Robert Plato

Numerische Mathematik kompakt

Grundlagenwissen für Studium und Praxis

4. Auflage

► Mit Online-Service

STUDIUM

VIEWEG+
TEUBNER

Robert Plato

Numerische Mathematik kompakt

Robert Plato

Numerische Mathematik kompakt

Grundlagenwissen für Studium und Praxis

4., aktualisierte Auflage

STUDIUM

VIEWEG+
TEUBNER

Bibliografische Information der Deutschen Nationalbibliothek
Die Deutsche Nationalbibliothek verzeichnet diese Publikation in der
Deutschen Nationalbibliografie; detaillierte bibliografische Daten sind im Internet über
<http://dnb.d-nb.de> abrufbar.

Dr. Robert Plato
Fachbereich Mathematik
Universität Siegen
Walter-Flex-Straße 3
57068 Siegen

E-Mail: plato@mathematik.uni-siegen.de

1. Auflage 2000
2., überarbeitete Auflage 2004
3., aktualisierte und verbesserte Auflage 2006
4., aktualisierte Auflage 2010

Lektorat: Ulrike Schmickler-Hirzebruch | Nastassja Vanselow

Vieweg+Teubner ist Teil der Fachverlagsgruppe Springer Science+Business Media.
www.viewegteubner.de

Umschlaggestaltung: KünkelLopka Medienentwicklung, Heidelberg
Gedruckt auf säurefreiem und chlorfrei gebleichtem Papier.
Printed in Germany

ISBN 978-3-8348-1018-2

Vorwort zur vierten Auflage

Für diese Neuauflage habe ich Aktualisierungen, Korrekturen und stilistische Änderungen vorgenommen, außerdem werden im Kapitel zur diskreten Fouriertransformation einige Abschnitte anders präsentiert. Der auf Seite vi näher beschriebene Onlinesupport mit den Lösungshinweisen bleibt auch für diese Neuauflage bestehen.

Hinweise und Verbesserungsvorschläge zu diesem Lehrbuch erreichen mich nun unter der Email-Adresse `plato@mathematik.uni-siegen.de`.

Siegen, im Oktober 2009 Robert Plato

Vorwort zur zweiten Auflage

Für die zweite Auflage ist das Layout etwas verändert worden, und zur Vereinheitlichung der Notation sind einige Umbenennungen erfolgt. Die Literaturhinweise wurden aktualisiert, der Index erweitert und Fehler beseitigt. Die Abschnitte über positiv definite Matrizen und das GMRES-Verfahren wurden etwas modifiziert, wobei dies auf Anregungen von Prof. Dr. Rembert Reemtsen (TU Cottbus) beziehungsweise G. Fuß (TU Berlin) zurückgeht. Außerdem sind in einigen Kapiteln die einführenden Bemerkungen erweitert worden.

Unter der im Vorwort zur ersten Auflage genannten Adresse wird weiterhin ein Online-Service angeboten. Mittlerweile ist ein Übungsbuch ([82]) entstanden, das vollständige Lösungswege zu den meisten der in diesem Buch vorgestellten Übungsaufgaben sowie zu weiteren Aufgaben enthält. Außerdem werden dort noch ein paar spezielle Anwendungen wie etwa die digitale Audio- und Bildkompression etwas eingehender behandelt.

Danken möchte ich der Christian-Albrechts-Universität zu Kiel, wo ich die Möglichkeit hatte, die erste Auflage des vorliegenden Buches vier Semester lang in Vorlesungen einzusetzen. Außerdem möchte ich dem DFG Forschungszentrum “Mathematik für Schlüsseltechnologien” (FZT 86) in Berlin für Unterstützung und dem Vieweg Verlag für die erneut angenehme Zusammenarbeit danken.

Berlin, im Juni 2004 Robert Plato

Vorwort zur ersten Auflage

Das vorliegende Lehrbuch ist hervorgegangen aus zwei jeweils vierstündigen Vorlesungen über Numerische Mathematik, die ich seit 1997 wiederholt an der Technischen Universität Berlin gehalten habe. Diese Vorlesungen sind in erster Linie von Studierenden der Wirtschafts- und Technomathematik und zu einem kleineren Teil

von Studierenden des Diplomstudiengangs Mathematik sowie der Physik und Informatik besucht worden.

In seiner jetzigen Form richtet sich das Lehrbuch an Studierende und Absolventen der Mathematik sowie benachbarter Fächer wie Informatik, Natur- und Ingenieurwissenschaften an Universitäten und Fachhochschulen. In kompakter Form werden zahlreiche grundlegende und für die Anwendungen wichtige Themenkomplexe aus der Numerischen Mathematik behandelt:

- Interpolation, schnelle Fouriertransformation und Integration,
- direkte und iterative Lösung linearer Gleichungssysteme,
- iterative Verfahren für nichtlineare Gleichungssysteme,
- numerische Lösung von Anfangs- und Randwertproblemen bei gewöhnlichen Differentialgleichungen,
- Eigenwertaufgaben bei Matrizen,
- Approximationstheorie und Rechnerarithmetik.

Auf die Behandlung der Numerik partieller Differentialgleichungen sowie der nichtlinearen Optimierung wird aufgrund des angestrebten überschaubaren Umfangs verzichtet.

Das Bestreben dieses Lehrbuchs ist es, die vorliegenden Themen auf möglichst elementare und übersichtliche Weise zu behandeln. Dies gilt auch für die Herleitung der Approximationseigenschaften der vorgestellten numerischen Methoden, bei der jeweils lediglich Grundkenntnisse der Analysis und der linearen Algebra vorausgesetzt werden. Außerdem sind für viele der diskutierten Verfahren die jeweiligen Vorgehensweisen durch Bilder und Schemata veranschaulicht, was das Erlernen der auftretenden Zusammenhänge erleichtern sollte. Für zahlreiche der behandelten Verfahren werden die praktisch bedeutungsvollen Aufwandsbetrachtungen angestellt und Pseudocodes angegeben, die sich unmittelbar in Computerprogramme umsetzen lassen. Die etwa 120 vorgestellten Übungsaufgaben unterschiedlichen Schwierigkeitsgrads sind fast alle im Übungsbetrieb verwendet worden und daher praxiserprobt.

Ich selbst habe die Vorläufer dieses Lehrbuchs ohne weitere Themenauswahl als Vorlage für Vorlesungen über Numerische Mathematik 1 und 2 verwendet. Dabei wurden die ersten sechs Kapitel in Teil 1 und die Kapitel 7 bis einschließlich 13 in Teil 2 der Vorlesung behandelt. Möglich wäre es aber auch, im ersten Teil die Behandlung des sechsten Kapitels über numerische Integration deutlich abzukürzen. Stattdessen könnten dann im ersten Teil beispielsweise noch die Grundlagen über Einschrittverfahren zur numerischen Lösung von Anfangswertproblemen bei gewöhnlichen Differentialgleichungen (Kapitel 7) oder Relaxationsverfahren zur iterativen Lösung linearer Gleichungssysteme (Kapitel 10) vorgestellt werden.

Zu diesem Buch wird ein Online-Service angeboten, der unter

`http://www.math.tu-berlin.de/numerik/plato/viewegbuch`

abrufbar ist. Er umfasst Lösungshinweise zu den vorgestellten Übungsaufgaben und MATLAB-Programme zu einigen der in diesem Buch präsentierten Pseudocodes. Außerdem werden über diesen Online-Service im Laufe der Zeit Abschnitte über weitere in diesem Buch nicht behandelte Themen beziehungsweise eine Liste der eventuell

anfallenden Korrekturen angeboten. Anregungen, nützliche Hinweise und Verbesserungsvorschläge zu diesem Lehrbuch sind jederzeit willkommen und erreichen mich unter meiner Email-Adresse `plato@math.tu-berlin.de`.

Mein Dank gilt meinen Kollegen Prof. Dr. R. D. Grigorieff und Dipl. Math. Etienne Emmrich für viele nützliche Anregungen, die in der vorliegenden Fassung weitestgehend berücksichtigt sind. Den Vorlesungsteilnehmern Dipl. Inf. Till Tantau und cand. math. Olivier Pfeiffer sowie einigen weiteren Studierenden sind zahlreiche kleine aber wichtige Verbesserungen zu verdanken. Außerdem danke ich Prof. Dr. Chuck Groetsch, Prof. Dr. Martin Hanke-Bourgeois und Prof. Dr. Hans-Jürgen Reinhardt für die Unterstützung bei der Durchführung dieses Buchprojekts und Frau Ulrike Schmickler-Hirzebruch vom Verlag Vieweg für die stets angenehme Zusammenarbeit.

Berlin, im Mai 2000 Robert Plato

Inhaltsverzeichnis

1 Polynominterpolation

1.1 Allgemeine Vorbetrachtungen, landausche Symbole

Gegenstand dieses und der beiden nachfolgenden Kapitel sind Problemstellungen der folgenden Art:

> Aus einer vorab festgelegten Menge von Funktionen $\mathcal{M}_n$ bestimme man eine Funktion, die durch gegebene Punkte $(x_0, f_0), (x_1, f_1), \ldots, (x_n, f_n) \in \mathbb{R}^2$ verläuft.

Hierbei ist $\mathcal{M}_n \subset \{\psi : \mathbf{I} \to \mathbb{R}\}$ eine problembezogen ausgewählte Menge von Funktionen, wobei $\mathbf{I} \subset \mathbb{R}$ ein endliches oder unendliches Intervall mit paarweise verschiedenen *Stützstellen* $x_0, x_1, \ldots, x_n \in \mathbf{I}$ ist. Solche Problemstellungen werden im Folgenden kurz als (eindimensionale) *Interpolationsprobleme* bezeichnet.

Bemerkung 1.1. Interpolationsprobleme treten in unterschiedlichen Anwendungsbereichen auf. Einige davon werden – ohne weitere Spezifikation der Menge $\mathcal{M}_n$ – im Folgenden vorgestellt:

- Durch die Interpolation von zeit- oder ortsabhängigen Messwerten wird die näherungsweise Ermittlung auch von Daten für solche Zeiten oder Orte ermöglicht, für die keine Messungen vorliegen.
- Die Interpolation lässt sich ebenfalls sinnvoll einsetzen bei der effizienten näherungsweisen Bestimmung des Verlaufs solcher Funktionen $f : \mathbf{I} \to \mathbb{R}$, die nur aufwändig auszuwerten sind. Hier wird die genannte Funktion f vorab lediglich an den vorgegebenen Stützstellen ausgewertet. Zur näherungsweisen Bestimmung der Funktionswerte von f an weiteren Stellen werden dann ersatzweise die entsprechenden Werte der interpolierenden Funktion aus $\mathcal{M}_n$ herangezogen, wobei hier $f_j = f(x_j)$ für $j = 0, 1, \ldots, n$ angenommen wird.
- Eine weitere wichtige Anwendung stellt das rechnergestützte Konstruieren (*Computer-Aided Design*, kurz *CAD*) dar, das beispielsweise zur Konstruktion von Schiffsrümpfen oder zur Festlegung von Schienenwegen verwendet wird. Mathematisch betrachtet geht es hierbei darum, interpolierende Funktionen mit hinreichend guten Glattheitseigenschaften zu verwenden.
- Es existieren weitere Anwendungen, deren Modellierung auf andere mathematische Problemstellungen führen wie etwa die numerische Integration oder die numerische Lösung von Anfangswertproblemen für gewöhnliche Differenzialgleichungen. Wie sich herausstellen wird, lassen sich hierfür unter Zuhilfenahme der Interpolation numerische Verfahren entwickeln. △

Für jedes der vorzustellenden Interpolationsprobleme sind im Prinzip die folgenden Themenkomplexe von Interesse:

- Existenz und Eindeutigkeit der interpolierenden Funktion aus der vorgegebenen Klasse von Funktionen $\mathcal{M}_n$. Dabei ist es aufgrund der vorliegenden $(n+1)$ Interpolationsbedingungen naheliegend, für $\mathcal{M}_n$ lineare Funktionenräume der Dimension $(n+1)$ heranzuziehen.
- Stabile Berechnung der Werte der interpolierenden Funktion an einer oder mehrerer Stellen.
- Aufwandsbetrachtungen für jedes der betrachteten Verfahren.
- Herleitung von Abschätzungen für den bezüglich einer gegebenen hinreichend glatten Funktion $f : [a,b] \to \mathbb{R}$ und der interpolierenden Funktion auf dem Intervall $[a,b]$ auftretenden größtmöglichen Fehler, wobei hier $f_j = f(x_j)$ für $j = 0,1,\ldots,n$ angenommen wird.

1.1.1 Landausche Symbole

Im Folgenden werden zunächst die landauschen Symbole $\mathcal{O}$ und o vorgestellt, mit denen sich bei Fehlerabschätzungen und Effizienzbetrachtungen die wichtigen Aussagen herausstellen lassen.

Definition 1.2. Gegeben seien zwei Funktionen $f,\ g : \mathbb{R}^N \supset \mathcal{D} \to \mathbb{R}$, und $x^* \in \mathbb{R}^N$ sei ein Häufungspunkt der Menge $\mathcal{D}$, es existiere also eine Folge $x^{(0)}, x^{(1)}, \ldots \subset \mathcal{D}$ mit $\max_{j=1,\ldots,N} |x_j^{(n)} - x_j^*| \to 0$ für $n \to \infty$.

(a) Die Notation

$$f(x) \;=\; \mathcal{O}(g(x)) \qquad \text{für} \quad \mathcal{D} \ni x \to x^*$$

ist gleichbedeutend mit der Existenz einer Konstanten $K \geq 0$ sowie einer Umgebung $\mathcal{U} = \{x \in \mathbb{R}^N : \max_{j=1,\ldots,N} |x_j - x_j^*| \leq \delta\}$ von x^* (mit einer Zahl $\delta > 0$), so dass die folgende Abschätzung gilt,

$$|f(x)| \;\leq\; K|g(x)| \qquad \text{für} \quad x \;\in\; \mathcal{U} \cap \mathcal{D}.$$

(b) Die Notation

$$f(x) \;=\; o(g(x)) \qquad \text{für} \quad \mathcal{D} \ni x \to x^*$$

wird verwendet, wenn für jede Zahl $\varepsilon > 0$ eine Umgebung $\mathcal{U}_\varepsilon = \{x \in \mathbb{R}^N : \max_{j=1,\ldots,N} |x_j - x_j^*| \leq \delta_\varepsilon\}$ (mit einer von ε abhängenden Zahl $\delta = \delta_\varepsilon > 0$) von x^* existiert, so dass Folgendes gilt,

$$|f(x)| \;\leq\; \varepsilon|g(x)| \qquad \text{für} \quad x \;\in\; \mathcal{U}_\varepsilon \cap \mathcal{D}.$$

Im eindimensionalen Fall $N = 1$ lassen sich diese Notationen auf die Situation $x^* = \infty$ übertragen, wobei nur die angegebenen Umgebungen durch Mengen der Form $\mathcal{U} = \{x \in \mathbb{R} : x \geq M\}$ mit Zahlen $M \in \mathbb{R}$ zu ersetzen sind.

Beispiel 1.3. (1) Wenn die Funktion g in einer Umgebung von x^* keine Nullstelle besitzt, ist $f(x) = o(g(x))$ für $\mathcal{D} \ni x \to x^*$ gleichbedeutend mit $f(x)/g(x) \to 0$ für $\mathcal{D} \ni x \to x^*$. Gilt zusätzlich noch $g(x^*) = 0$ und ist g an der Stelle x^* stetig, so impliziert (jeweils für $\mathcal{D} \ni x \to x^*$) die Aussage $f(x) = o(g(x))$ sinngemäß, dass $f(x)$ schneller gegen 0 konvergiert als $g(x)$ es tut.

(2) Es gilt $f(x) = \mathcal{O}(\mathbf{1})$ für $x \to x^*$ genau dann, wenn $f(x)$ in einer Umgebung von x^* beschränkt ist. Weiter gilt $f(x) = o(\mathbf{1})$ für $x \to x^*$ genau dann, wenn $f(x) \to 0$ für $\mathcal{D} \ni x \to x^*$ (Aufgabe 1.1). △

1.2 Existenz und Eindeutigkeit bei der Polynominterpolation

Im weiteren Verlauf dieses Kapitels werden zur Interpolation von $(n+1)$ beliebigen *Stützpunkten* $(x_0, f_0), (x_1, f_1), \ldots, (x_n, f_n) \in \mathbb{R}^2$ mit paarweise verschiedenen Stützstellen $x_0, \ldots, x_n$ speziell Funktionen aus der Menge

$$\Pi_n := \{\mathcal{P} \ : \ \mathcal{P} \text{ ist Polynom vom Grad } \leq n\}$$

herangezogen; es wird also ein Polynom $\mathcal{P}$ mit den folgenden Eigenschaften gesucht,

$$\left.\begin{array}{rcll} \mathcal{P} & \in & \Pi_n, & \\ \mathcal{P}(x_j) & = & f_j & \text{für} \quad j = 0, 1, \ldots, n. \end{array}\right\} \tag{1.1}$$

1.2.1 Die lagrangesche Interpolationsformel

Für den Nachweis der Existenz einer Lösung des Interpolationsproblems (1.1) lassen sich die folgenden Polynome verwenden.

Definition 1.4. Zu gegebenen $(n+1)$ paarweise verschiedenen Stützstellen $x_0, x_1, \ldots, x_n \in \mathbb{R}$ sind die $(n+1)$ *lagrangeschen Basispolynome* $L_0, L_1, \ldots, L_n \in \Pi_n$ folgendermaßen definiert,

$$L_k(x) = \prod_{\substack{s=0 \\ s \neq k}}^{n} \frac{x - x_s}{x_k - x_s} \qquad \text{für} \quad k = 0, 1, \ldots, n.$$

Bemerkung 1.5. Das lagrangesche Basispolynom L_k genügt offensichtlich den $(n+1)$ Interpolationsbedingungen

$$L_k(x_j) = \delta_{kj} := \begin{cases} 1 & \text{für } j = k, \\ 0 & \text{für } j \neq k. \end{cases}$$

Daraus resultiert auch unmittelbar die lineare Unabhängigkeit der lagrangeschen Basispolynome $L_0, L_1, \ldots, L_n$, so dass diese eine Basis des $(n+1)$-dimensionalen Raums Π_n der Polynome vom Grad $\leq n$ bilden. △

Das folgende Theorem behandelt die Frage der Existenz und Eindeutigkeit des interpolierenden Polynoms:

Theorem 1.6. *Zu beliebigen* $(n+1)$ *Stützpunkten* $(x_0, f_0), (x_1, f_1), \ldots, (x_n, f_n) \in \mathbb{R}^2$ *mit paarweise verschiedenen Stützstellen* $x_0, x_1, \ldots, x_n$ *existiert genau ein interpolierendes Polynom* $\mathcal{P} \in \Pi_n$ *(siehe Eigenschaft (1.1)). Es besitzt die Darstellung (*lagrangesche Interpolationsformel*)*

$$\mathcal{P}(x) = \sum_{k=0}^{n} f_k L_k(x). \tag{1.2}$$

BEWEIS. (a) Existenz: Für die Funktion $\mathcal{P}$ aus (1.2) gilt $\mathcal{P} \in \Pi_n$ und $\mathcal{P}(x_j) = \sum_{k=0}^{n} f_k \delta_{jk} = f_j$ für $j = 0, 1, \ldots, n$, wie man sofort nachrechnet.
(b) Eindeutigkeit: Wenn auch das Polynom $\mathcal{Q} \in \Pi_n$ den Interpolationsbedingungen genügt, wenn also $\mathcal{Q}(x_j) = f_j$ für $j = 0, 1, \ldots, n$ erfüllt ist, so gilt $\mathcal{Q} - \mathcal{P} \in \Pi_n$ und

$$(\mathcal{Q} - \mathcal{P})(x_j) = 0 \qquad \text{für} \quad j = 0, 1, \ldots, n.$$

Damit ist $\mathcal{Q} - \mathcal{P}$ ein Polynom vom Grad $\leq n$ mit mindestens $n+1$ paarweise verschiedenen Nullstellen, so dass (siehe beispielsweise Fischer [28], Abschnitt 1.3) notwendigerweise $\mathcal{Q} - \mathcal{P} \equiv 0$ beziehungsweise $\mathcal{Q} \equiv \mathcal{P}$ gilt. □

1.2.2 Eine erste Vorgehensweise zur Berechnung des interpolierenden Polynoms

Im Folgenden sollen Algorithmen zur Berechnung der Werte des interpolierenden Polynoms an einer oder mehrerer Stellen angegeben werden, wobei zur jeweiligen Bewertung auch Aufwandsbetrachtungen angestellt werden.

Definition 1.7. Jede der Grundoperationen Addition, Subtraktion, Multiplikation und Division sowie die Wurzelfunktion wird im Folgenden als *arithmetische Operation* bezeichnet.

Der jeweils zu betreibende Aufwand eines Verfahrens lässt sich über die Anzahl der durchzuführenden arithmetischen Operationen beschreiben. Der Einfachheit halber bleibt im Folgenden unberücksichtigt, dass ein Mikroprozessor zur Ausführung einer Division beziehungsweise zur Berechnung einer Quadratwurzel jeweils etwa vier mal so viel Zeit benötigt wie zur Durchführung einer Addition, einer Subtraktion oder einer Multiplikation (Überhuber [106], Abschnitt 5.5).

Wie sich herausstellt, ist die folgende Zielsetzung realistisch:

> Angestrebtes Ziel ist die Herleitung von Verfahren, für die das zu $(n+1)$ Stützpunkten gehörende interpolierende Polynom $\mathcal{P}$ (siehe (1.1)) nach einer Anlaufrechnung mit $\mathcal{O}(n^2)$ arithmetischen Operationen an jeder Stelle $x \in \mathbb{R}$ in $\mathcal{O}(n)$ arithmetischen Operationen ausgewertet werden kann. (1.3)

Hierbei sind Ausdrücke der Form "$\mathcal{O}(n^q)$" eine Kurzform für "$\mathcal{O}(n^q)$ für $n \to \infty$".

Eine erste Variante zur Bestimmung eines interpolierenden Polynoms mit dem in (1.3) angestrebten maximalen Aufwand basiert auf der folgenden Darstellung für die lagrangeschen Basispolynome,

$$L_k(x) = \prod_{\substack{s=0 \\ s\neq k}}^{n} \frac{x - x_s}{x_k - x_s} = \frac{\kappa_k}{x - x_k} q(x), \qquad k = 0, 1, \ldots, n, \tag{1.4}$$

$$\text{mit} \quad \kappa_k = \prod_{\substack{s=0 \\ s\neq k}}^{n} \frac{1}{x_k - x_s}, \qquad q(x) = \prod_{s=0}^{n} (x - x_s).$$

Die Zahlen $\kappa_0, \kappa_1, \ldots, \kappa_n$, die auch als *Stützkoeffizienten* bezeichnet werden, lassen sich mit einem Aufwand von insgesamt $\mathcal{O}(n^2)$ arithmetischen Operationen ermitteln. Sind diese Koeffizienten einmal berechnet, so lässt sich für jede Zahl $x \in \mathbb{R}$ der Wert $\mathcal{P}(x) = q(x)(\sum_{k=0}^{n} \kappa_k f_k/(x - x_k))$ in $\mathcal{O}(n)$ arithmetischen Operationen bestimmen, wie man sich leicht überlegt.

Diese Vorgehensweise zur Berechnung von $\mathcal{P}(x)$ lässt sich also mit in (1.3) angestrebten maximalen Aufwand realisieren und hat zudem den praxisrelevanten Vorteil, dass die in der Anlaufrechnung berechneten Koeffizienten $\kappa_0, \kappa_1, \ldots, \kappa_n$ nicht von den Stützwerten $f_0, f_1, \ldots, f_n$ abhängen. Bei einem Wechsel der Stützwerte f_0, $f_1, \ldots, f_n$ unter gleichzeitiger Beibehaltung der Stützstellen $x_0, x_1, \ldots, x_n$ ist also eine erneute Anlaufrechnung nicht erforderlich.

Bemerkung 1.8. Die Entwicklung des interpolierenden Polynoms $\mathcal{P} \in \Pi_n$ als Linearkombination der lagrangeschen Basispolynome in Kombination mit der in diesem Abschnitt 1.2.2 beschriebenen Vorgehensweise zur Auswertung von $\mathcal{P}(x)$ führt jedoch für nahe bei Stützstellen liegende Zahlen x zu Instabilitäten, was zurückzuführen ist auf auftretende Brüche mit betragsmäßig kleinen Nennern und Zählern.

Andererseits führt der Ansatz $\mathcal{P}(x) = \sum_{k=0}^{n} a_k x^k$ als Linearkombination der Monome zusammen mit den Interpolationsbedingungen auf ein lineares Gleichungssystem, dessen Lösung sich als zu aufwändig und zu empfindlich gegenüber Rundungsfehlern erweist. △

In Abschnitt 1.4 wird eine Darstellung des interpolierenden Polynoms bezüglich einer anderen Basis behandelt, mit der sich das interpolierende Polynom $\mathcal{P}$ mit dem in (1.3) angegebenen maximalen Aufwand stabil berechnen lässt.

1.3 Neville-Schema

Die Lösung für das Interpolationsproblem (1.1) kann schrittweise aus den interpolierenden Polynomen zu $m = 0, 1, \ldots$ Stützpunkten berechnet werden, wie sich im Folgenden herausstellt. Einerseits wird dieses Resultat für den Beweis der wesentlichen Aussage des nachfolgenden Abschnitts benötigt, andererseits erhält man dabei eine allgemein beliebte Vorgehensweise zur Auswertung des interpolierenden Polynoms an einigen wenigen Stellen.

Definition 1.9. Seien $k,\ m \in \mathbb{N}_0$. Zu den $(m+1)$ Stützpunkten $(x_k, f_k), (x_{k+1}, f_{k+1}), \ldots, (x_{k+m}, f_{k+m})$ bezeichne $\mathcal{P}_{k,k+1,\ldots,k+m}$ dasjenige (eindeutig bestimmte) Polynom vom Grad $\leq m$ mit der Eigenschaft

$$\mathcal{P}_{k,k+1,\ldots,k+m}(x_j) \;=\; f_j \qquad \text{für} \quad j \;=\; k, k+1, \ldots, k+m. \tag{1.5}$$

Für die vorgestellten Polynome $\mathcal{P}_{k,k+1,\ldots,k+m}$ besteht die folgende Rekursionsbeziehung:

Theorem 1.10. *Seien $(x_0, f_0), (x_1, f_1), \ldots, (x_n, f_n)$ vorgegebene Stützpunkte. Für die Interpolationspolynome $\mathcal{P}_{k,k+1,\ldots,k+m}$ (mit $k \geq 0$ und $m \geq 0$ mit $k+m \leq n$) aus (1.5) gilt die Rekursionsformel*

$$\mathcal{P}_k(x) \equiv f_k, \tag{1.6}$$

$$\mathcal{P}_{k,k+1,\ldots,k+m}(x) = \frac{(x-x_k)\mathcal{P}_{k+1,\ldots,k+m}(x) \;-\; (x-x_{k+m})\mathcal{P}_{k,\ldots,k+m-1}(x)}{x_{k+m} - x_k},\ m \geq 1. \tag{1.7}$$

BEWEIS. Die Identität (1.6) ist wegen $\mathcal{P}_k \in \Pi_0$ und $\mathcal{P}_k(x_k) = f_k$ offensichtlich richtig. Es bezeichne $\mathcal{Q}(x)$ die rechte Seite von (1.7), und $\mathcal{Q} = \mathcal{P}_{k,k+1,\ldots,k+m}$ ist dann nachzuweisen, was im Folgenden geschieht. Es gilt $\mathcal{P}_{k+1,\ldots,k+m} \in \Pi_{m-1}$ und $\mathcal{P}_{k,\ldots,k+m-1} \in \Pi_{m-1}$ und demnach $\mathcal{Q} \in \Pi_m$. Weiter gilt

$$\mathcal{Q}(x_k) \;=\; \frac{0 - (x_k - x_{k+m})f_k}{x_{k+m} - x_k} \;=\; f_k, \quad \mathcal{Q}(x_{k+m}) \;=\; \frac{(x_{k+m} - x_k)f_{k+m} - 0}{x_{k+m} - x_k} \;=\; f_{k+m},$$

und für $j = k+1, k+2, \ldots, k+m-1$ gilt

$$\mathcal{Q}(x_j) \;=\; \frac{(x_j - x_k)f_j - (x_j - x_{k+m})f_j}{x_{k+m} - x_k} \;=\; \frac{(-x_k + x_{k+m})f_j}{x_{k+m} - x_k} \;=\; f_j.$$

Aufgrund der Eindeutigkeit des interpolierenden Polynoms (Theorem 1.6) gilt daher notwendigerweise die Identität $\mathcal{Q} = \mathcal{P}_{k,k+1,\ldots,k+m}$. □

Die sich für die Werte $\mathcal{P}_{k,k+1,\ldots,k+m}(x)$ aus der Rekursionsformel (1.7) ergebenden Abhängigkeiten sind in Schema 1.1 dargestellt, das als *Neville-Schema* bezeichnet wird. Die Einträge in Schema 1.1 lassen sich beispielsweise spaltenweise jeweils von oben nach unten berechnen. Wie bereits erwähnt wird das resultierende Verfahren zur Auswertung des interpolierenden Polynoms $\mathcal{P}(x) = \mathcal{P}_{0\ldots n}(x)$ an einzelnen Stellen x verwendet, wobei jeweils $7n^2/2 + \mathcal{O}(n)$ arithmetische Operationen anfallen, wie man leicht nachzählt.

$$
\begin{array}{lllllll}
f_0 = \mathcal{P}_0(x) & & & & & & \\
 & \searrow & & & & & \\
f_1 = \mathcal{P}_1(x) & \to & \mathcal{P}_{01}(x) & & & & \\
 & \searrow & & \searrow & & & \\
f_2 = \mathcal{P}_2(x) & \to & \mathcal{P}_{12}(x) & \to & \mathcal{P}_{012}(x) & & \\
\vdots & & \vdots & & \vdots & \ddots & \\
f_{n-1} = \mathcal{P}_{n-1}(x) & \to & \mathcal{P}_{n-2,n-1}(x) & \to & \cdots & \cdots & \mathcal{P}_{0\ldots n-1}(x) \\
 & \searrow & & \searrow & & & \searrow \\
f_n = \mathcal{P}_n(x) & \to & \mathcal{P}_{n-1,n}(x) & \to & \cdots & \cdots & \mathcal{P}_{1\ldots n}(x) \to \mathcal{P}_{0\ldots n}(x)
\end{array}
$$

Schema 1.1: Neville-Schema

Beispiel 1.11. Man betrachte folgende Stützpunkte,

j	0	1	2
x_j	0	1	3
f_j	1	3	2

Für $x = 2$ sind die Werte des Neville-Schemas in Schema 1.2 angegeben.

$$
\begin{array}{lllll}
f_0 = \mathcal{P}_0(2) = 1 & & & & \\
f_1 = \mathcal{P}_1(2) = 3 & \mathcal{P}_{01}(2) = 5 & & & \\
f_2 = \mathcal{P}_2(2) = 2 & \mathcal{P}_{12}(2) = 5/2 & \mathcal{P}_{012}(2) = 10/3 &
\end{array}
$$

Schema 1.2: Neville-Schema zu Beispiel 1.11

Die Einträge in Schema 1.2 ergeben sich dabei folgendermaßen:

$$
\begin{aligned}
\mathcal{P}_{01}(2) &= \frac{(2-0)\mathcal{P}_1(2) - (2-1)\mathcal{P}_0(2)}{1-0} = \frac{2\cdot 3 - 1\cdot 1}{1} = 5, \\
\mathcal{P}_{12}(2) &= \frac{(2-1)\mathcal{P}_2(2) - (2-3)\mathcal{P}_1(2)}{3-1} = \frac{1\cdot 2 - (-1)\cdot 3}{2} = \frac{5}{2}, \\
\mathcal{P}_{012}(2) &= \frac{(2-0)\mathcal{P}_{12}(2) - (2-3)\mathcal{P}_{01}(2)}{3-0} = \frac{2\cdot 5/2 - (-1)\cdot 5}{3} = \frac{10}{3}.
\end{aligned}
$$

△

1.4 Die newtonsche Interpolationsformel, dividierte Differenzen

In diesem Abschnitt wird eine weitere Darstellung des interpolierenden Polynoms behandelt. Hierfür werden die folgenden Basispolynome benötigt.

Definition 1.12. Zu gegebenen paarweise verschiedenen $(n+1)$ Stützstellen $x_0, x_1, \ldots, x_n \in \mathbb{R}$ sind die speziellen $(n+1)$ *newtonschen Basispolynome* folgendermaßen erklärt:

$$1, \quad x - x_0, \quad (x-x_0)(x-x_1), \ \ldots\ldots, \ (x-x_0)(x-x_1)\ldots(x-x_{n-1}).$$

Das gesuchte interpolierende Polynom $\mathcal{P} \in \Pi_n$ mit $\mathcal{P}(x_j) = f_j$ für $j = 0, 1, \ldots, n$ (vergleiche (1.1)) soll nun als Linearkombination der newtonschen Basispolynome dargestellt werden, also in der Form

$$\left.\begin{aligned} \mathcal{P}(x) \;=\; a_0 \;+\; a_1(x-x_0) \;+\; a_2(x-x_0)(x-x_1) \;+\; \ldots \\ \ldots \;+\; a_n(x-x_0)(x-x_1)\ldots(x-x_{n-1}) \end{aligned}\right\} \tag{1.8}$$

mit noch zu bestimmenden Koeffizienten $a_0, a_1, \ldots, a_n$. Sind die Koeffizienten $a_0, a_1, \ldots, a_n$ erst einmal bestimmt, so kann für jede Zahl $x = \xi$ das Polynom (1.8) mit dem *Horner-Schema*

$$\mathcal{P}(\xi) \;=\; \big[\big[\ldots\big[a_n(\xi - x_{n-1}) \;+\; a_{n-1}\big](\xi - x_{n-2}) \;+\; \ldots \;+\; a_1\big](\xi - x_0) \;+\; a_0$$

ausgewertet werden, wobei die (insgesamt $3n$) arithmetischen Operationen von links nach rechts auszuführen sind.

Bemerkung 1.13. Die Koeffizienten $a_0, a_1, \ldots, a_n$ können im Prinzip aus den Gleichungen

$$\begin{aligned} f_0 &= \mathcal{P}(x_0) = a_0, \\ f_1 &= \mathcal{P}(x_1) = a_0 + a_1(x_1 - x_0), \\ f_2 &= \mathcal{P}(x_2) = a_0 + a_1(x_2 - x_0) + a_2(x_2 - x_0)(x_2 - x_1), \\ &\vdots \quad \vdots \quad \vdots \end{aligned}$$

gewonnen werden, wobei allerdings $n^3/3 + \mathcal{O}(n^2)$ arithmetische Operationen anfallen, wie man sich leicht überlegt. Im Folgenden soll eine günstigere Vorgehensweise vorgestellt werden, die eine Berechnung dieser Koeffizienten mit den angestrebten $\mathcal{O}(n^2)$ arithmetischen Operationen ermöglicht. △

Definition 1.14. Zu gegebenen Stützpunkten $(x_0, f_0), (x_1, f_1), \ldots, (x_n, f_n) \in \mathbb{R}^2$ sind die *dividierten Differenzen* folgendermaßen erklärt:

$$\begin{aligned} f[x_k] &:= f_k, \qquad k = 0, 1, \ldots, n, \\ f[x_k, \ldots, x_{k+m}] &:= \frac{f[x_{k+1}, \ldots, x_{k+m}] - f[x_k, \ldots, x_{k+m-1}]}{x_{k+m} - x_k}, \end{aligned}$$

für $0 \le k,\ m \le n$ mit $k + m \le n$.

Bemerkung 1.15. 1. Die dividierte Differenz $f[x_k, \ldots, x_{k+m}]$ hängt neben den Stützstellen $x_k, x_{k+1}, \ldots, x_{k+m}$ auch von den Stützwerten $f_k, f_{k+1}, \ldots, f_{k+m}$ ab.
2. Werden die *Stützwerte* etwa mit g_j anstelle f_j bezeichnet, so wird für die dividierten Differenzen naheliegenderweise die Bezeichnung $g[x_k, \ldots, x_{k+m}]$ verwendet.
3. Für die Berechnung aller dividierten Differenzen zu den Stützpunkten (x_0, f_0), $(x_1, f_1), \ldots, (x_n, f_n) \in \mathbb{R}^2$ sind lediglich $3n(n+1)/2$ arithmetische Operationen erforderlich. △

Die Abhängigkeiten zwischen den dividierten Differenzen sind in Schema 1.3 dargestellt.

$$
\begin{array}{lcccccccc}
f_0 & = & f[x_0] & & & & & & \\
 & & \searrow & & & & & & \\
f_1 & = & f[x_1] & \to & f[x_0,x_1] & & & & \\
 & & \searrow & & \searrow & & & & \\
f_2 & = & f[x_2] & \to & f[x_1,x_2] & \to & f[x_0,x_1,x_2] & & \\
\vdots & & \vdots & & \vdots & & \vdots & \ddots & \\
f_{n-1} & = & f[x_{n-1}] & \to & f[x_{n-2},x_{n-1}] & \to & \cdots & \cdots & f[x_0,\ldots,x_{n-1}] \\
 & & \searrow & & \searrow & & & & \searrow \\
f_n & = & f[x_n] & \to & f[x_{n-1},x_n] & \to & \cdots & \cdots & f[x_1,\ldots,x_n] \to f[x_0,\ldots,x_n]
\end{array}
$$

Schema 1.3: Abhängigkeiten zwischen den dividierten Differenzen

Beispielsweise gilt

$$
\begin{aligned}
f[x_0,x_1] &= \frac{f[x_1]-f[x_0]}{x_1-x_0}, \qquad & f[x_1,x_2] &= \frac{f[x_2]-f[x_1]}{x_2-x_1}, \\
f[x_0,x_1,x_2] &= \frac{f[x_1,x_2]-f[x_0,x_1]}{x_2-x_0}. &&
\end{aligned}
$$

Das nachfolgende Theorem liefert die wesentliche Aussage dieses Abschnitts 1.4.

Theorem 1.16 (Newtonsche Interpolationsformel). *Für das interpolierende Polynom $\mathcal{P} \in \Pi_n$ zu gegebenen $(n+1)$ Stützpunkten $(x_0, f_0), (x_1, f_1), \ldots, (x_n, f_n) \in \mathbb{R}^2$ gilt*

$$
\begin{aligned}
\mathcal{P}(x) &= f[x_0] + f[x_0,x_1](x-x_0) + \ldots \\
&\quad \ldots + f[x_0,\ldots,x_n](x-x_0)(x-x_1)\cdots(x-x_{n-1}). \qquad (1.9)
\end{aligned}
$$

BEWEIS. Dieser wird per vollständiger Induktion über n geführt. Die Aussage ist sicher richtig für $n = 0$ und beliebige Stützpunkte (x_0, f_0), und es sei nun angenommen, dass sie richtig ist für $n \in \mathbb{N}_0$ und beliebige Stützpunkte $(x_0, f_0), (x_1, f_1), \ldots, (x_n, f_n) \in \mathbb{R}^2$. Im Folgenden seien $(n+2)$ Stützpunkte $(x_0, f_0), (x_1, f_1), \ldots, (x_{n+1},$

$f_{n+1}) \in \mathbb{R}^2$ gegeben, und $\mathcal{P} \in \Pi_{n+1}$ bezeichne das zugehörige interpolierende Polynom. Mit der Notation aus Definition 1.9 gilt dann

$$\begin{aligned} \mathcal{P} - \mathcal{P}_{0,\ldots,n} &\in \Pi_{n+1}, \\ \mathcal{P}(x_j) - \mathcal{P}_{0,\ldots,n}(x_j) &= 0 \quad \text{für } j = 0,1,\ldots,n, \end{aligned}$$

und damit gilt $\mathcal{P}(x) - \mathcal{P}_{0,\ldots,n}(x) = a(x-x_0)\cdots(x-x_n)$ beziehungsweise

$$\mathcal{P}(x) = \mathcal{P}_{0,\ldots,n}(x) + a(x-x_0)\cdots(x-x_n) \tag{1.10}$$

mit einer geeigneten Konstanten $a \in \mathbb{R}$ (Übungsaufgabe; folgt aus der Eindeutigkeit des interpolierenden Polynoms (Theorem 1.6)). Nach Induktionsvoraussetzung gilt

$$\left.\begin{aligned} \mathcal{P}_{0,\ldots,n}(x) = f[x_0] + f[x_0,x_1](x-x_0) + \ldots \\ \ldots + f[x_0,\ldots,x_n](x-x_0)(x-x_1)\cdots(x-x_{n-1}), \end{aligned}\right\} \tag{1.11}$$

so dass wegen (1.10), (1.11) noch

$$a = f[x_0,\ldots,x_{n+1}] \tag{1.12}$$

nachzuweisen ist. Zu diesem Zweck verwendet man entsprechend Theorem 1.10 die Identität

$$\mathcal{P}(x) = \frac{(x-x_0)\mathcal{P}_{1,\ldots,n+1}(x) - (x-x_{n+1})\mathcal{P}_{0,\ldots,n}(x)}{x_{n+1}-x_0} \tag{1.13}$$

und führt in (1.13) einen Koeffizientenvergleich durch. Wegen der Identität (1.10) ist klar, dass a der führende Koeffizient von $\mathcal{P}$ ist, es gilt also

$$\mathcal{P} = \mathcal{Q} + ax^{n+1}$$

für ein gewisses Polynom $\mathcal{Q} \in \Pi_n$. Andererseits ist nach Induktionsvoraussetzung bekannt, dass das Polynom $\mathcal{P}_{1,\ldots,n+1}$ den führenden Koeffizienten $f[x_1,\ldots,x_{n+1}]$ sowie $\mathcal{P}_{0,\ldots,n}$ den führenden Koeffizienten $f[x_0,\ldots,x_n]$ besitzt; wegen (1.13) besitzt $\mathcal{P}$ also tatsächlich den führenden Koeffizienten

$$a = \frac{f[x_1,\ldots,x_{n+1}] - f[x_0,\ldots,x_n]}{x_{n+1}-x_0} \overset{\text{def}}{=} f[x_0,\ldots,x_{n+1}],$$

was identisch mit (1.12) ist und den Beweis komplettiert. □

1.5 Fehlerdarstellungen zur Polynominterpolation

Das folgende Theorem liefert für hinreichend glatte Funktionen eine Darstellung des bei der Polynominterpolation auftretenden Fehlers.

Theorem 1.17. *Die Funktion $f : [a,b] \to \mathbb{R}$ sei $(n+1)$-mal differenzierbar und sei $\mathcal{P} \in \Pi_n$ das Polynom mit $\mathcal{P}(x_j) = f(x_j)$ für $j = 0, 1, \ldots, n$. Für jedes $\overline{x} \in [a,b]$ gilt dann die Fehlerdarstellung*

$$f(\overline{x}) - \mathcal{P}(\overline{x}) = \frac{\omega(\overline{x})\, f^{(n+1)}(\xi)}{(n+1)!}, \tag{1.14}$$

mit einer Zwischenstelle $\xi = \xi(\overline{x}) \in [a,b]$ und

$$\omega(x) := (x - x_0) \cdots (x - x_n).$$

BEWEIS. Falls $\overline{x} = x_j$ für ein j gilt, so verschwinden beide Seiten der Gleichung (1.14). Sei nun $\overline{x} \notin \{x_0, x_1, \ldots, x_n\}$ und sei

$$\psi(x) := f(x) - \mathcal{P}(x) - K\,\omega(x),$$

wobei die Konstante K so gewählt sei, dass

$$\psi(\overline{x}) = 0$$

erfüllt ist. Im Folgenden soll eine spezielle Darstellung für die Konstante K hergeleitet werden. Hierzu beobachtet man, dass die Funktion ψ in dem Intervall $[a,b]$ mindestens $(n+2)$ paarweise verschiedene Nullstellen

$$x_0, \ldots, x_n, \overline{x}$$

besitzt. Eine wiederholte Anwendung des Theorems von Rolle zeigt: Die Funktion ψ' besitzt in dem Intervall $[a,b]$ mindestens $(n+1)$ paarweise verschiedene Nullstellen, die Funktion ψ'' besitzt in $[a,b]$ mindestens noch n paarweise verschiedene Nullstellen, und eine Fortführung dieses Arguments liefert die Existenz einer Nullstelle ξ der Funktion $\psi^{(n+1)}$ in dem Intervall $[a,b]$. Nun gilt aber

$$\mathcal{P}^{(n+1)} \equiv 0, \qquad \omega^{(n+1)} \overset{(*)}{\equiv} (n+1)!,$$

wobei man die Identität $(*)$ aufgrund des Umstands erhält, dass $\omega \in \Pi_{n+1}$ den führenden Koeffizienten eins besitzt. Insgesamt erhält man

$$\psi^{(n+1)}(\xi) = f^{(n+1)}(\xi) - K\,(n+1)! = 0$$

beziehungsweise $K = \frac{f^{(n+1)}(\xi)}{(n+1)!}$, was den Nachweis für die angegebene Fehlerdarstellung (1.14) komplettiert. □

Der Fehlerdarstellung (1.14) kann man unmittelbar entnehmen, dass beliebig oft differenzierbare Funktionen $f : [a,b] \to \mathbb{R}$ mit gleichmäßig beschränkten Ableitungen durch interpolierende Polynome gut approximiert werden (siehe das nachfolgende Theorem). Vorbereitend wird für eine Unterteilung

$$\Delta = \left\{ a = x_0^{(\Delta)} < x_1^{(\Delta)} < \ldots < x_{n(\Delta)}^{(\Delta)} = b \right\}$$

des vorgegebenen Intervalls $[a,b]$ das nachfolgende Maß für die Feinheit der Unterteilung Δ eingeführt,

$$\|\Delta\| := \max_{1\le j\le n(\Delta)} \{x_j^{(\Delta)} - x_{j-1}^{(\Delta)}\}.$$

Man beachte, dass das folgende Theorem auch für Intervallunterteilungen $\Delta^{(0)}, \Delta^{(1)}, \ldots$ mit der Eigenschaft $\|\Delta^{(m)}\| \not\to 0$ für $m \to \infty$ gültig ist.

Theorem 1.18. *Die Funktion $f : [a,b] \to \mathbb{R}$ sei unendlich oft differenzierbar, mit $\max_{x\in[a,b]} |f^{(s)}(x)| \le M$ für $s = 0, 1, \ldots$, mit einer endlichen Konstanten M. Weiter sei $\Delta^{(0)}, \Delta^{(1)}, \ldots$ eine Folge von Unterteilungen des Intervalls $[a,b]$ mit $n_m := n(\Delta^{(m)}) \to \infty$ für $m \to \infty$. Dann konvergiert die zugehörige Folge der interpolierenden Polynome $\mathcal{P}_m \in \Pi_{n_m}$ (welche bezüglich der Unterteilung $\Delta^{(m)}$ die zugehörigen Funktionswerte von f interpolieren) gleichmäßig gegen die Funktion f.*

BEWEIS. Mit der Notation aus Theorem 1.17 gilt $\max_{x\in[a,b]} |\omega(x)| \le (b-a)^{n_m+1}$ und somit

$$\max_{x\in[a,b]} |\mathcal{P}_m(x) - f(x)| \;\le\; M\frac{(b-a)^{n_m+1}}{(n_m+1)!} \;\to\; 0 \qquad \text{für} \quad m \to \infty. \qquad \square$$

Gleichmäßige Konvergenz der Interpolationspolynome erhält man auch unter geringeren Differenzierbarkeitsannahmen an die Funktion f (siehe Maess [69], Band 2). Im Allgemeinen kann man jedoch nicht erwarten, dass eine fest vorgegebene stetige Funktion auf einem kompakten Intervall umso besser durch ein interpolierendes Polynom approximiert wird, je feiner nur die Unterteilung der Stützstellen gewählt wird. Diese Aussage wird in dem folgenden Theorem 1.19 präzisiert, das hier ohne Beweis angegeben wird und insbesondere für Intervallunterteilungen $\Delta^{(0)}, \Delta^{(1)}, \ldots$ mit $\|\Delta^{(m)}\| \to 0$ für $m \to \infty$ von Bedeutung ist.

Theorem 1.19 (Faber). *Zu jeder Folge von Unterteilungen $\Delta^{(0)}, \Delta^{(1)}, \ldots$ des Intervalls $[a,b]$ gibt es eine stetige Funktion $f : [a,b] \to \mathbb{R}$, so dass die Folge der Polynome $\mathcal{P}_m \in \Pi_{n(\Delta^{(m)})}$ (welche bezüglich der Unterteilung $\Delta^{(m)}$ die zugehörigen Funktionswerte von f interpolieren) für $m \to \infty$ nicht gleichmäßig gegen die Funktion f konvergieren.*

Eine weitere, ohne Differenzierbarkeitsannahmen auskommende Fehlerdarstellung zur Polynominterpolation wird durch dividierte Differenzen ermöglicht:

Theorem 1.20. *Mit den Notationen von Theorem 1.17 mit einer beliebigen Funktion $f : [a,b] \to \mathbb{R}$ gilt im Fall $\overline{x} \notin \{x_0, \ldots, x_n\}$ die folgende Darstellung für den Interpolationsfehler,*

$$f(\overline{x}) - \mathcal{P}(\overline{x}) \;=\; f[x_0, \ldots, x_n, \overline{x}]\,\omega(\overline{x}).$$

BEWEIS. Mit $x_{n+1} := \overline{x}$ gilt aufgrund von Theorem 1.16 die Darstellung

$$\mathcal{P}_{0,\ldots,n+1}(x) \;=\; \underbrace{\mathcal{P}_{0,\ldots,n}(x)}_{=\;\mathcal{P}(x)} \;+\; f[x_0,\ldots,x_n,\overline{x}]\,\omega(x) \qquad \text{für} \quad x \in \mathbb{R},$$

und mit der Identität $f(\overline{x}) = \mathcal{P}_{0,\ldots,n+1}(\overline{x})$ folgt dann die Aussage des Theorems. □

Als Konsequenz aus den Theoremen 1.17 und 1.20 erhält man den folgenden Mittelwertsatz für höhere Ableitungen:

Korollar 1.21. *Zu jeder n-mal differenzierbaren Funktion $f : [a,b] \to \mathbb{R}$ und paarweise verschiedenen Stützstellen $x_0, x_1, \ldots, x_n \in [a,b]$ existiert eine Zwischenstelle $\xi = \xi(x) \in [a,b]$ mit*

$$f[x_0,\ldots,x_n] \;=\; \frac{f^{(n)}(\xi)}{n!},$$

wobei die Stützwerte durch $f_j = f(x_j)$ für $j = 0,1,\ldots,n$ festgelegt sind.

BEWEIS. Für $n = 0$ ist die Aussage trivialerweise richtig, und für $n \geq 1$ folgt sie unmittelbar aus einem Vergleich der rechten Seiten in den Theoremen 1.17 und 1.20, angewandt mit den Stützstellen $x_0,\ldots,x_{n-1}$ und für $\overline{x} = x_n$. □

1.6 Tschebyscheff-Polynome

In diesem Abschnitt wird unter anderem der Frage nachgegangen, für welche Stützstellen $x_0, x_1, \ldots, x_n \in [a,b]$ der Ausdruck $\max_{x\in[a,b]} |(x-x_0)\ldots(x-x_n)|$ am kleinsten wird, es ist also eine Lösung des Minimax-Problems

$$\max_{x\,\in\,[a,b]} |(x-x_0)\ldots(x-x_n)| \;\to\; \min \qquad\qquad \text{für} \quad x_0,x_1,\ldots,x_n \in [a,b]$$

zu bestimmen. Die Darstellung (1.14) lässt bei einer solchen "optimalen" Wahl der Stützstellen (falls diese zudem paarweise verschieden sind) einen minimalen Fehler bei der Polynominterpolation erwarten. Die Untersuchungen werden zunächst auf das Intervall $[a,b] = [-1,1]$ beschränkt; auf die allgemeine Situation für $[a,b]$ wird am Ende dieses Abschnitts eingegangen.

Es stellt sich im Folgenden heraus, dass solche optimalen Stützstellen $x_0, x_1, \ldots, x_n \in [-1,1]$ durch die Nullstellen des $(n+1)$-ten Tschebyscheff-Polynoms der ersten Art gegeben sind.

Definition 1.22. Die *Tschebyscheff-Polynome der ersten Art* sind folgendermaßen erklärt,

$$T_n(t) \;=\; \cos(n \arccos t), \quad t \in [-1,1] \qquad\qquad (n \;=\; 0,1,\ldots). \tag{1.15}$$

Theorem 1.23. *Für die Funktionen $T_0, T_1, \ldots$ aus (1.15) gelten die folgenden Aussagen:*

(a) $T_n(\cos\theta) = \cos n\theta \qquad$ *für* $\theta \in [0,\pi] \qquad (n = 0, 1, \ldots)$.

(b) Für $t \in [-1,1]$ *gilt* $T_0(t) = 1,\ T_1(t) = t$ *und*

$$T_{n+1}(t) = 2t\,T_n(t) - T_{n-1}(t), \qquad n = 1, 2, \ldots, \tag{1.16}$$

und Fortsetzung des Definitionsbereichs des Tschebyscheff-Polynoms T_n *auf ganz* $\mathbb{R}$ *mittels dieser Rekursionsformel liefert*

$$T_n \in \Pi_n. \tag{1.17}$$

(c) Der führende Koeffizient von T_n *ist für* $n \geq 1$ *gleich* 2^{n-1}.

(d) $\max\limits_{t\in[-1,1]} |T_n(t)| = 1.$

(e) Das Tschebyscheff-Polynom T_n *besitzt in dem Intervall* $[-1,1]$ *insgesamt* $(n+1)$ *Extrema:*

$$T_n(s_k^{(n)}) = (-1)^k \qquad \text{für } s_k^{(n)} := \cos\left(\tfrac{k\pi}{n}\right), \qquad k = 0, 1, \ldots, n. \tag{1.18}$$

(f) Das Tschebyscheff-Polynom T_n *besitzt* n *einfache Nullstellen, die allesamt in dem Intervall* $[-1,1]$ *liegen:*

$$T_n(t_k^{(n)}) = 0 \qquad \text{für } t_k^{(n)} := \cos\left(\tfrac{(2k-1)\pi}{2n}\right), \qquad k = 1, 2, \ldots, n. \tag{1.19}$$

BEWEIS. Die Aussage (a) ist offensichtlich richtig, und die Darstellungen für T_0 und T_1 in (b) ergeben sich sofort aus Teil (a). Für die Herleitung der Rekursionsformel (1.16) wird das folgende Additionstheorem benötigt,

$$\cos x + \cos y = 2\cos\left(\tfrac{x+y}{2}\right)\cos\left(\tfrac{x-y}{2}\right) \qquad \text{für } x,\ y \in \mathbb{R}. \tag{1.20}$$

Für $t = \cos\theta$ erhält man dann mit (1.20) sowie Teil (a) dieses Theorems die folgenden Identitäten,

$$2tT_n(t) - T_{n-1}(t) = 2\cos\theta\cos[n\theta] - \cos[(n-1)\theta] = \cos[(n+1)\theta] = T_{n+1}(t).$$

Teil (c) folgt unmittelbar aus der Rekursionsformel (b), und schließlich sind (d), (e) und (f) offensichtlich richtig. □

Das nachfolgende Theorem liefert die wesentliche Aussage dieses Abschnitts 1.6.

Theorem 1.24. *Für* $n \in \mathbb{N}_0$ *und mit der Notation aus (1.19) gilt die folgende Optimalitätseigenschaft:*

$$\min_{y_0,\ldots,y_n\in[-1,1]} \max_{t\in[-1,1]} |(t-y_0)\ldots(t-y_n)|$$

$$= \max_{t\in[-1,1]} |(t-t_1^{(n+1)})\ldots(t-t_{n+1}^{(n+1)})| \tag{1.21}$$

$$= \frac{1}{2^n}. \tag{1.22}$$

BEWEIS. Als Erstes beobachtet man, dass mit T_{n+1} entsprechend (1.15) die Darstellung

$$[\tfrac{1}{2^n}T_{n+1}](t) \;=\; (t-t_1^{(n+1)})\dots(t-t_{n+1}^{(n+1)}) \tag{1.23}$$

gilt, was sich unmittelbar aus Theorem 1.23, Teil (c) und (f) ergibt. Die Identität (1.22) folgt damit aus $\max_{t\in[-1,1]}|T_{n+1}(t)| = 1$ (Theorem 1.23, Teil (d)). Bei der Identität (1.21) ist die Abschätzung "$\le$" offensichtlich, und im Folgenden soll die Abschätzung "$\ge$" durch eine Widerspruchsannahme nachgewiesen werden. Angenommen, es gibt Zahlen $y_0, y_1, \dots, y_n \in [-1,1]$, so dass

$$\frac{1}{2^n} \;>\; \max_{t\in[-1,1]} |\omega(t)|, \qquad \omega(t) \;:=\; (t-y_0)\dots(t-y_n) \tag{1.24}$$

gilt. Dann besitzt das Polynom

$$\mathcal{P} \;:=\; \tfrac{1}{2^n}T_{n+1} - \omega$$

$(n+1)$ Nullstellen in $[-1,1]$, denn es liegen $(n+1)$ Vorzeichenwechsel vor, wie sich bei Betrachtung der $(n+2)$ aufsteigend angeordneten Extrema[1] von T_{n+1} zeigt,

$$\begin{array}{lcll}
[\tfrac{1}{2^n}T_{n+1}](s_0^{(n+1)}) &=& \tfrac{1}{2^n}, & \omega(s_0^{(n+1)}) < \tfrac{1}{2^n} \implies \mathcal{P}(s_0^{(n+1)}) > 0,\\
[\tfrac{1}{2^n}T_{n+1}](s_1^{(n+1)}) &=& -\tfrac{1}{2^n}, & \omega(s_1^{(n+1)}) > -\tfrac{1}{2^n} \implies \mathcal{P}(s_1^{(n+1)}) < 0,\\
[\tfrac{1}{2^n}T_{n+1}](s_2^{(n+1)}) &=& \tfrac{1}{2^n}, & \omega(s_2^{(n+1)}) < \tfrac{1}{2^n} \implies \mathcal{P}(s_2^{(n+1)}) > 0,\\
\vdots & & \vdots & \vdots \quad \vdots \qquad\quad \vdots
\end{array}$$

beziehungsweise allgemein

$$\mathcal{P}(s_k^{(n+1)})\,\mathcal{P}(s_{k-1}^{(n+1)}) \;<\; 0 \qquad \text{für} \quad k \;=\; 1,2,\dots,n+1.$$

Nun sind sowohl $T_{n+1}/2^n$ als auch ω jeweils Polynome vom Grad $= n+1$ und besitzen beide den führenden Koeffizienten 1, so dass notwendigerweise $\mathcal{P} \in \Pi_n$ gilt. Jedes Polynom vom Grad n mit $n+1$ paarweise verschiedenen Nullstellen muss jedoch identisch verschwinden, daher gilt $\mathcal{P} \equiv 0$ beziehungsweise

$$[\tfrac{1}{2^n}T_{n+1}] \;\equiv\; \omega,$$

was einen Widerspruch zur Annahme (1.24) darstellt. □

In Bild 1.1 ist der Verlauf des optimalen Polynoms vom Grad 10 dargestellt, und zum Vergleich ist noch das Polynom $\in \Pi_{10}$ mit äquidistanten Nullstellen und führendem Koeffizienten 1 abgebildet. Man beachte, dass sich bei dem optimalen Polynom die Abstände der einzelnen Nullstellen zueinander zu den beiden Rändern des Intervalls $[-1,1]$ hin verringern, was zu der Vermeidung von Oszillationen am Rand führt.

[1] diese sind in (1.18) angegeben

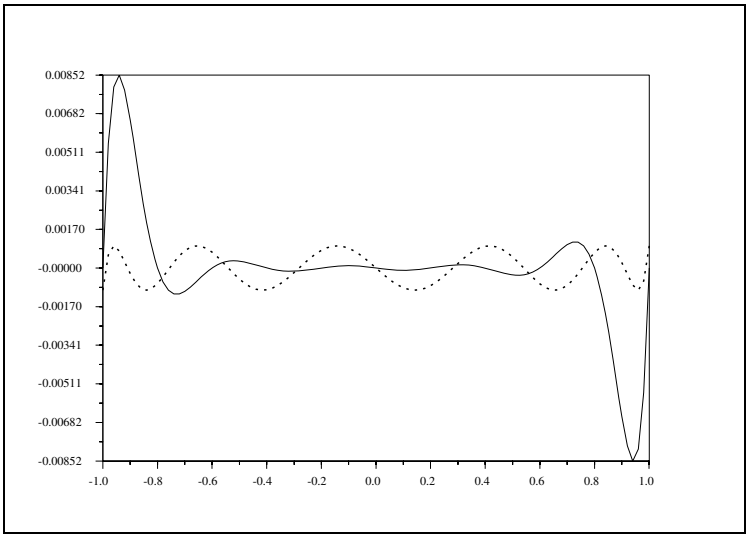


Bild 1.1: Darstellung von $\prod_{k=0}^{n}(x-x_k)$ und $\prod_{k=1}^{n+1}(x-t_k^{(n+1)})$ (letztere gestrichelt) für gleichabständige Nullstellen x_k beziehungsweise Tschebyscheff-Nullstellen $t_k^{(n+1)}$; für $n=10$

Der Fall $[a,b]=[-1,1]$ ist damit abgehandelt, und abschließend werden allgemeine Intervalle $[a,b]\subset\mathbb{R}$ betrachtet. Das nachfolgende Theorem[2] ist eine leichte Folgerung aus Theorem 1.24 verbunden mit der folgenden affin-linearen Transformation,

$$\psi:[-1,1]\ \to\ [a,b],\qquad t\ \mapsto\ \frac{1}{2}((b-a)t\ +\ a\ +\ b). \tag{1.25}$$

Theorem 1.25. *Mit der Funktion ψ aus (1.25) gilt die folgende Optimalitätseigenschaft,*

$$\begin{aligned}\min_{x_0,\ldots,x_n\in[a,b]}\ \max_{x\in[a,b]}\ &|(x-x_0)\ldots(x-x_n)|\\ &=\ \max_{x\in[a,b]}\ |(x-\psi(t_1^{(n+1)}))\ldots(x-\psi(t_{n+1}^{(n+1)}))| &(1.26)\\ &=\ \frac{(b-a)^{n+1}}{2\cdot 4^n}. &(1.27)\end{aligned}$$

[2] das auch noch bei anderen mathematischen Problemen zur Anwendung kommt

BEWEIS. Die Identität (1.27) ergibt sich folgendermaßen,

$$\begin{aligned}
&\max_{x\in[a,b]} |(x-\psi(t_1^{(n+1)}))\dots(x-\psi(t_{n+1}^{(n+1)}))| \\
&\quad= \max_{t\in[-1,1]} |(\psi(t)-\psi(t_1^{(n+1)}))\dots(\psi(t)-\psi(t_{n+1}^{(n+1)}))| \\
&\quad= \Big(\frac{b-a}{2}\Big)^{n+1} \max_{t\in[-1,1]} |(t-t_1^{(n+1)})\dots(t-t_{n+1}^{(n+1)})| \\
&\quad\overset{(*)}{=} \Big(\frac{b-a}{2}\Big)^{n+1} \frac{1}{2^n} \;=\; \frac{(b-a)^{n+1}}{2\cdot 4^n},
\end{aligned}$$

wobei man die Identität $(*)$ aus Theorem 1.24 erhält. Die Ungleichung "$\le$" in (1.26) ist offensichtlich richtig, und zum Beweis der Ungleichung "$\ge$" in (1.26) seien nun $x_0, x_1, \dots, x_n \in [a,b]$ beliebig. Dann gibt es eindeutig bestimmte Zahlen $y_0, y_1, \dots, y_n \in [-1,1]$ mit $\psi(y_j) = x_j$ für $j = 0,1,\dots,n$, und wie im ersten Teil des Beweises erhält man

$$\begin{aligned}
\max_{x\in[a,b]} |(x-x_0)\dots(x-x_n)| &= \max_{t\in[-1,1]} |\big(\psi(t)-\psi(y_0)\big)\dots\big(\psi(t)-\psi(y_n)\big)| \\
&= \Big(\frac{b-a}{2}\Big)^{n+1} \max_{t\in[-1,1]} |(t-y_0)\dots(t-y_n)| \\
&\overset{(*)}{\ge} \frac{(b-a)^{n+1}}{2\cdot 4^n},
\end{aligned}$$

wobei sich die Ungleichung $(*)$ erneut mit Theorem 1.24 ergibt. □

Abschließend werden in Bild 1.2 anhand einer Beispielfunktion die interpolierenden Polynome für gleichabständige und für "optimal" gewählte Stützstellen dargestellt.

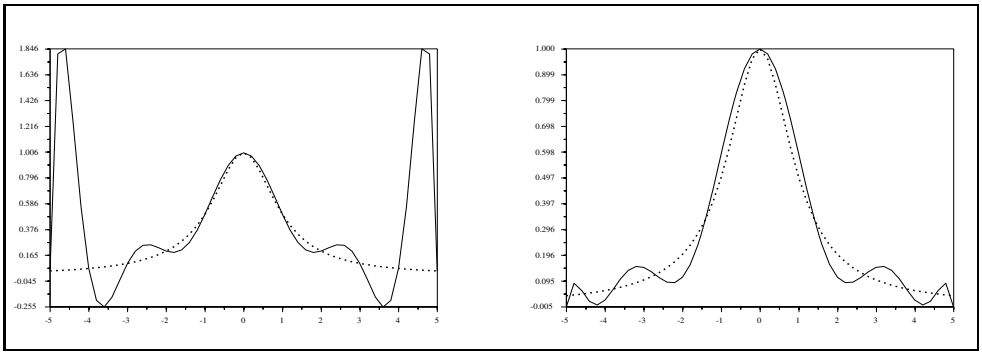


Bild 1.2: (Klassisches Beispiel von Runge) Interpolation der Funktion $f(x) = 1/(1+x^2)$, $x \in [-5,5]$ (gestrichelt) für äquidistante Stützstellen (links) beziehungsweise solchen Stützstellen, die sich aus linear transformierten Tschebyscheff-Nullstellen (rechts) ergeben; es ist $n = 6$. Man beachte die unterschiedlichen Skalierungen in den beiden Teilabbildungen links und rechts.

Weitere Themen und Literaturhinweise

Thematisch eng verwandt ist die *Hermite-Interpolation* (Aufgabe 1.3), die beispielsweise in Deuflhard/Hohmann [22], Mennicken/Wagenführer [71], Opfer [79], Schaback/Wendland [92], Schwarz/Klöckner [94], Freund/Hoppe [31], Weller [110] und in Werner [111] eingehend behandelt wird. Thematisch ebenfalls verwandt ist die *rationale Interpolation*, die beispielsweise in [71], [94], [31] und in [110] vorgestellt wird. Die Spline-Interpolation und die trigonometrische Interpolation sind Gegenstand der beiden folgenden Kapitel, und spezielle Darstellungen für die (vektorwertige) Polynominterpolation bezüglich äqudistanter Stützstellen sind in Abschnitt 8.3 angegeben.

Übungsaufgaben

Aufgabe 1.1. Für drei gegebene Funktionen $f,\ g,\ h : \mathbb{R}^N \supset \mathcal{D} \to \mathbb{R}$ und einen Häufungspunkt $x^* \in \mathbb{R}^N$ von $\mathcal{D}$ zeige man Folgendes:

(a) $f(x) = \mathcal{O}(g(x))$ für $\mathcal{D} \ni x \to x^* \quad \Longrightarrow \quad f(x) = \mathcal{O}(g(x))$ für $\mathcal{D} \ni x \to x^*$.

(b) $f(x) = \mathcal{O}(g(x)),\quad g(x) = \mathcal{O}(h(x))$ für $\mathcal{D} \ni x \to x^*$
$\Longrightarrow \quad f(x) = \mathcal{O}(h(x))$ für $\mathcal{D} \ni x \to x^*$.

(c) $f(x) = \mathcal{O}(\mathbf{1})$ für $\mathcal{D} \ni x \to x^* \quad \Longleftrightarrow \quad f(x) \to 0$ für $\mathcal{D} \ni x \to x^*$.

(d) $\mathcal{O}(f(x))\mathcal{O}(g(x)) = \mathcal{O}((fg)(x))$ für $\mathcal{D} \ni x \to x^*$.

(e) $\mathcal{O}(\mathcal{O}(f(x))) = \mathcal{O}(\mathcal{O}(f(x))) = \mathcal{O}(f(x))$ für $\mathcal{D} \ni x \to x^*$.

Aufgabe 1.2. Man zeige Folgendes: für gegebene paarweise verschiedene Stützstellen $x_0, x_1, \ldots, x_n \in \mathbb{R}$ ist die Abbildung $\mathbb{R}^{n+1} \to \Pi_n,\ (f_0, f_1, \ldots, f_n)^\top \mapsto \mathcal{P}$ (wobei $\mathcal{P}$ das jeweilige Interpolationspolynom gemäß (1.1) bezeichnet) linear.

Aufgabe 1.3. (Hermite-Interpolation) Man zeige: zu paarweise verschiedenen reellen Zahlen $x_0, x_1, \ldots, x_r$ sowie nichtnegativen ganzen Zahlen $m_0, m_1, \ldots, m_r \in \mathbb{N}_0$ mit $\sum_{j=0}^{r} m_j = n+1$ und vorgegebenen Zahlen $f_j^{(\nu)} \in \mathbb{R}$ für $\nu = 0, 1, \ldots, m_j - 1$ und $j = 0, 1, \ldots, r$ existiert genau ein Polynom $\mathcal{P} \in \Pi_n$ mit

$$\mathcal{P}^{(\nu)}(x_j) = f_j^{(\nu)} \qquad \text{für} \quad \begin{array}{l} \nu = 0, 1, \ldots, m_j - 1, \\ j = 0, 1, \ldots, r. \end{array}$$

Aufgabe 1.4. Zu paarweise verschiedenen reellen Zahlen $x_0, x_1, \ldots, x_n$ weise man für die zugehörigen lagrangeschen Basispolynome Folgendes nach:

(a) $$\sum_{k=0}^{n} L_k(x) \equiv 1;$$

(b) $$\sum_{k=0}^{n} L_k(0)\, x_k^s = \begin{cases} 1 & \text{für } s = 0, \\ 0 & \text{für } 1 \le s \le n, \\ (-1)^n x_0 x_1 \cdots x_n & \text{für } s = n+1. \end{cases}$$

Aufgabe 1.5. Zu den drei Stützpunkten $(x_j, \tan^2(x_j))$ für $j = 0, 1, 2$ mit den Stützstellen $x_0 = \pi/6$, $x_1 = \pi/4$ und $x_2 = \pi/3$ berechne man unter Verwendung des Schemas von Neville das zugehörige Interpolationspolynom.

Aufgabe 1.6. Zu gegebenen paarweise verschiedenen Stützstellen $x_0, x_1, \ldots, x_n \in \mathbb{R}$ und Stützwerten $f_0, f_1, \ldots, f_n \in \mathbb{R}$ weise man für die zugehörigen dividierten Differenzen Folgendes nach,

$$f[x_0, \ldots, x_n] \;=\; \sum_{j=0}^{n} f_j \Big/ \prod_{\substack{s=0 \\ s\neq j}}^{n} (x_j - x_s).$$

Aufgabe 1.7. Seien $(x_0, f_0), (x_1, f_1), \ldots, (x_n, f_n) \in \mathbb{R}^2$ und $(y_0, g_0), (y_1, g_1), \ldots, (y_n, g_n) \in \mathbb{R}^2$ Stützpunkte mit zugehörigen dividierten Differenzen $f[x_0, \ldots, x_n]$ und $g[y_0, \ldots, y_n]$. Man zeige: Wenn

$$\{(x_j, f_j),\, j = 0, 1, \ldots, n\} \;=\; \{(y_j, g_j),\, j = 0, 1, \ldots, n\}$$

erfüllt ist, so gilt $f[x_0, \ldots, x_n] \;=\; g[y_0, \ldots, y_n]$.

Aufgabe 1.8. Man bestimme in der newtonschen Darstellung das Interpolationspolynom zu den folgenden Stützpunkten:

j	0	1	2	3	4
x_j	-5	-2	-1	0	1
f_j	17	8	21	42	35

Im Folgenden bezeichnet $\mathcal{C}[a, b]$ die Menge der stetigen Funktionen $f : [a, b] \to \mathbb{R}$, und für $r = 1, 2, \ldots$ bezeichnet $\mathcal{C}^r[a, b]$ die Menge der r-fach stetig differenzierbaren Funktionen $f : [a, b] \to \mathbb{R}$.

Aufgabe 1.9. Man zeige, dass es zu jeder Funktion $f \in \mathcal{C}[a, b]$ und paarweise verschiedenen Stützstellen $x_0, x_1, \ldots, x_n \in [a, b]$ sowie für $\varepsilon > 0$ ein Polynom $\mathcal{P}$ gibt mit

$$\max_{x \in [a,b]} |\mathcal{P}(x) - f(x)| \;\leq\; \varepsilon, \qquad \mathcal{P}(x_j) \;=\; f(x_j) \qquad \text{für} \quad j \;=\; 0, 1, \ldots, n.$$

Aufgabe 1.10. Seien $\varphi_0, \varphi_1, \ldots, \varphi_n : \mathcal{C}[a, b] \to \mathbb{R}$ lineare Funktionale und $\mathcal{V} \subset \mathcal{C}[a, b]$ ein $(n + 1)$-dimensionaler linearer Teilraum.

(a) Man zeige, dass die verallgemeinerte Interpolationsaufgabe

$$\text{bestimme} \quad v \in \mathcal{V} \quad \text{mit} \quad \varphi_j(v) \;=\; \varphi_j(f) \qquad \text{für} \quad j \;=\; 0, 1, \ldots, n \tag{1.28}$$

genau dann für jedes $f \in \mathcal{C}[a, b]$ eindeutig lösbar ist, wenn die Funktion $f = 0$ nur $v = 0$ als verallgemeinerte Interpolierende besitzt.

(b) Sei die verallgemeinerte Interpolationsaufgabe (1.28) für jede Funktion $f \in \mathcal{C}[a, b]$ eindeutig lösbar und $\mathcal{L}_n : \mathcal{C}[a, b] \to \mathcal{V}$ der zugehörige Interpolationsoperator, das heißt, $\mathcal{L}_n f = v$. Man weise nach, dass $\mathcal{L}_n$ eine lineare Abbildung ist und für $f \in \mathcal{C}[a, b]$ gilt

$$\mathcal{L}_n f \;=\; f \quad \Longleftrightarrow \quad f \in \mathcal{V}.$$

Aufgabe 1.11. Für paarweise verschiedene Stützstellen $x_0, x_1, \ldots, x_n \in [a, b]$ bezeichne $\mathcal{L}_n : \mathcal{C}[a, b] \to \Pi_n$ den "Polynominterpolations-Operator", das heißt,

$$(\mathcal{L}_n f)(x_j) \;=\; f(x_j) \qquad \text{für} \quad j \;=\; 0, 1, \ldots, n \qquad (f \in \mathcal{C}[a, b]).$$

Man weise Folgendes nach:

$$\sup\big\{ \|\mathcal{L}_n f\|_\infty \;:\; f \in \mathcal{C}[a, b],\; \|f\|_\infty = 1 \big\} \;=\; \max_{x \in [a,b]} \left\{ \sum_{j=0}^{n} \prod_{\substack{s=0 \\ s\neq j}}^{n} \Big| \frac{x - x_s}{x_j - x_s} \Big| \right\},$$

wobei $\|\psi\|_\infty := \max\{|\psi(x)| \;:\; x \in [a, b]\}$ die Maximumnorm bezeichnet.

Aufgabe 1.12. Die *Tschebyscheff-Polynome der zweiten Art* $U_n \in \Pi_n$ sind definiert durch

$$U_0(x) := 1, \qquad U_1(x) := 2x, \qquad U_{n+1} := 2xU_n(x) - U_{n-1}(x), \quad n = 1, 2, \ldots .$$

(a) Man zeige $U_n(\cos\vartheta) = \dfrac{\sin((n+1)\vartheta)}{\sin\vartheta}$ für $\vartheta \in (0, \pi),\ n = 0, 1, \ldots$.

(b) Für $n = 0, 1, \ldots$ berechne man die beiden Werte $U_n(1)$ und $U_n(-1)$.

(c) Man zeige $T_n'(x) = nU_{n-1}(x)$ für $x \in [-1, 1],\ n = 1, 2, \ldots$.

Aufgabe 1.13 (*Numerische Aufgabe*). Mit einem Polynom vom Grad $\le n$ interpoliere man die Funktion $f(x) := 1/(25x^2 + 1),\ x \in [-1, 1]$,

- in äquidistanten Punkten $x_j = -1 + 2j/n, \quad j = 0, 1, \ldots, n$,
- in den Nullstellen $t_{j,n+1},\ j = 1, 2, \ldots, n+1$ des $(n+1)$-ten Tschebyscheff-Polynoms T_{n+1}.

Man wähle hierbei $n = 10$ und erstelle jeweils einen Ausdruck des Funktionsverlaufs.

2 Splinefunktionen

2.1 Einführende Bemerkungen

Bei der Polynominterpolation auf äquidistanten Gittern stellt sich mit wachsender Stützstellenzahl typischerweise ein oszillierendes Verhalten ein. Dies wird bei der in dem vorliegenden Abschnitt betrachteten Interpolation mittels Splinefunktionen vermieden. Für deren Einführung sei

$$\Delta \;=\; \big\{ a = x_0 < x_1 < \ldots < x_N = b \big\} \tag{2.1}$$

eine fest gewählte Zerlegung des Intervalls $[a,b]$, wobei man die Stützstellen $x_0, x_1, \ldots, x_N$ aus historischen Gründen auch als *Knoten* bezeichnet.

Definition 2.1. Eine *Splinefunktion der Ordnung* $\ell \in \mathbb{N}$ *zur Zerlegung* Δ ist eine Funktion $s \in \mathcal{C}^{\ell-1}[a,b]$, die auf jedem Intervall $[x_{j-1}, x_j]$ mit einem Polynom ℓ-ten Grades übereinstimmt. Der Raum dieser Splinefunktionen wird mit $S_{\Delta,\ell}$ bezeichnet, es gilt also

$$S_{\Delta,\ell} \;=\; \big\{ s \in \mathcal{C}^{\ell-1}[a,b] \,:\, s_{|[x_{j-1},x_j]} \;=\; p_{j\,|[x_{j-1},x_j]} \quad \text{für ein } p_j \in \Pi_\ell \quad (j \;=\; 1,\ldots,N) \big\}.$$

Anstelle Splinefunktion wird oft auch die Kurzbezeichnung Spline verwendet.

Bemerkung 2.2. Es ist offensichtlich $S_{\Delta,\ell}$ mit den üblichen Verknüpfungen ein linearer Raum. Für dessen Dimension gilt $\dim S_{\Delta,\ell} = N + \ell$, wie durch Abzählen der Freiheitsgrade intuitiv klar wird. △

In Bild 2.1 und Bild 2.2 sind Beispiele für lineare sowie quadratische Splines angegeben.

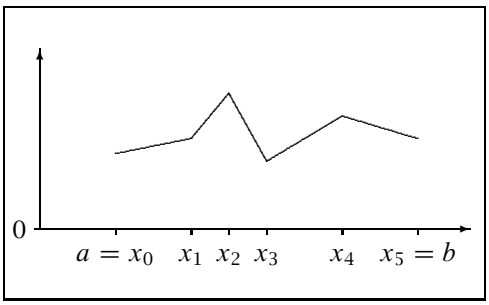


Bild 2.1: Ein linearer Spline auf $[a,b]$

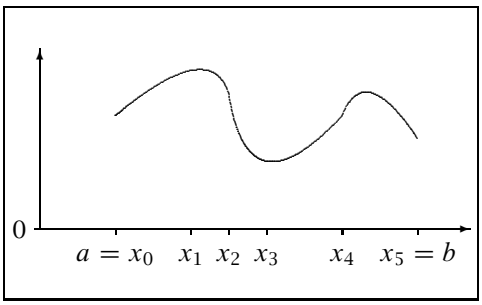


Bild 2.2: Ein quadratischer Spline auf $[a,b]$

Im Folgenden werden für interpolierende Splinefunktionen der Ordnung $\ell = 1$ (*lineare Splines* genannt) und Splinefunktionen der Ordnung $\ell = 3$ (*kubische Splines*) Algorithmen zur Berechnung sowie Fehlerabschätzungen hergeleitet. Splines der Ordnung $\ell = 2$ (*quadratische Splines*) spielen in der Praxis eine geringere Rolle und werden hier nicht behandelt.

2.2 Interpolierende lineare Splinefunktionen

2.2.1 Die Berechnung interpolierender linearer Splinefunktionen

Thema dieses Abschnitts ist die Berechnung linearer Splinefunktionen $s \in S_{\Delta,1}$ mit der Interpolationseigenschaft

$$s(x_j) = f_j \qquad \text{für} \quad j = 0, 1, \ldots, N, \tag{2.2}$$

wobei die Werte $f_0, f_1, \ldots, f_N \in \mathbb{R}$ vorgegeben sind. Für jeden Index $j \in \{0, 1, \ldots, N-1\}$ besitzt eine solche Funktion s auf dem Intervall $[x_j, x_{j+1}]$ die lokale Darstellung

$$s(x) = a_j + b_j(x - x_j) \qquad \text{für} \quad x \in [x_j, x_{j+1}], \tag{2.3}$$

und die Interpolationsbedingungen $s_j(x_j) = f_j$ und $s_j(x_{j+1}) = f_{j+1}$ ergeben unmittelbar

$$a_j = f_j, \qquad b_j = \frac{f_{j+1} - f_j}{x_{j+1} - x_j}. \tag{2.4}$$

Die Interpolationsbedingungen legen die Koeffizienten in dem allgemeinen Ansatz (2.3) in eindeutiger Weise fest und liefern den interpolierenden linearen Spline. Als Folgerung erhält man:

Theorem 2.3. *(Existenz und Eindeutigkeit des interpolierenden linearen Splines) Zu der Zerlegung $\Delta = \{a = x_0 < x_1 < \ldots < x_N = b\}$ und Werten $f_0, f_1, \ldots, f_N \in \mathbb{R}$ gibt es genau einen linearen Spline $s \in S_{\Delta,1}$ mit der Interpolationseigenschaft (2.2). Er besitzt die lokale Darstellung (2.3)–(2.4).*

Mit der Notation

$$\|u\|_\infty := \max_{x \in [a,b]} |u(x)|, \qquad u \in \mathcal{C}[a, b],$$

gilt für den Fehler bei der linearen Spline-Interpolation Folgendes:

Theorem 2.4. *Zu einer Funktion $f \in \mathcal{C}^2[a, b]$ sei $s \in S_{\Delta,1}$ der zugehörige interpolierende lineare Spline (siehe (2.2)). Dann gilt*

$$\|s - f\|_\infty \leq \tfrac{1}{8}\|f''\|_\infty h_{\max}^2 \qquad \textit{mit} \quad h_{\max} := \max_{j=0,\ldots,N-1} \{x_{j+1} - x_j\}.$$

BEWEIS. Für jeden Index $j \in \{1, 2, \ldots, N\}$ stimmt die Splinefunktion s auf dem Intervall $[x_{j-1}, x_j]$ mit demjenigen Polynom $\mathcal{P} \in \Pi_1$ überein, für das $\mathcal{P}(x_{j-1}) = f(x_{j-1})$ und $\mathcal{P}(x_j) = f(x_j)$ gilt, und Theorem 1.17 über den Fehler bei der Polynominterpolation liefert dann

$$\begin{aligned} |s(x) - f(x)| &\leq \frac{(x - x_{j-1})(x_j - x)}{2} \max_{\xi \in [x_{j-1}, x_j]} |f''(\xi)| \\ &\leq \frac{h_{\max}^2}{8} \|f''\|_\infty \qquad \text{für} \quad x \in [x_{j-1}, x_j]. \end{aligned}$$

Daraus folgt die angegebene Fehlerabschätzung. □

Bemerkung 2.5. Die wesentliche Aussage in Theorem 2.4 stellt $\|s - f\|_\infty = \mathcal{O}(h_{\max}^2)$ dar. △

2.3 Minimaleigenschaften kubischer Splinefunktionen

Im weiteren Verlauf wird die Interpolation mittels kubischer Splinefunktionen behandelt. Vor Behandlung der zugehörigen grundlegenden Themen wie Existenz, Eindeutigkeit, Berechnung und auftretender Fehler wird im vorliegenden Abschnitt zunächst eine für die Anwendungen wichtige Minimaleigenschaft interpolierender kubischer Splines vorgestellt (siehe Korollar 2.8 unten). Hierzu bezeichne im Folgenden

$$\|u\|_2 := \Big(\int_a^b |u(x)|^2 \, dx \Big)^{1/2}, \qquad u \in \mathcal{C}[a, b].$$

Lemma 2.6 (Holladay). *Wenn eine Funktion $f \in \mathcal{C}^2[a, b]$ und eine kubische Splinefunktion $s \in S_{\Delta,3}$ in den Knoten übereinstimmen,*

$$s(x_j) = f(x_j) \qquad \textit{für} \quad j = 0, 1, \ldots, N, \tag{2.5}$$

so gilt

$$\|f'' - s''\|_2^2 = \|f''\|_2^2 - \|s''\|_2^2 - 2([f' - s']s'')(x)|_{x=a}^{x=b}. \tag{2.6}$$

BEWEIS. Nach Definition von $\|\cdot\|_2$ gilt

$$\begin{aligned} \|f'' - s''\|_2^2 &= \int_a^b |f''(x) - s''(x)|^2 \, dx = \|f''\|_2^2 - 2\int_a^b (f''s'')(x) \, dx + \|s''\|_2^2 \\ &= \|f''\|_2^2 - 2\int_a^b ([f'' - s'']s'')(x) \, dx - \|s''\|_2^2, \end{aligned} \tag{2.7}$$

so dass man sich noch speziell mit dem mittleren Ausdruck in (2.7) zu befassen hat. Für $j = 1, 2, \ldots, N$ liefert partielle Integration

$$\begin{aligned}
&\int_{x_{j-1}}^{x_j} ([f'' - s'']s'')(x)\,dx \\
&= \big([f' - s']s''\big)(x)\big|_{x=x_{j-1}}^{x=x_j} - \int_{x_{j-1}}^{x_j} \big([f' - s']s'''\big)(x)\,dx \\
&= \text{——————} \text{ « } \text{——————} - \underbrace{\big([f - s]s'''\big)(x)|_{x=x_{j-1}+0}^{x=x_j-0}}_{=\,0} \\
&\qquad + \underbrace{\int_{x_{j-1}}^{x_j} ([f - s]s^{(4)})(x)\,dx}_{=\,0},
\end{aligned}$$

wobei der vorletzte Term aufgrund der Identität (2.5) verschwindet, und das letzte Integral verschwindet, da $s^{(4)} \equiv 0$ auf den Teilintervallen (x_{j-1}, x_j) gilt. Das Symbol —— « —— wird als Unterführungszeichen verwendet, es fungiert also als Platzhalter für den darüber stehenden Ausdruck. Anschließende Summation über $j = 1, 2, \ldots, N$ liefert aufgrund der Stetigkeit der Funktionen f', s', s'' auf dem Intervall $[a, b]$ die folgende Teleskopsumme und damit die Aussage des Lemmas,

$$\begin{aligned}
\int_a^b ([f'' - s'']s'')(x)\,dx &= \sum_{j=1}^{N} \big\{\big([f' - s']s''\big)(x_j) - \big([f' - s']s''\big)(x_{j-1})\big\} \\
&= \big([f' - s']s''\big)(b) - \big([f' - s']s''\big)(a).
\end{aligned}$$

□

Unter gewissen zusätzlichen Bedingungen vereinfacht sich die Aussage von Lemma 2.6:

Theorem 2.7. *Gegeben seien eine Funktion $f \in \mathcal{C}^2[a, b]$ und ein kubischer Spline $s \in S_{\Delta,3}$, die in den Knoten übereinstimmen, vergleiche (2.5). Dann gilt die Identität*

$$\|f''\|_2^2 - \|s''\|_2^2 = \|f'' - s''\|_2^2, \tag{2.8}$$

sofern eine der drei folgenden Bedingungen erfüllt ist:

(a) $s''(a) = s''(b) = 0$;

(b) $s'(a) = f'(a), \quad s'(b) = f'(b)$;

(c) $f'(a) = f'(b), \quad s'(a) = s'(b), \quad s''(a) = s''(b)$.

BEWEIS. In jedem der Fälle (a)–(c) verschwindet in (2.6) der Term $([f' - s']s'')(x)|_{x=a}^{x=b}$, und die Identität (2.6) geht dann über in die Identität (2.8). □

Korollar 2.8. *Zu gegebenen Werten $f_0, f_1, \ldots, f_N \in \mathbb{R}$ hat ein interpolierender kubischer Spline $s \in S_{\Delta,3}$ mit $s''(a) = s''(b) = 0$ unter allen hinreichend glatten interpolierenden Funktionen die geringste Krümmung, es gilt also*

$$\|s''\|_2 \le \|f''\|_2$$

für jede Funktion $f \in \mathcal{C}^2[a, b]$ mit $f(x_j) = f_j$ für $j = 0, 1, \ldots, N$.

BEWEIS. Die angegebene Abschätzung ergibt sich unmittelbar aus Theorem 2.7 für Splines mit der Eigenschaft (a) dort. □

Die in Korollar 2.8 angegebene Abschätzung gilt mit den entsprechenden Modifikationen in den zugehörigen Voraussetzungen auch für solche kubischen Splines, die den Bedingungen (b) oder (c) in Theorem 2.7 genügen.

Bemerkung 2.9. (1) Man weist über die Eigenschaft (2.8) leicht nach, dass jede der Bedingungen (a), (b) oder (c) in Theorem 2.7 die Eindeutigkeit des interpolierenden kubischen Splines impliziert (Aufgabe 2.3).

(2) Es stellt $\|f''\|_2$ lediglich eine Approximation an die mittlere Krümmung der Funktion f dar. Genauer ist die Krümmung von f in einem Punkt x gegeben durch $f''(x)/(1+f'(x)^2)^{3/2}$.

(3) Die in Korollar 2.8 vorgestellte Minimaleigenschaft stellt den Grund dafür dar, dass in der Praxis (beispielsweise bei der Konstruktion von Schiffsrümpfen oder der Festlegung von Schienenwegen) für die Interpolation oftmals kubische Splinefunktionen verwendet werden. △

In Bild 2.3 ist eine kubische Splinefunktion dargestellt.

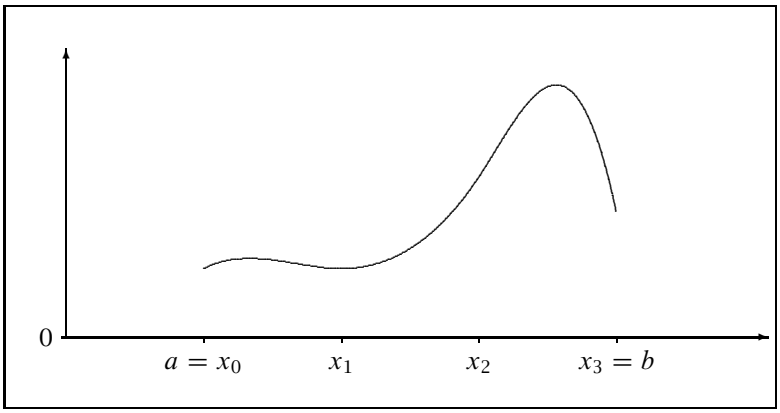


Bild 2.3: Ein kubischer Spline auf $[a,b]$ zu den Knoten $a = x_0 < x_1 < x_2 < x_3 = b$

2.4 Die Berechnung interpolierender kubischer Splinefunktionen

2.4.1 Vorüberlegungen

In dem vorliegenden Abschnitt wird die Berechnung interpolierender kubischer Splines behandelt. Ausgehend von dem lokalen Ansatz

$$\left.\begin{aligned} s(x) &= a_j + b_j(x-x_j) + c_j(x-x_j)^2 + d_j(x-x_j)^3 \\ &\text{für } x \in [x_j, x_{j+1}], \qquad j = 0,1,\ldots,N-1, \end{aligned}\right\} \tag{2.9}$$

für eine Funktion $s : [a,b] \to \mathbb{R}$ soll in diesem Abschnitt die Frage behandelt werden, wie man die Koeffizienten a_j, b_j, c_j und d_j für $j = 0,1,\ldots,N-1$ zu wählen hat, damit die Funktion s auf dem Intervall $[a,b]$ zweimal stetig differenzierbar ist[1] und darüber hinaus in den Knoten vorgegebene Werte $f_0, f_1, \ldots, f_N \in \mathbb{R}$ interpoliert,

$$s(x_j) = f_j \qquad \text{für} \quad j = 0,1,\ldots,N. \tag{2.10}$$

Das nachfolgende Lemma reduziert das genannte Problem auf die Lösung eines linearen Gleichungssystems, wobei die folgende Notation verwendet wird,

$$h_j := x_{j+1} - x_j \qquad \text{für} \quad j = 0,1,\ldots,N-1. \tag{2.11}$$

Lemma 2.10. *Falls $N+1$ reelle Zahlen $s_0'', s_1'', \ldots, s_N'' \in \mathbb{R}$ den folgenden $N-1$ gekoppelten Gleichungen*

$$h_{j-1}s_{j-1}'' + 2(h_{j-1}+h_j)s_j'' + h_j s_{j+1}'' = \overbrace{6\frac{f_{j+1}-f_j}{h_j} - 6\frac{f_j - f_{j-1}}{h_{j-1}}}^{=:\, g_j} \quad \textit{für} \quad j = 0,1,\ldots,N-1 \tag{2.12}$$

genügen, so liefert der lokale Ansatz (2.9) mit den Setzungen

$$c_j := \frac{s_j''}{2}, \qquad a_j := f_j, \qquad d_j := \frac{s_{j+1}'' - s_j''}{6h_j}, \tag{2.13}$$

$$b_j := \frac{f_{j+1}-f_j}{h_j} - \frac{h_j}{6}(s_{j+1}'' + 2s_j''), \tag{2.14}$$

für $j = 0,1,\ldots,N-1$ eine kubische Splinefunktion $s \in S_{\Delta,3}$, die die Interpolationsbedingungen (2.10) erfüllt.

BEWEIS. Mit den Notationen

$$p_j(x) = a_j + b_j(x-x_j) + c_j(x-x_j)^2 + d_j(x-x_j)^3 \in \Pi_3 \qquad (j = 0,1,\ldots,N-1)$$

erhält man für $j = 0,1,\ldots,N-1$ die folgenden Identitäten,

$$\begin{aligned} p_j(x_j) &= a_j = f_j, \\ p_{j+1}''(x_{j+1}) &= 2c_{j+1} = s_{j+1}'' = s_j'' + 6d_jh_j = p_j''(x_{j+1}) \qquad (j \le N-2) \end{aligned}$$

beziehungsweise

$$\begin{aligned} p_j(x_{j+1}) &= a_j + b_jh_j + c_jh_j^2 + d_jh_j^3 \\ &= f_j \;\text{---«---}\; \frac{s_j''}{2}h_j^2 + \frac{s_{j+1}''-s_j''}{6}h_j^2 \overset{(*)}{=} f_{j+1}, \end{aligned}$$

[1] und somit tatsächlich ein kubischer Spline ist

wobei die Identität $(*)$ eine Folgerung aus (2.14) darstellt. Die Stetigkeit der ersten Ableitung s' erhält man so,

$$p'_{j-1}(x_j) \;=\; b_{j-1} \;+\; 2c_{j-1}h_{j-1} \;+\; 3d_{j-1}h_{j-1}^2 \;\overset{(**)}{=}\; b_j \;=\; p'_j(x_j) \qquad (j \;=\; 1,2,\ldots,N-1),$$

wobei $(**)$ aus den Setzungen (2.13)–(2.14) und aus (2.12) resultiert. □

Bemerkung 2.11. (1) In der in Lemma 2.10 beschriebenen Situation bezeichnet man die $N+1$ reellen Zahlen $s''_0, s''_1, \ldots, s''_N \in \mathbb{R}$ als *Momente*. Diese stimmen mit den zweiten Ableitungen der Splinefunktion s in den Knoten x_j überein,

$$s''_j \;=\; s''(x_j) \qquad \text{für} \quad j \;=\; 0,1,\ldots,N.$$

(2) Mit Lemma 2.10 wird klar, dass sich die Koeffizienten in der Darstellung (2.9) unmittelbar aus den $N+1$ Momenten $s''_0, \ldots, s''_N$ ergeben. Diese $N+1$ Momente genügen den $N-1$ Bedingungen dieses Lemmas, womit also zwei Freiheitsgrade vorliegen. Aufgrund der Bedingungen (a)–(c) in Theorem 2.7 werden noch drei Möglichkeiten diskutiert, wofür abkürzend

$$s'_0 \;:=\; s'(x_0), \qquad\qquad s'_N \;:=\; s'(x_N)$$

gesetzt wird:

Natürliche Randbedingungen: $s''_0 \;=\; s''_N \;=\; 0$;

Vollständige Randbedingungen: $s'_0 \;=\; f'_0, \quad s'_N \;=\; f'_N$ für gegebene $f'_0,\ f'_N \in \mathbb{R}$;

Periodische Randbedingungen: $s'_0 \;=\; s'_N, \quad s''_0 \;=\; s''_N$.

Die Bezeichnung "natürliche Randbedingung" ist durch Korollar 2.8 gerechtfertigt.

(3) Division von (2.12) durch $3(h_{j-1}+h_j)$ führt auf die äquivalente Gleichung

$$\frac{h_{j-1}}{3(h_{j-1}\,+\,h_j)}s''_{j-1} \;+\; \frac{2}{3}s''_j \;+\; \frac{h_j}{3(h_{j-1}\,+\,h_j)}s''_{j+1} \;=\; 2\frac{f_{j+1}-f_j}{h_j(h_{j-1}\,+\,h_j)} \;-\; 2\frac{f_j-f_{j-1}}{h_{j-1}(h_{j-1}\,+\,h_j)}, \tag{2.15}$$

bei der die linke Seite eine Approximation an s''_j und die rechte Seite eine Differenzenapproximation an $f''(x_j)$ darstellt. Mehr hierzu finden Sie im Beweis von Lemma 2.15. △

In den folgenden Unterabschnitten 2.4.2–2.4.4 sollen die Bedingungen (2.12) für die Momente zusammen mit den unterschiedlichen Randbedingungen in Matrix-Vektor-Form angegeben werden.

2.4.2 Natürliche Randbedingungen

Die natürlichen Randbedingungen $s_0'' = s_N'' = 0$ führen zusammen mit (2.12) auf das folgende Gleichungssystem:

$$\begin{pmatrix} 2(h_0+h_1) & h_1 & 0 & \dots & 0 \\ h_1 & 2(h_1+h_2) & h_2 & \ddots & \vdots \\ 0 & h_2 & \ddots & \ddots & 0 \\ \vdots & \ddots & \ddots & \ddots & h_{N-2} \\ 0 & \dots & 0 & h_{N-2} & 2(h_{N-2}+h_{N-1}) \end{pmatrix} \begin{pmatrix} s_1'' \\ \vdots \\ s_{N-1}'' \end{pmatrix} = \begin{pmatrix} g_1 \\ \vdots \\ g_{N-1} \end{pmatrix}.$$

2.4.3 Vollständige Randbedingungen

Die vollständigen Randbedingungen

$$\begin{aligned} f_0' &\stackrel{!}{=} s_0' = b_0, \\ f_N' &\stackrel{!}{=} s_N' = b_{N-1} + 2c_{N-1}h_{N-1} + 3d_{N-1}h_{N-1}^2 \end{aligned}$$

führen mit (2.13)–(2.14) auf die beiden zusätzlichen Bedingungen

$$2h_0 s_0'' + h_0 s_1'' = -6f_0' + 6\frac{f_1 - f_0}{h_0} =: g_0, \tag{2.16}$$

$$h_{N-1}s_{N-1}'' + 2h_{N-1}s_N'' = 6f_N' - 6\frac{f_N - f_{N-1}}{h_{N-1}} =: g_N. \tag{2.17}$$

Diese Bedingungen (2.16)–(2.17) führen zusammen mit (2.12) auf das folgende Gleichungssystem:

$$\begin{pmatrix} 2h_0 & h_0 & 0 & \dots & \dots & 0 \\ h_0 & 2(h_0+h_1) & h_1 & \ddots & & \vdots \\ 0 & h_1 & \ddots & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & \ddots & \ddots & 0 \\ \vdots & & \ddots & \ddots & 2(h_{N-2}+h_{N-1}) & h_{N-1} \\ 0 & \dots & \dots & 0 & h_{N-1} & 2h_{N-1} \end{pmatrix} \begin{pmatrix} s_0'' \\ \vdots \\ s_N'' \end{pmatrix} = \begin{pmatrix} g_0 \\ \vdots \\ g_N \end{pmatrix}. \tag{2.18}$$

2.4.4 Periodische Randbedingungen

Die periodischen Randbedingungen

$$\begin{aligned} b_0 \;=\; s_0' \;&\stackrel{!}{=}\; s_N' \;=\; b_{N-1} \,+\, 2c_{N-1}h_{N-1} \,+\, 3d_{N-1}h_{N-1}^2, \\ s_0'' \;&\stackrel{!}{=}\; s_N'' \end{aligned}$$

führen mit (2.13)–(2.14) auf die zusätzliche Bedingung

$$2(h_{N-1} + h_0)s_0'' + h_0 s_1'' + h_{N-1}s_{N-1}'' \;=\; 6\frac{f_1 - f_0}{h_0} - 6\frac{f_N - f_{N-1}}{h_{N-1}} \;=:\; g_0. \tag{2.19}$$

Diese Bedingung (2.19) führt zusammen mit (2.12) auf das folgende Gleichungssystem:

$$\begin{pmatrix} 2(h_{N-1}+h_0) & h_0 & 0 & \dots & 0 & h_{N-1} \\ h_0 & 2(h_0+h_1) & h_1 & \ddots & & 0 \\ 0 & h_1 & \ddots & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & \ddots & \ddots & 0 \\ 0 & & \ddots & \ddots & \ddots & h_{N-2} \\ h_{N-1} & 0 & \dots & 0 & h_{N-2} & 2(h_{N-2}+h_{N-1}) \end{pmatrix} \begin{pmatrix} s_0'' \\ \vdots \\ s_{N-1}'' \end{pmatrix} = \begin{pmatrix} g_0 \\ \vdots \\ g_{N-1} \end{pmatrix}.$$

2.4.5 Existenz und Eindeutigkeit der betrachteten interpolierenden kubischen Splines

Für den Beweis der Existenz- und Eindeutigkeitsaussage für interpolierende kubische Splines wird das nachfolgende Lemma benötigt. Es wird hier in der nötigen Allgemeinheit formuliert wird, so dass es nochmals im Beweis des wichtigen Lemmas 2.15 angewandt werden kann. Vorbereitend wird die folgende Notation eingeführt,

$$\|z\|_\infty \;:=\; \max_{j=1,\dots,N} |z_j|, \qquad z \in \mathbb{R}^N.$$

Definition 2.12. Eine Matrix $A = (a_{jk}) \in \mathbb{R}^{N\times N}$ heißt *strikt diagonaldominant*, falls Folgendes gilt,

$$\sum_{\substack{k=1 \\ k\neq j}}^{N} |a_{jk}| \;<\; |a_{jj}| \qquad \text{für} \quad j \;=\; 1, 2, \dots, N.$$

Lemma 2.13. *Jede strikt diagonaldominante Matrix $A = (a_{jk}) \in \mathbb{R}^{N\times N}$ ist regulär und es gilt*

$$\|x\|_\infty \;\leq\; \max_{j=1,\dots,N} \left\{ \Big(|a_{jj}| - \sum_{\substack{k=1 \\ k\neq j}}^{N} |a_{jk}| \Big)^{-1} \right\} \|Ax\|_\infty \qquad \textit{für} \quad x \in \mathbb{R}^N. \tag{2.20}$$

BEWEIS. Für den Vektor $x \in \mathbb{R}^N$ sei der Index $j \in \{1, 2, \ldots, N\}$ so gewählt, dass $|x_j| = \|x\|_\infty$ gilt. Dann berechnet man

$$\begin{aligned}\|Ax\|_\infty &\geq |(Ax)_j| = \Big|\sum_{k=1}^{N} a_{jk}x_k\Big| \geq |a_{jj}||x_j| - \sum_{\substack{k=1\\k\neq j}}^{N} |a_{jk}||x_k| \\ &\geq |a_{jj}||x_j| - \sum_{\substack{k=1\\k\neq j}}^{N} |a_{jk}|\|x\|_\infty = \Big(|a_{jj}| - \sum_{\substack{k=1\\k\neq j}}^{N} |a_{jk}|\Big)\|x\|_\infty\end{aligned}$$

beziehungsweise

$$\|x\|_\infty \leq \Big(|a_{jj}| - \sum_{\substack{k=1\\k\neq j}}^{N} |a_{jk}|\Big)^{-1} \|Ax\|_\infty,$$

was die Ungleichung (2.20) nach sich zieht. Die Regularität der Matrix A folgt umgehend aus dieser Abschätzung (2.20). □

Offensichtlich ist jede der in den drei Abschnitten 2.4.2–2.4.4 betrachteten Matrizen strikt diagonaldominant. Als unmittelbare Folgerung aus dieser Beobachtung sowie den Lemmata 2.10 und 2.13 erhält man Folgendes:

Korollar 2.14. *Zur Zerlegung Δ und den Werten $f_0, f_1, \ldots, f_N \in \mathbb{R}$ gibt es jeweils genau einen interpolierenden kubischen Spline mit natürlichen beziehungsweise vollständigen (hier sind zusätzlich Zahlen $f'_0,\ f'_N \in \mathbb{R}$ vorgegeben) beziehungsweise periodischen Randbedingungen.*

2.5 Fehlerabschätzungen für interpolierende kubische Splines

Das folgende Lemma liefert eine Abschätzung für die Differenz der Momente von s und f in den Knoten x_j. Dabei werden wegen der einfacheren Vorgehensweise nur kubische Splines mit natürlichen Randbedingungen betrachtet. Vergleichbare Aussagen lassen sich auch für kubische Splines mit vollständigen oder periodischen Randbedingungen nachweisen (siehe beispielsweise Oevel [78], Mennicken/Wagenführer [71] und Freund/Hoppe [31]).

Lemma 2.15. *Zu einer gegebenen Funktion $f \in \mathcal{C}^4[a,b]$ mit $f''(a) = f''(b) = 0$ bezeichne $s \in S_{\Delta,3}$ den interpolierenden kubischen Spline[2] mit natürlichen Randbedingungen. Dann gilt*

$$\max_{j=1,\ldots,N-1} |s''(x_j) - f''(x_j)| \leq \frac{3}{4}\|f^{(4)}\|_\infty h_{\max}^2,$$
$$\text{mit } h_{\max} := \max_{j=0,\ldots,N-1}\{x_{j+1} - x_j\}.$$

[2] zur Zerlegung $\Delta = \{a = x_0 < \ldots < x_N = b\}$ und den Stützwerten $f_j = f(x_j)$ für $j = 0, 1, \ldots, N$

BEWEIS. Die Darstellung (2.15) für die Momente bedeutet in Matrixschreibweise

$$B \begin{pmatrix} s_1'' \\ \vdots \\ s_{N-1}'' \end{pmatrix} = \begin{pmatrix} \widehat{g}_1 \\ \vdots \\ \widehat{g}_{N-1} \end{pmatrix}, \tag{2.21}$$

wobei $\widehat{g}_j$ die rechte Seite von (2.15) bezeichnet, und die Matrix $B \in \mathbb{R}^{(N-1)\times(N-1)}$ besitzt die folgende Form,

$$B := \begin{pmatrix} \frac{2}{3} & \frac{h_1}{3(h_0+h_1)} & 0 & \dots & \dots & 0 \\ \frac{h_1}{3(h_1+h_2)} & \frac{2}{3} & \frac{h_2}{3(h_1+h_2)} & \ddots & & \vdots \\ 0 & \ddots & \ddots & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & \ddots & \ddots & 0 \\ \vdots & & \ddots & \frac{h_{N-3}}{3(h_{N-3}+h_{N-2})} & \frac{2}{3} & \frac{h_{N-2}}{3(h_{N-3}+h_{N-2})} \\ 0 & \dots & \dots & 0 & \frac{h_{N-2}}{3(h_{N-2}+h_{N-1})} & \frac{2}{3} \end{pmatrix},$$

mit der Notation $h_j = x_{j+1} - x_j$. Im Folgenden werden die Abbildungseigenschaften der Matrix B sowie die rechte Seite des Gleichungssystems (2.21) eingehender untersucht.

1. Durch Taylorentwicklung der Funktion f'' um den Punkt x_j erhält man die folgenden Darstellungen,

$$f''(x_{j-1}) = f''(x_j) - h_{j-1} f^{(3)}(x_j) + \frac{h_{j-1}^2}{2} f^{(4)}(\xi_j), \tag{2.22}$$

$$f''(x_{j+1}) = f''(x_j) + h_j f^{(3)}(x_j) + \frac{h_j^2}{2} f^{(4)}(\widehat{\xi}_j), \tag{2.23}$$

mit geeigneten Zwischenstellen ξ_j und $\widehat{\xi}_j$. Die Gleichung (2.22) wird dann mit dem Faktor $h_{j-1}/(3(h_{j-1}+h_j))$ und die Gleichung (2.23) mit dem Faktor $h_j/(3(h_{j-1}+h_j))$ multipliziert. Die beiden Ergebnisse werden anschließend addiert und resultieren in der folgenden Approximation an die zweite Ableitung $f''(x_j)$,

$$\begin{aligned} &\frac{h_{j-1}}{3(h_{j-1}+h_j)} f''(x_{j-1}) + \frac{2}{3} f''(x_j) + \frac{h_j}{3(h_{j-1}+h_j)} f''(x_{j+1}) \\ &\qquad = f''(x_j) + R_j + \delta_j, \\ &R_j := \frac{1}{3}(h_j - h_{j-1}) f^{(3)}(x_j), \\ &\delta_j := \frac{1}{6(h_{j-1}+h_j)} \left(h_{j-1}^3 f^{(4)}(\xi_j) + h_j^3 f^{(4)}(\widehat{\xi}_j) \right), \qquad j = 1, 2, \dots, N-1, \end{aligned}$$

beziehungsweise in Matrixschreibweise

$$B \begin{pmatrix} f''(x_1) \\ \vdots \\ f''(x_{N-1}) \end{pmatrix} = \begin{pmatrix} f''(x_1) \\ \vdots \\ f''(x_{N-1}) \end{pmatrix} + \begin{pmatrix} R_1 \\ \vdots \\ R_{N-1} \end{pmatrix} + \begin{pmatrix} \delta_1 \\ \vdots \\ \delta_{N-1} \end{pmatrix}. \tag{2.24}$$

2. Weiter ergibt eine Taylorentwicklung der Funktion f um den Punkt x_j die folgenden Darstellungen:

$$f(x_{j+1}) = f(x_j) + h_j f'(x_j) + \frac{h_j^2}{2} f''(x_j) + \frac{h_j^3}{6} f^{(3)}(x_j) + \frac{h_j^4}{24} f^{(4)}(\eta_j), \tag{2.25}$$

$$f(x_{j-1}) = f(x_j) - h_{j-1} f'(x_j) + \frac{h_{j-1}^2}{2} f''(x_j) - \frac{h_{j-1}^3}{6} f^{(3)}(x_j) + \frac{h_{j-1}^4}{24} f^{(4)}(\widehat{\eta}_j), \tag{2.26}$$

mit geeigneten Zwischenstellen η_j, $\widehat{\eta}_j \in [a,b]$. Eine Multiplikation der Gleichung (2.25) mit dem Faktor $2/h_j$ sowie Multiplikation der Gleichung (2.26) mit dem Faktor $2/h_{j-1}$ und jeweils anschließende Auflösung nach Termen mit $f(x_{j-1})$, $f(x_j)$ und $f(x_{j+1})$ führt auf die Gleichungen

$$\begin{aligned} 2\frac{f(x_{j+1}) - f(x_j)}{h_j} &= 2f'(x_j) + h_j f''(x_j) + \frac{h_j^2}{3} f^{(3)}(x_j) + \frac{h_j^3}{12} f^{(4)}(\eta_j), \\ -2\frac{f(x_j) - f(x_{j-1})}{h_{j-1}} &= -2f'(x_j) + h_{j-1} f''(x_j) - \frac{h_{j-1}^2}{3} f^{(3)}(x_j) + \frac{h_{j-1}^3}{12} f^{(4)}(\widehat{\eta}_j), \end{aligned}$$

und eine Addition dieser beiden Gleichungen sowie die anschließende Division durch $h_{j-1} + h_j$ resultiert in der folgenden Differenzenapproximation an die zweite Ableitung $f''(x_j)$,

$$\overbrace{2\frac{f_{j+1} - f_j}{h_j(h_{j-1} + h_j)} - 2\frac{f_j - f_{j-1}}{h_{j-1}(h_{j-1} + h_j)}}^{= \widehat{g}_j} = f''(x_j) + R_j + \widehat{\delta}_j, \quad j = 1, \ldots, N-1,$$

$$\widehat{\delta}_j := \frac{1}{12(h_{j-1} + h_j)} \big(h_j^3 f^{(4)}(\eta_j) + h_{j-1}^3 f^{(4)}(\widehat{\eta}_j) \big),$$

beziehungsweise in Vektorschreibweise

$$\begin{pmatrix} f''(x_1) \\ \vdots \\ f''(x_{N-1}) \end{pmatrix} = \begin{pmatrix} \widehat{g}_1 \\ \vdots \\ \widehat{g}_{N-1} \end{pmatrix} - \begin{pmatrix} R_1 \\ \vdots \\ R_{N-1} \end{pmatrix} - \begin{pmatrix} \widehat{\delta}_1 \\ \vdots \\ \widehat{\delta}_{N-1} \end{pmatrix}. \tag{2.27}$$

Verwendung der Identität (2.27) auf der rechten Seite von (2.24) und anschließende Subtraktion des Resultats von der Gleichung (2.21) führt auf eine Fehlerdarstellung der Form

$$B \begin{pmatrix} f''(x_1) - s''(x_1) \\ \vdots \\ f''(x_{N-1}) - s''(x_{N-1}) \end{pmatrix} = \begin{pmatrix} \delta_1 - \widehat{\delta}_1 \\ \vdots \\ \delta_{N-1} - \widehat{\delta}_{N-1} \end{pmatrix}.$$

Die Matrix B ist offensichtlich strikt diagonaldominant und somit aufgrund von Lemma 2.13 regulär, und mehr noch erhält man mit der Identität

$$\frac{2}{3} \;-\; \frac{h_j}{3(h_j+h_{j+1})} \;-\; \frac{h_{j+1}}{3(h_j+h_{j+1})} \;=\; \frac{1}{3}, \qquad j \;=\; 1,2,\ldots,N-1,$$

die Abschätzung

$$\begin{aligned} \max_{j=0,\ldots,N} |f''(x_j) - s''(x_j)| \;&\leq\; 3\max\left\{ |\delta_1| \;+\; |\widehat{\delta}_1|, \ldots, |\delta_{N-1}| \;+\; |\widehat{\delta}_{N-1}| \right\} \\ &\leq\; \tfrac{3}{4} h_{\max}^2 \, \| f^{(4)} \|_\infty, \end{aligned}$$

wobei in $(*)$ die Abschätzung

$$|\delta_j| \;+\; |\widehat{\delta}_j| \;\leq\; \frac{1}{4} \frac{h_{j-1}^3 \;+\; h_j^3}{h_{j-1} \;+\; h_j} \| f^{(4)} \|_\infty \;\leq\; \tfrac{1}{4} h_{\max}^2 \| f^{(4)} \|_\infty, \qquad j \;=\; 1,2,\ldots,N-1,$$

eingeht. Dies komplettiert den Beweis des Lemmas. □

Im folgenden Theorem werden die Approximationseigenschaften interpolierender kubischer Splines vorgestellt. Man beachte, dass die wesentliche Voraussetzung (2.28) für den Fehler der zweiten Ableitungen in den Knoten typischerweise erfüllt ist (siehe Lemma 2.15 und die davor angestellten Bemerkungen).

Theorem 2.16. *Sei $f \in \mathcal{C}^4[a,b]$, und sei $s \in S_{\Delta,3}$ ein interpolierender kubischer Spline*[3]*. Weiter bezeichne $h_j = x_{j+1} - x_j$ für $j = 0,1,\ldots,N-1$ und*

$$h_{\max} \;=\; \max_{j=0,\ldots,N-1} h_j, \qquad h_{\min} \;=\; \min_{j=0,\ldots,N-1} h_j.$$

Falls

$$\max_{j=0,\ldots,N} |s''(x_j) \;-\; f''(x_j)| \;\leq\; C \| f^{(4)} \|_\infty h_{\max}^2 \tag{2.28}$$

erfüllt ist mit einer Konstanten $C > 0$, so gelten mit der Zahl $c := \frac{h_{\max}}{h_{\min}}(C \;+\; \frac{1}{4})$ die folgenden Abschätzungen für jedes $x \in [a,b]$:

$$|s(x) \;-\; f(x)| \;\leq\; c \| f^{(4)} \|_\infty h_{\max}^4, \tag{2.29}$$

$$|s'(x) \;-\; f'(x)| \;\leq\; 2 \text{—«—} h_{\max}^3, \tag{2.30}$$

$$|s''(x) \;-\; f''(x)| \;\leq\; 2 \text{—«—} h_{\max}^2, \tag{2.31}$$

$$|s^{(3)}(x) \;-\; f^{(3)}(x)| \;\leq\; 2 \text{—«—} h_{\max} \qquad (x \neq x_j), \tag{2.32}$$

wobei der Ausdruck —«— hier jeweils für den Faktor $c\| f^{(4)} \|_\infty$ steht.

[3] zur Zerlegung $\Delta = \{a = x_0 < \ldots < x_N = b\}$ und den Stützwerten $f_j = f(x_j)$ für $j = 0,1,\ldots,N$

BEWEIS. Man weist zunächst die Fehlerabschätzung (2.32) für die dritten Ableitungen nach. Per Definition ist s'' auf jedem Intervall $[x_j, x_{j+1}]$ affin-linear, mithin gilt für $j = 0, 1, \ldots, N-1$

$$s^{(3)}(x) \equiv \frac{s''(x_{j+1}) - s''(x_j)}{h_j} \qquad \text{für} \quad x_j < x < x_{j+1}. \tag{2.33}$$

Eine Taylorentwicklung von f'' um den Punkt $x \in [x_j, x_{j+1}]$ liefert

$$\begin{aligned} f''(x_{j+1}) &= f''(x) + (x_{j+1} - x) f^{(3)}(x) + \frac{(x_{j+1} - x)^2}{2} f^{(4)}(\alpha_j), \\ f''(x_j) &= f''(x) + (x_j - x) f^{(3)}(x) + \frac{(x - x_j)^2}{2} f^{(4)}(\beta_j) \end{aligned}$$

mit gewissen Zwischenstellen $\alpha_j,\ \beta_j \in [x_j, x_{j+1}]$. Subtraktion der letzten beiden Gleichungen und anschließende Division durch h_j liefert

$$f^{(3)}(x) = \frac{f''(x_{j+1}) - f''(x_j)}{h_j} - \frac{(x_{j+1} - x)^2}{2h_j} f^{(4)}(\alpha_j) + \frac{(x - x_j)^2}{2h_j} f^{(4)}(\beta_j), \tag{2.34}$$

und die Subtraktion "(2.33)–(2.34)" ergibt

$$\begin{aligned} &s^{(3)}(x) - f^{(3)}(x) \\ &= \frac{s''(x_{j+1}) - f''(x_{j+1})}{h_j} - \frac{s''(x_j) - f''(x_j)}{h_j} + \frac{(x_{j+1} - x)^2 f^{(4)}(\alpha_j) - (x - x_j)^2 f^{(4)}(\beta_j)}{2h_j} \end{aligned}$$

und somit

$$\begin{aligned} |s^{(3)}(x) - f^{(3)}(x)| &\le \|f^{(4)}\|_\infty \frac{1}{\min\{h_0, \ldots, h_{N-1}\}} \left(C h_{\max}^2 + C h_{\max}^2 + \frac{h_{\max}^2}{2} \right) \\ &\le \underbrace{\frac{h_{\max}}{h_{\min}} \left(2C + \frac{1}{2} \right)}_{= 2c} \|f^{(4)}\|_\infty h_{\max}, \end{aligned}$$

wobei eine Abschätzung der Form

$$\begin{aligned} (x_{j+1} - x)^2 + (x - x_j)^2 &= (x_{j+1} - x_j)^2 - 2(x_{j+1} - x)(x - x_j) \\ &\le (x_{j+1} - x_j)^2 \le h_{\max}^2 \end{aligned}$$

für $x \in [x_j, x_{j+1}]$ eingeht. Die Fehlerabschätzung (2.32) für die dritten Ableitungen ist damit nachgewiesen.

Die weiteren Fehlerabschätzungen ergeben sich nun durch Integration. Zur Abschätzung der zweiten Ableitungen (2.31) wählt man zu einer gegebenen Zahl $x \in [a, b]$ den nächstgelegenen Knoten x_j, womit $|x - x_j| \le h_{\max}/2$ gilt. Der Hauptsatz der Differenzial- und Integralrechnung liefert

$$s''(x) - f''(x) = s''(x_j) - f''(x_j) + \int_{x_j}^{x} s^{(3)}(y) - f^{(3)}(y)\, dy$$

und somit

$$|s''(x) - f''(x)| \le C\|f^{(4)}\|_\infty h_{\max}^2 + 2c\|f^{(4)}\|_\infty |x - x_j| h_{\max} \le 2c\|f^{(4)}\|_\infty h_{\max}^2,$$

wobei noch die Eigenschaft $h_{\max}/h_{\min} \ge 1$ beziehungsweise $C \le c$ verwendet wurde. Damit ist auch (2.31) für die zweiten Ableitungen nachgewiesen. Zur Abschätzung (2.30) der ersten Ableitungen beachte man, dass die Stützstellen $a = x_0 < x_1 < \ldots < x_N = b$ Nullstellen der Funktion $s - f$ sind und somit die Funktion $s' - f'$ in jedem Teilintervall $[x_{j-1}, x_j]$ eine Nullstelle y_j besitzt. Wählt man zu einem gegebenen Punkt $x \in [a,b]$ die nächstgelegene Nullstelle y_j, so gilt $|x - y_j| \le h_{\max}$, und der Hauptsatz der Differenzial- und Integralrechnung liefert

$$\begin{aligned} |s'(x) - f'(x)| &= \left| \int_{y_j}^{x} s''(y) - f''(y)\, dy \right| \le 2c\|f^{(4)}\|_\infty h_{\max}^2 |x - y_j| \\ &\le 2c\|f^{(4)}\|_\infty h_{\max}^3. \end{aligned}$$

Damit ist auch die Fehlerabschätzung (2.30) für die ersten Ableitungen nachgewiesen. Abschließend wird der Fehler $s - f$ betrachtet. Für beliebiges $x \in [a,b]$ und den nächstgelegenen Knoten x_j erhält man

$$\begin{aligned} |s(x) - f(x)| &= \left| \int_{x_j}^{x} s'(y) - f'(y)\, dy \right| \le 2c\|f^{(4)}\|_\infty h_{\max}^3 |x - x_j| \\ &\le c\|f^{(4)}\|_\infty h_{\max}^4, \end{aligned}$$

womit auch die Fehlerabschätzung (2.29) nachgewiesen ist. □

Bemerkung 2.17. (a) Die wesentliche Aussage in Theorem 2.16 ist $\|s - f\|_\infty = \mathcal{O}(h_{\max}^4)$ für Zerlegungen Δ mit $h_{\max}/h_{\min} \le K$, wobei K eine von der Zerlegung Δ unabhängige Konstante bezeichnet. Diese Bedingung an den Quotienten $h_{\max}/h_{\min}$ stellt eine *Uniformitätsbedingung* an Δ dar.

(b) Konvergenz $\|s-f\|_\infty \to 0$ für $h_{\max} \to 0$ mit $h_{\max}/h_{\min} \le K$ erhält man auch unter geringeren Differenzierbarkeitseigenschaften. Für gleichmäßig stetige Funktionen $f : [a,b] \to \mathbb{R}$ wird ein entsprechendes Resultat in Mennicken/Wagenführer [71], Band 2 nachgewiesen. △

Weitere Themen und Literaturhinweise

Von einer gewissen Bedeutung sind in diesem Zusammenhang *B-Splines der Ordnung* $\ell \in \mathbb{N}_0$, bei denen es sich um spezielle nichtnegative und mit einem kompakten Träger versehene[4] Splinefunktionen der Ordnung ℓ aus den Räumen $S_{\Delta,\ell}$ handelt. Beispielsweise kann man mit ausgewählten B-Splines der Ordnung ℓ eine Basis für $S_{\Delta,\ell}$ erzeugen. Auf die Einführung von B-Splines wird hier im Sinne der angestrebten überschaubaren Darstellung verzichtet (ein paar weitere Anmerkungen finden Sie noch in Abschnitt 9.3.5) und stattdessen auf die folgende Auswahl von Lehrbüchern

[4] das heißt, diese verschwinden außerhalb eines endlichen Intervalls

verwiesen: de Boor [5], Deuflhard/Hohmann [22], Kress [63], Oevel [78], Mennicken /Wagenführer [71], Schaback/Wendland [92], Schwarz/Klöckner [94], Freund/Hoppe [31], Weller [110] und Werner [111]. Außerdem ist in diesem Zusammenhang die *Bézier-Interpolation* zu nennen, die beispielsweise in [63], [92], [94], [110] und [111] behandelt wird.

Übungsaufgaben

Aufgabe 2.1. Im Folgenden bezeichnet

$$\Delta \;=\; \{a \;=\; x_0 < x_1 < \ldots < x_N \;=\; b\} \tag{2.35}$$

wieder eine Zerlegung des Intervalls $[a,b]$. Weiter seien $f_0, f_1, \ldots, f_N \in \mathbb{R}$ gegebene Stützwerte, und s sei die zugehörige interpolierende *lineare* Splinefunktion. Im Folgenden bezeichnet $\mathcal{C}^1_\Delta[a,b]$ den Raum derjenigen stetigen Funktionen $f : [a,b] \to \mathbb{R}$, die stückweise stetig differenzierbar sind. Man zeige Folgendes:

(a) Für jede Funktion $f \in \mathcal{C}^1_\Delta[a,b]$ mit $f(x_j) = f_j$ für $j = 0,1,\ldots,N$ gilt:

(i) $\|f' - s'\|_2^2 \;=\; \|f'\|_2^2 \;-\; \|s'\|_2^2$.
(ii) Für eine beliebige (bzgl. Δ) lineare Splinefunktion ψ gilt $\|f' - s'\|_2 \;\leq\; \|f' - \psi'\|_2$.

(b) Die interpolierende lineare Splinefunktion s löst das Variationsproblem

$$\|f'\|_2 \to \min \qquad \text{für} \quad f \in \mathcal{C}^1_\Delta[a,b] \quad \text{mit} \quad f(x_j) \;=\; f_j \quad \text{für} \quad j \;=\; 0,1,\ldots,N.$$

Aufgabe 2.2. Gegeben seien eine Zerlegung (2.35) des Intervalls $[a,b]$ und Stützwerte $f_0, f_1, \ldots, f_N \in \mathbb{R}$.

(a) Man weise nach, dass es für jede Zahl $f_0' \in \mathbb{R}$ genau einen interpolierenden *quadratischen* Spline s gibt, der der Zusatzbedingung $s'(x_0) = f_0'$ genügt. Man gebe einen Algorithmus zur Berechnung von s an.

(b) Gesucht ist nun der interpolierende quadratische Spline s mit *periodischen* Randbedingungen $s'(x_0) = s'(x_N)$. Man treffe Aussagen über Existenz und Eindeutigkeit von s.

Aufgabe 2.3. Man weise die Aussage im ersten Teil von Bemerkung 2.9 nach.

Aufgabe 2.4. Auf dem Intervall $[-1,1]$ seien die Knoten $x_0 = -1$, $x_1 = 0$ und $x_2 = 1$ gegeben. Welche Eigenschaften eines natürlichen kubischen Splines bezüglich der zugehörigen Zerlegung besitzt die folgende Funktion, und welche besitzt sie nicht?

$$f(x) \;=\; \begin{cases} (x+1) \;+\; (x+1)^3 & \text{für } -1 \leq x \leq 0, \\ 4 + (x-1) \;+\; (x-1)^3 & \text{für } 0 < x \leq 1. \end{cases}$$

Aufgabe 2.5. Gegeben seien die Stützpunkte

k	0	1	2	3	4	5
x_k	-3	-2	-1	0	1	2
f_k	9	4	1	0	1	4

Man stelle das zugehörige lineare Gleichungssystem für die Momente der interpolierenden kubischen Splinefunktion mit natürlichen Randbedingungen auf.

Aufgabe 2.6. Gegeben seien eine äquidistante Zerlegung $\Delta = \{0 = x_0 < x_1 < \ldots < x_N = 1\}$ des Intervalls $[0,1]$, es gilt also $x_k = x_{k-1} + h$ für $k = 1, 2, \ldots, N$, mit $h = 1/N$. Man betrachte auf diesem Intervall die Funktion $f(x) = \sin(2\pi x)$ und die dazugehörende interpolierende kubische Splinefunktion $s \in S_{\Delta,3}$ mit natürlichen Randbedingungen. Wie groß muss die Zahl N gewählt werden, damit auf dem gesamten Intervall die Differenz zwischen s und f betragsmäßig kleiner als 10^{-12} ausfällt?

Aufgabe 2.7. Gegeben sei eine zweimal stetig differenzierbare Funktion $f : [a,b] \to \mathbb{R}$ und eine Zerlegung (2.35) des gegebenen Intervalls. Für den zugehörigen interpolierenden linearen Spline $s \in S_{\Delta,1}$ weise man mit Hilfe der taylorschen Formel die folgende Fehlerabschätzung nach:

$$|s'(x) - f'(x)| \leq \frac{1}{2}\|f''\|_\infty h_{\max} \qquad \text{für} \qquad x \in [a,b], \quad x \notin \{x_0, x_1, \ldots, x_N\},$$

wobei $h_{\max} := \max_{j=0,\ldots,N-1}\{x_{j+1} - x_j\}$ den maximalen Knotenabstand bezeichnet.

Aufgabe 2.8 (*Numerische Aufgabe*). Zur Interpolation beliebig verteilter Punkte (x_0, f_0), $(x_1, f_1), \ldots, (x_n, f_n) \in \mathbb{R}^2$ in der Ebene lassen sich *kubische Splinekurven* verwenden: Man bestimmt eine interpolierende kubische Splinefunktion s_1 zu den Werten $(t_0, x_0), (t_1, x_1), \ldots, (t_n, x_n) \in \mathbb{R}^2$ und eine zweite interpolierende kubische Splinefunktion s_2 zu den Werten $(t_0, f_0), (t_1, f_1), \ldots, (t_n, f_n) \in \mathbb{R}^2$. Hierbei wählt man

$$t_0 = 0, \qquad t_j = t_{j-1} + \sqrt{(x_j - x_{j-1})^2 + (f_j - f_{j-1})^2} \quad \text{für } j = 1, 2, \ldots, N.$$

Die gewünschte interpolierende kubische Splinekurve ist dann $(s_1(t), s_2(t))$ mit $t \in [0, t_N]$.

Diesen Ansatz wende man auf die folgenden Punkte an:

j	0	1	2	3	4	5	6	7	8
x_j	1.5	0.9	0.6	0.35	0.2	0.1	0.5	1.0	1.5
f_j	0.75	0.9	1.0	0.8	0.45	0.2	0.1	0.2	0.25

Dabei sollen die interpolierenden kubischen Splinefunktionen s_1 und s_2 natürliche Randbedingungen erfüllen. Man erstelle einen Ausdruck des sich ergebenden Kurvenverlaufs.

3 Diskrete Fouriertransformation und Anwendungen

In diesem Abschnitt wird zunächst die diskrete Fouriertransformation einführend behandelt, und anschließend werden einige Anwendungen präsentiert. Schließlich wird ein Verfahren zur "schnellen" diskreten Fouriertransformation vorgestellt. Zu den Anwendungen der diskreten Fouriertransformation gehört auch die trigonometrische Interpolation, daher wird das vorliegende Thema an dieser Stelle behandelt.

3.1 Diskrete Fouriertransformation

Definition 3.1. Zu einem gegebenem Datensatz von N komplexen Zahlen $f_0, f_1, \ldots, f_{N-1} \in \mathbb{C}$ wird der Datensatz $d_0, d_1, \ldots, d_{N-1}$ komplexer Zahlen definiert durch

$$d_k \;=\; \frac{1}{N}\sum_{j=0}^{N-1} f_j e^{-\mathrm{i}jk2\pi/N}, \qquad k \;=\; 0,1,\ldots,N-1 \qquad (\mathrm{i} \;=\; \sqrt{-1}) \quad (3.1)$$

als *diskrete Fouriertransformierte* oder auch als *diskrete Fourierkoeffizienten* von f_0, $f_1, \ldots, f_{N-1}$ bezeichnet. Es wird auch die folgende Notation verwendet,

$$\mathcal{F}(f_0,\ldots,f_{N-1}) \;:=\; (d_0,\ldots,d_{N-1}). \quad (3.2)$$

In Matrix-Vektorschreibweise ergibt sich die diskrete Fouriertransformierte durch die Multiplikation

$$\begin{pmatrix} d_0 \\ \vdots \\ d_{N-1} \end{pmatrix} \;=\; \frac{1}{N}\overline{V}\begin{pmatrix} f_0 \\ \vdots \\ f_{N-1} \end{pmatrix}, \quad (3.3)$$

wobei die Matrix $\overline{V} \in \mathbb{C}^{N\times N}$ konjugiert komplex ist zu der symmetrischen Matrix

$$V := (\omega^{kj})_{k,j=0..N-1} \;=\; \begin{pmatrix} 1 & 1 & 1 & \cdots & 1 \\ 1 & \omega & \omega^2 & \cdots & \omega^{N-1} \\ 1 & \omega^2 & \omega^4 & \cdots & \omega^{2(N-1)} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & \omega^{N-1} & \omega^{2(N-1)} & \cdots & \omega^{(N-1)^2} \end{pmatrix} \in \mathbb{C}^{N\times N}, \;\; \omega \;:=\; e^{\mathrm{i}2\pi/N}. \quad (3.4)$$

Im Folgenden bezeichnet $A^{\mathrm{H}} \in \mathbb{C}^{N\times M}$ die zu einer gegebenen Matrix $A \in \mathbb{C}^{M\times N}$ konjugiert komplexe und transponierte Matrix, $A^{\mathrm{H}} = \overline{A}^{\top}$. Im Fall $v = (v_1,\ldots,v_N)^{\top} \in \mathbb{C}^N$ beispielsweise bedeutet dies $v^{\mathrm{H}} = (\overline{v}_1,\ldots,\overline{v}_N)$.

Lemma 3.2. *Für die Spaltenvektoren der Matrix V in (3.4),*

$$v^{(k)} := (1, \omega^k, \omega^{2k}, \ldots, \omega^{(N-1)k})^\top \in \mathbb{C}^N, \qquad k = 0, 1, \ldots, N-1,$$

gilt

$$(v^{(k)})^{\mathrm{H}} v^{(\ell)} = \begin{cases} N & \text{für } k = \ell, \\ 0 & \text{für } k \neq \ell, \end{cases} \qquad k, \ \ell = 0, 1, \ldots, N-1; \tag{3.5}$$

die Spaltenvektoren von V sind also paarweise orthogonal zueinander.

BEWEIS. Im Fall $k = \ell$ erhält man wegen $|\omega| = 1$

$$(v^{(k)})^{\mathrm{H}} v^{(k)} = \sum_{s=0}^{N-1} \overline{\omega^{ks}} \omega^{ks} = \sum_{s=0}^{N-1} 1 = N,$$

und im Fall $k \neq \ell$ ergibt sich

$$\begin{aligned} (v^{(k)})^{\mathrm{H}} v^{(\ell)} &= \sum_{s=0}^{N-1} \overline{\omega^{ks}} \omega^{\ell s} = \sum_{s=0}^{N-1} \omega^{(\ell-k)s} = \sum_{s=0}^{N-1} (\omega^{(\ell-k)})^s \\ &= \frac{\omega^{(\ell-k)N} - 1}{\omega^{(\ell-k)} - 1} = \frac{e^{\mathrm{i}(\ell-k)2\pi} - 1}{e^{\mathrm{i}(\ell-k)2\pi/N} - 1} = 0, \end{aligned}$$

wobei der Nenner für $k, \ \ell \in \{0, 1, \ldots, N-1\}$ mit $k \neq \ell$ nicht verschwindet. □

Als unmittelbare Folgerung aus Lemma 3.2 erhält man das folgende Korollar.

Korollar 3.3. *1. (*Diskrete Fourierrücktransformation*) Für die Matrix $V \in \mathbb{C}^{N \times N}$ aus (3.4) gilt*

$$[\tfrac{1}{N} \overline{V}]^{-1} = V.$$

Jeder Datensatz $f_0, f_1, \ldots, f_{N-1}$ komplexer Zahlen lässt sich also aus seiner diskreten Fouriertransformierten $\mathcal{F}(f_0, \ldots, f_{N-1}) = (d_0, \ldots, d_{N-1})$ mittels

$$f_j = \sum_{k=0}^{N-1} d_k e^{\mathrm{i}jk2\pi/N}, \qquad j = 0, 1, \ldots, N-1, \tag{3.6}$$

zurückgewinnen. Es wird auch die folgende Notation verwendet,

$$\mathcal{F}^{-1}(d_0, \ldots, d_{N-1}) = (f_0, \ldots, f_{N-1}).$$

2. Mit der Notation aus (3.1) gilt $\sum_{k=0}^{N-1} |d_k|^2 = \frac{1}{N} \sum_{j=0}^{N-1} |f_j|^2$.

3.2 Anwendungen der diskreten Fouriertransformation

Für die nachfolgenden Betrachtungen ist eine Erweiterung der Definition (3.2) für Indizes $k \in \mathbb{Z}$ hilfreich,

$$d_k = \frac{1}{N} \sum_{j=0}^{N-1} f_j e^{-\mathrm{i}jk2\pi/N}, \qquad k \in \mathbb{Z}. \tag{3.7}$$

Wegen

$$e^{-\mathrm{i}j(k+N)\pi/N} = e^{-\mathrm{i}jk2\pi/N}\overbrace{e^{-\mathrm{i}jN\pi/N}}^{=1} = e^{-\mathrm{i}jk2\pi/N} \tag{3.8}$$

liegt in (3.7) Periodizität bezüglich k vor:

$$d_{k+N} \;=\; d_k, \qquad k \in \mathbb{Z}. \tag{3.9}$$

Mithilfe von (3.9) kann demnach aus den Werten $d_0, d_1, \ldots, d_{N-1}$ unmittelbar d_k für $k \in \mathbb{Z}$ ermittelt werden.

3.2.1 Fourierreihen

Jede riemann-integrierbare Funktion $f : [0, L] \to \mathbb{R}$ mit $f(0) = f(L)$ lässt sich in eine Fourierreihe entwickeln,

$$f(x) \;=\; \frac{a_0}{2} \;+\; \sum_{k=1}^{\infty} \Big[a_k \cos\big(k\tfrac{2\pi x}{L}\big) \;+\; b_k \sin\big(k\tfrac{2\pi x}{L}\big) \Big], \tag{3.10}$$

mit den *reellen Fourierkoeffizienten*

$$a_k \;=\; \frac{2}{L}\int_0^L f(y)\cos\big(k\tfrac{2\pi y}{L}\big)\,dy, \qquad b_k \;=\; \frac{2}{L}\int_0^L f(y)\sin\big(k\tfrac{2\pi y}{L}\big)\,dy, \tag{3.11}$$

für $k = 0, 1, \ldots$. Dabei konvergiert die Reihe in (3.10) im quadratischen Mittel. Mit der eulerschen Formel

$$e^{\pm\mathrm{i}k2\pi x/L} \;=\; \cos\big(k\tfrac{2\pi x}{L}\big) \;\pm\; \mathrm{i}\sin\big(k\tfrac{2\pi x}{L}\big), \qquad \mathrm{i} \;=\; \sqrt{-1} \qquad (k \in \mathbb{Z}),$$

erhält man die komplexe Fourierentwicklung

$$f(x) \;=\; \sum_{k=-\infty}^{\infty} c_k e^{\mathrm{i}k2\pi x/L} \tag{3.12}$$

mit den *komplexen Fourierkoeffizienten*

$$c_k \;=\; \frac{1}{L}\int_0^L f(y)e^{-\mathrm{i}k2\pi y/L}\,dy, \qquad k \in \mathbb{Z}, \tag{3.13}$$

wobei Konvergenz in (3.12) im quadratischen Mittel vorliegt. Zwischen den Koeffizienten in (3.11) und (3.13) besteht der folgende Zusammenhang (für $k \in \mathbb{N}_0$):

$$\begin{aligned} c_k &= \frac{a_k - \mathrm{i}b_k}{2}, & c_{-k} &= \frac{a_k + \mathrm{i}b_k}{2}, \\ a_k &= c_k + c_{-k}, & b_k &= \mathrm{i}(c_k - c_{-k}). \end{aligned}$$

3.2.2 Zusammenhang zwischen komplexen Fourierkoeffizienten und der diskreten Fouriertransformation

Im Folgenden wird beschrieben, inwiefern zu einer riemann-integrierbaren Funktion $f : [0, L] \to \mathbb{R}$ die diskreten Fourierkoeffizienten zur Approximation der komplexen Fourierkoeffizienten aus (3.13) herangezogen werden kann. Hierzu wird

$$f_j = f(x_j) \quad \text{mit} \quad x_j = jh,\ j = 0, 1, \ldots, N-1 \qquad (h = \frac{L}{N}) \tag{3.14}$$

betrachtet. Das in (3.13) auftretende Integral wird auf Teilintervallen durch eine einfache Rechteckregel approximiert:

$$\begin{aligned} c_k &= \frac{1}{L}\int_0^L f(y)e^{-\mathrm{i}k2\pi y/L}\,dy = \frac{1}{L}\sum_{j=0}^{N-1}\int_{jL/N}^{(j+1)L/N} f(y)e^{-\mathrm{i}k2\pi y/L}\,dy, \\ &\approx \frac{1}{L}\frac{L}{N}\sum_{j=0}^{N-1} f_j e^{-\mathrm{i}jk2\pi/N} = \frac{1}{N}\sum_{j=0}^{N-1} f_j e^{-\mathrm{i}jk2\pi/N} = d_k, \qquad k \in \mathbb{Z}, \end{aligned} \tag{3.15}$$

mit d_k aus (3.7). Solche "summierten Rechteckregeln" werden in Abschnitt 6.5.1 auf Seite 133 eingehender diskutiert.

Allerdings sind bei der Diskretisierung in (3.15) lediglich für betragsmäßig kleine Werte von k gute Approximationen zu erwarten, da nur dann die auftretenden Funktionen $e^{-\mathrm{i}k2\pi y/L}$ niederfrequent sind. Diese Einschränkung an k ist auch deswegen einleuchtend, da für glatte Funktionen f die Eigenschaft $c_k \to 0$ für $k \to \infty$ erfüllt ist, während für die Koeffizienten d_k aus (3.7) gemäß (3.9) eine Periodizität bzgl. k vorliegt.

Eine naheliegende Verwendung (für gerades N) ist

$$d_k \approx c_k, \quad k = 0, 1, \ldots, N/2-1, \qquad d_k = d_{k+N} \approx c_k, \quad k = -N/2, -N/2+1, \ldots, -1$$

beziehungsweise in kompakter Notation

$$(c_{-N/2}, \ldots, c_{-1}, c_0, \ldots, c_{N/2-1}) \approx (d_{N/2}, \ldots, d_{N-1}, d_0, \ldots, d_{N/2-1}).$$

Für die betrachtete Funktion f erhält man so die Näherung

$$f(x) \approx \sum_{k=-N/2}^{N/2-1} d_k e^{\mathrm{i}k2\pi x/L}. \tag{3.16}$$

Beispiel 3.4. Die *digitale Datenübertragung* liefert ein Beispiel für die praktische Anwendbarkeit der Eigenschaft (3.15). Hier ist es etwas vereinfacht dargestellt so, dass zu den N abgetasteteten Werten (3.14) eines analogen Signals $f : [0, L] \to \mathbb{R}$ die diskreten Fourierkoeffizienten $d_0, \ldots, d_{N-1}$ berechnet werden. Diese diskreten Fourierkoeffizienten werden anschließend an den gewünschten Zielort übermittelt, an dem mittels (3.16) (unter Zuhilfenahme der Periodizitätseigenschaft (3.9)) das analoge Signal f näherungsweise zurückgewonnen werden kann.

In diesem Zusammenhang spielen Glättung und Datenkompression eine Rolle. Zieht man nämlich zur Approximation einer Funktion f die Darstellung (3.16) heran, so werden dabei üblicherweise hochfrequente Anteile von f vernachlässigt, was einer *Glättung* der Funktion f gleichkommt. Dies lässt sich auch als *Datenkompression* interpretieren, da nur ein Teil der Fourierkoeffizienten bei der approximativen Rekonstruktion von f verwendet wird. △

3.2.3 Trigonometrische Interpolation, Teil 1

Zur Interpolation auf einem gegebenen Intervall $[0, L]$ mit $L > 0$ werden im Folgenden *trigonometrische Polynome* von der folgenden Form herangezogen,

$$p(x) \;=\; \sum_{k=-M}^{N-1-M} d_k \, e^{\mathrm{i}k2\pi x/L}, \qquad x \in \mathbb{R}, \tag{3.17}$$

mit einer Zahl $M \in \mathbb{N}_0$ und Koeffizienten $d_{-M}, d_{-M+1}, \ldots, d_{N-1-M}$. Weiter unten werden die Fälle $M = 0$ und $M = N/2$ etwas genauer betrachtet.

Interpolierende trigonometrische Polynome von der Form (3.17) erhält man folgendermaßen:

Theorem 3.5. *Zu äquidistanten Stützstellen $x_j = jL/N \in [0, L]$ und beliebigen Stützwerten $f_j \in \mathbb{C}$ für $j = 0, 1, \ldots, N-1$ mit $N \in \mathbb{N}$ besitzt das trigonometrische Polynom p aus (3.17) die Interpolationseigenschaft*

$$p(x_j) \;=\; f_j, \qquad j \;=\; 0, 1, \ldots, N-1, \tag{3.18}$$

falls die Koeffizienten $d_{-M}, d_{-M+1}, \ldots, d_{N-1-M}$ die Bedingung (3.7) erfüllen.

BEWEIS. Aus der Darstellung (3.17) erhält man unmittelbar

$$p(x_j) \;=\; \sum_{k=-M}^{N-1-M} d_k \, e^{\mathrm{i}jk2\pi/N} \;=\; \Bigg(\sum_{k=-M}^{-1} + \sum_{k=0}^{N-1-M} \Bigg) d_k \, e^{\mathrm{i}jk2\pi/N}, \tag{3.19}$$

und wir betrachten im Folgenden die erste dieser beiden Teilsummen. Hierfür erhält man Folgendes,

$$\sum_{k=-M}^{-1} d_k \, e^{\mathrm{i}jk2\pi/N} \;\overset{(*)}{=}\; \sum_{k=-M}^{-1} d_{k+N} \, e^{\mathrm{i}j(k+N)2\pi/N} \;\overset{(**)}{=}\; \sum_{k=N-M}^{N-1} d_k \, e^{\mathrm{i}jk2\pi/N}. \tag{3.20}$$

Dabei ergibt sich $(*)$ wie in (3.8) sowie mit der Setzung (3.7), und $(**)$ erhält man unmittelbar aus einer Umindizierung. Die Darstellungen (3.19) und (3.20) zusammen ergeben

$$p(x_j) \;=\; \sum_{k=0}^{N-1} d_k \, e^{\mathrm{i}jk2\pi/N} \;\overset{(*)}{=}\; f_j,$$

wobei sich die Darstellung $(*)$ aus der Definition in (3.1) ergibt. □

Ein trigonometrisches Polynom der Form (3.17), das zugleich die Interpolationseigenschaft (3.18) besitzt, bezeichnen wir als *trigonometrisches Interpolationspolynom*. Gemäß Theorem 3.5 gewinnt man die Koeffizienten eines solchen trigonometrischen Interpolationspolynoms mit der diskreten Fouriertransformierten (3.1) und anschließender periodischer Fortsetzung entsprechend (3.9).

3.2.4 Trigonometrische Interpolation, Teil 2

Wir betrachten nun trigonometrische Polynome $p = p_1$ aus (3.17) für den speziellen Fall $M = 0$:

$$p_1(x) \;=\; \sum_{k=0}^{N-1} d_k \, e^{\mathrm{i}k2\pi x/L}, \quad x \in \mathbb{R}. \tag{3.21}$$

Diese haben unter Umständen aufgrund eines oszillierenden Verhaltens schlechte Approximationseigenschaften, was anhand des folgenden Beispiels deutlich wird. In Abschnitt 3.2.5 wird allgemein beschrieben, warum dieses Verhalten nicht überraschend ist.

Beispiel 3.6. Man betrachte die Funktion $f : [0,1] \to \mathbb{R}$ definiert durch

$$f(x) \;=\; \begin{cases} x, & 0 \le x \le 1/2, \\ 1-x, & 1/2 \le x \le 1. \end{cases} \tag{3.22}$$

Für zwei verschiedene Werte von N sind in Bild 3.1 die zugehörigen trigonometrischen Interpolationspolynome der Form (3.21) dargestellt.

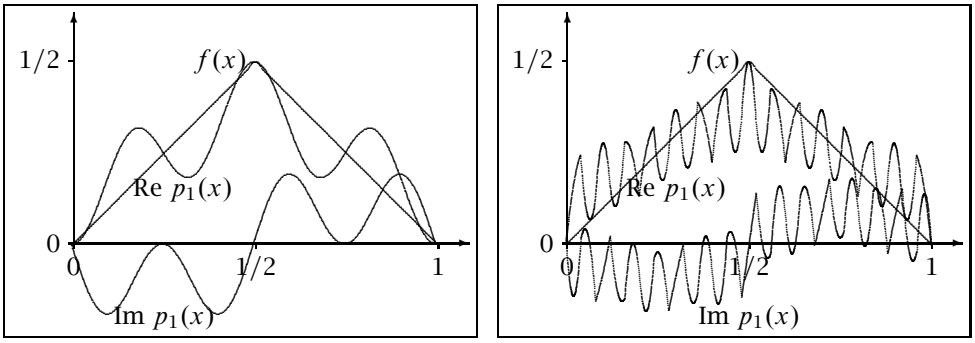


Bild 3.1: Darstellung der Funktion f und des Real- und Imaginärteils des trigonometrischen Interpolationspolynoms p_1 der Form (3.21); links für $N = 4$, rechts für $N = 16$ △

3.2.5 Trigonometrische Interpolation, Teil 3

Zur Gewinnung interpolierender trigonometrischer Polynome mit zugleich guten Approximationseigenschaften werden im Folgenden interpolierende trigonometrische

Polynome p von der Form (3.17) für den speziellen Fall $M = N/2$ betrachtet, wobei $N \in \mathbb{N}$ als gerade angenommen wird. In diesem speziellen Fall gilt $p = p_2$ mit

$$p_2(x) \;=\; \sum_{k=-N/2}^{N/2-1} d_k \, e^{\mathrm{i}k2\pi x/L}. \tag{3.23}$$

Man beachte, dass ein interpolierendes trigonometrisches Polynom der Form (3.23) mit der in (3.16) betrachteten Approximation übereinstimmt.

Ergeben sich die Stützwerte f_j aus den Werten einer hinreichend glatten periodischen Funktion an den Stützstellen x_j, so besitzt das trigonometrische Polynom p_2 aus (3.23) mit der Interpolationseigenschaft (3.18) auf dem gesamten Intervall $[0, L]$ gute Approximationseigenschaften, die in Theorem 3.8 unten präzisiert sind. Zunächst werden die Approximationseigenschaften anhand des folgenden Beispiels dargestellt.

Beispiel 3.7. Für die Funktion $f : [0,1] \to \mathbb{R}$ aus (3.22) sind in Bild 3.2 für zwei Werte von N jeweils die interpolierenden trigonometrischen Polynome p_2 aus (3.23), (3.18) dargestellt. △

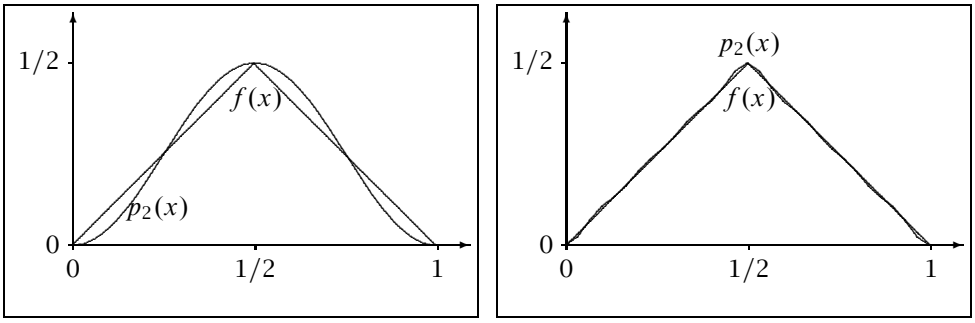


Bild 3.2: Darstellung der Funktionen f und p_2 aus (3.23), (3.18); links für $N = 4$, rechts für $N = 16$

Im Folgenden werden die Approximationseigenschaften des interpolierenden trigonometrischen Polynoms p_2 beschrieben.

Theorem 3.8. *Die Funktion $f : \mathbb{R} \to \mathbb{C}$ sei m-mal stetig differenzierbar und periodisch der Länge L, und es bezeichne $\|g\|_2 = (\int_0^L |g(x)|^2\,dx)^{1/2}$. Dann gilt für das trigonometrische Polynom p_2 aus (3.23) mit der Interpolationseigenschaft (3.18) (mit $f_j = f(x_j)$) die Fehlerabschätzung*

$$\|\,p_2 - f\,\|_2 \;\le\; c_m(\|\,f\,\|_2 \;+\; \|\,f^{(m)}\,\|_2)\,N^{-m}$$

mit einer gewissen Konstanten $c_m > 0$.

BEWEIS. Für einen elementaren Beweis unter expliziter Angabe der Konstanten c_m siehe Saranen/Vainikko [91]. □

Bemerkung 3.9. Es soll hier nochmals das interpolierende trigonometrische Polynom aus Abschnitt 3.2.4 betrachtet werden. Interpoliert ein solches trigonometrisches Polynom p_1 von der Form (3.21) auf dem Intervall $[0, L]$ an den äquidistanten Stützstellen $x_j = jL/N$, $j = 0, 1, \ldots, N-1$, eine gegebene m-mal stetig differenzierbare und L-periodische Funktion $f : \mathbb{R} \to \mathbb{C}$, so ist die Funktion $p_2(x) := p_1(x)e^{-\mathrm{i}N\pi x/L}$ von der Form (3.23) und interpoliert an den genannten Stützstellen die Funktion $f(x)e^{-\mathrm{i}N\pi x/L}$. Die letztgenannte Funktion oszilliert jedoch typischerweise stark. Genauer gilt

$$\frac{d^m}{dx^m}\big(f(x)e^{-\mathrm{i}N\pi x/L}\big) = e^{-\mathrm{i}N\pi x/L}\sum_{s=0}^{m}\binom{m}{s}\Big(\frac{-\mathrm{i}N\pi}{L}\Big)^s f^{(m-s)}(x)\Big),$$

wobei auf der rechten Seite dieser Gleichung der Term N^m dominiert und Theorem 3.8 hier somit lediglich

$$\|\,p_1 - f\,\|_2 = \|\,p_2 - f\,e^{-\mathrm{i}N\pi x/L}\,\|_2 = \mathcal{O}(\mathbf{1})$$

erwarten lässt. Dies wird durch Beispiel 3.6 bestätigt. △

Mit dem nächsten Beispiel wird der Effekt der Datenglättung demonstriert.

Beispiel 3.10. Für die Funktion $f : [0, 1] \to \mathbb{R}$ aus (3.22) ist in Bild 3.3 der mittels des trigonometrischen Interpolationspolynoms (3.23), (3.18) gewonnene Effekt der Datenglättung[1] veranschaulicht. △

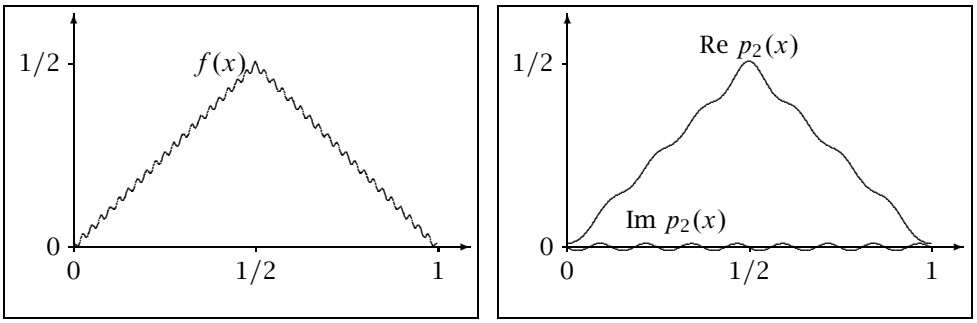


Bild 3.3: Links die Funktion f aus (3.22) mit kleinen aber hochfrequenten Störungen, und rechts das trigonometrische Interpolationspolynom p_2 für $N = 16$

3.2.6 Interpolierende reelle trigonometrische Polynome

Zur Interpolation der Stützpunkte (x_j, f_j) mit äquidistanten Stützstellen $x_j = jL/N \in [0, L]$ und *reellen* Zahlen $f_j \in \mathbb{R}$ für $j = 0, 1, \ldots, N-1$, werden im Folgenden

[1] siehe Beispiel 3.4

reelle trigonometrische Polynome der Form

$$T(x) \;=\; A_0 \;+\; 2\sum_{k=1}^{N/2-1}\big(A_k\cos\big(\tfrac{k2\pi x}{L}\big) \;+\; B_k\sin\big(\tfrac{k2\pi x}{L}\big)\big) \;+\; A_{N/2}\cos\big(\tfrac{N\pi x}{L}\big) \quad (3.24)$$

herangezogen mit geraden Zahlen N. Hierzu werden die folgenden Koeffizienten betrachtet:

$$A_k \;=\; \tfrac{1}{N}\sum_{j=0}^{N-1} f_j\cos\big(\tfrac{jk2\pi}{N}\big) \;\in\; \mathbb{R}, \qquad B_k \;=\; \tfrac{1}{N}\sum_{j=0}^{N-1} f_j\sin\big(\tfrac{jk2\pi}{N}\big) \;\in\; \mathbb{R}, \quad (3.25)$$
$$k \;=\; 0,1,\ldots,N/2.$$

Offenbar ist das trigonometrische Polynom T in (3.24) mit Koeffizienten A_k, B_k wie in (3.25) reellwertig. Das folgende elementare Lemma wird beim Beweis des nachfolgenden Theorems 3.12 benötigt und gibt darüber hinaus an, wie man die Zahlen in (3.25) mithilfe der diskreten Fouriertransformierten (3.1) erhält.

Lemma 3.11. *Zwischen den Zahlen A_k, B_k, $k = 0,1,\ldots,N-1$, in (3.25) einerseits und der diskreten Fouriertransformierten (3.3) bestehen die Zusammenhänge*

$$\begin{aligned} d_0 \;&=\; A_0, \qquad d_{-N/2} \;=\; A_{N/2},\\ d_k \;&=\; A_k \;-\; \mathrm{i}B_k, \qquad d_{-k} \;=\; A_k \;+\; \mathrm{i}B_k, \qquad k \;=\; 1,2,\ldots,N/2-1. \end{aligned} \quad (3.26)$$

BEWEIS. Gemäß (3.3) gilt

$$d_k \;=\; \tfrac{1}{N}\sum_{j=0}^{N-1} f_j\, e^{-\mathrm{i}jk2\pi/N} \;=\; \tfrac{1}{N}\sum_{j=0}^{N-1} f_j\big[\cos\big(\tfrac{jk2\pi x}{N}\big) \;-\; \mathrm{i}\sin\big(\tfrac{jk2\pi x}{N}\big)\big],$$
$$k \;=\; -N/2,\ldots,N/2-1,$$

woraus die angegebenen Identitäten unmittelbar folgen. □

Das folgende Theorem beschreibt die Interpolationseigenschaften des trigonometrischen Polynoms T aus (3.24)–(3.25).

Theorem 3.12. *Für das interpolierende trigonometrische Polynom p_2 mit (3.23), (3.18) und das reelle trigonometrische Polynom T aus (3.24)–(3.25) gilt* $\mathrm{Re}\; p_2(x) = T(x)$ *sowie $T(x_j) = f_j$ für $j = 0,1,\ldots,N-1$.*

BEWEIS. Mit dem trigonometrischen Polynom p_2 aus (3.23) gilt

$$\begin{aligned} p_2(x) \;&=\; d_0 \;+\; \sum_{k=1}^{N/2-1}\big[d_k\, e^{\mathrm{i}k2\pi x/L} \;+\; d_{-k}\, e^{-\mathrm{i}k2\pi x/L}\big] \;+\; d_{-N/2}e^{-\mathrm{i}N\pi x/L}\\ &\overset{(*)}{=}\; A_0 \;+\; \sum_{k=1}^{N/2-1}\big[(A_k \;-\; \mathrm{i}B_k)\,e^{\mathrm{i}k2\pi x/L} \;+\; (A_k \;+\; \mathrm{i}B_k)\,e^{-\mathrm{i}k2\pi x/L}\big] \;+\; A_{N/2}e^{-\mathrm{i}N\pi x/L}\\ &=\; A_0 \;+\; 2\sum_{k=1}^{N/2-1}\big[A_k\cos\big(\tfrac{k2\pi x}{L}\big) \;+\; B_k\sin\big(\tfrac{k2\pi x}{L}\big)\big] \;+\; A_{N/2}e^{-\mathrm{i}N\pi x/L}, \end{aligned}$$

wobei in $(*)$ noch Lemma 3.11 herangezogen wurde. Aus dieser Darstellung für p_2 ergeben sich unmittelbar die beiden Aussagen des Theorems. □

3.3 Schnelle Fourier-Transformation (FFT)

3.3.1 Einführende Bemerkungen

In diesem Abschnitt wird ein Verfahren zur "schnellen Fouriertransformation" (*Fast Fourier Transform*, kurz *FFT*) vorgestellt. Dieses Verfahren nutzt die spezielle Form der Transformation (3.1) aus und benötigt dabei lediglich $\mathcal{O}(N\log_2(N))$ komplexe Multiplikationen, wobei $\log_2$ den Logarithmus zur Basis 2 bezeichnet. Man beachte, dass die Berechnung der diskreten Fouriertransformierten (3.1) mittels einer Matrix-Vektor-Multiplikation entsprechend (3.3) insgesamt N^2 komplexe Multiplikationen erfordert.

3.3.2 Der grundlegende Zusammenhang

Von grundlegender Bedeutung für den FFT-Algorithmus ist das folgende Resultat.

Theorem 3.13. *Aus den diskreten Fouriertransformierten der beiden (komplexen) Datensätze $g_0, g_1, \ldots, g_{M-1}$ und $g_M, g_{M+1}, \ldots, g_{2M-1}$ der Längen M lässt sich die diskrete Fouriertransformierte des Datensatzes $g_0, g_M, g_1, g_{M+1}, \ldots, g_{M-1}, g_{2M-1}$ der Länge $2M$ folgendermaßen bestimmen:*

$$\begin{aligned}
&\tfrac{1}{2}\big(\mathcal{F}_k(g_0, g_1, \ldots, g_{M-1}) \;+\; e^{-\mathrm{i}k\pi/M}\,\mathcal{F}_k(g_M, g_{M+1}, \ldots, g_{2M-1})\big)\\
&\qquad = \;\; \mathcal{F}_k(g_0, g_M, g_1, g_{M+1}, .., g_{M-1}, g_{2M-1}) \quad \textit{für } k \;=\; 0, 1, \ldots, M-1,\\
&\tfrac{1}{2}\big(\mathcal{F}_k(g_0, g_1, \ldots, g_{M-1}) \;-\; e^{-\mathrm{i}k\pi/M}\,\mathcal{F}_k(g_M, g_{M+1}, \ldots, g_{2M-1})\big)\\
&\qquad = \;\; \mathcal{F}_{M+k}(g_0, g_M, g_1, g_{M+1}, \ldots, g_{M-1}, g_{2M-1}) \quad \textit{für } k \;=\; 0, 1, \ldots, M-1.
\end{aligned}$$

Hierbei bezeichnen $\mathcal{F}_k$ beziehungsweise $\mathcal{F}_{M+k}$ die k-te beziehungsweise $(M+k)$-te Komponente von $\mathcal{F}$.

BEWEIS. Für $k = 0, 1, \ldots, M-1$ gilt

$$\begin{aligned}
&\mathcal{F}_k(g_0, g_M, g_1, g_{M+1}, \ldots, g_{M-1}, g_{2M-1})\\
&\qquad = \;\; \frac{1}{2M}\Big(\sum_{j=0}^{M-1} g_j\, e^{-\mathrm{i}2jk2\pi/2M} \;+\; \sum_{j=0}^{M-1} g_{M+j}\, e^{-\mathrm{i}(2j+1)k2\pi/2M}\Big)\\
&\qquad = \;\; \frac{1}{2M}\Big(\sum_{j=0}^{M-1} g_j\, e^{-\mathrm{i}jk2\pi/M} \;+\; e^{-\mathrm{i}k\pi/M}\sum_{j=0}^{M-1} g_{M+j}\, e^{-\mathrm{i}jk2\pi/M}\Big).
\end{aligned}$$

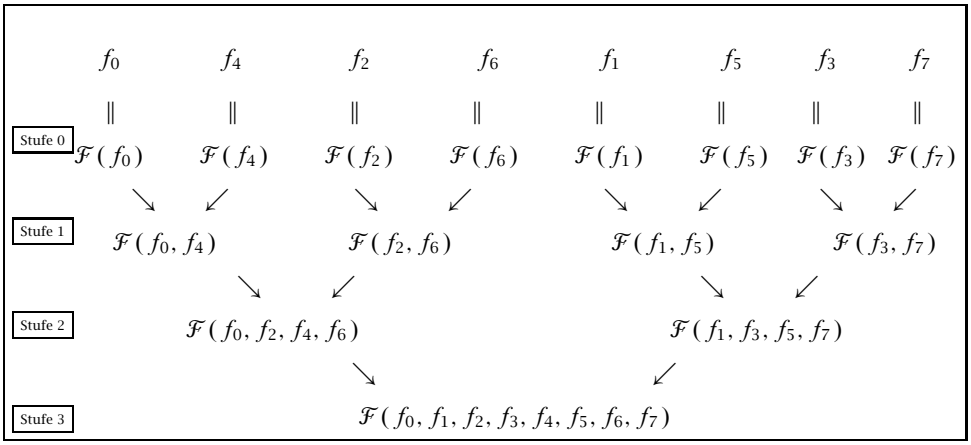
Die zweite Gleichung in Theorem 3.13 erhält man völlig analog, wobei noch

$$e^{-\mathrm{i}j(k+M)2\pi/2M} \;=\; e^{-\mathrm{i}jk2\pi/2M}\, e^{-\mathrm{i}j\pi} \;=\; (-1)^j\, e^{-\mathrm{i}jk2\pi/2M}$$

berücksichtigt wird. □

Für den Fall $N = 2^q$ mit $N \in \mathbb{N}$ kann die in Theorem 3.13 vorgestellte Eigenschaft genutzt werden, um die diskrete Fouriertransformierte eines komplexen Datensatzes $f_0, \ldots, f_{N-1}$ zu bestimmen. Dies soll zunächst anhand des nachfolgenden Beispiels erläutert werden.

Beispiel 3.14. In Schema 3.1 ist für den Spezialfall $N = 2^3$ dargestellt, wie man für $r = 0, 1, 2$ ausgehend von der Stufe r mit den diskreten Fouriertransformierten von Datensätzen der Länge 2^r zu den diskreten Fouriertransformierten von Datensätzen der Länge 2^{r+1} in der Stufe $r+1$ gelangt. Im Folgenden wird beschrieben, wie man in



Schema 3.1: Darstellung der schnellen Fouriertransformation im Fall $N = 2^3$

der Stufe 0 die angegebene Zuordnung $f_0, f_4, f_2, f_6, f_1, f_5, f_3, f_7$ auf die Positionen 0–7 erhält; für jede einzelne Positionsnummer $n \in \{0, 1, \ldots, 7\}$ wird die jeweilige Binärdarstellung $n = b_2 2^2 + b_1 2^1 + b_0 2^0$ ermittelt und in dieser anschließend die Reihenfolge der Binärziffern umgedreht. Die zugehörige Dezimalzahl $b_0 2^2 + b_1 2^1 + b_2 2^0$ liefert dann den gesuchten Index von f. Dieses Vorgehen der Bit-Umkehr ist in Tabelle 3.1 dargestellt. Die Begründung dafür, warum dieses Vorgehen die richtige Zuordnung liefert, wird in Abschnitt 3.3.4 nachgereicht. △

Position		Index von f	
Dezimal $\triangleq$	Binär	Binär revers $\triangleq$	Dezimal
0	000	000	0
1	001	100	4
2	010	010	2
3	011	110	6
4	100	001	1
5	101	101	5
6	110	011	3
7	111	111	7

Tabelle 3.1: Darstellung der Bit-Umkehr im Fall $N = 2^3$. Die Positionsangaben und Indizes betreffen von links aus gesehen die erste Zeile in Schema 3.1.

Für die Berechnung von $\mathcal{F}(f_0, f_1, \ldots, f_{N-1})$ lässt sich das Ergebnis aus Theorem 3.13 sowohl rekursiv (ohne Bit-Umkehr) als auch iterativ umsetzen. Im Folgenden soll der iterative Weg verfolgt werden, bei dem weniger Speicherplatz erforderlich ist. Die allgemeine Vorgehensweise hierzu ist in Definition 3.19 weiter unten beschrieben. Vorbereitend wird die Bit-Umkehr eingehender behandelt.

3.3.3 Bit-Umkehr

Im Folgenden wird die Bit-Umkehr in der allgemeinen Situation $N = 2^q$ betrachtet.

Definition 3.15. Für $q \in \mathbb{N}_0$ sei $n = \sum_{\ell=0}^{q-1} b_\ell 2^\ell$ die eindeutige Binärdarstellung einer Zahl $n \in \mathcal{M}_q = \{0, 1, \ldots, 2^q - 1\}$ mit Binärziffern (Bits) $b_\ell \in \{0, 1\}$. Die durch

$$\sigma_q : \mathcal{M}_q \;\to\; \mathcal{M}_q, \qquad \sum_{\ell=0}^{q-1} b_\ell\, 2^\ell \;\mapsto\; \sum_{\ell=0}^{q-1} b_{q-1-\ell}\, 2^\ell$$

definierte Abbildung bezeichnet man als *Bit-Umkehr*.

Die Situation $q = 0$ in Definition 3.15 wird dabei lediglich aus technischen Gründen zugelassen und bedeutet $M_0 = \{0\}$ und $\sigma_0(0) = 0$.

Bemerkung 3.16. Es gilt offensichtlich

$$\sigma_q\Big(\sum_{\ell=0}^{q-1} b_\ell\, 2^\ell\Big) \;=\; \sum_{\ell=0}^{q-1} b_\ell\, 2^{q-1-\ell}. \qquad \triangle$$

Das folgende Theorem liefert eine Vorgehensweise, mit der sich die Bit-Umkehr effizient realisieren lässt. Die Werte $\sigma_q(0), \sigma_q(1), \ldots, \sigma_q(2^q - 1)$ können damit mittels zwei geschachtelter for-Schleifen und ohne Durchführung von Multiplikationen berechnet werden.

Theorem 3.17. *Für die Bit-Umkehr $\sigma_q : \mathcal{M}_q \to \mathcal{M}_q$ gilt*

$$\sigma_q(2^r + n) \;=\; \sigma_q(n) \;+\; 2^{q-1-r}, \qquad \begin{array}{l} n = 0, 1, \ldots, 2^r - 1, \\ r = 0, 1, \ldots, q - 1. \end{array}$$

Beweis. Sei $r \in \{0, 1, \ldots, q-1\}$. Für $n \in \{0, 1, \ldots, 2^r - 1\}$ existiert eine eindeutige Binärdarstellung von der Form

$$n \;=\; \sum_{\ell=0}^{r-1} b_\ell 2^\ell,$$

und dann gilt $n + 2^r = \sum_{\ell=0}^{r-1} b_\ell 2^\ell \;+\; 2^r$ beziehungsweise

$$\sigma_q(n \;+\; 2^r) \;=\; \underbrace{\sum_{\ell=0}^{r-1} b_\ell 2^{q-1-\ell}}_{=\,\sigma_q(n)} + 2^{q-1-r}. \qquad \square$$

Für das Verständnis der Funktionsweise der Bit-Umkehr in der allgemeinen Situation $N = 2^q$ ist noch das folgende Resultat von Bedeutung.

Lemma 3.18. *Die Bit-Umkehr $\sigma_q : \mathcal{M}_q \to \mathcal{M}_q$ ist bijektiv mit $\sigma_q^{-1} = \sigma_q$. Weiter gilt für $r = 0, 1, \ldots$:*

$$\begin{aligned} \sigma_r(n) &= \sigma_{r+1}(2n), && n \in \mathcal{M}_r, \\ 2^r + \sigma_r(n) &= \sigma_{r+1}(2n+1), && \text{—«—}. \end{aligned}$$

BEWEIS. Ist elementar und wird hier nicht geführt (Aufgabe 3.7). □

3.3.4 Der FFT-Algorithmus in der Situation $N = 2^q$

Ausgehend von beliebigen gegebenen komplexen Zahlen $g_0, g_1, \ldots, g_{N-1} \in \mathbb{C}$ mit

$$N = 2^q \qquad \text{mit } q \in \mathbb{N}$$

führt der in Theorem 3.13 beschriebene Zusammenhang auf die in dem folgenden Algorithmus 3.19 beschriebenen Vorgehensweise. Wie sich herausstellen wird (siehe Korollar 3.23), stimmt der sich dabei ermittelte Vektor $\mathbf{d}^{[q,0]} \in \mathbb{C}^N$ mit der diskreten Fouriertransformierten $\mathcal{F}(g_{\sigma_q(0)}, \ldots, g_{\sigma_q(2^q-1)})$ überein. Damit wird dann auch unmittelbar klar, wie man die Zahlen $g_0, g_1, \ldots, g_{N-1} \in \mathbb{C}$ letztlich zu wählen hat, so dass der Vektor $\mathbf{d}^{[q,0]} \in \mathbb{C}^N$ tatsächlich mit der zu bestimmenden diskreten Fouriertransformierten $\mathcal{F}(f_0, \ldots, f_{N-1})$ eines gegebenen Datensatzes von N komplexen Zahlen $f_0, \ldots, f_{N-1}$ übereinstimmt.

Algorithmus 3.19 (FFT). Ausgehend von Zahlen $\mathbf{d}^{[0,j]} = g_j \in \mathbb{C},\ j = 0, \ldots, 2^q - 1$ bestimme man für Stufen $r = 1, 2, \ldots, q$ in der r-ten Stufe insgesamt 2^{q-r} Vektoren der Länge 2^r

$$\mathbf{d}^{[r,0]}, \mathbf{d}^{[r,1]}, \ldots, \mathbf{d}^{[r,2^{q-r}-1]} \in \mathbb{C}^{2^r}$$

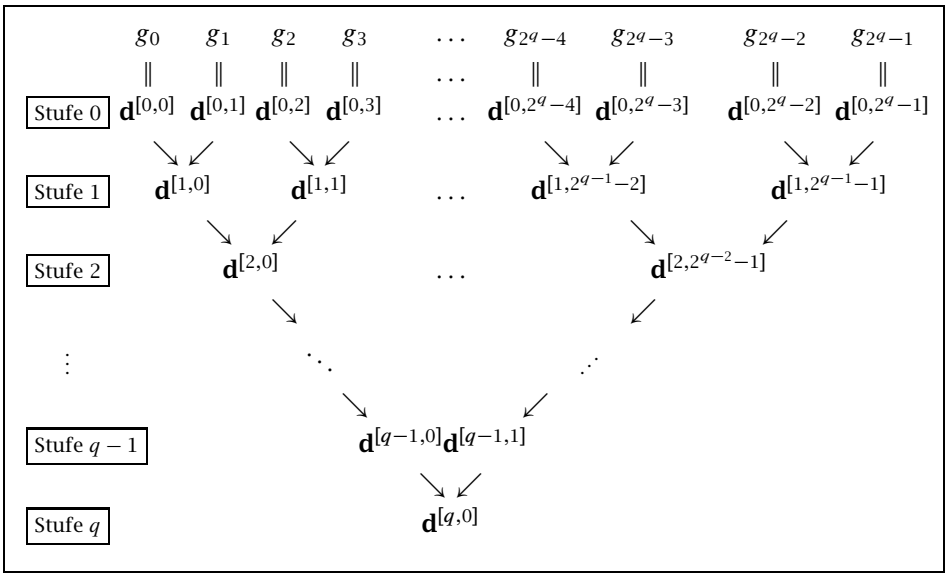
aus den Datensätzen der jeweils vorhergehenden Stufe $r-1$ gemäß der folgenden Vorschrift:

$$\begin{aligned} \mathbf{d}_k^{[r+1,j]} &:= \tfrac{1}{2}\big(\mathbf{d}_k^{[r,2j]} + \theta(r)^k \mathbf{d}_k^{[r,2j+1]}\big), \\ \mathbf{d}_{2^r+k}^{[r+1,j]} &:= \tfrac{1}{2}\big(\text{—«—} - \text{———«———}\big), \qquad \begin{aligned} k &= 0, \ldots, 2^r - 1, \\ j &= 0, \ldots, 2^{q-r-1} - 1, \\ r &= 0, \ldots, q-1, \end{aligned} \end{aligned}$$

mit den Zahlen $\theta(r) := e^{-\mathrm{i}\pi/2^r},\ r = 0, 1, \ldots, q-1$. △

Bemerkung 3.20. In Schema 3.2 ist die Vorgehensweise beim FFT-Algorithmus schematisch dargestellt. △

Mit dem nachfolgenden Theorem werden die Einträge der im Zuge des FFT-Algorithmus auftretenden Vektoren angegeben.



