Einführung in die Numerische Mathematik

Steffen Börm

Stand 21. Januar 2021

Inhaltsverzeichnis

1 Einleitung

Viele Gesetzmäßigkeiten in den Natur-, Ingenieur- und auch Wirtschaftswissenschaften werden mit Hilfe mathematischer Gleichungen beschrieben: Die auf Newton zurückgehenden Axiome der klassischen Mechanik werden in der Regel mit Hilfe von Differentialgleichungen ausgedrückt, das Verhalten einfacher Schaltungen mit Hilfe der Kirchhoff'schen Gesetze, während bei der Modellierung des zu erwartenden Gewinns eines Aktienpakets Integrale zum Einsatz kommen.

In einfachen Fällen lassen sich diese Gleichungen per Hand lösen, in der Praxis ist es jedoch wesentlich häufiger nicht möglich: Entweder lässt sich die Lösung nicht in Gestalt einer expliziten Formel darstellen, oder es sind schlicht zu viele Variablen im Spiel.

Die *numerische Mathematik* (oder kurz *Numerik*) beschäftigt sich mit der Frage, wie derartige Probleme möglichst schnell gelöst werden können. Dabei ist die Wahl des richtigen Verfahrens von entscheidender Bedeutung: Ein lineares Gleichungssystem mit der Cramer'schen Regel aufzulösen ist zwar theoretisch machbar, führt in der Praxis aber schon bei relativ kleinen Problemen zu einem inakzeptablen Zeitaufwand. Das Gauß'sche Eliminationsverfahren dagegen ist wesentlich effizienter und kann auf modernen Computern auch noch Systeme mit 10 000 Unbekannten in vertretbarer Zeit behandeln.

Ein weiterer wichtiger Gesichtspunkt ist die Genauigkeit der Berechnung: Bei Berechnungen auf Grundlage von Messdaten treten immer auch Messfehler auf, die das Ergebnis der Berechnung verfälschen. Bei der Modellierung eines physikalischen Prozesses wird man in der Regel von vereinfachenden Annahmen ausgehen und so einen Modellfehler einführen, der ebenfalls das Ergebnis verändert. Schließlich kommen bei der Durchführung der Berechnung auf einem Computer auch noch Rundungsfehler hinzu. Um die Qualität des Ergebnisses beurteilen zu können, müssen wir also auch untersuchen, welche Genauigkeit wir überhaupt erwarten dürfen.

Viele numerische Verfahren bieten die Möglichkeit, einen Kompromiss zwischen Geschwindigkeit und Genauigkeit einzugehen: Falls wir wissen, dass wir wegen Messfehlern oder anderen Ungenauigkeiten ohnehin nur ein Ergebnis mit 3% Genauigkeit erwarten dürfen, ist es unkritisch, einen zusätzlichen Fehler von 1% in Kauf zu nehmen, falls sich dadurch die Rechenzeit erheblich reduzieren lässt. Deshalb werden viele der im Rahmen dieser Vorlesung vorgestellten Verfahren Näherungsverfahren sein, bei denen wir sowohl Zeitbedarf als auch Genauigkeit analysieren.

Die folgenden Kapitel gliedern sich wie folgt:

- Kapitel 2 beschäftigt sich mit der Frage, wie sich Fehler bei Berechnungen fortpflanzen, wie genau das Ergebnis also sein kann, wenn gestörten Ausgangsdaten vorliegen. Als Anwendung der allgemeinen Untersuchung betrachten wir die Approximation der Menge $\mathbb{R}$ der reellen Zahlen durch *Maschinenzahlen*.

- Kapitel 3 untersucht die Behandlung linearer Gleichungssysteme. Im Mittelpunkt stehen hier effiziente Lösungsverfahren sowie ihre Umsetzung auf einem Computer. Da Computer grundsätzlich mit Maschinenzahlen arbeiten, stellt sich auch die Frage nach Verfahren, die auf die dadurch eingeführten Störungen nicht allzu empfindlich reagieren.

- Kapitel 4 behandelt die mit linearen Gleichungssystemen eng verwandten linearen Ausgleichsprobleme. Derartige Probleme treten häufig etwa bei der Auswertung von Messdaten oder bei der Optimierung der Parameter eines Systems auf. Ihre Lösung lässt sich auf die in Kapitel 3 eingeführten Verfahren zurückführen.

- Kapitel 5 konzentriert sich auf nichtlineare Gleichungssysteme, deren Anwendungsbereich sich von der Berechnung von Wurzeln bis hin zur Simulation strömungsdynamischer Vorgänge erstreckt. Mit Hilfe geeigneter Ansätze lässt sich das Lösen dieser Probleme auf die in Kapitel 3 vorgestellten Techniken zurückführen, allerdings kann in der Regel auch bei exakten Ausgangsdaten keine exakte Lösung mehr bestimmt werden, sondern nur eine beliebig gute Näherung.

- Kapitel 6 untersucht die Behandlung von Funktionen. Da sich allgemeine Funktionen nicht im Computer darstellen lassen, werden sie häufig durch Polynome ersetzt, und wir untersuchen die dabei erreichte Genauigkeit sowie effiziente Verfahren für die Konstruktion geeigneter Polynome und deren Auswertung.

- Kapitel 7 wendet die in Kapitel 6 eingeführten Techniken an, um Integrale näherungsweise zu berechnen. Die vorgestellten Methoden sind besonders interessant, da sie lediglich auf der Auswertung des Integranden in einigen Punkten des Integrationsgebiets beruhen und nicht auf die in der Praxis häufig nicht verfügbaren Stammfunktion.

- Kapitel 8 benutzt die in Kapitel 7 eingeführten Techniken, um gewöhnliche Differentialgleichungen näherungsweise zu lösen. Auch hier benötigen die gängigen Verfahren lediglich Auswertungen der die Gleichung beschreibenden Funktion, aber keine Stammfunktionen, so dass sie sich sehr einfach und flexibel einsetzen lassen.

- Kapitel 9 schließlich bietet einen Ausblick auf einige Anwendungen der in den vorangehenden Kapiteln behandelten Verfahren.

Danksagung

Ich bedanke mich bei Jens Burmeister, Knut Reimer, Dirk Boysen, Tobias Löffler, Christoph Gerken, Simon Groth, Max Amadeus Deppert, Christina Börst, Jens Liebenau, Sven Christophersen, Dennis Papesch, Torsten Knauf, Einhard Leichtfuß, Jelle Mathis Kuiper, Jonas Lorenzen, Maurice Mowinkel, Paul Luis Röhl und Jonna Matthiesen für Hinweise auf Fehler in früheren Fassungen dieses Skripts und für Verbesserungsvorschläge.

2 Kondition, Maschinenzahlen und Stabilität

2.1 Kondition

Bevor wir uns konkreten Verfahren zur Lösung mathematischer Probleme zuwenden können, müssen wir zunächst untersuchen, was wir von der Lösung überhaupt erwarten können: In der Praxis werden Berechnungen oft auf Messwerten basieren, und diese Messwerte sind praktisch immer mit einem gewissen Messfehler behaftet. Um das Ergebnis einer Berechnung beurteilen zu können, müssen wir deshalb wissen, wie sie auf Störungen der Eingabedaten reagiert.

Wir beschreiben dazu eine Berechnung durch eine Abbildung φ, die eine Menge von Eingabedaten X auf eine Menge von Ausgabedaten Y abbildet. Um die Größe von Fehlern beurteilen zu können, verwenden wir Vektorräume und Normen.

Definition 2.1 (Notationen) *Wir bezeichnen mit $\mathbb{K} \in \{\mathbb{R}, \mathbb{C}\}$ den Körper der reellen oder komplexen Zahlen.*

Teilintervalle der ganzen Zahlen schreiben wir als $[n : m] := \{z \in \mathbb{Z} \ : \ n \leq z \leq m\}$.

Definition 2.2 (Norm) *Sei $\mathcal{V}$ ein Vektorraum über dem Körper $\mathbb{K} \in \{\mathbb{R}, \mathbb{C}\}$ der reellen oder komplexen Zahlen, und sei $f : \mathcal{V} \to \mathbb{R}_{\geq 0}$ eine Abbildung, die die folgenden Eigenschaften besitzt:*

- *Für alle $\mathbf{x} \in \mathcal{V}$ gilt $f(\mathbf{x}) = 0$ genau dann, wenn $\mathbf{x} = \mathbf{0}$ gilt.*
- *Für alle $\mathbf{x} \in \mathcal{V}$ und $\alpha \in \mathbb{R}$ gilt $f(\alpha\mathbf{x}) = |\alpha| f(\mathbf{x})$.*
- *Für alle $\mathbf{x}, \mathbf{y} \in \mathcal{V}$ gilt $f(\mathbf{x} + \mathbf{y}) \leq f(\mathbf{x}) + f(\mathbf{y})$.*

Dann nennen wir f eine Norm. *In der Regel werden Normen in der Form $\|\mathbf{x}\|$ statt als $f(\mathbf{x})$ notiert.*

Ein typisches Beispiel für eine Norm ist die Maximum-Norm, die sich als einfache Verallgemeinerung des Betrags reeller Zahlen ergibt:

Beispiel 2.3 (Diskrete Maximum-Norm) *Sei $n \in \mathbb{N}$, und sei $\mathcal{V} = \mathbb{K}^n$ der Vektorraum der n-dimensionalen reellen oder komplexen Vektoren. Dann ist*

$$\|\cdot\|_\infty \colon \mathcal{V} \to \mathbb{R}_{\geq 0}, \qquad \mathbf{x} \mapsto \max\{|x_i| \ : \ i \in [1 : n]\},$$

eine Norm auf $\mathcal{V}$, die wir als die (diskrete) Maximum-Norm *bezeichnen.*

Falls bei einer Berechnung alle Parameter mehr oder weniger gleichberechtigt eingehen, kann es sinnvoll sein, eine Norm mit Hilfe der Summe zu konstruieren:

Beispiel 2.4 (Diskrete Summen-Norm) *Sei $n \in \mathbb{N}$, und sei $\mathcal{V} = \mathbb{K}^n$ der Vektorraum der n-dimensionalen reellen oder komplexen Vektoren. Dann ist*

$$\|\cdot\|_1 : \mathcal{V} \to \mathbb{R}_{\geq 0}, \qquad \mathbf{x} \mapsto \sum_{i=1}^{n} |x_i|,$$

eine Norm auf $\mathcal{V}$, die wir als die (diskrete) Summen-Norm *(oder die* Manhattan-Norm*) bezeichnen.*

Für geometrische Betrachtungen ist die euklidische Norm wichtiger, denn sie spiegelt dank des Satzes von Pythagoras unsere Vorstellung von der Länge einer Strecke wider:

Beispiel 2.5 (Euklidische Norm) *Sei $n \in \mathbb{N}$, und sei $\mathcal{V} = \mathbb{K}^n$ der Vektorraum der n-dimensionalen reellen oder komplexen Vektoren. Dann ist*

$$\|\cdot\|_2 \colon \mathcal{V} \to \mathbb{R}_{\geq 0}, \qquad \mathbf{x} \mapsto \left(\sum_{i=1}^{n} x_i^2 \right)^{1/2},$$

eine Norm auf $\mathcal{V}$, die wir als die euklidische Norm *bezeichnen.*

Wenn man einen Vektor geometrisch als Pfeil von dem Nullpunkt zu einem Endpunkt interpretiert, beschreibt $\|\mathbf{x} - \mathbf{y}\|_2$ den Abstand der Endpunkte zweier Vektoren $\mathbf{x}, \mathbf{y} \in \mathcal{V}$.

Normen sind nicht auf endlich-dimensionale Vektorräume beschränkt, sondern ermöglichen es uns auch, beispielsweise die Unterschiede zwischen Funktionen zu messen:

Beispiel 2.6 (Kontinuierliche Maximum-Norm) *Seien $a, b \in \mathbb{R}$ mit $a < b$ gegeben, sei $\mathcal{V} = C[a, b]$ der Raum der stetigen reellwertigen Funktionen auf $[a, b]$. Dann ist*

$$\|\cdot\|_{\infty,[a,b]} \colon \mathcal{V} \to \mathbb{R}_{\geq 0}, \qquad f \mapsto \max\{|f(x)| \ : \ x \in [a, b]\},$$

eine Norm auf V, die wir als die (kontinuierliche) Maximum-Norm *bezeichnen.*

Wenn die Menge X der Eingabedaten Teilmenge eines Vektorraums $\mathcal{V}$ mit der Norm $\|\cdot\|_V$ ist und der Vektor $\mathbf{x} \in X$ die exakten Eingabedaten sowie der Vektor $\widetilde{\mathbf{x}} \in X$ die gestörten Eingabedaten beschreibt, so gibt $\|\mathbf{x} - \widetilde{\mathbf{x}}\|_V$ an, wie „weit entfernt“ die Eingabedaten voneinander sind, ist also ein Maß für die Störung.

Falls die Menge Y der Ausgabedaten ebenfalls Teilmenge eines Vektorraums $\mathcal{W}$ mit der Norm $\|\cdot\|_W$ ist, können wir entsprechend untersuchen, wie weit die exakten Ausgabedaten $\mathbf{y} = \varphi(\mathbf{x})$ von den gestörten Ausgabedaten $\widetilde{\mathbf{y}} = \varphi(\widetilde{\mathbf{x}})$ entfernt sind, wir brauchen also eine Aussage über $\|\mathbf{y} - \widetilde{\mathbf{y}}\|_W$.

Für die Untersuchung einer Berechnung φ sind wir daran interessiert, abzuschätzen, wie sehr Eingabefehler im schlimmsten Fall verstärkt werden, also an der Größe

$$\sup_{\mathbf{x}, \widetilde{\mathbf{x}} \in X} \frac{\|\varphi(\mathbf{x}) - \varphi(\widetilde{\mathbf{x}})\|_W}{\|\mathbf{x} - \widetilde{\mathbf{x}}\|_V}.$$

Es gibt Situationen, in denen sich dieser Fehlerverstärkungsfaktor explizit berechnen lässt, im Allgemeinen werden wir aber nur eine Abschätzung angeben können, die für $\widetilde{\mathbf{x}}$ „in der Nähe“ von $\mathbf{x}$ gilt.

Definition 2.7 (Absolute Fehlerschranke) *Sei $\mathbf{x} \in X$, und sei $\epsilon \in \mathbb{R}_{>0} \cup \{\infty\}$. Eine Zahl $\alpha_{x,\epsilon} \in \mathbb{R}_{\geq 0}$, die*

$$\|\varphi(\mathbf{x}) - \varphi(\widetilde{\mathbf{x}})\|_W \leq \alpha_{x,\epsilon} \|\mathbf{x} - \widetilde{\mathbf{x}}\|_V \qquad \text{für alle } \widetilde{\mathbf{x}} \in X \text{ mit } \|\mathbf{x} - \widetilde{\mathbf{x}}\|_V < \epsilon$$

erfüllt, nennen wir eine absolute Fehlerschranke *für die ϵ-Umgebung von $\mathbf{x}$.*

In praktisch allen Anwendungsfällen sind wir weniger an dem *absoluten* als an dem *relativen* Fehler interessiert: Wenn wir den Abstand der Erde zur Sonne berechnen wollen, ist ein Fehler von 1 000 Kilometern eher vernachlässigbar als bei der Berechnung des Abstands zweier Städte in Europa. In aller Regel wird also eher der relative Fehler

$$\frac{\|\mathbf{x} - \widetilde{\mathbf{x}}\|_V}{\|\mathbf{x}\|_V}$$

von Interesse sein, und zwar sowohl bei der Bewertung der Ein- als auch der Ausgabedaten. Damit beschreibt die Größe

$$\sup_{\mathbf{x},\widetilde{\mathbf{x}} \in X} \frac{\|\varphi(\mathbf{x}) - \varphi(\widetilde{\mathbf{x}})\|_W}{\|\varphi(\mathbf{x})\|_W} \frac{\|\mathbf{x}\|_V}{\|\mathbf{x} - \widetilde{\mathbf{x}}\|_V}$$

die Verstärkung des relativen Fehlers. Auch für den relativen Fehler können wir lokale Fehlerschranken einführen:

Definition 2.8 (Relative Fehlerschranke) *Sei $\mathbf{x} \in X$, und sei $\epsilon \in \mathbb{R}_{>0} \cup \{\infty\}$. Eine Zahl $\kappa_{x,\epsilon} \in \mathbb{R}_{\geq 0}$, die*

$$\frac{\|\varphi(\mathbf{x}) - \varphi(\widetilde{\mathbf{x}})\|_W}{\|\varphi(\mathbf{x})\|_W} \leq \kappa_{x,\epsilon} \frac{\|\mathbf{x} - \widetilde{\mathbf{x}}\|_V}{\|\mathbf{x}\|_V} \qquad \text{für alle } \widetilde{\mathbf{x}} \in X \text{ mit } \|\mathbf{x} - \widetilde{\mathbf{x}}\|_V < \epsilon$$

erfüllt, nennen wir eine relative Fehlerschranke *für die ϵ-Umgebung von $\mathbf{x}$.*

Da wir es in der Praxis immer mit gestörten Eingabedaten zu tun haben, beschreibt die Fehlerverstärkung, welche Genauigkeit des Ergebnisses wir unter optimalen Bedingungen erwarten dürfen. Umgangssprachlich bezeichnen wir ein Problem als *gut konditioniert*, falls die betreffende relative Fehlerschranke einen nicht allzu großen Wert (auch ein Wert von tausend gilt hier oft noch als akzeptabel) aufweist.

Bei der Beurteilung der Kondition eines Problems ist die Wahl der richtigen Norm von entscheidender Bedeutung, wie das folgende Beispiel illustriert.

Beispiel 2.9 (Differenzenquotient) *In vielen Anwendungen ist man daran interessiert, Informationen über Ableitungen einer Funktion zu gewinnen. Ein beliebtes Mittel dabei sind Differenzenquotienten: Nach Definition gilt*

$$f'(x) = \lim_{h \to 0} \frac{f(x+h) - f(x)}{h}$$

für jede in einem Punkt $x \in \mathbb{R}$ differenzierbare Funktion $f : \mathbb{R} \to \mathbb{R}$, also bietet es sich an, ein festes $h \in \mathbb{R}_{>0}$ zu wählen und

$$\varphi(f) = \frac{f(x+h) - f(x)}{h}$$

als Approximation von $f'(x)$ zu verwenden. In diesem Fall ist die Wahl der richtigen Norm entscheidend: Mit der kontinuierlichen Maximum-Norm erhalten wir

$$|\varphi(f) - \varphi(\tilde{f})| = \frac{|f(x+h) - f(x) - \tilde{f}(x+h) + \tilde{f}(x)|}{h} \leq \frac{2\|f - \tilde{f}\|_{\infty,[x,x+h]}}{h},$$

also sind bezüglich dieser Norm nur absolute Fehlerschranken der Form $\alpha_{f,\epsilon} := 2/h$ zu erwarten, und derartige Fehlerschranken sind nur selten hilfreich, weil man in der Regel daran interessiert ist, mit sehr kleinen Werten von h zu arbeiten.

Wenn wir allerdings auf eine andere Norm ausweichen, verbessert sich die Lage deutlich: Nach dem Mittelwertsatz der Differentialrechnung existiert ein $\eta \in [x, x+h]$ mit

$$\varphi(f) = \frac{f(x+h) - f(x)}{h} = f'(\eta),$$

und aus der Linearität der Abbildung φ folgt sofort

$$|\varphi(f) - \varphi(\tilde{f})| = |\varphi(f - \tilde{f})| \leq \|f' - \tilde{f}'\|_{\infty,[x,x+h]}.$$

Wenn wir den Fehler also in der Maximumnorm der Ableitung *messen, ist die absolute Fehlerschranke völlig unabhängig von h. Insofern sind in diesem Fall Normen angemessen, mit denen sich auch die Ableitung beschränken lässt.*

2.2 Maschinenzahlen

Bei der Umsetzung einer Rechenvorschrift auf einem Computer kommt neben den Messfehlern in den Eingabedaten noch eine weitere Fehlerquelle hinzu: Da jeder real existierende Computer nur über eine endliche Speichermenge verfügt, kann er nicht unendlich viele unterschiedliche Zahlen darstellen.

Bei ganzen Zahlen lässt sich dieses Problem umgehen, indem man die Anzahl der Stellen beschränkt.

Definition 2.10 (Darstellung natürlicher Zahlen) *Sei $m \in \mathbb{N}$. Eine m-stellige Zahl $n \in \mathbb{N}_0$ wird als Summe*

$$n = \sum_{i=0}^{m-1} d_i \, b^i$$

mit einer Basis $b \in \mathbb{N}_{\geq 2}$ und Ziffern $d_0, \ldots, d_{m-1} \in [0 : b-1]$ dargestellt.

Die Ziffern lassen sich aus n durch wiederholte Division durch die Basis b gewinnen.

In der Praxis verwendet man in der Regel eine binäre Darstellung mit $b = 2$, üblich sind $m = 16, 32$ oder 64 Ziffern, in diesem Kontext *Bits* genannt, mit denen sich die Zahlen von 0 bis 65 535, 4 294 967 295 und 18 446 744 073 709 551 615 darstellen lassen.

Bemerkung 2.11 (Darstellung ganzer Zahlen) *Eine Zahl $z \in \mathbb{Z}$ kann als Produkt einer natürlichen Zahl mit einem Vorzeichen $\sigma \in \{-1, 1\}$ dargestellt werden. Diese Darstellung hätte allerdings den Nachteil, dass die Null doppelt auftaucht.*

Es gibt modifizierte Darstellungen wie die Zweierkomplementdarstellung*, die diesen Nachteil vermeiden.*

Brüche lassen sich relativ einfach darstellen, indem Zähler und Nenner als ganze Zahlen gespeichert werden. Wesentlich schwieriger wird es bei reellen Zahlen: Sei $x \in \mathbb{R}_{>0}$ gegeben. Indem wir

$$e := \min\{\alpha \in \mathbb{Z} \ : \ b^\alpha > x\} \in \mathbb{Z}$$

fixieren, erhalten wir

$$1/b \leq x b^{-e} < 1,$$

müssen also nur noch eine Zahl aus dem halboffenen Intervall $[1/b, 1)$ darstellen. Dazu verwenden wir wieder Ziffern, allerdings müssen wir diesmal unendlich viele Ziffern zulassen: Es gilt

$$\begin{aligned} x\, b^{-e} &= \sum_{i=1}^{\infty} d_i\, b^{-i}, \\ x &= b^e \sum_{i=1}^{\infty} d_i\, b^{-i} \end{aligned} \tag{2.1}$$

für geeignet gewählte $d_1, d_2, \ldots \in [0 : b-1]$. Aus $x\, b^{-e} \geq 1/b$ folgt insbesondere $d_1 > 0$. Die Darstellung (2.1) mit $d_1 > 0$ bezeichnen wir als *normalisierte Gleitkommadarstellung* der Zahl x.

Eine Darstellung von x im Computer erhalten wir wieder, indem wir uns auf endlich viele Ziffern beschränken:

Definition 2.12 (Maschinenzahl) *Seien $b \in \mathbb{N}_{\geq 2}$ und $m \in \mathbb{N}$ gegeben. Für alle $e \in \mathbb{Z}$, $\sigma \in \{-1, 1\}$ und $d_1, \ldots, d_m \in [0 : b-1]$ mit $d_1 \neq 0$ bezeichnen wir*

$$x := \sigma b^e \sum_{i=1}^{m} d_i\, b^{-i}$$

als (normalisierte) Maschinenzahl *mit dem* Exponenten *e, der* Mantisse *$(d_1, \ldots, d_m)$ und dem* Vorzeichen *σ.*

Die Null wird als Sonderfall hinzugenommen und ebenfalls als (nicht normalisierte) Maschinenzahl *bezeichnet.*

In diesem Kontext beschreiben die Basis *b und die* Mantissenlänge *m alle in dieser Form darstellbaren Maschinenzahlen. Die Menge aller Maschinenzahlen zu den erwähnten Parametern bezeichnen wir mit $\mathfrak{M}(b, m) \subseteq \mathbb{R}$.*

In der Praxis wird e in der Regel nur aus einer endlichen Teilmenge der Menge $\mathbb{Z}$ der ganzen Zahlen gewählt, und es werden spezielle Darstellungen für die Null und ähnliches hinzugenommen, aber diese Aspekte werden wir im Folgenden vernachlässigen.

Bemerkung 2.13 (IEEE-754-Standard) *In aktuellen Computern werden in der Regel Maschinenzahlen nach dem* IEEE-754-Standard *verwendet, vor allem Zahlen mit*

- einfacher Genauigkeit*, also mit $b = 2$, $m = 24$ und $e \in [-125 : 128]$, und mit*
- doppelter Genauigkeit*, also mit $b = 2$, $m = 53$ und $e \in [-1021 : 1024]$.*

Eine Maschinenzahl einfacher Genauigkeit wird durch 32 *Bits dargestellt: Da dank der Normalisierung immer $d_1 = 1$ gilt, brauchen nur die Ziffern $d_{24}, \dots, d_2$ abgespeichert zu werden, und das geschieht in den ersten* 23 *Bits. Es folgen* 8 *Bits für den Exponenten*[1]*, die als eine Zahl $\hat{e} \in [0 : 255]$ interpretiert werden, aus der sich der Exponent durch $e = \hat{e} - 126$ berechnen lässt. Dabei sind $\hat{e} = 0$ und $\hat{e} = 255$ für Sonderfälle wie die Darstellung der Null oder unendlicher Werte vorgesehen. Das letzte Bit ist schließlich für das Vorzeichen vorgesehen.*

Für eine Maschinenzahl doppelter Genauigkeit werden 64 *Bits verwendet:* 52 *Bits für die Ziffern $d_{53}, \dots, d_2$,* 11 *Bits für den Exponenten, interpretiert als Zahl $\hat{e} \in [0 : 2047]$, aus der der Exponent durch $e = \hat{e} - 1022$ hervor geht, wobei $\hat{e} = 0$ und $\hat{e} = 2047$ wieder für Sonderfälle vorgesehen sind, und schließlich wieder das letzte Bit für das Vorzeichen.*

Um die Maschinenzahlen als praktischen Ersatz für reelle Zahlen im Computer verwenden zu können, müssen wir dazu in der Lage sein, eine beliebige reelle Zahl durch eine Maschinenzahl näherungsweise darzustellen, also zu *approximieren*. Der einfachste Zugang besteht darin, direkt die normalisierte Gleitkommadarstellung zu verwenden und alle Ziffern ab der $(m+1)$-ten wegfallen zu lassen.

Definition 2.14 (Abrunden) *Seien $b \in \mathbb{N}_{\geq 2}$ und $m \in \mathbb{N}$ gegeben. Wir definieren die Abbildung $\mathrm{fl}_0 \colon \mathbb{R} \to \mathfrak{M}(b, m)$ in der folgenden Weise: Für jedes $x \in \mathbb{R} \setminus \{0\}$ existieren $\sigma \in \{-1, 1\}$, $e \in \mathbb{Z}$ und $(d_i)_{i \in \mathbb{N}}$ mit $d_1 \neq 0$ und*

$$x = \sigma b^e \sum_{i=1}^{\infty} d_i b^{-i}. \tag{2.2}$$

Diese Zahl wird von fl_0 auf

$$\mathrm{fl}_0(x) := \sigma b^e \sum_{i=1}^{m} d_i b^{-i}$$

abgebildet: Die Zahl x wird auf die im Betrag nächstkleinere Maschinenzahl abgerundet *(im IEEE-Standard „round towards zero“ genannt).*

Für $x = 0$ dagegen setzen wir einfach $\mathrm{fl}_0(0) = 0$.

[1] Anders als in diesem Skript werden im IEEE-754-Standard die Exponenten um eins kleiner gewählt, um eine Eins *vor* dem Komma zu erhalten.

Durch die Kombination des exponentiellen Terms b^e mit der Normalisierungsbedingung $d_1 \neq 0$ erhalten wir die Möglichkeit, die Approximation einer reellen Zahl durch eine Maschinenzahl in *relativer* statt absoluter Genauigkeit zu erreichen:

Lemma 2.15 (Abrundungsfehler) *Für alle $x \in \mathbb{R} \setminus \{0\}$ gilt*

$$\frac{|x - \mathrm{fl}_0(x)|}{|x|} \leq b^{1-m}.$$

Beweis. Seien $\sigma \in \{-1, 1\}$, $m \in \mathbb{N}$, $e \in \mathbb{Z}$ und eine Folge $(d_i)_{i=1}^{\infty}$ in $[0 : b-1]$ mit $d_1 \neq 0$ gegeben, die

$$x = \sigma b^e \sum_{i=1}^{\infty} d_i b^{-i}$$

erfüllen, und sei

$$\tilde{x} = \mathrm{fl}_0(x) = \sigma b^e \sum_{i=1}^{m} d_i b^{-i}.$$

Aus der Definition folgt bereits

$$\begin{aligned} |x - \tilde{x}| = b^e \sum_{i=m+1}^{\infty} d_i b^{-i} &\leq b^e \sum_{i=m+1}^{\infty} (b-1) b^{-i} = b^e \left(\sum_{i=m+1}^{\infty} b^{1-i} - \sum_{i=m+1}^{\infty} b^{-i} \right) \\ &= b^e \left(\sum_{i=m}^{\infty} b^{-i} - \sum_{i=m+1}^{\infty} b^{-i} \right) = b^e b^{-m} = b^{e-m}. \end{aligned}$$

Mit der Normierungsbedingung $d_1 \neq 0$ erhalten wir $|x| \geq b^{e-1}$ und damit bereits die Fehlerschranke

$$\frac{|x - \tilde{x}|}{|x|} \leq \frac{b^{e-m}}{b^{e-1}} = b^{1-m}.$$

■

Falls die Basis b eine gerade Zahl ist, können wir durch *kaufmännisches Runden* die Genauigkeit verbessern: Dabei wählen wir diejenige Maschinenzahl, die der darzustellenden Zahl am nächsten liegt.

Definition 2.16 (Runden) *Seien eine gerade Zahl $b \in \mathbb{N}_{\geq 2}$ und $m \in \mathbb{N}$ gegeben. Wir definieren die Abbildung $\mathrm{fl}\colon \mathbb{R} \to \mathfrak{M}(b, m)$ in der folgenden Weise: Für jedes $x \in \mathbb{R} \setminus \{0\}$ existieren $\sigma \in \{-1, 1\}$, $e \in \mathbb{Z}$ und $(d_i)_{i \in \mathbb{N}}$ mit*

$$x = \sigma b^e \sum_{i=1}^{\infty} d_i b^{-i}.$$

Diese Zahl wird von fl *auf*

$$\mathrm{fl}(x) := \sigma b^e \begin{cases} \sum_{i=1}^{m} d_i b^{-i} & \text{falls } d_{m+1} < b/2, \\ \sum_{i=1}^{m} d_i b^{-i} + b^{-m} & \text{ansonsten} \end{cases}$$

abgebildet: Die Zahl x wird auf die nächste Maschinenzahl gerundet *(im IEEE-Standard „round to nearest, ties away from zero“ genannt).*

Falls $d_{m+1} \geq b/2$ gilt, wird eins zu d_m addiert. Falls nun $d_m = b-1$ gilt, kommt es zu einem Übertrag, der zu d_{m-1} addiert wird. Falls $d_1 = d_2 = \ldots = d_m = b-1$ gelten, pflanzt sich der Übertrag bis in die erste Stelle fort, so dass wir insgesamt

$$\begin{aligned}\sum_{i=1}^{m} d_i b^{-i} + b^{-m} &= \sum_{i=1}^{m} (b-1) b^{-i} + b^{-m} = \sum_{i=1}^{m} b^{1-i} - \sum_{i=1}^{m} b^{-i} + b^{-m} \\ &= \sum_{i=0}^{m-1} b^{-i} - \sum_{i=1}^{m} b^{-i} + b^{-m} = 1 - b^{-m} + b^{-m} = 1\end{aligned}$$

erhalten. Also vergrößern wir einfach e um eins und haben auch in diesem Fall eine korrekte Maschinenzahl.

Für die Null verwenden wir wieder $\mathrm{fl}(0) = 0$.

Im Vergleich mit dem einfachen Abrunden können wir durch das Runden zur nächstgelegenen Maschinenzahl den Approximationsfehler halbieren:

Lemma 2.17 (Rundungsfehler) *Sei $b \in \mathbb{N}_{\geq 2}$ eine gerade Zahl. Für alle $x \in \mathbb{R} \setminus \{0\}$ gilt*

$$\frac{|x - \mathrm{fl}(x)|}{|x|} \leq \frac{b^{1-m}}{2}.$$

Beweis. Seien $\sigma \in \{-1, 1\}$, $m \in \mathbb{N}$, $e \in \mathbb{Z}$ und eine Folge $(d_i)_{i=1}^{\infty}$ in $[0, b-1]$ mit $d_1 \neq 0$ gegeben, die

$$x = \sigma b^e \sum_{i=1}^{\infty} d_i b^{-i}$$

erfüllen, und sei

$$\tilde{x} = \mathrm{fl}(x) = \sigma b^e \begin{cases} \sum_{i=1}^{m} d_i b^{-i} & \text{falls } d_{m+1} < b/2, \\ \sum_{i=1}^{m} d_i b^{-i} + b^{-m} & \text{ansonsten.} \end{cases}$$

Da das Vorzeichen σ für die Beträge $|x - \tilde{x}|$ und $|x|$ keine Rolle spielt, behandeln wir im Folgenden ohne Beschränkung der Allgemeinheit nur den Fall $\sigma = 1$.

Falls $d_{m+1} < b/2$ gilt, ist $\tilde{x} \leq x$. Aus der Geradzahligkeit von b folgt $d_{m+1} + 1 \leq b/2$ und wir erhalten

$$\begin{aligned}x - \tilde{x} = b^e \sum_{i=m+1}^{\infty} d_i b^{-i} &= b^e \left(d_{m+1} b^{-m-1} + \sum_{i=m+2}^{\infty} d_i b^{-i} \right) \\ &\leq b^e \left(d_{m+1} b^{-m-1} + \sum_{i=m+2}^{\infty} (b-1) b^{-i} \right) \\ &= b^e \left(d_{m+1} b^{-m-1} + \sum_{i=m+2}^{\infty} b^{1-i} - \sum_{i=m+2}^{\infty} b^{-i} \right)\end{aligned}$$

$$= b^e \left(d_{m+1} b^{-m-1} + \sum_{i=m+1}^{\infty} b^{-i} - \sum_{i=m+2}^{\infty} b^{-i} \right)$$
$$= b^e (d_{m+1} + 1) b^{-m-1} \leq b^e \frac{b}{2} b^{-m-1} = \frac{b^{e-m}}{2}.$$

Falls dagegen $d_{m+1} \geq b/2$ gilt, haben wir $\tilde{x} \geq x$ und gelangen zu der Abschätzung

$$\tilde{x} - x = b^e \left(b^{-m} - \sum_{i=m+1}^{\infty} d_i b^{-i} \right) \leq b^e \left(b^{-m} - d_{m+1} b^{-m-1} \right) \leq b^e \left(b^{-m} - \frac{b}{2} b^{-m-1} \right)$$
$$= b^e \left(b^{-m} - \frac{b^{-m}}{2} \right) = \frac{b^{e-m}}{2}.$$

Aus der Kombination beider Abschätzungen folgt in Kombination mit $|x| \geq b^{e-1}$ bereits $|x - \tilde{x}|/|x| \leq b^{1-m}/2$, also die gewünschte Schranke für den Rundungsfehler. ∎

Bemerkung 2.18 (Rundungsfehler für IEEE-754-Zahlen) *Mit dem Lemma 2.17 können wir den maximalen relativen Rundungsfehler für IEEE-754-Zahlen berechnen: für einfache Genauigkeit erhalten wir*

$$\frac{2^{1-24}}{2} = \frac{2^{-23}}{2} = 2^{-24} \approx 5,96 \times 10^{-8},$$

und für doppelte Genauigkeit

$$\frac{2^{1-53}}{2} = \frac{2^{-52}}{2} = 2^{-53} \approx 1,11 \times 10^{-16}.$$

Folgerung 2.19 (Maschinengenauigkeit) *Seien eine Mantissenlänge $m \in \mathbb{N}$ und eine geradzahlige Basis $b \in \mathbb{N}_{\geq 2}$ gegeben. Die Konstante*

$$\epsilon_{fl} := \frac{b^{1-m}}{2}$$

bezeichnen wir als die zu $\mathfrak{M}(b, m)$ gehörende Maschinengenauigkeit.

Zu jedem $x \in \mathbb{R}$ existiert ein $\epsilon \in \mathbb{R}$ mit $|\epsilon| \leq \epsilon_{fl}$ und

$$\mathrm{fl}(x) = (1 + \epsilon) x. \tag{2.3}$$

Beweis. Sei $x \in \mathbb{R}$. Falls $x = 0$ gilt, haben wir $\mathrm{fl}(x) = x$ und können $\epsilon = 0$ verwenden.

Ansonsten setzen wir $\tilde{x} = \mathrm{fl}(x)$ sowie

$$\epsilon := \frac{\tilde{x} - x}{x}$$

und erhalten

$$(1 + \epsilon) x = \left(1 + \frac{\tilde{x} - x}{x} \right) x = \frac{x + \tilde{x} - x}{x} x = \frac{\tilde{x}}{x} x = \tilde{x}$$

sowie mit Lemma 2.17 auch

$$|\epsilon| = \frac{|\tilde{x} - x|}{|x|} = \frac{|\operatorname{fl}(x) - x|}{|x|} \leq \frac{b^{1-m}}{2} = \epsilon_{\text{fl}}.$$

■

Die Darstellung (2.3) des relativen Fehlers bietet den Vorteil, im Gegensatz zu der Aussage des Lemmas 2.17 auch für $x = 0$ gültig zu sein. Deshalb wird ihr bei Untersuchungen der Fehlerfortpflanzung häufig der Vorzug gegenüber der Darstellung durch einen Quotienten gegeben.

Bemerkung 2.20 (Maschinenzahlen in C99) *Die Mathematik-Bibliothek des C99-Standards bietet uns die Möglichkeit, auf die einzelnen Komponenten von Maschinenzahlen zuzugreifen: Die Funktionen* `frexp` *und* `frexpf` *zerlegen eine Maschinenzahl in ihren Exponenten und den Mantissen-Anteil in* $[1/2, 1)$*, während die Funktionen* `ldexp` *und* `ldexpf` *eine Maschinenzahl mit einer Zweierpotenz multiplizieren, indem schlicht der Exponent vergrößert oder verkleinert wird.*

Da der Computer nur mit Maschinenzahlen rechnen kann, müssen insbesondere die üblichen Grundrechenarten durch gerundete Approximationen ersetzt werden, die wir mit $\oplus$, $\ominus$, $\odot$ und $\oslash$ bezeichnen und wie folgt definieren:

$$\begin{aligned} x \oplus y &= \operatorname{fl}(x + y), & x \ominus y &= \operatorname{fl}(x - y), \\ x \odot y &= \operatorname{fl}(xy), & x \oslash y &= \operatorname{fl}(x/y) \qquad \text{für alle } x, y \in \mathbb{R}. \end{aligned}$$

Die Verwendung gerundeter Zahlen hat Folgen, die wir nicht vergessen sollten, wenn wir mit einem Computer reelle Zahlen behandeln. Beispielsweise gilt das Assoziativgesetz nicht mehr: Wenn wir Maschinenzahlen zur Basis $b = 10$ mit der Mantissenlänge $m = 2$ verwenden und runden, gilt

$$\begin{aligned} (0.50 \oplus 0.50) \oplus 10.0 &= 1.0 \oplus 10.0 = 11.0, \\ 0.50 \oplus (0.50 \oplus 10.0) &= 0.50 \oplus 11.0 = 12.0, \end{aligned}$$

es spielt also eine Rolle, in welche Reihenfolge wir Additionen ausführen. Dieses Beispiel lässt sich ausbauen: Wenn wir zwanzigmal zu 10.0 die Zahl 0.5 (gerundet) hinzuaddieren, wird jedesmal aufgerundet und wir erhalten als Ergebnis 30. Wenn wir dagegen zwanzigmal 0.5 (gerundet) aufaddieren und erst im letzten Schritt 10 hinzuaddieren, erhalten wir das korrekte Ergebnis 20.

Durch das Rechnen mit gerundeten Zahlen kann es auch passieren, dass das Ergebnis einer Berechnung eine wesentlich geringere Genauigkeit als ϵ_{fl} erzielt. Wenn wir beispielsweise die Differenz

$$0.1347 - 0.1326 = 0.0021$$

berechnen, indem wir zunächst die beiden Zahlen auf der linken Seite in Maschinenzahlen umwandeln und dann die Differenz bilden, erhalten wir

$$0.13 - 0.13 = 0,$$

obwohl das korrekte Ergebnis 0.0021 sich eigentlich sogar exakt durch eine Maschinenzahl darstellen ließe. Dieser Effekt entsteht dadurch, dass sich die Ziffern der gerundeten Zahlen bei der Berechnung der Differenz gegenseitig wegheben und trägt deshalb den Namen *Auslöschung*. Er tritt in der Praxis vor allem bei der Berechnung von Differenzen auf und kann die Genauigkeit einer Berechnung mit Maschinenzahlen erheblich beeinträchtigen.

Selbstverständlich lässt sich der Auslöschungseffekt auch theoretisch untersuchen: Wenn wir mit $\tilde{x} := \mathrm{fl}(x)$ und $\tilde{y} := \mathrm{fl}(y)$ die Approximationen von $x, y \in \mathbb{R}$ durch Maschinenzahlen bezeichnen, finden wir mit Folgerung 2.19 zwei Zahlen $\epsilon_x, \epsilon_y \in \mathbb{R}$ mit

$$\tilde{x} = (1+\epsilon_x)x, \qquad \tilde{y} = (1+\epsilon_y)y$$

und $|\epsilon_x|, |\epsilon_y| \leq \epsilon_{\mathrm{fl}}$. Nach Definition der Addition $\oplus$ für Maschinenzahlen finden wir $\epsilon_+ \in \mathbb{R}$ mit $|\epsilon_+| \leq \epsilon_{\mathrm{fl}}$ und

$$\tilde{x} \oplus \tilde{y} = \mathrm{fl}(\tilde{x} + \tilde{y}) = (1+\epsilon_+)(\tilde{x} + \tilde{y}).$$

Insgesamt erhalten wir

$$\begin{aligned}\tilde{x} \oplus \tilde{y} &= (1+\epsilon_+)(\tilde{x} + \tilde{y}) = (1+\epsilon_+)((1+\epsilon_x)x + (1+\epsilon_y)y) \\ &= (1+\epsilon_+)(x + y + \epsilon_x x + \epsilon_y y).\end{aligned}$$

Im Fall $x + y = 0$ gilt (vgl. den Beweis von Folgerung 2.19) $\epsilon_x = \epsilon_y$, so dass wir das exakte Ergebnis finden.

Anderenfalls gelangen wir zu

$$\begin{aligned}\tilde{x} \oplus \tilde{y} &= (1+\epsilon_+)(x + y + \epsilon_x x + \epsilon_y y) = (1+\epsilon_+)\frac{x + y + \epsilon_x x + \epsilon_y y}{x+y}(x+y) \\ &= (1+\epsilon_+)\left(1 + \frac{\epsilon_x x + \epsilon_y y}{x+y}\right)(x+y) = (1+\delta)(x+y)\end{aligned}$$

mit der Hilfsgröße

$$\delta := \epsilon_+ + (1+\epsilon_+)\frac{\epsilon_x x + \epsilon_y y}{x+y},$$

die analog zu Folgerung 2.19 den relativen Fehler der Summe beschreibt.

Falls x und y dasselbe Vorzeichen aufweisen, gilt $|x+y| = |x| + |y|$ und damit

$$\begin{aligned}|\delta| &\leq |\epsilon_+| + (1+|\epsilon_+|)\frac{|\epsilon_x|\,|x| + |\epsilon_y|\,|y|}{|x+y|} = |\epsilon_+| + (1+|\epsilon_+|)\frac{|\epsilon_x|\,|x| + |\epsilon_y|\,|y|}{|x| + |y|} \\ &\leq \epsilon_{\mathrm{fl}} + (1+\epsilon_{\mathrm{fl}})\frac{\epsilon_{\mathrm{fl}}(|x| + |y|)}{|x| + |y|} = 2\epsilon_{\mathrm{fl}} + \epsilon_{\mathrm{fl}}^2.\end{aligned}$$

In diesem Fall weicht $\tilde{x} \oplus \tilde{y}$ also nur geringfügig von dem exakten Ergebnis $x + y$ ab.

Sehr viel ungünstiger ist die Situation, falls x und y unterschiedliche Vorzeichen besitzen, denn in diesem Fall kann

$$\left|\frac{\epsilon_x x + \epsilon_y y}{x+y}\right| = \frac{|\epsilon_x x + \epsilon_y y|}{|x+y|}$$

sehr groß werden, falls $|x+y| \ll |x| + |y|$ gilt. Das ist gerade der Effekt der Auslöschung.

2.3 Algorithmen

Wir beschreiben Berechnungen auf einem Computer, indem wir angeben, welche auf dem Computer verfügbaren Operationen er in welcher Reihenfolge mit welchen Daten ausführen soll. Für eine Darstellung der Berechnung, die so detailliert ist, dass ein Computer sie mit nur geringfügiger Hilfe eines Programmierers ausführen kann, ist die Bezeichnung *Algorithmus* üblich.

Da eine formal präzise Darstellung eines Algorithmus hier zu weit führen würde, soll lediglich kurz erklärt werden, wie wir einen Algorithmus notieren und wie er zu interpretieren ist. Die Umsetzung in die Form eines *Programms* in einer bestimmten Programmiersprache auf einem gegebenen Computer bezeichnet man als *Implementierung*.

Die Daten werden in der Form von *Variablen* repräsentiert: Jede Variable bezeichnet einen Speicherplatz des Computers, der beispielsweise eine ganze Zahl oder einen Buchstabe enthalten kann. Der *Typ* einer Variablen legt fest, wie der Speicherplatz interpretiert wird. Den aktuellen Inhalt des Speicherplatzes bezeichnen wir als den *Wert* der Variablen, und er kann sich im Laufe eines Algorithmus ändern.

In Formeln verwenden wir den Namen einer Variablen, um auf ihren Wert zu verweisen, beispielsweise bedeutet

$$x + y$$

in diesem Kontext, dass die Zahl, die sich im Speicherplatz der Variablen x befindet, und die Zahl, die sich unter y findet, addiert werden sollen. Falls x und y ganze Zahlen sind, soll die Addition exakt ausgeführt werden (Über- und Unterläufe vernachlässigen wir), falls x und y Maschinenzahlen sind, wird die gerundete Addition $\oplus$ verwendet.

Um Ergebnisse speichern zu können, sehen wir vor, dass Variablen ihren Wert im Laufe der Ausführung der Berechnung ändern dürfen. Diese Veränderung bewirkt der *Zuweisungsoperator* $\leftarrow$ in der folgenden Weise:

$$z \leftarrow x + y$$

bedeutet, dass die Summe von x und y in der oben beschriebenen Weise berechnet und das Ergebnis der neue Wert der Variablen z werden soll. Dabei geht deren vorheriger Wert verloren.

Bei der Durchführung einfacher Rechnungen verwenden wir die allgemein übliche mathematische Notation mit den gewohnten Regeln für die Einsparung von Klammern:

$$y \leftarrow 2z + x$$

bedeutet, dass der doppelte Wert der Variablen z zu x addiert und das Ergebnis unter y gespeichert wird.

Zur Abkürzung legen wir die Konvention fest, dass der Zuweisungsoperator immer als letzter ausgeführt wird, so dass

$$z \leftarrow z + 1$$

bedeutet, dass nach der Ausführung des Befehls der Wert der Variablen z um eins höher als vorher ist.

Folgen von Befehlen, die in einer bestimmten Reihenfolge ausgeführt werden sollen,

schreiben wir einfach untereinander:

$x \leftarrow y$
$y \leftarrow 2z$

bedeutet, dass zunächst x den Wert von y annehmen soll, und anschließend y den doppelten Wert von z.

Um Platz zu sparen, können Befehlsfolgen auch durch Semikola getrennt in dieselbe Zeile geschrieben werden:

$x \leftarrow y; \quad y \leftarrow 2z$

entspricht der obigen Berechnung, benötigt aber nur eine Zeile.

Ein wichtiges Hilfsmittel bei der Formulierung von Berechnungen sind Fallunterscheidungen. Wir formulieren sie wie folgt:

if $x < 0$ **then**
  $x \leftarrow -x$
end if

bedeutet, dass die Anweisungen zwischen „then“ und „end“ nur ausgeführt werden sollen, falls die Bedingung zwischen „if“ und „then“ wahr ist. Im Beispiel bewirkt das gerade, dass der Wert einer reellen Zahl x durch seinen Betrag ersetzt wird: Das Vorzeichen wird nur geändert, falls es ursprünglich negativ war.

Falls auch Anweisungen ausgeführt werden sollen, falls die Bedingung nicht erfüllt ist, verwenden wir die folgende Konstruktion:

if $x < y$ **then**
  $y \leftarrow y - x$
else
  $x \leftarrow x - y$
end if

In diesem Beispiel wird jeweils der kleinere der beiden Werte von dem größeren subtrahiert, ein wichtiger Schritt des Euklidischen Algorithmus für die Bestimmung des größten gemeinsamen Teilers zweier Zahlen.

Zur Formulierung der Bedingungen lassen wir beliebige mathematische Ausdrücke zu, bei denen in der bereits erwähnten Weise die Werte der Variablen eingesetzt werden. In bestimmten Situationen werden wir sogar umgangssprachliche Bedingungen wie „ϵ zu groß“ oder ähnliches zulassen, falls die mathematisch präzise Formulierung zu unhandlich sein sollte oder weiterer Diskussionen bedarf. In der Regel werden wir dann an geeigneter Stelle festlegen, wie die Bedingung mathematisch korrekt umzusetzen ist.

Sehr häufig wird eine Berechnung aus der wiederholten Durchführung einer Folge von Rechenschritten bestehen. Derartige Berechnungen schreiben wir als *Schleifen*, etwa in der Form einer *while-Schleife*:

$x \leftarrow 1$
while $y > 0$ **do**
  $x \leftarrow 2x; \quad y \leftarrow y - 1$
end while

Bei dieser Schleifenkonstruktion werden die Befehle zwischen „do“ und „end“, der so-

genannte *Rumpf* der Schleife, so lange ausgeführt, bis die Bedingung $y > 0$ nicht mehr gilt. Falls sie bereits bei Erreichen von „while“ nicht galt, wird die Schleife gar nicht ausgeführt. Das obige Beispiel gibt der Variablen x den Wert 2^y, falls der Wert von y eine nicht-negative ganze Zahl ist: Für $y = 0$ wird der Rumpf der Schleife nicht ausgeführt, und x behält den Wert 1. Für $y = 1$ wird der Rumpf der Schleife einmal ausgeführt, danach besitzt y den Wert 0 und x den Wert 2, womit die Schleife endet. Entsprechendes gilt für größere Werte von y.

Eine besonders einfache Variante der while-Schleife ist die *for-Schleife*, bei der der Rumpf so oft ausgeführt wird, wie es eine von Anfang an bekannte Zahl angibt:

$x \leftarrow 0$
for $i = 1, \ldots, n$ **do**
  $x \leftarrow x + i$
end for

berechnet die Summe der ersten n natürlichen Zahlen. Die Notation besagt, dass zunächst der Rumpf für $i = 1$ ausgeführt wird, dann für $i = 2$, und dass in dieser Weise fortgefahren wird, bis $i = n$ erreicht ist.

Falls die Reihenfolge keine Rolle spielt, in der der Rumpf für die verschiedenen Werte von i ausgeführt wird, verwenden wir die Notation

for $i \in [1 : n]$ **do**
  $x_i \leftarrow x_i + 2y_i$
end for

um anzudeuten, dass sich die Reihenfolge beliebig wählen lässt, um beispielsweise parallel arbeitende Rechenwerke innerhalb des Computers gleichmäßig auszulasten. Im Beispiel wird die Summe $\mathbf{x}+2\mathbf{y}$ zweier Vektoren $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$ berechnet, und dabei sind die einzelnen Komponenten völlig unabhängig voneinander.

Häufig wird ein komplizierterer Algorithmus sich aus einfacheren Teilalgorithmen zusammensetzen. Um diese Teile separat untersuchen zu können, hat sich das Konzept der *Prozedur* als nützlich erwiesen:

procedure vorzeichen(x, **var** σ)
if $x < 0$ **then**
  $\sigma \leftarrow -1$
else
  $\sigma \leftarrow 1$
end if

Diese Notation besagt, dass eine Prozedur des Namens „vorzeichen“ definiert werden soll, die als Argumente x und σ erhält. x darf nicht verändert werden, σ dagegen schon (es ist **var**iabel).

Wenn wir eine Prozedur verwenden wollen, schreiben wir ihren Namen in den Algorithmus, gefolgt von den Ausdrücken, die für die einzelnen Argumente eingesetzt werden sollen:

vorzeichen(-5, γ)

An allen Orten, an denen in der Definition der Prozedur x auftrat, wird nun -5 eingesetzt, und an den Orten, an denen σ auftrat, wird γ verwendet.

Eine Besonderheit sind *rekursive* Prozeduren, die sich selbst verwenden können. Als Beispiel verwenden wir eine Variante des Euklidischen Algorithmus für die Berechnung des größten gemeinsamen Teilers zweier Zahlen:

procedure ggt(x, y, **var** g)
if $x = y$ **then**
  $g \leftarrow x$
else if $x < y$ **then**
  ggt(x, $y - x$, g)
else
  ggt($x - y$, y, g)
end if

Zur Definition der Prozedur „ggt“ wird die Prozedur „ggt“ verwendet. Das ist kein Problem, solange dafür gesorgt ist, dass die Prozedur sich nicht unendlich lange selbst aufruft. In diesem Fall ist das dadurch sicher gestellt, dass jeder rekursive Aufruf entweder den Wert von x oder den von y reduziert, denn da es in der Menge der natürlichen Zahlen keine unendlichen strikt fallenden Folgen gibt, muss irgendwann der Fall $x = y$ eintreten.

Rekursive Prozeduren lassen sich in vielen Situationen einsetzen, um einen Algorithmus elegant zu formulieren, beispielsweise ergeben sie sich häufig unmittelbar aus Induktionsbeweisen: Die Anwendung der Induktionsvoraussetzung im Beweis wird zu einem rekursiven Prozeduraufruf im Programm.

Bei der praktischen Umsetzung eines Algorithmus in Form eines Programms ist zu beachten, dass man die korrekte Syntax der verwendeten Programmiersprache verwendet. Zuweisungen, Fallunterscheidungen, while- und for-Schleifen sowie Prozeduren finden sich in den meisten gängigen Programmiersprachen, so dass sie sich in der Regel direkt umsetzen lassen.

Hinzu kommen in der Regel *Variablendeklarationen*, mit denen der Programmierer dem Computer mitteilt, welche Variablen verwendet werden sollen und welche Werte diese Variablen annehmen dürfen. Beispielsweise legt `int x` in der Programmiersprache C fest, dass x eine Variable ist, die ganze Zahlen als Werte annehmen kann, während `float y` eine Variable y für Maschinenzahlen einfacher Genauigkeit deklariert.

2.4 Stabilität

Da ein Computer nicht mit „echten“ reellen Zahlen, sondern nur mit Maschinenzahlen arbeiten kann, werden die meisten Berechnungen, die wir mit Hilfe eines Computers durchführen, lediglich Näherungen der eigentlich beabsichtigten Berechnungen sein. Mathematisch beschreiben wir diesen Sachverhalt dadurch, dass neben der gewünschten Berechnung φ auch die vom Computer durchgeführte Berechnung $\tilde{\varphi}$ definiert wird. Wie wir bereits gesehen haben, hängt $\tilde{\varphi}$ von der Reihenfolge ab, in der die elementaren Operationen einer Berechnung durchgeführt werden.

Ein wichtiger Teil der Arbeit eines numerischen Mathematikers besteht darin, Berechnungen so zu arrangieren, also so in einen auf dem Computer durchführbaren Algorithmus umzusetzen, dass $\tilde{\varphi}$ möglichst nahe an φ liegt. Einen Algorithmus zur Durchführung einer bestimmten Berechnung bezeichnen wir als *stabil*, falls der von $\tilde{\varphi}$ eingeführte Fehler vergleichbar mit dem infolge der Konditionszahl zu erwartenden Fehler ist. Um die Stabilität eines numerischen Verfahrens zu messen, sind verschiedene Techniken üblich, auf die hier nicht näher eingegangen werden soll. Wir beschränken uns auf die bereits umrissene anschauliche Interpretation von Stabilität.

Als Beispiel untersuchen wir das Lösen der quadratischen Gleichung

$$x^2 - 2px - q = 0$$

mit $p^2 \geq q$. Die Lösungen ergeben sich per quadratischer Ergänzung: Aus

$$0 = x^2 - 2px - q = x^2 - 2px + p^2 - p^2 - q = (x-p)^2 - \left(p^2 + q\right)$$

erhalten wir

$$x_1 = p + \sqrt{p^2 + q}, \qquad\qquad x_2 = p - \sqrt{p^2 + q}.$$

Diese Formeln können wir nun direkt in einen ersten Algorithmus zur Berechnung der Nullstellen umsetzen:

$a \leftarrow p^2; \quad b \leftarrow a + q; \quad c \leftarrow \sqrt{b}$
$x_1 \leftarrow p + c; \quad x_2 \leftarrow p - c$

Wir konzentrieren uns auf die letzten beiden Schritte: Falls p nicht-negativ ist, tritt bei der Bestimmung von x_1 keine Auslöschung auf, bei der von x_2 hingegen kann es zu Problemen kommen, falls q nahe null ist und damit p und c fast gleich sind. Für negatives p können entsprechende Schwierigkeiten bei der Berechnung von x_1 auftreten, während diesmal x_2 unkritisch ist. Glücklicherweise lassen sich diese Schwierigkeiten mit einem alternativen Algorithmus vermeiden: Da x_1 und x_2 die Nullstellen des Polynoms $x^2 - 2px - q$ sind, muss insbesondere

$$x^2 - 2px - q = (x - x_1)(x - x_2) = x^2 - (x_1 + x_2)x + x_1 x_2$$

gelten, und wir erhalten per Koeffizientenvergleich die Beziehung $q = -x_1 x_2$. Aus x_1 können wir also x_2 beziehungsweise aus x_2 auch x_1 mit einer Division unmittelbar gewinnen und so die Subtraktion und damit die Gefahr der Auslöschung vermeiden. Diese Strategie setzt der folgende verbesserte Algorithmus um:

if $p \geq 0$ **then**
  $a \leftarrow p^2; \quad b \leftarrow a + q; \quad c \leftarrow \sqrt{b}$
  $x_1 \leftarrow p + c; \quad x_2 \leftarrow -q/x_1$
else
  $a \leftarrow p^2; \quad b \leftarrow a + q; \quad c \leftarrow \sqrt{b}$
  $x_2 \leftarrow p - c; \quad x_1 \leftarrow -q/x_2$
end if

In exakter Arithmetik wären beide Algorithmen völlig gleichwertig. Tatsächlich auf einem Computer ausführen können wir aber nur eine Fassung, bei der alle Rechenoperationen durch ihre gerundeten Gegenstücke ersetzt werden. Wir bezeichnen wie zuvor mit $\oplus$, $\ominus$, $\odot$ und $\oslash$ die gerundete Addition, Subktraktion, Multiplikation und Division. Mit $\operatorname{sqrt}(x) := \operatorname{fl}(\sqrt{x})$ bezeichnen wir die gerundete Wurzelfunktion. Dann nimmt der praktisch durchführbare Algorithmus die folgende Form an:

if $p \geq 0$ **then**
  $\tilde{a} \leftarrow p \odot p$; $\tilde{b} \leftarrow \tilde{a} \oplus q$; $\tilde{c} \leftarrow \operatorname{sqrt}(\tilde{b})$
  $\tilde{x}_1 \leftarrow p \oplus \tilde{c}$; $\tilde{x}_2 \leftarrow -q \oslash \tilde{x}_1$
else
  $\tilde{a} \leftarrow p \odot p$; $\tilde{b} \leftarrow \tilde{a} \oplus q$; $\tilde{c} \leftarrow \operatorname{sqrt}(\tilde{b})$
  $\tilde{x}_2 \leftarrow p \ominus \tilde{c}$; $\tilde{x}_1 \leftarrow -q \oslash \tilde{x}_2$
end if

Wir analysieren seine Stabilität für den Fall $p, q \geq 0$ mit Hilfe der Folgerung 2.19: Es existieren $\epsilon_a, \epsilon_b, \epsilon_c, \epsilon_{x_1}, \epsilon_{x_2} \in [-\epsilon_{\text{fl}}, \epsilon_{\text{fl}}]$ mit

$$\tilde{a} = (1+\epsilon_a)p^2, \qquad \tilde{b} = (1+\epsilon_b)(\tilde{a}+q), \qquad \tilde{c} = (1+\epsilon_c)\sqrt{\tilde{b}},$$
$$\tilde{x}_1 = (1+\epsilon_{x_1})(p+\tilde{c}), \qquad \tilde{x}_2 = (1+\epsilon_{x_2})(-q/\tilde{x}_1).$$

Wir konstruieren nun $\delta_{x_1}, \delta_{x_2} \in \mathbb{R}$ derart, dass

$$\tilde{x}_1 = (1+\delta_{x_1})x_1, \qquad\qquad \tilde{x}_2 = (1+\delta_{x_2})x_2$$

gelten, denn dann sind $|\delta_{x_1}|$ und $|\delta_{x_2}|$ die relativen Fehler der beiden Ergebnisse.

Auf dem Weg berechnen wir auch δ_a, δ_b und δ_c, mit denen wir $\tilde{a}$, $\tilde{b}$ und $\tilde{c}$ entsprechend durch a, b und c ausdrücken können.

Die ersten Schritte sind einfach:

$$\begin{aligned}
\tilde{a} &= (1+\epsilon_a)p^2 = (1+\epsilon_a)a = (1+\delta_a)a, && \delta_a := \epsilon_a,\\
\tilde{b} &= (1+\epsilon_b)(\tilde{a}+q) = (1+\epsilon_b)((1+\epsilon_a)a+q)\\
&= (1+\epsilon_b)(a+q+\epsilon_a a) = (1+\epsilon_b)(b+\epsilon_a a)\\
&= (1+\epsilon_b)\left(1+\epsilon_a\frac{a}{b}\right)b = (1+\delta_b)b, && \delta_b := \epsilon_a\frac{a}{b} + \epsilon_b\left(1+\epsilon_a\frac{a}{b}\right),\\
\tilde{c} &= (1+\epsilon_c)\sqrt{\tilde{b}} = (1+\epsilon_c)\sqrt{(1+\delta_b)b}\\
&= (1+\epsilon_c)\sqrt{1+\delta_b}\,\sqrt{b} = (1+\epsilon_c)\sqrt{1+\delta_b}\,c\\
&= (1+\delta_c)c && \delta_c := (1+\epsilon_c)\sqrt{1+\delta_b} - 1,\\
\tilde{x}_1 &= (1+\epsilon_{x_1})(p+\tilde{c}) = (1+\epsilon_{x_1})(p+(1+\delta_c)c)\\
&= (1+\epsilon_{x_1})(x_1+\delta_c c) = (1+\delta_{x_1})x_1, && \delta_{x_1} := \epsilon_{x_1} + (1+\epsilon_{x_1})\delta_c\frac{c}{x_1},\\
\tilde{x}_2 &= (1+\epsilon_{x_2})(-q/\tilde{x}_1) = (1+\epsilon_{x_2})\frac{-q}{(1+\delta_{x_1})x_1}\\
&= \frac{1+\epsilon_{x_2}}{1+\delta_{x_1}}x_2 = (1+\delta_{x_2})x_2, && \delta_{x_2} := \frac{1+\epsilon_{x_2}}{1+\delta_{x_1}} - 1.
\end{aligned}$$

Die Herausforderung besteht nun darin, obere Schranken für $|\delta_{x_1}|$ und $|\delta_{x_2}|$ zu finden. Wir erhalten direkt

$$\begin{aligned}
|\delta_a| &= |\epsilon_a| \leq \epsilon_{\text{fl}}, \\
|\delta_b| &= \left|\epsilon_a \frac{a}{b} + \epsilon_b \left(1 + \epsilon_a \frac{a}{b}\right)\right| \leq |\epsilon_a| \frac{|a|}{|b|} + |\epsilon_b| \left(1 + |\epsilon_a| \frac{|a|}{|b|}\right) \\
&\leq |\epsilon_a| + |\epsilon_b|(1 + |\epsilon_a|) \leq |\epsilon_a| + \frac{3}{2}|\epsilon_b| \leq \frac{5}{2}\epsilon_{\text{fl}},
\end{aligned}$$

wobei wir sowohl $|\epsilon_a| \leq \epsilon_{\text{fl}} \leq 1/2$ ausgenutzt haben als auch, dass wegen $q \geq 0$ auch $b \geq a$ und damit auch $|b| \geq |a|$ gilt.

Für die nächsten Schritte brauchen wir zwei einfache Hilfsaussagen.

Lemma 2.21 (Kehrwert und Wurzel) *Sei $x \in [-1/2, 1/2]$. Es gibt $y_1, y_2 \in \mathbb{R}$ mit*

$$\frac{1}{1+x} = 1 + y_1, \qquad |y_1| \leq 2\,|x|, \tag{2.4a}$$

$$\sqrt{1+x} = 1 + y_2, \qquad |y_2| \leq \sqrt{1/2}\,|x|. \tag{2.4b}$$

Beweis. Für den Kehrwert setzen wir

$$y_1 := \frac{1}{1+x} - 1 = \frac{-x}{1+x}$$

und erhalten direkt

$$|y_1| = \frac{1}{|1+x|}|x| \leq 2\,|x|.$$

Für die Wurzel betrachten wir die Funktion $f(x) = \sqrt{1+x}$ mit der Ableitung $f'(x) = \frac{1}{2\sqrt{1+x}}$. Für jedes $x \in [-1/2, 1/2]$ finden wir mit dem Mittelwertsatz der Differentialrechnung ein $\eta \in [-1/2, 1/2]$ mit

$$y_2 := \sqrt{1+x} - 1 = f(x) - f(0) = f'(\eta)\,x,$$

und wegen $|f'(\eta)| \leq \sqrt{1/2}$ folgt

$$|y_2| = |f'(\eta)|\,|x| \leq \sqrt{1/2}\,|x|.$$

Damit ist auch (2.4b) bewiesen. ∎

Indem wir (2.4b) auf δ_b anwenden, finden wir ein $\tilde{\delta}_b$ mit

$$\sqrt{1+\delta_b} = 1 + \tilde{\delta}_b, \qquad |\tilde{\delta}_b| \leq \sqrt{1/2}\,|\delta_b| \leq \frac{5}{2\sqrt{2}}\,\epsilon_{\text{fl}} \leq 2\epsilon_{\text{fl}}$$

und erhalten

$$\delta_c = (1+\epsilon_c)\sqrt{1+\delta_b} - 1 = (1+\epsilon_c)(1+\tilde{\delta}_b) - 1 = \tilde{\delta}_b + \epsilon_c(1+\tilde{\delta}_b),$$

$$
\begin{aligned}
|\delta_c| &\le |\tilde{\delta}_b| + |\epsilon_c|(1+|\tilde{\delta}_b|) \le 2\epsilon_{\text{fl}} + \epsilon_{\text{fl}}(1+2\epsilon_{\text{fl}}) \le 4\epsilon_{\text{fl}},\\
|\delta_{x_1}| &= \left|\epsilon_{x_1} + (1+\epsilon_{x_1})\delta_c \frac{c}{x_1}\right| \le |\epsilon_{x_1}| + (1+|\epsilon_{x_1}|)|\delta_c|\frac{|c|}{|x_1|}\\
&\le \epsilon_{\text{fl}} + (1+\epsilon_{\text{fl}})4\epsilon_{\text{fl}} \le 7\epsilon_{\text{fl}},
\end{aligned}
$$

wobei wir ausgenutzt haben, dass wegen $p, c \ge 0$ auch $|x_1| = |p + c| \ge |c|$ gilt. Der relative Fehler der Näherungslösung $\tilde{x}_1$ ist also weniger als das Siebenfache der Maschinengenauigkeit.

Für x_2 wenden wir (2.4a) auf δ_{x_1} an und finden ein $\tilde{\delta}_{x_1}$ mit

$$
\frac{1}{1+\delta_{x_1}} = 1 + \tilde{\delta}_{x_1}, \qquad\qquad |\tilde{\delta}_{x_1}| \le 2|\delta_{x_1}| \le 14\epsilon_{\text{fl}}.
$$

Es folgt

$$
\begin{aligned}
|\delta_{x_2}| &= \left|\frac{1+\epsilon_{x_2}}{1+\delta_{x_1}} - 1\right| = \left|(1+\epsilon_{x_2})(1+\tilde{\delta}_{x_1}) - 1\right| = \left|\tilde{\delta}_{x_1} + \epsilon_{x_2}(1+\tilde{\delta}_{x_1})\right|\\
&\le 14\epsilon_{\text{fl}} + \epsilon_{\text{fl}}\,(1+14\epsilon_{\text{fl}}) \le 14\epsilon_{\text{fl}} + 8\epsilon_{\text{fl}} = 22\epsilon_{\text{fl}},
\end{aligned}
$$

der relative Fehler der Näherungslösung $\tilde{x}_2$ ist also weniger als das 22fache der Maschinengenauigkeit. Damit können wir den Algorithmus als durchaus stabil ansehen.

Bemerkung 2.22 (Quadratische Terme) *Da ϵ_{fl}^2 in der Regel sehr viel kleiner als ϵ_{fl} ist, wird es bei der Analyse des Rundungsfehlers üblicherweise ignoriert. In diesem Fall erhielten wir $|\delta_b| \lesssim 2\epsilon_{\text{fl}}$, $|\delta_c| \lesssim 3\epsilon_{\text{fl}}$, $|\delta_{x_1}| \lesssim 4\epsilon_{\text{fl}}$ und schließlich $|\delta_{x_2}| \lesssim 5\epsilon_{\text{fl}}$, also eine sehr viel günstigere Abschätzung, die trotzdem in der Praxis der Realität sehr nahe kommt.*

Damit kann unser Algorithmus sogar als sehr stabil gelten.

Wie man sieht, ist schon für relativ einfache Algorithmen der Nachweis der Stabilität relativ aufwendig. In vielen Situationen lässt sich dieser Nachweis erheblich vereinfachen, indem man eine *Rückwärtsanalyse* durchführt: Sei $\mathbf{x} \in \mathcal{V}$, und seien

$$
\mathbf{y} := \varphi(\mathbf{x}), \qquad\qquad \widetilde{\mathbf{y}} := \tilde{\varphi}(\mathbf{x})
$$

die exakte und die durch den Algorithmus angenäherte Lösung. Die Rückwärtsanalyse beruht auf der Idee, die Existenz eines zweiten Vektors $\widetilde{\mathbf{x}} \in \mathcal{V}$ nachzuweisen, der

$$
\tilde{\varphi}(\mathbf{x}) = \varphi(\widetilde{\mathbf{x}}), \qquad\qquad \frac{\|\mathbf{x} - \widetilde{\mathbf{x}}\|_V}{\|\mathbf{x}\|_V} \le C_{\text{stab}}\epsilon_{\text{fl}}
$$

mit einer von $\mathbf{x}$ unabhängigen Konstanten C_{stab} erfüllt. Falls C_{stab} nicht allzu groß ist, bezeichnen wir den Algorithmus als *rückwärtsstabil*.

In diesem Fall lässt sich für $C_{\text{stab}}\epsilon_{\text{fl}}\|\mathbf{x}\|_V \le \epsilon$ die bereits bekannte Konditionsabschätzung

$$
\frac{\|\varphi(\mathbf{x}) - \tilde{\varphi}(\mathbf{x})\|_W}{\|\varphi(\mathbf{x})\|_W} = \frac{\|\varphi(\mathbf{x}) - \varphi(\widetilde{\mathbf{x}})\|_W}{\|\varphi(\mathbf{x})\|_W} \le \kappa_{x,\epsilon}\frac{\|\mathbf{x} - \widetilde{\mathbf{x}}\|_V}{\|\mathbf{x}\|_V} \le \kappa_{x,\epsilon}C_{\text{stab}}\epsilon_{\text{fl}}
$$

anwenden, um zu zeigen, dass der durch den Algorithmus eingeführte relative Fehler nicht allzu groß ist.

3 Lineare Gleichungssysteme

In diesem Kapitel beschäftigen wir uns mit Lösungsverfahren für lineare Gleichungssysteme: Für $n \in \mathbb{N}$ und Koeffizienten $a_{11}, \ldots, a_{1n}, a_{21}, \ldots, a_{nn} \in \mathbb{K}$ sowie $b_1, \ldots, b_n \in \mathbb{K}$ suchen wir $x_1, \ldots, x_n \in \mathbb{K}$ mit

$$\begin{aligned} a_{11}x_1 + a_{12}x_2 + \ldots + a_{1n}x_n &= b_1, \\ &\vdots \\ a_{n1}x_1 + a_{n2}x_2 + \ldots + a_{nn}x_n &= b_n. \end{aligned}$$

Indem wir die Koeffizienten zusammenfassen, erhalten wir

$$\mathbf{A} = \begin{pmatrix} a_{11} & \ldots & a_{1n} \\ \vdots & \ddots & \vdots \\ a_{n1} & \ldots & a_{nn} \end{pmatrix}, \qquad \mathbf{b} = \begin{pmatrix} b_1 \\ \vdots \\ b_n \end{pmatrix}, \qquad \mathbf{x} = \begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix}$$

mit der Matrix $\mathbf{A} \in \mathbb{K}^{n \times n}$ und den Vektoren $\mathbf{b}, \mathbf{x} \in \mathbb{K}^n$. Mit diesen Abkürzungen lässt sich das Gleichungssystem in der kompakten Form

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

schreiben. Das System ist genau dann für alle $\mathbf{b}$ lösbar, wenn die Matrix $\mathbf{A}$ regulär ist, also ihre Spalten- beziehungsweise Zeilenvektoren linear unabhängig sind.

Die Aufgabe, mit der wir uns in diesem Kapitel beschäftigen werden, lautet wie folgt:

> Gegeben seien eine reguläre Matrix $\mathbf{A} \in \mathbb{K}^{n \times n}$ und ein Vektor $\mathbf{b} \in \mathbb{K}^n$, finde einen Vektor $\mathbf{x} \in \mathbb{K}^n$ so, dass $\mathbf{A}\mathbf{x} = \mathbf{b}$ gilt.

3.1 Kondition

Bevor wir den Einfluss von Fehlern auf die Lösung eines linearen Gleichungssystems untersuchen, bietet es sich an, den Einfluss von Fehlern auf die Matrix-Vektor-Multiplikation zu untersuchen, also auf die Abbildung

$$\varphi : \mathbb{K}^n \to \mathbb{K}^m, \qquad \mathbf{x} \mapsto \mathbf{A}\mathbf{x}$$

mit einer Matrix $\mathbf{A} \in \mathbb{K}^{m \times n}$. Wir verwenden eine Norm $\|\cdot\|_V$ für den Definitionsbereich $\mathcal{V} = \mathbb{K}^n$ und eine Norm $\|\cdot\|_W$ für den Bildbereich $\mathcal{W} = \mathbb{K}^m$. Für die absolute Kondition sind wir an der Größe

$$\sigma_{x,\epsilon} = \sup_{\widetilde{\mathbf{x}} \in \mathbb{K}^n \setminus \{\mathbf{x}\}} \frac{\|\varphi(\widetilde{\mathbf{x}}) - \varphi(\mathbf{x})\|_W}{\|\widetilde{\mathbf{x}} - \mathbf{x}\|_V} = \sup_{\widetilde{\mathbf{x}} \in \mathbb{K}^n \setminus \{\mathbf{x}\}} \frac{\|\mathbf{A}(\widetilde{\mathbf{x}} - \mathbf{x})\|_W}{\|\widetilde{\mathbf{x}} - \mathbf{x}\|_V} = \sup_{\mathbf{z} \in \mathbb{K}^n \setminus \{\mathbf{0}\}} \frac{\|\mathbf{A}\mathbf{z}\|_W}{\|\mathbf{z}\|_V}$$

interessiert, denn sie beschreibt gerade die bestmögliche absolute Fehlerschranke.

Erinnerung 3.1 (Heine-Borel) *Sei $n \in \mathbb{N}$. Eine Menge $X \subseteq \mathbb{K}^n$ ist genau dann kompakt, wenn sie abgeschlossen und beschränkt ist.*

Lemma 3.2 (Induzierte Matrixnorm) *Seien $\|\cdot\|_V : \mathbb{K}^n \to \mathbb{R}_{\geq 0}$ und $\|\cdot\|_W : \mathbb{K}^m \to \mathbb{R}_{\geq 0}$ Normen auf $\mathcal{V} = \mathbb{K}^n$ und $\mathcal{W} = \mathbb{K}^m$. Dann ist*

$$\|\cdot\|_{W\leftarrow V} : \mathbb{K}^{m\times n} \to \mathbb{R}_{\geq 0}, \qquad \mathbf{A} \mapsto \sup_{\mathbf{z}\in\mathcal{V}\setminus\{\mathbf{0}\}} \frac{\|\mathbf{Az}\|_W}{\|\mathbf{z}\|_V},$$

eine Norm auf dem Raum $\mathbb{K}^{m\times n}$ der Matrizen. Diese Norm nennen wir die von $\mathcal{V}$ und $\mathcal{W}$ induzierte Matrixnorm*.*

Beweis. Sei $\mathbf{A} \in \mathbb{K}^{m\times n}$. Wir beweisen zunächst, dass $\|\mathbf{A}\|_{W\leftarrow V}$ wohldefiniert, dass also das Supremum endlich ist. Für alle $\mathbf{z} \in \mathcal{V} \setminus \{\mathbf{0}\}$ gilt

$$\frac{\|\mathbf{Az}\|_W}{\|\mathbf{z}\|_V} = \left\| \frac{\mathbf{Az}}{\|\mathbf{z}\|_V} \right\|_W = \left\| \mathbf{A} \frac{\mathbf{z}}{\|\mathbf{z}\|_V} \right\|_W,$$

und $\mathbf{z}/\|\mathbf{z}\|_V$ ist ein Einheitsvektor bezüglich der $\mathcal{V}$-Norm. Also genügt es, das Supremum über der Einheitssphäre

$$\mathcal{S}_V := \{\mathbf{z} \in \mathcal{V} \ : \ \|\mathbf{z}\|_V = 1\}$$

zu betrachten, denn wir haben

$$\|\mathbf{A}\|_{W\leftarrow V} = \sup_{\mathbf{z}\in\mathcal{S}_V} \|\mathbf{Az}\|_W. \tag{3.1}$$

Da die Einheitssphäre $\mathcal{S}_V$ beschränkt und abgeschlossen ist, ist sie nach dem Satz von Heine-Borel auch kompakt, und die Norm nimmt als stetige Abbildung auf $\mathcal{S}_V$ ein Maximum an. Also ist die induzierte Matrixnorm wohldefiniert, das Supremum in (3.1) ist ein Maximum. Wir müssen nur noch die Normaxiome nachprüfen.

Falls $\mathbf{A} = \mathbf{0}$ gilt, folgt unmittelbar $\|\mathbf{A}\|_{W\leftarrow V} = 0$. Falls umgekehrt $\|\mathbf{A}\|_{W\leftarrow V} = 0$ gilt, folgt $\|\mathbf{Az}\|_W = 0$ für alle $\mathbf{z} \in \mathcal{V}$, also $\mathbf{Az} = 0$, und damit $\mathbf{A} = 0$.

Sei $\alpha \in \mathbb{K}$, sei $\mathbf{B} \in \mathbb{K}^{m\times n}$. Dann gelten

$$\begin{aligned}
\|\alpha\mathbf{A}\|_{W\leftarrow V} &= \sup_{\mathbf{z}\in\mathcal{V}\setminus\{\mathbf{0}\}} \frac{\|\alpha\mathbf{Az}\|_W}{\|\mathbf{z}\|_V} = |\alpha| \sup_{\mathbf{z}\in\mathcal{V}\setminus\{\mathbf{0}\}} \frac{\|\mathbf{Az}\|_W}{\|\mathbf{z}\|_V} = |\alpha|\,\|\mathbf{A}\|_{W\leftarrow V}, \\
\|\mathbf{A}+\mathbf{B}\|_{W\leftarrow V} &= \sup_{\mathbf{z}\in\mathcal{V}\setminus\{\mathbf{0}\}} \frac{\|(\mathbf{A}+\mathbf{B})\mathbf{z}\|_W}{\|\mathbf{z}\|_V} = \sup_{\mathbf{z}\in\mathcal{V}\setminus\{\mathbf{0}\}} \frac{\|\mathbf{Az}+\mathbf{Bz}\|_W}{\|\mathbf{z}\|_V} \\
&\leq \sup_{\mathbf{z}\in\mathcal{V}\setminus\{\mathbf{0}\}} \frac{\|\mathbf{Az}\|_W + \|\mathbf{Bz}\|_W}{\|\mathbf{z}\|_V} \leq \sup_{\mathbf{z}\in\mathcal{V}\setminus\{\mathbf{0}\}} \frac{\|\mathbf{Az}\|_W}{\|\mathbf{z}\|_V} + \sup_{\mathbf{z}\in\mathcal{V}\setminus\{\mathbf{0}\}} \frac{\|\mathbf{Bz}\|_W}{\|\mathbf{z}\|_V} \\
&= \|\mathbf{A}\|_{W\leftarrow V} + \|\mathbf{B}\|_{W\leftarrow V},
\end{aligned}$$

also sind die Bedingungen Definition 2.2 in der Tat erfüllt. ∎

Für unsere Zwecke von zentraler Bedeutung ist, dass die induzierte Matrixnorm uns die Möglichkeit bietet, den Einfluss der Matrix-Vektor-Multiplikation auf die Norm eines Vektors quantitativ zu analysieren, denn damit können wir beispielsweise die Fortpflanzung von Fehlern abschätzen.

Lemma 3.3 (Verträglichkeit) *Sei $k \in \mathbb{N}$ gegeben. Seien $\|\cdot\|_U$, $\|\cdot\|_V$ und $\|\cdot\|_W$ Normen auf $\mathcal{U} = \mathbb{K}^k$, $\mathcal{V} = \mathbb{K}^n$ und $\mathcal{W} = \mathbb{K}^m$. Dann gelten*

$$\|\mathbf{A}\mathbf{y}\|_W \leq \|\mathbf{A}\|_{W\leftarrow V}\|\mathbf{y}\|_V \qquad \text{für alle } \mathbf{A} \in \mathbb{K}^{m\times n},\ \mathbf{y} \in \mathcal{V}, \tag{3.2a}$$

die Normen sind verträglich*, und*

$$\|\mathbf{A}\mathbf{B}\|_{W\leftarrow U} \leq \|\mathbf{A}\|_{W\leftarrow V}\|\mathbf{B}\|_{V\leftarrow U} \qquad \text{für alle } \mathbf{A} \in \mathbb{K}^{m\times n},\ \mathbf{B} \in \mathbb{K}^{n\times k}, \tag{3.2b}$$

die Normen sind submultiplikativ*.*

Beweis. Seien $\mathbf{A} \in \mathbb{K}^{m\times n}$ und $\mathbf{y} \in \mathcal{V} = \mathbb{K}^n \setminus \{\mathbf{0}\}$ gegeben. Dann gilt

$$\|\mathbf{A}\mathbf{y}\|_W = \frac{\|\mathbf{A}\mathbf{y}\|_W}{\|\mathbf{y}\|_V}\|\mathbf{y}\|_V \leq \left(\sup_{\mathbf{z}\in\mathcal{V}\setminus\{\mathbf{0}\}} \frac{\|\mathbf{A}\mathbf{z}\|_W}{\|\mathbf{z}\|_V}\right)\|\mathbf{y}\|_V = \|\mathbf{A}\|_{W\leftarrow V}\|\mathbf{y}\|_V,$$

also sind die Normen mit der induzierten Matrixnorm verträglich.

Sei $\mathbf{B} \in \mathbb{K}^{n\times k}$. Dann folgt mit (3.2a) die Abschätzung

$$\begin{aligned}\|\mathbf{A}\mathbf{B}\|_{W\leftarrow U} &= \sup_{z\in\mathcal{U}\setminus\{\mathbf{0}\}} \frac{\|\mathbf{A}\mathbf{B}\mathbf{z}\|_W}{\|\mathbf{z}\|_U} \leq \sup_{z\in\mathcal{U}\setminus\{\mathbf{0}\}} \frac{\|\mathbf{A}\|_{W\leftarrow V}\|\mathbf{B}\mathbf{z}\|_V}{\|\mathbf{z}\|_U} \\ &\leq \sup_{z\in\mathcal{U}\setminus\{\mathbf{0}\}} \frac{\|\mathbf{A}\|_{W\leftarrow V}\|\mathbf{B}\|_{V\leftarrow U}\|\mathbf{z}\|_U}{\|\mathbf{z}\|_U} = \|\mathbf{A}\|_{W\leftarrow V}\|\mathbf{B}\|_{V\leftarrow U},\end{aligned}$$

also sind die Matrixnormen submultiplikativ. ∎

In der Praxis ist es häufig schwierig, die induzierte Matrixnorm tatsächlich zu berechnen, statt sie lediglich abzuschätzen. Um so dankbarer ist man, wenn man eine Kombination von Normen $\|\cdot\|_V$ und $\|\cdot\|_W$ findet, für die man die induzierte Matrixnorm tatsächlich explizit angeben kann.

Lemma 3.4 (Zeilensummennorm) *Seien $\|\cdot\|_V$ und $\|\cdot\|_W$ die zu $\mathbb{K}^n$ und $\mathbb{K}^m$ gehörenden Maximumnormen (vgl. Beispiel 2.3). Dann gilt*

$$\|\mathbf{A}\|_{W\leftarrow V} = \|\mathbf{A}\|_\infty := \max\left\{\sum_{j=1}^n |a_{ij}| \ :\ i \in [1:m]\right\} \qquad \text{für alle } \mathbf{A} \in \mathbb{K}^{m\times n}.$$

Beweis. Sei $\mathbf{A} \in \mathbb{K}^{m\times n}$. Wir bezeichnen mit

$$\|\mathbf{A}\|_\infty := \max\left\{\sum_{j=1}^n |a_{ij}| \ :\ i \in [1:m]\right\}.$$

das Maximum der absoluten Zeilensummen der Matrix $\mathbf{A}$.

Sei $\mathbf{z} \in \mathbb{K}^n$. Dann gilt mit der Dreiecksungleichung

$$|(\mathbf{Az})_i| = \left| \sum_{j=1}^{n} a_{ij} z_j \right| \leq \sum_{j=1}^{n} |a_{ij}|\,|z_j| \leq \sum_{j=1}^{n} |a_{ij}|\,\|\mathbf{z}\|_\infty \leq \|\mathbf{A}\|_\infty \|\mathbf{z}\|_\infty,$$

und damit $\|\mathbf{Az}\|_\infty \leq \|\mathbf{A}\|_\infty \|\mathbf{z}\|_\infty$, also $\|\mathbf{A}\|_{W \leftarrow V} \leq \|\mathbf{A}\|_\infty$.

Sei nun ein $i \in [1:m]$ so fixiert, dass

$$\|\mathbf{A}\|_\infty = \sum_{j=1}^{n} |a_{ij}|$$

gilt. Wir definieren den Vektor $\mathbf{z} \in \mathcal{V}$ durch

$$z_j := \begin{cases} |a_{ij}|/a_{ij} & \text{falls } a_{ij} \neq 0, \\ 1 & \text{ansonsten} \end{cases} \qquad \text{für alle } j \in [1:n]$$

und erhalten $\|\mathbf{z}\|_\infty = 1$ und $a_{ij} z_j = |a_{ij}|$ sowie

$$\|\mathbf{A}\|_\infty = \sum_{j=1}^{n} |a_{ij}| = \sum_{j=1}^{n} a_{ij} z_j = (\mathbf{Az})_i,$$

also wegen $\|\mathbf{z}\|_\infty = 1$ auch

$$\|\mathbf{A}\|_\infty \leq \|\mathbf{Az}\|_\infty \leq \|\mathbf{A}\|_{W \leftarrow V} \|\mathbf{z}\|_\infty = \|\mathbf{A}\|_{W \leftarrow V},$$

also $\|\mathbf{A}\|_\infty \leq \|\mathbf{A}\|_{W \leftarrow V}$. Da wir $\|\mathbf{A}\|_{W \leftarrow V} \leq \|\mathbf{A}\|_\infty$ bereits bewiesen haben erhalten wir insgesamt

$$\|\mathbf{A}\|_{W \leftarrow V} = \|\mathbf{A}\|_\infty.$$

Das ist die gewünschte Gleichung. ∎

Wir haben bereits gesehen, dass $\|\mathbf{A}\|_{W \leftarrow V}$ die optimale absolute Fehlerschranke für die Matrix-Vektor-Multiplikation ist.

In der Praxis sind wir allerdings häufig eher an *relativen* Fehlerschranken interessiert, also an Abschätzungen der Form

$$\frac{\|\mathbf{A}\widetilde{\mathbf{x}} - \mathbf{Ax}\|_W}{\|\mathbf{Ax}\|_W} \leq \kappa_{x,\epsilon} \frac{\|\widetilde{\mathbf{x}} - \mathbf{x}\|_V}{\|\mathbf{x}\|_V}.$$

Für den Zähler des Bruchs auf der linken Seite können wir die Abschätzung des absoluten Fehlers einsetzen. Eine untere Schranke für den Nenner dürfen wir nur erwarten, falls die Matrix $\mathbf{A}$ injektiv ist, denn sonst wäre es ja möglich, dass $\|\mathbf{Ax}\|_W = 0$ gilt, aber $\|\mathbf{x}\|_V > 0$.

Da $\mathbf{A}$ eine quadratische Matrix ist, impliziert Injektivität bereits Regularität, wir dürfen also davon ausgehen, dass $\mathbf{A}^{-1}$ existiert. Unter dieser Annahme ist die Abschätzung des relativen Fehlers keine große Herausforderung.

Lemma 3.5 (Kondition) *Für alle* $\mathbf{x} \in \mathcal{V} \setminus \{\mathbf{0}\}$ *und alle* $\widetilde{\mathbf{x}} \in \mathcal{V}$ *gilt*

$$\frac{\|\mathbf{A}\widetilde{\mathbf{x}} - \mathbf{A}\mathbf{x}\|_W}{\|\mathbf{A}\mathbf{x}\|_W} \leq \|\mathbf{A}\|_{W\leftarrow V}\|\mathbf{A}^{-1}\|_{V\leftarrow W}\frac{\|\widetilde{\mathbf{x}} - \mathbf{x}\|_V}{\|\mathbf{x}\|_V}.$$

Beweis. Mit der definierenden Eigenschaft $\mathbf{A}^{-1}\mathbf{A} = \mathbf{I}$ der Inversen und der Verträglichkeit (3.2a) gilt

$$\begin{aligned}
\|\mathbf{x}\|_V = \|\mathbf{A}^{-1}\mathbf{A}\mathbf{x}\|_V &\leq \|\mathbf{A}^{-1}\|_{V\leftarrow W}\|\mathbf{A}\mathbf{x}\|_W, \\
\frac{\|\mathbf{x}\|_V}{\|\mathbf{A}^{-1}\|_{V\leftarrow W}} &\leq \|\mathbf{A}\mathbf{x}\|_W, \\
\|\mathbf{A}(\widetilde{\mathbf{x}} - \mathbf{x})\|_W &\leq \|\mathbf{A}\|_{W\leftarrow V}\|\widetilde{\mathbf{x}} - \mathbf{x}\|_V.
\end{aligned}$$

Wir kombinieren beide Abschätzungen und erhalten damit die Abschätzung

$$\begin{aligned}
\frac{\|\mathbf{A}\widetilde{\mathbf{x}} - \mathbf{A}\mathbf{x}\|_W}{\|\mathbf{A}\mathbf{x}\|_W} = \frac{\|\mathbf{A}(\widetilde{\mathbf{x}} - \mathbf{x})\|_W}{\|\mathbf{A}\mathbf{x}\|_W} &\leq \frac{\|\mathbf{A}\|_{W\leftarrow V}\|\widetilde{\mathbf{x}} - \mathbf{x}\|_V}{\|\mathbf{x}\|_V/\|\mathbf{A}^{-1}\|_{V\leftarrow W}} \\
&= \|\mathbf{A}\|_{W\leftarrow V}\|\mathbf{A}^{-1}\|_{V\leftarrow W}\frac{\|\widetilde{\mathbf{x}} - \mathbf{x}\|_V}{\|\mathbf{x}\|_V},
\end{aligned}$$

und das ist die gewünschte Ungleichung. ∎

Definition 3.6 (Konditionszahl) *Sei* $\mathbf{A} \in \mathbb{K}^{n\times n}$ *eine reguläre Matrix, und seien* $\|\cdot\|_V$ *und* $\|\cdot\|_W$ *zwei Normen. Die Größe*

$$\kappa_{W\leftarrow V}(\mathbf{A}) := \|\mathbf{A}\|_{W\leftarrow V}\|\mathbf{A}^{-1}\|_{V\leftarrow W}$$

bezeichnen wir als die Konditionszahl *der Matrix* $\mathbf{A}$*, sie beschreibt gemäß Lemma 3.5 die relative Fehlerverstärkung bei der Matrix-Vektor-Multiplikation.*

Für identische Normen verwenden wir die Abkürzung

$$\kappa_V(\mathbf{A}) := \|\mathbf{A}\|_{V\leftarrow V}\|\mathbf{A}^{-1}\|_{V\leftarrow V}.$$

Nun können wir uns wieder der Untersuchung des Lösens des linearen Gleichungssystems zuwenden, also der Berechnung von

$$\varphi : W \to V, \qquad \mathbf{b} \mapsto \mathbf{x} \text{ mit } \mathbf{A}\mathbf{x} = \mathbf{b}.$$

Durch Multiplizieren der definierenden Gleichung mit der Inversen $\mathbf{A}^{-1}$ erhalten wir direkt

$$\varphi(\mathbf{b}) = \mathbf{A}^{-1}\mathbf{b} \qquad \text{für alle } \mathbf{b} \in \mathbb{K}^n,$$

das Lösen des linearen Gleichungssystems entspricht also der Matrix-Vektor-Multiplikation mit $\mathbf{A}^{-1}$.

Somit können wir Lemma 3.5 anwenden, um zu folgern, dass für die Lösungen $\mathbf{x}, \widetilde{\mathbf{x}} \in \mathcal{V}$ der Gleichungssysteme

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

wegen $\mathbf{x} = \mathbf{A}^{-1}\mathbf{b}$ und $\widetilde{\mathbf{x}} = \mathbf{A}^{-1}\widetilde{\mathbf{b}}$ die Abschätzung

$$\frac{\|\widetilde{\mathbf{x}} - \mathbf{x}\|_V}{\|\mathbf{x}\|_V} \leq \|\mathbf{A}^{-1}\|_{V\leftarrow W}\|\mathbf{A}\|_{W\leftarrow V}\frac{\|\widetilde{\mathbf{b}} - \mathbf{b}\|_W}{\|\mathbf{b}\|_W} = \kappa_{W\leftarrow V}(\mathbf{A})\frac{\|\widetilde{\mathbf{b}} - \mathbf{b}\|_W}{\|\mathbf{b}\|_W}$$

gilt. Die Konditionszahl $\kappa_{W\leftarrow V}(\mathbf{A})$ beschreibt also auch die relative Fehlerverstärkung für das Lösen des Gleichungssystems.

Etwas schwieriger wird die Fehlerabschätzung, wenn wir den Fall untersuchen, dass auch die Matrix $\mathbf{A}$ gestört ist. In diesem Fall muss zunächst sicher gestellt werden, dass die gestörte Matrix regulär ist. Um die Regularität einer Matrix quantitativ erfassen zu können, benötigen wir die folgende Hilfsaussage:

Lemma 3.7 (Neumannsche Reihe) *Sei $\mathbf{X} \in \mathbb{K}^{n\times n}$ eine Matrix mit $\|\mathbf{X}\|_{V\leftarrow V} < 1$. Dann ist die Matrix $\mathbf{I} - \mathbf{X}$ regulär und die Norm ihrer Inversen lässt sich durch*

$$\|(\mathbf{I} - \mathbf{X})^{-1}\|_{V\leftarrow V} \leq \frac{1}{1 - \|\mathbf{X}\|_{V\leftarrow V}} \tag{3.3}$$

beschränken.

Beweis. Wir untersuchen die durch

$$\mathbf{Y}_m := \sum_{\ell=0}^{m} \mathbf{X}^\ell \qquad \text{für alle } m \in \mathbb{N}_0.$$

gegebene *Neumannsche Reihe* $(\mathbf{Y}_m)_{m=0}^{\infty}$. Dank der Dreiecksungleichung und der Submultiplikativität (3.2b) der Norm erhalten wir

$$\|\mathbf{Y}_m\|_{V\leftarrow V} \leq \sum_{\ell=0}^{m} \|\mathbf{X}^\ell\|_{V\leftarrow V} \leq \sum_{\ell=0}^{m} \|\mathbf{X}\|_{V\leftarrow V}^\ell \qquad \text{für alle } m \in \mathbb{N}_0.$$

Da $\|\mathbf{X}\|_{V\leftarrow V} < 1$ nach Voraussetzung gilt, dürfen wir die geometrische Summenformel anwenden, um

$$\|\mathbf{Y}_m\|_{V\leftarrow V} \leq \frac{1 - \|\mathbf{X}\|_{V\leftarrow V}^{m+1}}{1 - \|\mathbf{X}\|_{V\leftarrow V}} \leq \frac{1}{1 - \|\mathbf{X}\|_{V\leftarrow V}} \qquad \text{für alle } m \in \mathbb{N}_0$$

zu erhalten. Damit ist die Neumannsche Reihe absolut summierbar, also insbesondere auch summierbar mit einem Grenzwert $\mathbf{Y} \in \mathbb{K}^{n\times n}$, der

$$\|\mathbf{Y}\|_{V\leftarrow V} = \lim_{m\to\infty} \|\mathbf{Y}_m\|_{V\leftarrow V} \leq \frac{1}{1 - \|\mathbf{X}\|_{V\leftarrow V}}$$

erfüllt. Wir können elementar nachrechnen, dass

$$(\mathbf{I}-\mathbf{X})\mathbf{Y}_m = \sum_{\ell=0}^{m}\mathbf{X}^\ell - \sum_{\ell=1}^{m+1}\mathbf{X}^\ell = \mathbf{I}-\mathbf{X}^{m+1} \qquad \text{für alle } m\in\mathbb{N}_0$$

gilt, und für den Grenzwert folgt wegen (3.2b) und $\|\mathbf{X}\|_{V\leftarrow V}<1$

$$\lim_{m\to\infty}\|\mathbf{X}^m\|_{V\leftarrow V} \le \lim_{m\to\infty}\|\mathbf{X}\|_{V\leftarrow V}^m = 0,$$

also $\lim_{m\to\infty}\mathbf{X}^m=\mathbf{0}$, und damit auch

$$(\mathbf{I}-\mathbf{X})\mathbf{Y} = \lim_{m\to\infty}(\mathbf{I}-\mathbf{X})\mathbf{Y}_m = \lim_{m\to\infty}(\mathbf{I}-\mathbf{X}^{m+1}) = \mathbf{I},$$

also muss $\mathbf{I}-\mathbf{X}$ regulär sein mit $\mathbf{Y}=(\mathbf{I}-\mathbf{X})^{-1}$. ∎

Lemma 3.8 (Gestörte Matrix) *Sei $\mathbf{b}\in\mathcal{W}$. Sei $\widetilde{\mathbf{A}}\in\mathbb{K}^{n\times n}$ eine Matrix, für die*

$$\|\mathbf{A}^{-1}(\mathbf{A}-\widetilde{\mathbf{A}})\|_{V\leftarrow V}<1$$

gilt. Dann ist die Matrix $\widetilde{\mathbf{A}}$ regulär mit

$$\|\widetilde{\mathbf{A}}^{-1}\|_{V\leftarrow W} \le \frac{\|\mathbf{A}^{-1}\|_{V\leftarrow W}}{1-\|\mathbf{A}^{-1}(\mathbf{A}-\widetilde{\mathbf{A}})\|_{V\leftarrow V}}.$$

Beweis. Wir definieren die Matrix

$$\mathbf{X} := \mathbf{A}^{-1}(\mathbf{A}-\widetilde{\mathbf{A}}) \in \mathbb{K}^{n\times n}$$

und folgern aus der Voraussetzung, dass

$$\|\mathbf{X}\|_{V\leftarrow V}<1$$

gilt. Also können wir Lemma 3.7 anwenden und finden, dass die Matrix $\mathbf{I}-\mathbf{X}$ regulär ist und ihre Inverse die Abschätzung (3.3) erfüllt. Aufgrund der Regularität von

$$\mathbf{I}-\mathbf{X} = \mathbf{I}-\mathbf{A}^{-1}(\mathbf{A}-\widetilde{\mathbf{A}}) = \mathbf{I}-\mathbf{I}+\mathbf{A}^{-1}\widetilde{\mathbf{A}} = \mathbf{A}^{-1}\widetilde{\mathbf{A}}$$

muss auch $\widetilde{\mathbf{A}}$ regulär sein und

$$\begin{aligned}\|\widetilde{\mathbf{A}}^{-1}\|_{V\leftarrow W} &= \|\widetilde{\mathbf{A}}^{-1}\mathbf{A}\mathbf{A}^{-1}\|_{V\leftarrow W} \le \|\widetilde{\mathbf{A}}^{-1}\mathbf{A}\|_{V\leftarrow V}\|\mathbf{A}^{-1}\|_{V\leftarrow W}\\ &= \|(\mathbf{I}-\mathbf{X})^{-1}\|_{V\leftarrow V}\|\mathbf{A}^{-1}\|_{V\leftarrow W} \le \frac{\|\mathbf{A}^{-1}\|_{V\leftarrow W}}{1-\|\mathbf{X}\|_{V\leftarrow V}}\\ &= \frac{\|\mathbf{A}^{-1}\|_{V\leftarrow W}}{1-\|\mathbf{A}^{-1}(\mathbf{A}-\widetilde{\mathbf{A}})\|_{V\leftarrow V}}\end{aligned}$$

erfüllen. ∎

Mit Hilfe dieser Resultate können wir nun einen Störungssatz formulieren, der Störungen in der Matrix *und* der rechten Seite gleichzeitig berücksichtigt.

