Martin Hanke-Bourgeois

Grundlagen der Numerischen Mathematik und des Wissenschaftlichen Rechnens

3. Auflage

STUDIUM

VIEWEG+
TEUBNER

Martin Hanke-Bourgeois

Grundlagen der Numerischen Mathematik und des Wissenschaftlichen Rechnens

Aus dem Programm Analysis und Numerik

Jürgen Appell, Martin Väth
Elemente der Funktionalanalysis

Harro Heuser
Funktionalanalysis

Wolfgang Fischer und Ingo Lieb
Funktionentheorie

Lars Grüne, Oliver Junge
Gewöhnliche Differentialgleichungen

Harro Heuser
Gewöhnliche Differentialgleichungen

Günther J. Wirsching
Gewöhnliche Differentialgleichungen

Etienne Emmrich
Gewöhnliche und Operator-Differentialgleichungen

Matthias Bollhöfer, Volker Mehrmann
Numerische Mathematik

Martin Hanke-Bourgeois
Grundlagen der Numerischen Mathematik und des Wissenschaftlichen Rechnens

Gerhard Opfer
Numerische Mathematik für Anfänger

Robert Plato
Numerische Mathematik kompakt

Hans-Rudolf Schwarz, Norbert Köckler
Numerische Mathematik

Andreas Meister
Numerik linearer Gleichungssysteme

Stefan Sauter, Christoph Schwab
Randelementmethoden

Martin Hanke-Bourgeois

Grundlagen der Numerischen Mathematik und des Wissenschaftlichen Rechnens

3., 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.

Prof. Dr. Martin Hanke-Bourgeois
Geboren 1961 in Frankfurt/Main. Von 1980 bis 1987 Studium der Mathematik, 1989 Promotion an der Universität Karlsruhe (TH), 1994 Habilitation. Von 1995 bis 1997 Lehrstuhlvertretung an der Universität Kaiserslautern. Seit 1999 Professor für Angewandte Mathematik an der Johannes Gutenberg-Universität Mainz.

1. Auflage 2002
2., überarbeitete und erweiterte Auflage 2006
3., aktualisierte Auflage 2009

Dieses Werk ist ein Teil der Reihe Mathematische Leitfäden
(herausgegeben von Prof. Dr. Dr. h. c. mult. Gottfried Köthe; Prof. Dr. Klaus-Dieter Bierstedt, Universität Paderborn; Prof. Dr. Günther Trautmann, Universität Kaiserslautern)

Lektorat: Ulrike Schmickler-Hirzebruch | Nastassja Vanselow

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

Umschlaggestaltung: KünkelLopka Medienentwicklung, Heidelberg
Druck und buchbinderische Verarbeitung: STRAUSS GMBH, Mörlenbach
Gedruckt auf säurefreiem und chlorfrei gebleichtem Papier.
Printed in Germany

ISBN 978-3-8348-0708-3

Vorwort

Dieses Buch ist aus mehreren Vorlesungszyklen *Numerische Mathematik* hervorgegangen, die ich an den Universitäten in Karlsruhe, Kaiserslautern und Mainz gehalten habe. Im Gegensatz zu vielen anderen Lehrbüchern enthält es neben den üblichen Algorithmen der numerischen linearen Algebra und der Approximation eine umfassende Einführung in die Verfahren zur Lösung gewöhnlicher und partieller Differentialgleichungen. Dies bietet den Vorteil, daß die bereitgestellten Grundlagen auf die fortgeschritteneren Kapitel abgestimmt sind und die Notationen der einzelnen Kapitel weitgehend übereinstimmen.

Die Numerische Mathematik steht heute mehr denn je in der Verantwortung, sich den vielfältigen mathematischen Herausforderungen aus den Ingenieur- und Naturwissenschaften zu stellen. In diesem Kontext ist die Fähigkeit zur adäquaten mathematischen Modellierung, zur algorithmischen Umsetzung und zur effizienten Implementierung gefordert, was häufig mit dem Schlagwort *Wissenschaftliches Rechnen* umschrieben wird. Diese Entwicklung darf sich nicht allein auf die Forschung beschränken, der Anwendungscharakter muß vielmehr schon in der Lehre betont werden, damit die Studierenden auf die entsprechenden beruflichen Anforderungen vorbereitet werden. Aus diesem Grund enthält das Buch zahlreiche konkrete Anwendungen der jeweiligen Lehrinhalte.

An vielen Hochschulen werden numerische Verfahren zur Lösung von Differentialgleichungen unterrichtet, bevor die Studierenden Gelegenheit haben, die Lösungstheorie dieser Gleichungen kennenzulernen. Um dennoch ein tiefergehendes Verständnis zu ermöglichen, enthält das Buch drei Modellierungskapitel, die die wichtigsten Differentialgleichungen einführen. In meinen eigenen Vorlesungen habe ich solche Beispiele den jeweiligen Differentialgleichungskapiteln vorangestellt. Alternativ kann dieser Teil jedoch auch für eine unabhängige Modellierungsvorlesung oder ein ergänzendes Seminar verwendet werden.

Damit der Umfang des Buchs in einem vertretbaren Rahmen blieb, mußte an anderer Stelle gekürzt werden. Bisweilen habe ich deshalb bei der Auswahl des Stoffs auf gängige Resultate verzichtet, wenn mir ihr Anwendungsbezug nicht oder nicht mehr relevant zu sein schien. Natürlich sind diese Entscheidungen subjektiv gefärbt und lassen sich kontrovers diskutieren.

Die wichtigsten Algorithmen sind in einem Pseudocode formuliert, der sich an der Programmierumgebung MATLAB® orientiert.[1] MATLAB bietet den Vorteil, daß auch komplexere Algorithmen relativ schnell programmiert werden können. Dieser Zeitgewinn kann genutzt werden, um die Grenzen numerischer Verfahren experimentell auszutesten. Die Übungsaufgaben zu den einzelnen Kapiteln enthalten entsprechende Programmieraufgaben. MATLAB kann problemlos in den Übungen vorlesungsbegleitend eingeführt werden, als ergänzende Literatur empfehle ich die Bücher von Überhuber und Katzenbeisser [103] oder Higham und Higham [54]. Es sei angemerkt, daß auch fast alle Beispiele und Abbildungen des Buchs mit MATLAB erstellt wurden; The MathWorks, Inc., möchte ich an dieser Stelle für die Unterstützung danken.

Damit bin ich bei den Danksagungen angelangt, doch die Liste der vielen Studierenden, Kollegen und Freunde, die mit Rat und Tat zur Seite gestanden haben, ist derart umfangreich geworden, daß hier nicht alle erwähnt werden können. Ihnen allen ein herzliches Dankeschön. Besonderen Dank schulde ich meinen Mitarbeiter/innen Dr. M. Brühl, M. Geisel und B. Schappel, die unwahrscheinlich viel Zeit in die Korrektur des Manuskripts und die Auswahl der Übungsaufgaben investiert haben. Vor allem Herrn Dr. Brühl möchte ich dafür danken, daß ich in ihm einen Ansprechpartner hatte, den ich jederzeit in – zum Teil erschöpfende – Diskussionen über einzelne Abschnitte des Manuskripts verwickeln konnte. Zudem waren seine LaTeX-Kenntnisse eine unschätzbare Hilfe für mich.

Außerdem möchte ich namentlich den Kollegen M. Eiermann, M. Hochbruck, C. Lubich und C.-D. Munz für ihre Vorlesungsmanuskripte danken, die sich als äußerst hilfreich erwiesen haben. Frau Hochbruck hat darüber hinaus viele Teile des Manuskripts gelesen und in Vorlesungen „erprobt“. Ihre Ratschläge waren ausgesprochen hilfreich.

In die zweite und dritte Auflage wurden neben einigen neuen Aufgaben zahlreiche Korrekturen und kleinere Verbesserungen aufgenommen. Ich danke für die zahlreichen Vorschläge und Hinweise, die diesbezüglich an mich herangetragen worden sind. Über die sehr positiven Reaktionen auf dieses Buch habe ich mich sehr gefreut.

Mainz, im November 2008 Martin Hanke-Bourgeois

[1] MATLAB ist ein eingetragenes Warenzeichen von The MathWorks, Inc.

Inhalt

Einleitung

Die Aufgabe der *Numerischen Mathematik* besteht in der konkreten (zahlenmäßigen) Auswertung mathematischer Formeln beziehungsweise in der expliziten Lösung mathematischer Gleichungen; die Kapitelüberschriften dieses Buches geben einen Hinweis auf die vielfältigen Fragestellungen.

In der Regel ist das Ziel die Realisierung einer Rechenvorschrift (eines Algorithmus) auf einem Computer. Dabei ergeben sich drei wesentliche Nebenbedingungen.

- Die zur Verfügung stehende Zahlenmenge ist endlich und die einzelnen Rechenoperationen können nur im Rahmen der Maschinengenauigkeit erfolgen.
- Der Speichervorrat ist endlich; von wichtigen Ausnahmen abgesehen, können Funktionen einer reellen oder komplexen Variablen nur approximativ im Computer dargestellt werden.
- Die Rechenzeit ist beschränkt, so daß die meisten Probleme in der zur Verfügung stehenden Zeit nur näherungsweise gelöst werden können.

Jede dieser Einschränkungen resultiert in entsprechenden Fehlern (*Rundungsfehler*, *Diskretisierungsfehler* und *Verfahrensfehler*), die im Einzelfall diskutiert und abgeschätzt werden müssen. Für die jeweilige Aufgabenstellung ist dann zu entscheiden, welcher Algorithmus mit den zur Verfügung stehenden Ressourcen die vorgegebene Genauigkeit mit dem geringsten Aufwand erzielt.

Dieses Anforderungsprofil bildet eine Schnittstelle zwischen Mathematik auf der einen und zahlreichen Anwendungsfächern auf der anderen Seite. Die Automobilindustrie, um ein erstes Beispiel anzuführen, simuliert heute Bremsmanöver und Crashtests im Computer lange bevor der erste Prototyp eines Fahrzeugs gebaut wird; die Simulationen verwenden Programmpakete zur numerischen Lösung gewöhnlicher und partieller Differentialgleichungen. Ein zweites Beispiel ist die medizinische Diagnostik, die mittlerweile zahlreiche Verfahren einsetzt, die ohne Mathematik und insbesondere ohne numerische Methoden undenkbar wären. Eine solche Anwendung, auf die wir gleich zurückkommen werden, ist die Computertomographie. Als drittes Beispiel seien die Wettervorhersage und die Klimaforschung erwähnt, bei der enorme Datenmengen und komplizierte Strömungsprobleme auf sehr unterschiedlichen Längenskalen zu

bewältigen sind. Auch hier bestehen die wesentlichen Komponenten des mathematischen Modells aus partiellen Differentialgleichungen.

Die Rechnungen, die heute in der Wirtschaft, den Ingenieur- und den Naturwissenschaften am Computer durchgeführt werden, sind oftmals so komplex, daß eine vollständige Fehleranalyse nicht mehr möglich ist. In vielen Fällen werden numerische Ergebnisse berechnet und visualisiert, die allenfalls durch praktische Experimente plausibel gemacht werden können. Unter diesen Umständen ist es entscheidend, daß zumindest für repräsentative Modellgleichungen eine rigorose Analyse des Algorithmus vorgenommen wird.

Aus demselben Grund ist es auch unerläßlich, daß sich die Mathematik mit dem Modellbildungsprozeß *per se* auseinandersetzt. Umgekehrt muß die Modellierung gerade bei großen Simulationen auch wichtige numerische Fragestellungen berücksichtigen, etwa welche Vereinfachungen vorgenommen werden können, damit realistische Ergebnisse überhaupt erst berechenbar werden.

Die Numerische Mathematik muß sich somit heute nicht mehr nur mit dem Entwurf konkreter Algorithmen für überschaubare Teilprobleme und deren Fehleranalyse beschäftigen, sondern sich einem wesentlich vielfältigeren Aufgabenspektrum stellen, welches unter dem Begriff *Wissenschaftliches Rechnen* zusammengefaßt wird. Dieses Aufgabenspektrum reicht von der aktiven Mitarbeit bei der mathematischen und numerischen Modellierung, über die Auswahl und vor allem die Kombination sinnvoller Lösungsverfahren

für die einzelnen Module des Projekts, bis schließlich hin zu umfangreichen Testläufen mit real vorgegebenen Eingabedaten. Gerade dieses letzte Stadium eines interdisziplinären Vorhabens ist in seiner Bedeutung und seinen Schwierigkeiten nicht zu unterschätzen: Der Mathematiker muß insbesondere in der Lage sein, die Ergebnisse der Simulationen im physikalischen Kontext zu interpretieren und ein von den vorher durchgeführten Modellrechnungen abweichendes Verhalten des Programms zu erkennen.

Der Begriff des Wissenschaftlichen Rechnens wird gelegentlich noch weiter ausgelegt und umfaßt dann auch verschiedene Teilgebiete, die in die Informatik hineinreichen; stellvertretend seien hier Programmpakete zur Automatischen Differentiation und die Implementierung trickreicher Algorithmen für Hochleistungsrechner (Parallelrechner oder Vektorrechner) angeführt. Derartige Aspekte des Wissenschaftlichen Rechnens werden trotz ihrer offensichtlichen Bedeutung in diesem Buch nicht berücksichtigt, um dessen Umfang nicht über Gebühr zu strapazieren.

Am Beispiel der *Computertomographie* soll im folgenden die Arbeitsweise im Wissenschaftlichen Rechnen veranschaulicht werden. Abbildung 0.1 zeigt ei-

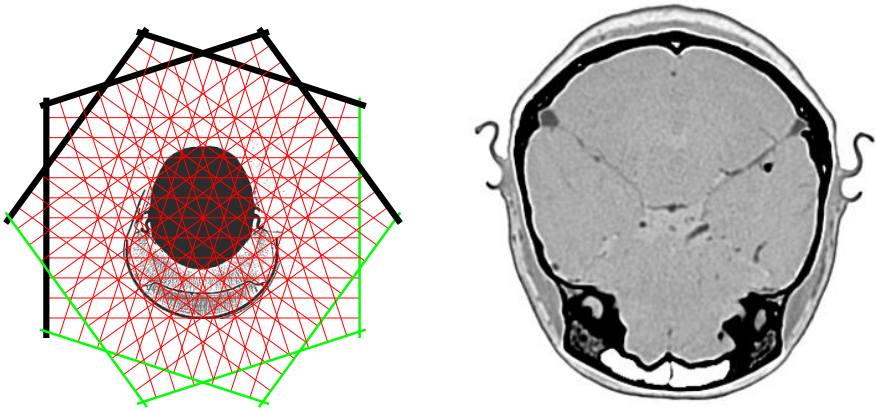


Abb. 0.1: Röntgentomographie: Scannergeometrie (links) und Rekonstruktion (rechts)

ne Röntgenaufnahme eines aktuellen Tomographen der Firma Siemens.[1] Für dieses Bild aus 512×512 Bildpunkten wurde der Kopf des Patienten mit 580×672 Röntgenstrahlen durchleuchtet, die wie in der linken Skizze in einer Querschnittsebene liegen; die dunklen Balken symbolisieren hier die Röntgenquelle, die helleren Balken die Detektoren.

Bei ihrem Weg durch den Körper werden die Röntgenstrahlen aufgrund ihrer kurzen Wellenlänge fast nicht gestreut, verlieren aber Energie. Die verbliebene Energie wird am Detektor gemessen. Für unsere Zwecke ist das physikalische Modell ausreichend, daß der Energieverlust im wesentlichen proportional zu der Energie des Röntgenstrahls ist; die nichtnegative Proportionalitätskonstante ist abhängig von der Ortsvariablen und kann als eine Dichtefunktion f des Körpers interpretiert werden. Ein hoher Energieverlust entspricht einem sehr dichten Gewebe (etwa einem Knochen), ein geringer Energieverlust entspricht dünnem Gewebe. Es ist diese Dichtefunktion f, die in Abbildung 0.1 rechts visualisiert wird.

Sei Γ der geradlinige Weg eines solchen Röntgenstrahls durch den Körper, x ein Punkt auf Γ und s die Bogenlängenparametrisierung von Γ. Ferner bezeichne $E(x)$ die Energie des Strahls im Punkt x und dE die Energieänderung entlang eines (infinitesimal) kleinen Wegstücks der Länge ds. Dann ergibt das obige Modell die Beziehung

$$dE \;=\; -E(x)f(x)\,ds$$

[1]Siemens (Erlangen) und insbesondere Herrn Dr. H. Bruder seien für die zur Verfügung gestellten Daten und die geduldige Beantwortung zahlreicher Rückfragen gedankt.

beziehungsweise die *Differentialgleichung*

$$\frac{d}{ds}E(x) \;=\; -f(x)E(x)\,, \qquad x = x(s)\,.$$

Division durch $E(x)$ ergibt

$$\frac{d}{ds}\log E(x) \;=\; \frac{1}{E(x)}\,\frac{d}{ds}E(x) \;=\; -f(x)\,,$$

und nach Integration über Γ folgt hieraus

$$\int_\Gamma f(x)\,ds \;=\; -\log E(D) + \log E(Q) \;=\; \log\frac{E(Q)}{E(D)}\,, \tag{0.1}$$

wobei $E(Q)$ die Energie des Röntgenstrahls an der Röntgenquelle und $E(D)$ die gemessene Energie am Detektor angibt.

Aus dieser Vielzahl von Integralmitteln über die einzelnen Linien ist nun die Dichtefunktion f zu rekonstruieren. Wir skizzieren im folgenden eine Möglichkeit hierfür, die im wesentlichen mit dem Algorithmus übereinstimmt, der von Sir Hounsfield in dem ersten Tomographen überhaupt implementiert wurde. Für diese Leistung wurde ihm 1979 gemeinsam mit dem Mathematiker Cormack der Medizin-Nobelpreis verliehen.

Bei dem Verfahren wird ein Quadrat Ω, in dem sich der Kopf befindet, durch ein Gitter in Teilquadrate Ω_k, $k = 1, \ldots, N$, unterteilt. Die Funktion f soll nun durch eine (bezüglich dieser Unterteilung) stückweise konstante Funktion

$$f \;\approx\; \tilde{f} \;=\; \sum_{k=1}^{N} s_k \chi_k$$

approximiert werden; dabei bezeichnen die χ_k die charakteristischen Funktionen der Teilquadrate Ω_k und die s_k geeignete nichtnegative Entwicklungskoeffizienten. Eingesetzt in (0.1) ergibt dieser Ansatz

$$\log\frac{E(Q)}{E(D)} \;\approx\; \int_\Gamma \tilde{f}(x)\,ds \;=\; \sum_{k=1}^{N} s_k \int_\Gamma \chi_k(x)\,ds \;=\; \sum_{k=1}^{N} |\Gamma \cap \Omega_k|\, s_k\,, \tag{0.2}$$

wobei $|\Gamma \cap \Omega_k|$ die Länge des Anteils von Γ in Ω_k angibt.

Betrachten wir alle $M = 580 \cdot 672$ Geraden Γ_j, für die Meßdaten b_j für die linke Seite von (0.2) zur Verfügung stehen, so ergibt sich (bei einer Unterteilung in $N = 512^2$ Pixel bzw. Teilquadrate) aus (0.2) ein riesiges überbestimmtes lineares Gleichungssystem für die gesuchten Koeffizienten $s_1, \ldots, s_N$. Aufgrund

der Herleitung bietet es sich an, dieses Gleichungssystem im Sinne eines *Ausgleichsproblems* zu lösen, also Koeffizienten s_k zu suchen, für die der „Datenfit“

$$\sum_{j=1}^{M} \Big(b_j - \sum_{k=1}^{N} |\Gamma_j \cap \Omega_k| \, s_k \Big)^2 \tag{0.3}$$

möglichst klein wird, vgl. Kapitel III.

Die Koeffizientenmatrix dieses Ausgleichsproblems enthält die Einträge $|\Gamma_j \cap \Omega_k|$, von denen die meisten Null sind, da jede Linie nur eine geringe Anzahl Teilquadrate schneidet. Doch selbst wenn nur die von Null verschiedenen Einträge abgespeichert werden, erfordert diese Matrix immer noch einen Speicherbedarf von knapp zwei Gigabyte und muß während der Rechnung auf Hintergrundspeicher ausgelagert werden. Für die Lösung des Ausgleichsproblems kommen daher allenfalls iterative Methoden in Betracht; in Abschnitt 15 werden wir ein solches Verfahren vorstellen und auf dieses Problem anwenden.

Da stückweise konstante Funktionen nur eine schlechte Approximationsgüte haben (vgl. Abschnitt 43), führt das obige numerische Modell zu einem relativ großen Verfahrensfehler. Für die Siemens-Rekonstruktion ergeben sich etwa in den Modellgleichungen (0.2) Abweichungen von bis zu 5%, obwohl die Meßgenauigkeit eher im Promillebereich liegt.

Heute wird daher in der Regel ein anderes Verfahren zur Berechnung von f verwendet, die sogenannte *gefilterte Rückprojektion*. Dieses Verfahren beruht auf tieferliegenden mathematischen Überlegungen und ermöglicht qualitativ hochwertige Rekonstruktionen mit einem deutlich reduzierten Aufwand. Interessanterweise steht die gefilterte Rückprojektion in engem Bezug zu rein theoretisch motivierten Arbeiten aus einer Zeit, als noch niemand an mögliche Anwendungen aus dem Bereich der Tomographie gedacht hat. Der Mathematiker Radon hat nämlich bereits 1917 eine Integraldarstellung für den Wert $f(x)$ der gesuchten Dichte im Punkt $x \in \Omega$ hergeleitet. Die heutigen Verfahren beruhen auf einer sehr ähnlichen Inversionsformel, die in zwei Schritten ausgewertet werden kann: Im ersten Schritt werden die Linienintegrale über alle parallel verlaufenden Geraden mit einer geeigneten Kernfunktion gefaltet; das Ergebnis dieser Faltung wird im anschließenden zweiten Schritt über alle Geraden gemittelt, die durch den Punkt x laufen (dieses Integralmittel wird Rückprojektion genannt und gibt dem Verfahren seinen Namen). Für die numerische Implementierung dieser Inversionsformel wird der kontinuierliche Faltungsoperator durch eine diskrete Faltung approximiert, die mit Hilfe der *schnellen Fouriertransformation* (FFT) effizient berechnet werden kann (vgl. die Abschnitte 53 und 54). Das abschließende Integralmittel kann durch eine einfache *Quadraturformel* (die Trapezsumme, vgl. Abschnitt 36) diskretisiert

werden. Für genauere Details und die mathematischen Grundlagen sei auf das Buch von Natterer [75] verwiesen.

Unabhängig davon, ob der erste oder der zweite Zugang zur Rekonstruktion von f gewählt wird, führt die Rechnung mit realen Daten zu teilweise überraschend schlechten Ergebnissen, da die numerische Lösung sehr sensibel auf Datenfehler reagiert: Das Problem ist *schlecht konditioniert*. Die Datenfehler setzen sich aus Meßungenauigkeiten, Modellfehlern in dem zugrunde gelegten (einfachen) physikalischen Modell und Approximationsfehlern bei der numerischen Diskretisierung zusammen. Diese Sensibilität ist in gewisser Weise mit den offensichtlichen Schwierigkeiten bei der *numerischen Differentiation* vergleichbar (Abschnitte 1.2 und 48) und erfordert spezielle Korrekturen der numerischen Algorithmen. Im Fall der gefilterten Rückprojektion erfolgt diese Korrektur durch einen zusätzlichen Filter für den Faltungskern, der den Fehlereinfluß dämpft. Diese Erkenntnisse sowie die Entwicklung geeigneter Filterfunktionen haben letztendlich den Durchbruch der Computertomographie ermöglicht. Ihre heutige Bedeutung für die medizinische Diagnostik ist somit das Resultat einer intensiven Kooperation von Mathematikern und Entwicklungsingenieuren und unterstreicht die Relevanz des Wissenschaftlichen Rechnens.

Abschließend sei noch erwähnt, daß die hohe Auflösung der Bilder aus der Röntgentomographie nicht zuletzt deshalb möglich ist, weil das physikalische Modell linear ist, die Meßwerte also linear von der Dichtefunktion f abhängen. Akustische, optische oder elektromagnetische Wellen mit einer höheren Wellenlänge werden im Körper gestreut, so daß ihr Verlauf vom Medium selbst abhängt; in diesem Fall geht die Linearität verloren und das lineare Ausgleichsproblem (0.3) muß letztendlich durch ein *nichtlineares Ausgleichsproblem* (Abschnitte 20 und 21) ersetzt werden.

Eine typische Anwendung dieser Art ist die *Ultraschalltomographie*, deren Potential gegenwärtig intensiv untersucht wird. Rekonstruktionen mit der Qualität, die wir heute aus der Röntgentomographie gewohnt sind, werden mit Ultraschallmessungen allerdings in absehbarer Zeit nicht möglich sein.

I Zentrale Grundbegriffe

Rundungsfehler stehen im Mittelpunkt des ersten Abschnitts dieses Kapitels. Insbesondere wird ihr Einfluß auf die Stabilität von numerischen Algorithmen anhand einiger ausgewählter Beispiele diskutiert. Daneben wird Matrix- und Vektornormen viel Platz eingeräumt, da ein sicherer Umgang mit diesen Begriffen im gesamten Rest dieses Buchs wesentlich ist.

Das Buch von Higham [55] ist eine ausgezeichnete Quelle mit interessanten Ergänzungen zu der Thematik dieses Kapitels. Aus der deutschsprachigen Literatur sei noch das Buch von Deuflhard und Hohmann [23] erwähnt.

1 Rundungsfehler, Kondition und Stabilität

Vor der Implementierung eines numerischen Algorithmus sollte man sich zunächst mit den möglichen Auswirkungen der Daten- und *Rundungsfehler* beschäftigen. Während die Rundungsfehler einer einzelnen Elementaroperation (Addition, Multiplikation, Standardfunktionen, etc.) in der Regel vernachlässigbar sind, kann ihre Fehlerfortpflanzung über mehrere Rechenschritte hinweg problematisch werden. Dies führt auf den Betriff der *Stabilität* eines Algorithmus.