Schema 3.2: Schema zur Vorgehensweise beim FFT-Algorithmus

Theorem 3.21. *Es gilt*

$$\mathbf{d}^{[r,j]} = \mathcal{F}(g_{j\,2^r+\sigma_r(0)}, g_{j\,2^r+\sigma_r(1)}, .., g_{j\,2^r+\sigma_r(2^r-1)}),\ j = 0,1,\ldots,2^{q-r}-1, \quad r = 0,1,\ldots,q. \tag{3.27}$$

BEWEIS. Es wird vollständige Induktion über r angewandt. Die Aussage (3.27) ist sicher richtig für $r = 0$, und im Folgenden sei (3.27) richtig für ein $0 \le r \le q-1$. Dann berechnet man unter Berücksichtigung von $2j\,2^r = j\,2^{r+1}$ Folgendes,

$$\begin{aligned}
\mathbf{d}_k^{[r+1,j]} &= \tfrac{1}{2}\big(\mathcal{F}_k(g_{2j\,2^r+\sigma_r(0)},\ldots,g_{2j\,2^r+\sigma_r(2^r-1)}) \\
&\qquad + \theta(r)^k\,\mathcal{F}_k(g_{(2j+1)2^r+\sigma_r(0)},\ldots,g_{(2j+1)2^r+\sigma_r(2^r-1)})\big) \\
&= \mathcal{F}_k(g_{s_0},\ldots,g_{s_{2^{r+1}-1}})
\end{aligned}$$

mit

$$\begin{aligned}
s_{2k} &:= j\,2^{r+1} + \overbrace{\sigma_r(k)}^{\sigma_{r+1}(2k)}, \\
s_{2k+1} &:= -\!\!\ll\!\!- + \underbrace{2^r + \sigma_r(k)}_{\sigma_{r+1}(2k+1)}, \qquad k = 0,1,\ldots,2^r-1,
\end{aligned}$$

unter Berücksichtigung von Lemma 3.18. Die angegebene Darstellung für $\mathbf{d}_{2^r+k}^{[r+1,j]}$ ergibt sich durch die gleiche Rechnung, mit $\theta(r)^k$ ersetzt durch $-\theta(r)^k$. Dies komplettiert den Beweis des Theorems. □

Bemerkung 3.22. Wenn man für eine fixierte Zahl r alle in (3.27) auftretenden *Argumente* $g_{j\,2^r\,+\,\sigma_r(k)}$ (für $k = 0, \ldots, 2^r - 1,\ j = 0, \ldots, 2^{q-r} - 1$) aufreiht mit j als äußerem Laufindex, so findet sich an der Position $j\,2^r\ +\ k$ die Zahl $g_{j\,2^r\,+\,\sigma_r(k)}$, deren Index man aus $j\,2^r\ +\ k \in \mathcal{M}_q$ durch Bit-Umkehr der ersten (zu den kleinsten Potenzen der Basis 2 gehörenden) r Bits erhält. Für $N = 8$ ist die Situation in Tabelle 3.2 dargestellt. △

Stufe r	Position der Argumente							
	0	1	2	3	4	5	6	7
0	g_{000} ⋮	g_{001} ⋮	g_{010} ⋮	g_{011} ⋮	g_{100} ⋮	g_{101} ⋮	g_{110} ⋮	g_{111}
1	g_{000}	g_{001} ⋮	g_{010}	g_{011} ⋮	g_{100}	g_{101} ⋮	g_{110}	g_{111}
2	g_{000}	g_{010}	g_{001}	g_{011} ⋮	g_{100}	g_{110}	g_{101}	g_{111}
3	g_{000}	g_{100}	g_{010}	g_{110}	g_{001}	g_{101}	g_{011}	g_{111}

Tabelle 3.2: Stufenweise Auflistung der Argumente aus (3.27) gemäß der in Bemerkung 3.22 angegebenen Reihenfolge am Beispiel $N = 2^3$. Die Indizes der Zahlen sind in Binärdarstellung angegeben.

Unter Beachtung von $\sigma_q \circ \sigma_q = \mathrm{id}$ erhält man als wesentliche Schlussfolgerung aus Theorem 3.21 das folgende Resultat:

Korollar 3.23. *Der FFT-Algorithmus liefert*

$$\mathbf{d}^{[q,0]} \ = \ \mathcal{F}(g_{\sigma_q(0)}, \ldots,\ g_{\sigma_q(2^q-1)}).$$

Die Setzung $g_k = f_{\sigma_q(k)}$, $k = 0, \ldots, 2^q - 1$, *führt daher auf* $\mathbf{d}^{[q,0]} = \mathcal{F}(f_0, \ldots, f_{2^q-1})$.

Die Bit-Umkehr liefert also tatsächlich die anfänglich richtige Zuordnung der Zahlen $f_0, f_1, \ldots, f_{N-1} \in \mathbb{C}$ auf die Positionen 0 bis $N - 1$.

3.3.5 Aufwandsbetrachtungen für den FFT-Algorithmus

Theorem 3.24. *Bei der schnellen Fouriertransformation zur Bestimmung der diskreten Fouriertransformierten eines Datensatzes der Länge* $N = 2^q$ *fallen nicht mehr als* $N\log_2(N)/2 + \mathcal{O}(N)$ *komplexe Multiplikationen an.*

BEWEIS. Wir verwenden im Folgenden die Notationen aus Algorithmus 3.25. Für $r \in \{0, 1, \ldots, q-1\}$ fallen beim Übergang von der r-ten zur $(r+1)$-ten Stufe des FFT-Algorithmus die folgenden komplexen Multiplikationen an:

- ausgehend von $\theta(r)$ erfordert die Berechnung der Zahlen $\theta(r)^2, \theta(r)^3, \ldots, \theta(r)^{2^r-1} \in \mathbb{C}$ insgesamt $2^r - 2$ ($\leq 2^r$) komplexe Multiplikationen;
- zur Bestimmung des Vektors $\mathbf{d}^{[r+1,j]} \in \mathbb{C}^{2^{r+1}}$ aus den beiden Vektoren $\mathbf{d}^{[r,2j]}$, $\mathbf{d}^{[r,2j+1]} \in \mathbb{C}^{2^r}$ sind 2^r komplexe Multiplikationen erforderlich, und dies jeweils

für die Indizes $j = 0, \ldots, 2^{q-r-1} - 1$. Dies summiert sich zu $2^r \times 2^{q-r-1} = 2^{q-1}$ komplexen Multiplikationen auf.

Beim Übergang von der r-ten zur $(r+1)$-ten Stufe des FFT-Algorithmus fallen demnach weniger als $2^{q-1} + 2^r$ komplexe Multiplikationen an. Berücksichtigt man noch die zu Beginn des FFT-Algorithmus notwendigen $q - 2$ ($\leq q$) komplexen Multiplikationen $\theta(r) = \theta(r+1)^2$, $r = q-2, q-3, \ldots, 1$, so erhält man abschließend für den gesamten FFT-Algorithmus die folgende obere Schranke für die erforderliche Zahl komplexer Multiplikationen:

$$\sum_{r=0}^{q-1} (2^{q-1} + 2^r) + q \leq q2^{q-1} + 2^q + q = \frac{N \log_2(N)}{2} + \mathcal{O}(N). \qquad \square$$

3.3.6 Pseudocode für den FFT-Algorithmus in der Situation $N = 2^q$

Abschließend wird der FFT-Algorithmus in Form eines Pseudocodes angegeben.

Algorithmus 3.25. Sei $N = 2^q$.

Eingabe $f(k) = f_k$, $k = 0, \ldots, N-1$ (** reeller oder komplexer Datensatz **)
Ausgabe $d(k) = d_k$, ——— « ——— (** diskrete Fouriertransformierte **)

```
for k = 0 : (N − 1)   d(k) = f(σ_q(k))/N   end
for r = 0 : (q − 1)          (** Übergang Stufe r → Stufe r + 1 **)
   M = 2^r;   θ = e^{−iπ/M};   (** M ≙ Datensatzlänge(r) **)
   for k = 0 : (M − 1)       (** k ≙ Position in den Datensätzen **)
      for j = 0 : 2^{q−r−1} − 1 (** 2^{q−r−1} ≙ (Anzahl Datensätze)(r + 1) **)
         x = θ^k d(2jM + M + k);
         d(2jM + M + k) = d(2jM + k) − x;
         d(2jM + k)     = d(2jM + k) + x;
      end
   end
end
```

△

Weitere Themen und Literaturhinweise

Die diskrete Fouriertransformation geht zurück auf Cooley/Tukey [12] und wird beispielsweise in Bärwolff [2], Bollhöfer/Mehrmann [6], Deuflhard/Hohmann [22], Hanke-Bourgeois [52], Oevel [78] und in Schwarz/Klöckner [94] einführend behandelt. In [52], [78] sowie in [82] werden auch die in der Bildverarbeitung bedeutungsvolle *zweidimensionale* diskrete Fourier- beziehungsweise Cosinustransformation

und deren Modifikationen für die Datenkompression beziehungsweise die Digitalisierung beschrieben. Diskrete Fouriertransformationen für die trigonometrische Interpolation auf nichtäquidistanten Gittern werden in Potts / Steidl / Tasche [84] behandelt.

Übungsaufgaben

Aufgabe 3.1. Für gerades N seien $(N+1)$ Stützstellen $x_0 < x_1 < \ldots < x_N$ und Stützwerte $f_0, f_1, \ldots, f_N \in \mathbb{C}$ gegeben, mit $x_N - x_0 < 2\pi$. Man zeige Folgendes:

(a) Es gibt genau ein trigonometrisches Polynom der Form

$$T(x) \;=\; \frac{A_0}{2} \;+\; \sum_{k=1}^{N/2} (A_k \cos kx \;+\; B_k \sin kx), \tag{3.28}$$

mit komplexen Koeffizienten A_k und B_k, das die Interpolationsbedingungen $T(x_j) = f_j$ für $j = 0, 1, \ldots, N$ erfüllt.

(b) Sind die Stützwerte $f_0, f_1, \ldots, f_N$ alle reell, so sind es auch alle Koeffizienten A_k, B_k des zugehörigen interpolierenden trigonometrischen Polynoms der Form (3.28).

Aufgabe 3.2. Sei N gerade. Man zeige:

(a) Für reelle Zahlen $x_1, x_2, \ldots, x_N$ ist die Funktion

$$t(x) \;=\; \prod_{s=1}^{N} \sin \frac{x - x_s}{2}$$

ein trigonometrisches Polynom von der Form (3.28) mit reellen Koeffizienten $A_k,\ B_k$.

(b) Man zeige mithilfe von Teil (a) der vorliegenden Aufgabe, dass das interpolierende trigonometrische Polynom zu den Stützstellen in Aufgabe 3.1 und zu den Stützwerten f_0, $f_1, \ldots, f_N$ identisch ist mit

$$T(x) \;=\; \sum_{k=0}^{N} \frac{f_k}{t_k(x_k)}\, t_k(x), \qquad \text{mit} \quad t_k(x) \;:=\; \prod_{\substack{s=0 \\ s \neq k}}^{N} \sin \frac{x - x_s}{2}.$$

Hinweis zu (a): Für $\mathcal{U}_n := \operatorname{span}\{\mathbf{1},\ \sin x,\ \cos x, \ldots,\ \sin nx,\ \cos nx\}$ weise man Folgendes nach:

- für beliebige Zahlen $b,\ c \in \mathbb{R}$ gilt $w(x) \;:=\; \sin \frac{x-b}{2} \sin \frac{x-c}{2} \;\in\; \mathcal{U}_1$;
- $g_1 \in \mathcal{U}_m,\ g_2 \in \mathcal{U}_n \quad\Longrightarrow\quad g_1 g_2 \in \mathcal{U}_{m+n}$.

Aufgabe 3.3. Es bezeichne nun $D_2 : \mathbb{C}^N \to \mathbb{C}^N$ die folgende lineare Abbildung:

$$D_2 c \;:=\; (-c_{j-1} \;+\; 2c_j \;+\; c_{j+1})_{j=0,\ldots,N-1}, \qquad \text{mit} \qquad c \;=\; (c_0, c_1, \ldots, c_{N-1})^\top,$$
$$c_{-1} \;:=\; c_{N-1}, \quad c_N := c_0,$$

und außerdem sei

$$M \;=\; \operatorname{diag}(\lambda_0, \lambda_1, \ldots, \lambda_{N-1}) \in \mathbb{C}^{N \times N} \quad \text{mit} \quad \lambda_k := 4 \sin^2(k\pi/N) \in \mathbb{R} \quad \text{für} \quad k = 0, 1, \ldots, N-1.$$

Man zeige Folgendes:

$$D_2 \;=\; \mathcal{F}^{-1} M \mathcal{F},$$
$$(D_2 - \lambda I)^{-1} \;=\; \mathcal{F}^{-1} (M - \lambda I)^{-1} \mathcal{F} \qquad (\lambda \in \mathbb{C}, \quad \lambda \neq \lambda_k \ \text{ für } k = 0, 1, \ldots, N-1).$$

Hierbei bezeichnet $\mathcal{F} : \mathbb{C}^N \to \mathbb{C}^N$ die diskrete Fouriertransformation.

Aufgabe 3.4. (a) Zu einem gegebenen Datensatz $f_0, f_1, \ldots, f_{N-1}$ komplexer Zahlen sei der Datensatz $\tilde{d}_0, \tilde{d}_1, \ldots, \tilde{d}_{N-1}$ komplexer Zahlen definiert durch

$$\tilde{d}_k \;=\; \frac{\gamma_k}{N} \sum_{j=0}^{N-1} f_j e^{-\mathrm{i}(2j+1)k\pi/N} \qquad \text{für} \quad k \;=\; 0, 1, \ldots, N-1 \tag{3.29}$$

mit gegebenen Koeffizienten $\gamma_k \neq 0$ für $k \;=\; 0, 1, \ldots, N-1$. Man zeige

$$f_j \;=\; \sum_{k=0}^{N-1} \frac{\tilde{d}_k}{\gamma_k} e^{\mathrm{i}(2j+1)k\pi/N} \qquad \text{für} \quad j \;=\; 0, 1, \ldots, N-1.$$

(b) Zu einem gegebenen Datensatz $f_0, f_1, \ldots, f_{n-1}$ reeller Zahlen mit $n \in \mathbb{N}$ sei der transformierte Datensatz $d_0, d_1, \ldots, d_{n-1}$ reeller Zahlen definiert durch

$$d_k \;=\; \frac{\gamma_k}{n} \sum_{j=0}^{n-1} f_j \cos\big(\frac{(2j+1)k\pi}{2n}\big) \qquad \text{für} \quad k \;=\; 0, 1, \ldots, n-1 \tag{3.30}$$

mit gegebenen Koeffizienten $\gamma_k \neq 0$ für $k \;=\; 0, 1, \ldots, n-1$. Man zeige:

$$f_j \;=\; \frac{d_0}{\gamma_0} \;+\; 2 \sum_{k=1}^{n-1} \frac{d_k}{\gamma_k} \cos\big(\frac{(2j+1)k\pi}{2n}\big) \quad \text{für} \quad j \;=\; 0, 1, \ldots, n-1. \tag{3.31}$$

Hinweis: Man verwende Teil (a) dieser Aufgabe mit den Setzungen $N = 2n$ und $f_{N-1-j} \;=\; f_j$ für $j \;=\; 0, 1, \ldots, n-1$ beziehungsweise $\gamma_{N-k} \;=\; \gamma_k$ für $k = 1, 2, \ldots, n$ und zeige für diese Situation noch $\tilde{d}_{N-k} = -\tilde{d}_k$ für $k = 1, 2, \ldots, n$.

Aufgabe 3.5. Für $n \in \mathbb{N}$ sei $f_0, f_1, \ldots, f_{n-1}$ ein gegebener Datensatz reeller Zahlen.

(a) Man zeige, dass mit den Koeffizienten d_k aus (3.30) für das trigonometrische Polynom

$$p(\theta) \;=\; \frac{d_0}{\gamma_0} \;+\; 2 \sum_{k=1}^{n-1} \frac{d_k}{\gamma_k} \cos k\theta \tag{3.32}$$

Folgendes gilt:

$$p\big(\frac{2j+1}{2n}\pi\big) \;=\; f_j \quad \text{für} \quad j \;=\; 0, 1, \ldots, n-1.$$

(b) Es sei $\mathcal{P} \in \Pi_{n-1}$ das Interpolationspolynom zu den Stützpunkten $(t_{j+1}^{(n)}, f_j)$ für $j = 0$, $1, \ldots, n-1$, wobei $t_{j+1}^{(n)} = \cos\left((2j+1)\pi\,/\,(2n)\right)$ die Nullstellen des Tschebyscheff-Polynoms T_n der ersten Art vom Grad n bezeichnet. Man zeige, dass mit den Koeffizienten d_k aus (3.30) Folgendes gilt:

$$\mathcal{P}(x) \;=\; \frac{d_0}{\gamma_0} \;+\; 2 \sum_{k=1}^{n-1} \frac{d_k}{\gamma_k} T_k(x). \tag{3.33}$$

Aufgabe 3.6 (*Numerische Aufgabe*). Man berechne entsprechend der Vorgehensweise in Teil (b) der Aufgabe 3.5 das Interpolationspolynom $\mathcal{P} \in \Pi_{n-1}$ zu den beiden Funktionen

$$f(x) \;=\; x^{1/3}, \qquad x \in [0, 64] \qquad \text{bzw.} \qquad f(x) \;=\; \log(x), \qquad x \in (0, 1]$$

für die Werte $n = 2^m$ für $m = 2, 4, \ldots, 10$ und mit den Stützstellen aus Teil (b) der Aufgabe 3.5, wobei hierfür das Intervall $[-1, 1]$ affin-linear auf $[0, 64]$ beziehungsweise $[0, 1]$ zu transformieren ist.

Die Koeffizienten $d_0, d_1, \ldots, d_{n-1}$ (mit den Faktoren $\gamma_k = 2$ für $k = 0, 1, \ldots, n-1$) des Interpolationspolynoms $\mathcal{P}$ in der Darstellung (3.33) berechne man mit der schnellen Fouriertransformation. Man berechne außerdem den auftretenden Fehler an (den linear zu transformierenden) Stellen $x_j = -1 + j/10$ für $j = 1, 2, \ldots, 20$. Zur Auswertung von $\mathcal{P}(x) = d_0/2 + \sum_{k=1}^{n-1} d_k T_k(x)$ verwende man die folgende Variante des Horner-Schemas:

$$\begin{aligned} b_n &:= b_{n+1} := 0, \qquad b_k := 2x\,b_{k+1} - b_{k+2} + d_k \qquad \text{für } k = n-1, n-2, \ldots, 0, \\ \mathcal{P}(x) &= (b_0 - b_2)/2. \end{aligned} \tag{3.34}$$

Man weise noch die Richtigkeit der Identität (3.34) nach.

Aufgabe 3.7. Man beweise Lemma 3.18.

4 Lösung linearer Gleichungssysteme

In diesem Abschnitt werden Verfahren zur Lösung linearer Gleichungssysteme $Ax = b$ vorgestellt, wobei $A = (a_{jk}) \in \mathbb{R}^{N\times N}$ eine gegebene Matrix und $b = (b_j) \in \mathbb{R}^N$ ein gegebener Vektor ist. Solche Gleichungssysteme treten in zahlreichen Anwendungen auf, wovon eine bereits aus Kapitel 2 über Splinefunktionen bekannt ist.

4.1 Gestaffelte lineare Gleichungssysteme

Typischerweise überführt man lineare Gleichungssysteme $Ax = b$ in eine gestaffelte Form, die dann einfach nach den Unbekannten aufzulösen ist. Solche gestaffelten linearen Gleichungssysteme werden zunächst kurz behandelt.

Definition 4.1. Matrizen $L,\ R \in \mathbb{R}^{N\times N}$ der Form

$$L = \begin{pmatrix} \ell_{11} & 0 & \cdots & 0 \\ \ell_{21} & \ell_{22} & \ddots & \vdots \\ \vdots & & \ddots & 0 \\ \ell_{N1} & \cdots & \cdots & \ell_{NN} \end{pmatrix}, \qquad R = \begin{pmatrix} r_{11} & r_{12} & \cdots & r_{1N} \\ 0 & r_{22} & & \vdots \\ \vdots & \ddots & \ddots & \vdots \\ 0 & \cdots & 0 & r_{NN} \end{pmatrix},$$

heißen *untere* beziehungsweise *obere Dreiecksmatrizen.*

Es sind die Matrizen L beziehungsweise R regulär genau dann, wenn $\det(L) = \prod_{j=1}^{N} \ell_{jj} \neq 0$ beziehungsweise $\det(R) = \prod_{j=1}^{N} r_{jj} \neq 0$ gilt.

4.1.1 Obere gestaffelte Gleichungssysteme

Für die obere Dreiecksmatrix $R = (r_{jk}) \in \mathbb{R}^{N\times N}$ mit $r_{jk} = 0$ für $j > k$ ist das entsprechende gestaffelte Gleichungssystem $Rx = z$ für einen gegebenen Vektor $z \in \mathbb{R}^N$ von der Form

$$\begin{aligned} r_{11}x_1 + r_{12}x_2 + \cdots + r_{1N}x_N &= z_1 \\ r_{22}x_2 + \cdots + r_{2N}x_N &= z_2 \\ \ddots \quad \vdots \quad \vdots \quad \vdots & \\ r_{NN}x_N &= z_N \end{aligned}$$

dessen Lösung $z \in \mathbb{R}^N$ für reguläres R zeilenweise von unten nach oben durch jeweiliges Auflösen nach der Unbekannten auf der Diagonalen berechnet werden kann, siehe Schema 4.1.

$$\texttt{for } j = N:-1:1 \qquad x_j \;=\; \Big(z_j - \sum_{k=j+1}^{N} r_{jk}x_k\Big)/r_{jj}\,; \qquad \texttt{end}$$

Schema 4.1: Rekursive Auflösung eines oberen gestaffelten Gleichungssystems $Rx = z$

Theorem 4.2. *Für die Auflösung eines oberen gestaffelten Gleichungssystems sind N^2 arithmetische Operationen erforderlich.*

BEWEIS. In den Stufen $j = N, N-1, \ldots, 1$ der Schleife aus Schema 4.1 sind zur Berechnung der Unbekannten x_j je $N - j$ Multiplikationen und genauso viele Subtraktionen sowie eine Division durchzuführen, insgesamt erhält man die folgende Anzahl von arithmetischen Operationen,

$$N \;+\; 2\sum_{j=1}^{N}(N-j) \;=\; N \;+\; 2\sum_{m=1}^{N-1} m \;=\; N \;+\; (N-1)N \;=\; N^2.$$

□

4.1.2 Untere gestaffelte Gleichungssysteme

Für die untere Dreiecksmatrix $L = (\ell_{jk}) \in \mathbb{R}^{N\times N}$ mit $\ell_{jk} = 0$ für $j < k$ ist das entsprechende gestaffelte Gleichungssystem $Lx = b$ mit einem gegebenen Vektor $b \in \mathbb{R}^N$ von der folgenden Form,

$$\begin{array}{lllll}
\ell_{11}x_1 & & & = & b_1 \\
\ell_{21}x_1 + & \ell_{22}x_2 & & = & b_2 \\
\vdots & \vdots & \ddots & \vdots & \vdots \\
\ell_{N1}x_1 + & \ell_{N2}x_2 + \cdots + & \ell_{NN}x_N & = & b_N
\end{array}$$

Dessen Lösung $x \in \mathbb{R}^N$ kann für eine reguläre Matrix L zeilenweise von oben nach unten durch jeweiliges Auflösen nach der Unbekannten auf der Diagonalen berechnet werden:

$$\texttt{for } j = 1:N \qquad x_j \;=\; \Big(b_j - \sum_{k=1}^{j-1} \ell_{jk}x_k\Big)/\ell_{jj}\,; \qquad \texttt{end}$$

Schema 4.2: Rekursive Auflösung eines regulären unteren gestaffelten Gleichungssystems $Lx = b$

Dabei sind genauso viele arithmetische Operationen durchzuführen wie im Fall des oberen gestaffelten Gleichungssystems, nämlich N^2 (vergleiche Theorem 4.2).

4.2 Der Gauß-Algorithmus

4.2.1 Einführende Bemerkungen

Seien wieder $A = (a_{jk}) \in \mathbb{R}^{N\times N}$ eine gegebene Matrix sowie $b = (b_j) \in \mathbb{R}^N$ ein gegebener Vektor. Im Folgenden wird der Gauß-Algorithmus beschrieben, der das Gleichungssystem $Ax = b$ in ein äquivalentes oberes gestaffeltes Gleichungssystem $Rx = z$ überführen soll, dessen Lösung $x \in \mathbb{R}^N$ dann leicht berechnet werden kann.

In der ersten Stufe des Gauß-Algorithmus wird das gegebene Gleichungssystem

$$\begin{aligned} a_{11}x_1 &+ a_{12}x_2 &+ \cdots + a_{1N}x_N &= b_1 \\ a_{21}x_1 &+ a_{22}x_2 &+ \cdots + a_{2N}x_N &= b_2 \\ \vdots & \quad \vdots & \quad \vdots & \quad \vdots \\ a_{N1}x_1 &+ a_{N2}x_2 &+ \cdots + a_{NN}x_N &= b_N \end{aligned}$$

durch Zeilenoperationen in ein äquivalentes Gleichungssystem der Form

$$\left.\begin{aligned} a_{11}x_1 + a_{12}x_2 &+ \cdots + a_{1N}x_N &= b_1 \\ a_{22}^{(2)}x_2 &+ \cdots + a_{2N}^{(2)}x_N &= b_2^{(2)} \\ \vdots & \quad \vdots & \quad \vdots \\ a_{N2}^{(2)}x_2 &+ \cdots + a_{NN}^{(2)}x_N &= b_N^{(2)} \end{aligned}\right\} \tag{4.1}$$

überführt. Falls $a_{11} \neq 0$ gilt, so kann dieses erreicht werden mit Zeilenoperationen

$$\texttt{neue Zeile } j := \texttt{alte Zeile } j - \ell_{j1} \cdot \texttt{alte Zeile } 1, \quad j = 2, 3, \ldots, N,$$

oder explizit

$$\underbrace{(a_{j1} - \ell_{j1}a_{11})}_{\stackrel{!}{=} 0} x_1 + \underbrace{(a_{j2} - \ell_{j1}a_{12})}_{=: a_{j2}^{(2)}} x_2 + \cdots + \underbrace{(a_{jN} - \ell_{j1}a_{1N})}_{=: a_{jN}^{(2)}} x_N = \underbrace{b_j - \ell_{j1}b_1}_{=: b_j^{(2)}}$$

mit der Setzung

$$\ell_{j1} := \frac{a_{j1}}{a_{11}}, \qquad j = 2, 3, \ldots, N.$$

Nach diesem Eliminierungsschritt verfährt man im nächsten Schritt ganz analog mit dem System der unteren $N - 1$ Gleichungen in (4.1). Diesen Eliminierungsprozess sukzessive durchgeführt auf die jeweils entstehenden Teilsysteme liefert zu $Ax = b$ äquivalente Gleichungssysteme

$$A^{(s)}x = b^{(s)}, \qquad s = 1, 2, \ldots, N,$$

wobei sich $A^{(s)} \in \mathbb{R}^{N\times N}$ und $b^{(s)} \in \mathbb{R}^N$ in der Reihenfolge

$$\begin{aligned} A &= A^{(1)} \to A^{(2)} \to \ldots \to A^{(N)} =: R \\ b &= b^{(1)} \to b^{(2)} \to \ldots \to b^{(N)} =: z \end{aligned}$$

ergeben mit Matrizen und Vektoren von der speziellen Form

$$A^{(s)} = \begin{pmatrix} a_{11}^{(1)} & a_{12}^{(1)} & \cdots & \cdots & \cdots & a_{1N}^{(1)} \\ & a_{22}^{(2)} & \cdots & \cdots & \cdots & a_{2N}^{(2)} \\ & & \ddots & & & \vdots \\ & & & a_{ss}^{(s)} & \cdots & a_{sN}^{(s)} \\ & & & \vdots & & \vdots \\ & & & a_{Ns}^{(s)} & \cdots & a_{NN}^{(s)} \end{pmatrix} \in \mathbb{R}^{N\times N}, \quad b^{(s)} = \begin{pmatrix} b_1^{(1)} \\ b_2^{(2)} \\ \vdots \\ b_s^{(s)} \\ \vdots \\ b_N^{(s)} \end{pmatrix} \in \mathbb{R}^N. \tag{4.2}$$

Hierbei wird vorausgesetzt, dass die auftretenden Diagonalelemente allesamt nicht verschwinden, $a_{ss}^{(s)} \neq 0$ für $s = 1, 2, \ldots, N$, da anderweitig der Gauß-Algorithmus abbricht beziehungsweise die Matrix R singulär ist.

Algorithmus 4.3. Ein Pseudocode für den Gauß-Algorithmus ist in dem folgenden Schema 4.3 angegeben. Dabei werden zur Illustration noch die Indizes $^{(1)}$, $^{(2)}$, ... mitgeführt. In jeder Implementierung werden dann entsprechend die Einträge der ursprünglichen Matrix A sowie in dem Vektor b überschrieben.

```
for s = 1 : N − 1                    (** A^(s) → A^(s+1), b^(s) → b^(s+1) **)
    for j = s + 1 : N                                    (** Zeile j **)
        ℓ_js = a_js^(s) / a_ss^(s);      b_j^(s+1) = b_j^(s) − ℓ_js b_s^(s);
        (a_{j,s+1}^(s+1), ..., a_jN^(s+1)) = (a_{j,s+1}^(s), ..., a_jN^(s)) − ℓ_js (a_{s,s+1}^(s), ..., a_sN^(s));
    end
end
```

Schema 4.3: Gauß-Algorithmus △

Theorem 4.4. *Für den Gauß-Algorithmus in Schema 4.3 sind*

$$\tfrac{2N^3}{3}\big(1 + \mathcal{O}\big(\tfrac{1}{N}\big)\big) \tag{4.3}$$

arithmetische Operationen erforderlich.

BEWEIS. In der s-ten Stufe des Gauß-Algorithmus sind $(N-s)^2 + (N-s)$ Multiplikationen und ebenso viele Additionen durchzuführen und außerdem sind $(N-s)$ Divisionen erforderlich, so dass insgesamt

$$2\sum_{s=1}^{N-1} s^2 \;+\; 3\sum_{s=1}^{N-1} s \;=\; \frac{(N-1)N(2N-1)}{3} \;+\; \frac{3N(N-1)}{2} \;=\; \frac{2N^3}{3}\big(1 \;+\; \mathcal{O}\big(\tfrac{1}{N}\big)\big)$$

arithmetische Operationen anfallen. □

Das folgende Theorem liefert eine Klasse von Matrizen $A \in \mathbb{R}^{N\times N}$, für die der Gauß-Algorithmus durchführbar ist.

Theorem 4.5. *Ist die Matrix $A = (a_{jk}) \in \mathbb{R}^{N\times N}$ strikt diagonaldominant, so ist der Gauß-Algorithmus zur Lösung von $Ax = b$ durchführbar.*

BEWEIS. Es wird mit vollständiger Induktion über $s = 1, 2, \ldots, N-1$ nachgewiesen, dass die Matrizen

$$B^{(s)} \;=\; \begin{pmatrix} a_{ss}^{(s)} & \cdots & a_{sN}^{(s)} \\ \vdots & & \vdots \\ a_{Ns}^{(s)} & \cdots & a_{NN}^{(s)} \end{pmatrix} \;\in\; \mathbb{R}^{(N-s+1)\times(N-s+1)} \tag{4.4}$$

strikt diagonaldominant sind. Für $B^{(1)} = A$ ist dies nach Voraussetzung richtig, und wir nehmen nun an, dass für ein $1 \le s \le N-2$ die Matrix $B^{(s)}$ strikt diagonaldominant ist. Dann gilt insbesondere $a_{ss}^{(s)} \neq 0$, somit ist der Gauß-Eliminationsschritt auf $B^{(s)}$ anwendbar und liefert die Matrix $B^{(s+1)} = (a_{jk}^{(s+1)})_{s+1\le j,k\le N} \in \mathbb{R}^{(N-s)\times(N-s)}$ mit

$$\big(a_{j,s+1}^{(s+1)}, \ldots, a_{jN}^{(s+1)}\big) \;=\; \big(a_{j,s+1}^{(s)}, \ldots, a_{jN}^{(s)}\big) - \ell_{js}\big(a_{s,s+1}^{(s)}, \ldots, a_{sN}^{(s)}\big), \quad j = s+1, \ldots, N,$$

mit den Koeffizienten

$$\ell_{js} \;=\; a_{js}^{(s)}/a_{ss}^{(s)}, \qquad j \;=\; s+1, s+2, \ldots, N.$$

Man erhält nun die strikte Diagonaldominanz der Matrix $B^{(s+1)}$: für $j = s+1, \ldots, N$ ergibt sich

$$\begin{aligned} \sum_{\substack{k=s+1\\k\neq j}}^{N} |a_{jk}^{(s+1)}| \;&\le\; \sum_{\substack{k=s+1\\k\neq j}}^{N} |a_{jk}^{(s)}| \;+\; |\ell_{js}| \sum_{\substack{k=s+1\\k\neq j}}^{N} |a_{sk}^{(s)}| \\ &<\; |a_{jj}^{(s)}| \;-\; |a_{js}^{(s)}| \;+\; \frac{|a_{js}^{(s)}|}{|a_{ss}^{(s)}|}\big(|a_{ss}^{(s)}| \;-\; |a_{sj}^{(s)}|\big) \\ &=\; |a_{jj}^{(s)}| \;-\; |\ell_{js}||a_{sj}^{(s)}| \;\le\; |a_{jj}^{(s+1)}|, \end{aligned}$$

was den Beweis komplettiert. □

4.2.2 Gauß-Algorithmus mit Pivotsuche

Zu Illustrationszwecken betrachten wir für $\varepsilon \in \mathbb{R}$ die reguläre Matrix

$$A_\varepsilon = \begin{pmatrix} \varepsilon & 1 \\ 1 & 0 \end{pmatrix}.$$

Für jeden Vektor $b \in \mathbb{R}^2$ ist der Gauß-Algorithmus zur Staffelung von $A_0 x = b$ nicht durchführbar, und für $0 \neq \varepsilon \approx 0$ erhält man in der ersten Stufe des Gauß-Algorithmus zur Staffelung von $A_\varepsilon x = b$ das Element $\ell_{21} = 1/\varepsilon$, was bei der Berechnung der Lösung zugehöriger Gleichungssysteme zu Fehlerverstärkungen führen kann. Zur Vermeidung solcher numerischen Instabilitäten bietet sich die folgende Vorgehensweise an:

Algorithmus 4.6. (Gauß-Algorithmus mit *Pivotstrategie*). Im Folgenden wird der Übergang $A^{(s)} \to A^{(s+1)}$ um eine Pivotstrategie ergänzt.

(a) Man bestimme zunächst einen Index $p \in \{s, s+1, \ldots, N\}$ mit

$$|a_{ps}^{(s)}| \geq |a_{js}^{(s)}| \qquad \text{für} \quad j = s, s+1, \ldots, N.$$

Das Element $a_{ps}^{(s)}$ wird als *Pivotelement* bezeichnet.

(b) Transformiere $A^{(s)} \to \widehat{A}^{(s)} = \left(\widehat{a}_{jk}^{(s)}\right) \in \mathbb{R}^{N\times N}$ sowie $b^{(s)} \to \widehat{b}^{(s)} = \left(\widehat{b}_j^{(s)}\right) \in \mathbb{R}^N$ durch Vertauschung der p-ten und der s-ten Zeile von $A^{(s)}$ beziehungsweise $b^{(s)}$:

$$\begin{aligned} &\left(\widehat{a}_{ps}^{(s)}, \ldots, \widehat{a}_{pN}^{(s)}\right) = \left(a_{ss}^{(s)}, \ldots, a_{sN}^{(s)}\right), \qquad \left(\widehat{a}_{ss}^{(s)}, \ldots, \widehat{a}_{sN}^{(s)}\right) = \left(a_{ps}^{(s)}, \ldots, a_{pN}^{(s)}\right), \\ &\widehat{b}_s^{(s)} = b_p^{(s)}, \qquad \widehat{b}_p^{(s)} = b_s^{(s)}, \end{aligned}$$

die anderen Einträge bleiben unverändert.

(c) Der nachfolgende Eliminationsschritt $\widehat{A}^{(s)} \to A^{(s+1)}$, $\widehat{b}^{(s)} \to b^{(s+1)}$ geht wie bisher so vonstatten, dass die Matrix $A^{(s+1)}$ die Form (4.2) erhält. △

Die in Algorithmus 4.6 vorgestellte Pivotsuche wird etwas genauer auch als *Spaltenpivotsuche* bezeichnet. Es existieren noch andere Privotstrategien (siehe Aufgabe 4.6).

4.3 Die Faktorisierung $PA = LR$

Typischerweise ist für eine gegebene reguläre Matrix $A \in \mathbb{R}^{N\times N}$ das Gleichungssystem $Ax = b$ für unterschiedliche rechte Seiten b zu lösen. Dies kann effizient mit einer Faktorisierung der Form $PA = LR$ geschehen, wobei $P \in \mathbb{R}^{N\times N}$ eine Permutationsmatrix[1] sowie $L \in \mathbb{R}^{N\times N}$ eine untere beziehungsweise $R \in \mathbb{R}^{N\times N}$ eine

[1] für deren Einführung siehe den nachfolgenden Abschnitt 4.3.1

obere Dreiecksmatrix ist: man hat für jede rechte Seite b jeweils nur nacheinander die beiden gestaffelten Gleichungssysteme

$$Lz = Pb, \qquad Rx = z,$$

zu lösen. Eine solche Faktorisierung $PA = LR$ gewinnt man mit dem Gauß-Algorithmus mit Spaltenpivotsuche; man hat nur die auftretenden Zeilenpermutationen und Zeilenoperationen geeignet zu verwenden. Die genaue Vorgehensweise wird am Ende dieses Abschnitts 4.3 beschrieben.

4.3.1 Permutationsmatrix

Es werden nun Permutationsmatrizen betrachtet, mit denen sich Zeilen- und Spaltenvertauschungen beschreiben lassen.

Definition 4.7. Man bezeichnet $P \in \mathbb{R}^{N\times N}$ als *Permutationsmatrix*, falls für eine bijektive Abbildung $\pi : \{1,\dots,N\} \to \{1,\dots,N\}$ (*Permutation* genannt) Folgendes gilt,

$$P = \left(\begin{array}{c|c|c} & & \\ \mathbf{e}_{\pi(1)} & \cdots & \mathbf{e}_{\pi(N)} \\ & & \end{array}\right), \tag{4.5}$$

wobei $\mathbf{e}_k \in \mathbb{R}^N$ den k-ten Einheitsvektor bezeichnet, das heißt, der k-te Eintrag des Vektors $\mathbf{e}_k$ ist gleich eins und die anderen Einträge sind gleich null.

Beispiel 4.8. Die folgende Matrix stellt eine Permutationsmatrix dar:

$$P = \begin{pmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix} \in \mathbb{R}^{4\times 4}.$$

△

Lemma 4.9. *Für eine Permutationsmatrix $P \in \mathbb{R}^{N\times N}$ mit zugehöriger Permutation π gilt die Darstellung*

$$P = \left(\begin{array}{c} \mathbf{e}_{\pi^{-1}(1)}^\top \\ \hline \vdots \\ \hline \mathbf{e}_{\pi^{-1}(N)}^\top \end{array}\right).$$

BEWEIS. Für $k = 1, 2, \dots, N$ gilt

$$\left(\begin{array}{c} \mathbf{e}_{\pi^{-1}(1)}^\top \\ \hline \vdots \\ \hline \mathbf{e}_{\pi^{-1}(N)}^\top \end{array}\right)\mathbf{e}_k = \begin{pmatrix} \mathbf{e}_{\pi^{-1}(1)}^\top \mathbf{e}_k \\ \vdots \\ \mathbf{e}_{\pi^{-1}(N)}^\top \mathbf{e}_k \end{pmatrix} = \mathbf{e}_{\pi(k)}.$$

□

Bei einer Permutationsmatrix treten also in jeder Zeile beziehungsweise jeder Spalte jeweils genau eine Eins und sonst nur Nullen auf.

Theorem 4.10. *Sei $P \in \mathbb{R}^{N\times N}$ eine Permutationsmatrix und π die zugehörige Permutation. Für Vektoren $a_1, a_2, \dots, a_N \in \mathbb{R}^M$ mit $M \in \mathbb{N}$ gilt*

$$P \begin{pmatrix} \frac{a_1^\top}{\vdots} \\ \hline a_N^\top \end{pmatrix} = \begin{pmatrix} \frac{a_{\pi^{-1}(1)}^\top}{\vdots} \\ \hline a_{\pi^{-1}(N)}^\top \end{pmatrix}, \quad \left(\begin{array}{c|c|c} a_1 & \dots & a_N \end{array}\right) P = \left(\begin{array}{c|c|c} a_{\pi(1)} & \dots & a_{\pi(N)} \end{array}\right). \tag{4.6}$$

BEWEIS. Die erste Identität erhält man wie folgt,

$$P \begin{pmatrix} \frac{a_1^\top}{\vdots} \\ \hline a_N^\top \end{pmatrix} = \sum_{j=1}^N \mathbf{e}_{\pi(j)} a_j^\top = \sum_{\ell=1}^N \mathbf{e}_\ell a_{\pi^{-1}(\ell)}^\top = \begin{pmatrix} \frac{a_{\pi^{-1}(1)}^\top}{\vdots} \\ \hline a_{\pi^{-1}(N)}^\top \end{pmatrix},$$

und die angegebene Spaltenpermutation folgt so:

$$\left(\begin{array}{c|c|c} a_1 & \dots & a_N \end{array}\right) P = \sum_{k=1}^N a_k \mathbf{e}_{\pi^{-1}(k)}^\top = \sum_{\ell=1}^N a_{\pi(\ell)} \mathbf{e}_\ell^\top = \left(\begin{array}{c|c|c} a_{\pi(1)} & \dots & a_{\pi(N)} \end{array}\right). \qquad \square$$

Bemerkung 4.11. Für eine gegebene Matrix A bewirkt also eine Multiplikation mit einer Permutationsmatrix von links eine Permutation der Zeilen von A, und eine Multiplikation mit einer Permutationsmatrix von rechts bewirkt eine Permutation der Spalten von A. In numerischen Implementierungen erfolgt die Abspeicherung einer Permutationsmatrix mit der zugehörigen Permutation π in Form eines Vektors $(\pi^{-1}(1), \dots, \pi^{-1}(N))^\top \in \mathbb{R}^N$ oder $(\pi(1), \dots, \pi(N))^\top \in \mathbb{R}^N$. △

Als unmittelbare Konsequenz aus der zweiten Identität in (4.6) erhält man noch das folgende Resultat.

Korollar 4.12. *Die Menge der Permutationsmatrizen $P \in \mathbb{R}^{N\times N}$ bildet zusammen mit der Matrizenmultiplikation eine Gruppe: für Permutationen $\pi_1, \pi_2 : \{1, \dots, N\} \to \{1, \dots, N\}$ gilt*

$$\left(\begin{array}{c|c|c} \mathbf{e}_{\pi_2(1)} & \dots & \mathbf{e}_{\pi_2(N)} \end{array}\right) \left(\begin{array}{c|c|c} \mathbf{e}_{\pi_1(1)} & \dots & \mathbf{e}_{\pi_1(N)} \end{array}\right) = \left(\begin{array}{c|c|c} \mathbf{e}_{\pi_2\circ\pi_1(1)} & \dots & \mathbf{e}_{\pi_2\circ\pi_1(N)} \end{array}\right).$$

Eine wichtige Rolle spielen im Folgenden elementare Permutationsmatrizen.

Definition 4.13. Eine *elementare Permutationsmatrix* ist von der Form (4.5) mit einer *Elementarpermutation* $\pi : \{1, \ldots, N\} \to \{1, \ldots, N\}$, die zwei Zahlen vertauscht und die restlichen Zahlen unverändert lässt, das heißt, es gibt Zahlen $1 \leq q, r \leq N$ mit

$$\pi(q) = r, \qquad \pi(r) = q, \qquad \pi(j) = j \quad \text{für} \quad j \notin \{q, r\}. \tag{4.7}$$

Bemerkung 4.14. Es sei $P \in \mathbb{R}^{N \times N}$ eine elementare Permutationsmatrix mit zugehöriger Elementarpermutation π von der Form (4.7). Dann gilt

$$P = \begin{pmatrix} 1 & & & & & & & \\ & \ddots & & & & & & \\ & & 1 & & & & & \\ & & & 0 & & & 1 & \\ & & & & 1 & & & \\ & & & & & \ddots & & \\ & & & & & & 1 & \\ & & & 1 & & & & \\ & & & & & & & 0 \\ & & & & & & & & 1 \\ & & & & & & & & & \ddots \\ & & & & & & & & & & 1 \end{pmatrix} \begin{matrix} \\ \\ \\ \leftarrow \text{ Zeile } q \\ \\ \\ \\ \leftarrow \text{ Zeile } r \\ \\ \\ \\ \end{matrix}$$

und es gilt $\pi^{-1} = \pi$ sowie $P^{-1} = P$. △

4.3.2 Eliminationsmatrizen

Es werden nun Eliminationsmatrizen betrachtet. Es wird sich herausstellen, dass sich mit solchen Matrizen Zeilenoperationen beschreiben lassen.

Definition 4.15. Jede Matrix von der Form

$$\begin{pmatrix} 1 & & & \\ & \ddots & & \\ & & 1 & \\ & -\ell_{s+1,s} & \ddots & \\ & \vdots & & \ddots \\ & -\ell_{Ns} & & 1 \end{pmatrix} \in \mathbb{R}^{N \times N} \tag{4.8}$$

mit $s \in \{1, 2, \ldots, N-1\}$ heißt *Eliminationsmatrix vom Index s*.

Bemerkung 4.16. 1. Eine Eliminationsmatrix vom Index s unterscheidet sich von der Einheitsmatrix also nur in der s-ten Spalte, und dort auch nur unterhalb der Diagonalen.
2. Die prinzipielle Vorgehensweise bei den Zeilenoperationen der s-ten Stufe des

Gauß-Algorithmus wird durch Multiplikation mit einer Eliminationsmatrix vom Index s beschrieben: für Vektoren $a_k \in \mathbb{R}^N$, $k = 1, 2, \ldots, N$ gilt

$$\begin{pmatrix} 1 & & & & \\ & \ddots & & & \\ & & 1 & & \\ & & -\ell_{s+1,s} & \ddots & \\ & & \vdots & & \ddots \\ & & -\ell_{Ns} & & 1 \end{pmatrix} \left(\begin{array}{c} a_1^\top \\ \hline \vdots \\ \hline a_N^\top \end{array}\right) = \left(\begin{array}{c} a_1^\top \\ \hline \vdots \\ \hline a_s^\top \\ \hline a_{s+1}^\top - \ell_{s+1,s} a_s^\top \\ \hline \vdots \\ \hline a_N^\top - \ell_{Ns} a_s^\top \end{array}\right).$$

3. Bei der Herleitung der Abbildungseigenschaften von Eliminationsmatrizen F_s der Form (4.8) ist die folgende Darstellung hilfreich,

$$F_s \;=\; I - \mathbf{f}_s \mathbf{e}_s^\top, \qquad \text{mit} \quad \mathbf{f}_s \;=\; \big(0, \ldots, 0, \ell_{s+1,s}, \ldots, \ell_{Ns}\big)^\top \;\in\; \mathbb{R}^N, \tag{4.9}$$

wobei $I \in \mathbb{R}^{N\times N}$ die Einheitsmatrix und $\mathbf{e}_s \in \mathbb{R}^N$ den s-ten Einheitsvektor bezeichnet.
4. Eine Eliminationsmatrix wird auch als Gauß-Transformation oder gelegentlich als Frobeniusmatrix bezeichnet. △

Die beiden folgenden Lemmata liefern Hilfsmittel für den Beweis von Theorem 4.19 über die Faktorisierung $PA = LR$.

Lemma 4.17. *Für $s = 1, 2, \ldots, N-1$ sind Eliminationsmatrizen F_s vom Index s regulär, und mit der Notation (4.8) für F_s gilt*

$$F_s^{-1} \;=\; \begin{pmatrix} 1 & & & & \\ & \ddots & & & \\ & & 1 & & \\ & & \ell_{s+1,s} & \ddots & \\ & & \vdots & & \ddots \\ & & \ell_{Ns} & & 1 \end{pmatrix} \qquad \textit{für} \quad s \;=\; 1, 2, \ldots, N-1,$$

$$F_1^{-1} \cdots F_{N-1}^{-1} \;=\; \begin{pmatrix} 1 & & & & \\ \ell_{21} & 1 & & & \\ \vdots & \ell_{32} & 1 & & \\ \vdots & \vdots & \ddots & \ddots & \\ \ell_{N1} & \ell_{N2} & \ldots & \ell_{N,N-1} & 1 \end{pmatrix}.$$

BEWEIS. Mit der Notation (4.9) für F_s berechnet man

$$(I + \mathbf{f}_s \mathbf{e}_s^\top) \underbrace{(I - \mathbf{f}_s \mathbf{e}_s^\top)}_{= F_s} \;=\; I + \mathbf{f}_s \mathbf{e}_s^\top - \mathbf{f}_s \mathbf{e}_s^\top - \mathbf{f}_s \underbrace{(\mathbf{e}_s^\top \mathbf{f}_s)}_{= 0 \in \mathbb{R}} \mathbf{e}_s^\top \;=\; I,$$

woraus die Regularität von F_s sowie die angegebene Darstellung für die Matrix F_s^{-1} folgt. Im Folgenden soll nun mit vollständiger Induktion

$$F_1^{-1} \cdots F_s^{-1} \;=\; I \;+\; \sum_{k=1}^{s} \mathbf{f}_k \mathbf{e}_k^\top, \qquad s \;=\; 1, 2, \ldots, N-1, \tag{4.10}$$

nachgewiesen werden, was im Fall $s = N-1$ gerade die letzte Darstellung des Lemmas liefert. Die Darstellung in (4.10) ist sicher richtig für $s = 1$, und wir nehmen nun an, dass sie richtig ist für ein $1 \le s \le N-2$. Dann erhält man wie behauptet

$$\begin{aligned} F_1^{-1} \cdots F_{s+1}^{-1} \;&=\; \big(I \;+\; \sum_{k=1}^{s} \mathbf{f}_k \mathbf{e}_k^\top\big)\big(I \;+\; \mathbf{f}_{s+1}\mathbf{e}_{s+1}^\top\big) \\ &=\; I \;+\; \mathbf{f}_{s+1}\mathbf{e}_{s+1}^\top \;+\; \sum_{k=1}^{s} \mathbf{f}_k \mathbf{e}_k^\top \;+\; \sum_{k=1}^{s} \mathbf{f}_k \overbrace{(\mathbf{e}_k^\top \mathbf{f}_{s+1})}^{=\,0\,\in\,\mathbb{R}} \mathbf{e}_{s+1}^\top. \end{aligned}$$

□

Lemma 4.18. *Sei F_s eine Eliminationsmatrix vom Index s in der Darstellung (4.9), und sei P eine elementare Permutationsmatrix mit zugehöriger Elementarpermutation π von der Form (4.7) mit Zahlen $s+1 \le q,\ r \le N$. Dann entsteht PF_sP aus F_s durch Vertauschen der Einträge q und r in der s-ten Spalte, das heißt,*

$$PF_sP \;=\; I \;-\; (P\mathbf{f}_s)\mathbf{e}_s^\top.$$

BEWEIS. Die Aussage ergibt sich unmittelbar:

$$PF_sP \;=\; \underbrace{P^2}_{=\,I} \;-\; (P\mathbf{f}_s)(\underbrace{\mathbf{e}_s^\top P}_{=\,\mathbf{e}_s^\top}),$$

wobei sowohl Bemerkung 4.14 als auch die zweite Identität in (4.6) für $M = 1$ sowie die Tatsache $q, r \ge s+1$ berücksichtigt sind. □

4.3.3 Die Faktorisierung $PA = LR$

Vorbereitend wird die bereits vorgestellte Vorgehensweise beim Gauß-Algorithmus mit Spaltenpivotstrategie[2] als Folge spezieller Matrix-Operationen beschrieben: es werden sukzessive Matrizen

$$A^{(s+1)} \;=\; F_s P_s A^{(s)} \qquad \text{für} \quad s \;=\; 1, 2, \ldots, N-1$$

[2] siehe hierzu Schema 4.3, Algorithmus 4.6 sowie Bemerkung 4.16

mit

$$F_s = \begin{pmatrix} 1 & & & & \\ & \ddots & & & \\ & & 1 & & \\ & & -\ell_{s+1,s} & \ddots & \\ & & \vdots & & \ddots \\ & & -\ell_{Ns} & & 1 \end{pmatrix}, \quad \begin{aligned} \ell_{js} &= \frac{a_{js}^{(s)}}{a_{p_s s}^{(s)}}, \quad j = s+1,\dots,N, \quad j \neq p_s, \\ \ell_{p_s s} &= \frac{a_{ss}^{(s)}}{a_{p_s s}^{(s)}}, \end{aligned} \tag{4.11}$$

$$P_s = \begin{pmatrix} 1 & & & & & & & \\ & \ddots & & & & & & \\ & & 1 & & & & & \\ & & & 0 & & & 1 & \\ & & & & 1 & & & \\ & & & & & \ddots & & \\ & & & & & & 1 & \\ & & & 1 & & & 0 & \\ & & & & & & & 1 \\ & & & & & & & \ddots \\ & & & & & & & & 1 \end{pmatrix} \begin{matrix} \\ \\ \\ \leftarrow \text{ Zeile } s \\ \\ \\ \\ \leftarrow \text{ Zeile } p_s \\ \\ \\ \end{matrix} \qquad A^{(s)} = \begin{pmatrix} * & \cdots & \cdots & \cdots & * \\ & \ddots & & & \vdots \\ & & * & \cdots & * \\ & & a_{ss}^{(s)} & \cdots & a_{sN}^{(s)} \\ & & \vdots & & \vdots \\ & & a_{Ns}^{(s)} & \cdots & a_{NN}^{(s)} \end{pmatrix}, \tag{4.12}$$

berechnet, wobei $p_s \geq s$ die Position derjenigen Zeile aus der Matrix $A^{(s)}$ mit dem Pivotelement bezeichnet. Es kann nun die Faktorisierung $PA = LR$ explizit angegeben werden.

Theorem 4.19. *Mit den Notationen (4.11)-(4.12) gilt für* $P = P_{N-1}\cdots P_1$, $R = A^{(N)}$ *sowie*

$$L = \begin{pmatrix} 1 & & & & \\ \widehat{\ell}_{21} & 1 & & & \\ \vdots & \widehat{\ell}_{32} & 1 & & \\ \vdots & \vdots & \ddots & \ddots & \\ \widehat{\ell}_{N1} & \widehat{\ell}_{N2} \dots & \widehat{\ell}_{N,N-1} & 1 \end{pmatrix}, \quad \textit{mit} \begin{pmatrix} 0 \\ \vdots \\ 0 \\ 1 \\ \widehat{\ell}_{s+1,s} \\ \vdots \\ \widehat{\ell}_{Ns} \end{pmatrix} := P_{N-1}\dots P_{s+1} \begin{pmatrix} 0 \\ \vdots \\ 0 \\ 1 \\ \ell_{s+1,s} \\ \vdots \\ \ell_{Ns} \end{pmatrix}, \tag{4.13}$$

die Identität $PA = LR$.

BEWEIS. Für $s = 1, 2, \dots$ gilt:

$$\begin{aligned} A^{(2)} &= F_1 P_1 A = F_1 (P_1 A), \\ A^{(3)} &= F_2 P_2 A^{(2)} = F_2 P_2 (F_1 \overbrace{P_2 P_2}^{=I} P_1 A) = F_2 (P_2 F_1 P_2)(P_2 P_1 A) \\ A^{(4)} &= F_3 P_3 A^{(3)} = F_3 P_3 \big(F_2 \underbrace{P_3 P_3}_{=I} P_2 F_1 P_2 \underbrace{P_3 P_3}_{=I} P_2 P_1 A\big) \\ &= F_3 (P_3 F_2 P_3)(P_3 P_2 F_1 P_2 P_3)(P_3 P_2 P_1 A), \end{aligned}$$

und so weiter, was schließlich auf

$$R \;=\; A^{(N)} \;=\; \widehat{F}_{N-1}\cdots\widehat{F}_1 PA \tag{4.14}$$

führt mit den Eliminationsmatrizen

$$\widehat{F}_s \;=\; P_{N-1}\cdots P_{s+1}F_s P_{s+1}\cdots P_{N-1} \overset{(*)}{=} \begin{pmatrix} 1 & & & & \\ & \ddots & & & \\ & & 1 & & \\ & & -\widehat{\ell}_{s+1,s} & \ddots & \\ & & \vdots & & \ddots \\ & & -\widehat{\ell}_{Ns} & & 1 \end{pmatrix},\ s = 1,\ldots,N-1,$$

wobei in der Identität $(*)$ noch Lemma 4.18 berücksichtigt ist. Eine Umformung von (4.14) liefert dann die Identität

$$PA \;=\; (\widehat{F}_1^{-1}\cdots\widehat{F}_{N-1}^{-1})R \overset{(**)}{=} LR,$$

wobei in $(**)$ noch Lemma 4.17 eingeht. Dies komplettiert den Beweis. □

Bemerkung 4.20. In praktischen Implementierungen werden die frei werdenden Anteile des unteren Dreiecks der Matrix A sukzessive überschrieben mit den Einträgen der unteren Dreiecksmatrix L, und in dem oberen Dreieck der Matrix A ergeben sich die Einträge der Dreiecksmatrix R. Die Permutationsmatrix P lässt sich einfach in Form eines Buchhaltungsvektors $r \in \mathbb{R}^N$ berechnen: es gilt

$$P\begin{pmatrix} b_1 \\ \vdots \\ b_N \end{pmatrix} \;=\; \begin{pmatrix} b_{r_1} \\ \vdots \\ b_{r_N} \end{pmatrix} \qquad \text{für} \quad \begin{pmatrix} r_1 \\ \vdots \\ r_N \end{pmatrix} := P\begin{pmatrix} 1 \\ \vdots \\ N \end{pmatrix},$$

was man unmittelbar aus Theorem 4.10 erschließt. △

Beispiel 4.21 (Oevel [78]). Die durch Theorem 4.19 vorgegebene Vorgehensweise soll anhand der Matrix

$$A \;=\; \begin{pmatrix} 0 & 0 & 1 & 1 \\ 2 & 2 & 2 & 2 \\ 1 & 2 & 2 & 2 \\ 1 & 2 & 3 & 6 \end{pmatrix} \in \mathbb{R}^{4\times 4}$$

exemplarisch vorgestellt werden. Nach Anhängen des für die Speicherung der Zeilenpermutationen zuständigen Buchhaltungsvektors geht man so vor (unterhalb der Treppe ergeben sich sukzessive die Einträge der unteren Dreiecksmatrix L aus (4.13)):

$$\begin{pmatrix} 0 & 0 & 1 & 1 \\ \textcircled{2} & 2 & 2 & 2 \\ 1 & 2 & 2 & 2 \\ 1 & 2 & 3 & 6 \end{pmatrix}, \begin{pmatrix} 1 \\ 2 \\ 3 \\ 4 \end{pmatrix} \quad \overset{\text{Zeilentausch}}{\longrightarrow} \quad \begin{pmatrix} 2 & 2 & 2 & 2 \\ 0 & 0 & 1 & 1 \\ 1 & 2 & 2 & 2 \\ 1 & 2 & 3 & 6 \end{pmatrix}, \begin{pmatrix} 2 \\ 1 \\ 3 \\ 4 \end{pmatrix}$$

$$\text{Elimination} \xrightarrow{} \left(\begin{array}{c|ccc} 2 & 2 & 2 & 2 \\ \hline 0 & 0 & 1 & 1 \\ 1/2 & \textcircled{1} & 1 & 1 \\ 1/2 & 1 & 2 & 5 \end{array}\right), \begin{pmatrix} 2 \\ 1 \\ 3 \\ 4 \end{pmatrix} \text{Zeilentausch} \xrightarrow{} \left(\begin{array}{c|ccc} 2 & 2 & 2 & 2 \\ \hline 1/2 & 1 & 1 & 1 \\ 0 & 0 & 1 & 1 \\ 1/2 & 1 & 2 & 5 \end{array}\right), \begin{pmatrix} 2 \\ 3 \\ 1 \\ 4 \end{pmatrix}$$

$$\text{Elimination} \xrightarrow{} \left(\begin{array}{c|ccc} 2 & 2 & 2 & 2 \\ \hline 1/2 & 1 & 1 & 1 \\ 0 & 0 & \textcircled{1} & 1 \\ 1/2 & 1 & 1 & 4 \end{array}\right), \begin{pmatrix} 2 \\ 3 \\ 1 \\ 4 \end{pmatrix} \text{Elimination} \xrightarrow{} \left(\begin{array}{c|ccc} 2 & 2 & 2 & 2 \\ \hline 1/2 & 1 & 1 & 1 \\ 0 & 0 & 1 & 1 \\ 1/2 & 1 & 1 & 3 \end{array}\right), \begin{pmatrix} 2 \\ 3 \\ 1 \\ 4 \end{pmatrix},$$

wobei das jeweils gewählte Pivotelement $*$ eingekreist dargestellt ist, $\textcircled{*}$. Es ergibt sich somit das folgende Resultat:

$$L = \begin{pmatrix} 1 & & & \\ 1/2 & 1 & & \\ 0 & 0 & 1 & \\ 1/2 & 1 & 1 & 1 \end{pmatrix}, \quad R = \begin{pmatrix} 2 & 2 & 2 & 2 \\ & 1 & 1 & 1 \\ & & 1 & 1 \\ & & & 3 \end{pmatrix}, \quad P \begin{pmatrix} b_1 \\ b_2 \\ b_3 \\ b_4 \end{pmatrix} = \begin{pmatrix} b_2 \\ b_3 \\ b_1 \\ b_4 \end{pmatrix}. \qquad \triangle$$

4.4 *LR*-Faktorisierung

In gewissen Situationen ist es möglich und zwecks Bewahrung etwaiger Bandstrukturen der Matrix A auch wünschenswert, auf eine Pivotstrategie zu verzichten und eine LR-Faktorisierung von der Form

$$A = \begin{pmatrix} 1 & & & \\ \ell_{21} & 1 & & \\ \vdots & \ddots & \ddots & \\ \ell_{N1} & \dots & \ell_{N,N-1} & 1 \end{pmatrix} \begin{pmatrix} r_{11} & r_{12} & \dots & r_{NN} \\ & r_{22} & & \vdots \\ & & \ddots & \vdots \\ & & & r_{NN} \end{pmatrix} \tag{4.15}$$

zu bestimmen. Ein direkter Ansatz zur Bestimmung einer solchen LR-Faktorisierung besteht darin, das Gleichungssystem (4.15) als N^2 Bestimmungsgleichungen für die N^2 gesuchten Größen r_{jk} $(j \le k)$ und ℓ_{jk} $(j > k)$ aufzufassen:

$$a_{jk} = \sum_{s=1}^{\min\{j,k\}} \ell_{js} r_{sk}, \qquad j,\ k = 1, 2, \dots, N. \tag{4.16}$$

Dabei gibt es verschiedene Reihenfolgen, mit denen man aus den Gleichungen in (4.16) die Einträge von L und R berechnen kann. Beispielsweise führt eine Berechnung der Zeilen von R und der Spalten von L entsprechend der *Parkettierung nach Crout*

$$\left(\begin{array}{c|c|c|c|c} 1a\ \rightarrow & & & & \\ \hline 1b & 2a\ \rightarrow & & & \\ \hline \downarrow & 2b & 3a\ \rightarrow & & \\ \hline & \downarrow & 3b & 4a\ \rightarrow & \\ \hline & & \downarrow & 4b & 5 \end{array}\right) \tag{4.17}$$

auf den in Schema 4.4 beschriebenen Algorithmus zur Bestimmung der LR-Faktorisierung.

```
for n = 1 : N
    for k = n : N     r_nk  =  a_nk − Σ_{s=1}^{n−1} ℓ_ns r_sk ;     end
    for j = n + 1 : N     ℓ_jn  =  (a_jn − Σ_{s=1}^{n−1} ℓ_js r_sn)/r_nn ;     end
end
```

Schema 4.4: LR-Faktorisierung nach Crout

Wie man leicht abzählt, fallen bei diesem Algorithmus $(2N^3/3)(1+\mathcal{O}(1/N))$ arithmetische Operationen an (Aufgabe 4.10).

4.5 Cholesky-Faktorisierung symmetrischer, positiv definiter Matrizen

4.5.1 Grundbegriffe

Gegenstand des vorliegenden Abschnitts sind die in der folgenden Definition betrachteten Matrizen.

Definition 4.22. Eine Matrix $A \in \mathbb{R}^{N\times N}$ heißt *symmetrisch*, falls $A = A^\top$ gilt. Sie heißt *positiv definit*, falls $x^\top Ax > 0$ für alle $0 \neq x \in \mathbb{R}^N$ gilt.

Beispielsweise sind die bei der kubischen Spline-Interpolation auftretenden Systemmatrizen zur Berechnung der Momente symmetrisch und positiv definit. Einzelheiten dazu werden in Abschnitt 4.5.3 nachgetragen. Für positiv definite Matrizen wird nun eine der LR-Faktorisierung ähnliche Faktorisierung mit einem geringeren Speicherplatzbedarf vorgestellt. Wir beginnen mit einem vorbereitenden Lemma.

Lemma 4.23. *Die Matrix $A \in \mathbb{R}^{N\times N}$ sei symmetrisch: Dann gilt:*

(a) Die Matrix A ist positiv definit genau dann, wenn alle Eigenwerte von A positiv sind.

(b) ——— « ——— *genau dann, wenn alle* Hauptuntermatrizen

$$\begin{pmatrix} a_{rr} & \cdots & a_{rs} \\ \vdots & \ddots & \vdots \\ a_{sr} & \cdots & a_{ss} \end{pmatrix} \in \mathbb{R}^{(s-r+1)\times(s-r+1)} \qquad \textit{für} \quad 1 \le r \le s \le N \tag{4.18}$$

von A positiv definit sind.

(c) Ist die Matrix A positiv definit, so gilt $\det(A) > 0$.

BEWEIS. (a) Ist die Matrix A positiv definit und $\lambda \in \mathbb{R}$ ein Eigenwert von A, so gilt für einen beliebigen Eigenvektor $0 \neq x \in \mathbb{R}^N$ von A zum Eigenwert λ Folgendes:

$$0 \;<\; x^\top A x \;=\; \lambda \underbrace{x^\top x}_{>\,0}$$

und damit $\lambda > 0$. Für den Nachweis der anderen Richtung der Äquivalenz benötigen wir die für symmetrische Matrizen A existierende Faktorisierung

$$\left.\begin{aligned} A \;=\; UDU^\top \qquad & U \in \mathbb{R}^{N\times N} \text{ regulär}, \qquad U^{-1} = U^\top, \\ & D \;=\; \operatorname{diag}(\lambda_1,\ldots,\lambda_N) \in \mathbb{R}^{N\times N}. \end{aligned}\right\} \tag{4.19}$$

Die Zahlen $\lambda_1,\ldots,\lambda_N \in \mathbb{R}$ sind dabei gerade die entsprechend ihrer Vielfachheit gezählten Eigenwerte der Matrix A, und diese seien nun allesamt als positiv angenommen. Dann ist die Matrix D positiv definit, denn es gilt

$$z^\top D z \;=\; \sum_{j=1}^{N} \lambda_j z_j^2 \;>\; 0 \qquad \text{für} \quad 0 \neq z = (z_j) \in \mathbb{R}^N.$$

Damit gilt auch

$$x^\top A x \;=\; (U^\top x)^\top D (U^\top x) \;>\; 0 \qquad \text{für} \quad 0 \neq x \in \mathbb{R}^N,$$

so dass die Matrix A ebenfalls positiv definit ist.

(b) Falls alle Hauptuntermatrizen von A positiv definit sind, so ist insbesondere auch die Matrix A positiv definit. Für den Nachweis der anderen Richtung der betrachteten Äquivalenz sei nun die Matrix A als positiv definit angenommen, und es sei $B \in \mathbb{R}^{(s-r+1)\times(s-r+1)}$ eine Hauptuntermatrix der Form (4.18). Die Matrix B ist offensichtlich symmetrisch, und sei nun $0 \neq x = (x_j)_{j=r}^{s} \in \mathbb{R}^{s-r+1}$. Für $z = (z_j)_{j=1}^{N} \in \mathbb{R}^N$ mit

$$z_j \;=\; \begin{cases} x_j, & r \le j \le s, \\ 0, & \text{sonst} \end{cases}$$

gilt dann $z \neq 0$ und

$$x^\top B x \;=\; \sum_{j,k=r}^{s} a_{jk}\, x_j\, x_k \;=\; \sum_{j,k=1}^{N} a_{jk}\, z_j\, z_k \;=\; z^\top A z \;>\; 0.$$

(c) Hier zieht man eine Faktorisierung von der Form (4.19) heran und erhält daraus wie angegeben

$$\det(A) \;=\; \det(U^{-1})\det(D)\det(U) \;=\; \det(D) \;=\; \prod_{j=1}^{N} \lambda_j \;>\; 0. \qquad \square$$

Theorem 4.24. *Die Matrix $A \in \mathbb{R}^{N\times N}$ sei symmetrisch und positiv definit. Dann gibt es genau eine untere Dreiecksmatrix $L = (\ell_{jk}) \in \mathbb{R}^{N\times N}$ mit $\ell_{jj} > 0$ für alle j und*

$$A \;=\; L L^\top. \tag{4.20}$$

Die Faktorisierung (4.20) wird als Cholesky-Faktorisierung *von A bezeichnet.*

BEWEIS. Der Beweis wird mit vollständiger Induktion über N geführt. Für $N = 1$ ist eine positiv definite Matrix $A = (\alpha) \in \mathbb{R}^{1\times 1}$ eine positive Zahl $\alpha > 0$, die eindeutig in der Form

$$\alpha \;=\; \ell \cdot \ell, \qquad \ell \;=\; \sqrt{\alpha},$$

geschrieben werden kann. Wir nehmen nun an, dass für eine ganze Zahl $N \geq 1$ die Aussage des Theorems richtig ist mit $N-1$ anstelle N und betrachten dann eine symmetrische, positiv definite Matrix $A \in \mathbb{R}^{N\times N}$. Diese lässt sich in der Form

$$\left(\begin{array}{c|c} A_{N-1} & b \\ \hline b^\top & a_{NN} \end{array}\right)$$

partitionieren mit einem Vektor $b \in \mathbb{R}^{N-1}$ und einer Matrix $A_{N-1} \in \mathbb{R}^{(N-1)\times(N-1)}$, die nach Lemma 4.23 positiv definit ist. Nach Induktionsvoraussetzung gibt es eine eindeutig bestimmte untere Dreiecksmatrix $L_{N-1} = (\ell_{jk}) \in \mathbb{R}^{(N-1)\times(N-1)}$ mit $\ell_{jj} > 0$ für $j = 1, 2, \ldots, N-1$ und

$$A_{N-1} \;=\; L_{N-1} L_{N-1}^\top.$$

Die gesuchte Matrix $L \in \mathbb{R}^{N\times N}$ setzt man nun in der Form

$$L \;=\; \left(\begin{array}{c|c} L_{N-1} & 0 \\ \hline c^\top & \alpha \end{array}\right)$$

an mit dem Ziel, einen Vektor $c \in \mathbb{R}^{N-1}$ und eine Zahl $\alpha > 0$ so zu bestimmen, dass

$$A = \left(\begin{array}{c|c} A_{N-1} & b \\ \hline b^\top & a_{NN} \end{array}\right) \stackrel{!}{=} \left(\begin{array}{c|c} L_{N-1} & 0 \\ \hline c^\top & \alpha \end{array}\right) \left(\begin{array}{c|c} L_{N-1}^\top & c \\ \hline 0^\top & \alpha \end{array}\right) \tag{4.21}$$

gilt. Gleichheit in (4.21) liegt genau dann vor, wenn

$$\begin{aligned} L_{N-1}c &= b \\ c^\top c + \alpha^2 &= a_{NN} \end{aligned} \tag{4.22}$$

gilt, und die erste dieser beiden Gleichungen besitzt sicher genau eine Lösung $c = L_{N-1}^{-1}b$, da $L_{N-1} \in \mathbb{R}^{(N-1)\times(N-1)}$ als untere Dreiecksmatrix mit nichtverschwindenden Diagonaleinträgen regulär ist. Auch die zweite Gleichung (4.22) besitzt eine Lösung $\alpha \in \mathbb{C}$, mit der dann die Faktorisierung (4.21) gültig ist. Wir zeigen abschließend $\alpha^2 > 0$; dann kann in (4.22) in eindeutiger Weise $\alpha > 0$ gewählt werden. Wegen (4.21) gilt

$$\det(A) = \det\left(\begin{array}{c|c} L_{N-1} & 0 \\ \hline c^\top & \alpha \end{array}\right) \det\left(\begin{array}{c|c} L_{N-1}^\top & c \\ \hline 0^\top & \alpha \end{array}\right) = \det(L_{N-1})^2\alpha^2$$

und wegen $\det(A) > 0$ (siehe Lemma 4.23) sowie der Regularität von L_{N-1} folgt wie behauptet $\alpha^2 > 0$. □

Bemerkung 4.25. Der im Beweis von Theorem 4.24 vorgestellte Algorithmus zur Berechnung einer Faktorisierung $A = LL^\top$ wird als *Quadratwurzelverfahren* bezeichnet. △

4.5.2 Die Berechnung einer Faktorisierung $A = LL^\top$ für positiv definite Matrizen $A \in \mathbb{R}^{N\times N}$

In einem direkten Ansatz zur Bestimmung einer solchen $LL^\top$-Faktorisierung fasst man die Matrix-Gleichung (4.20) als $N(N+1)/2$ Bestimmungsgleichungen für die $N(N+1)/2$ gesuchten Einträge ℓ_{jk} $(j \geq k)$ auf:

$$a_{jk} = \sum_{s=1}^{k} \ell_{js}\ell_{ks}, \qquad 1 \leq k \leq j \leq N. \tag{4.23}$$

Spaltenweise Berechnung der Einträge der unteren Dreiecksmatrix $L \in \mathbb{R}^{N\times N}$ aus den Gleichungen in (4.23) führt auf den in Schema 4.5 beschriebenen Algorithmus.

```
for n = 1 : N
    ℓ_nn = (a_nn − Σ_{k=1}^{n−1} ℓ_nk^2)^{1/2};
    for j = n+1 : N    ℓ_jn = (a_jn − Σ_{k=1}^{n−1} ℓ_jk ℓ_nk)/ℓ_nn;    end
end
```

Schema 4.5: $LL^{\top}$-Faktorisierung

Theorem 4.26. *Zur Berechnung einer Cholesky-Faktorisierung sind insgesamt* $(N^3/3)\cdot(1+\mathcal{O}(\frac{1}{N}))$ *arithmetische Operationen durchzuführen.*

BEWEIS. Nach Schema 4.5 summiert sich die Zahl der genannten Operationen zu

$$\begin{aligned}
&\sum_{n=1}^{N}\Big(2n-1+\sum_{j=n+1}^{N}(2n-1)\Big) = \sum_{n=1}^{N}\big((N+1-n)(2n-1)\big)\\
&= -\sum_{n=1}^{N}(N+1-n) + 2\sum_{n=1}^{N}(N+1-n)n\\
&= -\sum_{n=1}^{N} n + 2(N+1)\sum_{n=1}^{N} n - 2\sum_{n=1}^{N} n^2\\
&= (2N+1)\frac{N(N+1)}{2} - 2\frac{N(N+1)(2N+1)}{6} = \frac{N^3}{3}\big(1+\mathcal{O}\big(\tfrac{1}{N}\big)\big).
\end{aligned}$$

□

4.5.3 Eine Klasse positiv definiter Matrizen

Zu Beginn des vorliegenden Abschnitts 4.5 wurde bereits darauf hingewiesen, dass beispielsweise die bei der kubischen Spline-Interpolation auftretenden Systemmatrizen zur Berechnung der Momente symmetrisch und positiv definit sind. In diesem Abschnitt wird hierfür noch der Nachweis geliefert. Wir beginnen mit einem vorbereitenden Lemma.

Lemma 4.27. *Die Matrix* $A \in \mathbb{R}^{N\times N}$ *sei symmetrisch und strikt diagonaldominant, und sie besitze ausschließlich positive Diagonaleinträge. Dann ist die Matrix A positiv definit.*

BEWEIS. Gemäß Teil (a) von Lemma 4.23 genügt es nachzuweisen, dass alle Eigenwerte der Matrix A positiv sind. Zunächst stellt man fest, dass zu jedem Eigenwert $\lambda \in \mathbb{R}$ der Matrix $A = (a_{jk})$ notwendigerweise ein Index $j \in \{1, 2, \ldots, N\}$ mit

$$|a_{jj} - \lambda| \le \sum_{\substack{k=1\\ k\neq j}}^{N} |a_{jk}| \tag{4.24}$$

existieren muss[3], da ansonsten die Matrix $A-\lambda I$ strikt diagonaldominant und damit regulär wäre. Aus der Abschätzung (4.24) erhält man dann die Aussage des Lemmas,

$$a_{jj} - \lambda \;\leq\; |a_{jj}-\lambda| \;\leq\; \sum_{\substack{k=1\\k\neq j}}^{N} |a_{jk}| \qquad \text{bzw.} \qquad \lambda \;\geq\; a_{jj} - \sum_{\substack{k=1\\k\neq j}}^{N} |a_{jk}| \;>\; 0.$$

□

Beispiel 4.28. In Abschnitt 2.4 ab Seite 25 sind Verfahren zur Berechnung interpolierender kubischer Splinefunktionen mit natürlichen, vollständigen beziehungsweise periodischen Randbedingungen vorgestellt worden. Die dabei jeweils entstehenden linearen Gleichungssysteme zur Berechnung der Momente beinhalten Systemmatrizen, die den Bedingungen von Lemma 4.27 genügen und somit positiv definit sind. Diese linearen Gleichungssysteme lassen sich also jeweils mit einer Cholesky-Faktorisierung lösen. △

4.6 Bandmatrizen

Bei der Diskretisierung von gewöhnlichen oder partiellen Differenzialgleichungen oder auch der Berechnung der Momente kubischer Splinefunktionen ergeben sich lineare Gleichungssysteme $Ax = b$, bei denen $A = (a_{jk}) \in \mathbb{R}^{N\times N}$ eine *Bandmatrix* ist, das heißt, es gilt $a_{jk} = 0$ für $k < j-p$ oder $k > j+q$ mit gewissen Zahlen p, q:

$$A \;=\; \begin{pmatrix} a_{11} & \cdots & a_{1,q+1} & & & \\ \vdots & \ddots & & \ddots & & \\ a_{p+1,1} & & \ddots & & \ddots & \\ & \ddots & & \ddots & & a_{N-q,N} \\ & & \ddots & & \ddots & \vdots \\ & & & a_{N,N-p} & \cdots & a_{NN} \end{pmatrix}. \tag{4.25}$$

Bei solchen Problemstellungen lässt sich der zu betreibende Aufwand bei allen in diesem Kapitel angesprochenen Methoden verringern. (Ausgenommen sind Pivotstrategien, da sich hier die Bandstruktur nicht auf die Faktorisierung überträgt.)

Exemplarisch soll das Vorgehen für Bandmatrizen am Beispiel der LR-Faktorisie-

[3] Diese Eigenschaft wird nochmals in Theorem 12.9 auf Seite 327 verwendet.

rung demonstriert werden: der Ansatz

$$\begin{pmatrix} a_{11} & \cdots & a_{1,q+1} & & & \\ \vdots & \ddots & & \ddots & & \\ a_{p+1,1} & & \ddots & & \ddots & \\ & \ddots & & \ddots & & a_{N-q,N} \\ & & \ddots & & \ddots & \vdots \\ & & & a_{N,N-p} & \cdots & a_{NN} \end{pmatrix}$$

$$= \begin{pmatrix} 1 & & & & & \\ \ell_{21} & \ddots & & & & \\ \vdots & \ddots & \ddots & & & \\ \ell_{p+1,1} & & \ddots & \ddots & & \\ & \ddots & & \ddots & \ddots & \\ & & \ell_{N,N-p} & \cdots & \ell_{N,N-1} & 1 \end{pmatrix} \begin{pmatrix} r_{11} & \cdots & r_{1,q+1} & & & \\ & \ddots & & \ddots & & \\ & & \ddots & & r_{N-q,N} & \\ & & & \ddots & \vdots & \\ & & & & r_{NN} & \end{pmatrix}$$

beziehungsweise in Komponentenschreibweise

$$a_{jk} = \sum_{s=s_0}^{\min\{j,k\}} \ell_{js} r_{sk}, \quad \begin{array}{l} j = 1,\ldots,N, \quad k = \max\{1,\ j-p\},\ldots,\min\{\,j+q,\ N\,\}, \\ s_0 := \max\{1,\ j-p,\ k-q\,\} \end{array}$$

führt bei einer Parkettierung wie in (4.17) auf den in Schema 4.6 angegebenen Algorithmus zur Bestimmung der LR-Faktorisierung der Bandmatrix A.

```
for n = 1 : N
    for k = n : min{n + q, N}
        s0 = max{1, n − p, k − q};      rnk = ank − Σ_{s=s0}^{n−1} ℓns rsk;
    end
    for j = n + 1 : min{n + p, N}
        s0 = max{1, j − p, n − q};      ℓjn = (ajn − Σ_{s=s0}^{n−1} ℓjs rsn)/rnn;
    end
end
```

Schema 4.6: LR-Faktorisierung für Bandmatrizen

4.7 Normen und Fehlerabschätzungen

In diesem Abschnitt soll der Einfluss von Störungen[4] der Matrix $A \in \mathbb{R}^{N\times N}$ beziehungsweise des Vektors $b \in \mathbb{R}^N$ auf die Lösung des linearen Gleichungssystems

[4] Solche Störungen können durch Mess- oder Rundungsfehler verursacht werden.

$Ax = b$ untersucht werden, für die Einzelheiten sei auf Abschnitt 4.7.5 verwiesen. Zuvor werden in den nun folgenden Abschnitten 4.7.1–4.7.4 die nötigen Voraussetzungen geschaffen.

Dabei werden zunächst allgemeiner Vektoren aus $\mathbb{K}^N$ beziehungsweise Matrizen aus $\mathbb{K}^{N\times N}$ zugelassen, wobei entweder $\mathbb{K} = \mathbb{R}$ oder $\mathbb{K} = \mathbb{C}$ ist. Dies ermöglicht später die Herleitung von Schranken sowohl für Nullstellen von Polynomen als auch für Eigenwerte von Matrizen.

4.7.1 Normen

Definition 4.29. Sei $\mathcal{V}$ ein beliebiger Vektorraum über $\mathbb{K}$. Eine Abbildung $\|\cdot\| : \mathcal{V} \to \mathbb{R}_+$ heißt *Norm*, falls Folgendes gilt:

$$\begin{aligned} \|x + y\| &\le \|x\| + \|y\| && (x, y \in \mathcal{V}) && (\textit{Dreiecksungleichung});\\ \|\alpha x\| &= |\alpha|\|x\| && (x \in \mathcal{V},\quad \alpha \in \mathbb{K}) && (\textit{positive Homogenität});\\ \|x\| &= 0 \iff x = 0 && (x \in \mathcal{V}). \end{aligned}$$

Eine Norm $\|\cdot\| : \mathbb{K}^N \to \mathbb{R}_+$ wird auch als *Vektornorm* bezeichnet, und entsprechend wird eine Norm $\|\cdot\| : \mathbb{K}^{N\times N} \to \mathbb{R}_+$ auch *Matrixnorm* genannt.

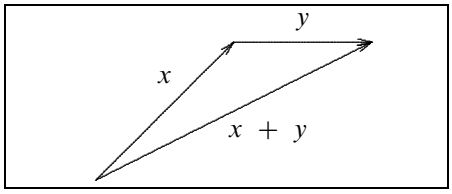


Bild 4.1: Illustration der Dreiecksungleichung

Lemma 4.30. *Für eine Norm $\|\cdot\| : \mathcal{V} \to \mathbb{R}_+$ gilt die* umgekehrte Dreiecksungleichung

$$|\|x\| - \|y\|| \le \|x - y\|, \qquad x, y \in \mathcal{V}.$$

BEWEIS. Zum einen gilt $\|x\| = \|x - y + y\| \le \|x - y\| + \|y\|$ und somit

$$\|x\| - \|y\| \le \|x - y\|. \tag{4.26}$$

Vertauschung von x und in y in (4.26) liefert dann

$$\|y\| - \|x\| \le \|x - y\|, \tag{4.27}$$

und (4.26)–(4.27) zusammen liefern die umgekehrte Dreiecksungleichung. □

Korollar 4.31. *Eine Norm $\|\cdot\| : \mathcal{V} \to \mathbb{R}_+$ ist stetig, das heißt, für beliebige Folgen $(x_n) \subset \mathcal{V}$ und Elemente $x \in \mathcal{V}$ folgt aus der Konvergenz $\|x_n - x\| \to 0$ für $n \to \infty$ auch $\|x_n\| \to \|x\|$ für $n \to \infty$.*

Im Folgenden werden einige spezielle Vektornormen vorgestellt.

Theorem 4.32. *Durch*

$$\begin{aligned}
\|x\|_2 &= \Big(\sum_{k=1}^{N} |x_k|^2\Big)^{1/2} && \text{(euklidische Norm);} \\
\|x\|_\infty &= \max_{k=1..N} |x_k| && \text{(Maximumnorm);} && (x \in \mathbb{K}^N); \\
\|x\|_1 &= \sum_{k=1}^{N} |x_k| && \text{(Summennorm);}
\end{aligned}$$

sind jeweils Normen auf $\mathbb{K}^N$ definiert.

BEWEIS. Der Nachweis dafür, dass die Maximum- und Summennorm tatsächlich die Normeigenschaften erfüllen, ist elementar und wird an dieser Stelle nicht geführt. Für die euklidische Norm resultiert die Dreiecksungleichung aus der cauchy-schwarzschen Ungleichung: für $x,\ y \in \mathbb{K}^N$ gilt

$$\begin{aligned}
\|x+y\|_2^2 &= (x+y)^{\mathrm{H}}(x+y) = \overbrace{x^{\mathrm{H}}x}^{=\|x\|_2^2} + \overbrace{2\mathrm{Re}\, x^{\mathrm{H}}y}^{\le 2\|x\|_2\|y\|_2} + \overbrace{y^{\mathrm{H}}y}^{=\|y\|_2^2} \\
&\le (\|x\|_2 + \|y\|_2)^2,
\end{aligned}$$

wobei Re z den Realteil einer komplexen Zahl $z \in \mathbb{C}$ bezeichnet. □

Man kann zeigen, dass je zwei verschiedene Normen $\|\cdot\|,\ |||\cdot||| : \mathbb{K}^N \to \mathbb{R}_+$ äquivalent in dem Sinne sind, dass es Konstanten $c_1,\ c_2 > 0$ gibt mit

$$c_1\|x\| \le |||x||| \le c_2\|x\|, \qquad x \in \mathbb{K}^N.$$

Konkret gelten für die in Theorem 4.32 aufgeführten Vektornormen die folgenden Abschätzungen:

$$\|x\|_\infty \le \|x\|_2 \le \sqrt{N}\|x\|_\infty, \tag{4.28}$$

$$\|x\|_\infty \le \|x\|_1 \le N\|x\|_\infty, \tag{4.29}$$

$$\|x\|_2 \le \|x\|_1 \le \sqrt{N}\|x\|_2. \tag{4.30}$$

Die (nicht zu verbessernden) Abschätzungen in (4.28)–(4.29) erhält man leicht, und die erste Abschätzung in (4.30) erhält man wie folgt (wobei o.B.d.A. $x \neq 0$ angenommen sei):

$$y := \frac{x}{\|x\|_1} \qquad \rightsquigarrow \qquad \|x\|_2 = \|x\|_1\|y\|_2 \le \|x\|_1\|y\|_1^{1/2} = \|x\|_1.$$

Die zweite Abschätzung in (4.30) schließlich folgt aus der cauchy-schwarzschen Ungleichung:

$$\|x\|_1 = \sum_{k=1}^{N} 1\cdot|x_k| \le \Big(\sum_{k=1}^{N} 1\Big)^{1/2}\Big(\sum_{k=1}^{N}|x_k|^2\Big)^{1/2} = \sqrt{N}\|x\|_2.$$

Somit werden für große Zahlen $N \in \mathbb{N}$ die jeweils zweiten Abschätzungen in (4.28)–(4.30) praktisch bedeutungslos aufgrund der Größe der auftretenden Koeffizienten.

Bemerkung 4.33. Allgemeiner ist für jedes $1 \le p < \infty$ durch

$$\|x\|_p := \Big(\sum_{k=1}^{N} |x_k|^p\Big)^{1/p}, \qquad x \in \mathbb{K}^N,$$

eine Norm auf $\mathbb{K}^N$ definiert mit der Eigenschaft $\lim_{p\to\infty} \|x\|_p = \|x\|_\infty$ für $x \in \mathbb{K}^N$. △

Im Folgenden werden drei spezielle Matrixnormen vorgestellt. Dabei erhält nur die letzte der drei Normen eine besondere Indizierung, für die beiden anderen werden später eigene Bezeichnungen vergeben (siehe Theorem 4.40).

Theorem 4.34. *Durch*

$$\begin{aligned}
\|A\| &= \max_{j=1..N} \sum_{k=1}^{N} |a_{jk}| && \text{(Zeilensummennorm);}\\
\|A\| &= \max_{k=1..N} \sum_{j=1}^{N} |a_{jk}| && \text{(Spaltensummennorm);} \quad (A = (a_{jk}) \in \mathbb{K}^{N\times N})\\
\|A\|_{\mathrm{F}} &= \Big(\sum_{j,k=1}^{N} |a_{jk}|^2\Big)^{1/2} && \text{(Frobeniusnorm)}
\end{aligned}$$

sind jeweils Normen auf $\mathbb{K}^{N\times N}$ definiert.

BEWEIS. Der Nachweis dafür, dass die Zeilen- beziehungsweise die Spaltensummennorm tatsächlich die Normeigenschaften erfüllen, ist elementar und wird an dieser Stelle nicht geführt. Jede Matrix $A \in \mathbb{K}^{N\times N}$ lässt sich als Vektor der Länge N^2 auffassen, und die Frobeniusnorm fällt dann mit der euklidischen Vektornorm in Theorem 4.32 zusammen, so dass die Frobeniusnorm tatsächlich auch die Normeigenschaften erfüllt. □

Definition 4.35. Eine Matrixnorm $\|\cdot\| : \mathbb{K}^{N\times N} \to \mathbb{R}_+$ nennt man

(a) *submultiplikativ*, falls

$$\|AB\| \le \|A\|\|B\| \qquad \big(A, B \in \mathbb{K}^{N\times N}\big);$$

(b) mit einer gegebenen Vektornorm $\|\cdot\| : \mathbb{K}^N \to \mathbb{R}_+$ *verträglich*, falls

$$\|Ax\| \le \|A\|\|x\| \qquad \big(A \in \mathbb{K}^{N\times N}, \quad x \in \mathbb{K}^N\big).$$

Definition 4.36. Sei $\|\cdot\| : \mathbb{K}^N \to \mathbb{R}_+$ eine Vektornorm. Die *induzierte Matrixnorm* ist definiert durch

$$\|A\| = \max_{0\neq x\in\mathbb{K}^N} \frac{\|Ax\|}{\|x\|}, \qquad A \in \mathbb{K}^{N\times N}. \tag{4.31}$$

Aufgrund der positiven Homogenität der Vektornorm gilt die Identität $\|A\| = \max_{x\in\mathbb{K}^N, \|x\|=1} \|Ax\|$ für jede Matrix $A \in \mathbb{K}^{N\times N}$. Wegen der Kompaktheit der Sphäre $\{x \in \mathbb{K}^N : \|x\| = 1\}$ sowie der Stetigkeit der Norm[5] wird das Maximum in (4.31) tatsächlich angenommen.

Die wesentlichen Eigenschaften induzierter Matrixnormen sind im Folgenden zusammengefasst:

Theorem 4.37. *Die durch eine Vektornorm induzierte Matrixnorm besitzt die in Definition 4.29 angegebenen Normeigenschaften, und sie ist sowohl submultiplikativ als auch verträglich mit der zugrunde liegenden Vektornorm. Es gilt $\|I\| = 1$.*

BEWEIS. Die Normeigenschaften der induzierten Matrixnorm sind leicht nachzuprüfen, gleiches gilt für die Verträglichkeit. Zum Nachweis der Submultiplikativität seien nun $\|\cdot\| : \mathbb{K}^N \to \mathbb{R}_+$ die Vektornorm mit induzierter Matrixnorm $\|\cdot\| : \mathbb{K}^{N\times N} \to \mathbb{R}_+$. Für $A,\ B \in \mathbb{K}^{N\times N}$ und $x \in \mathbb{K}^N$ mit $Bx \neq 0$ gilt dann

$$\frac{\|ABx\|}{\|x\|} = \frac{\|A(Bx)\|}{\|Bx\|}\frac{\|Bx\|}{\|x\|} \leq \|A\|\|B\|,$$

und im Fall $0 \neq x \in \mathbb{K}^N,\ Bx = 0$ gilt sicher auch $0 = \|ABx\|/\|x\| \leq \|A\|\|B\|$, so dass man insgesamt $\|AB\| \leq \|A\|\|B\|$ erhält. Die Identität $\|I\| = 1$ schließlich ist unmittelbar klar. □

4.7.2 Spezielle Matrixnormen

Definition 4.38. Für jede Matrix $B \in \mathbb{K}^{N\times N}$ bezeichnet

$$\begin{aligned} \sigma(B) &= \{\lambda \in \mathbb{C} : \lambda \text{ ist Eigenwert von } B\}, \\ r_\sigma(B) &= \max_{\lambda\in\sigma(B)} |\lambda| \end{aligned}$$

das *Spektrum* von B beziehungsweise den *Spektralradius* von B.

Theorem 4.39. *(a) Für eine Matrix $A \in \mathbb{C}^{N\times N}$ und die durch eine Vektornorm induzierte Matrixnorm $\|\cdot\| : \mathbb{C}^{N\times N} \to \mathbb{R}_+$ gilt*

$$\|A\| \geq r_\sigma(A). \tag{4.32}$$

(b) Ist $A \in \mathbb{R}^{N\times N}$ und sind alle Eigenwerte von A reell, so gilt die Ungleichung (4.32) auch für reelle Matrixnormen $\|\cdot\| : \mathbb{R}^{N\times N} \to \mathbb{R}_+$.

BEWEIS. (a) Sei $0 \neq x \in \mathbb{C}^N$ Eigenvektor zum Eigenwert $\lambda \in \mathbb{C}$ einer Matrix $A \in \mathbb{C}^{N\times N}$,

$$Ax = \lambda x.$$

Mit der zugehörigen Vektornorm $\|\cdot\| : \mathbb{C}^N \to \mathbb{R}_+$ gilt dann

$$\|A\| \geq \frac{\|Ax\|}{\|x\|} = \frac{|\lambda|\|x\|}{\|x\|} = |\lambda|.$$

[5] siehe Korollar 4.31

(b) In der vorliegenden Situation folgt die Behauptung wie in Teil (a) dieses Beweises, wobei dann jeweils "$\mathbb{C}$" durch "$\mathbb{R}$" zu ersetzen ist. □

Mit dem folgenden Theorem werden für die durch die Vektornormen $\|\cdot\|_\infty$ und $\|\cdot\|_1$ jeweils induzierten Matrixnormen handliche Darstellungen geliefert.

Theorem 4.40. *Für $A = (a_{jk}) \in \mathbb{K}^{N\times N}$ gilt*

$$\|A\|_\infty = \max_{j=1..N} \sum_{k=1}^{N} |a_{jk}| \qquad (\textit{Zeilensummennorm, siehe Theorem 4.34});$$

$$\|A\|_1 = \max_{k=1..N} \sum_{j=1}^{N} |a_{jk}| \qquad (\textit{Spaltensummennorm, } \text{———} \text{ « } \text{———}).$$

BEWEIS. Es wird zunächst die angegebene Darstellung für $\|A\|_\infty$ nachgewiesen. Für $x \in \mathbb{K}^N$ gilt

$$\|Ax\|_\infty = \max_{j=1..N} \Big| \sum_{k=1}^{N} a_{jk}x_k \Big| \le \max_{j=1..N} \sum_{k=1}^{N} |a_{jk}||x_k| \le \Big(\max_{j=1..N} \sum_{k=1}^{N} |a_{jk}| \Big) \|x\|_\infty,$$

und für den Nachweis der umgekehrten Abschätzung sei $j \in \{1, 2, \ldots, N\}$ beliebig aber fest. Für $x = (x_k) \in \mathbb{K}^N$ mit

$$x_k = \begin{cases} |a_{jk}|/a_{jk}, & \text{falls } a_{jk} \neq 0, \\ 1, & \text{sonst,} \end{cases} \qquad (k = 1, 2, \ldots, N)$$

gilt dann $\|x\|_\infty = 1$ und somit

$$\|A\|_\infty \ge \|Ax\|_\infty \ge \Big| \sum_{k=1}^{N} \underbrace{a_{jk}x_k}_{=|a_{jk}|} \Big| = \sum_{k=1}^{N} |a_{jk}|, \tag{4.33}$$

und aufgrund der freien Wahl des Indexes $j \in \{1, 2, \ldots, N\}$ in der Abschätzung (4.33) folgt die Darstellung für $\|A\|_\infty$.

Nun soll die Darstellung für $\|A\|_1$ nachgewiesen werden. Für $x \in \mathbb{K}^N$ gilt

$$\begin{aligned} \|Ax\|_1 &= \sum_{j=1}^{N} \Big| \sum_{k=1}^{N} a_{jk}x_k \Big| \le \sum_{j=1}^{N} \sum_{k=1}^{N} |a_{jk}||x_k| = \sum_{k=1}^{N} \Big(\sum_{j=1}^{N} |a_{jk}| \Big) |x_k| \\ &\le \Big(\max_{k=1..N} \sum_{j=1}^{N} |a_{jk}| \Big) \sum_{k=1}^{N} |x_k| = \Big(\max_{k=1..N} \sum_{j=1}^{N} |a_{jk}| \Big) \|x\|_1, \end{aligned}$$

und für den Nachweis der umgekehrten Abschätzung sei $n \in \{1, 2, \ldots, N\}$ beliebig aber fest. Mit dem n-ten Einheitsvektor $\mathbf{e}_n = (\delta_{kn})_k \in \mathbb{K}^N$ erhält man wegen $\|\mathbf{e}_n\|_1 = 1$ somit

$$\|A\|_1 \ge \|A\mathbf{e}_n\|_1 = \sum_{j=1}^{N} \Big| \sum_{k=1}^{N} a_{jk}\delta_{kn} \Big| = \sum_{j=1}^{N} |a_{jn}|, \tag{4.34}$$

und aufgrund der freien Wahl des Indexes $n \in \{1, 2, \ldots, N\}$ in der Abschätzung (4.34) folgt die Darstellung für $\|A\|_1$. □

Im Folgenden können die Betrachtungen wieder auf den reellen Fall beschränkt werden[6], $\mathbb{K} = \mathbb{R}$. Als unmittelbare Konsequenz aus Theorem 4.40 erhält man:

Korollar 4.41. *Für Matrizen $A \in \mathbb{R}^{N\times N}$ gilt*

$$\|A\|_\infty = \|A^\top\|_1, \qquad \|A\|_1 = \|A^\top\|_\infty.$$

Das folgende Theorem liefert für die durch die euklidische Vektornorm $\|\cdot\|_2$ induzierte Matrixnorm eine alternative Darstellung.

Theorem 4.42. *Für $A \in \mathbb{R}^{N\times N}$ gilt*

$$\|A\|_2 = r_\sigma(A^\top A)^{1/2} \qquad \text{(Spektralnorm)}.$$

BEWEIS. Es ist $A^\top A \in \mathbb{R}^{N\times N}$ eine symmetrische, positiv semidefinite Matrix, so dass es ein vollständiges System $u_1, \ldots, u_N \in \mathbb{R}^N$ von orthonormalen Eigenvektoren von $A^\top A$ gibt, das heißt,

$$A^\top A u_k = \lambda_k u_k, \qquad k = 1, 2, \ldots, N,$$

mit $\{\lambda_1, \ldots, \lambda_N\} = \sigma(A^\top A) \subset [0, \infty)$, und $u_k^\top u_\ell = \delta_{k\ell}$. Sei nun $x \in \mathbb{R}^N$ mit $\|x\|_2 = 1$ beliebig. Wegen der Orthonormalität der Eigenvektoren erhält man mit der Darstellung $x = \sum_{k=1}^N c_k u_k$ Folgendes,

$$\|Ax\|_2^2 = x^\top A^\top A x = \sum_{k=1}^N \lambda_k c_k^2 \overset{(*)}{\leq} \Big(\max_{k=1,\ldots,N} \lambda_k\Big) \sum_{k=1}^N c_k^2 = r_\sigma(A^\top A)\|x\|_2^2,$$

und in $(*)$ wird Gleichheit angenommen für einen Eigenvektor x zu einem maximalen Eigenwert von $A^\top A$. □

Die Bezeichnung "Spektralnorm" begründet sich in der folgenden Identität (4.35) für symmetrische Matrizen:

Theorem 4.43. *Sei $A \in \mathbb{R}^{N\times N}$ eine symmetrische Matrix, $A = A^\top$. Dann gilt*

$$\|A\|_2 = r_\sigma(A). \tag{4.35}$$

Für jede andere durch eine Vektornorm induzierte Matrixnorm $\|\cdot\| : \mathbb{R}^{N\times N} \to \mathbb{R}_+$ gilt

$$r_\sigma(A) \leq \|A\|. \tag{4.36}$$

[6] siehe die einführenden Bemerkungen in diesem Abschnitt 4.7

BEWEIS. Wegen $\sigma(A^2) = \{\lambda^2 : \lambda \in \sigma(A)\}$ gilt $r_\sigma(A^2) = r_\sigma(A)^2$ und daher

$$\|A\|_2 = r_\sigma(A^\top A)^{1/2} = r_\sigma(A^2)^{1/2} = \big(r_\sigma(A)^2\big)^{1/2} = r_\sigma(A).$$

Der zweite Teil des Theorems folgt nun mit Theorem 4.39. □

Beispiel 4.44. Die symmetrische Matrix

$$A = \begin{pmatrix} 1 & 3 \\ 3 & 2 \end{pmatrix}$$

besitzt die Eigenwerte $\lambda_{1/2} = (3 \pm \sqrt{37})/2$, so dass $\|A\|_2 = (3 + \sqrt{37})/2 \approx 4.541$ gilt. Weiter gilt $\|A\|_1 = \|A\|_\infty = 5$. Nebenbei zeigt dieses Beispiel, dass die in (4.28) angegebene Abschätzung $\|x\|_\infty \le \|x\|_2,\ x \in \mathbb{R}^N$, sich nicht auf die jeweils induzierten Matrixnormen überträgt. Als ein weiteres Beispiel betrachte man die nichtsymmetrische Matrix $A \in \mathbb{R}^{2\times 2}$ definiert durch

$$A = \begin{pmatrix} 0 & 1 \\ 0 & 1 \end{pmatrix} \qquad \left(\Longrightarrow \quad A^\top A = \begin{pmatrix} 0 & 0 \\ 0 & 2 \end{pmatrix} \right).$$

Hier gilt $\|A\|_1 = 2$ und $r_\sigma(A) = 1 = \|A\|_\infty$ sowie $\|A\|_2 = \sqrt{2}$, so dass auf die Voraussetzung "$A = A^\top$" in Theorem 4.43 nicht verzichtet werden kann. △

Das folgende Theorem liefert einfache Abschätzungen für die Spektralnorm.

Theorem 4.45. *Für jede Matrix $A \in \mathbb{R}^{N\times N}$ gelten die beiden folgenden Abschätzungen,*

$$\|A\|_2 \le \big(\|A\|_\infty \|A\|_1\big)^{1/2}, \qquad \|A\|_2 \le \|A\|_F.$$

BEWEIS. Die erste Abschätzung erhält man als Korollar zu Theorem 4.43,

$$\begin{aligned} \|A\|_2 &= r_\sigma(A^\top A)^{1/2} = \|A^\top A\|_2^{1/2} \overset{(*)}{\le} \|A^\top A\|_\infty^{1/2} \\ &\le (\|A^\top\|_\infty \|A\|_\infty)^{1/2} \overset{(**)}{=} (\|A\|_1 \|A\|_\infty)^{1/2}, \end{aligned}$$

wobei $(*)$ aus Theorem 4.43 und $(**)$ aus Korollar 4.41 folgt.

Die zweite Abschätzung resultiert aus der cauchy-schwarzschen Ungleichung,

$$\begin{aligned} \|Ax\|_2 &= \Big(\sum_{j=1}^N \Big| \sum_{k=1}^N a_{jk} x_k \Big|^2 \Big)^{1/2} \le \Big(\sum_{j=1}^N \Big(\sum_{k=1}^N |a_{jk}|^2 \Big) \Big(\sum_{s=1}^N |x_s|^2 \Big) \Big)^{1/2} \\ &= \|A\|_F \|x\|_2 \qquad \text{für} \quad x \in \mathbb{R}^N. \end{aligned}$$

□

4.7.3 Die Konditionszahl einer Matrix

Bei Stabilitätsuntersuchungen für lineare Gleichungssysteme spielt der nachfolgende Begriff eine besondere Rolle.

Definition 4.46. Sei $A \in \mathbb{R}^{N\times N}$ eine reguläre Matrix und $\|\cdot\| : \mathbb{R}^{N\times N} \to \mathbb{R}_+$ eine Matrixnorm. Die Zahl

$$\operatorname{cond}(A) \;=\; \|A\|\,\|A^{-1}\|$$

wird als *Konditionszahl* der Matrix A bezeichnet.

Das folgende Theorem liefert eine alternative Darstellung der Konditionszahl, die unter anderem eine geometrische Deutung ermöglicht (siehe Bemerkung 4.48).

Theorem 4.47. *Sei $A \in \mathbb{R}^{N\times N}$ eine reguläre Matrix und $\|\cdot\| : \mathbb{R}^N \to \mathbb{R}_+$ eine Vektornorm. Für die induzierte Konditionszahl gilt dann*

$$\operatorname{cond}(A) \;=\; \Big(\max_{\|x\|=1} \|Ax\|\Big)\Big/\Big(\min_{\|x\|=1} \|Ax\|\Big). \tag{4.37}$$

BEWEIS. Die Darstellung (4.37) erhält man wie folgt,

$$\begin{aligned}\|A^{-1}\| \;&=\; \max_{0\neq y\in\mathbb{R}^N} \frac{\|A^{-1}y\|}{\|y\|} \;\overset{(*)}{=}\; \max_{0\neq x\in\mathbb{R}^N} \frac{\|x\|}{\|Ax\|} \;=\; \max_{x\in\mathbb{R}^N,\|x\|=1} \frac{1}{\|Ax\|}\\ &=\; \Big(\min_{x\in\mathbb{R}^N,\|x\|=1} \|Ax\|\Big)^{-1},\end{aligned}$$

□

wobei die Identität $(*)$ aus der Substitution $y = Ax$ resultiert.

Bemerkung 4.48. Die Konditionszahl $\operatorname{cond}(A)$ gibt also die Bandbreite an, um die sich die Vektorlänge bei Multiplikation mit der Matrix A ändern kann. Aus der Darstellung (4.37) ergibt sich zudem die Ungleichung $\operatorname{cond}(A) \geq 1$. △

4.7.4 Störungsresultate für Matrizen

Lemma 4.49. *Für die durch eine Vektornorm induzierte Matrixnorm $\|\cdot\| : \mathbb{R}^{N\times N} \to \mathbb{R}_+$ und jede Matrix $B \in \mathbb{R}^{N\times N}$ mit $\|B\| < 1$ ist die Matrix $I + B$ regulär und es gilt*

$$\|(I+B)^{-1}\| \;\leq\; \frac{1}{1-\|B\|}.$$

BEWEIS. Die umgekehrte Dreiecksungleichung liefert für $x \in \mathbb{R}^N$

$$\begin{aligned}\|(I+B)x\| \;&=\; \|x + Bx\| \;\geq\; \|x\| - \|Bx\|\\ &\geq\; \|x\| - \|B\|\|x\| \;=\; (1-\|B\|)\|x\|,\end{aligned}$$

was die Regularität der Matrix $I + B$ impliziert. Die Substitution $y = (I + B)x$ in der vorangegangenen Abschätzung liefert dann auch

$$\|y\| \quad \geq \quad (1 - \|B\|)\|(I + B)^{-1}y\|, \qquad y \in \mathbb{R}^N,$$

was den Nachweis von Lemma 4.49 komplettiert. □

Als eine Konsequenz aus Lemma 4.49 erhält man die Offenheit der Menge der regulären Matrizen und die Stetigkeit der Matrixinversion.

Korollar 4.50. *Sei $\|\cdot\| : \mathbb{R}^{N\times N} \to \mathbb{R}_+$ die durch eine Vektornorm induzierte Matrixnorm, und $A \in \mathbb{R}^{N\times N}$ sei eine reguläre Matrix. Für jede Matrix $\Delta A \in \mathbb{R}^{N\times N}$ mit $\|\Delta A\| < 1/\|A^{-1}\|$ ist die Matrix $A + \Delta A$ regulär, und*

$$\|(A + \Delta A)^{-1}\| \quad \leq \quad \frac{1}{\|A^{-1}\|^{-1} - \|\Delta A\|}.$$

$$\|(A + \Delta A)^{-1} - A^{-1}\| \quad \leq \quad c\|\Delta A\| \qquad \text{für} \quad \|\Delta A\| \leq \frac{1}{2\|A^{-1}\|}, \quad \text{mit} \quad c = 2\|A^{-1}\|^2.$$

BEWEIS. Wegen $\|A^{-1}\Delta A\| \leq \|A^{-1}\|\|\Delta A\| < 1$ ist nach Lemma 4.49 die Matrix $A + \Delta A = A(I + A^{-1}\Delta A)$ regulär, und mit der Darstellung $(A + \Delta A)^{-1} = (I + A^{-1}\Delta A)^{-1}A^{-1}$ erhält man zudem

$$\|(A + \Delta A)^{-1}\| \quad \leq \quad \frac{\|A^{-1}\|}{1 - \|A^{-1}\Delta A\|} \quad \leq \quad \frac{\|A^{-1}\|}{1 - \|A^{-1}\|\|\Delta A\|}.$$

Die zweite Abschätzung des Korollars folgt unmittelbar aus der ersten Abschätzung zusammen mit der Darstellung

$$\begin{aligned} (A + \Delta A)^{-1} - A^{-1} \quad &= \quad -(A + \Delta A)^{-1}\Delta A A^{-1}, \\ \|\text{———} \ll \text{———}\| \quad &\leq \quad \frac{\|A^{-1}\|}{\|A^{-1}\|^{-1} - \|\Delta A\|}\|\Delta A\|. \end{aligned}$$

□

Korollar 4.51. *Sei $\|\cdot\| : \mathbb{R}^{N\times N} \to \mathbb{R}_+$ die durch eine Vektornorm induzierte Matrixnorm, und $A \in \mathbb{R}^{N\times N}$ sei eine reguläre Matrix.*

(a) Für jede Matrix $B \in \mathbb{R}^{N\times N}$ gilt: B ist singulär $\implies \frac{1}{\|A^{-1}\|} \leq \|A - B\|$;

(b) Es gilt

$$\frac{1}{\text{cond}(A)} \quad \leq \quad \min\left\{ \frac{\|A - B\|}{\|A\|} \quad : \quad B \in \mathbb{R}^{N\times N} \text{ ist singulär} \right\}. \tag{4.38}$$

BEWEIS. Aussage (a) ergibt sich durch Negation der ersten Aussage in Korollar 4.50, und Division in (a) durch $\|A\|$ liefert Aussage (b). □

Bemerkung 4.52. 1. Wegen der Stetigkeit der Matrixnorm (siehe Korollar 4.31) sowie der Abgeschlossenheit der Menge der singulären Matrizen aus $\mathbb{R}^{N\times N}$ (siehe Korollar 4.50) wird das Minimum in (4.38) tatsächlich auch angenommen.
2. Durch die Aussage (b) in Korollar 4.51 wird klar, dass $1/\text{cond}(A)$ eine untere Schranke für den relativen Abstand der Matrix A zur Menge der singulären Matrizen darstellt. △

4.7.5 Fehlerabschätzungen für fehlerbehaftete Gleichungssysteme

Es können nun die zentralen Theoreme dieses Abschnitts 4.7 formuliert werden.

Theorem 4.53 (Fehlerbehaftete rechte Seiten). *Mit $\|\cdot\|$ seien gleichzeitig sowohl eine Vektornorm auf $\mathbb{R}^N$ als auch die induzierte Matrixnorm auf $\mathbb{R}^{N\times N}$ bezeichnet. Es sei $A \in \mathbb{R}^{N\times N}$ eine reguläre Matrix, und $b,\ x \in \mathbb{R}^N$ und $\Delta b,\ \Delta x \in \mathbb{R}^N$ seien Vektoren mit*

$$Ax \;=\; b, \qquad A(x \;+\; \Delta x) \;=\; b \;+\; \Delta b. \tag{4.39}$$

Dann gelten für den absoluten beziehungsweise den relativen Fehler die folgenden Abschätzungen,

$$\|\Delta x\| \;\le\; \|A^{-1}\|\,\|\Delta b\|, \qquad\qquad \frac{\|\Delta x\|}{\|x\|} \;\le\; \operatorname{cond}(A)\,\frac{\|\Delta b\|}{\|b\|}. \tag{4.40}$$

Beweis. Aus (4.39) folgt unmittelbar $A\,\Delta x = \Delta b$ beziehungsweise $\Delta x = A^{-1}\Delta b$, woraus die erste Abschätzung in (4.40) resultiert. Aus dieser Abschätzung wiederum ergibt sich die zweite Abschätzung in (4.40),

$$\frac{\|\Delta x\|}{\|x\|} \overset{Ax=b}{\le} \|A^{-1}\|\,\frac{\|\Delta b\|}{\|b\|}\,\frac{\|Ax\|}{\|x\|} \;\le\; \operatorname{cond}(A)\,\frac{\|\Delta b\|}{\|b\|}.$$

□

Bemerkung 4.54. Fällt also die Konditionszahl einer Matrix A groß aus ($\operatorname{cond}(A) \gg 1$), so tut dies auch in (4.40) die obere Schranke für den relativen Fehler in der Lösung der fehlerbehafteten Version des linearen Gleichungssystems $Ax = b$. In einem solchen Fall spricht man von *schlecht konditionierten Gleichungssystemen* $Ax = b$. △

Vergleichbares wie in Theorem 4.53 gilt auch im Fall fehlerbehafteter Matrizen:

Theorem 4.55 (Fehlereinflüsse in der rechten Seite und der Matrix). *Mit $\|\cdot\|$ seien gleichzeitig sowohl eine Vektornorm als auch die induzierte Matrixnorm bezeichnet, $A \in \mathbb{R}^{N\times N}$ sei eine reguläre Matrix, und $\Delta A \in \mathbb{R}^{N\times N}$ sei eine Matrix mit $\|\Delta A\| < \|A^{-1}\|^{-1}$.*

Dann gilt für beliebige Vektoren $b,\ x \in \mathbb{R}^N$ und $\Delta b,\ \Delta x \in \mathbb{R}^N$ mit

$$Ax \;=\; b, \qquad (A \;+\; \Delta A)(x \;+\; \Delta x) \;=\; b \;+\; \Delta b, \tag{4.41}$$

die Abschätzung

$$\frac{\|\Delta x\|}{\|x\|} \;\le\; C\Big(\frac{\|\Delta A\|}{\|A\|} \;+\; \frac{\|\Delta b\|}{\|b\|}\Big) \quad \textit{mit} \quad C \;=\; \frac{1}{\frac{1}{\operatorname{cond}(A)} - \frac{\|\Delta A\|}{\|A\|}}.$$

BEWEIS. Aus (4.41) folgt unmittelbar

$$(A + \Delta A)\Delta x = \Delta b - \Delta A x,$$

und Korollar 4.50 liefert nun (neben der Regularität der Matrix $A + \Delta A$) die Abschätzung

$$\|\Delta x\| \leq \frac{1}{\|A^{-1}\|^{-1} - \|\Delta A\|} \big(\|\Delta b\| + \|\Delta A\|\|x\|\big).$$

Anschließende Division durch $\|x\|$ liefert wegen $\|b\| \leq \|A\|\|x\|$ die Aussage des Theorems. □

4.8 Orthogonalisierungsverfahren

In diesem Abschnitt soll für eine gegebene Matrix $A \in \mathbb{R}^{M\times N}$, $1 \leq N \leq M$, eine Faktorisierung der Form

$$A = QS \tag{4.42}$$

bestimmt werden mit einer orthogonalen Matrix Q,

$$Q \in \mathbb{R}^{M\times M}, \qquad Q^{-1} = Q^{\top}, \tag{4.43}$$

und S ist eine verallgemeinerte obere Dreiecksmatrix,

$$S = \left(\begin{array}{c} R \\ \hline \mathbf{0} \end{array}\right) \in \mathbb{R}^{M\times N}, \quad R = \left(\begin{array}{c} \\ \\ \end{array}\right) \in \mathbb{R}^{N\times N}, \quad \mathbf{0} = (0) \in \mathbb{R}^{(M-N)\times N}. \tag{4.44}$$

Eine solche Faktorisierung (4.42) ermöglicht beispielsweise die stabile Lösung von regulären aber eventuell schlecht konditionierten linearen Gleichungssystemen $Ax = b$ (für $M = N$); mehr hierzu in Abschnitt 4.8.4. Auch die stabile Lösung von Ausgleichsproblemen $\|Ax - b\|_2 \to \min$, $x \in \mathbb{R}^N$, ist mit einer solchen Faktorisierung möglich. Details hierzu finden Sie in Abschnitt 4.8.5.

4.8.1 Elementare Eigenschaften orthogonaler Matrizen

Vorbereitend werden einige Eigenschaften orthogonaler Matrizen vorgestellt.

Lemma 4.56. *Sei $Q \in \mathbb{R}^{M\times M}$ eine orthogonale Matrix. Dann ist auch $Q^{\top}$ eine orthogonale Matrix, und es gilt*

$$\|Qx\|_2 = \|x\|_2 = \|Q^{\top}x\|_2, \qquad x \in \mathbb{R}^M,$$

das heißt, Q und $Q^{\top}$ sind isometrisch *bezüglich der euklidischen Vektornorm.*

BEWEIS. Es gilt $(Q^\top)^{-1} = (Q^{-1})^{-1} = Q = (Q^\top)^\top$, somit ist auch $Q^\top$ eine orthogonale Matrix. Des Weiteren besitzt die Matrix Q die Isometrieeigenschaft:

$$\|Qx\|_2 \;=\; \big(x^\top \underbrace{Q^\top Q}_{=I}\, x\big)^{1/2} \;=\; (x^\top x)^{1/2} \;=\; \|x\|_2.$$

Diese beiden Aussagen ergeben dann die Identität $\|Q^\top x\|_2 = \|x\|_2$. □

Bezogen auf die euklidische Vektornorm $\|\cdot\|_2$ ändert sich die Konditionszahl einer quadratischen regulären Matrix nicht bei Multiplikation mit einer orthogonalen Matrix:

Korollar 4.57. *Sei $A \in \mathbb{R}^{N\times N}$ regulär, und $Q \in \mathbb{R}^{N\times N}$ sei eine orthogonale Matrix. Dann gilt*

$$\operatorname{cond}_2(QA) \;=\; \operatorname{cond}_2(A).$$

BEWEIS. Nach Lemma 4.56 gilt $\|QAx\|_2 = \|Ax\|_2$ für $x \in \mathbb{R}^N$, was unmittelbar auf $\|A\|_2 = \|QA\|_2$ führt. Weiter gilt nach Lemma 4.56 auch

$$\begin{aligned}\|A^{-1}Q^\top\|_2 \;&=\; \max_{0\neq x\in\mathbb{R}^N} \frac{\|A^{-1}Q^\top x\|_2}{\|x\|_2} \;=\; \max_{0\neq x\in\mathbb{R}^N} \frac{\|A^{-1}Q^\top x\|_2}{\|Q^\top x\|_2}\\ &\overset{(*)}{=}\; \max_{0\neq y\in\mathbb{R}^N} \frac{\|A^{-1}y\|_2}{\|y\|_2} \;=\; \|A^{-1}\|_2,\end{aligned}$$

wobei $(*)$ mit der Substitution $y = Q^\top x$ folgt. Insgesamt erhält man daraus

$$\operatorname{cond}_2(QA) \;=\; \|QA\|_2\,\|A^{-1}\underbrace{Q^{-1}}_{=Q^\top}\|_2 \;=\; \|A\|_2\,\|A^{-1}\|_2 \;=\; \operatorname{cond}_2(A).$$

□

Das folgende Resultat wird in Abschnitt 4.8.3 über die Gewinnung einer Faktorisierung $A = QS$ mittels spezieller und hintereinander auszuführender Transformationen benötigt.

Lemma 4.58. *Für orthogonale Matrizen $Q_1, Q_2 \in \mathbb{R}^{M\times M}$ ist auch Q_1Q_2 eine orthogonale Matrix.*

BEWEIS. Es gilt $(Q_1Q_2)^{-1} \;=\; Q_2^{-1}Q_1^{-1} \;=\; Q_2^\top Q_1^\top \;=\; (Q_1Q_2)^\top$. □

4.8.2 Die Faktorisierung $A = QR$ mittels Gram-Schmidt-Orthogonalisierung

Für eine quadratische reguläre Matrix $A \in \mathbb{R}^{N\times N}$ nimmt der Ansatz (4.42)–(4.44) die folgende Form an,

$$A \;=\; QR \tag{4.45}$$

mit einer orthogonalen Matrix $Q \in \mathbb{R}^{N\times N}$ und der oberen Dreiecksmatrix $R \in \mathbb{R}^{N\times N}$. Mit den Notationen

$$A = \left(\begin{array}{c|c|c} a_1 & \dots & a_N \end{array}\right), \quad Q = \left(\begin{array}{c|c|c} q_1 & \dots & q_N \end{array}\right), \quad R = \begin{pmatrix} r_{11} & \cdots & r_{1N} \\ & \ddots & \vdots \\ & & r_{NN} \end{pmatrix} \tag{4.46}$$

(mit Vektoren $a_k,\ q_k \in \mathbb{R}^N$) führt der Ansatz (4.45) auf die folgenden Forderungen,

$$a_k = \sum_{j=1}^{k} r_{jk} q_j, \qquad k = 1, 2, \dots, N, \tag{4.47}$$

$$q_1, \dots, q_N \in \mathbb{R}^N \qquad \text{paarweise orthonormal.} \tag{4.48}$$

Im Folgenden wird beschrieben, wie man mittels einer *Gram-Schmidt-Orthogonalisierung* eine solche Faktorisierung (4.47)–(4.48) gewinnt.

Algorithmus 4.59. Für eine gegebene reguläre Matrix $A \in \mathbb{R}^{N\times N}$ geht man bei der Gram-Schmidt-Orthogonalisierung schrittweise für $k = 1, 2, \dots, N$ so vor: ausgehend von bereits gewonnenen orthonormalen Vektoren $q_1, q_2, \dots, q_{k-1} \in \mathbb{R}^N$ mit

$$\operatorname{span}\{a_1, \dots, a_{k-1}\} = \operatorname{span}\{q_1, \dots, q_{k-1}\} =: \mathcal{M}_{k-1},$$

bestimmt man in Schritt $k \geq 1$ das Lot von a_k auf den linearen Unterraum $\mathcal{M}_{k-1} \subset \mathbb{R}^N$,

$$\widehat{q}_k := a_k - \sum_{j=1}^{k-1} (a_k^\top q_j) q_j, \tag{4.49}$$

und nach der Normierung

$$q_k := \frac{\widehat{q}_k}{\|\widehat{q}_k\|_2} \tag{4.50}$$

sind die Vektoren $q_1, \dots, q_k \in \mathbb{R}^N$ paarweise orthonormal mit

$$\operatorname{span}\{a_1, \dots, a_k\} = \operatorname{span}\{q_1, \dots, q_k\}. \qquad \Delta$$

Der Gleichung (4.49) entnimmt man unmittelbar die Darstellung

$$a_k = \underbrace{\|\widehat{q}_k\|_2}_{=:\, r_{kk}} q_k + \sum_{j=1}^{k-1} \underbrace{(a_k^\top q_j)}_{=:\, r_{jk}} q_j, \qquad k = 1, 2, \dots, N, \tag{4.51}$$

und mit den Notationen aus (4.50) beziehungsweise (4.51) erhält man nach Abschluss der Gram-Schmidt-Orthogonalisierung die gesuchte Faktorisierung (4.47)–(4.48)[7].

Der in Algorithmus 4.59 beschriebene Orthogonalisierungsprozess ist jedoch unter Umständen nicht gutartig (wenn etwa $\|\widehat{q}_k\|_2$ klein ausfällt), so dass zur Bestimmung einer QR-Faktorisierung andere Methoden vorzuziehen sind (mehr hierzu im folgenden Abschnitt 4.8.3).

[7] beziehungsweise in Matrixschreibweise und mit der Notation aus (4.46) die Faktorisierung $A = QR$

4.8.3 Die Faktorisierung $A = QS$ mittels Householder-Transformationen

Gegenstand dieses Abschnitts 4.8.3 ist die Bestimmung einer Faktorisierung der Form $A = QS$ entsprechend (4.43)–(4.44) mittels Householder-Transformationen, wobei wieder der allgemeine Fall $A \in \mathbb{R}^{M\times N}$ mit $M \geq N \geq 1$ zugelassen wird. In dem folgenden Unterabschnitt werden die nötigen Vorbereitungen getroffen.

Vorüberlegungen

Lemma 4.60. *Für eine Matrix*

$$\mathcal{H} \;=\; I \,-\, 2ww^{\top} \;\in\; \mathbb{R}^{s\times s} \quad \textit{mit} \qquad w \in \mathbb{R}^{s}, \quad w^{\top}w \;=\; 1 \tag{4.52}$$

mit $s \geq 1$ gilt Folgendes:

$$\mathcal{H}^{\top} \;=\; \mathcal{H} \qquad (\mathcal{H} \textit{ ist symmetrisch}) \tag{4.53}$$

$$\mathcal{H}^{2} \;=\; I \qquad \big(\mathcal{H} \textit{ ist} \text{ involutorisch}\big) \tag{4.54}$$

$$\mathcal{H}^{\top}\mathcal{H} \;=\; I \qquad (\mathcal{H} \textit{ ist orthogonal}). \tag{4.55}$$

BEWEIS. Die Identitäten (4.53)–(4.54) ergeben sich wie folgt,

$$\mathcal{H}^{\top} \;=\; I \,-\, 2(ww^{\top})^{\top} \;=\; I \,-\, 2ww^{\top} \;=\; \mathcal{H},$$

$$\mathcal{H}^{2} \;=\; (I \,-\, 2ww^{\top})(I \,-\, 2ww^{\top}) \;=\; I \,-\, 2ww^{\top} - 2ww^{\top} + 4w\,\overbrace{(w^{\top}w)}^{=\,1}\,w^{\top} \;=\; I,$$

und die Identität (4.55) folgt unmittelbar aus (4.53)–(4.54). □

Definition 4.61. Eine Abbildung

$$\mathbb{R}^{s} \;\to\; \mathbb{R}^{s}, \qquad x \;\mapsto\; \mathcal{H}x$$

mit einer Matrix $\mathcal{H} \in \mathbb{R}^{s\times s}$ der Form (4.52) mit $s \geq 1$ bezeichnet man als *Householder-Transformation*.

Eine Householder-Transformation mit einer Matrix $\mathcal{H} \in \mathbb{R}^{s\times s}$ der Form (4.52) bewirkt aufgrund der Identität $x - 2(w^{\top}x)w \;=\; x - (w^{\top}x)w - (w^{\top}x)w$ eine Spiegelung von x an der Hyperebene $\{z \in \mathbb{R}^{s} : \; z^{\top}w = 0\}$. Für den Fall $s = 2$ ist dies in Bild 4.2 veranschaulicht.

Bei der sukzessiven Triangulierung einer Matrix mittels Householder-Transformationen (siehe unten) ist in jedem Teilschritt (für unterschiedliche Werte von s) ein Vektor $w \in \mathbb{R}^{s}$, $\|w\|_2 = 1$, so zu bestimmen, dass die zugehörige Householder-Transformation einen gegebenen Vektor $x \in \mathbb{R}^{s}$ in ein Vielfaches des ersten Einheitsvektors $\mathbf{e}_1 = (1, 0, \ldots, 0)^{\top} \in \mathbb{R}^{s}$ abbildet. Das folgende Lemma gibt einen solchen Vektor $w \in \mathbb{R}^{s}$ an.

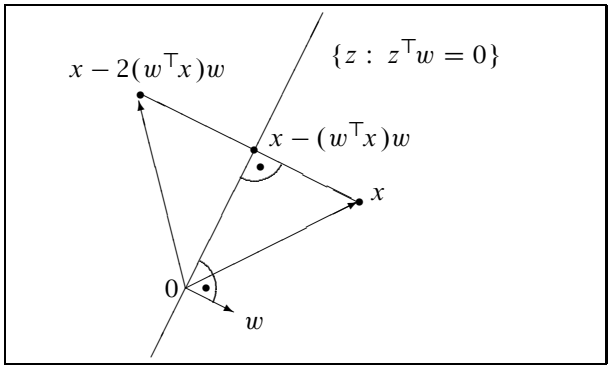


Bild 4.2: Darstellung der Householder-Spiegelung für den zweidimensionalen Fall

Lemma 4.62. *Gegeben sei ein Vektor* $0 \neq x \in \mathbb{R}^s$ *mit* $x \notin \operatorname{span}\{\mathbf{e}_1\}$. *Für*

$$w = \frac{x + \sigma \mathbf{e}_1}{\|x + \sigma \mathbf{e}_1\|_2} \quad \textit{mit} \quad \sigma = \pm \|x\|_2, \tag{4.56}$$

gilt

$$\|w\|_2 = 1, \tag{4.57}$$

$$(I - 2ww^\top)x = -\sigma \mathbf{e}_1. \tag{4.58}$$

BEWEIS. Wegen $x \notin \operatorname{span}\{\mathbf{e}_1\}$ verschwindet der Nenner in (4.56) nicht, so dass $w \in \mathbb{R}^s$ wohldefiniert ist und offensichtlich (4.57) gilt. Für den Nachweis der Identität (4.58) berechnet man

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

Daraus erhält man

$$2w^\top x = \frac{2(x + \sigma \mathbf{e}_1)^\top x}{\|x + \sigma \mathbf{e}_1\|_2} = \|x + \sigma \mathbf{e}_1\|_2,$$

was zusammen mit (4.56) die Darstellung

$$2w\, w^\top x = x + \sigma \mathbf{e}_1$$

liefert. Dies stimmt mit der Identität (4.58) überein. □

Bemerkung 4.63. Der Vektor $w \in \mathbb{R}^s$ in (4.56) entsteht also aus $x \in \mathbb{R}^s$ durch eine Modifikation des ersten Eintrags von x sowie einer anschließenden Normierung. Zur Vermeidung von Stellenauslöschungen wird in (4.56) $\sigma = \operatorname{sgn}(x_1)\|x\|_2$ gewählt. Hier bezeichnet für eine Zahl $y \in \mathbb{R}$

$$\operatorname{sgn}(y) = \begin{cases} 1, & \text{falls } y \geq 0, \\ -1, & \text{sonst.} \end{cases}$$

△

Triangulierung mittels Householder-Transformationen

Im Folgenden wird beschrieben, wie man ausgehend von der Matrix $A = A^{(1)} \in \mathbb{R}^{M\times N}$ sukzessive Matrizen der Form

$$A^{(k)} = \begin{pmatrix} a_{11}^{(k)} & a_{12}^{(k)} & \cdots & \cdots & \cdots & a_{1N}^{(k)} \\ & \ddots & & & & \vdots \\ & & a_{k-1,k-1}^{(k)} & \cdots & \cdots & a_{k-1,N}^{(k)} \\ & & & a_{kk}^{(k)} & \cdots & a_{kN}^{(k)} \\ & & & \vdots & & \vdots \\ & & & a_{Mk}^{(k)} & \cdots & a_{MN}^{(k)} \end{pmatrix} \in \mathbb{R}^{M\times N}, \quad k = 2, 3, \ldots, N_*, \tag{4.59}$$

bestimmt, so dass dann schließlich $A^{(N_*)} = S$ gilt mit einer verallgemeinerten oberen Dreiecksmatrix $S \in \mathbb{R}^{M\times N}$ von der Form (4.44). Hierbei wird die Bezeichnung

$$N_* = \begin{cases} N, & \text{falls } M = N, \\ N+1, & \text{falls } M > N, \end{cases}$$

verwendet. Die Matrizen in (4.59) werden dabei für $k = 1, 2, \ldots, N_* - 1$ sukzessive durch Transformationen der Form

$$A^{(k+1)} = \widehat{\mathcal{H}}_k A^{(k)}, \ \widehat{\mathcal{H}}_k = \left(\begin{array}{c|c} I_{k-1} & \mathbf{0} \\ \hline \mathbf{0} & \mathcal{H}_k \end{array}\right), \quad \begin{array}{l} \mathcal{H}_k = I_{M-(k-1)} - 2w_k w_k^\top, \\ w_k \in \mathbb{R}^{M-(k-1)}, \ \|w_k\|_2 = 1, \end{array}$$

gewonnen, wobei wieder $I_s \in \mathbb{R}^{s\times s}$ die Einheitsmatrix bezeichnet, und der Vektor $w_k \in \mathbb{R}^{M-(k-1)}$ ist so zu wählen, dass

$$\mathcal{H}_k \begin{pmatrix} a_{kk}^{(k)} \\ \vdots \\ a_{Mk}^{(k)} \end{pmatrix} = \begin{pmatrix} -\sigma_k \\ 0 \\ \vdots \\ 0 \end{pmatrix}$$

gilt; die genaue Form von $w_k \in \mathbb{R}^{M-k+1}$ und $\sigma_k \in \mathbb{R}$ entnimmt man Lemma 4.62. Nach Lemma 4.60 sind die Matrizen $\widehat{\mathcal{H}}_1, \ldots, \widehat{\mathcal{H}}_{N_*-1}$ orthogonal und symmetrisch, so dass man mit

$$S = \widehat{\mathcal{H}}_{N_*-1}\widehat{\mathcal{H}}_{N_*-2}\cdots\widehat{\mathcal{H}}_1 A, \qquad Q = \widehat{\mathcal{H}}_1\widehat{\mathcal{H}}_2\cdots\widehat{\mathcal{H}}_{N_*-1},$$

die gewünschte Faktorisierung $A = QS$ erhält, wobei Q nach Lemma 4.58 tatsächlich eine Orthogonalmatrix ist.

Bemerkung 4.64. (a) Praktisch geht man für $k = 1, 2, \ldots, N_* - 1$, so vor, dass man das Diagonalelement $a_{kk}^{(k+1)}$ gesondert abspeichert und in der Matrix $A^{(k+1)}$ den frei werdenden Platz in der k-ten Spalte unterhalb der Diagonalen dazu verwendet, den Vektor w_k abzuspeichern.

(b) Die nötigen Matrixmultiplikationen der Form

$$(I - 2ww^\top)B = B - wv^\top, \qquad v^\top := 2w^\top B$$

führt man so aus, dass zunächst der Vektor v berechnet und anschließend die Matrix B modifiziert ("aufdatiert") wird. △

4.8.4 Anwendung 1: Stabile Lösung schlecht konditionierter Gleichungssysteme $Ax = b$

Für eine reguläre aber eventuell schlecht konditionierte Matrix $A \in \mathbb{R}^{N\times N}$ ermöglicht eine Faktorisierung der Form $A = QR$ mit einer orthogonalen Matrix $Q \in \mathbb{R}^{N\times N}$ und einer oberen Dreiecksmatrix $R \in \mathbb{R}^{N\times N}$ eine stabile Lösung zugehöriger linearer Gleichungssysteme. Dies liegt daran, dass für einen gegebenen Vektor $b \in \mathbb{R}^N$ das Gleichungssystem $Ax = b$ äquivalent ist zu dem gestaffelten Gleichungssystem

$$Rx = Q^\top b,$$

wobei die Matrix R bezüglich der Norm $\|\cdot\|_2$ keine schlechtere Konditionszahl als die Matrix A aufweist und die Norm des Vektors $Q^\top b$ nicht größer als die des Vektors b ist:[8]

$$\begin{aligned} \mathrm{cond}_2(R) &= \mathrm{cond}_2(Q^\top A) = \mathrm{cond}_2(A), \\ \|Q^\top b\|_2 &= \|b\|_2. \end{aligned}$$

4.8.5 Anwendung 2: Lineare Ausgleichsrechnung

Lineare (unrestringierte) Ausgleichsprobleme sind von der Form

$$\|Ax - b\|_2 \to \min \qquad \text{für} \quad x \in \mathbb{R}^N, \tag{4.60}$$

mit gegebener Matrix $A \in \mathbb{R}^{M\times N}$ und gegebenem Vektor $b \in \mathbb{R}^M$. Zunächst soll ein konkretes lineares Ausgleichsproblem vorgestellt werden.

Beispiel 4.65. Im Folgenden ist diejenige Gerade in $\mathbb{R}^2$ gesucht, die im quadratischen Mittel den geringsten vertikalen Abstand zu vorgegebenen Stützpunkten $(y_j, f_j) \in \mathbb{R}^2$, $j = 1, 2, \ldots, M$ besitzt, mit paarweise verschiedenen reellen Zahlen $y_1, y_2, \ldots, y_M$; diese bezeichnet man als *Ausgleichsgerade*. Wegen der allgemeinen Darstellung

[8] siehe Lemma 4.56 und Korollar 4.57 für die Einzelheiten

$\{cy + d : y \in \mathbb{R}\}$ mit gewissen Koeffizienten $c, d \in \mathbb{R}$ für Geraden in $\mathbb{R}^2$ lautet das zu lösende Minimierungsproblem folglich

$$\sum_{j=1}^{M} (cy_j + d - f_j)^2 \to \min, \qquad c,\ d \in \mathbb{R}, \tag{4.61}$$

das man in der Form (4.60) schreiben kann,

$$\left\| \begin{pmatrix} y_1 & 1 \\ \vdots & \vdots \\ y_M & 1 \end{pmatrix} \begin{pmatrix} c \\ d \end{pmatrix} - \begin{pmatrix} f_1 \\ \vdots \\ f_M \end{pmatrix} \right\|_2 \to \min \qquad \text{für } c,\ d \in \mathbb{R}.$$

Von allgemeinerer Form ist das Problem, Koeffizienten $a_0, \ldots, a_{N-1} \in \mathbb{R}$ so zu bestimmen, dass für das Polynom $p(y) = \sum_{k=0}^{N-1} a_k y^k$ der Ausdruck

$$\sum_{j=1}^{M} (p(y_j) - f_j)^2 \tag{4.62}$$

minimal wird (mit $M \geq N$). Die zugehörige Lösung bezeichnet man als *Ausgleichspolynom*. Dieses Problem kann ebenfalls in der Form (4.60) geschrieben werden:

$$\left\| \begin{pmatrix} y_1^0 & y_1^1 & \ldots & y_1^{N-1} \\ \vdots & \vdots & \vdots & \vdots \\ y_M^0 & y_M^1 & \ldots & y_M^{N-1} \end{pmatrix} \begin{pmatrix} a_0 \\ \vdots \\ a_{N-1} \end{pmatrix} - \begin{pmatrix} f_1 \\ \vdots \\ f_M \end{pmatrix} \right\|_2 \to \min \quad \text{für } a_0, a_1, \ldots, a_{N-1} \in \mathbb{R}.$$

Für einen kleinen Grad $N-1$ und eine große Stützpunkteanzahl M tritt bei dem Ausgleichspolynom üblicherweise nicht ein solches oszillierendes Verhalten auf, wie man es von dem interpolierenden Polynom (vom Grad $\leq M-1$) zu erwarten hat. △

Mit dem nachfolgenden Theorem wird klar, wie mittels Faktorisierungen der Form $A = QS$ lineare Ausgleichsprobleme effizient gelöst werden können.

Theorem 4.66. *Für die Matrix $A \in \mathbb{R}^{M\times N}$, $1 \leq N \leq M$, mit maximalem Rang N sei eine Faktorisierung $A = QS$ gegeben mit einer orthogonalen Matrix $Q \in \mathbb{R}^{M\times M}$ und der verallgemeinerten oberen Dreiecksmatrix $S \in \mathbb{R}^{M\times N}$ entsprechend (4.44),*

$$S = \begin{pmatrix} R \\ \hline \mathbf{0} \end{pmatrix} \in \mathbb{R}^{M\times N}, \quad R = \begin{pmatrix} \ddots \\ \ddots \end{pmatrix} \in \mathbb{R}^{N\times N}, \quad \mathbf{0} = (0) \in \mathbb{R}^{(M-N)\times N}.$$

Zu gegebenem Vektor $b \in \mathbb{R}^M$ sei $Q^\top b$ wie folgt partitioniert,

$$Q^\top b =: \begin{pmatrix} y_1 \\ \hline y_2 \end{pmatrix} \in \mathbb{R}^M, \qquad y_1 \in \mathbb{R}^N, \qquad y_2 \in \mathbb{R}^{M-N}.$$

Dann ist für einen Vektor $x_* \in \mathbb{R}^N$ Folgendes äquivalent: es löst x_* das lineare Ausgleichsproblem

$$\| Ax - b \|_2 \to \min \qquad \textit{für } x \in \mathbb{R}^N,$$

genau dann, wenn $Rx_* = y_1$ erfüllt ist.

BEWEIS. Für einen beliebigen Vektor $x \in \mathbb{R}^N$ gilt

$$\begin{aligned} \|Ax-b\|_2^2 &= \|QSx-QQ^\top b\|_2^2 = \|Sx-Q^\top b\|_2^2 \\ &= \left\| \begin{pmatrix} R \\ \mathbf{0} \end{pmatrix} x - \begin{pmatrix} y_1 \\ y_2 \end{pmatrix} \right\|_2^2 = \|Rx-y_1\|_2^2 + \|y_2\|_2^2, \end{aligned}$$

woraus die Aussage des Theorems folgt:

$$\|Ax-b\|_2 \geq \|y_2\|_2; \qquad \|Ax-b\|_2 = \|y_2\|_2 \iff Rx = y_1. \qquad \square$$

Weitere Themen und Literaturhinweise

Der Gauß-Algorithmus zur Lösung linearer Gleichungssysteme lässt sich auch mit der (numerisch allerdings aufwändigen) Totalpivotsuche durchführen (Aufgabe 4.6). Mehr Einzelheiten zu der in Abschnitt 4.6 behandelten LR-Faktorisierung für Bandmatrizen werden beispielsweise in Schwarz/Klöckner [94], Weller [110] und Werner [111] vorgestellt. Untersuchungen zu den Auswirkungen von Störungen symmetrischer positiv definiter Matrizen auf ihre Cholesky-Faktorisierung findet man in Higham [55]. Eine QR-Faktorisierung für Bandmatrizen wird in Oevel [78] vorgestellt. Bei der Analyse schlecht konditionierter linearer Gleichungssysteme lässt sich die Singulärwertzerlegung einer Matrix verwenden (Aufgabe 4.16). Weitere Einzelheiten zu diesem Thema werden beispielsweise in Baumeister [3], Engl/Hanke/Neubauer [25], Golub/Van Loan [35], Hämmerlin/Hoffmann [48], Horn/Johnson [58], Kress [63], Louis [66] und in Rieder [86] behandelt. Zur stabilen Lösung schlecht konditionierter linearer Gleichungssysteme bietet sich die Verwendung von *Regularisierungsverfahren* an ([3], [25], [48], [63], [66], [86], Groetsch [42] und Hofmann [57]). Auch über *Matrixäquilibrierungen* lässt sich eine Reduktion der Konditionszahl erzielen (Aufgabe 4.18 und Schaback/Wendland [92]). Erwähnenswert ist auch der *Algorithmus von Strassen*, mit dem sich der numerische Aufwand bei der Multiplikation zweier $N \times N$-Matrizen (von normalerweise $\mathcal{O}(N^3)$ arithmetischen Operationen) auf $\mathcal{O}(N^{\log_2 7}) \approx \mathcal{O}(N^{2.807})$ arithmetische Operationen reduzieren lässt (siehe Strassen [100] beziehungsweise [48], [55] und Überhuber [106]). Mittels verfeinerter Techniken kann man den Aufwand weiter reduzieren; der aktuelle Stand ist $\mathcal{O}(N^{2.38})$ arithmetische Operationen (Pan [80]). Speziell auf *Parallel- und Vektorrechner* zugeschnittene Verfahren finden Sie in Golub/Ortega [37], Schwandt [93] und in [92] und [94].

Übungsaufgaben

Aufgabe 4.1. Man löse das lineare Gleichungssystem

$$\begin{pmatrix} 10^{-4} & 1 \\ 1 & 1 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \begin{pmatrix} 1 \\ 2 \end{pmatrix}$$

einmal mit dem Gauß-Algorithmus ohne Pivotsuche und einmal mit dem Gauß-Algorithmus inklusive Pivotsuche. Dabei verwende man jeweils eine dreistellige dezimale Gleitpunktarithmetik. (Hierbei ist nach jeder Operation das Zwischenergebnis auf drei gültige Dezimalstellen zu runden.)

Aufgabe 4.2. Zur Lösung eines linearen Gleichungssystems $Ax = b$ mit einer Tridiagonalmatrix

$$A = \begin{pmatrix} a_{11} & a_{12} & & & \\ a_{21} & \ddots & \ddots & & \\ & \ddots & \ddots & \ddots & \\ & & \ddots & \ddots & a_{N-1,N} \\ & & & a_{N,N-1} & a_{NN} \end{pmatrix} \in \mathbb{R}^{N\times N}$$

(es gilt $a_{jk} = 0$ für $k \le j-2$ oder $k \ge j+2$) vereinfache man den Gauß-Algorithmus in geeigneter Weise und gebe die zugehörige Anzahl der arithmetischen Operationen an.

Aufgabe 4.3. Es sei $A = (a_{jk}) \in \mathbb{R}^{N\times N}$ eine Bandmatrix von der Form (4.25) auf Seite 76. Zur Lösung von linearen Gleichungssystemen $Ax = b$ mit einer solchen Bandmatrix A gebe man einen modifizierten Gauß-Algorithmus an, der mit höchstens $p(3+2q)(N-1)$ arithmetischen Operationen auskommt.

Aufgabe 4.4. Zur Lösung eines linearen Gleichungssystems $Ax = b$ mit einer Matrix $A \in \mathbb{R}^{N\times N}$ wird der Gauß-Algorithmus betrachtet.

(a) Man zeige: ist die Matrix A symmetrisch, so sind auch die Matrizen $B^{(1)}, B^{(2)}, \ldots, B^{(N)}$ aus (4.4) auf Seite 61 allesamt symmetrisch.

(b) Man zeige weiter: ist die Matrix A symmetrisch und positiv definit, so sind auch die Matrizen $B^{(1)}, B^{(2)}, \ldots, B^{(N)}$ aus (4.4) alle symmetrisch und positiv definit und der Gauß-Algorithmus ist durchführbar.

(c) Man gebe einen auf symmetrische Matrizen zugeschnittenen Gauß-Algorithmus an und berechne die dabei anfallende Zahl der arithmetischen Operationen.

Aufgabe 4.5. Die Matrix $A = (a_{jk}) \in \mathbb{R}^{N\times N}$ sei *diagonaldominant*, das heißt,

$$|a_{jj}| \ge \sum_{\substack{k=1 \\ k\neq j}}^{N} |a_{jk}| \qquad \text{für} \quad j = 1, 2, \ldots, N,$$

und außerdem sei die Matrix A regulär. Man weise nach, dass der Gauß-Algorithmus ohne Pivotwahl durchführbar ist.

Aufgabe 4.6. Sei $P \in \mathbb{R}^{N\times N}$ eine Permutationsmatrix und π die zugehörige Permutation. Man zeige:

(a) Die Spaltenvektoren von P sind paarweise orthonormal zueinander, $P^{-1} = P^{\top}$.

(b) Mit der Darstellung (4.5) gilt

$$P^{-1} = \left(\mathbf{e}_{\pi^{-1}(1)} \;\middle|\; \ldots \;\middle|\; \mathbf{e}_{\pi^{-1}(N)} \right).$$

Aufgabe 4.7 (*Numerische Aufgabe*). Man schreibe einen Code, der den Gauß-Algorithmus einmal ohne Pivot-, einmal mit Spaltenpivot- und schließlich mit *Totalpivotsuche* durchführt.

Bei letzterem werden – ausgehend von der Notation in Algorithmus 4.6 – beim Übergang $A^{(s)} \to A^{(s+1)}$ zunächst Indizes $p, q \in \{s, s+1, \ldots, N\}$ mit

$$|a_{pq}^{(s)}| \geq |a_{jk}^{(s)}|, \qquad j, k = s, s+1, \ldots, N,$$

bestimmt und $a_{pq}^{(s)}$ als *Pivotelement* verwendet. Man teste das Programm anhand des Beispiels $Ax = b$ mit

$$\begin{aligned} a_{jk} &= \frac{1}{j+k-1}, \qquad j, k = 1, 2, \ldots, N, \\ b_j &= \frac{1}{j+N-1}, \qquad j = 1, 2, \ldots, N. \end{aligned}$$

Für $N = 50, 100, 200$ und jede Pivotstrategie gebe man die Werte x_{10j}, $j = 1, 2, 3, \ldots, N/10$ aus.

Aufgabe 4.8. Man zeige: Eine Matrix $A \in \mathbb{R}^{N\times N}$ besitzt eine LR-Faktorisierung genau dann, wenn die Hauptuntermatrizen von A von der Form

$$\begin{pmatrix} a_{11} & \cdots & a_{1n} \\ \vdots & \ddots & \vdots \\ a_{n1} & \cdots & a_{nn} \end{pmatrix} \in \mathbb{R}^{n\times n} \qquad \text{für} \quad n = 1, 2, \ldots, N$$

alle regulär sind.

Aufgabe 4.9. Sei $A = (a_{jk}) \in \mathbb{R}^{N\times N}$ symmetrisch und positiv definit. Man zeige Folgendes:

(a) $a_{jj} > 0, \qquad j = 1, 2, \ldots, N,$
(b) $a_{jk}^2 < a_{jj}\, a_{kk}, \qquad j, k = 1, 2, \ldots, N, \quad j \neq k,$
(c) der betragsmäßig größte Eintrag von A liegt auf der Hauptdiagonalen.

Aufgabe 4.10. Man rechne nach, dass bei der Berechnung einer LR-Faktorisierung einer gegebenen Matrix $A \in \mathbb{R}^{N\times N}$ gemäß der Parkettierung von Crout insgesamt $(2N^3/3)(1 + \mathcal{O}(1/N))$ arithmetische Operationen anfallen.

Aufgabe 4.11. Man zeige Folgendes:

(a) Die Menge der skalierten (die Diagonaleinträge sind alle $= 1$) unteren Dreiecksmatrizen $L \in \mathbb{R}^{N\times N}$ bildet bezüglich der Matrixmultiplikation eine Untergruppe in $\mathbb{R}^{N\times N}$.
(b) Die Menge der regulären oberen Dreiecksmatrizen $R \in \mathbb{R}^{N\times N}$ bildet bezüglich der Matrixmultiplikation eine Untergruppe in $\mathbb{R}^{N\times N}$.
(c) Die Darstellung $A = LR$ einer nichtsingulären Matrix $A \in \mathbb{R}^{N\times N}$ als Produkt einer skalierten unteren Dreiecksmatrix L und einer regulären oberen Dreiecksmatrix R ist eindeutig (sofern sie existiert).

Aufgabe 4.12. Gegeben sei die Matrix

$$\begin{pmatrix} 1 & 2 & 3 & -4 \\ 2 & 8 & 6 & -14 \\ 3 & 6 & a & -15 \\ -4 & -14 & -15 & 30 \end{pmatrix}$$

mit einem reellen Parameter a. Man berechne die zugehörige LR-Faktorisierung beziehungsweise gebe an, für welchen Wert des Parameters a diese nicht existiert.

Aufgabe 4.13. Die Matrix $A \in \mathbb{R}^{N\times N}$ sei symmetrisch und positiv definit. Man gebe einen Algorithmus zur Gewinnung einer Faktorisierung $A = RR^\top$ an. Hierbei bezeichnet $R = (r_{jk}) \in \mathbb{R}^{N\times N}$ eine obere Dreiecksmatrix mit $r_{jj} > 0$ für alle j. Man begründe zudem die Durchführbarkeit dieses Verfahrens.

Aufgabe 4.14. Es sei $A = (a_{jk}) \in \mathbb{R}^{N\times N}$ eine symmetrische, positiv definite Bandmatrix der Bandbreite m, das heißt, $a_{jk} = 0$ für j,k mit $|j-k| \geq m$. Man weise nach, dass in der Cholesky-Faktorisierung $A = LL^\top$ die untere Dreiecksmatrix L eine Bandmatrix der Bandbreite m ist.

Aufgabe 4.15. Gegeben seien die Matrizen

$$A = \begin{pmatrix} 101 & 99 \\ 99 & 101 \end{pmatrix}, \qquad B = \begin{pmatrix} 101 & 99 \\ -99 & 101 \end{pmatrix}.$$

(a) Berechne die Konditionszahlen $\mathrm{cond}_\infty(A)$ und $\mathrm{cond}_\infty(B)$.

(b) Für die Vektoren

$$b = \begin{pmatrix} 1 \\ 1 \end{pmatrix}, \qquad \Delta b = \begin{pmatrix} \delta \\ \delta \end{pmatrix}, \qquad \Delta\widehat{b} = \begin{pmatrix} \delta \\ -\delta \end{pmatrix}$$

mit einer kleinen reellen Zahl $\delta > 0$ löse man die Gleichungssysteme

$$Ax = b, \qquad A(x + \Delta x) = b + \Delta b, \qquad A(x + \Delta\widehat{x}) = b + \Delta\widehat{b}.$$

Man vergleiche die jeweiligen relativen Fehler $\|\Delta x\|_\infty / \|x\|_\infty$ und $\|\Delta\widehat{x}\|_\infty / \|x\|_\infty$ mit der allgemeinen Fehlerabschätzung $\|\Delta x\|/\|x\| \leq \mathrm{cond}(A)\,\|\Delta b\|/\|b\|$.

Aufgabe 4.16. Für diese Aufgabe verwende man das folgende Theorem über die *Singulärwertzerlegung* einer Matrix:

Theorem 4.67. *Zu einer nichtsingulären Matrix $A \in \mathbb{R}^{N\times N}$ gibt es orthonormale Matrizen U, $V \in \mathbb{R}^{N\times N}$ und eine Diagonalmatrix $\Sigma = \mathrm{diag}\,(\sigma_1, \ldots, \sigma_N) \in \mathbb{R}^{N\times N}$ (mit $\sigma_1 \geq \sigma_2 \geq \ldots \geq \sigma_N > 0$), so dass*

$$A = V\Sigma U^\top.$$

(a) Man zeige: für jeden Vektor $x \in \mathbb{R}^N$ gilt ausgehend von der Darstellung als Linearkombination $x = \sum_{k=1}^N c_k u_k$ der paarweise orthonormalen Spaltenvektoren $u_1, u_2, \ldots, u_N \in \mathbb{R}^N$ der Matrix $U \in \mathbb{R}^{N\times N}$ Folgendes:

$$Ax = \sum_{k=1}^N c_k \sigma_k v_k,$$

wobei $v_1, v_2, \ldots, v_N \in \mathbb{R}^N$ die paarweise orthonormalen Spaltenvektoren der Matrix $V \in \mathbb{R}^{N\times N}$ bezeichnen.

(b) Man gebe die Werte von $\|A\|_2$, $\|A^{-1}\|_2$ sowie $\mathrm{cond}_2(A)$ über die Singulärwerte der Matrix A an.

(c) Zur Lösung von

$$A(x + \Delta x) = b + \Delta b$$

gebe man mithilfe der Matrix U diejenigen Vektoren $b \in \mathbb{R}^N$ beziehungsweise $\Delta b \in \mathbb{R}^N$ an, die in den Abschätzungen

$$\begin{aligned} \|b\|_2 &\leq \|A\|_2 \|x\|_2, \\ \|\Delta x\|_2 &\leq \|A^{-1}\|_2 \|\Delta b\|_2, \\ \frac{\|\Delta x\|_2}{\|x\|_2} &\leq \mathrm{cond}_2(A) \frac{\|\Delta b\|_2}{\|b\|_2}, \end{aligned}$$

Gleichheit ergeben.

Aufgabe 4.17. Für eine reguläre Matrix $A \in \mathbb{R}^{N\times N}$ sei $B \in \mathbb{R}^{N\times N}$ eine Näherung für A^{-1} und $\|\cdot\| : \mathbb{R}^{N\times N} \to \mathbb{R}$ eine beliebige submultiplikative Matrixnorm. Man zeige:

$$\begin{aligned} \frac{\|A^{-1} - B\|}{\|A^{-1}\|} &\leq \min\{\|AB - I\|, \|BA - I\|\}, \\ \|BA - I\| &\leq \operatorname{cond}(A)\|AB - I\| \leq \operatorname{cond}(A)^2\|BA - I\|. \end{aligned}$$

Zu Testzwecken betrachte man die beiden Matrizen

$$A = \begin{pmatrix} 9999 & 9998 \\ 10000 & 9999 \end{pmatrix}, \qquad B = \begin{pmatrix} 9999.9999 & -9997.0001 \\ -10001 & 9998 \end{pmatrix},$$

und berechne die Matrizen $BA - I \in \mathbb{R}^{N\times N}$ sowie $AB - I \in \mathbb{R}^{N\times N}$.

Aufgabe 4.18. (a) Es sei $B = (b_{jk}) \in \mathbb{R}^{N\times N}$ eine reguläre Matrix, die zudem *zeilenäquilibriert* ist, das heißt,

$$\sum_{k=1}^{N} |b_{jk}| = 1, \qquad j = 1, 2, \ldots, N.$$

Man zeige, dass für jede reguläre Diagonalmatrix $D \in \mathbb{R}^{N\times N}$ die folgende Abschätzung gilt,

$$\operatorname{cond}_\infty(B) \leq \operatorname{cond}_\infty(DB).$$

(b) Sei $A \in \mathbb{R}^{N\times N}$ eine reguläre Matrix. Man zeige: es gibt eine Diagonalmatrix $D \in \mathbb{R}^{N\times N}$, so dass DA zeilenäquilibriert ist, und dann gilt

$$\operatorname{cond}_\infty(DA) \leq \operatorname{cond}_\infty(A).$$

Aufgabe 4.19. Es sei $A = (a_{jk}) \in \mathbb{R}^{N\times N}$ eine reguläre Matrix. Zeige mithilfe der QR-Faktorisierung die *hadamardsche Determinantenabschätzung*

$$|\det(A)| \leq \prod_{k=1}^{n} \Big(\sum_{j=1}^{n} |a_{jk}|^2\Big)^{1/2}.$$

Aufgabe 4.20. Man zeige für eine nichtsinguläre Matrix $A \in \mathbb{R}^{N\times N}$ und Vektoren $u, v \in \mathbb{R}^N$:

(a) Im Fall $v^\top A^{-1}u \neq -1$ gilt die *Sherman-Morrison-Formel*

$$(A + uv^\top)^{-1} = A^{-1} - \frac{A^{-1}uv^\top A^{-1}}{1 + v^\top A^{-1}u}.$$

(b) Im Fall $v^\top A^{-1}u = -1$ ist die Matrix $A + uv^\top$ singulär.

Aufgabe 4.21. Transformieren Sie die Matrix

$$A = \begin{pmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 1 & 0 & 1 \\ 0 & 0 & 1 \end{pmatrix}$$

mittels Householder-Transformationen auf obere Dreiecksgestalt.

Aufgabe 4.22 (*Numerische Aufgabe*). Man schreibe einen Code zur Lösung eines linearen Gleichungssystems mittels Householder-Transformationen. Man teste das Programm anhand des Beispiels $Ax = b$ mit

$$A = \begin{pmatrix} \delta & 0 & \cdots & 0 & 1 \\ -1 & \delta & \ddots & \vdots & 1 \\ -1 & \ddots & \ddots & 0 & 1 \\ \vdots & & \ddots & \delta & 1 \\ -1 & -1 & \cdots & -1 & 1 \end{pmatrix} \in \mathbb{R}^{N\times N}, \qquad b = \begin{pmatrix} 1 + \delta \\ \delta \\ -1 + \delta \\ \vdots \\ 3 - N + \delta \\ 2 - N \end{pmatrix} \in \mathbb{R}^N,$$

mit $N = 20$ und $\delta = 0.1$. Man gebe den Lösungsvektor $x = (x_1, x_2, \ldots, x_N)^\top$ aus.

5 Nichtlineare Gleichungssysteme

5.1 Vorbemerkungen

Im Folgenden sei $F : \mathbb{R}^N \to \mathbb{R}^N$ eine gegebene Funktion und $x_* \in \mathbb{R}^N$ eine Nullstelle von F,

$$F(x_*) \;=\; 0,$$

die es zu bestimmen gilt. Typischerweise lässt sich ein solches nichtlineares Gleichungssystem nur approximativ lösen, was im Folgenden mittels Iterationsverfahren der Form

$$x_{n+1} \;=\; \Phi(x_n) \qquad \text{für} \quad n \;=\; 0, 1, \ldots \tag{5.1}$$

geschehen soll mit einer geeigneten stetigen *Iterationsfunktion* $\Phi : \mathbb{R}^N \to \mathbb{R}^N$. Dabei soll die Abbildung Φ so beschaffen sein, dass Konvergenz im folgenden Sinne vorliegt.

Definition 5.1. Sei $\Phi : \mathbb{R}^N \to \mathbb{R}^N$ eine Iterationsfunktion. Das Verfahren (5.1) zur Bestimmung von $x_* \in \mathbb{R}^N$ heißt (lokal) *konvergent*, wenn eine Zahl $\delta > 0$ existiert, so dass für alle Startwerte

$$x_0 \;\in\; \mathcal{B}(x_*; \delta), \qquad\qquad \mathcal{B}(x_*; \delta) \;:=\; \{\, y \in \mathbb{R}^N : \; \| y - x_* \| < \delta \,\}$$

gilt

$$\| x_n \;-\; x_* \| \to 0 \qquad \text{für} \quad n \to \infty. \tag{5.2}$$

Hier bezeichnet $\| \cdot \| : \mathbb{R}^N \to \mathbb{R}$ eine nicht näher spezifizierte Vektornorm.

Bemerkung 5.2. Da die Iterationsfunktion $\Phi : \mathbb{R}^N \to \mathbb{R}^N$ als stetig in x_* vorausgesetzt ist, handelt es sich aufgrund der Konvergenz (5.2) bei $x_* \in \mathbb{R}^N$ notwendigerweise um einen *Fixpunkt* von Φ,

$$\Phi(x_*) \;=\; x_*,$$

denn

$$x_* \;=\; \lim_{n\to\infty} x_{n+1} \;=\; \lim_{n\to\infty} \Phi(x_n) \;=\; \Phi\Big(\lim_{n\to\infty} x_n \Big) \;=\; \Phi(x_*).$$

Daher bezeichnet man das Verfahren (5.1) als *Fixpunktiteration*. △

Mehr noch als Konvergenz (5.2) ist wünschenswert, dass das Verfahren (5.1) eine möglichst hohe Konvergenzordnung im Sinne der folgenden Definition besitzt.

Definition 5.3. Sei $\Phi : \mathbb{R}^N \to \mathbb{R}^N$ eine Iterationsfunktion mit Fixpunkt $x_* \in \mathbb{R}^N$. Das Verfahren (5.1) heißt (lokal) *konvergent von (mindestens) der Ordnung* $p \geq 1$, wenn ein $\delta > 0$ existiert, so dass für alle Startwerte $x_0 \in \mathcal{B}(x_*; \delta)$ gilt

$$\|x_{n+1} - x_*\| \leq C\|x_n - x_*\|^p \qquad \text{für} \quad n = 0, 1, \ldots, \tag{5.3}$$

mit einer Konstanten $0 \leq C < \infty$, wobei im Fall $p = 1$ noch $C < 1$ gefordert wird. Bei Konvergenz der Ordnung $p = 1$ beziehungsweise $p = 2$ spricht man dann von (mindestens) *linearer* beziehungsweise *quadratischer* Konvergenz.

Das Verfahren (5.1) heißt *konvergent von genau* der Ordnung p, wenn es konvergent von der Ordnung p ist und keine höhere Konvergenzordnung besitzt.

Bemerkung 5.4. (a) Lineare Konvergenz impliziert für $x_0 \in \mathcal{B}(x_*; \delta)$

$$\|x_n - x_*\| \leq C^n\|x_0 - x_*\|, \qquad n = 0, 1, \ldots \tag{5.4}$$

mit einer Konstanten $0 < C < 1$. Insbesondere ist das Verfahren also lokal konvergent.

(b) Ein Verfahren der Konvergenzordnung $p > 1$ besitzt für jedes $1 \leq q \leq p$ formal auch die niedrigere Konvergenzordnung q: für Startwerte

$$x_0 \in \mathcal{B}(x_*; \widehat{\delta}), \qquad \widehat{\delta} := \min\left\{\delta, \left(\tfrac{1}{2C}\right)^{1/(p-1)}\right\} \quad \text{mit} \quad C \text{ aus (5.3)},$$

erhält man induktiv $\|x_n - x_*\| \leq 2^{-n}\|x_0 - x_*\|$ für $n = 0, 1, \ldots$, somit liegt lineare Konvergenz vor. Weiter berechnet man

$$\begin{aligned} \|x_{n+1} - x_*\| &\leq C\|x_n - x_*\|^p = C\overbrace{\|x_n - x_*\|}^{\leq \delta^{p-q}}\|x_n - x_*\|^q \\ &\leq \delta^{p-q}C\|x_n - x_*\|^q \qquad \text{für} \quad n = 0, 1, \ldots, \end{aligned}$$

was die angegebene Konvergenzordnung $1 < q \leq p$ liefert.

(c) Je höher die Konvergenzordnung eines Verfahrens, desto schneller werden die Iterierten den gesuchten Wert x_* approximieren, denn für Zahlen $0 \leq q < p$ sowie Startwerte x_0 hinreichend nahe bei x_* und n hinreichend groß gilt $\|x_n - x_*\| \ll 1$ und damit $\|x_n - x_*\|^p \ll \|x_n - x_*\|^q$. △

5.2 Der eindimensionale Fall

5.2.1 Ein allgemeines Resultat

Das folgende Theorem befasst sich mit Verfahren (5.1) im eindimensionalen Fall $N = 1$ und liefert Konvergenzresultate für hinreichend gute Startwerte x_0.

Theorem 5.5. *Sei $\Phi : \mathbb{R} \to \mathbb{R}$ eine Iterationsfunktion mit Fixpunkt $x_* \in \mathbb{R}$, die zudem in x_* insgesamt p-mal differenzierbar sei mit $p \in \mathbb{N}$. Weiter sei*

$$\left\{ \begin{array}{ll} \Phi^{(k)}(x_*) = 0, \quad k = 1, 2, \ldots, p-1, & \text{falls } p \geq 2 \\ |\Phi'(x_*)| < 1, & \text{falls } p = 1 \end{array} \right\}$$

erfüllt. Dann ist das Verfahren (5.1) lokal mindestens konvergent von der Ordnung p. Wenn weiterhin $\Phi^{(p)}(x_*) \neq 0$ gilt, so liegt die genaue Konvergenzordnung p vor.

BEWEIS. Eine Taylorentwicklung der Funktion Φ im Punkt x_* liefert

$$\begin{aligned}\Phi(x) &= \sum_{k=0}^{p} \frac{\Phi^{(k)}(x_*)}{k!}(x - x_*)^k + \mathcal{O}(|x - x_*|^p) \\ &= \underbrace{\Phi(x_*)}_{= x_*} + \frac{\Phi^{(p)}(x_*)}{p!}(x - x_*)^p + \mathcal{O}(|x - x_*|^p) \qquad \text{für } x \to x_*,\end{aligned}$$

und somit

$$\frac{\Phi(x) - x_*}{(x - x_*)^p} \to \frac{\Phi^{(p)}(x_*)}{p!} \qquad \text{für } x \to x_*. \tag{5.5}$$

Folglich existiert zu jedem $\varepsilon > 0$ eine Zahl $\delta > 0$ mit

$$|\Phi(x) - x_*| \leq \big(\frac{|\Phi^{(p)}(x_*)|}{p!} + \varepsilon\big)|x - x_*|^p \qquad \text{für } x \in \mathcal{B}(x_*; \delta), \tag{5.6}$$

wobei im Fall $p = 1$ noch $\varepsilon > 0$ so klein zu wählen ist, dass die Ungleichung $|\Phi'(x_*)| + \varepsilon < 1$ erfüllt ist. Wenn man nun

$$x_0 \in \mathcal{B}(x_*; \widehat{\delta}) \qquad \big(\widehat{\delta} := \min\{\delta, \big(\tfrac{1}{2C}\big)^{1/(p-1)}\}, \qquad C := \frac{|\Phi^{(p)}(x_*)|}{p!} + \varepsilon\big)$$

wählt[1], so gilt auch $x_n \in \mathcal{B}(x_*; \widehat{\delta})$ für $n = 1, 2, \ldots$, und (5.6) liefert dann die angegebene Konvergenzordnung $\geq p$. Unter der Zusatzbedingung $\Phi^{(p)}(x_*) \neq 0$ gibt es wegen der Konvergenzaussage (5.5) für $0 < \varepsilon < |\Phi^{(p)}(x_*)|/p!$ eine Zahl $\delta > 0$ mit

$$|\Phi(x) - x_*| \geq \big(\frac{|\Phi^{(p)}(x_*)|}{p!} - \varepsilon\big)|x - x_*|^p \qquad \text{für } x \in \mathcal{B}(x_*; \delta),$$

was die genaue Konvergenzordnung p liefert. □

5.2.2 Das Newton-Verfahren im eindimensionalen Fall

Zur Bestimmung einer Nullstelle $x_* \in \mathbb{R}$ einer gegebenen Funktion $f : \mathbb{R} \to \mathbb{R}$ wird im Folgenden das Newton-Verfahren

$$x_{n+1} = x_n - \frac{f(x_n)}{f'(x_n)} =: \Phi(x_n), \qquad n = 0, 1, \ldots \tag{5.7}$$

betrachtet. Die geometrische Bedeutung des Newton-Verfahrens ist in Bild 5.1 veranschaulicht.

In dem nachfolgenden Theorem wird unter verschiedenen Voraussetzungen jeweils die Konvergenzordnung von Verfahren (5.7) angegeben[2].

[1] vergleiche hierzu die Argumentation in Teil (b) der Bemerkung 5.4
[2] unter Heranziehung von Theorem 5.5

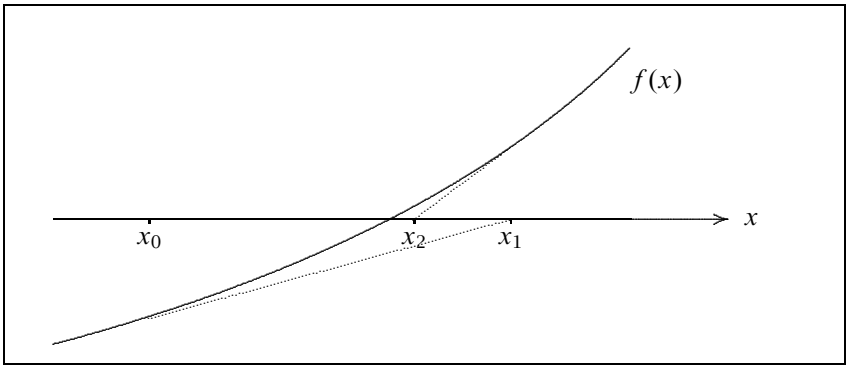


Bild 5.1: Veranschaulichung der Vorgehensweise beim Newton-Verfahren

Theorem 5.6. *Die Funktion $f : \mathbb{R} \to \mathbb{R}$ besitze eine Nullstelle $x_* \in \mathbb{R}$ und sei in einer Umgebung von x_* hinreichend oft differenzierbar.*

(a) Im Fall $f'(x_*) \neq 0$ konvergiert das Newton-Verfahren (5.7) mindestens quadratisch. (Falls $f''(x_*) = 0$ gilt, so ist es sogar konvergent von der Ordnung $\geq p = 3$.)

(b) Ist hingegen x_* eine m-fache Nullstelle von f mit einer Zahl $m \geq 2$, gilt also

$$f(x) \;=\; (x - x_*)^m g(x), \qquad\qquad g(x_*) \neq 0,$$

und ist die Funktion g zweimal differenzierbar in x_*, so ist die Iterationsfunktion Φ aus (5.7) differenzierbar in x_* mit

$$\Phi'(x_*) \;=\; 1 - \frac{1}{m}. \tag{5.8}$$

Das Newton-Verfahren (5.7) ist in diesem Fall also (genau) linear konvergent.

BEWEIS. Die Aussagen ergeben sich mit Theorem 5.5 angewandt auf die Funktion $\Phi(x) := x - f(x)/f'(x)$ sowie mit den folgenden Darstellungen: im Fall (a) hat man

$$\Phi' \;=\; 1 - \frac{(f')^2 - f\,f''}{(f')^2} \;=\; \frac{f\,f''}{(f')^2}, \qquad \Phi'' \;=\; \frac{(f')^3 f'' \,+\, f(f')^2 f''' - 2ff'(f'')^2}{(f')^4},$$

so dass also

$$\Phi(x_*) \;=\; x_*, \qquad\qquad \Phi'(x_*) \;=\; 0, \qquad\qquad \Phi''(x_*) \;=\; \frac{f''(x_*)}{f'(x_*)}$$

gilt. Im Fall (b) erhält man

$$f'(x) \;=\; m(x - x_*)^{m-1} g(x) \;+\; (x - x_*)^m g'(x)$$

und somit

$$\Phi(x) \;=\; x - \frac{f(x)}{f'(x)} \;=\; x - \frac{(x - x_*)g(x)}{mg(x) \,+\, (x - x_*)g'(x)} \;=:\; x - \frac{Z(x)}{N(x)},$$

$$\Phi'(x) \;=\; 1 - \frac{[g(x) \,+\, (x - x_*)g'(x)]N(x) \;-\; Z(x)[(m+1)g'(x) \,+\, (x - x_*)g''(x)]}{N(x)^2}.$$

Dies liefert schließlich (5.8), also $0 < \Phi'(x_*) < 1$ und insbesondere auch $\Phi'(x_*) \neq 0$.

□

5.3 Der banachsche Fixpunktsatz

In Abschnitt 5.2.1 ist das allgemeine Verfahren (5.1) im eindimensionalen Fall $N = 1$ und für hinreichend glatte Iterationsfunktionen $\Phi : \mathbb{R} \to \mathbb{R}$ sowie hinreichend gute Startwerte x_0 betrachtet worden. Im folgenden Theorem nun wird lineare Konvergenz für das allgemeine Verfahren (5.1) nachgewiesen für den mehrdimensionalen Fall $N \geq 1$ und ohne Differenzierbarkeitsbedingungen an Φ, und als Startvektor werden beliebige Elemente x_0 der zugrunde gelegten Menge zugelassen; überdies erhalt man die Existenz eines eindeutigen Fixpunktes. Dafür ist allerdings die globale Kontraktionseigenschaft (5.9) eine relativ schwer wiegende Forderung an die Iterationsfunktion Φ.