Satz 3.9 (Störungssatz) *Seien* $\mathbf{b}, \widetilde{\mathbf{b}} \in \mathcal{W}$ *mit* $\mathbf{b} \neq \mathbf{0}$ *gegeben. Sei* $\widetilde{\mathbf{A}} \in \mathbb{K}^{n\times n}$ *eine Matrix mit*

$$\|\mathbf{A}^{-1}(\mathbf{A} - \widetilde{\mathbf{A}})\|_{V\leftarrow V} < 1. \tag{3.4}$$

Dann ist die Matrix $\widetilde{\mathbf{A}}$ *regulär, und die Lösungen* $\mathbf{x}, \widetilde{\mathbf{x}} \in \mathcal{V}$ *der Gleichungssysteme*

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

erfüllen die Abschätzung

$$\frac{\|\mathbf{x} - \widetilde{\mathbf{x}}\|_V}{\|\mathbf{x}\|_V} \leq \frac{\kappa_{W\leftarrow V}(\mathbf{A})}{1 - \|\mathbf{A}^{-1}(\mathbf{A} - \widetilde{\mathbf{A}})\|_{V\leftarrow V}} \left(\frac{\|\mathbf{A} - \widetilde{\mathbf{A}}\|_{W\leftarrow V}}{\|\mathbf{A}\|_{W\leftarrow V}} + \frac{\|\mathbf{b} - \widetilde{\mathbf{b}}\|_W}{\|\mathbf{b}\|_W} \right).$$

Beweis. Mit der Dreiecksungleichung erhalten wir

$$\|\mathbf{x} - \widetilde{\mathbf{x}}\|_V = \|\mathbf{x} - \widetilde{\mathbf{A}}^{-1}\mathbf{b} + \widetilde{\mathbf{A}}^{-1}(\mathbf{b} - \widetilde{\mathbf{b}})\|_V \leq \|\mathbf{x} - \widetilde{\mathbf{A}}^{-1}\mathbf{b}\|_V + \|\widetilde{\mathbf{A}}^{-1}\|_{V\leftarrow W}\|\mathbf{b} - \widetilde{\mathbf{b}}\|_W.$$

Für den ersten Term erhalten wir mit (3.2a) die Abschätzung

$$\begin{aligned}
\|\mathbf{x} - \widetilde{\mathbf{A}}^{-1}\mathbf{b}\|_V &= \|\widetilde{\mathbf{A}}^{-1}\widetilde{\mathbf{A}}\mathbf{x} - \widetilde{\mathbf{A}}^{-1}\mathbf{A}\mathbf{A}^{-1}\mathbf{b}\|_V = \|\widetilde{\mathbf{A}}^{-1}\widetilde{\mathbf{A}}\mathbf{x} - \widetilde{\mathbf{A}}^{-1}\mathbf{A}\mathbf{x}\|_V \\
&\leq \|\widetilde{\mathbf{A}}^{-1}\|_{V\leftarrow W}\|(\widetilde{\mathbf{A}} - \mathbf{A})\mathbf{x}\|_W \leq \|\widetilde{\mathbf{A}}^{-1}\|_{V\leftarrow W}\|\widetilde{\mathbf{A}} - \mathbf{A}\|_{W\leftarrow V}\|\mathbf{x}\|_V \\
&\leq \|\widetilde{\mathbf{A}}^{-1}\|_{V\leftarrow W}\|\mathbf{A}\|_{W\leftarrow V}\frac{\|\widetilde{\mathbf{A}} - \mathbf{A}\|_{W\leftarrow V}}{\|\mathbf{A}\|_{W\leftarrow V}}\|\mathbf{x}\|_V,
\end{aligned}$$

für den zweiten Term

$$\begin{aligned}
\|\widetilde{\mathbf{A}}^{-1}\|_{V\to W}\|\mathbf{b} - \widetilde{\mathbf{b}}\|_W &\leq \|\widetilde{\mathbf{A}}^{-1}\|_{V\leftarrow W}\frac{\|\mathbf{b} - \widetilde{\mathbf{b}}\|_W}{\|\mathbf{b}\|_W}\|\mathbf{A}\mathbf{x}\|_W \\
&\leq \|\widetilde{\mathbf{A}}^{-1}\|_{V\leftarrow W}\frac{\|\mathbf{b} - \widetilde{\mathbf{b}}\|_W}{\|\mathbf{b}\|_W}\|\mathbf{A}\|_{W\leftarrow V}\|\mathbf{x}\|_V.
\end{aligned}$$

Nun können wir Lemma 3.8 anwenden und finden

$$\begin{aligned}
\frac{\|\mathbf{x} - \widetilde{\mathbf{x}}\|_V}{\|\mathbf{x}\|_V} &\leq \|\widetilde{\mathbf{A}}^{-1}\|_{V\leftarrow W}\|\mathbf{A}\|_{W\leftarrow V}\left(\frac{\|\mathbf{A} - \widetilde{\mathbf{A}}\|_{W\leftarrow V}}{\|\mathbf{A}\|_{W\leftarrow V}} + \frac{\|\mathbf{b} - \widetilde{\mathbf{b}}\|_W}{\|\mathbf{b}\|_W} \right) \\
&\leq \frac{\|\mathbf{A}\|_{W\leftarrow V}\|\mathbf{A}^{-1}\|_{V\leftarrow W}}{1 - \|\mathbf{A}^{-1}(\mathbf{A} - \widetilde{\mathbf{A}})\|_{V\leftarrow V}} \left(\frac{\|\mathbf{A} - \widetilde{\mathbf{A}}\|_{W\leftarrow V}}{\|\mathbf{A}\|_{W\leftarrow V}} + \frac{\|\mathbf{b} - \widetilde{\mathbf{b}}\|_W}{\|\mathbf{b}\|_W} \right).
\end{aligned}$$

Mit $\kappa_{W\leftarrow V}(\mathbf{A}) = \|\mathbf{A}\|_{W\leftarrow V}\|\mathbf{A}^{-1}\|_{V\leftarrow W}$ folgt die Behauptung. ∎

In der Praxis wird häufig die Störung der Matrix $\|\mathbf{A}^{-1}(\mathbf{A} - \widetilde{\mathbf{A}})\|_{V\leftarrow V}$ relativ klein sein, beispielsweise in der Größenordnung der Maschinengenauigkeit, so dass der entscheidende Term der Abschätzung für den relativen Fehler wieder die Konditionszahl $\kappa_{W\leftarrow V}(\mathbf{A})$ ist, die auch in dieser allgemeineren Situation die Verstärkung der Fehler beschreibt.

Bemerkung 3.10 *Wenn wir von der etwas stärkeren Voraussetzung*

$$\kappa_{W\leftarrow V}(\mathbf{A})\frac{\|\mathbf{A}-\widetilde{\mathbf{A}}\|_{W\leftarrow V}}{\|\mathbf{A}\|_{W\leftarrow V}} < 1$$

ausgehen, können wir

$$\|\mathbf{A}^{-1}(\mathbf{A}-\widetilde{\mathbf{A}})\|_{V\leftarrow V} \le \|\mathbf{A}^{-1}\|_{V\leftarrow W}\|\mathbf{A}-\widetilde{\mathbf{A}}\|_{W\leftarrow V} \le \kappa_{W\leftarrow V}(\mathbf{A})\frac{\|\mathbf{A}-\widetilde{\mathbf{A}}\|_{W\leftarrow V}}{\|\mathbf{A}\|_{W\leftarrow V}} < 1$$

folgern und dürfen Satz 3.9 anwenden. Ausschlaggebend für die Abschätzung ist also lediglich, dass der relative *Fehler der Matrix klein genug ist.*

Die Neumannsche Reihe ist nicht nur ein sehr nützliches Hilfsmittel für die Herleitung von Fehlerabschätzungen, sondern sie kann auch für die Konstruktion konkreter Lösungsverfahren für lineare Gleichungssysteme verwendet werden. Ein besonders einfaches Beispiel ist die *Jacobi-Iteration*, die eine Folge von Näherungslösungen des linearen Gleichungssystems bestimmt und dafür lediglich Matrix-Vektor-Multiplikationen mit der Matrix $\mathbf{A}$ und der Inversen ihres Diagonalanteils benötigt.

Übungsaufgabe 3.11 (Verträgliche Matrixnormen) *Seien $\|\cdot\|_V$ und $\|\cdot\|_W$ Normen auf den Räumen $\mathcal{V}=\mathbb{K}^n$ und $\mathcal{W}=\mathbb{K}^m$. Sei $\mathbf{A}\in\mathbb{K}^{m\times n}$.*

Sei $\|\cdot\|$ eine beliebige *Norm auf dem Raum $\mathbb{K}^{m\times n}$ der Matrizen.*

Beweisen Sie: Falls

$$\|\mathbf{A}\mathbf{y}\|_W \le \|\mathbf{A}\|\,\|\mathbf{y}\|_V \qquad \text{für alle } \mathbf{y}\in\mathcal{V}$$

gilt, folgt $\|\mathbf{A}\|_{W\leftarrow V}\le\|\mathbf{A}\|$.

Die induzierte Matrixnorm ist also die „kleinste Norm“ auf $\mathbb{K}^{m\times n}$, die mit den Normen auf $\mathcal{V}$ und $\mathcal{W}$ verträglich ist.

Übungsaufgabe 3.12 (Spaltensummennorm) *Seien $\|\cdot\|_V$ und $\|\cdot\|_W$ die zu $\mathcal{V}=\mathbb{K}^n$ und $\mathcal{W}=\mathbb{K}^m$ gehörenden Summennormen (vgl. Beispiel 2.4). Beweisen Sie*

$$\|\mathbf{A}\|_{W\leftarrow V} = \|\mathbf{A}\|_1 := \max\left\{\sum_{i=1}^{m}|a_{ij}| \;:\; j\in[1:n]\right\} \qquad \text{für alle } \mathbf{A}\in\mathbb{K}^{m\times n}.$$

Übungsaufgabe 3.13 (Maximumnorm) *Sei $\|\cdot\|_V$ die zu $\mathcal{V}=\mathbb{K}^n$ gehörende Summennorm und $\|\cdot\|_W$ die zu $\mathcal{W}=\mathbb{K}^m$ gehörende Maximumnorm. Beweisen Sie*

$$\|\mathbf{A}\|_{W\leftarrow V} = \max\{|a_{ij}| \;:\; j\in[1:m],\ i\in[1:n]\} \qquad \text{für alle } \mathbf{A}\in\mathbb{K}^{m\times n}.$$

Übungsaufgabe 3.14 (Summennorm) *Sei $\|\cdot\|_V$ die zu $\mathcal{V}=\mathbb{K}^n$ gehörende Maximumnorm und $\|\cdot\|_W$ die zu $\mathcal{W}=\mathbb{K}^m$ gehörende Summennorm. Beweisen Sie*

$$\|\mathbf{A}\|_{W\leftarrow V} \le \sum_{i=1}^{m}\sum_{j=1}^{n}|a_{ij}| \qquad \text{für alle } \mathbf{A}\in\mathbb{K}^{m\times n}.$$

Geben Sie $n, m \in \mathbb{N}$ und eine Matrix $\mathbf{A} \in \mathbb{K}^{m\times n}$ an mit

$$\|\mathbf{A}\|_{W\leftarrow V} < \sum_{i=1}^{m}\sum_{j=1}^{n}|a_{ij}|.$$

Hinweis: *Es gibt eine Matrix $\mathbf{A} \in \mathbb{R}^{2\times 3}$, die das Geforderte leistet. Vorzeichen spielen dabei eine entscheidende Rolle.*

Übungsaufgabe 3.15 (Frobenius-Norm) *Beweisen Sie, dass die durch*

$$\|\mathbf{A}\|_F := \left(\sum_{i=1}^{m}\sum_{j=1}^{n}|a_{ij}|^2\right)^{1/2} \qquad \text{für alle } \mathbf{A} \in \mathbb{K}^{m\times n}$$

gegebene Frobenius-Norm *eine Norm auf $\mathbb{K}^{m\times n}$ ist, die mit der Euklidischen Norm (vgl. Beispiel 2.5) verträglich ist, dass also*

$$\|\mathbf{A}\mathbf{y}\|_2 \leq \|\mathbf{A}\|_F\|\mathbf{y}\|_2 \qquad \text{für alle } \mathbf{A} \in \mathbb{K}^{m\times n},\ \mathbf{y} \in \mathbb{K}^n$$

gilt. Sei $\|\cdot\|_V$ eine beliebige Norm auf $\mathcal{V} = \mathbb{K}^n$. Beweisen Sie, dass die Frobenius-Norm auf $\mathbb{K}^{n\times n}$ nicht *die induzierte Matrixnorm $\|\cdot\|_{V\leftarrow V}$ ist.*

Hinweis: *Die Cauchy-Schwarz-Ungleichung ist nützlich. Für den letzten Aufgabenteil könnte ein Blick auf die Einheitsmatrix helfen.*

Übungsaufgabe 3.16 (Jacobi-Iteration) *Eine Matrix $\mathbf{A} \in \mathbb{K}^{n\times n}$ heißt* (streng) diagonaldominant*, falls*

$$\sum_{\substack{j=1\\ j\neq i}}^{n}|a_{ij}| < |a_{ii}| \qquad \text{für alle } i \in [1:n]$$

gilt. Sei $\mathbf{A} \in \mathbb{K}^{n\times n}$ nun streng diagonaldominant, und sei $\mathbf{D} \in \mathbb{K}^{n\times n}$ ihr Diagonalanteil, gegeben durch

$$d_{ij} := \begin{cases} a_{ii} & \text{falls } i = j,\\ 0 & \text{ansonsten} \end{cases} \qquad \text{für alle } i, j \in [1:n].$$

Beweisen Sie, dass $\mathbf{D}$ invertierbar ist und dass die Matrix $\mathbf{M} := \mathbf{I} - \mathbf{D}^{-1}\mathbf{A}$ die Ungleichung $\|\mathbf{M}\|_\infty < 1$ erfüllt. Dafür könnte Lemma 3.4 von Nutzen sein.

Folgern Sie daraus, dass streng diagonaldominante Matrizen invertierbar sind.

Die Jacobi-Iteration *für einen Anfangsvektor $\mathbf{x}^{(0)} \in \mathbb{K}^n$ berechnet durch*

$$\mathbf{x}^{(m+1)} := \mathbf{x}^{(m)} - \mathbf{D}^{-1}(\mathbf{A}\mathbf{x}^{(m)} - \mathbf{b}) \qquad \text{für alle } m \in \mathbb{N}_0$$

eine Folge von Näherungslösungen für das lineare Gleichungssystem $\mathbf{A}\mathbf{x} = \mathbf{b}$. Zeigen Sie, dass die Folge $(\mathbf{x}^{(m)})_{m=0}^{\infty}$ gegen die Lösung $\mathbf{x}$ konvergiert, falls $\mathbf{A}$ streng diagonaldominant ist. Dafür könnte man sich beispielsweise anschauen, wie sich der Fehlervektor $\mathbf{e}^{(m)} := \mathbf{x}^{(m)} - \mathbf{x}$ während der Iteration entwickelt.

3.2 Dreiecksmatrizen

Bevor wir uns der Behandlung allgemeiner linearer Gleichungssysteme zuwenden konzentrieren wir uns zunächst auf einen Spezialfall.

Definition 3.17 (Dreiecksmatrizen) *Sei* $\mathbf{A} \in \mathbb{K}^{n\times n}$ *eine Matrix.*

Wir bezeichnen $\mathbf{A}$ *als* untere Dreiecksmatrix*, falls*

$$a_{ij} = 0 \qquad \text{für alle } 1 \leq i < j \leq n$$

gilt, falls also alle Einträge oberhalb der Diagonalen der Matrix gleich null sind.

Wir bezeichnen $\mathbf{A}$ *als* obere Dreiecksmatrix*, falls*

$$a_{ij} = 0 \qquad \text{für alle } 1 \leq j < i \leq n$$

gilt, falls also alle Einträge unterhalb der Diagonalen der Matrix gleich null sind.

Falls $\mathbf{A}$ *eine Dreiecksmatrix ist, deren Diagonalelemente alle gleich eins sind, nennen wir sie* normiert*.*

Der Name „Dreiecksmatrix“ ergibt sich aus der Struktur der Matrix: Mit der Konvention, Nulleinträge in der Matrix weglassen zu dürfen, erhalten wir für untere beziehungsweise obere Dreiecksmatrizen die Darstellungen

$$\begin{pmatrix} a_{11} & & & \\ a_{21} & a_{22} & & \\ \vdots & \vdots & \ddots & \\ a_{n1} & a_{n2} & \dots & a_{nn} \end{pmatrix} \quad \text{und} \quad \begin{pmatrix} a_{11} & a_{12} & \dots & a_{1n} \\ & a_{22} & \dots & a_{2n} \\ & & \ddots & \vdots \\ & & & a_{nn} \end{pmatrix},$$

die die Bezeichnung motivieren: Nur das linke untere beziehungsweise rechte obere Dreieck der jeweiligen Matrix ist gefüllt.

Einige wichtige Berechnungen lassen sich für Dreiecksmatrizen besonders einfach durchführen. Als Beispiel untersuchen wir die Berechnung der Determinante.

Lemma 3.18 (Determinante) *Sei* $\mathbf{A} \in \mathbb{K}^{n\times n}$ *eine obere oder untere Dreiecksmatrix. Dann gilt*

$$\det(\mathbf{A}) = \prod_{i=1}^{n} a_{ii}.$$

Beweis. Wir betrachten zunächst den Fall, dass $\mathbf{A}$ eine obere Dreiecksmatrix ist und führen den Beweis per Induktion.

Induktionsanfang: Für $n = 1$ ist die Aussage trivial.

Induktionsvoraussetzung: Sei nun $n \in \mathbb{N}$ so gewählt, dass die Gleichung für alle oberen Dreiecksmatrizen $\mathbf{R} \in \mathbb{K}^{n\times n}$ gilt.

Induktionsschritt: Sei $\mathbf{R} \in \mathbb{K}^{(n+1)\times(n+1)}$ eine obere Dreiecksmatrix. Wir wenden den Entwicklungssatz von Laplace auf die erste Spalte der Matrix an: Nach Voraussetzung hat die Matrix $\mathbf{R}$ die Form

$$\mathbf{R} = \left(\begin{array}{c|ccc} r_{11} & r_{12} & \cdots & r_{1,n+1} \\ \hline & r_{22} & \cdots & r_{2,n+1} \\ & & \ddots & \vdots \\ & & & r_{n+1,n+1} \end{array}\right),$$

und bei Entwicklung nach der ersten Spalte erhalten wir

$$\det(\mathbf{R}) = r_{11} \det \begin{pmatrix} r_{22} & \cdots & r_{2,n+1} \\ & \ddots & \vdots \\ & & r_{n+1,n+1} \end{pmatrix} = r_{11} \det(\widehat{\mathbf{R}})$$

für die Teilmatrix

$$\widehat{\mathbf{R}} = \begin{pmatrix} r_{22} & \cdots & r_{2,n+1} \\ & \ddots & \vdots \\ & & r_{n+1,n+1} \end{pmatrix} \in \mathbb{K}^{n\times n}.$$

Wir wenden die Induktionsvoraussetzung auf diese Matrix an und erhalten

$$\det(\mathbf{R}) = r_{11} \det(\widehat{\mathbf{R}}) = r_{11} \prod_{i=2}^{n+1} r_{ii} = \prod_{i=1}^{n+1} r_{ii},$$

und damit ist der Induktionsschritt bewiesen.

Für untere Dreiecksmatrizen verfahren wir entsprechend, nur dass wir nach der ersten Zeile statt der ersten Spalte entwickeln. ∎

Bekanntlich ist eine Matrix genau dann invertierbar, wenn ihre Determinante ungleich null ist. Für Dreiecksmatrizen folgt daraus, dass sie genau dann invertierbar sind, wenn alle Diagonalelemente ungleich null sind. Das lässt sich auch direkt konstruktiv beweisen:

Lemma 3.19 (Invertierbarkeit) *Sei $\mathbf{A} \in \mathbb{K}^{n\times n}$ eine obere oder untere Dreiecksmatrix. Sie ist genau dann invertierbar, wenn alle Diagonalelemente ungleich null sind.*

Beweis. Wir beschränken uns auf den Fall oberer Dreiecksmatrizen, für untere Dreiecksmatrizen entsprechend vorgegangen werden.

Den Beweis führen wir per Induktion über $n \in \mathbb{N}$.

Induktionsanfang: Sei $n = 1$. Eine Matrix $\mathbf{R} \in \mathbb{K}^{1\times 1}$ is genau dann invertierbar wenn $\mathbf{R} \neq \mathbf{0}$ gilt.

Induktionsvoraussetzung: Sei $n \in \mathbb{N}$ so gewählt, dass jede obere Dreiecksmatrix $\mathbf{R} \in \mathbb{K}^{n\times n}$ genau dann invertierbar ist, wenn alle Diagonalelemente ungleich null sind.

Induktionsschritt: Sei $\mathbf{R} \in \mathbb{K}^{(n+1)\times(n+1)}$ eine obere Dreiecksmatrix. Wir definieren

$$\mathbf{R}_{**} := \begin{pmatrix} r_{11} & \cdots & r_{1n} \\ & \ddots & \vdots \\ & & r_{nn} \end{pmatrix}, \qquad \mathbf{R}_{*,n+1} := \begin{pmatrix} r_{1,n+1} \\ \vdots \\ r_{n,n+1} \end{pmatrix}$$

und erhalten die Blockdarstellung

$$\mathbf{R} = \left(\begin{array}{ccc|c} r_{11} & \cdots & r_{1n} & r_{1,n+1} \\ & \ddots & \vdots & \vdots \\ & & r_{nn} & r_{n,n+1} \\ \hline & & & r_{n+1,n+1} \end{array}\right) = \begin{pmatrix} \mathbf{R}_{**} & \mathbf{R}_{*,n+1} \\ & r_{n+1,n+1} \end{pmatrix}.$$

Seien $\mathbf{b}, \mathbf{x} \in \mathbb{K}^{n+1}$ gegeben. Wir zerlegen die Vektoren in

$$\mathbf{b} = \begin{pmatrix} \mathbf{b}_* \\ b_{n+1} \end{pmatrix}, \qquad \mathbf{x} = \begin{pmatrix} \mathbf{x}_* \\ x_{n+1} \end{pmatrix}$$

und stellen fest, dass $\mathbf{b} = \mathbf{R}\mathbf{x}$ genau dann gilt, wenn

$$\begin{pmatrix} \mathbf{b}_* \\ b_{n+1} \end{pmatrix} = \begin{pmatrix} \mathbf{R}_{**} & \mathbf{R}_{*,n+1} \\ & r_{n+1,n+1} \end{pmatrix} \begin{pmatrix} \mathbf{x}_* \\ x_{n+1} \end{pmatrix} = \begin{pmatrix} \mathbf{R}_{**}\mathbf{x}_* + \mathbf{R}_{*,n+1}x_{n+1} \\ r_{n+1,n+1}x_{n+1} \end{pmatrix}$$

gilt, also

$$\mathbf{R}_{**}\mathbf{x}_* = \mathbf{b}_* - \mathbf{R}_{*,n+1}x_{n+1}, \qquad r_{n+1,n+1}x_{n+1} = b_{n+1}. \tag{3.5}$$

Sei zunächst $\mathbf{R}$ als invertierbar vorausgesetzt, wir gehen also davon aus, dass für jedes $\mathbf{b} \in \mathbb{K}^{n+1}$ ein $\mathbf{x} \in \mathbb{K}^{n+1}$ mit $\mathbf{R}\mathbf{x} = \mathbf{b}$ existiert. Indem wir $b_{n+1} = 1$ einsetzen, folgt aus der zweiten Gleichung in (3.5) unmittelbar $r_{n+1,n+1} \neq 0$. Indem wir $b_{n+1} = 0$ setzen, folgt $x_{n+1} = 0$ aus der zweiten Gleichung und aus der ersten, dass für jedes $\mathbf{b}_* \in \mathbb{K}^n$ ein $\mathbf{x}_*$ mit $\mathbf{R}_{**}\mathbf{x}_* = \mathbf{b}_*$ existiert, dass also $\mathbf{R}_{**}$ invertierbar ist. Nach Induktionsvoraussetzung müssen dann auch die Diagonalelemente $r_{11}, r_{22}, \ldots, r_{nn}$ ungleich null sein.

Seien nun umgekehrt alle Diagonalelemente als ungleich null vorausgesetzt. Für ein beliebiges $\mathbf{b} \in \mathbb{K}^{n+1}$ können wir dann die zweite Gleichung in (3.5) erfüllen, indem wir $x_{n+1} = b_{n+1}/r_{n+1,n+1}$ setzen. Nach Induktionsvoraussetzung ist $\mathbf{R}_{**}$ invertierbar, so dass wir $\mathbf{x}_*$ als Lösung des ersten Gleichungssystems $\mathbf{R}_{**}\mathbf{x}_* = \mathbf{b}_* - \mathbf{R}_{*,n+1}x_{n+1}$ gewinnen können. ∎

Die im Beweis des Lemmas 3.19 verwendete Konstruktion der Lösung des Gleichungssystems lässt sich konkret in einer Programmiersprache umsetzen. Im Prinzip könnten wir die Berechnung exakt wie in der Theorie durchführen, also die Teilmatrizen $\mathbf{R}_{**}$ und $\mathbf{R}_{*,n+1}$ sowie den Vektor $\mathbf{b}_*$ konstruieren und rekursiv lösen. Das würde allerdings bedeuten, dass sehr viele Zwischenergebnisse anfallen, die gespeichert werden müssten.

Sehr viel effizienter ist es, die Matrizen und Vektoren *implizit* zu verwenden: Die Matrizen $\mathbf{R}_{**}$ und $\mathbf{R}_{*,n+1}$ sind lediglich Teile der Ausgangsmatrix $\mathbf{R}$, also können wir ihre Koeffizienten leicht bestimmen.

Die Komponente b_{n+1} der rechten Seite wird lediglich einmal zur Berechnung der Komponente x_{n+1} der Lösung verwendet, für den Rest des Algorithmus ist sie dann nicht mehr nötig. Deshalb bietet es sich an, den von b_{n+1} belegten Speicherplatz mit dem Wert x_{n+1} zu überschreiben, sobald er berechnet wurde.

Entsprechend ist $\mathbf{b}_*$ für die Berechnung des Vektors $\mathbf{x}_*$ nicht erforderlich, wir benötigen lediglich den Vektor $\mathbf{b}_* - \mathbf{R}_{*,n+1}x_{n+1}$, können also auch in diesem Fall die ursprüngliche

```
procedure rec_back_subst(R, n, var b);
b_n ← b_n/r_nn
if n > 1 then
  for i ∈ [1 : n − 1] do
    b_i ← b_i − r_in b_n
  end for
  rec_back_subst(R, n − 1, b)
end if
```

Abbildung 3.1: Rekursives Rückwärtseinsetzen zur Lösung von $\mathbf{Rx} = \mathbf{b}$. Die rechte Seite $\mathbf{b}$ wird mit der Lösung $\mathbf{x}$ überschrieben.

rechte Seite mit diesem Hilfsvektor überschreiben. Abbildung 3.1 stellt diese Berechnung in Pseudo-Code dar.

Dieser Algorithmus arbeitet *rekursiv*: Um ein Problem der Größe $n > 1$ zu lösen wird ein Problem der Größe $n - 1$ konstruiert und der Algorithmus auf dieses Teilproblem angewendet. Das entspricht der mathematischen Herleitung der Rechenvorschrift über eine vollständige Induktion, hat aber den Nachteil, dass in üblichen Programmiersprachen zusätzlicher Verwaltungsaufwand anfällt, den wir möglichst vermeiden sollten. In unserem Fall ist das relativ einfach möglich, da wir von Anfang an wissen, in welcher Reihenfolge die rekursiven Aufrufe stattfinden werden: Zunächst mit n, dann mit $n-1$, und so weiter bis wir 1 erreichen und x_1 berechnen. Indem wir die gerade aktuelle Dimension mit k bezeichnen gelangen wir zu dem in Abbildung 3.2 dargestellten Algorithmus, der ohne Rekursion auskommt.

```
procedure back_subst(R, n, var b);
for k = n, ..., 1 do
  b_k ← b_k/r_kk
  for i ∈ [1 : k − 1] do
    b_i ← b_i − r_ik b_k
  end for
end for
```

Abbildung 3.2: Rückwärtseinsetzen zur Lösung von $\mathbf{Rx} = \mathbf{b}$. Die rechte Seite $\mathbf{b}$ wird mit der Lösung $\mathbf{x}$ überschrieben.

Bei dieser Variante des Algorithmus ist wichtig, dass die äußere Schleife über k die Werte $n, n-1, \ldots, 1$ in der festgelegten Reihenfolge abarbeiten muss, während die innere Schleife über i die verschiedenen Werte einer beliebigen Reihenfolge aktualisieren kann. Das bedeutet, dass die Reihenfolge so arrangiert werden kann, dass die Berechnungen besonders effizient ablaufen. Beispielsweise könnten sie auf einem Parallelrechner sogar auf mehrere Prozessoren verteilt werden, um den Zeitbedarf zu reduzieren.

Da der Algorithmus die Komponenten $x_n, \ldots, x_1$ des Ergebnisvektors in umge-

kehrter Reihenfolge berechnet und für die Berechnung von x_j die bereits bekannten Werte $x_{j+1}, \ldots, x_n$ in die Matrix einsetzt, wird er üblicherweise mit dem Namen *Rückwärtseinsetzen* (engl. *backward substitution*) bezeichnet.

Für reguläre untere Dreiecksmatrizen $\mathbf{L} \in \mathbb{K}^{n\times n}$ können wir entsprechend vorgehen: Wenn wir zu $\mathbf{b} \in \mathbb{K}^n$ einen Vektor $\mathbf{x} \in \mathbb{K}^n$ mit

$$\mathbf{L}\mathbf{x} = \mathbf{b}$$

suchen, können wir

$$\mathbf{L}_{**} = \begin{pmatrix} l_{22} & & \\ \vdots & \ddots & \\ l_{n2} & \ldots & l_{nn} \end{pmatrix}, \qquad \mathbf{L}_{*1} = \begin{pmatrix} l_{21} \\ \vdots \\ l_{n1} \end{pmatrix}, \qquad \mathbf{x}_* := \begin{pmatrix} x_2 \\ \vdots \\ x_n \end{pmatrix}, \qquad \mathbf{b}_* := \begin{pmatrix} b_2 \\ \vdots \\ b_n \end{pmatrix}$$

einführen und das System in der Form

$$\begin{pmatrix} l_{11} & \\ \mathbf{L}_{*1} & \mathbf{L}_{**} \end{pmatrix} \begin{pmatrix} x_1 \\ \mathbf{x}_* \end{pmatrix} = \left(\begin{array}{c|ccc} l_{11} & & & \\ \hline l_{21} & l_{22} & & \\ \vdots & \vdots & \ddots & \\ l_{n1} & l_{n2} & \ldots & l_{nn} \end{array}\right) \left(\begin{array}{c} x_1 \\ \hline x_2 \\ \vdots \\ x_n \end{array}\right) = \left(\begin{array}{c} b_1 \\ \hline b_2 \\ \vdots \\ b_n \end{array}\right) = \begin{pmatrix} b_1 \\ \mathbf{b}_* \end{pmatrix}$$

schreiben. Diese System ist äquivalent zu den Gleichungen

$$l_{11}x_1 = b_1, \qquad\qquad \mathbf{L}_{*1}x_1 + \mathbf{L}_{**}\mathbf{x}_* = \mathbf{b}_*,$$

aus denen wir sofort $x_1 = b_1/l_{11}$ gewinnen und die zweite Gleichung in die Form

$$\mathbf{L}_{**}\mathbf{x}_* = \mathbf{b}_* - \mathbf{L}_{*1}x_1$$

überführen können. Für die Bestimmung der restlichen Komponenten muss also ein Gleichungssystem der reduzierten Dimension $n-1$ gelöst werden. Damit steht uns auch für reguläre untere Dreiecksmatrizen ein praktisches Lösungsverfahren zur Verfügung, das wir als Algorithmus in Abbildung 3.3 zusammenfassen. Aus naheliegenden Gründen hat sich der Name *Vorwärtseinsetzen* (engl. *forward substitution*) für diese Rechenvorschrift eingebürgert.

Natürlich wollen wir nicht nur ein Gleichungssystem lösen, sondern wir wollen es *schnell* lösen. Die „Geschwindigkeit“ eines Algorithmus wird in der Numerik üblicherweise an der Anzahl der arithmetischen Gleitkommaoperationen gemessen, da sie sich einfach abschätzen lässt und die Gesamtzahl der Operationen in der Regel zu ihr proportional ist.

Wieviele Gleitkommaoperationen benötigt nun das Rückwärtseinsetzen gemäß Abbildung 3.2? Der Rumpf der innersten Schleife (über i) benötigt zwei Operationen, nämlich die Multiplikation von r_{ij} mit b_j und die Subtraktion des Ergebnisses von b_i.

Im Rumpf der äußeren Schleife wird die innerste Schleife $n-k$ mal durchlaufen, so dass insgesamt $2(n-k)$ Operationen anfallen, hinzu kommt noch die Division von b_k durch r_{kk}, also benötigt dieser Rumpf $2(n-k)+1$ Operationen.

```
procedure for_subst(L, n, var b);
for k = 1, ..., n do
  b_k ← b_k / l_kk
  for i ∈ [k + 1 : n] do
    b_i ← b_i − l_ik b_k
  end for
end for
```

Abbildung 3.3: Vorwärtseinsetzen zur Berechnung von $\mathbf{x}$ aus $\mathbf{Lx} = \mathbf{b}$. Die rechte Seite $\mathbf{b}$ wird mit der Lösung $\mathbf{x}$ überschrieben.

Für die äußere Schleife fallen also

$$\sum_{k=1}^{n} (2(n-k)+1) = \sum_{\ell=0}^{n-1} (2\ell+1) = n + 2\sum_{\ell=0}^{n-1} \ell = n + n(n-1) = n^2$$

Operationen an, und das ist der Rechenaufwand des Rückwärtseinsetzens. Mit denselben Argumenten können wir zeigen, dass auch das Vorwärtseinsetzen in Algorithms 3.3 gerade n^2 Operationen benötigt.

Übungsaufgabe 3.20 (Gauß-Seidel-Iteration) *Sei $\mathbf{A} \in \mathbb{K}^{n\times n}$ streng diagonaldominant. Wir definieren Matrizen $\mathbf{D}, \mathbf{E}, \mathbf{F} \in \mathbb{K}^{n\times n}$ durch*

$$e_{ij} := \begin{cases} a_{ij} & \text{falls } j < i, \\ 0 & \text{ansonsten,} \end{cases} \qquad f_{ij} := \begin{cases} a_{ij} & \text{falls } j > i, \\ 0 & \text{ansonsten,} \end{cases}$$

$$d_{ij} := \begin{cases} a_{ii} & \text{falls } j = i, \\ 0 & \text{ansonsten,} \end{cases} \qquad \text{für alle } i, j \in [1:n].$$

Dann gilt $\mathbf{A} = \mathbf{D} + \mathbf{E} + \mathbf{F}$ und $\mathbf{D} + \mathbf{E}$ ist eine untere Dreiecksmatrix.

Zeigen Sie, dass $\mathbf{D} + \mathbf{E}$ invertierbar ist.

Die Gauß-Seidel-Iteration *für einen Anfangsvektor $\mathbf{x}^{(0)} \in \mathbb{K}^n$ berechnet durch*

$$\mathbf{x}^{(m+1)} = \mathbf{x}^{(m)} - (\mathbf{D}+\mathbf{E})^{-1}(\mathbf{A}\mathbf{x}^{(m)} - \mathbf{b}) \qquad \text{für alle } m \in \mathbb{N}_0$$

eine Folge von Näherungslösungen für das lineare Gleichungssystem $\mathbf{Ax} = \mathbf{b}$.

Zeigen Sie, dass

$$\begin{aligned} (\mathbf{D}+\mathbf{E})\mathbf{x}^{(m+1)} &= \mathbf{b} - \mathbf{F}\mathbf{x}^{(m)}, \\ \mathbf{D}\mathbf{x}^{(m+1)} &= \mathbf{b} - \mathbf{F}\mathbf{x}^{(m)} - \mathbf{E}\mathbf{x}^{(m+1)} \qquad \text{für alle } m \in \mathbb{N}_0 \end{aligned}$$

gilt, so dass sich $\mathbf{x}^{(m+1)}$ durch Vorwärtseinsetzen berechnen lässt.

Zeigen Sie, dass die durch $\mathbf{e}^{(m)} := \mathbf{x}^{(m)} - \mathbf{x}$ für alle $m \in \mathbb{N}_0$ definierten Fehlervektoren die Gleichung

$$\mathbf{D}\mathbf{e}^{(m+1)} = -\mathbf{F}\mathbf{e}^{(m)} - \mathbf{E}\mathbf{e}^{(m+1)} \qquad \text{für alle } m \in \mathbb{N}_0$$

erfüllen. Folgern Sie daraus, dass

$$\|\mathbf{e}^{(m+1)}\|_\infty \leq \varrho \|\mathbf{e}^{(m)}\|_\infty \qquad \text{für alle } m \in \mathbb{N}_0$$

gilt mit

$$\varrho := \max\left\{ \sum_{\substack{j=1 \\ j\neq i}}^{n} \frac{|a_{ij}|}{|a_{ii}|} \ : \ i \in [1:n] \right\} < 1,$$

dass also die Vektoren $\mathbf{x}^{(m)}$ *gegen die Lösung* $\mathbf{x}$ *konvergieren.*

3.3 LR-Zerlegung

Wir haben bereits gesehen, dass sich das Auflösen eines linearen Gleichungssystems, bei dem die Matrix Dreiecksgestalt besitzt, besonders einfach gestaltet. Für die Behandlung eines allgemeinen Gleichungssystems wäre es sinnvoll, wenn wir die Matrix in obere Dreiecksgestalt überführen könnten, ohne die Lösung zu verändern.

Am einfachsten lässt sich dieses Ziel mit der *Gauß-Elimination* erreichen: Als Beispiel untersuchen wir das zweidimensionale System

$$\begin{pmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \begin{pmatrix} b_1 \\ b_2 \end{pmatrix}.$$

Es entspricht den Gleichungen

$$\begin{aligned} a_{11}x_1 + a_{12}x_2 &= b_1, \\ a_{21}x_1 + a_{22}x_2 &= b_2. \end{aligned}$$

Wir wissen, dass sich an der Lösung des Systems nichts ändert, wenn wir eine Zeile mit einem von null verschiedenen Faktor multiplizieren. Wir wissen ebenfalls, dass sich die Lösung nicht verändert, wenn wir die Gleichungen addieren oder subtrahieren.

Unser Ziel ist es, die Matrix in eine obere Dreiecksgestalt zu bringen, wir müssen also den Eintrag a_{21} eliminieren. Falls $a_{11} \neq 0$ gilt, können wir das tun, indem wir die erste Zeile mit dem Faktor

$$l_{21} := \frac{a_{21}}{a_{11}}$$

multiplizieren und von der zweiten Zeile subtrahieren, denn dann erhalten wir wegen der Gleichung $a_{21} - l_{21}a_{11} = 0$ gerade

$$\begin{aligned} a_{11}x_1 + a_{12}x_2 &= b_1, \\ (a_{22} - l_{21}a_{12})\, x_2 &= b_2 - l_{21}b_1. \end{aligned}$$

In der zweiten Gleichung tritt x_1 nicht mehr auf, also können wir dieses System nun per Rückwärtseinsetzen lösen.

In Matrixschreibweise lässt sich unsere Transformation des Systems durch

$$\begin{pmatrix} a_{11} & a_{12} \\ & a_{22} - l_{21}a_{12} \end{pmatrix} = \begin{pmatrix} 1 & \\ -l_{21} & 1 \end{pmatrix} \begin{pmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{pmatrix}$$

darstellen, wir haben also die Matrix und die rechte Seite des ursprünglichen Gleichungssystems mit einer normierten unteren Dreiecksmatrix multipliziert, um sie auf obere Dreiecksgestalt zu bringen.

Wir bezeichnen die Einträge der oberen Dreiecksmatrix mit

$$r_{11} = a_{11}, \qquad r_{12} = a_{12}, \qquad r_{22} = a_{22} - l_{21}a_{12}$$

und bringen die Inverse der unteren Dreiecksmatrix auf die linke Seite der obigen Gleichung, um die Darstellung

$$\begin{pmatrix} r_{11} & r_{12} \\ & r_{22} \end{pmatrix} = \begin{pmatrix} 1 & \\ -l_{21} & 1 \end{pmatrix} \begin{pmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{pmatrix},$$
$$\begin{pmatrix} 1 & \\ l_{21} & 1 \end{pmatrix} \begin{pmatrix} r_{11} & r_{12} \\ & r_{22} \end{pmatrix} = \begin{pmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{pmatrix}$$

zu erhalten. Wir können unsere Transformation der Matrix auf obere Dreiecksgestalt also auch als eine *Zerlegung* der Matrix in ein Produkt aus einer unteren und einer oberen Dreiecksmatrix interpretieren.

Definition 3.21 (LR-Zerlegung) *Sei $\mathbf{A} \in \mathbb{K}^{n\times n}$ eine Matrix. Falls eine untere Dreiecksmatrix $\mathbf{L} \in \mathbb{K}^{n\times n}$ und eine obere Dreiecksmatrix $\mathbf{R} \in \mathbb{K}^{n\times n}$ die Gleichung*

$$\mathbf{LR} = \mathbf{A}$$

erfüllen, bezeichnen wir das Paar $(\mathbf{L}, \mathbf{R})$ als eine LR-Zerlegung *der Matrix $\mathbf{A}$.*

Bevor wir uns mit der oben bereits angedeuteten konkreten Konstruktion einer LR-Zerlegung beschäftigen, untersuchen wir zunächst, wie wir mit Hilfe einer derartigen Zerlegung Probleme lösen können.

Falls $(\mathbf{L}, \mathbf{R})$ eine LR-Zerlegung der Matrix $\mathbf{A}$ ist, können wir das Gleichungssystem $\mathbf{Ax} = \mathbf{b}$ lösen, indem wir

$$\mathbf{b} = \mathbf{Ax} = \mathbf{LRx} = \mathbf{Ly}, \qquad \mathbf{y} = \mathbf{Rx}$$

auflösen, also zunächst das Zwischenergebnis $\mathbf{y}$ als Lösung der Gleichung $\mathbf{Ly} = \mathbf{b}$ in unterer Dreiecksgestalt durch Vorwärtseinsetzen (siehe Abbildung 3.3) bestimmen und dann das Endergebnis $\mathbf{x}$ als Lösung der Gleichung $\mathbf{Rx} = \mathbf{y}$ in oberer Dreiecksgestalt durch Rückwärtseinsetzen (siehe Abbildung 3.2) gewinnen.

Die LR-Zerlegung lässt sich auch für andere Zwecke einsetzen, beispielsweise können wir mit Hilfe des Determinanten-Multiplikationssatzes auch die Determinante von $\mathbf{A}$ dank Lemma 3.18 gemäß

$$\det(\mathbf{A}) = \det(\mathbf{LR}) = \det(\mathbf{L})\det(\mathbf{R}) = \left(\prod_{i=1}^{n} l_{ii}\right)\left(\prod_{j=1}^{n} r_{jj}\right)$$

sehr effizient berechnen, sofern uns eine LR-Zerlegung zur Verfügung steht.

Nachdem wir nun wissen, dass eine LR-Zerlegung sehr nützlich sein kann, stellt sich natürlich die Frage, ob und, wenn ja, wie wir sie berechnen können. Theoretisch können wir die Zerlegung konstruieren, indem wir die definierende Gleichung $\mathbf{A} = \mathbf{LR}$ aufstellen und uns ihre einzelnen Komponenten

$$a_{ij} = (\mathbf{LR})_{ij} = \sum_{k=1}^{n} l_{ik} r_{kj} = \sum_{k=1}^{\min\{i,j\}} l_{ik} r_{kj}$$

anschauen (im letzten Schritt wurde ausgenutzt, dass $\mathbf{L}$ und $\mathbf{R}$ Dreiecksmatrizen sind, siehe Definition 3.17) und ähnlich wie im Fall des Vorwärtseinsetzens vorgehen. Auf diese Weise erhält man zwar einen Algorithmus zur Berechnung der Zerlegung, aber die zugrundeliegende Idee wird nicht deutlich erkennbar.

Stattdessen verwenden wir eine induktive Vorgehensweise, die einerseits nützliche zusätzliche Informationen bereitstellt und andererseits auch eine größere Nähe zu praktischen Algorithmen aufweist: Wir zerlegen die n-dimensionalen Matrizen $\mathbf{A}$, $\mathbf{L}$ und $\mathbf{R}$ in $(n-1)$-dimensionale Matrizen und einen Rest und gewinnen aus der definierenden Gleichung eine Induktionsvorschrift: Für

$$\mathbf{A}_{**} := \begin{pmatrix} a_{22} & \dots & a_{2n} \\ \vdots & \ddots & \vdots \\ a_{n2} & \dots & a_{nn} \end{pmatrix}, \qquad \mathbf{A}_{1*} := \begin{pmatrix} a_{12} & \dots & a_{1n} \end{pmatrix}, \qquad \mathbf{A}_{*1} := \begin{pmatrix} a_{21} \\ \vdots \\ a_{n1} \end{pmatrix},$$

$$\mathbf{L}_{**} := \begin{pmatrix} l_{22} & & \\ \vdots & \ddots & \\ l_{n2} & \dots & l_{nn} \end{pmatrix}, \qquad \mathbf{L}_{*1} := \begin{pmatrix} l_{21} \\ \vdots \\ l_{n1} \end{pmatrix},$$

$$\mathbf{R}_{**} := \begin{pmatrix} r_{22} & \dots & r_{2n} \\ & \ddots & \vdots \\ & & r_{nn} \end{pmatrix}, \qquad \mathbf{R}_{1*} := \begin{pmatrix} r_{12} & \dots & r_{1n} \end{pmatrix},$$

erhalten wir die Gleichungen

$$\begin{pmatrix} a_{11} & \mathbf{A}_{1*} \\ \mathbf{A}_{*1} & \mathbf{A}_{**} \end{pmatrix} = \left(\begin{array}{c|ccc} a_{11} & a_{12} & \dots & a_{1n} \\ \hline a_{21} & a_{22} & \dots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{n1} & a_{n2} & \dots & a_{nn} \end{array}\right) = \mathbf{A},$$

$$\begin{pmatrix} l_{11} & \\ \mathbf{L}_{*1} & \mathbf{L}_{**} \end{pmatrix} = \left(\begin{array}{c|ccc} l_{11} & & & \\ \hline l_{21} & l_{22} & & \\ \vdots & \vdots & \ddots & \\ l_{n1} & l_{n2} & \dots & l_{nn} \end{array}\right) = \mathbf{L},$$

$$\begin{pmatrix} r_{11} & \mathbf{R}_{1*} \\ & \mathbf{R}_{**} \end{pmatrix} = \left(\begin{array}{c|ccc} r_{11} & r_{12} & \dots & r_{1n} \\ \hline & r_{22} & \dots & r_{2n} \\ & & \ddots & \vdots \\ & & & r_{nn} \end{array}\right) = \mathbf{R},$$

und damit aus der definierenden Gleichung der LR-Zerlegung die Beziehung

$$\begin{pmatrix} a_{11} & \mathbf{A}_{1*} \\ \mathbf{A}_{*1} & \mathbf{A}_{**} \end{pmatrix} = \mathbf{A} = \mathbf{LR} = \begin{pmatrix} l_{11} & \\ \mathbf{L}_{*1} & \mathbf{L}_{**} \end{pmatrix} \begin{pmatrix} r_{11} & \mathbf{R}_{1*} \\ & \mathbf{R}_{**} \end{pmatrix}.$$

Diese Blockgleichung ist äquivalent mit

$$\begin{aligned} a_{11} &= l_{11} r_{11}, & \mathbf{A}_{1*} &= l_{11} \mathbf{R}_{1*}, \\ \mathbf{A}_{*1} &= \mathbf{L}_{*1} r_{11}, & \mathbf{A}_{**} &= \mathbf{L}_{*1} \mathbf{R}_{1*} + \mathbf{L}_{**} \mathbf{R}_{**}, \end{aligned}$$

also einem System von Matrixgleichungen, das wir von oben nach unten auflösen können, falls $a_{11} \neq 0$ gilt und wir ein $l_{11} \neq 0$ geeignet wählen:

$$\begin{gathered} r_{11} = a_{11}/l_{11}, \\ \mathbf{R}_{1*} = \frac{1}{l_{11}} \mathbf{A}_{1*}, \qquad \mathbf{L}_{*1} = \mathbf{A}_{*1} \frac{1}{r_{11}}, \\ \mathbf{L}_{**} \mathbf{R}_{**} = \widehat{\mathbf{A}} := \mathbf{A}_{**} - \mathbf{L}_{*1} \mathbf{R}_{1*}. \end{gathered}$$

In der Praxis verwendet man in der Regel $l_{11} = 1$ und $r_{11} = a_{11}$, dann wird eine LR-Zerlegung mit einer normierten Matrix $\mathbf{L}$ berechnet.

Bei der praktischen Umsetzung des Algorithmus empfiehlt es sich wieder, Speicherplätze geschickt zu überschreiben und so zu einer besonders sparsamen und damit effizienten Variante zu gelangen. In diesem Fall sehen wir, dass die Matrix $\mathbf{A}_{**}$ nur einmal benötigt wird, um die Matrix $\widehat{\mathbf{A}}$ aufzustellen. Also können wir auch erstere Matrix mit letzterer überschreiben, ohne das Ergebnis der Berechnung zu verändern. Entsprechend können wir $\mathbf{A}_{1*}$ mit $\mathbf{L}_{1*}$ und $\mathbf{A}_{*1}$ mit $\mathbf{R}_{*1}$ überschreiben. Wenn wir in dieser Weise fortfahren, wird nach und nach die obere Dreieckshälfte der Matrix mit der oberen Hälfte von $\mathbf{R}$ überschrieben, während die untere mit $\mathbf{L}$ überschrieben wird:

$$\begin{pmatrix} a_{11} & \dots & \dots & a_{1n} \\ a_{21} & a_{22} & \dots & a_{2n} \\ \vdots & \ddots & \ddots & \vdots \\ a_{n1} & \dots & a_{n,n-1} & a_{nn} \end{pmatrix} \quad \rightsquigarrow \quad \begin{pmatrix} r_{11} & \dots & \dots & r_{1n} \\ l_{21} & r_{22} & \dots & r_{2n} \\ \vdots & \ddots & \ddots & \vdots \\ l_{n1} & \dots & l_{n,n-1} & r_{nn} \end{pmatrix} \tag{3.6}$$

Da beides Dreiecksmatrizen sind und auf der Diagonale von $\mathbf{L}$ immer eins steht, sind dadurch die Faktoren $\mathbf{L}$ und $\mathbf{R}$ vollständig und sehr platzsparend beschrieben, und die Berechnung kann ohne Hilfsspeicher und Rekursion wie in Abbildung 3.4 erfolgen.

Natürlich interessieren wir uns auch in diesem Fall für den Rechenaufwand des Verfahrens. Die erste innere Schleife (Berechnung von a_{ik}) benötigt $n-k$ Divisionen, die zweite innere Schleife (Berechnung von a_{ij}) erfordert $(n-k)^2$ Multiplikationen und $(n-k)^2$ Subtraktionen, so dass wir einen Gesamtaufwand von

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

Operationen für den Rumpf der äußeren Schleife erhalten. Der Gesamtaufwand ist durch

$$\sum_{k=1}^{n} (n-k)(2(n-k)+1) = \sum_{\ell=0}^{n-1} \ell(2\ell+1) = 2\sum_{\ell=0}^{n-1} \ell^2 + \sum_{\ell=0}^{n-1} \ell$$

```
procedure decomp_lr(n, var A)
for k = 1, ..., n − 1 do
  if a_kk = 0 then
    Abbruch, A besitzt keine LR-Zerlegung
  end if
  for i ∈ [k + 1 : n] do
    a_ik ← a_ik/a_kk
  end for
  for i, j ∈ [i + 1 : n] do
    a_ij ← a_ij − a_ik a_kj
  end for
end for
```

Abbildung 3.4: Berechnung der LR-Zerlegung. Die Matrix **A** wird entsprechend (3.6) mit den Koeffizienten von **L** und **R** überschrieben.

gegeben (wobei wir $\ell = n - k$ substituiert haben). Man erkennt bereits, dass wir wohl ein kubisches Wachstum in n zu erwarten haben.

Lemma 3.22 (Summenformel) *Es gilt*

$$\sum_{i=1}^{n} i^2 = \frac{n}{6}(2n+1)(n+1) \qquad \text{für alle } n \in \mathbb{N}_0.$$

Beweis. Wir beweisen die Behauptung per Induktion über n.

Induktionsanfang: Für $n = 1$ gilt

$$\frac{n}{6}(2n+1)(n+1) = \frac{1 \cdot 3 \cdot 2}{6} = 1 = \sum_{i=1}^{1} i^2,$$

also ist der Induktionsanfang bewiesen.

Induktionsvoraussetzung: Sei $n \in \mathbb{N}$ so gegeben, dass die Gleichung gilt.

Induktionsschritt: Dann folgt

$$\begin{aligned}\sum_{i=1}^{n+1} i^2 &= \sum_{i=1}^{n} i^2 + (n+1)^2 = \frac{n}{6}(2n+1)(n+1) + (n+1)^2 \\ &= \Big(\frac{n}{6}(2n+1) + (n+1)\Big)(n+1) = \frac{n(2n+1)+6n+6}{6}(n+1) \\ &= \frac{2n^2+7n+6}{6}(n+1) = \frac{(2n+3)(n+2)}{6}(n+1) = \frac{n+1}{6}(2n+3)(n+2),\end{aligned}$$

und der Induktionsschritt ist bewiesen. ∎

Dieses Lemma können wir nun verwenden, um den Aufwand der Berechnung der LR-Zerlegung zu ermitteln: Wir benötigen gerade

$$\begin{aligned} 2\sum_{\ell=0}^{n-1}\ell^2+\sum_{\ell=0}^{n-1}\ell &= \frac{2(n-1)}{6}(2(n-1)+1)(n-1+1)+\frac{n}{2}(n-1)\\ &= \frac{n(n-1)(4n-2)}{6}+\frac{3n(n-1)}{6}\\ &= \frac{n(n-1)(4n+1)}{6}=\frac{n(4n^2-3n-1)}{6}\\ &= \frac{2}{3}n^3-\frac{n}{6}(3n+1)\leq\frac{2}{3}n^3 \end{aligned} \tag{3.7}$$

Operationen, um $\mathbf{L}$ und $\mathbf{R}$ zu bestimmen.

Bemerkung 3.23 (Landau-Notation) *Falls n groß genug ist, ist der quadratische Term in (3.7) vernachlässigbar, der Rechenaufwand wächst kubisch mit n. Für solche Betrachtungen ist die* Landau-Notation *gebräuchlich: Falls $f,g:\Omega\to\mathbb{R}_{>0}$ zwei Funktionen sind, schreiben wir*

$$f\in\mathcal{O}(g)$$

genau dann, wenn

$$\sup\left\{\frac{f(x)}{g(x)}\ :\ x\in\Omega\right\}<\infty \tag{3.8}$$

gilt, wenn sich also f gleichmäßig durch ein Vielfaches von g beschränken lässt.

Falls Definitionsbereich und Variablennamen aus dem Kontext ersichtlich sind, verzichtet man häufig auf die präzise Definition der Funktionen f und g und ersetzt sie durch die Ausdrücke, mit denen sie berechnet werden können. Im Fall der LR-Zerlegung würde man also etwa

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

schreiben, weil aus dem Kontext klar ist, dass $n\in\Omega:=\mathbb{N}$ die relevante Variable ist.

In derartigen Fällen sagt man auch „der Aufwand der LR-Zerlegung ist in $\mathcal{O}(n^3)$“.

Häufig gilt die Abschätzung (3.8) nur „in der Nähe“ eines Elements $a\in\Omega$. In diesem Fall verwendet man die Notation

$$f\in\mathcal{O}(g)\quad\text{für}\quad x\to a,$$

die bedeutet, dass es eine Umgebung $\Omega_a\subseteq\Omega$ von a gibt, auf der

$$\sup\left\{\frac{f(x)}{g(x)}\ :\ x\in\Omega_a\right\}<\infty$$

gilt. Derartige Fälle werden wir allerdings erst in Kapitel 5 kennen lernen.

Bisher sind wir davon ausgegangen, dass eine LR-Zerlegung existiert. Leider ist die Existenz einer derartigen Zerlegung nicht immer garantiert, beispielsweise gilt

$$\begin{pmatrix} l_{11} & \\ l_{21} & l_{22} \end{pmatrix} \begin{pmatrix} r_{11} & r_{12} \\ & r_{22} \end{pmatrix} = \begin{pmatrix} l_{11}r_{11} & l_{11}r_{12} \\ l_{21}r_{11} & l_{21}r_{12} + l_{22}r_{22} \end{pmatrix} \neq \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix} \tag{3.9}$$

für alle Koeffizienten, denn aus $l_{11}r_{12} \neq 0$ und $l_{21}r_{11} \neq 0$ folgen $l_{11} \neq 0$ sowie $r_{11} \neq 0$ und damit auch $l_{11}r_{11} \neq 0$, also können nicht alle Komponenten der linken und rechten Matrix übereinstimmen.

Die Ursache des Problems in diesem Beispiel besteht darin, dass der obere Diagonaleintrag verschwindet.

Definition 3.24 (Hauptuntermatrix) *Sei $\mathbf{A} \in \mathbb{K}^{n\times n}$, und sei $m \in [1:n]$. Die durch*

$$\mathbf{B} := \begin{pmatrix} a_{11} & \dots & a_{1m} \\ \vdots & \ddots & \vdots \\ a_{m1} & \dots & a_{mm} \end{pmatrix}$$

definierte Matrix bezeichnen wir als die m-te Hauptuntermatrix *von $\mathbf{A}$.*

Hauptuntermatrizen sind im Kontext der LR-Zerlegung von besonderem Interesse: Falls $(\mathbf{L}, \mathbf{R})$ eine LR-Zerlegung von $\mathbf{A} \in \mathbb{K}^{n\times n}$ ist, bilden für jedes $m \in [1:n]$ die m-ten Hauptuntermatrizen von $\mathbf{L}$ und $\mathbf{R}$ jeweils eine LR-Zerlegung der m-ten Hauptuntermatrix von $\mathbf{A}$.

Satz 3.25 (Existenz einer LR-Zerlegung) *Sei $\mathbf{A} \in \mathbb{K}^{n\times n}$, und sei für alle $m \in [1 : n-1]$ die m-te Hauptuntermatrix regulär. Dann besitzt $\mathbf{A}$ eine LR-Zerlegung.*

Beweis. Per Induktion über n.

Induktionsanfang: Für $n = 1$ setzen wir $l_{11} = 1$ und $r_{11} = a_{11}$ und sind fertig.

Induktionsvoraussetzung: Sei nun $n \in \mathbb{N}$ so gewählt, dass die Behauptung für alle Matrizen aus $\mathbb{K}^{n\times n}$ gilt.

Induktionsschritt: Sei $\mathbf{A} \in \mathbb{K}^{(n+1)\times(n+1)}$ eine Matrix, deren m-te Hauptuntermatrizen für $m \in [1:n]$ regulär sind. Wir definieren die Teilmatrizen

$$\mathbf{A}_{**} := \begin{pmatrix} a_{11} & \dots & a_{1,n} \\ \vdots & \ddots & \vdots \\ a_{n1} & \dots & a_{nn} \end{pmatrix}, \quad \mathbf{A}_{n+1,*} := \begin{pmatrix} a_{n+1,1} & \dots & a_{n+1,n} \end{pmatrix}, \quad \mathbf{A}_{*,n+1} := \begin{pmatrix} a_{1,n+1} \\ \vdots \\ a_{n,n+1} \end{pmatrix},$$

$$\mathbf{L}_{**} := \begin{pmatrix} \ell_{11} & & \\ \vdots & \ddots & \\ \ell_{n1} & \dots & \ell_{nn} \end{pmatrix}, \quad \mathbf{L}_{n+1,*} := \begin{pmatrix} \ell_{n+1,1} & \dots & \ell_{n+1,n} \end{pmatrix},$$

$$\mathbf{R}_{**} := \begin{pmatrix} r_{11} & \dots & r_{1,n} \\ & \ddots & \vdots \\ & & r_{nn} \end{pmatrix}, \qquad \mathbf{R}_{*,n+1} := \begin{pmatrix} r_{1,n+1} \\ \vdots \\ r_{n,n+1} \end{pmatrix}$$

und erhalten die zu $\mathbf{LR} = \mathbf{A}$ äquivalente Gleichung

$$\begin{pmatrix} \mathbf{L}_{**} & \\ \mathbf{L}_{n+1,*} & \ell_{n+1,n+1} \end{pmatrix} \begin{pmatrix} \mathbf{R}_{**} & \mathbf{R}_{*,n+1} \\ & r_{n+1,n+1} \end{pmatrix} = \begin{pmatrix} \mathbf{A}_{**} & \mathbf{A}_{*,n+1} \\ \mathbf{A}_{n+1,*} & a_{n+1,n+1} \end{pmatrix}.$$

Also müssen wir die Gleichungen

$$\mathbf{L}_{**}\mathbf{R}_{**} = \mathbf{A}_{**}, \tag{3.10a}$$

$$\mathbf{L}_{n+1,*}\mathbf{R}_{**} = \mathbf{A}_{n+1,*}, \qquad \mathbf{L}_{**}\mathbf{R}_{*,n+1} = \mathbf{A}_{*,n+1}, \tag{3.10b}$$

$$\ell_{n+1,n+1}r_{n+1,n+1} = a_{n+1,n+1} + \mathbf{L}_{n+1,*}\mathbf{R}_{*,n+1} \tag{3.10c}$$

lösen. Aus der Induktionsvoraussetzung folgt, dass $\mathbf{A}_{**}$ eine LR-Zerlegung $(\mathbf{L}_{**}, \mathbf{R}_{**})$ besitzen muss. Da $\mathbf{A}_{**}$ gerade die n-te Hauptuntermatrix von $\mathbf{A}$ ist, ist sie nach Voraussetzung insbesondere regulär, also müssen wegen $\mathbf{A}_{**} = \mathbf{L}_{**}\mathbf{R}_{**}$ auch die Matrizen $\mathbf{L}_{**}$ und $\mathbf{R}_{**}$ regulär sein. Wir können also

$$\begin{aligned} \mathbf{L}_{n+1,*} &:= \mathbf{A}_{n+1,*}\mathbf{R}_{**}^{-1}, & \mathbf{R}_{*,n+1} &:= \mathbf{L}_{**}^{-1}\mathbf{A}_{*,n+1}, \\ \ell_{n+1,n+1} &:= 1, & r_{n+1,n+1} &:= a_{n+1,n+1} - \mathbf{L}_{n+1,*}\mathbf{R}_{*,n+1} \end{aligned}$$

definieren und damit die Gleichungen (3.10) erfüllen. ∎

Bemerkung 3.26 (Durchführbarkeit) *Falls die m-ten Hauptuntermatrizen für alle $m \in [1 : n-1]$ regulär sind, ist der in Abbildung 3.4 dargestellte Algorithmus durchführbar: Aus dem Beweis von Satz 3.25 folgt, dass die m-te Hauptuntermatrix von $\mathbf{L}$ und die m-te Hauptuntermatrix von $\mathbf{R}$ eine LR-Zerlegung der m-ten Hauptuntermatrix von $\mathbf{A}$ beschreiben. Da letztere regulär ist, muss insbesondere auch die m-te Hauptuntermatrix von $\mathbf{R}$ regulär sein, also muss nach Lemma 3.18 $r_{mm} \neq 0$ für alle $m \in [1 : n-1]$ gelten.*

Nach Konstruktion stimmt a_{kk} in Zeile 3 von Abbildung 3.4 gerade mit r_{kk} überein, ist also ebenfalls von null verschieden. Damit ist der Algorithmus durchführbar.

Bemerkung 3.27 *Bei regulären Matrizen $\mathbf{A} \in \mathbb{K}^{n\times n}$ ist die Existenz einer LR-Zerlegung sogar äquivalent zur Regularität der Hauptuntermatrizen. Zum Nachweis fixieren wir eine LR-Zerlegung $(\mathbf{L}, \mathbf{R})$ von $\mathbf{A}$ und stellen fest, dass die m-ten Hauptuntermatrizen von $\mathbf{L}$ und $\mathbf{R}$ gerade eine LR-Zerlegung der m-ten Hauptuntermatrix von $\mathbf{A}$ definieren. Da $\mathbf{A} = \mathbf{LR}$ gilt und $\mathbf{A}$ regulär ist, müssen auch $\mathbf{L}$ und $\mathbf{R}$ regulär sein, und das impliziert nach Lemma 3.18 bereits, dass auch die m-ten Hauptuntermatrizen von $\mathbf{L}$ und $\mathbf{R}$ regulär sein müssen. Also ist es auch deren Produkt, die m-te Hauptuntermatrix von $\mathbf{A}$.*

Bemerkung 3.28 (Alternativer Algorithmus) *Aus dem Beweis des Satzes 3.25 lässt sich ebenfalls ein Algorithmus für die Konstruktion der LR-Zerlegung gewinnen: Wir beginnen mit der ersten Hauptuntermatrix und verwenden die Gleichungen (3.10), um eine LR-Zerlegung der nächstgrößeren Hauptuntermatrix zu berechnen.*

Die Gleichungen (3.10b) können dabei mit Hilfe des Vorwärtseinsetzens gelöst werden.

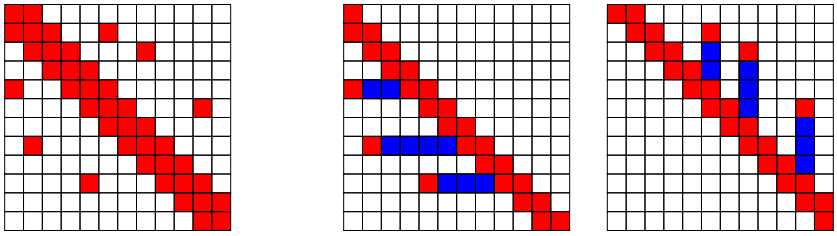

Abbildung 3.5: Skyline-Struktur der LR-Zerlegung

Bemerkung 3.29 *Bei nicht-regulären Matrizen kann eine LR-Zerlegung auch dann existieren, wenn nicht alle Hauptuntermatrizen regulär sind. Ein Beispiel ist durch*

$$\mathbf{A} = \begin{pmatrix} 1 & 1 & 1 \\ 1 & 1 & 1 \\ 1 & 1 & 1 \end{pmatrix}, \qquad \mathbf{L} = \begin{pmatrix} 1 & 0 & 0 \\ 1 & 0 & 0 \\ 1 & 0 & 0 \end{pmatrix}, \qquad \mathbf{R} = \begin{pmatrix} 1 & 1 & 1 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix}$$

gegeben: Es gilt $\mathbf{A} = \mathbf{LR}$*, aber die zweite Hauptuntermatrix* $\mathbf{A}_2$ *ist nicht invertierbar.*

Übungsaufgabe 3.30 (Skyline-Matrizen) *In der Praxis treten häufig Matrizen auf, in denen viele Einträge gleich null sind. Falls wir vorhersagen können, welche der Einträge in der LR-Zerlegung gleich null sind, lässt sich der erforderliche Rechenaufwand teilweise erheblich reduzieren.*

Sei $\mathbf{A} \in \mathbb{K}^{n\times n}$ *eine reguläre Matrix, und sei* $(\mathbf{L}, \mathbf{R})$ *eine LR-Zerlegung dieser Matrix. Beweisen Sie*

$$\begin{aligned} &\ell_{ij} \neq 0 \Rightarrow \text{es existiert ein } k \in [1:j] \text{ mit } a_{ik} \neq 0 \\ &r_{ij} \neq 0 \Rightarrow \text{es existiert ein } k \in [1:i] \text{ mit } a_{kj} \neq 0 \qquad \text{für alle } i,j \in [1:n]. \end{aligned}$$

Bei dem Beweis können Sie sich an dem des Satzes 3.25 orientieren und zeigen, dass wegen der Dreiecksstruktur der Matrizen $\mathbf{L}$ *und* $\mathbf{R}$ *in den Matrizen* $\mathbf{L}_{n+1,*}$ *und* $\mathbf{R}_{*,n+1}$ *nur an bestimmten Stellen von null verschiedene Einträge vorkommen können.*

Übungsaufgabe 3.31 (Dreiecksmatrizen) *Beweisen Sie die folgenden Aussagen.*

(a) Seien $\mathbf{L}_1, \mathbf{L}_2 \in \mathbb{K}^{n\times n}$ *untere Dreiecksmatrizen. Dann ist auch das Produkt* $\mathbf{L}_1\mathbf{L}_2$ *eine untere Dreiecksmatrix.*

(b) Seien $\mathbf{R}_1, \mathbf{R}_2 \in \mathbb{K}^{n\times n}$ *obere Dreiecksmatrizen. Dann ist auch das Produkt* $\mathbf{R}_1\mathbf{R}_2$ *eine obere Dreiecksmatrix.*

(c) Sei $\mathbf{L} \in \mathbb{K}^{n\times n}$ *eine reguläre untere Dreiecksmatrix. Dann ist auch die Inverse* $\mathbf{L}^{-1}$ *eine untere Dreiecksmatrix.*

(d) Sei $\mathbf{R} \in \mathbb{K}^{n\times n}$ *eine reguläre obere Dreiecksmatrix. Dann ist auch die Inverse* $\mathbf{R}^{-1}$ *eine obere Dreiecksmatrix.*

Übungsaufgabe 3.32 (Eindeutigkeit der LR-Zerlegung) *Sei $\mathbf{A} \in \mathbb{K}^{n\times n}$ eine reguläre Matrix. Seien $(\mathbf{L}_1, \mathbf{R}_1)$ und $(\mathbf{L}_2, \mathbf{R}_2)$ LR-Zerlegungen der Matrix $\mathbf{A}$.*

Beweisen Sie, dass eine reguläre Diagonalmatrix $\mathbf{D} \in \mathbb{K}^{n\times n}$ mit $\mathbf{L}_1 = \mathbf{L}_2\mathbf{D}$ und $\mathbf{R}_1 = \mathbf{D}^{-1}\mathbf{R}_2$ existiert.

Hinweis: *Übungsaufgabe 3.31 könnte nützlich sein.*

Übungsaufgabe 3.33 (Bandmatrizen) *Sei $\mathbf{A} \in \mathbb{K}^{n\times n}$ eine Matrix. Sei $k \in \mathbb{N}_0$ eine Zahl mit*

$$|i-j| > k \Rightarrow a_{ij} = 0 \qquad \text{für alle } i,j \in [1:n].$$

Sei $(\mathbf{L}, \mathbf{R})$ eine LR-Zerlegung der Matrix $\mathbf{A}$. Beweisen Sie, dass

$$\begin{aligned} |i-j| > k \Rightarrow \ell_{ij} = 0 \qquad & \text{für alle } i,j \in [1:n],\\ |i-j| > k \Rightarrow r_{ij} = 0 \qquad & \text{für alle } i,j \in [1:n] \end{aligned}$$

gelten. Können Sie einen Algorithmus angeben, mit dem sich eine derartige LR-Zerlegung in $\mathcal{O}(nk^2)$ Operationen berechnen lässt? Und Algorithmen für das Vorwärts- und Rückwärtseinsetzen in $\mathcal{O}(nk)$ Operationen?

Übungsaufgabe 3.34 (Konditionszahl) *Sei $\mathbf{A} \in \mathbb{K}^{n\times n}$ eine reguläre Matrix mit einer LR-Zerlegung $(\mathbf{L}, \mathbf{R})$.*

Seien $\|\cdot\|_U$, $\|\cdot\|_V$ und $\|\cdot\|_W$ Normen auf $\mathbb{K}^n$. Beweisen Sie

$$\|\mathbf{L}^{-1}\|_{W\leftarrow V} \geq \frac{\|\mathbf{R}\|_{V\leftarrow U}}{\|\mathbf{A}\|_{W\leftarrow U}}, \qquad \|\mathbf{R}^{-1}\|_{U\leftarrow V} \geq \frac{\|\mathbf{L}\|_{W\leftarrow V}}{\|\mathbf{A}\|_{W\leftarrow U}}.$$

Aus diesen Abschätzungen folgt

$$\kappa_{W\leftarrow V}(\mathbf{L}) \geq \frac{\|\mathbf{L}\|_{W\leftarrow V}\|\mathbf{R}\|_{V\leftarrow U}}{\|\mathbf{A}\|_{W\leftarrow U}}, \qquad \kappa_{V\leftarrow U}(\mathbf{R}) \geq \frac{\|\mathbf{L}\|_{W\leftarrow V}\|\mathbf{R}\|_{V\leftarrow U}}{\|\mathbf{A}\|_{W\leftarrow U}}.$$

Falls also das Produkt der Normen der Dreiecksmatrizen deutlich größer als die Norm der Matrix $\mathbf{A}$ ist, müssen wir mit schlecht konditionierten Gleichungssystemen rechnen.

Übungsaufgabe 3.35 (Sherman-Morrison-Woodbury-Formel) *Sei $\mathbf{A} \in \mathbb{K}^{n\times n}$ eine reguläre Matrix, seien $\mathbf{u} \in \mathbb{K}^{n\times 1}$ und $\mathbf{v} \in \mathbb{K}^{1\times n}$ gegeben. Beweisen Sie die Gleichungen*

$$\begin{pmatrix} 1 & \\ \mathbf{u} & \mathbf{I} \end{pmatrix} \begin{pmatrix} 1 & \mathbf{v} \\ & \mathbf{A}-\mathbf{uv} \end{pmatrix} = \begin{pmatrix} 1 & \mathbf{v} \\ \mathbf{u} & \mathbf{A} \end{pmatrix} = \begin{pmatrix} 1-\mathbf{vA}^{-1}\mathbf{u} & \mathbf{vA}^{-1} \\ & \mathbf{I} \end{pmatrix} \begin{pmatrix} 1 & \\ \mathbf{u} & \mathbf{A} \end{pmatrix}$$

und folgern Sie aus ihnen, dass $\mathbf{A}-\mathbf{uv}$ genau dann regulär ist, wenn $\mathbf{vA}^{-1}\mathbf{u} \neq 1$ gilt.

Zeigen Sie, dass in diesem Fall $(\mathbf{A}-\mathbf{uv})^{-1} = \mathbf{A}^{-1} + \frac{\mathbf{A}^{-1}\mathbf{uvA}^{-1}}{1-\mathbf{vA}^{-1}\mathbf{u}}$ gilt.

3.4 LR-Zerlegung mit Pivotsuche

Natürlich ist es unbefriedigend, dass sich ein einfaches Gleichungssystem

$$\begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \begin{pmatrix} b_1 \\ b_2 \end{pmatrix}$$

(vgl. (3.9)) nicht mit den bisherigen Verfahren behandeln lässt, da die erste Hauptuntermatrix nicht regulär ist und deshalb auch keine LR-Zerlegung existiert.

Dieses Beispiel legt allerdings auch eine möglich Lösung des Problems nahe: Wenn wir die beiden Zeilen der Matrix und der rechten Seite vertauschen, erhalten wir das System

$$\begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \begin{pmatrix} b_2 \\ b_1 \end{pmatrix},$$

für das wir mit größter Leichtigkeit eine LR-Zerlegung finden können.

Diese Beobachtung lässt uns hoffen, dass wir bei einer regulären Matrix $\mathbf{A}$ durch geschicktes Vertauschen von Zeilen dafür sorgen können, dass eine LR-Zerlegung der derart umsortierten Matrix existiert. Das Vertauschen von Zeilen beschreiben wir mathematisch durch eine Permutation:

Definition 3.36 (Permutation) *Sei $n \in \mathbb{N}$, und sei $\pi : [1:n] \to [1:n]$ eine Abbildung. Falls π bijektiv ist, nennen wir die Abbildung eine* n-stellige Permutation.

Das Vertauschen der Zeilen oder Spalten einer Matrix ist eine lineare Abbildung, die sich selbst durch eine Matrix darstellen lässt:

Lemma 3.37 (Permutationsmatrizen) *Zu jeder n-stelligen Permutation π definieren wir die Matrix $\mathbf{P}_\pi \in \mathbb{K}^{n\times n}$ durch*

$$(p_\pi)_{ij} = \begin{cases} 1 & \text{falls } j = \pi(i), \\ 0 & \text{ansonsten} \end{cases} \qquad \text{für alle } i,j \in [1:n].$$

Derartige Matrizen sind immer regulär, und für zwei n-stellige Permutationen π_1 und π_2 gilt

$$\mathbf{P}_{\pi_1}\mathbf{P}_{\pi_2} = \mathbf{P}_{\pi_2\circ\pi_1}. \tag{3.11}$$

Beweis. Wir beweisen zunächst (3.11). Nach Definition der Matrix-Multiplikation gilt für alle $i,j \in [1:n]$ die Gleichung

$$(\mathbf{P}_{\pi_1}\mathbf{P}_{\pi_2})_{ij} = \sum_{k=1}^{n} (p_{\pi_1})_{ik}(p_{\pi_2})_{kj} = (p_{\pi_1})_{i,\pi_1(i)}(p_{\pi_2})_{\pi_1(i),j} = (p_{\pi_2})_{\pi_1(i),j},$$

und aus der Definition von $\mathbf{P}_{\pi_2}$ folgt, dass dieser Term genau dann gleich eins ist, wenn $j = \pi_2(\pi_1(i)) = \pi_2 \circ \pi_1(i)$ gilt, anderenfalls ist er gleich null.

Für den Nachweis der Regularität wenden wir (3.11) auf $\pi_2 = \pi$ und $\pi_1 = \pi^{-1}$ an und erhalten $\mathbf{P}_{\pi^{-1}}\mathbf{P}_\pi = \mathbf{P}_{\pi\circ\pi^{-1}} = \mathbf{I}$, also ist $\mathbf{P}_{\pi^{-1}}$ eine Linksinverse von $\mathbf{P}_\pi$. Entsprechend

können wir nachweisen, dass es auch eine Rechtsinverse ist, also ist $\mathbf{P}_\pi$ insbesondere regulär. ∎

Permutationsmatrizen sind gerade so konstruiert, dass sie die Komponenten eines Vektors entsprechend der zugehörigen Permutation umsortieren:

$$\mathbf{P}_\pi \begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix} = \begin{pmatrix} x_{\pi(1)} \\ \vdots \\ x_{\pi(n)} \end{pmatrix}.$$

In unseren Anwendungen wird sich die Permutation π in der Regel aus Vertauschungen einzelner Komponenten zusammensetzen. Deshalb fixieren wir als Abkürzung für alle $i, j \in [1:n]$ die n-stellige Permutation π_{ij}^n, die durch

$$\pi_{ij}^n(i) = j, \qquad \pi_{ij}^n(j) = i, \qquad \pi_{ij}^n(k) = k \qquad \text{für alle } k \in [1:n] \setminus \{i, j\}$$

definiert ist. Unser Ziel ist es nun, für eine reguläre Matrix $\mathbf{A} \in \mathbb{K}^{n \times n}$ eine Zerlegung der Form

$$\mathbf{P}_\pi \mathbf{A} = \mathbf{L}\mathbf{R}$$

zu konstruieren, wobei π eine n-stellige Permutation, $\mathbf{L}$ eine untere und $\mathbf{R}$ eine obere Dreiecksmatrix ist.

Eine derartige Zerlegung bezeichnen wir als *LR-Zerlegung mit Pivotsuche* oder als *pivotisierte LR-Zerlegung*. Der Begriff *pivot* stammt aus dem Englischen und bedeutet soviel wie „Dreh- und Angelpunkt“.

Satz 3.38 (LR-Zerlegung mit Pivotsuche) *Sei $\mathbf{A} \in \mathbb{K}^{n \times n}$ eine reguläre Matrix. Dann können wir eine n-stellige Permutation π so finden, dass $\mathbf{P}_\pi \mathbf{A}$ eine LR-Zerlegung $(\mathbf{L}, \mathbf{R})$ besitzt, dass also*

$$\mathbf{P}_\pi \mathbf{A} = \mathbf{L}\mathbf{R}$$

mit einer unteren Dreiecksmatrix $\mathbf{L}$ und einer oberen Dreiecksmatrix $\mathbf{R}$ gilt.

Beweis. Wir beweisen die Aussage konstruktiv durch Induktion über $n \in \mathbb{N}$.

Induktionsanfang: Für $n = 1$ ist die Aussage trivial.

Induktionsvoraussetzung: Sei $n \in \mathbb{N}$ so fixiert, dass die Aussage für alle regulären Matrizen $\mathbf{A} \in \mathbb{K}^{n \times n}$ gilt.

Induktionsschritt: Sei $\mathbf{A} \in \mathbb{K}^{(n+1)\times(n+1)}$ regulär. Wir wählen ein $i \in [1 : n+1]$ so, dass $a_{i1} \neq 0$ gilt. Ein derartiges i existiert, weil ansonsten die erste Spalte gleich null wäre und damit die Matrix $\mathbf{A}$ nicht regulär. Nun vertauschen wir die i-te Zeile mit der ersten, arbeiten also mit

$$\mathbf{B} := \mathbf{P}_{\pi_{1i}^{n+1}} \mathbf{A}$$

weiter. Da $b_{11} = a_{i1} \neq 0$ gilt, können wir wie bei der Herleitung des Algorithmus für die ursprüngliche LR-Zerlegung verfahren: Wir setzen

$$\mathbf{B}_{**} := \begin{pmatrix} b_{22} & \dots & b_{2,n+1} \\ \vdots & \ddots & \vdots \\ b_{n+1,2} & \dots & b_{n+1,n+1} \end{pmatrix}, \quad \mathbf{B}_{1*} := \begin{pmatrix} b_{12} & \dots & b_{1,n+1} \end{pmatrix}, \quad \mathbf{B}_{*1} := \begin{pmatrix} b_{21} \\ \vdots \\ b_{n+1,1} \end{pmatrix}.$$

In einem ersten Schritt eliminieren wir die erste Spalte und erhalten

$$\mathbf{B} = \begin{pmatrix} b_{11} & \mathbf{B}_{1*} \\ \mathbf{B}_{*1} & \mathbf{B}_{**} \end{pmatrix} = \begin{pmatrix} 1 & \\ \mathbf{B}_{*1}\frac{1}{b_{11}} & \mathbf{I} \end{pmatrix} \begin{pmatrix} b_{11} & \mathbf{B}_{1*} \\ & \mathbf{B}_{**} - \mathbf{B}_{*1}\frac{1}{b_{11}}\mathbf{B}_{1*} \end{pmatrix}.$$

Wenn wir eine pivotisierte LR-Zerlegung des rechten unteren Matrixblocks

$$\widehat{\mathbf{B}} := \mathbf{B}_{**} - \mathbf{B}_{*1}\frac{1}{b_{11}}\mathbf{B}_{1*} \in \mathbb{K}^{n\times n}$$

finden könnten, wären wir unserem Ziel schon sehr nahe. Um die Induktionsvoraussetzung auf diesen Block anwenden zu können, müssen wir nachweisen, dass diese Matrix regulär ist.

Dazu wählen wir einen Vektor $\mathbf{y} \in \mathbb{K}^n$ aus dem Kern der Matrix $\widehat{\mathbf{B}}$, es gelte also $\widehat{\mathbf{B}}\mathbf{y} = \mathbf{0}$. Wir setzen

$$\mathbf{x} := \begin{pmatrix} -\mathbf{B}_{1*}\mathbf{y}/b_{11} \\ \mathbf{y} \end{pmatrix}$$

und erhalten

$$\mathbf{B}\mathbf{x} = \begin{pmatrix} 1 & \\ \mathbf{B}_{*1}\frac{1}{b_{11}} & \mathbf{I} \end{pmatrix} \begin{pmatrix} b_{11} & \mathbf{B}_{1*} \\ & \widehat{\mathbf{B}} \end{pmatrix} \begin{pmatrix} -\mathbf{B}_{1*}\mathbf{y}/b_{11} \\ \mathbf{y} \end{pmatrix} = \begin{pmatrix} 1 & \\ \mathbf{B}_{*1}\frac{1}{b_{11}} & \mathbf{I} \end{pmatrix} \begin{pmatrix} 0 \\ \mathbf{0} \end{pmatrix} = \mathbf{0}.$$

Da $\mathbf{B}$ regulär ist, muss $\mathbf{x}$ der Nullvektor sein, also folgt auch $\mathbf{y} = \mathbf{0}$. Damit ist $\widehat{\mathbf{B}}$ injektiv, also regulär.

Nach Induktionsvoraussetzung muss also eine Zerlegung

$$\mathbf{P}_{\hat{\pi}}\widehat{\mathbf{B}} = \widehat{\mathbf{L}}\widehat{\mathbf{R}}$$

mit einer n-stelligen Permutation $\hat{\pi}$ existieren. Wir setzen

$$\mathbf{L} := \begin{pmatrix} 1 & \\ \mathbf{P}_{\hat{\pi}}\mathbf{B}_{*1}\frac{1}{b_{11}} & \widehat{\mathbf{L}} \end{pmatrix}, \qquad \mathbf{R} := \begin{pmatrix} b_{11} & \mathbf{B}_{1*} \\ & \widehat{\mathbf{R}} \end{pmatrix}.$$

Daraus folgt

$$\begin{aligned} \mathbf{L}\mathbf{R} &= \begin{pmatrix} 1 & \\ \mathbf{P}_{\hat{\pi}}\mathbf{B}_{*1}\frac{1}{b_{11}} & \widehat{\mathbf{L}} \end{pmatrix} \begin{pmatrix} b_{11} & \mathbf{B}_{1*} \\ & \widehat{\mathbf{R}} \end{pmatrix} = \begin{pmatrix} b_{11} & \mathbf{B}_{1*} \\ \mathbf{P}_{\hat{\pi}}\mathbf{B}_{*1} & \mathbf{P}_{\hat{\pi}}\mathbf{B}_{*1}\frac{1}{b_{11}}\mathbf{B}_{1*} + \widehat{\mathbf{L}}\widehat{\mathbf{R}} \end{pmatrix} \\ &= \begin{pmatrix} b_{11} & \mathbf{B}_{1*} \\ \mathbf{P}_{\hat{\pi}}\mathbf{B}_{*1} & \mathbf{P}_{\hat{\pi}}(\mathbf{B}_{*1}\frac{1}{b_{11}}\mathbf{B}_{1*} + \widehat{\mathbf{B}}) \end{pmatrix} = \begin{pmatrix} 1 & \\ & \mathbf{P}_{\hat{\pi}} \end{pmatrix} \mathbf{B} = \begin{pmatrix} 1 & \\ & \mathbf{P}_{\hat{\pi}} \end{pmatrix} \mathbf{P}_{\pi_{1i}^{n+1}}\mathbf{A}. \end{aligned}$$

Das ist schon fast die gesuchte Zerlegung, wir müssen nur noch die beiden Permutationen zu einer Permutation zusammenfassen.

Dazu konstruieren wir eine $(n+1)$-stellige Permutation $\tilde{\pi}$, die die erste Komponente unverändert lässt und die restlichen wie $\hat{\pi}$ vertauscht. Sie ist gegeben durch

$$\tilde{\pi}(k) = \begin{cases} 1 & \text{falls } k = 1, \\ \hat{\pi}(k-1)+1 & \text{ansonsten} \end{cases} \qquad \text{für alle } k \in [1:n+1]$$

und erfüllt dank Lemma 3.37 die Gleichungen

$$\begin{pmatrix} 1 & \\ & \mathbf{P}_{\hat{\pi}} \end{pmatrix} = \mathbf{P}_{\tilde{\pi}}, \qquad \begin{pmatrix} 1 & \\ & \mathbf{P}_{\hat{\pi}} \end{pmatrix} \mathbf{P}_{\pi_{1i}^{n+1}} = \mathbf{P}_{\tilde{\pi}} \mathbf{P}_{\pi_{1i}^{n+1}} = \mathbf{P}_{\pi_{1i}^{n+1} \circ \tilde{\pi}}.$$

Damit erhalten wir

$$\mathbf{LR} = \mathbf{P}_{\pi} \mathbf{A},$$

mit $\pi := \pi_{1i}^{n+1} \circ \tilde{\pi}$, also die gesuchte Zerlegung. ∎

---

procedure pivot_lr(n, **var** $\mathbf{A}$, p)
for $k = 1, \ldots, n-1$ **do**
  Suche $i \in [k : n]$ so, dass $|a_{ik}|$ maximal ist
  if $a_{ik} = 0$ **then**
    Abbruch, $\mathbf{A}$ ist nicht regulär
  end if
  $p_k \leftarrow i$
  for $j \in [1 : n]$ **do**
    $h \leftarrow a_{kj}$; $a_{kj} \leftarrow a_{ij}$; $a_{ij} \leftarrow h$
  end for
  for $i \in [k+1 : n]$ **do**
    $a_{ik} \leftarrow a_{ik}/a_{kk}$
  end for
  for $i, j \in [k+1 : n]$ **do**
    $a_{ij} \leftarrow a_{ij} - a_{ik}a_{kj}$
  end for
end for

Abbildung 3.6: Algorithmus zur Berechnung der LR-Zerlegung mit Pivotsuche

Da der Beweis von Satz 3.38 konstruktiv ist, können wir die pivotisierte LR-Zerlegung berechnen, indem wir wieder die Induktion in einzelne Schritte auflösen. In dieser Weise erhalten wir den in Abbildung 3.6 dargestellten Algorithmus, der die Permutation π berechnet und die Matrix $\mathbf{A}$ mit der Matrix

$$\begin{pmatrix} r_{11} & \ldots & \ldots & r_{1n} \\ l_{21} & r_{22} & \ldots & r_{2n} \\ \vdots & \ddots & \ddots & \vdots \\ l_{n1} & \ldots & l_{n-1,n} & r_{rr} \end{pmatrix}$$

überschreibt, aus der wir wie gehabt die Faktoren $\mathbf{L}$ und $\mathbf{R}$ ablesen können. Die Permutation π stellen wir dabei durch einen Vektor p dar, dessen k-te Komponente angibt, mit welcher Zeile die k-te Zeile im k-ten Schritt des Verfahrens vertauscht wurde. Diese Darstellung ist einfacher handzuhaben als beispielsweise die Darstellung durch eine Wertetabelle.

Da Permutationsmatrizen nach Lemma 3.37 immer regulär sind, ist das Gleichungssystem $\mathbf{Ax} = \mathbf{b}$ äquivalent zu

$$\mathbf{LRx} = \mathbf{P}_\pi \mathbf{Ax} = \mathbf{P}_\pi \mathbf{b}$$

und das Auflösen des Gleichungssystems zerfällt nun in drei Schritte:

$$\widetilde{\mathbf{b}} := \mathbf{P}_\pi \mathbf{b}, \qquad \mathbf{Ly} = \widetilde{\mathbf{b}}, \qquad \mathbf{Rx} = \mathbf{y},$$

die kompakt in Abbildung 3.7 zusammengefasst sind.

```
procedure solve_lr(n, A, p, var b);
for i = 1, ..., n − 1 do
  j ← p_i;
  h ← b_i;   b_i ← b_j;   b_j ← h
end for
for j = 1, ..., n − 1 do
  for i ∈ [j + 1 : n] do
    b_i ← b_i − a_ij b_j
  end for
end for
for j = n, ..., 1 do
  b_j ← b_j / a_jj
  for i ∈ [1 : j − 1] do
    b_i ← b_i − a_ij b_j
  end for
end for
```

Abbildung 3.7: Lösen des Gleichungssystems per pivotisierter LR-Zerlegung

In diesem Algorithmus nutzen wir zusätzlich aus, dass auf der Diagonalen der Matrix $\mathbf{L}$ immer eins steht, so dass wir uns die Division durch das Diagonalelement, die eigentlich für das Vorwärtseinsetzen erforderlich wäre, sparen können.

Da das Vertauschen von Einträgen keine arithmetische Operation, sondern nur ein Kopiervorgang ist, erhöht sich der Rechenaufwand durch die Pivotsuche in unserer Zählweise nicht, auch in der Praxis führt sie zu keiner deutlichen Verlangsamung, verbessert allerdings die Stabilität des Algorithmus wesentlich.

Die Schleife für das Vorwärtseinsetzen erfordert in diesem Fall lediglich $n(n-1)$ Operationen, da keine Division durch die Diagonalelement erforderlich ist, so dass insgesamt $2n^2 - n$ Operationen für das Lösen des Gleichungssystems benötigt werden.

Bemerkung 3.39 (Determinante) *Um mit Hilfe der pivotisierten LR-Zerlegung auch die Determinante einer Matrix* $\mathbf{A}$ *bestimmen zu können, müssen wir die Determinante der Permutationsmatrix* $\mathbf{P}_\pi$ *berechnen. Dazu genügt es, die Determinanten der Vertauschungsmatrizen* $\mathbf{P}_{\pi_{ki}^{n-k+1}}$ *zu multiplizieren, und letztere sind gleich eins, falls* $k = i$ *gilt, und ansonsten gleich minus eins.*

3.5 Skalarprodukt und positiv definite Matrizen

In der Praxis treten häufig Matrizen auf, die besondere Eigenschaften aufweisen, die sich ausnutzen lassen, um Speicherplatz zu sparen oder Lösungsalgorithmen zu beschleunigen. Zwei besonders wichtige Eigenschaften einer Matrix, nämlich *selbstadjungiert* zu sein und *positiv definit* zu sein, lassen sich mit Hilfe des *Euklidischen Skalarprodukts* beschreiben, das durch

$$\langle \mathbf{x}, \mathbf{y} \rangle_2 := \sum_{i=1}^{n} \bar{x}_i y_i \qquad \text{für alle } \mathbf{x}, \mathbf{y} \in \mathbb{K}^n \tag{3.12}$$

definiert ist. Das Skalarprodukt ist wegen

$$\langle \mathbf{x}, \mathbf{x} \rangle_2 = \sum_{i=1}^{n} |x_i|^2 = \|\mathbf{x}\|_2^2 \qquad \text{für alle } \mathbf{x} \in \mathbb{K}^n \tag{3.13}$$

eng mit der bereits bekannten Euklidischen Norm verbunden. Im Fall $\mathbb{K} = \mathbb{R}$ ist es eine *Bilinearform*, im Fall $\mathbb{K} = \mathbb{C}$ eine *Sesquilinearform*, es erfüllt nämlich die Gleichungen

$$\begin{aligned} \langle \mathbf{x}, \mathbf{y} \rangle_2 &= \overline{\langle \mathbf{y}, \mathbf{x} \rangle_2}, \\ \langle \mathbf{x} + \lambda \mathbf{z}, \mathbf{y} \rangle_2 &= \langle \mathbf{x}, \mathbf{y} \rangle_2 + \bar{\lambda} \langle \mathbf{z}, \mathbf{y} \rangle_2, \\ \langle \mathbf{x}, \mathbf{y} + \lambda \mathbf{z} \rangle_2 &= \langle \mathbf{x}, \mathbf{y} \rangle_2 + \lambda \langle \mathbf{x}, \mathbf{z} \rangle_2 \qquad \text{für alle } \lambda \in \mathbb{K},\ \mathbf{x}, \mathbf{y}, \mathbf{z} \in \mathbb{K}^n. \end{aligned}$$

Aus der zweiten und dritten Gleichung lässt sich auch eine für Skalarprodukte gültige Variante der binomischen Gleichungen konstruieren, mit deren Hilfe wir die folgende sehr wichtige Abschätzung für Skalarprodukte beweisen können.

Lemma 3.40 (Cauchy-Schwarz) *Es gilt*

$$|\langle \mathbf{x}, \mathbf{y} \rangle_2| \leq \|\mathbf{x}\|_2 \|\mathbf{y}\|_2 \qquad \textit{für alle } \mathbf{x}, \mathbf{y} \in \mathbb{K}^n. \tag{3.14}$$

Beide Seiten in (3.14) sind genau dann gleich, wenn die Vektoren linear abhängig sind.

