf{x}=\mathbf{b}$ mit einer **positiv definiten Matrix** $\mathbf{A}\in\mathbb{R}^{n\times n}$ handelt es sich um ein Iterationsverfahren, das in höchstens n Schritten konvergiert. Da das **Ergebnis nach maximal** n **Schritten erreicht** wird, könnte man dieses auch als ein direktes Verfahren auffassen. In der Praxis ist es aber so, dass n normalerweise sehr groß ist und man das CG-Verfahren daher nach deutlich weniger als n Schritten stoppen wird. Damit handelt es sich in der praktischen Anwendung um ein **Iterationsverfahren**.

Als Vorbereitung für das CG-Verfahren benötigen wir einige neue Begriffe.

Definition 3.14. (positiv definite Matrix)
Eine Matrix $\mathbf{A} = [a_{i,j}] \in \mathbb{R}^{n \times n}$ *heißt* ***positiv definit****, wenn gilt:*

(1) $\mathbf{A}$ *ist symmetrisch, d.h.* $\mathbf{A}^T = \mathbf{A}$*, und*

(2) $\mathbf{x}^T \mathbf{A}\, \mathbf{x} = \sum_{i=1}^{n} \sum_{j=1}^{n} a_{i,j}\, x_i\, x_j > 0$ *für alle* $\mathbf{x} = \begin{bmatrix} x_1 \\ \vdots \\ x_n \end{bmatrix} \in \mathbb{R}^n \setminus \{\mathbf{0}\}$.

Der nächste Hilfssatz liefert uns weitere nützliche Informationen über symmetrische und positiv definite Matrizen.

Hilfssatz 3.15. (Informationen über symmetrische Matrizen)

(1) Eine ***symmetrische*** *Matrix* $\mathbf{A} \in \mathbb{R}^{n \times n}$ *hat* ***nur reelle Eigenwerte****, und es gibt eine Orthonormalbasis für* $\mathbb{R}^n$ *aus Eigenvektoren von* $\mathbf{A}$*.*

(2) Eine symmetrische Matrix $\mathbf{A} \in \mathbb{R}^{n \times n}$ *ist genau dann* ***positiv definit****, wenn alle ihre Eigenwerte* ***positiv*** *(also* > 0*) sind.*

Betrachten wir einige Beispiele für positiv definite Matrizen.

Beispiel 3.16. (positiv definite Matrizen)
Betrachten wir die beiden Matrizen

$$\mathbf{A} = \begin{bmatrix} \frac{3}{2} & 0 & \frac{1}{2} \\ 0 & 3 & 0 \\ \frac{1}{2} & 0 & \frac{3}{2} \end{bmatrix} \qquad \text{und} \qquad \mathbf{B} = \begin{bmatrix} \frac{3}{2} & 0 & \frac{1}{2} \\ 0 & -2 & 0 \\ \frac{1}{2} & 0 & \frac{3}{2} \end{bmatrix}.$$

Es gelten offenbar $\mathbf{A}^T = \mathbf{A}$ und $\mathbf{B}^T = \mathbf{B}$. Also sind $\mathbf{A}$ und $\mathbf{B}$ jeweils symmetrisch.

Wir zeigen nun, dass $\mathbf{A}$ positiv definit ist, indem wir $\mathbf{x}^T \mathbf{A}\, \mathbf{x}$ explizit berechnen und geeignet umformen:

$$\mathbf{x}^T \mathbf{A}\, \mathbf{x} = \begin{bmatrix} x_1; x_2; x_3 \end{bmatrix} \begin{bmatrix} \frac{3}{2} & 0 & \frac{1}{2} \\ 0 & 3 & 0 \\ \frac{1}{2} & 0 & \frac{3}{2} \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} x_1; x_2; x_3 \end{bmatrix} \begin{bmatrix} \frac{3}{2}\, x_1 + \frac{1}{2}\, x_3 \\ 3\, x_2 \\ \frac{1}{2}\, x_1 + \frac{3}{2}\, x_3 \end{bmatrix}$$

$$= \frac{3}{2}\, x_1^2 + \frac{1}{2}\, x_1\, x_3 + 3\, x_2^2 + \frac{1}{2}\, x_3\, x_1 + \frac{3}{2}\, x_3^2 = \frac{3}{2}\, x_1^2 + x_1\, x_3 + 3\, x_2^2 + \frac{3}{2}\, x_3^2$$

$$= x_1^2 + 3\,x_2^2 + x_3^2 + \frac{1}{2}\left(x_1^2 + 2\,x_1\,x_3 + x_3^2\right) = x_1^2 + 3\,x_2^2 + x_3^2 + \frac{1}{2}\,(x_1 + x_3)^2.$$

In der letzten Darstellung haben wir nun eine Summe von Quadraten, und somit gilt $\mathbf{x}^T\mathbf{A}\,\mathbf{x} \geq 0$ für alle $\mathbf{x} \in \mathbb{R}^3$. Die Gleichheit $\mathbf{x}^T\mathbf{A}\,\mathbf{x} = 0$ kann nur dann gelten, wenn $x_1 = 0$, $x_2 = 0$ und $x_3 = 0$ sind, also nur für $\mathbf{x} = \mathbf{0}$. Insofern gilt $\mathbf{x}^T\mathbf{A}\,\mathbf{x} > 0$ für alle $\mathbf{x} \in \mathbb{R}^3 \setminus \{\mathbf{0}\}$. Also ist $\mathbf{A}$ positiv definit.

Wir hätten genauso gut Hilfssatz 3.15 (2) nutzen können, um zu zeigen, dass $\mathbf{A}$ positiv definit ist. Wir berechnen also die Eigenwerte von $\mathbf{A}$:

$$\begin{aligned}
p_{\mathbf{A}}(\lambda) = \det(\mathbf{A} - \lambda\,\mathbf{E}_3) &= \det\left(\begin{bmatrix} \frac{3}{2} - \lambda & 0 & \frac{1}{2} \\ 0 & 3 - \lambda & 0 \\ \frac{1}{2} & 0 & \frac{3}{2} - \lambda \end{bmatrix}\right) \\
&= (3 - \lambda)\left(\frac{3}{2} - \lambda\right)^2 + 0 + 0 - \frac{1}{4}\,(3 - \lambda) - 0 - 0 \\
&= (3 - \lambda)\left[\left(\frac{3}{2} - \lambda\right)^2 - \frac{1}{4}\right] \\
&= (3 - \lambda)\left(\frac{3}{2} - \lambda - \frac{1}{2}\right)\left(\frac{3}{2} - \lambda + \frac{1}{2}\right) \\
&= (3 - \lambda)\,(1 - \lambda)\,(2 - \lambda),
\end{aligned}$$

wobei im vorletzten Schritt die dritte binomische Formel verwendet wurde. Also sind $\lambda_1 = 1$, $\lambda_2 = 2$ und $\lambda_3 = 3$ die Eigenwerte von $\mathbf{A}$. Da alle Eigenwerte von $\mathbf{A}$ positiv sind, folgt, dass die symmetrische Matrix $\mathbf{A}$ positiv definit ist.

Da es sich bei Hilfssatz 3.15 (1) um eine Äquivalenzaussage („genau dann, wenn“-Aussage) handelt, können wir diesen auch nutzen, um zu zeigen, dass die symmetrische Matrix $\mathbf{B}$ nicht positiv definit ist:

$$\begin{aligned}
p_{\mathbf{A}}(\lambda) = \det(\mathbf{A} - \lambda\,\mathbf{E}_3) &= \det\left(\begin{bmatrix} \frac{3}{2} - \lambda & 0 & \frac{1}{2} \\ 0 & -2 - \lambda & 0 \\ \frac{1}{2} & 0 & \frac{3}{2} - \lambda \end{bmatrix}\right) \\
&= (-2 - \lambda)\left(\frac{3}{2} - \lambda\right)^2 + 0 + 0 - \frac{1}{4}\,(-2 - \lambda) - 0 - 0 \\
&= (-2 - \lambda)\left[\left(\frac{3}{2} - \lambda\right)^2 - \frac{1}{4}\right] \\
&= (-2 - \lambda)\left(\frac{3}{2} - \lambda - \frac{1}{2}\right)\left(\frac{3}{2} - \lambda + \frac{1}{2}\right)
\end{aligned}$$

$$= (-2-\lambda)\,(1-\lambda)\,(2-\lambda),$$

wobei im vorletzten Schritt die dritte binomische Formel verwendet wurde. Also sind $\lambda_1 = 1$, $\lambda_2 = 2$ und $\lambda_3 = -2$ die Eigenwerte von $\mathbf{B}$. Da nicht alle Eigenwerte von $\mathbf{B}$ positiv sind, ist die symmetrische Matrix $\mathbf{B}$ nicht positiv definit. ♠

Hilfssatz 3.17. (positiv definit $\Longrightarrow$ invertierbar)
*Sei $\mathbf{A} \in \mathbb{R}^{n\times n}$ eine **positiv definite**, symmetrische Matrix. Dann ist $\mathbf{A}$ **regulär** (also **invertierbar**).*

Beweis von Hilfssatz 3.17: Eine Matrix $\mathbf{A} \in \mathbb{R}^{n\times n}$ ist genau dann invertierbar, also regulär, wenn das lineare Gleichungssystem $\mathbf{A}\,\mathbf{x} = \mathbf{0}$ nur die einzige Lösung $\mathbf{x} = \mathbf{0}$ hat. Wir betrachten daher das lineare Gleichungssystem $\mathbf{A}\,\mathbf{x} = \mathbf{0}$:

$$\mathbf{A}\,\mathbf{x} = \mathbf{0} \quad \Longrightarrow \quad \mathbf{x}^T\mathbf{A}\,\mathbf{x} = \underbrace{\mathbf{x}^T\,\mathbf{0}}_{=0} \quad \Longrightarrow \quad \mathbf{x}^T\mathbf{A}\,\mathbf{x} = 0 \tag{3.27}$$

Da $\mathbf{A}$ positiv definit ist, wissen wir aber, dass für alle $\mathbf{x} \neq \mathbf{0}$ gilt $\mathbf{x}^T\mathbf{A}\,\mathbf{x} > 0$. Weiter gilt offensichtlich $\mathbf{0}^T\mathbf{A}\,\mathbf{0} = 0$. Also folgt aus (3.27), dass $\mathbf{x} = \mathbf{0}$ ist.

Wir haben also gezeigt, dass aus $\mathbf{A}\,\mathbf{x} = \mathbf{0}$ folgt, dass $\mathbf{x} = \mathbf{0}$ ist. Folglich ist die Matrix $\mathbf{A}$ invertierbar, also regulär. □

Ab jetzt sei $\mathbf{A} = [a_{i,j}] \in \mathbb{R}^{n\times n}$ immer eine **positiv definite, symmetrische Matrix** (und somit regulär), und sei $\mathbf{b} \in \mathbb{R}^n$ die rechte Seite des zu lösenden linearen Gleichungssystems $\mathbf{A}\,\mathbf{x} = \mathbf{b}$. Die eindeutig bestimmte Lösung von $\mathbf{A}\,\mathbf{x} = \mathbf{b}$ sei mit $\widehat{\mathbf{x}}$ bezeichnet. Wir definieren nun das sogenannte **konjugierte Gradienten-Funktional** (oder **CG-Funktional**)

$$f : \mathbb{R}^n \to \mathbb{R}, \quad f(\mathbf{x}) := \frac{1}{2}\,\mathbf{x}^T\mathbf{A}\,\mathbf{x} - \mathbf{x}^T\,\mathbf{b}. \tag{3.28}$$

Wir zeigen nun, dass die eindeutig bestimmte Lösung $\widehat{\mathbf{x}}$ des linearen Gleichungssystems $\mathbf{A}\,\mathbf{x} = \mathbf{b}$ die **eindeutig bestimmte globale Minimalstelle des konjugierten Gradientenfunktionals** (3.28) ist.

Ausführlicher geschrieben erhalten wir für f aus (3.28)

$$f(\mathbf{x}) = f(x_1, x_2, \ldots, x_n) = \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n} a_{i,j}\,x_i\,x_j - \sum_{j=1}^{n} x_j\,b_j, \tag{3.29}$$

und wir sehen, dass das konjugierte Gradientenfunktional f ein (multivariates) Polynom vom Grad 2 auf $\mathbb{R}^n$, also in den Komponenten $x_1, x_2, \ldots, x_n$ des Vektors $\mathbf{x}$, ist. Um die Extremalstellen von f zu finden berechnen wir den Gradienten und die Hesse-Matrix von f (Details der Berechnung siehe Übungsblatt):

$$(\nabla f)(\mathbf{x}) = \mathbf{A}\,\mathbf{x} - \mathbf{b}, \tag{3.30}$$

$$(\mathbf{H}f)(\mathbf{x}) = \mathbf{A}. \tag{3.31}$$

Nullsetzen von ∇f liefert

$$\mathbf{0} = (\nabla f)(\mathbf{x}) = \mathbf{A}\,\mathbf{x} - \mathbf{b} \qquad \Longleftrightarrow \qquad \mathbf{A}\,\mathbf{x} = \mathbf{b}.$$

Da die Hesse-Matrix $(\mathbf{H}f)(\mathbf{x}) = \mathbf{A}$ positiv definit ist, folgt, dass f in der eindeutigen Lösung $\widehat{\mathbf{x}}$ von $\mathbf{A}\,\mathbf{x} = \mathbf{b}$ ein lokales Minimum hat. Es ist auch die einzige lokale Extremalstelle, da $\widehat{\mathbf{x}}$ der einzige kritische Punkt von f ist.

Wir zeigen nun, dass $\widehat{\mathbf{x}}$ sogar die eindeutig bestimmte globale Minimalstelle von f ist: Da f ein Polynom vom Grad 2 in $x_1, x_2, \ldots, x_n$ ist, lässt sich dieses exakt durch sein Taylor-Polynom vom Grad 2 mit dem Entwicklungspunkt $\widehat{\mathbf{x}}$ darstellen: Mit (3.30) und (3.31) folgt also

$$\begin{aligned} f(\mathbf{x}) &= \underbrace{f(\widehat{\mathbf{x}}) + (\mathbf{x}-\widehat{\mathbf{x}})^T (\nabla f)(\widehat{\mathbf{x}}) + \frac{1}{2}\,(\mathbf{x}-\widehat{\mathbf{x}})^T (\mathbf{H}f)(\widehat{\mathbf{x}})\,(\mathbf{x}-\widehat{\mathbf{x}})}_{\text{Taylorpolynom von } f \text{ vom Grad 2 mit Entwicklungspunkt } \widehat{\mathbf{x}}} \\ &= f(\widehat{\mathbf{x}}) + (\mathbf{x}-\widehat{\mathbf{x}})^T \big(\underbrace{\mathbf{A}\,\widehat{\mathbf{x}} - \mathbf{b}}_{=\mathbf{0}} \big) + \frac{1}{2}\,(\mathbf{x}-\widehat{\mathbf{x}})^T \mathbf{A}\,(\mathbf{x}-\widehat{\mathbf{x}}) \\ &= f(\widehat{\mathbf{x}}) + \frac{1}{2}\,(\mathbf{x}-\widehat{\mathbf{x}})^T \mathbf{A}\,(\mathbf{x}-\widehat{\mathbf{x}}). \end{aligned} \tag{3.32}$$

Da $\mathbf{A}$ positiv definit ist, gilt

$$(\mathbf{x}-\widehat{\mathbf{x}})^T \mathbf{A}\,(\mathbf{x}-\widehat{\mathbf{x}}) > 0 \qquad \text{für alle } \mathbf{x} \in \mathbb{R}^n \text{ mit } \big(\mathbf{x}-\widehat{\mathbf{x}} \neq \mathbf{0} \iff \mathbf{x} \neq \widehat{\mathbf{x}}\big),$$

und wir sehen an (3.32), dass gilt

$$f(\mathbf{x}) > f(\widehat{\mathbf{x}}) \qquad \text{für alle } \mathbf{x} \in \mathbb{R}^n \setminus \{\widehat{\mathbf{x}}\},$$

d.h. $\widehat{\mathbf{x}}$ ist die **eindeutig bestimmte globale Minimalstelle von** f.

Wir halten diese Erkenntnisse als Satz fest.

Satz 3.18. (globale Minimalstelle des CG-Funktionals)
Seien $\mathbf{A} \in \mathbb{R}^{n\times n}$ *eine* ***positiv definite, symmetrische Matrix*** *und* $\mathbf{b} \in \mathbb{R}^n$. *Die eindeutig bestimmte Lösung* $\widehat{\mathbf{x}}$ *des linearen Gleichungssystems* $\mathbf{A}\,\mathbf{x} = \mathbf{b}$ *ist die* ***eindeutig bestimmte globale Minimalstelle*** *des* ***konjugierten Gradienten Funktionals*** *(oder* ***CG-Funktionals****)*

$$f : \mathbb{R}^n \to \mathbb{R}, \quad f(\mathbf{x}) := \frac{1}{2}\,\mathbf{x}^T\mathbf{A}\,\mathbf{x} - \mathbf{x}^T\,\mathbf{b}. \tag{3.33}$$

Wir wollen nun versuchen, die Minimalstelle $\widehat{\mathbf{x}}$ des CG-Funktionals (3.33) mit einem einfachen **iterativen Prozess in maximal** n **Schritten** zu finden: (Im Folgenden schreiben wir der Einfachheit halber immer $\mathbf{x}_j, \mathbf{p}_j, \ldots$ statt $\mathbf{x}^{(j)}, \mathbf{p}^{(j)}, \ldots$, da wir in diesem Teilkapitel keine Komponenten der Vektoren benötigen und daher keine Verwirrung entstehen kann.) Wir wollen eine Folge von $(\mathbf{x}_j)_{j=1,\ldots,m}$ berechnen, die nach spätestens $m \le n$ Schritten $\mathbf{x}_m = \widehat{\mathbf{x}}$ erfüllt. Dabei wird $\mathbf{x}_{j+1}$ als Verbesserung von $\mathbf{x}_j$ mittels

$$\boxed{\mathbf{x}_{j+1} = \mathbf{x}_j + \alpha_j\,\mathbf{p}_j.}$$

berechnet, wobei der **Richtungsvektor** $\mathbf{p}_j \in \mathbb{R}^n \setminus \{\mathbf{0}\}$ und die **Schrittweite** $\alpha_j \in \mathbb{R}$ passend gewählt werden müssen. Für den Moment nehmen wir an, dass der Richtungsvektor $\mathbf{p}_j$ bereits gegeben ist, und machen uns nur um die Wahl der Schrittweite Gedanken. Es soll natürlich gelten

$$f(\mathbf{x}_{j+1}) = f\big(\mathbf{x}_j + \alpha_j\,\mathbf{p}_j\big) \le f(\mathbf{x}_j),$$

und wir werden α_j so wählen, dass die linke Seite minimal wird. Wir minimieren also f entlang der Geraden $\mathbf{x}_{j+1} = \mathbf{x}_j + \alpha\,\mathbf{p}_j$, $\alpha \in \mathbb{R}$. Sei also $\phi : \mathbb{R} \to \mathbb{R}$,

$$\phi(\alpha) := f\big(\mathbf{x}_j + \alpha\,\mathbf{p}_j\big) = \frac{1}{2}\,\big(\mathbf{x}_j + \alpha\,\mathbf{p}_j\big)^T\,\mathbf{A}\,\big(\mathbf{x}_j + \alpha\,\mathbf{p}_j\big) - \big(\mathbf{x}_j + \alpha\,\mathbf{p}_j\big)^T\,\mathbf{b}.$$

Die notwendige Bedingung $\phi'(\alpha) = 0$ liefert mit Hilfe der Kettenregel und (3.30)

$$\begin{aligned} 0 = \phi'(\alpha) &= \mathbf{p}_j^T\,(\nabla f)\big(\mathbf{x}_j + \alpha\,\mathbf{p}_j\big) = \mathbf{p}_j^T\,\big[\mathbf{A}\,\big(\mathbf{x}_j + \alpha\,\mathbf{p}_j\,\big) - \mathbf{b}\big] \\ &= \mathbf{p}_j^T\mathbf{A}\,\mathbf{x}_j + \alpha\,\mathbf{p}_j^T\mathbf{A}\,\mathbf{p}_j - \mathbf{p}_j^T\,\mathbf{b} = \mathbf{p}_j^T\,\big(\mathbf{A}\,\mathbf{x}_j - \mathbf{b}\big) + \alpha\,\mathbf{p}_j^T\mathbf{A}\,\mathbf{p}_j. \end{aligned} \tag{3.34}$$

Da $\mathbf{A}$ positiv definit ist, gilt für die zweite Ableitung von ϕ

$$\phi''(\alpha) = \mathbf{p}_j^T\mathbf{A}\,\mathbf{p}_j > 0.$$

Somit liegt bei jedem $\alpha \in \mathbb{R}$ mit $\phi'(\alpha) = 0$ ein lokales Minimum vor. Auflösen von (3.34) nach α ergibt für die Schrittlänge $\alpha_j = \alpha$ die Formel

$$0 = \mathbf{p}_j^T \left(\mathbf{A}\,\mathbf{x}_j - \mathbf{b}\right) + \alpha_j\, \mathbf{p}_j^T \mathbf{A}\,\mathbf{p}_j \quad \Longleftrightarrow \quad \boxed{\alpha_j = \frac{\mathbf{p}_j^T \left(\mathbf{b} - \mathbf{A}\,\mathbf{x}_j\right)}{\mathbf{p}_j^T \mathbf{A}\,\mathbf{p}_j} = \frac{\mathbf{p}_j^T\, \mathbf{r}_j}{\mathbf{p}_j^T \mathbf{A}\,\mathbf{p}_j}},$$

wobei wir das **Residuum im j-ten Schritt** durch

$$\boxed{\mathbf{r}_j := \mathbf{b} - \mathbf{A}\,\mathbf{x}_j}$$

bezeichnet haben.

Falls die Richtungsvektoren $\mathbf{p}_j$, $j = 1, 2, \ldots$, vorgegeben sind bzw. in jedem Iterationsschritt geeignet berechnet werden, erhalten wir also den folgenden generischen Minimierungsalgorithmus

Verfahren 3.19. (generischer Minimierungsalgorithmus)
Gegeben seien eine ***positive definite, symmetrische Matrix*** $\mathbf{A} \in \mathbb{R}^{n\times n}$ *und eine rechte Seite* $\mathbf{b} \in \mathbb{R}^n$. *Weiter seien ein Startvektor* $\mathbf{x}_1 \in \mathbb{R}^n$ *und ein erster Richtungsvektor* $\mathbf{p}_1 \in \mathbb{R}^n$ *gegeben.*

Initialisierung: Setze $\mathbf{r}_1 = \mathbf{b} - \mathbf{A}\,\mathbf{x}_1$ *(Residuum des Startvektors* $\mathbf{x}_1$*)*

Für $j = 1, 2, \ldots$ *führe nun jeweils die folgenden Schritte durch:*

(1) Berechne die Schrittweite $\alpha_j = \dfrac{\mathbf{p}_j^T\, \mathbf{r}_j}{\mathbf{p}_j^T \mathbf{A}\,\mathbf{p}_j}$.

(2) Berechne die verbesserte Näherung $\mathbf{x}_{j+1} = \mathbf{x}_j + \alpha_j\, \mathbf{p}_j$.

(3) Berechne das Residuum $\mathbf{r}_{j+1} = \mathbf{b} - \mathbf{A}\,\mathbf{x}_{j+1}$.

(4) Wähle einen neuen Richtungsvektor $\mathbf{p}_{j+1}$.

Wir beobachten noch, dass wir α_j so gewählt haben, dass der **Richtungsvektor $\mathbf{p}_j$ und das neue Residuum $\mathbf{r}_{j+1}$ zueinander orthogonal sind**, denn

$$\begin{aligned}
\mathbf{p}_j^T\, \mathbf{r}_{j+1} &= \mathbf{p}_j^T \left(\mathbf{b} - \mathbf{A}\,\mathbf{x}_{j+1}\right) = \mathbf{p}_j^T \left(\mathbf{b} - \mathbf{A}\,(\underbrace{\mathbf{x}_j + \alpha_j\, \mathbf{p}_j}_{=\mathbf{x}_{j+1}})\right) \\
&= \mathbf{p}_j^T \left(\mathbf{b} - \mathbf{A}\,\mathbf{x}_j - \alpha_j\, \mathbf{A}\,\mathbf{p}_j\right) = \mathbf{p}_j^T\, (\underbrace{\mathbf{b} - \mathbf{A}\,\mathbf{x}_j}_{=\mathbf{r}_j}) - \alpha_j\, \mathbf{p}_j^T \mathbf{A}\,\mathbf{p}_j \\
&= \mathbf{p}_j^T\, \mathbf{r}_j - \alpha_j\, \mathbf{p}_j^T \mathbf{A}\,\mathbf{p}_j = \mathbf{p}_j^T\, \mathbf{r}_j - \frac{\mathbf{p}_j^T\, \mathbf{r}_j}{\mathbf{p}_j^T \mathbf{A}\,\mathbf{p}_j} \cdot \left(\mathbf{p}_j^T \mathbf{A}\,\mathbf{p}_j\right) = 0, \qquad (3.35)
\end{aligned}$$

wobei wir im letzten Schritt die Formel für α_j eingesetzt haben.

Nun interessieren wir uns dafür, wie man die Richtungsvektoren $\mathbf{p}_j$, $j = 1, 2, \ldots, n$, nacheinander so wählt, dass das Verfahren 3.19 nach maximal n Schritten konvergiert. Es ist naheliegend, dass man $\mathbf{p}_1, \mathbf{p}_2, \ldots, \mathbf{p}_n \in \mathbb{R}^n$ als eine Basis von $\mathbb{R}^n$ wählt, damit gewährleistet ist, dass man die Lösung $\widehat{\mathbf{x}}$ von $\mathbf{A}\,\mathbf{x} = \mathbf{b}$ immer als Linearkombination der Vektoren $\mathbf{p}_1, \mathbf{p}_2, \ldots, \mathbf{p}_n$ darstellen kann.

Die naheliegendste Wahl von $\mathbf{p}_1, \mathbf{p}_2, \ldots, \mathbf{p}_n$ ist (vgl. (3.30))

$$\mathbf{p}_j = -\nabla f(\mathbf{x}_j) = -\big(\mathbf{A}\,\mathbf{x}_j - \mathbf{b}\big) = \mathbf{b} - \mathbf{A}\,\mathbf{x}_j = \mathbf{r}_j,$$

denn dann zeigt $\mathbf{p}_j$ in die **Richtung des steilsten Abstiegs** $-\nabla f(\mathbf{x}_j)$ von f im Punkt $\mathbf{x}_j$. An der obigen Rechnung sehen wir, dass dann $\mathbf{p}_j = \mathbf{r}_j$ gilt, d.h. $\mathbf{p}_j$ ist das Residuum der vorherigen Iterierten $\mathbf{x}_j$. Aus (3.35) folgt dann, das gilt $\mathbf{p}_j^T\,\mathbf{p}_{j+1} = 0$, d.h. zwei aufeinanderfolgende Richtungsvektoren $\mathbf{p}_j$ und $\mathbf{p}_{j+1}$ sind immer orthogonal zueinander. Allerdings erweist sich diese Wahl der $\mathbf{p}_j$ als ungünstig, und der mit dieser Wahl erhaltene Minimierungsalgorithmus, bekannt unter dem Namen „Methode des steilsten Abstiegs“, **konvergiert oft sehr langsam**. Wir werden dieses in einer Übungsaufgabe genauer untersuchen.

Die optimale Wahl der Richtungsvektoren $\mathbf{p}_1, \mathbf{p}_2, \ldots, \mathbf{p}_n \in \mathbb{R}^n$ stellen sogenannte **A-konjugierte Vektoren** dar:

Definition 3.20. (A-konjugierte Vektoren)
*Sei $\mathbf{A} \in \mathbb{R}^{n\times n}$ eine **positiv definite, symmetrische Matrix**. The Vektoren $\mathbf{p}_1, \mathbf{p}_2, \ldots, \mathbf{p}_m \in \mathbb{R}^n \setminus \{\mathbf{0}\}$ heißen **$\mathbf{A}$-konjugiert**, wenn gilt*

$$\mathbf{p}_j^T \mathbf{A}\,\mathbf{p}_k = 0 \qquad \text{für alle } j, k = 1, 2, \ldots, m \text{ with } j \neq k.$$

Was bedeutet es, wenn Vektoren A-konjugiert sind? Der nächste Hilfssatz liefert wichtige Informationen dazu. Den im Hilfssatz vorkommenden Begriff eines Skalarprodukts wiederholen wir auf dem Übungszettel.

Hilfssatz 3.21. (Eigenschaften A-konjugierter Vektoren)
*Sei $\mathbf{A} \in \mathbb{R}^{n\times n}$ eine **positiv definite, symmetrische Matrix**. Die Vektoren $\mathbf{p}_1, \mathbf{p}_2, \ldots, \mathbf{p}_m \in \mathbb{R}^n \setminus \{\mathbf{0}\}$ seien **$\mathbf{A}$-konjugiert**. Dann gelten:*

(1) Da $\mathbf{A}$ positiv definit ist, gilt $\mathbf{p}_j^T \mathbf{A}\,\mathbf{p}_j > 0$ für alle $j = 1, 2, \ldots, m$.

*(2) Die $\mathbf{A}$-konjugierten Vektoren $\mathbf{p}_1, \mathbf{p}_2, \ldots, \mathbf{p}_m$ sind **linear unabhängig**.*

(3) Es kann höchstens n $\mathbf{A}$-konjugierte Vektoren geben, d.h. es gilt $m \leq n$.

(4) Wir können ***ein A-Skalarprodukt*** *auf* $\mathbb{R}^n$ *mittels*

$$\langle \mathbf{x}, \mathbf{y} \rangle_{\mathbf{A}} := \mathbf{x}^T \mathbf{A}\, \mathbf{y}.$$

einführen. Bzgl. dieses Skalarprodukts sind die Vektoren $\mathbf{p}_1, \mathbf{p}_2, \ldots, \mathbf{p}_m$ *dann orthogonal, d.h. es gilt* $\langle \mathbf{p}_j, \mathbf{p}_k \rangle_{\mathbf{A}} = 0$ *für alle* $j, k = 1, 2, \ldots, m$ *mit* $j \neq k$*. Wir sprechen auch von* ***A-orthogonalen*** *Vektoren.*

Beweis von Hilfssatz 3.21:

(1) Aussage (1) folgt direkt aus der Definition von positiv definit.

(2) Um zu zeigen, dass die Vektoren $\mathbf{p}_1, \mathbf{p}_2, \ldots, \mathbf{p}_m$ linear unabhängig sind, betrachten wir (mit $k \in \{1, 2, \ldots, m\}$)

$$c_1\, \mathbf{p}_1 + c_2\, \mathbf{p}_2 + \ldots + c_k\, \mathbf{p}_k + \ldots + c_m\, \mathbf{p}_m = \mathbf{0}, \tag{3.36}$$

und multiplizieren von links mit $\mathbf{p}_k^T \mathbf{A}$. Dann erhalten wir

$$c_1\, \mathbf{p}_k^T \mathbf{A}\, \mathbf{p}_1 + c_2\, \mathbf{p}_k^T \mathbf{A}\, \mathbf{p}_2 + \ldots + c_k\, \mathbf{p}_k^T \mathbf{A}\, \mathbf{p}_k + \ldots + c_m\, \mathbf{p}_k^T \mathbf{A}\, \mathbf{p}_m = \mathbf{p}^T \mathbf{0} = 0.$$

Da $\mathbf{p}_k^T \mathbf{A}\, \mathbf{p}_j = 0$ für $j \neq k$ gilt, vereinfacht sich die Gleichung zu

$$c_k\, \mathbf{p}_k^T \mathbf{A}\, \mathbf{p}_k = 0 \qquad \Longleftrightarrow \qquad c_k = 0 \;\; (\text{wegen } \mathbf{p}_k^T \mathbf{A}\, \mathbf{p}_k > 0).$$

Da k beliebig war, folgt, dass alle Koeffizienten $c_1, c_2, \ldots, c_m$ in (3.36) alle null sein müssen. Daraus folgt, dass $\mathbf{p}_1, \mathbf{p}_2, \ldots, \mathbf{p}_m$ linear unabhängig sind.

(3) Aus Hilfssatz 3.21 (2) wissen wir, dass die $\mathbf{A}$-konjugierten Vektoren linear unabhängig sind. Da aber in $\mathbb{R}^n$ mehr als n Vektoren immer linear abhängig sind, folgt, dass $m \leq n$ gelten muss.

(4) Wir müssen zeigen, dass $\langle \cdot, \cdot \rangle_{\mathbf{A}} : \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}$ die folgenden Skalarprodukt-Eigenschaften erfüllt:

(S1) $\langle \mathbf{x}, \mathbf{y} + \mathbf{z} \rangle_{\mathbf{A}} = \langle \mathbf{x}, \mathbf{y} \rangle_{\mathbf{A}} + \langle \mathbf{x}, \mathbf{z} \rangle_{\mathbf{A}}$ für alle $\mathbf{x}, \mathbf{y}, \mathbf{z} \in \mathbb{R}^n$.
(S2) $\langle \mathbf{x}, \alpha\, \mathbf{y} \rangle_{\mathbf{A}} = \alpha\, \langle \mathbf{x}, \mathbf{y} \rangle_{\mathbf{A}}$ für alle $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$ und alle $\alpha \in \mathbb{R}$.
(S3) $\langle \mathbf{x}, \mathbf{y} \rangle_{\mathbf{A}} = \langle \mathbf{y}, \mathbf{x} \rangle_{\mathbf{A}}$ für alle $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$.
(S4) $\langle \mathbf{x}, \mathbf{x} \rangle_{\mathbf{A}} > 0$ für alle $\mathbf{x} \in \mathbb{R}^n \setminus \{\mathbf{0}\}$.
(S5) $\langle \mathbf{x}, \mathbf{x} \rangle_{\mathbf{A}} = 0 \quad \Longleftrightarrow \quad \mathbf{x} = \mathbf{0}$

Nachweis:

(S1) und (S2): Für alle $\alpha, \beta \in \mathbb{R}$ und alle $\mathbf{x}, \mathbf{y}, \mathbf{z} \in \mathbb{R}^n$ gilt wegen der Distributivgesetze (für die Matrizenmultiplikation und die skalare Multiplikation)

$$\langle \mathbf{x}, \alpha\, \mathbf{y} + \beta\, \mathbf{z} \rangle_{\mathbf{A}} = \mathbf{x}^T \mathbf{A} \left(\alpha\, \mathbf{y} + \beta\, \mathbf{z} \right)$$

$$= \alpha\, \mathbf{x}^T \mathbf{A}\, \mathbf{y} + \beta\, \mathbf{x}^T \mathbf{A}\, \mathbf{z}$$
$$= \alpha\, \langle \mathbf{x}, \mathbf{y} \rangle_{\mathbf{A}} + \beta\, \langle \mathbf{x}, \mathbf{z} \rangle_{\mathbf{A}}.$$

Setzen wir $\beta = 0$, so folgt (S2):

$$\langle \mathbf{x}, \alpha\, \mathbf{y} \rangle_{\mathbf{A}} = \alpha\, \langle \mathbf{x}, \mathbf{y} \rangle_{\mathbf{A}} \qquad \text{für alle } \mathbf{x} \in \mathbb{R}^n \text{ und alle } \alpha \in \mathbb{R}.$$

Setzen wir $\alpha = \beta = 1$, so folgt (S1):

$$\langle \mathbf{x}, \mathbf{y} + \mathbf{z} \rangle_{\mathbf{A}} = \langle \mathbf{x}, \mathbf{y} \rangle_{\mathbf{A}} + \langle \mathbf{x}, \mathbf{z} \rangle_{\mathbf{A}} \qquad \text{für alle } \mathbf{x}, \mathbf{y}, \mathbf{z} \in \mathbb{R}^n.$$

(S3) Nach Voraussetzung gibt $\mathbf{A}^T = \mathbf{A}$, und für jede reelle Zahl b gilt (als 1×1-Matrix) trivialerweise $b^T = b$. Damit finden wir für alle $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$

$$\langle \mathbf{x}, \mathbf{y} \rangle_{\mathbf{A}} = \big(\langle \mathbf{x}, \mathbf{y} \rangle_{\mathbf{A}} \big)^T = \big(\mathbf{x}^T \mathbf{A}\, \mathbf{y} \big)^T = \mathbf{y}^T \mathbf{A}^T \mathbf{x} = \mathbf{y}^T \mathbf{A}\, \mathbf{x} = \langle \mathbf{y}, \mathbf{x} \rangle_{\mathbf{A}},$$

wobei wir im vorletzten Schritt $\mathbf{A}^T = \mathbf{A}$ genutzt haben.

(S4) Da $\mathbf{A}$ positiv definit ist, gilt

$$\langle \mathbf{x}, \mathbf{x} \rangle_{\mathbf{A}} = \mathbf{x}^T \mathbf{A}\, \mathbf{x} > 0 \qquad \text{für alle } \mathbf{x} \in \mathbb{R}^n \setminus \{\mathbf{0}\}.$$

(S5) Sei $\mathbf{x} = \mathbf{0}$. Es gilt offenbar $\langle \mathbf{0}, \mathbf{0} \rangle_{\mathbf{A}} = \mathbf{0}^T \mathbf{A}\, \mathbf{0} = 0$, d.h. aus $\mathbf{x} = \mathbf{0}$ folgt $\langle \mathbf{x}, \mathbf{x} \rangle_{\mathbf{A}} = 0$. – Ist umgekehrt $\langle \mathbf{x}, \mathbf{x} \rangle_{\mathbf{A}} = 0$, dann folgt wegen Eigenschaft (S4) und $\langle \mathbf{0}, \mathbf{0} \rangle = 0$, dass $\mathbf{x} = \mathbf{0}$ gelten muss. □

Für $\mathbf{A}$-konjugierte Richtungsvektoren $\mathbf{p}_1, \mathbf{p}_2, \ldots, \mathbf{p}_n$ stoppt der generische Minimierungsalgorithmus (siehe Verfahren 3.19) nach höchstens n Iterationsschritten und liefert dann die exakte Lösung $\widehat{\mathbf{x}}$ von $\mathbf{A}\, \mathbf{x} = \mathbf{b}$. Natürlich wird man in der Praxis den Algorithmus nicht alle Iterationsschritte ausführen lassen, sondern stoppen, wenn eine hinreichend gute Genauigkeit erreicht ist.

Satz 3.22. (Verfahren 3.19 mit A-konjugierten Richtungsvektoren)
Seien $\mathbf{A} \in \mathbb{R}^{n \times n}$ *eine* ***positiv definite, symmetrische Matrix*** *und* $\mathbf{b} \in \mathbb{R}^n$. *Sei* $\mathbf{x}_1 \in \mathbb{R}^n$ *eine beliebiger Startvektor und seien die Richtungsvektoren* $\mathbf{p}_1, \mathbf{p}_2, \ldots, \mathbf{p}_n \in \mathbb{R}^n \setminus \{\mathbf{0}\}$ *in Verfahren 3.19* ***A-konjugiert****. Dann liefert der* ***generische Minimierungsalgorithmus*** *Verfahren 3.19* ***nach höchstens*** n ***Schritten*** *die eindeutig bestimmte Lösung* $\widehat{\mathbf{x}}$ *von* $\mathbf{A}\, \mathbf{x} = \mathbf{b}$.

Beweis von Satz 3.22: Da die n Richtungsvektoren $\mathbf{p}_1, \mathbf{p}_2, \ldots, \mathbf{p}_n$ $\mathbf{A}$-konjugiert sind, sind sie nach Hilfssatz 3.21 (2) linear unabhängig und bilden somit eine Basis von $\mathbb{R}^n$. Daher können wir den Vektor $\widehat{\mathbf{x}} - \mathbf{x}_1$ eindeutig als Linearkombination bzgl. dieser Basis darstellen: Es gibt also eindeutig bestimmte Koeffizienten

$\beta_1, \beta_2, \ldots, \beta_n \in \mathbb{R}$, so dass gilt:

$$\widehat{\mathbf{x}} - \mathbf{x}_1 = \sum_{j=1}^{n} \beta_j \, \mathbf{p}_j \qquad \Longleftrightarrow \qquad \widehat{\mathbf{x}} = \mathbf{x}_1 + \sum_{j=1}^{n} \beta_j \, \mathbf{p}_j \tag{3.37}$$

Nach Verfahren 3.19 gilt $\mathbf{x}_{j+1} = \mathbf{x}_j + \alpha_j \, \mathbf{p}_j$, $j = 1, 2, \ldots, n$. Daher kann man (durch wiederholte Anwendung der Formel $\mathbf{x}_{j+1} = \mathbf{x}_j + \alpha_j \, \mathbf{p}_j$) die Iterierte $\mathbf{x}_k$ wie folgt darstellen:

$$\mathbf{x}_k = \mathbf{x}_1 + \sum_{j=1}^{k-1} \alpha_j \, \mathbf{p}_j, \qquad k = 1, 2, \ldots, n, n+1. \tag{3.38}$$

Vergleicht man (3.37) und (3.38) für $k = n+1$, dann sieht man, dass es ausreichend ist zu zeigen, dass $\alpha_j = \beta_j$ für alle $j = 1, 2, \ldots, n$ gilt. Dann folgt $\mathbf{x}_{n+1} = \widehat{\mathbf{x}}$.

In Verfahren 3.19 berechnet sich α_i mit der Formel

$$\alpha_i = \frac{\mathbf{p}_i^T \, \mathbf{r}_i}{\mathbf{p}_i^T \, \mathbf{A} \, \mathbf{p}_i}, \qquad i = 1, 2, \ldots, n. \tag{3.39}$$

Um die Koeffizienten β_i, $i = 1, 2, \ldots, n$, zu berechnen, bemerken wir, dass aus (3.37) mit $\mathbf{b} = \mathbf{A} \, \widehat{\mathbf{x}}$ folgt

$$\mathbf{b} - \mathbf{A} \, \mathbf{x}_1 = \mathbf{A} \, \widehat{\mathbf{x}} - \mathbf{A} \, \mathbf{x}_1 = \mathbf{A} \left(\widehat{\mathbf{x}} - \mathbf{x}_1 \right) = \mathbf{A} \left(\sum_{j=1}^{n} \beta_j \, \mathbf{p}_j \right) = \sum_{j=1}^{n} \beta_j \, \mathbf{A} \, \mathbf{p}_j. \tag{3.40}$$

Multiplizieren von (3.40) von links mit $\mathbf{p}_i^T$ liefert:

$$\mathbf{p}_i^T \left(\mathbf{b} - \mathbf{A} \, \mathbf{x}_1 \right) = \mathbf{p}_i^T \left(\sum_{j=1}^{n} \beta_j \, \mathbf{A} \, \mathbf{p}_j \right) = \sum_{j=1}^{n} \beta_j \, \mathbf{p}_i^T \mathbf{A} \, \mathbf{p}_j = \beta_i \, \mathbf{p}_i^T \mathbf{A} \, \mathbf{p}_i, \tag{3.41}$$

wobei wir im letzten Schritt genutzt haben, dass $\mathbf{p}_i^T \mathbf{A} \, \mathbf{p}_j = 0$ für $j \neq i$ (da die Richtungsvektoren $\mathbf{p}_1, \mathbf{p}_2, \ldots, \mathbf{p}_n$ $\mathbf{A}$-konjugiert sind). Durch Auflösen von (3.41) nach β_i erhalten wir die folgende explizite Formel für β_i:

$$\beta_i = \frac{\mathbf{p}_i^T \left(\mathbf{b} - \mathbf{A} \, \mathbf{x}_1 \right)}{\mathbf{p}_i^T \mathbf{A} \, \mathbf{p}_i} = \frac{\mathbf{p}_i^T \, \mathbf{r}_1}{\mathbf{p}_i^T \mathbf{A} \, \mathbf{p}_i}, \qquad i = 1, 2, \ldots, n. \tag{3.42}$$

Diese unterschiedet sich auf den ersten Blick von (3.39). Allerdings folgt aus (3.38) für das Residuum $\mathbf{r}_i$

$$\mathbf{r}_i = \mathbf{b} - \mathbf{A} \, \mathbf{x}_i = \mathbf{b} - \mathbf{A} \left(\mathbf{x}_1 + \sum_{j=1}^{i-1} \alpha_j \, \mathbf{p}_j \right)$$

$$= \big(\mathbf{b} - \mathbf{A}\,\mathbf{x}_1\big) - \sum_{j=1}^{i-1} \alpha_j\,\mathbf{A}\,\mathbf{p}_j = \mathbf{r}_1 - \sum_{j=1}^{i-1} \alpha_j\,\mathbf{A}\,\mathbf{p}_j. \tag{3.43}$$

Multipliziert man in (3.43) von links mit $\mathbf{p}_i^T$ und nutzt anschließend $\mathbf{p}_i^T\mathbf{A}\,\mathbf{p}_j = 0$ für $j \neq i$ (da die Richtungsvektoren $\mathbf{p}_1, \mathbf{p}_2, \ldots, \mathbf{p}_n$ $\mathbf{A}$-konjugiert sind) aus, so folgt

$$\begin{aligned}\mathbf{p}_i^T\,\mathbf{r}_i &= \mathbf{p}_i^T \left(\mathbf{r}_1 - \sum_{j=1}^{i-1} \alpha_j\,\mathbf{A}\,\mathbf{p}_j\right) = \mathbf{p}_i^T\,\mathbf{r}_1 - \sum_{j=1}^{i-1} \alpha_j\,\mathbf{p}_i^T\mathbf{A}\,\mathbf{p}_j \\ &= \mathbf{p}_i^T\,\mathbf{r}_1, \qquad i = 1, 2, \ldots, n. \end{aligned} \tag{3.44}$$

Aus (3.42), (3.44) und (3.39) folgt nun

$$\beta_i = \frac{\mathbf{p}_i^T\,\mathbf{r}_1}{\mathbf{p}_i^T\mathbf{A}\,\mathbf{p}_i} = \frac{\mathbf{p}_i^T\,\mathbf{r}_i}{\mathbf{p}_i^T\mathbf{A}\,\mathbf{p}_i} = \alpha_i, \qquad i = 1, 2, \ldots, n.$$

Also ergibt sich aus (3.37) und aus (3.38) mit $k = n + 1$ und aus $\alpha_i = \beta_i$, $i = 1, 2, \ldots, n$, dass gilt $\mathbf{x}_{n+1} = \widehat{\mathbf{x}}$. Damit liefert das Verfahren nach höchstens n Schritten die eindeutig bestimmte Lösung $\widehat{\mathbf{x}}$ von $\mathbf{A}\,\mathbf{x} = \mathbf{b}$. □

Wir müssen noch diskutieren, **wie die A-konjugierten Richtungsvektoren gewählt werden sollen**. Am geschicktesten wäre es sicherlich, diese nicht vorab sondern in jedem einzelnen Iterationsschritt zu wählen. Wir hoffen dann auch, dass das Verfahren schon nach weniger als n Schritten abbricht.

Die erste Richtungsvektor ergibt sich aus $\mathbf{p}_1 = \mathbf{r}_1 = \mathbf{b} - \mathbf{A}\,\mathbf{x}_1$. Wir nehmen nun an, dass wir bereits die $\mathbf{A}$-konjugierten Richtungsvektoren $\mathbf{p}_1, \mathbf{p}_2, \ldots, \mathbf{p}_j$ bestimmt haben. Falls das Verfahren mit $\mathbf{x}_{j+1} = \widehat{\mathbf{x}}$ abbricht, sind wir fertig. Andernfalls müssen wir eine weiteren Richtungsvektor $\mathbf{p}_{j+1}$ bestimmen, der zu $\mathbf{p}_1, \mathbf{p}_2, \ldots, \mathbf{p}_j$ $\mathbf{A}$-konjugiert ist. Wir machen dazu den folgenden Ansatz

$$\mathbf{p}_{j+1} = \underbrace{\mathbf{b} - \mathbf{A}\,\mathbf{x}_{j+1}}_{=\,\mathbf{r}_{j+1}} + \sum_{k=1}^{j} \beta_{j,k}\,\mathbf{p}_k = \mathbf{r}_{j+1} + \sum_{k=1}^{j} \beta_{j,k}\,\mathbf{p}_k. \tag{3.45}$$

Der neue Richtungsvektor soll also das **Residuum $\mathbf{r}_{j+1}$ plus eine geeignete Linearkombination der vorherigen Richtungsvektoren** $\mathbf{p}_1, \mathbf{p}_2, \ldots, \mathbf{p}_j$ sein. Die Koeffizienten $\beta_{j,k}$, $k = 1, 2, \ldots, j$, sind durch die Bedingungen, dass $\mathbf{p}_{j+1}$ zu $\mathbf{p}_1, \mathbf{p}_2, \ldots, \mathbf{p}_j$ $\mathbf{A}$-konjugiert ist, eindeutig festgelegt. Wir multiplizieren also (3.45) von links jeweils mit $\mathbf{p}_i^T\mathbf{A}$, $i = 1, 2, \ldots, j$, und fordern $\mathbf{p}_i^T\mathbf{A}\,\mathbf{p}_{j+1} \stackrel{!}{=} 0$:

$$0 = \mathbf{p}_i^T\mathbf{A}\,\mathbf{p}_{j+1} = \mathbf{p}_i^T\mathbf{A} \left(\mathbf{r}_{j+1} + \sum_{k=1}^{j} \beta_{j,k}\,\mathbf{p}_k\right)$$

$$= \mathbf{p}_i^T \mathbf{A}\, \mathbf{r}_{j+1} + \sum_{k=1}^{j} \beta_{j,k}\, \mathbf{p}_i^T \mathbf{A}\, \mathbf{p}_k = \mathbf{p}_i^T \mathbf{A}\, \mathbf{r}_{j+1} + \beta_{j,i}\, \mathbf{p}_i^T \mathbf{A}\, \mathbf{p}_i, \tag{3.46}$$

wobei wir im letzten Schritt $\mathbf{p}_i^T \mathbf{A}\, \mathbf{p}_k = 0$ für $k \neq i$ genutzt haben. Auflösen von (3.46) nach $\beta_{j,i}$ liefert

$$\beta_{j,i} = -\frac{\mathbf{p}_i^T \mathbf{A}\, \mathbf{r}_{j+1}}{\mathbf{p}_i^T \mathbf{A}\, \mathbf{p}_i} = -\frac{(\mathbf{p}_i^T \mathbf{A}\, \mathbf{r}_{j+1})^T}{\mathbf{p}_i^T \mathbf{A}\, \mathbf{p}_i} = -\frac{\mathbf{r}_{j+1}^T \mathbf{A}\, \mathbf{p}_i}{\mathbf{p}_i^T \mathbf{A}\, \mathbf{p}_i}, \qquad i = 1, 2, \ldots, j. \tag{3.47}$$

Man kann zeigen, dass die Koeffizienten $\beta_{j,1}, \ldots, \beta_{j,j-1}$ in (3.47) **automatisch null** sind (siehe [7, Remark 6.10]), so dass

$$\boxed{\beta_{j+1} := \beta_{j,j} = -\frac{\mathbf{r}_{j+1}^T \mathbf{A}\, \mathbf{p}_j}{\mathbf{p}_j^T \mathbf{A}\, \mathbf{p}_j}} \tag{3.48}$$

der einzige Koeffizient ungleich null ist. Damit folgt aus (3.45) die folgende Formel für den neuen Richtungsvektor

$$\boxed{\mathbf{p}_{j+1} = \mathbf{r}_{j+1} + \beta_{j+1}\, \mathbf{p}_j.} \tag{3.49}$$

Der generische Minimierungsalgorithmus Verfahren 3.19 mit der Wahl (3.49) mit (3.48) für die Richtungsvektoren liefert das **Verfahren der konjugierten Gradienten** (oder **CG-Verfahren**):

Verfahren 3.23. (Verfahren der konjugierten Gradienten)
Gegeben seien eine ***positive definite, symmetrische Matrix*** $\mathbf{A} \in \mathbb{R}^{n \times n}$ *und eine rechte Seite* $\mathbf{b} \in \mathbb{R}^n$. *Weiter seien ein Startvektor* $\mathbf{x}_1 \in \mathbb{R}^n$ *gegeben.*

Initialisierung: Setze $\mathbf{p}_1 = \mathbf{r}_1 = \mathbf{b} - \mathbf{A}\, \mathbf{x}_1$.

Für $j = 1, 2, \ldots, n$ *führe die folgenden Schritte durch*

(1) Berechne die Schrittweite $\alpha_j = \dfrac{\mathbf{p}_j^T \mathbf{r}_j}{\mathbf{p}_j^T \mathbf{A}\, \mathbf{p}_j}$.

(2) Berechne die verbesserte Näherung $\mathbf{x}_{j+1} = \mathbf{x}_j + \alpha_j\, \mathbf{p}_j$.

(3) Berechne das Residuum $\mathbf{r}_{j+1} = \mathbf{b} - \mathbf{A}\, \mathbf{x}_{j+1}$.

(4) Berechne $\beta_{j+1} = -\dfrac{\mathbf{r}_{j+1}^T \mathbf{A}\, \mathbf{p}_j}{\mathbf{p}_j^T \mathbf{A}\, \mathbf{p}_j}$.

(5) Wähle den neuen Richtungsvektor $\mathbf{p}_{j+1} = \mathbf{r}_{j+1} + \beta_{j+1}\, \mathbf{p}_j$.

bis in Schritt (3) $\mathbf{r}_{j+1} = \mathbf{0}$ *auftritt. Dann stoppe.*

Wir zeigen nun, dass die in Verfahren 3.23 gewählten Vektoren $\mathbf{p}_1, \mathbf{p}_2, \ldots$ in der Tat $\mathbf{A}$-konjugiert sind und dass das Verfahren nach spätestens n Iterationsschritten abbricht und die Lösung $\widehat{\mathbf{x}}$ von $\mathbf{A}\,\mathbf{x} = \mathbf{b}$ liefert.

Satz 3.24. (Eigenschaften des CG-Verfahrens)
Gegeben seien eine ***positive definite, symmetrische Matrix*** $\mathbf{A} \in \mathbb{R}^{n\times n}$ *und eine rechte Seite* $\mathbf{b} \in \mathbb{R}^n$*. Weiter seien ein beliebiger Startvektor* $\mathbf{x}_1 \in \mathbb{R}^n$ *gegeben. Die im* ***CG-Verfahren*** *(siehe Verfahren 3.23) vorkommenden* ***Richtungsvektoren*** $\mathbf{p}_j$ *und* ***Residuen*** $\mathbf{r}_j$ *haben die folgenden Eigenschaften:*

$$\begin{aligned}
&(1) \quad \mathbf{p}_i^T\,\mathbf{A}\,\mathbf{p}_j = 0 \qquad \text{für } \; j = 1, 2 \ldots, i-1,\\
&(2) \quad \mathbf{r}_i^T\,\mathbf{r}_j = 0 \qquad \text{für } \; j = 1, 2 \ldots, i-1,\\
&(3) \quad \mathbf{p}_j^T\,\mathbf{r}_j = \mathbf{r}_j^T\mathbf{r}_j \qquad \text{für } \; j = 1, 2 \ldots, i.
\end{aligned}$$

Das CG-Verfahren ***bricht nach höchstens*** n ***Schritten ab****. Wenn das CG-Verfahren mit* j *Schritten abbricht, dann sind* $\mathbf{r}_{j+1} = \mathbf{b} - \mathbf{A}\,\mathbf{x}_{j+1} = \mathbf{0}$ *and* $\mathbf{p}_{j+1} = \mathbf{0}$*, d.h.* $\mathbf{x}_{j+1}$ *ist die* ***eindeutig bestimmte Lösung*** $\widehat{\mathbf{x}}$ *von* $\mathbf{A}\,\mathbf{x} = \mathbf{b}$*, also* $\mathbf{x}_{j+1} = \widehat{\mathbf{x}}$*. Also liefert das CG-Verfahren nach höchstens* n *Schritten die eindeutige Lösung* $\widehat{\mathbf{x}}$ *von* $\mathbf{A}\,\mathbf{x} = \mathbf{b}$*.*

Der nachfolgende Beweis von Satz 3.24 ist technisch und wird nicht in der Vorlesung besprochen. Er ist für mathematisch Interessierte der Vollständigkeit halber im Skript aufgeführt.

Beweis von Satz 3.24: Wir beweisen zunächst Eigenschaften (1) bis (3) mit vollständiger Induktion über i:

Induktionsanfang: Für $i = 1$ gibt es für (1) und (2) nichts zu zeigen; (3) folgt automatisch, da $\mathbf{p}_1 = \mathbf{r}_1$. Also betrachten wir als Anfangsschritt für alle drei Eigenschaften $i = 2$. Wir finden

$$\begin{aligned}
\mathbf{p}_2^T\,\mathbf{A}\,\mathbf{p}_1 &= (\mathbf{r}_2 + \beta_2\,\mathbf{p}_1)^T\,\mathbf{A}\,\mathbf{p}_1 = \mathbf{r}_2^T\,\mathbf{A}\,\mathbf{p}_1 + \beta_2\,\mathbf{p}_1^T\,\mathbf{A}\,\mathbf{p}_1\\
&= \mathbf{r}_2^T\,\mathbf{A}\,\mathbf{p}_1 - \frac{\mathbf{r}_2^T\,\mathbf{A}\,\mathbf{p}_1}{\mathbf{p}_1^T\,\mathbf{A}\,\mathbf{p}_1}\,\mathbf{p}_1^T\,\mathbf{A}\,\mathbf{p}_1 = 0,
\end{aligned}$$

wobei wir die Formel für β_2 eingesetzt haben. Dieses zeigt, dass (1) für $i = 2$ wahr ist. – Um (2) für $i = 2$ zu überprüfen, nutzen wir $\mathbf{p}_1 = \mathbf{r}_1$ und die Formel für α_1:

$$\mathbf{r}_2^T\,\mathbf{r}_1 = (\mathbf{b} - \mathbf{A}\,\mathbf{x}_2)^T\,\mathbf{r}_1 = \big(\mathbf{b} - \mathbf{A}\,(\mathbf{x}_1 + \alpha_1\,\mathbf{p}_1)\big)^T\,\mathbf{r}_1$$

$$= \big(\underbrace{(\mathbf{b} - \mathbf{A}\,\mathbf{x}_1)}_{=\mathbf{r}_1} - \alpha_1\,\mathbf{A}\,\mathbf{p}_1 \big)^T \mathbf{r}_1 = (\mathbf{r}_1 - \alpha_1\,\mathbf{A}\,\mathbf{p}_1)^T\,\mathbf{r}_1 \overset{\mathbf{r}_1 = \mathbf{p}_1}{\overset{\downarrow}{=}} (\mathbf{r}_1 - \alpha_1\,\mathbf{A}\,\mathbf{p}_1)^T\,\mathbf{p}_1$$

$$= \mathbf{r}_1^T\,\mathbf{p}_1 - \alpha_1\,\mathbf{p}_1^T\mathbf{A}\,\mathbf{p}_1 = \mathbf{r}_1^T\,\mathbf{p}_1 - \frac{\mathbf{p}_1^T\,\mathbf{r}_1}{\mathbf{p}_1^T\mathbf{A}\,\mathbf{p}_1}\,\mathbf{p}_1^T\mathbf{A}\,\mathbf{p}_1 = \mathbf{r}_1^T\,\mathbf{p}_1 - \mathbf{p}_1^T\,\mathbf{r}_1 = 0,$$

womit (2) für $i = 2$ nachgewiesen ist. – Schließlich gilt

$$\mathbf{p}_2^T\,\mathbf{r}_2 = (\mathbf{r}_2 + \beta_2\,\mathbf{p}_1)^T\,\mathbf{r}_2 = \mathbf{r}_2^T\,\mathbf{r}_2 + \beta_2\,\mathbf{p}_1^T\,\mathbf{r}_2,$$

und (3) ist für $i = 2$ nachgewiesen, wenn wir zeigen können, dass $\mathbf{p}_1^T\,\mathbf{r}_2 = 0$ ist. Dazu gehen wir wie folgt vor:

$$\mathbf{p}_1^T\,\mathbf{r}_2 = \mathbf{p}_1^T\,(\mathbf{b} - \mathbf{A}\,\mathbf{x}_2) = \mathbf{p}_1^T\,\big(\mathbf{b} - \mathbf{A}\,(\mathbf{x}_1 + \alpha_1\,\mathbf{p}_1)\big) = \mathbf{p}_1^T\,\big(\underbrace{(\mathbf{b} - \mathbf{A}\,\mathbf{x}_1)}_{=\mathbf{r}_1} - \alpha_1\,\mathbf{A}\,\mathbf{p}_1\big)$$

$$= \mathbf{p}_1^T\,(\mathbf{r}_1 - \alpha_1\,\mathbf{A}\,\mathbf{p}_1) = \mathbf{p}_1^T\,\mathbf{r}_1 - \alpha_1\,\mathbf{p}_1^T\,\mathbf{A}\,\mathbf{p}_1 = \mathbf{p}_1^T\,\mathbf{r}_1 - \frac{\mathbf{p}_1^T\,\mathbf{r}_1}{\mathbf{p}_1^T\,\mathbf{A}\,\mathbf{p}_1}\,\mathbf{p}_1^T\,\mathbf{A}\,\mathbf{p}_1 = 0,$$

wobei wir die Formel für α_1 im letzten Schritt verwendet haben. Damit ist (3) für $i = 2$ überprüft, und wir haben einen Induktionsanfang für alle drei Aussagen (1), (2) und (3) für $i = 2$.

Induktionsvoraussetzung: Wir nehmen an, dass die Aussagen (1), (2) und (3) alle für alle $j = 2, \ldots, i$ mit einem $i \geq 2$ gelten.

Induktionsschritt $i \curvearrowright i+1$: Wir zeigen zuerst, dass (2) mit $i+1$ (statt i) gilt. Per Definition gilt

$$\mathbf{r}_{i+1} = \mathbf{b} - \mathbf{A}\,\mathbf{x}_{i+1} = \mathbf{b} - \mathbf{A}\,\big(\mathbf{x}_i + \alpha_i\,\mathbf{p}_i\big) = \underbrace{\mathbf{b} - \mathbf{A}\,\mathbf{x}_i}_{=\mathbf{r}_i} - \alpha_i\,\mathbf{A}\,\mathbf{p}_i = \mathbf{r}_i - \alpha_i\,\mathbf{A}\,\mathbf{p}_i. \tag{3.50}$$

Mit (3.50), der Induktionsvoraussetzung (1) und (2) und $\mathbf{r}_j = \mathbf{p}_j - \beta_j\,\mathbf{p}_{j-1}$ (wegen $\mathbf{p}_j = \mathbf{r}_j + \beta_j\,\mathbf{p}_{j-1}$) gilt für $j = 1, 2, \ldots, i-1$ direkt

$$\begin{aligned}
\mathbf{r}_{i+1}^T\,\mathbf{r}_j &\overset{(3.50)}{\overset{\downarrow}{=}} \big(\mathbf{r}_i - \alpha_i\,\mathbf{A}\,\mathbf{p}_i\big)^T\,\mathbf{r}_j = \mathbf{r}_i^T\,\mathbf{r}_j - \alpha_i\,\mathbf{p}_i^T\mathbf{A}\,\mathbf{r}_j \\
&= \mathbf{r}_i^T\,\mathbf{r}_j - \alpha_i\,\mathbf{p}_i^T\mathbf{A}\,\big(\mathbf{p}_j - \beta_j\,\mathbf{p}_{j-1}\big) \\
&= \underbrace{\mathbf{r}_i^T\,\mathbf{r}_j}_{=0} - \alpha_i\,\underbrace{\mathbf{p}_i^T\mathbf{A}\,\mathbf{p}_j}_{=0} - \alpha_i\,\beta_j\,\underbrace{\mathbf{p}_i^T\mathbf{A}\,\mathbf{p}_{j-1}}_{=0} = 0.
\end{aligned}$$

Für $j = i$ nutzen wir (3.50), $\mathbf{r}_i = \mathbf{p}_i - \beta_i\,\mathbf{p}_{i-1}$ (wegen $\mathbf{p}_i = \mathbf{r}_i + \beta_i\,\mathbf{p}_{i-1}$), die Formel für α_i und die Identität

$$\mathbf{p}_i^T\mathbf{A}\,\mathbf{r}_i = \mathbf{p}_i^T\mathbf{A}\,\big(\mathbf{p}_i - \beta_i\,\mathbf{p}_{i-1}\big) = \mathbf{p}_i^T\mathbf{A}\,\mathbf{p}_i - \beta_i\,\underbrace{\mathbf{p}_i^T\mathbf{A}\,\mathbf{p}_{i-1}}_{=0} = \mathbf{p}_i^T\mathbf{A}\,\mathbf{p}_i \tag{3.51}$$

(wobei die Induktionsvoraussetzung (1) im letzten Schritt genutzt wurde) und die Induktionsvoraussetzung (3), um zu folgern, dass gilt

$$\mathbf{r}_{i+1}^T \mathbf{r}_i \overset{(3.50)}{\downarrow} \big(\mathbf{r}_i - \alpha_i \mathbf{A}\,\mathbf{p}_i\big)^T \mathbf{r}_i = \mathbf{r}_i^T \mathbf{r}_i - \alpha_i \mathbf{p}_i^T \mathbf{A}\,\mathbf{r}_i = \mathbf{r}_i^T \mathbf{r}_i - \frac{\mathbf{p}_i^T \mathbf{r}_i}{\mathbf{p}_i^T \mathbf{A}\,\mathbf{p}_i} \mathbf{p}_i^T \mathbf{A}\,\mathbf{r}_i$$

$$\overset{(3.51)}{\downarrow} \mathbf{r}_i^T \mathbf{r}_i - \frac{\mathbf{p}_i^T \mathbf{r}_i}{\mathbf{p}_i^T \mathbf{A}\,\mathbf{p}_i} \mathbf{p}_i^T \mathbf{A}\,\mathbf{p}_i = \mathbf{r}_i^T \mathbf{r}_i - \mathbf{p}_i^T \mathbf{r}_i = 0.$$

(Im letzten Schritt wurde (3) mit $j = i$ genutzt.) Damit ist (2) für $i+1$ bewiesen.

Als Nächstes zeigen wir Aussage (1) für $i + 1$. Zunächst gilt

$$\mathbf{p}_{i+1}^T \mathbf{A}\,\mathbf{p}_j = \big(\mathbf{r}_{i+1} + \beta_{i+1}\,\mathbf{p}_i\big)^T \mathbf{A}\,\mathbf{p}_j = \mathbf{r}_{i+1}^T \mathbf{A}\,\mathbf{p}_j + \beta_{i+1}\,\mathbf{p}_i^T \mathbf{A}\,\mathbf{p}_j. \tag{3.52}$$

Für $j = i$ folgt aus (3.52) mit der Formel für β_{i+1}

$$\begin{aligned}\mathbf{p}_{i+1}^T \mathbf{A}\,\mathbf{p}_i &= \mathbf{r}_{i+1}^T \mathbf{A}\,\mathbf{p}_i + \beta_{i+1}\,\mathbf{p}_i^T \mathbf{A}\,\mathbf{p}_i \\ &= \mathbf{r}_{i+1}^T \mathbf{A}\,\mathbf{p}_i - \frac{\mathbf{r}_{i+1}^T \mathbf{A}\,\mathbf{p}_i}{\mathbf{p}_i^T \mathbf{A}\,\mathbf{p}_i} \mathbf{p}_i^T \mathbf{A}\,\mathbf{p}_i = \mathbf{r}_{i+1}^T \mathbf{A}\,\mathbf{p}_i - \mathbf{r}_{i+1}^T \mathbf{A}\,\mathbf{p}_i = 0.\end{aligned}$$

Für $j = 1, 2, \ldots, i-1$ bemerken wir zunächst, dass, falls $\alpha_j = 0$ ist, (mit der Formel für α_j) direkt $\mathbf{p}_j^T \mathbf{r}_j = 0$ folgt. Weiter gilt dann nach der Induktionsvoraussetzung (3) $0 = \mathbf{p}_j^T \mathbf{r}_j = \mathbf{r}_j^T \mathbf{r}_j$ was impliziert, dass $\mathbf{r}_j = \mathbf{0}$ ist. Also würde die Iteration stoppen, wenn $\alpha_j = 0$ ist, und dann ist nichts mehr zu zeigen. Also nehmen wir an, dass $\alpha_j \neq 0$ gilt. Dann folgt aus (3.50)

$$\mathbf{r}_{i+1} = \mathbf{r}_i - \alpha_i \mathbf{A}\,\mathbf{p}_i \qquad \Longleftrightarrow \qquad \mathbf{A}\,\mathbf{p}_j = \frac{1}{\alpha_j}\big(\mathbf{r}_j - \mathbf{r}_{j+1}\big). \tag{3.53}$$

Mit der Formel für $\mathbf{p}_{i+1}$, der Induktionsvoraussetzung (1) und Formel (2) für $i+1$ (welche wir bereits bewiesen haben) folgt daraus für $j = 1, 2, \ldots, i-1$

$$\mathbf{p}_{i+1}^T \mathbf{A}\,\mathbf{p}_j = \big(\mathbf{r}_{i+1} + \beta_{i+1}\,\mathbf{p}_i\big)^T \mathbf{A}\,\mathbf{p}_j = \mathbf{r}_{i+1}^T \mathbf{A}\,\mathbf{p}_j + \beta_{i+1} \underbrace{\mathbf{p}_i^T \mathbf{A}\,\mathbf{p}_j}_{=0}$$

$$= \mathbf{r}_{i+1}^T \mathbf{A}\,\mathbf{p}_j \overset{(3.53)}{\downarrow} \frac{1}{\alpha_j} \mathbf{r}_{i+1}^T \big(\mathbf{r}_j - \mathbf{r}_{j+1}\big) = \frac{1}{\alpha_j}\big(\underbrace{\mathbf{r}_{i+1}^T \mathbf{r}_j}_{=0} - \underbrace{\mathbf{r}_{i+1}^T \mathbf{r}_{j+1}}_{=0}\big) = 0.$$

Dieses beweist Aussage (1) für $i + 1$.

Um (3) für $i + 1$ nachzuweisen, müssen wir nur $\mathbf{p}_{i+1}^T \mathbf{r}_{i+1} = \mathbf{r}_{i+1}^T \mathbf{r}_{i+1}$ zeigen, denn für $j = 1, 2, \ldots, i$ gilt $\mathbf{p}_j^T \mathbf{r}_j = \mathbf{r}_j^T \mathbf{r}_j$ per Induktionsvoraussetzung (3). Um

$\mathbf{p}_{i+1}^T \mathbf{r}_{i+1} = \mathbf{r}_{i+1}^T \mathbf{r}_{i+1}$ zu zeigen, nutzen wir, dass $f(\mathbf{x}) = \frac{1}{2}\mathbf{x}^T\mathbf{A}\mathbf{x} - \mathbf{b}^T\mathbf{x}$ per Konstruktion sein Minimum entlang der Geraden $\mathbf{x}_{i+1} + t\,\mathbf{p}_i$, $t \in \mathbb{R}$, bei $\mathbf{x}_{i+1}$, also bei $t = 0$, annimmt. Dieses bedeutet, dass die Ableitung der Funktion $\phi : \mathbb{R} \to \mathbb{R}$, $\phi(t) = f(\mathbf{x}_{i+1} + t\,\mathbf{p}_i)$ (wobei f der CG-Funktional ist), die Bedingung $\phi'(0) = 0$ erfüllt. Mit der Kettenregel folgt daraus (vgl. (3.30))

$$0 = \phi'(0) = \mathbf{p}_i^T\,(\nabla f)(\mathbf{x}_{i+1}) = \mathbf{p}_i^T\,\big(\mathbf{A}\,\mathbf{x}_{i+1} - \mathbf{b}\big) = -\mathbf{p}_i^T\,\mathbf{r}_{i+1}.$$

Daraus folgt

$$\mathbf{p}_{i+1}^T\,\mathbf{r}_{i+1} = \big(\mathbf{r}_{i+1} + \beta_{i+1}\,\mathbf{p}_i\big)^T\,\mathbf{r}_{i+1} = \mathbf{r}_{i+1}^T\,\mathbf{r}_{i+1} + \beta_{i+1}\,\underbrace{\mathbf{p}_i^T\,\mathbf{r}_{i+1}}_{=0} = \mathbf{r}_{i+1}^T\,\mathbf{r}_{i+1},$$

womit (3) für $i+1$ bewiesen ist.

Das Verfahren 3.23 stoppt, wenn $\mathbf{r}_{j+1} = \mathbf{0}$ ist. Ist $\mathbf{r}_{j+1} \neq \mathbf{0}$, so entsteht ein weiterer $\mathbf{A}$-konjugierter Richtungsvektor $\mathbf{p}_{j+1}$, und mit diesem neuen Richtungsvektor wird ein weiterer Iterationsschritt durchgeführt. Da jeder Iterationsschritt einen Richtungsvektor $\mathbf{p}_{j+1}$ generiert und da die Richtungsvektoren nach Aussage (1) $\mathbf{A}$-konjugiert und damit (nach Hilfssatz 3.21 (2)) linear unabhängig sind, können höchstens $n-1$ Iterationsschritte durchgeführt werden, in denen ein Richtungsvektor $\mathbf{p}_{j+1} \neq 0$ entsteht. (In $\mathbb{R}^n$ sind mehr als n Vektoren immer linear abhängig.) Also tritt spätestens im n-ten Iterationsschritt die Situation auf, dass $\mathbf{p}_{j+1} = \mathbf{0}$ ist. Dann folgt mit (3)

$$0 = \mathbf{0}^T\,\mathbf{r}_{j+1} = \mathbf{p}_{j+1}^T\,\mathbf{r}_{j+1} = \mathbf{r}_{j+1}^T\,\mathbf{r}_{j+1} = \|\mathbf{r}_{j+1}\|_2^2 \qquad \Longrightarrow \qquad \mathbf{r}_{j+1} = \mathbf{0}.$$

Also gilt

$$\mathbf{0} = \mathbf{r}_{j+1} = \mathbf{b} - \mathbf{A}\,\mathbf{x}_{j+1} \quad \Longleftrightarrow \quad \mathbf{b} = \mathbf{A}\,\mathbf{x}_{j+1} \quad \Longleftrightarrow \quad \mathbf{x}_{j+1} = \mathbf{A}^{-1}\,\mathbf{b} = \widehat{\mathbf{x}}.$$

d.h. spätestens im n-ten Schritt erhalten wir die eindeutige Lösung $\widehat{\mathbf{x}}$ des linearen Gleichungssystems $\mathbf{A}\,\mathbf{x} = \mathbf{b}$. □

Man kann das CG-Verfahren (Verfahren 3.23) noch verbessern, soweit die Implementierung betroffen ist. Dazu werden einige Schritte im Verfahren so umgeformt, dass sich der Rechenaufwand reduziert. Wegen Satz 3.24 (3) gilt

$$\alpha_j = \frac{\mathbf{p}_j^T\,\mathbf{r}_j}{\mathbf{p}_j^T\mathbf{A}\,\mathbf{p}_j} = \frac{\mathbf{r}_j^T\,\mathbf{r}_j}{\mathbf{p}_j^T\mathbf{A}\,\mathbf{p}_j} = \frac{\|\mathbf{r}_j\|_2^2}{\mathbf{p}_j^T\mathbf{A}\,\mathbf{p}_j}\,. \tag{3.54}$$

Weiter gilt wegen $\mathbf{x}_{j+1} = \mathbf{x}_j + \alpha_j\,\mathbf{p}_j$

$$\mathbf{r}_{j+1} = \mathbf{b} - \mathbf{A}\,\mathbf{x}_{j+1} = \mathbf{b} - \mathbf{A}\,\big(\mathbf{x}_j + \alpha_j\,\mathbf{p}_j\big) = \underbrace{\mathbf{b} - \mathbf{A}\,\mathbf{x}_j}_{=\mathbf{r}_j} - \alpha_j\,\mathbf{A}\,\mathbf{p}_j = \mathbf{r}_j - \alpha_j\,\mathbf{A}\,\mathbf{p}_j$$

$$\Longrightarrow \qquad \mathbf{r}_{j+1} = \mathbf{r}_j - \alpha_j \, \mathbf{A}\, \mathbf{p}_j \qquad \Longleftrightarrow \qquad \alpha_j \, \mathbf{A}\, \mathbf{p}_j = \mathbf{r}_j - \mathbf{r}_{j+1}. \tag{3.55}$$

Mit (3.55) folgt $\mathbf{A}\, \mathbf{p}_j = \alpha_j^{-1} \, (\mathbf{r}_j - \mathbf{r}_{j+1})$, so dass

$$\begin{aligned} \beta_{j+1} &= - \frac{\mathbf{r}_{j+1}^T \, \mathbf{A}\, \mathbf{p}_j}{\mathbf{p}_j^T \mathbf{A}\, \mathbf{p}_j} = - \frac{\mathbf{r}_{j+1}^T \, \alpha_j^{-1} \, (\mathbf{r}_j - \mathbf{r}_{j+1})}{\mathbf{p}_j^T \mathbf{A}\, \mathbf{p}_j} = - \frac{1}{\alpha_j} \, \frac{\overbrace{\mathbf{r}_{j+1}^T \, \mathbf{r}_j}^{=0} - \mathbf{r}_{j+1}^T \, \mathbf{r}_{j+1}}{\mathbf{p}_j^T \mathbf{A}\, \mathbf{p}_j} \\ &= \frac{1}{\alpha_j} \, \frac{\mathbf{r}_{j+1}^T \, \mathbf{r}_{j+1}}{\mathbf{p}_j^T \mathbf{A}\, \mathbf{p}_j} = \frac{\mathbf{p}_j^T \mathbf{A}\, \mathbf{p}_j}{\mathbf{p}_j^T \, \mathbf{r}_j} \, \frac{\mathbf{r}_{j+1}^T \, \mathbf{r}_{j+1}}{\mathbf{p}_j^T \mathbf{A}\, \mathbf{p}_j} = \frac{\mathbf{r}_{j+1}^T \mathbf{r}_{j+1}}{\mathbf{p}_j^T \mathbf{r}_j} = \frac{\mathbf{r}_{j+1}^T \mathbf{r}_{j+1}}{\mathbf{r}_j^T \mathbf{r}_j} = \frac{\|\mathbf{r}_{j+1}\|_2^2}{\|\mathbf{r}_j\|_2^2}, \end{aligned} \tag{3.56}$$

wobei wir Satz 3.24 (2), die Formel für α_j und zuletzt Satz 3.24 (3) verwendet haben. Der Vorteil an jeweils der Darstellung in (3.54), (3.55) und (3.56) ist, dass jeweils nur eine Matrix-Multiplikation $\mathbf{A}\, \mathbf{p}_j$ involviert ist, und deren Ergebnisvektor kann man nach der ersten Berechnung von $\mathbf{A}\, \mathbf{p}_j$ abspeichern und später damit weiter rechnen. Ansonsten sind nur Skalarprodukte zur Berechnung erforderlich. Mit (3.54), (3.55) und (3.56) erhalten wie die nachfolgende numerisch effizientere Version des CG-Verfahrens.

Verfahren 3.25. (Verfahren der konjugierten Gradienten)
Gegeben seien eine ***positive definite, symmetrische Matrix*** $\mathbf{A} \in \mathbb{R}^{n \times n}$ *und eine rechte Seite* $\mathbf{b} \in \mathbb{R}^n$. *Weiter seien ein Startvektor* $\mathbf{x}_1 \in \mathbb{R}^n$ *gegeben.*

Initialisierung: Setze $\mathbf{p}_1 = \mathbf{r}_1 = \mathbf{b} - \mathbf{A}\, \mathbf{x}_1$.

Für $j = 1, 2, \ldots, n$ *führe die folgenden Schritte durch*

(1) Berechne die Schrittweite $\alpha_j = \dfrac{\|\mathbf{r}_j\|_2^2}{\mathbf{p}_j^T \mathbf{A}\, \mathbf{p}_j}$.

(2) Berechne die verbesserte Näherung $\mathbf{x}_{j+1} = \mathbf{x}_j + \alpha_j \, \mathbf{p}_j$.