Theorem 5.7. *Sei $\mathcal{M} \subset \mathbb{R}^N$ eine abgeschlossene Teilmenge, und die Abbildung $\Phi : \mathcal{M} \to \mathcal{M}$ sei bezüglich einer Vektornorm $\|\cdot\| : \mathbb{R}^N \to \mathbb{R}$ eine* Kontraktion*, das heißt, für eine Konstante $0 < L < 1$ sei*

$$\|\Phi(x) - \Phi(y)\| \leq L\|x - y\|, \qquad x, y \in \mathcal{M}, \tag{5.9}$$

erfüllt. Dann gilt Folgendes:

- *Φ besitzt genau einen Fixpunkt $x_* \in \mathcal{M}$;*
- *Für jeden Startwert $x_0 \in \mathcal{M}$ liefert die* Fixpunktiteration[3]

$$x_{n+1} = \Phi(x_n), \qquad n = 0, 1, \ldots \tag{5.10}$$

eine gegen x_* konvergierende Folge, und es gilt genauer

$$\|x_n - x_*\| \leq \frac{L}{1-L}\|x_n - x_{n-1}\| \leq \frac{L^n}{1-L}\|x_1 - x_0\|, \qquad n = 1, 2, \ldots . \tag{5.11}$$

BEWEIS. Sind $x_*, \widehat{x}_* \in \mathcal{M}$ Fixpunkte von Φ, so gilt

$$\|x_* - \widehat{x}_*\| = \|\Phi(x_*) - \Phi(\widehat{x}_*)\| \leq L\|x_* - \widehat{x}_*\|$$

beziehungsweise $(1 - L)\|x_* - \widehat{x}_*\| \leq 0$, was $x_* = \widehat{x}_*$ bedeutet. Im Folgenden soll die Existenz eines Fixpunktes von Φ nachgewiesen werden, was mithilfe der Fixpunktiteration geschieht. Die dabei erzielten Zwischenergebnisse liefern dann auch unmittelbar die Abschätzungen (5.11). Sei also der Startvektor $x_0 \in \mathcal{M}$ beliebig, und $(x_n) \subset \mathbb{R}^N$ bezeichne die zugehörige Folge der Fixpunktiteration (5.10). Mithilfe einer Teleskopsumme erhält man dann für $n, k \in \mathbb{N}_0$ unter Verwendung von

[3] vergleiche (5.1)

$\|x_{j+1} - x_j\| \le L\|x_j - x_{j-1}\|$ für $j = 1, 2, \ldots$ die folgenden Abschätzungen:

$$\begin{aligned}
\|x_{n+k} - x_n\| &= \Big\| \sum_{\ell=n}^{n+k-1} x_{\ell+1} - x_\ell \Big\| \le \sum_{\ell=n}^{n+k-1} \|x_{\ell+1} - x_\ell\| \\
&\le \Big(\sum_{\ell=n}^{n+k-1} L^{\ell-n} \Big) \|x_{n+1} - x_n\| = \Big(\sum_{\ell=0}^{k-1} L^{\ell} \Big) \|x_{n+1} - x_n\| \\
&\le \frac{1-L^k}{1-L} \|x_{n+1} - x_n\| \le \frac{1}{1-L} \|x_{n+1} - x_n\| \\
&\le \frac{L}{1-L} \|x_n - x_{n-1}\| \le \frac{L^n}{1-L} \|x_1 - x_0\|.
\end{aligned}$$

Damit gilt insbesondere

$$\|x_{n+k} - x_n\| \le \frac{L}{1-L} \|x_n - x_{n-1}\| \le \frac{L^n}{1-L} \|x_1 - x_0\|, \qquad n, k \ge 0, \quad (5.12)$$

und somit ist $(x_n) \subset \mathbb{R}^N$ Cauchyfolge mit einem Grenzwert, der zudem Fixpunkt von Φ ist[4] und daher mit $x_* \in \mathcal{M}$ übereinstimmt. Der Grenzübergang "$k \to \infty$" in (5.12) liefert die angegebene Abschätzung (5.11). □

Bemerkung 5.8. (a) Der Ausdruck $(L^n/(1-L))\|x_1 - x_0\|$ in (5.11) kann für jedes n vor Beginn der Iteration bestimmt werden (nur x_1 wird hierzu benötigt) und ermöglicht eine *a priori-Fehlerabschätzung* für den Approximationsfehler $\|x_n - x_*\|$.

(b) Der mittlere Ausdruck $(L/(1-L))\|x_n - x_{n-1}\|$ in (5.11) hingegen kann im n-ten Iterationsschritt bestimmt werden und ermöglicht eine *a posteriori-Fehlerabschätzung* für den Approximationsfehler $\|x_n - x_*\|$.

(c) Praktisch geht man so vor: für eine vorgegebene Fehlerschranke $\varepsilon > 0$ wird die Iteration in Schritt $n = n(\varepsilon)$ abgebrochen, falls erstmalig

$$\frac{L}{1-L} \|x_n - x_{n-1}\| \le \varepsilon$$

gilt, und die a posteriori-Fehlerabschätzung garantiert dann die gewünschte Fehlerabschätzung $\|x_n - x_*\| \le \varepsilon$. Die a priori-Fehlerabschätzung ermöglicht die Abschätzung

$$n(\varepsilon) \le \lceil a \rceil, \qquad a = \frac{\log\big(\frac{\|x_1 - x_0\|}{(1-L)\varepsilon}\big)}{\log(1/L)} \quad (5.13)$$

für die Anzahl der nötigen Iterationsschritte, wobei $\lceil a \rceil$ die kleinste ganze Zahl $\ge a$ bezeichnet. △

[4] was aus der Bemerkung 5.2 folgt unter Beachtung der Tatsache, dass wegen der Kontraktionseigenschaft (5.9) die Abbildung Φ insbesondere stetig ist

Beispiel 5.9. Für

$$\begin{aligned} f(x) &:= x - e^{-x}, \qquad x \in \mathbb{R}, \\ f(x_*) &= 0 \quad \text{für } x_* \approx 0.56714329 \end{aligned}$$

soll die Nullstelle x_* bestimmt werden unter Anwendung der Fixpunktiteration (5.1) mit der Iterationsfunktion

$$\Phi(x) := e^{-x}, \qquad x \in \mathbb{R}.$$

Auf dem Intervall $\mathcal{M} = [0.5, 0.69]$ ist die Eigenschaft $\Phi(\mathcal{M}) \subset \mathcal{M}$ ebenso erfüllt wie die Kontraktionseigenschaft (5.9) mit

$$L = \max_{x \in [0.5, 0.69]} |\Phi'(x)| = \max_{x \in [0.5, 0.69]} e^{-x} = e^{-1/2} \approx 0.606531.$$

In der folgenden Tabelle sind einige der durch das Verfahren (5.1) gewonnenen Iterierten aufgelistet, wobei als Startwert $x_0 = 0.55$ gewählt ist und in der vorliegenden Situation das Verfahren von der speziellen Form $x_{n+1} = e^{-x_n}$, $n = 0, 1, \ldots$, ist.

n	x_n	n	x_n	n	x_n
		⋮	⋮	⋮	⋮
0	0.55000000	10	0.56708394	20	0.56714309
1	0.57694981	11	0.56717695	21	0.56714340
2	0.56160877	12	0.56712420	22	0.56714323
3	0.57029086	13	0.56715412	23	0.56714332
4	0.56536097	14	0.56713715	24	0.56714327
⋮	⋮	⋮	⋮	⋮	⋮

Die Situation soll für $n = 12$ genauer betrachtet werden. Die Fehlerabschätzung (5.11) liefert in diesem Fall

$$1.91 \cdot 10^{-5} \approx |x_{12} - x_*| \leq 8.13 \cdot 10^{-5} \leq 1.70 \cdot 10^{-4},$$

so dass die a posteriori-Abschätzung den wirklichen Fehler etwa um den Faktor 4 überschätzt, und die a priori-Abschätzung überschätzt den wirklichen Fehler etwa um den Faktor 10.

Das praktische Vorgehen soll nun für die spezielle Fehlerschranke $\varepsilon = 0.0076$ illustriert werden. Die a posteriori-Abschätzung liefert $n(\varepsilon) = 4$ als Stoppindex, $|x_4 - x_*| \leq \varepsilon$. Die Abschätzung (5.13) liefert mit $n(\varepsilon) \leq 16$ eine Überschätzung. Schließlich ist anzumerken, dass schon in Schritt 2 der (im Allgemeinen unbekannte) Approximationsfehler die Schranke ε unterschreitet, $|x_2 - x_*| \approx 0.0055 \leq \varepsilon$. △

5.4 Das Newton-Verfahren im mehrdimensionalen Fall

Für eine gegebene Funktion $F : \mathbb{R}^N \to \mathbb{R}^N$ soll nun die Konvergenz des Newton-Verfahrens zur Lösung des Gleichungssystems $F(x) = 0$ im mehrdimensionalen Fall $N \geq 1$ untersucht werden.[5]

[5] Für den eindimensionalen Fall sowie hinreichend gute Startwerte x_0 ist dies bereits in Abschnitt 5.2.2 geschehen.

5.4.1 Einige Begriffe aus der Analysis

In diesem Abschnitt werden einige Hilfsmittel aus der Analysis bereitgestellt. Im Folgenden wird mit $\|\cdot\|$ sowohl eine (beliebig aber fest gewählte) Vektornorm auf $\mathbb{R}^N$ als auch die induzierte Matrixnorm bezeichnet.

Bekanntlich heißt eine Funktion $F : \mathbb{R}^N \to \mathbb{R}^N$ in einem Punkt $x \in \mathbb{R}^N$ *differenzierbar*, falls eine lineare Abbildung $\mathcal{D}_x F : \mathbb{R}^N \to \mathbb{R}^N$ existiert mit der Eigenschaft

$$\frac{\|F(x+h) - F(x) - (\mathcal{D}_x F)(h)\|}{\|h\|} \to 0 \qquad \text{für} \quad \mathbb{R}^N \ni h \to 0.$$

Die Abbildung $\mathcal{D}_x F$ ist so eindeutig festgelegt und wird durch die Jacobi-Matrix repräsentiert,

$$(\mathcal{D}_x F)(z) = \mathcal{J}(x)z, \qquad \mathcal{J}(x) := \begin{pmatrix} \frac{\partial F_1}{\partial x_1}(x) & \frac{\partial F_1}{\partial x_2}(x) & \cdots & \frac{\partial F_1}{\partial x_N}(x) \\ \frac{\partial F_2}{\partial x_1}(x) & \frac{\partial F_2}{\partial x_2}(x) & \cdots & \frac{\partial F_2}{\partial x_N}(x) \\ \vdots & \vdots & & \vdots \\ \frac{\partial F_N}{\partial x_1}(x) & \frac{\partial F_N}{\partial x_2}(x) & \cdots & \frac{\partial F_N}{\partial x_N}(x) \end{pmatrix} \in \mathbb{R}^{N\times N}.$$

Die Funktion $F : \mathbb{R}^N \to \mathbb{R}^N$ heißt auf einer Menge $\mathcal{M} \subset \mathbb{R}^N$ differenzierbar, falls sie in jedem Punkt $x \in \mathcal{M}$ differenzierbar ist. Eine Menge $\mathcal{M} \subset \mathbb{R}^N$ heißt *konvex*, falls für je zwei Elemente $x,\ y \in \mathcal{M}$ auch die Verbindungsstrecke von x nach y zu $\mathcal{M}$ gehört, das heißt,

$$\{x + t(y-x) : 0 \le t \le 1\} \subset \mathcal{M}, \qquad x,\ y \in \mathcal{M}.$$

Im folgenden Lemma wird als Nachtrag zu Abschnitt 5.3 eine hinreichende Bedingung für die in Theorem 5.7 auftretende Kontraktionsbedingung (5.9) angegeben (für $\Phi = F$).

Lemma 5.10. *Eine gegebene Funktion $F : \mathbb{R}^N \to \mathbb{R}^N$ sei auf einer offenen konvexen Menge $\mathcal{M} \subset \mathbb{R}^N$ differenzierbar, und für eine Konstante $0 \le L < \infty$ gelte*

$$\|\mathcal{D}_x F\| \le L, \qquad x \in \mathcal{M},$$

wobei $\mathcal{D}_x F$ mit der zugehörigen Jacobi-Matrix $\mathcal{J}(x)$ identifiziert wird. Dann gilt die Abschätzung

$$\|F(x) - F(y)\| \le L\|x - y\|, \qquad x,\ y \in \mathcal{M}.$$

BEWEIS. Die Aussage des Lemmas ergibt sich unmittelbar aus dem Mittelwertsatz $F(x) - F(y) = \int_0^1 \mathcal{D}_{y+t(x-y)} F(x-y)\, dt$. □

Das nachfolgende Lemma über eine Variante der Taylorentwicklung für Funktionen mehrerer Veränderlicher wird beim Beweis des darauf folgenden Konvergenzresultats für das Newton-Verfahren benötigt.

Lemma 5.11. *Eine gegebene Funktion $F : \mathbb{R}^N \to \mathbb{R}^N$ sei auf der offenen konvexen Menge $\mathcal{M} \subset \mathbb{R}^N$ differenzierbar, und für eine Konstante $0 \le L < \infty$ gelte*

$$\| \mathcal{D}_x F - \mathcal{D}_y F \| \le L \| x - y \|, \qquad x, y \in \mathcal{M}.$$

Dann gilt die Abschätzung

$$\| F(x) - F(y) - (\mathcal{D}_y F)(x-y) \| \le \frac{L}{2} \| x - y \|^2, \qquad x, y \in \mathcal{M}.$$

BEWEIS. Nach Voraussetzung ist für beliebige $x, y \in \mathcal{M}$ die Funktion

$$\varphi : [0,1] \to \mathbb{R}^N, \qquad t \mapsto F(y + t(x-y))$$

stetig differenzierbar auf dem Intervall $[0,1]$, und die Kettenregel liefert

$$\varphi'(t) = (\mathcal{D}_{y + t(x-y)} F)(x-y), \qquad 0 \le t \le 1.$$

Für $0 \le t \le 1$ erhält man so die Abschätzung

$$\begin{aligned} \| \varphi'(t) - \varphi'(0) \| &= \| (\mathcal{D}_{y+t(x-y)} F)(x-y) - (\mathcal{D}_y F)(x-y) \| \\ &\le \| \mathcal{D}_{y+t(x-y)} F - \mathcal{D}_y F \| \| x - y \| \le Lt \| x - y \|^2. \end{aligned}$$

Wegen

$$\begin{aligned} \Delta &:= F(x) - F(y) - (\mathcal{D}_y F)(x-y) = \varphi(1) - \varphi(0) - \varphi'(0) \\ &= \int_0^1 \varphi'(t) - \varphi'(0)\, dt \end{aligned}$$

erhält man so schließlich die Aussage des Lemmas,

$$\| \Delta \| \le \int_0^1 \| \varphi'(t) - \varphi'(0) \|\, dt \le L \| x - y \|^2 \int_0^1 t\, dt = \frac{L}{2} \| x - y \|^2. \qquad \square$$

5.4.2 Das Newton-Verfahren und seine Konvergenz

Im Folgenden wird das Newton-Verfahren

$$x_{n+1} = x_n - (\mathcal{D}_{x_n} F)^{-1}(F(x_n)), \qquad n = 0, 1, \ldots, \tag{5.14}$$

zur Bestimmung einer Nullstelle der Funktion F betrachtet.

Bemerkung 5.12. In numerischen Implementierungen des Newton-Verfahrens geht man in den Schritten $n = 0, 1, \ldots$ jeweils so vor: Ausgehend von der bereits berechneten Iterierten $x_n \in \mathbb{R}^N$ löst man zunächst das lineare Gleichungssystem $(\mathcal{D}_{x_n} F)\Delta_n = -F(x_n)$ und erhält anschließend $x_{n+1} = x_n + \Delta_n$, so dass auf die aufwändige Matrixinversion $(\mathcal{D}_{x_n} F)^{-1}$ verzichtet werden kann. △

Das nachfolgende Theorem liefert unter gewissen Voraussetzungen quadratische Konvergenz sowie eine Menge von zulässigen Startvektoren x_0, die Existenz einer Nullstelle x_* wird vorausgesetzt.

Theorem 5.13. *Eine gegebene Funktion $F : \mathbb{R}^N \to \mathbb{R}^N$ sei auf der offenen konvexen Menge $\mathcal{M} \subset \mathbb{R}^N$ differenzierbar, und $x_* \in \mathcal{M}$ sei eine Nullstelle von F. Wenn für gewisse Zahlen $r,\ \beta,\ L > 0$ Folgendes gilt,*

$$\mathcal{B}(x_*; r) \subset \mathcal{M}, \qquad \mathcal{D}_{x_*} F \text{ ist invertierbar}, \qquad \|(\mathcal{D}_{x_*} F)^{-1}\| \le \beta,$$
$$\|\mathcal{D}_x F - \mathcal{D}_y F\| \le L\|x - y\|, \qquad x,\ y \in \mathcal{M},$$

so ist für jeden Startwert

$$x_0 \in \mathcal{B}(x_*; \delta) \qquad \text{mit} \quad \delta := \min\{r,\ \tfrac{1}{2\beta L}\}$$

das Newton-Verfahren (5.14) wohldefiniert, und es liegt lokale quadratische Konvergenz vor: für die Iterierten gilt

$$\|x_{n+1} - x_*\| \le \beta L\|x_n - x_*\|^2 \le \tfrac{1}{2}\|x_n - x_*\|, \qquad n = 0, 1, \ldots . \tag{5.15}$$

BEWEIS. Zunächst wird gezeigt, dass für jeden Vektor $x \in \mathbb{R}^N$ die folgende Implikation gilt:

$$\|x - x_*\| < \delta \quad \Longrightarrow \quad \mathcal{D}_x F \text{ ist invertierbar}, \quad \|(\mathcal{D}_x F)^{-1}\| \le 2\beta. \tag{5.16}$$

Die Voraussetzung $\|x - x_*\| < \delta$ impliziert nämlich

$$\eta := \|(\mathcal{D}_{x_*} F)^{-1}\|\,\|\mathcal{D}_x F - \mathcal{D}_{x_*} F\| \le \beta L\|x - x_*\| \le \beta L\delta \le \tfrac{1}{2},$$

und Korollar 4.50 liefert dann die Invertierbarkeit von $\mathcal{D}_x F$ sowie die angegebene Abschätzung (5.16),

$$\|(\mathcal{D}_x F)^{-1}\| \le \frac{\|(\mathcal{D}_{x_*} F)^{-1}\|}{1 - \eta} \le \frac{\beta}{1/2} = 2\beta.$$

Die Wohldefiniertheit des Newton-Verfahrens (5.14) folgt dann aus der Abschätzung (5.16) zusammen mit der folgenden Aussage

$$x_n \in \mathcal{B}(x_*; \delta), \qquad n = 0, 1, \ldots, \tag{5.17}$$

die nun mit vollständiger Induktion nachgewiesen wird; nebenbei werden sich dann auch die Abschätzungen (5.15) ergeben.

Nach Voraussetzung gilt $x_0 \in \mathcal{B}(x_*;\delta)$, und für ein $n \in \mathbb{N}_0$ sei nun bereits $x_n \in \mathcal{B}(x_*;\delta)$ gezeigt. Wegen (5.16) ist dann $\mathcal{D}_{x_n}F$ invertierbar und x_{n+1} somit wohldefiniert, und es gilt

$$x_{n+1} \;=\; x_n - (\mathcal{D}_{x_n}F)^{-1}(F(x_n)) \;=\; x_n - (\mathcal{D}_{x_n}F)^{-1}(F(x_n) - F(x_*))$$

beziehungsweise (unter Anwendung von Lemma 5.11)

$$\begin{aligned} x_{n+1} - x_* &= x_n - x_* - (\mathcal{D}_{x_n}F)^{-1}(F(x_n) - F(x_*)) \\ &= (\mathcal{D}_{x_n}F)^{-1}\big(F(x_*) - F(x_n) - (\mathcal{D}_{x_n}F)(x_* - x_n)\big); \\ \|x_{n+1} - x_*\| &= \|\text{———————— « ————————}\| \\ &\le 2\beta\tfrac{L}{2}\|x_n - x_*\|^2 = \beta L\underbrace{\|x_n - x_*\|}_{\le 1/(2\beta L)}^{2} \le \tfrac{1}{2}\|x_n - x_*\|, \end{aligned}$$

woraus $x_{n+1} \in \mathcal{B}(x_*;\delta)$ folgt, und der vorhergehenden Zeile entnimmt man auch noch die Abschätzungen (5.15), was den Beweis von Theorem 5.13 komplettiert. □

5.4.3 Nullstellenbestimmung bei Polynomen

Für Polynome liefert das (eindimensionale) Newton-Verfahren unter günstigen Umständen die größte Nullstelle:

Theorem 5.14. *Gegeben sei ein reelles Polynom $p(x) \in \Pi_r$, das eine reelle Nullstelle λ_1 besitze, so dass $\lambda_1 \ge \operatorname{Re}\xi$ für jede andere Nullstelle $\xi \in \mathbb{C}$ von p gilt.*[6] *Dann sind für jeden Startwert $x_0 > \lambda_1$ die Iterierten des Newton-Verfahrens*

$$x_{n+1} \;=\; x_n - \frac{p(x_n)}{p'(x_n)}, \qquad n \;=\; 0, 1, \ldots,$$

streng monoton fallend, und

$$|x_n - \lambda_1| \;\to\; 0 \qquad \textit{für} \quad n \to \infty.$$

BEWEIS. Es bezeichne $\lambda_1 \ge \lambda_2 \ge \ldots \ge \lambda_\ell$ die reellen Nullstellen sowie $\xi_1, \overline{\xi}_1, \ldots, \xi_m, \overline{\xi}_m$ (mit $\ell + 2m = r$) die komplexen Nullstellen des Polynoms p, das o.B.d.A. den führenden Koeffizienten eins besitze. Ganz allgemein erhält man mit den Wurzeln η_k eines Polynoms $q \in \Pi_r$ mit führendem Koeffizienten eins die folgenden Darstellungen für q und q',

$$q(x) \;=\; \prod_{k=1}^{r}(x - \eta_k) \qquad\qquad q'(x) \;=\; \sum_{k=1}^{r}\prod_{\substack{j=1\\ j\ne k}}^{r}(x - \eta_j) \;=\; \Big(\sum_{k=1}^{r}\frac{1}{x-\eta_k}\Big)q(x),$$

[6] Hier bezeichnet wieder $\operatorname{Re} z$ den Realteil einer komplexen Zahl $z \in \mathbb{C}$.

und somit gilt in der vorliegenden Situation

$$\begin{aligned} p(x) &= \prod_{k=1}^{\ell} (x-\lambda_k) \prod_{j=1}^{m} (x-\xi_j)(x-\overline{\xi}_j), \\ p'(x) &= \Big(\sum_{k=1}^{\ell} \frac{1}{x-\lambda_k} + 2 \sum_{j=1}^{m} \frac{x-\operatorname{Re}\xi_j}{(x-\xi_j)(x-\overline{\xi}_j)} \Big) p(x). \end{aligned} \tag{5.18}$$

Nun gilt für jedes $\xi \in \mathbb{C}\backslash\mathbb{R}$

$$\begin{aligned} (x-\xi)(x-\overline{\xi}) &= x^2 - 2x\operatorname{Re}\xi + |\xi|^2 > x^2 - 2x\operatorname{Re}\xi + (\operatorname{Re}\xi)^2 \\ &= (x-\operatorname{Re}\xi)^2 \geq 0, \qquad x \in \mathbb{R}, \end{aligned}$$

so dass in jedem Fall

$$p(x) > 0, \qquad p'(x) > 0 \qquad \text{für } x > \lambda_1$$

und damit

$$x - \frac{p(x)}{p'(x)} < x \qquad \text{für } x > \lambda_1$$

gilt. Andererseits gilt aber wegen der Darstellung (5.18) sowie wegen der Ungleichung

$$\Big(\sum_{k=1}^{\ell} \frac{1}{x-\lambda_k} + 2 \sum_{j=1}^{m} \frac{x-\operatorname{Re}\xi_j}{(x-\xi_j)(x-\overline{\xi}_j)} \Big) > \frac{1}{x-\lambda_1} \qquad \text{für } x > \lambda_1$$

auch

$$x - \frac{p(x)}{p'(x)} > \lambda_1 \qquad \text{für } x > \lambda_1.$$

Mittels vollständiger Induktion erschließt man, dass für einen Startwert $x_0 > \lambda_1$ das Newton-Verfahren eine streng monoton fallende Folge $x_1, x_2, \ldots$ mit $x_k > \lambda_1$ liefert, und dann liegt notwendigerweise Konvergenz vor mit einem Grenzwert, der als Fixpunkt der stetigen Iterationsabbildung (vergleiche den Beweis von Theorem 5.6) auch Nullstelle von p ist und somit mit λ_1 übereinstimmt. □

Beispiel 5.15. Als Beispiel sei ein Polynom $p \in \Pi_{11}$ betrachtet, dessen Nullstellen in der komplexen Ebene wie in Bild 5.2 verteilt seien.

Hier liefert das Newton-Verfahren für einen hinreichend großen Startwert näherungsweise die Nullstelle λ_1, und anschließende Anwendung des gleichen Verfahrens auf das *deflationierte* Polynom $p_1(x) = p(x)/(x-\lambda_1)$ liefert eine Näherung für die Nullstelle λ_2 (wobei als Startwert $x_0 = \lambda_1$ verwendet werden kann). Ganz analog lässt sich eine Approximation für λ_3 gewinnen. Theorem 5.14 liefert jedoch keine Aussage darüber, wie die Nullstellen λ_4 und λ_5 numerisch bestimmt werden können.

△

Für die praktische Umsetzung von Theorem 5.14 wird noch ein hinreichend großer Startwert benötigt. Das folgende Lemma liefert untere Schranken für mögliche Startwerte.

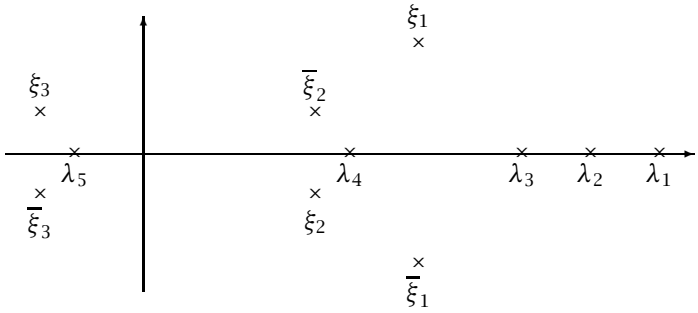


Bild 5.2: Beispiel für die Verteilung der Nullstellen eines Polynoms elften Grades in der komplexen Ebene

Lemma 5.16. *Gegeben sei das Polynom*

$$p(x) \;=\; a_0 \;+\; a_1 x \;+\; \ldots \;+\; a_{r-1} x^{r-1} \;+\; x^r,$$

und $\xi \in \mathbb{C}$ sei eine beliebige Nullstelle von $p(x)$.

(a) Es gelten die beiden Abschätzungen

$$|\xi| \;\le\; \max\big\{1,\; \sum_{k=0}^{r-1} |a_k|\big\}, \qquad\qquad |\xi| \;\le\; \max\big\{|a_0|,\; 1 + \max_{1\le k\le r-1} |a_k|\big\}.$$

(b) Im Fall $a_k \neq 0$ für $k = 1, \ldots, r-1$ gelten die beiden Abschätzungen

$$|\xi| \;\le\; \max\big\{\frac{|a_0|}{|a_1|},\; \max_{1\le k\le r-1} 2\frac{|a_k|}{|a_{k+1}|}\big\}, \qquad\qquad |\xi| \;\le\; \sum_{k=0}^{r-1} \frac{|a_k|}{|a_{k+1}|}.$$

(c) Schließlich gilt noch

$$|\xi| \;\le\; q^{1/r}, \qquad \textit{falls} \quad q \;:=\; \sum_{k=0}^{r-1} |a_k| \;<\; 1.$$

BEWEIS. Die *frobeniussche Begleitmatrix* zu dem Polynom p ist folgendermaßen definiert,

$$A \;:=\; \begin{pmatrix} 0 & & & -a_0 \\ 1 & \ddots & & \vdots \\ & \ddots & 0 & \vdots \\ & & 1 & -a_{r-1} \end{pmatrix} \;\in\; \mathbb{R}^{r\times r}.$$

Für das zugehörige charakteristische Polynom gilt die Identität

$$\det(\lambda I - A) \;=\; p(\lambda) \qquad \text{für} \quad \lambda \in \mathbb{C}, \tag{5.19}$$

wie im Folgenden nachgewiesen wird. Entwicklung der Determinante der Matrix $\lambda I - A$ nach der letzten Zeile liefert

$$\begin{aligned}\det(\lambda I - A) &= \det \begin{pmatrix} \lambda & & & a_0 \\ -1 & \ddots & & \vdots \\ & \ddots & \lambda & a_{r-2} \\ & & -1 & \lambda + a_{r-1} \end{pmatrix} \\ &= (\lambda + a_{r-1}) \underbrace{\det \begin{pmatrix} \lambda & & \\ -1 & \ddots & \\ & \ddots & \ddots \\ & & -1 \quad \lambda \end{pmatrix}}_{= \lambda^{r-1}} + \det \begin{pmatrix} \lambda & & & a_0 \\ -1 & \ddots & & \vdots \\ & \ddots & \lambda & \vdots \\ & & -1 & a_{r-2} \end{pmatrix},\end{aligned}$$

und erneute Entwicklung der auftretenden Determinanten nach jeweils der letzten Zeile liefert

$$\begin{aligned}\det \begin{pmatrix} \lambda & & & a_0 \\ -1 & \ddots & & \vdots \\ & \ddots & \lambda & \vdots \\ & & -1 & a_k \end{pmatrix} &= a_k \det \begin{pmatrix} \lambda & & \\ -1 & \ddots & \\ & \ddots & \ddots \\ & & -1 \quad \lambda \end{pmatrix} + \det \begin{pmatrix} \lambda & & & a_0 \\ -1 & \ddots & & \vdots \\ & \ddots & \lambda & \vdots \\ & & -1 & a_{k-1} \end{pmatrix} \\ &= a_k \lambda^k + \det \begin{pmatrix} \lambda & & a_0 \\ -1 & \ddots & \vdots \\ & \ddots & \lambda \quad \vdots \\ & & -1 \quad a_{k-1} \end{pmatrix},\end{aligned}$$

für $k = r-2,\ r-3, \ldots, 2$, und schließlich gilt

$$\det \begin{pmatrix} \lambda & a_0 \\ -1 & a_1 \end{pmatrix} = a_1 \lambda + a_0,$$

was den Beweis der Identität (5.19) komplettiert.

Aufgrund von (5.19) nun stimmt die Menge der Nullstellen des Polynoms p mit der Menge $\sigma(A)$ der Eigenwerte der Matrix A überein. Weiter gilt $r_\sigma(A) \le \|A\|$ für jede durch eine komplexe Vektornorm induzierte Matrixnorm, vergleiche Bemerkung 4.39, und wegen

$$\|A\|_1 = \max\Big\{1, \sum_{k=0}^{r-1} |a_k|\Big\}, \qquad \|A\|_\infty = \max\Big\{|a_0|,\ 1 + \max_{1 \le k \le r-1} |a_k|\Big\},$$

ergeben sich die Abschätzungen in (a). Für den Nachweis der Abschätzungen in (b) sei nun

$$D := \operatorname{diag}(a_1, \ldots, a_{r-1}, 1).$$

Die Matrix $D^{-1}AD \in \mathbb{R}^{r\times r}$ ist ähnlich zu der Matrix A, was $\sigma(D^{-1}AD) = \sigma(A)$ beziehungsweise $r_\sigma(D^{-1}AD) = r_\sigma(A)$ nach sich zieht. Weiter hat man die explizite Darstellung (es gilt $a_r = 1$)

$$D^{-1}AD = \begin{pmatrix} 0 & & & & -a_0/a_1 \\ a_1/a_2 & \ddots & & & -a_1/a_2 \\ & a_2/a_3 & \ddots & & -a_2/a_3 \\ & & \ddots & 0 & \vdots \\ & & & a_{r-1}/a_r & -a_{r-1}/a_r \end{pmatrix} \in \mathbb{R}^{r\times r},$$

so dass also die beiden Identitäten

$$\|D^{-1}AD\|_\infty = \max\Big\{\frac{|a_0|}{|a_1|}, \max_{1\le k\le r-1} 2\frac{|a_k|}{|a_{k+1}|}\Big\}, \qquad \|D^{-1}AD\|_1 = \sum_{k=0}^{r-1}\frac{|a_k|}{|a_{k+1}|}$$

gelten, und analog zu (a) ergeben sich die in (b) angegebenen Abschätzungen. Schließlich erhält man (c) folgendermaßen: wegen (a) ist in jedem Fall $|\xi| \le 1$ erfüllt, und weiter gilt für jede Zahl $x \in \mathbb{C}$ mit $q^{1/r} < |x| \le 1$ die Abschätzung

$$|p(x)| \ge |x|^r - \sum_{k=0}^{r-1}|a_k||x|^k > q - \sum_{k=0}^{r-1}|a_k| = 0,$$

so dass sogar $|\xi|^r \le q$ gilt. Dies komplettiert den Beweis des Lemmas. □

Eine Anwendung der vier Abschätzungen in (a) und (b) aus Lemma 5.16 auf einige spezielle Polynome liefert die in der folgenden Tabelle angegebenen Resultate.

$p(x)$	$\lvert\xi\rvert \le$			
$x^2 + 1 = (x-\mathrm{i})(x+\mathrm{i})$	1	1	–	–
$x^2 - 2x + 1 = (x-1)^2$	3	3	4	2.5

Weitere Themen und Literaturhinweise

Die numerische Lösung nichtlinearer Gleichungen wird ausführlich in Deuflhard [20] behandelt. Abschnitte über die numerische Lösung solcher Gleichungen findet man außerdem in jedem der im Literaturverzeichnis aufgeführten Lehrbücher über numerische Mathematik, beispielsweise in Deuflhard/Hohmann [22], Oevel [78], Schaback/Wendland [92] und in Werner [111]. Als eine Variante des in diesem Kapitel vorgestellten Newton-Verfahrens ist das *gedämpfte Newton-Verfahren*

$$x_{n+1} = x_n - \gamma_n(\mathcal{D}_{x_n}F)^{-1}(F(x_n)) \qquad \text{für } n = 0, 1, \ldots$$

zu nennen, mit einer der Konvergenzbeschleunigung dienenden und geeignet zu wählenden variablen Schrittweite γ_n. Eine weitere Variante des Newton-Verfahrens stellen die *Quasi-Newton-Verfahren* $x_{n+1} = x_n - A_n^{-1} F(x_n),\ n = 0, 1, \ldots$ dar, wobei die (numerisch aufwändig zu berechnenden) Jacobi-Matrizen $\mathcal{D}_{x_n} F$ durch einfacher zu gewinnende Matrizen $A_n \approx \mathcal{D}_{x_n} F$ ersetzt werden. Einzelheiten zu den beiden genannten Varianten werden beispielsweise in [20] beziehungsweise in Freund/Hoppe [31], Geiger/Kanzow [32], Großmann/Terno [44], Kosmol [62], Mennicken/Wagenführer [71], Nash/Sofer [75], Schwetlick [95] sowie in Aufgabe 5.6 vorgestellt. Weitere Varianten wie das *Sekantenverfahren* beruhen auf Approximationen der Ableitungen durch Differenzenquotienten.

Übungsaufgaben

Aufgabe 5.1. Gegeben sei die Gleichung

$$x + \ln x = 0,$$

deren eindeutige Lösung x_* im Intervall $[0.5, 0.6]$ liegt. Zur approximativen Lösung dieser Gleichung betrachte man die folgenden fünf Iterationsverfahren:

$$x_{n+1} := -\ln x_n, \qquad x_{n+1} := e^{-x_n}, \qquad x_{n+1} := (x_n + e^{-x_n})/2, \tag{5.20}$$

$$x_{n+1} := \frac{a x_n + e^{-x_n}}{a+1}, \qquad x_{n+1} := \frac{a_n x_n + e^{-x_n}}{a_n + 1}. \tag{5.21}$$

Welche der drei in (5.20) angegebenen Verfahren sind brauchbar? Man bestimme in (5.21) Werte $a \in \mathbb{R}$ beziehungsweise $a_0, a_1, \ldots \in \mathbb{R}$ so dass sich jeweils ein Verfahren von mindestens zweiter Ordnung ergibt.

Aufgabe 5.2. Die Funktion $\ln(x)$ soll an der Stelle $x = a > 0$ näherungsweise berechnet werden. Dies kann beispielsweise mit dem Newton-Verfahren zur Bestimmung einer Nullstelle der Funktion

$$f(x) = e^x - a$$

geschehen. Man gebe die zugehörige Iterationsvorschrift an und weise quadratische Konvergenz nach. Kann man die Konvergenzordnung $p = 3$ erwarten? Schließlich berechne man für $a = 1$ und Startwert $x_0 = 1$ die ersten vier Iterierten $x_1, \ldots, x_4$. Auf wie viele Nachkommastellen genau stimmen diese mit dem tatsächlichen Wert $0 = \ln(1)$ überein?

Aufgabe 5.3. Zu einer kontraktiven Funktion $\Phi : \mathbb{R}^N \to \mathbb{R}^N$ mit Kontraktionskonstante $0 < L < 1$ bezeichne $x_* \in \mathbb{R}^N$ den Fixpunkt von Φ, und der Vektor $x_0 \in \mathbb{R}^N$ sei beliebig. Die Folge $(x_n^\delta)_{n \in \mathbb{N}_0}$ sei gegeben durch

$$\begin{aligned} x_0^\delta &:= x_0 + \Delta x_0, \\ x_{n+1}^\delta &:= \Phi(x_n^\delta) + \Delta x_{n+1}, \qquad n = 0, 1, \ldots, \end{aligned}$$

wobei $\| \Delta x_n \| \le \delta$ für $n \in \mathbb{N}_0$ gelte bezüglich einer gegebenen Vektornorm $\| \cdot \| : \mathbb{R}^N \to \mathbb{R}$ und einer gewissen Fehlerschranke δ. Man zeige Folgendes:

$$\| x_n^\delta - x_* \| \le \frac{\delta}{1-L} + \frac{L^n}{1-L}\big((L+2)\delta + \| x_1^\delta - x_0^\delta \| \big), \qquad n = 0, 1, \ldots .$$

Aufgabe 5.4. Es sei die Abbildung $\Phi : \mathbb{R}^2 \to \mathbb{R}^2$ definiert durch

$$\Phi\begin{pmatrix} x \\ y \end{pmatrix} \;=\; \frac{1}{2}\begin{pmatrix} 1 \;+\; \dfrac{\sin x}{4} \;+\; y \\ 1 \;+\; \sin y \;+\; x \end{pmatrix}.$$

(a) Man untersuche die Kontraktionseigenschaft von Φ jeweils bezüglich $\|\cdot\|_\infty$ und $\|\cdot\|_2$.
(b) Man berechne den Fixpunkt $(\xi,\eta)^\top \in \mathbb{R}^2$ der Abbildung Φ mittels der gewöhnlichen Fixpunktiteration, für den Startwert $(x_0, y_0)^\top = (0,0)^\top$. Wie oft ist bei Verwendung der a priori-Fehlerabschätzung zu iterieren, bis

$$\|(x_n, y_n)^\top - (\xi,\eta)^\top\|_2 \;\le\; 10^{-2}$$

garantiert werden kann? Die entsprechende Frage stellt sich bei Anwendung der a posteriori-Fehlerabschätzung.

Aufgabe 5.5. Gegeben sei das nichtlineare Gleichungssystem

$$\left.\begin{aligned} uv \;+\; u - v - 1 \;&=\; 0, \\ uv \;&=\; 0. \end{aligned}\right\} \tag{5.22}$$

(a) Man bestimme die exakten Lösungen des nichtlinearen Gleichungssystems (5.22).
(b) Für die Startwerte

$$x_0 \;=\; \begin{pmatrix} 0 \\ 0 \end{pmatrix} \quad \text{und} \quad x_0 \;=\; \begin{pmatrix} 1 \\ 1 \end{pmatrix}$$

führe man jeweils den ersten Iterationsschritt des Newton-Verfahrens durch.

Aufgabe 5.6. Für eine reguläre Matrix $A \in \mathbb{R}^{N\times N}$ ist die inverse Matrix $X = A^{-1}$ offensichtlich eine Lösung der nichtlinearen Gleichung

$$X^{-1} - A \;=\; 0. \tag{5.23}$$

Das Newton-Verfahren zur Lösung der Gleichung (5.23) führt auf das *Verfahren von Schulz*

$$X_{n+1} \;:=\; X_n \;+\; X_n(I \;-\; AX_n), \qquad n \;=\; 0, 1, \ldots .$$

Man zeige: für jede Startmatrix $X_0 \in \mathbb{R}^{N\times N}$ mit $\|I - AX_0\| \le q < 1$ (mit einer gegebenen submultiplikativen Matrixnorm $\|\cdot\| : \mathbb{R}^{N\times N} \to \mathbb{R}$) konvergiert die Matrixfolge $X_0, X_1, \ldots \subset \mathbb{R}^{N\times N}$ gegen die Matrix A^{-1} mit den Abschätzungen

$$\|X_n - A^{-1}\| \;\le\; \frac{\|X_0\|}{1-q}\|I - AX_n\| \;\le\; \frac{\|X_0\|}{1-q}q^{(2^n)} \qquad \text{für} \quad n \;=\; 0, 1, \ldots .$$

Aufgabe 5.7 (*Numerische Aufgabe*). Man schreibe einen Code zur Lösung eines nichtlinearen Gleichungssystems mittels der folgenden Variante des Newton-Verfahrens:

$$x_{n+1} \;=\; x_n \;-\; A_n F(x_n) \qquad \text{für} \quad n \;=\; 0, 1, \ldots,$$

mit

$$A_{kp+j} \;=\; (\mathcal{D}_{x_{kp}} F)^{-1} \qquad \text{für} \quad \begin{aligned} j \;&=\; 0, 1, \ldots, p-1, \\ k \;&=\; 0, 1, \ldots . \end{aligned}$$

Hierbei bezeichnet $\mathcal{D}_x F$ die Jacobi-Matrix der Abbildung F im Punkt x. Man breche die Iteration ab, falls die Bedingung $\| x_n - x_{n-1} \|_2 \leq \texttt{tol}$ erstmalig erfüllt ist oder falls $n = n_{\max}$ gilt. Hier sind $p \in \mathbb{N}$, $n_{\max} \in \mathbb{N}_0$ und $\texttt{tol} > 0$ frei wählbare Parameter.

Man teste das Programm anhand des Beispiels

$$F\binom{u}{v} := \begin{pmatrix} \sin(u)\cos(v) \\ u^2 + v^2 - 3 \end{pmatrix} = \binom{0}{0},$$

mit den Parametern $\texttt{tol} = 10^{-4}$ und $n_{\max} = 100$ sowie mit den folgenden Startwerten beziehungsweise den folgenden Werten von p:

(a) $x_0 = \binom{1}{1}$, $p = 1$; (b) $x_0 = \binom{1}{1}$, $p = 5$;

(c) $x_0 = \binom{3}{3}$, $p = 1$; (d) $x_0 = \binom{3}{3}$, $p = 5$.

Aufgabe 5.8. Die Funktion $f \in C^1[a,b]$ sei streng monoton wachsend und konvex mit Nullstelle $x_* \in [a,b]$. Man zeige, dass für jeden Startwert $x_0 \in [x_*, b]$ die Näherungen x_n des Newton-Verfahrens gegen x_* konvergieren mit

$$x_{n+1} \leq x_n, \qquad n = 0, 1, \ldots .$$

6 Numerische Integration von Funktionen

Zahlreiche Anwendungen wie etwa die Bestimmung von Flächen oder Normalverteilungen führen letztlich auf das Problem der Berechnung von Integralen

$$\mathcal{I}(f) := \int_a^b f(x)\,dx \tag{6.1}$$

mit gewissen Funktionen $f \in C[a,b]$. Oftmals ist jedoch die Berechnung des Integrals (6.1) nicht möglich, da beispielsweise die Stammfunktion von f nicht berechnet werden kann oder die Funktionswerte von f als Resultat von Messungen nur an endlich vielen Stellen vorliegen.

Beispiel 6.1. Die Preise von Kaufoptionen auf europäischen Finanzmärkten lassen sich unter gewissen vereinfachenden Annahmen (zum Beispiel konstanten Volatilitäten) mit der Black-Scholes-Formel explizit angeben. Für Details sei auf Günther/Jüngel [45] oder Hanke-Bourgeois [52] verwiesen. In unserem Zusammenhang ist von Interesse, dass dabei Auswertungen der Fehlerfunktion

$$\operatorname{erf}(x) \;=\; \frac{2}{\sqrt{\pi}} \int_0^x \exp(-t^2)\,dt \qquad \text{für} \quad x \geq 0$$

erforderlich sind. Deren Werte lassen sich jedoch lediglich näherungsweise bestimmen. △

Man ist an einfachen Methoden zur näherungsweisen Berechnung des Integrals (6.1) interessiert, und hierzu werden im Folgenden *Quadraturformeln*

$$\mathcal{I}_n(f) \;=\; (b-a) \sum_{k=0}^{n} \sigma_k f(x_k), \tag{6.2}$$

herangezogen mit paarweise verschiedenen Stützstellen $x_0, x_1, \ldots, x_n \in [a,b]$ und reellen Gewichten $\sigma_0, \sigma_1, \ldots, \sigma_n \in \mathbb{R}$.

Definition 6.2. Die Zahl $r \in \mathbb{N}_0$ heißt *Genauigkeitsgrad* der Quadraturformel $\mathcal{I}_n$, wenn

$$\begin{aligned} \mathcal{I}_n(x^m) &= \mathcal{I}(x^m) \qquad \text{für} \quad m = 0, 1, \ldots, r, \\ \mathcal{I}_n(x^{r+1}) &\neq \mathcal{I}(x^{r+1}) \end{aligned} \tag{6.3}$$

erfüllt ist. Der Genauigkeitsgrad einer Quadraturformel $\mathcal{I}_n$ ist per Definition *mindestens* $r \in \mathbb{N}_0$, falls (6.3) gilt.

Bemerkung 6.3. (a) $\mathcal{I}_n : C[a,b] \to \mathbb{R}$ ist offensichtlich eine lineare Abbildung, es gilt also

$$\mathcal{I}_n(\alpha f + \beta g) \;=\; \alpha \mathcal{I}_n(f) \,+\, \beta \mathcal{I}_n(g) \qquad \forall\, f,\, g \in C[a,b], \qquad \alpha,\, \beta \in \mathbb{R}.$$

(b) Wegen der Linearität der Quadraturformel $\mathcal{I}_n$ und des Integrals $\mathcal{I}$ gilt:

$\mathcal{I}_n$ besitzt den Genauigkeitsgrad r

$$\iff \quad \begin{cases} \mathcal{I}_n(\mathcal{P}) \;=\; \mathcal{I}(\mathcal{P}) \text{ für alle Polynome } \mathcal{P} \text{ vom Grad } \le r, \text{ und} \\ \mathcal{I}_n(\mathcal{P}) \;\neq\; \mathcal{I}(\mathcal{P}) \text{ für } \textit{ein} \text{ Polynom } \mathcal{P} \text{ vom (genauen) Grad } = r+1 \end{cases}$$

$$\iff \quad \begin{cases} \mathcal{I}_n(\mathcal{P}) \;=\; \mathcal{I}(\mathcal{P}) \text{ für alle Polynome } \mathcal{P} \text{ vom Grad } \le r, \text{ und} \\ \mathcal{I}_n(\mathcal{P}) \;\neq\; \mathcal{I}_n(\mathcal{P}) \text{ für } \textit{alle} \text{ Polynom } \mathcal{P} \text{ vom (genauen) Grad } = r+1 \end{cases}$$

△

6.1 Interpolatorische Quadraturformeln

Definition 6.4. *Interpolatorische Quadraturformeln* $\mathcal{I}_n(f)$ sind folgendermaßen erklärt: nach einer Festlegung von $n \in \mathbb{N}_0$ sowie $(n+1)$ paarweise verschiedenen Stützstellen $x_0, x_1, \ldots, x_n \in [a,b]$ wird als Näherung für $\mathcal{I}(f)$ der Wert

$$\mathcal{I}_n(f) \;:=\; \int_a^b \mathcal{Q}_n(x)\,dx$$

herangezogen, wobei $\mathcal{Q}_n \in \Pi_n$ das interpolierende Polynom zu den Stützpunkten $(x_0, f(x_0)), (x_1, f(x_1)), \ldots, (x_n, f(x_n)) \in \mathbb{R}^2$ bezeichnet.

Bemerkung 6.5. Der Genauigkeitsgrad einer interpolatorischen Quadraturformel $\mathcal{I}_n$ ist offensichtlich mindestens n. △

Im Folgenden soll eine explizite Darstellung für $\mathcal{I}_n(f)$ hergeleitet werden. Daraus resultiert dann auch die Darstellung (6.2) für die Quadraturformel $\mathcal{I}_n(f)$ aus Definition 6.4.

Theorem 6.6. *Eine interpolatorische Quadraturformel $\mathcal{I}_n$ besitzt die Gestalt*

$$\mathcal{I}_n(f) \;=\; (b-a)\sum_{k=0}^{n} \sigma_k f(x_k) \quad \textit{mit} \quad \sigma_k \;:=\; \int_0^1 \prod_{\substack{m=0 \\ m\neq k}}^{n} \frac{t-t_m}{t_k-t_m}\,dt, \quad t_m \;:=\; \frac{x_m-a}{b-a}. \tag{6.4}$$

BEWEIS. Mit der lagrangeschen Interpolationsformel

$$\mathcal{Q}_n \;=\; \sum_{k=0}^{n} f(x_k) L_k \qquad \text{mit} \qquad L_k(x) \;=\; \prod_{\substack{m=0 \\ m\neq k}}^{n} \frac{x-x_m}{x_k-x_m}$$

erhält man $\mathcal{I}_n(f) = \sum_{k=0}^n f(x_k) \int_a^b L_k(x)\,dx$, und aus der nachfolgenden Rechnung resultiert dann die Aussage des Theorems,

$$\frac{1}{b-a}\int_a^b L_k(x)\,dx \;=\; \frac{1}{b-a}\int_a^b \prod_{\substack{m=0 \\ m\neq k}}^{n} \frac{x-x_m}{x_k-x_m}\,dx \;\overset{(*)}{=}\; \int_0^1 \prod_{\substack{m=0 \\ m\neq k}}^{n} \frac{t-t_m}{t_k-t_m}\,dt \;=\; \sigma_k,$$

wobei man die Identität $(*)$ mit der Substitution $x = (b-a)t + a$ erhält. □

Bemerkung 6.7. (a) Der Vorteil in der Darstellung (6.4) ist in der Unabhängigkeit der Gewichte σ_k sowohl von den Intervallgrenzen a und b als auch von der Funktion f begründet. Letztlich hängen die Gewichte nur von der relativen Verteilung der Stützstellen im Intervall $[a,b]$ ab.

(b) Für jede interpolatorische Quadraturformel $\mathcal{I}_n(f) = (b-a)\sum_{k=0}^n \sigma_k f(x_k)$ gilt

$$\sum_{k=0}^{n} \sigma_k \;=\; 1, \tag{6.5}$$

da ihr Genauigkeitsgrad mindestens $n \geq 0$ beträgt und somit $(b-a)\sum_{k=0}^n \sigma_k = \mathcal{I}_n(\mathbf{1}) = \mathcal{I}(\mathbf{1}) = b-a$ gilt. △

6.2 Spezielle interpolatorische Quadraturformeln

6.2.1 Abgeschlossene Newton-Cotes-Formeln

Die *Newton-Cotes-Formeln* ergeben sich durch die Wahl äquidistanter Stützstellen bei interpolatorischen Quadraturformeln. Wenn zusätzlich Intervallanfang und -ende Stützstellen sind, also $x_0 = a$, $x_n = b$ gilt, so spricht man von *abgeschlossenen Newton-Cotes-Formeln*. Speziell gilt hier also (für $n \geq 1$)

$$x_k \;:=\; a \;+\; kh, \qquad k \;=\; 0,1,\ldots,n, \qquad\qquad h \;=\; \frac{b-a}{n}.$$

Lemma 6.8. *Für die Gewichte $\sigma_0,\sigma_1,\ldots,\sigma_n$ der abgeschlossenen Newton-Cotes-Formeln gilt*

$$\sigma_k \;=\; \frac{1}{n}\int_0^n \prod_{\substack{m=0\\ m\neq k}}^{n} \frac{s-m}{k-m}\,ds \qquad \textit{für}\;\; k \;=\; 0,1,\ldots,n. \tag{6.6}$$

BEWEIS. Aus der Identität (6.4) erhält man aufgrund von $t_k = k/n$ für die Gewichte die angegebene Darstellung,

$$\sigma_k \;=\; \int_0^1 \prod_{\substack{m=0\\ m\neq k}}^{n} \frac{t-m/n}{(k-m)/n}\,dt \;=\; \frac{1}{n}\int_0^n \prod_{\substack{m=0\\ m\neq k}}^{n} \frac{s-m}{k-m}\,ds,$$

wobei man die zweite Gleichung aus der Substitution $t = s/n$ erhält. □

Die Darstellung (6.6) und die folgende Symmetrieeigenschaft der Gewichte der abgeschlossenen Newton-Cotes-Formeln ermöglichen die in den nachfolgenden Beispielen angestellten einfachen Berechnungen.

Lemma 6.9. *Für die Gewichte $\sigma_0,\sigma_1,\ldots,\sigma_n$ der abgeschlossenen Newton-Cotes-Formeln gilt*

$$\sigma_{n-k} \;=\; \sigma_k \qquad \textit{für}\;\; k \;=\; 0,1,\ldots,n. \tag{6.7}$$

BEWEIS. Für die lagrangeschen Basispolynome L_k gilt

$$L_{n-k}(x) \;=\; L_k(b+a-x), \qquad x \in [a,b], \tag{6.8}$$

denn $L_{n-k} \in \Pi_n$ und $\mathcal{Q}(x) := L_k(b+a-x) \in \Pi_n$, und

$$\begin{aligned} \mathcal{Q}(x_{n-j}) \;&=\; L_k\big(b+a-\big(a+(n-j)\tfrac{b-a}{n}\big)\big) \;=\; L_k\big(a+j\tfrac{b-a}{n}\big) \\ &=\; L_k(x_j) \;=\; \delta_{kj} \;=\; L_{n-k}(x_{n-j}) \qquad \text{für} \quad j \;=\; 0,1,\ldots,n, \end{aligned}$$

und die Eindeutigkeit des interpolierenden Polynoms resultiert in der Identität (6.8). Daraus erhält man

$$\begin{aligned} \sigma_{n-k} \;&=\; \tfrac{1}{b-a}\int_a^b L_{n-k}(x)\,dx \;=\; \tfrac{1}{b-a}\int_a^b L_k(b+a-x)\,dx \\ &\overset{(*)}{=}\; \tfrac{1}{b-a}\int_a^b L_k(t)\,dt \;=\; \sigma_k, \end{aligned}$$

wobei man $(*)$ mit der Substitution $x = b+a-t$ erhält. □

Beispiel 6.10. (a) Für $n = 1$ erhält man die *Trapezregel*,

$$\mathcal{I}_1(f) \;=\; (b-a)\frac{f(a)\;+\;f(b)}{2} \;\approx\; \int_a^b f(x)\,dx,$$

denn (6.5) und (6.7) liefern $\sigma_0 + \sigma_1 = 1$ und $\sigma_0 = \sigma_1$, somit $\sigma_0 \;=\; \sigma_1 \;=\; \frac{1}{2}$.

(b) Für $n = 2$ erhält man die *Simpson-Regel*

$$\mathcal{I}_2(f) \;=\; (b-a)\tfrac{1}{6}\big(f(a)\;+\;4f\big(\tfrac{a+b}{2}\big)\;+\;f(b)\big) \;\approx\; \int_a^b f(x)\,dx,$$

denn die Eigenschaften (6.5)–(6.7) ergeben Folgendes,

$$\sigma_0 \;=\; \tfrac{1}{2}\int_0^2 \tfrac{s-1}{0-1}\tfrac{s-2}{0-2}\,ds \;=\; \tfrac{1}{6}, \qquad \sigma_2 \;=\; \sigma_0, \qquad \sigma_1 \;=\; 1-\sigma_0-\sigma_2 \;=\; \tfrac{2}{3}.$$

Die geometrische Bedeutung der Trapez- und der Simpson-Regel ist in Bild 6.1 beziehungsweise Bild 6.2 dargestellt.

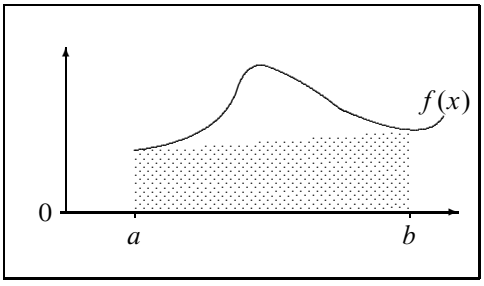


Bild 6.1: Illustration zur Trapezregel

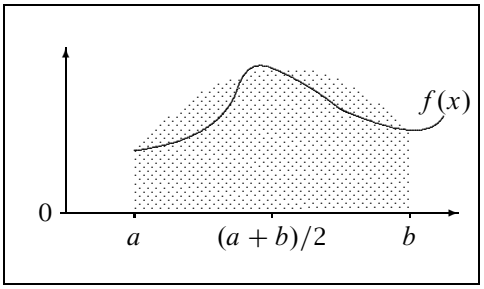


Bild 6.2: Illustration zur Simpson-Regel

(c) Der Fall $n = 3$ führt auf die *Newtonsche 3/8-Regel*

$$\mathcal{I}_3(f) \;=\; (b-a)\tfrac{1}{8}\big(f(a) \;+\; 3f\big(\tfrac{2a+b}{3}\big) \;+\; 3f\big(\tfrac{a+2b}{3}\big) \;+\; f(b)\big) \;\approx\; \int_a^b f(x)\,dx.$$

(d) In der Situation $n = 4$ erhält man die *Milne-Regel*

$$\begin{aligned}\mathcal{I}_4(f) \;&=\; \tfrac{b-a}{90}\big(7f(a) + 32f\big(\tfrac{3a+b}{4}\big) + 12f\big(\tfrac{2(a+b)}{4}\big) + 32f\big(\tfrac{a+3b}{4}\big) + 7f(b)\big)\\ &\approx \int_a^b f(x)\,dx.\end{aligned}$$

(e) Der Fall $n = 8$ liefert die folgende Quadraturformel,

$$\begin{aligned}\mathcal{I}_8(f) \;=\; \tfrac{b-a}{28350}\big(&989f(x_0) \;+\; 5888f(x_1) \;-\; 928f(x_2) \;+\; 10496f(x_3) \;-\; 4540f(x_4)\\ &+\; 10496f(x_5) \;-\; 928f(x_6) \;+\; 5888f(x_7) \;+\; 989f(x_8)\big)\\ \approx\; &\int_a^b f(x)\,dx.\end{aligned}$$

△

Zu der zuletzt betrachteten Quadraturformel $\mathcal{I}_8(f)$ ist Folgendes anzumerken:

- Es treten negative Gewichte auf, wie überhaupt für $n \geq 8$ bei den abgeschlossenen Newton-Cotes-Formeln. Dies widerspricht der Vorstellung des Integrals als Grenzwert einer Summe von Funktionswerten mit positiven Gewichten.
- Die Summe der Beträge der Gewichte übersteigt den Wert eins, was zu einer Verstärkung von Rundungsfehlern führt. Es gilt das folgende Theorem, das hier ohne Beweis angegeben wird.

Theorem 6.11 (Satz von Kusmin). *Die Gewichte $\sigma_0^{(n)}, \sigma_1^{(n)}, \ldots, \sigma_n^{(n)}$ der abgeschlossenen Newton-Cotes-Formeln $\mathcal{I}_n$ besitzen die Eigenschaft*

$$\sum_{k=0}^{n} |\sigma_k^{(n)}| \;\to\; \infty \qquad \textit{für} \quad n \to \infty.$$

Aus den beiden genannten Gründen werden abgeschlossene Newton-Cotes-Formeln nur für kleine Werte von n angewandt.

6.2.2 Andere interpolatorische Quadraturformeln

Beispiel 6.12. • Eine Rechteckregel lautet $\mathcal{I}_0(f) = (b-a)f(a)$ (hier ist $n = 0$ und $x_0 = a$), und eine weitere Rechteckregel ist $\mathcal{I}_0(f) = (b-a)f(b)$ (hier ist $n = 0$ und $x_0 = b$).

• Die Mittelpunktregel ist von der Form $\mathcal{I}_0(f) = (b-a)f\big(\frac{a+b}{2}\big)$ (hier ist $n = 0$ und $x_0 = (a+b)/2$).

Die geometrische Bedeutung der ersten Rechteck- und der Mittelpunktregel ist in Bild 6.3 beziehungsweise Bild 6.4 dargestellt. △

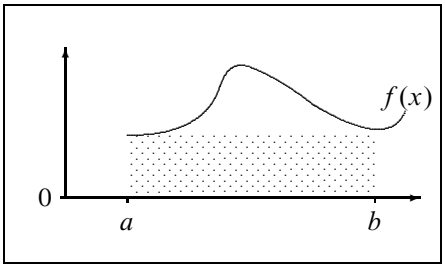


Bild 6.3: Darstellung der Rechteckregel

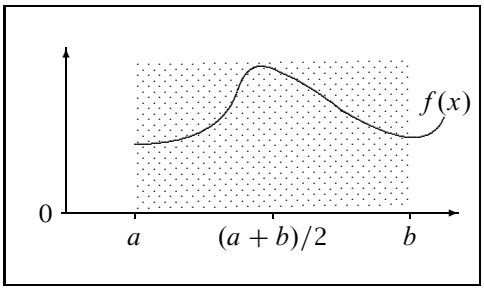


Bild 6.4: Darstellung der Mittelpunktregel

6.3 Der Fehler bei der interpolatorischen Quadratur

Im Folgenden wird eine Abschätzung für den bei der interpolatorischen Quadratur auftretenden Fehler vorgestellt. Insbesondere wird dabei deutlich, dass die interpolatorischen Quadraturformeln lediglich für kurze Intervalle $[a,b]$ (also für $b-a \ll 1$) gute Näherungen an das zu bestimmende Integral darstellen.

Vorbereitend wird noch folgende Sprechweise eingeführt: eine reellwertige Funktion ψ heißt *von einem Vorzeichen* auf dem Intervall $[c,d]$, wenn (sie dort definiert ist und) $\psi(x) \geq 0$ für alle $x \in [c,d]$ oder $\psi(x) \leq 0$ für alle $x \in [c,d]$ gilt.

Theorem 6.13. *Die interpolatorische Quadraturformel $\mathcal{I}_n(f) = (b-a)\sum_{k=0}^{n} \sigma_k f(x_k)$ besitze mindestens den Genauigkeitsgrad $r \geq n$, und die Funktion $f : [a,b] \to \mathbb{R}$ sei $(r+1)$-mal stetig differenzierbar. Dann gilt die folgende Fehlerabschätzung,*

$$|\mathcal{I}(f) - \mathcal{I}_n(f)| \;\leq\; c_r \frac{(b-a)^{r+2}}{(r+1)!} \max_{\xi \in [a,b]} |f^{(r+1)}(\xi)| \tag{6.9}$$

mit

$$c_r \;:=\; \min_{t_{n+1},\ldots,t_r \in [0,1]} \int_0^1 \prod_{k=0}^{r} |t - t_k|\, dt, \qquad t_k \;:=\; \frac{x_k - a}{b-a}, \qquad k \;=\; 0,1,\ldots,n. \tag{6.10}$$

Wenn mit den Werten $t_0, t_1, \ldots, t_n$ aus (6.10) für eine bestimmte Wahl von $t_{n+1}, \ldots, t_r \in [0,1]$ das Produkt $\prod_{k=0}^{r}(t - t_k)$ von einem Vorzeichen in $[0,1]$ ist, so gilt mit einer Zwischenstelle $\xi \in [a,b]$ die folgende Fehlerdarstellung,

$$\mathcal{I}(f) - \mathcal{I}_n(f) \;=\; c_r' \frac{(b-a)^{r+2}}{(r+1)!} f^{(r+1)}(\xi) \tag{6.11}$$

$$\text{mit} \quad c_r' \;:=\; \int_0^1 \prod_{k=0}^{r} (t - t_k)\, dt. \tag{6.12}$$

BEWEIS. 1. Seien $x_{n+1}, \ldots, x_r \in [a,b]$ beliebig aber so, dass $x_0, x_1, \ldots, x_r$ paarweise verschieden sind. Es soll in diesem ersten Teil des Beweises die unten stehende

Fehlerdarstellung (6.15) nachgewiesen werden. Sei dazu $\mathcal{Q}_r \in \Pi_r$ das zu den Stützpunkten $(x_0, f(x_0)), \ldots, (x_r, f(x_r))$ gehörende interpolierende Polynom. Aufgrund der Darstellung (6.2) für $\mathcal{I}_n$ erhält man

$$\mathcal{I}_n(f) \;=\; (b-a)\sum_{k=0}^{n} \sigma_k f(x_k) \;=\; (b-a)\sum_{k=0}^{n} \sigma_k \mathcal{Q}_r(x_k) \;=\; \mathcal{I}_n(\mathcal{Q}_r) \;=\; \mathcal{I}(\mathcal{Q}_r),$$

und somit

$$\mathcal{I}(f) - \mathcal{I}_n(f) \;=\; \mathcal{I}(f) - \mathcal{I}(\mathcal{Q}_r) \;=\; \int_a^b f(x) - \mathcal{Q}_r(x)\,dx. \tag{6.13}$$

Weiter gilt (siehe Theorem 1.17 auf Seite 11)

$$f(x) - \mathcal{Q}_r(x) \;=\; \frac{(\omega\,\nu)(x)\, f^{(r+1)}(\xi(x))}{(r+1)!}, \qquad x \in [a,b], \tag{6.14}$$

mit

$$\omega(x) \;:=\; (x-x_0)\cdots(x-x_n), \qquad \nu(x) \;:=\; (x-x_{n+1})\cdots(x-x_r),$$

und einer geeigneten Zwischenstellenfunktion $\xi : [a,b] \to [a,b]$. Man beachte, dass die rechte Seite der Gleichung (6.14) als Differenz zweier stetiger Funktionen selbst stetig und damit integrierbar ist. Weiter sei noch angemerkt, dass ω bereits durch die Quadraturformel festgelegt ist, während die Nullstellen von ν noch variieren können.

Aus (6.13) und (6.14) erhält man

$$\mathcal{I}(f) - \mathcal{I}_n(f) \;=\; \frac{1}{(r+1)!}\int_a^b (\omega\,\nu)(x)\, f^{(r+1)}(\xi(x))\,dx. \tag{6.15}$$

2. Es soll nun die Fehlerabschätzung (6.9) bewiesen werden, und hierzu seien x_{n+1}, $\ldots, x_r \in [a,b]$ beliebig. Dann wählt man Zahlen

$$x_{n+1}^{(m)}, \ldots, x_r^{(m)} \;\in\; [a,b], \qquad m \;=\; 1, 2, \ldots,$$

so dass Folgendes gilt,

$$\begin{aligned} &x_0, x_1, \ldots, x_n,\ x_{n+1}^{(m)}, \ldots, x_r^{(m)} \quad \text{paarweise verschieden,} \\ &x_k^{(m)} \to x_k \qquad \text{für} \quad m \to \infty \qquad (k \;=\; n+1, \ldots, r). \end{aligned}$$

Mit der Notation

$$\nu_m(x) \;=\; \prod_{k=n+1}^{r} (x - x_k^{(m)})$$

erhält man aus der Identität (6.15) angewandt mit $\nu = \nu_m$ sowie einem anschließenden Grenzübergang $m \to \infty$ Folgendes:

$$\begin{aligned} &|\mathcal{I}(f) - \mathcal{I}_n(f)| \\ &\;\le\; \frac{1}{(r+1)!} \max_{\xi\in[a,b]} |f^{(r+1)}(\xi)| \int_a^b |(\omega\,\nu_m)(x)|\,dx \\ &\;\le\; \text{———————} \ll \text{———————} \Big(\int_a^b |(\omega\,\nu)(x)|\,dx + \underbrace{\int_a^b |\omega(x)|\,|\nu_m(x) - \nu(x)|\,dx}_{\to\ 0 \text{ für } m\to\infty} \Big), \end{aligned}$$

wobei die Konvergenz des zweiten Terms aus der auf dem Intervall $[a,b]$ vorliegenden gleichmäßigen Konvergenz $\nu_m \to \nu$ für $m \to \infty$ resultiert. Somit erhält man

$$|\mathcal{I}(f) - \mathcal{I}_n(f)| \;\le\; \widehat{c}_r \frac{1}{(r+1)!} \max_{x\in[a,b]} |f^{(r+1)}(x)|$$

mit

$$\begin{aligned} \widehat{c}_r &\overset{(*)}{:=} \min_{x_{n+1},\dots,x_r \in [a,b]} \int_a^b \prod_{k=0}^{r} |x - x_k|\, dx \\ &\overset{(**)}{=} (b-a)^{r+2} \min_{t_{n+1},\dots,t_r \in [0,1]} \int_0^1 \prod_{k=0}^{r} |t - t_k|\, dt, \end{aligned}$$

wobei das Minimum in der Setzung $(*)$ aus Stetigkeitsgründen tatsächlich existiert, und $(**)$ resultiert aus der Substitution $x = (b-a)t + a$. Die Abschätzung (6.9) ist damit nachgewiesen.

3. Für den Nachweis von (6.11) betrachte man die Zahlen $x_k = (b-a)t_k + a$ für $k = n+1,\dots,r$, so dass entsprechend der Voraussetzung die Funktion $\omega\,\nu$ auf dem Intervall $[a,b]$ von einem Vorzeichen ist, etwa

$$(\omega\,\nu)(x) \;\ge\; 0, \qquad x \in [a,b].$$

Eine dem zweiten Teil dieses Beweises entsprechende Vorgehensweise liefert

$$\begin{aligned} \mathcal{I}(f) - \mathcal{I}_n(f) \;\le\; \frac{1}{(r+1)!}\Big(& \max_{\xi\in[a,b]} f^{(r+1)}(\xi) \int_a^b (\omega\,\nu)(x)\, dx \\ & + \max_{\xi\in[a,b]} |f^{(r+1)}(\xi)| \overbrace{\int_a^b |\omega(x)|\,|\nu_m(x) - \nu(x)|\, dx}^{\to\ 0 \text{ für } m\to\infty} \Big) \\ \to\; \frac{1}{(r+1)!} & \max_{\xi\in[a,b]} f^{(r+1)}(\xi) \int_a^b (\omega\,\nu)(x)\, dx \qquad \text{für } m \to \infty, \end{aligned}$$

und analog folgt

$$\mathcal{I}(f) - \mathcal{I}_n(f) \;\ge\; \frac{1}{(r+1)!} \min_{\xi\in[a,b]} f^{(r+1)}(\xi) \int_a^b (\omega\,\nu)(x)\, dx.$$

Die Anwendung des Zwischenwertsatzes auf die stetige Funktion $f^{(r+1)}$ liefert eine Zwischenstelle $\xi \in [a,b]$ mit

$$\mathcal{I}(f) - \mathcal{I}_n(f) \;=\; \frac{1}{(r+1)!} f^{(r+1)}(\xi) \int_a^b (\omega\,\nu)(x)\, dx, \tag{6.16}$$

und eine abschließende Substitution $x = (b-a)t + a$ ergibt die Identität (6.11). □

Beispiel 6.14. 1. (Rechteckregeln) Für $f \in C^1[a,b]$ gelten die Fehlerdarstellungen

$$\int_a^b f(x)\, dx - (b-a) f(a) \;=\; \frac{(b-a)^2}{2} f'(\xi_0), \tag{6.17}$$

$$\int_a^b f(x)\, dx - (b-a) f(b) \;=\; -\frac{(b-a)^2}{2} f'(\xi_1) \tag{6.18}$$

mit gewissen Zwischenstellen $\xi_0,\ \xi_1 \in [a,b]$. Die Darstellung (6.17) beispielsweise erhält man aus Theorem 6.13 angewandt mit $n = r = 0$ und $x_0 = a$ beziehungsweise $t_0 = 0$ unter Berücksichtigung von

$$\prod_{k=0}^{0}(t-t_k) \;=\; t \;\geq\; 0 \qquad \text{für} \quad 0 \leq t \leq 1, \qquad\qquad c_0' \;=\; \int_0^1 t\,dt \;=\; \frac{t^2}{2}\Big|_{t=0}^{t=1} \;=\; \frac{1}{2}.$$

Analog leitet man die Darstellung (6.18) her.

2. (Trapezregel) In diesem Fall gilt für $f \in C^2[a,b]$

$$\mathcal{I}(f) - \mathcal{I}_1(f) \;=\; -\frac{(b-a)^3}{12} f''(\xi) \tag{6.19}$$

mit einer Zwischenstelle $\xi \in [a,b]$. Dies folgt aus Theorem 6.13 angewandt mit $n = r = 1,\ x_0 = a,\ x_1 = b$ beziehungsweise $t_0 = 0,\ t_1 = 1$ unter Berücksichtigung von

$$\begin{aligned}\prod_{k=0}^{1}(t-t_k) \;&=\; t(t-1) \;\leq\; 0 \qquad \text{für} \quad 0 \leq t \leq 1,\\ c_1' \;&=\; \int_0^1 t(t-1)\,dt \;=\; \frac{t^3}{3} - \frac{t^2}{2}\Big|_{t=0}^{t=1} \;=\; -\frac{1}{6}.\end{aligned}$$

△

In dem vorangegangenen Beispiel wurde für $n = 0$ sowie für $n = 1$ verwendet, dass $\mathcal{I}_n$ jeweils mindestens den Genauigkeitsgrad $r = n$ besitzt. Analog kann man natürlich bei der Simpson-Regel (hier ist $n = 2$) vorgehen. Dort kann man sich jedoch zu Nutze machen, dass in diesem Fall der Genauigkeitsgrad $r = 3$ vorliegt, was im folgenden Abschnitt für eine allgemeinere Situation nachgewiesen wird.

6.4 Der Genauigkeitsgrad abgeschlossener Newton-Cotes-Formeln $\mathcal{I}_n$ für gerade Zahlen n

Das folgende Lemma wird für den Beweis von Theorem 6.16 benötigt, das die wesentliche Aussage dieses Abschnitts 6.4 darstellt.

Lemma 6.15. *Sei $n \in \mathbb{N}$ gerade, $h = (b-a)/n$, und $x_k = a + kh$ für $k = 0, 1, \ldots, n$. Für die Funktion*

$$F(x) \;:=\; \int_a^x \prod_{k=0}^{n}(y - x_k)\,dy, \qquad x \in [a,b], \tag{6.20}$$

gilt

$$F(a) \;=\; F(b) \;=\; 0, \qquad\qquad F(x) \;>\; 0 \qquad \textit{für} \quad a < x < b. \tag{6.21}$$

Der Beweis von Lemma 6.15 wird am Ende von Abschnitt 6.4 nachgetragen.

Theorem 6.16. *Die abgeschlossenen Newton-Cotes-Formeln $\mathcal{I}_n$ besitzen für gerades $n \geq 2$ den Genauigkeitsgrad $r = n + 1$.*

BEWEIS. Er gliedert sich in zwei Teile.

1. Offensichtlich ist der Genauigkeitsgrad von $\mathcal{I}_n$ mindestens n, siehe Bemerkung 6.5. Des Weiteren gilt $\mathcal{I}((x-(a+b)/2)^{n+1}) = 0$, denn der Integrand ist eine ungerade Funktion bezüglich des Intervallmittelpunkts[1] $(a+b)/2$. Im Folgenden wird

$$\mathcal{I}_n\big(\big(x - \tfrac{a+b}{2}\big)^{n+1}\big) \;=\; 0 \tag{6.22}$$

nachgewiesen, woraus sich dann unmittelbar ergibt, dass der Genauigkeitsgrad von $\mathcal{I}_n$ mindestens $r = n+1$ beträgt. Für den Nachweis von (6.22) setzt man $h = (b-a)/n$ und $x_k = a + kh$ für $k = 0, 1, \ldots, n$, so dass dann Folgendes gilt,

$$\begin{aligned} x_{n/2} \;&=\; \tfrac{a+b}{2} \;=\; a \,+\, \tfrac{n}{2}h, \\ x_{n-k} - \tfrac{a+b}{2} \;&=\; -\big(x_k - \tfrac{a+b}{2}\big), \qquad k \;=\; 0, 1, \ldots, \tfrac{n}{2} - 1. \end{aligned}$$

Aufgrund der Symmetrieeigenschaft $\sigma_{n-k} = \sigma_k$ für $k = 0, 1, \ldots, n$ (siehe (6.7)) erhält man daher

$$\begin{aligned} &\mathcal{I}_n\big(\big(x - \tfrac{a+b}{2}\big)^{n+1}\big) \\ &= (b-a)\Big(\sum_{k=0}^{n/2-1} \sigma_k \big(\big(x_k - \tfrac{a+b}{2}\big)^{n+1} + \big(x_{n-k} - \tfrac{a+b}{2}\big)^{n+1}\big) + \sigma_{n/2}\big(x_{n/2} - \tfrac{a+b}{2}\big)^{n+1}\Big) \\ &= (b-a)\Big(\sum_{k=0}^{n/2-1} \sigma_k \cdot 0 \;+\; \sigma_{n/2} \cdot 0\Big) \;=\; 0, \end{aligned}$$

was gerade die Aussage (6.22) darstellt.

2. Im Folgenden wird

$$\mathcal{I}_n(x^{n+2}) \;\neq\; \mathcal{I}(x^{n+2}) \tag{6.23}$$

nachgewiesen, woraus sich zusammen mit dem ersten Teil des Beweises die Aussage des Theorems über den Genauigkeitsgrad von $\mathcal{I}_n$ ergibt. Für den Nachweis von (6.23) betrachtet man für das Monom $f(x) = x^{n+2}$ und für eine beliebige Zahl $x_{n+1} \in [a, b]$ mit $x_{n+1} \neq x_k$ für $k = 0, 1, \ldots, n$ die Fehlerformel (6.15) und integriert anschließend

[1] Eine Erläuterung der Bezeichnung "ungerade bezüglich des Intervallmittelpunkts" findet sich im Beweisteil 2) von Lemma 6.15.

partiell:

$$\mathcal{I}(x^{n+2}) - \mathcal{I}_n(x^{n+2}) = \frac{1}{(n+2)!}\int_a^b \Big[\prod_{k=0}^{n+1}(x-x_k)\Big]\underbrace{\Big(\frac{d^{n+2}}{dx^{n+2}}x^{n+2}\Big)(\xi(x))}_{\equiv (n+2)!}\,dx$$

$$= \int_a^b \prod_{k=0}^{n+1}(x-x_k)\,dx = \int_a^b F'(x)(x-x_{n+1})\,dx \qquad (F \text{ wie in (6.20)})$$

$$= F(x)(x-x_{n+1})\big|_{x=a}^{x=b} - \int_a^b F(x)\Big(\frac{d}{dx}(x-x_{n+1})\Big)\,dx$$

$$\overset{(6.21)}{=} 0 - 0 - \int_a^b F(x)\cdot 1\,dx = -\int_a^b F(x)\,dx \overset{(6.21)}{\neq} 0.$$

Dies komplettiert den Beweis des Theorems. □

Beispiel 6.17 (Simpson-Regel). Hier gilt für $f \in C^4[a,b]$ die Fehlerdarstellung

$$\int_a^b f(x)\,dx - \frac{b-a}{6}\Big(f(a) + 4f\Big(\frac{a+b}{2}\Big) + f(b)\Big) = -\frac{(b-a)^5}{2880}f^{(4)}(\xi) \qquad (6.24)$$

mit einer Zwischenstelle $\xi \in [a,b]$, was aus Theorem 6.13 angewandt mit $r=3$, $n=2$, $x_0=a$, $x_1=(a+b)/2$, $x_2=b$ beziehungsweise $t_0=0$, $t_1=1/2$, $t_2=1$ resultiert. Für die Wahl $t_3=1/2$ erhält man nämlich (bezüglich der Notation siehe wieder Theorem 6.13)

$$\prod_{k=0}^{3}(t-t_k) = t\,(t-\tfrac{1}{2})^2(t-1) \le 0 \qquad \text{für } t \in [0,1],$$

$$c_3' = \int_0^1 t(t-\tfrac{1}{2})^2(t-1)\,dt = -\frac{1}{120},$$

und mit Theorem 6.13 ergibt sich die in (6.24) angegebene Fehlerdarstellung,

$$\mathcal{I}(f) - \mathcal{I}_2(f) = -\frac{(b-a)^5}{4!}\frac{1}{120}f^{(4)}(\xi) = -\frac{(b-a)^5}{2880}f^{(4)}(\xi). \qquad \triangle$$

6.4.1 Der Beweis von Lemma 6.15

Die Identität $F(a)=0$ ist offensichtlich richtig, und für den Nachweis der weiteren Aussagen des Lemmas sei der Integrand in (6.20) wie folgt bezeichnet,

$$\omega(y) = \prod_{k=0}^{n}(y-x_k), \qquad y \in \mathbb{R}.$$

1) Es wird im Folgenden die Positivität der Funktion F auf der linken Hälfte des Intervalls $[a,b]$ nachgewiesen,

$$F(x) = \int_a^x \omega(y)\,dy > 0, \qquad a < x \le \frac{a+b}{2}. \qquad (6.25)$$

Vorbereitendes hierzu wird in 1a)–1b) hergeleitet.

1a) Das Polynom ω mit genauem Grad $n+1$ besitzt die paarweise verschiedenen Nullstellen $x_0, x_1, \ldots, x_n$. Wegen $\omega(y) \to -\infty$ für $y \to -\infty$ (da ω ungeraden Grad besitzt) gilt also

$$\begin{aligned} \omega(y) &< 0 \quad \text{für} \quad y < a, \\ \omega(a+\tau) &> 0 \quad \text{für} \quad 0 < \tau < h, \\ \omega(x_1+\tau) &< 0 \quad \text{für} \quad 0 < \tau < h, \\ &\vdots \quad \vdots \quad \vdots \end{aligned}$$

siehe Bild 6.5 für eine Darstellung des Verlaufs der Funktion ω.

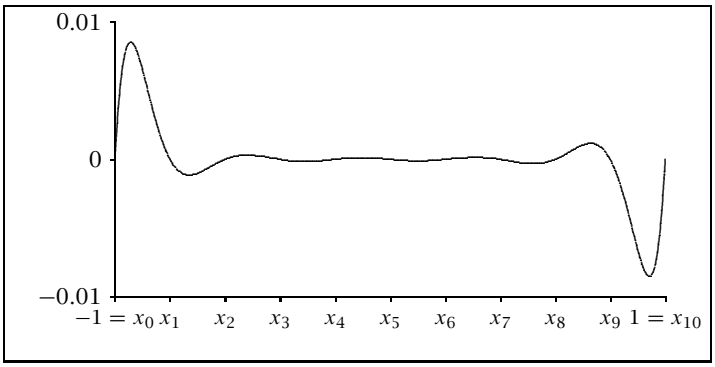


Bild 6.5: Beispiel für den Verlauf der Funktion ω

Allgemein gilt

$$\begin{aligned} \omega(x_{2j}+\tau) &> 0, \\ \omega(x_{2j+1}+\tau) &< 0 \quad \text{für} \quad 0 < \tau < h, \qquad j = 0, 1, \ldots, \frac{n}{2} - 1. \end{aligned} \tag{6.26}$$

1b) Weiter gilt

$$|\omega(y+h)| < |\omega(y)| \quad \text{für} \quad a \le y \le \frac{a+b}{2} - h, \qquad y \notin \{x_0, \ldots, x_{n/2-1}\}, \tag{6.27}$$

denn

$$\frac{\omega(y+h)}{\omega(y)} = \frac{\prod_{k=0}^{n}(y+h-x_k)}{\prod_{k=0}^{n}(y-x_k)} = \frac{(y+h-a)\prod_{k=0}^{n-1}(y-x_k)}{(y-b)\prod_{k=0}^{n-1}(y-x_k)} = \frac{y+h-a}{y-b},$$

und wegen der Annahmen in (6.27) gilt

$$|y+h-a| < \frac{b-a}{2}, \qquad |y-b| > \frac{b-a}{2}.$$

1c) Man erhält nun schließlich die in (6.25) angegebene Positivität der Funktion F: mit der Eigenschaft (6.26) erhält man unmittelbar

$$\int_{x_{2j}}^{x_{2j}+\tau} \omega(y)\, dy > 0, \qquad 0 < \tau \le h \qquad \text{mit} \quad 0 \le j \le \frac{n/2-1}{2}, \tag{6.28}$$

und die Abschätzung (6.27) liefert Folgendes,

$$\int_{x_{2j}}^{x_{2j+1}+\tau} \omega(y)\,dy \;=\; \int_{x_{2j}}^{x_{2j+\tau}} \overbrace{\omega(y) + \underbrace{\omega(y+h)}_{=-|\omega(y+h)|}}^{>0}\,dy \;+\; \overbrace{\int_{x_{2j+\tau}}^{x_{2j+1}} \omega(y)\,dy}^{\geq 0} \;>\; 0,$$
$$0 < \tau \leq h \qquad \big(0 \leq j \leq \tfrac{n/2-2}{2}\big).$$

2) Schließlich soll nachgewiesen werden, dass die Funktion F gerade bezüglich des Intervallmittelpunkts ist,

$$F\big(\tfrac{a+b}{2} + \tau\big) \;=\; F\big(\tfrac{a+b}{2} - \tau\big) \qquad \text{für} \quad 0 \leq \tau \leq \tfrac{b-a}{2}. \tag{6.29}$$

Daraus resultiert dann aufgrund von $F(a) = 0$ unmittelbar $F(b) = 0$, und (6.25) zieht die Ungleichung $F(x) > 0$ für $(a+b)/2 \leq x < b$ nach sich. Dies komplettiert den Nachweis der Aussagen in (6.21).

2a) Für den Beweis der Identität (6.29) wird benötigt, dass die Funktion ω ungerade bezüglich des Intervallmittelpunkts $(a+b)/2 = x_{n/2}$ ist: wegen $\frac{a+b}{2} - x_k = -(\frac{a+b}{2} - x_{n-k})$ für $k = 0, 1, \ldots, n$ gilt nämlich

$$\omega\big(\tfrac{a+b}{2} + y\big) \;=\; \prod_{k=0}^{n} \big(\tfrac{a+b}{2} + y - x_k\big) \;=\; -\prod_{k=0}^{n} \big(\tfrac{a+b}{2} - y - x_{n-k}\big)$$
$$\overset{(*)}{=} -\prod_{k=0}^{n} \big(\tfrac{a+b}{2} - y - x_k\big) \;=\; -\omega\big(\tfrac{a+b}{2} - y\big) \qquad \text{für} \quad 0 \leq y \leq \tfrac{b-a}{2},$$

wobei man $(*)$ mit der Indextransformation $k \to n-k$ erhält.

2b) Mit 2a) folgt schließlich die Identität (6.29):

$$F\big(\tfrac{a+b}{2} + \tau\big) \;=\; \int_0^{(a+b)/2-\tau} \omega(x)\,dx \;+\; \int_{(a+b)/2-\tau}^{(a+b)/2+\tau} \omega(x)\,dx$$
$$= \; F\big(\tfrac{a+b}{2} - \tau\big) + 0. \qquad \square$$

6.5 Summierte Quadraturformeln

Zur numerischen Berechnung des Integrals $\mathcal{I}(f) = \int_a^b f(x)\,dx$ kann man beispielsweise das Intervall $[a,b]$ mit Stützstellen

$$x_k \;=\; a + kh \qquad \text{für} \quad k = 0, 1, \ldots, N \qquad \big(h = \tfrac{b-a}{N}\big) \tag{6.30}$$

versehen und die bisher betrachteten Quadraturformeln zur numerischen Berechnung der Integrale

$$\int_{x_{k-1}}^{x_k} f(x)\,dx, \qquad k = 1, 2, \ldots, N$$

verwenden. Die Resultate werden schließlich aufsummiert, und die so gewonnenen Formeln bezeichnet man als *summierte Quadraturformeln*. Im Folgenden werden einige Beispiele und die jeweils zugehörigen Fehlerdarstellungen vorgestellt.

6.5.1 Summierte Rechteckregeln

Zwei Rechteckregeln sind in Beispiel 6.12 vorgestellt worden. Die *summierten Rechteckregeln* mit den äquidistanten Stützstellen aus (6.30) lauten dann entsprechend

$$\mathcal{T}_0(h) \;=\; h\sum_{k=0}^{N-1} f(x_k) \;\approx\; \int_a^b f(x)\,dx, \tag{6.31}$$

$$\widehat{\mathcal{T}}_0(h) \;=\; h\sum_{k=1}^{N} f(x_k) \quad \text{——— « ———}. \tag{6.32}$$

Die geometrische Bedeutung der summierten Rechteckregel (6.31) ist in Bild 6.6 dargestellt. Ihre approximativen Eigenschaften sind in dem nachfolgenden Theorem festgehalten.

Theorem 6.18. *Die Funktion $f : [a,b] \to \mathbb{R}$ sei einmal stetig differenzierbar auf dem Intervall $[a,b]$. Dann gibt es Zwischenstellen $\xi,\ \widehat{\xi} \in [a,b]$ mit*

$$\int_a^b f(x)\,dx \;-\; \mathcal{T}_0(h) \;=\; \frac{b-a}{2}\,h\,f'(\xi), \tag{6.33}$$

$$\text{——— « ———} \;-\; \widehat{\mathcal{T}}_0(h) \;=\; -\frac{b-a}{2}\,h\,f'(\widehat{\xi}), \tag{6.34}$$

mit $h = (b-a)/N$, und mit $\mathcal{T}_0(h)$ und $\widehat{\mathcal{T}}_0(h)$ wie in (6.31) beziehungsweise (6.32).

BEWEIS. Es wird hier nur die Fehlerdarstellung (6.33) betrachtet, den Nachweis für (6.34) führt man ganz analog. Für $\mathcal{T}_0(h)$ liefert Beispiel 6.14 die Existenz einer Zwischenstelle $\xi_k \in [a,b]$ mit

$$\int_{x_{k-1}}^{x_k} f(x)\,dx \;-\; hf(x_{k-1}) \;=\; \frac{h^2}{2}f'(\xi_k), \qquad k \;=\; 1,2,\ldots,N,$$

und Summation über k liefert

$$\int_a^b f(x)\,dx \;-\; \mathcal{T}_0(h) \;=\; \sum_{k=1}^{N} \frac{h^2}{2} f'(\xi_k) \;=\; \frac{b-a}{2}\,h\,\frac{1}{N}\sum_{k=1}^{N} f'(\xi_k).$$

Aufgrund der Ungleichungen

$$\min_{x\in[a,b]} f'(x) \;\le\; \frac{1}{N}\sum_{k=1}^{N} f'(\xi_k) \;\le\; \max_{x\in[a,b]} f'(x)$$

existiert nach Anwendung des Zwischenwertsatzes auf die Funktion f' eine Zwischenstelle $\xi \in [a,b]$ mit

$$f'(\xi) \;=\; \frac{1}{N}\sum_{k=1}^{N} f'(\xi_k),$$

was die Fehlerdarstellung (6.33) liefert. □

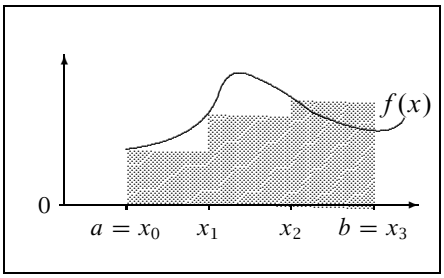


Bild 6.6: Summierte Rechteckregel

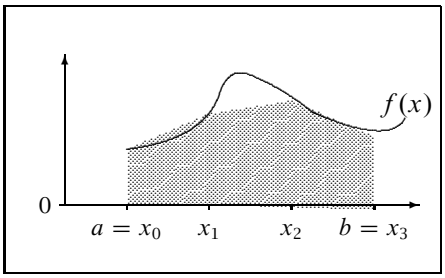


Bild 6.7: Summierte Trapezregel

6.5.2 Summierte Trapezregel

Die von der (in Beispiel 6.10 definierten) Trapezregel abgeleitete *summierte Trapezregel* mit den Stützstellen aus (6.30) lautet

$$\mathcal{T}_1(h) \;=\; \frac{h}{2}\big(f(a) \,+\, 2\sum_{k=1}^{N-1} f(x_k) \,+\, f(b)\big) \;\approx\; \int_a^b f(x)\,dx. \qquad (6.35)$$

Die geometrische Bedeutung der summierten Trapezregel (6.35) ist in Bild 6.7 veranschaulicht. Das nachfolgende Theorem liefert eine Fehlerdarstellung für diese summierte Quadraturformel.

Theorem 6.19. *Die Funktion $f : [a,b] \to \mathbb{R}$ sei auf dem Intervall $[a,b]$ zweimal stetig differenzierbar. Dann gibt es eine Zwischenstelle $\xi \in [a,b]$ mit*

$$\int_a^b f(x)\,dx \,-\, \mathcal{T}_1(h) \;=\; -\frac{b-a}{12}h^2 f''(\xi),$$

mit $h = (b-a)/N$ und $\mathcal{T}_1(h)$ wie in (6.35).

BEWEIS. Der Beweis verläuft entsprechend dem Beweis von Theorem 6.18: es gibt (siehe Beispiel 6.14) Zwischenstellen $\xi_k \in [a,b]$ mit

$$\int_{x_{k-1}}^{x_k} f(x)\,dx \;-\; \frac{h}{2}[f(x_{k-1}) \,+\, f(x_k)] \;=\; -\frac{h^3}{12}f''(\xi_k), \qquad k \;=\; 1,2,\ldots,N,$$

und Summation über k liefert

$$\begin{aligned}\int_a^b f(x)\,dx \,-\, \mathcal{T}_1(h) \;&=\; -\sum_{k=1}^{N}\frac{h^3}{12}f''(\xi_k) \;=\; -\frac{b-a}{12}h^2\,\frac{1}{N}\sum_{k=1}^{N} f''(\xi_k)\\ &=\; -\frac{b-a}{12}h^2 f''(\xi)\end{aligned}$$

für eine Zwischenstelle $\xi \in [a,b]$, wobei man die Existenz einer solchen Zwischenstelle durch Anwendung des Zwischenwertsatzes auf die Funktion f'' erhält. □

6.5.3 Summierte Simpson-Regel

Die von der (in Beispiel 6.10 vorgestellten) Simpson-Regel abgeleitete *summierte Simpson-Regel* lautet

$$\mathcal{T}_2(h) \;=\; \frac{h}{6}\big(f(a) \;+\; 4\sum_{k=1}^{N} f(x_{k-1/2}) \;+\; 2\sum_{k=1}^{N-1} f(x_k) \;+\; f(b)\big) \;\approx\; \int_a^b f(x)\,dx, \tag{6.36}$$

mit den äquidistanten Stützstellen $x_k = a + kh,\ k \geq 0$, und mit $h = (b-a)/N$. Das nachfolgende Theorem liefert eine Fehlerdarstellung für die summierte Simpson-Regel.

Theorem 6.20. *Die Funktion $f : [a,b] \to \mathbb{R}$ sei auf dem Intervall $[a,b]$ viermal stetig differenzierbar. Dann gibt es eine Zwischenstelle $\xi \in [a,b]$ mit*

$$\int_a^b f(x)\,dx \;-\; \mathcal{T}_2(h) \;=\; -\tfrac{b-a}{2880} h^4 f^{(4)}(\xi),$$

mit $h = (b-a)/N$ und $\mathcal{T}_2(h)$ wie in (6.36).

BEWEIS. Der Beweis verläuft wiederum entsprechend dem Beweis von Theorem 6.18. Für $k = 1, 2, \ldots, N$ gibt es (siehe Beispiel 6.17) Zwischenstellen $\xi_k \in [x_{k-1}, x_k]$ mit

$$\int_{x_{k-1}}^{x_k} f(x)\,dx \;-\; \frac{h}{6}\big[\, f(x_{k-1}) \;+\; 4f(x_{k-1/2}) \;+\; f(x_k)\big] \;=\; -\frac{h^5}{2880} f^{(4)}(\xi_k),$$

und Summation über k liefert

$$\begin{aligned}\int_a^b f(x)\,dx \;-\; \mathcal{T}_2(h) \;&=\; -\sum_{k=1}^{N} \frac{h^5}{2880} f^{(4)}(\xi_k) \;=\; -\frac{b-a}{2880} h^4 \frac{1}{N} \sum_{k=1}^{N} f^{(4)}(\xi_k) \\ &=\; -\frac{b-a}{2880} h^4 f^{(4)}(\xi)\end{aligned}$$

für eine Zwischenstelle $\xi \in [a,b]$, wobei man die Existenz einer solchen Zwischenstelle durch Anwendung des Zwischenwertsatzes auf die Funktion $f^{(4)}$ erhält. □

Bemerkung 6.21. Zwar ist die Zahl der erforderlichen Funktionsaufrufe bei der summierten Simpson-Regel doppelt so hoch wie bei den summierten Rechteckregeln oder der summierten Trapezregel. Für hinreichend glatte Funktionen f ist die Anwendung der summierten Simpson-Regel dennoch vorzuziehen, da sich beispielsweise gegenüber der summierten Trapezregel die Genauigkeit quadriert. △

6.6 Asymptotik der summierten Trapezregel

In dem vorliegenden Abschnitt 6.6 wird für die summierte Trapezregel (6.35) eine asymptotische Entwicklung vorgestellt, die beim Einsatz von Extrapolationsverfahren (siehe Abschnitt 6.7) Gewinn bringend eingesetzt werden kann.

6.6.1 Die Asymptotik

Für die summierte Trapezregel $\mathcal{T}_1(h)$ aus (6.35) wird im folgenden Theorem eine asymptotische Entwicklung angegeben, die gewisse Ähnlichkeiten mit einer Taylorentwicklung von $\mathcal{T}_1$ im Punkt $h = 0$ aufweist. (Man beachte jedoch, dass $\mathcal{T}_1(h)$ nur für diskrete positive Werte von h definiert ist.)

Theorem 6.22. *Sei $f \in C^{2r+2}[a,b]$, $r \geq 0$. Für die summierte Trapezregel*

$$\mathcal{T}_1(h) \;=\; \frac{h}{2}\big(f(a) \;+\; 2\sum_{k=1}^{N-1} f(x_k) \;+\; f(b)\big) \quad\approx\quad \int_a^b f(x)\,dx \qquad \big(h \;=\; \frac{b-a}{N}\big)$$

(vergleiche (6.35)) gilt die folgende Darstellung:

$$\mathcal{T}_1(h) \;=\; \tau_0 \;+\; \tau_1 h^2 \;+\; \ldots \;+\; \tau_r h^{2r} \;+\; R_{r+1}(h), \tag{6.37}$$

mit

$$\tau_0 \;=\; \int_a^b f(x)\,dx, \qquad R_{r+1}(h) \;=\; \mathcal{O}(h^{2r+2}) \quad \textit{für} \quad h \to 0, \tag{6.38}$$

und gewissen Koeffizienten $\tau_1, \tau_2, \ldots, \tau_r \in \mathbb{R}$.

BEWEIS. Siehe Abschnitt 6.9. □

Es fällt auf, dass in (6.37) Terme mit ungeraden Potenzen von h nicht auftreten, was man sich zu Nutze machen kann. Mehr hierzu finden Sie in dem nachfolgenden Abschnitt 6.7 über Extrapolationsmethoden.

6.7 Extrapolationsverfahren

6.7.1 Grundidee

Der vorliegende Abschnitt über Extrapolationsverfahren lässt sich inhaltlich Kapitel 1 über die Polynominterpolation zuordnen. Er wird erst hier präsentiert, da mit der vorgestellten Asymptotik der summierten Trapezregel nun eine spezielle Anwendung vorliegt.

Für eine gegebene Funktion[2] $\mathcal{T}(h)$, $h > 0$, liege mit gewissen Koeffizienten $\tau_0, \tau_1, \ldots, \tau_r \in \mathbb{R}$ das folgende asymptotische Verhalten vor,

$$\mathcal{T}(h) \;=\; \tau_0 \;+\; \tau_1 h^\gamma \;+\; \tau_2 h^{2\gamma} \;+\; \ldots \;+\; \tau_r h^{r\gamma} \;+\; \mathcal{O}(h^{(r+1)\gamma}) \qquad \text{für} \quad h \to 0, \tag{6.39}$$

mit einer Zahl $\gamma > 0$ und dem gesuchten Wert $\tau_0 = \lim_{h\to 0+} \mathcal{T}(h)$. Für eine Nullfolge positiver, paarweiser verschiedener Schrittweiten h sei $\mathcal{T}(h)$ bestimmbar.

Wegen (6.39) gilt zunächst nur

$$\mathcal{T}(h) \;=\; \tau_0 \;+\; \mathcal{O}(h^\gamma) \qquad \text{für} \quad h \to 0.$$

[2] die typischerweise ein numerisches Verfahren repräsentiert, das zu zulässigen Diskretisierungsparametern h jeweils eine Approximation für eine gesuchte Größe $\tau_0 \in \mathbb{R}$ liefert

Mithilfe des im Folgenden vorzustellenden Extrapolationsverfahrens erhält man ohne großen Mehraufwand genauere Approximationen an die gesuchte Größe τ_0 (siehe Theorem 6.26 unten). Der Ansatz des Extrapolationsverfahrens ist folgender: zu ausgewählten positiven Stützstellen $h_0, h_1, \ldots, h_n$ wird das eindeutig bestimmte Polynom $\mathcal{P}_{0,\ldots,n} \in \Pi_n$ mit

$$\mathcal{P}_{0,\ldots,n}(h_j^\gamma) \;=\; \mathcal{T}(h_j), \qquad j \;=\; 0, 1, \ldots, n,$$

herangezogen[3] und der Wert

$$\mathcal{P}_{0,\ldots,n}(0) \;\approx\; \mathcal{T}(0)$$

als Approximation für $\mathcal{T}(0)$ verwendet. Im Zusammenhang mit der summierten Trapezregel wird diese Vorgehensweise als *Romberg-Integration* bezeichnet und geht auf Romberg [87] zurück.

Beispiel 6.23. Die prinzipielle Vorgehensweise bei der Extrapolation ist für $n = 3$ in Bild 6.8 dargestellt. △

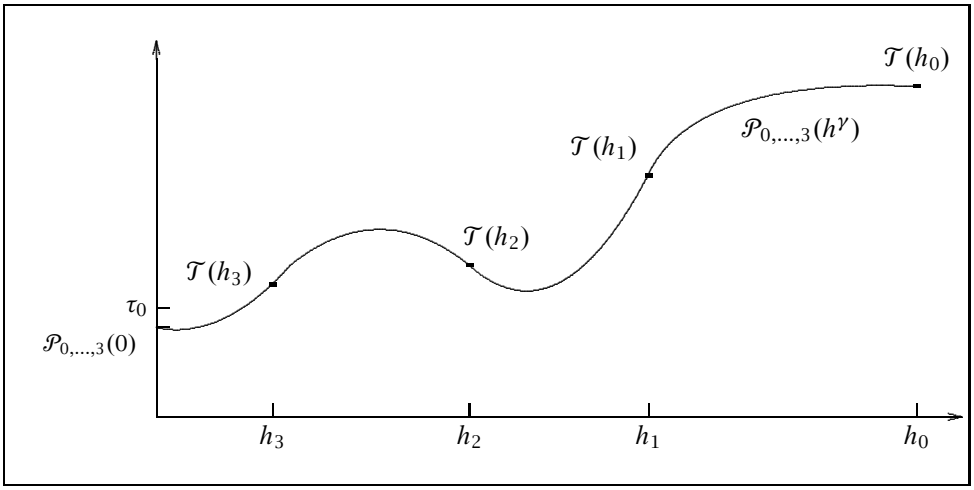


Bild 6.8: Darstellung der Vorgehensweise bei der Extrapolation; es ist $\mathcal{P}_{0,\ldots,3} \in \Pi_3$

6.7.2 Neville-Schema

Der Wert $\mathcal{P}_{0,\ldots,n}(0) \approx \mathcal{T}(0)$ lässt sich mit dem Neville-Schema berechnen. Für positive, paarweise verschiedene Schrittweiten $h_0,\ h_1, \ldots$ sei hierzu $\mathcal{P}_{k,\ldots,k+m} \in \Pi_m$ dasjenige Polynom mit

$$\mathcal{P}_{k,\ldots,k+m}(h_j^\gamma) \;=\; \mathcal{T}(h_j), \qquad j \;=\; k, k+1, \ldots, k+m, \tag{6.40}$$

[3] Für ein Polynom $\mathcal{P}$ wird die Funktion $h \to \mathcal{P}(h^\gamma)$ als *Polynom in* h^γ bezeichnet.

und es bezeichne

$$T_{k,\dots,k+m} \;:=\; \mathcal{P}_{k,\dots,k+m}(0). \tag{6.41}$$

Die Werte $T_{k,\dots,k+m}$ lassen sich mit dem Neville-Schema (1.7) rekursiv berechnen:

Theorem 6.24. *Für die Werte $T_{k,\dots,k+m}$ aus (6.41) gilt $T_k = \mathcal{T}(h_k)$ und*

$$T_{k,\dots,k+m} \;=\; T_{k+1,\dots,k+m} \;+\; \frac{T_{k+1,\dots,k+m} - T_{k,\dots,k+m-1}}{\left(\frac{h_k}{h_{k+m}}\right)^{\gamma} - 1} \qquad (m \geq 1, \quad k \geq 0).$$

BEWEIS. Mit der Darstellung (1.7) auf Seite 6 berechnet man leicht

$$\begin{aligned} T_{k,\dots,k+m} &= \frac{-h_k^{\gamma} T_{k+1,\dots,k+m} \;+\; h_{k+m}^{\gamma} T_{k,\dots,k+m-1}}{h_{k+m}^{\gamma} - h_k^{\gamma}} \\ &= T_{k+1,\dots,k+m} \;-\; h_{k+m}^{\gamma} \frac{T_{k+1,\dots,k+m} - T_{k,\dots,k+m-1}}{h_{k+m}^{\gamma} - h_k^{\gamma}} \\ &= \text{——} \,\text{«}\, \text{——} \;+\; \frac{T_{k+1,\dots,k+m} - T_{k,\dots,k+m-1}}{\left(\frac{h_k}{h_{k+m}}\right)^{\gamma} - 1}. \end{aligned}$$

□

Beispiel 6.25. Die zur summierten Trapezregel $\mathcal{T}_1(h)$ (hier gilt $\gamma = 2$) gehörenden Werte T_0, T_1 und T_{01} lauten für die Schrittweiten $h_0 = b - a$ und $h_1 = (b-a)/2$ folgendermaßen,

$$\begin{aligned} T_0 &= \tfrac{b-a}{2}\big(f(a) + f(b)\big), \qquad T_1 = \tfrac{b-a}{2}\big(\tfrac{f(a)}{2} + f\big(\tfrac{a+b}{2}\big) + \tfrac{f(b)}{2}\big), \\ T_{01} &= T_1 + \tfrac{T_1 - T_0}{4-1} \\ &= \tfrac{b-a}{2}\big(\tfrac{f(a)}{2} + f\big(\tfrac{a+b}{2}\big) + \tfrac{f(b)}{2}\big) + \tfrac{b-a}{6}\big(f\big(\tfrac{a+b}{2}\big) - \tfrac{f(b)}{2} - \tfrac{f(a)}{2}\big) \\ &= \tfrac{b-a}{6}\big(f(a) + 4f\big(\tfrac{a+b}{2}\big) + f(b)\big), \end{aligned}$$

so dass T_{01} der Simpson-Regel zur Approximation des Integrals $\int_a^b f(x)\,dx$ entspricht. △

6.7.3 Verfahrensfehler bei der Extrapolation

Die betrachteten Schrittweiten $h_0, h_1, \dots$ seien nun so gewählt, dass bezüglich einer Grundschrittweite $\widehat{h} > 0$ Folgendes gilt,

$$h_j \;=\; \widehat{h}/n_j \qquad \text{für} \quad j = 0, 1, \dots, \quad \text{mit} \quad 1 < n_0 \leq n_1 < \dots . \tag{6.42}$$

Mit dem folgenden Theorem, das einen Spezialfall der in Bulirsch [9] betrachteten Situation darstellt, wird beschrieben, wie gut die Werte $T_{k,\dots,k+m} = \mathcal{P}_{k,\dots,k+m}(0)$ den gesuchten Wert $\tau_0 = \lim_{h \to 0+} \mathcal{T}(h)$ approximieren.

Theorem 6.26. *Sei $\mathcal{T}(h),\ h > 0$, eine Funktion mit der asymptotischen Entwicklung (6.39), mit gewissen Zahlen $\gamma > 0$ und $r \in \mathbb{N}$. Für eine Folge $h_0,\ h_1, \ldots$ von Schrittweiten mit der Eigenschaft (6.42) erfülle das Polynom $\mathcal{P}_{k,\ldots,k+m} \in \Pi_m$ die Interpolationsbedingung (6.40), und $T_{k,\ldots,k+m}$ sei wie in (6.41). Dann gilt im Fall $0 \le m \le r-1$ die asymptotische Entwicklung*

$$T_{k,\ldots,k+m} \;=\; \tau_0 \;+\; (-1)^m \frac{\tau_{m+1}}{n_k^\gamma \cdots n_{k+m}^\gamma} \widehat{h}^{(m+1)\gamma} \;+\; \mathcal{O}\big(\widehat{h}^{(m+2)\gamma}\big) \qquad \textit{für} \quad \widehat{h} \to 0.$$

BEWEIS. O. B. d. A. darf $k = 0$ angenommen werden. Gemäß der lagrangeschen Interpolationsformel gilt

$$\mathcal{P}_{0,\ldots,m}(h^\gamma) \;=\; \sum_{j=0}^{m} \mathcal{T}(h_j) \Big[\prod_{\substack{s=0 \\ s\neq j}}^{m} \frac{h^\gamma - h_s^\gamma}{h_j^\gamma - h_s^\gamma} \Big] \qquad \text{für} \quad h \in \mathbb{R},$$

und somit

$$T_{0,\ldots,m} \;=\; \mathcal{P}_{0,\ldots,m}(0) \;=\; \sum_{j=0}^{m} c_{m,j}\, \mathcal{T}(h_j), \tag{6.43}$$

$$\text{mit } c_{m,j} \;:=\; \prod_{\substack{s=0 \\ s\neq j}}^{m} \frac{h_s^\gamma}{h_s^\gamma - h_j^\gamma} \;=\; \prod_{\substack{s=0 \\ s\neq j}}^{m} \frac{1}{1-(n_s/n_j)^\gamma}. \tag{6.44}$$

Nun gilt zum einen

$$\mathcal{T}(h_j) \;=\; \sum_{k=0}^{m+1} \tau_k h_j^{k\gamma} \;+\; \mathcal{O}(h_j^{(m+2)\gamma}), \tag{6.45}$$

und des Weiteren gilt nach Aufgabe 1.4 aus Kapitel 1 Folgendes,

$$\sum_{j=0}^{m} c_{m,j} h_j^{k\gamma} \;=\; \left\{ \begin{array}{cc} 1 & \text{für } k = 0, \\ 0 & \text{für } k = 1, \ldots, m, \\ (-1)^m h_0^\gamma \ldots h_m^\gamma & \text{für } k = m+1. \end{array} \right\} \tag{6.46}$$

Die beiden Identitäten (6.45) und (6.46) eingesetzt in (6.43) ergeben dann

$$\begin{aligned} T_{0,\ldots,m} \;&=\; \sum_{j=0}^{m} c_{m,j} \Big(\sum_{k=0}^{m+1} \tau_k h_j^{k\gamma} \;+\; \mathcal{O}(h_j^{(m+2)\gamma}) \Big) \\ &=\; \sum_{k=0}^{m+1} \Big(\sum_{j=0}^{m} c_{m,j}\, h_j^{k\gamma} \Big) \tau_k \qquad + \quad \sum_{j=0}^{m} c_{m,j}\, \mathcal{O}(h_j^{(m+2)\gamma}) \\ &= \tau_0 \;+\; (-1)^m \tau_{m+1} h_0^\gamma \ldots h_m^\gamma \;+\; \underbrace{\qquad\qquad « \qquad\qquad}_{= \mathcal{O}(\widehat{h}^{(m+2)\gamma})} \end{aligned}$$

unter Beachtung der Tatsache, dass die Koeffizienten $c_{m,j}$ aus (6.44) nicht von $\widehat{h}$ abhängen. Dies komplettiert den Beweis des Theorems. □

Bemerkung 6.27. Prominente Unterteilungen sind:

- $h_j = h_{j-1}/2$ für $j = 1, 2, \ldots$ mit $h_0 = \widehat{h}$ (*Romberg-Folge*)
- $h_0 = \widehat{h}$, $h_1 = \frac{\widehat{h}}{2}$, $h_2 = \frac{\widehat{h}}{3}$, $h_3 = \frac{\widehat{h}}{4}$, $h_4 = \frac{\widehat{h}}{6}$, $h_5 = \frac{\widehat{h}}{8}$, $h_6 = \frac{\widehat{h}}{12}$, $h_7 = \frac{\widehat{h}}{16}$, $h_8 = \frac{\widehat{h}}{24}, \ldots,$

  mit der Notation aus (6.42) allgemein $n_j = 2n_{j-2}$ für $j \geq 4$ (*Bulirsch-Folge*)

- $h_{j-1} = \widehat{h}/j$ für $j = 1, 2, \ldots$ (*harmonische Folge*) △

Beispiel 6.28. Speziell soll ausgehend von der Basisunterteilung $\widehat{h} = (b-a)/N$ noch die Romberg-Folge $h_j = \widehat{h}/2^j$ für $j = 0, 1, \ldots$ genauer betrachtet werden. Hier ist die Bedingung (6.42) mit $n_j = 2^j$ erfüllt, und unter den Bedingungen von Theorem 6.26 erhält man für $n \leq r-1$

$$T_{0,\ldots,n} \;=\; \tau_0 \;+\; \Big(\tfrac{(-1)^n}{2^{n(n+1)\gamma/2}}\tau_{n+1}\Big)\widehat{h}^{(n+1)\gamma} \;+\; \mathcal{O}\big(\widehat{h}^{(n+2)\gamma}\big).$$

Zur Veranschaulichung soll das Resultat noch speziell für die summierte Trapezregel

$$\mathcal{T}_1(h) \;=\; \int_a^b f(x)\,dx \;+\; \mathcal{O}(h^2)$$

betrachtet werden, mit $n = 2$. Mit der in Schema 6.1 angedeuteten Vorgehensweise erhält man so mit wenig Aufwand die sehr viel genauere Approximation

$$T_{012} \;=\; \int_a^b f(x)\,dx \;+\; \tfrac{\tau_3}{64}\widehat{h}^6 \;+\; \mathcal{O}(\widehat{h}^8). \qquad \triangle$$

$$\begin{array}{lcl}
\mathcal{T}_1(h_0) & = & T_0 \\
 & & \searrow \\
\mathcal{T}_1(h_1) & = & T_1 \to T_{01} \\
 & & \searrow \quad \searrow \\
\mathcal{T}_1(h_2) & = & T_2 \to T_{12} \to T_{012}
\end{array}$$

Schema 6.1: Neville-Schema zu Beispiel 6.28

6.8 Gaußsche Quadraturformeln

6.8.1 Einleitende Bemerkungen

Thema des vorliegenden Abschnitts ist die möglichst genaue numerische Berechnung gewichteter Integrale

$$\mathcal{I}(f) \;:=\; \int_a^b f(x)\varrho(x)\,dx \tag{6.47}$$

wobei $f : [a,b] \to \mathbb{R}$ eine vorgegebene Funktion und ϱ eine gegebene Gewichtsfunktion ist, siehe die folgende Definition. Hierbei werden zur Vereinfachung der Notation endliche Intervalle betrachtet, $-\infty < a \le b < \infty$. Die nachfolgenden Betrachtungen lassen sich jedoch auf unendliche Intervalle übertragen.

Definition 6.29. Es wird

$$\varrho : [a,b] \to (0, \infty]$$

Gewichtsfunktion genannt, wenn sie auf dem offenen Intervall (a, b) stückweise stetig sowie über $[a,b]$ integrabel ist.

Zur numerischen Berechnung des Integrals (6.47) werden wieder interpolatorische Quadraturformeln

$$\mathcal{I}_n(f) \;=\; \sum_{k=1}^{n} \sigma_k f(\lambda_k), \tag{6.48}$$

herangezogen, wobei im Unterschied zur Formel (6.2) teils aus historischen Gründen hier jedoch

- die Stützstellen mit λ_k bezeichnet werden,
- die Summation bei $k = 1$ beginnt,
- der Faktor $b - a$ fehlt.

In diesem Abschnitt wird beschrieben, für welche Wahl der Stützstellen $\lambda_1, \lambda_2, \ldots, \lambda_n$ und Gewichte $\sigma_1, \sigma_2, \ldots, \sigma_n$ der Genauigkeitsgrad der zugehörigen interpolatorischen Quadraturformel einen möglichst hohen Wert annimmt. Die Begriffe *interpolatorische Quadraturformel* und *Genauigkeitsgrad* sind hierbei ganz kanonisch auf Integrale mit Gewichten zu übertragen (wobei allerdings in den nachfolgenden Betrachtungen auch der Fall $\varrho \equiv 1$ von Interesse ist). Die resultierenden Formeln werden dann als *gaußsche Quadraturformeln* bezeichnet. Bei der Herleitung dieser Formeln werden orthogonale Polynome benötigt.

6.8.2 Orthogonale Polynome

Definition 6.30. Zu gegebener Gewichtsfunktion $\varrho : [a,b] \to (0, \infty]$ bezeichne

$$\langle p, q \rangle \;=\; \int_a^b p(x) q(x) \varrho(x)\, dx, \qquad \|p\| \;=\; \langle p, p \rangle^{1/2} \qquad \text{für} \quad p,\ q \in \Pi.$$

Die Abbildung $\langle \cdot, \cdot \rangle : \Pi \times \Pi \to \mathbb{R}$ definiert ein Skalarprodukt auf dem Raum aller reellen Polynome Π, insbesondere ist also $\langle \cdot, \cdot \rangle$ linear in jedem seiner Argumente bei jeweils festem anderem Argument, und es gilt $\langle p, p \rangle > 0$ für $0 \ne p \in \Pi$. Wir führen noch die folgende Notation ein.

Definition 6.31. 1. Zwei Polynome $p,\ q \in \Pi$ heißen *orthogonal* zueinander, wenn $\langle p, q \rangle = 0$ gilt.

2. Das *orthogonale Komplement von* $\Pi_n \subset \Pi$ ist gegeben durch

$$\Pi_n^{\perp} \;:=\; \{ p \in \Pi : \langle p, q \rangle \;=\; 0 \quad \forall\, q \in \Pi_n \}, \qquad n \;=\; 0, 1, \ldots .$$

Offensichtlich ist $\Pi_n^\perp$ ein linearer Unterraum von Π. Eine spezielle Folge paarweise orthogonaler Polynome erhält man durch Gram-Schmidt-Orthogonalisierung der Monome $1,\ x,\ x^2, \ldots$:

$$p_0 = 1, \tag{6.49}$$

$$p_n = x^n - \sum_{m=0}^{n-1} \frac{\langle x^n, p_m \rangle}{\| p_m \|^2} p_m, \qquad n = 1, 2, \ldots . \tag{6.50}$$

Nach Konstruktion ist also p_n ein Polynom vom genauen Grad n mit führendem Koeffizienten eins, und es gilt

$$p_n \in \Pi_{n-1}^\perp. \tag{6.51}$$

Mit dem nachfolgenden Theorem wird eine Vorgehensweise vorgestellt, mit der sich diese Orthogonalpolynome effizient berechnen lassen.

Theorem 6.32. *Die Orthogonalpolynome in (6.49), (6.50) genügen der* Drei-Term-Rekursion

$$\begin{aligned} p_0 &= 1, \qquad p_1 = x - \beta_0, \\ p_{n+1} &= (x - \beta_n) p_n - \gamma_n^2 p_{n-1}, \qquad n = 1, 2, \ldots, \end{aligned}$$

mit den Koeffizienten

$$\beta_n = \frac{\langle x p_n, p_n \rangle}{\| p_n \|^2} \quad \textit{für } n = 1, 2, \ldots, \qquad \gamma_n^2 = \frac{\| p_n \|^2}{\| p_{n-1} \|^2} \quad \textit{für } n = 1, 2, \ldots .$$

BEWEIS. Offenbar ist die angegebene Darstellung richtig für p_0 und p_1. Für $n \geq 1$ setzen wir

$$q_{n+1} := (x - \beta_n) p_n - \gamma_n^2 p_{n-1}$$

und zeigen im Folgenden $q_{n+1} = p_{n+1}$. Dazu beobachtet man, dass q_{n+1} (ebenso wie p_{n+1}) ein Polynom mit genauem Grad $n+1$ ist und den führenden Koeffizienten eins besitzt, und somit gilt

$$r := p_{n+1} - q_{n+1} \in \Pi_n. \tag{6.52}$$

Wir zeigen nun, dass q_{n+1} (ebenso wie p_{n+1}) im orthogonalen Komplement von Π_n liegt, so dass dann auch

$$r = p_{n+1} - q_{n+1} \in \Pi_n^\perp \tag{6.53}$$

gilt. Die Beziehungen (6.52) und (6.53) zusammen ergeben dann $\| r \|^2 = \langle r, r \rangle = 0$ und damit wie behauptet $p_{n+1} = q_{n+1}$.

Wie angekündigt wird nun

$$q_{n+1} \in \Pi_n^\perp \tag{6.54}$$

nachgewiesen. Aufgrund der Identität $\langle p_n, p_{n-1} \rangle = 0$ und der Definition von β_n gilt

$$\langle q_{n+1}, p_n \rangle \;=\; \langle x p_n, p_n \rangle \;-\; \beta_n \| p_n \|^2 \;=\; 0. \tag{6.55}$$

Weiter erhält man wieder wegen $\langle p_n, p_{n-1} \rangle = 0$ sowie aufgrund der Definition von γ_n Folgendes,

$$\langle q_{n+1}, p_{n-1} \rangle \;=\; \langle p_n, x p_{n-1} \rangle \;-\; \gamma_n^2 \| p_{n-1} \|^2 \;=\; \langle p_n, x p_{n-1} - p_n \rangle \;=\; 0, \tag{6.56}$$

wobei das letzte Gleichheitszeichen aus der Tatsache folgt, dass $x p_{n-1} - p_n$ ein Polynom vom Grad $\le n-1$ darstellt. Ferner ist q_{n+1} auch orthogonal zu jedem Polynom vom Grad $\le n-2$, denn es gilt

$$\langle q_{n+1}, q \rangle \;=\; \underbrace{\langle p_n, xq \rangle}_{=\,0} - \beta_n \underbrace{\langle p_n, q \rangle}_{=\,0} - \gamma_n^2 \underbrace{\langle p_{n-1}, q \rangle}_{=\,0} \;=\; 0 \quad \forall\, q \in \Pi_{n-2}. \tag{6.57}$$

Wegen $\Pi_n = \operatorname{span}\{p_n,\ p_{n-1}\} \oplus \Pi_{n-2}$ folgt aus (6.55)–(6.57) die nachzuweisende Eigenschaft (6.54), mit der man wie bereits beschrieben $p_{n+1} = q_{n+1}$ erhält. □

Das folgende Theorem liefert Aussagen über die Nullstellen der betrachteten Orthogonalpolynome.

Theorem 6.33. *Die Nullstellen $\lambda_1, \lambda_2, \ldots, \lambda_n$ des n-ten Orthogonalpolynoms p_n in (6.50) sind einfach und liegen alle im offenen Intervall (a, b). Sie besitzen die Darstellung*[4]

$$\lambda_k \;=\; \frac{\langle x L_k, L_k \rangle}{\| L_k \|^2}, \qquad L_k(x) \;:=\; \prod_{\substack{s=1 \\ s \neq k}}^{n} \frac{x - \lambda_s}{\lambda_k - \lambda_s} \qquad \textit{für}\ \ k \;=\; 1, 2, \ldots, n. \tag{6.58}$$

BEWEIS. Es seien $a < \lambda_1 < \ldots < \lambda_m < b$ $(0 \le m \le n)$ diejenigen Nullstellen von p_n in dem offenen Intervall (a, b), an denen p_n sein Vorzeichen wechselt, also diejenigen Nullstellen von p_n in (a, b) mit ungerader Vielfachheit. Im Folgenden wird $m = n$ nachgewiesen. Wäre $m \le n-1$, so hätte nämlich das Polynom

$$q(x) \;:=\; \prod_{k=1}^{m} (x - \lambda_k)$$

den Grad $0 \le m \le n-1$, so dass wegen (6.51)

$$\langle p_n, q \rangle \;=\; 0 \tag{6.59}$$

folgt. Nun ist aber das Polynom $p_n(x) q(x)$ nach Konstruktion von einem Vorzeichen auf $[a, b]$, so dass

$$\langle p_n, q \rangle \;=\; \int_a^b p_n(x) q(x) \varrho(x)\, dx \;\neq\; 0$$

[4] wobei $L_1, \ldots, L_n \in \Pi_{n-1}$ die den Nullstellen $\lambda_1, \ldots, \lambda_n$ zugeordneten lagrangeschen Basispolynome darstellen

gilt im Widerspruch zu (6.59).

Um zur Darstellung (6.58) zu gelangen, faktorisiert man p_n in der Form

$$p_n(x) \;=\; (x-\lambda_k)\widehat{q}(x),$$

mit einem geeigneten Polynom $\widehat{q} \in \Pi_{n-1}$ und erhält daraus

$$0 \;=\; \langle\, p_n, \widehat{q}\,\rangle \;=\; \langle\, x\widehat{q}, \widehat{q}\,\rangle \;-\; \lambda_k \langle\, \widehat{q}, \widehat{q}\,\rangle.$$

Hieraus folgt wegen $\langle\, \widehat{q}, \widehat{q}\,\rangle \neq 0$

$$\lambda_k \;=\; \frac{\langle\, x\widehat{q}, \widehat{q}\,\rangle}{\|\widehat{q}\|^2} \;=\; \frac{\langle\, xL_k, L_k\,\rangle}{\|L_k\|^2},$$

wobei sich die letzte Gleichung daraus ergibt, dass die Polynome $\widehat{q}$ und L_k bis auf einen konstanten Faktor übereinstimmen. □

Beispiel 6.34. In Tabelle 6.1 sind für verschiedene Intervalle und Gewichtsfunktionen die Bezeichnungen der zugehörigen orthogonalen Polynome aufgelistet.

Intervall	$\varrho(x)$	zugehörige orthogonale Polynome
$[-1,1]$	$\mathbf{1}$	Legendre-Polynome
$[-1,1]$	$1/\sqrt{1-x^2}$	Tschebyscheff-Polynome der ersten Art T_n
$[-1,1]$	$(1-x)^\alpha(1+x)^\beta,\ \alpha>-1,\ \beta>-1$	Jacobi-Polynome
$(-\infty,\infty)$	e^{-x^2}	Hermite-Polynome
$(-\infty,\infty)$	$e^{-x^2}x^\alpha,\quad \alpha>-1$	Laguerre-Polynome

Tabelle 6.1: Verschiedene Systeme von Orthogonalpolynomen

Man beachte, dass in den beiden zuletzt genannten Beispielen anders als bisher angenommen unendliche Intervalle betrachtet werden; hierzu sei auf die Bemerkung eingangs dieses Abschnitts 6.8 verwiesen. △

6.8.3 Optimale Wahl der Stützstellen und Gewichte

Das folgende Theorem beschreibt, unter welchen Bedingungen an n Stützstellen und Gewichte der Genauigkeitsgrad einer Quadraturformel $2n-1$ beträgt.

Theorem 6.35. *Für ein $n \in \mathbb{N}$ seien $\lambda_1, \ldots, \lambda_n \in \mathbb{R}$ paarweise verschiedene Zahlen, und weiter seien $\sigma_1, \ldots, \sigma_n \in \mathbb{R}$ beliebig. Dann und nur dann gilt*

$$\langle\, p, \mathbf{1}\,\rangle \;=\; \sum_{k=1}^{n} \sigma_k\, p(\lambda_k) \qquad \textit{für} \quad p \;\in\; \Pi_{2n-1}, \tag{6.60}$$

wenn die folgenden Bedingungen (a) und (b) erfüllt sind,

(a) die Zahlen $\lambda_1, \ldots, \lambda_n \in \mathbb{R}$ stimmen mit den Nullstellen des n-ten orthogonalen Polynoms p_n (siehe (6.58)) überein,

(b) die Gewichte $\sigma_1, \sigma_2, \ldots, \sigma_n$ haben die Gestalt

$$\sigma_k \;=\; \langle\, L_k, \mathbf{1} \,\rangle \qquad \textit{für} \quad k \;=\; 1, 2, \ldots, n,$$

wobei $L_1, L_2, \ldots, L_n \in \Pi_{n-1}$ die den Zahlen $\lambda_1, \lambda_2, \ldots, \lambda_n$ zugeordneten lagrangeschen Basispolynome darstellen[5]*.*

Unter diesen Bedingungen gilt auch $\sigma_k \;=\; \langle\, L_k, L_k \,\rangle > 0$ für $k = 1, 2, \ldots, n$.

BEWEIS. "$\Longrightarrow$" Es gelte (6.60), und zum Beweis von (a) setzen wir

$$q(x) \;:=\; (x - \lambda_1) \cdots (x - \lambda_n)$$

und weisen im Folgenden die Identität $q = p_n$ nach. Hierzu wendet man die Identität (6.60) auf das Polynom $p(x) := x^m q(x)$ mit $m \in \{0, 1, \ldots, n-1\}$ an und erhält

$$\langle\, q, x^m \,\rangle \;=\; \langle\, x^m q, \mathbf{1} \,\rangle \;=\; \sum_{k=1}^{n} \sigma_k \, \lambda_k^m \underbrace{q(\lambda_k)}_{=\,0} \;=\; 0 \qquad \text{für} \quad m \;=\; 0, 1, \ldots, n-1,$$

was insgesamt

$$q \in \Pi_{n-1}^{\perp}$$

und damit $q - p_n \in \Pi_{n-1}^{\perp}$ nach sich zieht. Außerdem ist q ein Polynom mit genauem Grad n und führendem Koeffizienten eins, so dass sich die Eigenschaft $q - p_n \in \Pi_{n-1}$ ergibt, was schließlich (wie im Beweis des vorigen Theorems 6.32) $q = p_n$ liefert. Teil (b) ergibt sich wegen $L_j(\lambda_k) = \delta_{jk}$ unmittelbar aus der Identität (6.60) angewandt mit $p = L_j$.

"$\Longleftarrow$" Es gelte nun (a), (b), und $p \in \Pi_{2n-1}$ sei beliebig. Dann lässt sich das Polynom p in der Form[6]

$$p \;=\; q p_n \;+\; r$$

schreiben mit gewissen Polynomen $q,\ r \in \Pi_{n-1}$. Wegen $p_n(\lambda_k) \;=\; 0$ gilt dann

$$p(\lambda_k) \;=\; r(\lambda_k), \qquad k \;=\; 1, 2, \ldots, n,$$

und mit der lagrangeschen Interpolationsformel erhält man

$$r(x) \;=\; \sum_{k=1}^{n} r(\lambda_k) L_k(x) \;=\; \sum_{k=1}^{n} p(\lambda_k) L_k(x).$$

Dies führt dann auf die angegebene Identität (6.60):

$$\langle\, p, \mathbf{1} \,\rangle \;=\; \underbrace{\langle\, q, p_n \,\rangle}_{=\,0} + \langle\, r, \mathbf{1} \,\rangle \;=\; \sum_{k=1}^{n} p(\lambda_k) \langle\, L_k, \mathbf{1} \,\rangle \;=\; \sum_{k=1}^{n} \sigma_k \, p(\lambda_k).$$

Die angegebene Darstellung $\sigma_k = \langle\, L_k, L_k \,\rangle > 0$ für die Gewichte ergibt sich aus der Darstellung (6.60) angewandt auf das Polynom $p = L_k^2$. □

[5] vergleiche (6.58)

[6] nach Polynomdivision mit Rest

Bemerkung 6.36. Man beachte, dass hier (im Unterschied zu den abgeschlossenen Newton-Cotes-Formeln) die Gewichte in jedem Fall positiv ausfallen[7]. △

Definition 6.37. Die Quadraturformel

$$\mathcal{I}_n(f) := \sum_{k=1}^{n} \sigma_k f(\lambda_k) \qquad \text{für} \quad f \in C[a,b], \tag{6.61}$$

mit den Stützstellen $\lambda_1, \ldots, \lambda_n \in \mathbb{R}$ und Gewichten $\sigma_1, \ldots, \sigma_n$ wie in (a) und (b) aus Theorem 6.35 bezeichnet man als *gaußsche Quadraturformel.*

Als eine unmittelbare Konsequenz aus Theorem 6.35 erhält man:

Korollar 6.38. *Die gaußsche Quadraturformel (6.61) ist interpolatorisch und besitzt mindestens den Genauigkeitsgrad* $r = 2n-1$.

BEWEIS. Zu einer gegebenen Funktion $f \in C[a,b]$ sei $\mathcal{Q}_{n-1} \in \Pi_{n-1}$ das interpolierende Polynom zu den Stützpunkten $(\lambda_1, f(\lambda_1)), (\lambda_2, f(\lambda_2)), \ldots, (\lambda_n, f(\lambda_n))$. Aus der Eigenschaft (6.60) erhält man die erste Aussage,

$$\sum_{k=1}^{n} \sigma_k f(\lambda_k) = \sum_{k=1}^{n} \sigma_k \mathcal{Q}_{n-1}(\lambda_k) = \langle \mathcal{Q}_{n-1}, \mathbf{1} \rangle,$$

und die angegebene untere Schranke für den Genauigkeitsgrad folgt ebenfalls unmittelbar aus (6.60). □

Mit dem folgenden Resultat wird die Fehleraussage aus Theorem 6.13 (siehe Seite 125) auf die vorliegende Situation der gewichteten Integrale übertragen.

Theorem 6.39. *Für den Fehler bei der Gaußquadratur (6.61) gilt unter der Voraussetzung* $f \in C^{2n}[a,b]$ *die Darstellung*

$$\mathcal{I}(f) - \mathcal{I}_n(f) = \Big(\tfrac{1}{(2n)!} \int_a^b p_n^2(x)\, \varrho(x)\, dx\Big) f^{(2n)}(\xi) \tag{6.62}$$

$$= \tfrac{(b-a)^{2n+1}}{(2n)!} \Big(\int_0^1 \Big[\prod_{k=1}^{n} (t-t_k)^2 \Big] \varrho((b-a)t+a)\, dt \Big) f^{(2n)}(\xi) \tag{6.63}$$

mit $t_k := \frac{\lambda_k - a}{b-a}$ *für* $k = 1, 2, \ldots, n$, *und mit einer geeigneten Zwischenstelle* $\xi \in [a,b]$.

BEWEIS. Der Genauigkeitsgrad bei der Gaußquadratur (6.61) beträgt nach Korollar 6.38 mindestens $r = 2n-1$. Wählt man zu den Stützstellen $\lambda_1, \lambda_2, \ldots, \lambda_n$ nun die weiteren Stützstellen $\lambda_{n+1} = \lambda_1, \ldots, \lambda_{2n} = \lambda_n$, so ist

$$\prod_{k=1}^{2n} (x - \lambda_k) = \prod_{k=1}^{n} (x - \lambda_k)^2 = p_n^2(x)$$

von einem Vorzeichen, und man erhält dann die Resultate (6.62)–(6.63) mit der gleichen Vorgehensweise wie in den Teilen 1 und 3 des Beweises von Theorem 6.13. □

[7] vergleiche hierzu die Anmerkungen vor Theorem 6.11

Bemerkung 6.40. 1. Als unmittelbare Konsequenz aus Theorem 6.39 ergibt sich, dass der Genauigkeitsgrad der gaußschen Quadraturformeln genau $r = 2n-1$ beträgt. Dies ist optimal; für die Situation $\varrho = 1$ siehe hierzu Aufgabe 6.2.

2. Man kann auch summierte gaußsche Quadraturformeln betrachten und anwenden; die Resultate aus Abschnitt 6.5 lassen sich ganz kanonisch übertragen. △

6.8.4 Nullstellen von orthogonalen Polynomen als Eigenwerte

Für größere Werte von n steht man noch vor dem Problem, die Nullstellen des n-ten orthogonalen Polynoms p_n sowie die Gewichte $\sigma_1, \ldots, \sigma_n$ zu bestimmen. Dazu gehen wir im Folgenden davon aus, dass die Koeffizienten β_j und γ_j in der Rekursion

$$p_0 = 1, \qquad p_1 = x - \beta_0, \tag{6.64}$$

$$p_{j+1} = (x-\beta_j)p_j - \gamma_j^2 p_{j-1}, \qquad j = 1, 2, \ldots, \tag{6.65}$$

explizit bekannt sind und betrachten dann die symmetrische Matrix

$$\mathbf{J} = \begin{pmatrix} \beta_0 & -\gamma_1 & 0 & \ldots & 0 \\ -\gamma_1 & \beta_1 & -\gamma_2 & \ddots & \vdots \\ 0 & -\gamma_2 & \ddots & \ddots & 0 \\ \vdots & \ddots & \ddots & \ddots & -\gamma_{n-1} \\ 0 & \ldots & 0 & -\gamma_{n-1} & \beta_{n-1} \end{pmatrix} \in \mathbb{R}^{n\times n}. \tag{6.66}$$

Theorem 6.41. *Die Nullstellen $\lambda_1, \lambda_2, \ldots, \lambda_n$ des n-ten Orthogonalpolynoms p_n stimmen mit den Eigenwerten der Matrix $\mathbf{J}$ überein, und die Gewichte ergeben sich daraus folgendermaßen:*

$$\sigma_k = \langle \mathbf{1}, \mathbf{1} \rangle / \Big(\sum_{j=0}^{n-1} \tau_j^2 \, p_j^2(\lambda_k) \Big) \qquad \textit{für} \quad k = 1, 2, \ldots, n, \tag{6.67}$$

mit den Zahlen

$$\tau_j := \begin{cases} 1 & \textit{für } j = 0, \\ (-1)^j / (\gamma_1 \gamma_2 \cdots \gamma_j) & \textit{für } j = 1, 2, \ldots, n-1. \end{cases}$$

BEWEIS. Es wird zunächst Folgendes nachgewiesen,

$$\mathbf{J} v^{(k)} = \lambda_k v^{(k)} \qquad \text{für} \quad k = 1, 2, \ldots, n, \tag{6.68}$$

mit dem Vektor

$$v^{(k)} = \big(\underbrace{\tau_0 \, p_0(\lambda_k)}_{=\,1}, \; \tau_1 \, p_1(\lambda_k), \ldots, \; \tau_{n-1} \, p_{n-1}(\lambda_k) \big)^\top \in \mathbb{R}^n.$$

Es ist

$$\begin{aligned}(\mathbf{J}v^{(k)})_1 &= \beta_0 \cdot 1 - \gamma_1 \tau_1 p_1(\lambda_k) = \beta_0 + p_1(\lambda_k) = \beta_0 + \lambda_k - \beta_0 \\ &= \lambda_k = \lambda_k v_1^{(k)},\end{aligned}$$

und weiter erhält man aus den Rekursionsformeln (6.65) mit $x = \lambda_k$ Folgendes (wobei in der nachfolgenden Situation $j = n-1$ noch $\gamma_n := \tau_n := 0$ gesetzt wird und $p_n(\lambda_k) = 0$ zu beachten ist):

$$\begin{aligned}(\mathbf{J}v^{(k)})_{j+1} &= -\gamma_j \tau_{j-1} p_{j-1}(\lambda_k) + \beta_j \tau_j p_j(\lambda_k) - \gamma_{j+1} \tau_{j+1} p_{j+1}(\lambda_k) \\ &= \underbrace{\frac{(-1)^j}{\gamma_1 \cdots \gamma_j}}_{=\tau_j} \left[\gamma_j^2 p_{j-1}(\lambda_k) + \beta_j p_j(\lambda_k) + p_{j+1}(\lambda_k)\right] \\ &= \tau_j \lambda_k p_j(\lambda_k) = \lambda_k v_{j+1}^{(k)} \qquad \text{für } j = 1, 2, \ldots, n-1,\end{aligned}$$

und (6.68) ist damit bewiesen. Im Folgenden soll noch die Darstellung (6.67) nachgewiesen werden. Die Identität (6.68) bedeutet noch, dass $v^{(k)}$ Eigenvektor zum Eigenwert λ_k der Matrix $\mathbf{J}$ ist. Gemäß Theorem 6.33 sind diese Eigenwerte paarweise verschieden, und aus der Symmetrie der Matrix $\mathbf{J}$ erhält man dann

$$v^{(k)\top} v^{(\ell)} = 0 \qquad \text{für } k \neq \ell. \tag{6.69}$$

Aufgrund der paarweisen Orthogonalität der Polynome $p_0, p_1, \ldots$ sowie wegen Theorem 6.35 gilt

$$\delta_{j0} \langle \mathbf{1}, \mathbf{1} \rangle = \langle p_j, \mathbf{1} \rangle = \sum_{\ell=1}^{n} \sigma_\ell p_j(\lambda_\ell) \qquad \text{für } j = 0, 1, \ldots, n-1, \tag{6.70}$$

und Multiplikation von (6.70) mit $\tau_j^2 p_j(\lambda_k)$ sowie anschließende Summation über j liefert

$$\begin{aligned}\langle \mathbf{1}, \mathbf{1} \rangle &= \sum_{j=0}^{n-1} \sum_{\ell=1}^{n} \sigma_\ell \tau_j^2 p_j(\lambda_k) p_j(\lambda_\ell) = \sum_{\ell=1}^{n} \sigma_\ell \sum_{j=0}^{n-1} \tau_j^2 p_j(\lambda_k) p_j(\lambda_\ell) \\ &= \sum_{\ell=1}^{n} \sigma_\ell (v^{(k)})^\top v^{(\ell)} = \sigma_k (v^{(k)})^\top v^{(k)},\end{aligned}$$

wobei in der letzten Gleichheit noch die Orthogonalitätsbeziehung (6.69) eingeht. Dies liefert die Aussage (6.67). □

Bemerkung 6.42. Die gesuchten Eigenwerte der Matrix $\mathbf{J}$ aus (6.66) können für größere Werte von n nur numerisch berechnet werden. Entsprechende Methoden werden in Kapitel 13 vorgestellt. △

6.9 Nachtrag: Beweis der Asymptotik für die summierte Trapezregel

6.9.1 Bernoulli-Polynome

Definition 6.43. Die *Bernoulli-Polynome* B_k sind rekursiv erklärt: $B_0(x) \equiv 1$, und für $k = 1, 2, \ldots$ gilt

$$B_k(x) \;=\; A_k \;+\; k\int_0^x B_{k-1}(t)\,dt, \qquad x \in [0,1], \tag{6.71}$$

$$\text{mit} \quad A_k \;:=\; -k\int_0^1 \Big(\int_0^x B_{k-1}(t)\,dt\Big)\,dx. \tag{6.72}$$

Beispielsweise gilt

$$B_1(x) \;=\; x - \tfrac{1}{2}, \qquad B_2(x) \;=\; x^2 - x + \tfrac{1}{6}, \tag{6.73}$$

$$B_3(x) \;=\; x^3 - \tfrac{3}{2}x^2 + \tfrac{1}{2}x, \qquad B_4(x) \;=\; x^4 - 2x^3 + x^2 - \tfrac{1}{30}.$$

Theorem 6.44. *Für die Bernoulli-Polynome B_k aus (6.71)–(6.72) gelten die folgenden Aussagen:*

(a) (äquivalente Formulierung) Es gilt $B_k \in \Pi_k$ für $k = 0, 1, \ldots$, und

$$B_k'(x) \;=\; kB_{k-1}(x), \qquad \int_0^1 B_k(x)\,dx \;=\; 0 \qquad \textit{für} \quad k \;=\; 1, 2, \ldots\,. \tag{6.74}$$

(b) Es gilt $B_1(0) = -1/2,\ B_1(1) = 1/2$, und

$$A_k \;=\; B_k(0) \;=\; B_k(1) \qquad \textit{für} \quad k \;=\; 2, 3, \ldots\,.$$

(c) Die Funktion B_{2k} ist gerade bezüglich $x = 1/2$, und B_{2k+1} ist ungerade bezüglich $x = 1/2$, es gilt also

$$B_{2k}\big(\tfrac{1}{2} + x\big) \;=\; B_{2k}\big(\tfrac{1}{2} - x\big) \qquad \textit{für} \quad 0 \le x \le \tfrac{1}{2},$$

$$B_{2k+1}\big(\tfrac{1}{2} + x\big) \;=\; -B_{2k+1}\big(\tfrac{1}{2} - x\big) \quad \text{——— « ———};$$

(d) $B_{2k+1}(0) \;=\; B_{2k+1}(1) \;=\; 0$ *für* $k \;=\; 1, 2, \ldots\,.$

BEWEIS. "(a)" gilt offensichtlich.

"(b)": Die Aussage für B_1 resultiert unmittelbar aus (6.73). Für $k \ge 2$ folgt $A_k = B_k(0)$ aus der Definition (6.71), und wegen der Mittelwerteigenschaft in (6.74) erhält man

$$B_k(1) \;=\; A_k \;+\; k\int_0^1 B_{k-1}(x)\,dx \;=\; A_k \;+\; k\cdot 0 \;=\; A_k.$$

"(c)" wird mit vollständiger Induktion nachgewiesen. $B_0 \equiv 1$ ist eine gerade Funktion bezüglich $x = 1/2$, und wir nehmen nun an, dass B_{2k} eine bezüglich $x = 1/2$ gerade Funktion ist. Dann gilt

$$\begin{aligned} B_{2k+1}(x) &= A_{2k+1} + (2k+1)\int_0^x B_{2k}(t)\,dt \\ &= \underbrace{A_{2k+1} + (2k+1)\int_0^{1/2} B_{2k}(t)\,dt}_{=:\ \widehat{A}_{2k+1}} + (2k+1)\underbrace{\int_{1/2}^x B_{2k}(t)\,dt}_{=:\ Q(x)},\quad 0 \le x \le 1. \end{aligned}$$

Nun ist Q ungerade bezüglich $x = 1/2$, denn

$$\begin{aligned} Q\big(\tfrac{1}{2}+x\big) &= \int_{1/2}^{1/2+x} B_{2k}(t)\,dt = \int_0^x B_{2k}\big(\tfrac{1}{2}+t\big)\,dt = \int_0^x B_{2k}\big(\tfrac{1}{2}-t\big)\,dt \\ &= \int_{1/2}^{1/2-x} B_{2k}(t)(-1)\,dt = -Q\big(\tfrac{1}{2}-x\big). \end{aligned}$$

Damit gilt aber notwendigerweise $\int_0^1 Q(x)\,dx = 0$, und wegen $\int_0^1 B_{2k+1}(x)\,dx = 0$, vergleiche (6.74), ist $\widehat{A}_{2k+1} = 0$ und somit $B_{2k+1} = (2k+1)Q$ eine bezüglich $x = 1/2$ ungerade Funktion.

Sofort ergibt sich nun, dass B_{2k+2} bezüglich $x = 1/2$ eine gerade Funktion ist:

$$\begin{aligned} B_{2k+2}\big(\tfrac{1}{2}+x\big) &= A_{2k+2} + (2k+2)\int_0^{1/2+x} B_{2k+1}(t)\,dt \\ &= A_{2k+2} + (2k+2)\int_0^{1/2-x} B_{2k+1}(t)\,dt + (2k+2)\underbrace{\int_{1/2-x}^{1/2+x} B_{2k+1}(t)\,dt}_{=\ 0} \\ &= B_{2k+2}\big(\tfrac{1}{2}-x\big) \qquad \text{für}\quad 0 \le x \le \tfrac{1}{2}. \end{aligned}$$

"(d)" Die erste Identität in (d) ist schon in (b) festgehalten, und die dritte Gleichheit ergibt sich aus der Tatsache, dass B_{2k+1} bezüglich $x = 1/2$ eine ungerade Funktion ist:

$$B_{2k+1}(1) = B_{2k+1}\big(\tfrac{1}{2}+\tfrac{1}{2}\big) = -B_{2k+1}\big(\tfrac{1}{2}-\tfrac{1}{2}\big) = -B_{2k+1}(0) = -B_{2k+1}(1).$$

□

Definition 6.45. Die Werte $B_{2k}(0)$, $k = 0, 1, \ldots$, heißen *bernoullische Zahlen*.

Die ersten bernoullischen Zahlen sind

$$B_0(0) = 1, \qquad B_2(0) = \frac{1}{6}, \qquad B_4(0) = -\frac{1}{30}, \qquad B_6(0) = \frac{1}{42}, \qquad B_8(0) = \frac{1}{30}.$$

Die bernoullischen Zahlen spielen beim Beweis von Theorem 6.22 eine Rolle.

6.9.2 Der Beweis von Theorem 6.22

Im Folgenden wird der Beweis von Theorem 6.22 geführt, und hierzu setzt man die Bernoulli-Polynome B_k von dem Intervall $[0,1]$ ausgehend 1-periodisch fort,

$$S_k(x) := B_k(x-m) \qquad \text{für } m \le x < m+1, \qquad m = 0, 1, \dots .$$

Es ist S_0 eine Sägezahnfunktion, die Funktion S_1 ist stückweise stetig differenzierbar, und für $k \ge 2$ ist S_k stetig differenzierbar, und es gilt

$$S_k'(x) = kS_{k-1}(x) \qquad \text{für } m < x < m+1, \qquad m \in \mathbb{N}_0 \qquad (k = 1, 2, \dots).$$

Im weiteren Verlauf wird nachgewiesen, dass die Darstellung (6.37) richtig ist mit τ_0 wie in (6.38) und für

$$\tau_k := \frac{B_{2k}(0)}{(2k)!}\big(f^{(2k-1)}(b) - f^{(2k-1)}(a)\big), \qquad k = 1, 2, \dots, r, \tag{6.75}$$

$$R_{r+1}(h) := \Big(\frac{1}{(2r+2)!}\int_a^b \big[S_{2r+2}(0) - S_{2r+2}\big(\tfrac{x-a}{h}\big)\big] f^{(2r+2)}(x)\,dx\Big) h^{2r+2}. \tag{6.76}$$

Aus (6.76) folgt dann

$$|R_{r+1}(h)| \le \frac{2(b-a)}{(2r+2)!} \max_{y\in[0,1]} |B_{2r+2}(y)| \max_{x\in[a,b]} |f^{(2r+2)}(x)|\, h^{2r+2}$$

und damit die zweite Darstellung in (6.38). Zum Beweis der Darstellung (6.37) mit den Koeffizienten aus (6.38), (6.75) und (6.76) wird zur Vereinfachung zunächst die Intervall-Transformation $[a,b] \to [0,N]$ vorgenommen: sei

$$g(t) := f(a+th), \qquad 0 \le t \le N.$$

Die Identität (6.37) mit den Koeffizienten aus (6.38), (6.75) und (6.76) ist dann äquivalent zu der *euler-maclaurinschen Summenformel*

$$\left.\begin{aligned} \frac{g(0)}{2} + g(1) + \dots + g(N-1) + \frac{g(N)}{2} - \int_0^N g(t)\,dt \\ = \sum_{k=1}^{r} \frac{B_{2k}(0)}{(2k)!}\big(g^{(2k-1)}(N) - g^{(2k-1)}(0)\big) + C_{r+1} \end{aligned}\right\} \tag{6.77}$$

mit dem Fehlerterm

$$C_{r+1} := \frac{1}{(2r+2)!}\int_0^N \big(S_{2r+2}(0) - S_{2r+2}(t)\big) g^{(2r+2)}(t)\,dt, \tag{6.78}$$

denn

$$\mathcal{T}_1(h) = h\big(\frac{g(0)}{2} + g(1) + \dots + g(N-1) + \frac{g(N)}{2}\big),$$

$$\int_a^b f(x)\,dx = h\int_0^N g(t)\,dt, \qquad f^{(j)}(a+th)h^j = g^{(j)}(t), \qquad 0 \le t \le N.$$

Es soll nun die Identität (6.77)–(6.78) nachgewiesen werden:

$$\begin{aligned}\tfrac{1}{2}(g(1)+g(0)) - \int_0^1 g(t)\,dt &= B_1(t)g(t)\big|_{t=0}^{t=1} - \int_0^1 B_0(t)g(t)\,dt \\ &= \int_0^1 B_1(t)g'(t)\,dt = \int_0^1 S_1(t)g'(t)\,dt,\end{aligned}$$

und analog gilt

$$\tfrac{1}{2}(g(j+1)+g(j)) - \int_j^{j+1} g(t)\,dt = \int_j^{j+1} S_1(t)g'(t)\,dt, \qquad j = 0,1,\ldots,N-1,$$

so dass man

$$\tfrac{g(0)}{2} + g(1) + \ldots + g(N-1) + \tfrac{g(N)}{2} - \int_0^N g(t)\,dt = \int_0^N S_1(t)g'(t)\,dt$$

erhält. Das letzte Integral wird weiter partiell integriert,

$$\begin{aligned}\int_0^N S_1(t)g'(t)\,dt &= \tfrac{1}{2!}S_2(t)g'(t)\big|_{t=0}^{t=N} - \tfrac{1}{2!}\int_0^N S_2(t)g''(t)\,dt \\ &= \tfrac{B_2(0)}{2!}(g'(N)-g'(0)) - \tfrac{1}{2!}\int_0^N S_2(t)g''(t)\,dt,\end{aligned}$$

und partielle Integration des letzten Integrals liefert wiederum

$$\begin{aligned}-\tfrac{1}{2!}\int_0^N S_2(t)g''(t)\,dt &= -\tfrac{1}{3!}S_3(t)g''(t)\big|_{t=0}^{t=N} + \tfrac{1}{3!}\int_0^N S_3(t)g'''(t)\,dt \\ &= \underbrace{-\tfrac{B_3(0)}{3!}}_{=0}(g''(N)-g''(0)) + \tfrac{1}{3!}\int_0^N S_3(t)g'''(t)\,dt \\ &= \tfrac{1}{3!}\int_0^N S_3(t)g'''(t)\,dt.\end{aligned}$$

Wiederholte partielle Integration liefert schließlich die Identität (6.77) mit der folgenden Konstanten,

$$\begin{aligned}C_{r+1} &= \tfrac{1}{(2r+2)!}S_{2r+2}(0)[g^{(2r+1)}(N)-g^{(2r+1)}(0)] - \tfrac{1}{(2r+2)!}\int_0^N S_{2r+2}(t)g^{(2r+2)}(t)\,dt \\ &= \tfrac{1}{(2r+2)!}S_{2r+2}(0)\int_0^N g^{(2r+2)}(t)\,dt - \tfrac{1}{(2r+2)!}\int_0^N S_{2r+2}(t)g^{(2r+2)}(t)\,dt \\ &= \tfrac{1}{(2r+2)!}\int_0^N \big(S_{2r+2}(0)-S_{2r+2}(t)\big)g^{(2r+2)}(t)\,dt,\end{aligned}$$

was mit der Setzung (6.78) übereinstimmt. □

Weitere Themen und Literaturhinweise

Eine Auswahl existierender Lehrbücher mit Abschnitten über numerische Integration bildet Freund/Hoppe [31], Hämmerlin/Hoffmann [48], Kress [63], Krommer/Überhuber [64], Oevel [78] und Werner [111]. Insbesondere in [64] werden viele weitere Themen wie die numerische Berechnung uneigentlicher und mehrdimensionaler Integrale beziehungsweise die symbolische Integration behandelt. Orthogonale Polynome werden ausführlich in Hanke-Bourgeois [52] behandelt.

Übungsaufgaben

Aufgabe 6.1. Gegeben sei eine Unterteilung $\Delta : a \le x_0 < x_1 < \ldots < x_n \le b$ des Intervalls $[a,b]$. Man zeige, dass es eindeutig bestimmte Zahlen $a_0, a_1, \ldots, a_n \in \mathbb{R}$ gibt mit

$$\sum_{k=0}^{n} a_k \mathcal{P}(x_k) \;=\; \int_a^b \mathcal{P}(x)\,dx \qquad \text{für alle} \quad \mathcal{P} \in \Pi_n.$$

Aufgabe 6.2. Zu einer beliebigen Unterteilung $a \le x_0 < \ldots < x_n \le b$ des Intervalls $[a,b]$ bezeichne $\mathcal{I}_n(f) \;=\; (b-a)\sum_{k=0}^{n} \sigma_k f(x_k)$ eine Quadraturformel. Man zeige, dass ihr Genauigkeitsgrad $\le 2n+1$ ist, es gibt also ein Polynom $\mathcal{P} \in \Pi_{2n+2}$ mit $\mathcal{I}_n(\mathcal{P}) \ne \int_a^b \mathcal{P}(x)\,dx$.

Aufgabe 6.3. Man bestimme die Koeffizienten $a_0, a_1, a_2 \in \mathbb{R}$ durch Taylorabgleich so, dass die Quadraturformel $Qf = a_0 f(a) + a_1 f(\frac{a+b}{2}) + a_2 f(b)$ zur näherungsweisen Berechnung des Integrals $\int_a^b f(x)\,dx$ einen möglichst hohen Genauigkeitsgrad besitzt.

Aufgabe 6.4. Zu einer periodischen stetigen Funktion $f : \mathbb{R} \to \mathbb{R}$ und den Stützstellen $x_j = 2\pi j/(N+1)$ mit $j = 0, 1, \ldots, N$ für gerades $N \in \mathbb{N}$ bezeichne Tf das interpolierende trigonometrische Polynom von der Form $(Tf)(x) \;=\; \frac{A_0}{2} + \sum_{k=1}^{N/2}(A_k \cos kx + B_k \sin kx)$. Weiter bezeichne $Qf := \int_0^{2\pi} (Tf)(x)\,dx$. Man zeige, dass sich Qf schreiben lässt als $Qf = \sum_{k=0}^{N} a_k f(x_k)$ mit (von f unabhängigen) *positiven* Gewichten $a_k > 0$ für $k = 0, 1, \ldots, N$.

Aufgabe 6.5. Man weise mithilfe der euler-maclaurinschen Summenformel für $N \in \mathbb{N}$ die folgende Identität nach,

$$\sum_{k=1}^{N} k^3 \;=\; \big(\tfrac{N(N+1)}{2}\big)^2.$$

Aufgabe 6.6. Das Funktionensystem $(U_n)_{n\in\mathbb{N}_0}$ der Tschebyscheff-Polynome der zweiten Art bildet bezüglich des Skalarprodukts $\langle u, v\rangle \;=\; \int_{-1}^{1} u(x)\,v(x)\,\sqrt{1-x^2}\,dx$ ein Orthogonalsystem.

Aufgabe 6.7 (*Numerische Aufgabe*). Man berechne die vier bestimmten Integrale

$$\int_0^{0.5} \frac{1}{16x^2+1}\,dx, \qquad \int_0^2 e^{-x^2}\,dx, \qquad \int_0^{\pi/2} \big(\cos\tfrac{x}{2}\big)^2 \sin 3x\,dx, \qquad \int_0^{\pi/2} \sqrt{|\cos 2x|}\,dx,$$

numerisch durch Extrapolation der Trapezsummen $\mathcal{T}_1(h_j)$ unter Anwendung der Romberg-Schrittweite $h_0 = b-a$ und $h_j = h_{j-1}/2$ für $j = 1, 2, \ldots$. Genauer: mit den Bezeichnungen aus (6.40)–(6.41) mit $\mathcal{T} = \mathcal{T}_1$ und $\gamma = 2$ berechne man für $k = 0, 1, \ldots$ die Werte

$$T_{k-m,\ldots,k} \qquad \text{für} \quad m \;=\; 0, 1, \ldots, \min\{k, m_*\}. \tag{6.79}$$

Man breche mit $k =: k_*$ ab, falls

$$m_* + 1 \;\le\; k \;\le\; 12, \qquad |T_{k-m_*,\ldots,k} \;-\; T_{k-m_*+1,\ldots,k}| \;\le\; \varepsilon$$

oder aber

$$k = 13$$

erfüllt ist (mit $m_* = 4$ und $\varepsilon = 10^{-8}$). Man gebe für jedes der vier zu berechnenden Integrale die Werte (6.79) für $k = 0, 1, \ldots, k_*$ in einem Tableau aus, jeweils auf acht Nachkommastellen genau.

7 Explizite Einschrittverfahren für Anfangswertprobleme bei gewöhnlichen Differenzialgleichungen

Viele Anwendungen wie beispielsweise die Berechnung der Flugbahn eines Raumfahrzeugs beim Wiedereintritt in die Erdatmosphäre oder Räuber-Beute-Modelle führen auf Anfangswertprobleme für Systeme von gewöhnlichen Differenzialgleichungen. Ebenso resultieren gewisse Diskretisierungen von Anfangswertproblemen für partielle Differenzialgleichungen in Anfangswertproblemen für Systeme von gewöhnlichen Differenzialgleichungen. Ein konkretes Beispiel hierzu wird in Abschnitt 8.9.4 auf Seite 229 vorstellt. Solche Anfangswertprobleme für Systeme von gewöhnlichen Differenzialgleichungen sind Gegenstand des vorliegenden und des nächsten Kapitels.

Definition 7.1. Ein *Anfangswertproblem für ein System von N gewöhnlichen Differenzialgleichungen 1. Ordnung* ist von der Form

$$y' = f(t,y), \qquad t \in [a,b], \tag{7.1}$$

$$y(a) = y_0, \tag{7.2}$$

mit einem gegebenen endlichen Intervall $[a,b]$, einem Vektor $y_0 \in \mathbb{R}^N$ und einer Funktion

$$f : [a,b] \times \mathbb{R}^N \to \mathbb{R}^N, \tag{7.3}$$

und gesucht ist eine differenzierbare Funktion $y : [a,b] \to \mathbb{R}^N$ mit den Eigenschaften (7.1)–(7.2).

Die Notation in (7.1) ist eine übliche Kurzform für $y'(t) = f(t,y(t))$, $t \in [a,b]$. Differenzierbarkeit bedeutet hier komponentenweise Differenzierbarkeit, und es ist $y'(t) = (y_1'(t), \ldots, y_N'(t))^\top \in \mathbb{R}^N$.

7.1 Ein Existenz- und Eindeutigkeitssatz

Die Existenz und Eindeutigkeit der Lösung ist auch bei Anfangswertproblemen für Systeme von gewöhnlichen Differenzialgleichungen eine grundlegende Fragestellung. Diese ist Gegenstand des nächsten Theorems, wobei die folgende Lipschitzbedingung für Funktionen f von der Form (7.3) eine wesentliche Rolle spielt,

$$\| f(t,u) - f(t,v) \| \le L \| u - v \|, \qquad t \in [a,b], \qquad u, v \in \mathbb{R}^N, \tag{7.4}$$

mit einer Konstanten $L > 0$, wobei hier und im Folgenden $\| \cdot \| : \mathbb{R}^N \to \mathbb{R}$ eine beliebige Vektornorm bezeichnet.

Neben der angesprochenen Existenz- und Eindeutigkeitsaussage für Anfangswertprobleme von der Form (7.1)–(7.2) liefert das folgende Theorem ein ebenso wichtiges Resultat zur stetigen Abhängigkeit von den Anfangswerten.

Theorem 7.2. *Es sei $f : [a,b] \times \mathbb{R}^N \to \mathbb{R}^N$ eine stetige Funktion, die die Lipschitzbedingung (7.4) erfülle. Dann gelten die beiden folgenden Aussagen:*

(a) (Picard/Lindelöf) Das Anfangswertproblem (7.1)–(7.2) besitzt genau eine stetig differenzierbare Lösung $y : [a,b] \to \mathbb{R}^N$.

(b) Für differenzierbare Funktionen $y,\ \widehat{y} : [a,b] \to \mathbb{R}^N$ mit

$$\begin{aligned} y' &= f(t,y), \qquad t \in [a,b]; \qquad & y(a) &= y_0 \\ \widehat{y}' &= f(t,\widehat{y}), \qquad \text{---}\,\text{“}\,\text{---} & \widehat{y}(a) &= \widehat{y}_0 \end{aligned}$$

gilt die Abschätzung

$$\| y(t) - \widehat{y}(t) \| \ \leq \ e^{L(t-a)} \| y_0 - \widehat{y}_0 \|, \qquad t \in [a,b]. \tag{7.5}$$

Einen Beweis hierzu finden Sie beispielsweise in Heuser [54], Abschnitt 12. Auch unter anderen Voraussetzungen an die Funktion f sind Existenz- und Eindeutigkeitsaussagen für das Anfangswertproblem (7.1)–(7.2) möglich. Zur Vereinfachung der Notation wird Folgendes angenommen:

> In diesem und dem folgenden Kapitel 8 wird ohne weitere Spezifikation an die Funktion f angenommen, dass jedes der betrachteten Anfangswertprobleme von der Form (7.1)–(7.2) jeweils eine eindeutig bestimmte Lösung $y : [a,b] \to \mathbb{R}^N$ besitzt.

An einigen Stellen erweist sich das folgende Resultat über die Glattheit der Lösung des Anfangswertproblems (7.1)–(7.2) als nützlich, das man mit der Kettenregel erhält.

Theorem 7.3. *Für eine p-mal stetig partiell differenzierbare Funktion mit $p \geq 1$ ist die Lösung des Anfangswertproblems (7.1)–(7.2) mindestens $(p+1)$-mal stetig partiell differenzierbar.*

Bemerkung 7.4. In der Situation von Theorem 7.3 lassen sich die höheren Ableitungen der Lösung angeben. Beispielsweise berechnet man im eindimensionalen Fall $N = 1$ sowie für $p = 1$ sofort Folgendes:

$$y''(t) \ = \ \frac{\partial f}{\partial t}(t,y(t)) \ + \ \frac{\partial f}{\partial y}(t,y(t))y'(t) \ = \ \Big(\frac{\partial f}{\partial t} \ + \ \frac{\partial f}{\partial y} f\Big)(t,y(t)). \tag{7.6}$$

Δ

In den meisten Fällen lässt sich die Lösung des Anfangswertproblems (7.1)–(7.2) nicht exakt berechnen, so dass man auf numerische Verfahren zurückgreift. Solche Verfahren werden in diesem und dem darauf folgenden Kapitel vorgestellt, wobei es die Zielsetzung der meisten dieser Verfahren ist, zu der Lösung $y : [a,b] \to \mathbb{R}^N$ des Anfangswertproblems (7.1)–(7.2) schrittweise für $\ell = 0, 1, \ldots$ Approximationen

$$u_\ell \approx y(t_\ell), \qquad \ell = 0, 1, \ldots, n,$$

zu gewinnen auf einem noch nicht näher spezifizierten Gitter

$$\begin{aligned} \Delta &= \{a = t_0 < t_1 < \ldots < t_n \leq b\}, \\ h_\ell &:= t_{\ell+1} - t_\ell \quad \text{für} \quad \ell = 0, 1, \ldots, n-1. \end{aligned} \tag{7.7}$$

7.2 Theorie der Einschrittverfahren

Im Folgenden werden Einschrittverfahren einführend behandelt.

Definition 7.5. Ein (explizites) *Einschrittverfahren* zur approximativen Bestimmung einer Lösung des Anfangswertproblems (7.1)–(7.2) ist von der Gestalt

$$u_{\ell+1} = u_\ell + h_\ell \varphi(t_\ell, u_\ell; h_\ell), \qquad \ell = 0, 1, \ldots, n-1; \qquad u_0 := y_0 \tag{7.8}$$

mit einer *Verfahrensfunktion* $\varphi : [a,b] \times \mathbb{R}^N \times \mathbb{R}_+ \to \mathbb{R}^N$ und einem noch nicht näher spezifizierten Gitter beziehungsweise Schrittweiten der Form (7.7).

Bemerkung 7.6. (1) Die Approximation u_ℓ hängt von $u_{\ell-1}$ nicht jedoch (unmittelbar) von $u_{\ell-2}$, $u_{\ell-3}, \ldots$ ab, was die Bezeichnung "Einschrittverfahren" rechtfertigt. Im anschließenden Kapitel 8 werden dann Mehrschrittverfahren behandelt.

(2) Ein Einschrittverfahren ist durch seine Verfahrensfunktion φ festgelegt, die Schrittweiten hingegen sind noch frei wählbar. Zur Vereinfachung der Notation wird dennoch im Folgenden bei Einschrittverfahren auf die Verfahrensvorschrift (7.8) verwiesen, obwohl Eigenschaften von φ behandelt werden.

(3) Ebenfalls zwecks einer vereinfachten Notation wird als Definitionsbereich einer Verfahrensfunktion φ immer $[a,b] \times \mathbb{R}^N \times \mathbb{R}_+$ angegeben, obwohl bei den meisten noch vorzustellenden speziellen Einschrittverfahren der Ausdruck $\varphi(t,u;h)$ lediglich für Schrittweiten $h \leq b - t$ wohldefiniert ist.

(4) Eine wichtige Rolle spielen in der Praxis auch *implizite* Einschrittverfahren, die durch die Definition (7.8) nicht unmittelbar erfasst sind. Solche impliziten Einschrittverfahren werden gemeinsam mit den Mehrschrittverfahren in Kapitel 8 behandelt.

△

Die wichtigste Kennzahl eines Einschrittverfahrens ist seine Konvergenzordnung:

Definition 7.7. Ein Einschrittverfahren (7.8) zur Lösung des Anfangswertproblems $y' = f(t,y),\ y(a) = y_0$ besitzt die *Konvergenzordnung* $p \geq 1$, falls sich der *globale Verfahrensfehler* abschätzen lässt in der Form

$$\max_{\ell=0,\ldots,n} \|u_\ell - y(t_\ell)\| \ \leq\ C h_{\max}^p, \qquad h_{\max} := \max_{\ell=0,\ldots,n-1} \{t_{\ell+1} - t_\ell\},$$

mit einer von dem gewählten Gitter Δ unabhängigen Konstanten $C \geq 0$.

Für die Bestimmung der Konvergenzordnung eines Einschrittverfahrens spielt der folgende Begriff eine maßgebliche Rolle.

Definition 7.8. Für ein Einschrittverfahren (7.8) zur Lösung des Anfangswertproblems $y' = f(t,y),\ y(a) = y_0$ bezeichnet

$$\eta(t,h) := \underbrace{y(t) + h\varphi(t,y(t);h)}_{\text{Verfahrensvorschrift}} - y(t+h) \qquad \text{für} \quad t \in [a,b], \qquad 0 \leq h \leq b-t,$$

den *lokalen Verfahrensfehler im Punkt* $(t+h, y(t+h))$ bezüglich der Schrittweite h.

Andere sinnvolle Definitionen des lokalen Verfahrensfehlers sind ebenfalls möglich (siehe Aufgabe 7.3).

Definition 7.9. Ein Einschrittverfahren (7.8) zur Lösung des Anfangswertproblems $y' = f(t,y),\ y(a) = y_0$ besitzt die *Konsistenzordnung* $p \geq 1$, falls für den lokalen Verfahrensfehler die Ungleichung

$$\|\eta(t,h)\| \ \leq\ C h^{p+1} \qquad \text{für} \quad t \in [a,b], \qquad 0 \leq h \leq b-t, \tag{7.9}$$

erfüllt ist mit einer (von t und h unabhängigen) Konstanten $C \geq 0$.

Die *Konsistenzordnung* bezeichnet man oft nur kurz als *Ordnung* eines Einschrittverfahrens. Es wird nun die wesentliche Abschätzung für den bei Einschrittverfahren auftretenden globalen Verfahrensfehler vorgestellt, wofür die folgende Lipschitzbedingung an die Verfahrensfunktion benötigt wird,

$$\left.\begin{aligned} \|\varphi(t,u;h) - \varphi(t,v;h)\| \ \leq\ L_\varphi \|u - v\| \quad \text{für} \ \ t \in [a,b], \quad 0 < h \leq b-t, \\ u,\ v \ \in \ \mathbb{R}^N. \end{aligned}\right\} \tag{7.10}$$

Bei allen in diesem Kapitel vorzustellenden speziellen Einschrittverfahren ist eine solche Lipschitzbedingung (7.10) erfüllt, falls die Funktion f der Lipschitzbedingung (7.4) genügt.

Theorem 7.10. *Ein Einschrittverfahren (7.8) zur Lösung des Anfangswertproblems $y' = f(t,y),\ y(a) = y_0$ besitze die Konsistenzordnung $p \geq 1$ und erfülle die Lipschitzbedingung (7.10). Dann liegt die Konvergenzordnung p vor. Genauer gilt*

$$\max_{\ell=0,\ldots,n} \|u_\ell - y(t_\ell)\| \ \leq\ K h_{\max}^p, \qquad h_{\max} := \max_{\ell=0,\ldots,n-1} \{t_{\ell+1} - t_\ell\}, \tag{7.11}$$

mit der Konstanten $K = \frac{C}{L_\varphi}(e^{L_\varphi(b-a)} - 1)$, wobei C aus der Abschätzung (7.9) herrührt.

BEWEIS. Mit den Setzungen

$$\begin{aligned} e_\ell &= u_\ell - y_\ell, \qquad y_\ell := y(t_\ell), \qquad && \ell = 0,1,\dots,n, \\ \eta_\ell &= \eta(t_\ell, h_\ell), && \ell = 0,1,\dots,n-1, \end{aligned}$$

gilt für $\ell = 0,1,\dots,n-1$

$$\begin{aligned} y_{\ell+1} &= y_\ell + h_\ell \varphi(t_\ell, y_\ell; h_\ell) - \eta_\ell, \\ u_{\ell+1} &= u_\ell + h_\ell \varphi(t_\ell, u_\ell; h_\ell), \end{aligned}$$

und daher

$$e_{\ell+1} = e_\ell + h_\ell\big(\varphi(t_\ell, u_\ell; h_\ell) - \varphi(t_\ell, y_\ell; h_\ell)\big) + \eta_\ell$$

beziehungsweise

$$\begin{aligned} \|e_{\ell+1}\| &\le \|e_\ell\| + h_\ell \|\varphi(t_\ell, u_\ell; h_\ell) - \varphi(t_\ell, y_\ell; h_\ell)\| + \|\eta_\ell\| \\ &\le (1 + h_\ell L_\varphi)\|e_\ell\| + h_\ell C h_{\max}^p, \end{aligned}$$

und das nachfolgende Lemma 7.12 liefert wegen $e_0 = 0$ unmittelbar die Aussage des Theorems. □

Bemerkung 7.11. Lipschitzbedingung (7.10) und Konsistenzordnung p zusammen gewährleisten also die Konvergenzordnung p des Einschrittverfahrens (7.8). △

7.2.1 Ein elementares Resultat zur Fehlerakkumulation

Lemma 7.12. *Für Zahlen $L > 0$, $a_\ell \ge 0$, $h_\ell > 0$ und $b \ge 0$ sei*

$$a_{\ell+1} \le (1 + h_\ell L) a_\ell + h_\ell b, \qquad \ell = 0,1,\dots,n-1,$$

erfüllt. Dann gelten die Abschätzungen

$$a_\ell \le \frac{e^{Lx_\ell} - 1}{L} b + e^{Lx_\ell} a_0 \qquad \text{mit} \quad x_\ell := \sum_{j=0}^{\ell-1} h_j \qquad (\ell = 0,1,\dots,n).$$

BEWEIS. Der Fall $\ell = 0$ ist klar, und den Induktionsschritt $\ell \to \ell + 1$ führt man wie folgt:

$$\begin{aligned} a_{\ell+1} &\le \overbrace{(1 + h_\ell L)}^{\le\, e^{h_\ell L}} \big(\tfrac{e^{Lx_\ell} - 1}{L} b + e^{Lx_\ell} a_0\big) + h_\ell b \\ &\le \big(\frac{e^{L(x_\ell + h_\ell)} - 1 - h_\ell L}{L} + h_\ell\big) b + e^{L(x_\ell + h_\ell)} a_0 \\ &= \frac{e^{Lx_{\ell+1}} - 1}{L} b + e^{Lx_{\ell+1}} a_0. \end{aligned}$$

□

7.3 Spezielle Einschrittverfahren

7.3.1 Einschrittverfahren der Konsistenzordnung $p = 1$

Beispiel 7.13. Das *Euler-Verfahren* ist von der Form

$$u_{\ell+1} \;=\; u_\ell \;+\; h_\ell f(t_\ell, u_\ell), \qquad \ell \;=\; 0, 1, \ldots, n-1; \qquad u_0 \;:=\; y_0. \tag{7.12}$$

Andere übliche Bezeichnungen für das Verfahren (7.12) sind *eulersches Polygonzugverfahren* oder *vorwärtsgerichtete Euler-Formel*.

In Bild 7.1 ist die Vorgehensweise des Euler-Verfahrens veranschaulicht. Dabei stellen die Funktionen y, $\widehat{y}$ beziehungsweise $\overline{y}$ Lösungen der Differenzialgleichung $y' = f(t, y)$ dar mit den Anfangswerten $y(t_0) = y_0$, $\widehat{y}(t_1) = u_1$ beziehungsweise $\overline{y}(t_2) = u_2$. Die gestrichelten Linien stellen Tangenten dar und illustrieren die Bestimmung der jeweils nächsten Approximation.