Für eine genauere Untersuchung der einzelnen Rundungsfehler und ihrer Fortpflanzung approximiert man die Rechnerarithmetik durch ein Modell, nach dem jede Elementaroperation auf dem Rechner anstelle des exakten Resultats die hierzu nächstgelegene Maschinenzahl liefert:

$$a \boxdot b = \square(a \circ b) .$$

Hierbei sind a und b Maschinenzahlen, $\circ$ steht für eine der mathematischen Grundoperationen und $\boxdot$ für die entsprechende Realisierung auf dem Rechner; die Operation $\square(x)$ bezeichnet die Rundung von x zur nächstgelegenen Maschinenzahl.

Eine weitere Annahme besagt, daß diese Rundungsoperation den tatsächlichen

Wert innerhalb einer maximalen relativen Genauigkeit bestimmt,

$$\square(x) = x(1+\varepsilon) \qquad \text{mit } |\varepsilon| \le \mathsf{eps}\,. \tag{1.1}$$

Die sogenannte *Maschinengenauigkeit* eps ist dabei folgendermaßen definiert:

$$\mathsf{eps} \;=\; \inf\{\, x > 0 \;:\; 1 \boxplus x \neq 1 \,\}\,.$$

Der genaue Wert ist rechnerabhängig; in der Regel ist eps eine negative Potenz von 2, etwa 2^{-d}. Wie man leicht einsieht, wird diese zweite Modellannahme (1.1) allerdings falsch, wenn eine von Null verschiedene Zahl x auf Null gerundet wird; man bezeichnet eine solche Situation als *Underflow*. Ähnlich ist die Situation beim *Overflow*, also wenn der Betrag eines Rechenergebnisses größer als die größte zur Verfügung stehende Maschinenzahl wird. In beiden Fällen sollte bei einer guten Programmiersprache zumindest eine Warnung ausgegeben werden. Sofern also weder Overflow noch Underflow auftreten, ist nach diesem Modell der relative Rundungsfehler durch

$$\frac{|\,\square(x) - x|}{|x|} \;\le\; \mathsf{eps}$$

beschränkt.

Unter den beiden genannten Modellannahmen sind alle Elementaroperationen auf dem Rechner in der folgenden Weise realisiert:

$$a \boxtimes b \;=\; (a \circ b)(1+\varepsilon) \qquad \text{mit } |\varepsilon| \le \mathsf{eps}\,. \tag{1.2}$$

Diese Voraussetzung erlaubt eine einfache und doch recht genaue Rundungsfehleranalyse numerischer Algorithmen, wie wir im folgenden skizzieren wollen. Als Ergänzung sei auf das lesenswerte Buch von Overton [79] verwiesen, das dem IEEE-Rechnerarithmetikstandard gewidmet ist, den beispielsweise auch MATLAB verwendet.

Bevor wir uns im weiteren der Stabilität eines Algorithmus zuwenden, führen wir zunächst den Begriff der *Kondition* eines gegebenen Problems ein. Dabei beschränken wir uns vorerst auf Probleme, bei denen eine reellwertige Funktion $x \mapsto f(x)$ für verschiedene $x \in \mathbb{R}$ ausgewertet werden soll. Aufgrund von Daten- oder Rundungsfehlern werde nun die Funktion f nicht an der Stelle x sondern an der Stelle $\tilde{x} = x + \triangle x$ ausgewertet; wie wirkt sich dieser Eingangsfehler auf das Ergebnis aus?

Bezeichnen wir mit $\triangle y = f(x+\triangle x) - f(x)$ den fortgepflanzten *absoluten Fehler*, so gilt für $f \in C^1(\mathbb{R})$ nach dem Mittelwertsatz

$$\triangle y \;=\; f(x+\triangle x) - f(x) \;=\; f'(\xi)\triangle x\,,$$

wobei ξ in dem Intervall zwischen x und $x + \triangle x$ liegt. Ist die Ableitung Lipschitz-stetig, dann gilt sogar[1]

$$\triangle y \;=\; f'(x)\triangle x + O(|\triangle x|^2)\,. \tag{1.3}$$

Um die Sache zu vereinfachen, vernachlässigen wir den quadratischen Term und verwenden die Größe $|f'(x)|$ als ein Maß für die Fehlerverstärkung des absoluten Eingangsfehlers. Dabei ist der *relative Fehler* üblicherweise von größerer Bedeutung; für $xy \neq 0$ folgt jedoch unmittelbar aus (1.3)

$$\frac{\triangle y}{y} \;\approx\; f'(x)\frac{\triangle x}{f(x)} \;=\; \big(f'(x)\frac{x}{f(x)}\big)\frac{\triangle x}{x}\,. \tag{1.4}$$

Definition 1.1. Die Zahl $\kappa_{\rm abs} = |f'(x)|$ heißt *absolute Konditionszahl* des Problems $x \mapsto f(x)$. Für $xf(x) \neq 0$ ist $\kappa_{\rm rel} = |f'(x)x/f(x)|$ die *relative Konditionszahl* dieses Problems.

Die beiden Konditionszahlen beschreiben also die Verstärkung des absoluten bzw. relativen Eingangsfehlers bei der Auswertung der Funktion f. Ein Problem ist *schlecht konditioniert*, falls eine der Konditionszahlen deutlich größer als Eins ist, ansonsten heißt es *gut konditioniert*.

Beispiele. 1. Bei der *Addition* wird eine reelle Zahl x zu a addiert: $f(x) = x + a$. Für $x \notin \{0, -a\}$ ergibt sich die zugehörige relative Konditionszahl

$$\kappa_{\rm rel} \;=\; \Big|\frac{f'(x)x}{f(x)}\Big| \;=\; \Big|\frac{x}{x+a}\Big|\,.$$

Die relative Konditionszahl ist somit groß, wenn $|x + a| \ll |x|$ gilt, also wenn $x \approx -a$. Diesen schlecht konditionierten Fall bezeichnet man als *Auslöschung*: Für $a = -1$, $x = 1.000\,001$ und Eingangsfehler $\triangle x = 0.001$ ist beispielsweise

$$\begin{aligned} x + a &= 0.000\,001\,, \\ (x + \triangle x) + a &= 0.001\,001\,. \end{aligned}$$

Der absolute Fehler im Ergebnis ist gleich dem Eingangsfehler, der relative Fehler wird hingegen um den Faktor 10^6 verstärkt.

2. Als zweites Beispiel betrachten wir die *Multiplikation* zweier Zahlen, also $f(x) = ax$. In diesem Fall lautet die absolute Konditionszahl

$$\kappa_{\rm abs} \;=\; |f'(x)| \;=\; |a|\,.$$

[1] Hier und im weiteren wird die *O-Notation* $a_\varepsilon = O(b_\varepsilon)$ verwendet, wenn eine positive Konstante C existiert, so daß die Ungleichung $|a_\varepsilon| \leq Cb_\varepsilon$ für alle ε aus einer vereinbarten Grundmenge $E \subset \mathbb{R}^+$ gültig ist. Die *o-Notation* $a_\varepsilon = o(b_\varepsilon)$ besagt, daß $a_\varepsilon/b_\varepsilon \to 0$ bei vereinbartem Grenzübergang $\varepsilon \to \varepsilon_0$. Gilt $a_\varepsilon = O(b_\varepsilon)$ und $b_\varepsilon = O(a_\varepsilon)$, so schreiben wir auch $a_\varepsilon \sim b_\varepsilon$.

Die absolute Kondition ist daher schlecht wenn $|a| \gg 1$ ist; in diesem Fall ergibt sich eine starke absolute Fehlerverstärkung. Der relative Fehler bleibt allerdings gleich ($\kappa_{\rm rel} = 1$). $\Diamond$

Betrachten wir nun die Implementierung eines Algorithmus $\boxed{f}$ zur Lösung eines gegebenen Problems $x \mapsto f(x) = y$ mit $x \in \mathcal{D}(f) \subset \mathbb{R}$. Wir wollen annehmen, daß x und y von Null verschieden sind. Im Verlauf des Algorithmus müssen Rundungsfehler mit relativer Größe eps in Kauf genommen werden. Man kann daher zufrieden sein, wenn die Genauigkeit des Ergebnisses im Rahmen dessen liegt, was aufgrund der Kondition des Problems ohnehin bei gerundeten Eingangsgrößen erwartet werden muß, vgl. (1.4), also wenn

$$\left|\frac{\boxed{f}(x) - f(x)}{f(x)}\right| \le C_V\, \kappa_{\rm rel}\, \mathsf{eps} \tag{1.5}$$

gilt mit einem mäßig großen $C_V > 0$, das weitestgehend von x unabhängig ist. Diese Form der Stabilitätsanalyse wird *Vorwärtsanalyse* genannt und der Algorithmus $\boxed{f}$ heißt *vorwärts stabil*, wenn (1.5) erfüllt ist. Als Beispiel seien die Grundrechenarten des Rechners angeführt: Gemäß unserer Modellannahme (1.2) sind sie allesamt vorwärts stabil; in der Praxis gilt dies nur unter der Einschränkung, daß weder Underflow noch Overflow auftreten.

Bei der *Rückwärtsanalyse* interpretiert man die berechnete Näherung als exakte Lösung eines Problems mit gestörten Eingangsdaten, also $\boxed{f}(x) = f(x+\triangle x)$ und untersucht die zugehörige Störung $|\triangle x|$. Gibt es mehrere Urbilder $x+\triangle x$, so wählt man eines mit kleinster Störung $\triangle x$. Gilt dann

$$\left|\frac{\triangle x}{x}\right| \le C_R\, \mathsf{eps} \tag{1.6}$$

und ist C_R nicht zu groß, so wird der Algorithmus $\boxed{f}$ *rückwärts stabil* genannt. Gibt es *kein* entsprechendes Urbild, dann ist $\boxed{f}$ *nicht* rückwärts stabil.

Für einen rückwärts stabilen Algorithmus ergibt sich somit nach (1.4) und (1.6) mit $\tilde{x} = x + \triangle x$

$$\left|\frac{\boxed{f}(x) - f(x)}{f(x)}\right| = \left|\frac{f(\tilde{x}) - f(x)}{f(x)}\right| \lesssim \kappa_{\rm rel}\left|\frac{\tilde{x} - x}{x}\right| \le C_R\, \kappa_{\rm rel}\, \mathsf{eps}\,.$$

Im Wesentlichen – d. h. bis auf den Einfluß des Approximationsfehlers in (1.4) – ist also jeder rückwärts stabile Algorithmus auch vorwärts stabil; man kann in (1.5) etwa $C_V = C_R$ wählen. Die Umkehrung ist jedoch im allgemeinen nicht richtig.

Die exakte Untersuchung der Kondition eines Problems sowie der Stabilität eines Lösungsalgorithmus ist im allgemeinen sehr aufwendig. Wir wollen dies

im folgenden weitgehend vermeiden und beschränken uns statt dessen auf eine Auswahl einprägsamer Beispiele, die die generelle Vorgehensweise illustrieren sollen.

Die konkreten Rechnungen in diesen Beispielen wurden in der Programmierumgebung MATLAB durchgeführt, auf die wir auch im weiteren Verlauf des Buches noch oft zurückgreifen werden. In dieser Programmierumgebung beträgt die Maschinengenauigkeit $\mathsf{eps} = 2^{-52} \approx 2.2 \cdot 10^{-16}$, und $2^{1024} \approx 1.7977 \cdot 10^{308}$ ist ungefähr die größte zur Verfügung stehende Maschinenzahl. Bei den dargestellten numerischen Ergebnissen sind die korrekten Ziffern jeweils dunkler gedruckt.

1.1 Auslöschung

Der bereits genannte Auslöschungseffekt tritt bei der Auswertung der Funktion

$$f(x) \;=\; x^3\Big(\frac{x}{x^2-1} \;-\; \frac{1}{x}\Big)\,, \qquad x > 1\,,$$

für große Argumente x auf. Die obige Klammerung der Funktion führt mit $x = 1.2 \cdot 10^7$ zunächst auf die Zwischenergebnisse

$$\begin{aligned}
\eta_1 \;&=\; \mathsf{8.333333333333391} \cdot 10^{-8} \;\approx\; x/(x^2-1)\,,\\
\eta_2 \;&=\; \mathsf{8.333333333333334} \cdot 10^{-8} \;\approx\; 1/x\,,\\
\eta_3 \;&=\; \mathsf{1.728} \cdot 10^{21} \;=\; x^3\,,
\end{aligned}$$

im Rahmen der Maschinengenauigkeit. Die anschließende Subtraktion

$$\eta_4 \;=\; \eta_1 \boxminus \eta_2 \;=\; \mathsf{5.691\ldots} \cdot 10^{-22}$$

ergibt jedoch nur noch eine korrekte Dezimalstelle, so daß schließlich ein völlig verkehrtes Ergebnis

$$\boxed{f}(x) \;=\; \eta_3 \boxdot \eta_4 \;=\; \mathsf{0.983\ldots}$$

berechnet wird. Wie man leicht einsieht ist die Funktion f im gesamten Definitionsgebiet $x > 1$ größer als Eins. Der Algorithmus ist also nicht rückwärts stabil. Man beachte, daß für dieses Problem

$$\kappa_{\mathrm{rel}} \;=\; \frac{2}{x^2-1}$$

sehr nahe bei Null ist, Eingangsfehler also stark gedämpft werden sollten. Der Algorithmus ist also auch nicht vorwärts stabil.

Für große x kann $f(x)$ mittels der Identität

$$f(x) \;=\; \frac{1}{1-x^{-2}}$$

stabil ausgewertet werden. Eine Implementierung dieser Darstellung liefert den korrekten Wert $\boxed{f}(x) = \texttt{1.000000000000007} = \square\, f(x)$.

1.2 Numerische Differentiation

Ähnliche Phänomene treten üblicherweise bei der numerischen Differentiation auf. Gegeben sei ein vorwärts stabiler Algorithmus $\boxed{F}$ zur Auswertung der stetig differenzierbaren Funktion F und gesucht ist eine Näherung für $f = F'$ an der Stelle x. Eine Möglichkeit hierzu besteht in der Berechnung des Differenzenquotienten

$$f(x) \;\approx\; \boxed{f}(x) \;=\; \frac{\boxed{F}(x+h) - \boxed{F}(x)}{h} \qquad \text{mit } |h| \text{ klein}\,.$$

Bei der Auswertung des Zählers ergibt sich nach (1.5) ein Ergebnis $F'(x)h + O(h^2) + O(F(x)\,\mathsf{eps})$, wobei der letzte Term den ersten für kleine h dominieren kann – wieder ein Auslöschungseffekt. Nach Division durch h erhält man dann einen relativ großen Fehler der Größenordnung $O(\mathsf{eps}/h)$.

Ist hingegen die Funktion F geschlossen gegeben, etwa $F(x) = x^2$, dann kann prinzipiell die Ableitung $f(x) = 2x$ im Rahmen der Rechengenauigkeit exakt ausgewertet werden. Die numerische Differentiation ist also instabil.

1.3 Nullstellenaufgabe

Die Funktion $g : \mathbb{R} \to \mathbb{R}$ sei stetig differenzierbar mit einer einfachen Nullstelle in $\widehat{x}$, das heißt es gilt

$$y = g(\widehat{x}) = 0 \qquad \text{und} \qquad g'(\widehat{x}) \neq 0\,.$$

Dann existieren offene Intervalle $\mathcal{I}$ um $\widehat{x}$ und $\mathcal{J}$ um $y = 0$ und eine lokale Umkehrfunktion $g^{-1} : \mathcal{J} \to \mathcal{I}$ von g mit $g^{-1}(0) = \widehat{x}$. Die Lösung der Nullstellenaufgabe $g(x) = 0$ kann also als Auswertung der Funktion $f = g^{-1}$ an der Stelle $y = 0$ interpretiert werden.

Nach Definition 1.1 ergibt sich hierfür die absolute Konditionszahl

$$\kappa_{\text{abs}} \;=\; |f'(0)| \;=\; \Big|\frac{1}{g'(\widehat{x})}\Big|\,.$$

Die Nullstellenaufgabe ist daher schlecht konditioniert, falls der Graph von g die x-Achse mit einem sehr kleinen Winkel schneidet (vgl. Abbildung 1.1).

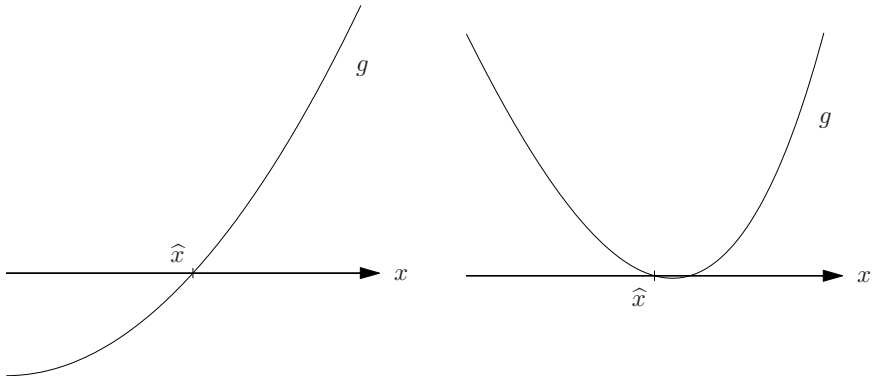


Abb. 1.1: Gut und schlecht konditionierte Nullstellenaufgabe

1.4 Quadratische Gleichungen

In dem Beispiel des vorigen Abschnitts wählen wir das Polynom $g(x) = ax^2 + bx + c$ mit Parametern $a, c \neq 0$ und $b^2 - 4ac > 0$. Die Nullstellen von g sind dann die Lösungen

$$x_{1/2} = \frac{-b \pm \sqrt{b^2 - 4ac}}{2a} \tag{1.7}$$

der quadratischen Gleichung $ax^2 + bx + c = 0$. Bei der Auswertung dieser Formel können sich verschiedene Probleme ergeben.

(i) $\quad b^2 \gg |4ac|\,, \quad$ so daß $\quad \sqrt{b^2 - 4ac} \approx |b|\,:$

In diesem Fall tritt im Zähler je nach Vorzeichen von b bei x_1 oder bei x_2 Auslöschung auf. Eine Implementierung der obigen Formel ist in diesem Fall also schlecht konditioniert, besonders dann, wenn auch noch $|a|$ klein ist. Besser ist der folgende Algorithmus, der Auslöschung vermeidet:

$$\text{für } b > 0: \qquad x_2 = \frac{-b - \sqrt{b^2 - 4ac}}{2a}\,, \quad x_1 = \frac{c}{ax_2}\,,$$

$$\text{für } b < 0: \qquad x_1 = \frac{-b + \sqrt{b^2 - 4ac}}{2a}\,, \quad x_2 = \frac{c}{ax_1}\,.$$

(ii) $\quad b^2 \approx 4ac\,:$

Dies kann zu Auslöschung im Radikanden der Wurzel führen. Aus dem vorherigen Beispiel wissen wir, daß in diesem Fall die Kondition der Nullstellenauf-

gabe,

$$\kappa_{\text{abs}} \;=\; \frac{1}{|2ax_{1/2}+b|} \;=\; \frac{1}{\sqrt{b^2-4ac}}\,,$$

sehr groß und daher die Nullstellenaufgabe *per se* schlecht konditioniert ist.

Im Fall (i) ist das Problem gut konditioniert ($\kappa_{\text{abs}} \approx 1/|b|$), aber die Formel (1.7) ist instabil; im Fall (ii) liegt die Schwierigkeit nicht in der fehlenden Stabilität des Algorithmus, sondern in der schlechten Kondition der Nullstellenaufgabe.

1.5 Berechnung der Eulerschen Zahl

Zur Berechnung der Eulerschen Zahl e kann man den Grenzprozeß

$$e = \lim_{n\to\infty} \big(1+\frac{1}{n}\big)^n$$

numerisch implementieren. Eine Tayloranalyse der zugehörigen Näherungen ergibt

$$\begin{aligned} e_n \;&=\; \big(1+\frac{1}{n}\big)^n \;=\; \exp\big(n\log(1+\frac{1}{n})\big) \;=\; \exp\big(n\,(\frac{1}{n}+O(n^{-2}))\big) \\ &=\; \exp\big(1+O(n^{-1})\big) \;=\; e+O(n^{-1}) \end{aligned}$$

und daher wird man erwarten, daß eine Approximation der Zahl e im Bereich der Maschinengenauigkeit für $n_* \approx 1/\mathsf{eps}$ erreicht wird.

Tabelle 1.1 zeigt die berechneten Näherungen $\boxed{e_n}$ für verschiedene n. Erstaunlicherweise erreicht die Genauigkeit nur etwa eine Größenordnung von $\sqrt{\mathsf{eps}} \approx 1.5\cdot 10^{-8}$ und wird danach wieder schlechter; für den Index $n_* = 10^{15} \approx 1/\mathsf{eps}$ ist keine einzige Ziffer von $\boxed{e_n}$ korrekt.

Die Erklärung liegt natürlich im Einfluß der Rundungsfehler, denn bei der Auswertung des Klammerausdrucks $1\boxplus\frac{1}{n}$ bleiben nur wenige signifikante Ziffern von $1/n$ erhalten. Die berechnete Näherung $\boxed{e_n}$ kann für $n \ll 1/\mathsf{eps}$ folgendermaßen abgeschätzt werden:

$$\begin{aligned} \boxed{e_n} \approx\; \exp\big(n\log(1\boxplus\frac{1}{n})\big) \;&=\; \exp\big(n\log(1+\frac{1}{n}+O(\mathsf{eps}))\big) \\ =\; \exp\big(1+O(n^{-1})+O(n\,\mathsf{eps})\big) \;&=\; e+O(n^{-1})+O(n\,\mathsf{eps})\,. \end{aligned}$$

Es ergeben sich somit zwei Fehlerkomponenten: ein Approximationsfehler mit Ordnung $O(n^{-1})$ und ein fortgepflanzter Datenfehler der Größe $O(n\,\mathsf{eps})$. Insgesamt ist daher lediglich eine absolute Genauigkeit von etwa $\max\{n^{-1}, n\,\mathsf{eps}\}$ erreichbar, und dies entspricht recht gut den Fehlerwerten in der Tabelle. Das Optimum wird für $n^{-1} \sim n\,\mathsf{eps}$ erreicht, also für $n \sim 1/\sqrt{\mathsf{eps}}$.

Tab. 1.1: Approximationen von e

n	$\boxed{e_n}$	$\lvert e - \boxed{e_n}\rvert$
10	2.593742460	$1.2 \cdot 10^{-1}$
10^2	2.704813829	$1.3 \cdot 10^{-2}$
10^3	2.716923932	$1.4 \cdot 10^{-3}$
10^4	2.718145926	$1.4 \cdot 10^{-4}$
10^5	2.718268237	$1.4 \cdot 10^{-5}$
10^6	2.718280469	$1.4 \cdot 10^{-6}$
10^7	2.718281694	$1.3 \cdot 10^{-7}$
10^8	2.718281798	$3.0 \cdot 10^{-8}$
10^9	2.718282052	$2.2 \cdot 10^{-7}$
10^{10}	2.718282053	$2.2 \cdot 10^{-7}$
10^{11}	2.718282053	$2.2 \cdot 10^{-7}$
10^{12}	2.718523496	$2.4 \cdot 10^{-4}$
10^{13}	2.716110034	$2.2 \cdot 10^{-3}$
10^{14}	2.716110034	$2.2 \cdot 10^{-3}$
10^{15}	3.035035206	$3.2 \cdot 10^{-1}$
e	2.718281828	

1.6 Rationale Funktionen

Das folgende Beispiel ist dem Buch von Higham [55] entnommen. Abbildung 1.2 zeigt die berechneten Funktionswerte der rationalen Funktion

$$f(x) \;=\; \frac{(((4x-59)x+324)x-751)x+622}{(((x-14)x+72)x-151)x+112} \tag{1.8}$$

an 300 *aufeinanderfolgenden* Maschinenzahlen in der Nähe von $x = 1.606$. Die gezackte Linie in der Mitte der Abbildung gibt eine hochgenaue Approximation an den exakten Wert der Funktion in diesem Bereich. Die runden Kreise in der Abbildung sind hingegen die Werte, die sich bei der Auswertung von f mit dem sogenannten *Hornerschema* ergeben, d. h. mit dem durch die Klammerung in (1.8) definierten Algorithmus.

Die Funktion f hat in der unmittelbaren Nachbarschaft des betrachteten Intervalls ein lokales Maximum. Entsprechend schlecht konditioniert ist die Umkehrfunktion f^{-1} (vgl. Abschnitt 1.3), was sich in der Abbildung widerspiegelt, da das Urbild eines gegebenen Funktionswerts nicht verläßlich bestimmt werden kann. Auffallend ist bei dieser Abbildung jedoch, daß die Lage der berechneten Näherungen im wesentlichen auf sieben Geradenstücke beschränkt ist, deren Definitionsbereiche zudem stark überlappen. Der dargestellte Bereich der y-Achse umfaßt etwa 80 aufeinanderfolgende Maschinenzahlen, von denen jeweils offensichtlich nur einige wenige, relativ weit auseinanderliegende Maschinenzahlen für $\boxed{f}(x)$ in Betracht kommen.

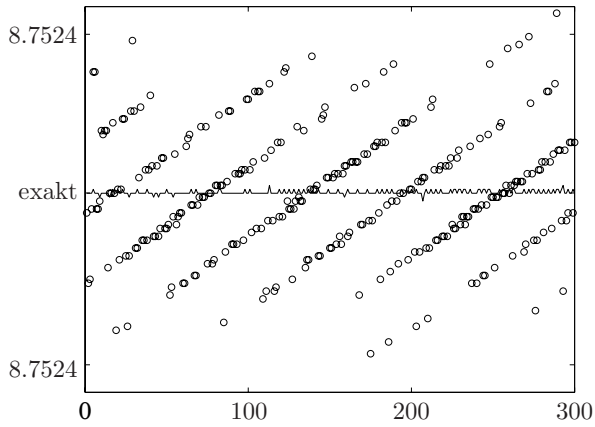


Abb. 1.2: Berechnete Funktionswerte von f aus (1.8), vgl. den Text zur Erläuterung der Abbildung

2 Vektor- und Matrixnormen

Die Überlegungen und Definitionen des vorigen Abschnitts lassen sich prinzipiell auch auf vektorwertige Funktionen mehrerer Variablen übertragen. Bei n Variablen und m Funktionskomponenten führt dies jedoch auf nm absolute und relative Konditionszahlen. Um diese Zahlenflut zu vermeiden, wird in der Regel ein einfacherer Zugang gewählt, der zu einer einzigen Konditionszahl für die *schlimmstmögliche* Fehlerfortpflanzung führt, vgl. Definition 2.9 weiter unten. Hierzu ist es erforderlich, für Vektoren und Matrizen geeignete Normen einzuführen.