(3) Berechne das Residuum $\mathbf{r}_{j+1} = \mathbf{r}_j - \alpha_j \, \mathbf{A}\, \mathbf{p}_j$.

(4) Berechne $\beta_{j+1} = \dfrac{\|\mathbf{r}_{j+1}\|_2^2}{\|\mathbf{r}_j\|_2^2}$.

(5) Wähle den neuen Richtungsvektor $\mathbf{p}_{j+1} = \mathbf{r}_{j+1} + \beta_{j+1} \, \mathbf{p}_j$.

bis in Schritt (3) $\mathbf{r}_{j+1} = \mathbf{0}$ *auftritt. Dann stoppe.*

Ein Iterationsschritt des CG-Verfahrens benötigt $\mathcal{O}(n^2)$ elementare Rechenoperationen. Da das CG-Verfahren nach spätestens n Schritten stoppt, werden insgesamt höchstens $\mathcal{O}(n^3)$ elementare Rechenoperationen ausgeführt.

Betrachten wir ein Beispiel. Zur Berechnung nutzen wir dabei Verfahren 3.25.

Beispiel 3.26. (Verfahren der konjugierten Gradienten)

Wir wollen das LGS $\mathbf{A}\,\mathbf{x} = \mathbf{b}$ mit $\mathbf{A} = \begin{bmatrix} \frac{3}{2} & 0 & \frac{1}{2} \\ 0 & 3 & 0 \\ \frac{1}{2} & 0 & \frac{3}{2} \end{bmatrix}$ und $\mathbf{b} = \begin{bmatrix} 1 \\ 1 \\ -1 \end{bmatrix}$ per Hand mit dem CG-Verfahren lösen. Als Startvektor wählen wir $\mathbf{x}_1 = \mathbf{0}$.

In Beispiel 3.16 haben wir schon überprüft, dass die symmetrische Matrix $\mathbf{A}$ positiv definit ist, so dass es Sinn macht, dass CG-Verfahren anzuwenden.

Schritt 1: Wir haben

$$\mathbf{x}_1 = \mathbf{0} = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix} \quad \text{und} \quad \mathbf{p}_1 = \mathbf{r}_1 = \mathbf{b} - \mathbf{A}\,\mathbf{x}_1 = \mathbf{b} - \mathbf{A}\,\mathbf{0} = \mathbf{b} - \mathbf{0} = \begin{bmatrix} 1 \\ 1 \\ -1 \end{bmatrix}.$$

Mit $\|\mathbf{r}_1\|_2^2 = 3$ und

$$\mathbf{A}\,\mathbf{p}_1 = \begin{bmatrix} \frac{3}{2} & 0 & \frac{1}{2} \\ 0 & 3 & 0 \\ \frac{1}{2} & 0 & \frac{3}{2} \end{bmatrix} \begin{bmatrix} 1 \\ 1 \\ -1 \end{bmatrix} = \begin{bmatrix} 1 \\ 3 \\ -1 \end{bmatrix}, \qquad \mathbf{p}_1^T\,\mathbf{A}\,\mathbf{p}_1 = \begin{bmatrix} 1; 1; -1 \end{bmatrix} \begin{bmatrix} 1 \\ 3 \\ -1 \end{bmatrix} = 5$$

bekommen wir

$$\alpha_1 = \frac{\|\mathbf{r}_1\|_2^2}{\mathbf{p}_1^T\,\mathbf{A}\,\mathbf{p}_1} = \frac{3}{5}.$$

Damit ist die Iterierte

$$\mathbf{x}_2 = \mathbf{x}_1 + \alpha_1\,\mathbf{p}_1 = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix} + \frac{3}{5} \begin{bmatrix} 1 \\ 1 \\ -1 \end{bmatrix} = \frac{3}{5} \begin{bmatrix} 1 \\ 1 \\ -1 \end{bmatrix},$$

und das zugehörige Residuum ist

$$\mathbf{r}_2 = \mathbf{r}_1 - \alpha_1\,\mathbf{A}\,\mathbf{p}_1 = \begin{bmatrix} 1 \\ 1 \\ -1 \end{bmatrix} - \frac{3}{5} \begin{bmatrix} 1 \\ 3 \\ -1 \end{bmatrix} = \frac{1}{5} \begin{bmatrix} 2 \\ -4 \\ -2 \end{bmatrix}.$$

Wir finden $\|\mathbf{r}_2\|_2^2 = \dfrac{24}{25}$ and damit

$$\beta_2 = \frac{\|\mathbf{r}_2\|_2^2}{\|\mathbf{r}_1\|_2^2} = \frac{\frac{24}{25}}{3} = \frac{8}{25},$$

so dass der neue Richtungsvektor $\mathbf{p}_2$ wie folgt lautet

$$\mathbf{p}_2 = \mathbf{r}_2 + \beta_2\,\mathbf{p}_1 = \frac{1}{5}\begin{bmatrix} 2 \\ -4 \\ -2 \end{bmatrix} + \frac{8}{25}\begin{bmatrix} 1 \\ 1 \\ -1 \end{bmatrix} = \frac{1}{25}\begin{bmatrix} 18 \\ -12 \\ -18 \end{bmatrix} = \frac{6}{25}\begin{bmatrix} 3 \\ -2 \\ -3 \end{bmatrix}.$$

Schritt 2: Da $\|\mathbf{r}_2\|_2 = \sqrt{\frac{24}{25}} \neq 0$ ist, führen wir einen zweiten Schritt des CG-Verfahrens durch.

$$\mathbf{A}\,\mathbf{p}_2 = \begin{bmatrix} \frac{3}{2} & 0 & \frac{1}{2} \\ 0 & 3 & 0 \\ \frac{1}{2} & 0 & \frac{3}{2} \end{bmatrix} \frac{6}{25}\begin{bmatrix} 3 \\ -2 \\ -3 \end{bmatrix} = \frac{6}{25}\begin{bmatrix} 3 \\ -6 \\ -3 \end{bmatrix} = \frac{18}{25}\begin{bmatrix} 1 \\ -2 \\ -1 \end{bmatrix}.$$

Mit

$$\mathbf{p}_2^T\mathbf{A}\,\mathbf{p}_2 = \frac{6}{25}\,[3;-2;-3]\,\frac{18}{25}\begin{bmatrix} 1 \\ -2 \\ -1 \end{bmatrix} = \frac{18\cdot 6}{25^2}\cdot 10 = \frac{216}{125}$$

und $\|\mathbf{r}_2\|_2^2 = \frac{24}{25}$ (aus dem ersten Schritt des CG-Verfahrens) bekommen wir

$$\alpha_2 = \frac{\|\mathbf{r}_2\|_2^2}{\mathbf{p}_2^T\mathbf{A}\,\mathbf{p}_2} = \frac{\frac{24}{25}}{\frac{216}{125}} = \frac{24}{25}\cdot\frac{125}{216} = \frac{5}{9}.$$

Daher ist die zweite Iterierte durch

$$\mathbf{x}_3 = \mathbf{x}_2 + \alpha_2\,\mathbf{p}_2 = \frac{3}{5}\begin{bmatrix} 1 \\ 1 \\ -1 \end{bmatrix} + \frac{5}{9}\,\frac{6}{25}\begin{bmatrix} 3 \\ -2 \\ -3 \end{bmatrix} = \frac{1}{15}\begin{bmatrix} 9 \\ 9 \\ -9 \end{bmatrix} + \frac{1}{15}\begin{bmatrix} 6 \\ -4 \\ -6 \end{bmatrix} = \begin{bmatrix} 1 \\ \frac{1}{3} \\ -1 \end{bmatrix}$$

gegeben, und das neue Residuum ist

$$\mathbf{r}_3 = \mathbf{r}_2 - \alpha_2\,\mathbf{A}\,\mathbf{p}_2 = \frac{1}{5}\begin{bmatrix} 2 \\ -4 \\ -2 \end{bmatrix} - \frac{5}{9}\,\frac{18}{25}\begin{bmatrix} 1 \\ -2 \\ -1 \end{bmatrix} = \frac{1}{5}\begin{bmatrix} 2 \\ -4 \\ -2 \end{bmatrix} - \frac{2}{5}\begin{bmatrix} 1 \\ -2 \\ -1 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}.$$

Da $\mathbf{r}_3 = \mathbf{0}$ ist, bricht das CG-Verfahren nach zwei Schritten ab und liefert die korrekte Lösung $\widehat{\mathbf{x}} = \mathbf{x}_3 = \begin{bmatrix} 1 \\ \frac{1}{3} \\ -1 \end{bmatrix}$. ♠

Wir lernen jetzt noch einige Informationen über die **Konvergenz des Verfahren der konjugierten Gradienten (CG-Verfahren)**. Dabei wird es sich aber aus Zeitgründen nur um einen kurzen Überblick handeln. Für die Details, weitere Zwischenergebnisse und die Beweise wird z.B. auf [7, Teilkapitel 6.3] verwiesen.

In Hilfssatz 3.21 (4) haben wir gelernt, dass für jede positiv definite, symmetrische Matrix $\mathbf{A} \in \mathbb{R}^{n\times n}$ durch

$$\langle \mathbf{x}, \mathbf{y} \rangle_{\mathbf{A}} := \mathbf{x}^T \mathbf{A}\, \mathbf{y} \tag{3.57}$$

ein Skalarprodukt auf $\mathbb{R}^n$ definiert ist, das sogenannte **A-Skalarprodukt**, und bzgl. dieses $\mathbf{A}$-Skalarprodukts sind die Richtungsvektoren $\mathbf{p}_1, \mathbf{p}_2, \ldots, \mathbf{p}_m$ im CG-Verfahren **orthogonal**, da sie nach Satz 3.24 (1) $\mathbf{A}$-konjugiert sind. Das durch (3.57) definierte $\mathbf{A}$-Skalarprodukt induziert eine Norm auf $\mathbb{R}^n$ mittels

$$\|\mathbf{x}\|_{\mathbf{A}} := \sqrt{\langle \mathbf{x}, \mathbf{x} \rangle_{\mathbf{A}}} = \sqrt{\mathbf{x}^T \mathbf{A}\, \mathbf{x}}. \tag{3.58}$$

Man nennt (3.58) die **A-Norm**. Dass es sich bei (3.58) um eine Norm handelt, zeigen wir am Ende dieses Teilkapitels. Dieses folgt relativ leicht aus den Eigenschaften des Skalarprodukts (3.57).

Definition 3.27. (i-ter Krylovraum)
*Seien $\mathbf{A} \in \mathbb{R}^{n\times n}$ und $\mathbf{p} \in \mathbb{R}^n$. Für $i \in \mathbb{N}$ ist der **i-te Krylovraum zu $\mathbf{A}$ und $\mathbf{p}$** wie folgt definiert*

$$\mathcal{K}_i(\mathbf{p}, \mathbf{A}) = \mathrm{Span}\big\{\mathbf{p}, \mathbf{A}\,\mathbf{p}, \mathbf{A}^2\,\mathbf{p}, \ldots, \mathbf{A}^{i-1}\,\mathbf{p}\big\}.$$

Dabei ist $\mathbf{A}^k$ mit $k \in \mathbb{N}$ als $\mathbf{A}^k = \underbrace{\mathbf{A}\,\mathbf{A}\cdots\mathbf{A}}_{k\text{-mal}}$ zu interpretieren.

(Für $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_m \in \mathbb{R}^n$ bezeichnet $\mathrm{Span}\{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_m\}$ die Menge aller Linearkombinationen der Vektoren $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_m$. $\mathrm{Span}\{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_m\}$ wird auch die lineare Hülle von $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_m$ genannt.)

Der i-te Krylovraum $\mathcal{K}_i(\mathbf{p}, \mathbf{A})$ ist per Definition ein **Untervektorraum von** $\mathbb{R}^n$, und es gilt $\dim\big(\mathcal{K}_i(\mathbf{p}, \mathbf{A})\big) \leq i$, da der Untervektorraum $\mathcal{K}_i(\mathbf{p}, \mathbf{A})$ von den i Vektoren $\mathbf{p}, \mathbf{A}\,\mathbf{p}, \mathbf{A}^2\,\mathbf{p}, \ldots, \mathbf{A}^{i-1}\,\mathbf{p}$ aufgespannt wird. Diese müssen aber nicht linear unabhängig sein: Ist $\mathbf{p}$ zum Beispiel ein Eigenvektor zu einem Eigenwert λ von $\mathbf{A}$, dann folgt für alle $i \in \mathbb{N}$

$$\mathcal{K}_i(\mathbf{p}, \mathbf{A}) = \mathrm{Span}\big\{\mathbf{p}, \lambda\,\mathbf{p}, \lambda^2\,\mathbf{p}, \ldots, \lambda^{i-1}\,\mathbf{p}\big\} = \mathrm{Span}\{\mathbf{p}\} = \mathcal{K}_1(\mathbf{p}, \mathbf{A}),$$

und $\mathcal{K}_i(\mathbf{p}, \mathbf{A})$ ist für jedes $i \in \mathbb{N}$ nur eindimensional.

Betrachten wir aber eine positiv definite, symmetrische Matrix $\mathbf{A}$ und die im Verlauf des CG-Verfahrens generierten Richtungsvektoren $\mathbf{p}_1, \mathbf{p}_2, \ldots, \mathbf{p}_m$, so können wir sehr viel stärkere Aussagen über die Krylovräume zu $\mathbf{A}$ und dem ersten Richtungsvektor $\mathbf{p}_1$ treffen. Dieses lernen wir im nachfolgenden Hilfssatz.

Hilfssatz 3.28. (Krylovräume des CG-Verfahrens)
*Seien $\mathbf{A} \in \mathbb{R}^{n\times n}$ eine **positiv definite, symmetrische Matrix** und $\mathbf{b} \in \mathbb{R}^n$. Seien $\mathbf{p}_1, \mathbf{p}_2, \ldots, \mathbf{p}_m$ und $\mathbf{r}_1, \mathbf{r}_2, \ldots, \mathbf{r}_m$ die im Verlaufe des CG-Verfahrens zur Lösung von $\mathbf{A}\,\mathbf{x} = \mathbf{b}$ berechneten **Richtungsvektoren** und **Residuen**, wobei das CG-Verfahren nach $m \leq n$ Schritten mit $\mathbf{r}_{m+1} = \mathbf{0}$ (und vorher $\mathbf{r}_m \neq \mathbf{0}$) stoppe. Dann gilt für die **Krylovräume zu $\mathbf{A}$ und $\mathbf{p}_1$***

$$\mathcal{K}_i(\mathbf{p}_1, \mathbf{A}) = \mathrm{Span}\big\{\mathbf{p}_1, \mathbf{p}_2, \ldots, \mathbf{p}_i\big\} = \mathrm{Span}\big\{\mathbf{r}_1, \mathbf{r}_2, \ldots, \mathbf{r}_i\big\} \tag{3.59}$$

für $i = 1, 2, \ldots, m$. Insbesondere folgt daraus $\dim\big(\mathcal{K}_i(\mathbf{p}_1, \mathbf{A})\big) = i$ für jedes $i = 1, 2, \ldots, m$, da die Vektoren $\mathbf{p}_1, \mathbf{p}_2, \ldots, \mathbf{p}_m$ per Konstruktion $\mathbf{A}$-konjugiert und somit linear unabhängig sind.

Warum ist Hilfssatz 3.28 interessant? Um dieses zu sehen, müssen wir die Iterierten $\mathbf{x}_i$ des CG-Verfahrens genauer untersuchen: Die Lösung $\widehat{\mathbf{x}}$ des linearen Gleichungssystems $\mathbf{A}\,\mathbf{x} = \mathbf{b}$ und die Iterierten $\mathbf{x}_i$ des CG-Verfahrens können jeweils wie folgt geschrieben werden:

$$\widehat{\mathbf{x}} = \mathbf{x}_1 + \sum_{j=1}^{m} \alpha_j\,\mathbf{p}_j \qquad \text{bzw.} \qquad \mathbf{x}_i = \mathbf{x}_1 + \sum_{j=1}^{i-1} \alpha_j\,\mathbf{p}_j, \quad i = 1, 2, \ldots, m+1. \tag{3.60}$$

Dieses folgt, wenn man die Beziehungen $\mathbf{x}_{j+1} = \mathbf{x}_j + \alpha_j\,\mathbf{p}_j$, $j = 1, 2, \ldots, m$, nacheinander ineinander einsetzt. Nach Hilfssatz 3.28 ist es möglich, ein beliebiges $\mathbf{x}$ aus dem affin linearen Krylovraum

$$\mathbf{x}_1 + \mathcal{K}_{i-1}(\mathbf{p}_1, \mathbf{A}) := \big\{\mathbf{x}_1 + \mathbf{z} \;:\; \mathbf{z} \in \mathcal{K}_{i-1}(\mathbf{p}, \mathbf{A})\big\} \tag{3.61}$$

wie folgt darzustellen:

$$\mathbf{x} = \mathbf{x}_1 + \sum_{j=1}^{i-1} \beta_j\,\mathbf{p}_j. \tag{3.62}$$

Insbesondere gilt mit (3.62) wegen (3.60) die Beziehung $\mathbf{x}_i \in \mathbf{x}_1 + \mathcal{K}_{i-1}(\mathbf{p}_1, \mathbf{A})$ für $i = 1, 2, \ldots, m+1$. Aus (3.60) folgt

$$\widehat{\mathbf{x}} - \mathbf{x}_i = \left(\mathbf{x}_1 + \sum_{j=1}^{m} \alpha_j\,\mathbf{p}_j\right) - \left(\mathbf{x}_1 + \sum_{j=1}^{i-1} \alpha_j\,\mathbf{p}_j\right) = \sum_{j=i}^{m} \alpha_j.$$

Da die Richtungsvektoren $\mathbf{p}_1, \mathbf{p}_2, \ldots, \mathbf{p}_m$ $\mathbf{A}$-konjugiert sind, folgt in der $\mathbf{A}$-Norm

$$\begin{aligned}
\left\|\widehat{\mathbf{x}} - \mathbf{x}_i\right\|_{\mathbf{A}}^2 &= \left\|\sum_{j=i}^{m} \alpha_j \, \mathbf{p}_j\right\|_{\mathbf{A}}^2 = \left(\sum_{j=i}^{m} \alpha_j \, \mathbf{p}_j\right)^T \mathbf{A} \left(\sum_{k=i}^{m} \alpha_k \, \mathbf{p}_k\right) \\
&= \sum_{j=i}^{m} \sum_{k=i}^{m} \alpha_j \, \alpha_k \, \mathbf{p}_j^T \mathbf{A} \, \mathbf{p}_k \leq \sum_{j=i}^{m} \sum_{k=i}^{m} \alpha_j \, \alpha_k \, \mathbf{p}_j^T \mathbf{A} \, \mathbf{p}_k + \sum_{k=1}^{i-1} |\alpha_k - \beta_k|^2 \underbrace{\mathbf{p}_k^T \mathbf{A} \, \mathbf{p}_k}_{\geq 0} \\
&= \sum_{j=i}^{m} \sum_{k=i}^{m} \alpha_j \, \alpha_k \, \mathbf{p}_j^T \mathbf{A} \, \mathbf{p}_k + \sum_{j=1}^{i-1} \sum_{k=1}^{i-1} (\alpha_j - \beta_j) \, (\alpha_k - \beta_k) \underbrace{\mathbf{p}_j^T \mathbf{A} \, \mathbf{p}_k}_{=0 \text{ für } j \neq k} \\
&\quad + 2 \sum_{k=1}^{i-1} \sum_{j=i}^{m} (\alpha_k - \beta_k) \, \alpha_j \underbrace{\mathbf{p}_j^T \mathbf{A} \, \mathbf{p}_k}_{=0 \text{ weil } j \neq k} \\
&= \left(\sum_{j=i}^{m} \alpha_j \, \mathbf{p}_j + \sum_{j=1}^{i-1} (\alpha_j - \beta_j) \, \mathbf{p}_j\right)^T \mathbf{A} \left(\sum_{k=i}^{m} \alpha_k \, \mathbf{p}_k + \sum_{k=1}^{i-1} (\alpha_k - \beta_k) \, \mathbf{p}_k\right) \\
&= \left\|\sum_{j=i}^{m} \alpha_j \, \mathbf{p}_j + \sum_{j=1}^{i-1} (\alpha_j - \beta_j) \, \mathbf{p}_j\right\|_{\mathbf{A}}^2 = \left\|\sum_{j=1}^{m} \alpha_j \, \mathbf{p}_j - \sum_{j=1}^{i-1} \beta_j \, \mathbf{p}_j\right\|_{\mathbf{A}}^2 \\
&= \left\| \underbrace{\mathbf{x}_1 + \sum_{j=1}^{m} \alpha_j \, \mathbf{p}_j}_{=\widehat{\mathbf{x}}} - \underbrace{\left(\mathbf{x}_1 + \sum_{j=1}^{i-1} \beta_j \, \mathbf{p}_j\right)}_{=\mathbf{x}} \right\|_{\mathbf{A}}^2 = \left\|\widehat{\mathbf{x}} - \mathbf{x}\right\|_{\mathbf{A}}^2,
\end{aligned}$$

wobei $\mathbf{x}$ der beliebige durch (3.62) gegebene Vektor in $\mathbf{x}_1 + \mathcal{K}_{i-1}(\mathbf{p}_1, \mathbf{A})$ ist. Da die obige Rechnung für jedes beliebige $\mathbf{x} \in \mathbf{x}_1 + \mathcal{K}_{i-1}(\mathbf{p}_1, \mathbf{A})$ gilt, folgt

$$\left\|\widehat{\mathbf{x}} - \mathbf{x}_i\right\|_{\mathbf{A}}^2 \leq \left\|\widehat{\mathbf{x}} - \mathbf{x}\right\|_{\mathbf{A}}^2 \qquad \text{for all } \mathbf{x} \in \mathbf{x}_1 + \mathcal{K}_{i-1}(\mathbf{p}_1, \mathbf{A}). \tag{3.63}$$

Die Ungleichung (3.63) bedeutet, dass unter allen Vektoren im affin linearen Unterraum $\mathbf{x}_1 + \mathcal{K}_{i-1}(\mathbf{p}_1, \mathbf{A})$ die **Iterierte $\mathbf{x}_i$ des CG-Verfahrens in der A-Norm den kleinsten Abstand von der Lösung $\widehat{\mathbf{x}}$ von $\mathbf{A}\,\mathbf{x} = \mathbf{b}$ hat**. Wir halten dieses als Satz fest.

Satz 3.29. (Bestapproximationen in den affinen Krylovräumen)
*Seien $\mathbf{A} \in \mathbb{R}^{n \times n}$ eine **positiv definite, symmetrische Matrix** und $\mathbf{b} \in \mathbb{R}^n$. Das CG-Verfahren zur Lösung von $\mathbf{A}\,\mathbf{x} = \mathbf{b}$ mit dem Startvektor $\mathbf{x}_1$ stoppe nach $m \leq n$ Schritten. Die Iterierte $\mathbf{x}_i$, $i \in \{1, 2, \ldots, m+1\}$, des **CG-Verfahrens** ist die **Bestapproximation** der Lösung $\widehat{\mathbf{x}}$ von $\mathbf{A}\,\mathbf{x} = \mathbf{b}$ **in dem***

affinen Krylovraum $\mathbf{x}_1 + \mathcal{K}_{i-1}(\mathbf{p}_1, \mathbf{A})$*, definiert durch* (3.61)*, bzgl. der* $\mathbf{A}$*-Norm* $\|\cdot\|_{\mathbf{A}}$*. Dieses bedeutet, dass gilt*

$$\|\widehat{\mathbf{x}} - \mathbf{x}_i\|_{\mathbf{A}} \leq \|\widehat{\mathbf{x}} - \mathbf{x}\|_{\mathbf{A}} \qquad \text{für alle } \mathbf{x} \in \mathbf{x}_1 + \mathcal{K}_{i-1}(\mathbf{p}_1, \mathbf{A}). \tag{3.64}$$

Wir bemerken, dass wir (3.64) äquivalent wie folgt schreiben können:

$$\boxed{\|\widehat{\mathbf{x}} - \mathbf{x}_i\|_{\mathbf{A}} = \min_{\mathbf{x} \in \mathbf{x}_1 + \mathcal{K}_{i-1}(\mathbf{p}_1, A)} \|\widehat{\mathbf{x}} - \mathbf{x}\|_{\mathbf{A}}} \tag{3.65}$$

Da die affinen Krylovräume $\mathbf{x}_1 + \mathcal{K}_{i-1}(\mathbf{p}_1, \mathbf{A})$ mit wachsendem i mehr Vektoren enthalten (wegen $\dim\big(\mathcal{K}_{i-1}(\mathbf{p}_1, \mathbf{A})\big) = i - 1$) ist es absolut plausibel, dass aus (3.65) folgt, dass der absolute Fehler (in der $\mathbf{A}$-Norm) der Bestapproximation $\mathbf{x}_i$ in $\mathbf{x}_1 + \mathcal{K}_{i-1}(\mathbf{p}_1, \mathbf{A})$ in der Regel kleiner sein sollte als der absolute Fehler (in der $\mathbf{A}$-Norm) der Bestapproximation $\mathbf{x}_{i-1}$ in dem kleineren affinen Krylovraum $\mathbf{x}_1 + \mathcal{K}_{i-2}(\mathbf{p}_1, \mathbf{A})$. Die Folge der absoluten Fehler $\|\widehat{\mathbf{x}} - \mathbf{x}_i\|_{\mathbf{A}}$ in der $\mathbf{A}$-Norm sollte also mit wachsendem i streng monoton fallen. Dieses ist in der Tat der Fall:

Hilfssatz 3.30. (Fehler der Iterierten des CG-Verfahrens nimmt ab)
Seien $\mathbf{A} \in \mathbb{R}^{n \times n}$ *eine* ***positiv definite, symmetrische Matrix*** *und* $\mathbf{b} \in \mathbb{R}^n$*. Das CG-Verfahren zur Lösung von* $\mathbf{A}\,\mathbf{x} = \mathbf{b}$ *mit dem Startvektor* $\mathbf{x}_1$ *stoppe nach* $m \leq n$ *Schritten. Für die* ***Iterierten*** $\mathbf{x}_i$*,* $i \in \{1, 2, \ldots, m+1\}$*,* ***des CG-Verfahrens*** *gilt dann*

$$\|\widehat{\mathbf{x}} - \mathbf{x}_{i+1}\|_{\mathbf{A}} < \|\widehat{\mathbf{x}} - \mathbf{x}_i\|_{\mathbf{A}} \qquad \text{für alle } i = 1, 2, \ldots, m. \tag{3.66}$$

Weiter kann man den folgenden nicht-trivialen Satz beweisen (vgl. [7, Teilkapitel 6.3] für die Herleitung):

Satz 3.31. (Fehlerabschätzung für das CG-Verfahren)
Sei $\mathbf{A} \in \mathbb{R}^{n \times n}$ *eine* ***positiv definite, symmetrische Matrix****, und seien*

$$\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_n > 0$$

die n *nicht notwendigerweise verschiedenen positiven, reellen Eigenwerte von* $\mathbf{A}$*. Sei* $\mathbf{b} \in \mathbb{R}^n$*, und sei* $\widehat{\mathbf{x}}$ *die eindeutig bestimmte Lösung von* $\mathbf{A}\,\mathbf{x} = \mathbf{b}$*. Das* ***CG-Verfahren*** *zur Lösung von* $\mathbf{A}\,\mathbf{x} = \mathbf{b}$ *mit dem Startvektor* $\mathbf{x}_1$ *stoppe*

nach $m \leq n$ Schritten. Dann erfüllen die Iterierten $\mathbf{x}_i$, $i \in \{1, 2, \ldots, m+1\}$, des CG-Verfahrens die folgende ***(a priori) Fehlerabschätzung***

$$\|\widehat{\mathbf{x}} - \mathbf{x}_i\|_{\mathbf{A}} \leq 2 \, \|\widehat{\mathbf{x}} - \mathbf{x}_1\|_{\mathbf{A}} \left(\frac{\sqrt{\mathrm{cond}_2(\mathbf{A})} - 1}{\sqrt{\mathrm{cond}_2(\mathbf{A})} + 1} \right)^{i-1}, \qquad (3.67)$$

wobei $\mathrm{cond}_2(\mathbf{A}) = \|\mathbf{A}\|_2 \, \|\mathbf{A}^{-1}\|_2 = \dfrac{\lambda_1}{\lambda_n}$ *die* ***Konditionszahl der Matrix $\mathbf{A}$ bzgl. der 2-Norm*** *ist.*

Wir beobachten zunächst, dass für den Quotienten in (3.67) gilt

$$0 \leq \frac{\sqrt{\mathrm{cond}_2(\mathbf{A})} - 1}{\sqrt{\mathrm{cond}_2(\mathbf{A})} + 1} < 1.$$

Ist die Matrix $\mathbf{A}$ allerdings fast singulär (also wenn einige der Eigenwerte fast null sind, d.h. $\lambda_{d+1} \approx \ldots \approx \lambda_n \approx 0$), dann wird $\mathrm{cond}_2(\mathbf{A})$ extrem groß, d.h. die Matrix ist sehr schlecht konditioniert. Für den Quotienten in (3.67) gilt dann

$$\frac{\sqrt{\mathrm{cond}_2(\mathbf{A})} - 1}{\sqrt{\mathrm{cond}_2(\mathbf{A})} + 1} \approx 1,$$

d.h. das CG-Verfahren konvergiert sehr langsam. Man sollte dann versuchen die Matrix durch eine geeignete Transformation so zu verändern, dass die Konditionszahl der neuen Matrix deutlich kleiner ist und erst danach das transformierte lineare Gleichungssystem mit dem CG-Verfahren lösen.

Zuletzt liefern wir noch den Nachweis, dass es sich bei der $\mathbf{A}$-Norm (vgl. (3.58)) wirklich um eine Norm für $\mathbb{R}^n$ handelt:

Nachweis: Da $\langle \mathbf{x}, \mathbf{y} \rangle_{\mathbf{A}} = \mathbf{x}^T \mathbf{A} \, \mathbf{y}$ für $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$ (nach Hilfssatz 3.21 (4)) ein Skalarprodukt auf $\mathbb{R}^n$ definiert, erfüllt dieses die **Cauchy-Schwarzsche Ungleichung**:

$$\left| \langle \mathbf{x}, \mathbf{y} \rangle_{\mathbf{A}} \right| \leq \sqrt{\langle \mathbf{x}, \mathbf{x} \rangle_{\mathbf{A}}} \, \sqrt{\langle \mathbf{y}, \mathbf{y} \rangle_{\mathbf{A}}} \qquad \text{für alle } \mathbf{x}, \mathbf{y} \in \mathbb{R}^n. \qquad (3.68)$$

Dieses wollen wir nutzen, um die Norm-Eigenschaften zu überprüfen.

Aus den Eigenschaften (S4) und (S5) eines Skalarprodukts folgt

$$\|\mathbf{x}\|_{\mathbf{A}} = \sqrt{\langle \mathbf{x}, \mathbf{x} \rangle_{\mathbf{A}}} \geq 0 \qquad \text{für alle } \mathbf{x} \in \mathbb{R}^n.$$

Also definiert

$$\| \cdot \|_{\mathbf{A}} : \mathbb{R}^n \to [0; \infty[\,, \qquad \|\mathbf{x}\|_{\mathbf{A}} := \sqrt{\langle \mathbf{x}, \mathbf{x} \rangle_{\mathbf{A}}},$$

eine Funktion mit der Zielmenge $[0; \infty[$.

(1) Es gelte $\|\mathbf{x}\|_\mathbf{A} = \sqrt{\langle \mathbf{x}, \mathbf{x} \rangle_\mathbf{A}} = 0 \iff \langle \mathbf{x}, \mathbf{x} \rangle_\mathbf{A} = 0$. Dann folgt aus den Eigenschaften (S4) und (S5) eines Skalarprodukts, dass $\mathbf{x} = \mathbf{0}$ ist.

(2) Für jedes $\mathbf{x} \in \mathbb{R}^n$ und jedes $\alpha \in \mathbb{R}$ gilt wegen der Eigenschaften (S2) und (S3) des Skalarprodukts

$$\begin{aligned} \|\alpha\,\mathbf{x}\|_\mathbf{A} &= \sqrt{\langle \alpha\,\mathbf{x}, \alpha\,\mathbf{x} \rangle_\mathbf{A}} = \sqrt{\alpha\,\langle \alpha\,\mathbf{x}, \mathbf{x} \rangle_\mathbf{A}} = \sqrt{\alpha^2\,\langle \mathbf{x}, \mathbf{x} \rangle_\mathbf{A}} \\ &= \sqrt{|\alpha|^2\,\langle \mathbf{x}, \mathbf{x} \rangle_\mathbf{A}} = |\alpha|\,\sqrt{\langle \mathbf{x}, \mathbf{x} \rangle_\mathbf{A}} = |\alpha|\,\|\mathbf{x}\|_\mathbf{A}. \end{aligned}$$

(3) Mit $\|\mathbf{x}\|_\mathbf{A} = \sqrt{\langle \mathbf{x}, \mathbf{x} \rangle_\mathbf{A}}$ können wir die Cauchy-Schwarzsche Ungleichung (3.68) wie folgt schreiben:

$$|\langle \mathbf{x}, \mathbf{y} \rangle_\mathbf{A}| \leq \|\mathbf{x}\|_\mathbf{A}\,\|\mathbf{y}\|_\mathbf{A} \qquad \text{for all } \mathbf{x}, \mathbf{y} \in \mathbb{R}^n. \tag{3.69}$$

Mit (3.69) und den Eigenschaften des Skalarprodukts folgt

$$\begin{aligned} \|\mathbf{x}+\mathbf{y}\|_\mathbf{A} &= \sqrt{\langle \mathbf{x}+\mathbf{y}, \mathbf{x}+\mathbf{y} \rangle_\mathbf{A}} \\ &= \sqrt{\langle \mathbf{x}+\mathbf{y}, \mathbf{x} \rangle_\mathbf{A} + \langle \mathbf{x}+\mathbf{y}, \mathbf{y} \rangle_\mathbf{A}} \\ &= \sqrt{\langle \mathbf{x}, \mathbf{x} \rangle_\mathbf{A} + \langle \mathbf{y}, \mathbf{x} \rangle_\mathbf{A} + \langle \mathbf{x}, \mathbf{y} \rangle_\mathbf{A} + \langle \mathbf{y}, \mathbf{y} \rangle_\mathbf{A}} \\ &= \sqrt{\|\mathbf{x}\|_\mathbf{A}^2 + 2\,\langle \mathbf{x}, \mathbf{y} \rangle_\mathbf{A} + \|\mathbf{y}\|_\mathbf{A}^2} \\ &\leq \sqrt{\|\mathbf{x}\|_\mathbf{A}^2 + 2\,|\langle \mathbf{x}, \mathbf{y} \rangle_\mathbf{A}| + \|\mathbf{y}\|_\mathbf{A}^2} \\ &\overset{(3.69)}{\leq} \sqrt{\|\mathbf{x}\|_\mathbf{A}^2 + 2\,\|\mathbf{x}\|_\mathbf{A}\,\|\mathbf{y}\|_\mathbf{A} + \|\mathbf{y}\|_\mathbf{A}^2} \\ &= \sqrt{\big(\|\mathbf{x}\|_\mathbf{A} + \|\mathbf{y}\|_\mathbf{A}\big)^2} \\ &= \|\mathbf{x}\|_\mathbf{A} + \|\mathbf{y}\|_\mathbf{A} \qquad \text{for all} \quad \mathbf{x}, \mathbf{y} \in \mathbb{R}^n; \end{aligned}$$

also ist die Dreiecksungleichung nachgewiesen.

Da alle drei Normeigenschaften erfüllt sind, ist $\|\cdot\|_\mathbf{A}$ eine Norm. □

KAPITEL 4

Lösung nicht-linearer Gleichungen

Eine nicht-lineare Gleichung in einer Variablen kann in der Regel als Nullstellengleichung einer reellwertigen Funktion einer Variablen geschrieben werden. Beispielsweise ist $x^2 = a$ mit $a > 0$ äquivalent zu der Nullstellengleichung $x^2 - a = 0$. Wir suchen nun also die Nullstellen der Funktion $f : \mathbb{R} \to \mathbb{R}$, $f(x) = x^2 - a$. Analog kann man ebenfalls ein System von n nicht-linearen Gleichungen mit n Unbekannten als ein Nullstellenproblem einer Funktion $\mathbf{f} : D \to \mathbb{R}^n$ mit $D \subseteq \mathbb{R}^n$ schreiben. Gesucht sind dann alle $\mathbf{x} \in D$ mit $\mathbf{f}(\mathbf{x}) = \mathbf{0}$. In diesem Kapitel interessieren wir uns dafür, **wie man solche Nullstellenprobleme numerisch löst**, d.h. Näherungswerte für die Nullstellen berechnet. Dieses geschieht mit **geeigneten Iterationsverfahren**.

Das Problem der Nullstellenberechnung tritt in Anwendungsproblemen häufig als ein Teilproblem auf, z.B bei der numerischen Lösung einer Differentialgleichung mit einem impliziten Einschrittverfahren (vgl. Kapitel 7).

Wir betrachten zunächst den eindimensionalen Fall, d.h. **wir suchen die Nullstelle(n) einer reellwertigen Funktion einer Variablen**. Ein wichtiges Hilfsmittel für die Existenz einer Nullstelle ist der Zwischenwertsatz, den Sie vermutlich aus Ihrer Mathematikvorlesung kennen:

Satz 4.1. (Zwischenwertsatz)
*Sei $f : [c; d] \to \mathbb{R}$ eine **stetige** Funktion, und seien a und b zwei beliebige Punkte in den Intervall $[c; d]$ mit der Eigenschaft $c \leq a < b \leq d$. Dann gibt es zu jedem Wert y zwischen $f(a)$ und $f(b)$ einen Punkt $z \in [a; b]$ mit $f(z) = y$.*

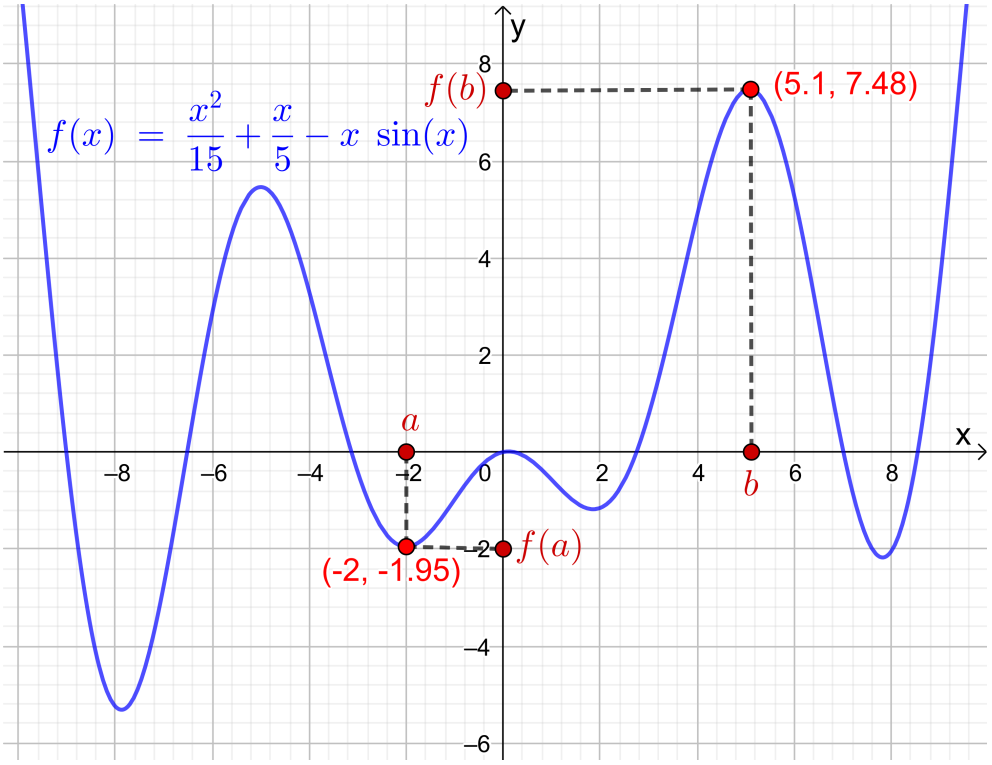


Abb. 4.1: Veranschaulichung des Zwischenwertsatzes: Da f stetig ist, werden im Intervall $[a;b]$ alle Werte zwischen $f(a)$ und $f(b)$ als Funktionswerte angenommen.

Der Zwischenwertsatz ist in Abbildung 4.1 veranschaulicht. Anschaulich bedeutet die Stetigkeit von f auf dem Intervall $[c;d]$, dass man den Graphen von f ohne Absetzen durchzeichnen kann. Da also der Graph von f die Punkte $\big(a;f(a)\big)$ und $\big(b;f(b)\big)$ mit einer durchgehenden Kurve verbindet, muss diese insbesondere auch für jeden Wert y zwischen $f(a)$ und $f(b)$ einen x-Wert z haben, für den $\big(z;f(z)\big) = (z;y)$, also $f(z) = y$, gilt.

In diesem Kapitel brauchen wir insbesondere den folgenden Sonderfall des Zwischenwertsatzes, bei dem der Funktionswert 0 als Zwischenwert auftritt:

Folgerung 4.2. (Existenz einer Nullstelle)
*Sei $f : [c;d] \to \mathbb{R}$ eine **stetige** Funktion, und seien a und b zwei beliebige Punkte in den Intervall $[c;d]$ mit den Eigenschaften $c \leq a < b \leq d$ und*

$$f(a) \leq 0 \leq f(b) \qquad \textit{oder} \qquad f(a) \geq 0 \geq f(b).$$

*Dann hat f in $[a;b]$ **mindestens eine Nullstelle**.*

4.1 Bisektionsverfahren

Wir betrachten eine Funktion f, die auf dem Intervall $[a;b]$ definiert und stetig ist. Es gelte weiter

$$f(a) \cdot f(b) < 0,$$

d.h. $f(a)$ und $f(b)$ sind beide ungleich null und haben unterschiedliche Vorzeichen. Dann nimmt f nach dem Zwischenwertsatz in dem Intervall $[a;b]$ mindestens einmal den Wert 0 an. Es gibt also ein $z \in]a;b[$ mit $f(z) = 0$. (Anders ausgedrückt, f wechselt in dem Intervall $]a;b[$ mindestens einmal sein Vorzeichen.) – Wie kann man auf einfache Weise eine Näherung für eine solche Nullstelle z berechnen?

Die Idee des **Bisektionsverfahrens** zur Nullstellenfindung ist wie folgt: Wir teilen das Intervall $[a;b]$ in zwei gleich große Teilintervalle auf und behalten das Teilintervall, in dem f garantiert eine Nullstelle hat. Die Wahl des zu behaltenden Teilintervalls erfolgt, indem wir die Funktionswerte an den beiden (neuen) Teilintervallenden multiplizieren und das Teilintervall nehmen, bei dem wir dabei einen nicht-positiven Wert für das Produkt der Funktionswerte erhalten.

Genauer sieht die **Vorgehensweise des Bisektionsverfahrens** wie folgt aus:

Schritt 1: Wir setzen $a_1 := a$ und $b_1 := b$.
(Nach Voraussetzung gilt $f(a_1) \cdot f(b_1) < 0$.)

Wir definieren $c_1 := \dfrac{a_1 + b_1}{2}$. (Dann erhalten wir mit $[a_1; c_1]$ und $[c_1; b_1]$ zwei gleich große Teilintervalle der Länge $(b-a)/2$.)

Ist $f(c_1) \cdot f(b_1) \leq 0$, so setzen wir $a_2 := c_1$ und $b_2 := b_1$.

Andernfalls (also wenn $f(c_1) \cdot f(b_1) > 0$ und damit $f(a_1) \cdot f(c_1) \leq 0$ ist) setzen wir $a_2 := a_1$ und $b_2 := c_1$.

(Wir haben also jetzt ein Teilintervall $[a_2; b_2]$ der Länge $(b-a)/2$ gefunden, in dem garantiert eine Nullstelle von f liegt.)

Schritt 2: (Nach Voraussetzung gilt $f(a_2) \cdot f(b_2) \leq 0$.)

Wir definieren $c_2 := \dfrac{a_2 + b_2}{2}$. (Dann erhalten wir mit $[a_2; c_2]$ und $[c_2; b_2]$ zwei gleich große Teilintervalle der Länge $(b-a)/4$.)

Ist $f(c_2) \cdot f(b_2) \leq 0$, so setzen wir $a_3 := c_2$ und $b_3 := b_2$.

Andernfalls (also wenn $f(c_2) \cdot f(b_2) > 0$ und damit $f(a_2) \cdot f(c_2) \leq 0$ ist) setzen wir $a_3 := a_2$ und $b_3 := c_2$.

(Wir haben also jetzt ein Teilintervall $[a_3; b_3]$ der Länge $(b-a)/4$ gefunden, in dem garantiert eine Nullstelle von f liegt.)

$\vdots$

Schritt n: (Nach Voraussetzung gilt $f(a_n) \cdot f(b_n) \leq 0$.)

Wir definieren $c_n := \dfrac{a_n + b_n}{2}$. (Dann erhalten wir mit $[a_n; c_n]$ und $[c_n; b_n]$ zwei gleich große Teilintervalle der Länge $(b-a)/2^n$.)

Ist $f(c_n) \cdot f(b_n) \leq 0$, so setzen wir $a_{n+1} := c_n$ und $b_{n+1} := b_n$.

Andernfalls (also wenn $f(c_n) \cdot f(b_n) > 0$ und damit $f(a_n) \cdot f(c_n) \leq 0$ ist) setzen wir $a_{n+1} := a_n$ und $b_{n+1} := c_n$.

(Wir haben also jetzt ein Teilintervall $[a_{n+1}; b_{n+1}]$ der Länge $(b-a)/2^n$ gefunden, in dem garantiert eine Nullstelle von f liegt.)

$\vdots$

Abbruchkriterium/Stoppkriterium: Wir wissen nach n Schritten des Bisektionsverfahrens, dass eine Nullstelle von f im Intervall $[a_{n+1}; b_{n+1}]$ der Länge $(a-b)/2^n$ liegt. Also haben wir nach n Schritten eine Näherung $\widetilde{z} := c_n$ der Nullstelle bestimmt, deren absoluter Fehler höchstens $(a-b)/2^n$ beträgt. – Soll also eine Näherung $\widetilde{z}$ einer Nullstelle z von f mit einem absoluten Fehler $|\widetilde{z} - z| \leq \varepsilon$ (für eine vorgegebene absolute Fehlerschranke $\varepsilon > 0$) gefunden werden, so stoppen wir das Bisektionsverfahren, sobald $(b-a)/2^n \leq \varepsilon$ ist, denn dann gilt nach Konstruktion für die Näherung $\widetilde{z} = c_n$

$$|\widetilde{z} - z| = |c_n - z| \leq \frac{b-a}{2^n} \leq \varepsilon,$$

weil z und $\widetilde{z} = c_n$ beide im Intervall $[a_{n+1}; b_{n+1}]$ der Länge $(b-a)/2^n$ liegen.

Ist eine absolute Fehlerschranke $\varepsilon > 0$ für die Näherung $\widetilde{z} = c_n$ der Nullstelle z vorgegeben ist, so können wir durch Auflösen von

$$|\widetilde{z} - z| = |c_n - z| \leq \frac{b-a}{2^n} \leq \varepsilon$$

nach n berechnen, nach wie vielen Iterationsschritten die gewünschte absolute Fehlerschranke ε mit Sicherheit erreicht wird:

$$\frac{b-a}{2^n} \leq \varepsilon \quad \Big| \cdot 2^n \quad \Longleftrightarrow \quad b - a \leq \varepsilon \cdot 2^n \quad \Big| : \varepsilon \quad \Longleftrightarrow \quad \frac{b-a}{\varepsilon} \leq 2^n = e^{\ln(2) \cdot n}$$

$$\overset{\ln'(x)>0}{\Longleftrightarrow} \quad \ln\left(\frac{b-a}{\varepsilon}\right) \leq \ln(2) \cdot n \quad \Big| : \ln(2) \quad \overset{\ln(2)>0}{\Longleftrightarrow} \quad \frac{\ln\left(\dfrac{b-a}{\varepsilon}\right)}{\ln(2)} \leq n,$$

wobei wir bei der Umformung von der ersten in die zweite Zeile genutzt haben, dass der natürliche Logarithmus ln streng monoton wachsend ist und das Anwenden von ln daher eine Äquivalenzumformung ist und die Richtung der Ungleichung (wegen des streng monotonen Wachstums von ln) erhalten bleibt.

$$\text{Für jedes } n \geq \frac{\ln\left(\frac{b-a}{\varepsilon}\right)}{\ln(2)} \text{ wird die absolute Fehlerschranke } \varepsilon \text{ erreicht.} \tag{4.1}$$

Wir halten das Bisektionsverfahren nun als Algorithmus fest. Dabei nutzen wir die **Signum-Funktion** (oder **Vorzeichenfunktion**)

$$\operatorname{sgn} : \mathbb{R} \to \mathbb{R}, \qquad \operatorname{sgn}(x) := \begin{cases} -1 & \text{für } x < 0, \\ 0 & \text{für } x = 0, \\ 1 & \text{für } x > 0, \end{cases}$$

um zu vermeiden, dass sehr kleine Werte für $f(c_n) \cdot f(b_n)$ vom Computer als 0 interpretiert werden.

Verfahren 4.3. (Bisektionsverfahren)
*Voraussetzungen: Die Funktion f sei **stetig** auf dem Intervall $[a;b]$ und es gelte $f(a) \cdot f(b) < 0$. (Dann hat f mindestens eine Nullstelle z in $]a;b[$.)*

Sei $\varepsilon > 0$ die gewünschte absolute Fehlerschranke für die Näherung von z.

Initialisierung: *Seien $a_1 := a$ und $b_1 := b$.*

Algorithmus: *Für $n = 1, 2, 3, \ldots$ führe folgenden Prozess durch*

(1) Definiere $c_n := \dfrac{a_n + b_n}{2}$.

(2) Falls $\operatorname{sgn}\big(f(c_n)\big) \cdot \operatorname{sgn}\big(f(b_n)\big) \leq 0$ *ist, setze $a_{n+1} := c_n$ und $b_{n+1} := b_n$. Andernfalls setze $a_{n+1} := a_n$ und $b_{n+1} := c_n$.*

bis $b_{n+1} - a_{n+1} \leq \varepsilon$ gilt.

Dann ist $\widetilde{z} := c_n$ aus dem letzten Schritt des Algorithmus die Näherung mit der absoluten Fehlerschranke ε für die unbekannte Nullstelle z, also $|\widetilde{z} - z| \leq \varepsilon$.

Betrachten wir ein Beispiel.

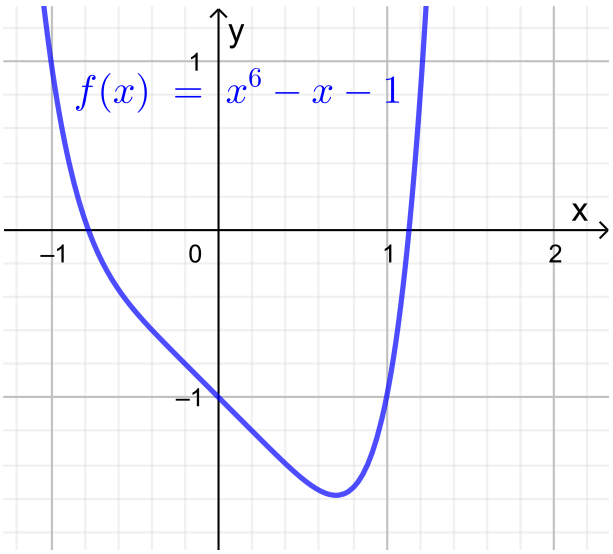


Abb. 4.2: Graph der Funktion $f : \mathbb{R} \to \mathbb{R}$, $f(x) = x^6 - x - 1$.

Beispiel 4.4. (Bisektionsverfahren)

Gesucht ist eine Näherung der größten reellen Nullstelle der Funktion

$$f : \mathbb{R} \to \mathbb{R}, \qquad f(x) = x^6 - x - 1 = x\,(x^5 - 1) - 1,$$

mit der garantierten absoluten Fehlerschranke $\varepsilon = 0{,}001 = 10^{-3}$.

Der Graph der Funktion f ist in Abbildung 4.2 gezeichnet. Als Polynom ist f beliebig oft differenzierbar und insbesondere stetig. Es gilt

$$f(2) = 2\,(2^5 - 1) - 1 = 61 > 0,$$
$$f(1) = 1\,(1^5 - 1) - 1 = -1 < 0.$$

Wegen $f(1) = -1 < 0 < 61 = f(2)$ liegt nach dem Zwischenwertsatz eine Nullstelle von f in dem Intervall $[a; b] = [1; 2]$.

Warum liegt im Intervall $[1; 2]$ nur die größte reelle Nullstelle von f und keine weitere Nullstelle von f? Um dieses nachzuweisen, berechnen wir die Ableitung

$$f'(x) = 6\,x^5 - 1.$$

Für die Ableitung gilt

$$f'(x) = 6\,x^5 - 1 \geq 6 \cdot 1^5 - 1 = 5 > 0 \qquad \text{für alle } x \geq 1.$$

Also ist f auf den Intervall $[1; \infty[$ streng monoton wachsend. Daraus folgt, dass f in $[1; \infty[$ höchstens einmal die x-Achse schneidet und damit höchstens eine Nullstelle in $[1; \infty[$ hat. Daher enthält das Intervall $[1; 2]$ nur genau eine Nullstelle

n	a_n	b_n	c_n	$b_n - c_n$	$f(c_n)$	$f(b_n)$	$f(b_n) \cdot f(c_n)$
1	1,00000	2,00000	1,50000	0,50000	$8{,}89 \cdot 10^{0}$	$6{,}10 \cdot 10^{1}$	$5{,}42 \cdot 10^{2}$
2	1,00000	1,50000	1,25000	0,25000	$1{,}56 \cdot 10^{0}$	$8{,}89 \cdot 10^{0}$	$1{,}39 \cdot 10^{1}$
3	1,00000	1,25000	1,12500	0,12500	$-9{,}77 \cdot 10^{-2}$	$1{,}56 \cdot 10^{0}$	$-1{,}53 \cdot 10^{-1}$
4	1,12500	1,25000	1,18750	0,06250	$6{,}17 \cdot 10^{-1}$	$1{,}56 \cdot 10^{0}$	$9{,}65 \cdot 10^{-1}$
5	1,12500	1,18750	1,15625	0,03125	$2{,}33 \cdot 10^{-1}$	$6{,}17 \cdot 10^{-1}$	$1{,}44 \cdot 10^{-1}$
6	1,12500	1,15625	1,14063	0,01563	$6{,}16 \cdot 10^{-2}$	$2{,}33 \cdot 10^{-1}$	$1{,}44 \cdot 10^{-2}$
7	1,12500	1,14063	1,13281	0,00781	$-1{,}96 \cdot 10^{-2}$	$6{,}16 \cdot 10^{-2}$	$-1{,}21 \cdot 10^{-3}$
8	1,13281	1,14063	1,13672	0,00391	$2{,}06 \cdot 10^{-2}$	$6{,}16 \cdot 10^{-2}$	$1{,}27 \cdot 10^{-3}$
9	1,13281	1,13672	1,13477	0,00195	$4{,}27 \cdot 10^{-4}$	$2{,}06 \cdot 10^{-2}$	$8{,}80 \cdot 10^{-6}$
10	1,13281	1,13477	1,13379	0,00098	$-9{,}60 \cdot 10^{-3}$	$4{,}27 \cdot 10^{-4}$	$-4{,}10 \cdot 10^{-6}$

Tabelle 4.1: Bisektionsverfahren zur Berechnung der größten Nullstelle von $f(x) = x^6 - x - 1$ mit den Startwerten $a_1 = 1$ und $b_1 = 2$.

von f, und diese ist die einzige Nullstelle von f in $[1; \infty[$ und damit die größte Nullstelle von f.

Wir führen daher das Bisektionsverfahren mit den Startwerten $a_1 = a = 1$ und $b_1 = b = 2$ durch. Die Werte für a_n, b_n, c_n und $b_n - c_n$, sowie $f(c_n)$, $f(b_n)$ und $f(b_n) \cdot f(c_n)$ für $n = 1, 2, \ldots, 10$, sind in Tabelle 4.1 angegeben. Dabei wurden a_n, b_n, c_n auf eine Gleitkommadarstellung mit einer 6-stelligen Mantisse gerundet angegeben. Die Werte von $f(c_n)$ und $f(b_n)$ bzw. $f(b_n) \cdot f(c_n)$ wurden dabei nur auf eine Gleitkommadarstellung mit 3-stelliger Mantisse gerundet angegeben, da hier nur die Größenordnung bzw. das Vorzeichen interessant sind.

Wir sehen, dass im zehnten (10.) Iterationsschritt gilt

$$|z - c_{10}| \leq b_{11} - a_{11} = b_{10} - c_{10} = 0{,}00098 \leq 0{,}001.$$

Also wird die gewünschte absolute Fehlerschranke $\varepsilon = 0{,}001 = 10^{-3}$ nach 10 Iterationsschritten erreicht. Wir erhalten als Näherung für die (größte reelle) Nullstelle von f nach 10 Iterationsschritten $\widetilde{z} = c_{10} = 1{,}13379$. (Zum Vergleich: Der auf eine Gleitkommadarstellung mit 10-stelliger Mantisse gerundete Wert der Nullstelle ist $z \doteq 1{,}134724138$.)

Nach (4.1) wird die absolute Fehlerschranke $\varepsilon = 0{,}001 = 10^{-3}$ spätestens nach

$$n \geq \frac{\ln\left(\frac{2-1}{0{,}001}\right)}{\ln(2)} = \frac{\ln(10^3)}{\ln(2)} = 9{,}97,$$

also nach spätestens 10 Iterationsschritten, garantiert erreicht. Die theoretischen Überlegungen bestätigen also, was wir bereits an Tabelle 4.1 gesehen hatten. ♠

4.2 Newton-Verfahren

Das Newton-Verfahren berechnet eine Folge von Annäherungen an eine Nullstelle einer stetig differenzierbaren Funktion mit Hilfe der Tangenten an den Graphen der Funktion. Das **Newton-Verfahren** ist in Abbildung 4.3 illustriert und funktioniert im Detail wie folgt:

Sei $f :]a; b[\to \mathbb{R}$ eine stetig differenzierbare Funktion, die in einen Punkt $z \in]a; b[$ eine Nullstelle hat, also $f(z) = 0$. Als Startwert für das Newton-Verfahren brauchen wir eine hinreichend gute Näherung x_0 für die Nullstelle z. Diese Näherung x_0 kann zum Beispiel mit Hilfe eines Plots des Graphen von f bestimmt werden.

Wir nehmen nun die Tangente in $\big(x_0; f(x_0)\big)$ an den Graphen von f und bestimmen deren Schnittpunkt mit der x-Achse. Da die Tangente in $\big(x_0; f(x_0)\big)$ an den Graphen dicht bei x_0 eine gute Näherung der Funktion f ist, erwarten wir, dass der Schnittpunkt x_1 dieser Tangente mit der x-Achse eine verbesserte Näherung der Nullstelle z ist.

Die **Tangente in $\big(x_0; f(x_0)\big)$ an den Graphen von** f ist durch das lineare Taylor-Polynom p_1 von f mit dem Entwicklungspunkt x_0 gegeben:

$$p_1(x) = f(x_0) + f'(x_0)\,(x - x_0).$$

Für die neue (verbesserte) Näherung x_1 der Nullstelle z gilt $p_1(x_1) = 0$, da x_1 als Schnittpunkt des linearen Taylor-Polynoms p_1 mit der x-Achse gegeben ist. Wir suchen also x_1 mit

$$p_1(x_1) = f(x_0) + f'(x_0)\,(x_1 - x_0) = 0, \tag{4.2}$$

wobei wir benötigen, dass die Ableitung $f'(x_0) \neq 0$ ist. (Falls $f'(x_0) = 0$ gilt, ist die Tangente parallel zur x-Achse und schneidet diese nicht oder ist identisch mit der x-Achse, falls $f(x_0) = 0$. In letzterem Fall ist der Startwert x_0 aber bereits eine Nullstelle von f.) Auflösen von (4.2) nach x_1 liefert:

$$\begin{aligned} f(x_0) + f'(x_0)\,(x_1 - x_0) = 0 \quad &\Longleftrightarrow \quad f'(x_0)\,(x_1 - x_0) = -f(x_0) \\ \overset{f'(x_0)\neq 0}{\Longleftrightarrow} \quad x_1 - x_0 = -\frac{f(x_0)}{f'(x_0)} \quad &\Longleftrightarrow \quad x_1 = x_0 - \frac{f(x_0)}{f'(x_0)} \end{aligned}$$

Die neue Näherung für die Nullstelle ist also gegeben durch

$$x_1 = x_0 - \frac{f(x_0)}{f'(x_0)}\,. \tag{4.3}$$

Wir können diese Vorgehensweise wiederholen, wobei wir nun eine verbesserte Näherung von x_1 für die Nullstelle z berechnen wollen. Dazu ersetzen wir in (4.3)

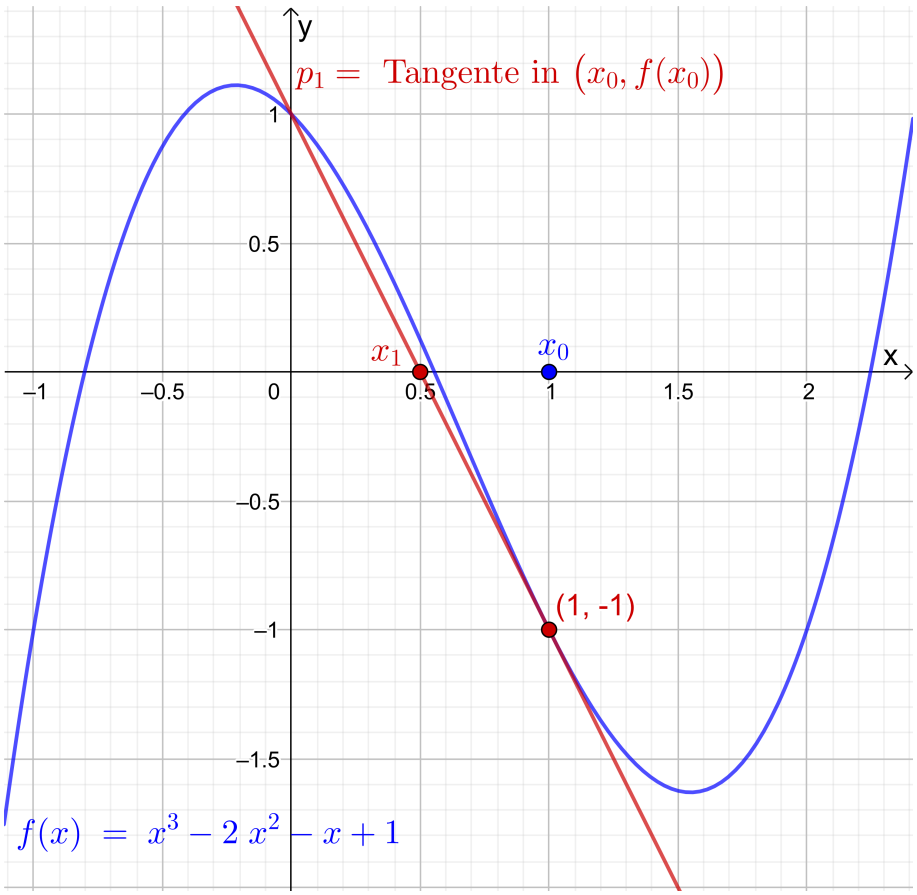


Abb. 4.3: Illustration des Newton-Verfahrens: Für den Startwert x_0 legen wir die Tangente p_1 an den Graphen im Punkt $\big(x_0; f(x_0)\big)$. Die neue Näherung x_1 für die Nullstelle ist der Schnittpunkt der Tangente mit der x-Achse.

x_1 durch die neue Näherung x_2 und den Startwert x_0 durch x_1. Somit erhalten wir aus (4.3) als Formel für die neue Näherung:

$$x_2 = x_1 - \frac{f(x_1)}{f'(x_1)}$$

Setzt man dieses Verfahren fort, so erhält man nach $n+1$ Schritten:

$$\boxed{x_{n+1} = x_n - \frac{f(x_n)}{f'(x_n)}}$$

Wir halten die Vorgehensweise des Newton-Verfahrens als Verfahren fest.

Verfahren 4.5. (Newton-Verfahren)
*Sei $f :]a; b[\to \mathbb{R}$ eine **stetig differenzierbare** Funktion, und sei z eine Nullstelle von f (d.h. es gilt $f(z) = 0$). Sei x_0 eine gute Näherung für z. Weiter*

*gelte $f'(x) \neq 0$ für alle x dicht bei z. (Insbesondere gilt dann $f'(x_0) \neq 0$.) Das **Newton-Verfahren** berechnet Näherungen für die Nullstelle z mit dem folgenden Iterationsverfahren:*

$$x_{n+1} = x_n - \frac{f(x_n)}{f'(x_n)}, \qquad n = 0, 1, 2, \ldots \tag{4.4}$$

Betrachten wir zunächst ein Beispiel.

Beispiel 4.6. (Newton-Verfahren)
Gesucht ist eine Näherung der größten reellen Nullstelle der Funktion

$$f : \mathbb{R} \to \mathbb{R}, \qquad f(x) = x^6 - x - 1,$$

die bereits in Beispiel 4.4 betrachtet wurde (siehe auch Abbildung 4.2). In Beispiel 4.4 hatten wir nachgewiesen, dass die größte reelle Nullstelle z im Intervall $[1; 2]$ liegt. Wir nehmen daher als Startwert für das Newton-Verfahren $x_0 = 1{,}5$.

Die Ableitung von f ist durch

$$f'(x) = 6\,x^5 - 1$$

gegeben, und somit lautet die Formel für das Newton-Verfahren:

$$x_{n+1} = x_n - \frac{f(x_n)}{f'(x_n)} = x_n - \frac{x_n^6 - x_n - 1}{6\,x_n^5 - 1}, \qquad n = 0, 1, 2, \ldots$$

In Tabelle 4.2 sind die ersten 6 Iterationsschritte (auf eine Gleitkommadarstellung mit 9-stelliger Mantisse gerundet) angegeben. Außer den Näherungen x_n der Nullstelle z wurden die Funktionswerte $f(x_n)$ angegeben. In der beiden letzten Spalten der Tabelle wurden in der Zeile für x_n den Fehler $x_{n-1} - z$ der vorherigen Näherung x_{n-1} und zusätzlich $x_{n-1} - x_n$ angegeben, wobei diese Werte (ebenso wie $f(x_n)$) auf eine Gleitkommadarstellung mit 3-stelliger Mantisse gerundet wurden, da hier nur die Größenordnung interessant ist.

Der wahre Wert der Nullstelle (mit Rundung auf eine Gleitkommadarstellung mit 10-stelliger Mantisse) ist $z \doteq 1{,}134724138$, und wir sehen, dass die Näherung $x_5 \doteq 1{,}13472415$ bereits 8 signifikante Ziffern hat.

Wir werden noch sehen, dass $x_{n-1} - x_n$ eine gute Näherung für $x_{n-1} - z$ liefert. Da bei einer unbekannten Nullstelle z der absolute Fehler $|x_n - z|$ nicht berechenbar ist, ist es wichtig einen guten Näherungswert für diesen zu haben, denn nur dann

n	x_n	$f(x_n)$	$x_{n-1} - x_n$	$x_{n-1} - z$
0	1,5	$8{,}89 \cdot 10^{1}$		
1	1,30049088	$2{,}54 \cdot 10^{1}$	$2{,}00 \cdot 10^{-1}$	$3{,}65 \cdot 10^{-1}$
2	1,18148042	$5{,}38 \cdot 10^{-1}$	$1{,}19 \cdot 10^{-1}$	$1{,}66 \cdot 10^{-1}$
3	1,13945559	$4{,}92 \cdot 10^{-2}$	$4{,}20 \cdot 10^{-2}$	$4{,}68 \cdot 10^{-2}$
4	1,13477763	$5{,}50 \cdot 10^{-4}$	$4{,}68 \cdot 10^{-3}$	$4{,}73 \cdot 10^{-3}$
5	1,13472415	$7{,}11 \cdot 10^{-8}$	$5{,}35 \cdot 10^{-5}$	$5{,}35 \cdot 10^{-5}$
6	1,13472414	$1{,}55 \cdot 10^{-15}$	$6{,}91 \cdot 10^{-9}$	$6{,}91 \cdot 10^{-9}$

Tabelle 4.2: Newton-Verfahren zur Berechnung der größten reellen Nullstelle der Funktion $f(x) = x^6 - x - 1$ mit dem Startwert $x_0 = 1{,}5$.

können wir eine Aussage darüber treffen, wann die Näherung x_n die gewünschte absolute Fehlerschranke (voraussichtlich) erreicht hat.

An der letzten Spalte lesen wir ab, dass sich der absolute Fehler des Newton-Verfahrens (in diesem Beispiel) zunächst bei $n = 1, 2, 3$ nur moderat verkleinert, wohingegen er ab $n = 4$ rasant abnimmt. ♠

Was kann man über den **absoluten Fehler** des Newton-Verfahrens aussagen? Der Satz von Taylor liefert uns für die Darstellung von $f(x_n)$ durch das konstante Taylor-Polynom mit dem Entwicklungspunkt z, dass es ein z_n zwischen x_n und z gibt, so dass gilt

$$f(x_n) = f(z) + f'(z_n)\,(x_n - z).$$

Da $f(z) = 0$ ist folgt

$$f(x_n) = f'(z_n)\,(x_n - z),$$

und Auflösen nach $x_n - z$ liefert unter der Annahme $f'(z_n) \neq 0$:

$$f(x_n) = f'(z_n)\,(x_n - z) \qquad \overset{f'(z_n)\neq 0}{\Longleftrightarrow} \qquad \frac{f(x_n)}{f'(z_n)} = x_n - z$$

Ist x_n dicht genug bei z, so folgt für z_n (welches zwischen x_n und z liegt) $z_n \approx x_n$ und somit $f'(z_n) \approx f'(x_n)$. Damit erhalten wir:

$$x_n - z = \frac{f(x_n)}{f'(z_n)} \approx \frac{f(x_n)}{f'(x_n)} = x_n - \Bigg(\underbrace{x_n - \frac{f(x_n)}{f'(x_n)}}_{= x_{n+1}} \Bigg) = x_n - x_{n+1}$$

Also gilt, falls x_n dicht genug bei z liegt für den Fehler von x_n

$$\boxed{x_n - z \approx x_n - x_{n+1}.} \tag{4.5}$$

Daraus folgt für den absoluten Fehler von x_n:

$$\boxed{|x_n - z| \approx |x_n - x_{n+1}|} \tag{4.6}$$

Da die Nullstelle z in der Regel nicht bekannt ist, kann man den absoluten Fehler $|x_n - z|$ von x_n nicht direkt berechnen. Formel (4.6) ist sehr hilfreich, um den absoluten Fehler $|x_n - z|$ der Näherung x_n angenähert zu berechnen.

Beispiel 4.7. (absoluter Fehler des Newton-Verfahrens)
Betrachten wir die Funktion

$$f : \mathbb{R} \to \mathbb{R}, \qquad f(x) = x^6 - x - 1,$$

deren größte Nullstelle in Beispiel 4.6 mit den Newton-Verfahren angenähert berechnet wurde. In den letzten beiden beiden Spalten der Tabelle in Beispiel 4.6 lesen wir ab für $n \geq 3$ ab:

$$\begin{aligned} x_3 - z &= 4{,}73 \cdot 10^{-3}, & x_3 - x_4 &= 4{,}68 \cdot 10^{-3}, \\ x_4 - z &= 5{,}35 \cdot 10^{-5}, & x_4 - x_5 &= 5{,}35 \cdot 10^{-5}, \\ x_5 - z &= 6{,}91 \cdot 10^{-9}, & x_5 - x_6 &= 6{,}91 \cdot 10^{-9}. \end{aligned}$$

Dieses zeigt, dass die Näherung $x_n - z \approx x_n - x_{n+1}$ aus (4.5) in diesem Beispiel bereits ab $n = 3$ gut und ab $n = 4$ sehr gut ist. ♠

Wir wollen nun allgemein das **Verhalten des absoluten Fehlers** des Newton-Verfahrens weiter untersuchen. Dazu setzen wir voraus, dass die Funktion $f :]a; b[\to \mathbb{R}$, deren Nullstelle z wir berechnen wollen, zweimal stetig differenzierbar ist und dass für die Ableitung in der Nullstelle z gilt

$$f'(z) \neq 0. \tag{4.7}$$

Die Bedingung (4.7) bedeutet, dass die Tangente an den Graphen von f im Punkt z nicht parallel zur x-Achse ist. Wegen der Stetigkeit der Ableitung f' folgt daraus, dass $f'(x) \neq 0$ für alle x gilt, die dicht genug bei z liegen.

Wir nutzen nun den Satz von Taylor (oder die Taylorsche Formel), um $f(z)$ durch das lineare Taylor-Polynom von f mit dem Entwicklungspunkt x_n darzustellen. Nach dem Satz von Taylor gibt es ein z_n zwischen z und x_n, so dass gilt

$$f(z) = f(x_n) + f'(x_n)\,(z - x_n) + \frac{1}{2}\,f''(z_n)\,(z - x_n)^2. \tag{4.8}$$

Da z eine Nullstelle von f ist, gilt $f(z) = 0$, und somit folgt aus (4.8)

$$0 = f(x_n) + f'(x_n)\,(z - x_n) + \frac{1}{2}\,f''(z_n)\,(z - x_n)^2. \tag{4.9}$$

Wir teilen in (4.9) durch $f'(x_n) \neq 0$ und formen weiter um:

$$\begin{aligned}
& 0 = \frac{f(x_n)}{f'(x_n)} + (z - x_n) + \frac{1}{2}\,\frac{f''(z_n)}{f'(x_n)}\,(z - x_n)^2 \\
\iff \qquad & 0 = z - \underbrace{\left(x_n - \frac{f(x_n)}{f'(x_n)}\right)}_{= x_{n+1}} + \frac{f''(z_n)}{2\,f'(x_n)}\,(z - x_n)^2 \\
\iff \qquad & 0 = z - x_{n+1} + \frac{f''(z_n)}{2\,f'(x_n)}\,(z - x_n)^2 \\
\iff \qquad & x_{n+1} - z = \frac{f''(z_n)}{2\,f'(x_n)}\,(z - x_n)^2 = \frac{f''(z_n)}{2\,f'(x_n)}\,(x_n - z)^2
\end{aligned}$$

Also gilt für den Fehler von x_{n+1}

$$x_{n+1} - z = \frac{f''(z_n)}{2\,f'(x_n)}\,(x_n - z)^2. \tag{4.10}$$

Falls x_n dicht bei z liegt so gilt in (4.10) angenähert:

$$\boxed{\frac{f''(z_n)}{2\,f'(x_n)} \approx \frac{f''(z)}{2\,f'(z)} =: M} \tag{4.11}$$

Setzt man (4.11) in (4.10) ein, so erhält man angenähert:

$$\boxed{x_{n+1} - z \approx M\,(x_n - z)^2} \tag{4.12}$$

Multiplikation von (4.12) mit M liefert:

$$\boxed{M\,(x_{n+1} - z) \approx M^2\,(x_n - z)^2 = \big[M\,(x_n - z)\big]^2} \tag{4.13}$$

Sind alle Näherungen x_n, $n = 0, 1, 2, \ldots$ dicht bei z, so dass (4.11) für alle $n = 0, 1, 2, \ldots$ gilt, so folgt durch wiederholte Anwendung von (4.13):

$$M\,(x_{n+1} - z) \approx \big[M\,(x_n - z)\big]^2 \approx \Big[\big[M\,(x_{n-1} - z)\big]^2\Big]^2 = \big[M\,(x_{n-1} - z)\big]^4$$

$$\approx \left[\left[M\,(x_{n-2}-z)\right]^2\right]^4 = \left[M\,(x_{n-2}-z)\right]^8 = \left[M\,(x_{n-2}-z)\right]^{2^3}$$
$$\approx \ldots \approx \left[M\,(x_1-x)\right]^{2^n} \approx \left[M\,(x_0-z)\right]^{2^{n+1}}$$

Also erhalten wir, falls alle x_n, $n = 0, 1, 2, \ldots$, dicht bei z liegen :

$$\boxed{M\,(x_n-z) \approx \left[M\,(x_0-z)\right]^{2^n}, \qquad n = 0, 1, 2, \ldots} \tag{4.14}$$

An Formel (4.14) sehen wir, dass der absolute Fehler $|x_n - z|$ von x_n gegen 0 strebt, wenn gilt

$$\left|M\,(x_0-z)\right| < 1 \qquad \Longleftrightarrow \qquad \boxed{|x_0-z| < \frac{1}{|M|} = \left|\frac{2\,f'(z)}{f''(z)}\right|.} \tag{4.15}$$

Wir sehen an (4.15), dass der Startwert x_0 sehr dicht bei z liegen muss, wenn $|M|$ sehr groß ist. Wird der Startwert x_0 also nicht hinreichend dicht bei z gewählt, so dass (4.15) nicht erfüllt ist, so wird das Newton-Verfahren normalerweise nicht gegen die Nullstelle z konvergieren.

Wie man einen guten Startwert x_0 wählt, hängt vom konkreten Beispiel ab: Dieses kann durch Zeichnen des Graphen der Funktion passieren, oder der Startwert ist bei praktischen Anwendungen durch physikalische Überlegungen zu der Problemstellung gegeben. Liegt kein guter Startwert vor, so kann es sinnvoll sein, zunächst ein paar Schritte mit dem Bisektionsverfahren durchzuführen, um einen besseren Startwert zu erhalten.

Wir illustrieren die vorherigen theoretischen Überlegungen an einem Beispiel.

Beispiel 4.8. (absoluter Fehler des Newton-Verfahrens)
Betrachten wir die Funktion $f : \mathbb{R} \to \mathbb{R},\ f(x) = x^6 - x - 1,$ deren größte reelle Nullstelle in Beispiel 4.6 mit den Newton-Verfahren angenähert berechnet wurde. Die erste und zweite Ableitung dieser Funktion sind

$$f'(x) = 6\,x^5 - 1, \qquad f''(x) = 30\,x^4.$$

Mit der Nullstelle $z \doteq 1{,}134724138$ erhalten wir in (4.11) (mit Rundung auf eine Gleitkommadarstellung mit einer 3-stelligen Mantisse)

$$M = \frac{f''(z)}{2\,f'(z)} = \frac{30\,z^4}{2\,(6\,z^5-1)} \doteq 2{,}42.$$

Liegen die x_n dicht genug bei z, so gilt nach (4.12)

$$x_{n+1} - z \approx M\,(x_n-z)^2 = 2{,}42\,(x_n-z)^2. \tag{4.16}$$

Betrachten wir beispielsweise $n = 3$ mit $x_3 - z = 4{,}73 \cdot 10^{-3}$ (vgl. Tabelle 4.2). Nach (4.16) sollte gelten (mit Rundung auf eine Gleitkommadarstellung mit 3-stelliger Mantisse)

$$x_4 - z \approx 2{,}42\,(x_3 - z)^2 \doteq 2{,}42 \cdot \left(4{,}73 \cdot 10^{-3}\right)^2 \doteq 5{,}41 \cdot 10^{-5}.$$

In Tabelle 4.2 finden wir $x_4 - z = 5{,}35 \cdot 10^{-5}$. Bereits für $n = 4$ liefert (4.16) also eine gute Vorhersage für den absoluten Fehler.

Untersuchen wir noch, ob die Konvergenzbedingung (4.15) für den Startwert $x_0 = 1{,}5$ in diesem Beispiel erfüllt ist: Wir finden

$$|x_0 - z| \doteq |1{,}5 - 1{,}134724138| \doteq 0{,}365 < \frac{1}{|M|} \doteq \frac{1}{2{,}42} = 0{,}414,$$

d.h. die Konvergenzbedingung ist für den Startwert $x_0 = 1{,}5$ erfüllt. ♠

Wir halten in einer Bemerkung fest, was wir über die Konvergenz des Newton-Verfahrens gelernt haben:

Bemerkung 4.9. (Konvergenz des Newton-Verfahrens)
Sei $f :]a; b[\to \mathbb{R}$ eine stetig differenzierbare Funktion mit einer Nullstelle z und sei $f'(z) \neq 0$.