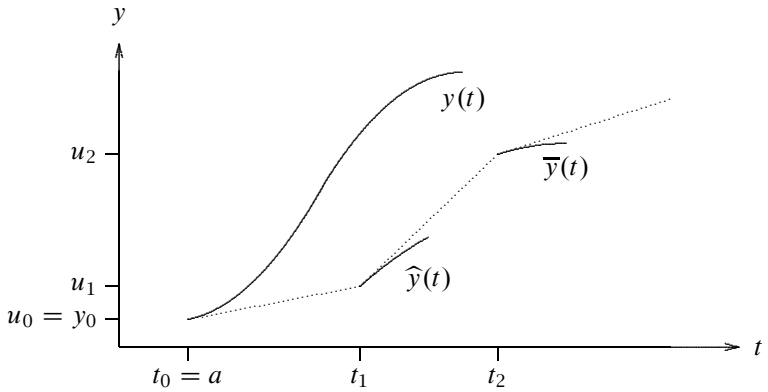


Bild 7.1: Vorgehensweise beim Euler-Verfahren △

Theorem 7.14. *Für eine stetig partiell differenzierbare Funktion $f : [a, b] \times \mathbb{R}^N \to \mathbb{R}^N$ besitzt das Euler-Verfahren die Konsistenzordnung $p = 1$.*

BEWEIS. Eine Taylorentwicklung der Lösung des Anfangswertproblems $y' = f(t, y)$, $y(a) = y_0$ liefert

$$y(t+h) \;=\; y(t) \;+\; y'(t)h \;+\; (y_j''(\tau_j))_{j=1}^N \frac{h^2}{2}$$

mit geeigneten Zwischenstellen $\tau_j \in [a, b]$, und daraus erhält man für den lokalen Verfahrensfehler

$$\eta(t,h) \;=\; y(t) \;+\; h \underbrace{f(t, y(t))}_{=\, y'(t)} \;-\; y(t+h) \;=\; -(y_j''(\tau_j))_{j=1}^N \frac{h^2}{2}$$

beziehungsweise

$$\|\eta(t,h)\|_\infty \;\le\; Ch^2, \qquad \text{mit} \quad C \;=\; \tfrac{1}{2}\max_{\tau\in[a,b]}\|y''(\tau)\|_\infty,$$

wobei die zweimalige stetige Differenzierbarkeit der Lösung y aus Theorem 7.3 folgt. □

7.3.2 Einschrittverfahren der Konsistenzordnung $p = 2$

Zur Herleitung von Einschrittverfahren (7.8) der Konsistenzordnung $p = 2$ wird für die Verfahrensfunktion der Ansatz

$$\left.\begin{aligned}\varphi(t,u;h) \;&=\; a_1 f(t,u) \;+\; a_2 f(t \;+\; b_1 h, u \;+\; b_2 h f(t,u)),\\ &\qquad t\in[a,b],\quad 0\le h\le b-t,\quad u\in\mathbb{R}^N,\end{aligned}\right\} \tag{7.13}$$

betrachtet mit noch festzulegenden Konstanten $a_j,\ b_j \in \mathbb{R}$.

Theorem 7.15. *Ein Einschrittverfahren (7.8) mit einer Verfahrensfunktion der Form (7.13) ist konsistent von der Ordnung $p = 2$, falls die Funktion $f : [a,b]\times\mathbb{R}^N \to \mathbb{R}^N$ zweimal stetig partiell differenzierbar ist und für die Koeffizienten Folgendes gilt:*

$$a_1 \;+\; a_2 \;=\; 1, \qquad a_2 b_1 \;=\; \tfrac{1}{2}, \qquad a_2 b_2 \;=\; \tfrac{1}{2}. \tag{7.14}$$

BEWEIS. Der Beweis wird für den eindimensionalen Fall $N = 1$ geführt. Taylorentwicklungen sowohl von $\varphi(t,y(t);\cdot)$ im Punkt $h = 0$ als auch von der Lösung y in t zusammen mit Theorem 7.3 ergeben

$$\begin{aligned}\varphi(t,y(t);h) \;&=\; \big[\,\overbrace{(a_1+a_2)}^{=\,1} f \;+\; h\big(\overbrace{a_2b_1}^{=1/2}\tfrac{\partial f}{\partial t} \;+\; \overbrace{a_2b_2}^{=1/2} f\tfrac{\partial f}{\partial y}\big)\big](t,y(t)) \;+\; \mathcal{O}(h^2),\\ y(t+h) \;&=\; y(t) \;+\; y'(t)h \;+\; y''(t)\tfrac{h^2}{2} \;+\; \mathcal{O}(h^3)\\ &\overset{(7.6)}{=}\; y(t) \;+\; \underbrace{\big[hf \;+\; \big(\tfrac{\partial f}{\partial t} \;+\; f\tfrac{\partial f}{\partial y}\big)\tfrac{h^2}{2}\big](t,y(t))}_{=\;h\varphi(t,y(t);h)\;+\;\mathcal{O}(h^3)} \;+\; \mathcal{O}(h^3),\end{aligned}$$

woraus für den lokalen Verfahrensfehler unmittelbar

$$\eta(t,h) \;=\; y(t) \;+\; h\varphi(t,y(t);h) \;-\; y(t+h) \;=\; \mathcal{O}(h^3)$$

folgt. □

Bemerkung 7.16. Der eine Freiheitsgrad in (7.14) kann nicht zur Gewinnung eines Verfahrens der Konsistenzordnung $p = 3$ verwendet werden. △

Es werden nun zwei Beispiele für Einschrittverfahren von der Form (7.13) vorgestellt.

Beispiel 7.17. Die Verfahrensfunktion für das *modifizierte Euler-Verfahren* lautet

$$\varphi(t,u;h) \;=\; f(t \,+\, \tfrac{h}{2}, u \,+\, \tfrac{h}{2} f(t,u)), \qquad t \in [a,b], \quad 0 \le h \le b-t, \quad u \in \mathbb{R}^N,$$

wobei φ aus dem Ansatz (7.13) hervorgeht für $a_1 = 0,\ a_2 = 1$ und $b_1 = b_2 = 1/2$, und das zugehörige Einschrittverfahren (7.8) besitzt nach Theorem 7.15 für eine hinreichend glatte Funktion f daher die Konsistenzordnung $p = 2$. Das Verfahren selbst lässt sich folgendermaßen formulieren,

$$u_{\ell+1/2} \;=\; u_\ell \,+\, \tfrac{h_\ell}{2} f(t_\ell, u_\ell), \qquad t_{\ell+1/2} := t_\ell \,+\, \tfrac{h_\ell}{2},$$

$$u_{\ell+1} \;=\; u_\ell \,+\, h_\ell f(t_{\ell+1/2}, u_{\ell+1/2}), \qquad \ell \;=\; 0,1,\ldots,n-1.$$

Die Wirkungsweise des modifizierten Euler-Verfahrens ist in Bild 7.2 veranschaulicht. Dabei stellen die Funktionen $y,\ \widehat{y},\ \overline{y}$ beziehungsweise $\widetilde{y}$ Lösungen der Differenzialgleichung $y' = f(t,y)$ dar mit den Anfangswerten $y(t_0) = y_0,\ \widehat{y}(t_{1/2}) = u_{1/2},\ \overline{y}(t_1) = u_1$ beziehungsweise $\widetilde{y}(t_{3/2}) = u_{3/2}$. Die Näherung u_1 erhält man von u_0 ausgehend auf einer Geraden der Steigung $\widehat{y}'(t_{1/2})$.

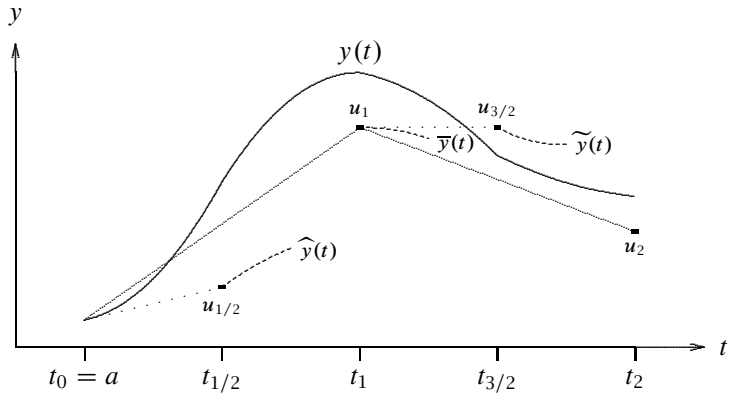


Bild 7.2: Vorgehensweise beim modifizierten Euler-Verfahren △

Beispiel 7.18. Die Verfahrensfunktion für das *Verfahren von Heun* lautet

$$\varphi(t,u;h) \;=\; \tfrac{1}{2}\big[\, f(t,u) \,+\, f(t \,+\, h, u \,+\, h f(t,u))\,\big], \quad t \in [a,b], \quad 0 \le h \le b-t, \quad u \in \mathbb{R}^N,$$

wobei φ aus der allgemeinen Form (7.13) hervorgeht für $a_1 = a_2 = 1/2$ und $b_1 = b_2 = 1$. Das zugehörige Einschrittverfahren (7.8) besitzt also für eine hinreichend glatte Funktion f ebenfalls die Konsistenzordnung $p = 2$. Der Algorithmus selbst

lässt sich folgendermaßen formulieren,

$$\begin{aligned} v_{\ell+1} &= u_\ell + h_\ell f(t_\ell, u_\ell), \\ w_{\ell+1} &= u_\ell + h_\ell f(t_{\ell+1}, v_{\ell+1}), \end{aligned} \qquad u_{\ell+1} = \tfrac{1}{2}(v_{\ell+1} + w_{\ell+1}), \quad \ell = 0, 1, \ldots, n-1.$$

△

7.3.3 Einschrittverfahren der Konsistenzordnung $p = 4$

Beispiel 7.19. Die Verfahrensfunktion für das *klassische Runge-Kutta-Verfahren* lautet

$$\varphi(t, u; h) = \tfrac{1}{6}[k_1 + 2k_2 + 2k_3 + k_4], \qquad t \in [a, b], \quad 0 \le h \le b - t, \quad u \in \mathbb{R}^N,$$

mit

$$\begin{aligned} k_1 &:= f(t, u), & k_2 &:= f(t + \tfrac{h}{2}, u + \tfrac{h}{2}k_1), \\ k_3 &:= f(t + \tfrac{h}{2}, u + \tfrac{h}{2}k_2), & k_4 &:= f(t + h, u + hk_3). \end{aligned}$$

Durch Taylorentwicklung lässt sich nachweisen, dass das klassische Runge-Kutta-Verfahren für eine hinreichend oft differenzierbare Funktion f die Konsistenzordnung $p = 4$ besitzt. △

Bei jedem der vorgestellten speziellen expliziten Einschrittverfahren ist für die Anwendbarkeit des Konvergenzresultats aus Theorem 7.10 jeweils noch die Lipschitzeigenschaft (7.10) nachzuprüfen. Hier stellt man leicht fest, dass diese Lipschitzbedingung (7.10) jeweils genau dann erfüllt ist, wenn die Funktion f der Lipschitzbedingung (7.4) genügt.

7.4 Rundungsfehleranalyse

In diesem Abschnitt 7.4 werden die Auswirkungen von fehlerbehafteten Anfangswerten und Rundungsfehlern bei Einschrittverfahren (7.8) untersucht. Hierzu sei im Folgenden angenommen, dass eine fehlerbehaftete Verfahrensvorschrift von der folgenden Form

$$\left.\begin{aligned} v_{\ell+1} &= v_\ell + h_\ell \varphi(t_\ell, v_\ell; h_\ell) + \rho_\ell, \quad \ell = 0, \ldots, n-1; \quad v_0 := y_0 + e_0, \\ \|\rho_\ell\| &\le \delta, \quad \text{———} \text{ « } \text{———} \quad \|e_0\| \le \varepsilon, \end{aligned}\right\} \tag{7.15}$$

vorliegt mit gewissen Vektoren $e_0, \rho_\ell \in \mathbb{R}^N$, und $\|\cdot\|$ bezeichnet eine nicht weiter spezifizierte Vektornorm.

Theorem 7.20. *Zur Lösung des Anfangswertproblems $y' = f(t, y)$, $y(a) = y_0$ sei durch (7.8) ein Einschrittverfahren mit der Konsistenzordnung $p \ge 1$ gegeben, das die*

Lipschitzbedingung (7.10) erfülle. Dann gelten für die durch die fehlerbehaftete Verfahrensvorschrift von der Form (7.15) gewonnenen Approximationen die folgenden Abschätzungen,

$$\max_{\ell=0,\dots,n} \|v_\ell - y(t_\ell)\| \;\le\; K\big(h_{\max}^p + \frac{\delta}{h_{\min}}\big) + e^{L_\varphi(b-a)}\varepsilon \tag{7.16}$$

$$\text{mit}\quad h_{\max} = \max_{\ell=0,\dots,n-1} h_\ell, \qquad h_{\min} = \min_{\ell=0,\dots,n-1} h_\ell,$$

mit der Konstanten $K := \frac{\max\{C,1\}}{L_\varphi}[e^{L_\varphi(b-a)} - 1]$*, für* C *aus (7.9).*

BEWEIS. Die Vorgehensweise im Beweis von Theorem 7.10 ist nur geringfügig zu modifizieren: mit den Setzungen

$$\begin{aligned} e_\ell &= v_\ell - y_\ell, \qquad y_\ell := y(t_\ell), \qquad && \ell = 0,1,\dots,n,\\ \eta_\ell &= \eta(t_\ell, h_\ell), && \ell = 0,1,\dots,n-1, \end{aligned}$$

gilt für $\ell = 0,1,\dots,n-1$

$$\begin{aligned} y_{\ell+1} &= y_\ell + h_\ell\varphi(t_\ell, y_\ell; h_\ell) - \eta_\ell,\\ v_{\ell+1} &= v_\ell + h_\ell\varphi(t_\ell, v_\ell; h_\ell) + \rho_\ell, \end{aligned}$$

und daher

$$e_{\ell+1} = e_\ell + h_\ell[\varphi(t_\ell, v_\ell; h_\ell) - \varphi(t_\ell, y_\ell; h_\ell)] + \rho_\ell + \eta_\ell$$

beziehungsweise

$$\begin{aligned} \|e_{\ell+1}\| &\le \|e_\ell\| + h_\ell\|\varphi(t_\ell, v_\ell; h_\ell) - \varphi(t_\ell, y_\ell; h_\ell)\| + \|\eta_\ell\| + \|\rho_\ell\|\\ &\le (1 + h_\ell L_\varphi)\|e_\ell\| + h_\ell\big(Ch_{\max}^p + \frac{\delta}{h_{\min}}\big), \end{aligned}$$

und Korollar 7.12 liefert zusammen mit der Abschätzung $\|e_0\| \le \varepsilon$ unmittelbar die Aussage des Theorems. □

Bemerkung 7.21. Die rechte Seite in der Abschätzung (7.16) setzt sich aus drei Termen zusammen: der erste Term $Kh_{\max}^p$ resultiert aus dem globalen Verfahrensfehler des Einschrittverfahrens, und der zweite Term $\delta/h_{\min}$ korrespondiert zu den akkumulierten Rundungsfehlern. Der Term $e^{L_\varphi(b-a)}\varepsilon$ schließlich rührt von einem fehlerbehafteten Anfangswert her. △

Als unmittelbare Folgerung aus Theorem 7.20 erhält man im Fall eines exakt gegebenen Anfangswerts ($\varepsilon = 0$) und konstanter Schrittweite:

Korollar 7.22. *Es liege die Situation aus Theorem 7.20 vor mit* $v_0 = y_0$ *und* $h_\ell = h$ *für* $\ell = 0,1,\dots,n-1$*. Dann gilt mit der Konstanten* $K := \frac{\max\{C,1\}}{L_\varphi}[e^{L_\varphi b} - 1]$ *die Fehlerabschätzung*

$$\max_{\ell=0,\dots,n} \|v_\ell - y(t_\ell)\| \;\le\; K\big(h^p + \frac{\delta}{h}\big). \tag{7.17}$$

Mit der Wahl $h = h_{\text{opt}} = (\delta/p)^{1/(p+1)}$ *erhält man*

$$\max_{\ell=0,\dots,n} \|v_\ell - y(t_\ell)\| \leq \frac{2K}{p^{p/(p+1)}} \delta^{p/(p+1)}.$$

Die Situation in Abschätzung (7.17) ist in Bild 7.3 veranschaulicht.

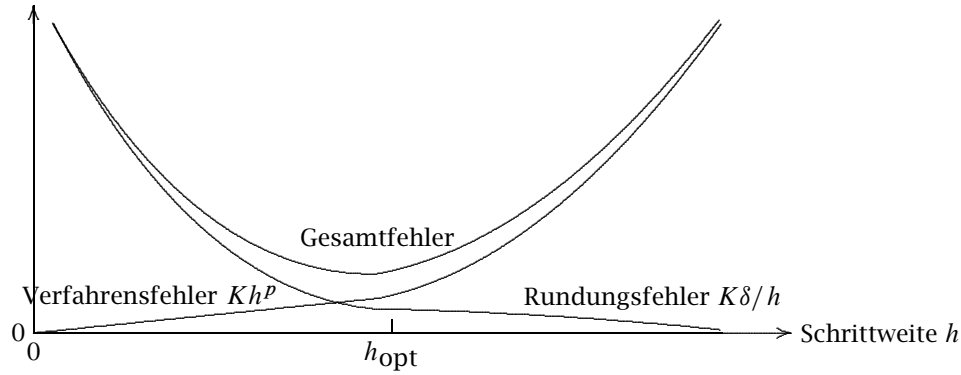


Bild 7.3: Einfluss des Rundungsfehlers in Abhängigkeit von der Schrittweite h (vergleiche Korollar 7.22)

7.5 Asymptotische Entwicklung der Approximationen

7.5.1 Einführende Bemerkungen

Zur Lösung des Anfangswertproblems $y' = f(t,y),\ y(a) = y_0$ werden in dem vorliegenden Abschnitt 7.5.1 Einschrittverfahren (7.8) bezüglich unterschiedlicher Gitter betrachtet, die der Einfachheit halber jeweils gleichabständige Knoten besitzen sollen,

$$h > 0, \qquad t_\ell = a + \ell h \qquad \text{für} \quad \ell = 0,1,\dots,n, \quad \text{mit} \quad 0 < n \leq \frac{b-a}{h}. \tag{7.18}$$

Im Folgenden ist es von Vorteil, die Schrittweitenabhängigkeit der Approximationen des Einschrittverfahrens (7.8) explizit anzugeben. Dies geschieht durch die folgende Notation,

$$u_h(t_{\ell+1}) := u_h(t_\ell) + h\varphi(t_\ell, u_h(t_\ell); h), \quad \ell = 0,\dots,n-1; \quad u_h(0) := y_0, \tag{7.19}$$

mit $t_\ell = t_\ell(h)$ entsprechend (7.18). Es ist dann

$$u_h(t) \quad \text{definiert für alle} \quad a < t \leq b, \quad h \in \mathbb{H}_t := \left\{ \frac{t-a}{m} : m = 1,2,\dots \right\}. \tag{7.20}$$

Die Funktion u_h wird als *Gitterfunktion* bezeichnet. Besitzt das zugrunde liegende Einschrittverfahren die Konsistenzordnung $p \geq 1$ und genügt die Verfahrensfunktion der Stabilitätsbedingung (7.10), so gilt nach Theorem 7.10 an jeder Stelle $a < t \leq b$

$$u_h(t) \;=\; y(t) \;+\; \mathcal{O}(h^p) \qquad \text{für} \quad \mathbb{H}_t \ni h \to 0. \tag{7.21}$$

In Abhängigkeit von der vorliegenden Konsistenzordnung und den Differenzierbarkeitseigenschaften der beteiligten Funktionen lässt sich die Darstellung (7.21) in Form einer asymptotischen Entwicklung präzisieren:

Theorem 7.23. *Bezüglich des Anfangswertproblems $y' = f(t,y),\ y(a) = y_0$ besitze eine gegebene Verfahrensfunktion $\varphi : [a,b] \times \mathbb{R}^N \times \mathbb{R}_+ \to \mathbb{R}^N$ die Konsistenzordnung $p \geq 1$ und genüge der Stabilitätsbedingung (7.10). Weiter seien die Funktionen f und φ jeweils $(p+r)$-mal stetig partiell differenzierbar. Für gewisse Koeffizientenfunktionen $c_{p+j} \in C^{r+1-j}([a,b],\mathbb{R}^N)$ mit $c_{p+j}(a) = 0$ für $j = 0,1,\ldots,r-1$ gilt dann die folgende asymptotische Entwicklung:*

$$\left.\begin{aligned} u_h(t) = y(t) + c_p(t)h^p + c_{p+1}(t)h^{p+1} + \ldots + c_{p+r-1}(t)h^{p+r-1} + \mathcal{O}(h^{p+r}), \\ t \in [a,b], \qquad h \in \mathbb{H}_t, \end{aligned}\right\} \tag{7.22}$$

wobei die angegebenen Konvergenzraten gleichmäßig in t auftreten.

Hierbei bezeichnet $C^s(D,\mathbb{R}^N)$ die Menge der s-mal stetig partiell differenzierbaren Funktionen $\psi : D \to \mathbb{R}^N$, wobei $D \subset \mathbb{R}^M$ gelte für ein $M \geq 1$. Auf der Basis solcher asymptotischer Entwicklungen lassen sich Verfahren höherer Ordnung gewinnen[1].

Die Existenz einer solchen Asymptotik ist erstmals in Gragg [38] nachgewiesen worden. In den folgenden Abschnitten 7.5.2 und 7.5.3 wird eine später entwickelte, auf Hairer/Lubich [49] und Deuflhard/Bornemann [21] basierende Methode zur Herleitung für die genannte asymptotische Entwicklung (7.22) vorgestellt.

7.5.2 Herleitung der asymptotischen Entwicklung des globalen Verfahrensfehlers, 1. Teil

Eine asymptotische Entwicklung (7.22) erhält man mittels nur zu diesem Anlass konstruierter spezieller Einschrittverfahren höherer Ordnung. Grundlage dafür bildet die folgende Rekursionsvorschrift, bei der die Verfahrensfunktion $\psi^* : [a,b] \times \mathbb{R}^N \times \mathbb{R}_+ \to \mathbb{R}^N$ aus einer Verfahrensfunktion ψ hervorgeht mittels

$$\psi^*(t,u;h) \;:=\; \psi(t,u - h^q c_q(t);h) \;+\; [c_q(t+h) - c_q(t)]h^{q-1}, \tag{7.23}$$

mit einer Zahl $q \geq 1$ und einer im Moment nicht näher spezifizierten Funktion $c_q : [a,b] \to \mathbb{R}^N$.

Lemma 7.24. *Bezüglich des Anfangswertproblems $y' = f(t,y),\ y(a) = y_0$ besteht zwischen den zu den Verfahrensfunktionen ψ und ψ^* gehörenden Gitterfunktionen v_h und v_h^* der folgende Zusammenhang,*

$$v_h^*(t) \;=\; v_h(t) \;+\; c_q(t)h^q, \qquad t \in [a,b], \qquad h \in \mathbb{H}_t.$$

[1] siehe Abschnitt 7.6 über Extrapolationsmethoden

BEWEIS. Offensichtlich gilt $v_h(0) = v_h^*(0) = y_0$, und dann erhält man induktiv für $t = h,\ 2h, \ldots$, die Aussage des Lemmas

$$\begin{aligned} v_h^*(t+h) &= v_h^*(t) + h\psi^*(t, v_h^*(t); h) \\ &= v_h(t) + h^q c_q(t) + h\psi(t, v_h(t); h) + [c_q(t+h) - c_q(t)]h^q \\ &= \underbrace{v_h(t) + h\psi(t, v_h(t); h)}_{= v_h(t+h)} + c_q(t+h)h^q. \end{aligned}$$

□

Bemerkung 7.25. Lemma 7.24 lässt sich sukzessive auf die folgenden Verfahrensfunktionen anwenden (das Schema ist zeilenweise zu lesen)

$$\left.\begin{array}{lllllll} \psi &=& \varphi, & q &=& p, & \varphi^{[1]} := \psi^* \\ \psi &=& \varphi^{[1]}, & q &=& p+1, & \varphi^{[2]} := \psi^* \\ \vdots & & \vdots & \vdots & & \vdots & \vdots \\ \psi &=& \varphi^{[r-1]}, & q &=& p+r-1, & \varphi^{[r]} := \psi^* \end{array}\right\} \tag{7.24}$$

Mit der Notation $u_{0,h} = u_h$ sowie $u_{s,h}$ für die zu $\varphi^{[s]}$ gehörende Gitterfunktion ($s = 1, 2 \ldots$) gilt nach Lemma 7.24

$$u_{s+1,h}(t) = u_{s,h}(t) + c_{p+s}(t)h^{p+s}, \qquad s = 0, 1, \ldots, r-1,$$

beziehungsweise

$$u_{r,h}(t) = u_h(t) + c_p(t)h^p + c_{p+1}(t)h^{p+1} + \ldots + c_{p+r-1}(t)h^{p+r-1}. \tag{7.25}$$

Für die komplette Herleitung der asymptotischen Entwicklung (7.22) sind nun "lediglich" noch konkrete Funktionen $c_p, \ldots, c_{p+r-1}$ zu ermitteln, so dass

$$u_{r,h}(t) - y(t) = \mathcal{O}(h^{p+r}) \qquad \text{für} \quad \mathbb{H}_t \ni h \to 0 \tag{7.26}$$

gilt beziehungsweise die zugehörige Verfahrensfunktion $\varphi^{[r]}$ aus dem Schema (7.24) die Konsistenzordnung $p + r$ besitzt. △

Die angestellten Bemerkungen legen es nahe, eine Funktion c_q zu wählen, so dass mittels der Rekursionsvorschrift (7.23) aus einer Verfahrensfunktion ψ mit der Konsistenzordnung q eine Verfahrensfunktion ψ^* erzeugt wird, die die Konsistenzordnung $q + 1$ besitzt. Die Einzelheiten dazu werden im Folgenden vorgestellt, wobei als Erstes eine Darstellung für den zu der zugrunde liegenden Verfahrensvorschrift φ gehörenden lokalen Verfahrensfehler geliefert wird:

Lemma 7.26. *Unter den in Theorem 7.23 genannten Bedingungen gilt für den zugrunde liegenden lokalen Verfahrensfehler die Entwicklung*

$$y(t+h) - y(t) - h\varphi(t, y(t); h) = d_{p+1}(t)h^{p+1} + \mathcal{O}(h^{p+2}) \qquad \textit{für} \quad h \to 0,$$

mit einer Funktion $d_{p+1} \in C^r([a, b], \mathbb{R}^N)$, wobei die angegebenen Konvergenzraten gleichmäßig in t auftreten.

BEWEIS. Die Behauptung folgt unmittelbar aus einer Taylorentwicklung der Funktion $g(h) = y(t+h) - y(t) - h\varphi(t, y(t); h)$ in $h = 0$,

$$\begin{aligned} y(t+h) - y(t) - h\varphi(t, y(t); h) &= \sum_{\ell=0}^{p+1} d_\ell(t) h^\ell + \mathcal{O}(h^{p+2}) \\ &= d_{p+1}(t) h^{p+1} + \mathcal{O}(h^{p+2}), \end{aligned}$$

da wegen der vorliegenden Konsistenzordnung q notwendigerweise $d_0(t) = \ldots = d_p(t) = 0$ gilt. Für die Funktion d_{p+1} gilt die Darstellung $d_{p+1}(t) = \frac{y^{(p+1)}(t)}{(p+1)!} - \frac{1}{p!}\frac{\partial^p \varphi}{\partial h^p}(t, y(t); 0)$. □

7.5.3 Herleitung der asymptotischen Entwicklung des globalen Verfahrensfehlers, 2. Teil

In Vorbereitung auf das nächste Lemma sei $\psi : [a,b] \times \mathbb{R}^N \times \mathbb{R}_+ \to \mathbb{R}^N$ eine beliebige Verfahrensfunktion, die bezüglich des Anfangswertproblems $y' = f(t,y),\ y(a) = y_0$ die Konsistenzordnung $q \geq 1$ besitzt mit der folgenden Darstellung für den lokalen Verfahrensfehler,

$$y(t+h) - y(t) - h\psi(t, y(t); h) = d_{q+1}(t) h^{q+1} + \mathcal{O}(h^{q+2}) \quad \text{für } h \to 0, \tag{7.27}$$

mit einer Funktion $d_{q+1} : [a,b] \to \mathbb{R}^N$, wobei die angegebenen Konvergenzraten gleichmäßig in t auftreten.

Des Weiteren wird die Konsistenzbedingung

$$\psi(t, u; 0) = f(t,u) \qquad \text{für } (t,u) \in [a,b] \times \mathbb{R}^N \tag{7.28}$$

vorausgesetzt. In allen praxisrelevanten Fällen liegt die vorausgesetzte Konsistenzordnung in der verallgemeinerten Form der Aufgabe 7.3 auf Seite 178 vor, so dass dann (7.28) automatisch erfüllt ist.

In den weiteren Überlegungen spielt das folgende Anfangswertproblem für ein inhomogenes lineares System gewöhnlicher Differenzialgleichungen eine technische Rolle,

$$c_q'(t) = \mathcal{D}_y f(t, y(t)) c_q(t) + d_{q+1}(t), \qquad t \in [a,b], \qquad c_q(a) = 0. \tag{7.29}$$

Hierbei bezeichnet $\mathcal{D}_y f(t,u) = (\frac{\partial f_i}{\partial y_j}(t,u))_{i,j=1}^N \in \mathbb{R}^{N \times N}$ die Funktionalmatrix der Abbildung $y \to f(t,y)$ an der Stelle $u \in \mathbb{R}^N$. Entsprechend wird diese Notation im Folgenden für Verfahrensfunktionen verwendet.

Mit dieser Wahl der Funktion c_q erhält man unter hinreichend guten Differenzierbarkeitseigenschaften der beteiligten Funktionen durch die Rekursionsvorschrift (7.23) eine Verfahrensfunktion ψ^* mit der Konsistenzordnung $q+1$.

Lemma 7.27. *Eine Verfahrensfunktion $\psi \in C^3([a,b] \times \mathbb{R}^N \times \mathbb{R}_+, \mathbb{R}^N)$ besitze die Konsistenzordnung $q \geq 1$ mit einem lokalen Verfahrensfehler von der Form (7.27), und die Konsistenzbedingung (7.28) sei erfüllt. Weiter sei $d_{q+1} \in C^s([a,b], \mathbb{R}^N)$ für*

ein $s \geq 1$ erfüllt, und die Abbildungen $t \mapsto \frac{\partial^2 \psi}{\partial h \, \partial y_j}(t, y(t); 0)$ und $t \mapsto \frac{\partial^2 \psi}{\partial y_i \, \partial y_j}(t, y(t); 0)$ seien für alle Indizes i, j mindestens $(s-1)$-mal stetig partiell differenzierbar auf $[a, b]$.

Unter diesen Voraussetzungen besitzt die Verfahrensfunktion ψ^* aus (7.23) mit $c_q \in C^{s+1}([a,b], \mathbb{R}^N)$ aus (7.29) die Konsistenzordnung[2] $q+1$. Im Fall $s \geq 2$ besitzt der zugehörige lokale Verfahrensfehler η^* die Darstellung

$$\eta^*(t,h) \;=\; d_{q+2}(t)\,h^{q+2} \;+\; \mathcal{O}(h^{q+3}) \qquad \textit{für} \quad h \to 0$$

gleichmäßig in t, mit einer Funktion[3] $d_{q+2} \in C^{s-1}([a,b], \mathbb{R}^N)$.

BEWEIS. Der lokale Verfahrensfehler bezüglich ψ^* besitzt die folgende Form,

$$\begin{aligned}
\eta^*(t,h) &:= y(t+h) - y(t) - h\psi^*(t, y(t); h) \\
&= \text{———} \ll \text{———} - h\psi(t, y(t) - h^q c_q(t); h) - [c_q(t+h) - c_q(t)]h^q \\
&= \underbrace{\text{———} \ll \text{———} - h\psi(t, y(t); h)}_{= \eta(t,h)} - \text{———} \ll \text{———} \\
&\quad - hR(t,h), \quad \text{mit } R(t,h) := \big[\psi(t, y(t) - h^q c_q(t); h) - \psi(t, y(t); h)\big].
\end{aligned}$$

Es soll zunächst der Fall $q \geq 2$ behandelt werden. Taylorentwicklungen liefern

$$\begin{aligned}
R(t,h) &= -\mathcal{D}_y \psi(t, y(t); h) h^q c_q(t) + \overbrace{\mathcal{O}(h^{2q})}^{= \mathcal{O}(h^{q+2})}, \qquad (7.30) \\
c_q(t+h) - c_q(t) &= h c_q'(t) + \tfrac{1}{2} c_q''(t) h^2 + \mathcal{O}(h^3) \quad \text{für} \quad h \to 0,
\end{aligned}$$

und zur Bearbeitung der Identität (7.30) verwendet man eine weitere Taylorentwicklung,

$$\mathcal{D}_y \psi(t, y(t); h) \;=\; \underbrace{\mathcal{D}_y \psi(t, y(t); 0)}_{= \mathcal{D}_y f(t, y(t))} + \frac{\partial \mathcal{D}_y \psi}{\partial h}(t, y(t); 0) h \;+\; \mathcal{O}(h^2) \quad \text{für } h \to 0,$$

mit der Matrix $\frac{\partial \mathcal{D}_y \psi}{\partial h}(t, y(t); 0) = \big(\frac{\partial^2 \psi_i}{\partial h \, \partial y_j}(t, y(t); h)\big)_{i,j=1}^N \in \mathbb{R}^{N \times N}$, wobei ψ_i die i-te Komponente der vektorwertigen Funktion ψ bezeichnet. Insgesamt erhält man

$$\begin{aligned}
\eta^*(t,h) &= \overbrace{\big[d_{q+1}(t) + \mathcal{D}_y f(t, y(t)) c_q(t) - c_q'(t)\big]}^{= 0} h^{q+1} \\
&\quad + \underbrace{\big[\tfrac{\partial \mathcal{D}_y \psi}{\partial h}(t, y(t); 0) c_q(t) - \tfrac{1}{2} c_q''(t)\big]}_{=: d_{q+2}(t)} h^{q+2} + \mathcal{O}(h^{q+3}) \quad \text{für } h \to 0,
\end{aligned}$$

[2] bezüglich des gleichen Anfangswertproblems $y' = f(t,y)$, $y(a) = y_0$

[3] Die spezielle Form von d_{q+2} ist im Beweis angegeben.

wobei die angegebenen Konvergenzraten gleichmäßig in t auftreten. Im Fall $q = 1$ verwendet man anstelle (7.30) die folgende Taylorentwicklung zweiter Ordnung,

$$R(t,h) \;=\; -\mathcal{D}_y\psi(t,y(t);h)hc_1(t) + (c_1(t)^\top \underbrace{\mathcal{D}_y^2\psi_i(t,y(t);h)}_{=\mathcal{D}_y^2\psi_i(t,y(t);0)+\mathcal{O}(h)} c_1(t))_{i=1}^N h^2 + \mathcal{O}(h^3)$$

für $h \to 0$, mit der Hessematrix $\mathcal{D}_y^2\psi_i(t,y(t);h) = \big(\frac{\partial^2\psi_i}{\partial y_k\,\partial y_l}(t,y(t);h)\big)_{k,l=1}^N$, wobei ψ_i die i-te Komponente von ψ bezeichnet. Man erhält so die Darstellung

$$\eta^*(t,h) \;=\; \overbrace{\big[\big(c_1(t)^\top\mathcal{D}_y^2\psi_i(t,y(t);0)c_1(t)\big)_{i=1}^N}^{=:\;d_3(t)} \;+\; \tfrac{\partial\mathcal{D}_y\psi}{\partial h}(t,y(t);0)c_1(t) \;-\; \tfrac{1}{2}c_1''(t)\big]h^3 + \mathcal{O}(h^4) \qquad \text{für} \quad h\to 0,$$

wobei die angegebenen Konvergenzraten gleichmäßig in t auftreten. □

Es sind nun alle Hilfsmittel zur Komplettierung des Beweises des Theorems über die asymptotische Entwicklung des globalen Verfahrensfehlers zusammengestellt:

BEWEIS VON THEOREM 7.23. Die Aussage des Theorems folgt unmittelbar aus den in Bemerkung 7.25 angestellten Vorüberlegungen, wobei noch für jede Anwendung von Lemma 7.27 dessen Voraussetzungen nachzuprüfen sind, was im Folgenden geschieht. Es ist so, dass mit der Verfahrensfunktion φ auch jede der in (7.24) betrachteten Funktionen $\varphi^{[s]}$ der Stabilitätsbedingung (7.10) genügt. Weiter gelten die Identitäten

$$f(t,u) \;=\; \varphi(t,u;0) \;=\; \varphi^{[1]}(t,u;0) \;=\; \ldots \;=\; \varphi^{[r-1]}(t,u;0)$$

sowie

$$\frac{\partial^2\varphi^{[r-1]}}{\partial h\,\partial y_j}(t,u;0) \;=\; \ldots \;=\; \frac{\partial^2\varphi^{[1]}}{\partial h\,\partial y_j}(t,u;0)$$
$$= \begin{cases} \dfrac{\partial^2\varphi}{\partial h\,\partial y_j}(t,u;0) & , \text{ falls } p\ge 2, \\ \text{——} \,\text{«}\, \text{——} - \displaystyle\sum_{i=1}^N \frac{\partial^2\varphi}{\partial y_i\,\partial y_j}(t,u;0), & -\,\text{«}\,-\, p=1, \end{cases}$$

so dass Lemma 7.27 tatsächlich jeweils anwendbar ist. Theorem 7.23 ist damit vollständig bewiesen. □

7.5.4 Asymptotische Entwicklungen des lokalen Verfahrensfehlers

Es werden nun die vorgestellten asymptotische Entwicklungen des globalen Verfahrensfehlers zur Gewinnung von Verfahren höherer Ordnung eingesetzt. Zuvor wird noch eine asymptotische Entwicklung für den *lokalen* Verfahrensfehler angegeben, die sich bei der Konstruktion von Schrittweitensteuerungen verwenden lässt:

Theorem 7.28. *Unter den Bedingungen von Theorem 7.23 gilt für jede fixierte Zahl $\ell \in \mathbb{N}$ die folgende Entwicklung für den lokalen Verfahrensfehler[4]:*

$$\underbrace{u_h(a+\ell h)}_{= u_\ell} = y(a+\ell h) + b_{p+1}h^{p+1} + b_{p+2}h^{p+2} + \ldots + b_{p+r-1}h^{p+r-1} + \mathcal{O}(h^{p+r}) \quad (7.31)$$

für $h > 0$, mit gewissen von der Zahl ℓ abhängenden vektoriellen Koeffizienten $b_{p+1}, \ldots, b_{p+r-1} \in \mathbb{R}^N$.

BEWEIS. Aus Theorem 7.23 erhält man unter Verwendung der Taylorentwicklungen

$$c_{p+j}(a+\ell h) = \sum_{k=1}^{r-j-1} c_{p+j}^{(k)}(a)\frac{(\ell h)^k}{k!} + \mathcal{O}(h^{r-j}) \qquad \text{für} \quad j = 0, 1, \ldots, r-1$$

unmittelbar die Aussage des Theorems,

$$\begin{aligned} u_h(a+\ell h) &= y(a+\ell h) + \sum_{j=0}^{r-1} c_{p+j}(a+\ell h)h^{p+j} + \mathcal{O}(h^{p+r}) \\ &= y(a+\ell h) + \sum_{s=1}^{r-1} \Big(\underbrace{\sum_{k=1}^{s} c_{p+s-k}^{(k)}(a)\frac{\ell^k}{k!}}_{=:\, b_{p+s}} \Big) h^{p+s} + \mathcal{O}(h^{p+r}). \end{aligned}$$

□

7.6 Extrapolationsmethoden für Einschrittverfahren

Im Folgenden wird ein Einschrittverfahren (7.19)[5] mit der Konsistenzordnung $p \geq 1$ und einer asymptotischen Entwicklung von der Form[6] $u_h(t) = y(t) + c_p(t)h^p + c_{p+1}(t)h^{p+1} + \ldots + c_{p+r-1}(t)h^{p+r-1} + \mathcal{O}(h^{p+r})$ herangezogen. Bei fixiertem $t \in [a, b]$ werden Extrapolationsverfahren für $h \to 0$ betrachtet mit dem Ziel der Gewinnung von Verfahren höherer Ordnung. Zur Approximation von $y(t)$ betrachte man für eine feste Stelle $t \in [a, b]$ zu Schrittweiten $h_{[0]} > h_{[1]} > \ldots$ aus $\mathbb{H}_t$ (siehe (7.20)) und einer Zahl $0 \leq m \leq r$ das vektorwertige Polynom $\mathcal{P}_{0,\ldots,m}$ von der Form

$$\mathcal{P}_{0,\ldots,m}(h) = d_0 + d_p h^p + d_{p+1}h^{p+1} + \ldots + d_{p+m-1}h^{p+m-1}, \quad h \in \mathbb{R}, \quad (7.32)$$

mit vektoriellen Koeffizienten $d_0, d_p, d_{p+1}, \ldots, d_{p+m-1} \in \mathbb{R}^N$, wobei diese $m+1$ Koeffizienten so zu bestimmen sind, dass die $m+1$ Interpolationsbedingungen

$$\mathcal{P}_{0,\ldots,m}(h_{[k]}) = u_{h_{[k]}}(t) \qquad \text{für} \quad k = 0, 1, \ldots, m, \quad (7.33)$$

[4] Anders als bei der asymptotischen Entwicklung des globalen Verfahrensfehlers hängt die betrachte Stelle hier von h ab.

[5] zur approximativen Bestimmung der Lösung des Anfangswertproblems $y' = f(t, y),\ y(a) = y_0$

[6] siehe (7.22)

erfüllt sind. Die betrachteten Schrittweiten seien dabei so gewählt, dass bezüglich einer Grundschrittweite $h \in \mathbb{H}_t$ Folgendes gilt,

$$h_{[k]} \;=\; h/n_k \qquad \text{für} \quad k \;=\; 0,1,\ldots, \quad \text{mit} \quad 1 \;\le\; n_0 \;\le\; n_1 \;\le\; \ldots\,. \tag{7.34}$$

Als Näherung für $y(t)$ wird $\mathcal{P}_{0,\ldots,m}(0)$ herangezogen. Durch diese Extrapolation nach $h \to 0$ erhält man ein Verfahren der Ordnung $p+m$, es gilt $\mathcal{P}_{0,\ldots,m}(0) \;=\; y(t) \;+\; \mathcal{O}(h^{p+m})$. Die genauen Approximationseigenschaften sind in dem folgenden Theorem angegeben.

Theorem 7.29. *Gegeben sei ein Einschrittverfahren (7.19)[5] mit einer asymptotischen Entwicklung von der Form (7.22). In der Situation (7.34) gilt für das (existierende und eindeutig bestimmte) Polynom $\mathcal{P}_{0,\ldots,m}$ von der Form (7.32) mit der Interpolationseigenschaft (7.33) die folgende Fehlerdarstellung*

$$\mathcal{P}_{0,\ldots,m}(0) \;=\; y(t) \;+\; \sum_{s=p+m}^{p+r-1} B_s\, c_s(t)\, h^s \;+\; \mathcal{O}(h^{p+r}), \tag{7.35}$$

mit von t und h unabhängigen Matrizen $B_{p+m},\ldots,B_{p+r-1} \in \mathbb{R}^{N\times N}$.

BEWEIS. Der Beweis wird zunächst für den eindimensionalen Fall ($N = 1$) geführt. Die Menge der Polynome von der Form (7.32) stimmt (für $N = 1$) überein mit $\{P \in \Pi_{p+m-1} \;:\; P^{(\nu)}(0) = 0 \text{ für } \nu = 1,2,\ldots,p-1\}$, und die angegebene Existenz und Eindeutigkeit folgt dann aus der des hermiteschen Interpolationsproblems, vergleiche Aufgabe 1.3 auf Seite 18. Im Folgenden wird die angegebene Fehlerdarstellung für $\mathcal{P}_{0,\ldots,m}(0) - y(t) = d_0 - y(t)$ hergeleitet. Hierzu schreibt man die Interpolationsbedingungen (7.33) in Form eines linearen Gleichungssystems

$$\underbrace{\begin{pmatrix} 1 & 1/n_0^p & 1/n_0^{p+1} & \cdots & 1/n_0^{p+m-1} \\ 1 & 1/n_1^p & 1/n_1^{p+1} & \cdots & 1/n_1^{p+m-1} \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & 1/n_m^p & 1/n_m^{p+1} & \cdots & 1/n_m^{p+m-1} \end{pmatrix}}_{=:\; A_m \in \mathbb{R}^{(m+1)\times(m+1)}} \begin{pmatrix} d_0 \\ h^p d_p \\ h^{p+1} d_{p+1} \\ \vdots \\ h^{p+m-1} d_{p+m-1} \end{pmatrix} \;=\; \begin{pmatrix} u_{h_{[0]}}(t) \\ u_{h_{[1]}}(t) \\ \vdots \\ u_{h_{[m]}}(t) \end{pmatrix}, \tag{7.36}$$

wobei die auftretende Matrix wegen der Eindeutigkeit des Polynoms $\mathcal{P}_{0,\ldots,m}$ regulär ist. Auf der anderen Seite führt eine Auswertung der asymptotischen Entwicklung

(7.22) an den Stellen $h_{[0]}, h_{[1]}, \ldots, h_{[m]}$ in Matrix-Vektor-Darstellung auf Folgendes,

$$\overbrace{\begin{pmatrix} 1 & 1/n_0^p & 1/n_0^{p+1} \cdots 1/n_0^{p+m-1} \\ 1 & 1/n_1^p & 1/n_1^{p+1} \cdots 1/n_1^{p+m-1} \\ \vdots & \vdots & \vdots \quad\quad \vdots \\ 1 & 1/n_m^p & 1/n_m^{p+1} \cdots 1/n_m^{p+m-1} \end{pmatrix}}^{= A_m} \begin{pmatrix} y(t) \\ c_p(t)h^p \\ c_{p+1}(t)h^{p+1} \\ \vdots \\ c_{p+m-1}(t)h^{p+m+1} \end{pmatrix} = \begin{pmatrix} u_{h_{[0]}}(t) \\ u_{h_{[1]}}(t) \\ \vdots \\ u_{h_{[m]}}(t) \end{pmatrix} - r_h(t), \quad (7.37)$$

$$\text{mit} \quad r_h(t) := \sum_{s=p+m}^{p+r-1} \begin{pmatrix} 1/n_0^s \\ 1/n_1^s \\ \vdots \\ 1/n_m^s \end{pmatrix} c_s(t)h^s + \mathcal{O}(h^{p+r}),$$

mit der gleichen Matrix wie in (7.36). Subtrahiert man nun das System (7.37) von dem Gleichungssystem (7.36), so führt dies auf

$$\underbrace{\begin{pmatrix} 1 & 1/n_0^p & 1/n_0^{p+1} \cdots 1/n_0^{p+m-1} \\ 1 & 1/n_1^p & 1/n_1^{p+1} \cdots 1/n_1^{p+m-1} \\ \vdots & \vdots & \vdots \quad\quad \vdots \\ 1 & 1/n_m^p & 1/n_m^{p+1} \cdots 1/n_m^{p+m-1} \end{pmatrix}}_{= A_m} \begin{pmatrix} d_0 - y(t) \\ (d_p - c_p(t))h^p \\ (d_{p+1} - c_{p+1}(t))h^{p+1} \\ \vdots \\ (d_{p+m-1} - c_{p+m-1}(t))h^{p+m+1} \end{pmatrix} = r_h(t). \quad (7.38)$$

Multipliziert man noch A_m^{-1} auf beiden Seiten, so führt eine Betrachtung der ersten Gleichung des entstehenden Systems für den eindimensionalen Fall $N = 1$ auf die Behauptung (unter Beachtung der Unabhängigkeit der Matrix A_m von h und t). Im allgemeinen Fall $N \geq 1$ sind in der Matrix A_m und in den in r_h auftretenden Vektoren die skalaren Einträge $1/n_j^q \in \mathbb{R}$ jeweils durch die Matrizen $(1/n_j^q)I \in \mathbb{R}^{N\times N}$ zu ersetzen, ansonsten bleibt die Argumentation die Gleiche. □

Bemerkung 7.30. (a) Im Fall einer Konsistenzordnung $p = 1$ und der eindimensionalen Situation $N = 1$ ist die Aussage von Theorem 7.29 eine unmittelbare Konsequenz aus Theorem 6.26 in Kapitel 6 über numerische Integration.

(b) Die in dem genannten Kapitel 6 angegebenen speziellen Unterteilungsfolgen lassen sich auch als Schrittweiten $h_{[0]} > h_{[1]} > \ldots$ verwenden. △

Beispiel 7.31. Mit den genannten Bezeichnungen wird nun der Spezialfall der Konsistenzordnung $p = 1$ und die Schrittweiten $h_{[0]} = h$, $h_{[1]} = h/2$, $h_{[2]} = h/4$

betrachtet mit der typischerweise kleinen Grundschrittweite $h > 0$. Man erhält dann die Fehlerdarstellung

$$\mathcal{P}_{012}(0) \;=\; y(t) \;+\; \mathcal{O}(h^3),$$

mit einem kubisch in h fallenden Fehler. Der erforderliche Aufwand zur Berechnung von $\mathcal{P}_{012}(0)$ entsprechend dem Neville-Schema (1.7) auf Seite 6 dagegen beträgt $n + 2n + 4n = 7n = \mathcal{O}(1/h)$ Schritte des vorliegenden Einschrittverfahrens, so dass der dafür erforderliche Aufwand lediglich linear in $n = \mathcal{O}(1/h)$ wächst. △

Beispiel 7.32. In der speziellen Situation $u_h(t) = y(t) + c_p(t)h^p + c_{p+1}(t)h^{p+1} + \mathcal{O}(h^{p+2})$ für $h \to 0$ und $h_{[0]} = h, h_{[1]} = h/n_1$ berechnet sich der Wert $\mathcal{P}_{01}(0)$ zu

$$\mathcal{P}_{01}(0) \;=\; u_{h/n_1}(t) \;+\; \frac{u_{h/n_1}(t) - u_h(t)}{n_1^p - 1},$$

was man wahlweise mit dem Neville-Schema (1.7) oder über das Gleichungssystem (7.36) im Beweis von Theorem 7.29 erhält. Das Gleichungssystem (7.38) aus dem angesprochenen Beweis liefert die Fehlerdarstellung

$$\mathcal{P}_{01}(0) \;=\; y(t) \;-\; \beta c_{p+1}(t)\, h^{p+1} \;+\; \mathcal{O}(h^{p+2})$$

mit dem Koeffizienten $\beta := (1-1/n_1)/(n_1^p - 1)$, Details werden hier nicht ausgeführt (Aufgabe 7.10 auf Seite 180).

Für die nachfolgenden Betrachtungen über Schrittweitensteuerungen wird hier noch der Spezialfall $t = a + \ell h$ mit fixiertem $\ell \in \mathbb{N}$ genauer untersucht. Eine Taylorentwicklung der Funktion c_{p+1} im Punkt $t = a$ liefert wegen der Identität $c_{p+1}(a) = 0$ die Abschätzung $c_{p+1}(a + \ell h) = \mathcal{O}(h)$ und somit

$$\mathcal{P}_{01}(0) \;=\; y(a + \ell h) \;+\; \mathcal{O}(h^{p+2}) \qquad \text{für} \quad h \to 0. \qquad \triangle$$

7.7 Schrittweitensteuerung

7.7.1 Verfahrensvorschrift

Zur Lösung des Anfangswertproblems $y' = f(t, y),\ y(a) = y_0$ wird für eine gegebene Verfahrensfunktion $\varphi : [a, b] \times \mathbb{R}^N \times \mathbb{R}_+ \to \mathbb{R}^N$ mit der Konsistenzordnung $p \geq 1$ die folgende Vorschrift herangezogen,

$$\left.\begin{aligned} w \;&=\; u_\ell \;+\; \tfrac{h_\ell}{2}\varphi(t_\ell, u_\ell; \tfrac{h_\ell}{2}), \\ u_{\ell+1} \;&=\; w \;+\; \tfrac{h_\ell}{2}\varphi(t_\ell + \tfrac{h_\ell}{2}, w; \tfrac{h_\ell}{2}), \quad t_{\ell+1} := t_\ell + h_\ell, \quad \ell = 0, 1, \ldots . \end{aligned}\right\} \tag{7.39}$$

Im Folgenden wird eine *adaptive* Wahl der Schrittweiten h_ℓ vorgestellt mit dem Ziel einer effizienten Fehlerkontrolle. Einführende Erläuterungen hierzu findet man im folgenden Abschnitt 7.7.2, und in den nachfolgenden Abschnitten 7.7.3 und 7.7.4 wird die genaue Vorgehensweise zur Wahl der Schrittweiten h_ℓ beschrieben.

Bemerkung 7.33. Der Schritt $(t_\ell, u_\ell) \to (t_{\ell+1}, u_{\ell+1})$ in der Verfahrensvorschrift (7.39) entspricht zwei Schritten $(t_\ell, u_\ell) \to (t_{\ell+1/2}, u_{\ell+1/2}) \to (t_{\ell+1}, u_{\ell+1})$ des Einschrittverfahrens (7.8) mit halber Schrittweite $h_\ell/2$. Diese Approximation $u_{\ell+1} \approx y(t_{\ell+1}) \in \mathbb{R}^N$ wird für eine Fehlerschätzung benötigt, daher kann man auch gleich die Verfahrensvorschrift (7.39) anstelle des ursprünglichen Einschrittverfahrens (7.8) verwenden. △

7.7.2 Problemstellung

Im Folgenden soll ausgehend von einer gegebenen Stelle $t_\ell \in [a, b]$ und einer gegebenen Approximation $u_\ell \approx y(t_\ell) \in \mathbb{R}^N$ eine Schrittweite $h_\ell > 0$ bestimmt werden, für die

$$\|u_{\ell+1} - z(t_\ell + h_\ell)\| \approx \varepsilon \tag{7.40}$$

erfüllt ist, wobei $u_{\ell+1} \in \mathbb{R}^N$ aus einem Schritt des gegenwärtig betrachteten Verfahrens (7.39) hervorgeht und $\varepsilon > 0$ eine vorgegebene Fehlerschranke darstellt, und $z : [t_\ell, b] \to \mathbb{R}^N$ bezeichnet die Lösung des Anfangswertproblems

$$z' = f(t, z), \qquad t \in [t_\ell, b]; \qquad\qquad z(t_\ell) = u_\ell. \tag{7.41}$$

Weiter bezeichnet $\|\cdot\|$ in (7.40) eine nicht näher spezifizierte Vektornorm.

Bemerkung 7.34. (a) Die Forderung (7.40) zeigt, dass die noch zu beschreibende Schrittweitensteuerung auf einer Vorgabe des *lokalen* Verfahrensfehlers beruht. Damit erhofft man sich ein vernünftiges Verhalten des *globalen* Verfahrensfehlers.

(b) Die Forderung (7.40) stellt man aus den folgenden Gründen:

- der lokale Verfahrensfehler $\|u_{\ell+1} - z(t_\ell + h_\ell)\|$ soll die vorgegebene Schranke ε nicht übersteigen. Dies wird durch die Wahl einer hinreichend kleinen Schrittweite h_ℓ erreicht.
- Aus Effizienzgründen und zur Vermeidung der Akkumulation von Rundungsfehlern wird man die Schrittweite h_ℓ jedoch nicht so klein wählen wollen, dass $\|u_{\ell+1} - z(t_\ell + h_\ell)\| \ll \varepsilon$ gilt.

(c) Zu beachten ist zudem, dass die Lösung des Anfangswertproblems (7.41) nicht bekannt ist und erst noch numerisch zu bestimmen ist. △

Zur Vereinfachung der Notation führen wir die folgende Bezeichnung für einen von dem Punkt (t_ℓ, u_ℓ) ausgehenden Schritt der Verfahrensvorschrift (7.39) mit Länge h ein,

$$u_{2\times h/2} := w + \tfrac{h}{2}\varphi(t_\ell + \tfrac{h}{2}, w; \tfrac{h}{2}) \quad \text{mit} \quad w = u_\ell + \tfrac{h}{2}\varphi(t_\ell, u_\ell; \tfrac{h}{2}). \tag{7.42}$$

Zur Bestimmung einer Schrittweite h_ℓ, für die die Forderung (7.40) ungefähr erfüllt ist, wird ausgehend von einer nicht zu kleinen Startschrittweite $h^{(0)}$ für $k = 0, 1 \ldots$, so vorgegangen:

- Zunächst berechnet man $u_{2\times h^{(k)}/2}$.

- Anschließend ermittelt man eine Schätzung für den Fehler $\|u_{2\times h^{(k)}/2} - z(t_\ell + h^{(k)})\|$ und bricht den Iterationsprozess mit $k_\varepsilon := k$ ab, falls diese Schätzung kleiner gleich ε ausfällt.
- Andernfalls, falls diese Schätzung größer als ε ist, wird eine neue Testschrittweite $h^{(k+1)} < h^{(k)}$ bestimmt.

Abschließend verfährt man mit $h_\ell = h^{(k_\varepsilon)}$ und $t_{\ell+1} = t_\ell + h^{(k_\varepsilon)}$ fort. Einzelheiten zu der genannten Fehlerschätzung und der Bestimmung einer neuen Testschrittweite werden in den nachfolgenden Abschnitten 7.7.3–7.7.4 beschrieben.

7.7.3 Vorgehensweise bei gegebener Testschrittweite $h^{(k)}$

Für eine Testschrittweite $h^{(k)} > 0$, $k \in \mathbb{N}_0$, bestimmt man entsprechend einem Schritt der Verfahrensvorschrift (7.42) den Vektor $u_{2\times h^{(k)}/2} \in \mathbb{R}^N$. Anschließend wird zur Überprüfung der Eigenschaft $\|u_{2\times h^{(k)}/2} - z(t_\ell + h^{(k)})\| \approx \varepsilon$ der Wert $z(t_\ell + h^{(k)})$ durch $z_{h^{(k)}} \in \mathbb{R}^N$ geschätzt, wobei

$$z_h := u_{2\times h/2} - \frac{v_h - u_{2\times h/2}}{2^p - 1} \quad \text{mit} \quad v_h := u_\ell + h\varphi(t_\ell, u_\ell; h), \qquad h > 0. \tag{7.43}$$

Dabei erhält man die Approximation (7.43) mittels lokaler Extrapolation entsprechend Beispiel 7.32 mit $n_1 = 2$. Der Fehler $\|u_{2\times h^{(k)}/2} - z(t_\ell + h^{(k)})\|$ berechnet sich dann näherungsweise zu

$$\delta^{(k)} := \|u_{2\times h^{(k)}/2} - z_{h^{(k)}}\| = \frac{\|v_{h^{(k)}} - u_{2\times h^{(k)}/2}\|}{2^p - 1}. \tag{7.44}$$

Ist dann die Abschätzung $\delta^{(k)} \le \varepsilon$ erfüllt, so gibt man sich (vergleiche (7.40) mit $t_{\ell+1} = t_\ell + h^{(k)}$) mit der Schrittweite $h_\ell = h^{(k)}$ zufrieden und verfährt wie in Abschnitt 7.7.2 beschrieben fort (mit ℓ um eins erhöht). Die vorliegende Situation ist in Bild 7.4 veranschaulicht.

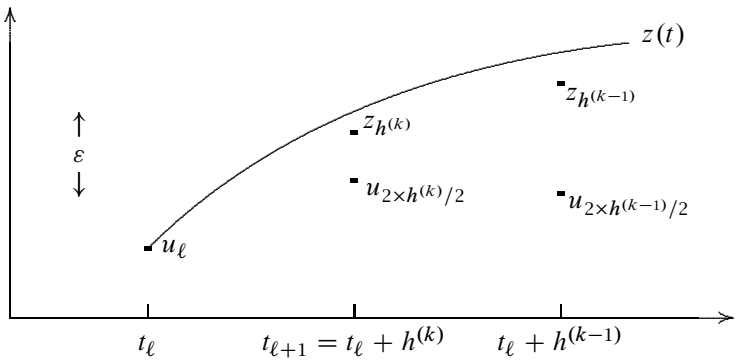


Bild 7.4: Illustration zur Schrittweitensteuerung

7.7.4 Bestimmung einer neuen Testschrittweite $h^{(k+1)}$ im Fall $\delta^{(k)} > \varepsilon$

Gilt mit der Notation aus (7.44) jedoch $\delta^{(k)} > \varepsilon$, so wiederholt man die in Abschnitt 7.7.3 vorgestellte Vorgehensweise mit k um eins erhöht, mit einer neuen Testschrittweite $h^{(k+1)} < h^{(k)}$. Bei der Festlegung einer solchen neuen Testschrittweite $h^{(k+1)}$ bedient man sich einer näherungsweisen Darstellung des Fehlers $u_{2\times h/2} - z(t_\ell + h)$:

Lemma 7.35. *Mit den Notationen (7.41)–(7.44) gilt unter den Bedingungen von Theorem 7.23 über die Asymptotik des globalen Verfahrensfehlers (dort für $r = 2$) Folgendes,*

$$\|u_{2\times h/2} - z(t_\ell + h)\| \;=\; \Big(\frac{h}{h^{(k)}}\Big)^{p+1}\delta^{(k)} \;+\; \mathcal{O}\big((h^{(k)})^{p+2}\big), \quad 0 < h \le h^{(k)}. \tag{7.45}$$

Gilt also $h^{(k)} \ll \varepsilon^{1/(p+2)}$, so gewinnt man aus der Darstellung (7.45) unter Vernachlässigung des Restglieds die neue Testschrittweite

$$h^{(k+1)} \;:=\; \Big(\frac{\varepsilon}{\delta^{(k)}}\Big)^{1/(p+1)} h^{(k)} \tag{7.46}$$

und wiederholt damit die Vorgehensweise in Abschnitt 7.7.3, mit k um eins erhöht.

Beweis von Lemma 7.35. Gemäß Theorem 7.28 existiert ein von h unabhängiger Vektor $b_{p+1} \in \mathbb{R}^N$ mit

$$u_{2\times h/2} - z(t_\ell + h) \;=\; b_{p+1}h^{p+1} \;+\; \mathcal{O}(h^{p+2}), \qquad h > 0, \tag{7.47}$$

und im Folgenden wird eine Approximation für b_{p+1} geliefert. Mithilfe von Beispiel 7.32 erhält man mit z_h aus (7.43) Folgendes,

$$z_h \;-\; z(t_\ell + h) \;=\; \mathcal{O}(h^{p+2}),$$

und dies eingesetzt in (7.47) führt auf

$$u_{2\times h/2} \;-\; z_h \;=\; b_{p+1}h^{p+1} \;+\; \mathcal{O}(h^{p+2}). \tag{7.48}$$

Wegen der Identität $\delta^{(k)} = \|u_{2\times h^{(k)}/2} - z_{h^{(k)}}\|$ bedeutet die Darstellung (7.48) insbesondere $\|b_{p+1}\|(h^{(k)})^{p+1} = \delta^{(k)} + \mathcal{O}((h^{(k)})^{p+2})$ beziehungsweise

$$\|b_{p+1}\| \;=\; \frac{\delta^{(k)}}{(h^{(k)})^{p+1}} \;+\; \mathcal{O}(h^{(k)}). \tag{7.49}$$

Die Darstellung (7.49) eingesetzt in (7.47) liefert die Aussage des Lemmas,

$$\begin{aligned}\|u_{2\times h/2} - z(t_\ell + h)\| &= \Big(\frac{h}{h^{(k)}}\Big)^{p+1}\delta^{(k)} \;+ \mathcal{O}(h^{(k)})h^{p+1} \;+\; \mathcal{O}(h^{p+2}) \\ &= \text{\_\_\_\_} \,\text{«}\, \text{\_\_\_\_} \;+ \mathcal{O}((h^{(k)})^{p+2}), \qquad 0 < h \le h^{(k)}. \end{aligned}$$

□

Bemerkung 7.36. (1) Für den Startschritt empfiehlt sich eine Wahl $h^{(0)} = \varepsilon^q$ mit einer Konstanten $1 < q < 1/(p+2)$.

(2) Zur der in diesem Abschnitt 7.7 vorgestellten Schrittweitenstrategie existieren Alternativen. Ebenfalls sinnvoll ist zum Beispiel ein Abbruchkriterium der Form $c_1\varepsilon \le \delta^{(k_\varepsilon)} \le c_2\varepsilon$. Ist diese Bedingung etwa für ein k noch nicht erfüllt, so setzt man $h^{(k+1)}$ entsprechend (7.46), wobei hier eine Schrittweitenvergrößerung $h^{(k+1)} > h^{(k)}$ eintreten kann.

(3) Nicht behandelt wird hier die Frage, ob das in diesem Abschnitt 7.7 beschriebene Abbruchkriterium nach einer endlichen Wahl von Versuchsschrittweiten abbricht oder nicht (beziehungsweise ob $k_\varepsilon < \infty$ gilt). △

7.7.5 Pseudocode zur Schrittweitensteuerung

Die in Abschnitt 7.7 beschriebene Vorgehensweise wird abschließend in Form eines Pseudocodes zusammengefasst, wobei wieder $\varphi : [a,b] \times \mathbb{R}^N \times \mathbb{R}_+ \to \mathbb{R}^N$ eine Verfahrensfunktion der Konsistenzordnung $p \ge 1$ zur Lösung des Anfangswertproblems (7.1)–(7.2) ist.

Algorithmus 7.37. Seien $t_0 = a, \quad u_0 = y_0, \quad \ell = 0, \quad h^{(0)} > 0, \quad \varepsilon > 0.$

```
repeat   k = 0;
  repeat
    if k = 0  then h = h^(0)  else h = (ε/δ)^(1/(p+1)) h  end;
    w = u_ℓ + (h/2) φ(t_ℓ, u_ℓ; h/2);      u_{ℓ+1} = w + (h/2) φ(t_ℓ + h/2, w; h/2);
    v = u_ℓ + h φ(t_ℓ, u_ℓ; h);      δ = ‖v − u_{ℓ+1}‖ / (2^p − 1);      k = k + 1;
  until δ ≤ ε;
  t_{ℓ+1} = t_ℓ + h;      ℓ = ℓ + 1;
until t_ℓ ≥ b;
```

△

Weitere Themen und Literaturhinweise

Die Theorie der Anfangswertprobleme für gewöhnliche Differenzialgleichungssysteme wird beispielsweise in Heuser [54] und in Dallmann/Elster [15] einführend behandelt, und eine Auswahl existierender Literatur über Einschrittverfahren zur numerischen Lösung solcher Probleme bildet Deuflhard/Bornemann [21], Grigorieff [41], Hairer/Nørsett/Wanner [50], Kress [63], Reinhardt [85], Strehmel/Weiner [101], Stoer/Bulirsch [99] und Weller [110]. Insbesondere in [21], [50] und [101] findet man

auch weitergehende Ausführungen über die hier nur beiläufig behandelten Runge-Kutta-Verfahren. In März [68] und in [101] findet man Einführungen über die hier nicht behandelten *Algebro-Differenzialgleichungssysteme*, bei denen es sich um spezielle implizite Differenzialgleichungssysteme von der Form $f(t, y(t), y'(t)) = 0$ handelt.

Übungsaufgaben

Aufgabe 7.1. Man forme das Anfangswertproblem

$$\begin{aligned} y_1'' &= t^2 - y_1' - y_2^2, \\ y_2'' &= t + y_2' + y_1^3, \\ y_1(0) &= 0, \quad y_2(0) = 1, \quad y_1'(0) = 1 \quad y_2'(0) = 0 \end{aligned}$$

in ein Anfangswertproblem für ein System erster Ordnung um.

Aufgabe 7.2. (a) Für das Anfangswertproblem

$$y' = (1 + |y|)^{-1} \quad \text{auf } [0, b], \qquad y(0) = y_0, \tag{7.50}$$

weise man Existenz und Eindeutigkeit der Lösung nach.
(b) Seien y und v Lösungen der Differenzialgleichung in (7.50) mit den Anfangswerten $y(0) = y_0$ beziehungsweise $v(0) = v_0$. Man weise Folgendes nach,

$$|y(t) - v(t)| \le e^{|t|} |y_0 - v_0| \qquad \text{für } t \in [0, b].$$

Aufgabe 7.3. Für ein Einschrittverfahren (7.8) zur Lösung des Anfangswertproblems $y' = f(t, y),\ y(a) = y_0$ lässt sich der lokale Verfahrensfehler allgemeiner auch für beliebige Punkte $(t, y) \in [a, b] \times \mathbb{R}^N$ definieren,

$$\eta(t, h) := y + h\varphi(t, y; h) - z(t + h), \qquad 0 \le h \le b - t,$$

wobei $z : [t, b] \to \mathbb{R}^N$ die Lösung des Anfangswertproblems $z' = f(s, z),\ s \in [t, b]$ mit Anfangswert $z(t) = y$ bezeichnet. Entsprechend lässt sich der Begriff Konsistenzordnung $p \ge 1$ aus Definition 7.9 für beliebige Punkte $(t, y) \in [a, b] \times \mathbb{R}^N$ verallgemeinern. Man zeige: Für jedes Einschrittverfahren (7.8) zur Lösung des Anfangswertproblems $y' = f(t, y),\ y(a) = y_0$ mit einer verallgemeinerten Konsistenzordnung $p \ge 1$ gilt die *Konsistenzbedingung*

$$\varphi(t, y; 0) = f(t, y) \qquad \text{für } (t, y) \in [a, b] \times \mathbb{R}^N.$$

Aufgabe 7.4. Man betrachte das Anfangswertproblem

$$y' = g(t), \qquad t \in [a, b], \tag{7.51}$$

$$y(a) = 0, \tag{7.52}$$

mit einer gegebenen hinreichend glatten Funktion $g : [a, b] \to \mathbb{R}$. Wendet man das Euler-Verfahren mit konstanter Schrittweite $h = (b - a)/N$ auf das Anfangswertproblem (7.51)–(7.52) an, so erhält man eine Näherungsformel für das Integral $\int_a^b g(t)\, dt$. Gleiches gilt für das Verfahren von Heun. Man gebe beide Nährungsformeln für das Integral sowie jeweils obere Schranken für den von der Zahl h abhängenden Integrationsfehler an.

Aufgabe 7.5. Gegeben sei das Anfangswertproblem

$$y' = t - t^3, \qquad y(0) = 0.$$

Zur Schrittweite h sollen mit dem Euler-Verfahren Näherungswerte u_ℓ für $y(t_\ell)$, $t_\ell = \ell h$, berechnet werden. Man gebe $y(t_\ell)$ und u_ℓ explizit an und zeige, dass an jeder Stelle t der Fehler $e_h(t) = u_h(t) - y(t)$ für $h = t/n \to 0$ gegen Null konvergiert.

Aufgabe 7.6 (*Numerische Aufgabe*). Man löse die van der Pol'sche Differenzialgleichung

$$y'' - \lambda(1 - y^2)y' + y = 0, \qquad y(0) = 2, \quad y'(0) = 0$$

für $\lambda = 0$ und $\lambda = 12$ numerisch jeweils mit dem Euler-Verfahren, dem modifizierten Euler-Verfahren sowie dem klassischen Runge-Kutta-Verfahren. Dabei verwende man jeweils einmal die konstante Schrittweite $h = 0.025$ und einmal die konstante Schrittweite $h = 0.0025$ und gebe tabellarisch die Näherungswerte an den Gitterpunkten $t = 0.5, 1.0, 1.5, \ldots, 15$, an.

Aufgabe 7.7 (Taylor-Verfahren). Für eine p-fach differenzierbare Funktion $f : [a, b] \times \mathbb{R} \to \mathbb{R}$ mit $p \in \mathbb{N}$ sei $f^{(0)} := f$ und

$$f^{[j]} := \frac{\partial f^{[j-1]}}{\partial t} + \frac{\partial f^{[j-1]}}{\partial y} f \qquad \text{für} \quad j = 1, 2, \ldots, p.$$

Zur Lösung des Anfangswertproblems $y' = f(t, y)$, $y(a) = y_0$ ist dann über die Verfahrensfunktion

$$\varphi(t, y; h) := \sum_{j=1}^{p} \frac{h^{j-1}}{j!} f^{(j-1)}(t, y) \tag{7.53}$$

ein Einschrittverfahren $u_{\ell+1} = u_\ell + h\varphi(t_\ell, u_\ell; h)$ der Ordnung p definiert. Nun zur Aufgabenstellung. Gegeben sei das Anfangswertproblem

$$y' = 1 - y \quad \text{auf } [0, 1], \qquad y(0) = 0. \tag{7.54}$$

(a) Man bestimme für jede Zahl $p \in \mathbb{N}$ die zugehörige Verfahrensfunktion φ.
(b) Man löse das Anfangswertproblem (7.54) für $p = 2$ und $h = 1/n$ näherungsweise mit dem zur Verfahrensfunktion (7.53) gehörenden Einschrittverfahren und schätze den Fehler bei $b = 1$ ab.

Aufgabe 7.8. Man zeige, dass das durch die Verfahrensfunktion

$$\varphi(t, y; h) = \frac{1}{6}(k_1 + 4k_2 + k_3),$$

$$k_1 = f(t, y), \quad k_2 = f(t + \tfrac{h}{2}, y + \tfrac{h}{2}k_1), \quad k_3 = f(t + h, y + h(-k_1 + 2k_2)),$$

gegebene Einschrittverfahren (*einfache Kutta-Regel*) die Konvergenzordnung $p = 3$ besitzt.

Aufgabe 7.9. Zur Lösung des Anfangswertproblems $y' = f(t, y)$, $y(a) = y_0$ sei für jedes $p > 0$ ein Einschrittverfahren p-ter Ordnung gegeben, welches für jeden Schritt die Rechenzeit pT_0 benötigt und in $t = b$ den Wert der gesuchten Funktion approximiert mit einem Fehler Kh^p. Die Konstanten K und T_0 sollen vom jeweiligen Verfahren unabhängig sein. Man bestimme für p und einen vorgeschriebenen Fehler $\varepsilon \le K$ in $t = b$ die größtmögliche Schrittweite $h = h(p, \varepsilon)$ und die zugehörige Gesamtrechenzeit $T = T(p, \varepsilon)$. Wie verhält sich T in Abhängigkeit von p und welches ist die optimale Konsistenzordnung $p_{\text{opt}} = p_{\text{opt}}(\varepsilon)$? Wie verhält sich p_{opt} in Abhängigkeit von ε? Der Einfachheit halber sei angenommen, dass die Zahlen p und N (wobei der Zusammenhang $h = (b - a)/N$ besteht) reell gewählt werden dürfen.

Aufgabe 7.10. Man weise die in Beispiel 7.32 getroffenen Aussagen nach.

Aufgabe 7.11 (*Numerische Aufgabe*). Man löse numerisch die Differenzialgleichung

$$y' \;=\; -200\,t\,y^2, \quad t \geq -3, \qquad\qquad y(-3) \;=\; \frac{1}{901},$$

mit dem Standard-Runge-Kutta-Verfahren der Ordnung $p = 4$ unter Verwendung der in Abschnitt 7.7 beschriebenen Schrittweitensteuerung. Zur Berechnung jeder neuen Schrittweite h_ℓ starte man mit $h^{(0)} = h_{\ell-1}$ (beziehungsweise im Fall $k = 0$ mit $h^{(0)} := 0.02$) und korrigiere gemäß Abschnitt 7.7 solange, bis (siehe Bemerkung 7.36) $\varepsilon/3 \leq \delta^{(k)} \leq 3\varepsilon$ oder $k = 20$ erfüllt ist, wobei $\varepsilon = 10^{-7}$ gilt. Für $\ell = 1, 2, \ldots, 50$ gebe man jeweils die Näherungswerte in t_ℓ sowie $y(t_\ell)$, $h_{\ell-1}$ und die Anzahl der Versuche k zur Bestimmung der Schrittweite h_ℓ an.

8 Mehrschrittverfahren für Anfangswertprobleme bei gewöhnlichen Differenzialgleichungen

Mit den in diesem Kapitel behandelten Mehrschrittverfahren zur näherungsweisen Bestimmung einer Lösung des Anfangswertproblems (7.1)–(7.2) (in Kurzschreibweise $y' = f(t,y),\ y(a) = y_0$) erhält man auf einfache Weise Verfahren höherer Konvergenzordnung.

8.1 Grundlegende Begriffe

8.1.1 Mehrschrittverfahren

Definition 8.1. Ein *m-Schrittverfahren* zur näherungsweisen Bestimmung einer Lösung des Anfangswertproblems $y' = f(t,y),\ y(a) = y_0$ besitzt auf einem äquidistantem Gitter die Form

$$\sum_{j=0}^{m} \alpha_j u_{\ell+j} \;=\; h\varphi(t_\ell, u_\ell, \ldots, u_{\ell+m}; h), \qquad \ell = 0,1,\ldots,n-m, \tag{8.1}$$

mit

- Koeffizienten $\alpha_j \in \mathbb{R}$ mit $\alpha_m \neq 0$ und einer Funktion

$$\varphi : [a,b] \times (\mathbb{R}^N)^{m+1} \times \mathbb{R}_+ \;\to\; \mathbb{R}^N, \tag{8.2}$$

- Gitterpunkten beziehungsweise Schrittweiten

$$t_\ell = a + \ell h \quad \text{für } \ell = 0,1,\ldots,n, \quad \text{mit } h = \frac{b-a}{n}, \tag{8.3}$$

- nicht näher spezifizierten Startwerten $u_0, \ldots, u_{m-1} \in \mathbb{R}^N$.

Ein m-Schrittverfahren bezeichnet man allgemeiner auch als *Mehrschrittverfahren*.

Bemerkung 8.2. (a) Üblicherweise setzt man $u_0 := y_0$, und die weiteren Startwerte $u_1, u_2, \ldots, u_{m-1} \in \mathbb{R}^N$ sind in einer *Anlaufrechnung* zu ermitteln.

(b) Nach der Anlaufrechnung wird für jedes $\ell \in \{0,1,\ldots,n-m\}$ so verfahren, dass aus den dann bereits bestimmten Näherungen $u_\ell, \ldots, u_{\ell+m-1} \in \mathbb{R}^N$ gemäß der Verfahrensvorschrift (8.1) die Näherung $u_{\ell+m} \in \mathbb{R}^N$ berechnet wird mit der Zielsetzung

$$u_{\ell+m} \;\approx\; y(t_{\ell+m}).$$

Hier bezeichnet $y : [a,b] \to \mathbb{R}^N$ die Lösung des Anfangswertproblems $y' = f(t,y)$, $y(a) = y_0$.

(c) Wie schon bei den Einschrittverfahren wird zwecks einer vereinfachten Notation der Definitionsbereich einer Funktion φ immer wie in (8.2) angegeben, obwohl bei den meisten noch vorzustellenden speziellen m-Schrittverfahren der Ausdruck $\varphi(t, v_0, \ldots, v_{m-1}; h)$ lediglich für Schrittweiten $h \leq (b-t)/m$ wohldefiniert ist.

(d) Hängt in der Verfahrensvorschrift (8.1) die rechte Seite tatsächlich von der Unbekannten $u_{\ell+m}$ ab, so spricht man von einem *impliziten* m-Schrittverfahren. Ist andererseits die Funktion φ unabhängig von $u_{\ell+m}$, so liegt ein *explizites* m-Schrittverfahren vor.

(e) Auf variablen Gittern, die hier nicht weiter behandelt werden, sind m-Schrittverfahren von der Form

$$\sum_{j=0}^{m} \alpha_j u_{\ell+j} \;=\; h_{\ell+m}\varphi\big(t_\ell, \ldots, t_{\ell+m}, u_\ell, \ldots, u_{\ell+m}; h_{\ell+m}\big), \qquad \ell = 0, 1, \ldots, n-m.$$

(f) Ist in der Verfahrensvorschrift (8.1) die Funktion φ von der speziellen Form

$$\varphi(t, v_0, \ldots, v_m; h) \;=\; \sum_{j=0}^{m} \beta_j f(t + jh, v_j),$$

so wird (8.1) als *lineares m-Schrittverfahren* bezeichnet. △

Beispiel 8.3. Ein spezielles lineares 2-Schrittverfahren ist die *Mittelpunktregel*,

$$u_{\ell+2} \;=\; u_\ell \;+\; 2hf(t_{\ell+1}, u_{\ell+1}), \qquad \ell = 0, 1, \ldots, n-2. \tag{8.4}$$

Ausführlich werden spezielle Mehrschrittverfahren in Abschnitt 8.3 behandelt. △

8.1.2 Konvergenz- und Konsistenzordnung

Die Approximationseigenschaften eines Mehrschrittverfahrens werden durch seine Konvergenzordnung beschrieben.

Definition 8.4. Ein Mehrschrittverfahren von der Form (8.1) zur Lösung des Anfangswertproblems $y' = f(t, y),\ y(a) = y_0$ besitzt die *Konvergenzordnung* $p \geq 1$, falls sich zu jeder Konstanten $c \geq 0$ und beliebigen Startwerten $u_0, \ldots, u_{m-1} \in \mathbb{R}^N$ mit $\|u_k - y(t_k)\| \leq ch^p$ für $k = 0, 1, \ldots, m-1$ der *globale Verfahrensfehler* in der Form

$$\max_{\ell=m,\ldots,n} \|u_\ell - y(t_\ell)\| \;\leq\; Kh^p$$

abschätzen lässt mit einer von der Schrittweite h unabhängigen Konstanten $K \geq 0$.

Hier und im Folgenden bezeichnet $\|\cdot\| : \mathbb{R}^N \to \mathbb{R}$ eine nicht näher spezifizierte Vektornorm. In Analogie zu den Einschrittverfahren spielen bei der Bestimmung der Konvergenzordnung eines Mehrschrittverfahrens die folgenden Begriffe eine wichtige Rolle.

Definition 8.5. Für ein Mehrschrittverfahren (8.1) zur Lösung des Anfangswertproblems $y' = f(t,y),\ y(a) = y_0$ bezeichnet

$$\left.\begin{aligned} \eta(t,h) \ := \ \Big[\sum_{j=0}^{m} \alpha_j y(t+jh) \Big] - h\varphi(t, y(t), y(t+h), \ldots, y(t+mh); h), \\ 0 \ < \ h \ \le \ \frac{b-t}{m}, \end{aligned} \right\} \tag{8.5}$$

den *lokalen Verfahrensfehler im Punkt* $(t, y(t))$ (bezüglich der Schrittweite h).

Definition 8.6. Ein Mehrschrittverfahren (8.1) zur Lösung des Anfangswertproblems $y' = f(t,y),\ y(a) = y_0$ besitzt die *Konsistenzordnung* $p \ge 1$, falls für eine Konstante C und eine hinreichend kleine Zahl $H > 0$ der lokale Verfahrensfehler die folgende Abschätzung erfüllt,

$$\| \eta(t,h) \| \ \le \ C h^{p+1}, \qquad a \le t \le b, \quad 0 \le h \le H.$$

Die *Konsistenzordnung* wird oft nur kurz als *Ordnung* eines Mehrschrittverfahrens bezeichnet.

8.1.3 Nullstabilität, Lipschitzbedingung

Bei der Behandlung der Konvergenzordnung eines Mehrschrittverfahrens wird auch die folgende Lipschitzbedingung an die Funktion $\varphi : [a,b] \times (\mathbb{R}^N)^{m+1} \times \mathbb{R}_+ \to \mathbb{R}^N$ aus der Verfahrensvorschrift (8.1) eine Rolle spielen,

$$\| \varphi(t, v_0, \ldots, v_m; h) - \varphi(t, w_0, \ldots, w_m; h) \| \le L_\varphi \sum_{j=0}^{m} \| v_j - w_j \| \quad (v_j, w_j \in \mathbb{R}^N). \tag{8.6}$$

Bemerkung 8.7. (a) Falls $f : [a,b] \times \mathbb{R}^N \to \mathbb{R}^N$ eine stetige Funktion ist, die die Lipschitzbedingung (7.4) erfüllt, so ist für lineare Mehrschrittverfahren die Lipschitzbedingung (8.6) erfüllt mit der Lipschitzkonstanten $L_\varphi = L \max_{j=0,\ldots,m} |\beta_j|$.

(b) Falls die Lipschitzbedingung (8.6) erfüllt ist, so ist für hinreichend kleine Schrittweiten h die Existenz und Eindeutigkeit der Approximationen des m-Schrittverfahrens (8.1) gewährleistet, da man die Bestimmungsgleichung für $u_{\ell+m}$ als Fixpunktgleichung schreiben kann, die für $0 < h < 1/(\alpha_m L_\varphi)$ einer Kontraktionsbedingung genügt. △

Schließlich ist bei den Konvergenzbetrachtungen für Mehrschrittverfahren die folgende Eigenschaft von Bedeutung.

Definition 8.8. Ein m-Schrittverfahren (8.1) zur Lösung von $y' = f(t,y),\ y(a) = y_0$ heißt *nullstabil*, falls für das *erzeugende Polynom*

$$\rho(\xi) \ := \ \alpha_m \xi^m \ + \ \alpha_{m-1} \xi^{m-1} \ + \ \ldots \ + \ \alpha_0 \ \in \ \Pi_m \tag{8.7}$$

die folgende *dahlquistsche Wurzelbedingung* erfüllt ist,

$$\begin{aligned} \rho(\xi) &= 0 \quad \Longrightarrow \quad |\xi| \le 1; \\ \rho(\xi) &= 0, \quad |\xi| = 1 \quad \Longrightarrow \quad \xi \text{ ist einfache Nullstelle von } \rho. \end{aligned}$$

8.1.4 Übersicht

Die nächsten Abschnitte des vorliegenden Kapitels behandeln die folgenden wichtigen Themen:

- Kriterien zur Bestimmung der Konvergenzordnung von allgemeinen Mehrschrittverfahren,
- Kriterien zur Bestimmung der Konsistenzordnung sowie Überprüfung der Nullstabilität allgemeiner Mehrschrittverfahren,
- Behandlung spezieller Mehrschrittverfahren.

8.2 Der globale Verfahrensfehler bei Mehrschrittverfahren

8.2.1 Das Konvergenztheorem

Es wird nun das wesentliche Konvergenzresultat für Mehrschrittverfahren vorgestellt.

Theorem 8.9. *Ein m-Schrittverfahren (8.1) für das Anfangswertproblem $y' = f(t,y)$, $y(a) = y_0$ sei nullstabil und die Funktion φ genüge der Lipschitzbedingung (8.6). Dann existieren Konstanten $K \ge 0$ und $H > 0$, so dass für $0 < h = (b-a)/n \le H$ die folgende Abschätzung gilt,*

$$\max_{\ell=0,\dots,n} \|u_\ell - y(t_\ell)\| \le K\Big[\max_{k=0,\dots,m-1} \|u_k - y(t_k)\| + \big(\max_{a\le t\le b-mh} \|\eta(t,h)\| \big)/h \Big]. \quad (8.8)$$

BEWEIS. Zur Vereinfachung der Notation nehmen wir im Folgenden $\alpha_m = 1$ an und betrachten den skalaren Fall $N = 1$. Mit den Setzungen

$$\begin{aligned} e_\ell &= u_\ell - y_\ell, \qquad y_\ell := y(t_\ell), \qquad \ell = 0,1,\dots,n, \\ \eta_\ell &= \eta(t_\ell, h), \qquad\qquad\qquad\quad \ell = 0,1,\dots,n-m, \end{aligned}$$

gelten für $\ell = 0,\dots,n-m$ die folgenden Darstellungen

$$\begin{aligned} \sum_{j=0}^{m} \alpha_j y_{\ell+j} &= h\varphi(t, y_\ell, \dots, y_{\ell+m}; h) + \eta_\ell, \\ \sum_{j=0}^{m} \alpha_j u_{\ell+j} &= h\varphi(t_\ell, u_\ell, \dots, u_{\ell+m}; h), \end{aligned}$$

und daher

$$\sum_{j=0}^{m} \alpha_j e_{\ell+j} \;=\; \underbrace{h[\varphi(t_\ell, u_\ell, \ldots, u_{\ell+m}; h) - \varphi(t, y_\ell, \ldots, y_{\ell+m}; h)]}_{=:\ \delta_\ell} - \eta_\ell. \quad (8.9)$$

Dieses lässt sich folgendermaßen schreiben,

$$\underbrace{\begin{pmatrix} e_{\ell+1} \\ e_{\ell+2} \\ \vdots \\ e_{\ell+m} \end{pmatrix}}_{=:\ E_{\ell+1}} = \underbrace{\begin{pmatrix} 0 & 1 & & \\ & \ddots & \ddots & \\ & & 0 & 1 \\ -\alpha_0 & \cdots & \cdots & -\alpha_{m-1} \end{pmatrix}}_{=:\ A} \underbrace{\begin{pmatrix} e_\ell \\ e_{\ell+1} \\ \vdots \\ e_{\ell+m-1} \end{pmatrix}}_{=:\ E_\ell} + \underbrace{\begin{pmatrix} 0 \\ \vdots \\ 0 \\ \delta_\ell - \eta_\ell \end{pmatrix}}_{=:\ F_\ell} \quad (8.10)$$

mit der Matrix $A \in \mathbb{R}^{m\times m}$ und den Vektoren $E_\ell,\ F_\ell \in \mathbb{R}^m$. Aus der Darstellung (8.10) erhält man mittels vollständiger Induktion die Beziehung

$$E_\ell \;=\; A^\ell E_0 \;+\; \sum_{\nu=0}^{\ell-1} A^{\ell-1-\nu} F_\nu, \qquad \ell \;=\; 0, 1, \ldots, n-m+1. \quad (8.11)$$

Zur Abschätzung der rechten Seite von (8.11) beobachtet man, dass die Wurzeln des erzeugenden Polynoms ρ mit den Eigenwerten der Matrix A übereinstimmen[1], und aufgrund der Nullstabilität erhält man aus dem nachzutragenden Lemma 8.15 die Beschränktheit der Potenzen der Matrix A, das heißt,

$$\|A^k\|_\infty \;\le\; C, \qquad k \;=\; 0, 1, \ldots, \quad (8.12)$$

mit einer Konstanten $C > 0$. Aus (8.11)–(8.12) resultiert die Abschätzung

$$\|E_\ell\|_\infty \;\le\; C\big[\|E_0\|_\infty \;+\; \sum_{\nu=0}^{\ell-1} \|F_\nu\|_\infty\big], \qquad \ell \;=\; 0, 1, \ldots, n-m+1. \quad (8.13)$$

Wegen (8.9) und (8.10) gilt

$$\begin{aligned} \|F_\nu\|_\infty \;=\; |\delta_\nu - \eta_\nu| \;&\le\; |\eta_\nu| \;+\; hL_\varphi \sum_{j=0}^{m} |e_{\nu+j}| \\ &\le\; \max_{j=0,\ldots,n-m} |\eta_j| \;+\; hL_\varphi m \|E_\nu\|_\infty \;+\; hL_\varphi \|E_{\nu+1}\|_\infty, \end{aligned}$$

und Summation ergibt

$$\sum_{\nu=0}^{\ell-1} \|F_\nu\|_\infty \;\le\; n[\max_{j=0,\ldots,n-m} |\eta_j|] \;+\; hc_1 \sum_{\nu=0}^{\ell-1} \|E_\nu\|_\infty \;+\; hL_\varphi \|E_\ell\|_\infty$$
$$\text{mit} \quad c_1 \;:=\; L_\varphi(m+1).$$

[1] Details hierzu findet man im Beweis von Lemma 5.16 im Kapitel über nichtlineare Gleichungssysteme.

Dies eingesetzt in (8.13) führt für $0 < h \le H$ mit einer Konstanten $H < 1/(CL_\varphi)$ auf folgende Abschätzung,

$$\|E_\ell\|_\infty \le \frac{C}{1-CL_\varphi H}\big(\|E_0\|_\infty + n[\max_{j=0,\dots,n-m}|\eta_j|]\big) + \frac{Cc_1}{1-CL_\varphi H}h\sum_{\nu=0}^{\ell-1}\|E_\nu\|_\infty,$$
$$\ell = 1,2,\dots,n-m+1.$$

Das ebenfalls noch nachzutragende diskrete Lemma von Gronwall 8.14 liefert dann die Behauptung, wenn man noch

$$\|E_0\|_\infty = \max_{\ell=0,\dots,m-1}|u_\ell - y(t_\ell)|, \qquad |u_\ell - y(t_\ell)| \le \|E_\ell\|_\infty,$$

berücksichtigt. □

Bemerkung 8.10. Dem Beweis von Theorem 8.9 entnimmt man noch, dass im Falle expliziter Verfahren $H = \infty$ als obere Schranke für die Schrittweiten gewählt werden kann und die wesentliche Fehlerabschätzung (8.8) für jede Schrittweite $h = (b-a)/n$ formal richtig ist. Es ist jedoch zu beachten, dass bei den noch zu behandelnden steifen Differenzialgleichungen (siehe Kapitel 8.9) der Fehler bei expliziten Verfahren erst für kleine Schrittweiten $h > 0$ klein ausfällt, was wegen der dort typischerweise großen Lipschitzkonstanten nicht im Widerspruch zur Fehlerabschätzung (8.8) steht. Hier ist der Einsatz impliziter Verfahren sinnvoller. Einzelheiten dazu werden in Abschnitt 8.9 vorgestellt. △

Als unmittelbare Folgerung aus Theorem 8.9 erhält man das folgende Korollar.

Korollar 8.11. *Ein nullstabiles m-Schrittverfahren (8.1) mit der Konsistenzordnung $p \ge 1$ und einer der Lipschitzbedingung (8.6) genügenden Funktion φ besitzt die Konvergenzordnung p.*

Es folgt ein Resultat über fehlerbehaftete Mehrschrittverfahren.

Korollar 8.12 (Rundungs- und Verfahrensfehleranalyse). *Ein m-Schrittverfahren (8.1) zur Lösung des Anfangswertproblems $y' = f(t,y),\ y(a) = y_0$ besitze die Konsistenzordnung $p \ge 1$ und sei nullstabil, und die Funktion φ genüge der Lipschitzbedingung (8.6). Für die Startwerte sei*

$$\max_{\ell=0,\dots,m-1}\|v_\ell - y(t_\ell)\| \le ch^p + \delta_1$$

erfüllt mit einer von h unabhängigen Konstanten $c \ge 0$. Für die Lösung der Gleichungen

$$\sum_{j=0}^{m}\alpha_j v_{\ell+j} = h\varphi(t_\ell, v_\ell, \dots, v_{\ell+m}; h) + \rho_\ell, \qquad \ell = 0,1,\dots,n-m,$$
$$\|\rho_\ell\| \le \delta_2, \qquad \text{———— « ————},$$

gilt dann die Fehlerabschätzung

$$\max_{\ell=0,\dots,n} \|v_\ell - y(t_\ell)\| \;\le\; K\big(h^p \;+\; \delta_1 \;+\; \frac{\delta_2}{h}\big)$$

mit einer von h unabhängigen Konstanten $K \ge 0$. Mit der Wahl $h = \delta_2^{1/(p+1)}$ erhält man

$$\max_{\ell=0,\dots,n} \|v_\ell - y(t_\ell)\| \;\le\; K\big(2\delta_2^{p/(p+1)} \;+\; \delta_1\big).$$

BEWEIS. Verläuft wie der Beweis von Theorem 8.9. Man hat dort nur $\eta_\ell = \eta(t_\ell, h_\ell) + \rho_\ell$ zu setzen. □

8.2.2 Hilfsresultat 1: Das Lemma von Gronwall

Als erster Nachtrag zum Beweis von Theorem 8.9 wird in diesem Abschnitt das diskrete Lemma von Gronwall vorgestellt. Vorbereitend hierzu wird die folgende kontinuierliche Fassung betrachtet.

Lemma 8.13 (Gronwall). *Für die riemann-integrierbare Funktion $\Phi : [0,T] \to \mathbb{R}$ sowie für Konstanten $\alpha,\ \beta \in \mathbb{R}$ mit $\beta > 0$ sei*

$$\Phi(t) \;\le\; \alpha \;+\; \beta \int_0^t \Phi(s)\, ds, \qquad t \;\in\; [0,T],$$

erfüllt. Dann gilt

$$\Phi(t) \;\le\; \alpha e^{\beta t}, \qquad t \;\in\; [0,T]. \tag{8.14}$$

BEWEIS. Mit der Notation

$$M \;:=\; \sup_{0\le t\le T} \Phi(t)$$

wird im Folgenden per Induktion über $n = 0, 1, \dots$ die folgende Abschätzung bewiesen,

$$\Phi(t) \;\le\; \alpha \sum_{\ell=0}^{n} \frac{(\beta t)^\ell}{\ell!} \;+\; M\frac{(\beta t)^{n+1}}{(n+1)!}, \qquad t \;\in\; [0,T]. \tag{8.15}$$

Der Grenzübergang $n \to \infty$ in (8.15) liefert dann die Abschätzung (8.14). Die Abschätzung (8.15) ist richtig für $n = 0$,

$$\Phi(t) \;\le\; \alpha \;+\; \beta \int_0^t \Phi(s)\, ds \;\le\; \alpha \;+\; \beta \int_0^t M\, ds \;=\; \alpha \;+\; M\beta t, \qquad t \;\in\; [0,T].$$

Wir nehmen nun an, dass für ein $n \in \mathbb{N}$ die Abschätzung (8.15) richtig ist mit $n-1$ anstelle n. Dann gilt

$$\begin{aligned}\Phi(t) \;&\le\; \alpha \;+\; \beta\int_0^t \Phi(s)\,ds \;\le\; \alpha \;+\; \beta\Big(\alpha\sum_{\ell=0}^{n-1}\frac{\beta^\ell}{\ell!}\int_0^t s^\ell\,ds \;+\; M\frac{\beta^n}{n!}\int_0^t s^n\,ds\Big)\\
&\le\; \alpha \;+\; \alpha\sum_{\ell=0}^{n-1}\frac{\beta^{\ell+1}}{\ell!}\frac{t^{\ell+1}}{\ell+1} \;+\; M\frac{\beta^{n+1}}{n!}\frac{t^{n+1}}{n+1} \;=\; \alpha \;+\; \alpha\sum_{\ell=1}^{n}\frac{(\beta t)^\ell}{\ell!} \;+\; M\frac{(\beta t)^{n+1}}{(n+1)!}\\
&=\; \alpha\sum_{\ell=0}^{n}\frac{(\beta t)^\ell}{\ell!} \;+\; M\frac{(\beta t)^{n+1}}{(n+1)!}, \qquad t \in [0,T],\end{aligned}$$

was den Beweis des Gronwall-Lemmas komplettiert. □

Eine unmittelbare Konsequenz aus dem Lemma von Gronwall ist das Resultat (7.5) über die stetige Abhängigkeit von den Anfangswerten bei einem Anfangswertproblem $y' = f(t,y),\ y(a) = y_0$. Hier soll das Lemma von Gronwall zum Beweis der folgenden diskreten Variante verwendet werden.

Lemma 8.14 (Diskrete Variante des Lemmas von Gronwall). *Es seien positive Zahlen $h_0,\dots,h_{r-1} > 0$ sowie Konstanten $\alpha \ge 0$ und $\beta \ge 0$ gegeben. Für Zahlen $v_0,\dots,v_r \in \mathbb{R}$ seien die folgenden Ungleichungen erfüllt,*

$$|v_0| \le \alpha, \qquad |v_\ell| \;\le\; \alpha \;+\; \beta\sum_{j=0}^{\ell-1} h_j|v_j| \qquad \text{für}\quad \ell = 1,2,\dots,r.$$

Dann gilt die folgende Abschätzung,

$$|v_\ell| \;\le\; \alpha\exp\big(\beta\sum_{j=0}^{\ell-1} h_j\big), \qquad \ell = 0,1,\dots,r.$$

BEWEIS. Es soll Lemma 8.13 angewandt werden, und hierzu betrachtet man mit der Notation $x_0 := 0$ und $x_{\ell+1} := x_\ell + h_\ell$ für $\ell = 0,1,\dots,r-1$ die Treppenfunktion

$$\Phi \;:=\; \sum_{\ell=0}^{r-1}|v_\ell|\chi_{[x_\ell,\,x_{\ell+1})} \;+\; |v_r|\chi_{\{x_r\}} \;:\; [0,T] \;\to\; \mathbb{R} \qquad (T := x_r),$$

wobei χ_M die charakteristische Funktion bezüglich einer gegebenen Menge M bezeichnet, es gilt also $\chi_M \equiv 1$ auf M und $\equiv 0$ außerhalb von M. Für beliebige $\ell \in \{0, 1,\dots,r-1\}$ und $t \in [x_\ell, x_{\ell+1})$, sowie auch für $\ell = r$ und $t = x_r$ gilt dann

$$\begin{aligned}\Phi(t) \;&=\; |v_\ell| \;\le\; \alpha \;+\; \beta\sum_{j=0}^{\ell-1} h_j|v_j| \;=\; \alpha \;+\; \beta\sum_{j=0}^{\ell-1}\int_{x_j}^{x_{j+1}}\Phi(s)\,ds\\
&=\; \alpha \;+\; \beta\int_0^{x_\ell}\Phi(s)\,ds \;\le\; \alpha \;+\; \beta\int_0^t\Phi(s)\,ds.\end{aligned}$$

Das Lemma von Gronwall liefert nun

$$|v_\ell| \;=\; \Phi(x_\ell) \;\le\; \alpha e^{\beta x_\ell} \;=\; \alpha\exp\bigl(\beta\sum_{j=0}^{\ell-1} h_j\bigr) \qquad \text{für} \quad \ell \;=\; 0,1,\ldots,r.$$

Dies komplettiert den Nachweis der Aussage der diskreten Variante des Lemmas von Gronwall. □

8.2.3 Beschränktheit der Matrixfolge A, A^2, $A^3, \ldots$

Das nachfolgende Lemma liefert den zweiten und letzten Nachtrag zum Beweis von Theorem 8.9. Zuvor führen wir noch die folgende Notation ein: einem Eigenwert $\lambda \in \mathbb{C}$ einer Matrix $A \in \mathbb{R}^{N\times N}$ *entsprechen nur lineare Elementarteiler*, falls die geometrische Vielfachheit von λ mit der algebraischen Vielfachheit übereinstimmt. Äquivalent dazu ist, dass alle zu λ gehörenden Jordanblöcke trivial sind.

Lemma 8.15. *Für eine gegebene Matrix $A \in \mathbb{R}^{N\times N}$ ist die Folge der Matrizen $A, A^2, A^3, \ldots$ beschränkt genau dann,*

(i) wenn der Spektralradius von A kleiner gleich eins ausfällt, $r_\sigma(A) \le 1$;

(ii) und wenn jedem Eigenwert $\lambda \in \mathbb{C}$ von A mit $|\lambda| = 1$ nur lineare Elementarteiler entsprechen.

BEWEIS. Für den Nachweis der Äquivalenz wird eine zu A ähnliche Matrix $J \in \mathbb{C}^{N\times N}$ in jordanscher Normalform herangezogen,

$$J \;=\; \begin{pmatrix} J_1 & & \\ & \ddots & \\ & & J_r \end{pmatrix}, \qquad J_\ell \;=\; \begin{pmatrix} \lambda_\ell & 1 & & \\ & \ddots & \ddots & \\ & & \ddots & 1 \\ & & & \lambda_\ell \end{pmatrix} \in \mathbb{C}^{N_\ell\times N_\ell}, \quad \ell \;=\; 1,2,\ldots,r,$$

wobei $N_\ell \ge 1$ und $\sum_{\ell=1}^{r} N_\ell = N$ gilt. Im Fall $N_\ell = 1$ bedeutet diese Notation $J_\ell = (\lambda_\ell) \in \mathbb{C}^{1\times 1}$.

Seien nun zuerst die Bedingungen (i) und (ii) erfüllt, es gilt also

$$|\lambda_\ell| \;\le\; 1; \qquad \text{im Fall} \quad |\lambda_\ell| = 1 \quad \text{sei} \;\; N_\ell \;=\; 1 \qquad\qquad (\ell \;=\; 1,2,\ldots,r\,). \tag{8.16}$$

Man wählt nun $\varepsilon > 0$ so klein, dass für jedes $\ell \in \{1,2,\ldots,r\}$ im Fall $N_\ell \ge 2$ die Ungleichung $|\lambda_\ell| + \varepsilon \le 1$ erfüllt ist, was aufgrund von (8.16) möglich ist. Dann betrachtet man

$$\widehat{J} \;=\; D^{-1}JD, \qquad\qquad D \;=\; \operatorname{diag}(1,\varepsilon,\varepsilon^2,\ldots,\varepsilon^{N-1}) \;\in\; \mathbb{R}^{N\times N},$$

und erhält unter Beachtung von $\widehat{J} = (\varepsilon^{k-j} J_{jk})$ Folgendes,

$$\widehat{J} \;=\; \begin{pmatrix} \widehat{J}_1 & & \\ & \ddots & \\ & & \widehat{J}_r \end{pmatrix}, \quad \widehat{J}_\ell \;=\; \begin{pmatrix} \lambda_\ell & \varepsilon & & \\ & \ddots & \ddots & \\ & & \ddots & \varepsilon \\ & & & \lambda_\ell \end{pmatrix} \in \mathbb{C}^{N_\ell\times N_\ell}, \quad \ell \;=\; 1,2,\ldots,r, \tag{8.17}$$

beziehungsweise $\widehat{J}_\ell = (\lambda_\ell) \in \mathbb{C}^{1\times 1}$ im Fall $N_\ell = 1$. Aufgrund der Konstruktion gilt

$$\|\widehat{J}\|_\infty \;=\; \max_{\ell=1,\dots,r} \|\widehat{J}_\ell\|_\infty \;\le\; 1$$

und daher

$$\|\widehat{J}^\nu\|_\infty \;\le\; 1, \qquad \nu \;=\; 1, 2, \dots \;.$$

Die Ähnlichkeit der Matrizen A und J impliziert $A = T^{-1}JT$ mit einer regulären Matrix $T \in \mathbb{C}^{N\times N}$, und damit gilt

$$A^\nu \;=\; T_1^{-1}\widehat{J}^\nu T_1, \qquad \nu \;=\; 0, 1, \dots, \quad \text{mit} \quad T_1 \;:=\; D^{-1}T.$$

Daher ist also auch die Matrixfolge $A^1,\ A^2, \dots$ beschränkt.

Wir nehmen nun umgekehrt an, dass eine der beiden Bedingungen (i), (ii) nicht erfüllt ist. Wenn die Bedingung (i) nicht erfüllt ist, so gilt für ein $1 \le \ell \le r$ die Ungleichung $|\lambda_\ell| > 1$, und dann betrachte man im Fall $N_\ell \ge 2$ etwa die Vektorfolge

$$\begin{pmatrix} \lambda_\ell & 1 & & \\ & \ddots & \ddots & \\ & & \ddots & 1 \\ & & & \lambda_\ell \end{pmatrix}^\nu \begin{pmatrix} 1 \\ 0 \\ \vdots \\ 0 \end{pmatrix} \;=\; \begin{pmatrix} \lambda_\ell^\nu \\ 0 \\ \vdots \\ 0 \end{pmatrix}, \qquad \nu \;=\; 0, 1, \dots,$$

und für $N_\ell = 1$ gilt $J_\ell^\nu = (\lambda_\ell^\nu) \in \mathbb{C}^{1\times 1}$. Falls (ii) nicht erfüllt ist, so gilt für ein $1 \le \ell \le r$ sowohl $|\lambda_\ell| = 1$ als auch $N_\ell \ge 2$, und hier betrachte man beispielsweise

$$\begin{pmatrix} \lambda_\ell & 1 & & \\ & \lambda_\ell & \ddots & \\ & & \ddots & 1 \\ & & & \lambda_\ell \end{pmatrix}^\nu \begin{pmatrix} 0 \\ 1 \\ 0 \\ \vdots \\ 0 \end{pmatrix} \;=\; \begin{pmatrix} \nu\lambda_\ell^{\nu-1} \\ \lambda_\ell^\nu \\ 0 \\ \vdots \\ 0 \end{pmatrix}, \qquad \nu \;=\; 0, 1, \dots \;.$$

In jedem Fall ist wegen

$$J^\nu \;=\; \begin{pmatrix} J_1^\nu & & \\ & \ddots & \\ & & J_r^\nu \end{pmatrix}, \qquad \nu \;=\; 0, 1, \dots,$$

dann die Matrix J und damit auch die zu J ähnliche Matrix A nicht potenzbeschränkt. Die Aussage des Lemmas ist damit vollständig nachgewiesen. □

8.2.4 Die Konsistenzordnung linearer Mehrschrittverfahren

Zum Abschluss der allgemeinen Betrachtungen über Mehrschrittverfahren wird in dem folgenden Lemma ein einfaches Kriterium zur Bestimmung der Konsistenzordnung eines linearen Mehrschrittverfahrens vorgestellt.

Lemma 8.16. *Sind für das lineare m-Schrittverfahren*

$$\sum_{j=0}^{m} \alpha_j u_{\ell+j} \;=\; h \sum_{j=0}^{m} \beta_j f(t_{\ell+j}, u_{\ell+j}), \qquad \ell \;=\; 0, 1, \ldots, n-m,$$

mit einer p-mal stetig partiell differenzierbaren Funktion $f : [a,b] \times \mathbb{R}^N \to \mathbb{R}^N$ (für eine Zahl $p \geq 1$) die Gleichungen

$$\sum_{j=0}^{m} \{ j^{\nu} \alpha_j \;-\; \nu j^{\nu-1} \beta_j \} \;=\; 0, \qquad \nu \;=\; 0, 1, \ldots, p, \tag{8.18}$$

erfüllt, so ist das m-Schrittverfahren konsistent von der Ordnung p. Für eine $(p+1)$-mal stetig partiell differenzierbare Funktion f gilt mehr noch die Darstellung

$$\left.\begin{aligned} \eta(t,h) \;&=\; C_{p+1}\, y^{(p+1)}(t) h^{p+1} \;+\; \mathcal{O}(h^{p+2}) \qquad \text{für} \quad h \to 0, \\ &\text{mit} \quad C_{p+1} \;:=\; \sum_{j=0}^{m} \Big[\frac{j^{p+1}\alpha_j}{(p+1)!} \;-\; \frac{j^p \beta_j}{p!} \Big]. \end{aligned}\right\} \tag{8.19}$$

BEWEIS. Die Lösung des Anfangswertproblems $y' = f(t,y),\ y(a) = y_0$ ist nach Theorem 7.3 $(p+1)$-mal stetig partiell differenzierbar. Taylorentwicklung der Funktionen y und y' in dem Punkt $t \in [a, b-mh]$ ergibt

$$\begin{aligned} y(t+jh) \;&=\; \sum_{\nu=0}^{p} \frac{y^{(\nu)}(t)}{\nu!} j^{\nu} h^{\nu} \;+\; \mathcal{O}(h^{p+1}), \\ y'(t+jh) \;&=\; \sum_{\nu=0}^{p-1} \frac{y^{(\nu+1)}(t)}{\nu!} j^{\nu} h^{\nu} \;+\; \mathcal{O}(h^{p}) \;=\; \sum_{\nu=0}^{p} \nu \frac{y^{(\nu)}(t)}{\nu!} j^{\nu-1} h^{\nu-1} \;+\; \mathcal{O}(h^{p}). \end{aligned}$$

Für den lokalen Verfahrensfehler folgt daraus

$$\begin{aligned} \eta(t,h) \;&=\; \sum_{j=0}^{m} \alpha_j y(t+jh) \;-\; h \sum_{j=0}^{m} \beta_j f(t+jh, y(t+jh)) \\ &=\; \sum_{j=0}^{m} [\alpha_j y(t+jh) \;-\; h\beta_j y'(t+jh)] \\ &=\; \sum_{\nu=0}^{p} [\underbrace{\sum_{j=0}^{m} j^{\nu}\alpha_j \;-\; \nu j^{\nu-1}\beta_j}_{=\;0}] \frac{y^{(\nu)}(t)}{\nu!} h^{\nu} \;+\; \mathcal{O}(h^{p+1}), \quad 0 \;<\; h \leq \frac{b-t}{m}. \end{aligned} \tag{8.20}$$

Die Darstellung (8.19) folgt durch die gleiche Entwicklung wie in (8.20), mit p ersetzt durch $p+1$. □

Bemerkung 8.17. (a) Die ersten beiden Gleichungen des Systems (8.18) bedeuten ausgeschrieben

$$\underbrace{\sum_{j=0}^{m} \alpha_j}_{=\,\rho(1)} = 0 \quad \text{für } \nu = 0, \qquad \underbrace{\sum_{j=1}^{m} j\alpha_j}_{=\,\rho'(1)} = \sum_{j=0}^{m} \beta_j \quad \text{für } \nu = 1,$$

wobei $\rho(\xi) = \alpha_m \xi^m + \ldots + \alpha_0$ das zugehörige erzeugende Polynom bezeichnet. Insbesondere implizieren Nullstabilität und Konsistenzordnung $p \geq 1$ notwendigerweise $\rho'(1) \neq 0$.

(b) Die Approximationen $u_0, \ldots, u_{n-m}$ des Mehrschrittverfahrens (8.1) bleiben unverändert, wenn die Verfahrensvorschrift (8.1) mit einer beliebigen Konstanten $\neq 0$ multipliziert wird; in diesem Sinne sind also sowohl der lokale Verfahrensfehler $\eta(t,h)$ als auch die Konstante C_{p+1} in (8.19) nicht eindeutig festgelegt. Als (die von p abhängige) *Fehlerkonstante* bezeichnet man die normierte Größe $C_{p+1}/\rho'(1)$.

(c) Die Konsistenzordnung der noch zu betrachtenden speziellen linearen Mehrschrittverfahren lässt sich auch anders als mit Lemma 8.16 bestimmen, wie sich noch herausstellen wird. △

8.3 Spezielle lineare Mehrschrittverfahren – Vorbereitungen

Die meisten der vorzustellenden Mehrschrittverfahren beruhen auf der Anwendung interpolatorischer Quadraturformeln auf äquidistanten Gittern. Vorbereitend werden in diesem Abschnitt Darstellungen für Interpolationspolynome auf äquidistanten Gittern geliefert.

Definition 8.18. Für einen gegebenen Datensatz $g_0, g_1, \ldots, g_r \in \mathbb{R}^N$ sind die *Rückwärtsdifferenzen* $\nabla^k g_\nu \in \mathbb{R}^N$ für $0 \leq k \leq \nu \leq r$ rekursiv erklärt durch

$$\begin{aligned} \nabla^0 g_\nu &= g_\nu, && \nu = 0,1,\ldots,r, \\ \nabla^k g_\nu &= \nabla^{k-1} g_\nu - \nabla^{k-1} g_{\nu-1}, && \nu = k, k+1, \ldots, r \qquad (k = 1,2,\ldots,r). \end{aligned}$$

Die bei den Rückwärtsdifferenzen auftretenden Zusammenhänge sind in Schema 8.1 dargestellt.

$$
\begin{array}{lllllllll}
\nabla^0 g_0 = g_0 & & & & & & & & \\
& \searrow & & & & & & & \\
\nabla^0 g_1 = g_1 & \to & \nabla^1 g_1 & & & & & & \\
& \searrow & & \searrow & & & & & \\
\nabla^0 g_2 = g_2 & \to & \nabla^1 g_2 & \to & \nabla^2 g_2 & & & & \\
\vdots & & \vdots & & \vdots & \ddots & & & \\
\nabla^0 g_{r-1} = g_{r-1} & \to & \nabla^1 g_{r-1} & \to & \cdots & \cdots & \nabla^{r-1} g_{r-1} & & \\
& \searrow & & \searrow & & & & \searrow & \\
\nabla^0 g_r = g_r & \to & \nabla^1 g_r & \to & \cdots & \cdots & \nabla^{r-1} g_r & \to & \nabla^r g_r
\end{array}
$$

Schema 8.1: Abhängigkeiten der Rückwärtsdifferenzen

Lemma 8.19. *Für die Rückwärtsdifferenzen $\nabla^k g_\nu \in \mathbb{R}^N$ eines gegebenen Datensatzes $g_0, g_1, \ldots, g_r \in \mathbb{R}^N$ gilt*

$$\nabla^k g_\nu = \sum_{j=0}^{k} (-1)^j \binom{k}{j} g_{\nu-j}, \qquad 0 \le k \le \nu \le r. \tag{8.21}$$

BEWEIS. Es bezeichne S den Rückwärtsshift,

$$S g_\nu := g_{\nu-1}, \qquad \nu = 1, 2, \ldots, r.$$

Wenn man dann die Operatoren $(I - S)^k$ und S^j in naheliegender Weise rekursiv erklärt, so erhält man mit dem binomischen Satz

$$\nabla^k g_\nu = (I - S)^k g_\nu = \sum_{j=0}^{k} (-1)^j \binom{k}{j} S^j g_\nu = \sum_{j=0}^{k} (-1)^j \binom{k}{j} g_{\nu-j}. \qquad \square$$

Die folgenden Darstellungen für das Interpolationspolynom und den zugehörigen Interpolationsfehler bei äquidistanten Stützstellen dienen als Vorbereitung auf die Behandlung spezieller Mehrschrittverfahren.

Lemma 8.20. *Gegeben seien insgesamt $r+1$ äquidistante Stützstellen $x_\ell = x_0 + \ell h$ für $\ell = 0, 1, \ldots, r$, mit Zahlen $x_0 \in \mathbb{R}$ und $h > 0$. Dann besitzt das zu gegebenen Vektoren $g_0, g_1, \ldots, g_r \in \mathbb{R}^N$ gehörende eindeutig bestimmte (vektorwertige) interpolierende Polynom $\mathcal{P} \in \Pi_r^N$ die Darstellung*

$$\mathcal{P}(x_r + sh) = \sum_{k=0}^{r} (-1)^k \binom{-s}{k} \nabla^k g_r, \qquad s \in \mathbb{R}. \tag{8.22}$$

Hierbei gelten die folgenden Identitäten,

$$\binom{-s}{k} = \frac{(-s)(-s-1)\cdots(-s-k+1)}{k!} = \frac{(-1)^k}{k!}s(s+1)\cdots(s+k-1), \quad (8.23)$$

und es bezeichnet

$$\Pi_r^N := \{\mathcal{P}(t) = \sum_{k=0}^{r} a_k t^k, \quad \text{mit} \quad a_k \in \mathbb{R}^N\}.$$

BEWEIS VON LEMMA 8.20. Für die newtonsche Darstellung des Polynoms $\mathcal{P}$ erhält man unter Verwendung von (8.23) und den Resultaten aus Abschnitt 1.4 Folgendes,

$$\begin{aligned}\mathcal{P}(x_r+sh) &= a_0 + a_1(x_r+sh-x_r) + \ldots + a_r(x_r+sh-x_r)\cdots(x_r+sh-x_1)\\ &= \sum_{k=0}^{r} a_k \prod_{j=0}^{k-1}(x_r + sh - x_{r-j}) = \sum_{k=0}^{r} a_k \prod_{j=0}^{k-1}(x_r + sh - (x_r - jh))\\ &= \sum_{k=0}^{r} a_k h^k \prod_{j=0}^{k-1}(s+j) = \sum_{k=0}^{r} a_k h^k (-1)^k k! \binom{-s}{k} \qquad (8.24)\end{aligned}$$

mit den dividierten Differenzen

$$a_k = g[x_r, \ldots, x_{r-k}] \in \mathbb{R}^N, \qquad k = 0, 1, \ldots, r. \qquad (8.25)$$

Die Aussage des Lemmas erhält man nun aus (8.24)–(8.25) zusammen mit der folgenden Darstellung für die dividierten Differenzen,

$$g[x_\ell, \ldots, x_{\ell-k}] = \frac{\nabla^k g_\ell}{k! h^k}, \qquad 0 \le k \le \ell \le r,$$

die man mittels vollständiger Induktion über $k = 0, 1, \ldots, r$ erhält:

$$\begin{aligned} g[x_\ell] &= g_\ell = \nabla^0 g_\ell, \qquad \ell = 0, 1, \ldots, r;\\ g[x_\ell, \ldots, x_{\ell-k}] &= \frac{g[x_\ell, \ldots, x_{\ell-k+1}] - g[x_{\ell-1}, \ldots, x_{\ell-k}]}{kh}\\ &= \frac{\nabla^{k-1} g_\ell - \nabla^{k-1} g_{\ell-1}}{((k-1)! h^{k-1})kh} = \frac{\nabla^k g_\ell}{k! h^k}, \qquad \ell = k, k+1, \ldots, r. \end{aligned}$$

□

Lemma 8.21. *Zu einer gegebenen Funktion $g \in C^{r+1}([c,d], \mathbb{R}^N)$ und zu den äquidistanten Stützstellen $x_\ell = x_0 + \ell h \in [c,d]$ für $\ell = 0, 1, \ldots, r$ bezeichne $\mathcal{P} \in \Pi_r^N$ das zugehörige (vektorwertige) interpolierende Polynom. Der Interpolationsfehler in $x_r + sh \in [c,d]$ besitzt die Darstellung*

$$\left.\begin{aligned} g(x_r+sh) - \mathcal{P}(x_r+sh) &= (-1)^{r+1}\binom{-s}{r+1} F(s) h^{r+1},\\ F(s) &= (g_j^{(r+1)}(\xi_j(s)))_{j=1,\ldots,N} \in \mathbb{R}^N, \end{aligned}\right\} \quad (8.26)$$

mit geeigneten Zwischenstellen $\xi_j(s) \in [c,d]$ für $j = 1, 2, \ldots, N$.

BEWEIS. Aus Abschnitt 1.5 ist die folgende Fehlerdarstellung bekannt,

$$g_j(x_r + sh) - \mathcal{P}_j(x_r + sh) = \frac{\omega(x_r + sh)\, g_j^{(r+1)}(\xi_j(s))}{(r+1)!},$$

mit $\omega(x) = (x - x_0)\cdots(x - x_r)$, und $\mathcal{P}_j$ bezeichnet die j-te Komponente des vektorwertigen Polynoms $\mathcal{P}$. Die Aussage des Lemmas folgt dann mit der Darstellung (8.23),

$$\begin{aligned}\omega(x_r + sh) &= \prod_{j=0}^{r} (x_r + sh - (x_r - jh)) = h^{r+1} \prod_{j=0}^{r} (s + j) \\ &= h^{r+1} (-1)^{r+1} \binom{-s}{r+1} (r+1)! \,.\end{aligned}$$

□

8.4 Adams-Verfahren

8.4.1 Der Ansatz

Zur Herleitung der ersten Klasse von Mehrschrittverfahren beobachtet man, dass die Lösung $y : [a, b] \to \mathbb{R}^N$ des Anfangswertproblems $y' = f(t, y),\ y(a) = y_0$ auch der folgenden Integralrelation genügt,

$$y(t_{\ell+m}) - y(t_{\ell+m-1}) = \int_{t_{\ell+m-1}}^{t_{\ell+m}} f(t, y(t))\, dt, \qquad \ell = 0, 1, \ldots, n-m, \quad (8.27)$$

was man durch Integration der Differenzialgleichung $y' = f(t, y(t))$ von $t_{\ell+m-1}$ bis $t_{\ell+m}$ erhält. Adams-Verfahren gewinnt man nun durch Ersetzen des Integranden durch geeignete Polynome $\mathcal{P}$,

$$u_{\ell+m} - u_{\ell+m-1} = \int_{t_{\ell+m-1}}^{t_{\ell+m}} \mathcal{P}(t)\, dt, \qquad \ell = 0, 1, \ldots, n-m. \quad (8.28)$$

Je nach der speziellen Wahl von $\mathcal{P}$ erhält man explizite beziehungsweise implizite Verfahren. Im Folgenden werden Einzelheiten hierzu vorgestellt.

8.4.2 Adams-Bashfort-Verfahren

Definition 8.22. Für $m \geq 1$ erhält man das m-schrittige *Adams-Bashfort-Verfahren* durch den Ansatz (8.28) mit

$$\left.\begin{aligned} \mathcal{P} \in \Pi_{m-1}^N, \qquad \mathcal{P}(t_j) &= f_j, \qquad j = \ell, \ell+1, \ldots, \ell+m-1, \\ f_j := f(t_j, u_j), \qquad & \text{---------- « ----------}. \end{aligned}\right\} \quad (8.29)$$

Die vorliegende Situation ist in Bild 8.1 veranschaulicht.

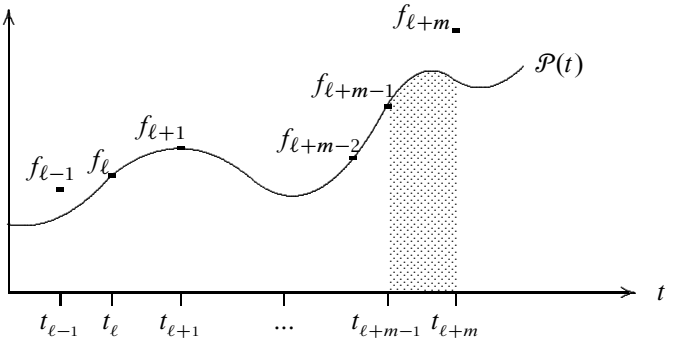


Bild 8.1: Vorgehensweise des m-schrittigen Adams-Bashfort-Verfahrens im eindimensionalen Fall

Das folgende Theorem liefert eine explizite Darstellung für das Adams-Bashfort-Verfahren:

Theorem 8.23. *Das m-schrittige Adams-Bashfort-Verfahren hat die Gestalt*

$$u_{\ell+m} - u_{\ell+m-1} = h \sum_{k=0}^{m-1} \gamma_k \nabla^k f_{\ell+m-1}, \qquad \ell = 0, 1, \ldots, n-m, \tag{8.30}$$

mit den von m unabhängigen Koeffizienten

$$\gamma_k := (-1)^k \int_0^1 \binom{-s}{k} ds, \qquad k = 0, 1, \ldots, \tag{8.31}$$

die sich rekursiv berechnen durch

$$\tfrac{1}{k+1}\gamma_0 + \tfrac{1}{k}\gamma_1 + \tfrac{1}{k-1}\gamma_2 + \ldots + \tfrac{1}{2}\gamma_{k-1} + \gamma_k = 1 \quad \textit{für } k = 0, 1, \ldots . \tag{8.32}$$

BEWEIS. Die Darstellung (8.30)–(8.31) folgt umgehend aus Lemma 8.20 mit $x_j = t_{\ell+j}$ für $j = 0, 1, \ldots, m-1$,

$$\int_{t_{\ell+m-1}}^{t_{\ell+m}} \mathcal{P}(t)\, dt = h \int_0^1 \mathcal{P}(t_{\ell+m-1} + sh)\, ds = h \sum_{k=0}^{m-1} \underbrace{(-1)^k \int_0^1 \binom{-s}{k} ds}_{\gamma_k} \nabla^k f_{\ell+m-1}. \tag{8.33}$$

Bei dem Nachweis der Rekursionsformel (8.32) für die Koeffizienten γ_k bedient man sich der erzeugenden Funktion

$$\begin{aligned} G(t) &:= \sum_{k=0}^{\infty} \gamma_k t^k = \sum_{k=0}^{\infty} (-t)^k \int_0^1 \binom{-s}{k} ds \overset{(*)}{=} \int_0^1 \Big[\sum_{k=0}^{\infty} \binom{-s}{k} (-t)^k \Big] ds \\ &= \int_0^1 (1-t)^{-s}\, ds = -\frac{1}{\ln(1-t)} (1-t)^{-s} \Big|_{s=0}^{s=1} \\ &= -\frac{t}{(1-t)\ln(1-t)}, \qquad -1 < t < 1. \end{aligned} \tag{8.34}$$

Hier folgt $(*)$ durch Vertauschen von Reihenentwicklung und Integration, was aufgrund der bei festem $-1 < t < 1$ gleichmäßigen Konvergenz von $\sum_{k=0}^{\infty} \binom{-s}{k}(-t)^k$ bezüglich $s \in [0,1]$ (in unserer Situation gilt[2] $|\binom{-s}{k}| \leq 1$) zulässig ist. Die Darstellung (8.34) für $G(t)$ liefert

$$G(t)\frac{-\ln(1-t)}{t} \;=\; \frac{1}{1-t}, \qquad |t| < 1,$$

beziehungsweise in Potenzreihenschreibweise

$$\big(\gamma_0 + \gamma_1 t + \gamma_2 t^2 + \ldots\big)\big(1 + \tfrac{t}{2} + \tfrac{t^2}{3} + \ldots\big) = \big(1 + t + t^2 + \ldots\big), \tag{8.35}$$

und ein Vergleich der Koeffizienten von $t^0, t^1, t^2, \ldots$ auf den beiden Seiten der Gleichung (8.35) ergibt die Aussage (8.32). □

Bemerkung 8.24. (a) Das m-schrittige Adams-Bashfort-Verfahren (8.30) lässt sich in eindeutiger Weise in der Form

$$u_{\ell+m} - u_{\ell+m-1} \;=\; h\sum_{j=0}^{m-1} \beta_{m,j}\, f_{\ell+j}, \qquad \ell \;=\; 0,1,\ldots,n-m, \tag{8.36}$$

schreiben mit den von der Zahl m abhängigen Koeffizienten $\beta_{m,0}, \beta_{m,1}, \ldots, \beta_{m,m-1} \in \mathbb{R}$, denn (8.21) ergibt unmittelbar

$$\begin{aligned}\sum_{k=0}^{m-1} \gamma_k \nabla^k f_{\ell+m-1} \;&=\; \sum_{k=0}^{m-1}\sum_{j=0}^{k} (-1)^j \tbinom{k}{j} \gamma_k f_{\ell+m-1-j} \\ &=\; \sum_{j=0}^{m-1} \underbrace{\Big[\,(-1)^j \sum_{k=j}^{m-1} \tbinom{k}{j} \gamma_k\,\Big]}_{=:\,\beta_{m,m-1-j}} f_{\ell+m-1-j}.\end{aligned}$$

(b) Aus der Rekursionsformel (8.32) berechnen sich die ersten vier Koeffizienten $\gamma_0, \ldots, \gamma_3 \in \mathbb{R}$ zu

$$\gamma_0 \;=\; 1, \qquad \gamma_1 \;=\; \tfrac{1}{2}, \qquad \gamma_2 \;=\; \tfrac{5}{12}, \qquad \gamma_3 \;=\; \tfrac{3}{8}.$$

Für $m = 1, \ldots, 4$ lauten die m-schrittigen Adams-Bashfort-Verfahren in der klassischen Darstellung eines linearen Mehrschrittverfahrens folgendermaßen,

$$\begin{aligned}
&m = 1: \; u_{\ell+1} = u_\ell + h f_\ell, && \ell = 0, \ldots, n-1;\\
&m = 2: \; u_{\ell+2} = u_{\ell+1} + \tfrac{h}{2}(3 f_{\ell+1} - f_\ell), && \text{\quad\quad} \ell = 0, \ldots, n-2;\\
&m = 3: \; u_{\ell+3} = u_{\ell+2} + \tfrac{h}{12}(23 f_{\ell+2} - 16 f_{\ell+1} + 5 f_\ell), && \text{\quad\quad} \ell = 0, \ldots, n-3;\\
&m = 4: \; u_{\ell+4} = u_{\ell+3} + \tfrac{h}{24}(55 f_{\ell+3} - 59 f_{\ell+2} + 37 f_{\ell+1} - 9 f_\ell), && \text{\quad\quad} \ell = 0, \ldots, n-4.
\end{aligned}$$

Insbesondere erhält man im Fall $m = 1$ das klassische Euler-Verfahren. △

[2] siehe (8.23)

Das folgende Theorem stellt die wesentlichen Eigenschaften der Adams-Bashfort-Verfahren heraus:

Theorem 8.25. *Das m-schrittige Adams-Bashfort-Verfahren ist nullstabil. Im Fall $f \in C^m([a,b]\times\mathbb{R}^N,\mathbb{R}^N)$ besitzt es die Konsistenzordnung $p=m$, und die Fehlerkonstante lautet γ_m.*

BEWEIS. Das zugehörige erzeugende Polynom ist

$$\rho(\xi) \;=\; \xi^{m-1}(\xi-1),$$

so dass die dahlquistsche Wurzelbedingung offensichtlich erfüllt ist. Für den Nachweis der Konsistenzordnung betrachtet man den lokalen Verfahrensfehler,

$$\begin{aligned}
\eta(t,h) &\overset{(*)}{=} y(t+mh) \;-\; y(t+(m-1)h) \;-\; h\sum_{j=0}^{m-1}\beta_{m,j}\,y'(t+jh)\\
&\overset{(**)}{=} \text{——————} \;«\; \text{——————} \;-\; h\sum_{k=0}^{m-1}\gamma_k\nabla^k y'(t+(m-1)h)\\
&\overset{(\bullet)}{=} \int_{t+(m-1)h}^{t+mh} y'(s) \;-\; \mathcal{P}(s)\,ds,
\end{aligned}$$

mit

$$\mathcal{P}\in\Pi_{m-1}^N, \qquad \mathcal{P}(t+jh) \;=\; y'(t+jh), \qquad \text{für} \quad j \;=\; 0,1,\ldots,m-1,$$

wobei $\nabla^k y'(t+(m-1)h)$ die Rückwärtsdifferenzen bezüglich der Folge $y'(t), y'(t+h),\ldots,y'(t+(m-1)h)$ bezeichnen. Die Identitäten $(*)$ und $(**)$ resultieren dabei unmittelbar aus der Verfahrensdarstellung (8.36) sowie der daran anschließenden Begründung, und die Identität $(\bullet)$ folgt mit Lemma 8.20 (siehe auch (8.33) im Beweis von Theorem 8.23). Mit der Darstellung (8.26) für den Interpolationsfehler erhält man dann

$$\begin{aligned}
\eta(t,h) &= h\int_0^1 y'(t+(m-1+s)h) \;-\; \mathcal{P}(t+(m-1+s)h)\,ds\\
&= h^{m+1}(-1)^m\int_0^1 \binom{-s}{m}\,F(s)\,ds \;=\; \mathcal{O}(h^{m+1}) \qquad \text{für} \quad h\to 0,\\
&\qquad \text{mit} \quad F(s) \;=\; \left(y_j^{(m+1)}(\xi_j(s))\right)_{j=1,..,N}, \qquad \xi_j(s)\in[t,t+mh].
\end{aligned}$$

Im Fall $f\in C^{m+1}([a,b]\times\mathbb{R}^N,\mathbb{R}^N)$ verwendet man

$$y_j^{(m+1)}(\xi_j(s)) \;=\; y_j^{(m+1)}(t) \;+\; \mathcal{O}(h) \qquad \text{für} \quad h\to 0$$

und folgert mit der Definition (8.31) für γ_m die folgende Darstellung für den lokalen Verfahrensfehler,

$$\eta(t,h) \;=\; \gamma_m y^{(m+1)}(t)h^{m+1} \;+\; \mathcal{O}(h^{m+2}) \qquad \text{für} \quad h\to 0.$$

Wegen $\rho'(1)=1$ ist γ_m die Fehlerkonstante. □

8.4.3 Adams-Moulton-Verfahren

Definition 8.26. Für $m \geq 1$ erhält man das m-schrittige *Adams-Moulton-Verfahren* durch den Ansatz (8.28) mit

$$\left.\begin{aligned} \mathcal{P} \in \Pi_m^N, \qquad \mathcal{P}(t_j) &= f_j, \qquad j = \ell, \ell+1, \ldots, \ell+m, \\ f_j &:= f(t_j, u_j), \qquad \text{\quad\quad « \quad\quad .} \end{aligned}\right\} \tag{8.37}$$

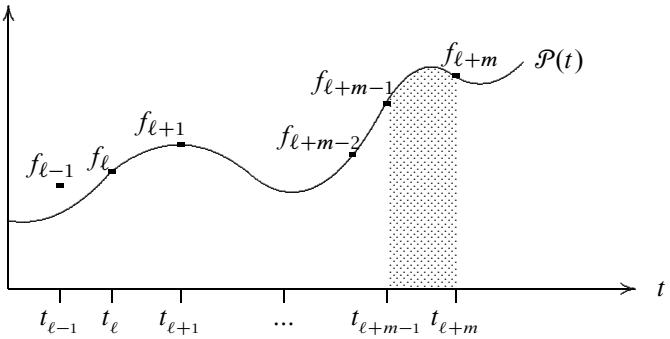


Bild 8.2: Vorgehensweise des m-schrittigen Adams-Moulton-Verfahrens im eindimensionalen Fall

Die folgenden Resultate über das Adams-Moulton-Verfahren lassen sich genauso wie die Resultate über die Adams-Bashfort-Verfahren erzielen. Daher wird hier auf die jeweiligen Nachweise verzichtet.

Theorem 8.27. *Das m-schrittige Adams-Moulton-Verfahren hat die Gestalt*

$$u_{\ell+m} - u_{\ell+m-1} = h \sum_{k=0}^{m} \gamma_k^* \nabla^k f_{\ell+m}, \qquad \ell = 0, 1, \ldots, n-m,$$

mit den von m unabhängigen Koeffizienten

$$\gamma_k^* := (-1)^k \int_{-1}^{0} \binom{-s}{k} ds, \qquad \text{für } k = 0, 1, \ldots,$$

die sich rekursiv berechnen durch $\gamma_0^* = 1$ und

$$\frac{1}{k+1}\gamma_0^* + \frac{1}{k}\gamma_1^* + \frac{1}{k-1}\gamma_2^* + \ldots + \frac{1}{2}\gamma_{k-1}^* + \gamma_k^* = 0 \quad \text{für } k = 1, 2, \ldots . \tag{8.38}$$

Bemerkung 8.28. (a) Das m-schrittige Adams-Moulton-Verfahren lässt sich in eindeutiger Weise in der Form

$$u_{\ell+m} - u_{\ell+m-1} = h \sum_{j=0}^{m} \beta_{m,j}^* f_{\ell+j}, \qquad \ell = 0, 1, \ldots, n-m,$$

schreiben mit den von der Zahl m abhängigen Koeffizienten

$$\beta^*_{m,m-j} \;=\; (-1)^j \sum_{k=j}^{m} \binom{k}{j} \gamma^*_k, \qquad j \;=\; 0,1,\ldots,m.$$

(b) Aus der Rekursionsformel (8.38) berechnen sich die ersten vier Koeffizienten $\gamma^*_0,\ldots,\gamma^*_3$ zu

$$\gamma^*_0 \;=\; 1, \qquad \gamma^*_1 \;=\; -\tfrac{1}{2}, \qquad \gamma^*_2 \;=\; -\tfrac{1}{12}, \qquad \gamma^*_3 \;=\; -\tfrac{1}{24}.$$

Für $m = 1,2,3$ lauten die m-schrittigen Adams-Moulton-Verfahren in der klassischen Darstellung eines linearen Mehrschrittverfahrens folgendermaßen,

$$\begin{aligned}
&m=1: u_{\ell+1} = u_\ell \;+\; \tfrac{h}{2}(f_{\ell+1} \;+\; f_\ell), && \ell \;=\; 0,\ldots,n-1;\\
&m=2: u_{\ell+2} = u_{\ell+1} \;+\; \tfrac{h}{12}(5f_{\ell+2} \;+\; 8f_{\ell+1} \;-\; f_\ell), && \text{——«——}\ n-2;\\
&m=3: u_{\ell+3} = u_{\ell+2} \;+\; \tfrac{h}{24}(9f_{\ell+3} \;+\; 19f_{\ell+2} \;-\; 5f_{\ell+1} \;+\; f_\ell), && \text{——«——}\ n-3.
\end{aligned}$$

Das für $m = 1$ gewonnene Verfahren wird als *Trapezregel* bezeichnet. △

Das folgende Resultat stellt die wesentlichen Eigenschaften der Adams-Moulton-Verfahren heraus:

Theorem 8.29. *Das m-schrittige Adams-Moulton-Verfahren ist nullstabil. Im Fall $f \in C^{m+1}([a,b]\times\mathbb{R}^N,\mathbb{R}^N)$ besitzt es die Konsistenzordnung $p = m+1$, und die Fehlerkonstante lautet γ^*_{m+1}.*

Bemerkung 8.30. Ein m-schrittiges Adams-Moulton-Verfahren besitzt demnach eine höhere Konvergenzordnung als ein m-schrittiges Adams-Bashfort-Verfahren. Der dafür zu zahlende Preis besteht in der numerischen Lösung eines nichtlinearen Gleichungssystems für die Näherung $u_{\ell+m} \in \mathbb{R}^N$. Approximationen hierfür lassen sich mittels gewisser Fixpunktiterationen gewinnen, die in Abschnitt 8.7 vorgestellt werden. △

8.5 Nyström- und Milne-Simpson-Verfahren

8.5.1 Der Ansatz

Zur Herleitung einer zweiten Klasse von Mehrschrittverfahren integriert man die Differenzialgleichung $y' = f(t,y(t))$ von $t_{\ell+m-2}$ bis $t_{\ell+m}$,

$$y(t_{\ell+m}) \;-\; y(t_{\ell+m-2}) \;=\; \int_{t_{\ell+m-2}}^{t_{\ell+m}} f(t,y(t))\,dt, \qquad \ell \;=\; 0,1,\ldots,n-m, \quad (8.39)$$

und spezielle Verfahren gewinnt man nun durch Ersetzen des Integranden durch geeignete Polynome $\mathcal{P}$,

$$u_{\ell+m} - u_{\ell+m-2} = \int_{t_{\ell+m-2}}^{t_{\ell+m}} \mathcal{P}(t)\,dt, \qquad \ell = 0,1,\ldots,n-m. \tag{8.40}$$

Je nach der speziellen Wahl von $\mathcal{P}$ erhält man explizite beziehungsweise implizite Verfahren. Einzelheiten hierzu werden im Verlauf des vorliegenden Abschnitts 8.5 vorgestellt.

8.5.2 Nyström-Verfahren

Definition 8.31. Für $m \geq 2$ erhält man das m-schrittige *Nyström-Verfahren* durch den Ansatz (8.40) mit

$$\mathcal{P} \in \Pi_{m-1}^N, \qquad \mathcal{P}(t_j) = f_j, \qquad j = \ell, \ell+1, \ldots, \ell+m-1,$$
$$f_j := f(t_j, u_j), \qquad \text{———— « ————}.$$

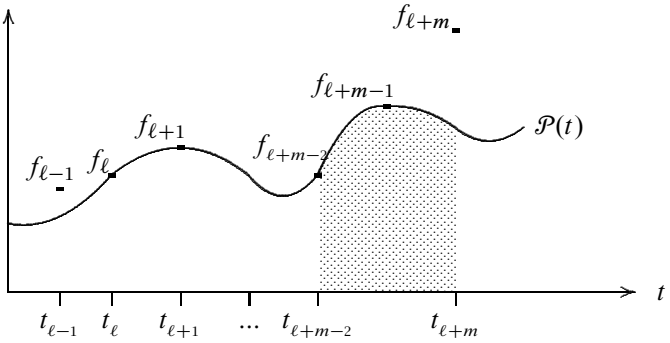


Bild 8.3: Vorgehensweise des m-schrittigen Nyström-Verfahrens im eindimensionalen Fall

Die folgenden Resultate für die Nyström-Verfahren lassen sich genauso wie die Resultate über die Adams-Bashfort-Verfahren herleiten. Auf die jeweiligen Nachweise wird daher wiederum verzichtet.

Theorem 8.32. *Das m-schrittige Nyström-Verfahren hat die Gestalt*

$$u_{\ell+m} - u_{\ell+m-2} = h \sum_{k=0}^{m-1} \varkappa_k \nabla^k f_{\ell+m-1}, \qquad \ell = 0,1,\ldots,n-m,$$

mit den von m unabhängigen Koeffizienten

$$\varkappa_k := (-1)^k \int_{-1}^{1} \binom{-s}{k}\,ds, \qquad k = 0,1,\ldots,$$

die sich rekursiv berechnen durch $\varkappa_0 = 2$ *und*

$$\frac{1}{k+1}\varkappa_0 + \frac{1}{k}\varkappa_1 + \frac{1}{k-1}\varkappa_2 + \ldots + \frac{1}{2}\varkappa_{k-1} + \varkappa_k = 1 \quad \text{für } k = 1, 2, \ldots . \tag{8.41}$$

Bemerkung 8.33. (a) Das m-schrittige Nyström-Verfahren lässt sich in eindeutiger Weise in der Form

$$u_{\ell+m} - u_{\ell+m-2} = h \sum_{j=0}^{m-1} \beta_{m,j} f_{\ell+j}, \qquad \ell = 0, 1, \ldots, n-m,$$

schreiben mit den von der Zahl m abhängigen Koeffizienten

$$\beta_{m,m-1-j} = (-1)^j \sum_{k=j}^{m-1} \binom{k}{j} \varkappa_k, \qquad j = 0, 1, \ldots, m-1.$$

(b) Aus (8.41) berechnen sich die ersten fünf Koeffizienten $\varkappa_0, \ldots, \varkappa_4$ zu

$$\varkappa_0 = 2, \qquad \varkappa_1 = 0, \qquad \varkappa_2 = \frac{1}{3}, \qquad \varkappa_3 = \frac{1}{3}, \qquad \varkappa_4 = \frac{29}{30}.$$

Für $m = 2, 3, 4$ lauten die m-schrittigen Nyström-Verfahren in der klassischen Darstellung eines linearen Mehrschrittverfahrens folgendermaßen,

$$\begin{aligned}
&m = 2: \quad u_{\ell+2} = u_\ell + 2h f_{\ell+1}, && \ell = 0, \ldots, n-2;\\
&m = 3: \quad u_{\ell+3} = u_{\ell+1} + \tfrac{h}{3}(7 f_{\ell+2} - 2 f_{\ell+1} + f_\ell), && \text{____ « ____ } n-2;\\
&m = 4: \quad u_{\ell+4} = u_{\ell+2} + \tfrac{h}{3}(8 f_{\ell+3} - 5 f_{\ell+2} + 4 f_{\ell+1} - f_\ell), && \text{____ « ____ } n-4.
\end{aligned}$$

Für $m = 2$ erhält man also die Mittelpunktregel. △

Das folgende Resultat stellt die wesentlichen Eigenschaften der Nyström-Verfahren heraus:

Theorem 8.34. *Das m-schrittige Nyström-Verfahren ist nullstabil. Für $f \in C^m([a,b] \times \mathbb{R}^N, \mathbb{R}^N)$ besitzt es die Konsistenzordnung $p = m$. Die Fehlerkonstante lautet $\varkappa_m/2$.*

8.5.3 Milne-Simpson-Verfahren

Definition 8.35. Für $m \geq 2$ erhält man das m-schrittige *Milne-Simpson-Verfahren* durch den Ansatz (8.40) mit

$$\begin{aligned}
&\mathcal{P} \in \Pi_m^N, \quad \mathcal{P}(t_j) = f_j, && j = \ell, \ell+1, \ldots, \ell+m,\\
&f_j := f(t_j, u_j), && \text{_____ « _____ }.
\end{aligned}$$

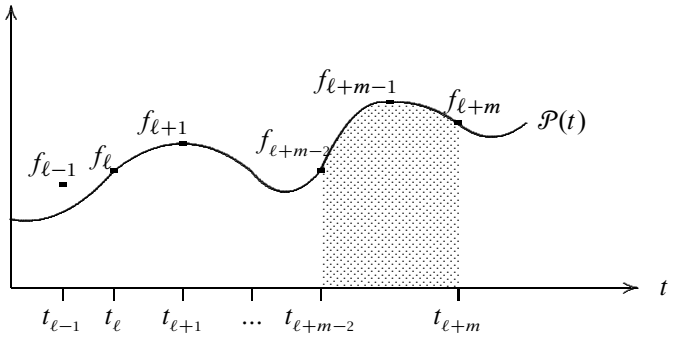


Bild 8.4: Vorgehensweise des m-schrittigen Milne-Simpson-Verfahrens im eindimensionalen Fall

Die folgenden Resultate für die Milne-Simpson-Verfahren ergeben sich genauso wie die Resultate über die Adams-Bashfort-Verfahren. Auf die einzelnen Beweisführungen wird daher auch hier verzichtet.

Theorem 8.36. *Für $m \geq 2$ hat das m-schrittige Milne-Simpson-Verfahren die Gestalt*

$$u_{\ell+m} - u_{\ell+m-2} = h \sum_{k=0}^{m} \varkappa_k^* \nabla^k f_{\ell+m}, \qquad \ell = 0, 1, \ldots, n-m,$$

mit den von der Zahl m unabhängigen Koeffizienten

$$\varkappa_k^* := (-1)^k \int_{-2}^{0} \binom{-s}{k} ds, \qquad k = 0, 1, \ldots,$$

die sich rekursiv berechnen durch $\varkappa_0^* = 2$, $\varkappa_1^* = -2$ und

$$\frac{1}{k+1}\varkappa_0^* + \frac{1}{k}\varkappa_1^* + \frac{1}{k-1}\varkappa_2^* + \ldots + \frac{1}{2}\varkappa_{k-1}^* + \varkappa_k^* = 0 \quad \text{für } k = 2, 3, \ldots . \tag{8.42}$$

Bemerkung 8.37. (a) Das m-schrittige Milne-Simpson-Verfahren (8.42) lässt sich in eindeutiger Weise in der Form

$$u_{\ell+m} - u_{\ell+m-2} = h \sum_{j=0}^{m} \beta_{m,j}^* f_{\ell+j}, \qquad \ell = 0, 1, \ldots, n-m,$$

schreiben mit den von der Zahl m abhängigen Koeffizienten

$$\beta_{m,m-j}^* = (-1)^j \sum_{k=j}^{m} \binom{k}{j} \varkappa_j^*, \qquad j = 0, 1, \ldots, m-1.$$

(b) Aus (8.41) berechnen sich die ersten fünf Koeffizienten $\varkappa_0^*, \ldots, \varkappa_4^*$ zu

$$\varkappa_0^* = 2, \quad \varkappa_1^* = -2, \quad \varkappa_2^* = \frac{1}{3}, \quad \varkappa_3^* = 0, \quad \varkappa_4^* = -\frac{1}{90}.$$

Für $m = 2$ beziehungsweise $m = 4$ lauten die m-schrittigen Milne-Simpson-Verfahren in der klassischen Darstellung eines linearen Mehrschrittverfahrens folgendermaßen,

$$\begin{aligned} m = 2: \quad u_{\ell+2} &= u_\ell + \frac{h}{3}(f_{\ell+2} + 4f_{\ell+1} + f_\ell), \qquad 0 \le \ell \le n-2; \\ m = 4: \quad u_{\ell+4} &= u_{\ell+2} + \frac{h}{90}(29f_{\ell+4} + 124f_{\ell+3} + 24f_{\ell+2} + 4f_{\ell+1} - f_\ell), \\ &\qquad\qquad 0 \le \ell \le n-4. \end{aligned}$$

Für $m = 2$ erhält man das *Verfahren von Milne*, das der Simpson-Regel zur numerischen Integration entspricht. △

Theorem 8.38. *Für $m \ge 2$ ist das m-schrittige Milne-Simpson-Verfahren nullstabil. Wir unterscheiden nun die Fälle $m = 2$ und $m \ge 4$:*[3]

(a) Für eine hinreichend glatte Funktion f besitzt das (zweischrittige) Verfahren von Milne die Konsistenzordnung $p = 4$, und die Fehlerkonstante lautet $-1/180$.

(b) Für $m \ge 4$ und eine hinreichend glatte Funktion f besitzt das m-schrittige Milne–Simpson-Verfahren die Konsistenzordnung $p = m + 1$, und die Fehlerkonstante lautet $\varkappa^*_{m+1}/2$.

Bemerkung 8.39. Ganz allgemein erhält man für jede Zahl $q \ge 3$ weitere Klassen von Mehrschrittverfahren durch Integration der Differenzialgleichung $y' = f(t, y)$ von $t_{\ell+m-q}$ bis $t_{\ell+m}$,

$$y(t_{\ell+m}) - y(t_{\ell+m-q}) = \int_{t_{\ell+m-q}}^{t_{\ell+m}} f(t, y(t))\, dt, \qquad \ell = 0, 1, \ldots, n-m,$$

sowie durch anschließendes Ersetzen des Integranden durch geeignete Polynome $\mathcal{P}$,

$$u_{\ell+m} - u_{\ell+m-q} = \int_{t_{\ell+m-q}}^{t_{\ell+m}} \mathcal{P}(t)\, dt, \qquad \ell = 0, 1, \ldots, n-m. \tag{8.43}$$

Bei allen auf solchen Ansätzen (mit $q \ge 1$) beruhenden Ein- und Mehrschrittverfahren wird für jeden Index ℓ die Vorgehensweise in (8.43) als *Integrationsschritt* bezeichnet. △

8.6 BDF-Verfahren

Im Folgenden werden die (impliziten) rückwärtigen Differenziationsformeln behandelt, die kurz als BDF-Verfahren (backward differentiation formulas) bezeichnet werden.

[3] Für $m = 3$ erhält man das gleiche Verfahren wie für $m = 2$.

8.6.1 Der Ansatz

Definition 8.40. Für $m \geq 1$ ist die Vorgehensweise bei dem m-schrittigen *BDF-Verfahren* für $\ell = 0, \ldots, n-m$ folgendermaßen: ausgehend von den schon berechneten Approximationen $u_j \approx y(t_j)$ für $j = \ell, \ldots, \ell+m-1$, bestimmt man die Näherung $u_{\ell+m} \approx y(t_{\ell+m})$ dahingehend, dass für das Interpolationspolynom

$$\mathcal{P} \in \Pi_m^N, \qquad \mathcal{P}(t_j) \;=\; u_j, \qquad j \;=\; \ell, \ell+1, \ldots, \ell+m, \tag{8.44}$$

Folgendes erfüllt ist,

$$\mathcal{P}'(t_{\ell+m}) \stackrel{!}{=} f_{\ell+m}, \qquad \text{mit} \quad f_{\ell+m} \;:=\; f(t_{\ell+m}, u_{\ell+m}). \tag{8.45}$$

Der Vektor $u_{\ell+m} \in \mathbb{R}^N$ wird also durch die zusätzliche Bedingung (8.45) festgelegt. Die vorliegende Situation ist in Bild 8.5 veranschaulicht.

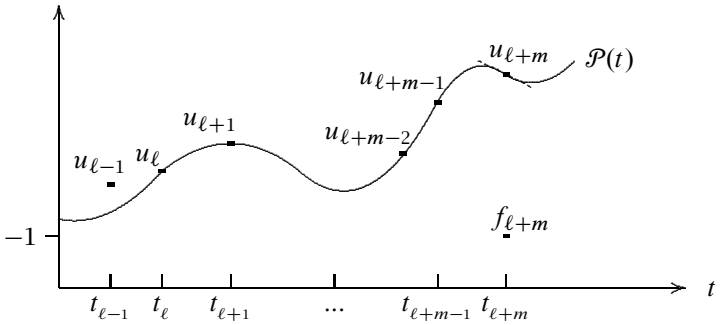


Bild 8.5: Vorgehensweise des m-schrittigen BDF-Verfahrens im eindimensionalen Fall

Theorem 8.41. *Das m-schrittige BDF-Verfahren hat die Gestalt*

$$\sum_{k=1}^{m} \tfrac{1}{k} \nabla^k u_{\ell+m} \;=\; h f_{\ell+m}, \qquad \ell \;=\; 0, 1, \ldots, n-m. \tag{8.46}$$

BEWEIS. Für das Polynom $\mathcal{P}$ aus (8.44) erhält man nach Lemma 8.20 auf Seite 193 die folgende Darstellung,

$$\mathcal{P}(t_{\ell+m} + sh) \;=\; \sum_{k=0}^{m} (-1)^k \tbinom{-s}{k} \nabla^k u_{\ell+m}, \qquad s \in \mathbb{R}, \tag{8.47}$$

mit noch freiem $u_{\ell+m} \in \mathbb{R}^N$. Zur Anpassung an die Bedingung (8.45) wird (8.47) differenziert,

$$\mathcal{P}'(t_{\ell+m}) \;=\; \frac{1}{h} \frac{d}{ds} \mathcal{P}(t_{\ell+m} + sh)\big|_{s=0} \;=\; \frac{1}{h} \sum_{k=0}^{m} (-1)^k \frac{d}{ds} \tbinom{-s}{k}\big|_{s=0} \nabla^k u_{\ell+m},$$

und wegen $\binom{-s}{0} = 1$ sowie[4]

$$\begin{aligned}\frac{d}{ds}\binom{-s}{k}\Big|_{s=0} &= \frac{d}{ds}\frac{(-s)(-s-1)\cdots(-s-k+1)}{k!}\Big|_{s=0} = -\frac{(-1)(-2)\cdots(-k+1)}{k!}\\ &= (-1)^k\frac{1\cdot 2\cdots(k-1)}{k!} = \frac{(-1)^k}{k}\end{aligned}$$

für $k \geq 1$ erhält man die Äquivalenz der Aussagen (8.44)–(8.45) beziehungsweise (8.46). □

Bemerkung 8.42. (a) Das m-schrittige BDF-Verfahren (8.46) lässt sich in eindeutiger Weise in der Form

$$\sum_{j=0}^{m}\alpha_{m,j}u_{\ell+j} = hf_{\ell+m}, \qquad \ell = 0,1,\ldots,n-m,$$

schreiben mit den von der Zahl m abhängigen Koeffizienten $\alpha_{m,0},\ldots,\alpha_{m,m} \in \mathbb{R}$, denn die Darstellung (8.21) liefert

$$\sum_{k=1}^{m}\frac{1}{k}\nabla^k u_{\ell+m} = \sum_{k=1}^{m}\frac{1}{k}\sum_{j=0}^{k}(-1)^j\binom{k}{j}u_{\ell+m-j} = \sum_{j=0}^{m}\Big[\underbrace{(-1)^j\sum_{k=\max\{j,1\}}^{m}\frac{1}{k}\binom{k}{j}}_{=:\,\alpha_{m,m-j}}\Big]u_{\ell+m-j}.$$

(b) Für $m = 1,\ldots,5$ lauten die m-schrittigen BDF-Verfahren in der klassischen Darstellung eines linearen Mehrschrittverfahrens folgendermaßen (jeweils für $\ell \leq n-m$):

$m=1$:	$u_{\ell+1} - u_\ell = hf_{\ell+1}$;
$m=2$:	$\frac{1}{2}(3u_{\ell+2} - 4u_{\ell+1} + u_\ell) = hf_{\ell+2}$;
$m=3$:	$\frac{1}{6}(11u_{\ell+3} - 18u_{\ell+2} + 9u_{\ell+1} - 2u_\ell) = hf_{\ell+3}$;
$m=4$:	$\frac{1}{12}(25u_{\ell+4} - 48u_{\ell+3} + 36u_{\ell+2} - 16u_{\ell+1} + 3u_\ell) = hf_{\ell+4}$;
$m=5$:	$\frac{1}{60}(137u_{\ell+5} - 300u_{\ell+4} + 300u_{\ell+3} - 200u_{\ell+2} + 75u_{\ell+1} - 12u_\ell) = hf_{\ell+5}$.

Insbesondere erhält man im Fall $m = 1$ das *implizite Euler-Verfahren.* △

Man kann Folgendes nachweisen (siehe etwa Abschnitt III.3 in Hairer/Nørsett/Wanner [50]):

Theorem 8.43. *Das m-schrittige BDF-Verfahren ist genau für $1 \leq m \leq 6$ nullstabil. Für hinreichend glatte Funktionen f besitzt es die Konsistenzordnung $p = m$, und die Fehlerkonstante lautet $-1/(m+1)$.*

Mehrschrittverfahren	Ordnung	Fehlerkonstante
m-schrittige Adams-Bashfort-Verfahren, $m \geq 1$	m	γ_m
___ « ___ Adams-Moulton-Verfahren, $m \geq 1$	$m+1$	γ^*_{m+1}
___ « ___ Nyström-Verfahren, $m \geq 2$	m	$\varkappa_m/2$
___ « ___ Milne-Simpson-Verfahren, $m \geq 4$	$m+1$	$\varkappa^*_{m+1}/2$
___ « ___ BDF-Verfahren, $1 \leq m \leq 6$	m	$-1/(m+1)$

Tabelle 8.1: Übersicht der Konsistenzordnungen und Fehlerkonstanten für spezielle nullstabile m-Schrittverfahren

8.6.2 Tabellarische Übersicht über spezielle Mehrschrittverfahren

8.7 Prädiktor-Korrektor-Verfahren

Implizite m-Schrittverfahren von der Form (8.1) mit $\alpha_m = 1$ implementiert man in der Form eines Prädiktor-Korrektor-Schemas. Bei im Folgenden fixiertem ℓ geht man dabei folgendermaßen vor:

- mittels einer Fixpunktiteration (dem *Korrektor*, engl. corrector) bestimmt man $u^{[1]}_{\ell+m}, \ldots, u^{[M-1]}_{\ell+m} \in \mathbb{R}^N$ und schließlich $u_{\ell+m} := u^{[M]}_{\ell+m} \in \mathbb{R}^N$;
- den Startwert $u^{[0]}_{\ell+m} \in \mathbb{R}^N$ verschafft man sich durch ein zunächst nicht näher spezifiziertes explizites m-Schrittverfahren (den sogenannten *Prädiktor*, engl. predictor),

Die folgende Definition präzisiert dieses Vorgehen.

Definition 8.44. Gegeben seien

- ein implizites m-Schrittverfahren von der Form (8.1) mit $\alpha_m = 1$ (der Korrektor);
- ein explizites m-Schrittverfahren (der Prädiktor) mit Koeffizienten $\alpha^*_0, \alpha^*_1, \ldots, \alpha^*_{m-1}$ sowie der Funktion

$$\varphi_* : [a,b] \times (\mathbb{R}^N)^m \times [0,H] \;\to\; \mathbb{R}^N.$$

Bei dem zugehörigen Prädiktor-Korrektor-Verfahren geht man für $\ell = 0, \ldots, n-m$ so vor: für fixiertes ℓ bestimmt man $u^{[0]}_{\ell+m}, \ldots, u^{[M-1]}_{\ell+m}$, $u^{[M]}_{\ell+m} =: u_{\ell+m} \in \mathbb{R}^N$ entspre-

[4] siehe (8.23)

chend den folgenden Bestimmungsgleichungen,

$$u_{\ell+m}^{[0]} + \sum_{j=0}^{m-1} \alpha_j^* u_{\ell+j} = h\varphi_*\big(t_\ell, u_\ell, \ldots, u_{\ell+m-1}; h\big), \tag{8.48-a}$$

$$u_{\ell+m}^{[\nu]} + \sum_{j=0}^{m-1} \alpha_j u_{\ell+j} = h\varphi\big(t_\ell, u_\ell, \ldots, u_{\ell+m-1}, u_{\ell+m}^{[\nu-1]}; h\big), \quad \nu = 1, 2, \ldots, M, \tag{8.48-b}$$

$$u_{\ell+m} = u_{\ell+m}^{[M]}.$$

Hier setzt man $u_0 = y_0$, und die übrigen Startwerte $u_\ell = u_\ell^{(0)} \approx y(t_\ell)$, $\ell = 1, \ldots, m-1$, hat man in einer (an dieser Stelle nicht näher spezifizierten) Anlaufrechnung zu bestimmen.

Das folgende Lemma macht deutlich, dass sich das vorgestellte Prädiktor-Korrektor-Verfahren als nichtlineares explizites m-Schrittverfahren von der Form (8.1) darstellen lässt.

Lemma 8.45. *Gegeben sei ein Prädiktor-Korrektor-Verfahren entsprechend Definition 8.44. Für die gewonnenen Approximationen $u_\ell \approx y(t_\ell) \in \mathbb{R}^N$ eines Prädiktor-Korrektor-Verfahrens gilt die Darstellung*

$$u_{\ell+m} + \sum_{j=0}^{m-1} \alpha_j u_{\ell+j} = h\psi^{[M]}(t_\ell, u_\ell, \ldots, u_{\ell+m-1}; h), \quad \ell = 0, 1, \ldots, n-m, \tag{8.49}$$

wobei die Funktion $\psi^{[M]} : [a, b] \times (\mathbb{R}^N)^m \times [0, H] \to \mathbb{R}^N$ wie folgt rekursiv definiert ist,

$$\psi^{[\nu]}(t, v_0, \ldots, v_{m-1}; h) = \varphi(t, v_0, \ldots, v_{m-1}, v_m^{[\nu-1]}; h), \quad \nu = 1, \ldots, M, \tag{8.50}$$

mit

$$\left.\begin{aligned} v_m^{[0]} &= h\varphi_*(t, v_0, \ldots, v_{m-1}; h) - \sum_{j=0}^{m-1} \alpha_j^* v_j, \\ v_m^{[\nu-1]} &= h\psi^{[\nu-1]}(\underline{\quad} \text{«} \underline{\quad}) - \sum_{j=0}^{m-1} \alpha_j v_j, \qquad \nu = 2, \ldots, M. \end{aligned}\right\} \tag{8.51}$$

BEWEIS. Für den Nachweis der Darstellung (8.49) setzt man in (8.50)–(8.51)

$$v_0 := u_\ell, \qquad v_1 := u_{\ell+1}, \ \ldots, \ v_{m-1} := u_{\ell+m-1},$$

und durch Vergleich von (8.48) und (8.51) erkennt man mittels vollständiger Induktion leicht

$$v_m^{[\nu]} = u_{\ell+m}^{[\nu]}, \qquad \nu = 0, 1, \ldots, M,$$

wobei $v_m^{[M]}$ entsprechend (8.51) definiert sei. Dies bedeutet nichts anderes als

$$u_{\ell+m}^{[\nu]} + \sum_{j=0}^{m-1} \alpha_j u_{\ell+j} = h\psi^{[\nu]}(t_\ell, u_\ell, \ldots, u_{\ell+m-1}; h), \quad \nu = 1, 2, \ldots, M. \tag{8.52}$$

Für $\nu = M$ erhält man aus (8.52) schließlich die Darstellung (8.49). □

Gegenstand des folgenden Theorems sind die Konsistenzordnung und Nullstabilität von Prädiktor-Korrektor-Verfahren.

Theorem 8.46. *Gegeben sei ein Prädiktor-Korrektor-Verfahren von der Form in Definition 8.44, welches die folgenden Eigenschaften besitze:*

- *der Prädiktor besitze die Konsistenzordnung $p_* \geq 1$, und die Funktion φ_* genüge einer Lipschitzbedingung der Form (8.6);*
- *der Korrektor sei nullstabil und besitze die Konsistenzordnung $p \geq p_* + M$, und die Funktion φ genüge einer Lipschitzbedingung der Form (8.6).*

Dann ist das Prädiktor-Korrektor-Verfahren nullstabil und besitzt die Konsistenzordnung $p_* + M$, und die zugehörige Funktion $\psi^{[M]}$ genügt der Lipschitzbedingung (8.6).

BEWEIS. Die zu den Funktionen φ beziehungsweise φ_* gehörenden Lipschitzkonstanten seien mit L beziehungsweise L_* bezeichnet.

(a) Die Nullstabilität folgt unmittelbar aus der Darstellung (8.49).

(b) Wir zeigen im Folgenden für $\nu = 1, 2, \ldots, M$ induktiv, dass die Funktion $\psi^{[\nu]}$ aus (8.50) einer Lipschitzbedingung der Form (8.6) genügt mit einer gewissen Lipschitzkonstanten $L^{[\nu]}$. Tatsächlich erhält man für $w_m^{[\nu]}$ entsprechend $v_m^{[\nu]}$ aus (8.50), (8.51) Folgendes (für $0 < h \leq H$),

$$\begin{aligned}
&\| \psi^{[1]}(t, v_0, \ldots, v_{m-1}; h) - \psi^{[1]}(t, w_0, \ldots, w_{m-1}; h) \| \\
&\quad \leq L\Big[\Big(\sum_{j=0}^{m-1} \| v_j - w_j \| \Big) + \| v_m^{[0]} - w_m^{[0]} \| \Big] \\
&\quad \leq L\Big[\Big(\sum_{j=0}^{m-1} \| v_j - w_j \| \Big) \Big(1 + \max_{j=0,\ldots,m-1} |\alpha_j^*| \Big) \\
&\qquad\qquad + h \| \varphi_*(t, v_0, \ldots, v_{m-1}; h) - \varphi_*(t, w_0, \ldots, w_{m-1}; h) \| \Big] \\
&\quad \leq \underbrace{L\Big[1 + \max_{j=0,\ldots,m-1} |\alpha_j^*| + HL_* \Big]}_{=:\, L^{[1]}} \Big(\sum_{j=0}^{m-1} \| v_j - w_j \| \Big),
\end{aligned}$$

und genauso erhält man für $\nu = 2, 3, \ldots, M$:

$$\begin{aligned}
&\|\psi^{[\nu]}(t, v_0, \ldots, v_{m-1}; h) - \psi^{[\nu]}(t, w_0, \ldots, w_{m-1}; h)\| \\
&\quad \le L\Big[\Big(\sum_{j=0}^{m-1} \|v_j - w_j\|\Big)\Big(1 + \max_{j=0,\ldots,m-1} |\alpha_j|\Big) \\
&\qquad\qquad + h\|\psi^{[\nu-1]}(t, v_0, \ldots, v_{m-1}; h) - \psi^{[\nu-1]}(t, w_0, \ldots, w_{m-1}; h)\|\Big] \\
&\quad \le \underbrace{L\Big[1 + \max_{j=0,\ldots,m-1} |\alpha_j| + HL^{[\nu-1]}\Big]}_{=:\, L^{[\nu]}} \Big(\sum_{j=0}^{m-1} \|v_j - w_j\|\Big).
\end{aligned}$$

(c) Für den Nachweis der angegebenen Konsistenzordnung definiert man

$$\left.\begin{aligned}
\eta^*(t,h) &= y(t+mh) + \sum_{j=0}^{m-1} \alpha_j^* y(t+jh) - h\varphi_*(t, y(t), \ldots, y(t+(m-1)h)), \\
\eta^{[\nu]}(t,h) &= y(t+mh) + \sum_{j=0}^{m-1} \alpha_j y(t+jh) - h\psi^{[\nu]}(\underline{\qquad} \text{ « } \underline{\qquad}), \\
&\qquad\qquad\qquad\qquad\qquad\qquad \nu = 1, 2, \ldots, M,
\end{aligned}\right\} \tag{8.53}$$

womit η_* der lokale Verfahrensfehler des Prädiktors ist, und $\eta^{[M]}(t,h)$ stellt den lokalen Verfahrensfehler des Prädiktor-Korrektor-Verfahrens dar. Im Folgenden wird mittels vollständiger Induktion Folgendes gezeigt,

$$\|\eta^{[\nu]}(t,h)\| = \mathcal{O}(h^{p_*+\nu+1}) \quad \text{für} \quad h \to 0 \qquad (\nu = 1, 2, \ldots, M), \tag{8.54}$$

und für $\nu = M$ erhält man die angegebene Konsistenzordnung für das Prädiktor-Korrektor-Verfahren. Für den Nachweis von (8.54) zieht man für $\nu = 1, 2, \ldots, M$ die Definition (8.50) von $\psi^{[\nu]}$ heran,

$$\begin{aligned}
&\psi^{[\nu]}\big(t, y(t), y(t+h), \ldots, y(t+(m-1)h); h\big) \\
&\qquad = \varphi\big(t, y(t), y(t+h), \ldots, y(t+(m-1)h), v_m^{[\nu]}; h\big),
\end{aligned} \tag{8.55}$$

mit

$$\begin{aligned}
v_m^{[0]} &= h\varphi_*\big(t, y(t), \ldots, y(t+(m-1)h)\big) - \sum_{j=0}^{m-1} \alpha_j^* y(t+jh) \\
&\overset{(8.53)}{=} y(t+mh) - \eta_*(t,h), \\
v_m^{[\nu-1]} &\overset{\nu \ge 1}{=} h\psi^{[\nu-1]}\big(t, y(t), \ldots, y(t+(m-1)h)\big) - \sum_{j=0}^{m-1} \alpha_j y(t+jh) \\
&\overset{(8.53)}{=} y(t+mh) - \eta^{[\nu-1]}(t,h).
\end{aligned}$$

Dies eingesetzt in (8.55) ergibt unter Verwendung der Notation $\eta^{[0]} = \eta_*$

$$\begin{aligned}
&\psi^{[\nu]}(t, y(t), \dots, y(t+(m-1)h); h) \\
&\quad = \varphi(t, y(t), \dots, y(t+(m-1)h), y(t+mh) - \eta^{[\nu-1]}(t,h); h) \\
&\quad = \varphi(\sim\!\sim\!\sim, y(t+mh); h) \\
&\qquad\quad + [\varphi(\sim\!\sim\!\sim, y(t+mh) - \eta^{[\nu-1]}(t,h); h) - \varphi(\sim\!\sim\!\sim, y(t+mh); h)]
\end{aligned}$$

wobei $\sim\!\sim\!\sim$ für "$t, y(t), \dots, y(t+(m-1)h)$" steht. Bezeichnet noch $\eta(t,h)$ den lokalen Verfahrensfehler des Korrektors, so erhält man aus der letzten Darstellung zusammen mit (8.53) die folgenden Abschätzungen,

$$\begin{aligned}
\|\eta^{[1]}(t,h)\| &\le \|\eta(t,h)\| + hL\|\eta^*(t,h)\|, \\
\|\eta^{[\nu]}(t,h)\| &\le \text{__«__} + hL\|\eta^{[\nu-1]}(t,h)\|, \qquad \nu = 2,3,\dots,M,
\end{aligned}$$

beziehungsweise mit vollständiger Induktion

$$\|\eta^{[\nu]}(t,h)\| = \mathcal{O}(h^{p+1}) + h\mathcal{O}(h^{p_*+\nu}) = \mathcal{O}(h^{p_*+\nu+1}) \quad \text{für } h \to 0 \ (\nu = 1,2,\dots,M),$$

was mit der Aussage (8.54) übereinstimmt. □

Bemerkung 8.47. In der typischen Situation $p-1 = p_* = m$ ist nach Theorem 8.46 ein Korrektorschritt ausreichend, man wählt also $M = 1$. △

8.7.1 Linearer Prädiktor/Linearer Korrektor

Typischerweise sind sowohl Prädiktor als auch Korrektor lineare Mehrschrittverfahren, es gilt also

$$\begin{aligned}
\varphi_*(t, v_0, \dots, v_{m-1}; h) &= \sum_{j=0}^{m-1} \beta_j^* f(t+jh, v_j), \\
\varphi(t, v_0, \dots, v_m; h) &= \sum_{j=0}^{m} \beta_j f(t+jh, v_j).
\end{aligned}$$

In dieser speziellen Situation wird das Prädiktor-Korrektor-Verfahren in Form eines Pseudocodes dargestellt.

Algorithmus 8.48. Für ein gegebenes lineares implizites m-Schrittverfahren von der Form (8.1) mit $\alpha_m = 1$ (der Korrektor) sowie ein explizites lineares m-Schrittverfahren mit Koeffizienten $\alpha_j^*,\ \beta_j^*,\ j = 0,\dots,m-1$ (der Prädiktor) nimmt das zugehörige Prädiktor-Korrektor-Verfahren die folgende Gestalt an:

$$
\begin{array}{l}
\texttt{for } \ell = 0,1,\ldots,n-m \\
\mathbf{P}\ u_{\ell+m}^{[0]} + \sum_{j=0}^{m-1} \alpha_j^* u_{\ell+j} = h \sum_{j=0}^{m-1} \beta_j^* f_{\ell+j}; \\
\quad \texttt{for } \nu = 1,\ldots,M: \\
\qquad f_{\ell+m}^{[\nu-1]} = f(t_{\ell+m}, u_{\ell+m}^{[\nu-1]}) \qquad \mathbf{E} \\
\qquad u_{\ell+m}^{[\nu]} + \sum_{j=0}^{m-1} \alpha_j u_{\ell+j} = h\big[\sum_{j=0}^{m-1} \beta_j f_{\ell+j}\big] + h\beta_m f_{\ell+m}^{[\nu-1]} \qquad \mathbf{C} \\
\quad u_{\ell+m} = u_{\ell+m}^{[M]} \\
\mathbf{E}\ f_{\ell+m} = f(t_{\ell+m}, u_{\ell+m}^{[M]})
\end{array}
$$

Wie üblich ist hier $u_0 := y_0$, und die weiteren Startwerte $u_1, \ldots, u_\ell \in \mathbb{R}^N$ sind in einer nicht näher spezifizierten Anlaufrechnung zu berechnen, und schließlich setzt man $f_\ell := f(t_\ell, u_\ell)$ für $\ell = 0, \ldots, m-1$. Das resultierende Verfahren bezeichnet man als $\mathbf{P(EC)^M E}$-*Verfahren*, wobei **E** für "evaluate" steht. △

Bemerkung 8.49. Zur Einsparung einer Funktionsauswertung kann man in Algorithmus 8.48 die Setzung $f_{\ell+m} = f(t_{\ell+m}, u_{\ell+m}^{[M]})$ zu $f_{\ell+m} := f_{\ell+m}^{[M-1]}$ modifizieren. Das resultierende Gesamtverfahren bezeichnet man als $\mathbf{P(EC)^M}$-*Verfahren*, welches hier nicht weiter diskutiert werden soll und auch nicht als Mehrschrittverfahren von der Form (8.1) darstellbar ist. △

8.8 Lineare homogene Differenzengleichungen

8.8.1 Die Testgleichung

In diesem Abschnitt soll das Verhalten spezieller Mehrschrittverfahren zu Illustrationszwecken anhand der Testgleichung

$$y'(t) = \lambda y(t), \qquad t \geq 0 \qquad (\lambda \in \mathbb{R}),$$

untersucht werden. Ein allgemeines lineares m-Schrittverfahren nimmt hier die Form

$$\sum_{j=0}^{m} \gamma_j u_{\ell+j} = 0, \qquad \ell = 0, 1, \ldots, \tag{8.56}$$

an mit $\gamma_j = \alpha_j - h\lambda\beta_j$ für $j = 0, 1, \ldots, m$. Im Folgenden wird beschrieben, wie man die Lösungen $(u_\ell)_{\ell \in \mathbb{N}_0}$ der Differenzengleichung (8.56) erhält.

8.8.2 Existenz und Eindeutigkeit bei linearen homogenen Differenzengleichungen

Definition 8.50. Im Folgenden bezeichne

$$s(\mathbb{K}) = \{u = (u_\ell)_{\ell \in \mathbb{N}_0} \mid u_\ell \in \mathbb{K}\} \tag{8.57}$$

den Raum der Folgen, mit $\mathbb{K} = \mathbb{C}$ oder $\mathbb{K} = \mathbb{R}$. Eine Abbildung $L : s(\mathbb{K}) \to s(\mathbb{K})$ von der Form

$$(Lu)_\ell \;=\; \sum_{j=0}^{m} \gamma_j u_{\ell+j}, \qquad \ell \;=\; 0, 1, \ldots \tag{8.58}$$

mit gegebenen Koeffizienten $\gamma_0, \gamma_1, \ldots, \gamma_m \in \mathbb{R},\ \gamma_m \neq 0$, bezeichnet man als *linearen Differenzenoperator m-ter Ordnung*. Die Gleichung $Lu = 0$ nennt man zugehörige *homogene Differenzengleichung*. Schließlich bezeichnet

$$\mathcal{N}(L) \;=\; \{u \;=\; (u_\ell)_{\ell\in\mathbb{N}_0} \in s(\mathbb{K}) \quad | \quad Lu = 0\} \tag{8.59}$$

den *Nullraum* von L.