Im weiteren ist es meist unerheblich, ob die Vektoren bzw. Matrizen reelle oder komplexe Einträge besitzen; in diesem Fall schreiben wir der Einfachheit halber $\mathbb{K}$ für den entsprechenden Zahlenkörper und drücken dadurch aus, daß die entsprechenden Resultate in gleicher Weise für $\mathbb{K} = \mathbb{R}$ und $\mathbb{K} = \mathbb{C}$ gelten.

Somit bezeichnet $\mathbb{K}^n$ den Vektorraum der n-dimensionalen Vektoren

$$x = [x_i] = \begin{bmatrix} x_1 \\ \vdots \\ x_n \end{bmatrix}, \qquad x_i \in \mathbb{K}\,,$$

über $\mathbb{K}$ und $\mathbb{K}^{m\times n}$ den entsprechenden Vektorraum der $m \times n$-Matrizen

$$A = [a_{ij}] = \begin{bmatrix} a_{11} & \cdots & a_{1n} \\ \vdots & & \vdots \\ a_{m1} & \cdots & a_{mn} \end{bmatrix}, \qquad a_{ij} \in \mathbb{K}\,.$$

Für $x \in \mathbb{K}^n$ unterscheiden wir zwischen

$$x^T = [x_1, x_2, \ldots, x_n] \in \mathbb{K}^{1\times n} \quad \text{und} \quad x^* = [\overline{x}_1, \overline{x}_2, \ldots, \overline{x}_n] \in \mathbb{K}^{1\times n};$$

bei x^* sind die Einträge komplex konjugiert; für $\mathbb{K} = \mathbb{R}$ stimmen die beiden „Zeilenvektoren“ x^T und x^* überein.[2] A^T und A^* in $\mathbb{K}^{n\times m}$ sind entsprechend definiert. Stimmen A und A^* überein, so heißt A hermitesch. Ist $A \in \mathbb{K}^{n\times n}$ invertierbar, so bezeichnen wir mit A^{-1} die Inverse von A, A^{-*} ist eine Kurzschreibweise für die Inverse von A^*. Stimmen A^* und A^{-1} überein, so heißt A unitär. Der Nullvektor und die Nullmatrix werden jeweils mit 0 bezeichnet, I ist die Einheitsmatrix in $\mathbb{K}^{n\times n}$.

Wie üblich identifizieren wir die Matrix $A \in \mathbb{K}^{m\times n}$ mit der linearen Abbildung

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

Mit $\mathcal{R}(A)$ bezeichnen wir den *Bildraum* (engl.: *range*) dieser Abbildung und mit $\mathcal{N}(A)$ ihren *Kern* (Nullraum), also den Unterraum, der von A auf das Nullelement abgebildet wird. Der Kern ist trivial, $\mathcal{N}(A) = \{0\}$, falls A vollen Spaltenrang hat, d. h. falls $\operatorname{Rang} A = n$.

Im Raum $\mathbb{K}^n$ greifen wir gelegentlich auf die *kartesische Basis* $\{e_1, \ldots, e_n\}$ zurück, wobei $e_i = [\delta_{ij}]_{j=1}^n$ den Vektor bezeichnet, der in der i-ten Komponente eine Eins und ansonsten nur Nulleinträge enthält;

$$\delta_{ij} = \begin{cases} 1, & i = j, \\ 0, & i \neq j, \end{cases}$$

ist das sogenannte *Kronecker-Symbol*. Für die *lineare Hülle* von k Vektoren $z_1, \ldots, z_k$ verwenden wir die Notation $\operatorname{span}\{z_1, \ldots, z_k\}$ (engl.: *to span*, umfassen).

Definition 2.1. Sei X ein Vektorraum über $\mathbb{K}$. Eine Abbildung $\|\cdot\| : X \to \mathbb{R}$ heißt *Norm*, wenn die folgenden Eigenschaften erfüllt sind:

(i)	$\|x\| > 0$	für alle $x \in X \setminus \{0\}$,
(ii)	$\|\alpha x\| = \lvert\alpha\rvert \, \|x\|$	für alle $x \in X$, $\alpha \in \mathbb{K}$,
(iii)	$\|x + y\| \leq \|x\| + \|y\|$	für alle $x, y \in X$.

Die drei Bedingungen werden (in dieser Reihenfolge) mit *Definitheit*, *Homogenität* und *Dreiecksungleichung* bezeichnet.

Beispiele. Die gängigsten Normen in $X = \mathbb{K}^n$ sind die

[2] Für das euklidische Innenprodukt zweier Vektoren $x, y \in \mathbb{K}^n$ verwenden wir die konsistente Schreibweise x^*y.

- Betragssummennorm: $\|x\|_1 = \sum_{i=1}^{n} |x_i|\,,$
- Euklidnorm: $\|x\|_2 = \Big(\sum_{i=1}^{n} |x_i|^2\Big)^{1/2} = \sqrt{x^* x}\,,$
- Maximumnorm: $\|x\|_\infty = \max_{1\le i\le n} |x_i|\,.$

Mögliche Normen in $X = \mathbb{K}^{m\times n}$ sind die

- Spaltensummennorm: $\|A\|_1 = \max_{1\le j\le n} \sum_{i=1}^{m} |a_{ij}|\,,$
- Zeilensummennorm: $\|A\|_\infty = \max_{1\le i\le m} \sum_{j=1}^{n} |a_{ij}|\,,$
- Frobeniusnorm: $\|A\|_F = \Big(\sum_{i=1}^{m}\sum_{j=1}^{n} |a_{ij}|^2\Big)^{1/2}.$

Die wichtige Spektralnorm $\|A\|_2$ wird weiter unten eingeführt, vgl. (2.2). ◊

Jede Norm in $X = \mathbb{K}^n$ oder $X = \mathbb{K}^{m\times n}$ induziert eine *Metrik*

$$d(x,y) = \|x-y\|\,, \qquad x,y \in X\,,$$

und damit einen Konvergenzbegriff: Die Folge $\{x_k\} \subset X$ ist konvergent mit Grenzelement $x \in X$ genau dann, wenn $d(x_k, x)$ für $k \to \infty$ gegen Null konvergiert. Glücklicherweise ist dieser Konvergenzbegriff in den Räumen $X = \mathbb{K}^n$ und $X = \mathbb{K}^{m\times n}$ nur auf den ersten Blick von der Wahl der Norm abhängig; tatsächlich ist in diesen Räumen die Frage, ob eine Folge $\{x_k\} \subset X$ konvergent ist oder nicht, völlig unabhängig von der gewählten Norm.

Satz 2.2. *Alle Normen in $\mathbb{K}^n$ sind äquivalent zur Maximumnorm, d. h. für jede Norm $\|\cdot\|$ in $\mathbb{K}^n$ gibt es positive Konstanten $c, C > 0$ mit*

$$c\|x\|_\infty \le \|x\| \le C\|x\|_\infty \qquad \textit{für alle } x \in \mathbb{K}^n\,.$$

Beweis. Seien $x = [x_i]$ und $y = [y_i]$ beliebige Vektoren im $\mathbb{K}^n$ und $\|\cdot\|$ eine Norm im $\mathbb{K}^n$. Dann ist

$$x - y = \sum_{i=1}^{n} (x_i - y_i) e_i$$

und aus der Dreiecksungleichung folgt

$$\big|\, \|x\| - \|y\| \,\big| \le \|x-y\| \le \sum_{i=1}^{n} |x_i - y_i|\, \|e_i\| \le \|x-y\|_\infty \sum_{i=1}^{n} \|e_i\|\,.$$

Folglich ist $\|\cdot\| : \mathbb{K}^n \to \mathbb{R}$ eine Lipschitz-stetige Funktion mit Lipschitz-Konstante $L = \sum_{i=1}^n \|e_i\|$. Als solche nimmt die Funktion $\|\cdot\|$ auf der kompakten Einheitssphäre $\{x \in \mathbb{K}^n : \|x\|_\infty = 1\}$ sowohl ihr Maximum C wie auch ihr Minimum c an; wegen der Normeigenschaft ist das Minimum strikt positiv. Daher folgt für beliebiges $z \in \mathbb{K}^n \setminus \{0\}$

$$c \le \Big\| \frac{z}{\|z\|_\infty} \Big\| \le C$$

beziehungsweise

$$c\|z\|_\infty \le \|z\| \le C\|z\|_\infty ,$$

was zu zeigen war. □

Als Korollar folgt damit unmittelbar, daß zwei beliebige Normen im $\mathbb{K}^n$ zueinander äquivalent sind. Ferner gilt ein entsprechendes Resultat für Normen in $\mathbb{K}^{m\times n}$, vgl. Aufgabe 5.

Wir beweisen nun die obige Behauptung über die Unabhängigkeit des Konvergenzbegriffs.

Korollar 2.3. *In $\mathbb{K}^n$ und $\mathbb{K}^{m\times n}$ sind für jede Norm Folgen genau dann konvergent, wenn sie komponentenweise konvergent sind.*

Beweis. Wir beweisen den Satz hier nur für $X = \mathbb{K}^n$ und verweisen für den Beweis des Falls $X = \mathbb{K}^{m\times n}$ wieder auf Aufgabe 5.

Nehmen wir an, es gibt ein $x \in X$ und eine Norm $\|\cdot\|$ in X, so daß

$$\|x^{(k)} - x\| \longrightarrow 0 , \qquad k \to \infty .$$

Aus Satz 2.2 folgt hieraus

$$\|x^{(k)} - x\|_\infty \le \frac{1}{c} \|x^{(k)} - x\| \longrightarrow 0 , \qquad k \to \infty .$$

Da $\max_i |x_i^{(k)} - x_i|$ somit für $k \to \infty$ gegen Null konvergiert, liegt komponentenweise Konvergenz gegen x vor.

Ist umgekehrt die Folge $\{x^{(k)}\}$ komponentenweise konvergent gegen x, dann folgt $\|x^{(k)} - x\|_\infty \to 0$ für $k \to \infty$ und für eine beliebige andere Norm in $\mathbb{K}^n$ folgt aus Satz 2.2

$$\|x^{(k)} - x\| \le C\|x^{(k)} - x\|_\infty \longrightarrow 0 , \qquad k \to \infty .$$

Somit konvergiert die Folge auch in dieser Norm gegen x. □

Der Vektorraum $\mathbb{K}^{n\times n}$ unterscheidet sich vom $\mathbb{K}^n$ dadurch, daß noch eine Multiplikation AB für $A, B \in \mathbb{K}^{n\times n}$ definiert ist. Daher sind noch die folgenden beiden Begriffe wesentlich.

Definition 2.4. Eine Norm $\|\cdot\|_M$ auf $\mathbb{K}^{n\times n}$ heißt *submultiplikativ*, falls

$$\|AB\|_M \le \|A\|_M \|B\|_M \qquad \text{für alle } A, B \in \mathbb{K}^{n\times n};$$

eine Norm $\|\cdot\|_M$ auf $\mathbb{K}^{n\times n}$ heißt *verträglich* mit einer (Vektor-)Norm $\|\cdot\|$ auf $\mathbb{K}^n$, wenn

$$\|Ax\| \le \|A\|_M \|x\| \qquad \text{für alle } A \in \mathbb{K}^{n\times n} \text{ und alle } x \in \mathbb{K}^n.$$

Beispiele. Die Norm $|||A||| = \max_{ij} |a_{ij}|$ ist nicht submultiplikativ, denn für

$$A = \begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix} \qquad \text{ist} \qquad A^2 = \begin{bmatrix} 2 & 2 \\ 2 & 2 \end{bmatrix}$$

und $|||A^2||| = 2 \not\le 1 = |||A|||^2$. Die sogenannte *Gesamtnorm* $\|A\|_G = n \max_{ij} |a_{ij}|$ ist hingegen submultiplikativ, vgl. Aufgabe 5.

Die Frobeniusnorm ist mit der Euklidnorm verträglich. Um dies nachzurechnen, verwendet man zunächst die Cauchy-Schwarz-Ungleichung und erhält

$$(Ax)_i^2 = \Big(\sum_{j=1}^n a_{ij}x_j\Big)^2 \le \sum_{j=1}^n |a_{ij}|^2 \sum_{j=1}^n |x_j|^2 = \Big(\sum_{j=1}^n |a_{ij}|^2\Big) \|x\|_2^2$$

für $1 \le i \le n$. Durch Summation über i folgt dann die Behauptung:

$$\|Ax\|_2^2 = \sum_{i=1}^n (Ax)_i^2 \le \sum_{i=1}^n \Big(\sum_{j=1}^n |a_{ij}|^2\Big) \|x\|_2^2 = \|A\|_F^2 \|x\|_2^2 .$$

◊

Definition und Satz 2.5. *Sei $\|\cdot\|$ eine Norm in $\mathbb{K}^n$. Dann ist*

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

eine Norm auf $\mathbb{K}^{n\times n}$, die durch $\|\cdot\|$ induzierte Norm*. Diese Norm ist submultiplikativ und ist mit der Ausgangsnorm verträglich. Ist $\|\cdot\|_M$ eine andere mit $\|\cdot\|$ verträgliche Norm, dann gilt $|||A||| \le \|A\|_M$ für alle $A \in \mathbb{K}^{n\times n}$.*

Beweis. Die Normeigenschaften sind leicht nachgerechnet. Für die Submultiplikativität betrachten wir eine beliebige Matrix $A \in \mathbb{K}^{n\times n}$ und eine Matrix

$B \neq 0$ und erhalten

$$\begin{aligned} |\!|\!| AB |\!|\!| &= \sup_{x\neq 0} \frac{\|ABx\|}{\|x\|} = \sup_{Bx\neq 0} \frac{\|ABx\|}{\|x\|} = \sup_{Bx\neq 0} \Big(\frac{\|ABx\|}{\|Bx\|} \frac{\|Bx\|}{\|x\|} \Big) \\ &\leq \sup_{Bx\neq 0} \frac{\|ABx\|}{\|Bx\|} \sup_{Bx\neq 0} \frac{\|Bx\|}{\|x\|} \leq \sup_{y\neq 0} \frac{\|Ay\|}{\|y\|} \sup_{x\neq 0} \frac{\|Bx\|}{\|x\|} \\ &= |\!|\!| A |\!|\!| \, |\!|\!| B |\!|\!| \,. \end{aligned}$$

Folglich ist die induzierte Norm submultiplikativ. Die Verträglichkeit mit der Ausgangsnorm folgt unmittelbar aus der Definition: Demnach ist nämlich

$$|\!|\!| A |\!|\!| = \sup_{x\neq 0} \frac{\|Ax\|}{\|x\|} \geq \frac{\|Ax\|}{\|x\|}$$

für jedes $x \neq 0$ beziehungsweise $\|Ax\| \leq |\!|\!| A |\!|\!| \, \|x\|$. Sei abschließend $\|\cdot\|_M$ eine andere mit $\|\cdot\|$ verträgliche Norm. Dann ist nach Definition 2.5 $|\!|\!| A |\!|\!| = \|Ax\|$ für ein gewisses $x \in \mathbb{K}^n$ mit $\|x\| = 1$ und aus der Verträglichkeit der zweiten Norm folgt daher

$$|\!|\!| A |\!|\!| = \|Ax\| \leq \|A\|_M \|x\| = \|A\|_M \,.$$

Damit ist der Beweis vollständig. □

Beispiel 2.6. Sei $A \in \mathbb{K}^{n\times n}$ und $x \in \mathbb{K}^n$. Dann ergibt eine einfache Rechnung, daß

$$\begin{aligned} \|Ax\|_1 &= \sum_{i=1}^n |\,(Ax)_i\,| = \sum_{i=1}^n \Big| \sum_{j=1}^n a_{ij} x_j \Big| \\ &\leq \sum_{i=1}^n \sum_{j=1}^n |a_{ij}|\,|x_j| = \sum_{j=1}^n |x_j| \sum_{i=1}^n |a_{ij}| \\ &\leq \sum_{j=1}^n |x_j| \max_{1\leq j\leq n} \sum_{i=1}^n |a_{ij}| = \|x\|_1 \|A\|_1 \,. \end{aligned}$$

Also ist die Spaltensummennorm mit der Betragssummennnorm verträglich, das heißt

$$\frac{\|Ax\|_1}{\|x\|_1} \leq \|A\|_1 \qquad \text{für alle } x \neq 0 \,. \tag{2.1}$$

Wir wollen nun zeigen, daß $\|A\|_1$ die kleinstmögliche obere Schranke ist, also daß die Betragssummennorm die Spaltensummennorm induziert. Dazu müssen

wir ein $x \in \mathbb{K}^n \setminus \{0\}$ finden, für das Gleichheit in (2.1) gilt. Wir wählen hierzu den Spaltenindex j, für den

$$\|A\|_1 = \sum_{i=1}^{n} |a_{ij}|$$

gilt und definieren x als den zugehörigen kartesischen Basisvektor e_j: Aufgrund dieser Konstruktion folgt unmittelbar, daß

$$\|A\|_1 \;=\; \|Ae_j\|_1 \;=\; \frac{\|Ae_j\|_1}{\|e_j\|_1}\,,$$

was zu zeigen war. ◊

In ähnlicher Weise weist man nach, daß die Zeilensummennorm durch die Maximumnorm induziert wird, vgl. Aufgabe 6.

Bemerkung. Die Definitionen 2.4 und 2.5 lassen sich sinngemäß auf Normen in $\mathbb{K}^{m\times n}$ mit $m \neq n$ übertragen. In diesem Fall müssen dann natürlich sowohl in $\mathbb{K}^m$ als auch in $\mathbb{K}^n$ zugehörige Normen spezifiziert werden. Man überprüft beispielsweise sofort anhand der Rechnung in Beispiel 2.6, daß die Spaltensummennorm in $\mathbb{K}^{m\times n}$ durch die Kombination beider Betragssummennormen in $\mathbb{K}^m$ und $\mathbb{K}^n$ induziert wird. ◊

Die vermutlich wichtigste Norm in $\mathbb{K}^n$ ist die Euklidnorm. Wir werden uns daher im Rest dieses Abschnitts mit der durch die Euklidnormen in $\mathbb{K}^n$ und $\mathbb{K}^m$ induzierten Norm in $\mathbb{K}^{m\times n}$ beschäftigen, der sogenannten *Spektralnorm*

$$\|A\|_2 = \max_{\|x\|_2=1} \|Ax\|_2 = \max_{\|x\|_2=1} \big((Ax)^*(Ax)\big)^{1/2} = \max_{\|x\|_2=1} (x^*A^*Ax)^{1/2}\,. \tag{2.2}$$

Die Spektralnorm ist nicht mit dem *Spektralradius* $\varrho(A)$ einer quadratischen Matrix $A \in \mathbb{K}^{n\times n}$ zu verwechseln: Ist $\sigma(A) \subset \mathbb{C}$ die Menge aller Eigenwerte (das *Spektrum*) von A, dann ist der Spektralradius gegeben durch

$$\varrho(A) \;=\; \max\{\,|\lambda| \,:\, \lambda \in \sigma(A)\,\}\,, \tag{2.3}$$

also dem Betrag des betragsgrößten Eigenwerts der Matrix A. Lediglich für einige spezielle Matrizen stimmen $\|A\|_2$ und $\varrho(A)$ überein, zum Beispiel für hermitesche Matrizen $A = A^* \in \mathbb{K}^{n\times n}$. Aber auch im allgemeinen Fall besteht zwischen den beiden Begriffen Spektralnorm und Spektralradius ein gewisser Zusammenhang:

Satz 2.7. *Für jede Matrix $A \in \mathbb{K}^{m\times n}$ ist $\|A\|_2 = \|A^*\|_2 = \big(\varrho(A^*A)\big)^{1/2}$.*

Beweis. Aus $\|A\|_2 = 0$ folgt unmittelbar $A = 0$ und der Beweis ist in diesem Fall trivial. Sei also im weiteren $\mu = \|A\|_2 > 0$ und $x \in \mathbb{K}^n$ mit $\|x\|_2 = 1$ ein Vektor, für den das Maximum in (2.2) angenommen wird. Für $y = Ax/\mu$ ergibt sich dann

$$\begin{aligned} \|A^*y - \mu x\|_2^2 &= \|A^*y\|_2^2 - (A^*y)^*(\mu x) - (\mu x)^* A^*y + \mu^2 \|x\|_2^2 \\ &= \|A^*y\|_2^2 - \mu\, y^*Ax - \mu\, x^*A^*y + \mu^2 \\ &= \|A^*y\|_2^2 - 2\mu \operatorname{Re} y^*Ax + \mu^2 . \end{aligned}$$

Nach Konstruktion ist aber $Ax = \mu y$ und $\|y\|_2 = \|Ax\|_2/\mu = \|A\|_2/\mu = 1$, so daß

$$\|A^*y - \mu x\|_2^2 = \|A^*y\|_2^2 - 2\mu^2 \|y\|_2^2 + \mu^2 = \|A^*y\|_2^2 - \mu^2 . \tag{2.4}$$

Da die linke Seite von (2.4) nichtnegativ ist, folgt hieraus

$$\|A\|_2^2 = \mu^2 \le \|A^*y\|_2^2 \le \|A^*\|_2^2 \|y\|_2^2 = \|A^*\|_2^2 .$$

Somit ist $\|A\|_2 \le \|A^*\|_2$. Das gleiche Argument auf A^* anstelle von A angewandt ergibt entsprechend $\|A^*\|_2 \le \|A\|_2$, das heißt $\|A\|_2$ und $\|A^*\|_2$ stimmen überein.

Wiederum eingesetzt in (2.4) folgt schließlich

$$\|A^*y - \mu x\|_2^2 = \|A^*y\|_2^2 - \mu^2 \le \|A^*\|_2^2 \|y\|_2^2 - \mu^2 = \|A\|_2^2 - \mu^2 = 0 .$$

Also ist $A^*y = \mu x$ beziehungsweise $A^*Ax = \mu^2 x$. Mit anderen Worten, x ist ein Eigenvektor von A^*A zum Eigenwert $\mu^2 = \|A\|_2^2$. Es verbleibt nun lediglich noch zu zeigen, daß $\|A\|_2^2$ der betragsgrößte Eigenwert von A^*A ist. Dies folgt jedoch unmittelbar aus (2.2), denn ist z mit $\|z\|_2 = 1$ ein beliebiger Eigenvektor von A^*A zum Eigenwert λ, dann ergibt (2.2)

$$\mu^2 = \|A\|_2^2 \ge |z^*(A^*A)z| = |z^*\lambda z| = |\lambda|\, \|z\|_2^2 = |\lambda| . \qquad \square$$

Satz 2.7 erklärt, warum die durch die Euklidnorm induzierte Norm Spektralnorm genannt wird, allerdings ist nicht das Spektrum von A sondern das von A^*A entscheidend. An dieser Stelle sei darauf hingewiesen, daß jede hermitesche Matrix (wie A^*A) eine Orthonormalbasis aus Eigenvektoren besitzt und die zugehörigen Eigenwerte allesamt reell sind. Ist die Matrix zudem positiv (semi)definit (wie A^*A), d. h. ist x^*Ax für alle $x \neq 0$ positiv (nichtnegativ), dann sind die Eigenwerte positiv (nichtnegativ). Mit Hilfe dieser Eigenschaft ist ein etwas einfacherer Beweis von Satz 2.7 möglich (vgl. Aufgabe 14).

Die Berechnung der Spektralnorm einer Matrix ist aufgrund von Satz 2.7 wesentlich aufwendiger als die der Zeilen- oder Spaltensummennorm; schließlich

müssen hierzu die Eigenwerte von A^*A bestimmt werden. Häufig ist man daher mit guten Abschätzungen zufrieden. In vielen Fällen ist zum Beispiel die folgende obere Schranke für die Spektralnorm ausreichend.

Satz 2.8. *Für $A \in \mathbb{K}^{m\times n}$ gilt $\|A\|_2 \le \big(\|A\|_1 \|A\|_\infty\big)^{1/2}$.*

Beweis. Nach Satz 2.7 ist $\|A\|_2^2$ der größte Eigenwert von A^*A; sei x ein zugehöriger Eigenvektor mit $\|x\|_1 = 1$. Da nach Beispiel 2.6 bzw. der anschließenden Bemerkung die Spaltensummennorm durch die Betragssummennormen in $\mathbb{K}^m$ und $\mathbb{K}^n$ induziert wird, gilt

$$\|A\|_2^2 = \|A^*Ax\|_1 \le \|A^*\|_1 \|Ax\|_1 \le \|A^*\|_1 \|A\|_1 \|x\|_1 = \|A^*\|_1 \|A\|_1 .$$

Wegen $\|A^*\|_1 = \|A\|_\infty$ ist dies gerade die Behauptung. □

Wie angekündigt verwenden wir nun Normen, um die Kondition eines Problems $x \mapsto F(x)$ abzuschätzen, wenn $x \in \mathcal{D}(F)$ und $F(x)$ jeweils n-dimensionale Vektoren sind. Wir beschränken uns auf den Spezialfall, in dem ein lineares Gleichungssystem $Az = b$ zu lösen ist, wobei $A \in \mathbb{K}^{n\times n}$ invertierbar sei. In diesem Fall ist also $F(b) = A^{-1}b$. Bei einem Eingangsfehler $\triangle b$ ergibt sich

$$z = A^{-1}b \qquad \text{und} \qquad z + \triangle z = A^{-1}(b + \triangle b) = A^{-1}b + A^{-1}\triangle b ,$$

das heißt die berechnete Lösung $z + \triangle z$ enthält den fortgepflanzten Fehler

$$\triangle z = A^{-1}\triangle b .$$

Sind nun $\|\cdot\|_M$ und $\|\cdot\|$ ein verträgliches Matrix-/Vektornormpaar, dann folgt

$$\begin{aligned} \frac{\|\triangle z\|}{\|z\|} &= \frac{\|A^{-1}\triangle b\|}{\|z\|} \le \|A^{-1}\|_M \frac{\|\triangle b\|}{\|b\|} \frac{\|Az\|}{\|z\|} \\ &\le \|A^{-1}\|_M \|A\|_M \frac{\|\triangle b\|}{\|b\|} . \end{aligned} \tag{2.5}$$

Definition 2.9. Der Faktor

$$\operatorname{cond}_M(A) = \|A^{-1}\|_M \|A\|_M$$

wird als *Kondition* der Matrix A bezüglich der Norm $\|\cdot\|_M$ bezeichnet.