(1) Falls der Startwert x_0 des Newton-Verfahrens **dicht genug bei** z liegt, **konvergiert** das Newton-Verfahren gegen z, also $\lim\limits_{n\to\infty} x_n = z$.

(2) Falls die Iterierten x_n dicht genug bei z liegen, gilt für den **(absoluten) Fehler** von x_n die Näherung

$$x_n - z \approx x_n - x_{n+1} \qquad \Longrightarrow \qquad |x_n - z| \approx |x_n - x_{n+1}|.$$

(3) Falls f sogar zweimal stetig differenzierbar ist und falls die Iterierten x_n dicht genug bei z liegen, gilt für den **(absoluten) Fehler** von x_{n+1}

$$\left.\begin{array}{l} x_{n+1} - z \approx M\,(x_n - z)^2 \\ \Longrightarrow \quad |x_{n+1} - z| \approx |M|\,|x_n - z|^2 \end{array}\right\} \qquad \text{mit} \qquad M = \frac{f''(z)}{2\,f'(z)}\,.$$

Das Newton-Verfahren **konvergiert normalerweise gegen die Nullstelle** z, wenn der Startwert x_0 die folgende Bedingung erfüllt:

$$|x_0 - z| < \frac{1}{|M|} = \left|\frac{2\,f'(z)}{f''(z)}\right|$$

4.3 Sekanten-Verfahren

Beim **Sekantenverfahren** wird die Näherung der Nullstelle einer stetig differenzierbaren Funktion mit Hilfe von zwei Näherungen und der Sekante durch die beiden zugehörigen Punkte auf dem Graphen bestimmt. Genauer funktioniert dieses wie folgt (siehe auch Abbildung 4.4):

Sei $f :]a; b[\to \mathbb{R}$ eine stetig differenzierbare Funktion mit einer Nullstelle z, also $f(z) = 0$. Seien x_0 und x_1 zwei Näherungswerte für z. Diese können entweder beide auf einer Seite der Nullstelle liegen oder auf gegenüberliegenden Seiten der Nullstelle. Wir legen nun die Sekante durch die beiden Punkte $\big(x_0; f(x_0)\big)$ und $\big(x_1; f(x_1)\big)$ auf dem Graphen von f und bestimmen den Schnittpunkt x_2 der Sekante mit der x-Achse. Dieser Schnittpunkt ist die neue Näherung für die Nullstelle z. Die Gleichung der Sekante ist

$$p(x) = f(x_1) + \frac{f(x_1) - f(x_0)}{x_1 - x_0} \cdot (x - x_1).$$

(Es liegt ein Polynom ersten Grades vor, und in der Tat gelten $p(x_1) = f(x_1)$ und $p(x_0) = f(x_1) + \big(f(x_1) - f(x_0)\big) \cdot (-1) = f(x_0)$.) Wir lösen $p(x_2) = 0$ nach x_2 auf, um die Schnittstelle x_2 der Sekante mit der x-Achse zu bestimmen:

$$\begin{aligned}
0 = p(x_2) &= f(x_1) + \frac{f(x_1) - f(x_0)}{x_1 - x_0} \cdot (x_2 - x_1) \\
\iff \qquad -f(x_1) &= \frac{f(x_1) - f(x_0)}{x_1 - x_0} \cdot (x_2 - x_1) \\
\iff \qquad -f(x_1) \cdot \frac{x_1 - x_0}{f(x_1) - f(x_0)} &= x_2 - x_1 \\
\iff \qquad x_1 - f(x_1) \cdot \frac{x_1 - x_0}{f(x_1) - f(x_0)} &= x_2
\end{aligned}$$

Also erhalten wir als neue Näherung für die Nullstelle

$$x_2 = x_1 - f(x_1) \cdot \frac{x_1 - x_0}{f(x_1) - f(x_0)}\,. \tag{4.17}$$

Wir wiederholen diese Vorgehensweise mit den beiden Näherungen x_1 und x_2 für die Nullstelle z und erhalten als Schnittpunkt der Sekante durch $\big(x_1; f(x_1)\big)$ und $\big(x_2, f(x_2)\big)$ mit der x-Achse (durch Ersetzen in (4.17) von x_2 durch x_3, von x_1 durch x_2 und von x_0 durch x_1)

$$x_3 = x_2 - f(x_2) \cdot \frac{x_2 - x_1}{f(x_2) - f(x_1)}\,. \tag{4.18}$$

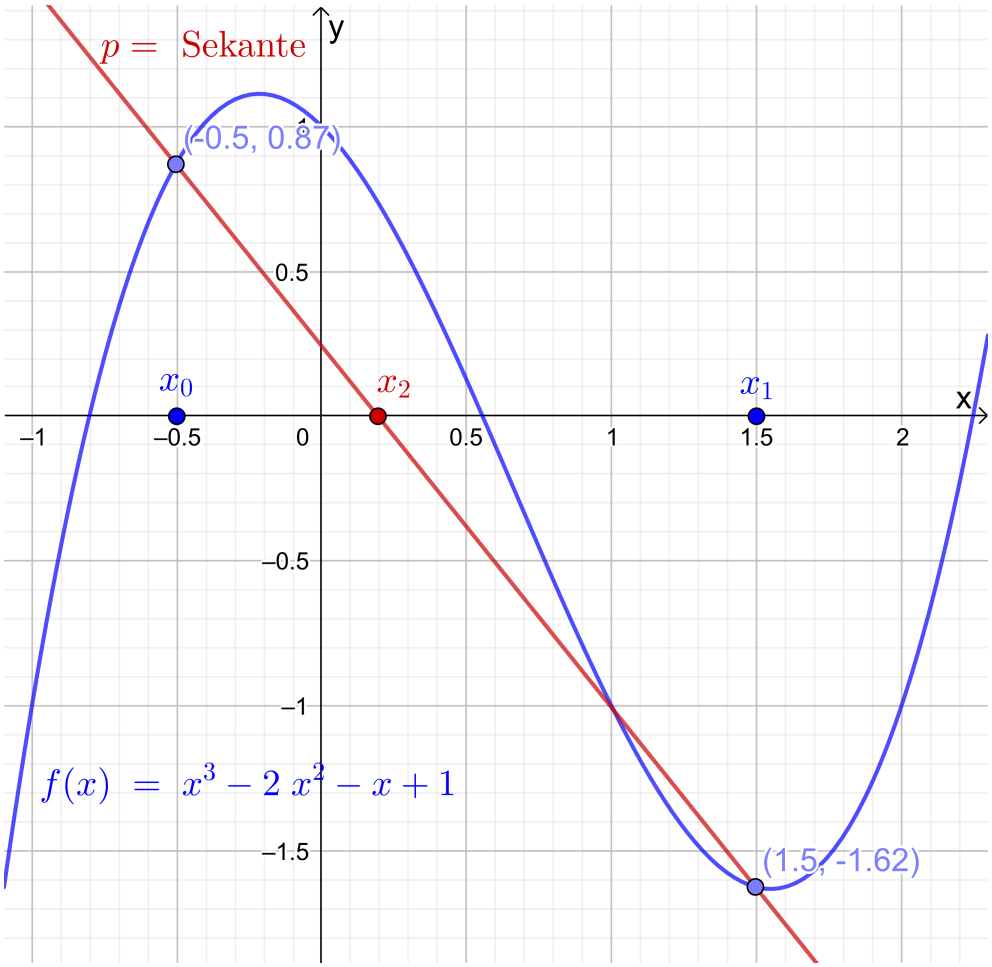


Abb. 4.4: Illustration des Sekantenverfahrens: Für die Startwerte x_0 und x_1 legen wir die Sekante p durch $\big(x_0; f(x_0)\big)$ und $\big(x_1; f(x_1)\big)$. Die neue Näherung x_2 für die Nullstelle ist der Schnittpunkt der Sekante mit der x-Achse.

Wir können diese Vorgehensweise nun mit den Näherungen x_2 und x_3 für die Nullstelle z fortsetzen. Das Wiederholen dieses Prozesses liefert nach n Schritten

$$x_{n+1} = x_n - f(x_n) \cdot \frac{x_n - x_{n-1}}{f(x_n) - f(x_{n-1})}\,. \tag{4.19}$$

(Natürlich sind (4.17) und (4.18) als Sonderfälle von (4.19) für $n = 1$ bzw. $n = 2$ in (4.19) enthalten.) Wir halten das Sekantenverfahren als Verfahren fest.

Verfahren 4.10. (Sekantenverfahren)
Sei $f :]a; b[\to \mathbb{R}$ eine stetig differenzierbare Funktion, die in z eine Nullstelle hat, und seien x_0 und x_1 zwei verschiedene (hinreichend gute) Näherungswerte für die Nullstelle z. Das ***Sekantenverfahren*** *berechnet Näherungen für die*

Nullstelle z mit dem folgenden Iterationsverfahren

$$x_{n+1} = x_n - f(x_n) \cdot \frac{x_n - x_{n-1}}{f(x_n) - f(x_{n-1})}, \qquad n = 1, 2, \ldots. \tag{4.20}$$

Das Sekantenverfahren ist ein **zweistufiges** Iterationsverfahren, weil zur Berechnung der neuen Näherung x_{n+1} zwei Näherungswerte x_n und x_{n-1} benötigt werden. Das Bisektionsverfahren ist ebenfalls ein zweistufiges Iterationsverfahren. Allerdings konvergiert das Sekantenverfahren normalerweise deutlich schneller als das Bisektionsverfahren. – Das Newton-Verfahren ist dagegen ein **einstufiges** Iterationsverfahren, da zur Berechnung der neuen Näherung x_{n+1} nur die Näherung x_n benötigt wird.

Betrachten wir als Beispiel wieder die Funktion, deren größte Nullstelle in den Beispielen 4.4 und 4.6 bereits mit den Bisektionsverfahren bzw. mit dem Newton-Verfahren bestimmt wurde.

Beispiel 4.11. (Sekanten-Verfahren)
Gesucht ist eine Näherung der größten reellen Nullstelle der Funktion

$$f : \mathbb{R} \to \mathbb{R}, \qquad f(x) = x^6 - x - 1,$$

die bereits in Beispiel 4.4 und 4.6 betrachtet wurde. In Beispiel 4.4 hatten wir uns überlegt, dass die größte Nullstelle im Intervall $[1; 2]$ liegt.

Wir nehmen daher als Startwerte für das Sekantenverfahren $x_0 = 2$ und $x_1 = 1$. Die Iterierten x_n sind für $n = 0, 1, 2, \ldots, 8$ in Tabelle 4.3 auf eine Gleitkommadarstellung mit einer 9-stelligen Mantisse gerundet aufgelistet. Weiter sind in Tabelle 4.3 $f(x_n)$, der Fehler $x_{n-1} - z$, sowie $x_{n-1} - x_n$ als Näherung für $x_{n-1} - z$ jeweils auf eine Gleitkommadarstellung mit einer 3-stelligen Mantisse gerundet angegeben.

Die gesuchte Nullstelle ist $z \doteq 1{,}134724138$, und wir sehen, dass nach 8 Iterationsschritten mit $x_8 \doteq 1{,}13472414$ eine Näherung mit 8 signifikanten Ziffern vorliegt. An der zweitletzten Spalte beobachten wir (wie auch beim Newton-Verfahren), dass der absolute Fehler zunächst nur langsam abnimmt, aber dann ab $n = 5$ rasant keiner wird. ♠

Die Fehleranalyse des Sekantenverfahrens ist mathematisch komplizierter als die des Newton-Verfahrens, und wir geben daher nur die Ergebnisse an. (Für die Herleitung dieser Ergebnisse siehe beispielsweise [13, Teilkapitel 5.3.1].)

n	x_n	$f(x_n)$	$x_{n-1} - x_n$	$x_{n-1} - z$
0	2,00000000	$6{,}10 \cdot 10^{1}$		
1	1,00000000	$-1{,}00 \cdot 10^{0}$	$1{,}00 \cdot 10^{0}$	
2	1,01612903	$-9{,}15 \cdot 10^{-1}$	$-1{,}61 \cdot 10^{-2}$	$-1{,}35 \cdot 10^{-1}$
3	1,19057777	$6{,}57 \cdot 10^{-1}$	$-1{,}74 \cdot 10^{-1}$	$-1{,}19 \cdot 10^{-1}$
4	1,11765583	$-1{,}68 \cdot 10^{-1}$	$-7{,}29 \cdot 10^{-2}$	$-5{,}59 \cdot 10^{-2}$
5	1,13253155	$-2{,}24 \cdot 10^{-2}$	$-1{,}49 \cdot 10^{-2}$	$-1{,}71 \cdot 10^{-2}$
6	1,13481681	$9{,}54 \cdot 10^{-4}$	$-2{,}29 \cdot 10^{-3}$	$-2{,}19 \cdot 10^{-3}$
7	1,13472365	$-5{,}07 \cdot 10^{-6}$	$9{,}32 \cdot 10^{-5}$	$9{,}27 \cdot 10^{-5}$
8	1,13472414	$-1{,}13 \cdot 10^{-9}$	$-4{,}92 \cdot 10^{-7}$	$-4{,}92 \cdot 10^{-7}$

Tabelle 4.3: Sekantenverfahren zur Berechnung der größten Nullstelle der Funktion $f(x) = x^6 - x - 1$ mit den Startwerten $x_0 = 2$ und $x_1 = 1$.

Bemerkung 4.12. (Konvergenz des Sekantenverfahrens)
Sei $f :]a; b[\to \mathbb{R}$ eine zweimal stetig differenzierbare Funktion mit einer Nullstelle z, also $f(z) = 0$, und sei $f'(z) \neq 0$.

(1) Wenn x_0 und x_1 **dicht genug** bei der Nullstelle z liegen, dann **konvergiert das Sekantenverfahren gegen** z, und es gilt

$$\lim_{n\to\infty} \frac{|x_{n+1} - z|}{|x_n - z|^r} = \left| \frac{f''(z)}{2\, f'(z)} \right|^{r-1} = |M|^{r-1} \qquad \text{mit} \quad M := \frac{f''(z)}{2\, f'(z)},$$

wobei $r = (\sqrt{5} + 1)/2 \doteq 1{,}62$. Für x_n dicht genug bei z gilt daher

$$|x_{n+1} - z| \approx |M|^{r-1}\, |x_n - z|^r \quad \text{mit} \quad r = (\sqrt{5} + 1)/2 \doteq 1{,}62. \tag{4.21}$$

(2) Aus (4.21) folgt durch Multiplizieren mit $|M|$

$$|M|\, |x_{n+1} - z| \approx \big(|M|\, |x_n - z|\big)^r \quad \text{mit} \quad r = (\sqrt{5} + 1)/2 \doteq 1{,}62,$$

und wiederholte Anwendung dieser Formel liefert

$$|M|\, |x_{n+1} - z| \approx \big(|M|\, |x_1 - z|\big)^{r^n} \quad \text{mit} \quad r = (\sqrt{5} + 1)/2 \doteq 1{,}62.$$

Da r^n mit wachsendem n beliebig groß wird, kann man nur erwarten, dass das **Sekantenverfahren konvergiert, wenn für** x_1 **gilt**

$$|M|\, |x_1 - z| < 1 \quad \Longleftrightarrow \quad |x_1 - z| < \frac{1}{|M|} = \left| \frac{2\, f'(z)}{f''(z)} \right|. \tag{4.22}$$

(3) Aus (4.21) kann man folgern, dass für x_n dicht genug bei z gilt:

$$x_{n-1} - z \approx x_{n-1} - x_n \qquad \Longrightarrow \qquad |x_{n-1} - z| \approx |x_{n-1} - x_n| \quad (4.23)$$

Wir machen uns die Informationen zur Konvergenz des Sekantenverfahrens am Beispiel 4.11 klar.

Beispiel 4.13. (Sekanten-Verfahren)

In Beispiel 4.11 wurde die größte reelle Nullstelle $z \doteq 1{,}134724138$ der Funktion $f : \mathbb{R} \to \mathbb{R},\ f(x) = x^6 - x - 1,$ mit dem Sekantenverfahren mit den Startwerten $x_0 = 2$ und $x_1 = 1$ berechnet. Wir haben die Konstante

$$M = \frac{f''(z)}{2\, f'(z)} = \frac{30\, z^4}{2\,(6\, z^5 - 1)} \doteq 2{,}42 \qquad \Longrightarrow \qquad |M| \doteq 2{,}42.$$

In Tabelle 4.3 lesen wir ab: $|x_5 - z| \doteq 2{,}19 \cdot 10^{-3}$ und $|x_4 - z| \doteq 1{,}71 \cdot 10^{-2}$. Daher gilt für die rechte Seite in (4.21) mit $n = 4$

$$|M|^{r-1}\, |x_4 - z|^r \doteq (2{,}42)^{0{,}62} \cdot (1{,}71 \cdot 10^{-2})^{1{,}62} \doteq 2{,}37 \cdot 10^{-3}.$$

Die Näherung (4.21) ist also bereits für $n + 1 = 5$ ziemlich gut erfüllt.

Wie sieht es mit der Konvergenzbedingung (4.22) aus? Wir erhalten für $x_1 = 1$

$$|x_1 - z| \doteq |1 - 1{,}134724138| \doteq 0{,}135 < \frac{1}{|M|} \doteq \frac{1}{2{,}42} = 0{,}414,$$

d.h. die Konvergenzbedingung ist für die Startwerte $x_0 = 2$ und $x_1 = 1$ erfüllt.

Betrachten wir noch (4.23) exemplarisch für $n = 5$: In Tabelle 4.3 lesen wir ab, dass gilt $x_5 - z \doteq -2{,}19 \cdot 10^{-3}$, und $x_5 - x_6 \doteq -2{,}29 \cdot 10^{-3}$ liefert in der Tat eine gute Näherung für $x_5 - z$. Für $n = 6$ und $n = 7$ sind die Näherungen $x_{n-1} - x_n \approx x_{n-1} - z$ in der Tabelle 4.3 ähnlich gut. ♠

4.4 Fixpunktiteration zur Lösung nicht-linearer Gleichungen

Natürlich kann man auch die Fixpunktiteration aus dem Banachschen Fixpunktsatz (siehe Satz 3.3) nutzen, um die Lösung einer Gleichung oder Nullstellengleichung zu finden, nachdem man diese vorher in eine geeignete Fixpunktgleichung

umgewandelt hat. Dieses soll hier nur kurz für Fall einer Gleichung mit einer Unbekannten diskutiert werden.

Aus dem **Banachschen Fixpunktsatz** (siehe Satz 3.3) folgt das nachfolgende Satz für eine reellwertige Funktion einer Variablen, die stetig differenzierbar ist:

Satz 4.14. (eindim. Fixpunktsatz mit verschärften Voraussetzungen)
*Sei $g :]c;d[\to \mathbb{R}$ eine **stetig differenzierbare** Funktion (d.h. g und g' sind stetig auf $]c;d[$). Es sei $[a;b] \subseteq]c;d[$, und g habe die Eigenschaften*

$$a \leq g(x) \leq b \qquad \text{für alle } x \in [a;b] \tag{4.24}$$

$$\text{und} \qquad \lambda = \max_{x \in [a;b]} |g'(x)| < 1. \tag{4.25}$$

Dann gelten:

*(1) Es gibt genau eine Lösung z der Gleichung $x = g(x)$, d.h. g hat **genau einen einzigen Fixpunkt** z in $[a;b]$.*

*(2) Für jeden Startwert $x_0 \in [a;b]$ **konvergieren die Iterierten***

$$x_{n+1} = g(x_n), \qquad n = 0, 1, 2, \ldots,$$

***gegen den Fixpunkt** z.*

*(3) **A posteriori Fehlerabschätzung:** Mit der Konstante λ aus (4.25) gilt*

$$|x_n - z| \leq \frac{\lambda}{1-\lambda} |x_n - x_{n-1}| \qquad \text{für alle } n = 1, 2, \ldots. \tag{4.26}$$

*(4) **A priori Fehlerabschätzung:** Mit der Konstante λ aus (4.25) gilt*

$$|x_n - z| \leq \frac{\lambda^n}{1-\lambda} |x_1 - x_0| \qquad \text{für alle } n = 1, 2, \ldots. \tag{4.27}$$

Beweis von Satz 4.14: Aus (4.24) folgt, dass $g(x) \in [a;b]$ für alle $x \in [a;b]$ gilt. Somit können wir g als Funktion $g : [a;b] \to [a;b]$ auffassen. Weiter folgt mit dem Mittelwertsatz der Differentialrechnung, dass es zu allen $x, y \in [a;b]$ mit $x \neq y$ ein z zwischen x und y gibt mit

$$g(x) - g(y) = g'(z)\,(x - y) \qquad \Longrightarrow \qquad |g(x) - g(y)| = |g'(z)|\,|x - y|$$

$$\Longrightarrow \qquad |g(x) - g(y)| \leq \underbrace{\left(\max_{t \in [a;b]} |g'(t)| \right)}_{= \lambda \text{ nach (4.25)}} |x - y| = \lambda \, |x - y|.$$

Da $x, y \in [a; b]$ beliebig waren, gilt also

$$|g(x) - g(y)| \leq \lambda \, |x - y| \qquad \text{für alle } x, y \in [a; b]$$

mit der Konstante $\lambda < 1$ aus (4.25). Damit ist die Funktion g eine Kontraktion auf dem Intervall $[a; b]$. Es sind also die Voraussetzungen von Banachschen Fixpunktsatz (siehe Satz 3.3) erfüllt. Damit folgen die Aussagen (1) bis (4) aus dem Banachschen Fixpunktsatz. □

Da die Bedingung (4.24) in der Praxis nicht immer leicht zu überprüfen ist bzw. da es nicht immer einfach ist, ein passendes Intervall $[a; b]$ mit der Eigenschaft (4.24) anzugeben, benötigen wir noch den folgenden Satz.

Satz 4.15. (Kriterium für Konvergenz der Fixpunktiteration)

Sei $g :]c; d[\to \mathbb{R}$ stetig differenzierbar, und g habe einen Fixpunkt z im Intervall $]c; d[$. Dann gelten die folgenden Aussagen:

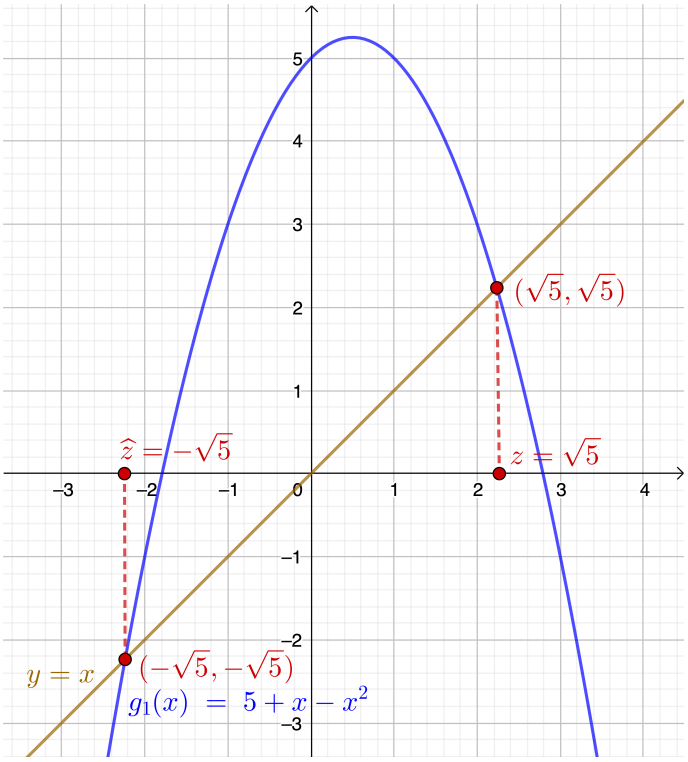
(1) Wenn $|g'(z)| < 1$ ist, dann gibt es ein Intervall $[a; b] \subseteq]c; d[$, in dem z liegt und für welches die Voraussetzungen (4.24) und (4.25) und somit auch alle Schlussfolgerungen aus Satz 4.14 erfüllt sind.

(2) Wenn $|g'(z)| > 1$ ist, dann konvergiert die Fixpunktiteration $x_{n+1} = g(x_n)$, $n = 0, 1, 2, \ldots$, nicht gegen z.

(3) Ist $|g'(z)| = 1$, so können wir keine Aussage treffen. (Falls die Fixpunktiteration in diesem Fall konvergieren sollte, wird die Konvergenz so langsam sein, dass das Verfahren nicht praktisch anwendbar ist.)

Praktische Anwendung der Fixpunktiteration: Satz 4.14 wird selten direkt angewendet, da es schwierig ist, ein Intervall $[a; b]$ mit (4.24) zu finden. **Statt dessen nutzt man in der Regel Satz 4.15 wie folgt:** Man schreibt die zu lösende Gleichung $f(x) = 0$ so in eine Fixpunktgleichung $x = g(x)$ um, dass für alle x in der Nähe des Fixpunkts gilt $|g'(x)| \leq \lambda < 1$ mit einer positiven Konstante $\lambda < 1$. Wenn man nun die Fixpunktiteration mit einem hinreichend guten Näherungswert x_0 für den Fixpunkt startet, kann man erwarten, dass die Fixpunktiteration gegen den Fixpunkt z konvergiert.

Betrachten wir einige Beispiele.



Grafische Bestimmung der Fixpunkte: Man findet die Fixpunkte grafisch, indem man die Schnittpunkte des Graphen der Funktion mit der Winkelhalbierenden $y = x$ (in braun gezeichnet) bestimmt.

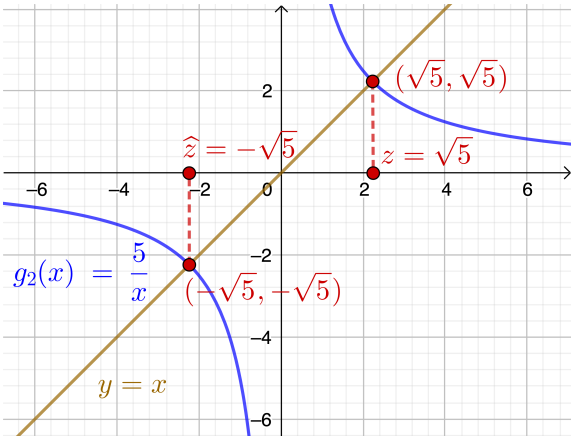


Abb. 4.5: Die Funktionen $g_1 : \mathbb{R} \to \mathbb{R}$, $g_1(x) = 5 + x - x^2$, und $g_2 : \mathbb{R} \setminus \{0\} \to \mathbb{R}$, $g_2(x) = \frac{5}{x}$, mit ihren Fixpunkten $z = \sqrt{5}$ und $\widehat{z} = -\sqrt{5}$.

Beispiel 4.16. (Fixpunktgleichungen und Fixpunktiterationen)
Wir betrachten die Gleichung $x^2 - 5 = 0$ mit der positiven Lösung $z = \sqrt{5} \doteq 2{,}2361$ (und der negativen Lösung $\widehat{z} = -\sqrt{5}$). Diese lässt sich mit Hilfe von

$$x^2 - 5 = 0 \quad \Longleftrightarrow \quad x^2 = 5 \quad \Longleftrightarrow \quad 5 - x^2 = 0 \quad \Longleftrightarrow \quad 1 - \frac{x^2}{5} = 0$$

jeweils in die folgenden vier Fixpunktgleichungen umformen:

$$\text{(a)} \quad x = 5 + x - x^2 \qquad \big(\text{addiere } x \text{ zu } 5 - x^2 = 0\big) \tag{4.28}$$

$$\text{(b)} \quad x = \frac{5}{x} \qquad \big(\text{dividiere } x^2 = 5 \text{ durch } x \neq 0\big) \tag{4.29}$$

$$\text{(c)} \quad x = 1 + x - \frac{x^2}{5} \qquad \left(\text{addiere } x \text{ zu } 1 - \frac{x^2}{5} = 0\right) \tag{4.30}$$

$$\text{(d)} \quad x = \frac{1}{2}\left(x + \frac{5}{x}\right) \qquad \left(\text{multipliziere (4.29) mit } \frac{1}{2} \text{ und addiere dann } \frac{1}{2}x\right) \tag{4.31}$$

Definieren wir nun die Funktionen

$$\text{(a)} \quad g_1 : \mathbb{R} \to \mathbb{R}, \quad g_1(x) = 5 + x - x^2,$$

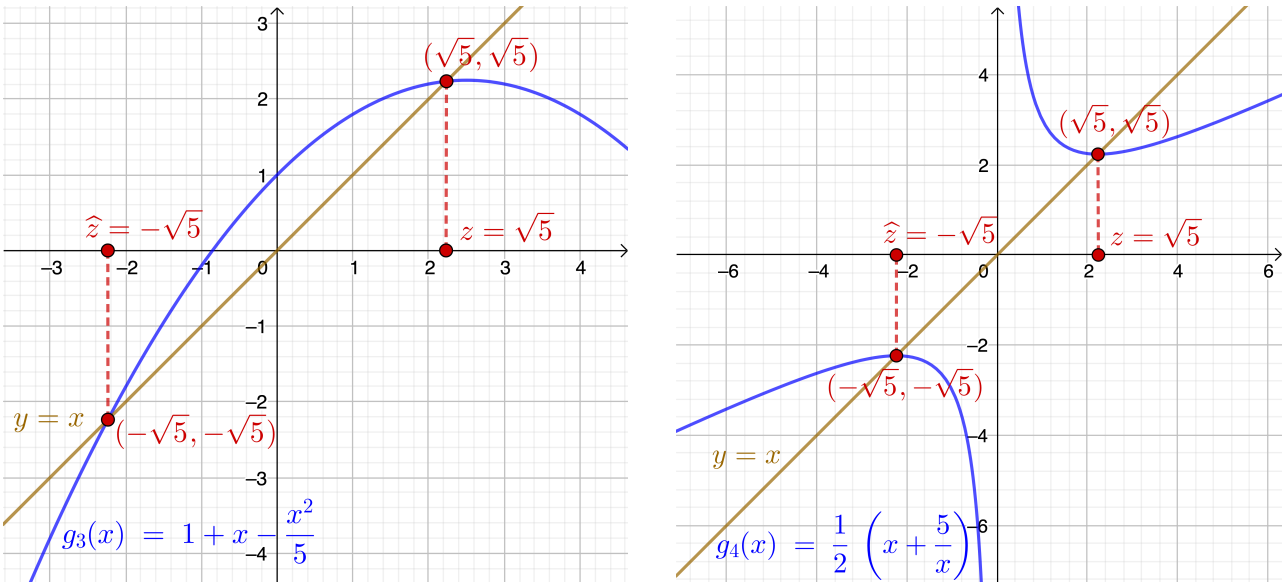


Abb. 4.6: Die Funktionen $g_3 : \mathbb{R} \to \mathbb{R}$, $g_3(x) = 1 + x - \frac{x^2}{5}$, und $g_4 : \mathbb{R} \setminus \{0\} \to \mathbb{R}$, $g_4(x) = \frac{1}{2}\left(x + \frac{5}{x}\right)$, mit ihren Fixpunkten $z = \sqrt{5}$ und $\widehat{z} = -\sqrt{5}$.

$$\text{(b)} \quad g_2 : \mathbb{R} \setminus \{0\} \to \mathbb{R}, \quad g_2(x) = \frac{5}{x},$$

$$\text{(c)} \quad g_3 : \mathbb{R} \to \mathbb{R}, \quad g_3(x) = 1 + x - \frac{x^2}{5},$$

$$\text{(d)} \quad g_4 : \mathbb{R} \setminus \{0\} \to \mathbb{R}, \quad g_4(x) = \frac{1}{2}\left(x + \frac{5}{x}\right),$$

so folgt aus (4.28), (4.29), (4.30) und (4.31), dass wir $x^2 - 5 = 0$ jeweils als die Fixpunktgleichung $g_k(x) = x$ mit $k \in \{1, 2, 3, 4\}$ schreiben können.

Die Funktionen g_k, $k = 1, 2, 3, 4$, mit ihren Fixpunkten $z = \sqrt{5}$ und $\widehat{z} = -\sqrt{5}$ sind in Abbildungen 4.5 und 4.6 gezeichnet. Man findet die Fixpunkte einer Funktion, indem man die Schnittpunkte des Graphen der Funktion mit der Winkelhalbierenden $y = x$ bestimmt.

Wir versuchen die positive Lösung $z = \sqrt{5} \doteq 2{,}236067977$ der quadratischen Gleichung $x^2 - 5 = 0$ jeweils mittels der Fixpunktiterationen $x_{n+1} = g_k(x_n)$ mit $n \in \mathbb{N}_0$ (für $k \in \{1, 2, 3, 4\}$) mit dem Startwert $x_0 = 2{,}5$ angenähert zu berechnen. Wir berechnen also die folgenden Fixpunktiterationen:

$$\text{(a)} \quad x_{n+1} = 5 + x_n - x_n^2 \quad \text{für} \quad g_1(x) = 5 + x - x^2, \tag{4.32}$$

$$\text{(b)} \quad x_{n+1} = \frac{5}{x_n} \quad \text{für} \quad g_2(x) = \frac{5}{x}, \tag{4.33}$$

n	x_n für g_1	x_n für g_2	x_n für g_3	x_n für g_4
0	2,500000	2,5	2,500000	2,500000
1	1,250000	2,0	2,250000	2,250000
2	4,687500	2,5	2,237500	2,236111
3	−12,28516	2,0	2,236219	2,236068
4	−158,2102	2,5	2,236084	2,236068
5	−25.183,68	2,0	2,236070	2,236068
6	−634.243.100	2,5	2,236068	2,236068

Tabelle 4.4: Fixpunktiteration für die Funktionen g_1, g_2, g_3 und g_4 aus Beispiel 4.16 mit dem Startwert $x_0 = 2{,}5$.

$$\text{(c)} \quad x_{n+1} = 1 + x_n - \frac{x_n^2}{5} \quad \text{für} \quad g_3(x) = 1 + x - \frac{x^2}{5}\,, \tag{4.34}$$

$$\text{(d)} \quad x_{n+1} = \frac{1}{2}\left(x_n + \frac{5}{x_n}\right) \quad \text{für} \quad g_4(x) = \frac{1}{2}\left(x + \frac{5}{x}\right). \tag{4.35}$$

Die Iterierten $x_{n+1} = g_k(x_n)$ für $n = 0, 1, 2, \ldots, 6$ sind jeweils in Tabelle 4.4 auf eine Gleitkommadarstellung mit 7-stelliger Mantisse gerundet aufgelistet.

Basierend auf der Berechnung von x_n für $n = 1, 2, \ldots, 6$ lässt sich vermuten, dass nur die Fixpunktiterationen für g_3 und g_4 gegen $z = \sqrt{5}$ konvergieren werden, denn hier wird im sechsten bzw. dritten Iterationsschritt eine Näherung für $z = \sqrt{5}$ mit 6 signifikanten Ziffern erreicht. Bei g_1 ist die Folge der Iterierten wohl unbeschränkt und divergiert. Bei g_2 pendelt die Folge der Iterierten immer zwischen 2,5 und 2 hin und her und ist somit ebenfalls divergent.

Können wir das beobachtete Verhalten mit Hilfe von Satz 4.15 erklären?

(a) Die Ableitung von $g_1(x) = 5 + x - x^2$ ist $g_1'(x) = 1 - 2\,x$. Also gilt

$$g_1'(z) = g_1'(\sqrt{5}) = 1 - 2\sqrt{5} \doteq -3{,}47 \quad \Longrightarrow \quad |g_1'(z)| \doteq 3{,}47 > 1,$$

und nach Satz 4.15 (2) wird die Fixpunktiteration divergieren.

(b) Die Ableitung von $g_2(x) = \dfrac{5}{x} = 5\,x^{-1}$ ist $g_2'(x) = -5\,x^{-2} = \dfrac{-5}{x^2}$. Also gilt

$$g_2'(z) = g_2'(\sqrt{5}) = \frac{-5}{(\sqrt{5})^2} = -1 \quad \Longrightarrow \quad |g_1'(z)| = |-1| = 1,$$

und nach Satz 4.15 (3) können wir keine Aussage treffen. Wir beobachten aber, dass die Iterierten nicht konvergieren, sondern abwechselnd die Werte 2,5 und 2,0 annehmen. (Man kann sich leicht überlegen, dass sich dieser Prozess fortsetzt, wenn wir x_n für $n = 7, 8, \ldots$, berechnen.)

(c) Die Ableitung von $g_3(x) = 1 + x - \frac{x^2}{5}$ ist $g_3'(x) = 1 - \frac{2\,x}{5}$. Also gilt

$$g_3'(z) = g_3'(\sqrt{5}) = 1 - \frac{2\sqrt{5}}{5} \doteq 0{,}106 \quad \Longrightarrow \quad |g_1'(z)| \doteq 0{,}106 < 1,$$

und nach Satz 4.15 (1) sollte die Fixpunktiteration für einen Startwert, der dicht genug bei $z = \sqrt{5}$ liegt, konvergieren. Dieses ist für den Startwert $x_0 = 2{,}5$ offenbar der Fall.

(d) Für $g_4(x) = \frac{1}{2}\left(x + \frac{5}{x}\right) = \frac{1}{2}\,x + \frac{5}{2}\,x^{-1}$ erhalten wir die Ableitung

$$g_4'(x) = \frac{1}{2} - \frac{5}{2}\,x^{-2} = \frac{1}{2} - \frac{5}{2}\,\frac{1}{x^2}\,.$$

Also gilt

$$g_4'(z) = g_4'(\sqrt{5}) = \frac{1}{2} - \frac{5}{2}\,\frac{1}{(\sqrt{5})^2} = 0 \quad \Longrightarrow \quad |g_1'(z)| = 0 < 1,$$

und nach Satz 4.15 (1) sollte die Fixpunktiteration für einen Startwert, der dicht genug bei $z = \sqrt{5}$ liegt, konvergieren. Dieses ist für den Startwert $x_0 = 2{,}5$ offenbar der Fall.

Wir können also in drei der vier Beispiele die Konvergenz bzw. Divergenz der Fixpunktiteration mit Hilfe von Satz 4.15 erklären. ♠

4.5 Newton-Verfahren für Gleichungssysteme

Wir lernen nun die Verallgemeinerung des Newton-Verfahrens zur Bestimmung der Lösungen von $\mathbf{f}(\mathbf{x}) = \mathbf{0}$ kennen, wobei $\mathbf{f} : D \to \mathbb{R}^n$, mit $D \subseteq \mathbb{R}^n$ offen und konvex, eine einmal stetig differenzierbare Funktion ist, deren Ableitungsmatrix in der Nähe der Nullstelle $\mathbf{z}$ (mit $\mathbf{f}(\mathbf{z}) = \mathbf{0}$) regulär ist. Wie in Teilkapitel 3.3 schreiben wir für die (vektoriellen) Iterierten $\mathbf{x}_k$ statt $\mathbf{x}^{(k)}$, da wir bei der Einführung des Newton-Verfahrens nur dan ganzen Vektor der Iterierten, aber nicht seine einzelnen Komponenten brauchen.

Sei also $\mathbf{f} : D \to \mathbb{R}^n$, mit $D \subseteq \mathbb{R}^n$ offen und konvex, eine zweimal stetig differenzierbare Funktion, die in $\mathbf{z} \in D$ ein Nullstelle habe. Nach dem Satz von Taylor (oder der Taylorschen Formel) gilt für das Taylor-Polynom von $\mathbf{f}$ mit dem Entwicklungspunkt $\mathbf{x} \in D$

$$\mathbf{0} = \mathbf{f}(\mathbf{z}) = \mathbf{f}(\mathbf{x}) + \big((\mathbf{Jf})(\mathbf{x})\big)\,(\mathbf{z}-\mathbf{x}) + \mathcal{O}\big(\|\mathbf{x}-\mathbf{z}\|^2\big),$$

wobei $(\mathbf{Jf})(\mathbf{x}) \in \mathbb{R}^{n\times n}$ die Jacobi-Matrix von $\mathbf{f}$ in $\mathbf{x}$ ist und wobei $\mathcal{O}\big(\|\mathbf{x}-\mathbf{z}\|^2\big)$ den Einfluss des Restglieds erfasst. Ist $\mathbf{x} = \mathbf{x}_k$ dicht bei der Nullstelle $\mathbf{z}$, so ist das Restglied der Ordnung $\mathcal{O}\big(\|\mathbf{x}_k-\mathbf{z}\|^2\big)$ vernachlässigbar, und es gilt genähert

$$\mathbf{0} \approx \mathbf{f}(\mathbf{x}_k) + \big((\mathbf{Jf})(\mathbf{x}_k)\big)\,(\mathbf{z}-\mathbf{x}_k).$$

Auflösen der Gleichung $\mathbf{0} = \mathbf{f}(\mathbf{x}_k) + \big((\mathbf{Jf})(\mathbf{x}_k)\big)\,(\mathbf{z}-\mathbf{x}_k)$ nach $\mathbf{z}$ liefert eine neue Näherung $\mathbf{x}_{k+1}$ für $\mathbf{z}$

$$\begin{aligned}
&\mathbf{0} = \mathbf{f}(\mathbf{x}_k) + \big((\mathbf{Jf})(\mathbf{x}_k)\big)\,(\mathbf{z}-\mathbf{x}_k) \quad\Longleftrightarrow\quad -\mathbf{f}(\mathbf{x}_k) = \big((\mathbf{Jf})(\mathbf{x}_k)\big)\,(\mathbf{z}-\mathbf{x}_k) \quad\Longleftrightarrow\\
&-\big((\mathbf{Jf})(\mathbf{x}_k)\big)^{-1}\mathbf{f}(\mathbf{x}_k) = \mathbf{z}-\mathbf{x}_k \quad\Longleftrightarrow\quad \mathbf{x}_k - \big((\mathbf{Jf})(\mathbf{x}_k)\big)^{-1}\mathbf{f}(\mathbf{x}_k) = \mathbf{z}, \qquad (4.36)
\end{aligned}$$

wobei wir für den Schritt in die zweite Zeile benötigen, dass $(\mathbf{Jf})(\mathbf{x}_k)$ regulär (also invertierbar) ist, so dass die inverse Matrix $\big((\mathbf{Jf})(\mathbf{x}_k)\big)^{-1}$ existiert. Wir erhalten also aus (4.36) als neue Näherung für die Nullstelle

$$\boxed{\mathbf{x}_{k+1} = \mathbf{x}_k - \big((\mathbf{Jf})(\mathbf{x}_k)\big)^{-1}\mathbf{f}(\mathbf{x}_k).} \qquad (4.37)$$

In der Praxis wird man aber nicht die inverse Matrix $\big((\mathbf{Jf})(\mathbf{x})\big)^{-1}$ berechnen, sondern mit einem geeigneten Verfahren (z.B. aus Kapiteln 2 oder 3) zunächst das lineare Gleichungssystem (siehe zweite Formel in der ersten Zeile von (4.36) mit $\mathbf{z} = \mathbf{x}_{k+1}$)

$$\boxed{\big((\mathbf{Jf})(\mathbf{x}_k)\big)\,\underbrace{(\mathbf{x}_{k+1}-\mathbf{x}_k)}_{=:\,\mathbf{z}_k} = -\mathbf{f}(\mathbf{x}_k)}$$

lösen und dann die neue Näherung der Nullstelle mittels

$$\boxed{\mathbf{x}_{k+1} = \mathbf{x}_k + \mathbf{z}_k, \quad \text{wobei} \quad \mathbf{z}_k \text{ die Lösung von } \big((\mathbf{Jf})(\mathbf{x}_k)\big)\,\mathbf{z}_k = -\mathbf{f}(\mathbf{x}_k) \text{ ist,}}$$

berechnen. Wir halten das Newton-Verfahren für nicht-lineare (oder lineare) Gleichungssysteme als numerisches Verfahren fest.

Verfahren 4.17. (Newton-Verfahren für Gleichungssysteme)
*Sei $D \subseteq \mathbb{R}^n$ offen und konvex, und sei $\mathbf{f} : D \to \mathbb{R}^n$ **stetig differenzierbar** und habe eine Nullstelle $\mathbf{z}$ in D. Für alle $\mathbf{x} \in D$ sei die Jacobi-Matrix $(\mathbf{Jf})(\mathbf{x})$ regulär (also invertierbar). Weiter seien eine Fehlerschranke $\epsilon > 0$, eine Norm $\|\cdot\|$ für $\mathbb{R}^n$ und ein Startvektor $\mathbf{x}_0$ vorgegeben.*

Für $k = 1, 2, \ldots$ führe folgende Schritte durch

(1) Löse das lineare Gleichungssystem $\big((\mathbf{Jf})(\mathbf{x}_k)\big)\,\mathbf{z}_k = -\mathbf{f}(\mathbf{x}_k)$.

(2) Berechne die neue Näherung $\mathbf{x}_{k+1} = \mathbf{x}_k + \mathbf{z}_k$.

bis $\|\mathbf{x}_{k+1} - \mathbf{x}_k\| \leq \epsilon$ ist. Dann stoppe.

Betrachten wir ein Beispiel für das Newton-Verfahren für Gleichungssysteme.

Beispiel 4.18. (Newton-Verfahren für Gleichungssysteme)
Gesucht ist die Lösung $\mathbf{z} = [u; v]^T$ des Systems der beiden quadratischen Gleichungen

$$\begin{aligned} x^2 + y^2 + 0{,}6\,y - 0{,}16 &= 0, \\ x^2 - y^2 + x - 1{,}6\,y - 0{,}14 &= 0, \end{aligned} \tag{4.38}$$

für die $u > 0$ und $v > 0$ gilt. Hier haben wir also

$$\mathbf{f} : \mathbb{R}^2 \to \mathbb{R}^2, \quad \mathbf{f}(x; y) = \begin{bmatrix} f_1(x; y) \\ f_2(x; y) \end{bmatrix} \quad \text{mit} \quad \begin{cases} f_1(x; y) = x^2 + y^2 + 0{,}6\,y - 0{,}16, \\ f_2(x; y) = x^2 - y^2 + x - 1{,}6\,y - 0{,}14, \end{cases}$$

Wir berechnen zunächst die Jacobi-Matrix

$$(\mathbf{Jf})(x; y) = \begin{bmatrix} (\partial_x f_1)(x; y) & (\partial_y f_1)(x; y) \\ (\partial_x f_2)(x; y) & (\partial_y f_2)(x; y) \end{bmatrix} = \begin{bmatrix} 2\,x & 2\,y + 0{,}6 \\ 2\,x + 1 & -2\,y - 1{,}6 \end{bmatrix}.$$

Wir berechnen die Determinante

$$\begin{aligned} &\det\big((\mathbf{Jf})(x; y)\big) = 2\,x\,(-2\,y - 1{,}6) - (2\,x + 1)\,(2\,y + 0{,}6) \\ &= -4\,x\,y - 3{,}2\,x - 4\,x\,y - 2\,y - 1{,}2\,x - 0{,}6 = -\big(8\,x\,y + 4{,}4\,x + 2\,y + 0{,}6\big). \end{aligned}$$

Für alle $(x; y)$ mit $x \geq 0$ und $y \geq 0$ gilt $\det\big((\mathbf{Jf})(x; y)\big) \leq -0{,}6$, d.h. insbesondere ist $\det\big((\mathbf{Jf})(x; y)\big) \neq 0$, so dass die Jacobi-Matrix $(\mathbf{Jf})(x; y)$ für alle $(x; y) \in [0; \infty[\times [0; \infty[$ regulär, also invertierbar, ist.

k	x_k	y_k	$\|\mathbf{x}_{k-1} - \mathbf{z}\|_2$	$\|\mathbf{x}_{k-1} - \mathbf{x}_k\|_2$
0	0,6000000000	0,2500000000		
1	0,3450404858	0,1531376518	$3{,}531 \cdot 10^{-1}$	$2{,}727 \cdot 10^{-1}$
2	0,2775310555	0,1224629827	$8{,}050 \cdot 10^{-2}$	$7{,}415 \cdot 10^{-2}$
3	0,2718851108	0,1196643843	$6{,}347 \cdot 10^{-3}$	$6{,}301 \cdot 10^{-3}$
4	0,2718445085	0,1196433787	$4{,}572 \cdot 10^{-5}$	$4{,}571 \cdot 10^{-5}$
5	0,2718445063	0,1196433776	$2{,}460 \cdot 10^{-9}$	$2{,}460 \cdot 10^{-9}$

Tabelle 4.5: Die ersten fünf Iterationsschritte $\mathbf{x}_k = [x_k; y_k]^T$ des Newton-Verfahrens zur Bestimmung der Nullstelle $\mathbf{z} = [u; v]^T$ mit $u, v > 0$ des nicht-linearen Gleichungssystems (4.38) aus Beispiel 4.18.

In $(k+1)$-ten Schritt des Newton-Verfahrens müssen wir nun zuerst das LGS

$$\begin{bmatrix} 2\,x_k & 2\,y_k + 0{,}6 \\ 2\,x_k + 1 & -2\,y_k - 1{,}6 \end{bmatrix} \begin{bmatrix} s_k \\ t_k \end{bmatrix} = - \begin{bmatrix} f_1(x_k; y_k) \\ f_2(x_k; y_k) \end{bmatrix}$$

zur Bestimmung von $\mathbf{z}_k = [s_k; t_k]^T$ lösen und berechnen danach mit

$$\begin{bmatrix} x_{k+1} \\ y_{k+1} \end{bmatrix} = \begin{bmatrix} x_k \\ y_k \end{bmatrix} + \begin{bmatrix} s_k \\ t_k \end{bmatrix}$$

die neue Näherung der Nullstelle.

Wir verwenden die Startwerte $x_0 = 0{,}6$ und $y_0 = 0{,}25$, also $\begin{bmatrix} x_0 \\ y_0 \end{bmatrix} = \begin{bmatrix} 0{,}6 \\ 0{,}25 \end{bmatrix}$.

Schritt 1: Das LGS, welches wir im ersten Schritt lösen müssen, lautet

$$\begin{bmatrix} 1{,}2 & 1{,}1 \\ 2{,}2 & -2{,}1 \end{bmatrix} \begin{bmatrix} s_0 \\ t_0 \end{bmatrix} = - \begin{bmatrix} 0{,}4125 \\ 0{,}3575 \end{bmatrix}.$$

Dessen Lösung ist $s_0 \doteq -0{,}254960$ und $t_0 \doteq -0{,}096862$, und mittels

$$\begin{bmatrix} x_1 \\ y_1 \end{bmatrix} = \begin{bmatrix} x_0 \\ y_0 \end{bmatrix} + \begin{bmatrix} s_0 \\ t_0 \end{bmatrix} \doteq \begin{bmatrix} 0{,}6 \\ 0{,}25 \end{bmatrix} + \begin{bmatrix} -0{,}254960 \\ -0{,}096862 \end{bmatrix} = \begin{bmatrix} 0{,}345040 \\ 0{,}153138 \end{bmatrix}$$

erhalten wir die neue Näherung.

Analog führt man weitere Schritte durch.

In Tabelle 4.5 sind die Ergebnisse der ersten fünf Iterationsschritte (mit Rundung auf eine Gleitkommadarstellung mit 10-stelliger Mantisse) aufgelistet. Zusätzlich zum absoluten Fehler $\|\mathbf{x}_{k-1} - \mathbf{z}\|_2$ ist auch die Näherung $\|\mathbf{x}_{k-1} - \mathbf{x}_k\|_2$, also

$$\left\| \begin{bmatrix} x_{k-1} \\ y_{k-1} \end{bmatrix} - \begin{bmatrix} u \\ v \end{bmatrix} \right\|_2 \approx \left\| \begin{bmatrix} x_{k-1} \\ y_{k-1} \end{bmatrix} - \begin{bmatrix} x_k \\ y_k \end{bmatrix} \right\|_2$$

für den absoluten Fehler $\|\mathbf{x}_{k-1} - \mathbf{z}\|_2$ aufgelistet, mit welcher wir den absoluten Fehler der Näherungen bei unbekannten $\mathbf{z}$ „überwachen“ können. Wir sehen in Tabelle 4.5, dass nach bereits vier Iterationsschritten die Näherung der Nullstelle 8 signifikante Ziffern hat. ♠

Der nachfolgende Satz gibt Informationen über die Konvergenz des Newton-Verfahrens für nicht-lineare (oder lineare) Gleichungssysteme.

Satz 4.19. (Konvergenz des Newton-Verf. für Gleichungssysteme)
*Sei $D \subseteq \mathbb{R}^n$ offen, und sei $\|\cdot\|$ eine Norm für $\mathbb{R}^n$. Die Funktion $\mathbf{f} : D \to \mathbb{R}^n$ sei **zweimal stetig differenzierbar** und habe eine Nullstelle $\mathbf{z}$ in D, und $(\mathbf{Jf})(\mathbf{z})$ sei regulär. Dann gibt es einen Radius $\delta > 0$, so dass $\{\mathbf{x} \in \mathbb{R}^n : \|\mathbf{x} - \mathbf{z}\| \leq \delta\} \subseteq D$ ist und dass das Newton-Verfahren für jeden Startvektor $\mathbf{x}_0 \in \{\mathbf{x} \in \mathbb{R}^n : \|\mathbf{x} - \mathbf{z}\| \leq \delta\}$ **durchführbar ist und gegen $\mathbf{z}$ konvergiert**. Für die Iterierten $\mathbf{x}_k$, $k = 0, 1, 2, \ldots$, des Newton-Verfahrens gilt*

$$\|\mathbf{x}_{k+1} - \mathbf{z}\| \leq c \, \|\mathbf{x}_k - \mathbf{z}\|^2 \tag{4.39}$$

mit einer (von k unabhängigen) Konstanten $c \geq 0$.

Natürlich gelten Verfahren 4.17 und Satz 4.19 auch, wenn $n = 1$ ist. Für diesen Sonderfall erhalten wir das „normale“ Newton-Verfahren aus Teilkapitel 4.2.

Wir nehmen für den Moment an, dass die Iterierten $\mathbf{x}_k$ des Newton-Verfahrens gegen die Nullstelle $\mathbf{z}$ konvergieren. Aus der Iterationsvorschrift (siehe (4.37))

$$\mathbf{x}_{k+1} = \mathbf{x}_k - \big((\mathbf{Jf})(\mathbf{x}_k)\big)^{-1} \mathbf{f}(\mathbf{x}_k), \qquad k \in \mathbb{N}_0,$$

folgt $\mathbf{x}_k - \mathbf{x}_{k+1} = \big((\mathbf{Jf})(\mathbf{x}_k)\big)^{-1} \mathbf{f}(\mathbf{x}_k)$. Falls $\mathbf{x}_k$ dicht bei der Nullstelle $\mathbf{z}$ liegt, so gilt genähert $\big((\mathbf{Jf})(\mathbf{x}_k)\big)^{-1} = \big((\mathbf{Jf})(\mathbf{z})\big)^{-1}$, also

$$\mathbf{x}_k - \mathbf{x}_{k+1} \approx \big((\mathbf{Jf})(\mathbf{z})\big)^{-1} \mathbf{f}(\mathbf{x}_k), \tag{4.40}$$

und die (komponentenweise) Näherung von $\mathbf{f}(\mathbf{x}_k)$ durch das Taylorpolynom vom

Grad 1 mit dem Entwicklungspunkt $\mathbf{z}$ liefert

$$\mathbf{f}(\mathbf{x}_k) \approx \underbrace{\mathbf{f}(\mathbf{z})}_{=\mathbf{0}} + \big((\mathbf{Jf})(\mathbf{z})\big)\,(\mathbf{x}_k - \mathbf{z}) = \big((\mathbf{Jf})(\mathbf{z})\big)\,(\mathbf{x}_k - \mathbf{z}). \tag{4.41}$$

Einsetzen von (4.41) in (4.40) liefert

$$\mathbf{x}_k - \mathbf{x}_{k+1} \approx \big((\mathbf{Jf})(\mathbf{z})\big)^{-1}\,\big((\mathbf{Jf})(\mathbf{z})\big)\,(\mathbf{x}_k - \mathbf{z}) = \mathbf{x}_k - \mathbf{z}.$$

Also gilt

$$\boxed{\mathbf{x}_k - \mathbf{x}_{k+1} \approx \mathbf{x}_k - \mathbf{z}} \quad \Longrightarrow \quad \boxed{\|\mathbf{x}_k - \mathbf{x}_{k+1}\| \approx \|\mathbf{x}_k - \mathbf{z}\|,}$$

d.h. $\|\mathbf{x}_k - \mathbf{x}_{k+1}\|$ liefert für $\mathbf{x}_k$ dicht bei $\mathbf{z}$ eine **gute Näherung für den absoluten Fehler** $\|\mathbf{x}_k - \mathbf{z}\|$. Wir halten dieses in einer Bemerkung fest.

Bemerkung 4.20. (a posteriori Fehlerabschätzung für das Newton–Verfahren für Gleichungssysteme)
Im Folgenden seien die Voraussetzungen wie in Satz 4.19, und alle $\mathbf{x}_k$ liegen in $\{\mathbf{x} \in \mathbb{R}^n \,:\, \|\mathbf{x} - \mathbf{z}\| \le \delta\} \subseteq D$ mit einem passend gewählten $\delta > 0$ und die Iterierten $\mathbf{x}_k$ des Newton-Verfahren (siehe Verfahren 4.17) konvergieren gegen die Nullstelle $\mathbf{z} \in D$. Wenn die Iterierten $\mathbf{x}_k$ dicht genug bei $\mathbf{z}$ liegen, so gilt die Näherung

$$\mathbf{x}_k - \mathbf{x}_{k+1} \approx \mathbf{x}_k - \mathbf{z}.$$

Daraus folgt die nachfolgende **Näherung für den absoluten Fehler**

$$\|\mathbf{x}_k - \mathbf{x}_{k+1}\| \approx \|\mathbf{x}_k - \mathbf{z}\|.$$

Da die Nullstelle $\mathbf{z}$ in der Regel nicht bekannt ist, verwenden wir in der Praxis die Näherung $\|\mathbf{x}_k - \mathbf{x}_{k+1}\|$ zur Berechnung des absoluten Fehlers $\|\mathbf{x}_k - \mathbf{z}\|$.

Beispiel 4.21. (Newton-Verfahren für Gleichungssysteme)
In Beispiel 4.18 wurde die Lösung $\mathbf{z} = [u; v]^T$ mit $u > 0$ und $v > 0$ des Systems der beiden quadratischen Gleichungen

$$x^2 + y^2 + 0{,}6\,y - 0{,}16 = 0,$$
$$x^2 - y^2 + x - 1{,}6\,y - 0{,}14 = 0,$$

in Tabelle 4.5 mit dem Newton-Verfahren für Gleichungssysteme berechnet. An den letzten beiden Spalten von Tabelle 4.5 sieht man, dass die Näherung

$$\|\mathbf{x}_{k-1} - \mathbf{x}_k\| \approx \|\mathbf{x}_{k-1} - \mathbf{z}\|$$

ab $n = 3$ gut und ab $n = 4$ sehr gut erfüllt ist. ♠

Der Beweis von Satz 4.19 soll der Vollständigkeit halber für mathematisch Interessierte erklärt werden.

Beweis von Satz 4.19: Da D offen ist, gibt es ein $\delta_0 > 0$, so dass die abgeschlossene Kugel $\{\mathbf{x} \in \mathbb{R}^n \ : \ \|\mathbf{x} - \mathbf{z}\| \leq \delta\}$ mit Mittelpunkt $\mathbf{z}$ und Radius $\delta > 0$ für alle $\delta \leq \delta_0$ in D liegt. Da $(\mathbf{Jf})(\mathbf{z})$ regulär ist, gilt $\det\big((\mathbf{Jf})(\mathbf{z})\big) \neq 0$. Die reellwertige Funktion $g : D \to \mathbb{R}$, $g(\mathbf{x}) = \det\big((\mathbf{Jf})(\mathbf{x})\big)$, ist stetig (weil $\mathbf{f}$ stetig differenzierbar ist) und erfüllt $g(\mathbf{z}) \neq 0$. Daraus folgt (wegen der Stetigkeit von g), dass ein Radius $\widehat{\delta} > 0$ mit $0 < \widehat{\delta} < \delta_0$ existiert, so dass $g(\mathbf{x}) \neq 0$ für alle $\mathbf{x} \in \{\mathbf{x} \in \mathbb{R}^n \ : \ \|\mathbf{x} - \mathbf{z}\| \leq \widehat{\delta}\} \subseteq D$ gilt und dass weiter mit einer positiven Konstante c_1 gilt

$$\|\big((\mathbf{Jf})(\mathbf{z})\big)^{-1}\| \leq c_1 \qquad \text{für alle } \mathbf{x} \in \{\mathbf{x} \in \mathbb{R}^n \ : \ \|\mathbf{x} - \mathbf{z}\| \leq \widehat{\delta}\}. \tag{4.42}$$

Sei nun $\mathbf{x}_k \in \{\mathbf{x} \in \mathbb{R}^n \ : \ \|\mathbf{x}-\mathbf{z}\| \leq \delta\}$. Der Satz von Taylor liefert (komponentenweise) für das Taylorpolynom vom Grad 1 von $\mathbf{f}$ mit dem Entwicklungspunkt $\mathbf{x}_k$

$$\begin{aligned} \mathbf{0} = \mathbf{f}(\mathbf{z}) &= \mathbf{f}(\mathbf{x}_k) + \big((\mathbf{Jf})(\mathbf{x}_k)\big)\,(\mathbf{z} - \mathbf{x}_k) + \mathcal{O}(\|\mathbf{x}_k - \mathbf{z}\|^2) \\ \iff \qquad \mathbf{f}(\mathbf{x}_k) &+ \big((\mathbf{Jf})(\mathbf{x}_k)\big)\,(\mathbf{z} - \mathbf{x}_k) = \mathcal{O}(\|\mathbf{x}_k - \mathbf{z}\|^2) \end{aligned} \tag{4.43}$$

Damit folgt aus (4.43) mit einer positiven Konstante c_2

$$\|\mathbf{f}(\mathbf{x}_k) + \big((\mathbf{Jf})(\mathbf{x}_k)\big)\,(\mathbf{z} - \mathbf{x}_k)\| \leq c_2\,\|\mathbf{x}_k - \mathbf{z}\|^2. \tag{4.44}$$

Mit der Iterationsvorschrift (siehe (4.37)) bekommen wir

$$\begin{aligned} \mathbf{x}_{k+1} - \mathbf{z} &= \big[\mathbf{x}_k - \big((\mathbf{Jf})(\mathbf{x}_k)\big)^{-1}\,\mathbf{f}(\mathbf{x}_k)\big] - \mathbf{z} \\ &= -\big((\mathbf{Jf})(\mathbf{x}_k)\big)^{-1}\,\mathbf{f}(\mathbf{x}_k) + \mathbf{x}_k - \mathbf{z} \\ &= \big((\mathbf{Jf})(\mathbf{x}_k)\big)^{-1}\,\big[-\mathbf{f}(\mathbf{x}_k) - \big((\mathbf{Jf})(\mathbf{x}_k)\big)\,(\mathbf{z} - \mathbf{x}_k)\big]. \end{aligned}$$

Damit erhalten wir für den absoluten Fehler

$$\begin{aligned} \|\mathbf{x}_{k+1} - \mathbf{z}\| &= \Big\|\big((\mathbf{Jf})(\mathbf{x}_k)\big)^{-1}\,\big[-\mathbf{f}(\mathbf{x}_k) - \big((\mathbf{Jf})(\mathbf{x}_k)\big)\,(\mathbf{z} - \mathbf{x}_k)\big]\Big\| \\ &= \underbrace{\Big\|\big((\mathbf{Jf})(\mathbf{x}_k)\big)^{-1}\Big\|}_{\leq c_1 \text{ nach (4.42)}} \cdot \underbrace{\Big\| -\mathbf{f}(\mathbf{x}_k) - \big((\mathbf{Jf})(\mathbf{x}_k)\big)\,(\mathbf{z} - \mathbf{x}_k)\Big\|}_{= \|\mathbf{f}(\mathbf{x}_k) + ((\mathbf{Jf})(\mathbf{x}_k))\,(\mathbf{z}-\mathbf{x}_k)\| \leq c_2\,\|\mathbf{x}_k - \mathbf{z}\|^2 \text{ nach (4.44)}} \\ &\leq c_1\,c_2\,\|\mathbf{x}_k - \mathbf{z}\|^2, \end{aligned}$$

wobei wir im letzten Schritt (4.42) und (4.44) angewendet haben. Also gilt

$$\|\mathbf{x}_{k+1} - \mathbf{z}\| \leq c_1\,c_2\,\|\mathbf{x}_k - \mathbf{z}\|^2, \tag{4.45}$$

womit (4.39) bewiesen ist.

Da $\mathbf{x}_k \in \{\mathbf{x} \in \mathbb{R}^n \ : \ \|\mathbf{x} - \mathbf{z}\| \leq \delta\} \subseteq D$ ist, gilt insbesondere $\|\mathbf{x}_k - \mathbf{z}\| \leq \delta$. Damit folgt aus (4.45)

$$\|\mathbf{x}_{k+1} - \mathbf{z}\| \leq c_1 \, c_2 \underbrace{\|\mathbf{x}_k - \mathbf{z}\|}_{\leq \delta} \cdot \|\mathbf{x}_k - \mathbf{z}\| \leq c_1 \, c_2 \, \delta \, \|\mathbf{x}_k - \mathbf{z}\| \tag{4.46}$$

Indem wir den Radius $\delta > 0$ falls erforderlich noch weiter verkleinern, können wir immer erreichen, dass $c_1 \, c_2 \, \delta =: \kappa < 1$ ist. (Das Verkleinern des Radius $\delta > 0$ vergrößert die Konstanten c_1 und c_2 nicht. Also kann hier nichts schief gehen.) Dann gilt mit einem $\kappa < 1$

$$\|\mathbf{x}_{k+1} - \mathbf{z}\| \leq \kappa \, \|\mathbf{x}_k - \mathbf{z}\|,$$

und es folgt, dass $\mathbf{x}_k$ gegen $\mathbf{z}$ konvergiert, weil sich der Abstand zur Nullstelle mit jedem Iterationsschritt mindestens um den Faktor $\kappa < 1$ verkleinert. □

4.6 Konvergenzordnung von Iterationsverfahren

Als Letztes lernen wir das Konzept der Konvergenzordnung eines Iterationsverfahrens kennen.

Definition 4.22. (Konvergenzordnung eines Iterationsverfahrens)
*Sei $\|\cdot\|$ eine Norm auf $\mathbb{R}^n$. Sei $\big(\mathbf{x}_k\big)_{k \in \mathbb{N}_0}$ die Folge (in $\mathbb{R}^n$) der von einem Iterationsverfahren $\mathbf{x}_{k+1} = \mathbf{f}(\mathbf{x}_k)$ erzeugten Iterierten, wobei $\mathbf{f} : D \to \mathbb{R}^n$ mit $D \subseteq \mathbb{R}^n$ die Iterationsvorschrift ist. Das Iterationsverfahren **konvergiere gegen $\mathbf{z}$**. Wir sagen, dass das Iterationsverfahren*

*(1) (mindestens) die **Konvergenzordnung 1** hat bzw. (mindestens) **linear konvergent** ist, wenn es eine positive Kostante $c < 1$ und einen Index k_0 gibt, so dass gilt*

$$\|\mathbf{x}_{k+1} - \mathbf{z}\| \leq c \, \|\mathbf{x}_k - \mathbf{z}\| \qquad \text{für alle } k \geq k_0.$$

*(2) (mindestens) die **Konvergenzordnung 2** hat bzw. **quadratisch konvergent** ist, wenn es eine positive Kostante $c > 0$ gibt, so dass gilt*

$$\|\mathbf{x}_{k+1} - \mathbf{z}\| \leq c \, \|\mathbf{x}_k - \mathbf{z}\|^2 \qquad \text{für alle } k = 0, 1, 2, \ldots.$$

(3) (mindestens) die ***Konvergenzordnung*** $p > 1$ *hat, wenn es eine positive Kostante* $c > 0$ *gibt, so dass gilt*

$$\|\mathbf{x}_{k+1} - \mathbf{z}\| \leq c \, \|\mathbf{x}_k - \mathbf{z}\|^p \qquad \text{für alle } k = 0, 1, 2, \ldots.$$

(4) ***superlinear konvergent*** *ist, wenn es eine positive Nullfolge* $(c_k)_{k \in \mathbb{N}_0}$ *gibt (d.h.* $c_k > 0$ *für alle* k *und* $\lim_{k \to \infty} c_k = 0$*), so dass gilt*

$$\|\mathbf{x}_{k+1} - \mathbf{z}\| \leq c_k \, \|\mathbf{x}_k - \mathbf{z}\| \qquad \text{für alle } k = 0, 1, 2, \ldots.$$

(Falls $n = 1$ *ist, wird die Norm* $\| \cdot \|$ *immer durch den Absolutbetrag ersetzt.)*

Natürlich ist Definition 4.22 (2) ein Sonderfall von Definition 4.22 (3).

Welche Konvergenzordnungen haben die Iterationsverfahren, die wir in dieser Vorlesung bisher kennengelernt haben?

- Die **Fixpunktiteration** aus dem Banachschen Fixpunktsatz (siehe Satz 3.3) hat die **Konvergenzordnung 1** bzw. ist **linear konvergent**, denn mit der Kontraktionskonstante q mit $0 < q < 1$ gilt wegen der Kontraktionseigenschaften (mit dem Fixpunkt $\widehat{\mathbf{x}}$)

$$\|\mathbf{x}_{k+1} - \widehat{\mathbf{x}}\| = \|\mathbf{f}(\mathbf{x}_k) - \mathbf{f}(\widehat{\mathbf{x}})\| \leq q \, \|\mathbf{x}_k - \widehat{\mathbf{x}}\| \quad \text{für alle } k = 0, 1, 2, \ldots$$

- Das **Jacobi-Verfahren** und das **Gauß-Seidel-Verfahren** basieren auf der Fixpunktiteration und haben damit ebenfalls mindestens die **Konvergenzordnung 1** bzw. sind ebenfalls mindestens **linear konvergent**.
- Das **Verfahren der konjugierten Gradienten (CG-Verfahren)** bricht nach spätestens n Schritten ab, so dass es hier nur begrenzt Sinn macht (nämlich nur für großes n), von einer Konvergenzordnung zu sprechen. Man kann dann aus Hilfssatz 3.30 folgern, dass das CG-Verfahren mindestens **linear konvergent** ist.
- Das **Bisektionsverfahren** konvergiert immer, aber wir können ihm keine Konvergenzordnung zuweisen.
- Das **Newton-Verfahren** (siehe Satz 4.19 und Bemerkung 4.9) ist wegen (4.39) **quadratisch konvergent** bzw. hat die **Konvergenzordnung 2**.
- Das **Sekantenverfahren** hat nach (4.21) in Bemerkung 4.12 die **Konvergenzordnung** $r = (\sqrt{5} + 1)/2 \doteq 1{,}62$.

KAPITEL 5

Numerische Eigenwertberechnung

In diesem Kapitel interessieren, wir uns dafür, wie man die Eigenwerte und die zugehörigen Eigenvektoren einer quadratischen Matrix $\mathbb{R}^{n\times n}$ numerisch bestimmt.

Zur Erinnerung: Ein Wert $\lambda \in \mathbb{C}$ ist ein **Eigenwert** einer Matrix $\mathbf{A} \in \mathbb{R}^{n\times n}$, wenn es einen Vektor $\mathbf{x} \in \mathbb{C}^n \setminus \{\mathbf{0}\}$ gibt, so dass gilt

$$\mathbf{A}\,\mathbf{x} = \lambda\,\mathbf{x}. \tag{5.1}$$

Man nennt einen Vektor $\mathbf{x}$, der (5.1) erfüllt, dann einen **Eigenvektor zum Eigenwert** λ. An der Umformung

$$\mathbf{A}\,\mathbf{x} = \lambda\,\mathbf{x} \quad \Longleftrightarrow \quad \mathbf{A}\,\mathbf{x} - \lambda\,\mathbf{x} = \mathbf{0} \quad \Longleftrightarrow \quad \big(\mathbf{A} - \lambda\,\mathbf{E}_n\big)\,\mathbf{x} = \mathbf{0} \tag{5.2}$$

sehen wir direkt, dass das homogene lineare Gleichungssystem $\big(\mathbf{A}-\lambda\,\mathbf{E}_n\big)\,\mathbf{x} = \mathbf{0}$ mit der quadratischen Matrix $\mathbf{A}-\lambda\,\mathbf{E}_n$ hat nur dann weitere Lösungen außer dem Nullvektor hat, wenn die Matrix $\mathbf{A}-\lambda\,\mathbf{E}_n$ **singulär** (also nicht regulär, d.h. nicht invertierbar) ist. Also ist $\lambda \in \mathbb{C}$ genau dann ein Eigenwert von $\mathbf{A}$, wenn gilt $\det(\mathbf{A} - \lambda\,\mathbf{E}_n) = 0$. Daher bestimmt man die Eigenwerte von $\mathbf{A}$ als Nullstellen des **charakteristischen Polynoms**

$$p_{\mathbf{A}}(\lambda) = \det\big(\mathbf{A} - \lambda\,\mathbf{E}_n\big). \tag{5.3}$$

Hat man die Eigenwerte bestimmt, so findet man alle Eigenvektoren zum Eigenwert λ, in dem man (nach (5.2)) das homogene lineare Gleichungssystem

$$\big(\mathbf{A} - \lambda\,\mathbf{E}_n\big)\,\mathbf{x} = \mathbf{0} \tag{5.4}$$

löst. Die Lösungsmenge des homogenen linearen Gleichungssystems (5.4) ist ein ein Untervektorraum von $\mathbb{C}^n$, der so sogenannte **Eigenraum zum Eigenwert** λ

$$E_{\mathbf{A}}(\lambda) = \big\{\mathbf{x} \in \mathbb{C}^n \;:\; \big(\mathbf{A} - \lambda\,\mathbf{E}_n\big)\,\mathbf{x} = \mathbf{0}\big\}.$$

$E_{\mathbf{A}}(\lambda)$ enthält alle Eigenvektoren zum Eigenwert λ, sowie den Nullvektor.

Für große Matrixen $\mathbf{A} \in \mathbb{R}^{n \times n}$ ist es nicht praktikabel die Eigenwerte als Nullstellen des charakteristischen Polynoms (5.3) zu bestimmen, sondern man nutzt geeignete numerische Verfahren zur Bestimmung der Eigenwerte und Eigenvektoren. Einige solche Verfahren besprechen wir in diesem Kapitel.

5.1 Grundlegende Techniken

Der erste Satz liefert eine grundlegende Information über die Lage der Eigenwerte.

Satz 5.1. (Gerschgorin-Kreise)
Die Eigenwerte einer Matrix $\mathbf{A} = [a_{j,k}] \in \mathbb{R}^{n \times n}$ *liegen alle in der Vereinigung* $\bigcup_{j=1}^{n} K_j$ *der* ***Gerschgorin-Kreise***

$$K_j := \left\{ \lambda \in \mathbb{C} \ : \ |\lambda - a_{j,j}| \leq \sum_{\substack{k=1,\\ k \neq j}}^{n} |a_{j,k}| \right\}, \qquad j = 1, 2, \ldots, n. \tag{5.5}$$

Für mathematisch Interessierte ist der Beweis angegeben.

Beweis von Satz 5.1: Für jeden Eigenwert $\lambda \in \mathbb{C}$ von $\mathbf{A} \in \mathbb{R}^{n \times n}$ können wir einen Eigenvektor $\mathbf{x} \in \mathbb{C}^n \setminus \{0\}$ finden, der $\|\mathbf{x}\|_\infty = \max_{1 \leq i \leq n} |x_i| = 1$ erfüllt. Aus $\big(\mathbf{A} - \lambda\, \mathbf{E}_n\big)\, \mathbf{x} = \mathbf{0} \iff \mathbf{A}\, \mathbf{x} - \lambda\, \mathbf{x} = \mathbf{0}$ folgt, dass gilt:

$$\left(\sum_{k=1}^{n} a_{i,k}\, x_k \right) - \lambda\, x_i = 0, \qquad i = 1, 2, \ldots, n$$

$$\iff \qquad \big(a_{i,i} - \lambda\big)\, x_i + \sum_{\substack{k=1,\\ k \neq i}}^{n} a_{i,k}\, x_k = 0, \qquad i = 1, 2, \ldots, n$$

Umsortieren liefert

$$\big(\lambda - a_{i,i}\big)\, x_i = \sum_{\substack{k=1,\\ k \neq i}}^{n} a_{i,k}\, x_k = 0, \qquad i = 1, 2, \ldots, n. \tag{5.6}$$

Sei nun $i = j$ ein Index, für den $|x_j| = \|\mathbf{x}\|_\infty = 1$ gilt. Dann folgt für $i = j$ aus (5.6) durch Anwenden des Absolutbetrags, Ausnutzen von $|x_j| = 1$ und geeignet

nach oben Abschätzen

$$|\lambda - a_{j,j}| = |\lambda - a_{j,j}| \underbrace{|x_j|}_{=1} = \left|\left(\lambda - a_{j,j}\right) x_j\right| \overset{(5.6)}{=} \left| \sum_{\substack{k=1,\\ k\neq j}}^{n} a_{j,k}\, x_k \right|$$

$$\leq \sum_{\substack{k=1,\\ k\neq j}}^{n} |a_{j,k}|\, |x_k| \leq \underbrace{\left(\max_{1\leq i\leq n} |x_i|\right)}_{= \|\mathbf{x}\|_\infty = 1} \sum_{\substack{k=1,\\ k\neq j}}^{n} |a_{j,k}| \leq \sum_{\substack{k=1,\\ k\neq j}}^{n} |a_{j,k}|.$$

Für den Eigenwert $\lambda \in \mathbb{C}$ gilt also

$$|\lambda - a_{j,j}| \leq \sum_{\substack{k=1,\\ k\neq j}}^{n} |a_{j,k}|,$$

wobei j ein Index mit $|x_j| = \|\mathbf{x}\|_\infty$ für den Eigenvektor $\mathbf{x}$ zum Eigenwert λ ist. Also wissen wir, dass der Eigenwert λ in dem durch (5.5) definierten Gerschgorin-Kreis K_j liegt, wobei j ein Index mit $|x_j| = \|\mathbf{x}\|_\infty = 1$ für den Eigenvektor $\mathbf{x}$ zum Eigenwert λ ist. Da der Eigenwert λ in den obigen Überlegungen beliebig war, folgt, dass alle Eigenwerte in der Vereinigung der Gerschgorin-Kreise K_j, $j = 1, 2, \ldots, n$, enthalten sind. □

Der Rayleigh-Quotient ermöglicht uns, einen Eigenwert angenähert zu bestimmen, wenn wir iterativ eine (hinreichend gute) Näherung eines zugehörigen Eigenvektors bestimmen können.