Beweis. Seien $\mathbf{x}, \mathbf{y} \in \mathbb{K}^n$. Für $\mathbf{x} = \mathbf{0}$ ist die Aussage trivial. Sei also $\mathbf{x} \neq \mathbf{0}$. Der Ausgangspunkt unserer Argumentation ist die Ungleichung

$$\begin{aligned} 0 \leq \|\mathbf{y} - \lambda \mathbf{x}\|_2^2 &= \langle \mathbf{y} - \lambda \mathbf{x}, \mathbf{y} - \lambda \mathbf{x} \rangle_2 \\ &= \langle \mathbf{y}, \mathbf{y} \rangle_2 - \langle \mathbf{y}, \lambda \mathbf{x} \rangle_2 - \langle \lambda \mathbf{x}, \mathbf{y} \rangle_2 + \langle \lambda \mathbf{x}, \lambda \mathbf{x} \rangle_2 \\ &= \|\mathbf{y}\|_2^2 - \lambda \langle \mathbf{y}, \mathbf{x} \rangle_2 - \bar{\lambda} \langle \mathbf{x}, \mathbf{y} \rangle_2 + |\lambda|^2 \|\mathbf{x}\|_2^2, \end{aligned}$$

die aufgrund der Nicht-Negativität der Norm für alle $\lambda \in \mathbb{K}$ gilt. Um diese Eigenschaft möglichst gut auszunutzen, wählen wir λ so, dass der Ausdruck möglichst klein wird. Für $\mathbb{K} = \mathbb{R}$ können wir die rechte Seite als Polynom in λ interpretieren und nach der Nullstelle seiner Ableitung suchen, um das Minimum zu bestimmen. So erhalten wir

$\lambda = \langle \mathbf{x}, \mathbf{y} \rangle_2 / \|\mathbf{x}\|_2^2$, und diese Wahl des Parameters führt auch im allgemeinen Fall zu der Gleichung

$$\begin{aligned}\|\mathbf{y} - \lambda\mathbf{x}\|_2^2 &= \|\mathbf{y}\|_2^2 - \frac{\langle \mathbf{x}, \mathbf{y} \rangle_2 \langle \mathbf{y}, \mathbf{x} \rangle_2}{\|\mathbf{x}\|_2^2} - \frac{\overline{\langle \mathbf{x}, \mathbf{y} \rangle_2} \langle \mathbf{x}, \mathbf{y} \rangle_2}{\|\mathbf{x}\|_2^2} + \frac{|\langle \mathbf{x}, \mathbf{y} \rangle_2|^2}{\|\mathbf{x}\|_2^4} \|\mathbf{x}\|_2^2 \\ &= \|\mathbf{y}\|_2^2 - \frac{|\langle \mathbf{x}, \mathbf{y} \rangle_2|^2}{\|\mathbf{x}\|_2^2} - \frac{|\langle \mathbf{x}, \mathbf{y} \rangle_2|^2}{\|\mathbf{x}\|_2^2} + \frac{|\langle \mathbf{x}, \mathbf{y} \rangle_2|^2}{\|\mathbf{x}\|_2^2} = \|\mathbf{y}\|_2^2 - \frac{|\langle \mathbf{x}, \mathbf{y} \rangle_2|^2}{\|\mathbf{x}\|_2^2}. \end{aligned} \tag{3.15}$$

Damit erhalten wir insbesondere

$$0 \leq \|\mathbf{x}\|_2^2 \|\mathbf{y} - \lambda\mathbf{x}\|_2^2 = \|\mathbf{x}\|_2^2 \|\mathbf{y}\|_2^2 - |\langle \mathbf{x}, \mathbf{y} \rangle_2|^2,$$

also (3.14). Falls $|\langle \mathbf{x}, \mathbf{y} \rangle_2| = \|\mathbf{x}\|_2 \|\mathbf{y}\|_2$ gilt, folgt aus (3.15) auch

$$\|\mathbf{y} - \lambda\mathbf{x}\|_2^2 = \|\mathbf{y}\|_2^2 - \frac{\|\mathbf{x}\|_2^2 \|\mathbf{y}\|_2^2}{\|\mathbf{x}\|_2^2} = \|\mathbf{y}\|_2^2 - \|\mathbf{y}\|_2^2 = 0,$$

also $\mathbf{y} = \lambda\mathbf{x}$, und damit der zweite Teil der Behauptung. ∎

Aufgrund dieser Gleichung können wir den *Winkel* $\angle(\mathbf{x}, \mathbf{y})$ zwischen zwei reellen Vektoren durch

$$\cos \angle(\mathbf{x}, \mathbf{y}) := \frac{\langle \mathbf{x}, \mathbf{y} \rangle_2}{\|\mathbf{x}\|_2 \|\mathbf{y}\|_2} \qquad \text{für alle } \mathbf{x}, \mathbf{y} \in \mathbb{R}^n \setminus \{\mathbf{0}\} \tag{3.16}$$

definieren, denn die rechte Seite dieser Gleichung liegt immer im Intervall $[-1, 1]$, also im Bild des Cosinus. Insbesondere stehen zwei Vektoren *senkrecht* aufeinander, wenn ihr Skalarprodukt verschwindet.

Eine weitere nützliche Eigenschaft des Skalarprodukts besteht darin, dass sich mit seiner Hilfe die Identität zweier Vektoren beweisen lässt.

Lemma 3.41 (Testvektoren) *Seien $\mathbf{x}, \mathbf{y} \in \mathbb{K}^n$ derart, dass*

$$\langle \mathbf{z}, \mathbf{x} \rangle_2 = \langle \mathbf{z}, \mathbf{y} \rangle_2 \qquad \textit{für alle } \mathbf{z} \in \mathbb{K}^n$$

gilt. Dann folgt $\mathbf{x} = \mathbf{y}$.

Beweis. Wir setzen $\mathbf{z} := \mathbf{x} - \mathbf{y}$ und stellen fest, dass

$$0 = \langle \mathbf{z}, \mathbf{x} \rangle_2 - \langle \mathbf{z}, \mathbf{y} \rangle_2 = \langle \mathbf{z}, \mathbf{x} - \mathbf{y} \rangle_2 = \langle \mathbf{z}, \mathbf{z} \rangle_2 = \|\mathbf{z}\|_2^2 = \|\mathbf{x} - \mathbf{y}\|_2^2$$

gilt, also folgt $\mathbf{x} = \mathbf{y}$. ∎

Nun können wir uns der ersten Eigenschaft zuwenden, die den Umgang mit einer Matrix erheblich vereinfachen kann: Die Eigenschaft, *selbstadjungiert* zu sein.

Definition 3.42 (Adjungierte Matrix) *Sei $\mathbf{A} \in \mathbb{K}^{m\times n}$. Die durch*

$$b_{ij} := \bar{a}_{ji} \qquad \text{für alle } i \in [1:n],\ j \in [1:m]$$

definierte Matrix $\mathbf{B} \in \mathbb{K}^{n\times m}$ bezeichnen wir als die Adjungierte *oder* adjungierte Matrix *von $\mathbf{A}$ und notieren sie als $\mathbf{A}^*$.*

Im Fall $\mathbb{K} = \mathbb{R}$ gilt $\mathbf{A}^* = \mathbf{A}^T$, die adjungierte ist also die transponierte Matrix. Im Fall $\mathbb{K} = \mathbb{C}$ gilt $\mathbf{A}^* = \mathbf{A}^H$, die adjungierte ist dann die konjugierte transponierte Matrix.

Definition 3.43 (Selbstadjungierte Matrix) *Sei $\mathbf{A} \in \mathbb{K}^{n\times n}$. Falls $\mathbf{A} = \mathbf{A}^*$ gilt, heißt $\mathbf{A}$* selbstadjungiert.

Im Falle $\mathbb{K} = \mathbb{R}$ ist die Matrix $\mathbf{A}$ genau dann selbstadjungiert, wenn sie symmetrisch ist. Im Falle $\mathbb{K} = \mathbb{C}$ ist sie selbstadjungiert, wenn sie Hermitesch ist.

Neben vielen für die mathematische Analyse wichtigen Eigenschaften besitzen selbstadjungierte Matrizen auch den praktischen Vorteil, dass wir wegen $a_{ji} = \bar{a}_{ij}$ nur ihre obere oder untere Dreieckshälfte abzuspeichern brauchen: Falls wir a_{ij} für alle $i \geq j$ kennen, können wir a_{ji} einfach rekonstruieren. Wenn wir nur eine Dreieckshälfte abspeichern, reduziert sich der Speicherbedarf von n^2 auf $\frac{(n+1)n}{2}$, er wird also fast halbiert.

Leider reicht es nicht aus, die Matrix effizient abzuspeichern, wir müssen auch Gleichungssysteme lösen können. Die einfache oder pivotisierte LR-Zerlegung wäre dafür unattraktiv, da sie wieder n^2 Koeffizienten erfordern würde.

Um zu einem effizienteren Ansatz zu gelangen untersuchen wir die Eigenschaften adjungierter Matrizen etwas näher. Besonders nützlich ist ihre Beziehung zu dem Skalarprodukt, mit deren Hilfe sich viele Beweise sehr elegant gestalten lassen.

Lemma 3.44 (Adjungierte und Skalarprodukt) *Sei $\mathbf{A} \in \mathbb{K}^{m\times n}$. Es gilt*

$$\langle \mathbf{Ax}, \mathbf{y}\rangle_2 = \langle \mathbf{x}, \mathbf{A}^*\mathbf{y}\rangle_2 \qquad \text{für alle } \mathbf{A} \in \mathbb{K}^{n\times m},\ \mathbf{x} \in \mathbb{K}^n,\ \mathbf{y} \in \mathbb{K}^m, \tag{3.17}$$

wobei auf der linken Seite das Skalarprodukt auf $\mathbb{K}^m$ und auf der rechten das auf $\mathbb{K}^n$ verwendet wird.

Beweis. Seien $\mathbf{x} \in \mathbb{K}^n$ und $\mathbf{y} \in \mathbb{K}^m$ gegeben. Wir setzen $\mathbf{B} := \mathbf{A}^*$. Dann gilt

$$\langle \mathbf{Ax}, \mathbf{y}\rangle_2 = \sum_{i=1}^{m} \overline{(\mathbf{Ax})_i} y_i = \sum_{i=1}^{m}\sum_{j=1}^{n} \bar{a}_{ij}\bar{x}_j y_i = \sum_{j=1}^{n}\sum_{i=1}^{m} \bar{x}_j b_{ji} y_i = \sum_{j=1}^{n} \bar{x}_j (\mathbf{By})_j = \langle \mathbf{x}, \mathbf{By}\rangle_2.$$

Dank $\mathbf{B} = \mathbf{A}^*$ ist das bereits die gewünschte Gleichung. ∎

Mit der Gleichung (3.17) in Kombination mit Lemma 3.41 lassen sich häufig sehr kompakt nützliche Aussagen über Produkte von Matrizen beweisen. Ein Beispiel ist das folgende Lemma.

Lemma 3.45 (Adjungierte und Produkt sowie Inverse) *Seien $\mathbf{A} \in \mathbb{K}^{m\times n}$ und $\mathbf{B} \in \mathbb{K}^{n\times k}$ gegeben. Dann gilt $(\mathbf{AB})^* = \mathbf{B}^*\mathbf{A}^*$.*

Sei $\mathbf{A} \in \mathbb{K}^{n\times n}$ regulär. Dann ist auch $\mathbf{A}^*$ regulär und es gilt $(\mathbf{A}^*)^{-1} = (\mathbf{A}^{-1})^*$.

Beweis. Sei $\mathbf{x} \in \mathbb{K}^m$. Für alle $\mathbf{z} \in \mathbb{K}^k$ gilt nach Lemma 3.44

$$\langle \mathbf{z}, (\mathbf{AB})^*\mathbf{x}\rangle_2 = \langle \mathbf{ABz}, \mathbf{x}\rangle_2 = \langle \mathbf{Bz}, \mathbf{A}^*\mathbf{x}\rangle_2 = \langle \mathbf{z}, \mathbf{B}^*\mathbf{A}^*\mathbf{x}\rangle_2,$$

also folgt mit Lemma 3.41 bereits $(\mathbf{AB})^*\mathbf{x} = \mathbf{B}^*\mathbf{A}^*\mathbf{x}$. Da wir diese Gleichung für beliebige $\mathbf{x}$ bewiesen haben, folgt die erste Behauptung.

Für eine reguläre Matrix $\mathbf{A} \in \mathbb{K}^{n\times n}$ erhalten wir mit dieser Aussage

$$\mathbf{I} = \mathbf{I}^* = (\mathbf{A}^{-1}\mathbf{A})^* = \mathbf{A}^*(\mathbf{A}^{-1})^*, \qquad \mathbf{I} = \mathbf{I}^* = (\mathbf{AA}^{-1})^* = (\mathbf{A}^{-1})^*\mathbf{A}^*,$$

also ist $(\mathbf{A}^{-1})^*$ sowohl eine Rechts- als auch eine Linksinverse der Matrix $\mathbf{A}^*$. ∎

Nun können wir uns wieder der Frage nach einem effizienten Lösungsverfahren für Gleichungssysteme mit einer selbstadjungierten Matrix zuwenden.

Falls $\mathbf{A}$ selbstadjungiert und $(\mathbf{L}, \mathbf{R})$ eine LR-Zerlegung der Matrix ist, gilt nach Lemma 3.45

$$\mathbf{LR} = \mathbf{A} = \mathbf{A}^* = (\mathbf{LR})^* = \mathbf{R}^*\mathbf{L}^*,$$

und $\mathbf{R}^*$ ist eine untere sowie $\mathbf{L}^*$ eine obere Dreiecksmatrix, also ist auch $(\mathbf{R}^*, \mathbf{L}^*)$ eine LR-Zerlegung von $\mathbf{A}$. Ein genauerer Blick auf die Konstruktion von LR-Zerlegungen legt nahe, dass sich zwei LR-Zerlegungen derselben Matrix nur durch eine unterschiedliche Wahl der Diagonalelemente unterscheiden können (vgl. Übungsaufgabe 3.32), so dass wir $\mathbf{R} = \mathbf{DL}^*$ mit einer Diagonalmatrix $\mathbf{D}$ folgern können. Damit haben wir die Zerlegung $\mathbf{A} = \mathbf{LDL}^*$ erhalten, die sich in $\frac{(n+1)n}{2}$ Speicherplätzen (n für $\mathbf{D}$ und $\frac{(n-1)n}{2}$ für die Matrix $\mathbf{L}$ mit Diagonalelementen gleich eins) darstellen lässt und mit der sich das Gleichungssystem wieder durch Vorwärts- und Rückwärtseinsetzen lösen lässt.

Eine derartige Zerlegung existiert nur dann, wenn eine LR-Zerlegung *ohne Pivotsuche* existiert, und eine Pivotsuche können wir nicht einsetzen, weil sie im Allgemeinen die Selbstadjungiertheit der Matrix zunichte machen würde.

Glücklicherweise gibt es eine wichtige Klasse von Matrizen, für die immer eine LR-Zerlegung existiert:

Definition 3.46 (Positiv definite Matrix) *Sei $\mathbf{A} \in \mathbb{K}^{n\times n}$ eine Matrix. Falls für alle Vektoren $\mathbf{y} \in \mathbb{K}^n \setminus \{\mathbf{0}\}$ die Beziehung*

$$\langle \mathbf{y}, \mathbf{Ay}\rangle_2 = \sum_{i=1}^{n}\sum_{j=1}^{n} \bar{y}_i a_{ij} y_j \in \mathbb{R}_{>0}$$

gilt, heißt die Matrix $\mathbf{A}$ positiv definit.

Für unsere Zwecke besitzen positiv definite Matrizen nützliche Eigenschaften:

Lemma 3.47 (Positiv definite Matrix) *Sei $\mathbf{A} \in \mathbb{K}^{n\times n}$ eine positiv definite Matrix. Dann ist $\mathbf{A}$ regulär, und alle Hauptuntermatrizen sind ebenfalls positiv definit, also insbesondere auch regulär.*

Die Inverse $\mathbf{A}^{-1}$ ist ebenfalls positiv definit.

Beweis. Falls $\mathbf{y} \in \mathbb{K}^n$ ein Vektor aus dem Kern von $\mathbf{A}$ ist, gilt $\mathbf{Ay} = \mathbf{0}$ und damit insbesondere

$$\langle \mathbf{y}, \mathbf{Ay} \rangle_2 = \langle \mathbf{y}, \mathbf{0} \rangle_2 = 0,$$

also muss nach Definition $\mathbf{y} = \mathbf{0}$ gelten, der Kern enthält also nur den Nullvektor. Mit dem Dimensionssatz folgt, dass die Matrix $\mathbf{A}$ regulär sein muss.

Sei $m \in [1 : n]$, und sei $\mathbf{B} \in \mathbb{K}^{m \times m}$ die m-te Hauptuntermatrix von $\mathbf{A}$. Sei $\mathbf{y} \in \mathbb{K}^m \setminus \{\mathbf{0}\}$. Wir definieren den Hilfsvektor $\mathbf{z} \in \mathbb{K}^n \setminus \{\mathbf{0}\}$ durch

$$z_i = \begin{cases} y_i & \text{falls } i \leq m, \\ 0 & \text{ansonsten} \end{cases} \qquad \text{für alle } 1 \leq i \leq n$$

und erhalten

$$\langle \mathbf{y}, \mathbf{By} \rangle_2 = \sum_{i=1}^{m} \sum_{j=1}^{m} \bar{y}_i a_{ij} y_j = \sum_{i=1}^{n} \sum_{j=1}^{n} \bar{z}_i a_{ij} z_j = \langle \mathbf{z}, \mathbf{Az} \rangle_2 > 0,$$

also muss auch $\mathbf{B}$ positiv definit sein.

Sei $\mathbf{y} \in \mathbb{K}^n \setminus \{\mathbf{0}\}$. Wir setzen $\mathbf{z} := \mathbf{A}^{-1}\mathbf{y}$ und halten fest, dass dann $\mathbf{Az} = \mathbf{y}$ gilt. Dann folgt

$$\langle \mathbf{y}, \mathbf{A}^{-1}\mathbf{y} \rangle_2 = \langle \mathbf{Az}, \mathbf{A}^{-1}\mathbf{Az} \rangle_2 = \langle \mathbf{Az}, \mathbf{z} \rangle_2 = \langle \mathbf{z}, \mathbf{Az} \rangle_2 > 0,$$

da $\mathbf{A}$ positiv definit ist und $\mathbf{z} \neq \mathbf{0}$ gilt. ∎

Aus Satz 3.25 folgt damit insbesondere, dass eine positiv definite Matrix immer eine LR-Zerlegung besitzt. Für eine *selbstadjungierte* positiv definite Matrix können wir sogar eine besonders einfache LR-Zerlegung konstruieren:

Definition 3.48 (Cholesky-Zerlegung) *Sei $\mathbf{A} \in \mathbb{K}^{n \times n}$ eine Matrix. Falls eine untere Dreiecksmatrix $\mathbf{L} \in \mathbb{K}^{n \times n}$ die Gleichung*

$$\mathbf{LL}^* = \mathbf{A} \tag{3.18}$$

erfüllt, bezeichnen wir sie als Cholesky-Faktor *der Matrix $\mathbf{A}$. Die Zerlegung (3.18) wird häufig als* Cholesky-Zerlegung *bezeichnet.*

Offenbar kann eine Zerlegung der Form (3.18) nur existieren, wenn $\mathbf{A}$ selbstadjungiert ist. Falls $\mathbf{L}$ ein Cholesky-Faktor einer Matrix $\mathbf{A}$ ist, gilt für beliebige Vektoren $\mathbf{y} \in \mathbb{K}^n$ nach (3.17) die Gleichung

$$\langle \mathbf{y}, \mathbf{Ay} \rangle_2 = \langle \mathbf{y}, \mathbf{LL}^*\mathbf{y} \rangle_2 = \langle \mathbf{L}^*\mathbf{y}, \mathbf{L}^*\mathbf{y} \rangle_2 = \|\mathbf{L}^*\mathbf{y}\|_2^2.$$

Falls $\mathbf{L}^*$ also regulär ist, und damit auch $\mathbf{L}$, muss $\mathbf{A}$ positiv definit sein. Wir können beweisen, dass diese beiden Eigenschaften bereits die Existenz des Cholesky-Faktors implizieren:

Satz 3.49 (Cholesky-Zerlegung) *Sei $\mathbf{A} \in \mathbb{K}^{n \times n}$ eine selbstadjungierte positiv definite Matrix. Dann existiert ein regulärer Cholesky-Faktor $\mathbf{L}$ der Matrix $\mathbf{A}$.*

Beweis. Wir gehen wieder induktiv vor.

Induktionsanfang: Da $\mathbf{A}$ positiv definit ist, gilt $a_{11} > 0$, so dass wir $\ell_{11} := \sqrt{a_{11}}$ setzen können.

Induktionsvoraussetzung: Sei $n \in \mathbb{N}$ so gegeben, dass für jede selbstadjungierte positiv definite Matrix $\mathbf{A} \in \mathbb{K}^{n\times n}$ ein regulärer Cholesky-Faktor existiert.

Induktionsschritt: Sei $\mathbf{A} \in \mathbb{K}^{(n+1)\times(n+1)}$ selbstadjungiert und positiv definit. Wir definieren

$$\mathbf{A}_{**} := \begin{pmatrix} a_{22} & \dots & a_{2,n+1} \\ \vdots & \ddots & \vdots \\ a_{n+1,2} & \dots & a_{n+1,n+1} \end{pmatrix}, \quad \mathbf{A}_{1*} := \begin{pmatrix} a_{12} & \dots & a_{1,n+1} \end{pmatrix}, \quad \mathbf{A}_{*1} := \begin{pmatrix} a_{21} \\ \vdots \\ a_{n+1,1} \end{pmatrix},$$
$$\mathbf{L}_{**} := \begin{pmatrix} \ell_{22} & & \\ \vdots & \ddots & \\ \ell_{n+1,2} & \dots & \ell_{n+1,n+1} \end{pmatrix}, \qquad \mathbf{L}_{*1} := \begin{pmatrix} \ell_{21} \\ \vdots \\ \ell_{n+1,1} \end{pmatrix}$$

und erhalten die Gleichung

$$\begin{pmatrix} a_{11} & \mathbf{A}_{1*} \\ \mathbf{A}_{*1} & \mathbf{A}_{**} \end{pmatrix} = \mathbf{A} = \mathbf{L}\mathbf{L}^* = \begin{pmatrix} \ell_{11} & \\ \mathbf{L}_{*1} & \mathbf{L}_{**} \end{pmatrix} \begin{pmatrix} \overline{\ell}_{11} & \mathbf{L}_{*1}^* \\ & \mathbf{L}_{**}^* \end{pmatrix},$$

die zu den Gleichungen

$$\begin{aligned} a_{11} &= |\ell_{11}|^2, & \mathbf{A}_{1*} &= \ell_{11}\mathbf{L}_{*1}^*, \\ \mathbf{A}_{*1} &= \mathbf{L}_{*1}\overline{\ell}_{11}, & \mathbf{A}_{**} &= \mathbf{L}_{*1}\mathbf{L}_{*1}^* + \mathbf{L}_{**}\mathbf{L}_{**}^* \end{aligned}$$

äquivalent ist. Da $\mathbf{A}$ positiv definit ist, ist nach Lemma 3.47 auch $a_{11} \in \mathbb{R}_{>0}$, also können wir $\ell_{11} := \sqrt{a_{11}}$ setzen, um die erste Gleichung zu erfüllen. Die zweite und dritte Gleichung sind äquivalent, da $\mathbf{A}_{*1} = \mathbf{A}_{1*}^*$ infolge der Selbstadjungiertheit der Matrix $\mathbf{A}$ gilt, und beide Gleichungen können erfüllt werden, indem wir $\mathbf{L}_{*1} := \mathbf{A}_{*1}\ell_{11}^{-1}$ setzen (da ℓ_{11} reell ist, haben wir $\ell_{11} = \overline{\ell}_{11}$).

Damit bleibt nur noch die Gleichung

$$\mathbf{L}_{**}\mathbf{L}_{**}^* = \mathbf{A}_{**} - \mathbf{L}_{*1}\mathbf{L}_{*1}^* = \mathbf{A}_{**} - \mathbf{A}_{*1}a_{11}^{-1}\mathbf{A}_{1*} =: \widehat{\mathbf{A}}$$

zu untersuchen. Um die Induktionsvoraussetzung auf $\widehat{\mathbf{A}}$ anwenden zu können, müssen wir nachweisen, dass diese Matrix selbstadjungiert und positiv definit ist. Ersteres ist offensichtlich, zum Nachweis der letzteren Eigenschaft verwenden wir die Hilfsmatrix

$$\mathbf{T} := \begin{pmatrix} -\mathbf{A}_{1*}a_{11}^{-1} \\ \mathbf{I} \end{pmatrix} \in \mathbb{K}^{(n+1)\times n}.$$

Für diese Matrix gilt

$$\mathbf{T}^*\mathbf{A}\mathbf{T} = \begin{pmatrix} -\mathbf{A}_{*1}a_{11}^{-1} & \mathbf{I} \end{pmatrix} \begin{pmatrix} a_{11} & \mathbf{A}_{1*} \\ \mathbf{A}_{*1} & \mathbf{A}_{**} \end{pmatrix} \begin{pmatrix} -a_{11}^{-1}\mathbf{A}_{1*} \\ \mathbf{I} \end{pmatrix} = -\mathbf{A}_{*1}a_{11}^{-1}\mathbf{A}_{1*} + \mathbf{A}_{**} = \widehat{\mathbf{A}}.$$

Sei nun $\widehat{\mathbf{y}} \in \mathbb{K}^n \setminus \{\mathbf{0}\}$. Für $\mathbf{y} := \mathbf{T}\widehat{\mathbf{y}} \in \mathbb{K}^{n+1} \setminus \{\mathbf{0}\}$ erhalten wir dank (3.17)

$$\langle \widehat{\mathbf{A}}\widehat{\mathbf{y}}, \widehat{\mathbf{y}} \rangle_2 = \langle \mathbf{T}^* \mathbf{A} \mathbf{T} \widehat{\mathbf{y}}, \widehat{\mathbf{y}} \rangle_2 = \langle \mathbf{A}\mathbf{T}\widehat{\mathbf{y}}, \mathbf{T}\widehat{\mathbf{y}} \rangle_2 = \langle \mathbf{A}\mathbf{y}, \mathbf{y} \rangle_2 > 0,$$

also ist $\widehat{\mathbf{A}}$ positiv definit und wir können die Induktionsvoraussetzung anwenden. ∎

Da der Beweis von Satz 3.49 konstruktiv ist, können wir direkt einen Algorithmus zur Berechnung des Cholesky-Faktors gewinnen, indem wir die Induktion auflösen und, wie schon im Falle der LR-Zerlegung, die Matrix $\mathbf{A}_{**}$ mit der im Zuge der Konstruktion verwendeten Hilfsmatrix $\widehat{\mathbf{A}}$ überschreiben. Die korrespondierende Rechenvorschrift ist in Abbildung 3.8 zusammengefasst.

---

procedure decomp_cholesky(n, **var** $\mathbf{A}$)
for $i = 1, \ldots, n$ **do**
  if $a_{ii} \notin \mathbb{R}_{>0}$ **then**
    Abbruch, $\mathbf{A}$ ist nicht selbstadjungiert positiv definit
  end if
  $a_{ii} \leftarrow \sqrt{a_{ii}}$
  for $j \in [i+1 : n]$ **do**
    $a_{ji} \leftarrow a_{ji}/a_{ii}$
  end for
  for $j \in [i+1 : n],\ k \in [i+1 : j]$ **do**
    $a_{jk} \leftarrow a_{jk} - a_{ji}\bar{a}_{ki}$
  end for
end for

Abbildung 3.8: Berechnung der Cholesky-Zerlegung. Die linke untere Hälfte der Matrix $\mathbf{A}$ wird mit dem Cholesky-Faktor $\mathbf{L}$ überschrieben.

Da dieser Algorithmus lediglich Einträge aus der linken unteren Dreieckshälfte der Matrix $\mathbf{A}$ verwendet, benötigt er lediglich $(n+1)n/2$ Speicherplätze, also genausoviel wie die Darstellung der ursprünglichen selbstadjungierten Matrix.

Wenden wir uns nun der Analyse des Rechenaufwands des Verfahrens zu. Für jedes $i \in [1 : n]$ werden eine Wurzel, $n - i$ Divisionen und $(n - i + 1)(n - i)/2$ Produkte und ebensoviele Subtraktionen berechnet, es fallen also insgesamt

$$\sum_{i=1}^{n} 1 + (n-i) + (n-i+1)(n-i) = \sum_{\ell=0}^{n-1} 1 + \ell + (\ell+1)\ell = \sum_{\ell=0}^{n-1} (\ell+1)^2 = \sum_{\ell=1}^{n} \ell^2$$

Gleitkommaoperationen an. Mit Lemma 3.22 folgt, dass die Berechnung des Cholesky-Faktors

$$\frac{n}{6}(2n+1)(n+1) = \frac{n(2n^2+3n+1)}{6} = \frac{1}{3}n^3 + \frac{n}{6}(3n+1) \leq \frac{1}{3}n^3 + \frac{2}{3}n^2$$

Operationen benötigt. Die Cholesky-Zerlegung benötigt also nicht nur ungefähr halb soviel Speicher wie die LR-Zerlegung, sie lässt sich für größere Werte von n auch ungefähr doppelt so schnell berechnen.

```
procedure adjback_subst(L, n, var b);
for j = n, ..., 1 do
  b_j ← b_j / ℓ_jj
  for i ∈ [1 : j − 1] do
    b_i ← b_i − ℓ̄_ji b_j
  end for
end for
```

Abbildung 3.9: Adjungiertes Rückwärtseinsetzen zur Lösung von $\mathbf{L}^*\mathbf{x} = \mathbf{b}$. Die rechte Seite $\mathbf{b}$ wird mit der Lösung $\mathbf{x}$ überschrieben.

Um das Gleichungssystem $\mathbf{Ax} = \mathbf{b}$ zu lösen, lösen wir zwei Dreieckssysteme

$$\mathbf{Ly} = \mathbf{b}, \qquad \mathbf{L}^*\mathbf{x} = \mathbf{y}.$$

Die erste Gleichung können wir wie gehabt durch Vorwärtseinsetzen behandeln, für die zweite Gleichung müssen wir rückwärts in die adjungierte Matrix einsetzen (siehe den Algorithmus in Abbildung 3.9). Insgesamt benötigen wir dafür $2n^2$ Operationen.

Übungsaufgabe 3.50 (Minimierungsaufgabe) *Sei $\mathbf{A} \in \mathbb{K}^{n\times n}$ selbstadjungiert und positiv definit, und sei $\mathbf{b} \in \mathbb{K}^n$. Wir definieren die Funktion*

$$J\colon \mathbb{K}^n \to \mathbb{R}, \qquad \mathbf{x} \mapsto \langle \mathbf{x}, \mathbf{Ax}\rangle_2 - 2\operatorname{Re}\langle \mathbf{b}, \mathbf{x}\rangle_2.$$

Beweisen Sie, dass

$$J(\mathbf{x}) \leq J(\mathbf{y}) \qquad \text{für alle } \mathbf{y} \in \mathbb{K}^n$$

genau dann gilt, wenn $\mathbf{Ax} = \mathbf{b}$ gilt.

Hinweis: *Mit der Cholesky-Zerlegung $\mathbf{A} = \mathbf{LL}^*$ lässt sich die* Energienorm $\|\mathbf{x}\|_A := \|\mathbf{L}^*\mathbf{x}\|_2$ *definieren. Wenn $\mathbf{x}^*$ die Lösung des linearen Gleichungssystems $\mathbf{Ax}^* = \mathbf{b}$ ist, in welcher Beziehung steht dann $\|\mathbf{x} - \mathbf{x}^*\|_A^2$ zu $J(\mathbf{x})$?*

Übungsaufgabe 3.51 (Richardson-Iteration) *Sei $\mathbf{A} \in \mathbb{K}^{n\times n}$ selbstadjungiert und positiv definit, und sei $\mathbf{b} \in \mathbb{K}^n$. Sei $\mathbf{x} \in \mathbb{K}^n$ die Lösung des linearen Gleichungssystems $\mathbf{Ax} = \mathbf{b}$.*

Die Richardson-Iteration *für den* Dämpfungsparameter $\theta \in \mathbb{R}_{>0}$ *und den* Startvektor $\mathbf{x}^{(0)} \in \mathbb{K}^n$ *ist definiert durch*

$$\mathbf{x}^{(m+1)} := \mathbf{x}^{(m)} + \theta(\mathbf{b} - \mathbf{Ax}^{(m)}) \qquad \text{für alle } m \in \mathbb{N}_0.$$

Beweisen Sie, dass $\lim_{m\to\infty} \mathbf{x}^{(m)} = \mathbf{x}$ gilt, falls $\theta < 2/\|\mathbf{A}\|_2$ gesichert ist.

Hinweis: *Mit Hilfe des Satzes von Heine-Borel lässt sich zeigen, dass ein $\alpha \in \mathbb{R}_{>0}$ mit $\alpha\|\mathbf{y}\|_2^2 \leq \langle \mathbf{y}, \mathbf{Ay}\rangle_2$ für alle $\mathbf{x} \in \mathbb{K}^n$ existiert. Mit der Cauchy-Schwarz-Ungleichung erhält man eine obere Schranke für diesen Ausdruck und kann die Entwicklung der Fehler $\mathbf{e}^{(m)} := \mathbf{x}^{(m)} - \mathbf{x}$ untersuchen.*

Übungsaufgabe 3.52 (Cauchy-Schwarz-Ungleichung) *Sei $\mathbf{A} \in \mathbb{K}^{n\times n}$ eine selbstadjungierte und positiv definite Matrix. Zeigen Sie*

$$|\langle \mathbf{x}, \mathbf{y}\rangle_2|^2 \leq \langle \mathbf{x}, \mathbf{A}\mathbf{x}\rangle_2 \langle \mathbf{y}, \mathbf{A}^{-1}\mathbf{y}\rangle_2 \qquad \text{für alle } \mathbf{x}, \mathbf{y} \in \mathbb{K}^n.$$

Hinweis: *Der Beweis lässt sich sehr kurz führen, wenn man die Cholesky-Zerlegung und die Cauchy-Schwarz-Ungleichung (3.14) verwendet. Ein allgemeinerer Beweis beruht darauf, dass man $0 \leq \langle \mathbf{x} - \lambda\mathbf{A}^{-1}\mathbf{y}, \mathbf{A}(\mathbf{x} - \lambda\mathbf{A}^{-1}\mathbf{y})\rangle_2$ für alle $\lambda \in \mathbb{K}$ ausnutzt.*

Übungsaufgabe 3.53 (Injektiv und surjektiv) *Eine Matrix $\mathbf{A} \in \mathbb{K}^{m\times n}$ kann immer als lineare Abbildung von $\mathbb{K}^n$ nach $\mathbb{K}^m$ interpretiert werden, und in dieser Interpretation kann sie injektiv oder surjektiv sein.*

Beweisen Sie:

- *Falls $\mathbf{A}$ surjektiv ist, ist $\mathbf{A}^*$ injektiv.*
- *Falls $\mathbf{A}$ injektiv ist, ist $\mathbf{A}^*$ surjektiv.*

Hinweis: *Beide Beweise lassen sich mit Lemma 3.44 führen. Ohne Beweis kann die folgende Aussage verwendet werden: Für einen echten Teilraum $\mathcal{V} \subsetneq \mathbb{K}^n$ findet man einen Vektor $\mathbf{x} \in \mathbb{K}^n \setminus \mathcal{V}$ derart, dass $\langle \mathbf{x}, \mathbf{y}\rangle_2 = 0$ für alle $\mathbf{y} \in \mathcal{V}$ gilt.*

Übungsaufgabe 3.54 (Spektralnorm) *Die von der Euklidischen Norm induzierte Matrixnorm*

$$\|\mathbf{A}\|_2 := \max\left\{ \frac{\|\mathbf{A}\mathbf{z}\|_2}{\|\mathbf{z}\|_2} \;:\; \mathbf{z} \in \mathbb{K}^n \setminus \{\mathbf{0}\} \right\} \qquad \text{für alle } \mathbf{A} \in \mathbb{K}^{m\times n},\ m, n \in \mathbb{N},$$

wird als Spektralnorm *bezeichnet. Sei $\mathbf{A} \in \mathbb{K}^{m\times n}$.*

(a) Beweisen Sie

$$\|\mathbf{A}\|_2 = \max\left\{ \frac{|\langle \mathbf{y}, \mathbf{A}\mathbf{z}\rangle_2|}{\|\mathbf{y}\|_2\|\mathbf{z}\|_2} \;:\; \mathbf{z} \in \mathbb{K}^n \setminus \{\mathbf{0}\},\ \mathbf{y} \in \mathbb{K}^m \setminus \{\mathbf{0}\} \right\}.$$

(b) Beweisen Sie $\|\mathbf{A}\|_2 = \|\mathbf{A}^*\|_2$.

(c) Beweisen Sie $\|\mathbf{A}\|_2 = \|\mathbf{A}^*\mathbf{A}\|_2^{1/2}$.

(d) Beweisen Sie $\|\mathbf{A}\|_2 = \|\mathbf{A}\mathbf{A}^*\|_2^{1/2}$.

Übungsaufgabe 3.55 (inf-sup-Bedingung) *Sei $\mathbf{A} \in \mathbb{K}^{n\times n}$.*

(a) Sei $\mathbf{A}$ regulär. Sei $\alpha := 1/\|\mathbf{A}\|_2$. Beweisen Sie

$$\alpha \leq \inf\left\{ \sup\left\{ \frac{|\langle \mathbf{y}, \mathbf{A}\mathbf{z}\rangle_2|}{\|\mathbf{y}\|_2\,\|\mathbf{z}\|_2} \;:\; \mathbf{y} \in \mathbb{K}^n \setminus \{\mathbf{0}\} \right\} \;:\; \mathbf{z} \in \mathbb{K}^n \setminus \{\mathbf{0}\} \right\}. \tag{3.19}$$

(b) Sei $\alpha \in \mathbb{R}_{>0}$ mit (3.19) gegeben. Zeigen Sie, dass $\mathbf{A}$ regulär ist mit $\|\mathbf{A}^{-1}\|_2 \leq 1/\alpha$.

Übungsaufgabe 3.56 (Dualität) *Beweisen Sie für beliebige Vektoren* $\mathbf{x} \in \mathbb{K}^n$ *die Gleichungen*

$$\|\mathbf{x}\|_\infty = \max\left\{ \frac{|\langle \mathbf{x}, \mathbf{y}\rangle_2|}{\|\mathbf{y}\|_1} \ : \ \mathbf{y} \in \mathbb{K}^n \setminus \{\mathbf{0}\} \right\},$$
$$\|\mathbf{x}\|_1 = \max\left\{ \frac{|\langle \mathbf{x}, \mathbf{y}\rangle_2|}{\|\mathbf{y}\|_\infty} \ : \ \mathbf{y} \in \mathbb{K}^n \setminus \{\mathbf{0}\} \right\}.$$

Folgern Sie daraus für beliebige Matrizen $\mathbf{A} \in \mathbb{K}^{m\times n}$ *die Gleichungen*

$$\|\mathbf{A}\|_\infty = \max\left\{ \frac{|\langle \mathbf{y}, \mathbf{A}\mathbf{x}\rangle_2|}{\|\mathbf{y}\|_1\|\mathbf{x}\|_\infty} \ : \ \mathbf{x} \in \mathbb{K}^m \setminus \{\mathbf{0}\},\ \mathbf{y} \in \mathbb{K}^n \setminus \{\mathbf{0}\} \right\},$$
$$\|\mathbf{A}\|_1 = \max\left\{ \frac{|\langle \mathbf{y}, \mathbf{A}\mathbf{x}\rangle_2|}{\|\mathbf{y}\|_\infty\|\mathbf{x}\|_1} \ : \ \mathbf{x} \in \mathbb{K}^m \setminus \{\mathbf{0}\},\ \mathbf{y} \in \mathbb{K}^n \setminus \{\mathbf{0}\} \right\}$$

und insbesondere $\|\mathbf{A}\|_\infty = \|\mathbf{A}^*\|_1$.

Hinweis: *Offenbar ist es wichtig, geeignete Vektoren* $\mathbf{y}$ *zu wählen. Einerseits können kanonische Einheitsvektoren nützlich sein, also Vektoren, bei denen ein Koeffizient gleich eins ist und alle anderen gleich null sind. Andererseits können Vektoren hilfreich sein, deren Koeffizienten alle den Betrag eins aufweisen.*

Übungsaufgabe 3.57 (Positiv semidefinite Matrizen) *Sei* $\mathbf{A} \in \mathbb{K}^{n\times n}$ *selbstadjungiert. Wir setzen voraus, dass* $\mathbf{A}$ positiv semidefinit *ist, wir haben also*

$$\langle \mathbf{y}, \mathbf{A}\mathbf{y}\rangle_2 \in \mathbb{R}_{\geq 0} \qquad \text{für alle } \mathbf{x} \in \mathbb{K}^n.$$

(a) Sei $\mathbf{x} \in \mathbb{K}^n$. *Beweisen Sie*

$$2\lambda\|\mathbf{A}\mathbf{x}\|_2^2 \leq \langle \mathbf{x}, \mathbf{A}\mathbf{x}\rangle_2 + \lambda^2\langle \mathbf{A}\mathbf{x}, \mathbf{A}^2\mathbf{x}\rangle_2 \qquad \text{für alle } \lambda \in \mathbb{R}_{\geq 0}.$$

(b) Sei $\mathbf{x} \in \mathbb{K}^n$. *Folgern Sie aus Teil (a), dass aus* $\langle \mathbf{x}, \mathbf{A}\mathbf{x}\rangle_2 = 0$ *bereits* $\mathbf{A}\mathbf{x} = \mathbf{0}$ *folgt.*

(c) Seien $\mathbf{x}, \mathbf{y} \in \mathbb{K}^n$. *Beweisen Sie die verallgemeinerte Cauchy-Schwarz-Ungleichung*

$$\langle \mathbf{x}, \mathbf{A}\mathbf{y}\rangle_2 \leq \sqrt{\langle \mathbf{x}, \mathbf{A}\mathbf{x}\rangle_2}\sqrt{\langle \mathbf{y}, \mathbf{A}\mathbf{y}\rangle_2}.$$

Hinweise: *Für Teil (a) könnte sich ein Blick auf Vektoren der Form* $\mathbf{y} = \mathbf{x} + \lambda\mathbf{A}\mathbf{x}$ *lohnen. Für Teil (c) könnte man den Beweis des Lemmas 3.40 anpassen.*

Übungsaufgabe 3.58 (Pivotisierte Cholesky-Zerlegung) *Sei* $\mathbf{A} \in \mathbb{K}^{n\times n}$ *selbstadjungiert und positiv semidefinit.*

(a) Beweisen Sie, dass ein $k \in [1:n]$ *mit* $a_{kk} > 0$ *existiert, falls* $\mathbf{A} \neq \mathbf{0}$ *gilt.*

(b) Beweisen Sie, dass eine Permutationsmatrix $\mathbf{P}_\pi \in \mathbb{K}^{n\times n}$ *und eine linke untere Dreiecksmatrix* $\mathbf{L} \in \mathbb{K}^{n\times n}$ *existieren mit*

$$\mathbf{P}_\pi\mathbf{A}\mathbf{P}_\pi^* = \mathbf{L}\mathbf{L}^*.$$

Hinweis: *Für Teil (a) ist die Übungsaufgabe 3.57 nützlich. Für Teil (b) kann der Beweis des Satzes 3.49 als Anregung dienen.*

3.6 Orthogonale Zerlegungen

Die bisher diskutierten Lösungsverfahren basieren auf Dreieckszerlegungen der Matrix $\mathbf{A}$. Obwohl das Lösen der Systeme $\mathbf{Ly} = \mathbf{b}$, $\mathbf{Rx} = \mathbf{y}$ theoretisch zu der exakten Lösung führen sollte, treten in der praktischen Durchführung auf dem Computer Rundungsfehler auf, wenn mit Maschinenzahlen gearbeitet werden muss. Bei sehr großen und schlecht konditionierten Gleichungssystemen können diese Rundungsfehler so sehr verstärkt werden, dass ein sehr ungenaues Ergebnis entsteht. Wir sollten also untersuchen, welchen Einfluss Rundungsfehler auf die berechnete Lösung haben.

Das können wir mit Hilfe der Rückwärtsanalyse tun: Im ersten Schritt lösen wir $\mathbf{Ly} = \mathbf{b}$. Man kann zeigen, dass die mit Rundungsfehlern berechnete Näherungslösung $\widetilde{\mathbf{y}}$ die exakte Lösung eines gestörten Gleichungssystems $\widetilde{\mathbf{L}}\widetilde{\mathbf{y}} = \widetilde{\mathbf{b}}$ ist, wobei die relativen Fehler

$$\frac{\|\mathbf{L} - \widetilde{\mathbf{L}}\|_\infty}{\|\mathbf{L}\|_\infty} \leq C_1\, \epsilon_{\text{fl}}, \qquad \frac{\|\mathbf{b} - \widetilde{\mathbf{b}}\|_\infty}{\|\mathbf{b}\|_\infty} \leq C_1\, \epsilon_{\text{fl}}$$

mit einer geeigneten Konstanten C_1 erfüllen, so dass wir mit der Satz 3.9 auf

$$\frac{\|\mathbf{y} - \widetilde{\mathbf{y}}\|_\infty}{\|\mathbf{y}\|_\infty} \leq 2\, C_L \kappa_\infty(\mathbf{L}) \epsilon_{\text{fl}}$$

mit einer geeigneten Konstanten C_R schließen dürfen. Im zweiten Schritt lösen wir $\mathbf{Rx} = \widetilde{\mathbf{y}}$. Die mit Rundungsfehlern berechnete Näherungslösung $\widetilde{\mathbf{x}}$ ist dann die exakte Lösung eines gestörten Gleichungssystems $\widetilde{\mathbf{R}}\widetilde{\mathbf{x}} = \widehat{\mathbf{y}}$, wobei wieder für die relativen Fehler

$$\frac{\|\mathbf{R} - \widetilde{\mathbf{R}}\|_\infty}{\|\mathbf{R}\|_\infty} \leq C_2\, \epsilon_{\text{fl}}, \qquad \frac{\|\widetilde{\mathbf{y}} - \widehat{\mathbf{y}}\|_\infty}{\|\widetilde{\mathbf{y}}\|_\infty} \leq C_2\, \epsilon_{\text{fl}}$$

mit einem geeigneten C_2 gilt. Wenn wir von $\|\mathbf{y}\|_\infty \approx \|\widetilde{\mathbf{y}}\|_\infty$ ausgehen, erhalten wir mit der Dreiecksungleichung

$$\begin{aligned}\frac{\|\mathbf{y} - \widehat{\mathbf{y}}\|_\infty}{\|\mathbf{y}\|_\infty} &\leq \frac{\|\mathbf{y} - \widetilde{\mathbf{y}}\|_\infty}{\|\mathbf{y}\|_\infty} + \frac{\|\widetilde{\mathbf{y}} - \widehat{\mathbf{y}}\|_\infty}{\|\mathbf{y}\|_\infty} \\ &\approx \frac{\|\mathbf{y} - \widetilde{\mathbf{y}}\|_\infty}{\|\mathbf{y}\|_\infty} + \frac{\|\widetilde{\mathbf{y}} - \widehat{\mathbf{y}}\|_\infty}{\|\widetilde{\mathbf{y}}\|_\infty} \leq 2\, C_L \kappa_\infty(\mathbf{L}) \epsilon_{\text{fl}} + C_2\, \epsilon_{\text{fl}},\end{aligned}$$

so dass mit Satz 3.9 die Abschätzung

$$\frac{\|\mathbf{x} - \widetilde{\mathbf{x}}\|_\infty}{\|\mathbf{x}\|_\infty} \lesssim C_R \kappa_\infty(\mathbf{R})\, (2\, C_L \kappa_\infty(\mathbf{L}) \epsilon_{\text{fl}} + 2\, C_2 \epsilon_{\text{fl}}) \approx 2\, C_L C_R \kappa_\infty(\mathbf{L}) \kappa_\infty(\mathbf{R}) \epsilon_{\text{fl}}$$

für eine geeignete Konstante C_L folgt. Die relative Genauigkeit der mit Hilfe der LR-Zerlegung berechneten Lösung hängt also entscheidend von den Konditionszahlen der Matrizen $\mathbf{L}$ und $\mathbf{R}$ ab, statt von der Konditionszahl der Matrix $\mathbf{A}$, die für die ursprüngliche Aufgabenstellung entscheidend ist. Leider können die Konditionszahlen der beiden Dreiecksmatrizen wesentlich ungünstiger als die der Matrix $\mathbf{A}$ sein.

Beispiel 3.59 (Schlecht konditionierte LR-Zerlegung) *Wir untersuchen die Matrix*

$$\mathbf{A}_\epsilon := \begin{pmatrix} 1 & 1/\epsilon \\ 1/\epsilon & 1 \end{pmatrix} \tag{3.20}$$

für ein $\epsilon \in (0,1)$. Ihre Inverse ist durch

$$\mathbf{A}_\epsilon^{-1} = \frac{1}{1-\epsilon^2} \begin{pmatrix} -\epsilon^2 & \epsilon \\ \epsilon & -\epsilon^2 \end{pmatrix}$$

gegeben, die Zeilensummennormen lassen sich dank Lemma 3.4 direkt berechnen:

$$\|\mathbf{A}_\epsilon\|_\infty = 1 + 1/\epsilon, \qquad \|\mathbf{A}_\epsilon^{-1}\|_\infty = \frac{\epsilon + \epsilon^2}{1-\epsilon^2},$$

so dass wir für die Konditionszahl

$$\kappa_\infty(\mathbf{A}_\epsilon) = \|\mathbf{A}_\epsilon\|_\infty \|\mathbf{A}_\epsilon^{-1}\|_\infty = \frac{(1+1/\epsilon)(\epsilon+\epsilon^2)}{1-\epsilon^2} = \frac{1+2\epsilon+\epsilon^2}{1-\epsilon^2} = \frac{(1+\epsilon)^2}{(1-\epsilon)(1+\epsilon)} = \frac{1+\epsilon}{1-\epsilon}$$

erhalten. Für $\epsilon \to 0$ strebt die Konditionszahl gegen eins, die Matrix ist also sehr gut konditioniert.

Eine LR-Zerlegung der Matrix $\mathbf{A}_\epsilon$ ist durch

$$\mathbf{L}_\epsilon := \begin{pmatrix} 1 & \\ 1/\epsilon & 1 \end{pmatrix}, \qquad \mathbf{R}_\epsilon := \begin{pmatrix} 1 & 1/\epsilon \\ & 1 - 1/\epsilon^2 \end{pmatrix}$$

gegeben, die Inversen die beiden Faktoren durch

$$\mathbf{L}_\epsilon^{-1} = \begin{pmatrix} 1 & \\ -1/\epsilon & 1 \end{pmatrix}, \qquad \mathbf{R}_\epsilon^{-1} = \begin{pmatrix} 1 & \epsilon/(1-\epsilon^2) \\ & -\epsilon^2/(1-\epsilon^2) \end{pmatrix}.$$

Um die Konditionszahlen berechnen zu können treffen wir die Annahme $\epsilon \leq 1/2$, die dafür sorgt, dass die Zeilensummennorm der Matrix $\mathbf{R}_\epsilon$ von deren zweiter Zeile bestimmt wird. Damit ergibt sich

$$\kappa_\infty(\mathbf{L}_\epsilon) = \|\mathbf{L}_\epsilon\|_\infty \|\mathbf{L}_\epsilon^{-1}\|_\infty = (1+1/\epsilon)^2 = \frac{(1+\epsilon)^2}{\epsilon^2},$$
$$\kappa_\infty(\mathbf{R}_\epsilon) = \|\mathbf{R}_\epsilon\|_\infty \|\mathbf{R}_\epsilon^{-1}\|_\infty = \frac{1-\epsilon^2}{\epsilon^2} \frac{1-\epsilon^2+\epsilon}{1-\epsilon^2} = \frac{1-\epsilon^2+\epsilon}{\epsilon^2}.$$

Wir stellen fest, dass für $\epsilon \to 0$ beide Konditionszahlen gegen unendlich streben.

Wie wir an diesem Beispiel sehen können, kann also eine LR-Zerlegung einer gutartigen Matrix zu ausgesprochen schlecht konditionierten Faktoren $\mathbf{L}$ und $\mathbf{R}$ führen, die befürchten lassen, dass keine genaue Lösung berechnet werden kann.

Um derartige Effekte zu vermeiden, ist es erstrebenswert, die Dreiecksmatrizen durch einen anderen Matrixtyp zu ersetzen, der günstigere Eigenschaften besitzt. Eine gute Wahl sind *unitäre* Matrizen:

Definition 3.60 (Unitäre Matrix) *Sei $\mathbf{Q} \in \mathbb{K}^{n\times n}$ eine Matrix. Falls $\mathbf{Q}^*\mathbf{Q} = \mathbf{I}$ gilt, heißt* $\mathbf{Q}$ unitär.

Im reellwertigen Fall $\mathbb{K} = \mathbb{R}$ werden unitäre Matrizen in der Regel als orthogonale Matrizen *bezeichnet.*

Lemma 3.61 (Unitäre Matrizen) *Sei $\mathbf{Q} \in \mathbb{K}^{n\times n}$ eine unitäre Matrix. Dann gelten:*

- $\langle \mathbf{Q}\mathbf{x}, \mathbf{Q}\mathbf{y}\rangle_2 = \langle \mathbf{x}, \mathbf{y}\rangle_2$ *für alle* $\mathbf{x}, \mathbf{y} \in \mathbb{K}^n$,
- $\|\mathbf{Q}\mathbf{x}\|_2 = \|\mathbf{x}\|_2$ *für alle* $\mathbf{x} \in \mathbb{K}^n$ *und*
- $\mathbf{Q}$ *ist regulär mit* $\mathbf{Q}^{-1} = \mathbf{Q}^*$, *und* $\mathbf{Q}^*$ *ist ebenfalls unitär.*
- *Für zwei unitäre Matrizen* $\mathbf{Q}_1, \mathbf{Q}_2 \in \mathbb{K}^{n\times n}$ *ist auch das Produkt* $\mathbf{Q}_1\mathbf{Q}_2$ *unitär.*
- *Für* $m \in \mathbb{N}$, $\mathbf{A} \in \mathbb{K}^{n\times m}$ *und* $\mathbf{B} \in \mathbb{K}^{m\times n}$ *gelten die Gleichungen* $\|\mathbf{Q}\mathbf{A}\|_2 = \|\mathbf{A}\|_2$ *und* $\|\mathbf{B}\mathbf{Q}^*\|_2 = \|\mathbf{B}\|_2$.

Beweis. Seien $\mathbf{x}, \mathbf{y} \in \mathbb{K}^n$ gegeben. Dann gilt nach (3.17)

$$\langle \mathbf{Q}\mathbf{x}, \mathbf{Q}\mathbf{y}\rangle_2 = \langle \mathbf{Q}^*\mathbf{Q}\mathbf{x}, \mathbf{y}\rangle_2 = \langle \mathbf{x}, \mathbf{y}\rangle_2,$$

also die erste Gleichung. Aus ihr folgt

$$\|\mathbf{Q}\mathbf{x}\|_2 = \sqrt{\langle \mathbf{Q}\mathbf{x}, \mathbf{Q}\mathbf{x}\rangle_2} = \sqrt{\langle \mathbf{x}, \mathbf{x}\rangle_2} = \|\mathbf{x}\|_2, \tag{3.21}$$

und damit die zweite Gleichung. Aus dieser Gleichung ergibt sich, dass der Kern von $\mathbf{Q}$ nur den Nullvektor enthalten kann, also muss nach Dimensionssatz das Bild von $\mathbf{Q}$ der gesamte Raum $\mathbb{K}^n$ sein. Somit finden wir für ein beliebiges $\mathbf{z} \in \mathbb{K}^n$ einen Vektor $\widehat{\mathbf{z}} \in \mathbb{K}^n$ mit $\mathbf{z} = \mathbf{Q}\widehat{\mathbf{z}}$ und folgern

$$\mathbf{Q}\mathbf{Q}^*\mathbf{z} = \mathbf{Q}\mathbf{Q}^*\mathbf{Q}\widehat{\mathbf{z}} = \mathbf{Q}\widehat{\mathbf{z}} = \mathbf{z},$$

also $\mathbf{Q}\mathbf{Q}^* = \mathbf{I}$, so dass sich zusammen mit der definierenden Eigenschaft $\mathbf{Q}^*\mathbf{Q} = \mathbf{I}$ insgesamt $\mathbf{Q}^* = \mathbf{Q}^{-1}$ ergibt. Außerdem haben wir $(\mathbf{Q}^*)^*\mathbf{Q}^* = \mathbf{Q}\mathbf{Q}^* = \mathbf{I}$, also ist $\mathbf{Q}^*$ unitär.

Seien nun $\mathbf{Q}_1, \mathbf{Q}_2 \in \mathbb{K}^{n\times n}$ unitäre Matrizen. Für das Produkt $\mathbf{Q}_{12} := \mathbf{Q}_1\mathbf{Q}_2$ gilt mit Lemma 3.45

$$\mathbf{Q}_{12}^*\mathbf{Q}_{12} = \mathbf{Q}_2^*\mathbf{Q}_1^*\mathbf{Q}_1\mathbf{Q}_2 = \mathbf{Q}_2^*\mathbf{Q}_2 = \mathbf{I},$$

also ist $\mathbf{Q}_{12}$ unitär.

Seien nun $m \in \mathbb{N}$, $\mathbf{A} \in \mathbb{K}^{n\times m}$ und $\mathbf{B} \in \mathbb{K}^{m\times n}$. Dank (3.21) gilt

$$\begin{aligned}\|\mathbf{Q}\mathbf{A}\|_2 &= \max\left\{\frac{\|\mathbf{Q}\mathbf{A}\mathbf{x}\|_2}{\|\mathbf{x}\|_2} \;:\; \mathbf{x} \in \mathbb{K}^m \setminus \{\mathbf{0}\}\right\} \\ &= \max\left\{\frac{\|\mathbf{A}\mathbf{x}\|_2}{\|\mathbf{x}\|_2} \;:\; \mathbf{x} \in \mathbb{K}^m \setminus \{\mathbf{0}\}\right\} = \|\mathbf{A}\|_2,\end{aligned}$$

und weil darüber hinaus $\mathbf{Q}$ surjektiv ist, folgt auch

$$\begin{aligned}\|\mathbf{B}\mathbf{Q}^*\|_2 &= \max\left\{\frac{\|\mathbf{B}\mathbf{Q}^*\mathbf{x}\|_2}{\|\mathbf{x}\|_2} \;:\; \mathbf{x}\in\mathbb{K}^n\setminus\{\mathbf{0}\}\right\}\\ &= \max\left\{\frac{\|\mathbf{B}\mathbf{Q}^*\mathbf{Q}\mathbf{y}\|_2}{\|\mathbf{Q}\mathbf{y}\|_2} \;:\; \mathbf{y}\in\mathbb{K}^n\setminus\{\mathbf{0}\}\right\}\\ &= \max\left\{\frac{\|\mathbf{B}\mathbf{y}\|_2}{\|\mathbf{y}\|_2} \;:\; \mathbf{y}\in\mathbb{K}^n\setminus\{\mathbf{0}\}\right\} = \|\mathbf{B}\|_2.\end{aligned}$$

Die Surjektivität bedeutet nämlich gerade, dass wir für jeden Vektor $\mathbf{x}\in\mathbb{K}^n\setminus\{\mathbf{0}\}$ ein Urbild $\mathbf{y}\in\mathbb{K}^n\setminus\{\mathbf{0}\}$ mit $\mathbf{x}=\mathbf{Q}\mathbf{y}$ finden können. ∎

Beispiel 3.62 (QR-Zerlegung) *Wir können die Matrix $\mathbf{A}_\epsilon$ aus Beispiel 3.59 auch mit eine unitären Transformation auf obere Dreiecksgestalt bringen: Wenn wir*

$$\mathbf{Q}_\epsilon := \frac{1}{\sqrt{1+1/\epsilon^2}}\begin{pmatrix}1 & 1/\epsilon\\ -1/\epsilon & 1\end{pmatrix}$$

setzen, erhalten wir

$$\begin{aligned}\mathbf{Q}_\epsilon^*\mathbf{Q}_\epsilon &= \frac{1}{1+1/\epsilon^2}\begin{pmatrix}1 & -1/\epsilon\\ 1/\epsilon & 1\end{pmatrix}\begin{pmatrix}1 & 1/\epsilon\\ -1/\epsilon & 1\end{pmatrix}\\ &= \frac{1}{1+1/\epsilon^2}\begin{pmatrix}1+1/\epsilon^2 & \\ & 1/\epsilon^2+1\end{pmatrix} = \mathbf{I},\\ \mathbf{Q}_\epsilon\mathbf{A}_\epsilon &= \frac{1}{\sqrt{1+1/\epsilon^2}}\begin{pmatrix}1 & 1/\epsilon\\ -1/\epsilon & 1\end{pmatrix}\begin{pmatrix}1 & 1/\epsilon\\ 1/\epsilon & 1\end{pmatrix}\\ &= \frac{1}{\sqrt{1+1/\epsilon^2}}\begin{pmatrix}1+1/\epsilon^2 & 2/\epsilon\\ & 1-1/\epsilon^2\end{pmatrix} =: \mathbf{R}_\epsilon,\end{aligned}$$

also bringt die unitäre Matrix $\mathbf{Q}_\epsilon$ die Matrix $\mathbf{A}_\epsilon$ auf obere Dreiecksgestalt. Aus

$$\begin{aligned}\|\mathbf{Q}_\epsilon\|_\infty &= \frac{1}{\sqrt{1+1/\epsilon^2}}\left\|\begin{pmatrix}1 & 1/\epsilon\\ -1/\epsilon & 1\end{pmatrix}\right\|_\infty = \frac{1+1/\epsilon}{\sqrt{1+1/\epsilon^2}} = \frac{\epsilon+1}{\sqrt{\epsilon^2+1}},\\ \|\mathbf{R}_\epsilon\|_\infty &= \frac{1}{\sqrt{1+1/\epsilon^2}}\left\|\begin{pmatrix}1+1/\epsilon^2 & 2/\epsilon\\ & 1-1/\epsilon^2\end{pmatrix}\right\|_\infty = \frac{1+1/\epsilon^2+2/\epsilon}{\sqrt{1+1/\epsilon^2}} = \frac{(\epsilon+1)^2}{\epsilon\sqrt{\epsilon^2+1}},\\ \|\mathbf{R}_\epsilon^{-1}\|_\infty &= \frac{\sqrt{1+1/\epsilon^2}}{(1/\epsilon^2+1)(1/\epsilon^2-1)}\left\|\begin{pmatrix}1-1/\epsilon^2 & -2/\epsilon\\ & 1+1/\epsilon^2\end{pmatrix}\right\|_\infty = \frac{\epsilon\sqrt{\epsilon^2+1}(1-\epsilon^2+2\epsilon)}{(1+\epsilon^2)(1-\epsilon^2)}\end{aligned}$$

folgen

$$\begin{aligned}\kappa_\infty(\mathbf{Q}_\epsilon) &= \|\mathbf{Q}_\epsilon\|_\infty\|\mathbf{Q}_\epsilon^*\|_\infty = \frac{(1+\epsilon)^2}{1+\epsilon^2},\\ \kappa_\infty(\mathbf{R}_\epsilon) &= \|\mathbf{R}_\epsilon\|_\infty\|\mathbf{R}_\epsilon^{-1}\|_\infty = \frac{(1+\epsilon)^2(1+2\epsilon-\epsilon^2)}{(1+\epsilon^2)(1-\epsilon^2)}\end{aligned}$$

Für $\epsilon\to 0$ streben beide Konditionszahlen gegen eins, verhalten sich also sehr gutartig.

Unser Ziel ist es nun, allgemeine Matrizen mit Hilfe unitärer Transformationen auf obere Dreiecksgestalt zu bringen.

Definition 3.63 (QR-Zerlegung) *Sei $\mathbf{A} \in \mathbb{K}^{m\times n}$ eine Matrix. Falls eine unitäre Matrix $\mathbf{Q} \in \mathbb{K}^{m\times m}$ und eine obere Dreiecksmatrix $\mathbf{R} \in \mathbb{K}^{m\times n}$ die Gleichung*

$$\mathbf{QR} = \mathbf{A}$$

erfüllen, bezeichnen wir das Paar $(\mathbf{Q}, \mathbf{R})$ als eine QR-Zerlegung *der Matrix $\mathbf{A}$.*

Ähnlich wie die LR-Zerlegung lässt sich auch die QR-Zerlegung einer quadratischen Matrix direkt zum Lösen eines Gleichungssystems verwenden: Falls $\mathbf{QR} = \mathbf{A}$ gilt, können wir das System $\mathbf{Ax} = \mathbf{b}$ lösen, indem wir

$$\mathbf{b} = \mathbf{Ax} = \mathbf{QRx} = \mathbf{Qy}, \qquad \mathbf{y} = \mathbf{Rx} \tag{3.22}$$

auflösen: Da $\mathbf{Q}$ unitär ist, gilt $\mathbf{Q}^{-1} = \mathbf{Q}^*$, also können wir die erste Gleichung mit $\mathbf{y} = \mathbf{Q}^*\mathbf{b}$ lösen. Die zweite Gleichung behandeln wir wie üblich per Rückwärtseinsetzen, also etwa mit dem Algorithmus aus Abbildung 3.2.

Bemerkung 3.64 (Fehlerfortpflanzung) *Sei $(\mathbf{Q}, \mathbf{R})$ eine QR-Zerlegung einer invertierbaren Matrix $\mathbf{A}$.*
Mit Lemma 3.61 finden wir

$$\begin{aligned} &\|\mathbf{Q}\|_2 = \|\mathbf{I}\|_2 = 1, && \|\mathbf{Q}^{-1}\|_2 = \|\mathbf{I}\|_2 = 1, && \kappa_2(\mathbf{Q}) = 1, \\ &\|\mathbf{R}\|_2 = \|\mathbf{A}\|_2, && \|\mathbf{R}^{-1}\|_2 = \|\mathbf{A}^{-1}\|_2, && \kappa_2(\mathbf{R}) = \kappa_2(\mathbf{A}), \end{aligned}$$

also dürfen wir darauf hoffen, dass sich Fehler bei der Lösungsstrategie (3.22) so günstig wie möglich fortpflanzen.

Der erste Schritt bei der Konstruktion einer LR-Zerlegung besteht darin, eine Transformation durchzuführen, die in der ersten Spalte der Matrix Nulleinträge unterhalb der Diagonalen entstehen lässt. Für die Konstruktion einer QR-Zerlegung wäre es sinnvoll, wenn wir dieses Ziel auch mit einer unitären Transformation erreichen könnten.

Definition 3.65 (Householder-Spiegelung) *Sei $\mathbf{v} \in \mathbb{K}^n \setminus \{\mathbf{0}\}$ ein Vektor. Die Matrix*

$$\mathbf{Q}_v := \mathbf{I} - 2\frac{\mathbf{vv}^*}{\mathbf{v}^*\mathbf{v}} \tag{3.23}$$

heißt Householder-Spiegelung *zu $\mathbf{v}$.*

In dieser Definition verwenden wir eine formal etwas unsaubere Notation: Wir identifizieren den Vektor $\mathbf{v} \in \mathbb{K}^n$ in naheliegender Weise mit einer Matrix mit n Zeilen und einer Spalte, so dass $\mathbf{vv}^* \in \mathbb{K}^{n\times n}$ und $\mathbf{v}^*\mathbf{v} \in \mathbb{K}^{1\times 1}$ Matrizen sind. Die letztere Matrix

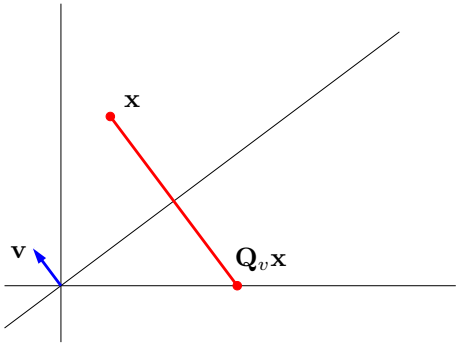

Abbildung 3.10: Geometrische Interpretation der Householder-Spiegelung

identifizieren wir mit einer Zahl, nämlich ihrem Koeffizienten, so dass die Division durch $\mathbf{v}^*\mathbf{v}$ wegen

$$\mathbf{v}^*\mathbf{v} = \sum_{i=1}^{n} \bar{v}_i v_i = \langle \mathbf{v}, \mathbf{v} \rangle_2 = \|\mathbf{v}\|_2^2 \neq 0$$

wohldefiniert ist. In dieser Weise lässt sich die Anwendung der Householder-Spiegelung auf einen Vektor $\mathbf{z} \in \mathbb{K}^n$ in der Form

$$\mathbf{Q}_v\mathbf{z} = \mathbf{z} - 2\frac{\mathbf{v}\mathbf{v}^*}{\mathbf{v}^*\mathbf{v}}\mathbf{z} = \mathbf{z} - 2\mathbf{v}\frac{\mathbf{v}^*\mathbf{z}}{\mathbf{v}^*\mathbf{v}} = \mathbf{z} - 2\mathbf{v}\frac{\langle \mathbf{v}, \mathbf{z} \rangle_2}{\langle \mathbf{v}, \mathbf{v} \rangle_2} \tag{3.24}$$

darstellen.

Lemma 3.66 (Spiegelung) *Sei $\mathbf{v} \in \mathbb{K}^n \setminus \{\mathbf{0}\}$. Dann ist die Householder-Spiegelung $\mathbf{Q}_v \in \mathbb{K}^{n\times n}$ eine unitäre selbstadjungierte Matrix.*

Beweis. Aus der Definition folgt mit Lemma 3.45 direkt, dass die Matrix $\mathbf{Q}_v$ selbstadjungiert ist. Mit (3.24) erhalten wir

$$\mathbf{Q}_v\mathbf{v} = \mathbf{v} - 2\mathbf{v}\frac{\langle \mathbf{v}, \mathbf{v} \rangle_2}{\langle \mathbf{v}, \mathbf{v} \rangle_2} = \mathbf{v} - 2\mathbf{v} = -\mathbf{v}. \tag{3.25}$$

Sei $\mathbf{z} \in \mathbb{K}^n$. Nach (3.24) und (3.25) gilt

$$\begin{aligned}\mathbf{Q}_v\mathbf{Q}_v\mathbf{z} &= \mathbf{Q}_v\mathbf{z} - 2\mathbf{v}\frac{\langle \mathbf{v}, \mathbf{Q}_v\mathbf{z} \rangle_2}{\langle \mathbf{v}, \mathbf{v} \rangle_2} = \mathbf{Q}_v\mathbf{z} - 2\mathbf{v}\frac{\langle \mathbf{Q}_v\mathbf{v}, \mathbf{z} \rangle_2}{\langle \mathbf{v}, \mathbf{v} \rangle_2} = \mathbf{Q}_v\mathbf{z} - 2\mathbf{v}\frac{\langle -\mathbf{v}, \mathbf{z} \rangle_2}{\langle \mathbf{v}, \mathbf{v} \rangle_2} \\ &= \mathbf{Q}_v\mathbf{z} + 2\mathbf{v}\frac{\langle \mathbf{v}, \mathbf{z} \rangle_2}{\langle \mathbf{v}, \mathbf{v} \rangle_2} = \mathbf{z} - 2\mathbf{v}\frac{\langle \mathbf{v}, \mathbf{z} \rangle_2}{\langle \mathbf{v}, \mathbf{v} \rangle_2} + 2\mathbf{v}\frac{\langle \mathbf{v}, \mathbf{z} \rangle_2}{\langle \mathbf{v}, \mathbf{v} \rangle_2} = \mathbf{z},\end{aligned}$$

also folgt $\mathbf{Q}_v^*\mathbf{Q}_v = \mathbf{Q}_v\mathbf{Q}_v = \mathbf{I}$, somit ist $\mathbf{Q}_v$ unitär. ∎

Für unsere Anwendung ist entscheidend, dass wir die Transformation so wählen können, dass sie in einer Spalte der Matrix $\mathbf{A}$ alle Einträge unterhalb der Diagonalen eliminiert. Mathematisch gesehen entspricht das einer Transformation, die den Spaltenvektor auf einen überwiegend aus Nullen bestehenden Vektor überführt.

Wir untersuchen den allgemeinen Fall: Sei $\mathbf{x} \in \mathbb{K}^n$ ein beliebiger Vektor, den wir mit Hilfe einer Householder-Spiegelung auf ein Vielfaches eines zweiten Vektors $\mathbf{y} \in \mathbb{K}^n \setminus \{\mathbf{0}\}$ abbilden wollen. Wir suchen also nach einem $\mathbf{v} \in \mathbb{K}^n \setminus \{\mathbf{0}\}$, das

$$\mathbf{Q}_v\mathbf{x} = \mathbf{x} - 2\mathbf{v}\frac{\langle \mathbf{v}, \mathbf{x}\rangle_2}{\langle \mathbf{v}, \mathbf{v}\rangle_2} = \alpha\mathbf{y}$$

für ein $\alpha \in \mathbb{K}$ erfüllt. Aus dieser Gleichung folgt sofort

$$2\mathbf{v}\frac{\langle \mathbf{v}, \mathbf{x}\rangle_2}{\langle \mathbf{v}, \mathbf{v}\rangle_2} = \mathbf{x} - \alpha\mathbf{y}.$$

Da die Länge des Householder-Vektors keinen Einfluss auf $\mathbf{Q}_v$ hat, können wir einfach den Ansatz $\mathbf{v} = \mathbf{x} - \alpha\mathbf{y}$ verwenden und müssen lediglich α so wählen, dass

$$2\frac{\langle \mathbf{v}, \mathbf{x}\rangle_2}{\langle \mathbf{v}, \mathbf{v}\rangle_2} = 1$$

gilt. Diese Gleichung entspricht

$$\begin{aligned} 2\|\mathbf{x}\|_2^2 - 2\bar{\alpha}\langle \mathbf{y}, \mathbf{x}\rangle_2 &= 2\langle \mathbf{x} - \alpha\mathbf{y}, \mathbf{x}\rangle_2 = 2\langle \mathbf{v}, \mathbf{x}\rangle_2 = \langle \mathbf{v}, \mathbf{v}\rangle_2 \\ &= \langle \mathbf{x} - \alpha\mathbf{y}, \mathbf{x} - \alpha\mathbf{y}\rangle_2 = \|\mathbf{x}\|_2^2 - \bar{\alpha}\langle \mathbf{y}, \mathbf{x}\rangle_2 - \alpha\langle \mathbf{x}, \mathbf{y}\rangle_2 + |\alpha|^2\|\mathbf{y}\|_2^2, \\ 0 &= |\alpha|^2\|\mathbf{y}\|_2^2 + \bar{\alpha}\langle \mathbf{y}, \mathbf{x}\rangle_2 - \alpha\overline{\langle \mathbf{y}, \mathbf{x}\rangle_2} - \|\mathbf{x}\|_2^2. \end{aligned}$$

Der Imaginärteil der rechten Seite kann nur verschwinden, falls $\bar{\alpha}\langle \mathbf{y}, \mathbf{x}\rangle_2$ reell ist, falls also α ein reelles Vielfaches des Vorzeichens von $\langle \mathbf{y}, \mathbf{x}\rangle_2$ ist. Der Realteil kann nur verschwinden, falls $|\alpha|^2 = \|\mathbf{x}\|_2^2/\|\mathbf{y}\|_2^2$ gilt. Indem wir das Vorzeichen so wählen, dass Auslöschungseffekte bei der Berechnung des Householder-Vektors im Fall $\mathbf{x} \approx -\mathbf{y}$ möglichst vermieden werden, erhalten wir das folgende Lemma.

Lemma 3.67 *Seien $\mathbf{x}, \mathbf{y} \in \mathbb{K}^n \setminus \{\mathbf{0}\}$ gegeben. Mit*

$$\mathbf{v} := \mathbf{x} - \alpha\mathbf{y}, \qquad \alpha := -\operatorname{sgn}(\langle \mathbf{y}, \mathbf{x}\rangle_2)\frac{\|\mathbf{x}\|_2}{\|\mathbf{y}\|_2}$$

gilt $\mathbf{v} \neq \mathbf{0}$ und wir erhalten

$$\mathbf{Q}_v\mathbf{x} = \alpha\mathbf{y},$$

die Householder-Spiegelung bildet also $\mathbf{x}$ auf ein Vielfaches von $\mathbf{y}$ ab.

Beweis. Wir setzen $z := \langle \mathbf{y}, \mathbf{x}\rangle_2$. Das Vorzeichen von z definieren wir durch

$$\operatorname{sgn}(z) := \begin{cases} z/|z| & \text{falls } z \neq 0, \\ 1 & \text{ansonsten,} \end{cases}$$

es erfüllt die Gleichungen

$$\overline{\operatorname{sgn}(z)}z = |z|, \qquad |\operatorname{sgn}(z)| = 1.$$

Daraus folgt

$$\begin{aligned}
\langle \mathbf{v}, \mathbf{x} \rangle_2 &= \langle \mathbf{x} - \alpha \mathbf{y}, \mathbf{x} \rangle_2 = \langle \mathbf{x}, \mathbf{x} \rangle_2 - \bar{\alpha} \langle \mathbf{y}, \mathbf{x} \rangle_2 \\
&= \langle \mathbf{x}, \mathbf{x} \rangle_2 + \frac{\|\mathbf{x}\|_2}{\|\mathbf{y}\|_2} \overline{\operatorname{sgn}(\langle \mathbf{y}, \mathbf{x} \rangle_2)} \langle \mathbf{y}, \mathbf{x} \rangle_2 = \|\mathbf{x}\|_2^2 + \frac{\|\mathbf{x}\|_2}{\|\mathbf{y}\|_2} |\langle \mathbf{y}, \mathbf{x} \rangle_2| > 0, \\
\langle \mathbf{v}, \mathbf{v} \rangle_2 &= \langle \mathbf{x} - \alpha \mathbf{y}, \mathbf{x} - \alpha \mathbf{y} \rangle_2 = \langle \mathbf{x}, \mathbf{x} \rangle_2 - \bar{\alpha} \langle \mathbf{y}, \mathbf{x} \rangle_2 - \alpha \langle \mathbf{x}, \mathbf{y} \rangle_2 + |\alpha|^2 \langle \mathbf{y}, \mathbf{y} \rangle_2 \\
&= \langle \mathbf{x}, \mathbf{x} \rangle_2 + \frac{\|\mathbf{x}\|_2}{\|\mathbf{y}\|_2} \overline{\operatorname{sgn}(\langle \mathbf{y}, \mathbf{x} \rangle_2)} \langle \mathbf{y}, \mathbf{x} \rangle_2 + \frac{\|\mathbf{x}\|_2}{\|\mathbf{y}\|_2} \operatorname{sgn}(\langle \mathbf{y}, \mathbf{x} \rangle_2) \overline{\langle \mathbf{y}, \mathbf{x} \rangle_2} + \frac{\|\mathbf{x}\|_2^2}{\|\mathbf{y}\|_2^2} \langle \mathbf{y}, \mathbf{y} \rangle_2 \\
&= \|\mathbf{x}\|_2^2 + \frac{\|\mathbf{x}\|_2}{\|\mathbf{y}\|_2} |\langle \mathbf{y}, \mathbf{x} \rangle_2| + \frac{\|\mathbf{x}\|_2}{\|\mathbf{y}\|_2} |\langle \mathbf{y}, \mathbf{x} \rangle_2| + \|\mathbf{x}\|_2^2 \\
&= 2 \left(\|\mathbf{x}\|_2^2 + \frac{\|\mathbf{x}\|_2}{\|\mathbf{y}\|_2} |\langle \mathbf{y}, \mathbf{x} \rangle_2| \right) = 2 \langle \mathbf{v}, \mathbf{x} \rangle_2 > 0
\end{aligned}$$

und damit $\mathbf{v} \neq \mathbf{0}$ sowie dank (3.24) auch die Gleichung

$$\mathbf{Q}_v \mathbf{x} = \mathbf{x} - 2\mathbf{v} \frac{\langle \mathbf{v}, \mathbf{x} \rangle_2}{\langle \mathbf{v}, \mathbf{v} \rangle_2} = \mathbf{x} - \mathbf{v} = \alpha \mathbf{y},$$

die zu beweisen war. ∎

Wir sind daran interessiert, möglichst viele Nulleinträge in der ersten Spalte der Matrix $\mathbf{A}$ zu erhalten. Dieses Ziel können wir nun mit Hilfe des Lemmas 3.67 erreichen: Wir wenden das Lemma an, indem wir für $\mathbf{x} \in \mathbb{K}^n$ die erste Spalte der Matrix $\mathbf{A}$ und für $\mathbf{y} \in \mathbb{K}^n$ den ersten kanonischen Einheitsvektor einsetzen und folgern, dass für

$$\mathbf{x} := \begin{pmatrix} a_{11} \\ a_{21} \\ \vdots \\ a_{n1} \end{pmatrix}, \qquad \mathbf{y} := \begin{pmatrix} 1 \\ 0 \\ \vdots \\ 0 \end{pmatrix}, \qquad \mathbf{v} := \begin{pmatrix} a_{11} + \|\mathbf{x}\|_2 \operatorname{sgn}(a_{11}) \\ a_{21} \\ \vdots \\ a_{n1} \end{pmatrix} \tag{3.26}$$

die Householder-Spiegelung $\mathbf{Q}_v$ gerade $\mathbf{x}$ auf ein Vielfaches des ersten kanonischen Einheitsvektors abbildet, also mit Ausnahme der ersten alle Zeilen der ersten Spalte von $\mathbf{A}$ eliminiert.

Durch wiederholte Anwendung dieser Technik können wir eine beliebige Matrix auf obere Dreiecksgestalt bringen.

Satz 3.68 (Existenz einer QR-Zerlegung) *Jede Matrix $\mathbf{A} \in \mathbb{K}^{m \times n}$ besitzt eine QR-Zerlegung.*

Beweis. Per Induktion über m.

Induktionsanfang: Für $m = 1$ setzen wir $q_{11} = 1$ und $\mathbf{R} = \mathbf{A}$ und sind fertig.

Induktionsvoraussetzung: Sei nun $m \in \mathbb{N}$ so gewählt, dass die Behauptung für alle Matrizen aus $\mathbb{K}^{m \times n}$ mit $n \in \mathbb{N}$ gilt.

Induktionsschritt: Sei $n \in \mathbb{N}$ und $\mathbf{A} \in \mathbb{K}^{(m+1) \times n}$. Falls die erste Spalte von $\mathbf{A}$ gleich null ist, können wir einen beliebigen Householder-Vektor $\mathbf{v} \in \mathbb{K}^{m+1} \setminus \{\mathbf{0}\}$ verwenden, um $\mathbf{Q}_v \mathbf{0} = \mathbf{0}$ zu erhalten.

Anderenfalls wenden wir Lemma 3.67 auf die Vektoren

$$\mathbf{x} := \begin{pmatrix} a_{11} \\ a_{21} \\ \vdots \\ a_{m+1,1} \end{pmatrix}, \qquad \mathbf{y} := \begin{pmatrix} 1 \\ 0 \\ \vdots \\ 0 \end{pmatrix}, \qquad \mathbf{v} = \begin{pmatrix} a_{11} + \operatorname{sgn}(a_{11})\|\mathbf{x}\|_2 \\ a_{21} \\ \vdots \\ a_{m+1,1} \end{pmatrix},$$

an, um eine unitäre Householder-Spiegelung $\mathbf{Q}_v \in \mathbb{K}^{(m+1)\times(m+1)}$ zu erhalten. Damit lässt sich das Produkt von $\mathbf{Q}_v$ mit $\mathbf{A}$ in der Form

$$\mathbf{B} := \mathbf{Q}_v\mathbf{A} = \mathbf{Q}_v \begin{pmatrix} a_{11} & a_{12} & \dots & a_{1n} \\ a_{21} & a_{22} & \dots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{m+1,1} & a_{m+1,2} & \dots & a_{m+1,n} \end{pmatrix} = \begin{pmatrix} \alpha & b_{12} & \dots & b_{1n} \\ 0 & b_{22} & \dots & b_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ 0 & b_{m+1,2} & \dots & b_{m+1,n} \end{pmatrix}$$

darstellen, wobei

$$\alpha = \begin{cases} -\operatorname{sgn}(a_{11})\|\mathbf{x}\|_2 & \text{falls } \mathbf{x} \neq \mathbf{0}, \\ 0 & \text{ansonsten} \end{cases}$$

gilt. Da $\mathbf{Q}_v^{-1} = \mathbf{Q}_v^* = \mathbf{Q}_v$ gilt, folgt aus dieser Gleichung $\mathbf{A} = \mathbf{Q}_v\mathbf{B}$. Falls $n = 1$ gelten sollte, ist $\mathbf{B}$ bereits eine obere Dreiecksmatrix und wir sind fertig.

Anderenfalls zerlegen wir $\mathbf{B}$ in der bekannten Weise in die Teilmatrizen

$$\mathbf{B}_{**} := \begin{pmatrix} b_{22} & \dots & b_{2n} \\ \vdots & \ddots & \vdots \\ b_{m+1,2} & \dots & b_{m+1,n} \end{pmatrix}, \qquad \mathbf{B}_{1*} := \begin{pmatrix} b_{12} & \dots & b_{1n} \end{pmatrix}$$

und erhalten

$$\mathbf{B} = \left(\begin{array}{c|ccc} \alpha & b_{12} & \dots & b_{1n} \\ \hline 0 & b_{22} & \dots & b_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ 0 & b_{m+1,2} & \dots & b_{m+1,n} \end{array}\right) = \begin{pmatrix} \alpha & \mathbf{B}_{1*} \\ \mathbf{0} & \mathbf{B}_{**} \end{pmatrix}$$

mit $\mathbf{B}_{**} \in \mathbb{K}^{m\times(n-1)}$. Wir wenden die Induktionsvoraussetzung auf $\mathbf{B}_{**}$ an und finden eine unitäre Matrix $\mathbf{Q}_{**} \in \mathbb{K}^{m\times m}$ sowie eine obere Dreiecksmatrix $\mathbf{R}_{**} \in \mathbb{K}^{m\times(n-1)}$ mit $\mathbf{B}_{**} = \mathbf{Q}_{**}\mathbf{R}_{**}$. Mit

$$\mathbf{Q} := \mathbf{Q}_v \begin{pmatrix} 1 & \mathbf{0} \\ \mathbf{0} & \mathbf{Q}_{**} \end{pmatrix}, \qquad \mathbf{R} := \begin{pmatrix} \alpha & \mathbf{B}_{1*} \\ \mathbf{0} & \mathbf{R}_{**} \end{pmatrix}$$

folgt aus $\mathbf{Q}_v = \mathbf{Q}_v^*$ und $\mathbf{A} = \mathbf{Q}_v^*\mathbf{B} = \mathbf{Q}_v\mathbf{B}$ bereits

$$\mathbf{A} = \mathbf{Q}_v\mathbf{B} = \mathbf{Q}_v \begin{pmatrix} \alpha & \mathbf{B}_{1*} \\ \mathbf{0} & \mathbf{B}_{**} \end{pmatrix} = \mathbf{Q}_v \begin{pmatrix} \alpha & \mathbf{B}_{1*} \\ \mathbf{0} & \mathbf{Q}_{**}\mathbf{R}_{**} \end{pmatrix} = \mathbf{Q}_v \begin{pmatrix} 1 & \mathbf{0} \\ \mathbf{0} & \mathbf{Q}_{**} \end{pmatrix} \begin{pmatrix} \alpha & \mathbf{B}_{1*} \\ \mathbf{0} & \mathbf{R}_{**} \end{pmatrix} = \mathbf{Q}\mathbf{R},$$

und mit Lemma 3.61 folgt, dass $\mathbf{Q}$ als Produkt unitärer Matrizen selbst unitär ist. ∎

Der Beweis von Satz 3.68 ist konstruktiv, so dass wir ihn direkt in einen Algorithmus umsetzen können.

Auch in diesem Fall müssen wir natürlich wieder untersuchen, wie aufwendig die Berechnung der QR-Zerlegung ist.

Lemma 3.69 (Rechenaufwand) *Sei $\mathbf{A} \in \mathbb{K}^{m\times n}$, sei $k := \min\{n, m\}$.*

Die QR-Zerlegung kann in

$$\left(4mn + 2n + \frac{17}{3}\right) k - (2m + 2n + 1)k^2 + \frac{4}{3}k^3 \tag{3.27a}$$

Operationen berechnet werden. Falls $k = n$ gilt, sind es

$$2mn^2 - \frac{2}{3}n^3 + n^2 + \frac{17}{3}n \tag{3.27b}$$

Operationen, und falls $k = m = n$ gilt, sind es

$$\frac{4}{3}n^3 + n^2 + \frac{17}{3}n. \tag{3.27c}$$

Beweis. Für die i-te Householder-Spiegelung muss die Norm einer Spalte mit $m - i + 1$ Elementen berechnet werden, so dass $m - i + 1$ Multiplikationen, $m - i$ Additionen und eine Wurzel anfallen, insgesamt also $2(m - i + 1)$ Operationen. Indem wir die Norm zu dem ersten Element der Spalte addieren, erhalten wir den Householder-Vektor, der damit nach $2(m-i+1)+1$ Operationen zur Verfügung steht. Dank Lemma 3.67 können wir mit einer Multiplikation und einer Vorzeichenberechnung auch die Transformation der i-ten Spalte durchführen. Es bietet sich an, den Faktor

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

nur einmal zu berechnen, da sich dann die Householder-Spiegelung kurz als

$$\mathbf{Q}_v\mathbf{z} = \mathbf{z} - \tau\langle\mathbf{v}, \mathbf{z}\rangle_2\mathbf{v} \tag{3.28}$$

schreiben lässt. Da uns $\|\mathbf{x}\|_2^2$ und $\|\mathbf{x}\|_2$ bereits zur Verfügung stehen, sind dafür nur drei zusätzliche Operationen erforderlich. Insgesamt benötigen wir also $2(m - i + 1) + 6 = 2(m-i+4)$ Operationen, um die Householder-Spiegelung zu berechnen und auf die i-te Spalte anzuwenden.

Für die verbliebenen $n-i$ Spalten sind gemäß der Gleichung (3.28) ein Skalarprodukt und eine Linearkombination zu berechnen. Ersteres erfordert $2(m-i+1)-1$ Operationen, letzteres $2(m - i + 1)$, und die Multiplikation mit dem Faktor τ eine weitere, so dass insgesamt $4(m-i+1)$ Operationen für jede Spalte anfallen, also $4(m-i+1)(n-i)$ für alle Spalten.

Wir gelangen zu dem Ergebnis, dass die Berechnung der QR-Zerlegung

$$\sum_{i=1}^{k} 2(m - i + 4) + 4(m - i + 1)(n - i)$$

$$
\begin{aligned}
&= \sum_{i=1}^{k} 2m - 2i + 8 + 4mn - 4in + 4n - 4mi + 4i^2 - 4i \\
&= \sum_{i=1}^{k} (4mn + 2m + 4n + 8) - (4n + 4m + 6)i + 4i^2 \\
&= (4mn + 2m + 4n + 8)k - (4n + 4m + 6)\frac{k(k+1)}{2} + 4\frac{k}{6}(2k+1)(k+1) \\
&= (4mn + 2m + 4n + 8)k - (2n + 2m + 3)k - (2n + 2m + 3)k^2 + \frac{2}{3}(2k^3 + 3k^2 + k) \\
&= \left(4mn + 2n + \frac{17}{3}\right)k - (2n + 2m + 1)k^2 + \frac{4}{3}k^3
\end{aligned}
$$

Operationen benötigt. Das ist (3.27a).

Für $k = n$ vereinfacht sich dieser Ausdruck zu

$$4mn^2 + 2n^2 + \frac{17}{3}n - 2n^3 - 2mn^2 - n^2 + \frac{4}{3}n^3 = 2mn^2 - \frac{2}{3}n^3 + n^2 + \frac{17}{3}n,$$

so dass wir (3.27b) erhalten.

Für $k = m = n$ können wir weiter vereinfachen, um

$$2n^3 - \frac{2}{3}n^3 + n^2 + \frac{17}{3}n = \frac{4}{3}n^3 + n^2 + \frac{17}{3}n$$

zu erhalten, also (3.27c). ∎

Für große Matrizen ist nur der kubische Term von Bedeutung, so dass wir festhalten können, dass die QR-Zerlegung ungefähr den doppelten Rechenaufwand der LR-Zerlegung (vgl. (3.7)) erfordert.

Falls $m \geq n$ gilt, falls also mehr Zeilen als Spalten auftreten, tritt die Anzahl der Zeilen m lediglich als *linearer* Faktor in der Aufwandsabschätzung auf, so dass sich die QR-Zerlegung einer Matrix mit sehr vielen Zeilen unter Umständen immer noch effizient berechnen lässt, sofern die Anzahl der Spalten nicht zu groß ist.

In professionellen Implementierungen wird der Householder-Algorithmus häufig etwas umformuliert, um die gesamte Zerlegung in dem für die ursprüngliche Matrix verwendeten Speicherbereich unterbringen zu können. Ein Blick auf die definierende Gleichung (3.23) einer Householder-Spiegelung zu einem Vektor $\mathbf{v}$ zeigt, dass wir den Vektor für beliebige $\zeta \in \mathbb{K} \setminus \{0\}$ durch eine skalierte Variante $\zeta\mathbf{v}$ ersetzen können, ohne die Spiegelung zu ändern: Der Faktor kürzt sich aus Zähler und Nenner heraus, so dass wir

$$\mathbf{Q}_{\zeta v} = \mathbf{I} - 2\frac{\zeta\mathbf{v}\bar{\zeta}\mathbf{v}^*}{\bar{\zeta}\mathbf{v}^*\zeta\mathbf{v}} = \mathbf{I} - 2\frac{\mathbf{v}\mathbf{v}^*}{\mathbf{v}^*\mathbf{v}} = \mathbf{Q}_v$$

erhalten. Wir wählen den Parameter ζ so, dass der Vektor $\zeta\mathbf{v}$ besonders einfach wird: Nach (3.26) wird sein erster Eintrag v_1 nur dann gleich null sein, wenn $\mathbf{x}$ gleich null ist, und in diesem Fall benötigen wir keine Householder-Spiegelung. Anderenfalls können wir

$\mathbf{v}$ so skalieren, dass der erste Eintrag gleich eins ist, indem wir $\zeta = 1/v_1$ verwenden und die Spiegelung mit dem Vektor

$$\widehat{\mathbf{v}} = \begin{pmatrix} 1 \\ v_2/v_1 \\ \vdots \\ v_n/v_1 \end{pmatrix}$$

statt mit $\mathbf{v}$ durchführen. Die erste Komponente dieses Vektors brauchen wir nicht abzuspeichern, da sie immer gleich eins ist.

Um die Householder-Spiegelung $\mathbf{Q}_{\hat{v}} = \mathbf{Q}_v$ effizient auf Vektoren $\mathbf{z}$ anwenden zu können, bringen wir (3.24) in die Form

$$\mathbf{Q}_{\hat{v}}\mathbf{z} = \mathbf{z} - 2\frac{\langle \widehat{\mathbf{v}}, \mathbf{z}\rangle_2}{\|\widehat{\mathbf{v}}\|_2^2}\widehat{\mathbf{v}} = \mathbf{z} - \widehat{\tau}\langle \widehat{\mathbf{v}}, \mathbf{z}\rangle_2 \widehat{\mathbf{v}}$$

mit dem Faktor

$$\widehat{\tau} = \frac{2}{\|\widehat{\mathbf{v}}\|_2^2} = \frac{2|v_1|^2}{\|\mathbf{v}\|_2^2}.$$

Mit den Vektoren aus (3.26) erhalten wir

$$\begin{aligned} v_1 &= a_{11} + \|\mathbf{x}\|_2 \operatorname{sgn}(a_{11}), \\ |v_1|^2 &= |a_{11}|^2 + 2\|\mathbf{x}\|_2|a_{11}| + \|\mathbf{x}\|_2^2 = (\|\mathbf{x}\|_2 + |a_{11}|)^2, \\ \|\mathbf{v}\|_2^2 &= |v_1|^2 + |a_{21}|^2 + \ldots + |a_{n1}|^2 \\ &= |a_{11}|^2 + 2\|\mathbf{x}\|_2|a_{11}| + \|\mathbf{x}\|_2^2 + |a_{21}|^2 + \ldots + |a_{n1}|^2 = 2\|\mathbf{x}\|_2(\|\mathbf{x}\|_2 + |a_{11}|) \end{aligned}$$

und damit

$$\widehat{\tau} = \frac{2|v_1|^2}{\|\mathbf{v}\|_2^2} = \frac{2(\|\mathbf{x}\|_2 + |a_{11}|)^2}{2\|\mathbf{x}\|_2(\|\mathbf{x}\|_2 + |a_{11}|)} = \frac{\|\mathbf{x}\|_2 + |a_{11}|}{\|\mathbf{x}\|_2}$$

Sollte $\mathbf{x} = \mathbf{0}$ gelten, setzen wir $\widehat{\tau} = 0$ und können den Vektor $\widehat{\mathbf{v}}$ völlig beliebig wählen oder auf die Anwendung der Householder-Spiegelung ganz verzichten.

Im Zuge des Gesamtalgorithmus müssen wir auf jede Spalte der Matrix eine Householder-Spiegelung anwenden. Den im i-ten Schritt verwendeten (entsprechend den obigen Überlegungen skalierten) Vektor bezeichnen wir mit $\widehat{\mathbf{v}}^{(i)}$. Ähnlich wie bei der LR- und der Cholesky-Zerlegung können wir alle Komponenten dieses Vektors mit Ausnahme der ersten in den Matrixeinträgen speichern, die durch die Householder-Spiegelung zu null geworden sind. Die erste Komponente brauchen wir nicht zu speichern, da sie dank unserer Skalierung immer gleich eins ist.

Unser Algorithmus kann also $\mathbf{A}$ mit der Matrix

$$\begin{pmatrix} r_{11} & \dots & \dots & r_{1n} \\ \hat{v}_2^{(1)} & r_{22} & & \vdots \\ \vdots & \hat{v}_2^{(2)} & \ddots & \vdots \\ \vdots & \vdots & \ddots & r_{nn} \\ \vdots & \vdots & \vdots & \hat{v}_2^{(n)} \\ \vdots & \vdots & \vdots & \vdots \\ \hat{v}_m^{(1)} & \hat{v}_{m-1}^{(2)} & \dots & \hat{v}_{m-n+1}^{(n)} \end{pmatrix} \tag{3.29}$$

überschreiben, im Interesse der Effizienz können wir die zugehörigen Skalierungsfaktoren $\widehat{\tau}^{(i)}$ in einem zusätzlichen Vektor $\mathbf{t} \in \mathbb{K}^n$ aufbewahren.

Bei quadratischen Matrizen können wir die so gewonnene QR-Zerlegung verwenden, um das Gleichungssystem $\mathbf{Ax} = \mathbf{b}$ nach $\mathbf{x}$ aufzulösen. Wenn die Matrizen $\mathbf{Q}$ und $\mathbf{R}$ in der in (3.29) beschriebenen Darstellung gegeben sind, müssen wir zunächst mit dem in Abbildung 3.11 angegebenen Algorithmus den Vektor $\mathbf{y} = \mathbf{Q}^*\mathbf{b}$ bestimmen, um anschließend die Lösung $\mathbf{x}$ aus $\mathbf{Rx} = \mathbf{y}$ durch Rückwärtseinsetzen zu gewinnen.

```
procedure trans_qr(n, m, A, t, var b);
for i = 1, ..., min{n, m} do
  if t_i ≠ 0 then
    γ ← b_i;   τ ← t_i
    for k ∈ [i + 1 : m] do
      γ ← γ + ā_ki b_k
    end for
    δ ← τγ
    b_i ← b_i − δ
    for k ∈ [i + 1 : m] do
      b_k ← b_k − δ a_ki
    end for
  end if
end for
```

Abbildung 3.11: Berechnung von $\mathbf{y} = \mathbf{Q}^*\mathbf{b}$ für eine QR-Zerlegung auf Householder-Basis. Dabei wird $\mathbf{b}$ mit dem Vektor $\mathbf{y}$ überschrieben.

Die Multiplikation mit $\mathbf{Q}^*$ lässt sich effizient gestalten, indem wir sie als Folge der einzelnen Householder-Projektionen $\mathbf{Q}_{\hat{v}^{(1)}}, \dots, \mathbf{Q}_{\hat{v}^{(n)}}$ darstellen, die wir bei vorberechnetem $\tau^{(i)}$ sehr einfach durchführen können. Der resultierende Algorithmus ist in Abbildung 3.11 angegeben. Mit $k := \min\{n, m\}$ stellen wir fest, dass die Berechnung von $\mathbf{y}$ gerade

$$\sum_{i=1}^{k} 4(m-i) + 2 = 4mk - 2k(k+1) + 2k = 2k(2m-k)$$

Operationen erfordert. Im für uns interessanten Fall $m = n = k$ sind das gerade $2n^2$ Operationen. Das Rückwärtseinsetzen erfordert weitere n^2 Operationen, so dass insgesamt $3n^2$ Operationen benötigt werden. Bei der Auswertung sind also für ein Verfahren auf Grundlage der QR-Zerlegung anderthalb mal so viele Operationen wie bei der LR-Zerlegung erforderlich.

Bemerkung 3.70 (Komplexe Householder-Spiegelung) *Für komplexwertige Aufgabenstellungen, also im Fall $\mathbb{K} = \mathbb{C}$, ist häufig eine Variante der Householder-Spiegelung von Interesse, die einen Vektor $\mathbf{x} \in \mathbb{C}^n \setminus \{\mathbf{0}\}$ auf ein* reelles *Vielfaches des ersten kanonischen Einheitsvektors $\delta = (1, 0, \ldots, 0) \in \mathbb{C}^n$ abbildet. Dazu suchen wir einen Vektor $\mathbf{v} \in \mathbb{C}^n$ und eine Zahl $\tau \in \mathbb{C}$ derart, dass*

$$\mathbf{H}_v := \mathbf{I} - \tau \mathbf{v}\mathbf{v}^*$$

unitär ist und $\mathbf{H}_v\mathbf{x} = -\alpha\delta$ mit $\alpha \in \mathbb{R}$ erfüllt.