Aus der Abschätzung (2.5) wird ersichtlich, daß die Kondition einer Matrix eine Art relative Konditionszahl für die Lösung eines linearen Gleichungssystems ist, vgl. Definition 1.1. Sie beschreibt die *schlimmstmögliche* Fortpflanzung des Eingangsfehlers beim Lösen des linearen Gleichungssystems. Falls übrigens $\|\cdot\|_M$ durch eine Vektornorm induziert wird, kann man Beispiele für b und $\triangle b$ konstruieren, für die in (2.5) Gleichheit herrscht, vgl. Aufgabe 16.

Aufgaben

1. Zeigen Sie, daß unter der Modellannahme (1.2) die Multiplikation $f(x) = ax$ rückwärts stabil ist. Unter welchen Einschränkungen gilt dies auch für die Addition $g(x) = x + a$?

2. Vergleichen Sie die beiden Rechenvorschriften

$$f(x) \;=\; \sin x - \sin y \qquad \text{und} \qquad g(x) \;=\; 2 \sin \frac{x-y}{2} \cos \frac{x+y}{2}$$

für festes $y \in (0, \pi/2)$ und $x \approx y$. In exakter Arithmetik stimmen f und g überein. Zeigen Sie, daß zwar beide Vorschriften rückwärts stabil sind, aber nur g ein Resultat im Bereich der Maschinengenauigkeit liefert.

3. Schreiben Sie ein Programm, daß die ersten $n = 10^6$ Terme der Reihe

$$\sum_{k=0}^{n} \frac{(-1)^k}{2k+1} \;=\; \frac{\pi}{4}$$

aufsummiert. Summieren Sie einmal in aufsteigender und einmal in absteigender Reihenfolge. Erwarten Sie einen Unterschied? Interpretieren Sie Ihre numerischen Ergebnisse.

4. Plotten Sie das Polynom

$$p(x) \;=\; 223200658\, x^3 \;-\; 1083557822\, x^2 \;+\; 1753426039\, x \;-\; 945804881$$

(vgl. Rump [94]) im Intervall $[1.61801916, 1.61801917]$. Wieviele Nullstellen liegen in diesem Intervall? Überprüfen Sie Ihre Vermutung mit einem Computeralgebraprogramm.

5. Zeigen Sie:
(a) Die Gesamtnorm

$$\|A\|_G \;=\; \sqrt{mn} \max_{1 \le i \le m} \max_{1 \le j \le n} |a_{ij}|$$

ist eine Norm auf $\mathbb{K}^{m \times n}$, die für $m = n$ submultiplikativ ist.
(b) Alle Normen auf $\mathbb{K}^{m \times n}$ sind zueinander äquivalent und Konvergenz in $\mathbb{K}^{m \times n}$ ist bezüglich jeder Norm äquivalent zur komponentenweisen Konvergenz.

6. Zeigen Sie, daß die Maximumnormen in $\mathbb{K}^m$ und $\mathbb{K}^n$ die Zeilensummennorm in $\mathbb{K}^{m \times n}$ induzieren.

7. Weisen Sie nach, daß zu jeder submultiplikativen Norm $\|\cdot\|_M$ auf $\mathbb{K}^{n \times n}$ eine Norm auf $\mathbb{K}^n$ existiert, mit der die Norm $\|\cdot\|_M$ verträglich ist. Gibt es auch immer eine Norm in $\mathbb{K}^n$, die $\|\cdot\|_M$ induziert?

8. Beweisen Sie die folgenden Eigenschaften der Spektralnorm:
(a) Ist $A \in \mathbb{K}^{m \times n}$ beliebig, dann gilt $\|A\|_2 = \max |y^* A x|$, wobei das Maximum über alle $x \in \mathbb{K}^n$ und alle $y \in \mathbb{K}^m$ mit $\|x\|_2 = \|y\|_2 = 1$ gebildet wird.
(b) Ist $A \in \mathbb{K}^{n \times n}$ hermitesch, dann gilt $\|A\|_2 = \max |x^* A x|$, wobei $x \in \mathbb{K}^n$ alle Vektoren mit Euklidnorm $\|x\|_2 = 1$ durchläuft.

9. Betrachten Sie für Matrizen $A \in \mathbb{K}^{n\times n}$ den Spektralradius $\varrho(A)$ und den *numerischen Radius*

$$r(A) = \sup_{0\neq x\in\mathbb{C}^n} \Big| \frac{x^*Ax}{x^*x} \Big| .$$

Ist einer dieser beiden Radien eine Norm in $\mathbb{K}^{n\times n}$? Überprüfen Sie gegebenenfalls die Verträglichkeit mit der Euklidnorm.

10. Sei $\|\cdot\|$ eine Vektornorm und $|\!|\!|\cdot|\!|\!|$ die induzierte Matrixnorm.
(a) Zeigen Sie, daß für jede nichtsinguläre Matrix $S \in \mathbb{C}^{n\times n}$ durch $\|x\|_S = \|Sx\|$ eine Vektornorm definiert wird und $|\!|\!|A|\!|\!|_S = |\!|\!|SAS^{-1}|\!|\!|$ die zugehörige induzierte Matrixnorm ist.
(b) Sei $A \in \mathbb{K}^{n\times n}$ und $V^{-1}AV = J$ die Jordan-Normalform von A. Betrachten Sie die Zeilensummennorm $|\!|\!|\cdot|\!|\!| = \|\cdot\|_\infty$ und die Matrix $S = D^{-1}V^{-1}$ mit

$$D = \begin{bmatrix} \varepsilon & & 0 \\ & \varepsilon^2 & \\ & \ddots & \\ 0 & & \varepsilon^n \end{bmatrix} \qquad \text{und} \qquad 0 < \varepsilon < 1 .$$

Weisen Sie die Ungleichung $|\!|\!|A|\!|\!|_S \leq \varrho(A) + \varepsilon$ nach.

11. Unter $\mathrm{Spur}(B)$ einer Matrix $B \in \mathbb{K}^{n\times n}$ versteht man die Summe aller Diagonalelemente von B. Beweisen Sie, daß für $A \in \mathbb{K}^{m\times n}$

$$\|A\|_F^2 = \mathrm{Spur}(A^*A) = \sum_{\lambda\in\sigma(A^*A)} \lambda .$$

Verwenden sie dies für einen Beweis der Abschätzung $\|A\|_2 \leq \|A\|_F$ für alle $A \in \mathbb{K}^{m\times n}$. (Alternativ folgt dies auch unmittelbar aus der Veträglichkeit der Frobeniusnorm mit der Euklidnorm.)

12. Betrachten Sie die Bidiagonalmatrix

$$L = \begin{bmatrix} 1 & & & \\ 1 & 1 & & \\ & \ddots & \ddots & \\ & & 1 & 1 \end{bmatrix} \in \mathbb{R}^{n\times n} .$$

(a) Weisen Sie die Abschätzung $1 \leq \|L\|_2 \leq 2$ nach.
(b) Zeigen Sie, daß Konstanten $c, C > 0$ existieren mit $cn \leq \|L^{-1}\|_2 \leq Cn$.
Für den genauen Wert von $\|L^{-1}\|_2$ vergleiche Aufgabe VI.14.

13. Seien $\|\cdot\|_M$ und $\|\cdot\|$ ein verträgliches Matrix-/Vektornormpaar in $\mathbb{K}^{n\times n}$ und $\mathbb{K}^n$. Zeigen Sie:
(a) $\|I\|_M \geq 1$ und $\mathrm{cond}_M(A) \geq 1$ für jede nichtsinguläre Matrix $A \in \mathbb{K}^{n\times n}$.
(b) Es gilt $\varrho(A) \leq \|A\|_M$ für alle $A \in \mathbb{K}^{n\times n}$.
(c) Ist $\|\cdot\|_M$ die durch $\|\cdot\|$ induzierte Matrixnorm, dann gilt $\|I\|_M = 1$ und $\mathrm{cond}_M(A) = 1$ genau dann, wenn A ein Vielfaches einer Isometrie ist, d. h. wenn $\|Ax\| = \mu\|x\|$ für alle $x \in \mathbb{K}^n$ und ein $\mu > 0$.

14. (a) Führen Sie einen alternativen Beweis von Satz 2.7, in dem Sie verwenden, daß eine Orthonormalbasis des $\mathbb{K}^n$ aus Eigenvektoren von A^*A existiert.
(b) Beweisen Sie für den Fall, daß A hermitesch und invertierbar ist die Darstellung

$$\operatorname{cond}_2(A) = |\lambda_1/\lambda_n|,$$

wobei λ_1 den betragsgrößten und λ_n den betragskleinsten Eigenwert von A bezeichnet.

15. Sei $\|\cdot\|$ eine Vektornorm in $\mathbb{K}^n$ und $\|\cdot\|_M$ eine mit ihr verträgliche Matrixnorm in $\mathbb{K}^{n\times n}$ sowie $A \in \mathbb{K}^{n\times n}$ eine nichtsinguläre Matrix. Zeigen Sie:
(a) Ist $\|\cdot\|_M$ submultiplikativ und $B \in \mathbb{K}^{n\times n}$ mit $\|A-B\|_M/\|A\|_M < 1/\operatorname{cond}_M(A)$, dann ist B nichtsingulär und es gilt

$$\frac{\|I\|_M}{1+\|A^{-1}(B-A)\|_M} \le \|B^{-1}A\|_M \le \frac{\|I\|_M}{1-\|A^{-1}(B-A)\|_M}.$$

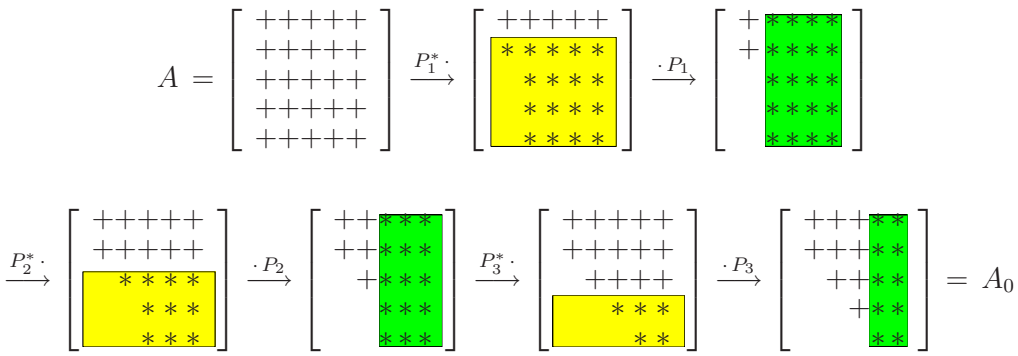
(b) Es existiert eine singuläre Matrix $B \in \mathbb{K}^{n\times n}$ mit $\|A-B\|_2/\|A\|_2 = 1/\operatorname{cond}_2(A)$.

16. Sei $\|\cdot\|$ eine Vektornorm in $\mathbb{K}^n$, $\|\cdot\|_M$ eine verträgliche Matrixnorm in $\mathbb{K}^{n\times n}$ und $A \in \mathbb{K}^{n\times n}$ nichtsingulär.
(a) Konstruieren Sie im Fall, daß $\|\cdot\|_M$ die induzierte Matrixnorm ist, Vektoren b und $\triangle b$ in $\mathbb{K}^n$, so daß in (2.5) Gleichheit herrscht.
(b) Zeigen Sie, daß unter der Voraussetzung $\|\triangle A\|_M\|A^{-1}\|_M < 1$ das gestörte lineare Gleichungssystem $(A+\triangle A)\tilde{z} = b+\triangle b$ eindeutig lösbar ist und daß für $0 \neq b \in \mathbb{K}^n$ und $z = A^{-1}b$ die Störung der Lösung, $\triangle z = \tilde{z} - z$, folgender Abschätzung genügt:

$$\frac{\|\triangle z\|}{\|z\|} \le \frac{\operatorname{cond}_M(A)\|I\|_M}{1-\operatorname{cond}_M(A)\frac{\|\triangle A\|_M}{\|A\|_M}}\Big(\frac{\|\triangle A\|_M}{\|A\|_M}+\frac{\|\triangle b\|}{\|b\|}\Big).$$

Hinweis: Aufgabe 15.

Algebraische Gleichungen



II Lineare Gleichungssysteme

Lineare Gleichungssysteme haben eine seltene Ausnahmestellung in der Mathematik, denn sie können durch Gauß-Elimination mit endlich vielen Elementaroperationen explizit gelöst werden (exakte Arithmetik vorausgesetzt). Diese Ausnahmestellung mag dazu verleiten, lineare Gleichungssysteme vom mathematischen Standpunkt aus als trivial anzusehen, und in der Praxis werden daher oft „irgendwelche“ Routinen aus einer Programmbibliothek zur Lösung solcher Systeme aufgerufen.

Da jedoch der Rechenaufwand in der Regel kubisch mit der Anzahl der Unbekannten anwächst, ist es gerade bei hochdimensionalen Gleichungssystemen entscheidend, strukturelle Eigenschaften zu berücksichtigen. In den folgenden Abschnitten werden daher verschiedene Algorithmen für spezielle Matrizenklassen vorgestellt.

Zur Vertiefung der hier vorgestellten Resultate sei auf das umfassende Buch von Golub und Van Loan [34] verwiesen.

3 Ein Beispiel aus der Mechanik

Wir eröffnen dieses Kapitel mit einem Beispiel aus der (linearen) Elastizitätstheorie. Probleme dieser Art treten in einer Vielzahl technischer Anwendungen auf und führen in der Praxis leicht zu Gleichungssystemen mit einigen tausend Gleichungen.

Die konkrete Aufgabenstellung in dem hier betrachteten Beispiel besteht darin, den Einfluß der Schwerkraft auf eine Brücke zu untersuchen. Die Brücke wird durch das zweidimensionale *Tragwerk* aus Abbildung 3.1 mit 18 elastischen *Stäben* und acht *Gelenken* modelliert, das an zwei *Lagern* rechts und links verankert ist.[1]

[1]Die Ähnlichkeit dieser Konstruktion mit der Brücke aus der Animation `truss` in MATLAB ist beabsichtigt, vgl. Abschnitt 22 in Kapitel V. Ähnliche Beispiele finden sich auch in [62, 86].

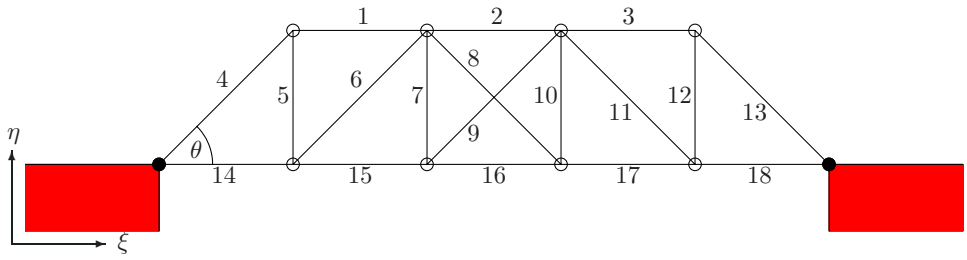


Abb. 3.1: Gerüst der Brücke

Wir assoziieren die Gelenke des Tragwerks (in der Abbildung durch nicht ausgefüllte Kreise dargestellt) mit Punkten $z_i = [\xi_i, \eta_i]^T \in \mathbb{R}^2$ und zählen dabei den Index i von links nach rechts hoch, d. h. wir bezeichnen mit $z_1, \ldots, z_4$ die vier Gelenke in der oberen Reihe und mit $z_5, \ldots, z_8$ die vier Gelenke in der unteren Reihe. Die Stäbe numerieren wir wie in der Abbildung angegeben: im unbelasteten (abgebildeten) Zustand haben sie die Längen $l_1, \ldots, l_{18}$.

Unter der Annahme, daß die Masse zu gleichen Teilen in den Gelenken konzentriert ist, während die Stäbe masselos sind, können wir uns unter dem Tragwerk auch ein *Masse-Feder-System* (mit sehr steifen Federn) vorstellen. Wir nehmen weiterhin an, daß äußere Kräfte $p_1, \ldots, p_8 \in \mathbb{R}^2$ an den Gelenken z_1 bis z_8 angreifen – im Beispiel wird dies die Schwerkraft sein – und wollen die entstehende Deformation berechnen. Die äußeren Kräfte verteilen sich nämlich als innere Kräfte auf die einzelnen Stäbe, die dadurch gestreckt oder gestaucht werden und somit für die Deformation des gesamten Tragwerks verantwortlich sind.

Die Berechnung der inneren Kräfte erfolgt mit Hilfe des *Schnittprinzips*: Ein einzelnes Gelenk wird freigeschnitten, indem alle Verbindungsstäbe zu Nachbargelenken durchtrennt werden; die inneren Kräfte kompensieren den Wegfall der festen Verbindungen an den Schnittenden. Dieses Vorgehen ist in Abbildung 3.2 für das erste Gelenk illustriert. Da Stäbe Kräfte nur in ihrer Längsrichtung aufnehmen können, zeigen diese inneren Kräfte in die jeweilige Stabrichtung; sie können daher durch skalare Größen f_k, $k = 1, \ldots, 18$, repräsentiert werden, deren Vorzeichen für *Zugkräfte* (vom Gelenk weg zeigend) positiv und *Druckkräfte* negativ gesetzt wird.

Das freigeschnittene System soll sich im *statischen Gleichgewicht* befinden. Daraus ergibt sich etwa für das erste Gelenk die Gleichung

$$0 = p_1 + f_1 \begin{bmatrix} 1 \\ 0 \end{bmatrix} + f_4 \begin{bmatrix} -c \\ -s \end{bmatrix} + f_5 \begin{bmatrix} 0 \\ -1 \end{bmatrix},$$

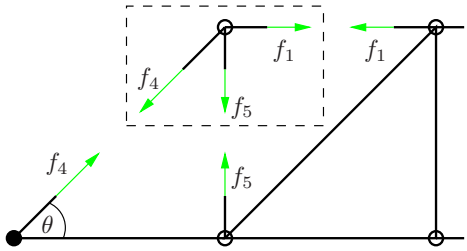


Abb. 3.2:
Freigeschnittenes Gelenk 1

wobei $c = \cos\theta$, $s = \sin\theta$ und θ der in Abbildung 3.1 eingezeichnete Winkel zwischen den einzelnen Querstreben und der horizontalen ξ-Achse ist. Entsprechend ergibt die Kräftebilanz für die anderen Gelenke

$$
\begin{aligned}
0 &= p_2 + f_1 \begin{bmatrix} -1 \\ 0 \end{bmatrix} + f_2 \begin{bmatrix} 1 \\ 0 \end{bmatrix} + f_6 \begin{bmatrix} -c \\ -s \end{bmatrix} + f_7 \begin{bmatrix} 0 \\ -1 \end{bmatrix} + f_8 \begin{bmatrix} c \\ -s \end{bmatrix}, \\
0 &= p_3 + f_2 \begin{bmatrix} -1 \\ 0 \end{bmatrix} + f_3 \begin{bmatrix} 1 \\ 0 \end{bmatrix} + f_9 \begin{bmatrix} -c \\ -s \end{bmatrix} + f_{10} \begin{bmatrix} 0 \\ -1 \end{bmatrix} + f_{11} \begin{bmatrix} c \\ -s \end{bmatrix}, \\
0 &= p_4 + f_3 \begin{bmatrix} -1 \\ 0 \end{bmatrix} + f_{12} \begin{bmatrix} 0 \\ -1 \end{bmatrix} + f_{13} \begin{bmatrix} c \\ -s \end{bmatrix}, \\
0 &= p_5 + f_5 \begin{bmatrix} 0 \\ 1 \end{bmatrix} + f_6 \begin{bmatrix} c \\ s \end{bmatrix} + f_{14} \begin{bmatrix} -1 \\ 0 \end{bmatrix} + f_{15} \begin{bmatrix} 1 \\ 0 \end{bmatrix}, \\
0 &= p_6 + f_7 \begin{bmatrix} 0 \\ 1 \end{bmatrix} + f_9 \begin{bmatrix} c \\ s \end{bmatrix} + f_{15} \begin{bmatrix} -1 \\ 0 \end{bmatrix} + f_{16} \begin{bmatrix} 1 \\ 0 \end{bmatrix}, \\
0 &= p_7 + f_8 \begin{bmatrix} -c \\ s \end{bmatrix} + f_{10} \begin{bmatrix} 0 \\ 1 \end{bmatrix} + f_{16} \begin{bmatrix} -1 \\ 0 \end{bmatrix} + f_{17} \begin{bmatrix} 1 \\ 0 \end{bmatrix}, \\
0 &= p_8 + f_{11} \begin{bmatrix} -c \\ s \end{bmatrix} + f_{12} \begin{bmatrix} 0 \\ 1 \end{bmatrix} + f_{17} \begin{bmatrix} -1 \\ 0 \end{bmatrix} + f_{18} \begin{bmatrix} 1 \\ 0 \end{bmatrix}.
\end{aligned}
$$

Sammeln wir die acht mal zwei vorgegebenen äußeren Kraftkomponenten in der obigen Reihenfolge im Vektor $p \in \mathbb{R}^{16}$ und setzen $f = [f_k] \in \mathbb{R}^{18}$, so kann die Kräftebilanz in dem Gleichungssystem

$$p = Ef \tag{3.1}$$

mit

$$E \;=\; \begin{bmatrix}
-1 & & c & & & & & & & & & & & & \\
 & & s & 1 & & & & & & & & & & & \\
1 & -1 & & & c & & -c & & & & & & & & \\
 & & & & s & 1 & s & & & & & & & & \\
 & 1 & -1 & & & & & c & & -c & & & & & \\
 & & & & & & & s & 1 & s & & & & & \\
 & & 1 & & & & & & & & -c & & & & \\
 & & & & & & & & & 1 & s & & & & \\
 & & & & -c & & & & & & & 1 & -1 & & \\
 & & & -1 & -s & & & & & & & & & & \\
 & & & & & & -c & & & & & & 1 & -1 & \\
 & & & & & -1 & & -s & & & & & & & \\
 & & & & & & c & & & & & & & 1 & -1 \\
 & & & & & & -s & & -1 & & & & & & \\
 & & & & & & & & & c & & & & & 1 & -1 \\
 & & & & & & & & & -s & -1 & & & &
\end{bmatrix}$$

zusammengefaßt werden. E heißt *Gleichgewichtsmatrix*; der Übersichtlichkeit halber sind die Nulleinträge von E nicht dargestellt. Das so bestimmte Kräftegleichgewicht ist allerdings nur für kleine Auslenkungen eine akzeptable Näherung, da Veränderungen der Gelenkpositionen und ihre Auswirkungen auf die Richtungen der Kräfte nicht berücksichtigt worden sind.

Durch die Deformation des Tragwerks ändert sich die Länge l_k des k-ten Stabs in $l_k + d_k$. Diese Deformation des Stabs ist mit der inneren Kraft f_k über ein Materialgesetz gekoppelt. Im einfachsten Modell linear-elastischer Stäbe ist dies das *Hookesche Gesetz*, wonach die innere Kraft in einem Stab proportional zu seiner *Verzerrung*, d. h. seiner relativen Längenänderung, ist:

$$f_k \;=\; \eta\,\frac{d_k}{l_k}\,. \tag{3.2}$$

Der Proportionalitätsfaktor η (der sogenannte *Youngsche Elastizitätsmodul*) beschreibt die Steifigkeit des Stabs.

Die Längenänderungen berechnen sich aus den alten Positionen z_i und den neuen Positionen $z_i + x_i$ der einzelnen Gelenke; x_i, $i = 1, \ldots, 8$, sind die sogenannten *Verschiebungen*. Verbindet etwa der k-te Stab die Gelenke z_i und z_j, dann errechnet sich folglich die neue Länge $l_k + d_k$ aus

$$\begin{aligned}
(l_k + d_k)^2 \;&=\; \|z_i + x_i - z_j - x_j\|_2^2 \\
&=\; \|z_i - z_j\|_2^2 \;+\; 2(z_i - z_j)^*(x_i - x_j) \;+\; \|x_i - x_j\|_2^2\,.
\end{aligned}$$

Durch Vernachlässigung der quadratischen Terme d_k^2 und $\|x_i - x_j\|_2^2$ auf der linken und rechten Seite ergibt sich wegen $l_k = \|z_i - z_j\|_2$ in erster Näherung

$$d_k \;\approx\; \frac{(z_i - z_j)^*(x_i - x_j)}{\|z_i - z_j\|_2} \;=\; \Big(\frac{z_i - z_j}{\|z_i - z_j\|_2}\Big)^*(x_i - x_j)\,. \tag{3.3}$$

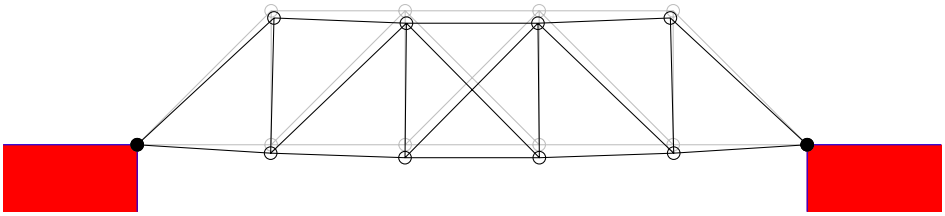


Abb. 3.3: Deformation der Brücke aufgrund der Schwerkraft

Der Vektor $(z_i - z_j)/\|z_i - z_j\|_2$ in (3.3) ist gerade wieder der Richtungsvektor des k-ten Stabs mit entsprechendem Vorzeichen. Werden die Verschiebungen wieder in einem Vektor $x \in \mathbb{R}^{16}$ zusammengefaßt (und zwar in der gleichen Reihenfolge wie im Vektor p), ergibt sich aus (3.3) die Beziehung

$$d = E^* x \tag{3.4}$$

zwischen dem Vektor $d = [d_k] \in \mathbb{R}^{18}$ und dem Verschiebungsvektor x mit derselben Gleichgewichtsmatrix E wie zuvor.