Definition 5.2. (Rayleigh-Quotient)
*Der **Rayleigh-Quotient** eines Vektors $\mathbf{x} \in \mathbb{R}^n \setminus \{\mathbf{0}\}$ bzgl. einer reellen Matrix $\mathbf{A} \in \mathbb{R}^{n\times n}$ ist der Skalar*

$$R(\mathbf{x}) := \frac{\mathbf{x}^T \mathbf{A}\, \mathbf{x}}{\mathbf{x}^T \mathbf{x}}.$$

Wir beobachten: Ist $\mathbf{x} \in \mathbb{R}^n \setminus \{\mathbf{0}\}$ ein Eigenvektor zum Eigenwert $\lambda \in \mathbb{R}$ von $\mathbf{A} \in \mathbb{R}^{n\times n}$, also $\mathbf{A}\,\mathbf{x} = \lambda\,\mathbf{x}$, dann gilt

$$R(\mathbf{x}) = \frac{\mathbf{x}^T \mathbf{A}\, \mathbf{x}}{\mathbf{x}^T \mathbf{x}} = \frac{\lambda\, \mathbf{x}^T \mathbf{x}}{\mathbf{x}^T \mathbf{x}} = \lambda. \tag{5.7}$$

Wir wollen uns in der nachfolgenden Diskussion auf **reelle symmetrische** Matrizen $\mathbf{A} \in \mathbb{R}^{n\times n}$ beschränken. Es gelte also $\mathbf{A}^T = \mathbf{A}$. Solche Matrizen haben

die Eigenschaft, dass sie immer n (nicht notwendigerweise verschiedene) **reelle Eigenwerte** $\lambda_1, \lambda_2, \ldots, \lambda_n \in \mathbb{R}$ haben und dass es weiter zu diesen Eigenwerten n **reelle Eigenvektoren** $\mathbf{w}_1, \mathbf{w}_2, \ldots, \mathbf{w}_n \in \mathbb{R}^n$ gibt, die **paarweise orthogonal** und **normiert** sind. Es gilt also $\mathbf{A}\,\mathbf{w}_j = \lambda_j\,\mathbf{w}_j$ für $j = 1, 2, \ldots, n$ und

$$\mathbf{w}_j^T\,\mathbf{w}_k = 0 \quad \text{für alle } j, k = 1, 2, \ldots, n \text{ mit } j \neq k \quad (\mathbf{w}_j \text{ und } \mathbf{w}_k \text{ sind ortogonal}),$$
$$\mathbf{w}_j^T\,\mathbf{w}_j = \|\mathbf{w}_j\|_2^2 = 1 \quad \text{für alle } j = 1, 2, \ldots, n \quad (\mathbf{w}_j \text{ ist normiert}).$$

Mit dem Kronecker-Delta $\delta_{j,k}$, definiert als 1 für $j = k$ und 0 sonst, können wir diese Eigenschaften auch kompakter als

$$\mathbf{w}_j^T\,\mathbf{w}_k = \delta_{j,k} \qquad \text{für alle } j, k = 1, 2 \ldots, n \tag{5.8}$$

schreiben. Aus der paarweisen Orthogonalität (siehe (5.8)) der reellen Eigenvektoren $\mathbf{w}_1, \mathbf{w}_2, \ldots \mathbf{w}_n \in \mathbb{R}^n$ folgt insbesondere, dass diese linear unabhängig sind. Also bilden $\mathbf{w}_1, \mathbf{w}_2, \ldots \mathbf{w}_n$ eine Basis für $\mathbb{R}^n$. Man nennt eine solche Basis auch eine **Orthonormalbasis** wegen der Eigenschaft (5.8). Also kann jeder Vektor $\mathbf{x} \in \mathbb{R}^n$ eindeutig als Linearkombination

$$\mathbf{x} = \sum_{k=1}^{n} c_k\,\mathbf{w}_k, \tag{5.9}$$

dargestellt werden, und die Koeffizienten $c_1, c_2, \ldots, c_n \in \mathbb{R}$ können unter Ausnutzung von (5.8) über

$$\mathbf{w}_j^T\,\mathbf{x} = \mathbf{w}_j^T \left(\sum_{k=1}^{n} c_k\,\mathbf{w}_k \right) = \sum_{k=1}^{n} c_k \underbrace{\mathbf{w}_j^T\,\mathbf{w}_k}_{=\delta_{j,k}} = \sum_{k=1}^{n} c_k\,\delta_{j,k} = c_j \quad \Longrightarrow \quad c_j = \mathbf{w}_j^T\,\mathbf{x}$$

leicht bestimmt werden. Weiter kann man mit Hilfe von (5.8) auch leicht die Euklidische Norm von $\mathbf{x}$ mit der Darstellung (5.9) berechnen:

$$\|\mathbf{x}\|_2^2 = \mathbf{x}^T\mathbf{x} = \left(\sum_{j=1}^{n} c_j\,\mathbf{w}_j \right)^T \left(\sum_{k=1}^{n} c_k\,\mathbf{w}_k \right) = \sum_{j=1}^{n}\sum_{k=1}^{n} c_j\,c_k \underbrace{\mathbf{w}_j^T\,\mathbf{w}_k}_{=\delta_{j,k}}$$

$$= \sum_{j=1}^{n}\sum_{k=1}^{n} c_j\,c_k\,\delta_{j,k} = \sum_{k=1}^{n} c_k^2 \qquad \Longrightarrow \qquad \|\mathbf{x}\|_2 = \left(\sum_{k=1}^{n} c_k^2 \right)^{1/2}$$

Nach diesen Vorbereitungen formulieren wir das Konvergenzresultat für den Rayleigh-Quotienten.

Satz 5.3. (Konvergenz des Rayleigh-Quotienten)
*Sei $\mathbf{A} \in \mathbb{R}^{n\times n}$ eine **symmetrische Matrix**, und sei $\big(\mathbf{x}^{(j)}\big)_{j\in\mathbb{N}}$ eine Folge von Vektoren in $\mathbb{R}^n$, die gegen einen Eigenvektor $\mathbf{w}_J$ of $\mathbf{A}$ mit Eigenwert λ_J konvergiert (also $\lim_{j\to\infty} \mathbf{x}^{(j)} = \mathbf{w}_J$) und normalisiert ist (also $\|\mathbf{x}^{(j)}\|_2 = 1$ für alle $j = 1, 2, \ldots$ erfüllt). Dann gilt für den in Definition 5.2 definierten Rayleight-Quotienten*

$$\lim_{j\to\infty} R(\mathbf{x}^{(j)}) = R(\mathbf{w}_J) = \lambda_J. \tag{5.10}$$

Weiter gilt

$$\big|R(\mathbf{x}^{(j)}) - R(\mathbf{w}_J)\big| = \mathcal{O}\big(\|\mathbf{x}^{(j)} - \mathbf{w}_J\|_2^2\big), \tag{5.11}$$

*d.h. die Konvergenzordnung der Iteration $\big(R(\mathbf{x}^{(j)})\big)_{j\in\mathbb{N}}$ ist **quadratisch**.*

Beweis von (5.10) **in Satz 5.3:** Per definition ist der Rayleigh-Quotient eine stetige Funktion, und somit gilt

$$\lim_{j\to\infty} R(\mathbf{x}^{(j)}) = R\left(\lim_{j\to\infty} \mathbf{x}^{(j)}\right) = R(\mathbf{w}_J) = \frac{\mathbf{w}_J^T \overbrace{\mathbf{A}\,\mathbf{w}_J}^{=\lambda_J\,\mathbf{w}_J}}{\mathbf{w}_J^T\,\mathbf{w}_J} = \frac{\lambda_J\,\mathbf{w}_J^T\,\mathbf{w}_J}{\mathbf{w}_J^T\,\mathbf{w}_J} = \lambda_J, \tag{5.12}$$

wobei wir im ersten Schritt die Stetigkeit von R und im vorletzten Schritt $\mathbf{A}\,\mathbf{w}_J = \lambda_J\,\mathbf{w}_J$ (da $\mathbf{w}_J$ ein Eigenvektor zum Eigenwert λ_J ist) ausgenutzt haben. □

Für mathematisch Interessierte ist auch der aufwendigere Beweis von (5.11) in Satz 5.3 angegeben.

Beweis von (5.11) **in Satz 5.3:** Wie nutzen aus, dass $\mathbf{A}$ symmetrisch ist und daher n reelle (nicht notwendigerweise verschiedene) Eigenwerte $\lambda_1, \lambda_2, \ldots, \lambda_n \in \mathbb{R}$ und eine Orthonormalbasis aus n zugehörigen rellen Eigenvektoren $\mathbf{w}_1, \mathbf{w}_2, \ldots, \mathbf{w}_n$ hat. Es gilt also $\mathbf{A}\,\mathbf{w}_k = \lambda_k\,\mathbf{w}_k$ für $k = 1, 2, \ldots, n$ und die Eigenvektoren sind paarweise orthogonal und normiert (siehe (5.8)). Bzgl. dieser Orthonormalbasis hat $\mathbf{x}^{(j)}$ dann die eindeutige Darstellung

$$\mathbf{x}^{(j)} = \sum_{k=1}^{n} c_k\,\mathbf{w}_k, \tag{5.13}$$

wobei die eindeutig bestimmten Koeffizienten $c_k = c_k^{(j)}$ auch von j abhängen. Wir unterdrücken dieses in der Notation aber, damit die Formeln nicht zu unübersichtlich werden. Wegen $\mathbf{A}\,\mathbf{w}_k = \lambda_k\,\mathbf{w}_k$ für $k = 1, 2, \ldots, n$ folgt aus (5.13)

$$\mathbf{A}\,\mathbf{x}^{(j)} = \mathbf{A}\left(\sum_{k=1}^{n} c_k\,\mathbf{w}_k\right) = \sum_{k=1}^{n} c_k\,\underbrace{\mathbf{A}\,\mathbf{w}_k}_{=\lambda_k\,\mathbf{w}_k} = \sum_{k=1}^{n} c_k\,\lambda_k\,\mathbf{w}_k. \tag{5.14}$$

Mit $\mathbf{w}_j^T\,\mathbf{w}_k = 0$ für $j \neq k$, $\mathbf{w}_k\,\mathbf{w}_k^T = \|\mathbf{w}_k\|_2^2 = 1$ für alle $k = 1, 2, \ldots, n$ und $\|\mathbf{x}^{(j)}\|_2^2 = (\mathbf{x}^{(j)})^T\,\mathbf{x}^{(j)} = 1$ for all $j \in \mathbb{N}$, folgt aus (5.14) und (5.13)

$$\begin{aligned} R\big(\mathbf{x}^{(j)}\big) &= \frac{(\mathbf{x}^{(j)})^T\,\mathbf{A}\,\mathbf{x}^{(j)}}{(\mathbf{x}^{(j)})^T\,\mathbf{x}^{(j)}} = (\mathbf{x}^{(j)})^T\,\mathbf{A}\,\mathbf{x}^{(j)} = \left(\sum_{i=1}^{n} c_i\,\mathbf{w}_i\right)^T \left(\sum_{k=1}^{n} c_k\,\lambda_k\,\mathbf{w}_k\right) \\ &= \sum_{k=1}^{n}\sum_{i=1}^{n} c_i\,c_k\,\lambda_k\,\underbrace{\mathbf{w}_i^T\,\mathbf{w}_k}_{=\delta_{i,k}} = \sum_{k=1}^{n}\sum_{i=1}^{n} c_i\,c_k\,\lambda_k\,\delta_{i,k} = \sum_{k=1}^{n} \lambda_k\,c_k^2. \end{aligned}$$

Mit $R(\mathbf{w}_J) = \lambda_J$ (siehe hintere Umformungen in (5.12)), folgt daraus

$$R(\mathbf{x}^{(j)}) - R(\mathbf{w}_J) = \sum_{k=1}^{n} \lambda_k\,c_k^2 - \lambda_J = \lambda_J\,(c_J^2 - 1) + \sum_{\substack{k=1,\\ k\neq J}}^{n} \lambda_k\,c_k^2.$$

Wir schätzen $R(\mathbf{x}^{(j)}) - R(\mathbf{w}_J)$ geeignet ab:

$$\begin{aligned} \big|R(\mathbf{x}^{(j)}) - R(\mathbf{w}_J)\big| &= \left|\lambda_J\,(c_J^2 - 1) + \sum_{\substack{k=1,\\ k\neq J}}^{n} \lambda_k\,c_k^2\right| \leq |\lambda_J|\,|c_J^2 - 1| + \sum_{\substack{k=1,\\ k\neq J}}^{n} |\lambda_k|\,c_k^2 \\ &\leq \left(\max_{1\leq i\leq n} |\lambda_i|\right)\left(|c_J^2 - 1| + \sum_{\substack{k=1,\\ k\neq J}}^{n} c_k^2\right) \\ &= \left(\max_{1\leq i\leq n} |\lambda_i|\right)\left(|(2\,c_J - 2) + (c_J^2 - 2\,c_J + 1)| + \sum_{\substack{k=1,\\ k\neq J}}^{n} c_k^2\right) \\ &= \left(\max_{1\leq i\leq n} |\lambda_i|\right)\left(|2\,(c_J - 1) + (c_J - 1)^2| + \sum_{\substack{k=1,\\ k\neq J}}^{n} c_k^2\right) \\ &\leq \left(\max_{1\leq i\leq n} |\lambda_i|\right)\left(2\,|c_J - 1| + (c_J - 1)^2 + \sum_{\substack{k=1,\\ k\neq J}}^{n} c_k^2\right). \end{aligned} \tag{5.15}$$

Mit $\|\mathbf{x}^{(j)}\|_2 = 1$, $\|\mathbf{w}_J\|_2 = 1$ und (5.13) und $\mathbf{w}_i^T\,\mathbf{w}_k = 0$ für $i \neq k$ folgt einerseits

$$\big\|\mathbf{x}^{(j)} - \mathbf{w}_J\big\|_2^2 = \big(\mathbf{x}^{(j)} - \mathbf{w}_J\big)^T\,\big(\mathbf{x}^{(j)} - \mathbf{w}_J\big) = \underbrace{\|\mathbf{x}^{(j)}\|_2^2}_{=1} + \underbrace{\|\mathbf{w}_J\|_2^2}_{=1} - 2\,\mathbf{w}_J^T\,\mathbf{x}^{(j)}$$

$$= 2 - 2\,\mathbf{w}_J^T \left(\sum_{k=1}^{n} c_k\,\mathbf{w}_k \right) = 2 - 2 \sum_{k=1}^{n} c_k \underbrace{\mathbf{w}_J^T\,\mathbf{w}_k}_{=\delta_{k,J}} = 2 - 2\,c_J = 2\,(1 - c_J). \quad (5.16)$$

Andererseits folgt aus (5.13) auch

$$\left\|\mathbf{x}^{(j)} - \mathbf{w}_J\right\|_2^2 = \left\| \sum_{k=1}^{n} c_k\,\mathbf{w}_k - \mathbf{w}_J \right\|_2^2 = \left\| (c_J - 1)\,\mathbf{w}_J + \sum_{\substack{k=1,\\ k\neq J}}^{n} c_k\,\mathbf{w}_k \right\|_2^2$$

$$= \left((c_J - 1)\,\mathbf{w}_J + \sum_{\substack{i=1,\\ i\neq J}}^{n} c_i\,\mathbf{w}_i \right)^T \left((c_J - 1)\,\mathbf{w}_J + \sum_{\substack{k=1,\\ k\neq J}}^{n} c_k\,\mathbf{w}_k \right)$$

$$= (c_J - 1)^2 \underbrace{\mathbf{w}_J^T\,\mathbf{w}_J}_{=1} + 2\,(c_J - 1) \sum_{\substack{k=1,\\ k\neq J}}^{n} c_k \underbrace{\mathbf{w}_J^T\,\mathbf{w}_k}_{=0} + \sum_{\substack{i=1,\\ i\neq J}}^{n} \sum_{\substack{k=1,\\ k\neq J}}^{n} c_i\,c_k \underbrace{\mathbf{w}_i^T\,\mathbf{w}_k}_{=\delta_{i,k}}$$

$$= (c_J - 1)^2 + \sum_{\substack{i=1,\\ i\neq J}}^{n} \sum_{\substack{k=1,\\ k\neq J}}^{n} c_i\,c_k\,\delta_{i,k} = (c_J - 1)^2 + \sum_{\substack{k=1,\\ k\neq J}}^{n} c_k^2. \quad (5.17)$$

Aus (5.16) und (5.17) erhalten wir also die folgenden beiden Darstellungen von $\left\|\mathbf{x}^{(j)} - \mathbf{w}_J\right\|_2^2$:

$$\left\|\mathbf{x}^{(j)} - \mathbf{w}_J\right\|_2^2 = 2\,(1 - c_J) \quad \text{und} \quad \left\|\mathbf{x}^{(j)} - \mathbf{w}_J\right\|_2^2 = (c_J - 1)^2 + \sum_{\substack{k=1,\\ k\neq J}}^{n} c_k^2 \quad (5.18)$$

Ersetzt man in der letzten Zeile von (5.15) die entsprechenden Terme mit Hilfe von (5.18) jeweils durch $\left\|\mathbf{x}^{(j)} - \mathbf{w}_J\right\|_2^2$ so erhält man

$$\left|R(\mathbf{x}^{(j)}) - R(\mathbf{w}_J)\right| \leq \left(\max_{1\leq i\leq n} |\lambda_i| \right) \left(\left\|\mathbf{x}^{(j)} - \mathbf{w}_J\right\|_2^2 + \left\|\mathbf{x}^{(j)} - \mathbf{w}_J\right\|_2^2 \right)$$

$$= 2 \left(\max_{1\leq i\leq n} |\lambda_i| \right) \left\|\mathbf{x}^{(j)} - \mathbf{w}_J\right\|_2^2,$$

womit die quadratische Konvergenz (5.11) bewiesen ist. □

Betrachten wir ein Beispiel für die Anwendung von Satz 5.3.

Beispiel 5.4. (Konvergenz des Rayleigh-Quotienten)
In Beispiel 3.16 haben wir Eigenwerte der symmetrischen Matrix

$$\mathbf{A} = \begin{bmatrix} \frac{3}{2} & 0 & \frac{1}{2} \\ 0 & 3 & 0 \\ \frac{1}{2} & 0 & \frac{3}{2} \end{bmatrix}$$

berechnet und fanden, dass diese $\lambda_1 = 1$, $\lambda_2 = 2$ und $\lambda_3 = 3$ sind. Wir berechnen zunächst die zugehörigen Eigenräume:

Um den Eigenraum zum Eigenwert $\lambda_1 = 1$ zu berechnen, lösen wir das lineare Gleichungssystem $(\mathbf{A} - 1\,\mathbf{E}_3)\,\mathbf{x} = \mathbf{0}$:

$$\left[\begin{array}{ccc|c} \frac{1}{2} & 0 & \frac{1}{2} & 0 \\ 0 & 2 & 0 & 0 \\ \frac{1}{2} & 0 & \frac{1}{2} & 0 \end{array}\right] \overset{\substack{Z_3 \to Z_3 - Z_1 \\ Z_2 \to \frac{1}{2} Z_2 \\ \text{dann: } Z_1 \to 2\, Z_1}}{\Longleftrightarrow} \left[\begin{array}{ccc|c} 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{array}\right] \quad \Longrightarrow \quad \begin{cases} x_1 = -x_3, \\ x_2 = 0, \\ x_3 \in \mathbb{R} \text{ beliebig} \end{cases}$$

$$\Longrightarrow \quad \text{Eigenraum zu } \lambda_1 = 1\text{:} \quad E_{\mathbf{A}}(1) = \left\{ \begin{bmatrix} -\alpha \\ 0 \\ \alpha \end{bmatrix} \;:\; \alpha \in \mathbb{R} \right\} = \text{Span}\left(\begin{bmatrix} -1 \\ 0 \\ 1 \end{bmatrix} \right)$$

Um den Eigenraum zum Eigenwert $\lambda_2 = 2$ zu berechnen, lösen wir das lineare Gleichungssystem $(\mathbf{A} - 2\,\mathbf{E}_3)\,\mathbf{x} = \mathbf{0}$:

$$\left[\begin{array}{ccc|c} -\frac{1}{2} & 0 & \frac{1}{2} & 0 \\ 0 & 1 & 0 & 0 \\ \frac{1}{2} & 0 & -\frac{1}{2} & 0 \end{array}\right] \overset{\substack{Z_3 \to Z_3 + Z_1 \\ \text{dann: } Z_1 \to -2\, Z_1}}{\Longleftrightarrow} \left[\begin{array}{ccc|c} 1 & 0 & -1 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{array}\right] \quad \Longrightarrow \quad \begin{cases} x_1 = x_3, \\ x_2 = 0, \\ x_3 \in \mathbb{R} \text{ beliebig} \end{cases}$$

$$\Longrightarrow \quad \text{Eigenraum zu } \lambda_2 = 2\text{:} \quad E_{\mathbf{A}}(2) = \left\{ \begin{bmatrix} \alpha \\ 0 \\ \alpha \end{bmatrix} \;:\; \alpha \in \mathbb{R} \right\} = \text{Span}\left(\begin{bmatrix} 1 \\ 0 \\ 1 \end{bmatrix} \right)$$

Um den Eigenraum zum Eigenwert $\lambda_3 = 3$ zu berechnen, lösen wir das lineare Gleichungssystem $(\mathbf{A} - 3\,\mathbf{E}_3)\,\mathbf{x} = \mathbf{0}$:

$$\left[\begin{array}{ccc|c} -\frac{3}{2} & 0 & \frac{1}{2} & 0 \\ 0 & 0 & 0 & 0 \\ \frac{1}{2} & 0 & -\frac{3}{2} & 0 \end{array}\right] \overset{\substack{Z_3 \to Z_3 + \frac{1}{3} Z_1 \\ \text{dann: } Z_1 \to -\frac{2}{3} Z_1}}{\Longleftrightarrow} \left[\begin{array}{ccc|c} 1 & 0 & -\frac{1}{3} & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & -\frac{4}{3} & 0 \end{array}\right]$$

$$\overset{\substack{Z_1 \to Z_1 - \frac{1}{4} Z_3 \\ \text{dann: } Z_3 \to -\frac{3}{4} Z_3}}{\Longleftrightarrow} \left[\begin{array}{ccc|c} 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 \end{array}\right] \quad \Longrightarrow \quad \begin{cases} x_1 = 0, \\ x_2 \in \mathbb{R} \text{ beliebig}, \\ x_3 = 0 \end{cases}$$

$$\Longrightarrow \quad \text{Eigenraum zu } \lambda_3 = 3\text{:} \quad E_{\mathbf{A}}(3) = \left\{ \begin{bmatrix} 0 \\ \alpha \\ 0 \end{bmatrix} \;:\; \alpha \in \mathbb{R} \right\} = \text{Span}\left(\begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix} \right)$$

Wir erhalten hier also die folgende Orthonormalbasis von $\mathbb{R}^n$ aus Eigenvektoren:

$$B = \big(\mathbf{w}_1, \mathbf{w}_2, \mathbf{w}_3\big) \qquad \text{mit} \qquad \mathbf{w}_1 = \begin{bmatrix} -\frac{1}{\sqrt{2}} \\ 0 \\ \frac{1}{\sqrt{2}} \end{bmatrix}, \quad \mathbf{w}_2 = \begin{bmatrix} \frac{1}{\sqrt{2}} \\ 0 \\ \frac{1}{\sqrt{2}} \end{bmatrix}, \quad \mathbf{w}_3 = \begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix}.$$

Gegeben sei nun die Folge

$$\big(\mathbf{x}^{(j)}\big)_{j\in\mathbb{N}} \qquad \text{mit} \qquad \mathbf{x}^{(j)} = \frac{1}{\sqrt{2}\left(1+\frac{1}{j^2}\right)^{1/2}} \begin{bmatrix} 1+\frac{1}{j} \\ 0 \\ 1-\frac{1}{j} \end{bmatrix}.$$

Dann gelten

$$\lim_{j\to\infty} \mathbf{x}^{(j)} = \lim_{j\to\infty} \mathbf{x}^{(j)} \frac{1}{\sqrt{2}\left(1+\frac{1}{j^2}\right)^{1/2}} \begin{bmatrix} 1+\frac{1}{j} \\ 0 \\ 1-\frac{1}{j} \end{bmatrix} = \frac{1}{\sqrt{2}} \begin{bmatrix} 1 \\ 0 \\ 1 \end{bmatrix} = \mathbf{w}_2,$$

$$\begin{aligned} \|\mathbf{x}^{(j)}\|_2^2 &= \left(\frac{1}{\sqrt{2}\left(1+\frac{1}{j^2}\right)^{1/2}} \right)^2 \left[\left(1+\frac{1}{j}\right)^2 + \left(1-\frac{1}{j}\right)^2 \right] \\ &= \frac{1}{2\left(1+\frac{1}{j^2}\right)} \left[\left(1+\frac{2}{j}+\frac{1}{j^2}\right) + \left(1-\frac{2}{j}+\frac{1}{j^2}\right) \right] \\ &= \frac{1}{2\left(1+\frac{1}{j^2}\right)} \left[2+\frac{2}{j^2} \right] = \frac{1}{2\left(1+\frac{1}{j^2}\right)} 2 \left[1+\frac{1}{j^2} \right] = 1, \end{aligned}$$

d.h. die Voraussetzungen in Satz 5.3 sind erfüllt. Nach Satz 5.3 gilt

$$\lim_{j\to\infty} R\big(\mathbf{x}^{(j)}\big) = \lambda_2 = 2,$$

weil $\lim_{j\to\infty} \mathbf{x}^{(j)} = \mathbf{w}_2$ ein Eigenvektor zu $\lambda_2 = 2$ ist.

Wir berechnen nun den Grenzwert $\lim_{j\to\infty} R\big(\mathbf{x}^{(j)}\big)$ direkt:

$$R\big(\mathbf{x}^{(j)}\big) \overset{\|\mathbf{x}^{(j)}\|_2=1}{\overset{\downarrow}{=}} \big(\mathbf{x}^{(j)}\big)^T \mathbf{A}\, \mathbf{x}^{(j)} = \frac{1}{2\left(1+\frac{1}{j^2}\right)} \begin{bmatrix} 1+\frac{1}{j} \\ 0 \\ 1-\frac{1}{j} \end{bmatrix}^T \begin{bmatrix} \frac{3}{2} & 0 & \frac{1}{2} \\ 0 & 3 & 0 \\ \frac{1}{2} & 0 & \frac{3}{2} \end{bmatrix} \begin{bmatrix} 1+\frac{1}{j} \\ 0 \\ 1-\frac{1}{j} \end{bmatrix}$$

$$
\begin{aligned}
&= \frac{1}{2\left(1+\frac{1}{j^2}\right)} \left[1+\tfrac{1}{j}\,;\ 0\,;\ 1-\tfrac{1}{j}\right] \begin{bmatrix} 2+\frac{1}{j} \\ 0 \\ 2-\frac{1}{j} \end{bmatrix} \\
&= \frac{1}{2\left(1+\frac{1}{j^2}\right)} \left[\left(1+\tfrac{1}{j}\right)\left(2+\tfrac{1}{j}\right) + \left(1-\tfrac{1}{j}\right)\left(2-\tfrac{1}{j}\right)\right] \\
&= \frac{1}{2\left(1+\frac{1}{j^2}\right)} \left[\left(2+\tfrac{3}{j}+\tfrac{1}{j^2}\right) + \left(2-\tfrac{3}{j}+\tfrac{1}{j^2}\right)\right] \\
&= \frac{1}{2\left(1+\frac{1}{j^2}\right)} \left[2\left(2+\tfrac{1}{j^2}\right)\right] = \frac{2+\frac{1}{j^2}}{1+\frac{1}{j^2}}
\end{aligned}
$$

Also finden wir

$$\lim_{j\to\infty} R\big(\mathbf{x}^{(j)}\big) = \lim_{j\to\infty} \frac{2+\frac{1}{j^2}}{1+\frac{1}{j^2}} = \frac{2}{1} = 2 = \lambda_2,$$

wie erwartet. ♠

5.2 Von-Mises-Vektoriteration (Potenzmethode)

In diesem Teilkapitel betrachten wir nur **reelle** Matrizen $\mathbf{A} \in \mathbb{R}^{n\times n}$, aber wir nehmen nicht mehr an, dass $\mathbf{A}$ eine symmetrische Matrix ist. Statt dessen nehmen wir aber an, dass $\mathbf{A}$ einen **dominanten Eigenwert** (also betraglich größten Eigenwert) hat, d.h. dass es einen Eigenwert gibt, den wir λ_1 nennen, so dass

$$|\lambda_1| > |\lambda_2| \geq |\lambda_3| \geq \ldots \geq |\lambda_n|. \tag{5.19}$$

Für solche Matrizen wollen wir im Folgenden den dominanten Eigenwert λ_1 und einen zugehörigen Eigenvektor bestimmen.

Weiter setzen wir in diesem Teilkapitel voraus, dass **A nur reelle Eigenwerte** $\lambda_1, \lambda_2, \ldots, \lambda_n \in \mathbb{R}$ hat und dass es **n linear unabhängige zugehörige reelle Eigenvektoren** $\mathbf{w}_1, \mathbf{w}_2, \ldots, \mathbf{w}_n \in \mathbb{R}^n$ gibt. Es gelte also $\mathbf{A}\,\mathbf{w}_j = \lambda_j\,\mathbf{w}_j$ für $j = 1, 2, \ldots, n$. Weiter sollen die Eigenvektoren $\mathbf{w}_1, \mathbf{w}_2, \ldots, \mathbf{w}_n$ **normiert** sein, d.h. $\|\mathbf{w}_1\|_2 = \|\mathbf{w}_2\|_2 = \ldots = \|\mathbf{w}_n\|_2 = 1$. (Beachten Sie, dass dieses zusätzliche Annahmen sind, die nicht automatisch erfüllt sind: Ein reelle Matrix $\mathbf{A} \in \mathbb{R}^{n\times n}$ hat mit Vielfachheit gezählt nicht notwendigerweise n reelle Eigenwerte, sondern es können unter den n Eigenwerten auch komplexe, nicht-reelle Eigenwerte auftreten. Zudem kann es passieren, dass die Dimension eines Eigenraums zu einem

Eigenwert kleiner ist als die (algebraische) Vielfachheit dieses Eigenwerts. Tritt letztere Situation auf, so gibt es keine Möglichkeit, n linear unabhängige Eigenvektoren zu finden.) Beachten Sie auch, dass wir nicht annehmen dürfen, dass die Eigenvektoren $\mathbf{w}_1, \mathbf{w}_2, \ldots, \mathbf{w}_n \in \mathbb{R}^n$ paarweise orthogonal sind, da $\mathbf{A}$ nicht als symmetrisch vorausgesetzt wurde. Da die n Eigenvektoren $\mathbf{w}_1, \mathbf{w}_2, \ldots, \mathbf{w}_n \in \mathbb{R}^n$ aber linear unabhängig sind, bilden Sie eine Basis von $\mathbb{R}^n$, und jedes $\mathbf{x} \in \mathbb{R}^n$ kann eindeutig als Linearkombination

$$\mathbf{x} = \sum_{j=1}^{n} c_j \, \mathbf{w}_j. \tag{5.20}$$

mit eindeutig bestimmten Koeffizienten $c_1, c_2, \ldots, c_n \in \mathbb{R}$ dargestellt werden. Wenden wir $\mathbf{A}$ auf $\mathbf{x}$, gegeben durch (5.20), wiederholt an, so folgt mit $\mathbf{A}\,\mathbf{w}_j = \lambda_j \, \mathbf{w}_j$ und damit $\mathbf{A}^m \, \mathbf{w}_j = \lambda_j^m \, \mathbf{w}_j$ für $j = 1, 2, \ldots, n$ und $m \in \mathbb{N}$

$$\begin{aligned} \mathbf{A}^m \, \mathbf{x} &= \mathbf{A}^m \left(\sum_{j=1}^{n} c_j \, \mathbf{w}_j \right) = \sum_{j=1}^{n} c_j \, \mathbf{A}^m \, \mathbf{w}_j = \sum_{j=1}^{n} c_j \, \lambda_j^m \, \mathbf{w}_j \\ &= \lambda_1^m \left(c_1 \, \mathbf{w}_1 + \sum_{j=2}^{n} c_j \left(\frac{\lambda_j}{\lambda_1} \right)^m \mathbf{w}_j \right) =: \lambda_1^m \big(c_1 \, \mathbf{w}_1 + \mathbf{R}_m \big), \end{aligned} \tag{5.21}$$

mit dem Restglied

$$\mathbf{R}_m := \sum_{j=2}^{n} c_j \left(\frac{\lambda_j}{\lambda_1} \right)^m \mathbf{w}_j. \tag{5.22}$$

Die Vektorfolge $\big(\mathbf{R}_m\big)_{m \in \mathbb{N}}$ strebt für $m \to \infty$ gegen den Nullvektor $\mathbf{0}$, denn mit $\|\mathbf{w}_j\|_2 = 1$ für $j = 1, 2 \ldots, n$, und $|\lambda_j / \lambda_1| < 1$ für $j = 2, \ldots, n$ (wegen (5.19)) folgt mit der Dreiecksungleichung

$$\|\mathbf{R}_m\|_2 = \left\| \sum_{j=2}^{n} c_j \left(\frac{\lambda_j}{\lambda_1} \right)^m \mathbf{w}_j \right\|_2 \leq \sum_{j=2}^{n} |c_j| \left| \frac{\lambda_j}{\lambda_1} \right|^m \underbrace{\|\mathbf{w}_j\|_2}_{=1} = \sum_{j=2}^{n} |c_j| \left| \frac{\lambda_j}{\lambda_1} \right|^m \overset{m \to \infty}{\longrightarrow} 0,$$

wobei wir genutzt haben, dass $\left(\left| \frac{\lambda_j}{\lambda_1} \right|^m \right)_{m \in \mathbb{N}}$ als geometrische Folge mit $\left| \frac{\lambda_j}{\lambda_1} \right| < 1$ für jedes $j = 2, 3, \ldots, n$ gegen null strebt. Falls $c_1 \neq 0$ ist, folgt mit $\lim_{m \to \infty} \mathbf{R}_m = 0$ aus (5.21)

$$\lim_{m \to \infty} \frac{1}{\lambda_1^m} \, \mathbf{A}^m \, \mathbf{x} = \lim_{m \to \infty} \big(c_1 \, \mathbf{w}_1 + \mathbf{R}_m \big) = c_1 \, \mathbf{w}_1 + \underbrace{\lim_{m \to \infty} \mathbf{R}_m}_{= \mathbf{0}} = c_1 \, \mathbf{w}_1. \tag{5.23}$$

Dabei ist $c_1 \, \mathbf{w}_1 \neq \mathbf{0}$ (als Vielfaches des Eigenvektors $\mathbf{w}_1$) ebenfalls ein Eigenvektor von $\mathbf{A}$ zum Eigenwert λ_1. Da wir bis jetzt aber λ_1 nicht kennen und somit die Folge

$\big((\mathbf{A}^m\,\mathbf{x})/\lambda_1^m\big)_{m\in\mathbb{N}}$ nicht berechnen können, ist die Grenzwertbeziehung (5.23) zur Bestimmung eines Eigenvektors zu λ_1 zunächst nur von begrenztem Wert. Ein zusätzliches Problem liegt darin, dass die Norm von $\mathbf{A}^m\,\mathbf{x}$ gegen null strebt, falls $|\lambda_1|<1$ ist und gegen ∞ strebt, falls $|\lambda_1|>1$ ist. Beide Problem kann man aber durch geeignetes **Normieren** beheben:

Betrachten wir beispielsweise die Euklidische Norm von $\mathbf{A}^m\,\mathbf{x}$. Dann folgt aus (5.21) mit $\|\mathbf{w}_j\|_2=1$ für $j=1,2,\ldots,n$,

$$\begin{aligned}\|\mathbf{A}^m\mathbf{x}\|_2^2 &= \big\|\lambda_1^m\,\big(c_1\,\mathbf{w}_1+\mathbf{R}_m\big)\big\|_2^2 = \lambda_1^{2m}\big(c_1\,\mathbf{w}_1+\mathbf{R}_m\big)^T\,\big(c_1\,\mathbf{w}_1+\mathbf{R}_m\big)\\ &= \lambda_1^{2m}\big(|c_1|^2\,\underbrace{\mathbf{w}_1^T\,\mathbf{w}_1}_{=1}+2\,c_1\,\mathbf{w}_1^T\,\mathbf{R}_m+\underbrace{\mathbf{R}_m^T\,\mathbf{R}_m}_{=\|\mathbf{R}_m\|_2^2}\big) = \lambda_1^{2m}\big(|c_1|^2+r_m\big),\end{aligned}\tag{5.24}$$

wobei $r_m\in\mathbb{R}$ durch

$$r_m := 2\,c_1\,\mathbf{w}_1^T\,\mathbf{R}_m+\|\mathbf{R}_m\|_2^2\tag{5.25}$$

definiert ist. Aus $\lim_{m\to\infty}\mathbf{R}_m=\mathbf{0}$ folgt $\lim_{m\to\infty}r_m=0$. Daher folgt aus (5.24)

$$\|\mathbf{A}^m\,\mathbf{x}\|_2=|\lambda_1|^m\,\sqrt{|c_1|^2+r_m},\qquad\text{mit}\quad\lim_{m\to\infty}r_m=0.\tag{5.26}$$

Aus (5.26) folgt nun mit $\lim_{m\to\infty}\sqrt{|c_1|^2+r_m}=\sqrt{|c_1|^2+0}=|c_1|$

$$\frac{\|\mathbf{A}^{m+1}\,\mathbf{x}\|_2}{\|\mathbf{A}^m\,\mathbf{x}\|_2}=\frac{|\lambda_1|^{m+1}\,\sqrt{|c_1|^2+r_{m+1}}}{|\lambda_1|^m\,\sqrt{|c_1|^2+r_m}}=|\lambda_1|\,\frac{\sqrt{|c_1|^2+r_{m+1}}}{\sqrt{|c_1|^2+r_m}}\overset{m\to\infty}{\longrightarrow}|\lambda_1|\,\frac{|c_1|}{|c_1|}=|\lambda_1|,\tag{5.27}$$

was uns bis auf das Vorzeichen den dominanten Eigenwert λ_1 liefert. Um das Vorzeichen von λ_1 und einen zugehörigen Eigenvektor zu finden, verfeinern wir die Methode wie folgt:

Verfahren 5.5. (Von-Mises-Vektoriteration/Potenzmethode)
*Die reelle Matrix $\mathbf{A}\in\mathbb{R}^{n\times n}$ habe (mit Vielfachheit gezählt) n reelle Eigenwerte $\lambda_1,\lambda_2,\ldots,\lambda_n\in\mathbb{R}$ und ein zugehöriges System von n linear unabhängigen normierten reellen Eigenvektoren $\mathbf{w}_1,\mathbf{w}_2,\ldots,\mathbf{w}_n\in\mathbb{R}^n$, die dann eine Basis von $\mathbb{R}^n$ bilden. ($\mathbf{w}_1,\mathbf{w}_2,\ldots,\mathbf{w}_n$ seien also linear unabhängig, und erfüllen $\|\mathbf{w}_j\|_2=1$ und $A\,\mathbf{w}_j=\lambda_j\,\mathbf{w}_j$ für alle $j=1,2,\ldots,n$.) Die Matrix $\mathbf{A}$ haben einen **dominanten Eigenwert** λ_1, d.h. es gilt $|\lambda_1|>|\lambda_j|$ für alle $j=2,3,\ldots,n$. Die **Von-Mises-Vektoriteration** (oder **Potenzmethode**) ist das folgende Verfahren:*

Initialisierung: Wähle $\mathbf{x}^{(0)}=\sum_{j=1}^{n}c_j\,\mathbf{w}_j$ with $c_1\neq0$, und setze $\mathbf{y}^{(0)}:=\dfrac{\mathbf{x}^{(0)}}{\|\mathbf{x}^{(0)}\|_2}$.

Für $m=1,2,\ldots$ berechnen wir nun

(1) $\mathbf{x}^{(m)} := \mathbf{A}\,\mathbf{y}^{(m-1)}$,

(2) $\mathbf{y}^{(m)} := \sigma_m \dfrac{\mathbf{x}^{(m)}}{\|\mathbf{x}^{(m)}\|_2}$, *wobei* $\sigma_m \in \{-1, 1\}$ *so gewählt wird, dass gilt* $\mathbf{y}^{(m)^T}\,\mathbf{y}^{(m-1)} \geq 0$.

Die Wahl von σ_m bedeutet, dass der Winkel zwischen $\mathbf{y}^{(m-1)}$ und $\mathbf{y}^{(m)}$ in $[0; \pi/2]$ liegt, d.h. dass wir ein „Umklappen“ vermeiden, wenn wir von $\mathbf{y}^{(m-1)}$ zu $\mathbf{y}^{(m)}$ übergehen. Man kann das Vorzeichen σ_m mittels

$$\sigma_m := \operatorname{sgn}\left((\mathbf{y}^{(m-1)})^T\,\mathbf{x}^{(m)}\right) = \operatorname{sgn}\left((\mathbf{x}^{(m)})^T\,\mathbf{y}^{(m-1)}\right)$$

berechnen, wie man sich leicht durch eine genauere Inspektion von Schritt (2) in Verfahren 5.5 überlegt.

Die Bedingung $c_1 \neq 0$ ist meist erfüllt. Dieses passiert (bei großen Matrizen) schon durch unvermeidliche numerische Rundungsfehler. Insofern stellt die Annahme $c_1 \neq 0$ keine wirkliche Einschränkung dar.

Der nächste Satz liefert uns wichtige Informationen über die Konvergenz der Von-Mises-Vektoriteration.

Satz 5.6. (Konvergenz der Von-Mises-Vektoriteration)
Seien die Voraussetzungen und die Notation wie in Verfahren 5.5. Dann haben die Iterierten der ***Von-Mises-Vektoriteration*** *(Verfahren 5.5) die folgenden Eigenschaften*

(1) $\|\mathbf{x}^{(m)}\|_2 \to |\lambda_1|$ *für* $m \to \infty$,

(2) $\mathbf{y}^{(m)}$ *konvergiert gegen einen normierten Eigenvektor von* $\mathbf{A}$ *zum dominanten Eigenwert* λ_1

(3) $\sigma_m \to \operatorname{sgn}(\lambda_1)$ *für* $m \to \infty$, *d.h.* $\sigma_m = \operatorname{sgn}(\lambda_1)$ *für großes* m.

Die Euklidische Norm kann bei der Normalisierung in Verfahren 5.5 durch eine andere Norm ersetzt werden, ohne sich an dem Konvergenzverhalten etwas ändert. Oft wird die ∞-Norm verwendet, weil diese billiger zu berechnen ist.

Betrachten wir zunächst ein Beispiel.

Beispiel 5.7. (Von-Mises-Vektoriteration/Potenzmethode)

Wir betrachten die reelle 3×3-Matrix $\mathbf{A} = \begin{bmatrix} 0 & -2 & 2 \\ -2 & -3 & 2 \\ -3 & -6 & 5 \end{bmatrix}$.

Diese hat die reellen Eigenwerte $\lambda_1 = 2$, $\lambda_2 = 1$, and $\lambda_3 = -1$. Da wir drei verschiedene Eigenwerte gefunden haben und jeder damit einen Eigenraum der Dimension 1 hat, gibt es auch ein Basis aus reellen Eigenvektoren. Der dominante Eigenwert ist $\lambda_1 = 2$, und ein zugehöriger normierter Eigenvektor ist durch

$$\mathbf{w}_1 = \frac{1}{\sqrt{2}} \begin{bmatrix} 1 \\ 0 \\ 1 \end{bmatrix}$$

gegeben. (Berechnen Sie $\mathbf{A}\,\mathbf{w}_1$, um sich zu überzeugen, dass dieses stimmt.) Normierte Eigenvektoren zu den Eigenwerten $\lambda_2 = 1$ und $\lambda_3 = -1$ sind wie folgt gegeben: (Wir berechnen die Eigenwerte und die Eigenräume von $\mathbf{A}$ in einer Übungsaufgabe.)

$$\mathbf{w}_2 = \frac{1}{\sqrt{5}} \begin{bmatrix} 2 \\ -1 \\ 0 \end{bmatrix} \qquad \text{bzw.} \qquad \mathbf{w}_3 = \frac{1}{\sqrt{2}} \begin{bmatrix} 0 \\ 1 \\ 1 \end{bmatrix}.$$

Wir berechnen nun die ersten zwei Schritte der Von-Mises-Vektoriteration mit dem Startvektor $\mathbf{x}^{(0)} = [1;\ 1;\ 1]^T$ per Hand. Es gilt

$$\mathbf{x}^{(0)} = \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix} = -\begin{bmatrix} 1 \\ 0 \\ 1 \end{bmatrix} + \begin{bmatrix} 2 \\ -1 \\ 0 \end{bmatrix} + 2\begin{bmatrix} 0 \\ 1 \\ 1 \end{bmatrix} = -\sqrt{2}\,\mathbf{w}_1 + \sqrt{5}\,\mathbf{w}_2 + 2\sqrt{2}\,\mathbf{w}_3,$$

so dass $c_1 = -\sqrt{2} \neq 0$ erfüllt ist und wir einen zulässigen Startvektor haben.

Wir berechnen vorab: $\mathbf{y}^{(0)} = \dfrac{\mathbf{x}^{(0)}}{\|\mathbf{x}^{(0)}\|_2} = \dfrac{1}{\sqrt{3}} \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix}$

Schritt 1: Wir berechnen

$$\mathbf{x}^{(1)} = \mathbf{A}\,\mathbf{y}^{(0)} = \frac{1}{\sqrt{3}} \begin{bmatrix} 0 & -2 & 2 \\ -2 & -3 & 2 \\ -3 & -6 & 5 \end{bmatrix} \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix} = \frac{1}{\sqrt{3}} \begin{bmatrix} 0 \\ -3 \\ -4 \end{bmatrix}.$$

Wegen

$$\sigma_1 = \operatorname{sgn}\left((\mathbf{y}^{(0)})^T\,\mathbf{x}^{(1)}\right) = \operatorname{sgn}\left(\frac{1}{\sqrt{3}} \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix}^T \frac{1}{\sqrt{3}} \begin{bmatrix} 0 \\ -3 \\ -4 \end{bmatrix}\right) = \operatorname{sgn}\left(\frac{-7}{3}\right) = -1$$

m	1	2	3	4	5	6
$\mathbf{y}^{(m)}$	$\begin{bmatrix}0{,}0000\\0{,}6000\\0{,}8000\end{bmatrix}$	$\begin{bmatrix}0{,}6667\\-0{,}3333\\0{,}6667\end{bmatrix}$	$\begin{bmatrix}0{,}4983\\0{,}2491\\0{,}8305\end{bmatrix}$	$\begin{bmatrix}0{,}7062\\-0{,}0504\\0{,}7062\end{bmatrix}$	$\begin{bmatrix}0{,}6602\\0{,}0660\\0{,}7482\end{bmatrix}$	$\begin{bmatrix}0{,}7071\\-0{,}0114\\0{,}7071\end{bmatrix}$
$\sigma_m \, \lVert\mathbf{x}^{(m)}\rVert_2$	$-2{,}8868$	$0{,}60000$	$4{,}0139$	$1{,}6463$	$2{,}2923$	$1{,}9296$

Tabelle 5.1: Ergebnisse der ersten sechs Iterationsschritte der Von-Mises-Vektoriteration aus Beispiel 5.7.

und

$$\lVert\mathbf{x}^{(1)}\rVert_2 = \frac{1}{\sqrt{3}} \sqrt{(-3)^2 + (-4)^2} = \frac{5}{\sqrt{3}}$$

bekommen wir

$$\mathbf{y}^{(1)} = \sigma_1 \frac{\mathbf{x}^{(1)}}{\lVert\mathbf{x}^{(1)}\rVert_2} = -\frac{\sqrt{3}}{5} \cdot \frac{1}{\sqrt{3}} \begin{bmatrix}0\\-3\\-4\end{bmatrix} = \frac{1}{5} \begin{bmatrix}0\\3\\4\end{bmatrix}.$$

Schritt 2: Wir berechnen

$$\mathbf{x}^{(2)} = \mathbf{A}\,\mathbf{y}^{(1)} = \frac{1}{5} \begin{bmatrix}0 & -2 & 2\\-2 & -3 & 2\\-3 & -6 & 5\end{bmatrix} \begin{bmatrix}0\\3\\4\end{bmatrix} = \frac{1}{5} \begin{bmatrix}2\\-1\\2\end{bmatrix}.$$

Wegen

$$\sigma_2 = \operatorname{sgn}\left((\mathbf{y}^{(1)})^T \mathbf{x}^{(2)}\right) = \operatorname{sgn}\left(\frac{1}{5} \begin{bmatrix}0\\3\\4\end{bmatrix}^T \frac{1}{5} \begin{bmatrix}2\\-1\\2\end{bmatrix}\right) = \operatorname{sgn}\left(\frac{5}{25}\right) = 1$$

und

$$\lVert\mathbf{x}^{(2)}\rVert_2 = \frac{1}{5} \sqrt{2^2 + (-1)^2 + 2^2} = \frac{\sqrt{9}}{5} = \frac{3}{5}$$

erhalten wir

$$\mathbf{y}^{(2)} = \sigma_2 \frac{\mathbf{x}^{(2)}}{\lVert\mathbf{x}^{(2)}\rVert_2} = 1 \cdot \frac{5}{3} \cdot \frac{1}{5} \begin{bmatrix}2\\-1\\2\end{bmatrix} = \frac{1}{3} \begin{bmatrix}2\\-1\\2\end{bmatrix}.$$

In Tabelle 5.1 sind die Ergebnisse der ersten sechs Iterationsschritte der Von-Mises-Vektoriteration mit dem Startvektor $\mathbf{x}^{(0)} = [1;\ 1;\ 1]^T$ (auf eine Gleitkommadarstellung mit 4-stelliger Mantisse bei $\mathbf{y}^{(m)}$ gerundet und auf eine Gleitkommadarstellung mit 5-stelliger Mantisse bei $\sigma_m \|\mathbf{x}^{(m)}\|_2$) gerundet) angeben. Die Berechnung erfolgte mit Matlab.

Nach Satz 5.6 strebt $\sigma_m \|\mathbf{x}^{(m)}\|_2$ für $m \to \infty$ gegen der dominanten Eigenwert $\lambda_1 = 2$, und $\mathbf{y}^{(m)}$ strebt für $m \to \infty$ gegen einen zugehörigen normierten Eigenvektor zu $\lambda_1 = 2$, hier gegen $\mathbf{w}_1 = (\sqrt{2})^{-1}\,[1;\,0;\,1]^T \doteq [0{,}707107;\ 0;\ 0{,}707107]^T$. Wir sehen, dass das Ergebnis nach sechs Iterationsschritten nicht besonders gut ist. Nach 15 Iterationsschritten bekommt man die folgenden Näherungen:

$$\lambda_1 \approx \sigma_{15} \|\mathbf{x}^{(15)}\|_2 \doteq 2{,}0002 \qquad \text{und} \qquad \mathbf{w}_1 \approx \mathbf{y}^{(15)} \doteq \begin{bmatrix} 0{,}7071 \\ 0{,}0001 \\ 0{,}7071 \end{bmatrix}.$$

Dieses sind gute Näherungen für $\lambda_1 = 2$ und $\mathbf{w}_1 = (\sqrt{2})^{-1}\,[1;\,0;\,1]^T$. ♠

Als Letztes beweisen wir für mathematisch Interessierte noch Satz 5.6.

Beweis von Satz 5.6: Nach der Von-Mises-Vektoriteration gilt für $k \in \mathbb{N}$

$$\mathbf{y}^{(k)} = \sigma_k \frac{\mathbf{x}^{(k)}}{\|\mathbf{x}^{(k)}\|_2} = \sigma_k \frac{\mathbf{A}\,\mathbf{y}^{(k-1)}}{\|\mathbf{A}\,\mathbf{y}^{(k-1)}\|}\,. \tag{5.28}$$

Wendet man (5.28) wiederholt für $k = m, m-1, \ldots, 2, 1$ an, so bekommt man

$$\begin{aligned} \mathbf{y}^{(m)} &= \sigma_m \frac{\mathbf{A}\,\mathbf{y}^{(m-1)}}{\|\mathbf{A}\,\mathbf{y}^{(m-1)}\|_2} = \sigma_m \frac{\mathbf{A}\left(\sigma_{m-1} \frac{\mathbf{A}\,\mathbf{y}^{(m-2)}}{\|\mathbf{A}\,\mathbf{y}^{(m-2)}\|_2}\right)}{\left\|\sigma_{m-1} \frac{\mathbf{A}^2\,\mathbf{y}^{(m-2)}}{\|\mathbf{A}\,\mathbf{y}^{(m-2)}\|_2}\right\|_2} = \sigma_m\,\sigma_{m-1} \frac{\mathbf{A}^2\,\mathbf{y}^{(m-2)}}{\|\mathbf{A}^2\,\mathbf{y}^{(m-2)}\|_2} \\ &= \sigma_m\,\sigma_{m-1} \cdots \sigma_1 \frac{\mathbf{A}^m\,\mathbf{y}^{(0)}}{\|\mathbf{A}^m\,\mathbf{y}^{(0)}\|_2} = \sigma_m\,\sigma_{m-1} \cdots \sigma_1 \frac{\mathbf{A}^m\,\mathbf{x}^{(0)}}{\|\mathbf{A}^m\,\mathbf{x}^{(0)}\|_2}\,, \end{aligned} \tag{5.29}$$

für $m = 1, 2, \ldots$, wobei wir $\mathbf{y}^{(0)} = \mathbf{x}^{(0)}/\|\mathbf{x}^{(0)}\|_2$ im letzten Schritt genutzt haben. (Dabei kürzt sich $\|\mathbf{x}^{(0)}\|_2$ weg.)

Aus (5.29) folgt durch Multiplizieren mit $\mathbf{A}$ von vorne

$$\mathbf{x}^{(m+1)} = \mathbf{A}\,\mathbf{y}^{(m)} = \sigma_m\,\sigma_{m-1} \cdots \sigma_1 \frac{\mathbf{A}^{m+1}\,\mathbf{x}^{(0)}}{\|\mathbf{A}^m\,\mathbf{x}^{(0)}\|_2}\,,$$

und Anwenden der 2-Norm und Ausnutzen von (5.27) with $\mathbf{x} = \mathbf{x}^{(0)}$ liefern

$$\|\mathbf{x}^{(m+1)}\|_2 = \frac{\|\mathbf{A}^{m+1}\,\mathbf{x}^{(0)}\|_2}{\|\mathbf{A}^m\,\mathbf{x}^{(0)}\|_2} = |\lambda_1| \frac{\sqrt{|c_1|^2 + r_{m+1}}}{\sqrt{|c_1|^2 + r_m}} \overset{m\to\infty}{\longrightarrow} |\lambda_1|.$$

Einsetzen der Darstellungen (5.21) und (5.26) in die Darstellung (5.29) von $\mathbf{y}^{(m)}$ ergeben für hinreichend großes m

$$\begin{aligned}\mathbf{y}^{(m)} &= \sigma_m\,\sigma_{m-1}\,\cdots\,\sigma_1\,\frac{\lambda_1^m\,(c_1\,\mathbf{w}_1+\mathbf{R}_m)}{|\lambda_1|^m\,\sqrt{|c_1|^2+r_m}}\\ &= \sigma_m\,\sigma_{m-1}\,\cdots\,\sigma_1\,\big[\operatorname{sgn}(\lambda_1)\big]^m\,\operatorname{sgn}(c_1)\,\frac{|c_1|}{\sqrt{|c_1|^2+r_m}}\,\mathbf{w}_1+\boldsymbol{\varrho}_m,\end{aligned}\tag{5.30}$$

wobei der Restgliedterm $\boldsymbol{\varrho}_m$ wie folgt definiert ist

$$\boldsymbol{\varrho}_m := \sigma_m\,\sigma_{m-1}\,\cdots\,\sigma_1\,\big[\operatorname{sgn}(\lambda_1)\big]^m\,\frac{\mathbf{R}_m}{\sqrt{|c_1|^2+r_m}}\,.$$

Da $c_1\neq 0$ ist und da $\mathbf{R}_m\to\mathbf{0}$ und $r_m\to 0$ für $m\to\infty$ gelten, folgt $\boldsymbol{\varrho}_m\to\mathbf{0}$ für $m\to\infty$. Aus (5.30) und $\boldsymbol{\varrho}_m\to\mathbf{0}$ für $m\to\infty$ folgt, dass $\mathbf{y}^{(m)}$ in der Tat gegen einen Eigenvektor of $\mathbf{A}$ zum Eigenwert λ_1 konvergiert, wenn $\sigma_m=\operatorname{sgn}(\lambda_1)$ für alle $m\geq m_0$ erfüllt ist. Diese Bedingung ist aber wegen $(\mathbf{y}^{(m-1)})^T\,\mathbf{y}^{(m)} = (\mathbf{y}^{(m)})^T\,\mathbf{y}^{(m-1)}\geq 0$ gemäß der ersten Zeile in (5.30) mit $\|\mathbf{w}_1\|_2^2=1$ erfüllt:

$$\begin{aligned}0\leq (\mathbf{y}^{(m-1)})^T\,\mathbf{y}^{(m)} &= \sigma_m\,\sigma_{m-1}^2\,\cdots\,\sigma_1^2\,\frac{\lambda_1^{2m-1}\,\big(c_1\,\mathbf{w}_1+\mathbf{R}_{m-1}\big)^T\,\big(c_1\,\mathbf{w}_1+\mathbf{R}_m\big)}{|\lambda_1|^{2m-1}\,\sqrt{|c_1|^2+r_{m-1}}\,\sqrt{|c_1|^2+r_m}}\\ &= \sigma_m\,\operatorname{sgn}(\lambda_1)\,\frac{|c_1|^2+\mathbf{R}_{m-1}^T\,\mathbf{R}_m+c_1\,\mathbf{w}_1^T\,(\mathbf{R}_{m-1}+\mathbf{R}_m)}{|c_1|^2\sqrt{\big(1+\frac{r_{m-1}}{|c_1|^2}\big)\big(1+\frac{r_m}{|c_1|^2}\big)}}\end{aligned}\tag{5.31}$$

Weil $\mathbf{R}_m\to\mathbf{0}$ und $r_m\to 0$ für $m\to\infty$ gelten, strebt der Bruch in der zweiten Zeile von (5.31) für $m\to\infty$ gegen 1, and somit folgt aus (5.31) für hinreichend großes m, dass

$$0\leq\sigma_m\,\operatorname{sgn}(\lambda_1)\qquad\Longleftrightarrow\qquad\sigma_m=\operatorname{sgn}(\lambda_1),$$

gilt. Damit haben wir alle Aussagen von Satz 5.6 bewiesen. □

5.3 Transformation in Hessenberg-Form

In diesem Teilkapitel lernen wir, wie man eine quadratische Matrix $\mathbf{A}\in\mathbb{R}^{n\times n}$ so in eine sogenannte Hessenberg-Matrix umformen kann, dass sich die Eigenwerte nicht ändern. Bei diesem Prozess verwenden wir orthogonale Householder-Matrizen, welche Sie schon aus Teilkapitel 2.4 kennen. Die Transformation der

Matrix in Hessenberg-Form dient als Vorbereitungsschritt für das QR-Verfahren zur Berechnung aller Eigenwerte, welches wir im nächsten Teilkapitel besprechen.

Ist eine Matrix $\mathbf{B} = [b_{j,k}] \in \mathbb{R}^{n\times n}$ eine obere Dreiecksmatrix (d.h. $b_{j,k} = 0$ für alle $j > k$), so sind die Eigenwerte von $\mathbf{B}$ genau die Einträge auf der Diagonalen von $\mathbf{B}$. Dieses folgt, indem man das charakteristische Polynom

$$p_{\mathbf{B}}(\lambda) = \det\left(\mathbf{B} - \lambda\,\mathbf{E}_n\right)$$

mit dem Laplaceschen Entwicklungssatz wiederholt nach der ersten Spalte entwickelt. Dieses ergibt

$$p_{\mathbf{B}}(\lambda) = \det\left(\mathbf{B} - \lambda\,\mathbf{E}_n\right) = (\lambda - b_{1,1})\,(\lambda - b_{2,2})\,\cdots\,(\lambda - b_{n,n}),$$

d.h. die Eigenwerte von $\mathbf{B}$ sind $\lambda_1 = b_{1,1}$, $\lambda_2 = b_{2,2}$, $\ldots$, $\lambda_n = b_{n,n}$.

Daher wäre es zur Bestimmung der Eigenwerte einer Matrix $\mathbf{A} \in \mathbb{R}^{n\times n}$ wünschenswert, das man die Matrix $\mathbf{A}$ mittels einer Transformation, welche die Eigenwerte unverändert lässt, in eine obere Dreiecksmatrix überführen kann. Dieses ist auch möglich. Die **Schur-Zerlegung** (welche wir in diesem Kurs nicht besprechen) besagt, dass es zu jeder quadratischen Matrix $\mathbf{A} \in \mathbb{R}^{n\times n}$ eine **orthogonale Matrix** $\mathbf{S} \in \mathbb{R}^{n\times n}$ (d.h. $\mathbf{S}^{-1} = \mathbf{S}^T$) gibt, so dass gilt

$$\mathbf{S}^T\mathbf{A}\,\mathbf{S} = \mathbf{S}^{-1}\mathbf{A}\,\mathbf{S} = \mathbf{R} \qquad \Longleftrightarrow \qquad \mathbf{A} = \mathbf{S}\,\mathbf{R}\,\mathbf{S}^T = \mathbf{S}\,\mathbf{R}\,\mathbf{S}^{-1}, \tag{5.32}$$

wobei $\mathbf{R} \in \mathbb{R}^{n\times n}$ eine obere Dreiecksmatrix ist. Die Äquivalenzumformung in (5.32) erfolgt, indem man von links mit $\mathbf{S}$ und von rechts mit $\mathbf{S}^{-1} = \mathbf{S}^T$ multipliziert und dann $\mathbf{S}\,\mathbf{S}^{-1} = \mathbf{E}_n$ ausnutzt. Wie wir gleich zeigen werden, lässt eine Transformation der Form $\mathbf{S}^{-1}\,\mathbf{A}\,\mathbf{S}$ die Eigenwerte unverändert.

Hilfssatz 5.8. (Ähnlichkeitstransformation erhält Eigenwerte)
*Sei $\mathbf{A} \in \mathbb{R}^{n\times n}$, und sei $\mathbf{S} \in \mathbb{R}^{n\times n}$ eine invertierbare Matrix. Die Eigenwerte von $\mathbf{A}$ bleiben unter einer **Ähnlichkeitstransformation** $\mathbf{S}^{-1}\mathbf{A}\,\mathbf{S}$ unverändert, d.h. die Matrizen $\mathbf{A}$ und $\mathbf{S}^{-1}\mathbf{A}\,\mathbf{S}$ haben **genau dieselben Eigenwerte** (einschließlich der Vielfachheiten).*

Beweis von Hilfssatz 5.8: Sei $\mathbf{B} := \mathbf{S}^{-1}\,\mathbf{A}\,\mathbf{S}$. Dann folgt

$$\begin{aligned} p_{\mathbf{B}}(\lambda) &= \det(\mathbf{B} - \lambda\,\mathbf{E}_n) = \det\left(\mathbf{S}^{-1}\,\mathbf{A}\,\mathbf{S} - \lambda\,\mathbf{E}_n\right) \\ &= \det\left(\mathbf{S}^{-1}\,\mathbf{A}\,\mathbf{S} - \lambda\,\mathbf{S}^{-1}\,\mathbf{S}\right) = \det\left(\mathbf{S}^{-1}\,\mathbf{A}\,\mathbf{S} - \lambda\,\mathbf{S}^{-1}\,\mathbf{E}_n\,\mathbf{S}\right) \\ &= \det\left(\mathbf{S}^{-1}\,(\mathbf{A} - \lambda\,\mathbf{E}_n)\,\mathbf{S}\right) = \det(\mathbf{S}^{-1})\,\det(\mathbf{A} - \lambda\,\mathbf{E}_n)\,\det(\mathbf{S}) \end{aligned}$$

$$= \frac{1}{\det(\mathbf{S})} \det(\mathbf{A} - \lambda\,\mathbf{E}_n)\,\det(\mathbf{S}) = \det(\mathbf{A} - \lambda\,\mathbf{E}_n) = p_{\mathbf{A}}(\lambda),$$

wobei wir $\mathbf{S}^{-1}\,\mathbf{S} = \mathbf{E}_n$, den Multiplikationssatz $\det(\mathbf{C}\,\mathbf{B}) = \det(\mathbf{B})\,\det(\mathbf{C})$ für Determinanten und $\det(\mathbf{S}^{-1}) = 1/\det(\mathbf{S})$ genutzt haben. Nach der obigen Rechnung haben die Matrizen $\mathbf{A}$ und $\mathbf{B} = \mathbf{S}^{-1}\,\mathbf{A}\,\mathbf{S}$ das gleiche charakteristische Polynom und somit die gleichen Eigenwerte (einschließlich der Vielfachheiten). □

Als ersten Schritt, um eine Matrix $\mathbf{A} \in \mathbb{R}^{n\times n}$ mit einem Iterationsverfahren in eine obere Dreiecksmatrix zu bringen, transformieren wir $\mathbf{A}$ in eine sogenannte **Hessenberg-Matrix**. In einem zweiten Schritt erzeugen wir mit Ähnlichkeitstransformationen dann eine Folge von Hessenberg-Matrizen, die gegen eine obere Dreiecksmatrix konvergieren.

Definition 5.9. (Hessenberg-Matrix)
*Eine Matrix $\mathbf{A} = [a_{i,j}] \in \mathbb{R}^{n\times n}$ heißt eine **Hessenberg-Matrix**, wenn alle Einträge unterhalb der ersten unteren Nebendiagonalen null sind, also wenn gilt $a_{i,j} = 0$ für alle $i > j + 1$. $\mathbf{A}$ hat dann die folgende Form*

$$\mathbf{A} = \begin{bmatrix} * & * & \cdots & * & * & * \\ * & * & \cdots & * & * & * \\ 0 & * & \ddots & \vdots & \vdots & \vdots \\ 0 & 0 & \ddots & * & * & * \\ \vdots & \vdots & \ddots & * & * & * \\ 0 & 0 & \cdots & 0 & * & * \end{bmatrix}.$$

In diesem Teilkapitel lernen wir, wie man $\mathbf{A} \in \mathbb{R}^{n\times n}$ durch Ähnlichkeitstransformationen mit Householder-Matrizen in eine Hessenberg-Matrix überführt.

Zur Erinnerung: Householder-Matrizen wurden in Definition 2.25 in Teilkapitel 2.4 eingeführt. Eine **Householder-Matrix** ist eine Matrix in $\mathbb{R}^{n\times n}$ der Form

$$\mathbf{H}(\mathbf{w}) := \mathbf{E}_n - 2\,\mathbf{w}\,\mathbf{w}^T, \quad \text{wobei } \mathbf{w} \in \mathbb{R}^n \text{ mit } \mathbf{w}^T\mathbf{w} = \|\mathbf{w}\|_2^2 = 1 \text{ oder } \mathbf{w} = \mathbf{0}.$$

Householder Matrizen $\mathbf{H}(\mathbf{w})$ sind **orthogonal**, d.h. es gilt $\big(\mathbf{H}(\mathbf{w})\big)^T = \big(\mathbf{H}(\mathbf{w})\big)^{-1}$, und **symmetrisch**, d.h. es gilt $\big(\mathbf{H}(\mathbf{w})\big)^T = \mathbf{H}(\mathbf{w})$.

Satz 5.10. (Householder-Transformation in Hessenberg-Form)
*Sei $\mathbf{A} \in \mathbb{R}^{n\times n}$ Dann existieren $n-2$ **Householder-Matrizen** $\mathbf{H}(\mathbf{w}_1)$, $\mathbf{H}(\mathbf{w}_2)$, $\ldots$, $\mathbf{H}(\mathbf{w}_{n-2})$, so dass $\mathbf{S}^T\,\mathbf{A}\,\mathbf{S}$ mit $\mathbf{S} := \mathbf{H}(\mathbf{w}_1)\,\mathbf{H}(\mathbf{w}_2)\cdots\mathbf{H}(\mathbf{w}_{n-2})$, eine **Hessenberg-Matrix** ist.*

Der Beweis ist konstruktiv und liefert direkt die Vorgehensweise zur praktischen Bestimmung der Householder-Matrizen.

Beweis von Satz 5.10: Der Beweis erfolgt mit vollständiger Induktion.

Seien $\mathbf{A}^{(0)} = \mathbf{A}$, und $\mathbf{A}^{(k-1)} = \big(\mathbf{H}(\mathbf{w}_{k-1})\big)^T \mathbf{A}^{(k-2)}\, \mathbf{H}(\mathbf{w}_{k-1})$ werde rekursiv mit der Startmatrix $\mathbf{A}^{(0)}$ berechnet. Wir behaupten, dass $\mathbf{A}^{(k-1)}$ die folgende Form

$$\mathbf{A}^{(k-1)} = \left[\begin{array}{c|c} \mathbf{A}_k^{(k-1)} & \mathbf{B}^{(k-1)} \\ \hline \mathbf{C}^{(k-1)} & \mathbf{D}^{(k-1)} \end{array}\right] \tag{5.33}$$

hat, wobei die Teilmatrix $\mathbf{A}_k^{(k-1)} \in \mathbb{R}^{k\times k}$ eine $k \times k$-Hessenberg-Matrix ist und $\mathbf{C}^{(k-1)} = \big[\mathbf{0}; \mathbf{0}; \ldots; \mathbf{0}; \mathbf{c}_k\big]$ eine $(n-k) \times k$-Matrix ist. Die Matrizen $\mathbf{B}^{(k-1)}$ und $\mathbf{D}^{(k-1)}$ sind eine beliebige $k \times (n-k)$-Matrix bzw. eine beliebige $(n-k) \times (n-k)$-Matrix.

Beweis der Behauptung mit vollstäniger Induktion:

Induktionsanfang: Für $k = 1$ hat $\mathbf{A}^{(0)} = \mathbf{A}$ eine wie oben beschriebene Zerlegung, denn $\mathbf{A}_1^{(0)} = [a_{1,1}] \in \mathbb{R}^{1\times 1}$ ist eine 1×1 Hessenberg-Matrix und $\mathbf{C}^{(0)} = [\mathbf{c}_0] \in \mathbb{R}^{(n-1)\times 1}$ mit $\mathbf{c}_0 = [a_{2,1}; a_{3,1}; \ldots; a_{n,1}]^T$ hat die passende Form. Also haben wir eine Induktionsverankerung für $k = 1$.

Induktionsschritt $k \curvearrowright k+1$: Es gelte (5.33) für ein $k \in \mathbb{N}$. Sei $\mathbf{H}_k = \mathbf{H}(\mathbf{u}_k)$ eine $(n-k) \times (n-k)$ Householder-Matrix mit der Eigenschaft, dass $\mathbf{H}_k\, \mathbf{c}_k = \|\mathbf{c}_k\|_2\, \mathbf{e}_1$, wobei $\mathbf{e}_1$ der erste Vektor des Standardbasis von $\mathbb{R}^{n-k}$ ist. Wir definieren nun $\mathbf{w}_k := \begin{bmatrix} \mathbf{0} \\ \mathbf{u}_k \end{bmatrix}$ in $\mathbb{R}^n$, wobei $\mathbf{0}$ der Nullvektor in $\mathbb{R}^k$ ist. Dann gilt

$$\mathbf{H}(\mathbf{w}_k) = \mathbf{E}_n - 2\, \mathbf{w}_k\, \mathbf{w}_k^T = \left[\begin{array}{c|c} \mathbf{E}_k & \mathbf{O}_{k\times(n-k)} \\ \hline \mathbf{O}_{(n-k)\times k} & \mathbf{H}_k \end{array}\right].$$

(Die Matrizen $\mathbf{O}_{k\times(n-k)}$ und $\mathbf{O}_{(n-k)\times k}$ sind dabei die $k \times (n-k)$-Nullmatrix bzw. die $(n-k) \times k$-Nullmatrix.) Die Ähnlichkeitstransformation mit der Householder-Matrix $\mathbf{H}(\mathbf{w}_k)$ ergibt (mit $\big(\mathbf{H}(\mathbf{w}_k)\big)^{-1} = \big(\mathbf{H}(\mathbf{w}_k)\big)^T = \mathbf{H}(\mathbf{w}_k)$)

$$\big(\mathbf{H}(\mathbf{w}_k)\big)^T\, \mathbf{A}^{(k-1)}\, \mathbf{H}(\mathbf{w}_k) = \mathbf{H}(\mathbf{w}_k)\, \mathbf{A}^{(k-1)}\, \mathbf{H}(\mathbf{w}_k)$$

$$= \left[\begin{array}{c|c} \mathbf{E}_k & \mathbf{O}_{k\times(n-k)} \\ \hline \mathbf{O}_{(n-k)\times k} & \mathbf{H}_k \end{array}\right] \cdot \left[\begin{array}{c|c} \mathbf{A}_k^{(k-1)} & \mathbf{B}^{(k-1)} \\ \hline \mathbf{C}^{(k-1)} & \mathbf{D}^{(k-1)} \end{array}\right] \cdot \left[\begin{array}{c|c} \mathbf{E}_k & \mathbf{O}_{k\times(n-k)} \\ \hline \mathbf{O}_{(n-k)\times k} & \mathbf{H}_k \end{array}\right]$$

$$= \left[\begin{array}{c|c} \mathbf{A}_k^{(k-1)} & \mathbf{B}^{(k-1)} \\ \hline \mathbf{H}_k\,\mathbf{C}^{(k-1)} & \mathbf{H}_k\,\mathbf{D}^{(k-1)} \end{array}\right] \cdot \left[\begin{array}{c|c} \mathbf{E}_k & \mathbf{O}_{k\times(n-k)} \\ \hline \mathbf{O}_{(n-k)\times k} & \mathbf{H}_k \end{array}\right]$$

$$= \left[\begin{array}{c|c} \mathbf{A}_k^{(k-1)} & \mathbf{B}^{(k-1)}\,\mathbf{H}_k \\ \hline \mathbf{H}_k\,\mathbf{C}^{(k-1)} & \mathbf{H}_k\,\mathbf{D}^{(k-1)}\,\mathbf{H}_k \end{array}\right] =: \mathbf{A}^{(k)}.$$

Nach Konstruktion der $(n-k)\times(n-k)$ Householder-Matrix $\mathbf{H}_k$ und mit $\mathbf{C}^{(k-1)} = \big[\mathbf{0};\mathbf{0};\ldots;\mathbf{0};\mathbf{c}_k\big]$ gilt $\mathbf{H}_k\,\mathbf{C}^{(k-1)} = \big[\mathbf{0};\mathbf{0};\ldots,\mathbf{0};c_k\,\mathbf{e}_1\big]$. Zerlegt man $\mathbf{A}^{(k)}$ wie in (5.33) mit $k-1$ durch k ersetzt, so findet man, dass $\mathbf{A}_{k+1}^{(k)}$ eine $(k+1)\times(k+1)$ Hessenberg-Matrix ist und dass $\mathbf{C}^{(k)}$ eine $(n-k-1)\times(k+1)$-Matrix von der Form $\big[\mathbf{0};\mathbf{0};\ldots;\mathbf{0};\mathbf{c}_{k+1}\big]^T$ ist. Damit ist der Induktionsschritt abgeschlossen.

Nach dem Induktionsprinzip gilt die Aussage für alle $k = 1, 2, \ldots, n-2$. (Mehr als $n-2$ Schritte ergeben keinen Sinn, da die Hessenberg-Form (spätestens) nach $n-2$ Schritten erreicht wird.)

Nach Konstruktion gilt mit $\mathbf{A}^{(0)} = \mathbf{A}$ die Formel

$$\begin{aligned}\mathbf{A}^{(n-2)} &= \big(\mathbf{H}(\mathbf{w}_{n-2})\big)^T \cdot \ldots \cdot \big(\mathbf{H}(\mathbf{w}_2)\big)^T \big(\mathbf{H}(\mathbf{w}_1)\big)^T \mathbf{A}\,\mathbf{H}(\mathbf{w}_1)\,\mathbf{H}(\mathbf{w}_2) \cdot \ldots \cdot \mathbf{H}(\mathbf{w}_{n-2}) \\ &= \big(\mathbf{H}(\mathbf{w}_1)\,\mathbf{H}(\mathbf{w}_2) \cdot \ldots \cdot \mathbf{H}(\mathbf{w}_{n-2})\big)^T \mathbf{A}\,\mathbf{H}(\mathbf{w}_1)\,\mathbf{H}(\mathbf{w}_2) \cdot \ldots \cdot \mathbf{H}(\mathbf{w}_{n-2}),\end{aligned}$$

und mit $\mathbf{S} := \mathbf{H}(\mathbf{w}_1)\,\mathbf{H}(\mathbf{w}_2) \cdot \ldots \cdot \mathbf{H}(\mathbf{w}_{n-2})$ folgt $\mathbf{S}^T\mathbf{A}\,\mathbf{S} = \mathbf{A}^{(n-2)}$ mit der Hessenberg-Matrix $\mathbf{A}^{(n-2)}$. □

Bemerkung 5.11. (Bestimmung der Householder-Matrizen)
Dem Beweis von Satz 5.10 entnehmen wir, was im **k-ten Schritt der Householder-Transformation auf Hessenberg-Form** zu tun ist: Wir benötigen einen Vektor $\mathbf{u}_k \in \mathbb{R}^{n-k}$, so dass $\mathbf{H}(\mathbf{u}_k)$ eine $(n-k)\times(n-k)$ Householder-Matrix der Eigenschaft

$$\mathbf{H}(\mathbf{u}_k)\,\mathbf{c}_k = \|\mathbf{c}_k\|_2\,\mathbf{e}_1,$$

wobei $\mathbf{c}_k \in \mathbb{R}^{n-k}$ der Vektor mit den untersten $n-k$ Einträgen der k-ten Spalte der Matrix $\mathbf{A}^{(k-1)}$ aus dem vorherigen Schritt sind. Gemäß Hilfssatz 2.28 definieren wir $\mathbf{u}_k \in \mathbb{R}^{n-k}$ mit

$$\mathbf{v}_k := \mathbf{c}_k - \|\mathbf{c}_k\|_2\,\mathbf{e}_1 \qquad \text{und} \qquad \mathbf{u}_k := \frac{\mathbf{v}_k}{\|\mathbf{v}_k\|_2}.$$

Die $n \times n$ Householder-Matrix $\mathbf{H}(\mathbf{w}_k)$ für die Householder-Transformation erhält man dann durch Einbettung von $\mathbf{H}(\mathbf{u}_k)$ wie folgt:

$$\mathbf{H}(\mathbf{w}_k) = \left[\begin{array}{c|c} \mathbf{E}_k & \mathbf{O}_{k\times(n-k)} \\ \hline \mathbf{O}_{(n-k)\times k} & \mathbf{H}_k \end{array}\right].$$

Wir führen die Householder-Transformation auf Hessenberg-Form an einem Beispiel per Hand durch.

Beispiel 5.12. (Householder-Transformation auf Hessenberg-Form)
Wir wollen die Householder-Transformation auf Hessenberg-Form für die folgende 4×4-Matrix durchführen:

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

Dazu benötigen wir (höchstens) $4 - 2 = 2$ Schritte:

Schritt 1 (erste Spalte von $\mathbf{A}^{(0)} = \mathbf{A}$): $\mathbf{A}^{(0)} = \mathbf{A} = \begin{bmatrix} 1 & 0 & 4 & 0 \\ 0 & 3 & 3 & 4 \\ 4 & 3 & 3 & 4 \\ 0 & 4 & 4 & -3 \end{bmatrix}$

Die zu konstruierende 3×3 Householder-Matrix soll $\begin{bmatrix} 0 \\ 4 \\ 0 \end{bmatrix}$ auf $c_1 \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix}$ abbilden.