Mit dem Ansatz $\mathbf{v} = \mathbf{x} + \alpha\delta$ und einem Blick auf die letzten Komponenten des Vektors

$$-\alpha\delta \stackrel{!}{=} \mathbf{H}_v\mathbf{x} = \mathbf{x} - \tau\mathbf{v}\mathbf{v}^*\mathbf{x} = \mathbf{x} - \tau(\mathbf{x} + \alpha\delta)(\mathbf{x} + \alpha\delta)^*\mathbf{x} = \mathbf{x} - \tau(\mathbf{x} + \alpha\delta)(\|\mathbf{x}\|_2^2 + \alpha x_1)$$

erhalten wir $\tau = \frac{1}{\|x\|_2^2 + \alpha x_1}$. Damit $\mathbf{H}_v$ unitär sein kann, muss nach Lemma 3.61 $|\alpha| = \|\alpha\delta\|_2 = \|\mathbf{H}_v\mathbf{x}\|_2 = \|\mathbf{x}\|_2$ gelten, also $\alpha = \pm\|\mathbf{x}\|_2$.

Wir zerlegen $x_1 = a + ib$ mit $a, b \in \mathbb{R}$ und vermeiden Auslöschung, indem wir $\alpha = \operatorname{sgn}(a)\|\mathbf{x}\|_2$ wählen. Die Norm des Householder-Vektors beträgt dann

$$\|\mathbf{v}\|_2^2 = (\mathbf{x} + \alpha\delta)^*(\mathbf{x} + \alpha\delta) = \|\mathbf{x}\|_2^2 + \alpha\bar{x}_1 + \alpha x_1 + \alpha^2 = 2(\|\mathbf{x}\|_2^2 + \alpha a).$$

Damit $\mathbf{H}_v$ unitär ist, muss

$$\begin{aligned}\mathbf{I} = \mathbf{H}_v^*\mathbf{H}_v &= (\mathbf{I} - \tau\mathbf{v}\mathbf{v}^*)^*(\mathbf{I} - \tau\mathbf{v}\mathbf{v}^*) = \mathbf{I} - \bar{\tau}\mathbf{v}\mathbf{v}^* - \tau\mathbf{v}\mathbf{v}^* + |\tau|^2\mathbf{v}\mathbf{v}^*\mathbf{v}\mathbf{v}^* \\ &= \mathbf{I} - \left(\bar{\tau} + \tau - |\tau|^2\|\mathbf{v}\|_2^2\right)\mathbf{v}\mathbf{v}^* = \mathbf{I} - \left(2\operatorname{Re}(\tau) - |\tau|^2\|\mathbf{v}\|_2^2\right)\mathbf{v}\mathbf{v}^*\end{aligned}$$

gelten. Wir haben

$$\tau = \frac{1}{\|\mathbf{x}\|_2^2 + \alpha x_1} = \frac{1}{\|\mathbf{x}\|_2^2 + \alpha(a + ib)} = \frac{2}{\|\mathbf{v}\|_2^2 + 2\alpha ib} = \frac{2(\|\mathbf{v}\|_2^2 - 2\alpha ib)}{\|\mathbf{v}\|_2^4 + 4\alpha^2 b^2}$$

und erhalten

$$\begin{aligned}2\operatorname{Re}(\tau) - |\tau|^2\|\mathbf{v}\|_2^2 &= \frac{4\|\mathbf{v}\|_2^2}{\|\mathbf{v}\|_2^4 + 4\alpha^2 b^2} - \frac{4\|\|\mathbf{v}\|_2^4 + 16\alpha^2 b^2}{(\|\mathbf{v}\|_2^4 + 4\alpha^2 b^2)^2}\|\mathbf{v}\|_2^2 \\ &= \frac{4\|\mathbf{v}\|_2^2(\|\mathbf{v}\|_2^4 + 4\alpha^2 b^2) - 4\|\mathbf{v}\|_2^6 - 16\alpha^2 b^2\|\mathbf{v}\|_2^2}{(\|\mathbf{v}\|_2^4 + 4\alpha^2 b^2)^2} = 0.\end{aligned}$$

Also ist $\mathbf{H}_v$ unitär und leistet das Gewünschte. Der Einsatz dieser modifizierten Householder-Spiegelung kann beispielsweise sinnvoll sein, weil sich mit ihr komplexe Divisionen während des Rückwärtseinsetzens vermeiden lassen.

Da τ für $b \neq 0$ nicht reell ist, anders als bei der konventionellen Householder-Spiegelung, ist $\mathbf{H}_v$ nicht selbstadjungiert, so dass bei der Implementierung sorgfältig darauf geachtet werden muss, ob mit $\mathbf{H}_v = \mathbf{I} - \tau\mathbf{v}\mathbf{v}^*$ oder mit $\mathbf{H}_v^* = \mathbf{I} - \bar{\tau}\mathbf{v}\mathbf{v}^*$ multipliziert werden muss.

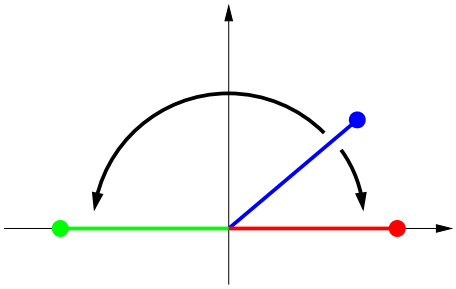

Abbildung 3.12: Geometrische Interpretation der Givens-Rotation

Bemerkung 3.71 (Givens-Rotation) *Die Householder-Spiegelung ist keineswegs die einzige unitäre Transformation, mit der wir Nulleinträge in einer Matrix erzeugen können. Eine Alternative sind* Givens-Rotationen*: Im zweidimensionalen Fall besitzen sie die Form*

$$\mathbf{G} = \begin{pmatrix} \bar{c} & \bar{s} \\ -s & c \end{pmatrix}$$

für $c, s \in \mathbb{K}$ mit $|c|^2 + |s|^2 = 1$. Für einen beliebigen Vektor $\mathbf{x} \in \mathbb{K}^2$ erhalten wir

$$\mathbf{G}\mathbf{x} = \begin{pmatrix} \bar{c} & \bar{s} \\ -s & c \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \begin{pmatrix} \bar{c}x_1 + \bar{s}x_2 \\ -sx_1 + cx_2 \end{pmatrix},$$

und im Falle $\mathbf{x} \neq \mathbf{0}$ können wir

$$s := \frac{x_2}{\sqrt{|x_1|^2 + |x_2|^2}}, \qquad c := \frac{x_1}{\sqrt{|x_1|^2 + |x_2|^2}}$$

setzen und finden

$$\mathbf{G}\mathbf{x} = \begin{pmatrix} \sqrt{|x_1|^2 + |x_2|^2} \\ 0 \end{pmatrix},$$

haben also einen Eintrag des Vektors eliminiert. Bei längeren Vektoren können wir einen Eintrag nach dem anderen eliminieren und so, wie bei der Householder-Spiegelung, ein Vielfaches des ersten kanonischen Einheitsvektors erhalten.

Bemerkung 3.72 (Effiziente Speicherung) *Mit Hilfe einer einfachen Modifikation können wir eine Givens-Rotation durch eine einzige Zahl beschreiben: Wir können c und s als Cosinus und Sinus des Rotationswinkels interpretieren. Falls wir $c \neq 0$ voraussetzen, ist der Tangens des Winkels dann durch*

$$t = \frac{s}{c} = \frac{x_2}{x_1}$$

gegeben und kann verwendet werden, um Cosinus und Sinus zu rekonstruieren, allerdings unter Verzicht auf das Vorzeichen: Aus $s = tc$ und $|c|^2 + |s|^2 = 1$ folgt

$$1 = |c|^2 + |s|^2 = |c|^2 + |t|^2|c|^2 = (1 + |t|^2)|c|^2,$$

und wir können

$$c = \frac{1}{\sqrt{1+|t|^2}}, \qquad s = tc$$

verwenden. Es folgt

$$\begin{pmatrix} \bar{c} & \bar{s} \\ -s & c \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \begin{pmatrix} \bar{c}x_1 + \bar{s}x_2 \\ -sx_1 + cx_2 \end{pmatrix} = \begin{pmatrix} \bar{c}(x_1 + |x_2|^2/\bar{x}_1) \\ c(-x_2 + x_2) \end{pmatrix} = \begin{pmatrix} \bar{c}(|x_1|^2 + |x_2|^2)/\bar{x}_1 \\ 0 \end{pmatrix},$$

wir erreichen also die gewünschte Null in der zweiten Zeile, allerdings nicht mehr einen reellen Eintrag in der ersten.

Übungsaufgabe 3.73 (Unitäre Matrizen) *Zeigen Sie, dass für eine selbstadjungierte Matrix* $\mathbf{A} \in \mathbb{K}^{n\times n}$ *genau dann*

$$\langle \mathbf{x}, \mathbf{A}\mathbf{x}\rangle_2 = 0 \qquad \text{für alle } \mathbf{x} \in \mathbb{K}^n$$

gilt, wenn $\mathbf{A}$ *die Nullmatrix ist.*

Hinweis: *Man könnte* $\mathbf{x} = \mathbf{y} + \mathbf{A}\mathbf{y}$ *einsetzen.*

Folgern Sie daraus, dass für eine Matrix $\mathbf{Q} \in \mathbb{K}^{n\times n}$

$$\|\mathbf{Q}\mathbf{x}\|_2 = \|\mathbf{x}\|_2 \qquad \text{für alle } \mathbf{x} \in \mathbb{K}^n$$

genau dann gilt, wenn $\mathbf{Q}$ *unitär ist.*

Hinweis: *Lemma 3.44.*

Übungsaufgabe 3.74 (Determinante) *Um mit der QR-Zerlegung auch Determinanten berechnen zu können, müssen wir etwas über die Determinanten der Householder-Spiegelungen wissen.*

(a) Seien $\mathbf{a}, \mathbf{b} \in \mathbb{K}^n$. *Beweisen Sie* $\det(\mathbf{I} + \mathbf{a}\mathbf{b}^*) = 1 + \mathbf{b}^*\mathbf{a}$.

(b) Sei $\mathbf{v} \in \mathbb{K}^n \setminus \{\mathbf{0}\}$. *Sei* $\mathbf{Q}_v$ *die zugehörige Householder-Spiegelung. Beweisen Sie* $\det(\mathbf{Q}_v) = -1$.

(c) Seien $\mathbf{C} \in \mathbb{K}^{n\times n}$ *invertierbar. Seien* $\mathbf{a}, \mathbf{b} \in \mathbb{K}^n$. *Beweisen Sie* $\det(\mathbf{C} + \mathbf{a}\mathbf{b}^*) = 1 + \mathbf{b}^*\mathbf{C}^{-1}\mathbf{a}$.

Hinweis: *Bei Teil (a) hilft eine unitäre Matrix* $\mathbf{Q}$, *die* $\mathbf{a}$ *auf ein Vielfaches des ersten kanonischen Einheitsvektors abbildet. Mit dem Determinanten-Multiplikationssatz und Lemma 3.61 kann dann die Matrix* $\mathbf{Q}(\mathbf{I} + \mathbf{a}\mathbf{b}^*)\mathbf{Q}^*$ *näher untersucht werden.*

Übungsaufgabe 3.75 (Gram-Schmidt-Orthonormalisierung) *Sei* $\mathbf{A} \in \mathbb{K}^{n\times n}$ *regulär. Beweisen Sie, dass* $\mathbf{A}^*\mathbf{A}$ *eine Cholesky-Zerlegung* $\mathbf{A}^*\mathbf{A} = \mathbf{L}\mathbf{L}^*$ *besitzt.*

Zeigen Sie, dass $\mathbf{Q} := \mathbf{A}(\mathbf{L}^{-1})^*$ *unitär ist und dass* $\mathbf{A} = \mathbf{Q}\mathbf{L}^*$ *gilt.*

Hinweis: *Lemma 3.45.*

4 Lineare Ausgleichsprobleme

Bisher haben wir lineare Gleichungssysteme untersucht, die lösbar waren. Nun wenden wir uns solchen Problemen zu, für die potentiell keine exakte Lösung mehr existiert: Wir wählen $n, m \in \mathbb{N}$ und suchen zu einer Matrix $\mathbf{A} \in \mathbb{K}^{m \times n}$ und einem Vektor $\mathbf{b} \in \mathbb{K}^m$ einen Vektor $\mathbf{x} \in \mathbb{K}^n$ mit

$$\begin{aligned} a_{11}x_1 + a_{12}x_2 + \ldots + a_{1n}x_n &= b_1, \\ &\vdots \\ a_{m1}x_1 + a_{m2}x_2 + \ldots + a_{mn}x_n &= b_m. \end{aligned}$$

Falls $\mathbf{b}$ nicht im Bild von $\mathbf{A}$ liegt, bezeichnet man das System als *überbestimmt*, in diesem Fall existiert offenbar keine Lösung. Falls der Kern von $\mathbf{A}$ nicht nur den Nullvektor enthält, bezeichnet man das System als *unterbestimmt*, in diesem Fall kann eine Lösung existieren, ist aber nicht eindeutig. Ein System kann gleichzeitig über- und unterbestimmt sein, beispielsweise für $\mathbf{A} = \mathbf{0}$ und $\mathbf{b} \neq \mathbf{0}$.

Bei einem überbestimmten System können wir das Gleichungssystem abschwächen, indem wir nicht $\mathbf{Ax} = \mathbf{b}$ verlangen, sondern nur nach einem $\mathbf{x}$ suchen, das den *Defekt* $\mathbf{Ax} - \mathbf{b}$ in einem geeigneten Sinne minimiert. In der Regel misst man die Größe des Defekts in einer Norm $\|\cdot\|$ und erhält die folgende Aufgabe:

> Gegeben seien eine Matrix $\mathbf{A} \in \mathbb{K}^{m \times n}$ und ein Vektor $\mathbf{b} \in \mathbb{K}^m$. Finde einen Vektor $\mathbf{x} \in \mathbb{K}^n$ so, dass $\|\mathbf{Ax} - \mathbf{b}\|$ minimal ist.

Um eindeutige Lösbarkeit auch bei unterbestimmten Systemen sicher zu stellen, kann zusätzlich gefordert werden, dass unter allen Lösungen der Aufgabe der Vektor $\mathbf{x}$ mit minimaler Norm gewählt werden soll.

Man bezeichnet derartige Aufgaben als *lineare Ausgleichsprobleme*.

4.1 Motivation: Lineare Regression

Ein wichtiges Anwendungsgebiet linearer Ausgleichsprobleme ist die *lineare Regression*, bei der man versucht, eine wissenschaftliche Hypothese durch Messungen zu bestätigen: Man geht davon aus, dass zwischen gemessenen Daten ein Zusammenhang besteht, der durch eine von einigen Parametern abhängige Formel beschrieben wird. Die Aufgabe besteht darin, diese Parameter so zu wählen, dass die mit der Formel vorhergesagten Resultate möglichst gut zu den Messwerten passen.

Mathematisch gesehen ist die Formel eine Abbildung, die die Abhängigkeit bestimmter Ausgabegrößen von bestimmten Eingabegrößen mit Hilfe möglichst weniger Parameter beschreibt.

Im allgemeinen Fall sind Punkte $\xi_1, \ldots, \xi_m \in D$ in einer Menge D von Eingabegrößen sowie zugehörige Ausgabegrößen $b_1, \ldots, b_m \in \mathbb{K}$ gegeben. Für Funktionen $g_1, \ldots, g_n : D \to \mathbb{K}$ suchen wir nun Parameter $x_1, \ldots, x_n \in \mathbb{K}$ so, dass

$$g_x : D \to \mathbb{K}, \qquad \xi \mapsto x_1 g_1(\xi) + \ldots + x_n g_n(\xi)$$

in den Punkten $\xi_1, \ldots, \xi_m$ möglichst nahe an $b_1, \ldots, b_m$ liegt. Damit beschreibe g_x die vermutete Gesetzmäßigkeit, die den Zusammenhang zwischen Eingabegrößen und Ausgabegrößen beschreibt.

Unser Ziel ist es, die Abweichungen zwischen der Vorhersage $g_x(\xi_i)$ und den Messwerten b_i

$$|g_x(\xi_i) - b_i|$$

für alle $i \in [1 : m]$ möglichst klein zu halten. Mehrere Fehler gleichzeitig zu minimieren, ist in der Regel nicht möglich, wir müssen stattdessen die Fehler in geeigneter Weise zu einem Gesamtfehler zusammenfassen. Eine für die Theorie sehr hilfreiche Wahl besteht darin, die *Summe der Fehlerquadrate* zu minimieren, also die Größe

$$f(x_1, \ldots, x_n) := \sum_{i=1}^{m} |g_x(\xi_i) - \mu_i|^2.$$

Falls diese Summe klein ist, muss auch jeder Einzelfehler klein sein, insofern sollten die Parameter $x_1, \ldots, x_n$ zu einer guten Approximation der Messwerte führen.

Da g_x von den Parametern $x_1, \ldots, x_n$ linear abhängt, können wir eine Matrix $\mathbf{A} \in \mathbb{K}^{m \times n}$ durch

$$a_{ij} := g_j(\xi_i) \qquad \text{für alle } i \in [1 : m],\ j \in [1 : n]$$

definieren und erhalten

$$g_x(\xi_i) = x_1 g_1(\xi_i) + \ldots + x_n g_n(\xi_i) = \sum_{j=1}^{n} a_{ij} x_j = (\mathbf{A}\mathbf{x})_i \qquad \text{für alle } i \in [1 : m],$$

$$f(x_1, \ldots, x_n) = \sum_{i=1}^{m} |g_x(\xi_i) - b_i|^2 = \sum_{i=1}^{m} |(\mathbf{A}\mathbf{x})_i - \mathbf{b}_i|^2 = \|\mathbf{A}\mathbf{x} - \mathbf{b}\|_2^2.$$

Die optimale Parameterkombination ist also durch die Lösung eines linearen Ausgleichsproblems beschrieben.

Beispiel 4.1 (Ohmsches Gesetz) *Der Zusammenhang zwischen dem Strom I und der Spannung U in einem elektrischen Leiter wird durch das Ohmsche Gesetz $U = IR$ beschrieben, wobei R den Widerstand angibt.*

Wenn wir für verschiedene Spannungen $U_1, \ldots, U_m$ die fließenden Ströme $I_1, \ldots, I_m$ gemessen haben, können wir versuchen, den Widerstand R zu berechnen, indem wir die Summe der Fehlerquadrate

$$f(R) := \sum_{i=1}^{m} |U_i - I_i R|^2$$

minimieren. Wir haben also ein lineares Regressionsproblem mit dem Parameter R und den Messwerten $U_1, \ldots, U_m$ und $I_1, \ldots, I_m$ zu lösen.

4.2 Normalengleichung

Der Schwierigkeitsgrad eines linearen Ausgleichsproblem hängt wesentlich von der verwendeten Norm ab. Als besonders günstig erweisen sich hier Hilbertraum-Normen, also solche Normen, die sich aus einem Skalarprodukt $\langle \cdot, \cdot \rangle$ gemäß der Form

$$\|\mathbf{x}\| = \sqrt{\langle \mathbf{x}, \mathbf{x} \rangle} \qquad \text{für alle } \mathbf{x} \in \mathbb{K}^n$$

ergeben. Das naheliegende Skalarprodukt ist im Falle des Vektorraums $\mathbb{K}^n$ das in (3.12) eingeführte Euklidische Skalarprodukt $\langle \cdot, \cdot \rangle_2$, das die Euklidische Norm $\| \cdot \|_2$ induziert.

Unser Ziel ist es, zu einer gegebenen Matrix $\mathbf{A} \in \mathbb{K}^{m \times n}$ und einem gegebenen Vektor $\mathbf{b} \in \mathbb{K}^m$ einen Vektor $\mathbf{x} \in \mathbb{K}^n$ so zu finden, dass die Euklidische Norm des *Defekts*

$$\mathbf{A}\mathbf{x} - \mathbf{b}$$

so klein wie möglich wird. Dazu untersuchen wir zunächst, ob wir den Defekt reduzieren können, indem wir ein Vielfaches eines anderen Vektors hinzuaddieren.

Lemma 4.2 (Optimalität) *Sei $\mathbf{A} \in \mathbb{K}^{m \times n}$, sei $\mathbf{b} \in \mathbb{K}^m$, und seien $\mathbf{x}, \mathbf{z} \in \mathbb{K}^n$. Dann gilt die Minimalitätseigenschaft*

$$\|\mathbf{A}\mathbf{x} - \mathbf{b}\|_2 \leq \|\mathbf{A}(\mathbf{x} + \lambda \mathbf{z}) - \mathbf{b}\|_2 \qquad \text{für alle } \lambda \in \mathbb{K} \tag{4.1a}$$

genau dann, wenn die Gleichung

$$\langle \mathbf{A}\mathbf{z}, \mathbf{A}\mathbf{x} - \mathbf{b} \rangle_2 = 0 \tag{4.1b}$$

gilt. Wir können $\mathbf{x}$ also nicht durch Hinzuaddieren eines Vielfachens des Vektors $\mathbf{z}$ verbessern, falls $\mathbf{A}\mathbf{z}$ senkrecht auf dem Defekt $\mathbf{A}\mathbf{x} - \mathbf{b}$ steht.

Beweis. Falls $\mathbf{A}\mathbf{z} = \mathbf{0}$ gilt, gelten (4.1a) und (4.1b) immer.

Also setzen wir nun $\mathbf{A}\mathbf{z} \neq \mathbf{0}$ voraus. Für alle $\lambda \in \mathbb{K}$ gilt

$$\begin{aligned}
\|\mathbf{A}(\mathbf{x} + \lambda \mathbf{z}) - \mathbf{b}\|_2^2 &= \|(\mathbf{A}\mathbf{x} - \mathbf{b}) + \lambda \mathbf{A}\mathbf{z}\|_2^2 \\
&= \langle (\mathbf{A}\mathbf{x} - \mathbf{b}) + \lambda \mathbf{A}\mathbf{z}, (\mathbf{A}\mathbf{x} - \mathbf{b}) + \lambda \mathbf{A}\mathbf{z} \rangle_2 \\
&= \langle \mathbf{A}\mathbf{x} - \mathbf{b}, \mathbf{A}\mathbf{x} - \mathbf{b} \rangle_2 + \langle \lambda \mathbf{A}\mathbf{z}, \lambda \mathbf{A}\mathbf{z} \rangle_2 + \langle \mathbf{A}\mathbf{x} - \mathbf{b}, \lambda \mathbf{A}\mathbf{z} \rangle_2 + \langle \lambda \mathbf{A}\mathbf{z}, \mathbf{A}\mathbf{x} - \mathbf{b} \rangle_2 \\
&= \|\mathbf{A}\mathbf{x} - \mathbf{b}\|_2^2 + |\lambda|^2 \|\mathbf{A}\mathbf{z}\|_2^2 + \lambda \langle \mathbf{A}\mathbf{x} - \mathbf{b}, \mathbf{A}\mathbf{z} \rangle_2 + \bar{\lambda} \langle \mathbf{A}\mathbf{z}, \mathbf{A}\mathbf{x} - \mathbf{b} \rangle_2.
\end{aligned} \tag{4.2}$$

Gelte (4.1a). Um den maximalen Nutzen aus dieser Ungleichung zu ziehen, wählen wir λ in (4.2) so, dass die rechte Seite möglichst klein wird. Eine Kurvendiskussion legt es nahe,

$$\lambda := -\frac{\langle \mathbf{A}\mathbf{z}, \mathbf{A}\mathbf{x} - \mathbf{b} \rangle_2}{\|\mathbf{A}\mathbf{z}\|_2^2}$$

zu verwenden, so dass wir

$$\|\mathbf{A}(\mathbf{x} + \lambda \mathbf{z}) - \mathbf{b}\|_2^2 = \|\mathbf{A}\mathbf{x} - \mathbf{b}\|_2^2 + \frac{|\langle \mathbf{A}\mathbf{z}, \mathbf{A}\mathbf{x} - \mathbf{b} \rangle_2|^2}{\|\mathbf{A}\mathbf{z}\|_2^2} - 2\frac{|\langle \mathbf{A}\mathbf{z}, \mathbf{A}\mathbf{x} - \mathbf{b} \rangle_2|^2}{\|\mathbf{A}\mathbf{z}\|_2^2}$$

$$= \|\mathbf{Ax} - \mathbf{b}\|_2^2 - \frac{|\langle \mathbf{Az}, \mathbf{Ax} - \mathbf{b}\rangle_2|^2}{\|\mathbf{Az}\|_2^2}$$

erhalten. Aus (4.1a) folgt unmittelbar

$$\frac{|\langle \mathbf{Az}, \mathbf{Ax} - \mathbf{b}\rangle_2|^2}{\|\mathbf{Az}\|_2^2} \leq 0,$$

also insbesondere (4.1b).

Gelte nun umgekehrt (4.1b). Einsetzen in (4.2) führt zu

$$\|\mathbf{A}(\mathbf{x} + \lambda\mathbf{z}) - \mathbf{b}\|_2^2 = \|\mathbf{Ax} - \mathbf{b}\|_2^2 + |\lambda|^2\|\mathbf{Az}\|_2^2 \geq \|\mathbf{Ax} - \mathbf{b}\|_2^2 \qquad \text{für alle } \lambda \in \mathbb{K},$$

also ist (4.1a) gezeigt. ∎

Falls (4.1b) für *alle* Vektoren $\mathbf{z} \in \mathbb{K}^n$ gelten sollte, können wir mit Hilfe des Lemmas 3.41 ein lineares Gleichungssystem herleiten, das äquivalent mit der Minimierungsaufgabe ist.

Satz 4.3 (Normalengleichung) *Sei $\mathbf{A} \in \mathbb{K}^{m\times n}$, sei $\mathbf{b} \in \mathbb{K}^m$, und sei $\mathbf{x} \in \mathbb{K}^n$. Dann gilt die Minimalitätseigenschaft*

$$\|\mathbf{Ax} - \mathbf{b}\|_2 \leq \|\mathbf{Ay} - \mathbf{b}\|_2 \qquad \textit{für alle } \mathbf{y} \in \mathbb{K}^n \tag{4.3a}$$

genau dann, wenn $\mathbf{x}$ *die* Normalengleichung

$$\mathbf{A}^*\mathbf{Ax} = \mathbf{A}^*\mathbf{b} \tag{4.3b}$$

löst. Das Minimierungsproblem ist also äquivalent mit einem linearen Gleichungssystem.

Beweis. Gelte (4.3a). Sei $\mathbf{z} \in \mathbb{K}^n$. Aus (4.3a) folgt

$$\|\mathbf{Ax} - \mathbf{b}\|_2 \leq \|\mathbf{A}(\mathbf{x} + \lambda\mathbf{z}) - \mathbf{b}\|_2 \qquad \text{für alle } \lambda \in \mathbb{K},$$

also (4.1a). Mit Lemma 3.44) und Lemma 4.2 folgt

$$\langle \mathbf{z}, \mathbf{A}^*(\mathbf{Ax} - \mathbf{b})\rangle_2 = \langle \mathbf{Az}, \mathbf{Ax} - \mathbf{b}\rangle_2 = 0 \qquad \text{für alle } \mathbf{z} \in \mathbb{K}^n.$$

Mit Lemma 3.41 folgt daraus $\mathbf{A}^*(\mathbf{Ax} - \mathbf{b}) = \mathbf{0}$, also die Normalengleichung (4.3b).

Gelte nun (4.3b). Sei $\mathbf{y} \in \mathbb{K}^n$. Wir setzen $\mathbf{z} := \mathbf{y} - \mathbf{x}$. Mit Lemma 3.44 folgt

$$\langle \mathbf{Az}, \mathbf{Ax} - \mathbf{b}\rangle_2 = \langle \mathbf{z}, \mathbf{A}^*(\mathbf{Ax} - \mathbf{b})\rangle_2 = \langle \mathbf{z}, \mathbf{0}\rangle_2 = 0,$$

also (4.1b). Mit Lemma 4.2 folgt (4.1a), und für $\lambda = 1$ erhalten wir

$$\|\mathbf{Ay} - \mathbf{b}\|_2 = \|\mathbf{A}(\mathbf{x} + \lambda\mathbf{z}) - \mathbf{b}\|_2 \leq \|\mathbf{Ax} - \mathbf{b}\|_2,$$

also die Minimalitätseigenschaft (4.3a). ∎

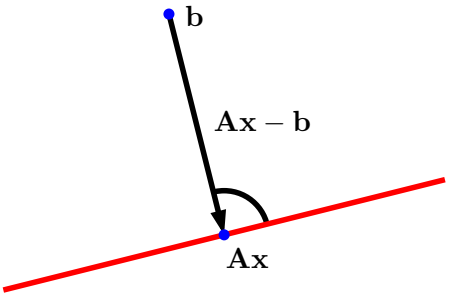

Abbildung 4.1: Geometrische Interpretation der Normalengleichung

Die Normalengleichung lässt sich geometrisch interpretieren: Aus (4.1b) folgt, dass $\mathbf{x}$ genau dann das Ausgleichsproblem löst, wenn der Defekt $\mathbf{Ax}-\mathbf{b}$ senkrecht auf dem Bild der Matrix $\mathbf{A}$ steht, wenn also $\mathbf{Ax}$ der Fußpunkt des von $\mathbf{b}$ auf das Bild gefällten Lots ist (vgl. Abbildung 4.1).

Offenbar kann die Lösung eines linearen Ausgleichsproblems nur dann eindeutig sein, wenn der Kern der Matrix $\mathbf{A}$ nur die Null enthält, denn ansonsten könnte man zu jeder Lösung ein von null verschiedenes Element des Kerns addieren und würde eine weitere Lösung erhalten. Deshalb gehen wir davon aus, dass $\mathbf{A}$ injektiv, sein Kern also trivial ist. Insbesondere ist dann $m \geq n$. In diesem Fall können wir eine Aussage über Existenz und Eindeutigkeit der Lösung gewinnen:

Erinnerung 4.4 (Dimensionsformel bzw. Rangsatz) *Seien $\mathcal{V}$ und $\mathcal{W}$ endlich-dimensionale Vektorräume. Sei $L\colon \mathcal{V} \to \mathcal{W}$ eine lineare Abbildung. Dann gilt*

$$\dim(\mathrm{Bild}(L)) + \dim(\mathrm{Kern}(L)) = \dim(\mathcal{V}).$$

Folgerung 4.5 (Lösbarkeit) *Für alle $\mathbf{A} \in \mathbb{K}^{m\times n}$ und $\mathbf{b} \in \mathbb{K}^m$ existiert ein $\mathbf{x} \in \mathbb{K}^n$ mit*

$$\mathbf{A}^*\mathbf{Ax} = \mathbf{A}^*\mathbf{b},$$

das Ausgleichsproblem ist also immer lösbar.

Falls $\mathbf{A}$ injektiv ist, ist die Lösung eindeutig bestimmt.

Beweis. Wir definieren den Bildraum $\mathcal{R} \subseteq \mathbb{K}^n$ der Matrix $\mathbf{A}^*$ und den Kern $\mathcal{N} \subseteq \mathbb{K}^n$ der Matrix $\mathbf{A}$ durch

$$\mathcal{R} := \{\mathbf{A}^*\mathbf{z} \ : \ \mathbf{z} \in \mathbb{K}^m\}, \qquad \mathcal{N} := \{\mathbf{y} \in \mathbb{K}^n \ : \ \mathbf{Ay} = \mathbf{0}\}.$$

Sei $\mathbf{x} \in \mathcal{R}$ und $\mathbf{y} \in \mathcal{N}$. Nach Definition finden wir $\mathbf{z} \in \mathbb{K}^m$ mit $\mathbf{x} = \mathbf{A}^*\mathbf{z}$, und es folgt mit Lemma 3.44

$$\langle \mathbf{x}, \mathbf{y}\rangle_2 = \langle \mathbf{A}^*\mathbf{z}, \mathbf{y}\rangle_2 = \langle \mathbf{z}, \mathbf{Ay}\rangle_2 = \langle \mathbf{z}, \mathbf{0}\rangle_2 = 0.$$

Für $\mathbf{x} \in \mathcal{R} \cap \mathcal{N}$ folgt insbesondere

$$\|\mathbf{x}\|_2^2 = \langle \mathbf{x}, \mathbf{x}\rangle_2 = 0.$$

Wir definieren die lineare Abbildung

$$L : \mathcal{R} \to \mathcal{R}, \qquad \mathbf{x} \mapsto \mathbf{A}^*\mathbf{A}\mathbf{x}.$$

Sei $\mathbf{x} \in \mathcal{R}$ mit $L\mathbf{x} = \mathbf{0}$ gegeben. Dann folgt mit Lemma 3.44

$$0 = \langle \mathbf{x}, L\mathbf{x}\rangle_2 = \langle \mathbf{x}, \mathbf{A}^*\mathbf{A}\mathbf{x}\rangle_2 = \langle \mathbf{A}\mathbf{x}, \mathbf{A}\mathbf{x}\rangle_2 = \|\mathbf{A}\mathbf{x}\|_2^2,$$

also $\mathbf{x} \in \mathcal{N}$, und damit bereits $\mathbf{x} = \mathbf{0}$. Also ist L injektiv, es gilt $\dim(\mathrm{Kern}(L)) = 0$.

Mit der Dimensionsformel (vgl. Erinnerung 4.4) folgt

$$\dim(\mathcal{R}) = \dim(\mathrm{Bild}(L)) + \dim(\mathrm{Kern}(L)) = \dim(\mathrm{Bild}(L)),$$

also muss L auch bijektiv sein.

Sei nun $\mathbf{b} \in \mathbb{K}^m$. Dann gilt $\mathbf{A}^*\mathbf{b} \in \mathcal{R}$, also existiert nach dem zuvor Gezeigten ein $\mathbf{x} \in \mathcal{R}$ mit

$$\mathbf{A}^*\mathbf{A}\mathbf{x} = L\mathbf{x} = \mathbf{A}^*\mathbf{b}.$$

Damit ist die Lösbarkeit bewiesen.

Sei $\mathbf{A}$ injektiv, und seien $\mathbf{x}, \widetilde{\mathbf{x}} \in \mathbb{K}^n$ mit

$$\mathbf{A}^*\mathbf{A}\mathbf{x} = \mathbf{A}^*\mathbf{b}, \qquad \mathbf{A}^*\mathbf{A}\widetilde{\mathbf{x}} = \mathbf{A}^*\mathbf{b}$$

gegeben. Durch Subtraktion der Gleichungen folgt

$$\mathbf{A}^*\mathbf{A}(\mathbf{x} - \widetilde{\mathbf{x}}) = \mathbf{0},$$

und wir erhalten

$$0 = \langle \mathbf{x} - \widetilde{\mathbf{x}}, \mathbf{A}^*\mathbf{A}(\mathbf{x} - \widetilde{\mathbf{x}})\rangle_2 = \langle \mathbf{A}(\mathbf{x} - \widetilde{\mathbf{x}}), \mathbf{A}(\mathbf{x} - \widetilde{\mathbf{x}})\rangle_2 = \|\mathbf{A}(\mathbf{x} - \widetilde{\mathbf{x}})\|_2^2.$$

Da $\mathbf{A}$ als injektiv vorausgesetzt ist, folgt daraus bereits $\mathbf{x} = \widetilde{\mathbf{x}}$. ∎

Übungsaufgabe 4.6 (Bild und Kern) *Sei $\mathbf{A} \in \mathbb{K}^{n\times n}$ selbstadjungiert.*

(a) Beweisen Sie $\langle \mathbf{x}, \mathbf{y}\rangle_2 = 0$ für alle $\mathbf{x} \in \mathrm{Bild}(\mathbf{A})$ und alle $\mathbf{y} \in \mathrm{Kern}(\mathbf{A})$.

(b) Beweisen Sie $\mathbb{K}^n = \mathrm{Bild}(\mathbf{A}) \oplus \mathrm{Kern}(\mathbf{A})$.

Übungsaufgabe 4.7 (Minimumnormlösung) *Sei $\mathbf{A} \in \mathbb{K}^{m\times n}$, sei $\mathbf{b} \in \mathbb{K}^m$. Im Beweis der Folgerung 4.5 haben wir gesehen, dass wir eine Lösung $\mathbf{x}$ des linearen Ausgleichsproblems (4.3a) in der Form $\mathbf{x} = \mathbf{A}^*\mathbf{z}$ mit $\mathbf{z} \in \mathbb{K}^m$ finden können.*

Sei $\widetilde{\mathbf{x}} \in \mathbb{K}^n$ eine weitere Lösung des Ausgleichsproblems (4.3a). Beweisen Sie die Ungleichung $\|\mathbf{x}\|_2 \leq \|\widetilde{\mathbf{x}}\|_2$, dass $\mathbf{x}$ also unter allen Lösungen die minimale Norm aufweist.

4.3 Kondition

Neben der prinzipiellen Lösbarkeit des Ausgleichsproblems interessieren wir uns natürlich auch wieder für die Anfälligkeit der Lösung für Störungen. Damit die Lösung eindeutig ist, setzen wir voraus, dass die Matrix $\mathbf{A}$ injektiv ist.

Um ein Gegenstück des Lemmas 3.5 zu erhalten, benötigen wir eine Möglichkeit, die Norm $\|\mathbf{Ax}\|_2$ des Matrix-Vektor-Produkts nach oben und unten durch $\|\mathbf{x}\|_2$ abzusätzen.

Lemma 4.8 (Normschranken) *Für alle $\mathbf{A} \in \mathbb{K}^{m\times n}$ sind*

$$\alpha_2(\mathbf{A}) := \min\{\|\mathbf{Ay}\|_2 \ : \ \mathbf{y} \in \mathbb{K}^n,\ \|\mathbf{y}\|_2 = 1\},$$
$$\beta_2(\mathbf{A}) := \max\{\|\mathbf{Ay}\|_2 \ : \ \mathbf{y} \in \mathbb{K}^n,\ \|\mathbf{y}\|_2 = 1\}$$

wohldefinierte reelle Zahlen, die

$$\alpha_2(\mathbf{A})\|\mathbf{z}\|_2 \le \|\mathbf{Az}\|_2 \le \beta_2(\mathbf{A})\|\mathbf{z}\|_2 \qquad \text{für alle } \mathbf{z} \in \mathbb{K}^n \tag{4.4}$$

erfüllen. Falls $\mathbf{A}$ injektiv ist, gilt $\alpha_2(\mathbf{A}) > 0$.

Beweis. Die n-dimensionale Einheitssphäre

$$\mathcal{S} := \{\mathbf{y} \in \mathbb{K}^n \ : \ \|\mathbf{y}\|_2 = 1\}$$

ist nach dem Satz von Heine-Borel (vgl. Erinnerung 3.1) kompakt.

Also muss die stetige Abbildung $\mathbf{y} \mapsto \|\mathbf{Ay}\|_2$ auf dieser Menge ein Minimum und ein Maximum annehmen, nämlich gerade $\alpha_2(\mathbf{A})$ und $\beta_2(\mathbf{A})$.

Sei nun $\mathbf{z} \in \mathbb{K}^n$. Falls $\mathbf{z} = \mathbf{0}$ gilt, folgt (4.4) unmittelbar.

Ansonsten definieren wir $\mathbf{y} := \mathbf{z}/\|\mathbf{z}\|_2 \in \mathcal{S}$ und erhalten

$$\alpha_2(\mathbf{A})\|\mathbf{z}\|_2 \le \|\mathbf{Ay}\|_2\|\mathbf{z}\|_2 = \|\mathbf{Az}\|_2, \qquad \|\mathbf{Az}\|_2 = \|\mathbf{Ay}\|_2\|\mathbf{z}\|_2 \le \beta_2(\mathbf{A})\|\mathbf{z}\|_2.$$

Sei nun $\mathbf{A}$ injektiv, und sei $\mathbf{y} \in \mathbb{K}^n$ mit $\|\mathbf{y}\|_2 = 1$ ein Vektor, für den $\|\mathbf{Ay}\|_2 = \alpha_2(\mathbf{A})$ gilt. Da $\mathbf{A}$ injektiv und $\mathbf{y}$ nicht der Nullvektor ist, folgt $0 < \|\mathbf{Ay}\|_2 = \alpha_2(\mathbf{A})$. ∎

Der Faktor $\beta_2(\mathbf{A})$ ist uns schon bekannt: Es gilt

$$\beta_2(\mathbf{A}) = \max\left\{\frac{\|\mathbf{Az}\|_2}{\|\mathbf{z}\|_2} \ : \ \mathbf{z} \in \mathbb{K}^n \setminus \{\mathbf{0}\}\right\} = \|\mathbf{A}\|_2, \tag{4.5}$$

die neue Bezeichnung für die Matrixnorm wurde nur eingeführt, um die Ungleichung (4.4) etwas einheitlicher zu gestalten.

Falls $\mathbf{A} \in \mathbb{K}^{n\times n}$ regulär ist, haben wir

$$\begin{aligned}\alpha_2(\mathbf{A}) &= \min\left\{\frac{\|\mathbf{Az}\|_2}{\|\mathbf{z}\|_2} \ : \ \mathbf{z} \in \mathbb{K}^n \setminus \{\mathbf{0}\}\right\} = \min\left\{\frac{\|\mathbf{y}\|_2}{\|\mathbf{A}^{-1}\mathbf{y}\|_2} \ : \ \mathbf{y} \in \mathbb{K}^n \setminus \{\mathbf{0}\}\right\} \\ &= \frac{1}{\max\left\{\frac{\|\mathbf{A}^{-1}\mathbf{y}\|_2}{\|\mathbf{y}\|_2} \ : \ \mathbf{y} \in \mathbb{K}^n \setminus \{\mathbf{0}\}\right\}} = \frac{1}{\|\mathbf{A}^{-1}\|_2}\end{aligned} \tag{4.6}$$

dank der Substitution $\mathbf{z} = \mathbf{A}^{-1}\mathbf{y}$.

Lemma 4.9 (Kondition) *Sei $\mathbf{A} \in \mathbb{K}^{m\times n}$ injektiv. Dann ist $\mathbf{A}^*\mathbf{A}$ regulär. Seien $\mathbf{b}, \widetilde{\mathbf{b}} \in \mathbb{K}^m$ gegeben, und seien $\mathbf{x}, \widetilde{\mathbf{x}} \in \mathbb{K}^n$ die Lösungen der Ausgleichsprobleme*

$$\|\mathbf{A}\mathbf{x} - \mathbf{b}\|_2 \leq \|\mathbf{A}\mathbf{y} - \mathbf{b}\|_2, \tag{4.7a}$$

$$\|\mathbf{A}\widetilde{\mathbf{x}} - \widetilde{\mathbf{b}}\|_2 \leq \|\mathbf{A}\mathbf{y} - \widetilde{\mathbf{b}}\|_2 \qquad \text{für alle } \mathbf{y} \in \mathbb{K}^n. \tag{4.7b}$$

Sei außerdem $\mathbf{x} \neq \mathbf{0}$. Mit der Matrix $\mathbf{P} := \mathbf{A}(\mathbf{A}^*\mathbf{A})^{-1}\mathbf{A}^*$ gelten dann $\mathbf{P}\mathbf{b} \neq \mathbf{0}$ und

$$\frac{\|\mathbf{x} - \widetilde{\mathbf{x}}\|_2}{\|\mathbf{x}\|_2} \leq \frac{\beta_2(\mathbf{A})}{\alpha_2(\mathbf{A})} \frac{\|\mathbf{P}(\mathbf{b} - \widetilde{\mathbf{b}})\|_2}{\|\mathbf{P}\mathbf{b}\|_2}. \tag{4.8}$$

Beweis. Da $\mathbf{A}$ injektiv ist, gilt mit Lemma 3.44)

$$\langle \mathbf{z}, \mathbf{A}^*\mathbf{A}\mathbf{z}\rangle_2 = \langle \mathbf{A}\mathbf{z}, \mathbf{A}\mathbf{z}\rangle_2 = \|\mathbf{A}\mathbf{z}\|_2^2 > 0 \qquad \text{für alle } \mathbf{z} \in \mathbb{K}^n \setminus \{\mathbf{0}\},$$

also muss auch $\mathbf{A}^*\mathbf{A}$ positiv definit sein, also nach Lemma 3.47 auch regulär.

Aufgrund von Satz 4.3 und Lemma 4.8 gilt

$$\begin{aligned}\|\mathbf{x} - \widetilde{\mathbf{x}}\|_2 &= \|(\mathbf{A}^*\mathbf{A})^{-1}\mathbf{A}^*\mathbf{b} - (\mathbf{A}^*\mathbf{A})^{-1}\mathbf{A}^*\widetilde{\mathbf{b}}\|_2 = \|(\mathbf{A}^*\mathbf{A})^{-1}\mathbf{A}^*(\mathbf{b} - \widetilde{\mathbf{b}})\|_2 \\ &\leq \frac{1}{\alpha_2(\mathbf{A})}\|\mathbf{A}(\mathbf{A}^*\mathbf{A})^{-1}\mathbf{A}^*(\mathbf{b} - \widetilde{\mathbf{b}})\|_2 = \frac{\|\mathbf{P}(\mathbf{b} - \widetilde{\mathbf{b}})\|_2}{\alpha_2(\mathbf{A})},\end{aligned}$$

und für die Norm der Lösung erhalten wir entsprechend die Schranke

$$\|\mathbf{x}\|_2 = \|(\mathbf{A}^*\mathbf{A})^{-1}\mathbf{A}^*\mathbf{b}\|_2 \geq \frac{1}{\beta_2(\mathbf{A})}\|\mathbf{A}(\mathbf{A}^*\mathbf{A})^{-1}\mathbf{A}^*\mathbf{b}\|_2 = \frac{\|\mathbf{P}\mathbf{b}\|_2}{\beta_2(\mathbf{A})}.$$

Indem wir beide Abschätzungen kombinieren, finden wir

$$\frac{\|\mathbf{x} - \widetilde{\mathbf{x}}\|_2}{\|\mathbf{x}\|_2} \leq \frac{\beta_2(\mathbf{A})}{\alpha_2(\mathbf{A})} \frac{\|\mathbf{P}(\mathbf{b} - \widetilde{\mathbf{b}})\|_2}{\|\mathbf{P}\mathbf{b}\|_2},$$

und das ist die gesuchte Abschätzung. ∎

Der erste Faktor der Abschätzung (4.8) lässt sich auf etwas uns bereits Bekanntes zurückführen: Wäre $\mathbf{A}$ eine quadratische Matrix, so würde die Injektivität bereits die Existenz von $\mathbf{A}^{-1}$ nach sich ziehen, und wir hätten mit (4.6) und (4.5) die Gleichung

$$\frac{\beta_2(\mathbf{A})}{\alpha_2(\mathbf{A})} = \|\mathbf{A}\|_2\|\mathbf{A}^{-1}\|_2 = \kappa_2(\mathbf{A}).$$

Dieser Faktor ist also lediglich eine Verallgemeinerung der uns bereits von der Untersuchung linearer Gleichungssysteme in Lemma 3.5 her bekannten Konditionszahl. Dementsprechend bezeichnen wir im hier betrachteten Fall injektiver rechteckiger Matrizen auch die Größe

$$\kappa_2(\mathbf{A}) := \frac{\beta_2(\mathbf{A})}{\alpha_2(\mathbf{A})} \tag{4.9}$$

als Konditionszahl der Matrix $\mathbf{A}$ bezüglich der euklidischen Norm.

Um den zweiten Faktor interpretieren zu können, müssen wir die Matrix $\mathbf{P} = \mathbf{A}(\mathbf{A}^*\mathbf{A})^{-1}\mathbf{A}^*$ untersuchen, die sowohl im Zähler als auch im Nenner dieses Faktors auftritt. Bei ihr handelt es sich um eine *orthogonale Projektion*.

Definition 4.10 (Orthogonale Projektion) *Sei* $\mathbf{P} \in \mathbb{K}^{m\times m}$*. Falls die Gleichungen* $\mathbf{P}^* = \mathbf{P}$ *und* $\mathbf{P}^2 = \mathbf{P}$ *gelten, nennen wir* $\mathbf{P}$ *eine* orthogonale Projektion.

Die in Lemma 4.9 definierte Matrix $\mathbf{P} = \mathbf{A}(\mathbf{A}^*\mathbf{A})^{-1}\mathbf{A}^*$ ist in der Tat eine orthogonale Projektion, denn es gelten

$$\mathbf{P}^* = (\mathbf{A}(\mathbf{A}^*\mathbf{A})^{-1}\mathbf{A}^*)^* = \mathbf{A}^{**}((\mathbf{A}^*\mathbf{A})^{-1})^*\mathbf{A}^* = \mathbf{A}(\mathbf{A}^*\mathbf{A})^{-1}\mathbf{A}^* = \mathbf{P}$$

wegen Lemma 3.45 und

$$\mathbf{P}^2 = \mathbf{A}(\mathbf{A}^*\mathbf{A})^{-1}\mathbf{A}^*\mathbf{A}(\mathbf{A}^*\mathbf{A})^{-1}\mathbf{A}^* = \mathbf{A}(\mathbf{A}^*\mathbf{A})^{-1}\mathbf{A}^* = \mathbf{P}.$$

Eine wichtige Eigenschaft orthogonaler Projektionen besteht darin, dass sie einen beliebigen Vektor auf dessen beste Approximation in ihrem Bildraum abbilden.

Lemma 4.11 (Bestapproximation) *Sei* $\mathbf{P} \in \mathbb{K}^{m\times m}$ *eine orthogonale Projektion. Dann gilt*

$$\|\mathbf{x} - \mathbf{P}\mathbf{x}\|_2^2 + \|\mathbf{P}\mathbf{x} - \mathbf{y}\|_2^2 = \|\mathbf{x} - \mathbf{y}\|_2^2 \qquad \text{für alle } \mathbf{x} \in \mathbb{K}^m,\ \mathbf{y} \in \operatorname{Bild}\mathbf{P}.$$

Aus dieser Gleichung folgen

$$\|\mathbf{x} - \mathbf{P}\mathbf{x}\|_2 \le \|\mathbf{x} - \mathbf{y}\|_2 \qquad \text{für alle } \mathbf{x} \in \mathbb{K}^m,\ \mathbf{y} \in \operatorname{Bild}\mathbf{P},$$

also dass $\mathbf{P}\mathbf{x}$ *die bestmögliche Approximation des Vektors* $\mathbf{x}$ *in* $\operatorname{Bild}\mathbf{P}$ *ist, und*

$$\|\mathbf{P}\mathbf{x}\|_2 \le \|\mathbf{x}\|_2 \qquad \text{für alle } \mathbf{x} \in \mathbb{K}^m,$$

also dass die Matrixnorm von $\mathbf{P}$ *durch eins beschränkt ist.*

Beweis. Seien $\mathbf{x} \in \mathbb{K}^m$ und $\mathbf{y} \in \operatorname{Bild}\mathbf{P}$ gegeben. Nach Voraussetzung muss ein $\mathbf{z} \in \mathbb{K}^m$ mit $\mathbf{y} = \mathbf{P}\mathbf{z}$ existieren, und mit diesem $\mathbf{z}$ erhalten wir wegen $\mathbf{P}^2 = \mathbf{P}$ die Gleichung

$$\mathbf{P}\mathbf{y} = \mathbf{P}\mathbf{P}\mathbf{z} = \mathbf{P}^2\mathbf{z} = \mathbf{P}\mathbf{z} = \mathbf{y},$$

also ist $\mathbf{y}$ eines seiner eigenen Urbilder. Daraus folgt nun mit (3.17) die Beziehung

$$\begin{aligned}
\|\mathbf{x} - \mathbf{y}\|_2^2 &= \|\mathbf{x} - \mathbf{P}\mathbf{x} + \mathbf{P}\mathbf{x} - \mathbf{y}\|_2^2 = \|\mathbf{x} - \mathbf{P}\mathbf{x} + \mathbf{P}(\mathbf{x} - \mathbf{y})\|_2^2 \\
&= \|\mathbf{x} - \mathbf{P}\mathbf{x}\|_2^2 + \langle \mathbf{x} - \mathbf{P}\mathbf{x}, \mathbf{P}(\mathbf{x} - \mathbf{y})\rangle_2 + \langle \mathbf{P}(\mathbf{x} - \mathbf{y}), \mathbf{x} - \mathbf{P}\mathbf{x}\rangle_2 + \|\mathbf{P}(\mathbf{x} - \mathbf{y})\|_2^2 \\
&= \|\mathbf{x} - \mathbf{P}\mathbf{x}\|_2^2 + \langle \mathbf{P}^*(\mathbf{x} - \mathbf{P}\mathbf{x}), \mathbf{x} - \mathbf{y}\rangle_2 + \langle \mathbf{x} - \mathbf{y}, \mathbf{P}^*(\mathbf{x} - \mathbf{P}\mathbf{x})\rangle_2 + \|\mathbf{P}\mathbf{x} - \mathbf{y}\|_2^2 \\
&= \|\mathbf{x} - \mathbf{P}\mathbf{x}\|_2^2 + \langle \mathbf{P}(\mathbf{x} - \mathbf{P}\mathbf{x}), \mathbf{x} - \mathbf{y}\rangle_2 + \langle \mathbf{x} - \mathbf{y}, \mathbf{P}(\mathbf{x} - \mathbf{P}\mathbf{x})\rangle_2 + \|\mathbf{P}\mathbf{x} - \mathbf{y}\|_2^2 \\
&= \|\mathbf{x} - \mathbf{P}\mathbf{x}\|_2^2 + \langle \mathbf{P}\mathbf{x} - \mathbf{P}^2\mathbf{x}, \mathbf{x} - \mathbf{y}\rangle_2 + \langle \mathbf{x} - \mathbf{y}, \mathbf{P}\mathbf{x} - \mathbf{P}^2\mathbf{x}\rangle_2 + \|\mathbf{P}\mathbf{x} - \mathbf{y}\|_2^2 \\
&= \|\mathbf{x} - \mathbf{P}\mathbf{x}\|_2^2 + \|\mathbf{P}\mathbf{x} - \mathbf{y}\|_2^2.
\end{aligned}$$

Indem wir den zweiten Term durch null nach unten abschätzen erhalten wir die erste Ungleichung, indem wir $\mathbf{y} = \mathbf{0}$ einsetzen und den ersten Term durch null abschätzen die zweite. ∎

Somit bildet $\mathbf{P}$ gerade jeden Vektor auf seine beste Approximation im Bild von $\mathbf{P}$ ab. Das Bild der Matrix $\mathbf{P}$ lässt sich leicht bestimmen: Nach Definition muss es im Bild der Matrix $\mathbf{A}$ enthalten sein, und für jedes $\mathbf{x} \in \operatorname{Bild} \mathbf{A}$ finden wir ein $\mathbf{z} \in \mathbb{K}^n$ mit $\mathbf{x} = \mathbf{A}\mathbf{z}$, also

$$\mathbf{P}\mathbf{x} = \mathbf{A}(\mathbf{A}^*\mathbf{A})^{-1}\mathbf{A}^*\mathbf{A}\mathbf{z} = \mathbf{A}\mathbf{z} = \mathbf{x},$$

so dass das Bild der Matrix $\mathbf{A}$ auch eine Teilmenge des Bildes der Matrix $\mathbf{P}$ sein muss, also folgt $\operatorname{Bild} \mathbf{A} = \operatorname{Bild} \mathbf{P}$.

Demzufolge ordnet die orthogonale Projektion $\mathbf{P}$ jeder rechten Seite $\mathbf{b}$ ihre beste Approximation im Bild der Matrix $\mathbf{A}$ zu. Der Unterschied zwischen dieser Approximation und dem ursprünglichen Vektor $\mathbf{b}$ lässt sich geometrisch als *Winkel* interpretieren:

Definition 4.12 (Winkel) *Sei $\mathcal{V} \subseteq \mathbb{K}^m$ ein nicht-trivialer Vektorraum, und sei $\mathbf{x} \in \mathbb{K}^m \setminus \{\mathbf{0}\}$. Der* Winkel *zwischen $\mathbf{x}$ und $\mathcal{V}$ ist definiert durch*

$$\cos \angle(\mathbf{x}, \mathcal{V}) := \max_{\mathbf{y} \in \mathcal{V} \setminus \{\mathbf{0}\}} \frac{|\langle \mathbf{x}, \mathbf{y} \rangle_2|}{\|\mathbf{x}\|_2 \|\mathbf{y}\|_2}.$$

Mit Hilfe dieser Definition lässt sich das vorläufige Resultat aus Lemma 4.9 nun etwas anschaulicher darstellen:

Satz 4.13 (Kondition) *Sei $\mathbf{A} \in \mathbb{K}^{m \times n}$ injektiv, seien $\mathbf{b}, \widetilde{\mathbf{b}} \in \mathbb{K}^m$ gegeben, und seien $\mathbf{x}, \widetilde{\mathbf{x}} \in \mathbb{K}^n$ die Lösungen der Ausgleichsprobleme (4.7) zu diesen rechten Seiten. Sei außerdem $\mathbf{x} \neq \mathbf{0}$. Dann gilt*

$$\frac{\|\mathbf{x} - \widetilde{\mathbf{x}}\|_2}{\|\mathbf{x}\|_2} \leq \frac{\kappa_2(\mathbf{A})}{\cos \angle(\mathbf{b}, \operatorname{Bild} \mathbf{A})} \frac{\|\mathbf{b} - \widetilde{\mathbf{b}}\|_2}{\|\mathbf{b}\|_2}$$

mit der verallgemeinerten Konditionszahl $\kappa_2(\mathbf{A})$ aus (4.9).

Beweis. Wir wenden Definition 4.12 auf $\mathcal{V} = \operatorname{Bild} \mathbf{A} = \operatorname{Bild} \mathbf{P}$ an und erhalten

$$\begin{aligned}\cos \angle(\mathbf{b}, \operatorname{Bild} \mathbf{A}) &= \max_{\mathbf{y} \in \operatorname{Bild} \mathbf{A} \setminus \{\mathbf{0}\}} \frac{|\langle \mathbf{b}, \mathbf{y} \rangle_2|}{\|\mathbf{b}\|_2 \|\mathbf{y}\|_2} = \max_{\mathbf{y} \in \operatorname{Bild} \mathbf{A} \setminus \{\mathbf{0}\}} \frac{|\langle \mathbf{b}, \mathbf{P}\mathbf{y} \rangle_2|}{\|\mathbf{b}\|_2 \|\mathbf{y}\|_2} \\ &= \max_{\mathbf{y} \in \operatorname{Bild} \mathbf{A} \setminus \{\mathbf{0}\}} \frac{|\langle \mathbf{P}^*\mathbf{b}, \mathbf{y} \rangle_2|}{\|\mathbf{b}\|_2 \|\mathbf{y}\|_2} = \max_{\mathbf{y} \in \operatorname{Bild} \mathbf{A} \setminus \{\mathbf{0}\}} \frac{|\langle \mathbf{P}\mathbf{b}, \mathbf{y} \rangle_2|}{\|\mathbf{b}\|_2 \|\mathbf{y}\|_2}.\end{aligned}$$

Nach Lemma 4.9 gilt $\mathbf{P}\mathbf{b} \neq \mathbf{0}$. Mit der Cauchy-Schwarz-Ungleichung aus Lemma 3.40, folgt, dass das Maximum gerade für $\mathbf{y} = \mathbf{P}\mathbf{b}$ angenommen wird, und wir erhalten

$$\cos \angle(\mathbf{b}, \operatorname{Bild} \mathbf{A}) = \frac{\|\mathbf{P}\mathbf{b}\|_2}{\|\mathbf{b}\|_2}, \qquad \frac{1}{\|\mathbf{P}\mathbf{b}\|_2} = \frac{1}{\cos \angle(\mathbf{b}, \operatorname{Bild} \mathbf{A})} \frac{1}{\|\mathbf{b}\|_2}.$$

Aus Lemma 4.11 ergibt sich

$$\|\mathbf{P}(\mathbf{b} - \widetilde{\mathbf{b}})\|_2 \leq \|\mathbf{b} - \widetilde{\mathbf{b}}\|_2,$$

und durch Einsetzen in Lemma 4.9 folgt die gesuchte Aussage über den Fehler. ∎

Anders als bei den linearen Gleichungssystemen (vgl. Lemma 3.5) geht bei den linearen Ausgleichsproblemen auch der Winkel zwischen der rechten Seite und dem Bild der Matrix $\mathbf{A}$ in die Fehlerverstärkung ein: Wenn der Winkel groß ist, führen kleine relative Störungen der rechten Seite $\mathbf{b}$ zu großen relativen Störungen der Projektion $\mathbf{Pb}$. Falls $\mathbf{b} \in \text{Bild}\,\mathbf{A}$ gilt, ist der Winkel gleich null und die Abschätzung stimmt mit dem Ergebnis aus Lemma 3.5 überein.

4.4 Lösen per Normalengleichung

Wenden wir uns nun der Berechnung der Lösung eines linearen Ausgleichsproblems zu. Die Lösung $\mathbf{x}$ des Ausgleichsproblem ist laut Satz 4.3 durch die Normalengleichung

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

gegeben.

Wir gehen davon aus, dass $\mathbf{A}$ injektiv ist, damit die Lösung eindeutig bestimmt ist. In diesem Fall ist die Matrix $\mathbf{G} := \mathbf{A}^*\mathbf{A}$ selbstadjungiert und positiv definit, so dass wir ihre Cholesky-Zerlegung (siehe Definition 3.48)

$$\mathbf{G} = \mathbf{L}\mathbf{L}^*$$

berechnen können. Die Berechnung der Lösung $\mathbf{x}$ kann dann mittels

$$\mathbf{L}\mathbf{L}^*\mathbf{x} = \mathbf{A}^*\mathbf{b}$$

in der bekannten Weise durch Vorwärts- und Rückwärtseinsetzen erfolgen. Der resultierende Algorithmus ist in Abbildung 4.2 zusammengefasst.

Untersuchen wir nun den Rechenaufwand dieser Methode. Da $\mathbf{G}$ selbstadjungiert ist, müssen wir nur die linke untere Dreieckshälfte berechnen, wofür

$$\sum_{i=1}^{n}\sum_{j=1}^{i} 2m = 2m\sum_{i=1}^{n} i = mn(n+1)$$

Operationen erforderlich sind. Die Berechnung von $\mathbf{A}^*\mathbf{b}$ erfordert $2mn$ Operationen.

Wir haben bereits gesehen, dass die Cholesky-Faktorisierung

$$\frac{1}{3}n^3 + \frac{n}{6}(3n+1)$$

Operationen erfordert und dass für Vorwärts- und Rückwärtseinsetzen jeweils n^2 Operationen anzusetzen sind. Insgesamt müssen wir also

$$mn(n+3) + \frac{1}{3}n^3 + \frac{n}{6}(3n+1) + 2n^2 = mn(n+3) + \frac{1}{3}n^3 + \frac{n}{6}(15n+1)$$

Operationen ausführen, um $\mathbf{x}$ zu bestimmen. Der Rechenaufwand wächst somit linear mit m und kubisch mit n.

In vielen praktischen Anwendungen ist m wesentlich größer als n, so dass das lineare Wachstum in m bedeutet, dass das Verfahren sehr effizient sein kann.

```
procedure normal_equation(A, b, m, n, var x);
for i ∈ [1 : n] do
  for j ∈ [1 : i] do
    g_ij ← 0
    for k ∈ [1 : m] do
      g_ij ← g_ij + ā_ki a_kj
    end for
  end for
  x_i ← 0
  for k ∈ [1 : m] do
    x_i ← x_i + ā_ki b_k
  end for
end for
decomp_cholesky(G, n)
for_subst(G, n, x)
adjback_subst(G, n, x)
```

Abbildung 4.2: Lösen eines Ausgleichsproblem mit Hilfe der Normalengleichung.

Übungsaufgabe 4.14 (Kondition) *Sei $\mathbf{A} \in \mathbb{K}^{m\times n}$ injektiv. Dann ist die* Gramsche Matrix *$\mathbf{A}^*\mathbf{A} \in \mathbb{K}^{n\times n}$ positiv definit und selbstadjungiert, so dass wir mit Satz 3.49 einen regulären Cholesky-Faktor $\mathbf{L} \in \mathbb{K}^{n\times n}$ finden mit $\mathbf{A}^*\mathbf{A} = \mathbf{L}\mathbf{L}^*$.*

(a) Beweisen Sie $\beta_2(\mathbf{A}) = \|\mathbf{L}\|_2$.

(b) Beweisen Sie $\alpha_2(\mathbf{A}) = 1/\|\mathbf{L}^{-1}\|_2$.

(c) Beweisen Sie $\kappa_2(\mathbf{A}^*\mathbf{A}) = \kappa_2(\mathbf{A})^2$.

Hinweis: *Übungsaufgabe 3.54 kann sehr hilfreich sein.*

Übungsaufgabe 4.15 (Nicht-injektive Matrix) *Sei $\mathbf{A} \in \mathbb{K}^{m\times n}$ so gegeben, dass die ersten $k \in [1 : n-1]$ Spalten linear unabhängig sind und die restlichen $n-k$ Spalten sich durch die ersten k darstellen lassen, dass also*

$$\mathbf{A} = \widehat{\mathbf{A}} \begin{pmatrix} \mathbf{I} & \mathbf{B} \end{pmatrix}$$

mit einer regulären Matrix $\widehat{\mathbf{A}} \in \mathbb{K}^{m\times k}$ und $\mathbf{B} \in \mathbb{K}^{k\times(n-k)}$ gilt.

(a) Beweisen Sie, dass eine reguläre linke untere Dreiecksmatrix $\widehat{\mathbf{L}} \in \mathbb{K}^{k\times k}$ und eine Matrix $\mathbf{C} \in \mathbb{K}^{(n-k)\times k}$ existieren mit

$$\begin{pmatrix} \widehat{\mathbf{L}} & \\ \mathbf{C} & \mathbf{0} \end{pmatrix} \begin{pmatrix} \widehat{\mathbf{L}}^* & \mathbf{C}^* \\ & \mathbf{0} \end{pmatrix} = \mathbf{A}^*\mathbf{A}.$$

(b) Beweisen Sie, dass für jedes $\mathbf{b} \in \mathbb{K}^m$ Vektoren $\mathbf{y} \in \mathbb{K}^k$ und $\mathbf{x} \in \mathbb{K}^k$ existieren mit

$$\begin{pmatrix} \widehat{\mathbf{L}} & \\ \mathbf{C} & \mathbf{0} \end{pmatrix} \begin{pmatrix} \mathbf{y} \\ \mathbf{0} \end{pmatrix} = \mathbf{A}^*\mathbf{b}, \qquad \begin{pmatrix} \widehat{\mathbf{L}}^* & \mathbf{C}^* \\ & \mathbf{0} \end{pmatrix} \begin{pmatrix} \mathbf{x} \\ \mathbf{0} \end{pmatrix} = \begin{pmatrix} \mathbf{y} \\ \mathbf{0} \end{pmatrix}.$$

4.5 Orthogonale Zerlegung

Die Normalengleichung hat den Vorteil, dass man mit ihr das Lösen eines linearen Ausgleichsproblems auf das Lösen eines quadratischen Gleichungssystems zurückführen kann, das immer lösbar ist. Allerdings gibt es Fälle, in denen die Konditionszahl der Gramschen Matrix $\mathbf{G} = \mathbf{A}^*\mathbf{A}$ erheblich größer als die der Matrix $\mathbf{A}$ ist (vgl. Übungsaufgabe 4.14), so dass es zu numerischen Instabilitäten kommen kann, die wir natürlich gerne vermeiden würden.

Den Schlüssel dazu bieten, wie schon bei linearen Gleichungssystemen, die orthogonalen Faktorisierungen. Nach Satz 3.68 besitzen auch rechteckige Matrizen wie $\mathbf{A}$ eine QR-Zerlegung

$$\mathbf{A} = \mathbf{QR}.$$

Der große Vorteil einer unitären Transformation besteht darin, dass sie Normen unverändert lässt (siehe Lemma 3.61), so dass wir

$$\|\mathbf{Ax} - \mathbf{b}\|_2^2 = \|\mathbf{QRx} - \mathbf{QQ}^*\mathbf{b}\|_2^2 = \|\mathbf{Rx} - \mathbf{Q}^*\mathbf{b}\|_2^2$$

erhalten. Damit ist ein Vektor $\mathbf{x} \in \mathbb{K}^n$ genau dann eine Lösung des linearen Ausgleichsproblems (4.3a), wenn er

$$f(\mathbf{x}) = \|\mathbf{Rx} - \mathbf{Q}^*\mathbf{b}\|_2^2$$

minimiert. Da $\mathbf{R}$ eine obere Dreiecksmatrix ist, erhalten wir

$$\mathbf{R} = \begin{pmatrix} \widehat{\mathbf{R}} \\ \mathbf{0} \end{pmatrix}, \qquad \widehat{\mathbf{R}} := \begin{pmatrix} r_{11} & \dots & r_{1n} \\ & \ddots & \vdots \\ & & r_{nn} \end{pmatrix}.$$

Wir zerlegen $\mathbf{c} := \mathbf{Q}^*\mathbf{b}$ passend in

$$\mathbf{c} = \mathbf{Q}^*\mathbf{b} = \begin{pmatrix} \widehat{\mathbf{b}} \\ \mathbf{b}_0 \end{pmatrix}, \qquad \widehat{\mathbf{b}} = \begin{pmatrix} c_1 \\ \vdots \\ c_n \end{pmatrix}, \qquad \mathbf{b}_0 = \begin{pmatrix} c_{n+1} \\ \vdots \\ c_m \end{pmatrix}$$

und gelangen so zu der Darstellung

$$f(\mathbf{x}) = \|\widehat{\mathbf{R}}\mathbf{x} - \widehat{\mathbf{b}}\|_2^2 + \|\mathbf{b}_0\|_2^2 \qquad \text{für alle } \mathbf{x} \in \mathbb{K}^n.$$

Der zweite Term der rechten Seite ist von $\mathbf{x}$ völlig unabhängig, er spielt also insbesondere für dessen Berechnung auch keine Rolle. Der erste Term dagegen kann verwendet werden, um $\mathbf{x}$ direkt zu bestimmen: Wir gehen davon aus, dass $\mathbf{A}$ injektiv ist. Dann muss auch $\mathbf{R}$ und damit auch $\widehat{\mathbf{R}}$ injektiv sein, also muss die quadratische Matrix $\widehat{\mathbf{R}}$ auch regulär sein. Demzufolge können wir $\mathbf{x}$ als Lösung des linearen Gleichungssystems

$$\widehat{\mathbf{R}}\mathbf{x} = \widehat{\mathbf{b}}$$

bestimmen und so mit

$$f(\mathbf{x}) = \|\mathbf{b}_0\|_2^2$$

```
procedure ls_qr(A, m, n, var b, x);
decomp_qr(m, n, A, r)
trans_qr(m, n, A, r, b)
back_subst(A, n, b)
```

Abbildung 4.3: Lösen eines Ausgleichsproblem mit Hilfe der QR-Zerlegung.

den Fehler minimieren.

Da wir vorausgesetzt haben, dass $\mathbf{A}$ injektiv ist, muss insbesondere $m \geq n$ gelten. In diesem Fall benötigt die Berechnung der QR-Zerlegung gerade

$$2mn^2 + \frac{n}{6}(25 - 4n^2 + 3n)$$

Operationen, die Berechnung von $\mathbf{c} = \mathbf{Q}^*\mathbf{b}$ erfolgt in

$$3n(2m - n + 1) + 2n$$

Operationen, und das Rückwärtseinsetzen erfordert n^2 Operationen, so dass wir insgesamt

$$\begin{aligned} &2mn^2 + \frac{n}{6}(25 - 4n^2 - 3n) + 3n(2m - n + 1) + 2n + n^2 \\ &\quad = mn(2n + 6) + n\left(\frac{25 - 4n^2 - 3n}{6} - 3n + 3 + 2 + n\right) \\ &\quad = mn(2n + 6) + n\left(\frac{25 - 4n^2 - 15n}{6} + 5\right) \leq mn(2n + 6) + 6n \end{aligned}$$

Operationen benötigen. Für große Werte von m wird also das Lösen des Ausgleichsproblems mit Hilfe der QR-Zerlegung ungefähr doppelt soviele Rechenoperationen wie das Lösen mit Hilfe der Normalengleichung benötigen, bietet dafür aber den Vorteil höherer Stabilität.

Übungsaufgabe 4.16 (Orthogonale Projektion) *Sei $\mathcal{V} \subseteq \mathbb{K}^m$ ein beliebiger Teilraum.*

(a) Eine Matrix $\mathbf{V} \in \mathbb{K}^{m\times n}$ nennen wir isometrisch, *falls $\mathbf{V}^*\mathbf{V} = \mathbf{I}$ gilt.*

Sei $\mathbf{V} \in \mathbb{K}^{m\times n}$ isometrisch. Beweisen Sie, dass $\mathbf{P} := \mathbf{V}\mathbf{V}^*$ eine orthogonale Projektion mit $\operatorname{Bild}\mathbf{P} = \operatorname{Bild}\mathbf{V}$ ist.

(b) Sei $k := \dim\mathcal{V}$. Beweisen Sie, dass eine isometrische Matrix $\mathbf{V} \in \mathbb{K}^{m\times k}$ mit $\operatorname{Bild}\mathbf{V} = \mathcal{V}$ existiert.

(c) Beweisen Sie, dass eine orthogonale Projektion $\mathbf{P} \in \mathbb{K}^{m\times m}$ mit $\operatorname{Bild}\mathbf{P} = \mathcal{V}$ existiert.

Hinweis: *Der Basisergänzungssatz kann für Teil (b) hilfreich sein.*

Übungsaufgabe 4.17 (Lösbarkeit) *Sei* $\mathbf{A} \in \mathbb{K}^{m\times n}$*, sei* $\mathcal{R} := \operatorname{Bild} \mathbf{A}$*.*

1. *Sei* $\mathbf{b} \in \mathcal{R}$*. Beweisen Sie*

$$\langle \mathbf{b}, \mathbf{y}\rangle_2 = 0 \qquad \text{für alle } \mathbf{y} \in \operatorname{Kern} \mathbf{A}^*. \tag{4.10}$$

2. *Sei* $\mathbf{b} \in \mathbb{K}^m$ *mit (4.10) gegeben. Beweisen Sie, dass daraus* $\mathbf{b} \in \mathcal{R}$ *folgt, dass also eine Lösung* $\mathbf{x} \in \mathbb{K}^n$ *des linearen Gleichungssystems* $\mathbf{Ax} = \mathbf{b}$ *existiert.*

Hinweis: *Für Teil (b) könnte es sich lohnen, die orthogonale Projektion* $\mathbf{P} \in \mathbb{K}^{m\times m}$ *auf* $\mathcal{R}$ *zu verwenden, die nach Übungsaufgabe 4.16 existiert. Liegt dann* $\mathbf{b} - \mathbf{Pb}$ *in* $\operatorname{Kern} \mathbf{A}^*$*?*

Übungsaufgabe 4.18 (Fehlerfortpflanzung) *Sei* $(\mathbf{Q}, \mathbf{R})$ *eine QR-Zerlegung einer injektiven Matrix* $\mathbf{A} \in \mathbb{K}^{m\times n}$*. Dann gilt*

$$\mathbf{R} = \begin{pmatrix} \widehat{\mathbf{R}} \\ \mathbf{0} \end{pmatrix} \qquad \text{mit} \qquad \widehat{\mathbf{R}} = \begin{pmatrix} r_{11} & \cdots & r_{1n} \\ & \ddots & \vdots \\ & & r_{nn} \end{pmatrix}.$$

(a) Wir definieren

$$\widehat{\mathbf{Q}} := \begin{pmatrix} q_{11} & \cdots & q_{1n} \\ \vdots & \ddots & \vdots \\ q_{m1} & \cdots & q_{mn} \end{pmatrix} \in \mathbb{K}^{m\times n}.$$

Beweisen Sie, dass $\widehat{\mathbf{Q}}^*\widehat{\mathbf{Q}} = \mathbf{I}$ *und* $\mathbf{A} = \widehat{\mathbf{Q}}\widehat{\mathbf{R}}$ *gelten, und folgern Sie daraus, dass* $\mathbf{P} := \widehat{\mathbf{Q}}\widehat{\mathbf{Q}}^*$ *eine orthogonale Projektion auf das Bild der Matrix* $\mathbf{A}$ *ist.*

(b) Sei $\mathbf{b} \in \mathbb{K}^m$*. Beweisen Sie, dass* $\mathbf{x} := \widehat{\mathbf{R}}^{-1}\widehat{\mathbf{Q}}^*\mathbf{b}$ *das korrespondierende lineare Ausgleichsproblem löst, also die folgende Eigenschaft besitzt:*

$$\|\mathbf{Ax} - \mathbf{b}\|_2 \le \|\mathbf{Ay} - \mathbf{b}\|_2 \qquad \text{für alle } \mathbf{y} \in \mathbb{K}^n.$$

(c) Seien $\mathbf{b}, \widetilde{\mathbf{b}} \in \mathbb{K}^m$ *mit* $\widehat{\mathbf{Q}}^*\mathbf{b} \neq \mathbf{0}$ *gegeben. Nach Teil (b) sind* $\mathbf{x} := \widehat{\mathbf{R}}^{-1}\widehat{\mathbf{Q}}^*\mathbf{b}$ *und* $\widetilde{\mathbf{x}} := \widehat{\mathbf{R}}^{-1}\widehat{\mathbf{Q}}^*\widetilde{\mathbf{b}}$ *Lösungen der linearen Ausgleichsprobleme mit den rechten Seiten* $\mathbf{b}$ *und* $\widetilde{\mathbf{b}}$*. Beweisen Sie* $\mathbf{Pb} \neq \mathbf{0}$ *und*

$$\frac{\|\mathbf{x} - \widetilde{\mathbf{x}}\|_2}{\|\mathbf{x}\|_2} \le \kappa_2(\widehat{\mathbf{R}}) \frac{\|\mathbf{P}(\mathbf{b} - \widetilde{\mathbf{b}})\|_2}{\|\mathbf{Pb}\|_2}.$$

(d) Beweisen Sie $\kappa_2(\mathbf{A}) = \kappa_2(\widehat{\mathbf{R}})$*.*

Anmerkung: *Wir haben gesehen, dass wir das lineare Ausgleichsproblem lösen können, ohne die Normalengleichung zu benutzen. Diese Aufgabe zeigt, dass wir auch die Fehlerabschätzung des Lemmas 4.9 beweisen können, ohne auf die Normalengleichung zurückzugreifen.*

4.6 Verallgemeinerung

Bisher haben wir uns mit überbestimmten Systemen beschäftigt: Es gab mehr Gleichungen als Unbekannte, so dass wir lediglich darauf hoffen durften, den Gesamtfehler möglichst gering zu halten.

Nun interessieren wir uns für unterbestimmte Systeme: Sei $\mathbf{A} \in \mathbb{K}^{m\times n}$ surjektiv, aber nicht unbedingt injektiv. Nach Definition der Surjektivität gilt $\operatorname{Bild}\mathbf{A} = \mathbb{K}^m$, also insbesondere $m \leq n$ und $\mathbf{b} \in \operatorname{Bild}\mathbf{A}$ für jede rechte Seite $\mathbf{b} \in \mathbb{K}^m$, somit können wir immer ein $\mathbf{x} \in \mathbb{K}^n$ mit

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

finden. Da $\mathbf{A}$ nicht unbedingt injektiv ist, muss $\mathbf{x}$ allerdings nicht eindeutig bestimmt sein, stattdessen wird es im Allgemeinen eine Menge

$$\mathcal{S} := \{\mathbf{x} \in \mathbb{K}^n \ : \ \mathbf{A}\mathbf{x} = \mathbf{b}\}$$

von unendlich vielen Lösungen geben. Da der Verlust der Eindeutigkeit verschiedene unerwünschte Nebenwirkungen hat, würden wir gerne durch Hinzufügen einer weiteren Bedingung festlegen, welche der Lösungen aus $\mathcal{S}$ die „richtige“ ist.

Besonders einfach ist die Suche nach der *Minimumnormlösung*: Wir verlangen, dass

$$\|\mathbf{x}\|_2 \leq \|\mathbf{y}\|_2 \qquad \text{für alle } \mathbf{y} \in \mathcal{S}$$

gilt, dass die gesuchte Lösung also die minimale Norm unter allen möglichen Lösungen besitzen soll. Da $\mathcal{S}$ ein affiner Teilraum ist (zwei Elemente unterscheiden sich nur durch einen Vektor aus dem Kern der Matrix $\mathbf{A}$), existiert genau ein $\mathbf{x}$ mit der obigen Eigenschaft, wir haben also die Eindeutigkeit wieder hergestellt.

Beispiel 4.19 (Minimumnormlösung) *Als Beispiel für die Vorzüge der Minimumnormlösung untersuchen wir das besonders einfache Gleichungssystem*

$$x_1 - x_2 = 1, \qquad \begin{pmatrix}1 & -1\end{pmatrix}\begin{pmatrix}x_1\\x_2\end{pmatrix} = 1.$$

Offenbar gilt $x_2 = x_1 - 1$, so dass sich die Menge der Lösungen kompakt als

$$\mathcal{S} := \{\mathbf{x} \in \mathbb{R}^2 \ : \ x_2 = x_1 - 1\}$$

schreiben lässt und gerade einen eindimensionalen affinen Teilraum des $\mathbb{R}^2$ beschreibt.

Die Minimumnormlösung lässt sich bestimmen, indem wir

$$\|\mathbf{x}\|_2^2 = x_1^2 + x_2^2 = x_1^2 + (x_1 - 1)^2 = x_1^2 + x_1^2 - 2x_1 + 1 = 2x_1^2 - 2x_1 + 1$$

minimieren. Das Minimum ist als Nullstelle der Ableitung charakterisiert, also durch

$$4x_1 - 2 = 0,$$

so dass wir $x_1 = 1/2$ und $x_2 = -1/2$ erhalten.

Ein Vorzug dieser Lösung gegenüber anderen lässt sich einfach erkennen: Beispielsweise ist auch $\widehat{\mathbf{x}} := (1\,000\,000, 999\,999)$ *eine Lösung des linearen Gleichungssystems, aber wenn wir die Komponenten dieses Vektors etwa durch gerundete vierstellige Maschinenzahlen annähern, erhalten wir* $\widetilde{\mathbf{x}} = (1\,000\,000, 1\,000\,000)$*, und dieser Vektor erfüllt die ursprüngliche Gleichung nicht mehr, weil es zu Auslöschungseffekten kommt. Das Problem entsteht dadurch, dass* $\widehat{\mathbf{x}}$ *einen Anteil aus dem Kern der Systemmatrix enthält, der für die Lösung völlig irrelevant ist, aber trotzdem durch Ziffern der Maschinenzahlen dargestellt werden muss. Die Minimumnormlösung besitzt keinen Anteil aus dem Kern, so dass dieser Effekt nicht auftritt.*

Es stellt sich die Frage, wie sich die Minimumnormlösung $\mathbf{x}$ praktisch berechnen lässt. Wir greifen dazu auf eine unitäre Zerlegung zurück, und zwar auf eine Variante der QR-Zerlegung:

Definition 4.20 (LQ-Zerlegung) *Sei* $\mathbf{A} \in \mathbb{K}^{m\times n}$ *eine Matrix. Falls eine unitäre Matrix* $\mathbf{Q} \in \mathbb{K}^{n\times n}$ *und eine untere Dreiecksmatrix* $\mathbf{L} \in \mathbb{K}^{m\times n}$ *die Gleichung*

$$\mathbf{LQ} = \mathbf{A}$$

erfüllen, bezeichnen wir das Paar $(\mathbf{Q}, \mathbf{L})$ *als eine* LQ-Zerlegung *der Matrix* $\mathbf{A}$*.*

Aus der Existenz der QR-Zerlegung für beliebige Matrizen folgt direkt, dass auch eine LQ-Zerlegung immer existiert:

Lemma 4.21 (Existenz einer LQ-Zerlegung) *Jede Matrix* $\mathbf{A} \in \mathbb{K}^{m\times n}$ *besitzt eine LQ-Zerlegung.*

Beweis. Die Adjungierte $\mathbf{A}^* \in \mathbb{K}^{n\times m}$ besitzt nach Satz 3.68 eine QR-Zerlegung $(\widehat{\mathbf{Q}}, \widehat{\mathbf{R}})$ mit einer unitären Matrix $\widehat{\mathbf{Q}} \in \mathbb{K}^{n\times n}$ und einer oberen Dreiecksmatrix $\widehat{\mathbf{R}} \in \mathbb{K}^{n\times m}$. Dann ist $\mathbf{Q} := \widehat{\mathbf{Q}}^*$ nach Lemma 3.61 ebenfalls unitär, $\mathbf{L} := \widehat{\mathbf{R}}^*$ ist eine untere Dreiecksmatrix, und wegen Lemma 3.45 gilt

$$\mathbf{A} = (\mathbf{A}^*)^* = (\widehat{\mathbf{Q}}\widehat{\mathbf{R}})^* = \widehat{\mathbf{R}}^*\widehat{\mathbf{Q}}^* = \mathbf{LQ},$$

also ist $(\mathbf{Q}, \mathbf{L})$ eine LQ-Zerlegung der Matrix $\mathbf{A}$. ∎

Praktisch berechnen lässt sich die LQ-Zerlegung, indem man Householder-Spiegelungen oder Givens-Rotationen von rechts statt von links mit der Matrix $\mathbf{A}$ multipliziert, also Linearkombinationen von Spalten statt von Zeilen bildet, um die gewünschte Form zu erreichen.

Sei nun also eine LQ-Zerlegung $(\mathbf{Q}, \mathbf{L})$ der Matrix $\mathbf{A}$ gegeben. Durch Einsetzen in die definierende Gleichung erhalten wir

$$\mathbf{b} = \mathbf{Ax} = \mathbf{LQx}.$$

Wir führen die transformierte Größe

$$\mathbf{z} := \mathbf{Qx}$$

ein und beobachten, dass sich die untere Dreiecksmatrix $\mathbf{L}$ in der Form

$$\mathbf{L} = \begin{pmatrix} \widehat{\mathbf{L}} & \mathbf{0} \end{pmatrix}, \qquad \widehat{\mathbf{L}} = \begin{pmatrix} l_{11} & & \\ \vdots & \ddots & \\ l_{m1} & \dots & l_{mm} \end{pmatrix}$$

zerlegen lässt. Wir spalten $\mathbf{z}$ entsprechend in

$$\widehat{\mathbf{z}} = \begin{pmatrix} z_1 \\ \vdots \\ z_m \end{pmatrix}, \qquad \mathbf{z}_0 = \begin{pmatrix} z_{m+1} \\ \vdots \\ z_n \end{pmatrix}$$

und folgern

$$\mathbf{b} = \mathbf{L}\mathbf{z} = \begin{pmatrix} \widehat{\mathbf{L}} & \mathbf{0} \end{pmatrix} \begin{pmatrix} \widehat{\mathbf{z}} \\ \mathbf{z}_0 \end{pmatrix} = \widehat{\mathbf{L}}\widehat{\mathbf{z}}.$$

Da $\mathbf{A}$ surjektiv ist, muss auch $\mathbf{L}$ surjektiv sein, und das kann nur eintreten, wenn $\widehat{\mathbf{L}}$ surjektiv ist. Da $\widehat{\mathbf{L}}$ eine quadratische Matrix ist, muss sie deshalb regulär sein, also ist durch die Gleichung $\mathbf{b} = \widehat{\mathbf{L}}\widehat{\mathbf{z}}$ bereits der Vektor $\widehat{\mathbf{z}}$ eindeutig definiert.

Der Vektor $\mathbf{z}_0$ hat offenbar keinerlei Einfluss auf die Gleichung, kann also auf den ersten Blick beliebig gewählt werden. Da wir allerdings die Minimumnormlösung suchen, sind wir auch daran interessiert

$$\|\mathbf{x}\|_2^2 = \|\mathbf{Q}\mathbf{x}\|_2^2 = \|\mathbf{z}\|_2^2 = \|\widehat{\mathbf{z}}\|_2^2 + \|\mathbf{z}_0\|_2^2$$

zu minimieren. $\widehat{\mathbf{z}}$ steht bereits fest, also müssen wir nur noch dafür sorgen, dass $\|\mathbf{z}_0\|_2^2$ möglichst klein wird. Offenbar ist hier das Minimum durch $\mathbf{z}_0 = \mathbf{0}$ gegeben.

Damit haben wir einen Lösungsalgorithmus für das unterbestimmte System gefunden: Zunächst berechnen wir die LQ-Zerlegung der Matrix $\mathbf{A}$, dann lösen wir $\widehat{\mathbf{L}}\widehat{\mathbf{z}} = \mathbf{b}$ durch Vorwärtseinsetzen, und dann konstruieren wir $\mathbf{x} = \mathbf{Q}^*(\widehat{\mathbf{z}}, \mathbf{0})$. Der Algorithmus entspricht im Wesentlichen dem für das überbestimmte Problem, nur dass diesmal eine LQ- statt einer QR-Zerlegung verwendet und die Transformation $\mathbf{Q}^*$ auf die Hilfslösung $\mathbf{z}$ statt auf die rechte Seite $\mathbf{b}$ angewendet wird. Der Rechenaufwand ist deshalb praktisch identisch, wobei die Parameter n und m allerdings ihre Rollen tauschen.

Bei gut konditionierten Probleme wäre es erstrebenswert, auch ein Gegenstück der Normalengleichung für den unterbestimmten Fall zur Verfügung zu haben, da das Auflösen der Normalengleichung weniger Rechenoperationen als der Zugang über die unitäre Zerlegung benötigt. Wir haben bereits gesehen, dass die Minimumnormlösung die Gleichung

$$\mathbf{x} = \mathbf{Q}^* \begin{pmatrix} \widehat{\mathbf{z}} \\ \mathbf{0} \end{pmatrix}$$

mit einem $\widehat{\mathbf{z}} \in \mathbb{K}^m$ erfüllen muss. Da $\widehat{\mathbf{L}}$ regulär ist, ist nach Lemma 3.45 auch $\widehat{\mathbf{L}}^*$ regulär, also existiert ein $\widehat{\mathbf{x}} \in \mathbb{K}^m$ mit

$$\widehat{\mathbf{z}} = \widehat{\mathbf{L}}^*\widehat{\mathbf{x}}.$$

Indem wir diese Gleichung in die Gleichung für $\mathbf{x}$ einsetzen, erhalten wir

$$\mathbf{x} = \mathbf{Q}^* \begin{pmatrix} \widehat{\mathbf{z}} \\ \mathbf{0} \end{pmatrix} = \mathbf{Q}^* \begin{pmatrix} \widehat{\mathbf{L}}^* \widehat{\mathbf{x}} \\ \mathbf{0} \end{pmatrix} = \mathbf{Q}^* \begin{pmatrix} \widehat{\mathbf{L}}^* \\ \mathbf{0} \end{pmatrix} \widehat{\mathbf{x}} = \mathbf{Q}^* \mathbf{L}^* \widehat{\mathbf{x}} = \mathbf{A}^* \widehat{\mathbf{x}},$$

die Minimumnormlösung muss also im Bild der adjungierten Matrix $\mathbf{A}^*$ liegen. Wir setzen in die zu lösende Gleichung ein und erhalten

$$\mathbf{A}\mathbf{A}^* \widehat{\mathbf{x}} = \mathbf{b}. \tag{4.11}$$

Diese Gleichung ähnelt der Normalengleichung (4.3b) und besitzt ähnliche Eigenschaften:

Satz 4.22 (Lösbarkeit) *Sei* $\mathbf{A}$ *surjektiv. Dann ist* $\mathbf{A}\mathbf{A}^*$ *selbstadjungiert und positiv definit, also insbesondere invertierbar. Es existiert genau eine Minimumnormlösung* $\mathbf{x} \in \mathcal{S}$*, die durch*

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

gegeben ist.

Beweis. Offenbar ist $\mathbf{A}\mathbf{A}^*$ selbstadjungiert. Um nachzuweisen, dass die Matrix auch positiv definit ist, wählen wir einen Vektor $\mathbf{x} \in \mathbb{K}^m$ und berechnen

$$\langle \mathbf{A}\mathbf{A}^*\mathbf{x}, \mathbf{x} \rangle_2 = \langle \mathbf{A}^*\mathbf{x}, \mathbf{A}^*\mathbf{x} \rangle_2 = \|\mathbf{A}^*\mathbf{x}\|_2^2 \geq 0.$$

Nehmen wir nun $\mathbf{A}^*\mathbf{x} = \mathbf{0}$ an. Da $\mathbf{A}$ surjektiv ist, gilt $\mathbf{x} \in \mathbb{K}^m = \operatorname{Bild} \mathbf{A}$, also finden wir einen Vektor $\mathbf{y} \in \mathbb{K}^n$ mit $\mathbf{x} = \mathbf{A}\mathbf{y}$. Daraus folgt

$$0 = \langle \mathbf{y}, \mathbf{A}^*\mathbf{x} \rangle_2 = \langle \mathbf{A}\mathbf{y}, \mathbf{x} \rangle_2 = \langle \mathbf{x}, \mathbf{x} \rangle_2 = \|\mathbf{x}\|_2^2,$$

also insbesondere $\mathbf{x} = \mathbf{0}$. Damit ist der Nullvektor das einzige Element des Kerns der Matrix $\mathbf{A}^*$, also ist diese Matrix injektiv und damit $\mathbf{A}\mathbf{A}^*$ positiv definit.

Wir haben bereits gesehen, dass die Minimumnormlösung $\mathbf{x}$ im Bild der Matrix $\mathbf{A}^*$ liegen muss, und jetzt wissen wir auch, dass es nur genau ein Urbild $\widehat{\mathbf{x}}$ mit $\mathbf{x} = \mathbf{A}^*\widehat{\mathbf{x}}$ geben kann, das durch

$$\mathbf{A}\mathbf{A}^* \widehat{\mathbf{x}} = \mathbf{b}$$

definiert ist. Da $\mathbf{A}\mathbf{A}^*$ invertierbar ist, gelten

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

und das ist die gesuchte Darstellung der Lösung. ∎

Diese Darstellung der Lösung lässt sich, ähnlich wie im Fall der Normalengleichung, mit Hilfe einer Cholesky-Zerlegung der Matrix $\mathbf{A}\mathbf{A}^*$ effizient berechnen, allerdings tritt auch hier das Problem auf, dass bei einem schlecht konditionierten Problem eventuelle Eingabe- und Rundungsfehler ungünstiger als bei dem Ansatz auf Grundlage der LQ-Zerlegung verstärkt werden.

Übungsaufgabe 4.23 (Moore-Penrose-Pseudoinverse) *Sei $\mathbf{A} \in \mathbb{K}^{m\times n}$. Eine Matrix $\mathbf{A}^+ \in \mathbb{K}^{n\times m}$ heißt* Moore-Penrose-Pseudoinverse *der Matrix $\mathbf{A}$, falls sie die Gleichungen*

$$\mathbf{A}\mathbf{A}^+\mathbf{A} = \mathbf{A}, \tag{4.12a}$$
$$\mathbf{A}^+\mathbf{A}\mathbf{A}^+ = \mathbf{A}^+, \tag{4.12b}$$
$$(\mathbf{A}^+\mathbf{A})^* = \mathbf{A}^+\mathbf{A}, \tag{4.12c}$$
$$(\mathbf{A}\mathbf{A}^+)^* = \mathbf{A}\mathbf{A}^+ \tag{4.12d}$$

erfüllt. Sei $\mathbf{b} \in \mathbb{K}^m$.

(a) Beweisen Sie, dass aus (4.12a) und (4.12c) folgt, dass

$$\langle \mathbf{A}\mathbf{A}^+\mathbf{b} - \mathbf{b}, \mathbf{A}\mathbf{z}\rangle_2 = 0 \qquad \text{für alle } \mathbf{z} \in \mathbb{K}^n.$$

(b) Beweisen Sie, dass $\mathbf{x} := \mathbf{A}^+\mathbf{b}$ die folgende Minimalitätsbedingung erfüllt:

$$\|\mathbf{A}\mathbf{x} - \mathbf{b}\|_2 \leq \|\mathbf{A}\mathbf{y} - \mathbf{b}\|_2 \qquad \text{für alle } \mathbf{y} \in \mathbb{K}^n. \tag{4.13}$$

(c) Beweisen Sie, dass aus (4.12b) und (4.12d) folgt, dass

$$\langle \mathbf{A}^+\mathbf{b}, \mathbf{y}\rangle_2 = 0 \qquad \text{für alle } \mathbf{y} \in \operatorname{Kern} \mathbf{A}.$$

(d) Beweisen Sie, dass $\mathbf{x}^+ := \mathbf{A}^+\mathbf{b}$ die Minimumnormlösung ist, dass also

$$\|\mathbf{x}^+\|_2 \leq \|\mathbf{x}\|_2 \qquad \text{für alle } \mathbf{x} \in \mathbb{K}^n \text{ mit (4.13).}$$

Hinweis: *Lemma 3.44 und Satz 4.3 können hilfreich sein.*

Zusatzfragen: *Ist $\mathbf{A}^+$ durch (4.12) bereits eindeutig festgelegt? Kann man auch ohne Satz 4.3, also direkt mit (4.13), kurz beweisen, dass zwei Lösungen des Ausgleichsproblems sich nur durch einen Vektor aus dem Kern der Matrix $\mathbf{A}$ unterscheiden?*

Übungsaufgabe 4.24 (Existenz der Pseudoinversen) *Sei $\mathbf{A} \in \mathbb{K}^{m\times n}$. Wie in Folgerung 4.5 definieren wir*

$$\mathcal{R} := \{\mathbf{A}^*\mathbf{z} \ : \ \mathbf{z} \in \mathbb{K}^m\} \subseteq \mathbb{K}^n, \qquad \mathcal{N} := \{\mathbf{y} \in \mathbb{K}^n \ : \ \mathbf{A}\mathbf{y} = \mathbf{0}\} \subseteq \mathbb{K}^n.$$

Wir wissen bereits, dass

$$L\colon \mathcal{R} \to \mathcal{R}, \qquad \mathbf{x} \mapsto \mathbf{A}^*\mathbf{A}\mathbf{x},$$

eine bijektive lineare Abbildung des Teilraums $\mathcal{R}$ in sich selbst ist.

Beweisen Sie, dass die Abbildung $\mathbf{A}^+ := L^{-1}\mathbf{A}^*$ die Bedingungen (4.12) erfüllt.

Hinweis: *Ohne Beweis darf verwendet werden, dass $\mathbb{K}^n = \mathcal{R} \oplus \mathcal{N}$ gilt. Der zugehörige Beweis ist in Übungsaufgabe 4.26 skizziert.*

Übungsaufgabe 4.25 (Singulärwertzerlegung) *Sei* $\mathbf{A} \in \mathbb{K}^{m\times n}$.

(a) Beweisen Sie, dass $\mathbf{u} \in \mathbb{K}^m$ *und* $\mathbf{v} \in \mathbb{K}^n$ *existieren mit* $\|\mathbf{u}\|_2 = 1$, $\|\mathbf{v}\|_2 = 1$ *und*

$$|\langle \mathbf{u}, \mathbf{A}\mathbf{v}\rangle_2| \geq |\langle \mathbf{y}, \mathbf{A}\mathbf{x}\rangle_2| \quad \text{für alle } \mathbf{x} \in \mathbb{K}^n,\ \mathbf{y} \in \mathbb{K}^m \text{ mit } \|\mathbf{x}\|_2 = 1,\ \|\mathbf{y}\|_2 = 1.$$

(b) Beweisen Sie, dass ein $\sigma \in \mathbb{K}$ *so existiert, dass die Vektoren aus Teil (a) die Gleichungen* $\sigma\mathbf{u} = \mathbf{A}\mathbf{v}$ *und* $\bar{\sigma}\mathbf{v} = \mathbf{A}^*\mathbf{u}$ *erfüllen.*

(c) Beweisen Sie, dass unitäre Matrizen $\mathbf{U} \in \mathbb{K}^{m\times m}$ *und* $\mathbf{V} \in \mathbb{K}^{n\times n}$ *so existieren, dass* $\mathbf{U}^*\mathbf{A}\mathbf{V}$ *eine Diagonalmatrix ist.*

Hinweis: *Bei Teil (a) empfiehlt sich der Satz von Heine-Borel, bei Teil (b) die Cauchy-Schwarz-Ungleichung, und bei Teil (c) können Householder-Spiegelungen helfen.*

Übungsaufgabe 4.26 (Direkte Summe) *Sei* $\mathbf{A} \in \mathbb{K}^{m\times n}$. *Sei* $\mathbf{P} \in \mathbb{K}^{n\times n}$ *die orthogonale Projektion auf* $\mathrm{Bild}(\mathbf{A}^*)$.