Bezeichnet L die Diagonalmatrix mit den Längen l_k auf der Diagonalen, so lautet das Hookesche Gesetz (3.2)

$$f = \eta L^{-1} d\,,$$

und aus (3.1) und (3.4) folgt somit die Gleichung

$$p = Ax \qquad \text{mit} \quad A = \eta E L^{-1} E^* \in \mathbb{R}^{16\times 16} \tag{3.5}$$

zwischen dem Vektor p der äußeren Kräfte und dem Verschiebungsvektor x. Die Matrix A ist die *Steifigkeitsmatrix* des Tragwerks. Sie ist invertierbar (vgl. Aufgabe 1), so daß nach (3.5) zu jedem Kraftvektor p genau ein resultierender Verschiebungsvektor x gehört.

Die Beziehung (3.5) modelliert die Realität im Rahmen der *linearen* Elastizitätstheorie, da einerseits bei der Berechnung der Längenänderungen d_k quadratische Anteile vernachlässigt wurden und andererseits die Winkeländerungen bei der Aufstellung des Kräftegleichgewichts nicht eingeflossen sind.

Beschreibt p die Schwerkraft, so weisen alle acht Kraftvektoren p_i nach unten; genauer ist

$$p_i = \begin{bmatrix} 0 \\ -mg \end{bmatrix}, \tag{3.6}$$

wobei g die Erdbeschleunigung und m die Masse eines Gelenks bezeichnet. Die zugehörige Lösung x des linearen Gleichungssystems (3.5) enthält die einzelnen Verschiebungsvektoren x_i für die jeweiligen Gelenke der Brücke unter dem Einfluß der Schwerkraft. Abbildung 3.3 illustriert das Ergebnis.

4 Die LR-Zerlegung

Wir wenden uns nun dem wichtigsten Algorithmus zur Lösung linearer Gleichungssysteme zu, der *Gauß-Elimination*. Wir gehen davon aus, daß das Verfahren an sich bekannt ist und betonen vorrangig die Interpretation der Gauß-Elimination als Faktorisierung der Koeffizientenmatrix in zwei Dreiecksmatrizen, die sogenannte *LR-Zerlegung*.

Sei zunächst $x = [x_1, \ldots, x_n]^T \in \mathbb{K}^n$ ein beliebiger Vektor, dessen k-te Komponente $x_k \neq 0$ ist; sei dabei ein fest gewählter Index. Mit e_k bezeichnen wir wieder den k-ten kartesischen Einheitsvektor in $\mathbb{K}^n$ und definieren damit die $n \times n$-Matrix

$$\begin{aligned} L_k \;&=\; I - l_k e_k^*\,, \qquad l_k \;=\; [\,0\,,\ldots,0\,,l_{k+1,k}\,,\ldots,l_{nk}\,]^T \\ &\text{mit} \quad l_{jk} = x_j/x_k\,, \quad j = k+1,\ldots,n\,. \end{aligned} \tag{4.1}$$

Man rechnet unmittelbar nach, daß

$$L_k x \;=\; \begin{bmatrix} 1 & 0 & & & \cdots & 0 \\ 0 & \ddots & \ddots & & & \vdots \\ & \ddots & 1 & 0 & & \\ & & -l_{k+1,k} & 1 & \ddots & \\ \vdots & & \vdots & \ddots & \ddots & 0 \\ 0 & \cdots & -l_{nk} & \cdots & 0 & 1 \end{bmatrix} \begin{bmatrix} x_1 \\ \vdots \\ x_k \\ x_{k+1} \\ \vdots \\ x_n \end{bmatrix} \;=\; \begin{bmatrix} x_1 \\ \vdots \\ x_k \\ 0 \\ \vdots \\ 0 \end{bmatrix}. \tag{4.2}$$

Diese sogenannten *Eliminationsmatrizen* können also genutzt werden, um die unteren $n-k$ Einträge eines Spaltenvektors zu Null zu transformieren.

Sei nun $A = A_1 = [a_{ij}]$ eine $n \times n$-Matrix und $x = [a_{i1}] \in \mathbb{K}^n$ die erste Spalte von A_1. Wenn $a_{11} \neq 0$ ist, läßt sich $L_1 = I - l_1 e_1^*$ wie in (4.1) mit $k = 1$ definieren und man erhält

$$L_1 A_1 = \begin{bmatrix} 1 & 0 & \cdots & & 0 \\ -l_{21} & 1 & 0 & \cdots & 0 \\ -l_{31} & 0 & 1 & & 0 \\ \vdots & & \ddots & \ddots & 0 \\ -l_{n1} & 0 & \cdots & 0 & 1 \end{bmatrix} \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ a_{31} & a_{32} & \cdots & a_{3n} \\ \vdots & \vdots & & \vdots \\ a_{n1} & a_{n2} & \cdots & a_{nn} \end{bmatrix} = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ 0 & a_{22}^{(2)} & \cdots & a_{2n}^{(2)} \\ 0 & a_{32}^{(2)} & \cdots & a_{3n}^{(2)} \\ \vdots & \vdots & & \vdots \\ 0 & a_{n2}^{(2)} & \cdots & a_{nn}^{(2)} \end{bmatrix}$$

mit gewissen neuen Einträgen $a_{ij}^{(2)}$, $i,j \geq 2$. Die resultierende Matrix nennen wir

$$A_2 = L_1 A_1 \,. \tag{4.3}$$

Dies ist der erste Schritt der Gauß-Elimination. Wenn $a_{22}^{(2)} \neq 0$ ist, wird im zweiten Schritt $x = [a_{12}, a_{22}^{(2)}, \ldots, a_{n2}^{(2)}]^T$, also die zweite Spalte von A_2 gewählt. Mit der zugehörigen Matrix L_2 aus (4.1) ergibt sich dann entsprechend

$$A_3 = L_2 A_2 = \begin{bmatrix} 1 & 0 & & & \cdots & 0 \\ 0 & 1 & & & & \vdots \\ & -l_{32} & 1 & & & \\ & -l_{42} & 0 & 1 & & \\ \vdots & \vdots & & \ddots & \ddots & 0 \\ 0 & -l_{n2} & & \cdots & 0 & 1 \end{bmatrix} A_2 = \begin{bmatrix} a_{11} & a_{12} & a_{13} & \cdots & a_{1n} \\ 0 & a_{22}^{(2)} & a_{23}^{(2)} & \cdots & a_{2n}^{(2)} \\ 0 & 0 & a_{33}^{(3)} & \cdots & a_{3n}^{(3)} \\ 0 & 0 & a_{43}^{(3)} & \cdots & a_{4n}^{(3)} \\ \vdots & \vdots & \vdots & & \vdots \\ 0 & 0 & a_{n3}^{(3)} & \cdots & a_{nn}^{(3)} \end{bmatrix}.$$

In dieser Weise fortfahrend (immer vorausgesetzt, daß das sogenannte *Pivotelement* $a_{ii}^{(i)}$ von Null verschieden ist) erhält man nach $n-1$ Transformationen schließlich eine obere Dreiecksmatrix $R = A_n$ und es gilt

$$R = L_{n-1} A_{n-1} = L_{n-1} L_{n-2} \cdots L_1 A$$

beziehungsweise

$$A = LR \qquad \text{mit} \qquad L = L_1^{-1} L_2^{-1} \cdots L_{n-1}^{-1} \,. \tag{4.4}$$

Die inversen Matrizen L_i^{-1} sowie die Matrix L können dabei explizit angegeben werden:

Lemma 4.1. *Es ist* $L_i^{-1} = I + l_i e_i^*$ *und* $L = I + l_1 e_1^* + \ldots + l_{n-1} e_{n-1}^*$.

Beweis. Aufgrund der Nulleinträge in den Vektoren l_j und e_i ist

$$e_i^* l_j = 0 \qquad \text{für} \quad 1 \leq i \leq j \leq n \,. \tag{4.5}$$

Daraus folgt zunächst die erste Behauptung, denn

$$(I - l_i e_i^*)(I + l_i e_i^*) \;=\; I - l_i e_i^* + l_i e_i^* - l_i e_i^* l_i e_i^* \;=\; I - l_i e_i^* l_i e_i^* \;=\; I \,.$$

Die spezielle Form von L ergibt sich induktiv: Dazu nehmen wir an, daß

$$L_1^{-1} \cdots L_k^{-1} \;=\; I + l_1 e_1^* + \ldots + l_k e_k^*$$

für ein k mit $1 \le k < n$ gilt (für $k = 1$ ist dies nach dem ersten Teil des Lemmas erfüllt). Aus $L_{k+1}^{-1} = I + l_{k+1}e_{k+1}^*$ folgt dann

$$L_1^{-1} \cdots L_k^{-1} L_{k+1}^{-1} \;=\; (I + l_1 e_1^* + \ldots + l_k e_k^*)(I + l_{k+1} e_{k+1}^*)\,,$$

und wegen (4.5) ergibt dies

$$\begin{aligned} L_1^{-1} \ldots L_{k+1}^{-1} &= I + l_1 e_1^* + \ldots + l_k e_k^* + l_{k+1} e_{k+1}^* + \sum_{i=1}^{k} l_i e_i^* l_{k+1} e_{k+1}^* \\ &= I + l_1 e_1^* + \ldots + l_k e_k^* + l_{k+1} e_{k+1}^*\,. \end{aligned}$$

Damit ist die Induktionsbehauptung auch für $k+1$ erfüllt und die Aussage des Lemmas bewiesen. □

Wird im Verlauf der Gauß-Elimination ein Pivotelement $a_{ii}^{(i)}$, $1 \le i < n$, Null, dann bricht das Verfahren in dieser Form zusammen. Sind hingegen alle Pivotelemente für $i = 1, \ldots, n$ von Null verschieden, dann haben wir insgesamt das folgende Resultat bewiesen:

Satz 4.2. *Falls kein Pivotelement Null wird, bestimmt die Gauß-Elimination eine LR-Zerlegung*

$$A = LR = \begin{bmatrix} 1 & & & & \\ l_{21} & 1 & & 0 & \\ l_{31} & l_{32} & 1 & & \\ \vdots & \vdots & \ddots & \ddots & \\ l_{n1} & l_{n2} & \cdots & l_{n,n-1} & 1 \end{bmatrix} \begin{bmatrix} a_{11} & a_{12} & a_{13} & \cdots & a_{1n} \\ & a_{22}^{(2)} & a_{23}^{(2)} & \cdots & a_{2n}^{(2)} \\ & & a_{33}^{(3)} & \cdots & a_{3n}^{(3)} \\ & 0 & & \ddots & \vdots \\ & & & & a_{nn}^{(n)} \end{bmatrix}$$

mit invertierbaren Dreiecksmatrizen L und R.

Gleichungssysteme mit Dreiecksmatrizen können unmittelbar durch Vorwärts- bzw. Rückwärtssubstitution gelöst werden. Somit ermöglicht die LR-Zerlegung in einfacher Weise die Lösung eines linearen Gleichungssystems $Ax = b$, vgl. Algorithmus 4.1.

Aufwand. Der Hauptaufwand von Algorithmus 4.1 besteht in der LR-Zerlegung. Anhand von (4.2) und der speziellen Gestalt von A_k sieht man, daß die Matrix-Matrix-Multiplikation $A_{k+1} = L_k A_k$ genau $(n-k)^2$ Multiplikationen kostet. Dazu kommen noch die $n-k$ Divisionen aus (4.1) zur Berechnung von l_k. Somit ergibt sich also für die LR-Zerlegung ein Aufwand von

$$\sum_{k=1}^{n-1} (n-k+1)(n-k) \;=\; \sum_{j=1}^{n-1} (j+1)j \;=\; \frac{1}{3}\,n^3 - \frac{1}{3}\,n$$

Initialisierung: $A \in \mathbb{K}^{n\times n}$ erfülle die Voraussetzung von Satz 4.2 und $b \in \mathbb{K}^n$ sei gegeben
faktorisiere $A = LR$ mit der Gauß-Elimination
$Ly = b$ % bestimme y durch Vorwärtssubstitution
$Rx = y$ % berechne x durch Rückwärtssubstitution

Ergebnis: $x = A^{-1}b$, denn $Ax = (LR)x = L(Rx) = b$

Algorithmus 4.1: Lösung linearer Gleichungssysteme (erste Fassung)

Multiplikationen/Divisionen. Demgegenüber sind die Kosten für die Berechnung der eigentlichen Lösung x vernachlässigbar: Bei der Vorwärtssubstitution ist zunächst für jeden Eintrag von L, der von Null und Eins verschieden ist, eine Multiplikation erforderlich; die gleiche Anzahl Multiplikationen wird für die Rückwärtssubstitution mit R gebraucht zuzüglich n Divisionen durch die Diagonalelemente. Insgesamt ergibt sich also ein zusätzlicher Aufwand von $n^2 - n$ Multiplikationen und n Divisionen für die Berechnung von x. ◊

Beispiel. Das Verfahren soll an einem einfachen Beispiel vorgeführt werden:

$$A = \begin{bmatrix} 1 & 0 & 1 \\ 2 & 2 & 2 \\ 1 & 8 & 0 \end{bmatrix} \overset{k=1}{=} \begin{bmatrix} 1 & 0 & 0 \\ 2 & 1 & 0 \\ 1 & 0 & 1 \end{bmatrix} \begin{bmatrix} 1 & 0 & 1 \\ 0 & 2 & 0 \\ 0 & 8 & -1 \end{bmatrix} \overset{k=2}{=} \begin{bmatrix} 1 & 0 & 0 \\ 2 & 1 & 0 \\ 1 & 4 & 1 \end{bmatrix} \begin{bmatrix} 1 & 0 & 1 \\ 0 & 2 & 0 \\ 0 & 0 & -1 \end{bmatrix} = LR\,.$$

Ist $b = [1, 1, 1]^T$ die rechte Seite des zugehörigen Gleichungssystems, so ergibt sich zunächst y aus $Ly = b$:

$$\begin{bmatrix} 1 & 0 & 0 \\ 2 & 1 & 0 \\ 1 & 4 & 1 \end{bmatrix} \begin{bmatrix} y_1 \\ y_2 \\ y_3 \end{bmatrix} = \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix}, \qquad \text{also} \quad y = \begin{bmatrix} 1 \\ -1 \\ 4 \end{bmatrix}.$$

Die Lösung x erhält man anschließend durch Rückwärtssubstitution aus dem Gleichungssystem $Rx = y$:

$$\begin{bmatrix} 1 & 0 & 1 \\ 0 & 2 & 0 \\ 0 & 0 & -1 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} 1 \\ -1 \\ 4 \end{bmatrix}, \qquad \text{also} \quad x = \begin{bmatrix} 5 \\ -1/2 \\ -4 \end{bmatrix}.$$

◊

Leider sind die Voraussetzungen von Satz 4.2 nicht immer erfüllt. Klar ist beispielsweise, daß sie bei einer singulären Matrix $A \in \mathbb{K}^{n\times n}$ nicht erfüllt sein können, denn sonst ergibt sich aus Satz 4.2 unmittelbar

$$0 = \det A = \det L \det R = 1 \prod_{i=1}^{n} a_{ii}^{(i)}\,,$$

also ein Widerspruch, da die rechte Seite von Null verschieden ist, wenn alle Pivotelemente ungleich Null sind. Es gibt aber auch *nichtsinguläre* Matrizen, die keine LR-Faktorisierung besitzen, zum Beispiel

$$A = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix} \neq \begin{bmatrix} 1 & 0 \\ l_{21} & 1 \end{bmatrix} \begin{bmatrix} r_{11} & r_{12} \\ 0 & r_{22} \end{bmatrix} = \begin{bmatrix} r_{11} & r_{12} \\ l_{21} r_{11} & l_{21} r_{12} + r_{22} \end{bmatrix}.$$

Durch Vergleich der $(1,1)$-Elemente von A und LR ergibt sich zwangsläufig $r_{11} = 0$ im Widerspruch zu der Bedingung $l_{21} r_{11} = 1$ für die Übereinstimmung der $(2,1)$-Elemente. Tatsächlich bricht die Gauß-Elimination in diesem Beispiel bereits im ersten Schritt zusammen, da das erste Pivotelement a_{11} Null ist.

4.1 Spaltenpivotsuche

Selbst wenn eine LR-Zerlegung existiert, können numerische Instabilitäten auftreten, falls im Verlauf der Gauß-Elimination ein Pivotelement sehr klein (also fast Null) wird.

Beispiel. Wir betrachten das 2×2 lineare Gleichungssystem $Ax = b$ mit

$$A = \begin{bmatrix} 10^{-3} & -1 \\ 1 & 2 \end{bmatrix} \quad \text{und} \quad b = \begin{bmatrix} -4 \\ 6 \end{bmatrix}.$$

Die Matrix A ist gut konditioniert: Aus der expliziten Formel für die Inverse von A erhält man

$$A^{-1} = \frac{1}{1 + 0.002} \begin{bmatrix} 2 & 1 \\ -1 & 10^{-3} \end{bmatrix} \approx \begin{bmatrix} 1.996 & 0.998 \\ -0.998 & 0.001 \end{bmatrix}$$

und somit ist $\operatorname{cond}_\infty(A) = \|A\|_\infty \|A^{-1}\|_\infty \approx 3 \cdot 3 = 9$. Gemäß (2.5) dürfen wir daher bei einem vorwärts stabilen Lösungsalgorithmus mit nicht viel mehr als einer Stelle Genauigkeitsverlust bei der berechneten Lösung rechnen. Für das obige b lautet die auf drei Stellen nach dem Komma gerundete Lösung

$$x = A^{-1} b \approx \begin{bmatrix} -1.996 \\ 3.998 \end{bmatrix}.$$

Nehmen wir nun an, wir hätten einen Rechner mit dreistelliger Dezimaldarstellung, also mit Maschinengenauigkeit $\mathsf{eps} \approx 5 \cdot 10^{-3}$. Dann ergibt die Gauß-Elimination zunächst

$$L^{-1} \boxdot A = \begin{bmatrix} 1 & 0 \\ -1000 & 1 \end{bmatrix} \boxdot \begin{bmatrix} 0.001 & -1 \\ 1 & 2 \end{bmatrix} = \begin{bmatrix} 0.001 & -1 \\ 0 & 1000 \end{bmatrix} = R;$$

dabei ist das Element r_{22} von R gerundet worden:

$$r_{22} \;=\; 1000 \boxplus 2 \;=\; \square(1002) \;=\; 1000\,.$$

Bei der Vorwärtssubstitution zur Berechnung von y ergibt sich ein weiterer Rundungsfehler,

$$y_1 = -4\,, \quad y_2 \;=\; 6 \boxminus (1000 \boxdot y_1) \;=\; 6 \boxplus 4000 \;=\; \square(4006) \;=\; 4010\,,$$

der dann schließlich auf die indiskutable Näherungslösung von $Rx = y$ führt:

$$\begin{aligned} x_2 \;&=\; 4010 \boxslash 1000 \;=\; 4.01\,, \\ x_1 \;&=\; 1000 \boxdot (-4 \boxplus x_2) \;=\; 1000 \boxdot 0.01 \;=\; 10\,. \end{aligned}$$

x_1 ist also völlig falsch, noch nicht einmal das Vorzeichen ist korrekt. Somit ist der Algorithmus 4.1 *nicht* vorwärts stabil. Der Grund dafür ist das kleine Pivotelement a_{11}, das einen großen Faktor $l_{21} = 1000$ und damit starke Fehlerverstärkung bewirkt. ◊

Zur Stabilisierung der Gauß-Elimination vertauscht man daher vor jedem Eliminationsschritt die i-te Zeile und eine andere Zeile mit dem Ziel, ein möglichst großes Pivotelement zu erhalten. In unserem Fall würden beispielsweise die beiden Zeilen von A vertauscht. Dann ergibt sich bei gleicher Rechengenauigkeit

$$\begin{bmatrix} 1 & 0 \\ -0.001 & 1 \end{bmatrix} \boxdot \begin{bmatrix} 1 & 2 \\ 0.001 & -1 \end{bmatrix} = \begin{bmatrix} 1 & 2 \\ 0 & -1 \end{bmatrix}$$

und dann durch Vorwärts- bzw. Rückwärtssubstitution (man beachte, daß natürlich auch die Komponenten der rechten Seite $[b_1, b_2]^T$ vertauscht werden müssen)

$$y_1 = 6\,, \qquad y_2 \;=\; -4 \boxminus 0.006 \;=\; -4.01\,,$$

und

$$x_2 \;=\; 4.01\,, \qquad x_1 = 6 \boxminus (2 \boxdot x_2) = 6 \boxminus 8.02 = -2.02\,.$$

Die Genauigkeit dieses Ergebnisses ist nun im Bereich des erwarteten Fehlers.

Offensichtlich ist jedoch das Attribut „groß“ für ein Pivotelement ein sehr relativer Begriff. Man rechnet nämlich leicht nach, daß sich genau der gleiche Vektor x ergibt, wenn die erste Zeile des Gleichungssystems vorab mit 10000 multipliziert wird und damit a_{11} im ersten Schritt das größte Element der ersten Spalte ist. Bewährt hat sich daher die *Spaltenpivotsuche* (*partial pivoting*), bei der im i-ten Teilschritt das Element $a_{ki}^{(i)}$ $(i \le k \le n)$ als Pivotelement

gewählt wird, das *relativ zur Betragssummennorm der jeweiligen Zeile* am betragsgrößten ist.

Mit Spaltenpivotsuche wird die Matrixformulierung der Gauß-Elimination komplizierter. Werden vor dem i-ten Eliminationsschritt beispielsweise die i-te und die j-te Zeile ($j > i$) der Matrix A_i vertauscht, dann kann dies durch die zugehörige Permutationsmatrix

$$P_i = \begin{bmatrix} 1 & & & & & & \\ & \ddots & & & & & \\ & & 1 & & & & \\ & & & 0 & & 1 & \\ & & & & 1 & & \\ & & & & & \ddots & \\ & & & & & & 1 \\ & & & 1 & & 0 & \\ & & & & & & 1 \\ & & & & & & & \ddots \\ & & & & & & & & 1 \end{bmatrix} \tag{4.6}$$

beschrieben werden (die eingezogenen Linien grenzen den Bereich zwischen i-ter und j-ter Zeile bzw. Spalte ab). Es gelten nämlich die folgenden Eigenschaften:

- Multiplikation einer Matrix A mit P_i von *links* entspricht einer Vertauschung der i-ten und j-ten *Zeile* von A;
- Multiplikation einer Matrix A mit P_i von *rechts* entspricht einer Vertauschung der i-ten und j-ten *Spalte* von A;
- insbesondere ist $P_i^2 = I$.

Werden also vor dem i-ten Eliminationsschritt die i-te und j-te Zeile von A vertauscht, bedeutet dies, daß in dem Eliminationsschritt die Matrix L_i von links an $P_i A_i$ heranmultipliziert wird, also

$$A_{i+1} = L_i P_i A_i \,. \tag{4.7}$$

Lemma 4.3. *Sei $k < i$, P_i durch* (4.6) *und L_k durch* (4.1) *gegeben. Dann ist $P_i L_k = L'_k P_i$, wobei L'_k bis auf eine Vertauschung von l_{ik} und l_{jk} wieder die Form* (4.1) *hat.*

Beweis. Wegen $P_i^2 = I$ gilt

$$P_i L_k = P_i L_k P_i^2 = (P_i L_k P_i) P_i \,,$$

und daraus folgt die gewünschte Matrixgleichung mit $L'_k = P_i L_k P_i$. Aufgrund der genannten Rechenregeln mit P_i folgt weiterhin, daß

$$L'_k = (P_i L_k) P_i = \left[\begin{array}{ccc|cc|c} \ddots & & & & & \\ & 1 & & & & \\ & -l_{k+1,k} & 1 & & & \\ & \vdots & & \ddots & & \\ \hdashline & -l_{jk} & & 0 & 1 & \\ & \vdots & & & \ddots & \\ & -l_{ik} & & 1 & 0 & \\ \hdashline & \vdots & & & & \ddots \\ & -l_{nk} & & & & 1 \end{array}\right] P_i = \left[\begin{array}{ccc|cc|c} \ddots & & & & & \\ & 1 & & & & \\ & -l_{k+1,k} & 1 & & & \\ & \vdots & & \ddots & & \\ \hdashline & -l_{jk} & & 1 & 0 & \\ & \vdots & & & \ddots & \\ & -l_{ik} & & 0 & 1 & \\ \hdashline & \vdots & & & & \ddots \\ & -l_{nk} & & & & 1 \end{array}\right].$$

Somit hat L'_k die Form (4.1), lediglich l_{jk} und l_{ik} sind vertauscht. □

Wir können nun den folgenden Satz für die Gauß-Elimination mit Spaltenpivotsuche beweisen.

Satz 4.4. *Ist A nichtsingulär, dann bestimmt die Gauß-Elimination mit Spaltenpivotsuche eine LR-Zerlegung $PA = \tilde{L}R$, wobei R wie zuvor die reduzierte obere Dreiecksmatrix A_n bezeichnet und $P = P_{n-1} \cdots P_1$ eine Permutationsmatrix ist. Die linke untere Dreiecksmatrix $\tilde{L}$ ergibt sich durch Vertauschen geeigneter Elemente in den Spalten der Matrix L aus Lemma* 4.1.

Beweis. Zunächst sei angenommen, daß die Gauß-Elimination mit Spaltenpivotsuche nicht zusammenbricht. Dann ergibt sich aus (4.7) durch sukzessive Anwendung von Lemma 4.3

$$\begin{aligned} R &= A_n = L_{n-1}P_{n-1}A_{n-1} = L_{n-1}P_{n-1}L_{n-2}P_{n-2}L_{n-3}P_{n-3}\cdots A \\ &= L_{n-1}\tilde{L}_{n-2}P_{n-1}P_{n-2}L_{n-3}P_{n-3}\cdots A \\ &= L_{n-1}\tilde{L}_{n-2}\tilde{L}_{n-3}P_{n-1}P_{n-2}P_{n-3}\cdots A\,. \end{aligned}$$

Hierbei ist $\tilde{L}_{n-2} = L'_{n-2}$ und $\tilde{L}_{n-3}$ die Matrix, die sich gemäß Lemma 4.3 nach dem Vertauschen von L_{n-3} mit P_{n-2} und P_{n-1} ergibt. Durch Auflösen der Rekursion erhält man schließlich

$$R = L_{n-1}\cdots\tilde{L}_1 P_{n-1}\cdots P_1 A = L_{n-1}\cdots\tilde{L}_1 PA\,.$$

Die Matrix $\tilde{L}$ aus der Formulierung des Satzes ist somit wie in Lemma 4.1 das Produkt der Inversen von $\tilde{L}_1$ bis L_{n-1}, d. h. auch die Elemente von $\tilde{L}$ unterscheiden sich von den Elementen von L aus Lemma 4.1 lediglich durch Permutationen innerhalb der einzelnen Spalten.