Wir konstruieren die 3×3 Householder-Matrix:

$$\mathbf{c}_1 = \begin{bmatrix} 0 \\ 4 \\ 0 \end{bmatrix} \qquad \text{und} \qquad c_1 = \|\mathbf{c}_1\|_2 = 4,$$

$$\mathbf{v}_1 = \mathbf{c}_1 - c_1\,\mathbf{e}_1 = \begin{bmatrix} 0 \\ 4 \\ 0 \end{bmatrix} - 4 \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix} = \begin{bmatrix} -4 \\ 4 \\ 0 \end{bmatrix} \qquad \text{und} \qquad \|\mathbf{v}_1\|_2 = 4\sqrt{2},$$

$$\mathbf{u}_1 = \frac{\mathbf{v}_1}{\|\mathbf{v}_1\|_2} = \frac{1}{4\sqrt{2}} \begin{bmatrix} -4 \\ 4 \\ 0 \end{bmatrix} = \frac{1}{\sqrt{2}} \begin{bmatrix} -1 \\ 1 \\ 0 \end{bmatrix}$$

Die 3×3 Householder-Matrix ist also

$$\mathbf{H}_1 = \mathbf{H}(\mathbf{u}_1) = \mathbf{E}_3 - 2\,\mathbf{u}_1\,\mathbf{u}_1^T = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} - 2 \cdot \frac{1}{(\sqrt{2})^2} \begin{bmatrix} -1 \\ 1 \\ 0 \end{bmatrix} \begin{bmatrix} -1;\ 1;\ 0 \end{bmatrix}$$

$$= \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} - \begin{bmatrix} 1 & -1 & 0 \\ -1 & 1 & 0 \\ 0 & 0 & 0 \end{bmatrix} = \begin{bmatrix} 0 & 1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 1 \end{bmatrix}.$$

Die zugehörige 4×4 Householder-Matrix ist dann

$$\mathbf{H}(\mathbf{w}_1) = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix},$$

und wir erhalten

$$\mathbf{A}^{(1)} := \mathbf{H}(\mathbf{w}_1)^T\,\mathbf{A}^{(0)}\,\mathbf{H}(\mathbf{w}_1) = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} 1 & 0 & 4 & 0 \\ 0 & 3 & 3 & 4 \\ 4 & 3 & 3 & 4 \\ 0 & 4 & 4 & -3 \end{bmatrix} \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}$$

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

Schritt 2 (zweite Spalte von $\mathbf{A}^{(1)}$): $\mathbf{A}^{(1)} = \begin{bmatrix} 1 & 4 & 0 & 0 \\ 4 & 3 & 3 & 4 \\ 0 & 3 & 3 & 4 \\ 0 & 4 & 4 & -3 \end{bmatrix}$

Die zu konstruierende 2×2 Householder-Matrix soll $\begin{bmatrix} 3 \\ 4 \end{bmatrix}$ auf $c_2 \begin{bmatrix} 1 \\ 0 \end{bmatrix}$ abbilden.

Wir konstruieren die 2×2-Householder-Matrix:

$$\mathbf{c}_2 = \begin{bmatrix} 3 \\ 4 \end{bmatrix} \qquad \text{und} \qquad c_2 = \|\mathbf{c}_2\|_2 = \sqrt{25} = 5,$$

$$\mathbf{v}_2 = \mathbf{c}_2 - c_2\,\mathbf{e}_1 = \begin{bmatrix} 3 \\ 4 \end{bmatrix} - 5 \begin{bmatrix} 1 \\ 0 \end{bmatrix} = \begin{bmatrix} -2 \\ 4 \end{bmatrix} \qquad \text{und} \qquad \|\mathbf{v}_2\|_2 = \sqrt{20} = 2\sqrt{5},$$

$$\mathbf{u}_2 = \frac{\mathbf{v}_2}{\|\mathbf{v}_2\|_2} = \frac{1}{2\sqrt{5}} \begin{bmatrix} -2 \\ 4 \end{bmatrix} = \frac{1}{\sqrt{5}} \begin{bmatrix} -1 \\ 2 \end{bmatrix}$$

Die 2×2 Householder-Matrix ist also

$$\mathbf{H}_2 = \mathbf{H}(\mathbf{u}_2) = \mathbf{E}_2 - 2\,\mathbf{u}_2\,\mathbf{u}_2^T = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} - 2 \cdot \frac{1}{(\sqrt{5})^2} \begin{bmatrix} -1 \\ 2 \end{bmatrix} \begin{bmatrix} -1;\ 2 \end{bmatrix}$$

$$= \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} - \frac{2}{5} \begin{bmatrix} 1 & -2 \\ -2 & 4 \end{bmatrix} = \begin{bmatrix} \frac{3}{5} & \frac{4}{5} \\ \frac{4}{5} & -\frac{3}{5} \end{bmatrix}.$$

Die zugehörige 4×4 Householder-Matrix ist dann

$$\mathbf{H}(\mathbf{w}_2) = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & \frac{3}{5} & \frac{4}{5} \\ 0 & 0 & \frac{4}{5} & -\frac{3}{5} \end{bmatrix},$$

und wir erhalten

$$\mathbf{A}^{(2)} := \mathbf{H}(\mathbf{w}_2)^T\,\mathbf{A}^{(1)}\,\mathbf{H}(\mathbf{w}_2) = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & \frac{3}{5} & \frac{4}{5} \\ 0 & 0 & \frac{4}{5} & -\frac{3}{5} \end{bmatrix} \begin{bmatrix} 1 & 4 & 0 & 0 \\ 4 & 3 & 3 & 4 \\ 0 & 3 & 3 & 4 \\ 0 & 4 & 4 & -3 \end{bmatrix} \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & \frac{3}{5} & \frac{4}{5} \\ 0 & 0 & \frac{4}{5} & -\frac{3}{5} \end{bmatrix}$$

$$= \begin{bmatrix} 1 & 4 & 0 & 0 \\ 4 & 3 & 3 & 4 \\ 0 & 5 & 5 & 0 \\ 0 & 0 & 0 & 5 \end{bmatrix} \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & \frac{3}{5} & \frac{4}{5} \\ 0 & 0 & \frac{4}{5} & -\frac{3}{5} \end{bmatrix} = \begin{bmatrix} 1 & 4 & 0 & 0 \\ 4 & 3 & 5 & 0 \\ 0 & 5 & 3 & 4 \\ 0 & 0 & 4 & -3 \end{bmatrix}.$$

Wir beobachten, dass $\mathbf{A}^{(2)}$ in der Tat in eine Hessenberg-Matrix ist. ♠

Der Rechenaufwand für die Householder-Transformation auf Hessenberg-Form ist $\mathcal{O}(n^3)$.

5.4 QR-Verfahren zur Eigenwertberechnung

Das **QR-Verfahren** zur Eigenwertberechnung ist ein Iterationsverfahren, welches alle Eigenwerte auf einmal berechnet. das QR-Verfahren profitiert davon (also konvergiert in der Regel schneller), wenn man die Matrix vorab mittels Householder-Transformation in eine Hessenberg-Matrix transformiert.

Verfahren 5.13. (QR-Verfahren zur Eigenwertberechnung)
Sei $\mathbf{A} \in \mathbb{R}^{n \times n}$.

Initialisierung: Setzte $\mathbf{A}_0 := \mathbf{A}$.

Für $m = 0, 1, 2, \ldots$ *führen wir folgende Schritte aus:*

(1) Bestimme die ***QR-Zerlegung*** $\mathbf{A}_m = \mathbf{Q}_m \mathbf{R}_m$ *von* $\mathbf{A}_m$ *(mit einer orthogonalen Matrix* $\mathbf{Q}_m$ *und einer oberen Dreiecksmatrix* $\mathbf{R}_m$*).*

(2) Berechne $\mathbf{A}_{m+1} := \mathbf{R}_m \mathbf{Q}_m$ *(vertausche Reihenfolge von* $\mathbf{Q}_m$ *und* $\mathbf{R}_m$*).*

Da $\mathbf{Q}_m$ eine orthogonale Matrix ist, also $\mathbf{Q}_m^T = \mathbf{Q}_m^{-1}$ erfüllt, folgt aus der QR-Zerlegung $\mathbf{A}_m = \mathbf{Q}_m \mathbf{R}_m$ von $\mathbf{A}_m$

$$\mathbf{A}_m = \mathbf{Q}_m \mathbf{R}_m \quad \Longleftrightarrow \quad \mathbf{Q}_m^T \mathbf{A}_m = \mathbf{R}_m \quad \Longleftrightarrow \quad \mathbf{Q}_m^T \mathbf{A}_m \mathbf{Q}_m = \mathbf{R}_m \mathbf{Q}_m$$

und wir lesen ab, dass gilt

$$\mathbf{A}_{m+1} = \mathbf{R}_m \mathbf{Q}_m = \mathbf{Q}_m^T \mathbf{A}_m \mathbf{Q}_m = \mathbf{Q}_m^{-1} \mathbf{A}_m \mathbf{Q}_m. \tag{5.34}$$

Die Formel (5.34) für die Berechnung der Matrix $\mathbf{A}_{m+1}$ aus der Matrix $\mathbf{A}_m$ aus dem vorherigen Schritt des QR-Verfahrens zeigt, dass $\mathbf{A}_{m+1}$ aus $\mathbf{A}_m$ durch eine Ähnlichkeitstransformation berechnet wird und dass somit $\mathbf{A}_{m+1}$ und $\mathbf{A}_m$ die gleichen Eigenwerte haben (vgl. Hilfssatz 5.8). Durch wiederholte Anwendung dieser Argumentation sieht man, dass **alle im QR-Verfahren berechneten Matrizen** $\mathbf{A}_m$**,** $m = 0, 1, 2, \ldots$ **die gleichen Eigenwerte haben**. Insbesondere haben sie alle die gleichen Eigenwerte wie die ursprüngliche Matrix $\mathbf{A}_0 = \mathbf{A}$.

Wir lernen zunächst zwei Hilfssätze kennen, die uns nützliche Informationen über das QR-Verfahren und die QR-Zerlegung liefern.

Hilfssatz 5.14. (Eigenschaften des QR-Verfahrens)
Sei $\mathbf{A} \in \mathbb{R}^{n \times n}$*, und seien die Bezeichnungen wie in Verfahren 5.13.*

(1) Falls $\mathbf{A}_m$ *eine* ***Hessenberg-Matrix*** *ist, so ist* $\mathbf{A}_{m+1}$ ***ebenfalls eine Hessenberg-Matrix****. Insbesondere folgt, dass für eine* ***Hessenberg-Matrix*** $\mathbf{A}_0 = \mathbf{A}$ *alle im QR-Verfahren berechneten Matrizen* $\mathbf{A}_m$ ***ebenfalls Hessenberg-Matrizen*** *sind.*

(2) Falls $\mathbf{A}_m$ *eine* ***tridiagonale Matrix*** *ist, so ist* $\mathbf{A}_{m+1}$ ***ebenfalls eine tridiagonale Matrix****. Insbesondere folgt, dass für eine* ***tridiagonale Matrix*** $\mathbf{A}_0 = \mathbf{A}$ *alle im QR-Verfahren berechneten Matrizen* $\mathbf{A}_m$ ***ebenfalls tridiagonale Matrizen*** *sind.*

Erklärung: Eine quadratische Matrix $\mathbf{A}$ heißt **tridiagonal**, falls die einzigen Einträge ungleich null in $\mathbf{A}$ auf der Diagonalen, der unteren Nebendiagonalen und der oberen Nebendiagonalen auftreten. Also ist $\mathbf{A} = [a_{j,k}] \in \mathbb{R}^{n\times n}$ tridiagonal, wenn gilt $a_{j,k} = 0$ wenn $j > k+1$ oder $j+1 < k$.

Hilfssatz 5.15. (Eindeutigkeit der QR-Zerlegung)
*Sei $\mathbf{A} \in \mathbb{R}^{n\times n}$ eine reguläre Matrix. Eine Zerlegung $\mathbf{A} = \mathbf{Q}\,\mathbf{R}$ mit einer **orthogonalen Matrix** $\mathbf{Q}$ und einer **oberen Dreiecksmatrix** $\mathbf{R}$ ist bis auf die Vorzeichen der Diagonaleinträge von $\mathbf{Q}$ und $\mathbf{R}$ eindeutig bestimmt.*

Nun können wie den Satz über die Konvergenz der QR-Verfahrens formulieren.

Satz 5.16. (Konvergenz des QR-Verfahrens)
*Sei $\mathbf{A} \in \mathbb{R}^{n\times n}$ regulär. $\mathbf{A}$ habe weiter n reelle Eigenwerte $\lambda_1, \lambda_2, \ldots, \lambda_n \in \mathbb{R}$ und n linear unabhängige zugehörige reelle Eigenvektoren $\mathbf{w}_1, \mathbf{w}_2, \ldots, \mathbf{w}_n \in \mathbb{R}^n$ (also $\mathbf{A}\,\mathbf{w}_i = \lambda_i\,\mathbf{w}_i$ für $i = 1, 2, \ldots, n$). Sei $\mathbf{T} = [\mathbf{w}_1; \mathbf{w}_2; \ldots; \mathbf{w}_n] \in \mathbb{R}^{n\times n}$ die Matrix, deren Spaltenvektoren die Eigenvektoren $\mathbf{w}_1, \mathbf{w}_2, \ldots, \mathbf{w}_n$ sind. Wir nehmen an, dass ihre Inverse $\mathbf{T}^{-1}$ eine LR-Zerlegung ohne Pivot-Strategie (also ohne erforderlichen Zeilentausch) besitzt. Dann haben die im **QR-Verfahren mit der Startmatrix $\mathbf{A}_0 := \mathbf{A}$ berechneten Matrizen** $\mathbf{A}_m = [a_{ij}^{(m)}] \in \mathbb{R}^{n\times n}$ die folgenden Eigenschaften:*

(1) Die Einträge in den Matrizen $\mathbf{A}_m$ unterhalb der Diagonalen konvergieren gegen null, also $a_{i,j}^{(m)} \overset{m\to\infty}{\longrightarrow} 0$ für alle $i, j = 1, 2, \ldots, n$ mit $i > j$.

(2) Die Folgen $\left(\mathbf{A}_{2m}\right)_{m\in\mathbb{N}_0}$ und $\left(\mathbf{A}_{2m+1}\right)_{m\in\mathbb{N}_0}$ konvergieren jeweils gegen eine obere Dreiecksmatrix.

*(3) Die **Diagonaleinträge** der Matrizen $\mathbf{A}_m$ **konvergieren gegen die Eigenwerte von $\mathbf{A}$**. Genauer gilt: $a_{i,i}^{(m)} \overset{m\to\infty}{\longrightarrow} \lambda_{\pi(i)}$ für alle $i = 1, 2, \ldots, n$, wobei $\pi : \{1, 2, \ldots, n\} \to \{1, 2, \ldots, n\}$ eine Permutation der Zahlen $1, 2, \ldots, n$ ist.*

Weiter konvergieren die im QR-Verfahren berechneten Matrizen $\mathbf{Q}_m$ gegen eine orthogonale Diagonalmatrix, d.h. gegen eine Diagonalmatrix, bei der alle Einträge auf der Diagonalen 1 oder -1 sind.

Die Beweise der in diesem Teilkapitel vorgestellten Resultate, insbesondere der Beweis von Satz 5.16 sind sehr technisch aber geben wenig tiefere Einsichten in das QR-Verfahren. Wir lassen die Beweise daher weg und verweisen auf beispielsweise [7, Teilkapitel 7.6].

m	1	2	3	4	5	6	7	8
$(\mathbf{A}_{m+1})_{1,1}$	1,2222	1,3267	1,3657	1,3820	1,3904	1,3953	1,3983	1,4004
$(\mathbf{A}_{m+1})_{2,2}$	1,1056	1,1380	1,1474	1,1525	1,1554	1,1566	1,1568	1,1566
$(\mathbf{A}_{m+1})_{3,3}$	0,9069	0,8837	0,8722	0,8624	0,8555	0,8512	0,8487	0,8472
$(\mathbf{A}_{m+1})_{4,4}$	0,7652	0,6516	0,6147	0,6030	0,5988	0,5970	0,5962	0,5958

Tabelle 5.2: Diagonaleinträge der vom QR-Verfahren berechneten Matrizen $\mathbf{A}_m$, $m = 1, 2, \ldots, 8$, für Beispiel 5.17.

Betrachten wir ein numerisches Beispiel.

Beispiel 5.17. (QR-Verfahren zur Eigenwertberechnung)
Wir betrachten die folgende symmetrische Matrix

$$\mathbf{A} = \begin{bmatrix} 1 & 0 & 0{,}25 & 0{,}25 \\ 0 & 1 & 0 & 0{,}25 \\ 0{,}25 & 0 & 1 & 0 \\ 0{,}25 & 0{,}25 & 0 & 1 \end{bmatrix}.$$

Diese hat die folgenden vier verschiedenen Eigenwerte (welche man beispielsweise durch geschicktes Umformen des charakteristischen Polynoms bestimmen kann):

$$\lambda_1 = 1 + \sqrt{\frac{3+\sqrt{5}}{32}} \doteq 1{,}4045, \qquad \lambda_2 = 1 + \sqrt{\frac{3-\sqrt{5}}{32}} \doteq 1{,}1545,$$

$$\lambda_3 = 1 - \sqrt{\frac{3-\sqrt{5}}{32}} \doteq 0{,}8455, \qquad \lambda_4 = 1 - \sqrt{\frac{3+\sqrt{5}}{32}} \doteq 0{,}5955.$$

Wir führen mit Matlab die ersten 8 Iterationsschritte des QR-Verfahrens durch. In Tabelle 5.2 sind die Diagonaleintäge der Matrizen $\mathbf{A}_m$ der ersten acht Iterationsschritte in der Gleitkommadarstellung mit Exponent 0 auf vier Nachkommastellen gerundet angegeben. Wir wissen nach Satz 5.16, dass diese Diagonaleinträge gegen die Eigenwerte von $\mathbf{A}$ konvergieren. Nach 8 Iterationsschritten hat die Näherung der Eigenwerte $\lambda_1, \lambda_2, \lambda_4$ jeweils drei signifikante Ziffern, und die Näherung von λ_3 hat zwei signifikante Ziffern. ♠

KAPITEL 6

Numerische Integration

In diesem Kapitel lernen wir Verfahren zur numerischen Berechnung von Integralen, also zur **numerischen Integration** oder **Quadratur** kennen. Als Vorbereitung für aufwendigere numerische Integrationsverfahren benötigen wir einige Informationen über **Interpolation mit Polynomen**.

Warum benötigt man numerische Integrationsverfahren? Viele Integrale, wie z.B.

$$\int_0^1 e^{x^2}\,\mathrm{d}x \qquad \text{oder} \qquad \int_0^\pi x^\pi \sin\left(\sqrt{x}\right)\mathrm{d}x$$

sind nicht (mit Hilfe des Hauptsatzes der Differential- und Integralrechnung) elementar berechenbar, weil die Integranden keine (elementar berechenbare) Stammfunktion haben. In diesem Fall braucht man eine numerische Integrationsformel, um das Integral angenähert zu berechnen. – Integrale müssen in vielen Anwendungsproblemen berechnet werden. Unter anderem spielen numerische Integrationsformeln bei der numerischen Lösung gewöhnlicher Differentialgleichungen eine wichtige Rolle, wie wir noch in Kapitel 7 sehen werden.

6.1 Interpolation

In diesem Teilkapitel interessieren wir uns für die **Interpolation** einer Funktion **durch ein Polynom** passenden Grades und lernen die für theoretische Fragestellungen sehr wichtige **Interpolationsformel von Lagrange** kennen.

Wie beginnen damit, dass wir zunächst erklären, was der Begriff „Interpolation“ überhaupt bedeutet.

Problemstellung 6.1. (Interpolationsproblem)
Sei $n \in \mathbb{N}_0$. Gegeben sind $n+1$ Datenpunkte $(x_0; y_0), (x_1; y_1), \ldots, (x_n; y_n)$ in der $(x; y)$-Ebene mit ***paarweise verschiedenen*** *$x_0, x_1, \ldots, x_n$ (d.h. $x_i \neq x_j$, wenn $i \neq j$ ist). Gesucht ist eine (geeignete) Funktion g, deren Graph durch diese Punkte geht, also $g(x_i) = y_i$ für $i = 0, 1, 2, \ldots, n$ erfüllt. Man sagt dann, die Funktion g* ***interpoliert*** *die Datenpunkte $(x_i; y_i)$, $i = 0, 1, 2, \ldots, n$, und nennt g die* ***Interpolierende*** *(oder* ***Interpolante****) dieser Datenpunkte.*

In der Regel kommen die Datenpunkte $(x_i; y_i)$, $i = 0, 1, 2, \ldots, n$, vom Graphen einer (unbekannten) Funktion f, also $y_i = f(x_i)$, $i = 0, 1, 2, \ldots, n$, oder von einem Anwendungsproblem, bei dem die y_i als Funktionswerte an der Punkten x_i aufgefasst werden können. Hier sind einige Beispiele.

Beispiel 6.2. (Datensätze für Interpolation)

(a) $(x_0; y_0), (x_1; y_1), \ldots, (x_n; y_n)$ mit $y_i =$ Temperatur zum Zeitpunkt x_i.

(b) $(x_0; y_0), (x_1; y_1), \ldots, (x_n; y_n)$ mit $y_i = \cos(x_i)$, $i = 0, 1, 2, \ldots, n$.

(c) $(x_0; y_0), (x_1; y_1), \ldots, (x_n; y_n)$ mit $x_i =$ „Körpergröße gerundet auf ganze cm“ und $y_i =$ „durchschnittliches Gewicht in der Bevölkerung der BRD bei der Körpergröße x_i“.

Überlegen Sie sich einige weitere Beispiele aus dem täglichen Leben. ♠

Betrachten wir nun ein Beispiel zur Interpolation.

Beispiel 6.3. (Interpolationsprobleme)
Gegeben seien die Datenpunkte $(x_0; y_0) = (0; 1)$ und $(x_1; y_1) = \big(\ln(2); -2\big)$.

(a) Gesucht ist ein lineares Polynom $y(x) = a\,x + b$ (mit passend zu wählenden Konstanten a und b), das die gegebenen Datenpunkte interpoliert.

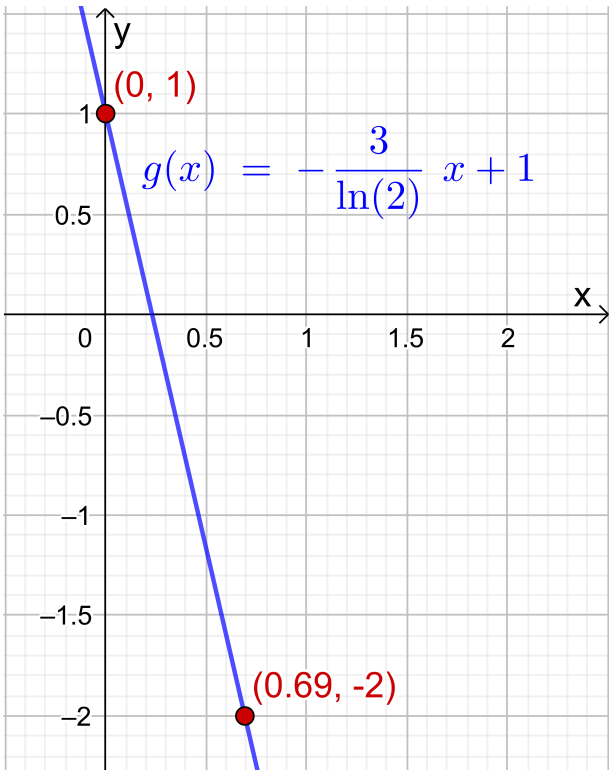
Um das Interpolationsproblem zu lösen, nutzt man die Interpolationsbedingungen $y(x_k) = y_k$ für $k = 0, 1$ aus:

$$1 = y(0) = a \cdot 0 + b \qquad \Longrightarrow \qquad b = 1,$$

$$-2 = y\big(\ln(2)\big) = a \cdot \ln(2) + b \qquad \Longrightarrow \qquad a = \frac{-2 - b}{\ln(2)} \overset{b=1}{\underset{\downarrow}{=}} a = \frac{-3}{\ln(2)}$$

Also ist das interpolierende lineare Polynom

$$y(x) = \frac{-3}{\ln(2)} \cdot x + 1.$$



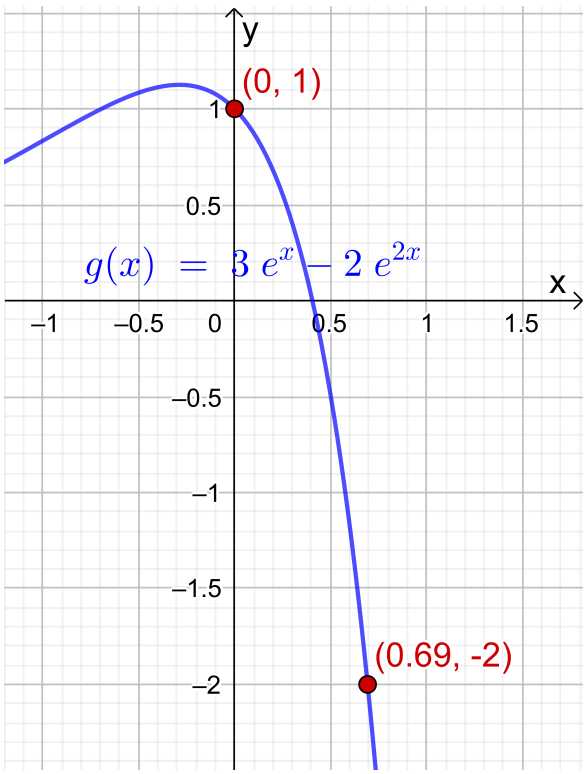


Abb. 6.1: Die Datenpunkte $(0;1)$ und $\big(\ln(2);-2\big) \doteq (0{,}69;-2)$ und die Interpolierenden aus Beispiel 6.3.

(b) Gesucht ist eine Funktion der Form $y(x) = a\,e^{x} + b\,e^{2x}$ (mit passend zu wählenden Konstanten a und b), die die gegebenen Datenpunkte interpoliert.

Um das Interpolationsproblem zu lösen, nutzt man die Interpolationsbedingungen $y(x_k) = y_k$ für $k = 0, 1$ aus:

$$1 = y(0) = a\,e^{0} + b\,e^{2\cdot 0} = a + b \qquad \Longleftrightarrow \qquad a + b = 1,$$

$$-2 = y\big(\ln(2)\big) = a\,\underbrace{e^{\ln(2)}}_{=2} + b\,e^{2\ln(2)} = a\cdot 2 + b\cdot \underbrace{(e^{\ln(2)})^{2}}_{=2^2=4} = 2\,a + 4\,b$$

$$\Longleftrightarrow \qquad 2\,a + 4\,b = -2 \qquad \Longleftrightarrow \qquad a + 2\,b = -1.$$

Wir erhalten also die beiden Gleichungen:

$$\begin{aligned} a + b &= 1 \qquad \text{(I)} \\ a + 2\,b &= -1 \qquad \text{(II)} \end{aligned}$$

Wir ziehen die Gleichung (I) von Gleichung (II) ab und erhalten:

$$a + 2\,b - (a + b) = -1 - 1 \qquad \Longleftrightarrow \qquad b = -2$$

Einsetzen von $b = -2$ in (I) liefert:

$$a + b = 1 \quad \overset{b=-2}{\Longleftrightarrow} \quad a + (-2) = 1 \quad \Longleftrightarrow \quad a = 3$$

Also ist die Interpolierende

$$y(x) = 3\,e^x - 2\,e^{2x}.$$

Die Graphen der Interpolierenden aus Teil (a) und (b) sind in Abbildung 6.1 gezeichnet. Natürlich sehen diese (abgesehen davon, dass beide Funktionen die Datenpunkte interpolieren) völlig unterschiedlich aus.

Wir beobachten: In jedem der beiden Interpolationsprobleme hatten wir genauso viele Konstanten wie Gleichungen. ♠

Im Folgenden interessieren wir uns nur für **Interpolation mit Polynomen**. Zur Vereinfachung der Notation führen wir die Bezeichnung $\mathbb{P}_n$ für die **Menge aller reellen Polynome (in einer Variablen) von Grad** $\leq n$ ein, also

$$\boxed{\mathbb{P}_n = \big\{p(x) = a_0 + a_1\,x + a_2\,x^2 + \ldots + a_n\,x^n \;:\; a_0, a_1, \ldots, a_n \in \mathbb{R}\big\}.}$$

Da ein Polynom vom Grad n von der Form

$$p(x) = a_0 + a_1\,x + a_2\,x^2 + \ldots + a_n\,x^n$$

mit $n+1$ Konstanten $a_0, a_1, a_2, \ldots, a_n \in \mathbb{R}$ ist, vermuten wir, dass man $n+1$ Datenpunkte $(x_i; y_i)$, $i = 0, 1, 2, \ldots, n$, in der $(x;y)$-Ebene mit paarweise verschiedenen $x_0, x_1, \ldots, x_n$ (also $x_i \neq x_j$, wenn $i \neq j$) durch ein Polynom vom Grad $\leq n$ interpolieren kann. Wir werden dieses nun genauer untersuchen.

In Verallgemeinerung von Beispiel 6.3 (a) kann man zeigen, dass das **lineare Interpolationspolynom** $P_1(x) = a\,x + b$ für die zwei Datenpunkte $(x_0; y_0)$ und $(x_1; y_1)$ mit $x_0 \neq x_1$ von der folgenden Form ist:

$$\boxed{P_1(x) = \frac{y_1 - y_0}{x_1 - x_0}\,x + \frac{y_0\,x_1 - x_0\,y_1}{x_1 - x_0}} \tag{6.1}$$

Wir werden diese Formel auf einem Übungsblatt nachrechnen. Dort überprüfen wir auch, dass sich (6.1) in die folgende Form bringen lässt:

$$\boxed{P_1(x) = y_0\,L_0(x) + y_1\,L_1(x) \quad \text{mit} \quad L_0(x) := \frac{x - x_1}{x_0 - x_1}\,, \quad L_1(x) := \frac{x - x_0}{x_1 - x_0}} \tag{6.2}$$

Die Funktionen L_0 und L_1 heißen die **Lagrange-Polynome vom Grad** 1 und haben die folgenden interessanten Eigenschaften:

$$L_0(x_0) = \frac{x_0 - x_1}{x_0 - x_1} = 1, \qquad L_0(x_1) = \frac{x_1 - x_1}{x_0 - x_1} = 0,$$
$$L_1(x_0) = \frac{x_0 - x_0}{x_1 - x_0} = 0, \qquad L_1(x_1) = \frac{x_1 - x_0}{x_1 - x_0} = 1.$$

Damit folgt aus $P_1(x) = y_0\, L_0(x) + y_1\, L_1(x)$ direkt

$$P_1(x_0) = y_0\, L_0(x_0) + y_1\, L_1(x_0) = y_0 \cdot 1 + y_1 \cdot 0 = y_0,$$
$$P_1(x_1) = y_0\, L_0(x_1) + y_1\, L_1(x_1) = y_0 \cdot 0 + y_1 \cdot 1 = y_1.$$

Es ist anschaulich klar, dass es zu gegebenen Datenpunkten $(x_0; y_0)$ und $(x_1; y_1)$ mit $x_0 \neq x_1$ **genau ein** lineares Interpolationspolynom gibt, da man durch zwei verschiedene Punkte in der $(x; y)$-Ebene nur genau eine Gerade legen kann.

Betrachten wir ein Beispiel.

Beispiel 6.4. (lineares Interpolationspolynom)
Gesucht ist das lineare Interpolationspolynom zu den Datenpunkten $(x_0; y_0) = (1; 1)$ und $(x_1; y_1) = (4; 2)$. Nach (6.1) ist das lineare Interpolationspolynom

$$P_1(x) = \frac{2-1}{4-1}\, x + \frac{1 \cdot 4 - 1 \cdot 2}{4-1} = \frac{1}{3}\, x + \frac{2}{3}\,.$$

Mit (6.2) können wir das lineare Interpolationspolynom auch wie folgt schreiben:

$$P_1(x) = 1 \cdot \frac{x-4}{1-4} + 2 \cdot \frac{x-1}{4-1} = \frac{x-4}{-3} + 2 \cdot \frac{x-1}{3}$$

Die gegebenen Datenpunkte stammen von der Quadratwurzelfunktion

$$f : [0; \infty[\to \mathbb{R}, \qquad f(x) = \sqrt{x},$$

denn es gilt

$$\big(1; f(1)\big) = \big(1; \sqrt{1}\big) = (1; 1) \qquad \text{und} \qquad \big(4; f(4)\big) = \big(4; \sqrt{4}\big) = (4; 2).$$

In Abbildung 6.2 ist die Quadratwurzelfunktion zusammen mit dem linearen Interpolationspolynom P_1 gezeichnet. Wir sehen, dass das lineare Interpolationspolynom P_1 die Funktion $f(x) = \sqrt{x}$ auf dem Intervall $\left[\frac{1}{2}; \frac{9}{2}\right] = [0{,}5;\ 4{,}5]$ gar nicht so schlecht annähert, aber außerhalb dieses Intervalls liefert das lineare Interpolationspolynom nur eine schlechte Näherung. So erhalten wir beispielsweise

$$P_1(3) = \frac{1}{3} \cdot 3 + \frac{2}{3} = \frac{5}{3} \doteq 1{,}6667 \qquad \text{und} \qquad \sqrt{3} \doteq 1,7321,$$

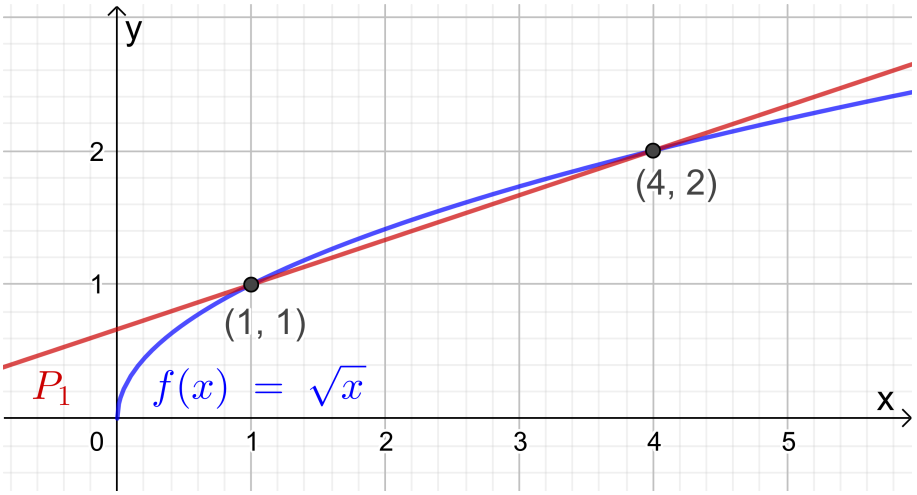


Abb. 6.2: Die Funktion $f : [0; \infty[\to \mathbb{R}$, $f(x) = \sqrt{x}$, und ihr lineares Interpolationspolynom P_1 für die Datenpunkte $(1; 1)$ und $(4; 2)$.

d.h. für $x = 3$ haben wir einen relativen Fehler von

$$\mathrm{Rel}\big(P_1(3)\big) = \frac{|P_1(3) - \sqrt{3}|}{|\sqrt{3}|} = \frac{\left|\frac{5}{3} - \sqrt{3}\right|}{\sqrt{3}} \doteq 0{,}038,$$

also einen relativen Fehler von knapp 4 %. ♠

Wir wollen nun den Fall der Interpolation von drei Datenpunkten $(x_0; y_0)$, $(x_1; y_1)$ und $(x_2; y_2)$ in der $(x; y)$-Ebene mit $x_0 \neq x_1$, $x_0 \neq x_2$ und $x_1 \neq x_2$ mit einem (höchstens) quadratischen Polynom betrachten. Das **quadratische Interpolationspolynom** $P_2 \in \mathbb{P}_2$ ist durch die folgende Formel gegeben:

$$P_2(x) = y_0\, L_0(x) + y_1\, L_1(x) + y_2\, L_2(x) \quad \text{mit} \quad L_0(x) := \frac{(x - x_1)\,(x - x_2)}{(x_0 - x_1)\,(x_0 - x_2)},$$
$$L_1(x) := \frac{(x - x_0)\,(x - x_2)}{(x_1 - x_0)\,(x_1 - x_2)}, \quad L_2(x) := \frac{(x - x_0)\,(x - x_1)}{(x_2 - x_0)\,(x_2 - x_1)}. \tag{6.3}$$

Die Funktionen L_0, L_1, L_2 heißen die **Lagrange-Polynome vom Grad** 2, und (6.3) wird als die **Interpolationsformel von Lagrange** bezeichnet.

Warum ist die Formel (6.3) **richtig?** Wir bemerken zunächst durch Inspektion, dass die Funktionen L_0, L_1, L_2 jeweils Polynome vom Grad 2 sind. Also muss P_2, definiert durch (6.3), ein Polynom vom Grad ≤ 2 sein. Weiter überprüft man

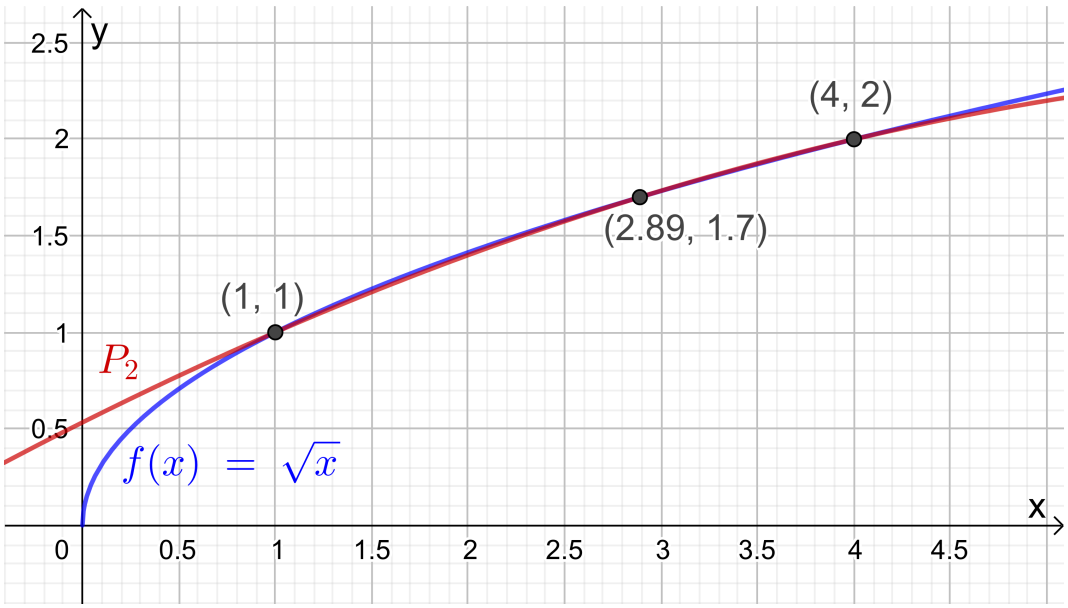


Abb. 6.3: Die Funktion $f : [0; \infty[\to \mathbb{R}$, $f(x) = \sqrt{x}$, und ihr quadratisches Interpolationspolynom P_2 für die Datenpunkte $(1; 1)$, $(4; 2)$ und $(2{,}89; 1{,}7)$.

leicht (dieses ist eine Übungsaufgabe), dass gilt

$$L_j(x_i) = \delta_{i,j} \qquad \text{für alle } i, j = 0, 1, 2, \tag{6.4}$$

wobei $\delta_{i,j}$ das **Kronecker-Delta** ist, welches wie folgt definiert ist:

$$\delta_{i,j} := \begin{cases} 1 & \text{für } i = j, \\ 0 & \text{für } i \neq j. \end{cases} \tag{6.5}$$

Dann folgt mit (6.4)

$$\begin{aligned} P_2(x_0) &= y_0\, L_0(x_0) + y_1\, L_1(x_0) + y_2\, L_2(x_0) = 1 \cdot y_0 + 0 \cdot y_1 + 0 \cdot y_2 = y_0, \\ P_2(x_1) &= y_0\, L_0(x_1) + y_1\, L_1(x_1) + y_2\, L_2(x_1) = 0 \cdot y_0 + 1 \cdot y_1 + 0 \cdot y_2 = y_1, \\ P_2(x_2) &= y_0\, L_0(x_2) + y_1\, L_1(x_2) + y_2\, L_2(x_2) = 0 \cdot y_0 + 0 \cdot y_1 + 1 \cdot y_2 = y_2, \end{aligned}$$

und wir sehen, dass P_2 die Datenpunkte $(x_0; y_0)$, $(x_1; y_1)$ und $(x_2; y_2)$ interpoliert.

Betrachten wir ein Beispiel.

Beispiel 6.5. (quadratisches Interpolationspolynom)
Gesucht ist ein (höchstens) quadratisches Polynom, das die Datenpunkte

$$(x_0; y_0) = (1; 1), \qquad (x_1; y_1) = (4; 2) \qquad \text{und} \qquad (x_2; y_2) = (2{,}89; 1{,}7)$$

interpoliert. Es gilt $(x_i; y_i) = \left(x_i; \sqrt{x_i}\right)$ für $i = 0, 1, 2$, d.h. die Datenpunkte stammen von der Quadratwurzelfunktion $f :]0; \infty[\to \mathbb{R}$, $f(x) = \sqrt{x}$.

Nach (6.3) ist ein (höchstens) quadratische Interpolationspolynom

$$P_2(x) = 1 \cdot \frac{(x-4)\,(x-2{,}89)}{(1-4)\,(1-2{,}89)} + 2 \cdot \frac{(x-1)\,(x-2{,}89)}{(4-1)\,(4-2{,}89)} + 1{,}7 \cdot \frac{(x-1)\,(x-4)}{(2{,}89-1)\,(2{,}89-4)}$$

In Abbildung 6.3 ist die Quadratwurzelfunktion zusammen mit dem (höchstens) quadratischen Interpolationspolynom gezeichnet. Verglichen mit dem linearen Interpolationspolynom (siehe Beispiel 6.4 und Abbildung 6.2) hat sich die Näherung von $f(x) = \sqrt{x}$ deutlich verbessert. Mit dem bloßen Auge sieht man auf dem Intervall [0,8; 4,4] in der Veranschaulichung der Graphen in Abbildung 6.3 fast keinen Unterschied mehr zwischen $f(x) = \sqrt{x}$ und P_2. Nun gilt in $x = 3$

$$\begin{aligned} P_2(3) &= 1 \cdot \frac{(3-4)\,(3-2{,}89)}{(1-4)\,(1-2{,}89)} + 2 \cdot \frac{(3-1)\,(3-2{,}89)}{(4-1)\,(4-2{,}89)} + 1{,}7 \cdot \frac{(3-1)\,(3-4)}{(2{,}89-1)\,(2{,}89-4)} \\ &\doteq 1{,}7334 \end{aligned}$$

d.h. für $x = 3$ haben wir einen relativen Fehler von

$$\mathrm{Rel}\big(P_2(3)\big) = \frac{\sqrt{3} - P_2(3)}{\sqrt{3}} \doteq 0{,}00078,$$

also einen relativen Fehler von knapp unter 0,08 %. (Natürlich liegt 3 dicht bei $2,98$, und für einen anderen Punkt wie $x = 2$ hätten wir vermutlich eine etwas schlechtere Näherung bekommen.) ♠

Wir gehen nun noch einen Schritt weiter und wollen $n+1$ Datenpunkte $(x_i; y_i)$, $i = 0, 1, 2, \ldots, n$, in der $(x; y)$-Ebene mit paarweise verschiedenen Punkten x_i, $i = 0, 1, 2, \ldots, n$, durch ein Polynom $P_n \in \mathbb{P}_n$ (also vom Grad $\leq n$) interpolieren. (Dass die Datenpunkte „paarweise verschieden“ sind, bedeutet, dass $x_i \neq x_j$ gilt, wenn $i \neq j$ ist.) In Analogie zu (6.2) und (6.3) bekommt man dann den folgenden Satz, den wir auch beweisen werden:

Satz 6.6. (Interpolationsformel von Lagrange)
Seien $(x_i; y_i)$, $i = 0, 1, 2, \ldots, n$, genau $n+1$ Datenpunkte mit paarweise verschiedenen $x_0, x_1, x_2, \ldots, x_n$ (d.h. $x_i \neq x_j$ wenn $i \neq j$). Dann ist

$$P_n(x) = y_0\, L_0(x) + y_1\, L_1(x) + y_2\, L_2(x) + \ldots + y_n\, L_n(x) \tag{6.6}$$

*mit den **Lagrange-Polynomen vom Grad** n (bzgl. $x_0, x_1, \ldots, x_n$)*

$$L_i(x) = \frac{(x-x_0) \cdot \ldots \cdot (x-x_{i-1}) \cdot (x-x_{i+1}) \cdot \ldots \cdot (x-x_n)}{(x_i-x_0) \cdot \ldots \cdot (x_i-x_{i-1}) \cdot (x_i-x_{i+1}) \cdot \ldots \cdot (x_i-x_n)}, \tag{6.7}$$

$i = 0, 1, 2, \ldots, n$, *ein* ***Polynom vom Grad*** $\leq n$***, das die Datenpunkte interpoliert****. Die Darstellung* (6.6) *eines interpolierenden Polynoms vom Grad* $\leq n$ *nennt man die* ***Interpolationsformel von Lagrange****.*

Schauen wir uns die Formel (6.7) für die Lagrange-Polynome genauer an: Im Zähler von L_i steht ein Polynom vom exakten Grad n, bei dem alle Linearfaktoren $(x - x_j)$, $j = 0, 1, 2, \ldots, n$, mit Ausnahme von $(x - x_i)$ multipliziert werden. Im Nenner steht eine Konstante, die man durch Multiplikation der $(x_i - x_j)$, $j = 0, 1, 2, \ldots, n$, mit Ausnahme von $(x_i - x_i)$ erhält. Nach Formel (6.7) gilt also

$$
\begin{aligned}
L_0(x) &= \frac{(x - x_1) \cdot \ldots \cdot (x - x_n)}{(x_0 - x_1) \cdot \ldots \cdot (x_0 - x_n)}\,, \\
L_1(x) &= \frac{(x - x_0) \cdot (x - x_2) \cdot \ldots \cdot (x - x_n)}{(x_1 - x_0) \cdot (x_1 - x_2) \cdot \ldots \cdot (x_1 - x_n)}\,, \\
L_2(x) &= \frac{(x - x_0) \cdot (x - x_1) \cdot (x - x_3) \cdot \ldots \cdot (x - x_n)}{(x_2 - x_0) \cdot (x_2 - x_1) \cdot (x_2 - x_3) \cdot \ldots \cdot (x_2 - x_n)}\,, \\
&\vdots \\
L_{n-1}(x) &= \frac{(x - x_0) \cdot (x - x_1) \cdot \ldots \cdot (x - x_{n-2}) \cdot (x - x_n)}{(x_{n-1} - x_0) \cdot (x_{n-1} - x_1) \cdot \ldots \cdot (x_{n-1} - x_{n-2}) \cdot (x_{n-1} - x_n)}\,, \\
L_n(x) &= \frac{(x - x_0) \cdot (x - x_1) \cdot \ldots \cdot (x - x_{n-1})}{(x_n - x_0) \cdot (x_n - x_1) \cdot \ldots \cdot (x_n - x_{n-1})}\,.
\end{aligned}
$$

Die Formeln (6.2) und (6.3) sind jeweils der Sonderfall von (6.6) und (6.7) für den Fall $n = 1$ bzw. $n = 2$. (Inspizieren Sie bitte (6.6) und (6.7) noch einmal, um sich dieses klar zu machen.)

Wir werden im Beweis von Satz 6.6 noch zeigen, dass die in (6.7) definierten Lagrange-Polynome vom Grad n die Eigenschaft

$$L_j(x_i) = \delta_{i,j} \qquad \text{für alle } i, j = 0, 1, 2, \ldots, n \tag{6.8}$$

haben, wobei $\delta_{i,j}$ das in (6.5) definierte Kronecker-Delta ist. Mit (6.8) zeigt man dann leicht, dass (6.6) die Datenpunkte interpoliert.

Betrachten wir zunächst ein Beispiel.

Beispiel 6.7. (Interpolationspolynom vom Grad ≤ 4)
Wir wollen die Cosinusfunktion $f : \mathbb{R} \to \mathbb{R}$, $f(x) = \cos(x)$, durch das Interpolationspolynom $P_4 \in \mathbb{P}_4$ (also vom Grad ≤ 4) bzgl. der folgenden fünf Datenpunkte

(mit paarweise verschiedenen x_i, $i = 0, 1, \ldots, 4$) interpolieren:

$$\begin{aligned}
(x_0; y_0) &= \left(-\tfrac{\pi}{2}; \cos\left(-\tfrac{\pi}{2}\right)\right) = \left(-\tfrac{\pi}{2}; 0\right), \\
(x_1; y_1) &= \left(-\tfrac{\pi}{4}; \cos\left(-\tfrac{\pi}{4}\right)\right) = \left(-\tfrac{\pi}{4}; \tfrac{\sqrt{2}}{2}\right), \\
(x_2; y_2) &= (0; \cos(0)) = (0; 1), \\
(x_3, y_3) &= \left(\tfrac{\pi}{4}; \cos\left(\tfrac{\pi}{4}\right)\right) = \left(\tfrac{\pi}{4}; \tfrac{\sqrt{2}}{2}\right), \\
(x_4, y_4) &= \left(\tfrac{\pi}{2}; \cos\left(\tfrac{\pi}{2}\right)\right) = \left(\tfrac{\pi}{2}; 0\right).
\end{aligned}$$

Wir stellen zuerst die Lagrange-Polynome von Grad 4 auf:

$$\begin{aligned}
L_0(x) &= \frac{\left(x + \frac{\pi}{4}\right) \cdot (x - 0) \cdot \left(x - \frac{\pi}{4}\right) \cdot \left(x - \frac{\pi}{2}\right)}{\left(-\frac{\pi}{2} + \frac{\pi}{4}\right) \cdot \left(-\frac{\pi}{2} - 0\right) \cdot \left(-\frac{\pi}{2} - \frac{\pi}{4}\right) \cdot \left(-\frac{\pi}{2} - \frac{\pi}{2}\right)} \\
&= \frac{32}{3\,\pi^4} \cdot \left(x + \frac{\pi}{4}\right) \cdot x \cdot \left(x - \frac{\pi}{4}\right) \cdot \left(x - \frac{\pi}{2}\right), \\
L_1(x) &= \frac{\left(x + \frac{\pi}{2}\right) \cdot (x - 0) \cdot \left(x - \frac{\pi}{4}\right) \cdot \left(x - \frac{\pi}{2}\right)}{\left(-\frac{\pi}{4} + \frac{\pi}{2}\right) \cdot \left(-\frac{\pi}{4} - 0\right) \cdot \left(-\frac{\pi}{4} - \frac{\pi}{4}\right) \cdot \left(-\frac{\pi}{4} - \frac{\pi}{2}\right)} \\
&= -\frac{128}{3\,\pi^4} \cdot \left(x + \frac{\pi}{2}\right) \cdot x \cdot \left(x - \frac{\pi}{4}\right) \cdot \left(x - \frac{\pi}{2}\right), \\
L_2(x) &= \frac{\left(x + \frac{\pi}{2}\right) \cdot \left(x + \frac{\pi}{4}\right) \cdot \left(x - \frac{\pi}{4}\right) \cdot \left(x - \frac{\pi}{2}\right)}{\left(0 + \frac{\pi}{4}\right) \cdot \left(0 + \frac{\pi}{2}\right) \cdot \left(0 - \frac{\pi}{4}\right) \cdot \left(0 - \frac{\pi}{2}\right)} \\
&= \frac{64}{\pi^4} \cdot \left(x + \frac{\pi}{2}\right) \cdot \left(x + \frac{\pi}{4}\right) \cdot \left(x - \frac{\pi}{4}\right) \cdot \left(x - \frac{\pi}{2}\right), \\
L_3(x) &= \frac{\left(x + \frac{\pi}{2}\right) \cdot \left(x + \frac{\pi}{4}\right) \cdot (x - 0) \cdot \left(x - \frac{\pi}{2}\right)}{\left(\frac{\pi}{4} + \frac{\pi}{2}\right) \cdot \left(\frac{\pi}{4} + \frac{\pi}{4}\right) \cdot \left(\frac{\pi}{4} - 0\right) \cdot \left(\frac{\pi}{4} - \frac{\pi}{2}\right)} \\
&= -\frac{128}{3\,\pi^4} \cdot \left(x + \frac{\pi}{2}\right) \cdot \left(x + \frac{\pi}{4}\right) \cdot x \cdot \left(x - \frac{\pi}{2}\right), \\
L_4(x) &= \frac{\left(x + \frac{\pi}{2}\right) \cdot \left(x + \frac{\pi}{4}\right) \cdot (x - 0) \cdot \left(x - \frac{\pi}{4}\right)}{\left(\frac{\pi}{2} + \frac{\pi}{2}\right) \cdot \left(\frac{\pi}{2} + \frac{\pi}{4}\right) \cdot \left(\frac{\pi}{2} - 0\right) \cdot \left(\frac{\pi}{2} - \frac{\pi}{4}\right)} \\
&= \frac{32}{3\,\pi^4} \cdot \left(x + \frac{\pi}{2}\right) \cdot \left(x + \frac{\pi}{4}\right) \cdot x \cdot \left(x - \frac{\pi}{4}\right).
\end{aligned}$$

Mit diesen ist das Interpolationspolynom vom Grad ≤ 4 gegeben durch

$$\begin{aligned}
P_4(x) &= 0 \cdot L_0(x) + \frac{\sqrt{2}}{2} \cdot L_1(x) + 1 \cdot L_2(x) + \frac{\sqrt{2}}{2} \cdot L_3(x) + 0 \cdot L_4(x) \\
&= \frac{\sqrt{2}}{2} L_1(x) + L_2(x) + \frac{\sqrt{2}}{2} L_3(x) = L_2(x) + \frac{\sqrt{2}}{2} \big(L_1(x) + L_3(x)\big).
\end{aligned}$$

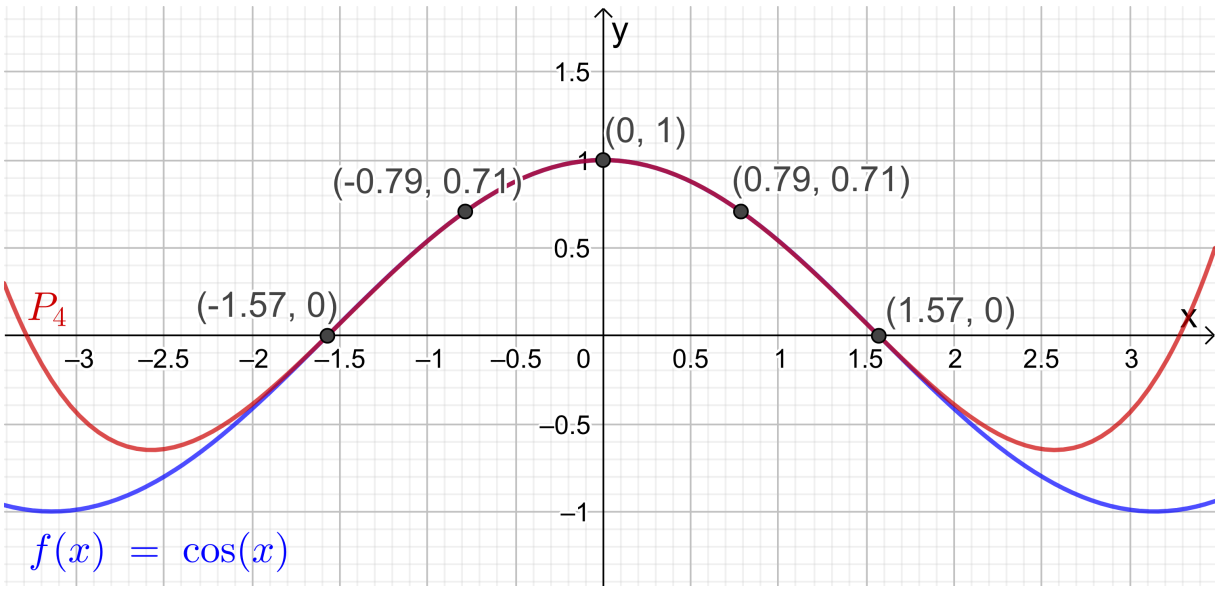


Abb. 6.4: $\cos(x)$ und sein Interpolationspolynom P_4.

Wir können P_4 in diesem konkreten Beispiel noch weiter vereinfachen:

$$\begin{aligned} L_1(x) + L_3(x) &= -\frac{128}{3\,\pi^4} \cdot \left(x + \frac{\pi}{2}\right) \cdot x \cdot \left(x - \frac{\pi}{4}\right) \cdot \left(x - \frac{\pi}{2}\right) \\ &\quad - \frac{128}{3\,\pi^4} \cdot \left(x + \frac{\pi}{2}\right) \cdot \left(x + \frac{\pi}{4}\right) \cdot x \cdot \left(x - \frac{\pi}{2}\right) \\ &= -\frac{128}{3\,\pi^4} \cdot \left(x + \frac{\pi}{2}\right) \cdot x \cdot \left(x - \frac{\pi}{2}\right) \cdot \underbrace{\left[\left(x - \frac{\pi}{4}\right) + \left(x + \frac{\pi}{4}\right)\right]}_{=2\,x} \\ &= -\frac{256}{3\,\pi^4} \cdot x^2 \cdot \left(x + \frac{\pi}{2}\right) \cdot \left(x - \frac{\pi}{2}\right) = -\frac{256}{3\,\pi^4} \cdot x^2 \cdot \left(x^2 - \frac{\pi^2}{4}\right), \end{aligned}$$

wobei wir im letzten Schritt die dritte binomische Formel verwendet haben. Ebenfalls mit der dritten binomischen Formel folgt

$$\begin{aligned} L_2(x) &= \frac{64}{\pi^4} \cdot \left(x + \frac{\pi}{2}\right) \cdot \left(x + \frac{\pi}{4}\right) \cdot \left(x - \frac{\pi}{4}\right) \cdot \left(x - \frac{\pi}{2}\right) \\ &= \frac{64}{\pi^4} \cdot \left(x + \frac{\pi}{2}\right) \cdot \left(x - \frac{\pi}{2}\right) \cdot \left(x + \frac{\pi}{4}\right) \cdot \left(x - \frac{\pi}{4}\right) \\ &= \frac{64}{\pi^4} \cdot \left(x^2 - \frac{\pi^4}{4}\right) \cdot \left(x^2 - \frac{\pi^2}{16}\right). \end{aligned}$$

Also erhalten wir für das Interpolationspolynom

$$P_4(x) = \frac{64}{\pi^4} \cdot \left(x^2 - \frac{\pi^4}{4}\right) \cdot \left(x^2 - \frac{\pi^2}{16}\right) - \frac{128\sqrt{2}}{3\,\pi^4} \cdot x^2 \cdot \left(x^2 - \frac{\pi^2}{4}\right).$$

Die Funktion $f(x) = \cos(x)$ und das Interpolationspolynom P_4 vom Grad ≤ 4 sind in Abbildung 6.4 gezeichnet. ♠

Wir beweisen nun die Interpolationsformel von Lagrange.

Beweis von Satz 6.6: Wir untersuchen zunächst die Lagrange-Polynome (6.7): Der Nenner von L_i, $i = 0, 1, 2, \ldots, n$, ist jeweils eine Konstante. Im Zähler von L_i, $i = 0, 1, 2, \ldots, n$, gibt es jeweils n lineare Faktoren; also muss L_i ein Polynom vom exakten Grad n sein. An der Formel (6.6), also an

$$P_n(x) = y_0\, L_0(x) + y_1\, L_1(x) + y_2\, L_2(x) + \ldots + y_n\, L(x),$$

sieht man direkt, dass P_n dann (als Linearkombination von $L_0, L_1, \ldots, L_n$) auch ein Polynom vom Grad $\leq n$ sein muss.

Es gilt für die Lagrange-Polynome

$$L_i(x_k) = \frac{(x_k - x_1) \cdot \ldots \cdot (x_k - x_{i-1}) \cdot (x_k - x_{i+1}) \cdot \ldots \cdot (x_k - x_n)}{(x_i - x_1) \cdot \ldots \cdot (x_i - x_{i-1}) \cdot (x_i - x_{i+1}) \cdot \ldots \cdot (x_i - x_n)} = 0$$
$$\text{für } i, k = 0, 1, 2, \ldots, n+1 \text{ mit } i \neq k,$$

weil im Zähler dann ein Term, nämlich $(x_k - x_k)$, gleich 0 ist; und es gilt

$$L_i(x_i) = \frac{(x_i - x_1) \cdot \ldots \cdot (x_i - x_{i-1}) \cdot (x_i - x_{i+1}) \cdot \ldots \cdot (x_i - x_n)}{(x_i - x_1) \cdot \ldots \cdot (x_i - x_{i-1}) \cdot (x_i - x_{i+1}) \cdot \ldots \cdot (x_i - x_n)} = 1$$
$$\text{für } i = 0, 1, 2, \ldots, n.$$

Damit folgt also

$$L_i(x_k) = \delta_{i,k} \qquad \text{für alle } i, k = 0, 1, 2, \ldots, n,$$

wobei $\delta_{i,k}$ das in (6.5) definierte Kronecker-Delta ist. Daraus folgt direkt

$$P_n(x_k) = y_0\, L_0(x_k) + y_1\, L_1(x_k) + \ldots + y_n\, L(x_k) = y_k \quad \text{für } k = 0, 1, 2, \ldots, n,$$

denn nur für $i = k$ gilt $L_i(x_k) = 1$, und für alle i mit $i \neq k$ gilt $L_i(x_k) = 0$. Also interpoliert des Polynom P_n die Datenpunkte $(x_i; y_i)$, $i = 0, 1, 2, \ldots, n$. □

Gibt es **mehr als ein Polynom in** $\mathbb{P}_n$ (also vom Grad $\leq n$), **welches gegebene Datenpunkte** $(x_i; y_i)$**,** $i = 0, 1, 2, \ldots, n$**,** in der $(x; y)$-Ebene mit paarweise verschiedenen $x_0, x_1, \ldots, x_n$ **interpoliert?** Hierüber gibt der nachfolgende Satz Auskunft.

Satz 6.8. (Existenz und Eindeutigkeit des Interpolationspolynoms)
Seien $(x_i; y_i)$, $i = 0, 1, 2, \ldots, n$, genau $n+1$ Datenpunkte in der $(x; y)$-Ebene mit paarweise verschiedenen $x_0, x_1, \ldots, x_n$ (d.h. $x_i \neq x_j$ wenn $i \neq j$). Dann existiert ***genau ein Polynom P_n in $\mathbb{P}_n$ (also vom Grad $\leq n$)*** *mit*

$$P_n(x_i) = y_i \qquad \textit{für alle } i = 0, 1, 2, \ldots, n, \tag{6.9}$$

d.h. das Interpolationspolynom $P_n \in \mathbb{P}_n$ ist ***eindeutig bestimmt****.*

Beweis von Satz 6.8: Die Interpolationsformel von Lagrange (siehe Satz 6.6) liefert uns, dass mindestens ein Polynom $P_n \in \mathbb{P}_n$ (nämlich das durch (6.6) und (6.7) gegebene Polynom) existiert, welches die $n+1$ Datenpunkte $(x_i; y_i)$, $i = 0, 1, 2, \ldots, n$, interpoliert. Wir müssen also nur noch zeigen, dass dieses Interpolationspolynom eindeutig bestimmt ist.

Dazu nehmen wir an, dass $P_n, Q_n \in \mathbb{P}_n$ jeweils $P_n(x_i) = y_i$ für $i = 0, 1, 2, \ldots, n$ bzw. $Q_n(x_i) = y_i$ für $i = 0, 1, 2, \ldots, n$ erfüllen. Die Differenzfunktion $F_n(x) = P_n(x) - Q_n(x)$ ist dann ebenfalls ein Polynom in $\mathbb{P}_n$ und sie erfüllt

$$F_n(x_i) = P_n(x_i) - Q_n(x_i) = y_i - y_i = 0 \qquad \text{für alle } i = 0, 1, 2, \ldots, n.$$

Damit hat das Polynom $F_n \in \mathbb{P}_n$ aber $n+1$ Nullstellen. Hat ein Polynom in $\mathbb{P}_n$, vom Grad $\leq n$, mehr als n Nullstellen, so muss es das Nullpolynom sein. Also folgt, dass F_n das Nullpolynom ist, d.h.

$$\begin{aligned} & F_n(x) = P_n(x) - Q_n(x) = 0 \quad \text{für alle } x \in \mathbb{R} \\ \iff \qquad & P_n(x) = Q_n(x) \quad \text{für alle } x \in \mathbb{R}. \end{aligned}$$

Also sind P_n und Q_n gleich, und das interpolierende Polynom in $\mathbb{P}_n$ ist damit eindeutig bestimmt. □

6.2 Dividierte Differenzen und die Interpolationsformel von Newton*

Dieses Teilkapitel wird nicht behandelt und ist damit auch nicht klausurrelevant. Es liefert aber wichtige Informationen über Polynominterpolation und wurde daher ins Skript aufgenommen.

*Dieses Teilkapitel wird nicht behandelt und ist damit auch nicht klausurrelevant.

Die Berechnung des interpolierenden Polynoms $P_n \in \mathbb{P}_n$ mit der Interpolationsformel von Lagrange ist relativ aufwendig, denn man muss die $n+1$ Lagrange-Polynome $L_0, L_1, \ldots, L_n$ auswerten, was jeweils $2n-1$ Multiplikationen/Divisionen und zusätzlich $2n-2$ Additionen/Subtraktionen erfordert. In der Tat sind die Lagrange-Polynome und die Interpolationsformel von Lagrange vor allem für theoretische Untersuchungen wichtig. Für die praktische Berechnung des interpolierenden Polynoms P_n verwendet man die **Interpolationsformel von Newton**, die wir in diesem Teilkapitel kennenlernen werden. Als Vorbereitung dafür benötigen wir aber zunächst **dividierte Differenzen**.

Definition 6.9. (dividierte Differenzen)
Sei $n \in \mathbb{N}$. Sei $f : [a;b] \to \mathbb{R}$ eine Funktion, und seinen $x_0, x_1, \ldots, x_n$ genau $n+1$ verschiedene Punkte in $[a;b]$. Dann definieren wir rekursiv :

(1) $f[x_i, x_{i+1}] := \dfrac{f(x_{i+1}) - f(x_i)}{x_{i+1} - x_i}$

heißt eine ***dividierte Differenz erster Ordnung*** *von f.*

(2) $f[x_{i-1}, x_i, x_{i+1}] := \dfrac{f[x_i, x_{i+1}] - f[x_{i-1}, x_i]}{x_{i+1} - x_{i-1}}$

heißt eine ***dividierte Differenz zweiter Ordnung*** *von f.*

$\vdots$

(n) $f[x_0, x_1, \ldots, x_{n-1}, x_n] := \dfrac{f[x_1, \ldots, x_n] - f[x_0, \ldots, x_{n-1}]}{x_n - x_0}$

heißt eine ***dividierte Differenz n-ter Ordnung*** *von f.*

Betrachten wir zunächst ein Beispiel.

Beispiel 6.10. (dividierte Differenzen)
Sei $f : [0;\infty[\to \mathbb{R}$, $f(x) = \sqrt{x}$, und seien $x_0 = 1$, $x_1 = 4$, $x_2 = 2{,}89$. Die zugehörigen Funktionswerte sind dann

$$f(x_0) = f(1) = \sqrt{1} = 1,$$
$$f(x_1) = f(4) = \sqrt{4} = 2,$$
$$f(x_2) = f(2{,}89) = \sqrt{2{,}89} = 1{,}7,$$

und die dividierten Differenzen erster und zweiter Ordnung lauten

$$f[x_0, x_1] = \frac{f(x_1) - f(x_0)}{x_1 - x_0} = \frac{2-1}{4-1} = \frac{1}{3} \doteq 0{,}3333,$$

$$f[x_1, x_2] = \frac{f(x_2) - f(x_1)}{x_2 - x_1} = \frac{1{,}7 - 2}{2{,}89 - 4} = \frac{-0{,}3}{-1{,}11} = \frac{30}{111} \doteq 0{,}2703,$$

$$f[x_0, x_1, x_2] = \frac{f[x_1, x_2] - f[x_0, x_1]}{x_2 - x_0} = \frac{\frac{30}{111} - \frac{1}{3}}{2{,}89 - 1} = \frac{\frac{90 - 111}{333}}{1{,}89} = \frac{\frac{-21}{333}}{1{,}89}$$

$$= \frac{\frac{-7}{111}}{1{,}89} = -\frac{700}{111 \cdot 189} = -\frac{100}{111 \cdot 27} = -\frac{100}{2997} \doteq -0{,}0334.$$

Mit nur drei Punkten können wir keine dividierten Differenzen höherer Ordnung bilden. ♠

Welche Eigenschaften haben die dividierten Differenzen? Damit befasst sich der nächste Satz.

Satz 6.11. (Eigenschaften der dividierten Differenzen)
Seien $f :]c; d[\to \mathbb{R}$ eine Funktion und $[a; b] \subseteq]c; d[$, und seien $x_0, x_1, \ldots, x_n$ genau $n + 1$ verschiedene Punkte in $[a; b]$. Dann gelten:

(1) Verändert man in in einer dividierten Differenz die Reihenfolge der Punkte (d.h. permutiert man die Punkte), so ändert sich der Wert der dividierten Differenz nicht.

(2) Aus dem Mittelwertsatz der Differentialrechnung folgt: Ist f auf $]c; d[$ stetig differenzierbar, so gibt es einen Punkt z zwischen x_i und x_{i+1}, so dass gilt

$$f[x_i, x_{i+1}] = \frac{f(x_{i+1}) - f(x_i)}{x_{i+1} - x_i} = f'(z).$$

(3) Ist f auf $]c; d[$ n-mal stetig differenzierbar, so gibt es einen Punkt z im kleinsten Intervall, das alle $x_0, x_1, \ldots, x_n$ enthält, so dass gilt

$$f[x_0, x_1, \ldots, x_{n-1}, x_n] = \frac{1}{n!} f^{(n)}(z).$$

Nach dieser Vorbereitung können wir schließlich die Interpolationsformel von Newton einführen.

Satz 6.12. (Interpolationsformel von Newton)
Sei $n \in \mathbb{N}_0$. Sei $f : [a; b] \to \mathbb{R}$ eine Funktion, und seien $x_0, x_1, \ldots, x_n$ ge-

nau $n+1$ verschiedene Punkte in $[a;b]$. Seien $\big(x_i; f(x_i)\big)$, $i = 0, 1, \ldots, n$, die zugehörigen Datenpunkte in der $(x;y)$-Ebene für die Interpolation der Funktion f. Die ***(jeweils eindeutig bestimmten) interpolierenden Polynome*** *$P_j \in \mathbb{P}_j$* ***(vom Grad $\leq j$) mit*** *$j = 0, 1, 2, \ldots, n$, die jeweils die Datenpunkte $\big(x_i; f(x_i)\big)$, $i = 0, 1, \ldots, j$, interpolieren, können mit Hilfe der dividierten Differenzen von f wie folgt berechnet werden:*

$$
\begin{aligned}
P_0(x) &= f(x_0) \\
P_1(x) &= f(x_0) + (x - x_0)\, f[x_0, x_1], \\
P_2(x) &= f(x_0) + (x - x_0)\, f[x_0, x_1] + (x - x_0)\,(x - x_1)\, f[x_0, x_1, x_2], \\
P_3(x) &= f(x_0) + (x - x_0)\, f[x_1, x_2] + (x - x_0)\,(x - x_1)\, f[x_0, x_1, x_2] \\
&\qquad + (x - x_0)\,(x - x_1)\,(x - x_2)\, f[x_0, x_1, x_2, x_3], \\
&\vdots \\
P_n(x) &= f(x_0) + (x - x_0)\, f[x_0, x_1] + \ldots \\
&\qquad + (x - x_0) \cdot \ldots \cdot (x - x_{n-2})\, f[x_0, x_1, \ldots, x_{n-1}] \\
&\qquad + (x - x_0)\,(x - x_1) \cdot \ldots \cdot (x - x_{n-1})\, f[x_0, x_1, \ldots, x_n].
\end{aligned}
$$

Warum ist die Interpolationsformel von Newton so nützlich? Es gilt für jedes $j = 0, 1, 2, \ldots, n-1$ die Rekursionsformel

$$\boxed{P_{j+1}(x) = P_j(x) + (x - x_0) \cdot \ldots \cdot (x - x_j)\, f[x_0, x_1, \ldots, x_{j+1}],} \tag{6.10}$$

denn nach Satz 6.12 gilt

$$
\begin{aligned}
P_{j+1}(x) &= \underbrace{f(x_0) + \ldots + (x - x_0) \cdot \ldots \cdot (x - x_{j-1})\, f[x_0, x_1, \ldots, x_j]}_{= P_j(x)} \\
&\qquad + (x - x_0) \cdot \ldots \cdot (x - x_j)\, f[x_0, x_1, \ldots, x_j, x_{j+1}] \\
&= P_j(x) + (x - x_0) \cdot \ldots \cdot (x - x_j)\, f[x_0, x_1, \ldots, x_n, x_{j+1}].
\end{aligned}
$$

Die Formel (6.10) besagt das Folgende: Wollen wir die Interpolation verbessern und fügen daher einen neuen Datenpunkt $\big(x_{j+1}; f(x_{j+1})\big)$ zu $\big(x_i; f(x_i)\big)$, $i = 0, 1, \ldots, j$, hinzu, so brauchen wir das Interpolationspolynom nicht neu zu berechnen, sondern müssen nur zu $P_j(x)$ einfach den Term

$$(x - x_0) \cdot \ldots \cdot (x - x_j)\, f[x_0, x_1, \ldots, x_j, x_{j+1}]$$

addieren, um $P_{j+1}(x)$ zu erhalten.

Betrachten wir zunächst ein Beispiel.

Beispiel 6.13. (Interpolationsformel von Newton)
Sei $f : [0;\infty[\to \mathbb{R}$, $f(x) = \sqrt{x}$, und seien $x_0 = 1$, $x_1 = 4$ und $x_2 = 2{,}89$. Die zugehörigen Funktionswerte sind dann $f(x_0) = 1$, $f(x_1) = 2$, $f(x_2) = 1{,}7$. Nach Beispiel 6.10 sind die dividierten Differenzen erster und zweiter Ordnung

$$f[x_0, x_1] = \frac{1}{3}\,, \qquad f[x_1, x_2] = \frac{30}{111}\,, \qquad f[x_0, x_1, x_2] = -\frac{100}{2997}\,.$$

Das konstante Interpolationspolynom $P_0 \in \mathbb{P}_0$ des Datenpunkts $(1;1)$, das lineare Interpolationspolynom $P_1 \in \mathbb{P}_1$ der Datenpunkte $(1;1)$ und $(4;2)$ bzw. das quadratische Interpolationspolynom $P_2 \in \mathbb{P}_2$ der Datenpunkte $(1;1)$, $(4;2)$ und $(2{,}89;1{,}7)$ sind also nach der Interpolationsformel von Newton gegeben durch

$$\begin{aligned} P_0(x) &= f(x_0) = 1. \\ P_1(x) &= P_0(x) + (x - x_0)\, f[x_0, x_1] = 1 + \frac{1}{3}\,(x-1) = \frac{1}{3}\,x + \frac{2}{3}\,, \qquad (6.11) \\ P_2(x) &= P_1(x) + (x - x_0)\,(x - x_1)\, f[x_0, x_1, x_2] \\ &= 1 + \frac{1}{3}\,(x-1) - \frac{100}{2997}\,(x-1)\,(x-4). \qquad (6.12) \end{aligned}$$

Natürlich ist (6.11) identisch mit der Formel für das lineare Interpolationspolynom, die wir in Beispiel 6.4 mit der Interpolationsformel von Lagrange erhalten haben. Ebenso liefert (6.12) das gleiche quadratische Interpolationspolynom, welches in Beispiel 6.5 berechnet wurde; allerdings ist dieses (ohne Vereinfachungen) nicht offensichtlich. ♠

Was passiert in Satz 6.12 und Definition 6.9, wenn die Datenpunkte $(x_i; y_i)$, $i = 0, 1, 2, \ldots, n$, nur als Punkte in der Ebene gegeben sind, die eine unbekannte Funktion (angenähert) beschreiben, und nicht als Punkte auf dem Graphen einer bekannten Funktion f? Natürlich funktioniert die Berechnung der Interpolierenden ganz analog. Dieses wird in der nächsten Bemerkung erklärt.

Bemerkung 6.14. (dividierte Differenzen für $(x_i; y_i)$, $i = 0, 1, \ldots, n$)
Sei $n \in \mathbb{N}_0$, und seien $(x_i; y_i)$, $i = 0, 1, \ldots, n$, genau $n+1$ Datenpunkte in der $(x;y)$-Ebene mit paarweise verschiedenen $x_0, x_1, \ldots, x_n$.

(1) Dann berechnen wir die **dividierten Differenzen** wie folgt:

$$y[x_i, x_{i+1}] := \frac{y_{i+1} - y_i}{x_{i+1} - x_i}\,, \qquad i = 0, 1, \ldots, n-1,$$

$$y[x_{i-1}, x_i, x_{i+1}] := \frac{y[x_i, x_{i+1}] - y[x_{i-1}, x_i]}{x_{i+1} - x_{i-1}}, \qquad i = 1, \ldots, n-1,$$

$$\vdots$$

$$y[x_0, x_1, \ldots, x_{n-1}, x_n] := \frac{y[x_1, \ldots, x_n] - y[x_0, \ldots, x_{n-1}]}{x_n - x_0}.$$

(2) Dann werden die **interpolierenden Polynome** $P_j \in \mathbb{P}_j$ (vom Grad $\leq j$) mit $j = 0, 1, 2, \ldots, n$, die jeweils die Datenpunkte $\big(x_i; y_i\big)$, $i = 0, 1, \ldots, j$, interpolieren, wie folgt berechnet:

$$\begin{aligned}
P_0(x) &= y_0 \\
P_1(x) &= y_0 + (x - x_0)\, y[x_0, x_1], \\
P_2(x) &= y_0 + (x - x_0)\, y[x_0, x_1] + (x - x_0)\,(x - x_1)\, y[x_0, x_1, x_2], \\
P_3(x) &= y_0 + (x - x_0)\, y[x_1, x_2] + (x - x_0)\,(x - x_1)\, y[x_0, x_1, x_2] \\
&\quad + (x - x_0)\,(x - x_1)\,(x - x_2)\, y[x_0, x_1, x_2, x_3], \\
&\vdots \\
P_n(x) &= y_0 + (x - x_0)\, y[x_0, x_1] + \ldots \\
&\quad + (x - x_0) \cdot \ldots \cdot (x - x_{n-2})\, y[x_0, x_1, \ldots, x_{n-1}] \\
&\quad + (x - x_0)\,(x - x_1) \cdot \ldots \cdot (x - x_{n-1})\, y[x_0, x_1, \ldots, x_n].
\end{aligned}$$

(3) Auch hier gilt für jedes $j = 0, 1, \ldots, n-1$ die **rekursive Beziehung**

$$P_{j+1}(x) = P_j(x) + (x - x_0) \cdot \ldots \cdot (x - x_j)\, y[x_0, x_1, \ldots, x_{j+1}].$$

6.3 Der Fehler der Polynominterpolation*

Dieses Teilkapitel wird nicht behandelt und ist damit auch nicht klausurrelevant. Es liefert aber wichtige Informationen über Polynominterpolation und wurde daher ins Skript aufgenommen. Ein Resultat dieses Teilkapitels wird auch benötigt, um die Konvergenzordnung der Trapezregel für numerische Integration in Teilkapitel 6.4 zu beweisen.

Sei f eine $(n+1)$-mal stetig differenzierbare Funktion, und seien $x_0, x_1, \ldots, x_n \in \mathbb{R}$

*Dieses Teilkapitel wird nicht behandelt und ist damit auch nicht klausurrelevant.

paarweise verschieden. Wir interessieren uns nun dafür, **wie gut das interpolierende Polynom** $P_n \in \mathbb{P}_n$ der Datenpunkte $\big(x_i; f(x_i)\big)$, $i = 0, 1, 2, \ldots, n$, **die Funktion** f auf einen geeigneten Intervall, welches $x_0, x_1, \ldots, x_n$ enthält, **approximiert d.h. annähert**. Zunächst lernen wir den folgenden Satz kennen.