(a) Beweisen Sie

$$\langle \mathbf{x} - \mathbf{P}\mathbf{x}, \mathbf{y}\rangle = 0 \qquad \text{für alle } \mathbf{x} \in \mathbb{K}^n,\ \mathbf{y} \in \mathrm{Bild}(\mathbf{A}^*).$$

(b) Folgern Sie daraus

$$\mathbf{x} - \mathbf{P}\mathbf{x} \in \mathrm{Kern}(\mathbf{A}) \qquad \text{für alle } \mathbf{x} \in \mathbb{K}^n.$$

Damit folgt $\mathbb{K}^n = \mathrm{Bild}(\mathbf{A}^*) + \mathrm{Kern}(\mathbf{A})$.

(c) Beweisen Sie

$$\langle \mathbf{x}, \mathbf{y}\rangle_2 = 0 \qquad \text{für alle } \mathbf{x} \in \mathrm{Bild}(\mathbf{A}^*),\ \mathbf{y} \in \mathrm{Kern}(\mathbf{A}).$$

Zusammen mit (b) folgt $\mathbb{K}^n = \mathrm{Bild}(\mathbf{A}^*) \oplus \mathrm{Kern}(\mathbf{A})$.

(d) Sei $\mathbf{x} \in \mathbb{K}^n$ *gegeben mit*

$$\langle \mathbf{x}, \mathbf{y}\rangle_2 = 0 \qquad \text{für alle } \mathbf{y} \in \mathrm{Kern}(\mathbf{A}).$$

Beweisen Sie $\mathbf{x} \in \mathrm{Bild}(\mathbf{A}^*)$.

Also muss die Minimumnormlösung immer im Bild der Matrix $\mathbf{A}^*$ *liegen.*

5 Nichtlineare Gleichungssysteme

In der Praxis treten neben linearen Gleichungssystemen, die wir bereits in Kapitel 3 behandelt haben, häufig auch nichtlineare Gleichungssysteme auf. Analog zu den linearen Gleichungssystemen können wir sie in der Form

$$\begin{aligned} f_1(x_1, \dots, x_n) &= b_1, \\ f_2(x_1, \dots, x_n) &= b_2, \\ &\vdots \\ f_n(x_1, \dots, x_n) &= b_n \end{aligned}$$

darstellen, wobei $\mathbf{b} \in \mathbb{K}^n$ und Funktionen $f_1, \dots, f_n : \mathbb{K}^n \to \mathbb{K}$ gegeben und ein Vektor $\mathbf{x} \in \mathbb{K}^n$ gesucht sind.

Zur Vereinfachung der Notation fassen wir die Funktionen und die rechte Seite $\mathbf{b}$ zu einer vektorwertigen Funktion $f : \mathbb{K}^n \to \mathbb{K}^n$ zusammen, die durch

$$f(\mathbf{x}) = \begin{pmatrix} f_1(x_1, \dots, x_n) - b_1 \\ \vdots \\ f_n(x_1, \dots, x_n) - b_n \end{pmatrix}, \qquad \text{für alle } \mathbf{x} \in \mathbb{K}^n$$

gegeben ist. Damit nimmt das nichtlineare Gleichungssystem die Form eines Nullstellenproblems an: Die Aufgabe, mit der wir uns in diesem Kapitel beschäftigen werden, lautet wie folgt:

> Gegeben sei eine Funktion $f : \mathbb{K}^n \to \mathbb{K}^n$. Finde einen Vektor $\mathbf{x} \in \mathbb{K}^n$ so, dass $f(\mathbf{x}) = \mathbf{0}$ gilt.

Derartige Gleichungssysteme sind *erheblich* schwieriger zu lösen als lineare Gleichungssysteme. Der Satz von Abel-Ruffini beispielsweise besagt, dass allgemeine Polynomgleichungen fünften oder höheren Grades Lösungen besitzen können, die nicht durch Grundrechenarten und Wurzeln dargestellt, also mit einer endlichen Anzahl von Operationen berechnet werden können.

5.1 Kondition

Bevor wir Verfahren zur Berechnung von Nullstellen diskutieren sollten wir zunächst einen Blick auf die zu erwartende Kondition werfen, um einschätzen zu können, wie eine Nullstelle auf Störungen der Eingabedaten reagiert.

Wir beschränken uns bei der Diskussion auf den eindimensionalen Fall: Sei $f : \mathbb{K} \to \mathbb{K}$ eine Funktion, die in $x \in \mathbb{K}$ eine m-fache Nullstelle besitzt, wobei $m \in \mathbb{N}$ gilt und f in x mindestens m-mal differenzierbar sein soll.

Sei $\tilde{f} : \mathbb{K} \to \mathbb{K}$ eine zweite Funktion, die in $\tilde{x} \in \mathbb{K}$ eine Nullstelle besitzt.

Erinnerung 5.1 (Taylor) *Sei $f \in C^m[a,b]$, und seien $x, y \in [a,b]$ gegeben. Dann kann mit dem Hauptsatz der Integral- und Differentialrechnung und partieller Integration die Gleichung*

$$f(y) = \sum_{\nu=0}^{m-1} \frac{(y-x)^\nu}{\nu!} f^{(\nu)}(x) + (y-x)^m \int_0^1 \frac{(1-t)^{m-1}}{(m-1)!} f^{(m)}\big(x + t(y-x)\big)\, dt \tag{5.1a}$$

bewiesen werden. Mit dem Mittelwertsatz der Integralrechnung finden wir davon ausgehend ein $\eta \in [a,b]$ (genauer gesagt zwischen x und y) mit

$$f(y) = \sum_{\nu=0}^{m-1} \frac{(y-x)^\nu}{\nu!} f^{(\nu)}(x) + \frac{(y-x)^m}{m!} f^{(m)}(\eta). \tag{5.1b}$$

Wir analysieren die Fehlerfortpflanzung, indem wir den Fehler $\tilde{x} - x$ mit Hilfe der Taylor-Entwicklung aus den Funktionen f und $\tilde{f}$ herausziehen: Es gilt

$$f(\tilde{x}) = f(x) + (\tilde{x}-x)f'(x) + \ldots + \frac{(\tilde{x}-x)^{m-1}}{(m-1)!} f^{(m-1)}(x) + \frac{(\tilde{x}-x)^m}{m!} f^{(m)}(\eta)$$

für ein $\eta \in [x, \tilde{x}]$. Da x eine m-fache Nullstelle der Funktion f ist, verschwinden die ersten m Summanden, es bleibt nur

$$f(\tilde{x}) = \frac{(\tilde{x}-x)^m}{m!} f^{(m)}(\eta).$$

Da $\tilde{x}$ eine Nullstelle der Funktion $\tilde{f}$ ist, können wir auch

$$f(\tilde{x}) - \tilde{f}(\tilde{x}) = \frac{(\tilde{x}-x)^m}{m!} f^{(m)}(\eta)$$

schreiben und die linke Seite durch die Maximumnorm des Fehlers abschätzen, um

$$\|f - \tilde{f}\|_\infty \geq \frac{|\tilde{x}-x|^m}{m!} |f^{(m)}(\eta)|$$

und daraus auch

$$|\tilde{x}-x|^m \leq \frac{\|f-\tilde{f}\|_\infty m!}{|f^{(m)}(\eta)|},$$

$$|\tilde{x}-x| \leq \left(\frac{\|f-\tilde{f}\|_\infty m!}{|f^{(m)}(\eta)|}\right)^{1/m}$$

zu erhalten. Die rechte Seite dieser Abschätzung wird um so schlechter, je höher m ist: Bei einer einfachen Nullstelle führt eine durch ein $\epsilon \in \mathbb{R}_{>0}$ gestörte Funktion f nur zu einer zu ϵ proportionalen Störung der Lösung, bei einer doppelten Nullstelle ist die Störung bereits zu $\sqrt{\epsilon}$ proportional, bei einer dreifachen Nullstelle zu $\sqrt[3]{\epsilon}$.

Auf ein gut konditioniertes Problem dürfen wir also nur hoffen, wenn die Nullstelle einfach und die erste Ableitung „möglichst unterschiedlich“ von null ist.

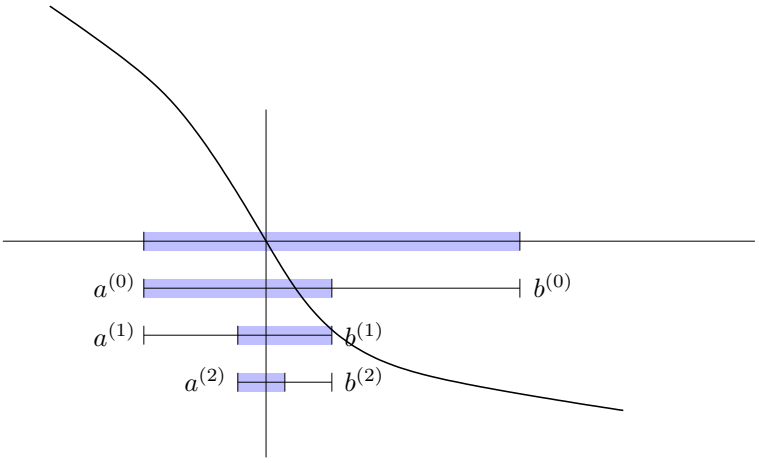

Abbildung 5.1: Geometrische Interpretation des Bisektionsverfahrens

5.2 Bisektionsverfahren

Auch das erste von uns untersuchte Verfahren arbeitet nur im eindimensionalen Fall, sogar nur im Reellen: Eine stetige Funktion $f : \mathbb{R} \to \mathbb{R}$ ist gegeben, und wir suchen nach einem $x \in \mathbb{R}$ mit $f(x) = 0$. Um uns dieser Aufgabe überhaupt nähern zu können, benötigen wir zusätzliche Informationen über die Eigenschaften der Funktion f.

Erinnerung 5.2 (Zwischenwertsatz) *Sei $f \in C[a, b]$, und sei $y \in [f(a), f(b)]$. Dann existiert ein $x \in [a, b]$ mit $f(x) = y$.*

Ein besonders einfaches und zuverlässiges Verfahren beruht auf dem Zwischenwertsatz für stetige Funktionen: Falls falls $a, b \in \mathbb{R}$ mit $a < b$ so gegeben sind, dass

$$f(a)f(b) < 0$$

gilt, dass also $f(a)$ und $f(b)$ entgegengesetzte Vorzeichen besitzen, muss nach diesem Satz ein $x \in (a, b)$ mit $f(x) = 0$ existieren.

Die Idee des *Bisektionsverfahrens* besteht darin, das Intervall $[a, b]$, in dem eine Nullstelle liegen muss, in zwei Teilintervalle zu zerlegen. In einem dieser Teilintervalle muss dann ebenfalls eine Nullstelle liegen, so dass wir mit ihm die Suche fortsetzen können.

Um die Durchmesser der Teilintervalle so klein wie möglich zu machen, bietet es sich an, das ursprüngliche Intervall in der Mitte zu teilen. Wir setzen also

$$c := \frac{b + a}{2}$$

und prüfen, ob $f(a)f(c) < 0$ oder $f(c)f(b) < 0$ gelten. Im ersten Fall setzen wir die Suche im Intervall $[a, c]$ fort, im zweiten in $[c, b]$.

Der entstehende Algorithmus ist in Abbildung 5.2 zusammengefasst und in Abbildung 5.1 illustriert.

```
procedure bisection(f, var a, b);
f_a ← f(a);   f_b ← f(b)
repeat
  c ← (b + a)/2
  f_c ← f(c)
  if f_a f_c < 0 then
    b ← c;   f_b ← f_c
  else if f_c f_b < 0 then
    a ← c;   f_a ← f_c
  end if
until b − a < ϵ oder f_a f_b ≥ 0
```

Abbildung 5.2: Bisektionsverfahren zur Bestimmung einer Nullstelle einer stetigen Funktion f. Ausgehend von einem Intervall $[a, b]$ mit $f(a)f(b) < 0$ findet der Algorithmus ein Intervall der Länge $< \epsilon$, das eine Nullstelle enthält.

Der Bisektionsalgorithmus ist auf den ersten Blick sehr elegant: Er funktioniert bereits für eine stetige Funktion, und er berechnet in jedem Schritt Unter- und Obergrenzen der Lösung, so dass sich einfach feststellen lässt, wann die Nullstelle hinreichend genau angenähert wurde. Außerdem ist garantiert, dass sich die Länge des Intervalls in jedem Schritt halbiert, dass also die Nullstelle immer enger eingeschlossen wird.

Auf den zweiten Blick zeigt sich der wesentliche Nachteil dieses Ansatzes: Er funktioniert nur, falls uns geeignete Startwerte zur Verfügung stehen, falls wir also ein Intervall finden, in dem eine ungerade Anzahl von Nullstellen enthalten sind, denn bei einer geraden Anzahl würde in beiden Endpunkten dasselbe Vorzeichen auftreten und die Voraussetzungen des Verfahrens wären verletzt.

In gewisser Weise ist das Bisektionsverfahren typisch für Algorithmen zur Behandlung nichtlinearer Probleme:

- Es arbeitet *iterativ*, berechnet also in jedem Schritt eine neue Näherungslösung aus einer vorangehenden,
- es kann eine beliebig genaue Näherungslösung finden, aber in der Regel niemals die exakte Lösung, weil dafür unendlich viele Schritte erforderlich wären, und
- die erste Näherungslösung, von der das Verfahren ausgeht, muss geeignete Eigenschaften aufweisen, die wir sicherstellen müssen.

Die ersten beiden Punkte sind in der Regel eher Vor- als Nachteile, weil wir im Computer ohnehin nur mit näherungsweisen Zahldarstellungen arbeiten und durch die geeignete Wahl des Abbruchkriteriums dafür sorgen können, dass die berechnete Näherung gerade gut genug ist, so dass kein unnötiger Rechenaufwand anfällt. Der dritte Punkt dagegen ist kritisch, denn die Bereitstellung guter Ausgangswerte erweist sich in der Praxis häufig als schwierig.

5.3 Iterationsverfahren

Bei linearen Gleichungssystemen haben wir Verfahren (wie die pivotisierte LR-Zerlegung oder die QR-Zerlegung) kennen gelernt, die für eine beliebige reguläre Matrix eine Lösung berechnen. Diese Verfahren benötigen lediglich die Grundrechenarten Addition, Subtraktion, Multiplikation und Division sowie im Fall der QR-Zerlegung die Berechnung von Quadratwurzeln.

Einen ähnlichen Grad von Allgemeinheit können wir bei nichtlinearen Gleichungssystemen nicht erreichen: Wie bereits erwähnt besagt der Satz von Abel-Ruffini, dass es Polynome fünften Grades gibt, deren Nullstellen sich nicht durch Grundrechenarten und Wurzeln ausdrücken lassen, wir werden also auch keinen Algorithmus konstruieren können, der nur mit Grundrechenarten und Wurzeln alle Nullstellen bestimmt.

Allerdings wissen wir, dass wir infolge der Approximation von Zahlen durch Gleitkommadarstellungen ohnehin nicht auf eine exakte Darstellung von Nullstellen hoffen können, der Computer wird uns in der Regel nur eine Näherung der Nullstelle geben.

Also reicht es uns auch, ein Verfahren zu finden, das eine Näherung der Lösung eines nichtlinearen Gleichungssystems bestimmt. Natürlich wollen wir die Genauigkeit der Näherung kontrollieren können: Bei der Auswertung der Daten eines Computertomographen wird eine höhere Genauigkeit als bei der Berechnung einer dreidimensionalen Grafik für ein Computerspiel gefordert werden.

Einen guten Ansatz zur Lösung dieser Aufgabe bieten *Iterationsverfahren*: Aus der Funktion f konstruieren wir eine *Iterationsfunktion* $\Phi : \mathbb{K}^n \to \mathbb{K}^n$, die aus einer gegebenen Näherungslösung $\mathbf{x}^{(0)} \in \mathbb{K}^n$ eine bessere Näherungslösung

$$\mathbf{x}^{(1)} := \Phi(\mathbf{x}^{(0)})$$

berechnet. Durch wiederholte Anwendung der Iterationsfunktion erhalten wir eine Folge $(\mathbf{x}^{(m)})_{m=0}^{\infty}$ von Näherungslösungen, die durch die Vorschrift

$$\mathbf{x}^{(m+1)} := \Phi(\mathbf{x}^{(m)}) \qquad \text{für alle } m \in \mathbb{N}_0 \tag{5.2}$$

definiert ist und für $m \to \infty$ gegen eine exakte Lösung $\mathbf{x}^* \in \mathbb{K}^n$ konvergiert.

Da die Iterationsfunktion die Lösung verbessern soll, dürfen wir zumindest $\Phi(\mathbf{x}^*) = \mathbf{x}^*$ erwarten, exakte Lösungen sollten also *Fixpunkte* der Iterationsfunktion sein. Für die Untersuchung dieser Fixpunkte ist der Mittelwertsatz oft sehr hilfreich.

Erinnerung 5.3 (Mittelwertsatz) *Seien $f, g \in C[a,b]$ mit $g \geq 0$ gegeben. Dann existiert ein $\eta \in [a,b]$ mit*

$$\int_a^b f(x)g(x)\,dx = f(\eta)\int_a^b g(x)\,dx. \tag{5.3a}$$

In Kombination mit dem Hauptsatz der Integral- und Differentialrechnung folgt, dass für alle $f \in C^1[a,b]$ ein $\eta \in [a,b]$ existiert mit

$$f'(\eta) = \frac{f(b) - f(a)}{b - a}. \tag{5.3b}$$

Beispiel 5.4 *Wir suchen eine Nullstelle des Polynoms*

$$p(x) = x^6 - x - 1$$

im Intervall $[1, 2]$. *Aus* $p(1) = -1$ *und* $p(2) = 61$ *folgt mit dem Zwischenwertsatz (vgl. Erinnerung 5.2), dass eine Nullstelle* $x^* \in [1, 2]$ *existieren muss* [1].

Die Nullstelle x^* *sollte ein Fixpunkt jeder brauchbaren Iterationsfunktion sein, es sollte also* $\Phi(x^*) = x^*$ *gelten. Demnach können wir Kandidaten für Iterationsfunktionen konstruieren, indem wir die Gleichung* $p(x) = 0$ *als Fixpunktgleichung schreiben. Eine erste Möglichkeit besteht darin,*

$$p(x) = 0 \quad \Longleftrightarrow \quad x = x^6 - 1, \qquad x = \Phi_1(x) := x^6 - 1$$

zu verwenden. Eine zweite Möglichkeit ist

$$p(x) = 0 \quad \Longleftrightarrow \quad x^6 = x + 1 \quad \Longleftrightarrow \quad x = \sqrt[6]{x+1}, \qquad x = \Phi_2(x) := \sqrt[6]{x+1}.$$

Wir untersuchen zunächst Φ_1. *Sei* $x^{(0)} \in [1, 2]$ *eine Näherung von* x^*, *und sei* $x^{(1)} := \Phi_1(x^{(0)})$ *die von* Φ_1 *berechnete neue Näherung. Aus dem Mittelwertsatz der Differentialrechnung folgt*

$$|x^{(1)} - x^*| = |\Phi_1(x^{(0)}) - \Phi_1(x^*)| = |\Phi_1'(\eta)|\ |x^{(0)} - x^*|$$

für ein $\eta \in [x^{(0)}, x^*]$. *Leider gilt* $\Phi_1'(\eta) = 6\eta^5 \geq 6$ *für alle* $\eta \in \mathbb{R}_{\geq 1}$, *also folgt*

$$|x^{(1)} - x^*| \geq 6\ |x^{(0)} - x^*|,$$

die neue Näherung wird also mindestens sechsmal weiter von der richtigen Lösung entfernt sein als die alte. Diese Iterationsvorschrift ist offenbar ungeeignet.

Wenden wir uns nun Φ_2 *zu. Auch in diesem Fall erhalten wir für* $x^{(1)} := \Phi_2(x^{(0)})$ *die Gleichung*

$$|x^{(1)} - x^*| = |\Phi_2'(\eta)|\ |x^{(0)} - x^*|$$

für ein $\eta \in [x^{(0)}, x^*]$, *aber dank*

$$\Phi_2'(\eta) = \frac{(\eta+1)^{-5/6}}{6} = \frac{1}{6(\eta+1)^{5/6}} \leq \frac{1}{6\,2^{5/6}} < \frac{1}{6}$$

für alle $\eta \in \mathbb{R}_{\geq 1}$ *erhalten wir diesmal*

$$|x^{(1)} - x^*| \leq \frac{1}{6}|x^{(0)} - x^*|,$$

die neue Näherung wird also mindestens sechsmal näher an der richtigen Lösung liegen als die alte. Diese Iterationsvorschrift ist offenbar empfehlenswert.

[1] Dieses Beispiel habe ich dem Buch „Numerik für Ingenieure und Naturwissenschaftler“ von W. Dahmen und A. Reusken entnommen.

Eine Iterationsfunktion ist also um so besser geeignet, je kleiner ihre Ableitung in der Nähe der gesuchten Nullstelle ist. Im Allgemeinen muss Φ nicht einmal differenzierbar sein, sondern es reicht bereits, wenn die Iterationsfunktion Lipschitz-stetig mit einer Kostanten kleiner als eins ist. Für derartige *Kontraktionen* lässt sich beweisen, dass die Folgenglieder immer näher aneinander rücken. Um daraus zu folgern, dass die Folge auch konvergiert, benötigen wir den Begriff der *Cauchy-Folge*.

Erinnerung 5.5 (Cauchy-Folgen) *Eine Folge $(\mathbf{x}^{(m)})_{m=0}^{\infty}$ wird als* Cauchy-Folge *bezeichnet, falls für jedes $\epsilon \in \mathbb{R}_{>0}$ ein $m_0 \in \mathbb{N}$ existiert mit*

$$\|\mathbf{x}^{(m+n)} - \mathbf{x}^{(m)}\| \leq \epsilon \qquad \text{für alle } m, n \in \mathbb{N}_0 \text{ mit } m \geq m_0.$$

Falls $(\mathbf{x}^{(m)})_{m=0}^{\infty}$ eine Cauchy-Folge ist, deren Folgenglieder in einer abgeschlossenen Menge $U \subseteq \mathbb{K}^n$ liegen, konvergiert die Folge gegen einen Grenzwert $\mathbf{x}^* \in U$.

Mit Hilfe der Cauchy-Folgen kann nun der *Fixpunktsatz von Banach* bewiesen werden, den wir hier in einer Version vorstellen, die explizite Aussagen über die Qualität der Näherungslösungen zulässt.

Satz 5.6 (Banach) *Sei $U \subseteq \mathbb{K}^n$ eine abgeschlossene Menge, für die $\Phi(U) \subseteq U$ gilt, die also durch Φ in sich selbst abgebildet wird. Sei $L \in [0, 1)$ so gegeben, dass*

$$\|\Phi(\mathbf{x}) - \Phi(\mathbf{y})\| \leq L\|\mathbf{x} - \mathbf{y}\| \qquad \text{für alle } \mathbf{x}, \mathbf{y} \in U \tag{5.4}$$

gilt. Dann besitzt Φ genau einen Fixpunkt $\mathbf{x}^* \in U$ und die durch (5.2) definierte Folge $(\mathbf{x}^{(m)})_{m=0}^{\infty}$ konvergiert gegen ihn. Es gelten sogar

$$\|\mathbf{x}^{(m)} - \mathbf{x}^*\| \leq \frac{L^m}{1-L}\|\mathbf{x}^{(1)} - \mathbf{x}^{(0)}\|,$$
$$\|\mathbf{x}^{(m)} - \mathbf{x}^*\| \leq \frac{1}{1-L}\|\mathbf{x}^{(m+1)} - \mathbf{x}^{(m)}\| \qquad \text{für alle } m \in \mathbb{N}_0.$$

Beweis. Sei $m \in \mathbb{N}$. Nach Voraussetzung gilt

$$\|\mathbf{x}^{(m+1)} - \mathbf{x}^{(m)}\| = \|\Phi(\mathbf{x}^{(m)}) - \Phi(\mathbf{x}^{(m-1)})\| \leq L\|\mathbf{x}^{(m)} - \mathbf{x}^{(m-1)}\|,$$

und mit einer einfachen Induktion folgt

$$\|\mathbf{x}^{(m+1)} - \mathbf{x}^{(m)}\| \leq L^m\|\mathbf{x}^{(1)} - \mathbf{x}^{(0)}\| \qquad \text{für alle } m \in \mathbb{N}_0.$$

Seien nun $m \in \mathbb{N}_0$ und $n \in \mathbb{N}$. Wir erhalten per Teleskopsumme

$$\begin{aligned}\|\mathbf{x}^{(m+n)} - \mathbf{x}^{(m)}\| &= \left\|\sum_{i=0}^{n-1} \mathbf{x}^{(m+i+1)} - \mathbf{x}^{(m+i)}\right\| \leq \sum_{i=0}^{n-1} \|\mathbf{x}^{(m+i+1)} - \mathbf{x}^{(m+i)}\| \\ &\leq \sum_{i=0}^{n-1} L^{m+i}\|\mathbf{x}^{(1)} - \mathbf{x}^{(0)}\| = L^m \frac{1-L^n}{1-L}\|\mathbf{x}^{(1)} - \mathbf{x}^{(0)}\|\end{aligned}$$

$$\leq \frac{L^m}{1-L}\|\mathbf{x}^{(1)} - \mathbf{x}^{(0)}\|. \tag{5.5}$$

Aus dieser Abschätzung folgt, dass $(\mathbf{x}^{(m)})_{m=0}^{\infty}$ eine Cauchy-Folge sein muss: Da $\lim_{m\to\infty} L^m = 0$ gilt, finden wir zu jedem $\epsilon \in \mathbb{R}_{>0}$ einen Punkt der Folge, ab dem zwei beliebige Iterierte höchstens einen Abstand von ϵ voneinander haben.

Da $\mathbb{K}^n$ ein Banach-Raum ist, ist die abgeschlossene Teilmenge U vollständig, also muss die Cauchy-Folge $(\mathbf{x}^{(m)})_{m=0}^{\infty}$ einen Grenzwert $\mathbf{x}^* \in U$ besitzen. Sei $\epsilon \in \mathbb{R}_{>0}$. Dann existiert ein $m \in \mathbb{N}_0$ mit $\|\mathbf{x}^{(m)} - \mathbf{x}^*\| \leq \epsilon$ und $\|\mathbf{x}^{(m+1)} - \mathbf{x}^*\| \leq \epsilon$ und wir erhalten mit der Dreiecksungleichung und (5.4) die Abschätzung

$$\begin{aligned}\|\mathbf{x}^* - \Phi(\mathbf{x}^*)\| &= \|\mathbf{x}^* - \mathbf{x}^{(m+1)} + \Phi(\mathbf{x}^{(m)}) - \Phi(\mathbf{x}^*)\| \\ &\leq \|\mathbf{x}^* - \mathbf{x}^{(m+1)}\| + \|\Phi(\mathbf{x}^{(m)}) - \Phi(\mathbf{x}^*)\| \quad \leq \epsilon + L\|\mathbf{x}^{(m)} - \mathbf{x}^*\| \leq \epsilon + L\,\epsilon.\end{aligned}$$

Da ϵ beliebig gewählt wurde, folgt $\mathbf{x}^* = \Phi(\mathbf{x}^*)$, der Grenzwert ist also tatsächlich ein Fixpunkt der Funktion Φ.

Sei nun $\mathbf{x}^+ \in U$ ein beliebiger Fixpunkt von Φ. Wieder aus (5.4) folgt

$$\|\mathbf{x}^* - \mathbf{x}^+\| = \|\Phi(\mathbf{x}^*) - \Phi(\mathbf{x}^+)\| \leq L\|\mathbf{x}^* - \mathbf{x}^+\|, \qquad (1-L)\|\mathbf{x}^* - \mathbf{x}^+\| \leq 0.$$

Wegen $1 - L > 0$ muss $\mathbf{x}^* = \mathbf{x}^+$ gelten, also ist $\mathbf{x}^*$ der einzige Fixpunkt von Φ in U.

Aus der Stetigkeit der Norm und (5.5) folgt

$$\|\mathbf{x}^* - \mathbf{x}^{(m)}\| = \lim_{n\to\infty} \|\mathbf{x}^{(m+n)} - \mathbf{x}^{(m)}\| \leq \frac{L^m}{1-L}\|\mathbf{x}^{(1)} - \mathbf{x}^{(0)}\| \qquad \text{für alle } m \in \mathbb{N}_0,$$

also die erste Fehlerabschätzung.

Zum Nachweis der zweiten Fehlerabschätzung wählen wir wieder $m \in \mathbb{N}_0$ und erhalten mit (5.4)

$$\begin{aligned}\|\mathbf{x}^{(m)} - \mathbf{x}^*\| &= \|\mathbf{x}^{(m)} - \mathbf{x}^{(m+1)} + \mathbf{x}^{(m+1)} - \mathbf{x}^*\| \leq \|\mathbf{x}^{(m)} - \mathbf{x}^{(m+1)}\| + \|\mathbf{x}^{(m+1)} - \mathbf{x}^*\| \\ &= \|\mathbf{x}^{(m)} - \mathbf{x}^{(m+1)}\| + \|\Phi(\mathbf{x}^{(m)}) - \Phi(\mathbf{x}^*)\| \\ &\leq \|\mathbf{x}^{(m)} - \mathbf{x}^{(m+1)}\| + L\|\mathbf{x}^{(m)} - \mathbf{x}^*\|,\end{aligned}$$

also auch

$$(1-L)\|\mathbf{x}^{(m)} - \mathbf{x}^*\| \leq \|\mathbf{x}^{(m+1)} - \mathbf{x}^{(m)}\|$$

und damit die zweite Fehlerabschätzung. ∎

Der Fixpunktsatz gilt für beliebige Normen auf $\mathbb{K}^n$, wir können also je nach Bedarf die Norm so anpassen, dass beispielsweise das Konvergenzgebiet U möglichst groß oder die Kontraktionszahl L möglichst klein wird.

Falls die Iterationsfunktion eine Kontraktion ist, folgt aus Banachs Fixpunktsatz, dass der Fehler, also der Unterschied zwischen der Iterierten $\mathbf{x}^{(m)}$ und der Lösung $\mathbf{x}^*$ pro Schritt mindestens um den Faktor L reduziert wird. Man spricht in diesem Fall von *linearer Konvergenz*.

Attraktiver wäre es natürlich, wenn wir den Fehler um mehr als lediglich einen linearen Faktor reduzieren könnten, und in der Tat ist es möglich, in bestimmten Situationen Iterationsfunktionen zu konstruieren, die diese Eigenschaft besitzen.

5.4 Newton-Verfahren

Unser Ziel ist es, eine Iterationsfunktion zu konstruieren, die den Fehler um mehr als einen linearen Faktor reduziert. Wir untersuchen dazu zunächst den skalarwertigen Fall, nämlich eine zweimal differenzierbare Funktion $f\colon \mathbb{R} \to \mathbb{R}$. Indem wir eine Taylor-Entwicklung (siehe Erinnerung 5.1) um x im Punkt x^* auswerten, erhalten wir

$$0 = f(x^*) = f(x) + f'(x)(x^* - x) + \frac{f''(\eta)}{2}(x^* - x)^2$$

für ein $\eta \in [x, x^*]$. Wenn wir davon ausgehen, dass $f'(x) \neq 0$ gilt, finden wir per Division

$$0 = \frac{f(x)}{f'(x)} + x^* - x + \frac{f''(\eta)}{2f'(x)}(x^* - x)^2.$$

Durch Umsortieren ergibt sich

$$x - \frac{f(x)}{f'(x)} = x^* + \frac{f''(\eta)}{2f'(x)}(x^* - x)^2. \tag{5.6}$$

Falls x nahe genug an der Lösung x^* liegt, ist also

$$\Phi(x) := x - \frac{f(x)}{f'(x)} \tag{5.7}$$

eine sehr gute Approximation von x^*. Diese Gleichung definiert die Iterationsfunktion der *Newton-Iteration* zur Nullstellenbestimmung. Aus (5.6) folgt

$$|\Phi(x) - x^*| = \frac{|f''(\eta)|}{2|f'(x)|}|x - x^*|^2.$$

Wir nehmen an, dass wir einen Radius $\varrho \in \mathbb{R}_{>0}$ und Konstanten $\beta, \gamma \in \mathbb{R}_{\geq 0}$ finden mit

$$\begin{aligned} \sup\left\{ \frac{1}{|f'(\xi)|} \;:\; \xi \in [x^* - \varrho, x^* + \varrho] \right\} &\leq \beta, \\ \sup\{|f''(\xi)| \;:\; \xi \in [x^* - \varrho, x^* + \varrho]\} &\leq \gamma. \end{aligned}$$

Dann erhalten wir die Fehlerabschätzung

$$|\Phi(x) - x^*| \leq \frac{\beta\gamma}{2}|x - x^*|^2 \qquad \text{für alle } x \in \mathbb{R} \text{ mit } |x - x^*| \leq \varrho, \tag{5.8}$$

und mit $r := \min\left\{\frac{1}{\beta\gamma}, \varrho\right\}$ folgt

$$|\Phi(x) - x^*| \leq \frac{\beta\gamma}{2}|x - x^*|^2 \leq \frac{1}{2}|x - x^*| \qquad \text{für alle } x \in \mathbb{R} \text{ mit } |x - x^*| \leq r,$$

das Iterationsverfahren wird also konvergieren, und wegen (5.8) wird die Konvergenzgeschwindigkeit immer besser, je besser die Näherungslösung ist.

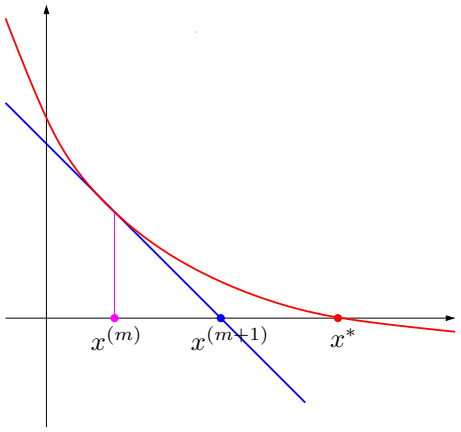

Abbildung 5.3: Geometrische Interpretation des Newton-Verfahrens

Da der Fehler der neuen Iterierten gemäß (5.8) quadratisch von dem Fehler der alten Iterierten abhängt, spricht man hier von *quadratischer Konvergenz*. Sie kann erheblich besser sein als lineare Konvergenz: Sei $x^{(0)} \in U$ mit $|x^{(0)} - x^*| \leq r$ gegeben. Dann gilt

$$|x^{(1)} - x^*| \leq \frac{\beta\gamma}{2}|x^{(0)} - x^*|^2 \leq \frac{\beta\gamma}{2}r^2 \leq \frac{\beta\gamma}{2}\frac{r}{\beta\gamma} = \frac{r}{2},$$
$$|x^{(2)} - x^*| \leq \frac{\beta\gamma}{2}|x^{(1)} - x^*|^2 \leq \frac{\beta\gamma}{2}\frac{r^2}{4} \leq \frac{\beta\gamma}{2}\frac{r}{4\,\beta\gamma} = \frac{r}{8},$$
$$|x^{(3)} - x^*| \leq \frac{\beta\gamma}{2}|x^{(2)} - x^*|^2 \leq \frac{\beta\gamma}{2}\frac{r^2}{64} \leq \frac{\beta\gamma}{2}\frac{r}{64\,\beta\gamma} = \frac{r}{128},$$

und mit einer einfachen Induktion folgt

$$|x^{(m)} - x^*| \leq 2r\,2^{-2^m} \qquad \text{für alle } m \in \mathbb{N}_0,$$

abgesehen von dem Vorfaktor wird also der Fehler in jedem Schritt quadriert. Offenbar lässt sich mit einem quadratisch konvergenten Verfahren sehr schnell eine sehr hohe Genauigkeit erreichen, sofern ein hinreichend guter Startwert zur Verfügung steht.

Bemerkung 5.7 (Geometrische Interpretation) *Das Newton-Verfahren lässt sich auch geometrisch interpretieren: Wir approximieren die Funktion f durch ihre Tangente im aktuellen Punkt $x^{(m)}$, die durch*

$$t\colon \mathbb{R} \to \mathbb{R}, \qquad x \mapsto f(x^{(m)}) + f'(x^{(m)})(x - x^{(m)})$$

gegeben ist. Da wir die Nullstelle x^* von f nicht direkt bestimmen können, berechnen wir stattdessen die Nullstelle $x^{(m+1)}$ der Approximation t, die durch

$$0 = t(x^{(m+1)}) = f(x^{(m)}) + f'(x^{(m)})(x^{(m+1)} - x^{(m)})$$

gegeben ist. Für $f'(x^{(m)}) \neq 0$ können wir diese Gleichung nach $x^{(m+1)}$ auflösen und erhalten die Iterationsvorschrift (5.7) des Newton-Verfahrens (siehe Abbildung 5.3).

Beispiel 5.8 (Wurzelberechnung) *Mit Hilfe des Newton-Verfahrens können wir uns der Frage der Berechnung der Wurzel einer positiven Zahl $a \in \mathbb{R}_{>0}$ widmen. Beispielsweise können wir $\sqrt{a}$ als Nullstelle der Funktion*

$$f\colon \mathbb{R}_{>0} \to \mathbb{R}_{>0}, \qquad x \mapsto x^2 - a,$$

darstellen, für die sich die Iterationsfunktion

$$\Phi(x) = x - \frac{f(x)}{f'(x)} = x - \frac{x^2 - a}{2x} = \frac{2x^2 - x^2 + a}{2x} = \frac{1}{2}\left(x + \frac{a}{x}\right) \qquad \text{für alle } x \in \mathbb{R}_{>0}$$

ergibt, die dem Newton-Heron-Verfahren *entspricht. Dieses Verfahren bietet den Vorteil, für beliebige Startwerte $x > 0$ zu konvergieren: Sei $x \in \mathbb{R}_{>0}$ der Startwert, und sei*

$$y := \Phi(x) = \frac{1}{2}\left(x + \frac{a}{x}\right)$$

die nächste Iterierte. Für den Fehler erhalten wir

$$y - \sqrt{a} = \frac{1}{2}\left(x + \frac{a}{x}\right) - \sqrt{a} = \frac{1}{2x}\left(x^2 - 2x\sqrt{a} + a\right) = \frac{(x - \sqrt{a})^2}{2x} \geq 0,$$

insbesondere wird also $y \geq \sqrt{a}$ gelten, nach dem ersten Schritt sind demnach alle Iterierten größer oder gleich $\sqrt{a}$.

Falls wir $x \geq \sqrt{a}$ voraussetzen, finden wir

$$y - \sqrt{a} = \frac{(x - \sqrt{a})^2}{2x} \leq \min\left\{\frac{x - \sqrt{a}}{2}, \frac{(x - \sqrt{a})^2}{2\sqrt{a}}\right\},$$

also mindestens lineare Konvergenz ab der zweiten Iterierten.

Für die Umsetzung auf einem Computer kann die folgende Variante geschickter sein: Anstelle der Wurzel berechnen wir die reziproke Wurzel *$b := 1/\sqrt{a}$, denn aus ihr können wir anschließend durch $ab = a/\sqrt{a} = \sqrt{a}$ die Wurzel rekonstruieren.*

Wir beschreiben b als Nullstelle der Funktion

$$f\colon \mathbb{R}_{>0} \to \mathbb{R}, \qquad x \mapsto a - \frac{1}{x^2} = a - x^{-2},$$

und wollen wieder das Newton-Verfahren anwenden. Durch Einsetzen von f und f' erhalten wir

$$\Phi(x) = x - \frac{f(x)}{f'(x)} = x - \frac{a - x^{-2}}{2x^{-3}} = x - \frac{ax^3 - x}{2} = \frac{x}{2}(3 - ax^2) \qquad \text{für alle } x \in \mathbb{R}_{>0}.$$

Bemerkenswert an dieser Iterationsvorschrift ist, dass ihre Auswertung lediglich Subtraktionen und Multiplikationen erfordert und deshalb auf modernen Computern sehr schnell ausgeführt werden kann.

Viele nichtlineare Gleichungen beinhalten mehr als eine Unbekannte, also benötigen wir eine mehrdimensionale Variante des Newton-Verfahrens. Sei nun also $f \in C^2(\mathbb{K}^n, \mathbb{K}^n)$ eine zweimal stetig differenzierbare Funktion von $\mathbb{K}^n$ nach $\mathbb{K}^n$. Wir können die Suche nach einer Nullstelle $\mathbf{x}^* \in \mathbb{K}^n$ dieser Funktion auf den eindimensionalen Fall zurückführen, indem wir die Strecke von $\mathbf{x}$ zu $\mathbf{x}^*$ durch

$$\lambda\colon [0,1] \to \mathbb{R}^n, \qquad s \mapsto \mathbf{x} + s(\mathbf{x}^* - \mathbf{x}),$$

mit $\lambda(0) = \mathbf{x}$, $\lambda(1) = \mathbf{x}^*$ und $\lambda'(s) = \mathbf{x}^* - \mathbf{x}$ parametrisieren und die Hilfsfunktion

$$g\colon [0,1] \to \mathbb{R}^n, \qquad s \mapsto f(\lambda(s)),$$

einführen, deren Ableitung infolge der Kettenregel durch

$$g'(s) = Df(\lambda(s))(\mathbf{x}^* - \mathbf{x}) \qquad \text{für alle } s \in [0,1]$$

gegeben ist. Hier bezeichnet

$$Df(\mathbf{y}) = \begin{pmatrix} \frac{\partial f_1}{\partial y_1}(\mathbf{y}) & \cdots & \frac{\partial f_1}{\partial y_n}(\mathbf{y}) \\ \vdots & \ddots & \vdots \\ \frac{\partial f_n}{\partial y_1}(\mathbf{y}) & \cdots & \frac{\partial f_n}{\partial y_n}(\mathbf{y}) \end{pmatrix} \in \mathbb{K}^{n\times n}$$

die *Jacobi-Matrix* der Funktion f. Wir wenden den Hauptsatz der Integral- und Differentialrechnung auf g an, um

$$\begin{aligned} \mathbf{0} = f(\mathbf{x}^*) = g(1) &= g(0) + \int_0^1 g'(s)\,ds = g(0) + g'(0) + \int_0^1 g'(s) - g'(0)\,ds \\ &= f(x) + Df(\mathbf{x})(\mathbf{x}^* - \mathbf{x}) + \int_0^1 \big(Df(\lambda(s)) - Df(\mathbf{x})\big)(\mathbf{x}^* - \mathbf{x})\,ds \end{aligned}$$

zu erhalten. Wir gehen davon aus, dass die Matrix $Df(\mathbf{x})$ invertierbar ist und multiplizieren die Gleichung mit ihrer Inversen, so dass sich

$$\mathbf{0} = Df(\mathbf{x})^{-1} f(\mathbf{x}) + \mathbf{x}^* - \mathbf{x} + Df(\mathbf{x})^{-1} \int_0^1 \big(Df(\lambda(s)) - Df(\mathbf{x})\big)(\mathbf{x}^* - \mathbf{x})\,ds$$

ergibt. Durch Umsortieren folgt

$$\mathbf{x} - Df(\mathbf{x})^{-1} f(\mathbf{x}) = \mathbf{x}^* + Df(\mathbf{x})^{-1} \int_0^1 \big(Df(\lambda(s)) - Df(\mathbf{x})\big)(\mathbf{x}^* - \mathbf{x})\,ds. \tag{5.9}$$

Diese Gleichung ist das mehrdimensionale Gegenstück der Gleichung (5.6).

Um λ in der beschriebenen Weise definieren zu können, muss zu zwei Punkten $\mathbf{x}$ und $\mathbf{x}^*$ des Definitionsbereichs auch die Verbindungsstrecke im Definitionsbereich liegen.

Definition 5.9 (Konvexe Menge) *Sei $U \subseteq \mathbb{R}^n$. Die Menge U heißt* konvex*, falls*

$$\mathbf{x} + s(\mathbf{y} - \mathbf{x}) \in U \qquad \textit{für alle } \mathbf{x}, \mathbf{y} \in U,\ s \in [0,1] \tag{5.10}$$

gilt, falls also die Verbindungsstrecke beliebiger Punkte aus U wieder in U liegt.

Falls das Integral klein ist, gilt

$$\mathbf{x} - Df(\mathbf{x})^{-1} f(\mathbf{x}) \approx \mathbf{x}^*,$$

und damit können wir das mehrdimensionale Newton-Verfahren durch

$$\Phi(\mathbf{x}) := \mathbf{x} - Df(\mathbf{x})^{-1} f(\mathbf{x}) \tag{5.11}$$

definieren. Während das eindimensionale Verfahren nur durchführbar ist, falls $f'(x) \neq 0$ gilt, ist das mehrdimensionale Verfahren nur durchführbar, falls die Jacobi-Matrix $Df(\mathbf{x})$ regulär ist. Im eindimensionalen Fall stimmen beide Bedingungen und Formeln überein.

Satz 5.10 (Quadratische Konvergenz) *Seien $\|\cdot\|_V$ und $\|\cdot\|_W$ Normen auf $\mathbb{K}^n$. Sei eine konvexe Umgebung $U \subseteq \mathbb{K}^n$ von $\mathbf{x}^*$ so gewählt, dass $Df(\mathbf{x})$ für alle $\mathbf{x} \in U$ regulär ist und es Konstanten $\beta, \gamma \in \mathbb{R}_{\geq 0}$ gibt, die*

$$\|Df(\mathbf{x})^{-1}\|_{V \leftarrow W} \leq \beta \qquad \text{für alle } \mathbf{x} \in U, \tag{5.12a}$$

$$\|Df(\mathbf{x}) - Df(\mathbf{y})\|_{W \leftarrow V} \leq \gamma \|\mathbf{x} - \mathbf{y}\|_V \qquad \text{für alle } \mathbf{x}, \mathbf{y} \in U \tag{5.12b}$$

erfüllen. Dann gilt

$$\|\mathbf{x}^* - \Phi(\mathbf{x})\|_V \leq \frac{\beta\gamma}{2} \|\mathbf{x}^* - \mathbf{x}\|_V^2 \qquad \text{für alle } \mathbf{x} \in U.$$

Falls ein $r \in \mathbb{R}_{>0}$ so existiert, dass die offene Kugel $K := \{\mathbf{x} \in \mathbb{K}^n \,:\, \|\mathbf{x} - \mathbf{x}^*\|_V < r\}$ in U enthalten ist und $r \leq \frac{2}{\beta\gamma}$ gilt, konvergiert die Newton-Iteration für jeden Startvektor aus K, und alle Iterierten liegen ebenfalls in K.

Beweis. Sei $\mathbf{x} \in U$. Ausgehend von (5.9) erhalten wir mit (3.2a), (5.12a) und der Dreiecksungleichung für Integrale die Abschätzung

$$\begin{aligned}
\|\Phi(\mathbf{x}) - \mathbf{x}^*\|_V &= \left\| Df(\mathbf{x})^{-1} \int_0^1 \big(Df(\lambda(s)) - Df(\mathbf{x})\big)(\mathbf{x}^* - \mathbf{x})\, ds \right\|_V \\
&\leq \|Df(\mathbf{x})^{-1}\|_{V \leftarrow W} \left\| \int_0^1 \big(Df(\lambda(s)) - Df(\mathbf{x})\big)(\mathbf{x}^* - \mathbf{x})\, ds \right\|_W \\
&\leq \beta \int_0^1 \big\| \big(Df(\lambda(s)) - Df(\mathbf{x})\big)(\mathbf{x}^* - \mathbf{x}) \big\|_W \, ds \\
&\leq \beta \int_0^1 \|Df(\lambda(s)) - Df(\mathbf{x})\|_{W \leftarrow V} \, \|\mathbf{x}^* - \mathbf{x}\|_V \, ds.
\end{aligned}$$

Nun können wir die Lipschitz-Bedingung (5.12b) sowie die Definition der Abbildung λ verwenden, um zu

$$\begin{aligned}
\|\Phi(\mathbf{x}) - \mathbf{x}^*\|_V &\leq \beta \int_0^1 \|Df(\lambda(s)) - Df(\mathbf{x})\|_{W \leftarrow V} \, \|\mathbf{x}^* - \mathbf{x}\|_V \, ds \\
&\leq \beta\gamma \int_0^1 \|\lambda(s) - \mathbf{x}\|_V \|\mathbf{x}^* - \mathbf{x}\|_V \, ds
\end{aligned}$$

$$= \beta\gamma \int_0^1 s\|\mathbf{x}^* - \mathbf{x}\|_V^2 \, ds = \frac{\beta\gamma}{2}\|\mathbf{x}^* - \mathbf{x}\|_V^2$$

zu gelangen und den ersten Teil des Beweises abzuschließen.

Zum Nachweis der Konvergenz seien $r \in \mathbb{R}_{>0}$ mit $r \leq \frac{2}{\beta\gamma}$ und $K \subseteq U$ gegeben. Sei $\mathbf{x}^{(0)} \in K$. Wir definieren

$$\delta := \frac{\beta\gamma}{2}\|\mathbf{x}^{(0)} - \mathbf{x}^*\|_V$$

und halten fest, dass wegen $\|\mathbf{x}^{(0)} - \mathbf{x}^*\|_V < r \leq \frac{2}{\beta\gamma}$ insbesondere $\delta < 1$ gilt.

Soeben haben wir bewiesen, dass daraus einerseits

$$\|\mathbf{x}^* - \mathbf{x}^{(1)}\|_V = \|\mathbf{x}^* - \Phi(\mathbf{x}^{(0)})\|_V \leq \frac{\beta\gamma}{2}\|\mathbf{x}^* - \mathbf{x}^{(0)}\|_V^2 = \delta\|\mathbf{x}^* - \mathbf{x}^{(0)}\|_V < r$$

folgt, also $\mathbf{x}^{(1)} \in K$, und wir andererseits auch

$$\frac{\beta\gamma}{2}\|\mathbf{x}^* - \mathbf{x}^{(1)}\|_V = \frac{\beta\gamma}{2}\|\mathbf{x}^* - \Phi(\mathbf{x}^{(0)})\|_V \leq \left(\frac{\beta\gamma}{2}\right)^2 \|\mathbf{x}^* - \mathbf{x}^{(0)}\|_V^2 \leq \delta^2$$

festhalten dürfen. Mit einer einfachen Induktion erhalten wir

$$\mathbf{x}^{(m)} \in K, \qquad \frac{\beta\gamma}{2}\|\mathbf{x}^* - \mathbf{x}^{(m)}\|_V \leq \delta^{2^m} \qquad \text{für alle } m \in \mathbb{N}_0,$$

also wegen $\delta < 1$ insbesondere Konvergenz gegen $\mathbf{x}^*$. ∎

Bemerkung 5.11 (Affine Invarianz) *Eine nützliche Eigenschaft des Newton-Verfahrens ist seine Invarianz unter linearen Abbildungen: Falls $\mathbf{A} \in \mathbb{R}^{n\times n}$ eine reguläre Matrix ist, sind die Newton-Iterierten für f und $\tilde{f} := \mathbf{A}f$ identisch, falls derselbe Anfangsvektor $\mathbf{x}^{(0)}$ gewählt wurde.*

Unsere Bedingungen (5.12a) und (5.12b) weisen diese Invarianzeigenschaft nicht auf. Allerdings können wir sie durch eine kombinierte Bedingung der Form

$$\big\|Df(\mathbf{x})^{-1}\big(Df(\mathbf{x}) - Df(\mathbf{y})\big)\big\|_{V\leftarrow V} \leq \alpha\|\mathbf{x} - \mathbf{y}\|_V \qquad \textit{für alle } \mathbf{x}, \mathbf{y} \in U$$

ersetzen, die die Invarianz bewahrt und mit der sich der Beweis analog führen lässt.

Bemerkung 5.12 (Praktische Umsetzung) *In der Praxis wird man bei den meisten Anwendungen nicht die Inverse $Df(\mathbf{x})^{-1}$ berechnen, um den Iterationsschritt*

$$\mathbf{x}^{(m+1)} = \mathbf{x}^{(m)} - Df(\mathbf{x}^{(m)})^{-1} f(\mathbf{x}^{(m)})$$

durchzuführen. Stattdessen löst man das lineare Gleichungssystem

$$Df(\mathbf{x}^{(m)})\mathbf{r}^{(m)} = -f(\mathbf{x}^{(m)})$$

und berechnet die neue Iterierte durch

$$\mathbf{x}^{(m+1)} = \mathbf{x}^{(m)} + \mathbf{r}^{(m)}.$$

Falls das lineare Gleichungssystem mit Hilfe einer Faktorisierung gelöst wird, ist der Rechenaufwand für die Bestimmung der Faktorisierung in der Regel wesentlich höher als der für das Lösen des Systems (etwa per Rückwärts- und Vorwärtseinsetzen). Deshalb kann es sinnvoll sein, den Rechenaufwand zu reduzieren, indem man dieselbe Matrix (und damit auch dieselbe Faktorisierung) wiederholt verwendet und beispielsweise hofft, dass $Df(\mathbf{x}^{(m)})$ eine gute Näherung von $Df(\mathbf{x}^{(m+1)})$ darstellt.

Bemerkung 5.13 (Gedämpftes Verfahren) *Die quadratische Konvergenz des Newton-Verfahrens hat auch eine Schattenseite: Falls der Startvektor $\mathbf{x}^{(0)}$ nicht nahe genug an der Nullstelle $\mathbf{x}^*$ liegt, können die Iterierten sich immer weiter von der Nullstelle entfernen.*

Um zu verhindern, dass das Newton-Verfahren bei einem ungünstig gewählten Startpunkt divergiert, verwendet man häufig das modifizierte Verfahren

$$\mathbf{x}^{(m+1)} = \mathbf{x}^{(m)} + \alpha^{(m)}\mathbf{r}^{(m)},$$

bei dem man den Skalierungsparameter $\alpha^{(m)} \in \mathbb{K}$ so wählt, dass $\|f(\mathbf{x}^{(m+1)})\|$ oder $\|Df(\mathbf{x}^{(m+1)})f(\mathbf{x}^{(m+1)})\|$ möglichst klein wird. Auf diese Weise lässt sich unter bestimmten Bedingungen globale Konvergenz erzwingen.

Es gibt verschiedene Techniken zur Wahl von $\alpha^{(m)}$, beispielsweise könnte man den Parameter mit Hilfe eines Bisektionsverfahrens zu bestimmen versuchen oder einfach so lange reduzieren, bis die Norm kleiner wird.

6 Approximation von Funktionen

Das Newton-Verfahren, das wir als Methode zur Behandlung nichtlinearer Gleichungssysteme kennen gelernt haben, basiert auf der Idee, eine potentiell komplizierte Funktion durch eine einfachere Funktion zu ersetzen: Statt die Nullstelle einer Funktion f zu berechnen, ersetzen wir f durch eine Tangente, und die Nullstelle der Tangente verwenden wir dann als Approximation der Nullstelle von f.

Da vergleichbare Ansätze in der numerischen Mathematik sehr häufig anzutreffen sind, lohnt es sich, sie in einem abstrakten Rahmen zu untersuchen. Als Beispiel dient uns die *Polynominterpolation*: Wir wollen eine Funktion $f\colon \mathbb{R} \to \mathbb{R}$ approximieren. Bekannt sind einzelne Werte $f_i = f(x_i)$ in Punkten $x_0 < \ldots < x_m$. Unser Ziel ist es, ein Polynom zu konstruieren, das in diesen Punkten dieselben Werte wie f annimmt und damit hoffentlich die Funktion auch in anderen Punkten approximiert. Wir suchen das Polynom in dem Vektorraum

$$\Pi_m := \operatorname{span}\{x \mapsto a_0 + a_1 x + \ldots + a_m x^m \;:\; a_0, \ldots, a_m \in \mathbb{R}\}$$

der Polynome höchstens m-ten Grades. Zur Abkürzung verzichtet man häufig auf das Wort „höchstens“ und spricht von Polynomen m-ten Grades, selbst wenn der Grad niedriger ist.

Mit dieser Notation nimmt die zu lösende Aufgabe die folgende Form an:

> Gegeben Punkte $x_0 < \ldots < x_m \in \mathbb{R}$ und Funktionswerte $f_0, \ldots, f_m \in \mathbb{R}$,
> finde ein Polynom $p \in \Pi_m$ so, dass $p(x_i) = f_i$ für alle $i \in [0 : m]$ gilt.

6.1 Existenz und Eindeutigkeit

Bevor wir uns der konkreten Konstruktion des Polynoms $p \in \Pi_m$ widmen können, müssen wir zunächst klären, ob ein derartiges Polynom überhaupt existiert und, falls es existiert, ob es durch die Werte in den Punkten $x_0, \ldots, x_m$ eindeutig bestimmt ist.

Die Frage nach der Existenz lässt sich sehr einfach beantworten, indem wir ein passendes Polynom $p \in \Pi_m$ explizit angeben. Wir verwenden dazu die *Lagrange-Polynome* $\ell_j \in \Pi_m$, die durch

$$\ell_j(x) = \prod_{\substack{k=0 \\ k \neq j}}^{m} \frac{x - x_k}{x_j - x_k} \qquad \text{für alle } j \in [0 : m] \text{ und } x \in \mathbb{R} \tag{6.1}$$

definiert sind. Da nach Voraussetzung aus $j \neq k$ auch $x_j \neq x_k$ folgt, sind alle Faktoren auf der rechten Seite der Gleichung wohldefiniert, und da alle diese Faktoren Polynome

ersten Grades sind und es m Faktoren gibt, muss auch $\ell_j \in \Pi_m$ gelten. Also ist ℓ_j für alle $j \in [0:m]$ ein wohldefiniertes Polynom m-ten Grades.

Die zentrale Eigenschaft der Lagrange-Polynome besteht darin, dass sie in den Punkten $x_0, \dots, x_m$ bestimmte Werte annehmen: Es gilt

$$\ell_j(x_j) = \prod_{\substack{k=0 \\ k \neq j}}^{m} \frac{x_j - x_k}{x_j - x_k} = \prod_{\substack{k=0 \\ k \neq j}}^{m} 1 = 1,$$

und für jedes $i \in [0:m]$ mit $i \neq j$ erhalten wir

$$\ell_j(x_i) = \prod_{\substack{k=0 \\ k \neq j}}^{m} \frac{x_i - x_k}{x_j - x_k} = \frac{x_i - x_i}{x_j - x_i} \prod_{\substack{k=0 \\ k \neq i, k \neq j}}^{m} \frac{x_i - x_k}{x_j - x_k} = 0,$$

da jedes derartige i auch in dem Produkt über alle $k \neq j$ auftreten muss. Beide Gleichungen lassen sich in der Form

$$\ell_j(x_i) = \begin{cases} 1 & \text{falls } i = j, \\ 0 & \text{ansonsten} \end{cases} \qquad \text{für alle } i, j \in [0:m]$$

zusammenfassen. Infolge dieser Gleichung wird nun $f_j \ell_j$ ein Polynom aus Π_m sein, das in dem Punkt x_j den Wert f_j annimmt und in allen anderen Punkten gleich null ist, und die Summe aller derartigen Polynome

$$p := \sum_{j=0}^{m} f_j \ell_j \tag{6.2}$$

muss demzufolge in allen Punkten den richtigen Wert annehmen: Es gilt

$$p(x_i) = \sum_{j=0}^{m} f_j \ell_j(x_i) = f_i \ell_i(x_i) = f_i \qquad \text{für alle } i \in [0:m],$$

also haben wir ein Polynom gefunden, das unser Problem löst.

Satz 6.1 (Existenz und Eindeutigkeit) *Es existiert genau ein $p \in \Pi_m$, das $p(x_i) = f_i$ für alle $i \in [0:m]$ erfüllt, und es ist durch (6.2) gegeben.*

Insbesondere ist Π_m ein $(m+1)$-dimensionaler Vektorraum.

Beweis. Wir definieren den linearen Operator

$$\Phi \colon \Pi_m \to \mathbb{R}^{m+1}, \qquad p \mapsto \begin{pmatrix} p(x_0) \\ \vdots \\ p(x_m) \end{pmatrix},$$

der jedem Polynom dessen Werte in den Interpolationspunkten zuordnet.

Für jeden beliebigen Vektor

$$\mathbf{f} = \begin{pmatrix} f_0 \\ \vdots \\ f_m \end{pmatrix} \in \mathbb{R}^{m+1}$$

erfüllt das durch (6.2) definierte Polynom $p \in \Pi_m$ dann $\Phi[p] = \mathbf{f}$, also ist Φ surjektiv.

Nach unserer Definition wird der Raum Π_m von den Polynomen $x \mapsto x^i$ für $i \in [0:m]$ aufgespannt, kann also höchstens die Dimension $m+1$ besitzen. Mit dem Dimensionssatz der linearen Algebra folgt

$$m+1 \geq \dim \Pi_m = \dim \operatorname{Bild} \Phi + \dim \operatorname{Kern} \Phi = m+1+\dim \operatorname{Kern} \Phi,$$

und wir lesen $\dim \Pi_m = m+1$ und $\dim \operatorname{Kern} \Phi = 0$ ab. Also ist Φ injektiv, und damit auch bijektiv. ∎

Bemerkung 6.2 (Polynombasis) *Da nach Definition die $(m+1)$-elementige Menge*

$$\{x \mapsto x^i \ : \ i \in [0:m]\}$$

den $(m+1)$-dimensionalen Polynomraum Π_m aufspannt, muss es sich um eine Basis handeln. Wir nennen sie die Monombasis. *Jedes Polynom $p \in \Pi_m$ können wir mit Hilfe des Satzes von Taylor in dieser Basis darstellen, denn es gilt*

$$p(x) = \sum_{i=0}^{m} \frac{p^{(i)}(0)}{i!} x^i \qquad \text{für alle } x \in \mathbb{R},$$

da $p^{(m+1)} = 0$ gilt und der Restterm deshalb entfällt.

Satz 6.1 besagt, dass auch die Lagrange-Polynome eine Basis des Polynomraums Π_m bilden, die wir naheliegenderweise als Lagrange-Basis *bezeichnen.*

Übungsaufgabe 6.3 (Duale Basis) *Seien lineare Abbildungen*

$$\lambda_i \colon \Pi_m \to \mathbb{R} \qquad \text{für alle } i \in [0:m]$$

so gegeben, dass für jedes Polynom $p \in \Pi_m$ aus

$$\lambda_i(p) = 0 \qquad \text{für alle } i \in [0:m]$$

bereits $p = 0$ folgt.

(a) Beweisen Sie, dass Polynome $q_0, \ldots, q_m \in \Pi_m$ existieren mit

$$\lambda_i(q_j) = \begin{cases} 1 & \text{falls } i = j, \\ 0 & \text{ansonsten} \end{cases} \qquad \text{für alle } i, j \in [0:m].$$

(b) Beweisen Sie mit den in (a) gefundenen Polynomen $q_0, \ldots, q_m$ die Gleichung

$$p = \sum_{i=0}^{m} \lambda_i(p) q_i \qquad \text{für alle } p \in \Pi_m.$$

(c) Beweisen Sie, dass die Polynome $q_0, \ldots, q_m \in \Pi_m$ aus Teil (a) eine Basis des Raums Π_m sind.

6.2 Effiziente Auswertung

Im Prinzip können wir das Polynom p direkt mit Hilfe der Formel (6.2) in einem beliebigen Punkt x auswerten. Die Auswertung eines Lagrange-Polynoms erfordert $2m$ Subtraktionen, m Divisionen und $m-1$ Multiplikationen, also $4m-1$ Operationen. Die Auswertung von p erfordert $m+1$ Multiplikationen mit den Werten $f_0, \ldots, f_m$, m Additionen, und $m+1$ Auswertungen der Lagrange-Polynome, so dass insgesamt

$$(m+1)(4m-1) + 2m + 1 = 4m(m+1) + m = m(4m+5)$$

Operationen anfallen. Das ist ein relativ hoher Aufwand für eine einfache Operation wie die Auswertung eines Polynoms.

Deshalb sind wir daran interessiert, einen Algorithmus zur Auswertung von p zu konstruieren, der schneller arbeitet. Wir führen Zwischenergebnisse ein: Für $i, j \in [0:m]$ mit $i \leq j$ bezeichnen wir mit $p_{i,j} \in \Pi_{j-i}$ das nach Satz 6.1 existierende Polynom, das

$$p_{i,j}(x_k) = f_k \qquad \text{für alle } k \in [i:j] \tag{6.3}$$

erfüllt. Diese Polynome nehmen also nur auf einer Teilmenge der Punkte die gewünschten Werte an.

Unser Ansatz beruht auf zwei Beobachtungen: Erstens gilt für $i = j$ offenbar $p_{i,j} \equiv f_i$, so dass sich diese Polynome sehr einfach konstruieren lassen. Zweitens können wir aus Polynomen niedrigen Grades solche höheren Grades gewinnen, indem wir die folgende Aussage verwenden.

Lemma 6.4 (Aitken-Rekurrenz) *Für alle $i, j \in [0:m]$ mit $i < j$ gilt*

$$p_{i,j}(x) = \frac{x_j - x}{x_j - x_i} p_{i,j-1}(x) + \frac{x - x_i}{x_j - x_i} p_{i+1,j}(x) \qquad \textit{für alle } x \in \mathbb{R}.$$

Beweis. Nach Definition gilt $p_{i,j-1}, p_{i+1,j} \in \Pi_{j-i-1}$, also ist das durch

$$q(x) := \frac{x_j - x}{x_j - x_i} p_{i,j-1}(x) + \frac{x - x_i}{x_j - x_i} p_{i+1,j}(x) \qquad \text{für alle } x \in \mathbb{R}$$

definierte Polynom höchstens vom Grad $j - i$.

Da $p_{i,j}$ denselben Grad besitzt, folgt aus dem Identitätssatz für Polynome, dass es ausreicht, die Gleichheit von $p_{i,j}$ und q in $j - i + 1$ verschiedenen Punkten nachzuweisen. Es ist naheliegend, dafür die Punkte x_k mit $k \in [i:j]$ zu verwenden.

Sei also $k \in [i:j]$ gewählt. Wir unterscheiden drei Fälle:

Fall 1: Für $k = i$ gelten nach Definition $p_{i,j-1}(x_k) = f_i$ sowie $p_{i,j}(x_k) = f_i$, also auch

$$q(x_k) = \frac{x_j - x_i}{x_j - x_i} p_{i,j-1}(x_i) + \frac{x_i - x_i}{x_j - x_i} p_{i+1,j}(x_i) = p_{i,j-1}(x_i) = f_i = p_{i,j}(x_k).$$

Fall 2: Für $k = j$ gelten nach Definition $p_{i+1,j}(x_k) = f_j$ sowie $p_{i,j}(x_k) = f_j$, also auch

$$q(x_k) = \frac{x_j - x_j}{x_j - x_i} p_{i,j-1}(x_j) + \frac{x_j - x_i}{x_j - x_i} p_{i+1,j}(x_j) = p_{i+1,j}(x_j) = f_j = p_{i,j}(x_k).$$

Fall 3: Für $i < k < j$ gelten nach Definition $p_{i,j-1}(x_k) = f_k$, $p_{i+1,j}(x_k) = f_k$ sowie $p_{i,j}(x_k) = f_k$, so dass wir

$$\begin{aligned} q(x_k) &= \frac{x_j - x_k}{x_j - x_i} p_{i,j-1}(x_k) + \frac{x_k - x_i}{x_j - x_i} p_{i+1,j}(x_k) = \frac{x_j - x_k}{x_j - x_i} f_k + \frac{x_k - x_i}{x_j - x_i} f_k \\ &= \frac{x_j - x_k + x_k - x_i}{x_j - x_i} f_k = \frac{x_j - x_i}{x_j - x_i} f_k = f_k = p_{i,j}(x_k) \end{aligned}$$

erhalten. Damit stimmen $p_{i,j}$ und q in $j - i + 1$ Punkten überein, sind also identisch. ■

Die in Lemma 6.4 angegebene Formel zur Berechnung von $p_{i,j}$ lässt sich wiederholt anwenden, um die Auswertung eines Polynoms m-ten Grades auf die Auswertung von zwei Polynomen $(m-1)$-ten Grades, von drei Polynomen $(m-2)$-ten Grades und schließlich von $m+1$ Polynomen nullten Grades zurückzuführen. Letztere lassen sich, wie bereits gezeigt, einfach handhaben.

Das resultierende *Neville-Aitken-Verfahren* lässt sich in der Form eines Dreiecksschemas darstellen, bei dem die bekannten Werte $f_0, \ldots, f_m$ auf der linken Seite und der zu bestimmende Wert $p(x) = p_{0,m}(x)$ auf der rechten Seite steht:

$$\begin{array}{c|ccccccc} f_0 & p_{0,0}(x) & & & & & & \\ f_1 & p_{1,1}(x) & p_{0,1}(x) & & & & & \\ f_2 & p_{2,2}(x) & p_{1,2}(x) & p_{0,2}(x) & & & & \\ \vdots & \vdots & \vdots & \vdots & \ddots & & & \\ f_{m-1} & p_{m-1,m-1}(x) & p_{m-2,m-1}(x) & p_{m-3,m-1}(x) & \cdots & p_{0,m-1}(x) & & \\ f_m & p_{m,m}(x) & p_{m-1,m}(x) & p_{m-2,m}(x) & \cdots & p_{1,m}(x) & p_{0,m}(x) \end{array}$$

In diesem Schema wird der rechts stehende Wert jeweils aus seinem linken Nachbarn und dessen oberem Nachbarn berechnet.

Bei der Implementierung des Verfahrens können wir wieder Speicherplatz sparen, indem wir geschickt Werte überschreiben: Falls wir die Werte einer Spalte des Neville-Aitken-Schemas von unten nach oben berechnen, wird $p_{i+1,j}(x)$ nicht mehr benötigt, sobald $p_{i,j}(x)$ berechnet worden ist.

---

procedure eval_poly(m, x, **var f**);
for $k = 1, \ldots, m$ **do**
  for $i = m, \ldots, k$ **do**
    $h \leftarrow (x_i - x) f_{i-1} + (x - x_{i-k}) f_i$
    $f_i \leftarrow h/(x_i - x_{i-k})$
  end for
end for

Abbildung 6.1: Auswertung des Interpolationspolynoms $p = p_{0,m}$ in einem Punkt x. Zu Beginn enthält der Vektor $\mathbf{f} = (f_0, \ldots, f_m)$ die zu interpolierenden Werte, am Ende gilt $p_{0,m}(x) = f_m$.

Wir gehen davon aus, dass die zu interpolierenden Werte in einem Vektor $\mathbf{f} = (f_0, \ldots, f_m)$ abgespeichert sind. Bei der Berechnung der ersten Spalte verfahren wir wie beschrieben und ersetzen f_m durch den Wert $p_{m-1,m}(x)$, dann f_{m-1} durch $p_{m-2,m-1}(x)$ und so weiter, bis f_1 durch $p_{0,1}(x)$ ersetzt wurde. Insgesamt findet sich $p_{i-1,i}(x)$ jeweils in f_i für $i \in [1:m]$.

Bei der Berechnung der zweiten Spalte ersetzen wir f_m durch den Wert $p_{m-2,m}(x)$, der sich aus $p_{m-1,m}(x)$, zu finden in f_m, und $p_{m-2,m-1}(x)$, zu finden in f_{m-1}, berechnet. Wir fahren so fort, bis f_2 durch $p_{0,2}(x)$ ersetzt wurde. Insgesamt findet sich nun $p_{i-2,i}(x)$ jeweils in f_i für $i \in [2:m]$.

Bei der Berechnung der k-ten Spalte ersetzen wir wieder f_m durch den Wert $p_{m-k,m}(x)$, der sich aus $p_{m-k+1,m}(x)$, zu finden in f_m, und $p_{m-k,m-1}(x)$, zu finden in f_{m-1}, berechnet. Wir fahren fort, bis f_{m-k} durch $p_{0,k}(x)$ ersetzt wurde. Insgesamt enthält nun f_i jeweils den Wert $p_{i-k,i}(x)$ für $i \in [k:m]$.

Der resultierende Algorithmus ist in Abbildung 6.1 zusammengefasst. Der Rumpf der inneren Schleife benötigt 7 Operationen, diese Schleife wird $(m-k+1)$-mal für $k = 1, \ldots, m$ ausgeführt, also erhalten wir eine Zahl von

$$\sum_{k=1}^{m} 7(m-k+1) = \sum_{\ell=1}^{m} 7\ell = \frac{7}{2} m(m+1)$$

Operationen, wenn wir $\ell = m - k + 1$ substituieren. Der neue Algorithmus ist also immerhin etwas schneller als der direkte Zugang über die Lagrange-Polynome.

Vor allem aber lässt er sich modifizieren, um eine besonders effiziente Auswertung von p zu ermöglichen: Grundlage ist das *Horner-Schema*, ein schnelles Verfahren zur Auswertung von Polynomen der Form

$$q(x) = a_0 + a_1 x + a_2 x^2 + \ldots + a_m x^m.$$

Wir stellen fest, dass alle Summanden außer dem ersten x enthalten, also können wir die Variable ausklammern:

$$q(x) = a_0 + x\left(a_1 + a_2 x + \ldots + a_m x^{m-1}\right).$$

Im Inneren der Klammer gilt dasselbe, also können wir auch hier x ausklammern und erhalten

$$q(x) = a_0 + x\left(a_1 + x\left(a_2 + \ldots + a_m x^{m-2}\right)\right).$$

Falls wir in dieser Weise fortfahren, erhalten wir eine Darstellung der Formel, die sich mit m Multiplikationen und m Additionen auswerten lässt, also mit einem Gesamtaufwand von $2m$. Das ist sehr viel weniger, als die bisherigen Algorithmen benötigen.

Leider liegt uns das Interpolationspolynom nicht als Summe von Monomen vor, so dass wir das Horner-Schema nicht direkt anwenden können. Wir können aber versuchen, eine Darstellung des Interpolationspolynoms p zu finden, die sich mit einem geeignet verallgemeinerten Horner-Schema handhaben lässt. Dazu vergleichen wir die Polynome $p_{0,i}$ und $p_{0,i-1}$ miteinander. Nach Definition stimmen sie in den Punkten $x_0, \ldots, x_{i-1}$

überein, also besitzt ihre Differenz $p_{0,i} - p_{0,i-1}$ in diesen Punkten Nullstellen, die wir herausdividieren können, um eine Darstellung der Form

$$p_{0,i}(x) - p_{0,i-1}(x) = (x - x_0) \dots (x - x_{i-1}) d_i \tag{6.4}$$

mit einer Konstanten $d_i \in \mathbb{R}$ zu erhalten. Indem wir induktiv vorgehen, erhalten wir

$$p_{0,1}(x) = p_{0,0}(x) + (x - x_0)d_1, \qquad p_{0,2}(x) = p_{0,1}(x) + (x - x_0)(x - x_1)d_2,$$
$$p_{0,m}(x) = p_{0,m-1}(x) + (x - x_0) \dots (x - x_{m-1})d_m$$

und somit die Gleichung

$$p(x) = d_0 + (x - x_0)d_1 + (x - x_0)(x - x_1)d_2 + \dots + (x - x_0) \dots (x - x_{m-1})d_m,$$

die den Namen *Newton-Darstellung* trägt.

Diese Formel lässt sich nun wieder per Horner-Schema auswerten: Wir klammern statt der Variablen x den Linearfaktor $x - x_0$ aus, um

$$p(x) = d_0 + (x - x_0)\Big(d_1 + (x - x_1)d_2 + \dots + (x - x_1) \dots (x - x_{m-1})d_m\Big)$$

zu erhalten, klammern dann $x - x_1$ aus, um

$$p(x) = d_0 + (x - x_0)\Big(d_1 + (x - x_1)\big(d_2 + \dots + (x - x_2) \dots (x - x_{m-1})d_m\big)\Big) \tag{6.5}$$

zu erhalten, und gelangen zu einer Formel, die sich in $3m$ Operationen auswerten lässt.

Lemma 6.5 (Newton-Basis) *Wir definieren für alle $i, j \in [0 : m]$ mit $i \leq j$ die* Newton-Basispolynome

$$n_{i,j}(x) := \begin{cases} 1 & \text{falls } i = j, \\ (x - x_i)\, n_{i+1,j}(x) & \text{ansonsten} \end{cases} \qquad \text{für alle } x \in \mathbb{R}. \tag{6.6}$$

Dann ist $\{n_{i,k} \ : \ k \in [i : j]\}$ eine Basis des Raums Π_{j-i}.

Beweis. Aus der Definition folgt mit einer trivialen Induktion

$$n_{i,j} \in \Pi_{j-i} \qquad \text{für alle } i, j \in [0 : m],\ i \leq j,$$

wir müssen also zeigen, dass die $j - i + 1$ Polynome $n_{i,i}, \dots, n_{i,j}$ linear unabhängig sind.

Diesen Beweis führen wir per Induktion über $j - i$.

Induktionsanfang: Sei $i = j \in [0 : m]$. Dann ist $n_{i,i} = 1$ eine Basis des Raums Π_0.

Induktionsvoraussetzung: Sei $n \in \mathbb{N}_0$ so gegeben, dass für alle $i, j \in [0 : m]$ mit $j - i = n$ die Polynome $n_{i,i}, \dots, n_{i,j}$ linear unabhängig sind.

Induktionsschritt: Seien $i, j \in [0 : m]$ mit $j - i = n + 1$ gegeben. Seien Koeffizienten $a_i, \dots, a_j \in \mathbb{R}$ mit

$$0 = a_i n_{i,i} + a_{i+1} n_{i,i+1} + \dots + a_j n_{i,j}$$

gegeben. Nach Definition gilt

$$\begin{aligned} 0 &= a_i n_{i,i}(x) + a_{i+1} n_{i,i+1}(x) + \ldots + a_j n_{i,j}(x) \\ &= a_i + (x - x_i)\big(a_{i+1} n_{i+1,i+1}(x) + \ldots + a_j n_{i+1,j}(x)\big) \qquad \text{für alle } x \in \mathbb{R}, \end{aligned}$$

und durch Einsetzen von $x = x_i$ folgt $a_i = 0$. Dann muss auch

$$0 = a_{i+1} n_{i+1,i+1}(x) + \ldots + a_j n_{i+1,j}(x) \qquad \text{für alle } x \in \mathbb{R} \setminus \{x_i\}$$

gelten, also nach dem Identitätssatz für Polynome

$$0 = a_{i+1} n_{i+1,i+1} + \ldots + a_j n_{i+1,j}.$$

Nach Induktionsvoraussetzung sind $n_{i+1,i+1}, \ldots, n_{i+1,j}$ wegen $j - (i+1) = n+1-1 = n$ linear unabhängig, also erhalten wir $a_{i+1} = a_{i+2} = \ldots = a_j = 0$. Da $a_i = 0$ bereits gezeigt ist, sind wir fertig. ∎

Um das Polynom $p = p_{0,m}$ mit dem Horner-Schema auswerten zu können, können wir Lemma 6.5 anwenden, um Koeffizienten $d_0, \ldots, d_m \in \mathbb{R}$ mit

$$p = \sum_{k=0}^{m} d_k \, n_{0,k} \tag{6.7}$$

zu finden. Das Horner-Schema verwendet die Zwischenstufen

$$q_\ell := \sum_{k=m-\ell}^{m} d_k \, n_{m-\ell,k} \in \Pi_\ell \qquad \text{für alle } \ell \in [0:m]$$

definieren. Offenbar gelten $q_0 = d_m$ und $q_m = p$, und mit der Rekurrenzgleichung

$$\begin{aligned} q_{\ell+1}(x) &= \sum_{k=m-\ell-1}^{m} d_k \, n_{m-\ell-1,k}(x) \\ &= d_{m-\ell-1} + (x - x_{m-\ell-1}) \sum_{k=m-\ell}^{m} d_k \, n_{m-\ell,k}(x) \\ &= d_{m-\ell-1} + (x - x_{m-\ell-1}) \, q_\ell(x) \qquad \text{für alle } \ell \in [0:m-1],\ x \in \mathbb{R} \end{aligned}$$

können wir ausgehend von q_0 der Reihe nach $q_1, q_2, \ldots, q_m$ berechnen, wobei jeder Schritt nur 3 arithmetische Operationen erfordert, also genügen insgesamt $3m$ Operationen für die Auswertung des Interpolationspolynoms.

Allerdings lässt sich dieses verallgemeinerte Horner-Schema nur einsetzen, wenn uns die Koeffizienten $d_0, \ldots, d_m$ zur Verfügung stehen. Ein klassischer Zugang zu ihrer Berechnung auf *dividierten Differenzen*, die sich aus der in Lemma 6.4 bewiesenen Aitken-Rekurrenz herleiten lassen. Als Vorbereitung benötigen wir eine alternative Darstellung der Newton-Basis.

Lemma 6.6 (Newton-Basis von rechts) *Für alle $i, j \in [0:m]$ mit $i \leq j$ gilt*

$$n_{i,j}(x) = \begin{cases} 1 & \text{falls } i = j, \\ n_{i,j-1}(x)\,(x - x_{j-1}) & \text{ansonsten} \end{cases} \qquad \text{für alle } x \in \mathbb{R}. \tag{6.8}$$

Beweis. Per Induktion über $j - i$.

Induktionsanfang: Sei $i = j \in [0:m]$. Dann gilt $n_{i,j} = 1$.

Induktionsvoraussetzung: Sei $n \in \mathbb{N}_0$ so gegeben, dass (6.8) für alle $i, j \in [0:m]$ mit $j - i = n$ gilt.

Induktionsschritt: Seien $i, j \in [0:m]$ mit $j - i = n + 1$ gegeben. Da dann $j > i$ gilt, folgt mit Definition (6.6)

$$n_{i,j}(x) = (x - x_i) n_{i+1,j}(x) \qquad \text{für alle } x \in \mathbb{R}.$$

Wegen $j - (i+1) = n$ können wir die Induktionsvoraussetzung anwenden. Falls $i+1 = j$ gilt, folgt

$$n_{i,j}(x) = (x - x_i)\, n_{i+1,j}(x) = x - x_i = n_{i,j-1}(x)\,(x - x_{j-1}) \qquad \text{für alle } x \in \mathbb{R}.$$

Anderenfalls, also für $i + 1 < j$, haben wir

$$\begin{aligned} n_{i,j}(x) &= (x - x_i)\, n_{i+1,j}(x) = (x - x_i)\, n_{i+1,j-1}(x)\,(x - x_{j-1}) \\ &= n_{i,j-1}(x)\,(x - x_{j-1}) \qquad \text{für alle } x \in \mathbb{R}. \end{aligned}$$

Damit ist die gewünschte Gleichung bewiesen. ∎

Lemma 6.7 (Dividierte Differenzen) *Seien $i, j \in [0:m]$ mit $i < j$ gegeben. Seien $d_{i,i}, \ldots, d_{i,j-1} \in \mathbb{R}$ und $d_{i+1,i+1}, \ldots, d_{i+1,j} \in \mathbb{R}$ gegeben mit*

$$p_{i,j-1} = \sum_{k=i}^{j-1} d_{i,k}\, n_{i,k}, \qquad p_{i+1,j} = \sum_{k=i+1}^{j} d_{i+1,k}\, n_{i+1,k}.$$

Dann gilt

$$p_{i,j} = \sum_{k=i}^{j} d_{i,k}\, n_{i,k}, \qquad \text{mit} \qquad d_{i,j} := \frac{d_{i+1,j} - d_{i,j-1}}{x_j - x_i}. \tag{6.9}$$

Beweis. Gemäß Lemma 6.5 existieren Koeffizienten $\hat{d}_{i,i}, \ldots, \hat{d}_{i,j} \in \mathbb{R}$ mit

$$p_{i,j} = \sum_{k=i}^{j} \hat{d}_{i,k}\, n_{i,k}.$$

Da das Polynom

$$q := \sum_{k=i}^{j-1} \hat{d}_{i,k}\, n_{i,k}$$

in Π_{j-i-1} liegt und in Punkten $x \in \{x_i, \dots, x_{j-1}\}$ wegen $n_{i,j}(x) = 0$ mit $p_{i,j-1}$ übereinstimmt, muss nach Identitätssatz $q = p_{i,j-1}$ gelten. Da $\{n_{i,i}, \dots, n_{i,j-1}\}$ nach Lemma 6.5 eine Basis ist, folgt per Koeffizientenvergleich $\hat{d}_{i,k} = d_{i,k}$ für alle $k \in [i : j-1]$.

Wir kombinieren Lemma 6.4 mit (6.6) und (6.8) und erhalten

$$\begin{aligned}
p_{i,j}(x) &= \frac{x - x_i}{x_j - x_i} p_{i+1,j}(x) + \frac{x_j - x}{x_j - x_i} p_{i,j-1}(x) \\
&= \frac{x - x_i}{x_j - x_i} \sum_{k=i+1}^{j} d_{i+1,k} n_{i+1,k}(x) - \frac{x - x_j}{x_j - x_i} \sum_{k=i}^{j-1} d_{i,k} n_{i,k}(x) \\
&= \frac{1}{x_j - x_i} \left((x - x_i) \sum_{k=i+1}^{j} d_{i+1,k} n_{i+1,k}(x) - (x - x_j) \sum_{k=i}^{j-1} d_{i,k} n_{i,k}(x) \right) \\
&= \frac{1}{x_j - x_i} \left(\sum_{k=i+1}^{j} d_{i+1,k} n_{i,k}(x) - \sum_{k=i}^{j-1} (x - x_k + x_k - x_j) d_{i,k} n_{i,k}(x) \right) \\
&= \frac{1}{x_j - x_i} \left(\sum_{k=i+1}^{j} d_{i+1,k} n_{i,k}(x) - \sum_{k=i}^{j-1} d_{i,k} n_{i,k+1}(x) - \sum_{k=i}^{j-1} (x_k - x_j) d_{i,k} n_{i,k}(x) \right) \\
&= \frac{1}{x_j - x_i} \left(\sum_{k=i+1}^{j} d_{i+1,k} n_{i,k}(x) - \sum_{k=i+1}^{j} d_{i,k-1} n_{i,k}(x) - \sum_{k=i}^{j-1} (x_k - x_j) d_{i,k} n_{i,k}(x) \right) \\
&= \frac{1}{x_j - x_i} \left(\sum_{k=i+1}^{j} (d_{i+1,k} - d_{i,k-1}) n_{i,k}(x) - \sum_{k=i}^{j-1} (x_k - x_j) d_{i,k} n_{i,k}(x) \right).
\end{aligned}$$

Da $\{n_{i,i}, \dots, n_{i,j}\}$ nach Lemma 6.5 eine Basis ist, folgt per Koeffizientenvergleich

$$\hat{d}_{i,k} = \begin{cases} \frac{d_{i+1,j} - d_{i,j-1}}{x_j - x_i} & \text{falls } k = j, \\ d_{ii} & \text{falls } k = i, \\ \frac{d_{i+1,k} - d_{i,k-1} - (x_k - x_j) d_{i,k}}{x_j - x_i} & \text{ansonsten} \end{cases} \qquad \text{für alle } k \in [i : j].$$

Damit ist der Beweis abgeschlossen, uns fehlte ja nur noch der Fall $k = j$.

Der Vollständigkeit halber möchte ich den letzten Fall $i < k < j$ noch etwas näher betrachten. Wir wissen bereits, dass in diesem Fall $\hat{d}_{i,k} = d_{i,k}$ gilt, so dass die letzte Gleichung durch Subtraktion von $d_{i,k}$ auf beiden Seiten die Form

$$0 = \frac{d_{i+1,k} - d_{i,k-1} - (x_k - x_j + x_j - x_i) d_{i,k}}{x_j - x_i} = \frac{d_{i+1,k} - d_{i,k-1} - (x_k - x_i) d_{i,k}}{x_j - x_i}$$

annimmt, aus der unmittelbar

$$d_{i,k} = \frac{d_{i+1,k} - d_{i,k-1}}{x_k - x_i}$$

folgt, also erhalten wir die Dividierte-Differenzen-Formel für alle $k \in [i+1 : j]$. ∎

Mit Hilfe der Gleichung (6.9 können wir die $d_{i,j}$ analog zu $p_{i,j}(x)$ mit einem Dreiecksschema berechnen:

$$\begin{array}{c|ccccccc}
f_0 & d_{0,0} & & & & & & \\
f_1 & d_{1,1} & d_{0,1} & & & & & \\
f_2 & d_{2,2} & d_{1,2} & d_{0,2} & & & & \\
\vdots & \vdots & \vdots & \vdots & \ddots & & & \\
f_{m-1} & d_{m-1,m-1} & d_{m-2,m-1} & d_{m-3,m-1} & \dots & d_{0,m-1} & & \\
f_m & d_{m,m} & d_{m-1,m} & d_{m-2,m} & \dots & d_{1,m} & d_{0,m}
\end{array}$$

Dieser Ansatz geht auf Newton zurück und ist unter dem deshalb naheliegenden Namen *Newtons dividierte Differenzen* bekannt.

Indem wir die Werte in derselben Weise wie zuvor überschreiben, also in der k-ten Iteration der äußeren Schleife für $i \in [k:m]$ die dividierte Differenz $d_{i-k,i}$ in f_i schreiben, können wir der Reihe nach $d_0 = d_{0,0}, \dots, d_m = d_{0,m}$ effizient mit dem in Abbildung 6.2 dargestellten Algorithmus berechnen.

```
procedure newton_setup(m, var f);
for k = 1,...,m do
  for i = m,...,k do
    f_i ← (f_i − f_{i−1})/(x_i − x_{i−k})
  end for
end for
```

Abbildung 6.2: Berechnung der Koeffizienten $d_0, \dots, d_m$ der Newton-Darstellung des Interpolationspolynoms p. Zu Beginn enthält der Vektor $\mathbf{f} = (f_0, \dots, f_m)$ die zu interpolierenden Werte, am Ende gilt $\mathbf{f} = (d_0, \dots, d_m)$.

Der Rumpf der inneren Schleife benötigt in dieser Variante nur 3 Operationen und wird $(m-k+1)$-mal für jedes $k = 1, \dots, m$ ausgeführt. Damit erhalten wir einen Gesamtaufwand von

$$\sum_{k=1}^{m} 3(m-k+1) = \sum_{\ell=1}^{m} 3\ell = \frac{3}{2} m(m+1)$$

Operationen. Wenn der Koeffizientenvektor $\mathbf{d} = (d_0, \dots, d_m)$ bekannt ist, können wir das Horner-Schema verwenden, um mit dem in Abbildung 6.3 angegebenen Algorithmus das Polynom $p = p_{0,m}$ für ein beliebiges $x \in \mathbb{R}$ auszuwerten.

Offenbar benötigt der Rumpf der Schleife gerade 3 Operationen, und da die Schleife m-mal durchlaufen wird, erhalten wir den erwarteten Gesamtaufwand von $3m$ Operationen. Falls wir uns also in einer Situation befinden, in der wir ein Interpolationspolynom sehr häufig auswerten müssen, empfiehlt es sich, in einer Vorbereitungsphase einmal die dividierten Differenzen des Polynoms zu bestimmen und dann die Auswertungen in lediglich linearem Aufwand durchzuführen.

```
procedure newton_eval(m, x, d, var y);
y ← d_m
for k = m − 1, ..., 0 do
  y ← d_k + (x − x_k)y
end for
```

Abbildung 6.3: Auswertung eines Polynoms in der Newton-Darstellung.

Bemerkung 6.8 (Algebraische Interpretation) *Die dividierten Differenzen lassen sich auch mit Hilfe der in Kapitel 3 vorgestellten Verfahren herleiten: Indem wir die definierende Gleichung (6.7) in den Punkten $x_0, \ldots, x_m$ auswerten, erhalten wir die Gleichungen*

$$f_i = p(x_i) \qquad = \sum_{j=0}^{m} d_j \underbrace{\prod_{k=0}^{j-1} (x_i - x_k)}_{=:a_{ij}} \qquad \text{für alle } i \in [0:m].$$

Da $a_{ij} = 0$ für $i < j$ gilt, erhalten wir das lineare Gleichungssystem

$$\begin{pmatrix} n_{0,0}(x_0) & & & \\ n_{0,0}(x_1) & n_{0,1}(x_1) & & \\ \vdots & \ddots & \ddots & \\ n_{0,0}(x_m) & \cdots & n_{m-1,m}(x_m) & n_{0,m}(x_m) \end{pmatrix} \begin{pmatrix} d_0 \\ d_1 \\ \vdots \\ d_m \end{pmatrix} = \begin{pmatrix} f_0 \\ f_1 \\ \vdots \\ f_m \end{pmatrix},$$

das sich durch Vorwärtseinsetzen lösen lässt.

Wenn wir das System, wie in Kapitel 3 vorgeschlagen, Spalte für Spalte lösen, können wir die einzelnen Spalten mit Lemma 6.6 der Reihe nach berechnen und müssen die gesamte Matrix nicht abspeichern.

Verglichen mit den dividierten Differenzen hat dieser Ansatz den Nachteil, dass ein zusätzlicher Vektor für die jeweils aktuelle Spalte der Matrix benötigt wird, aber den Vorteil, dass lediglich m der auf modernen Prozessoren zeitaufwendigen Divisionen benötigt werden, während die dividierten Differenzen $\frac{m}{2}(m+1)$ Divisionen brauchen.

Bemerkung 6.9 (Bézier-Kurven) *In der Computergrafik ist man häufig an Kurven interessiert, die ästhetisch ansprechend sind und sich gut bearbeiten lassen. Die Polynominterpolation ist dabei nicht unbedingt die beste Wahl: Da die Lagrange-Polynome höheren Grades stark oszillieren (an fast jedem Interpolationspunkt wechselt das Vorzeichen), sind die entstehenden Kurven durch die Werte f_i nur schwer zu steuern.*

Deshalb verwendet man häufig Bézier-Kurven*, die zwar nicht interpolieren, aber andere Vorteile bieten. Konstruiert werden sie in der Regel mit dem* Algorithmus von de Casteljau*, der auf Konvexkombinationen beruht: Für ein $x \in [0,1]$ berechnen wir*

$$q_{i,j}(x) := \begin{cases} f_i & \text{falls } i = j, \\ (1-x)\, q_{i,j-1}(x) + x\, q_{i+1,j}(x) & \text{ansonsten} \end{cases} \qquad \text{für alle } 0 \le i \le j \le m$$

wieder mit Hilfe eines Dreiecksschemas, das Polynom $q := q_{0,m}$ *ist das* Bézier-Polynom *zu den* Kontrollpunkten $f_0, \dots, f_m \in \mathbb{R}$. *In der Praxis verwendet man häufig Kontrollpunkte aus dem Vektorraum* $\mathbb{R}^d$, *dann sind die Polynome* $q_{i,j}$ *vektorwertig.*

Ähnlich wie bei Lagrange-Polynomen lässt sich auch das Bézier-Polynom in geschlossener Form darstellen: Für $i, j \in \mathbb{N}_0$ *mit* $i \le j$ *sind die* Bernstein-Polynome *durch*

$$b_{i,j}(x) := \binom{j}{i} x^i (1-x)^{j-i} \qquad \text{für alle } x \in \mathbb{R}$$

definiert. Sie erfüllen die Gleichung

$$\begin{aligned}(1-x)\, b_{i+1,j}(x) + x\, b_{i,j}(x) &= \binom{j}{i+1} x^{i+1} (1-x)^{j-i} + \binom{j}{i} x^{i+1} (1-x)^{j-i} \\ &= \binom{j+1}{i+1} x^{i+1} (1-x)^{j-i} = b_{i+1,j+1}(x) \qquad \text{für alle } 0 \le i \le j-1,\ x \in \mathbb{R}.\end{aligned}$$

Wir beweisen induktiv

$$q_{i,j}(x) = \sum_{k=i}^{j} f_k\, b_{k-i,j-i}(x) \qquad \text{für alle } 0 \le i \le j \le m,\ x \in \mathbb{R}.$$

Für $i = j$ *ist die Aussage wegen* $b_{0,0} = 1$ *trivial. Wenn wir annehmen, dass* $n \in \mathbb{N}_0$ *so gegeben ist, dass die Aussage für* $i, j \in [0:m]$ *mit* $j - i = n$ *gilt, erhalten wir für* $i, j \in [0:m]$ *mit* $j - i = n+1$ *die Gleichung*

$$\begin{aligned}q_{i,j}(x) &= (1-x)\, q_{i,j-1}(x) + x\, q_{i+1,j}(x) \\ &= (1-x) \sum_{k=i}^{j-1} f_k\, b_{k-i,j-i-1}(x) + x \sum_{k=i+1}^{j} f_k\, b_{k-i-1,j-i-1}(x) \\ &= (1-x) f_i\, b_{0,j-i-1}(x) + \sum_{k=i+1}^{j-1} f_k \left((1-x)\, b_{k-i,j-i-1}(x) + x\, b_{k-i-1,j-i-1}(x)\right) \\ &\qquad + x\, f_j\, b_{j-i-1,j-i-1}(x) \\ &= f_i\, b_{0,j-i}(x) + \sum_{k=i+1}^{j-1} f_k\, b_{k-i,j-i}(x) + f_j\, b_{j-i,j-i}(x) \\ &= \sum_{k=i}^{j} f_k\, b_{k-i,j-i}(x) \qquad \text{für alle } x \in \mathbb{R}.\end{aligned}$$

Diese Darstellung hat nützliche Konsequenzen: Da die Bernstein-Polynome nicht-negativ sind, kann ein Computergrafiker, der einen Kontrollpunkt $f_k \in \mathbb{R}^d$ *auf seinem Monitor in eine Richtung schiebt, erwarten, dass die gesamte Kurve in diese Richtung wandert.*

Eine Weiterentwicklung der Bézier-Polynome sind B-Splines, *die aus* stückweisen *Polynome zusammengesetzt werden, so dass* f_k *lediglich das* lokale *Verhalten der Kurve beschreibt, nicht ihren gesamten Verlauf.*

6.3 Qualität der Approximation

Letztendlich sind wir vor allem daran interessiert, mit Hilfe der Interpolation eine Funktion $f \in C[a,b]$, $a < b$, zu approximieren. Also müssen wir uns auch der Frage widmen, wie genau diese Approximation ist.

Wir gehen davon aus, dass $x_0, \ldots, x_m \in [a,b]$ gilt und dass die Interpolationswerte durch die Auswertung der Funktion f gemäß

$$f_i = f(x_i) \qquad \text{für alle } i \in [0:m]$$

gegeben sind. Als Hilfsmittel für die Darstellung des Approximationsfehlers verwenden wir die in (6.9) eingeführten dividierten Differenzen, deren Abhängigkeit von der Funktion f wir durch die Notation

$$d_{i,j}[f] := \begin{cases} f(x_i) & \text{falls } i = j, \\ \frac{d_{i+1,j}[f] - d_{i,j-1}[f]}{x_j - x_i} & \text{ansonsten} \end{cases} \qquad \text{für alle } f \in C[a,b],\ i,j \in [0:m],\ i \leq j \tag{6.10}$$

zum Ausdruck bringen.

Die dividierten Differenzen stehen in enger Beziehung zu den Ableitungen der zu interpolierenden Funktion f. Falls für ein $j \in [1:m]$ die Funktion auf $[x_{j-1}, x_j]$ stetig differenzierbar ist, besagt der Mittelwertsatz der Differentialrechnung (siehe Erinnerung 5.3) gerade, dass wir ein $\eta \in (x_{j-1}, x_j)$ mit

$$d_{j-1,j}[f] = \frac{d_{j,j}[f] - d_{j-1,j-1}[f]}{x_j - x_{j-1}} = \frac{f(x_j) - f(x_{j-1})}{x_j - x_{j-1}} = f'(\eta)$$

finden können. Eine ähnliche Aussage werden wir nun für höhere Ableitungen herleiten.

Lemma 6.10 (Nullstellen) *Sei $k \in \mathbb{N}$, und sei $g \in C^k[a,b]$ eine Funktion, die im Intervall $[a,b]$ mindestens $k+1$ verschiedene Nullstellen besitzt. Dann besitzt die k-te Ableitung $g^{(k)}$ mindestens eine Nullstelle in diesem Intervall.*

Beweis. Per Induktion über $k \in \mathbb{N}$.