Zu klären bleibt schließlich noch, daß die Gauß-Elimination mit Spaltenpivotsuche nicht abbricht, also daß alle Pivotelemente nach der Spaltenpivotsuche von Null verschieden sind. Wäre etwa das Pivotelement im i-ten Teilschritt

tatsächlich Null, dann wären zwangsläufig wegen der Auswahlregel des Pivotelements *alle* Elemente $a_{ji}^{(i)}$, $j \geq i$, gleich Null, das heißt

$$A_i = \left[\begin{array}{ccc:ccc} a_{11} & \cdots & a_{1i} & & \cdots & a_{1n} \\ & \ddots & \vdots & & & \vdots \\ \hdashline & & 0 & a_{i,i+1}^{(i)} & \cdots & a_{i,n}^{(i)} \\ & 0 & \vdots & \vdots & & \vdots \\ & & 0 & a_{n,i+1}^{(i)} & \cdots & a_{nn}^{(i)} \end{array}\right] .$$

Die Determinante des rechten unteren quadratischen Blocks ist demnach Null und daher auch die Determinante von A_i. Nach dem Produktsatz für Determinanten folgt daraus aber

$$0 = \det A_i = \det\left(L_{i-1}P_{i-1}\cdots L_1P_1A\right) = \Big(\prod_{j=1}^{i-1}\underbrace{\det L_j}_{=1}\prod_{j=1}^{i-1}\underbrace{\det P_j}_{=\pm 1}\Big)\det A$$

im Widerspruch zu der Voraussetzung, daß A nichtsingulär ist. □

Beispiel. Das folgende Beispiel illustriert die Berechnung der Matrizen P, $\tilde{L}$ und R bei der Spaltenpivotsuche. Betrachtet wird die Matrix

$$A = \begin{bmatrix} 1 & 1 & 0 & 2 \\ 1/2 & 1/2 & 2 & -1 \\ -1 & 0 & -1/8 & -5 \\ 2 & -6 & 9 & 12 \end{bmatrix} .$$

Obwohl das $(1,1)$-Element kleiner als das $(4,1)$-Element ist, werden vor dem ersten Eliminationsschritt keine Zeilen vertauscht, da das Pivotelement auf der Grundlage der *relativen* Größe ausgewählt wird, bezogen auf die Betragssummennorm der jeweiligen Zeilen; unter diesem Gesichtspunkt ist das $(1,1)$-Element relativ am größten. Der erste Eliminationsschritt lautet daher

$$A = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 1/2 & 1 & 0 & 0 \\ -1 & 0 & 1 & 0 \\ 2 & 0 & 0 & 1 \end{bmatrix}\begin{bmatrix} 1 & 1 & 0 & 2 \\ 0 & 0 & 2 & -2 \\ 0 & 1 & -1/8 & -3 \\ 0 & -8 & 9 & 8 \end{bmatrix} = L_1^{-1}A_2 .$$

Das $(2,2)$-Element von A_2 ist Null, also muß ein anderes Pivotelement gewählt werden. Ein Vergleich der dritten und vierten Zeile ergibt die relativen Größen $1/(1+3+1/8) = 8/33$ für das $(3,2)$-Element und $8/(8+9+8) = 8/25$ für das $(4,2)$-Element. Folglich wird das $(4,2)$-Element ausgewählt und durch

Vertauschung der zweiten und vierten Zeile an die $(2,2)$-Position permutiert. Der zweite Eliminationsschritt lautet dann

$$P_2A_2 = \begin{bmatrix} 1 & 1 & 0 & 2 \\ 0 & -8 & 9 & 8 \\ 0 & 1 & -1/8 & -3 \\ 0 & 0 & 2 & -2 \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & -1/8 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} 1 & 1 & 0 & 2 \\ 0 & -8 & 9 & 8 \\ 0 & 0 & 1 & -2 \\ 0 & 0 & 2 & -2 \end{bmatrix} = L_2^{-1}A_3\,.$$

Vor dem letzten Eliminationsschritt müssen die dritte und vierte Zeile von A_3 aufgrund der Spaltenpivotsuche vertauscht werden und man erhält

$$P_3A_3 = \begin{bmatrix} 1 & 1 & 0 & 2 \\ 0 & -8 & 9 & 8 \\ 0 & 0 & 2 & -2 \\ 0 & 0 & 1 & -2 \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 1/2 & 1 \end{bmatrix} \begin{bmatrix} 1 & 1 & 0 & 2 \\ 0 & -8 & 9 & 8 \\ 0 & 0 & 2 & -2 \\ 0 & 0 & 0 & -1 \end{bmatrix} = L_3^{-1}R\,.$$

Um die Einträge von $\tilde{L}$ in der korrekten Reihenfolge zu erhalten, ist es am einfachsten, die strikten unteren Diagonaleinträge von L einfach in den entsprechenden Matrixeinträgen von A zu belassen, wie es in der folgenden Darstellung in den hellgrau hinterlegten Feldern illustriert wird. Bei einer Vertauschung zweier dunkel hinterlegter Zeilen aufgrund der Pivotstrategie werden die Einträge von L dann in der korrekten Weise mitvertauscht:

$$\begin{bmatrix} 1 & 1 & 0 & 2 \\ 1/2 & 1/2 & 2 & -1 \\ -1 & 0 & -1/8 & -5 \\ 2 & -6 & 9 & 12 \end{bmatrix} \longrightarrow \begin{bmatrix} 1 & 1 & 0 & 2 \\ 1/2 & 0 & 2 & -2 \\ -1 & 1 & -1/8 & -3 \\ 2 & -8 & 9 & 8 \end{bmatrix}$$

$$\longrightarrow \begin{bmatrix} 1 & 1 & 0 & 2 \\ 2 & -8 & 9 & 8 \\ -1 & 1 & -1/8 & -3 \\ 1/2 & 0 & 2 & -2 \end{bmatrix} \longrightarrow \begin{bmatrix} 1 & 1 & 0 & 2 \\ 2 & -8 & 9 & 8 \\ -1 & -1/8 & 1 & -2 \\ 1/2 & 0 & 2 & -2 \end{bmatrix}$$

$$\longrightarrow \begin{bmatrix} 1 & 1 & 0 & 2 \\ 2 & -8 & 9 & 8 \\ 1/2 & 0 & 2 & -2 \\ -1 & -1/8 & 1 & -2 \end{bmatrix} \longrightarrow \begin{bmatrix} 1 & 1 & 0 & 2 \\ 2 & -8 & 9 & 8 \\ 1/2 & 0 & 2 & -2 \\ -1 & -1/8 & 1/2 & -1 \end{bmatrix}$$

Somit ist

$$PA = \begin{bmatrix} 1 & 1 & 0 & 2 \\ 2 & -6 & 9 & 12 \\ 1/2 & 1/2 & 2 & -1 \\ -1 & 0 & -1/8 & -5 \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 2 & 1 & 0 & 0 \\ 1/2 & 0 & 1 & 0 \\ -1 & -1/8 & 1/2 & 1 \end{bmatrix} \begin{bmatrix} 1 & 1 & 0 & 2 \\ 0 & -8 & 9 & 8 \\ 0 & 0 & 2 & -2 \\ 0 & 0 & 0 & -1 \end{bmatrix}.$$

die gesuchte Faktorisierung $PA = \tilde{L}R$. ◊

Initialisierung: $A \in \mathbb{K}^{n\times n}$ erfülle die Voraussetzung von Satz 4.2 und $b \in \mathbb{K}^n$ sei gegeben
faktorisiere $PA = \tilde{L}R$ mit der Gauß-Elimination mit Spaltenpivotsuche
$\tilde{L}y = Pb$ % bestimme y durch Vorwärtssubstitution
$Rx = y$ % berechne x durch Rückwärtssubstitution

Ergebnis: $x = A^{-1}b$, denn $PAx = (\tilde{L}R)x = \tilde{L}(Rx) = Pb$

Algorithmus 4.2: Lösung linearer Gleichungssysteme

Für spezielle Matrizen kann auf die Spaltenpivotsuche verzichtet werden, da ohnehin niemals Zeilen vertauscht werden müssen. Hierzu gehören die strikt diagonaldominanten Matrizen.

Definition 4.5. Eine Matrix $A \in \mathbb{K}^{n\times n}$ heißt *strikt diagonaldominant*, falls

$$|a_{ii}| > \sum_{\substack{j=1\\ j\neq i}}^{n} |a_{ij}| \qquad \text{für alle } i = 1,\ldots,n\,.$$

Satz 4.6. *Ist A strikt diagonaldominant, dann wählt die Spaltenpivotsuche in jedem Eliminationsschritt der Gauß-Elimination das Diagonalelement $a_{ii}^{(i)}$ als Pivotelement aus. Insbesondere existiert also eine LR-Zerlegung von A und A ist nichtsingulär.*

Beweis. Betrachten wir zunächst die Auswahl des ersten Pivotelements. Da A strikt diagonaldominant ist, gilt

$$\| [a_{11},\ldots,a_{1n}]^T \|_1 \;=\; |a_{11}| + \sum_{j=2}^{n} |a_{1j}| \;<\; 2\,|a_{11}|\,,$$

d. h. das $(1,1)$-Element ist betragsmäßig mehr als halb so groß wie die Betragssummennorm der ersten Zeile. Entsprechend ergibt sich, daß alle anderen Einträge a_{i1} für $i > 1$ betragsmäßig höchstens halb so groß sind wie die Betragssummennormen der jeweiligen Zeilen:

$$2\,|a_{i1}| \;\leq\; |a_{i1}| \;+\; |a_{ii}| \;\leq\; \| [a_{i1},\ldots,a_{in}]^T \|_1\,.$$

Folglich wird vor dem ersten Eliminationsschritt das $(1,1)$-Element als Pivotelement ausgewählt.

Der Beweis läuft nun induktiv durch, wenn wir zeigen können, daß die rechte untere $(n-1)\times(n-1)$-Submatrix von A_2 wieder strikt diagonaldominant ist.

Dazu schreiben wir den ersten Eliminationsschritt in der Blockform

$$\left[\begin{array}{c|c} a_{11} & b^T \\ \hline a & B \end{array}\right] = \left[\begin{array}{c|c} 1 & 0 \\ \hline \dfrac{a}{a_{11}} & I \end{array}\right] \left[\begin{array}{c|c} a_{11} & b^T \\ \hline 0 & B - \dfrac{ab^T}{a_{11}} \end{array}\right] .$$

Der rechte untere Block von A_2 ist also gegeben durch $B - ab^T/a_{11}$, so daß das (i,j)-Element $a_{ij}^{(2)}$ von A_2 für $2 \le i,j \le n$ die folgende Form hat:

$$a_{ij}^{(2)} \;=\; a_{ij} - \frac{a_{i1}a_{1j}}{a_{11}}\,, \qquad 2 \le i,j \le n\,.$$

Die Untermatrix von A_2 ist demnach genau dann strikt diagonaldominant, wenn

$$\Big|\, a_{ii} - \frac{a_{i1}a_{1i}}{a_{11}} \,\Big| \;>\; \sum_{\substack{j=2\\ j\neq i}}^{n} \Big|\, a_{ij} - \frac{a_{i1}a_{1j}}{a_{11}} \,\Big|\,, \qquad i = 2,\ldots,n\,. \tag{4.8}$$

Wegen der strikten Diagonaldominanz von A gilt tatsächlich

$$\begin{aligned} \sum_{\substack{j=2\\ j\neq i}}^{n} \Big|\, a_{ij} - \frac{a_{i1}a_{1j}}{a_{11}} \,\Big| \;&\le\; \sum_{\substack{j=2\\ j\neq i}}^{n} |a_{ij}| \;+\; \Big|\, \frac{a_{i1}}{a_{11}} \,\Big| \sum_{\substack{j=2\\ j\neq i}}^{n} |a_{1j}| \\ &<\; \sum_{\substack{j=2\\ j\neq i}}^{n} |a_{ij}| \;+\; |a_{i1}|\, \frac{|a_{11}| \;-\; |a_{1i}|}{|a_{11}|} \\ &<\; |a_{ii}| \;-\; |a_{i1}| \;+\; |a_{i1}| \;-\; \frac{|a_{i1}a_{1i}|}{|a_{11}|} \;=\; |a_{ii}| \;-\; \frac{|a_{i1}a_{1i}|}{|a_{11}|}\,, \end{aligned}$$

und demnach folgt (4.8) aus der umgekehrten Dreiecksungleichung. Somit ist A_2 ebenfalls strikt diagonaldominant, was zu zeigen war.

Nach Satz 4.2 ergibt die Gauß-Elimination also eine Faktorisierung $A = LR$ mit nichtsingulären Matrizen L und R und damit ist auch A nichtsingulär. □

4.2 Totalpivotsuche

Mit der Spaltenpivotsuche ist die Gauß-Elimination in der Regel sehr zuverlässig, obwohl immer noch Beispiele konstruiert werden können, bei denen selbst diese Pivotwahl nicht stabil ist. In solchen Ausnahmefällen kann man statt dessen eine andere Pivotstrategie verfolgen, die sogenannte *Totalpivotsuche* (*total pivoting*).

Dabei wählt man vor dem i-ten Eliminationsschritt aus dem gesamten rechten unteren Matrixblock (also aus den Indizes (j,k) mit $i \le j,k \le n$) das Element $a_{jk}^{(i)}$ als Pivotelement aus, das betragsmäßig am größten ist. Das entsprechende Element, etwa $a_{jk}^{(i)}$, wird an die (i,i)-Position gebracht, indem wie zuvor die Zeilen j und i und zusätzlich noch die Spalten k und i vertauscht werden. Letzteres kann formal dadurch beschrieben werden, daß A mit einer Permutationsmatrix Q_i von rechts multipliziert wird (Q_i sieht wie die Permutationsmatrix in (4.6) aus, wobei k die Rolle von j übernimmt). Entsprechend zu (4.7) ergibt dies die Matrixtransformation

$$A_{i+1} = L_i P_i A_i Q_i \,,$$

und man erhält schließlich eine LR-Zerlegung der Matrix PAQ mit $Q = Q_1 \cdots Q_{n-1}$.

Wird die Totalpivotsuche zur Lösung eines linearen Gleichungssystems verwendet, dann entsprechen Spaltenvertauschungen Permutationen des Vektors x. Der Ergebnisvektor ist also nicht in der richtigen Reihenfolge und muß abschließend zurückpermutiert werden.

Satz 4.7. *Die Gauß-Elimination mit Totalpivotsuche ist rückwärts stabil.*

Für einen Beweis dieses Resultats sei auf das Buch von Wilkinson [108] verwiesen. Die Totalpivotsuche wird in der Praxis nur selten eingesetzt, da die Suche nach dem betragsgrößten Element im i-ten Schritt einem Aufwand $O(i^2)$ entspricht; der Gesamtaufwand der Pivotsuche ist also nicht mehr gegenüber der eigentlichen Rechnung vernachlässigbar.

4.3 Nachiteration

Zum Abschluß beschreiben wir noch eine einfache Möglichkeit zur Verbesserung der Genauigkeit einer mit der LR-Zerlegung berechneten Näherungslösung x des Gleichungssystems $A\widehat{x} = b$. Dieses Prinzip der *Nachiteration* aus Algorithmus 4.3 nutzt geschickt aus, daß die exakte Lösung $\widehat{x}$ in der Form

$$\widehat{x} \;=\; A^{-1}b \;=\; x + A^{-1}(b - Ax)$$

geschrieben werden kann; $r = b - Ax$ ist das sogenannte *Residuum* der Näherungslösung x.

Aufwand. Jeder Nachiterationsschritt ist relativ billig, da die Vorwärts- und Rückwärtssubstitution mit rund n^2 Multiplikationen auskommt (zum Vergleich: Die Berechnung der LR-Zerlegung ist eine Größenordnung teurer). ◊

Initialisierung: $A \in \mathbb{K}^{n\times n}$ sei nichtsingulär; gesucht ist die Lösung $\widehat{x}$ von $Ax = b$

bestimme Näherungslösung x mit Algorithmus 4.2

while x nicht genau genug **do** % Nachiteration
  berechne $r = b - Ax$ mit *doppelter Genauigkeit* % beachte: $b - Ax = A(\widehat{x} - x)$
  löse $Az = r$ wie in Algorithmus 4.2 % verwende die berechnete LR-Zerlegung
  $x = x + z$
end while

Algorithmus 4.3: Nachiteration

Eine heuristische Überlegung mag die Nachiteration erläutern. Dazu nehmen wir an, daß die Kondition der Matrix die Größenordnung $\mathrm{cond}_\infty(A) \approx 10^q$ hat, während die Rechnerarithmetik auf etwa d Dezimalstellen genau ist. Unter der Voraussetzung $d > q$ erwarten wir aufgrund der Fehlerabschätzung (2.5) für eine vorwärts stabile LR-Zerlegung etwa $d - q$ Stellen Genauigkeit in der Näherungslösung x. Da im Anschluß an die Berechnung von x das Residuum r mit doppelter Genauigkeit berechnet wird, können wir für das Residuum nach der Rundung wieder d Stellen Genauigkeit voraussetzen. Mit dem gleichen Argument wie zuvor werden dann von der exakten Lösung $\widehat{z} = \widehat{x} - x$ von $Az = r$ wieder $d - q$ Stellen korrekt berechnet, d. h. nach einem Schritt der Nachiteration kennen wir $2(d-q)$ Dezimalstellen von $\widehat{x}$. Entsprechend erhalten wir nach k Nachiterationsschritten $\widehat{x}$ auf $\min\{d, (k+1)(d-q)\}$ Stellen genau.

Es ist dabei von entscheidender Bedeutung, daß das Residuum so genau wie möglich berechnet wird: Daher muß r unbedingt mit doppelter Genauigkeit und mit der Originalmatrix A berechnet werden, nicht mit dem Produkt LR.

5 Die Cholesky-Zerlegung

Wir betrachten als nächstes eine „Blockversion“ der LR-Zerlegung. Dazu partitionieren wir ein gegebenes $A \in \mathbb{K}^{n\times n}$ in der Form

$$A = \begin{bmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{bmatrix} \qquad \text{mit nichtsingulärem } A_{11} \in \mathbb{K}^{p\times p}\,.$$

Demzufolge ist $A_{12} \in \mathbb{K}^{p\times(n-p)}$, $A_{21} \in \mathbb{K}^{(n-p)\times p}$ und $A_{22} \in \mathbb{K}^{(n-p)\times(n-p)}$. Bei der Block-$LR$-Zerlegung von A gehen wir analog zum vorigen Abschnitt vor und faktorisieren

$$A \;=\; \begin{bmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{bmatrix} \;=\; \begin{bmatrix} I & 0 \\ A_{21}A_{11}^{-1} & I \end{bmatrix} \begin{bmatrix} A_{11} & A_{12} \\ 0 & S \end{bmatrix}\,. \tag{5.1}$$

Der $(2,2)$-Block S der rechten oberen Block-Dreiecksmatrix kann explizit ausgerechnet werden,

$$S = A_{22} - A_{21}A_{11}^{-1}A_{12} \in \mathbb{K}^{(n-p)\times(n-p)} \tag{5.2}$$

und wird *Schur-Komplement* von A_{11} in A genannt.

Die Lösung eines linearen Gleichungssystems $Ax = b$ kann entsprechend durch (Block-)Vorwärts- und Rückwärtssubstitution erfolgen: Dazu werden die Vektoren x und $b \in \mathbb{K}^n$ konform in ihre ersten p Komponenten $x_1, b_1 \in \mathbb{K}^p$ und die restlichen Komponenten $x_2, b_2 \in \mathbb{K}^{n-p}$ unterteilt; die Vorwärtssubstitution ergibt dann Hilfsvektoren

$$\begin{aligned} y_1 &= b_1\,, \\ y_2 &= b_2 - A_{21}A_{11}^{-1}b_1\,, \end{aligned}$$

aus denen durch anschließende Rückwärtssubstitution das Ergebnis berechnet wird:

$$\begin{aligned} x_2 &= S^{-1}y_2 = S^{-1}(b_2 - A_{21}A_{11}^{-1}b_1)\,, \\ x_1 &= A_{11}^{-1}(b_1 - A_{12}x_2)\,. \end{aligned}$$

Letzteres ist allerdings nur möglich, wenn S nichtsingulär ist.

Lemma 5.1. *A sei hermitesch und positiv definit und $1 \le p \le n$. Dann ist die Submatrix $A_{11} \in \mathbb{K}^{p\times p}$ invertierbar und sowohl A_{11} als auch S sind hermitesch und positiv definit.*

Beweis. Wegen

$$\begin{bmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{bmatrix} = A = A^* = \begin{bmatrix} A_{11}^* & A_{21}^* \\ A_{12}^* & A_{22}^* \end{bmatrix}$$

ergibt sich

$$A_{11} = A_{11}^*\,, \qquad A_{22} = A_{22}^* \qquad \text{und} \qquad A_{12} = A_{21}^*\,.$$

Folglich ist A_{11} hermitesch und für einen beliebigen Vektor $x \in \mathbb{K}^p$ gilt

$$0 \le \begin{bmatrix} x \\ 0 \end{bmatrix}^* \begin{bmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{bmatrix} \begin{bmatrix} x \\ 0 \end{bmatrix} = \begin{bmatrix} x \\ 0 \end{bmatrix}^* \begin{bmatrix} A_{11}x \\ A_{21}x \end{bmatrix} = x^*A_{11}x$$

mit Gleichheit genau für $x = 0$. Das bedeutet, daß A_{11} ebenfalls positiv definit ist und A_{11}^{-1} existiert. S ist somit wohldefiniert mit

$$S^* = A_{22}^* - A_{12}^*A_{11}^{-*}A_{21}^* = A_{22} - A_{21}A_{11}^{-1}A_{12} = S\,.$$

Für den Nachweis, daß S positiv definit ist, definieren wir für ein beliebiges $y \in \mathbb{K}^{n-p}$ den Vektor $x = -A_{11}^{-1} A_{12}\, y \in \mathbb{K}^p$ und erhalten

$$\begin{aligned} 0 \;&\leq\; \begin{bmatrix} x \\ y \end{bmatrix}^* \begin{bmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} \;=\; \begin{bmatrix} x \\ y \end{bmatrix}^* \begin{bmatrix} A_{11}x + A_{12}y \\ A_{21}x + A_{22}y \end{bmatrix} \\ &=\; \begin{bmatrix} x \\ y \end{bmatrix}^* \begin{bmatrix} -A_{12}y + A_{12}y \\ -A_{21}A_{11}^{-1}A_{12}y + A_{22}y \end{bmatrix} \;=\; \begin{bmatrix} x \\ y \end{bmatrix}^* \begin{bmatrix} 0 \\ Sy \end{bmatrix} \\ &=\; y^* S y\,, \end{aligned}$$

wobei wiederum Gleichheit nur für $y = 0$ gelten kann. Somit ist S positiv definit. □

Im folgenden betrachten wir eine Variante der LR-Zerlegung.

Definition 5.2. Eine Faktorisierung $A = LL^*$ mit linker unterer Dreiecksmatrix L mit positiven Diagonaleinträgen heißt *Cholesky-Zerlegung* von A.

Eine notwendige Bedingung für die Existenz einer Cholesky-Zerlegung gibt das folgende Resultat.

Proposition 5.3. *Hat A eine Cholesky-Zerlegung, dann ist A hermitesch und positiv definit.*

Beweis. Aus $A = LL^*$ folgt unmittelbar $A^* = (L^*)^* L^* = LL^* = A$; also ist A hermitesch. Ferner ist

$$x^* A x \;=\; x^* L L^* x \;=\; (L^* x)^* L^* x \;=\; \|L^* x\|_2^2 \;\geq\; 0$$

für alle $x \in \mathbb{K}^n$ mit Gleichheit genau für $x = 0$, da L positive Diagonaleinträge haben soll und somit nichtsingulär ist. Folglich ist A positiv definit. □

Tatsächlich sind diese Bedingungen an A auch hinreichend.

Satz 5.4. *Ist A hermitesch und positiv definit, dann existiert eine Cholesky-Zerlegung von A.*

Beweis. Der Beweis wird induktiv über die Dimension n der Matrix geführt, wobei für $n = 1$ die „Matrix“ nur aus einem Element a_{11} besteht, das positiv sein muß, da A positiv definit ist. Also kann man für $n = 1$ einfach $L = [\,\sqrt{a_{11}}\,]$ setzen.