Bemerkung 8.51. Mit den natürlichen Verknüpfungen bildet $s(\mathbb{K})$ einen linearen Vektorraum über $\mathbb{K}$, und eine Abbildung $L : s(\mathbb{K}) \to s(\mathbb{K})$ von der Form (8.58) ist linear. △

Theorem 8.52. *Zu gegebenem Differenzenoperator (8.58) und Startwerten* $u_0^{(0)}, \ldots, u_{m-1}^{(0)} \in \mathbb{K}$ *gibt es genau eine Folge* $u \in s(\mathbb{K})$ *mit*

$$Lu \;=\; 0, \qquad u_\ell \;=\; u_\ell^{(0)} \qquad \textit{für} \quad \ell \;=\; 0, 1, \ldots, m-1. \tag{8.60}$$

BEWEIS. Für eine Folge $u \in s(\mathbb{K})$ bedeutet $Lu = 0$ Folgendes,

$$u_{\ell+m} \;=\; -\Big(\sum_{j=0}^{m-1} \gamma_j u_{\ell+j}\Big)/\gamma_m, \qquad \ell \;=\; 0, 1, \ldots, \tag{8.61}$$

woraus unmittelbar Existenz und Eindeutigkeit einer Folge $(u_\ell)_{\ell\in\mathbb{N}_0} \in s(\mathbb{K})$ mit der Eigenschaft (8.60) resultieren. □

Theorem 8.53. *Für jeden linearen Differenzenoperator L der Ordnung m gilt die Identität* $\dim \mathcal{N}(L) = m$.

BEWEIS. Für $\nu = 1, 2, \ldots, m$ sei die Folge $u^{[\nu]} \in s(\mathbb{K})$ folgendermaßen definiert,

$$Lu^{[\nu]} \;=\; 0, \qquad u_\ell^{[\nu]} \;=\; \begin{cases} 1, & \text{für } \ell \;=\; \nu - 1, \\ 0, & \text{für } \ell \;\in\; \{0, \ldots, m-1\}\backslash\{\nu - 1\}. \end{cases}$$

Diese m Folgen bilden eine Basis von $\mathcal{N}(L)$, wie im Folgenden nachgewiesen wird.
(i) Die Folgen $u^{[1]}, \ldots, u^{[m]}$ sind linear unabhängig, denn für gegebene Koeffizienten $c_1, \ldots, c_m \in \mathbb{K}$ gilt:

$$\sum_{\nu=1}^{m} c_\nu u^{[\nu]} \;=\; 0 \;\rightsquigarrow\; 0 \;=\; \Big(\sum_{\nu=1}^{m} c_\nu u^{[\nu]}\Big)_\ell \;=\; \sum_{\nu=1}^{m} c_\nu u_\ell^{[\nu]} \;=\; c_{\ell+1}, \quad \ell \;=\; 0, \ldots, m-1.$$

(ii) Andererseits gilt

$$\mathcal{N}(L) \ \subset \ \text{span}\,\{u^{[1]},\ldots,u^{[m]}\},$$

denn für eine beliebige Folge $u \in \mathcal{N}(L)$ gelten mit $c_\nu := u_{\nu-1}$ für $\nu = 1,\ldots,m$ die Identitäten

$$\Big(\sum_{\nu=1}^{m} c_\nu u^{[\nu]}\Big)_\ell \ = \ \sum_{\nu=1}^{m} c_\nu u_\ell^{[\nu]} \ = \ c_{\ell+1} \ = \ u_\ell, \qquad \ell \ = \ 0,1,\ldots,m-1,$$

beziehungsweise $u = \sum_{\nu=1}^{m} c_\nu u^{[\nu]}$ aufgrund von Theorem 8.52. □

8.8.3 Die komplexwertige allgemeine Lösung der homogenen Differenzengleichung $Lu = 0$

Zur Bestimmung einer Basis des m-dimensionalen Raums der komplexwertigen Lösungsfolgen der Gleichung $Lu = 0$ mit gegebenem Differenzenoperator L der Form (8.58) macht man zunächst den Ansatz $u = (\xi^\ell)_{\ell\in\mathbb{N}_0}$ mit $\xi \in \mathbb{C}$ und erhält

$$(Lu)_\ell \ = \ \sum_{j=0}^{m} \gamma_j \xi^{\ell+j} \ \overset{(*)}{=} \ \xi^\ell \sum_{j=0}^{m} \gamma_j \xi^j, \qquad \ell \ = \ 0,1,\ldots,$$

so dass die Gleichung $Lu = 0$ erfüllt ist, falls $\xi \in \mathbb{C}$ eine Nullstelle des *charakteristischen Polynoms*

$$\psi(\xi) \ = \ \gamma_m \xi^m \ + \ \gamma_{m-1}\xi^{m-1} \ + \ \ldots \ + \ \gamma_0 \tag{8.62}$$

ist. Diese Aussage (und insbesondere die Identität $(*)$) ist auch wahr für $\xi = 0$, wobei der genannte Ansatz hier $u = (1,0,0,\ldots) \in s(\mathbb{C})$ bedeutet.

Im Falle einer s-fachen Nullstelle $\xi \in \mathbb{C}$ mit $s \geq 2$ ist dieser Ansatz jedoch nicht hinreichend allgemein. Es stellt sich Folgendes heraus:

- gilt $\xi \neq 0$, so ist für jedes $0 \leq \nu \leq s-1$ auch $u = (\ell^\nu \xi^\ell)_{\ell\in\mathbb{N}_0}$ Lösung der Gleichung $Lu = 0$.
- Gilt andererseits $\xi = 0$, so ist für jedes $0 \leq \nu \leq s-1$ auch $u = (\overbrace{0,\ldots,0}^{\nu-\text{mal}},1,0,0,\ldots) \in s(\mathbb{C})$ Lösung der Gleichung $Lu = 0$.

Das allgemeine Resultat hierzu ist in dem folgenden Theorem festgehalten.

Theorem 8.54. *Zu gegebenem Differenzenoperator L der Form (8.58) seien $\xi_1,\ldots,\xi_r \in \mathbb{C}$ die paarweise verschiedenen Nullstellen des charakteristischen Polynoms (8.62) mit den jeweiligen Vielfachheiten $m_1,\ldots,m_r \in \mathbb{N}$. Für beliebige Polynome $\mathcal{P}_k \in \Pi_{m_k-1}, k = 1,2,\ldots,r$ (mit komplexen Koeffizienten) sowie gegebenenfalls Zahlen $a_j \in \mathbb{C},\ j = 0,1,\ldots,m_{k_*-1}$, ist je nach der Situation*

(i) $\xi_k \neq 0$ *für* $k = 1,\ldots,r$; *(ii)* $\xi_{k_*} = 0$ *für ein* $1 \leq k_* \leq r$;

durch

$$\left.\begin{aligned} &\text{(i) } u_\ell &&= \sum_{k=1}^{r} \mathcal{P}_k(\ell)\xi_k^\ell, && \ell = 0,1,\dots, \ (\xi_k \neq 0 \ \textit{für alle}\ k)\\ &\text{(ii) } u_\ell &&= \sum_{\substack{k=1\\k\neq k_*}}^{r} \text{——«——} + \sum_{j=0}^{m_{k_*}-1} a_j\delta_{j\ell}, \text{——«——} && (\xi_{k_*} = 0 \ \textit{für ein}\ k_*) \end{aligned}\right\} \tag{8.63}$$

eine Folge $u \in s(\mathbb{C})$ mit $Lu = 0$ definiert. Umgekehrt lässt sich jede Lösung $u \in s(\mathbb{C})$ der Gleichung $Lu = 0$ in der Form (8.63) darstellen.

BEWEIS. Im Folgenden verwenden wir die Notation

$$\omega_\nu(x) := x(x-1)\cdots(x-\nu+1) = \prod_{s=0}^{\nu-1}(x-s), \qquad x \in \mathbb{R},$$

so dass ω_ν ein Polynom vom genauen Grad ν mit den Nullstellen $0,1,\dots,\nu-1$ ist. Weiter sei noch festgehalten, dass für $k = 1,2,\dots,r$ die Eigenschaft $\psi^{(\nu)}(\xi_k) = 0$ für $\nu = 0,1,\dots,m_k-1$ gleichbedeutend mit

$$\sum_{j=\nu}^{m} \gamma_j \omega_\nu(j)\xi_k^{j-\nu} = 0, \qquad \nu = 0,1,\dots,m_k-1, \tag{8.64}$$

ist. Dies gilt mit der Konvention $0^0 = 1$ auch für den Fall $\xi_k = 0$ und bedeutet hier nichts anderes als $\gamma_0 = \gamma_1 = \ldots = \gamma_{m_k-1} = 0$. Im Folgenden soll das spezielle System

$$(u^{[k,\nu]})_{\substack{k=1,\dots,r\\ \nu=0,\dots,m_k-1}} \subset s(\mathbb{C})$$

definiert durch

$$u^{[k,\nu]} = (\omega_\nu(\ell)\xi_k^{\ell-\nu})_{\ell\in\mathbb{N}_0} \qquad \text{für} \quad \begin{matrix} k \in \{1,\dots,r\},\\ \nu \in \{0,\dots,m_k-1\}\end{matrix} \tag{8.65}$$

betrachtet werden, wobei diese spezielle Wahl von $u^{[k,\nu]}$ einen kurzen Beweis der linearen Unabhängigkeit ermöglicht. Die Elemente $u^{[k,\nu]} \in s(\mathbb{C})$ lassen sich folgendermaßen darstellen:

- Für $\xi_k \neq 0$ gilt die Identität

$$u^{[k,\nu]} = \underbrace{\xi_k^{-\nu}}_{\text{const.}} (\omega_\nu(\ell)\xi_k^\ell)_{\ell\geq 0}, \tag{8.66}$$

  und aufgrund der speziellen Form von ω_ν gilt $u_\ell^{[k,\nu]} = 0$ für $\ell = 0,1,\dots,\nu-1$.
- Mit der Konvention $0\times\infty = 0$ bedeutet die Darstellung (8.65) im Falle $\xi_{k_*} = 0$ Folgendes,

$$u^{[k_*,\nu]} = \nu!(\delta_{\ell\nu})_{\ell\geq 0}, \qquad \nu = 0,1,\dots,m_{k_*}-1. \tag{8.67}$$

Die Tatsache $\dim \mathcal{N}(L) = m$ ist aufgrund von Theorem 8.53 bereits bekannt, und des Weiteren gilt $\sum_{k=1}^{r} m_k = m$. Im Folgenden wird nachgewiesen, dass das System (8.65) eine Basis von $\mathcal{N}(L)$ bildet. Mit den Darstellungen (8.66)–(8.67) für dieses System erhält man die Darstellungen (8.63), wenn man noch berücksichtigt, dass sich jedes Polynom $\mathcal{P} \in \Pi_n$ in eindeutiger Weise in der Form $P(x) = \sum_{s=0}^{n} a_s \omega_s(x)$ darstellen lässt.

Für den Nachweis der Basiseigenschaften des Systems (8.65) wird als Erstes für fixierte $k \in \{1, \dots, r\}$ und $\nu \in \{0, \dots, m_k - 1\}$ die Identität $Lu^{[k,\nu]} = 0$ nachgewiesen. Hierzu beobachtet man, dass für festes ℓ die Funktion $\mathbb{C} \to \mathbb{C},\ j \mapsto \omega_\nu(\ell + j)$ ein Polynom ν-ten Grades in j darstellt, so dass es Koeffizienten $a_{\nu,\ell,s} \in \mathbb{C}$ für $s = 0, 1, \dots, \nu$ gibt mit

$$\omega_\nu(\ell + j) \;=\; \sum_{s=0}^{\nu} a_{\nu,\ell,s}\omega_s(j), \qquad j \;=\; 0, 1, \dots .$$

Damit gilt

$$\begin{aligned}(Lu^{[k,\nu]})_\ell \;&=\; \sum_{j=0}^{m} \gamma_j u_{\ell+j}^{[k,\nu]} \;=\; \sum_{j=0}^{m} \gamma_j \omega_\nu(\ell + j)\xi_k^{\ell+j-\nu} \\ &=\; \sum_{j=0}^{m} [\gamma_j \Big(\sum_{s=0}^{\nu} a_{\nu,\ell,s}\omega_s(j)\Big)\xi_k^{\ell+j-\nu}] \;=\; \xi_k^\ell \sum_{s=0}^{\nu} [a_{\nu,\ell,s} \underbrace{\Big(\sum_{j=0}^{m} \gamma_j \omega_s(j)\xi_k^{j-\nu} \Big)}_{\overset{(8.64)}{=}\ 0}] \;=\; 0.\end{aligned}$$

Es ist nun noch die lineare Unabhängigkeit der Familie (8.65) nachzuweisen. Hierzu seien $(c_{k\nu})_{\substack{k=1,\dots,r \\ \nu=0,\dots,m_k-1}} \subset \mathbb{C}$ Koeffizienten mit

$$\sum_{\substack{k=1,\dots,r \\ \nu=0,\dots,m_k-1}} c_{k\nu} u^{[k,\nu]} \;=\; 0.$$

Dies bedeutet

$$0 \;=\; \sum_{\substack{k=1,\dots,r \\ \nu=0,\dots,m_k-1}} c_{k\nu} u_\ell^{[k,\nu]} \;=\; \sum_{\substack{k=1,\dots,r \\ \nu=0,\dots,m_k-1}} c_{k\nu}\omega_\nu(\ell)\xi_k^{\ell-\nu}, \quad \ell \;=\; 0, 1, \dots, m-1,$$

beziehungsweise in Matrixschreibweise

$$Bc \;=\; \sum_{k=1}^{r} B_k c_k \;=\; 0$$

mit Matrizen und Vektoren

$$B \;=\; \left(\begin{array}{c|c|c} & & \\ B_1 & \dots & B_r \\ & & \end{array} \right) \in \mathbb{C}^{m\times m}, \qquad c \;=\; \left(\begin{array}{c} c_1 \\ \hline \vdots \\ \hline c_r \end{array} \right) \in \mathbb{C}^m,$$

wobei $B_k \in \mathbb{C}^{m\times m_k}$ und $c_k \in \mathbb{C}^{m_k}$ wie folgt erklärt sind,

$$B_k = \underbrace{\begin{pmatrix} \omega_0(0) & 0 & \dots & 0 \\ \omega_0(1)\xi_k^1 & \omega_1(1) & \ddots & \vdots \\ \vdots & & \ddots & 0 \\ \omega_0(m_k-1)\xi_k^{m_k-1} & \dots & \dots & \omega_{m_k-1}(m_k-1) \\ \vdots & & & \vdots \\ \omega_0(m-1)\xi_k^{m-1} & \dots & \dots & \omega_{m_k-1}(m-1)\xi_k^{m-m_k} \end{pmatrix}}_{(\omega_\nu(\ell)\xi_k^{\ell-\nu})_{\substack{\ell=0,\dots,m-1\\ \nu=0,\dots,m_k-1}}}, \qquad c_k = \begin{pmatrix} c_{k0} \\ \vdots \\ c_{k,m_k-1} \end{pmatrix}.$$

Die lineare Unabhängigkeit der Familie (8.65) ergibt sich nun aus der Regularität der Matrix $B \in \mathbb{C}^{m\times m}$, die im Folgenden nachgewiesen wird.

Hierzu beobachtet man, dass für ein Polynom

$$p(\xi) = \sum_{j=0}^{m-1} d_j \xi^j,$$

mit den paarweise verschiedenen Nullstellen $\xi_1, \xi_2, \dots, \xi_r \in \mathbb{C}$ und den jeweiligen Vielfachheiten $m_1, \dots, m_r \in \mathbb{N}$ nur[5] $p \equiv 0$ beziehungsweise $d_0 = \dots = d_{m-1} = 0$ gelten kann, denn wegen $\sum_{k=1}^r m_k = m$ besitzt das Polynom $p \in \Pi_{m-1}$ mindestens m Nullstellen (entsprechend ihren Vielfachheiten gezählt). Wegen

$$p^{(\nu)}(\xi_k) = \sum_{\ell=\nu}^{m-1} d_\ell \omega_\nu(\ell)\xi_k^{\ell-\nu}, \qquad \nu = 0, 1, \dots, m_k-1, \qquad k = 1, 2, \dots, r,$$

ist dies gleichbedeutend damit, dass das Gleichungssystem $Ad = 0$ nur die triviale Lösung besitzen kann, wobei

$$A = \left(\begin{array}{c} A_1 \\ \hline \vdots \\ \hline A_r \end{array}\right) \in \mathbb{C}^{m\times m},$$

[5] siehe beispielsweise Fischer [28], Abschnitt 1.3

und die Matrix $A_k \in \mathbb{C}^{m_k \times m}$ ist folgendermaßen erklärt,

$$A_k = \underbrace{\begin{pmatrix} \omega_0(0) & \omega_0(1)\xi_k^1 & \dots & \omega_0(m_k-1)\xi_k^{m_k-1} & \dots & \omega_0(m-1)\xi_k^{m-1} \\ 0 & \omega_1(1) & & \vdots & & \vdots \\ \vdots & \ddots & \ddots & \vdots & & \vdots \\ 0 & \dots & 0 & \omega_{m_k-1}(m_k-1) & \dots & \omega_{m_k-1}(m_k-1)\xi_k^{m-m_k} \end{pmatrix}}_{(\omega_\nu(\ell)\xi_k^{\ell-\nu})_{\substack{\nu=0,\dots,m_k-1 \\ \ell=0,\dots,m-1}}}.$$

Dies zieht die Regularität der Matrix A nach sich. Wegen der Eigenschaft $B = A^\top$ folgt daraus die behauptete Regularität der Matrix B. □

Eine erste Konsequenz aus Theorem 8.54 ist die folgende Aussage:

Korollar 8.55. *Sei L ein Differenzenoperator der Form (8.58). Genau dann hat jede Lösung $u \in s(\mathbb{C})$ der Gleichung $Lu = 0$ die Eigenschaft $\sup_{\ell=0,1,\dots} |u_\ell| < \infty$, wenn für die paarweise verschiedenen Nullstellen $\xi_1, \dots, \xi_r \in \mathbb{C}$ des charakteristischen Polynoms (8.62) Folgendes gilt,*

$$|\xi_k| < 1 \quad \textit{oder} \quad \left\{ \begin{array}{l} |\xi_k| = 1, \\ \xi_k \ \textit{einfache Nullstelle} \end{array} \right\} \qquad (k = 1, 2, \dots, r).$$

8.8.4 Die reellwertige allgemeine Lösung der homogenen Differenzengleichung $Lu = 0$

In erster Linie ist man an den reellen Lösungen der Differenzengleichung $Lu = 0$ interessiert. Hierzu bedient man sich für $\lambda \in \mathbb{C}$ der Polarkoordinatendarstellung $\lambda = \rho e^{\mathrm{i}\varphi} \in \mathbb{C},\ \rho > 0,\ \varphi \in [0, 2\pi)$, und erhält unmittelbar die Darstellung

$$\lambda^\ell = \rho^\ell e^{\mathrm{i}\ell\varphi} = \rho^\ell(\cos(\ell\varphi) + \mathrm{i}\sin(\ell\varphi)), \qquad \ell = 0, 1, \dots .$$

Berücksichtigt man noch, dass aufgrund der reellen Koeffizienten von $\psi(\xi) = \gamma_m\xi^m + \gamma_{m-1}\xi^{m-1} + \dots + \gamma_0$ mit jeder Nullstelle $\xi \in \mathbb{C}$ von ψ auch $\psi(\overline{\xi}) = 0$ gilt, erhält man als zweite Konsequenz aus Theorem 8.54 die allgemeine Form der reellen Lösungsfolgen der Gleichung $Lu = 0$:

Theorem 8.56. *Zu gegebenem Differenzenoperator L von der Form (8.58) seien $\xi_1, \dots, \xi_{r_1} \in \mathbb{R}$ sowie $\lambda_1, \overline{\lambda_1}, \dots, \lambda_{r_2}, \overline{\lambda_{r_2}} \in \mathbb{C}\backslash\mathbb{R}$ die paarweise verschiedenen Nullstellen des charakteristischen Polynoms (8.62), mit den jeweiligen Vielfachheiten $m_1, \dots, m_{r_1}$ und $n_1, \dots, n_{r_2} \in \mathbb{N}$, sowie den Polarkoordinatendarstellungen $\lambda_k = \rho_k e^{i\varphi_k} \in \mathbb{C}$, mit $\rho_k > 0,\ \varphi_k \in (0, 2\pi)$. Für beliebige Polynome*

$$\mathcal{P}_k \in \Pi_{m_k-1} \quad \textit{für} \quad k = 1, \dots, r_1, \qquad \mathcal{Q}_k,\ \widehat{\mathcal{Q}}_k \in \Pi_{n_k-1} \quad \textit{für} \quad k = 1, \dots, r_2,$$

sowie gegebenenfalls Zahlen $a_0, \dots, a_{m_{k_*}-1} \in \mathbb{R}$ *ist je nach der Situation*

(i) $\xi_k \neq 0$ *für* $k = 1, \dots, r_1$; *(ii)* $\xi_{k_*} = 0$ *für ein* $1 \le k_* \le r_1$;

durch

$$\text{(i) } u = \Big(\sum_{k=1}^{r_1} \mathcal{P}_k(\ell)\xi_k^\ell + \sum_{k=1}^{r_2} \rho_k^\ell \big[\mathcal{Q}_k(\ell)\cos(\ell\varphi_k) + \widehat{\mathcal{Q}}_k(\ell)\sin[\ell\varphi_k] \big] \Big)_\ell$$

$$\text{(ii) } u = \Big(\sum_{\substack{k=1 \\ k\neq k_*}}^{r_1} \text{---«---} + \text{----------«----------} + \sum_{j=0}^{m_{k_*}-1} a_j \delta_{j\ell} \Big)_\ell$$

eine Folge $u \in s(\mathbb{R})$ *mit* $Lu = 0$ *definiert. Umgekehrt lässt sich jede Lösung* $u \in s(\mathbb{R})$ *der Gleichung* $Lu = 0$ *in der Form (i) beziehungsweise (ii) darstellen.*

8.8.5 Eine spezielle Differenzengleichung

Zur näherungsweisen Lösung des Anfangswertproblems $y' = f(t,y)$, $y(a) = y_0$ wird im Folgenden zu Testzwecken das Zweischrittverfahren

$$u_{\ell+2} - 4u_{\ell+1} + 3u_\ell = -2hf(t_\ell, u_\ell), \qquad \ell = 0, 1, \dots, n-2, \tag{8.68}$$

untersucht.

Theorem 8.57. *(a) Das Verfahren (8.68) besitzt unter den üblichen Glattheitsvoraussetzungen an die Funktion* f *die Konsistenzordnung* $p = 2$. *Es ist jedoch nicht nullstabil.*

(b) Die Anwendung des Verfahrens (8.68) auf die Testgleichung

$$y'(t) = -y(t), \qquad t \in [0, b], \qquad y(0) = 1, \tag{8.69}$$

mit der Schrittweite $h = b/n > 0$ *sowie den Startwerten* $u_0 = 1$ *und* $u_1 = e^{-h}$ *liefert Folgendes,*

$$u_\ell = \big(\underbrace{e^{-t_\ell}}_{y(t_\ell)} + \tfrac{h^3}{6}\, e^{t_\ell/3}\, 3^\ell \big)(1 + \mathcal{O}(h)) \qquad \text{für } h \to 0, \quad \ell = 0, 1, \dots, n, \tag{8.70}$$

wobei (8.70) gleichmäßig in ℓ *gilt, es hängt also* $\mathcal{O}(h)$ *nicht von* ℓ *ab.*

Wegen der fehlenden Nullstabilität ist also keine Konvergenz des Verfahrens (8.68) zu erwarten, und anhand der Testgleichung lässt sich das genaue Divergenzverhalten beobachten: an jeder festen Stelle $t = \ell h$ verhält sich u_ℓ für $\ell = t/h \to \infty$ wie $t^3 e^{t/3} 3^\ell/(6\ell^3)$. Für die feste Schrittweite $h = 0.01$ sind die durch das Verfahren (8.68) gelieferten Resultate in Tabelle 8.2 vorgestellt.

ℓ	t_ℓ	$y(t_\ell)$	u_ℓ	$e^{-t_\ell} + \frac{h^3}{6} e^{t_\ell/3} 3^\ell$
2	0.02	9.802×10^{-1}	9.802×10^{-1}	9.802×10^{-1}
3	0.03	9.704×10^{-1}	9.704×10^{-1}	9.704×10^{-1}
⋮	⋮	⋮	⋮	⋮
7	0.07	9.324×10^{-1}	9.328×10^{-1}	9.328×10^{-1}
8	0.08	9.231×10^{-1}	9.242×10^{-1}	9.242×10^{-1}
⋮	⋮	⋮	⋮	⋮
13	0.13	8.781×10^{-1}	1.148×10^{0}	1.156×10^{0}
14	0.14	8.694×10^{-1}	1.682×10^{0}	1.705×10^{0}
⋮	⋮	⋮	⋮	⋮
20	0.20	8.187×10^{-1}	6.050×10^{2}	6.22×10^{2}
21	0.21	8.106×10^{-1}	1.819×10^{3}	1.871×10^{3}
⋮	⋮	⋮	⋮	⋮
30	0.30	7.408×10^{-1}	3.688×10^{7}	3.792×10^{7}
31	0.31	7.334×10^{-1}	1.110×10^{8}	1.142×10^{8}
⋮	⋮	⋮	⋮	⋮
100	1.00	3.679×10^{-1}	1.164×10^{41}	1.199×10^{41}

Tabelle 8.2: Illustration des Differenzenverfahrens (8.68), mit der Schrittweite $h = 0.01$ angewandt auf die Testgleichung (8.69) für $b = 1$

BEWEIS VON THEOREM 8.57. (a) Die angegebene Konsistenzordnung ergibt sich unmittelbar aus Lemma 8.16. Das zu dem Verfahren (8.68) gehörende erzeugende Polynom ist $\rho(\xi) = \xi^2 - 4\xi + 3$ mit den Wurzeln $\xi_{1/2} = 2 \pm \sqrt{4-3} = 2 \pm 1$ beziehungsweise $\xi_1 = 3,\ \xi_2 = 1$, so dass also keine Nullstabilität vorliegt.

(b) Anwendung des Verfahrens (8.68) auf die Testgleichung $y' = -y$ führt auf die Differenzengleichung

$$u_{\ell+2} \ - \ 4u_{\ell+1} \ + \ (3-2h)u_\ell \ = \ 0, \qquad \ell \ = \ 0, 1, \ldots, n-2. \tag{8.71}$$

Das zugehörige charakteristische Polynom lautet

$$\psi(\xi) \ = \ \xi^2 \ - \ 4\xi \ + \ 3 \ - \ 2h, \qquad \xi \in \mathbb{C},$$

mit den Nullstellen

$$\xi_{1/2} \ = \ 2 \ \pm \ \sqrt{4 \ - \ (3 \ - \ 2h)} \ = \ 2 \ \pm \ \sqrt{1 \ + \ 2h}.$$

Die allgemeine Lösung von (8.71) ist demnach

$$u_\ell \ = \ c_1 \xi_1^\ell \ + \ c_2 \xi_2^\ell, \qquad \ell \ = \ 0, 1, \ldots \ . \tag{8.72}$$

Anpassung dieser allgemeinen Lösung an die exakten Anfangsbedingungen $u_0 = 1,\ u_1 = e^{-h}$ ergibt $u_0 \ = \ c_1 + c_2 \ = \ 1,\ u_1 \ = \ c_1\xi_1 + c_2\xi_2 \ = \ e^{-h}$ beziehungsweise

$$c_1 \ = \ \frac{\xi_2 - e^{-h}}{\xi_2 - \xi_1}, \qquad c_2 \ = \ \frac{e^{-h} - \xi_1}{\xi_2 - \xi_1}. \tag{8.73}$$

Zur Beschreibung des Verhaltens von u_ℓ aus (8.72) mit Koeffizienten wie in (8.73) verwendet man

$$\sqrt{1+x} \;=\; 1 \,+\, \tfrac{1}{2}x \;-\; \tfrac{1}{8}x^2 \,+\, \tfrac{1}{16}x^3 \,+\, \mathcal{O}(x^4) \qquad \text{für} \quad x \to 0$$

und erhält für die Nullstellen die folgenden Taylorentwicklungen,

$$\xi_1 \;=\; 2 \,+\, (1 \,+\, h \,+\, \mathcal{O}(h^2)) \;=\; 3 \,+\, h \,+\, \mathcal{O}(h^2) \qquad \text{für} \quad h \to 0 \tag{8.74}$$

beziehungsweise

$$\left.\begin{aligned} \xi_2 \;&=\; 2 - (1 \,+\, h \,-\, \tfrac{1}{2}h^2 \,+\, \tfrac{1}{2}h^3 \,+\, \mathcal{O}(h^4)) \\ &=\; 1 \,-\, h \,+\, \tfrac{1}{2}h^2 \,-\, \tfrac{1}{2}h^3 \,+\, \mathcal{O}(h^4) \\ &=\; e^{-h} \,-\, \tfrac{1}{3}h^3 \,+\, \mathcal{O}(h^4) \qquad \text{für} \quad h \to 0. \end{aligned}\right\} \tag{8.75}$$

Für die Koeffizienten c_1, c_2 aus (8.73) erhält man mit den Darstellungen (8.74)–(8.75) und wegen $\xi_2 - \xi_1 = -2 + \mathcal{O}(h)$ Folgendes,

$$\begin{aligned} c_1 \;&=\; \frac{-\frac{1}{3}h^3 \,+\, \mathcal{O}(h^4)}{-2 + \mathcal{O}(h)} \;=\; \tfrac{1}{6}h^3 \,+\, \mathcal{O}(h^4) \qquad \text{für} \quad h \to 0, \\ c_2 \;&=\; \frac{\xi_2 - \xi_1 \,+\, \mathcal{O}(h^3)}{\xi_2 - \xi_1} \;=\; 1 \,+\, \mathcal{O}(h^3) \qquad \text{für} \quad h \to 0. \end{aligned}$$

Die Lösungsfolge $u \in s(\mathbb{R})$ der Differenzengleichung (8.71) mit $u_0 = 1,\ u_1 = e^{-h}$ nimmt somit folgende Gestalt an,

$$u_\ell \;=\; \tfrac{1}{6}h^3(1 + \mathcal{O}(h))(3 + h + \mathcal{O}(h^2))^\ell + (1 \,+\, \mathcal{O}(h^3))(e^{-h} + \mathcal{O}(h^3))^\ell \tag{8.76}$$

für $h \to 0$. Zur Behandlung des zweiten Summanden der rechten Seite in (8.76) berechnet man noch

$$[e^{-h} \,+\, \mathcal{O}(h^3)]^\ell \;=\; e^{-t_\ell}[1 \,+\, \mathcal{O}(h^3)e^h]^\ell \;\overset{(*)}{=}\; e^{-t_\ell}[1 \,+\, \mathcal{O}(h^2)] \qquad \text{für} \quad h \to 0,$$

wobei sich $(*)$ unter Berücksichtigung von $\log(1+x) = \mathcal{O}(x)$ und $e^x = 1 + \mathcal{O}(x)$ für $x \to 0$ aus

$$\ell \log\bigl(1 \,+\, \mathcal{O}(h^3)e^h\bigr) \;=\; \ell\mathcal{O}(h^3)e^h \;=\; \mathcal{O}(h^2) \qquad \text{für} \quad h \to 0 \tag{8.77}$$

ergibt. Den ersten Summanden der rechten Seite in (8.76) behandelt man ganz ähnlich,

$$\begin{aligned} [3 \,+\, h \,+\, \mathcal{O}(h^2)]^\ell \;&=\; 3^\ell[1 \,+\, \tfrac{1}{3}h \,+\, \mathcal{O}(h^2)]\ell \;=\; 3^\ell[e^{h/3} \,+\, \mathcal{O}(h^2)]^\ell \\ &=\; 3^\ell e^{t_\ell/3}[1 \,+\, \mathcal{O}(h^2)e^{-h/3}]^\ell \;\overset{(**)}{=}\; 3^\ell e^{t_\ell/3}[1 \,+\, \mathcal{O}(h)] \quad \text{für} \quad h \to 0, \end{aligned}$$

wobei man $(**)$ genauso wie (8.77) erhält. Daraus resultiert die Darstellung (8.70),

$$\begin{aligned} u_\ell \;&=\; e^{-t_\ell}(1 \,+\, \mathcal{O}(h^2)) \,+\, \tfrac{1}{6}h^3\, e^{t_\ell/3}\, 3^\ell(1 \,+\, \mathcal{O}(h)) \\ &=\; \Bigl(\underbrace{e^{-t_\ell}}_{=\,y(t_\ell)} \,+\, \tfrac{1}{6}h^3\, e^{t_\ell/3}\, 3^\ell\Bigr)(1 + \mathcal{O}(h)) \qquad \text{für} \quad h \to 0, \qquad \ell \;=\; 0, 1, \ldots, n. \end{aligned}$$

Dies komplettiert den Beweis. □

8.9 Steife Differenzialgleichungen

8.9.1 Einführende Bemerkungen

In vielen Anwendungen wie etwa der chemischen Reaktionskinetik treten Anfangswertprobleme für spezielle Differenzialgleichungen $y' = f(t,y),\ t \in [a,b]$ auf, bei denen ein Gleichgewichtszustand $\psi : [a,b] \to \mathbb{R}^N$ existiert, dem sich jede Lösung $y : [a,b] \to \mathbb{R}^N$ der Differenzialgleichung *unabhängig von der Lage des Anfangswerts* schnell annähert, das heißt, außerhalb eines kleinen Intervalls $[a, a+\varepsilon]$ gilt $y \approx \psi$. Solche Differenzialgleichungen werden als "steif" bezeichnet und erfordern eine besondere numerische Behandlung, wie sich herausstellen wird. Im Folgenden wird zunächst der Begriff "steife Differenzialgleichung" etwas präzisiert.

Definition 8.58. Ein Anfangswertproblem $y' = f(t,y),\ y(a) = y_0$ genügt einer *oberen Lipschitzbedingung* bezüglich eines gegebenen Skalarprodukts $\langle\cdot,\cdot\rangle : \mathbb{R}^N \times \mathbb{R}^N \to \mathbb{R}$, wenn es eine stetige Funktion $M : [a,b] \to \mathbb{R}$ gibt mit

$$\langle f(t,u) - f(t,v), u - v\rangle \ \le\ M(t)\|u - v\|^2, \qquad u,\ v \in \mathbb{R}^N. \tag{8.78}$$

Gilt $M(t) \le 0$ für jede Zahl $t \in [a,b]$, so bezeichnet man das gegebene Anfangswertproblem als *dissipativ*.

Hier und im Folgenden bezeichnet $\|\cdot\| : \mathbb{R}^N \to \mathbb{R}$ die durch das Skalarprodukt induzierte Norm. Im weiteren Verlauf sollen Anfangswertprobleme $y' = f(t,y)$, $y(a) = y_0$ betrachtet werden, die

(a) zum einen dissipativ sind oder zumindest einer oberen Lipschitzbedingung genügen mit $M(t)$ von moderater positiver Größe, beispielsweise $M(t) \le 1$;

(b) zum anderen die folgende Eigenschaft besitzen,

$$m(t) \ :=\ \inf_{\substack{u,v \in \mathbb{R}^N \\ u \ne v}} \frac{\langle f(t,u) - f(t,v), u - v\rangle}{\|u - v\|^2} \ \ll\ 0 \qquad \text{für}\quad t \in [a,b]. \tag{8.79}$$

Eine Anfangswertproblem $y' = f(t,y),\ y(a) = y_0$ mit den in (a) und (b) beschriebenen Eigenschaften bezeichnet man als *steif*.

Bemerkung 8.59. Bei steifen Differenzialgleichungen kann aufgrund der Abschätzung

$$\frac{|\langle f(t,u) - f(t,v), u - v\rangle|}{\|u - v\|^2} \ \le\ \frac{\|f(t,u) - f(t,v)\|}{\|u - v\|}$$

die Funktion $f : [a,b] \times \mathbb{R}^N \to \mathbb{R}^N$ die Lipschitzbedingung (7.4) höchstens noch mit einer groß ausfallenden Lipschitzkonstanten $L \ge |m(t)|$ erfüllen, so dass die Konvergenzsätze 7.10 und 8.9 für Einschritt- beziehungsweise Mehrschrittverfahren wegen der auftretenden großen Konstanten erst für kleine Schrittweiten $h > 0$ sinnvolle Resultate liefern. △

In dem folgenden Beispiel wird anhand einer einfachen steifen Differenzialgleichung das Verhalten sowohl des expliziten als auch des impliziten Euler-Verfahrens getestet. Wie sich zeigt, liefert das explizite Euler-Verfahren erst für sehr kleine Integrationsschritte vernünftige Ergebnisse, was aufgrund der vorigen Bemerkung 8.59 auch nicht sonderlich überraschend ist.

Beispiel 8.60. Das Anfangswertproblem

$$y' = \lambda y - (1+\lambda)e^{-t}, \qquad t \in [0,1], \qquad y(0) = y_0, \tag{8.80}$$

besitzt die Lösung

$$y(t) = e^{-t} + (y_0 - 1)e^{\lambda t}, \qquad t \in [0,1].$$

Für $\lambda \in \mathbb{R}$, $\lambda \ll 0$ gilt demnach $y(t) \approx e^{-t}$ bereits für kleine Werte $0 < t \ll 1$. Tatsächlich ist das Anfangswertproblem (8.80) für $\lambda \in \mathbb{R}$ mit $\lambda \ll 0$ steif, mit $M(t) \equiv m(t) \equiv -|\lambda|$.

Im Folgenden werden für die beiden Werte $\lambda = -10$ (das Anfangswertproblem (8.80) ist in dieser Situation nicht steif) und $\lambda = -1000$ (dann ist das Anfangswertproblem (8.80) steif) jeweils sowohl für das explizite als auch das implizite Eulerverfahren numerische Ergebnisse präsentiert. In allen vier Fällen werden gleichabständige Gitter unterschiedlicher Feinheit verwendet, und zwar solche mit den Knotenabständen

$$h = 2^{-k} \quad \text{für} \quad k = 2j, \qquad j = 2,3,\ldots,6.$$

Die Resultate sind in Tabelle 8.3 wiedergegeben. Der Anfangswert ist jeweils $y_0 = 1$, und die Lösung des Anfangswertproblems (8.80) ist dann unabhängig von λ und lautet $y(t) = e^{-t}$ für $t \in [0,1]$. Man beachte, dass im Falle des expliziten Eulerverfahrens der Fehler an der Stelle $t = 1$ für kleiner gewählte Schrittweiten zunächst über alle Schranken hinauswächst. Für die Schrittweiten $h = 2^{-10}$ und $h = 2^{-12}$ werden vernünftige Ergebnisse erzielt. △

Wie sich in Beispiel 8.60 gezeigt hat, liefert das implizite Euler-Verfahren hier trotz der in Bemerkung 8.59 angestellten Beobachtungen für alle kleinen Schrittweiten $h > 0$ vernünftige Ergebnisse. Dieses Verhalten ist kein Zufall, wie sich in Abschnitt 8.9.3 herausstellen wird.

8.9.2 Existenz und Eindeutigkeit der Lösung bei Anfangswertproblemen für Differenzialgleichungen mit oberer Lipschitzeigenschaft

Für Anfangswertprobleme bei gewöhnlichen Differenzialgleichungen mit oberer Lipschitzeigenschaft sollen zunächst die Fragen "Existenz und Eindeutigkeit einer Lösung" sowie die "stetige Abhängigkeit von den Anfangswerten" diskutiert werden. Zwar kann unter diesen Voraussetzungen nicht auf Theorem 7.2 von Picard/Lindelöf auf Seite 155 zurückgegriffen werden, eine stetige Abhängigkeit von den Anfangswerten (und damit insbesondere die Eindeutigkeit der Lösung) liegt dennoch vor:

$\lambda = -10$			$\lambda = -1000$		
h	$u_h(1) - y(1)$ expl. Eulerverf.	$u_h(1) - y(1)$ impl. Eulerverf.	h	$u_h(1) - y(1)$ expl. Eulerverf.	$u_h(1) - y(1)$ impl. Eulerverf.
0.0625	-1.247×10^{-3}	1.308×10^{-3}	0.0625	1.283×10^{24}	1.175×10^{-5}
0.0156	-3.174×10^{-4}	3.212×10^{-4}	0.0156	2.865×10^{69}	2.892×10^{-6}
0.039	-7.971×10^{-5}	7.994×10^{-5}	0.039	8.014×10^{112}	7.202×10^{-7}
0.010	-1.995×10^{-5}	1.996×10^{-5}	0.010	-1.797×10^{-7}	1.799×10^{-7}
0.002	-4.989×10^{-6}	4.990×10^{-6}	0.002	-4.495×10^{-8}	4.496×10^{-8}

Tabelle 8.3: Numerische Ergebnisse für das explizite/implizite Eulerverfahren. Dabei bezeichnet $u_h(1)$ jeweils die gewonnenen Approximationen für $y(1)$.

Theorem 8.61. *Die Funktion $f : [a,b] \times \mathbb{R}^N \to \mathbb{R}^N$ genüge der oberen Lipschitzbedingung (8.78) bezüglich eines gegebenen Skalarprodukts $\langle \cdot, \cdot \rangle$ und einer gegebenen Funktion M. Dann gilt für differenzierbare Funktionen $y,\ \widehat{y} : [a,b] \to \mathbb{R}^N$ mit*

$$\begin{aligned} y' &= f(t,y), \qquad t \in [a,b], & y(a) &= y_0, \\ \widehat{y}' &= f(t,\widehat{y}), \qquad \text{—— « ——} & \widehat{y}(a) &= \widehat{y}_0, \end{aligned}$$

die Abschätzung

$$\| y(t) - \widehat{y}(t) \| \ \le\ \exp\Big(\int_a^t M(s)\, ds \Big)\, \| y_0 - \widehat{y}_0 \|, \qquad t \in [a,b]. \tag{8.81}$$

BEWEIS. Die Funktion

$$\Phi(t) \ = \ \| (y - \widehat{y})(t) \|^2, \qquad t \in [a,b],$$

ist differenzierbar auf dem Intervall $[a,b]$, und es gilt

$$\left.\begin{aligned} \Phi'(t) &\overset{(*)}{=} 2\langle (y-\widehat{y})'(t), (y-\widehat{y})(t) \rangle \ = \ 2\langle f(t,y(t)) - f(t,\widehat{y}(t)), (y-\widehat{y})(t) \rangle \\ &\le \ 2M(t)\, \| (y-\widehat{y})(t) \|^2 \ = \ 2M(t)\Phi(t), \qquad t \in [a,b], \end{aligned}\right\} \tag{8.82}$$

wobei die letzte Abschätzung aus der oberen Lipschitzbedingung (8.78) resultiert. Die Identität $(*)$ folgt unmittelbar aus dem nachfolgenden Lemma 8.62. Die Abschätzung (8.82) zusammen mit der weiter unten nachzutragenden Variante des Gronwall-Lemmas liefert die Behauptung (8.81). □

Es sind noch zwei Hilfsresultate nachzutragen.

Lemma 8.62. *Es seien $\langle\cdot,\cdot\rangle : \mathbb{R}^N \times \mathbb{R}^N \to \mathbb{R}$ ein Skalarprodukt mit induzierter Norm $\|\cdot\| : \mathbb{R}^N \to \mathbb{R}$ und $u : [a,b] \to \mathbb{R}^N$ eine differenzierbare Funktion. Dann ist die Funktion*

$$\Phi(t) \;=\; \|u(t)\|^2, \qquad t \in [a,b],$$

differenzierbar auf dem Intervall $[a,b]$, mit

$$\Phi'(t) \;=\; 2\langle u'(t), u(t)\rangle, \qquad t \in [a,b].$$

BEWEIS. Die Aussage ergibt sich zum Beispiel folgendermaßen,

$$\begin{aligned}\frac{\Phi(t+h)-\Phi(t)}{h} &= \frac{\|u(t+h)\|^2 - \|u(t)\|^2}{h}\\ &= \frac{\langle u(t+h), u(t+h)-u(t)\rangle}{h} + \frac{\langle u(t+h)-u(t), u(t)\rangle}{h}\\ &\to 2\langle u'(t), u(t)\rangle \qquad \text{für} \quad h \to 0.\end{aligned}$$

□

Das folgende Resultat stellt eine Variante des Gronwall-Lemmas dar:

Lemma 8.63. *Für die differenzierbare Funktion $\Phi : [a,b] \to \mathbb{R}$ sei*

$$\Phi'(t) \;\le\; c(t)\,\Phi(t), \qquad t \in [a,b],$$

erfüllt mit der stetigen Funktion $c : [a,b] \to \mathbb{R}$. Dann gilt

$$\Phi(t) \;\le\; \exp\Big(\int_a^t c(s)\,ds\Big)\Phi(a), \qquad t \in [a,b]. \tag{8.83}$$

BEWEIS. Mit der Notation

$$\beta(t) \;:=\; \exp\Big(-\int_a^t c(s)\,ds\Big), \qquad t \in [a,b],$$

erhält man auf dem Intervall $[a,b]$ Folgendes,

$$(\Phi\beta)' \;=\; \Phi'\beta + \Phi\beta' \;=\; \Phi'\beta - c\Phi\beta \;=\; \beta(\Phi' - c\Phi) \;\le\; 0,$$

so dass die Funktion $\Phi\beta$ auf dem Intervall $[a,b]$ monoton fallend ist und damit insbesondere $\Phi(t)\beta(t) \le \Phi(a)$ gilt für $t \in [a,b]$, was gerade die Aussage (8.83) darstellt. □

In gewissen Situationen gewährleistet auch die obere Lipschitzeigenschaft (8.78) die Existenz der Lösungen der zugehörigen Anfangswertprobleme, so zum Beispiel bei Anfangswertproblemen für *autonome* Differenzialgleichungen

$$y' \;=\; f(y), \qquad t \in [a,b], \qquad y(a) \;=\; y_0, \tag{8.84}$$

was in dem folgenden Theorem ohne Beweis festgehalten wird (siehe Strehmel/Weiner [101]).

Theorem 8.64. *Genügt die (von t unabhängige) Funktion $f : \mathbb{R}^N \to \mathbb{R}^N$ einer oberen Lipschitzbedingung (8.78), gilt also*

$$\langle f(u) - f(v), u - v \rangle \;\le\; M \| u - v \|^2, \qquad u,\ v \;\in\; \mathbb{R}^N, \tag{8.85}$$

mit einer Konstanten $M \in \mathbb{R}$, so besitzt das Anfangswertproblem (8.84) genau eine Lösung.

Beispiel 8.65. Das autonome Anfangswertproblem

$$y' \;=\; -y^3, \qquad t \in [a,b], \qquad\qquad y(a) \;=\; y_0 \;\in\; \mathbb{R},$$

ist dissipativ (bezüglich des Skalarprodukts $\langle u, v \rangle = uv$ für $u,\ v \in \mathbb{R}$) und besitzt nach Theorem 8.64 eine eindeutige Lösung. Man beachte, dass Theorem 7.2 hier nicht anwendbar ist, denn die Funktion $f(y) = -y^3$ für $y \in \mathbb{R}$ genügt keiner globalen Lipschitzbedingung von der Form (7.4). △

Zum Abschluss dieses einführenden Abschnitts werden untere und obere Lipschitzschranken für stetig partiell differenzierbare Funktionen angegeben.

Lemma 8.66. *Die Funktion $f : [a,b] \times \mathbb{R}^N \to \mathbb{R}^N$ sei stetig partiell differenzierbar.*
(a) Mit der Notation aus (8.79) gilt

$$m(t) \;=\; \inf_{0 \neq w \in \mathbb{R}^N} \frac{\langle \mathcal{D}_y f(t,u) w, w \rangle}{\| w \|^2} \qquad \textit{für} \quad t \in [a,b], \qquad u,\ w \in \mathbb{R}^N. \tag{8.86}$$

(b) Die Funktion f genügt bezüglich einer gegebenen Funktion $M : [a,b] \to \mathbb{R}$ der oberen Lipschitzbedingung (8.78) genau dann, wenn Folgendes gilt,

$$\langle \mathcal{D}_y f(t,u) w, w \rangle \;\le\; M(t) \| w \|^2 \qquad \textit{für} \quad t \in [a,b], \qquad u,\ w \in \mathbb{R}^N.$$

BEWEIS. Der Mittelwertsatz für vektorwertige Funktionen bedeutet

$$f(t,u) \;-\; f(t,v) \;=\; \Big(\int_0^1 \mathcal{D}_y f(t, v + s(u-v))\, ds \Big)\, (u - v)$$

beziehungsweise

$$\langle f(t,u) - f(t,v), u - v \rangle \;=\; \Big\langle \Big[\int_0^1 \mathcal{D}_y f(t, v + s(u-v))\, ds \Big]\, (u-v), u - v \Big\rangle . \tag{8.87}$$

Auf der anderen Seite gilt

$$\mathcal{D}_y f(t,u) w \;=\; \lim_{h \to 0} \frac{1}{h} \big(f(t, u + hw) \;-\; f(t,u) \big), \qquad u,\ w \in \mathbb{R}^N, \quad t \in [a,b]. \tag{8.88}$$

Aus den Darstellungen (8.87) und (8.88) erhält man unmittelbar die Aussagen (a) und (b) des Lemmas. □

8.9.3 Das implizite Euler-Verfahren für steife Differenzialgleichungen

In diesem Abschnitt wird für das in Beispiel 8.60 auftretende günstige Verhalten des impliziten Euler-Verfahrens bei der Lösung steifer Anfangswertprobleme eine mathematische Erklärung geliefert. Das folgende Lemma dient dabei als Vorbereitung.

Lemma 8.67. *Die Funktion $f : [a,b] \times \mathbb{R}^N \to \mathbb{R}^N$ genüge der oberen Lipschitzbedingung (8.78) mit $M(t) \equiv M$. Je nach der Situation (i) $M \le 0$ beziehungsweise (ii) $M > 0$ gilt dann für beliebige $u,\ v \in \mathbb{R}^N$ sowie $t \in [a,b]$ die folgende Abschätzung,*

$$\left.\begin{array}{ll} \text{(i) } M \le 0: & \|u - v\| \le \|u - v - h(f(t,u) - f(t,v))\| \quad \forall\ h > 0, \\ \text{(ii) } M > 0: & \text{—«—} \le (1 + \varkappa h) \text{—————«—————} \quad \forall\ 0 < h \le H, \end{array}\right\} \tag{8.89}$$

mit der Zahl $0 < H < 1/M$ und der Konstanten $\varkappa := M/(1 - HM)$ in der Situation (ii).

BEWEIS. Nach Voraussetzung gilt

$$h\langle f(t,u) - f(t,v), u - v\rangle \le hM\|u - v\|^2$$

beziehungsweise

$$\begin{aligned} (1 - hM)\|u - v\|^2 &\le \langle u - v, u - v\rangle - h\langle f(t,u) - f(t,v), u - v\rangle \\ &= \langle u - v - h\big(f(t,u) - f(t,v)\big), u - v\rangle \\ &\le \|u - v - h\big(f(t,u) - f(t,v)\big)\|\,\|u - v\|. \end{aligned}$$

Die Behauptung im Fall $M \le 0$ folgt daraus unmittelbar, und im Fall $M > 0$ ergibt sie sich nach der weiteren Rechnung

$$\frac{1}{1 - hM} = 1 + \frac{M}{1 - hM}h \le 1 + \overbrace{\frac{M}{1 - HM}}^{=\varkappa} h. \qquad \square$$

Für gleichabständige Knoten $t_\ell = a + \ell h,\ \ell = 0, 1, \ldots, n$, mit $h = (b - a)/n$ ist das implizite Euler-Verfahren zur Lösung von $y' = f(t,y),\ y(a) = y_0$ von der Form (vergleiche Bemerkung 8.42)

$$u_{\ell+1} = u_\ell + hf(t_{\ell+1}, u_{\ell+1}), \qquad \ell = 0, 1, \ldots, n-1, \qquad u_0 := y_0, \tag{8.90}$$

und besitzt für eine hinreichend glatte Funktion f die Konsistenzordnung $p = 1$, das heißt, für den lokalen Verfahrensfehler (vergleiche (8.5) auf Seite 183)

$$\eta(t,h) = y(t + h) - y(t) - hf(t + h, y(t + h)), \qquad 0 < h \le b - t,$$

gilt die Abschätzung

$$\|\eta(t,h)\| \le Ch^2, \qquad 0 \le h \le b - t,$$

mit einer von h und t unabhängigen Konstanten $C \geq 0$.

Das folgende Theorem liefert die wesentliche Konvergenzaussage für das implizite Euler-Verfahren zur Lösung steifer Differenzialgleichungen. Man beachte, dass die Konstante K hier im Falle $M \leq 0$ moderat ausfällt.

Theorem 8.68. *Erfüllt die Funktion $f : [a,b] \times \mathbb{R}^N \to \mathbb{R}^N$ die obere Lipschitzbedingung (8.78) mit $M(t) \equiv M$, so gilt für den globalen Verfahrensfehler des impliziten Euler-Verfahrens (8.90) die folgende Abschätzung,*

$$\max_{\ell=0,\dots,n} \|u_\ell - y(t_\ell)\| \;\leq\; Kh, \qquad \text{mit } K := \begin{cases} C(b-a), & \text{falls } M \leq 0 \\ \frac{C}{M}\big(e^{M(b-a)/(1-HM)} - 1\big), & \text{dto. } M > 0 \end{cases} \tag{8.91}$$

mit der Einschränkung $0 < h \leq H < 1/M$ im Fall $M > 0$.

BEWEIS. Mit den Setzungen

$$\begin{aligned} e_\ell &= u_\ell - y_\ell, \qquad y_\ell := y(t_\ell), & \ell &= 0,1,\dots,n, \\ \eta_\ell &= \eta(t_\ell, h), & \ell &= 0,1,\dots,n-1, \end{aligned}$$

gilt für $\ell = 0,1,\dots,n-1$

$$\begin{aligned} y_{\ell+1} &= y_\ell + hf(t_{\ell+1}, y_{\ell+1}) + \eta_\ell, \\ u_{\ell+1} &= u_\ell + hf(t_{\ell+1}, u_{\ell+1}), \end{aligned}$$

und daher

$$e_{\ell+1} - h[f(t_{\ell+1}, u_{\ell+1}) - f(t_{\ell+1}, y_{\ell+1})] = e_\ell - \eta_\ell. \tag{8.92}$$

Im Fall $M \leq 0$ erhält man aus (8.89) und (8.92)

$$\begin{aligned} \|e_{\ell+1}\| &\leq \|e_{\ell+1} - h\big(f(t_{\ell+1}, u_{\ell+1}) - f(t_{\ell+1}, y_{\ell+1})\big)\| = \|e_\ell - \eta_\ell\| \\ &\leq \|e_\ell\| + \|\eta_\ell\| \leq \|e_\ell\| + Ch^2. \end{aligned}$$

Wegen $e_0 = 0$ erhält man mittels vollständiger Induktion die angegebene Abschätzung (8.91) für den Fall $M \leq 0$. Für $M > 0$ geht man vergleichbar vor: wiederum aus (8.89) und (8.92) erhält man mit $\varkappa := M/(1-MH)$ die folgenden Abschätzungen,

$$\begin{aligned} \|e_{\ell+1}\| &\leq (1+\varkappa h)\|e_{\ell+1} - h\big(f(t_{\ell+1}, u_{\ell+1}) - f(t_{\ell+1}, y_{\ell+1})\big)\| \\ &\leq (1+\varkappa h)\big(\|e_\ell\| + \|\eta_\ell\|\big) \leq (1+\varkappa h)\|e_\ell\| + \tfrac{1}{1-MH}\|\eta_\ell\|, \end{aligned}$$

und mit Lemma 7.12 erhält man die Abschätzung (8.91) auch für den Fall $M > 0$. Dies komplettiert den Beweis des Theorems. □

8.9.4 Steife Differenzialgleichungen in den Anwendungen

Die Linienmethode bei der Wärmeleitungsgleichung

Ein Anfangsrandwertproblem für die räumlich eindimensionale Wärmeleitungsgleichung ist gegeben durch

$$\begin{aligned} \frac{\partial u}{\partial t} &= \frac{\partial^2 u}{\partial x^2}, && 0 < x < L, \quad 0 < t < T, \\ u(0,t) &= u(L,t) = 0, && t \in [0,T], \\ u(x,0) &= f(x), && x \in [0,L], \end{aligned}$$

wobei $f : [0,L] \to \mathbb{R}$ eine gegebene Funktion ist. Die Funktion $u : [0,L] \times [0,T] \to \mathbb{R}$ soll numerisch bestimmt werden. Für äquidistante Gitterpunkte

$$x_j = j\Delta x, \qquad j = 1,2,\ldots,N-1 \qquad (\Delta x = L/N),$$

und eine hinreichend glatte Funktionen u ergibt eine Approximation von $\frac{\partial^2 u}{\partial x^2}(x_j,t)$, $1 \le j \le N-1$, durch zentrale Differenzenquotienten 2. Ordnung Folgendes (Details werden später vorgestellt, siehe Lemma 9.6):

$$\frac{\partial^2 u}{\partial x^2}(x_j,t) = \frac{u(x_{j+1},t) - 2u(x_j,t) + u(x_{j-1},t)}{(\Delta x)^2} + \mathcal{O}((\Delta x)^2).$$

Vernachlässigung des Terms $\mathcal{O}((\Delta x)^2)$ führt auf das folgende gekoppelte System von $N-1$ gewöhnlichen Differenzialgleichungen für $y_j(t) \approx u(x_j,t)$,

$$\left.\begin{aligned} y_j'(t) &= \frac{1}{(\Delta x)^2}\big(y_{j+1}(t) - 2y_j(t) + y_{j-1}(t)\big), && 0 < t < T, \\ y_j(0) &= f(x_j), && j = 1,2,\ldots,N-1, \end{aligned}\right\} \tag{8.93}$$

(mit $y_0(t) := y_N(t) := 0$) beziehungsweise in kompakter Form

$$y'(t) = -Ay(t), \qquad 0 < t < T, \qquad y(0) = w_0,$$

mit

$$y(t) = \big(y_1(t),\ldots,y_{N-1}(t)\big)^\top, \qquad w_0 = \big(f(x_1),\ldots,f(x_{N-1})\big)^\top,$$

$$A = \frac{1}{(\Delta x)^2}\begin{pmatrix} 2 & -1 & & \\ -1 & \ddots & \ddots & \\ & \ddots & \ddots & -1 \\ & & -1 & 2 \end{pmatrix} \in \mathbb{R}^{(N-1)\times(N-1)}.$$

Die vorgestellte Vorgehensweise, die Wärmeleitungsgleichung durch ein System gewöhnlicher Differenzialgleichungen bezüglich der Zeit t mittels Diskretisierung in Ortsrichtung x zu approximieren, wird als *Linienmethode* bezeichnet.

Die Eigenwerte λ_k der symmetrischen Matrix A lassen sich explizit berechnen (eine Herleitung wird in Lemma 9.12 nachgereicht),

$$\lambda_k \;=\; \frac{4}{(\Delta x)^2}\sin^2\Big(\frac{k\pi}{2N}\Big) \;>\; 0 \qquad \text{für} \quad k \;=\; 1,2,\ldots,N-1,$$

so dass das System (8.93) bezüglich des Skalarprodukts $\langle u,v\rangle \;=\; \sum_{j=1}^{N-1} u_j v_j$ dissipativ ist. Wegen

$$\lambda_{N-1} \quad\approx\quad \frac{4}{(\Delta x)^2}$$

ist es für kleine Ortsschrittweiten Δx sehr steif.

Weitere Themen und Literaturhinweise

Die auf Seite 177 genannten Lehrbücher zum Thema Einschrittverfahren enthalten allesamt auch Einführungen über Mehrschrittverfahren zur numerischen Lösung nichtsteifer Anfangswertprobleme. Im Folgenden werden einige weitere Themenkreise ansatzweise vorgestellt.

(a) Asymptotische Entwicklungen des globalen Verfahrensfehlers existieren auch für Mehrschrittverfahren. Wie sich herausstellt, liegen für spezielle Mehrschrittverfahren wie etwa die implizite Trapezregel oder das explizite Zweischrittverfahren von Gragg [38] asymptotische Entwicklungen in h^2 vor, bei denen man wie schon bei der summierten Trapezregel angepasste Extrapolationsverfahren verwendet, etwa das Gragg-Bulirsch-Stoer-Verfahren aus Bulirsch/Stoer [10]. Es besteht auch die Möglichkeit einer simultanen Anwendung von Extrapolationsverfahren und Schrittweitensteuerungsstrategien. Einzelheiten hierzu findet man beispielsweise in Deuflhard [18], [19] und in Hairer/Nørsett/Wanner [50].

(b) Für stetig partiell differenzierbare Funktionen f lässt sich eine obere Lipschitzbedingung auch noch sinnvoll definieren, falls die zugrunde liegende Vektornorm $\|\cdot\| : \mathbb{R}^N \to \mathbb{R}_+$ *nicht* durch ein Skalarprodukt induziert ist. Hierzu bedient man sich der *logarithmischen Norm* $\mu[\cdot] : \mathbb{R}^{N\times N} \to \mathbb{R}$, die folgendermaßen definiert ist,

$$\mu[A] \;:=\; \lim_{h\to 0+} \frac{\|I + hA\| - 1}{h}, \qquad A \in \mathbb{R}^{N\times N}, \tag{8.94}$$

wobei $\|\cdot\| : \mathbb{R}^{N\times N} \to \mathbb{R}_+$ die durch die zugrunde liegende Vektornorm induzierte Matrixnorm bezeichnet. Die logarithmische Norm ist unabhängig voneinander von Dahlquist [14] und Lozinski [67] eingeführt worden. Deren allgemeine Eigenschaften sowie konkrete Darstellungen für einige durch geläufige Vektornormen induzierte logarithmische Normen werden in den Aufgaben 8.11–8.16 vorgestellt. Mithilfe logarithmischer Normen lassen sich zum Beispiel Aussagen über die stetige Abhängigkeit von den Anfangswerten treffen. Gilt etwa bezüglich einer gegebenen Funktion $M : [a,b] \to \mathbb{R}$ eine verallgemeinerte obere Lipschitzbedingung von der Form

$$\mu[\mathcal{D}_y f(t,u)] \;\le\; M(t) \qquad \text{für} \quad t \in [a,b], \qquad u \in \mathbb{R}^N,$$

so behält die Fehlerabschätzung (8.81) über die stetige Abhängigkeit von den Anfangswerten ihre Gültigkeit (Dekker/Verwer [16]).

(c) Neben dem impliziten Euler-Verfahren eignen sich viele andere implizite Ein- und Mehrschrittverfahren zur numerischen Lösung steifer Anfangswertprobleme. Ausführliche Behandlungen dieses Themas findet man beispielsweise in Deuflhard/Bornemann [21], Hairer/Wanner [51] oder Strehmel/Weiner [101].

Übungsaufgaben

Aufgabe 8.1. Man zeige, dass ein lineares m-Schrittverfahren genau dann für alle Anfangswertprobleme mit hinreichend glatten Funktionen $f : [a,b] \times \mathbb{R}^N \to \mathbb{R}^N$ die Konsistenzordnung p besitzt, wenn mit der Notation

$$L[y(t),h] := \sum_{j=0}^{m} [\alpha_j y(t+jh) - h\beta_j y'(t+jh)]$$

die Beziehungen $L[t^0,h] = L[t^1,h] = \ldots = L[t^p,h] = 0$ erfüllt sind.

Aufgabe 8.2. Man bestimme mithilfe des Gleichungssystems (8.18) die (genaue) Konsistenzordnung des Zweischrittverfahrens

$$u_{\ell+2} - u_\ell = \frac{h}{3}\big(f(t_{\ell+2},u_{\ell+2}) + 4f(t_{\ell+1},u_{\ell+1}) + f(t_\ell,u_\ell)\big).$$

Für das Mehrschrittverfahren

$$u_{\ell+3} + \gamma(u_{\ell+2} - u_{\ell+1}) - u_\ell = h\frac{3+\gamma}{2}\big(f(t_{\ell+2},u_{\ell+2}) + f(t_{\ell+1},u_{\ell+1})\big)$$

bestimme man die von $\gamma \in \mathbb{R}$ abhängige Konsistenzordnung p. Für welche Werte von $\gamma \in \mathbb{R}$ ist das Verfahren nullstabil?

Aufgabe 8.3. Man zeige, dass für jede Zahl $m \in \mathbb{N}$ (bis auf Normierung) genau ein m-schrittiges lineares Verfahren

$$\sum_{j=0}^{m} \alpha_j u_{\ell+j} = h\sum_{j=0}^{m} \beta_j f(t_{\ell+j},u_{\ell+j})$$

mit der Konsistenzordnung $2m$ existiert, aber keines mit der Konsistenzordnung $2m+1$.

Hinweis: Für $p = 2m$ und $p = 2m+1$ betrachte man jeweils das Konsistenz-Gleichungssystem (8.18) für die Unbekannten α_j, $j = 0,1,\ldots,m$, und $-\beta_j$, $j = 0,1,\ldots,m$, und argumentiere wie zum Ende des Beweises von Theorem 8.54.

Aufgabe 8.4. (a) Für die homogene Differenzengleichung

$$u_{\ell+3} - 4u_{\ell+2} + 5u_{\ell+1} - 2u_\ell = 0, \qquad \ell = 0,1,\ldots$$

gebe man die allgemeine Lösung an.

(b) Man löse folgende Differenzengleichungen:

$$\begin{aligned}
u_{\ell+2} - 2u_{\ell+1} - 3u_\ell &= 0, \quad u_0 = 0, \quad u_1 = 1,\\
u_{\ell+1} - u_\ell &= 2^\ell, \quad u_0 = 0,\\
u_{\ell+1} - u_\ell &= \ell, \quad u_0 = 0,\\
u_{\ell+2} - 2tu_{\ell+1} + u_\ell &= 0, \quad u_0 = 1, \quad u_1 = t \in (-1,1).
\end{aligned}$$

Aufgabe 8.5. (a) Man zeige, dass jede Lösung $y(t)$ der skalaren Differenzialgleichung 2. Ordnung

$$y'' \;=\; f(t,y), \qquad t \in [a,b], \tag{8.95}$$

der folgenden Identität genügt (für $t,\, t \pm h \in [a,b]$):

$$\begin{aligned} &y(t+h) \;-\; 2y(t) \;+\; y(t-h) \\ &\quad=\; h^2 \int_0^1 (1-s)\big(f(t+sh, y(t+sh)) \;+\; f(t-sh, y(t-sh))\big)\, ds. \end{aligned} \tag{8.96}$$

(b) Zur numerischen Lösung einer Anfangswertaufgabe für (8.95) setze man in (8.96) $t = t_{\ell+m-1}$ und ersetze die Funktion $f(s, y(s))$ durch dasjenige Polynom $P \in \Pi_{m-1}$, welches die Stützpunkte $(t_{\ell+j},\, f_{\ell+j}),\ j = 0, \ldots, m-1$ interpoliert, wobei die übliche Notation $f_{\ell+j} = f(t_{\ell+j}, u_{\ell+j})$ verwendet wird. Daraus leite man die expliziten linearen *Störmer-Verfahren*

$$u_{\ell+m} \;-\; 2u_{\ell+m-1} \;+\; u_{\ell+m-2} \;=\; h^2 \sum_{k=0}^{m-1} \sigma_k \nabla^k f_{\ell+m-1}, \qquad \ell \;=\; 0, 1, \ldots, n-m$$

mit den Koeffizienten

$$\sigma_k \;=\; (-1)^k \int_0^1 (1-s)\Big(\binom{-s}{k} \;+\; \binom{s}{k} \Big)\, ds$$

her. Für $m = 2$ und $m = 3$ gebe man die Verfahren an.

Aufgabe 8.6. Man beweise: Für ein nullstabiles lineares Mehrschrittverfahren der Konsistenzordnung p gilt

$$\xi_1(h\lambda) \;=\; e^{h\lambda} \;+\; \mathcal{O}(h^{p+1}) \qquad \text{für} \quad h \to 0,$$

wobei $\xi_1(h\lambda)$ die Nullstelle des Polynoms

$$Q(\xi, h\lambda) \;=\; \rho(\xi) - h\lambda\sigma(\xi)$$

mit $\xi_1(h\lambda) \to \xi_1(0) = 1$ für $h\lambda \to 0$ bezeichnet. Hier ist ρ das erzeugende Polynom, und $\sigma(\xi) := \beta_m \xi^m \;+\; \ldots \;+\; \beta_0 \in \Pi_m$.

Aufgabe 8.7. Für die Fälle $m = 1, 2, 3$ rechne man die auf Seite 206 angegebenen expliziten Darstellungen der BDF-Formeln nach und und überprüfe jeweils die Nullstabilität.

Aufgabe 8.8. Das zweischrittige Verfahren

$$u_{\ell+2} \;+\; 4u_{\ell+1} \;-\; 5u_\ell \;=\; h\big(4 f(t_{\ell+1}, u_{\ell+1}) \;+\; 2 f(t_\ell, u_\ell) \big) \tag{8.97}$$

besitzt unter den üblichen Glattheitsvoraussetzungen die Konsistenzordnung $p = 3$. Ist es nullstabil? Man wende es mit der Schrittweite $h > 0$ und Startwerten $u_0 = 1$ und $u_1 = e^{-h}$ auf die Testgleichung $y' = -y,\ \ y(0) = 1$ an und zeige, dass mit $t \neq 0$ und $h = h_\ell = t/\ell$ für $\ell \to \infty$ Folgendes gilt:

$$u_\ell \;=\; (1 + \mathcal{O}(h^4))\big(e^{-t/\ell} \;+\; \mathcal{O}(h^4) \big)^\ell \;-\; \tfrac{1}{216} h^4 (1 \;+\; \mathcal{O}(h))(-5 \;-\; 3h \;+\; \mathcal{O}(h^2))^\ell,$$

und dabei der erste Summand für $\ell \to \infty$ gegen e^{-t} konvergiert und der zweite Summand sich für große ℓ verhält wie

$$-\frac{t^4}{216} \frac{(-5)^\ell}{\ell^4} e^{3t/5}.$$

Aufgabe 8.9 (*Numerische Aufgabe*). Man löse numerisch das Anfangswertproblem

$$y' \;=\; -y, \qquad y(0) \;=\; 1,$$

mit dem

- zweischrittigen Verfahren (8.97), einmal mit den Startwerten $u_0 = 1,\ u_1 = e^{-h}$ und dann auch mit den Startwerten $u_0 = 1,\ u_1 = \lambda_1 := -2 - 3h + \sqrt{9 + 6h + 4h^2}$;
- und für $\gamma = 0$ und $\gamma = 9$ mit dem dreischrittigen Verfahren

$$u_{\ell+3} + \gamma(u_{\ell+2} - u_{\ell+1}) - u_\ell = h\frac{3+\gamma}{2}\big(f(t_{\ell+2}, u_{\ell+2}) + f(t_{\ell+1}, u_{\ell+1})\big)$$

(vergl. Aufgabe 8.2) mit den Startwerten $u_0 = 1,\ u_1 = e^{-h}$ und $u_2 = e^{-2h}$.

Die Schrittweite sei jeweils $h = 0.01$. Geben Sie tabellarisch zu den Gitterpunkten $t = t_\ell = \ell h,\ \ell = 2, 3, \ldots, 100$ die exakte Lösung $y(t)$, die Näherung $u_h(t)$, den Fehler $u_h(t) - y(t)$ und im Falle des ersten Verfahrens $-\frac{t^4}{216}\frac{(-5)^\ell}{\ell^4}e^{3t/5}$ an.

Aufgabe 8.10 (*Numerische Aufgabe*). Man löse das Anfangswertproblem

$$\begin{aligned} y'(t) &= \lambda y(t), \qquad t \in [0, 15], \\ y(0) &= 1, \end{aligned}$$

für $\lambda = -1$ und $\lambda = 1$ jeweils mit den beiden folgenden Prädiktor-Korrektor-Verfahren:

1. Das *Verfahren von Milne* besitzt Prädiktor und Korrektor

$$\begin{aligned} u_{\ell+4}^{(0)} &= u_\ell + \frac{4}{3}h\big(2f_{\ell+3} - f_{\ell+2} + 2f_{\ell+1}\big) \\ u_{\ell+4}^{(\nu+1)} &= u_{\ell+2} + \frac{1}{3}h\big(f_{\ell+4}^{(\nu)} + 4f_{\ell+3} + f_{\ell+2}\big), \qquad \nu = 0, 1, \ldots . \end{aligned}$$

2. Das *Verfahren von Hamming* besitzt den gleichen Prädiktor wie das Verfahren von Milne, und der Korrektor ist hier

$$u_{\ell+4}^{(\nu+1)} - \frac{9}{8}u_{\ell+3} + \tfrac{1}{8}u_{\ell+1} = \frac{3}{8}h\big(f_{\ell+4}^{(\nu)} + 2f_{\ell+3} - f_{\ell+2}\big).$$

Hierbei bedeutet $f_\ell = f(t_\ell, u_\ell)$ und $f_{\ell+4}^{(\nu)} = f(t_{\ell+4}, u_{\ell+1}^{(\nu)})$. Für die Anlaufrechnung verwende man das klassische Runge-Kutta-Verfahren und für die Korrektoriteration das Abbruchkriterium

$$\frac{|u_{\ell+4}^{(\nu+1)} - u_{\ell+4}^{(\nu)}|}{|u_{\ell+4}^{(\nu)}|} \le 10^{-5}.$$

Man verwende jeweils die Schrittweite $h = 0.1$ und gebe tabellarisch zu den Gitterpunkten $t = 0.1, 0.2, 0.3, \ldots, 1.0, 2.0, 3.0, \ldots, 15$, die exakte Lösung $y(t)$, die Näherung $u_h(t)$, den Fehler $u_h(t) - y(t)$ und die Anzahl der durchgeführten Iterationsschritte an.

Aufgabe 8.11. Für die Matrix

$$A = \begin{pmatrix} -10 & 12 \\ 12 & -20 \end{pmatrix} \in \mathbb{R}^{2\times 2}$$

berechne man die logarithmischen Normen $\mu_\infty[A],\ \mu_1[A]$ und $\mu_2[A]$.

Aufgabe 8.12. Diskretisierung der Wärmeleitungsgleichung mit Neumann-Randbedingungen

$$\begin{aligned} \frac{\partial u}{\partial t} &= \frac{\partial^2 u}{\partial x^2} + f(x,t), \qquad 0 \le x \le 1, \qquad a \le t \le b, \\ \frac{\partial u}{\partial x}(0,t) &= \frac{\partial u}{\partial x}(1,t) = 0, \qquad \text{—— « ——}, \\ u(x,0) &= g(x), \qquad 0 \le x \le 1, \end{aligned}$$

führt mithilfe zentraler Differenzenquotienten erster und zweiter Ordnung (bei äquidistanter Ortsschrittweite $\Delta x = 1/N$) auf ein Anfangswertproblem für ein System von $N+1$ gewöhnlichen Differenzialgleichungen

$$y'(t) \;=\; Ay(t) \,+\, z(t), \qquad y(a) \;=\; z_0$$

mit einer geeigneten Matrix $A \in \mathbb{R}^{(N+1)\times(N+1)}$. Man gebe eine Matrixnorm an, so dass für die zugehörige logarithmische Norm $\mu[A] \le 0$ gilt.

Aufgabe 8.13. Man weise

$$\mu[A] \;=\; \lim_{h\to+0} \frac{\ln\|e^{hA}\|}{h} \qquad \text{für} \quad A \in \mathbb{R}^{N\times N}$$

nach. *Hinweis:* Zunächst zeige man

$$\mu[A] \;=\; \lim_{h\to+0} \frac{\|e^{hA}\|-1}{h}.$$

Aufgabe 8.14. Man weise nach, dass für Matrizen $A,\ B \in \mathbb{R}^{N\times N}$ und nichtnegative Zahlen $c \in \mathbb{R},\ c \ge 0$ Folgendes gilt,

$$\mu[cA] \;=\; c\mu[A], \qquad \mu[A+B] \;\le\; \mu[A] \,+\, \mu[B].$$

Aufgabe 8.15. Man zeige:

(a) Ist die Norm $\|\cdot\| : \mathbb{K}^N \to \mathbb{R}$ durch ein Skalarprodukt $\langle\cdot,\cdot\rangle : \mathbb{K}^N \times \mathbb{K}^N \to \mathbb{R}$ induziert, so gilt für die zugehörige logarithmische Norm die Darstellung

$$\mu[A] \;=\; \max_{x\in\mathbb{K}^N\,:\,|x|=1} \operatorname{Re}\langle Ax, x\rangle \qquad \text{für} \quad A \in \mathbb{K}^{N\times N},$$

wobei man im reellen Fall $\mathbb{K} = \mathbb{R}$ den Ausdruck $\operatorname{Re}\langle Ax, x\rangle$ durch $\langle Ax, x\rangle$ ersetzen kann. (Die Definition (8.94) für logarithmische Normen lässt sich auch für komplexe Matrizen beziehungsweise für Normen auf komplexen Räumen verwenden.)

(b) Für eine durch eine Vektornorm $\|\cdot\| : \mathbb{C}^N \to \mathbb{R}$ induzierte logarithmische Norm $\mu[\cdot] : \mathbb{C}^{N\times N} \to \mathbb{R}$ gilt die Ungleichung

$$\mu[A] \;\ge\; \max_{\lambda\in\sigma(A)} \operatorname{Re}\lambda \qquad \text{für} \quad A \in \mathbb{C}^{N\times N}.$$

Gilt hier im Allgemeinen Gleichheit?

Aufgabe 8.16. Sei $\mu_\infty[\cdot] : \mathbb{R}^{N\times N} \to \mathbb{R}$ die zur Maximumnorm $\|\cdot\|_\infty : \mathbb{R}^N \to \mathbb{R}$ gehörende logarithmische Norm. Man weise für $0 \ne A \in \mathbb{R}^{N\times N}$ die folgende Äquivalenz nach:

$$\mu_\infty[A] \;\le\; 0 \qquad \Longleftrightarrow \qquad \|I + \mu A\|_\infty \;\le\; 1 \qquad \forall\ 0 < \mu \;\le\; \|A\|_\infty.$$

9 Randwertprobleme bei gewöhnlichen Differenzialgleichungen

9.1 Problemstellung, Existenz, Eindeutigkeit

9.1.1 Problemstellung

Viele praxisrelevante Fragestellungen führen auf Randwertprobleme für gewöhnliche Differenzialgleichungen.

Beispiel 9.1. Die zeitlich stationäre Temperaturverteilung in einem dünnen Metallstab wird beschrieben durch das folgende Randwertproblem:

$$\begin{aligned} c\frac{\partial^2 u}{\partial x^2} &= f(x), \qquad a < x < b, \\ u(a) &= \alpha, \qquad u(b) = \beta, \end{aligned}$$

wobei $f : [a,b] \to \mathbb{R}$ eine gegebene Funktion ist, die anliegende, zeitlich unabhängige Wärmequellen darstellt. Die Funktion $u : [a,b] \to \mathbb{R}$ beschreibt die zeitlich unabhängige Temperaturverteilung in dem Stab und ist gesucht. Die Temperaturen (hier mit α beziehungsweise β bezeichnet) an den beiden Rändern sind vorgegeben, und $c > 0$ stellt eine Materialkonstante dar. △

Randwertprobleme für gewöhnliche Differenzialgleichungen sind Gegenstand des vorliegenden Kapitels.

Definition 9.2. Ein *Randwertproblem für eine gewöhnliche Differenzialgleichung zweiter Ordnung* mit separierten Randbedingungen ist von der Form

$$u'' = f(x,u,u'), \qquad x \in [a,b], \tag{9.1}$$

$$u(a) = \alpha, \qquad u(b) = \beta, \tag{9.2}$$

auf einem endlichen Intervall $[a,b]$ und mit gegebenen Zahlen $\alpha,\ \beta \in \mathbb{R}$ sowie einer Funktion $f : [a,b] \times \mathbb{R}^2 \to \mathbb{R}$, und gesucht ist eine zweimal stetig differenzierbare Funktion $u : [a,b] \to \mathbb{R}$ mit den Eigenschaften (9.1)–(9.2).

Die Notation in (9.1) ist eine übliche Kurzform für $u''(x) = f(x,u(x),u'(x)),\ \ x \in [a,b]$. Oft werden solche Randwertprobleme auch in abgeschwächter Form betrachtet, bei der eine stetige Lösung $u : [a,b] \to \mathbb{R}$ der Differenzialgleichung $u'' = f(x,u,u')$ lediglich auf dem offenen Intervall (a,b) gesucht wird (und die zweimalige stetige Differenzierbarkeit von u lediglich dort gefordert wird). Zur Vereinfachung der Situation werden Randwertprobleme im weiteren Verlauf in der spezielleren Fassung (9.1)–(9.2) betrachtet.

Bemerkung 9.3. In den Anwendungen treten auch Randwertprobleme für gewöhnliche Differenzialgleichungen höherer Ordnung und für Systeme von gewöhnlichen Differenzialgleichungen auf:

- Ein Randwertproblem für eine gewöhnliche Differenzialgleichung n-ter Ordnung mit linearen Randbedingungen ist von der Form

$$u^{(n)} = f(x, u, u', \ldots, u^{(n-1)}), \qquad x \in [a, b], \tag{9.3}$$

$$\sum_{k=0}^{n-1} \left(c_{jk} u^{(k)}(a) + d_{jk} u^{(k)}(b) \right) = \alpha_j, \qquad j = 0, 1, \ldots, n-1 \tag{9.4}$$

mit einer gegebenen Funktion $f : [a, b] \times \mathbb{R}^n \to \mathbb{R}$ und gegebenen reellen Koeffizienten c_{jk}, d_{jk} und $\alpha_j \in \mathbb{R}$ sowie einer zu bestimmenden n-mal stetig differenzierbaren Funktion $u : [a, b] \to \mathbb{R}$.

- Ein Randwertproblem für ein System von n gewöhnlichen Differenzialgleichungen erster Ordnung mit linearen Randbedingungen ist von der Form

$$U' = F(x, U), \qquad x \in [a, b], \tag{9.5}$$

$$AU(a) + BU(b) = U_0 \tag{9.6}$$

mit einer gegebenen Funktion $F : [a, b] \times \mathbb{R}^n \to \mathbb{R}^n$ und Matrizen $A, B \in \mathbb{R}^{n \times n}$ und einem Vektor $U_0 \in \mathbb{R}^n$, und mit einer zu bestimmenden differenzierbaren vektorwertigen Funktion $U : [a, b] \to \mathbb{R}^n$.

Jedes Randwertproblem von der Form (9.3)–(9.4) lässt sich mit den Setzungen $U_1 = u$, $U_2 = u', \ldots, U_n = u^{(n-1)}$ in ein Randwertproblem für ein System von n gewöhnlichen Differenzialgleichungen erster Ordnung von der Form (9.5)–(9.6) überführen. △

Die folgenden Betrachtungen beschränken sich auf die in (9.1)–(9.2) betrachteten Randwertprobleme für gewöhnliche Differenzialgleichungen zweiter Ordnung.

9.1.2 Existenz und Eindeutigkeit der Lösung

Wie schon bei Anfangswertproblemen für gewöhnliche Differenzialgleichungen ist auch bei Randwertproblemen zunächst die Frage der Existenz und Eindeutigkeit der Lösung zu behandeln.

Beispiel 9.4. Die homogene lineare gewöhnliche Differenzialgleichung zweiter Ordnung

$$u''(x) + u(x) = 0, \qquad a < x < b,$$

besitzt die allgemeine Lösung $u(x) = c_1 \sin x + c_2 \cos x$ für $x \in [a, b]$, mit Koeffizienten c_1, $c_2 \in \mathbb{R}$, wobei aus der Theorie der gewöhnlichen Differenzialgleichungen bekannt ist, dass hierfür keine weiteren Lösungen existieren. Im Folgenden sollen verschiedene Randbedingungen (auf unterschiedlichen Grundintervallen) betrachtet werden.

(a) Das Randwertproblem

$$u'' + u = 0 \quad \text{auf} \quad [0, \pi/2], \qquad u(0) = 0, \qquad u(\pi/2) = 1,$$

besitzt die eindeutige Lösung $u(x) = \sin x$, $x \in [0, \pi/2]$.

(b) Bei dem Randwertproblem

$$u'' + u = 0 \quad \text{auf} \quad [0, \pi], \qquad u(0) = 0, \qquad u(\pi) = 0,$$

stellt jede Funktion von der Gestalt $u(x) = c_1 \sin x$, $x \in [0, \pi]$, mit $c_1 \in \mathbb{R}$ eine Lösung dar.

(c) Schließlich existiert für das Randwertproblem

$$u'' + u = 0 \quad \text{auf} \quad [0, \pi], \qquad u(0) = 0, \qquad u(\pi) = 1,$$

keine Lösung. △

Durch das vorangegangene Beispiel 9.4 wird deutlich, dass es bei Randwertproblemen für gewöhnliche Differenzialgleichungen keine so allgemein gültige Existenz- und Eindeutigkeitsaussage wie bei Anfangswertproblemen gibt. Unter gewissen Zusatzbedingungen lassen sich jedoch Existenz und Eindeutigkeit nachweisen. Ein entsprechendes Resultat für die in (9.5)–(9.6) beschriebene allgemeine Situation bei Systemen von gewöhnlichen Differenzialgleichungen erster Ordnung findet man beispielsweise in Stoer/Bulirsch [99]. Es wird nun noch ein Spezialfall des Randwertproblems (9.1)–(9.2) bei gewöhnlichen Differenzialgleichungen zweiter Ordnung betrachtet. Es handelt sich hierbei um das folgende *sturm-liouvillesche Randwertproblem* mit homogenen Randbedingungen,

$$-u''(x) + r(x)u(x) = \varphi(x), \qquad a \le x \le b, \tag{9.7}$$

$$u(a) = u(b) = 0, \tag{9.8}$$

wobei $r, \varphi : [a, b] \to \mathbb{R}$ vorgegebene stetige Funktionen sind. Hier gilt die folgende Aussage:

Theorem 9.5. *Das Randwertproblem (9.7)–(9.8) besitzt für stetige Funktionen $r, \varphi : [a, b] \to \mathbb{R}$ eine eindeutig bestimmte Lösung $u \in C^2[a, b]$, falls r nicht-negativ ist, $r(x) \ge 0$ für $x \in [a, b]$.*

BEWEIS. Siehe Kress [63], Theorem 11.4. □

Zur numerischen Lösung von solchen Randwertproblemen (9.7)–(9.8) und allgemeiner von Randwertproblemen von der Form (9.1)–(9.2) werden im Folgenden *Differenzenverfahren*, *Variationsmethoden (Galerkin-Verfahren)* und *Einfachschießverfahren* vorgestellt.

9.2 Differenzenverfahren

9.2.1 Numerische Differenziation

In dem folgenden Lemma wird der später benötigte zentrale Differenzenquotient zweiter Ordnung (zur Approximation der zweiten Ableitung einer Funktion von einer Veränderlichen) definiert und seine Approximationseigenschaften behandelt. Bei dieser Gelegenheit werden gleich noch die gängigen Differenzenquotienten zur Approximation der ersten Ableitung vorgestellt.

Lemma 9.6. *(a) Für $u \in C^2[a,b]$ gelten mit geeigneten Zahlen $\theta_1,\ \theta_2 \in [0,1]$ die Beziehungen*

$$\frac{u(x+h)-u(x)}{h} = u'(x) + u''(x+\theta_1 h)\frac{h}{2} \quad (\textit{vorwärts gerichteter Differenzenquotient})$$

$$\frac{u(x)-u(x-h)}{h} = u'(x) - u''(x-\theta_2 h)\frac{h}{2} \quad (\textit{rückwärts} \text{ ——— « ———}).$$

(b) Für $u \in C^3[a,b]$ gilt mit einer geeigneten Zahl $\theta \in [-1,1]$ Folgendes,

$$\frac{u(x+h)-u(x-h)}{2h} = u'(x) + u^{(3)}(x+\theta h)\frac{h^2}{6} \quad \textit{(zentraler Differenzenquotient 1. Ordnung).}$$

(c) Für $u \in C^4[a,b]$ gilt mit einer geeigneten Zahl $\theta \in [-1,1]$ Folgendes,

$$\frac{u(x+h) - 2u(x) + u(x-h)}{h^2} = u''(x) - u^{(4)}(x+\theta h)\frac{h^2}{12} \quad \textit{(zentraler Differenzenquotient 2. Ordnung).}$$

Die rechts vorgestellten Bezeichnungen beziehen sich auf die linke Seite der jeweiligen Gleichung.

BEWEIS. Die Aussagen erhält man mittels geeigneter Taylorentwicklungen der Funktion u in x.
(a) Hier verwendet man

$$u(x \pm h) = u(x) \pm u'(x)h + u''(x \pm \theta_{1/2}h)\frac{h^2}{2}.$$

(b) Eine weitere Taylorentwicklung der Funktion u in x liefert mit geeigneten Zahlen $\theta_1,\ \theta_2 \in [0,1]$

$$u(x \pm h) = u(x) \pm u'(x)h + u''(x)\frac{h^2}{2} \pm u^{(3)}(x \pm \theta_{1/2}h)\frac{h^3}{6},$$

und eine Subtraktion führt auf die angegebene Darstellung,

$$\begin{aligned}\frac{u(x+h) - u(x-h)}{2h} &= 0 + u'(x)h + 0 + \big(u^{(3)}(x+\theta_1 h) + u^{(3)}(x-\theta_2 h)\big)\frac{h^2}{12} \\ &\overset{(*)}{=} u'(x)h + u^{(3)}(x+\theta h)\frac{h^2}{6},\end{aligned}$$

mit einer Zahl $\theta \in [-1,1]$, wobei man die Identität $(*)$ mithilfe des Mittelwertsatzes erhält.
(c) Ganz entsprechend erhält man mit geeigneten Zahlen $\theta_1,\ \theta_2 \in [0,1]$ auch

$$u(x \pm h) \;=\; u(x) \;\pm\; u'(x)h \;+\; u''(x)\frac{h^2}{2} \;\pm\; u^{(3)}(x)\frac{h^3}{6} \;+\; u^{(4)}(x \pm \theta_{1/2}h)\frac{h^4}{24},$$

und daraus erhält man für eine Zahl $\theta \in [-1,1]$ die folgende Identität,

$$\begin{aligned}
&\frac{u(x+h) \;-\; 2u(x) \;+\; u(x-h)}{h^2} \\
&\quad=\; 0 \;+\; u''(x) \;+\; 0 \;+\; \big(u^{(4)}(x \pm \theta_1 h) \;+\; u^{(4)}(x \pm \theta_2 h)\big)\frac{h^2}{24} \\
&\quad=\; u''(x) \;+\; u^{(4)}(x+\theta h)\frac{h^2}{12}.
\end{aligned}$$

□

9.2.2 Der Ansatz für Differenzenverfahren

Im Folgenden wird der Ansatz für Differenzenverfahren vorgestellt, wobei dies anhand des speziellen Randwertproblems $-u'' + ru = \varphi,\ u(a) = u(b) = 0$ mit der nichtnegativen Funktion $r \geq 0$ geschieht[1]. Das zugrunde liegende Intervall $[a,b]$ wird mit Gitterpunkten versehen, die hier äquidistant gewählt seien,

$$x_j \;=\; a \;+\; jh, \qquad j \;=\; 0,1,\ldots,N \qquad\quad \text{mit} \quad h \;=\; \frac{b-a}{N}. \tag{9.9}$$

Eine Betrachtung des genannten Randwertproblems $-u'' + ru = \varphi,\ u(a) = u(b) = 0$ an diesen Gitterpunkten bei einer gleichzeitigen Approximation der Werte $u''(x_1)$, $\ldots, u''(x_{N-1})$ durch jeweils entsprechende zentrale Differenzenquotienten 2. Ordnung führt auf das folgende gekoppelte System von $N-1$ linearen Gleichungen,

$$\left.\begin{aligned}
&\frac{-v_{j+1} \;+\; 2v_j \;-\; v_{j-1}}{h^2} \;+\; r(x_j)v_j \;=\; \varphi(x_j), \qquad j \;=\; 1,2,\ldots,N-1, \\
&\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad (v_0 \;=\; v_N \;=\; 0)
\end{aligned}\right\} \tag{9.10}$$

für die Approximationen $v_j \approx u(x_j),\ j = 1,\ldots,N-1$. Setzt man noch

$$r_j \;=\; r(x_j), \qquad \varphi_j \;=\; \varphi(x_j), \qquad\quad j \;=\; 1,2,\ldots,N-1,$$

[1] vergleiche (9.7)–(9.8)

so erhält man für das Gleichungssystem (9.10) die folgende Matrix-Vektor-Darstellung

$$\underbrace{\frac{1}{h^2}\begin{pmatrix} 2+r_1h^2 & -1 & & \\ -1 & 2+r_2h^2 & \ddots & \\ & \ddots & \ddots & -1 \\ & & -1 & 2+r_{N-1}h^2 \end{pmatrix}}_{=:\ A\ \in\ \mathbb{R}^{(N-1)\times(N-1)}} \begin{pmatrix} v_1 \\ v_2 \\ \vdots \\ v_{N-1} \end{pmatrix} = \begin{pmatrix} \varphi_1 \\ \varphi_2 \\ \vdots \\ \varphi_{N-1} \end{pmatrix} \in \mathbb{R}^{N-1}. \quad (9.11)$$

Daraus erhält man unmittelbar die folgende Fehlerdarstellung:

Theorem 9.7. *Für das Differenzenschema (9.10) zur Lösung des Randwertproblems (9.7)–(9.8) mit $r \geq 0$ gilt mit der Notation $u_j := u(x_j)$ und der Matrix A aus (9.11) die Fehlerdarstellung*

$$\frac{1}{h^2}A\begin{pmatrix} v_1 - u_1 \\ \vdots \\ v_{N-1} - u_{N-1} \end{pmatrix} = -\frac{h^2}{12}\begin{pmatrix} u^{(4)}(x_1+\theta_1 h) \\ \vdots \\ u^{(4)}(x_{N-1}+\theta_{N-1}h) \end{pmatrix}. \quad (9.12)$$

BEWEIS. Die Aussage folgt unmittelbar aus der zu (9.10) äquivalenten Darstellung (9.11) und der aus Teil (c) in Lemma 9.6 resultierenden Identität

$$\frac{1}{h^2}A\begin{pmatrix} u_1 \\ \vdots \\ u_{N-1} \end{pmatrix} = \begin{pmatrix} \varphi_1 \\ \vdots \\ \varphi_{N-1} \end{pmatrix} + \frac{h^2}{12}\begin{pmatrix} u^{(4)}(x_1+\theta_1 h) \\ \vdots \\ u^{(4)}(x_{N-1}+\theta_{N-1}h) \end{pmatrix}. \qquad \square$$

Für den Nachweis der eindeutigen Lösbarkeit des Gleichungssystems (9.10) und die gleichzeitige Herleitung eine Normabschätzung des Fehlers in (9.12) wird im Folgenden

- die Regularität der Matrix $A \in \mathbb{R}^{(N-1)\times(N-1)}$ nachgewiesen sowie
- eine Abschätzung der Form $h^2 \| A^{-1} \|_\infty \leq C$ geliefert mit einer von der Zahl N unabhängigen Konstanten $C > 0$.

Hierzu sind ein paar Vorbereitungen erforderlich.

9.2.3 Das Konvergenzresultat für Differenzenverfahren

Definition 9.8. Für zwei Matrizen $S = (s_{jk}),\ T = (t_{jk}) \in \mathbb{R}^{N\times N}$ schreibt man

$$S \leq T \quad :\Longleftrightarrow \quad s_{jk} \leq t_{jk} \quad \text{für} \quad j,k = 1,2,\ldots,N,$$

beziehungsweise äquivalent dazu $T \geq S$. Eine Matrix $S \in \mathbb{R}^{N\times N}$ heißt *nichtnegativ*, wenn $S \geq 0$ gilt.

Im Folgenden werden die unmittelbar erforderlichen Resultate über nichtnegative Matrizen geliefert. Weitere Eigenschaften solcher Matrizen werden in Abschnitt 9.2.4 vorgestellt.

Lemma 9.9. *Für gegebene Matrizen $S,\ T \in \mathbb{R}^{N\times N}$ gelten die folgenden Implikationen,*

$$0 \leq S \leq T \quad \Longrightarrow \quad \|S\|_\infty \leq \|T\|_\infty; \tag{9.13}$$

$$T \geq 0 \quad \Longrightarrow \quad \|T\|_\infty = \left\| T \begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix} \right\|_\infty . \tag{9.14}$$

BEWEIS. Mit den Notationen $S = (s_{jk}),\ T = (t_{jk}) \in \mathbb{R}^{N\times N}$ erhält man die Aussage (9.13) folgendermaßen,

$$\begin{aligned} \|S\|_\infty &= \max_{j=1,\dots,N} \sum_{k=1}^{N} |s_{jk}| = \max_{j=1,\dots,N} \sum_{k=1}^{N} s_{jk} \leq \max_{j=1,\dots,N} \sum_{k=1}^{N} t_{jk} \\ &= \max_{j=1,\dots,N} \sum_{k=1}^{N} |t_{jk}| = \|T\|_\infty . \end{aligned}$$

Aus den letzten beiden Identitäten resultiert dann auch die Aussage (9.14). □

Das folgende Theorem liefert die wesentlichen Hilfsmittel für den Beweis der nachfolgenden Fehlerabschätzung bei Differenzenverfahren zur Lösung von Randwertproblemen. Der Beweis von Teil (a) dieses Theorems wird in Abschnitt 9.2.4 nachgereicht.

Theorem 9.10. *(a) Die Matrix $A \in \mathbb{R}^{(N-1)\times(N-1)}$ aus (9.11) ist regulär, und im Ordnungssinn gilt (vergleiche Definition 9.8)*

$$0 \leq A^{-1} \leq A_0^{-1}, \quad A_0 := \begin{pmatrix} 2 & -1 & & \\ -1 & \ddots & \ddots & \\ & \ddots & \ddots & -1 \\ & & -1 & 2 \end{pmatrix} \in \mathbb{R}^{(N-1)\times(N-1)} \text{ ist regulär;} \tag{9.15}$$

(b) es gilt

$$\|A^{-1}\|_\infty \leq \|A_0^{-1}\|_\infty \leq \frac{(b-a)^2}{8}\, h^{-2}. \tag{9.16}$$

BEWEIS. Der Beweis von Teil (a) wird nachgetragen, hier wird nur der Nachweis für Teil (b) geführt. Das spezielle Randwertproblem

$$-z''(x) = 1, \qquad a < x < b, \qquad\qquad z(a) = z(b) = 0,$$

besitzt die Lösung

$$z(x) \;=\; \tfrac{1}{2}(x-a)\,(b-x), \qquad a \le x \le b,$$

so dass insbesondere $z \in C^4[a,b]$ und $z^{(4)} \equiv 0$ gilt. Aus der Fehlerdarstellung für den zentralen Differenzenquotienten 2. Ordnung erhält man deshalb

$$A_0^{-1}\begin{pmatrix}1\\ \vdots \\ 1\end{pmatrix} \;=\; \frac{1}{h^2}\begin{pmatrix}z_1\\ \vdots \\ z_{N-1}\end{pmatrix}, \tag{9.17}$$

mit der Notation $z_j = z(x_j)$. Die zweite Abschätzung in (9.16) folgt nun unmittelbar aus (9.14) sowie Teil (a) dieses Theorems, und die erste Abschätzung in (9.16) erhält man sofort aus (9.13) sowie wiederum aus Teil (a) dieses Theorems. □

Die vorherige Aussage ermöglicht die Herleitung der folgenden Fehlerabschätzung für Differenzenverfahren zur Lösung von Randwertproblemen.

Theorem 9.11. *Gegeben sei das Randwertproblem (9.7)–(9.8) mit $r \ge 0$, für dessen Lösung $u \in C^4[a,b]$ erfüllt sei. Dann gilt*

$$\max_{j=0,\dots,N} |v_j - u(x_j)| \;\le\; Mh^2,$$

mit der Konstanten $M := \frac{(b-a)^2}{96}\|u^{(4)}\|_\infty$ und den Notationen aus (9.9) und (9.10).

BEWEIS. Die Aussage folgt unmittelbar aus den Theoremen 9.7 und 9.10. □

9.2.4 Vorbereitungen für den Beweis von Teil (a) des Theorems 9.10

Die Regularität der Matrix A_0 aus (9.15) ist eine unmittelbare Konsequenz aus der Tatsache, dass die Eigenwerte von Tridiagonalmatrizen mit konstanten Einträgen entlang der Haupt- und der Nebendiagonalen direkt angegeben werden können:

Lemma 9.12. *Eine Tridiagonalmatrix*

$$A \;=\; \begin{pmatrix} a & b & & \\ c & a & \ddots & \\ & \ddots & \ddots & b \\ & & c & a \end{pmatrix} \;\in\; \mathbb{R}^{(N-1)\times(N-1)}$$

mit Zahlen $a,\ b,\ c \in \mathbb{R},\ b\cdot c > 0$, besitzt die folgenden Eigenwerte,

$$\lambda_k \;=\; a \;+\; 2\,\mathrm{sgn}(c)\,\sqrt{bc}\,\cos\big(\frac{k\pi}{N}\big), \qquad k \;=\; 1,2,\dots,N-1.$$

Die zugehörigen Eigenvektoren sind im Beweis angegeben.

BEWEIS. Zur Vereinfachung der Notation wird im Folgenden der Fall $a = 0$ betrachtet. (Die Aussage in der allgemeinen Situation erhält man danach durch Betrachten der Matrix $A - aI$.) Mit den Setzungen

$$D := \frac{\pi}{N}, \qquad M := \left(\frac{c}{b}\right)^{1/2},$$
$$x^{[k]} := (x_\ell^{[k]})_{\ell=1}^{N-1} \in \mathbb{R}^{N-1} \quad \text{mit} \quad x_\ell^{[k]} := M^{\ell/2}\sin(k\ell D)$$

erhält man unter Verwendung der Darstellung

$$x_\ell^{[k]} = \frac{M^\ell}{2\mathrm{i}}\left(e^{\mathrm{i}k\ell D} - e^{-\mathrm{i}k\ell D}\right), \qquad \ell = 1, 2, \ldots, N-1, \tag{9.18}$$

für $j = 1, 2, \ldots, N-1$ Folgendes,

$$\begin{aligned}(Ax^{[k]})_j &= \frac{1}{2\mathrm{i}}\Big[cM^{j-1}e^{\mathrm{i}(j-1)kD} + bM^{j+1}e^{\mathrm{i}(j+1)kD} \\ &\qquad - \left(cM^{j-1}e^{-\mathrm{i}(j-1)kD} + bM^{j+1}e^{-\mathrm{i}(j+1)kD}\right)\Big] \\ &= \frac{M^j}{2\mathrm{i}}\Big[\left(cM^{-1}e^{-\mathrm{i}kD} + bMe^{\mathrm{i}kD}\right)e^{\mathrm{i}jkD} - \left(cM^{-1}e^{\mathrm{i}kD} + bMe^{-\mathrm{i}kD}\right)e^{-\mathrm{i}jkD}\Big],\end{aligned}$$

wobei diese Vorgehensweise auch in den Fällen $j = 1$ und $j = N-1$ zulässig ist, da die rechte Seite der Gleichung in (9.18) für $\ell = 0$ und $\ell = N$ verschwindet. Wegen $cM^{-1} = bM = \operatorname{sgn}(c)\sqrt{bc}$ berechnet man daraus mit der Abkürzung $\sigma = \operatorname{sgn}(c)$ Folgendes,

$$\begin{aligned}&(Ax^{[k]})_j \\ &= \frac{M^j}{2\mathrm{i}}\left[\left(\sigma\sqrt{bc}e^{-\mathrm{i}kD} + \sigma\sqrt{bc}e^{\mathrm{i}kD}\right)e^{\mathrm{i}jkD} - \left(\sigma\sqrt{bc}e^{-\mathrm{i}kD} + \sigma\sqrt{bc}e^{\mathrm{i}kD}\right)e^{\mathrm{i}jkD}\right] \\ &= \frac{M^j}{2\mathrm{i}}\left[\sigma\sqrt{bc}\left(e^{\mathrm{i}kD} + e^{-\mathrm{i}kD}\right)\right]\left(e^{\mathrm{i}jkD} - e^{\mathrm{i}jkD}\right) = \left(2\sigma\sqrt{bc}\cos(kD)\right)x_j^{[k]}. \qquad \square\end{aligned}$$

Für Matrizen A, deren Eigenwerte allesamt im offenen Einheitskreis liegen, lässt sich die Inverse der Matrix $I - A$ als *neumannsche Reihe* darstellen. Genauer gilt Folgendes:

Theorem 9.13. *Für eine Matrix $A \in \mathbb{R}^{N\times N}$ sind die folgenden Aussagen äquivalent:*

(a) $\sigma(A) \subset \{\lambda \in \mathbb{C} : |\lambda| < 1\}$;

(b) Es existiert eine Vektornorm $\|\cdot\| : \mathbb{R}^N \to \mathbb{R}$, so dass für die zugehörige Matrixnorm gilt $\|A\| < 1$;

(c) Die Reihe $\sum_{\nu=0}^{\infty} A^\nu$ ist konvergent;

(d) Es gilt $A^\nu \to 0$ für $\nu \to \infty$.

Wenn eine der (und damit alle) Bedingungen erfüllt ist, so gilt

$$(I - A)^{-1} \;=\; \sum_{\nu=0}^{\infty} A^{\nu}. \tag{9.19}$$

BEWEIS. $(a) \implies (b)$: Für jede Zahl $\varepsilon > 0$ existiert[2] eine verallgemeinerte Jordan-Faktorisierung der Form $A = T^{-1}\widehat{J}T$ mit einer regulären Matrix $T \in \mathbb{C}^{N\times N}$ sowie

$$\widehat{J} \;=\; \begin{pmatrix} \widehat{J}_1 & & \\ & \ddots & \\ & & \widehat{J}_r \end{pmatrix}, \qquad \widehat{J}_k \;=\; \begin{pmatrix} \lambda_k & \varepsilon & & \\ & \ddots & \ddots & \\ & & \ddots & \varepsilon \\ & & & \lambda_k \end{pmatrix} \in \mathbb{C}^{N_k\times N_k}, \qquad k \;=\; 1, 2, \ldots, r,$$

mit $N_k \geq 1$, $\sum_{k=1}^{r} N_k = N$. Im Fall $N_k = 1$ bedeutet dies $\widehat{J}_k = [\lambda_k] \in \mathbb{C}^{1\times 1}$. Hier sei nun $\varepsilon > 0$ hinreichend klein gewählt, so dass für jeden Index $k \in \{1, 2, \ldots, r\}$ die Ungleichung $|\lambda_k| + \varepsilon < 1$ erfüllt ist, was wegen Voraussetzung (a) möglich ist. Aufgrund der Konstruktion gilt

$$\|\widehat{J}\|_\infty \;=\; \max_{k=1,\ldots,r} \|\widehat{J}_k\|_\infty \;<\; 1.$$

Man setzt dann

$$\|x\|_T \;:=\; \|Tx\|_\infty, \qquad\qquad x \in \mathbb{R}^N,$$

und weist leicht nach, dass $\|\cdot\|_T$ eine Norm auf $\mathbb{R}^N$ darstellt. Für die zugehörige Matrixnorm ist dann tatsächlich $\|A\|_T < 1$ erfüllt, denn für jeden Vektor $x \in \mathbb{R}^N$ gilt

$$\|Ax\|_T \;=\; \|TAx\|_\infty \;=\; \|\widehat{J}Tx\|_\infty \;\leq\; \|\widehat{J}\|_\infty \|Tx\|_\infty \;=\; \|\widehat{J}\|_\infty \|x\|_T.$$

$(b) \implies (c)$: Die Behauptung folgt unmittelbar aus der absoluten Konvergenz,

$$\sum_{\nu=0}^{\infty} \|A^{\nu}\| \;\leq\; \sum_{\nu=0}^{\infty} \|A\|^{\nu} \;<\; \infty.$$

$(c) \implies (d)$: In jedem mit einer Norm versehenen Vektorraum folgt aus der Konvergenz einer Reihe $\sum_{j=0}^{\infty} x_j$ die Konvergenz seiner Summanden gegen null, $x_j \to 0$ $(j \to \infty)$.
$(d) \implies (a)$: Wenn $\lambda \in \mathbb{C}$ ein Eigenwert von A mit $|\lambda| \geq 1$ ist, so erhält man mit einem zugehörigen Eigenvektor $x \in \mathbb{C}^N$ und für jede Vektornorm $\|\cdot\| : \mathbb{C}^N \to \mathbb{R}$

$$\|A^{\nu}x\| \;=\; \|\lambda^{\nu}x\| \;=\; |\lambda|^{\nu}\|x\| \quad\geq\quad \|x\|$$

beziehungsweise $\|A^{\nu}\| \geq 1$ für $\nu = 1, 2, \ldots$ im Widerspruch zur Annahme (d).

Schließlich gilt unter den Bedingungen (a)–(d)

$$(I - A)\sum_{\nu=0}^{n-1} A^{\nu} \;=\; \sum_{\nu=0}^{n} (A^{\nu} - A^{\nu+1}) \;=\; I - A^{n} \;\to\; I \qquad \text{für} \quad n \to \infty,$$

woraus man die Darstellung (9.19) erhält. □

[2] siehe den Beweis von Lemma 8.15

Weitere Eigenschaften nichtnegativer Matrizen

Es folgen einige Aussagen über nichtnegative Matrizen.

Lemma 9.14. *Für nichtnegative Matrizen $S,\ T \in \mathbb{R}^{N\times N}$ sind sowohl $S+T \in \mathbb{R}^{N\times N}$ als auch $S\,T \in \mathbb{R}^{N\times N}$ nichtnegative Matrizen. Weiter gilt für Matrizen $S_1,\ S_2 \in \mathbb{R}^{N\times N}$ und $T_1,\ T_2 \in \mathbb{R}^{N\times N}$ mit $0 \le S_1 \le S_2$ und $0 \le T_1 \le T_2$ auch $0 \le S_1T_1 \le S_2T_2$. Konvergente Folgen nichtnegativer Matrizen besitzen nichtnegative Grenzwerte.*

BEWEIS. Ist elementar und wird hier nicht geführt. □

Theorem 9.15. *Für Matrizen $S,\ T \in \mathbb{R}^{N\times N}$ und $\lambda \in \mathbb{R}$ gilt die folgende Implikation,*

$$\left\{\begin{array}{r} 0 \le S \le T, \\ \lambda > r_\sigma(T) \end{array}\right\} \implies \left\{\begin{array}{l} \lambda > r_\sigma(S), \\ 0 \le (\lambda I - S)^{-1} \le (\lambda I - T)^{-1}. \end{array}\right\} \tag{9.20}$$

BEWEIS. Zunächst wird der Spezialfall $\lambda = 1 > r_\sigma(T)$ betrachtet. Es ist $\sum_{\nu=0}^{\infty} S^\nu$ konvergent, denn unter Anwendung von Lemma 9.9, Lemma 9.14 und Theorem 9.13 erhält man

$$\| \sum_{\nu=n_0}^{n_1} S^\nu \|_\infty \le \| \sum_{\nu=n_0}^{n_1} T^\nu \|_\infty \to 0 \qquad \text{für} \quad n_0 \le n_1, \qquad n_0,\ n_1 \to \infty.$$

Wiederum nach Theorem 9.13 folgt daraus $1 > r_\sigma(S)$ sowie die Darstellbarkeit der Inversen der Matrix $I - S$ als neumannsche Reihe, $(I-S)^{-1} = \sum_{\nu=0}^{\infty} S^\nu$. Daraus resultiert schließlich der zweite Teil der Aussage (9.20) für den Spezialfall $\lambda = 1$,

$$(I-S)^{-1} = \sum_{\nu=0}^{\infty} S^\nu \le \sum_{\nu=0}^{\infty} T^\nu = (I-T)^{-1}.$$

Die Aussage für die allgemeine Situation $\lambda > 0$ erhält man durch Betrachtung von $\lambda^{-1}S$ und $\lambda^{-1}T$: es gilt $\lambda^{-1}S \le \lambda^{-1}T$ sowie $1 > r_\sigma(\lambda^{-1}T)$, mit der schon bewiesenen Aussage (9.20) für den Spezialfall $\lambda = 1$ erhält man die Regularität der Matrix $I - \lambda^{-1}S$ sowie $(I-\lambda^{-1}S)^{-1} \le (I-\lambda^{-1}T)^{-1}$ und daraus wiederum unmittelbar die Aussage (9.20) in ihrer ganzen Allgemeinheit. □

Als unmittelbare Konsequenz erhält man das folgende Resultat.

Theorem 9.16. *Für Matrizen $A,\ B \in \mathbb{R}^{N\times N}$ mit $0 \le A \le B$ gilt $r_\sigma(A) \le r_\sigma(B)$.*

BEWEIS. Diese Aussage erhält man unmittelbar durch Anwendung von Theorem 9.15 für $\lambda = r_\sigma(A) + \varepsilon$ mit $\varepsilon > 0,\ \varepsilon \to 0$. □

Das folgende Resultat für nichtnegative Matrizen wird im nachfolgenden Kapitel 10 benötigt.

Theorem 9.17. *Für jede Matrix $B \in \mathbb{R}^{N\times N}$ mit $B \geq 0$ und jede Zahl $\lambda > 0$ gilt die folgende Äquivalenz,*

$$\lambda > r_\sigma(B) \quad \Longleftrightarrow \quad \left\{ \begin{array}{l} \lambda I - B \quad \textit{ist regulär,} \\ (\lambda I - B)^{-1} \geq 0. \end{array} \right\} \tag{9.21}$$

BEWEIS. Die Implikation "$\Longrightarrow$" folgt unmittelbar aus Theorem 9.15 angewandt mit $S = 0$. Für den Nachweis der Implikation "$\Longleftarrow$" wird zunächst der Spezialfall $\lambda = 1$ betrachtet. Ist die Matrix $I - B$ regulär und gilt $(I - B)^{-1} \geq 0$, so folgt

$$\begin{aligned} 0 \;\leq\; \sum_{\nu=0}^{n-1} B^\nu \;&=\; \sum_{\nu=0}^{n-1} B^\nu (I-B)(I-B)^{-1} \;=\; \sum_{\nu=0}^{n-1} (B^\nu - B^{\nu+1})(I-B)^{-1} \\ &=\; (I - \underbrace{B^n}_{\geq 0}) \underbrace{(I-B)^{-1}}_{\geq 0} \;\leq\; (I-B)^{-1}, \end{aligned}$$

beziehungsweise insbesondere

$$0 \;\leq\; \sum_{\nu=0}^{n-1} B^\nu \;\leq\; (I-B)^{-1}, \qquad n \;=\; 1, 2, \ldots .$$

Wegen $\sum_{\nu=0}^{n-1} B^\nu \leq \sum_{\nu=0}^{n} B^\nu$ für $n = 1, 2, \ldots$ ist also $\sum_{\nu=0}^{\infty} B^\nu$ notwendigerweise konvergent und damit gilt[3] die Ungleichung $r_\sigma(B) < 1$.

Die allgemeine Situation $\lambda > 0$ für die Implikation "$\Longleftarrow$" in der Aussage (9.21) lässt sich auf den Fall $\lambda = 1$ zurückführen,

$$\begin{aligned} \lambda > r_\sigma(B) \quad &\Longleftrightarrow \quad 1 > r_\sigma(\lambda^{-1} B), \\ \lambda I - B \;\text{ regulär, }\; (\lambda I - B)^{-1} \geq 0 \quad &\Longleftrightarrow \quad I - \lambda^{-1} B \;\text{ regulär, }\; (I - \lambda^{-1} B)^{-1} \geq 0. \end{aligned}$$

Dies komplettiert den Beweis des Theorems. □

Als Konsequenz aus Theorem 9.17 erhält man das folgende klassische Resultat.

Theorem 9.18 (Satz von Perron). *Für jede Matrix $A \in \mathbb{R}^{N\times N}$ mit $A \geq 0$ ist die Zahl $\lambda = r_\sigma(A)$ ein Eigenwert von A.*

BEWEIS. Wäre die Matrix $\lambda I - A$ regulär, so ergäbe sich

$$0 \;\overset{(*)}{\leq}\; ((\lambda + \varepsilon) I - A)^{-1} \;\overset{(**)}{\to}\; (\lambda I - A)^{-1} \qquad \text{für} \quad 0 < \varepsilon \to 0,$$

wobei die Ungleichung $(*)$ aus Theorem 9.17 resultiert, und $(**)$ folgt mit Korollar 4.50 über die Stetigkeit der Matrixinversion. Daraus erhält man $(\lambda I - A)^{-1} \geq 0$ im Widerspruch zur Aussage von Theorem 9.17. □

[3] vergleiche Theorem 9.13

9.2.5 Nachweis der Aussage in Teil (a) von Theorem 9.10

Für den Nachweis der Aussage (9.15) betrachtet man die folgenden Matrizen D, D_0, S und $S_0 \in \mathbb{R}^{(N-1)\times(N-1)}$,

$$D = 2I + h^2 \operatorname{diag}(r_1, \ldots, r_{N-1}), \qquad D_0 = 2I,$$

$$S = \begin{pmatrix} 0 & \frac{1}{2 + r_1 h^2} & & \\ \frac{1}{2 + r_2 h^2} & \ddots & \ddots & \\ & \ddots & 0 & \frac{1}{2 + r_{N-2} h^2} \\ & & \frac{1}{2 + r_{N-1} h^2} & 0 \end{pmatrix}, \qquad S_0 = \begin{pmatrix} 0 & \frac{1}{2} & & \\ \frac{1}{2} & \ddots & \ddots & \\ & \ddots & \ddots & \frac{1}{2} \\ & & \frac{1}{2} & 0 \end{pmatrix}$$

und erhält damit die Darstellungen

$$A = D(I - S), \qquad A_0 = D_0(I - S_0).$$

Mit Lemma 9.12 erhält man

$$\sigma(S_0) = \left\{ \cos\left(\tfrac{k\pi}{N}\right) : k = 1, \ldots, N-1 \right\} \subset \{x : -1 < x < 1\},$$

und offensichtlich gilt $0 \leq S \leq S_0$, so dass nach Theorem 9.15 die Matrizen $I - S_0$ und $I - S$ regulär sind und mehr noch

$$0 \leq (I - S)^{-1} \leq (I - S_0)^{-1}$$

gilt. Weiterhin sind die Matrizen D und D_0 offensichtlich regulär mit $D^{-1} \leq D_0^{-1}$. Insgesamt erhält man also die Regularität der Matrix A sowie

$$0 \leq A^{-1} = (I - S)^{-1} D^{-1} \leq (I - S_0)^{-1} D_0^{-1} = A_0^{-1},$$

was den Beweis von Teil (a) des Theorems 9.10 komplettiert. □

Bemerkung 9.19. Der vorgestellte Beweis lässt sich noch kompakter führen mithilfe der im anschließenden Kapitel behandelten Theorie der M-Matrizen (siehe insbesondere Aufgabe 10.7). △

9.3 Galerkin-Verfahren

In dem vorliegenden Abschnitt werden Galerkin-Verfahren behandelt, die bei speziellen Problemstellungen und bei Verwendung geeigneter Ansatzräume bessere Approximationseigenschaften als Differenzenverfahren besitzen.

9.3.1 Einführende Bemerkungen

Im Folgenden wird der Ansatz für Galerkin-Verfahren zur approximativen Lösung von Randwertproblemen vorgestellt. Exemplarisch soll dies zunächst anhand des speziellen sturm-liouvilleschen Randwertproblems $-u'' + ru = \varphi,\ u(a) = u(b) = 0$, mit der nichtnegativen Funktion $r : [a,b] \to \mathbb{R}_+$ geschehen[4]. Hierzu wird dieses Randwertproblem als Operatorgleichung $\mathcal{L}u = \varphi$ geschrieben mit

$$\left.\begin{aligned} \mathcal{L} : C[a,b] \supset \mathcal{D}_{\mathcal{L}} \to C[a,b], \qquad u \mapsto -u'' + ru, \\ \mathcal{D}_{\mathcal{L}} = \{u \in C^2[a,b] : u(a) = u(b) = 0\}, \end{aligned}\right\} \tag{9.22}$$

und im weiteren Verlauf bezeichne noch

$$\langle u, v \rangle_2 := \int_a^b u(x)\, v(x)\, dx, \qquad u,\ v \in C[a,b], \tag{9.23}$$

das L_2-Skalarprodukt, und $\mathcal{S} \subset \mathcal{D}_{\mathcal{L}}$ sei ein linearer Unterraum mit $\dim \mathcal{S} < \infty$. Als Raum $\mathcal{S}$ kann hier beispielsweise der Raum der kubischen Splines mit natürlichen Randbedingungen verwendet werden. In der vorliegenden speziellen Situation ist die Galerkin-Approximation $\widehat{s} \in \mathcal{S}$ folgendermaßen erklärt[5]:

$$\widehat{s} \in \mathcal{S}, \qquad \langle \mathcal{L}\widehat{s}, \psi \rangle_2 = \langle \varphi, \psi \rangle_2 \qquad \text{für alle} \quad \psi \in \mathcal{S}. \tag{9.24}$$

Interessiert ist man an der Verwendung von solchen Räumen $\mathcal{S}$, für die einerseits der Fehler $\widehat{s} - u$ bezüglich der L^2-Norm $\|\cdot\|_2$ oder anderer gängiger Normen möglichst klein ausfällt, und andererseits soll die zugehörige Galerkin-Approximation mit möglichst wenig Aufwand bestimmt werden können.

Im weiteren Verlauf werden die folgenden Themen abgehandelt:

- Galerkin-Verfahren werden in einer allgemeinen Form und für eine große Klasse von Problemstellungen definiert sowie ihre Konvergenzeigenschaften behandelt (übernächster Abschnitt 9.3.3).
- Die Bedeutung der in Abschnitt 9.3.3 erzielten Konvergenzresultate sollen anhand des sturm-liouvilleschen Differenzialoperators $\mathcal{L}u = -u'' + ru$ aus (9.22) erläutert werden. Die dafür benötigten Eigenschaften von $\mathcal{L}$ werden in dem nachfolgenden Abschnitt 9.3.2 hergeleitet.

9.3.2 Eigenschaften des Differenzialoperators $\mathcal{L}u = -u'' + ru$

Im Folgenden werden einige Eigenschaften des Differenzialoperators $\mathcal{L}u = -u'' + ru$ aus (9.22) vorgestellt. Als Erstes geht es darum, das anhand des Modellbeispiels aus (9.22) betrachtete Galerkin-Verfahren dahingehend sinnvoll zu verallgemeinern, dass eine Verwendung des Raums $\mathcal{S}$ der linearen Splinefunktionen infrage kommt, der aufgrund der fehlenden Differenzierbarkeitseigenschaften nicht in dem Definitionsbereich $\mathcal{D}_{\mathcal{L}}$ des sturm-liouvilleschen Differenzialoperators enthalten ist. Dabei ist

[4] vergleiche (9.7)–(9.8)

[5] Die konkrete Art der Berechnung wird in Abschnitt 9.3.4 behandelt.

die folgende symmetrische Bilinearform hilfreich,

$$\left.\begin{aligned} [\![u,v]\!] \;&:=\; \int_a^b (u'v' \,+\, ruv)(x)\,dx, \qquad\qquad u,\ v \in C^1_\Delta[a,b], \\ C^1_\Delta[a,b] \;&=\; \{u : [a,b] \to \mathbb{R} :\ u \text{ ist stückweise stetig differenzierbar}\}. \end{aligned}\right\} \tag{9.25}$$

Hierbei heißt eine Funktion $u : [a,b] \to \mathbb{R}$ *stückweise stetig differenzierbar*, falls sie auf dem Intervall $[a,b]$ stetig ist und eine Zerlegung $\Delta = \{a = x_0 < x_1 < \ldots < x_N = b\}$ existiert, so dass auf jedem der offenen Teilintervalle $(x_0, x_1),\ (x_1, x_2), \ldots,$ (x_{N-1}, x_N) die Ableitung der Funktion u existiert und dort eine stetige Funktion darstellt. Das Symbol Δ in $C^1_\Delta[a,b]$ bezieht sich *nicht* auf eine vorab festgelegte Zerlegung.

Die Bedeutung des in (9.25) auftretenden Integrals mit stückweise stetig differenzierbaren Funktionen u, v wird klar mit der folgenden Setzung,

$$\int_a^b u'(x)\,v'(x)\,dx \;=\; \sum_{k=1}^{M} \int_{z_{k-1}}^{z_k} u'(x)\,v'(x)\,dx, \tag{9.26}$$

wobei die Zahlen $a = z_0 < z_1 < \ldots < z_M = b$ so gewählt sind, dass die Funktion $u'v'$ auf jedem der offenen Teilintervalle $(z_0, z_1),\ (z_1, z_2), \ldots, (z_{M-1}, z_M)$ definiert und stetig ist. Wegen der fehlenden Setzung der Funktion $u'v'$ an den Stellen $z_0, \ldots, z_M$ sind die die Integrale auf der rechten Seite von (9.26) als uneigentliche Integrale zu verstehen. Entsprechend ist für stückweise stetig differenzierbare Funktionen $u : [a,b] \to \mathbb{R}$ der Wert $\|u'\|_2 = (\int_a^b u'(x)^2\,dx)^{1/2}$ zu verstehen.

Mit dem folgenden Lemma wird der Zusammenhang zwischen der angegebenen Bilinearform und dem sturm-liouvilleschen Differenzialoperator $\mathcal{L}$ beschrieben:

Lemma 9.20. *Es gilt*

$$\begin{aligned} [\![u,v]\!] \;=\; \langle \mathcal{L}u, v\rangle_2 \qquad & \textit{für}\ \ u \in \mathcal{D}_{\mathcal{L}},\quad v \in \mathcal{D}, \\ & \textit{mit}\ \ \mathcal{D} \;=\; \{u \in C^1_\Delta[a,b] \,:\, u(a) \,=\, u(b) \,=\, 0\}. \end{aligned} \tag{9.27}$$

BEWEIS. Auch für stückweise stetig differenzierbare Funktionen sind die Regeln der partiellen Integration anwendbar, und so erhält man

$$\begin{aligned} \langle \mathcal{L}u, v\rangle_2 \;&=\; \int_a^b (-u'' \,+\, ru)(x)\,v(x)dx \\ &=\; -(u'v)(x)|_a^b \,+\, \int_a^b (u'v' \,+\, ruv)(x)\,dx \\ &=\; 0 \,+\, \int_a^b (u'v' \,+\, ruv)(x)\,dx \;=\; [\![u,v]\!]. \end{aligned}$$

□

Bemerkung 9.21. Man beachte, dass der Ausdruck $[\![u,v]\!]$ auch für Funktionen $u \in \mathcal{D}\backslash\mathcal{D}_{\mathcal{L}}$ definiert ist. Aufgrund der Identität (9.27) stellt die Bilinearform $[\![\cdot,\cdot]\!]$ somit

bezüglich des ersten Eingangs eine Fortsetzung der Bilinearform $\langle \mathcal{L}\cdot, \cdot \rangle_2$ dar. Diese Eigenschaft ermöglicht die Erweiterung des in (9.24) anhand des sturm-liouvilleschen Differenzialoperators $\mathcal{L}u = -u'' + ru$ eingeführten Galerkin-Verfahrens auch auf solche Ansatzräume $\mathcal{S} \subset \mathcal{D}$, die nicht in $\mathcal{D}_{\mathcal{L}}$ enthalten sind (vergleiche Definition 9.28 unten). △

Als unmittelbare Konsequenz aus Theorem 9.20 und der Symmetrie der Bilinearform $[\![\cdot, \cdot]\!]$ erhält man die Symmetrie des sturm-liouvilleschen Differenzialoperators $\mathcal{L}$.

Korollar 9.22. *Der sturm-liouvillesche Differenzialoperator $\mathcal{L}$ in (9.22) ist symmetrisch, es gilt also*

$$\langle \mathcal{L}u, v \rangle_2 = \langle u, \mathcal{L}v \rangle_2 \qquad \text{für } u, v \in \mathcal{D}_{\mathcal{L}}.$$

BEWEIS. Die Behauptung folgt unmittelbar aus Lemma 9.20,

$$\langle \mathcal{L}u, v \rangle_2 = [\![u, v]\!] = [\![v, u]\!] = \langle \mathcal{L}v, u \rangle_2 = \langle u, \mathcal{L}v \rangle_2. \qquad \square$$

In dem nächsten Theorem werden die (später benötigte) positive Definitheit der Abbildung $u \mapsto [\![u, u]\!]$ nachgewiesen und gängige obere und untere Schranken für $[\![u, u]\!]$ hergeleitet. (Diese Schranken ermöglichen die Herleitung konkreter Fehlerabschätzungen für die Galerkin-Approximation.) Das folgende Lemma liefert hierfür die technischen Hilfsmittel.

Lemma 9.23. *Mit der Notation $\|u\|_2 = \langle u, u \rangle_2^{1/2}$ gilt die* friedrichsche Ungleichung

$$\|u\|_2 \leq (b-a)\|u'\|_2 \qquad \text{für } u \in C^1_\Delta[a,b] \text{ mit } u(a) = 0. \tag{9.28}$$

BEWEIS. Aufgrund der Eigenschaft $u(a) = 0$ gilt

$$u(x) = \int_a^x u'(t)\,dt \qquad \text{für } x \in [a,b], \tag{9.29}$$

da der Haupsatz der Differenzial- und Integralrechnung auch für stückweise stetig differenzierbare Funktionen gültig ist. Ausgehend von (9.29) liefert eine Anwendung der cauchy-schwarzschen Ungleichung die folgende Abschätzung,

$$u(x)^2 \leq \int_a^x 1^2\,dt \cdot \int_a^x u'(t)^2\,dt = (x-a)\int_a^x u'(t)^2\,dt \leq (b-a)\overbrace{\int_a^b u'(t)^2\,dt}^{= \|u'\|_2^2} \quad \text{für } x \in [a,b],$$

und die angegebene Ungleichung (9.28) resultiert nun unmittelbar aus der trivialen Abschätzung $\|v\|_2 = (\int_a^b v(s)^2\,ds)^{1/2} \leq (b-a)^{1/2}\|v\|_\infty$ für $v \in C[a,b]$. $\square$

Mithilfe des vorhergehenden Lemmas lassen sich obere und untere Schranken für $[\![u, u]\!]$ herleiten, die die Grundlage für nachfolgende konkrete Fehlerabschätzungen darstellen.

Theorem 9.24. *Es gelten die Ungleichungen*

$$\|u'\|_2^2 \;\le\; [\![u,u]\!] \;\le\; \kappa_1\|u'\|_2^2 \qquad \textit{für}\;\; u \in C^1_\Delta[a,b] \;\; \textit{mit}\;\; u(a)=0, \tag{9.30}$$

mit der Konstanten $\kappa_1 \;=\; 1 \;+\; \|r\|_\infty (b-a)^2$.

BEWEIS. Die angegebenen Ungleichungen erhält man folgendermaßen,

$$[\![u,u]\!] = \int_a^b ((u')^2 \;+\; ru^2)(s)\,ds \;\overset{(*)}{\ge}\; \int_a^b u'(s)^2\,ds \;=\; \|u'\|_2^2,$$

$$\text{—«—} = \text{————«————} \le \|u'\|_2^2 \;+\; \|r\|_\infty \|u\|_2^2 \;\overset{(**)}{\le}\; \kappa_1\|u'\|_2^2,$$

wobei die Abschätzungen $(*)$ und $(**)$ aus der Nichtnegativität $r \ge 0$ beziehungsweise der friedrichschen Ungleichung resultieren. □

Die später benötigten Eigenschaften des speziellen Differenzialoperators $\mathcal{L}u = -u'' + ru$ stehen nun allesamt zur Verfügung.

9.3.3 Galerkin-Verfahren – ein allgemeiner Ansatz

Galerkin-Verfahren lassen sich in den unterschiedlichsten Situationen einsetzen und werden hier daher in genügender Allgemeinheit betrachtet. Zunächst werden die entsprechenden Annahmen zusammengetragen.

Voraussetzungen 9.25. (a) In einem reellen Vektorraum $\mathcal{V}$ wird die lineare Gleichung

$$\mathcal{L}u \;=\; \varphi \qquad \text{mit} \;\; \mathcal{L} : \mathcal{V} \supset \mathcal{D}_{\mathcal{L}} \to \mathcal{V} \;\text{linear}\,, \qquad \varphi \in \mathcal{V}$$

betrachtet, wobei $\mathcal{D}_{\mathcal{L}}$ ein linearer Unterraum von $\mathcal{V}$ ist. Diese Gleichung $\mathcal{L}u = \varphi$ besitze eine Lösung $u_* \in \mathcal{D}_{\mathcal{L}}$. Weiter sei $\langle\cdot,\cdot\rangle : \mathcal{V}\times\mathcal{V} \to \mathbb{R}$ eine Bilinearform auf $\mathcal{V}$.

(b) Es bezeichne $[\![\cdot,\cdot]\!] : \mathcal{D}\times\mathcal{D} \to \mathbb{R}$ eine zweite Bilinearform auf einem linearen Unterraum $\mathcal{D} \subset \mathcal{V}$, wobei $\mathcal{D}$ eine Obermenge des Definitionsbereichs $\mathcal{D}_{\mathcal{L}}$ der Abbildung $\mathcal{L}$ darstellt, $\mathcal{D}_{\mathcal{L}} \subset \mathcal{D}$. Diese zweite Bilinearform $[\![\cdot,\cdot]\!]$ sei positiv definit,

$$[\![u,u]\!] > 0 \qquad \text{für} \;\; 0 \ne u \in \mathcal{D},$$

und zwischen den beiden genannten Bilinearformen bestehe der folgende Zusammenhang,

$$[\![u,v]\!] \;=\; \langle \mathcal{L}u, v\rangle \qquad \text{für} \;\; u \in \mathcal{D}_{\mathcal{L}}, \;\; v \in \mathcal{D}. \tag{9.31}$$

Beispiel 9.26. Der im vorangegangenen Abschnitt 9.3.2 betrachtete Differenzialoperator $\mathcal{L}u = -u'' + ru$ erfüllt mit den in dem dortigen Zusammenhang betrachteten Bilinearformen die in Voraussetzung 9.25 genannten Bedingungen (mit den Notationen $\mathcal{V} = C[a,b]$ und $\langle\cdot,\cdot\rangle = \langle\cdot,\cdot\rangle_2$). △

Bemerkung 9.27. (a) Unter den in Voraussetzung 9.25 genannten Bedingungen ist der Operator $\mathcal{L}$ notwendigerweise injektiv. Falls nämlich $\mathcal{L}u = 0$ erfüllt ist für eine Funktion $u \in \mathcal{D}_{\mathcal{L}}$, so gilt

$$0 \;=\; \langle \mathcal{L}u, u \rangle \;=\; [\![u, u]\!] \quad \rightsquigarrow \quad u \;=\; 0.$$

(b) Die Abbildung $\mathcal{D} \ni u \mapsto [\![u, u]\!]^{1/2}$ bezeichnet man als *Energienorm*. Tatsächlich erfüllt sie die Normeigenschaften, was offensichtlich ist im Fall einer symmetrischen Bilinearform $[\![\cdot, \cdot]\!]$, die dann ein Skalarprodukt darstellt. Man kann aber auch für den nichtsymmetrischen Fall die Normeigenschaften der Energienorm nachweisen (Aufgabe 9.10).

(c) Die Eigenschaft (9.31) dient in den nachfolgenden Betrachtungen lediglich dazu, Galerkin-Verfahren in einer relativ allgemeinen Form zu erklären. Es existiert jedoch ein weiterer Anwendungsbereich, der hier kurz angesprochen werden soll. Aufgrund der Eigenschaft (9.31) stellt die Lösung $u_* \in \mathcal{D}_{\mathcal{L}}$ der Operatorgleichung $\mathcal{L}u = \varphi$ auch eine Lösung der *Variationsgleichung*

$$\text{finde } u \in \mathcal{D} \quad \text{mit} \quad [\![u, v]\!] \;=\; \langle \varphi, v \rangle \qquad \text{für alle} \quad v \in \mathcal{D} \tag{9.32}$$

dar. Diese Variationsgleichung (9.32) erlangt in denjenigen Anwendungen eine eigenständige Bedeutung, bei denen die Gleichung $\mathcal{L}u = \varphi$ entgegen der Voraussetzung 9.25 nicht in $\mathcal{D}$ lösbar ist, die Variationsgleichung (9.32) jedoch eine Lösung $u_* \in \mathcal{D}$ besitzt. Solche Lösungen bezeichnet man dann als *verallgemeinerte* oder *schwache Lösung* von $\mathcal{L}u = \varphi$. Die nachfolgenden Resultate gelten auch für schwache Lösungen. △

Definition 9.28. Es seien die in Voraussetzung 9.25 genannten Bedingungen erfüllt. Zur approximativen Lösung der Gleichung $\mathcal{L}u = \varphi$ ist für einen gegebenen linearen Unterraum $\mathcal{S} \subset \mathcal{D}$ mit $\dim \mathcal{S} < \infty$ die *Galerkin-Approximation* $\widehat{s} \in \mathcal{S}$ wie folgt erklärt,

$$\widehat{s} \in \mathcal{S}, \qquad [\![\widehat{s}, \psi]\!] \;=\; \langle \varphi, \psi \rangle \qquad \text{für alle} \quad \psi \in \mathcal{S}. \tag{9.33}$$

Dieses Verfahren wird als *Galerkin-Verfahren* beziehungsweise im Falle der Symmetrie der Bilinearform $[\![\cdot, \cdot]\!]$ auch als *Ritz-Verfahren* bezeichnet.

Bemerkung 9.29. (a) Wenn $\mathcal{S} \subset \mathcal{D}_{\mathcal{L}}$ gilt, so kann man (9.33) in der folgenden klassischen und der (aus dem in (9.24) angegebenen Beispiel) bereits bekannten Form schreiben,

$$\widehat{s} \in \mathcal{S}, \qquad \langle \mathcal{L}\widehat{s}, \psi \rangle \;=\; \langle \varphi, \psi \rangle \qquad \text{für alle} \quad \psi \in \mathcal{S}.$$

(b) Die Galerkin-Approximation ist eindeutig bestimmt. Sind nämlich $\widehat{s}, s \in \mathcal{S}$ zwei Galerkin-Approximationen, so gilt insbesondere $\widehat{s} - s \in \mathcal{S}$ und dann $[\![\widehat{s} - s, \widehat{s} - s]\!] = 0$, so dass aufgrund von Teil (b) der Annahme 9.25 notwendigerweise $\widehat{s} = s$ gilt.

(c) Wenn $u_* \in \mathcal{D}_{\mathcal{L}}$ die Lösung der Gleichung $\mathcal{L}u = \varphi$ bezeichnet, so gilt für jedes Element $\widehat{s} \in \mathcal{S}$:

$$\widehat{s} \text{ ist Galerkin-Approximation} \iff [\widehat{s} - u_*, \psi] = 0 \quad \text{für alle} \quad \psi \in \mathcal{S}. \tag{9.34}$$

Dies folgt unmittelbar aus den Darstellungen (9.32) und (9.33).

(d) Allgemeiner als in (9.33) kann man für lineare Räume $\mathcal{S}_1 \subset \mathcal{D}$ und $\mathcal{S}_2 \subset \mathcal{V}$ mit $\dim \mathcal{S}_1 = \dim \mathcal{S}_2 < \infty$ Approximationen $\widehat{s} \in \mathcal{S}_1$ von der folgenden Form betrachten,

$$\widehat{s} \in \mathcal{S}_1, \qquad [\![\widehat{s}, \psi]\!] = \langle \varphi, \psi \rangle \quad \text{für} \quad \psi \in \mathcal{S}_2. \tag{9.35}$$

In diesem Zusammenhang wird $\mathcal{S}_1$ als *Ansatzraum* und $\mathcal{S}_2$ als *Testraum* bezeichnet. Bei Galerkin-Verfahren stimmen demnach Ansatz- und Testraum überein. △

Die folgende Minimaleigenschaft der Galerkin-Approximation bildet die Grundlage für die Herleitung konkreter Fehlerabschätzungen bei Galerkin-Verfahren. Man beachte, dass hier die Symmetrie der Bilinearform $[\![\cdot,\cdot]\!]$ benötigt wird.

Theorem 9.30. *Es seien die in Voraussetzung 9.25 genannten Bedingungen erfüllt, und zusätzlich sei die Bilinearform $[\![\cdot,\cdot]\!] : \mathcal{D} \times \mathcal{D} \to \mathbb{R}$ symmetrisch. Dann minimiert die Galerkin-Approximation $\widehat{s} \in \mathcal{S}$ in dem Raum $\mathcal{S} \subset \mathcal{D}$ den Fehler bezüglich der Energienorm, es gilt also*

$$[\widehat{s} - u_*, \widehat{s} - u_*] = \min_{s \in \mathcal{S}} [s - u_*, s - u_*]. \tag{9.36}$$

BEWEIS. Die Aussage erhält man durch folgende Rechnung, bei der $s \in \mathcal{S}$ beliebig gewählt ist,

$$\begin{aligned}
&[\widehat{s} - u_*, \widehat{s} - u_*] \\
&\quad = [s - u_*, \widehat{s} - u_*] + \overbrace{[\widehat{s} - s, \widehat{s} - u_*]}^{= 0 \text{ nach } (9.34)} \\
&\quad = [s - u_*, s - u_*] + [s - u_*, \widehat{s} - s] \\
&\quad = \text{———«———} - \underbrace{[\widehat{s} - s, \widehat{s} - s]}_{\geq 0} + \underbrace{[\widehat{s} - u_*, \widehat{s} - s]}_{= 0} \\
&\quad \leq \text{———«———} .
\end{aligned}$$

□

Die in Theorem 9.30 vorgestellte Minimaleigenschaft der Galerkin-Approximation bezüglich der Energienorm ist ein erster Schritt zur Herleitung konkreter Fehlerabschätzungen für das Galerkin-Verfahren. Ausgangspunkt weiterer Fehlerabschätzungen ist das folgende triviale Resultat, das man in den Anwendungen typischerweise mit speziellen Normen $|\!|\!| \cdot |\!|\!| : \mathcal{D} \to \mathbb{R}_+$ einsetzt.

Theorem 9.31. *Es seien die in Voraussetzung 9.25 genannten Bedingungen erfüllt mit einer symmetrischen Bilinearform $[\![\cdot,\cdot]\!]$, und bezüglich einer nichtnegativen Abbildung $|||\cdot||| : \mathcal{D} \to \mathbb{R}_+$ gelte*

$$c_1 |||u|||^2 \;\le\; [\![u,u]\!] \;\le\; c_2 |||u|||^2 \qquad \textit{für alle} \quad u \in \mathcal{D} \tag{9.37}$$

mit gewissen Konstanten $c_2 \ge c_1 > 0$. Dann gilt

$$|||\widehat{s} - u_*||| \;\le\; c \min_{s\in\mathcal{S}} |||s - u_*||| \qquad \textit{mit} \quad c \;=\; \sqrt{\frac{c_2}{c_1}}. \tag{9.38}$$

BEWEIS. Die Aussage folgt unmittelbar aus der Eigenschaft (9.36). □

In der Situation (9.38) nennt man das Galerkin-Verfahren *quasioptimal* bezüglich $|||\cdot|||$, da die Galerkin-Approximation bis auf einen konstanten Faktor aus dem Raum $\mathcal{S}$ die optimale Approximation an u_* darstellt.

Auch für *nichtsymmetrische* Bilinearformen $[\![\cdot,\cdot]\!]$ erhält man unter vergleichbaren Bedingungen die Quasioptimalität der Galerkin-Approximation.

Theorem 9.32 (Lemma von Céa). *Es seien die in Voraussetzung 9.25 genannten Bedingungen erfüllt und bezüglich einer Abbildung $|||\cdot||| : \mathcal{D} \to \mathbb{R}_+$ gelte*

$$c_1 |||u|||^2 \;\le\; [\![u,u]\!] \quad \textit{für } u \in \mathcal{D}, \qquad [\![u,v]\!] \;\le\; c_2 |||u||| |||v||| \quad \textit{für } u,\, v \in \mathcal{D} \tag{9.39}$$

mit gewissen Konstanten $c_2 \ge c_1 > 0$. Dann gilt $|||\widehat{s} - u_*||| \le c \min_{s\in\mathcal{S}} |||s - u_*|||$ mit $c = c_2/c_1$, das Galerkin-Verfahren ist also quasioptimal bezüglich $|||\cdot|||$.

BEWEIS. Die Aussage erhält man durch folgende Rechnung, bei der $s \in \mathcal{S}$ beliebig gewählt ist,

$$\begin{aligned} c_1 |||\widehat{s} - u_*|||^2 &\overset{(*)}{\le} [\![\widehat{s} - u_*,\, \widehat{s} - u_*]\!] \\ &= [\![\widehat{s} - u_*,\, s - u_*]\!] + \underbrace{[\![\widehat{s} - u_*,\, \widehat{s} - s]\!]}_{=\,0} \\ &\overset{(**)}{\le} c_2 |||\widehat{s} - u_*||| \, |||s - u_*|||, \end{aligned}$$

wobei man die Abschätzungen $(*)$ und $(**)$ jeweils unmittelbar aus den Bedingungen in (9.39) erhält. Eine Division durch $|||\widehat{s} - u_*|||$ liefert nun (im Fall $|||\widehat{s} - u_*||| \neq 0$, andernfalls ist die Aussage sowieso trivial) die Quasioptimalität. □

Bemerkung 9.33. Typischerweise ist in Theorem 9.32 die Abbildung $|||\cdot|||$ eine Norm, und die erste der beiden Bedingungen in (9.39) wird dann als *Koerzivität* der Bilinearform $[\![\cdot,\cdot]\!]$ bezüglich $|||\cdot|||$ bezeichnet. Die zweite Bedingung in (9.39) stellt eine Beschränktheitsbedingung an die Bilinearform $[\![\cdot,\cdot]\!]$ dar. △

9.3.4 Systemmatrix

Zur konkreten Berechnung der Galerkin-Approximation benötigt man noch eine Basis für den Raum $\mathcal{S}$:

Lemma 9.34. *Es seien die in Voraussetzung 9.25 genannten Bedingungen erfüllt und das System $s_1, \ldots, s_N \in \mathcal{S}$ bilde eine Basis von $\mathcal{S}$. Es ist das Element $s = \sum_{k=1}^{N} c_k s_k \in \mathcal{S}$ mit den Koeffizienten $c_1, \ldots, c_N \in \mathbb{R}$ genau dann Galerkin-Approximation, wenn die Koeffizienten $c_1, \ldots, c_N \in \mathbb{R}$ dem folgenden linearen Gleichungssystem genügen,*

$$\begin{pmatrix} [\![s_1, s_1]\!] & \ldots & [\![s_N, s_1]\!] \\ \vdots & \ddots & \vdots \\ [\![s_1, s_N]\!] & \ldots & [\![s_N, s_N]\!] \end{pmatrix} \begin{pmatrix} c_1 \\ \vdots \\ c_N \end{pmatrix} = \begin{pmatrix} \langle \varphi, s_1 \rangle \\ \vdots \\ \langle \varphi, s_N \rangle \end{pmatrix}. \tag{9.40}$$

BEWEIS. Nach Definition (9.33) ist mit der gegebenen Basis von $\mathcal{S}$ ein Element $s \in \mathcal{S}$ genau dann Galerkin-Approximation, wenn $s \in \mathcal{S}$ und $[\![s, s_j]\!] = \langle \varphi, s_j \rangle$ für $j = 1, 2, \ldots, N$ gilt. Mit dem Ansatz $s = \sum_{k=1}^{N} c_k s_k \in \mathcal{S}$ ist dies gleichbedeutend mit

$$\sum_{k=1}^{N} [\![s_k, s_j]\!] c_k \;=\; \langle \varphi, s_j \rangle, \qquad j \;=\; 1, 2, \ldots, N.$$

Die Matrixversion hierzu ist identisch mit (9.40). □

Bemerkung 9.35. (a) Die in (9.40) auftretende Matrix wird als *Systemmatrix* oder auch als *Steifigkeitsmatrix* bezeichnet und ist regulär aufgrund der Eindeutigkeit der Galerkin-Approximation (siehe Teil (b) von Bemerkung 9.29). Daraus erhält man auch unmittelbar die Existenz der Galerkin-Approximation.

(b) Das Gleichungssystem (9.40) stellt lediglich eine "Halbdiskretsierung" der gegebenen Operatorgleichung $\mathcal{L}u = \varphi$ dar, denn sowohl die Einträge in der Systemmatrix als auch die Komponenten des Vektors auf der rechten Seite des Gleichungssystems sind in der Regel nicht exakt bekannt und müssen numerisch berechnet werden. Im Fall der beiden speziellen Bilinearformen aus Voraussetzung 9.25 kann dies beispielsweise mittels Quadraturformeln geschehen.

Allgemein bezeichnet man solche Verfahren, bei denen die Einträge in der Systemmatrix beziehungsweise der rechten Seite des Gleichungssystems (9.40) durch exakt auswertbare Näherungsformeln approximiert werden, als *volldiskrete Galerkin-Verfahren.* △

9.3.5 Finite-Elemente-Methode

In der Praxis ist der zugrunde liegende Raum $\mathcal{V}$ typischerweise ein Funktionenraum und man verwendet als Basis des zum Galerkin-Verfahren gehörenden Raums $\mathcal{S}$ oft Funktionen $s_1, \ldots, s_N \in \mathcal{S}$ mit einem jeweils kleinen Träger, es gilt also $s_k = 0$ außerhalb einer vom jeweiligen Index k abhängenden Menge und $s_k \cdot s_j = 0$ für einen Großteil der Indizes. In diesem Fall wird das zugehöriges Galerkin-Verfahren auch als *Finite-Elemente-Methode* bezeichnet.

Beispiel 9.36. Zu der Zerlegung $\Delta = \{a = x_0 < x_1 < \ldots < x_N = b\}$ eines Intervalls $[a, b]$ sei $\mathcal{S}$ der Raum der linearen Splines, $\mathcal{S} = S_{\Delta,1}$. Eine Basis dieses $(N+1)$-dimensionalen Vektorraums erhält man durch *Hutfunktionen* (lineare B-Splines), die folgendermaßen erklärt sind,

$$s_j(x) = \left\{ \begin{array}{ll} \frac{1}{h_{j-1}}(x - x_{j-1}), & \text{falls} \quad x \in [x_{j-1}, x_j], \\ \frac{1}{h_j}(x_{j+1} - x), & \text{falls} \quad x \in [x_j, x_{j+1}], \\ 0 & \text{sonst} \end{array} \right\} \quad j = 0, 1, \ldots, N, \qquad (9.41)$$

wobei $h_j = x_{j+1} - x_j$, $j = 0, 1, \ldots, N-1$ die Knotenabstände bezeichnet. In (9.41) sind in den Fällen "$j = 0$" beziehungsweise "$j = N$" die Situationen "$x \in [x_{-1}, x_0]$" beziehungsweise "$x \in [x_N, x_{N+1}]$" ohne Relevanz. Die vorliegende Situation ist in Bild 9.1 veranschaulicht.

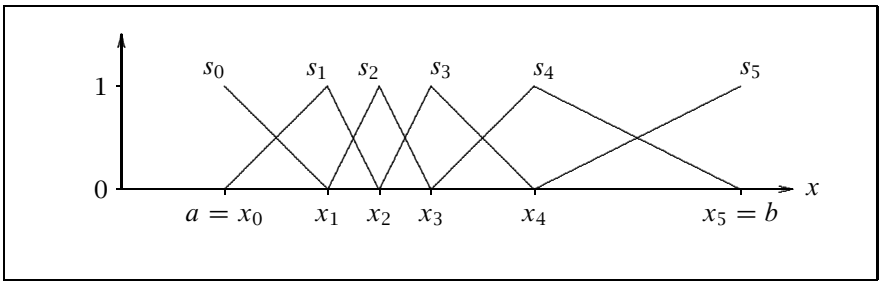


Bild 9.1: Darstellung der Hutfunktionen an einem Beispiel

Für das Referenzbeispiel (9.22) mit den homogenen Randbedingungen verwendet man sinnvollerweise Räume $\mathcal{S}$ mit in den Randpunkten a und b verschwindenden Funktionen, beispielsweise also den Raum der linearen Splines $S_{\Delta,1}$ mit Nullrandbedingungen, $\mathcal{S} = \{s \in S_{\Delta,1} : s(a) = s(b) = 0\}$. Eine Basis dieses $(N-1)$-dimensionalen Vektorraums bilden die Hutfunktionen $s_1, \ldots, s_{N-1}$. △

Beispiel 9.37. Mit der Notation $x_j = a + jh \in \mathbb{R}$ für $j = -3, -2, \ldots, N+3$ mit $h = (b-a)/N$ sei $\mathcal{S}$ der Raum der kubischen Splines zur äquidistanten Zerlegung $\Delta = \{a = x_0 < x_1 < \ldots < x_N = b\}$ des Intervalls $[a, b]$. Eine Basis dieses $(N+3)$-dimensionalen Vektorraums $\mathcal{S} = S_{\Delta,3}$ erhält man beispielsweise, indem man hilfsweise auf dem Intervall $[x_{-3}, x_{N+3}]$ und zur Zerlegung $\widehat{\Delta} =$

$\{x_{-3} < x_{-2} < \ldots < x_{N+3}\}$ die eindeutig bestimmten kubischen Splinefunktionen $s_{-1}, s_0, \ldots, s_N, s_{N+1} \in S_{\widehat{\Delta},3}$ mit natürlichen Randbedingungen und den Funktionswerten $s_j(x_j) = 2/3$, $s_j(x_{j\pm1}) = 1/6$ und $s_j(x_\ell) = 0$ in den restlichen Knoten heranzieht. Bei diesen Funktionen handelt es sich um spezielle kubische *B-Splines*, deren explizite Form beispielsweise in Oevel [78] angegeben ist. Durch Einschränkung der Definitionsbereiche dieser B-Splines auf das Intervall $[a, b]$ erhält man eine Familie von Funktionen, die eine Basis von $\mathcal{S} = S_{\Delta,3}$ darstellt. Die vorliegende Situation ist in Bild 9.2 veranschaulicht.

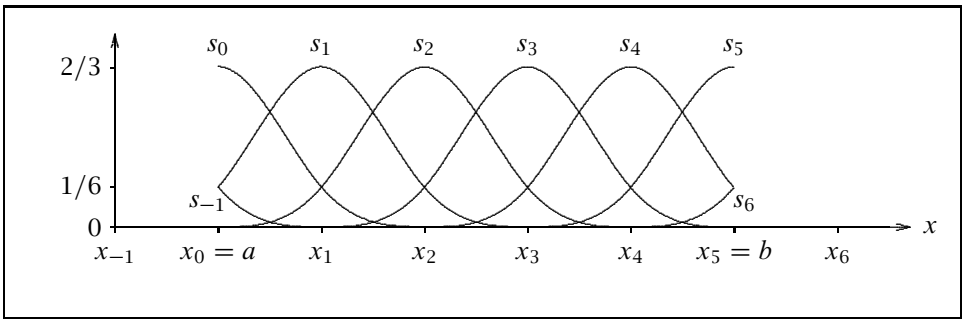


Bild 9.2: Darstellung von kubischen B-Splines anhand eines Beispiels ($N = 5$)

Ist bei Verwendung der Finite-Elemente-Methode der zugrunde liegende Operator $\mathcal{L}$ ein Differenzialoperator, so besitzt die zugehörige Systemmatrix bei richtiger Anordnung der Basiselemente typischerweise eine Bandstruktur, so dass sich das entsprechende Gleichungssystem (9.40) mit verhältnismäßig geringem Aufwand lösen lässt. Die Situation wird im nachfolgenden Abschnitt verdeutlicht.

9.3.6 Anwendungen

Im Folgenden wird nun wieder das spezielle sturm-liouvillesche Randwertproblem aus Abschnitt 9.3.1 betrachtet:

> Es bezeichne $\mathcal{L} : C[a, b] \supset \mathcal{D}_\mathcal{L} \to C[a, b]$ den speziellen Differenzialoperator aus (9.22). Weiter bezeichnet $\langle \cdot, \cdot \rangle_2$ das L^2-Skalarprodukt (siehe (9.23)), und $[\![\cdot, \cdot]\!] : C^1_\Delta[a, b] \times C^1_\Delta[a, b] \to \mathbb{R}$ sei die Bilinearform (9.25). Die Gleichung $\mathcal{L}u = \varphi$ besitze eine Lösung $u_* \in \mathcal{D}_\mathcal{L}$. (9.42)

Ausgehend von der in (9.42) beschriebenen Situation werden nun die Approximationseigenschaften des Galerkin-Verfahrens bezüglich spezieller Ansatzräume $\mathcal{S}$ vorgestellt. Vorbereitend wird die folgende allgemeine Abschätzung festgehalten, die eine unmittelbare Konsequenz aus den bereits gewonnenen Resultaten ist.

Korollar 9.38. *Ausgehend von der in (9.42) beschriebenen Situation sei zu einem vorgegebenen Ansatzraum $\mathcal{S} \subset \mathcal{D} = \{u \in C^1_\Delta[a, b] : u(a) = u(b) = 0\}$ die zugehörige*

Galerkin-Approximation mit $\widehat{s} \in \mathcal{S}$ bezeichnet. Hier gilt die folgende Fehlerabschätzung,

$$\|\widehat{s}' - u_*'\|_2 \;\leq\; \kappa \min_{s\in\mathcal{S}} \|s' - u_*'\|_2 \tag{9.43}$$

mit $\kappa = (1 + \|r\|_\infty (b-a)^2)^{1/2}$.

BEWEIS. Die Aussage folgt unmittelbar aus den Theoremen 9.24, 9.30 und 9.31. □

Im Folgenden werden für $\mathcal{S}$ lineare beziehungsweise kubische Splineräume mit Nullrandbedingungen herangezogen. Für die Abschätzung der rechten Seite von (9.43) lassen sich in dieser Situation die bereits bekannten Schranken für den jeweils bei der Interpolation auftretenden Fehler verwenden.

Korollar 9.39. *Zu einer gegebenen Zerlegung $\Delta = \{a = x_0 < x_1 < \ldots < x_N = b\}$ bezeichne $\mathcal{S}$ den Raum der linearen Splinefunktion mit Nullrandbedingungen,*

$$\mathcal{S} \;=\; \{s \in S_{\Delta,1} \;:\; s(a) \;=\; s(b) \;=\; 0\}. \tag{9.44}$$

Mit den Notationen aus (9.42) gilt für die zugehörige Galerkin-Approximation $\widehat{s} \in \mathcal{S}$ die folgende Abschätzung,

$$\|\widehat{s}' - u_*'\|_2 \;\leq\; c h_{\max} \|u_*''\|_\infty$$

mit einer Konstanten $c \geq 0$, wobei $u_* \in C^2[a,b]$ angenommen wird.

BEWEIS. Dieses Resultat erhält man als unmittelbare Konsequenz aus Korollar 9.38 unter Berücksichtigung von Aufgabe 2.7. □

Bemerkung 9.40. In der Situation von Korollar 9.39 ist man auch an Abschätzungen für den Fehler $\widehat{s} - u_*$ interessiert, die aber mit den in diesem Abschnitt hergeleiteten Techniken nicht mit der optimalen Ordnung hergeleitet werden können. Mit einer etwas genaueren Wahl der zugrunde liegenden Räume und mit einer verfeinerten Technik (die als *Dualitäts-* oder *Aubin-Nitsche-Trick* bezeichnet wird) lässt sich aber für das Galerkin-Verfahren mit dem Ansatzraum aus (9.44) zur Lösung des sturm-liouvilleschen Randwertproblems mit homogenen Randbedingungen (9.7)–(9.8) die Abschätzung $\|\widehat{s} - u_*\|_2 = \mathcal{O}(h_{\max}^2)$ nachweisen. △

In der vorliegenden Situation (9.42), (9.44) mit den Hutfunktionen $s_1, \ldots, s_{N-1}$ (siehe Beispiel 9.36) als Basis von $\mathcal{S}$ soll noch die zugehörige Systemmatrix betrachtet werden. Wegen $s_k s_j = 0$ für $|k-j| \geq 2$ gilt auch

$$[\![s_k, s_j]\!] \;=\; 0 \qquad \text{für} \quad |k-j| \geq 2,$$

so dass die zugehörige Systemmatrix eine Tridiagonalmatrix darstellt, deren Einträge folgendes Aussehen besitzen:

$$\begin{aligned}
[\![s_j, s_{j-1}]\!] &= [\![s_{j-1}, s_j]\!] \\
&= -\frac{1}{h_{j-1}} - \frac{1}{h_{j-1}^2} \int_{x_{j-1}}^{x_j} (x - x_{j-1})(x_j - x) r(x)\, dx, \quad j = 2, 3, \ldots, N-1, \\
[\![s_j, s_j]\!] &= \frac{1}{h_{j-1}} + \frac{1}{h_{j-1}^2} \int_{x_{j-1}}^{x_j} (x - x_{j-1})^2 r(x)\, dx + \frac{1}{h_j} \\
&\qquad + \frac{1}{h_j^2} \int_{x_j}^{x_{j+1}} (x_{j+1} - x)^2 r(x)\, dx, \qquad j = 1, 2, \ldots, N-1,
\end{aligned}$$

mit $h_j = x_{j+1} - x_j$ für $j = 0, 1, \ldots, N-1$.

Beispiel 9.41. Für die spezielle Situation (9.22)–(9.25) werde zu der Zerlegung $\Delta = \{a = x_0 < x_1 < \ldots < x_N = b\}$ der Raum $\mathcal{S}$ der kubischen Splines mit Nullrandbedingungen betrachtet,

$$\mathcal{S} = \{s \in S_{\Delta,3} : s(a) = s(b) = 0\}.$$

Mit der Notation $h_j = x_{j+1} - x_j$ für $j = 0, 1, \ldots, N-1$ sei die Uniformitätsbedingung

$$\max_{j=0,\ldots,N-1} h_j \leq K \min_{j=0,\ldots,N-1} h_j$$

erfüllt mit einer Konstanten $K \geq 0$ von moderater Größe. Dann gilt für die zugehörige Galerkin-Approximation $\widehat{s} \in \mathcal{S}$ die folgende Abschätzung,

$$\|\widehat{s}' - u_*'\|_2 \leq c h_{\max}^3 \|u_*^{(4)}\|_\infty \qquad (h_{\max} := \max_{j=0,\ldots,N-1} h_j),$$

mit der Konstanten $c = (1 + \|r\|_\infty (b-a))^{1/2} 2K$, wobei $u_* \in C^4[a,b]$ und $u''(a) = u''(b) = 0$ vorausgesetzt wird. Dieses Resultat ist eine unmittelbare Konsequenz aus Korollar 9.38 und Theorem 2.16, wobei man in (9.43) den die Funktion u_* interpolierenden kubischen Spline s mit natürlichen Randbedingungen betrachtet. △

Bemerkung 9.42. Auch in der Situation von Beispiel 9.41 ist man an Abschätzungen für den Fehler $\widehat{s} - u_*$ interessiert. Unter leicht modifizierten Bedingungen lässt sich auch hier mit dem bereits angesprochenen Aubin-Nitsche-Trick die Abschätzung $\|\widehat{s} - u_*\|_2 = \mathcal{O}(h_{\max}^4)$ nachweisen. △

9.3.7 Das Energiefunktional

Als Ergänzung zu der in der Voraussetzung 9.25 beschriebenen allgemeinen Situation wird im Folgenden das Energiefunktional vorgestellt, mit dem sich einerseits die Lösung der Gleichung $\mathcal{L}u = \varphi$ und andererseits die zugehörige Galerkin-Approximation charakterisieren lassen.

Definition 9.43. In der Situation von Voraussetzung 9.25 ist das zugehörige *Energiefunktional* $\mathcal{J} : \mathcal{D} \to \mathbb{R}$ folgendermaßen erklärt,

$$\mathcal{J}(u) \;=\; \tfrac{1}{2}[\![u,u]\!] - \langle u,\varphi\rangle \qquad \text{für} \quad u \in \mathcal{D}.$$

Das folgende Theorem zeigt, dass sich der Wert des Energiefunktionals nur um eine Konstante von dem Fehler in der Energienorm unterscheidet.

Theorem 9.44. *Es seien die in Voraussetzung 9.25 genannten Bedingungen erfüllt mit einer symmetrischen Bilinearform* $[\![\cdot,\cdot]\!]$*. Dann gilt*

$$\mathcal{J}(u) \;=\; \tfrac{1}{2}\big([\![u-u_*,u-u_*]\!] - [\![u_*,u_*]\!]\big) \qquad \textit{für} \quad u \in \mathcal{D},$$

wobei wieder $u_* \in \mathcal{D}_{\mathcal{L}}$ *die Lösung der Gleichung* $\mathcal{L}u = \varphi$ *bezeichnet.*

BEWEIS. Man erhält die Aussage des Theorems durch folgende Rechnung,

$$\begin{aligned} 2\mathcal{J}(u) \;&=\; [\![u,u]\!] - 2\langle u,\varphi\rangle \;=\; [\![u,u]\!] - 2\langle u,\mathcal{L}u_*\rangle \;=\; [\![u,u]\!] - 2[\![u,u_*]\!] \\ &=\; \big([\![u,u]\!] - 2[\![u,u_*]\!] + [\![u_*,u_*]\!]\big) - [\![u_*,u_*]\!] \\ &=\; [u-u_*,\,u-u_*] - [\![u_*,u_*]\!], \qquad\qquad u \in \mathcal{D}. \end{aligned}$$

□

Als unmittelbare Konsequenz der Theoreme 9.30 und 9.44 erhält man die folgende Minimaleigenschaft.

Korollar 9.45. *In der Situation von Theorem 9.44 gilt*

$$\begin{aligned} \mathcal{J}(u_*) \;&=\; \min_{u\in\mathcal{D}} \mathcal{J}(u) \;=\; -\tfrac{1}{2}[\![u_*,u_*]\!], \\ \mathcal{J}(\widehat{s}) \;&=\; \min_{s\in\mathcal{S}} \mathcal{J}(s), \end{aligned}$$

wobei $\widehat{s} \in \mathcal{S}$ *die Galerkin-Approximation zu einem gegebenem Ansatzraum* $\mathcal{S}$ *bezeichnet.*

Bemerkung 9.46. Die Ergebnisse in Theorem 9.44 und Korollar 9.45 behalten ihre Gültigkeit für den Fall, dass die Gleichung $\mathcal{L}u = \varphi$ entgegen der Annahme 9.25 nicht in $\mathcal{D}_{\mathcal{L}}$ lösbar ist, jedoch eine verallgemeinerte Lösung $u_* \in \mathcal{D}$ existiert. Demnach ist ein Element $u \in \mathcal{D}$ genau dann verallgemeinerte Lösung der Gleichung $\mathcal{L}u = \varphi$, wenn es das Energiefunktional minimiert. △

9.4 Einfachschießverfahren

Eine weitere Möglichkeit zur Lösung von Randwertproblemen bei gewöhnlichen Differenzialgleichungen bietet das im Folgenden vorgestellte Einfachschießverfahren, das anhand des allgemeinen Randwertproblems $u'' = f(x, u, u')$, $u(a) = \alpha$, $u(b) = \beta$ betrachtet wird[6]. Im Folgenden wird ohne weitere Spezifikation an die Funktion f beziehungsweise an die Randbedingungen angenommen, dass für das vorliegende Randwertproblem eine eindeutig bestimmte Lösung $u : [a, b] \to \mathbb{R}$ existiert.

Ausgangspunkt des Einfachschießverfahrens ist die Betrachtung korrespondierender Anfangswertprobleme für die vorliegende gewöhnliche Differenzialgleichung 2. Ordnung,

$$u'' = f(x, u, u'), \qquad x \in [a, b], \tag{9.45}$$

$$u(a) = \alpha, \qquad u'(a) = s, \tag{9.46}$$

deren Lösung für jede Zahl $s \in \mathbb{R}$ existiere und mit

$$u(\cdot, s) : [a, b] \to \mathbb{R} \tag{9.47}$$

bezeichnet wird. Dabei ist $s = s_* \in \mathbb{R}$ so zu bestimmen, dass $u(b, s_*) = \beta$ gilt und damit die Funktion $u(\cdot, s_*) : [a, b] \to \mathbb{R}$ die Lösung des vorgegebenen Randwertproblems $u'' = f(x, u, u')$, $u(a) = \alpha$, $u(b) = \beta$ darstellt, also $u(\cdot, s_*) = u(\cdot)$ auf dem Intervall $[a, b]$ erfüllt ist. Diese Bestimmung von s_* erfolgt typischerweise iterativ, was die Bezeichnung *Einfachschießverfahren* begründet und in Bild 9.3 illustriert ist.

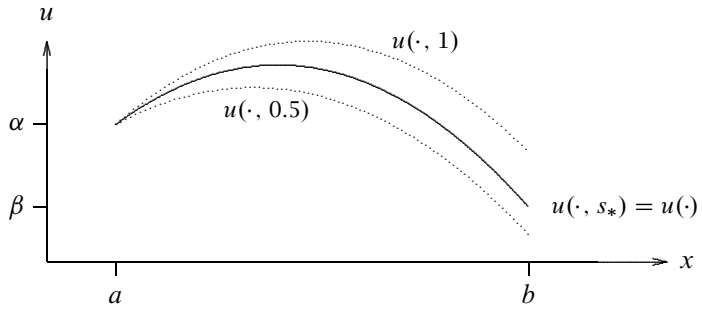


Bild 9.3: Veranschaulichung der Situation beim Einfachschießverfahren

Die nach dem vorliegenden Ansatz entstandene Problemstellung ist äquivalent zu einer Bestimmung der (eindeutig bestimmten) Nullstelle $s_* \in \mathbb{R}$ der nichtlinearen Funktion

$$F(s) := u(b, s) - \beta, \qquad s \in \mathbb{R}. \tag{9.48}$$

Zur näherungsweisen Lösung dieses Nullstellenproblems lassen sich die in Kapitel 5 vorgestellten Iterationsverfahren einsetzen, von denen im Folgenden zwei Verfahren genauer betrachtet werden.

[6] vergleiche (9.1) – (9.2) auf Seite 235

9.4.1 Numerische Realisierung des Einfachschießverfahrens mit dem Newton-Verfahren

Eine Möglichkeit zur numerischen Realisierung des Einfachschießverfahrens besteht in der Anwendung des Newton-Verfahrens,

$$s_{n+1} \;=\; s_n \;-\; \frac{F(s_n)}{F'(s_n)}, \qquad n \;=\; 0, 1, \ldots \,. \tag{9.49}$$

Dabei sind in jedem Schritt des Newton-Verfahrens (9.49) zum einen eine Auswertung der Funktion F und damit das Lösen eines Anfangswertproblems der Form (9.45)–(9.46) erforderlich, was wiederum numerisch mit einem der in den Kapiteln 7 und 8 vorgestellten Ein- beziehungsweise Mehrschrittverfahren geschieht.

Des Weiteren fällt in jedem Schritt des Newton-Verfahrens (9.49) eine Auswertung der Ableitung

$$F'(s) \;=\; \frac{\partial u}{\partial s}(b\,,\,s), \qquad s \in \mathbb{R},$$

an. An jeder Stelle s erhält man eine solche Ableitung $F'(s)$ als die Lösung eines Anfangswertproblems für eine (von s abhängende) gewöhnliche Differenzialgleichung 2. Ordnung:

Lemma 9.47. *Bei hinreichend guten Differenzierbarkeitseigenschaften der beteiligten Funktionen stellt für jeden Wert $s \in \mathbb{R}$ die Funktion*

$$v \;:=\; \frac{\partial u}{\partial s}(\cdot, s) \,:\, [a, b] \;\to\; \mathbb{R}$$

die Lösung eines Anfangswertproblems für eine spezielle lineare gewöhnliche Differenzialgleichung 2. Ordnung dar,

$$\begin{aligned} v''(x) \;&=\; g_1(x,s)\,v(x) \;+\; g_2(x,s)\,v'(x), \qquad x \;\in\; [a,b], \\ &\qquad\qquad v(a) \;=\; 0, \quad v'(a) \;=\; 1. \end{aligned} \tag{9.50}$$

Die spezielle Form der Funktionen $g_1(\cdot, s),\ g_2(\cdot, s) : [a, b] \to \mathbb{R}$ ist im Beweis angegeben.

BEWEIS. Die Aussage erhält man unter Anwendung der Kettenregel,

$$\begin{aligned} v''(x) \;&=\; \frac{\partial^3 u}{\partial s\,\partial x^2}(x\,,\,s) \;=\; \frac{d}{ds} f\Big(x, u(x\,,\,s), \frac{\partial u}{\partial x}(x\,,\,s)\Big) \\ &= \underbrace{\frac{\partial f}{\partial u}\Big(x, u(x\,,\,s), \frac{\partial u}{\partial x}(x\,,\,s)\Big)}_{=:\; g_1(x,s)}\, v(x) \;+\; \underbrace{\frac{\partial f}{\partial u'}\Big(x, u(x\,,\,s), \frac{\partial u}{\partial x}(x\,,\,s)\Big)}_{=:\; g_2(x,s)}\, v'(x), \quad x \in [a,b], \end{aligned}$$

beziehungsweise

$$u(a\,,\,\cdot) \equiv \alpha \quad\rightsquigarrow\quad v(a) \;=\; 0, \qquad\qquad \frac{\partial u}{\partial x}(a\,,\,s) \;=\; s \quad\rightsquigarrow\quad v'(a) \;=\; 1.$$

□

Zu beachten ist noch, dass die im Anschluss von (9.49) beschriebene Anwendung spezieller Ein- oder Mehrschrittverfahren zur numerischen Berechnung von $F(s)$ gleichzeitig Approximationen für die Funktionen $u(\cdot, s)$ und $\frac{\partial u}{\partial x}(\cdot, s)$ auf einem Gitter $a = x_0 < x_1 < \ldots < x_m = b$ liefert. Damit sind auch die Werte der Funktionen $g_1(\cdot, s)$ und $g_2(\cdot, s)$ an den genannten Gitterpunkten näherungsweise bekannt, was die approximative Lösung des Anfangswertproblems (9.50) mittels spezieller Ein- oder Mehrschrittverfahren bezüglich des gleichen Gitters ermöglicht.

9.4.2 Numerische Realisierung des Einfachschießverfahrens mit einer Fixpunktiteration

Eine weitere Möglichkeit zur numerische Realisierung des Einfachschießverfahrens besteht in der Anwendung einer Fixpunktiteration,

$$s_{n+1} \;=\; s_n - \gamma F(s_n), \qquad n = 0, 1, \ldots, \tag{9.51}$$

mit einem Startwert $s_0 \in \mathbb{R}$ und einem Parameter $\gamma > 0$. In Aufgabe 9.13 sind Bedingungen angegeben, die eine Kontraktionseigenschaft und damit Konvergenz der Fixpunktiteration (9.51) gewährleisten.

Weitere Themen und Literaturhinweise

Die Theorie der Randwertprobleme für gewöhnliche Differenzialgleichungssysteme wird beispielsweise in Heuser [54] und in Dallmann/Elster [15] einführend behandelt. Dort findet man auch zahlreiche Beispiele für spezielle Randwertprobleme. Eine Auswahl existierender Lehrbücher mit Abschnitten über die numerische Lösung von Randwertproblemen bildet Emmrich [24], Golub/Ortega [37], Kress [63], Schwarz/Klöckner [94], Stoer/Bulirsch [99] und Weller [110]. Ausführliche Erläuterungen über die Finite-Elemente-Methode in mehreren Raumdimensionen zur Lösung von Randwertproblemen für partielle Differenzialgleichungen findet man beispielsweise in Bärwolff [2], Braess [7], Goering/Roos/Tobiska [33], Großmann/Roos [43], Hanke-Bourgeois [52], Knabner/Angermann [61], Jung/Langer [59], Suttmeier [103] und in Schwetlick/Kretzschmar [96]. Den Aubin-Nitsche-Trick zur Herleitung von Fehlerabschätzungen für das Galerkin-Verfahren findet man in [7] oder Finckenstein [26], Band 2. Die Theorie der nichtnegativen Matrizen wird beispielsweise in Berman/Plemmons [4] und in Horn/Johnson [58] behandelt.

Einfachschießverfahren lassen sich problemlos auf allgemeinere Randwertprobleme (etwa mit nichtlinearen Randbedingungen) übertragen. Gelegentlich stellen sich bei Einfachschießverfahren jedoch Instabilitäten gegenüber Datenstörungen ein (dieser Effekt wird in Aufgabe 9.14 anhand eines Randwertproblems für eine einfache lineare Differenzialgleichung 2. Ordnung demonstriert), weswegen in der Praxis auch *Mehrfachschießverfahren* eingesetzt werden, die hier jedoch nicht weiter behandelt werden. Eine Einführung hierzu findet man etwa [99], wo auch ein Vergleich der einzelnen zur Lösung von Randwertproblemen bei gewöhnlichen Differenzialgleichungen verwendeten Verfahren angestellt wird.