Induktionsanfang: Sei $g \in C^1[a,b]$ eine Funktion, die in $[a,b]$ mindestens zwei Nullstellen $a \leq x_0 < x_1 \leq b$ besitzt. Nach dem Satz von Rolle besitzt dann g' mindestens eine Nullstelle $\eta \in (x_0, x_1)$.

Induktionsvoraussetzung: Gelte die Aussage nun für $k \in \mathbb{N}$.

Induktionsschritt: Sei $g \in C^{k+1}[a,b]$ eine Funktion mit mindestens $k+2$ verschiedenen Nullstellen $x_0, \ldots, x_{k+1} \in [a,b]$, die wir gemäß

$$a \leq x_0 < x_1 < \ldots < x_{k+1} \leq b$$

als sortiert voraussetzen dürfen. Nach dem Satz von Rolle muss zwischen je zwei dieser Nullstellen eine Nullstelle der Ableitung g' liegen, wir finden also $\eta_0, \ldots, \eta_k$ mit

$$x_i < \eta_i < x_{i+1}, \qquad g'(\eta_i) = 0 \qquad \text{für alle } i \in [0:k].$$

Demnach besitzt g' mindestens $k+1$ verschiedene Nullstellen in (a,b). Indem wir die Induktionsvoraussetzung auf g' anwenden, finden wir eine Nullstelle $\eta \in (a,b)$ von $(g')^{(k)} = g^{(k+1)}$. ∎

Indem wir die richtige Nullstelle einer Hilfsfunktion wählen, können wir nun eine beliebige dividierte Differenz mit Hilfe der Ableitung der Funktion f beschreiben:

Lemma 6.11 (Dividierte Differenz) *Seien $i,j \in [0:m]$ gegeben. Falls $f \in C^{j-i}[a,b]$ gilt, gibt es ein $\eta \in (a,b)$ mit*

$$d_{i,j}[f] = \frac{f^{(j-i)}(\eta)}{(j-i)!}.$$

Beweis. Als Vorbereitung beweisen wir $n_{i,j}^{(j-i)} = (j-i)!$ per Induktion über $j-i$.

Induktionsanfang: Für $j = i$ ist die Aussage trivial.

Induktionsvoraussetzung: Sei $n \in \mathbb{N}_0$ so gegeben, dass die Aussage für alle $i,j \in [0:m]$ mit $j - i = n$ gilt.

Induktionsschritt: Seien $i,j \in [0:m]$ gegeben mit $j - i = n+1$. Wir definieren $p(x) := x - x_i$ und erhalten mit Lemma 6.5 und der Leibniz-Regel

$$\begin{aligned} n_{i,j}^{(j-i)}(x) = (p\, n_{i+1,j})^{(j-i)}(x) &= \sum_{k=0}^{j-i} \binom{j-i}{k} p^{(k)}(x)\, n_{i+1,j}^{(j-i-k)}(x) \\ &= (x - x_i)\, n_{i+1,j}^{(j-i)}(x) + (j-i)\, n_{i+1,j}^{(j-i-1)}(x) \qquad \text{für alle } x \in \mathbb{R}. \end{aligned}$$

Da $n_{i+1,j}^{(j-i)} = 0$ gilt und wir $n_{i+1,j}^{(j-i-1)} = (j-i-1)!$ nach Induktionsvoraussetzung haben, folgt die Behauptung.

Sei $\ell := j - i$. Wir untersuchen die Funktion $r := f - p_{i,j}$. Da f eine ℓ-mal stetig differenzierbare Funktion und $p_{i,j}$ ein Polynom ist, ist auch r ℓ-mal stetig differenzierbar.

Nach Definition von $p_{i,j}$ gilt

$$r(x_k) = f(x_k) - p_{i,j}(x_k) = f_k - f_k = 0 \qquad \text{für alle } k \in [i:j],$$

also besitzt r mindestens $\ell + 1$ Nullstellen in $[a,b]$. Aus Lemma 6.10 folgt, dass $r^{(\ell)}$ noch mindestens eine Nullstelle $\eta \in (a,b)$ besitzen muss. Für diese Nullstelle gilt

$$0 = r^{(\ell)}(\eta) = f^{(\ell)}(\eta) - p_{i,j}^{(\ell)}(\eta).$$

Wir müssen also nun die ℓ-te Ableitung des Polynoms $p_{i,j}$ untersuchen.

Mit Lemma 6.7 gilt

$$p_{i,j}(x) = p_{i,j-1}(x) + d_{i,j}[f] n_{i,j}(x) \qquad \text{für alle } x \in \mathbb{R}.$$

Da $p_{i,j-1} \in \Pi_{j-i-1}$ gilt, haben wir $p_{i,j-1}^{(\ell)} \equiv 0$, also verschwindet der erste Term. Für den zweiten Term benutzen wir unsere Vorbetrachtung und erhalten

$$p_{i,j}^{(\ell)}(\eta) = \ell!\; d_{i,j}[f].$$

Damit folgt

$$f^{(\ell)}(\eta) = (j-i)!\, d_{i,j}[f],$$

und Dividieren durch die Fakultät ergibt die gesuchte Gleichung. ∎

Aus dieser Aussage erhalten wir nun mit einem einfachen Trick die entscheidende Aussage über den Approximationsfehler:

Satz 6.12 (Interpolationsfehler) *Sei $x \in [a,b]$. Falls $f \in C^{m+1}[a,b]$ gilt, gibt es ein $\eta \in (a,b)$ mit*

$$f(x) - p(x) = (x-x_0)\dots(x-x_m)\frac{f^{(m+1)}(\eta)}{(m+1)!}.$$

Beweis. Für $x \in \{x_0,\dots,x_m\}$ ist die Aussage offensichtlich: Sowohl die linke als auch die rechte Seite sind gleich null.

Sei nun also $x \in [a,b] \setminus \{x_0,\dots,x_m\}$. Dann können wir $x_{m+1} := x$ als weiteren Interpolationspunkt mit $f_{m+1} := f(x_{m+1})$ hinzunehmen, und das korrespondierende Interpolationspolynom $p_{0,m+1} \in \Pi_{m+1}$ erfüllt insbesondere auch $p_{0,m+1}(x) = f(x)$.

Gemäß (6.4) erhalten wir

$$f(x) - p(x) = p_{0,m+1}(x) - p_{0,m}(x) = (x-x_0)\dots(x-x_m)d_{0,m+1}[f],$$

und aus Lemma 6.11 folgt die Behauptung. ∎

Bei dieser Darstellung des Fehlers muss man sehr sorgfältig auf die Abhängigkeiten der einzelnen Parameter achten: Insbesondere ist η nicht nur von f und $x_0,\dots,x_m$, sondern auch von x abhängig, und unsere Existenzaussage bietet nur wenige Informationen über die Natur dieser Abhängigkeit.

Für die Praxis ist eine auf einem nicht weiter bestimmten Zwischenpunkt η basierende Fehlerdarstellung häufig wenig hilfreich, deshalb schätzen wir den Fehler ab, indem wir zum Maximum übergehen:

Folgerung 6.13 (Interpolationsfehler) *Sei $f \in C^{m+1}[a,b]$ und sei $p \in \Pi_m$ das Interpolationspolynom in den Punkten $x_0,\dots,x_m \in [a,b]$. Mit*

$$\omega\colon \mathbb{R} \to \mathbb{R}, \qquad x \mapsto (x-x_0)\dots(x-x_m),$$

bezeichnen wir das sogenannte Stützstellenpolynom *und erhalten die Abschätzung*

$$\|f-p\|_{\infty,[a,b]} \leq \|\omega\|_{\infty,[a,b]} \frac{\|f^{(m+1)}\|_{\infty,[a,b]}}{(m+1)!} \tag{6.11}$$

des maximalen Interpolationsfehlers.

Beweis. Folgt direkt aus Satz 6.12. ∎

Die Abschätzung (6.11) ist insofern besonders nützlich, als sie es uns ermöglicht, die Eigenschaften des Interpolationsverfahrens und die Eigenschaften der zu interpolierenden Funktion separat zu untersuchen:

Bemerkung 6.14 (Stützstellenpolynom) *Der Term* $\|\omega\|_{\infty,[a,b]}$ *hängt nur von der Wahl der Interpolationspunkte ab. Eine einfache obere Abschätzung können wir erhalten, indem wir* $|x - x_i| \leq b - a$ *für alle* $x_i \in [a, b]$ *ausnutzen und so*

$$\|\omega\|_{\infty,[a,b]} \leq \prod_{i=0}^{m} |x - x_i| \leq (b - a)^{m+1} \tag{6.12}$$

erhalten. Es stellt sich die Frage, ob man durch eine geschickte Wahl der Interpolationspunkte dafür sorgen kann, dass dieser Term deutlich kleiner wird. Damit werden wir uns im nächsten Abschnitt eingehend befassen.

Bemerkung 6.15 (Regularität) *Der Term*

$$\frac{\|f^{(m+1)}\|_{\infty,[a,b]}}{(m+1)!}$$

beschreibt, wie „glatt" die Funktion f ist: Je kleiner die Norm der $(m+1)$*-ten Ableitung ist, um so besser lässt sich f approximieren. In der Praxis hat sich für Aussagen über die Ableitungen einer Funktion den Begriff der* Regularität *eingebürgert. Eine Funktion besitzt* hohe Regularität*, falls hohe Ableitungen sich durch kleine Konstanten beschränken lassen, und* niedrige Regularität*, falls die Konstanten groß oder die hohen Ableitungen gar nicht definiert sind.*

6.4 Tschebyscheff-Interpolation

Unser Ziel ist es, die Stützstellen $x_0, \ldots, x_m$ so zu wählen, dass das zugehörige Stützstellenpolynom ω möglichst klein wird, denn dann ist auch der Interpolationsfehler klein.

Tatsächlich ist es möglich, die Interpolationspunkte so zu wählen, dass $\|\omega\|_{\infty,[a,b]}$ minimal wird. Die Analyse lässt sich wesentlich vereinfachen, indem wir uns zunächst auf ein *Referenzintervall* $[-1, 1]$ konzentrieren und später allgemeine Intervalle darauf zurückführen.

Grundlage unserer Untersuchung ist eine Familie von Polynomen, die sich durch für unsere Zwecke besonders nützliche Eigenschaften auszeichnen:

Definition 6.16 (Tschebyscheff-Polynome) *Wir definieren eine Familie von Polynomen* $T_m \in \Pi_m$ *durch*

$$T_m(x) = \begin{cases} 1 & \text{falls } m = 0, \\ x & \text{falls } m = 1, \\ 2xT_{m-1}(x) - T_{m-2}(x) & \text{ansonsten} \end{cases} \qquad \text{für alle } m \in \mathbb{N}_0, x \in \mathbb{R}.$$

Das Polynom T_m *nennen wir das* m*-te* Tschebyscheff-Polynom*.*

Ihre besonderen Eigenschaften verdanken die Tschebyscheff-Polynome ihrer engen Beziehung zu trigonometrischen Funktionen:

Lemma 6.17 (Trigonometrische Darstellung) *Sei $m \in \mathbb{N}_0$. Es gilt*

$$T_m(x) = \cos(m \arccos x) \qquad \text{für alle } x \in [-1,1].$$

Beweis. Wir definieren $C_m \in C[-1,1]$ durch

$$C_m(x) := \cos(m \arccos x) \qquad \text{für alle } x \in [-1,1]$$

und müssen $T_m = C_m$ beweisen. Dazu verwenden wir ein abschnittsweise Induktion.

Induktionsanfang: Für $m = 0$ und $m = 1$ ist die Gleichung offensichtlich erfüllt.

Induktionsvoraussetzung: Sei nun also $m \in \mathbb{N}$ so gewählt, dass $T_m = C_m$ und $T_{m-1} = C_{m-1}$ gelten.

Induktionsschritt: Um $T_{m+1} = C_{m+1}$ zu beweisen verwenden wir das Additionstheorem

$$\cos(\alpha + \beta) = \cos(\alpha)\cos(\beta) - \sin(\alpha)\sin(\beta).$$

Für ein $x \in [-1,1]$ fixieren wir $\xi = \arccos x$ und erhalten

$$\begin{aligned} C_{m+1}(x) + C_{m-1}(x) &= \cos(m\xi + \xi) + \cos(m\xi - \xi) \\ &= \cos(m\xi)\cos(\xi) - \sin(m\xi)\sin(\xi) \\ &\qquad + \cos(m\xi)\cos(-\xi) - \sin(m\xi)\sin(-\xi) \\ &= 2\cos(m\xi)\cos(\xi) - \sin(m\xi)\sin(\xi) + \sin(m\xi)\sin(\xi) \\ &= 2\cos(m\xi)\cos(\xi) = 2xC_m(x), \end{aligned}$$

also folgt dank der Induktionsvoraussetzung

$$C_{m+1}(x) = 2xC_m(x) - C_{m-1}(x) = 2xT_m(x) - T_{m-1}(x) = T_{m+1}(x),$$

und damit ist die Induktion vollständig. ∎

Aus dieser Darstellung der Tschebyscheff-Polynome folgt unmittelbar

$$\|T_{m+1}\|_{\infty,[-1,1]} = 1. \tag{6.13}$$

Wir würden nun gerne aus T_{m+1} das Stützstellenpolynom ω konstruieren. Offenbar gilt $\omega \in \Pi_{m+1}$ und

$$\omega(x) = (x - x_0)\dots(x - x_m) = \alpha_0 + \alpha_1 x + \dots + \alpha_m x^m + x^{m+1} \qquad \text{für alle } x \in \mathbb{R}$$

mit geeigneten $\alpha_0, \dots, \alpha_m \in \mathbb{R}$, der $(m+1)$-te Koeffizient ist also gerade gleich eins. Derartige Polynome nennen wir *normiert*. Die Differenz zweier normierter Polynome aus Π_{m+1} ist offenbar ein Polynom aus Π_m.

Aus der Induktion in Definition 6.16 folgt, dass für $m \in \mathbb{N}_0$ der $(m+1)$-te Koeffizient von T_{m+1} gerade gleich 2^m ist. Also muss $2^{-m}T_{m+1}$ ein normiertes Polynom aus Π_{m+1} sein. Dank Lemma 6.17 können wir die Nullstellen des Tschebyscheff-Polynoms T_{m+1} explizit berechnen: Mit den durch

$$\hat{x}_i := \cos\left(\pi \frac{2(m-i)+1}{2m+2}\right) \qquad \text{für alle } i \in [0:m] \tag{6.14}$$

definierten *Tschebyscheff-Punkten* in $[-1,1]$ gilt

$$\begin{aligned} T_{m+1}(\hat{x}_i) &= \cos\left((m+1)\pi\frac{2(m-i)+1}{2m+2}\right) \\ &= \cos(\pi(m-i)+\pi/2) = 0 \qquad \text{für alle } i \in [0:m]. \end{aligned}$$

Da der Cosinus in $[0,\pi]$ streng monoton fällt, sind diese Nullstellen paarweise verschieden. Das zugehörige Stützstellenpolynom

$$\hat{\omega}(x) = (x-\hat{x}_0)\dots(x-\hat{x}_m) \qquad \text{für alle } x \in \mathbb{R}. \tag{6.15}$$

besitzt dieselben Nullstellen wir T_{m+1}, also auch dieselben wie das normierte Polynom $2^{-m}T_{m+1}$. Da $\hat{\omega} - 2^{-m}T_{m+1} \in \Pi_m$ gilt und die Differenz in den $m+1$ Nullstellen verschwindet, folgt mit dem Identitätssatz $\hat{\omega} = 2^{-m}T_{m+1}$.

Folgerung 6.18 (Tschebyscheff-Interpolation) *Sei $f \in C^{m+1}[-1,1]$, und sei $p \in \Pi_m$ das Interpolationspolynom von f in den Punkten $\hat{x}_0,\dots,\hat{x}_m \in [-1,1]$ aus (6.14). Dann gilt*

$$\|f-p\|_{\infty,[-1,1]} \leq 2^{-m}\frac{\|f^{(m+1)}\|_{\infty,[-1,1]}}{(m+1)!}.$$

Beweis. Wegen (6.15) ist $\hat{\omega} = 2^{-m}T_{m+1}$ das Stützstellenpolynom zu den Punkten $\hat{x}_0,\dots,\hat{x}_m$. Dank (6.13) gilt $\|\hat{\omega}\|_{\infty,[-1,1]} = 2^{-m}$, und durch Einsetzen in Folgerung 6.13 erhalten wir das gewünschte Resultat. ∎

Die Verwendung der Nullstellen $\hat{x}_0,\dots,\hat{x}_m$ des Tschebyscheff-Polynoms T_{m+1} zur Definition eines Interpolationsverfahrens führt nicht nur zu einer relativ guten Fehlerabschätzung, man kann sogar beweisen, dass es keine andere Wahl der Interpolationspunkte geben kann, für die $\|\omega\|_{\infty,[-1,1]}$ kleiner wird:

Lemma 6.19 (Optimalität) *Seien $x_0 < x_1 < \ldots < x_m$ beliebige Punkte, und sei ω das zugehörige Stützstellenpolynom. Dann gilt*

$$\|\omega\|_{\infty,[-1,1]} \geq 2^{-m} = \|\hat{\omega}\|_{\infty,[-1,1]}.$$

Beweis. Wir führen den Beweis per Kontraposition: Wir wählen ein Polynom $\omega \in \Pi_{m+1}$ mit $\|\omega\|_{\infty,[-1,1]} < 2^{-m}$ und werden nun zeigen, dass sein $(m+1)$-ter Koeffizient nicht gleich eins ist, ω also kein Stützstellenpolynom sein kann.

Dazu wählen wir die Punkte

$$\xi_i := \cos\left(\pi\frac{2i}{2m+2}\right) \qquad \text{für alle } i \in [0:m+1].$$

Es gilt

$$T_{m+1}(\xi_i) = \cos\left((m+1)\pi\frac{2i}{2m+2}\right) = \cos(\pi i) = (-1)^i \qquad \text{für alle } i \in [0:m+1].$$

Nach Voraussetzung ist

$$|\omega(\xi_i)| < 2^{-m} = 2^{-m}|T_{m+1}(\xi_i)| = |\hat{\omega}(\xi_i)| \qquad \text{für alle } i \in [0 : m+1],$$

also gilt für $r := \omega - \hat{\omega} \in \Pi_{m+1}$

$$\begin{aligned} r(\xi_i) &< 0 && \text{für alle geraden } i \in [0 : m+1], \\ r(\xi_i) &> 0 && \text{für alle ungeraden } i \in [0 : m+1]. \end{aligned}$$

Nach dem Zwischenwertsatz für stetige Funktionen muss somit in jedem Intervall (ξ_i, ξ_{i+1}) das Polynom r eine Nullstelle z_i besitzen, wir haben also $m+1$ Nullstellen des Polynoms r konstruiert.

Wäre nun $r \in \Pi_m$, würde mit dem Identitätssatz $r = 0$ folgen. Per Kontraposition schließen wir aus $r(\xi_0) \neq 0$, dass $r \notin \Pi_m$ gilt, also kann ω nicht normiert sein. ∎

Mit einer etwas technischen Modifikation des Beweises kann man zeigen, dass nicht nur kein normiertes Polynom $(m+1)$-ten Grades eine geringere Norm als 2^{-m} besitzen kann, sondern sogar, dass $\hat{\omega}$ das *einzige* Polynom ist, das diese Eigenschaft besitzt.

Auf dem Intervall $[-1, 1]$ sind also die Punkte $\hat{x}_0, \ldots, \hat{x}_m$ aus (6.14) die bestmögliche Wahl, wenn es darum geht $\|\hat{\omega}\|_{\infty,[-1,1]}$ zu minimieren.

Mit diesem Wissen können wir uns allgemeinen Intervallen zuwenden. Seien nun $a, b \in \mathbb{R}$ mit $a < b$ gegeben. Wir transformieren $[-1, 1]$ mit Hilfe der Abbildung

$$\Phi_{[a,b]} \colon [-1,1] \to [a,b], \qquad x \mapsto \frac{b+a}{2} + \frac{b-a}{2}x, \tag{6.16}$$

das Referenzintervall in das gewünschte allgemeine Intervall. Da

$$\Phi'_{[a,b]}(x) = \frac{b-a}{2} \qquad \text{für alle } x \in [-1,1]$$

gilt, ist $\Phi_{[a,b]}$ streng monoton wachsend und damit insbesondere injektiv. Mit Hilfe dieser Transformation können wir aus Funktionen auf $[a, b]$ Funktionen auf $[-1, 1]$ machen und dabei die für unsere Fehlerabschätzungen wichtigen Ableitungen unter Kontrolle halten:

Lemma 6.20 (Transformierte Ableitungen) *Seien $a, b \in \mathbb{R}$ mit $a < b$ gegeben, seien $m \in \mathbb{N}_0$ und $f \in C^m[a, b]$. Dann gilt*

$$(f \circ \Phi_{[a,b]})^{(m)}(x) = \left(\frac{b-a}{2}\right)^m f^{(m)} \circ \Phi_{[a,b]}(x) \qquad \textit{für alle } x \in [-1,1]. \tag{6.17}$$

Beweis. Per Induktion über m.

Induktionsanfang: Für $m = 0$ ist die Aussage trivial.

Induktionsvoraussetzung: Sei $m \in \mathbb{N}_0$ so gegeben, dass (6.17) gilt.

Induktionsschritt: Sei $f \in C^{m+1}[a, b]$ und $x \in [-1, 1]$. Dann gilt nach Induktionsvoraussetzung

$$(f \circ \Phi_{[a,b]})^{(m)}(x) = \left(\frac{b-a}{2}\right)^m f^{(m)} \circ \Phi_{[a,b]}(x).$$

Aus der Kettenregel folgt

$$\begin{aligned}(f \circ \Phi_{[a,b]})^{(m+1)}(x) &= \left(\frac{b-a}{2}\right)^m (f^{(m)} \circ \Phi_{[a,b]})'(x) \\ &= \left(\frac{b-a}{2}\right)^m \Phi'_{[a,b]}(x) f^{(m+1)} \circ \Phi_{[a,b]}(x) \\ &= \left(\frac{b-a}{2}\right)^{m+1} f^{(m+1)} \circ \Phi_{[a,b]}(x),\end{aligned}$$

und damit ist die Induktion vollständig. ∎

Wir definieren die transformierten Interpolationspunkte durch

$$x_i := \Phi_{[a,b]}(\hat{x}_i) = \frac{b+a}{2} + \frac{b-a}{2}\hat{x}_i \qquad \text{für alle } i \in [0:m].$$

Da $\Phi_{[a,b]}$ streng monoton wachsend ist, gilt $a \leq x_0 < x_1 < \ldots < x_m \leq b$, so dass wir wie zuvor Interpolationspolynome konstruieren können.

Satz 6.21 (Allgemeine Tschebyscheff-Interpolation) *Sei $f \in C^{m+1}[a,b]$ und sei $p \in \Pi_m$ das Interpolationspolynom in den transformierten Punkten $x_0, \ldots, x_m \in [a,b]$. Dann gilt*

$$\|f - p\|_{\infty,[a,b]} \leq 2\left(\frac{b-a}{4}\right)^{m+1} \frac{\|f^{(m+1)}\|_{\infty,[a,b]}}{(m+1)!}.$$

Beweis. Wir definieren $\hat{f} := f \circ \Phi_{[a,b]} \in C^{m+1}[-1,1]$ und konstruieren ein Interpolationspolynom $\hat{p} \in \Pi_m$ in den Tschebyscheff-Punkten $\hat{x}_0, \ldots, \hat{x}_m$.

Wir vergleichen $p \circ \Phi_{[a,b]}$ mit $\hat{p}$. Es gilt

$$\begin{aligned}p \circ \Phi_{[a,b]}(\hat{x}_i) &= p(\Phi_{[a,b]}(\hat{x}_i)) = p(x_i) = f(x_i) \\ &= f(\Phi_{[a,b]}(\hat{x}_i)) = \hat{f}(\hat{x}_i) = \hat{p}(\hat{x}_i) \qquad \text{für alle } i \in [0:m],\end{aligned}$$

also ist $p \circ \Phi_{[a,b]}$ ein Polynom aus Π_m, das in $m+1$ Punkten mit $\hat{p}$ übereinstimmt. Daraus folgt mit dem Identitätssatz $\hat{p} = p \circ \Phi_{[a,b]}$.

Da das Bild von $\Phi_{[a,b]}$ gerade das Intervall $[a,b]$ ist, gilt

$$\|f - p\|_{\infty,[a,b]} = \|f \circ \Phi_{[a,b]} - p \circ \Phi_{[a,b]}\|_{\infty,[-1,1]} = \|\hat{f} - \hat{p}\|_{\infty,[-1,1]},$$

und mit Folgerung 6.18 und Lemma 6.20 erhalten wir

$$\begin{aligned}\|f - p\|_{\infty,[a,b]} = \|\hat{f} - \hat{p}\|_{\infty,[-1,1]} &\leq 2^{-m} \frac{\|\hat{f}^{(m+1)}\|_{\infty,[-1,1]}}{(m+1)!} \\ &= 2^{-m}\left(\frac{b-a}{2}\right)^{m+1} \frac{\|f^{(m+1)} \circ \Phi_{[a,b]}\|_{\infty,[-1,1]}}{(m+1)!} \\ &= 2\left(\frac{b-a}{4}\right)^{m+1} \frac{\|f^{(m+1)}\|_{\infty,[a,b]}}{(m+1)!}.\end{aligned}$$

Das ist die gewünschte Abschätzung. ∎

6.5 Stabilität und Bestapproximation

Die bisher von uns untersuchten Fehlerdarstellungen und -abschätzungen stellten relativ hohe Anforderungen: Die Abschätzung des Interpolationsfehlers mit einem Polynom aus Π_m gelingt mit den bisherigen Aussagen nur, wenn die zu interpolierende Funktion $(m+1)$-mal stetig differenzierbar ist. Wir werden nun Techniken untersuchen, mit denen sich wesentlich allgemeinere Aussagen gewinnen lassen.

Zu gegebenen Interpolationspunkten $x_0 < x_1 < \ldots < x_m$ mit Lagrange-Polynomen $\ell_0, \ldots, \ell_m \in \Pi_m$ führen wir den *Interpolationsoperator* ein, der eine beliebige Funktion auf ihr Interpolationspolynom abbildet. Dank (6.2) ist dieser Operator linear und besitzt die Darstellung

$$\mathfrak{I}\colon C[a,b] \to \Pi_m, \qquad f \mapsto \sum_{i=0}^{m} f(x_i)\ell_i.$$

Das Interpolationspolynom einer Funktion $f \in C[a,b]$ kann mit seiner Hilfe als $\mathfrak{I}[f] \in \Pi_m$ geschrieben werden.

Den Ausgangspunkt der allgemeinen Theorie bildet die Beobachtung, dass bei der Interpolation eines Polynoms $q \in \Pi_m$ in $m+1$ Punkten das Polynom reproduziert wird: Interpolationspolynom $\mathfrak{I}[q]$ und Polynom q stimmen in den $m+1$ Interpolationspunkten überein, müssen also nach Identitätssatz gleich sein. Diese Eigenschaft lässt sich als

$$\mathfrak{I}[q] = q \qquad \text{für alle } q \in \Pi_m \tag{6.18}$$

schreiben und impliziert $\mathfrak{I}^2 = \mathfrak{I} \circ \mathfrak{I} = \mathfrak{I}$, der Interpolationsoperator ist also eine Projektion auf den Raum Π_m der Polynome.

Aus dieser Eigenschaft der Interpolation können wir eine allgemeine Fehlerabschätzung gewinnen: Wir ersetzen f durch ein Polynom q, das die Funktion gut approximiert. Sofern der Interpolationsoperator nicht allzu empfindlich auf diese Störung reagiert, folgt, dass q eine gute Näherung der Interpolationspolynoms ist, und per Dreiecksungleichung muss es dann auch eine gute Näherung der Funktion f sein.

Die „Empfindlichkeit“ des Interpolationsoperators gegenüber Störungen können wir durch seine Operatornorm quantifizieren.

Definition 6.22 (Lebesgue-Konstante) *Die Zahl*

$$\Lambda := \max\left\{ \sum_{i=0}^{m} |\ell_i(x)| \;:\; x \in [a,b] \right\}$$

heißt die Lebesgue-Konstante *oder* Stabilitätskonstante *des Interpolationsoperators* $\mathfrak{I}$.

Lemma 6.23 (Stabilität) *Sei* Λ *die Lebesgue-Konstante des Interpolationsoperators* $\mathfrak{I}$. *Es gilt*

$$\|\mathfrak{I}[f]\|_{\infty,[a,b]} \leq \Lambda \|f\|_{\infty,[a,b]} \qquad \textit{für alle } f \in C[a,b].$$

Beweis. Die Aussage folgt direkt aus der Darstellung (6.2) des Interpolationspolynoms durch die Lagrange-Polynome: Für alle $x \in [a,b]$ gilt dank der Dreiecksungleichung

$$|\mathfrak{I}[f](x)| = \left|\sum_{i=0}^{m} f(x_i)\ell_i(x)\right| \leq \sum_{i=0}^{m} |f(x_i)||\ell_i(x)| \leq \|f\|_{\infty,[a,b]} \sum_{i=0}^{m} |\ell_i(x)| \leq \Lambda \|f\|_{\infty,[a,b]}.$$

Also muss auch das Maximum $\|\mathfrak{I}[f]\|_{\infty,[a,b]}$ diese Ungleichung erfüllen. ∎

Die Lebesgue-Konstante ist nicht nur eine Konstante, die die Abschätzung aus Lemma 6.23 erfüllt, sie ist die beste solche Konstante, siehe Übungsaufgabe 6.27.

Beispiel 6.24 (Tschebyscheff-Interpolation) *Es lässt sich (mit erheblichem Aufwand, siehe* The Chebyshev Polynomials *von T. Rivlin, Wiley New York, 1990) zeigen, dass für die Tschebyscheff-Interpolation*

$$\Lambda \leq \frac{2}{\pi}\log(m+1) + 1$$

gilt, und dass diese Lebesgue-Konstante relativ nahe an der liegt, die man bei optimaler Wahl der Interpolationspunkte theoretisch erreichen könnte.

Indem wir die Stabilitätsaussage von Lemma 6.23 mit der Projektionseigenschaft (6.18) kombinieren, erhalten wir die folgende *Bestapproximationsaussage*:

Satz 6.25 (Bestapproximation) *Sei eine Funktion $f \in C[a,b]$ gegeben, und sei Λ die Lebesgue-Konstante des Interpolationsoperator $\mathfrak{I}$. Dann gilt*

$$\|f - \mathfrak{I}[f]\|_{\infty,[a,b]} \leq (1+\Lambda)\|f-q\|_{\infty,[a,b]} \qquad \text{für alle } q \in \Pi_m,$$

der Interpolationsfehler unterscheidet sich also nur um einen festen Faktor von dem bestmöglichen Fehler.

Beweis. Sei $q \in \Pi_m$. Dank (6.18) gilt $\mathfrak{I}[q] = q$, und wir erhalten

$$\begin{aligned}\|f - \mathfrak{I}[f]\|_{\infty,[a,b]} &= \|f - q + q - \mathfrak{I}[f]\|_{\infty,[a,b]} = \|f - q + \mathfrak{I}[q] - \mathfrak{I}[f]\|_{\infty,[a,b]} \\ &\leq \|f-q\|_{\infty,[a,b]} + \|\mathfrak{I}[f-q]\|_{\infty[a,b]}.\end{aligned}$$

Mit Hilfe von Lemma 6.23 folgt daraus

$$\|f - \mathfrak{I}[f]\|_{\infty,[a,b]} \leq \|f-q\|_{\infty,[a,b]} + \Lambda\|f-q\|_{\infty,[a,b]} = (1+\Lambda)\|f-q\|_{\infty,[a,b]},$$

also die zu zeigende Abschätzung. ∎

Fehlerabschätzungen wie diese sind außerordentlich nützlich: Einerseits werden wieder die Eigenschaften des Interpolationsverfahrens (ausgedrückt durch die Lebesgue-Konstante Λ) und die Eigenschaften der Funktion (ausgedrückt durch $\|f-q\|_{\infty,[a,b]}$) getrennt analysiert, andererseits können wir das approximierende Polynom q völlig beliebig wählen und so auch noch in Situationen etwas beweisen, in denen Aussagen wie die von Satz 6.21 an mangelnder Differenzierbarkeit von f scheitern.

Beispiel 6.26 (Niedrige Regularität) *Mit Folgerung 6.13 können wir nur dann eine Aussage über den Interpolationsfehler gewinnen, wenn $f \in C^{m+1}[a,b]$ gilt. Mit Satz 6.25 hingegen können wir auch dann noch eine Aussage treffen, falls $f \in C^n[a,b]$ für ein $n \in [1:m+1]$ gelten sollte: Wir verwenden Satz 6.21 als Existenzresultat. Da das Tschebyscheff-Interpolationspolynom $(n-1)$-ten Grades $p \in \Pi_{n-1}$ die Abschätzung*

$$\|f-p\|_{\infty,[a,b]} \leq 2\left(\frac{b-a}{4}\right)^n \frac{\|f^{(n)}\|_{\infty,[a,b]}}{n!}$$

erfüllt und da $p \in \Pi_{n-1} \subseteq \Pi_m$ gilt, muss wegen Satz 6.25 auch für unseren allgemeinen Interpolationsoperator die Abschätzung

$$\|f-\mathfrak{I}[f]\|_{\infty,[a,b]} \leq (1+\Lambda)\|f-p\|_{\infty,[a,b]} \leq 2(1+\Lambda)\left(\frac{b-a}{4}\right)^n \frac{\|f^{(n)}\|_{\infty,[a,b]}}{n!}$$

gelten. Auch bei reduzierter Regularität können wir also noch Aussagen über den Fehler treffen, und der Grad m des Interpolationsoperators gibt nur eine obere Grenze für die höchste nutzbare Regularität an.

Übungsaufgabe 6.27 (Lebesgue-Konstante) *Sei Λ die in Definition 6.22 eingeführte Lebesgue-Konstante.*

Beweisen Sie, dass es eine stetige Funktion $f \in C[a,b]$ gibt mit

$$\|\mathfrak{I}[f]\|_{\infty,[a,b]} = \Lambda\|f\|_{\infty,[a,b]}.$$

Hinweis: *Wie wäre es mit einer stückweise linearen Funktion, die in den Interpolationspunkten x_i die Werte 1 oder -1 annimmt?*

Übungsaufgabe 6.28 (Approximation von Funktionalen) *Sei J eine stetige lineare Funktion, die $C[a,b]$ auf $\mathbb{R}$ abbildet. Um $J(f)$ für eine Funktion $f \in C[a,b]$ zu approximieren, ersetzen wir f durch das Interpolationspolynom $\mathfrak{I}[f]$ und erhalten*

$$J(\mathfrak{I}[f]) = J\left(\sum_{i=0}^m f(x_i)\ell_i\right) = \sum_{i=0}^m f(x_i)J(\ell_i) \qquad \text{für alle } f \in C[a,b],$$

also definieren wir die Gewichte $w_i := J(\ell_i)$ *für alle $i \in [0:m]$ und wollen das approximative Funktional*

$$\widetilde{J}\colon C[a,b] \mapsto \mathbb{R}, \qquad f \mapsto \sum_{i=0}^m w_i f(x_i)$$

untersuchen. Wir setzen

$$C := \sup\left\{\frac{|J(f)|}{\|f\|_{\infty,[a,b]}} \;:\; f \in C[a,b]\right\}, \qquad \widetilde{C} := \sum_{i=0}^m |w_i|.$$

Beweisen Sie

$$|J(f)-\widetilde{J}(f)| \leq (C+\widetilde{C})\|f-q\|_{\infty,[a,b]} \qquad \text{für alle } f \in C[a,b],\ q \in \Pi_m.$$

6.6 Extrapolation

Traditionell verwendet man die Interpolation, um *zwischen* bekannten Werten einer Funktion Näherungswerte zu konstruieren. Wir können sie aber auch benutzen, um Näherungswerte *außerhalb* des durch die Interpolationspunkte beschriebenen Intervalls zu berechnen. Dann spricht man von *Extrapolation*.

Eine zentrale Anwendung der Extrapolation ist die Verbesserung von Näherungslösungen: Häufig hängen numerische Verfahren von Parametern ab, die die Genauigkeit der Näherung bestimmen, und die exakte Lösung ergibt sich, wenn man diese Parameter gegen null streben lässt. Unser Ziel ist es, den Grenzwert zu approximieren, ohne das zugrundeliegende Verfahren zu modifizieren.

Wir beschreiben dabei das Näherungsverfahren abstrakt durch eine Funktion

$$g\colon [0, h_{\max}] \to \mathbb{R},$$

die jedem Parameter $h \in (0, h_{\max}]$ die entsprechende Näherungslösung $g(h)$ zuordnet. Die Funktion soll stetig im Nullpunkt sein, wir gehen aber davon aus, dass wir sie in diesem Punkt nicht direkt auswerten können, beispielsweise weil der Rechenaufwand dafür inakzeptabel wäre oder andere Probleme auftreten.

Unser Ziel ist es, aus Auswertungen in Punkten des verbleibenden Intervalls $(0, h_{\max}]$ eine Näherung des Funktionswerts $g(0)$ zu konstruieren.

Definition 6.29 (Asymptotische Entwicklung) *Seien $m \in \mathbb{N}_0$, $C_{ae} \in \mathbb{R}_{\geq 0}$, $h_{max} \in \mathbb{R}_{>0}$ und $q \in \Pi_m$. Falls*

$$\|g - q\|_{\infty,[0,h]} \leq C_{ae} h^{m+1} \qquad \text{für alle } h \in (0, h_{max}] \tag{6.19}$$

gilt, nennen wir (C_{ae}, h_{max}, q) eine asymptotische Entwicklung *der Funktion g.*

Man erkennt, dass aus (6.19) insbesondere $g(0) = q(0)$ folgt, indem wir den Grenzwert $h \to 0$ betrachten.

Asymptotische Entwicklungen lassen sich in vielen praktischen Fällen aus der Taylor-Entwicklung berechnen. Besonders naheliegend ist in diesem Kontext das Beispiel des *numerischen Differenzierens*, also der numerischen Approximation der Ableitungen einer Funktion.

Beispiel 6.30 (Differenzenquotient) *Als erstes Beispiel untersuchen wir die Approximation der Ableitung durch einen Differenzenquotienten. Sei $f \in C^1[a, b]$, sei $x \in (a, b)$, dann ist*

$$g(h) := \frac{f(x+h) - f(x)}{h},$$

für hinreichend kleine Werte von h ein Approximation der Ableitung $f'(x)$.

Falls $f \in C^{m+2}[a, b]$ und $h \in (0, b - x]$ gilt, erhalten wir mit dem Satz von Taylor die Gleichung

$$f(x+h) = f(x) + hf'(x) + \sum_{i=1}^{m} h^{i+1} \frac{f^{(i+1)}(x)}{(i+1)!} + h^{m+2} \frac{f^{(m+2)}(\eta)}{(m+2)!}$$

für ein $\eta \in [x, x+h]$. Durch Umsortieren und Division durch h folgt

$$g(h) = \frac{f(x+h) - f(x)}{h} = f'(x) + \sum_{i=1}^{m} h^i \frac{f^{(i+1)}(x)}{(i+1)!} + h^{m+1} \frac{f^{(m+2)}(\eta)}{(m+2)!},$$

und mit

$$q(h) := f'(x) + \sum_{i=1}^{m} h^i \frac{f^{(i+1)}(x)}{(i+1)!} \qquad \text{für alle } h \in (0, b-x]$$

und den Konstanten

$$C_{ae} := \frac{\|f^{(m+2)}\|_{\infty,[a,b]}}{(m+2)!}, \qquad h_{max} := b - x$$

folgt (6.19), also ist (C_{ae}, h_{max}, q) eine asymptotische Entwicklung der Funktion

$$g\colon [0, h_{max}] \to \mathbb{R}, \qquad h \mapsto \begin{cases} f'(x) & \text{falls } h = 0, \\ \frac{f(x+h)-f(x)}{h} & \text{ansonsten.} \end{cases}$$

Indem wir die Abschätzung dieses Beispiels für $m = 0$ untersuchen stellen wir fest, dass wir von diesem Differenzenquotienten nur erwarten dürfen, dass er proportional zu h gegen die Ableitung strebt, also relativ langsam. Unser Ziel ist es nun, die Konvergenz zu beschleunigen.

Nach Voraussetzung können wir die Funktion g in Punkten $h \in (0, h_{\max}]$ auswerten, also liegt es nahe, sie durch ein Polynom $p \in \Pi_m$ zu interpolieren. Dazu wählen wir Interpolationspunkte $0 < h_m < h_{m-1} < \ldots < h_0 \leq h_{\max}$ und bestimmen $p \in \Pi_m$ so, dass

$$p(h_i) = g(h_i) \qquad \text{für alle } i \in [0:m]$$

gilt. Die Auswertung von p im Nullpunkt verwenden wir als Approximation des gesuchten Funktionswerts $g(0)$.

Zum Nachweis einer Fehlerabschätzung bezeichen wir den zu den Punkten $h_m, \ldots, h_0$ gehörenden Interpolationsoperator mit $\mathfrak{I}$ und mit Λ seine Lebesgue-Konstante (vgl. Definition 6.22) und erhalten

$$\begin{aligned} |p(0) - g(0)| &= |p(0) - q(0)| \leq \|p - q\|_{\infty,[0,h_0]} = \|\mathfrak{I}[g] - \mathfrak{I}[q]\|_{\infty,[0,h_0]} \\ &= \|\mathfrak{I}[g-q]\|_{\infty,[0,h_0]} \leq \Lambda \|g - q\|_{\infty,[0,h_0]} \leq \Lambda C_{\mathrm{ae}} h_0^{m+1}. \end{aligned}$$

Es sieht also so aus, als würden wir mit der *praktisch berechenbaren* Größe $p(0)$ eine Näherung der Ableitung erhalten, die für $h_0 \to 0$ wie h_0^{m+1} konvergiert.

Allerdings trifft das noch nicht ganz zu: Die Lebesgue-Zahl Λ hängt von $h_m, \ldots, h_0$ ab, könnte sich also ungünstig verhalten, falls wir $h_m, \ldots, h_1$ für verschiedene Werte von h_0 ungeschickt wählen. Glücklicherweise lässt sich dieses Problem relativ einfach lösen, und die Behandlung der Stabilitätskonstant lässt sich besonders elegant gestalten, wenn wir

bedenken, dass wir gar nicht an der Konvergenz auf dem gesamten Intervall interessiert sind, sondern nur an dem Verhalten im Punkt null.

Um auszuschließen, dass eine ungeschickte Wahl der Interpolationspunkte die Stabilität beeinträchtigt, fixieren wir Interpolationspunkte $\alpha_0, \dots, \alpha_m \in (0,1]$ und leiten aus ihnen Interpolationspunkte auf dem Intervall $(0,h]$ mittels der Skalierung

$$h_i = \alpha_i h \qquad \text{für alle } i \in [0:m] \tag{6.20}$$

her, wobei wir nun $h \in \mathbb{R}_{>0}$ beliebig wählen können. Den korrespondierenden Interpolationsoperator bezeichnen wir mit $\mathfrak{I}_h$.

Lemma 6.31 (Skalierte Interpolation) *Mit $\alpha_0, \dots, \alpha_m \in (0,1]$ und der Konstanten*

$$\Lambda_0 := \sum_{i=0}^{m} \prod_{\substack{j=0 \\ j\neq i}}^{m} \frac{|0-\alpha_j|}{|\alpha_i - \alpha_j|}$$

erhalten wir für alle $h \in \mathbb{R}_{>0}$ die Abschätzung

$$|\mathfrak{I}_h[f](0)| \leq \Lambda_0 \|f\|_{\infty,[0,h]} \qquad \text{für alle } f \in C[0,h].$$

Beweis. Sei $h \in \mathbb{R}_{>0}$. Für die Lagrange-Polynomen $\ell_{h,i}$, $i \in [0:m]$, zu den durch (6.20) definierten Punkten gilt wegen (6.1) die Gleichung

$$\ell_{h,i}(0) = \prod_{\substack{j=0 \\ j\neq i}}^{m} \frac{0-h_j}{h_i - h_j} = \prod_{\substack{j=0 \\ j\neq i}}^{m} \frac{0h - \alpha_j h}{\alpha_i h - \alpha_j h} = \prod_{\substack{j=0 \\ j\neq i}}^{m} \frac{0-\alpha_j}{\alpha_i - \alpha_j} \qquad \text{für alle } i \in [0:m],$$

also sind die Werte der Lagrange-Polynome in null von h unabhängig. Damit ist auch

$$\Lambda_{h,0} := \sum_{i=0}^{m} |\ell_{h,i}(0)| = \sum_{i=0}^{m} \prod_{\substack{j=0 \\ j\neq i}}^{m} \frac{|0-\alpha_j|}{|\alpha_i - \alpha_j|} = \Lambda_0$$

von h unabhängig.

Sei nun $f \in C[0,h]$ gegeben. Mit der Dreiecksungleichung folgt

$$\begin{aligned} |\mathfrak{I}_h[f](0)| &= \left| \sum_{i=0}^{m} f(h_i)\ell_{h,i}(0) \right| \leq \sum_{i=0}^{m} |f(h_i)|\,|\ell_{h,i}(0)| \\ &\leq \|f\|_{\infty,[0,h]} \sum_{i=0}^{m} |\ell_{h,i}(0)| = \Lambda_{h,0}\|f\|_{\infty,[0,h]} = \Lambda_0 \|f\|_{\infty,[0,h]}. \end{aligned}$$

Die Stabilität der Auswertung in null ist also unabhängig von der Wahl des Skalierungsparameters h. ∎

Nun können wir das Extrapolationsverfahren formal sauber einführen: Aus unserer Voraussetzung $\alpha_0, \dots, \alpha_m \in (0,1]$ folgt $h_0, \dots, h_m \in (0,h]$, wir können also g in diesen

Interpolationspunkten auswerten, sofern $h \leq h_{\max}$ gilt. Also können wir auch $g_h := \mathfrak{I}_h[g]$ konkret berechnen und

$$g_h(0) = \mathfrak{I}_h[g](0) \tag{6.21}$$

als verbesserte Näherung des Funktionswerts $g(0)$ verwenden.

Satz 6.32 (Extrapolation) *Die Funktion g besitze eine asymptotische Entwicklung der Form (6.19), und die Approximation $g_h(0)$ sei gemäß (6.21) und (6.20) definiert. Dann gilt*

$$|g_h(0) - g(0)| \leq \Lambda_0 C_{ae} h^{m+1} \qquad \text{für alle } h \in (0, h_{max}].$$

Beweis. Sei $h \in (0, h_{\max}]$. Dann folgt aus Lemma 6.31 und (6.19) bereits

$$\begin{aligned} |g_h(0) - g(0)| &= |\mathfrak{I}_h[g](0) - q(0)| = |\mathfrak{I}_h[g - q](0)| \\ &\leq \Lambda_0 \|g - q\|_{\infty,[0,h]} \leq \Lambda_0 C_{\mathrm{ae}} h^{m+1}. \end{aligned}$$

Wir können also den gesuchten Wert $g(0)$ mit sehr hoher Genauigkeit approximieren, indem wir h hinreichend klein wählen. ∎

Durch geeignete Wahl der Funktion g und des Polynoms q lassen sich spezielle Eigenschaften des numerischen Näherungsverfahrens ausnutzen. Als Beispiel verwenden wir einen modifizierten Differenzenquotienten zur Annäherung der Ableitung:

Beispiel 6.33 (Zentraler Differenzenquotient) *Sei $f \in C^1[a, b]$, sei $x \in (a, b)$, dann ist der* zentrale Differenzenquotient

$$g(h) := \frac{f(x + h) - f(x - h)}{2h}$$

für hinreichend kleine Werte von h eine Approximation der Ableitung $f'(x)$.

Falls $f \in C^{2m+3}[a, b]$ und $h \in (0, h_{max}]$ für $h_{max} := \min\{x - a, b - x\}$ gilt, erhalten wir analog zu dem Beispiel 6.30 die Darstellung

$$g(h) = f'(x) + \sum_{i=1}^{m} h^{2i} \frac{f^{(2i+1)}(x)}{(2i + 1)!} + h^{2m+2} \frac{f^{(2m+3)}(\eta)}{(2m + 3)!} \tag{6.22}$$

mit einem $\eta \in [x - h, x + h]$. Diese Darstellung enthält also „Lücken“, alle ungeradzahligen h-Potenzen fehlen. Es bietet sich an, diese Eigenschaft nicht ungenutzt zu lassen.

Wir ersetzen dazu h durch $\hat{h} := h^2$ und g durch

$$\hat{g} \colon [0, h_{max}^2] \to \mathbb{R}, \qquad \hat{h} \mapsto g(\sqrt{\hat{h}}),$$

so dass (6.22) die Gestalt

$$\hat{g}(\hat{h}) = f'(x) + \sum_{i=1}^{m} \hat{h}^i \frac{f^{(2i+1)}(x)}{(2i + 1)!} + \hat{h}^{m+1} \frac{f^{(2m+3)}(\eta)}{(2m + 3)!}$$

annimmt. Die „Lücken“ sind verschwunden, wir definieren

$$\hat{q}(\hat{h}) := f'(x) + \sum_{i=1}^{m} \hat{h}^i \frac{f^{(2i+1)}(x)}{(2i+1)!} \qquad \text{für alle } \hat{h} \in \mathbb{R}$$

und erhalten mit (6.22)

$$\|\hat{g} - \hat{q}\|_{\infty,[0,h_{max}^2]} \leq C_{ae} \hat{h}^{m+1} \qquad \text{für alle } \hat{h} \in [0, \hat{h}_{max}]$$

für die Konstanten

$$C_{ae} := \frac{\|f^{(2m+3)}\|_{\infty,[a,b]}}{(2m+3)!}, \qquad \hat{h}_{max} := h_{max}^2.$$

Damit ist $(C_{ae}, \hat{h}_{max}, \hat{q})$ eine asymptotische Entwicklung der Funktion $\hat{g}$, auf die wir die entwickelten Techniken anwenden können.

Allerdings gibt es eine Besonderheit: Wir arbeiten mit $\hat{h} = h^2$ anstelle von h, also wird schließlich auch in der Abschätzung des Satzes 6.32 der Fehler proportional zu $\hat{h}^{m+1} = h^{2m+2}$ fallen, wir haben also erheblich an Genauigkeit gewonnen.

Die Auswertung der Funktion $\hat{g}$ lässt sich besonders einfach gestalten, wenn wir dafür sorgen, dass sie nur in Punkten erfolgt, deren Wurzel wir bereits kennen. Dazu wählen wir $\alpha_0, \ldots, \alpha_m \in (0,1]$ und definieren $\hat{\alpha}_i := \alpha_i^2$ für $i \in [0:m]$. Für ein gegebenes $h \in (0, h_{max}]$ und $\hat{h} := h^2$ sind dann die Interpolationspunkte durch

$$h_i := \alpha_i h, \qquad \hat{h}_i := \hat{\alpha}_i \hat{h} = h_i^2 \qquad \text{für alle } i \in [0:m]$$

gegeben, und wir erhalten

$$\hat{g}(\hat{h}_i) = g(h_i) \qquad \text{für alle } i \in [0:m],$$

brauchen also keine Wurzeln zu berechnen, um $\hat{g}$ auszuwerten.

Beispiel 6.34 (Romberg-Quadratur) *Neben der Ableitung ist auch das Integral einer Funktion f als Grenzwert definiert, also bietet es sich an, die Anwendung der Extrapolationstechnik auf die Berechnung von Integralen zu untersuchen. Als Ausgangspunkt dient uns die* summierte Trapezregel

$$T_k(f) := \frac{1}{k} \left(\frac{1}{2} f(0) + \sum_{i=1}^{k-1} f(i/k) + \frac{1}{2} f(1) \right),$$

mit der sich das Integral

$$I(f) := \int_0^1 f(x)\, dx$$

einer Funktion $f \in C[0,1]$ approximieren lässt.

Da $T_k(f)$ nur für $k \in \mathbb{N}$ definiert ist, führen wir

$$H := \{1/k \ : \ k \in \mathbb{N}\}$$

ein und können die Funktion g nur auf H untersuchen:

$$g\colon H \to \mathbb{R}, \qquad\qquad h \mapsto T_{1/h}(f).$$

Trotz des eingeschränkten Definitionsbereichs lässt sich der Grenzwert

$$\lim_{\substack{h\to 0\\ h\in H}} g(h) = \lim_{k\to\infty} T_k(f)$$

per Extrapolation approximieren.

Dafür benötigen wir die Existenz einer asymptotischen Entwicklung, die sich in diesem Fall mit Hilfe der Euler-Maclaurin-Summenformel *gewinnen lässt: Falls $f \in C^{2m+2}[0,1]$ gilt, existieren von f abhängende Konstanten $\beta_2, \ldots, \beta_{2m}$ und C_{ae} mit*

$$\left| g(h) - I(f) - \sum_{i=1}^{m} \beta_{2i} h^{2i} \right| \le C_{ae} h^{2m+2} \qquad \text{für alle } h \in H.$$

Wie schon im Fall des zentralen Differenzenquotienten treten keine ungeraden Potenzen von h in der Summe auf, also bietet es sich an, wieder $\hat{h} := h^2$ zu substituieren, um zu

$$\left| g(\sqrt{\hat{h}}) - I(f) - \sum_{i=1}^{m} \beta_{2i} \hat{h}^i \right| \le C_{ae} \hat{h}^{m+1} \qquad \text{für alle } \hat{h} \in \widehat{H} := \{h^2 \ : \ h \in H\}$$

zu gelangen.

Die Romberg-Quadratur *beruht auf der Beobachtung, dass sich $T_{2k}(f) = g(h/2)$ aus $T_k(t) = g(h)$ effizient mit der Formel*

$$T_{2k}(f) = \frac{1}{2} T_k(f) + \frac{1}{2k} \sum_{i=1}^{k} f\left(\frac{2i-1}{2k}\right)$$

berechnen lässt, so dass es sich anbietet, die Extrapolationspunkte so zu wählen, dass g nur in $\alpha_i = 2^{-i}/k$ ausgewertet werden muss. Als quadrierte Interpolationspunkte bieten sich also $\hat{\alpha}_i = 4^{-i}/k^2$ an. Wertet man das so entstehende Interpolationspolynom mit dem Neville-Aitken-Verfahren im Nullpunkt aus, ergibt sich das Romberg-Quadraturverfahren.

Aus $I_{0,0} = T_k(f)$ und $I_{1,1} = T_{2k}(f)$ erhalten wir

$$I_{0,1} = \frac{4^{-1}/k^2 - 0}{4^{-1}/k^2 - 1/k^2} I_{0,0} + \frac{0 - 1/k^2}{4^{-1}/k^2 - 1/k^2} I_{1,1} = \frac{-1}{3} I_{0,0} + \frac{4}{3} I_{1,1}.$$

Wenn wir $I_{2,2} = T_{4k}(f)$ hinzu nehmen, ergeben sich

$$I_{1,2} = \frac{4^{-2}/k^2 - 0}{4^{-2}/k^2 - 4^{-1}/k^2} I_{1,1} + \frac{0 - 4^{-1}/k^2}{4^{-2}/k^2 - 4^{-1}/k^2} I_{2,2} = \frac{-1}{3} I_{1,1} + \frac{4}{3} I_{2,2},$$
$$I_{0,2} = \frac{4^{-2}/k^2 - 0}{4^{-2}/k^2 - 1/k^2} I_{0,1} + \frac{0 - 1/k^2}{4^{-2}/k^2 - 1/k^2} I_{1,2} = \frac{-1}{15} I_{0,1} + \frac{16}{15} I_{1,2}.$$

6.7 Stückweise Polynome

In diesem Abschnitt bezeichnen wir mit $\mathfrak{I}_{[a,b]}$ einen beliebigen Interpolationsoperator m-ten Grades auf dem Intervall $[a,b]$. Aus der Kombination der Abschätzungen (6.11) und (6.12) folgt die Fehlerabschätzung

$$\|f - \mathfrak{I}_{[a,b]}[f]\|_{\infty,[a,b]} \leq (b-a)^{m+1} \frac{\|f^{(m+1)}\|_{\infty,[a,b]}}{(m+1)!} \qquad \text{für alle } f \in C^{m+1}[a,b]. \tag{6.23}$$

Wenn wir uns die Aufgabe stellen, diesen Fehler unter eine vorgegebene Schranke $\epsilon \in \mathbb{R}_{>0}$ zu senken, bieten sich uns nur wenige Möglichkeiten: Falls

$$\frac{\|f^{(m+1)}\|_{\infty,[a,b]}}{(m+1)!} \leq C\, c^{m+1} \qquad \text{für alle } m \in \mathbb{N}_0 \tag{6.24}$$

mit Konstanten $C, c \in \mathbb{R}_{\geq 0}$ gelten sollte und wir beispielsweise $(b-a)c \leq 1/2$ sicherstellen können, folgt aus der obigen Abschätzung

$$\|f - \mathfrak{I}_{[a,b]}[f]\|_{\infty,[a,b]} \leq C(b-a)^{m+1}c^{m+1} = C((b-a)c)^{m+1} = C\,2^{-(m+1)},$$

wir dürfen also darauf hoffen, dass wir durch Erhöhen des Grades m um eins den Fehler mindestens halbieren. Mit der Wahl

$$m = \left\lceil \log_2 \frac{C}{\epsilon} - 1 \right\rceil = \lceil \log_2 C - \log_2 \epsilon - 1 \rceil$$

erhalten wir

$$\|f - \mathfrak{I}_{[a,b]}[f]\|_{\infty,[a,b]} \leq C\,2^{-(m+1)} \leq C\frac{\epsilon}{C} = \epsilon,$$

also die gewünschte Genauigkeit. Die Bedingung (6.24) impliziert insbesondere, dass die Funktion f auf einer Umgebung des Intervalls $[a,b]$ analytisch ist, die Voraussetzungen für diesen Ansatz zur Verbesserung der Genauigkeit sind also relativ hoch.

Wir können auch einen alternativen Zugang wählen: Falls $f \in C^{m+1}[a,b]$ für ein $m \in \mathbb{N}_0$ gilt, konstruieren wir nicht ein Interpolationspolynom für das gesamte Intervall $[a,b]$, sondern zerlegen es in Teilintervalle, auf denen dann interpoliert wird.

Definition 6.35 (Stückweise Polynome) *Seien $k \in \mathbb{N}$ und $a = y_0 < y_1 < \ldots < y_k = b$ gegeben. Dann ist*

$$\Pi_{m,(y_0,\ldots,y_k)} := \{f\colon [a,b] \to \mathbb{R} \;:\; f|_{(y_{i-1},y_i]} \in \Pi_m \text{ für alle } i \in [1:m]\}$$

ein Vektorraum, den wir als den Raum der stückweisen Polynome *auf der durch $(y_0,\ldots,y_k)$ gegebenen Zerlegung des Intervalls $[a,b]$ bezeichnen.*

Der bisher untersuchte Interpolationsoperator $\mathfrak{I}_{[a,b]}$ bildet auf den Raum Π_m der Polynome ab, jetzt suchen wir einen Interpolationsoperator, der auf den Raum der stückweisen Polynome abbildet. Es liegt nahe, ihn aus den Operatoren $\mathfrak{I}_{[y_{i-1},y_i]}$ zusammenzusetzen:

Definition 6.36 (Stückweise Interpolation) *Wir definieren den Operator*

$$\mathfrak{I}_{m,(y_0,\ldots,y_k)} \colon C[a,b] \to \Pi_{m,(y_0,\ldots,y_k)}$$

punktweise durch

$$\mathfrak{I}_{m,(y_0,\ldots,y_k)}[f](x) := \begin{cases} \mathfrak{I}_{[y_0,y_1]}[f](x) & \text{falls } x \in [y_0,y_1], \\ \mathfrak{I}_{[y_{i-1},y_i]}[f](x) & \text{falls } x \in (y_{i-1},y_i] \end{cases} \qquad \text{für alle } x \in [a,b].$$

Wir nennen ihn den stückweisen Interpolationsoperator *für die durch $y_0,\ldots,y_k$ gegebene Zerlegung des Intervalls $[a,b]$.*

Im einfachsten Fall zerlegen wir $[a,b]$ in k gleichgroße Intervalle, indem wir

$$y_i := a + \frac{b-a}{k} i \qquad \text{für alle } i \in [0:k]$$

setzen und die Intervalle $[y_{i-1},y_i]$ untersuchen. In diesem Fall spricht man von einer *äquidistanten* Zerlegung des Intervalls und bezeichnet mit

$$h := \frac{b-a}{k}$$

die *Schrittweite*. Unter diesen Bedingungen erhalten wir die Fehlerabschätzung

$$\|f - \mathfrak{I}_{m,(y_0,\ldots,y_k)}[f]\|_{\infty,[a,b]} \leq h^{m+1} \frac{\|f^{(m+1)}\|_{\infty,[a,b]}}{(m+1)!} \qquad \text{für alle } f \in C^{m+1}[a,b],$$

indem wir die allgemeine Abschätzung (6.23) auf die Teilintervalle $[y_{i-1},y_i]$ anwenden und ausnutzen, dass $y_i - y_{i-1} = h$ gilt.

Nun können wir auch untersuchen, wie sich mit Hilfe des Parameters k die Genauigkeit steuern lässt: Wenn wir

$$C := (b-a)^{m+1} \frac{\|f^{(m+1)}\|_{\infty,[a,b]}}{(m+1)!}$$

setzen, erhalten wir für den Fehler die Abschätzung

$$\|f - \mathfrak{I}_{m,(y_0,\ldots,y_k)}[f]\|_{\infty,[a,b]} \leq C\, k^{-(m+1)}.$$

Um den Fehler unter die Schranke ϵ zu drücken, müssen wir also

$$C\, k^{-(m+1)} \leq \epsilon, \qquad k^{m+1} \geq \frac{C}{\epsilon}, \qquad k \geq \left(\frac{C}{\epsilon}\right)^{1/(m+1)}$$

erreichen, etwa indem wir

$$k = \left\lceil (C/\epsilon)^{1/(m+1)} \right\rceil$$

wählen. Mit diesem Ansatz können wir also jede beliebige Genauigkeit erreichen, *ohne* besonders hohe Regularität der zu interpolierenden Funktion f voraussetzen zu müssen. Allerdings ist eine hohe Regularität auch hier durchaus von Vorteil: Der Parameter k wächst wie $\epsilon^{-1/(m+1)}$, also um so langsamer, je größer wir m wählen dürfen.

Übungsaufgabe 6.37 (Lokal verfeinertes Gitter) *Mit Hilfe stückweiser Polynome lassen sich auch Funktionen approximieren, die Singularitäten aufweisen. Besonders effizient ist dieser Ansatz, wenn wir statt einer äquidistanten eine der Singularität angepasste Unterteilung des Intervalls verwenden.*

Sei $q \in (0,1)$ gegeben, und sei für alle $i \in \mathbb{N}_0$ und $m \in \mathbb{N}_0$ mit $\mathfrak{I}_{m,[q^{i+1},q^i]}$ der auf das Intervall $[q^{i+1}, q^i]$ transformierte Tschebyscheff-Interpolationsoperator m-ten Grades bezeichnet.

(a) Sei für $a, b \in \mathbb{R}$ mit $a < b$ durch $\Phi_{[a,b]}$ die aus (6.16) bekannte Transformation bezeichnet, und sei $\mathfrak{I}_m$ der Tschebyscheff-Interpolationoperator m-ten Grades auf dem Referenzintervall $[-1,1]$. Sei $f \in C(0,1]$. Beweisen Sie

$$\mathfrak{I}_{m,[q^{i+1},q^i]}[f] = \mathfrak{I}[f \circ \Phi_{[q^{i+1},q^i]}] \circ \Phi^{-1}_{[q^{i+1},q^i]} \qquad \text{für alle } i, m \in \mathbb{N}_0.$$

(b) Wir definieren die Skalierungsfunktionen

$$S_i \colon \mathbb{R} \to \mathbb{R}, \qquad x \mapsto q^i x \qquad \text{für alle } i \in \mathbb{N}_0.$$

Sei $f \in C(0,1]$. Beweisen Sie

$$\|f - \mathfrak{I}_{m,[q^{i+1},q^i]}[f]\|_{\infty,[q^{i+1},q^i]} = \|f \circ S_i - \mathfrak{I}_{m,[q,1]}[f \circ S_i]\|_{\infty,[q,1]} \quad \text{für alle } i, m \in \mathbb{N}_0.$$

(c) Sei $f(x) := \log(x)$ für alle $x \in (0,1]$. Beweisen Sie

$$\|f - \mathfrak{I}_{m,[q^{i+1},q^i]}[f]\|_{\infty,[q^{i+1},q^i]} = \|f - \mathfrak{I}_{m,[q,1]}[f]\|_{\infty,[q,1]} \qquad \text{für alle } i, m \in \mathbb{N}_0.$$

(d) Sei f wie in Teil (c) gegeben. Beweisen Sie

$$\|f - \mathfrak{I}_{m,[q^i,q^{i-1}]}[f]\|_{\infty,[q^i,q^{i-1}]} \leq \frac{2}{m+1}\left(\frac{1-q}{4q}\right)^{m+1} \qquad \text{für alle } i, m \in \mathbb{N}_0.$$

Hinweise: *Bei Teil (a) kann der Eindeutigkeitssatz für Polynome helfen. Bei Teil (b) lohnt sich ein Blick auf die Beziehung zwischen S_i und $\Phi_{[q^{i+1},q^i]}$.*

Bonusaufgabe: *Was geschieht, wenn wir für ein $\alpha \in \mathbb{N}$ die durch $f(x) = x^{-\alpha}$ für alle $x \in (0,1]$ gegebene Funktion betrachten?*

Übungsaufgabe 6.38 (Fehlerschätzung) *Sei $f \in C^2[-1,1]$* konvex*, es gelte also $f''(x) \geq 0$ für alle $x \in [-1,1]$. Wir approximieren f durch die linearen Polynome*

$$p_1(x) := \frac{1-x}{2} f(-1) + \frac{1+x}{2} f(1),$$
$$p_2(x) := f(0) + x f'(0) \qquad \text{für alle } x \in [-1,1],$$

also durch das Interpolationspolynom zu den Endpunkten und durch das Taylor-Polynom zu dem Mittelpunkt des Intervalls.

(a) Beweisen Sie

$$p_2(x) \leq f(x) \leq p_1(x) \qquad \text{für alle } x \in [-1,1].$$

(b) Beweisen Sie

$$\|f - p_1\|_{\infty,[-1,1]} \leq \|p_2 - p_1\|_{\infty,[-1,1]}.$$

6.8 Splines

In vielen Anwendungen ist man daran interessiert, dass das stückweise Interpolationspolynom stetig oder sogar differenzierbar ist. Die Forderung nach Stetigkeit lässt sich dabei besonders einfach erfüllen: Falls wir dafür sorgen, dass jeder Interpolationsoperator $\mathfrak{I}_{[y_{i-1},y_i]}$ die Randpunkte y_{i-1} und y_i als Interpolationspunkte enthält, ist sichergestellt, dass

$$\mathfrak{I}_{[y_{i-1},y_i]}[f](y_i) = f(y_i) = \mathfrak{I}_{[y_i,y_{i+1}]}[f](y_i) \qquad \text{für alle } i \in [1:k-1]$$

gilt, so dass die stückweisen Interpolationspolynome stetig ineinander übergehen und deshalb auch $\mathfrak{I}_{m,(y_0,\ldots,y_m)}[f]$ stetig sein muss. Offensichtlich muss für diese Konstruktion $m \geq 1$ gelten, denn ansonsten stünde uns pro Intervall nur ein Interpolationspunkt zur Verfügung, der nicht gleichzeitig der linke und der rechte Randpunkt sein kann.

Allgemein kann man sich überlegen, dass sich stückweise Polynomen m-ten Grades zu einer $(m-1)$-mal stetig differenzierbaren Funktion zusammensetzen lassen.

Definition 6.39 (Splineraum) *Seien $a = y_0 < y_1 < \ldots < y_k = b$ und $m \in \mathbb{N}$ gegeben. Der Vektorraum*

$$\mathcal{S}_{m,(y_0,\ldots,y_k)} := \{u \in C^{m-1}[a,b] \ : \ u|_{[y_{i-1},y_i)} \in \Pi_m \text{ für alle } i \in [1:k]\}$$

wird als der Splineraum *m-ten Grades bezeichnet, Funktionen aus diesem Raum nennt man* Splines*.*

Offenbar ist der Splineraum ein Teilraum des Raums $\Pi_{m,(y_0,\ldots,y_k)}$ der stückweisen Polynome. Wir interessieren uns für die Frage, wie sich mit Splines eine Funktion interpolieren lässt, deren Werte in den Punkten $y_0, \ldots, y_k$ bekannt sind.

Als Beispiel untersuchen wir die weit verbreiteten *kubischen Splines*, also Funktionen aus dem Raum $\mathcal{S}_{3,(y_0,\ldots,y_k)}$. Wir gehen von gegebenen Werten $f_0, \ldots, f_k \in \mathbb{R}$ aus und suchen ein $s \in \mathcal{S}_{3,(y_0,\ldots,y_k)}$ mit

$$s(y_i) = f_i \qquad \text{für alle } i \in [0:k].$$

Zur Vereinfachung beschränken wir uns hier auf eine äquidistante Unterteilung

$$y_i = a + hi, \qquad h = \frac{b-a}{k}, \qquad \text{für alle } i \in [0:k].$$

Da wir wissen, dass $s|_{[y_{i-1},y_i]} \in \Pi_3$ für alle $i \in [1:k]$ gelten muss, bietet es sich an, eine geeignete Basis für diesen Raum zu konstruieren.

Ein erster Ansatz könnte eine Lagrange-Basis sein, allerdings müssten wir dann einerseits vier geeignete Interpolationspunkte in jedem Intervall wählen und andererseits wäre es recht umständlich, die Stetigkeitsbedingungen sich zu stellen, schließlich soll s zweimal stetig differenzierbar sein.

Deshalb verwenden wir eine Basis, die es uns erlaubt, die Funktionswerte und die Werte der ersten Ableitung in den Punkten y_i explizit vorzugeben. Die Funktionswerte sind durch unsere Interpolationsaufgabe gegeben, die Ableitungswerte müssen wir so wählen, dass s nicht nur einmal, sondern sogar zweimal stetig differenzierbar wird.

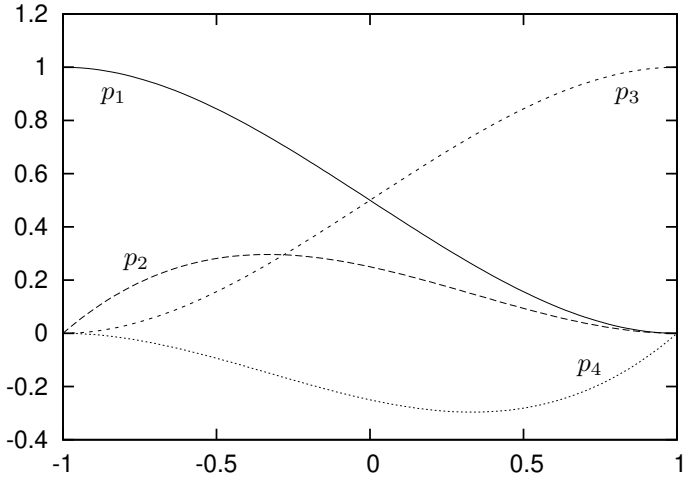

Abbildung 6.4: Kubische Hermite-Basis

Lemma 6.40 (Hermite-Basis) *Die Polynome*

$$p_1(x) := \frac{1}{4}(x-1)^2(x+2), \qquad p_2(x) := \frac{1}{4}(x-1)^2(x+1),$$
$$p_3(x) := \frac{1}{4}(x+1)^2(2-x), \qquad p_4(x) := \frac{1}{4}(x+1)^2(x-1),$$

erfüllen die Gleichungen

$$\begin{aligned}
p_1(-1) &= 1, & p_1'(-1) &= 0, & p_1(1) &= 0, & p_1'(1) &= 0,\\
p_2(-1) &= 0, & p_2'(-1) &= 1, & p_2(1) &= 0, & p_2'(1) &= 0,\\
p_3(-1) &= 0, & p_3'(-1) &= 0, & p_3(1) &= 1, & p_3'(1) &= 0,\\
p_4(-1) &= 0, & p_4'(-1) &= 0, & p_4(1) &= 0, & p_4'(1) &= 1.
\end{aligned}$$

Beweis. Die Eigenschaften lassen sich direkt nachrechnen.

Die Idee hinter der Konstruktion ist einfach: p_1 muss eine doppelte Nullstelle in $x = 1$ besitzen, also muss ein Faktor $(x-1)^2$ vorkommen, und es sind nur noch zwei Koeffizienten $a, b \in \mathbb{R}$ mit $p_1(x) = (x-1)^2(a+bx)$ zu bestimmen.

p_2 muss eine doppelte Nullstelle in $x = 1$ und eine einfache in $x = -1$ besitzen, also ist hier sogar nur noch ein Koeffizient $a \in \mathbb{R}$ mit $p_2(x) = (x-1)^2(x+1)a$ zu wählen.

Da das Intervall $[-1, 1]$ spiegelsymmetrisch ist, können wir $p_3(x) = p_1(-x)$ und $p_4(x) = -p_2(-x)$ setzen. ∎

Die Basis $\{p_1, p_2, p_3, p_4\}$ ähnelt der Lagrange-Basis, allerdings können wir mit ihrer Hilfe nicht nur Funktionswerte, sondern auch Ableitungen interpolieren: Für ein $f \in C^1[-1, 1]$ ist

$$p := f(-1)\,p_1 + f'(-1)\,p_2 + f(1)\,p_3 + f'(1)\,p_4$$

gerade dasjenige kubische Polynom, das

$$p(-1) = f(-1), \qquad p'(-1) = f'(-1), \qquad p(1) = f(1), \qquad p'(-1) = f'(-1)$$

erfüllt, also Werte und Ableitungswerte in den Punkten -1 und 1 exakt reproduziert. Derartige Basen bezeichnet man als *Hermite-Basen*, ein Interpolationverfahren, das nicht nur Funktionswerte, sondern auch Ableitungswerte verwendet, nennen wir *Hermite-Interpolationsverfahren.*

Für die Konstruktion unseres Splines ist dieser Ansatz sehr nützlich, denn er ermöglicht es uns, zumindest die Differenzierbarkeit des Splines s explizit sicherzustellen, indem wir dafür sorgen, dass die Ableitungen in jedem Punkt $y_1, \ldots, y_{k-1}$ stetig ineinander übergehen: Wir wählen $d_0, \ldots, d_k \in \mathbb{R}$ geeignet und konstruieren s so, dass

$$s'(y_i) = d_i \qquad \text{für alle } i \in [0:k]$$

gilt. Dazu verwenden wir die Hermite-Basis: Auf einem Intervall $[y_{i-1}, y_i]$ konstruieren wir eine Basis mit Hilfe der Abbildung

$$\Phi_i \colon [-1,1] \to [y_{i-1}, y_i], \qquad x \mapsto \frac{y_i + y_{i-1}}{2} + \frac{y_i - y_{i-1}}{2} x = y_i + \frac{h}{2}(x-1),$$

und erhalten transformierte Basisfunktionen

$$p_{i,1} := p_1 \circ \Phi_i^{-1}, \qquad p_{i,2} := \frac{h}{2} p_2 \circ \Phi_i^{-1},$$
$$p_{i,3} := p_3 \circ \Phi_i^{-1}, \qquad p_{i,4} := \frac{h}{2} p_4 \circ \Phi_i^{-1}.$$

Lemma 6.41 (Transformierte Hermite-Basis) *Für jedes $i \in [1:k]$ gelten*

$$\begin{aligned}
&p_{i,1}(y_{i-1}) = 1, && p'_{i,1}(y_{i-1}) = 0, && p_{i,1}(y_i) = 0, && p'_{i,1}(y_i) = 0,\\
&p_{i,2}(y_{i-1}) = 0, && p'_{i,2}(y_{i-1}) = 1, && p_{i,2}(y_i) = 0, && p'_{i,2}(y_i) = 0,\\
&p_{i,3}(y_{i-1}) = 0, && p'_{i,3}(y_{i-1}) = 0, && p_{i,3}(y_i) = 1, && p'_{i,3}(y_i) = 0,\\
&p_{i,4}(y_{i-1}) = 0, && p'_{i,4}(y_{i-1}) = 0, && p_{i,4}(y_i) = 0, && p'_{i,4}(y_i) = 1.
\end{aligned}$$

Beweis. Wir können elementar

$$\Phi_i^{-1}(y) = 1 + \frac{2}{h}(y - y_i) \qquad \text{für alle } y \in [y_{i-1}, y_i]$$

nachrechnen, also folgt $(\Phi_i^{-1})' = 2/h$ und wir erhalten mit der Kettenregel

$$p'_{i,1} = \frac{2}{h} p'_1 \circ \Phi_i^{-1}, \qquad p'_{i,2} = p'_2 \circ \Phi_i^{-1},$$
$$p'_{i,3} = \frac{2}{h} p'_3 \circ \Phi_i^{-1}, \qquad p'_{i,4} = p'_4 \circ \Phi_i^{-1}.$$

Durch Einsetzen von $\Phi_i^{-1}(y_{i-1}) = -1$ und $\Phi_i^{-1}(y_i) = 1$ erhalten wir mit Lemma 6.40 die gewünschten Aussagen. ∎

Damit steht uns eine Hermite-Basis auf dem Intervall $[y_{i-1}, y_i]$ zur Verfügung, mit deren Hilfe wir s durch

$$\begin{aligned} s(x) = f_{i-1}\, p_{i,1}(x) + d_{i-1}\, p_{i,2}(x) \\ + f_i\, p_{i,3}(x) + d_i\, p_{i,4}(x) \qquad \text{für alle } i \in [1:k], x \in [y_{i-1}, y_i] \end{aligned} \tag{6.25}$$

darstellen können. Damit sind insbesondere $s(y_i) = f_i$ und $s'(y_i) = d_i$ sichergestellt, also die Stetigkeit der Funktion und ihrer ersten Ableitung.

Nach Definition 6.39 müssen wir allerdings auch die Stetigkeit der zweiten Ableitung sicherstellen. Dazu berechnen wir die Ableitungen der Hermite-Basisfunktionen:

$$\begin{aligned} p_1''(x) &= \frac{3}{2}x, & p_2''(x) &= \frac{3}{2}x - \frac{1}{2}, \\ p_3''(x) &= p_1''(-x) = -\frac{3}{2}x, & p_4''(x) &= -p_2''(-x) = \frac{3}{2}x + \frac{1}{2}. \end{aligned}$$

Für die transformierten Basisfunktionen folgt

$$\begin{aligned} p_{i,1}'' &= \frac{4}{h^2} p_1'' \circ \Phi_i^{-1}, & p_{i,2}'' &= \frac{2}{h} p_2'' \circ \Phi_i^{-1}, \\ p_{i,3}'' &= \frac{4}{h^2} p_3'' \circ \Phi_i^{-1}, & p_{i,4}'' &= \frac{2}{h} p_4'' \circ \Phi_i^{-1}. \end{aligned}$$

Die Werte der Basispolynome und ihrer Ableitungen in den Randpunkten sind in der folgenden Tabelle zusammengefasst:

	$p_{i,j}$		$p'_{i,j}$		$p''_{i,j}$	
j	y_{i-1}	y_i	y_{i-1}	y_i	y_{i-1}	y_i
1	1	0	0	0	$-6h^{-2}$	$6h^{-2}$
2	0	0	1	0	$-4h^{-1}$	$2h^{-1}$
3	0	1	0	0	$6h^{-2}$	$-6h^{-2}$
4	0	0	0	1	$-2h^{-1}$	$4h^{-1}$

Um sicherzustellen, dass die zweite Ableitung in einem Punkt y_i stetig ist, muss also die Gleichung

$$\begin{aligned} &f_{i-1}p_{i,1}''(y_i) + d_{i-1}p_{i,2}''(y_i) + f_i p_{i,3}''(y_i) + d_i p_{i,4}''(y_i) = \lim_{y \nearrow y_i} s''(y) \\ &= \lim_{y \searrow y_i} s''(y) = f_i p_{i+1,1}''(y_i) + d_i p_{i+1,2}''(y_i) + f_{i+1} p_{i+1,3}''(y_i) + d_{i+1} p_{i+1,4}''(y_i) \end{aligned}$$

gelten. Dank $\Phi_i(1) = y_i$ und $\Phi_{i+1}(-1) = y_i$ können wir alle Koeffizienten dieser Gleichung explizit berechnen und erhalten

$$\frac{6}{h^2} f_{i-1} + \frac{2}{h} d_{i-1} - \frac{6}{h^2} f_i + \frac{4}{h} d_i = -\frac{6}{h^2} f_i - \frac{4}{h} d_i + \frac{6}{h^2} f_{i+1} - \frac{2}{h} d_{i+1}.$$

Nun trennen wir die bekannten von den unbekannten Größen und finden

$$\frac{2}{h} d_{i-1} + \frac{8}{h} d_i + \frac{2}{h} d_{i+1} = \frac{6}{h^2} f_{i+1} - \frac{6}{h^2} f_{i-1},$$

$$d_{i-1} + 4d_i + d_{i+1} = \frac{3}{h}(f_{i+1} - f_{i-1}), \qquad (6.26)$$
$$\frac{d_{i-1} + 4d_i + d_{i+1}}{6} = \frac{f_{i+1} - f_{i-1}}{2h}.$$

Die letzte Gleichung dient nur der Illustration: Auf ihrer rechten Seite steht ein Differenzenquotient, der für $h \to 0$ gegen $f'(y_i)$ konvergiert (es ist nämlich gerade der zentrale Differenzenquotient, den wir in Beispiel 6.33 kennen gelernt haben), während ihre linke Seite in diesem Fall gegen d_i strebt, so dass wir $d_i \to f'(y_i)$ erhalten. Da $d_i = s'(y_i)$ gilt, ist das ein sehr willkommenes Ergebnis.

Mit Hilfe der Gleichung (6.26) können wir die Unbekannten $d_1, \ldots, d_{k-1} \in \mathbb{R}$ so bestimmen, dass die durch (6.25) definierte Funktion zweimal stetig differenzierbar und damit ein Element des Splineraums $\mathcal{S}_{3,(y_0,\ldots,y_k)}$ ist.

Die Unbekannten d_0 und d_k dagegen sind durch diese Gleichung noch nicht bestimmt, wir können also die Ableitung des Splines s an Start- und Endpunkt beliebig festlegen. Eine Möglichkeit sind die *natürlichen Randbedingungen*

$$s''(a) = 0, \qquad s''(b) = 0, \qquad (6.27)$$

die sich mit unserer transformierten Hermite-Basis in der Form

$$\begin{aligned}
0 = s''(a) &= f_0 p''_{1,1}(y_0) + d_0 p''_{1,2}(y_0) + f_1 p''_{1,3}(y_0) + d_1 p''_{1,4}(y_0) \\
&= -\frac{6}{h^2} f_0 - \frac{4}{h} d_0 + \frac{6}{h^2} f_1 - \frac{2}{h} d_1, \\
0 = s''(b) &= f_{k-1} p''_{k,1}(y_k) + d_{k-1} p''_{k,2}(y_k) + f_k p''_{k,3}(y_k) + d_k p''_{k,4}(y_k) \\
&= \frac{6}{h^2} f_{k-1} + \frac{2}{h} d_{k-1} - \frac{6}{h^2} f_k + \frac{4}{h} d_k
\end{aligned}$$

schreiben lassen. Wir trennen wieder bekannte und unbekannte Größen und erhalten

$$\begin{aligned}
\frac{1}{h}(4d_0 + 2d_1) &= \frac{6}{h^2}(f_1 - f_0), & \frac{1}{h}(2d_{k-1} + 4d_k) &= \frac{6}{h^2}(f_k - f_{k-1}), \\
2d_0 + d_1 &= \frac{3}{h}(f_1 - f_0), & d_{k-1} + 2d_k &= \frac{3}{h}(f_k - f_{k-1}), \\
\frac{2d_0 + d_1}{3} &= \frac{f_1 - f_0}{h}, & \frac{d_{k-1} + 2d_k}{3} &= \frac{f_k - f_{k-1}}{h}.
\end{aligned}$$

Auch hier dient die letzte Zeile nur der Illustration: Für $h \to 0$ konvergieren die rechten Seiten gegen $f'(a)$ beziehungsweise $f'(b)$, während die linken gegen $d_0 = s'(a)$ beziehungsweise $d_k = s'(b)$ streben. Zusammen mit (6.26) erhalten wir insgesamt das lineare Gleichungssystem

$$\begin{pmatrix} 2 & 1 & & & \\ 1 & 4 & 1 & & \\ & \ddots & \ddots & \ddots & \\ & & 1 & 4 & 1 \\ & & & 1 & 2 \end{pmatrix} \begin{pmatrix} d_0 \\ d_1 \\ \vdots \\ d_{k-1} \\ d_k \end{pmatrix} = \frac{3}{h} \begin{pmatrix} -1 & 1 & & & \\ -1 & 0 & 1 & & \\ & \ddots & \ddots & \ddots & \\ & & -1 & 0 & 1 \\ & & & -1 & 1 \end{pmatrix} \begin{pmatrix} f_0 \\ f_1 \\ \vdots \\ f_{k-1} \\ f_k \end{pmatrix},$$

das wir auflösen müssen, um die Unbekannten $d_0, \ldots, d_k \in \mathbb{R}$ zu bestimmen.

Lemma 6.42 (Lösbarkeit) *Sei*

$$\mathbf{M} := \begin{pmatrix} 2 & 1 & & & \\ 1 & 4 & 1 & & \\ & \ddots & \ddots & \ddots & \\ & & 1 & 4 & 1 \\ & & & 1 & 2 \end{pmatrix} \in \mathbb{R}^{n\times n}.$$

Dann gilt

$$\langle \mathbf{x}, \mathbf{M}\mathbf{x}\rangle_2 \geq \|\mathbf{x}\|_2^2 \qquad \text{für alle } \mathbf{x} \in \mathbb{K}^n.$$

Insbesondere ist die Matrix $\mathbf{M}$ *positiv definit und selbstadjungiert.*

Beweis. Sei $\mathbf{x} \in \mathbb{K}^n$. Es gilt

$$\begin{aligned}
\langle \mathbf{x}, \mathbf{M}\mathbf{x}\rangle_2 &= \bar{x}_1(2x_1 + x_2) + \sum_{i=2}^{n-1} \bar{x}_i(x_{i-1} + 4x_i + x_{i+1}) + \bar{x}_n(x_{n-1} + 2x_n) \\
&= 2|x_1|^2 + \bar{x}_1 x_2 + \sum_{i=2}^{n-1} \bar{x}_i x_{i-1} + 4\sum_{i=2}^{n-1} |x_i|^2 + \sum_{i=2}^{n-1} \bar{x}_i x_{i+1} + \bar{x}_n x_{n-1} + 2|x_n|^2 \\
&= 2|x_1|^2 + 4\sum_{i=2}^{n-1} |x_i|^2 + 2|x_n|^2 + \sum_{i=2}^{n} \bar{x}_i x_{i-1} + \sum_{i=2}^{n} \bar{x}_{i-1} x_i \\
&= |x_1|^2 + 2\sum_{i=2}^{n-1} |x_i|^2 + |x_n|^2 + \sum_{i=2}^{n} (|x_{i-1}|^2 + x_{i-1}\bar{x}_i + \bar{x}_{i-1}x_i + |x_i|^2) \\
&= |x_1|^2 + 2\sum_{i=2}^{n-1} |x_i|^2 + |x_n|^2 + \sum_{i=2}^{n} |x_{i-1} + x_i|^2 \\
&\geq |x_1|^2 + 2\sum_{i=2}^{n-1} |x_i|^2 + |x_n|^2 \geq \|\mathbf{x}\|_2^2,
\end{aligned}$$

wobei wir im vorletzten Schritt die Ungleichung

$$0 \leq |a+b|^2 = (a+b)(\bar{a}+\bar{b}) = |a|^2 + a\bar{b} + b\bar{a} + |b|^2 \qquad \text{für alle } a, b \in \mathbb{K}$$

verwendet haben. ∎

Aus diesem Resultat folgt, dass wir die Ableitungen $d_0, \ldots, d_k$ beispielsweise mit Hilfe der Cholesky- oder LR-Zerlegung berechnen können, um damit eine Darstellung der Splinefunktion s zu erhalten.

Da das Lösen des linearen Gleichungssystems lediglich $\sim k+1$ Rechenoperationen erfordert, können wir die Koeffizienten $d_0, \ldots, d_k$ relativ effizient bestimmen. Sobald die Koeffizienten bekannt sind, lässt sich die Splinefunktion für beliebiges x auswerten, indem das $i \in [1:k]$ mit $x \in [y_{i-1}, y_i]$ ermittelt und dann die Hermite-Basis ausgewertet wird. Ersteres können wir im allgemeinsten Fall mit einem Bisektionsalgorithmus in $\sim \log_2 k$ Operationen bewältigen, der Aufwand für letzteres ist konstant.

Übungsaufgabe 6.43 (Dimension) *Seien $a = y_0 < y_1 < \ldots < y_k = b$ und $m \in \mathbb{N}$ gegeben. Wir definieren die Funktionen*

$$p_i \colon \mathbb{R} \to \mathbb{R}, \qquad x \mapsto \begin{cases} (x - y_i)^m & \text{falls } x > y_i, \\ 0 & \text{ansonsten} \end{cases} \qquad \text{für alle } i \in [1 : k-1].$$

Wir wollen beweisen, dass diese Funktionen gerade den Unterschied zwischen dem „gewöhnlichen“ Polynomraum Π_m und dem Splineraum $\mathcal{S}_{m,(y_0,\ldots,y_k)}$ ausmachen.

(a) Beweisen Sie, dass $p_i \in \mathcal{S}_{m,(y_0,\ldots,y_k)}$ für alle $i \in [1 : k-1]$ gilt.

(b) Beweisen Sie, dass $\mathcal{S}_{m,(y_0,\ldots,y_k)} = \Pi_m + \operatorname{span}\{p_1, \ldots, p_{k-1}\}$ gilt.

(c) Beweisen Sie $\dim \mathcal{S}_{m,(y_0,\ldots,y_k)} = m + k$.

Hinweis: *Bei den Teilen (a) und (b) kann die Taylor-Entwicklung in den Punkten y_i nützlich sein, mit der sich Polynome exakt darstellen lassen.*

Übungsaufgabe 6.44 (Duale Basis) *Sei $m \in \mathbb{N}_0$, und seien $b_0, \ldots, b_m \in \Pi_m$ sowie lineare Abbildungen $\lambda_0, \ldots, \lambda_m : \Pi_m \to \mathbb{R}$ gegeben. Beweisen Sie: Falls*

$$\lambda_i(b_j) = \begin{cases} 1 & \text{falls } i = j, \\ 0 & \text{ansonsten} \end{cases} \qquad \text{für alle } i, j \in [0 : m] \tag{6.28}$$

erfüllt ist, ist $\{b_0, \ldots, b_m\}$ eine Basis des Raums Π_m und es gilt

$$p = \sum_{i=0}^{m} \lambda_i(p) b_i \qquad \text{für alle } p \in \Pi_m, \tag{6.29}$$

die Werte $\lambda_i(p)$ sind also gerade die Koeffizienten der Darstellung des Polynoms p in dieser Basis.

Zeigen Sie außerdem, dass aus (6.29) umgekehrt folgt, dass $\{b_0, \ldots, b_m\}$ eine Basis ist und (6.28) gilt. (Hinweis: Ist die Abbildung $(\alpha_0, \ldots, \alpha_m) \mapsto \sum_{i=0}^{m} \alpha_i b_i$ surjektiv?)

Übungsaufgabe 6.45 (Beispiele für duale Basen) *Beweisen Sie, dass die folgenden Kombinationen die Bedingung (6.28) erfüllen:*

- *Die Monome $b_i(x) = x^i$ und die Taylor-Koeffizienten $\lambda_i(p) = p^{(i)}(0)/i!$.*
- *Die Lagrange-Polynome $b_i(x) = \ell_i(x)$ und die Punktauswertungen $\lambda_i(p) = p(x_i)$.*
- *Die Newton-Polynome $b_i(x) = (x - x_0) \ldots (x - x_{i-1})$ und die dividierten Differenzen $\lambda_i(p) = d_{0,i}(p)$.*
- *Im Fall $m = 3$ die Polynome $b_0 = p_1, b_1 = p_2, b_2 = p_3, b_3 = p_4$ und die Punktauswertungen $\lambda_0(p) = p(-1), \lambda_1(p) = p'(-1), \lambda_2(p) = p(1), \lambda_3(p) = p'(1)$.*

7 Numerische Integration

Mit der Suche nach Nullstellen haben wir bereits einen Problemtyp kennen gelernt, der sich mit rein analytischen oder algebraischen Methoden nicht zufriedenstellend behandeln lässt, so dass numerische Näherungsverfahren zum Einsatz kommen müssen.

Dieses Kapitel beschäftigt sich mit einer weiteren Klasse von Problemen, die sich im Allgemeinen nur mit Hilfe numerischer Techiken behandeln lässt, nämlich mit der Integration: Sofern keine Stammfunktion des Integranden bekannt ist, lässt sich ein Integral im Allgemeinen nicht konkret berechnen.

Also sind numerische Verfahren von Interesse. Wir suchen nach Algorithmen, die die folgende Aufgabe lösen:

Gegeben seien $a, b \in \mathbb{R}$ mit $a < b$ und eine Funktion $f \in C[a, b]$, berechne

$$\mathcal{I}_{[a,b]}(f) := \int_a^b f(x)\,dx.$$

Traditionell bezeichnet man die betreffenden Algorithmen als *Quadraturverfahren*, da sie anschaulich die Fläche unter einer Kurve bestimmen und diese Fläche in Vielfachen eines Einheitsquadrats gemessen wird.

7.1 Quadratur per Interpolation

Wir untersuchen zunächst Quadraturverfahren auf dem *Referenzintervall* $[-1, 1]$. Integrale auf allgemeinen Intervallen können wir mit einer einfachen Transformation auf diesen Fall zurückführen.

Zur Abkürzung bezeichnen wir das Integral mit

$$\mathcal{I}\colon C[-1,1] \to \mathbb{R}, \qquad f \mapsto \int_{-1}^{1} f(x)\,dx.$$

Da das Riemann-Integral mit Hilfe einer gewichteten Summe von Werten von f definiert wird, bietet es sich an, nach Quadraturverfahren zu suchen, die mit ähnlich geringen Voraussetzungen auskommen.

Definition 7.1 (Quadraturformel) *Seien ein $m \in \mathbb{N}_0$, $x_0, \ldots, x_m \in [-1, 1]$ und $w_0, \ldots, w_m \in \mathbb{R}$ gegeben. Die Abbildung*

$$\mathcal{Q}\colon C[-1,1] \to \mathbb{R}, \qquad f \mapsto \sum_{i=0}^{m} w_i f(x_i) \tag{7.1}$$

nennen wir dann die Quadraturformel *der Stufe m zu den* Stützstellen $x_0, \ldots, x_m$ *und den* Gewichten $w_0, \ldots, w_m$.

Beispiel 7.2 (Mittelpunktregel) *Die besonders einfache Quadraturformel*

$$\mathcal{M}\colon C[-1,1] \to \mathbb{R}, \qquad f \mapsto 2\,f(0),$$

heißt Mittelpunktregel.

Mit partieller Integration erhalten wir

$$\begin{aligned}\mathcal{I}(f) - \mathcal{M}(f) &= \int_{-1}^{1} f(x) - f(0)\,dx \\ &= [x(f(x) - f(0))]_{-1}^{1} - \int_{-1}^{1} x f'(x)\,dx \\ &= f(1) - f(0) + f(-1) - f(0) - \int_{-1}^{1} x f'(x)\,dx,\end{aligned}$$

und mit dem Hauptsatz der Integral- und Differentialrechnung folgt

$$\begin{aligned}\mathcal{I}(f) - \mathcal{M}(f) &= \int_{0}^{1} f'(x)\,dx + \int_{-1}^{0} -f'(x)\,dx - \int_{-1}^{1} x f'(x)\,dx \\ &= \int_{0}^{1} (1-x) f'(x)\,dx - \int_{-1}^{0} (1+x) f'(x)\,dx,\end{aligned}$$

woraus wir mit einer weiteren partiellen Integration die Gleichung

$$\begin{aligned}\mathcal{I}(f) - \mathcal{M}(f) &= \left[-\frac{(1-x)^2}{2} f'(x)\right]_0^1 + \int_0^1 \frac{(1-x)^2}{2} f''(x)\,dx \\ &\quad - \left[\frac{(1+x)^2}{2} f'(x)\right]_{-1}^0 + \int_{-1}^0 \frac{(1+x)^2}{2} f''(x)\,dx \\ &= \int_0^1 \frac{(1-x)^2}{2} f''(x)\,dx + \int_{-1}^0 \frac{(1+x)^2}{2} f''(x)\,dx \qquad (7.2)\end{aligned}$$

erhalten. Mit Hilfe des Mittelwertsatzes der Integralrechnung finden wir $\eta_+ \in [0,1]$ *und* $\eta_- \in [-1,0]$ *mit*

$$\mathcal{I}(f) - \mathcal{M}(f) = f''(\eta_+) \int_0^1 \frac{(1-x)^2}{2}\,dx + f''(\eta_-) \int_{-1}^0 \frac{(1+x)^2}{2}\,dx = \frac{1}{6}(f''(\eta_+) + f''(\eta_-)),$$

und mit dem Zwischenwertsatz für stetige Funktionen finden wir ein $\eta \in [\eta_-, \eta_+]$ *mit*

$$\mathcal{I}(f) - \mathcal{M}(f) = \frac{f''(\eta)}{3}.$$

Damit haben wir eine explizite Darstellung des Quadraturfehlers gewonnen.

Falls f vektorwertig ist, können wir ausgehend von (7.2) immerhin noch

$$\|\mathcal{I}(f) - \mathcal{M}(f)\| \le \int_0^1 \frac{(1-x)^2}{2} \|f''(x)\|\,dx + \int_{-1}^0 \frac{(1+x)^2}{2} \|f''(x)\|\,dx$$

$$\leq \int_0^1 \frac{(1-x)^2}{2}\|f''\|_{\infty,[-1,1]}\,dx + \int_{-1}^0 \frac{(1+x)^2}{2}\|f''\|_{\infty,[-1,1]}\,dx$$
$$= \frac{1}{6}\|f''\|_{\infty,[-1,1]} + \frac{1}{6}\|f''\|_{\infty,[-1,1]} = \frac{1}{3}\|f''\|_{\infty,[-1,1]}$$

erhalten, indem wir die Dreiecksungleichung für Integrale anwenden.

Der Beweis der Fehlerdarstellung für die Mittelpunktregel deutet bereits an, wie man allgemein vorgehen könnte: Es gilt

$$\mathcal{M}(f) = \int_{-1}^1 f(0)\,dx,$$

die Quadraturformel kann also als Integral einer Approximation von f interpretiert werden, in diesem Fall als das Integral einer konstanten Approximation.

Dieser Ansatz lässt sich verallgemeinern: Statt f nur in einem Punkt auszuwerten, können wir mehrere Punkte $-1 \leq x_0 < \ldots < x_m \leq 1$ fixieren und auf der Grundlage der Werte $f(x_0), \ldots, f(x_m)$ versuchen, eine Approximation von f zu konstruieren, die sich einfach integrieren lässt. Eine naheliegende Wahl besteht darin, einen polynomiellen Interpolanten zu verwenden, also ein $p \in \Pi_m$ mit

$$p(x_i) = f(x_i) \qquad \text{für alle } i \in [0:m].$$

Gemäß (6.2) lässt sich p mit Hilfe der Lagrange-Polynome $\ell_0, \ldots, \ell_m$ in der Form

$$p(x) = \sum_{i=0}^m f(x_i)\ell_i(x) \qquad \text{für alle } x \in \mathbb{R}$$

darstellen, so dass das Integral von p die Form

$$\mathcal{I}(p) = \int_{-1}^1 p(x)\,dx = \sum_{i=0}^m f(x_i)\int_{-1}^1 \ell_i(x)\,dx$$

annimmt und wir mit

$$w_i := \int_{-1}^1 \ell_i(x)\,dx \qquad \text{für alle } i \in [0:m] \tag{7.3}$$

eine Quadraturformel

$$\mathcal{Q}(f) := \sum_{i=0}^m w_i f(x_i)$$

erhalten. Für ein Polynom $f \in \Pi_m$ folgt aus dem Identitätssatz sofort $f = p$, also auch $\mathcal{Q}(f) = \mathcal{I}(f)$, die so definierte Quadraturformel wird also immerhin für Polynome das korrekte Integral berechnen.