Sei nun die Behauptung für alle quadratischen Matrizen der Dimension $n-1$ korrekt und A eine beliebige $n \times n$ Matrix. Dann partitionieren wir

$$A = \begin{bmatrix} a_{11} & A_{12} \\ A_{21} & A_{22} \end{bmatrix} \qquad \text{mit } A_{22} \in \mathbb{K}^{(n-1)\times(n-1)} \text{ und } A_{12} = A_{21}^*\,.$$

Nach Lemma 5.1 ist das Schur-Komplement $S = A_{22} - A_{21}A_{12}/a_{11}$ von a_{11} in A hermitesch und positiv definit. Aufgrund der Induktionsannahme hat S daher eine Cholesky-Zerlegung $S = L_S L_S^*$. Mit $l_{11} = \sqrt{a_{11}}$,

$$L = \begin{bmatrix} l_{11} & 0 \\ A_{21}/l_{11} & L_S \end{bmatrix} \qquad \text{und} \qquad L^* = \begin{bmatrix} l_{11} & A_{12}/l_{11} \\ 0 & L_S^* \end{bmatrix}$$

folgt

$$LL^* = \begin{bmatrix} l_{11} & 0 \\ A_{21}/l_{11} & L_S \end{bmatrix} \begin{bmatrix} l_{11} & A_{12}/l_{11} \\ 0 & L_S^* \end{bmatrix} = \begin{bmatrix} l_{11}^2 & A_{12} \\ A_{21} & B \end{bmatrix}$$

mit

$$B = \frac{1}{l_{11}^2} A_{21}A_{12} + L_S L_S^* = \frac{1}{a_{11}} A_{21}A_{12} + S = A_{22}\,.$$

Also ist

$$LL^* = \begin{bmatrix} a_{11} & A_{12} \\ A_{21} & A_{22} \end{bmatrix} = A$$

eine Cholesky-Zerlegung von A. □

Die numerische Berechnung der Einträge von L kann sukzessive durch zeilenweisen Koeffizientenvergleich des Produkts $A = LL^*$ erfolgen,

$$\begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & & \vdots \\ a_{n1} & a_{n2} & \cdots & a_{nn} \end{bmatrix} = \begin{bmatrix} l_{11} & & & 0 \\ l_{21} & l_{22} & & \\ \vdots & & \ddots & \\ l_{n1} & l_{n2} & \cdots & l_{nn} \end{bmatrix} \begin{bmatrix} l_{11} & \overline{l_{21}} & \cdots & \overline{l_{n1}} \\ & l_{22} & & \overline{l_{n2}} \\ & & \ddots & \vdots \\ 0 & & & l_{nn} \end{bmatrix}.$$

Die Einträge von L ergeben sich somit in der folgenden Weise:

$$\begin{aligned} a_{11} &= |l_{11}|^2 && l_{11} = a_{11}^{1/2}\,; \\ a_{21} &= l_{21}\overline{l_{11}} && l_{21} = a_{21}/\overline{l_{11}}\,, \\ a_{22} &= |l_{21}|^2 + |l_{22}|^2 && l_{22} = \left(a_{22} - |l_{21}|^2\right)^{1/2}\,; \\ a_{31} &= l_{31}\overline{l_{11}} && l_{31} = a_{31}/\overline{l_{11}}\,, \\ a_{32} &= l_{31}\overline{l_{21}} + l_{32}\overline{l_{22}} && l_{32} = (a_{32} - l_{31}\overline{l_{21}})/\overline{l_{22}}\,, \\ a_{32} &= |l_{31}|^2 + |l_{32}|^2 + |l_{33}|^2 && l_{33} = \left(a_{33} - |l_{31}|^2 - |l_{32}|^2\right)^{1/2}\,; \\ &\vdots && \vdots \end{aligned}$$

Die Lösbarkeit dieser (nichtlinearen) Gleichungen ist durch den Existenzbeweis (Satz 5.4) gewährleistet, d. h. alle Quadratwurzeln existieren und die resultierenden Diagonalelemente l_{ii} von L sind ungleich Null. Aus diesem Algorithmus folgt unmittelbar

Korollar 5.5. *Die Cholesky-Zerlegung einer hermiteschen und positiv definiten Matrix A ist eindeutig bestimmt.*

Aufwand. Die Berechnung von l_{ij} (mit $i \geq j$) erfordert insgesamt j Multiplikationen, Divisionen oder Wurzeln. Demnach ergibt sich ein Gesamtaufwand von

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

Die Berechnung der Cholesky-Zerlegung ist somit um den Faktor Zwei billiger als die LR-Zerlegung. ◊

Beispiel. Gegeben sei

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

Dann ergeben sich die Einträge von L wie oben skizziert:

$$\begin{aligned} l_{11} &= \sqrt{1} = 1\,, & l_{31} &= 1/1 = 1\,, \\ l_{21} &= 2/1 = 2\,, & l_{32} &= (2-2)/1 = 0\,, \\ l_{22} &= \sqrt{5-4} = 1\,, & l_{33} &= \sqrt{10-1} = 3\,. \end{aligned}$$

Also ist

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

◊

Die Cholesky-Zerlegung kann natürlich für die gleichen Zwecke wie die LR-Zerlegung eingesetzt werden mit dem Vorteil, daß sie nur etwa halb so viel kostet wie die LR-Zerlegung. Bedeutsamer ist allerdings die Tatsache, daß LL^* immer hermitesch und positiv definit ist, obwohl das berechnete L in der Praxis aufgrund von Rundungsfehlern nur eine Näherung an den exakten Cholesky-Faktor ist. Wird hingegen die LR-Faktorisierung einer hermiteschen, positiv definiten Matrix A berechnet, dann ist aufgrund der Rundungsfehler *nicht* gewährleistet, daß das (exakte) Produkt LR hermitesch und positiv definit ist.

6 Toeplitz-Systeme

Wir beginnen diesen Abschnitt mit einem Beispiel aus der Signalverarbeitung.

Beispiel 6.1. Eine Antenne empfängt in regelmäßigen Abständen störanfällige Signalwerte $y_i \in \mathbb{C}$, $i \in \mathbb{Z}$. In vielen Anwendungen enthalten aufeinanderfolgende Signalwerte Redundanzen und sind deshalb nicht völlig unkorreliert. Diese Korrelationen können oftmals mittels eines sogenannten *endlichen linearen Filters* ausgedrückt werden,

$$y_i \approx \tilde{y}_i = \sum_{k=1}^{n} \xi_k y_{i-k} \,. \tag{6.1}$$

Falls die Koeffizienten $\xi_k \in \mathbb{C}$ auf der rechten Seite bekannt sind, können so fehlerbehaftete Signalwerte mittels (6.1) rekonstruiert werden. Leider sind jedoch diese Koeffizienten in der Regel unbekannt.

Einen Ansatz zur Berechnung geeigneter Koeffizienten ξ_k findet sich in der stochastischen Literatur: Hierzu werden die Signalwerte y_i durch Zufallsvariablen modelliert und der lineare Filter wird so bestimmt, daß das Fehlerfunktional

$$\Phi(\xi_1, \ldots, \xi_n) = \mathcal{E}\big|\tilde{y}_i - y_i\big|^2$$

minimiert wird, wobei $\mathcal{E}$ der Erwartungswert und $\tilde{y}_i$ durch (6.1) definiert ist. Der resultierende Filter heißt *Wiener-Filter*. Durch Ausquadrieren ergibt sich

$$\begin{aligned}\Phi(\xi_1, \ldots, \xi_n) \;&=\; \mathcal{E}\,\Big|\sum_{k=1}^{n} \xi_k y_{i-k} - y_i\Big|^2 \\ &=\; \mathcal{E}|y_i|^2 - \sum_{k=1}^{n} \overline{\xi_k}\mathcal{E}(\overline{y_{i-k}} y_i) - \sum_{k=1}^{n} \xi_k \mathcal{E}(\overline{y_i} y_{i-k}) + \sum_{j,k=1}^{n} \overline{\xi_j}\xi_k \mathcal{E}(\overline{y_{i-j}} y_{i-k})\,,\end{aligned}$$

wobei lediglich vorausgesetzt werden muß, daß die Varianz $\mathcal{E}|y_i|^2$ aller Meßwerte endlich ist.

Mit den folgenden Vektoren $x, b \in \mathbb{C}^n$ und der Matrix $T \in \mathbb{C}^{n\times n}$,

$$x = \big[\,\xi_k\,\big]_{k=1}^{n}\,, \quad b = \big[\,\mathcal{E}(\overline{y_i} y_{i-k})\,\big]_{k=1}^{n}\,, \quad T = \big[\,\mathcal{E}(\overline{y_{i-j}} y_{i-k})\,\big]_{j,k=1}^{n}\,,$$

kann das Fehlerfunktional auch folgendermaßen geschrieben werden:

$$\Phi(x) = \mathcal{E}|y_i|^2 - x^*b - b^*x + x^*Tx = \mathcal{E}|y_i|^2 - 2\,\mathrm{Re}\,x^*b + x^*Tx\,. \tag{6.2}$$

Man beachte, daß die sogenannte *Kovarianzmatrix* T hermitesch und (zumindest) positiv semidefinit ist, denn

$$x^*Tx \;=\; \sum_{j,k=1}^{n} \overline{\xi_j}\xi_k\, \mathcal{E}(\overline{y_{i-j}}y_{i-k}) \;=\; \mathcal{E}\,\Big|\sum_{k=1}^{n} \xi_k y_{i-k}\Big|^2 \;\geq\; 0\,.$$

Wir werden im weiteren darüber hinaus annehmen, daß die untere Schranke Null nur für $\xi_k = 0$, $k = 1, \ldots, n$, also nur für $x = 0$ angenommen werden kann. In diesem Fall ist T positiv definit.

Setzen wir $\widehat{x} = T^{-1}b$, dann ergibt eine einfache Rechnung, daß

$$\begin{aligned}\Phi(x) - \Phi(\widehat{x}) \;&=\; x^*Tx - 2\,\mathrm{Re}\,x^*b \;-\; \widehat{x}^*T\widehat{x} + 2\,\mathrm{Re}\,\widehat{x}^*b \\ &=\; (x-\widehat{x})^*T(x-\widehat{x}) + 2\,\mathrm{Re}\,x^*T\widehat{x} - 2\widehat{x}^*T\widehat{x} + 2\,\mathrm{Re}(\widehat{x}^*b - x^*b) \\ &=\; (x-\widehat{x})^*T(x-\widehat{x})\,.\end{aligned}$$

Da T positiv definit ist, ist der letzte Ausdruck nichtnegativ und genau dann Null, wenn $x = \widehat{x}$ ist. Mit anderen Worten: Die eindeutige Lösung des linearen Gleichungssystems

$$Tx = b \tag{6.3}$$

ist das globale Minimum des Funktionals (6.2).

In der Signalverarbeitung spielen Signale aus sogenannten *stationären stochastischen Prozessen* eine besondere Rolle, für die die Korrelation zwischen den Signalwerten y_i und y_{i-k} nicht vom aktuellen Zeitpunkt (also dem Index i) abhängt. Mit anderen Worten,

$$\mathcal{E}(\overline{y_{i-j}}y_{i-k}) \;=\; t_{k-j} \qquad \text{für alle } i, j, k \in \mathbb{Z}$$

und gewisse $t_k \in \mathbb{C}$, $k \in \mathbb{Z}$. Für diese stationären Prozesse hat die Kovarianzmatrix T die spezielle Form

$$T = \begin{bmatrix} t_0 & t_1 & \cdots & t_{n-1} \\ t_{-1} & t_0 & \ddots & \vdots \\ \vdots & \ddots & \ddots & t_1 \\ t_{1-n} & \cdots & t_{-1} & t_0 \end{bmatrix}. \tag{6.4}$$

Eine Matrix dieser Form heißt *Toeplitz-Matrix*. Die rechte Seite des linearen Gleichungssystems (6.3) hat die spezielle Form $b = [t_1, \ldots, t_n]^T$; man nennt dies die *Yule-Walker-Gleichung*. ◊

Gleichungssysteme $Tx = b$ mit Toeplitz-Matrizen treten in verschiedenen Anwendungen auf, nicht nur in der Signalverarbeitung. Sie können zudem besonders effizient (mit nur $2n^2$ Multiplikationen) gelöst werden. Bevor wir uns im folgenden derartigen Lösungsverfahren zuwenden, untersuchen wir zunächst die Toeplitz-Struktur etwas genauer.

Sei $A \in \mathbb{K}^{n\times n}$ eine beliebige Matrix und E die Permutationsmatrix

$$E = \begin{bmatrix} 0 & & 1 \\ & \iddots & \\ 1 & & 0 \end{bmatrix}. \tag{6.5}$$

Bei der Transformation $A \mapsto EA$ wird die Reihenfolge der Zeilen von A und bei der Transformation $A \mapsto AE$ die Reihenfolge der Spalten von A umgekehrt. Folglich wird bei der Abbildung $A \mapsto EA^TE$ das (i,j)-Element von A auf die Position $(n+1-j, n+1-i)$ verschoben, und zwar für jedes $1 \le i,j \le n$; das folgende Schema soll diese Permutation des (i,j)-Elements verdeutlichen:

$$(i,j) \overset{A^T}{\rightsquigarrow} (j,i) \overset{A^TE}{\rightsquigarrow} (j, n+1-i) \overset{E(A^TE)}{\rightsquigarrow} (n+1-j, n+1-i)\,.$$

Diese Transformation entspricht einer Spiegelung an der *Antidiagonalen*. Für $A = I$ ergibt sich somit wieder I, denn $EA^TE = E^2 = I$. Matrizen wie die Einheitsmatrix, die invariant unter einer solchen Spiegelung sind, nennt man *persymmetrisch*.

Proposition 6.2. *Toeplitz-Matrizen $T \in \mathbb{K}^{n\times n}$ sind persymmetrisch, d. h. es gilt $T = ET^TE$.*

Beweis. Der Beweis ist unmittelbar klar, denn sowohl das (i,j)-Element als auch das $(n+1-j, n+1-i)$-Element von T liegen beide auf der $(i-j)$-ten Nebendiagonalen und sind daher aufgrund der Definition einer Toeplitz-Matrix gleich. □

Wir beschränken uns im weiteren auf den Fall, daß T reell, symmetrisch und positiv definit ist. Ferner sei das Gleichungssystem (6.3) so skaliert, daß $t_0 = 1$ ist; da T positiv definit sein soll, kann dies immer erzwungen werden. Im folgenden schreiben wir T_n für T und bezeichnen mit T_k, $1 \le k < n$, die Hauptuntermatrizen von T. Nach Lemma 5.1 ist T_k für jedes k symmetrisch und positiv definit. Ferner sei

$$b^{(k)} = \begin{bmatrix} b_1 \\ \vdots \\ b_k \end{bmatrix}, \quad t^{(k)} = \begin{bmatrix} t_1 \\ \vdots \\ t_k \end{bmatrix}, \qquad 1 \le k \le n\,,$$

und $x^{(k)}$ die (eindeutig bestimmte) Lösung von

$$T_k x^{(k)} = b^{(k)}\,. \tag{6.6}$$

Schließlich führen wir noch die Notation

$$\overleftarrow{t^{(k)}} \;=\; E t^{(k)} \;=\; \begin{bmatrix} t_k \\ \vdots \\ t_1 \end{bmatrix} \tag{6.7}$$

für die Spiegelung von $t^{(k)}$ ein.

6.1 Der Levinson-Algorithmus

Wir beschreiben nun den rekursiven *Levinson-Algorithmus* zur Lösung des Toeplitz-Systems $Tx = b$. Dazu nehmen wir an, die Lösung $x^{(k)}$ von (6.6) sei für ein $1 \le k < n$ bekannt und bestimmen die Lösung $x^{(k+1)}$ des nächstgrößeren Systems

$$T_{k+1} x^{(k+1)} \;=\; \left[\begin{array}{ccc:c} & T_k & & \begin{matrix} t_k \\ \vdots \\ t_1 \end{matrix} \\ \hdashline t_k & \cdots & t_1 & 1 \end{array}\right] \left[\begin{array}{c} v \\ \hdashline \mu \end{array}\right] \;=\; \left[\begin{array}{c} b^{(k)} \\ \hdashline b_{k+1} \end{array}\right]. \tag{6.8}$$

Aus (6.8) ergibt sich einerseits

$$v \;=\; T_k^{-1}\big(b^{(k)} - \mu\, \overleftarrow{t^{(k)}}\big) \;=\; x^{(k)} - \mu T_k^{-1} \overleftarrow{t^{(k)}} \tag{6.9}$$

und andererseits

$$\begin{aligned} \mu \;&=\; b_{k+1} - \overleftarrow{t^{(k)}}{}^{*} v \;=\; b_{k+1} - \overleftarrow{t^{(k)}}{}^{*}\big(\, x^{(k)} - \mu T_k^{-1} \overleftarrow{t^{(k)}}\,\big) \\ &=\; b_{k+1} - \overleftarrow{t^{(k)}}{}^{*} x^{(k)} + \mu\, \overleftarrow{t^{(k)}}{}^{*} T_k^{-1} \overleftarrow{t^{(k)}}\,. \end{aligned} \tag{6.10}$$

An dieser Stelle führen wir noch die Lösung $y^{(k)}$ des Toeplitz-Gleichungssystems

$$T_k y^{(k)} \;=\; t^{(k)} \tag{6.11}$$

ein; (6.11) ist gerade die Yule-Walker-Gleichung aus Beispiel 6.1. $y^{(k)}$ steht in unmittelbarem Bezug zu (6.9) und (6.10), denn wegen Proposition 6.2 und der Symmetrie von T_k gilt

$$\overleftarrow{y^{(k)}} \;=\; E y^{(k)} \;=\; E T_k^{-1} t^{(k)} \;=\; T_k^{-1} E t^{(k)} \;=\; T_k^{-1} \overleftarrow{t^{(k)}}\,. \tag{6.12}$$

Initialisierung: Die Toeplitz-Matrix T sei so normiert, daß $t_0 = 1$ ist
$x^{(1)} = b_1\,,\ y^{(1)} = t_1\,,\ \sigma_1 = 1 - t_1^2$
`for` $k = 1, \ldots, n-1$ `do`
berechne $x^{(k+1)} = \begin{bmatrix} v \\ \mu \end{bmatrix}$ gemäß (6.13)
berechne $y^{(k+1)} = \begin{bmatrix} z \\ \zeta \end{bmatrix}$ gemäß (6.15)
$\sigma_{k+1} = (1 - \zeta^2)\sigma_k$ % vgl. (6.16)
`end for`
Ergebnis: $x^{(n)} = T^{-1}b$

Algorithmus 6.1: Levinson-Algorithmus

Mit anderen Worten: Die Lösung $x^{(k+1)}$ von (6.8) ergibt sich aus

$$x^{(k+1)} = \begin{bmatrix} v \\ \mu \end{bmatrix} \qquad \text{mit} \qquad \begin{aligned} \mu &= \big(b_{k+1} - \overleftarrow{t^{(k)}}{}^{*} x^{(k)}\big)/\sigma_k\,, \\ v &= x^{(k)} - \mu\, \overleftarrow{y^{(k)}}\,, \end{aligned} \tag{6.13}$$

wobei σ_k das Schur-Komplement von T_k in T_{k+1} ist,

$$\sigma_k = 1 - \overleftarrow{t^{(k)}}{}^{*} T_k^{-1} \overleftarrow{t^{(k)}} = 1 - \overleftarrow{t^{(k)}}{}^{*} \overleftarrow{y^{(k)}} = 1 - t^{(k)*} y^{(k)}\,, \tag{6.14}$$

und somit nach Lemma 5.1 positiv ist.

Um (6.13) implementieren zu können, bedarf es noch der Kenntnis von $y^{(k)}$. Da aber (6.11) ein Gleichungssystem von derselben Form wie (6.6) ist und die Lösung von (6.11) für jedes $k = 1, \ldots, n$ als Zwischenresultat benötigt wird, können wir $y^{(k)}$ genau wie $x^{(k)}$ rekursiv berechnen: In Analogie zu (6.13) ergibt dies

$$y^{(k+1)} = \begin{bmatrix} z \\ \zeta \end{bmatrix} \qquad \text{mit} \qquad \begin{aligned} \zeta &= \big(t_{k+1} - \overleftarrow{t^{(k)}}{}^{*} y^{(k)}\big)/\sigma_k\,, \\ z &= y^{(k)} - \zeta \overleftarrow{y^{(k)}}\,. \end{aligned} \tag{6.15}$$

Die Vektoren $y^{(k)}$ können nun für $k = 1, \ldots, n$ aus (6.15) berechnet werden und ermöglichen dabei gleichzeitig die Berechnung der Vektoren $x^{(k)}$ aus (6.13).

Alternativ zu (6.14) kann σ_{k+1} auch aus der Rekursion

$$\sigma_{k+1} = \big(1 - \zeta^2\big)\sigma_k \tag{6.16}$$

mit ζ aus (6.15) berechnet werden, was aus (6.14) und (6.15) folgt:

$$\begin{aligned}\sigma_{k+1} &= 1 - t^{(k+1)*} y^{(k+1)} = 1 - t^{(k)} z - t_{k+1}\zeta \\ &= 1 - t^{(k)*}\big(y^{(k)} - \zeta \overleftarrow{y^{(k)}}\big) - t_{k+1}\zeta \\ &= \sigma_k - \zeta\big(t_{k+1} - t^{(k)*}\overleftarrow{y^{(k)}}\big) = \sigma_k - \zeta^2\sigma_k\,.\end{aligned}$$

Algorithmus 6.1 verwendet diese etwas billigere Implementierung.

Aufwand. Der k-te Schleifendurchlauf von Algorithmus 6.1 kostet $4k + O(1)$ Multiplikationen. Folglich ergibt sich ein Gesamtaufwand von $2n^2 + O(n)$ Multiplikationen zur Berechnung von $T^{-1}b$. ◊

Beispiel. Zu lösen sei das Toeplitz-System

$$\begin{bmatrix} 2 & 1 & 1/2 \\ 1 & 2 & 1 \\ 1/2 & 1 & 2 \end{bmatrix} x = \begin{bmatrix} 2 \\ -8 \\ -28 \end{bmatrix}.$$

1. Normalisieren des Gleichungssystems durch Division durch 2 ergibt

$$T = \begin{bmatrix} 1 & 1/2 & 1/4 \\ 1/2 & 1 & 1/2 \\ 1/4 & 1/2 & 1 \end{bmatrix}, \qquad b = \begin{bmatrix} 1 \\ -4 \\ -14 \end{bmatrix}.$$

Demnach ist $t_1 = 1/2$ und $t_2 = 1/4$.

2. Initialisierungen: $\quad x^{(1)} = 1\,, \quad y^{(1)} = 1/2\,, \quad \sigma_1 = 3/4\,.$

3. Der Schleifendurchlauf $k = 1$ ergibt zunächst

$$\mu = (-4 - 1/2)\cdot 4/3 = -6\,, \quad v = 1 + 6/2 = 4\,, \quad x^{(2)} = \begin{bmatrix} 4 \\ -6 \end{bmatrix}.$$

An dieser Stelle sollte man überprüfen, daß $T_2 x^{(2)} = [1, -4]^T = b^{(2)}$ erfüllt ist. Weiterhin folgt

$$\zeta = \big(1/4 - (1/2)\cdot(1/2)\big)\cdot 4/3 = 0\,, \qquad z = 1/2\,,$$

also

$$y^{(2)} = \begin{bmatrix} 1/2 \\ 0 \end{bmatrix} \qquad \text{und} \qquad \sigma_2 = 1 \cdot \sigma_1 = 3/4\,. \tag{6.17}$$

Auch hier ergibt die Probe das korrekte Ergebnis $T_2 y^{(2)} = [1/2, 1/4]^T = t^{(2)}$.

4. Der Schleifendurchlauf $k = 2$ führt dann auf die Lösung:

$$\mu = \left(-14 - \begin{bmatrix}1/4\\1/2\end{bmatrix}^* \begin{bmatrix}4\\-6\end{bmatrix}\right) \cdot 4/3 = -12 \cdot 4/3 = -16\,,$$
$$v = \begin{bmatrix}4\\-6\end{bmatrix} - (-16)\begin{bmatrix}0\\1/2\end{bmatrix} = \begin{bmatrix}4\\-6\end{bmatrix} + \begin{bmatrix}0\\8\end{bmatrix} = \begin{bmatrix}4\\2\end{bmatrix}.$$

Also ist $x = x^{(3)} = [4, 2, -16]^T$. ◊

6.2 Der Algorithmus von Trench

Man kann sogar mit nur rund $\frac{3}{2}n^2$ Multiplikationen die gesamte Inverse T^{-1} einer symmetrischen und positiv definiten Toeplitz-Matrix $T \in \mathbb{R}^{n\times n}$ bestimmen. Grundlage für diesen *Algorithmus von Trench* ist Proposition 6.2: Aus ihr folgt, daß

$$T^{-1} = \left(ET^TE\right)^{-1} = ET^{-T}E = ET^{-1}E \tag{6.18}$$

ebenfalls persymmetrisch ist. Im allgemeinen ist T^{-1} jedoch *keine* Toeplitz-Matrix.

Für die Herleitung des Algorithmus von Trench zerlegen wir wieder

$$T = \begin{bmatrix} T_{n-1} & t^{\overleftarrow{(n-1)}} \\ t^{\overleftarrow{(n-1)*}} & 1 \end{bmatrix} \quad \text{und} \quad T^{-1} = T_n^{-1} = \begin{bmatrix} X & x \\ x^* & \xi \end{bmatrix}. \tag{6.19}$$

Aus der hintersten Spalte der Blockidentität $TT^{-1} = I$ ergibt sich

$$T_{n-1}x + \xi t^{\overleftarrow{(n-1)}} = 0\,, \qquad t^{\overleftarrow{(n-1)*}}x + \xi = 1\,. \tag{6.20}$$

Hieraus folgt unmittelbar $x = -\xi T_{n-1}^{-1} t^{\overleftarrow{(n-1)}}$, d. h. x ist durch die Lösung der Yule-Walker-Gleichung gegeben, vgl. (6.12),

$$x = -\xi y^{\overleftarrow{(n-1)}}\,.$$

Eingesetzt in die zweite Gleichung in (6.20) folgt aus (6.14)

$$1 = \xi - \xi\, t^{\overleftarrow{(n-1)*}} y^{\overleftarrow{(n-1)}} = \xi\sigma_{n-1}\,.$$

Also haben wir mit

$$\xi = \frac{1}{\sigma_{n-1}}, \qquad x = -\frac{1}{\sigma_{n-1}} \overleftarrow{y^{(n-1)}} \tag{6.21}$$

die letzte Zeile und Spalte von T_n^{-1} bestimmt. Aufgrund der Persymmetrie von T^{-1} erhalten wir hieraus unmittelbar die erste Zeile und Spalte von X.