Übungsaufgaben

Aufgabe 9.1. Im Folgenden wird das Randwertproblem

$$\begin{aligned} u''(x) \;+\; p(x)u'(x) \;+\; r(x)u(x) \;&=\; \varphi(x), \qquad x \in [a,b], \\ u(a) \;=\; \alpha, \quad u(b) \;&=\; \beta, \end{aligned}$$

betrachtet mit Zahlen $\alpha,\ \beta \in \mathbb{R}$ und Funktionen $p,\ r,\ \varphi \in C[a,b]$ mit $r(x) \le 0$ für $x \in [a,b]$. Approximation der Ableitungen u' und u'' durch zentrale Differenzenquotienten erster beziehungsweise zweiter Ordnung auf einem äquidistanten Gitter $x_j = a + j(b-a)/N$ für $j = 1, 2, \ldots, N-1$ führt mit einer gewissen Matrix $A \in \mathbb{R}^{(N-1)\times(N-1)}$ und einem gewissen Vektor $b \in \mathbb{R}^{N-1}$ auf ein lineares Gleichungssystem $Av = b$ für $v = (v_1, v_2, \ldots, v_{N-1})^\top \in \mathbb{R}^{N-1}$, mit den Näherungen $v_j \approx u(x_j)$. Man gebe A und b an und zeige, dass das Gleichungssystem für hinreichend kleine Werte von h eindeutig lösbar ist.

Aufgabe 9.2. Für eine Matrix $A \in \mathbb{R}^{N\times N}$ sei eine *reguläre Zerlegung* gegeben, also eine Zerlegung der Form

$$A \;=\; B - P, \qquad B,\ P \;\in\; \mathbb{R}^{N\times N}, \qquad B \text{ regulär}, \quad B^{-1} \;\ge\; 0, \quad P \;\ge\; 0.$$

Dann gilt die folgende Äquivalenz:

$$A \text{ regulär}, \quad A^{-1} \;\ge\; 0 \quad\Longleftrightarrow\quad I - B^{-1}P \text{ regulär}, \quad (I - B^{-1}P)^{-1} \;\ge\; 0.$$

Ist eine dieser beiden Bedingungen erfüllt, so gilt $r_\sigma(B^{-1}P) < 1$.

Aufgabe 9.3. Eine Matrix $A \in \mathbb{R}^{N\times N}$ sei regulär mit einer nichtnegativen Inversen, $A^{-1} \ge 0$. Man zeige: für jede reguläre Zerlegung $A \;=\; B - P$ der Matrix A gilt

$$r_\sigma(B^{-1}P) \;=\; \frac{r_\sigma(A^{-1}P)}{1 + r_\sigma(A^{-1}P)}.$$

Aufgabe 9.4. Gegeben sei eine reguläre Matrix $A \in \mathbb{R}^{N\times N}$ mit $A^{-1} \ge 0$ und zwei regulären Zerlegungen $A \;=\; B_1 - P_1 \;=\; B_2 - P_2$, wobei $P_1 \le P_2$ gelte. Man weise die Ungleichungen $r_\sigma(B_1^{-1}P_1) \le r_\sigma(B_2^{-1}P_2) < 1$ nach.

Aufgabe 9.5. Für eine Funktion $\varphi \in C[0,1]$ betrachte man das Randwertproblem

$$u'' \;=\; \varphi(x), \qquad u(0) \;=\; u(1) \;=\; 0. \tag{9.52}$$

(a) Man zeige, dass sich die Lösung von (9.52) in der Form

$$u(x) \;=\; \int_0^1 G(x,\xi)\,\varphi(\xi)\,d\xi, \qquad x \in [0,1],$$

schreiben lässt mit der greenschen Funktion

$$G(x,\xi) \;=\; \begin{cases} \xi(x-1), & \text{falls } \xi \;\le\; x, \\ x(\xi-1), & \text{sonst.} \end{cases}$$

(b) Die Funktionen u beziehungsweise $u + \Delta u$ seien Lösungen des Randwertproblems (9.52) beziehungsweise der fehlerbehafteten Version

$$(u + \Delta u)'' \;=\; \varphi \;+\; \Delta\varphi \qquad \text{auf } [0,1], \qquad (u + \Delta u)(0) \;=\; (u + \Delta u)(1) \;=\; 0,$$

mit $\Delta\varphi \in C[0,1],\ |\Delta\varphi(x)| \le \varepsilon$ für $x \in [0,1]$. Man zeige $|\Delta u(x)| \le \varepsilon x(1-x)/2$ für $x \in [0,1]$.

(c) Das Differenzenverfahren mit zentralen Differenzenquotienten zweiter Ordnung liefert als Lösung eines lineares Gleichungssystems $A_0 v = b$ Näherungswerte v_j für $u(x_j)$ mit $x_j = j/N,\ j = 1, 2, \ldots, N-1$. Für die fehlerbehaftete Variante

$$A_0(v + \Delta v) = b + \Delta b \qquad \text{mit} \quad \Delta b \in \mathbb{R}^{N-1}, \quad \|\Delta b\|_\infty \le \varepsilon$$

weise man Folgendes nach,

$$|\Delta v_j| \le \frac{\varepsilon}{2} x_j (1 - x_j) \qquad \text{für} \quad j = 1, 2, \ldots, N-1.$$

Aufgabe 9.6. Die lineare Abbildung $\Delta : \mathbb{R}^{N+1} \to \mathbb{R}^{N-1}$ sei definiert durch

$$(\Delta v)_j := b_j v_{j-1} - a_j v_j + c_j v_{j+1}, \qquad j = 1, 2, \ldots, N-1,$$

mit Koeffizienten $b_j > 0,\ c_j > 0$ und $a_j \ge b_j + c_j$ für $j = 1, 2, \ldots, N-1$.

(a) Man beweise das folgende *diskrete Maximumprinzip*: Wenn für den Vektor $v = (v_0, \ldots, v_N)^\top \in \mathbb{R}^{N+1}$ mit $\Delta v \ge 0$ die folgende Bedingung erfüllt ist,

$$v_{j_*} = \max_{j=0,\ldots,N} v_j \qquad \text{für ein } 1 \le j_* \le N-1,$$

so gilt $v_0 = v_1 = \ldots = v_N$.

(b) Man beweise die *inverse Monotonie* der Abbildung $-\Delta$: Wenn für Zahlen u_j und $v_j \in \mathbb{R}$ $(j = 0, \ldots, N)$ die Bedingungen

$$-\Delta u \le -\Delta v, \qquad u_0 \le v_0, \quad u_N \le v_N,$$

erfüllt sind, so gilt $u \le v$.

Aufgabe 9.7. Gegeben sei eine Zerlegung $\Delta = \{a = x_0 < x_1 < \ldots < x_N = b\}$ des Intervalls $[a, b]$, und $h_{\max} = \max_{j=0,\ldots,N-1}\{x_{j+1} - x_j\}$ bezeichne den maximalen Knotenabstand. Man zeige: für jede Funktion $f \in C^1_\Delta[a, b]$ mit $f(x_0) = f(x_1) = \ldots = f(x_N) = 0$ gilt die Abschätzung $\|f\|_2 \le h_{\max} \|f'\|_2$.

Aufgabe 9.8. Gegeben sei der Differenzialoperator

$$\mathcal{L} : C[a, b] \supset \mathcal{D}_{\mathcal{L}} \to C[a, b], \qquad u \mapsto -(pu')' + ru,$$
$$\mathcal{D}_{\mathcal{L}} = \{u \in C^2[a, b] : u(a) = \alpha u(b) + u'(b) = 0\},$$

mit $p \in C^1[a, b],\ r \in C[a, b],\ p(x) \ge p_0 > 0,\ r(x) \ge 0$ für $x \in [a, b]$ und mit $\alpha \ge 0$. Die Bilinearform $[\![\cdot, \cdot]\!]$ auf $C^1_\Delta[a, b]$ sei durch

$$[\![u, v]\!] = \int_a^b [pu'v' + ruv]\, dx + \alpha(puv)(b), \qquad u,\ v \in C^1_\Delta[a, b],$$

definiert, und $\langle \cdot, \cdot \rangle_2$ sei das L_2-Skalarprodukt auf $C[a, b]$. Man zeige Folgendes:

(a) Die Bilinearform $[\![\cdot, \cdot]\!]$ stellt eine Fortsetzung der Abbildung $\langle \mathcal{L}\cdot, \cdot \rangle_2$ dar, und bezüglich des Skalarprodukts $\langle \cdot, \cdot \rangle_2$ ist die Abbildung $\mathcal{L}$ symmetrisch.

(b) Man zeige $c_1 \|u\|^2_\infty \le [\![u, u]\!] \le c_2 \|u'\|^2_\infty$ für $u \in C^1_\Delta[a, b]$ mit $u(a) = 0$, mit geeigneten Konstanten c_1 und c_2.

Aufgabe 9.9. Gegeben sei der folgende Differenzialoperator vierter Ordnung,

$$\mathcal{L} : C[a,b] \supset \mathcal{D}_{\mathcal{L}} \to C[a,b], \qquad u \mapsto (pu'')'' + ru,$$
$$\mathcal{D}_{\mathcal{L}} = \{u \in C^4[a,b] : u(a) = u'(a) = u''(b) = u'''(b) = 0\},$$

mit $p \in C^2[a,b]$, $r \in C[a,b]$, $p(x) \geq p_0 > 0$, $r(x) \geq 0$ für $x \in [a,b]$, und $\langle \cdot, \cdot \rangle_2$ sei das L_2-Skalarprodukt auf $C[a,b]$.

(a) Man zeige, dass die Abbildung $\mathcal{L}$ symmetrisch und positiv definit bezüglich $\langle \cdot, \cdot \rangle_2$ ist.
(b) Auf dem Raum $C^2_\Delta[a,b] = \{u \in C^1[a,b] \to \mathbb{R} : u' \text{ stückweise stetig differenzierbar}\}$ bestimme man eine Bilinearform $[\![\cdot, \cdot]\!]$, die eine Fortsetzung der Abbildung $\langle \mathcal{L}\cdot, \cdot \rangle_2$ darstellt und für die Abschätzungen von der Form $c_1 \|u\|^2_\infty \leq [\![u,u]\!] \leq c_2 \|u''\|^2_\infty$ gelten für $u \in C^2_\Delta[a,b]$ mit $u(a) = u'(a) = 0$.

Aufgabe 9.10. Man zeige: Für eine positiv definite Bilinearform $[\![\cdot, \cdot]\!] : \mathcal{D} \times \mathcal{D} \to \mathbb{R}$ auf einem reellen Vektorraum $\mathcal{D}$ gilt die verallgemeinerte cauchy-schwarzsche Ungleichung,

$$|[\![u,v]\!] + [\![v,u]\!]| \leq 2[\![u,u]\!]^{1/2} [\![v,v]\!]^{1/2} \qquad \text{für } u, v \in \mathcal{D}.$$

Daraus leite man die Dreiecksungleichung für die zugehörige Norm $\mathcal{D} \ni u \mapsto [\![u,u]\!]^{1/2}$ her.

Aufgabe 9.11 (Fehlerquadratmethode). Es seien $\mathcal{V}$ und $\mathcal{W}$ reelle Vektorräume, die Abbildung $\mathcal{L} : \mathcal{V} \to \mathcal{W}$ sei injektiv und linear, und $\langle \cdot, \cdot \rangle$ sei ein Skalarprodukt auf $\mathcal{W}$ mit der zugehörigen Norm $\| \cdot \|$. Außerdem seien $u_* \in \mathcal{V}$ und $\varphi \in \mathcal{W}$. Man weise die Äquivalenz der folgenden drei Aussagen nach:

(i) u_* löst die Minimierungsaufgabe $\| \mathcal{L}u - \varphi \| \to \min$ für $u \in \mathcal{V}$.
(ii) Es gilt $\langle \mathcal{L}u_*, \mathcal{L}v \rangle = \langle \varphi, \mathcal{L}v \rangle$ für jedes $v \in \mathcal{V}$.
(iii) Es gilt $\mathcal{L}u_* - \varphi \in \mathcal{R}(\mathcal{L})^\perp$, dem orthogonalen Komplement des Bildraums von $\mathcal{L}$ bezüglich $\langle \cdot, \cdot \rangle$.

Ist weiter der Vektorraum $\mathcal{V}$ endlich-dimensional mit Basis $v_1, \ldots, v_N$ und gilt die Identität $u_* = \sum_{k=1}^N c_k v_k$ mit gewissen Koeffizienten $c_1, \ldots, c_N$, so ist jede der Eigenschaften (i), (ii) und (iii) äquivalent zu der Eigenschaft $Ac = b$ mit den Notationen

$$A = (\langle \mathcal{L}v_j, \mathcal{L}v_k \rangle)_{j,k=1}^N, \qquad b = (\langle \varphi, \mathcal{L}v_j \rangle)_{j=1}^N, \qquad c = (c_1, \ldots, c_N)^\top.$$

Aufgabe 9.12. Gegeben sei das Randwertproblem

$$\mathcal{L}u = -u'' + xu = -x^3 + x^2 + 2, \quad x \in [0,1], \qquad u(0) = u(1) = 0.$$

Wie lautet das ritzsche Gleichungssystem, wenn als Ansatzfunktionen trigonometrische Polynome von der Form $v_k(x) = \sqrt{2} \sin k\pi x$, $k = 1, 2, \ldots, N$ verwendet werden?

Aufgabe 9.13. Man betrachte das Randwertproblem $u'' = f(x, u, u')$, $u(a) = \alpha$, $u(b) = \beta$ mit einer stetig partiell differenzierbaren Funktion $f : [a,b] \times \mathbb{R}^2 \to \mathbb{R}$, die die folgenden Bedingungen erfülle,

$$0 < \frac{\partial f}{\partial u}(x, v_1, v_2) \leq K, \qquad \left| \frac{\partial f}{\partial u'}(x, v_1, v_2) \right| \leq L, \qquad (x, v_1, v_2) \in [a,b] \times \mathbb{R}^2,$$

mit gewissen Konstanten K, $L \geq 0$. Sei $u(\cdot, s)$ Lösung des zugehörigen Anfangswertproblems (9.45)–(9.46).

(a) Für die Ableitung der zum Einfachschießverfahren korrespondierenden Funktion $F(s) = u(b, s) - \beta$ weise man die Ungleichungen $0 < \kappa_1 \le F'(s) \le \kappa_2$ für $s \in \mathbb{R}$ nach, mit den Konstanten

$$\kappa_1 := \frac{1}{L}\big(1 - \exp(-L(b-a))\big),$$
$$\kappa_2 := \frac{2\exp(L\frac{b-a}{2})}{C}\sinh\big(C\frac{b-a}{2}\big), \quad \text{mit} \quad C := L\sqrt{1+\frac{4K}{L^2}}.$$

(b) Man weise nach, dass das Iterationsverfahren

$$s^{(n+1)} = \Phi(s^{(n)}) := s^{(n)} - \gamma F(s^{(n)}) \qquad \text{für} \quad n = 0, 1, \ldots,$$

für jeden Startwert $s^{(0)}$ und jeden Wert $0 < \gamma < 2/\kappa_2$ gegen die (einzige) Nullstelle s_* der Funktion F konvergiert. Für $\gamma = 2/(\kappa_1 + \kappa_2)$ weise man die folgende a priori-Fehlerabschätzung nach:

$$|s^{(n)} - s_*| \le \big(\frac{\kappa_2 - \kappa_1}{\kappa_2 + \kappa_1}\big)^n \frac{|F(s^{(0)})|}{\kappa_1}, \qquad n = 0, 1, \ldots .$$

Aufgabe 9.14. Zur Lösung des Randwertproblems

$$u'' = 100u \quad \text{auf } [0,3], \qquad u(0) = 1, \quad u(3) = e^{-30},$$

betrachte man die Lösung $u(\cdot, s)$ des Anfangswertproblems $u'' = 100u$, $u(0) = 1$, $u'(0) = s$. Man berechne $u(3, s_\varepsilon)$ für $s_\varepsilon = s_*(1+\varepsilon)$, wobei s_* die Lösung der Gleichung $u(3, s_*) = e^{-30}$ bezeichnet und $\varepsilon > 0$ beliebig ist. Ist in diesem Fall das Einfachschießverfahren eine geeignete Methode zur Lösung des vorliegenden Randwertproblems?

Aufgabe 9.15 (*Numerische Aufgabe*). Man löse numerisch das Randwertproblem

$$u''(x) + 6x(1-x)u'(x) + u(x)^2 = x^4 + 10x^3 - 17x^2 + 6x - 2, \quad x \in [0,1],$$
$$u(0) = u(1) = 0,$$

mit dem Einfachschießverfahren. Zur Nullstellensuche verwende man das Newton-Verfahren einmal mit Startwert $s^{(0)} = 1$ und einmal mit $s^{(0)} = 20$. Die jeweiligen Anfangswertprobleme löse man numerisch mit dem expliziten Eulerverfahren mit Schrittweite $h = 1/30$. Man gebe die Näherungen v_j zu den Gitterpunkten $x_j = jh$, $j = 0, 1, \ldots, 30$, tabellarisch an.

10 Gesamtschritt-, Einzelschritt- und Relaxationsverfahren zur Lösung linearer Gleichungssysteme

10.1 Iterationsverfahren zur Lösung linearer Gleichungssysteme

Zur Lösung linearer Gleichungssysteme

$$Ax = b \qquad \left(A \in \mathbb{R}^{N\times N} \text{ regulär}, \qquad b \in \mathbb{R}^N\right) \tag{10.1}$$

mit der eindeutigen Lösung $x_* = A^{-1}b \in \mathbb{R}^N$ werden in den beiden folgenden Kapiteln 10 und 11 einige spezielle Iterationsverfahren vorgestellt. Dabei hat man sich unter einem *Iterationsverfahren* ganz allgemein ein Verfahren vorzustellen, bei dem – ausgehend von einem beliebigen Startvektor $x^{(0)} \in \mathbb{R}^N$ – sukzessive Vektoren $x^{(1)}, x^{(2)}, \ldots \in \mathbb{R}^N$ berechnet werden gemäß der zum jeweiligen Verfahren gehörenden Iterationsvorschrift.

10.1.1 Hintergrund zum Einsatz iterativer Verfahren bei linearen Gleichungssystemen

Iterative Verfahren werden unter anderem zur schnellen approximativen Lösung linearer Gleichungssysteme (10.1) eingesetzt. Im Vergleich dazu benötigen die in Kapitel 4 vorgestellten direkten Verfahren zur Lösung eines Gleichungssystems von der Form (10.1) im Allgemeinen[1] $cN^3 + \mathcal{O}(N^2)$ arithmetische Operationen mit einer gewissen Konstanten $c > 0$. Demgegenüber setzt sich bei jedem der vorzustellenden Iterationsverfahren ein einzelner Iterationsschritt typischerweise wie folgt zusammen:

- es treten ein oder zwei Matrix-Vektor-Multiplikationen auf, die mit jeweils N^2 Multiplikationen zu Buche schlagen,
- zudem sind mehrere kleine Operationen notwendig wie etwa die Berechnung von Skalarprodukten oder Summen von Vektoren, bei denen insgesamt $\mathcal{O}(N)$ arithmetische Operationen anfallen.

Insgesamt erfordert die Durchführung eines Iterationsschrittes also $\mathcal{O}(N^2)$ arithmetische Operationen. Liefert nun das Iterationsverfahren nach einer vertretbaren Anzahl von $n \ll N$ Iterationsschritten hinreichend gute Approximationen $x^{(n)} \approx x_*$, so beträgt der Gesamtaufwand insgesamt also deutlich weniger als die oben genannten $cN^3 + \mathcal{O}(N^2)$ arithmetischen Operationen.

[1] also bei voll besetzter Matrix A ohne spezielle Struktur

Weitere zu beachtende Aspekte im Zusammenhang mit dem Einsatz iterativer Verfahren sind in der nachfolgenden Bemerkung aufgeführt.

Bemerkung 10.1. (a) Bereits bei der numerischen Lösung *nichtlinearer* Gleichungssysteme in Kapitel 5 sind einige Iterationsverfahren vorgestellt worden, dort vor dem Hintergrund fehlender direkter Methoden. Natürlich lassen sich einige der dort vorgestellten Resultate – so zum Beispiel der banachsche Fixpunktsatz (Theorem 5.7) – zur approximativen Lösung linearer Gleichungssysteme verwenden. In den beiden folgenden Kapiteln 10–11 wird sich jedoch Folgendes herausstellen:

- Für gewisse Fixpunktiterationen lassen sich auch bei fehlender Kontraktionseigenschaft noch Konvergenzresultate nachweisen, und dies größtenteils bei beliebiger Wahl des Startwerts $x^{(0)} \in \mathbb{R}^N$.
- Für Gleichungssysteme $Ax = b$ mit speziellen Eigenschaften – etwa Monotonie oder Symmetrie von A – lassen sich besonders effiziente Methoden einsetzen.

(b) In den Anwendungen treten häufig Fragestellungen auf, deren Modellierung und anschließende Diskretisierung auf große lineare Gleichungssysteme $Ax = b$ mit *schwach besetzten* (ein Großteil der N^2 Einträge ist also identisch null) Matrizen $A \in \mathbb{R}^{N\times N}$ führen. Ein Modellbeispiel hierzu ist in Abschnitt 10.2.1 angegeben. Die bereits getroffenen Aussagen über direkte und iterative Löser lassen sich mit entsprechenden Modifikationen bezüglich des Aufwands übertragen. △

10.2 Lineare Fixpunktiteration

Eine Klasse von Iterationsverfahren zur approximativen Bestimmung der Lösung x_* der Gleichung (10.1) gewinnt man durch Umformulierung von $Ax = b$ in eine Fixpunktgleichung der Form

$$x \;=\; \mathcal{H}\,x \;+\; z, \tag{10.2}$$

mit einer geeigneten zunächst nicht näher spezifizierten *Iterationsmatrix* $\mathcal{H} \in \mathbb{R}^{N\times N}$ sowie einem geeigneten Vektor $z \in \mathbb{R}^N$. Es sei nur angenommen, dass die Lösung $x_* \in \mathbb{R}^N$ der Gleichung (10.1) zugleich einziger Fixpunkt von (10.2) ist. Die zur Fixpunktgleichung (10.2) gehörende *lineare Fixpunktiteration* lautet dann

$$x^{(n+1)} \;=\; \mathcal{H}\,x^{(n)} \;+\; z, \qquad n \;=\; 0, 1, \ldots, \tag{10.3}$$

wobei $x^{(0)} \in \mathbb{R}^N$ ein frei wählbarer Startvektor ist. Im Folgenden werden für lineare Fixpunktiterationen der Form (10.3) Resultate für (globale) Konvergenz im Sinne der folgenden Definition geliefert.

Definition 10.2. Das Verfahren (10.3) zur Bestimmung von $x_* \in \mathbb{R}^N$ heißt *konvergent*, wenn für jeden Startwert $x^{(0)} \in \mathbb{R}^N$ Folgendes gilt,

$$\|\,x^{(n)} - x_*\| \;\to\; 0 \qquad \text{für} \quad n \to \infty. \tag{10.4}$$

Hier bezeichnet $\|\cdot\| : \mathbb{R}^N \to \mathbb{R}$ eine nicht näher spezifizierte Vektornorm. Ein nicht konvergentes Verfahren (10.3) nennt man *divergent*.

Theorem 10.3. *Das stationäre Iterationsverfahren (10.3) ist konvergent genau dann, wenn die Ungleichung $r_\sigma(\mathcal{H}) < 1$ erfüllt ist.*

BEWEIS. Nach Voraussetzung gilt $x_* = \mathcal{H}x_* + z$, und somit gelten die Fehlerdarstellungen $x^{(n+1)} - x_* = \mathcal{H}(x^{(n)} - x_*)$ beziehungsweise

$$x^{(n)} - x_* = \mathcal{H}^n(x^{(0)} - x_*), \qquad n = 0, 1, \dots . \tag{10.5}$$

Konvergenz ist demnach gleichbedeutend mit $\mathcal{H}^n \to 0$ für $n \to \infty$. Dies wiederum ist nach Theorem 9.13 äquivalent zur Eigenschaft $r_\sigma(\mathcal{H}) < 1$. □

Bemerkung 10.4. Ebenfalls nach Theorem 9.13 ist das stationäre Iterationsverfahren (10.3) konvergent genau dann, wenn eine Vektornorm $\|\cdot\| : \mathbb{R}^N \to \mathbb{R}$ existiert, so dass für die zugehörige Matrixnorm die Ungleichung $\|\mathcal{H}\| < 1$ erfüllt ist. Für spezielle Matrizen A und spezielle Verfahren (10.3) ist es jedoch häufig so, dass dieses Kriterium für gängige und leicht zu berechnende Normen nicht erfüllt ist, obwohl die (oft auch nachweisbare) Ungleichung $r_\sigma(\mathcal{H}) < 1$ erfüllt ist und somit Konvergenz vorliegt. △

Die Konvergenz der linearen Fixpunktiteration (10.3) ist umso besser, je kleiner der Spektralradius $r_\sigma(\mathcal{H})$ ausfällt:

Theorem 10.5. *Zu einer beliebigen Matrix $\mathcal{H} \in \mathbb{R}^{N\times N}$ und jeder Zahl $\varepsilon > 0$ existiert eine Vektornorm $\|\cdot\| : \mathbb{R}^N \to \mathbb{R}$, mit der für das stationäre Iterationsverfahren (10.3) die folgende Abschätzung gilt,*

$$\|x^{(n)} - x_*\| \le (r_\sigma(\mathcal{H}) + \varepsilon)^n \|x^{(0)} - x_*\|, \qquad n = 0, 1, \dots .$$

BEWEIS. Die Aussage ist eine unmittelbare Konsequenz aus der Darstellung (10.5) und dem folgenden Lemma. □

Lemma 10.6. *Zu jeder Matrix $\mathcal{H} \in \mathbb{R}^{N\times N}$ und jeder Zahl $\varepsilon > 0$ existiert eine Vektornorm $\|\cdot\| : \mathbb{R}^N \to \mathbb{R}$, so dass für die zugehörige Matrixnorm die folgende Ungleichung gilt:*

$$\|\mathcal{H}\| \le r_\sigma(\mathcal{H}) + \varepsilon.$$

BEWEIS. Mit der Notation $a := 1/(r_\sigma(\mathcal{H}) + \varepsilon)$ erhält man $r_\sigma(a\mathcal{H}) = a\, r_\sigma(\mathcal{H}) < 1$, und Theorem 9.13 liefert dann die Existenz einer Vektornorm $\|\cdot\| : \mathbb{R}^N \to \mathbb{R}$, so dass für die zugehörige Matrixnorm die Ungleichung $\|a\mathcal{H}\| < 1$ erfüllt ist. Daraus erhält man unmittelbar die Aussage des Lemmas. □

Als unmittelbare Konsequenz aus Lemma 10.6 erhält man Folgendes:

Korollar 10.7. *Für jede Matrix $\mathcal{H} \in \mathbb{R}^{N\times N}$ gilt*

$$r_\sigma(\mathcal{H}) = \inf\{\|\mathcal{H}\| : \textit{die Matrixnorm ist durch eine reelle Vektornorm induziert}\}. \tag{10.6}$$

In Aufgabe 10.1 wird ein Kriterium dafür angegeben, wann in (10.6) das Minimum angenommen wird.

10.2.1 Ein Modellbeispiel

Problemstellung

Im Folgenden wird ein Beispiel vorgestellt, bei dem die noch vorzustellenden iterativen Verfahren sinnvoll angewendet werden können[2]. Es handelt sich hierbei um ein dirichletsches Randwertproblem für die *Poisson-Gleichung*,

$$-\frac{\partial^2 u}{\partial x^2} - \frac{\partial^2 u}{\partial y^2} = f \qquad \text{auf } \Omega := (0,1)^2, \tag{10.7}$$

$$u = 0 \qquad \text{auf } \Gamma := \text{Rand von } [0,1]^2, \tag{10.8}$$

wobei $f : [0,1]^2 \to \mathbb{R}$ eine gegebene stetige Funktion ist, und die Funktion $u : [0,1]^2 \to \mathbb{R}$ ist zu bestimmen. Im Folgenden wird vorausgesetzt, dass das Randwertproblem (10.7)–(10.8) eine eindeutig bestimmte stetige und im Inneren von $[0,1]^2$ zweimal stetig differenzierbare Lösung $u : [0,1]^2 \to \mathbb{R}$ besitzt.[3]

Der Ansatz für Differenzenverfahren

Zur numerischen Lösung des Randwertproblems (10.7)–(10.8) mittels Differenzenverfahren wird das zugrunde liegende Intervall $[0,1]^2$ mit Gitterpunkten versehen, die hier äquidistant gewählt seien,

$$x_j = jh, \qquad y_k = kh, \qquad j,\, k = 0, 1, \ldots, M \qquad \left(h = \tfrac{1}{M}\right). \tag{10.9}$$

Die inneren Gitterpunkte sind in Bild 10.1 dargestellt.

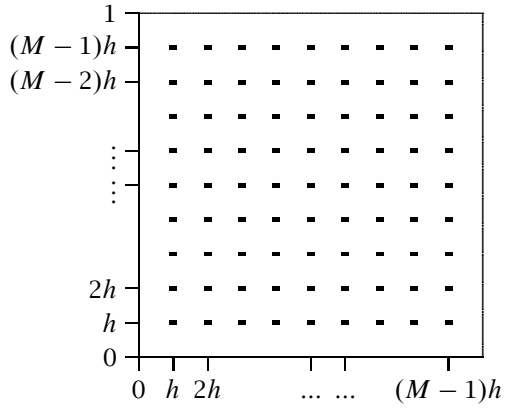


Bild 10.1: Darstellung des gegebenen Gitters

Bezüglich dieses Gitters (10.9) wird das Randwertproblem (10.7)–(10.8) in zweierlei Hinsicht diskretisiert: die Poisson-Gleichung (10.7) wird lediglich an den inneren

[2] vergleiche Bemerkung 10.1

[3] Unter zusätzlichen Voraussetzungen an f ist diese Annahme erfüllt (Hackbusch [46], Kapitel 3).

Gitterpunkten (x_j, y_k), $1 \leq j, k \leq M-1$, betrachtet, und die partiellen Ableitungen werden dort jeweils durch zentrale Differenzenquotienten 2. Ordnung approximiert,

$$\left.\begin{aligned}
-\frac{\partial^2 u}{\partial x^2}(x_j, y_k) &= \frac{-u(x_{j-1}, y_k) + 2u(x_j, y_k) - u(x_{j+1}, y_k)}{h^2} + \mathcal{O}(h^2), \\
-\frac{\partial^2 u}{\partial y^2}(x_j, y_k) &= \frac{-u(x_j, y_{k-1}) + 2u(x_j, y_k) - u(x_j, y_{k+1})}{h^2} + \mathcal{O}(h^2), \\
& \qquad\qquad j, k = 1, 2, \ldots, M-1,
\end{aligned}\right\} \quad (10.10)$$

wobei hier $u \in C^4([0,1]^2)$ angenommen wird. Vernachlässigung des Restglieds in (10.10) führt auf das folgende gekoppelte System von $N = (M-1)^2$ linearen Gleichungen,

$$\frac{-U_{j-1,k} - U_{j,k-1} + 4U_{j,k} - U_{j,k+1} - U_{j+1,k}}{h^2} = f_{j,k}, \quad j, k = 1, \ldots, M-1, \quad (10.11)$$

für die Approximationen

$$U_{j,k} \approx u(x_j, y_k), \qquad j, k = 1, 2, \ldots, M-1,$$

wobei in (10.11) noch

$$\begin{aligned}
U_{j,0} = U_{0,k} &= 0, & j, k &= 1, 2, \ldots, M-1, \\
f_{j,k} &= f(x_j, y_k), & &\text{——— « ———}
\end{aligned}$$

gesetzt ist. Zu jedem Gitterpunkt (x_j, y_k) korrespondiert in natürlicher Weise sowohl die Unbekannte $U_{j,k}$ als auch eine Gleichung aus (10.11).

Ordnet man in Bild 10.1 diese Gitterpunkte beziehungsweise die entsprechenden Unbekannten und Gleichungen zeilenweise (von links nach rechts) und dann aufwärts an, so erhält man die folgende Matrixdarstellung für die Gleichungen (10.11),

$$\underbrace{\frac{1}{h^2}\left(\begin{array}{cccc|cccc|cccc|cccc}
4 & -1 & & & -1 & & & & & & & & & & & \\
-1 & \ddots & \ddots & & & \ddots & & & & & & & & & & \\
& \ddots & \ddots & -1 & & & \ddots & & & & & & & & & \\
& & -1 & 4 & & & & -1 & & & & & & & & \\
\hline
-1 & & & & 4 & -1 & & & \ddots & & & & & & & \\
& \ddots & & & -1 & \ddots & \ddots & & & \ddots & & & & & & \\
& & \ddots & & & \ddots & \ddots & -1 & & & \ddots & & & & & \\
& & & -1 & & & -1 & 4 & & & & \ddots & & & & \\
\hline
& & & & \ddots & & & & \ddots & & & & -1 & & & \\
& & & & & \ddots & & & & \ddots & & & & \ddots & & \\
& & & & & & \ddots & & & & \ddots & & & & \ddots & \\
& & & & & & & \ddots & & & & \ddots & & & & -1 \\
\hline
& & & & & & & & -1 & & & & 4 & -1 & & \\
& & & & & & & & & \ddots & & & -1 & \ddots & \ddots & \\
& & & & & & & & & & \ddots & & & \ddots & \ddots & -1 \\
& & & & & & & & & & & -1 & & & -1 & 4
\end{array}\right)}_{=:\,A}
\left(\begin{array}{c}
U_{1,1} \\ \vdots \\ \vdots \\ U_{M-1,1} \\ \hline U_{1,2} \\ \vdots \\ \vdots \\ U_{M-1,2} \\ \hline \vdots \\ \vdots \\ \vdots \\ \vdots \\ \hline U_{1,M-1} \\ \vdots \\ \vdots \\ U_{M-1,M-1}
\end{array}\right)
=
\left(\begin{array}{c}
f_{1,1} \\ \vdots \\ \vdots \\ f_{M-1,1} \\ \hline f_{1,2} \\ \vdots \\ \vdots \\ f_{M-1,2} \\ \hline \vdots \\ \vdots \\ \vdots \\ \vdots \\ \hline f_{1,M-1} \\ \vdots \\ \vdots \\ f_{M-1,M-1}
\end{array}\right)$$

Die zugrunde liegende Matrix $A \in \mathbb{R}^{N\times N}$ mit $N = (M-1)^2$ ist schwach besetzt und dient im Folgenden als ein Referenzbeispiel für die vorzustellenden speziellen Klassen von Matrizen.

Bemerkung 10.8. In dem Differenzenschema (10.11) treten auf der linken Seite der Gleichung für jeden Index (j,k) die Näherungen zum Gitterpunkt (x_j, y_k) und seinen vier Nachbarn auf, weshalb man hier von einer *Fünfpunkteformel* oder auch von einem *Fünfpunktestern* spricht. Die zur Gewinnung der Matrixdarstellung angegebene Reihung der Gitterpunkte wird als *lexikografische Anordnung* bezeichnet. △

10.3 Einige spezielle Klassen von Matrizen und ihre Eigenschaften

In Vorbereitung auf die nachfolgenden Abschnitte 10.4 und 10.5 über das Gesamt- und das Einzelschrittverfahren sollen zunächst einige spezielle Klassen von Matrizen betrachtet werden.

10.3.1 Irreduzible Matrizen

Auch im Folgenden liegt das Hauptaugenmerk auf reellen Matrizen. Aus technischen Gründen wie etwa anstehenden Spektralbetrachtungen werden nun jedoch auch komplexe Matrizen und Normen zugelassen.

Definition 10.9. Eine Matrix $B = (b_{jk}) \in \mathbb{C}^{N\times N}$ heißt *reduzibel*, falls Mengen $\mathcal{J},\ \mathcal{K} \subset \{1,2,\ldots,N\}$ mit folgenden Eigenschaften existieren:

$$\begin{aligned} &\mathcal{J} \neq \varnothing, \qquad \mathcal{K} \neq \varnothing, \qquad \mathcal{J} \cap \mathcal{K} = \varnothing, \qquad \mathcal{J} \cup \mathcal{K} = \{1,2,\ldots,N\}, \\ &b_{jk} = 0 \qquad \forall\, j \in \mathcal{J}, \quad k \in \mathcal{K}. \end{aligned} \tag{10.12}$$

Andernfalls heißt die Matrix *irreduzibel*.

Beispiel 10.10. Die Matrix

$$\begin{pmatrix} 1 & 2 & 0 \\ -1 & 1 & 0 \\ 3 & 0 & 1 \end{pmatrix}$$

ist reduzibel: man betrachte $\mathcal{J} = \{1,2\}$ und $\mathcal{K} = \{3\}$. △

Die Bezeichnung "reduzibel" begründet sich in der folgenden Eigenschaft:

Bemerkung 10.11. Die Lösung eines gegebenen nichtsingulären Gleichungssystems $Ax = b$ mit einer reduziblen Matrix $A = (a_{jk}) \in \mathbb{C}^{N\times N}$ lässt sich in zwei kleinere Teilaufgaben zerlegen (die Notation sei entsprechend Definition 10.9 gewählt):

(i) man bestimmt zunächst die Unbekannten x_j, $j \in \mathcal{J}$, des linearen Gleichungssystems

$$\sum_{k=1}^{N} a_{jk} x_k \;=\; \sum_{k\in\mathcal{J}} a_{jk} x_k \;\stackrel{!}{=}\; b_j, \qquad j \in \mathcal{J}.$$

(ii) Anschließend bestimmt man die Unbekannten x_j, $j \in \mathcal{K}$, des linearen Gleichungssystems

$$\sum_{k\in\mathcal{K}} a_{jk} x_k \;\stackrel{!}{=}\; b_j - \sum_{k\in\mathcal{J}} a_{jk} x_k, \qquad j \in \mathcal{K}.$$

△

Beispiel 10.12. Eine Tridiagonalmatrix ist irreduzibel genau dann, wenn jeder ihrer Nebendiagonaleinträge von null verschieden ist. △

BEWEIS. Die Tridiagonalmatrix sei mit $B = (b_{jk}) \in \mathbb{C}^{N\times N}$ bezeichnet.

"$\Longrightarrow$": Für einen beliebigen Index $j_* \in \{1,\dots,N-1\}$ sind die Mengen $\mathcal{J} = \{1,\dots,j_*\}$ und $\mathcal{K} = \{j_*+1,\dots,N\}$ nichtleer und disjunkt mit $\mathcal{J}\cup\mathcal{K} = \{1,\dots,N\}$. Da für beliebige Indizes $j \in \mathcal{J}$ und $k \in \mathcal{K}$ mit $|j-k| \geq 2$ ohnehin $b_{jk} = 0$ gilt, ist aufgrund der Irreduzibilität der Matrix B notwendigerweise $b_{j_*,j_*+1} \neq 0$. Die Eigenschaft $b_{j_*+1,j_*} \neq 0$ erschließt man nach Vertauschen von $\mathcal{J}$ und $\mathcal{K}$ genauso.

"$\Longleftarrow$": Für beliebige Mengen $\mathcal{J}$, $\mathcal{K} \subset \{1,2,\dots,N\}$ von der Form (10.12) existieren notwendigerweise Indizes $j \in \mathcal{J}$, $k \in \mathcal{K}$, die benachbart sind, es gilt also $k = j+1$ oder $k = j-1$. Für solche Indizes gilt aufgrund der Annahme $b_{jk} \neq 0$, und infolgedessen ist die Matrix B irreduzibel. □

Beispiel 10.13. Die zu dem vorgestellten Modellbeispiel aus Abschnitt 10.2.1 gehörende Matrix ist irreduzibel diagonaldominant (Aufgabe 10.5). △

Die folgenden elementaren Eigenschaften werden ebenfalls noch benötigt.

Lemma 10.14. *Die Matrix $B \in \mathbb{C}^{N\times N}$ sei irreduzibel.*

(a) Für jede Diagonalmatrix $D \in \mathbb{C}^{N\times N}$ ist mit B auch die Matrix $B + D$ irreduzibel.

(b) Für Zahlen $c_{jk} \in \mathbb{R}$ mit $c_{jk} \neq 0$ für $j \neq k$ ist mit $B = (b_{jk})$ auch die Matrix $(c_{jk} b_{jk}) \in \mathbb{C}^{N\times N}$ irreduzibel.

BEWEIS. Ist eine Matrix irreduzibel, so ändert sich diese Eigenschaft aufgrund der Definition offenkundig nicht, wenn man die Diagonaleinträge beliebig abändert. Entsprechendes gilt, wenn die nichtverschwindenden Nichtdiagonaleinträge beliebig zu nichtverschwindenden Einträgen abgeändert werden. □

Definition 10.15. Eine Matrix $B = (b_{jk}) \in \mathbb{C}^{N\times N}$ heißt *irreduzibel diagonaldominant*, falls B irreduzibel ist und weiter Folgendes gilt,

$$\left.\begin{array}{rcll} \displaystyle\sum_{\substack{k=1\\k\neq j}}^{N} |b_{jk}| & \leq & |b_{jj}|, & j = 1,2,\ldots,N, \\ \text{—} \ll \text{—} & < & \text{—} \ll \text{—} & \text{für mindestens ein } j \in \{1,2,\ldots,N\}. \end{array}\right\} \tag{10.13}$$

Theorem 10.16. *Eine irreduzibel diagonaldominante Matrix $B = (b_{jk}) \in \mathbb{C}^{N\times N}$ ist regulär.*

BEWEIS. Angenommen, es gibt einen Vektor $0 \neq x \in \mathbb{C}^N$ mit $Bx = 0$. Für die Indexmengen

$$\mathcal{J} := \{j \,:\, |x_j| = \|x\|_\infty\}, \qquad \mathcal{K} := \{j \,:\, |x_j| < \|x\|_\infty\},$$

gilt dann offensichtlich $\mathcal{J} \cup \mathcal{K} = \{1,\ldots,N\}$ sowie $\mathcal{J} \cap \mathcal{K} = \varnothing$ und $\mathcal{J} \neq \varnothing$. Es gilt jedoch auch $\mathcal{K} \neq \varnothing$, denn andernfalls wäre $|x_j| = \|x\|_\infty$ $(\neq 0)$ für alle j, und zusammen mit der Abschätzung

$$|b_{jj}||x_j| \overset{Bx=0}{\leq} \sum_{\substack{k=1\\k\neq j}}^{N} |b_{jk}||x_k|, \qquad j = 1,2,\ldots,N,$$

ergäbe dies einen Widerspruch zur zweiten Annahme in (10.13). Nach Annahme existieren nun Indizes $j_* \in \mathcal{J}$, $k_* \in \mathcal{K}$ mit $b_{j_*k_*} \neq 0$. Daraus folgt

$$|b_{j_*j_*}| \overset{Bx=0}{\leq} \sum_{\substack{k=1\\k\neq j_*}}^{N} \underbrace{|b_{j_*k}|}_{\neq 0 \text{ für } k=k_*} \underbrace{\frac{|x_k|}{|x_{j_*}|}}_{<1 \text{ für } k=k_*} < \sum_{\substack{k=1\\k\neq j_*}}^{N} |b_{j_*k}|$$

im Widerspruch zur ersten Annahme in (10.13). □

10.4 Das Gesamtschrittverfahren

Für eine gegebene Matrix $A \in \mathbb{R}^{N\times N}$ sowie einen Vektor $b \in \mathbb{R}^N$ bedeutet das lineare Gleichungssystem $Ax = b$ ausgeschrieben Folgendes,

$$\sum_{k=1}^{N} a_{jk}x_k = b_j, \qquad j = 1,2,\ldots,N, \tag{10.14}$$

und im Folgenden sollen verschiedene Fixpunktformulierungen für das Gleichungssystem (10.14) angegeben werden unter Verwendung der folgenden Zerlegung der

Matrix A,

$$\underbrace{\begin{pmatrix} a_{11} & \dots\dots & a_{1N} \\ \vdots & \ddots & \vdots \\ \vdots & \ddots & \vdots \\ a_{N1} & \dots\dots & a_{NN} \end{pmatrix}}_{=:\,A} = \underbrace{\begin{pmatrix} a_{11} & & \\ & \ddots & \\ & & a_{NN} \end{pmatrix}}_{=:\,D} + \underbrace{\begin{pmatrix} & & \\ a_{21} & & \\ \vdots & \ddots & \\ a_{N1} & \cdots & a_{N,N-1} \end{pmatrix}}_{=:\,L} + \underbrace{\begin{pmatrix} & a_{12} & \cdots & a_{1N} \\ & & \ddots & \vdots \\ & & & a_{N-1,N} \\ & & & \end{pmatrix}}_{=:\,R}. \tag{10.15}$$

Eine äquivalente Fixpunktformulierung für (10.14) lautet

$$\begin{aligned} a_{jj}x_j &= b_j - \sum_{\substack{k=1 \\ k\neq j}}^{N} a_{jk}x_k, \qquad j = 1,2,\dots,N, \\ &\rightsquigarrow \quad Dx = b - (L+R)x, \end{aligned}$$

und die zugehörige Fixpunktiteration ist in der folgenden Definition angegeben.

Definition 10.17. Für eine Matrix $A = (a_{jk}) \in \mathbb{R}^{N\times N}$ mit nichtverschwindenden Diagonaleinträgen, $a_{jj} \neq 0$ für $j = 1,2,\dots,N$, ist das *Gesamtschrittverfahren*, auch *Jacobi-Verfahren* genannt, zur Lösung des Gleichungssystems $Ax = b$ von der folgenden Form,

$$\left.\begin{aligned} x_j^{(n+1)} &= \frac{1}{a_{jj}}\Big(b_j - \sum_{\substack{k=1 \\ k\neq j}}^{N} a_{jk}x_k^{(n)}\Big), \quad j = 1,2,\dots,N, \quad n = 0,1,\dots, \\ &\rightsquigarrow \quad x^{(n+1)} = D^{-1}b - D^{-1}(L+R)x^{(n)}, \quad \text{\_\_\_\_\_«\_\_\_\_\_}. \end{aligned}\right\} \tag{10.16}$$

Die zugehörige Iterationsmatrix hat die Gestalt

$$\mathcal{H}_{\mathbf{Ges}} = -D^{-1}(L+R) = -\begin{pmatrix} 0 & a_{12}/a_{11} & \cdots & a_{1N}/a_{11} \\ a_{21}/a_{22} & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & a_{N-1,N}/a_{N-1,N-1} \\ a_{N1}/a_{NN} & \cdots & a_{N,N-1}/a_{NN} & 0 \end{pmatrix}.$$

Erste hinreichende Kriterien für die Konvergenz des Gesamtschrittverfahrens liefert das folgende Theorem.

Theorem 10.18. *Das Gesamtschrittverfahren ist durchführbar und konvergent, falls eine der beiden folgenden Bedingungen erfüllt ist,*

$$\sum_{\substack{k=1\\k\neq j}}^{N} |a_{jk}| \;<\; |a_{jj}|, \qquad j \;=\; 1,2,\ldots,N \quad \left(\begin{array}{ll} \Longleftrightarrow & A \text{ ist strikt diagonaldominant} \\ \Longleftrightarrow & \|\mathcal{H}_{\mathbf{Ges}}\|_\infty \;<\; 1 \end{array} \right);$$

oder

$$\sum_{\substack{j=1\\j\neq k}}^{N} |a_{jk}| \;<\; |a_{kk}|, \qquad k \;=\; 1,2,\ldots,N \quad \left(\Longleftrightarrow \quad A^\top \text{ ist strikt diagonaldominant} \right).$$

BEWEIS. Jede der genannten Bedingungen impliziert unmittelbar $a_{jj} \neq 0$ für $j = 1,2,\ldots,N$ und damit die Durchführbarkeit des Gesamtschrittverfahrens. Die erste Bedingung bedeutet $\|\mathcal{H}_{\mathbf{Ges}}\|_\infty < 1$, was $r_\sigma(\mathcal{H}_{\mathbf{Ges}}) < 1$ und somit die Konvergenz des Gesamtschrittverfahrens nach sich zieht. Die zweite Bedingung bedeutet $\|(L+R)D^{-1}\|_1 < 1$, wegen der Ähnlichkeit der Matrizen $\mathcal{H}_{\mathbf{Ges}}$ und $(L+R)D^{-1}$ folgt daraus $\sigma(\mathcal{H}_{\mathbf{Ges}}) = \sigma((L+R)D^{-1})$ beziehungsweise

$$r_\sigma(\mathcal{H}_{\mathbf{Ges}}) \;=\; r_\sigma((L+R)D^{-1}) \;\leq\; \|(L+R)D^{-1}\|_1 \;<\; 1,$$

was den Beweis des Theorems komplettiert. □

In wichtigen Anwendungen treten Matrizen auf[4], für die keine der beiden jeweils relativ starken Voraussetzungen von Theorem 10.18 erfüllt ist, wohl aber die schwächere Voraussetzung des folgenden Theorems, die ebenfalls die Konvergenz des Gesamtschrittverfahrens impliziert.

Theorem 10.19. *Für irreduzibel diagonaldominante Matrizen $A \in \mathbb{R}^{N\times N}$ ist das Gesamtschrittverfahren durchführbar und konvergent.*

BEWEIS. Würde ein Diagonaleintrag der Matrix A verschwinden, so wären aufgrund der vorausgesetzten Diagonaldominanz von A alle Einträge in der gleichen Zeile der Matrix A identisch null im Widerspruch zur vorausgesetzten Irreduzibilität von A. Für die Konvergenz des Verfahrens ist die Ungleichung $r_\sigma(\mathcal{H}_{\mathbf{Ges}}) < 1$ beziehungsweise für jede Zahl $\lambda \in \mathbb{C}$ mit $|\lambda| \geq 1$ die Regularität der Matrix $\mathcal{H}_{\mathbf{Ges}} - \lambda I$ nachzuweisen, wofür nach Theorem 10.16 die irreduzible Diagonaldominanz der Matrix $\mathcal{H}_{\mathbf{Ges}} - \lambda I$ hinreichend ist. Letzteres wird in den folgenden Beweisteilen (i) und (ii) nachgewiesen.

(i) Mit Lemma 10.14 erschließt man, dass mit der Matrix A auch die Matrizen $L+R$, $\mathcal{H}_{\mathbf{Ges}} = -D^{-1}(L+R)$ und $\mathcal{H}_{\mathbf{Ges}} - \lambda I$ irreduzibel sind.

(ii) Des Weiteren erfüllt die Matrix $(b_{jk}) := \mathcal{H}_{\mathbf{Ges}} - \lambda I$ auch die Bedingungen (10.13),

$$\sum_{\substack{k=1\\k\neq j}}^{N} \underbrace{\frac{|a_{jk}|}{|a_{jj}|}}_{|b_{jk}|} \;\overset{(*)}{\leq}\; 1 \;\leq\; |\lambda| \;=\; |b_{jj}|, \qquad j \;=\; 1,2,\ldots,N,$$

[4] zum Beispiel die Matrix A aus (9.11), die man nach Anwendung des Differenzenschemas auf das spezielle in Abschnitt 9.2 betrachtete Randwertproblem 2. Ordnung erhält

und für zumindest einen Index j liegt in $(*)$ Ungleichheit vor aufgrund der Voraussetzung an die Matrix A. Damit ist die Matrix $\mathcal{H}_{\mathbf{Ges}} - \lambda I$ in der Tat irreduzibel diagonaldominant, was den Beweis des Theorems komplettiert. □

10.5 Das Einzelschrittverfahren

Betrachtet man für das Gesamtschrittverfahren die komponentenweise Darstellung in (10.16), so ist es naheliegend, zur Berechnung von $x_j^{(n+1)}$ anstelle von $x_1^{(n)}, \ldots, x_{j-1}^{(n)}$ die bereits berechneten und vermeintlich besseren Approximationen $x_1^{(n+1)}, \ldots, x_{j-1}^{(n+1)}$ zu verwenden. Die zugehörige Fixpunktiteration ist in der folgenden Definition festgehalten.

Definition 10.20. Für eine Matrix $A = (a_{jk}) \in \mathbb{R}^{N\times N}$ mit $a_{jj} \neq 0$ für $j = 1, \ldots, N$ ist das *Einzelschrittverfahren*, auch *Gauß-Seidel-Verfahren* genannt, zur Lösung der Gleichung $Ax = b$ von der folgenden Form,

$$x_j^{(n+1)} = \frac{1}{a_{jj}}\Big(b_j - \sum_{k=1}^{j-1} a_{jk} x_k^{(n+1)} - \sum_{k=j+1}^{N} a_{jk} x_k^{(n)}\Big),\ j = 1, 2, \ldots, N, n = 0, 1, \ldots,$$

$$\text{bzw.} \qquad (D + L)x^{(n+1)} = b - Rx^{(n)}, \text{ ——— « ———}.$$

Die zugehörige Iterationsmatrix hat die Gestalt

$$\mathcal{H}_{\mathbf{Ein}} = -(D + L)^{-1} R.$$

10.5.1 Der Betrag einer Matrix

Für die Herleitung von Konvergenzergebnissen für das Einzelschrittverfahren werden einige weitere ordnungstheoretische Resultate für Matrizen benötigt, die im Folgenden vorgestellt werden.

Definition 10.21. Der Betrag einer Matrix $A = (a_{jk}) \in \mathbb{R}^{N\times N}$ ist durch $|A| := (|a_{jk}|) \in \mathbb{R}^{N\times N}$ definiert.

Lemma 10.22. *Für Matrizen $A, B \in \mathbb{R}^{N\times N}$ gelten die folgenden Aussagen:*

$$|A| \geq 0, \qquad |A + B| \leq |A| + |B|, \qquad |AB| \leq |A|\,|B|,$$
$$\|A\|_\infty = \||A|\|_\infty = \||A|\mathbf{e}\|_\infty, \quad \textit{mit } \mathbf{e} = (1, \ldots, 1)^\top \in \mathbb{R}^N,$$
$$|A| \leq |B| \implies \|A\|_\infty \leq \|B\|_\infty.$$

BEWEIS. Es soll exemplarisch nur der Nachweis für die dritte Abschätzung in der ersten Zeile der Aussage geführt werden. Hierzu setzt man $A = (a_{jk})$, $B = (b_{jk})$, $|AB| = (c_{jk})$, $|A||B| = (d_{jk}) \in \mathbb{R}^{N\times N}$ und erhält

$$c_{jk} \;=\; \Big|\sum_{\ell=1}^{N} a_{j\ell}b_{\ell k}\Big| \;\le\; \sum_{\ell=1}^{N} |a_{j\ell}||b_{\ell k}| \;=\; d_{jk}.$$

Die Beweise der anderen Aussagen sind ähnlich einfach und werden wie bereits angekündigt hier nicht geführt. □

Insbesondere ist die Abbildung $|\cdot| : \mathbb{R}^{N\times N} \to \mathbb{R}^{N\times N}$ stetig. Die folgende Eigenschaft stellt eine Variante von Lemma 9.9 dar und wird im Beweis des nächsten Theorems benötigt.

Lemma 10.23. *Für Matrizen $S,\ T \in \mathbb{R}^{N\times N}$ und $\lambda \in \mathbb{R}$ gilt die Implikation*

$$|S| \le T, \quad 1 > r_\sigma(T) \implies I - S \text{ regulär}, \quad |(I-S)^{-1}| \le (I-T)^{-1}. \quad (10.17)$$

BEWEIS. Es sind die Reihen $\sum_{\nu=0}^{\infty} S^\nu$ und $\sum_{\nu=0}^{\infty} |S|^\nu$ konvergent, denn unter Anwendung von Lemma 10.22 erhält man

$$\begin{aligned}\Big\|\sum_{\nu=n_0}^{n_1} S^\nu\Big\|_\infty \;&=\; \Big\|\Big|\sum_{\nu=n_0}^{n_1} S^\nu\Big|\Big\|_\infty \;\le\; \Big\|\sum_{\nu=n_0}^{n_1} |S|^\nu\Big\|_\infty \\ &\le\; \Big\|\sum_{\nu=n_0}^{n_1} T^\nu\Big\|_\infty \;\to\; 0 \qquad \text{für } n_0 \le n_1, \qquad n_0,\, n_1 \to \infty.\end{aligned}$$

Daher ist die Matrix $I - S$ regulär und es gilt die Identität $(I-S)^{-1} = \sum_{\nu=0}^{\infty} S^\nu$, so dass man schließlich die Aussage (10.17) erhält,

$$|(I-S)^{-1}| \;\le\; \sum_{\nu=0}^{\infty} |S|^\nu \;\le\; \sum_{\nu=0}^{\infty} T^\nu \;=\; (I-T)^{-1}.$$ □

Entsprechend der (partiellen) Ordnung für Matrizen wird noch eine Ordnung für Vektoren eingeführt. Für zwei Vektoren $x = (x_j),\ y = (y_j) \in \mathbb{R}^N$ schreibt man

$$x \le y \qquad :\Longleftrightarrow \qquad x_j \le y_j \qquad \text{für } j = 1, 2, \ldots, N.$$

Synonym für $x \le y$ wird die Notation $y \ge x$ verwendet. Ein Vektor $x \in \mathbb{R}^N$ heißt *nichtnegativ*, falls $x \ge 0$ gilt.

10.5.2 Konvergenzergebnisse für das Einzelschrittverfahren

Ein erstes hinreichendes Kriterium für die Konvergenz des Einzelschrittverfahrens liefert das folgende Theorem.

Theorem 10.24. *Für jede strikt diagonaldominante Matrix $A \in \mathbb{R}^{N\times N}$ gilt*

$$\|\mathcal{H}_{\mathbf{Ein}}\|_\infty \;\le\; \|\mathcal{H}_{\mathbf{Ges}}\|_\infty \;<\; 1.$$

BEWEIS. Die strikte Diagonaldominanz der Matrix A ist offensichtlich äquivalent zu $\|\mathcal{H}_{\mathbf{Ges}}\|_\infty < 1$ (siehe Theorem 10.18). Im Folgenden wird die zweite Ungleichung in

$$0 \;\le\; |\mathcal{H}_{\mathbf{Ein}}|\mathbf{e} \;\le\; \|\mathcal{H}_{\mathbf{Ges}}\|_\infty \mathbf{e} \tag{10.18}$$

mit $\mathbf{e} = (1,\dots,1)^\top$ nachgewiesen (die erste Ungleichung in (10.18) ist offensichtlich richtig), woraus dann $\|\mathcal{H}_{\mathbf{Ein}}\|_\infty = \||\mathcal{H}_{\mathbf{Ein}}|\mathbf{e}\|_\infty \le \|\mathcal{H}_{\mathbf{Ges}}\|_\infty$ und somit die Aussage des Theorems folgt. Für den Nachweis der zweiten Ungleichung von (10.18) beobachtet man

$$\begin{aligned}
\mathcal{H}_{\mathbf{Ges}} &= -D^{-1}(L+R) \qquad \rightsquigarrow \qquad |\mathcal{H}_{\mathbf{Ges}}| = |D^{-1}L| + |D^{-1}R|,\\
\mathcal{H}_{\mathbf{Ein}} &= -(I + D^{-1}L)^{-1}D^{-1}R,
\end{aligned}$$

und Lemma 10.23 angewandt mit $S = -D^{-1}L$ und $T = |D^{-1}L|$ liefert (in der Abschätzung $(\bullet)$)

$$\begin{aligned}
|\mathcal{H}_{\mathbf{Ein}}| &\le |(I+D^{-1}L)^{-1}||D^{-1}R| \overset{(\bullet)}{\le} (I-|D^{-1}L|)^{-1}\overbrace{|\mathcal{H}_{\mathbf{Ges}}|-|D^{-1}L|}\,|D^{-1}R|\\
&= (I-|D^{-1}L|)^{-1}\big(|\mathcal{H}_{\mathbf{Ges}}| - I + (I-|D^{-1}L|)\big)\\
&= I + (I-|D^{-1}L|)^{-1}(|\mathcal{H}_{\mathbf{Ges}}| - I)
\end{aligned}$$

und daher

$$\begin{aligned}
|\mathcal{H}_{\mathbf{Ein}}|\mathbf{e} &\le \mathbf{e} + (I-|D^{-1}L|)^{-1}(\overbrace{|\mathcal{H}_{\mathbf{Ges}}|\mathbf{e}}^{\le \|\mathcal{H}_{\mathbf{Ges}}\|_\infty \mathbf{e}} - \mathbf{e})\\
&= \mathbf{e} + \underbrace{(\|\mathcal{H}_{\mathbf{Ges}}\|_\infty - 1)}_{\le 0}\underbrace{(I-|D^{-1}L|)^{-1}}_{\overset{(*)}{\ge} I}\mathbf{e}\\
&\le \mathbf{e} + (\|\mathcal{H}_{\mathbf{Ges}}\|_\infty - 1)\mathbf{e} = \|\mathcal{H}_{\mathbf{Ges}}\|_\infty \mathbf{e},
\end{aligned}$$

was (10.18) bedeutet und den Beweis komplettiert. Die Ungleichung $(*)$ folgt dabei mit Lemma 10.23 angewandt mit $S = 0$ und $T = |D^{-1}L|$. □

Bemerkung 10.25. Für strikt diagonaldominante Matrizen $A \in \mathbb{R}^{N\times N}$ kann man nach Theorem 10.24 bei der Anwendung des Einzelschrittverfahrens eine schnellere Konvergenz als bei der Anwendung des Gesamtschrittverfahrens erwarten. △

Ein weiteres hinreichendes Kriterium für die Konvergenz des Einzelschrittverfahrens liefert das folgende Theorem[5].

Theorem 10.26. *Für irreduzibel diagonaldominante Matrizen $A \in \mathbb{R}^{N\times N}$ ist das Einzelschrittverfahren durchführbar und konvergent.*

[5] siehe die Anmerkungen vor Theorem 10.19

BEWEIS. Das Nichtverschwinden der Diagonaleinträge von A (was die Durchführbarkeit des Einzelschrittverfahrens gewährleistet) ist bereits in Theorem 10.19 gezeigt worden. Für die Konvergenz des Einzelschrittverfahrens ist die Ungleichung $r_\sigma(\mathscr{H}_{\mathbf{Ein}}) < 1$ beziehungsweise für jede Zahl $\lambda \in \mathbb{C}$ mit $|\lambda| \geq 1$ die Regularität der Matrix $\mathscr{H}_{\mathbf{Ein}} - \lambda I$ nachzuweisen. Hierzu sei in Erinnerung gerufen, dass $\mathscr{H}_{\mathbf{Ein}} = -(D+L)^{-1}R$ gilt, und die Regularität der Matrix $\mathscr{H}_{\mathbf{Ein}} - \lambda I$ ist damit äquivalent zur Regularität von $\lambda D + \lambda L + R$. Für die letztgenannte Eigenschaft ist nach Theorem 10.16 die irreduzible Diagonaldominanz der Matrix $\lambda D + \lambda L + R$ hinreichend, die im Folgenden nachgewiesen wird.

(i) Mit Lemma 10.14 erschließt man, dass mit A auch die Matrix $\lambda D + \lambda L + R$ irreduzibel ist.

(ii) Weiter erfüllt die Matrix $(b_{jk}) := \lambda D + \lambda L + R$ auch die Bedingungen (10.13),

$$\sum_{k=1}^{j-1} |\lambda a_{jk}| + \sum_{k=j+1}^{N} |a_{jk}| \leq |\lambda| \sum_{\substack{k=1 \\ k\neq j}}^{N} |a_{jk}| \overset{(*)}{\leq} |\lambda||a_{jj}|, \quad j = 1,2,\ldots,N,$$

und in $(*)$ gilt nach Voraussetzung an die Matrix A für ein Index j die Ungleichheit. Damit ist die Matrix $\lambda D + \lambda L + R$ in der Tat irreduzibel diagonaldominant, was den Beweis komplettiert. □

10.6 Das Relaxationsverfahren und erste Konvergenzresultate

Im Folgenden wird das Relaxationsverfahren vorgestellt. Hier ist im $(n+1)$-ten Iterationsschritt die Vorgehensweise so, dass – ausgehend von dem Vektor $x^{(n)}$ – für die Indizes $j = 1,\ldots,N$ jeweils zunächst hilfsweise die Zahl $\widehat{x}_j^{(n+1)}$ gemäß der Vorschrift des Einzelschrittverfahrens ermittelt wird aus den bereits berechneten Werten $x_1^{(n+1)},\ldots,x_{j-1}^{(n+1)}$, $x_{j+1}^{(n)},\ldots,x_N^{(n)}$, und anschließend berechnet sich $x_j^{(n+1)}$ als eine durch einen Parameter $\omega \in \mathbb{R}$ festgelegte Linearkombination von $\widehat{x}_j^{(n+1)}$ und $x_j^{(n)}$. Im Einzelnen sieht die Iterationsvorschrift folgendermaßen aus,

$$\left.\begin{aligned} \widehat{x}_j^{(n+1)} &= \frac{1}{a_{jj}}\Big(b_j - \sum_{k=1}^{j-1} a_{jk}x_k^{(n+1)} - \sum_{k=j+1}^{N} a_{jk}x_k^{(n)}\Big) \\ x_j^{(n+1)} &= \omega\,\widehat{x}_j^{(n+1)} + (1-\omega)x_j^{(n)} \end{aligned}\right\}, \quad \begin{aligned} &j = 1,2,\ldots,N \\ &(n = 0,1,\ldots). \end{aligned}$$

Die zugehörige Fixpunktiteration ist in der folgenden Definition angegeben.

Definition 10.27. Für eine Matrix $A = (a_{jk}) \in \mathbb{R}^{N\times N}$ mit nichtverschwindenden Diagonaleinträgen, $a_{jj} \neq 0$ für $j = 1,2,\ldots,N$, ist das *Relaxationsverfahren* zur

Lösung der Gleichung $Ax = b$ von der folgenden Form,

$$a_{jj}x_j^{(n+1)} = \omega\big(b_j - \sum_{k=1}^{j-1} a_{jk}x_k^{(n+1)} - \sum_{k=j+1}^{N} a_{jk}x_k^{(n)}\big) + (1-\omega)a_{jj}x_j^{(n)},$$
$$\text{für } j = 1,2,\ldots,N \qquad (n = 0,1,\ldots),$$

beziehungsweise in Matrix-Vektor-Schreibweise

$$(D + \omega L)x^{(n+1)} = \omega b + ((1-\omega)D - \omega R)x^{(n)} \qquad (n = 0,1,\ldots).$$

Die zugehörige Iterationsmatrix hat die Gestalt

$$\mathcal{H}(\omega) = (D + \omega L)^{-1}((1-\omega)D - \omega R). \tag{10.19}$$

Für $\omega = 1$ stimmt das Relaxationsverfahren mit dem Einzelschrittverfahren überein, $\mathcal{H}(1) = \mathcal{H}_{\mathbf{Ein}}$. Für $\omega < 1$ spricht man von *Unter-*, im Falle $\omega > 1$ von *Überrelaxation.*

In dem vorliegenden Abschnitt werden für zwei Klassen von Matrizen allgemeine Konvergenzresultate zum Relaxationsverfahren hergeleitet. Eine optimale Wahl des Relaxationsparameters ω wird dabei nicht weiter diskutiert. Die erzielten Resultate sind aber bereits für den Fall $\omega = 1$ (Einzelschrittverfahren) von Interesse.

Bemerkung 10.28. Eine besondere Bedeutung erlangt das Relaxationsverfahren für die spezielle Klasse der konsistent geordneten Matrizen A, die im nächsten Abschnitt 10.7 behandelt werden. Für solche Matrizen A lässt sich der Spektralradius der Iterationsmatrix $\mathcal{H}(\omega)$ als Funktion des Relaxationsparameters ω genau ermitteln beziehungsweise die Wahl von ω optimieren. △

Für allgemeine Matrizen $A \in \mathbb{R}^{N\times N}$ mit nichtverschwindenden Diagonalelementen gilt das folgende Resultat, mit dem sich die Wahl vernünftiger Relaxationsparameter schnell einschränken lässt.

Theorem 10.29 (Kahan). *Für die Iterationsmatrix des Relaxationsverfahrens gilt*

$$r_\sigma(\mathcal{H}(\omega)) \geq |\omega - 1|, \qquad \omega \in \mathbb{R}.$$

BEWEIS. Mit der Bezeichnung $\lambda_1,\ldots,\lambda_N \in \mathbb{C}$ für die entsprechend ihrer Vielfachheit gezählten Eigenwerte von $\mathcal{H}(\omega)$ gilt aufgrund der Darstellung (10.19) Folgendes,

$$\prod_{j=1}^{N} \lambda_j = \det \mathcal{H}(\omega) = \underbrace{\det(I - \omega D^{-1}L)^{-1}}_{=\,1} \det\big((1-\omega)I - \omega D^{-1}R\big) = (1-\omega)^N,$$

so dass notwendigerweise $|\lambda_j| \geq |1-\omega|$ für mindestens einen Index $1 \leq j \leq N$ gilt. □

Korollar 10.30. *Das Relaxationsverfahren ist höchstens für* $0 < \omega < 2$ *konvergent.*

BEWEIS. Für $\omega \notin (0, 2)$ gilt nach Theorem 10.29 die Ungleichung $r_\sigma(\mathcal{H}(\omega)) \geq 1$, so dass nach Theorem 10.3 keine Konvergenz vorliegen kann. □

Ein erstes hinreichendes Kriterium für die Konvergenz des Relaxationsverfahrens liefert das folgende Theorem.

Theorem 10.31 (Ostrowski, Reich). *Für eine symmetrische, positiv definite Matrix* $A \in \mathbb{R}^{N\times N}$ *ist das zugehörige Relaxationsverfahren für jeden Wert* $0 < \omega < 2$ *durchführbar und konvergent,*

$$r_\sigma(\mathcal{H}(\omega)) \;<\; 1 \qquad \textit{für} \quad 0 < \omega < 2.$$

BEWEIS. Aufgrund der Definitheit der Matrix A gilt $a_{jj} = \mathbf{e}_j^\top A \mathbf{e}_j > 0$ für alle j, was insbesondere die Durchführbarkeit des Relaxationsverfahrens nach sich zieht. Für den Nachweis der Konvergenz berechnet man zunächst

$$\begin{aligned}\mathcal{H}(\omega) &= I - \omega(D + \omega L)^{-1}A = I - \big(\tfrac{1}{\omega}D + L\big)^{-1}A \\ &= I - 2\big(2A^{-1}\big(\tfrac{1}{\omega}D + L\big)\big)^{-1} = I - 2(Q + I)^{-1} = (Q - I)(Q + I)^{-1}, \\ &\qquad \text{mit} \quad Q := 2A^{-1}\big(\tfrac{1}{\omega}D + L\big) - I.\end{aligned}$$

Im Folgenden wird

$$\sigma(Q) \subset \{\lambda \in \mathbb{C} : \operatorname{Re}\lambda > 0\} \tag{10.20}$$

nachgewiesen. Wegen

$$\sigma(\mathcal{H}(\omega)) = \Big\{\frac{\lambda - 1}{\lambda + 1} : \lambda \in \sigma(Q)\Big\}$$

und

$$\Big|\frac{\lambda - 1}{\lambda + 1}\Big|^2 = \frac{(\operatorname{Re}\lambda - 1)^2 + (\operatorname{Im}\lambda)^2}{(\operatorname{Re}\lambda + 1)^2 + (\operatorname{Im}\lambda)^2} < 1 \qquad \text{für} \quad \operatorname{Re}\lambda > 0$$

erhält man dann die Aussage des Theorems. Für den Nachweis von (10.20) betrachtet man $\lambda \in \mathbb{C}$ und $0 \neq x \in \mathbb{C}^N$ mit $Qx = \lambda x$ und erhält zunächst

$$\lambda Ax = 2\big(\tfrac{1}{\omega}D + L\big)x - Ax.$$

Skalare Multiplikation mit dem Vektor x liefert

$$\begin{aligned}(\operatorname{Re}\lambda)\overbrace{x^{\mathrm{H}}Ax}^{>0} &= 2\operatorname{Re} x^{\mathrm{H}}\big(\tfrac{1}{\omega}D + L\big)x - x^{\mathrm{H}}Ax \\ &= x^{\mathrm{H}}\big(\tfrac{2}{\omega}D + L + \underbrace{L^{\mathrm{H}}}_{=R}\big)x - x^{\mathrm{H}}(D + L + R)x = \overbrace{\big(\tfrac{2}{\omega} - 1\big)x^{\mathrm{H}}Dx}^{>0,\ \text{da } a_{jj}>0\ \forall\ j},\end{aligned}$$

und daraus folgt $\operatorname{Re}\lambda > 0$. □

10.6.1 M-Matrizen

Im Folgenden wird eine weitere Klasse von Matrizen vorgestellt, bei denen das Relaxationsverfahren einsetzbar ist.

Definition 10.32. Eine Matrix $A = (a_{jk}) \in \mathbb{R}^{N\times N}$ heißt *M-Matrix*, falls Folgendes gilt:

(a) Die Matrix A ist regulär und besitzt eine Inverse mit ausschließlich nichtnegativen Einträgen, $A^{-1} \geq 0$.

(b) Alle Einträge der Matrix A außer denen auf der Diagonalen sind nichtpositiv, $a_{jk} \leq 0$ für alle Indizes j, k mit $j \neq k$.

M-Matrizen lassen sich folgendermaßen charakterisieren:

Theorem 10.33. *Für eine Matrix $A = (a_{jk}) \in \mathbb{R}^{N\times N}$ gilt die folgende Äquivalenz,*

$$A \text{ ist M-Matrix} \quad \Longleftrightarrow \quad \left\{ \begin{array}{ll} a_{jj} > 0 & \text{für } j = 1,\ldots,N, \\ a_{jk} \leq 0 & \text{für alle } j,\ k \text{ mit } j \neq k, \\ r_\sigma(D^{-1}(L+R)) < 1, & \end{array} \right\} \tag{10.21}$$

mit der Zerlegung $A = D + L + R$ in Diagonal-, unteren und oberen Anteil entsprechend (10.15). Die Inverse jeder M-Matrix A besitzt die nichtnegative neumannsche Reihenentwicklung

$$A^{-1} = \sum_{\nu=0}^{\infty} \underbrace{\big(-D^{-1}(L+R)\big)^{\nu}}_{\geq 0} \underbrace{D^{-1}}_{\geq 0} \geq 0. \tag{10.22}$$

BEWEIS. "$\Longleftarrow$" Mit der Identität $I - D^{-1}A = -D^{-1}(L+R)$ und den Voraussetzungen

$$D \text{ regulär}, \quad D^{-1} \geq 0, \quad -(L+R) \geq 0, \qquad r_\sigma(-D^{-1}(L+R)) < 1,$$

erhält man unter Anwendung von Theorem 9.13 die Regularität der Matrix A sowie die nichtnegative neumannsche Reihenentwicklung (10.22) für die Inverse A^{-1}, womit die Richtung "$\Longleftarrow$" nachgewiesen ist. Für den Nachweis der anderen Implikation "$\Longrightarrow$" sei nun A eine M-Matrix. Wenn $a_{kk} \leq 0$ für ein $k \in \{1,\ldots,N\}$ gilt, so erhält man für den Vektor $a^{(k)} = (a_{jk})_j \in \mathbb{R}^N$ die Ungleichung $a^{(k)} \leq 0$ und daraus den Widerspruch k-ter Einheitsvektor $\mathbf{e}_k = A^{-1}a^{(k)} \leq 0$. Für den Nachweis der Ungleichung $r_\sigma(B) < 1$ mit $B := -D^{-1}(L+R)$ stellt man Folgendes fest,

$$B \geq 0, \qquad I - B = D^{-1}A \text{ regulär}, \qquad (I-B)^{-1} = A^{-1}D \geq 0,$$

und Theorem 9.17 liefert die behauptete Ungleichung $r_\sigma(B) < 1$. □

Beispiel 10.34. Die Matrix zu dem in Abschnitt 10.2.1 vorgestellten Modellbeispiel ist eine M-Matrix, denn als irreduzibel diagonaldominante Matrix gilt für sie nach Theorem 10.19 die Ungleichung $r_\sigma(D^{-1}(L+R)) < 1$. △

Theorem 10.35. *Für eine M-Matrix $A \in \mathbb{R}^{N\times N}$ ist das Relaxationsverfahren durchführbar und für jeden Parameter $0 < \omega \leq 1$ konvergent,*

$$r_\sigma(\mathcal{H}(\omega)) \;<\; 1 \qquad \textit{für} \quad 0 < \omega \leq 1.$$

BEWEIS. Die Durchführbarkeit ist aufgrund des Nichtverschwindens der Diagonaleinträge der Matrix A (siehe Theorem 10.33) gewährleistet. Im Folgenden wird

$$\mathcal{H}(\omega) \;\geq\; 0, \qquad I - \mathcal{H}(\omega) \quad \text{regulär}, \qquad (I - \mathcal{H}(\omega))^{-1} \;\geq\; 0, \tag{10.23}$$

nachgewiesen. Die Aussage des Theorems erhält man dann unmittelbar mit Theorem 9.17. Nach Voraussetzung gilt (mit der Zerlegung $D + L + R = A$ aus (10.15))

$$D \text{ regulär}, \quad D \;\geq\; 0, \quad D^{-1} \;\geq\; 0, \quad R \;\leq\; 0, \quad L \;\leq\; 0.$$

Damit ist insbesondere die Matrix $D + \omega L$ regulär, und die Eigenschaft $\mathcal{H}(\omega) \geq 0$ resultiert dann aus $(D + \omega L)^{-1} \geq 0$, was man wie folgt einsieht,

$$\begin{aligned} -\omega D^{-1}L \;&\geq\; 0, \qquad \sigma(-\omega D^{-1}L) \;=\; \{0\}, \\ &\overset{(*)}{\rightsquigarrow} \qquad \big(I \,-\, (-\omega D^{-1}L)\big)^{-1} \;=\; (I \,+\, \omega D^{-1}L)^{-1} \;\geq\; 0, \end{aligned}$$

wobei man die Schlussfolgerung $(*)$ mit Theorem 9.17 erhält. Die beiden anderen Aussagen in (10.23) ergeben sich folgendermaßen,

$$\begin{aligned} I \,-\, \mathcal{H}(\omega) \;&=\; (D + \omega L)^{-1}\big(D \,+\, \omega L - (1-\omega)D \,+\, \omega R\big) \;=\; \overbrace{\omega(D + \omega L)^{-1}A}^{\text{regulär}}, \\ (I \,-\, \mathcal{H}(\omega))^{-1} \;&=\; \tfrac{1}{\omega}A^{-1}(D + \omega L) \;=\; \tfrac{1}{\omega}A^{-1}\big(A \,-\, (1-\omega)L \,-\, R\big) \\ &=\; \tfrac{1}{\omega}I \,+\, \underbrace{\tfrac{1}{\omega}A^{-1}}_{\geq\, 0}\underbrace{\big(-(1-\omega)L \,-\, R\big)}_{\geq\, 0} \;\geq\; 0. \end{aligned}$$

Dies komplettiert den Beweis von Theorem 10.35. □

Bemerkung 10.36. Beim Relaxationsverfahren für M-Matrizen gelten spezieller die Abschätzungen $r_\sigma(\mathcal{H}(\omega_2)) \leq r_\sigma(\mathcal{H}(\omega_1)) < 1$ für $0 < \omega_1 \leq \omega_2 \leq 1$ (Aufgabe 10.10), so dass innerhalb des Parameterintervalls $0 < \omega \leq 1$ die Wahl $\omega = 1$ optimal ist. △

10.7 Das Relaxationsverfahren für konsistent geordnete Matrizen

Es soll nun noch eine Klasse von Matrizen behandelt werden, bei denen sich der Spektralradius der zugehörigen Iterationsmatrix $\mathcal{H}(\omega)$ als Funktion des Relaxationsparameters ω genau ermitteln beziehungsweise die Wahl von ω optimieren lässt.

Definition 10.37. Eine Matrix $A = (a_{jk}) \in \mathbb{R}^{N\times N}$ mit $a_{jj} \neq 0$ für alle j heißt *konsistent geordnet*, falls die Eigenwerte der Matrix

$$\mathcal{J}(\alpha) := \alpha D^{-1}L + \alpha^{-1}D^{-1}R \quad \in \quad \mathbb{C}^{N\times N}, \qquad 0 \neq \alpha \in \mathbb{C}, \tag{10.24}$$

unabhängig von α sind, wenn also die Identität $\sigma(\mathcal{J}(\alpha)) = \sigma(\mathcal{J}(1))$ gilt für $0 \neq \alpha \in \mathbb{C}$. Hierbei bezeichnet $A = D + L + R$ die Zerlegung in Diagonal-, unteren und oberen Anteil entsprechend (10.15).

Beispiel 10.38. Eine Block-Tridiagonalmatrix

$$A = \begin{pmatrix} D_1 & C_1 & & \\ B_1 & \ddots & \ddots & \\ & \ddots & \ddots & C_{M-1} \\ & & B_{M-1} & D_M \end{pmatrix} \in \mathbb{R}^{N\times N}$$

mit regulären Diagonalmatrizen $D_k \in \mathbb{R}^{N_k\times N_k}$, $k = 1, 2, \ldots, M$ (mit $\sum_{k=1}^{M} N_k = N$) ist konsistent geordnet. (Die Nebendiagonalmatrizen seien hierbei von entsprechender Dimension, es gilt also $B_k \in \mathbb{R}^{N_{k+1}\times N_k}$ und $C_k \in \mathbb{R}^{N_k\times N_{k+1}}$ für $k = 1, 2, \ldots, M-1$.) △

BEWEIS. Hier gilt

$$D^{-1}L = \begin{pmatrix} 0 & & & \\ D_2^{-1}B_1 & \ddots & & \\ & \ddots & \ddots & \\ & & D_M^{-1}B_{M-1} & 0 \end{pmatrix}, \quad D^{-1}R = \begin{pmatrix} 0 & D_1^{-1}C_1 & & \\ & \ddots & \ddots & \\ & & \ddots & D_{M-1}^{-1}C_{M-1} \\ & & & 0 \end{pmatrix},$$

und somit

$$\mathcal{J}(\alpha) = \begin{pmatrix} 0 & \alpha^{-1}D_1^{-1}C_1 & & \\ \alpha D_2^{-1}B_1 & \ddots & \ddots & \\ & \ddots & \ddots & \alpha^{-1}D_{M-1}^{-1}C_{M-1} \\ & & \alpha D_M^{-1}B_{M-1} & 0 \end{pmatrix} \in \mathbb{R}^{N\times N}.$$

Mit einer geeigneten Transformationsmatrix S_α von Diagonalgestalt erhält man schließlich die Ähnlichkeit der Matrizen $\mathcal{J}(1)$ und $\mathcal{J}(\alpha)$:

$$\mathcal{S}_\alpha := \begin{pmatrix} \alpha^0 I_{N_1} & & & \\ & \alpha^1 I_{N_2} & & \\ & & \ddots & \\ & & & \alpha^{M-1} I_{N_M} \end{pmatrix}$$

$$\rightsquigarrow \quad \mathcal{S}_\alpha \mathcal{J}(1) = \begin{pmatrix} 0 & \alpha^0 D_1^{-1} C_1 & & \\ \alpha D_2^{-1} B_1 & \ddots & \ddots & \\ & \ddots & \ddots & \alpha^{M-2} D_{M-1}^{-1} C_{M-1} \\ & & \alpha^{M-1} D_M^{-1} B_{M-1} & 0 \end{pmatrix}$$

beziehungsweise $\mathcal{S}_\alpha \mathcal{J}(1) \mathcal{S}_\alpha^{-1} = \mathcal{J}(\alpha)$. □

Beispiel 10.39. Die Matrix aus dem Modellbeispiel in Abschnitt 10.2.1 ist konsistent geordnet (Aufgabe 10.14). △

Das folgende Theorem 10.41 stellt eine Beziehung her zwischen den Eigenwerten von $\mathcal{H}_{\mathbf{Ges}} = -D^{-1}(L + R)$ und denen von $\mathcal{H}(\omega)$. Zuvor wird die folgende Eigenschaft konsistent geordneter Matrizen festgehalten:

Lemma 10.40. *Bei konsistent geordneten Matrizen $A \in \mathbb{R}^{N \times N}$ liegt die Menge der Eigenwerte $\sigma(\mathcal{H}_{\mathbf{Ges}}) \subset \mathbb{C}$ der zum Gesamtschrittverfahren gehörenden Iterationsmatrix $\mathcal{H}_{\mathbf{Ges}}$ symmetrisch zum Ursprung, es gilt also $\sigma(\mathcal{H}_{\mathbf{Ges}}) = \sigma(-\mathcal{H}_{\mathbf{Ges}})$.*

BEWEIS. Mit der Notation (10.24) gilt $\mathcal{J}(1) = -\mathcal{H}_{\mathbf{Ges}}$ und $\mathcal{J}(-1) = \mathcal{H}_{\mathbf{Ges}}$, woraus die Aussage unmittelbar folgt. □

Theorem 10.41. *Die Matrix $A \in \mathbb{R}^{N \times N}$ sei konsistent geordnet, und sei $0 \neq \omega \in \mathbb{R}$. Weiter sei $0 \neq \lambda \in \mathbb{C}$ eine beliebige Zahl und $\sqrt{\lambda} \in \mathbb{C}$ eine der beiden Wurzeln von λ. Dann gilt die folgende Äquivalenz:*

$$\lambda \in \sigma(\mathcal{H}(\omega)) \qquad \Longleftrightarrow \qquad \frac{\lambda + \omega - 1}{\omega \sqrt{\lambda}} \in \sigma(\mathcal{H}_{\mathbf{Ges}}). \tag{10.25}$$

BEWEIS. Sei $0 \neq \lambda \in \mathbb{C}$ und $\sqrt{\lambda} \in \mathbb{C}$ eine der beiden Wurzeln von λ. Es gilt dann

$$\begin{aligned}
\lambda I - \mathcal{H}(\omega) &= (D+\omega L)^{-1}\big(\lambda(D+\omega L) - (1-\omega)D + \omega R\big) \\
&= (D+\omega L)^{-1}\big((\lambda+\omega-1)D + \omega(\lambda L + R)\big) \\
&= (I+\omega D^{-1}L)^{-1}\big((\lambda+\omega-1)I + \omega\lambda^{1/2}\big(\lambda^{1/2}D^{-1}L + \lambda^{-1/2}D^{-1}R\big)\big) \\
&= \omega\lambda^{1/2}\underbrace{(I+\omega D^{-1}L)^{-1}}_{\text{regulär}}\Big(\frac{\lambda+\omega-1}{\omega\lambda^{1/2}}I + \big(\lambda^{1/2}D^{-1}L + \lambda^{-1/2}D^{-1}R\big)\Big)
\end{aligned}$$

beziehungsweise

$$\begin{aligned}
&\lambda \in \sigma(\mathcal{H}(\omega)) \iff \\
&\quad \frac{\lambda+\omega-1}{\omega\lambda^{1/2}} \in \sigma\big(\underbrace{-\lambda^{1/2}D^{-1}L - \lambda^{-1/2}D^{-1}R}_{= \mathcal{J}(-\lambda^{1/2})}\big) = \sigma\big(\underbrace{-D^{-1}L - D^{-1}R}_{= \mathcal{J}(-1) = \mathcal{H}_{\mathbf{Ges}}}\big),
\end{aligned}$$

was mit der im Theorem angegebenen Äquivalenz übereinstimmt. □

Korollar 10.42 (Der Fall $\omega = 1$). *Für jede konsistent geordnete Matrix $A \in \mathbb{R}^{N\times N}$ gilt*

$$r_\sigma(\mathcal{H}_{\mathbf{Ein}}) = r_\sigma(\mathcal{H}_{\mathbf{Ges}})^2.$$

Für eine konsistent geordnete Matrix $A \in \mathbb{R}^{N\times N}$ sind demnach Gesamt- und Einzelschrittverfahren entweder beide konvergent oder divergent, und im Fall der Konvergenz ist das Einzelschrittverfahren doppelt so schnell wie das Gesamtschrittverfahren.

Mit dem folgenden Theorem wird das Verhalten von $r_\sigma(\mathcal{H}(\omega))$ in Abhängigkeit von ω beschrieben. Eine entsprechende Veranschaulichung liefert Bild 10.2 auf Seite 290.

Theorem 10.43. *Die Matrix $A \in \mathbb{R}^{N\times N}$ sei konsistent geordnet, und die Eigenwerte der Matrix $\mathcal{H}_{\mathbf{Ges}} = -D^{-1}(L+R)$ seien allesamt reell und betragsmäßig kleiner als eins, es sei also $\sigma(D^{-1}(L+R)) \subset (-1,1)$ erfüllt. Dann gilt*

$$r_\sigma(\mathcal{H}(\omega)) = \begin{cases} \frac{1}{4}\big(\omega\varrho_{\mathbf{Ges}} + \sqrt{\omega^2\varrho_{\mathbf{Ges}}^2 - 4(\omega-1)}\big)^2, & 0 < \omega \leq \omega_*, \\ \omega - 1, & \omega_* \leq \omega \leq 2, \end{cases}$$

mit $\varrho_{\mathbf{Ges}} := r_\sigma(D^{-1}(L+R))$ *und* $\omega_* := \dfrac{2}{1+\sqrt{1-\varrho_{\mathbf{Ges}}^2}}$.

BEWEIS. Sei $0 < \omega \leq 2$ mit $\omega \neq 1$ fest gewählt.[6]

[6] Die Situation $\omega = 1$ ist bereits mit Korollar 10.42 abgeklärt.

(a) In einem ersten Schritt werden (vergleiche Theorem 10.41) für jede Zahl $\mu \in \mathbb{R}$ die Lösungen $\lambda \in \mathbb{C}$ der Gleichung

$$\lambda - \omega\mu\sqrt{\lambda} + \omega - 1 = 0, \tag{10.26}$$

bestimmt. In der Tat besitzt die Gleichung (10.26) zwei Lösungen $\lambda_{1/2} = \lambda_{1/2}(\mu) \in \mathbb{C}$, für die entsprechend der Annahme $\omega \neq 1$ notwendigerweise $\lambda_{1/2} \neq 0$ gilt. Explizite Darstellungen sind

$$\left.\begin{aligned} \lambda_{1/2} &:= \tfrac{1}{4}\big(\omega\mu \pm \sqrt{\omega^2\mu^2 - 4(\omega-1)}\big)^2, \\ \sqrt{\lambda_{1/2}} &:= \tfrac{1}{2}\big(\text{———— « ————}\big), \end{aligned}\right\} \tag{10.27}$$

und daraus erhält man

$$|\lambda_{1/2}| = \left\{\begin{array}{ll} \tfrac{1}{4}\big(\omega\mu \pm \sqrt{\omega^2\mu^2 - 4(\omega-1)}\big)^2, & \mu^2 \geq \dfrac{4(\omega-1)}{\omega^2}, \\ \omega - 1, & \mu^2 < \text{—— « ——}, \end{array}\right\} \tag{10.28}$$

wobei der zweite Fall in (10.28) daraus resultiert, dass für $\mu^2 < 4(\omega-1)/\omega^2$ der Radikand in (10.27) negativ ist und für eine Zahl $z \in \mathbb{C}$ die Identität $|z^2| = (\operatorname{Re} z)^2 + (\operatorname{Im} z)^2$ gilt. Es seien noch die folgenden Eigenschaften festgehalten,

$$|\lambda_1| \geq |\lambda_2| \quad \text{für} \quad \mu \geq 0, \tag{10.29}$$

$$|\lambda_1(\mu)| \geq |\lambda_1(\widehat{\mu})| \quad \text{für} \quad \mu \geq \widehat{\mu} \geq 0. \tag{10.30}$$

(b) Basierend auf den Ergebnissen aus Teil (a) wird nun

$$r_\sigma(\mathcal{H}(\omega)) = \left\{\begin{array}{ll} \tfrac{1}{4}\big(\omega\varrho_{\mathbf{Ges}} + \sqrt{\omega^2\varrho_{\mathbf{Ges}}^2 - 4(\omega-1)}\big)^2, & \varrho_{\mathbf{Ges}}^2 \geq \dfrac{4(\omega-1)}{\omega^2}, \\ \omega - 1, & \varrho_{\mathbf{Ges}}^2 < \text{—— « ——}, \end{array}\right\} \tag{10.31}$$

nachgewiesen. Zum einen ist wegen $\sigma(\mathcal{H}_{\mathbf{Ges}}) \subset \mathbb{R}$ und der Nullsymmetrie der Menge $\sigma(\mathcal{H}_{\mathbf{Ges}}) \subset \mathbb{R}$ der Spektralradius

$$\mu_* := \varrho_{\mathbf{Ges}}$$

der Matrix $\mathcal{H}_{\mathbf{Ges}}$ bereits ein Eigenwert von $\mathcal{H}_{\mathbf{Ges}}$, so dass nach Theorem 10.41 die Zahl $\lambda_1(\mu_*)$ aus (10.27) einen Eigenwert von $\mathcal{H}(\omega)$ liefert, was unmittelbar die Ungleichung "$\geq$" in (10.31) ergibt. In (10.31) gilt aber auch die andere Ungleichung "$\leq$", denn für einen beliebigen Eigenwert $\lambda \in \sigma(\mathcal{H}(\omega))$ wählt man $\mu \in \mathbb{R}$ entsprechend (10.26), wobei $\sqrt{\lambda}$ so gewählt sei, dass $\mu \geq 0$ gilt. Wiederum mit Theorem 10.41 erhält man $\mu \in \sigma(\mathcal{H}_{\mathbf{Ges}})$, und wegen Teil (a) gilt (mit der Notation aus (10.27)) $\lambda_1(\mu) = \lambda$ oder $\lambda_2(\mu) = \lambda$. Wegen der Monotonieeigenschaften (10.29)–(10.30) erhält man so

$$|\lambda| \leq |\lambda_1(\mu)| \leq |\lambda_1(\mu_*)|$$

und daher die Ungleichung "$\leq$" in (10.31).

(c) Den Rest erhält man nun aus (10.31) und der folgenden Äquivalenz,

$$\varrho_{\mathbf{Ges}}^2 \;\geq\; \frac{4(\omega-1)}{\omega^2} \qquad \Longleftrightarrow \qquad \omega \;\leq\; \frac{2}{1+\sqrt{1-\varrho_{\mathbf{Ges}}^2}}, \tag{10.32}$$

die im Folgenden noch nachgewiesen wird. Die Situation ist klar im Fall $\varrho_{\mathbf{Ges}} = 0$, und für $\varrho_{\mathbf{Ges}} \neq 0$ ist die linke Seite der Äquivalenz (10.32) gleichbedeutend mit

$$\omega^2\varrho_{\mathbf{Ges}}^2 \;\geq\; 4(\omega-1) \qquad \Longleftrightarrow \qquad \omega^2 \;-\; \frac{4}{\varrho_{\mathbf{Ges}}^2}\omega \;+\; \frac{4}{\varrho_{\mathbf{Ges}}^2} \;\geq\; 0$$

$$\Longleftrightarrow \quad \omega \;\leq\; \frac{2}{\varrho_{\mathbf{Ges}}^2}\big(1-\sqrt{1-\varrho_{\mathbf{Ges}}^2}\big) \quad \text{oder} \quad \omega \;\geq\; \underbrace{\frac{2}{\varrho_{\mathbf{Ges}}^2}\big(1+\sqrt{1-\varrho_{\mathbf{Ges}}^2}\big)}_{>\,2,\ \text{scheidet aus}}$$

$$\Longleftrightarrow \quad \omega \;\leq\; \text{---------} \;\text{«}\; \text{---------} \;=\; \frac{2}{1+\sqrt{1-\varrho_{\mathbf{Ges}}^2}} \;=\; \omega_* \;\in\; [1,2].$$

Dies komplettiert den Beweis des Theorems. □

Der Verlauf des Spektralradius $r_\sigma(\mathcal{H}(\omega))$ in Abhängigkeit des Relaxationsparameters ω ist in Bild 10.2 dargestellt.

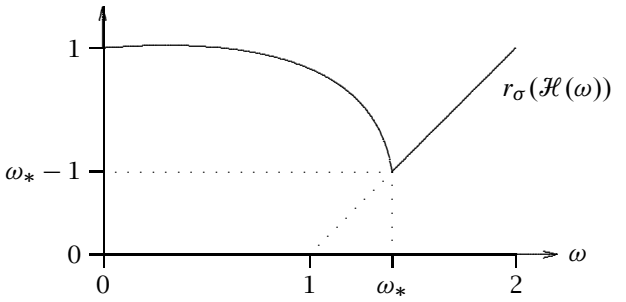


Bild 10.2: Darstellung des Verlaufs der Funktion $\omega \mapsto r_\sigma(\mathcal{H}(\omega))$

Bemerkung 10.44. Typischerweise ist der Spektralradius $\varrho_{\mathbf{Ges}}$ und somit der optimale Relaxationsparameter ω_* nicht genau bekannt. Wegen

$$\lim_{\omega\to\omega_*^-} \frac{dr_\sigma(\mathcal{H}(\omega))}{d\omega} \;=\; -\infty, \qquad \lim_{\omega\to\omega_*^+} \frac{dr_\sigma(\mathcal{H}(\omega))}{d\omega} \;=\; 1,$$

wählt man den Relaxationsparameter ω besser etwas zu groß als etwas zu klein. △

Weitere Themen und Literaturhinweise

Die hier vorgestellten Iterationsverfahren und Klassen von Matrizen werden in zahlreichen Lehrbüchern behandelt, so beispielsweise in Berman/Plemmons [4], Finckenstein [26], Golub/Ortega [37], Hämmerlin/Hoffmann [48], Hackbusch [47], Hanke-Bourgeois [52], Kress [63], Meister [70], Oevel [78], Schaback/Wendland [92], Schwarz/Klöckner [94], Stoer/Bulirsch [99] und Windisch [112]. Insbesondere in [47] finden Sie auch Ausführungen über die hier außer in Aufgabe 10.15 nicht weiter betrachteten *Block-Relaxationsverfahren*. Informationen über die hier nicht behandelte *Zweigitteriteration* beziehungsweise die allgemeineren *Mehrgitterverfahren* findet man beispielsweise in [47] und in [63].

Übungsaufgaben

Aufgabe 10.1. Für jede Matrix $\mathcal{H} \in \mathbb{R}^{N\times N}$ sind die folgenden Aussagen äquivalent:

(i) es existiert eine Vektornorm $\|\cdot\| : \mathbb{C}^N \to \mathbb{R}$, so dass für die induzierte Matrixnorm gilt $\|\mathcal{H}\| = r_\sigma(\mathcal{H})$;
(ii) jedem Eigenwert $\lambda \in \mathbb{C}$ von $\mathcal{H}$ mit $|\lambda| = r_\sigma(\mathcal{H})$ entsprechen nur lineare Elementarteiler.

Aufgabe 10.2. (a) Welche der drei Matrizen

$$\begin{pmatrix} 2 & 0 & 1 \\ 1 & 2 & 0 \\ 0 & 1 & 2 \end{pmatrix}, \qquad \begin{pmatrix} 2 & 0 & 1 \\ 1 & 1 & 0 \\ 0 & 1 & 1 \end{pmatrix}, \qquad \begin{pmatrix} 1 & 0 & 1 \\ 1 & 1 & 0 \\ 0 & 1 & 1 \end{pmatrix}$$

ist strikt diagonaldominant? Soweit dies möglich ist, ziehe man daraus jeweils Schlussfolgerungen über die Konvergenz des Gesamtschrittverfahrens.

(b) Zu Testzwecken soll für jede der genannten Matrizen sowie jeweils der rechten Seite $b = (0,0,0)^\top$ das dazugehörige lineare Gleichungssystem näherungsweise mit dem Gesamtschrittverfahren gelöst werden. Als Startvektor verwende man jeweils $x^{(0)} = (1,1,1)^\top$. Man gebe jeweils eine allgemeine Darstellung der n-ten Iterierten $x^{(n)} \in \mathbb{R}^3$ an und diskutiere die Ergebnisse im Hinblick auf Konvergenz.

Aufgabe 10.3. Gegeben seien die Matrizen

$$A = \begin{pmatrix} 0 & 1 & 0 & 1 \\ 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & 1 \\ 1 & 0 & 1 & 0 \end{pmatrix}, \qquad B = \begin{pmatrix} 0 & 1 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 2 & 0 & 0 \\ 2 & 2 & 0 & 0 & 1 \end{pmatrix}.$$

Man zeige, dass A irreduzibel beziehungsweise B reduzibel ist.

Aufgabe 10.4. Zu gegebener Matrix $A = (a_{jk}) \in \mathbb{R}^{N\times N}$ und beliebigen Indizes $j, k \in \{1,\dots,N\}$ mit $j \neq k$ heißt eine Familie von Indizes $j_0, j_1, \dots, j_M \in \{1,2,\dots,N\}$ mit $j_0 = j$, $j_M = k$ eine *j und k verbindende Kette*, falls $a_{j_{r-1},j_r} \neq 0$ gilt für $r = 1,2,\dots,M$.

Man zeige Folgendes: Eine Matrix $A \in \mathbb{R}^{N\times N}$ ist irreduzibel genau dann, wenn für alle Indizes $j, k \in \{1, \ldots, N\}$ mit $j \neq k$ eine j und k verbindende Kette existiert.

Aufgabe 10.5. Man zeige, dass die zu dem vorgestellten Modellbeispiel aus Abschnitt 10.2.1 gehörende Matrix irreduzibel diagonaldominant ist.

Aufgabe 10.6. Sei $A = (a_{jk}) \in \mathbb{R}^{N\times N}$ eine irreduzibel diagonaldominante Matrix mit $a_{jj} > 0$ für $j = 1, 2, \ldots, N$. Man zeige:

(a) Für alle Eigenwerte $\lambda \in \mathbb{C}$ von A gilt $\mathrm{Re}\,\lambda > 0$.
(b) Ist die Matrix A symmetrisch, so ist sie auch positiv definit.

Aufgabe 10.7. Für zwei Matrizen $A, \widehat{A} \in \mathbb{R}^{N\times N}$ betrachte man Zerlegungen $A = D + L + R$ beziehungsweise $\widehat{A} = \widehat{D} + \widehat{L} + \widehat{R}$ jeweils in Diagonal- sowie unteren und oberen Anteil. Man zeige: wenn A eine M-Matrix ist und die Ungleichungen $0 \leq D \leq \widehat{D}$ sowie $L + R \leq \widehat{L} + \widehat{R} \leq 0$ erfüllt sind, so ist auch $\widehat{A}$ eine M-Matrix und es gilt $0 \leq \widehat{A}^{-1} \leq A^{-1}$.

Aufgabe 10.8. Für eine Matrix $A = (a_{jk}) \in \mathbb{R}^{N\times N}$ beweise man die Äquivalenz der folgenden vier Aussagen:

(i) A ist M-Matrix;
(ii) $A + sI$ ist M-Matrix für alle $s \geq 0$;
(iii) es gibt eine Matrix $B \in \mathbb{R}^{N\times N}$ mit $B \geq 0$ und eine Zahl $s > r_\sigma(B)$, so dass die Identität $A = sI - B$ gilt;
(iv) die Nichtdiagonaleinträge a_{jk}, $j \neq k$, der Matrix A sind nichtpositiv, und alle Eigenwerte von A besitzen einen positiven Realteil, $\sigma(A) \subset \{\lambda \in \mathbb{C} : \mathrm{Re}\,\lambda > 0\}$.

Aufgabe 10.9. Gegeben sei das lineare Randwertproblem

$$-u''(x) + \frac{1}{1+x}u'(x) = \varphi(x), \qquad 0 < x < 1, \qquad u(0) = 0, \quad u(1) = 0. \tag{10.33}$$

Diskretisierung von (10.33) mit zentralen Differenzenquotienten zweiter beziehungsweise erster Ordnung bei konstanter Gitterweite $h = 1/N$ führt auf ein lineares Gleichungssystem $Av = b$. Man zeige Folgendes:

(a) Für $h < 2$ ist $A \in \mathbb{R}^{(N-1)\times(N-1)}$ eine M-Matrix.
(b) Für die Hilfsfunktion

$$\theta(x) = -\frac{(1+x)^2}{2}\ln(1+x) + \tfrac{2}{3}x(x+2)\ln 2$$

und mit den Notationen $v_j = \theta(x_j)$, $x_j = jh$ für $j = 1, 2, \ldots, N-1$ und $\mathbf{e} = (1, \ldots, 1)^\top \in \mathbb{R}^{N-1}$ gilt die Abschätzung

$$\| Av - \mathbf{e} \|_\infty \leq \tfrac{1}{4}h^2$$

(und damit $(Av)_j \geq 1 - h^2/4$ für $j = 1, 2, \ldots, N-1$).
(c) Für eine von h unabhängige Konstante M gilt $\| A^{-1} \|_\infty \leq M$.
(d) Für die Lösung u von (10.33) und die Lösung v_* des Gleichungssystems $Av = b$ gilt mit der Notation $z = (u(x_j))_{j=1}^{N-1}$ und einer von h unabhängigen Konstanten K die Abschätzung $\| v_* - z \|_\infty \leq Kh^2$.

Aufgabe 10.10. Für eine gegebene M-Matrix $A \in \mathbb{R}^{N\times N}$ weise man die folgenden Abschätzungen nach:

$$r_\sigma(\mathcal{H}(\omega_2)) \leq r_\sigma(\mathcal{H}(\omega_1)) < 1 \qquad \text{für } 0 < \omega_1 \leq \omega_2 \leq 1.$$

Aufgabe 10.11. Im Folgenden wird das Randwertproblem

$$u''(x) \;+\; p(x)u'(x) \;+\; r(x)u(x) \;=\; \varphi(x), \qquad x \in [a,b], \qquad u(a) \;=\; u(b) \;=\; 0,$$

betrachtet mit Funktionen $p,\ r,\ \varphi \in C[a,b]$ mit $r(x) \le 0$ für $x \in [a,b]$. Eine Diskretisierung der Ableitungen mittels zentraler Differenzenquotienten bei konstanter Schrittweite $h = (b-a)/N$ führt mit den Notationen $x_j = a + jh,\ p_j = p(x_j)$ und $r_j = r(x_j),\ \varphi_j = \varphi(x_j)$ für $j = 1, 2, \ldots, N-1$ sowie

$$A \;=\; \frac{1}{h^2}\begin{pmatrix} 2 & -(1-\frac{h}{2}p_1) & & & \\ -(1+\frac{h}{2}p_2) & 2 & -(1-\frac{h}{2}p_2) & & \\ & -(1+\frac{h}{2}p_3) & \ddots & \ddots & \\ & & \ddots & 2 & -(1-\frac{h}{2}p_{N-2}) \\ & & & -(1+\frac{h}{2}p_{N-1}) & 2 \end{pmatrix}$$

und $D = \operatorname{diag}(r_1, r_2, \ldots, r_{N-1}),\ c = (\varphi_j)_{j=1}^{N-1}$, auf das Gleichungssystem $(A+D)v = c$.

(a) Man zeige, dass $A + D$ eine M-Matrix ist, falls Folgendes erfüllt ist,

$$h \max_{x\in[a,b]} |p(x)| \le 2, \qquad \inf\{\operatorname{Re}\lambda : \lambda \in \sigma(A)\} \;+\; \inf_{x\in[a,b]} r(x) \;>\; 0.$$

(b) Im Fall $p(x) \equiv 0$ und $h \le (b-a)/2$ ist $A + D$ eine M-Matrix, wenn Folgendes erfüllt ist,

$$\inf_{x\in[a,b]} r(x) \;>\; -\Big(\frac{\pi}{b-a}\Big)^2 \;+\; \frac{h^2}{12}\Big(\frac{\pi}{b-a}\Big)^4.$$

Aufgabe 10.12. Ist die Matrix

$$A \;=\; \frac{1}{h^2}\begin{pmatrix} 2 & -1 & & \\ -1 & \ddots & \ddots & \\ & \ddots & \ddots & -1 \\ & & -1 & 2 \end{pmatrix} \;\in\; \mathbb{R}^{(N-1)\times(N-1)}$$

mit $h = 1/N$ positiv definit beziehungsweise eine M-Matrix beziehungsweise konsistent geordnet? Man bestimme als Funktion von h die Eigenwerte von $I - D^{-1}A$ und den zugehörigen Spektralradius $r_\sigma(I - D^{-1}A)$, den optimalen Parameter ω_* für das Relaxationsverfahren sowie den Spektralradius $r_\sigma(\mathcal{H}(\omega_*))$ der entsprechenden Iterationsmatrix $\mathcal{H}(\omega_*)$.

Aufgabe 10.13. Man zeige, dass reguläre Dreiecksmatrizen konsistent geordnet sind.

Aufgabe 10.14. Gegeben sei eine Block-Tridiagonalmatrix von der speziellen Form

$$A \;=\; \begin{pmatrix} B & b_1 D & & \\ a_1 D & \ddots & \ddots & \\ 0 & \ddots & \ddots & b_{M-1}D \\ & & a_{M-1}D & B \end{pmatrix} \;\in\; \mathbb{R}^{N\times N}$$

mit der Diagonalmatrix $D = \operatorname{diag}(b_{11}, \ldots, b_{KK})$ wobei $0 \neq b_{jj}$ die Diagonaleinträge von $B \in \mathbb{R}^{K\times K}$ bezeichne. Mit der Zerlegung $B = D + L + R$ entsprechend (10.15) und mit

$$\mathcal{J}(\alpha) \;=\; \alpha D^{-1}L \;+\; \alpha^{-1}D^{-1}R, \qquad 0 \neq \alpha \in \mathbb{C}$$

gelte $\mathcal{J}(\alpha) = S_\alpha \mathcal{J}(1) S_\alpha^{-1}$ für $0 \neq \alpha \in \mathbb{C}$ mit einer geeigneten Transformationsmatrix $S_\alpha \in \mathbb{R}^{N\times N}$. Man zeige, dass die Matrix A konsistent geordnet ist.

Aufgabe 10.15. Es sei

$$A = \begin{pmatrix} A_{11} & \cdots & A_{1M} \\ \vdots & \ddots & \vdots \\ A_{M1} & \cdots & A_{MM} \end{pmatrix}$$

eine quadratische Matrix mit quadratischen Diagonalblöcken $A_{jj},\ j = 1, 2, \ldots, M$, und die Block-Diagonalmatrix $D = \operatorname{diag}(A_{11}, \ldots, A_{NN})$ sei nichtsingulär. Weiter bezeichne

$$L \;=\; \begin{pmatrix} & & \\ A_{21} & & \\ \vdots & \ddots & \\ A_{M1} & \cdots & A_{M,M-1} \end{pmatrix}, \qquad R \;=\; \begin{pmatrix} A_{12} & \cdots & A_{1M} \\ & \ddots & \vdots \\ & & A_{M-1,M} \\ & & \end{pmatrix},$$

und

$$\mathcal{H}(\omega) \;=\; (D + \omega L)^{-1}((1-\omega)D - \omega R) \qquad (\omega \neq 0).$$

In den folgenden Teilaufgaben (a) und (b) seien für eine Zahl $p > 1$ die Eigenwerte von

$$\mathcal{J}(\alpha) \;=\; \alpha D^{-1}L \;+\; \alpha^{-(p-1)}D^{-1}R, \qquad 0 \neq \alpha \in \mathbb{C}, \tag{10.34}$$

unabhängig von α, es gelte also $\sigma(\mathcal{J}(\alpha)) = \sigma(\mathcal{J}(1))$ für $\alpha \neq 0$. Man weise Folgendes nach:

(a) Ist $\mu \in \sigma(D^{-1}(L+R))$ erfüllt und die Zahl $\lambda \in \mathbb{C}$ eine Lösung der Gleichung

$$(\lambda \;+\; \omega \;-\; 1)^p \;=\; \lambda^{p-1}\omega^p\mu^p, \tag{10.35}$$

so gilt $\lambda \in \sigma(\mathcal{H}(\omega))$. Ist umgekehrt $0 \neq \lambda \in \sigma(\mathcal{H}(\omega))$ und erfüllt μ die Gleichung (10.35), dann ist $\mu \in \sigma(D^{-1}(L+R))$.

(b) Für $\mu \neq 0$ gilt

$$\mu \;\in\; \sigma(D^{-1}(L+R)) \quad\Longleftrightarrow\quad \mu^p \;\in\; \sigma(\mathcal{H}(1)),$$

und $r_\sigma(D^{-1}(L+R))^p = r_\sigma(\mathcal{H}(1))$.

(c) Sei nun A von der speziellen Gestalt

$$A \;=\; \begin{pmatrix} A_{11} & 0 & \cdots & 0 & A_{1M} \\ A_{21} & \ddots & \ddots & & 0 \\ 0 & \ddots & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & \ddots & 0 \\ 0 & \cdots & 0 & A_{M,M-1} & A_{MM} \end{pmatrix}.$$

Man zeige, dass mit $p = M \geq 2$ die Eigenwerte der Matrix $\mathcal{J}(\alpha)$ aus (10.34) unabhängig von α sind.

Aufgabe 10.16 (*Numerische Aufgabe*). Zur numerischen Lösung des Randwertproblems

$$u''(x) + u(x) = e^x, \quad x \in [0, \pi/2], \qquad u(0) = u(\pi/2) = 0,$$

betrachte man auf einem äquidistanten Gitter der Weite $h = \frac{\pi}{2N}$ das zugehörige Differenzenschema

$$v_{j+1} - (2 - h^2)v_j + v_{j-1} = h^2 e^{z_j}, \qquad j = 1, 2, \ldots, N-1, \tag{10.36}$$

mit $z_j = jh$. Für $N = 30$ beziehungsweise $N = 200$ bestimme man eine approximative Lösung von (10.36) mithilfe des *Relaxationsverfahrens* mit den folgenden Relaxationsparametern, $\omega = 0.1, 0.2, 0.3, \ldots, 2.0, 2.1$, wobei die Iteration jeweils abgebrochen werden soll, wenn mehr als 1000 Iterationen (für $N = 200$ mehr als 2000 Iterationen) benötigt werden oder falls

$$\| x^{(n)} - x^{(n-1)} \|_\infty \leq 10^{-5}$$

ausfällt. Als Startwert wähle man jeweils $x^{(0)} = 0$. Für jede Wahl von ω gebe man die Anzahl der benötigten Iterationsschritte n, $\| x^{(n)} - x^{(n-1)} \|_\infty$ und den Fehler $\max_{j=1,\ldots,N-1} |x_j^{(n)} - u(z_j)|$ tabellarisch an.

11 Verfahren der konjugierten Gradienten und GMRES-Verfahren

11.1 Vorbetrachtungen

Ziel der nachfolgenden Betrachtungen ist erneut die approximative Lösung eines regulären linearen Gleichungssystems

$$Ax \;=\; b \qquad (A \in \mathbb{R}^{N\times N} \text{ regulär}, \qquad b \in \mathbb{R}^N)$$

(mit der eindeutigen Lösung $x_* \;=\; A^{-1}b \in \mathbb{R}^N$), und hierzu seien

$$\{0\} \;\subset\; \mathcal{D}_1 \;\subset\; \mathcal{D}_2 \;\subset\; \ldots \;\subset\; \mathbb{R}^N \tag{11.1}$$

zunächst nicht weiter spezifizierte (endlich oder unendlich viele) lineare Unterräume. Im Folgenden werden zwei Ansätze zur Bestimmung von (unterschiedlichen) Vektorfolgen $x_n \in \mathcal{D}_n,\ n = 1, 2, \ldots$, vorgestellt.[1]

Definition 11.1. (a) Für gegebene Ansatzräume (11.1) hat der *Ansatz des orthogonalen Residuums* zur Bestimmung von Vektoren $x_1,\ x_2, \ldots \in \mathbb{R}^N$ die folgende Form,

$$\left.\begin{array}{rcl} x_n & \in & \mathcal{D}_n, \\ Ax_n - b & \in & \mathcal{D}_n^{\perp} \end{array}\right\} \qquad n \;=\; 1, 2, \ldots\,. \tag{11.2}$$

(b) Der *Ansatz des minimalen Residuums* zur Bestimmung von Vektoren $x_1,\ x_2, \ldots \in \mathbb{R}^N$ hat für gegebene Ansatzräume (11.1) die folgende Form,

$$\left.\begin{array}{rcl} x_n & \in & \mathcal{D}_n, \\ \| Ax_n - b \|_2 & = & \min\limits_{x\in\mathcal{D}_n} \| Ax - b \|_2 \end{array}\right\} \qquad n \;=\; 1, 2, \ldots\,. \tag{11.3}$$

Hierbei bezeichnet wie üblich

$$\mathcal{M}^{\perp} \;:=\; \left\{ y \in \mathbb{R}^N :\; y^{\top}x = 0 \quad \text{für jedes } \; x \in \mathcal{M} \right\}, \qquad \mathcal{M} \subset \mathbb{R}^N \text{ beliebig},$$

das orthogonale Komplement einer Menge $\mathcal{M}$, und $\| \cdot \|_2$ bezeichnet wieder die euklidische Vektornorm. Schließlich bezeichnet im Folgenden zu jedem $x \in \mathbb{R}^N$ der Vektor $Ax - b$ das zugehörige *Residuum*[2], was die Bezeichnungen für die beiden in Definition 11.1 vorgestellten Ansätze erklärt.

[1] Im Unterschied zum vorigen Kapitel 10 wird nun wieder die etwas knappere Tiefstellung für den Laufindex n gewählt. Dies ist hier ohne weiteres möglich, da die einzelnen Einträge in den vektorwertigen Iterierten im Folgenden keine spezielle Rolle spielen.

[2] In der Literatur findet man die Bezeichnung "Residuum" oft auch für den Vektor $b - Ax$ anstelle $Ax - b$.

Bemerkung 11.2. Natürliche Fragestellungen im Zusammenhang mit den beiden vorgestellten Ansätzen sind jeweils Existenz und Eindeutigkeit der Vektoren x_n. Zudem gilt es, Algorithmen zur Bestimmung dieser Vektoren anzugeben sowie Abschätzungen für den Fehler $\|x_n - x_*\|$ herzuleiten bezüglich gängiger Normen. Schließlich sind spezielle Ansatzräume für $\mathcal{D}_1, \mathcal{D}_2, \ldots$ auszuwählen. △

Bei der Wahl spezieller Ansatzräume in (11.1) werden die im Folgenden definierten Krylovräume eine hervorgehobene Rolle spielen:

Definition 11.3. Zu gegebener Matrix $A \in \mathbb{R}^{N\times N}$ und einem Vektor $b \in \mathbb{R}^N$ ist die Folge der *Krylovräume* wie folgt erklärt,

$$\mathcal{K}_n(A,b) \;=\; \operatorname{span}\{b, Ab, \ldots, A^{n-1}b\} \;\subset\; \mathbb{R}^N, \qquad n \;=\; 0, 1, \ldots .$$

Offensichtlich sind die Krylovräume aufsteigend, es gilt $\{0\} = \mathcal{K}_0(A,b) \subset \mathcal{K}_1(A,b) \subset \ldots$. Weitere Eigenschaften von eher technischer Natur werden zu einem späteren Zeitpunkt vorgestellt[3].

11.1.1 Ausblick

In dem vorliegenden Kapitel werden nun die beiden in Definition 11.1 angegebenen Ansätze mit den speziellen Räumen $\mathcal{D}_n = \mathcal{K}_n(A,b)$ behandelt.[4]

(a) Der Ansatz des orthogonalen Residuums mit den Räumen $\mathcal{D}_n = \mathcal{K}_n(A,b)$ wird für symmetrische, positiv definite Matrizen $A \in \mathbb{R}^{N\times N}$ betrachtet. Dies führt auf das klassische Verfahren der konjugierten Gradienten. Einzelheiten hierzu werden in den Abschnitten 11.2–11.4 vorgestellt.

Für allgemeine (also indefinite oder nichtsymmetrische) reguläre Matrizen $A \in \mathbb{R}^{N\times N}$ kann man zu den Normalgleichungen $A^\top Ax = A^\top b$ übergehen und hierfür das angesprochene Verfahren der konjugierten Gradienten betrachten. Einige Details zu diesem Ansatz finden sich in Abschnitt 11.5.

(b) Schließlich wird für die Räume $\mathcal{D}_n = \mathcal{K}_n(A,b)$ der Ansatz des minimalen Residuums betrachtet. Dies führt auf das (in Abschnitt 11.6 behandelte) GMRES-Verfahren, welches universell einsetzbar ist, weitere Voraussetzungen an die Matrix A wie etwa Symmetrie entfallen hier.

11.2 Der Ansatz des orthogonalen Residuums für positiv definite Matrizen

In dem vorliegenden Abschnitt 11.2 wird der Ansatz des orthogonalen Residuums für allgemeine Ansatzräume der Form (11.1) betrachtet unter der zusätzlichen Annahme, dass $A \in \mathbb{R}^{N\times N}$ eine symmetrische, positiv definite Matrix ist.

[3] siehe Lemma 11.31 auf Seite 319

[4] Diese Verfahren werden allgemein als *Krylovraummethoden* bezeichnet.

11.2.1 Existenz, Eindeutigkeit und Minimaleigenschaft

Im Folgenden wird für eine gegebene symmetrische, positiv definite Matrix $A \in \mathbb{R}^{N\times N}$ die Existenz und Eindeutigkeit der zum Ansatz des orthogonalen Residuums (11.2) gehörenden Vektoren x_n diskutiert. Hierzu werden die folgenden Notationen eingeführt:

$$\begin{aligned} \langle x,\, y\rangle_2 &= x^\top y, \qquad x,\, y \in \mathbb{R}^N, \\ \langle x,\, y\rangle_A &= x^\top A y, \qquad \text{——«——}, \qquad \|x\|_A \;=\; (x^\top A x)^{1/2}, \qquad x \in \mathbb{R}^N. \end{aligned}$$

Bemerkung 11.4. (1) Die neue Notation $\langle\cdot,\,\cdot\rangle_2$ für das klassische skalare Produkt wird wegen der gelegentlich einfacheren Lesbarkeit eingeführt.

(2) Wie man leicht nachrechnet, bildet im Falle einer symmetrischen, positiv definiten Matrix $A \in \mathbb{R}^{N\times N}$ die Abbildung $\langle\cdot,\,\cdot\rangle_A$ ein Skalarprodukt auf $\mathbb{R}^N$, und $\|\cdot\|_A$ stellt offensichtlich die zugehörige Norm dar; diese bezeichnet man als *A-Norm*.

(3) Aufgrund der Natur des Ansatzes des orthogonalen Residuums erhält man Fehlerabschätzungen zunächst nur bezüglich der A-Norm. Fehlerabschätzungen bezüglich der natürlicheren euklidischen Norm $\|\cdot\|_2$ werden dann noch über die Äquivalenz von Normen hergeleitet. △

Das folgende Resultat liefert für den Ansatz des orthogonalen Residuums neben Existenz und Eindeutigkeit auch eine Minimaleigenschaft, mit der zu einem späteren Zeitpunkt[5] noch konkrete Fehlerabschätzungen hergeleitet werden.

Theorem 11.5. *Zu gegebener symmetrischer, positiv definiter Matrix $A \in \mathbb{R}^{N\times N}$ sind für $n = 1, 2, \ldots$ die Vektoren x_n aus dem Ansatz des orthogonalen Residuums (11.2) – mit allgemeinen Ansatzräumen $\mathcal{D}_n$ entsprechend (11.1) – eindeutig bestimmt, und es gilt*

$$\|x_n - x_*\|_A \;=\; \min_{x\in\mathcal{D}_n} \|x - x_*\|_A, \qquad n \;=\; 1, 2, \ldots\,. \tag{11.4}$$

BEWEIS. Bei fest gewähltem Index n betrachtet man für den Nachweis der Eindeutigkeit zwei Vektoren $x_n,\ \widehat{x}_n$ mit der Eigenschaft (11.2). Hier gilt

$$\langle \underbrace{A(x_n - \widehat{x}_n)}_{\in\, \mathcal{D}_n^\perp},\ \underbrace{x_n - \widehat{x}_n}_{\in\, \mathcal{D}_n}\rangle_2 \;=\; 0 \qquad \rightsquigarrow \qquad x_n \;=\; \widehat{x}_n.$$

Für den Nachweis der Existenz setzt man mit einer beliebigen Basis $d_0, d_1, \ldots, d_{m-1}$ von $\mathcal{D}_n$ (mit $m := \dim \mathcal{D}_n$) wie folgt an,

$$x_n \;=\; \sum_{k=0}^{m-1} \alpha_k\, d_k \tag{11.5}$$

[5] siehe Abschnitt 11.4

und erhält damit

$$
\begin{aligned}
x_n \text{ genügt (11.2)} \quad &\Longleftrightarrow \quad Ax_n - b \in \mathcal{D}_n^{\perp} && (11.6)\\
&\Longleftrightarrow \quad \langle Ax_n - b,\, d_j \rangle_2 = 0 \qquad \text{für} \quad j = 0, 1, \ldots, m-1,\\
&\Longleftrightarrow \quad \sum_{k=0}^{m-1} \langle Ad_k,\, d_j \rangle_2 \alpha_k = \langle b,\, d_j \rangle_2 \quad \text{---------- (} \quad \text{----------}, && (11.7)
\end{aligned}
$$

was ein lineares System von m Gleichungen für die m Koeffizienten $\alpha_0, \ldots, \alpha_{m-1}$ darstellt. Infolgedessen und aufgrund der Eindeutigkeit der Lösung – diese wurde im ersten Teil dieses Beweises bereits nachgewiesen – ist dieses Gleichungssystem also lösbar. Schließlich ist noch die Minimaleigenschaft (11.4) nachzuweisen. Hierzu berechnet man für einen beliebigen Vektor $x \in \mathcal{D}_n$ Folgendes,

$$
\begin{aligned}
\|x - x_*\|_A^2 &= \|x_n - x_* + x - x_n\|_A^2\\
&= \|x_n - x_*\|_A^2 + 2\overbrace{\langle \underbrace{A(x_n - x_*)}_{\in \mathcal{D}_n^{\perp}},\, \underbrace{x - x_n}_{\in \mathcal{D}_n} \rangle_2}^{=0} + \|x - x_n\|_A^2\\
&\geq \|x_n - x_*\|_A^2.
\end{aligned}
$$

Dies komplettiert den Beweis des Theorems. □

11.2.2 Der Ansatz des orthogonalen Residuums (11.2) für gegebene A-konjugierte Basen

Mit dem Beweis von Theorem 11.5 ist bereits eine Möglichkeit zur Durchführung des Ansatzes des orthogonalen Residuums vorgestellt worden; ausgehend von einer Basis $d_0, \ldots, d_{m-1}$ für $\mathcal{D}_n$ hat man nur das durch den Ansatz (11.5) entstehende Gleichungssystem (11.7) zu lösen. Im Folgenden soll ein Spezialfall behandelt werden, bei dem dieses Gleichungssystem (11.7) von Diagonalgestalt ist.

Definition 11.6. Es sei $A \in \mathbb{R}^{N \times N}$ eine symmetrische, positiv definite Matrix. Gegebene Vektoren $d_0, d_1, \ldots, d_{n_*-1} \in \mathbb{R}^N \setminus \{0\}$ mit $n_* \leq N$ heißen *A-konjugiert*, falls Folgendes gilt,

$$\langle Ad_k,\, d_j \rangle_2 = 0 \qquad \text{für} \quad k \neq j.$$

Bemerkung 11.7. A-Konjugiertheit ist also gleichbedeutend mit paarweiser Orthogonalität bezüglich des Skalarprodukts $\langle \cdot, \cdot \rangle_A$. △

Unter Fortführung des Ansatzes (11.5)–(11.7) lässt sich im Falle symmetrischer positiv definiter Matrizen $A \in \mathbb{R}^{N \times N}$ der Ansatz des orthogonalen Residuums (11.2) besonders einfach verwirklichen, falls eine A-konjugierte Basis von $\mathcal{D}_n$ gegeben ist. Genauer gilt Folgendes:

Theorem 11.8. *Für eine gegebene symmetrische, positiv definite Matrix $A \in \mathbb{R}^{N\times N}$ und A-konjugierte Vektoren $d_0, d_1, \ldots, d_{n_*-1} \in \mathbb{R}^N \setminus \{0\}$ mit $n_* \le N$ gelte*

$$\mathcal{D}_n \;=\; \operatorname{span}\{d_0, d_1, \ldots, d_{n-1}\}, \qquad n \;=\; 0, 1, \ldots, n_*.$$

Dann erhält man für den Ansatz des orthogonalen Residuums (11.2) die folgenden Darstellungen für $n \;=\; 1, 2, \ldots, n_*$:

$$x_n \;=\; \sum_{k=0}^{n-1} \alpha_k d_k, \quad \textit{mit} \quad \alpha_k \;=\; -\frac{\langle r_k,\, d_k\rangle_2}{\langle Ad_k,\, d_k\rangle_2}, \tag{11.8}$$

$$r_k \;:=\; Ax_k - b, \quad k \ge 1, \quad r_0 \;:=\; -b. \tag{11.9}$$

BEWEIS. Aus der Vorgehensweise des Ansatzes (11.5)–(11.7) (mit $m \;=\; n$) im Beweis von Theorem 11.5 erhält man im Fall der nun vorliegenden A-Konjugiertheit zunächst Folgendes,

$$x_n \;=\; \sum_{k=0}^{n-1} \alpha_k d_k, \quad \text{mit} \quad \alpha_k \;:=\; \frac{\langle b,\, d_k\rangle_2}{\langle Ad_k,\, d_k\rangle_2} \qquad (n \;=\; 1, 2, \ldots, n_*), \tag{11.10}$$

und die Zahl α_k in (11.10) stimmt mit der aus (11.8) überein, was für $k \;=\; 0$ klar ist und für $k \ge 1$ so folgt:

$$\langle \underbrace{b - Ax_n}_{=\; -r_n},\, d_n\rangle_2 \;=\; \langle b,\, d_n\rangle_2 - \sum_{k=0}^{n-1} \alpha_k \underbrace{\langle Ad_k,\, d_n\rangle_2}_{=\; 0} \;=\; \langle b,\, d_n\rangle_2, \quad n \;=\; 0, 1, \ldots, n_*.$$

Dies komplettiert den Beweis. □

Bemerkung 11.9. (a) Der Darstellung (11.8) entnimmt man, dass die Zahl α_k unabhängig von n ist und somit Folgendes gilt,

$$x_{n+1} \;=\; x_n + \alpha_n d_n, \quad r_{n+1} \;=\; r_n + \alpha_n Ad_n \quad (n \;=\; 0, \ldots, n_* - 1;\; x_0 := 0), \tag{11.11}$$

womit sich die Durchführung des Verfahrens (11.8) weiter vereinfacht. Man beachte, dass die Berechnung des Matrix-Vektor-Produkts Ad_n für die Bestimmung von α_n sowieso erforderlich ist, und mittels (11.11) erhält man dann das Residuum r_{n+1} auf einfache Weise, also ohne Berechnung eines weiteren Matrix-Vektor-Produkts. (Die meisten Abbruchkriterien basieren auf den Werten des Residuums, weshalb dieses von Bedeutung ist.)

(b) Aufgrund der ersten Identität in (11.11) bezeichnet man den Vektor d_n als *Suchrichtung*, und die Zahl α_n wird als *Schrittweite* bezeichnet. Diese Bezeichnungsweise verwendet man im Übrigen auch bei anderen Verfahren der Form (11.11).

(c) Ebenfalls mit der ersten Identität in (11.11) wird klar, das im Prinzip eine simultane Berechnung der Suchrichtungen und Approximationen in der Reihenfolge $d_0, x_1, d_1, x_2, \ldots$ möglich ist. In der Praxis wird im Fall $\mathcal{D}_n \;=\; \mathcal{K}_n(A, b)$ auch so vorgegangen. Einzelheiten werden im nachfolgenden Abschnitt 11.3 behandelt.

(d) Für vorgegebene Suchrichtungen in der Vorschrift (11.11) sind die Schrittweiten aus (11.8) optimal in dem folgenden Sinne,

$$\|x_{n+1} - x_*\|_A = \min_{t \in \mathbb{R}} \|x_n + t d_n - x_*\|_A.$$

Der Nachweis dafür ist elementar und wird hier nicht geführt. △

11.3 Das CG-Verfahren für positiv definite Matrizen

11.3.1 Einleitende Bemerkungen

Für den Ansatz des orthogonalen Residuums sollen im Folgenden nun speziell Krylovräume als Ansatzräume herangezogen werden.

Definition 11.10. Zu gegebener symmetrischer, positiv definiter Matrix $A \in \mathbb{R}^{N\times N}$ ist das *Verfahren der konjugierten Gradienten* gegeben durch Ansatz (11.2) mit der speziellen Wahl

$$\mathcal{D}_n = \mathcal{K}_n(A,b), \qquad n = 0,1,\ldots . \tag{11.12}$$

Dieses Verfahren bezeichnet man auch kurz als *CG-Verfahren*, wobei die Notation "CG" von der englischen Bezeichnung "method of conjugate gradients" herrührt. Der Grund für die Bezeichnungsweise "konjugierte Gradienten" wird später geliefert[6]. Für die praktische Durchführung des CG-Verfahrens liefert Theorem 11.8 einen ersten Ansatz. Die noch ausstehende Konstruktion A-konjugierter Suchrichtungen in dem Raum $\mathcal{K}_n(A,b)$ ist das Thema des folgenden Abschnitts 11.3.2.

11.3.2 Die Berechnung A-konjugierter Suchrichtungen in $\mathcal{K}_n(A,b)$

Das folgende Lemma behandelt die Berechnung A-konjugierter Suchrichtungen in $\mathcal{K}_n(A,b)$ für $n = 0,1,\ldots$. Ausgehend von den Notationen aus Theorem 11.8 wird für jetzt fixierten Index n dabei so vorgegangen, dass – ausgehend von einer bereits konstruierten A-konjugierten Basis $d_0,\ldots,d_{n-1}$ für $\mathcal{K}_n(A,b)$ – eine A-konjugierte Basis für $\mathcal{K}_{n+1}(A,b)$ gewonnen wird durch eine Gram-Schmidt-Orthogonalisierung der Vektoren $d_0,\ldots,d_{n-1},\ -r_n \in \mathbb{R}^N$ bezüglich des Skalarprodukts $\langle\cdot,\cdot\rangle_A$. Wie sich im Beweis von Lemma 11.11 herausstellt, genügt hierfür eine Gram-Schmidt-Orthogonalisierung der beiden Vektoren $d_{n-1},\ -r_n \in \mathbb{R}^N$.

Lemma 11.11. *Zu gegebener symmetrischer, positiv definiter Matrix $A \in \mathbb{R}^{N\times N}$ und mit den Notationen aus Theorem 11.8 seien die Suchrichtungen speziell wie folgt gewählt: $d_0 := b$ sowie*

$$d_n := -r_n + \beta_{n-1} d_{n-1}, \quad \beta_{n-1} := \frac{\langle Ar_n,\, d_{n-1}\rangle_2}{\langle Ad_{n-1},\, d_{n-1}\rangle_2}, \quad n = 1,2,\ldots,n_*-1, \tag{11.13}$$

[6] siehe Bemerkung 11.15

wobei n_* den ersten Index mit $r_{n_*} = 0$ bezeichnet. Mit dieser Wahl sind die Vektoren $d_0, d_1, \ldots, d_{n_*-1} \in \mathbb{R}^N$ A-konjugiert und es gilt

$$\left.\begin{aligned} \operatorname{span}\{d_0, \ldots, d_{n-1}\} &= \operatorname{span}\{b, r_1, r_2, \ldots, r_{n-1}\} = \mathcal{K}_n(A,b), \\ & \qquad n = 1, 2, \ldots, n_*. \end{aligned}\right\} \tag{11.14}$$

BEWEIS. Mittels vollständiger Induktion über $n = 1, 2, \ldots, n_*$ werden sowohl die A-Konjugiertheit der Vektoren $d_0, d_1, \ldots, d_{n-1} \in \mathbb{R}^N$ als auch die beiden Identitäten in (11.14) nachgewiesen. Wegen

$$\operatorname{span}\{d_0\} = \operatorname{span}\{b\} = \mathcal{K}_1(A,b)$$

ist der Induktionsanfang klar, und im Folgenden sei angenommen, dass die Vorschrift (11.13) ein System $d_0 = b,\ d_1,\ d_2, \ldots,\ d_{n-1}$ von A-konjugierten Vektoren mit der Eigenschaft (11.14) liefert mit einem fixierten Index $1 \le n \le n_* - 1$. Gemäß (11.2) gilt $r_n \in \mathcal{K}_n(A,b)^\perp$, und im Fall $r_n \neq 0$ sind demnach die Vektoren $d_0, \ldots, d_{n-1},\ -r_n$ linear unabhängig. Eine Gram-Schmidt-Orthogonalisierung dieser Vektoren bezüglich des Skalarprodukts $\langle \cdot, \cdot \rangle_A$ liefert den Vektor

$$d_n := -r_n + \sum_{k=0}^{n-1} \frac{\langle Ar_n,\, d_k \rangle_2}{\langle Ad_k,\, d_k \rangle_2} d_k \overset{(*)}{=} -r_n + \beta_{n-1} d_{n-1}, \tag{11.15}$$

wobei man die Identität $(*)$ aus den Eigenschaften $A(\mathcal{K}_{n-1}(A,b)) \subset \mathcal{K}_n(A,b)$ sowie $r_n \in \mathcal{K}_n(A,b)^\perp$ erschließt:

$$\langle Ar_n,\, d_k \rangle_2 = \langle r_n,\, Ad_k \rangle_2 = 0, \qquad k = 0, 1, \ldots, n-2.$$

Nach Konstruktion sind die Vektoren $d_0, \ldots, d_{n-1},\ d_n$ A-konjugiert und es gilt $\operatorname{span}\{d_0, \ldots, d_{n-1}, d_n\} = \operatorname{span}\{b, r_1, r_2, \ldots, r_n\}$. Aufgrund der zweiten Identität in (11.11) gilt zudem $\operatorname{span}\{b, r_1, r_2, \ldots, r_n\} \subset \mathcal{K}_{n+1}(A,b)$, so dass aus Dimensionsgründen auch hier notwendigerweise Gleichheit vorliegt. Dies komplettiert den Beweis des Lemmas. □

Bemerkung 11.12. Mit dem durch Lemma 11.11 beschriebenen Abbruch wird gleichzeitig die Lösung des Gleichungssystems $Ax = b$ geliefert, es gilt also $x_{n_*} = x_*$. Dabei gilt notwendigerweise

$$n_* \le N,$$

denn aufgrund der linearen Unabhängigkeit der beiden Vektorsysteme in (11.14) erhält man $\dim \mathcal{K}_n(A,b) = n$ für $n = 0, 1, \ldots, n_*$. △

Als unmittelbare Konsequenz aus dem Beweis von Lemma 11.11 erhält man für die Schrittweiten noch die folgende Darstellung, wie man sie üblicherweise auch in numerischen Implementierungen verwendet:

Lemma 11.13. *In der Situation von Lemma 11.11 gelten die Darstellungen*

$$\alpha_n = \frac{\|r_n\|_2^2}{\langle Ad_n,\, d_n\rangle_2}, \qquad n = 0,1,\ldots,n_*-1, \tag{11.16}$$

$$\beta_{n-1} = \frac{\|r_n\|_2^2}{\|r_{n-1}\|_2^2}, \qquad n = 1,2,\ldots,n_*-1 \qquad (r_0 := -b). \tag{11.17}$$

BEWEIS. Mit $r_n \in \mathcal{K}_n(A,b)^\perp$ sowie der Setzung (11.13) für die Suchrichtung d_n erhält man $-\langle r_n,\, d_n\rangle_2 = \|r_n\|_2^2$, und zusammen mit (11.8) liefert dies (11.16). Diese Darstellung (11.16) für α_n zusammen mit der Identität[7] $r_n = r_{n-1} + \alpha_{n-1}Ad_{n-1}$ liefert schließlich Folgendes,

$$\|r_n\|_2^2 = \underbrace{\langle r_n,\, r_{n-1}\rangle_2}_{=\,0} + \alpha_{n-1}\langle r_n,\, Ad_{n-1}\rangle_2 = \beta_{n-1}\|r_{n-1}\|_2^2,$$

und daher gilt auch die angegebene Darstellung (11.17) für β_{n-1}. Dies komplettiert den Beweis des Lemmas. □

11.3.3 Der Algorithmus zum CG-Verfahren

Trägt man die Resultate aus Theorem 11.8, Darstellung (11.11), Lemma 11.11 sowie Lemma 11.13 zusammen, so ergibt sich der folgende Algorithmus für das Verfahren der konjugierten Gradienten.

Algorithmus 11.14. *Schritt 0:* Setze $r_0 = -b$.
Für $n = 0,\,1,\ldots$:

(a) Wenn $r_n = 0$, so Abbruch, $n = n_*$.

(b) Wenn andererseits $r_n \neq 0$, so verfahre man in *Schritt* $n+1$ wie folgt,

$$d_n = \begin{cases} -r_n + \beta_{n-1}d_{n-1}, & \beta_{n-1} = \dfrac{\|r_n\|_2^2}{\|r_{n-1}\|_2^2}, \quad \text{wenn } n \geq 1 \\ -r_0, & \text{wenn } n = 0 \end{cases}$$

$$x_{n+1} = x_n + \alpha_n d_n, \qquad \alpha_n = \frac{\|r_n\|_2^2}{\langle Ad_n,\, d_n\rangle_2},$$

$$r_{n+1} = r_n + \alpha_n Ad_n.$$

△

Bemerkung 11.15. Die in Definition 11.10 eingeführte Bezeichnung "Verfahren der konjugierten Gradienten" hat ihre Ursache in den beiden folgenden Eigenschaften:

- Für jeden Index n ist das Residuum r_n identisch mit dem Gradienten des Energiefunktionals $\mathcal{J}(x) = \frac{1}{2}\langle Ax,\, x\rangle_2 - \langle x,\, b\rangle_2$ an der Stelle x_n, es gilt also $r_n = \nabla\mathcal{J}(x_n)$; siehe hierzu Aufgabe 11.2.

[7] vergleiche (11.11)

- Es gilt

$$\langle r_n, r_k \rangle_2 = 0 \qquad \text{für} \quad n \neq k.$$

Dies folgt unmittelbar aus den Eigenschaften (11.2) sowie (11.14). △

11.4 Die Konvergenzgeschwindigkeit des CG-Verfahrens

Mit Bemerkung 11.12 wird klar, dass das CG-Verfahren als direktes Verfahren interpretiert werden kann, das nach endlich vielen Schritten die exakte Lösung von $Ax = b$ liefert, $x_{n_*} = x_*$. Aufgrund der eingangs von Abschnitt 10 angestellten Bemerkungen sind jedoch auch die Approximationseigenschaften der Iterierten $x_1, x_2, \ldots$ von Interesse. Aus diesem Grund werden in dem vorliegenden Abschnitt ausgehend von der Optimalitätseigenschaft (11.4) konkrete Fehlerabschätzungen für das Verfahren der konjugierten Gradienten hergeleitet. Hierbei ist das folgende Lemma nützlich.

Lemma 11.16. *Zu einer gegebenen symmetrischen, positiv definiten Matrix $A \in \mathbb{R}^{N\times N}$ sei $(\lambda_k, v_k)_{k=1,\ldots,N}$ ein vollständiges System von (positiven) Eigenwerten $\lambda_k > 0$ und zugehörigen orthonormalen Eigenvektoren $v_k \in \mathbb{R}^N$, es liegt also folgende Situation vor:*

$$Av_k = \lambda_k v_k, \qquad v_j^\top v_k = \delta_{jk}, \qquad j,k = 1,2,\ldots,N.$$

Mit der Entwicklung $x = \sum_{k=1}^N c_k v_k \in \mathbb{R}^N$ gelten für jedes Polynom p die folgenden Darstellungen:

$$p(A)x = \sum_{k=1}^N c_k p(\lambda_k) v_k,$$

$$\|p(A)x\|_2 = \Big(\sum_{k=1}^N c_k^2\, p(\lambda_k)^2\Big)^{1/2}, \qquad \|p(A)x\|_A = \Big(\sum_{k=1}^N c_k^2\, \lambda_k\, p(\lambda_k)^2\Big)^{1/2}.$$

Insbesondere gilt also

$$m^{1/2}\|x\|_2 \le \|x\|_A \le M^{1/2}\|x\|_2, \quad x \in \mathbb{R}^N \qquad \left(\begin{array}{l} m := \min\limits_{k=1,\ldots,N} \lambda_k, \\ M := \max\limits_{k=1,\ldots,N} \lambda_k \end{array}\right). \tag{11.18}$$

BEWEIS. Mit der angegebenen Entwicklung für $x \in \mathbb{R}^N$ bezüglich der vorgegebenen Basis erhält man unmittelbar Folgendes,

$$A^\nu x = \sum_{k=1}^N c_k \lambda_k^\nu v_k, \qquad \nu = 0, 1, \ldots,$$

und daraus folgt die erste Identität des Lemmas. Weiter berechnet man

$$\begin{aligned}\|\,p(A)x\,\|_2 &= \langle\!\langle \sum_{j=1}^{N} c_j\,p(\lambda_j)\,v_j\,,\ \sum_{k=1}^{N} c_k\,p(\lambda_k)\,v_k\rangle\!\rangle_2^{1/2}\\ &= \Big(\sum_{j,k=1}^{N} c_j\,c_k\,p(\lambda_j)\,p(\lambda_k)\,\underbrace{\langle\, v_j\,,\,v_k\,\rangle_2}_{=\,\delta_{jk}}\Big)^{1/2} \;=\; \Big(\sum_{k=1}^{N} c_k^2\,p(\lambda_k)^2\Big)^{1/2},\end{aligned}$$

und analog erhält man

$$\begin{aligned}\|\,p(A)x\,\|_A &= \langle\!\langle A\Big(\sum_{j=1}^{N} c_j\,p(\lambda_j)v_j\Big)\,,\ \sum_{k=1}^{N} c_k\,p(\lambda_k)\,v_k\rangle\!\rangle_2^{1/2}\\ &= \langle\!\langle \sum_{j=1}^{N} c_j\,\lambda_j\,p(\lambda_j)\,v_j\,,\ \sum_{k=1}^{N} c_k\,p(\lambda_k)\,v_k\rangle\!\rangle_2^{1/2}\\ &= \Big(\sum_{j,k=1}^{N} c_j\,c_k\,\lambda_j\,p(\lambda_j)\,p(\lambda_k)\,\underbrace{\langle\, v_j\,,\,v_k\,\rangle_2}_{=\,\delta_{jk}}\Big)^{1/2} \;=\; \Big(\sum_{k=1}^{N} c_k^2\,\lambda_k\,p(\lambda_k)^2\Big)^{1/2}.\end{aligned}$$

□

Den ersten Schritt auf dem Weg zur Herleitung spezieller Abschätzungen für $\|\,x_n - x_*\,\|_A$ liefert das folgende Theorem.

Theorem 11.17. *Zu einer gegebenen symmetrischen, positiv definiten Matrix $A \in \mathbb{R}^{N\times N}$ gelten für das CG-Verfahren die folgenden Fehlerabschätzungen:*

$$\|\,x_n - x_*\,\|_A \;\le\; \Big(\inf_{p\in\Pi_n,\ p(0)=1}\ \sup_{\lambda\in\sigma(A)} |\,p(\lambda)\,|\Big)\,\|\,x_*\,\|_A \qquad \textit{für}\quad n = 0,1,\ldots,n_*.$$

BEWEIS. Für jedes Polynom $p \in \Pi_n$ mit $p(0) = 1$ ist $q(t) := (1-p(t))/t$ ein Polynom vom Grad höchstens $n-1$, und somit gilt mit der Setzung $x := q(A)b$ Folgendes,

$$x \in \mathcal{K}_n(A,b), \qquad x - x_* \;=\; -p(A)x_*.$$

Mit Lemma 11.16 und der Entwicklung $x_* = \sum_{k=1}^{N} c_k v_k \in \mathbb{R}^N$ erhält man

$$\begin{aligned}\|\,x_n - x_*\,\|_A \;\le\; \overbrace{\|\,x - x_*\,\|_A}^{=\,\|\,p(A)x_*\,\|_A} &= \Big(\sum_{k=1}^{N} c_k^2\,\lambda_k\,p(\lambda_k)^2\Big)^{1/2}\\ &\le \sup_{\lambda\in\sigma(A)} |\,p(\lambda)\,|\,\underbrace{\Big(\sum_{k=1}^{N} c_k^2\,\lambda_k\Big)^{1/2}}_{=\,\|\,x_*\,\|_A}.\end{aligned}$$

Dies komplettiert den Beweis. □

Zur Herleitung spezieller Abschätzungen des Fehlers $x_n - x_*$ mittels Theorem 11.17 werden im Folgenden Tschebyscheff-Polynome der ersten Art herangezogen[8], die auf dem Intervall $[-1,1]$ die Darstellung $T_n(t) = \cos(n \arccos t)$ besitzen. Das folgende Lemma wird für die Herleitung der genannten speziellen Fehlerabschätzungen benötigt:

Lemma 11.18. *Für die Tschebyscheff-Polynome der ersten Art $T_0, T_1, \ldots$ gilt*

$$\begin{aligned} T_n(t) &= \tfrac{1}{2}\big[(t + \sqrt{t^2-1})^n + (t - \sqrt{t^2-1})^n\big] \quad \text{für } t \in \mathbb{R},\ |t| \geq 1, \\ T_n\big(\tfrac{\kappa+1}{\kappa-1}\big) &\geq \tfrac{1}{2}\big(\tfrac{\sqrt{\kappa}+1}{\sqrt{\kappa}-1}\big)^n \quad \text{für } \kappa \in \mathbb{R},\ \kappa > 1. \end{aligned} \tag{11.19}$$

BEWEIS. Auf dem Intervall $[-1,1]$ besitzt T_n die folgende Darstellung,

$$\begin{aligned} T_n(t) \overset{t=:\cos\theta}{=} \cos n\theta &= \tfrac{1}{2}[e^{in\theta} + e^{-in\theta}] \\ &= \tfrac{1}{2}\big[(\cos\theta + i\sin\theta)^n + (\cos\theta - i\sin\theta)^n\big] \\ &= \underbrace{\tfrac{1}{2}\big[(t + i\sqrt{1-t^2})^n + \big(t - i\sqrt{1-t^2}\big)^n\big]}_{=:\ p(t)} \quad \text{mit } t \in [-1,1]. \end{aligned} \tag{11.20}$$

Die nachfolgende Darstellung zeigt, dass die in (11.20) definierte Funktion $p(t)$ ein Polynom (vom Höchstgrad n) darstellt,

$$p(t) = \tfrac{1}{2}\sum_{j=0}^{n} \binom{n}{j} t^{n-j} i^j \underbrace{(\sqrt{1-t^2})^j}_{\in \Pi_j \text{ für } j/2 \in \mathbb{N}_0} \overbrace{(1 + (-1)^j)}^{=\ 0 \text{ für } j/2 \notin \mathbb{N}_0}, \qquad t \in \mathbb{R}.$$

Zusammenfassend lässt sich feststellen, dass T_n und p zwei Polynome darstellen, die auf dem Intervall $[-1,1]$ übereinstimmen, daher gilt notwendigerweise auch

$$T_n(t) = p(t) \quad \text{für } t \in \mathbb{R}.$$

Die im Lemma angegebene Darstellung von $T_n(t)$ für $|t| \geq 1$ folgt dann unmittelbar aus der Identität $i\sqrt{1-t^2} = \sqrt{t^2-1}$.

Für den Nachweis der Ungleichung (11.19) berechnet man für $\kappa \geq 1$

$$\begin{aligned} \frac{\kappa+1}{\kappa-1} \pm \sqrt{\Big(\frac{\kappa+1}{\kappa-1}\Big)^2 - 1} &= \frac{\kappa+1 \pm \sqrt{(\kappa+1)^2 - (\kappa-1)^2}}{\kappa-1} = \frac{\kappa+1 \pm 2\sqrt{\kappa}}{\kappa-1} \\ &= \frac{(\sqrt{\kappa} \pm 1)^2}{\kappa-1} = \frac{\sqrt{\kappa} \pm 1}{\sqrt{\kappa} \mp 1}, \end{aligned}$$

[8] vergleiche Definition 1.22

und daraus resultiert die Behauptung,

$$T_n\big(\tfrac{\kappa+1}{\kappa-1}\big) \;=\; \tfrac{1}{2}\Big[\big(\tfrac{\sqrt{\kappa}+1}{\sqrt{\kappa}-1}\big)^n \;+\; \big(\tfrac{\sqrt{\kappa}-1}{\sqrt{\kappa}+1}\big)^n\Big] \;\geq\; \tfrac{1}{2}\big(\tfrac{\sqrt{\kappa}+1}{\sqrt{\kappa}-1}\big)^n.$$

□

Es werden nun die Resultate für die Konvergenzgeschwindigkeit des Verfahrens der konjugierten Gradienten vorgestellt.

Theorem 11.19. *Zu einer gegebenen symmetrischen, positiv definiten Matrix $A \in \mathbb{R}^{N\times N}$ gelten für das CG-Verfahren die folgenden Fehlerabschätzungen:*

$$\|x_n - x_*\|_A \;\leq\; 2\gamma^n \|x_*\|_A, \qquad\qquad n \;=\; 0,1,\ldots,$$

$$\|x_n - x_*\|_2 \;\leq\; 2\sqrt{\kappa_A}\,\gamma^n \|x_*\|_2, \qquad\qquad \text{——— « ———}$$

mit den Notationen $\kappa_A := \mathrm{cond}_2(A)$ und $\gamma := \frac{\sqrt{\kappa_A}-1}{\sqrt{\kappa_A}+1}$.

BEWEIS. Für den Nachweis der ersten Abschätzung wird im Normalfall $\kappa_A > 1$ Theorem 11.17 angewandt mit dem folgenden Polynom,

$$p(\lambda) \;:=\; \frac{T_n[(M+m-2\lambda)/(M-m)]}{T_n[(M+m)/(M-m)]}, \qquad \lambda \in \mathbb{R},$$

wobei die Zahlen m und M wie schon in (11.18) den kleinsten beziehungsweise größten Eigenwert der Matrix A bezeichnen. Offensichtlich gilt $p \in \Pi_n$ und $p(0) = 1$, wegen $\sigma(A) \subset [m, M]$ und

$$\max_{m\leq\lambda\leq M} |p(\lambda)| \;=\; |T_n\big(\tfrac{M+m}{M-m}\big)|^{-1} \;=\; |T_n\big(\tfrac{\kappa_A+1}{\kappa_A-1}\big)|^{-1} \;\overset{(11.19)}{\leq}\; 2\gamma^n$$

erhält man die erste Abschätzung des Theorems für die Situation $\kappa_A > 1$. (Der degenerierte Fall $\kappa_A = 1$ ist gleichbedeutend mit $A = \lambda I$ für ein $\lambda > 0$ und führt auf $x_1 = x_*$.) Die zweite Abschätzung des Theorems ist eine unmittelbare Konsequenz aus der ersten Abschätzung und der Normäquivalenz (11.18). □

11.5 Das CG-Verfahren für die Normalgleichungen

Ist das reguläre lineare Gleichungssystem $Ax = b$ symmetrisch indefinit oder aber nichtsymmetrisch, so kann man zu den Normalgleichungen

$$A^\top A x \;=\; A^\top b$$

übergehen und hierauf das klassische CG-Verfahren anwenden. Diesen Ansatz bezeichnet man als *CGNR-Verfahren*.

Bemerkung 11.20. (a) Als unmittelbare Konsequenz aus Theorem 11.5 ergibt sich für die Iterierten des CGNR-Verfahrens die Minimaleigenschaft

$$\| Ax_n - b \|_2 \;=\; \min_{x \in \mathcal{K}_n(A^\top A, A^\top b)} \| Ax - b \|_2. \tag{11.21}$$

Diese Eigenschaft (11.21) begründet den Buchstaben "R" in der Notation CGNR, da in dieser Variante das Residuum minimiert wird, und der Buchstabe "N" steht für "Normalgleichungen". Aufgrund der Eigenschaft (11.21) ist auch unmittelbar klar, dass das CGNR-Verfahren für die spezielle Wahl $\mathcal{D}_n = \mathcal{K}_n(A^\top A, A^\top b)$, $n = 0, 1, \ldots$, mit dem Ansatz des minimalen Residuums (11.3) übereinstimmt.

(b) Einen Algorithmus zur Bestimmung der Iterierten des CGNR-Verfahrens erhält man durch Übertragung des Algorithmus 11.14 angewandt auf die Normalgleichungen $A^\top Ax = A^\top b$. Dabei sind in jedem Iterationsschritt zwei Matrix-Vektor-Multiplikationen erforderlich (zur Berechnung von Ad_n und $A^\top Ad_n$). Man beachte, dass die numerisch kostspielige Berechnung der Matrix $A^\top A$ dafür nicht erforderlich ist.

(c) Als Konsequenz aus Theorem 11.19 erhält man für das CGNR-Verfahren die folgenden Fehlerabschätzungen:

$$\| Ax_n - b \|_2 \;\le\; 2\gamma^n \| b \|_2, \qquad n \;=\; 0, 1, \ldots,$$

$$\| x_n - x_* \|_2 \;\le\; 2\kappa_A \gamma^n \| x_* \|, \qquad \text{——— «} \text{———}$$

mit den Notationen $\kappa_A := \mathrm{cond}_2(A)$ und $\gamma := \frac{\kappa_A - 1}{\kappa_A + 1}$. Man beachte, dass die in Theorem 11.19 auftretenden Größen $\sqrt{\kappa_A}$ hier durch κ_A ersetzt werden mussten, was sich bei schlecht konditionierten Problemen ($\kappa_A \gg 1$) als ungünstig erweist. △

11.6 Arnoldi-Prozess

11.6.1 Vorbetrachtungen zum GMRES-Verfahren

Eine weitere Möglichkeit zur Lösung eines regulären linearen Gleichungssystems $Ax = b$ mit symmetrisch indefiniter oder aber nichtsymmetrischer Matrix $A \in \mathbb{R}^{N \times N}$ liefert das GMRES-Verfahren:

Definition 11.21. Das *GMRES-Verfahren* ist definiert durch den Ansatz des minimalen Residuums (11.3) mit der speziellen Wahl $\mathcal{D}_n = \mathcal{K}_n(A, b)$, es gilt also

$$x_n \;\in\; \mathcal{K}_n(A, b), \quad \| Ax_n - b \|_2 \;=\; \min_{x \in \mathcal{K}_n(A,b)} \| Ax - b \|_2, \quad n \;=\; 1, 2, \ldots . \tag{11.22}$$

Die Abkürzung "GMRES" hat ihren Ursprung in der englischen Bezeichnung "generalized minimal residual method". Ursprünglich wurde dieses Verfahren für symmetrische Matrizen A betrachtet und dabei mit *MINRES* bezeichnet. Für $n = 1, 2, \ldots$ ist die grundsätzliche Vorgehensweise zur Realisierung des GMRES-Verfahrens folgendermaßen:

(a) Mittels des gleich zu beschreibenden Arnoldi-Prozesses wird bezüglich des euklidischen Skalarprodukts eine Orthogonalbasis von $\mathcal{K}_n(A,b)$ erzeugt.

(b) Mittels dieser Orthogonalbasis lässt sich das Minimierungsproblem (11.22) als ein einfacheres Minimierungsproblem formulieren, das schnell gelöst werden kann. Details hierzu werden in Abschnitt 11.7 vorgestellt.

Der vorliegende Abschnitt 11.6 befasst sich mit dem in (a) angesprochenen Arnoldi-Prozess.

11.6.2 Arnoldi-Prozess

Die Vorgehensweise beim Arnoldi-Prozess ist schnell beschrieben: ausgehend von einem gegebenen normierten Vektor $q_1 \in \mathbb{R}^N$ wird bezüglich des klassischen Skalarprodukts $\langle \cdot, \cdot \rangle_2$ eine Folge paarweise orthonormaler Vektoren $q_1, q_2, \ldots$ generiert durch Gram-Schmidt-Orthogonalisierung der Vektoren $q_1, Aq_1, Aq_2, \ldots$[9]. Der folgende Algorithmus beschreibt die genaue Vorgehensweise.

Algorithmus 11.22 (Arnoldi-Prozess). Ausgehend von einem gegebenem Vektor $0 \neq b \in \mathbb{R}^N$ setzt man $q_1 = b/\|b\|_2 \in \mathbb{R}^N$ und geht folgendermaßen vor für $n = 1, 2, \ldots$:

(a) (Orthogonalisierung) Man setzt

$$h_{jn} := (Aq_n)^\top q_j \in \mathbb{R}, \qquad j = 1, 2, \ldots, n, \tag{11.23}$$

$$\widehat{q}_{n+1} := Aq_n - \sum_{j=1}^{n} h_{jn} q_j \in \mathbb{R}^N. \tag{11.24}$$

(b) (Normierung) Im Fall $\widehat{q}_{n+1} = 0$ bricht der Prozess ab; der Abbruchindex wird mit $n_* = n$ bezeichnet. Wenn andererseits $\widehat{q}_{n+1} \neq 0$ gilt, so setzt man

$$h_{n+1,n} := \|\widehat{q}_{n+1}\|_2 \in \mathbb{R}, \qquad q_{n+1} := \frac{1}{\|\widehat{q}_{n+1}\|_2} \widehat{q}_{n+1} \in \mathbb{R}^N. \tag{11.25}$$

△

Bemerkung 11.23. (a) Der Arnoldi-Prozess hat eine eigenständige Bedeutung und kann beispielsweise auch zur numerischen Behandlung von Eigenwertproblemen eingesetzt werden; mehr Details hierzu später[10].

(b) Den Setzungen (11.23)–(11.24) entnimmt man, dass der Arnoldi-Prozess genau dann abbricht, wenn erstmalig $Aq_n \in \operatorname{span}\{q_1, \ldots, q_n\}$ gilt. △

Das folgende Lemma stellt die wichtigsten Eigenschaften im Zusammenhang mit dem Arnoldi-Prozess zusammen.

[9] Die zu orthogonalisierenden Vektoren werden also erst im Verlauf des Prozesses generiert und sind nicht von vornherein gegeben.

[10] siehe Bemerkung 11.27

Lemma 11.24. *Die durch den Arnoldi-Prozess erzeugten Vektoren $q_1, q_2, \ldots, q_{n_*} \in \mathbb{R}^N$ sind paarweise orthonormal, und es gilt*

$$\operatorname{span}\{q_1, q_2, \ldots, q_n\} \;=\; \operatorname{span}\{q_1, \ldots, q_{n-1},\, Aq_{n-1}\} \;=\; \mathcal{K}_n(A,b) \qquad (11.26)$$

für $n = 1, 2, \ldots, n_*$. Ist die Matrix A regulär, so gilt für die eindeutige Lösung $x_* \in \mathbb{R}^N$ des Gleichungssystems $Ax = b$ Folgendes,

$$x_* \;\in\; \mathcal{K}_{n_*}(A,b). \qquad (11.27)$$

BEWEIS. Die paarweise Orthogonalität erhält man mittels vollständiger Induktion über n (unter Verwendung von (11.23)):

$$q_{n+1}^\top q_j \;=\; \frac{1}{h_{n+1,n}}[(Aq_n)^\top q_j - h_{jn}] \;=\; 0, \qquad \begin{array}{l} j = 1, 2, \ldots, n, \\ n = 1, 2, \ldots, n_* - 1. \end{array}$$

Schließlich gewährleistet die Setzung (11.25) die Eigenschaft $\|q_{n+1}\|_2 = 1$. Die beiden Identitäten in (11.26) sollen nun mit vollständiger Induktion über n nachgewiesen werden. Wegen $q_1 = b/\|b\|_2$ ist die Behauptung richtig für $n = 1$, und es wird nun der Induktionsschritt $1 \le n-1 \to n \le n_*$ geführt. Aufgrund von $n \le n_*$ sind die Vektoren $q_1, \ldots, q_{n-1}, Aq_{n-1} \in \mathbb{R}^N$ linear unabhängig, so dass nach Konstruktion die erste Identität in (11.26) richtig ist. Die zweite Identität in (11.26) erhält man so: die Relation "$\subset$" folgt aus $Aq_{n-1} \in A(\mathcal{K}_{n-1}(A,b)) \subset \mathcal{K}_n(A,b)$; die Identität "$=$" ergibt sich dann aus Dimensionsgründen:

$$n \;=\; \dim \operatorname{span}\{q_1, \ldots, q_{n-1},\, Aq_{n-1}\} \;\le\; \dim \mathcal{K}_n(A,b) \;\le\; n.$$

Die Aussage in (11.27) erhält man so: nach Definition von n_* in Algorithmus 11.22 gilt $Aq_{n_*} \in \operatorname{span}\{q_1, \ldots, q_{n_*}\} = \mathcal{K}_{n_*}(A,b)$, und per Konstruktion gilt

$$Aq_j \;\in\; \mathcal{K}_{j+1}(A,b) \;\subset\; \mathcal{K}_{n_*}(A,b), \qquad j \;=\; 1, 2, \ldots, n_* - 1,$$

so dass insgesamt $A(\mathcal{K}_{n_*}(A,b)) \subset \mathcal{K}_{n_*}(A,b)$ gilt beziehungsweise aus Dimensionsgründen die Abbildung $A : \mathcal{K}_{n_*}(A,b) \to \mathcal{K}_{n_*}(A,b)$ bijektiv ist, und wegen $b \in \mathcal{K}_{n_*}(A,b)$ gilt dann – wie in (11.27) angegeben – notwendigerweise auch $x_* \in \mathcal{K}_{n_*}(A,b)$. Dies komplettiert den Beweis. □

Bemerkung 11.25. (a) Mit der Aussage (11.26) wird klar, dass $\dim \mathcal{K}_n(A,b) = n$ für $n = 1, 2, \ldots, n_*$ gilt. Einige weitere Eigenschaften von Krylovräumen werden zu einem späteren Zeitpunkt vorgestellt[11]. Der Arnoldi-Prozess bricht also notwendigerweise nach höchstens N Schritten ab, $n_* \le N$.

(b) In Schritt n des Arnoldi-Prozesses sind $2N(N-1)$ arithmetische Operationen zur Berechnung von Aq_n erforderlich. Zudem fallen noch $(3+2n)N$ arithmetische Operationen zur Bestimmung von $h_{jn} \in \mathbb{R}$, $j = 1, \ldots, n+1$ und $q_{n+1} \in \mathbb{R}^N$ an. Im Fall $n_* = N$ ergeben sich insgesamt $3N^3 + \mathcal{O}(N^2)$ arithmetische Operationen.

[11] siehe Lemma 11.31

(c) Ist die Matrix A symmetrisch, $A = A^\top$, so gilt für $j \le n-2$ die Identität $h_{jn} = q_n^\top A q_j = 0$ aufgrund der Eigenschaften $Aq_j \in \mathcal{K}_{j+1}(A,b) \subset \mathcal{K}_{n-1}(A,b)$ und $q_n \in \mathcal{K}_n(A,b)^\perp$. Die Gram-Schmidt-Orthogonalisierung (11.23)–(11.24) geht hier also über in eine Drei-Term-Rekursion (das heißt, für die Berechnung von q_{n+1} werden nur q_n und q_{n-1} benötigt):

$$\widehat{q}_{n+1} \;:=\; Aq_n - h_{nn}q_n - h_{n-1,n}q_{n-1}, \qquad n = 1,2,\ldots,n_*.$$

Diesen Spezialfall für den Arnoldi-Prozess bezeichnet man als *Lanczos-Prozess.* △

Matrixversion des Arnoldi-Prozesses

Für die weiteren Anwendungen ist die folgende Matrixversion des Arnoldi-Prozesses von Bedeutung.

Theorem 11.26. *Für eine gegebene Matrix $A \in \mathbb{R}^{N\times N}$ und einen Vektor $0 \neq b \in \mathbb{R}^N$ gelten mit den Notationen aus dem Arnoldi-Prozess die folgenden Identitäten:*

$$A \underbrace{\begin{pmatrix} | & & | \\ q_1 & \cdots & q_n \\ | & & | \end{pmatrix}}_{=:\, Q_n \in \mathbb{R}^{N\times n}} = \begin{pmatrix} | & & | & | \\ q_1 & \cdots & q_n & q_{n+1} \\ | & & | & | \end{pmatrix} \underbrace{\begin{pmatrix} h_{11} & \cdots & h_{1n} \\ h_{21} & \ddots & \vdots \\ & \ddots & h_{nn} \\ & & h_{n+1,n} \end{pmatrix}}_{=:\, H_n \in \mathbb{R}^{(n+1)\times n}}, \quad n = 1,\ldots,n_*-1, \tag{11.28}$$

beziehungsweise im letzten Schritt

$$A \underbrace{\begin{pmatrix} | & & | \\ q_1 & \cdots & q_{n_*} \\ | & & | \end{pmatrix}}_{=:\, Q_{n_*} \in \mathbb{R}^{N\times n_*}} = \begin{pmatrix} | & & | \\ q_1 & \cdots & q_{n_*} \\ | & & | \end{pmatrix} \underbrace{\begin{pmatrix} h_{11} & \cdots & \cdots & h_{1n_*} \\ h_{21} & \ddots & & \vdots \\ & \ddots & \ddots & \vdots \\ & & h_{n_*,n_*-1} & h_{n_*n_*} \end{pmatrix}}_{=:\, H_{n_*} \in \mathbb{R}^{n_*\times n_*}}. \tag{11.29}$$

BEWEIS. Es genügt der Nachweis von (11.29), da die Matrixprodukte in (11.28) für $n = 1,\ldots,n_*-1$ jeweils gerade die ersten n Spalten der beiden Matrixprodukte von (11.29) darstellen. Ein Vergleich der n_* Spalten der Matrixprodukte in (11.29) führt auf $Aq_n = \sum_{j=1}^{n+1} h_{jn}q_j$ beziehungsweise

$$h_{n+1,n}q_{n+1} \;=\; Aq_n - \sum_{j=1}^{n} h_{jn}q_j, \qquad n = 1,2,\ldots,n_*-1,$$

sowie auf $Aq_{n_*} = \sum_{j=1}^{n_*} h_{jn_*}q_j$. Dies entspricht exakt den Setzungen (11.23)–(11.25) des Arnoldi-Prozesses. □

Bemerkung 11.27. (a) In Kurzform bedeuten die Darstellungen (11.28)–(11.29) Folgendes,

$$AQ_n \;=\; Q_{n+1}H_n \qquad (n \;=\; 1,2,\ldots,n_*-1), \qquad\qquad AQ_{n_*} \;=\; Q_{n_*}H_{n_*}. \tag{11.30}$$

(b) Bricht der Arnoldi-Prozess nicht vorzeitig ab, gilt also $n_* = N$, so erhält man eine Faktorisierung der Form

$$Q_N^\top A Q_N \;=\; \begin{pmatrix} h_{11} & \ldots & \ldots & h_{1N} \\ h_{21} & \ddots & & \vdots \\ & \ddots & \ddots & \vdots \\ & & h_{N,N-1} & h_{NN} \end{pmatrix} \in \mathbb{R}^{N\times N}, \qquad Q_N^\top \;=\; Q_N^{-1} \in \mathbb{R}^{N\times N},$$

so dass die Matrix A durch orthogonale Ähnlichkeitstransformationen auf *obere Hessenbergform* gebracht worden ist, das heißt, die resultierende Matrix unterscheidet sich von einer oberen Dreiecksmatrix lediglich durch die nichtverschwindenden Einträge auf der unteren Nebendiagonalen; eine solche Matrix bezeichnet man als *Hessenbergmatrix*. Eine Hessenbergform ist bei der numerischen Behandlung von Eigenwertproblemen von Vorteil, siehe Kapitel 13; dort werden auch andere orthogonale Ähnlichkeitstransformationen (Householder-Transformationen, Givens-Rotationen) zur Gewinnung einer Hessenbergform vorgestellt. △

11.7 Realisierung von GMRES auf der Basis des Arnoldi-Prozesses

11.7.1 Einführende Bemerkungen

Im Folgenden wird eine Methode zur Umsetzung des GMRES-Verfahrens vorgestellt, die die durch den Arnoldi-Prozess generierten Orthogonalbasen der Krylovräume $\mathcal{K}_1(A,b),\ \mathcal{K}_2(A,b),\ldots$ verwendet.

Theorem 11.28. *Mit den Notationen aus dem Arnoldi-Prozess gelten für die Vektoren $x_1,\ x_2,\ldots \in \mathbb{R}^N$ aus dem GMRES-Verfahren genau dann die Darstellungen*

$$x_n \;=\; Q_n z_n, \qquad\qquad n \;=\; 1,2,\ldots,n_*, \tag{11.31}$$

wenn für $n = 1,2,\ldots,n_*$ der Vektor $z_n \in \mathbb{R}^n$ das folgende Minimierungsproblem löst,

$$\|\,H_n z \;-\; c_n\,\|_2 \;\to\; \min \quad \textit{für}\ z \in \mathbb{R}^n, \quad \textit{mit}\ c_n \;:=\; \begin{pmatrix} \|\,b\,\|_2 \\ 0 \\ \vdots \\ 0 \end{pmatrix} \in \mathbb{R}^{\min\{n+1,n_*\}}. \tag{11.32}$$

BEWEIS. Für jeden Index $n \leq n_* - 1$ und jeden Vektor $z \in \mathbb{R}^n$ gilt

$$\| AQ_n z - b \|_2 \;=\; \| Q_{n+1} H_n z \;-\; Q_{n+1} c_n \|_2 \;=\; \| H_n z \;-\; c_n \|_2, \tag{11.33}$$

wobei die Norm $\| \cdot \|_2$ in (11.33) die ersten beiden Male auf $\mathbb{R}^N$ und und im dritten Fall auf $\mathbb{R}^{n+1}$ operiert; die letzte Identität in (11.33) resultiert aus der Isometrieeigenschaft $\| Q_n y \|_2 = \| y \|_2$. Für den Index $n = n_*$ verhält sich die Situation nicht viel anders; man hat nur in dem mittleren Ausdruck von (11.33) die beiden auftretenden Indizes $n + 1$ jeweils durch n zu ersetzen. □

11.7.2 Allgemeine Vorgehensweise zur Lösung des Minimierungsproblems (11.32)

Im vorigen Abschnitt 11.7.1 ist auf der Basis des Arnoldi-Prozesses das Problem der Bestimmung der Approximationen $x_1, x_2, \ldots \in \mathbb{R}^N$ des GMRES-Verfahrens reduziert worden auf die Lösung des linearen Ausgleichsproblems (11.32). Im Folgenden wird dargestellt, wie man die dabei auftretende Matrix H_n mit oberer Hessenbergstruktur schnell in eine orthogonale Matrix und eine verallgemeinerte obere Dreiecksmatrix von der folgenden Form faktorisiert:

- Für $n = 1, 2, \ldots, n_* - 1$ bestimmt man sukzessive Faktorisierungen der Form

$$H_n \;=\; T_n \begin{pmatrix} R_n \\ \hline \mathbf{0}^\top \end{pmatrix}, \quad T_n \in \mathbb{R}^{(n+1)\times(n+1)}, \quad R_n \;=\; \begin{pmatrix} * & \cdots & * \\ & \ddots & \vdots \\ & & * \end{pmatrix} \in \mathbb{R}^{n\times n}, \; \mathbf{0} \in \mathbb{R}^n. \tag{11.34}$$
$$T_n^{-1} \;=\; T_n^\top,$$

Nach der Bestimmung solcher Faktorisierungen kann das jeweilige Ausgleichsproblem (11.32) unmittelbar gelöst werden durch die Auflösung des folgenden gestaffelten Gleichungssystems:[12]

$$R_n z \;=\; y \in \mathbb{R}^n, \quad \text{mit} \quad \begin{pmatrix} y \\ \hline * \end{pmatrix} := T_n^\top \begin{pmatrix} \| b \|_2 \\ 0 \\ \vdots \\ 0 \end{pmatrix} \in \mathbb{R}^{n+1} \qquad (n \;=\; 1, 2, \ldots, n_* - 1).$$

- Für den Index $n = n_*$ verhält sich die Situation nur geringfügig anders. Hier bestimmt man eine Faktorisierung der Form

$$H_{n_*} \;=\; T_{n_*} R_{n_*}, \quad T_{n_*} \in \mathbb{R}^{n_*\times n_*}, \quad R_{n_*} \;=\; \begin{pmatrix} * & \cdots & * \\ & \ddots & \vdots \\ & & * \end{pmatrix} \in \mathbb{R}^{n_*\times n_*}, \tag{11.35}$$
$$T_{n_*}^{-1} \;=\; T_{n_*}^\top,$$

[12] Eine einführende Behandlung dieser Vorgehensweise finden Sie in Abschnitt 4.8.5.

und die Lösung des linearen Ausgleichsproblems (11.32) (die in dieser Situation gleichzeitig die Lösung von $Ax = b$ darstellt) kann dann leicht über das folgende gestaffelte Gleichungssystem bestimmt werden,

$$R_{n_*} z \;=\; T_{n_*}^\top \begin{pmatrix} \|\, b \,\|_2 \\ 0 \\ \vdots \\ 0 \end{pmatrix} \;\in\; \mathbb{R}^{n_*}.$$

Im folgenden Abschnitt 11.7.3 wird beschrieben, wie man auf effiziente Art Faktorisierungen der Form (11.34)–(11.35) gewinnt.

11.7.3 Detaillierte Beschreibung der Vorgehensweise zur Lösung des Minimierungsproblems (11.32)

Im Folgenden wird beschrieben, wie man für fixierten Index $n \le n_*$ ausgehend von einer Faktorisierung der Form

$$H_{n-1} \;=\; T_{n-1} \left(\begin{array}{c} R_{n-1} \\ \hline \mathbf{0}^\top \end{array}\right) \;\in\; \mathbb{R}^{n\times(n-1)}, \qquad \mathbf{0} \in \mathbb{R}^{n-1},$$

verfährt, um im Fall $n \le n_* - 1$ eine Faktorisierung der Form (11.34) und im Fall $n = n_*$ eine Faktorisierung von der Gestalt (11.35) zu erhalten.