Definition 7.3 (Interpolatorische Quadraturformel) *Wir nennen eine Quadraturformel $\mathcal{Q}$ mit den Stützstellen $x_0, \ldots, x_m$ und Gewichten $w_0, \ldots, w_m$* interpolatorisch*, falls die Stützstellen paarweise verschieden sind und die Gleichungen*

$$w_i = \int_{-1}^{1} \ell_i(x)\,dx \qquad \text{für alle } i \in [0:m] \tag{7.4}$$

gelten, wobei ℓ_i das i-te Lagrange-Polynom (6.1) bezeichnet.

Für die praktische Berechnung der Gewichte ist die Formel (7.4 eher unpraktisch, insbesondere bei höheren Polynomgraden, da die Integration der Lagrange-Polynome durchaus umständlich werden kann. Glücklicherweise gibt es eine Alternative: Wenn $\mathcal{Q}$ eine interpolatorische Quadraturformel mit den Interpolationspunkten $x_0, \ldots, x_m \in [a,b]$ ist, müssen insbesondere alle Polynome m-ten Grades exakt integriert werden, denn sie werden durch die Interpolation exakt reproduziert. Für jedes $p \in \Pi_m$ gilt also

$$\sum_{k=0}^{m} w_k\, p(x_k) = \mathcal{Q}(p) = \mathcal{I}(p) = \int_{-1}^{1} p(x)\,dx.$$

Wir können Polynome p wählen, deren Integrale besonders einfach zu berechnen sind, beispielsweise die Monome $p(x) = x^j$ für $j \in [0:m]$. Es folgt

$$\sum_{k=0}^{m} x_k^j w_k = \int_{-1}^{1} x^j\,dx = \begin{cases} \frac{2}{j+1} & \text{falls } j \text{ gerade,} \\ 0 & \text{falls } j \text{ ungerade} \end{cases} \qquad \text{für alle } j \in [0:m]. \tag{7.5}$$

Wir haben ein lineares Gleichungssystem mit $m+1$ Gleichungen für die $m+1$ Gewichte $w_0, \ldots, w_m$ gefunden. Es stellt sich lediglich die Frage, ob die durch

$$a_{jk} := x_k^j \qquad \text{für alle } j,k \in [0:m]$$

gegebene *Vandermonde-Matrix* $A \in \mathbb{R}^{[0:m]\times[0:m]}$ regulär ist. Diese Frage lässt sich elegant mit Hilfe der Lagrange-Polynome beantworten: Für $i \in [0:m]$ ist $\ell_i \in \Pi_m$, und da die Monome eine Basis des Polynomraums bilden, existieren Koeffizienten $b_{i0}, \ldots, b_{im} \in \mathbb{R}$ mit

$$\ell_i(x) = \sum_{j=0}^{m} b_{ij}\, x^j \qquad \text{für alle } x \in \mathbb{R}.$$

Wir fassen diese Koeffizienten zu einer Matrix $B \in \mathbb{R}^{[0:m]\times[0:m]}$ zusammen und erhalten

$$(BA)_{ik} = \sum_{j=0}^{m} b_{ij}a_{jk} = \sum_{j=0}^{m} b_{ij}x_k^j = \ell_i(x_k) = \begin{cases} 1 & \text{falls } i = k, \\ 0 & \text{ansonsten} \end{cases} \qquad \text{für alle } i,k \in [0:m],$$

also gilt $BA = I$, und damit muss A injektiv sein, also als quadratische Matrix auch regulär. Also können wir das lineare Gleichungssystem (7.5) lösen, um die Gewichte zu bestimmen, ohne Lagrange-Polynome per Hand zu integrieren.

Die Stützstellen $x_0, \dots, x_m$ haben wir bei dieser Konstruktion noch nicht näher festgelegt. Eine besonders naheliegende Wahl besteht darin,

$$x_i := -1 + \frac{2i}{m} \qquad \text{für alle } i \in [0:m] \tag{7.6}$$

zu verwenden, die Stützstellen also äquidistant zu wählen.

Die Kombination dieser Stützstellen mit den nach (7.3) festgelegten Gewichten führt zu einer speziellen Klasse von Quadraturformeln:

Definition 7.4 (Newton-Cotes-Quadraturformeln) *Sei $m \in \mathbb{N}$, und seien Stützstellen $x_0, \dots, x_m$ sowie Gewichte $w_0, \dots, w_m$ durch (7.6) und (7.3) definiert. Dann heißt*

$$\mathcal{N}_m \colon C[-1,1] \to \mathbb{R}, \qquad f \mapsto \sum_{i=0}^{m} w_i f(x_i),$$

die m-te Newton-Cotes-Quadraturformel.

Sie besitzt die Eigenschaft, dass $\mathcal{N}_m(p) = \mathcal{I}(p)$ für alle Polynome $p \in \Pi_m$ gilt.

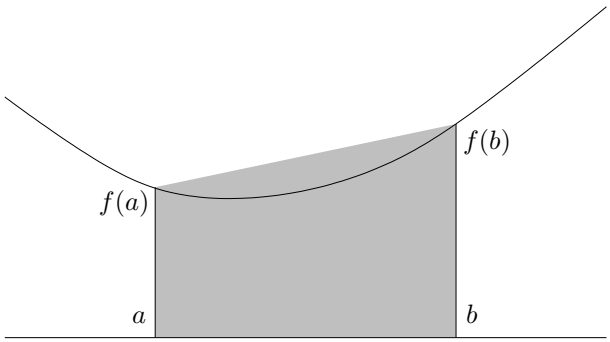

Abbildung 7.1: Approximation des Integrals mit Hilfe der Trapezregel

Beispiel 7.5 (Trapezregel) *Ein wichtiges Beispiel für eine Newton-Cotes-Formel ist die* Trapezregel

$$\mathcal{T} \colon C[-1,1] \to \mathbb{R}, \qquad f \mapsto f(-1) + f(1),$$

die ihren Namen der Vorstellung verdankt, dass die dem Integral entsprechende Fläche unter der Kurve durch ein Trapez approximiert wird (siehe Abbildung 7.1).

Übungsaufgabe 7.6 (Trapezregel) *Sei $f \in C^2[-1,1]$. Beweisen Sie, dass ein $\eta \in [-1,1]$ mit*

$$\mathcal{I}(f) - \mathcal{T}(f) = -\frac{2}{3} f''(\eta)$$

existiert.

Hinweis: Der Beweis lässt sich analog zu Beispiel 7.2 führen, indem man den Fehler als Integral über $f-p$ mit einem geeigneten $p \in \Pi_1$ schreibt. Die Funktion $\varphi(x) = (x^2-1)/2$ erfüllt $\varphi'' = 1$, also kann ein Produkt mit φ'' eingefügt und zweimal partiell integriert werden.

Übungsaufgabe 7.7 (vektorwertige Trapezregel) *Sei $\mathcal{V}$ ein normierter Vektorraum und sei $f \in C^2([-1,1],\mathcal{V})$. Integral und Trapezregel sind auch für vektorwertige Funktionen definiert, so dass sich der Quadraturfehler weitestgehend mit denselben Methoden wie in Übungsaufgabe 7.6 analysieren lässt. Der Mittelwertsatz der Integralrechnung steht uns allerdings nicht mehr zur Verfügung, so dass wir lediglich die Abschätzung des Fehlers erwarten dürfen: Beweisen Sie*

$$\|\mathcal{I}(f) - \mathcal{T}(f)\|_V \leq \frac{2}{3}\|f''\|_{\infty,[a,b]}.$$

Ein weiteres Beispiel ist die zweite Newton-Cotes-Formel, die auch als *Simpsonregel* bezeichnet wird und durch

$$\mathcal{S}\colon C[-1,1] \to \mathbb{R}, \qquad f \mapsto \frac{f(-1) + 4\,f(0) + f(1)}{3} \tag{7.7}$$

gegeben ist.

Übungsaufgabe 7.8 (Simpson-Regel) *Sei $f \in C^4[-1,1]$, sei p das dazugehörige quadratische Interpolationspolynom in den Punkten $x_0 = -1$, $x_1 = 0$ und $x_2 = 1$. Sei*

$$\varphi\colon [0,1] \to \mathbb{R}, \qquad x \mapsto \frac{(x-1)^3(3x+1)}{72}.$$

Wir wollen den Fehler der Simpson-Regel näher bestimmen.

(a) Beweisen Sie

$$\varphi^{(4)} = 1, \qquad \varphi(1) = \varphi'(1) = \varphi''(1) = 0, \qquad \varphi'(0) = 0.$$

(b) Folgern Sie daraus mit partieller Integration

$$\int_0^1 f(x) - p(x)\,dx = \int_0^1 f^{(4)}(x)\,\varphi(x)\,dx - \big(f'(0) - p'(0)\big)\varphi''(0) - f'''(0)\,\varphi(0),$$
$$\int_{-1}^0 f(x) - p(x)\,dx = \int_{-1}^0 f^{(4)}(x)\,\varphi(-x)\,dx + \big(f'(0) - p'(0)\big)\varphi''(0) + f'''(0)\,\varphi(0).$$

(c) Beweisen Sie, dass ein $\eta \in [-1,1]$ existiert mit

$$\mathcal{I}(f) - \mathcal{S}(f) = \int_{-1}^1 f(x) - p(x)\,dx = -\frac{f^{(4)}(\eta)}{90}.$$

Hinweis: *Sie können sich an dem Beispiel 7.2 orientieren.*

Übungsaufgabe 7.9 (3/8-Regel) *Berechnen Sie die Gewichte der dritten Newton-Cotes-Quadraturformel $\mathcal{N}_3$.*

7.2 Fehleranalyse

Die in den Beispielen 7.2 und 7.5 angegebenen Fehlerdarstellungen haben zwar den Vorteil, explizite Aussagen über die Genauigkeit einer Quadraturformel zu ermöglichen, sie müssen allerdings in der Regel für jede neue Quadraturformel „in Handarbeit“ hergeleitet werden. Außerdem haben sie den Nachteil, dass sie keine Aussagen für den Fall eines nur einmal differenzierbaren Integranden ermöglichen.

Deshalb sind wir an einer verallgemeinerten Theorie interessiert, die zwar keine exakten Fehlerdarstellungen mehr bietet, aber dafür wesentlich flexibler anwendbar ist.

Die Theorie basiert auf zwei Komponenten: Erstens können wir ausnutzen, dass Quadraturformeln für bestimmte Teilräume, etwa Polynome eines gewissen Grades, das exakte Integral berechnen. Zweitens bietet uns ein geeigneter Stabilitätsbegriff die Möglichkeit, Aussagen darüber zu treffen, wie eine Quadraturformel reagiert, wenn wir ihr Argument durch eine Approximation aus dem zuvor erwähnten Teilraum ersetzen.

Wir wenden uns zunächst der Stabilität zu:

Lemma 7.10 (Stabilität) *Sei $\mathcal{Q}$ eine Quadraturformel der Stufe m mit Stützstellen $x_0, \dots, x_m$ und Gewichten $w_0, \dots, w_m$. Dann heißt*

$$C_Q := \sum_{i=0}^{m} |w_i|$$

die Stabilitätskonstante *von $\mathcal{Q}$. Sie erfüllt*

$$|\mathcal{Q}(f)| \leq C_Q \|f\|_{\infty,[-1,1]} \qquad \text{für alle } f \in C[-1,1].$$

Beweis. Sei $f \in C[-1,1]$. Die Stabilitätsabschätzung ergibt sich direkt aus der Dreiecksungleichung:

$$|\mathcal{Q}(f)| = \left| \sum_{i=0}^{m} w_i f(x_i) \right| \leq \sum_{i=0}^{m} |w_i| \, |f(x_i)| \leq \sum_{i=0}^{m} |w_i| \, \|f\|_{\infty,[-1,1]} = C_Q \|f\|_{\infty,[-1,1]}.$$

∎

Da $\mathcal{Q}$ linear ist, erhalten wir aus dieser Abschätzung sofort

$$|\mathcal{Q}(f) - \mathcal{Q}(\tilde{f})| = |\mathcal{Q}(f - \tilde{f})| \leq C_Q \|f - \tilde{f}\|_{\infty,[-1,1]} \qquad \text{für alle } f, \tilde{f} \in C[-1,1],$$

Störungen des Integranden werden also schlimmstenfalls mit einem Faktor von C_Q verstärkt.

Ideal wäre es, wenn wir $\tilde{f}$ so wählen könnten, dass $\mathcal{Q}(\tilde{f}) = \mathcal{I}(\tilde{f})$ gilt, denn dann wären wir dem Ziel, das Integral zu approximieren, einen erheblichen Schritt näher gekommen. Für $\tilde{f} \in \Pi_1$ beispielsweise gilt $\tilde{f}'' = 0$, so dass aus den Fehlerdarstellungen der Beispiele 7.2 und 7.5 bereits

$$\mathcal{I}(\tilde{f}) = \mathcal{M}(\tilde{f}) = \mathcal{T}(\tilde{f})$$

folgt, lineare Polynome werden demnach sowohl von der Mittelpunkt- als auch von der Trapezregel exakt integriert. Auf eine besonders gute Approximation des Integrals dürfen wir also hoffen, falls Polynome möglichst hohen Grades exakt integriert werden.

Definition 7.11 (Exakte Quadraturformeln) *Sei $\mathcal{Q}$ eine Quadraturformel. Wir nennen sie* exakt *für Polynome n-ten Grades (oder kurz Quadraturformel n-ten Grades), falls*

$$\mathcal{Q}(p) = \mathcal{I}(p) = \int_{-1}^{1} p(x)\,dx \qquad \text{für alle } p \in \Pi_n$$

gilt, falls also die Formel für Polynome n-ten Grades das exakte Integral berechnet.

Interpolatorische Quadraturformeln sind mindestens für Polynome m-ten Grades exakt, da diese bei der Interpolation exakt reproduziert werden.

Nun können wir den beschriebenen Plan in die Tat umsetzen und eine polynomielle Approximation mit der Stabilitätsabschätzung kombinieren, um eine einfache Abschätzung für die Genauigkeit der Quadraturformel zu erhalten.

Lemma 7.12 (Bestapproximation) *Sei $n \in \mathbb{N}_0$. Sei eine Funktion $f \in C[a,b]$ gegeben, und sei $\mathcal{Q}$ eine Quadraturformel mit der Stabilitätskonstanten C_Q, die für Polynome n-ten Grades exakt ist. Dann gilt*

$$|\mathcal{I}(f) - \mathcal{Q}(f)| \leq (2 + C_Q)\|f - q\|_{\infty,[-1,1]} \qquad \text{für alle } q \in \Pi_n,$$

der Quadraturfehler wird also dadurch bestimmt, wie gut sich f durch Polynome n-ten Grades approximieren lässt.

Beweis. Sei $q \in \Pi_n$, und sei $e := f - q$ der Approximationsfehler. Aus der Monotonie des Integrals folgt

$$|\mathcal{I}(e)| = \left|\int_{-1}^{1} e(x)\,dx\right| \leq \int_{-1}^{1} |e(x)|\,dx \leq \int_{-1}^{1} \|e\|_{\infty,[-1,1]}\,dx = 2\|e\|_{\infty,[-1,1]},$$

und aus Lemma 7.10 folgt

$$|\mathcal{Q}(e)| \leq C_Q\|e\|_{\infty,[-1,1]}.$$

Da $\mathcal{Q}$ eine Formel n-ten Grades ist, gilt $\mathcal{Q}(q) = \mathcal{I}(q)$, und wir erhalten

$$\begin{aligned}|\mathcal{I}(f) - \mathcal{Q}(f)| &= |\mathcal{I}(f) - \mathcal{I}(q) + \mathcal{Q}(q) - \mathcal{Q}(f)| \leq |\mathcal{I}(f-q)| + |\mathcal{Q}(q-f)| \\ &= |\mathcal{I}(e)| + |\mathcal{Q}(e)| \leq 2\|e\|_{\infty,[-1,1]} + C_Q\|e\|_{\infty,[-1,1]} \\ &= (2 + C_Q)\|e\|_{\infty,[-1,1]} = (2 + C_Q)\|f - q\|_{\infty,[-1,1]},\end{aligned}$$

und das ist die gesuchte Abschätzung. ∎

Aus diesem Resultat können wir eine Abschätzung für den Quadraturfehler gewinnen, indem wir ein geeignetes Polynom einsetzen, beispielsweise das uns aus Folgerung 6.18 bereits bekannte Tschebyscheff-Interpolationspolynom. Dabei ist zu beachten, dass wir das Interpolationspolynom n-ten Grades verwenden, obwohl die interpolatorische Quadraturformel lediglich auf der Interpolation m-ten Grades beruht. Dieses Interpolationspolynom wird im Algorithmus nicht benötigt, es spielt nur als theoretisches Hilfsmittel in dem Beweis eine Rolle.

Lemma 7.13 (Genauigkeit) *Sei $n \in \mathbb{N}_0$, und sei $\mathcal{Q}$ eine Quadraturformel n-ten Grades. Dann gilt*

$$|\mathcal{I}(f) - \mathcal{Q}(f)| \leq \frac{2 + C_Q}{2^n} \frac{\|f^{(n+1)}\|_{\infty,[-1,1]}}{(n+1)!} \qquad \text{für alle } f \in C^{n+1}[-1,1].$$

Beweis. Wir verwenden Folgerung 6.18, um ein Polynom $q \in \Pi_n$ zu finden, das die Abschätzung

$$\|f - q\|_{\infty,[-1,1]} \leq 2^{-n} \frac{\|f^{(n+1)}\|_{\infty,[-1,1]}}{(n+1)!}.$$

erfüllt, nämlich gerade das Polynom, das f in den Tschebyscheff-Punkten n-ten Grades interpoliert. Durch Einsetzen dieses Polynoms in Lemma 7.12 erhalten wir die gewünschte Abschätzung. ∎

Übungsaufgabe 7.14 (Interpolatorische Quadratur) *Sei $\mathcal{Q}$ eine Quadraturformel m-ter Stufe mit paarweise verschiedenen Stützstellen, die für Polynome m-ten Grades exakt ist. Beweisen Sie, dass $\mathcal{Q}$ dann eine interpolatorische Quadraturformel ist.*

7.3 Transformierte und zusammengesetzte Quadraturformeln

Die Fehlerdarstellungen und -abschätzungen, die wir bisher kennen gelernt haben, waren erstens nur für das Referenzintervall $[-1,1]$ formuliert und boten uns zweitens keine Möglichkeit, die Genauigkeit zu verbessern: Falls die zweite Ableitung des Integranden groß ist, wird die Mittelpunktregel nur eine grobe Näherung des Integrals berechnen, und die Fehlerdarstellung aus Beispiel 7.2 impliziert, dass sich daran auch nichts ändern lässt.

Beide Probleme lassen sich lösen, indem wir Transformationen von allgemeinen Intervallen $[a,b]$ auf das Referenzintervall untersuchen. Einerseits erschließen sich uns dadurch Näherungsverfahren für allgemeine Integrale, andererseits ändern sich durch die Transformation die in den Fehlerabschätzungen auftretenden Ableitungen, so dass wir den Fehler beeinflussen können.

Um eine Approximation eines Integrals über dem Intervall $[a,b]$ zu erhalten, verwenden wir den Transformationssatz: Mit der Abbildung

$$\Phi_{[a,b]} \colon [-1,1] \to [a,b] \qquad x \mapsto \frac{b+a}{2} + \frac{b-a}{2} x$$

können wir $[-1,1]$ auf ein beliebiges Intervall $[a,b]$ abbilden, und für eine m-stufige Quadraturformel $\mathcal{Q}$ der Gestalt (7.1) folgt

$$\begin{aligned}\mathcal{I}_{[a,b]}(f) &:= \int_a^b f(x)\,dx = \int_{-1}^1 \Phi'_{[a,b]}(\hat{x}) f(\Phi_{[a,b]}(\hat{x}))\,d\hat{x} = \frac{b-a}{2} \int_{-1}^1 f \circ \Phi_{[a,b]}(\hat{x})\,d\hat{x} \\ &= \frac{b-a}{2} \mathcal{I}(f \circ \Phi_{[a,b]}) \approx \frac{b-a}{2} \mathcal{Q}(f \circ \Phi_{[a,b]}) = \frac{b-a}{2} \sum_{i=0}^m w_i f(\Phi_{[a,b]}(x_i)).\end{aligned}$$

Eine gute Quadraturformel für das Intervall $[-1,1]$ lässt sich somit mit Hilfe der Transformation $\Phi_{[a,b]}$ einfach zu einer guten Quadraturformel für ein Intervall $[a,b]$ ausbauen.

Definition 7.15 (Transformierte Quadraturformel) *Sei $\mathcal{Q}$ eine m-stufige Quadraturformel mit Stützstellen $x_0,\dots,x_m$ und Gewichten $w_0,\dots,w_m$. Seien $a,b\in\mathbb{R}$ mit $a<b$ gegeben. Dann heißt die Abbildung*

$$\mathcal{Q}_{[a,b]}\colon C[a,b]\to\mathbb{R}, \qquad f\mapsto\frac{b-a}{2}\sum_{i=0}^{m}w_i f(\Phi_{[a,b]}(x_i))$$

die zu $\mathcal{Q}$ gehörende transformierte Quadraturformel *auf $[a,b]$.*

Offenbar lässt sich die transformierte Quadraturformel mit

$$\hat{w}_i:=\frac{b-a}{2}w_i, \qquad \hat{x}_i:=\Phi_{[a,b]}(x_i)=\frac{b+a}{2}+\frac{b-a}{2}x_i \qquad \text{für alle } i\in[0:m]$$

in der vertrauten Form

$$\mathcal{Q}_{[a,b]}(f)=\sum_{i=0}^{m}\hat{w}_i f(\hat{x}_i) \qquad \text{für alle } f\in C[a,b]$$

darstellen, die für die konkrete Umsetzung im Rechner vorteilhaft sein kann.

Um eine Aussage über die Genauigkeit der transformierten Quadraturformel zu gewinnen, können wir uns auf die Analyse der Ableitungen der bei der Transformation verwendeten Funktion $\hat{f}:=f\circ\Phi_{[a,b]}$ stützen, die wir in Lemma 6.20 gewonnen haben: Es gilt

$$\hat{f}^{(n)}(x)=\left(\frac{b-a}{2}\right)^n f^{(n)}\circ\Phi_{[a,b]}(x) \qquad \text{für alle } x\in[-1,1].$$

Zunächst wenden wir diese Gleichung auf die explizite Fehlerdarstellung aus Beispiel 7.2 an: Für die Mittelpunktregel gelten die Gleichungen

$$\begin{aligned}\mathcal{I}_{[a,b]}(f)-\mathcal{M}_{[a,b]}(f)&=\frac{b-a}{2}\left(\mathcal{I}(\hat{f})-\mathcal{M}(\hat{f})\right)=\frac{b-a}{2}\frac{\hat{f}''(\eta)}{3}\\&=\frac{b-a}{2}\frac{(b-a)^2}{4}\frac{f''\circ\Phi_{[a,b]}(\eta)}{3}=\frac{(b-a)^3}{24}f''(\hat{\eta})\end{aligned}\tag{7.8}$$

mit einem $\eta\in[-1,1]$, wobei wir $\hat{\eta}:=\Phi_{[a,b]}(\eta)\in[a,b]$ verwenden, sowie im Fall einer vektorwertigen Funktion

$$\begin{aligned}\|\mathcal{I}_{[a,b]}(f)-\mathcal{M}_{[a,b]}(f)\|&=\frac{b-a}{2}\|\mathcal{I}(\hat{f})-\mathcal{M}(\hat{f})\|\\&\le\frac{b-a}{2}\frac{\|\hat{f}''\|_{\infty,[-1,1]}}{3}=\frac{(b-a)^3}{24}\|f''\|_{\infty,[a,b]}.\end{aligned}\tag{7.9}$$

Entsprechend können wir auch mit der Trapezregel aus Beispiel 7.5 verfahren und erhalten

$$\mathcal{I}_{[a,b]}(f) - \mathcal{T}_{[a,b]}(f) = -\frac{(b-a)^3}{12} f''(\hat{\eta})$$

für ein $\hat{\eta} \in [a,b]$. Auch die allgemeine Fehlerabschätzung aus Lemma 7.13 können wir auf diese Weise auf beliebige Intervalle übertragen:

Lemma 7.16 (Genauigkeit) *Sei $n \in \mathbb{N}_0$, sei $\mathcal{Q}$ eine Quadraturformel n-ten Grades, und seien $a, b \in \mathbb{R}$ mit $a < b$ gegeben. Dann gilt*

$$\left|\mathcal{I}_{[a,b]}(f) - \mathcal{Q}_{[a,b]}(f)\right| \leq 4(2 + C_Q) \left(\frac{b-a}{4}\right)^{n+2} \frac{\|f^{(n+1)}\|_{\infty,[a,b]}}{(n+1)!} \quad \text{für alle } f \in C^{n+1}[a,b].$$

Beweis. Sei $f \in C^{n+1}[a,b]$ gegeben und $\hat{f} := f \circ \Phi_{[a,b]}$. Wegen Lemma 6.20 gilt

$$\hat{f}^{(n+1)}(x) = \left(\frac{b-a}{2}\right)^{n+1} f^{(n+1)} \circ \Phi_{[a,b]}(x) \qquad \text{für alle } x \in [-1,1],$$

und damit

$$\|\hat{f}^{(n+1)}\|_{\infty,[-1,1]} = \left(\frac{b-a}{2}\right)^{n+1} \|f^{(n+1)}\|_{\infty,[a,b]}.$$

Aus Lemma 7.13 erhalten wir nun

$$\begin{aligned}
\left|\mathcal{I}_{[a,b]}(f) - \mathcal{Q}_{[a,b]}(f)\right| &= \left|\frac{b-a}{2}\mathcal{I}(\hat{f}) - \frac{b-a}{2}\mathcal{Q}(\hat{f})\right| \\
&= \frac{b-a}{2}\left|\mathcal{I}(\hat{f}) - \mathcal{Q}(\hat{f})\right| \\
&\leq \frac{b-a}{2}\frac{2+C_Q}{2^n}\frac{\|\hat{f}^{(n+1)}\|_{\infty,[-1,1]}}{(n+1)!} \\
&= \frac{b-a}{2}\frac{2+C_Q}{2^n}\left(\frac{b-a}{2}\right)^{n+1}\frac{\|f^{(n+1)}\|_{\infty,[a,b]}}{(n+1)!} \\
&= \frac{2+C_Q}{2^n}\left(\frac{b-a}{2}\right)^{n+2}\frac{\|f^{(n+1)}\|_{\infty,[a,b]}}{(n+1)!} \\
&= 4(2+C_Q)\left(\frac{b-a}{4}\right)^{n+2}\frac{\|f^{(n+1)}\|_{\infty,[a,b]}}{(n+1)!},
\end{aligned}$$

und das ist die gewünschte Abschätzung. ∎

Offenbar ist der Fehler einer Quadraturformel um so geringer, je kleiner das Intervall ist, auf das wir sie anwenden. Um die Genauigkeit zu verbessern, wäre es also erstrebenswert, die Intervalle zu verkleinern.

Dabei können wir wie bei der Approximation mit stückweisen Polynomen vorgehen: Wir zerlegen das Intervall $[a,b]$ in Teilintervalle und approximieren den Integranden auf jedem der Teilintervalle durch ein Polynom. Dann integrieren wir diese stückweise

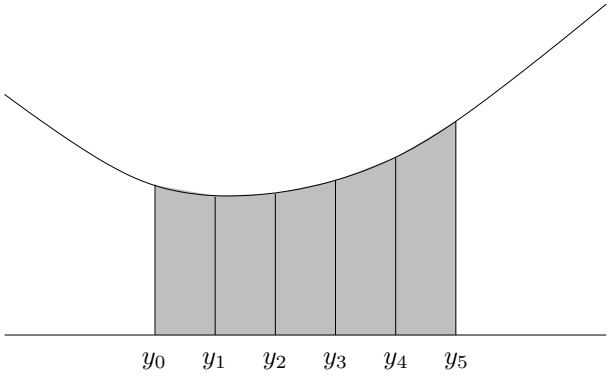

Abbildung 7.2: Approximation des Integrals mit Hilfe der summierten Trapezregel

Approximation. Aufgrund der Eigenschaften des Integrals können wir diesen Ansatz besonders elegant formulieren, indem wir das Integral über das stückweise Polynom als Summe der Integrale über die Teilintervalle schreiben.

Die einfachste Lösung besteht wieder darin, das gegebene Intervall $[a, b]$ äquidistant zu zerlegen, also in $k \in \mathbb{N}$ gleich große Teilintervalle: Wir führen

$$y_i := a + hi, \qquad h := \frac{b-a}{k} \qquad \text{für alle } i \in [0:m] \tag{7.10}$$

ein und zerlegen das Integral über $[a, b]$ in k Integrale über die Teilintervalle $[y_{i-1}, y_i]$, die wir dann mit Hilfe der passend transformierten Quadraturformel approximieren:

$$\int_a^b f(x)\,dx = \sum_{i=1}^{k} \int_{y_{i-1}}^{y_i} f(x)\,dx \approx \sum_{i=1}^{k} \mathcal{Q}_{[y_{i-1},y_i]}(f).$$

In dieser Weise können wir zu einer beliebigen Quadraturformel eine *summierte Quadraturformel* definieren, die mit einem k-mal höheren Rechenaufwand eine hoffentlich k^{n+1}-mal bessere Genauigkeit erreicht.

Definition 7.17 (Summierte Quadraturformel) *Sei $\mathcal{Q}$ eine Quadraturformel, und seien $a, b \in \mathbb{R}$ mit $a < b$ gegeben. Sei $k \in \mathbb{N}$, und seien $y_0, \ldots, y_k \in [a, b]$ wie in (7.10) definiert. Dann nennen wir*

$$\mathcal{Q}_{[a,b],k} \colon C[a,b] \to \mathbb{R}, \qquad f \mapsto \sum_{i=1}^{k} \mathcal{Q}_{[y_{i-1},y_i]}(f),$$

die zu $\mathcal{Q}$, $[a, b]$ und k gehörende summierte *oder* zusammengesetzte Quadraturformel.

Aus Fehlerdarstellungen und -abschätzungen für die zugrundeliegende Quadraturformel können wir Darstellungen und Abschätzungen für die summierte Formel gewinnen.

Im Fall der Mittelpunktregel etwa gilt

$$\begin{aligned}\mathcal{M}_{[a,b],k}(f) &= \sum_{i=1}^{k} \mathcal{M}_{[y_{i-1},y_i]}(f) = \sum_{i=1}^{k} (y_i - y_{i-1})\, f\left(\frac{y_{i-1}+y_i}{2}\right) \\ &= h \sum_{i=1}^{k} f\big(a + h(i-1/2)\big)\end{aligned}$$

und wir erhalten mit (7.8) die Gleichung

$$\begin{aligned}\mathcal{I}_{[a,b]}(f) - \mathcal{M}_{[a,b],k}(f) &= \sum_{i=1}^{k} \mathcal{I}_{[y_{i-1},y_i]}(f) - \mathcal{M}_{[y_{i-1},y_i]}(f) \\ &= \sum_{i=1}^{k} \frac{(y_i - y_{i-1})^3}{24} f''(\eta_i) = \frac{(b-a)^3}{24\,k^3} \sum_{i=1}^{k} f''(\eta_i)\end{aligned}$$

für geeignete $\eta_1, \ldots, \eta_k \in [a,b]$. Mit Hilfe des Zwischenwertsatzes für stetige Funktionen finden wir ein $\eta \in [a,b]$ mit

$$\frac{1}{k} \sum_{i=1}^{k} f''(\eta_i) = f''(\eta)$$

und gelangen zu der Fehlergleichung

$$\mathcal{I}_{[a,b]}(f) - \mathcal{M}_{[a,b],k}(f) = \frac{(b-a)^3}{24\,k^2} f''(\eta),$$

der Fehler wird also wie $1/k^2$ sinken, wenn wir k erhöhen. Auf diesem Weg können wir eine beliebig hohe Genauigkeit erreichen.

Für die Trapezregel erhalten wir entsprechend

$$\begin{aligned}\mathcal{T}_{[a,b],k}(f) &= \sum_{i=1}^{k} \mathcal{T}_{[y_{i-1},y_i]}(f) = \sum_{i=1}^{k} \frac{y_i - y_{i-1}}{2} \big(f(y_{i-1}) + f(y_i)\big) \\ &= \frac{h}{2} \sum_{i=1}^{k} \big(f(a+(i-1)h) + f(a+ih)\big) = \frac{h}{2} \left(f(a) + 2 \sum_{i=1}^{k-1} f(a+ih) + f(b) \right)\end{aligned}$$

mit der Fehlergleichung

$$\mathcal{I}_{[a,b]}(f) - \mathcal{T}_{[a,b],k}(f) = -\frac{(b-a)^3}{12\,k^2} f''(\eta)$$

für ein $\eta \in [a,b]$. Wie man sieht, sind der Rechenaufwand der Mittelpunkt- und der Trapezregel vergleichbar, falls k groß ist: Für die Mittelpunktregel müssen k Werte aufaddiert werden, für die Trapezregel sind es nur $k+1$, da die Werte für $y_1, \ldots, y_{k-1}$ doppelt auftreten und deshalb nur einmal berechnet werden müssen.

Für allgemeine Quadraturformeln, die für Polynome n-ten Grades exakt sind, lässt sich eine entsprechende Fehlerabschätzung gewinnen.

Satz 7.18 (Genauigkeit der summierten Quadratur) *Sei $\mathcal{Q}$ eine Quadraturformel n-ten Grades, und seien $a, b \in \mathbb{R}$ mit $a < b$ gegeben. Sei $k \in \mathbb{N}$. Dann gilt*

$$|\mathcal{I}_{[a,b]}(f) - \mathcal{Q}_{[a,b],k}(f)| \leq 4\frac{2+C_Q}{k^{n+1}} \left(\frac{b-a}{4}\right)^{n+2} \frac{\|f^{(n+1)}\|_{\infty,[a,b]}}{(n+1)!} \quad \text{für alle } f \in C^{n+1}[a,b].$$

Beweis. Sei $f \in C^{n+1}[a,b]$. Indem wir Lemma 7.16 auf die Teilintervalle $[y_{i-1}, y_i]$ anwenden, erhalten wir

$$\begin{aligned}
|\mathcal{I}_{[a,b]}(f) - \mathcal{Q}_{[a,b],k}(f)| &\leq \sum_{i=1}^{k} |\mathcal{I}_{[y_{i-1},y_i]}(f) - \mathcal{Q}_{[y_{i-1},y_i]}(f)| \\
&\leq \sum_{i=1}^{k} 4(2+C_Q) \left(\frac{y_i - y_{i-1}}{4}\right)^{n+2} \frac{\|f^{(n+1)}\|_{\infty,[y_{i-1},y_i]}}{(n+1)!} \\
&= 4(2+C_Q) \left(\frac{b-a}{4k}\right)^{n+2} \sum_{i=1}^{k} \frac{\|f^{(n+1)}\|_{\infty,[y_{i-1},y_i]}}{(n+1)!} \\
&\leq 4\frac{2+C_Q}{k^{n+1}} \left(\frac{b-a}{4}\right)^{n+2} \frac{1}{k} \sum_{i=1}^{k} \frac{\|f^{(n+1)}\|_{\infty,[a,b]}}{(n+1)!} \\
&= 4\frac{2+C_Q}{k^{n+1}} \left(\frac{b-a}{4}\right)^{n+2} \frac{\|f^{(n+1)}\|_{\infty,[a,b]}}{(n+1)!},
\end{aligned}$$

und das ist die gewünschte Abschätzung. ∎

Wir können also auch bei einer allgemeinen Quadraturformel eine beliebig hohe Genauigkeit erreichen, indem wir die Anzahl k der Teilintervalle hinreichend groß wählen. Die notwendige Anzahl der Teilintervalle ist dabei um so kleiner, je höher der Grad der Quadraturformel ist.

Bemerkung 7.19 (Niedrige Regularität) *Ein Vorteil des Beweisansatzes mit Hilfe von Exaktheit und Stabilität besteht darin, dass sich auch dann noch Aussagen gewinnen lassen, falls der Integrand nicht beliebig oft differenzierbar ist.*

Falls beispielsweise f nur einmal differenzierbar ist und wir die Mittelpunktregel anwenden, können wir aus Satz 7.18 immer noch eine Abschätzung der Form

$$|\mathcal{I}_{[a,b]}(f) - \mathcal{Q}_{[a,b],k}(f)| \leq \frac{2+C_Q}{k} \frac{(b-a)^2}{4} \|f'\|_{\infty,[a,b]} = \frac{(b-a)^2}{k} \|f'\|_{\infty,[a,b]}$$

erhalten, denn eine für Polynome ersten Grades exakte Quadraturformel ist insbesondere auch für solche nullten Grades exakt, so dass wir den Satz auf $n = 0$ anwenden und so eine brauchbare Aussage erhalten können.

Übungsaufgabe 7.20 (Summierte Simpson-Regel) *Beweisen Sie, dass ein $\eta \in [a,b]$ existiert mit*

$$\mathcal{I}_{[a,b]}(f) - \mathcal{S}_{[a,b],k}(f) = -\frac{(b-a)^5}{2880\,k^4} f^{(4)}(\eta) \qquad \text{für alle } f \in C^4[a,b],\ k \in \mathbb{N}.$$

Hinweis: *Die Übungsaufgabe 7.8 ist natürlich hilfreich.*

7.4 Gauß-Quadratur

Eine Quadraturformel ist dann besonders effizient, wenn sie mit einer möglichst geringen Anzahl von Auswertungen des Integranden, also mit einer möglichst geringen Stufe m, für Polynome möglichst hohen Grades n exakt ist, denn dann wird sich, etwa bei einer summierten Formel, besonders schnell eine hohe Genauigkeit erreichen lassen.

Nach Konstruktion ist eine m-stufige Newton-Cotes-Formel exakt für Polynome m-ten Grades, wir können also immerhin Quadraturformeln beliebig hohen Grades konstruieren. Allerdings stellt sich die Frage, ob die willkürliche Wahl äquidistanter Stützstellen besonders geschickt ist.

Beispielsweise haben wir gesehen, dass die Mittelpunktregel, eine Quadraturformel der Stufe $m = 0$, denselben Grad $n = 1$ wie die Trapezregel, eine Formel der Stufe $m = 1$, erreicht. Offenbar ist es also möglich, mit einem m-stufigen Verfahren einen höheren Grad als m zu erreichen.

Es empfiehlt sich ein genauerer Blick auf die Mittelpunktregel, um zu verstehen, wieso sie einen unerwartet hohen Grad erreicht. Entscheidend ist hier die einfache Gleichung

$$\int_{-1}^{1} x\,dx = \left[\frac{x^2}{2}\right]_{-1}^{1} = 0,$$

die $\mathcal{I}(p) = 0$ für das Polynom $p(x) = x$ impliziert. Offenbar gilt auch $\mathcal{M}(p) = 0$, so dass wir

$$\mathcal{I}(p) = 0 = \mathcal{M}(p)$$

erhalten. Damit ist $\mathcal{M}$ auch für lineare Polynome exakt.

Entscheidend bei dieser Untersuchung sind zwei Eigenschaften von p: Es ist ein Polynom, dessen Integral gleich null ist, und es ist ein Polynom, das in allen Stützstellen von $\mathcal{M}$ verschwindet. Falls wir Polynome höheren Grades finden können, die ähnliche Eigenschaften besitzen, dürfen wir darauf hoffen, entsprechend verallgemeinerte Quadraturformeln hohen Grades konstruieren zu können.

Lemma 7.21 (Legendre-Polynome) *Es gibt für jedes $m \in \mathbb{N}_0$ ein Polynom $L_m \in \Pi_m \setminus \{0\}$ mit*

$$\int_{-1}^{1} L_m(x)p(x)\,dx = 0 \qquad \text{für alle } p \in \Pi_{m-1} \text{ falls } m > 0. \tag{7.11}$$

Die Menge $\{L_0, \ldots, L_m\}$ ist eine Basis des Raums Π_m.

Beweis. Per Induktion.

Induktionsanfang: Für $m = 0$ wählen wir $L_0 = 1$ und sind fertig.

Induktionsvoraussetzung: Sei nun $n \in \mathbb{N}_0$ so gewählt, dass die Voraussetzung für alle $m \in [0 : n]$ gilt.

Induktionsschritt: Wir verwenden die Gram-Schmidt-Orthogonalisierung im Raum $L^2[-1, 1]$ zur Konstruktion von L_{n+1}, setzen also

$$L_{n+1}(x) := x^{n+1} - \sum_{m=0}^{n} \frac{\int_{-1}^{1} y^{n+1} L_m(y)\,dy}{\int_{-1}^{1} L_m^2(y)\,dy} L_m(x) \qquad \text{für alle } x \in \mathbb{R}.$$

Da $L_m \in \Pi_m$ für alle $m \in [0:n]$ gilt und offenbar das $(n+1)$-te Monom nicht in Π_n enthalten ist, muss $L_{n+1} \neq 0$ gelten.

Sei $k \in [0:n]$ gegeben. Dann gilt

$$\begin{aligned}\int_{-1}^{1} L_{n+1}(x)L_k(x)\,dx &= \int_{-1}^{1} x^{n+1}L_k(x)\,dx - \sum_{m=0}^{n} \frac{\int_{-1}^{1} y^{n+1}L_m(y)\,dy}{\int_{-1}^{1} L_m^2(y)\,dy} \int_{-1}^{1} L_m(x)L_k(x)\,dx \\ &= \int_{-1}^{1} x^{n+1}L_k(x)\,dx - \frac{\int_{-1}^{1} y^{n+1}L_k(y)\,dy}{\int_{-1}^{1} L_k^2(y)\,dy} \int_{-1}^{1} L_k^2(x)\,dx = 0.\end{aligned}$$

Sei nun $p \in \Pi_n$ gegeben. Da $\{L_0, \ldots, L_n\}$ eine Basis von Π_n ist, existieren $\alpha_0, \ldots, \alpha_n \in \mathbb{K}$ mit

$$p = \sum_{k=0}^{n} \alpha_k L_k,$$

und wir erhalten

$$\int_{-1}^{1} L_{n+1}(x)p(x)\,dx = \sum_{k=0}^{n} \alpha_k \int_{-1}^{1} L_{n+1}(x)L_k(x)\,dx = \sum_{k=0}^{n} \alpha_k 0 = 0,$$

also die gewünschte Orthogonalitätsbeziehung.

Wir haben bereits gesehen, dass L_{n+1} linear unabhängig von $L_0, \ldots, L_n$ sein muss, denn letztere Polynome liegen in Π_n, L_{n+1} hingegen nicht. Damit ist $\{L_0, \ldots, L_{n+1}\}$ eine Menge von $n+2$ linear unabhängigen Funktionen aus dem $(n+2)$-dimensionalen Raum Π_{n+1}, also muss es auch eine Basis sein. Somit ist die Induktion vollständig. ∎

Ein Polynom mit den in Lemma 7.21 beschriebenen Eigenschaften nennt man *Legendre-Polynom* (gelegentlich wird zusätzlich noch eine geeignete Skalierung gefordert). Man kann leicht nachprüfen, dass $L_1(x) = x$ gilt, also haben wir bei der Untersuchung der Mittelpunktregel bereits mit einem Legendre-Polynom gearbeitet. Ein wichtiger Punkt bei diesem Beispiel besteht darin, dass die Stützstelle der Mittelpunktregel gerade eine Nullstelle von L_1 ist. Um die Theorie verallgemeinern zu können, müssen wir auch etwas über die Nullstellen von L_m wissen.

Lemma 7.22 (Nullstellen) *Sei $m \in \mathbb{N}$. Dann besitzt ein $L_m \in \Pi_m \setminus \{0\}$ mit der Eigenschaft (7.11) gerade genau m Nullstellen, die alle in $[-1,1]$ liegen.*

Beweis. Sei $L_m \in \Pi_m \setminus \{0\}$ ein Polynom, das (7.11) erfüllt, und seien $\xi_1, \ldots, \xi_k$ mit $k \in [0:m]$ seine Nullstellen in $[-1,1]$ und $\mu_1, \ldots, \mu_k$ deren Vielfachheiten.

Unser Ziel ist es, ein Polynom $p \neq 0$ zu konstruieren, das dafür sorgt, dass $L_m p$ nur Nullstellen geradzahliger Vielfachheit besitzt. Daraus folgt, dass $L_m p$ in $[-1,1]$ nie das Vorzeichen wechseln kann, und damit sein Integral von null verschieden sein muss.

Dazu setzen wir

$$p(x) := \prod_{\substack{i=1 \\ \mu_i \text{ ungerade}}}^{k} x - \xi_i \qquad \text{für alle } x \in \mathbb{K}.$$

Offenbar gilt $p \in \Pi_k$, und nach Konstruktion besitzt das Produkt $L_m p$ nur Nullstellen geradzahliger Vielfachheit. Da weder p noch L_m das Nullpolynom ist, ist $L_m p \neq 0$, und da an keiner seiner Nullstellen ein Vorzeichenwechsel passieren kann, muss

$$\int_{-1}^{1} L_m(x)p(x)\,dx \neq 0$$

gelten. Aus (7.11) folgt nun, dass $p \notin \Pi_{m-1}$ gelten muss, also insbesondere $k \geq m$. Da L_m höchstens m Nullstellen besitzen kann, erhalten wir $k = m$, also liegen alle m Nullstellen von L_m in $[-1, 1]$ und sind einfach. ∎

Nun können wir unseren Plan in die Tat umsetzen und Quadraturformeln besonders hohen Grades definieren.

Definition 7.23 (Gauß-Quadratur) *Sei $m \in \mathbb{N}_0$. Seien $x_0, \ldots, x_m \in [-1, 1]$ die nach Lemma 7.22 existierenden $m + 1$ einfachen Nullstellen eines Legendre-Polynoms L_{m+1}, und seien $w_0, \ldots, w_m$ die gemäß (7.3) definierten Gewichte. Dann nennen wir*

$$\mathcal{G}_m \colon C[-1, 1] \to \mathbb{R}, \qquad f \mapsto \sum_{i=0}^{m} w_i f(x_i),$$

die m-te Gauß-Quadraturformel.

Natürlich sind wir daran interessiert, nachzuweisen, dass sich unsere Mühen gelohnt haben und die Gauß-Quadraturformeln für Polynome eines hohen Grades exakt sind.

Lemma 7.24 (Exaktheitsgrad) *Sei $m \in \mathbb{N}_0$. Dann ist $\mathcal{G}_m$ exakt für Polynome des Grades $2m + 1$.*

Beweis. Sei $p \in \Pi_{2m+1}$ ein beliebiges Polynom. Per Polynomdivision können wir $q, r \in \Pi_m$ so finden, dass

$$p = L_{m+1}q + r$$

gilt. Dann haben wir

$$\begin{aligned}
\mathcal{I}(p) &= \int_{-1}^{1} L_{m+1}(x)q(x)\,dx + \int_{-1}^{1} r(x)\,dx = 0 + \int_{-1}^{1} r(x)\,dx = \mathcal{I}(r),\\
\mathcal{G}_m(p) &= \sum_{i=0}^{m} w_i L_{m+1}(x_i)q(x_i) + \sum_{i=0}^{m} w_i r(x_i) = \sum_{i=0}^{m} w_i\, 0\, q(x_i) + \sum_{i=0}^{m} w_i r(x_i) = \mathcal{G}_m(r).
\end{aligned}$$

Für die erste Gleichung nutzen wir (7.11) aus, für die zweite, dass die x_i gerade Nullstellen von L_{m+1} sind.

Wir haben bereits gesehen, dass die nach (7.3) bestimmten Gewichte dazu führen, dass Polynome m-ten Grades exakt integriert werden, also gilt $\mathcal{I}(r) = \mathcal{G}_m(r)$ und es folgt

$$\mathcal{I}(p) = \mathcal{I}(r) = \mathcal{G}_m(r) = \mathcal{G}_m(p).$$

Damit ist der Beweis auch schon vollständig. ∎

Es stellt sich die Frage, ob wir eine Quadraturformel finden können, die für Polynome eines noch höheren Grades exakt ist. Das ist nicht der Fall:

Lemma 7.25 (Maximale Exaktheit) *Es gibt keine m-stufige Quadraturformel, die exakt für Polynome des Grades $2m+2$ ist.*

Beweis. Sei $\mathcal{Q}$ eine m-stufige Quadraturformel mit den Quadraturpunkten $x_0, \ldots, x_m \in [-1,1]$. Wir untersuchen das Polynom

$$p(x) := \prod_{i=0}^{m} (x - x_i)^2 \qquad \text{für alle } x \in \mathbb{R}.$$

Es liegt offenbar in Π_{2m+2} und erfüllt $p(x_i) = 0$ für alle $i \in [0:m]$. Damit muss $\mathcal{Q}(p) = 0$ gelten. Da p aber nicht-negativ und nicht null ist, gilt auch $\mathcal{I}(p) > 0 = \mathcal{Q}(p)$, also kann $\mathcal{Q}$ nicht exakt für Polynome $(2m+2)$-ten Grades sein. ∎

Insofern erreichen Gauß-Formeln den höchsten bei $m+1$ Stützstellen möglichen Grad, und damit auch die optimale Effizienz.

Mit einem ähnlichen Trick lässt sich nachweisen, dass die Gewichte der Gauß-Quadraturformeln strikt positiv sind und dass die Stabilitätskonstante den optimalen Wert annimmt:

Lemma 7.26 (Quadraturgewichte) *Sei $m \in \mathbb{N}_0$, und seien $w_0, \ldots, w_m$ die Gewichte einer Quadraturformel $\mathcal{Q}$, die exakt für Polynome des Grades $2m$ ist. Dann gilt*

$$w_i > 0 \qquad \text{für alle } i \in [0:m]$$

und die Stabilitätskonstante erfüllt $C_Q = 2$. Das ist die kleinste Stabilitätskonstante, die für eine Quadraturformel mindestens nullten Grades möglich ist.

Beweis. Sei $i \in [0:m]$. Wir definieren ein $p \in \Pi_{2m}$ durch

$$p(x) = \prod_{\substack{j=0 \\ j \neq i}}^{m} (x - x_j)^2 \qquad \text{für alle } x \in \mathbb{R}.$$

Offenbar ist p nicht null und nicht-negativ und besitzt höchstens den Grad $2m$. Da $\mathcal{Q}$ für Polynome dieses Grades exakt ist, folgt

$$0 < \mathcal{I}(p) = \mathcal{Q}(p) = \sum_{k=0}^{m} w_k p(x_k) = w_i p(x_i),$$

und aus $p(x_i) > 0$ können wir auf $w_i > 0$ schließen.

Für die Funktion $f := 1$ erhalten wir

$$2 = \mathcal{I}(f) = \mathcal{Q}(f) = \sum_{i=0}^{m} w_i f(x_i) = \sum_{i=0}^{m} w_i = \sum_{i=0}^{m} |w_i| = C_Q.$$

Für eine beliebige Quadraturformel $\mathcal{Q}_0$ mindestens nullten Grades muss

$$|\mathcal{Q}_0(f)| = |\mathcal{I}(f)| = 2 = 2\|f\|_{\infty,[-1,1]}$$

gelten, also kann wegen Lemma 7.10 die Stabilitätskonstante C_{Q_0} nicht kleiner als 2 sein. In dieser Hinsicht weist $\mathcal{Q}$ die optimale Stabilitätskonstante auf. ∎

Folgerung 7.27 (Gewichte) *Sei $m \in \mathbb{N}_0$. Alle Gewichte der m-ten Gauß-Quadraturformel $\mathcal{G}_m$ sind positiv und die Stabilitätskonstante ist $C_{G_m} = 2$.*

Beweis. Wir kombinieren Lemma 7.24 mit Lemma 7.26. ∎

Beispiel 7.28 (Gauß-Formel dritten Grades) *Wir untersuchen die Konstruktion einer Gauß-Quadraturformel der Stufe $m = 1$ und dritten Grades. Dazu brauchen wir ein $L_2 \in \Pi_2 \setminus \{0\}$ mit*

$$0 = \int_{-1}^{1} L_2(x)\,dx, \qquad 0 = \int_{-1}^{1} L_2(x)\,x\,dx.$$

Mit dem Ansatz $L_2(x) = \alpha + \beta x + \gamma x^2$ erhalten wir die Gleichungen

$$0 = \int_{-1}^{1} \alpha + \beta x + \gamma x^2\,dx = \left[\alpha x + \beta\frac{x^2}{2} + \gamma\frac{x^3}{3}\right]_{x=-1}^{1} = 2\alpha + \frac{2}{3}\gamma,$$
$$0 = \int_{-1}^{1} \alpha x + \beta x^2 + \gamma x^3\,dx = \left[\alpha\frac{x^2}{2} + \beta\frac{x^3}{3} + \gamma\frac{x^4}{4}\right]_{x=-1}^{1} = \frac{2}{3}\beta,$$

aus denen wir unmittelbar $\beta = 0$ und $\alpha = -\gamma/3$ folgern können. Also ist $L_2(x) = x^2 - 1/3$ ein Legendre-Polynom, und seine Nullstellen sind durch $x_0 = -1/\sqrt{3}, x_1 = 1/\sqrt{3}$ gegeben.

Aus Symmetriegründen muss $w_0 = w_1$ gelten, und mit Folgerung 7.27 folgt sofort $w_0 = w_1 = 1$. Damit ist die einstufige Gauß-Quadraturformel durch

$$\mathcal{G}_1(f) := f\left(-\frac{1}{\sqrt{3}}\right) + f\left(\frac{1}{\sqrt{3}}\right)$$

gegeben, und es lässt sich leicht nachprüfen, dass sie quadratische und kubische Polynome exakt integriert, während das beispielsweise der ähnlich aussehenden Trapezregel bereits bei quadratischen Polynomen nicht mehr gelingt.

Bemerkung 7.29 (Eindeutigkeit) *Sei $\mathcal{Q}$ eine beliebige m-stufige Quadraturformel mit den Quadraturpunkten $x_0, \ldots, x_m$. Falls $\mathcal{Q}$ exakt für Polynome der Ordnung $2m+1$ ist, können wir durch Wahl geeigneter Polynome $q \in \Pi_m$ und Anwendung der Exaktheitsbedingung auf $qL_{m+1} \in \Pi_{2m+1}$ folgern, dass $L_{m+1}(x_i) = 0$ für alle $i \in [0 : m]$ gelten muss. Damit können sich die Quadraturpunkte nur durch ihre Reihenfolge von denen der Gauß-Quadraturformel $\mathcal{G}_m$ unterscheiden. In diesem Sinn ist $\mathcal{G}_m$ die einzige Quadraturformel, die die optimale Ordnung $2m+1$ erreicht.*

Bemerkung 7.30 (Verallgemeinerte Integrale) *Ein genauerer Blick auf die Herleitung der Gauß-Quadraturformel enthüllt, dass wir nur wenige Eigenschaften des Riemann-Integrals verwendet haben: Es ist wichtig, dass das Integral des Quadrats eines von null verschiedenen Polynoms existiert und positiv ist, und es ist wichtig, dass das Integral linear ist.*

Tatsächlich lassen sich verallgemeinerte Gauß-Quadraturformeln auch für Integrale mit allgemeineren Maßen oder, unter bestimmten Bedingungen, sogar für uneigentliche Integrale konstruieren.

8 Gewöhnliche Differentialgleichungen

Viele naturwissenschaftliche Phänomene werden mit Hilfe von Differentialgleichungen beschrieben. Beispielsweise besagen die Axiome der klassischen Mechanik, dass die Geschwindigkeit die Ableitung des Orts nach der Zeit ist, und die Beschleunigung wiederum die Ableitung der Geschwindigkeit. Eine auf ein Objekt wirkende Kraft verursacht eine Beschleunigung, die die Geschwindigkeit beeinflusst, und schließlich auch den Ort des Objekts.

Solche Zusammenhänge führen häufig zu *Anfangswertproblemen*: Wir beschreiben das zu untersuchende System durch m zeitabhängige Parameter, also durch eine Funktion $y : [a, b] \to V := \mathbb{K}^m$, die jedem Zeitpunkt zwischen a und b einen Satz von Parametern zuordnet, der das System zu diesem Zeitpunkt beschreibt.

Bei einem *Anfangswert*problem gehen wir davon aus, dass der Zustand zum Anfangszeitpunkt a bekannt ist, dass also $y(a) = y_0$ für ein $y_0 \in V$ gilt. Die Veränderung des Systems wird über eine Funktion $f : [a, b] \times V \to V$ beschrieben, die jedem Zeitpunkt und jedem Zustand des Systems die Ableitung der Funktion y zu diesem Zeitpunkt angibt, also die Veränderung, die das System zu diesem Zeitpunkt in diesem Zustand erfährt.

Unsere Aufgabe besteht darin, aus dem Anfangszustand und der Veränderungsfunktion Informationen über das Verhalten des Systems während des gesamten Zeitraums $[a, b]$ zu gewinnen. Mathematisch formuliert nimmt diese Aufgabe die folgende Form an:

> Gegeben seien $a, b \in \mathbb{R}$ mit $a < b$, ein Anfangswert $y_0 \in V$ und eine Funktion $f \in C([a, b] \times V, V)$, berechne die Lösung $y \in C^1([a, b], V)$ der Gleichungen
>
> $$y(a) = y_0, \qquad y'(t) = f(t, y(t)) \qquad \text{für alle } t \in [a, b]. \tag{8.1}$$

8.1 Theoretische Grundlagen

Eine elegante Möglichkeit zur Beschreibung der Lösung eines Anfangswertproblems besteht darin, die Differentialgleichung auf eine Integralgleichung zurückzuführen:

Lemma 8.1 (Integralformulierung) *Sei $y_0 \in V$, und sei $f \in C([a, b] \times V, V)$. Eine Funktion $y \in C([a, b], V)$, die*

$$y(t) = y_0 + \int_a^t f(s, y(s))\, ds \qquad \textit{für alle } t \in [a, b] \tag{8.2}$$

erfüllt, ist einmal stetig differenzierbar und Lösung des Anfangswertproblems (8.1).

Umgekehrt ist jede Lösung des Anfangswertproblems auch eine Lösung der Integralgleichung (8.2).

Beweis. Sei zunächst $y \in C([a,b],V)$ eine Lösung der Integralgleichung (8.2). Sei $t \in [a,b)$ und $h \in (0, b-t]$. Dann gilt

$$\begin{aligned}\frac{y(t+h)-y(t)}{h} - f(t,y(t)) &= \frac{1}{h}\left(y_0 + \int_a^{t+h} f(s,y(s))\,ds - y_0 - \int_a^t f(s,y(s))\,ds \right.\\ &\qquad \left. - \int_t^{t+h} f(t,y(t))\,ds\right)\\ &= \frac{1}{h}\int_t^{t+h} f(s,y(s)) - f(t,y(t))\,ds\end{aligned}$$

und mit der Dreiecksungleichung für Integrale folgt

$$\left\|\frac{y(t+h)-y(t)}{h} - f(t,y(t))\right\| \leq \frac{1}{h}\int_t^{t+h} \|f(s,y(s)) - f(t,y(t))\|\,ds.$$

Da f und y stetig sind, ist auch $g(s) := f(s,y(s)) - f(t,y(t))$ eine auf $[t,t+h]$ stetige Funktion, so dass sich

$$\left\|\frac{y(t+h)-y(t)}{h} - f(t,y(t))\right\| \leq \frac{1}{h}\int_t^{t+h} \|g\|_{\infty,[t,t+h]}\,ds = \|g\|_{\infty,[t,t+h]}$$

ergibt. Wegen $g(t) = 0$ konvergiert die rechte Seite für $h \to 0$ gegen null, so dass wir $y'(t) = f(t,y(t))$ erhalten. Für den rechten Randpunkt $t = b$ können wir mit dem linksseitigen Differenzenquotienten entsprechend verfahren. Damit ist y stetig differenzierbar und auch Lösung des Anfangswertproblems (8.1).

Sei nun $y \in C^1([a,b],V)$ eine Lösung des Anfangswertproblems. Nach Hauptsatz der Differential- und Integralrechnung gilt

$$y(t) = y(a) + \int_a^t y'(s)\,ds = y(a) + \int_a^t f(s,y(s))\,ds \qquad \text{für alle } t \in [a,b],$$

also ist y auch eine Lösung der Integralgleichung (8.2). ∎

Erinnerung 8.2 (Picard-Iteration) *Die Gleichung (8.2) können wir als Fixpunktgleichung*

$$y = \Phi[y]$$

auf der Menge $C([a,b],V)$ lesen, indem wir eine Iterationsfunktion

$$\Phi : C([a,b],V) \to C([a,b],V), \qquad y \mapsto \left(t \mapsto y_0 + \int_a^t f(s,y(s))\,ds\right), \tag{8.3}$$

definieren, die stetige Funktionen auf stetige Funktionen abbildet.

In Anlehnung an den Fixpunktsatz 5.6 von Banach bietet es sich dann an, die Lösung y durch eine Iteration

$$y^{(m+1)} := \Phi[y^{(m)}] \qquad \text{für alle } m \in \mathbb{N}_0$$

zu approximieren. Falls f Lipschitz-stetig im zweiten Argument ist, lässt sich beweisen, dass diese Iteration für alle Ausgangsnäherungen $y^{(0)} \in C([a,b],V)$ gegen den eindeutig bestimmten Fixpunkt konvergiert, der nach Lemma 8.1 stetig differenzierbar ist und das Anfangswertproblem (8.1) löst.

Um uns Schwierigkeiten bei Existenz und Eindeutigkeit der Lösungen im Folgenden zu ersparen setzen wir von nun an immer voraus, dass f im zweiten Argument Lipschitz-stetig ist, dass also eine Konstante $L_f \in \mathbb{R}_{\geq 0}$ existiert, die

$$\|f(t,x) - f(t,z)\| \leq L_f \|x - z\| \qquad \text{für alle } t \in [a,b] \text{ und } x, z \in V \tag{8.4}$$

erfüllt. Falls $L_f(b-a) < 1$ gilt, folgt daraus bereits, dass die in (8.3) definierte Iterationsfunktion eine Kontraktion bezüglich der Maximum-Norm ist, so dass eine Lösung der Integralgleichung (8.2) existiert, also auch eine des Anfangswertproblems (8.1).

Eine Existenz- und Eindeutigkeitsaussage für beliebige Intervalle können wir erhalten, indem wir die Maximum-Norm durch eine *gewichtete Maximum-Norm*[1] ersetzen, die wie folgt definiert ist:

$$\|u\|_{e,[a,b]} := \max\{\|u(t)\| e^{-2L_f(t-a)} \ : \ t \in [a,b]\} \qquad \text{für alle } u \in C([a,b],V).$$

Für Φ erhalten wir dann

$$\begin{aligned}
\|\Phi(x) - \Phi(z)\| e^{-2L_f(t-a)} &\leq e^{-2L_f(t-a)} \int_a^t \|f(s,x(s)) - f(s,z(s))\| \, ds \\
&\leq e^{-2L_f(t-a)} \int_a^t L_f \|x(s) - z(s)\| \, ds \\
&\leq e^{-2L_f(t-a)} \int_a^t L_f e^{2L_f(s-a)} \|x - z\|_{e,[a,b]} \, ds \\
&= e^{-2L_f(t-a)} \frac{1}{2} e^{2L_f(t-a)} \|x - z\|_{e,[a,b]} \\
&= \frac{1}{2} \|x - z\|_{e,[a,b]} \qquad \text{für alle } x, z \in C([a,b],V),
\end{aligned}$$

so dass $\|\Phi(x) - \Phi(z)\|_{e,[a,b]} \leq \|x - z\|_{e,[a,b]}/2$ folgt und Φ deshalb unabhängig von der Länge des Intervalls immer eine Kontraktion ist.

Übungsaufgabe 8.3 (Picard-Iteration) *Gegeben sei das Anfangswertproblem*

$$y(0) = 1, \qquad y'(t) = y(t) \qquad \text{für alle } t \in \mathbb{R}.$$

Führen Sie die Picard-Iteration für diese Gleichung durch, um ausgehend von der Anfangsnäherung $y^{(0)} := 1$ eine Folge $(y^{(m)})_{m=0}^{\infty}$ zu konstruieren, die gegen die Lösung y konvergiert. Kommt Ihnen die Folge bekannt vor?

[1] Die Idee, eine gewichtete Norm zu verwenden, findet sich in Abschnitt 1.6 des Buchs „Gewöhnliche Differentialgleichungen“ von Wolfgang Walter, Springer 1993.

In Hinblick auf numerische Verfahren ist auch die Frage nach der Stabilität der Lösung von Interesse, also die Frage danach, wie sich die Lösung ändert, wenn wir den Anfangswert y_0 oder die rechte Seite f stören.

Erinnerung 8.4 (Grønwall-Ungleichung) *Für $\alpha \in \mathbb{R}_{\geq 0}$, eine stetige Funktion $\beta \in C([a,b], \mathbb{R}_{\geq 0})$ und eine stetige Funktion $u \in C([a,b], \mathbb{R}_{\geq 0})$, die die Ungleichung*

$$u(t) \leq \alpha + \int_a^t \beta(s) u(s)\, ds \qquad \text{für alle } t \in [a,b] \tag{8.5a}$$

erfüllt, gilt

$$u(t) \leq \alpha \exp\left(\int_a^t \beta(s)\, ds \right) \qquad \text{für alle } t \in [a,b]. \tag{8.5b}$$

Mit Hilfe dieser Ungleichung können wir untersuchen, wie die Lösung eines Anfangswertproblems auf Störungen der Parameter reagiert: Seien $f, g \in C([a,b] \times V, V)$ sowie $y_0, z_0 \in V$ gegeben, und sei f Lipschitz-stetig mit (8.4). Seien $y, z \in C^1([a,b], V)$ Lösungen der Anfangswertprobleme

$$\begin{aligned} y(a) &= y_0, & y'(t) &= f(t, y(t)), \\ z(a) &= z_0, & z'(t) &= g(t, z(t)) \end{aligned} \qquad \text{für alle } t \in [a,b].$$

Mit Lemma 8.1 erhalten wir

$$\|y(t) - z(t)\| \leq \|y_0 - z_0\| + \int_a^b \|f(s, z(s)) - g(s, z(s))\|\, ds + L_f \int_a^t \|y(s) - z(s)\|\, ds \qquad \text{für alle } t \in [a,b].$$

Damit erfüllt die Differenz $u(t) := \|y(t) - z(t)\|$ die Ungleichung (8.5a) mit

$$\alpha := \|y_0 - z_0\| + \int_a^b \|f(s, z(s)) - g(s, z(s))\|\, ds, \qquad \beta := L_f,$$

so dass aus (8.5b) bereits

$$\|y(t) - z(t)\| \leq \left(\|y_0 - z_0\| + \int_a^b \|f(s, z(s)) - g(s, z(s))\|\, ds \right) e^{L_f (t-a)} \qquad \text{für alle } t \in [a,b]$$

folgt. Insbesondere hängt die Lösung Lipschitz-stetig von dem Anfangswert y_0 und der Funktion f ab.

8.2 Einfache Lösungsverfahren

Wir widmen uns der Frage, wie sich Lösungen des Anfangswertproblems (8.1) durch allgemeine numerische Verfahren approximieren lassen. Mit Hilfe der Integralformulierung (8.2) können wir das Lösen des Anfangswertproblems auf ein Integrationsproblem zurückführen: Falls uns eine Quadraturformel für das Intervall $[a, b]$ zur Verfügung steht, erhalten wir

$$y(b) = y_0 + \int_a^b f(s, y(s))\,ds \approx y_0 + \sum_{i=0}^{m} w_i f(s_i, y(s_i)).$$

Leider können wir die Summe nicht auswerten, denn dazu müssten wir die unbekannte Funktion y in den Quadraturpunkten s_i kennen.

Durch eine geschickte Wahl der Quadraturformel lässt sich dieses Problem allerdings umgehen: Wir verwenden beispielsweise nur einen Quadraturpunkt $s_0 = a$, denn $y(a) = y_0$ ist uns als Anfangswert bekannt. Es ist naheliegend, das Gewicht so zu wählen, dass wenigstens eine Quadraturformel nullten Grades entsteht, also als $w_0 = b - a$. Damit erhalten wir die Näherung

$$y(b) \approx \tilde{y}(b) := y(a) + (b - a) f(a, y(a)). \tag{8.6}$$

Natürlich interessieren wir uns wieder für die Größe des bei diesem Ansatz entstehenden Fehlers.

Lemma 8.5 (Fehlerabschätzung) *Sei $y \in C^2([a, b], V)$ Lösung der Gleichung (8.1), und sei*

$$\tilde{y}(b) := y(a) + (b - a) f(a, y(a))$$

die durch das explizite Euler-Verfahren berechnete Näherung. Dann gilt

$$\|y(b) - \tilde{y}(b)\|_V \leq \frac{(b - a)^2}{2} \|y''\|_{\infty,[a,b]}.$$

Beweis. Wegen (8.2) gilt

$$\begin{aligned} y(b) - \tilde{y}(b) &= y(a) + \int_a^b f(s, y(s))\,ds - y(a) - \int_a^b f(a, y(a))\,ds \\ &= \int_a^b f(s, y(s)) - f(a, y(a))\,ds = \int_a^b y'(s) - y'(a)\,ds. \end{aligned}$$

Mit Hilfe des Hauptsatzes der Differential- und Integralrechnung erhalten wir

$$y'(s) - y'(a) = \int_a^s y''(r)\,dr,$$

also per Dreiecksungleichung für Integrale

$$\|y'(s) - y'(a)\|_V = \left\| \int_a^s y''(r)\,ds \right\| \leq \int_a^s \|y''(r)\|_V\,dr \leq (s - a)\|y''\|_{\infty,[a,b]}.$$

Indem wir in die ursprüngliche Gleichung einsetzen und erneut die Dreiecksungleichung ausnutzen, erhalten wir

$$\begin{aligned}\|y(b) - \tilde{y}(b)\|_V &\leq \int_a^b \|y'(s) - y'(a)\|_V \, ds \\ &\leq \int_a^b (s-a)\|y''\|_{\infty,[a,b]} \, ds = \frac{(b-a)^2}{2}\|y''\|_{\infty,[a,b]}.\end{aligned}$$

Das ist die gewünschte Abschätzung. ∎

Es lässt sich zeigen, dass es Funktionen y gibt, für die diese Abschätzung sich nicht verbessern lässt. Auf einem festen Intervall werden wir also wohl nicht eine beliebig genaue Approximation erreichen können. Wie bereits im Fall der Quadraturverfahren lässt sich das Problem umgehen, indem wir das Intervall in Teilintervalle zerlegen und die Näherung auf diesen Teilintervallen anwenden. Dazu fixieren wir $k \in \mathbb{N}$ und Punkte

$$a = t_0 < t_1 < \ldots < t_k = b.$$

Aus der Integralformulierung (8.2) folgt

$$\begin{aligned}y(t_i) &= y_0 + \int_{t_0}^{t_i} f(s, y(s)) \, ds = y_0 + \int_{t_0}^{t_{i-1}} f(s, y(s)) \, ds + \int_{t_{i-1}}^{t_i} f(s, y(s)) \, ds \\ &= y(t_{i-1}) + \int_{t_{i-1}}^{t_i} f(s, y(s)) \, ds \qquad \text{für alle } i \in [1:k],\end{aligned}$$

so dass wir nun die Quadraturformel lediglich auf den Teilintervallen $[t_{i-1}, t_i]$ anwenden müssen, deren Länge wir steuern können. Damit können wir, wie im vorangehenden Kapitel gezeigt, auch die Genauigkeit der Approximation steuern.

Wenn wir die Schrittweiten mit

$$h_i := t_i - t_{i-1} \qquad \text{für alle } i \in [1:k]$$

bezeichnen, nimmt die Formel (8.6) für den ersten Schritt die Form

$$y(t_1) \approx \tilde{y}(t_1) := y(t_0) + h_1 f(t_0, y(t_0))$$

an. Wenn wir nun eine Näherung von $y(t_2)$ in derselben Weise berechnen wollen, müssten wir eigenlich von dem exakten Startwert $y(t_1)$ ausgehen, der uns aber nicht zur Verfügung steht. Also ersetzen wir ihn durch die Näherung $\tilde{y}(t_1)$ aus dem vorangehenden Schritt und erhalten so

$$y(t_2) \approx \tilde{y}(t_2) := \tilde{y}(t_1) + h_2 f(t_1, \tilde{y}(t_1)).$$

Entsprechend können wir fortfahren und erhalten einen ersten Algorithmus zur Approximation der Lösung einer gewöhnlichen Differentialgleichung:

Definition 8.6 (Explizites Euler-Verfahren) *Das durch*

$$\begin{aligned}\tilde{y}(t_0) &:= y_0,\\ \tilde{y}(t_i) &:= \tilde{y}(t_{i-1}) + h_i f(t_{i-1}, \tilde{y}(t_{i-1})) \qquad \text{für alle } i \in [1:k]\end{aligned}$$

definierte Verfahren zur näherungsweisen Berechnung der Lösung y der gewöhnlichen Differentialgleichung (8.1) nennt man das explizite Euler-Verfahren.

Das explizite Euler-Verfahren hat den Vorteil, dass es sehr einfach zu implementieren ist: Wir können uns direkt „von links nach rechts“ durch das Intervall $[a, b]$ hindurcharbeiten und der Reihe nach Näherungslösungen mit je einer Auswertung der rechten Seite f und einer Linearkombination bestimmen.

Bemerkung 8.7 (Implizites Euler-Verfahren) *Statt das Integral durch den Wert der Funktion im linken Randpunkt zu approximieren, können wir auch den rechten Randpunkt verwenden. Dann haben wir*

$$y(b) = y(a) + \int_a^b f(s, y(s))\, ds \approx y(a) + (b-a) f(b, y(b)),$$

also können wir unsere Näherung $\tilde{y}(b)$ als Lösung der Fixpunktgleichung

$$x = \Phi(x), \qquad \Phi\colon V \to V,\ x \mapsto y(a) + (b-a) f(b, x),$$

beschreiben. Falls f Lipschitz-stetig im zweiten Argument ist, erhalten wir

$$\begin{aligned}\|\Phi(x_1) - \Phi(x_2)\|_V &= \|y(a) + (b-a) f(b, x_1) - y(a) - (b-a) f(b, x_2)\|_V\\ &= |b-a|\, \|f(b, x_1) - f(b, x_2)\|_V\\ &\leq L_f |b-a|\, \|x_1 - x_2\|_V \qquad \text{für alle } x_1, x_2 \in V,\end{aligned}$$

so dass für hinreichend kleine Intervalle $L_f|b-a| < 1$ gilt und wir eine Kontraktion erhalten. Damit wäre das implizite Euler-Verfahren immerhin durchführbar, und wir können ähnlich wie in Lemma 8.5 eine Aussage über den Approximationsfehler beweisen.

Bei beiden Varianten des Euler-Verfahrens ist wegen der relativ langsamen Konvergenz häufig eine sehr kleine Schrittweite notwendig, wir werden also die Anzahl k der Teilintervalle sehr groß wählen müssen, und damit wird auch der Rechenaufwand erheblich sein. Ähnlich wie bei Quadraturverfahren bietet es sich an, nach Verfahren höheren Grades zu suchen.

Eine Möglichkeit besteht darin, von einem Quadraturverfahren höheren Grades auszugehen. Ein einfaches Beispiel ist die Mittelpunktregel, also

$$y(b) = y(a) + \int_a^b f(s, y(s))\, ds \approx y(a) + (b-a) f\left(\frac{b+a}{2}, y\left(\frac{b+a}{2}\right)\right).$$

Da uns $y((b+a)/2)$ nicht zur Verfügung steht, nähern wir es mit Hilfe des bereits bekannten expliziten Euler-Verfahrens an, verwenden also

$$y\left(\frac{b+a}{2}\right) \approx y(a) + \left(\frac{b+a}{2} - a\right) f(a, y(a)) = y(a) + \frac{b-a}{2} f(a, y(a)),$$

um die Näherung

$$y(b) \approx y(a) + (b-a) f\left(\frac{b+a}{2}, y(a) + \frac{b-a}{2} f(a, y(a))\right)$$

zu erhalten. Wie schon im Fall des Euler-Verfahrens wenden wir diesen Ansatz nicht für das gesamte Intervall $[a, b]$ an, sondern arbeiten wieder auf den Teilintervallen $[t_{i-1}, t_i]$:

Definition 8.8 (Runge-Verfahren) *Das durch*

$$\begin{aligned} \tilde{y}(t_0) &= y_0, \\ \tilde{y}(t_i) &= \tilde{y}(t_{i-1}) + h_i f\left(t_{i-1} + \frac{h_i}{2}, \tilde{y}(t_{i-1}) + \frac{h_i}{2} f(t_{i-1}, \tilde{y}(t_{i-1}))\right) \\ &\qquad \text{für alle } i \in [1:k] \end{aligned}$$

definierte Verfahren zur näherungsweisen Berechnung der Lösung y der gewöhnlichen Differentialgleichung (8.1) nennt man das Runge-Verfahren.

Gegenüber dem Euler-Verfahren hat dieser Ansatz den Nachteil, dass pro Schritt zwei Auswertungen der Funktion f erforderlich sind. Die Fehleranalyse zeigt allerdings, dass sich dadurch auch die Konvergenz deutlich verbessert:

Lemma 8.9 (Fehlerabschätzung) *Sei $y \in C^3([a,b], V)$ Lösung der Gleichung (8.1), und sei f im zweiten Argument Lipschitz-stetig, es gelte also (8.4). Die Näherung des Runge-Verfahrens bezeichnen wir mit*

$$\tilde{y}(b) := y(a) + (b-a) f\left(\frac{b+a}{2}, y(a) + \frac{b-a}{2} f(a, y(a))\right).$$

Dann gilt

$$\|y(b) - \tilde{y}(b)\|_V \leq \frac{(b-a)^3}{8} \left(\frac{1}{3}\|y'''\|_{\infty,[a,b]} + L_f \|y''\|_{\infty,[a,b]}\right).$$

Beweis. Sei $c = (b+a)/2 = a + (b-a)/2$. Wir unterscheiden zwischen dem Quadraturfehler

$$e_Q := y(b) - (y(a) + (b-a) f(c, y(c)))$$

und dem Fehler

$$e_A := (b-a)\left(f(c, y(c)) - f\left(c, y(a) + \frac{b-a}{2} f(a, y(a))\right)\right),$$

der durch die Approximation von $f(c, y(c))$ durch das Euler-Verfahren entsteht. Es gilt

$$y(b) - \tilde{y}(b) = e_Q + e_A, \qquad \|y(b) - \tilde{y}(b)\|_V \leq \|e_Q\|_V + \|e_A\|_V,$$

also können wir beide Fehler einzeln abschätzen.

Für den Quadraturfehler folgt aus (7.9) direkt

$$\|e_Q\|_V \leq \frac{(b-a)^3}{24} \|y'''\|_{\infty,[a,b]}. \tag{8.7a}$$

Den Fehler e_A schätzen wir mit Hilfe der Lipschitz-Stetigkeit ab:

$$\|e_A\|_V \leq (b-a) L_f \left\| y(c) - \left(y(a) + \frac{b-a}{2} f(a, y(a)) \right) \right\|_V.$$

Auf den hinteren Term können wir Lemma 8.5 mit dem Intervall $[a, c]$ statt $[a, b]$ anwenden, um

$$\|e_A\|_V \leq (b-a) L_f \frac{(b-a)^2}{8} \|y''\|_{\infty,[a,b]} = L_f \frac{(b-a)^3}{8} \|y''\|_{\infty,[a,b]} \tag{8.7b}$$

zu erhalten. Durch Addition der Abschätzungen (8.7a) und (8.7b) erhalten wir das gewünschte Ergebnis. ∎

Ein etwas komplizierteres Beispiel verwendet die Trapezregel für die Näherung des Integrals, also

$$y(b) = y(a) + \int_a^b f(s, y(s))\, ds \approx y(a) + \frac{b-a}{2} (f(b, y(b)) + f(a, y(a))).$$

Zur Definition einer Näherung $\widehat{y}(b)$ verwenden wir somit die Gleichung

$$\widehat{y}(b) = y(a) + \frac{b-a}{2} (f(b, \widehat{y}(b)) + f(a, y(a))),$$

die es uns leider nicht ermöglicht, $\widehat{y}(b)$ direkt zu bestimmen. Mit der Funktion

$$\Phi \colon V \to V, \qquad z \mapsto y(a) + \frac{b-a}{2} (f(b, z) + f(a, y(a))),$$

können wir die Gleichung allerdings als Fixpunktgleichung

$$\widehat{y}(b) = \Phi(\widehat{y}(b))$$

schreiben. Entsprechend Satz 5.6 besitzt diese Gleichung eine Lösung, falls Φ eine Kontraktion ist. Indem wir wieder voraussetzen, dass f die Lipschitz-Bedingung (8.4) erfüllt, erhalten wir

$$\|\Phi(x) - \Phi(z)\|_V = \frac{b-a}{2} \|f(b, x) - f(b, z)\|_V \leq \frac{b-a}{2} L_f \|x - z\|_V \quad \text{für alle } x, z \in V,$$

die Abbildung Φ wird also eine Kontraktion sein, falls die Schrittweite klein genug ist, um

$$\frac{b-a}{2}L < 1$$

sicherzustellen. Damit können wir $\widehat{y}(b)$ durch eine Näherung $\tilde{y}(b)$ ersetzen, die wir mit einigen Schritten der Fixpunktiteration berechnen. Sofern wir von einem guten Startwert ausgehen und die Schrittweite klein genug ist, könnte schon ein einziger Iterationsschritt ausreichen.

Als Startwert verwenden wir die Näherung

$$y(a) + (b-a)f(a, y(a))$$

des expliziten Euler-Verfahrens, so dass sich nach einem Schritt der Fixpunktiteration die Näherung

$$\begin{aligned}\tilde{y}(b) &:= \Phi(y(a) + (b-a)f(a, y(a))) \\ &= y(a) + \frac{b-a}{2}(f(b, y(a) + (b-a)f(a, y(a))) + f(a, y(a)))\end{aligned}$$

ergibt, die wir als Näherung des Fixpunkts $\widehat{y}(b)$ verwenden können. So erhalten wir ein weiteres Näherungsverfahren:

Definition 8.10 (Heun-Verfahren) *Das durch*

$$\begin{aligned}\tilde{y}(t_0) &= y_0, \\ \tilde{y}(t_i) &= \tilde{y}(t_{i-1}) + \frac{h_i}{2}\left(f(t_i, \tilde{y}(t_{i-1}) + h_i f(t_{i-1}, \tilde{y}(t_{i-1}))) + f(t_{i-1}, \tilde{y}(t_{i-1}))\right) \\ &\qquad \text{für alle } i \in [1:k]\end{aligned}$$

definierte Verfahren zur näherungsweisen Berechnung der Lösung y der gewöhnlichen Differentialgleichung (8.1) nennt man das Heun-Verfahren.

Übungsaufgabe 8.11 (Fehlerabschätzung) *Sei $y \in C^3([a,b], V)$ Lösung der Gleichung (8.1), und sei f im zweiten Argument Lipschitz-stetig, es gelte also (8.4). Die Näherung des Heun-Verfahrens bezeichnen wir mit*

$$\tilde{y}(b) := y(a) + \frac{b-a}{2}\left(f(b, y(a) + (b-a)f(a, y(a))) + f(a, y(a))\right).$$

Beweisen Sie die Abschätzung

$$\|y(b) - \tilde{y}(b)\|_V \leq \frac{(b-a)^3}{4}\left(\frac{1}{3}\|y'''\|_{\infty,[a,b]} + L_f\|y''\|_{\infty,[a,b]}\right).$$

Hinweis: Analog zu Lemma 8.9. Durch Transformation lässt sich aus Übungsaufgabe 7.7 eine Abschätzung des Quadraturfehlers gewinnen. Lemma 8.5 erlaubt es uns, auch den Fehler der gestörten Quadraturformel zu beschränken.

8.3 Konsistenz und Konvergenz

Bisher sind wir bei der Analyse des Fehlers davon ausgegangen, dass wir den exakten Anfangswert kennen. In der Praxis ist der Anfangswert das Ergebnis einer Näherung aus dem vorangehenden Zeitschritt, so dass sich ein einmal entstandener Fehler durch das gesamte Verfahren hindurch fortpflanzt.

Um eine Aussage über den Gesamtfehler treffen zu können, untersuchen wir eine etwas verallgemeinerte Formulierung der bisher betrachteten Algorithmen:

Definition 8.12 (Einschrittverfahren) *Wir definieren*

$$\Delta_{a,b} := \{(t,h) \ : \ t \in [a,b], h \in (0, b-t]\},$$

um sicherzustellen, dass $t+h \in [a,b]$ für alle $(t,h) \in \Delta_{a,b}$ gesichert ist. Eine Funktion

$$\Phi : \Delta_{a,b} \times V \to V$$

bezeichnen wir als Verfahrensfunktion. *Mit ihrer Hilfe können wir Näherungswerte*

$$\begin{aligned} &\tilde{y}(t_0) := y_0, \\ &\tilde{y}(t_i) := \tilde{y}(t_{i-1}) + h_i \Phi(t_{i-1}, h_i, \tilde{y}(t_{i-1})) \qquad \text{für alle } i \in [1:k] \end{aligned}$$

der Lösung y der gewöhnlichen Differentialgleichung (8.1) definieren. Derartige Näherungsverfahren bezeichnen wir als Einschrittverfahren.

Alle bisher diskutierten Ansätze sind Einschrittverfahren mit den entsprechenden Verfahrensfunktionen

$$\begin{aligned} &\Phi(t,h,x) = f(t,x) && \text{(explizites Euler-Verfahren)}, \\ &\Phi(t,h,x) = f\left(t + \frac{h}{2}, x + \frac{h}{2} f(t,x)\right) && \text{(Runge-Verfahren)}, \\ &\Phi(t,h,x) = \frac{1}{2}\left(f(t+h, x+hf(t,x)) + f(t,x)\right) && \text{(Heun-Verfahren)}. \end{aligned}$$

Die Verwendung einer Verfahrensfunktion gibt uns unter anderem auch die Freiheit, Lösungen zu verschiedenen Anfangswerten zu konstruieren und miteinander zu vergleichen. Die Abhängigkeit von Anfangswerten und Anfangszeitpunkten lässt sich elegant durch Abbildungen beschreiben:

Definition 8.13 (Propagator) *Für $a, b \in \mathbb{R}$ mit $a \leq b$ definieren wir die Abbildung*

$$\Psi_{b,a} : V \to V$$

wie folgt: Für jedes $x \in V$ existiert genau eine Lösung $y \in C^1([a,b], V)$ des Anfangswertproblems

$$y(a) = x, \qquad y'(t) = f(t, y(t)) \qquad \text{für alle } t \in [a,b].$$

Wir setzen $\Psi_{b,a}(x) := y(b)$.

Die Abbildung $\Psi_{b,a}$ *nennen wir den* Propagator *der gewöhnlichen Differentialgleichung zu dem Anfangszeitpunkt* a *und dem Endzeitpunkt* b. *Sie ordnet einem Anfangwert zu dem Zeitpunkt* a *einen Endwert zu dem Zeitpunkt* b *zu.*

Infolge der eindeutigen Lösbarkeit des Anfangswertproblems spielt es keine Rolle, ob wir die Lösung der Differentialgleichung auf einem Gesamtintervall betrachten oder Lösungen auf Teilintervallen zusammensetzen.

Lemma 8.14 (Fortsetzungseigenschaft) *Seien* $a, b, c \in \mathbb{R}$ *mit* $a \leq b \leq c$ *gegeben. Dann gilt*

$$\Psi_{c,a}(x) = \Psi_{c,b}(\Psi_{b,a}(x)) \qquad \text{für alle } x \in V.$$

Beweis. Seien $y \in C^1([a,c], V)$, $z_1 \in C^1([a,b], V)$ und $z_2 \in C^1([b,c], V)$ Lösungen der Anfangswertprobleme

$$\begin{aligned} y(a) &= x, & y'(t) &= f(t, y(t)) & &\text{für alle } t \in [a,c], & &(8.8a)\\ z_1(a) &= x, & z_1'(t) &= f(t, z_1(t)) & &\text{für alle } t \in [a,b], & &(8.8b)\\ z_2(b) &= \Psi_{b,a}(x) & z_2'(t) &= f(t, z_2(t)) & &\text{für alle } t \in [b,c]. & &(8.8c) \end{aligned}$$

Gemäß Definition 8.13 gelten dann $\Psi_{c,a}(x) = y(c)$, $\Psi_{b,a}(x) = z_1(b)$ und $\Psi_{c,b}(\Psi_{b,a}(x)) = z_2(c)$.

Ein Blick auf (8.8a) zeigt, dass $y|_{[a,b]}$ auch eine Lösung des Anfangswertproblems (8.8b) ist, also muss $y|_{[a,b]} = z_1$ gelten.

Insbesondere gilt $y(b) = z_1(b) = \Psi_{b,a}(x)$, also ist $y|_{[b,c]}$ auch eine Lösung des Anfangswertproblems (8.8c), so dass insbesondere $y|_{[b,c]} = z_2$ gilt. Damit folgt

$$\Psi_{c,a}(x) = y(c) = z_2(c) = \Psi_{c,b}(\Psi_{b,a}(x)),$$

also die Behauptung. ∎

Das Gegenstück eines Propagators $\Psi_{b,a}$ bildet seine Approximation durch ein Einschrittverfahren.

Definition 8.15 (Diskreter Propagator) *Sei* $\Phi : \Delta_{a,b} \times V \to V$ *eine Verfahrensfunktion. Für alle* $i, j \in [0:k]$ *mit* $i \leq j$ *definieren wir Abbildungen* $\widetilde{\Psi}_{t_j,t_i} : V \to V$ *per Induktion über* $j - i$ *durch*

$$\widetilde{\Psi}_{t_j,t_i}(x) := \begin{cases} x & \text{falls } i = j, \\ x + h_j \Phi(t_i, h_j, x) & \text{falls } i + 1 = j, \\ \widetilde{\Psi}_{t_j,t_{j-1}}(\widetilde{\Psi}_{t_{j-1},t_i}(x)) & \text{ansonsten.} \end{cases}$$

Die Abbildungen $\widetilde{\Psi}_{t_j,t_i}$ *nennen wir den* diskreten Propagator *des durch* Φ *gegebenen Einschrittverfahrens zu dem Anfangszeitpunkt* t_i *und dem Endzeitpunkt* t_j. *Sie ordnen einem Anfangswert* $\tilde{y}(t_i) = x$ *zu dem Zeitpunkt* t_i *den durch das Verfahren berechneten Näherungswert* $\tilde{y}(t_j)$ *zu dem Zeipunkt* t_j *zu.*

Aus der Definition ergibt sich unmittelbar eine mit der in Lemma 8.14 bewiesenen vergleichbare Eigenschaft, allerdings aus naheliegenden Gründen nur in den diskreten Zeitpunkten.

Lemma 8.16 (Fortsetzungseigenschaft) *Seien $i, j, \ell \in [0:k]$ mit $i \leq j \leq \ell$ gegeben. Dann gilt*

$$\widetilde{\Psi}_{t_\ell,t_i}(x) = \widetilde{\Psi}_{t_\ell,t_j}(\widetilde{\Psi}_{t_j,t_i}(x)) \qquad \text{für alle } x \in V.$$

Beweis. Wir führen den Beweis per Induktion über $\ell - j \in \mathbb{N}_0$.

Induktionsanfang: Falls $\ell - j = 0$ gilt, folgt die Aussage unmittelbar aus Definition 8.15.

Induktionsvoraussetzung: Sei nun $n \in \mathbb{N}_0$ so gewählt, dass die Gleichung für alle $i, j, \ell \in [0:k]$ mit $i \leq j \leq \ell$ mit $\ell - j = n$ gilt.

Induktionsschritt: Seien $i, j, \ell \in [0:k]$ mit $i \leq j \leq \ell$ und $\ell - j = n + 1$ gegeben. Nach Definition gilt

$$\widetilde{\Psi}_{t_\ell,t_j}(x) = \widetilde{\Psi}_{t_\ell,t_{\ell-1}}(\widetilde{\Psi}_{t_{\ell-1},t_j}(x)) \qquad \text{für alle } x \in V.$$

Indem wir die Induktionsvoraussetzung auf das Tripel $i, j, \ell - 1$ anwenden, erhalten wir

$$\begin{aligned}\widetilde{\Psi}_{t_\ell,t_j}(\widetilde{\Psi}_{t_j,t_i}(x)) &= \widetilde{\Psi}_{t_\ell,t_{\ell-1}}(\widetilde{\Psi}_{t_{\ell-1},t_j}(\widetilde{\Psi}_{t_j,t_i}(x))) \\ &= \widetilde{\Psi}_{t_\ell,t_{\ell-1}}(\widetilde{\Psi}_{t_{\ell-1},t_i}(x)) = \widetilde{\Psi}_{t_\ell,t_i}(x) \qquad \text{für alle } x \in V,\end{aligned}$$

und damit ist die Induktionsbehauptung bewiesen. ∎

Mit Hilfe dieser Aussagen können wir nun eine Abschätzung für den Fehler konstruieren, indem wir zwischen die Näherungslösung und die exakte Lösung eine Näherungslösung einfügen, die von einem *exakten* Anfangswert in einem Zwischenpunkt ausgeht: Für $i \leq j \leq \ell$, $y_i \in V$ wie im Lemma gilt

$$\begin{aligned}\Psi_{t_\ell,t_i}(x) - \widetilde{\Psi}_{t_\ell,t_i}(x) &= \Psi_{t_\ell,t_j}(\Psi_{t_j,t_i}(x)) - \widetilde{\Psi}_{t_\ell,t_j}(\widetilde{\Psi}_{t_j,t_i}(x)) \\ &= \Psi_{t_\ell,t_j}(\Psi_{t_j,t_i}(x)) - \widetilde{\Psi}_{t_\ell,t_j}(\Psi_{t_j,t_i}(x)) \\ &\quad + \widetilde{\Psi}_{t_\ell,t_j}(\Psi_{t_j,t_i}(x)) - \widetilde{\Psi}_{t_\ell,t_j}(\widetilde{\Psi}_{t_j,t_i}(x)).\end{aligned}$$

Die erste Differenz beschreibt den Fehler zwischen der exakten Lösung und der Näherungslösung, der auf dem kürzeren Intervall $[t_j, t_\ell]$ entsteht. Diesen Fehler können wir per Induktion beschränken.

Die zweite Differenz hängt davon ab, wie empfindlich das Näherungsverfahren auf Störungen des Anfangswerts zu dem Zeitpunkt t_j reagiert. Diese zweite Differenz können wir mit einer geeigneten Stabilitätsabschätzung in den Griff bekommen.

Lemma 8.17 (Stabilität) *Sei die Verfahrensfunktion Φ Lipschitz-stetig im dritten Argument, es existiere also $L_\Phi \in \mathbb{R}_{\geq 0}$ mit*

$$\|\Phi(t,h,x) - \Phi(t,h,z)\|_V \leq L_\Phi \|x - z\|_V \qquad \text{für alle } (t,h) \in \Delta_{a,b},\ x, z \in V. \tag{8.9}$$

Seien $i, j \in [0:k]$ mit $i \leq j$ und $x, z \in V$ gegeben. Dann gilt

$$\|\widetilde{\Psi}_{t_j,t_i}(x) - \widetilde{\Psi}_{t_j,t_i}(z)\|_V \leq e^{L_\Phi(t_j - t_i)} \|x - z\|_V.$$

Beweis. Wir führen den Beweis per Induktion über $j - i \in \mathbb{N}_0$.

Induktionsanfang: Für $i = j$ ist die Aussage wegen $\widetilde{\Psi}_{t_i,t_i}(x) = x$ und $\widetilde{\Psi}_{t_i,t_i}(z) = z$ trivial.

Induktionsvoraussetzung: Sei nun $n \in \mathbb{N}_0$ so gewählt, dass die Behauptung für alle $i, j \in [0:k]$ mit $j - i = n$ gilt.

Induktionsschritt: Seien $i, j \in [0:k]$ mit $j-i = n+1$ gegeben. Dann gelten $(j-1)-i = n$ und $j-1 \in [0:k]$, also können wir die Induktionsvoraussetzung anwenden und erhalten

$$\|\widetilde{\Psi}_{t_{j-1},t_i}(x) - \widetilde{\Psi}_{t_{j-1},t_i}(z)\|_V \leq e^{L_\Phi(t_{j-1}-t_i)}\|x - z\|_V. \tag{8.10}$$

Nach Definition erhalten wir

$$\begin{aligned}\widetilde{\Psi}_{t_j,t_i}(x) - \widetilde{\Psi}_{t_j,t_i}(z) &= \widetilde{\Psi}_{t_{j-1},t_i}(x) + h_j\Phi(t_{j-1}, h_j, \widetilde{\Psi}_{t_{j-1},t_i}(x)) \\ &\quad - \widetilde{\Psi}_{t_{j-1},t_i}(z) + h_j\Phi(t_{j-1}, h_j, \widetilde{\Psi}_{t_{j-1},t_i}(z)) \\ &= (\widetilde{\Psi}_{t_{j-1},t_i}(x) - \widetilde{\Psi}_{t_{j-1},t_i}(z)) \\ &\quad + h_j(\Phi(t_{j-1}, h_j, \widetilde{\Psi}_{t_{j-1},t_i}(x)) - \Phi(t_{j-1}, h_j, \widetilde{\Psi}_{t_{j-1},t_i}(z))),\end{aligned}$$

so dass sich mit Dreiecksungleichung und (8.9) die Abschätzung

$$\begin{aligned}\|\widetilde{\Psi}_{t_j,t_i}(x) - \widetilde{\Psi}_{t_j,t_i}(z)\|_V &\leq \|\widetilde{\Psi}_{t_{j-1},t_i}(x) - \widetilde{\Psi}_{t_{j-1},t_i}(z)\|_V \\ &\quad + L_f h_j\|\widetilde{\Psi}_{t_{j-1},t_i}(x) - \widetilde{\Psi}_{t_{j-1},t_i}(z)\|_V \\ &= (1 + L_f h_j)\|\widetilde{\Psi}_{t_{j-1},t_i}(x) - \widetilde{\Psi}_{t_{j-1},t_i}(z)\|_V\end{aligned}$$

ergibt. An der Reihendarstellung der Exponentialfunktion lesen wir

$$1 + L_\Phi h_j \leq e^{L_\Phi h_j}$$

ab und erhalten so in Kombination mit (8.10) und $t_j = t_{j-1} + h_j$ die Abschätzung

$$\begin{aligned}\|\widetilde{\Psi}_{t_j,t_i}(x) - \widetilde{\Psi}_{t_j,t_i}(z)\|_V &\leq (1 + L_\Phi h_j)e^{L_\Phi(t_{j-1}-t_i)}\|x - z\|_V \\ &\leq e^{L_\Phi h_j}e^{L_\Phi(t_{j-1}-t_i)}\|x - z\|_V \\ &= e^{L_\Phi(h_j+t_{j-1}-t_i)}\|x - z\|_V = e^{L_\Phi(t_j-t_i)}\|x - z\|_V,\end{aligned}$$

mit der die Induktion vollständig ist. ∎

Um den Fehler auf dem Intervall $[t_i, t_j]$ abzuschätzen, zerlegen wir es in das kürzere Intervall $[t_i, t_{j-1}]$, das sich per Induktion behandeln lässt, und das Intervall $[t_{j-1}, t_j]$, auf dem sich die durch die numerische Approximation entstandenen Fehler gemäß Lemma 8.17 verstärken.