Satz 6.15. (Interpolationsfehler)
*Sei $f :]c; d[\to \mathbb{R}$ eine $(n+1)$-mal stetig differenzierbare Funktion, sei $[a; b] \subseteq]c; d[$, und seien $x_0, x_1, \ldots, x_n \in [a; b]$ paarweise verschieden. Sei $P_n \in \mathbb{P}_n$ das **interpolierende Polynom der Datenpunkte** $\big(x_i; f(x_i)\big)$, $i = 0, 1, 2, \ldots, n$. Dann gibt es zu jedem $x \in [a; b]$ einen Punkt c_x zwischen dem Minimum und Maximum von $x_0, x_1, \ldots, x_n$ und x, so dass der **Fehler der Näherung** $P_n(x)$ **des Funktionswertes** $f(x)$ durch*

$$P_n(x) - f(x) = - \frac{(x - x_0)\,(x - x_1) \cdot \ldots \cdot (x - x_n)}{(n+1)!}\, f^{(n+1)}(c_x) \qquad (6.13)$$

gegeben ist.

Betrachten wir zwei Beispiele, um uns klar zu machen, was (6.13) bedeutet.

Beispiel 6.16. (Interpolationsfehler bei linearer Interpolation)
Sei $f : \mathbb{R} \to \mathbb{R}$, $f(x) = e^x$, und seien $x_0, x_1 \in [0; 1]$ mit $0 \leq x_0 < x_1 \leq 1$. Nach (6.13) gibt es dann zu jedem $x \in [0; 1]$ ein c_x zwischen dem Minimum und dem Maximum von x_0, x_1 und x mit

$$P_1(x) - e^x = - \frac{(x - x_0)\,(x - x_1)}{2!}\, f''(c_x) = - \frac{(x - x_0)\,(x - x_1)}{2}\, e^{c_x}, \qquad (6.14)$$

wobei wir genutzt haben, dass $f'(x) = e^x$ und $f''(x) = e^x$ ist, und wobei P_1 das lineare Interpolationspolynom bzgl. der Datenpunkte $(x_0; e^{x_0}), (x_1; e^{x_1})$ ist.

Um den Fehler (6.14) im Folgenden weiter zu untersuchen, nehmen wir ab jetzt an, dass $x_0 < x < x_1$ gilt. Dann folgt mit $-(x - x_1) = x_1 - x$ aus (6.14), dass

$$P_1(x) - e^x = \frac{(x - x_0)\,(x_1 - x)}{2}\, e^{c_x}, \qquad (6.15)$$

wobei c_x nun in $[x_0; x_1]$ liegen muss (da $x_0 = \min\{x_0, x_1, x\}$, $x_1 = \max\{x_0, x_1, x\}$). Wegen $x - x_0 > 0$ und $x_1 - x > 0$ und $e^{c_x} > 0$ sehen wir an (6.15), dass der Interpolationsfehler immer positiv ist. Falls $[x_0; x_1]$ ein sehr kleines Intervall ist, so ist e^x dort annähernd konstant. Der Fehler (6.15) verhält sich dann annähernd wie ein quadratisches Polynom.

Wir schätzen den Fehler (6.15) nun weiter ab. Wenn wir den Absolutbetrag auf beiden Seiten von (6.15) anwenden, erhalten wir

$$|P_1(x) - e^x| = \left| \frac{(x - x_0)(x_1 - x)}{2} e^{c_x} \right| = \frac{(x - x_0)(x_1 - x)}{2} e^{c_x}. \tag{6.16}$$

Wegen dem streng monoton wachsenden Wachstumsverhalten der Exponentialfunktion folgt aus $x_0 \leq c_x \leq x_1$, dass $e^{x_0} \leq e^{c_x} \leq e^{x_1}$. Somit folgt aus (6.16)

$$\frac{(x - x_0)(x_1 - x)}{2} e^{x_0} \leq |P_1(x) - e^x| \leq \frac{(x - x_0)(x_1 - x)}{2} e^{x_1}. \tag{6.17}$$

Die untere und die obere Schranke für den Fehler in (6.17) hängen immer noch von x ab. Um aus (6.17) eine Abschätzung herzuleiten, die von x unabhängig ist, nutzen wir, dass gilt

$$\max_{x \in [x_0; x_1]} \frac{(x - x_0)(x_1 - x)}{2} = \frac{h^2}{8} \qquad \text{mit} \qquad h = x_1 - x_0. \tag{6.18}$$

Dieses folgt daraus, dass $(x - x_0)(x_1 - x)$ eine nach unten geöffnete Parabel mir den Nullstellen x_0, x_1 ist. Das globale Maximum in $[x_0; x_1]$ muss dann in der Mitte zwischen seinen Nullstellen, also bei $\frac{1}{2}(x_1 + x_0)$ auftreten. Damit erhalten wir für $x = \frac{1}{2}(x_1 + x_0)$ also $x - x_0 = \frac{1}{2}(x_1 + x_0) - x_0 = \frac{1}{2}(x_1 - x_0) = \frac{h}{2}$ und $x_1 - x = x_1 - \frac{1}{2}(x_1 + x_0) = \frac{1}{2}(x_1 - x_0) = \frac{h}{2}$, woraus (6.18) direkt folgt.

Nutzt man (6.18) für die obere Schranke in (6.17) aus, so gilt mit $h = x_1 - x_0$

$$|P_1(x) - e^x| \leq \left(\max_{x \in [x_0; x_1]} \frac{(x - x_0)(x_1 - x)}{2} \right) e^{x_1} = \frac{h^2}{8} e^{x_1}. \tag{6.19}$$

Da wir nur $x_0, x_1 \in [0; 1]$ betrachten, können wir e^{x_1} weiter durch $e^{x_1} \leq e^1 = e$ abschätzen und es folgt mit $h = x_1 - x_0$

$$|P_1(x) - e^x| \leq \frac{e}{8} h^2 \qquad \text{für alle } x \text{ mit } 0 \leq x_0 \leq x \leq x_1 \leq 1. \tag{6.20}$$

Betrachten wir ein Zahlenbeispiel: Seien $x_0 = 0{,}82$ und $x_1 = 0{,}83$. Dann sind (mit Rundung auf eine Gleitkommadarstellung mit 7-stelliger Mantisse)

$$f(x_0) = e^{0{,}82} \doteq 2{,}270500, \qquad f(x_1) = e^{0{,}83} \doteq 2{,}293319,$$

und die lineare Interpolierende ist (nach der Interpolationsformel von Lagrange)

$$\begin{aligned} P_1(x) &\doteq 2{,}270500 \cdot \frac{x - 0{,}83}{0{,}082 - 0{,}83} + 2{,}293319 \cdot \frac{x - 0{,}82}{0{,}083 - 0{,}82} \\ &= \frac{2{,}270500 \cdot (0{,}83 - x) + 2{,}293319 \cdot (x - 0{,}82)}{0{,}01}. \end{aligned}$$

Für $x = 0{,}826$ finden wir die Näherung

$$P_1(0{,}826) \doteq 2{,}284191.$$

Der wahre Wert ist

$$f(0{,}826) = e^{0{,}826} \doteq 2{,}284164,$$

und der Interpolationsfehler ist somit

$$P_1(0{,}826) - e^{0{,}826} \doteq 2{,}284191 - 2{,}284164 = 0{,}0000274 = 2{,}74 \cdot 10^{-5}. \quad (6.21)$$

Laut (6.20) sollte mit $h = x_1 - x_0 = 0{,}83 - 0{,}82 = 0{,}01$ gelten

$$|P_1(0{,}826) - e^{0{,}826}| \leq \frac{e}{8} \cdot (0{,}01)^2 \doteq 0{,}0000340 = 3{,}40 \cdot 10^{-5}, \quad (6.22)$$

und wir sehen, dass der Interpolationsfehler (6.21) innerhalb der durch (6.22) gegebenen absoluten Fehlerschranke liegt. ♠

Beispiel 6.17. (Interpolationsfehler bei quadratischer Interpolation)
Sei $f : \mathbb{R} \to \mathbb{R}$, $f(x) = e^x$, und seien $x_0, x_1, x_2 \in [0;1]$ paarweise verschieden. Nach (6.13) gibt es dann zu jedem $x \in [0;1]$ ein c_x zwischen dem Minimum und dem Maximum von x_0, x_1, x_2 und x mit

$$\begin{aligned} P_2(x) - e^x &= -\frac{(x-x_0)\,(x-x_1)\,(x-x_2)}{3!}\, f'''(c_x) \\ &= -\frac{(x-x_0)\,(x-x_1)\,(x-x_2)}{6}\, e^{c_x}, \end{aligned} \quad (6.23)$$

wobei wir genutzt haben, dass $f'(x) = e^x$, $f'(x) = e^x$, $f'''(x) = e^x$ ist, und wobei $P_2 \in \mathbb{P}_2$ das (höchstens) quadratische Interpolationspolynom bzgl. der Datenpunkte $(x_0; e^{x_0})$, $(x_1; e^{x_1})$, $(x_2; e^{x_2})$ ist.

Wir nehmen nun an, dass $0 \leq x_0 < x_1 < x_2 \leq 1$ gilt und dass die drei Punkte x_0, x_1, x_2 jeweils gleiche Abstände zum Nachbarpunkt haben, also $h = x_2 - x_1 = x_1 - x_0$. Erfüllt x nun $0 \leq x_0 < x < x_2 \leq 1$. Dann liefert das Anwenden des Absolutbetrags auf (6.23)

$$\begin{aligned} |P_2(x) - e^x| &= \left| -\frac{(x-x_0)\,(x-x_1)\,(x-x_2)}{6}\, e^{c_x} \right| \\ &\leq \left| \frac{(x-x_0)\,(x-x_1)\,(x-x_2)}{6} \right| e, \end{aligned} \quad (6.24)$$

wobei wir im zweiten Schritt $0 < e^{c_x} < e^1 = e$ genutzt haben. (Dieses folgt, weil c_x in $[x_0; x_2]$ und damit in $[0;1]$ liegt und damit $e^{c_x} \leq e^1 = e$ gilt, da die Exponentialfunktion streng monoton wachsend ist.)

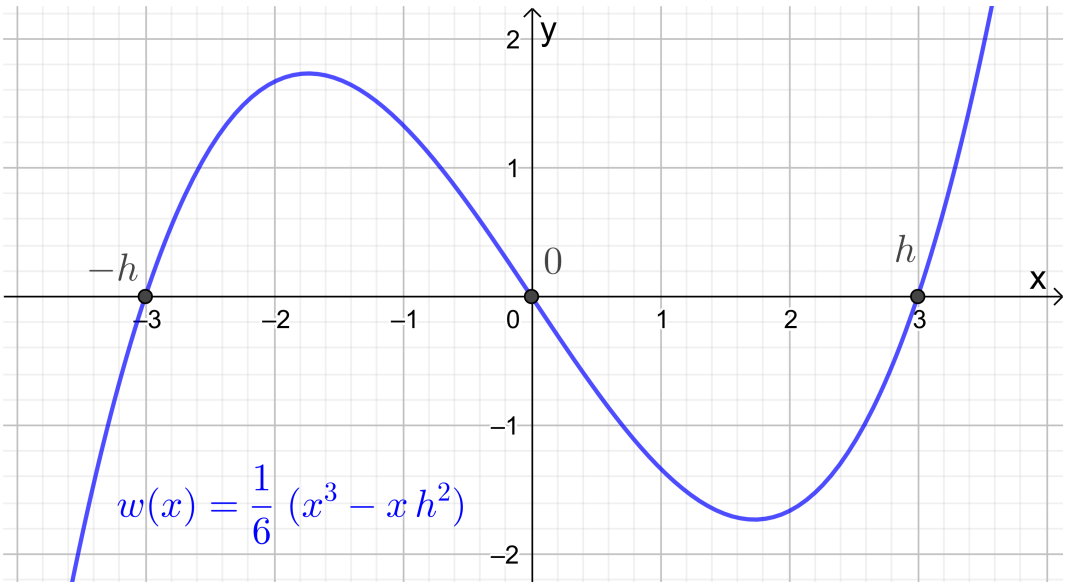


Abb. 6.5: Die Funktion $w(x) = \frac{(x+h)\,x\,(x-h)}{6} = \frac{x^3 - x\,h^2}{6}$ mit $h = 3$.

Um den Fehler weiter abzuschätzen, nutzen wir, dass gilt

$$\max_{x\in[x_0;x_2]} \left| \frac{(x-x_0)\,(x-x_1)\,(x-x_2)}{6} \right| = \frac{h^3}{9\sqrt{3}}\,. \tag{6.25}$$

Diese Abschätzung folgt, indem man zunächst x durch $x + x_1$ ersetzt, also

$$\begin{aligned} w(x) &= \frac{(x+x_1-x_0)\,(x+x_1-x_1)\,(x+x_1-x_2)}{6} \\ &= \frac{(x+h)\,x\,(x-h)}{6} = \frac{1}{6}\,x\,(x^2-h^2) = \frac{1}{6}\,(x^3 - x\,h^2)\,, \end{aligned}$$

wobei wir $x_1 - x_0 = x_2 - x_1 = h$ und danach die dritte binomische Formel $(x+h)\,(x-h) = x^2 - h^2$ genutzt haben. Das Intervall $[x_0; x_2]$ geht dann über in das Intervall $[x_0 - x_1; x_2 - x_1] = [-h; h]$. Mit

$$0 = w'(x) = \frac{3\,x^2 - h^2}{6} = \frac{1}{2}\left(x^2 - \frac{h^2}{3}\right) = \frac{1}{2}\left(x + \frac{h}{\sqrt{3}}\right)\left(x - \frac{h}{\sqrt{3}}\right)$$

sehen wir unter Berücksichtigung der Vorzeichenwechsel der Ableitung w', dass w auf $[-h; h]$ in $-\frac{h}{\sqrt{3}}$ sein globales Maximum und in $\frac{h}{\sqrt{3}}$ sein globales Minimum annimmt (siehe auch den Graphen in Abbildung 6.5). (Eigentlich folgt aus den Vorzeichenwechseln der Ableitung nur, dass lokale Extrema vorliegen, aber zusammen mit $w(-h) = w(h) = 0$ kann man folgern, dass dieses auch globale Extrema sind.) Also folgt

$$\max_{x\in[x_0;x_2]} \left| \frac{(x-x_0)\,(x-x_1)\,(x-x_2)}{6} \right| = \max_{x\in[-h;h]} \left| \frac{1}{6}\,x\,(x^2-h^2) \right|$$

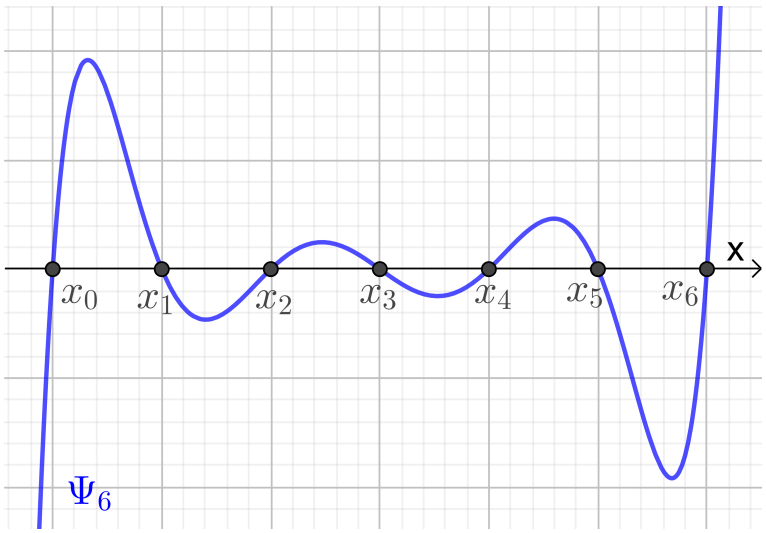


Abb. 6.6: Die Funktion Ψ_6 (siehe (6.28)) für $n+1 = 7$ Punkte $x_0 < x_1 < \ldots < x_6$ mit gleichen Abständen.

$$= \left| \frac{1}{6} \left(\pm \frac{h}{\sqrt{3}} \right) \left(\left(\pm \frac{h}{\sqrt{3}} \right)^2 - h^2 \right) \right| = \frac{1}{6} \left| \pm \frac{h}{\sqrt{3}} \left(-\frac{2\,h^2}{3} \right) \right| = \frac{1}{9\sqrt{3}}\, h^3.$$

Wenden wir (6.25) in (6.24) an, so erhalten wir für $x \in [x_0; x_2]$

$$|P_2(x) - e^x| \leq \left| \frac{(x - x_0)\,(x - x_1)\,(x - x_2)}{6} \right| e \leq \frac{e}{9\sqrt{3}}\, h^3 \doteq 0{,}174 \cdot h^3. \qquad (6.26)$$

Ist beispielsweise $h = 0{,}01$, dann gilt für $x \in [x_0; x_2] \subseteq [0; 1]$ mit $x_2 - x_1 = x_1 - x_0 = h = 0{,}01$

$$|P_2(x) - e^x| \leq 0{,}174 \cdot (0{,}01)^3 = 1{,}74 \cdot 10^{-7}. \qquad (6.27)$$

Vergleichen wir mit der Abschätzung (6.22) für den Interpolationsfehler der linearen Interpolierenden mit $x_1 - x_0 = h = 0{,}01$, so sehen wir, dass sich in (6.27) der Interpolationsfehler bei dem quadratischen Interpolationspolynom ungefähr um den Faktor $0{,}5 \cdot 10^{-2}$ verkleinert hat. ♠

In der Darstellung (6.13) des Interpolationsfehlers tritt das Polynom

$$\Psi_n(x) = (x - x_0)\,(x - x_1) \cdot \ldots \cdot (x - x_n) \qquad (6.28)$$

vom Grad $n + 1$ auf. Dieses ist der wichtigste Faktor, wenn wir den Fehler betrachten. In Abbildung 6.6 haben wir Ψ_6 für paarweise verschiedene Punkte $x_0 < x_1 < \ldots < x_6$ gezeichnet, die jeweils den gleichen Abstand haben. Man sieht direkt, dass $\Psi_6(x)$ und damit auch der Interpolationsfehler massiv größer

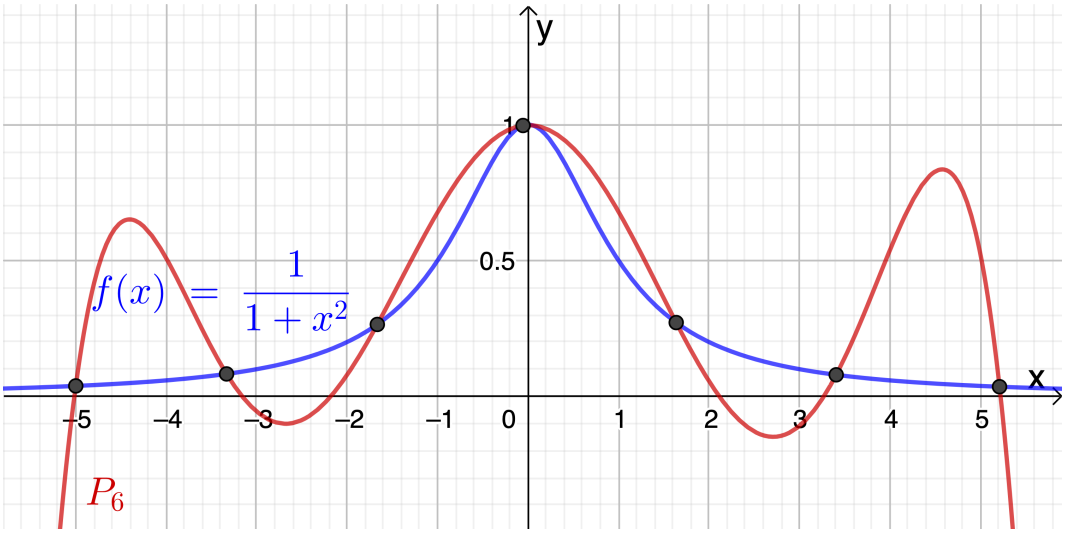


Abb. 6.7: Interpolierendes Polynom P_6 von $f(x) = \frac{1}{1+x^2}$ für Datenpunkte vom Graphen von f mit $-5 = x_0 < x_1 < \ldots < x_6 = 5$ mit gleichen Abstand.

werden, wenn man sich den Enden des Intervalls $[x_0; x_6]$ nähert. Wir vermuten daher, dass Interpolationspunkte mit gleichen Abständen vermutlich keine besonders gute Wahl sind.

Zum Abschluss betrachten wir noch ein Beispiel, an dem man sieht, dass das interpolierende Polynom P_n einer Funktion f bzgl. der Datenpunkte $\big(x_i; f(x_i)\big)$, $i = 0, 1, \ldots, n$, (mit paarweise verschiedenen $x_0, x_1, \ldots, x_n$) mit wachsendem n nicht immer gegen die Funktion f strebt.

Beispiel 6.18. (interpolierendes Polynom strebt nicht gegen f)
Gegeben sei die rationale Funktion

$$f : \mathbb{R} \to \mathbb{R}, \qquad f(x) = \frac{1}{1+x^2}\,.$$

Wir betrachten das interpolierende Polynom P_n bzgl. der Datenpunkte $\big(x_i; f(x_i)\big)$, $i = 0, 1, \ldots, n$, wobei die $-5 = x_0 < x_1 < \ldots < x_n = 5$ gleiche Abstände haben, also

$$x_i = -5 + i\,h, \qquad i = 0, 1, \ldots, n, \qquad \text{wobei} \quad h = \frac{10}{n}\,.$$

Dann strebt $P_n(x)$ in vielen Punkten $x \in [-5; 5]$ mit wachsendem n nicht gegen $f(x)$. Insbesondere für x mit $|x| > 4$ (also $x < -4$ oder $x > 4$) ist die Annäherung von $f(x)$ durch $P_n(x)$ sehr schlecht. In Abbildung 6.7 ist dieses für P_6 illustriert. Mit wachsendem n wird dieser Effekt noch schlimmer. ♠

6.4 Elementare Quadraturformeln: Trapezregel

Die Grundidee zur Konstruktion **numerischer Integrationsformeln** oder **Quadraturformeln** ist, den **Integranden** f im bestimmten Integral

$$I(f) := \int_a^b f(x)\,\mathrm{d}x \qquad \text{mit } f \in \mathcal{C}\big([a;b]\big) \tag{6.29}$$

durch eine geeignete Approximation zu ersetzen, die leicht exakt zu integrieren ist. Wir beginnen in diesem Teilkapitel mit der Herleitung der einfachsten elementaren Quadraturformel, nämlich der Trapezregel.

Um die Trapezregel zu bekommen, ersetzen wir den Integranden f in (6.29) durch sein lineares Interpolationspolynom $P_1 \in \mathbb{P}_1$ bzgl. der Datenpunkte $\big(a; f(a)\big)$ und $\big(b; f(b)\big)$ von f an den Intervallenden:

$$P_1(x) = f(a)\,\frac{x-b}{a-b} + f(b)\,\frac{x-a}{b-a} \tag{6.30}$$

Einsetzen von (6.30) in (6.29) und anschließendes Berechnen des Integrals über das Interpolationspolynom P_1 liefern

$$\begin{aligned}\int_a^b f(x)\,\mathrm{d}x \approx \int_a^b P_1(x)\,\mathrm{d}x &= \int_a^b \left(f(a)\,\frac{x-b}{a-b} + f(b)\,\frac{x-a}{b-a} \right) \mathrm{d}x \\ &= \left[f(a)\,\frac{(x-b)^2}{2\,(a-b)} + f(b)\,\frac{(x-a)^2}{2\,(b-a)} \right]_{x=a}^{x=b} \\ &= f(b)\,\frac{b-a}{2} - f(a)\,\frac{a-b}{2} = \frac{f(a)+f(b)}{2}\,(b-a).\end{aligned}$$

Wir erhalten also die sogenannte **Trapezregel**

$$\boxed{T_1(f) := (b-a)\,\frac{f(a)+f(b)}{2}\,,}$$

die nun (ohne vorherige Berechnung des linearen Interpolationspolynoms) direkt bei Kenntnis der Funktionswerte $f(a)$ und $f(b)$ angewendet werden kann.

Verfahren 6.19. (Trapezregel)
Sei $f : [a;b] \to \mathbb{R}$ eine stetige Funktion. Die ***Trapezregel***

$$T_1(f) := (b-a)\,\frac{f(a)+f(b)}{2} \tag{6.31}$$

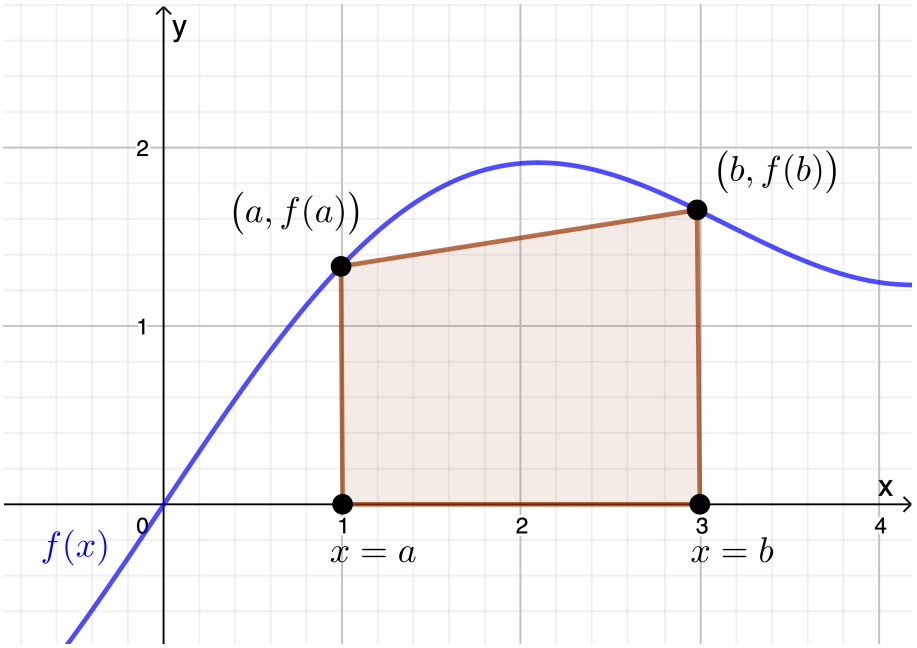


Abb. 6.8: Veranschaulichung der Trapezregel T_1 aus Verfahren 6.19.

liefert eine Nährung für das Integral $I(f) = \int_a^b f(x)\,\mathrm{d}x$.

Die Trapezregel hat eine geometrische Anschauung, der sie Ihren Namen verdankt (siehe Abbildung 6.8): Die Fläche zwischen dem Graphen des linearen Interpolationspolynoms $P_1(x)$ und der x-Achse von $x = a$ bis $x = b$ bildet ein Trapez.

Betrachten wir ein Beispiel.

Beispiel 6.20. (Trapezregel)
Wir wollen das Integral

$$I(f) = \int_0^1 \frac{1}{1+x}\,\mathrm{d}x$$

mit der Trapezregel angenähert berechnen:

$$T_1(f) = (1-0)\,\frac{f(0)+f(1)}{2} = \frac{1}{2}\left(\frac{1}{1+0} + \frac{1}{1+1}\right) = \frac{3}{4} = 0{,}75.$$

Wir berechnen das Integral direkt

$$I(f) = \int_0^1 \frac{1}{1+x}\,\mathrm{d}x = \Big[\ln(1+x)\Big]_{x=0}^{x=1} = \ln(2) - \ln(1) = \ln(2) \doteq 0{,}6931471806.$$

Der absolute Fehler der Näherung ist $|T_1(f) - I(f)| = \left|\frac{3}{4} - \ln(2)\right| \doteq 0{,}0569$. ♠

Hilfssatz 6.21. (Trapezregel ist exakt für Polynome vom Grad ≤ 1)
*Für jedes Polynom $p_1(x) = c\,x + d$ mit Konstanten $c, d \in \mathbb{R}$ gilt für die durch (6.31) definierte Trapezregel $T_1(p_1) = I(p_1)$, d.h. die Trapezregel **integriert Polynome von Grad** ≤ 1 exakt.*

Beweis von Hilfssatz 6.21: Dieses zeigt man, indem man $T_1(p_1)$ und $I(p_1)$ konkret berechnet und umformt, bis man sieht, dass sie gleich sind. – Alternativ kann man dieses auch mit Hilfe der Eigenschaften des linearen Interpolationspolynoms begründen. □

Wie kann man mit Hilfe der Trapezregel eine bessere numerische Integrationsformel bekommen? Man könnte das **Integrationsintervall $[a;b]$ in mehrere gleichlange Teilintervalle zerlegen** und dann **auf jedem Teilintervall die Trapezregel nutzen**. Betrachten wir dieses zunächst für ein Beispiel.

Beispiel 6.22. („zusammengesetzte“ Trapezregel)
Wir wollen das Integral

$$I(f) = \int_0^1 \frac{1}{1+x}\,\mathrm{d}x = \ln(2) \doteq 0{,}6931471806.$$

in zwei Teilintegrale über $\left[0;\frac{1}{2}\right]$ und $\left[\frac{1}{2};1\right]$ zerlegen und diese dann jeweils mit der Trapezregel angenähert berechnen:

$$I(f) = \int_0^1 \frac{1}{1+x}\,\mathrm{d}x = \int_0^{1/2} \frac{1}{1+x}\,\mathrm{d}x + \int_{1/2}^1 \frac{1}{1+x}\,\mathrm{d}x,$$

$$\begin{aligned} T_2(f) &= \left(\frac{1}{2} - 0\right) \frac{f(0) + f\left(\frac{1}{2}\right)}{2} + \left(1 - \frac{1}{2}\right) \frac{f\left(\frac{1}{2}\right) + f(1)}{2} \\ &= \frac{1}{2} \cdot \frac{1 + \frac{2}{3}}{2} + \frac{1}{2} \cdot \frac{\frac{2}{3} + \frac{1}{2}}{2} = \frac{1}{4} \cdot \frac{5}{3} + \frac{1}{4} \cdot \frac{7}{6} = \frac{17}{24} \doteq 0{,}70833 \end{aligned}$$

Der absolute Fehler der Näherung ist $|T_2(f) - I(f)| = \left|\frac{17}{24} - \ln(2)\right| \doteq 0{,}0152$. Gegenüber dem Ergebnis $T_1(f)$ aus Beispiel 6.20 beträgt der absolute Fehler von $T_2(f)$ ungefähr nur $\frac{1}{4}$ des absoluten Fehlers von $T_1(f)$. ♠

Wir wollen die Idee aus dem letzten Beispiel nun verallgemeinern und das **Integrationsintervall $[a;b]$ in n gleich große Teilintervalle unterteilen und**

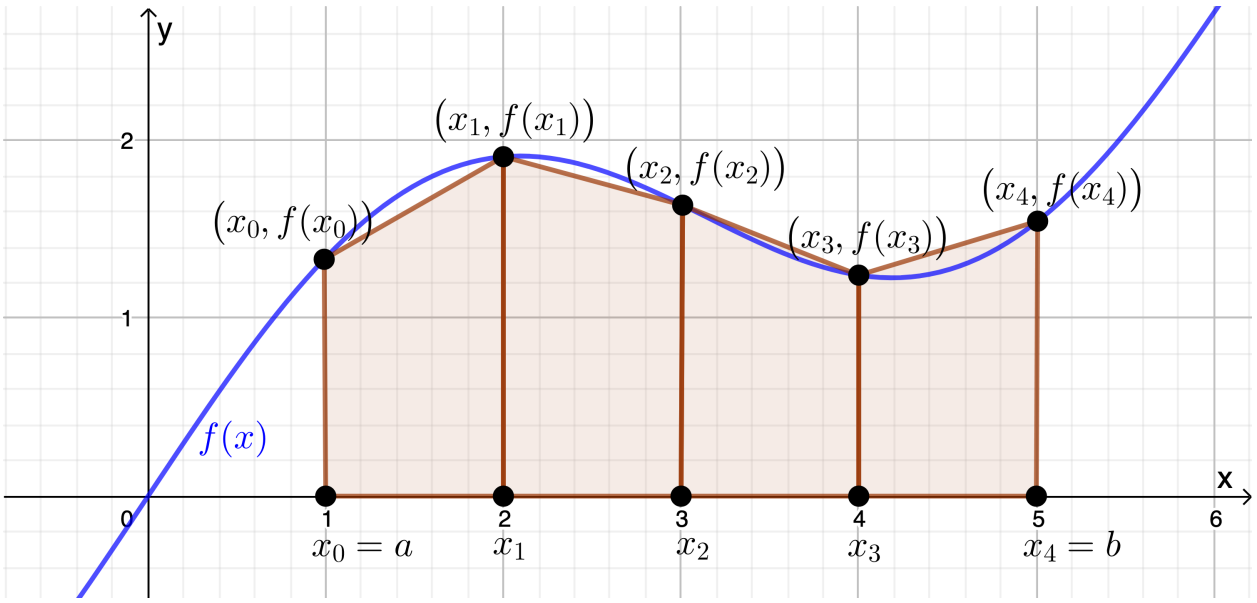


Abb. 6.9: Veranschaulichung der Trapezrege T_n für numerische Integration aus Verfahren 6.23 (hier mit einer Unterteilung in $n = 4$ gleich lange Teilintervalle).

auf jedem Teilintervall die Trapezregel anwenden: Sei also

$$h := \frac{b-a}{n}\,.$$

Dann erhält man mit

$$x_k := a + k\,h, \qquad k = 0, 1, 2, \ldots, n,$$

durch $[x_k; x_{k+1}]$, $k = 0, 1, 2, \ldots, n-1$, eine Unterteilung von $[a; b]$ in n gleich große Teilintervalle der Länge h. Das Integral (6.29) ist dann (mit $x_0 = a$ und $x_n = b$) entsprechend die Summe der Integrale über diese Teilintervalle

$$\begin{aligned} I(f) &= \int_a^b f(x)\,\mathrm{d}x = \int_{x_0}^{x_n} f(x)\,\mathrm{d}x \\ &= \int_{x_0}^{x_1} f(x)\,\mathrm{d}x + \int_{x_1}^{x_2} f(x)\,\mathrm{d}x + \ldots + \int_{x_{n-1}}^{x_n} f(x)\,\mathrm{d}x. \end{aligned}$$

Wir nutzen nun auf jedem dieser Teilintervalle der Länge h die Trapezregel (6.31) und erhalten damit

$$\begin{aligned} I(h) &\approx h\,\frac{f(x_0)+f(x_1)}{2} + h\,\frac{f(x_1)+f(x_2)}{2} + \ldots + h\,\frac{f(x_{n-1})+f(x_n)}{2} \\ &= \frac{h}{2}\Big[\big(f(x_0)+f(x_1)\big) + \big(f(x_1)+f(x_2)\big) + \ldots + \big(f(x_{n-1})+f(x_n)\big)\Big] \end{aligned}$$

$$= h \left[\frac{1}{2} f(x_0) + f(x_1) + f(x_2) + \ldots + f(x_{n-1}) + \frac{1}{2} f(x_n) \right].$$

Dieses ist die **Trapezregel für numerische Integration**, und wir halten diese als Verfahren fest. Die Trapezregel für numerische Integration ist in Abbildung 6.9 mit $n = 4$ illustriert.

Verfahren 6.23. (Trapezregel für numerische Integration)
*Sei $f : [a; b] \to \mathbb{R}$ eine stetige Funktion. Für $n \in \mathbb{N}$ sei $h := \frac{1}{n}(b - a)$ und die **Knoten(punkte)** seien $x_k := a + k\,h$, $k = 0, 1, 2, \ldots, n$. Dann liefert die **Trapezregel für numerische Integration***

$$T_n(f) := h \left[\frac{1}{2} f(x_0) + \sum_{k=1}^{n-1} f(x_k) + \frac{1}{2} f(x_n) \right] \tag{6.32}$$
$$= h \left[\frac{1}{2} f(x_0) + f(x_1) + f(x_2) + \ldots + f(x_{n-1}) + \frac{1}{2} f(x_n) \right]$$

eine Näherung für das Integral $I(f) = \int_a^b f(x)\,\mathrm{d}x$.

Der Index n von $T_n(f)$ steht für die Anzahl der gleichlangen Teilintervalle, in die $[a; b]$ unterteilt wurde. Natürlich ist die „einfache Trapezregel“ in Verfahren 6.19 der Sonderfall $n = 1$ von Verfahren 6.23.

Aus Hilfssatz 6.21 folgt direkt, dass die Trapezregel T_n für numerische Integration für alle Polynome vom Grad ≤ 1 exakt ist.

Satz 6.24. (Trapezregel für numerische Integration ist exakt auf $\mathbb{P}_1$)
*Für jedes Polynom $p_1 \in \mathbb{P}_1$, also $p_1(x) = c\,x + d$ mit Konstanten $c, d \in \mathbb{R}$, gilt für die durch (6.32) definierte Trapezregel für numerische Integration $T_n(p_1) = I(p_1)$, d.h. die Trapezregel T_n für numerische Integration **integriert Polynome in $\mathbb{P}_1$ (also von Grad ≤ 1) exakt**.*

Beweis von Satz 6.24 Sei p_1 ein beliebiges Polynom vom Grad ≤ 1. Dann gilt nach der Konstruktion der Trapezregel T_n

$$T_n(p_1) = \sum_{k=0}^{n-1} \int_{x_k}^{x_{k+1}} P_{1,k}(x)\,\mathrm{d}x,$$

n	$h = \frac{b-a}{n} = \frac{1}{n}$	$T_n(f)$	$\lvert T_n(f) - I(f)\rvert$	$\frac{\lvert T_{n/2}(f) - I(f)\rvert}{\lvert T_n(f) - I(f)\rvert}$
$2 = 2^1$	0,5	0,731370252	$1{,}55 \cdot 10^{-2}$	
$4 = 2^2$	0,25	0,742984098	$3{,}84 \cdot 10^{-3}$	4,02
$8 = 2^3$	0,125	0,745865615	$9{,}59 \cdot 10^{-4}$	4,01
$16 = 2^4$	0,0625	0,746584597	$2{,}40 \cdot 10^{-4}$	4,00
$32 = 2^5$	0,03125	0,746764255	$5{,}99 \cdot 10^{-5}$	4,00
$64 = 2^6$	0,015625	0,746809164	$1{,}50 \cdot 10^{-5}$	4,00
$128 = 2^7$	0,0078125	0,746820391	$3{,}74 \cdot 10^{-6}$	4,00

Tabelle 6.1: Trapezregel $T_n(f)$ mit $f(x) = e^{-x^2}$ und $[a; b] = [0; 1]$ zur Berechnung des Integrals $\int_0^1 e^{-x^2}\,\mathrm{d}x$.

wobei $P_{1,k}$ das lineare Interpolationspolynom von p_1 bzgl. der beiden Datenpunkte $\big(x_k; p_1(x_k)\big)$ und $\big(x_{k+1}, p_1(x_{k+1})\big)$ ist. Da die zwei Interpolationsbedingungen $P_{1,k}(x_k) = p_1(x_k)$ und $P_{1,k}(x_{k+1}) = p_1(x_{k+1})$ gelten und da p_1 ebenfalls ein Polynom vom Grad ≤ 1 ist, welches die Interpolationsbedingungen trivialerweise erfüllt, folgt aus der Eindeutigkeit des linearen Interpolationspolynoms, dass $P_{1,k} = p_1$ gelten muss. Also folgt

$$T_n(p_1) = \sum_{k=0}^{n-1} \int_{x_k}^{x_{k+1}} P_{1,k}(x)\,\mathrm{d}x, = \sum_{k=0}^{n-1} \int_{x_k}^{x_{k+1}} p_1(x)\,\mathrm{d}x = \int_a^b p_1(x)\,\mathrm{d}x = I(p_1),$$

und die Trapezregel integriert p_1 offensichtlich exakt. □

Betrachten wir ein Beispiel.

Beispiel 6.25. (Trapezregel für numerische Integration)
Wir berechnen das Integral

$$\int_0^1 e^{-x^2}\,\mathrm{d}x \doteq 0{,}746824132812427$$

für $n = 2^j$ mit $j = 1, 2, \ldots, 7$ mit der Trapezregel für numerische Integration T_n Es sind also $[a; b] = [0; 1]$, $f(x) = e^{-x^2}$ und $h = (1-0)/n = 1/n$, und die Knoten sind dann $x_k = a + k\,h = 0 + k \cdot \frac{1}{n} = \frac{k}{n}$, $k = 0, 1, 2, \ldots, n$. Wir müssen also für

$n = 2^j$ mit $j = 1, 2, \ldots, 7$ jeweils die folgende gewichtete Summe berechnen:

$$T_n(f) = \frac{1}{n} \left[\frac{e^{-0^2}}{2} + \sum_{k=1}^{n-1} e^{-(k/n)^2} + \frac{e^{-1^2}}{2} \right] = \frac{1}{n} \left[\frac{1}{2} + \sum_{k=1}^{n-1} e^{-(k/n)^2} + \frac{e^{-1}}{2} \right]$$

Die Ergebnisse sind in Tabelle 6.1 auf eine Gleitkommadarstellung mit einer 9-stelligen Mantisse gerundet angegeben. Weiter haben wir den absoluten Fehler $|T_n(f) - I(f)|$ auf eine Gleitkommadarstellung mit einer 3-stelligen Mantisse gerundet aufgelistet, sowie den Quotienten $|T_{n/2}(f) - I(f)|/|T_n(f) - I(f)|$.

An den Quotienten $|T_{n/2}(f) - I(f)|/|T_n(f) - I(f)|$ sehen wir, dass sich bei einer Verdoppelung von n und damit einer Halbierung des Abstands h der Knotenpunkte x_k, $k = 0, 1, \ldots, n$, (und ungefähr einer Verdoppelung der Anzahl der Knotenpunkte) der numerische Integrationsfehler $|T_n(f) - I(f)|$ ungefähr durch den Faktor 4 geteilt wird. Daher vermuten wir, dass der numerische Integrationsfehler möglicherweise die Ordnung $\mathcal{O}(h^2)$ hat. ♠

Wie der nachfolgende Satz zeigt, ist das im vorigen Beispiel vermutete Konvergenzverhalten von der Ordnung $\mathcal{O}(h^2)$ richtig.

Satz 6.26. (Konvergenz der Trapezregel für numerische Integration)
*Sei $n \in \mathbb{N}$, und sei $f : [a; b] \to \mathbb{R}$ eine zweimal stetig differenzierbare Funktion. Dann gilt für den **absoluten Fehler** $E_n^T(f) := |T_n(f) - I(f)|$ der Approximation des Integrals*

$$I(f) = \int_a^b f(x)\,\mathrm{d}x$$

*durch die in Verfahren 6.23 definierte **Trapezregel für numerische Integration** $T_n(f)$ die Fehlerabschätzung*

$$E_n^T(f) = |T_n(f) - I(f)| \leq \frac{b-a}{12}\, h^2 \max_{x \in [a;b]} |f''(x)|. \tag{6.33}$$

Für mathematisch Interessierte zeigen wir den Beweis von Satz 6.26.

Beweis von Satz 6.26: Wir erinnern uns, dass wir die Trapezregel $T_1(f)$ zu Beginn des Kapitels als das Integral über das lineare Interpolationspolynom P_1 bzgl. der Datenpunkte $\big(a; f(a)\big)$, $\big(b; f(b)\big)$ eingeführt haben. Entsprechend gilt für das Teilintervall $[x_k; x_{k+1}]$, $k \in \{0, 1, 2, \ldots, n-1\}$, bei der Trapezregel für

numerische Integration $T_n(f)$ also

$$\int_{x_k}^{x_{k+1}} f(x)\,\mathrm{d}x \approx \int_{x_k}^{x_{k+1}} P_{1,k}(x)\,\mathrm{d}x \qquad \text{mit dem linearen Interpolationspolynom}$$

$$P_{1,k}(x) = f(x_k)\,\frac{x-x_{k+1}}{x_k-x_{k+1}} + f(x_{x+1})\,\frac{x-x_k}{x_{k+1}-x_k} \quad \text{bzgl.} \quad \begin{cases} \big(x_k; f(x_k)\big), \\ \big(x_{k+1}; f(x_{k+1})\big), \end{cases}$$

wobei das Integral auf der rechten Seite gerade der Beitrag zu $T_n(f)$ durch die Trapezregel auf dem Teilintervall $[x_k; x_{k+1}]$ ist. Es gilt mit der Dreiecksungleichung

$$\begin{aligned} E_n^T(f) = |T_n(f) - I(f)| &= \left| \sum_{k=0}^{n-1} \int_{x_k}^{x_{k+1}} P_{1,k}(x)\,\mathrm{d}x - \sum_{k=0}^{n-1} \int_{x_k}^{x_{k+1}} f(x)\,\mathrm{d}x \right| \\ &= \left| \sum_{k=0}^{n-1} \int_{x_k}^{x_{k+1}} \big[P_{1,k}(x) - f(x)\big]\,\mathrm{d}x \right| \le \sum_{k=0}^{n-1} \left| \int_{x_k}^{x_{k+1}} \big[P_{1,k}(x) - f(x)\big]\,\mathrm{d}x \right| \\ &\le \sum_{k=0}^{n-1} \int_{x_k}^{x_{k+1}} \big|P_{1,k}(x) - f(x)\big|\,\mathrm{d}x. \end{aligned} \tag{6.34}$$

Die Integrale in der dritten Zeile von (6.34) sind alle von der Form

$$\int_{\alpha}^{\alpha+h} \big|P_1(x) - f(x)\big|\,\mathrm{d}x \quad \text{mit} \quad P_1(x) = f(\alpha)\,\frac{(\alpha+h)-x}{h} + f(\alpha+h)\,\frac{x-\alpha}{h}.$$

Nach Satz 6.15 gilt (da f zweimal stetig differenzierbar ist) mit einem (von $x \in [\alpha; \alpha+h]$ abhängigen) Punkt c_x in $[\alpha; \alpha+h]$

$$\begin{aligned} |P_1(x) - f(x)| &= \left| -\frac{(x-\alpha)\big(x-(\alpha+h)\big)}{2!}\,f''(c_x) \right| \\ &= \frac{|f''(c_x)|}{2}\,(x-\alpha)\big((\alpha+h)-x\big) \\ &\le \frac{1}{2}\left(\max_{t\in[\alpha;\alpha+h]} |f''(t)|\right)(x-\alpha)\big((\alpha+h)-x\big). \end{aligned} \tag{6.35}$$

Wir formen nun $(x-\alpha)\big((\alpha+h)-x\big)$ geeignet um:

$$\begin{aligned} \big(x-\alpha\big)\big((\alpha+h)-x\big) &= \left(x - \left(\alpha+\frac{h}{2}\right) + \frac{h}{2}\right)\left(\left(\alpha+\frac{h}{2}\right) - x + \frac{h}{2}\right) \\ &= \left(\frac{h}{2} + \left[x - \left(\alpha+\frac{h}{2}\right)\right]\right)\left(\frac{h}{2} - \left[x - \left(\alpha+\frac{h}{2}\right)\right]\right) \end{aligned}$$

$$= \left(\frac{h}{2}\right)^2 - \left[x - \left(\alpha + \frac{h}{2}\right)\right]^2 = \frac{h^2}{4} - \left[x - \left(\alpha + \frac{h}{2}\right)\right]^2, \tag{6.36}$$

Einsetzen von (6.36) in (6.35) und Integrieren über $[\alpha; \alpha + h]$ ergibt:

$$\begin{aligned}
&\int_\alpha^{\alpha+h} \left|P_1(x) - f(x)\right| \mathrm{d}x \\
&\leq \frac{1}{2} \left(\max_{t\in[\alpha;\alpha+h]} |f''(t)|\right) \int_\alpha^{\alpha+h} \left(\frac{h^2}{4} - \left[x - \left(\alpha + \frac{h}{2}\right)\right]^2\right) \mathrm{d}x \\
&= \frac{1}{2} \left(\max_{t\in[\alpha;\alpha+h]} |f''(t)|\right) \left(\left[\frac{h^2}{4}\, x\right]_{x=\alpha}^{x=\alpha+h} - \left[\frac{1}{3}\left(x - \left(\alpha + \frac{h}{2}\right)\right)^3\right]_{x=\alpha}^{x=\alpha+h}\right) \\
&= \frac{1}{2} \left(\max_{t\in[\alpha;\alpha+h]} |f''(t)|\right) \left(\left[\frac{h^2}{4}\left((\alpha+h) - \alpha\right)\right] - \left[\frac{1}{3}\left(\frac{h}{2}\right)^3 - \frac{1}{3}\left(-\frac{h}{2}\right)^3\right]\right) \\
&= \frac{1}{2} \left(\max_{t\in[\alpha;\alpha+h]} |f''(t)|\right) \left(\frac{1}{4}\, h^3 - \frac{1}{12}\, h^3\right) = \frac{1}{12}\, h^3 \left(\max_{t\in[\alpha;\alpha+h]} |f''(t)|\right)
\end{aligned} \tag{6.37}$$

Anwenden der Abschätzung (6.37) auf jedes der Integrale in (6.34) liefert

$$\begin{aligned}
E_n^T(f) &\leq \sum_{k=0}^{n-1} \int_{x_k}^{x_{k+1}} \left|P_{1,k}(x) - f(x)\right| \mathrm{d}x \leq \sum_{k=0}^{n-1} \frac{1}{12}\, h^3 \left(\max_{t\in[x_k;x_{k+1}]} |f''(t)|\right) \\
&\leq \frac{1}{12}\, h^2 \left(\max_{t\in[a;b]} |f''(t)|\right) \underbrace{\sum_{k=0}^{n-1} h}_{\substack{=n\,h \\ =b-a}} = \frac{b-a}{12}\, h^2 \left(\max_{t\in[a;b]} |f''(t)|\right),
\end{aligned}$$

womit (6.33) bewiesen ist. □

6.5 Elementare Quadraturformeln: Simpson-Regel

Um eine elementare, aber bessere, Quadraturformel als die Trapezregel für numerische Integration zu bekommen, gehen wir analog zur Herleitung der Trapezregel vor, aber ersetzen den Integranden durch ein (höchstens) quadratisches Interpolationspolynom aus $\mathbb{P}_2$. Genauer ersetzen wir in

$$I(f) = \int_a^b f(x)\, \mathrm{d}x \qquad \text{mit } f \in \mathcal{C}\big([a;b]\big) \tag{6.38}$$

den Integranden f durch das (höchstens) quadratische Interpolationspolynom $P_2 \in \mathbb{P}_2$ bzgl. der Datenpunkte $\big(a; f(a)\big)$, $\big(c; f(c)\big)$, $\big(b; f(b)\big)$, wobei $c := (a+b)/2$ der Punkt genau in der Mitte zwischen a und b ist. Das quadratische Interpolationspolynom $P_2 \in \mathbb{P}_2$ hat dann die folgende Form:

$$P_2(x) = f(a)\,\frac{(x-b)\,(x-c)}{(a-b)\,(a-c)} + f(c)\,\frac{(x-a)\,(x-b)}{(c-a)\,(c-b)} + f(b)\,\frac{(x-a)\,(x-c)}{(b-a)\,(b-c)}$$

Also erhalten wir die folgende Näherungsformel für das Integral

$$\begin{aligned} I(f) \approx \int_a^b P_2(x)\,\mathrm{d}x &= f(a)\,\int_a^b \frac{(x-b)\,(x-c)}{(a-b)\,(a-c)}\,\mathrm{d}x \\ &+ f(c)\,\int_a^b \frac{(x-a)\,(x-b)}{(c-a)\,(c-b)}\,\mathrm{d}x + f(b)\,\int_a^b \frac{(x-a)\,(x-c)}{(b-a)\,(b-c)}\,\mathrm{d}x. \end{aligned} \tag{6.39}$$

Die drei Integrale lassen sich exakt berechnen. Zuvor ist es aber zweckmäßig $h := (b-a)/2$ einzuführen. Dann gelten $h = c-a = b-c$ und $b-a = 2\,h$, und somit folgt auch $c = a+h$, $b = c+h = a+2\,h$. Dann führen wir in jedem der drei Integrale die Substitution $t = x-a \iff x = t+a$, $\mathrm{d}t = \mathrm{d}x$, durch, damit in dem Ergebnis nicht mehr a, b, c, sondern nur h auftaucht. Wir erhalten mit dieser Vorgehensweise mit $x-b = t+a-b = t-2\,h$, $x-c = t+a-c = t-h$ und $x-a = t$ jeweils

$$\begin{aligned} \int_a^b \frac{(x-b)\,(x-c)}{(a-b)\,(a-c)}\,\mathrm{d}x &= \frac{1}{2\,h^2}\int_0^{2h} (t-2h)\,(t-h)\,\mathrm{d}t \\ &= \frac{1}{2\,h^2}\int_0^{2h} \big(t^2 - 3\,h\,t + 2\,h^2\big)\,\mathrm{d}t = \frac{1}{2\,h^2}\left[\frac{1}{3}\,t^3 - \frac{3}{2}\,h\,t^2 + 2\,h^2\,t\right]_{t=0}^{t=2h} \\ &= \frac{1}{2\,h^2}\left[\frac{8}{3}\,h^3 - 6\,h^3 + 4\,h^3\right] = \frac{1}{2\,h^2}\cdot\frac{2}{3}\,h^3 = \frac{h}{3}\,, \end{aligned}$$

$$\begin{aligned} \int_a^b \frac{(x-a)\,(x-b)}{(c-a)\,(c-b)}\,\mathrm{d}x &= \frac{-1}{h^2}\int_0^{2h} t\,(t-2h)\,\mathrm{d}t = \frac{-1}{h^2}\int_0^{2h} \big(t^2 - 2\,h\,t\big)\,\mathrm{d}t \\ &= \frac{-1}{h^2}\left[\frac{1}{3}\,t^3 - h\,t^2\right]_{t=0}^{t=2h} = \frac{-1}{h^2}\left[\frac{8}{3}\,h^3 - 4\,h^3\right] = \frac{-1}{h^2}\cdot\frac{-4}{3}\,h^3 = \frac{4\,h}{3}\,, \end{aligned}$$

$$\begin{aligned} \int_a^b \frac{(x-a)\,(x-c)}{(b-a)\,(b-c)}\,\mathrm{d}x &= \frac{1}{2\,h^2}\int_0^{2h} t\,(t-h)\,\mathrm{d}t = \frac{1}{2\,h^2}\int_0^{2h} \big(t^2 - h\,t\big)\,\mathrm{d}t \\ &= \frac{1}{2\,h^2}\left[\frac{1}{3}\,t^3 - \frac{1}{2}\,h\,t^2\right]_{t=0}^{t=2h} = \frac{1}{2\,h^2}\left[\frac{8}{3}\,h^3 - 2\,h^3\right] = \frac{1}{2\,h^2}\cdot\frac{2}{3}\,h^3 = \frac{h}{3}\,. \end{aligned}$$

Einsetzen der Werte für die Integrale in (6.39) liefert die **Simpson-Regel**

$$I(f) \approx f(a) \cdot \frac{h}{3} + f(c) \cdot \frac{4\,h}{3} + f(b) \cdot \frac{h}{3} = \frac{h}{3}\left[(a) + 4\,f(c) + f(b)\right]$$

$$= \boxed{\frac{h}{3}\left[f(a) + 4\,f\left(\frac{a+b}{2}\right) + f(b)\right] =: S_2(f),} \tag{6.40}$$

wobei wir in der zweiten Zeile von (6.40) noch $c = (a+b)/2$ eingesetzt haben.

Wir halten die Simpson-Regel als Verfahren fest:

Verfahren 6.27. (Simpson-Regel)
Seien $f : [a;b] \to \mathbb{R}$ eine stetige Funktion und $h := \frac{b-a}{2}$. Die ***Simpson-Regel***

$$S_2(f) := \frac{h}{3}\left[f(a) + 4\,f\left(\frac{a+b}{2}\right) + f(b)\right] \tag{6.41}$$

liefert eine Nährung für das Integral $I(f) = \displaystyle\int_a^b f(x)\,\mathrm{d}x$.

Betrachten wir zunächst ein Beispiel.

Beispiel 6.28. (Simpson-Regel)
Wir wollen das Integral

$$I(f) = \int_0^1 \frac{1}{1+x}\,\mathrm{d}x = \ln(2) \doteq 0{,}6931471806$$

(welches bereits in Beispiel 6.20 mit der Trapezregel und in Beispiel 6.22 mit der „zusammengesetzten“ Trapezregel berechnet wurde) mit der Simpson-Regel angenähert berechnen: Es gilt $h = (1-0)/2 = \frac{1}{2}$ und damit $h/3 = 1/6$ und

$$S_2(f) = \frac{1}{6}\left[f(0) + 4\,f\left(\frac{1}{2}\right) + f(1)\right] = \frac{1}{6}\left[\frac{1}{1+0} + 4 \cdot \frac{1}{1+\frac{1}{2}} + \frac{1}{1+1}\right]$$

$$= \frac{1}{6}\left[1 + \frac{8}{3} + \frac{1}{2}\right] = \frac{25}{36} \doteq 0{,}69444.$$

Der exakte Wert des Integrals ist $I(f) = \ln(2) \doteq 0{,}6931471806$, und der absolute Fehler der Simpson-Regel ist somit $|S_2(f) - I(f)| = |\frac{25}{36} - \ln(2)| \doteq 0{,}00130$. Wir sehen, dass der absolute Fehler deutlich kleiner ist als in Beispielen 6.20 und

6.22. Dabei ist der Vergleich mit Beispiel 6.22 eher angebracht, weil dort auch drei Knoten verwendet werden wie in der Simpson-Regel. Aber selbst verglichen mit Beispiel 6.22 ist der Fehler ungefähr um einen Faktor 10 kleiner. ♠

Analog zu Hilfssatz 6.21 gilt der folgende Hilfssatz für die Simpson-Regel. Dabei erscheint es zunächst verblüffend, dass die Simpson-Regel sogar alle Polynome vom Grad ≤ 3 und nicht nur alle Polynome vom Grad ≤ 2 exakt integriert.

Hilfssatz 6.29. (Simpson-Regel ist exakt für Polynome in $\mathbb{P}_3$)
Für jedes Polynom $p_3 \in \mathbb{P}_3$, also $p_3(x) = c_3\,x^3 + c_2\,x^2 + c_1\,x + c_0$ mit den Konstanten $c_0, c_1, c_2, c_3 \in \mathbb{R}$, gilt für die durch (6.41) definierte Simpson-Regel $S_2(p_3) = I(p_3)$, d.h. die Simpson-Regel ***integriert Polynome in $\mathbb{P}_3$ (also von Grad ≤ 3) exakt****.*

Beweis von Hilfssatz 6.21: Da die Polynome $q_0(x) = 1$, $q_1(x) = x - a$, $q_2(x) = (x-a)^2$ und $q_3(x) = (x-a)^3$ eine Basis des Polynomraums $\mathbb{P}_3$ der Polynome vom Grad ≤ 3 bilden, kann man jedes Polynom p_3 im $\mathbb{P}_3$ in der Form

$$\begin{aligned} p_3(x) &= c_0\,q_0(x) + c_1\,q_1(x) + c_2\,q_2(x) + c_3\,q_3(x) \\ &= c_0 + c_1\,(x-a) + c_2\,(x-a)^2 + c_3\,(x-a)^3 \end{aligned}$$

schreiben. Wegen der Linearität des Integrals und der Simpson-Regel, gelten

$$\begin{aligned} I(p_3) &= c_0\,I(q_0) + c_1\,I(q_1) + c_2\,I(q_2) + c_3\,I(p_3), \\ S_2(p_3) &= c_0\,S_2(q_0) + c_1\,S_2(q_1) + c_2\,S_2(q_2) + c_3\,S_2(p_3), \end{aligned}$$

so dass es reicht, zu zeigen, dass $I(q_\ell) = S_2(q_\ell)$ für $\ell = 0, 1, 2, 3$. Wir erhalten

$$\begin{aligned} I(q_\ell) &= \int_a^b (x-a)^\ell \,\mathrm{d}x = \left[\frac{1}{\ell+1}\,(x-a)^{\ell+1}\right]_{x=a}^{x=b} \\ &= \frac{1}{\ell+1}\,(b-a)^{\ell+1} = \frac{1}{\ell+1}\,(2h)^{\ell+1}, \qquad \ell \in \mathbb{N}_0, \end{aligned} \tag{6.42}$$

wobei wir im letzten Schritt $b - a = 2h$ genutzt haben. Anderseits liefert die Simpson-Regel mit $b - a = 2\,h$ und $\frac{a+b}{2} - a = \frac{b-a}{2} = h$

$$S_2(q_\ell) = \frac{h}{3}\left[(a-a)^\ell + 4\left(\frac{a+b}{2} - a\right)^\ell + (b-a)^\ell\right]$$

$$= \frac{h}{3}\left[0^\ell + 4\,h^\ell + (2\,h)^\ell\right] = \begin{cases} \dfrac{4+2^\ell}{3}\,h^{\ell+1} & \text{für } \ell \in \mathbb{N}, \\ \dfrac{h\,(1+4+1)}{3} = 2\,h & \text{für } \ell = 0. \end{cases} \tag{6.43}$$

Also finden wir durch den Vergleich von (6.42) und (6.43)

$$\begin{aligned} I(q_0) &= 2\,h = S_2(q_0), \\ I(q_1) &= \frac{1}{2}\,(2h)^2 = 2\,h^2 = \frac{4+2^1}{3}\,h^2 = S_2(q_1), \\ I(q_2) &= \frac{1}{3}\,(2h)^3 = \frac{8}{3}\,h^3 = \frac{4+2^2}{3}\,h^3 = S_2(q_2), \\ I(q_3) &= \frac{1}{4}\,(2h)^4 = 4\,h^4 = \frac{4+2^3}{3}\,h^4 = S_2(q_3), \end{aligned}$$

d.h. die Simpson-Regel ist in der Tat für Polynome in $\mathbb{P}_3$ exakt. □

Analog zur Vorgehensweise bei der Trapezregel wollen wir nun mit der Simpson-Regel eine „zusammengesetzte“ Simpson-Regel bauen, indem wir das **Intervall** $[a;b]$ **in** n **gleich große Teilintervalle zerlegen und auf jedem Teilintervall die Simpson-Regel nutzen**: Sei also

$$h := \frac{b-a}{2n}\,.$$

Dann erhält man mit

$$x_k := a + k\,h, \qquad k = 0, 1, 2, \ldots, 2n,$$

durch $[x_k; x_{k+2}]$, $k = 0, 2, 4, \ldots, 2n-2$, eine Unterteilung von $[a;b]$ in n gleich große Teilintervalle der Länge $2h$. Das Integral (6.38) ist dann entsprechend die Summe der Integrale über diese Teilintervalle, also

$$\begin{aligned} I(f) &= \int_a^b f(x)\,\mathrm{d}x = \int_{x_0}^{x_{2n}} f(x)\,\mathrm{d}x \\ &= \int_{x_0}^{x_2} f(x)\,\mathrm{d}x + \int_{x_2}^{x_4} f(x)\,\mathrm{d}x + \ldots + \int_{x_{2n-2}}^{x_{2n}} f(x)\,\mathrm{d}x. \end{aligned}$$

Wir nutzen nun auf jedem dieser Teilintervalle der Länge $2h$ die Simpson-Regel (6.41) und erhalten damit

$$I(f) \approx \frac{h}{3}\left[f(x_0) + 4\,f(x_1) + f(x_2)\right] + \frac{h}{3}\left[f(x_2) + 4\,f(x_3) + f(x_4)\right]$$

$$
\begin{aligned}
&\quad + \ldots + \frac{h}{3}\left[f(x_{2n-2}) + 4\,f(x_{2n-1}) + f(x_{2n})\right] \\
&= \frac{h}{3}\left[f(x_0) + 4\,f(x_1) + 2\,f(x_2) + 4\,f(x_3) + 2\,f(x_4)\right. \\
&\quad \left. + \ldots + 2\,f(x_{2n-2}) + 4\,f(x_{2n-1}) + f(x_{2n})\right] \\
&= \boxed{\frac{h}{3}\left[f(x_0) + 2\sum_{k=1}^{n-1} f(x_{2k}) + 4\sum_{k=1}^{n} f(x_{2k-1}) + f(x_{2n})\right] =: S_{2n}(f)}
\end{aligned}
$$

Dieses ist die **Simpson-Regel für numerische Integration**, und wir halten diese als Verfahren fest. Die Simpson-Regel für numerische Integration ist seit zwei Jahrhunderten eine der beliebtesten elementaren Integrationsregeln.

Verfahren 6.30. (Simpson-Regel für numerische Integration)
*Sei $f : [a;b] \to \mathbb{R}$ eine stetige Funktion. Für $n \in \mathbb{N}$ sei $h := \frac{1}{2n}(b-a)$, und die **Knoten(punkte)** seien $x_k := a + k\,h$, $k = 0, 1, 2, \ldots, 2n$. Dann liefert die **Simpson-Regel für numerische Integration***

$$
\begin{aligned}
S_{2n}(f) &:= \frac{h}{3}\left[f(x_0) + 2\sum_{k=1}^{n-1} f(x_{2k}) + 4\sum_{k=1}^{n} f(x_{2k-1}) + f(x_{2n})\right] \qquad (6.44) \\
&= \frac{h}{3}\left[f(x_0) + 4\,f(x_1) + 2\,f(x_2) + 4\,f(x_3) + 2\,f(x_4)\right. \\
&\quad \left. + \ldots + 2\,f(x_{2n-2}) + 4\,f(x_{2n-1}) + f(x_{2n})\right]
\end{aligned}
$$

eine Näherung für das Integral $I(f) = \int_a^b f(x)\,\mathrm{d}x$.

Betrachten wir zunächst ein Beispiel.

Beispiel 6.31. (Simpson-Regel für numerische Integration)
Wir berechnen das Integral

$$\int_0^1 e^{-x^2}\,\mathrm{d}x \doteq 0{,}746824132812427$$

(welches bereits in Beispiel 6.25 mit der Trapezregel für numerische Integration T_n berechnet wurde) für $2n = 2^j$ mit $j = 1, 2, \ldots, 7$ mit der Simpson-Regel für numerische Integration S_{2n}: Es sind also $[a;b] = [0;1]$, $f(x) = e^{-x^2}$, $h = \frac{1-0}{2n} = \frac{1}{2n}$, und die Knoten sind dann $x_k = a + k\,h = 0 + k \cdot \frac{1}{2n} = \frac{k}{2n}$, $k = 0, 1, 2, \ldots, 2n$.

$2n$	$h = \frac{b-a}{2n} = \frac{1}{2n}$	$S_{2n}(f)$	$\lvert S_{2n}(f) - I(f)\rvert$	$\frac{\lvert S_n(f) - I(f)\rvert}{\lvert S_{2n}(f) - I(f)\rvert}$
$2 = 2^1$	0,5	0,74718042891	$3{,}56 \cdot 10^{-4}$	
$4 = 2^2$	0,25	0,74685537979	$3{,}12 \cdot 10^{-5}$	11,4
$8 = 2^3$	0,125	0,74682612053	$1{,}99 \cdot 10^{-6}$	15,7
$16 = 2^4$	0,0625	0,74682425744	$1{,}25 \cdot 10^{-7}$	15,9
$32 = 2^5$	0,03125	0,74682414061	$7{,}79 \cdot 10^{-9}$	16,0
$64 = 2^6$	0,015625	0,74682413330	$4{,}87 \cdot 10^{-10}$	16,0
$128 = 2^7$	0,0078125	0,74682413284	$3{,}04 \cdot 10^{-11}$	16,0

Tabelle 6.2: Simpson-Regel $S_{2n}(f)$ mit $f(x) = e^{-x^2}$ und $[a;b] = [0;1]$ zur Berechnung des Integrals $\int_0^1 e^{-x^2}\,\mathrm{d}x$.

Wir müssen also für $2n = 2^j$ mit $j = 1, 2, \ldots, 7$ jeweils die folgende gewichtete Summe berechnen:

$$\begin{aligned} S_{2n}(f) &= \frac{1}{6n}\left[e^{-0^2} + 2\sum_{k=1}^{n-1} e^{-((2k)/(2n))^2} + 4\sum_{k=1}^{n} e^{-((2k-1)/(2n))^2} + e^{-1^2}\right] \\ &= \frac{1}{6n}\left[1 + 2\sum_{k=1}^{n-1} e^{-(k/n)^2} + 4\sum_{k=1}^{n} e^{-((k-\frac{1}{2})/n)^2} + e^{-1}\right] \end{aligned}$$

Die Ergebnisse sind in Tabelle 6.2 auf eine Gleitkommadarstellung mit einer 11-stelligen Mantisse gerundet angegeben. Weiter haben wir den absoluten Fehler $|S_{2n}(f) - I(f)|$ und den Quotienten $|S_n(f) - I(f)|/|S_{2n}(f) - I(f)|$ auf eine Gleitkommadarstellung mit einer 3-stelligen Mantisse gerundet aufgelistet. – Ein Vergleich der Tabelle 6.2 mit Tabelle 6.1 zeigt, dass die Simpson-Regel S_{2n} eine deutlich bessere Näherung liefert als die Trapezregel T_{2n} mit der gleichen Knotenzahl.

An den Quotienten $|S_n(f) - I(f)|/|S_{2n}(f) - I(f)|$ sehen wir, dass sich bei einer Verdoppelung von n und damit einer Halbierung des Abstands h der Knotenpunkte x_k, $k = 0, 1, \ldots, 2n$, (und ungefähr einer Verdoppelung der Anzahl der Knotenpunkte) der Integrationsfehler ungefähr durch den Faktor 16 geteilt wird. Daher vermuten wir, dass der numerische Integrationsfehler $|S_{2n}(f) - I(f)|$ möglicherweise die Ordnung $\mathcal{O}(h^4)$ hat. ♠

Der nachfolgende Satz zeigt, dass das im vorigen Beispiel vermutete Konvergenzverhalten von der Ordnung $\mathcal{O}(h^4)$ richtig ist.

Satz 6.32. (Konvergenz der Simpson-Regel für num. Integration)
Sei $n \in \mathbb{N}$, und sei $f : [a;b] \to \mathbb{R}$ eine viermal stetig differenzierbare Funktion. Dann gilt für den ***absoluten Fehler*** $E_{2n}^S(f) := |S_{2n}(f) - I(f)|$ *der Approximation des Integrals*

$$I(f) = \int_a^b f(x)\,\mathrm{d}x$$

durch die in Verfahren 6.30 definierte ***Simpson-Regel für numerische Integration*** $S_{2n}(f)$ *die Fehlerabschätzung*

$$E_{2n}^S(f) = |S_{2n}(f) - I(f)| \leq \frac{b-a}{180}\, h^4 \max_{x\in[a;b]} |f^{(4)}(x)|.$$

Aus Hilfssatz 6.29 folgt schließlich noch, dass die Simpson-Regel für numerische Integration auf $\mathbb{P}_3$ exakt ist.

Satz 6.33. (Simpson-Regel für num. Integration ist exakt auf $\mathbb{P}_3$)
Für jedes Polynom $p_3(x) = c_3\,x^3 + c_2\,x^2 + c_1\,x + c_0$ in $\mathbb{P}_3$ mit Konstanten $c_0, c_1, c_2, c_3 \in \mathbb{R}$ gilt für die durch (6.44) definierte Simpson-Regel für numerische Integration $S_{2n}(p_3) = I(p_3)$, d.h. die Simpson-Regel für numerische Integration ***integriert Polynome in $\mathbb{P}_3$ (also von Grad ≤ 3) exakt****.*

6.6 Gauß Quadratur

Bei der Konstruktion der Trapezregel und der Simpson-Regel für die numerische Integration des Integrals

$$I(f) = \int_a^b f(x)\,\mathrm{d}x \qquad \text{mit} \qquad f \in \mathcal{C}\big([a;b]\big) \tag{6.45}$$

wurde der Integrand in (6.45) auf jedem Teilintervall jeweils durch das interpolierende Polynom (bzgl. äquidistanter Knotenpunkte) vom Grad 1 bzw. 2 ersetzt. Die so erhaltenen Integrationsformeln hatten die Eigenschaft, dass sie alle Polynome in $\mathbb{P}_1$ (bei der Trapezregel) bzw. alle Polynome in $\mathbb{P}_3$ (bei der Simpson-Regel) exakt integrierten. Nun wollen wir numerische Integrationsformeln finden, die **alle Polynome bis zum einem möglichst hohen Grad exakt integrieren**.

Warum ist eine solche Vorgehensweise sinnvoll? Stetige Funktion lassen

n	$\varrho_n(f)$	n	$\varrho_n(f)$
1	$5{,}30 \cdot 10^{-2}$	6	$7{,}82 \cdot 10^{-6}$
2	$1{,}79 \cdot 10^{-2}$	7	$4{,}62 \cdot 10^{-7}$
3	$6{,}63 \cdot 10^{-4}$	8	$9{,}64 \cdot 10^{-8}$
4	$4{,}63 \cdot 10^{-4}$	9	$8{,}05 \cdot 10^{-9}$
5	$1{,}62 \cdot 10^{-5}$	10	$9{,}16 \cdot 10^{-10}$

Tabelle 6.3: Fehler der Minimax-Approximation von $f : [0;1] \to \mathbb{R}$, $f(x) = e^{-x^2}$.

sich sehr gut durch Polynome approximieren, wie das nachfolgende Beispiel zeigt. Daher können wir hoffen, dass eine Integrationsformel, die Polynome bis zu einem hinreichend hohen Grad exakt integriert, auch stetige Funktionen gut integriert.

Genauer gilt für eine stetige Funktion $f : [a;b] \to \mathbb{R}$, dass es ein **eindeutig bestimmtes Polynom** $q_n \in \mathbb{P}_n$ (also vom Grad $\leq n$) gibt, so dass gilt

$$\max_{x\in[a;b]} |q_n(x) - f(x)| = \min_{p\in\mathbb{P}_n} \left(\max_{x\in[a;b]} |p(x) - f(x)| \right) =: \varrho_n(f). \tag{6.46}$$

Bei $\varrho_n(f)$ in (6.46) handelt es sich um die maximale betragliche Abweichung von q_n von f, wobei q_n das Polynom in $\mathbb{P}_n$ ist, für welches diese betragliche Abweichung am kleinsten ist. Man nennt q_n in (6.46) auch die **Minimax-Approximation von f in $\mathbb{P}_n$** und $\varrho_n(f)$ in (6.46) den **Minimax-Fehler** (der Minimax-Approximation von f in $\mathbb{P}_n$). Im nachfolgenden Beispiel wurde $\varrho_n(f)$ für eine konkrete Funktion berechnet, und wir sehen, dass $\varrho_n(f)$ rapide gegen null strebt.

Beispiel 6.34. (Approximation einer stetigen Funktion in $\mathbb{P}_n$)
Sei $f : [0;1] \to \mathbb{R}$, $f(x) = e^{-x^2}$. In Tabelle 6.3 wurde der Minimax-Fehler (6.46) angegeben. Wir sehen, dass dieser rapide gegen null strebt. Allerdings nimmt der Minimax-Fehler nicht mit einer gleichmäßigen Rate ab. ♠

Kommen wir nach dieser Motivation zu unserer Ausgangsfragestellung zurück: Wir möchten also numerische Integrationsformeln konstruieren, die alle Polynome in $\mathbb{P}_m$ mit einen niedrigen bis moderaten Grad m exakt integrieren. Wir beschränken uns bei der nachfolgenden Diskussion auf das Intervall $[a;b] = [-1;1]$. Der Transfer auf andere Intervalle ist unkompliziert und wird am Ende besprochen.

Wir wollen also für stetige Funktionen $f : [-1;1] \to \mathbb{R}$ das Integral

$$I(f) = \int_{-1}^{1} f(x)\,\mathrm{d}x \tag{6.47}$$

mit einer **numerischen Integrationsformel** oder **Quadraturformel**

$$Q_n(f) := \sum_{j=1}^{n} w_j\, f(x_j) \tag{6.48}$$

angenähert berechnen, welche alle Polynome bis zu einem möglichst hohen Grad m exakt integriert. In (6.48) nennen wir die n paarweise verschiedenen Punkte $x_1, x_2, \ldots, x_n$ die **Knoten(punkte)** der Integrationsformel Q_n, und die Zahlen $w_1, w_2, \ldots, w_n$ nennen wir die zugehörigen **Gewichte**. Üblicherweise verlangt man auch noch, dass die **Gewichte** $w_1, w_2, \ldots, w_n$ **positiv** sind. Es gibt allerdings auch numerische Integrationsformeln mit positiven und negativen Gewichten.

Die Quadraturformel (6.48) hat also $2n$ Parameter, nämlich $x_1, x_2, \ldots, x_n$ und $w_1, w_2, \ldots, w_n$, die wir passend wählen können, um zum erreichen, dass (6.48) Polynome bis zu einem möglichst hohen Grad m exakt integriert. Der Raum $\mathbb{P}_m$ der Polynome vom Grad $\leq m$ hat die Dimension $\dim(\mathbb{P}_m) = m+1$, denn die $m+1$ Monome $p_\ell(x) = x^\ell$, $\ell = 0, 1, \ldots, m$, bilden eine Basis für $\mathbb{P}_m$. Daher vermuten wir, dass es mit einer Quadraturformel (6.48) bei passender Wahl der insgesamt $2n = \dim(\mathbb{P}_{2n-1})$ Parameter (n Knoten(punkte) und n Gewichte) möglich sein sollte, alle Polynome in $\mathbb{P}_{2n-1}$ also vom Grad $\leq 2n-1$ exakt zu integrieren. Formeln mit dieser Eigenschaft nennt man **Gauß Quadraturformeln**. Wir halten dieses direkt für den Fall eines beliebigen Intervalls $[a;b]$ als Definition fest.

Definition 6.35. (Gauß Quadraturformel)
Sei Q_n eine Quadraturformel

$$Q_n(f) = \sum_{j=1}^{n} w_j\, f(x_j), \qquad f \in \mathcal{C}\big([a;b]\big),$$

*mit den **n (verschiedenen) Knoten(punkten)** $x_1, x_2, \ldots, x_n$ und den **n zugehörigen Gewichten** $w_1, w_2, \ldots, w_n$, als Näherung des Integrals*

$$I(f) = \int_a^b f(x)\,\mathrm{d}x, \qquad f \in \mathcal{C}\big([a;b]\big).$$

Wenn gilt $Q_n(p) = I(p)$ für alle $p \in \mathbb{P}_{2n-1}$, also

$$\sum_{j=1}^{n} w_j\, p(x_j) = \int_a^b p(x)\,\mathrm{d}x \qquad \text{für alle } p \in \mathbb{P}_{2n-1},$$

dann nennen wir Q_n eine ***Gauß Quadraturformel*** *oder* ***Gauß Quadratur****.*

Wenn wir prüfen möchten, ob eine Integrationsformel (6.48) alle Polynome in $\mathbb{P}_m$ exakt integriert, reicht es diese für die Monome $p_\ell(x) = x^\ell$, $\ell = 0, 1, \ldots, m$, zu überprüfen, denn für ein beliebiges Polynom $p \in \mathbb{P}_m$ gilt

$$p(x) = \sum_{\ell=0}^{m} a_\ell \, x^\ell,$$

und damit folgen

$$I(p) = \int_{-1}^{1} p(x) \, \mathrm{d}x = \int_{-1}^{1} \left(\sum_{\ell=0}^{m} a_\ell \, x^\ell \right) \mathrm{d}x = \sum_{\ell=0}^{m} a_\ell \left(\int_{-1}^{1} x^\ell \, \mathrm{d}x \right) = \sum_{\ell=0}^{m} a_\ell \, I(x^\ell) \tag{6.49}$$

und

$$Q_n(p) = \sum_{j=1}^{n} w_j \left(\sum_{\ell=0}^{m} a_\ell \, x_j^\ell \right) = \sum_{\ell=0}^{m} a_\ell \left(\sum_{j=1}^{n} w_j x_j^\ell \right) = \sum_{\ell=0}^{m} a_\ell \, Q_n(x^\ell). \tag{6.50}$$

Gilt also $Q_n(x^\ell) = I(x^\ell)$ für alle $\ell = 0, 1, 2, \ldots, m$, so folgt aus (6.49) und (6.50) sofort, dass für alle $p \in \mathbb{P}_m$ ebenfalls $Q_n(p) = I(p)$ gilt.

Wir studieren das Problem der Konstruktion von Gauß Quadraturformeln zunächst für $n = 1$ und $n = 2$ separat.