• Wie bisher auch soll zunächst die Situation $n \le n_* - 1$ behandelt werden. Da die Hessenbergmatrix H_n eine einfache Erweiterung von H_{n-1} darstellt, ist die folgende orthogonale Transformation von H_n naheliegend,

$$\left(\begin{array}{c|c} T_{n-1}^\top & \mathbf{0} \\ \hline \mathbf{0}^\top & 1 \end{array}\right) \underbrace{\left(\begin{array}{c|c} H_{n-1} & \begin{matrix} h_{1n} \\ \vdots \\ h_{nn} \end{matrix} \\ \hline \mathbf{0}^\top & h_{n+1,n} \end{array}\right)}_{=\,H_n} = \left(\begin{array}{c|c} T_{n-1}^\top H_{n-1} & T_{n-1}^\top \begin{pmatrix} h_{1n} \\ \vdots \\ h_{nn} \end{pmatrix} \\ \hline \mathbf{0}^\top & h_{n+1,n} \end{array}\right)$$

$$= \left(\begin{array}{c|c} R_{n-1} & \begin{matrix} r_{1n} \\ \vdots \\ r_{n-1,n} \end{matrix} \\ \hline \mathbf{0}^\top & * \\ \hline \mathbf{0}^\top & * \end{array}\right), \quad \text{mit} \quad \begin{pmatrix} r_{1n} \\ \vdots \\ r_{n-1,n} \\ * \end{pmatrix} := T_{n-1}^\top \begin{pmatrix} h_{1n} \\ \vdots \\ h_{nn} \end{pmatrix}. \tag{11.36}$$

Die untere der beiden mit "$*$" bezeichneten Zahlen stimmt mit $h_{n+1,n}$ überein, was im Folgenden aber keine Rolle mehr spielt. Man beachte, dass bei dieser Transformation

tatsächlich nur eine Matrix-Vektor-Multiplikation (von der Gestalt $T_{n-1}^\top x$) zur Berechnung des letzten Spaltenvektors anfällt, da die Dreiecksmatrix R_{n-1} als bekannt angenommen ist. Nun ist noch der Vektor $(*,*)^\top \in \mathbb{R}^2$ orthogonal in ein Vielfaches des ersten Einheitsvektors zu transformieren, ohne dabei den Rest der in (11.36) auftretenden Matrix zu verändern. Hierzu wird der Vektor $w^{[n]} \in \mathbb{R}^2$, $\|w^{[n]}\|_2 = 1$, gemäß Lemma 4.62 auf Seite 92 so bestimmt, dass für die Householdermatrix $W^{[n]} = I_2 - 2w^{[n]}(w^{[n]})^\top \in \mathbb{R}^{2\times 2}$ Folgendes gilt,

$$W^{[n]}\begin{pmatrix}*\\*\end{pmatrix} = \begin{pmatrix}r_{nn}\\0\end{pmatrix} \quad \text{bzw. äquivalent} \quad \left(\begin{array}{c|c} I_{n-1} & \\ \hline & W^{[n]} \end{array}\right)\begin{pmatrix} r_{1n}\\ \vdots \\ r_{n-1,n}\\ *\\ *\end{pmatrix} = \begin{pmatrix} r_{1n}\\ \vdots \\ r_{n-1,n}\\ r_{nn}\\ 0\end{pmatrix},$$

wobei wieder $I_s \in \mathbb{R}^{s\times s}$ die Einheitsmatrix bezeichnet. So hat man bereits die gewünschte Faktorisierung gewonnen,

$$\underbrace{\left(\begin{array}{c|c} I_{n-1} & \mathbf{0} \\ \hline \mathbf{0}^\top & W^{[n]} \end{array}\right)\left(\begin{array}{c|c} T_{n-1}^\top & \mathbf{0} \\ \hline \mathbf{0}^\top & 1 \end{array}\right)}_{=:\ T_n^\top} H_n = \begin{pmatrix} * & \cdots & * \\ & \ddots & \vdots \\ & & * \\ & & 0 \end{pmatrix} \in \mathbb{R}^{(n+1)\times n}.$$

• Nun soll noch die Situation $n = n_*$ behandelt werden, die sich geringfügig von dem Fall $n \le n_* - 1$ unterscheidet. Hier führt man die folgende Transformation aus,

$$T_{n_*-1}^\top \underbrace{\left(\begin{array}{c|c} H_{n_*-1} & \begin{matrix} h_{1n_*}\\ \vdots \\ h_{n_*n_*}\end{matrix}\end{array}\right)}_{=\ H_{n_*}} = \left(\begin{array}{c|c} R_{n_*-1} & \begin{matrix} r_{1n_*}\\ \vdots \\ r_{n_*-1,n_*}\\ r_{n_*n_*}\end{matrix} \\ \hline \mathbf{0}^\top & \end{array}\right) =: \quad R_{n_*},$$

$$\text{mit} \quad \begin{pmatrix} r_{1n_*}\\ \vdots \\ r_{n_*n_*}\end{pmatrix} := T_{n_*-1}^\top \begin{pmatrix} h_{1n_*}\\ \vdots \\ h_{n_*n_*}\end{pmatrix},$$

bei der lediglich eine Matrix-Vektor-Multiplikation von der Art $T_{n_*-1}^\top x$ anfällt. Die gewünschte Faktorisierung liegt nun schon vor; eine anschließende Elimination ist hier nicht erforderlich, so dass die Wahl $T_{n_*} = T_{n_*-1}$ zum Ziel führt.

Bemerkung 11.29. (a) Eine unmittelbare Folgerung aus der vorgestellten Vorgehensweise sind die folgenden Darstellungen,

$$R_n = \begin{pmatrix} r_{11} & \dots & r_{1n} \\ & \ddots & \vdots \\ & & r_{nn} \end{pmatrix} \in \mathbb{R}^{n\times n}, \qquad n = 1,2,\dots,n_*,$$

$$T_n^\top = S_n^{[n]}\, S_{n-1}^{[n]} \cdots S_1^{[n]}, \quad \text{mit } S_j^{[n]} := \left(\begin{array}{c|c|c} I_{j-1} & & \\ \hline & W^{[j]} & \\ \hline & & I_{n-j} \end{array}\right) \in \mathbb{R}^{(n+1)\times(n+1)},$$
$$j = 1,2,\dots,n, \qquad n = 1,2,\dots,n_*-1,$$

beziehungsweise $T_{n_*} = T_{n_*-1}$. Naheliegenderweise verwendet man diese Faktorisierungen von $T_n^\top$ für die numerischen Berechnungen, die Berechnung eines Matrix-Vektor-Produkts von der Form $T_{n-1}^\top x$ wird also über n zweidimensionale Matrix-Vektor-Multiplikationen realisiert.

(b) Man beachte, dass bei der Lösung des Minimierungsproblems (11.32) in jedem Schritt n lediglich $\mathcal{O}(N)$ arithmetische Operationen erforderlich sind, so dass die numerische Hauptlast auf dem Arnoldi-Prozess ruht. Insgesamt lässt sich festhalten, dass für jeden Schritt des GMRES-Verfahrens lediglich eine Matrix-Vektor-Multiplikation sowie Operationen niedrigen Aufwands benötigt werden, es fallen also $2N^2 + \mathcal{O}(N)$ arithmetische Operationen pro Iterationsschritt an. Dies ist ein Gewinn gegenüber dem CGNR-Verfahren, bei dem zwei Matrix-Vektor-Multiplikationen pro Iterationsschritt erforderlich sind. Auf der anderen Seite ist anzumerken, dass GMRES sich nicht wie das CGNR-Verfahren als einfache Zweitermrekursion realisieren lässt und der Speicherplatzbedarf wegen der benötigten Matrizen R_n und orthogonalen Vektoren q_n, $n = 1,2,\dots$ höher ausfällt. Schließlich gestaltet sich die Gewinnung von Fehlerabschätzungen für das GMRES-Verfahren schwieriger, wie sich im nachfolgenden Abschnitt herausstellen wird. △

11.7.4 MATLAB-Programm für GMRES

Im Folgenden wird ein MATLAB-Programm (das auch unter OCTAVE lauffähig ist) für das GMRES-Verfahren auf der Basis des Arnoldi-Prozesses angegeben. Die Matrix $A \in \mathbb{R}^{N\times N}$ sowie der Vektor $b \in \mathbb{R}^N$ sind dabei als gegeben angenommen. Der Algorithmus bricht in dieser Variante mit dem Schritt $n = n_*$ ab, er fungiert hier also als direkter Löser. Auf der Webseite zu diesem Buch finden Sie weitergehende Erläuterungen zu diesem Programm.

```
% ................. gmres.m .....................

x = zeros(N,1);    res = zeros(N,1);    u = zeros(2,1);
d = zeros(2,1);    Q = zeros(N,N);    R = zeros(N,N);
w = zeros(2,N);    y = zeros(N,1);    y(1) = norm(b);
```

```
h = zeros(N,1);    goahead = 1;    n = 1;
                          %(*** Ende der Initialisierungen ***)
Q(:,1) = b/norm(b);    myeps = 0.000001;
          %(*** Start der Iteration; n = Iterationsschritt ***)
while (goahead == 1)    v = A*Q(:,n);    z = v;
   for j= 1:n
      h(j) = Q(:,j)'*v;    z = z - h(j)*Q(:,j);
   end
   qhat = z;    normqhat = norm(qhat);
   if ( (normqhat <= myeps) | (n==N) )    goahead = 0;
   else    h(n+1)=normqhat;    Q(:,n+1) = qhat/normqhat;
   end
    %(*** Anwendung vergangener orthogonaler Transformation ***)
   for j= 1:n-1
      u = h(j:j+1);    d = w(:,j);    h(j:j+1) = u - 2*(d'*u)*d;
   end
   if (n >= 2)    R(1:n-1,n) = h(1:n-1);    end
         %(*** Berechnung der neuen orthog. Transformation ***)
   if (goahead == 0)    R(n,n) = h(n);
   else
      u = h(n:n+1);
      if (abs(u(1)) <= myeps) sigma = norm(u);
      else sigma = norm(u)*u(1)/abs(u(1));
      end
      u(1) = u(1) + sigma;    w(:,n) = u/norm(u);
      R(n,n) = -sigma;    R(n+1,n) = 0;
      u = y(n:n+1);    d = w(:,n);    y(n:n+1) = u - 2*(d'*u)*d;
   end
        %(*** Auflösung des gestaffelten Gleichungssystems ***)
   for j= n:-1:1
      sum = y(j);
      for k= j+1:n
         sum = sum - R(j,k)*z(k);
      end
      z(j) = sum/R(j,j);
   end
   x = Q(:,1:n)*z(1:n);    res(n) = norm(A*x - b);
   if (goahead==1) n = n+1; end
                  %(*** Ende des n-ten Iterationsschrittes ***)
end                                %(*** Ende der Iteration ***)
```

Allgemeine Informationen zu MATLAB, OCTAVE und anderen Programmsystemen finden Sie in Abschnitt 11.10.

11.8 Konvergenzgeschwindigkeit des GMRES-Verfahrens

In der Praxis liefert das GMRES-Verfahren schon nach wenigen Iterationsschritten gute Approximationen $x_n \approx x_*$. Ausgangspunkt für etwaige Fehlerabschätzungen zum GMRES-Verfahren ist das folgende Theorem über das Abklingverhalten der Norm des Residuums, das ein Analogon zu Theorem 11.17 für das Verfahren der konjugierten Gradienten darstellt.

Theorem 11.30. *Die Matrix $A \in \mathbb{R}^{N\times N}$ sei diagonalisierbar, $T^{-1}AT = \operatorname{diag}(\lambda_1,\dots,\lambda_N) =: D \in \mathbb{C}^{N\times N}$ mit einer regulären Matrix $T \in \mathbb{R}^{N\times N}$. Dann gilt für das GMRES-Verfahren Folgendes,*

$$\|Ax_n - b\|_2 \le \operatorname{cond}_2(T)\Big(\inf_{p\in\Pi_n,\ p(0)=1}\ \max_{k=1,..,N} |p(\lambda_k)|\Big)\|b\|_2,\ n = 1,2,\dots,n_* \quad (11.37)$$

mit dem Stoppindex n_* aus dem Arnoldi-Prozess (siehe Algorithmus 11.22).

BEWEIS. Sei $p \in \Pi_n$ irgendein Polynom mit $p(0) = 1$. Dann ist $q(t) := (1 - p(t))/t$ ein Polynom vom Grad höchstens $n-1$, und somit gilt

$$x := q(A)b \in \mathcal{K}_n(A,b) \qquad \text{bzw.} \qquad Ax - b = -p(A)b.$$

Aus der Minimaleigenschaft des GMRES-Verfahrens folgt

$$\|Ax_n - b\|_2 \le \|Ax - b\|_2 = \|p(A)b\|_2 \le \|p(A)\|_2\|b\|_2,$$

und unter Berücksichtigung von $p(A) = Tp(D)T^{-1}$ schätzt man weiter wie folgt ab,

$$\|p(A)\|_2 \le \|T\|_2\|p(D)\|_2\|T^{-1}\|_2 = \operatorname{cond}_2(T)\max_{k=1,..,N}|p(\lambda_k)|.$$

Daraus resultiert die Aussage des Theorems. □

Anders als beim Verfahren der konjugierten Gradienten für symmetrische positiv definite Matrizen A, deren Eigenwerte auf der positiven reellen Achse liegen, ist eine Abschätzung des in (11.37) auftretenden Ausdrucks $\inf_{p\in\Pi_n,\ p(0)=1}\max_{k=1,..,N}|p(\lambda_k)|$ in der vorliegenden allgemeinen Situation nicht ohne weiteres möglich. Weitergehende Untersuchungen für symmetrische aber indefinite Matrizen A findet man in Fischer [27]. In speziellen Situationen ist die Verwendung des CGNR-Verfahrens gegenüber der des GMRES-Verfahrens vorzuziehen (Aufgabe 11.5).

11.9 Nachtrag 1: Krylovräume

Abschließend wird das Wachstumsverhalten von Krylovräumen etwas eingehender betrachtet: mit einer eindeutig bestimmten Zahl $0 \le n_* \le N$ gilt Folgendes,

$$\left.\begin{aligned} &\underbrace{\mathcal{K}_0(A,b)}_{=\{0\}} \subsetneqq \mathcal{K}_1(A,b) \subsetneqq .. \subsetneqq \mathcal{K}_{n_*-1}(A,b) \subsetneqq \mathcal{K}_{n_*}(A,b) = \mathcal{K}_{n_*+1}(A,b) = .. \\ &x_* \in \mathcal{K}_{n_*}(A,b), \qquad x_* \notin \mathcal{K}_n(A,b) \quad \text{für } n = 0,1,\dots,n_*-1, \end{aligned}\right\} \quad (11.38)$$

wobei sich die in (11.38) angegebenen Eigenschaften unmittelbar aus dem nachfolgenden Lemma ergeben.

Lemma 11.31. *Für eine reguläre Matrix $A \in \mathbb{R}^{N\times N}$ sowie einen Vektor $b \in \mathbb{R}^N$ sind für jede Zahl $n \geq 1$ die folgenden Aussagen äquivalent:*

(a) Die Vektoren $b, Ab, \ldots, A^n b$ sind linear abhängig;

(b) $\mathcal{K}_n(A,b) = \mathcal{K}_{n+1}(A,b)$;

(c) $A(\mathcal{K}_n(A,b)) \subset \mathcal{K}_n(A,b)$;

(d) es existiert ein linearer Unterraum $\mathcal{M} \subset \mathbb{R}^N$ mit $\dim \mathcal{M} \leq n$, für den $b \in \mathcal{M}$ gilt und der bezüglich der Matrix A invariant ist, $A(\mathcal{M}) \subset \mathcal{M}$;

(e) für $x_* = A^{-1}b$ gilt $x_* \in \mathcal{K}_n(A,b)$.

BEWEIS. Die Äquivalenzen ergeben sich folgendermaßen:
$(a) \Rightarrow (b)$: Nach Voraussetzung existieren eine Zahl $0 \leq m \leq n$ und Konstanten γ_0, $\gamma_1, \ldots, \gamma_{m-1} \in \mathbb{R}$ mit

$$A^m b = \sum_{\nu=0}^{m-1} \gamma_\nu A^\nu b.$$

Daraus resultiert dann

$$A^n b = \sum_{\nu=n-m}^{n-1} \gamma_{\nu-(n-m)} A^\nu b \in \mathcal{K}_n(A,b).$$

$(b) \Rightarrow (c)$: Dies folgt sofort aus $A(\mathcal{K}_n(A,b)) \subset \mathcal{K}_{n+1}(A,b) = \mathcal{K}_n(A,b)$.
$(c) \Rightarrow (d)$: Man wähle $\mathcal{M} = \mathcal{K}_n(A,b)$.
$(d) \Rightarrow (a)$: Nach Voraussetzung gilt $A^\nu b \in \mathcal{M}$ für $\nu = 0, 1, \ldots$, und daher gilt $\dim \operatorname{span}\{b, Ab, \ldots, A^n b\} \leq n$ beziehungsweise (a).
$(c) \Rightarrow (e)$: Aus Dimensionsgründen ist die Abbildung $A : \mathcal{K}_n(A,b) \to \mathcal{K}_n(A,b)$ eine Bijektion, und wegen $b \in \mathcal{K}_n(A,b)$ besitzt somit die Gleichung $Ax = b$ in dem Krylovraum $\mathcal{K}_n(A,b)$ eine Lösung, die wegen der Injektivität von A notwendigerweise mit x_* übereinstimmt.
$(e) \Rightarrow (a)$: Nach Annahme gilt $x_* \in \operatorname{span}\{b, Ab, \ldots, A^{n-1}b\}$ beziehungsweise $b = Ax_* \in \operatorname{span}\{b^2, b^3, \ldots, A^n b\}$, woraus die behauptete lineare Abhängigkeit folgt. □

11.10 Nachtrag 2: Interaktive Programmsysteme mit Multifunktionalität

Bei dem auf Seite 316 angesprochenen Programmsystem MATLAB handelt es sich um ein interaktives Programmsystem mit Multifunktionalität. Die interaktive Arbeitsweise von Programmsystemen im Allgemeinen erlaubt dabei jeweils eine komfortable und rasche Bearbeitung unterschiedlicher und schnell wechselnder Problemstellungen. Unter Multifunktionalität ist dabei vorrangig Numerik-, Computeralgebra- und Grafik-Funktionalität zu verstehen. Hierbei bedeutet *Numerik-Funktionalität* die Bereitstellung von fertigen Routinen beispielsweise zur Lösung linearer und nichtlinearer Gleichungssysteme, zur Durchführung der schnellen Fouriertransformation

oder zur numerischen Lösung von Anfangswertproblemen bei gewöhnlichen Differenzialgleichungen. Unter dem Begriff *Computeralgebra-Funktionalität* versteht man die Fähigkeit eines Programmsystems, die aus der Analysis oder Algebra bekannten Vorgehensweisen vorzunehmen. Dazu gehört beispielsweise die Berechnung der Ableitung oder der Stammfunktion von Funktionen einer Veränderlichen ebenso wie die exakte Berechnung von Eigenwerten von Matrizen oder die exakte Bestimmung von Nullstellen von Polynomen. In allen Fällen sind Variablen als Koeffizienten zugelassen. Solche Berechnungen mittels Computeralgebra-Systemen bezeichnet man allgemein als *symbolisches Rechnen*. Naturgemäß können Programmsysteme mit Computeralgebra-Funktionalität auch nur solche Probleme lösen, für die überhaupt analytisch Lösungen angegeben werden können. Die Nullstellen von Polynomen vom Grad ≥ 5 etwa können im Allgemeinen auch mit solchen Programmsystemen nicht bestimmt werden. *Grafik-Funktionalität* ermöglicht beispielsweise die Darstellung von Funktionen $f : \mathbb{R}^2 \supset \mathcal{D} \to \mathbb{R}$ als Niveauflächen in dreidimensionalen Abbildungen.

Bei MATLAB handelt es sich um das derzeit wohl bekannteste interaktive Programmsystem mit ausgeprägter Numerik- und Grafik-Funktionalität. Es bietet interaktiv Routinen zu allen in dem vorliegenden Buch vorgestellten und weiteren Problemstellungen an. Zusätzlich existieren Module beispielsweise für das symbolische Rechnen, wobei dieses auf dem Computeralgebra-System MAPLE basiert. Eingesetzt wird MATLAB in Lehre und Forschung an Hochschulen und in Unternehmen beispielswiese aus der chemischen Industrie, der Automobil- oder der Stahlindustrie.

Weitere interaktive Programmsysteme mit Multifunktionalität sind OCTAVE und SCILAB (vorwiegend Numerik- und Grafik-Funktionalität) sowie MAPLE, MuPAD und MATHEMATICA (vorrangig Computeralgebra- und Grafik-Funktionalität; Numerik-Funktionalität ist ebenfalls vorhanden). Andere Computeralgebra-Systeme sind KANT V4 (Computational Algebraic Number Theory, [13]; ein Programmsystem mit Computeralgebra-Funktionalität für den Bereich Zahlentheorie), MACSYMA sowie REDUCE. Internet-Adressen zu den einzelnen Programmsystemen finden Sie im Online-Service zu diesem Buch.

Weitere Themen und Literaturhinweise

Das Verfahren der konjugierten Gradienten geht zurück auf Hestenes/Stiefel [53] und wird zum Beispiel in den Lehrbüchern Fischer [27], Hackbusch [47], Kelley [60], Meister [70], Schwarz/Klöckner [94] und Stoer/Bulirsch [99] behandelt. Varianten des Verfahren vom Typ der konjugierten Gradienten werden beispielsweise in [70] sowie in Ashby/Manteuffel/Saylor [1], Freund/Golub/Nachtigal [30], Hanke-Bourgeois [52], Nachtigal/S. Reddy/L. Reddy [74], Saad/Schultz [90], Sonneveld [97], Stoer [98] und in Vuik/van der Vorst [108] diskutiert. Resultate zur Konvergenzgeschwindigkeit des GMRES-Verfahrens findet man beispielsweise in [52] sowie in Greenbaum/Pták/Strakoš [40], Liesen [65], Moret [73], Nevanlinna [77], Plato/Vainikko [83] und in van der Vorst/Vuik [107]. Eine Einführung zu MATLAB findet man beispielsweise in Gramlich/Werner [39], und Anwendungen von MATLAB in der Finanzmathematik werden in Günther/Jüngel [45] vorgestellt.

Übungsaufgaben

Aufgabe 11.1. Zu gegebener symmetrischer, positiv definiter Matrix $A \in \mathbb{R}^{N\times N}$ und einem Vektor $b \in \mathbb{R}^N$ habe die Zahl n_* die Bedeutung aus (11.38). Man zeige: $x_* = A^{-1}b$ ist Linearkombination von n_* Eigenvektoren der Matrix A.

Aufgabe 11.2. Zu gegebener symmetrischer, positiv definiter Matrix $A \in \mathbb{R}^{N\times N}$ und einem Vektor $b \in \mathbb{R}^N$ zeige man: für jeden Index n ist das Residuum $r_n = Ax_n - b$ identisch mit dem Gradienten des Energiefunktionals $\mathcal{J}(x) = \frac{1}{2}\langle Ax, x\rangle_2 - \langle x, b\rangle_2$ an der Stelle x_n, es gilt also $r_n = \nabla\mathcal{J}(x_n)$.

Aufgabe 11.3. Zu gegebener symmetrischer, positiv definiter Matrix $A \in \mathbb{R}^{N\times N}$ zeige man:

(a) Man weise für das CG-Verfahren für $n = 1, 2, \ldots, n_*$ die folgenden Darstellungen nach:

$$\begin{aligned} x_n &= q_n(A)b \quad \text{mit} \quad q_n \in \Pi_{n-1} \text{ geeignet,} \\ r_n &= -p_n(A)b \quad \text{mit} \quad p_n(t) = 1 - tq_n(t). \end{aligned}$$

(b) Der zur Entwicklung $q_n(t) = \sum_{k=0}^{n-1} c_k t^k$ gehörende Koeffizientenvektor $(c_0, \ldots, c_{n-1})^\top \in \mathbb{R}^n$ ist Lösung des linearen Gleichungssystems

$$\begin{pmatrix} b^\top Ab & b^\top A^2b & \cdots & b^\top A^nb \\ b^\top A^2b & b^\top A^3b & \cdots & b^\top A^{n+1}b \\ \vdots & \vdots & \ddots & \vdots \\ b^\top A^nb & b^\top A^{n+1}b & \cdots & b^\top A^{2n-1}b \end{pmatrix} \begin{pmatrix} c_0 \\ c_1 \\ \vdots \\ c_{n-1} \end{pmatrix} = \begin{pmatrix} b^\top b \\ b^\top Ab \\ \vdots \\ b^\top A^{n-1}b \end{pmatrix}.$$

Aufgabe 11.4. Zu gegebener symmetrischer, positiv definiter Matrix $A \in \mathbb{R}^{N\times N}$ weise man für das CG-Verfahren die folgenden Beziehungen nach (für $n = 0, 1, \ldots, n_*$, mit der Zahl n_* aus (11.38)):

$$\begin{aligned} r_n^\top d_n &= -\|r_n\|_2^2, & d_n &= -\|r_n\|_2^2 \sum_{k=0}^{n} \frac{r_k}{\|r_k\|_2^2}, \\ \|d_n\|_2^2 &= \|r_n\|_2^4 \sum_{k=0}^{n} \frac{1}{\|r_k\|_2^2}, & d_n^\top d_k &= \frac{\|r_n\|_2^2}{\|r_k\|_2^2}\|d_k\|_2^2 \quad \text{für} \quad k \le n, \\ \|x_{n+1}\|_2 &\ge \|x_n\|_2 \quad (n \le n_* - 1), & \|r_n\|_2 &\le \|d_n\|_2. \end{aligned}$$

Aufgabe 11.5. Es bezeichne

$$A = \begin{pmatrix} 0 & & & 1 \\ 1 & 0 & & \\ & \ddots & \ddots & \\ & & 1 & 0 \end{pmatrix} \in \mathbb{R}^{N\times N}, \qquad b = \begin{pmatrix} 1 \\ 0 \\ \vdots \\ 0 \end{pmatrix} \in \mathbb{R}^N, \qquad x_* = \begin{pmatrix} 0 \\ \vdots \\ 0 \\ 1 \end{pmatrix} \in \mathbb{R}^N,$$

so dass x_* die eindeutige Lösung des linearen Gleichungssystems $Ax = b$ darstellt; die Matrix A wird als *zirkulant* bezeichnet. Man zeige:

(a) Das GMRES-Verfahren liefert die Vektoren $x_1 = x_2 = \ldots = x_{N-1} = 0$ und $x_N = x_*$, das heißt, das GMRES-Verfahren liefert in den Schritten $n = 1, 2, \ldots, N-1$ keine Approximationen an die Lösung x_*, auch eine schrittweise Verbesserung tritt nicht auf.

(b) Dagegen liefert das CGNR-Verfahren nach nur einem Iterationsschritt die exakte Lösung, $x_1 = x_*$.

12 Eigenwertprobleme

12.1 Einleitung

Die mathematische Formulierung des Schwingungsverhalten von mechanischen oder elektrischen Systemen sowie die anschließende Diskretisierung des Modells führt auf das Problem der Bestimmung der Eigenwerte von Matrizen (kurz als *Eigenwertproblem* bezeichnet), was in der Regel numerisch geschieht. Für solche Eigenwertprobleme werden in dem vorliegenden Kapitel Störungs-, Einschließungs- und Variationssätze vorgestellt, und in dem darauf folgenden Kapitel 13 werden numerische Verfahren zur Lösung von Eigenwertproblemen behandelt.

12.2 Störungstheorie für Eigenwertprobleme

In diesem Abschnitt wird diskutiert, inwieweit sich Änderungen der Einträge einer Matrix $A \in \mathbb{R}^{N\times N}$ auf die Menge ihrer Eigenwerte auswirken. Etwas genauer werden zu gegebener Matrix $\Delta A \in \mathbb{R}^{N\times N}$ Abschätzungen dafür angegeben, wie groß der Abstand der Eigenwerte der gestörten Matrix $A + \Delta A \in \mathbb{R}^{N\times N}$ zum jeweils nächstgelegenen Eigenwert von A höchstens sein kann. Umgekehrte Abschätzungen – wie groß also der Abstand der Eigenwerte von A zum jeweils nächstgelegenen Eigenwert von $A + \Delta A$ höchstens sein kann – sind von geringerem praktischem Interesse und werden hier auch nicht behandelt.

Das folgende Theorem liefert ein entsprechendes Resultat für diagonalisierbare Matrizen. Die allgemeine Situation wird in Theorem 12.5 betrachtet.

12.2.1 Diagonalisierbare Matrizen

Theorem 12.1 (Bauer/Fike). *Die Matrix $A \in \mathbb{R}^{N\times N}$ sei diagonalisierbar,*

$$T^{-1}AT \;=\; \operatorname{diag}(\lambda_1,\ldots,\lambda_N) \;\in\; \mathbb{C}^{N\times N}, \tag{12.1}$$

mit der regulären Matrix $T \in \mathbb{C}^{N\times N}$ und den Eigenwerten $\lambda_1,\ldots,\lambda_N \in \mathbb{C}$ von A. Dann gelten für die Eigenwerte der Matrix $A + \Delta A \in \mathbb{R}^{N\times N}$ mit $\Delta A \in \mathbb{R}^{N\times N}$ beliebig die folgenden Abschätzungen,

$$\forall\, \mu \;\in\; \sigma(A+\Delta A): \qquad \min_{k=1..N} |\mu - \lambda_k| \;\le\; \operatorname{cond}_p(T)\,\|\Delta A\|_p. \tag{12.2}$$

Hier ist $1 \le p \le \infty$, und $\|\cdot\|_p$ bezeichnet die durch die gleichnamige Vektornorm induzierte Matrixnorm.

BEWEIS. Sei $\mu \in \mathbb{C}$ ein Eigenwert der Matrix $A + \Delta A$. Falls μ gleichzeitig ein Eigenwert der Matrix A ist, so folgt die Aussage unmittelbar, im Folgenden sei also $\mu \notin \sigma(A)$

angenommen. Wegen

$$A \;=\; TDT^{-1}, \qquad\qquad D \;:=\; \operatorname{diag}(\lambda_1,\ldots,\lambda_N),$$

gilt

$$(A-\mu I)^{-1} \;=\; T(D-\mu I)^{-1}T^{-1},$$

und somit

$$\begin{aligned}\|(A-\mu I)^{-1}\|_p \;&\le\; \operatorname{cond}_p(T)\|(D-\mu I)^{-1}\|_p \;\overset{(*)}{=}\; \operatorname{cond}_p(T)\max_{k=1..N}|\mu-\lambda_k|^{-1}\\ &=\; \operatorname{cond}_p(T)\big(\min_{k=1..N}|\mu-\lambda_k|\big)^{-1}\end{aligned}$$

beziehungsweise

$$\min_{k=1..N}|\mu-\lambda_k| \;\le\; \frac{\operatorname{cond}_p(T)}{\|(A-\mu I)^{-1}\|_p}, \tag{12.3}$$

wobei in $(*)$ die für Diagonalmatrizen $\widehat{D}=\operatorname{diag}(d_1,\ldots,d_N)\in\mathbb{C}^{N\times N}$ gültige Identität $\|\widehat{D}\|_p=\max_{k=1..N}|d_k|$ eingeht. Zur weiteren Abschätzung von (12.3) betrachtet man einen Eigenvektor $x\in\mathbb{C}^N$ von $A+\Delta A$ zum Eigenwert μ und erhält

$$(A+\Delta A)x \;=\; \mu x \qquad \text{bzw.}\quad Ax-\mu x \;=\; -\Delta Ax \qquad \text{bzw.}\quad x \;=\; -(A-\mu I)^{-1}\Delta Ax$$

und somit

$$1 \;\le\; \|(A-\mu I)^{-1}\Delta A\|_p \;\le\; \|(A-\mu I)^{-1}\|_p\|\Delta A\|_p, \tag{12.4}$$

und dies in (12.3) verwendet liefert die Aussage des Theorems. □

Das Eigenwertproblem ist für symmetrische Matrizen also stabil in dem Sinne, dass für $p=2$ die Konstante $\operatorname{cond}_2(T)$ in (12.2) minimal ausfällt:

Korollar 12.2. *Für eine symmetrische Matrix $A\in\mathbb{R}^{N\times N}$ und jede Matrix $\Delta A\in\mathbb{R}^{N\times N}$ gilt*

$$\min_{\lambda\in\sigma(A)}|\mu-\lambda| \;\le\; \|\Delta A\|_2 \qquad \textit{für}\quad \mu\;\in\;\sigma(A+\Delta A).$$

BEWEIS. In der vorliegenden Situation darf in (12.1) die Matrix $T\in\mathbb{R}^{N\times N}$ orthogonal gewählt werden, $T^{-1}=T^\top$ (vergleiche auch Abschnitt 12.6.2). Dann gilt insbesondere $\|T\|_2=\|T^{-1}\|_2=1$, so dass die Behauptung unmittelbar aus Theorem 12.1 folgt. □

Bemerkung 12.3. Die Transformationsmatrix T in (12.1) besteht aus N linear unabhängigen Eigenvektoren von A. Die Konditionszahl von T wird dann größer ausfallen, wenn etwa zwei von diesen Eigenvektoren fast parallel sind. △

Beispiel 12.4. Die diagonalisierbare Matrix

$$A = \begin{pmatrix} a & 1 \\ \eta & a \end{pmatrix}, \qquad 0 < \eta \ll 1, \tag{12.5}$$

besitzt die Eigenwerte $\lambda_{1/2} = a \pm \sqrt{\eta}$, und als Transformationsmatrix und deren Inverse ergibt sich zum Beispiel

$$T = \begin{pmatrix} 1 & 1 \\ \sqrt{\eta} & -\sqrt{\eta} \end{pmatrix}, \qquad T^{-1} = \frac{1}{2}\begin{pmatrix} 1 & 1/\sqrt{\eta} \\ 1 & -1/\sqrt{\eta} \end{pmatrix}, \qquad \mathrm{cond}_\infty(T) = 1 + \frac{1}{\sqrt{\eta}},$$

so dass in diesem Beispiel die Abschätzung (12.2) für $0 < \eta \ll 1$ eine große Schranke liefert. In einem solchen Fall ist die Anwendung des folgenden Theorems 12.5 vorzuziehen. In diesem Zusammenhang sei auch auf Beispiel 12.7 verwiesen, wo gerade der Grenzfall $\eta = 0$ aus (12.5) behandelt wird. △

12.2.2 Der allgemeine Fall

Mit der Schur-Faktorisierung[1] einer gegebenen Matrix A lässt sich in der allgemeinsten Situation für A die Empfindlichkeit der Menge der Eigenwerte von A gegenüber Störungen in den Einträgen dieser Matrix messen.

Theorem 12.5. *Für die Eigenwerte von $A + \Delta A \in \mathbb{R}^{N\times N}$ mit beliebigen Matrizen $A,\ \Delta A \in \mathbb{R}^{N\times N}$ gelten die folgenden Abschätzungen,*

$$\forall \mu \in \sigma(A + \Delta A) : \min_{\lambda\in\sigma(A)} |\mu - \lambda| \le c \max\left\{ \|\Delta A\|_2,\ \|\Delta A\|_2^{1/N} \right\},\ c := \max\{\theta, \theta^{1/N}\},$$
$$\theta := \sum_{s=0}^{N-1} \|R\|_2^s.$$

Hier ist $R \in \mathbb{C}^{N\times N}$ der nichtdiagonale Anteil aus einer Schur-Faktorisierung $Q^{-1}AQ = D + R \in \mathbb{C}^{N\times N}$ mit einer Diagonalmatrix $D \in \mathbb{C}^{N\times N}$. Die Matrixnorm $\|\cdot\|_2$ ist induziert durch die euklidische Vektornorm.

BEWEIS. Sei $\mu \in \mathbb{C}$ ein Eigenwert der Matrix $A + \Delta A$. Falls μ auch Eigenwert der Matrix A ist, folgt die angegebene Abschätzung unmittelbar. Im Folgenden sei also $\mu \notin \sigma(A)$. Aus der auch für nichtdiagonalisierbare Matrizen gültigen Abschätzung (12.4) erhält man mit $A = Q(D + R)Q^{-1}$

$$1 \le \|(D - \mu I + R)^{-1}\|_2 \|\Delta A\|_2. \tag{12.6}$$

Weiter gilt

$$\left.\begin{aligned} (D - \mu I + R)^{-1} &= (I + (D - \mu I)^{-1}R)^{-1}(D - \mu I)^{-1} \\ &\overset{(*)}{=} \sum_{s=0}^{N-1} (-(D - \mu I)^{-1}R)^s (D - \mu I)^{-1}, \end{aligned}\right\} \tag{12.7}$$

[1] Eine Definition wird in Abschnitt 12.6 ab Seite 332 nachgetragen.

wobei in $(*)$ eingeht, dass $(D-\mu I)^{-1}R$ eine strikte obere Dreiecksmatrix darstellt (also verschwindende Diagonaleinträge besitzt) und somit $((D-\mu I)^{-1}R)^N = 0$ gilt, wobei man Letzteres leicht nachrechnet. Weiter ist noch zu beachten, dass für Matrizen $B \in \mathbb{C}^{N\times N}$ mit $B^p = 0$ für ein $p \in \mathbb{N}_0$ Folgendes gilt, $(I-B)^{-1} = \sum_{s=0}^{p-1} B^s$, was man ebenfalls leicht nachrechnet.

Für die abzuschätzende Größe

$$\varepsilon \;:=\; \min_{\lambda\in\sigma(A)} |\mu-\lambda| \;=\; \frac{1}{\|(D-\mu I)^{-1}\|_2}$$

erhält man somit aus (12.6)-(12.7)

$$1 \;\le\; \frac{1}{\varepsilon}\sum_{s=0}^{N-1}\Big(\frac{\|R\|_2}{\varepsilon}\Big)^s \|\Delta A\|_2 \;\le\; \underbrace{\sum_{s=0}^{N-1}\|R\|_2^s}_{=\;\theta} \left\{ \begin{array}{ll} 1/\varepsilon, & \text{falls } \varepsilon \ge 1 \\ 1/\varepsilon^N, & \text{sonst} \end{array} \right\} \|\Delta A\|_2,$$

woraus sich unmittelbar die Aussage des Theorems ergibt. □

Bemerkung 12.6. In der typischen Situation $\|\Delta A\|_2 \ll 1$ geht die Abschätzung in Theorem 12.5 über in

$$\min_{\lambda\in\sigma(A)} |\mu-\lambda| \;\le\; c\,\|\Delta A\|_2^{1/N} \qquad \text{für} \quad \mu \in \sigma(A+\Delta A). \tag{12.8}$$

Für diagonalisierbare Matrizen A ist diese Abschätzung (12.8) bezüglich des Terms ΔA aufgrund von $\|\Delta A\|_2 \ll \|\Delta A\|_2^{1/N}$ für große N schwächer als die Abschätzung (12.2). △

Beispiel 12.7. Für die Matrix

$$A \;=\; \begin{pmatrix}\lambda & 1\\ 0 & \lambda\end{pmatrix} \;=\; \underbrace{\begin{pmatrix}\lambda & 0\\ 0 & \lambda\end{pmatrix}}_{=:\;D} + \underbrace{\begin{pmatrix}0 & 1\\ 0 & 0\end{pmatrix}}_{=:\;R}$$

mit einer beliebigen Zahl $\lambda \in \mathbb{C}$ gilt mit der Notation aus Theorem 12.5 die Identität $\|R\|_2 = 1$ und somit $\theta = 2$. Für

$$A \;+\; \Delta A \;=\; \begin{pmatrix}\lambda & 1\\ \delta & \lambda\end{pmatrix}, \qquad 0 < \delta \;\ll\; 1,$$

ist dann $\|\Delta A\|_2 = \delta$ erfüllt, so dass Theorem 12.5 die Abschätzung

$$\forall\, \mu \;\in\; \sigma(A \;+\; \Delta A): \qquad \min_{\lambda\in\sigma(A)} |\mu-\lambda| \;\le\; 2\sqrt{\delta}$$

liefert. Tatsächlich gilt

$$\sigma(A) \;=\; \{\lambda\}, \quad \sigma(A+\Delta A) \;=\; \{\lambda \;\pm\; \sqrt{\delta}\}, \quad \max_{\mu\in\sigma(A+\Delta A)} \min_{\lambda\in\sigma(A)} |\mu-\lambda| \;=\; \sqrt{\delta}. \qquad \square$$

Sowohl Theorem 12.1 als auch Theorem 12.5 liefern Abschätzungen für die Empfindlichkeit der *Menge* $\sigma(A)$ gegenüber Störungen in den Einträgen der Matrix A. Aussagen über die Änderung der entsprechend ihrer *algebraischen Vielfachheit* gezählten Eigenwerte werden jedoch erst durch das folgende Theorem möglich, das hier ohne Beweis angegeben wird.

Theorem 12.8. *Für eine Matrix $A \in \mathbb{R}^{N\times N}$ mit den Eigenwerten $\lambda_1, \ldots, \lambda_N \in \mathbb{C}$ existiert zu jedem $\varepsilon > 0$ ein $\delta > 0$ mit der folgenden Eigenschaft: Zu jeder Matrix $\Delta A \in \mathbb{R}^{N\times N}$ mit $\|\Delta A\| \le \delta$ gibt es eine Nummerierung $\mu_1, \ldots, \mu_N \in \mathbb{C}$ der Eigenwerte von $A + \Delta A$ mit*

$$\max_{k=1,\ldots,N} |\mu_k - \lambda_k| \;\le\; \varepsilon. \tag{12.9}$$

Hierbei bezeichnet $\|\cdot\| : \mathbb{R}^{N\times N} \to \mathbb{R}_+$ eine Matrixnorm.

BEWEIS. Siehe Mennicken/Wagenführer [71] oder Werner [111]. □

12.3 Lokalisierung von Eigenwerten

Im Folgenden wird ein wichtiges Einschließungsresultat für Eigenwerte vorgestellt.

Theorem 12.9. *(a) Für eine Matrix $A = (a_{jk}) \in \mathbb{R}^{N\times N}$ gilt*

$$\sigma(A) \;\subset\; \bigcup_{j=1}^{N} \mathcal{G}_j,$$

mit den Gerschgorin-Kreisen

$$\mathcal{G}_j \;:=\; \Big\{ z \in \mathbb{C} \;:\; |z - a_{jj}| \;\le\; \sum_{\substack{k=1\\k\neq j}}^{N} |a_{jk}| \Big\}, \qquad j \;=\; 1, 2, \ldots, N.$$

(b) Wenn genauer die Vereinigung von q Gerschgorin-Kreisen

$$K^{(1)} := \mathcal{G}_{j_1} \cup \ldots \cup \mathcal{G}_{j_q} \qquad (j_\ell \neq j_m \quad \text{für } \ell \neq m)$$

disjunkt von der Menge der Vereinigung $K^{(2)}$ der restlichen $N - q$ Gerschgorin-Kreise ist, so enthält $K^{(1)}$ genau q Eigenwerte von A und $K^{(2)}$ enthält genau $N - q$ Eigenwerte von A (jeweils entsprechend ihrer algebraischen Vielfachheit gezählt).

BEWEIS. (a) Für $\lambda \in \mathbb{C}$ ist die Bedingung $\lambda \notin \bigcup_{j=1}^{N} \mathcal{G}_j$ gleichbedeutend mit

$$|\lambda - a_{jj}| \;>\; \sum_{\substack{k=1\\k\neq j}}^{N} |a_{jk}|, \qquad j \;=\; 1, 2, \ldots, N,$$

was wiederum gerade die strikte Diagonaldominanz der Matrix $A - \lambda I \in \mathbb{C}^{N\times N}$ impliziert. Daher ist $A - \lambda I \in \mathbb{C}^{N\times N}$ nichtsingulär[2], die Zahl λ also kein Eigenwert von A. Damit ist Teil (a) nachgewiesen. Für den Nachweis der Aussage in (b) zerlegt man die Matrix $A = (a_{jk})$ in die Summe eines diagonalen und eines nichtdiagonalen Anteils, $A = D + M$ mit

$$D \;=\; \operatorname{diag}(a_{11},\ldots,a_{NN}) \;\in\; \mathbb{R}^{N\times N}, \qquad M \;=\; A - D,$$

und betrachtet in $\mathbb{R}^{N\times N}$ die Strecke von D nach A,

$$A(t) \;=\; D \;+\; tM \;=\; \begin{pmatrix} a_{11} & ta_{12} & \cdots & \cdots & ta_{1N} \\ ta_{21} & a_{22} & ta_{23} & \cdots & ta_{2N} \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ \vdots & & \ddots & \ddots & ta_{N-1,N} \\ ta_{N1} & \cdots & \cdots & ta_{N,N-1} & a_{NN} \end{pmatrix} \in \mathbb{R}^{N\times N}, \quad 0 \le t \le 1,$$

so dass $A(0) = D$ und $A(1) = A$ gilt. In den folgenden Punkten (i)–(iii) werden nun einige Vorbereitungen getroffen für den anschließend in Punkt (iv) beschriebenen entscheidenden Beweisschritt.

(i) Als Erstes soll

$$\sigma(A(t)) \;\subset\; K^{(1)} \,\cup\, K^{(2)} \qquad \text{für} \quad 0 \;\le\; t \;\le\; 1 \tag{12.10}$$

nachgewiesen werden. Hierzu bezeichne $\mathcal{G}_1(t),\ldots,\mathcal{G}_N(t)$ die zu $A(t)$ gehörenden Gerschgorin-Kreise,

$$\mathcal{G}_j(t) \;=\; \Big\{ z \in \mathbb{C} \;:\; |z - a_{jj}| \;\le\; t \sum_{\substack{k=1 \\ k\neq j}}^{N} |a_{jk}| \Big\}, \qquad j \;=\; 1,2,\ldots,N.$$

Offensichtlich gilt

$$\mathcal{G}_j(t) \;\subset\; \mathcal{G}_j, \qquad j \;=\; 1,2,\ldots,N \qquad \text{für} \quad 0 \le t \le 1,$$

und mit Teil (a) dieses Theorems erhält man (12.10).

(ii) Von den insgesamt N Eigenwerten von $D = A(0)$ befinden sich die q Eigenwerte $a_{j_1 j_1},\ldots,a_{j_q j_q}$ in der Menge $K^{(1)}$ und die restlichen $N-q$ Eigenwerte liegen in $K^{(2)}$, was unmittelbar aus der Eigenschaft $a_{jj} \in \mathcal{G}_j$ für $j = 1,2,\ldots,N$ folgt.

(iii) Weiter beobachtet man vorbereitend noch

$$\varepsilon \;:=\; \operatorname{dist}(K^{(1)}, K^{(2)}) \;>\; 0, \tag{12.11}$$

was aus der Disjunktheitsvoraussetzung $K^{(1)} \cap K^{(2)} \;=\; \varnothing$ und der Abgeschlossenheit der Mengen $K^{(1)}$ und $K^{(2)}$ folgt.

[2] siehe Lemma 2.13 auf Seite 29

(iv) Die Eigenschaften (12.10)–(12.11) und die Schlussfolgerung in (ii) zusammen mit der stetigen Abhängigkeit der Eigenwerte gegenüber Matrixstörungen ergeben nun Teil (b) des Theorems, wie im Folgenden noch detailliert nachgewiesen wird. Hierzu bezeichne

$$t_0 \;:=\; \sup\big\{\, t \in [0,1] : \;\; \text{genau } q \text{ Eigenwerte von } A(t) \text{ liegen in } K^{(1)} \,\big\}. \qquad (12.12)$$

Die Menge in (12.12) enthält $t = 0$ und ist somit nichtleer. Wenn $\lambda_1(t_0), \ldots, \lambda_N(t_0) \in \mathbb{C}$ die der algebraischen Vielfachheit nach gezählten Eigenwerte von $A(t_0)$ bezeichnen, so existiert nach Theorem 12.8 zu ε aus (12.11) eine Zahl $\delta > 0$ und eine Nummerierung $\lambda_1(t), \ldots, \lambda_N(t) \in \mathbb{C}$ der Eigenwerte von $A(t)$ mit

$$\max_{k=1,\ldots,N} |\lambda_k(t) - \lambda_k(t_0)| \;<\; \varepsilon \qquad \text{für} \quad t \in [0,1], \qquad |t - t_0| < \delta. \qquad (12.13)$$

Aus der Eigenschaft (12.13) folgt zweierlei: zum einen wird das Maximum in (12.12) angenommen, denn gemäß der Definition von t_0 gibt es ein $t \in [0,1]$ mit $t_0 - \delta < t \le t_0$, so dass die Menge $K^{(1)}$ genau q Eigenwerte von $A(t)$ enthält, und genau $N - q$ Eigenwerte von $A(t)$ sind in $K^{(2)}$ enthalten. (Die Situation ist in Bild 12.1 veranschaulicht.) Wegen (12.13), (12.10) und (12.11) enthält die Menge $K^{(1)}$ mithin auch genau q Eigenwerte von $A(t_0)$.

Zum anderen ist noch $t_0 = 1$ nachzuweisen; wegen $A(1) = A$ ergibt sich daraus die Aussage des Theorems. Wäre $t_0 < 1$, so enthielte für jedes $t \in [0,1]$ mit $t_0 < t \le t_0 + \delta$ die Menge $K^{(1)}$ genau q Eigenwerte von $A(t)$ (wieder aufgrund der Eigenschaften (12.13), (12.10) und (12.11)). Dies stellt einen Widerspruch zur Definition (12.12) dar und komplettiert den Beweis der Aussage des Theorems. □

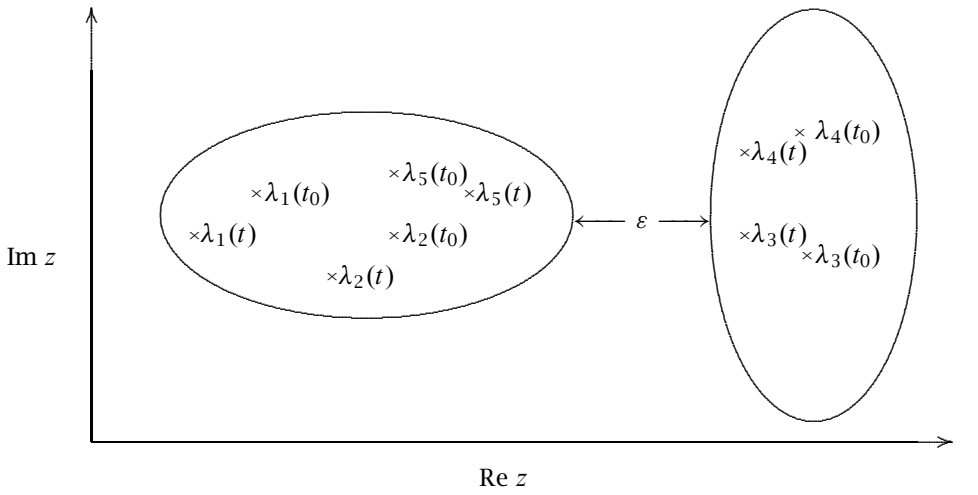


Bild 12.1: Veranschaulichung zweier Situationen im Beweis von Theorem 12.9 am Beispiel $N = 5$. Dargestellt ist die Verteilung der Eigenwerte von $A(t_0)$ und $A(t)$ für t mit $|t - t_0| \le \delta$. Die Ellipsen sollen die Mengen K_1 beziehungsweise K_2 umfassen.

Beispiel 12.10. Für die Matrix

$$A = \begin{pmatrix} 5 & 1/2 & 0 & 1/2 \\ 1/2 & 3 & 0 & 0 \\ 1/2 & 0 & 1 & 1/2 \\ 1/2 & 0 & 0 & 6 \end{pmatrix} \in \mathbb{R}^{4\times 4}$$

ist die Lage der Gerschgorin-Kreise in Bild 12.2 dargestellt. Aus Theorem 12.9 folgert man dann, dass es reelle Eigenwerte $0 \le \lambda_3 \le 2$ und $2.5 \le \lambda_2 \le 3.5$ gibt (komplexe Eigenwerte reeller Matrizen treten automatisch als konjugiert komplexe Paare auf). Die beiden anderen Eigenwerte liegen entweder im Intervall $[4, 6.5]$ oder sind durch ein komplex konjugiertes Paar in $\mathcal{G}_3 \cup \mathcal{G}_4$ gegeben.

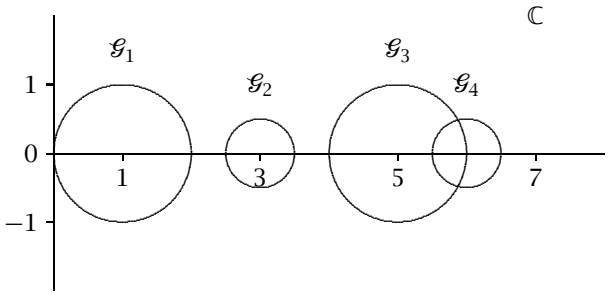


Bild 12.2: Gerschgorin-Kreise für Beispiel 12.10 △

12.4 Variationsformulierung für Eigenwerte von symmetrischen Matrizen

Im Folgenden spielen orthogonale Komplemente von Mengen $\mathcal{L} \subset \mathbb{R}^N$ eine Rolle,

$$\mathcal{L}^\perp := \{ y \in \mathbb{R}^N \ : \ y^\top x = 0 \quad \text{für jedes } x \in \mathcal{L} \}.$$

Es ist $\mathcal{L}^\perp \subset \mathbb{R}^N$ ein linearer Unterraum. Falls $\mathcal{L} \subset \mathbb{R}^N$ ein linearer Unterraum ist, so gilt $\mathcal{L} \oplus \mathcal{L}^\perp = \mathbb{R}^N$.

Theorem 12.11 (Courant/Fischer). *Für eine symmetrische Matrix $A \in \mathbb{R}^{N\times N}$ mit Eigenwerten $\lambda_1 \ge \lambda_2 \ge \ldots \ge \lambda_N$ gilt Folgendes,*

$$\lambda_{k+1} = \min_{\substack{\mathcal{L}\subset\mathbb{R}^N \text{ linear} \\ \dim \mathcal{L} \le k}} \max_{0\neq x\in\mathcal{L}^\perp} \frac{x^\top A x}{x^\top x} = \min_{y_1,\ldots,y_k\in\mathbb{R}^N} \max_{\substack{0 \neq x \in \mathbb{R}^N \\ x^\top y_\ell = 0,\ \ell=1,\ldots,k}} \frac{x^\top A x}{x^\top x}, \quad (12.14)$$

$$\lambda_{N-k} = \max_{\substack{\mathcal{L}\subset\mathbb{R}^N \text{ linear} \\ \dim \mathcal{L} \le k}} \min_{0\neq x\in\mathcal{L}^\perp} \frac{x^\top A x}{x^\top x} = \max_{y_1,\ldots,y_k\in\mathbb{R}^N} \min_{\substack{0 \neq x \in \mathbb{R}^N \\ x^\top y_\ell = 0,\ \ell=1,\ldots,k}} \frac{x^\top A x}{x^\top x}, \quad (12.15)$$

jeweils für $k = 0, 1, \ldots, N-1$.

BEWEIS. Es wird nur der Nachweis für (12.14) geführt, die Aussage (12.15) ergibt sich ganz entsprechend. Die zweite Identität in (12.14) ist unmittelbar einsichtig, und im Folgenden soll die erste Identität in (12.14) nachgewiesen werden. Dazu sei $u_1, \ldots, u_N \in \mathbb{R}^N$ ein vollständiges System von Eigenvektoren (zu den Eigenwerten $\lambda_1, \ldots, \lambda_N$), die aufgrund der Symmetrie der Matrix A zudem noch als paarweise orthonormal angenommen werden dürfen[3].

Zum Beweis der Ungleichung "$\leq$" in (12.14) sei $\mathcal{L} \subset \mathbb{R}^N$ ein beliebiger linearer Unterraum mit $\dim \mathcal{L} \leq k$. Dann gilt $\dim \mathcal{L}^\perp \geq N - k$, und wegen $\dim \operatorname{span}\{u_1, \ldots, u_{k+1}\} = k + 1$ existiert ein Vektor

$$x \in \operatorname{span}\{u_1, \ldots, u_{k+1}\} \cap \mathcal{L}^\perp, \qquad x^\top x = 1. \tag{12.16}$$

Für den Vektor x aus (12.16) gilt insbesondere die Darstellung

$$x = \sum_{\ell=1}^{k+1} \alpha_\ell u_\ell, \qquad \sum_{\ell=1}^{k+1} |\alpha_\ell|^2 = 1,$$

mit gewissen Koeffizienten $\alpha_1, \ldots, \alpha_{k+1} \in \mathbb{R}$. Weiter gilt $Ax = \sum_{\ell=1}^{k+1} \lambda_\ell \alpha_\ell u_\ell$ sowie

$$x^\top A x = \sum_{\ell=1}^{k+1} \lambda_\ell |\alpha_\ell|^2 \geq \lambda_{k+1} \sum_{\ell=1}^{k+1} |\alpha_\ell|^2 = \lambda_{k+1},$$

was wegen $x \in \mathcal{L}^\perp$ gerade die angegebene Abschätzung "$\leq$" in (12.14) liefert.

Für den Beweis der Abschätzung "$\geq$" in (12.14) sei speziell $\mathcal{L} := \operatorname{span}\{u_1, \ldots, u_k\}$ gewählt. Für jeden Vektor $x \in \mathcal{L}^\perp$ mit $x^\top x = 1$ gibt es eine Darstellung

$$x = \sum_{\ell=k+1}^{N} \alpha_\ell u_\ell, \qquad \sum_{\ell=k+1}^{N} |\alpha_\ell|^2 = 1,$$

mit gewissen Koeffizienten $\alpha_{k+1}, \ldots, \alpha_N \in \mathbb{R}$. Daraus erhält man die Identität $Ax = \sum_{\ell=k+1}^{N} \lambda_\ell \alpha_\ell u_\ell$, und weiter

$$x^\top A x = \sum_{\ell=k+1}^{N} \lambda_\ell |\alpha_\ell|^2 \leq \lambda_{k+1} \sum_{\ell=k+1}^{N} |\alpha_\ell|^2 = \lambda_{k+1},$$

was gerade die Abschätzung "$\geq$" in (12.14) liefert. □

Als unmittelbare Folgerung aus Theorem 12.11 erhält man:

Korollar 12.12 (Satz von Rayleigh/Ritz). *Unter den Bedingungen von Theorem 12.11 gilt*

$$\lambda_1 = \max_{0 \neq x \in \mathbb{R}^N} \frac{x^\top A x}{x^\top x}, \qquad \lambda_N = \min_{0 \neq x \in \mathbb{R}^N} \frac{x^\top A x}{x^\top x}.$$

[3] siehe auch (12.18) im Nachtrag zu diesem Kapitel

Bemerkung 12.13. Den Ausdruck

$$R(x) \;=\; \frac{x^\top Ax}{x^\top x}, \qquad 0 \neq x \;\in\; \mathbb{R}^N,$$

bezeichnet man als *Rayleigh-Quotienten*.

12.5 Störungsresultate für Eigenwerte symmetrischer Matrizen

Ein Störungsresultat für die Eigenwerte symmetrischer Matrizen ist bereits in Korollar 12.2 vorgestellt werden. Für den Spezialfall symmetrischer Störungen liefert das folgende Theorem eine Verschärfung des genannten Resultats.

Theorem 12.14. *Seien $A,\ \Delta A \in \mathbb{R}^{N\times N}$ symmetrische Matrizen, und für $B \in \{A,\ \Delta A,\ A+\Delta A\}$ bezeichne $\lambda_1(B) \geq \lambda_2(B) \geq \ldots \geq \lambda_N(B)$ die monoton fallend angeordneten Eigenwerte der Matrix B. Dann gilt*

$$\lambda_k(A) \;+\; \lambda_N(\Delta A) \;\leq\; \lambda_k(A+\Delta A) \;\leq\; \lambda_k(A) \;+\; \lambda_1(\Delta A), \qquad k \;=\; 1,2,\ldots,N,$$

und damit insbesondere

$$|\lambda_k(A+\Delta A) \;-\; \lambda_k(A)| \;\leq\; \|\Delta A\|_2, \qquad k \;=\; 1,2,\ldots,N. \tag{12.17}$$

BEWEIS. Theorem 12.11 und Korollar 12.12 ergeben für $k = 0,1,\ldots,N-1$

$$\begin{aligned}
\lambda_{k+1}(A+\Delta A) &= \min_{\substack{\mathcal{N}\subset\mathbb{R}^N \text{ linear}\\ \dim\mathcal{N}\leq k}} \max_{0\neq x\in\mathcal{N}^\perp} \Big\{\frac{x^\top Ax}{x^\top x} \;+\; \frac{x^\top \Delta A x}{x^\top x}\Big\} \\
&\leq \text{———} \text{«} \text{———} \Big\{\frac{x^\top Ax}{x^\top x}\Big\} + \lambda_1(\Delta A) \;=\; \lambda_{k+1}(A) + \lambda_1(\Delta A), \\
\lambda_{N-k}(A+\Delta A) &= \max_{\substack{\mathcal{N}\subset\mathbb{R}^N \text{ linear}\\ \dim\mathcal{N}\leq k}} \min_{0\neq x\in\mathcal{N}^\perp} \Big\{\frac{x^\top Ax}{x^\top x} \;+\; \frac{x^\top \Delta A x}{x^\top x}\Big\} \\
&\geq \text{———} \text{«} \text{———} \Big\{\frac{x^\top Ax}{x^\top x}\Big\} + \lambda_N(\Delta A) \;=\; \lambda_{N-k}(A) + \lambda_N(\Delta A).
\end{aligned}$$

Die Abschätzung (12.17) folgt nun unmittelbar aus der Identität $r_\sigma(\Delta A) = \|\Delta A\|_2$, siehe (4.35) auf Seite 83. □

12.6 Nachtrag: Faktorisierungen von Matrizen

Im Folgenden werden einige aus der linearen Algebra bekannte Matrix-Faktorisierungen in Erinnerung gerufen. Detaillierte Erläuterungen hierzu findet man zum Beispiel in Fischer [28] oder im Fall der Schur-Faktorisierung in Bunse/Bunse-Gerstner [11] oder Opfer [79].

12.6.1 Symmetrische Matrizen

Eine Matrix $A \in \mathbb{R}^{N\times N}$ heißt *symmetrisch*, falls $A = A^\top$ gilt. Es existiert dann eine Orthonormalbasis $u_1, \ldots, u_N \in \mathbb{R}^N$ bestehend aus Eigenvektoren von A. Bezeichnet man die zugehörigen Eigenwerte mit $\lambda_1, \ldots, \lambda_N \in \mathbb{R}$, so liegt folgende Situation vor:

$$\left.\begin{aligned} Au_k &= \lambda_k u_k, \\ u_k^\top u_\ell &= \delta_{k\ell}, \qquad k, \ell = 1, 2, \ldots, N. \end{aligned}\right\} \tag{12.18}$$

Theorem 12.15. *Die Matrix $A \in \mathbb{R}^{N\times N}$ sei symmetrisch mit Zerlegung (12.18). Dann gilt*

$$A = UDU^\top \quad \textit{mit } D := \operatorname{diag}(\lambda_1, \ldots, \lambda_N) \in \mathbb{R}^{N\times N}, \quad U = \left(\begin{array}{c|c|c} u_1 & \ldots & u_N \end{array}\right) \in \mathbb{R}^{N\times N}.$$

BEWEIS. Jeder Vektor $x \in \mathbb{R}^N$ besitzt die Darstellung

$$x = \sum_{\ell=1}^{N} \alpha_\ell u_\ell$$

mit gewissen Koeffizienten $\alpha_1, \ldots, \alpha_N \in \mathbb{R}$, und dann gilt

$$UDU^\top x = \sum_{\ell=1}^{N} \alpha_\ell UDU^\top u_\ell = \sum_{\ell=1}^{N} \alpha_\ell U \underbrace{D\mathbf{e}_\ell}_{\lambda_\ell \mathbf{e}_\ell} = \sum_{\ell=1}^{N} \alpha_\ell \lambda_\ell u_\ell = Ax.$$

□

12.6.2 Diagonalisierbare Matrizen

Die Matrix $A \in \mathbb{R}^{N\times N}$ heißt *diagonalisierbar*, falls eine Faktorisierung der Form

$$T^{-1}AT = \operatorname{diag}(\lambda_1, \ldots, \lambda_N) \in \mathbb{C}^{N\times N}, \tag{12.19}$$

existiert mit einer regulären Matrix $T \in \mathbb{C}^{N\times N}$. Die Zahlen $\lambda_1, \ldots, \lambda_N \in \mathbb{C}$ stellen dann die Eigenwerte der Matrix A dar, und der k-te Spaltenvektor $u_k \in \mathbb{R}^N$ von $T = (u_1|\ldots|u_N) \in \mathbb{C}^{N\times N}$ ist ein Eigenvektor der Matrix A zum Eigenwert λ_k.

12.6.3 Schur-Faktorisierung

Jede Matrix $A \in \mathbb{R}^{N\times N}$ ist ähnlich zu einer Dreiecksmatrix, wobei die Transformationsmatrix $Q \in \mathbb{C}^{N\times N}$ unitär gewählt werden kann, das heißt, $Q^{-1} = Q^{\mathrm{H}}$. Die entsprechende Faktorisierung

$$Q^{-1}AQ = R \qquad \big(Q \in \mathbb{C}^{N\times N} \text{ unitär}, \quad R \in \mathbb{C}^{N\times N} \text{ untere Dreiecksmatrix}\big) \tag{12.20}$$

wird als *Schur-Faktorisierung* bezeichnet.

Weitere Themen und Literaturhinweise

Eine Auswahl existierender Lehrbücher mit Abschnitten über Variationsformulierungen sowie Störungsresultate für die Eigenwerte symmetrischer und nichtsymmetrischer Matrizen bildet Deuflhard/Hohmann [22], Golub/Van Loan [35], Hämmerlin/Hoffmann [48], Hanke-Bourgeois [52], Horn/Johnson [58], Kress [63], Mennicken/Wagenführer [71], Oevel [78], Parlett [81], Schaback/Wendland [92], Stoer/Bulirsch [99] und Werner [111]. Variationsformulierungen und Störungsresultate für Singulärwertzerlegungen findet man in [58], [35] und in Baumeister [3].

Übungsaufgaben

Aufgabe 12.1. (a) Gegeben seien die (komplexen) Tridiagonalmatrizen

$$A = \begin{pmatrix} a_1 & b_2 & & 0 \\ c_2 & a_2 & \ddots & \\ & \ddots & \ddots & b_N \\ 0 & & c_N & a_N \end{pmatrix}, \qquad B = \begin{pmatrix} -a_1 & b_2 & & 0 \\ c_2 & -a_2 & \ddots & \\ & \ddots & \ddots & b_N \\ 0 & & c_N & -a_N \end{pmatrix}.$$

Man zeige: Die komplexe Zahl λ ist ein Eigenwert der Matrix A genau dann, wenn $-\lambda$ ein Eigenwert der Matrix B ist.

(b) Für die reelle symmetrische Tridiagonalmatrix

$$A = \begin{pmatrix} a_1 & b_2 & & 0 \\ b_2 & a_2 & \ddots & \\ & \ddots & \ddots & b_N \\ 0 & & b_N & a_N \end{pmatrix} \in \mathbb{R}^{N\times N}$$

sei

$$a_k = -a_{N+1-k} \quad \text{für } k = 1,2,\ldots,N, \qquad b_k = b_{N+2-k} \quad \text{für } k = 2,3,\ldots,N,$$

erfüllt. Man weise Folgendes nach: eine Zahl $\lambda \in \mathbb{C}$ ist Eigenwert der Matrix A genau dann, wenn $-\lambda$ ein Eigenwert von A ist.

(c) Man zeige, dass die Eigenwerte der Tridiagonalmatrix

$$A = \begin{pmatrix} 0 & \overline{b}_2 & & 0 \\ b_2 & 0 & \ddots & \\ & \ddots & \ddots & \overline{b}_N \\ 0 & & b_N & 0 \end{pmatrix} \in \mathbb{C}^{N\times N}$$

symmetrisch zur Zahl 0 liegen und Folgendes gilt,

$$\det(A) = \begin{cases} (-1)^{N/2}|b_2 b_4 \ldots b_N|^2, & \text{falls } N \text{ gerade,} \\ 0 & \text{sonst.} \end{cases}$$

Aufgabe 12.2. Es sei $A \in \mathbb{R}^{N\times N}$ eine Matrix von der Form

$$A \;=\; (I \,-\, 2vv^{\top})\, D\, (I - 2vv^{\top}) \quad \text{mit} \quad D \;=\; \operatorname{diag}(\lambda_1, \ldots, \lambda_N) \;\in\; \mathbb{R}^{N\times N},$$
$$v \;\in\; \mathbb{R}^N, \qquad v^{\top}v \;=\; 1.$$

Man zeige:

(a) Die Matrix A ist symmetrisch, und für $k \;=\; 1, 2, \ldots, N$ ist die Zahl λ_k ein Eigenwert von A mit der k-ten Spalte aus der Matrix $I - 2vv^T$ als zugehörigem Eigenvektor.
(b) Ist speziell $v = (1, 1, \ldots, 1)^{\top}/\sqrt{N}$, so erhält man mit der Notation $A = (a_{jk})$ Folgendes,

$$a_{jk} \;=\; \frac{1}{N}(N\lambda_k\delta_{jk} \,-\, 2\lambda_j \,-\, 2\lambda_k \,+\, 2r), \quad \text{mit} \quad r \;=\; \frac{2}{N}\sum_{s=1}^{N}\lambda_s.$$

Aufgabe 12.3. Für eine symmetrische Matrix $A \in \mathbb{R}^{N\times N}$ und einen Vektor $x = (x_k) \in \mathbb{R}^N$ mit $x_k \neq 0$ für $k \;=\; 1, 2, \ldots, N$ bezeichne

$$d_k \;:=\; \frac{(Ax)_k}{x_k} \qquad \text{für} \quad k \;=\; 1, 2, \ldots, N.$$

Man zeige: für jede Zahl $\mu \in \mathbb{R}$ enthält das Intervall $[\mu-\varrho, \mu+\varrho]$ mit $\varrho := \max_{1\le k\le N} |d_k - \mu|$ mindestens einen Eigenwert λ der Matrix A.

Aufgabe 12.4. Zu gegebener Jordanmatrix

$$A \;:=\; \begin{pmatrix} \lambda & 1 & & 0 \\ & \lambda & \ddots & \\ & & \ddots & 1 \\ 0 & & & \lambda \end{pmatrix} \;\in\; \mathbb{C}^{N\times N}$$

und einer Störungsmatrix $B \in \mathbb{C}^{N\times N}$ bezeichne $\lambda_k(\theta),\ k \;=\; 1, 2, \ldots, N$, die Eigenwerte der fehlerbehafteten Matrix $A + \theta B$, mit $\theta \in \mathbb{C}$. Man weise mit dem Satz von Gerschgorin (der auch für komplexe Matrizen richtig ist) Folgendes nach:

(a) $$\max_{1\le k\le N} |\lambda_k(\theta) \,-\, \lambda| \;\le\; (\|B\|_{\infty} \,+\, 1)|\theta|^{1/N} \qquad \text{für} \quad |\theta| \le 1.$$

(b) Die Abschätzung in (a) ist in Bezug auf den Exponenten $1/N$ von $|\theta|$ nicht zu verbessern.

Aufgabe 12.5. Sei $A = (a_{jk}) \in \mathbb{R}^{N\times N}$ eine irreduzible Matrix, und $\mathcal{G} = \bigcup_{j=1}^{N} \mathcal{G}_j$ bezeichne die Vereinigung der Gerschgorin-Kreise. Man zeige: für jeden Eigenwert λ der Matrix A mit $\lambda \in \partial\mathcal{G}$ gilt auch $\lambda \in \partial\mathcal{G}_j$ für $j = 1, 2, \ldots, N$, und alle Komponenten eines zu λ gehörenden Eigenvektors sind betragsmäßig gleich groß.

Aufgabe 12.6. Man zeige Folgendes: Für eine symmetrische Matrix $A \in \mathbb{R}^{N\times N}$ enthält jedes Intervall der Form $[\mu - \|Ax - \mu x\|_2, \mu + \|Ax - \mu x\|_2]$ mit einer Zahl $\mu \in \mathbb{R}$ und einem Vektor $x \in \mathbb{R}^N$ mit $\|x\|_2 = 1$ mindestens einen Eigenwert der Matrix A.

Aufgabe 12.7. Für eine symmetrische Matrix $A \in \mathbb{R}^{N\times N}$ mit den Eigenwerten $\lambda_1 \ge \lambda_2 \ge \ldots \ge \lambda_N$ weise man Folgendes nach:

$$\lambda_k \;=\; \max_{\substack{M\subset\mathbb{R}^N \text{ linear} \\ \dim M = k}} \; \min_{0\neq x\in M} \frac{x^{\top}Ax}{x^{\top}x}, \qquad k \;=\; 1, 2, \ldots, N,$$

$$\lambda_{N-k+1} \;=\; \min_{\substack{M\subset\mathbb{R}^N \text{ linear} \\ \dim M = k}} \; \max_{0\neq x\in M} \frac{x^{\top}Ax}{x^{\top}x}, \qquad \text{——— « ———}.$$

Aufgabe 12.8. Seien $A,\ \Delta A \in \mathbb{R}^{N\times N}$ symmetrische Matrizen, und für $B \in \{A,\ \Delta A,\ A+\Delta A\}$ bezeichne $\lambda_1(B) \geq \lambda_2(B) \geq \ldots \geq \lambda_N(B)$ die angeordneten Eigenwerte der Matrix B.

(a) Durch Angabe einer geeigneten Matrix ΔA zeige man, dass die Abschätzungen[4]

$$\lambda_k(A) + \lambda_N(\Delta A) \leq \lambda_k(A + \Delta A) \leq \lambda_k(A) + \lambda_1(\Delta A) \qquad \text{für } k = 1, 2, \ldots, N,$$

nicht zu verbessern sind.

(b) Falls die Matrix ΔA positiv definit ist, so gilt

$$\lambda_k(A) \leq \lambda_k(A + \Delta A) \qquad \text{für } k = 1, 2, \ldots, N.$$

Aufgabe 12.9. Es besitze eine symmetrische Matrix $A \in \mathbb{R}^{N\times N}$ mit monoton fallend angeordneten Eigenwerten $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_N$ eine *rechte* untere Dreiecksform,

$$A = \begin{pmatrix} 0 & \cdots & 0 & a_{1N} \\ \vdots & \iddots & \iddots & \vdots \\ 0 & \iddots & & \vdots \\ a_{N1} & a_{N2} & \cdots & a_{NN} \end{pmatrix}, \qquad \text{mit } a_{jk} = a_{kj} \qquad \text{für alle } j, k.$$

Man zeige: es gilt $\lambda_k \geq 0$ für alle Indizes $k \leq \lfloor N/2 \rfloor$, und außerdem gilt $\lambda_k \leq 0$ für alle Indizes $k \geq \lceil N/2 \rceil + 1$. Hierbei bezeichnet $\lfloor x \rfloor$ die größte ganze Zahl $\leq x$, und $\lceil x \rceil$ ist die kleinste ganze Zahl $\geq x$.

[4] siehe Theorem 12.14

13 Numerische Verfahren für Eigenwertprobleme

13.1 Einführende Bemerkungen

Im Folgenden werden verschiedene numerische Verfahren zur approximativen Bestimmung von Eigenwerten quadratischer Matrizen vorgestellt. Dabei basiert eine Klasse von Algorithmen auf der Anwendung von Ähnlichkeitstransformationen, eine zweite auf Vektoriterationen.

13.1.1 Ähnlichkeitstransformationen

In dem vorliegenden Abschnitt werden Verfahren vorgestellt, von denen jedes auf der Hintereinanderausführung von Ähnlichkeitstransformationen beruht,

$$\left.\begin{aligned} A = A^{(1)} &\to A^{(2)} \to A^{(3)} \to \dots \\ A^{(m+1)} &= S_m^{-1} A^{(m)} S_m, \qquad m = 1, 2, \dots, \quad \text{mit} \quad S_m \in \mathbb{R}^{N\times N} \text{ regulär} \end{aligned}\right\} \tag{13.1}$$

mit der Zielsetzung, für hinreichend große Werte von m auf effiziente Weise gute Approximationen für die Eigenwerte von $A^{(m)}$ zu gewinnen.[1]

Im weiteren Verlauf werden die folgenden speziellen Verfahren von der Form (13.1) behandelt.

- Mittels $N-2$ *Householder-Ähnlichkeitstransformationen* (siehe Abschnitt 13.2) lässt sich eine obere Hessenbergmatrix $A^{(N-1)}$ erzeugen, wobei obere beziehungsweise untere Hessenbergmatrizen allgemein folgende Gestalt besitzen,

$$\begin{pmatrix} \times & \dots & \dots & \dots & \times \\ \times & \times & & & \times \\ 0 & \times & \ddots & & \vdots \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ 0 & \dots & 0 & \times & \times \end{pmatrix} \quad \text{bzw.} \quad \begin{pmatrix} \times & \times & 0 & \dots & 0 \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ \vdots & & \ddots & \ddots & 0 \\ \vdots & & & \ddots & \times \\ \times & \dots & \dots & \dots & \times \end{pmatrix} \in \mathbb{R}^{N\times N}.$$

  Eine Matrix $B = (b_{jk})$ ist demnach genau dann eine *obere Hessenbergmatrix*, falls $b_{jk} = 0$ gilt für $j \geq k+2$. Entsprechend ist $B = (b_{jk})$ genau dann eine *untere Hessenbergmatrix*, falls $b_{jk} = 0$ für $j \leq k-2$ gilt.

  Die Hessenbergstruktur ist insofern von Vorteil, als sich hier mit dem Newton-Verfahren beziehungsweise auch mit dem QR-Verfahren effizient die Nullstellen des zugehörigen charakteristischen Polynoms bestimmen lassen (siehe Abschnitt 13.3 beziehungsweise Abschnitt 13.5).

[1] Diese Eigenwerte stimmen aufgrund der durchgeführten Ähnlichkeitstransformationen mit denen der Matrix $A = A^{(1)}$ überein.

- Mit *Givensrotationen* (siehe Abschnitt 13.4 für Einzelheiten) lassen sich Matrizen $A^{(m)}$ erzeugen, deren Nichtdiagonaleinträge für wachsendes m in einem zu spezifizierenden Sinn betragsmäßig immer kleiner werden, so dass dann die Diagonaleinträge von $A^{(m)}$ gute Approximationen an die Eigenwerte von A darstellen.
- *QR-Verfahren* (siehe Abschnitt 13.5) liefern Matrizen $A^{(m)}$, deren Einträge im unteren Dreieck für hinreichend große Werte von m betragsmäßig klein ausfallen, und dann approximieren die Diagonaleinträge von $A^{(m)}$ die Eigenwerte der Matrix A, wie sich herausstellen wird.

Mit der folgenden Bemerkung wird deutlich, warum man aus Stabilitätsgründen in (13.1) sinnvollerweise orthogonale Matrizen S_m wählt.

Bemerkung 13.1. Im Folgenden sei die Matrix $A \in \mathbb{R}^{N\times N}$ als diagonalisierbar angenommen, $T^{-1}AT = D$ mit der regulären Matrix $T \in \mathbb{R}^{N\times N}$ und der Diagonalmatrix $D \in \mathbb{R}^{N\times N}$. Bekanntermaßen[2] bildet dann bezüglich einer gegebenen Vektornorm $\|\cdot\|_p$ die Zahl $\mathrm{cond}_p(T)$ eine Fehlerkonstante für den Fehler in den Eigenwerten von A gegenüber kleinen Störungen in der Matrix A,

$$\max_{\mu\in\sigma(A+\Delta A)} \min_{\lambda\in\sigma(A)} |\mu-\lambda| \;\le\; \mathrm{cond}_p(T)\|\Delta A\|_p.$$

Dementsprechend bildet also nach dem $(m-1)$-ten Schritt des Verfahrens (13.1) aufgrund von

$$\widehat{T}_m^{-1} A^{(m)} \widehat{T}_m \;=\; D \qquad \text{mit} \quad \widehat{T}_m \;:=\; S_{1\ldots m}^{-1} T, \qquad S_{1\ldots m} \;:=\; S_m \cdots S_1,$$

die Konditionszahl $\mathrm{cond}_p(\widehat{T}_m)$ eine Fehlerkonstante für den Fehler der Eigenwerte $\lambda \in \sigma(A^{(m)}) = \sigma(A)$ gegenüber kleinen Störungen in der Matrix $A^{(m)}$. Wegen der Ungleichung $\mathrm{cond}_p(\widehat{T}_m) \le \mathrm{cond}_p(S_{1\ldots m})\,\mathrm{cond}_p(T)$ ist demnach bezüglich der Norm $\|\cdot\| = \|\cdot\|_2$ die Verwendung orthogonaler Transformationen empfehlenswert:

$$S_k^{-1} \;=\; S_k^{\top} \quad \forall\, k \qquad \Longrightarrow \qquad \mathrm{cond}_2(\widehat{T}_m) \;=\; \mathrm{cond}_2(T). \qquad \triangle$$

Für die einzelnen Verfahren gibt es noch weitere Gründe, die Transformationsmatrizen S_m orthogonal zu wählen. Details hierzu werden später vorgestellt.

13.1.2 Vektoriteration

Bei der zweiten Klasse numerischer Verfahren zur Bestimmung der Eigenwerte von Matrizen handelt es sich um Vektoriterationen, die allgemein von der folgenden Form sind,

$$z^{(m+1)} \;=\; C z^{(m)}, \qquad m \;=\; 1, 2, \ldots \qquad (z^{(0)} \in \mathbb{R}^N, \quad C \in \mathbb{R}^{N\times N} \text{ geeignet}),$$

mit der Zielsetzung, aus den Vektoren $z^{(m)} \in \mathbb{R}^N$ Informationen über einzelne Eigenwerte oder auch nur den Spektralradius $r_\sigma(A)$ einer gegebenen Matrix $A \in \mathbb{R}^{N\times N}$ zu gewinnen. Details hierzu werden in Abschnitt 13.7 vorgestellt.

[2] siehe Theorem 12.1

13.2 Transformation auf Hessenbergform

Es sollen zunächst Transformationen der Form $A^{(m+1)} = S_m^{-1} A^{(m)} S_m,\ m = 1, 2, \ldots, N-2$, vorgestellt werden, mit denen sukzessive Matrizen von der Form

$$A^{(m)} = \left(\begin{array}{ccccccc} \times & \ldots & \ldots & \ldots & \ldots & \ldots & \times \\ \times & \ddots & & & & & \vdots \\ 0 & \ddots & \ddots & & & & \vdots \\ \vdots & \ddots & \times & \times & \times & \ldots & \times \\ \vdots & & 0 & \times & \times & \ldots & \times \\ \vdots & & \vdots & \vdots & \vdots & & \vdots \\ 0 & \ldots & 0 & \times & \times & \ldots & \times \end{array}\right) \begin{array}{l} \left.\begin{array}{c}\\\\\\\\\end{array}\right\} m \\ \left.\begin{array}{c}\\\\\end{array}\right\} N-m \end{array} \quad \underbrace{\qquad\qquad}_{m} \quad \underbrace{\qquad\qquad}_{N-m}$$

$$= \left(\begin{array}{c|c|c} & & \\ A_1^{(m)} & & A_2^{(m)} \\ & & \\ \hline & & \\ \mathbf{0} & a^{(m)} & A_3^{(m)} \\ & & \end{array}\right) \tag{13.2}$$

erzeugt werden mit der Hessenbergmatrix $A_1^{(m)} \in \mathbb{R}^{m \times m}$ und den im Allgemeinen vollbesetzten Matrizen $A_2^{(m)} \in \mathbb{R}^{m \times (N-m)}$ und $A_3^{(m)} \in \mathbb{R}^{(N-m) \times (N-m)}$, sowie mit einem gewissen Vektor $a^{(m)} \in \mathbb{R}^{N-m}$. Die Matrix $A^{(N-1)}$ schließlich besitzt Hessenberggestalt.

Das Vorgehen ist hier, in dem Schritt $A^{(m)} \to A^{(m+1)} = S_m^{-1} A^{(m)} S_m$ mit einer Householdertransformation (Abschnitt 13.2.1) den Vektor $a^{(m)}$ aus (13.2) in ein Vielfaches des Einheitsvektors $(1, 0, \ldots, 0)^\top \in \mathbb{R}^{N-m}$ zu transformieren und dabei das aus Nulleinträgen bestehende Trapez in der Matrix $A^{(m)}$ zu erhalten.

Die Transformationsmatrizen $S_1, \ldots, S_{N-1}$ sind hier orthogonal, was aus Stabilitätsgründen von Vorteil ist[3]. Ein weiterer Vorteil besteht darin, dass für symmetrische Matrizen $A \in \mathbb{R}^{N \times N}$ die Matrix $A^{(N-1)} \in \mathbb{R}^{N \times N}$ ebenfalls symmetrisch und somit notwendigerweise (als Hessenbergmatrix) tridiagonal ist, das heißt, $A^{(N-1)}$ ist dünn besetzt, was beispielsweise für die Anwendung des Newton-Verfahrens zur Bestimmung der Nullstellen des charakteristischen Polynoms der Matrix $A^{(N-1)}$ von praktischem Vorteil ist.

13.2.1 Householder-Ähnlichkeitstransformationen zur Gewinnung von Hessenbergmatrizen

Eine Möglichkeit zur Transformation auf Hessenbergform über ein Schema der Form $A^{(m+1)} = S_m^{-1} A^{(m)} S_m,\ m = 1, 2, \ldots, N-2$, besteht in der Anwendung von Householder-Transformationen,

$$S_m = \left(\begin{array}{c|c} I_m & \mathbf{0} \\ \hline \mathbf{0} & \mathcal{H}_m \end{array}\right), \qquad \left.\begin{array}{l} \mathcal{H}_m = I_{N-m} - 2 w_m w_m^\top \in \mathbb{R}^{(N-m)\times(N-m)}, \\ w_m \in \mathbb{R}^{N-m}, \qquad w_m^\top w_m = 1, \end{array}\right\} \tag{13.3}$$

[3] siehe hierzu Bemerkung 13.1

wobei $I_s \in \mathbb{R}^{s\times s}$ mit $s \geq 1$ die Einheitsmatrix bezeichnet und der Vektor $w_m \in \mathbb{R}^{N-m}$ so gewählt wird, dass[4]

$$\mathcal{H}_m a \;=\; \sigma_m \mathbf{e}_m \tag{13.4}$$

gilt mit einem Koeffizienten $\sigma_m \in \mathbb{R}$. Nach Lemma 4.60 auf Seite 91 ist die Matrix S_m orthogonal und symmetrisch, und mit (13.2)–(13.4) erhält man hier Matrizen $A^{(m)}$ der Gestalt (13.2) beziehungsweise

$$A^{(m+1)} \;=\; S_m A^{(m)} S_m \;=\; \left(\begin{array}{c|c|c} & & \\ A_1^{(m)} & & A_2^{(m)}\mathcal{H}_m \\ & & \\ \hline & & \\ \mathbf{0} & \sigma_m \mathbf{e}_m & \mathcal{H}_m A_3^{(m)} \mathcal{H}_m \\ & & \end{array}\right). \tag{13.5}$$

Von Interesse ist der bei dieser Vorgehensweise anfallende Gesamtaufwand zur Berechnung der Matrix $A^{(N-1)}$:

Theorem 13.2. *Die Transformation auf obere Hessenberggestalt mittels Householder-Ähnlichkeitstransformationen von der Form (13.5) lässt sich mit*

$$\tfrac{10N^3}{3}\big(1 \;+\; \mathcal{O}\big(\tfrac{1}{N}\big)\big)$$

arithmetischen Operationen realisieren.

BEWEIS. Zu einem gegebenen Vektor $w_m \in \mathbb{R}^{N-m}$ lässt sich jede Matrix-Vektor-Multiplikation von der Form $H_m x = (I - 2w_m w_m^\top)x \;=\; x - 2(w_m^\top x)w_m$ mit $x \in \mathbb{R}^{N-m}$ in $2(N-m)$ Additionen und ebenso vielen Multiplikationen realisieren, insgesamt also in $4(N-m)$ arithmetischen Operationen. Der gleiche Aufwand ist für jede Multiplikation $x^\top H_m = (H_m x)^\top$ erforderlich. Dem Schema (13.5) entnimmt man, dass bei dem Schritt $A^{(m)} \to A^{(m+1)}$ insgesamt $2(N-m) + m = N - m + N$ solcher Matrix-Vektor-Multiplikationen erforderlich sind und dafür demnach $4(N-m)^2 + 4(N-m)N$ arithmetische Operationen anfallen. Bei Durchführung des gesamten Schemas von $A = A^{(1)}$ bis hin zur Berechnung von $A^{(N-1)}$ summiert sich dies zu

$$4\sum_{m=1}^{N-2}\big((N-m)^2 \;+\; N(N-m)\big) \;=\; 4\underbrace{\sum_{m=2}^{N-1} m^2}_{\frac{(N-1)N(2N-1)}{6} - 1} \;+\; 4N\underbrace{\sum_{m=2}^{N-1} m}_{\frac{(N-1)N}{2} - 1} \;=\; \tfrac{10N^3}{3}\big(1 + \mathcal{O}\big(\tfrac{1}{N}\big)\big)$$

arithmetischen Operationen. Die Berechnung der Vektoren $w_1, \ldots, w_{N-2}$ erfordert nochmals die dagegen nicht weiter ins Gewicht fallenden $\mathcal{O}(N^2)$ Additionen und ebenso viele Multiplikationen sowie $\mathcal{O}(N)$ Divisionen und genauso viele Quadratwurzeln. □

[4] Die genaue Form des Vektors $w_m \in \mathbb{R}^{N-m}$ ist in Lemma 4.62 auf Seite 92 angegeben.

13.2.2 Der symmetrische Fall

Falls die Matrix $A \in \mathbb{R}^{N\times N}$ symmetrisch ist, so erhält man aufgrund der Orthogonalität der Transformationsmatrizen für $A^{(m)}$ die Form

$$A^{(m)} = \left[\begin{array}{ccccc|ccc} \times & \times & 0 & \dots & 0 & \dots & 0 \\ \times & \ddots & \ddots & \ddots & & & \vdots \\ 0 & \ddots & \ddots & \times & 0 & \dots & 0 \\ \vdots & \ddots & \times & \times & \times & \dots & \times \\ \vdots & & 0 & \times & \times & \dots & \times \\ \vdots & & \vdots & \vdots & \vdots & & \vdots \\ 0 & \dots & 0 & \times & \times & \dots & \times \end{array}\right] \begin{matrix} \left.\begin{matrix}\\\\\\\end{matrix}\right\} m \\ \left.\begin{matrix}\\\\\\\end{matrix}\right\} N-m \end{matrix} = \left(\begin{array}{ccc|c} & A_1^{(m)} & & \mathbf{0} \\ & & & a^{(m)\top} \\ \hline \mathbf{0} & a^{(m)} & & A_3^{(m)} \end{array}\right) \tag{13.6}$$

$$\underbrace{\qquad\qquad}_{m} \quad \underbrace{\qquad\qquad}_{N-m}$$

mit der Tridiagonalmatrix $A_1^{(m)} \in \mathbb{R}^{m\times m}$ und der im Allgemeinen vollbesetzten Matrix $A_3^{(m)} \in \mathbb{R}^{(N-m)\times(N-m)}$, sowie mit einem gewissen Vektor $a^{(m)} \in \mathbb{R}^{N-m}$. Die Matrix $A^{(N-1)}$ schließlich besitzt Tridiagonalgestalt. Die entsprechende Householder-Ähnlichkeitstransformation liefert eine Matrix $A^{(m+1)}$ mit der folgenden Struktur,

$$A^{(m+1)} = S_m A^{(m)} S_m = \left(\begin{array}{cc|c} & & \mathbf{0} \\ & A_1^{(m)} & \\ & & \sigma_m \mathbf{e}_m^\top \\ \hline \mathbf{0} & \sigma_m \mathbf{e}_m & \mathcal{H}_m A_3^{(m)} \mathcal{H}_m \end{array}\right). \tag{13.7}$$

Für zugrunde liegende symmetrische Matrizen A ist der bei dieser Vorgehensweise anfallende Gesamtaufwand zur Berechnung von $A^{(N-1)}$ etwas geringer als für nichtsymmetrische Matrizen aus $\mathbb{R}^{N\times N}$:

Theorem 13.3. *Bei symmetrischen Matrizen $A \in \mathbb{R}^{N\times N}$ lässt sich durch Householder-Ähnlichkeitstransformationen eine Tridiagonalmatrix gewinnen mit einem Aufwand von*

$$\frac{8N^3}{3}\left(1 + \mathcal{O}\left(\tfrac{1}{N}\right)\right)$$

arithmetischen Operationen.

BEWEIS. Es sind die gleichen Überlegungen wie beim Beweis von Theorem 13.2 anzustellen, so dass hier nur die wenigen Modifikationen herausgestellt werden. So entnimmt man dem Schema (13.7), dass bei dem Schritt $A^{(m)} \to A^{(m+1)}$ insgesamt

$2(N-m)$ Matrix-Vektor-Multiplikationen mit Householdermatrizen $\in \mathbb{R}^{(N-m)\times(N-m)}$ erforderlich sind und dafür demnach $8(N-m)^2$ arithmetische Operationen anfallen. Bei Durchführung des gesamten Schemas von $A = A^{(1)}$ bis hin zur Berechnung von $A^{(N-1)}$ summiert sich dies zu

$$8\sum_{m=1}^{N-2}(N-m)^2 \;=\; 8\sum_{m=2}^{N-1}m^2 \;=\; 8\big(\tfrac{(N-1)N(2N-1)}{6}-1\big) \;=\; \tfrac{8N^3}{3}\big(1+\mathcal{O}\big(\tfrac{1}{N}\big)\big)$$

arithmetischen Operationen. Die Berechnung der Vektoren $w_1,\ldots,w_{N-2}$ erfordert nochmals die vergleichsweise nicht weiter ins Gewicht fallenden $\mathcal{O}(N^2)$ arithmetischen Operationen. □

13.3 Newton-Verfahren zur Berechnung der Eigenwerte von Hessenbergmatrizen

Im vorangegangenen Abschnitt 13.2 sind Methoden vorgestellt worden, mit denen man zu einer gegebenen Matrix $A \in \mathbb{R}^{N\times N}$ eine obere Hessenbergmatrix $B \in \mathbb{R}^{N\times N}$ gewinnt, deren Eigenwerte mit denen von A übereinstimmen, $\sigma(B) = \sigma(A)$. In dem vorliegenden Abschnitt wird geschildert, wie sich die Eigenwerte von Hessenbergmatrizen effizient näherungsweise bestimmen lassen.

Hierzu bedient man sich des Newton-Verfahrens $\mu_{m+1} = \mu_m - p(\mu_m)/p'(\mu_m)$, $m = 0, 1, \ldots$, zur iterativen Bestimmung der Nullstellen des zugehörigen charakteristischen Polynoms[5] $p(\mu) = \det(B-\mu I)$, dessen Nullstellen mit den Eigenwerten der Matrix $B \in \mathbb{R}^{N\times N}$ übereinstimmen. Bei vollbesetzten Matrizen ist diese Vorgehensweise mit $cN^3 + \mathcal{O}(N^2)$ arithmetischen Operationen pro Iterationsschritt (mit einer gewissen Konstanten $c > 0$) recht aufwändig. Bei Hessenbergmatrizen B jedoch lässt sich für jedes μ der Aufwand zur Berechnung der Werte $p(\mu)$ und $p'(\mu)$ auf jeweils $\mathcal{O}(N^2)$ arithmetische Operationen reduzieren, wie sich im Folgenden herausstellen wird.

13.3.1 Der nichtsymmetrische Fall. Die Methode von Hyman

Das charakteristische Polynom $p(\mu)$ einer Hessenbergmatrix und die zugehörige Ableitung $p'(\mu)$ lassen sich jeweils über die Auflösung spezieller gestaffelter linearer Gleichungssysteme berechnen, wie sich im Folgenden herausstellen wird.

Theorem 13.4. *Sei $B = (b_{jk}) \in \mathbb{R}^{N\times N}$ eine obere Hessenbergmatrix mit $b_{j,j+1} \neq 0$ für $j = 1, 2, \ldots, N-1$ und charakteristischem Polynom $p(\mu) = \det(B-\mu I)$, $\mu \in \mathbb{R}$. Im Folgenden sei $\mu \in \mathbb{R}$ fest gewählt und kein Eigenwert von B, und es bezeichne $x = x(\mu) = (x_k(\mu)) \in \mathbb{R}^N$ den eindeutig bestimmten Vektor mit*

$$(B-\mu I)x \;=\; \mathbf{e}_1, \tag{13.8}$$

[5] Entsprechende Konvergenzresultate finden Sie in Abschnitt 5.4.3.

mit $\mathbf{e}_1 = (1, 0, \ldots, 0)^\top \in \mathbb{R}^N$. *Dann gelten die folgenden Darstellungen,*

$$p(\mu) \;=\; \frac{(-1)^{N-1} b_{21} b_{32} \cdots b_{N,N-1}}{x_N(\mu)}, \qquad \frac{p(\mu)}{p'(\mu)} \;=\; \frac{1}{x_N(\mu)} \Big/ \frac{d}{d\mu}\Big(\frac{1}{x_N(\mu)}\Big). \tag{13.9}$$

BEWEIS. Anwendung der cramerschen Regel auf die Gleichung (13.8) liefert die erste Aussage in (13.9),

$$\begin{aligned} x_N \;&=\; \det \begin{pmatrix} b_{11}-\mu & b_{12} & \cdots & b_{1,N-1} & 1 \\ b_{21} & b_{22}-\mu & & \vdots & 0 \\ & b_{32} & \ddots & \vdots & \vdots \\ & & \ddots & b_{N-1,N-1}-\mu & \vdots \\ & & & b_{N,N-1} & 0 \end{pmatrix} \Big/ p(\mu) \\ &\overset{(*)}{=}\; (-1)^{N-1} \underbrace{\det \begin{pmatrix} b_{21} & b_{22}-\mu & \cdots & b_{2,N-1} \\ & b_{32} & \ddots & \vdots \\ & & \ddots & b_{N-1,N-1}-\mu \\ & & & b_{N,N-1} \end{pmatrix}}_{= b_{21} b_{32} \cdots b_{N,N-1}} \Big/ p(\mu), \end{aligned}$$

wobei man die Identität $(*)$ durch Entwicklung der auftretenden Determinante nach der letzten Spalte erhält. Dies ergibt die erste Identität in (13.9), und eine anschließende Differenziation liefert die zweite Aussage in (13.9). □

Bemerkung 13.5. In Theorem 13.4 stellt die Bedingung an das Nichtverschwinden der unteren Nebendiagonaleinträge keine ernsthafte Restriktion dar: im Fall $b_{j,j+1} = 0$ für ein $j \in \{1, 2, \ldots, N-1\}$ lässt sich das Problem auf die Bestimmung der Eigenwerte zweier Teilmatrizen von oberer Hessenbergstruktur reduzieren. △

Die für (13.9) erforderliche N-te Komponente der Lösung des Gleichungssystems (13.8) und deren Ableitung erhält man jeweils über die Lösung gestaffelter linearer Gleichungssysteme:

Theorem 13.6. *Mit den Bezeichnungen aus Theorem 13.4 erhält man die beiden Werte* $1/x_N(\mu)$ *und* $\frac{d}{d\mu}(\frac{1}{x_N(\mu)})$ *aus den folgenden (durch Umformung und Differenziati-*

on von (13.8) entstandenen) gestaffelten linearen Gleichungssystemen

$$\left.\begin{array}{rrrrrrrrrrrr}
(b_{11}-\mu)v_1 + & b_{12}v_2 & + & \cdots & + & b_{1,N-1}v_{N-1} & + & b_{1N} & = & \frac{1}{x_N(\mu)} \\
b_{21}v_1 + (b_{22}-\mu)v_2 + & \cdots & + & b_{2,N-1}v_{N-1} & + & b_{2N} & = & 0 \\
\ddots & \ddots & & \vdots & & \vdots & & & & \\
b_{N-1,N-2}v_{N-2} - (b_{N-1,N-1}-\mu)v_{N-1} + & b_{N-1,N} & = & 0 \\
b_{N,N-1}v_{N-1} & + b_{NN} - \mu = & 0
\end{array}\right\} \tag{13.10}$$

beziehungsweise

$$\left.\begin{array}{rrrrrrrrrrrr}
(b_{11}-\mu)z_1 + & b_{12}z_2 & + & \cdots & + & b_{1,N-1}z_{N-1} & - & v_1 & = & \frac{d}{d\mu}\left(\frac{1}{x_N(\mu)}\right) \\
b_{21}z_1 + (b_{22}-\mu)z_2 + & \cdots & + & b_{2,N-1}z_{N-1} & - & v_2 & = & 0 \\
\ddots & \ddots & & \vdots & & \vdots & & & & \\
b_{N-1,N-2}z_{N-2} - (b_{N-1,N-1}-\mu)z_{N-1} - & v_{N-1} & = & 0 \\
b_{N,N-1}z_{N-1} & - 1 & = & 0
\end{array}\right\} \tag{13.11}$$

die man rekursiv nach den Unbekannten $v_{N-1}, v_{N-2}, \ldots, v_1, 1/x_N(\mu)$ beziehungsweise $z_{N-1}, z_{N-2}, \ldots, z_1, \frac{d}{d\mu}(\frac{1}{x_N(\mu)})$ auflöst.

BEWEIS. Die Aussage (13.10) erhält man (für $v_k = x_k(\mu)/x_N(\mu)$), indem die einzelnen Zeilen des Gleichungssystems (13.8) durch $x_N(\mu)$ dividiert werden. Differenziation der Gleichungen in (13.10) nach μ liefert für $z_k = (\frac{dv_k}{d\mu})(\mu)$ unmittelbar (13.11). □

13.3.2 Das Newton-Verfahren zur Berechnung der Eigenwerte tridiagonaler Matrizen

Ist die in Abschnitt 13.3.1 behandelte Matrix $B \in \mathbb{R}^{N\times N}$ symmetrisch, so ist sie notwendigerweise tridiagonal. In diesem Fall lassen sich die Werte $p(\mu) = \det(B - \mu I)$ und $p'(\mu)$ auf einfache Weise rekursiv berechnen:

Lemma 13.7. *Zu gegebenen Zahlen $a_1, \ldots, a_N \in \mathbb{R}$ und $b_2, \ldots, b_N \in \mathbb{R}$ gelten für die charakteristischen Polynome*

$$p_n(\mu) \;=\; \det(J_n - \mu I), \qquad J_n \;=\; \begin{pmatrix} a_1 & b_2 & & \\ b_2 & \ddots & \ddots & \\ & \ddots & \ddots & b_n \\ & & b_n & a_n \end{pmatrix}, \qquad n \;=\; 1, 2, \ldots, N,$$

die folgenden Rekursionsformeln

$$\left.\begin{aligned} p_1(\mu) &= a_1 - \mu, \\ p_n(\mu) &= (a_n - \mu) p_{n-1}(\mu) \;-\; b_n^2 p_{n-2}(\mu), \qquad n \;=\; 2, 3, \ldots, N, \end{aligned}\right\} \tag{13.12}$$

mit der Notation $p_0(\mu) := 1$. Für die Ableitungen gelten die Rekursionsformeln

$$\begin{aligned} p_1'(\mu) &= -1, \\ p_n'(\mu) &= -p_{n-1}(\mu) \;+\; (a_n - \mu) p_{n-1}'(\mu) \;-\; b_n^2 p_{n-2}'(\mu), \qquad n \;=\; 2, 3, \ldots, N. \end{aligned}$$

BEWEIS. Die angegebene Darstellung für p_1 ergibt sich unmittelbar, und weiter gilt

$$p_2(\mu) \;=\; \det \begin{pmatrix} a_1 - \mu & b_2 \\ b_2 & a_2 - \mu \end{pmatrix} \;=\; \underbrace{(a_1 - \mu)}_{=\; p_1(\mu)} (a_2 - \mu) \;-\; b_2^2,$$

was die angegebene Darstellung für p_2 ist. Für $n \geq 3$ erhält man

$$p_n(\mu) \;=\; \det \begin{pmatrix} a_1 - \mu & b_2 & & & \\ b_2 & \ddots & \ddots & & \\ & \ddots & a_{n-2} - \mu & b_{n-1} & \\ & & b_{n-1} & a_{n-1} - \mu & b_n \\ & & & b_n & a_n - \mu \end{pmatrix}$$

$$\overset{(*)}{=} \;(a_n - \mu) p_{n-1}(\mu) \;-\; b_n \underbrace{\det \begin{pmatrix} a_1 - \mu & b_2 & & & \\ b_2 & \ddots & \ddots & & \\ & \ddots & a_{n-3} - \mu & b_{n-2} & \\ & & b_{n-2} & a_{n-2} - \mu & b_{n-1} \\ & & & 0 & b_n \end{pmatrix}}_{\overset{(**)}{=}\; b_n p_{n-2}(\mu)},$$

wobei sich die Identitäten $(*)$ beziehungsweise $(**)$ durch Determinantenentwicklung nach der letzten Spalte beziehungsweise der letzten Zeile ergeben. Dies komplettiert den Beweis der Identität (13.12). Die angegebenen Rekursionsformeln für die Ableitungen der Polynome p_n erhält man unmittelbar durch Differenziation der Terme in (13.12). □

13.4 Jacobi-Verfahren zur Nichtdiagonaleinträge-Reduktion bei symmetrischen Matrizen

In dem folgenden Abschnitt 13.4.1 wird spezifiziert, inwieweit bei quadratischen Matrizen B die Diagonaleinträge Approximationen an die Eigenwerte von B darstellen (für den Fall, dass die Nichtdiagonaleinträge von B betragsmäßig klein ausfallen). Anschließend werden in Abschnitt 13.4.2 zu einer gegebenen symmetrischen Matrix $A \in \mathbb{R}^{N\times N}$ spezielle Verfahren von der Form $A^{(m+1)} = S_m^{-1} A^{(m)} S_m,\ m = 1, 2, \ldots$ behandelt, mit denen man sukzessive solche zu A ähnlichen Matrizen B mit betragsmäßig kleinen Nichtdiagonaleinträgen erzeugt.

13.4.1 Approximation der Eigenwerte durch Diagonaleinträge

Vor der Einführung des Jacobi-Verfahrens und den zugehörigen Konvergenzbetrachtungen sind ein paar Ergänzungen zu den in Kapitel 12 vorgestellten allgemeinen Störungsresultaten für Eigenwerte erforderlich.

Definition 13.8. Für eine symmetrische Matrix $B = (b_{jk}) \in \mathbb{R}^{N\times N}$ ist die Zahl $\mathcal{S}(B) \in \mathbb{R}_+$ folgendermaßen erklärt,

$$\mathcal{S}(B) := \sum_{\substack{j,k=1 \\ j\neq k}}^{N} b_{jk}^2. \tag{13.13}$$

Offensichtlich gilt

$$\mathcal{S}(B) = \|B\|_{\mathrm{F}}^2 - \sum_{k=1}^{N} b_{kk}^2 = \|B - D\|_{\mathrm{F}}^2, \quad \text{mit } D := \operatorname{diag}(b_{11}, \ldots, b_{NN}), \tag{13.14}$$

wobei $\|\cdot\|_{\mathrm{F}}$ die Frobeniusnorm bezeichnet. Der Wert $\mathcal{S}(B)$ wird im Folgenden als Maß dafür verwendet, wie weit die Matrix B von einer Diagonalgestalt entfernt ist. Bei Matrizen mit (gegenüber der Diagonalen) betragsmäßig kleinen Nichtdiagonaleinträgen stellen die Diagonaleinträge Approximationen für die Eigenwerte dar. Genauer gilt Folgendes:

Theorem 13.9. *Seien $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_N$ die Eigenwerte der symmetrischen Matrix $B = (b_{jk}) \in \mathbb{R}^{N\times N}$, und seien $b_{j_1 j_1} \geq b_{j_2 j_2} \geq \ldots \geq b_{j_N j_N}$ die der Größe nach angeordneten Diagonaleinträge von B. Dann gilt*

$$|b_{j_r j_r} - \lambda_r| \leq \sqrt{\mathcal{S}(B)}, \qquad r = 1, 2, \ldots, N.$$

BEWEIS. Mit der Notation $D := \operatorname{diag}(b_{11}, \ldots, b_{NN})$ erhält man

$$\max_{r=1,\ldots,N} |b_{j_r j_r} - \lambda_r| \overset{(*)}{\leq} \|B - D\|_2 \overset{(**)}{\leq} \|B - D\|_{\mathrm{F}} = \sqrt{\mathcal{S}(B)},$$

wobei die Ungleichung $(*)$ aus Theorem 12.14 angewandt mit $A = B,\ \Delta A = D - B$ folgt. Die Abschätzung $(**)$ resultiert aus der allgemeinen Ungleichung $\|\cdot\|_2 \leq \|\cdot\|_{\mathrm{F}}$ (siehe Theorem 4.45), und die letzte Identität ist eine unmittelbare Konsequenz aus den Definitionen für $\|\cdot\|_{\mathrm{F}}$ und $\mathcal{S}(\cdot)$, vergleiche die Darstellung (13.14). □

13.4.2 Givensrotationen zur Reduktion der Nichtdiagonaleinträge

Im Folgenden wird das Verfahren von Jacobi zur approximativen Bestimmung der Eigenwerte symmetrischer Matrizen $A \in \mathbb{R}^{N\times N}$ über die Reduktion der Nichtdiagonaleinträge vorgestellt, $\mathcal{S}(A) > \mathcal{S}(A^{(2)}) > \ldots$. Dieses Verfahren ist von der Form $A^{(m+1)} = S_m^{-1} A^{(m)} S_m,\ m = 1, 2, \ldots$ mit $A^{(1)} = A$, wobei die einzelnen Ähnlichkeitstransformationen von der allgemeinen Form

$$\widehat{B} := \Omega_{pq}^{-1} B\, \Omega_{pq}, \quad \Omega_{pq} = \begin{pmatrix} 1 & & & & & & & & \\ & \ddots & & & & & & & \\ & & 1 & & & & & & \\ & & & c & & & -s & & \\ & & & & 1 & & & & \\ & & & & & \ddots & & & \\ & & & & & & 1 & & \\ & & & s & & & & c & \\ & & & & & & & & 1 \\ & & & & & & & & & \ddots \\ & & & & & & & & & & 1 \end{pmatrix} \begin{matrix} \\ \\ \\ \leftarrow p \\ \\ \\ \\ \leftarrow q \\ \\ \\ \\ \end{matrix} \in \mathbb{R}^{N\times N} \tag{13.15}$$
$$\begin{matrix} & & & \uparrow & & \uparrow & & & \\ & & & p & & q & & & \end{matrix}$$

sind mit einer symmetrischen Matrix $B \in \mathbb{R}^{N\times N}$ und mit speziell zu wählenden Indizes $p \neq q$ und reellen Zahlen

$$c,\ s \in \mathbb{R}, \qquad c^2 + s^2 = 1. \tag{13.16}$$

Im Folgenden soll zunächst ein allgemeiner Zusammenhang zwischen den Zahlen $\mathcal{S}(\widehat{B})$ und $\mathcal{S}(B)$ hergestellt werden. Hierzu beobachtet man, dass wegen der besonderen Struktur der Matrix Ω_{pq} Folgendes gilt,

$$\widehat{B} = B + \begin{pmatrix} 0 & 0 & 0 & & & \\ 0 & 0 & 0 & & & \\ 0 & 0 & 0 & & & \\ & & & & & \\ & & & & & \\ & & & & & \end{pmatrix} \begin{matrix} \\ \\ \leftarrow p \\ \\ \leftarrow q \\ \\ \end{matrix} \in \mathbb{R}^{N\times N},$$
$$\begin{matrix} & & \uparrow & \uparrow & & \\ & & p & q & & \end{matrix}$$

wobei in der Matrix $\widehat{B} = (\widehat{b}_{jk})$ die Einträge mit den Indizes $(p,p),(q,q)$ und (p,q) von besonderer Bedeutung sind:

$$\widehat{b}_{pp} = c^2 b_{pp} + 2csb_{pq} + s^2 b_{qq}, \tag{13.17}$$

$$\widehat{b}_{qq} = s^2 b_{pp} - 2csb_{pq} + c^2 b_{qq}, \tag{13.18}$$

$$\widehat{b}_{pq} = \widehat{b}_{qp} = cs(b_{qq} - b_{pp}) + (c^2 - s^2) b_{pq}, \tag{13.19}$$

$$\widehat{b}_{jk} = b_{jk}, \qquad j,k \notin \{p,q\}, \tag{13.20}$$

wobei $B = (b_{jk})$. Weiter gilt noch

$$\widehat{b}_{jp} = \widehat{b}_{pj} = cb_{jp} + sb_{jq}, \qquad \widehat{b}_{jq} = \widehat{b}_{qj} = -sb_{jp} + cb_{jq} \qquad \text{für } j \notin \{p,q\}.$$

Im folgenden Theorem 13.11 wird ein Zusammenhang zwischen den Zahlen $\mathcal{S}(\widehat{B})$ und $\mathcal{S}(B)$ hergestellt, für dessen Beweis das folgende Resultat über die Invarianz der Frobeniusnorm gegenüber orthogonalen Transformationen benötigt wird.

Lemma 13.10. *Für jede Matrix $B \in \mathbb{R}^{N\times N}$ und jede orthogonale Matrix $Q \in \mathbb{R}^{N\times N}$ gilt die Identität $\| Q^{-1}BQ \|_{\mathrm{F}} = \| B \|_{\mathrm{F}}$.*

BEWEIS. Zunächst sei an die aus der linearen Algebra bekannte *Spur* einer Matrix $A = (a_{jk}) \in \mathbb{R}^{N\times N}$ erinnert, $\mathrm{spur}(A) = \sum_{k=1}^{N} a_{kk}$. Die Aussage folgt nun unmittelbar aus den beiden folgenden Identitäten,

$$\| A \|_{\mathrm{F}}^2 = \mathrm{spur}(A^{\top}A), \qquad \mathrm{spur}(ST) = \mathrm{spur}(TS) \qquad \text{für alle } A, S, T \in \mathbb{R}^{N\times N},$$

deren elementaren Nachweise hier nicht geliefert werden. □

Das folgende Theorem stellt einen Zusammenhang zwischen den Zahlen $\mathcal{S}(\widehat{B})$ und $\mathcal{S}(B)$ her.

Theorem 13.11. *Für eine symmetrische Matrix $B = (b_{jk}) \in \mathbb{R}^{N\times N}$ gilt mit den Bezeichnungen aus (13.15) Folgendes,*

$$\mathcal{S}(\widehat{B}) = \mathcal{S}(B) - 2(b_{pq}^2 - \widehat{b}_{pq}^2).$$

BEWEIS. Eine Anwendung von Lemma 13.10 und den Identitäten (13.14) und (13.20) liefert

$$\begin{aligned} \mathcal{S}(\widehat{B}) &= \| \widehat{B} \|_{\mathrm{F}}^2 - \sum_{k=1}^{N} \widehat{b}_{kk}^2 \\ &= \underbrace{\Big(\| B \|_{\mathrm{F}}^2 - \sum_{k=1}^{N} b_{kk}^2 \Big)}_{\mathcal{S}(B)} + b_{pp}^2 + b_{qq}^2 - \widehat{b}_{pp}^2 - \widehat{b}_{qq}^2. \end{aligned} \tag{13.21}$$

Zur Verarbeitung der letzten vier Summanden in (13.21) verwendet man die Identitäten (13.17)–(13.19) in der folgenden Matrixschreibweise,

$$\underbrace{\begin{pmatrix} \widehat{b}_{pp} & \widehat{b}_{pq} \\ \widehat{b}_{pq} & \widehat{b}_{qq} \end{pmatrix}}_{=:\,\widehat{b}} = \begin{pmatrix} c & s \\ -s & c \end{pmatrix} \underbrace{\begin{pmatrix} b_{pp} & b_{pq} \\ b_{pq} & b_{qq} \end{pmatrix}}_{=:\,b} \begin{pmatrix} c & -s \\ s & c \end{pmatrix}.$$

Die entstehenden Matrizen b und $\widehat{b} \in \mathbb{R}^{2\times 2}$ sind also orthogonal ähnlich zueinander, und daher erhält man unter Anwendung von Lemma 13.10

$$\underbrace{\widehat{b}_{pp}^2 + \widehat{b}_{qq}^2 + 2\widehat{b}_{pq}^2}_{=\,\|\widehat{b}\|_{\mathrm{F}}^2} = \underbrace{b_{pp}^2 + b_{qq}^2 + 2b_{pq}^2}_{=\,\|b\|_{\mathrm{F}}^2}. \tag{13.22}$$

und die Identitäten (13.21)–(13.22) ergeben dann die Aussage des Theorems. □

Mit Lemma 13.11 wird offensichtlich, dass (bei festem Index (p,q)) im Fall $\widehat{b}_{pq} = 0$ die Zahl $\mathcal{S}(\widehat{B})$ die größtmögliche Verringerung gegenüber $\mathcal{S}(B)$ zu verzeichnen hat.

Korollar 13.12. *Wählt man in (13.15) die Zahlen c und s so, dass $\widehat{b}_{pq} = 0$ erfüllt ist, dann gilt*

$$\mathcal{S}(\widehat{B}) = \mathcal{S}(B) - 2b_{pq}^2.$$

Das folgende Theorem stellt eine Wahl der Zahlen c und s vor, mit der man $\widehat{b}_{pq} = 0$ erhält.

Theorem 13.13. *In (13.15) erhält man den Eintrag $\widehat{b}_{pq} = 0$ durch folgende Wahl der Zahlen c und s (o.B.d.A. sei $b_{pq} \neq 0$)*

$$c = \sqrt{\frac{1+C}{2}},\quad s = \operatorname{sgn}(b_{pq})\sqrt{\frac{1-C}{2}} \quad \text{mit } C = \frac{b_{pp}-b_{qq}}{\left((b_{pp}-b_{qq})^2 + 4b_{pq}^2\right)^{1/2}}. \tag{13.23}$$

Beweis. Mit (13.19) folgt

$$\begin{aligned} \widehat{b}_{pq} &= \operatorname{sgn}(b_{pq})\sqrt{\frac{1-C^2}{4}}(b_{qq}-b_{pp}) + Cb_{pq} \\ &\overset{(*)}{=} \frac{\operatorname{sgn}(b_{pq})|b_{pq}|(b_{qq}-b_{pp})}{\left((b_{pp}-b_{qq})^2 + 4b_{pq}^2\right)^{1/2}} + \frac{b_{pp}-b_{qq}}{\left((b_{pp}-b_{qq})^2 + 4b_{pq}^2\right)^{1/2}} b_{pq} = 0, \end{aligned}$$

wobei die Identität $(*)$ sich so ergibt:

$$\begin{aligned} \sqrt{\frac{1-C^2}{4}} &= \frac{1}{2}\Big(\frac{(b_{pp}-b_{qq})^2 + 4b_{pq}^2 - (b_{pp}-b_{qq})^2}{(b_{pp}-b_{qq})^2 + 4b_{pq}^2}\Big)^{1/2} \\ &= \frac{|b_{pq}|}{((b_{pp}-b_{qq})^2 + 4b_{pq}^2)^{1/2}}. \end{aligned}$$

□

Bemerkung 13.14. 1. Offensichtlich gilt in (13.23) $|C| < 1$, so dass dort die Zahl s wohldefiniert ist. Ebenso offensichtlich gilt dann $c^2 + s^2 = 1$, womit die Matrix Ω_{pq} in (13.15) orthogonal ist.
2. Bei einer Wahl von c und s entsprechend (13.23) tritt üblicherweise für gewisse Indizes $(j,k) \notin \{(p,q),\ (q,p)\}$ üblicherweise auch der Fall ein, dass $\widehat{b}_{jk} \neq 0$ gilt, obwohl eventuell $b_{jk} = 0$ erfüllt ist. △

Im Folgenden soll noch die spezielle Wahl des Indexes (p,q) diskutiert werden. Korollar 13.12 legt nahe, (p,q) so zu wählen, dass $|b_{pq}|$ maximal wird. In diesem Fall erhält man die folgende Abschätzung:

Theorem 13.15. *Für Indizes (p,q) mit $p \neq q$ sei*

$$|b_{pq}| \quad \geq \quad |b_{jk}| \qquad \text{für} \quad j,k \ = \ 1,2,\ldots,N, \quad j \neq k, \tag{13.24}$$

erfüllt. Mit den Bezeichnungen aus (13.15) und Einträgen c und s entsprechend Theorem 13.13 gilt dann die Abschätzung

$$\mathcal{S}(\widehat{B}) \ \leq \ (1-\varepsilon_N)\mathcal{S}(B), \qquad \text{mit} \quad \varepsilon_N \ := \ \frac{2}{N(N-1)}.$$

BEWEIS. Wegen (13.24) gilt die Abschätzung

$$\mathcal{S}(B) \ = \ \sum_{\substack{j,k=1 \\ j\neq k}}^{N} b_{jk}^2 \ \leq \ N(N-1)b_{pq}^2,$$

da die Anzahl der Nichtdiagonaleinträge $N(N-1)$ beträgt. Die Aussage folgt nun mit Korollar 13.12. □

13.4.3 Zwei spezielle Jacobi-Verfahren

Im Folgenden werden für das zu Beginn von Abschnitt 13.4 bereits vorgestellte Jacobi-Verfahren zwei unterschiedliche Möglichkeiten der Wahl der Indizes (p_1,q_1), $(p_2,q_2),\ldots$ behandelt.

Das klassische Jacobi-Verfahren

Algorithmus 13.16 (Klassisches Jacobi-Verfahren). Für eine gegebene symmetrische Matrix $A \in \mathbb{R}^{N\times N}$ setze man $A^{(1)} := A$.

```
for m = 1, 2, ...:
  bestimme Indizes p,q mit |a_pq^(m)| ≥ |a_jk^(m)| für j,k = 1,...,N, j ≠ k;
  A^(m+1) := Ω_pq^{-1} A^(m) Ω_pq;
            (* für Ω_pq aus (13.15) mit c und s wie in (13.23) *)
end
```

△

Bemerkung 13.17. 1. Nach Theorem 13.15 konvergiert für die Matrizen $A^{(m)}$ des klassischen Jacobi-Verfahrens die Messgröße $\mathcal{S}(A^{(m)}) \to 0$ linear. Genauer gilt

$$\mathcal{S}(A^{(m)}) \;\le\; (1-\varepsilon_N)^m \mathcal{S}(A) \quad \text{für } m = 1,2,\dots \quad \Big(\varepsilon_N = \frac{2}{N(N-1)}, \quad A = A^{(1)}\Big).$$

Ist eine absolute Genauigkeit $\eta > 0$ vorgegeben, mit der die Eigenwerte der vorgegebenen Matrix A bestimmt werden sollen, so ist gemäß Theorem 13.9 nach

$$m \;\ge\; 2\frac{\log(\sqrt{\mathcal{S}(A)}/\eta)}{-\log(1-\varepsilon_N)} \;\approx\; N^2 \log\big(\sqrt{\mathcal{S}(A)}/\eta\big)$$

Schritten die gewünschte Genauigkeit erreicht, $\sqrt{\mathcal{S}(A^{(m)})} \le \eta$. Für das Erreichen einer vorgegebenen Genauigkeit sind somit cN^2 Iterationsschritte durchzuführen.
2. In jedem Schritt des klassischen Jacobi-Verfahrens fallen etwa $4N$ Multiplikationen und $2N$ Additionen sowie $\mathcal{O}(1)$ Divisionen und Quadratwurzelberechnungen an, insgesamt also $6N(1+\mathcal{O}(1/N))$ arithmetische Operationen. Hinzu kommt in jedem Schritt der weitaus höher ins Gewicht fallende Aufwand zur Bestimmung des betragsmäßig größten Elements, wofür $N(N-1)/2$ Vergleichsoperationen erforderlich sind. △

Das zyklische Jacobi-Verfahren

Mit Bemerkung 13.17 wird klar, dass beim klassischen Jacobi-Verfahren $cN^4+\mathcal{O}(N^3)$ Operationen für das Erreichen einer vorgegebenen Genauigkeit durchzuführen sind (mit einer Konstanten $c > 0$), was die Anwendung dieses Verfahrens nur für kleine Matrizen zulässt. Daher ist die folgende Variante des Jacobi-Verfahrens in Betracht zu ziehen, die auf die Bestimmung des jeweils betragsmäßig größten Eintrags verzichtet:

Algorithmus 13.18 (Zyklisches Jacobi-Verfahren). Für eine gegebene symmetrische Matrix $A \in \mathbb{R}^{N\times N}$ setze man $A^{(1)} := A$.

```
for m = 0,1,...:   B := A^(m);
  for p = 1 : N-1
    for q = p+1 : N   B := Ω_pq^{-1} B Ω_pq; end
        (* für Ω_pq aus (13.15) mit c und s wie in (13.23) *)
  end
  A^(m+1) := B;
end
```

Bemerkung 13.19. 1. Das zyklische Jacobi-Verfahren ist von der allgemeinen Form $A^{(m+1)} = S_m^{-1} A^{(m)} S_m$, $m = 1, 2, \ldots$ mit

$$\begin{aligned} S_m &= \big(\Omega_{12}\Omega_{13}\cdots\Omega_{1N}\big)\big(\Omega_{23}\Omega_{24}\cdots\Omega_{2N}\big)\cdots\big(\Omega_{N-2,N-1}\Omega_{N-2,N}\big)\Omega_{N-1,N} \\ &= \prod_{p=1}^{N-1}\Big(\prod_{q=p+1}^{N}\Omega_{pq}\Big), \end{aligned}$$

wobei die Einträge $c = c(p,q,j)$ und $s = s(p,q,j)$ von Ω_{pq} entsprechend Theorem 13.13 gewählt sind.
2. In einem Schritt $A^{(m)} \to A^{(m+1)}$ des zyklischen Jacobi-Verfahrens werden $N(N-1)/2$ Jacobi-Transformationen (13.15) mit insgesamt $3N^3(1+\mathcal{O}(1/N))$ arithmetischen Operationen durchgeführt. Typischerweise ist nach $m = \mathcal{O}(1)$ Schritten die Zahl $\mathcal{S}(A^{(m)})$ hinreichend klein (man beachte hierzu das nachfolgende Theorem 13.20), so dass man mit einem Gesamtaufwand von $\mathcal{O}(N^3)$ arithmetischen Operationen auskommt. △

Das zyklische Jacobi-Verfahren konvergiert im Falle einfacher Eigenwerte quadratisch im Sinne des folgenden Theorems. Eine Beweisidee dazu und Hinweise auf die entsprechende Originalliteratur findet man in Parlett [81].

Theorem 13.20. *Falls alle Eigenwerte der symmetrischen Matrix $A \in \mathbb{R}^{N\times N}$ einfach auftreten, so gilt für die Matrizen $A^{(m)}$ des zyklischen Jacobi-Verfahrens*

$$\mathcal{S}(A^{(m+1)}) \;\le\; \frac{\mathcal{S}(A^{(m)})^2}{\delta} \qquad \text{für } m = 1, 2, \ldots, \quad \text{mit } \delta := \min_{\lambda,\,\mu\,\in\,\sigma(A),\;\lambda\neq\mu} |\lambda - \mu|.$$

13.5 Das *QR*-Verfahren

13.5.1 Eindeutigkeit und Stetigkeit der *QR*-Faktorisierung einer Matrix

Für das in den folgenden Abschnitten 13.5.2–13.5.3 behandelte QR-Verfahren zur approximativen Bestimmung der Eigenwerte einer Matrix werden die folgenden Aussagen über Eindeutigkeit und Stetigkeit der QR-Faktorisierung einer Matrix benötigt.

Lemma 13.21 (Eindeutigkeit der QR-Faktorisierung). *Für Orthogonalmatrizen Q_1, $Q_2 \in \mathbb{R}^{N\times N}$ und reguläre obere Dreiecksmatrizen R_1, $R_2 \in \mathbb{R}^{N\times N}$ sei*

$$Q_1 R_1 \;=\; Q_2 R_2$$

erfüllt. Dann existiert eine Vorzeichenmatrix $S = \operatorname{diag}(\sigma_1, \ldots, \sigma_N) \in \mathbb{R}^{N\times N}$ *mit $\sigma_k \in \{-1, 1\}$, so dass Folgendes gilt,*

$$Q_2 \;=\; Q_1 S, \qquad R_2 \;=\; S R_1.$$

BEWEIS. Nach Voraussetzung gilt

$$Q_1^{-1} Q_2 \;=\; R_1 R_2^{-1} \;=:\; S.$$

Es sind Produkte und Inverse von orthogonalen Matrizen wieder orthogonal, und entsprechendes gilt für obere Dreiecksmatrizen. Folglich ist S sowohl obere Dreiecksmatrix als auch orthogonal,

$$S^{-1} \;=\; S^\top, \qquad S \;=\; \left(\begin{array}{cc} \ast & \ast \\ & \ddots \\ 0 & \ast \end{array}\right) \;\in\; \mathbb{R}^{N\times N}. \tag{13.25}$$

Damit kann S nur eine Diagonalmatrix sein, $S = \operatorname{diag}(\sigma_1, \ldots, \sigma_N) \in \mathbb{R}^{N\times N}$, und wieder wegen $S^{-1} = S^\top$ erhält man $\sigma_k = 1/\sigma_k$ für $k = 1, 2, \ldots, N$, woraus die Aussage des Lemmas folgt. □

Definition 13.22. Für Matrizen $A_m = (a_{jk}^{(m)})$ und $A = (a_{jk}) \in \mathbb{R}^{N\times N}$ schreibt man

$$A_m \to A \quad \text{für } m \to \infty \quad :\Longleftrightarrow \quad a_{jk}^{(m)} \to a_{jk} \quad \text{für } m \to \infty \quad (j, k = 1, 2, \ldots, N).$$

Bekanntermaßen gilt $A_m \to A$ für $m \to \infty$ genau dann, wenn $\|A_m - A\| \to 0$ für $m \to \infty$ für irgendeine Matrixnorm $\|\cdot\| : \mathbb{R}^{N\times N} \to \mathbb{R}$ erfüllt ist. Für die Konvergenzbetrachtungen des noch vorzustellenden QR-Verfahrens wird das folgende Resultat über die lokale Lipschitzstetigkeit der QR-Faktorisierung benötigt. Im Folgenden ist $\mathcal{O}(\Delta_m)$ eine Kurzschreibweise für $\mathcal{O}(\|\Delta_m\|_2)$.

Lemma 13.23 (Stetigkeit der QR-Faktorisierung). *Für Orthogonalmatrizen $Q_m,\ Q \in \mathbb{R}^{N\times N}$ und obere Dreiecksmatrizen $R_m,\ R \in \mathbb{R}^{N\times N}$ sei*

$$\overbrace{Q_m R_m - Q R}^{=:\ \Delta_m} \;\to\; 0 \qquad \textit{für } m \to \infty \tag{13.26}$$

erfüllt, und die Matrix $Q R \in \mathbb{R}^{N\times N}$ sei regulär. Dann existieren Vorzeichenmatrizen

$$S_m \;=\; \operatorname{diag}(\sigma_1^{(m)}, \ldots, \sigma_N^{(m)}) \;\in\; \mathbb{R}^{N\times N} \qquad \textit{mit } \sigma_k^{(m)} \in \{-1, 1\}, \tag{13.27}$$

mit

$$Q_m S_m \;=\; Q \;+\; \mathcal{O}(\Delta_m), \qquad S_m R_m \;=\; R \;+\; \mathcal{O}(\Delta_m) \qquad \textit{für } m \to \infty. \tag{13.28}$$

BEWEIS. Es ist die Matrix R regulär, da Q und QR reguläre Matrizen sind, und somit können wir

$$\widehat{R}_m \;:=\; R_m R^{-1}$$

betrachten. Als Erstes beobachtet man

$$\widehat{R}_m^\top \widehat{R}_m \;=\; I \;+\; \mathcal{O}(\Delta_m) \qquad \text{für} \quad m \to \infty, \tag{13.29}$$

was sich wie folgt ergibt,

$$\begin{aligned}
\widehat{R}_m^\top \widehat{R}_m \;&=\; (R^{-1})^\top R_m^\top R_m R^{-1} \;=\; (R^\top)^{-1}(Q_m R_m)^\top (Q_m R_m) R^{-1} \\
&\overset{(*)}{=}\; (R^\top)^{-1}\big((QR)^\top + \mathcal{O}(\Delta_m)\big)\big(QR + \mathcal{O}(\Delta_m)\big) R^{-1} \\
&=\; \underbrace{(R^\top)^{-1} R^\top R R^{-1}}_{=\, I} + \mathcal{O}(\Delta_m) \qquad \text{für} \quad m \to \infty,
\end{aligned}$$

wobei in $(*)$ noch zu beachten ist, dass $\| B \|_2 = \| B^\top \|_2$ gilt für beliebige Matrizen $B \in \mathbb{R}^{N\times N}$. Im Folgenden wird mithilfe von (13.29) nachgewiesen, dass für gewisse Vorzeichenmatrizen $S_m \in \mathbb{R}^{N\times N}$ von der Form (13.27) Folgendes gilt,

$$S_m \widehat{R}_m \;=\; I \;+\; \mathcal{O}(\Delta_m) \qquad \text{für} \quad m \to \infty. \tag{13.30}$$

Aus (13.30) folgert man dann nämlich die Darstellung (13.28),

$$\begin{aligned}
S_m R_m \;&=\; S_m \widehat{R}_m R \;=\; R \;+\; \mathcal{O}(\Delta_m), \\
Q_m S_m \;&\overset{(\bullet)}{=}\; (Q_m R_m)(S_m R_m)^{-1} \;\overset{(\bullet\bullet)}{=}\; (QR + \mathcal{O}(\Delta_m))(R^{-1} + \mathcal{O}(\Delta_m)) \\
&=\; Q \;+\; \mathcal{O}(\Delta_m) \qquad \text{für} \quad m \to \infty.
\end{aligned}$$

Hierbei ist in $(\bullet)$ zu beachten, dass nach Voraussetzung $S_m^2 = I$ gilt, und für hinreichend große m ist die Matrix R_m regulär, was sich beispielsweise aus (13.26), der Regularität von QR und der Eigenschaft $\| Q_m^{-1} \|_2 = 1$ ergibt. Die Identität $(\bullet\bullet)$ ist eine Folgerung aus Korollar 4.50.

Im Folgenden wird nun die Konvergenzaussage (13.30) nachgewiesen. Inverse und Produkte von oberen Dreiecksmatrizen bilden wieder obere Dreiecksmatrizen, somit ist insbesondere $\widehat{R}_m$ eine obere Dreiecksmatrix. Man erhält dann folgende Zerlegung,

$$\widehat{R}_m \;=\; \begin{pmatrix} \widehat{r}_{11}^{(m)} & \times & \dots & \times \\ & \widehat{r}_{22}^{(m)} & \ddots & \vdots \\ & & \ddots & \times \\ & & & \widehat{r}_{NN}^{(m)} \end{pmatrix} \;=:\; \underbrace{\operatorname{diag}(\widehat{r}_{11}^{(m)}, \dots, \widehat{r}_{NN}^{(m)})}_{=:\, D_m} \;+\; \underbrace{\begin{pmatrix} 0 & \times & \dots & \times \\ \vdots & \ddots & \ddots & \vdots \\ \vdots & & \ddots & \times \\ 0 & \dots & \dots & 0 \end{pmatrix}}_{=:\, U_m}. \tag{13.31}$$

Mit den Bezeichnungen aus (13.31) wird nun

$$D_m^2 \;=\; I \;+\; \mathcal{O}(\Delta_m), \qquad U_m \;=\; \mathcal{O}(\Delta_m) \qquad \text{für} \quad m \to \infty \tag{13.32}$$

nachgewiesen, woraus dann mit den Vorzeichenmatrizen $S_m = \operatorname{diag}(\operatorname{sgn}(\widehat{r}_{11}^{(m)}), \ldots, \operatorname{sgn}(\widehat{r}_{NN}^{(m)}))$ unmittelbar (13.30) folgt. Zum Nachweis von (13.32) beobachtet man als Erstes

$$\widehat{R}_m \;=\; (\widehat{R}_m^{\top})^{-1} \;+\; B_m \qquad \text{mit} \quad B_m \;:=\; (\widehat{R}_m^{\top})^{-1}(\widehat{R}_m^{\top}\widehat{R}_m \;-\; I\,).$$

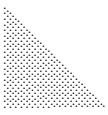
Mit (13.29) folgt

$$B_m \;=\; \mathcal{O}(\Delta_m),$$

wobei noch zu beachten ist, dass (13.29) die Beschränktheit der Matrixfolge $\widehat{R}_0^{-1}$, $\widehat{R}_1^{-1}, \ldots$ nach sich zieht, $\|\,\widehat{R}_m^{-1}\,\|_2 = \|\,(\widehat{R}_m^{\top}\widehat{R}_m)^{-1}\,\|_2^{1/2} \to 1$ für $m \to \infty$.

Zum Zweiten ist offensichtlich $\widehat{R}_m^{\top}$ eine untere Dreiecksmatrix, und Inverse von unteren Dreiecksmatrizen sind wieder untere Dreiecksmatrizen, so dass $(\widehat{R}_m^{\top})^{-1}$ eine untere Dreiecksmatrix ist. Daher stimmt notwendigerweise das strikte obere Dreieck von B_m mit dem strikten oberen Dreieck von U_m überein. Insgesamt erhält man damit folgende Darstellung,

$$B_m \;=\; \left(\quad\right) \;+\; U_m \;=\; \mathcal{O}(\Delta_m).$$



Es ist nun klar, dass sich daraus die zweite Identität in (13.32) ergibt, und abschließend wird die erste Identität in (13.32) nachgewiesen,

$$\begin{aligned} D_m^2 \;&=\; D_m^{\top} D_m \;=\; (\widehat{R}_m^{\top} - U_m^{\top})(\widehat{R}_m - U_m) \\ &=\; \underbrace{\widehat{R}_m^{\top}\widehat{R}_m}_{=I+\mathcal{O}(\Delta_m)} \;-\; \underbrace{\widehat{R}_m^{\top}U_m}_{=\mathcal{O}(\Delta_m)} \;-\; \underbrace{U_m^{\top}\widehat{R}_m}_{=\mathcal{O}(\Delta_m)} \;+\; \underbrace{U_m^{\top}U_m}_{\mathcal{O}(\Delta_m)} \;=\; I \;+\; \mathcal{O}(\Delta_m) \quad \text{für } m \to \infty. \end{aligned}$$

Damit ist (13.32) und somit auch (13.30) nachgewiesen, und man erhält die Stetigkeitsaussage (13.28). □

13.5.2 Definition des QR-Verfahrens

Der folgende Algorithmus beschreibt in Form eines Pseudocodes das QR-Verfahren zur approximativen Bestimmung der Eigenwerte einer Matrix A.

Algorithmus 13.24 (QR-Verfahren). Sei $A \in \mathbb{R}^{N\times N}$ eine beliebige reguläre Matrix.

```
A^(1) := A;
for m = 1, 2, ...:
    bestimme Faktorisierung A^(m) = Q_m R_m
    mit Q_m ∈ R^{N×N} orthogonal und R_m ∈ R^{N×N}
    von oberer Dreiecksgestalt;
    A^(m+1) := R_m Q_m ∈ R^{N×N};
end
```

△

Wie sich gleich herausstellen wird, approximieren die Diagonaleinträge von $A^{(m)}$ unter geeigneten Bedingungen für $m \to \infty$ die Eigenwerte der Matrix A. Hierbei werden die folgenden Darstellungen für die Iterationsmatrizen $A^{(m)}$ und die Potenzen A^m benötigt.

Lemma 13.25. *Mit den Bezeichnungen aus Algorithmus 13.24 sowie der Notation*

$$Q_{1\dots m} := Q_1 Q_2 \cdots Q_m, \qquad R_{m\dots 1} := R_m R_{m-1} \cdots R_1, \tag{13.33}$$

gilt

$$\begin{aligned} A^{(m+1)} &= Q_m^{-1} A^{(m)} Q_m, && m = 1, 2, \dots, \\ \text{—«—} &= Q_{1\dots m}^{-1} A Q_{1\dots m}, && \text{———«———}, \\ A^m &= Q_{1\dots m} R_{m\dots 1}, && \text{———«———}. \end{aligned}$$

BEWEIS. Die erste Identität ist unmittelbar einsichtig, und daraus resultiert dann die zweite Identität,

$$\begin{aligned} A^{(m+1)} = Q_m^{-1} A^{(m)} Q_m &= Q_m^{-1} Q_{m-1}^{-1} A^{(m-1)} Q_{m-1} Q_m \\ &= \dots = Q_m^{-1} \cdots Q_1^{-1} A Q_1 \cdots Q_m. \end{aligned}$$

Die dritte Identität erhält man mittels vollständiger Induktion unter Verwendung des folgenden Arguments,

$$\begin{aligned} Q_{1\dots m} R_{m\dots 1} &= Q_{1\dots m-1} Q_m R_m R_{m-1\dots 1} = Q_{1\dots m-1} A^{(m)} R_{m-1\dots 1} \\ &\overset{(*)}{=} A Q_{1\dots m-1} R_{m-1\dots 1} \qquad \text{für } m \geq 1, \end{aligned}$$

wobei in $(*)$ die gerade bewiesene zweite Identität eingeht. Damit ist Lemma 13.25 vollständig bewiesen. □

Wie sich im Verlauf des Beweises für den folgenden zentralen Konvergenzsatz herausstellen wird, hat die QR-Faktorisierung $A^m = Q_{1\dots m} R_{m\dots 1}$ für die m-te Potenz der Matrix A insofern eine besondere Bedeutung, als dass sich die Matrix $R_{m\dots 1}$ bis auf die Vorzeichenwahl als ein Produkt von drei Matrizen darstellen lässt, bei der die Diagonalmatrix $\operatorname{diag}(\lambda_1^m, \dots, \lambda_N^m)$ den dominanten Faktor darstellt. Weiter zeigt sich schließlich, dass die Matrix $A^{(m)} = Q_{1\dots m-1}^{-1} A^m R_{m-1\dots 1}^{-1}$ dann eine Normierung von A^m darstellt, bei der sich auf der Diagonalen die Werte $\lambda_1, \dots, \lambda_N$ herauskristallisieren.

13.5.3 Konvergenz des QR-Verfahrens für betragsmäßig einfache Eigenwerte

Unter gewissen Bedingungen konvergieren für $m \to \infty$ die Diagonaleinträge von $A^{(m)}$ gegen die betragsmäßig fallend sortierten Eigenwerte von A, wobei die Konvergenzgeschwindigkeit von der betragsmäßig betrachteten Trennung der Eigenwerte abhängt:

Theorem 13.26. *Die Matrix $A \in \mathbb{R}^{N\times N}$ sei regulär und diagonalisierbar mit betragsmäßig einfachen Eigenwerten $\lambda_1, \ldots, \lambda_N \in \mathbb{R}$, die o.B.d.A. betragsmäßig fallend angeordnet seien,*

$$|\lambda_1| > |\lambda_2| > \ldots > |\lambda_N| > 0, \tag{13.34}$$

und die Inverse der Matrix $T = \big(v_1 | \ldots | v_N\big) \in \mathbb{R}^{N\times N}$ mit Eigenvektoren $v_k \in \mathbb{R}^N$ zu λ_k besitze ohne Zeilenvertauschung eine LR-Faktorisierung.[6] *Dann gilt für das in Algorithmus 13.24 beschriebene QR-Verfahren*

$$A^{(m)} = S_m U S_m + \mathcal{O}(q^m) \quad \text{für } m \to \infty, \qquad \text{mit } q := \max_{k=1..N-1} \Big|\frac{\lambda_{k+1}}{\lambda_k}\Big|,$$

mit geeigneten Matrizen von der Form

$$S_m = \operatorname{diag}(\sigma_1^{(m)}, \ldots, \sigma_N^{(m)}) \in \mathbb{R}^{N\times N}, \quad \big(\sigma_k^{(m)} \in \{-1, 1\}\big), \qquad U = \begin{pmatrix} \lambda_1 & \times & \ldots & \times \\ & \ddots & \ddots & \vdots \\ & & \ddots & \times \\ & & & \lambda_N \end{pmatrix} \in \mathbb{R}^{N\times N}. \tag{13.35}$$

Insbesondere approximieren also die Diagonaleinträge von $A^{(m)} = (a_{jk}^{(m)})$ die betragsmäßig fallend sortierten Eigenwerte von A,

$$\max_{k=1..N} |a_{kk}^{(m)} - \lambda_k| = \mathcal{O}(q^m) \quad \text{für } m \to \infty.$$

BEWEIS. Für die Eigenvektormatrix $T \in \mathbb{R}^{N\times N}$ aus der Voraussetzung des Theorems betrachte man eine QR-Faktorisierung,

$$T = Q\widehat{R}, \quad Q \in \mathbb{R}^{N\times N} \text{ orthogonal}, \quad \widehat{R} = \begin{pmatrix} \\ \\ \\ \end{pmatrix} \in \mathbb{R}^{N\times N}. \tag{13.36}$$

Es wird nun Folgendes nachgewiesen,

$$A^{(m)} = S_m(\widehat{R}D\widehat{R}^{-1})S_m + \mathcal{O}(q^m) \quad \text{für } m \to \infty \tag{13.37-a}$$

mit Matrizen $S_m \in \mathbb{R}^{N\times N}$ von der Form (13.27) und der Diagonalmatrix

$$D := \operatorname{diag}(\lambda_1, \ldots, \lambda_N) \in \mathbb{R}^{N\times N}. \tag{13.37-b}$$

Die Aussage des Theorems erhält man danach mit der Matrix $U := \widehat{R}D\widehat{R}^{-1}$. Für den Nachweis von (13.37) benötigt man die vorausgesetzte Faktorisierung der Form

$$T^{-1} = LR, \quad L = \begin{pmatrix} 1 & & & \\ \times & \ddots & & \\ \vdots & \ddots & \ddots & \\ \times & \ldots & \times & 1 \end{pmatrix} \in \mathbb{R}^{N\times N}, \quad R = \begin{pmatrix} \\ \\ \\ \end{pmatrix} \in \mathbb{R}^{N\times N}, \tag{13.38}$$

[6] Eine detaillierte Formulierung finden Sie in (13.38) im Beweis. Eine Erläuterung dazu liefert die anschließende Bemerkung 13.27.

und beobachtet als Erstes, dass

$$L_m := D^m L D^{-m} = I + \mathcal{O}(q^m) \quad \text{für} \quad m \to \infty \tag{13.39}$$

gilt, denn mit der Notation $L = (L_{jk})$ gilt $L_m = ((\lambda_j/\lambda_k)^m L_{jk})$, und dann folgt (13.39) aus der Ungleichung $|\lambda_j/\lambda_k| \le q$ für $j \ge k+1$. Im Weiteren wird eine QR-Faktorisierung von $\widehat{R} L_m \in \mathbb{R}^{N\times N}$ benötigt,

$$\widehat{R} L_m =: \widehat{Q}_m \widehat{R}_m, \qquad \widehat{Q}_m \in \mathbb{R}^{N\times N} \text{ orthogonal}, \qquad \widehat{R}_m = \begin{pmatrix} \ddots \\ & \ddots \end{pmatrix} \in \mathbb{R}^{N\times N}.$$

Man erhält aus (13.39) die Konvergenz $\widehat{Q}_m \widehat{R}_m = \widehat{R} + \mathcal{O}(q^m) = I\widehat{R} + \mathcal{O}(q^m)$ für $m \to \infty$, und Lemma 13.23 über die Stetigkeit der QR-Faktorisierung liefert dann mit einer entsprechenden Vorzeichenwahl in den Spalten der Matrix $\widehat{Q}_m$ beziehungsweise den Zeilen der Matrix $\widehat{R}_m$ Folgendes,

$$\widehat{Q}_m = I + \mathcal{O}(q^m), \qquad \widehat{R}_m = \widehat{R} + \mathcal{O}(q^m) \quad \text{für} \quad m \to \infty. \tag{13.40}$$

Diese Konvergenzaussage ist der erste Schritt beim Nachweis von (13.37).

Im zweiten Schritt ergeben sich für die Potenzen A^m, $m \ge 1$, die beiden folgenden QR-Faktorisierungen,

$$A^m = TD^mT^{-1} \overset{(*)}{=} Q\widehat{R}D^m LR \overset{(13.39)}{=} Q \underbrace{\widehat{R}L_m}_{\widehat{Q}_m\widehat{R}_m} D^m R = \underbrace{Q\widehat{Q}_m}_{\text{orthog.}} \underbrace{\widehat{R}_m D^m R}_{\text{Dreieck}}, \tag{13.41}$$

$$A^m = Q_{1\dots m} R_{m\dots 1}, \tag{13.42}$$

wobei in der ersten Identität von (13.41) die Faktorisierung $A = TDT^{-1}$ eingeht, und die Identität $(*)$ resultiert aus (13.36) und (13.38). Die Identität (13.42) erhält man aus Lemma 13.25. Die Eindeutigkeit der QR-Faktorisierung (vergleiche Lemma 13.21) liefert dann

$$\begin{aligned} Q_{1\dots m} &= Q\widehat{Q}_m S_{m+1}, \\ R_{m\dots 1} &= S_{m+1}\widehat{R}_m D^m R, \quad \text{mit} \quad S_{m+1} = \operatorname{diag}(\sigma_1^{(m+1)}, \dots, \sigma_N^{(m+1)}) \in \mathbb{R}^{N\times N}, \\ &\qquad \big(\sigma_k^{(m+1)} \in \{-1, 1\} \text{ geeignet}\big). \end{aligned}$$

Daraus erhält man

$$\begin{aligned} Q_m &= Q_{1\dots m-1}^{-1} Q_{1\dots m} = S_m \widehat{Q}_{m-1}^{-1} \overbrace{Q^{-1}Q}^{=I} \widehat{Q}_m S_{m+1}, \\ R_m &= R_{m\dots 1} R_{m-1\dots 1}^{-1} = S_{m+1}\widehat{R}_m \underbrace{D^m R R^{-1} (D^{-1})^{m-1}}_{=D} \widehat{R}_{m-1}^{-1} S_m, \end{aligned}$$

und daraus wiederum

$$A^{(m)} = Q_m R_m = S_m \underbrace{\widehat{Q}_{m-1}^{-1}}_{\to I} \underbrace{\widehat{Q}_m}_{\to I} \underbrace{S_{m+1}^2}_{=I} \underbrace{\widehat{R}_{m+1}}_{\to \widehat{R}} D \underbrace{\widehat{R}_m^{-1}}_{\to \widehat{R}^{-1}} S_m,$$

wobei die angegebenen Konvergenzeigenschaften mit der Rate $\mathcal{O}(q^m)$ gelten, wie man der Darstellung (13.40) entnimmt. Daraus erhält man schließlich die Identität (13.37), $S_m A^{(m)} S_m = \widehat{R}\, D\, \widehat{R}^{-1} + \mathcal{O}(q^m)$ für $m \to \infty$. Dies komplettiert den Beweis des Theorems. □

Bemerkung 13.27. (a) Die Bedingung der Existenz einer LR-Faktorisierung für die Inverse der in Theorem 13.26 beschriebenen Eigenvektormatrix T ist äquivalent zu der Eigenschaft

$$\operatorname{span}\{\mathbf{e}_1, \ldots, \mathbf{e}_n\} \cap \operatorname{span}\{v_{n+1}, \ldots, v_N\} = \{0\} \qquad \text{für} \quad n = 1, 2, \ldots, N-1,$$

siehe Aufgabe 13.2. Hier bezeichnet $\mathbf{e}_k \in \mathbb{R}^N$ den k-ten Einheitsvektor. Wegen der fehlenden Kenntnis der Eigenvektoren $v_1, \ldots, v_N$ ist diese Bedingung praktisch nicht nachprüfbar.

(b) Im Falle komplexer Eigenwerte, $\sigma(A) \not\subset \mathbb{R}$, ist die Bedingung (13.34) des Satzes nicht erfüllt und auch die Aussage des zugehörigen Theorems verliert ihre Gültigkeit. Einzelheiten über die erforderlichen Modifikationen finden Sie beispielsweise in Oevel [78] und in Stoer/Bulirsch [99].

(c) Bei vollbesetzten Matrizen erfordert jeder Schritt des QR-Verfahrens wegen der notwendigen Berechnung einer QR-Faktorisierung $cN^3 + \mathcal{O}(N^2)$ arithmetische Operationen. Daher ist es zweckmäßiger, zunächst eine Ähnlichkeitstransformation auf Hessenberggestalt gemäß Abschnitt 13.2 durchzuführen und die entstehende Matrix mit dem QR-Verfahren zu bearbeiten. Weitere Einzelheiten hierzu werden im folgenden Abschnitt 13.5.4 vorgestellt.

(d) Eine alternative Präsentation des QR-Verfahrens findet man in Kress [63] (siehe auch Watkins [109]). △

13.5.4 Praktische Durchführung des QR-Verfahrens für Hessenbergmatrizen

Ausgehend von dem letzten Aspekt der Bemerkung 13.27 wird im Folgenden für den Spezialfall einer Hessenbergmatrix $A \in \mathbb{R}^{N\times N}$ eine effiziente Vorgehensweise zur Berechnung der Iterierten[7] $A^{(2)}, A^{(3)}, \ldots$ des QR-Verfahrens beschrieben.

Prinzipielles Vorgehen bei der Durchführung des Schritts $A^{(m)} \to A^{(m+1)}$

Zur Durchführung des Schritts $A^{(m)} \to A^{(m+1)}$ hat man nach Definition zunächst eine QR-Faktorisierung $A^{(m)} = Q_m R_m$ für die Hessenbergmatrix $A^{(m)} = (a_{jk}^{(m)})$ zu bestimmen, was sukzessive in der folgenden Form geschieht,

$$\left.\begin{aligned} A^{(m)} &= A^{(m,1)} \to A^{(m,2)} \to \ldots \to A^{(m,N)} =: R_m, \\ A^{(m,k+1)} &= S_{mk}^{\top} A^{(m,k)}, \qquad k = 1, 2, \ldots, N-1, \end{aligned}\right\} \tag{13.43}$$

[7] die allesamt von Hessenbergform sind, siehe Übungsaufgabe 13.3

mit dem Ziel der schrittweisen Elimination der unteren Nebendiagonaleinträge,

$$A^{(m,k)} = \begin{pmatrix} a_{11}^{(m,k)} & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & a_{1N}^{(m,k)} \\ 0 & \ddots & & & & & & \vdots \\ \vdots & \ddots & a_{k-1,k-1}^{(m,k)} & \cdots & \cdots & \cdots & \cdots & a_{k-1,N}^{(m,k)} \\ \vdots & & 0 & a_{kk}^{(m,k)} & \cdots & \cdots & \cdots & a_{kN}^{(m,k)} \\ \vdots & & 0 & a_{k+1,k}^{(m)} & a_{k+1,k+1}^{(m)} & \cdots & \cdots & a_{k+1,N}^{(m)} \\ \vdots & & & 0 & a_{k+2,k+1}^{(m)} & a_{k+2,k+2}^{(m)} & & a_{k+2,N}^{(m)} \\ \vdots & & & & \ddots & \ddots & \ddots & \vdots \\ 0 & \cdots & \cdots & \cdots & \cdots & 0 & a_{N,N-1}^{(m)} & a_{NN}^{(m)} \end{pmatrix} \begin{matrix} \\ \\ \\ \leftarrow \text{Zeile } k \\ \leftarrow \text{Zeile } k{+}1 \\ \\ \\ \\ \end{matrix} \quad (13.44)$$

$$\uparrow$$
$$\text{Spalte } k$$

wobei die verwendete Notation für die Einträge der Matrix $A^{(m,k)}$ dadurch gerechtfertigt ist, dass die Matrizen $A^{(m,k)}$ und $A^{(m)}$ in den Zeilen $k+1, k+2, \ldots, N$ übereinstimmen. Das angesprochene Ziel wird erreicht, wenn man im Zuge der Transformation (13.43) spezielle Givensrotationen $S_{mk} \in \mathbb{R}^{N\times N}$ von der Form

$$S_{mk} = \begin{pmatrix} 1 & & & & & & & \\ & \ddots & & & & & & \\ & & 1 & & & & & \\ & & & c & -s & & & \\ & & & s & c & & & \\ & & & & & 1 & & \\ & & & & & & \ddots & \\ & & & & & & & 1 \end{pmatrix} \begin{matrix} \\ \\ \\ \leftarrow \text{ Zeile } k \\ \leftarrow \text{ Zeile } k+1 \\ \\ \\ \\ \end{matrix}$$

$$\uparrow$$
$$\text{Spalte } k$$

verwendet mit den folgenden Setzungen für die Zahlen $c,\ s \in \mathbb{R}$,

$$\begin{aligned} ca + sb &= \frac{1}{\sqrt{a^2+b^2}} \\ -sa + cb &= 0 \end{aligned} \quad \text{bzw.} \quad \begin{aligned} c &= \frac{a}{\sqrt{a^2+b^2}} \\ s &= \frac{b}{\sqrt{a^2+b^2}} \end{aligned} \quad \text{mit} \quad \begin{aligned} a &:= a_{kk}^{(m,k)}, \\ b &:= a_{k+1,k}^{(m)}, \end{aligned}$$

wobei noch $b \neq 0$ angenommen wird. Gilt andernfalls $b = 0$, so ist keine Transformation erforderlich und man kann $c = 1,\ s = 0$ setzen. In jedem Fall gilt $c^2 + s^2 = 1$ und S_{mk} ist somit eine Orthogonalmatrix.

Mit diesen Notationen ändert sich bei einer Transformation von der Form $A^{(m,k)} \to S_{mk}^{\top} A^{(m,k)}$ lediglich die in (13.44) gekennzeichnete Teilmatrix $\begin{pmatrix} * & \cdots & * \\ * & \cdots & * \end{pmatrix} \in$

$\mathbb{R}^{2\times(N-k+1)}$ zu

$$\begin{pmatrix} c & s \\ -s & c \end{pmatrix} \boxed{\begin{matrix} * & \cdots & \cdots & * \\ * & \cdots & \cdots & * \end{matrix}} = \boxed{\begin{matrix} * & \cdots & \cdots & * \\ 0 & * & \cdots & * \end{matrix}} \in \mathbb{R}^{2\times(N-k+1)}.$$

Nach der Gewinnung einer QR-Faktorisierung $A^{(m)} = Q_m R_m$ für die Hessenbergmatrix $A^{(m)}$ besteht der zweite Teil bei der Durchführung des Schritts $A^{(m)} \to A^{(m+1)}$ des QR-Verfahrens in der Berechnung des Matrixprodukts $A^{(m+1)} = R_m Q_m$ mit $Q_m := S_{m1} S_{m2} \cdots S_{m,N-1}$.

Die Durchführung des Schritts $A^{(m)} \to A^{(m+1)}$ in der Praxis

Zur Speicherplatzersparnis führt man in der Praxis die beiden genannten Teile des Schritts $A^{(m)} \to A^{(m+1)}$ simultan in der folgenden Form durch,

$$\left.\begin{aligned} A^{(m)} &= B^{(m,1)} \to B^{(m,2)} \to \ldots \to B^{(m,N)} =: A^{(m+1)}, \\ B^{(m,k+1)} &= S_{mk}^\top B^{(m,k)} S_{mk}, \qquad k = 1, 2, \ldots, N-1, \end{aligned}\right\} \tag{13.45}$$

wobei im Detail so vorgegangen wird:

Algorithmus 13.28 (QR-Verfahren für Hessenbergmatrizen). Man berechnet

$$B^{(m,k)} \overset{(k,1)}{\to} S_{mk}^\top B^{(m,k)} \overset{(k,2)}{\to} S_{mk}^\top B^{(m,k)} S_{mk} =: B^{(m,k+1)}, \quad k = 1, \ldots, N-1, \tag{13.46}$$

wobei nach dem Schritt $(k,1)$ in der Matrix $S_{mk}^\top B^{(m,k)}$ die Einträge mit den Indizes $k+1, k+1$ beziehungsweise $k+2, k+1$ mit den Werten $a_{k+1,k+1}^{(m,k+1)}$ beziehungsweise $a_{k+2,k+1}^{(m)}$ übereinstimmen und diese für die Berechnung der Givensrotation $S_{m,k+1}$ zwischenzuspeichern sind. △

Die in dem Algorithmus 13.28 gewählte Reihenfolge bei der Durchführung der Matrizenmultiplikationen führt aufgrund der Assoziativität des Matrixprodukts dennoch tatsächlich auf die Matrix

$$B^{(m,N)} = S_{m,N-1}^\top S_{m,N-2}^\top \cdots S_{m1}^\top A^{(m)} S_{m1} S_{m2} \cdots S_{m,N-1} = A^{(m+1)}.$$

Mit dem folgenden Lemma wird klar, dass sich die Werte $a_{k+1,k+1}^{(m,k+1)}$ beziehungsweise $a_{k+2,k+1}^{(m)}$ nach dem Schritt $(k,1)$ tatsächlich an den genannten Positionen stehen. (Bei dem darauf folgenden Schritt $(k,2)$ aus (13.46) werden diese überschrieben.)

Lemma 13.29. *Die in (13.46) nach dem Schritt $(k,1)$ entstehende Matrix $S_{mk}^\top B^{(m,k)}$ ist von Hessenbergform. Deren Einträge stimmen in den Spalten $k+1, k+2, \ldots, N$*

mit denen der Matrix $A^{(m,k+1)}$ aus (13.43) überein,

$$S_{mk}^{\top} B^{(m,k)} = \begin{pmatrix} * & \cdots & * & a_{1,k+1}^{(m,k+1)} & & \cdots & & \cdots & a_{1N}^{(m,k+1)} \\ * & \ddots & \vdots & \vdots & & & & & \vdots \\ 0 & \ddots & * & \vdots & & & & & \vdots \\ \vdots & \ddots & * & a_{k+1,k+1}^{(m,k+1)} & & & & & a_{k+1,N}^{(m,k+1)} \\ \vdots & & \ddots & a_{k+2,k+1}^{(m)} & a_{k+2,k+2}^{(m)} & & & & a_{k+2,N}^{(m)} \\ \vdots & & & \ddots & \ddots & & \ddots & & \vdots \\ 0 & \cdots & \cdots & \cdots & 0 & & a_{N,N-1}^{(m)} & & a_{NN}^{(m)} \end{pmatrix}.$$

BEWEIS. Es gilt offensichtlich $S_{mk}^{\top} S_{m,k-1}^{\top} \cdots S_{m1}^{\top} A^{(m)} = A^{(m,k+1)}$ und somit auch $S_{mk}^{\top} B^{(m,k)} = A^{(m,k+1)} S_{m,1} \cdots S_{m,k-1}$. Im Folgenden wird mittels vollständiger Induktion über $\ell = 1, 2, \ldots, k$ die Darstellung

$$A^{(m,k+1)} S_{m1} \cdots S_{m,\ell-1}$$

$$= \begin{pmatrix} * & \cdots & * & * & * & * & \cdots & * & a_{1,k+1}^{(m,k+1)} & & \cdots & & a_{1,N}^{(m,k+1)} \\ * & \ddots & \vdots & \vdots & \vdots & \vdots & & \vdots & \vdots & & & & \vdots \\ 0 & \ddots & * & \vdots & \vdots & \vdots & & \vdots & \vdots & & & & \vdots \\ \vdots & \ddots & * & * & \vdots & \vdots & & \vdots & \vdots & & & & \vdots \\ \vdots & & \ddots & 0 & * & \vdots & & \vdots & \vdots & & & & \vdots \\ \vdots & & & \ddots & 0 & * & & \vdots & \vdots & & & & \vdots \\ \vdots & & & & \ddots & \ddots & \ddots & \vdots & \vdots & & & & \vdots \\ \vdots & & & & & \ddots & \ddots & * & \vdots & & & & \vdots \\ \vdots & & & & & & \ddots & 0 & a_{k+1,k+1}^{(m,k+1)} & & & & a_{k+1,N}^{(m,k+1)} \\ \vdots & & & & & & & \ddots & a_{k+2,k+1}^{(m)} & a_{k+2,k+2}^{(m)} & & & a_{k+2,N}^{(m)} \\ \vdots & & & & & & & & \ddots & \ddots & \ddots & & \vdots \\ 0 & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & 0 & & a_{N,N-1}^{(m)} & a_{NN}^{(m)} \end{pmatrix} \begin{matrix} \\ \\ \\ \\ \leftarrow \text{Zeile } \ell \\ \leftarrow \text{Zeile } \ell+1 \\ \\ \\ \\ \\ \leftarrow \text{Zeile } k+1 \\ \\ \\ \\ \end{matrix} \tag{13.47}$$

$$\uparrow \text{Spalte } \ell \qquad \uparrow \text{Spalte } k+1$$

nachgewiesen, so dass die Einträge in den Spalten $k+1, k+2, \ldots, N$ mit denen der Matrix $A^{(m,k+1)}$ übereinstimmen. Die Aussage des Lemmas folgt dann aus (13.47) mit $\ell = k$. Die Identität (13.47) ist offensichtlich richtig für $\ell = 1$. Ausgehend von der Darstellung (13.47) mit einem $\ell \leq k-1$ bedeutet die Multiplikation $(A^{(m,k+1)} S_{m,1} \cdots S_{m,\ell-1}) S_{m,\ell}$ eine Transformation der in (13.47) gekennzeichneten Teilmatrix,

$$\begin{pmatrix} * & * \\ \vdots & \vdots \\ * & \vdots \\ 0 & * \end{pmatrix} \rightarrow \begin{pmatrix} * & * \\ \vdots & \vdots \\ * & \vdots \\ 0 & * \end{pmatrix} \begin{pmatrix} c & -s \\ s & c \end{pmatrix} = \begin{pmatrix} * & * \\ \vdots & \vdots \\ \vdots & \vdots \\ * & * \end{pmatrix} \in \mathbb{R}^{(\ell+1)\times 2},$$

so dass auch der Induktionsschritt abgeschlossen ist. □

Bemerkung 13.30. Mit dem Beweis wird auch deutlich, dass für $k = 1, 2, \ldots, N-1$ nach dem ersten Teilschritt $(k, 1)$ aus (13.46) die entstehende Matrix $S_{mk}^{\top} B^{(m,k)}$ von Hessenberggestalt ist. Für die Matrizen $B^{(m,2)}, \ldots, B^{(m,N-1)}$ gelten die folgenden Darstellungen,

$$B^{(m,k)} = \begin{pmatrix} * & \cdots & \cdots & \cdots & \cdots & \cdots & * \\ * & \ddots & & & & & \vdots \\ 0 & \ddots & \ddots & & & & \vdots \\ \vdots & \ddots & \ddots & \ddots & & & \vdots \\ \vdots & & 0 & \ddots & \ddots & & \vdots \\ \vdots & & & * & \ddots & \ddots & \vdots \\ \vdots & & & & 0 & \ddots & \vdots \\ \vdots & & & & & \ddots & \vdots \\ 0 & \cdots & \cdots & \cdots & \cdots & 0 & * \end{pmatrix} \begin{matrix} \\ \\ \\ \\ \leftarrow \text{ Zeile } k \\ \leftarrow \text{ Zeile } k+1 \\ \\ \\ \end{matrix} \quad (k = 2, 3, \ldots, N-1),$$

$$\uparrow \quad \uparrow$$
$$\text{Spalte } k \quad \text{Spalte } k+1$$

so dass die Matrizen $B^{(m,2)}, \ldots, B^{(m,N-1)}$ jeweils an der Position $(k+1, k-1)$ von einer Hessenberggestalt abweichen. Beim Übergang $B^{(m,k)} \to B^{(m,k+1)}$ wird zunächst der durch $\begin{pmatrix} * & \cdots & * \\ * & \cdots & * \end{pmatrix} \in \mathbb{R}^{2\times(N-k+2)}$ gekennzeichnete Block durch die Transformation

$$\begin{pmatrix} * & \cdots & * \\ * & \cdots & * \end{pmatrix} \to \begin{pmatrix} c & s \\ -s & c \end{pmatrix} \begin{pmatrix} * & \cdots & * \\ * & \cdots & * \end{pmatrix} \in \mathbb{R}^{2\times(N-k+2)} \tag{13.48}$$

überschrieben, und in der daraus entstehenden Matrix $S_{mk}^{\top} B^{(m,k)}$ wird anschließend mit der gekennzeichneten Teilmatrix $\in \mathbb{R}^{(k+2)\times 2}$ die Transformation

$$\begin{pmatrix} * & * \\ \vdots & \vdots \\ * & * \end{pmatrix} \to \begin{pmatrix} * & * \\ \vdots & \vdots \\ * & * \end{pmatrix} \begin{pmatrix} c & -s \\ s & c \end{pmatrix} \tag{13.49}$$

durchgeführt. △

Mit der vorangegangenen Bemerkung lässt sich leicht der bei der Durchführung des Schritts $A^{(m)} \to A^{(m+1)}$ anfallende Gesamtaufwand ermitteln.

Theorem 13.31. *Für Hessenbergmatrizen A lässt sich das Schema (13.45) zur Durchführung des Schritts $A^{(m)} \to A^{(m+1)}$ des QR-Verfahrens mit*

$$6N^2\big(1 + \mathcal{O}\big(\tfrac{1}{N}\big)\big)$$

arithmetischen Operationen realisieren.

BEWEIS. Eine Transformation der Form (13.48) erfordert $(N-k+2)\times 2\times 2 = 4(N-k+2)$ Multiplikationen und $2(N-k+2)$ Additionen, insgesamt fallen dabei also $6(N-k+2)$ arithmetische Operationen an. Entsprechend erfordert eine Transformation der Form (13.49) $6(k+2)$ arithmetische Operationen, und der Schritt k aus (13.46) – bestehend aus den beiden Transformationen (13.48)–(13.49) – benötigt also $6(N+2)$ arithmetische Operationen. Für die $N-1$ Schritte aus (13.46) sind demnach $6N^2(1+\mathcal{O}(1/N))$ arithmetische Operationen durchzuführen. Die Berechnung der Givensrotationen erfordert nochmals die dagegen nicht weiter ins Gewicht fallende Berechnung von N Quadratwurzeln und $2N$ Quotienten. □

13.6 Das *LR*-Verfahren

Alternativ zum QR-Verfahren kann man auch folgendermaßen vorgehen:

Algorithmus 13.32 (*LR-Verfahren*). Sei $A \in \mathbb{R}^{N\times N}$ eine reguläre Matrix.

```
A^(1) := A;
for m = 1,2,...:
    bestimme Faktorisierung A^(m) = L_m R_m
    mit L_m bzw. R_m ∈ R^(N×N) von unterer
    bzw. oberer Dreiecksgestalt;
    A^(m+1) := R_m L_m ∈ R^(N×N);
end
```

△

Für das LR-Verfahren lassen sich dem QR-Verfahren vergleichbare Resultate erzielen. Einzelheiten finden Sie beispielsweise in Stoer/Bulirsch [99].

13.7 Die Vektoriteration

13.7.1 Definition und Eigenschaften der Vektoriteration

Definition 13.33. Für eine gegebene Matrix $B \in \mathbb{R}^{N\times N}$ lautet die *Vektoriteration* folgendermaßen:

$$z^{(m+1)} = Bz^{(m)}, \qquad m = 0,1,\ldots \qquad (z^{(0)} \in \mathbb{R}^N). \tag{13.50}$$

Die Vektoriteration ermöglicht unter günstigen Umständen die Bestimmung des betragsmäßig größten Eigenwerts der Matrix B. Das nachfolgende Theorem liefert hierzu ein Konvergenzresultat für diagonalisierbare Matrizen $B \in \mathbb{R}^{N\times N}$ mit Eigenwerten $\lambda_1, \lambda_2, \ldots, \lambda_N \in \mathbb{C}$. Hierzu sei noch folgende Sprechweise eingeführt: für einen Index $1 \le k_* \le N$ besitzt ein gegebener Vektor $x \in \mathbb{C}^N$ *einen Anteil in* $\mathcal{N}(B-\lambda_{k_*}I)$,

falls in der eindeutigen Zerlegung[8] $x = \sum_{k=1}^{N} x_k$ mit $x_k \in \mathcal{N}(B - \lambda_k I)$ der Vektor x_{k_*} nicht verschwindet, $x_{k_*} \neq 0$.

Theorem 13.34. *Für die diagonalisierbare Matrix $B \in \mathbb{R}^{N\times N}$ mit Eigenwerten $\lambda_1, \ldots, \lambda_N \in \mathbb{C}$ gelte $\lambda_1 = \lambda_2 = \ldots = \lambda_r$, $|\lambda_r| > |\lambda_{r+1}| \geq \ldots \geq |\lambda_N|$ mit $r \leq N - 1$.[9] Falls der Startvektor $z^{(0)} \in \mathbb{R}^N$ einen Anteil in $\mathcal{N}(B - \lambda_1 I)$ besitzt, gilt für die Vektoriteration (13.50)*

$$\frac{\|z^{(m+1)}\|}{\|z^{(m)}\|} = |\lambda_1| + \mathcal{O}(q^m) \qquad \textit{für } m \to \infty, \qquad \textit{mit } q := \Big|\frac{\lambda_{r+1}}{\lambda_1}\Big| < 1,$$

mit einer beliebigen Vektornorm $\|\cdot\| : \mathbb{C}^N \to \mathbb{R}$.

BEWEIS. Es gibt eine Darstellung der Form $z^{(0)} = x_1 + \sum_{k=r+1}^{N} x_k$ mit $x_k \in \mathcal{N}(B - \lambda_k I)$, und dann gilt allgemein

$$z^{(m)} = \lambda_1^m x_1 + \sum_{k=r+1}^{N} \lambda_k^m x_k = \lambda_1^m \big(x_1 + \sum_{k=r+1}^{N} \big(\tfrac{\lambda_k}{\lambda_1}\big)^m x_k\big), \qquad m = 0, 1, \ldots . \tag{13.51}$$

Daraus erhält man nacheinander

$$\begin{aligned}
\lambda_1^{-m} z^{(m)} &= x_1 + \sum_{k=r+1}^{N} \big(\tfrac{\lambda_k}{\lambda_1}\big)^m x_k = x_1 + \mathcal{O}(q^m) \qquad \text{für } m \to \infty, \\
|\lambda_1|^{-m} \|z^{(m)}\| &= \|x_1\| + \mathcal{O}(q^m) \qquad \text{für } m \to \infty, \\
|\lambda_1|^{-1} \frac{\|z^{(m+1)}\|}{\|z^{(m)}\|} &= \frac{\|x_1\| + \mathcal{O}(q^{m+1})}{\|x_1\| + \mathcal{O}(q^m)} \overset{(*)}{=} 1 + \mathcal{O}(q^m) \qquad \text{für } m \to \infty,
\end{aligned} \tag{13.52}$$

wobei die Identität $(*)$ wegen $x_1 \neq 0$ gilt. Die Identität (13.52) liefert dann unmittelbar die Aussage des Theorems. □

Bemerkung 13.35. In Theorem 13.34 stellt die Bedingung "$z^{(0)} \in \mathbb{R}^N$ besitzt einen Anteil in $\mathcal{N}(B - \lambda_1 I)$" keine wesentliche Einschränkung dar. Selbst falls $z^{(0)}$ doch keinen Anteil in $\mathcal{N}(B - \lambda_1 I)$ besitzt, so wird sich im Verlauf der Iteration aufgrund von Rundungsfehlern die in dem Beweis von Theorem 13.34 benötigte Eigenschaft einstellen, dass die Vektoren $z^{(m)}$ Anteile in $\mathcal{N}(B - \lambda_1 I)$ besitzen. △

Das folgende Theorem liefert eine Folge reeller Zahlen, die im Falle symmetrischer Matrizen gegen den betragsmäßig größten Eigenwert konvergiert (und nicht gegen den Betrag davon).

Theorem 13.36. *Die Matrix $B \in \mathbb{R}^{N\times N}$ sei symmetrisch, und für ihre Eigenwerte $\lambda_1, \ldots, \lambda_N \in \mathbb{R}$ sei $\lambda_1 = \lambda_2 = \ldots = \lambda_r$, $|\lambda_r| > |\lambda_{r+1}| \geq \ldots \geq |\lambda_N|$ mit $r \leq$*

[8] bekanntlich gilt in der vorliegenden Situation $\mathbb{R}^N = \oplus_{k=1}^{N} \mathcal{N}(B - \lambda_k I)$

[9] Im Fall $r = N$ liegt die triviale Situation $B = \lambda_1 I$ vor.

$N-1$ erfüllt[10]*. Falls der Startvektor $z^{(0)} \in \mathbb{R}^N$ einen Anteil in $\mathcal{N}(B-\lambda_1 I)$ besitzt, so konvergiert die zur Vektoriteration gehörende Folge der* Rayleigh-Quotienten

$$r_m \;=\; \frac{(z^{(m)})^\top z^{(m+1)}}{\|z^{(m)}\|_2^2}, \qquad m \;=\; 1,2,\ldots$$

gegen den Eigenwert λ_1,

$$r_m \;=\; \lambda_1 \;+\; \mathcal{O}(q^{2m}) \qquad \textit{für} \quad m\to\infty, \qquad \textit{mit} \quad q \;:=\; \Big|\frac{\lambda_{r+1}}{\lambda_1}\Big| \;<\; 1.$$

BEWEIS. Wie im Beweis von Theorem 13.34 erhält man (vergleiche (13.51))

$$z^{(m)} \;=\; \lambda_1^m\big(x_1 \;+\; \sum_{k=r+1}^{N} \big(\tfrac{\lambda_k}{\lambda_1}\big)^m x_k\big), \qquad m \;=\; 0,1,\ldots,$$

wobei hier o.B.d.A. angenommen werden darf, dass die Eigenvektoren $x_1, x_{r+1}, x_{r+2}, \ldots, x_N \in \mathbb{R}^N$ paarweise orthogonal sind. Daraus erhält man

$$\begin{aligned}(z^{(m)})^\top z^{(m+1)} \;&=\; \lambda_1^{2m+1}\big(\|x_1\|_2^2 \;+\; \sum_{k=r+1}^{N} \big(\tfrac{\lambda_k}{\lambda_1}\big)^{2m+1}\|x_k\|_2^2\big),\\ \|z^{(m)}\|_2^2 \;&=\; \lambda_1^{2m}\big(\|x_1\|_2^2 \;+\; \sum_{k=r+1}^{N} \big(\tfrac{\lambda_k}{\lambda_1}\big)^{2m}\|x_k\|_2^2\big),\end{aligned}$$

und Quotientenbildung ergibt

$$\begin{aligned}r_m \;&=\; \lambda_1\frac{\|x_1\|_2^2 \;+\; \sum_{k=r+1}^{N}\big(\frac{\lambda_k}{\lambda_1}\big)^{2m+1}\|x_k\|_2^2,}{\|x_1\|_2^2 \;+\; \sum_{k=r+1}^{N}\big(\frac{\lambda_k}{\lambda_1}\big)^{2m}\|x_k\|_2^2} \;=\; \lambda_1\frac{\|x_1\|_2^2 \;+\; \mathcal{O}(q^{2m+1})}{\|x_1\|_2^2 \;+\; \mathcal{O}(q^{2m})}\\ &=\; \lambda_1 \;+\; \mathcal{O}(q^{2m}) \qquad \text{für} \quad m\to\infty,\end{aligned}$$

was die Aussage des Theorems liefert. □

13.7.2 Spezielle Vektoriterationen

Im Folgenden werden zwei spezielle Vektoriterationen vorgestellt.