Satz 8.18 (Konvergenz) *Sei Φ Lipschitz-stetig im dritten Argument, es gelte also die Abschätzung (8.9). Sei $y \in C^1([a,b], V)$ Lösung des Anfangswertproblems (8.1). Seien*

$$y_i := y(t_i) \qquad \text{für alle } i \in [0:k]$$

die Werte dieser Lösung in den diskreten Zeitpunkten des Näherungsverfahrens. Sei $K_\Phi \in \mathbb{R}_{\geq 0}$ *eine Konstante, die*

$$\|\Psi_{t_j,t_{j-1}}(y_{j-1}) - \widetilde{\Psi}_{t_j,t_{j-1}}(y_{j-1})\|_V \leq K_\Phi h_j \qquad \text{für alle } j \in [1:k], \tag{8.11}$$

erfüllt, es gelte also eine Fehlerabschätzung für einzelne Schritte *des Verfahrens ausgehend von der exakten Lösung. Dann folgt*

$$\|\Psi_{t_j,t_i}(y_i) - \widetilde{\Psi}_{t_j,t_i}(y_i)\|_V \leq \begin{cases} \frac{e^{L_\Phi(t_j-t_i)}-1}{L_\Phi} K_\Phi & \text{falls } L_\Phi > 0, \\ (t_j - t_i) K_\Phi & \text{ansonsten} \end{cases}$$
$$\text{für alle } i,j \in [0:k],\ i \leq j.$$

Beweis. Falls die Lipschitz-Bedingung (8.9) für $L_\Phi = 0$ gilt, gilt sie auch für jedes $L_\Phi > 0$, so dass wir uns zunächst auf den Fall $L_\Phi > 0$ beschränken dürfen.

Auch diesen Beweis führen wir per Induktion über $j - i \in \mathbb{N}_0$.

Induktionsanfang: Für $i = j$ ist die Aussage wegen $\Psi_{t_i,t_i}(y_i) = y_i = \widetilde{\Psi}_{t_i,t_i}(y_i)$ trivial.

Induktionsvoraussetzung: Sei nun $n \in \mathbb{N}_0$ so gewählt, dass die Behauptung für alle $i,j \in [0:k]$ mit $j - i = n$ gilt.

Induktionsschritt: Seien $i,j \in [0:k]$ mit $j - i = n+1$ gegeben. Wir verwenden Lemma 8.14 und Lemma 8.16, um den Gesamtfehler als Summe aus dem Fehler im ersten Schritt und dem Restfehler darzustellen, und erhalten die Zerlegung

$$\begin{aligned}
\Psi_{t_j,t_i}(y_i) - \widetilde{\Psi}_{t_j,t_i}(y_i) &= \Psi_{t_j,t_{j-1}}(\Psi_{t_{j-1},t_i}(y_i)) - \widetilde{\Psi}_{t_j,t_{j-1}}(\widetilde{\Psi}_{t_{j-1},t_i}(y_i)) \\
&= \Psi_{t_j,t_{j-1}}(\Psi_{t_{j-1},t_i}(y_i)) - \widetilde{\Psi}_{t_j,t_{j-1}}(\Psi_{t_{j-1},t_i}(y_i)) \\
&\quad + \widetilde{\Psi}_{t_j,t_{j-1}}(\Psi_{t_{j-1},t_i}(y_i)) - \widetilde{\Psi}_{t_j,t_{j-1}}(\widetilde{\Psi}_{t_{j-1},t_i}(y_i)) \\
&= \Psi_{t_j,t_{j-1}}(y_{j-1}) - \widetilde{\Psi}_{t_j,t_{j-1}}(y_{j-1}) \\
&\quad + \widetilde{\Psi}_{t_j,t_{j-1}}(\Psi_{t_{j-1},t_i}(y_i)) - \widetilde{\Psi}_{t_j,t_{j-1}}(\widetilde{\Psi}_{t_{j-1},t_i}(y_i)).
\end{aligned}$$

Auf den ersten Term können wir unsere Voraussetzung anwenden, für den zweiten Term erhalten wir mit dem Stabilitätslemma 8.17 und der Induktionsvoraussetzung eine Abschätzung. Insgesamt ergibt sich wir

$$\begin{aligned}
\|\Psi_{t_j,t_i}(y_i) - \widetilde{\Psi}_{t_j,t_i}(y_i)\|_V &\leq \|\Psi_{t_j,t_{j-1}}(y_{j-1}) - \widetilde{\Psi}_{t_j,t_{j-1}}(y_{j-1})\|_V \\
&\quad + \|\widetilde{\Psi}_{t_j,t_{j-1}}(\Psi_{t_{j-1},t_i}(y_i)) - \widetilde{\Psi}_{t_j,t_{j-1}}(\widetilde{\Psi}_{t_{j-1},t_i}(y_i))\|_V \\
&\leq K_\Phi h_j + e^{L_\Phi h_j} \|\Psi_{t_{j-1},t_i}(y_i) - \widetilde{\Psi}_{t_{j-1},t_i}(y_i)\|_V \\
&\leq K_\Phi h_j + e^{L_\Phi h_j} \frac{e^{L_\Phi(t_{j-1}-t_i)} - 1}{L_\Phi} K_\Phi \\
&\leq \frac{e^{L_\Phi h_j} - 1}{L_\Phi} K_\Phi + \frac{e^{L_\Phi(t_j - t_i)} - e^{L_\Phi h_j}}{L_\Phi} K_\Phi \\
&= \frac{e^{L_\Phi(t_j-t_i)} - 1}{L_\Phi} K_\Phi.
\end{aligned}$$

erhalten, wobei wir im vorletzten Schritt die Abschätzung $1 + L_\Phi h_j \leq e^{L_\Phi h_j}$ verwendet haben. Damit ist die Induktion abgeschlossen.

Falls die Lipschitz-Bedingung (8.9) auch mit $L_\Phi = 0$ gelten sollte, können wir mit Hilfe der Regel von l'Hôspital zu dem Grenzwert

$$\lim_{L_\Phi \to 0} \frac{e^{L_\Phi (t_j - t_i)} - 1}{L_\Phi} = \frac{(t_j - t_i) e^{0(t_j - t_i)}}{1} = t_j - t_i$$

übergehen und erhalten die gewünschte Abschätzung. ∎

Wie man sieht dürfen wir nur dann darauf hoffen, dass die Näherungslösung gegen die exakte Lösung konvergiert, wenn wir den Wert der Konstanten K_Φ reduzieren können.

Um das Verhalten dieser Konstanten genauer zu beschreiben, führen wir den Begriff der *Konsistenz* einer Verfahrensfunktion beziehungsweise eines Einschrittverfahrens ein:

Definition 8.19 (Konsistenz) *Sei eine Verfahrensfunktion Φ gegeben. Die durch*

$$\tau(t,h,x) := \frac{\Psi_{t+h,t}(x) - \widetilde{\Psi}_{t+h,t}(x)}{h} \qquad \text{für alle } (t,h) \in \Delta_{a,b},\ x \in V \tag{8.12}$$

definierte Abbildung $\tau : \Delta_{a,b} \times V \to V$ bezeichnen wir als den Konsistenzfehler*.*

Falls für eine Lösung y des Problems (8.1) Konstanten $C_{ko} \in \mathbb{R}_{\geq 0}$, $h_{ko} \in \mathbb{R}_{>0} \cup \{\infty\}$ und $n \in \mathbb{N}$ existieren, die

$$\|\tau(t,h,y(t))\|_V \leq C_{ko} h^n \qquad \text{für alle } (t,h) \in \Delta_{a,b},\ h \leq h_{ko},\ x \in V$$

erfüllen, nennen wir Φ konsistent von Ordnung n *mit dieser Lösung. In diesem Fall nennen wir auch das durch Φ definierte Einschrittverfahren konsistent von Ordnung n.*

Die etwas umständliche Definition, die Konsistenz nur für bestimmte Lösungen fordert, ist notwendig, da sie häufig davon abhängt, wie häufig die Lösung differenzierbar ist. Beispielsweise für das Runge-Verfahren können wir

$$C_{\mathrm{ko}} := \frac{1}{8}\left(\frac{1}{3}\|y'''\|_{\infty,[a,b]} + L\|y''\|_{\infty,[a,b]}\right), \qquad h_{\mathrm{ko}} := \infty, \qquad n := 2$$

setzen, falls y dreimal stetig differenzierbar ist, und die Konsistenz zweiter Ordnung aus Lemma 8.9 folgern.

Für das explizite Euler-Verfahren folgt Konsistenz erster Ordnung für zweimal stetig differenzierbare Lösungen y aus Lemma 8.5 mit den Konstanten

$$C_{\mathrm{ko}} := \frac{1}{2}\|y''\|_{\infty,[a,b]}, \qquad h_{\mathrm{ko}} := \infty, \qquad n := 1.$$

Falls wir wissen, dass ein Einschrittverfahren konsistent ist, nimmt die Fehlerabschätzung aus Satz 8.18 die folgende besonders einfache Form an:

Folgerung 8.20 (Konvergenz) *Die Verfahrensfunktion Φ erfülle die Lipschitz-Bedingung (8.9) und sei konsistent von Ordnung n mit der Lösung y des Problems (8.1). Sei $h := \max\{h_1, \ldots, h_k\} \leq h_{ko}$. Dann gilt*

$$\|y(t_j) - \tilde{y}(t_j)\|_V \leq \begin{cases} C_{ko}\frac{e^{L_\Phi(t_j-a)}-1}{L_\Phi}h^n & \text{falls } L_\Phi > 0, \\ C_{ko}(t_j-a)h^n & \text{ansonsten} \end{cases} \qquad \text{für alle } j \in [0:k].$$

Beweis. Nach (8.12) gilt

$$\|\Psi_{t_j,t_{j-1}}(y_{j-1}) - \widetilde{\Psi}_{t_j,t_{j-1}}(y_{j-1})\|_V \leq \|\tau(t_{j-1}, h_j, y_{j-1})\|_V h_j \leq C_{\mathrm{ko}} h_j^{n+1} \qquad \text{für alle } j \in [1:k],$$

also ist die Bedingung (8.11) mit $K_\Phi := C_{\mathrm{ko}}h^n$ erfüllt.

Mit Satz 8.18 folgt die Behauptung. ∎

Ein konsistentes Verfahren erlaubt es uns also, jede beliebige Genauigkeit zu erreichen. Der dafür erforderliche Aufwand hängt von der Konsistenzordnung n ab: Je höher die Ordnung ist, desto stärker wird der Fehler durch das Hinzufügen weiterer Teilintervalle reduziert.

8.4 Verfahren höherer Ordnung

Der Rechenaufwand eines Einschrittverfahrens ist in der Regel proportional zu der Anzahl k der Schritte, die durchgeführt werden müssen. Bei einem konsistenten Verfahren n-ter Ordnung besagt Folgerung 8.20, dass sich der Fehler der berechneten Näherungswerte wie k^{-n} verhält. Um mit möglichst geringem Aufwand eine gegebene Genauigkeit zu erreichen, ist es demnach sinnvoll, nach Verfahren möglichst hoher Ordnung zu suchen.

Eine Möglichkeit besteht darin, den bereits bei dem Runge-Verfahren verwendeten Ansatz zu verallgemeinern: Bei diesem Verfahren wurde zunächst eine Näherung der Lösung in der Mitte des Intervalls berechnet, und diese Näherung floss dann in die Berechnung einer besseren Näherung ein. Diesen Ansatz können wir fortführen: Wir berechnen eine erste Näherung, bestimmen mit ihrer Hilfe eine zweite Näherung, dann eine dritte, und so weiter. Wie bei Quadraturformeln ist es sinnvoll, davon auszugehen, dass die Näherungen sich als Linearkombinationen von Auswertungen der Funktion f ergeben, und indem wir die Ergebnisse dieser Auswertungen in Hilfsvariablen $k_1, \ldots, k_m \in V$ speichern, erhalten wir die Rechenvorschrift

$$\begin{aligned} k_1 &:= f(t + c_1 h, x), \\ k_2 &:= f(t + c_2 h, x + a_{21} h k_1), \\ k_3 &:= f(t + c_3 h, x + a_{31} h k_1 + a_{32} h k_2), \\ &\vdots \\ k_m &:= f(t + c_m h, x + a_{m1} h k_1 + \ldots + a_{m,m-1} h k_{m-1}) \end{aligned}$$

für die Hilfsvariablen, aus denen wir dann per

$$\Phi(t,h,x) := b_1 k_1 + \ldots + b_m k_m$$

den Wert der Verfahrensfunktion gewinnen können. Charakteristisch für dieses Verfahren sind die Zahl m der Hilfsgrößen, die Koeffizienten $c_1, \ldots, c_m \in [0,1]$, die festlegen, zu welchen Zeitpunkten im Intervall $[t, t+h]$ diese Größen gehören, die Koeffizienten $a_{21}, a_{31}, a_{32}, \ldots, a_{m,m-1} \in \mathbb{K}$, die die in den Zwischenschritten bestimmten Näherungen definieren, sowie die Koeffizienten $b_1, \ldots, b_m \in \mathbb{K}$, die die abschließende Linearkombination beschreiben.

Definition 8.21 (Runge-Kutta-Verfahren) *Sei $m \in \mathbb{N}$, seien Vektoren $\mathbf{b}, \mathbf{c} \in \mathbb{K}^m$ und eine Matrix $\mathbf{A} \in \mathbb{K}^{m \times m}$ gegeben. Falls*

$$c_i \in [0,1] \qquad \text{für alle } i \in [1:m]$$

und

$$a_{ij} = 0 \qquad \text{für alle } i,j \in [1:m] \text{ mit } i \leq j$$

gelten, definieren die Formeln

$$k_i(t,h,x) := f\left(t + c_i h, x + h \sum_{j=1}^{i-1} a_{ij} k_j(t,h,x)\right) \qquad \text{für alle } (t,h) \in \Delta_{a,b},\ x \in V, \text{ und alle } i \in [1:m],$$

$$\Phi(t,h,x) := \sum_{i=1}^{m} b_i k_i(t,h,x) \qquad \text{für alle } (t,h) \in \Delta_{a,b},\ x \in V$$

die Verfahrensfunktion eines Einschrittverfahrens. Dieses Verfahren bezeichnen wir als das Runge-Kutta-Verfahren *der Stufe m zu den Vektoren $\mathbf{b}, \mathbf{c}$ und der Matrix $\mathbf{A}$.*

Zur Abkürzung werden die Koeffizienten eines Runge-Kutta-Verfahren häufig in der Form

$$\begin{array}{c|cccc} c_1 & & & & \\ c_2 & a_{21} & & & \\ \vdots & \vdots & \ddots & & \\ c_m & a_{m1} & \ldots & a_{m,m-1} & \\ \hline & b_1 & \ldots & b_{m-1} & b_m \end{array}$$

zusammengefasst, an der sich auch m bequem ablesen lässt. Die Darstellung bezeichnet man als *Butcher-Schema*.

Das uns bereits bekannte explizite Euler-Verfahren ist in diesem Sinne ein einstufiges Runge-Kutta-Verfahren, das durch das Butcher-Schema

$$\begin{array}{c|c} 0 & \\ \hline & 1 \end{array}$$

beschrieben ist. Entsprechend sind das Runge- und das Heun-Verfahren durch

$$\begin{array}{c|cc} 0 & & \\ 1/2 & 1/2 & \\ \hline & 0 & 1 \end{array} \qquad\qquad \begin{array}{c|cc} 0 & & \\ 1 & 1 & \\ \hline & 1/2 & 1/2 \end{array}$$

als zweistufige Runge-Kutta-Verfahren charakterisiert.

Es stellt sich die Frage, wie sich Runge-Kutta-Verfahren höherer Ordnung konstruieren lassen. Einen Teil der Antwort erhalten wir, indem wir die gewöhnliche Differentialgleichung (8.1) mit der speziellen Funktion

$$f(t,x) = g(t) \qquad \text{für alle } t \in [a,b]$$

untersuchen. Aus Lemma 8.1 folgt unmittelbar, dass

$$y(t) = y_0 + \int_a^t g(s)\,ds \qquad \text{für alle } t \in [a,b]$$

gilt, das Lösen der Differentialgleichung ist also auf die Auswertung eines Integrals reduziert.

Die Verfahrensfunktion eines m-stufiges Runge-Kutta-Verfahren nimmt in diesem Fall die Form

$$\begin{aligned} k_i &= g(t + hc_i) && \text{für alle } i \in [1:m], \\ \Phi(t,h,x) &= \sum_{i=1}^m b_i k_i = \sum_{i=1}^m b_i g(t+hc_i) \end{aligned}$$

an, also die einer Quadraturformel für das Intervall $[t, t+h]$, bei der $c_1, \ldots, c_m$ die Stützstellen auf dem Einheitsintervall $[0,1]$ und $b_1, \ldots, b_m$ die zugehörigen Gewichte sind.

Deshalb liegt es nahe, bei der Suche nach neuen Runge-Kutta-Verfahren von Quadraturformeln auszugehen. Aus der Simpson-Quadraturformel (7.7) ergibt sich durch geeignete Wahl der Matrix $\mathbf{A}$ so das *klassische Runge-Kutta-Verfahren*, das durch das Butcher-Schema

$$\begin{array}{c|cccc} 0 & & & & \\ 1/2 & 1/2 & & & \\ 1/2 & 0 & 1/2 & & \\ 1 & 0 & 0 & 1 & \\ \hline & 1/6 & 1/3 & 1/3 & 1/6 \end{array}$$

definiert ist. Es lässt sich nachweisen, dass dieses Verfahren von vierter Ordnung ist, falls y hinreichend oft stetig differenzierbar ist.

Die Konstruktion von Runge-Kutta-Verfahren höherer Ordnung ist relativ schwierig, weil sich die Koeffizienten der Matrix $\mathbf{A}$ im allgemeinen Fall als Lösung eines unhandlichen nichtlinearen Gleichungssystems ergeben. Es lässt sich theoretisch beweisen, dass ein Runge-Kutta-Verfahren fünfter Ordnung mindestens sechsstufig sein muss, und dass

auch bei höheren Ordnungen die Anzahl der Stufen echt größer als die Ordnung sein muss.

Für die Konstruktion von Verfahren höherer Ordnung bietet es sich deshalb an, nach alternativen Ansätzen zu suchen. Eine Möglichkeit bieten Extrapolationsverfahren: Falls y hinreichend oft differenzierbar ist, können wir beispielsweise mit einem einfachen Verfahren Näherungslösungen $\widetilde{\Psi}_{t+h,t}(x)$ für die Schrittweite h und $\widetilde{\Psi}_{t+h,t+h/2}(\widetilde{\Psi}_{t+h/2,t}(x))$ für die Schrittweite $h/2$ berechnen und durch Extrapolation eine Näherung höherer Ordnung für $\Psi_{t+h,t}(x)$ erhalten. In dieser Weise lassen sich im Prinzip Verfahren beliebig hoher Ordnung relativ einfach konstruieren, allerdings mit im Vergleich zu Runge-Kutta-Verfahren wesentlich höherem Rechenaufwand.

8.5 Verfeinerungen und Erweiterungen

Sehr häufig wird sich die Lösung y der gewöhnlichen Differentialgleichung (8.1) nicht auf dem gesamten Definitionsgebiet $[a, b]$ einheitlich verhalten. Falls die Gleichung eine chemische Reaktion beschreibt, wird oft der Fall eintreten, dass über lange Zeit nur wenig passiert, bis dann plötzlich sehr schnell eine Veränderung eintritt. Bei der Simulation eines Mondflugs wird das Raumschiff in der Nähe der Erde oder des Monds seine Flugbahn deutlich stärker ändern als auf der Verbindungsstrecke. Es liegt nahe, diese Eigenschaften der Lösung auszunutzen.

Bemerkung 8.22 (Adaptive Schrittweite) *Für das explizite Euler-Verfahren erhalten wir aus Lemma 8.5 die Abschätzung*

$$\|\tau(t,h,y(t))\|_V = \frac{\|\Psi_{t+h,t}(y(t)) - \widetilde{\Psi}_{t+h,t}(y(t))\|_V}{h} \leq \frac{h}{2}\|y''\|_{\infty,[t,t+h]},$$

wir erhalten also eine Abschätzung des Konsistenzfehlers, die nur von der zweiten Ableitung auf dem relevanten Teilintervall abhängt.

Falls wir die Stützstellen $a = t_0 < t_1 < \ldots < t_k = b$ so wählen können, dass

$$\|\tau(t,h_i,y_{i-1})\|_V \leq \frac{h_i}{2}\|y''\|_{\infty,[t_{i-1},t_i]} \leq \epsilon \qquad \text{für alle } i \in [1:k] \tag{8.13}$$

gilt, folgt aus Satz 8.18 im Fall $L_\Phi > 0$ insbesondere die Abschätzung

$$\|\tilde{y}(t_j) - y(t_j)\|_V \leq \frac{e^{L_\Phi(t_j-a)-1}}{L_\Phi}\epsilon \qquad \text{für alle } j \in [0:k].$$

Die Voraussetzung (8.13) bedeutet, dass wir zu den Zeiten, zu denen y'' kleine Werte annimmt, mit relativ großen Schrittweiten arbeiten können, ohne den Gesamtfehler zu gefährden. Mit diesem Zugang können wir eine hohe Genauigkeit trotz geringen Rechenaufwands erreichen. Man spricht von einer adaptiven Diskretisierung, *weil die Punkte $t_0, \ldots, t_k$ an die Lösung y angepasst werden.*

Der praktischen Anwendung der adaptiven Schrittweitensteuerung steht im Wege, dass wir schon y nicht kennen, also erst recht nicht die höheren Ableitungen, von denen der Approximationsfehler abhängt. In der Praxis behilft man sich deshalb damit, den Fehler zu schätzen.

Bemerkung 8.23 (Fehlerschätzer) *In der Praxis wird die Größe des durch das Näherungsverfahren eingeführten Fehlers geschätzt, indem man unterschiedliche Verfahren miteinander vergleicht. Beispielsweise könnte man eine Näherung* $\widetilde{\Psi}^{(n)}_{t+h,t}(x)$ *mit einem Verfahren n-ter Ordnung und eine zweite Näherung* $\widetilde{\Psi}^{(n+1)}_{t+h,t}(x)$ *mit einem Verfahren* $(n+1)$*-ter Ordnung berechnen und beide vergleichen. Falls die Schrittweite klein genug ist, dürfen wir davon ausgehen, dass das „bessere“ Verfahren eine genauere Näherung bestimmen wird, so dass die Differenz zwischen beiden Näherungen proportional zu der Differenz von* $\widetilde{\Psi}^{(n)}_{t+h,t}(x)$ *zu der exakten Lösung* $\Psi_{t+h,t}(x)$ *ist. Quantitativ lässt sich der Ansatz mit Hilfe der Dreiecksungleichung motivieren: Es gilt*

$$\begin{aligned}\|\widetilde{\Psi}^{(n)}_{t+h,t}(x) - \Psi_{t+h,t}(x)\|_V \leq \|\widetilde{\Psi}^{(n)}_{t+h,t}(x) - \widetilde{\Psi}^{(n+1)}_{t+h,t}(x)\|_V \\ + \|\widetilde{\Psi}^{(n+1)}_{t+h,t}(x) - \Psi_{t+h,t}(x)\|_V,\end{aligned}$$

und falls wir voraussetzen, dass der Fehler des „besseren“ Verfahrens um einen Faktor $\alpha < 1$ *kleiner als der des „schlechteren“ ist, dass also*

$$\|\widetilde{\Psi}^{(n+1)}_{t+h,t}(x) - \Psi_{t+h,t}(x)\|_V \leq \alpha\|\widetilde{\Psi}^{(n)}_{t+h,t}(x) - \Psi_{t+h,t}(x)\|_V \tag{8.14}$$

gilt, erhalten wir

$$\begin{aligned}(1-\alpha)\|\widetilde{\Psi}^{(n)}_{t+h,t}(x) - \Psi_{t+h,t}(x)\|_V &\leq \|\widetilde{\Psi}^{(n)}_{t+h,t}(x) - \widetilde{\Psi}^{(n+1)}_{t+h,t}(x)\|_V, \\ \|\widetilde{\Psi}^{(n)}_{t+h,t}(x) - \Psi_{t+h,t}(x)\|_V &\leq \frac{1}{1-\alpha}\|\widetilde{\Psi}^{(n)}_{t+h,t}(x) - \widetilde{\Psi}^{(n+1)}_{t+h,t}(x)\|_V,\end{aligned}$$

unter diesen Bedingungen können also tatsächlich den Fehler durch eine praktisch berechenbare Größe beschränken. Unter etwas stärkeren Voraussetzungen ist es sogar möglich, die für eine vorgegebene Genauigkeit ϵ *optimale Schrittweite anzunähern.*

Kritisch bei dieser Methode ist, dass wir mit der sogenannten Sättigungsannahme *(8.14) arbeiten müssen, die im Wesentlichen beschreibt, dass die Schrittweite so klein und die Lösung y lokal so gutartig ist, dass das Verfahren höherer Ordnung wesentlich besser als das niedrigerer Ordnung arbeitet. Derartige zusätzliche Annahmen über die Lösung wollten wir aber eigentlich gerade mit Hilfe des Fehlerschätzers vermeiden.*

Ein gängiges Maß für den mit einem Verfahren verbundenen Rechenaufwand ist die Anzahl der Auswertungen der Funktion f. Wie wir bereits gesehen haben, sind bei Runge-Kutta-Verfahren der Ordnung m mindestens m solcher Auswertungen erforderlich, bei Extrapolationsverfahren in der Regel deutlich mehr. Durch geschicktes Wiederverwenden der Ergebnisse vorangegangener Schritte lässt sich der Aufwand deutlich reduzieren:

procedure multistep;
Berechne $\tilde{y}_i \approx y(t_i)$ und $f_i \approx f(t_i, y(t_i))$ für $i \in [0:m]$;
for $i = m, \ldots, k-1$ **do**
  $\tilde{y}_{i+1} \leftarrow \tilde{y}_i$;
  for $j \in [0:m]$ **do**
    $\tilde{y}_{i+1} \leftarrow \tilde{y}_{i+1} + w_{i,j} f_{i-j}$
  end for;
  $f_{i+1} \leftarrow f(t_{i+1}, \tilde{y}_{i+1})$
end for

Abbildung 8.1: Approximation der Lösung y per Mehrschrittverfahren.

Bemerkung 8.24 (Mehrschrittverfahren) *Bei der Konstruktion einer alternativen Möglichkeit zur Approximation der Lösung y können wir wieder von Lemma 8.1 ausgehen, also von*

$$y(t_{i+1}) = y(t_i) + \int_{t_i}^{t_{i+1}} f(s, y(s))\, ds,$$

allerdings ersetzen wir den Integranden jetzt durch einen Interpolanten, der durch Interpolationspunkte definiert ist, die teilweise außerhalb *des Intervalls $[t_i, t_{i+1}]$ liegen: Wir verwenden*

$$f(s, y(s)) \approx \sum_{j=0}^{m} f(t_{i-j}, y(t_{i-j}))\ell_{i,j}(s) \qquad \text{für alle } s \in [t_i, t_{i+1}],$$

wobei die Lagrange-Polynome durch

$$\ell_{i,j}(s) = \prod_{\substack{\ell=0 \\ \ell \neq j}}^{m} \frac{s - t_{i-\ell}}{t_{i-j} - t_{i-\ell}} \qquad \text{für alle } s \in [t_i, t_{i+1}]$$

definiert sind, also gerade zu den Stützstellen $t_{i-m}, \ldots, t_i$ passen. Damit erhalten wir die Approximation

$$y(t_{i+1}) \approx y(t_i) + \sum_{j=0}^{m} f(t_{i-j}, y(t_{i-j})) \int_{t_i}^{t_{i+1}} \ell_{i,j}(s)\, ds,$$

die sich mit den Abkürzungen

$$w_{i,j} := \int_{t_i}^{t_{i+1}} \ell_{i,j}(s)\, ds \qquad \text{für alle } i \in \mathbb{N}_{\geq m},\ j \in [0:m]$$

in der kompakten Form

$$y(t_{i+1}) \approx y(t_i) + \sum_{j=0}^{m} w_{i,j} f(t_{i-j}, y(t_{i-j}))$$

darstellen lässt. Indem wir die exakten Funktionswerte durch Näherungen aus den vorangehenden Schritten ersetzen und die Auswertungen von f in diesen Punkten zwischenspeichern, erhalten wir das in Abbildung 8.5 dargestellte Verfahren, das nur eine *Auswertung der Funktion f pro Schritt benötigt und das unter geeigneten Voraussetzungen trotzdem in einem geeigneten Sinne von m-ter Ordnung ist.*

Da bei der Berechnung der Näherung nun nicht nur die Näherung des vorangehenden Zeitschritts eingeht, sondern Näherungen mehrerer Schritte, tragen derartige Verfahren die Bezeichnung Mehrschrittverfahren.

In der Regel beschränkt man sich bei Mehrschrittverfahren auf feste Schrittweiten, da die Gewichte $w_{i,j}$ dann nicht von i abhängen, so dass sie nur einmal für $j \in [0:m]$ vorberechnet werden müssen. Die Analyse der Fehlerfortpflanzung ist für diese Verfahren etwas anspruchsvoller als für Einschrittverfahren, im Fall des hier vorgestellten Adams-Bashforth-Verfahrens *lässt sich für eine $(m+1)$-mal stetig differenzierbare Funktion allerdings beweisen, dass die Näherungen $\tilde{y}_i$ von m-ter Ordnung gegen die exakten Werte der Lösung konvergieren.*

Die bisher vorgestellten Konvergenzaussagen ermöglichen es uns bereits, jede hinreichend glatte Lösung im Prinzip zu approximieren. Allerdings gibt es Problemklassen, bei denen die Lösung zwar glatt ist, sie aber von den bisher diskutierten Verfahren trotzdem erst für sehr geringere Schrittweiten hinreichend gut angenähert werden kann. Die Ursache ist eine zu große Lipschitz-Konstante L_Φ der Verfahrensfunktion, die in Lemma 8.17 zu einer unattraktiv hohen Stabilitätskonstanten führt. Durch die Wahl eines passenden Verfahrens lässt sich dieses Problem elegant lösen:

Bemerkung 8.25 (Implizite Verfahren) *Wir untersuchen die besonders einfache Differentialgleichung*

$$y(0) = 1, \qquad y'(t) = \alpha y(t) \qquad \text{für alle } t \in [0,1],$$

für ein $\alpha \in \mathbb{R}$, deren Lösung durch

$$y(t) = e^{\alpha t} \qquad \text{für alle } t \in [0,1]$$

gegeben ist. Für das explizite Euler-Verfahren erhalten wir die Verfahrensfunktion

$$\Phi(t,h,x) = \alpha x \qquad \text{mit der Lipschitz-Konstanten} \qquad L_\Phi = |\alpha|,$$

die für großes α zu einem sehr großen Faktor in der Konvergenzabschätzung des Satzes 8.18 führt.

Indem wir für die Approximation der Integralformulierung des Problems den rechten Randpunkt des Integrals anstatt des bisher verwendeten linken Randpunkts verwenden, erhalten wir das implizite Euler-Verfahren, *das durch*

$$y(t+h) = y(t) + \int_t^{t+h} f(s, y(s))\, ds \approx y(t) + hf(t+h, y(t+h))$$

definiert wird, so dass für die Berechnung einer Näherung des Werts $y(t+h)$ das im Allgemeinen nichtlineare Gleichungssystem

$$0 = y(t) + hf(t+h, \tilde{y}(t+h)) - \tilde{y}(t+h)$$

gelöst werden muss, etwa mit den Verfahren aus Kapitel 5.

Für den hier untersuchten einfachen Modellfall erhalten wir

$$\begin{aligned} 0 &= y(t) + \alpha h\tilde{y}(t+h) - \tilde{y}(t+h) = y(t) - (1-\alpha h)\tilde{y}(t+h), \\ \tilde{y}(t+h) &= \frac{y(t)}{1-\alpha h} = y(t) + h\frac{\alpha}{1-\alpha h}y(t), \end{aligned}$$

und das entspricht gerade der Verfahrensfunktion

$$\Phi(t,h,x) = \frac{\alpha}{1-\alpha h}x \qquad \textit{mit der Lipschitz-Konstanten} \qquad L_\Phi = \frac{|\alpha|}{|1-\alpha h|}.$$

Interessant ist hier der Fall $\alpha < 0$, in dem wir

$$L_\Phi = \frac{|\alpha|}{1-\alpha h} \leq \min\left\{|\alpha|, \frac{1}{h}\right\}$$

erhalten, also eine Lipschitz-Konstante, die für große Schrittweiten unabhängig von α beschränkt ist und sich für kleiner Schrittweiten wie die des expliziten Verfahrens verhält. Demnach erlauben es uns implizite Verfahren bei bestimmten Anfangswertproblemen, mit wesentlich größeren Schrittweiten als ihre expliziten Gegenstück zu arbeiten.

Diese Situation tritt etwa bei der Behandlung sogenannter steifer Differentialgleichungen *auf, die beispielsweise bei der Behandlung der Simulation der elektromagnetischer Felder oder der Wärmeleitung eine wichtige Rolle spielen. Um sie behandeln zu können wurden für die meisten in diesem Kapitel diskutierten Verfahren auch implizite Varianten entwickelt.*

9 Anwendungsbeispiele

Obwohl in dieser Vorlesung lediglich die grundlegenden Verfahren für die numerische Behandlung einiger wichtiger mathematischer Aufgabenstellungen behandelt werden, lassen sich trotzdem schon eine Reihe „echter Probleme“ mit ihrer Hilfe lösen.

9.1 Resonanzfrequenzen und Eigenwerte

Schwingungsphänomene spielen beispielsweise bei der Ausbreitung elektromagnetischer oder akustischer Wellen eine wichtige Rolle. Sie sind auch in der Strukturmechanik wichtig, etwa um die Resonanzfrequenzen einer Brücke zu berechnen: Es ist beispielsweise in Deutschland nach §27(6) StVO verboten, im Gleichschritt über eine Brücke zu marschieren, denn falls die Schritte der Resonanzfrequenz nahe kommen, kann sich die Schwingung aufschaukeln und die Brücke zum Einsturz bringen.

Als Beispiel untersuchen wir eine an beiden Enden eingespannte Saite (etwa einer Gitarre). Mit $u(x,t)$ bezeichnen wir ihre Auslenkung aus der Ruhelage in einem Punkt x zu einem Zeitpunkt t.

Der Einfachheit halber beschränken wir uns auf eine Saite der Länge eins, die lediglich nach oben oder unten ausgelenkt werden kann. In diesem Fall suchen wir nach einer Funktion $u : [0,1] \times \mathbb{R} \to \mathbb{R}$, die jedem Punkt auf der Saite und jedem Zeitpunkt eine Auslenkung zuordnet. Da die Saite an beiden Enden eingespannt ist, erhalten wir die Randbedingungen

$$u(0,t) = u(1,t) = 0 \qquad \text{für alle } t \in \mathbb{R},$$

die wir schon im Fall der Wärmeleitungsgleichung verwendet haben.

Das Schwingungsverhalten der Saite wird durch die *Wellengleichung*

$$\frac{\partial^2 u}{\partial t^2}(x,t) = c\frac{\partial^2 u}{\partial x^2}(x,t) \qquad \text{für alle } x \in [0,1],\ t \in \mathbb{R} \tag{9.1}$$

beschrieben, in der der Parameter c von den Eigenschaften der Saite abhängt, beispielsweise von ihrer Dicke und ihrem Material.

Für lineare Differentialgleichungen dieser Art lässt sich die Lösung mit Hilfe eines Separationsansatzes darstellen: Wir gehen von

$$u(x,t) = u_0(x)\cos(\omega t) \qquad \text{für alle } x \in [0,1],\ t \in \mathbb{R}$$

aus, wobei $u_0 : [0,1] \to \mathbb{R}$ eine Abbildung und $\omega \in \mathbb{R}$ ein Parameter sind. ω bezeichnet man als *Frequenz* der Funktion u. In diesem Fall gilt

$$\frac{\partial u}{\partial t}(x,t) = -\omega u_0(x)\sin(\omega t),$$

$$\frac{\partial^2 u}{\partial t^2}(x,t) = -\omega^2 u_0(x)\cos(\omega t) \qquad \text{für alle } x \in [0,1],\ t \in \mathbb{R}.$$

Durch Einsetzen in die Wellengleichung (9.1) erhalten wir

$$-\omega^2 u_0(x)\cos(\omega t) = c u_0''(x)\cos(\omega t) \qquad \text{für alle } x \in [0,1],\ t \in \mathbb{R},$$

so dass wir den Cosinus-Faktor eliminieren können, um zu der Gleichung

$$\omega^2 u_0(x) = -c u_0''(x) \qquad \text{für alle } x \in [0,1] \tag{9.2}$$

zu gelangen. Diese Gleichung ähnelt einem *Eigenwertproblem*: Auf der rechten Seite steht ein linearer Operator, denn das zweimalige Differenzieren und die Multiplikation mit dem Faktor $-c$ sind lineare Operationen, und gesucht ist eine Funktion $u_0 \neq 0$, die bei Anwendung dieses Operators lediglich ihre Länge ändert. u_0 bezeichnet man deshalb als *Eigenfunktion* zu dem *Eigenwert* ω^2, ω heißt *Eigenfrequenz*.

In diesem speziellen Fall kann man die Gleichung (9.2) analytisch lösen, indem man den Ansatz $u_0(x) = \sin(\pi k x)$ wählt.

Im allgemeinen Fall verwendet man üblicherweise eine *Diskretisierung*, man ersetzt also das kontinuierliche Intervall $[0,1]$ durch diskrete Punkte. Wir wählen dazu ein $n \in \mathbb{N}$ die Punkte

$$x_i = ih, \qquad h := \frac{1}{n+1} \qquad \text{für alle } i \in \{0,\dots,n+1\}.$$

Wir wollen eine Eigenfunktion u_0 in diesen Punkten berechnen, sind also an den Werten

$$u_i = u_0(x_i) \qquad \text{für alle } i \in \{0,\dots,n+1\}$$

interessiert. Aus den Randbedingungen folgen die Gleichungen $u_0 = 0$ und $u_{n+1} = 0$, für die verbliebenen Werte können wir eine Approximation der zweiten Ableitung verwenden.

Lemma 9.1 (Differenzenquotient) *Sei $h \in \mathbb{R}_{>0}$, und sei $g \in C^4[-h,h]$. Dann existiert ein $\eta \in [-h,h]$ mit*

$$\frac{g(h) - 2g(0) + g(-h)}{h^2} = g''(0) + \frac{h^2}{12} g^{(4)}(\eta).$$

Beweis. Mit dem Satz von Taylor finden wir $\eta_+ \in [0,h]$ und $\eta_- \in [-h,0]$ mit

$$g(h) = g(0) + hg'(0) + \frac{h^2}{2}g''(0) + \frac{h^3}{6}g'''(0) + \frac{h^4}{24}g^{(4)}(\eta_+),$$
$$g(-h) = g(0) - hg'(0) + \frac{h^2}{2}g''(0) - \frac{h^3}{6}g'''(0) + \frac{h^4}{24}g^{(4)}(\eta_-).$$

Indem wir beide Gleichungen addieren und den Zwischenwertsatz verwenden, finden wir ein $\eta \in [\eta_-, \eta_+] \subseteq [-h,h]$ mit

$$g(h) + g(-h) = 2g(0) + h^2 g''(0) + \frac{h^4}{12}\frac{g^{(4)}(\eta_+) + g^{(4)}(\eta_-)}{2}$$

$$= 2g(0) + h^2 g''(0) + \frac{h^4}{12} g^{(4)}(\eta).$$

Durch Umsortieren der Terme und Dividieren durch h^2 ergibt sich die gewünschte Gleichung. ∎

Wir untersuchen (9.2) in einem Punkt x_i mit $i \in [1:n]$. Indem wir Lemma 9.1 auf $g(t) = u_0(x_i + t)$ anwenden, ergibt sich

$$\begin{aligned}\omega^2 u_i = \omega^2 u_0(x_i) = -c u_0''(x_i) &= -c\frac{u_0(x_i+h) - 2u_0(x_i) + u_0(x_i-h)}{h^2} - c\frac{h^2}{12} u_0(\eta) \\ &= c\frac{2u_i - u_{i+1} - u_{i-1}}{h^2} - c\frac{h^2}{12} u_0^{(4)}(\eta)\end{aligned}$$

mit einem $\eta \in [x_i - h, x_i + h]$. Falls h klein genug ist, können wir den zweiten Term vernachlässigen und erhalten die Approximation

$$\omega^2 u_i \approx c\frac{2u_i - u_{i+1} - u_{i-1}}{h^2} \qquad \text{für alle } i \in [1:n],$$

die wir wegen $u_0 = 0 = u_{n+1}$ kompakt als

$$\omega^2 \mathbf{u} \approx \mathbf{L}\mathbf{u}$$

mit

$$\mathbf{u} := \begin{pmatrix} u_1 \\ u_2 \\ \vdots \\ u_n \end{pmatrix}, \qquad \mathbf{L} := \frac{c}{h^2}\begin{pmatrix} 2 & -1 & & \\ -1 & \ddots & \ddots & \\ & \ddots & \ddots & -1 \\ & & -1 & 2 \end{pmatrix} \tag{9.3}$$

schreiben können. Näherungen $\tilde{\omega} \in \mathbb{R}$ und $\widetilde{\mathbf{u}} \in \mathbb{R}^n$ lassen sich dann als Lösung des Systems

$$\tilde{\omega}^2 \widetilde{\mathbf{u}} = \mathbf{L}\widetilde{\mathbf{u}}$$

berechnen. Dabei handelt es sich um ein *Eigenwertproblem*.

Wir werden Eigenwertprobleme in der folgenden allgemeinen Form untersuchen:

> Sei $\mathbf{A} \in \mathbb{K}^{n\times n}$ eine Matrix. Wir suchen einen *Eigenwert* $\lambda \in \mathbb{K}$ und einen *Eigenvektor* $\mathbf{e} \in \mathbb{K}^{\mathcal{I}} \setminus \{\mathbf{0}\}$ mit
>
> $$\mathbf{A}\mathbf{e} = \lambda \mathbf{e}.$$

Diese Aufgabenstellung wird in der Praxis oft genauer spezifiziert, beispielsweise werden häufig nur bestimmte Paare von Eigenwerten und Eigenvektoren benötigt.

Auf den ersten Blick könnte man auf die Idee verfallen, das Eigenwertproblem in ein nichtlineares Gleichungssystem zu überführen, beispielsweise

$$f(\mathbf{e}, \lambda) := \begin{pmatrix} (\mathbf{A} - \lambda \mathbf{I})\mathbf{e} \\ \|\mathbf{e}\|^2 - 1 \end{pmatrix} \stackrel{!}{=} 0,$$

auf das sich das Newton-Verfahren (siehe Abschnitt 5.4) anwenden ließe. Dafür bräuchten wir allerdings die Jacobi-Matrix der Abbildung f, die durch

$$Df(\mathbf{x}, \mu) = \begin{pmatrix} \mathbf{A} - \mu\mathbf{I} & -\mathbf{x} \\ 2\mathbf{x}^* & 0 \end{pmatrix} \qquad \text{für alle } \mathbf{x} \in \mathbb{R}^n,\ \mu \in \mathbb{R}$$

gegeben ist. Falls μ ein Eigenwert ist, liegt ein entsprechender Eigenvektor im Kern der Matrix $\mathbf{A} - \mu\mathbf{I}$, die damit nicht invertierbar sein kann. Nach Satz 3.25 besitzt die Jacobi-Matrix in diesem Fall keine LR-Zerlegung, so dass speziellere Techniken zum Einsatz kommen müssen.

Ein sowohl theoretisch als auch praktisch erfolgreicher Ansatz besteht darin, Eigenwerte als Extremwerte des *Rayleigh-Quotienten* zu definieren.

Definition 9.2 (Rayleigh-Quotient) *Sei* $\mathbf{A} \in \mathbb{K}^{n\times n}$. *Die* Rayleigh-Quotientenabbildung *zu* $\mathbf{A}$ *ist gegeben durch*

$$\Lambda_A : \mathbb{K}^n \setminus \{\mathbf{0}\} \to \mathbb{K}, \qquad \mathbf{x} \mapsto \frac{\langle \mathbf{x}, \mathbf{A}\mathbf{x}\rangle_2}{\langle \mathbf{x}, \mathbf{x}\rangle_2}.$$

Falls $\mathbf{e} \in \mathbb{K}^n \setminus \{\mathbf{0}\}$ ein Eigenvektor zu einem Eigenwert $\lambda \in \mathbb{K}$ einer Matrix $\mathbf{A} \in \mathbb{K}^{n\times n}$ ist, gilt für den Rayleigh-Quotienten

$$\Lambda_A(\mathbf{e}) = \frac{\langle \mathbf{e}, \mathbf{A}\mathbf{e}\rangle_2}{\langle \mathbf{e}, \mathbf{e}\rangle_2} = \frac{\langle \mathbf{e}, \lambda\mathbf{e}\rangle_2}{\langle \mathbf{e}, \mathbf{e}\rangle_2} = \lambda\frac{\langle \mathbf{e}, \mathbf{e}\rangle_2}{\langle \mathbf{e}, \mathbf{e}\rangle_2} = \lambda, \tag{9.4}$$

wir können mit dem Rayleigh-Quotienten also zu einem Eigenvektor den passenden Eigenwert berechnen.

Für selbstadjungierte Matrizen lässt sich der Rayleigh-Quotient allerdings auch zur Charakterisierung der Eigenwerte einsetzen:

Satz 9.3 (Courant-Fischer) *Sei* $\mathbf{A} \in \mathbb{K}^{n\times n}$ *eine selbstadjungierte Matrix. Dann bildet die Rayleigh-Quotientenabbildung in die Menge* $\mathbb{R}$ *der reellen Zahlen ab und besitzt sowohl ein Minimum als auch ein Maximum. Das Minimum ist der kleinste, das Maximum der größte Eigenwert der Matrix* $\mathbf{A}$.

Beweis. Sei $\mathbf{x} \in \mathbb{K}^n \setminus \{\mathbf{0}\}$. Da $\mathbf{A}$ selbstadjungiert ist, gilt mit Lemma 3.44

$$\Lambda_A(\mathbf{x}) = \frac{\langle \mathbf{x}, \mathbf{A}\mathbf{x}\rangle_2}{\langle \mathbf{x}, \mathbf{x}\rangle_2} = \frac{\overline{\langle \mathbf{A}\mathbf{x}, \mathbf{x}\rangle_2}}{\langle \mathbf{x}, \mathbf{x}\rangle_2} = \frac{\overline{\langle \mathbf{x}, \mathbf{A}^*\mathbf{x}\rangle_2}}{\langle \mathbf{x}, \mathbf{x}\rangle_2} = \frac{\overline{\langle \mathbf{x}, \mathbf{A}\mathbf{x}\rangle_2}}{\langle \mathbf{x}, \mathbf{x}\rangle_2} = \overline{\Lambda_A(\mathbf{x})},$$

und es folgt $\Lambda_A(\mathbf{x}) \in \mathbb{R}$.

Wir stellen fest, dass für jedes $\alpha \in \mathbb{K}\setminus\{0\}$ und jeden Vektor $\mathbf{x} \in \mathbb{K}^n\setminus\{\mathbf{0}\}$ die Gleichung

$$\Lambda_A(\alpha\mathbf{x}) = \frac{\langle \alpha\mathbf{x}, \mathbf{A}\alpha\mathbf{x}\rangle}{\langle \alpha\mathbf{x}, \alpha\mathbf{x}\rangle} = \frac{|\alpha|^2\langle \mathbf{x}, \mathbf{A}\mathbf{x}\rangle}{|\alpha|^2\langle \mathbf{x}, \mathbf{x}\rangle} = \Lambda_A(\mathbf{x})$$

gilt. Demnach genügt es, Minimum und Maximum auf der Einheitssphäre

$$\mathcal{S} := \{\mathbf{x} \in \mathbb{K}^n \ :\ \|\mathbf{x}\|_2 = 1\}$$

zu suchen. Diese Menge ist nach dem Satz von Heine-Borel kompakt, also muss die stetige und reellwertige Abbildung Λ_A auf ihr ein Minimum und ein Maximum annehmen. Wir bezeichnen das Maximum mit $\lambda \in \mathbb{R}$ und fixieren einen Vektor $\mathbf{e} \in \mathcal{S}$ mit $\Lambda_A(\mathbf{e}) = \lambda$.

Sei $\mathbf{y} \in \mathbb{K}^n$, und sei $\alpha \in (0, 1/\|\mathbf{y}\|_2)$, mit der Konvention $1/0 = \infty$. Wir haben $\|\mathbf{e} + \alpha\mathbf{y}\|_2 \geq \|\mathbf{e}\|_2 - \alpha\|\mathbf{y}\|_2 > 0$, also insbesondere $\mathbf{e} + \alpha\mathbf{y} \neq \mathbf{0}$. Es gilt

$$\begin{aligned}
\lambda\|\mathbf{e}\|_2^2 + 2\lambda\alpha \operatorname{Re}\langle \mathbf{y}, \mathbf{e}\rangle_2 + \lambda\alpha^2\|\mathbf{y}\|_2^2 &= \lambda(\langle \mathbf{e}, \mathbf{e}\rangle_2 + \langle \alpha\mathbf{y}, \mathbf{e}\rangle_2 + \langle \mathbf{e}, \alpha\mathbf{y}\rangle_2 + \langle \alpha\mathbf{y}, \alpha\mathbf{y}\rangle_2) \\
&= \lambda\langle \mathbf{e} + \alpha\mathbf{y}, \mathbf{e} + \alpha\mathbf{y}\rangle_2 \geq \langle \mathbf{e} + \alpha\mathbf{y}, \mathbf{A}(\mathbf{e} + \alpha\mathbf{y})\rangle_2 \\
&= \langle \mathbf{e}, \mathbf{A}\mathbf{e}\rangle_2 + \langle \alpha\mathbf{y}, \mathbf{A}\mathbf{e}\rangle_2 + \langle \mathbf{e}, \alpha\mathbf{A}\mathbf{y}\rangle_2 + \langle \alpha\mathbf{y}, \alpha\mathbf{A}\mathbf{y}\rangle_2 \\
&= \langle \mathbf{e}, \mathbf{A}\mathbf{e}\rangle_2 + 2\alpha \operatorname{Re}\langle \mathbf{y}, \mathbf{A}\mathbf{e}\rangle_2 + \alpha^2\langle \mathbf{y}, \mathbf{A}\mathbf{y}\rangle_2 \\
&= \lambda\|\mathbf{e}\|_2^2 + 2\alpha \operatorname{Re}\langle \mathbf{y}, \mathbf{A}\mathbf{e}\rangle_2 + \alpha^2\langle \mathbf{y}, \mathbf{A}\mathbf{y}\rangle_2, \\
\alpha^2(\langle \mathbf{y}, \lambda\mathbf{y}\rangle_2 - \langle \mathbf{y}, \mathbf{A}\mathbf{y}\rangle_2) &\geq 2\alpha(\operatorname{Re}\langle \mathbf{y}, \mathbf{A}\mathbf{e}\rangle_2 - \operatorname{Re}\langle \mathbf{y}, \lambda\mathbf{e}\rangle_2), \\
\alpha\langle \mathbf{y}, \lambda\mathbf{y} - \mathbf{A}\mathbf{y}\rangle_2 &\geq 2\operatorname{Re}\langle \mathbf{y}, \mathbf{A}\mathbf{e} - \lambda\mathbf{e}\rangle_2.
\end{aligned}$$

Wir setzen $\mathbf{y} := \mathbf{A}\mathbf{e} - \lambda\mathbf{e}$ und erhalten

$$\alpha\langle \mathbf{y}, \lambda\mathbf{y} - \mathbf{A}\mathbf{y}\rangle_2 \geq 2\|\mathbf{A}\mathbf{e} - \lambda\mathbf{e}\|_2^2.$$

Da α beliebig klein gewählte werden kann, folgt $\|\mathbf{A}\mathbf{e} - \lambda\mathbf{e}\|_2 = 0$, also $\mathbf{A}\mathbf{e} = \lambda\mathbf{e}$.

Mit (9.4) folgt, dass kein Eigenwert größer als λ sein kann, also muss λ der maximale Eigenwert sein.

Den Nachweis, dass das Minimum der Abbildung Λ_A der kleinste Eigenwert ist, können wir analog führen. ∎

Satz 9.4 (Hauptachsentransformation) *Sei $\mathbf{A} \in \mathbb{K}^{n\times n}$ eine selbstadjungierte Matrix. Dann existieren eine unitäre Matrix $\mathbf{Q} \in \mathbb{K}^{n\times n}$ und eine reellwertige Diagonalmatrix $\mathbf{D} \in \mathbb{R}^{n\times n}$ mit $\mathbf{Q}^*\mathbf{X}\mathbf{Q} = \mathbf{D}$.*

Beweis. Wir führen den Beweis per Induktion über n.

Induktionsanfang: Für $n = 1$ ist jede Matrix auch eine Diagonalmatrix, also können wir $\mathbf{Q} = 1$ und $\mathbf{D} = \mathbf{A}$ setzen.

Induktionsvoraussetzung: Sei $n \in \mathbb{N}$ so gegeben, dass die Aussage für alle selbstadjungierten Matrizen $\mathbf{A} \in \mathbb{K}^{n\times n}$ gilt.

Induktionsschritt: Sei $\mathbf{A} \in \mathbb{R}^{(n+1)\times(n+1)}$. Nach Satz 9.3 finden wir einen Eigenwert $\lambda \in \mathbb{R}$ und einen passenden Eigenvektor $\mathbf{e} \in \mathbb{K}^{n+1}$ mit $\|\mathbf{e}\|_2 = 1$. Sei $\delta_1 \in \mathbb{K}^{n+1}$ der erste kanonische Einheitsvektor $\delta_1 = (1, 0, \ldots, 0)$, und sei $\mathbf{Q}_1 \in \mathbb{K}^{(n+1)\times(n+1)}$ die Householder-Transformation, die $\mathbf{Q}_1^*\mathbf{e} = \alpha\delta_1$ mit $|\alpha| = 1$ erfüllt. Es folgt

$$\mathbf{Q}_1^*\mathbf{A}\mathbf{Q}_1\delta_1 = \frac{1}{\alpha}\mathbf{Q}_1^*\mathbf{A}\mathbf{e} = \lambda\frac{1}{\alpha}\mathbf{Q}_1^*\mathbf{e} = \lambda\delta_1,$$

und da $\mathbf{Q}_1^*\mathbf{A}\mathbf{Q}_1$ selbstadjungiert ist, erhalten wir

$$\mathbf{Q}_1^*\mathbf{A}\mathbf{Q}_1 = \begin{pmatrix} \lambda & \\ & \widehat{\mathbf{A}} \end{pmatrix}$$

mit einer selbstadjungierten Matrix $\widehat{\mathbf{A}} \in \mathbb{R}^{n\times n}$. Nach Induktionsvoraussetzung existieren eine unitäre Matrix $\widehat{\mathbf{Q}} \in \mathbb{K}^{n\times n}$ und eine Diagonalmatrix $\widehat{\mathbf{D}} \in \mathbb{R}^{n\times n}$ mit $\widehat{\mathbf{Q}}^* \widehat{\mathbf{A}} \widehat{\mathbf{Q}} = \widehat{\mathbf{D}}$, so dass wir insgesamt

$$\begin{pmatrix} 1 & \\ & \widehat{\mathbf{Q}} \end{pmatrix}^* \mathbf{Q}_1^* \mathbf{X} \mathbf{Q}_1 \begin{pmatrix} 1 & \\ & \widehat{\mathbf{Q}} \end{pmatrix} = \begin{pmatrix} \lambda & \\ & \widehat{\mathbf{Q}}^* \widehat{\mathbf{X}} \widehat{\mathbf{Q}} \end{pmatrix} = \begin{pmatrix} \lambda & \\ & \widehat{\mathbf{D}} \end{pmatrix}$$

erhalten. Mit

$$\mathbf{Q} := \mathbf{Q}_1 \begin{pmatrix} 1 & \\ & \widehat{\mathbf{Q}} \end{pmatrix}, \qquad \mathbf{D} := \begin{pmatrix} \lambda & \\ & \widehat{\mathbf{D}} \end{pmatrix}$$

folgt daraus die Behauptung. ∎

Bemerkung 9.5 (Orthonormalbasis) *Wenn wir mit $\delta_i \in \mathbb{K}^n$ die durch*

$$\delta_{i,j} = \begin{cases} 1 & \text{falls } j = i, \\ 0 & \text{ansonsten} \end{cases} \qquad \text{für alle } i,j \in [1:n]$$

gegebenen kanonischen Einheitsvektoren bezeichnen, können wir die Spalten einer unitären Matrix $\mathbf{Q} \in \mathbb{K}^{n\times n}$ in der Form

$$\mathbf{q}_i := \mathbf{Q}\delta_i \qquad \text{für alle } i \in [1:n]$$

schreiben. Mit Lemma 3.61 folgt

$$\begin{aligned} \langle \mathbf{q}_i, \mathbf{q}_j \rangle_2 &= \langle \mathbf{Q}\delta_i, \mathbf{Q}\delta_j \rangle_2 = \langle \mathbf{Q}^*\mathbf{Q}\delta_i, \delta_j \rangle_2 \\ &= \langle \delta_i, \delta_j \rangle_2 = \begin{cases} 1 & \text{falls } i = j, \\ 0 & \text{ansonsten} \end{cases} \qquad \text{für alle } i,j \in [1:n], \end{aligned}$$

also ist $(\mathbf{q}_i)_{i=1}^n$ eine Orthonormalbasis des Raums $\mathbb{K}^n$.

Falls, wie im Satz 9.4, $\mathbf{A} = \mathbf{Q}\mathbf{D}\mathbf{Q}^*$ mit einer Diagonalmatrix $\mathbf{D}$ gilt, folgt für alle $i \in [1:n]$ mit $\mathbf{D}\delta_i = d_{ii}\delta_i$ auch

$$\mathbf{A}\mathbf{q}_i = \mathbf{Q}\mathbf{D}\mathbf{Q}^*\mathbf{Q}\delta_i = \mathbf{Q}\mathbf{D}\delta_i = \mathbf{Q}\delta_i d_{ii} = d_{ii}\mathbf{q}_i,$$

also ist $\mathbf{q}_i$ ein Eigenvektor der Matrix $\mathbf{A}$ zu dem Eigenwert d_{ii}.

Sei nun $\mathbf{e} \in \mathbb{K}^n \setminus \{\mathbf{0}\}$ ein Eigenvektor der Matrix $\mathbf{A}$ zu einem Eigenwert $\lambda \in \mathbb{K}$, und sei $\widehat{\mathbf{e}} := \mathbf{Q}^*\mathbf{e}$. Dann gilt

$$\mathbf{D}\widehat{\mathbf{e}} = \mathbf{Q}^*\mathbf{Q}\mathbf{D}\mathbf{Q}^*\mathbf{e} = \mathbf{Q}^*\mathbf{A}\mathbf{e} = \lambda\mathbf{Q}^*\mathbf{e} = \lambda\widehat{\mathbf{e}}.$$

Da $\widehat{\mathbf{e}} \neq \mathbf{0}$ gilt, finden wir ein $i \in [1:n]$ mit $\hat{e}_i \neq 0$, und aus der i-ten Zeile der obigen Gleichung folgt $d_{ii}\hat{e}_i = (\mathbf{D}\widehat{\mathbf{e}})_i = \lambda\hat{e}_i$, also wegen $\hat{e}_i \neq 0$ auch $\lambda = d_{ii}$. Somit ist jeder Eigenwert der Matrix $\mathbf{A}$ ein Diagonalelement der Matrix $\mathbf{D}$.

Um Satz 9.4 anwenden zu können, gehen wir im Folgenden davon aus, dass $\mathbf{A}$ selbstadjungiert ist, so dass wir eine unitäre Matrix $\mathbf{Q} \in \mathbb{K}^{n\times n}$ und eine reelle Diagonalmatrix

$$\mathbf{D} = \begin{pmatrix} \lambda_1 & & \\ & \ddots & \\ & & \lambda_n \end{pmatrix}$$

mit $\mathbf{A} = \mathbf{Q}\mathbf{D}\mathbf{Q}^*$ fixieren können. Indem wir die Spalten der Matrix $\mathbf{Q}$ und die Diagonalelemente der Matrix $\mathbf{D}$ geeignet permutieren, können wir immer sicherstellen, dass

$$|\lambda_1| \geq |\lambda_2| \geq \ldots \geq |\lambda_n| \tag{9.5}$$

gilt, dass die Eigenwerte also dem Betrag nach absteigend sortiert sind.

Wir wählen einen beliebigen Vektor $\mathbf{x} \in \mathbb{K}^n$ und untersuchen, was geschieht, wenn wir ihn m-mal mit der Matrix $\mathbf{A}$ multiplizieren. Da $\mathbf{Q}^*\mathbf{Q} = \mathbf{I}$ gilt, haben wir

$$\mathbf{y} := \mathbf{A}^m\mathbf{x} = (\mathbf{Q}\mathbf{D}\mathbf{Q}^*)^m\mathbf{x} = \mathbf{Q}\mathbf{D}^m\mathbf{Q}^*\mathbf{x}.$$

Mit den transformierten Vektoren

$$\widehat{\mathbf{x}} := \mathbf{Q}^*\mathbf{x}, \qquad\qquad \widehat{\mathbf{y}} := \mathbf{Q}^*\mathbf{y}$$

schreibt sich diese Gleichung kurz als

$$\widehat{\mathbf{y}} = \mathbf{D}^m\widehat{\mathbf{x}} = \begin{pmatrix} \lambda_1^m\hat{x}_1 \\ \vdots \\ \lambda_n^m\hat{x}_n \end{pmatrix}.$$

Wenn m groß ist, dürfen wir erwarten, dass wegen der Anordnung (9.5) die ersten Einträge des Vektors deutlich größer als die letzten sind, dass also in $\mathbf{y}$ Anteile in Richtung der Eigenvektoren zu den größten Eigenwerten dominieren. Diese Beobachtung ist die Grundlage der *Vektoriteration* (auch bekannt als *Von-Mises-Iteration* oder im Englischen als *power iteration*): Wir wählen einen Anfangsvektor $\mathbf{x}^{(0)} \in \mathbb{K}^n$ und definieren die Iterationsvektoren durch

$$\mathbf{x}^{(m+1)} := \mathbf{A}\mathbf{x}^{(m)} \qquad\qquad \text{für alle } m \in \mathbb{N}_0. \tag{9.6}$$

Da die Norm eines Eigenvektors keine Rolle spielt, bietet es sich an, für die Konvergenzuntersuchung derartiger Iterationsverfahren Größen zu verwenden, die von der Norm unabhängig sind.

Besonders elegante Resultate können wir mit Hilfe des bereits in (3.16) eingeführten Winkels zwischen Vektoren erhalten: Er ist durch

$$\cos\angle(\mathbf{x},\mathbf{y}) = \frac{|\langle\mathbf{x},\mathbf{y}\rangle_2|}{\|\mathbf{x}\|_2\|\mathbf{y}\|_2}, \qquad\qquad \text{für alle } \mathbf{x},\mathbf{y} \in \mathbb{K}^n \setminus \{\mathbf{0}\}$$

definiert, Sinus und Tangens ergeben sich gemäß

$$\begin{aligned}\sin\angle(\mathbf{x},\mathbf{y}) &= \sqrt{1-\cos^2\angle(\mathbf{x},\mathbf{y})},\\ \tan\angle(\mathbf{x},\mathbf{y}) &= \frac{\sin\angle(\mathbf{x},\mathbf{y})}{\cos\angle(\mathbf{x},\mathbf{y})} \qquad \text{für alle } \mathbf{x},\mathbf{y}\in\mathbb{K}^n\setminus\{\mathbf{0}\}.\end{aligned}$$

Je kleiner Sinus und Tangens sind, desto näher liegen die Richtungen der Vektoren $\mathbf{x}$ und $\mathbf{y}$ einander.

Sei nun $\mathbf{e}\in\mathbb{K}^n$ die erste Spalte der Matrix $\mathbf{Q}$, also ein Eigenvektor der Matrix $\mathbf{A}$ zu dem betragsgrößten Eigenwert λ_1. Für den Tangens des Winkels zwischen den Iterationsvektoren $\mathbf{x}^{(m)}$ der Vektoriteration und dem Eigenvektor $\mathbf{e}$ erhalten wir die folgende Konvergenzaussage:

Satz 9.6 (Konvergenz) *Sei $\mathbf{x}^{(0)}\in\mathbb{R}^n$ mit $\langle\mathbf{x}^{(0)},\mathbf{e}\rangle_2\neq 0$ gegeben. Dann gilt*

$$\tan\angle(\mathbf{x}^{(m)},\mathbf{e}) \leq \left(\frac{|\lambda_2|}{|\lambda_1|}\right)^m \tan\angle(\mathbf{x}^{(0)},\mathbf{e}) \qquad \textit{für alle } m\in\mathbb{N}_0.$$

Insbesondere konvergiert der Winkel gegen null, falls λ_1 ein dominanter Eigenwert *ist, also $|\lambda_1|>|\lambda_2|$ gilt.*

Beweis. Wir bezeichnen mit $\delta := (1,0,\ldots,0)$ den ersten kanonischen Einheitsvektor und halten fest, dass $\mathbf{Q}\delta=\mathbf{e}$ nach Definition gilt.

Wie zuvor ist es wesentlich einfacher, mit transformierten Vektoren zu arbeiten, an denen sich die Anteile der verschiedenen Eigenvektoren unmittelbar ablesen lassen. Deshalb definieren wir

$$\widehat{\mathbf{x}}^{(m)} := \mathbf{Q}^*\mathbf{x}^{(m)} \qquad \text{für alle } m\in\mathbb{N}_0$$

und halten fest, dass

$$\widehat{\mathbf{x}}^{(m+1)} = \mathbf{Q}^*\mathbf{x}^{(m+1)} = \mathbf{Q}^*\mathbf{A}\mathbf{x}^{(m)} = \mathbf{Q}^*\mathbf{A}\mathbf{Q}\widehat{\mathbf{x}}^{(m)} = \mathbf{D}\widehat{\mathbf{x}}^{(m)} \qquad \text{für alle } m\in\mathbb{N}_0$$

gilt. Mit einer einfachen Induktion folgt $\widehat{\mathbf{x}}^{(m)}=\mathbf{D}^m\widehat{\mathbf{x}}^{(0)}$ für alle $m\in\mathbb{N}_0$.

Gemäß Lemma 3.61 gilt

$$\begin{aligned}\cos^2\angle(\mathbf{x}^{(m)},\mathbf{e}) &= \frac{|\langle\mathbf{x}^{(m)},\mathbf{e}\rangle_2|^2}{\|\mathbf{x}^{(m)}\|_2^2\|\mathbf{e}\|_2^2} = \frac{|\langle\mathbf{Q}\widehat{\mathbf{x}}^{(m)},\mathbf{Q}\delta\rangle_2|^2}{\|\mathbf{Q}\widehat{\mathbf{x}}^{(m)}\|_2^2\|\mathbf{Q}\delta\|_2^2}\\ &= \frac{|\langle\widehat{\mathbf{x}}^{(m)},\delta\rangle_2|^2}{\|\widehat{\mathbf{x}}^{(m)}\|_2^2\|\delta\|_2^2} = \frac{|\hat{x}_1^{(m)}|^2}{\sum_{i=1}^n|\hat{x}_i^{(m)}|^2} \qquad \text{für alle } m\in\mathbb{N}_0,\end{aligned}$$

so dass wir

$$\sin^2\angle(\mathbf{x}^{(m)},\mathbf{e}) = 1-\cos^2\angle(\mathbf{x}^{(m)},\mathbf{e}) = \frac{\sum_{i=1}^n|\hat{x}_i^{(m)}|^2-|\hat{x}_1^{(m)}|^2}{\sum_{i=1}^n|\hat{x}_i^{(m)}|^2} = \frac{\sum_{i=2}^n|\hat{x}_i^{(m)}|^2}{\sum_{i=1}^n|\hat{x}_i^{(m)}|^2},$$

$$\tan^2 \angle(\mathbf{x}^{(m)}, \mathbf{e}) = \frac{\sin^2 \angle(\mathbf{x}^{(m)}, \mathbf{e})}{\cos^2 \angle(\mathbf{x}^{(m)}, \mathbf{e})} = \frac{\sum_{i=2}^n |\hat{x}_i^{(m)}|^2}{|\hat{x}_1^{(m)}|^2}$$

erhalten. Sei nun $m \in \mathbb{N}_0$ fixiert. Für das Quadrat des Tangens erhalten wir

$$\begin{aligned}\tan^2 \angle(\mathbf{x}^{(m)}, \mathbf{e}) &= \frac{\sum_{i=2}^n |\hat{x}_i^{(m)}|^2}{|\hat{x}_1^{(m)}|^2} = \frac{\sum_{i=2}^n |\lambda_i^m \hat{x}_i^{(0)}|^2}{|\lambda_1^m \hat{x}_1^{(0)}|^2} \leq \frac{\sum_{i=2}^n |\lambda_2|^{2m} |\hat{x}_i^{(0)}|^2}{|\lambda_1|^{2m} |\hat{x}_1^{(0)}|^2} \\ &= \left(\frac{|\lambda_2|}{|\lambda_1|}\right)^{2m} \frac{\sum_{i=2}^n |\hat{x}_i^{(0)}|^2}{|\hat{x}_1^{(0)}|^2} = \left(\frac{|\lambda_2|}{|\lambda_1|}\right)^{2m} \tan^2 \angle(\mathbf{x}^{(0)}, \mathbf{e}),\end{aligned}$$

so dass die gewünschte Ungleichung folgt, indem wir die Wurzel ziehen. ∎

In unserem Fall ist es wenig sinnvoll, den *größten* Eigenwert zu approximieren, da er auch mit dem größten Diskretisierungsfehler verbunden ist. Stattdessen sind wir an dem *kleinsten* Eigenwert interessiert, der der kleinsten Eigenfrequenz entspricht, der *Grundschwingung* der Saite, die im Wesentlichen die Höhe ihres Tons bestimmt.

Dazu modifizieren wir das Verfahren geringfügig: Der betragskleinste Eigenwert der Matrix $\mathbf{A}$ ist der betragsgrößte Eigenwert ihrer Inversen $\mathbf{A}^{-1}$, also verwenden wir die *inverse Iteration*

$$\mathbf{x}^{(m+1)} := \mathbf{A}^{-1}\mathbf{x}^{(m)} \qquad \text{für alle } m \in \mathbb{N}_0.$$

Wir können noch einen Schritt weiter gehen: Wir wählen ein $\mu \in \mathbb{K}$ und stellen fest, dass der betragskleinste Eigenwert der Matrix $\mathbf{A} - \mu\mathbf{I}$ derjenige ist, der μ am nächsten liegt. Sein Kehrwert ist der betragsgrößte Eigenwert der Inversen $(\mathbf{A}-\mu\mathbf{I})^{-1}$. Die resultierende Variante

$$\mathbf{x}^{(m+1)} := (\mathbf{A} - \mu\mathbf{I})^{-1}\mathbf{x}^{(m)} \qquad \text{für alle } m \in \mathbb{N}_0$$

der Vektoriteration bezeichnet man als *inverse Iteration mit Shift*, μ wird als *Shift-Parameter* bezeichnet. Im Englischen bedeutet *to shift*, etwas zu verschieben, und bei dieser Iteration werden die Eigenwerte um μ „verschoben“.

Die Aussage des Satzes 9.6 überträgt sich direkt:

Folgerung 9.7 (Konvergenz der inversen Iteration) *Sei $\mu \in \mathbb{K}$, und seien die Eigenwerte gemäß*

$$|\lambda_1 - \mu| \leq |\lambda_2 - \mu| \leq \ldots \leq |\lambda_n - \mu| \tag{9.7}$$

angeordnet. Sei $\mathbf{e}$ wieder die erste Spalte der Matrix $\mathbf{Q}$, also ein normierter Eigenvektor der Matrix $\mathbf{A}$ zu dem Eigenwert λ_1.

Sei $\mathbf{x}^{(0)} \in \mathbb{R}^n$ mit $\langle \mathbf{x}^{(0)}, \mathbf{e}\rangle_2 \neq 0$ gegeben. Dann gilt

$$\tan \angle(\mathbf{x}^{(m)}, \mathbf{e}) \leq \left(\frac{|\lambda_1 - \mu|}{|\lambda_2 - \mu|}\right)^m \tan \angle(\mathbf{x}^{(0)}, \mathbf{e}) \qquad \textit{für alle } m \in \mathbb{N}_0.$$

Beweis. Sei $\widehat{\mathbf{A}} := (\mathbf{A} - \mu\mathbf{I})^{-1}$. Es gilt

$$\mathbf{A} - \mu\mathbf{I} = \mathbf{Q}\mathbf{D}\mathbf{Q}^* - \mu\mathbf{I} = \mathbf{Q}(\mathbf{D} - \mu\mathbf{I})\mathbf{Q}^*,$$
$$\widehat{\mathbf{A}} = (\mathbf{A} - \mu\mathbf{I})^{-1} = \mathbf{Q}(\mathbf{D} - \mu\mathbf{I})^{-1}\mathbf{Q}^*,$$

also erhalten wir $\widehat{\mathbf{A}} = \mathbf{Q}\widehat{\mathbf{D}}\mathbf{Q}^*$ mit

$$\widehat{\mathbf{D}} := \begin{pmatrix} \hat{\lambda}_1 & & \\ & \ddots & \\ & & \hat{\lambda}_n \end{pmatrix}, \qquad \hat{\lambda}_i := \frac{1}{\lambda_i - \mu} \qquad \text{für alle } i \in [1:n].$$

Aufgrund der Voraussetzung (9.7) gilt

$$|\hat{\lambda}_1| \geq |\hat{\lambda}_2| \geq \ldots \geq |\hat{\lambda}_n|,$$

so dass wir Satz 9.6 auf die Matrix $\widehat{\mathbf{A}}$ anwenden können, um

$$\begin{aligned} \tan\angle(\mathbf{x}^{(m)}, \mathbf{e}) &\leq \left(\frac{|\hat{\lambda}_2|}{|\hat{\lambda}_1|}\right)^m \tan\angle(\mathbf{x}^{(0)}, \mathbf{e}) \\ &= \left(\frac{|\lambda_1 - \mu|}{|\lambda_2 - \mu|}\right)^m \tan\angle(\mathbf{x}^{(0)}, \mathbf{e}) \qquad \text{für alle } m \in \mathbb{N}_0 \end{aligned}$$

zu erhalten. ∎

Um beurteilen zu können, wie gut unsere Näherung des Eigenvektors bereits ist, wäre es hilfreich, auch den Eigenwert zu kennen. Diese Aufgabe haben wir bereits gelöst: Der Rayleigh-Quotient $\Lambda_A(\mathbf{x}^{(m+1)})$ verspricht, eine gute Näherung des Eigenwerts zu sein.

Lemma 9.8 (Rayleigh-Quotient) *Sei $\mathbf{e} \in \mathbb{K}^n \setminus \{\mathbf{0}\}$ ein Eigenvektor der Matrix $\mathbf{A}$ zu dem Eigenwert λ. Dann gilt*

$$|\lambda - \Lambda_A(\mathbf{x})| \leq \|\lambda_1\mathbf{I} - \mathbf{A}\|_2 \sin^2\angle(\mathbf{x}, \mathbf{e}) \qquad \textit{für alle } \mathbf{x} \in \mathbb{R}^n \setminus \{\mathbf{0}\}.$$

Beweis. Sei $\mathbf{x} \in \mathbb{K}^n \setminus \{\mathbf{0}\}$, und sei $\alpha \in \mathbb{K}$.

Da $(\lambda\mathbf{I} - \mathbf{A}) = (\lambda\mathbf{I} - \mathbf{A})^*$ und $(\lambda\mathbf{I} - \mathbf{A})\mathbf{e} = \mathbf{0}$ gelten, erhalten wir mit der Cauchy-Schwarz-Ungleichung (3.14) und der Verträglichkeit (3.2a) die Abschätzung

$$\begin{aligned} |\lambda - \Lambda(\mathbf{x})| &= \left|\frac{\lambda\langle\mathbf{x},\mathbf{x}\rangle_2}{\langle\mathbf{x},\mathbf{x}\rangle_2} - \frac{\langle\mathbf{x},\mathbf{A}\mathbf{x}\rangle_2}{\langle\mathbf{x},\mathbf{x}\rangle_2}\right| = \left|\frac{\langle\mathbf{x},(\lambda\mathbf{I}-\mathbf{A})\mathbf{x}\rangle_2}{\langle\mathbf{x},\mathbf{x}\rangle_2}\right| \\ &= \left|\frac{\langle\mathbf{x},(\lambda\mathbf{I}-\mathbf{A})(\mathbf{x}-\alpha\mathbf{e}),\rangle_2}{\langle\mathbf{x},\mathbf{x}\rangle_2}\right| = \left|\frac{\langle(\lambda\mathbf{I}-\mathbf{A})\mathbf{x},\mathbf{x}-\alpha\mathbf{e}\rangle_2}{\langle\mathbf{x},\mathbf{x}\rangle_2}\right| \\ &= \left|\frac{\langle(\lambda\mathbf{I}-\mathbf{A})(\mathbf{x}-\alpha\mathbf{e}),\mathbf{x}-\alpha\mathbf{e}\rangle_2}{\langle\mathbf{x},\mathbf{x}\rangle_2}\right| \\ &\leq \frac{\|(\lambda_1\mathbf{I}-\mathbf{A})(\mathbf{x}-\alpha\mathbf{e})\|_2\|\mathbf{x}-\alpha\mathbf{e}\|_2}{\|\mathbf{x}\|_2^2} \leq \|\lambda_1\mathbf{I}-\mathbf{A}\|_2\frac{\|\mathbf{x}-\alpha\mathbf{e}\|_2^2}{\|\mathbf{x}\|_2^2}. \end{aligned}$$

Nun wählen wir $\alpha := \langle \mathbf{e}, \mathbf{x} \rangle_2 / \|\mathbf{e}\|_2^2$ und stellen fest, dass wegen die Gleichung $\langle \alpha\mathbf{e}, \mathbf{x} \rangle_2 = |\langle \mathbf{e}, \mathbf{x} \rangle_2|^2 / \|\mathbf{e}\|_2^2 = \|\mathbf{e}\|_2^2 |\langle \mathbf{x}, \mathbf{e} \rangle_2|^2 / \|\mathbf{e}\|_2^4 = \|\mathbf{e}\|_2^2 |\alpha|^2 = \|\alpha\mathbf{e}\|_2^2$ die Winkelfunktionen durch

$$\begin{aligned}
\cos^2 \angle(\mathbf{x}, \mathbf{e}) &= \frac{|\alpha|^2 \|\mathbf{e}\|_2^2}{\|\mathbf{x}\|_2^2}, \\
\sin^2 \angle(\mathbf{x}, \mathbf{e}) &= 1 - \cos^2 \angle(\mathbf{x}, \mathbf{e}) = \frac{\|\mathbf{x}\|_2^2 - |\alpha|^2 \|\mathbf{e}\|_2^2}{\|\mathbf{x}\|_2^2} = \frac{\|\mathbf{x}\|_2^2 - \|\alpha\mathbf{e}\|_2^2}{\|\mathbf{x}\|_2^2} \\
&= \frac{\|\mathbf{x}\|_2^2 - \langle \alpha\mathbf{e}, \mathbf{x} \rangle_2 - \langle \mathbf{x}, \alpha\mathbf{e} \rangle_2 + \|\alpha\mathbf{e}\|_2^2}{\|\mathbf{x}\|_2^2} = \frac{\|\mathbf{x} - \alpha\mathbf{e}\|_2^2}{\|\mathbf{x}\|_2^2}
\end{aligned}$$

gegeben sind. Damit folgt die gewünschte Abschätzung. ∎

Mit Hilfe des Rayleigh-Quotienten können wir in jedem Schritt unseres Verfahrens prüfen, ob die Norm des *Residuums*

$$\mathbf{r}^{(m)} := \mathbf{A}\mathbf{x}^{(m)} - \Lambda_A(\mathbf{x}^{(m)})\mathbf{x}^{(m)}$$

klein genug ist.

Übungsaufgabe 9.9 (Gestörte Matrix) *Aussagen über den Approximationsfehler der Iteration können wir mit einer* Rückwärtsanalyse *erhalten: Statt den Fehler direkt auszurechnen, stören wir* $\mathbf{A}$ *so, dass* $\mathbf{x}^{(m)}$ *ein exakter Eigenvektor der gestörten Matrix ist. Ergebnisse wie Satz 9.10 können dann benutzt werden, um Fehlerabschätzungen zu gewinnen.*

Für beliebige Vektoren $\mathbf{a}, \mathbf{b} \in \mathbb{K}^n$ *mit* $\mathbf{a}^*\mathbf{b} = \langle \mathbf{a}, \mathbf{b} \rangle_2 = 0$ *beweise man, dass die Matrix*

$$\mathbf{E} := \mathbf{a}\mathbf{b}^* + \mathbf{b}\mathbf{a}^*$$

die Gleichung $\|\mathbf{E}\|_2 = \|\mathbf{a}\|_2 \|\mathbf{b}\|_2$ *erfüllt.*

Indem man diese Aussage auf $\mathbf{x}^{(m)}$ *und* $\mathbf{r}^{(m)}$ *anwendet, beweise man, dass eine selbstadjungierte Matrix* $\widetilde{\mathbf{A}} \in \mathbb{K}^{n \times n}$ *mit*

$$\widetilde{\mathbf{A}}\mathbf{x}^{(m)} = \Lambda_A(\mathbf{x}^{(m)})\mathbf{x}^{(m)}, \qquad \|\mathbf{A} - \widetilde{\mathbf{A}}\|_2 \leq \|\mathbf{r}^{(m)}\|_2 / \|\mathbf{x}^{(m)}\|_2$$

existiert, dass unsere Näherung $\mathbf{x}^{(m)}$ *also ein* exakter *Eigenvektor der gestörten Matrix* $\widetilde{\mathbf{A}}$ *zu dem Eigenwert* $\Lambda_A(\mathbf{x}^{(m)})$ *ist.*

Falls das Residuum $\mathbf{r}^{(m)}$ klein ist, können wir mit Störungssätzen wie dem folgenden daraus Rückschlüsse auf die Genauigkeit der Eigenwertapproximation gezogen werden.

Satz 9.10 (Bauer-Fike) *Seien* $\mathbf{A}, \widetilde{\mathbf{A}} \in \mathbb{K}^{\mathcal{I} \times \mathcal{I}}$ *selbstadjungierte Matrizen. Zu jedem Eigenwert* $\tilde{\lambda}$ *der Matrix* $\widetilde{\mathbf{A}}$ *existiert ein Eigenwert* λ *der Matrix* $\mathbf{A}$ *mit*

$$|\lambda - \tilde{\lambda}| \leq \|\mathbf{A} - \widetilde{\mathbf{A}}\|_2.$$

Beweis. Sei $\tilde{\lambda}$ ein Eigenwert der Matrix $\widetilde{\mathbf{A}}$. Sei λ ein Eigenwert der Matrix $\mathbf{A}$, der $\tilde{\lambda}$ am nächsten liegt, der also

$$|\lambda - \tilde{\lambda}| \leq |\mu - \tilde{\lambda}| \qquad \text{für alle Eigenwerte } \mu \text{ von } \mathbf{A}$$

erfüllt. Da λ und $\tilde{\lambda}$ reell sind, ist das äquivalent dazu, dass $(\lambda - \tilde{\lambda})^2$ der minimale Eigenwert der Matrix $\mathbf{B} := (\mathbf{A} - \tilde{\lambda}\mathbf{I})^2$ ist.

Nach dem Satz 9.3 von Courant und Fischer ist $(\lambda - \tilde{\lambda})^2$ das Minimum des Rayleigh-Quotienten der Matrix $\mathbf{B}$, und mit Hilfe des Lemmas 3.44 erhalten wir für jeden Vektor $\mathbf{y} \in \mathbb{K}^n$ mit $\|\mathbf{y}\|_2 = 1$ die Abschätzung

$$\begin{aligned}
(\lambda - \tilde{\lambda})^2 &= \min\{\Lambda_B(\mathbf{x}) \ : \ \mathbf{x} \in \mathbb{K}^n \setminus \{\mathbf{0}\}\} \\
&\leq \Lambda_B(\mathbf{y}) = \langle \mathbf{y}, \mathbf{B}\mathbf{y}\rangle_2 = \langle \mathbf{y}, (\mathbf{A} - \tilde{\lambda}\mathbf{I})^2\mathbf{y}\rangle_2 \\
&= \langle (\mathbf{A} - \tilde{\lambda}\mathbf{I})\mathbf{y}, (\mathbf{A} - \tilde{\lambda}\mathbf{I})\mathbf{y}\rangle_2 = \|(\mathbf{A} - \tilde{\lambda}\mathbf{I})\mathbf{y}\|_2^2 \\
&= \|(\mathbf{A} - \widetilde{\mathbf{A}} + \widetilde{\mathbf{A}} - \tilde{\lambda}\mathbf{I})\mathbf{y}\|_2^2 \\
&= \|(\mathbf{A} - \widetilde{\mathbf{A}})\mathbf{y} + (\widetilde{\mathbf{A}} - \tilde{\lambda}\mathbf{I})\mathbf{y}\|_2^2.
\end{aligned}$$

Um den zweiten Term verschwinden zu lassen, wählen wir $\mathbf{y}$ nun als Eigenvektor der Matrix $\widetilde{\mathbf{A}}$ (weiterhin mit $\|\mathbf{y}\|_2 = 1$) zu dem Eigenwert $\tilde{\lambda}$, so dass wir

$$(\lambda - \tilde{\lambda})^2 \leq \|(\mathbf{A} - \widetilde{\mathbf{A}})\mathbf{y}\|_2^2 \leq \|\mathbf{A} - \widetilde{\mathbf{A}}\|_2^2 \|\mathbf{y}\|_2^2 = \|\mathbf{A} - \widetilde{\mathbf{A}}\|_2^2$$

finden und zu der gewünschten Abschätzung gelangen, indem wir die Wurzel ziehen. ∎

Nun können wir eine praxistaugliche Version der inversen Iteration mit Shift konstruieren:

- Im Interesse der Effizienz und Stabilität berechnen wir die Inverse $(\mathbf{A} - \mu\mathbf{I})^{-1}$ nicht explizit, sondern werten sie für einen Vektor $\mathbf{x}^{(m)}$ aus, indem wir das lineare Gleichungssystem $(\mathbf{A} - \mu\mathbf{I})\mathbf{w}^{(m)} = \mathbf{x}^{(m)}$ lösen.
- Die dabei auftretenden Hilfsvektoren $\mathbf{w}^{(m)}$ verwenden wir, um den Rayleigh-Quotienten $\hat{\lambda}$ für $\widehat{\mathbf{A}}$ effizient auszuwerten. Der Rayleigh-Quotient für $\mathbf{A}$ kann daraus wegen $\hat{\lambda} = 1/(\lambda - \mu)$ als $\lambda = 1/\hat{\lambda} + \mu$ rekonstruiert werden.
- Indem wir in jedem Schritt $\mathbf{x}^{(m)}$ so skalieren, dass ein Einheitsvektor entsteht, verhindern wir, dass die Koeffizienten den (endlichen) Bereich der Maschinenzahlen verlassen.
- Das Residuum verwenden wir für das Abbruchkriterium.

Der vollständige Algorithmus ist in Abbildung 9.1 zusammengefasst.

Bemerkung 9.11 (Rayleigh-Iteration) *Sofern uns bereits eine gute Näherung des Eigenvektors vorliegt, können wir die inverse Iteration erheblich beschleunigen, indem wir den bisher fixierten Shift-Parameter μ in jedem Schritt durch den jeweils aktuellen Rayleigh-Quotienten ersetzen (also $\mu = \lambda$ in Abbildung 9.1 verwenden). In diesem Fall verändert sich zwar das zu lösende Gleichungssystem von Schritt zu Schritt, aber dafür erhalten wir unter geeigneten Bedingungen ein* kubisch *konvergentes Verfahren.*

procedure invit($\mathbf{A}$, μ, ϵ, **var** λ, $\mathbf{x}$);
Bereite das Lösen der Gleichungssysteme vor (z.B. LR- oder QR-Zerlegung)
$\mathbf{x} \leftarrow \mathbf{x}/\|\mathbf{x}\|_2$
repeat
  Löse $(\mathbf{A} - \mu\mathbf{I})\mathbf{w} = \mathbf{x}$
  $\hat{\lambda} \leftarrow \langle \mathbf{x}, \mathbf{w} \rangle_2$
  $\lambda \leftarrow 1/\hat{\lambda} + \mu$
  $\mathbf{x} \leftarrow \mathbf{w}/\|\mathbf{w}\|_2$
until $\|\mathbf{A}\mathbf{x} - \lambda\mathbf{x}\|_2 \leq \epsilon$

Abbildung 9.1: Inverse Iteration mit Shift $\mu \in \mathbb{K}$. Die Iterationsvektoren stehen in $\mathbf{x}$, die zugehörigen Eigenwertapproximationen in λ.

9.2 Mechanik

Ein einfaches Beispiel für eine gewöhnliche Differentialgleichung ergibt sich aus der Mechanik: Eine Kugel rollt auf einer gekrümmten Bahn. Zur Vereinfachung ersetzen wir die Kugel durch einen Punkt einer gewissen Masse m, der sich unter dem Einfluss der Gravitation bewegt.

Mathematisch stellen wir die Bahn durch eine Abbildung

$$\gamma : \mathbb{R} \to \mathbb{R}^2$$

dar, die eine Kurve im zweidimensionalen Raum beschreibt. Da die Kugel die Bahn nicht verlassen soll, können wir ihre Position zu einem Zeitpunkt t durch

$$x(t) = \gamma(z(t)) \qquad \text{für alle } t \in \mathbb{R}$$

darstellen, wobei

$$z : \mathbb{R} \to \mathbb{R}$$

jedem Zeitpunkt einen Punkt im Parameterbereich der Kurve zuordnet.

Nach dem zweiten Newton'schen Gesetz entsteht eine Beschleunigung der Kugel durch einwirkende Kräfte. In unserem Fall wirkt die *Gravitationskraft*

$$G = m \begin{pmatrix} 0 \\ -g \end{pmatrix}$$

in Richtung Erdboden. Hinzu kommt eine *Rückstellkraft* $R(t)$, die verhindert, dass die Kugel die Bahn verlässt. Diese Kraft wirkt senkrecht zu der Tangentialrichtung der Bahn im Punkt $x(t)$, erfüllt also

$$\langle \gamma'(z(t)), R(t) \rangle_2 = 0 \qquad \text{für alle } t \in \mathbb{R}.$$

Die Beschleunigung zu einem Zeitpunkt t ist gerade die zweite Ableitung $x''(t)$, und das zweite Newton'sche Gesetz führt zu der Gleichung

$$mx''(t) = G + R(t) \qquad \text{für alle } t \in \mathbb{R}. \tag{9.8}$$

Unser Ziel ist es, diese Differentialgleichung so umzuformen, dass eine Gleichung entsteht, mit deren Hilfe wir z bestimmen können. Per Ketten- und Produktregel erhalten wir zunächst

$$\begin{aligned} x'(t) &= \gamma'(z(t))z'(t), \\ x''(t) &= \gamma''(z(t))(z'(t))^2 + \gamma'(z(t))z''(t) \end{aligned} \qquad \text{für alle } t \in \mathbb{R}.$$

Durch Einsetzen in (9.8) folgt

$$m(\gamma''(z(t))(z'(t))^2 + \gamma'(z(t))z''(t)) = G + R(t) \qquad \text{für alle } t \in \mathbb{R}.$$

Um $z''(t)$ zu bestimmen, müssten wir „durch $\gamma'(z(t))$ dividieren“. Das können wir nicht, da es sich um einen Vektor handelt. Wir können allerdings die Gleichung im Skalarprodukt mit diesem Vektor multiplizieren und so eine skalare Gleichung erhalten:

$$\begin{aligned} &m\langle\gamma'(z(t)), \gamma''(z(t))\rangle_2(z'(t))^2 \\ &+ m\langle\gamma'(z(t)), \gamma'(z(t))\rangle_2 z''(t) = \langle\gamma'(z(t)), G + R(t)\rangle_2 \end{aligned} \qquad \text{für alle } t \in \mathbb{R}.$$

Das Skalarprodukt eines Vektors mit sich selbst ist gerade das Quadrat seiner Norm, so dass sich der Faktor vor $z''(t)$ vereinfachen lässt. Da die Rückstellkraft $R(t)$ nur senkrecht zur Bahn wirkt, ist ihr Skalarprodukt mit $\gamma'(z(t))$ gleich null, und die Gleichung nimmt die Form

$$m\langle\gamma'(z(t)), \gamma''(z(t))\rangle_2(z'(t))^2 + m\|\gamma'(z(t))\|_2^2 z''(t) = \langle\gamma'(z(t)), G\rangle_2 \qquad \text{für alle } t \in \mathbb{R}$$

an, die wir nach $z''(t)$ auflösen können:

$$z''(t) = \frac{\langle\gamma'(z(t)), G - m\gamma''(z(t))(z'(t))^2\rangle_2}{m\|\gamma'(z(t))\|_2^2} \qquad \text{für alle } t \in \mathbb{R}.$$

Das ist schon fast eine gewöhnliche Differentialgleichung, allerdings steht noch die *zweite* Ableitung statt der ersten auf der linken Seite. Dieses Problem lösen wir, indem wir eine zweite Funktion

$$v(t) = z'(t) \qquad \text{für alle } t \in \mathbb{R}$$

einführen und die Gleichung in die Form eines Systems in z und v bringen:

$$\begin{aligned} z'(t) &= v(t), \\ v'(t) &= z''(t) = \frac{\langle\gamma'(z(t)), G - m\gamma''(z(t))(v(t))^2\rangle_2}{m\|\gamma'(z(t))\|_2^2} \end{aligned} \qquad \text{für alle } t \in \mathbb{R}.$$

Um die in (8.1) gegebene Gestalt eines Anfangswertproblems zu erhalten, definieren wir

$$y(t) = \begin{pmatrix} z(t) \\ v(t) \end{pmatrix}, \qquad f(t,a) = \begin{pmatrix} a_2 \\ \frac{\langle\gamma'(a_1), G - m\gamma''(a_1)a_2^2\rangle_2}{m\|\gamma'(a_1)\|_2^2} \end{pmatrix} \qquad \text{für alle } t \in \mathbb{R},\ a \in \mathbb{R}^2,$$

denn dann folgt

$$y'(t) = \begin{pmatrix} z'(t) \\ v'(t) \end{pmatrix} = f(t, y(t)) \qquad \text{für alle } t \in \mathbb{R}.$$

Wir können zu einem beliebigen Zeitpunkt eine Anfangsposition und eine Anfangsgeschwindigkeit vorgeben und mit den in Kapitel 8 beschriebenen Zeitschrittverfahren die Lösung approximieren, also die Bewegung der Kugel auf der Bahn simulieren.

Denkanstoß 9.12 (Spline-Kurve) *Damit f wohldefiniert ist, muss γ zweimal stetig differenzierbar sein (und γ' darf niemals null werden).*

Wie könnte man vorgehen, um γ als kubische Spline-Funktion, darzustellen?

Wie schwierig wäre es, ein Programm zu schreiben, bei dem der Anwender die Bahn durch Wahl der Koeffizienten der B-Spline-Darstellung frei definieren kann, und das dann die Bewegung der Kugel grafisch darstellt?

Denkanstoß 9.13 (Oberfläche) *Was passiert, wenn man die durch γ beschriebene Kurve durch eine durch*

$$\gamma : \mathbb{R}^2 \to \mathbb{R}^3$$

beschriebene Fläche ersetzt? Welche Eigenschaft muss die Rückstellkraft besitzen? Welche Rolle spielen die Tangentialvektoren $\partial_1\gamma$ und $\partial_2\gamma$?

9.3 Wärmeleitung

Wir wenden uns einem einfachen Beispiel für eine parabolische Differentialgleichung zu: Die *Wärmeleitungsgleichung*

$$\frac{\partial u}{\partial t}(x,t) = \frac{\partial^2 u}{\partial x^2}(x,t) \qquad \text{für alle } x \in [0,1],\ t \in \mathbb{R} \tag{9.9}$$

beschreibt die Ausbreitung von Wärme in einem Draht der Länge eins: $u(x,t)$ ist die Temperatur in dem Punkt x zu dem Zeitpunkt t. Wir gehen davon aus, dass die Temperatur am linken und rechten Randpunkt konstant gehalten wird, der Einfachheit halber auf Null-Temperatur:

$$u(0,t) = u(1,t) = 0 \qquad \text{für alle } t \in \mathbb{R}.$$

Da die Gleichung im Prinzip bereits die Gestalt einer gewöhnlichen Differentialgleichung aufweist, leider mit einem Differentialoperator auf der rechten Seite, bietet es sich an, nach einer Möglichkeit zu suchen, um sie in eine Form zu bringen, die sich mit den uns bereits bekannten Techniken behandeln lässt.

Dazu verwenden wir dieselbe Diskretisierung wie zuvor: Wir wählen $n \in \mathbb{N}$ und ersetzen das Intervall $[0,1]$ durch äquidistant verteilte Punkte

$$x_i = ih, \qquad h := \frac{1}{n+1} \qquad \text{für alle } i \in \{0, \dots, n+1\}.$$

Wir wollen die Lösung u in diesen Punkten berechnen, sind also an den Funktionen

$$u_i(t) = u(x_i, t) \qquad \text{für alle } i \in \{0, \dots, n+1\}$$

interessiert. Aus den Randbedingungen folgen die Gleichungen $u_0(t) = 0$ und $u_{n+1}(t) = 0$ für alle $t \in \mathbb{R}$, für die verbliebenen Funktionen können wir die in Lemma 9.1 eingeführte Näherung verwenden. Indem wir die einzelnen Funktionen zu einer vektorwertigen Funktion

$$\mathbf{y}(t) = \begin{pmatrix} u_1(t) \\ \vdots \\ u_n(t) \end{pmatrix} = \begin{pmatrix} u(x_1, t) \\ \vdots \\ u(x_n, t) \end{pmatrix} \qquad \text{für alle } t \in \mathbb{R}$$

zusammenfassen, erhalten wir mit der Matrix $\mathbf{L}$ aus (9.3) die Approximation

$$-\mathbf{L}\mathbf{y}(t) \approx \begin{pmatrix} \frac{\partial^2 u}{\partial x^2}(x_1, t) \\ \vdots \\ \frac{\partial^2 u}{\partial x^2}(x_n, t) \end{pmatrix} \qquad \text{für alle } t \in \mathbb{R}.$$

Wir setzen in (9.9) ein und erhalten

$$\mathbf{y}'(t) \approx -\mathbf{L}\mathbf{y}(t) \qquad \text{für alle } t \in \mathbb{R}.$$

Indem wir die Verfahren des Kapitels 8 auf das System

$$\widetilde{\mathbf{y}}'(t) = -\mathbf{L}\widetilde{\mathbf{y}}(t) \qquad \text{für alle } t \in \mathbb{R}$$

anwenden, können wir Näherungen der Lösung u in den Punkten $x_1, \dots, x_n$ berechnen.

Da die Matrix $\mathbf{L}$ positiv definit ist, besitzt die Matrix $-\mathbf{L}$ ausschließlich negative Eigenwerte. Eine genauere Analyse zeigt, dass der kleinste Eigenwert der Matrix $\mathbf{L}$ für $h \to 0$ gegen π^2 konvergiert, während die größten wie h^{-2} gegen unendlich streben. Das hat zur Folge, dass wir eine *steife* Differentialgleichung lösen müssen, so dass sich der Einsatz impliziter Zeitschrittverfahren empfiehlt.

Glücklicherweise ist die rechte Seite der Differentialgleichung linear, so dass sich die Durchführung der einzelnen Zeitschritte vereinfacht. Beispielsweise für das implizite Euler-Verfahren ist

$$\widetilde{\mathbf{y}}(t_i) = \widetilde{\mathbf{y}}(t_{i-1}) - h_i \mathbf{L}\widetilde{\mathbf{y}}(t_i)$$

aufzulösen, also das lineare Gleichungssystem

$$(\mathbf{I} + h_i \mathbf{L})\widetilde{\mathbf{y}}(t_i) = \widetilde{\mathbf{y}}(t_{i-1}).$$

Da $\mathbf{L}$ positiv definit ist, gilt dasselbe für $\mathbf{I} + h_i\mathbf{L}$, also lässt sich das Gleichungssystem beispielsweise durch eine LR-Zerlegung behandeln. Da die Matrix auch tridiagonal ist, ist die Anzahl der Rechenoperationen proportional zu n, so dass ein sehr effizientes Verfahren entsteht.

Index