Beispiel 6.36. (Gauß Quadratur mit $n = 1$ Knoten)
Wir betrachten also

$$\int_{-1}^{1} f(x) \, \mathrm{d}x \approx w_1 \, f(x_1) =: Q_1(f), \tag{6.51}$$

wobei das Gewicht w_1 und der Knoten(punkt) x_1 so gewählt werden sollen, dass die durch die rechte Seite von (6.51) gegebene Integrationsformel Q_1 alle Polynome bis zu einem möglichst hohen Grad exakt integriert.

Wir verlangen also zunächst in (6.51) Gleichheit für das konstante Polynom $p_0(x) = 1$, d.h. es soll gelten:

$$Q_1(p_0) = w_1 \, p_0(x_1) = w_1 \cdot 1 = w_1 \stackrel{!}{=} \int_{-1}^{1} p_0(x) \, \mathrm{d}x = \int_{-1}^{1} 1 \, \mathrm{d}x = 2 \quad \Longrightarrow \quad w_1 = 2 \tag{6.52}$$

Weiter verlangen wir in (6.51) Gleichheit für das lineare Polynom $p_1(x) = x$, d.h. es soll gelten:

$$Q_1(p_1) = w_1\, p_1(x_1) = w_1 \cdot x_1 \overset{!}{=} \int_{-1}^{1} p_1(x)\,\mathrm{d}x = \int_{-1}^{1} x\,\mathrm{d}x = \left[\frac{1}{2}\,x^2\right]_{x=-1}^{x=1} = 0$$

$$\Longrightarrow \qquad w_1\, x_1 = 0 \qquad \overset{w_1=2}{\Longrightarrow} \qquad 2\, x_1 = 0 \qquad \Longrightarrow \qquad x_1 = 0 \tag{6.53}$$

Mit $w_1 = 2$ und $x_1 = 0$ (aus (6.52) und (6.53)) erhalten wir also die Gauß Quadraturformel

$$\boxed{Q_1(f) = 2\, f(0),}$$

welche alle Polynome in $\mathbb{P}_1$ (also vom Grad $\leq 1 = 2 \cdot 1 - 1$) exakt integriert. ♠

Beispiel 6.37. (Gauß Quadratur mit $n = 2$ Knoten)
Wir betrachten also

$$\int_{-1}^{1} f(x)\,\mathrm{d}x \approx w_1\, f(x_1) + w_2\, f(x_2) =: Q_2(f), \tag{6.54}$$

wobei die Gewichte w_1 und w_2 und die Knoten(punkte) x_1 un x_2 so gewählt werden sollen, dass die durch die rechte Seite von (6.54) gegebene Integrationsformel Q_1 alle Polynome bis zu einem möglichst hohen Grad exakt integriert. – Wir verlangen also zunächst in (6.54) Gleichheit für das konstante Polynom $p_0(x) = 1$, d.h. es soll gelten:

$$Q_2(p_0) = w_1\, p_0(x_1) + w_2\, p_0(x_2) = w_1 \cdot 1 + w_2 \cdot 1 = w_1 + w_2 \overset{!}{=} \int_{-1}^{1} p_0(x)\,\mathrm{d}x$$

$$= \int_{-1}^{1} 1\,\mathrm{d}x = 2 \qquad \Longrightarrow \qquad w_1 + w_2 = 2 \tag{6.55}$$

Weiter verlangen wir in (6.54) Gleichheit für das lineare Polynom $p_1(x) = x$, d.h. es soll gelten:

$$Q_2(p_1) = w_1\, p_1(x_1) + w_2\, p_1(x_2) = w_1 \cdot x_1 + w_2 \cdot x_2 \overset{!}{=} \int_{-1}^{1} p_1(x)\,\mathrm{d}x$$

$$= \int_{-1}^{1} x\,\mathrm{d}x = \left[\frac{1}{2}\,x^2\right]_{x=-1}^{x=1} = 0 \qquad \Longrightarrow \qquad w_1\, x_1 + w_2\, x_2 = 0 \tag{6.56}$$

Weiter verlangen wir in (6.54) Gleichheit für das quadratische Polynom $p_2(x) = x^2$, d.h. es soll gelten:

$$Q_2(p_2) = w_1\, p_2(x_1) + w_2\, p_2(x_2) = w_1 \cdot x_1^2 + w_2 \cdot x_2^2 \overset{!}{=} \int_{-1}^{1} p_2(x)\,\mathrm{d}x$$

$$= \int_{-1}^{1} x^2 \,\mathrm{d}x = \left[\frac{1}{3}\,x^3\right]_{x=-1}^{x=1} = \frac{2}{3} \qquad \Longrightarrow \qquad w_1\,x_1^2 + w_2\,x_2^2 = \frac{2}{3} \quad (6.57)$$

Weiter verlangen wir in (6.54) Gleichheit für das kubische Polynom $p_3(x) = x^3$, d.h. es soll gelten:

$$Q_2(p_3) = w_1\,p_3(x_1) + w_2\,p_3(x_2) = w_1 \cdot x_1^3 + w_2 \cdot x_2^3 \stackrel{!}{=} \int_{-1}^{1} p_3(x)\,\mathrm{d}x$$

$$= \int_{-1}^{1} x^3 \,\mathrm{d}x = \left[\frac{1}{4}\,x^4\right]_{x=-1}^{x=1} = 0 \qquad \Longrightarrow \qquad w_1\,x_1^3 + w_2\,x_2^2 = 0 \quad (6.58)$$

Aus (6.55), (6.56), (6.57) und (6.58) erhalten wir die folgenden vier (teilweise nicht-linearen) Gleichungen in den vier Unbekannten w_1, w_2, x_1, x_2:

$$\begin{aligned} w_1 + w_2 &= 2, \\ w_1\,x_1 + w_2\,x_2 &= 0, \\ w_1\,x_1^2 + w_2\,x_2^2 &= \frac{2}{3}, \\ w_1\,x_1^3 + w_2\,x_2^3 &= 0. \end{aligned}$$

Dieses ist ein **nicht-lineares System von vier Gleichungen in vier Unbekannten**. Man kann zeigen, dass dieses die folgenden beiden Lösungen hat

$$\left(w_1 = w_2 = 1,\ x_1 = -\tfrac{\sqrt{3}}{3},\ x_1 = \tfrac{\sqrt{3}}{3}\right), \quad \left(w_1 = w_2 = 1,\ x_1 = \tfrac{\sqrt{3}}{3},\ x_1 = -\tfrac{\sqrt{3}}{3}\right),$$

welche bis auf eine Umnummerierung der Knoten identisch sind. Wir erhalten also die Gauß Quadraturformel

$$\boxed{Q_2(f) = 1 \cdot f\left(-\tfrac{\sqrt{3}}{3}\right) + 1 \cdot f\left(\tfrac{\sqrt{3}}{3}\right) = f\left(-\tfrac{\sqrt{3}}{3}\right) + f\left(\tfrac{\sqrt{3}}{3}\right),}$$

welche alle Polynome in $\mathbb{P}_3$ (also vom Grad $\leq 3 = 2 \cdot 2 - 1$) exakt integriert. ♠

Betrachten wir ein konkretes Beispiel.

Beispiel 6.38. (Gauß Quadratur mit Q_2)
Wir berechnen das Integral

$$I(f) = \int_{-1}^{1} e^x \,\mathrm{d}x = \left[e^x\right]_{x=-1}^{x=1} = e - e^{-1} \doteq 2{,}3504024$$

mit der in Beispiel 6.37 hergeleiteten Gauß Quadraturformel Q_2 angenähert:

$$Q_2(f) = e^{-\sqrt{3}/3} + e^{\sqrt{3}/3} \doteq 2{,}3426961$$

Der absolute Fehler ist $|Q_2(f) - I(f)| \doteq 0{,}00771$. Unter Berücksichtigung der Tatsache, dass nur zwei Knoten(punkte) verwendet wurden, ist der absolute Fehler ziemlich klein. ♠

Wie sieht es in Verallgemeinerung von Beispielen 6.36 und 6.37 mit der **Konstruktion von Gauß Quadraturformeln Q_n mit $n \in \mathbb{N}$ (die also auf $\mathbb{P}_{2n-1}$ exakt sind)** aus?

Wir wollen also für stetige Funktionen $f : [-1; 1] \to \mathbb{R}$ das Integral

$$I(f) = \int_{-1}^{1} f(x)\,\mathrm{d}x$$

mit einer numerischen Integrationsformel

$$Q_n(f) := \sum_{j=1}^{n} w_j\, f(x_j) \tag{6.59}$$

angenähert berechnen, und die Formel (6.59) soll Polynome bis zu einem möglichst hohen Grad m exakt berechnen. Die Quadraturformel (6.59) hat n Knoten(punkte) $x_1, x_2, \ldots, x_n$ und n zugehörige Gewichte $w_1, w_2, \ldots, w_n$. Wir können also insgesamt $2n$ Parameter wählen und erwarten daher, dass es möglich sein sollte, die $2n$ Monome $q_\ell(x) = x^\ell$, $\ell = 0, 1, 2, \ldots, 2n-1$, exakt zu integrieren. Damit würden dann auch auch alle Polynome in $\mathbb{P}_{2n-1}$ exakt integriert. Aus den Bedingungen $Q_n(x^\ell) = I(x^\ell)$ für alle $\ell = 0, 1, 2, \ldots, 2n-1$ erhalten wir ein **System mit** den folgenden $2n$ **nicht-linearen Gleichungen**:

$$\begin{aligned}
w_1 + w_2 + \ldots + w_n &= 2,\\
w_1\, x_1 + w_2\, x_2 + \ldots + w_n\, x_n &= 0,\\
w_1\, x_1^2 + w_2\, x_2^2 + \ldots + w_n\, x_n^2 &= \tfrac{2}{3},\\
w_1\, x_1^3 + w_2\, x_2^3 + \ldots + w_n\, x_n^3 &= 0,\\
&\vdots \quad \vdots \quad \vdots\\
w_1\, x_1^{2n-2} + w_2\, x_2^{2n-2} + \ldots + w_n\, x_n^{2n-2} &= \tfrac{2}{2n-1},\\
w_1\, x_1^{2n-1} + w_2\, x_2^{2n-1} + \ldots + w_n\, x_n^{2n-1} &= 0.
\end{aligned} \tag{6.60}$$

n	j	w_j	x_j
1	1	2,0000000000	0,0000000000
2	1	1,0000000000	−0,5773502692
	2	1,0000000000	0,5773502692
3	1	0,5555555556	−0,7745966692
	2	0,8888888889	0,0000000000
	3	0,5555555556	0,7745966692
4	1	0,3478548451	−0,8611363116
	2	0,6521451549	−0,3399810436
	3	0,6521451549	0,3399810436
	4	0,3478548451	0,8611363116
5	1	0,2369268851	−0,9061798459
	2	0,4786286705	−0.5384693101
	3	0,5688888889	0,0000000000
	4	0,4786286705	0,5384693101
	5	0,2369268851	0,9061798459
6	1	0,1713244924	−0,9324695142
	2	0,3607615730	−0,6612093865
	3	0,4679139346	−0,2386191861
6	4	0,4679139346	0,2386191861
	5	0,3607615730	0,6612093865
	6	0,1713244924	0,9324695142
7	1	0,1294849662	−0,9491079123
	2	0,2797053915	−0,7415311856
	3	0,3818300505	−0,4058451514
	4	0,4179591837	0,0000000000
	6	0,3818300505	0,4058451514
	6	0,2797053915	0,7415311856
	7	0,1294849662	0,9491079123
8	1	0,1012285363	−0,9602898565
	2	0,2223810345	−0,7966664774
	3	0,3137066459	−0,5255324099
	4	0,3626837834	−0,1834346425
	5	0,3626837834	0,1834346425
	6	0,3137066459	0,5255324099
	7	0,2223810345	0,7966664774
	8	0,1012285363	0,9602898565

Tabelle 6.4: Gewichte und Knoten der Gauß Quadraturformeln Q_n, $n = 1, \ldots, 8$, (mit einer Rundung auf eine Gleitkommadarstellung mit 10-stelliger Mantisse) zur numerischen Integration über $[-1; 1]$ mit einem Exaktheitsgrad von $2n - 1$.

Es ist in der Tat möglich, die n Knoten(punkte) $x_1, x_2, \ldots, x_n$ und n zugehörige Gewichte $w_1, w_2, \ldots, w_n$ so zu wählen, dass die $2n$ nicht-linearen Gleichungen in (6.60) alle erfüllt sind. Die Lösung dieses nicht-linearen Gleichungssystems ist allerdings (vor allem für größere Werte von n) eine sehr herausfordernde Aufgabe. Günstigerweise liegen diese Knoten und Gewichte aber in Tabellen vor, und in Tabelle 6.4 sind die Knoten und Gewichte der Gauß Quadraturformeln Q_n (mit einer Rundung auf eine Gleitkommadarstellung mit 10-stelliger Mantisse) für $n = 1, 2, \ldots, 8$ angegeben.

Betrachten wir ein Beispiel.

n	$Q_n(f)$	$\lvert Q_n(t) - I(f)\rvert$
1	2,0000000000	$3{,}50 \cdot 10^{-1}$
2	2,3426960879	$7{,}71 \cdot 10^{-3}$
3	2,3503369288	$6{,}55 \cdot 10^{-5}$
4	2,3504020921	$2{,}95 \cdot 10^{-7}$
5	2,3504023866	$7{,}08 \cdot 10^{-10}$

Tabelle 6.5: Gauß Quadratur $Q_n(f)$ zur Berechnung von $I(f) = \int_{-1}^{1} e^x \,\mathrm{d}x$.

Beispiel 6.39. (Gauß Quadratur)
Wir berechnen das Integral

$$I(f) = \int_{-1}^{1} e^x \,\mathrm{d}x = e - e^{-1} \doteq 2{,}350402387$$

aus Beispiel 6.38 nun mit den Gauß Quadraturformeln $Q_n(f)$ für $n = 1, 2, \ldots, 5$. Die Ergebnisse sind in Tabelle 6.5 zusammen mit den absoluten Fehlern angegeben. Wir sehen, das wir bereits mit fünf Knoten und fünf zugehörigen Gewichten eine sehr gute Näherung erreichen. ♠

Transfer für andere Integrationsintervalle: Wie berechnet man ein Integral

$$I(f) = \int_{a}^{b} f(x) \,\mathrm{d}x \qquad \text{mit stetigem} \qquad f : [a; b] \to \mathbb{R}, \tag{6.61}$$

bei dem $[a; b] \neq [-1; 1]$ ist? Hier hilft die **Substitutionsregel:** Mit der affin linearen Funktion

$$x = x(t) = \frac{b + a + t\,(b - a)}{2}, \qquad t \in [-1; 1], \tag{6.62}$$

wird das Intervall $[-1; 1]$ auf das Intervall $[a; b]$ abgebildet, denn $x(-1) = a$ und $x(1) = b$. Mit der Substitution $x = x(t)$ in (6.62) mit

$$\frac{\mathrm{d}x}{\mathrm{d}t} = \frac{b - a}{2} \qquad \Longleftrightarrow \qquad \mathrm{d}x = \frac{b - a}{2} \,\mathrm{d}t$$

folgt aus (6.61)

$$I(f) = \int_{a}^{b} f(x) \,\mathrm{d}x = \int_{-1}^{1} f\left(\frac{b + a + t\,(b - a)}{2}\right) \frac{b - a}{2} \,\mathrm{d}t$$

n	$Q_n(f)$	$\lvert Q_n(t) - I(f)\rvert$
1	0,7788007831	$3{,}20 \cdot 10^{-2}$
2	0,7465946883	$2{,}29 \cdot 10^{-4}$
3	0,7468145842	$9{,}55 \cdot 10^{-6}$
4	0,7468244681	$3{,}35 \cdot 10^{-7}$
5	0,7468241268	$6{,}01 \cdot 10^{-9}$
6	0,7468241329	$7{,}61 \cdot 10^{-11}$

Tabelle 6.6: Gauß Quadratur $Q_n(f)$ mit $f(x) = e^{-x^2}$ und $[a;b] = [0;1]$ zur Berechnung von $I(f) = \int_0^1 e^{-x^2}\,\mathrm{d}x$.

$$= \frac{b-a}{2} \int_{-1}^{1} \underbrace{f\left(\frac{b+a+t\,(b-a)}{2}\right)}_{=:\,\widetilde{f}(t)}\,\mathrm{d}t = \frac{b-a}{2} \int_{-1}^{1} \widetilde{f}(t)\,\mathrm{d}t \tag{6.63}$$

mit der neuen Funktion

$$\widetilde{f}(t) := f\big(x(t)\big) = f\left(\frac{b+a+t\,(b-a)}{2}\right), \qquad t \in [-1;1]. \tag{6.64}$$

Also können wir mittels

$$\boxed{\int_a^b f(x)\,\mathrm{d}x = \frac{b-a}{2} \int_{-1}^{1} \widetilde{f}(t)\,\mathrm{d}t \qquad \text{mit} \quad \widetilde{f}(t) := f\left(\frac{b+a+t\,(b-a)}{2}\right)} \tag{6.65}$$

Integrale über $[a;b]$ mit den eben eingeführten Gauß Quadraturformeln für numerische Integration über $[-1;1]$ berechnen. Nutzt man die Gauß Quadraturformel Q_n, um das Integral auf rechten Seite von (6.65) zu berechnen, so werden alle $f \in \mathbb{P}_n$ exakt integriert, denn durch die Transformation (6.64) wird ein Polynom $f \in \mathbb{P}_n$ wieder auf ein Polynom $\widetilde{f} \in \mathbb{P}_n$ abgebildet. Man erhält dann die folgende **Quadraturformel für eine Gauß Quadratur für das Intervall** $[a;b]$:

$$\boxed{\widetilde{Q}_n(f) = \frac{b-a}{2} \sum_{k=1}^{n} w_k\, f\big(x(t_k)\big),} \tag{6.66}$$

wobei t_k, $k = 1, 2, \ldots, n$, die Knoten und w_k, $k = 1, 2, \ldots, n$, die zugehörigen Gewichte der Gauß Quadratur Q_n für $[-1;1]$ sind. Die Gauß Quadratur (6.66)

für das Intervall $[a;b]$ hat dann also die Knoten $x_k := x(t_k)$, $k = 1, 2, \ldots, n$, und die zugehörigen Gewichte $\frac{b-a}{2}\, w_k$, $k = 1, 2, \ldots, n$.

Betrachten wir ein Beispiel dazu.

Beispiel 6.40. (Gauß Quadraturformel)
Wir berechnen das Integral

$$\int_0^1 e^{-x^2}\,\mathrm{d}x \doteq 0{,}746824132812427$$

(welches bereits in Beispiel 6.25 mit der Trapezregel für numerische Integration und in Beispiel 6.31 mit der Simpson-Regel für numerische Integration berechnet wurde) nun mit den Gauß Quadraturformeln Q_n. Hier ist $[a;b] = [0;1]$, und somit benötigen wir die affin lineare Transformation

$$x(t) = \frac{1 + 0 + t\,(1-0)}{2} = \frac{1+t}{2}\,.$$

Mit (6.65) erhalten wir

$$\int_0^1 e^{-x^2}\,\mathrm{d}x = \frac{1}{2}\int_{-1}^1 \widetilde{f}(t)\,\mathrm{d}t \quad \text{mit} \quad \widetilde{f}(t) := \exp\Big(-\big(x(t)\big)^2\Big) = \exp\Big(-\big(\tfrac{1+t}{2}\big)^2\Big)\,.$$

Also lautet die transformierte Formel für die Gauß Quadratur

$$Q_n(f) = \frac{1}{2}\sum_{k=1}^n w_k\, \exp\left(-\left(\frac{1+x_k}{2}\right)^2\right),$$

wobei $x_1, x_2, \ldots, x_n$ die Knoten(punkte) und $w_1, w_2, \ldots, w_n$ die zugehörigen Gewichte der Gauß Quadraturformel für $[-1;1]$ sind.

Die Ergebnisse sind in Tabelle 6.6 (mit Rundung auf eine Gleitkommadarstellung mit 10-stelliger Mantisse) angegeben. Mit Q_6, also mit nur $n = 6$ Knoten, erreichen wir ein ausgezeichnetes Ergebnis mit neun signifikanten Ziffern. Ein vergleichbares Ergebnis wird von der Simpson-Regel für numerische Integration erst für ein $64 < 2n \leq 128$ erreicht. Wir sehen, wie effizient die Gauß Quadratur (verglichen mit der populären Simpson-Regel für numerische Integration) ist. ♠

Zuletzt sollen noch zwei relevante Dinge kurz besprochen werden:

Der erste Aspekt betrifft die **Berechnung der Knoten der Gauß Quadraturformeln**: Es ist nicht nötig, die Knoten $x_1, x_2, \ldots, x_n$ über die $2n$ nicht-linearen Gleichungen in (6.60) zu bestimmen, denn die Knoten von Q_n sind genau die n

verschiedenen rellen Nullstellen des Legendre Polynoms P_n (und diese liegen auch alle im Intervall $[-1; 1]$). Die **Legendre Polynome** P_n, $n \in \mathbb{N}_0$, sind ein **System orthogonaler Polynome** mit den folgenden Eigenschaften:

(1) Für jedes $n \in \mathbb{N}_0$ gilt: P_n hat genau den Grad n.

(2) $\int_{-1}^{1} P_n(t)\, P_m(t)\, \mathrm{d}t = 0$ für alle $m, n \in \mathbb{N}_0$ mit $m \neq n$.

(3) $P_n(1) = 1$ für alle $n \in \mathbb{N}_0$.

Der zweite Aspekt betrifft eine Verallgemeinerung der Theorie: In manchen Anwendungen treten **Integrale mit einer Gewichtsfunktion** auf:

$$I(f) = \int_a^b f(x)\, w(x)\, \mathrm{d}t \qquad \text{mit stetigem} \qquad f : [a; b] \to \mathbb{R}, \tag{6.67}$$

wobei die **Gewichtsfunktion** $w : [a; b] \to \mathbb{R}$ nur nicht-negative Werte annimmt und höchstens an einzelnen Stellen in $[a; b]$ den Wert null annimmt. Ein Beispiel für ein Integral mit einer Gewichtsfunktion ist

$$\int_{-1}^{1} f(t)\, (1 - t^2)\, \mathrm{d}t \qquad \text{mit der Gewichtsfunktion} \quad w(t) = 1 - t^2.$$

In einem Anwendungsproblem könnte die Gewichtsfunktion w beispielsweise eine Dichtefunktion (Massendichte, Ladungsdichte) sein. Für Integrale (6.67) kann man dann auch Gauß Quadraturformeln konstruieren, bei denen der Effekt der Gewichtsfunktion direkt durch die Wahl der Knoten(punkte) und der zugehörigen Gewichte berücksichtigt ist.

6.7 Ausblick auf mehrdimensionale Quadratur: Tensorprodukt-Formeln

Die einfachste Möglichkeit, Quadraturformeln für die Integration über einen zweidimensionalen Integrationsbereich zu konstruieren, ist sogenannte **Tensorprodukt-Formeln** zu bilden. Dieses ist allerdings nur für zweidimensionale Mengen möglich, die sich als Integrationsbereich als **kartesisches Produkt zweier Intervalle** schreiben lassen. Solche Mengen sind beispielsweise

- Rechtecke $[a; b] \times [c; d]$,
- Kreisscheiben $\big\{\mathbf{x} \in \mathbb{R}^2 \ : \ \|\mathbf{x}\|_2 \leq R\big\}$, denn diese lassen sich in ebenen Polarkoordinaten $(\varrho; \phi)$ als $[0; R] \times [0; 2\pi]$ schreiben;

- die Sphäre $\mathbb{S}^2 := \{\mathbf{x} \in \mathbb{R}^3 \ : \ \|\mathbf{x}\|_2 = R\}$ in $\mathbb{R}^3$, denn diese lässt sich in Kugelkoordinaten mit festem Radius $r = R$ als $[0;\pi] \times [0;2\pi]$ schreiben;
- Zylindermäntel $\{[x;y;z]^T \in \mathbb{R}^3 \ : \ x^2 + y^2 = R^2, \ z \in [0;h]\}$, denn diese lassen sich in Zylinderkoordinaten dem mit festen Radius $\varrho = R$ als $[0;2\pi] \times [0;h]$ schreiben (bei einer Zylinderhöhe von h).

Wir erklären hier nur die Idee für Integration über ein **Rechteck** $[a;b] \times [c;d]$. Ist der Integrand $f : [a;b] \times [c;b] \to \mathbb{R}$ stetig, so können wir das Integral

$$I(f) = \int\limits_{[a;b]\times[c;d]} f(x;y)\,\mathrm{d}(x;y) \tag{6.68}$$

nach dem Satz von Fubini als

$$I(f) = \int_a^b \int_c^d f(x;y)\,\mathrm{d}y\,\mathrm{d}x = \int_c^d \int_a^b f(x;y)\,\mathrm{d}x\,\mathrm{d}y \tag{6.69}$$

berechnen. Ist Q_n eine Quadraturformel für numerische Integration über $[a;b]$ und ist $\widehat{Q}_m$ eine Quadraturformel für numerische Integration über $[c;d]$, also

$$Q_n(g) := \sum_{k=1}^{n} w_k\, g(x_k) \approx \int_a^b g(x)\,\mathrm{d}x \qquad \text{für stetiges } g : [a;b] \to \mathbb{R}, \tag{6.70}$$

$$\widehat{Q}_m(h) := \sum_{\ell=1}^{m} \widehat{w}_\ell\, h(y_\ell) \approx \int_c^d h(y)\,\mathrm{d}y \qquad \text{für stetiges } h : [c;d] \to \mathbb{R}, \tag{6.71}$$

so kann das Integral (6.68) numerisch (angenähert) berechnet werden, indem wir die beiden durch (6.70) und (6.71) gegebenen Quadraturformeln zur Berechnung der zwei Integrale in (6.69) wie folgt anwenden:

Für jedes feste $x \in [a;b]$ gilt mit der Quadraturformel (6.71)

$$\int_c^d f(x;y)\,\mathrm{d}y \approx \widehat{Q}_m\big(f(x;\cdot)\big) = \sum_{\ell=1}^{m} \widehat{w}_\ell\, f(x;y_\ell).$$

Anschließendes Anwenden der Quadraturformel (6.70) liefert nun

$$\begin{aligned}\int_a^b \int_c^d f(x;y)\,\mathrm{d}y\,\mathrm{d}x &\approx \int_a^b \left(\sum_{\ell=1}^{m} \widehat{w}_\ell\, f(x;y_\ell)\right) \mathrm{d}x \approx Q_n\left(\sum_{\ell=1}^{m} \widehat{w}_\ell\, f(\cdot;y_\ell)\right) \\ &= \sum_{k=1}^{n} w_k \sum_{\ell=1}^{m} \widehat{w}_\ell\, f(x_k;y_\ell) = \sum_{k=1}^{n}\sum_{\ell=1}^{m} w_k\, \widehat{w}_\ell\, f(x_k;y_\ell),\end{aligned}$$

d.h. für stetiges $f : [a; b] \times [c; b] \to \mathbb{R}$ gilt angenähert

$$\boxed{\int_a^b \int_c^d f(x; y)\,\mathrm{d}y\,\mathrm{d}x \approx \sum_{k=1}^{n} \sum_{\ell=1}^{m} w_k\,\widehat{w}_\ell\,f(x_k; y_\ell).} \tag{6.72}$$

Die Formel (6.72) ist eine **Tensorprodukt-Formel für numerische Integration über das Rechteck** $[a; b] \times [c; b]$ mit den Knoten(punkten)

$$(x_k; y_\ell), \qquad k = 1, 2, \ldots, n, \quad \ell = 1, 2, \ldots, m,$$

und den zugehörigen Gewichten

$$w_k\,\widehat{w}_\ell, \qquad k = 1, 2, \ldots, n, \quad \ell = 1, 2, \ldots, m.$$

Die Formel (6.72) wird eine Tensorprodukt-Formel genannt, weil sie als „Tensorprodukt“ der beiden eindimensionalen Integrationsformeln gebildet wird.

Betrachten wir ein Beispiel.

Beispiel 6.41. (Tensorprodukt-Quadraturformel)
Wir berechnen das Integral

$$\begin{aligned} I(f) &= \int\limits_{[0;1]\times[0;1]} y\,e^{xy}\,\mathrm{d}(x; y) = \int_0^1 \int_0^1 y\,e^{xy}\,\mathrm{d}x\,\mathrm{d}y \\ &= \int_0^1 \Big[e^{xy}\Big]_{x=0}^{x=1} \mathrm{d}y = \int_0^1 \big[e^y - 1\big]\,\mathrm{d}y = \Big[e^y - y\Big]_{y=0}^{y=1} \\ &= [e - 1] - [1 - 0] = e - 2 \doteq 0{,}7182818285 \end{aligned} \tag{6.73}$$

mit einer Tensorprodukt-Formel, die wir mit zwei Simpson-Regeln S_{2n} für numerische Integration bilden. Wir wenden also für jedes der beiden Intervalle über $[0; 1]$ (in $[0; 1] \times [0; 1]$) die gleiche Simpson-Regel S_{2n} für numerische Integration an

$$\begin{aligned} S_{2n}(g) &= \frac{1}{6n}\left[g(0) + 2\sum_{k=1}^{n-1} g\left(\tfrac{2k}{2n}\right) + 4\sum_{k=1}^{n} g\left(\tfrac{2k-1}{2n}\right) + g(1)\right] \\ &= \left[g(0) + 2\sum_{k=1}^{n-1} g\left(\tfrac{k}{n}\right) + 4\sum_{k=1}^{n} g\left(\tfrac{2k-1}{2n}\right) + g(1)\right] \end{aligned} \tag{6.74}$$

an. Damit erhalten wir durch eine Anwendung von S_{2n} auf das innere Integral in (6.73) mit $g(x) = y\,\exp(x\,y)$ zunächst

$$I(f) = \int_0^1 \int_0^1 y\,e^{xy}\,\mathrm{d}x\,\mathrm{d}y$$

$$\approx \frac{1}{6n}\int_0^1 \left(y + 2\sum_{k=1}^{n-1} y\,\exp\left(\tfrac{k}{n}\,y\right) + 4\sum_{k=1}^{n} y\,\exp\left(\tfrac{2k-1}{2n}\,y\right) + y\,e^y \right) \mathrm{d}y. \tag{6.75}$$

Nun wenden wir die Simpson-Regel S_{2n} erneut an, um das äußere Integral in (6.73), welches in in (6.75) noch angenähert werden muss, zu berechnen. Dabei ist die Funktion g in der Simpson-Regel für numerische Integration (6.74) durch

$$g(y) = y + 2\sum_{k=1}^{n-1} y\,\exp\left(\tfrac{k}{n}\,y\right) + 4\sum_{k=1}^{n} y\,\exp\left(\tfrac{2k-1}{2n}\,y\right) + y\,e^y$$

gegeben. Wir erhalten damit

$$\begin{aligned} I(f) &\approx \frac{1}{6n}\int_0^1 \left(y + 2\sum_{k=1}^{n-1} y\,\exp\left(\tfrac{k}{n}\,y\right) + 4\sum_{k=1}^{n} y\,\exp\left(\tfrac{2k-1}{2n}\,y\right) + y\,e^y \right) \mathrm{d}y \\ &= \frac{1}{(6n)^2}\Bigg[0 + 2\sum_{\ell=1}^{n-1}\left(\frac{\ell}{n} + 2\sum_{k=1}^{n-1}\frac{\ell}{n}\,\exp\left(\tfrac{k\ell}{n^2}\right) + 4\sum_{k=1}^{n}\frac{\ell}{n}\,\exp\left(\tfrac{(2k-1)\ell}{2n^2}\right) \right. \\ &\qquad \left. + \frac{\ell}{n}\,\exp\left(\tfrac{\ell}{n}\right)\right) + 4\sum_{\ell=1}^{n}\left(\frac{2\ell-1}{2n} + 2\sum_{k=1}^{n-1}\frac{2\ell-1}{2n}\,\exp\left(\tfrac{k(2\ell-1)}{2n^2}\right) \right. \\ &\qquad \left. + 4\sum_{k=1}^{n}\frac{2\ell-1}{2n}\,\exp\left(\tfrac{(2k-1)(2\ell-1)}{4n^2}\right) + \frac{2\ell-1}{2n}\,\exp\left(\tfrac{2\ell-1}{2n}\right)\right) \\ &\qquad + \left(1 + 2\sum_{k=1}^{n-1}\exp\left(\tfrac{k}{n}\right) + 4\sum_{k=1}^{n}\exp\left(\tfrac{2k-1}{2n}\right) + e \right)\Bigg] =: Q_{2n}(f). \end{aligned}$$

Die Ergebnisse sind in Tabelle 6.7 auf eine Gleitkommadarstellung mit 10-stelliger Mantisse gerundet zusammen mit dem absoluten Fehler $|Q_{2n}(f) - I(f)|$ für $2n = 2^j$ mit $j = 1, 2, \ldots, 6$ angegeben. Der absolute Fehler wurde dabei auf ein Gleitkommadarstellung mit 3-stelliger Mantisse gerundet angegeben. Für $2n = 64$, also mit $(64+1)^2 = 4.225$ Knotenpunkten, erhalten wir ein Ergebnis, welches acht signifikante Ziffern hat. Auch das Ergebnis für $2n = 32$, also mit $(32+1)^2 = 1.089$ Knotenpunkten, hat bereits sieben signifikante Ziffern und ist sehr gut.

Wir haben bei jeder neuen Berechnung den vorherigen Wert für $2n$ verdoppelt und damit den äquidistanten Abstand $h = 1/(2n)$ der Knoten halbiert. Dabei verkleinert sich wie bei der Simpson-Regel für numerische Integration der absolute Fehler um einen Faktor 1/16 (siehe letzte Spalte in Tabelle 6.7). Offenbar „erbt“ die Tensorprodukt-Quadraturformel diese Eigenschaft von der Simpson-Regel für numerische Integration. ♠

$2n$	$h = \frac{b-a}{2n} = \frac{1}{2n}$	$Q_{2n}(f)$	$\lvert Q_{2n}(f) - I(f)\rvert$	$\frac{\lvert Q_n(f) - I(f)\rvert}{\lvert Q_{2n}(f) - I(f)\rvert}$
$2 = 2^1$	0,5	0,7189670218	$6{,}85 \cdot 10^{-4}$	
$4 = 2^2$	0,25	0,7183246267	$4{,}28 \cdot 10^{-5}$	16,0
$8 = 2^3$	0,125	0,7182845133	$2{,}68 \cdot 10^{-6}$	16,0
$16 = 2^4$	0,0625	0,7182819965	$1{,}68 \cdot 10^{-7}$	16,0
$32 = 2^5$	0,03125	0,7182818390	$1{,}05 \cdot 10^{-8}$	16,0
$64 = 2^6$	0,015625	0,7182818291	$6{,}57 \cdot 10^{-10}$	16,0

Tabelle 6.7: Tensorprodukt-Quadraturformel Q_{2n} aus Beispiel 6.41 zur Berechnung des Integrals $\int\limits_{[0;1]\times[0;1]} y\, e^{xy}\, \mathrm{d}(x;y)$.

Natürlich hätten wir im letzten Beispiel für jedes der beiden Integrale auch eine unterschiedliche (eindimensionale) Quadraturformel zum Bilden einer Tensorprodukt-Formel verwenden können.

KAPITEL 7

Numerik für gewöhnliche Differentialgleichungen

Betrachten wir zum Einstieg ein Anwendungsproblem aus der Physik:

Sei $y(t)$ die zur Zeit t vorhandene Menge einer zerfallenden radioaktiven Substanz mit Zerfallskonstante $\lambda > 0$. Für eine kleine Zeitspanne $h := \Delta t \neq 0$ gilt

$$y(t+h) - y(t) \approx -\lambda\, y(t)\, h,$$

sofern $|h|$ klein genug ist. (Interpretation: Die Änderung der Menge der radioaktiven Substanz in dem kleinen Zeitintervall h ist also antiproportional zu der Menge der vorhanden Substanz.) Wir dividieren durch h und lassen h gegen 0 gehen:

$$\frac{y(t+h)-y(t)}{h} \approx -\lambda\, y(t) \qquad \Longrightarrow \qquad y'(t) = \lim_{h\to 0} \frac{y(t+h)-y(t)}{h} = -\lambda\, y(t).$$

Dieses führt auf die **gewöhnliche Differentialgleichung erster Ordnung**

$$y'(t) = -\lambda\, y(t).$$

Sie ist daher ein **mathematisches Modell für den radioaktiven Zerfall**. Alle Lösungen dieser Differentialgleichung sind von der Form

$$y(t) = c\, e^{-\lambda t}, \qquad \text{mit einer Konstante } c \in \mathbb{R}. \tag{7.1}$$

Damit man bestimmen kann, wie viel von der radioaktiven Substanz zum Zeitpunkt t vorhanden ist, braucht man die Information wie viel von der radioaktiven Substanz zu einem festen Zeitpunkt t_0 vorhanden war; wir brauchen also einen **Anfangswert** $y(t_0) = y_0$. Man nennt

$$y'(t) = -\lambda\, y(t) \qquad \text{mit} \qquad y(t_0) = y_0$$

ein **Anfangswertproblem**. Hat man die allgemeine Lösung von (7.1), so kann man die (in diesem Fall einzige) Lösung des Anfangswertproblems finden, indem man die Anfangsbedingung verwendet:

$$y_0 \stackrel{!}{=} y(t_0) = c\, e^{-\lambda t_0} \qquad \Longleftrightarrow \qquad c = y_0\, e^{\lambda t_0}$$

Also ist die Lösung des Anfangswertproblems

$$y(t) = y_0\, e^{\lambda t_0}\, e^{-\lambda t} = y_0\, e^{\lambda (t_0 - t)}.$$

In diesem Kapitel wiederholen wir zunächst einige Grundlagen zu gewöhnlichen Differentialgleichungen erster (und höherer) Ordnung und lernen dann die wichtigsten Resultate zur Existenz und Eindeutigkeit von Lösungen von gewöhnlichen Differentialgleichungen erster Ordnung kennen. Nach diesen Vorbereitungen führen wir das explizite Euler-Verfahren und danach das implizierte Euler-Verfahren ein. Beides sind Beispiele für sogenannte Einschrittverfahren. Wie besprechen die Begriffe der Konsistenz(-ordnung) und der Konvergenz(-ordnung) eines Einschrittverfahrens. Weitere wichtige Einschrittverfahren sind die (expliziten) Runge-Kutta-Verfahren, die wir ebenfalls ausführlich untersuchen. Zum Abschluss geben wir einen kurzen Ausblick auf Mehrschrittverfahren.

7.1 Wiederholung: Gewöhnliche Differentialgleichungen

Zur Erinnerung halten wir noch einmal fest, wie eine gewöhnliche Differentialgleichung erster Ordnung genau definiert ist.

Definition 7.1. (gewöhnliche Differentialgleichung 1. Ordnung)
Seien $D \subseteq \mathbb{R}^2$ offen und $f : D \to \mathbb{R}$ eine Funktion und $(t_0; y_0) \in D$.

*(1) Man nennt $y' = f(t; y)$ eine **gewöhnliche Differentialgleichung (DGL) erster Ordnung**. Dabei heißt t die unabhängige Variable, und y heißt die abhängige Variable.*

*(2) $y' = f(t; y)$ mit $y(t_0) = y_0$ heißt ein **Anfangswertproblem (AWP)**. (Dabei heißt $y(t_0) = y_0$ eine **Anfangsbedingung**.)*

*(3) Eine Funktion $y : I \to \mathbb{R}$ heißt eine **Lösung** von $y' = f(t; y)$, wenn*

(i) I ein offenes Intervall ist,

(ii) y auf I differenzierbar ist,

(iii) $\big(t; y(t)\big) \in D$ *für alle* $t \in I$ *gilt, und*

(iv) $y'(t) = f\big(t; y(t)\big)$ *für alle* $t \in I$ *gilt.*

(4) $y : I \to \mathbb{R}$ *heißt eine* ***Lösung des Anfangswertproblems*** $y' = f(t; y)$ *mit* $y(t_0) = y_0$*, wenn* y *eine Lösung von* $y' = f(t; y)$ *ist und die Anfangsbedingung* $y(t_0) = y_0$ *erfüllt.*

Ein wichtiger Sonderfall sind die linearen Differentialgleichungen erster Ordnung.

Definition 7.2. (lineare Differentialgleichung erster Ordnung)
Eine gewöhnliche Differentialgleichung erster Ordnung heißt ***linear****, wenn sie von der Form*

$$y' = a(t)\, y + b(t) \tag{7.2}$$

ist (oder sich durch elementare Umformungen auf diese Form bringen lässt). Dabei sind $a : I \to \mathbb{R}$ *und* $b : I \to \mathbb{R}$ *stetige Funktionen auf einem Intervall* I*.*

Ist $b(t) = 0$ *für alle* $t \in I$*, so heißt die Differentialgleichung (7.2)* ***homogen****, andernfalls nennt man sie* ***inhomogen****.*

Bei einer linearen gewöhnlichen Differentialgleichung erster Ordnung ist die Funktion f in Definition 7.1 also durch $f(t; y) = a(t)\, y + b(t)$ gegeben, d.h. f ist eine **affin lineare Funktion** in y.

Betrachten wir einige Beispiele.

Beispiel 7.3. (Differentialgleichungen erster Ordnung)

(a) $y' = \frac{3}{t}\, y$ ist linear und homogen. Hier ist $a(t) = 3/t$ und $b(t) = 0$.

(b) $y' = y + t^2$ ist linear und inhomogen. Hier ist $a(t) = 1$ und $b(t) = t^2$.

(c) $y' = y^2 + t$ ist nicht linear. Hier ist $f(t; y) = y^2 + t^2$

(d) $y' = e^t\, y + \frac{1}{t^2+1}$ ist linear und inhomogen. Hier ist $a(t) = e^t$ und $b(t) = \frac{1}{t^2+1}$.

(e) $y' = \sin(y) + \cos(t)$ ist nicht linear. Hier ist $f(t; y) = \sin(y) + \cos(t)$.

Sie kennen viele weitere Beispiele aus Ihren Mathematik-Vorlesungen. ♠

Natürlich kann es in einem Anwendungsproblem auch vorkommen, dass mehrere relevante physikalische Größen durch mehrere gewöhnliche Differentialgleichungen erster Ordnung in Beziehung zueinander gesetzt werden; dann erhält man ein System gewöhnlicher Differentialgleichungen erster Ordnung.

Definition 7.4. (Systeme gewöhnlicher DGLen und AWPe)
Seien $D \subseteq \mathbb{R}^{n+1}$ offen und $\mathbf{f} : D \to \mathbb{R}^n$.

(1) Die vektorielle Gleichung

$$\mathbf{y}' = \mathbf{f}(t; \mathbf{y}) \qquad \Longleftrightarrow \qquad \begin{cases} y_1' = f_1(t; y_1; y_2; \ldots; y_n), \\ y_2' = f_2(t; y_1; y_2; \ldots; y_n), \\ \vdots \qquad\quad \vdots \\ y_n' = f_n(t; y_1; y_2; \ldots; y_n), \end{cases} \tag{7.3}$$

*heißt ein **System gewöhnlicher Differentialgleichungen (DGLen) erster Ordnung**.*

*(2) Eine Funktion $\mathbf{y} : I \to \mathbb{R}^n$ heißt eine **Lösung von** (7.3), falls*

(i) I ein offenes Intervall ist,
(ii) $\mathbf{y}$ eine auf I stetig differenzierbare Funktion ist,
(iii) $\big(t; \mathbf{y}(t)\big) \in D$ für alle $t \in I$ gilt, und
(iv) $\mathbf{y}'(t) = \mathbf{f}\big(t; \mathbf{y}(t)\big)$ für alle $t \in I$ gilt.

*(3) **Anfangswertproblem:** Seien $(t_0; \mathbf{y}_0) \in D$ vorgegeben.*

*(i) Die Gleichung (7.3) zusammen mit der **Anfangsbedingung** $\mathbf{y}(t_0) = \mathbf{y}_0$ heißt dann ein **Anfangswertproblem (AWP)**.*
*(ii) Eine stetig differenzierbare Funktion $\mathbf{y} : I \to \mathbb{R}^n$, wobei I ein offenes Intervall ist, heißt eine **Lösung des Anfangswertproblems***

$$\mathbf{y}' = \mathbf{f}(t; \mathbf{y}) \qquad \textit{mit} \qquad \mathbf{y}(t_0) = \mathbf{y}_0,$$

falls $\mathbf{y}$ eine Lösung von (7.3) ist und zusätzlich $\mathbf{y}(t_0) = \mathbf{y}_0$ erfüllt.

Beispiel 7.5. (Räuber-Beute-Modell von Lotka-Volterra)
Im Räuber-Beute-Modell von Lotka-Volterra mit den Konstanten $\alpha, \beta \in]0; \infty[$,

$$y_1'(t) = \alpha \big(1 - y_2(t)\big)\, y_1(t), \tag{7.4}$$
$$y_2'(t) = \beta \big(y_1(t) - 1\big)\, y_2(t), \tag{7.5}$$

beschreibt $y_1(t)$ die Anzahl der Beutetiere (z.B. Mäuse) zum Zeitpunkt t und $y_2(t)$ die Anzahl der Raubtiere (z.B. Greifvögel) zum Zeitpunkt t. Dabei sind die Einheiten so gewählt, dass für $y_1 = y_2 = 1$ ein Gleichgewichtszustand eintritt, d.h. dann entspricht die Vermehrung genau der Abnahme durch Sterben und Gefressenwerden. Es sind $f_1(t; y_1; y_2) = \alpha\,(1 - y_2)\, y_1$, $f_2(t; y_1; y_2) = \beta\,(y_1 - 1)\, y_2$.

Was besagen (7.4) und (7.5)? Die Änderungsrate $y_1'(t)$ bzw. $y_2'(t)$ der Anzahl der Beutetiere bzw. der Raubtiere ist jeweils proportional zu der Anzahl der Beutetiere $y_1(t)$ bzw. der Raubtiere $y_2(t)$. Die Proportionalitätsfaktoren sind $\alpha\big(1-y_2(t)\big)$ bzw. $\beta\big(y_1(t)-1\big)$ und hängen jeweils von der anderen Funktion ab. Gilt zu einem Zeitpunkt $y_1(t)=y_2(t)=1$, so verschwinden beide Proportionalitätsfaktoren und die Zahlen der Beutetiere und der Raubtiere ändern sich zu dem Zeitpunkt nicht. Ist zu einem Zeitpunkt $y_1(t)>1$, so ist der Proportionalitätsfaktor $\beta\big(y_1(t)-1\big)$ in (7.5) positiv und die Anzahl $y_2(t)$ der Raubtiere nimmt zu. (Da viele Beutetiere zum Fressen da sind, können sich die Raubtiere gut vermehren.) Überschreitet die Anzahl der Raubtiere $y_2(t)$ irgendwann Wert 1, also $y_2(t)>1$, so wird der Proportionalitätsfaktor $\alpha\big(1-y_2(t)\big)$ in (7.4) negativ und die Anzahl der Beutetiere $y_1(t)$ nimmt ab. (Die vielen Raubtiere fressen mehr Beutetiere, als neue geboren werden.) Wenn die Anzahl $y_1(t)$ den Wert 1 unterschreitet, also $y_1(t)<1$, dann wird der Proportionalitätsfaktor $\beta\big(y_1(t)-1\big)$ in (7.5) negativ und die Anzahl der Raubtiere $y_2(t)$ nimmt ab (weil es nicht so viele Beutetiere zu fressen gibt). Wenn die Anzahl der Raubtiere den Wert 1 unterschreitet, also $y_2(t)<1$, dann wird der Proportionalitätsfaktor $\alpha\big(1-y_2(t)\big)$ in (7.4) positiv und die Anzahl $y_1(t)$ der Beutetiere nimmt wieder zu, bis ihre Anzahl irgendwann den Wert 1 überschreitet. (Es sind nicht mehr so viele Raubtiere da, so dass sich die Beutetiere schneller vermehren können, als sie gefressen werden.) Der „Kreislauf“ bei der Entwicklung der Beutetier- und Raubtier-Populationen fängt nun wieder von vorne an. ♠

Was passiert bei einer **gewöhnliche Differentialgleichung höherer Ordnung**

$$y^{(m)} = f(t; y; y'; y''; \ldots; y^{(m-1)}), \tag{7.6}$$

wobei $f : D \to \mathbb{R}$, mit $D \subseteq \mathbb{R}^{m+1}$ offen, eine Funktion und $m \geq 2$ ist?

Die gewöhnliche Differentialgleichung (7.6) m-ter Ordnung lässt sich wie folgt **in ein System m gewöhnlicher Differentialgleichungen erster Ordnung umschreiben**, so dass wir (7.6) mit numerischen Lösungsverfahren für Systeme gewöhnlicher Differentialgleichungen behandeln können:

$$\begin{aligned} z_1' &= z_2, \\ z_2' &= z_3, \\ &\vdots \\ z_{m-1}' &= z_m, \\ z_m'(t) &= f(t; z_1; z_2; \ldots; z_m). \end{aligned}$$

7.2 Existenz und Eindeutigkeit

In diesem Teilkapitel betrachten wir der Einfachheit halber „nur“ gewöhnliche Differentialgleichungen $y' = f(t; y)$ erster Ordnung und keine Systeme gewöhnlicher Differentialgleichungen. (Analoge Resultate gelten für Systeme gewöhnlicher Differentialgleichungen erster Ordnung; siehe z.B. [14].)

Satz 7.6. (Existenz und Eindeutigkeit der Lösung einer DGL)
Sei $[a; b] \times \mathbb{R}$ *mit* $-\infty < a < b < \infty$, *und sei* $f : [a; b] \times \mathbb{R} \to \mathbb{R}$ *eine stetige Funktion, die* ***gleichmäßig Lipschitz-stetig in der zweiten Variablen*** *ist, d.h. es existiert eine „Lipschitz-Konstante“* $L \geq 0$, *so dass gilt*

$$|f(t; v) - f(t; w)| \leq L\,|v - w| \quad \text{für alle } (t; v), (t; w) \in [a; b] \times \mathbb{R}. \tag{7.7}$$

Dann gibt es zu jedem Paar $(t_0; y_0) \in [a; b] \times \mathbb{R}$ *eine* ***eindeutig bestimmte*** *Funktion* $y : [a; b] \to \mathbb{R}$ *mit den folgenden Eigenschaften:*

(i) y *ist stetig differenzierbar in* $[a; b]$,

(ii) $y' = f\big(t; y(t)\big)$ *für alle* $t \in [a; b]$,

(iii) $y(t_0) = y_0$.

(In Worten: Es gibt eine eindeutig bestimmte stetig differenzierbare Lösung $y : [a; b] \to \mathbb{R}$ *des Anfangswertproblems* $y' = f(t; y)$ *mit* $y(t_0) = y_0$.*) Man nennt* (7.7) *eine* ***Lipschitz-Bedingung an*** f *(in der zweiten Variable).*

Betrachten wir die **Lipschitz-Bedingung** (7.7) genauer: Falls die partielle Ableitung $\frac{\partial f}{\partial y}$ von f nach der zweiten Variablen überall auf $[a; b] \times \mathbb{R}$ existiert und stetig und beschränkt ist, so folgt mit dem Mittelwertsatz der Differentialrechnung, dass die Lipschitz-Bedingung (7.7) erfüllt ist, denn: Nach dem Mittelwertsatz der Differentialrechnung gibt es für jedes $t \in [a; b]$ zu $v, w \in \mathbb{R}$ mit $v \neq w$ ein z zwischen v und w, so dass

$$f(t; v) - f(t; w) = \frac{\partial f(t; z)}{\partial y}\,(v - w) \quad \Longrightarrow \quad |f(t; v) - f(t; w)| = \left|\frac{\partial f(t; z)}{\partial y}\right|\,|v - w|.$$

Ist $\frac{\partial f}{\partial y}$ auf $[a; b] \times \mathbb{R}$ beschränkt, so gibt es eine Konstante $L \geq 0$ mit

$$\left|\frac{\partial f(t; y)}{\partial y}\right| \leq L \qquad \text{für alle } (t; y) \in [a; b] \times \mathbb{R}.$$

Anwenden dieser Abschätzung in der vorherigen Gleichung liefert

$$|f(t;v) - f(t;w)| = \left|\frac{\partial f(t;z)}{\partial y}\right| |v-w| \leq L\,|v-w|. \tag{7.8}$$

Da $v, w \in \mathbb{R}$ mit $v \neq w$ beliebig waren, liefert uns (7.8) die Lipschitz-Bedingung (7.7). (Falls $v = w$ gilt, sind beide Seiten in (7.7) null, und die Lipschitz-Bedingung ist für $v = w$ automatisch ebenfalls erfüllt.)

Der nächste Satz zeigt, dass die Lösung des Anfangswertproblems stetig von den Anfangswerten abhängt.

Satz 7.7. (Lösung einer DGL hängt stetig vom Anfangswert ab)
Seien $[a;b] \times \mathbb{R}$ *mit* $-\infty < a < b < \infty$, $t_0 \in [a;b]$, *und sei* $f : [a;b] \times \mathbb{R} \to \mathbb{R}$ *eine stetige Funktion, die* ***gleichmäßig Lipschitz-stetig in der zweiten Variablen*** *ist, d.h. es existiert eine „Lipschitz-Konstante“* $L \geq 0$, *so dass gilt*

$$|f(t;v) - f(t;w)| \leq L\,|v-w| \quad \textit{für alle } (t;v), (t;w) \in [a;b] \times \mathbb{R}.$$

Sei $y(\cdot\,;z) : [a;b] \to \mathbb{R}$ *(mit dem Parameter* $z \in \mathbb{R}$*) die eindeutig bestimmte Lösung des Anfangswertproblems*

$$y' = f(t;y) \qquad \textit{mit} \qquad y'(t_0) = z. \tag{7.9}$$

Dann gilt für die eindeutig bestimmten Lösungen $y(\cdot\,;z_1)$ *bzw.* $y(\cdot\,;z_2)$ *von* (7.9) *mit* $z = z_1$ *bzw.* $z = z_2$ *die Abschätzung*

$$|y(t;z_1) - y(t;z_2)| \leq e^{L|t-t_0|}\,|z_1 - z_2| \quad \textit{für alle } t \in [a;b] \textit{ und alle } z_1, z_2 \in \mathbb{R}. \tag{7.10}$$

Was besagt die Ungleichung (7.10)**?**

Die Funktionen $y(\cdot\,;z_1)$ bzw. $y(\cdot\,;z_2)$ sind die Lösungen der gleichen Differentialgleichung $y' = f(t;y)$ mit verschiedenen Anfangswerten, nämlich $y(t_0) = z_1$ bzw. $y(t_0) = z_2$. Auf der linken Seite von (7.10) steht die betragliche Abweichung der Lösungen $y(\cdot\,;z_1)$ bzw. $y(\cdot\,;z_2)$ zum Zeitpunkt t, und auf der rechten Seite von (7.10) steht eine obere Schranke für diese Abweichung, nämlich $e^{L|t-t_0|}\,|z_1 - z_2|$.

Sind z_1 und z_2 beliebig dicht beieinander, so wird wegen des Faktors $|z_1 - z_2|$ auch die obere Schranke beliebig klein. Wir haben also eine **stetige Abhängigkeit der Lösung der Differentialgleichung** $y' = f(t;y)$ **vom Anfangswert**.

Der andere Faktor $e^{L|t-t_0|}$ in der oberen Schranke $e^{L|t-t_0|}\,|z_1 - z_2|$ beschreibt,

wie sich die obere Schranke mit der Zeit verändert. Je weiter wir zeitlich von Zeitpunkt t_0 der Anfangsbedingung entfernt sind, desto größer wird $e^{L|t-t_0|}$. Physikalisch bedeutet dieses, dass sich **Lösungen** der gleichen Differentialgleichung $y' = f(t; y)$ **mit dicht beieinander liegenden Anfangswerten** zum Zeitpunkt t_0 **mit dem Fortschreiten der Zeit immer weiter voneinander entfernen können**.

7.3 Euler-Verfahren und allgemeiner Einschrittverfahren

In diesem Teilkapitel lernen wir zunächst das explizite und danach das implizite Euler-Verfahren kennen. Beide bieten einen guten Startpunkt, um die wesentlichen Ideen zu (expliziten und impliziten) Einschrittverfahren einzuführen.

In diesem Teilkapitel seien $[a; b] \times \mathbb{R}$ mit $-\infty < a < b < \infty$ und $t_0 \in [a; b]$, und die Funktion $f : [a; b] \times R \to \mathbb{R}$ sei stetig. Weiter gebe es eine Konstante $L \geq 0$, so dass gilt $|f(t; v) - f(t; w)| \leq L\,|v - w|$ für alle $(t; v), (t; w) \in [a; b] \times \mathbb{R}$. Nach Satz 7.6 hat dann das Anfangswertproblems

$$y' = f\big(t; y\big) \qquad \text{mit} \qquad y(t_0) = y_0 \tag{7.11}$$

eine eindeutige Lösung $y : [a; b] \to \mathbb{R}$.

Das **explizite Euler-Verfahren** ist das einfachste Verfahren zur numerischen Lösung eines Anfangswertproblems (7.11). Das Euler-Verfahren ergibt sich, indem in (7.11) die Ableitung $y'(t)$ durch den **Vorwärts-Differenzenquotienten**

$$y'(t) \approx \frac{y(t+h) - y(t)}{h} \tag{7.12}$$

mit einer **festen Schrittweite** $h \neq 0$ ersetzt wird. Einsetzen von (7.12) in (7.11) ergibt die Näherungsformel

$$\frac{y(t+h) - y(t)}{h} \approx f\big(t; y(t)\big),$$

und Auflösen nach $y(t + h)$ liefert

$$y(t+h) \approx y(t) + h\, f\big(t; y(t)\big). \tag{7.13}$$

Beginnend mit dem Wertepaar $(t_0; y_0)$ aus der Anfangsbedingung $y(t_0) = y_0$ berechnen wir nun mit der Näherung aus (7.13) die folgende Näherung für $y(t_1)$ mit $t_1 = t_0 + h$

$$y_1 = y_0 + h\, f(t_0; y_0).$$

Analog erhalten wir als Näherung für $y(t_2)$ mit $t_2 = t_1 + h = t_0 + 2\,h$

$$y_2 = y_1 + h\,f(t_1; y_1).$$

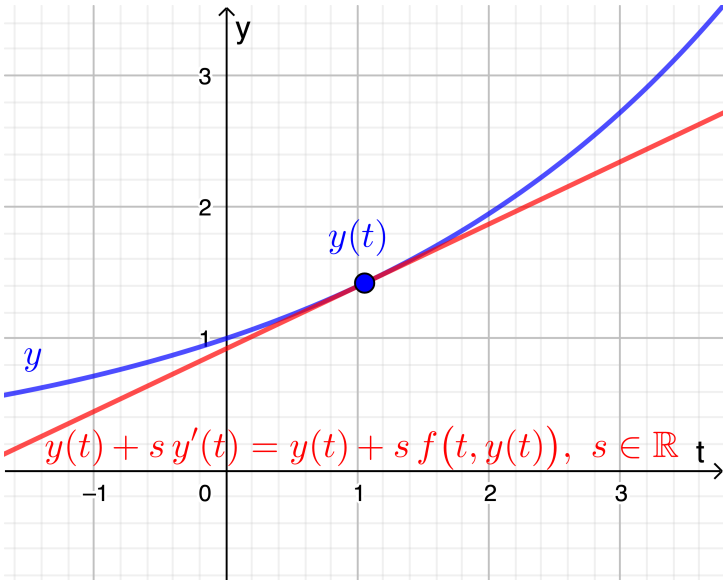
Wir setzen den Prozess fort und berechnen im $(k+1)$-ten Schritt die folgende Näherung für den Funktionswert $y(t_{k+1})$ (der eindeutigen Lösung y des Anfangswertproblems (7.11)) rekursiv (mit Hilfe der Näherung y_k aus dem vorherigen Schritt):

$$\boxed{y_{k+1} = y_k + h\,f(t_k; y_k), \qquad \text{wobei} \quad t_{k+1} = t_k + h = t_0 + (k+1)\,h.} \tag{7.14}$$

Die Formel (7.13) hat eine geometrische Interpretation (vgl. Abbildung rechts): Die Gerade

$$\begin{aligned} g(s) &= y(t) + s\,f\big(t; y(t)\big) \\ &= y(t) + s\,y'(t), \quad s \in \mathbb{R}, \end{aligned}$$

ist die Tangente an den Graphen von y im Punkt $\big(t; y(t)\big)$. Also wird $y(t+h)$ durch den entsprechenden Funktionswert $y(t) + h\,f\big(t; y(t)\big)$ der Tangente im Punkt $t+h$ angenähert.



Die mit (7.14) berechneten Näherungen y_k, $k = 1, 2, 3, \ldots$, stimmen in der Regel nicht mit den Funktionswerten $y(t_k)$, $k = 1, 2, 3, \ldots$, der eindeutigen Lösung y des Anfangswertproblems (7.11) überein.

In Abbildung 7.1 auf Seite 253 ist das explizite Euler-Verfahren für das Anfangswertproblem $y'(t) = \frac{1}{2}\,y(t)$ mit $y(0) = 1$ illustriert. In der Grafik in Abbildung 7.1 wurden die Näherungen $(t_k; y_k)$, $k = 0, 1, 2, 3$, der Punkte $\big(t_k; y(t_k)\big)$, $k = 0, 1, 2, 3$, auf den Graphen mit einem Polygonzug verbunden. Normalerweise wird man keinen Polygonzug (also keine stückweise lineare Interpolierende) sondern eine geeignete „glatte“ (d.h. mindestens stetig differenzierbare) Interpolierende (vgl. auch Teilkapitel 6.1) der Daten $(t_k; y_k)$, $k = 0, 1, 2, \ldots$, verwenden, um auch zwischen den Werten $(t_k; y_k)$, $k = 0, 1, 2, \ldots$, eine Näherung der eindeutigen Lösung y des Anfangswertproblems (7.11) zu bekommen.

Wir haben also das nachfolgende Iterationsverfahren hergeleitet:

Verfahren 7.8. (explizites Euler-Verfahren)
Seien $[a;b] \times \mathbb{R}$ *mit* $-\infty < a < b < \infty$, $t_0 \in [a;b]$, *und* $f : [a;b] \times \mathbb{R} \to \mathbb{R}$ *sei eine stetige Funktion, für die es eine Konstante* $L \geq 0$ *gilt, so dass gilt* $|f(t;v) - f(t;w)| \leq L\,|v - w|$ *für alle* $(t;v), (t;w) \in [a;b] \times \mathbb{R}$. *Das* ***explizite Euler-Verfahren*** *zur numerischen Lösung das Anfangswertproblems*

$$y' = f(t;y) \qquad mit \qquad y(t_0) = y_0 \tag{7.15}$$

konstruiert Näherungen y_k *der Funktionswerte* $y(t_k)$ *der eindeutigen Lösung* $y : [a;b] \to \mathbb{R}$ *von* (7.15) *an den* ***äquidistanten Gitterpunkten***

$$t_k = t_0 + k\,h, \qquad k = 0, 1, 2, \ldots,$$

mit der ***Schrittweite*** $h \neq 0$ *mit der folgenden* ***Iterationsformel****:*

Für $k = 0, 1, 2, \ldots$ *führe die folgenden Schritte durch:*

(1) Berechne $y_{k+1} := y_k + h\,f(t_k; y_k)$.
(2) Setze $t_{k+1} := t_k + h$.

Betrachten wir zwei Beispiele.

Beispiel 7.9. (explizites Euler-Verfahren)
Wir betrachten das Anfangswertproblem

$$y'(t) = -y(t) \qquad \text{mit} \qquad y(0) = 1,$$

dessen eindeutige Lösung durch $y(t) = e^{-t}$ gegeben ist. Hier ist also $f(t;y) = -y$. Das explizite Euler-Verfahren mit der Schrittweite $h \neq 0$ hat die Iterationsformel

$$y_{k+1} = y_k + h\,f(t_k; y_k) = y_k - h\,y_k, \qquad t_{k+1} = t_k + h, \qquad k = 0, 1, 2, \ldots,$$

mit den Startwerten $t_0 = 0$ und $y_0 = 1$.

Das explizite Euler-Verfahren wurde jeweils mit einer der Schrittweiten $h_1 = 0{,}2$, $h_2 = 0{,}1$ und $h_3 = 0{,}05$ durchgeführt, und in Tabelle 7.1 sind jeweils die berechneten Näherungswerte (gerundet auf eine Gleitkommadarstellung mit 5-stelliger Mantisse) für $y(t)$ mit $t \in \{1, 2, 3, 4, 5\}$ sowie deren absolute und relative Fehler angegeben.

Wir machen in Tabelle 7.1 folgende Beobachtungen:

- Es kann durchaus passieren, dass der absolute Fehler mit wachsendem k wieder kleiner wird. (Hier treten unter den angegebenen Werten die maxi-

h	k	t_k	y_k	$\lvert y_k - y(t_k)\rvert$	$\lvert y_k - y(t_k)\rvert / \lvert y(t_k)\rvert$
$h_1 = 0{,}2$	5	1,0	$3{,}2768 \cdot 10^{-1}$	$4{,}02 \cdot 10^{-2}$	0,109
	10	2,0	$1{,}0738 \cdot 10^{-1}$	$2{,}80 \cdot 10^{-2}$	0,207
	15	3,0	$3{,}5184 \cdot 10^{-2}$	$1{,}46 \cdot 10^{-2}$	0,293
	20	4,0	$1{,}1529 \cdot 10^{-2}$	$6{,}79 \cdot 10^{-3}$	0,371
	25	5,0	$3{,}7779 \cdot 10^{-3}$	$2{,}96 \cdot 10^{-3}$	0,439
$h_2 = 0{,}1$	10	1,0	$3{,}4867 \cdot 10^{-1}$	$1{,}92 \cdot 10^{-2}$	0,0552
	20	2,0	$1{,}2158 \cdot 10^{-1}$	$1{,}38 \cdot 10^{-2}$	0,102
	30	3,0	$4{,}2391 \cdot 10^{-2}$	$7{,}40 \cdot 10^{-3}$	0,149
	40	4,0	$1{,}4781 \cdot 10^{-2}$	$3{,}53 \cdot 10^{-3}$	0,193
	50	5,0	$5{,}1538 \cdot 10^{-3}$	$1{,}58 \cdot 10^{-3}$	0,234
$h_3 = 0{,}05$	20	1,0	$3{,}5849 \cdot 10^{-1}$	$9{,}39 \cdot 10^{-3}$	0,0255
	40	2,0	$1{,}2851 \cdot 10^{-1}$	$6{,}82 \cdot 10^{-3}$	0,0504
	60	3,0	$4{,}6070 \cdot 10^{-2}$	$3{,}72 \cdot 10^{-3}$	0,0747
	80	4,0	$1{,}6515 \cdot 10^{-2}$	$1{,}80 \cdot 10^{-3}$	0,0983
	100	5,0	$5{,}9205 \cdot 10^{-3}$	$8{,}17 \cdot 10^{-4}$	0,121

Tabelle 7.1: Explizites Euler-Verfahren für $y' = -y$ mit $y(0) = 1$.

malen absoluten Fehler bei $t = 1{,}0$ auf.) Der relative Fehler nimmt aber in diesem Beispiel für alle drei Schrittweiten mit wachsendem k zu.

- Die Schrittweite wurde jeweils halbiert, d.h. $h_2 = 0{,}1 = \frac{1}{2} h_1$ und $h_3 = 0{,}05 = \frac{1}{2} h_2$. Betrachtet man die absoluten und relativen Fehler in demselben Punkt $t \in \{1; 2; 3; 4; 5\}$ für die verschiedenen Schrittweiten, so sehen wir, dass sich mit der Halbierung der Schrittweite auch die absoluten und relativen Fehler ungefähr halbiert haben.

Abschließend kann man sagen, das die Näherungen selbst bei der kleinsten Schrittweite $h_3 = 0{,}05$ bei $t = 5$ bereits einen relativen Fehler von ca. 12 % haben, was keine gute Näherung ist. ♠

Beispiel 7.10. (explizites Euler-Verfahren)
Wir betrachten das Anfangswertproblem

h	k	t_k	y_k	$\lvert y_k - y(t_k)\rvert$	$\lvert y_k - y(t_k)\rvert / \lvert y(t_k)\rvert$
$h_1 = 0{,}2$	5	1,0	2,1592	$6{,}82 \cdot 10^{-2}$	0,0306
	10	2,0	3,1697	$2{,}39 \cdot 10^{-1}$	0,0701
	15	3,0	5,4332	$4{,}76 \cdot 10^{-1}$	0,0805
	20	4,0	9,1411	$7{,}65 \cdot 10^{-1}$	0,0772
	25	5,0	14,406	$1{,}09 \cdot 10^{0}$	0,0703
	30	6,0	21,303	$1{,}45 \cdot 10^{0}$	0,0637
$h_2 = 0{,}1$	10	1,0	2,1912	$3{,}63 \cdot 10^{-2}$	0,0163
	20	2,0	3,2841	$1{,}24 \cdot 10^{-1}$	0,0364
	30	3,0	5,6636	$2{,}46 \cdot 10^{-1}$	0,0416
	40	4,0	9,5125	$3{,}93 \cdot 10^{-1}$	0,0397
	50	5,0	14,939	$5{,}60 \cdot 10^{-1}$	0,0361
	60	6,0	22,013	$7{,}44 \cdot 10^{-1}$	0,0327
$h_3 = 0{,}05$	20	1,0	2,2087	$1{,}87 \cdot 10^{-2}$	0,00840
	40	2,0	3,3449	$6{,}34 \cdot 10^{-2}$	0,0186
	60	3,0	5,7845	$1{,}25 \cdot 10^{-1}$	0,0212
	80	4,0	9,7061	$1{,}99 \cdot 10^{-1}$	0,0201
	100	5,0	15,214	$2{,}84 \cdot 10^{-1}$	0,0183
	120	6,0	22,381	$3{,}76 \cdot 10^{-1}$	0,0165

Tabelle 7.2: Explizites Euler-Verfahren für $y' = \frac{y+t^2-2}{t+1}$ mit $y(0) = 2$.

$$y'(t) = \frac{y(t) + t^2 - 2}{t + 1} \qquad \text{mit} \qquad y(0) = 2,$$

dessen eindeutige Lösung durch

$$y(t) = t^2 + 2\,t + 2 - 2\,(t+1)\,\ln(t+1)$$

gegeben ist. Hier ist also $f(t; y) = \frac{y+t^2-2}{t+1}$. Das explizite Euler-Verfahren mit der Schrittweite $h \neq 0$ hat die Iterationsformel

$$y_{k+1} = y_k + h\,f(t_k; y_k) = y_k + \frac{h\,(y_k + t_k^2 - 2)}{t_k + 1}\,, \qquad t_{k+1} = t_k + h,$$

$k = 0, 1, 2, \ldots$, mit den Startwerten $t_0 = 0$ und $y_0 = 2$.

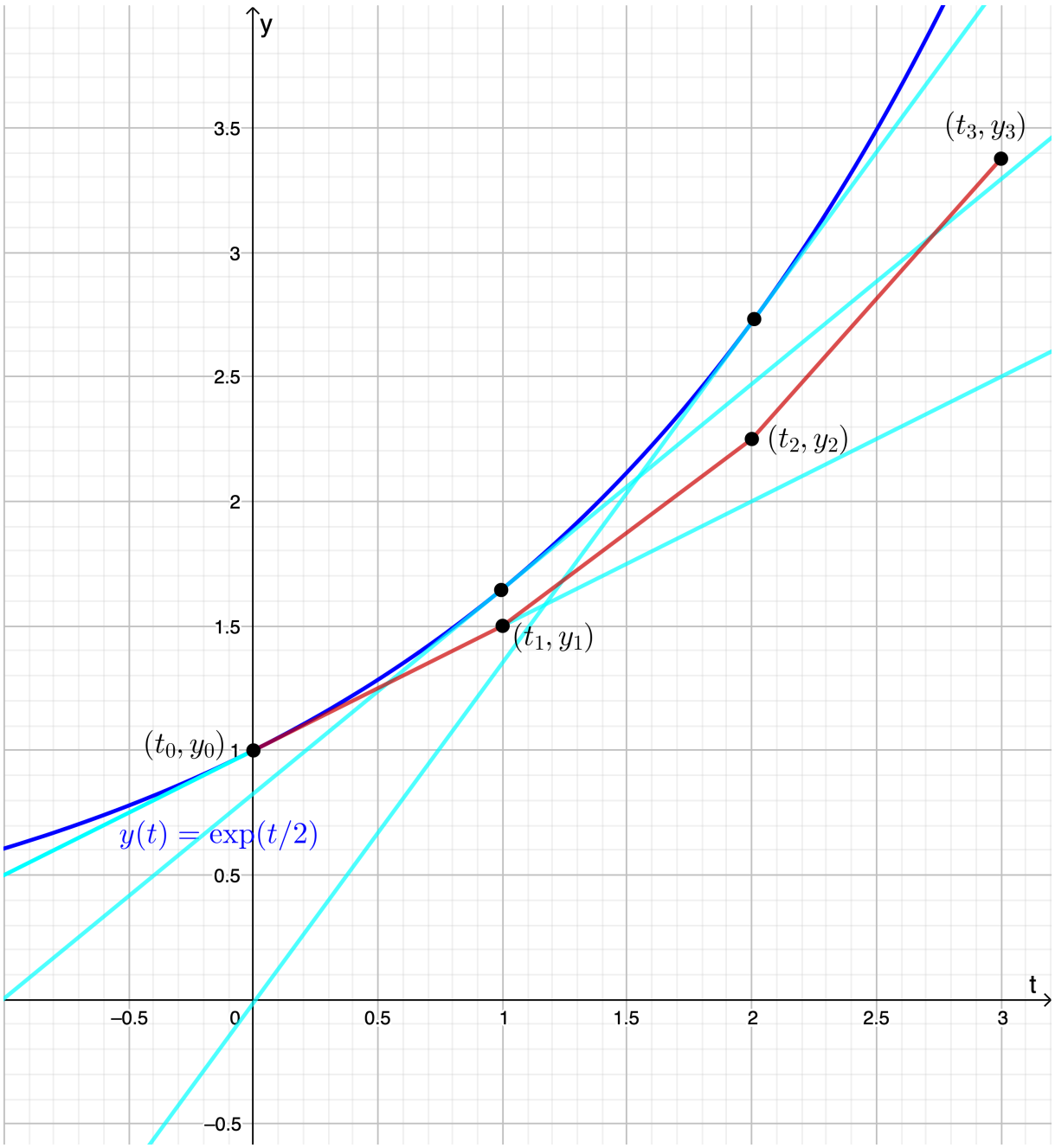


Abb. 7.1: Das Euler-Verfahren generiert Näherungswerte y_k für $y(t_k)$, hier verbunden durch einen Polygonzug (in rot), der sich mit wachsendem k oft weiter von der Lösung y des Anfangswertproblems (7.11) entfernt. Für die drei eingezeichneten Schritte des Euler-Verfahrens sind ebenfalls die Tangenten in den Punkten $\big(t_k; y(t_k)\big)$ an den Graphen von y mit den Steigungen $y'(t_k)$ eingezeichnet.

Das explizite Euler-Verfahren wurde jeweils mit einer der Schrittweiten $h_1 = 0{,}2$, $h_2 = 0{,}1$ und $h_3 = 0{,}05$ durchgeführt, und in Tabelle 7.2 sind jeweils die berechneten Näherungswerte (gerundet auf eine Gleitkommadarstellung mit 5-stelliger Mantisse) für $y(t)$ mit $t \in \{1; 2; 3; 4; 5; 6\}$ sowie deren absolute und relative Fehler angegeben.

Wir beobachten auch in diesem Beispiel, dass sich die absoluten und relativen Fehler ungefähr halbieren, wenn man die Schrittweite halbiert. Anders als in Beispiel 7.9 beobachten wir hier, dass mit wachsendem k die absoluten Fehler größer werden. Für jede der drei Schrittweiten werden die relativen Fehler in den angeben Punkten mit wachsendem k erst größer werden, aber dann ab $t = 3$ wieder kleiner. Auch hier ist die Qualität der Näherung selbst für die kleinste Schrittweite nicht besonders gut. ♠

Das explizite Euler-Verfahren ist das einfachste Einschrittverfahren, aber es ist wenig praxistauglich, da normalerweise die Qualität der berechneten Näherungen y_k für $y(t_k)$, $k = 0, 1, 2, \ldots$, nicht besonders gut ist.

Eine Verbesserung der expliziten Euler-Verfahrens erhält man, wenn man in der Herleitung des expliziten Euler-Verfahrens den Vorwärts-Differenzenquotienten durch den Rückwärts-Differenzenquotienten ersetzt. Dieses führt auf das **implizite Euler-Verfahren**: Zur numerischen Lösung eines Anfangswertproblems für eine gewöhnliche Differentialgleichung

$$y' = f(t; y) \qquad \text{mit} \qquad y(t_0) = y_0 \tag{7.16}$$

ersetzt man in (7.16) die Ableitung $y'(t)$ durch den **Rückwärts-Differenzenquotienten**

$$y'(t) \approx \frac{y(t) - y(t-h)}{h} \tag{7.17}$$

mit einer **festen Schrittweite** $h \neq 0$. Einsetzen von (7.17) in (7.16) ergibt

$$\frac{y(t) - y(t-h)}{h} \approx f\big(t; y(t)\big),$$

und Auflösen nach $y(t)$ liefert

$$y(t) \approx y(t-h) + h\, f\big(t; y(t)\big). \tag{7.18}$$

Beginnend mit einem Wertepaar $(t_0; y_0)$ berechnen wir nun mit der Näherung aus (7.18) die folgende Näherung für $y(t_1)$ mit $t_1 = t_0 + h \iff t_0 = t_1 - h$

$$y_1 = y_0 + h\, f(t_1; y_1).$$

Analog erhalten wir als Näherung für $y(t_2)$ mit $t_2 = t_1 + h = t_0 + 2\,h$

$$y_2 = y_1 + h\, f(t_2; y_2).$$

Wir setzen den Prozess fort und berechnen im $(k+1)$-ten Schritt die folgende Näherung für den Funktionswert $y(t_{k+1})$ (der Lösung y des Anfangswertproblems

(7.16)) rekursiv (mit Hilfe der Näherung y_k aus dem vorherigen Schritt):

$$\boxed{y_{k+1} = y_k + h\, f(t_{k+1}; y_{k+1}), \quad \text{wobei} \;\; t_{k+1} = t_k + h = t_0 + (k+1)\, h.} \tag{7.19}$$

Durch die Gleichung (7.19) ist y_{k+1} nur **implizit** gegeben (daher kommt der Name „implizites Euler-Verfahren"), denn, da y_{k+1} auch auf der rechten Seite von (7.19) in $f(t_{k+1}; y_{k+1})$ auftritt, müssen wir zur Bestimmung von y_{k+1} die **Gleichung**

$$y_{k+1} = y_k + h\, f(t_{k+1}; y_{k+1})$$

nach y_{k+1} **auflösen**. Dass dieses in jedem Schritt des impliziten Euler-Verfahrens möglich ist, bildet eine Voraussetzung dafür, dass wir das implizite Euler-Verfahren überhaupt anwenden können.

Auch hier gilt: Die mit den impliziten Euler-Verfahren berechneten Näherungen y_k, $k = 1, 2, 3, \ldots$, stimmen in der Regel nicht mit den Funktionswerten $y(t_k)$, $k = 1, 2, 3, \ldots$, der eindeutigen Lösung y des Anfangswertproblems (7.16) überein.