Definition 13.37. Für eine gegebene Matrix $A \in \mathbb{R}^{N\times N}$ ist die *von Mises-Iteration* folgendermaßen definiert,

$$z^{(m+1)} \;=\; Az^{(m)}, \qquad m \;=\; 0,1,\ldots \qquad (z^{(0)} \in \mathbb{R}^N).$$

[10] der Fall $r = N$ ist trivial, $B = \lambda_1 I$

Die von Mises-Iteration erhält man mit der speziellen Wahl $B = A$ aus der Vektoriteration (13.50), und die Eigenschaften der von Mises-Iteration entnimmt man daher unmittelbar Abschnitt 13.7.1.

Definition 13.38. Für eine gegebene Matrix $A \in \mathbb{R}^{N\times N}$ und eine Zahl $\mu \in \mathbb{R}\backslash\sigma(A)$ ist die *inverse Iteration von Wielandt* folgendermaßen erklärt,

$$(A - \mu I)z^{(m+1)} \;=\; z^{(m)}, \qquad m \;=\; 0, 1, \ldots \qquad (z^{(0)} \in \mathbb{R}^N).$$

Bemerkung 13.39. Die inverse Iteration von Wielandt erhält man mit der speziellen Wahl $B = (A - \mu I)^{-1}$ aus der Vektoriteration (13.50). Abschnitt 13.7.1 liefert daher für eine symmetrische Matrix $A \in \mathbb{R}^{N\times N}$ mit Eigenwerten $\lambda_1, \ldots, \lambda_N \in \mathbb{R}$ unmittelbar das Folgende: Ist k_* ein Index, für den für $k \;=\; 1, 2, \ldots, N$

$$\text{entweder} \quad \lambda_k \;=\; \lambda_{k_*} \quad \text{oder} \quad |\lambda_{k_*} - \mu| \;<\; |\lambda_k - \mu|$$

erfüllt ist, so gilt für die dazugehörende Folge der Rayleigh-Quotienten $r_m \to (\lambda_{k_*} - \mu)^{-1}$ beziehungsweise

$$r_m^{-1} \;+\; \mu \;\to\; \lambda_{k_*} \qquad \text{für} \quad m \;\to\; \infty. \qquad \triangle$$

Weitere Themen und Literaturhinweise

Die in diesem Kapitel vorgestellten und andere Algorithmen zur numerischen Bestimmung der Eigenwerte von Matrizen finden Sie beispielsweise in den in Kapitel 12 genannten Lehrbüchern und in Bunse/Bunse-Gerstner [11] und Trefethen/Bau [104]. Verfahren zur numerischen Berechnung der Singulärwertzerlegung einer Matrix werden in [11], Deuflhard/Hohmann [22], Golub/Van Loan [35], Stoer/Bulirsch [99] und in Werner [111] vorgestellt.

Übungsaufgaben

Aufgabe 13.1. Man weise nach, dass eine obere Hessenbergmatrix durch eine Ähnlichkeitstransformation mit einer Diagonalmatrix so umgeformt werden kann, dass die unteren Nebendiagonaleinträge nur die Werte 0 oder 1 annehmen.

Aufgabe 13.2. Man zeige unter Verwendung von Aufgabe 4.8 auf Seite 98 Folgendes: für eine gegebene reguläre Matrix $T = (v_1|\ldots|v_N) \in \mathbb{R}^{N\times N}$ besitzt die Inverse T^{-1} genau dann eine LR-Faktorisierung, wenn Folgendes gilt,

$$\operatorname{span}\{\mathbf{e}_1, \ldots, \mathbf{e}_n\} \;\cap\; \operatorname{span}\{v_{n+1}, \ldots, v_N\} \;=\; \{0\} \qquad \text{für} \quad n \;=\; 1, 2, \ldots, N-1,$$

wobei $\mathbf{e}_k \in \mathbb{R}^N$ den k-ten Einheitsvektor bezeichnet.

Aufgabe 13.3. Das QR-Verfahren erhält eine Hessenberg- oder Tridiagonalform: ist die reguläre Matrix A von Hessenberg- beziehungsweise Tridiagonalform, so besitzen auch die zu dem QR-Verfahren gehörenden Matrizen $A^{(2)},\ A^{(3)},\ldots$ eine Hessenberg- beziehungsweise Tridiagonalform.

Aufgabe 13.4. Es sei $A \in \mathbb{R}^{N\times N}$ eine symmetrische Matrix mit Eigenwerten $\lambda_1 = \lambda_2 = \ldots = \lambda_r,\ |\lambda_r| > |\lambda_{r+1}| \geq \ldots \geq |\lambda_N|$. Mit der Vektorfolge $z^{(m+1)} = Az^{(m)},\ m = 0, 1, \ldots$, werde die Folge der Rayleigh-Quotienten

$$r_m \;=\; \frac{(z^{(m)})^\top z^{(m+1)}}{\|z^{(m)}\|_2^2}, \qquad m \;=\; 0, 1, \ldots,$$

gebildet mit einem Startvektor $z^{(0)}$, der einen Anteil im Eigenraum der Matrix A zum Eigenwert λ_1 besitze. Man weise Folgendes nach: für einen Eigenvektor x zum Eigenwert λ_1 gilt

$$\operatorname{sgn}(r_m)^m \frac{z^{(m)}}{\|z^{(m)}\|_2} \;=\; x \;+\; \mathcal{O}\Big(\Big|\frac{\lambda_{r+1}}{\lambda_1}\Big|^m\Big) \qquad \text{für} \quad m \to \infty.$$

Aufgabe 13.5. Es sei $A \in \mathbb{R}^{N\times N}$ eine diagonalisierbare Matrix mit Eigenwerten $\lambda_1,\ \lambda_2, \ldots, \lambda_N$, für die $\lambda_2 = -\lambda_1 < 0$ und $|\lambda_2| > |\lambda_3| \geq \ldots \geq |\lambda_N|$ gelte. Für die Vektoriteration $z^{(m+1)} = Az^{(m)},\ m = 0, 1, \ldots$ weise man Folgendes nach ($\|\cdot\|$ bezeichne irgendeine Vektornorm):

(a) Falls $z^{(0)}$ einen Anteil im Eigenraum der Matrix A zum Eigenwert λ_1 besitzt, so gilt für einen Eigenvektor x_1 zum Eigenwert λ_1 Folgendes:

$$\frac{\lambda_1 z^{(2m)} \;+\; z^{(2m+1)}}{\|\lambda_1 z^{(2m)} \;+\; z^{(2m+1)}\|} \;=\; x_1 \;+\; \mathcal{O}\Big(\Big|\frac{\lambda_3}{\lambda_1}\Big|^{2m}\Big) \qquad \text{für} \quad m \to \infty.$$

(b) Falls $z^{(0)}$ einen Anteil im Eigenraum der Matrix A zum Eigenwert λ_2 besitzt, so gilt für einen Eigenvektor x_2 zum Eigenwert λ_2 Folgendes:

$$\frac{\lambda_1 z^{(2m)} \;-\; z^{(2m+1)}}{\|\lambda_1 z^{(2m)} \;-\; z^{(2m+1)}\|} \;=\; x_2 \;+\; \mathcal{O}\Big(\Big|\frac{\lambda_3}{\lambda_1}\Big|^{2m}\Big) \qquad \text{für} \quad m \to \infty.$$

Aufgabe 13.6. Es sei λ_1 eine einfache dominante Nullstelle des Polynoms

$$p(x) \;=\; \sum_{k=0}^{n} a_k x^k \quad \text{mit} \quad a_n \;=\; 1.$$

Zu vorgegebenen hinreichend allgemeinen Startwerten $x_{1-n},\ x_{2-n}, \ldots,\ x_0 \in \mathbb{R}\backslash\{0\}$ betrachte man die Folge

$$x_{m+n} \;=\; -\sum_{k=0}^{n-1} a_k x_{m+k}, \qquad m \;=\; 1, 2, \ldots\ .$$

Durch Anwendung der Vektoriteration auf die Transponierte der frobeniusschen Begleitmatrix zu $p(x)$ weise man Folgendes nach,

$$\frac{x_{m+1}}{x_m} \;=\; \lambda_1 \;+\; \mathcal{O}\Big(\Big|\frac{\lambda_2}{\lambda_1}\Big|^m\Big) \qquad \text{für} \quad m \to \infty,$$

wobei $\lambda_2 \in \mathbb{C}$ eine nach λ_1 betragsmäßig größte Nullstelle des Polynoms p sei.

Aufgabe 13.7 (*Numerische Aufgabe*). Für die Matrix $A = (a_{jk}) \in \mathbb{R}^{N\times N}$ mit

$$a_{jk} \;=\; \begin{cases} N-j+1, & \text{falls } k \le j, \\ N-k+1, & \text{sonst,} \end{cases}$$

bestimme man für $N = 50$ und $N = 100$ mit dem LR-Algorithmus numerisch jeweils sowohl den betragsmäßig kleinsten als auch den betragsmäßig größten Eigenwert. Sei $A_m = (a_{jk}^{(m)})$, $m = 0, 1, \ldots$, die hierbei erzeugte Matrixfolge. Man breche das Verfahren ab, falls $m = 100$ oder

$$\varepsilon_m \;:=\; \max_{k=1,\ldots,N} \frac{|a_{kk}^{(m-1)} - a_{kk}^{(m)}|}{|a_{kk}^{(m-1)}|} \;\le\; 0.05$$

erfüllt ist. Man gebe außer den gewonnenen Approximationen für die gesuchten Eigenwerte auch die Werte $\varepsilon_1, \varepsilon_2, \ldots$ an.

14 Restglieddarstellung nach Peano

14.1 Einführende Bemerkungen

Für ganz unterschiedliche Verfahren (zur Lösung auch ganz unterschiedlicher Problemstellungen wie etwa Interpolation sowie numerische Integration und Differenziation) existiert ein eleganter und einheitlicher Zugang zur Herleitung von Fehlerdarstellungen. Dieser Zugang, der zudem Verallgemeinerungen schon bekannter Fehlerdarstellungen für Funktionen f mit geringeren Differenzierbarkeitseigenschaften ermöglicht, soll in dem vorliegenden Kapitel 14 in Grundzügen vorgestellt werden.

Im Folgenden wird das lineare Funktional $\mathcal{R} : C^{-1}[a,b] \to \mathbb{R}$ definiert durch

$$\mathcal{R}f \;=\; \sum_{k=0}^{n} \alpha_k f(x_k) \;+\; \beta \int_a^b f(x)\,dx, \qquad f \in C^{-1}[a,b], \tag{14.1}$$

betrachtet. Dabei sind $x_0, x_1, \ldots, x_n \in [a,b]$ paarweise verschiedene Stützstellen, und α_k und $\beta \in \mathbb{R}$ sind gegebene Koeffizienten. Weiter bezeichnet $C^{-1}[a,b]$ den Raum der stückweise stetigen Funktionen auf $[a,b]$. Es sei angenommen, dass das Funktional $\mathcal{R}$ für ein $r \geq 0$ auf dem Raum der Polynome vom Höchstgrad r verschwindet,

$$\mathcal{R}p \;=\; 0 \qquad \forall\, p \in \Pi_r.$$

Beispiel 14.1. Zu gegebenen Stützstellen $x_0, x_1, \ldots, x_n \in [a,b]$ hat das Restglied bei der Polynominterpolation für einen ausgewählten Punkt $\overline{x} \in [a,b]$ die folgende Gestalt,

$$\mathcal{R}f \;=\; \sum_{k=0}^{n} f(x_k)\, L_k(\overline{x}) \;-\; f(\overline{x}), \qquad f \in C^{-1}[a,b], \tag{14.2}$$

mit den lagrangeschen Basispolynomen $L_k(x) = \prod_{j=0, j\neq k}^{n} \frac{x-x_j}{x_k-x_j}$. Bekanntermaßen gilt hier $\mathcal{R}_{|\Pi_n} = 0$, und für hinreichend glatte Funktionen f gilt die folgende Fehlerdarstellung[1]:

$$\mathcal{R}f \;=\; \frac{\omega(\overline{x})\, f^{(n+1)}(\xi)}{(n+1)!}, \qquad f \in C^{n+1}[a,b],$$

mit $\omega(x) := (x-x_0)\cdots(x-x_n)$. △

Beispiel 14.2. Für eine gegebene interpolatorische Quadraturformel und für hinreichend glatte Funktionen f hat das Restglied die folgende Gestalt,

$$\mathcal{R}f \;=\; (b-a) \sum_{k=0}^{n} \sigma_k f(x_k) \;-\; \int_a^b f(x)\,dx, \qquad f \in C^{-1}[a,b].$$

[1] siehe (1.14)

Per Definition ist für Quadraturformeln ein Genauigkeitsgrad von mindestens r gleichbedeutend mit der Eigenschaft $\mathcal{R}_{|\Pi_r} = 0$, und für Funktionen $f \in C^{m+1}[a,b]$ mit $n \le m \le r$ sind bereits Fehlerabschätzungen bekannt[2]. Auch hier stellt sich die Frage nach Fehlerdarstellungen für weniger glatte Funktionen f. △

14.2 Peano-Kerne

Im weiteren Verlauf werden die folgenden Notationen verwendet:
(a)

$$(x-t)_+^m := \begin{cases} (x-t)^m, & x \ge t, \\ 0, & x < t, \end{cases} \quad \text{für } m \ge 1, \qquad (x-t)_+^0 := \begin{cases} 1, & x \ge t, \\ 0, & x < t; \end{cases}$$

(b) für eine Funktion $\psi : [a,b] \times [c,d] \to \mathbb{R}$ mit der Eigenschaft $\psi(\cdot,t) \in C^{-1}[a,b]$ für jedes $t \in [c,d]$ bezeichnet

$$\mathcal{R}_x(\psi(x,t)) \;=\; \mathcal{R}(\psi(\cdot,t)), \qquad t \in [c,d].$$

Das Argument von $\mathcal{R}_x$ ist also jeweils als Funktion von x aufzufassen.

Definition 14.3. Gegeben sei ein Funktional $\mathcal{R} : C^{-1}[a,b] \to \mathbb{R}$ der Gestalt (14.1), welches auf dem Raum Π_r verschwindet. Dann bezeichnet man die Funktionen

$$K_m(t) := \tfrac{1}{m!}\mathcal{R}_x((x-t)_+^m), \qquad t \in [a,b] \qquad (m = 0,1,\ldots,r)$$

als *Peano-Kerne*.

Das folgende Theorem liefert die zentrale Aussage des vorliegenden Abschnitts. Der zugehörige Beweis beruht auf einer Approximation der Funktion f durch Polynome vom Grad $\le r$, die mittels Taylorentwicklungen gewonnen werden.

Theorem 14.4. *Gegeben sei ein Funktional $\mathcal{R} : C^{-1}[a,b] \to \mathbb{R}$ der Gestalt (14.1), welches auf dem Raum Π_r verschwindet. Für jedes $0 \le m \le r$ gilt*

$$\mathcal{R}f \;=\; \int_a^b f^{(m+1)}(t) K_m(t)\,dt, \qquad f \in C^{m+1}[a,b].$$

Falls weiterhin $\mathcal{R}(x^{r+1}) \ne 0$ erfüllt ist und der Peano-Kern K_r sein Vorzeichen nicht wechselt, so gilt die Darstellung

$$\mathcal{R}f \;=\; \varkappa f^{(r+1)}(\xi), \qquad f \in C^{r+1}[a,b], \tag{14.3}$$

mit einer geeigneten Zwischenstelle $\xi = \xi(f) \in [a,b]$ und der Konstanten $\varkappa = \frac{\mathcal{R}(x^{r+1})}{(r+1)!}$.

[2] siehe Theorem 6.13

BEWEIS. Eine Taylorentwicklung der Funktion f in dem linken Randpunkt a mit Integraldarstellung des Restglieds liefert

$$f(x) = \overbrace{f(a) + f'(a)(x-a) + \ldots + \frac{f^{(m)}(a)}{m!}(x-a)^m}^{=:\, p_m(x) \in \Pi_m} + r_m(x), \quad x \in [a,b],$$

$$\text{mit } r_m(x) := \frac{1}{m!}\int_a^x f^{(m+1)}(t)(x-t)^m\,dt = \frac{1}{m!}\int_a^b f^{(m+1)}(t)(x-t)_+^m\,dt, \quad x \in [a,b].$$

Somit erschließt man

$$\begin{aligned}\mathcal{R}f &= \mathcal{R}(p_m + r_m) = \overbrace{\mathcal{R}p_m}^{=0} + \mathcal{R}r_m = \frac{1}{m!}\mathcal{R}_x\Big(\int_a^b f^{(m+1)}(t)(x-t)_+^m\,dt\Big)\\ &\overset{(*)}{=} \frac{1}{m!}\int_a^b f^{(m+1)}(t)\mathcal{R}_x((x-t)_+^m)\,dt = \int_a^b f^{(m+1)}(t)K_m(t)\,dt,\\ &\qquad\qquad f \in C^{m+1}[a,b],\end{aligned}$$

wobei sich die Identität $(*)$ wie folgt berechnet,

$$\begin{aligned}&\mathcal{R}_x\Big(\int_a^b f^{(m+1)}(t)(x-t)_+^m\,dt\Big)\\ &\quad= \sum_{k=0}^n \alpha_k \int_a^b f^{(m+1)}(t)(x_k-t)_+^m\,dt + \beta\int_a^b\int_a^b f^{(m+1)}(t)(x-t)_+^m\,dt\,dx\\ &\quad= \int_a^b f^{(m+1)}(t)\underbrace{\Big(\sum_{k=0}^n \alpha_k(x_k-t)_+^m + \beta\int_a^b (x-t)_+^m\,dx\Big)}_{\mathcal{R}_x((x-t)_+^m)}\,dt.\end{aligned}$$

Damit ist der erste Teil der Aussage des Theorems bewiesen. Wechselt nun der Peano-Kern K_r sein Vorzeichen nicht, so liefert eine Anwendung des Mittelwertsatzes der Integralrechnung

$$\mathcal{R}f = \underbrace{\Big(\int_a^b K_r(t)\,dt\Big)}_{=:\,\varkappa} f^{(r+1)}(\xi), \qquad f \in C^{r+1}[a,b], \tag{14.4}$$

mit einer geeigneten Zwischenstelle $\xi = \xi(f) \in [a,b]$. Eine Anwendung der Identität (14.4) auf das Monom x^{r+1} liefert schließlich die behauptete Darstellung für die Konstante $\varkappa$,

$$\mathcal{R}(x^{r+1}) = \varkappa(r+1)!,$$

womit auch die Darstellung (14.3) bewiesen ist. □

Bemerkung 14.5. Auch für allgemeine Fehlerfunktionale der Form

$$\mathcal{R}f = \sum_{k=0}^{n_0}\alpha_{0k}f(x_{0k}) + \sum_{k=0}^{n_1}\alpha_{1k}f'(x_{1k}) + \ldots + \sum_{k=0}^{n_s}\alpha_{sk}f^{(s)}(x_{sk}) + \beta\int_a^b f(x)\,dx \quad (14.5)$$

$$\text{für} \quad f \in C^{m+1}[a,b]$$

gelten für $m = s, s+1, \ldots, r$ die Darstellungen aus Theorem 14.4 mit dem Peano-Kern aus Definition 14.3 (noch allgemeiner dürften auch Terme mit gewichteten Integralen von Ableitungen der Funktion f auftreten). Man hat sich nur zu überlegen, dass die Identität $(*)$ im Beweis von Theorem 14.4 auch in dieser allgemeinen Situation ihre Gültigkeit behält. △

14.3 Anwendungen

14.3.1 Interpolation

Theorem 14.6. *Zu gegebenen Stützstellen $x_0, x_1, \ldots, x_n \in [a,b]$ besitzt bei der Polynominterpolation das Restglied für eine ausgewählte Stelle $\overline{x} \in [a,b]$ die folgende Darstellung*[3]

$$\mathcal{R}f = \frac{1}{m!}\sum_{k=0}^{n} L_k(\overline{x})\int_{\overline{x}}^{x_k} f^{(m+1)}(t)(x_k - t)^m\,dt, \quad f \in C^{m+1}[a,b] \qquad (0 \le m \le n),$$

mit den lagrangeschen Basispolynomen $L_k(x) = \prod_{\substack{j=0\\ j\neq k}}^{n} \frac{x - x_j}{x_k - x_j}$.

BEWEIS. Nach Definition gilt für den Peano-Kern K_m die folgende Darstellung,

$$K_m(t) = \frac{1}{m!}\Big[\sum_{k=0}^{n}(x_k - t)_+^m L_k(\overline{x}) - (\overline{x} - t)_+^m\Big],$$

und daher

$$\begin{aligned}
\mathcal{R}f &= \frac{1}{m!}\sum_{k=0}^{n}\Big(L_k(\overline{x})\int_a^{x_k} f^{(m+1)}(t)(x_k - t)^m\,dt\Big) - \int_a^{\overline{x}} f^{(m+1)}(t)(\overline{x} - t)\,dt \\
&= \frac{1}{m!}\int_a^{\overline{x}} f^{(m+1)}(t)\overbrace{\Big[\sum_{k=0}^{n} L_k(\overline{x})(x_k - t)^m - (\overline{x} - t)^m\Big]}^{=\,0}\,dt \\
&\qquad + \frac{1}{m!}\sum_{k=0}^{n}\Big(L_k(\overline{x})\int_{\overline{x}}^{x_k} f^{(m+1)}(t)(x_k - t)^m\,dt\Big),
\end{aligned}$$

was in der behaupteten Darstellung resultiert. □

[3] vergleiche (14.2)

14.3.2 Numerische Integration

Beispiel 14.7 (Numerische Integration, Simpson-Regel). Das Restglied der Simpson-Regel zur numerischen Integration auf dem Intervall $[-1,1]$ hat die folgende Gestalt,

$$\mathcal{R}f \;=\; \tfrac{1}{3}f(-1) \;+\; \tfrac{4}{3}f(0) \;+\; \tfrac{1}{3}f(1) \;-\; \int_{-1}^{1} f(x)\,dx, \qquad f \in C^{-1}[-1,1],$$

und bekanntermaßen[4] ist $r = 3$ der Genauigkeitsgrad der Simpson-Regel. Daher gilt für $t \geq 0$ (und $m = 3$)

$$\begin{aligned}
K_3(t) &= \tfrac{1}{6}\mathcal{R}_x((x-t)_+^3) \\
&= \tfrac{1}{6}\Big[\tfrac{1}{3}(-1-t)_+^3 \;+\; \tfrac{4}{3}(0-t)_+^3 \;+\; \tfrac{1}{3}(1-t)_+^3 \;-\; \int_{-1}^{1}(x-t)_+^3\,dx\Big] \\
&= \tfrac{1}{6}\Big[\tfrac{1}{3}\cdot 0 \;-\; \tfrac{4}{3}\cdot 0 \;+\; \tfrac{1}{3}(1-t)^3 \;-\; \int_{t}^{1}(x-t)^3\,dx\Big] \\
&= \tfrac{1}{6}\big[\tfrac{1}{3}(1-t)^3 \;-\; \tfrac{1}{4}(1-t)^4\big] \;=\; \tfrac{1}{72}(1-t)^3(1+3t) \;\geq\; 0 \quad \text{für } t \in [0,1].
\end{aligned}$$

Weiter gilt nach Aufgabe 14.2 die folgende Identität,

$$K_3(-t) \;=\; K_3(t), \qquad t \in [0,1],$$

so dass der Peano-Kern K_3 auf dem Intervall $[-1,1]$ von einem Vorzeichen ist, $K_3(t) \geq 0$ für $t \in [-1,1]$. Also ist (14.3) anwendbar, und wegen

$$\frac{\mathcal{R}(x^4)}{4!} \;=\; \frac{1}{24}\Big(\tfrac{1}{3} \;+\; \tfrac{4}{3}\cdot 0 \;+\; \tfrac{1}{3} \;-\; \int_{-1}^{1} x^4\,dx\Big) \;=\; \frac{1}{90}$$

erhält man so die schon bekannte Fehlerdarstellung

$$\begin{aligned}
\tfrac{1}{3}f(-1) \;+\; \tfrac{4}{3}f(0) \;+\; \tfrac{1}{3}f(1) \;-\; \int_{-1}^{1} f(t)\,dt \;&=\; \frac{1}{90}f^{(4)}(\xi) \\
\text{für} \quad f \in C^4[-1,1], \qquad \xi \;&=\; \xi(f) \;\in\; [-1,1].
\end{aligned}$$

△

Weitere Themen und Literaturhinweise

Weitergehende Betrachtungen zur peanoschen Restglieddarstellung werden zum Beispiel in Hämmerlin/Hoffmann [48] und in Schaback/Wendland [92] angestellt.

Übungsaufgaben

Aufgabe 14.1. Man zeige, dass für allgemeine Fehlerfunktionale der Form (14.5) die Darstellung aus Theorem 14.4 mit dem Peano-Kern aus Definition 14.3 für Werte $m = s, s+1, \ldots, r$ ihre Gültigkeit behält.

[4] siehe Theorem 6.16

Aufgabe 14.2. Gegeben sei ein Funktional $\mathcal{R} : C^{-1}[a,b] \to \mathbb{R}$ der Gestalt (14.1), welches auf dem Raum Π_r verschwindet, und m sei eine ungerade Zahl mit $1 \le m \le r$. Man zeige: falls

$$\mathcal{R}f \;=\; \mathcal{R}\widehat{f} \qquad \text{für} \quad f \in C^{m+1}[a,b]$$
$$\text{mit} \quad \widehat{f}\big(\tfrac{a+b}{2}+x\big) \;:=\; f\big(\tfrac{a+b}{2}-x\big), \quad x \in [-\tfrac{b-a}{2}, \tfrac{b-a}{2}]$$

erfüllt ist, so ist der Peano-Kern K_m symmetrisch bezüglich des Intervallmittelpunkts, das heißt,

$$K_m\big(\tfrac{a+b}{2} + x\big) \;=\; K_m\big(\tfrac{a+b}{2} - x\big), \qquad x \in [0, \tfrac{b-a}{2}].$$

Aufgabe 14.3. Im Folgenden betrachte man die Quadraturformel $Qf := \int_{-1}^{1} P(x)\,dx$ zur näherungsweisen Berechnung des Integrals $\int_{-1}^{1} f(x)\,dx$, wobei für $f \in C^1[-1,1]$ das Polynom $P \in \Pi_5$ die Lösung der folgenden hermiteschen Interpolationsaufgabe bezeichnet,

$$P(x_j) \;=\; f(x_j), \qquad P'(x_j) \;=\; f'(x_j) \qquad \text{für} \quad j \;=\; 0,1,2,$$

mit $x_0 \;=\; -1,\; x_1 \;=\; 0$ und $x_2 \;=\; 1$.

(a) Man zeige

$$Qf \;=\; \tfrac{7}{15}f(-1) \;+\; \tfrac{1}{15}f'(-1) \;+\; \tfrac{16}{15}f(0) \;+\; \tfrac{7}{15}f(1) - \tfrac{1}{15}f'(1).$$

(b) Zeige: die Quadraturformel Q besitzt den Genauigkeitsgrad 5.

(c) Man berechne für $n = 5$ den Peano-Kern K_5 zu der Quadraturformel Q und zeige, dass dieser sein Vorzeichen nicht wechselt.

(d) Man bestimme unter Verwendung von (c) eine Fehlerdarstellung für die betrachtete Quadraturformel.

15 Approximationstheorie

15.1 Einführende Bemerkungen

Eine wichtige Fragestellung der numerischen Mathematik ist es, bezüglich einer festgelegten Norm für eine gegebene Funktion eine Bestapproximation aus einer Menge von Funktionen zu bestimmen sowie den auftretenden Fehler abzuschätzen. Vergleichbare Fragestellungen treten auch für Vektoren anstelle von Funktionen auf.

Beispiel 15.1. Die Frage der optimalen Wahl der Stützstellen bei der Polynominterpolation führt auf das Minimaxproblem[1]

$$\max_{x\in[a,b]} |(x-x_0)\dots(x-x_n)| \;\to\; \min \qquad \text{für} \quad x_0,x_1,\dots,x_n \;\in\; [a,b]. \tag{15.1}$$

Die Gesamtheit aller Funktionen von der Form $(x-x_0)\dots(x-x_n)$ stimmt überein mit dem Raum der Polynome vom Grad $n+1$ mit führendem Koeffizienten eins, so dass das Minimierungsproblem (15.1) äquivalent zu dem folgenden Approximationsproblem ist:

$$\|x^{n+1}-p\|_\infty \;=\; \max_{x\in[a,b]} |x^{n+1}-p(x)| \;\to\; \min \qquad \text{für} \quad p \;\in\; \Pi_n. \qquad \triangle$$

Beispiel 15.2. Lineare Ausgleichsprobleme besitzen die Form[2]

$$\|Ax-b\|_2 \;\to\; \min \qquad \text{für} \quad x\in\mathbb{R}^N,$$

mit gegebener Matrix $A \in \mathbb{R}^{M\times N}$ und gegebenem Vektor $b \in \mathbb{R}^M$. Diese können ebenfalls als Approximationsprobleme aufgefasst werden, bei dem aus der Menge $\{Ax : x \in \mathbb{R}^N\}$ eine Bestapproximation an den Vektor b (und anschließend ein Urbild unter A) zu bestimmen ist. $\triangle$

In dem vorliegenden Abschnitt wird in Grundzügen eine allgemeine Theorie über Bestapproximationen – im Folgenden kurz als Proxima bezeichnet – vorgestellt.

Definition 15.3. Für eine Teilmenge $\varnothing \neq \mathcal{M} \subset \mathcal{V}$ eines normierten Raums $(\mathcal{V}, \|\cdot\|)$ und ein gegebenes Element $v \in \mathcal{V}$ heißt u^* ein *$\mathcal{M}$-Proximum* an v, falls

$$u^* \;\in\; \mathcal{M}, \qquad \|u^*-v\| \;=\; \underbrace{\inf_{u\in\mathcal{M}} \|u-v\|}_{=:\,\mathbf{E}_v(\mathcal{M})}.$$

Die Zahl $\mathbf{E}_v(\mathcal{M})$ bezeichnet man als *Minimalabstand* des Elements v von der Teilmenge $\mathcal{M}$.

[1] Dieses Problem ist erstmalig in Abschnitt 1.6 behandelt worden unter gleichzeitiger Angabe einer Lösung.

[2] siehe hierzu Abschnitt 4.8.5 für eine erstmalige Behandlung, wo zugleich Lösungsvorschläge zu finden sind

Bemerkung 15.4. (a) Natürliche Fragestellungen in diesem Zusammenhang sind Existenz und Eindeutigkeit eines Proximums u^* sowie die Angabe von Verfahren zur Bestimmung von u^* und eventuell noch die Herleitung von Abschätzungen für den Minimalabstand.

(b) Das in Definition 15.3 beschriebene Problem ist ein Optimierungsproblem von der Form

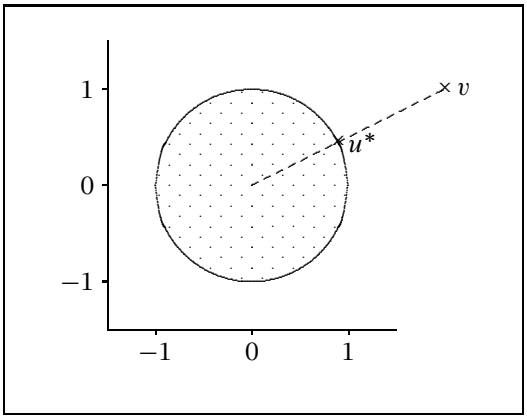
$$f(u) \to \min \qquad \text{für } u \in \mathcal{M} \subset \mathcal{V}, \tag{15.2}$$

mit dem speziellen Zielfunktional $f(u) = \|u-v\|$. Allgemeine Probleme von der Form (15.2) sind Gegenstand der *nichtlinearen Optimierung*, die ein weites Feld darstellt und hier nicht weiter verfolgt wird. Literaturhinweise zu diesem Thema finden Sie auf Seite 394. △

15.2 Existenz eines Proximums

In dem vorliegenden Abschnitt soll – im Anschluss an die Vorstellung zweier Beispiele – in einem allgemeinen Kontext die Frage der Existenz eines Proximums behandelt werden.

Beispiel 15.5. Man betrachte die folgende spezielle Situation:



$$\begin{aligned} \mathcal{V} &= \mathbb{R}^2, \\ \|v\| &= \|v\|_2 = (v_1^2 + v_2^2)^{1/2}, \\ \mathcal{M} &= \{x \in \mathbb{R}^2 : \|x\|_2 \le 1\}. \end{aligned}$$

Für den Vektor $v = (2,1)^\top$ ist $u^* = (2/\sqrt{5}, 1/\sqrt{5})^\top$ ein $\mathcal{M}$-Proximum (das hier zudem eindeutig bestimmt ist) an den Vektor v. △

Beispiel 15.6. Man betrachte nun die folgende Situation:

$$\begin{aligned} \mathcal{V} &= C[0,1], \qquad \|v\| = \|v\|_\infty = \max_{t\in[0,1]} |v(t)|, \\ \mathcal{M} &= \{e^{\beta t} : \beta > 0\}, \end{aligned}$$

und sei $v \equiv \frac{1}{2}$. Es ist $\|e^{\beta t} - v\|_\infty = e^\beta - \frac{1}{2} > \frac{1}{2}$ für $\beta > 0$, so dass $\mathbf{E}_v(\mathcal{M}) = \frac{1}{2}$ gilt und kein $\mathcal{M}$-Proximum an v existiert. △

Die folgende Definition und das nachfolgende Lemma dienen der Herleitung einer ersten Existenzaussage für Proxima.

Definition 15.7. Für eine Teilmenge $\varnothing \neq \mathcal{M} \subset \mathcal{V}$ eines normierten Raums $(\mathcal{V}, \|\cdot\|)$ und ein gegebenes Element $v \in \mathcal{V}$ heißt $(u_k)_{k\in\mathbb{N}}$ eine *$\mathcal{M}$-Minimalfolge* an v, wenn

$$(u_k)_{k\in\mathbb{N}} \subset \mathcal{M}, \qquad \|u_k - v\| \to \mathbf{E}_v(\mathcal{M}) \quad \text{für} \quad k \to \infty. \tag{15.3}$$

Lemma 15.8. *Für eine Teilmenge $\varnothing \neq \mathcal{M} \subset \mathcal{V}$ eines normierten Raums $(\mathcal{V}, \|\cdot\|)$ und ein gegebenes Element $v \in \mathcal{V}$ sei $(u_k)_{k\in\mathbb{N}}$ eine $\mathcal{M}$-Minimalfolge an v, die in $\mathcal{M}$ einen Häufungspunkt u^* besitze,*

$$u^* \in \mathcal{M}, \qquad \|u_k - u^*\| \to 0 \quad \textit{für} \quad \mathbb{N}_1 \ni k \to \infty \quad (\mathbb{N}_1 \subset \mathbb{N} \textit{ geeignet}). \tag{15.4}$$

Dann ist u^* ein $\mathcal{M}$-Proximum an v.

BEWEIS. Es gilt

$$\|u^* - v\| \leq \overbrace{\|u^* - u_k\|}^{\to 0 \text{ für } \mathbb{N}_1 \ni k \to \infty} + \overbrace{\|u_k - v\|}^{\to \mathbf{E}_v(\mathcal{M}) \text{ für } k\to\infty}$$

und infolgedessen notwendigerweise $\|u^* - v\| \leq \mathbf{E}_v(\mathcal{M})$. □

Als unmittelbare Konsequenz aus dem vorangegangenen Lemma erhält man das folgende Resultat.

Theorem 15.9. *Ist $\varnothing \neq \mathcal{M} \subset \mathcal{V}$ eine kompakte Teilmenge des normierten Raums $(\mathcal{V}, \|\cdot\|)$, so existiert zu jedem Vektor $v \in \mathcal{V}$ ein $\mathcal{M}$-Proximum an v.*

Korollar 15.10. *Ist $\mathcal{U} \subset \mathcal{V}$ ein endlich-dimensionaler linearer Unterraum des normierten Raums $(\mathcal{V}, \|\cdot\|)$, so existiert zu jedem Vektor $v \in \mathcal{V}$ ein $\mathcal{U}$-Proximum an v.*

BEWEIS. Die Menge

$$\mathcal{M} := \{u \in \mathcal{U} \,:\, \|u - v\| \leq \mathbf{E}_v(\mathcal{U}) + 1\} \subset \mathcal{U}$$

ist offensichtlich nichtleer und kompakt, nach Theorem 15.9 existiert also ein $\mathcal{M}$-Proximum u^* an v. Wegen

$$\|u^* - v\| = \inf_{u\in\mathcal{M}} \|u - v\| \leq \sup_{u\in\mathcal{M}} \|u - v\| \leq \mathbf{E}_v(\mathcal{U}) + 1 \leq \inf_{u\in\mathcal{U}\setminus\mathcal{M}} \|u - v\|$$

gilt dann notwendigerweise $\|u^* - v\| = \inf_{u\in\mathcal{U}} \|u - v\| = \mathbf{E}_v(\mathcal{U})$. □

Zusammenfassend kann man festhalten, dass sowohl in kompakten Teilmengen von normierten Räumen als auch in endlich-dimensionalen linearen Unterräumen von normierten Räumen die Existenz eines Proximums gewährleistet ist.

15.3 Eindeutigkeit eines Proximums

In den beiden folgenden Unterabschnitten 15.3.1 und 15.3.2 werden in einem allgemeinen Rahmen jeweils ein hinreichendes Kriterium für die Eindeutigkeit eines Proximums hergeleitet.

15.3.1 Einige Notationen; streng konvexe Mengen

Definition 15.11. Sei $(\mathcal{V}, \|\cdot\|)$ ein normierter Raum.

(a) Für $x \in \mathcal{V}$ und $r > 0$ ist die *abgeschlossene Kugel um x mit Radius r* gegeben durch

$$\mathcal{B}(x;r) \;=\; \{y \in \mathcal{V} \;:\; \|y-x\| \;\le\; r\}.$$

(b) Für eine Teilmenge $\mathcal{M} \subset \mathcal{V}$ bezeichnet

$$\mathcal{M}^\circ \;=\; \{x \;\in\; \mathcal{M} \;:\; \text{es existiert ein } \varepsilon > 0 \text{ mit } \mathcal{B}(x;\varepsilon) \;\subset\; \mathcal{M}\}$$

den offenen Kern von $\mathcal{M}$. Es heißt $\mathcal{M}$ *offen*, falls $\mathcal{M}^\circ = \mathcal{M}$ gilt. Schließlich heißt $\mathcal{M}$ *abgeschlossen*, falls $\mathcal{V}\backslash\mathcal{M}$ eine in $\mathcal{V}$ offene Menge ist.

Beispiel 15.12. In einem normierten Raum $(\mathcal{V}, \|\cdot\|)$ ist $\mathcal{B}(x;r)$ eine abgeschlossene Teilmenge und es gilt $\mathcal{B}(x;r)^\circ = \{y \in \mathcal{V} \;:\; \|y-x\| < r\}$. △

Definition 15.13. Eine Teilmenge $\mathcal{M} \subset \mathcal{V}$ des normierten Raums $(\mathcal{V}, \|\cdot\|)$ heißt *konvex*, falls für je zwei Elemente $x,\ y \in \mathcal{M}$ auch die Verbindungsstrecke von x nach y zu $\mathcal{M}$ gehört, das heißt,

$$\{x \;+\; \lambda(y-x) \;:\quad 0 \le \lambda \le 1\} \;\subset\; \mathcal{M}, \qquad x,\ y \in \mathcal{M}.$$

Es heißt $\mathcal{M}$ *streng konvex*, falls zu je zwei verschiedenen Punkten deren Verbindungsstrecke ohne die Endpunkte selbst zum offenen Kern von $\mathcal{M}$ gehört, das heißt,

$$\{x \;+\; \lambda(y-x) \;:\quad 0 < \lambda < 1\} \;\subset\; \mathcal{M}^\circ, \qquad x,\ y \in \mathcal{M},\quad x \ne y.$$

Offensichtlich ist eine streng konvexe Menge auch konvex.

Lemma 15.14. *Ist $\varnothing \ne \mathcal{M} \subset \mathcal{V}$ eine konvexe Teilmenge des normierten Raums $(\mathcal{V}, \|\cdot\|)$, so ist für jedes $v \in \mathcal{V}$ die Menge der $\mathcal{M}$–Proxima an v konvex.*

BEWEIS. Für zwei $\mathcal{M}$–Proxima u_1 und u_2 an v sowie jede Zahl $\lambda \in [0,1]$ gilt

$$\begin{aligned}\|(1-\lambda)u_1 \;+\; \lambda u_2 \;-\; v\| \;&\le\; (1-\lambda)\|u_1 - v\| \;+\; \lambda\|u_2 - v\| \\ &\le\; (1-\lambda)\mathbf{E}_v(\mathcal{M}) \;+\; \lambda\mathbf{E}_v(\mathcal{M}) \;=\; \mathbf{E}_v(\mathcal{M}).\end{aligned}$$

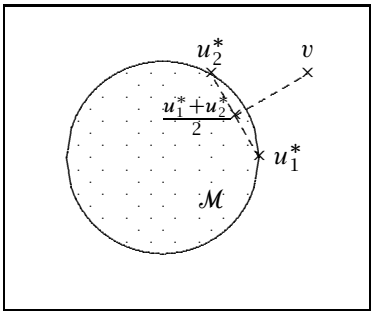
□

Die streng konvexen Mengen liefern die erste Klasse von Mengen, in denen Proxima eindeutig sind:

Proposition 15.15. *Ist $\varnothing \neq \mathcal{M} \subset \mathcal{V}$ eine streng konvexe Teilmenge des normierten Raums $(\mathcal{V}, \|\cdot\|)$, so existiert zu jedem Element $v \in \mathcal{V}$ höchstens ein $\mathcal{M}$-Proximum an v.*

BEWEIS. Seien u_1^* und u_2^* $\mathcal{M}$-Proxima an $v \in \mathcal{V}\backslash\mathcal{M}$ (im Fall $v \in \mathcal{M}$ ist die Situation klar), und nach Lemma 15.14 ist dann auch $\frac{1}{2}(u_1^* + u_2^*)$ ein $\mathcal{M}$-Proximum. Wenn nun $u_1^* \neq u_2^*$ gilt, so ist $\frac{1}{2}(u_1^* + u_2^*) \in \mathcal{M}^\circ$, und dann liegt für eine hinreichend klein gewählte Zahl $0 < \lambda < 1$ die folgende Situation vor,

$$\begin{aligned} u_\lambda &:= (1-\lambda)\frac{u_1^*+u_2^*}{2} + \lambda v \quad \in \quad \mathcal{M}, \\ \|u_\lambda - v\| &= (1-\lambda)\|\frac{u_1^*+u_2^*}{2} - v\| \\ &= (1-\lambda)\mathbf{E}_v(\mathcal{M}) \;<\; \mathbf{E}_v(\mathcal{M}), \end{aligned}$$



und daraus resultiert ein Widerspruch. □

15.3.2 Strikt normierte Räume

Definition 15.16. Der normierte Raum $(\mathcal{V}, \|\cdot\|)$ heißt *strikt normiert*, falls die abgeschlossene Einheitskugel $\mathcal{B}(0;1)$ streng konvex ist.

Strikte Normiertheit bedeutet also Folgendes:

$$\|x\| = \|y\| = 1, \quad x \neq y \quad \Longrightarrow \quad \|(1-\lambda)x + \lambda y\| < 1 \quad \text{für } 0 < \lambda < 1 \quad (x, y \in \mathcal{V}). \tag{15.5}$$

Ein normierter Raum $(\mathcal{V}, \|\cdot\|)$ ist demnach genau dann strikt normiert, wenn die Einheitssphäre $\{x \in \mathcal{V} : \|x\| = 1\}$ keine Strecken enthält. Vier spezielle Situationen sind in den folgenden Bildern 15.1–15.4 dargestellt.

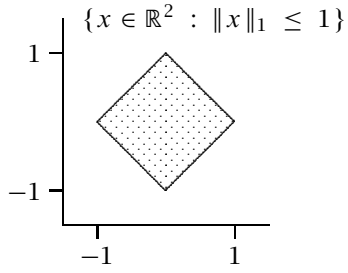


Bild 15.1: $(\mathbb{R}^2, \|\cdot\|_1)$, nicht strikt normiert

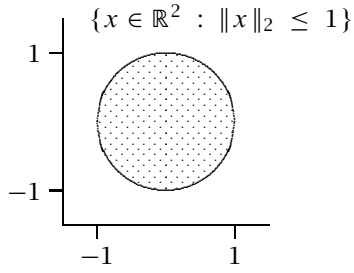


Bild 15.2: $(\mathbb{R}^2, \|\cdot\|_2)$, strikt normiert

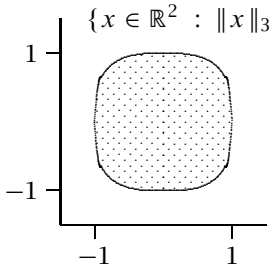


Bild 15.3: $(\mathbb{R}^2, \|\cdot\|_3)$, strikt normiert

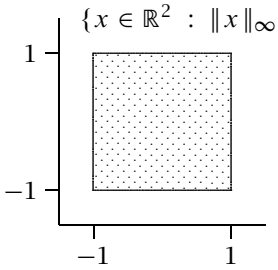


Bild 15.4: $(\mathbb{R}^2, \|\cdot\|_\infty)$, nicht strikt normiert

Bemerkung 15.17. Im Grunde genügt es, in (15.5) die Betrachtungen auf den Fall $\lambda = 1/2$ zu beschränken, was etwa für beliebige $0 < \lambda \leq 1/2$ aus der Darstellung $(1-\lambda)x + \lambda y = \frac{1}{2}(x + x + 2\lambda(y-x))$ resultiert. △

Nur in strikt normierten Räumen stellt zu je zwei linear unabhängigen Elementen die Dreiecksungleichung eine echte Ungleichung dar, wie sich mit dem folgenden Theorem herausstellt:

Theorem 15.18. *Ein normierter Raum* $(\mathcal{V}, \|\cdot\|)$ *ist genau dann strikt normiert, wenn für beliebige Elemente* $0 \neq x \in \mathcal{V}$ *und* $0 \neq y \in \mathcal{V}$ *Folgendes gilt,*

$$\|x\| + \|y\| = \|x+y\| \quad \Longrightarrow \quad \exists\, \alpha \in \mathbb{K} \ \textit{mit}\ y = \alpha x. \tag{15.6}$$

BEWEIS. "$\Longrightarrow$" Sei $\mathcal{V}$ strikt normiert, und für beliebige Elemente $0 \neq x \in \mathcal{V}$ und $0 \neq y \in \mathcal{V}$ gelte $\|x\| + \|y\| = \|x+y\|$. O.B.d.A. sei noch $\|x\| \leq \|y\|$, und dann berechnet man

$$\begin{aligned} \Big\| \frac{x}{\|x\|} + \frac{y}{\|y\|} \Big\| &\geq \Big\| \frac{x}{\|x\|} + \frac{y}{\|x\|} \Big\| - \Big\| \frac{y}{\|x\|} - \frac{y}{\|y\|} \Big\| \\ &= \frac{\|x\| + \|y\|}{\|x\|} - \|y\| \Big(\frac{1}{\|x\|} - \frac{1}{\|y\|} \Big) = 2, \end{aligned}$$

und die strikte Normiertheit impliziert

$$\frac{x}{\|x\|} = \frac{y}{\|y\|}, \qquad \rightsquigarrow \qquad y = \alpha x \quad \text{mit} \quad \alpha = \frac{\|y\|}{\|x\|}.$$

"$\Longleftarrow$" Umgekehrt sei nun die Eigenschaft (15.6) erfüllt, und für beliebige Vektoren $x,\ y \in \mathcal{V}$ mit $\|x\| = \|y\| = 1$ sowie $x \neq y$ und $\lambda \in (0,1)$ ist

$$\|\lambda x + (1-\lambda)y\| < 1 \tag{15.7}$$

nachzuweisen. Im Fall $\|x\| < 1$ oder $\|y\| < 1$ folgt (15.7) unmittelbar mit der Dreiecksungleichung, und im Folgenden gelte also $\|x\| = \|y\| = 1$. Falls nun im Widerspruch zur Ungleichung (15.7) die Identität $\|\lambda x + (1-\lambda)y\| \overset{(*)}{=} 1$ erfüllt ist, so gilt also

$$\|\lambda x\| + \|(1-\lambda)y\| = \|\lambda x + (1-\lambda)y\| = 1,$$

und aufgrund von (15.6) existiert ein $\alpha \in \mathbb{K}$ mit $\lambda x = \alpha(1-\lambda)y$, beziehungsweise

$$x \;=\; \beta y \qquad \text{für ein } \beta \in \mathbb{K}.$$

Im Folgenden wird $\beta = 1$ nachgewiesen, was einen Widerspruch zu $x \neq y$ liefert. Wegen $\|x\| = \|y\| = 1$ gilt $|\beta| = 1$, und weiter berechnet man noch

$$\begin{aligned} 0 \;&\overset{(*)}{=}\; |\lambda\beta + (1-\lambda)|^2 - 1 \;=\; |\lambda\beta|^2 \;+\; 2\lambda(1-\lambda)\mathrm{Re}\,\beta \;+\; |1-\lambda|^2 - 1 \\ &=\; 2\lambda(1-\lambda)(\mathrm{Re}\,\beta - 1) \end{aligned}$$

und daher $\mathrm{Re}\,\beta = 1$, was zusammen mit $|\beta| = 1$ den behaupteten Widerspruch $\beta = 1$ liefert. □

Das folgende Theorem liefert eine weitere Klasse von Mengen, in denen Proxima eindeutig sind.

Theorem 15.19. *Ist $\mathcal{U} \subset \mathcal{V}$ ein linearer Unterraum eines strikt normierten Raums $(\mathcal{V}, \|\cdot\|)$, so existiert zu jedem Element $v \in \mathcal{V}$ höchstens ein $\mathcal{U}$-Proximum an v.*

BEWEIS. Sind u_1^* und u_2^* zwei verschiedene $\mathcal{U}$-Proxima an $v \in \mathcal{V}$, so gilt aufgrund der strikten Normiertheit

$$\Big\| \underbrace{\tfrac{u_1^* + u_2^*}{2}}_{\in\,\mathcal{U}} - v \Big\| \;<\; \mathbf{E}_v(\mathcal{U}),$$

und daraus resultiert ein Widerspruch. □

Zusammenfassend kann man festhalten, dass sowohl in streng konvexen Teilmengen als auch in strikt normiert linearen Unterräumen die Eindeutigkeit eines Proximums gewährleistet ist.

15.4 Approximationstheorie in Räumen mit Skalarprodukt

15.4.1 Einige Grundlagen

Relativ günstige Verhältnisse in Bezug auf Eindeutigkeit, Existenz und Berechnung von Proxima liegen für solche Normen vor, die durch ein Skalarprodukt induziert werden. Einige grundlegende Resultate hierzu werden im Folgenden vorgestellt, wobei als Erstes die Eigenschaften eines Skalarprodukts in Erinnerung gerufen werden. Der Einfachheit halber werden wieder reelle Vektorräume zugrunde gelegt.

Definition 15.20. Ein *Skalarprodukt* auf einem Vektorraum $\mathcal{V}$ ist eine Abbildung $\langle\cdot,\cdot\rangle : \mathcal{V} \times \mathcal{V} \to \mathbb{R}$ mit folgenden Eigenschaften:

- für jeden Vektor $x \in \mathcal{V}$ sind die Abbildungen $\langle x,\cdot\rangle$ und $\langle\cdot, x\rangle : \mathcal{V} \to \mathbb{R}$ linear (Bilinearität);
- es gilt $\langle x, x\rangle > 0$ für jeden Vektor $0 \neq x \in \mathcal{V}$ (Definitheit);

- für beliebige Vektoren $x,\ y \in \mathcal{V}$ gilt $\langle x, y \rangle = \langle y, x \rangle$ (Symmetrie).

Ein Skalarprodukt bezeichnet man auch als *inneres Produkt.*

Theorem 15.21. *Ein* Skalarprodukt *auf einem reellem Vektorraum $\mathcal{V}$ induziert eine Norm mittels $\|x\| = \langle x, x \rangle^{1/2}$ für $x \in \mathcal{V}$.*

BEWEIS. Positive Definitheit und Homogenität der Norm sind jeweils unmittelbare Folgerungen aus der Definitheit und der Bilinearität des Skalarprodukts. Die Dreiecksungleichung für die Norm resultiert aus der *cauchy-schwarzschen Ungleichung*

$$|\langle x, y \rangle| \ \leq \ \|x\| \|y\|, \qquad x,\ y \in \mathcal{V}, \tag{15.8}$$

wobei in (15.8) Gleichheit genau dann vorliegt, wenn x und y linear abhängig sind. Einen Beweis für (15.8) finden Sie etwa in Fischer [28]. □

Beispiel 15.22. (a) Das klassische euklidische Skalarprodukt auf $\mathbb{R}^N$ ist gegeben durch $\langle x,\ y \rangle_2 = x^\top y$ für $x,\ y \in \mathbb{R}^N$.

(b) Für eine symmetrische, positiv definite Matrix $A \in \mathbb{R}^{N \times N}$ ist durch $\langle x\ ,\ y \rangle_A = x^\top A y$ für $x,\ y \in \mathbb{R}^N$ ein Skalarprodukt auf $\mathbb{R}^N$ definiert, welches im Zusammenhang mit dem Verfahren der konjugierten Gradienten[3] von Bedeutung ist.

(c) Zu gegebener Gewichtsfunktion $\varrho : [a, b] \to (0, \infty]$ stellt

$$\langle p, q \rangle \ := \ \int_a^b p(x) q(x) \varrho(x)\, dx, \qquad p,\ q \in \Pi,$$

ein Skalarprodukt auf dem Raum aller reellen Polynome Π dar.[4] △

Wichtige und elementare Identitäten in diesem Zusammenhang sind

$$\|x + y\|^2 \ = \ \|x\|^2 + 2\langle x, y \rangle + \|y\|^2, \quad x,\ y \in \mathcal{V}, \tag{15.9}$$

$$\|x + y\|^2 + \|x - y\|^2 \ = \ 2(\|x\|^2 + \|y\|^2), \qquad \text{—— « ——}, \tag{15.10}$$

wobei (15.10) als *Parallelogrammgleichung* bezeichnet wird. Als eine Folgerung aus dieser Identität erhält man die – für die Eindeutigkeit des Proximums in linearen Unterräumen relevante – strikte Normiertheit:

Theorem 15.23. *Ein Vektorraum mit einer durch ein Skalarprodukt induzierten Norm ist strikt normiert.*

BEWEIS. Die Aussage folgt unmittelbar aus der Parallelogrammgleichung (15.10) sowie aus der Eigenschaft (15.5) und Bemerkung 15.17. □

[3] siehe Abschnitt 11

[4] Solche Skalarprodukte treten im Abschnitt 6.8 über die Gaußquadratur auf.

15.4.2 Proxima in linearen Unterräumen

Im Folgenden spielen orthogonale Komplemente von Mengen $\mathcal{M} \subset \mathcal{V}$ eine Rolle,

$$\mathcal{M}^\perp := \{ y \in \mathcal{V} : \langle x, y \rangle = 0 \quad \text{für jedes } x \in \mathcal{M} \}.$$

Mit dem folgenden Theorem wird eine Charakterisierung für Proxima aus linearen Unterräumen vorgestellt.

Theorem 15.24. *Sei $\mathcal{U} \subset \mathcal{V}$ ein linearer Unterraum $\mathcal{U} \subset \mathcal{V}$ eines Vektorraums $\mathcal{V}$ mit innerem Produkt $\langle \cdot, \cdot \rangle$. Es ist ein Element $u^* \in \mathcal{U}$ genau dann ein $\mathcal{U}$-Proximum an einen gegebenes $v \in \mathcal{V}$, wenn $u^* - v \in \mathcal{U}^\perp$ gilt.*

BEWEIS. "$\Longleftarrow$" Im Fall $u^* - v \in \mathcal{U}^\perp$ berechnet man für ein beliebiges Element $u \in U$ mithilfe der Identität (15.9) Folgendes,

$$\begin{aligned} \|u - v\|^2 &= \|u^* - v + u - u^*\|^2 \\ &= \|u^* - v\|^2 + 2 \underbrace{\langle \overbrace{u^* - v}^{\in \mathcal{U}^\perp}, \overbrace{u - u^*}^{\in \mathcal{U}} \rangle}_{=0} + \underbrace{\|u - u^*\|^2}_{\geq 0} \geq \|u^* - v\|^2, \end{aligned}$$

so dass u^* ein $\mathcal{U}$-Proximum an den Vektor v darstellt.

"$\Longrightarrow$" Im Fall $u^* - v \notin \mathcal{U}^\perp$ existiert nach Definition ein Element $\psi \in \mathcal{U}$ mit $\langle u^* - v, \psi \rangle \neq 0$, o.B.d.A. sei $\langle u^* - v, \psi \rangle < 0$ erfüllt[5]. In dieser Situation erhält man für hinreichend kleine Zahlen $0 < t \ll 1$ Folgendes,

$$\|u^* + t\psi - v\|^2 = \|u^* - v\|^2 + \underbrace{2t\langle u^* - v, \psi \rangle + t^2 \|\psi\|^2}_{< 0 \text{ für } 0 < t \ll 1} < \|u^* - v\|^2,$$

so dass u^* kein $\mathcal{U}$-Proximum an den Vektor v darstellt. Dies komplettiert den Beweis des Theorems. □

Für die Situation $\mathcal{V} = \mathbb{R}^3$ ist die Aussage von Theorem 15.24 in Bild 15.5 veranschaulicht.

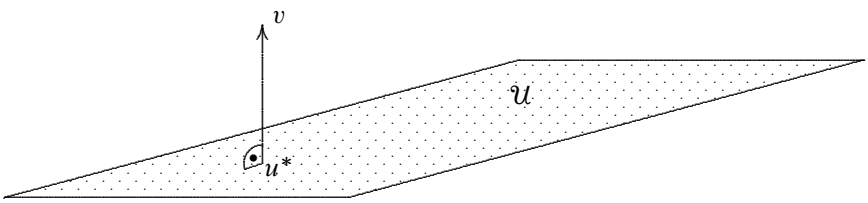


Bild 15.5: Darstellung der Aussage von Theorem 15.24 für $\mathcal{V} = \mathbb{R}^3$ und einen Unterraum $\mathcal{U}$ mit dim $\mathcal{U} = 2$

Mit dem folgenden Theorem wird für *endlich-dimensionale* lineare Unterräume mit gegebener Basis eine Methode zur Bestimmung des Proximums geliefert.

[5] andernfalls geht man von ψ zu $-\psi$ über

Theorem 15.25. *In einem Vektorraum $\mathcal{V}$ mit innerem Produkt $\langle \cdot, \cdot \rangle$ sei $\mathcal{U} \subset \mathcal{V}$ ein endlich-dimensionaler linearer Unterraum mit gegebener Basis $u_1, \ldots, u_m$, und es sei $u^* \in U$. Mit dem Ansatz $u^* = \sum_{k=1}^m \alpha_k u_k$ ist u^* genau dann ein $\mathcal{U}$-Proximum an ein gegebenes Element $v \in \mathcal{V}$, wenn die Koeffizienten $\alpha_1, \ldots, \alpha_m$ dem folgenden linearen Gleichungssystem genügen,*

$$\sum_{k=1}^{m} \langle u_k, u_j \rangle \alpha_k \;=\; \langle v, u_j \rangle_2, \qquad j \;=\; 1, 2, \ldots, m. \tag{15.11}$$

BEWEIS. Man hat nur zu berücksichtigen, dass für einen beliebigen Vektor $w \in \mathcal{V}$ die folgende Äquivalenz richtig ist[6]:

$$w \;\in\; \mathcal{U}^\perp \qquad \Longleftrightarrow \qquad \langle w, u_j \rangle \;=\; 0 \qquad \text{für} \quad j \;=\; 1, 2, \ldots, m. \qquad \square$$

Bemerkung 15.26. (a) Die im Zusammenhang mit Theorem 15.25 auftretende Matrix

$$\underbrace{\begin{pmatrix} \langle u_1, u_1 \rangle & \ldots & \langle u_m, u_1 \rangle \\ \vdots & \ddots & \vdots \\ \langle u_1, u_m \rangle & \ldots & \langle u_m, u_m \rangle \end{pmatrix}}_{=:\; G} \in \mathbb{R}^{m \times m}$$

wird als *gramsche Matrix* bezeichnet. Sie ist offensichtlich symmetrisch und wegen der Eindeutigkeit des Proximums auch regulär; schließlich liegt aufgrund der leicht nachzuweisenden Identität $\alpha^\top G \alpha \;=\; \| \sum_{k=1}^m \alpha_k u_k \|^2$ für $\alpha \;=\; (\alpha_1, \ldots, \alpha_m)^\top \in \mathbb{R}^m$ auch positive Definitheit vor. Das zugehörige Gleichungssystem (15.11) nennt man *Normalgleichungen* (für Proxima).

(b) Wenn mit den Bezeichnungen aus Theorem 15.25 die Vektoren $u_1, \ldots, u_m$ eine *orthonormale* Basis des Unterraums $\mathcal{U}$ bilden, so vereinfacht sich die Berechnung des Proximums zu

$$u^* \;=\; \sum_{k=1}^{m} \langle v, u_k \rangle u_k.$$

Diese Eigenschaft macht man sich beispielsweise beim Verfahren der konjugierten Gradienten zu Nutze. $\triangle$

Abschließend wird als weitere Anwendung von Theorem 15.24 eine Charakterisierung für Lösungen linearer Ausgleichsprobleme geliefert:

Korollar 15.27. *Zu gegebener Matrix $A \in \mathbb{R}^{M \times N}$ sowie gegebenem Vektor $b \in \mathbb{R}^M$ ist der Vektor $x_* \in \mathbb{R}^N$ genau dann eine Lösung des linearen Ausgleichsproblems $\| Ax - b \|_2 \to \min$ für $x \in \mathbb{R}^N$, wenn x_* zugleich eine Lösung der Normalgleichungen $A^\top A x = A^\top b$ darstellt.*

[6] siehe Übungsaufgabe 15.2

BEWEIS. Die Aussage folgt unmittelbar aus Theorem 15.24 unter Beachtung der Identität $\mathcal{R}(A)^\perp = \mathcal{N}(A^\top)$, wobei $\mathcal{R}(A)$ den Bildraum der Matrix A und $\mathcal{N}(A^\top)$ den Nullraum der transponierten Matrix $A^\top$ bezeichnet. □

15.5 Gleichmäßige Approximation stetiger Funktionen durch Polynome vom Höchstgrad $n-1$

Eine wichtige Rolle auf dem Raum $\mathbb{R}^N$ sowie auf Funktionenräumen kommt der gleichmäßigen Approximation zu, die mathematisch mittels Maximumnormen beschrieben wird. Solche Normen sind jedoch nicht durch Skalarprodukte induziert und somit die Resultate aus Abschnitt 15.4 nicht anwendbar. Auch strikte Normiertheit liegt in Vektorräumen mit Maximumnormen nicht vor, so dass Theorem 15.19 über die Eindeutigkeit von Proxima in linearen Unterräumen ebenfalls nicht anwendbar ist. Dennoch sind in solchen Räumen für *spezielle* lineare Unterräume Eindeutigkeitsaussagen möglich beziehungsweise existieren Lösungsverfahren.

Im Folgenden sollen speziell die Unterräume Π_{n-1} des Raums $C[a,b]$ betrachtet werden, wobei dieser mit einer gewichteten Maximumnorm von der Gestalt

$$\|\psi\|_{\infty,w} := \sup_{t\in[a,b]} |\psi(t)|w(t), \qquad \psi \in C[a,b], \tag{15.12}$$

versehen ist mit einer Gewichtsfunktion

$$w : [a,b] \to \mathbb{R} \text{ stetig}, \qquad w(t) > 0 \quad \text{für } t \in [a,b]. \tag{15.13}$$

Das folgende Theorem liefert in der vorliegenden Situation eine Charakterisierung für Π_{n-1}-Proxima an stetige Funktionen.

Theorem 15.28 (Alternantensatz). *Mit den Bezeichnungen (15.12) und (15.13) seien eine Funktion $f \in C[a,b]$ sowie ein Polynom $p^* \in \Pi_{n-1}$ mit $p^* \neq f$ gegeben. Dann sind die folgenden Aussagen (a) und (b) äquivalent:*

(a) p^* ist ein Π_{n-1}-Proximum an f, es gilt also

$$\|f - p^*\|_{\infty,w} = \min_{p\in\Pi_{n-1}} \|f - p\|_{\infty,w}.$$

(b) Es existiert eine Alternante $s_0, s_1, \ldots, s_n \in [a,b]$ *für f und p^*, das heißt,*

$$s_0 < s_1 < \ldots < s_n,$$
$$(f(s_k) - p^*(s_k))w(s_k) = -(f(s_{k-1}) - p^*(s_{k-1}))w(s_{k-1}) \quad \text{für } k = 1,2,\ldots,n,$$

und diese Alternante besitzt die Eigenschaft

$$|f(s_k) - p^*(s_k)|w(s_k) = \|f - p^*\|_{\infty,w} \qquad \text{für } k = 0,1,\ldots,n.$$

BEWEIS. "$(b) \Longrightarrow (a)$": Angenommen, es gibt ein Polynom $p \in \Pi_{n-1}$ mit der Eigenschaft

$$\underbrace{\sup_{t\in[a,b]} |f(t) - p(t)|w(t)}_{= \|f - p\|_{\infty,w}} < \underbrace{\sup_{t\in[a,b]} |f(t) - p^*(t)|w(t)}_{= \|f - p^*\|_{\infty,w}}. \tag{15.14}$$

Eine Veranschaulichung der vorliegenden Situation liefert Bild 15.6.

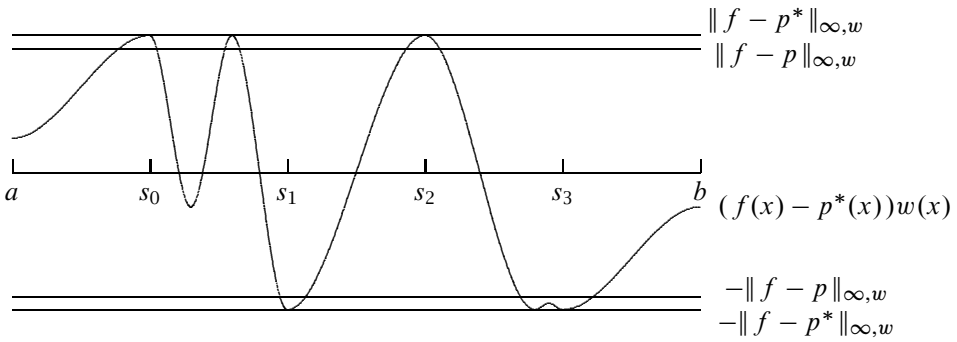


Bild 15.6: $(b) \Longrightarrow (a)$: Beweisveranschaulichung für den Spezialfall $n = 3$

Man betrachte nun die Funktion

$$\psi := (f - p^*)w - (f - p)w = (p - p^*)w.$$

Für $s_0, s_1, \ldots, s_n$ entsprechend (b) erhält man aus der Ungleichung (15.14) Folgendes: $f(s_k) - p^*(s_k) > 0$ impliziert $\psi(s_k) > 0$, und entsprechend impliziert $f(s_k) - p^*(s_k) < 0$ die Ungleichung $\psi(s_k) < 0$. Daher wechselt die Funktion ψ mindestens n-mal ihr Vorzeichen auf dem Intervall $[a, b]$, und damit hat $p - p^*$ mindestens n paarweise verschiedene Nullstellen, woraus $p = p^*$ folgt, was einen Widerspruch zur Ungleichung (15.14) darstellt.

"$(a) \Longrightarrow (b)$": Angenommen, es existiert keine Alternante für f und p^*. In diesem Fall kann das Intervall $[a, b]$ in $1 \le n_* \le n$ abgeschlossene Teilintervalle

$$I_k = [t_{k-1}, t_k], \qquad 1 \le k \le n_* \qquad \big(\text{mit } a = t_0 < t_1 < \ldots < t_{n_*} = b\big)$$

zerlegt werden, so dass Folgendes gilt:

- $(f(t_k) - p^*(t_k))w(t_k) = 0$ für $k = 1, 2, \ldots, n_* - 1$;
- für jeden Index $k \in \{1, 2, \ldots, n_*\}$ existiert ein $s_k \in I_k$ mit

$$|f(s_k) - p^*(s_k)|w(s_k) = \|f - p^*\|_{\infty,w} \neq 0,$$
$$\forall x \in I_k : -(f(x) - p^*(x))w(x) \neq (f(s_k) - p^*(s_k))w(s_k) \quad \text{für } k = 1, 2, \ldots, n_*;$$

- für jeden Index $k \in \{1, 2, \ldots, n_* - 1\}$ gilt

$$(f(s_k) - p^*(s_k))w(s_k) = -(f(s_{k+1}) - p^*(s_{k+1}))w(s_{k+1}).$$

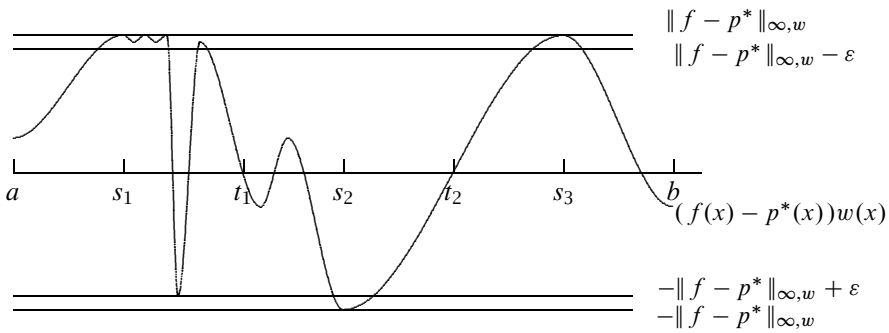


Bild 15.7: $(a) \Longrightarrow (b)$: Beweisveranschaulichung für den Spezialfall $n_* = 3$

O. B. d. A. darf noch angenommen werden, dass

$$\begin{aligned} f(s_k) - p^*(s_k) &> 0 \qquad \text{für } k \text{ ungerade,} \\ \text{\_\_\_\_\_ « \_\_\_\_\_} &< 0 \qquad \text{für } k \text{ gerade.} \end{aligned}$$

Dann existiert notwendigerweise eine Zahl $\varepsilon > 0$ mit

$$\begin{aligned} \inf_{t\in I_k} (f(t) - p^*(t))w(t) &\geq -\|f - p^*\|_{\infty,w} + \varepsilon \qquad \text{für } k \text{ ungerade,} \\ \sup_{t\in I_k} \text{\_\_\_\_\_ « \_\_\_\_\_} &\leq \|f - p^*\|_{\infty,w} - \varepsilon \qquad \text{für } k \text{ gerade,} \end{aligned}$$

und dann gibt es ein Polynom $\Delta p \in \Pi_{n_*-1}$ mit den folgenden Eigenschaften:

$\Delta p < 0$ auf I_k, falls k ungerade, $\Delta p > 0$ auf I_k, falls k gerade, $\|\Delta p\|_{\infty,w} \leq \varepsilon/2$,

wobei die letztgenannte Eigenschaft durch Multiplikation mit einer kleinen positiven Konstanten folgt. Eine Illustration der vorliegenden Situation findet sich in Bild 15.7. Für das Polynom

$$p := p^* - \Delta p \in \Pi_{n-1}$$

gilt dann $f - p = f - p^* + \Delta p$ und daher

$$\begin{aligned} (f - p)(t) &< (f - p^*)(t), \qquad t \in (t_{k-1}, t_k) \qquad \text{für } k \text{ ungerade,} \\ \inf_{t\in I_k} (f(t) - p(t))w(t) &\geq -\|f - p^*\|_{\infty,w} + \varepsilon/2 \qquad \text{\_\_\_\_\_ « \_\_\_\_\_} \\ (f - p)(t) &> (f - p^*)(t), \qquad t \in (t_{k-1}, t_k) \qquad \text{für } k \text{ gerade,} \\ \sup_{t\in I_k} (f(t) - p(t))w(t) &\leq \|f - p^*\|_{\infty,w} - \varepsilon/2 \qquad \text{\_\_\_\_\_ « \_\_\_\_\_} \end{aligned}$$

und infolgedessen ergibt sich der Widerspruch $\|f - p\|_{\infty,w} < \|f - p^*\|_{\infty,w}$. Dies komplettiert den Beweis. □

Bemerkung 15.29. Die Voraussetzungen des Alternantensatzes lassen sich abschwächen. So genügt es, von der Funktion w anstelle Positivität lediglich Nichtnegativität zu fordern, das heißt, $w(t) \geq 0$ für $t \in [a, b]$ [7], und außerdem kann die

[7] Dann stellt $\|\cdot\|_{\infty,w}$ im Allgemeinen keine Norm mehr dar, was aber hier keine Rolle spielt.

Bedingung "$p^* \neq f$" zu "$\|f - p^*\|_{\infty,w} > 0$" abgeschwächt werden. Weiter können – anstelle stetiger f – solche Funktionen $f : [a,b] \to \overline{\mathbb{R}} = \mathbb{R} \cup \{-\infty, \infty\}$ zugelassen werden, für die das Produkt fw eine auf dem Intervall $[a,b]$ stetige Funktion ergibt. Der Beweis lässt sich ohne Weiteres auf diese allgemeinere Situation übertragen, für die ebenfalls Anwendungen existieren (siehe Nemirovskiĭ/Polyak [76]). Beispiele hierzu werden in den Aufgaben 15.3 und 15.4 vorgestellt. △

15.6 Anwendungen des Alternantensatzes

15.6.1 Ein Beispiel

Beispiel 15.30. Zu einer gegebenen konvexen Funktion $f \in C^2[a,b]$ ist das Π_1-Proximum gesucht. Aus dem Mittelwertsatz der Differenzialrechnung erhält man eine Zwischenstelle $\xi \in (a,b)$ mit der Eigenschaft

$$f'(\xi) \;=\; \frac{f(b)-f(a)}{b-a},$$

und das Π_1-Proximum p^* ist dann gegeben durch

$$\begin{aligned} p^*(t) \;&:=\; \tfrac{1}{2}\Big(\tfrac{f(b)-f(a)}{b-a}(t-\xi) \;+\; f(\xi) \;+\; \tfrac{f(b)-f(a)}{b-a}(t-a) \;+\; f(a)\Big) \\ &=\; \tfrac{f(b)-f(a)}{b-a}\Big(t - \tfrac{a+\xi}{2}\Big) \;+\; \tfrac{f(a)\;+\;f(\xi)}{2}, \qquad t \in \mathbb{R}, \end{aligned}$$

denn die Punkte $s_0 = a,\ s_1 = \xi$ und $s_2 = b$ bilden eine Alternante,

$$-(p^*-f)(a) \;=\; (p^*-f)(\xi) \;=\; -(p^*-f)(b) \;=\; \|p^*-f\|_\infty.$$

Die vorliegende Situation ist in Bild 15.8 dargestellt. △

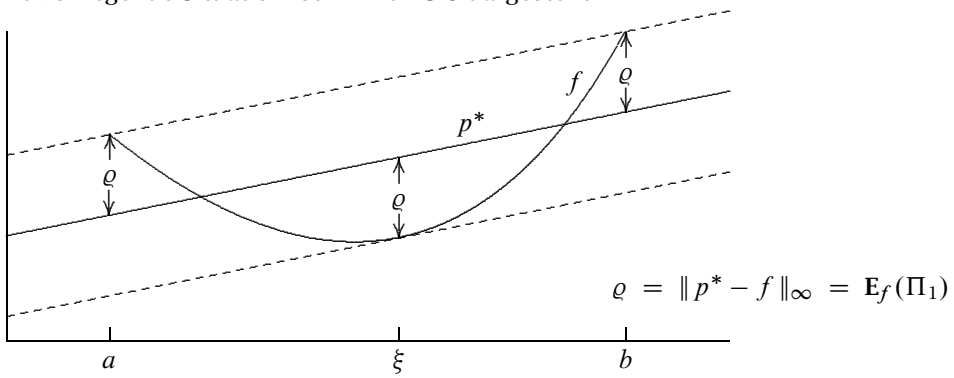


Bild 15.8: Veranschaulichung der in Beispiel 15.30 vorliegenden Situation

15.6.2 Eine erste Anwendung des Alternantensatzes

Theorem 15.31. *Für $n \geq 1$ ist das Polynom*

$$p^*(t) \;=\; t^n \;-\; \tfrac{1}{2^{n-1}} T_n(t), \qquad t \in \mathbb{R},$$

bezüglich der Maximumnorm ein Π_{n-1}-Proximum an die Funktion $f(t) = t^n,\ t \in [-1,1]$, mit

$$\| p^* - t^n \|_\infty \ = \ \min_{p \in \Pi_{n-1}} \| p - t^n \|_\infty \ = \ \frac{1}{2^{n-1}}.$$

Hierbei bezeichnet $T_n \in \Pi_n$ das n-te Tschebyscheff-Polynom der ersten Art, es gilt also $T_n(t) = \cos(n \arccos t),\ t \in [-1,1]$.

BEWEIS. Der führende Koeffizient von T_n ist 2^{n-1} (siehe Theorem 1.23 auf Seite 14), und somit gilt $p^* \in \Pi_{n-1}$. Weiter gilt offensichtlich $\| p^* - t^n \|_\infty = \frac{1}{2^{n-1}}$, und das System $s_k = \cos\left(\frac{(n-k)\pi}{n}\right),\ k = 0, 1, \ldots, n$, bildet aufgrund von

$$s_k^n \ - \ p^*(s_k) \ = \ \frac{1}{2^{n-1}} T_n(s_k) \ = \ \frac{(-1)^{n-k}}{2^{n-1}} \qquad \text{für} \quad k \ = \ 0, 1, \ldots, n,$$

eine Alternante, so dass aus Theorem 15.28 die Aussage des Theorems folgt. □

Als unmittelbare Konsequenz ergibt sich das folgende Resultat (vergleiche Theorem 1.24):

Korollar 15.32. *Für die Zahlen $t_k^{(n)} = \cos\left(\frac{(2k-1)\pi}{2n}\right),\ k = 1, 2, \ldots, n$ (mit $n \in \mathbb{N}$) gilt die folgende Optimalitätseigenschaft:*

$$\min_{y_1,\ldots,y_n \in [-1,1]} \max_{t \in [-1,1]} |(t - y_1) \ldots (t - y_n)| \overset{(*)}{=} \max_{t \in [-1,1]} |(t - t_1^{(n)}) \ldots (t - t_n^{(n)})| \overset{(**)}{=} \frac{1}{2^{n-1}}.$$

BEWEIS. Bei den Werten $t_1^{(n)}, \ldots, t_n^{(n)}$ handelt es sich um die Nullstellen des Tschebyscheff-Polynoms T_n, und der führende Koeffizient von T_n lautet 2^{n-1}; daraus resultiert die Identität $(**)$. Die Ungleichung "$\leq$" in $(*)$ ist offensichtlich richtig, und "$\geq$" schließlich erhält man wie folgt:

$$\frac{1}{2^{n-1}} \overset{(\bullet)}{=} \min_{p \in \Pi_{n-1}} \| \underbrace{p - t^n}_{\in\ \Pi_n} \|_\infty \ \leq \ \min_{y_1,\ldots,y_n \in [-1,1]} \max_{t \in [-1,1]} | \underbrace{(t - y_1) \ldots (t - y_n)}_{\in\ \Pi_n} |,$$

wobei die Identität $(\bullet)$ eine Konsequenz aus Theorem 15.31 ist. □

15.6.3 Eine zweite Anwendung des Alternantensatzes

Als eine weitere Anwendung des Alternantensatzes erhält man das folgende Resultat. Es liefert nachträglich die Optimalität der im Beweis von Theorem 11.19 über die Konvergenzraten beim Verfahren der konjugierten Gradienten verwendeten Polynome (bezogen auf das Intervall $[m, M]$).

Theorem 15.33. *Ausgehend von Zahlen $0 < m \le M$ gilt für das Polynom*

$$p^*(\lambda) := cT_n\big(\tfrac{M+m-2\lambda}{M-m}\big), \qquad c := T_n\big(\tfrac{M+m}{M-m}\big)^{-1} \qquad (\lambda \in \mathbb{R}),$$

Folgendes:

$$p^* \in \Pi_n, \qquad p^*(0) = 1, \tag{15.15}$$

$$\max_{m\le\lambda\le M} |p^*(\lambda)| = \min_{\substack{p\in\Pi_n\\ p(0)=1}} \max_{m\le\lambda\le M} |p(\lambda)| = T_n\big(\tfrac{M+m}{M-m}\big)^{-1}. \tag{15.16}$$

BEWEIS. Die Eigenschaft (15.15) ist offensichtlich richtig, und für den Nachweis von (15.16) betrachtet man die folgenden Darstellungen,

$$\max_{m\le\lambda\le M} |p^*(\lambda)| = \max_{m\le\lambda\le M} \lambda|\tfrac{1}{\lambda} - q^*(\lambda)| = c \quad \text{mit } q^*(\lambda) := \tfrac{1-p^*(\lambda)}{\lambda} \in \Pi_{n-1},$$

$$\min_{\substack{p\in\Pi_n\\ p(0)=1}} \max_{m\le\lambda\le M} |p(\lambda)| = \min_{q\in\Pi_{n-1}} \max_{m\le\lambda\le M} \lambda|\tfrac{1}{\lambda} - q(\lambda)|,$$

und erhält die Aussage des Theorems mittels Theorem 15.28 angewandt mit q^* anstelle p^* sowie

$$[a,b] = [m,M], \qquad w(\lambda) = \lambda, \qquad f(\lambda) = \tfrac{1}{\lambda},$$

unter Berücksichtigung der Tatsache, dass

$$\lambda_k := -\tfrac{M-m}{2}s_k + \tfrac{M+m}{2} \qquad \text{mit } s_k := \cos\big(\tfrac{k\pi}{n}\big), \qquad k = 0,1,\dots,n,$$

eine Alternante darstellt,

$$\lambda_k\big(\tfrac{1}{\lambda_k} - q^*(\lambda_k)\big) = p^*(\lambda_k) = T_n(s_k) = c(-1)^k \quad \text{für } k = 0,1,\dots,n. \qquad \square$$

Bemerkung 15.34. Zur Bestimmung eines solchen Π_{n-1}-Proximums lässt sich – auf der Grundlage des Alternantensatzes – ein Algorithmus angeben, das *Austauschverfahren von Remez*. Einzelheiten hierzu finden Sie beispielsweise in Hämmerlin/Hoffmann [48] und in Schaback/Wendland [92]. △

15.7 Haarsche Räume, Tschebyscheff-Systeme

Die Aussage des Alternantensatzes behält ihre Gültigkeit, wenn man anstelle des Raums Π_{n-1} der Polynome vom Grad $\le n-1$ haarsche Räume mit der Dimension n betrachtet. Die entsprechende Theorie wird im Folgenden vorgestellt. Von grundlegender Bedeutung sind dabei die folgenden Begriffe.

Definition 15.35. (a) Ein endlich-dimensionaler linearer Raum $\mathcal{U} \subset C[a,b]$ heißt *haarscher Raum*, falls jede Funktion $0 \not\equiv u \in \mathcal{U}$ höchstens $n-1$ paarweise verschiedene Nullstellen besitzt, wobei $n := \dim \mathcal{U}$.

(b) Ein linear unabhängiges Funktionensystem $u_1, \ldots, u_n \in C[a,b]$ heißt *Tschebyscheff-System*, falls $\mathcal{U} = \operatorname{span}\{u_1, \ldots, u_n\} \subset C[a,b]$ einen haarschen Raum bildet.

Beispiel 15.36. (a) Die Monome $1, x, x^2, \ldots, x^{n-1} \in C[a,b]$ bilden offensichtlich ein Tschebyscheff-System.

(b) Die Exponentialmonome $1, e^x, e^{2x}, \ldots, e^{(n-1)x} \in C[a,b]$ bilden ein Tschebyscheff-System.

BEWEIS. Hier betrachtet man

$$\mathcal{U} := \operatorname{span}\{1,\ e^x,\ e^{2x}, \ldots,\ e^{(n-1)x}\} = \{p \circ e^x \ :\ p \in \Pi_{n-1}\} \subset C[a,b].$$

Falls dann $u = p \circ e^x \in \mathcal{U}$ mindestens n paarweise verschiedene Nullstellen $a \le x_1 < \ldots < x_n \le b$ hat, so besitzt das Polynom $p \in \Pi_{n-1}$ die n paarweise verschiedenen Nullstellen $e^{x_1} < \ldots < e^{x_n}$, und somit gilt notwendigerweise $p \equiv 0$ beziehungsweise $u \equiv 0$. □

(c) Für $0 \le a < b < 2\pi$ bilden die trigonometrischen Monome $1, \sin x, \cos x, \ldots, \sin mx, \cos mx \in C[a,b]$ ein Tschebyscheff-System.

BEWEIS. Hierzu betrachtet man

$$\begin{aligned}
\mathcal{U} &:= \operatorname{span}\{1,\ \sin x,\ \cos x, \ldots,\ \sin mx,\ \cos mx\} \\
&= \Big\{ \sum_{k=0}^{m} \big(\alpha_k \sin kx + \beta_k \cos kx\big) \ :\ \alpha_k,\ \beta_k \in \mathbb{R} \Big\} \\
&= \Big\{ \sum_{k=-m}^{m} \gamma_k e^{\mathrm{i}kx} \ :\ \gamma_k \in \mathbb{C},\quad \operatorname{Re}\gamma_k = \operatorname{Re}\gamma_{-k},\quad \operatorname{Im}\gamma_k = -\operatorname{Im}\gamma_{-k} \Big\} \\
&\subset \{e^{-\mathrm{i}mx} q \circ e^{\mathrm{i}x} \ :\ q \in \Pi_{2m}\} \subset C[a,b].
\end{aligned}$$

Falls dann $u = e^{-\mathrm{i}mx}(q \circ e^{\mathrm{i}x}) \in \mathcal{U}$ mindestens $(2m+1)$ paarweise verschiedene Nullstellen $0 \le x_0 < \ldots < x_{2m} < 2\pi$ besitzt, so hat (aufgrund der Injektivität der Funktion $e^{\mathrm{i}x}$ auf dem Intervall $[0, 2\pi)$) das Polynom $q \in \Pi_{2m}$ mindestens $(2m+1)$ paarweise verschiedene Nullstellen und somit gilt notwendigerweise $q \equiv 0$ beziehungsweise $u \equiv 0$. □

△

15.7.1 Alternantensatz für haarsche Räume

Der Alternantensatz lässt sich auf haarsche Räume übertragen:

Theorem 15.37. *Für einen haarschen Raum $\mathcal{U} \subset C[a,b]$ der Dimension* $\dim \mathcal{U} = n$ *behält der Alternantensatz seine Gültigkeit, wenn dort "Π_{n-1}" durch "$\mathcal{U}$" ersetzt wird.*[8]

[8] Etwas genauer ist dort noch das Wort "Polynom" zu streichen, und sinnvollerweise wird man die Notation "p^*" durch "u^*" ersetzen.

BEWEIS. Der Beweis verläuft ähnlich dem des Alternantensatzes für Polynome, unter Verwendung des nachfolgenden Resultats über die eindeutige Lösbarkeit des Interpolationsproblems in haarschen Räumen. □

Theorem 15.38. *Zu einem haarschen Raum $\mathcal{U} \subset C[a,b]$ der Dimension* $\dim \mathcal{U} = n$ *und n Stützpunkten $(x_1, f_1), \ldots, (x_n, f_n)$, mit paarweise verschiedenen Stützstellen $x_1, x_2, \ldots, x_n \in [a,b]$ gibt es genau ein Element $u \in \mathcal{U}$ mit der Interpolationseigenschaft*

$$u(x_j) \;=\; f_j \qquad \text{für} \quad j \;=\; 1, 2, \ldots, n.$$

BEWEIS. Wird hier nicht geführt (Aufgabe 15.6). □

15.7.2 Eindeutigkeit des Proximums

Für haarsche Räume $\mathcal{U} \subset C[a,b]$ ist die Existenz von $\mathcal{U}$-Proxima an Funktionen $f \in C[a,b]$ aufgrund von Korollar 15.10 gewährleistet. Im Folgenden werden nun Eindeutigkeitsbetrachtungen geführt, der Einfachheit halber nur für die spezielle Gewichtsfunktion $w \equiv 1$.

Theorem 15.39. *Bezüglich der Maximumnorm $\|\cdot\|_\infty$ auf dem Intervall $[a,b]$ ist in einem haarschen Raum $\mathcal{U} \subset C[a,b]$ zu jedem $f \in C[a,b]$ das $\mathcal{U}$-Proximum an die Funktion f eindeutig bestimmt.*

BEWEIS. Für zwei $\mathcal{U}$-Proxima $u_1,\ u_2 \in \mathcal{U}$ an die Funktion f ist auch[9] die Funktion $\frac{1}{2}(u_1 + u_2)$ ein $\mathcal{U}$-Proximum an f, für die dann eine Alternante $a \le s_0 < s_1 < \ldots < s_n \le b,\ n := \dim \mathcal{U}$, existiert, das heißt,

$$\underbrace{\tfrac{1}{2}(u_1 - f)(s_k)}_{|| \;\le\; \mathbf{E}_f(\mathcal{U})} + \underbrace{\tfrac{1}{2}(u_2 - f)(s_k)}_{|| \;\le\; \mathbf{E}_f(\mathcal{U})} \;=\; \big(\tfrac{1}{2}(u_1 + u_2) - f\big)(s_k) \;=\; \tau(-1)^k \mathbf{E}_f(\mathcal{U}),$$
$$k \;=\; 0, 1, \ldots, n, \qquad \tau \in \{-1, 1\},$$

und daher

$$(u_1 - f)(s_k) \;=\; (u_2 - f)(s_k) \quad \text{bzw.} \quad \underbrace{(u_1 - u_2)}_{\in\; \mathcal{U}}(s_k) \;=\; 0 \quad \text{für } k \;=\; 0, 1, \ldots, n,$$

so dass notwendigerweise $u_1 \equiv u_2$ gilt. □

Bemerkung 15.40. Man beachte, dass der Vektorraum $C[a,b]$ zusammen mit der Maximumnorm $\|\cdot\|_\infty$ nicht strikt normiert ist, so dass Theorem 15.39 nicht unmittelbar aus Theorem 15.19 resultiert. △

15.7.3 Untere Schranken für den Minimalabstand

Ist für eine Approximation $u \in \mathcal{U}$ an eine Funktion f eine Alternante gegeben, an dessen Punkten jedoch der Abstand von u zur Funktion f nicht maximal und der Alternantensatz daher nicht anwendbar ist, so gewinnt man doch zumindest eine untere Schranke für den Minimalabstand $\mathbf{E}_f(\mathcal{U})$:

[9] siehe Lemma 15.14

Theorem 15.41 (de la Valleé-Poussin). *Seien $\mathcal{U} \subset C[a,b]$ ein haarscher Raum sowie $f \in C[a,b]$ und $u \in \mathcal{U}$. Wenn $a \le s_0 < s_1 < \ldots < s_n \le b$ mit $n := \dim \mathcal{U}$ eine Alternante bezüglich der Funktionen f und u darstellt, das heißt,*

$$(u-f)(s_k) \;=\; \tau\,\delta\,(-1)^k \qquad \text{für } \; k \;=\; 0,1,\ldots,n,$$

erfüllt ist mit geeigneten Zahlen $\tau \in \{-1,1\}$ und $0 < \delta \le \|u-f\|_\infty$, so gilt die folgende Abschätzung,

$$\delta \;\le\; \mathbf{E}_f(\mathcal{U}).$$

BEWEIS. Im Fall $\mathbf{E}_f(\mathcal{U}) < \delta$ würde man für das $\mathcal{U}$-Proximum u^* an f die Identität

$$u - u^* \;=\; u - f \;-\underbrace{(u^* - f)}_{\|\ \|_\infty \;=\; \mathbf{E}_f(\mathcal{U})}$$

erhalten, mit der Konsequenz

$$\operatorname{sgn}(\underbrace{u-u^*}_{\in\ \mathcal{U}})(s_k) \;=\; \operatorname{sgn}(u-f)(s_k) \;=\; \tau(-1)^k \qquad \text{für } \; k \;=\; 0,1,\ldots,n,$$

so dass die Funktion $u - u^*$ dann n Nullstellen besitzen würde und infolgedessen sich der Widerspruch $u \equiv u^*$ ergäbe. □

Bemerkung 15.42. In Ergänzung zu Theorem 15.41 kann man für den Minimalabstand noch die triviale obere Schranke $\mathbf{E}_f(\mathcal{U}) \le \|u-f\|_\infty$ angeben. △

Weitere Themen und Literaturhinweise

Ausführliche Behandlungen des Themas Approximationstheorie finden Sie beispielsweise in Hämmerlin/Hoffmann [48], Opfer [79] und in Schaback/Wendland [92]. Die in Abschnitt 15.3.2 vorgestellte Theorie der strikt normierten Räume lässt sich erweitern um die Theorie der gleichmäßig konvexen, vollständig normierten Räume $\mathcal{V}$, in denen für konvexe abgeschlossene Teilmengen $\varnothing \neq \mathcal{M} \subset \mathcal{V}$ die Existenz von $\mathcal{M}$-Proxima gewährleistet ist. Einzelheiten hierzu werden beispielsweise in Hirzebruch/Scharlau [56] vorgestellt. Dort werden auch (für mit einem Skalarprodukt versehene Räume) Orthonormalsysteme behandelt, die zur Bestimmung von Proxima in Unterräumen verwendet werden. Einführungen zu dem in Bemerkung 15.4 angesprochenen Thema "nichtlineare Optimierung" finden Sie beispielsweise in Dennis/Schnabel [17], Grossmann/Terno [44], Geiger/Kanzow [32], Nash/Sofer [75], Schaback/Wendland [92], Schwarz/Klöckner [94], Schwetlick [95], Tröltzsch [105] oder Werner [111].

Übungsaufgaben

Aufgabe 15.1. Man weise nach, dass der Vektorraum $C[a,b]$ zusammen mit der Maximumnorm $\|\cdot\|_\infty$ nicht strikt normiert ist.

Aufgabe 15.2. Man weise die Äquivalenz (15.11) nach.

Aufgabe 15.3. Man weise für die Folge von Funktionen

$$p_n(t) \;=\; \frac{(-1)^n}{2n+1}\,\frac{T_{2n+1}(\sqrt{t})}{\sqrt{t}}, \qquad t>0 \qquad (n \;=\; 0,1,\ldots)$$

Folgendes nach:

$$\begin{aligned} p_n &\overset{(*)}{\in} \Pi_n, && p_n(0)=1,\\ \max_{0\le t\le 1} |\,p_n(t)|\sqrt{t} &= \frac{1}{2n+1} && \text{für } n \;=\; 0,1,\ldots,\\ \max_{0\le t\le 1} |\,p_n(t)|\sqrt{t} &= \min_{\substack{p\in\Pi_n\\ p(0)=1}} \max_{0\le t\le 1} |\,p(t)|\sqrt{t}, \end{aligned}$$

wobei $(*)$ so zu verstehen ist, dass zu der Funktion p_n eine Fortsetzung nach 0 und darüber hinaus auf die negative Halbachse existiert, welche ein Polynom von Höchstgrad n darstellt.

Aufgabe 15.4. Man überlege sich, dass für die Folge von Funktionen

$$p_n(t) \;=\; \frac{1-T_{n+1}(1-2t)}{2(n+1)^2 t}, \qquad 0\neq t\in\mathbb{R} \qquad (n \;=\; 0,1,\ldots)$$

Folgendes gilt:

$$\begin{aligned} p_n &\in \Pi_n, && p_n(0)=1,\\ \max_{0\le t\le 1} |\,p_n(t)|\,t &= \frac{1}{(n+1)^2} && \text{für } n \;=\; 0,1,\ldots,\\ \max_{0\le t\le 1} |\,p_n(t)|\,t &\neq \min_{\substack{p\in\Pi_n\\ p(0)=1}} \max_{0\le t\le 1} |\,p(t)|\,t. \end{aligned}$$

Aufgabe 15.5. Es ist $p\equiv 0$ bezüglich der Maximumnorm ein Π_{n-1}-Proximum an die Funktion $f(t)=\sin 3t,\ t\in[0,2\pi]$ genau dann, wenn $n-1\le 2$ gilt.

Aufgabe 15.6. Man beweise Theorem 15.38.

16 Rechnerarithmetik

In dem vorliegenden Kapitel werden zunächst einige Grundlagen über die in Hard- und Software verwendeten reellen Zahlensysteme vorgestellt. Anschließend wird die Approximation reeller Zahlen durch Elemente solcher Zahlensysteme behandelt. Ein weiteres Thema bilden die arithmetischen Grundoperationen in diesen Zahlensystemen.

Bemerkung 16.1. Solche Umwandlungs- und Arithmetikfehler verursachen bei jedem numerischen Verfahren Fehler sowohl in den Eingangsdaten als auch bei der Durchführung des jeweiligen Verfahrens. Für verschiedene Situationen sind die Auswirkungen solcher Fehler in einem allgemeinen Kontext bereits diskutiert worden:

- der Einfluss fehlerbehafteter Matrizen und rechter Seiten auf die Lösung eines zugrunde liegenden linearen Gleichungssystems (Abschnitt 4.7.5),
- und bei Einschrittverfahren zur Lösung von Anfangswertproblemen für gewöhnliche Differenzialgleichungen die Auswirkungen der in jedem Integrationsschritt auftretenden eventuellen Fehler auf die Güte der Approximation an die Lösung der Differenzialgleichung (Abschnitt 7.4),
- und der Einfluss fehlerbehafteter Matrizen auf die Lösung von Eigenwertproblemen (Abschnitt 12.2). △

16.1 Zahlendarstellungen

Von grundlegender Bedeutung für die Realisierung von Zahlendarstellungen auf Rechnern ist die folgende aus der Analysis bekannte Darstellung.

Theorem 16.2. *Zu gegebener* Basis $\mathrm{b} \geq 2$ *lässt sich jede Zahl* $0 \neq x \in \mathbb{R}$ *in der Form*

$$x = \sigma \sum_{k=-\mathrm{e}+1}^{\infty} a_{k+\mathrm{e}} \mathrm{b}^{-k} = \sigma \Big(\sum_{k=1}^{\infty} a_k \mathrm{b}^{-k} \Big) \mathrm{b}^{\mathrm{e}}, \qquad \begin{array}{l} a_1,\, a_2, \ldots \in \{0, 1, \ldots, \mathrm{b}-1\}, \\ \mathrm{e} \in \mathbb{Z}, \qquad \sigma \in \{+, -\} \end{array} \tag{16.1}$$

darstellen mit einer nichtverschwindenden führenden Ziffer, $a_1 \neq 0$*. Zwecks Eindeutigkeit der Ziffern sei angenommen, dass es eine unendliche Teilmenge* $\mathbb{N}_1 \subset \mathbb{N}$ *gibt mit* $a_k \neq \mathrm{b} - 1$ *für* $k \in \mathbb{N}_1$*.*

BEWEIS. Siehe etwa Forster [29]. □

Bemerkung 16.3. (a) Die zweite Darstellung für x in (16.1) bezeichnet man als Gleitpunktdarstellung.

(b) Durch die abschließende Bedingung in Theorem 16.2 ist die Eindeutigkeit der Ziffern in den Darstellungen (16.1) gewährleistet. So wird zum Beispiel für die Zahl $0.9999\ldots = 1.0$ die letztere Darstellung gewählt.

(c) Praxisrelevante Zahlensysteme und ihre Ziffern sind in Tabelle 16.1 dargestellt. △

Zahlensystem	Basis b	mögliche Ziffern
Dezimalsystem	10	0, 1, 2, 3, 4, 5, 6, 7, 8, 9
Binärsystem	2	0, 1
Oktalsystem	8	0, 1, 2, 3, 4, 5, 6, 7
Hexadezimalsystem	16	0, 1, 2, 3, 4, 5, 6, 7, 8, 9, A, B, C, D, E, F

Tabelle 16.1: Praxisrelevante Zahlensysteme und ihre Ziffern

16.2 Allgemeine Gleitpunkt-Zahlensysteme

16.2.1 Grundlegende Begriffe

In jedem Prozessor beziehungsweise bei jeder Programmiersprache werden jeweils nur einige Systeme reeller Zahlen verarbeitet. Solche Systeme werden im Folgenden vorgestellt.

Definition 16.4. Zu gegebener Basis $\mathrm{b} \geq 2$ und *Mantissenlänge* $\mathrm{t} \in \mathbb{N}$ sowie für Exponentenschranken $\mathrm{e}_{\min} < 0 < \mathrm{e}_{\max}$ ist die Menge $\mathbb{F} = \mathbb{F}(\mathrm{b}, \mathrm{t}, \mathrm{e}_{\min}, \mathrm{e}_{\max}) \subset \mathbb{R}$ wie folgt erklärt,

$$\mathbb{F} := \left\{ \sigma \Big(\sum_{k=1}^{\mathrm{t}} a_k \mathrm{b}^{-k} \Big) \mathrm{b}^{\mathrm{e}} : \begin{array}{l} a_1, \ldots, a_{\mathrm{t}} \in \{0, 1, \ldots, \mathrm{b}-1\},\ a_1 \neq 0 \\ \mathrm{e} \in \mathbb{Z},\ \mathrm{e}_{\min} \leq \mathrm{e} \leq \mathrm{e}_{\max},\quad \sigma \in \{+, -\} \end{array} \right\} \cup \{0\}. \tag{16.2}$$

Die Menge $\widehat{\mathbb{F}}$ ist definiert als diejenige Obermenge von $\mathbb{F}$, bei der in der Liste von Parametern in (16.2) zusätzlich noch die Kombination "$\mathrm{e} = \mathrm{e}_{\min},\ a_1 = 0$" zugelassen ist.

Die Elemente von $\widehat{\mathbb{F}}$ (und damit insbesondere auch die Elemente von $\mathbb{F} \subset \widehat{\mathbb{F}}$) werden im Folgenden kurz als *Gleitpunktzahlen* bezeichnet. Zu jeder solchen Gleitpunktzahl

$$x \;=\; \sigma\, a\, \mathrm{b}^{\mathrm{e}} \;\in\; \mathbb{F} \qquad \text{mit} \quad a \;=\; \sum_{k=1}^{\mathrm{t}} a_k \mathrm{b}^{-k} \tag{16.3}$$

bezeichnet σ das *Vorzeichen*, es ist a die *Mantisse* mit den *Ziffern* $a_1, \ldots, a_{\mathrm{t}}$, und e ist der *Exponent*. Gleitpunktzahlen mit der Darstellung (16.3) bezeichnet man im Fall $a_1 \geq 1$ als *normalisiert*, andernfalls als *denormalisiert*.

Bemerkung 16.5. Die Menge $\mathbb{F} \subset \mathbb{R}$ stellt folglich ein System normalisierter Gleitpunktzahlen dar. Diese Normalisierung garantiert die Eindeutigkeit in der Darstellung (16.3). Im Spezialfall des kleinsten zugelassenen Exponenten $\mathrm{e} = \mathrm{e}_{\min}$ bleibt

diese Eindeutigkeit (mit Ausnahme der Zahl 0) jedoch erhalten, wenn auf die Normalisierung verzichtet wird, so dass bis auf die Zahl 0 auch alle Gleitpunktzahlen aus $\widehat{\mathbb{F}}$ eindeutig in der Form (16.3) darstellbar sind. △

Im weiteren Verlauf werden zunächst grundlegende Eigenschaften der Mengen $\mathbb{F}$ und $\widehat{\mathbb{F}}$ festgehalten (Abschnitte 16.2.2 und 16.2.3) und anschließend einige spezielle Systeme von Gleitpunktzahlen vorgestellt (Abschnitt 16.3).

16.2.2 Struktur des normalisierten Gleitpunkt-Zahlensystems $\mathbb{F}$

Im Folgenden werden für die Gleitpunktzahlen aus dem System $\mathbb{F} \subset \mathbb{R}$ zunächst Schranken angegeben und anschließend deren Verteilung auf der reellen Achse beschrieben. Wegen der Symmetrie von $\mathbb{F}$ um den Nullpunkt genügt es dabei, deren positive Elemente zu betrachten.

Theorem 16.6. *In dem System* $\mathbb{F} = \mathbb{F}(\mathrm{b}, \mathrm{t}, \mathrm{e}_{\min}, \mathrm{e}_{\max})$ *normalisierter Gleitpunktzahlen stellen*

$$x_{\min} := \mathrm{b}^{\mathrm{e}_{\min}-1}, \qquad x_{\max} := \mathrm{b}^{\mathrm{e}_{\max}}(1-\mathrm{b}^{-\mathrm{t}}),$$

das kleinste positive beziehungsweise das größte Element dar, es gilt also $x_{\min},\ x_{\max} \in \mathbb{F}$ *und*

$$x_{\min} = \min\{x \in \mathbb{F} \ : \ x > 0\}, \qquad x_{\max} = \max\{x \in \mathbb{F}\}.$$

BEWEIS. Für die Mantisse a einer beliebigen Gleitpunktzahl aus $\mathbb{F}$ gilt notwendigerweise

$$\mathrm{b}^{-1} \ \le \ a \ \le \ \sum_{k=1}^{\mathrm{t}} \mathrm{b}^{-k}(\mathrm{b}-1) \ \overset{(*)}{=} \ 1-\mathrm{b}^{-\mathrm{t}},$$

wobei die erste Ungleichung aus der Normalisierungseigenschaft $a_1 \ge 1$ und die zweite Ungleichung aus der Eigenschaft $a_k \le \mathrm{b}-1$ resultiert. Die Summe schließlich stellt eine Teleskopsumme dar, woraus die Identität $(*)$ folgt und der Beweis komplettiert ist. □

Bemerkung 16.7. Der durch das normalisierte Gleitpunkt-Zahlensystem $\mathbb{F}$ überdeckte Bereich sieht demnach wie folgt aus, $\mathbb{F} \subset [-x_{\max}, -x_{\min}] \cup \{0\} \cup [x_{\min}, x_{\max}]$, was in Bild 16.1 veranschaulicht ist.

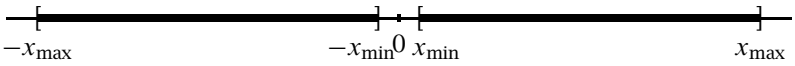


Bild 16.1: Darstellung des durch das normalisierte Gleitpunkt-Zahlensystem $\mathbb{F}$ überdeckten Bereiches △

Detaillierte Aussagen über die Verteilung der Gleitpunktzahlen aus den System $\mathbb{F}$ liefern das folgende Theorem und die anschließende Bemerkung.

Theorem 16.8. *In jedem der Intervalle* $[\mathrm{b}^{\mathrm{e}-1}, \mathrm{b}^{\mathrm{e}}]$, $\mathrm{e}_{\min} \le \mathrm{e} \le \mathrm{e}_{\max}$, *befinden sich gleich viele Gleitpunktzahlen aus dem System* $\mathbb{F}$, *bei einer jeweils äquidistanten Verteilung mit den konstanten Abständen* $\mathrm{b}^{\mathrm{e}-\mathrm{t}}$*:*

$$\mathbb{F} \cap [\mathrm{b}^{\mathrm{e}-1}, \mathrm{b}^{\mathrm{e}}] = \Big\{ \underbrace{(\mathrm{b}^{-1} + j\mathrm{b}^{-\mathrm{t}})\mathrm{b}^{\mathrm{e}}}_{\mathrm{b}^{\mathrm{e}-1} + j\mathrm{b}^{\mathrm{e}-\mathrm{t}}} : j = 0, 1, \ldots, M^\sharp \Big\}, \qquad M^\sharp := \mathrm{b}^{\mathrm{t}} - \mathrm{b}^{\mathrm{t}-1}.$$

BEWEIS. Im Folgenden werden die im Beweis von Theorem 16.6 zum Thema Mantissen angestellten Überlegungen fortgeführt. Die Mantissengesamtzahl beträgt $\mathrm{b}^{\mathrm{t}-1}(\mathrm{b}-1) = \mathrm{b}^{\mathrm{t}} - \mathrm{b}^{\mathrm{t}-1}$, und diese sind äquidistant über das gesamte abgeschlossene Intervall $[\mathrm{b}^{-1}, 1 - \mathrm{b}^{-\mathrm{t}}]$ verteilt mit jeweiligem Abstand $\mathrm{b}^{-\mathrm{t}}$, eine aufsteigende Anordnung der Mantissen sieht also wie folgt aus:

$$a = \mathrm{b}^{-1} + j\mathrm{b}^{-\mathrm{t}}, \qquad j = 0, 1, \ldots, M^\sharp - 1.$$

Hieraus resultiert die Aussage des Theorems. □

Bemerkung 16.9. Durch Theorem 16.8 wird die ungleichmäßige Verteilung der Gleitpunktzahlen auf der Zahlengeraden verdeutlicht. So tritt in dem System der normalisierten Gleitpunktzahlen $\mathbb{F}$ zwischen der größten negativen Zahl $-x_{\min}$ und der kleinsten positiven Zahl $x_{\min}$ eine (relativ betrachtet) große Lücke auf, und ferner werden die Abstände zwischen den Gleitpunktzahlen mit wachsender absoluter Größe zunehmend größer. Die beschriebene Situation für $\mathbb{F}$ ist in Bild 16.2 veranschaulicht.

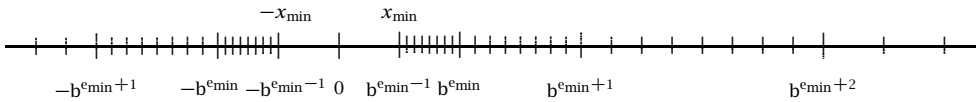


Bild 16.2: Verteilung der betragsmäßig kleinen normalisierten Gleitpunktzahlen des Systems $\mathbb{F}$ △

Eine wichtige Kenngröße des Gleitpunkt-Zahlensystems $\mathbb{F}$ ist der maximale relative Abstand der Zahlen aus dem Bereich $\{x \in \mathbb{R} : x_{\min} \le |x| \le x_{\max}\}$ zum jeweils nächstgelegenen Element aus $\mathbb{F}$. Hier gilt Folgendes:

Theorem 16.10.

$$\min_{z \in \mathbb{F}} \frac{|z - x|}{|x|} \le \underbrace{\tfrac{1}{2}\mathrm{b}^{-\mathrm{t}+1}}_{=:\ \mathrm{eps}} \qquad \textit{für } x \in \mathbb{R} \textit{ mit } x_{\min} \le |x| \le x_{\max}. \tag{16.4}$$

BEWEIS. Aus Symmetriegründen genügt es, die Betrachtungen auf positive Zahlen x zu beschränken, und im Folgenden werden die Betrachtungen auf eines der infrage kommenden Intervalle $[\mathrm{b}^{\mathrm{e}-1}, \mathrm{b}^{\mathrm{e}}]$ konzentriert. Nach Theorem 16.8 sind die Gleitpunktzahlen aus dem System $\mathbb{F}$ über das gesamte Intervall $[\mathrm{b}^{\mathrm{e}-1}, \mathrm{b}^{\mathrm{e}}]$ äquidistant verteilt mit den konstanten Abständen $\mathrm{b}^{\mathrm{e}-\mathrm{t}}$, und somit beträgt für eine beliebige reelle

Zahl x aus diesem Intervall der Abstand zum nächstgelegenen Element aus $\mathbb{F}$ höchstens $\frac{1}{2}\mathrm{b}^{\mathrm{e}-\mathrm{t}}$. Die Eigenschaft $\mathrm{b}^{\mathrm{e}-1} \leq x$ liefert schließlich die Aussage des Theorems. □

Bemerkung 16.11. Aus der Abschätzung (16.4) wird unmittelbar einsichtig, dass bei festgelegter Basis b die Genauigkeit des Gleitpunkt-Zahlensystems $\mathbb{F}$ ausschließlich von der Anzahl der Ziffern der Mantisse abhängt, während die Wahl der Exponentenschranken $\mathrm{e}_{\min}$ und $\mathrm{e}_{\max}$ die Größe des von dem Gleitpunkt-Zahlensystem $\mathbb{F}$ überdeckten Bereichs beeinflussen. △

Für die eindeutig bestimmte Zahl $n \in \mathbb{N}$ mit $0.5\times10^{-n} \leq \mathrm{eps} < 5\times10^{-n}$ spricht man im Zusammenhang mit dem System $\mathbb{F}$ von einer *n-stelligen Dezimalstellenarithmetik.*

16.2.3 Struktur des denormalisierten Gleitpunkt-Zahlensystems $\widehat{\mathbb{F}}$

Im Folgenden werden für das Obersystem $\widehat{\mathbb{F}} \supset \mathbb{F}$ die gegenüber dem System der normalisierten Gleitpunkt-Zahlensystems $\mathbb{F}$ zusätzlichen Eigenschaften beschrieben.

Theorem 16.12. *Auf dem Bereich* $(-\infty, -x_{\min}] \cup [x_{\min}, \infty)$ *stimmen die Gleitpunkt-Zahlensysteme* $\mathbb{F}$ *und* $\widehat{\mathbb{F}}$ *überein, und auf dem abgeschlossenen Intervall* $[-\mathrm{b}^{\mathrm{e}_{\min}}, \mathrm{b}^{\mathrm{e}_{\min}}] = [-\mathrm{b}x_{\min}, \mathrm{b}x_{\min}]$ *sind die Gleitpunktzahlen aus dem System* $\widehat{\mathbb{F}}$ *äquidistant verteilt mit konstanten Abständen* $\mathrm{b}^{\mathrm{e}_{\min}-\mathrm{t}}$*:*

$$\widehat{\mathbb{F}} \cap [-\mathrm{b}^{\mathrm{e}_{\min}}, \mathrm{b}^{\mathrm{e}_{\min}}] = \{ j\,\mathrm{b}^{\mathrm{e}_{\min}-\mathrm{t}} : j = -\mathrm{b}^{\mathrm{t}}, \ldots, \mathrm{b}^{\mathrm{t}} \}. \tag{16.5}$$

Insbesondere stellt

$$\widehat{x}_{\min} := \mathrm{b}^{\mathrm{e}_{\min}-\mathrm{t}}$$

das kleinste positive Element in $\widehat{\mathbb{F}}$ *dar.*

BEWEIS. Für die Mantisse a einer beliebigen denormalisierten Gleitpunktzahl aus $\widehat{\mathbb{F}}$ gilt $a_1 = 0$, und die Eigenschaft $a_k \leq \mathrm{b} - 1$ liefert

$$a \leq \sum_{k=2}^{\mathrm{t}} \mathrm{b}^{-k}(\mathrm{b}-1) = \mathrm{b}^{-1} - \mathrm{b}^{-\mathrm{t}},$$

beziehungsweise $\widehat{\mathbb{F}} \backslash \mathbb{F} \subset \{x \in \mathbb{R} : 0 < |x| < x_{\min}\}$, was identisch mit der ersten Aussage des Theorems ist. Im denormalisierten Fall sind die Mantissen über das gesamte abgeschlossene Intervall $[0, 1 - \mathrm{b}^{-\mathrm{t}}]$ äquidistant verteilt mit Mantissenabstand $\mathrm{b}^{-\mathrm{t}}$, eine aufsteigende Anordnung sieht hier wie folgt aus:

$$a = j\mathrm{b}^{-\mathrm{t}}, \qquad j = 0, 1, \ldots, \mathrm{b}^{\mathrm{t}} - 1.$$

Daraus erhält man die Aussage (16.5). □

Die beschriebene Situation für $\widehat{\mathbb{F}}$ ist in Bild 16.3 veranschaulicht.

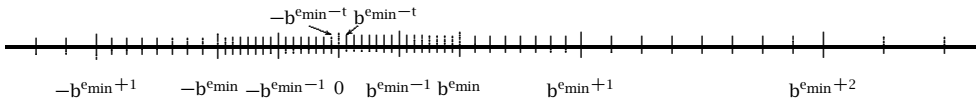


Bild 16.3: Verteilung der betragsmäßig kleinen Gleitpunktzahlen aus dem System $\widehat{\mathbb{F}}$

Bemerkung 16.13. Die in dem System der normalisierten Gleitpunktzahlen $\mathbb{F}$ (relativ gesehen) auftretenden großen Lücken zwischen der größten negativen Zahl $-x_{\min}$ und der Zahl 0 sowie zwischen 0 und der kleinsten positiven Zahl $x_{\min}$ sind in dem Gleitpunkt-Zahlensystem $\widehat{\mathbb{F}}$ aufgefüllt worden mit äquidistant verteilten denormalisierten Gleitpunktzahlen. Man beachte jedoch, dass auf der anderen Seite die *relativen* Abstände der Gleitpunktzahlen aus $\widehat{\mathbb{F}}$ zur Zahl 0 hin anwachsen bis hin zu

$$\min_{z\in\widehat{\mathbb{F}},\, z\neq\widehat{x}_{\min}} \frac{|z-\widehat{x}_{\min}|}{\widehat{x}_{\min}} \;=\; 1. \qquad \Delta$$

16.3 Gleitpunkt-Zahlensysteme in der Praxis

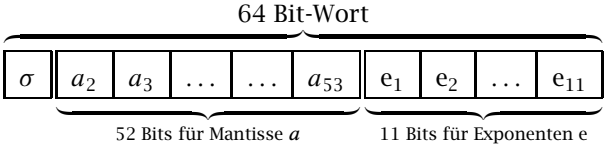
16.3.1 Die Gleitpunktzahlen des Standards IEEE 754

Zwei weitverbreitete Gleitpunkt-Zahlensysteme sind

- $\widehat{\mathbb{F}}(2,24,-125,128)$ (einfaches Grundformat),
- $\widehat{\mathbb{F}}(2,53,-1021,1024)$ (doppeltes Grundformat),

die beide Bestandteil des IEEE[1]-Standards 754 aus dem Jahr 1985 sind, in dem zugleich die Art der Repräsentation festgelegt ist. Einzelheiten hierzu werden im Folgenden erläutert, wobei mit dem gängigeren doppelten Grundformat begonnen wird. Neben den genannten Grundformaten existieren noch erweiterte Gleitpunkt-Zahlensysteme – im Folgenden kurz als *Weitformate* bezeichnet – die ebenfalls in einer einfachen und einer doppelten Version existieren und im Anschluss an die einfachen und doppelten Grundformate vorgestellt werden.

Beispiel 16.14 (IEEE, doppeltes Grundformat). Die Gleitpunktzahlen aus dem System $\widehat{\mathbb{F}}(2,53,-1021,1024)$ lassen sich in 64-Bit-Worten realisieren, wobei jeweils ein Bit zur Darstellung des Vorzeichens σ verwendet wird und 52 Bits die Mantisse sowie 11 Bits den Exponenten ausmachen,



[1] IEEE ist eine Abkürzung für "Institute of Electrical and Electronics Engineers".

Man beachte, dass bei normalisierten Gleitpunktzahlen für die führende Ziffer der Mantisse notwendigerweise $a_1 = 1$ gilt, so dass hier auf eine explizite Darstellung verzichtet werden kann. Mit den 11 Exponentenbits lassen sich wegen $2^{11} = 2048$ die 2046 Exponenten von $e_{min} = -1021$ bis $e_{max} = 1024$ kodieren. Dies geschieht in *Bias-Notation* (verschobene Notation), bei der der Exponent e durch die Dualzahldarstellung der Zahl $e - e_{min} + 1 \in \{1, \ldots, e_{max} - e_{min} + 1\} = \{1, \ldots, 2046\}$ repräsentiert wird. Von den beiden verbleibenden Bitkombinationen aus dem Exponentenbereich wird die Nullbitfolge $00\cdots0$ zur Umschaltung der Mantisse auf denormalisierte Gleitpunktzahlen ($e = e_{min}$, $a_1 = 0$) verwendet. Das verbleibende freie Bitmuster $11\cdots1$ verwendet man zur Umschaltung der Mantissenbits für die Darstellung symbolischer Ausdrücke wie $+\infty$, $-\infty$ oder *NaN-Ausdrücken*, wobei NaN eine Abkürzung für "Not a Number" ist und bestimmte arithmetische Gleitpunktoperationen wie "0/0", "$0\times\infty$" oder "$\infty - \infty$" symbolisiert. (Natürlich bleiben bei der Umschaltung zur Darstellung solcher symbolischen Ausdrücke die meisten Bitmuster der Mantisse unbelegt.)

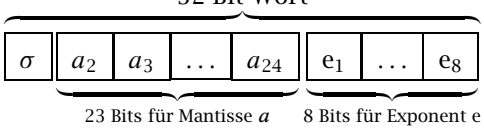
Die kleinste positive normalisierte sowie die größte Gleitpunktzahl sind hier

$$x_{min} = 2^{-1022} \approx 2.23\times10^{-308}, \qquad x_{max} \approx 2^{1024} \approx 1.80\times10^{308},$$

während $\widehat{x}_{min} = 2^{-1074} \approx 4.94\times10^{-324}$ die kleinste positive denormalisierte Gleitpunktzahl ist. Der relative Abstand einer beliebigen Zahl aus dem Bereich $\{x \in \mathbb{R} : x_{min} \le |x| \le x_{max}\}$ zum nächstgelegenen Element aus $\widehat{\mathbb{F}}(2, 53, -1021, 1024)$ beträgt höchstens

$$\text{eps} = 2^{-53} \approx 1.11\times10^{-16}. \qquad \triangle$$

Beispiel 16.15 (IEEE, einfaches Grundformat). Die Gleitpunktzahlen aus dem System $\widehat{\mathbb{F}}(2, 24, -125, 128)$ werden in 32-Bit-Worten kodiert, wovon jeweils 23 Bits für die Mantisse und 8 Bits für den Exponenten sowie ein Vorzeichenbit vergeben werden.



Aufgrund der Identität $2^8 = 256$ lassen sich mit den 8 Exponentenbits die 254 Exponenten von $e_{min} = -125$ bis $e_{max} = 128$ in Bias-Notation kodieren, und die beiden verbleibenden Bitkombinationen aus dem Exponentenbereich werden wie bei dem doppelten Grundformat verwendet. Die kleinste positive normalisierte sowie die größte Gleitpunktzahl sehen hier wie folgt aus,

$$x_{min} = 2^{-126} \approx 1.10\times10^{-38}, \qquad x_{max} \approx 2^{128} \approx 3.40\times10^{38},$$

und $\widehat{x}_{min} = 2^{-149} \approx 1.40\times10^{-45}$ ist die kleinste positive denormalisierte Gleitpunktzahl. Der relative Abstand einer beliebigen Zahl aus dem Bereich $\{x \in \mathbb{R} : x_{min} \le |x| \le x_{max}\}$ zum nächstgelegenen Element aus $\widehat{\mathbb{F}}(2, 24, -125, 128)$ beträgt höchstens $\text{eps} = 2^{-24} \approx 0.60\times10^{-7}$. $\triangle$

Beispiel 16.16 (IEEE, einfaches und doppeltes Weitformat). Neben dem genannten einfachen und doppelten Grundformat legt der IEEE-Standard 754 Gleitpunkt-Zahlensysteme im *Weitformat* fest – wiederum in einer einfachen und einer doppelten Fassung. Hierbei sind im Unterschied zu den Grundformaten lediglich Unterschranken für die verwendete Bitanzahl und die Mantissenlänge sowie Ober- und Unterschranken für den Exponenten vorgeschrieben. Ein typisches erweitertes Gleitpunkt-Zahlensystem aus der Klasse der doppelten Formate ist

$$\widehat{\mathbb{F}}(2, 64, -16381, 16384),$$

deren Elemente über 80-Bit-Worte dargestellt werden mit einem Vorzeichenbit, 64 Bits für die Mantisse sowie 15 Bits für den Exponenten. Die kleinste positive normalisierte sowie die größte Gleitpunktzahl lauten hier

$$x_{\min} \;=\; 2^{-16382} \;\approx\; 10^{-4932}, \qquad x_{\max} \;\approx\; 2^{16384} \;\approx\; 10^{4932},$$

und der maximale relative Abstand einer beliebigen reellen Zahl aus dem Bereich $\{x \in \mathbb{R} : x_{\min} \leq |x| \leq x_{\max}\}$ zum nächstgelegenen Element aus $\widehat{\mathbb{F}}(2, 64, -16381, 16384)$ liegt bei $\text{eps} \;=\; 2^{-64} \approx 5.42 \times 10^{-20}$. △

Die einfachen und doppelten Grundformate des IEEE-Standards 754 waren beziehungsweise sind in vielen gängigen Hardware- und Softwareprodukten implementiert, so zum Beispiel in den Prozessoren von Intel (486DX, Pentium), DEC (Alpha), IBM (RS/6000), Motorola (680x0) und Sun (SPARCstation) oder den Programmiersprachen C++ und Java und den Programmpaketen MATLAB und SCILAB.

16.3.2 Weitere Gleitpunkt-Zahlensysteme in der Praxis

Im Folgenden werden weitere in der Praxis verwendete Gleitpunkt-Zahlensysteme vorgestellt.

Beispiel 16.17 (Taschenrechner). Bei wissenschaftlichen Taschenrechnern werden zumeist dezimale Gleitpunkt-Zahlensysteme verwendet. Weitverbreitet ist das System $\mathbb{F}(10, 10, -98, 100)$, wobei intern mit einer längeren Mantisse (in einigen Fällen mit 12 Ziffern) gearbeitet wird. △

Beispiel 16.18 (Cray). Zwei gängige Gleitpunkt-Zahlensysteme auf Cray-Rechnern sind die Systeme $\mathbb{F}(2, 48, -16384, 8191)$ und $\mathbb{F}(2, 96, -16384, 8191)$. △

Beispiel 16.19 (IBM System/390). Auf Großrechnern von IBM existieren drei hexadezimale Gleitpunkt-Zahlensysteme: $\mathbb{F}(16, 6, -64, 63)$ (einfaches Format) sowie $\mathbb{F}(16, 14, -64, 63)$ (doppeltes Format) und $\mathbb{F}(16, 28, -64, 63)$ (erweitertes Format). Man beachte, dass bei allen drei Systemen lediglich die Mantissenlänge und somit die Genauigkeit variiert, der überdeckte Zahlenbereich hingegen bleibt unverändert. △

Die charakteristischen Größen der vorgestellten sowie einiger anderer praxisrelevanter Systeme von Gleitpunktzahlen sind in Tabelle 16.2 zusammengestellt.

Rechner o. Norm	Format	Basis b	♯ Ziff. t	Exponentlimit e_{min}	e_{max}	denorm	x_{max}	x_{min}	$\widehat{x}_{min}$	eps
IEEE	einfach	2	24	−125	128	ja	$3{\times}10^{38}$	$1{\times}10^{-38}$	$1{\times}10^{-45}$	$6{\times}10^{-8}$
– « –	doppelt	2	53	−1021	1024	ja	$2{\times}10^{308}$	$2{\times}10^{-308}$	$5{\times}10^{-324}$	$1{\times}10^{-16}$
– « –	erweit. doppelt	2	64	−16381	16384	ja	$1{\times}10^{4932}$	$1{\times}10^{-4932}$	$4{\times}10^{-4951}$	$5{\times}10^{-20}$
IBM 390	einfach	16	6	−64	63	nein	$7{\times}10^{75}$	$5{\times}10^{-79}$	–	$5{\times}10^{-7}$
— « —	doppelt	16	14	−64	63	nein	— « —	— « —	–	$1{\times}10^{-16}$
Taschenrechner (Bsp.)		10	10	−98	100	nein	$1{\times}10^{99}$	$1{\times}10^{-99}$	–	$1{\times}10^{-10}$

Tabelle 16.2: Übersicht praxisrelevanter Gleitpunkt-Zahlensysteme

16.4 Runden, Abschneiden

Ein erster Schritt bei der Durchführung von Algorithmen besteht in der Approximation reeller Zahlen durch Elemente aus dem Gleitpunkt-Zahlensystem $\mathbb{F}$. In den folgenden Abschnitten 16.4.1 und 16.4.2 werden hierzu zwei Möglichkeiten vorgestellt.

16.4.1 Runden

Die erste Variante zur Approximation reeller Zahlen aus dem Überdeckungsbereich eines gegebenen Gleitpunkt-Zahlensystems $\mathbb{F}$ liefert die folgende Definition:

Definition 16.20. Zu einem gegebenen Gleitpunkt-Zahlensystem $\mathbb{F} = \mathbb{F}(b, t, e_{min}, e_{max})$ mit b gerade ist die Funktion $\mathrm{rd} : \{x \in \mathbb{R} : x_{min} \le |x| \le x_{max}\} \to \mathbb{R}$ folgendermaßen erklärt,

$$\mathrm{rd}(x) = \left\{ \begin{array}{ll} \sigma\big(\sum_{k=1}^{t} a_k b^{-k}\big) b^e, & \text{falls } a_{t+1} \le \frac{b}{2} - 1 \\ \sigma\big(\text{—} \text{ « } \text{—} + b^{-t}\big) b^e, & \text{falls } a_{t+1} \ge \frac{b}{2} \end{array} \right\} \text{ für } x = \sigma\Big(\sum_{k=1}^{\infty} a_k b^{-k}\Big) b^e \tag{16.6}$$

mit einer normalisierten Darstellung für x entsprechend Theorem 16.2. Man bezeichnet $\mathrm{rd}(x)$ als den *auf* t *Stellen gerundeten Wert von* x.

Beispiel 16.21. Bezüglich der Basis $b = 10$ und der Mantissenlänge $t = 3$ gilt die Identität $\mathrm{rd}(0.9996) = 1.0 = 0.1{\times}10^1$. Dies verdeutlicht noch, dass sich beim Runden alle Ziffern ändern können. △

Der Rundungsprozess liefert das nächstliegende Element aus dem System $\mathbb{F}$:

Theorem 16.22. *Zu einem gegebenen Gleitpunkt-Zahlensystem* $\mathbb{F} = \mathbb{F}(b, t, e_{min}, e_{max})$ *gilt für jede Zahl* $x \in \mathbb{R}$ *mit* $x_{min} \le |x| \le x_{max}$ *die Eigenschaft* $\mathrm{rd}(x) \in \mathbb{F}$*, mit der Minimaleigenschaft* $|\mathrm{rd}(x) - x| = \min_{z \in \mathbb{F}} |z - x|$.

BEWEIS. Ausgehend von der Notation $x = \sigma(\sum_{k=1}^{\infty} a_k b^{-k})b^e$ erhält man durch elementare Abschätzungen die folgenden unteren und oberen Schranken für den Ausdruck $\sum_{k=1}^{\infty} a_k b^{-k}$:

$$\sum_{k=1}^{t} a_k b^{-k} \;\le\; \sum_{k=1}^{\infty} a_k b^{-k} \;\le\; \sum_{k=1}^{t} a_k b^{-k} + \overbrace{\sum_{k=t+1}^{\infty} \underbrace{(b-1)b^{-k}}_{b^{-k+1}-b^{-k}}}^{=\,b^{-t}}$$

und daraus folgt

$$\underbrace{\Big(\sum_{k=1}^{t} a_k b^{-k}\Big)}_{\le\, b^{-1}} b^e \;\le\; |x| \;\le\; \underbrace{\Big(\sum_{k=1}^{t} a_k b^{-k} + b^{-t}\Big)}_{\le\, 1} b^e.$$

Daher liegen die Schranken in dem Intervall $[b^{e-1}, b^e]$, so dass die beiden für $\mathrm{rd}(x)$ infrage kommenden Werte $\sigma\big(\sum_{k=1}^{t} a_k b^{-k}\big)b^e$ und $\sigma\big(\sum_{k=1}^{t} a_k b^{-k} + b^{-t}\big)b^e$ nach Theorem 16.8 die Nachbarn von x aus dem Gleitpunkt-Zahlensystem $\mathbb{F}$ darstellen. Daraus resultiert insbesondere $\mathrm{rd}(x) \in \mathbb{F}$, und im Folgenden wird die Ungleichung

$$|\mathrm{rd}(x) - x| \;\le\; b^{-t+e}/2 \tag{16.7}$$

nachgewiesen, wobei die obere Schranke in der Abschätzung (16.7) die Hälfte des Abstands der beiden Nachbarn zueinander darstellt, so dass (16.7) die behauptete Optimalität nach sich zieht. Zum Beweis von (16.7) unterscheidet man zwei Situationen. Im Fall "$a_{t+1} \le b/2 - 1$" berechnet man

$$\begin{aligned}
|\mathrm{rd}(x) - x| \;&=\; \Big(\sum_{k=t+1}^{\infty} a_k b^{-k}\Big)b^e \;=\; \big(a_{t+1}b^{-(t+1)} + \sum_{k=t+2}^{\infty} a_k b^{-k}\big)b^e \\
&\le\; \Big[\; \big(\tfrac{b}{2} - 1\big)b^{-(t+1)} \;+\; \sum_{k=t+2}^{\infty} \overbrace{(b-1)b^{-k}}^{b^{-k+1}-b^{-k}} \;\Big]b^e \\
&=\; \big[\; \text{——} \,\ll\, \text{——} \;+\; b^{-(t+1)}\big]b^e \;=\; \tfrac{1}{2}b^{-t+e},
\end{aligned}$$

und in der Situation "$a_{t+1} \ge b/2$" erhält man

$$\begin{aligned}
|\mathrm{rd}(x) - x| \;&=\; \Big(b^{-t} - \sum_{k=t+1}^{\infty} a_k b^{-k}\Big)b^e \;=\; \Big(b^{-t} - \overbrace{a_{t+1}b^{-(t+1)}}^{\ge\, b^{-t}/2} - \overbrace{\sum_{k=t+2}^{\infty} a_k b^{-k}}^{\ge\, 0}\Big)b^e \\
&\le\; \tfrac{1}{2}b^{-t+e}.
\end{aligned}$$

Aus diesen Abschätzungen schließlich erhält man die Ungleichung (16.7). □

Die Situation beim Runden ist in Bild 16.4 veranschaulicht.

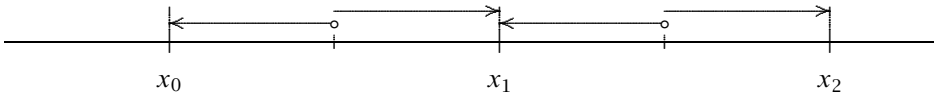


Bild 16.4: Es stellen $x_0,\ x_1$ und x_2 benachbarte Zahlen aus dem System $\mathbb{F}$ dar. Die Pfeile kennzeichnen jeweils Bereiche, aus denen nach $x_0,\ x_1$ beziehungsweise nach x_2 gerundet wird.

Als leichte Folgerung aus Theorem 16.22 erhält man das folgende Resultat.

Korollar 16.23. *In einem gegebenen Gleitpunkt-Zahlensystem* $\mathbb{F} = \mathbb{F}(\mathrm{b}, \mathrm{t}, \mathrm{e}_{\min}, \mathrm{e}_{\max})$ *gilt für jede Zahl* $x \in \mathbb{R}$ *mit* $x_{\min} \le |x| \le x_{\max}$ *die folgende Abschätzung für den relativen Rundungsfehler,*

$$\frac{|\mathrm{rd}(x) - x|}{|x|} \le \overbrace{\mathrm{b}^{-\mathrm{t}+1}/2}^{=\ \mathrm{eps}} \qquad \textit{für} \quad x \in \mathbb{R} \quad \textit{mit} \quad x_{\min} \le |x| \le x_{\max}. \tag{16.8}$$

Eine alternative Fehlerdarstellung ist

$$\mathrm{rd}(x) = x + \Delta x \qquad \textit{für ein} \ \Delta x \in \mathbb{R} \quad \textit{mit} \quad \frac{|\Delta x|}{|x|} \le \mathrm{eps}. \tag{16.9}$$

BEWEIS. Die Abschätzung (16.8) folgt aus dem Beweis von Theorem 16.22 oder direkt aus Theorem 16.10. Die Darstellung (16.9) ergibt sich mit der Setzung $\Delta x := \mathrm{rd}(x) - x$ unmittelbar aus der Abschätzung (16.8). □

Bemerkung 16.24. Auch auf dem Intervall $(-x_{\min}, x_{\min})$ stellt (16.6) eine sinnvolle (und dem IEEE-Standard 754 entsprechende) Definition für die Funktion rd dar, wenn man in (16.6) die normalisierte Darstellung für x ersetzt durch $x = \sigma(\sum_{k=1}^{\infty} a_k \mathrm{b}^{-k})\, \mathrm{b}^{\mathrm{e}_{\min}}$ mit $a_1 = 0$. Tatsächlich gilt $\mathrm{rd}(x) \in \widehat{\mathbb{F}}$ und $|\mathrm{rd}(x) - x| = \min_{z \in \widehat{\mathbb{F}}} |z - x|$ für $x \in (-x_{\min}, x_{\min})$, jedoch verliert die Aussage von Korollar 16.23 über den relativen Rundungsfehler für solche Werte von x ihre Gültigkeit, was unmittelbar aus Bemerkung 16.13 folgt. Der Fall $|x| > x_{\max}$ führt im IEEE-Standard 754 zu einem Overflow, genauer zu $\mathrm{rd}(x) = \infty$ beziehungsweise $\mathrm{rd}(x) = -\infty$. △

16.4.2 Abschneiden

Ein einfache Alternative zum Runden stellt das Abschneiden (english: truncate) dar:

Definition 16.25. Zu einem gegebenen Gleitpunkt-Zahlensystem $\mathbb{F} = \mathbb{F}(\mathrm{b}, \mathrm{t}, \mathrm{e}_{\min}, \mathrm{e}_{\max})$ ist die Funktion $\mathrm{tc} : \{x \in \mathbb{R} : x_{\min} \le |x| \le x_{\max}\} \to \mathbb{R}$ folgendermaßen erklärt,

$$\mathrm{tc}(x) = \sigma\Big(\sum_{k=1}^{\mathrm{t}} a_k \mathrm{b}^{-k}\Big)\mathrm{b}^{\mathrm{e}} \qquad \text{für} \quad x = \sigma\Big(\sum_{k=1}^{\infty} a_k \mathrm{b}^{-k}\Big)\mathrm{b}^{\mathrm{e}}.$$

Es wird $\mathrm{tc}(x)$ als die *auf* t *Stellen abgeschnittene Zahl* x bezeichnet. Die Situation beim Abschneiden ist in Bild 16.5 veranschaulicht.

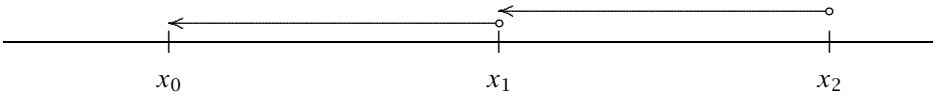


Bild 16.5: Es stellen x_0, x_1 und x_2 benachbarte Zahlen aus dem System $\mathbb{F}$ dar. Die Pfeile kennzeichnen jeweils Bereiche, aus denen nach x_0 beziehungsweise nach x_1 abgeschnitten wird.

Beispiel 16.26. Für die Basis $\mathrm{b} = 10$ und die Mantissenlänge $\mathrm{t} = 3$ gilt die Identität $\mathrm{tc}(0.9996) = 0.999{\times}10^0$. △

Theorem 16.27. *Zu einem gegebenen Gleitpunkt-Zahlensystem* $\mathbb{F} = \mathbb{F}(\mathrm{b}, \mathrm{t}, \mathrm{e}_{\min}, \mathrm{e}_{\max})$ *gelten für jede Zahl* $x \in \mathbb{R}$ *mit* $x_{\min} \le |x| \le x_{\max}$ *die Eigenschaft* $\mathrm{tc}(x) \in \mathbb{F}$ *und die folgende Fehlerabschätzung,*

$$\frac{|\mathrm{tc}(x) - x|}{|x|} \;\le\; \underbrace{2\mathrm{eps}}_{\mathrm{b}^{-\mathrm{t}+1}} \qquad \textit{für} \quad x \in \mathbb{R} \quad \textit{mit} \quad x_{\min} \le |x| \le x_{\max}. \tag{16.10}$$

Eine alternative Fehlerdarstellung ist

$$\mathrm{tc}(x) \;=\; x \,+\, \Delta x \qquad \textit{für ein } \Delta x \in \mathbb{R} \quad \textit{mit} \quad \frac{|\Delta x|}{|x|} \le 2\mathrm{eps}. \tag{16.11}$$

BEWEIS. Für eine beliebige Zahl $x \in \mathbb{R}$ mit $x_{\min} \le |x| \le x_{\max}$ weist man die Eigenschaft $\mathrm{tc}(x) \in \mathbb{F}$ entsprechend der Vorgehensweise im Beweis von Theorem 16.22 nach, und mit der Darstellung $x = \sigma(\sum_{k=1}^{\infty} a_k \mathrm{b}^{-k})\mathrm{b}^{\mathrm{e}}$ erhält man die Abschätzung (16.10) leicht durch

$$|\mathrm{tc}(x) - x| \;=\; \Big(\sum_{k=\mathrm{t}+1}^{\infty} a_k \mathrm{b}^{-k}\Big)\mathrm{b}^{\mathrm{e}} \;\le\; \Big(\sum_{k=\mathrm{t}+1}^{\infty} \overbrace{(\mathrm{b}-1)\mathrm{b}^{-k}}^{\mathrm{b}^{-k+1}-\mathrm{b}^{-k}}\Big)\mathrm{b}^{\mathrm{e}} \;=\; \mathrm{b}^{-\mathrm{t}+\mathrm{e}}$$

sowie der Eigenschaft $|x| \ge \mathrm{b}^{b-1}$. Die Darstellung (16.11) resultiert mit der Setzung $\Delta x := \mathrm{tc}(x) - x$ unmittelbar aus der Abschätzung (16.10). □

Bemerkung 16.28. Die Aussagen aus Bemerkung 16.24 lassen sich für die Abschneidefunktion tc übertragen. △

16.5 Arithmetik in Gleitpunkt-Zahlensystemen

In den folgenden Abschnitten werden arithmetische Grundoperationen in Gleitpunkt-Zahlensystemen vorgestellt und Abschätzungen für den bei der Hintereinanderausführung solcher Operationen entstehenden Gesamtfehler hergeleitet.

16.5.1 Arithmetische Grundoperationen in Gleitpunkt-Zahlensystemen

In einem gegebenen Gleitpunkt-Zahlensystem $\mathbb{F} = \mathbb{F}(b, t, e_{\min}, e_{\max})$ sehen naheliegende Realisierungen von Grundoperationen $\circ \in \{+, -, \times, /\}$ zum Beispiel so aus,

$$x \circ^* y = \mathrm{rd}(x \circ y) \quad \text{für } x, y \in \mathbb{F} \quad \text{mit } x_{\min} \le |x \circ^* y| \le x_{\max}, \tag{16.12}$$

$$\text{oder } x \circ^* y = \mathrm{tc}(x \circ y) \quad \text{——— « ———} \tag{16.13}$$

wobei im Fall der Division $y \neq 0$ angenommen ist.

Bemerkung 16.29. (a) Man beachte, dass für Operationen von der Gestalt (16.12) oder (16.13) sowohl Assoziativ- als auch Distributivgesetze keine Gültigkeit besitzen.

(b) Praktisch lassen sich (16.12) beziehungsweise (16.13) so realisieren, dass man zu gegebenen Zahlen $x, y \in \mathbb{F}$ anstelle des exakten Wertes $x \circ y$ eine Approximation $z \approx x \circ y \in \mathbb{R}$ mit $\mathrm{rd}(z) = \mathrm{rd}(x \circ y)$ beziehungsweise $\mathrm{tc}(z) = \mathrm{tc}(x \circ y)$ bestimmt. △

Für die folgenden Betrachtungen wird lediglich die Annahme getroffen, dass der bei arithmetischen Grundoperationen in Gleitpunkt-Zahlensystemen auftretende relative Fehler dieselbe Größenordnung wie der relative Rundungsfehler besitzt, eine weitere Spezifikation ist nicht erforderlich.

Definition 16.30. Zu einem gegebenen Gleitpunkt-Zahlensystem $\mathbb{F} = \mathbb{F}(b, t, e_{\min}, e_{\max})$ bezeichnen im Folgenden $+^*$, $-^*$, $\times^*$, $/^*$ Operationen mit den Eigenschaften

$$x \circ^* y \in \mathbb{F}, \quad x \circ^* y = x \circ y + \eta \quad \text{für ein } \eta \in \mathbb{R}, \quad \frac{|\eta|}{|x \circ y|} \le K \,\mathrm{eps} \tag{16.14}$$

$$\big(x, y \in \mathbb{F} \quad \text{mit} \quad x_{\min} \le |x \circ y| \le x_{\max}, \qquad \circ \in \{+, -, \times, /\}\big),$$

wobei im Fall der Division $y \neq 0$ angenommen ist, und $K \ge 0$ ist eine Konstante.

In den Fällen (16.12) beziehungsweise (16.13) gilt (16.14) mit $K = 1$ beziehungsweise $K = 2$. In den beiden nächsten Abschnitten werden Abschätzungen für den akkumulierten Fehler bei der Hintereinanderausführung von Grundoperationen in Gleitpunkt-Zahlensystemen hergeleitet.

16.5.2 Fehlerakkumulation bei der Hintereinanderausführung von Multiplikationen und Divisionen in Gleitpunkt-Zahlensystemen

Das folgende Lemma wird benötigt beim Beweis des darauf folgenden Theorems über die Fehlerausbreitung bei der Hintereinanderausführung von Multiplikationen und Divisionen in Gleitpunkt-Zahlensystemen.

Lemma 16.31. *Für Zahlen $\tau_1, \ldots, \tau_n \in \mathbb{R}$ mit $|\tau_k| \le \varepsilon$ für $k = 1, 2, \ldots, n$, und für Exponenten $\sigma_1, \sigma_2, \ldots, \sigma_n \in \{-1, 1\}$ gilt in der Situation $n\varepsilon < 1$ Folgendes,*

$$\prod_{k=1}^{n} (1 + \tau_k)^{\sigma_k} = 1 + \beta_n \qquad \textit{mit} \quad |\beta_n| \le \frac{n\varepsilon}{1 - n\varepsilon}. \tag{16.15}$$

BEWEIS. Es wird ein Induktionsbeweis über n geführt, und hierzu seien vorbereitend die folgenden elementaren Abschätzungen angegeben,

$$|(1+\tau_k)^{\sigma_k}| \le \frac{1+\varepsilon}{1-\varepsilon}, \qquad |(1+\tau_k)^{\sigma_k} - 1| \le \frac{\varepsilon}{1-\varepsilon} \quad \text{für } k = 1,2,\ldots,n. \tag{16.16}$$

Die zweite Abschätzung in (16.16) liefert den Induktionsanfang $n = 1$ für (16.15), und im Folgenden wird der Induktionsschritt "$n \to n+1$" geführt. Hierzu schreibt man

$$\prod_{k=1}^{n+1}(1+\tau_k)^{\sigma_k} - 1 = (1+\tau_{n+1})^{\sigma_{n+1}}\Big(\prod_{k=1}^{n}(1+\tau_k)^{\sigma_k} - 1\Big) + (1+\tau_{n+1})^{\sigma_{n+1}} - 1$$

und schätzt dann mit (16.15) und der Induktionsannahme folgendermaßen ab,

$$\begin{aligned}\Big|\prod_{k=1}^{n+1}(1+\tau_k)^{\sigma_k} - 1\Big| &\le \frac{1+\varepsilon}{1-\varepsilon}\,\frac{n\varepsilon}{1-n\varepsilon} + \frac{\varepsilon}{1-\varepsilon} \\ &= \frac{1}{1-\epsilon}\,\frac{n\varepsilon + n\varepsilon^2 + \varepsilon - n\varepsilon^2}{1-n\varepsilon} = \frac{1}{1-\epsilon}\,\frac{(n+1)\varepsilon}{1-n\varepsilon} \\ &= \frac{(n+1)\varepsilon}{1-(n+1)\varepsilon+n\varepsilon^2} \le \frac{(n+1)\varepsilon}{1-(n+1)\varepsilon},\end{aligned}$$

so dass die Darstellung für den Fall $n+1$ bewiesen und der Induktionsschritt damit abgeschlossen ist. □

Theorem 16.32. *Zu einem gegebenen Gleitpunkt-Zahlensystem* $\mathbb{F} = \mathbb{F}(\mathrm{b}, \mathrm{t}, \mathrm{e}_{\min}, \mathrm{e}_{\max})$ *seien Zahlen* $x_1, x_2, \ldots, x_n \in \mathbb{R}$ *und* $\Delta x_1, \Delta x_2, \ldots, \Delta x_n \in \mathbb{R}$ *gegeben mit*

$$x_k + \Delta x_k \in \mathbb{F}, \qquad \frac{|\Delta x_k|}{|x_k|} \le K\,\mathrm{eps} \quad \textit{für } k = 1,2,\ldots,n, \tag{16.17}$$

mit $(n-1)K\,\mathrm{eps} < 1/4$. *Weiter sei für Grundoperationen* $\circ_1, \ldots, \circ_{n-1} \in \{\times, /\}$ *die Eigenschaft (16.14) sowie* $x_{\min} \le |x_1 \circ_1 \ldots \circ_j x_j| \le x_{\max}$ *für* $j = 2,\ldots,n-1$ *erfüllt, wobei jeweils noch ein gewisser Abstand zu den Intervallrändern* $x_{\min}$ *und* $x_{\max}$ *gegeben sei*[2]. *Dann gilt die Fehlerdarstellung*

$$\begin{aligned}&(x_1 + \Delta x_1) \circ_1^* (x_2 + \Delta x_2) \circ_2^* \ldots \circ_{n-1}^* (x_n + \Delta x_n) \\ &\quad = x_1 \circ_1 x_2 \circ_2 \ldots \circ_{n-1} x_n + \eta, \\ &\qquad \textit{mit } \frac{|\eta|}{|x_1 \circ_1 \ldots \circ_{n-1} x_n|} \le \frac{(2n-1)K\,\mathrm{eps}}{1-(2n-1)K\,\mathrm{eps}}.\end{aligned}$$

BEWEIS. Ausgehend von der Fehlerdarstellung

$$x_k + \Delta x_k = x_k(1+\tau_k) \quad \text{mit } |\tau_k| \le K\,\mathrm{eps}, \qquad \text{für } k = 1,2,\ldots,n,$$

[2] Diese Bedingung wird in (16.19) im Beweis präzisiert.

berechnet man unter Anwendung von (16.14)

$$\begin{aligned}(x_1 + \Delta x_1) \circ_1^* (x_2 + \Delta x_2) &= (x_1(1+\tau_1)) \circ_1^* (x_2(1+\tau_2)) \\ &= (x_1 \circ_1 x_2)\big((1+\tau_1) \circ_1 (1+\tau_2)\big)(1+\alpha_1) \\ &\qquad \text{mit } |\alpha_1| \le K\,\mathrm{eps},\end{aligned}$$

und mit einer entsprechenden Vorgehensweise erhält man sukzessive die Darstellungen

$$\left.\begin{aligned}&(x_1 + \Delta x_1) \circ_1^* (x_2 + \Delta x_2) \circ_2^* \cdots \circ_{j-1}^* (x_j + \Delta x_j) \\ &= (x_1 \circ_1 x_2 \circ_2 \ldots \circ_{j-1} x_j)(1+\beta_{2j-1}) \\ &\text{mit } 1+\beta_{2j-1} = (1+\tau_1) \circ_1 (1+\tau_2) \circ_2 \cdots \circ_{j-1} (1+\tau_j) \prod_{k=1}^{j-1}(1+\alpha_k),\end{aligned}\right\} \quad (16.18)$$

für $j = 2, 3, \ldots, n$, mit $|\alpha_k| \le K\,\mathrm{eps}$ für alle k. Die Anwendbarkeit der Eigenschaft (16.14) wird zum Beispiel durch die Bedingung

$$\frac{1-(2n-2)K\,\mathrm{eps}}{1-(4n-4)K\,\mathrm{eps}} x_{\min} \le |x_1 \circ_1 \ldots \circ_{j-1} x_j| \le \big(1-(2n-2)K\,\mathrm{eps}\big)x_{\max}, \quad (16.19)$$

gewährleistet, denn sie zusammen mit Lemma 16.31 impliziert, dass die Resultate der Multiplikationen und Divisionen in dem Gleitpunkt-Zahlensystem allesamt in dem relevanten Bereich $\{x \in \mathbb{R} : x_{\min} \le |x| \le x_{\max}\}$ enthalten sind. Aus der Darstellung (16.18) folgt unter nochmaliger Anwendung von Lemma 16.31 die Aussage des Theorems. □

Bemerkung 16.33. (a) Theorem 16.32 impliziert die Gutartigkeit von Multiplikationen und Divisionen in Gleitpunkt-Zahlensystemen, relative Eingangsfehler werden nicht übermäßig verstärkt.

(b) Falls in der Situation von Theorem 16.32 etwa die Ungleichung $(2n-1)K\,\mathrm{eps} < 0.1 \le 1$ erfüllt ist, so gilt

$$\frac{|\eta|}{|x_1 \circ_1 \ldots \circ_{n-1} x_n|} \le \frac{(2n-1)K\,\mathrm{eps}}{0.9} \le (1.12K\,\mathrm{eps})(2n-1).$$

Mit jeder zusätzlichen maschinenarithmetischen Multiplikation oder Division kann sich also eine 12-prozentige Fehlerverstärkung einstellen. △

16.5.3 Fehlerverstärkung bei der Hintereinanderausführung von Additionen in einem gegebenen Gleitpunkt-Zahlensystem $\mathbb{F}$

Das folgende Theorem befasst sich mit der möglichen Fehlerverstärkung bei der Hintereinanderausführung von Additionen und Subtraktionen in einem gegebenen Gleitpunkt-Zahlensystem $\mathbb{F} = \mathbb{F}(\mathrm{b}, \mathrm{t}, \mathrm{e}_{\min}, \mathrm{e}_{\max})$. Dabei werden beliebige Vorzeichen zugelassen, so dass man sich auf die Betrachtung von Additionen beschränken kann. Erläuterungen zur Abschätzung (16.20) finden Sie in der darauf folgenden Bemerkung 16.35.

Theorem 16.34. *Zu einem gegebenen Gleitpunkt-Zahlensystem* $\mathbb{F} = \mathbb{F}(\mathrm{b}, \mathrm{t}, \mathrm{e}_{\min}, \mathrm{e}_{\max})$ *seien* $x_1, x_2, \ldots, x_n \in \mathbb{R}$ *und* $\Delta x_1, \Delta x_2, \ldots, \Delta x_n \in \mathbb{R}$ *Zahlen mit der Eigenschaft (16.17), und es bezeichne*

$$S_k^* := \sum_{j=1}^{k}{}^{*}(x_j + \Delta x_j), \qquad S_k := \sum_{j=1}^{k} x_j \qquad \text{für } k = 1, 2, \ldots, n,$$

wobei $\sum^*$ *für eine Hintereinanderausführung von Additionen in* $\mathbb{F}$ *von links nach rechts steht. Dann gilt die folgende Fehlerabschätzung,*

$$|S_k^* - S_k| \le \underbrace{\Big(\sum_{j=1}^{k} (1+\mathrm{eps})^{k-j} \big(2|x_j| + |S_j|\big) \Big)}_{=:\, M_k} \mathrm{eps} \quad \text{für } k = 1, 2, \ldots, n, \tag{16.20}$$

falls noch (mit der Notation $M_0 = 0$*) die Partialsummen innerhalb gewisser Schranken liegen:*

$$x_{\min} + (M_{k-1} + |x_k|)\mathrm{eps} \le |S_k| \le x_{\max} - (M_{k-1} + |x_k|)\mathrm{eps}, \quad k = 1, 2, \ldots, n. \tag{16.21}$$

BEWEIS. Es wird die Abschätzung (16.20) per Induktion über k bewiesen. Die Aussage in (16.20) ist sicher richtig für $k = 1$, und im Folgenden sei angenommen, dass sie für ein $k \ge 1$ richtig ist. Mit der Notation

$$\Delta S_j := S_j^* - S_j \qquad \text{für } j \ge 1, \qquad \Delta S_0 = 0,$$

berechnet man mit einer gewissen Zahl $\tau_k \in \mathbb{R}$, $|\tau_k| \le \mathrm{eps}$, Folgendes,

$$\begin{aligned} \Delta S_k &= S_k^* - S_k = S_{k-1}^* +^* (x_k + \Delta x_k) - S_k \\ &= (S_{k-1} + \Delta S_{k-1}) +^* (x_k + \Delta x_k) - S_k \\ &\overset{(*)}{=} (S_k + \Delta S_{k-1} + \Delta x_k)(1+\tau_k) - S_k \\ &= (1+\tau_k)\Delta S_{k-1} + \tau_k S_k + (1+\tau_k)\Delta x_k \end{aligned}$$

und daher

$$|\Delta S_k| \le (1+\mathrm{eps})|\Delta S_{k-1}| + \mathrm{eps}(|S_k| + 2|x_k|). \tag{16.22}$$

Die Identität $(*)$ folgt hierbei aus der Eigenschaft (16.14), wobei die Resultate der Additionen in dem Gleitpunkt-Zahlensystem aufgrund der Annahme (16.21) allesamt in dem relevanten Bereich $\{x \in \mathbb{R} : x_{\min} \le |x| \le x_{\max}\}$ enthalten sind. Die Aussage dieses Theorems ist nun eine unmittelbare Konsequenz aus der Abschätzung (16.22) und der Induktionsannahme. □

Bemerkung 16.35. (a) Der Faktor $(1 + \text{eps})^{k-j}$ in der Abschätzung (16.20) ist umso größer, je kleiner k ist. Daher wird man vernünftigerweise beim Aufsummieren mit den betragsmäßig kleinen Zahlen beginnen. Dies gewährleistet zudem, dass die Partialsummen S_k betragsmäßig nicht unnötig anwachsen.

(b) Theorem 16.34 liefert lediglich eine Abschätzung für den *absoluten* Fehler. Der *relative* Fehler $|S_n^* - S_n|/|S_n|$ jedoch kann groß ausfallen, falls $|S_n|$ klein gegenüber $\sum_{j=1}^{n-1}(|x_j| + |S_j|) + |x_n|$ ist. △

Weitere Themen und Literaturhinweise

Eine ausführliche Behandlung von Gleitpunkt-Zahlensystemen und der Grundarithmetiken finden Sie etwa in Überhuber [106] (Band 1), Goldberg [34] oder in Higham [55]. Insbesondere in [106] werden viele weitere interessante Themen wie beispielsweise spezielle Summationsalgorithmen für Gleitpunktzahlen, numerische Softwarepakete, die Anzahl der benötigten Taktzyklen zur Durchführung der vier Grundoperationen $+, -, \times, /$, die asymptotische Komplexität von Algorithmen und die konkrete Implementierung von arithmetischen Operationen behandelt. Dass letztere nicht immer einwandfrei verläuft, zeigt sich am Beispiel der fehlerhaften Pentium-Chips im Jahr 1994 (Moler [72]).

Literaturverzeichnis

[1] ASHBY, S. F., T. A. MANTEUFFEL und P. SAYLOR: *A taxomony for conjugate gradient methods.* SIAM J. Numer. Anal., 27(6):1542–1568, 1990.

[2] BÄRWOLFF, G.: *Numerik für Ingenieure, Physiker und Informatiker.* Elsevier, München, 2007.

[3] BAUMEISTER, J.: *Stable Solution of Inverse Problems.* Vieweg, Braunschweig/Wiesbaden, 1987.

[4] BERMAN, A. und R. PLEMMONS: *Nonnegative Matrices in the Mathematical Sciences.* SIAM, Philadelphia, 1. Auflage, Reprint, 1994.

[5] BOOR, C. DE: *A Practical Guide to Splines.* Springer, Heidelberg, Berlin, 1978.

[6] BOLLHÖFER, M. und V. MEHRMANN: *Numerische Mathematik. Eine projektorientierte Einführung für Ingenieure, Mathematiker und Naturwissenschaftler.* Vieweg, Wiesbaden, 2004.

[7] BRAESS, D.: *Finite Elemente.* Springer, Berlin, Heidelberg, New York, 3. Auflage, 2003.

[8] BRENAN, K. E., S. L. CAMPBELL und L. R. PETZOLD: *Numerical Solution of Initial-Value Problems in Differential-Algebraic Equations.* SIAM, Philadelphia, 1. Auflage, Reprint, 1996.

[9] BULIRSCH, R.: *Bemerkungen zur Romberg-Integration.* Numer. Math., 6:6–16, 1964.

[10] BULIRSCH, R. und J. STOER: *Numerical treatment of ordinary differential equations by extrapolation methods.* Numer. Math., 8:1–13, 1966.

[11] BUNSE, W. und A. BUNSE-GERSTNER: *Numerische Mathematik.* Teubner, Stuttgart, 1985.

[12] COOLEY, J. W. und J. W. TUKEY: *An algorithm for the machine calculation of complex Fourier series.* Math. of Computations, 19:297–301, 1965.

[13] DABERKOW, M., C. FIEKER, J. KLÜNERS, M. POHST, K. ROEGNER, M. SCHÖRNIG und K. WILDANGER: *KANT V4.* J. Symbolic Computation, 24:267–283, 1997.

[14] DAHLQUIST, G.: *Stability and error bounds in the numerical integration of ordinary differential equations.* Transactions of the Royal Institute of Technology, Stockholm, 130, 1959.

[15] DALLMANN, H. und K.-H. ELSTER: *Einführung in die höhere Mathematik III.* Gustav Fischer Verlag, Jena, 2. Auflage, 1992.

[16] DEKKER, K. und J. G. VERWER: *Stability of Runge-Kutta methods for stiff nonlinear differential equations.* North-Holland, Amsterdam, 1984.

[17] DENNIS, J. E. und R. B. SCHNABEL: *Numerical Methods for Unconstrained Optimization and Nonlinear Equations.* SIAM, Philadelphia, 1. Auflage, Reprint, 1996.

[18] DEUFLHARD, P.: *Order and step-size control in extrapolation methods.* Numer. Math., 41:399–422, 1983.

[19] DEUFLHARD, P.: *Recent progess in extrapolation methods for ordinary differential equations.* SIAM Review, 27:505–535, 1985.

[20] DEUFLHARD, P.: *Newton Methods for Nonlinear Problems.* Springer, Heidelberg, Berlin, 2004.

[21] DEUFLHARD, P. und F. BORNEMANN: *Numerische Mathematik 2.* de Gruyter, Berlin, 3. Auflage, 2008.

[22] DEUFLHARD, P. und A. HOHMANN: *Numerische Mathematik 1.* de Gruyter, Berlin, 4. Auflage, 2008.

[23] ELMAN, H.: *Iterative methods for linear systems.* In: J. GILBERT UND ANDERE (Herausgeber): *Advances in Numerical Analysis, Vol. III., Proceedings of the fifth summer school in numerical analysis, Lancaster University, UK, 1992*, pp. 69–118, Oxford, 1994. Clarendon Press.

[24] EMMRICH, E.: *Gewöhnliche und Operator-Differentialgleichungen.* Vieweg, Wiesbaden, 2004.

[25] ENGL, H. W., M. HANKE und A. NEUBAUER: *Regularization of Inverse Problems.* Kluwer, Dordrecht, 2. Auflage, 2000.

[26] FINCKENSTEIN, K. GRAF FINCK VON: *Einführung in die Numerische Mathematik, Band 1 und 2.* Carl Hanser Verlag, München, 1977/1978.

[27] FISCHER, B.: *Polynomial Based Iteration Methods for Symmetric Linear Systems.* Wiley-Teubner, Chichester, Stuttgart, 1996.

[28] FISCHER, G.: *Lineare Algebra.* Vieweg, Wiesbaden, 15. Auflage, 2005.

[29] FORSTER, O.: *Analysis 1.* Vieweg/Teubner, Wiesbaden, 9. Auflage, 2008.

[30] FREUND, R. W., G. H. GOLUB und N. M. NACHTIGAL: *Iterative solution of linear systems.* In: *Acta Numerica*, pp. 1–44, Cambridge, 1991. Cambridge Univ. Press.

[31] FREUND, R. W. und R. H. W. HOPPE: *Stoer/Bulirsch: Numerische Mathematik 1.* Springer, Berlin, 10. Auflage, 2007.

[32] GEIGER, C. und C. KANZOW: *Numerische Verfahren zur Lösung unrestringierter Optimierungsaufgaben.* Springer, Heidelberg, Berlin, 2002.

[33] GOERING, H., H. G. ROOS und L. TOBISKA: *Finite-Element-Methode.* Akademie-Verlag, Berlin, 3. Auflage, 1993.

[34] GOLDBERG, D.: *What every computer scientist should know about floating-point arithmetic.* ACM Computer Surveys, 23:5–48, 1991.

[35] GOLUB, G. und C. F. VAN LOAN: *Matrix Computations.* The Johns Hopkins University Press, Baltimore, London, 2. Auflage, 1993.

[36] GOLUB, G. und J. M. ORTEGA: *Wissenschaftliches Rechnen und Differentialgleichungen. Eine Einführung in die Numerische Mathematik.* Heldermann Verlag, Berlin, 1995.

[37] GOLUB, G. und J. M. ORTEGA: *Scientific Computing.* Teubner, Stuttgart, 1996.

[38] GRAGG, W. B.: *On extrapolation algorithms for ordinary initial value problems.* SIAM J. Numer. Anal., 2:384–403, 1965.

[39] GRAMLICH, G. und W. WERNER: *Numerische Mathematik mit Matlab.* dpunkt.verlag, Heidelberg, 2000.

[40] GREENBAUM, A., V. PTÁK und Z. STRAKOŠ: *Any nonincreasing convergence curve is possible for GMRES.* SIAM J. Matrix Anal. Appl., 17(3):465–465, 1996.

[41] GRIGORIEFF, R. D.: *Numerik gewöhnlicher Differentialgleichungen, Band 1 und 2*. Teubner, Stuttgart, 1972/77.

[42] GROETSCH, C. W.: *Inverse Problems in the Mathematical Sciences*. Vieweg, Braunschweig /Wiesbaden, 1993.

[43] GROSSMANN, CH. und H.-G. ROOS: *Numerische Behandlung partieller Differentialgleichungen*. Teubner, Stuttgart, 3. Auflage, 2005.

[44] GROSSMANN, CH. und J. TERNO: *Numerik der Optimierung*. Teubner, Stuttgart, 2. Auflage, 1997.

[45] GÜNTHER, M. und A. JÜNGEL: *Finanzderivate mit MATLAB*. Vieweg, Wiesbaden, 2003.

[46] HACKBUSCH, W.: *Theorie und Numerik elliptischer Differentialgleichungen*. Teubner, Stuttgart, 1986.

[47] HACKBUSCH, W.: *Iterative Lösung großer schwach besetzter Gleichungssysteme*. Teubner, Stuttgart, 1991.

[48] HÄMMERLIN, G. und K.-H. HOFFMANN: *Numerische Mathematik*. Springer, Berlin, 4. Auflage, 1994.

[49] HAIRER, E. und C. LUBICH: *Asymptotic expansion of the global error of fixed-stepsize methods*. Numer. Math., 45:345–360, 1984.

[50] HAIRER, E., S. P. NØRSETT und G. WANNER: *Solving Ordinary Differential Equations I, Nonstiff Problems*. Springer, Berlin, 2. Auflage, 1993.

[51] HAIRER, E. und G. WANNER: *Solving Ordinary Differential Equations II, Stiff Problems*. Springer, Berlin, 2. Auflage, 1996.

[52] HANKE-BOURGEOIS, M.: *Grundlagen der Numerischen Mathematik*. Teubner, Stuttgart, 2. Auflage, 2006.

[53] HESTENES, M.R. und E. STIEFEL: *Method of conjugate gradients for solving linear systems*. J. Res. Nat. Bur. Standards, Sec. B 49:409–432, 1952.

[54] HEUSER, H.: *Gewöhnliche Differentialgleichungen*. Teubner, Stuttgart, 4. Auflage, 2004.

[55] HIGHAM, N.: *Accuracy and Stability of Numerical Algorithms*. SIAM, Philadelphia, 2. Auflage, 2002.

[56] HIRZEBRUCH, F. und W. SCHARLAU: *Einführung in die Funktionalanalysis*. B. I. Wissenschaftsverlag, Mannheim/Wien/Zürich, 1971.

[57] HOFMANN, B.: *Mathematik Inverser Probleme*. Teubner, Stuttgart, Leipzig, 1999.

[58] HORN, R. A. und C. R. JOHNSON: *Matrix Analysis*. Cambridge University Press, Cambridge, 1. Auflage, Reprint, 1994.

[59] JUNG, M. und U. LANGER: *Methode der finiten Elemente für Ingenieure*. Teubner, Stuttgart, 2001.

[60] KELLEY, C. T.: *Iterative Methods for Linear and Nonlinear Equations*. SIAM, Philadelphia, 1995.

[61] KNABNER, P. und L. ANGERMANN: *Numerik partieller Differentialgleichungen*. Springer, Berlin, Heidelberg, 2000.

[62] KOSMOL, P.: *Methoden zur numerischen Behandlung nichtlinearer Gleichungen und Optimierungsaufgaben*. Teubner, Stuttgart, 1989.

[63] KRESS, R.: *Numerical Analysis.* Springer, Berlin, Heidelberg, New York, 1998.

[64] KROMMER, A. und C. ÜBERHUBER: *Computational Integration.* SIAM, Philadelphia, 1998.

[65] LIESEN, J.: *Computable convergence bounds for GMRES.* SIAM J. Matrix Analysis, 21(3):882–903, 2000.

[66] LOUIS, A. K.: *Inverse und schlecht gestellte Probleme.* Teubner, Stuttgart, 1989.

[67] LOZINSKIĬ, S. M.: *Error estimate for numerical integration of ordinary differential equations.* Izv. Vysš. Učebn. Zaved. Matematika, 6(6):52–90, 1958.

[68] MÄRZ, R.: *Numerical methods for differential algebraic equations.* In: ISERLES, A. (Herausgeber): *Acta Numerica Vol. 1*, pp. 141–198, Cambridge, 1992. Cambridge Univ. Press.

[69] MAESS, G.: *Vorlesungen über numerische Mathematik, Band 1 und 2.* Birkhäuser, Basel, 1985/88.

[70] MEISTER, A.: *Numerik linearer Gleichungssysteme.* Vieweg, Wiesbaden, 2. Auflage, 2005.

[71] MENNICKEN, R. und E. WAGENFÜHRER: *Numerische Mathematik, Band 1 und 2.* Vieweg, Braunschweig/Wiesbaden, 1977.

[72] MOLER, C.: *A Tale of Two Numbers.* SIAM News, 28(1):p. 1 and p. 16, 1995.

[73] MORET, I.: *A note on the superlinear convergence of GMRES.* SIAM J. Numer. Analysis, 34(2):513–516, 1997.

[74] NACHTIGAL, N. M., S. C. REDDY und L. N. REDDY: *How fast are nonsymmetric matrix iterations?* SIAM J. Matrix Anal. Appl., 13(3):778–795, 1992.

[75] NASH, S. G. und A. SOFER: *Linear and Nonlinear Programming.* McGraw-Hill, New York, 1996.

[76] NEMIROVSKIĬ, A.S. und B.T. POLYAK: *Iterative methods for solving linear ill-posed problems under precise information I.* Moscow Univ. Comput. Math. Cybern., 22(3):1–11, 1984.

[77] NEVANLINNA, O.: *Convergence of Iterations for Linear Equations.* Birkhäuser, Basel, 1993.

[78] OEVEL, W.: *Einführung in die Numerische Mathematik.* Spektrum, Heidelberg, 1996.

[79] OPFER, G.: *Numerische Mathematik für Anfänger.* Vieweg/Teubner, Wiesbaden, 5. Auflage, 2008.

[80] PAN, V.: *Complexity of computations with matrices and polynomials.* SIAM Review, 34:225–262, 1992.

[81] PARLETT, B. N.: *The Symmetric Eigenvalue Problem.* SIAM, Philadelphia, 1. Auflage, Reprint, 1988.

[82] PLATO, R.: *Übungsbuch zur Numerischen Mathematik.* Vieweg/Teubner, Wiesbaden, 2. Auflage, 2009.

[83] PLATO, R. und G. VAINIKKO: *The fast solution of periodic integral and pseudo-differential equations by GMRES.* Computational Methods in Applied Mathematics, 1(4):383–397, 2001.

[84] POTTS, D., G. STEIDL und M. TASCHE: *Fast Fourier transforms for nonequispaced data: A tutorial.* In: BENEDETTO, J. J. und P. FERREIRA (Herausgeber): *Modern Sampling Theory: Mathematics and Applications*, pp. 253–274, Basel, 2001. Birkhäuser.

[85] REINHARDT, H.-J.: *Numerik gewöhnlicher Differentialgleichungen.* de Gruyter, Berlin, 2008.

[86] RIEDER, A.: *Keine Probleme mit Inversen Problemen.* Vieweg, Wiesbaden, 2003.

[87] ROMBERG, W.: *Vereinfachte numerische Integration.* Det. Kong. Norske Videnskabers Selskab Forhandlinger, 28(7), Trondheim 1955.

[88] ROOS, H.-G. und H. SCHWETLICK: *Numerische Mathematik.* Teubner, Stuttgart, Leipzig, 1999.

[89] SAAD, Y. und M. H. SCHULTZ: *Conjugate gradient-like algorithms for solving nonsymmetric linear systems.* Math. of Comput., 44(170):417–424, 1985.

[90] SAAD, Y. und M. H. SCHULTZ: *GMRES: A generalized minimal residual algorithm for solving nonsymmetric linear systems.* SIAM J. Sci. Stat. Comput., 7(3):856–869, 1986.

[91] SARANEN, J. und G. VAINIKKO: *Periodic Integral and Pseudodifferential Equations with Numerical Approximation.* Springer, Berlin Heidelberg New York, 2001.

[92] SCHABACK, R. und H. WENDLAND: *Numerische Mathematik.* Springer, Berlin, Heidelberg, New York, 5. Auflage, 2004.

[93] SCHWANDT, H.: *Parallele Numerik.* Teubner, Stuttgart, 2003.

[94] SCHWARZ, H. und N. KÖCKLER: *Numerische Mathematik.* Vieweg/Teubner, Wiesbaden, 7. Auflage, 2009.

[95] SCHWETLICK, H.: *Numerische Lösung nichtlinearer Gleichungen.* Oldenbourg, München, 1979.

[96] SCHWETLICK, H. und H. KRETZSCHMAR: *Numerische Verfahren für Naturwissenschaftler und Ingenieure.* Fachbuchverlag Leipzig, 1991.

[97] SONNEVELD, P.: *CGS, a fast Lanczos-type solver for nonsymmetric linear systems.* SIAM J. Sci. Stat. Comput., 10(1):36–52, 1989.

[98] STOER, J.: *Solution of large systems of equations by conjugate gradient type methods.* In: BACHEM, A., M. GRÖTSCHEL und B. KORTE (Herausgeber): *Mathematical Programming. The State of the Art. Bonn 1982*, pp. 540–565, Berlin, New York, 1983. Springer.

[99] STOER, J. und R. BULIRSCH: *Numerische Mathematik 2.* Springer, Berlin, 5. Auflage, 2005.

[100] STRASSEN, V.: *Gaussian elimination is not optimal.* Numer. Math., 13:354–356, 1969.

[101] STREHMEL, K. und R. WEINER: *Numerik gewöhnlicher Differentialgleichungen.* Teubner, Stuttgart, 1995.

[102] STUMMEL, F. und K. HAINER: *Praktische Mathematik.* Teubner, Stuttgart, 1982.

[103] SUTTMEIER, F.-T.: *Numerical Solution of Variational Inequalities by Adaptive Finite Elements.* Vieweg/Teubner, Wiesbaden, 2008.

[104] TREFETHEN, L. N. und D. BAU: *Numerical Linear Algebra.* SIAM, Philadelphia, 1997.

[105] TRÖLTZSCH, F.: *Optimale Steuerung partieller Differentialgleichungen.* Vieweg, Wiesbaden, 2005.

[106] ÜBERHUBER, C. W.: *Numerical Computation, Band 1 und 2.* Springer, Berlin, Heidelberg, 1997.

[107] VORST, H. A. VAN DER und C. VUIK: *The superlinear convergence behaviour of GMRES.* Journal of Computational and Applied Mathematics, 48:327–341, 1993.

[108] VUIK, C. und H. A. VAN DER VORST: *A comparison of some GMRES-like methods.* Linear Algebra and its Applications, 160:131–162, 1992.

[109] WATKINS, D. S.: *Understanding the QR algorithm.* SIAM Review, 24:427–440, 1982.

[110] WELLER, F.: *Numerische Mathematik für Ingenieure und Naturwissenschaftler.* Vieweg, Braunschweig/Wiesbaden, 1996.

[111] WERNER, J.: *Numerische Mathematik, Band 1 und 2.* Vieweg, Braunschweig/Wiesbaden, 1990.

[112] WINDISCH, G.: *M-matrices in Numerical Analysis.* Teubner, Leipzig, 1989.

Index