Damit erhalten wir das nachfolgende Iterationsverfahren:

Verfahren 7.11. (implizites Euler-Verfahren)
Seien $[a; b] \times \mathbb{R}$ *mit* $-\infty < a < b < \infty$, $t_0 \in [a; b]$, *und* $f : [a; b] \times \mathbb{R} \to \mathbb{R}$ *sei eine stetige Funktion, für die es eine Konstante* $L \geq 0$ *gilt, so dass gilt* $|f(t; v) - f(t; w)| \leq L\, |v - w|$ *für alle* $(t; v), (t; w) \in [a; b] \times \mathbb{R}$. *Das* ***implizite Euler-Verfahren*** *zur numerischen Lösung das Anfangswertproblems*

$$y' = f(t; y) \qquad \textit{mit} \qquad y(t_0) = y_0 \tag{7.20}$$

konstruiert Näherungen y_k *der Funktionswerte* $y(t_k)$ *der eindeutigen Lösung* $y : [a; b] \to \mathbb{R}$ *von* (7.20) *an den* ***äquidistanten Gitterpunkten***

$$t_k = t_0 + k\, h, \qquad k = 0, 1, 2, \ldots,$$

mit der ***Schrittweite*** $h \neq 0$ *mit der folgenden* ***Iterationsformel****:*

Für $k = 0, 1, 2, \ldots$ *führe die folgenden Schritte durch:*

(1) Setze $t_{k+1} := t_k + h$.

(2) Löse $y_{k+1} := y_k + h\, f(t_{k+1}; y_{k+1})$ *nach* y_{k+1} *auf.*

Betrachten wir ein Beispiel.

h	implizites Euler-Verfahren: Näherung für $y(0{,}2)$	explizites Euler-Verfahren: Näherung für $y(0{,}2)$
$h_1 = 0{,}1$	$8{,}26 \cdot 10^{-3}$	81
$h_2 = 0{,}05$	$7{,}72 \cdot 10^{-4}$	256
$h_3 = 0{,}02$	$1{,}69 \cdot 10^{-5}$	1
$h_4 = 0{,}01$	$9{,}54 \cdot 10^{-7}$	0
$h_5 = 0{,}001$	$5{,}72 \cdot 10^{-9}$	$7{,}06 \cdot 10^{-10}$

Tabelle 7.3: Implizites und explizites Euler-Verfahren zur Berechnung des Funktionswerts $y(0{,}2)$ der Lösung des Anfangswertproblems $y' = -100\,y$ mit $y(0) = 1$ für verschiedene Schrittweiten h.

Beispiel 7.12. (implizites Euler-Verfahren)
Wir betrachten das Anfangswertproblem

$$y'(t) = -100\,y(t) \qquad \text{mit} \qquad y(0) = 1,$$

dessen eindeutige Lösung durch $y(t) = e^{-100t}$ gegeben ist. Hier ist $f(t; y) = -100\,y$. Das implizite Euler-Verfahren mit der Schrittweite $h \neq 0$ hat die Iterationsformel

$$t_{k+1} = t_k + h, \qquad y_{k+1} = y_k + h\,f(t_{k+1}, y_{k+1}) = y_k - 100\,h\,y_{k+1},$$

$k = 0, 1, 2, \ldots$, mit den Startwerten $t_0 = 0$ und $y_0 = 1$. Die implizite Gleichung für y_{k+1} können wir hier direkt nach y_{k+1} auflösen:

$$\begin{aligned} y_{k+1} = y_k - 100\,h\,y_{k+1} \qquad &\Longleftrightarrow \qquad y_{k+1} + 100\,h\,y_{k+1} = y_k \\ \Longleftrightarrow \qquad (1 + 100\,h)\,y_{k+1} = y_k \qquad &\Longleftrightarrow \qquad y_{k+1} = (1 + 100\,h)^{-1}\,y_k. \end{aligned}$$

Also erhalten wird für das implizite Euler-Verfahren die folgende explizite Iterationsformel:

$$t_{k+1} = t_k + h, \qquad y_{k+1} = (1 + 100\,h)^{-1}\,y_k, \qquad k = 0, 1, \ldots,$$

mit den Startwerten $t_0 = 0$ und $y_0 = 1$.

Das implizite Euler-Verfahren wurde jeweils mit den Schrittweiten $h_1 = 0{,}1$, $h_2 = 0{,}05$, $h_3 = 0{,}02$, $h_4 = 0{,}01$ und $h_5 = 0{,}001$ durchgeführt, um $y(0{,}2) = e^{-100 \cdot 0{,}2} \doteq 2{,}061 \cdot 10^{-9}$ zu berechnen. In Tabelle 7.3 sind jeweils die mit dem implizitem Euler-Verfahren sowie zum Vergleich die mit dem expliziten Euler-Verfahren berechneten Näherungwerte für $y(0{,}2)$ (auf eine Gleitkommadarstellung mit 3-stelliger Mantisse gerundet) angegeben.

Wir beobachten, dass das implizite Euler-Verfahren in diesem Beispiel deutlich besser arbeitet. Das explizite Euler-Verfahren liefert erst ab einer Schrittweite von $h = h_5 = 0{,}001$ eine Näherung, die sich überhaupt in einer „ähnlichen Größenordnung“ wie $y(0{,}2)$ bewegt. Einige der Werte für kleinere Schrittweiten sind sogar extrem groß (wie die 81 für $h_1 = 0{,}1$ oder die 256 für $h_2 = 0{,}05$). Beim impliziten Euler-Verfahren erhalten wir für den Wert $h_5 = 0{,}001$ nicht nur eine bessere Näherung als beim expliziten Euler-Verfahren. Auch für die größeren Schrittweiten, erhalten wir (anders als beim expliziten Euler-Verfahren) „kleine“ Näherungswerte. ♠

Das explizite und implizite Euler-Verfahren sind Beispiele für sogenannte Einschrittverfahren.

Definition 7.13. (Einschrittverfahren)
Seien $[a;b] \times \mathbb{R}$ *mit* $-\infty < a < b < \infty$, $t_0 \in [a;b]$, *und* $f : [a;b] \times \mathbb{R} \to \mathbb{R}$ *sei eine stetige Funktion, für die es eine Konstante* $L \geq 0$ *gilt, so dass gilt* $|f(t;v) - f(t;w)| \leq L\,|v - w|$ *für alle* $(t;v), (t;w) \in [a;b] \times \mathbb{R}$. *Ein* ***Einschrittverfahren*** *zur numerischen Lösung das Anfangswertproblems*

$$y' = f(t;y) \qquad \text{mit} \qquad y(t_0) = y_0 \tag{7.21}$$

konstruiert Näherungen y_k *der Funktionswerte* $y(t_k)$ *der eindeutigen Lösung* $y : [a;b] \to \mathbb{R}$ *von* (7.21) *an den* ***äquidistanten Gitterpunkten***

$$t_k = t_0 + k\,h, \qquad k = 0, 1, 2, \ldots,$$

mit der ***Schrittweite*** $h \neq 0$ *mit der* ***Iterationsformel***

$$y_{k+1} := y_k + h\,\Phi(t_k; y_k; y_{k+1}; h), \qquad k = 0, 1, 2, \ldots, \tag{7.22}$$

wobei die Funktion Φ *mit Hilfe der rechten Seite* $f : [a;b] \times \mathbb{R} \to \mathbb{R}$ *der Differentialgleichung in* (7.21) *definiert ist. Die Funktion* Φ *heißt die* ***Inkrementfunktion*** *des Einschrittverfahrens. Falls* $\Phi(t_k; y_k; y_{k+1}; h)$ *nicht von dem dritten Argument* y_{k+1} *abhängt, nennt man das Einschrittverfahren* (7.22) ***explizit***. *Andernfalls nennt man das Einschrittverfahren* (7.22) ***implizit***.

Beispiel 7.14. (Einschrittverfahren)
Seien $[a;b] \times \mathbb{R}$ mit $-\infty < a < b < \infty$, $t_0 \in [a;b]$, und $f : [a;b] \times \mathbb{R} \to \mathbb{R}$ sei eine stetige Funktion, für die es eine Konstante $L \geq 0$ gilt, so dass die Lipschitz-Bedingung $|f(t;v) - f(t;w)| \leq L\,|v - w|$ für alle $(t;v), (t;w) \in [a;b] \times \mathbb{R}$ gilt.

Für $(t_0; y_0) \in [a; b] \times \mathbb{R}$ betrachten wir das Anfangswertproblem

$$y' = f(t; y) \qquad \text{mit} \qquad y(t_0) = y_0. \tag{7.23}$$

(a) Das **explizite Euler-Verfahren** mit Schrittweite $h \neq 0$ zur Berechnung der Näherungen y_k der Werte $y(t_k)$ der eindeutigen Lösung $y : [a; b] \to \mathbb{R}$ von (7.23) an den äquidistanten Gitterpunkten $t_k = t_0 + k\,h$, $k = 0, 1, 2, \ldots$, hat (nach Verfahren 7.8) die Iterationsformel

$$y_{k+1} = y_k + h\, f(t_k; y_k).$$

Hier gilt also $\Phi(t_k; y_k; y_{k+1}; h) = f(t_k; y_k)$. Es handelt sich also um ein explizites Einschrittverfahren.

(b) Das **implizite Euler-Verfahren** mit Schrittweite $h \neq 0$ zur Berechnung der Näherungen y_k der Werte $y(t_k)$ der eindeutigen Lösung $y : [a; b] \to \mathbb{R}$ von (7.23) an den äquidistanten Gitterpunkten $t_k = t_0 + k\,h$, $k = 0, 1, 2, \ldots$, hat (nach Verfahren 7.11) die Iterationsformel

$$y_{k+1} = y_k + h\, f(t_{k+1}; y_{k+1}).$$

Hier gilt also $\Phi(t_k; y_k; y_{k+1}; h) = f(t_k + h; y_{k+1})$. Es handelt sich also um ein implizites Einschrittverfahren.

Wir lernen noch weitere Beispiele kennen. ♠

Beispiel 7.15. (Mittelpunktverfahren)
Seien $[a; b] \times \mathbb{R}$ mit $-\infty < a < b < \infty$, $t_0 \in [a; b]$, und $f : [a; b] \times \mathbb{R} \to \mathbb{R}$ sei eine stetige Funktion, für die es eine Konstante $L \geq 0$ gilt, so dass die Lipschitz-Bedingung $|f(t; v) - f(t; w)| \leq L\,|v - w|$ für alle $(t; v), (t; w) \in [a; b] \times \mathbb{R}$ gilt. Für $(t_0; y_0) \in [a; b] \times \mathbb{R}$ betrachten wir das Anfangswertproblem

$$y'(t) = f(t; y) \qquad \text{mit} \qquad y(t_0) = y_0. \tag{7.24}$$

Das **Mittelpunktverfahren** mit Schrittweite $h \neq 0$ zur Berechnung der Näherungen y_k der Werte $y(t_k)$ der eindeutigen Lösung $y : [a; b] \to \mathbb{R}$ von (7.24) an den äquidistanten Gitterpunkten $t_k = t_0 + k\,h$, $k = 0, 1, 2, \ldots$, ist durch die folgende Iterationsformel definiert:

$$\boxed{y_{k+1} := y_k + h\, f\left(t_k + \frac{h}{2}; \frac{1}{2}\,(y_k + y_{k+1})\right).} \tag{7.25}$$

Hier gilt also $\Phi(t_t; y_k; y_{k+1}; h) = f\left(t_k + \frac{h}{2}; \frac{1}{2}\,(y_k + y_{k+1})\right)$. Es handelt sich also um ein implizites Einschrittverfahren.

h	k	t_k	y_k	$\lvert y_k - y(t_k)\rvert$	$\lvert y_k - y(t_k)\rvert / \lvert y(t_k)\rvert$
$h_1 = 0{,}2$	5	1,0	$3{,}6665 \cdot 10^{-1}$	$1{,}23 \cdot 10^{-3}$	0,00335
	10	2,0	$1{,}3443 \cdot 10^{-1}$	$9{,}05 \cdot 10^{-4}$	0,00668
	15	3,0	$4{,}9289 \cdot 10^{-2}$	$4{,}98 \cdot 10^{-4}$	0,0100
	20	4,0	$1{,}8072 \cdot 10^{-2}$	$2,44 \cdot 10^{-4}$	0,0133
	25	5,0	$6{,}6259 \cdot 10^{-3}$	$1{,}12 \cdot 10^{-4}$	0,0166
$h_2 = 0{,}1$	10	1,0	$3{,}6757 \cdot 10^{-1}$	$3{,}07 \cdot 10^{-4}$	0,000834
	20	2,0	$1{,}3511 \cdot 10^{-1}$	$2{,}26 \cdot 10^{-4}$	0,00167
	30	3,0	$4{,}9663 \cdot 10^{-2}$	$1{,}24 \cdot 10^{-4}$	0,00250
	40	4,0	$1{,}8255 \cdot 10^{-2}$	$6{,}10 \cdot 10^{-5}$	0,00333
	50	5,0	$6{,}7099 \cdot 10^{-3}$	$2,81 \cdot 10^{-5}$	0,00416
$h_3 = 0{,}05$	20	1,00	$3{,}6780 \cdot 10^{-1}$	$7{,}67 \cdot 10^{-5}$	0,000208
	40	2,00	$1{,}3528 \cdot 10^{-1}$	$5,64 \cdot 10^{-5}$	0,000417
	60	3,00	$4{,}9756 \cdot 10^{-2}$	$3{,}11 \cdot 10^{-5}$	0,000625
	80	4,00	$1{,}8300 \cdot 10^{-2}$	$1{,}53 \cdot 10^{-5}$	0,000833
	100	5,00	$6{,}7309 \cdot 10^{-3}$	$7{,}02 \cdot 10^{-6}$	0,00104

Tabelle 7.4: Mittelpunktverfahren für $y' = -y$ mit $y(0) = 1$.

Wenden wir das Mittelpunktverfahren für das Anfangswertproblem

$$y'(t) = -y(t) \qquad \text{mit} \qquad y(0) = 1,$$

aus Beispiel 7.9 an, dessen Lösung durch $y(t) = e^{-t}$ gegeben ist. Die Formel (7.25) des Mittelpunktverfahrens lautet dann für das konkrete Beispiel mit $f(t; y) = -y$:

$$y_{k+1} = y_k - h\,\frac{1}{2}\,(y_k + y_{k+1}) \qquad \Longleftrightarrow \qquad y_{k+1} + \frac{h}{2}\,y_{k+1} = y_k - \frac{h}{2}\,y_k$$

$$\Longleftrightarrow \qquad \left(1 + \frac{h}{2}\right) y_{k+1} = \left(1 - \frac{h}{2}\right) y_k \qquad \Longleftrightarrow \qquad \boxed{y_{k+1} = \frac{1 - \frac{h}{2}}{1 + \frac{h}{2}}\,y_k}$$

Wir nutzen das Mittelpunktverfahren jeweils mit einer der Schrittweiten $h_1 = 0{,}2$, $h_2 = 0{,}1$ und $h_3 = 0{,}05$ zur Berechnung von Näherungwerten für $y(t)$ mit $t \in \{1; 2; 3; 4; 5\}$. In Tabelle 7.4 sind jeweils die berechneten Näherungwerte für $y(t)$ mit $t \in \{1; 2; 3; 4; 5\}$ (gerundet auf eine Gleitkommadarstellung mit 5-stelliger Mantisse) sowie deren absolute und relative Fehler angegeben. Ein Ver-

gleich der Tabelle 7.4 mit der Tabelle 7.1 zeigt, dass das Mittelpunktverfahren bei der gleichen Schrittweite deutlich bessere Näherungen liefert. ♠

7.4 Konsistenz und Konvergenz von Einschrittverfahren

Was bedeutet **Konvergenz für ein Einschrittverfahren**? Die Näherungswerte y_k sollten gegen die Funktionswerte $y(t_k)$ der Lösung der Anfangswertproblems streben, wenn die Schrittweite h gegen 0 strebt. Bevor wir dieses präzise definieren können, brauchen wir noch das Konzept der Konsistenz und der Ordnung eines Einschrittverfahrens. Wir beschränken uns in diesem Kapitel der Einfachheit halber auf **explizite** Einschrittverfahren.

Um die Voraussetzungen an die Funktion f in der gewöhnlichen Differentialgleichung $y' = f(t; y)$ bequem formulieren zu können, führen wir den folgenden **Funktionenraum** ein: Für $p \in \mathbb{N}_0$ ist $\mathcal{F}_p\big([c; d] \times \mathbb{R}\big)$ die Menge aller Funktionen auf $[c; d] \times \mathbb{R}$, deren partielle Ableitungen bis zu und einschließlich der Ordnung p existieren und stetig und auf $[c; d] \times \mathbb{R}$ beschränkt sind. (Eine Funktion $g : D \to \mathbb{R}$ heißt „auf D beschränkt“, wenn es eine Konstante $S > 0$ gibt, so dass $|g(\mathbf{x})| \leq S$ für alle $\mathbf{x} \in D$ gilt.) Ist $f \in \mathcal{F}_1\big([c; d] \times \mathbb{R}\big)$ so folgt (nach den Überlegungen auf Seiten 246 bis 247), dass es eine Konstante $L \geq 0$ mit $|f(s; v) - f(s; w)| \leq L\,|v - w|$ für alle $(s; v), (s; w) \in [c; d] \times \mathbb{R}$.

Um die Aussagen über die Konsistenz und Konvergenz von Einschrittverfahren technisch bequem (d.h. so dass alles immer passend definiert ist) formulieren zu können, betrachten wir die Differentialgleichung $y' = f(t; y)$ für Funktionen f definiert auf $I_\delta \times \mathbb{R}$ mit einem leicht größeren Intervall I_δ als $I = [a; b]$, nämlich auf $I_\delta = [a - \delta; b + \delta]$ mit einem $\delta > 0$. Um längliche Notation zu vermeiden, schreiben wir im Folgenden

$$\begin{aligned} I &:= [a; b] \qquad \text{mit} \qquad -\infty < a < b < \infty \qquad \text{und} \\ I_\delta &:= [a - \delta; b + \delta] \qquad \text{mit einem (kleinen) } \delta > 0. \end{aligned}$$

Auf der Zahlengeraden ist das Intervall $I_\delta = [a-\delta; b+\delta]$ gegenüber dem Intervall $I = [a; b]$ um die Entfernung $\delta > 0$ nach links und nach rechts verlängert worden.

In diesem Teilkapitel werden wir $f : I_\delta \times \mathbb{R}$ mit einer Lipschitz-Bedingung in der zweiten Variablen betrachten, so dass die Lösung des Anfangswertproblems $y' = f(t; y)$ mit $y(t_0) = y_0$ für jedes $(t_0; y_0) \in I_\delta \times \mathbb{R}$ existiert und auf ganz I_δ eindeutig definiert ist. Als Anfangswerte werden wir aber nur $(t_0; y_0) \in I \times \mathbb{R}$

betrachten. Dadurch ist sichergestellt, dass $t_0 + h$ für alle $|h| \leq \delta$ in I_δ liegt und dass $f(t+h; y)$ für alle $(t; y) \in I \times \mathbb{R}$ und alle $|h| \leq \delta$ immer definiert ist. Dadurch vermeiden wir komplizierte Einschränkungen an die Schrittweiten h.

Definition 7.16. (lokaler Diskretisierungsfehler und Konsistenz)
Seien $I = [a; b]$ und $I_\delta = [a-\delta; b+\delta]$ mit $\delta > 0$. Sei $f : I_\delta \times \mathbb{R} \to \mathbb{R}$ eine stetige Funktion, für die es eine Konstante $L \geq 0$ gilt, so dass gilt $|f(s; v) - f(s; w)| \leq L\,|v - w|$ für alle $(s; v), (s; w) \in I_\delta \times \mathbb{R}$. Für $(t; y) \in I \times \mathbb{R}$ sei $z : I_\delta \to \mathbb{R}$ die eindeutige Lösung des Anfangswertproblems

$$z'(s) = f\big(s; z(s)\big) \qquad \text{mit} \qquad z(t) = y. \tag{7.26}$$

*Weiter sei ein **explizites Einschrittverfahren** zur Berechnung von Näherungen z_k der Werte $z(s_k)$ der eindeutigen Lösung $z : I_\delta \to \mathbb{R}$ von (7.26) gegeben:*

$$\left.\begin{aligned} z_{k+1} &:= z_k + h\,\Phi(s_k; z_k; h), \\ s_{k+1} &:= t + (k+1)\,h = s_k + h, \end{aligned}\right\} \quad \text{für } k = 0, 1, \ldots, \text{ mit } s_0 = t,\ z_0 = y. \tag{7.27}$$

(1) Die Funktion

$$\Delta(t; y; h) = \Delta\big(t; z(t); h\big) := \begin{cases} \dfrac{z(t+h) - z(t)}{h} & \text{für } |h| \leq \delta \text{ mit } h \neq 0, \\ f(t; y) & \text{für } h = 0, \end{cases}$$

*ist der **(Vorwärts-)Differenzenquotient der eindeutigen Lösung** $z : I_\delta \to \mathbb{R}$ von (7.26) im Punkt $t \in I$ mit der Schrittweite h. Es gilt $\lim\limits_{h \to 0} \Delta(t; y; h) = f(t; y)$.*

(2) Aus der Iterationsvorschrift des expliziten Einschrittverfahrens (7.27) folgt für $k = 0$ mit $s_0 = t$ und $z_0 = y$

$$\Phi(s_0; z_0; h) = \frac{z_1 - z_0}{h} \qquad \Longleftrightarrow \qquad \Phi(t; y; h) = \frac{z_1 - z_0}{h}\,,$$

*d.h. $\Phi(s_0; z_0; h) = \Phi(t; y; h)$ ist der **(Vorwärts-)Differenzenquotient der vom expliziten Einschrittverfahren** (7.27) im ersten Iterationsschritt **berechneten Näherungslösung** von (7.26) im Punkt $t \in I$ mit der Schrittweite h.*

*(3) Der **lokale Diskretisierungsfehler** des expliziten Einschrittverfahrens (7.27) **im Punkt** $(t; y) \in I \times \mathbb{R}$ ist definiert als*

$$\tau(t; y; h) := \Delta(t; y; h) - \Phi(t; y; h), \qquad \text{wobei } |h| \leq \delta.$$

(4) Das explizite Einschrittverfahren (7.27) heißt ***konsistent****, wenn für alle* $(t;y) \in I \times \mathbb{R}$ *und für alle Funktionen* $f \in \mathcal{F}_1(I_\delta \times \mathbb{R})$ *gilt*

$$\lim_{h\to 0} \tau(t;y;h) = 0 \qquad \Longleftrightarrow \qquad \lim_{h\to 0} \Phi(t;y;h) = f(t;y),$$

wobei wir $\lim_{h\to 0} \Delta(t;y;h) = f(t;y)$ *ausgenutzt haben.*

(5) Das explizite Einschrittverfahren (7.27) ist ein ***Verfahren der Ordnung*** $p \in \mathbb{N}$*, wenn für jede Funktion* $f \in \mathcal{F}_p(I_\delta \times \mathbb{R})$ *gilt* $\tau(t;y;h) = \mathcal{O}(h^p)$ ***gleichmäßig für alle*** $(t;y) \in I \times \mathbb{R}$*, d.h. wenn für jede Funktion* $f \in \mathcal{F}_p(I_\delta \times \mathbb{R})$ *eine Konstante* K *existiert, so dass gilt*

$$|\tau(t;y;h)| \le K\,|h|^p \qquad \text{für alle } (t;y) \in I \times \mathbb{R} \text{ und alle } |h| \le \delta.$$

Natürlich ist ein Verfahren der Ordnung $p \in \mathbb{N}$ insbesondere konsistent.

Nun können wir zeigen, dass das explizite Euler-Verfahren konsistent ist.

Satz 7.17. (Konsistenz des expliziten Euler-Verfahrens)
Seien $I = [a;b]$ *und* $I_\delta = [a-\delta; b+\delta]$ *mit* $\delta > 0$*. Seien* $(t;y) \in I \times \mathbb{R}$ *und* $f \in \mathcal{F}_1(I_\delta \times \mathbb{R})$*. Das* ***explizite Euler-Verfahren*** *zur numerischen Lösung des Anfangswertproblems*

$$z'(s) = f\big(s; z(s)\big) \qquad \text{mit} \qquad z(t) = y \tag{7.28}$$

ist ***konsistent*** *und ist ein* ***Verfahren der Ordnung*** $p = 1$*.*

Beweis von Satz 7.17: Das explizite Euler-Verfahren hat die Inkrementfunktion $\Phi(t;y;h) = f(t;y)$. Damit gilt für alle $(t;y) \in I \times \mathbb{R}$ und für alle $f \in \mathcal{F}_1(I_\delta \times \mathbb{R})$

$$\lim_{h\to 0} \Phi(t;y;h) = \lim_{h\to 0} f(t;y) = f(t;y).$$

Also ist das explizite Euler-Verfahren konsistent.

Für den lokalen Diskretisierungsfehler gilt mit $\Phi(t;y;h) = f(t;y)$ für $0 < |h| \le \delta$

$$\begin{aligned}\tau(t;y;h) &= \Delta(t;y;h) - \Phi(t;y;h) = \frac{z(t+h)-z(t)}{h} - f(t;y) \\ &= \frac{1}{h}\Big(z(t+h) - \big[z(t) + h\, f\big(t; z(t)\big)\big]\Big) = \frac{1}{h}\Big(z(t+h) - \big[z(t) + h\, z'(t)\big]\Big),\end{aligned} \tag{7.29}$$

wobei wir in der zweiten Zeile $f\big(t; z(t)\big) = z'(t)$ genutzt haben. Für $f \in \mathcal{F}_1(I_\delta \times \mathbb{R})$ ist $z'(s) = f\big(s; z(s)\big)$ stetig differenzierbar, d.h. die Lösung z des Anfangswertproblems (7.28) ist zweimal stetig differenzierbar. Entwickeln wir $z(t+h)$ mit dem dem Satz von Taylor in ein Taylorpolynom vom Grad 1 mit dem Entwicklungspunkt t, so erhalten wir für alle h mit $|h| \leq \delta$

$$z(t+h) = z(t) + h\, z'(t) + \frac{h^2}{2}\, z''(t + \eta\, h)$$

mit einem $\eta \in]0;1[$. Einsetzen in (7.29) liefert

$$\begin{aligned}
\tau(t;y;h) &= \frac{1}{h}\Big(z(t+h) - \big[z(t) + h\, z'(t)\big]\Big) \\
&= \frac{1}{h}\left(z(t) + h\, z'(t) + \frac{h^2}{2}\, z''(t+\eta\, h) - \big[z(t) + h\, z'(t)\big]\right) \\
&= \frac{h}{2}\, z''(t + \eta\, h). \qquad (7.30)
\end{aligned}$$

Wegen wegen $z'(s) = f\big(s; z(s)\big)$ folgt mit der Kettenregel

$$z''(s) = \frac{\partial f\big(s; z(s)\big)}{\partial s} + \frac{\partial f\big(s; z(s)\big)}{\partial z}\, z'(s) = \frac{\partial f\big(s; z(s)\big)}{\partial s} + \frac{\partial f\big(s; z(s)\big)}{\partial z}\, f\big(s; z(s)\big).$$

Daraus folgt, dass für $f \in \mathcal{F}_1(I_\delta \times \mathbb{R})$ eine Konstante $S \geq 0$ existiert mit

$$|z''(s)| \leq \left|\frac{\partial f\big(s; z(s)\big)}{\partial s}\right| + \left|\frac{\partial f\big(s; z(s)\big)}{\partial z}\right| \big|f\big(s; z(s)\big)\big| \leq S \qquad \text{für alle } s \in I, \qquad (7.31)$$

denn f und seine partiellen Ableitungen $\frac{\partial f}{\partial s}$ und $\frac{\partial f}{\partial z}$ sind auf $I_\delta \times \mathbb{R}$ beschränkt (weil $f \in \mathcal{F}_1(I_\delta \times \mathbb{R})$ ist). Anwenden von (7.31) in (7.30) liefert

$$|\tau(t;y;h)| = \frac{|h|}{2}\, \underbrace{|z''(t+\eta\, h)|}_{\leq S} \leq \frac{S}{2}\, |h| \qquad \text{für alle } (t;y) \in I \times \mathbb{R} \text{ und alle } |h| \leq \delta.$$

Also ist das explizite Euler-Verfahren ein Verfahren der Ordnung $p = 1$. □

Als letztes neues Konzept benötigen wir noch den Begriff der Konvergenz von Einschrittverfahren.

Definition 7.18. (Konvergenz von Einschrittverfahren)
Seien $I = [a; b]$, $I_\delta = [a - \delta; b + \delta]$ mit $\delta > 0$ und $f \in \mathcal{F}_1(I_\delta \times \mathbb{R})$. Für $(t_0; y_0) \in I \times \mathbb{R}$ sei $y : I_\delta \to \mathbb{R}$ die eindeutige Lösung des Anfangswertproblems

$$y'(t) = f(t; y) \qquad \text{mit} \qquad y(t_0) = y_0. \tag{7.32}$$

*Weiter sei ein **explizites Einschrittverfahren** zur Berechnung von Näherungen y_k der Werte $y(t_k)$ der eindeutigen Lösung $y = y(t)$ von (7.32) gegeben:*

$$\left.\begin{aligned} y_{k+1} &:= y_k + h\,\Phi(t_k; y_k; h), \\ t_{k+1} &:= t_0 + (k+1)\,h = t_k + h, \end{aligned}\right\} \quad \text{für } k = 0, 1, \ldots, \text{ mit } y_0 = y(t_0). \tag{7.33}$$

(1) Für $t \in I$ erhalten wir für jedes $n \in \mathbb{N}$ mit den Schrittweite $h_n := \frac{t - t_0}{n}$ (und somit $t = t_0 + n\,h_n$) durch die Iterierte y_n eine Näherung für

$$y(t_n) = y(t_0 + n\,h_n) = y\left(t_0 + n \cdot \frac{t - t_0}{n}\right) = y\big(t_0 + (t - t_0)\big) = y(t).$$

*Daher ist für $t \in I$ der **globale Diskretisierungsfehler** des Einschrittverfahrens (7.33) definiert als*

$$e(t; h_n) = y_n - y(t) \quad \text{mit} \quad \left(h_n := \frac{t - t_0}{n} \iff t = t_n = t_0 + n\,h_n\right).$$

*(2) Das durch (7.33) definierte Einschrittverfahren heißt **konvergent**, wenn für alle $f \in \mathcal{F}_1(I_\delta \times \mathbb{R})$ gilt*

$$\lim_{n \to \infty} e(t; h_n) = 0 \qquad \text{für alle } t \in I \text{ und alle } (t_0; y_0) \in I \times \mathbb{R}.$$

Der nachfolgende Satz liefert uns wichtige Informationen über den Zusammenhang zwischen Konsistenz und Konvergenz. Er sieht wegen der technischen Voraussetzungen kompliziert aus, aber der zentrale Punkt ist, dass (unter geeigneten Voraussetzungen) **ein konsistentes Einschrittverfahren konvergent ist** und dass für ein **Einschrittverfahren der Ordnung $p > 0$ der globale Diskretisierungsfehler von der Ordnung p ist**, also $e(t; h_n) = \mathcal{O}(h_n^p)$ für alle $t \in I$ und alle $|h_n| \leq \delta$.

Satz 7.19. (Konvergenz von Einschrittverfahren)
Seien $I = [a;b]$, $I_\delta = [a-\delta; b+\delta]$ mit $\delta > 0$ und $f \in \mathcal{F}_1(I_\delta \times \mathbb{R})$. Für $(t_0; y_0) \in I \times \mathbb{R}$ sei $y : I_\delta \to \mathbb{R}$ die eindeutige Lösung des Anfangswertproblems

$$y'(t) = f(t;y) \qquad mit \qquad y(t_0) = y_0. \tag{7.34}$$

*Weiter sei ein **explizites Einschrittverfahren** zur Berechnung von Näherungen y_k der Werte $y(t_k)$ der eindeutigen Lösung $y = y(t)$ von (7.34) gegeben:*

$$\left.\begin{aligned} y_{k+1} &:= y_k + h\,\Phi(t_k; y_k; h), \\ t_{k+1} &:= t_0 + (k+1)\,h = t_k + h, \end{aligned}\right\} \quad \text{für } k = 0, 1, \ldots, \text{ mit } y_0 = y(t_0). \tag{7.35}$$

Die Inkrementfunktion Φ des Einschrittverfahrens (7.35) sei stetig auf

$$G = \big\{(t;y;h) \;:\; t \in I \text{ und } |y - y(t)| \le \gamma \text{ und } 0 \le |h| \le h_0\big\} \tag{7.36}$$

mit den Konstanten $0 < h_0 \le \delta$ und $\gamma > 0$. Weiter nehmen wir an, dass positive Konstanten M und N existieren, so dass gelten

$$|\Phi(t;v;h) - \Phi(t;w;h)| \le M\,|v - w| \quad \text{für alle } (t;v;h), (t;w;h) \in G \tag{7.37}$$

$$\text{und} \qquad \big|\tau\big(t; y(t); h\big)\big| = \big|\Delta\big(t; y(t); h\big) - \Phi\big(t; y(t); h\big)\big| \le N\,|h|^p \quad \text{für alle } t \in I \text{ und alle } |h| \le h_0. \tag{7.38}$$

*Dann existiert eine Konstante $\overline{h}$ mit $0 < \overline{h} \le h_0$, so dass für den **globalen Diskretisierungsfehler** die folgende Fehlerabschätzung gilt:*

$$|e(t;h_n)| \le |h_n|^p\, N\, \frac{e^{M|t-t_0|-1}}{M} \quad \text{für alle } t \in I \text{ und alle } h_n = \frac{t-t_0}{n},\ n = 1, 2, \ldots, \text{ mit } |h_n| \le \overline{h}. \tag{7.39}$$

*Man sagt dann das Einschrittverfahren habe die **Konvergenzordnung** p. Gilt in (7.36) $\gamma = \infty$, dann ist $\overline{h} = h_0$.*

Betrachten wir Satz 7.19 genauer. Wir beginnen mit der Abschätzung (7.39) für den globalen Diskretisierungsfehler: Aus (7.39) folgt

$$|e(t;h_n)| \le C\,|h_n|^p \quad \text{für alle } t \in I,\ h_n = \frac{t-t_0}{n},\ n \in \mathbb{N},\ \text{mit } |h_n| \le \overline{h}$$

mit der Konstante $C := \frac{N}{M} \max\big\{e^{M|a-t_0|-1}; e^{M|b-t_0|-1}\big\}$. Dieses besagt, dass der

globale Diskretisierungsfehler für $n \to \infty$ (und damit für $h_n \to 0$) wie $O(|h_n|^p)$ abnimmt. (Wir beobachten noch, dass der Faktor, mit dem $|h_n|^p$ in (7.39) multipliziert wird, von t abhängt und desto größer wird je weiter t von t_0 entfernt ist.) Die Voraussetzungen für dieses Resultat sind eine Lipschitz-Bedingung (7.37) für die zweite Variable der Inkrementfunktion Φ und dass der lokale Diskretisierungsfehler gemäß (7.38) von der Ordnung $\mathcal{O}(|h|^p)$ ist, also dass das Einschrittverfahren von der Ordnung p ist. Durch (7.36) ist eine Menge vorgegeben, auf der die Lipschitz-Bedingung an Φ gelten soll. Ist $\gamma = \infty$, so ist bei G die Bedingung an y automatisch erfüllt und wir erhalten

$$G = \big\{(t; y; h) \ : \ t \in I \text{ und } 0 \leq |h| \leq h_0\big\} = I \times \mathbb{R} \times [-h_0; h_0],$$

d.h. die einzige Einschränkung ist, dass die Schrittweiten h des expliziten Einschrittverfahrens $|h| \leq h_0$ erfüllen sollen. Da aber nur der Fall kleiner h interessant ist, ist dieses in der Praxis kein Problem. – Wenn $0 < \gamma < \infty$ ist, so besagt $|y - y(t)| \leq \gamma$, dass nur Werte für y in einem „Schlauch“ um mit „Radius“ γ um die eindeutige Lösung $y = y(t)$ betrachtet werden. Falls das explizite Einschrittverfahren konvergent ist, so werden die Näherungen y_k, $k \in \mathbb{N}_0$, für klein genuges h aber auch in diesem „Schlauch“ liegen.

Als Folgerung aus Satz 7.19 und Satz 7.17 erhalten wir die Konvergenz des expliziten Euler-Verfahrens.

Satz 7.20. (Konvergenz des expliziten Euler-Verfahrens)
*Seien $I = [a; b]$ und $I_\delta = [a - \delta; b + \delta]$ mit $\delta > 0$. Seien $(t_0; y_0) \in I \times \mathbb{R}$ und $f \in \mathcal{F}_1(I_\delta \times \mathbb{R})$. Das **explizite Euler-Verfahren** zur numerischen Lösung das Anfangswertproblems $y'(t) = f\big(t; y(t)\big)$ mit $y(t_0) = y_0$ ist **konvergent** mit der **Konvergenzordnung** $p = 1$.*

Beweis von Satz 7.20: Das explizite Euler-Verfahren hat die Inkrementfunktion $\Phi(t; y; h) = f(t; y)$. Nach Satz 7.17 ist das explizite Euler-Verfahren konsistent und ein Verfahren der Ordnung $p = 1$. Wir wählen $G = I \times \mathbb{R} \times [-\delta; \delta]$ mit $\gamma = \infty$ und $h_0 = \delta$. Für $f \in \mathcal{F}_1\big(I_\delta \times \mathbb{R}\big)$ gibt es eine Konstante $L \geq 0$, so dass gilt

$$|f(t; v) - f(t; w)| \leq L\,|v - w| \qquad \text{für alle } (t; v), (t; w) \in I_\delta \times \mathbb{R}.$$

Daraus folgt für die Inkrementfunktion $\Phi(t; y; h) = f(t; y)$ direkt

$$|\Phi(t; v; h) - \Phi(t; w; h)| \leq |f(t; v) - f(t; w)| \leq L\,|v - w|$$
$$\text{für alle } (t; v; h), (t; w; h) \in I \times \mathbb{R} \times [-\delta; \delta] = G.$$

Nach Satz 7.17 und dem zugehörigen Beweis gilt mit $y = y(t)$

$$\big|\tau\big(t; y(t); h\big)\big| \leq N\,|h| \quad \text{für alle } t \in I \text{ und alle } h \text{ mit } |h| \leq \delta.$$

Aus Satz 7.19 folgt dann, dass das explizite Euler-Verfahren konvergent ist und die Konvergenzordnung $p = 1$ hat. □

7.5 Explizite Runge-Kutta-Verfahren

Wir haben gesehen, dass das explizite Euler-Verfahren konvergent ist und die Konvergenzordnung $p = 1$ hat. **Je höher die Konvergenzordnung eines Verfahrens ist, desto schneller konvergiert es.** Daher stellt sich die **Frage, wie man Einschrittverfahren mit einer höheren Konvergenzordnung herleiten kann?**

Eine häufig eingesetzte Familie von Verfahren mit einer höheren Konvergenzordnung sind die sogenannten **Runge-Kutta-Verfahren**, die wir nun einführen.

Im Folgenden seien $I = [a; b]$, $I_\delta = [a - \delta; b + \delta]$ mit $\delta > 0$ und $f \in \mathcal{F}_2(I_\delta \times \mathbb{R})$. Für $(t; y) \in I \times \mathbb{R}$ sei $z : I_\delta \to \mathbb{R}$ die eindeutige Lösung des Anfangswertproblems

$$z'(s) = f\big(s; z(s)\big) \qquad \text{mit} \qquad z(t) = y. \tag{7.40}$$

Wir konstruieren nun **explizite Runge-Kutta-Verfahren mit der Konvergenzordnung** $p = 2$. Dieses sind explizite Einschrittverfahren von der Form

$$\left.\begin{aligned} y_{k+1} &:= y_k + h\,\Phi(t_k; y_k; h), \\ t_{k+1} &:= t + k\,h = t_k + h, \end{aligned}\right\} \quad \text{für } k = 0, 1, \ldots, \quad \text{mit } y_0 = y(t_0). \tag{7.41}$$

mit einer Inkrementfunktion der Form

$$\boxed{\Phi(t; y; h) = \gamma_1\, f(t; y) + \gamma_2\, f\big(t + \alpha\, h; y + \beta\, h\, f(t; y)\big),} \tag{7.42}$$

wobei die Konstanten $\alpha, \beta, \gamma_1, \gamma_2$ passend gewählt werden müssen, damit das Verfahren auch tatsächlich die Konvergenzordnung 2 bekommt. Dabei muss man sich die durch (7.42) definierte Inkrementfunktion $\Phi(t_k; y_k; h)$ als eine „durchschnittliche Steigung“ der Lösung y auf dem Intervall $[t_k; t_{k+1}]$ vorstellen.

Um mit Hilfe von Satz 7.19 die Konvergenzordnung eines Verfahrens der Form (7.41) mit der Inkrementfunktion (7.42) folgern zum können, betrachten wir den lokalen Diskretisierungsfehler $\tau(t; y; h)$ (vgl. Definition 7.16) und wollen die Konstanten $\alpha, \beta, \gamma_1, \gamma_2$ in (7.42) so wählen, dass dieser die Ordnung $p = 2$ hat: Für $t \in I$ und $|h| \le \delta$ ist der lokale Diskretisierungsfehler (mit $y = z(t)$)

$$\tau(t; y; h) = \Delta(t; y; h) - \Phi(t; y; h) = \frac{z(t + h) - z(t)}{h} - \Phi(t; y; h)$$

$$= \frac{1}{h}\big[z(t+h) - z(t)\big] - \gamma_1 f(t;y) - \gamma_2 f\big(t + \alpha h; y + \beta h f(t;y)\big), \qquad (7.43)$$

wobei $z : I_\delta \to \mathbb{R}$ die eindeutige Lösung des Anfangswertproblems (7.40) ist. Nun entwicklen wir $z(t+h)$ mit $|h| \leq \delta$ mit dem Satz von Taylor in ein Taylorpolynom vom Grad 2 mit dem Entwicklungspunkt t:

$$\begin{aligned} & z(t+h) = z(t) + z'(t)\, h + \frac{1}{2} z''(t)\, h^2 + \mathcal{O}(|h|^3) \\ \iff \qquad & z(t+h) - z(t) = z'(t)\, h + \frac{1}{2} z''(t)\, h^2 + \mathcal{O}(|h|^3) \end{aligned} \qquad (7.44)$$

(Dabei kann die durch den $\mathcal{O}(|h|^3)$ Term implizierte Konstante $K > 0$ in

$$\left| z(t+h) - \left(z(t) + z'(t)\, h + \frac{z''(t)}{2} h^2 \right) \right| \leq K\, |h|^3 \quad \text{für alle } (t;y) \in I \times \mathbb{R},\ |h| \leq \delta$$

einheitlich für alle $(t;y;h) \in I \times \mathbb{R} \times [a;b] \times [-\delta;\delta]$ gewählt werden, denn wegen $z'(s) = f\big(s; z(s)\big)$ ist $z : I_\delta \to \mathbb{R}$ dreimal stetig differenzierbar und seine Ableitungen bis einschließlich der Ordnung 3 lassen sich durch die partiellen Ableitungen bis einschließlich der Ordnung 2 von $f \in \mathcal{F}_2(I_\delta \times \mathbb{R})$ ausdrücken und diese sind auf $I_\delta \times \mathbb{R}$ beschränkt.) Wir haben $z'(s) = f\big(s; z(s)\big)$ und damit mit der Kettenregel

$$z''(s) = \frac{\partial f}{\partial s}\big(s; z(s)\big) + \frac{\partial f}{\partial z}\big(s; z(s)\big) \cdot z'(s) = \frac{\partial f}{\partial s}\big(s; z(s)\big) + \frac{\partial f}{\partial z}\big(s; z(s)\big) \cdot f\big(s; z(s)\big).$$

Einsetzen von $z'(t) = f\big(t; z(t)\big) = f(t;y)$ und der gerade berechneten Formel für $z''(s)$ mit $s = t$ (als Funktion von f und seinen partiellen Ableitungen) in (7.44) und Ersetzen von $z(t) = y$ (aus der Anfangsbedingung) liefert

$$z(t+h) - z(t) = h\, f(t;y) + \frac{1}{2} h^2 \left(\frac{\partial f(t;y)}{\partial s} + \frac{\partial f(t;y)}{\partial z}\, f(t;y) \right) + \mathcal{O}(|h|^3). \qquad (7.45)$$

Damit haben wir eine passende Darstellung für den ersten Term in (7.43) gefunden. – Wir benötigen noch eine passende Darstellung für den zweiten Term in (7.43): Dazu nutzen wir den Satz von Taylor, um $f\big(t + \alpha h; y + \beta h f(t;y)\big)$ in ein Taylorpolynom vom Grad 1 mit dem Entwicklungspunkt $(t;y)$ zu entwickeln:

$$\begin{aligned} & f\big(t + \alpha h; y + \beta h f(t;y)\big) \\ & \qquad = f(t;y) + \frac{\partial f(t;y)}{\partial s}\, \alpha h + \frac{\partial f(t;y)}{\partial z}\, \beta h f(t;y) + \mathcal{O}(|h|^2), \end{aligned} \qquad (7.46)$$

wobei die Konstante in $\mathcal{O}(|h|^2)$ wegen $f \in \mathcal{F}_2(I_\delta \times \mathbb{R})$ unabhängig von $(t;y) \in I \times \mathbb{R}$ und $|h| \leq \delta$ gewählt werden kann. Einsetzen von (7.45) und (7.46) in den

den lokalen Diskretisierungsfehler (7.43) und anschließendes Sortieren nach jeweils Vielfachen von $f(t;y)$ und seinen partiellen Ableitungen $\frac{\partial f(t;y)}{\partial s}$ und $\frac{\partial f(t;y)}{\partial z}$ liefert: Für $(t;y) \in I \times \mathbb{R}$ und alle $|h| \leq \delta$ gilt

$$\begin{aligned}\tau(t;y;h) &= f(t;y) + \frac{1}{2}\,h\,\frac{\partial f(t;y)}{\partial s} + \frac{1}{2}\,h\,\frac{\partial f(t;y)}{\partial z}\,f(t;y) - \gamma_1\,f(t;y)\\ &\quad - \gamma_2\,f(t;y) - \gamma_2\,\frac{\partial f(t;y)}{\partial s}\,\alpha\,h - \gamma_2\,\frac{\partial f(t;y)}{\partial z}\,\beta\,h\,f(t;y) + \mathcal{O}(|h|^2)\\ &= (1-\gamma_1-\gamma_2)\,f(t;y) + \frac{h}{2}\,(1-2\,\gamma_2\,\alpha)\,\frac{\partial f(t;y)}{\partial s}\\ &\quad + \frac{h}{2}\,(1-2\,\gamma_2\,\beta)\,\frac{\partial f(t;y)}{\partial z}\,f(t;y) + \mathcal{O}(|h|^2). \qquad (7.47)\end{aligned}$$

Wir sehen an (7.47), dass der lokale Diskretisierungsfehler von der Ordnung $\mathcal{O}(|h|^2)$ ist, wenn die folgenden Bedingungen alle erfüllt sind:

$$1-\gamma_1-\gamma_2 = 0, \qquad 1-2\,\gamma_2\,\alpha = 0, \qquad 1-2\,\gamma_2\,\beta = 0.$$

Die letzten beiden Gleichungen können nur erfüllt sein, wenn $\gamma_2 \neq 0$ ist. Auflösen der drei Gleichung nach γ_1 bzw. α bzw. β liefert dann die folgenden Formeln:

$$\boxed{\gamma_2 \in \mathbb{R}\setminus\{0\} \text{ ist beliebig wählbar}, \quad \gamma_1 = 1-\gamma_2, \quad \alpha = \beta = \frac{1}{2\,\gamma_2}\,.} \qquad (7.48)$$

Wählt man in dem expliziten Einschrittverfahren (7.41) mit der Inkrementfunktion (7.42) die Konstanten $\alpha, \beta, \gamma_1, \gamma_2$ so, dass (7.48) erfüllt ist, so erhält man also in (7.47) einen lokalen Diskretisierungsfehler $\tau(t;y;h)$ der Ordnung $\mathcal{O}(|h|^2)$ (denn alle Terme mit einer kleineren Ordnung als $\mathcal{O}(|h|^2)$ verschwinden), d.h. das Verfahren hat die Ordnung 2. Dann folgt aus Satz 7.19, dass das explizite Einschrittverfahren (7.41) mit der Inkrementfunktion (7.42) konvergent ist und die Konvergenzordnung 2 hat.

Beliebte Wahlen für γ_2 in (7.47) sind die Zahlen $\frac{1}{2}$, $\frac{3}{4}$ und 1. Wir halten die zugehörigen expliziten Einschrittverfahren, genannt explizite Runge-Kutta-Verfahren der Ordnung 2, als Verfahren fest:

Verfahren 7.21. (explizite Runge-Kutta-Verfahren der Ordnung 2)
Seien $[a;b] \times \mathbb{R}$ *mit* $-\infty < a < b < \infty$, $t_0 \in [a;b]$, *und* $f : [a;b] \times \mathbb{R} \to \mathbb{R}$ *sei eine stetige Funktion, für die es eine Konstante* $L \geq 0$ *gilt, so dass gilt* $|f(t;v) - f(t;w)| \leq L\,|v-w|$ *für alle* $(t;v), (t;w) \in [a;b] \times \mathbb{R}$. *Die nachfolgenden* ***expliziten Runge-Kutta-Verfahren der Ordnung 2*** *zur numerischen*

Lösung das Anfangswertproblems

$$y'(t) = f\big(t; y(t)\big) \qquad \textit{mit} \qquad y(t_0) = y_0 \tag{7.49}$$

konstruieren jeweils Näherungen y_k der Funktionswerte $y(t_k)$ der eindeutigen Lösung $y : [a; b] \to \mathbb{R}$ von (7.49) an den ***äquidistanten Gitterpunkten***

$$t_k = t_0 + k\,h, \qquad k = 0, 1, 2, \ldots,$$

mit der ***Schrittweite*** *h mit der folgenden* ***Iterationsformeln****:*

(a) ***Verfahren von Heun*** *(mit $\gamma_1 = \gamma_2 = \frac{1}{2}$, $\alpha = \beta = 1$):*
Für $k = 0, 1, 2, \ldots$ führe die folgenden Schritte durch:

(1) Berechne $y_{k+1} := y_k + \dfrac{h}{2}\Big[f(t_k; y_k) + f\big(t_k + h; y_k + h\,f(t_k; y_k)\big)\Big]$.

(2) Setze $t_{k+1} := t_k + h$.

(b) ***Methode von Ralston*** *(mit $\gamma_1 = \frac{1}{4}$, $\gamma_2 = \frac{3}{4}$, $\alpha = \beta = \frac{2}{3}$):*
Für $k = 0, 1, 2, \ldots$ führe die folgenden Schritte durch:

(1) Berechne

$$y_{k+1} := y_k + \frac{h}{4}\,f(t_k; y_k) + \frac{3\,h}{4}\,f\big(t_k + \tfrac{2}{3}\,h; y_k + \tfrac{2}{3}\,h\,f(t_k; y_k)\big).$$

(2) Setze $t_{k+1} := t_k + h$.

(c) ***Explizite Mittelpunktmethode*** *(mit $\gamma_1 = 0$, $\gamma_2 = 1$, $\alpha = \beta = \frac{1}{2}$):*
Für $k = 0, 1, 2, \ldots$ führe die folgenden Schritte durch:

(1) Berechne $y_{k+1} := y_k + h\,f\big(t_k + \frac{1}{2}\,h; y_k + \frac{1}{2}\,h\,f(t_k; y_k)\big)$.

(2) Setze $t_{k+1} := t_k + h$.

Bei den konstruierten expliziten Runge-Kutta-Verfahren der Ordnung 2 handelt es sich um explizite Einschrittverfahren. Man kann auch implizite Runge-Kutta-Verfahren konstruieren, aber dieses besprechen wir in dieser Lehrveranstaltung nicht.

Beispiel 7.22. (Verfahren von Heun)
Betrachten wir das Anfangswertproblem

$$y'(t) = -y(t) + 2\,\cos(t) \qquad \text{mit} \qquad y(0) = 1,$$

dessen eindeutige Lösung $y(t) = \sin(x) + \cos(t)$ ist. Die Funktion f ist also hier $f(t; y) = -y + 2\,\cos(t)$.

h	k	t_k	y_k	$\lvert y_k - y(t_k)\rvert$	$\lvert y_k - y(t_k)\rvert / \lvert y(t_k)\rvert$
$h = 0{,}1$	20	2,0	0,491216	$1{,}93 \cdot 10^{-3}$	0,00392
	40	4,0	−1,40790	$2{,}55 \cdot 10^{-3}$	0,00181
	60	6,0	0,680697	$5{,}81 \cdot 10^{-5}$	0,0000853
	80	8,0	0,841376	$2{,}48 \cdot 10^{-3}$	0,00294
	100	10,0	−1,38097	$2{,}13 \cdot 10^{-3}$	0,00154
$h = 0{,}05$	40	2,00	0,492682	$4{,}68 \cdot 10^{-4}$	0,000949
	80	4,00	−1,40982	$6{,}25 \cdot 10^{-4}$	0,000443
	120	6,00	0,680735	$2{,}01 \cdot 10^{-5}$	0,0000296
	160	8,00	0,843254	$6{,}04 \cdot 10^{-4}$	0,000716
	200	10,00	−1,38257	$5{,}23 \cdot 10^{-4}$	0,000378

Tabelle 7.5: Verfahren von Heun für $y' = -y + 2\,\cos(t)$ mit $y(0) = 1$.

Wir nutzen das Verfahren von Heun mit den Schrittweiten $h = 0{,}1$ bzw. $h = 0{,}05$, um Näherungswerte für $y(t)$ mit $t \in \{2; 4; 6; 8; 10\}$ zu berechnen. Das Verfahren von Heun (siehe Verfahren 7.21 (a)) hat für dieses konkrete Beispiel die Iterationsvorschrift

$$\begin{aligned}
y_{k+1} &= y_k + \frac{h}{2}\Big[f(t_k; y_k) + f\big(t_k + h; y_k + h\, f(t_k; y_k)\big)\Big] \\
&= y_k + \frac{h}{2}\Big[f(t_k; y_k) - \big(y_k + h\, f(t_k; y_k)\big) + 2\,\cos(t_k + h)\Big] \\
&= y_k + \frac{h}{2}\Big[-y_k + (1-h)\, f(t_k; y_k) + 2\,\cos(t_k + h)\Big] \\
&= y_k + \frac{h}{2}\Big[-y_k + (1-h)\big(-y_k + 2\,\cos(t_k)\big) + 2\,\cos(t_k + h)\Big] \\
&= y_k + \frac{h}{2}\Big[-(2-h)\, y_k + 2\,(1-h)\,\cos(t_k) + 2\,\cos(t_k + h)\Big],
\end{aligned}$$

$t_{k+1} = t_k + h$, $k = 0, 1, 2, \ldots$, mit dem Startwert $y_0 = 1$ für $t_0 = 0$. Die Näherungen für $y(t)$ mit $t \in \{2; 4; 6; 8; 10\}$ sind in Tabelle 7.5 (mit Rundung auf eine Gleitkommadarstellung mit 6-stelliger Mantisse) angegeben. Per Konstruktion (als Runge-Kutta-Verfahren der Ordnung 2) sollte das Verfahren von Heun die Konvergenzordnung $p = 2$ haben, d.h. es sollte für die absoluten Fehler der berechneten Näherungswerte y_k für $y(t_k)$ gelten $|y_k - y(t_k)| \leq K\,|h|^2$. Dieses ist zunächst nicht hilfreich, da wir die Konstante K nicht kennen. Es folgt aber bei

Halbierung der Schrittweite für die neuen Näherungswerte y_k für $y(t_k)$

$$|y_k - y(t_k)| \leq K\,(|h|/2)^2 = |y_k - y(t_k)| \leq \frac{1}{4}\,(K\,|h|^2),$$

d.h. bei der Halbierung der Schrittweite sollten die absoluten Fehler ungefähr ein Viertel der absoluten Fehler vor der Halbierung der Schrittweite betragen. Wir überprüfen dieses für die Werte aus Tabelle 7.5, indem wir die Quotienten der absoluten Fehler bilden (mit $h_2 = 0{,}05 = h_1/2$ wurde die Schrittweite halbiert):

t	2,0	4,0	6,0	8,0	10,0
$\dfrac{\text{absoluter Fehler der Näherung in } t \text{ für } h = 0{,}1}{\text{absoluter Fehler der Näherung in } t \text{ für } h = 0{,}05}$	4,12	4,08	2,89	4,11	4,07

Bis auf den absoluten Fehler der Näherung bei Fehler bei $t = 6$ verhalten sich die absoluten Fehler der Näherungen wie erwartet. ♠

Die expliziten Runge-Kutta-Verfahren in Verfahren 7.21 haben unter passenden Voraussetzungen an f die Konsistenzordnung und die Konvergenzordnung $p = 2$.

Satz 7.23. (Konvergenzordnung von Verfahren 7.21)
*Seien $I = [a; b]$ und $I_\delta = [a - \delta; b + \delta]$ mit $\delta > 0$. Seien $(t_0; y_0) \in I \times \mathbb{R}$ und $f \in \mathcal{F}_2(I_\delta \times \mathbb{R})$. Die in Verfahren 7.21 zur Lösung des Anfangswertproblems $y'(t) = f\big(t; y(t)\big)$ mit $y(t_0) = y_0$ eingeführten expliziten Runge-Kutta-Verfahren der Ordnung 2 sind **konvergent** und haben die **Konvergenzordnung** $p = 2$.*

Beweis von Satz 7.23: Wir wollen Satz 7.19 verwenden und überprüfen die Voraussetzungen: Sei $G := I \times \mathbb{R} \times [-\delta; \delta]$. Weil $f \in \mathcal{F}_2(I_\delta \times \mathbb{R})$ ist, existiert eine Konstante $L > 0$, so dass

$$|f(t; v) - f(t; w)| \leq L\,|v - w| \qquad \text{für alle } (t; v), (t; w) \in I_\delta \times \mathbb{R} \tag{7.50}$$

gilt. Für die durch (7.42) gegebene Inkrementfunktion

$$\Phi(t; y; h) = \gamma_1\, f(t; y) + \gamma_2\, f\big(t + \alpha\, h; y + \beta\, h\, f(t; y)\big),$$

folgt durch wiederholte Anwendung der Dreiecksungleichung und von (7.50)

$$|\Phi(t; v; h) - \Phi(t; w; h)| \leq |\gamma_1|\,|f(t; v) - f(t; w)|$$

$$
\begin{aligned}
&+ |\gamma_2| \left| f\big(t + \alpha\, h; v + \beta\, h\, f(t;v)\big) - f\big(t + \alpha\, h; w + \beta\, h\, f(t;w)\big) \right| \\
&\leq |\gamma_1|\, L\, |v - w| + |\gamma_2|\, L \left| \big(v + \beta\, h\, f(t;v)\big) - \big(w + \beta\, h\, f(t;w)\big) \right| \\
&= |\gamma_1|\, L\, |v - w| + |\gamma_2|\, L \left| (v - w) + \beta\, h \left(f(t;v) - f(t;w) \right) \right| \\
&\leq |\gamma_1|\, L\, |v - w| + |\gamma_2|\, L \left(|v - w| + |\beta|\, |h| \left| f(t;v) - f(t;w) \right| \right) \\
&\leq |\gamma_1|\, L\, |v - w| + |\gamma_2|\, L \left(|v - w| + |\beta|\, |h|\, L\, |v - w| \right) \\
&\leq \big(|\gamma_1| + |\gamma_2| + |\gamma_2|\, |\beta|\, \delta\, L \big)\, L\, |v - w| \quad \text{für alle } (t;v), (t;w) \in I \times \mathbb{R},\ |h| \leq \delta.
\end{aligned}
$$

Nach den Überlegungen zur Beginn der Teilkapitels gilt nach (7.47) für den lokalen Diskretisierungsfehler für alle Wahlen vom $\alpha, \beta, \gamma_1, \gamma_2$ gemäß (7.48)

$$\left| \tau\big(t; y(t); h\big) \right| \leq K\, |h|^2 \qquad \text{für alle } t \in I \text{ und alle } |h| \leq \delta.$$

Nach Satz 7.19 sind explizite Runge-Kutta-Verfahren der Ordnung $p = 2$, und insbesondere auch diejenigen in Verfahren 7.21, konvergent und haben die Konvergenzordnung $p = 2$. □

Das folgende Runge-Kutta-Verfahren der Ordnung $p = 4$ wird häufig in der Praxis eingesetzt.

Verfahren 7.24. (explizites Runge-Kutta-Verfahren der Ordnung 4)
Seien $[a;b] \times \mathbb{R}$ *mit* $-\infty < a < b < \infty$, $t_0 \in [a;b]$, *und* $f : [a;b] \times \mathbb{R} \to \mathbb{R}$ *sei eine stetige Funktion, für die es eine Konstante* $L \geq 0$ *gilt, so dass gilt* $|f(t;v) - f(t;w)| \leq L\, |v - w|$ *für alle* $(t;v), (t;w) \in [a;b] \times \mathbb{R}$. *Das nachfolgende* ***explizite Runge-Kutta-Verfahren der Ordnung 4*** *zur numerischen Lösung das Anfangswertproblems*

$$y'(t) = f\big(t; y(t)\big) \qquad \textit{mit} \qquad y(t_0) = y_0 \tag{7.51}$$

konstruiert Näherungen y_k *der Funktionswerte* $y(t_k)$ *der eindeutigen Lösung* $y : [a;b] \to \mathbb{R}$ *von* (7.51) *an den* ***äquidistanten Gitterpunkten***

$$t_k = t_0 + k\, h, \qquad k = 0, 1, 2, \ldots,$$

mit der ***Schrittweite*** h *mit der folgenden* ***Iterationsformel****:*

Für $k = 0, 1, 2, \ldots$ *führe die folgenden Schritte durch:*

(1) Berechne $\eta_1 := f(t_k; y_k)$.

(2) Berechne $\eta_2 := f\left(t_k + \dfrac{h}{2}\,;\, y_k + \dfrac{h}{2}\, \eta_1 \right)$.

(3) Berechne $\eta_3 := f\left(t_k + \frac{h}{2}\,;\, y_k + \frac{h}{2}\,\eta_2\right)$.

(4) Berechne $\eta_4 := f(t_k + h; y_k + h\,\eta_3)$.

(5) Berechne $y_{k+1} := y_k + \frac{h}{6}\left(\eta_1 + 2\,\eta_2 + 2\,\eta_3 + \eta_4\right)$.

(6) Setze $t_{k+1} = t_k + h$.

Auf dem Übungszettel betrachten wir ein Beispiel für Verfahren 7.24.

Was kann man über die Konvergenzordnung des expliziten Runge-Kutta-Verfahrens der Ordnung 4 (aus Verfahren 7.24) aussagen? Der nächste Satz beantwortet diese Frage.

Satz 7.25. (Konvergenzordnung von Verfahren 7.24)
Seien $I = [a; b]$ *und* $I_\delta = [a - \delta; b + \delta]$ *mit* $\delta > 0$. *Seien* $(t_0; y_0) \in I \times \mathbb{R}$ *und* $f \in \mathcal{F}_4(I_\delta \times \mathbb{R})$. *Das in Verfahren 7.24 zur Lösung des Anfangswertproblems* $y'(t) = f\big(t; y(t)\big)$ *mit* $y(t_0) = y_0$ *eingeführte explizite Runge-Kutta-Verfahren der Ordnung* 4 *ist* ***konvergent*** *und hat die* ***Konvergenzordnung*** $p = 4$.

7.6 Ausblick: Mehrschrittverfahren

Bisher haben wir nur Einschrittverfahren kennengelernt. Bei diesen hängt die Inkrementfunktion Φ in der Iterationsvorschrift $y_{k+1} = y_k + h\,\Phi(t_k; y_k; y_{k+1}; h)$ von t_k, y_k, h und bei einem impliziten Einschrittverfahren noch von y_{k+1} ab. Bei einem **expliziten Mehrschrittverfahren** wird die neue Näherung y_{k+r} nicht nur in Abhängigkeit von t_k, y_{k+r-1} und h, sondern noch zusätzlich in Abhängigkeit von den Näherungen $y_{k+r-2}, \ldots, y_{k+1}, y_k$, also **in Abhängigkeit von (allen oder einigen) Näherungen aus den r vorherliegenden Schritten** berechnet.

Definition 7.26. (Mehrschrittverfahren)
Seien $[a; b] \times \mathbb{R}$ *mit* $-\infty < a < b < \infty$, $t_0 \in [a; b]$, *und* $f : [a; b] \times \mathbb{R} \to \mathbb{R}$ *sei eine stetige Funktion, für die es eine Konstante* $L \geq 0$ *gilt, so dass gilt* $|f(t; v) - f(t; w)| \leq L\,|v - w|$ *für alle* $(t; v), (t; w) \in [a; b] \times \mathbb{R}$. *Ein* ***Mehrschrittverfahren*** *zur numerischen Lösung das Anfangswertproblems*

$$y' = f(t; y) \qquad \textit{mit} \qquad y(t_0) = y_0 \tag{7.52}$$

konstruiert Näherungen y_k der Funktionswerte $y(t_k)$ der eindeutigen Lösung $y : [a;b] \to \mathbb{R}$ von (7.52) an den ***äquidistanten Gitterpunkten***

$$t_k = t_0 + k\,h, \qquad k = 0, 1, 2, \ldots,$$

mit der ***Schrittweite*** $h \neq 0$ *mit einer* ***Iterationsformel*** *der Form*

$$\begin{aligned} y_{k+r} := {} & a_{r-1}\,y_{k+r-1} + a_{r-2}\,y_{k+r-2} + \ldots + a_1\,y_{k+1} + a_0\,y_k \\ & + h\,\Phi(t_k; y_k; y_{k+1}; \ldots; y_{k+r-1}; y_{k+r}; h), \qquad k = 0, 1, 2, \ldots, \end{aligned} \tag{7.53}$$

wobei $a_0, a_1, \ldots, a_{r-2}, a_{r-1} \in \mathbb{R}$ Konstanten sind und wobei die Funktion Φ mit Hilfe der rechten Seite $f : [a;b] \times \mathbb{R} \to \mathbb{R}$ der Differentialgleichung in (7.52) definiert ist. Die Funktion Φ heißt die ***Inkrementfunktion*** *des Mehrschrittverfahrens. Falls $\Phi(t_k; y_k; y_{k+1}; \ldots; y_{k+r-1}; y_{k+r}; h)$ nicht von dem vorletzten Argument y_{k+r} abhängt, nennt man das Mehrschrittverfahren (7.53)* ***explizit****. Andernfalls nennt man das Mehrschrittverfahren (7.53)* ***implizit****. Für das Mehrschrittverfahren (7.53) benötigt man als Startwerte t_0, y_0 und Näherungen $y_1, y_2, \ldots, y_{r-1}$ der Funktionswerte $y(t_1), y(t_2), \ldots, y(t_{r-1})$ der Lösung $y : [a;b] \to \mathbb{R}$ von (7.52).*

Wir erklären nun, wie **Mehrschrittverfahren mit Hilfe von Polynominterpolation konstruiert werden können**: Dabei seien $[a;b] \times \mathbb{R}$ mit $-\infty < a < b < \infty$, $t_0 \in [a;b]$, und $f : [a;b] \times \mathbb{R} \to \mathbb{R}$ sei eine stetige Funktion, für die es eine Konstante $L \geq 0$ gilt, so dass $|f(t;v) - f(t;w)| \leq L\,|v - w|$ für alle $(t;v), (t;w) \in [a;b] \times \mathbb{R}$ gilt. Dann hat das Anfangswertproblem

$$y'(t) = f\big(t; y(t)\big) \qquad \text{mit} \qquad y(t_0) = y_0 \tag{7.54}$$

eine eindeutige Lösung $y : [a;b] \to \mathbb{R}$. Wir beginnen damit, dass wir (7.54) mit Hilfe des Hauptsatzes der Differentialrechnung und Integralrechnung in eine Integralgleichung umschreiben: Mit $r \in \mathbb{N}$, $\ell \in \{1; 2; \ldots; r\}$ und $h \neq 0$ gilt

$$\begin{aligned} y(s + rh) &= y\big(s + (r-\ell)h\big) + \big[y(s + rh) - \big(s + (r-\ell)h\big)\big] \\ &= y\big(s + (r-\ell)h\big) + \int_{s+(r-\ell)h}^{s+rh} y'(t)\,\mathrm{d}t \\ &= y\big(s + (r-\ell)h\big) + \int_{s+(r-\ell)h}^{s+rh} f\big(t; y(t)\big)\,\mathrm{d}t, \end{aligned} \tag{7.55}$$

wobei natürlich $s + (r-\ell)h, s + rh \in [a;b]$ gelten muss. Insbesondere folgt aus (7.55) mit $s = t_k$ und damit $s + rh = t_k + rh = t_{k+r}$ und $s + (r-\ell)h =$

$t_k + (r-\ell)h = t_{k+r-\ell}$, dass gilt

$$y(t_{k+r}) = y(t_{k+r-\ell}) + \int_{t_{k+r-\ell}}^{t_{k+r}} f\big(t; y(t)\big)\,\mathrm{d}t, \qquad \text{wenn } t_{k+r-\ell}, t_{k+r} \in [a;b]. \quad (7.56)$$

Der Integrand in (7.56) soll nun durch das Interpolationspolynom P_m mit $m \leq r$ vom Grad $\leq m$ bzgl. der $m+1$ Datenpunkte

$$\big(t_{k+j}; f(t_{k+j}; y_{k+j})\big), \qquad j = 0, 1, \ldots, m,$$

ersetzt werden, wobei m oft $r-1$ oder $m = r$ ist. Das Interpolationspolynom P_m vom Grad $\leq m$ soll also die Bedingungen

$$P_m(t_{k+j}) = f(t_{k+j}; y_{k+j}) \qquad \text{für } j = 0, 1, 2, \ldots, m,$$

erfüllen. Mit der Interpolationsformel von Lagrange gilt dann

$$P_m(x) = \sum_{j=0}^{m} f(t_{k+j}; y_{k+j})\, L_j(t), \quad (7.57)$$

wobei $L_0, L_1, \ldots, L_m$ die Lagrange-Polynome vom Grad m bzgl. der $m+1$ Punkte t_{k+j}, $j = 0, 1, 2, \ldots, m$, sind. Einsetzen der Formel (7.57) für P_m in (7.56) ergibt

$$\begin{aligned} y(t_{k+r}) \approx y_{k+r} &:= y_{k+r-\ell} + \int_{t_{k+r-\ell}}^{t_{k+r}} P_m(t)\,\mathrm{d}t \\ &= y_{k+r-\ell} + \int_{t_{k+r-\ell}}^{t_{k+r}} \left(\sum_{j=0}^{m} f(t_{k+j}; y_{k+j})\, L_j(t) \right) \mathrm{d}t \\ &= y_{k+r-\ell} + \sum_{j=0}^{m} f(t_{k+j}; y_{k+j}) \int_{t_{k+r-\ell}}^{t_{k+r}} L_j(t)\,\mathrm{d}t, \end{aligned} \quad (7.58)$$

wobei wir auch noch $y(t_{k+r-\ell})$ durch seine Näherung $y_{k+r-\ell}$ ersetzt haben. Mit den (nicht von k abhängigen) Konstanten

$$w_j := \int_{t_{k+r-\ell}}^{t_{k+r}} L_j(t)\,\mathrm{d}t, \qquad j = 0, 1, 2, \ldots, m, \quad (7.59)$$

erhält man also die Iterationsvorschrift

$$\boxed{y_{k+r} = y_{k+r-\ell} + \sum_{j=0}^{m} w_j\, f(t_{k+j}; y_{k+j}), \qquad k = 0, 1, 2, \ldots.} \quad (7.60)$$

Man könnte zunächst meinen, dass die durch die Integrale über die Lagrange-Polynome definieren Konstanten (7.59) von k abhängen, denn mit einer Änderung von k ändern sich die $m+1$ Punkte t_{k+j}, $j = 0, 1, 2, \ldots, m$, und damit auch die Lagrange-Polynome $L_0, L_1, \ldots, L_m$. Das Integrationsintervall in (7.59) ändert sich aber auch. Da wir äquidistante Punkte $t_k = t_0 + k\,h$, $k = 0, 1, 2, \ldots$, verwenden, kann man nun durch eine einfache Substitution zeigen, dass die Werte w_j der Integrale in (7.59) nicht von k abhängen.

Als Beispiel konstruieren wir eine Adams-Bashforth-Methode der Ordnung 2:

Beispiel 7.27. (Adams-Bashforth-Methode der Ordnung 2)
Es seien $[a;b] \times \mathbb{R}$ mit $-\infty < a < b < \infty$, $t_0 \in [a;b]$, und $f : [a;b] \times \mathbb{R} \to \mathbb{R}$ sei eine stetige Funktion, für die es eine Konstante $L \geq 0$ gilt, so dass gilt $|f(t;v) - f(t;w)| \leq L\,|v - w|$ für alle $(t;v), (t;w) \in [a;b] \times \mathbb{R}$. Dann hat das Anfangswertproblem

$$y'(t) = f\big(t; y(t)\big) \qquad \text{mit} \qquad y(t_0) = y_0 \tag{7.61}$$

eine eindeutige Lösung $y : [a;b] \to \mathbb{R}$.

Es sei eine Schrittweite h gegeben, und wir nutzen die äquidistanten Gitterpunkte

$$t_{k+1} := t_k + h = t_0 + (k+1)\,h, \qquad k = 0, 1, 2, \ldots.$$

In dem obigen Ansatz wählen wir nun $r = 2$, $\ell = 1$ und $m = 1$. Dann erhalten wir in der ersten Zeile von (7.58)

$$y_{k+2} := y_{k+1} + \int_{t_{k+1}}^{t_{k+2}} P_1(t)\,\mathrm{d}t, \tag{7.62}$$

wobei P_1 das Interpolationspolynom bzgl. der Datenpunkte $\big(t_k; f(t_k; y_k)\big)$ und $\big(t_{k+1}; f(t_{k+1}; y_{k+1})\big)$ ist. Wir haben also das lineare Interpolationspolynom

$$\begin{aligned} P_1(t) &= f(t_k; y_k)\, \underbrace{\frac{t - t_{k+1}}{t_k - t_{k+1}}}_{= L_0(t)} + f(t_{k+1}; y_{k+1})\, \underbrace{\frac{t - t_k}{t_{k+1} - t_k}}_{= L_1(t)} \\ &= f(t_k; y_k)\, \frac{1}{(-h)}\,(t - t_{k+1}) + f(t_{k+1}; y_{k+1})\, \frac{1}{h}\,(t - t_k). \end{aligned} \tag{7.63}$$

Einsetzen von (7.63) in (7.62) liefert

$$y_{k+2} = y_{k+1} + \int_{t_{k+1}}^{t_{k+2}} \left(f(t_k; y_k)\, \frac{1}{(-h)}\,(t - t_{k+1}) + f(t_{k+1}; y_{k+1})\, \frac{1}{h}\,(t - t_k) \right) \mathrm{d}t$$

$$= y_{k+1} - f(t_k; y_k)\,\frac{1}{h}\int_{t_{k+1}}^{t_{k+2}} (t - t_{k+1})\,\mathrm{d}t + f(t_{k+1}; y_{k+1})\,\frac{1}{h}\int_{t_{k+1}}^{t_{k+2}} (t - t_k)\,\mathrm{d}t, \tag{7.64}$$

und das Berechnen der beiden Integrale ergibt

$$\int_{t_{k+1}}^{t_{k+2}} (t - t_{k+1})\,\mathrm{d}t = \left[\frac{1}{2}\,(t - t_{k+1})^2\right]_{t=t_{k+1}}^{t=t_{k+2}} = \frac{1}{2}\,(t_{k+2} - t_{k+1})^2 = \frac{1}{2}\,h^2, \tag{7.65}$$

$$\begin{aligned}\int_{t_{k+1}}^{t_{k+2}} (t - t_k)\,\mathrm{d}t &= \left[\frac{1}{2}\,(t - t_k)^2\right]_{t=t_{k+1}}^{t=t_{k+2}} = \frac{1}{2}\,(t_{k+2} - t_k)^2 - \frac{1}{2}\,(t_{k+2} - t_{k+1})^2 \\ &= \frac{1}{2}\,(2h)^2 - \frac{1}{2}\,h^2 = \frac{3}{2}\,h^2.\end{aligned} \tag{7.66}$$

Einsetzen von (7.65) und (7.66) in (7.64) ergibt

$$\begin{aligned}y_{k+2} &= y_{k+1} - f(t_k; y_k)\,\frac{1}{h}\,\frac{1}{2}\,h^2 + f(t_{k+1}; y_{k+1})\,\frac{1}{h}\,\frac{3}{2}\,h^2 \\ &= y_{k+1} - \frac{h}{2}\,f(t_k; y_k) + \frac{3h}{2}\,f(t_{k+1}; y_{k+1}) \\ &= y_{k+1} + \frac{h}{2}\,\big[3\,f(t_{k+1}; y_{k+1}) - f(t_k; y_k)\big].\end{aligned}$$

Wir ersetzen noch $k+1$ durch k und erhalten damit die folgende Iterationsformel:

$$\boxed{y_{k+1} = y_k + \frac{h}{2}\,\big[3\,f(t_k; y_k) - f(t_{k-1}; y_{k-1})\big], \quad t_{k+1} = t_k + h, \quad k = 0, 1, 2, \ldots} \tag{7.67}$$

Wir wollen diese **Adams-Bashforth-Methode der Ordnung** 2 nun anwenden, um das Anfangswertproblem

$$y'(t) = -y(t) + 2\,\cos(t) \qquad \text{mit} \qquad y(0) = 1,$$

dessen eindeutige Lösung $y(t) = \sin(x) + \cos(t)$ ist, numerisch zu lösen. Mit $f(t; y) = -y + 2\,\cos(t)$ lautet die konkrete Iterationsformel der Adams-Bashforth-Methode der Ordnung 2 zur Berechnung von y_{k+1} aus (7.67)

$$y_{k+1} = y_k + \frac{h}{2}\,\big[3\,\big(-y_k + 2\,\cos(t_k)\big) - \big(-y_{k-1} + 2\,\cos(t_{k-1})\big)\big].$$

Für die Schrittweiten $h = 0{,}1$ und $h = 0{,}05$ und die Startwerte $t_0 = 0$, $y_0 = 1$ und $y_1 = y(h) = \sin(h) + \cos(h)$ sind die Näherungswerte y_k (mit Rundung auf eine Gleitkommadarstellung mit 6-stelliger Mantisse) in Tabelle 7.6 angegeben.

Wir beobachten (siehe nachfolgende Tabelle), dass bei der Halbierung der Schrittweite die absoluten Fehler (mit Ausnahme des Wertes bei $t = 8$) ungefähr ein Viertel der absoluten Fehler vor der Halbierung der Schrittweite betragen.

h	k	t_k	y_k	$\lvert y_k - y(t_k)\rvert$	$\lvert y_k - y(t_k)\rvert / \lvert y(t_k)\rvert$
$h = 0{,}1$	20	2,0	0,491019	$2{,}13 \cdot 10^{-3}$	0,00432
	40	4,0	−1,41343	$2{,}98 \cdot 10^{-3}$	0,00211
	60	6,0	0,684661	$3{,}91 \cdot 10^{-3}$	0,00574
	80	8,0	0,843490	$3{,}68 \cdot 10^{-4}$	0,000436
	100	10,0	−1,38671	$3{,}61 \cdot 10^{-3}$	0,00261
$h = 0{,}05$	40	2,00	0,492597	$5{,}53 \cdot 10^{-4}$	0,00112
	80	4,00	−1,41117	$7{,}24 \cdot 10^{-4}$	0,000513
	120	6,00	0,681743	$9{,}88 \cdot 10^{-4}$	0,00145
	160	8,00	0,843737	$1{,}21 \cdot 10^{-4}$	0,000144
	200	10,00	−1,38398	$8{,}90 \cdot 10^{-4}$	0,000643

Tabelle 7.6: Adams-Bashforth-Methode der Ordnung 2 für das Anfangswertproblem $y' = -y + 2\cos(t)$ mit $y(0) = 1$.

t	2,0	4,0	6,0	8,0	10,0
$\dfrac{\text{absoluter Fehler der Näherung in } t \text{ für } h = 0{,}1}{\text{absoluter Fehler der Näherung in } t \text{ für } h = 0{,}05}$	3,86	4,11	3,96	3,03	4,06

Daher vermuten wir, dass sich der Fehler der Adams-Bashforth-Methode der Ordnung 2 wie $\mathcal{O}(|h|^2)$ verhält. ♠