Nun betrachten wir die $(n-1)\times(n-1)$-dimensionale linke obere Blockgleichung der Identität $TT^{-1} = I$: Die entsprechende Gleichung lautet

$$T_{n-1}X + \overleftarrow{t^{(n-1)}} x^* = I\,,$$

und somit ist wegen (6.12) und (6.21)

$$X = T_{n-1}^{-1} - T_{n-1}^{-1} \overleftarrow{t^{(n-1)}} x^* = T_{n-1}^{-1} - \overleftarrow{y^{(n-1)}} x^* = T_{n-1}^{-1} + \sigma_{n-1} x x^*\,. \tag{6.22}$$

Ersetzt man auf der rechten Seite von (6.22) T_{n-1}^{-1} durch $ET_{n-1}^{-1}E$ gemäß (6.18), so erhält man

$$X = ET_{n-1}^{-1}E + \sigma_{n-1}\, x x^*\,. \tag{6.23}$$

Andererseits ergibt sich unmittelbar aus (6.22) durch Multiplikation mit E von links und von rechts

$$EXE = ET_{n-1}^{-1}E + \sigma_{n-1}\, Exx^*E = ET_{n-1}^{-1}E + \sigma_{n-1}\, \overleftarrow{x}\,\overleftarrow{x}^*\,;$$

zusammen mit (6.23) folgt somit

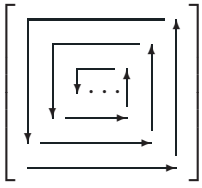
$$X = EXE + \sigma_{n-1}\left(xx^* - \overleftarrow{x}\,\overleftarrow{x}^*\right).$$

Dies bedeutet, daß die Elemente x_{ij} von X über die Einträge von $x = [\xi_i]$ wie folgt gekoppelt sind (man beachte die Dimension $n-1$ von X und x):

$$x_{ij} = x_{n-j,n-i} + \sigma_{n-1}\left(\xi_i\xi_j - \xi_{n-i}\xi_{n-j}\right). \tag{6.24}$$

Daher können nun alle Elemente von T^{-1} in der folgenden „spiralförmigen“ Reihenfolge berechnet werden:

$$T^{-1} = \left[\quad\right].$$



Dabei werden zunächst gemäß (6.19) in die unterste Zeile von T^{-1} der Vektor x^T und das Element ξ aus (6.21) eingetragen; damit ist auch die hinterste Spalte von T^{-1} wegen der Symmetrie der Matrix festgelegt. Aufgrund der Persymmetrie von T^{-1} können danach die erste Zeile und die erste Spalte von T^{-1} ausgefüllt werden. In den inneren Spiralschleifen werden als erstes bei dem Durchlaufen des unteren Pfeils von links nach rechts die Matrixeinträge x_{ij} der Submatrix X von T^{-1} aus der Gleichung (6.24) bestimmt; dazu benötigt man lediglich die bereits berechneten Einträge von T^{-1} aus der unmittelbar zuvor nach unten durchlaufenen Spalte. Der anschließende Pfeil nach oben zeigt die Matrixeinträge an, die danach wieder aufgrund der Symmetrie bekannt sind, und der Rest dieser Spiralschleife (Pfeil von rechts oben nach links unten ums Eck) durchläuft die Einträge, die wegen der Persymmetrie ebenfalls unmittelbar gegeben sind. In dieser Weise fährt man fort, bis alle Matrixelemente berechnet sind.

Beispiel. Am einfachsten macht man sich das Verfahren von Trench wieder an einem Beispiel klar. Für die Matrix aus dem vorigen Beispiel,

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

muß zuerst wieder $T = A/2$ gesetzt werden, damit $t_0 = 1$ ist. In (6.17) haben wir bereits $\sigma_2 = 3/4$ und $y^{(2)} = [1/2, 0]^T$ berechnet; damit sind die unterste Zeile und hinterste Spalte von $T^{-1} = 2A^{-1}$ durch (6.21) gegeben. Der Algorithmus von Trench geht dann folgendermaßen weiter:

$$\begin{aligned} 2A^{-1} \quad &= \quad \begin{bmatrix} x_{11} & x_{12} & 0 \\ x_{21} & x_{22} & -2/3 \\ 0 & -2/3 & 4/3 \end{bmatrix} \overset{\text{persymm.}}{=} \begin{bmatrix} 4/3 & -2/3 & 0 \\ -2/3 & x_{22} & -2/3 \\ 0 & -2/3 & 4/3 \end{bmatrix} \\ &\overset{(6.24)}{=} \quad \begin{bmatrix} 4/3 & -2/3 & 0 \\ -2/3 & 5/3 & -2/3 \\ 0 & -2/3 & 4/3 \end{bmatrix}, \end{aligned}$$

wobei sich im letzten Schritt der verbliebene Matrixeintrag x_{22} wie folgt aus (6.24) berechnet:

$$x_{22} \;=\; x_{11} + \sigma_2(\xi_2^2 - \xi_1^2) \;=\; 4/3 \;+\; 3/4\,\big(\,4/9 - 0\big) \;=\; 5/3\,.$$

Insgesamt ergibt sich also die Inverse

$$A^{-1} = \begin{bmatrix} 2/3 & -1/3 & 0 \\ -1/3 & 5/6 & -1/3 \\ 0 & -1/3 & 2/3 \end{bmatrix}.$$

$\Diamond$

7 Der Banachsche Fixpunktsatz

Wenn die Gleichungssysteme sehr groß sind, verbieten sich Eliminationsverfahren wegen ihres hohen Aufwands. Zudem sind die großen in der Praxis auftretenden Matrizen meist dünn besetzt, d. h. nur wenige Einträge einer Zeile sind ungleich Null. Typische Beispiele für solche Matrizen sind die Steifigkeitsmatrizen aus der Elastizitätstheorie, die wir in Abschnitt 3 kennengelernt haben (vgl. Aufgabe 2). Während die Matrix eines solchen Problems noch gut in den Speicher passen mag, trifft dies für die Faktoren L und R aus der Gauß-Elimination unter Umständen nicht mehr zu, da diese im allgemeinen nicht mehr so dünn besetzt sind (vgl. etwa Abbildung 93.4 zu Beispiel 93.6 für einen konkreten Fall). Unter diesen Umständen behilft man sich gerne mit Iterationsverfahren, die das Gleichungssystem zwar nicht exakt, aber hinreichend genau lösen.

Bevor wir konkrete Verfahren vorstellen können, beweisen wir zunächst ein zentrales Resultat, den Banachschen Fixpunktsatz:

Satz 7.1 (Banachscher Fixpunktsatz). *Sei $\Phi : \mathcal{K} \to \mathcal{K}$ eine (nichtlineare) bezüglich $\|\cdot\|$ kontrahierende Selbstabbildung (eine* Kontraktion*) einer abgeschlossenen Teilmenge $\mathcal{K} \subset \mathbb{K}^n$ mit Kontraktionsfaktor q, d. h.*

$$\|\Phi(x) - \Phi(z)\| \;\le\; q\,\|x - z\| \quad \text{für ein } q < 1 \text{ und alle } x, z \in \mathcal{K}\,. \tag{7.1}$$

Dann hat die Fixpunktgleichung $x = \Phi(x)$ genau eine Lösung $\widehat{x} \in \mathcal{K}$ ($\widehat{x}$ heißt Fixpunkt von Φ), und die Fixpunktiteration *$x^{(k+1)} = \Phi\big(x^{(k)}\big)$, $k = 0, 1, 2, \ldots$, konvergiert für jeden Startvektor $x^{(0)} \in \mathcal{K}$ gegen $\widehat{x}$ für $k \to \infty$. Darüber hinaus ist für $k \ge 1$*

$$(a)\;\; \|x^{(k)} - \widehat{x}\| \;\le\; q\,\|x^{(k-1)} - \widehat{x}\| \qquad (\textit{Monotonie})\,,$$

$$(b)\;\; \|x^{(k)} - \widehat{x}\| \;\le\; \frac{q^k}{1-q}\,\|x^{(1)} - x^{(0)}\| \qquad (\textit{A-priori-Schranke})\,,$$

$$(c)\;\; \|x^{(k)} - \widehat{x}\| \;\le\; \frac{q}{1-q}\|x^{(k)} - x^{(k-1)}\| \qquad (\textit{A-posteriori-Schranke})\,.$$

Beweis. 1. Wir wählen einen beliebigen Startwert $x^{(0)} \in \mathcal{K}$ und betrachten die durch $x^{(k+1)} = \Phi(x^{(k)})$, $k = 0, 1, 2, \ldots$, definierte Iterationsfolge. Aufgrund der Kontraktionseigenschaft von Φ gilt für beliebiges $k \in \mathbb{N}$, daß

$$\|x^{(k+1)} - x^{(k)}\| \;=\; \|\Phi(x^{(k)}) - \Phi(x^{(k-1)})\| \;\le\; q\,\|x^{(k)} - x^{(k-1)}\|\,. \tag{7.2}$$

Damit ergibt sich induktiv

$$\|x^{(k+1)} - x^{(k)}\| \;\le\; q^k\,\|x^{(1)} - x^{(0)}\|\,, \qquad k \in \mathbb{N}. \tag{7.3}$$

Als nächstes wird gezeigt, daß die Folge $\{x^{(k)}\}$ eine Cauchy-Folge ist. Dazu wählen wir $m, l \in \mathbb{N}$ mit $l > m$ und erhalten aus (7.3)

$$\begin{aligned} \|x^{(l)} - x^{(m)}\| &\leq \|x^{(l)} - x^{(l-1)}\| + \ldots + \|x^{(m+1)} - x^{(m)}\| \\ &\leq (q^{l-1} + q^{l-2} + \ldots + q^m)\, \|x^{(1)} - x^{(0)}\| \\ &\leq q^m \frac{1}{1-q} \|x^{(1)} - x^{(0)}\| . \end{aligned} \tag{7.4}$$

Da q^m für $m \to \infty$ gegen Null konvergiert, wird der letzte Ausdruck kleiner als jedes positive ε, wenn nur m hinreichend groß wird. Daher ist $\{x^{(k)}\}$ eine Cauchy-Folge mit Grenzwert x. Da alle Iterierten wegen der Selbstabbildungseigenschaft in der *abgeschlossenen* Menge $\mathcal{K}$ bleiben, gehört auch x zu $\mathcal{K}$.

2. Nun weisen wir nach, daß x ein Fixpunkt von Φ ist. Dazu beachte man zunächst, daß Φ (Lipschitz-)stetig ist. Folglich kann man in der Rekursion $x^{(k+1)} = \Phi(x^{(k)})$ den Grenzübergang $k \to \infty$ betrachten: Während die linke Seite gegen x konvergiert, konvergiert die rechte Seite wegen der Stetigkeit von Φ gegen $\Phi(x)$. Also ist $x = \Phi(x)$ beziehungsweise x ein Fixpunkt von Φ. Damit ist die Existenz eines Fixpunkts nachgewiesen.

3. Die Eindeutigkeit des Fixpunkts folgt aus der Kontraktionseigenschaft: Falls $\widehat{x}$ und x zwei Fixpunkte von Φ in $\mathcal{K}$ sind, ist

$$\|x - \widehat{x}\| \;=\; \|\Phi(x) - \Phi(\widehat{x})\| \;\leq\; q\, \|x - \widehat{x}\|$$

und dies kann wegen $q < 1$ nur gelten, wenn $\|x - \widehat{x}\| = 0$, also $x = \widehat{x}$ ist. Mit anderen Worten: Φ hat in $\mathcal{K}$ nur den einen Fixpunkt $\widehat{x}$ und die Iterationsfolge $\{x^{(k)}\}$ konvergiert für jedes $x^{(0)}$ gegen $\widehat{x}$.

4. Es verbleibt der Nachweis der drei Fehlerabschätzungen. Die erste Ungleichung (a) ergibt sich in ähnlicher Weise wie zuvor die Eindeutigkeit:

$$\|x^{(k)} - \widehat{x}\| \;=\; \|\Phi(x^{(k-1)}) - \Phi(\widehat{x})\| \;\leq\; q\, \|x^{(k-1)} - \widehat{x}\| .$$

Ungleichung (b) folgt leicht aus (7.4): Demnach ist für $m > k$

$$\|x^{(m)} - x^{(k)}\| \;\leq\; q^k \frac{1}{1-q} \|x^{(1)} - x^{(0)}\| ,$$

und die Behauptung ergibt sich durch Grenzübergang $m \to \infty$. Für Ungleichung (c) schätzen wir die linke Seite von (7.2) mit der umgekehrten Dreiecksungleichung und der Monotonieabschätzung (a) nach unten ab:

$$\begin{aligned} \|x^{(k+1)} - x^{(k)}\| &\geq \|x^{(k)} - \widehat{x}\| - \|x^{(k+1)} - \widehat{x}\| \\ &\geq \|x^{(k)} - \widehat{x}\| - q\, \|x^{(k)} - \widehat{x}\| = (1-q)\, \|x^{(k)} - \widehat{x}\| . \end{aligned}$$

Eingesetzt in (7.2) folgt daraus auch die letzte Behauptung dieses Satzes. □

Initialisierung: Sei $A = M - N$ mit invertierbarem M
  wähle beliebiges $x^{(0)} \in \mathbb{K}^n$
  `for` $k = 1, 2, \ldots$ `do`
    löse $Mx^{(k)} = Nx^{(k-1)} + b$
  `until` stop

Algorithmus 7.1: Allgemeine Fixpunktiteration für lineare Gleichungssysteme

Der Banachsche Fixpunktsatz läßt sich zur Konstruktion konvergenter Iterationsverfahren für die numerische Lösung nichtsingulärer linearer Gleichungssysteme $Ax = b$ mit $A \in \mathbb{K}^{n\times n}$ und $b \in \mathbb{K}^n$ verwenden: Hierzu wählt man eine additive Zerlegung von A,

$$A = M - N\,,$$

wobei M invertierbar sein soll und bringt die Gleichung $Ax = b$ auf „Fixpunktgestalt“

$$Mx = Nx + b \qquad \text{bzw.} \qquad x = Tx + c \tag{7.5}$$

mit $T = M^{-1}N$ und $c = M^{-1}b$. Die rechte Seite $Tx + c$ von (7.5) entspricht also der (hier affin linearen) Funktion $\Phi(x)$ aus Satz 7.1.

Es ist offensichtlich, daß ein solches Vorgehen nur dann sinnvoll ist, wenn Gleichungssysteme mit der Matrix M erheblich einfacher zu lösen sind als Gleichungssysteme mit A. Zur Konvergenz des resultierenden Algorithmus 7.1 gibt der Banachsche Fixpunktsatz die folgende Auskunft:

Satz 7.2. *Ist $|||\cdot|||$ eine Norm in $\mathbb{K}^{n\times n}$, die mit einer Vektornorm $\|\cdot\|$ verträglich ist, und ist $|||M^{-1}N||| < 1$, dann konvergiert Algorithmus* 7.1 *für jedes $x^{(0)} \in \mathbb{K}^n$ gegen $A^{-1}b$.*

Beweis. Wir setzen $\Phi(x) = Tx+c$ mit $T = M^{-1}N$ und $c = M^{-1}b$. Mit $\mathcal{K} = \mathbb{K}^n$ ist die Selbstabbildungsvoraussetzung aus Satz 7.1 offensichtlich erfüllt. Ferner ist wegen der Linearität von T

$$\|\Phi(x) - \Phi(z)\| \;=\; \|T(x - z)\| \;\le\; |||T|||\|x - z\|\,,$$

und wegen $|||T||| = |||M^{-1}N||| < 1$ ist Φ somit eine Kontraktion. Nach Satz 7.1 konvergiert die Folge $\{x^{(k)}\}$ aus Algorithmus 7.1 daher gegen den eindeutig bestimmten Fixpunkt $\widehat{x} = T\widehat{x} + c$. Aus (7.5) ist offensichtlich, daß $\widehat{x}$ das lineare Gleichungssystem $Ax = b$ löst. Umgekehrt ist jede Lösung des Gleichungssystems ein Fixpunkt von Φ. Da genau ein Fixpunkt existiert, ergibt sich hieraus die Invertierbarkeit von A. □

Korollar 7.3. *Sei A invertierbar und $T = M^{-1}N$. Dann konvergiert Algorithmus 7.1 genau dann für jedes $x^{(0)} \in \mathbb{K}^n$ gegen $\widehat{x} = A^{-1}b$, wenn für den Spektralradius $\varrho(T)$ von T die Ungleichung $\varrho(T) < 1$ erfüllt ist.*

Beweis. Falls $\varrho(T) < 1$ ist, existiert eine Norm $\|\cdot\|_\varepsilon$ in $\mathbb{K}^n$ und eine dadurch induzierte Norm $|\!|\!| \cdot |\!|\!|_\varepsilon$ in $\mathbb{K}^{n\times n}$ mit $|\!|\!|T|\!|\!|_\varepsilon \le \varrho(T) + \varepsilon < 1$, vgl. Aufgabe I.10. Damit ergibt sich die eine Beweisrichtung aus Satz 7.2.

Ist umgekehrt $\varrho(T) \ge 1$, dann existiert ein Eigenwert λ von A mit $|\lambda| \ge 1$ und zugehörigem Eigenvektor $z \neq 0$. Da $x = \widehat{x}$ ein Fixpunkt von $Tx + c$ ist, ergibt sich für $x^{(0)} = \widehat{x} + z$ und ein festes $k \ge 1$ der Iterationsfehler

$$x^{(k)} - \widehat{x} \;=\; Tx^{(k-1)} + c - \widehat{x} \;=\; Tx^{(k-1)} - T\widehat{x} \;=\; T(x^{(k-1)} - \widehat{x})$$

und durch Induktion folgt

$$x^{(k)} - \widehat{x} \;=\; T^k(x^{(0)} - \widehat{x}) \;=\; T^k z \;=\; \lambda^k z\,. \tag{7.6}$$

Wegen $\|\lambda^k z\| = |\lambda|^k \|z\| \ge \|z\| > 0$ kann $x^{(k)}$ also für $k \to \infty$ nicht gegen $\widehat{x} = A^{-1}b$ konvergieren. □

Dem Spektralradius von $T = M^{-1}N$ kommt also bei der Iteration aus Algorithmus 7.1 eine besondere Bedeutung zu: Gemäß Korollar 7.3 entscheidet $\varrho(T)$ über Konvergenz und Divergenz. Darüber hinaus bestimmt der Spektralradius aber auch noch die asymptotische Konvergenzgeschwindigkeit:

Satz 7.4. *Unter den Voraussetzungen von Korollar 7.3 gilt*

$$\max_{x^{(0)}} \limsup_{k\to\infty} \|\widehat{x} - x^{(k)}\|^{1/k} \;=\; \varrho(T)\,.$$

Beweis. Wie im Beweis von Korollar 7.3 sieht man sofort anhand eines Eigenpaars (λ, z) von T mit $|\lambda| = \varrho(T)$ und $z \neq 0$, daß

$$\max_{x^{(0)}} \limsup_{k\to\infty} \|x^{(k)} - \widehat{x}\|^{1/k} \;\ge\; \limsup_{k\to\infty} \|T^k z\|^{1/k} \;=\; \limsup_{k\to\infty} |\lambda|\,\|z\|^{1/k} \;=\; \varrho(T)\,.$$

Unter Verwendung der Norm $\|\cdot\|_\varepsilon$ und der induzierten (Matrix-)Norm $|\!|\!|\cdot|\!|\!|_\varepsilon$ aus dem Beweis von Korollar 7.3 ergibt sich ferner für jeden Startvektor $x^{(0)}$

$$\|x^{(k)} - \widehat{x}\|_\varepsilon \;=\; \|T^k(x^{(0)} - \widehat{x})\|_\varepsilon \;\le\; |\!|\!|T|\!|\!|_\varepsilon^k \|x^{(0)} - \widehat{x}\|_\varepsilon\,.$$

Wegen der Äquivalenz aller Normen im $\mathbb{K}^n$ existiert daher ein geeignetes $c_\varepsilon > 0$ mit

$$\|x^{(k)} - \widehat{x}\|^{1/k} \;\le\; \big(\,c_\varepsilon \|x^{(k)} - \widehat{x}\|_\varepsilon\,\big)^{1/k} \;\le\; |\!|\!|T|\!|\!|_\varepsilon \big(\,c_\varepsilon \|x^{(0)} - \widehat{x}\|_\varepsilon\,\big)^{1/k},$$

und die rechte Seite konvergiert gegen $\|\!|T|\!\|_\varepsilon$ für $k \to \infty$. Folglich ist

$$\varrho(T) \;\le\; \max_{x^{(0)}} \limsup_{k\to\infty} \|x^{(k)} - \widehat{x}\|^{1/k} \;\le\; \|\!|T|\!\|_\varepsilon \;\le\; \varrho(T) + \varepsilon$$

und da $\varepsilon > 0$ beliebig klein gewählt werden kann, folgt hieraus die Behauptung. □

Definition 7.5. Aufgrund von Satz 7.4 nennt man $\varrho(T)$ auch den (*asymptotischen*) *Konvergenzfaktor* der Iteration $x^{(k)} = Tx^{(k-1)} + c$. Die Zahl $r = -\log_{10} \varrho(T)$ gibt die (*asymptotische*) *Konvergenzrate* an.

Als Faustregel kann man sagen, daß etwa $1/r$ Iterationsschritte für eine zusätzliche signifikante Dezimalstelle des Grenzwerts benötigt werden. Bei dieser Heuristik ist allerdings Vorsicht angebracht, wie das folgende Beispiel zeigt.

Beispiel 7.6. T bezeichne die sogenannte *Shiftmatrix* der Dimension n,

$$T = \begin{bmatrix} 0 & 1 & & 0 \\ & \ddots & \ddots & \\ & & 0 & 1 \\ 0 & & & 0 \end{bmatrix}.$$

Bei einer Matrix-Vektor-Multiplikation Tx werden alle Einträge von x um eine Position nach oben „geshiftet“. Offensichtlich liegt T in Jordan-Normalform vor und man liest unmittelbar den Spektralradius $\varrho(T) = 0$ ab. Entsprechend ist $r = \infty$ und man wird eine extrem schnelle Konvergenz erwarten ($1/r = 0$). Ist jedoch der Ausgangsfehler $x^{(0)} - \widehat{x} = e_n$, dann ergibt sich aus (7.6)

$$x^{(n-1)} - \widehat{x} \;=\; T^{n-1}(x^{(0)} - \widehat{x}) \;=\; T^{n-1}e_n \;=\; e_1\,,$$

das heißt $\|x^{(n-1)} - \widehat{x}\| = \|x^{(0)} - \widehat{x}\| = 1$ sowohl für die Euklidnorm, die Maximumnorm als auch die Betragssummennorm. Mit anderen Worten: In den ersten $n-1$ Iterationsschritten tritt *überhaupt* keine Fehlerreduktion auf. Die Bedeutung der Konvergenzrate ist daher lediglich asymptotischer Natur. ◊

8 Drei einfache Iterationsverfahren

Das einfachste konkrete Beispiel eines Iterationsverfahrens zur Lösung eines Gleichungssystems $Ax = b$ mit $A = [a_{ij}] \in \mathbb{K}^{n\times n}$ und $b = [b_i] \in \mathbb{K}^n$ ist vermutlich das *Gesamtschrittverfahren* (oder *Jacobi-Verfahren*) aus Algorithmus 8.1, bei dem jeweils die i-te Gleichung nach der i-ten Unbekannten aufgelöst wird:

$$x_i^{(k+1)} \;=\; \frac{1}{a_{ii}} \left(b_i - \sum_{j\neq i} a_{ij} x_j^{(k)}\right), \qquad i = 1, \ldots, n\,; \tag{8.1}$$

Initialisierung: Gegeben sei das Gleichungssystem $Ax = b$ mit $a_{ii} \neq 0$, $i = 1, \ldots, n$
wähle beliebiges $x^{(0)} \in \mathbb{K}^n$
`for` $k = 0, 1, \ldots$ `do` % k: Iterationsindex
`for` $i = 1, \ldots, n$ `do`
$$x_i^{(k+1)} = \frac{1}{a_{ii}} \big(b_i - \sum_{j \neq i} a_{ij} x_j^{(k)}\big)$$
`end for`
`until` stop % end for k

Algorithmus 8.1: Gesamtschrittverfahren

dabei bezeichnet k den Iterationsindex. Damit das Verfahren durchführbar ist, müssen alle Diagonaleinträge von A ungleich Null sein.

Aufwand. Es ist einfach zu sehen, daß in jedem Iterationsschritt von Algorithmus 8.1 (d. h. für jedes k) genau eine Multiplikation oder Division mit jedem von Null verschiedenen Eintrag von A nötig ist. ◊

Die Frage nach der Konvergenz von Algorithmus 8.1 werden wir auf Satz 7.2 zurückführen. Dazu zerlegen wir A in

$$A = D - L - R\,, \tag{8.2}$$

wobei D eine Diagonalmatrix, L eine strikte linke untere und R eine strikte rechte obere Dreiecksmatrix ist. Dann können die n Gleichungen (8.1) als Vektorgleichung

$$x^{(k+1)} = D^{-1}\big(b + (L + R)x^{(k)}\big) \tag{8.3}$$

geschrieben werden. Das Gesamtschrittverfahren entspricht also der Fixpunktiteration aus Algorithmus 7.1 mit $M = D$ und $N = L + R$. Die entsprechende Iterationsmatrix $\mathcal{J} = M^{-1}N = D^{-1}(L + R)$ wird *Gesamtschrittoperator* genannt.

Bei dem ganz ähnlichen *Einzelschritt-* oder *Gauß-Seidel-Verfahren* (Algorithmus 8.2) setzt man in (8.1) alle bereits berechneten Komponenten von $x^{(k+1)}$ auf der rechten Seite ein. Der Aufwand ist somit der gleiche wie beim Gesamtschrittverfahren. Dazu der „Originalton“ von Carl Friedrich Gauß[2]:

> *„Ich empfehle Ihnen diesen Modus zur Nachahmung. Schwerlich werden Sie je wieder direct eliminiren, wenigstens nicht, wenn Sie mehr als zwei Unbekannte haben. Das indirecte Verfahren lässt sich halb im Schlafe ausführen, oder man kann während desselben an andere Dinge denken.“*

[2] aus einem Brief an Gerling aus dem Jahr 1823

Initialisierung: Gegeben sei das Gleichungssystem $Ax = b$ mit $a_{ii} \neq 0$, $i = 1, \ldots, n$
wähle beliebiges $x^{(0)} \in \mathbb{K}^n$
`for` $k = 0, 1, \ldots$ `do` % k: Iterationsindex
`for` $i = 1, \ldots, n$ `do`

$$x_i^{(k+1)} = \frac{1}{a_{ii}} \left(b_i - \sum_{j<i} a_{ij} x_j^{(k+1)} - \sum_{j>i} a_{ij} x_j^{(k)} \right)$$

`end for`
`until` stop % end for k

Algorithmus 8.2: Einzelschrittverfahren

Entsprechend zu (8.3) erhält man die Matrixformulierung des Einzelschrittverfahrens, indem man in der Rechenvorschrift für $x_i^{(k+1)}$ in Algorithmus 8.2 alle Komponenten von $x^{(k+1)}$ auf die linke Seite bringt. Demnach ist

$$a_{ii} x_i^{(k+1)} + \sum_{j<i} a_{ij} x_j^{(k+1)} = b_i - \sum_{j>i} a_{ij} x_j^{(k)}, \qquad i = 1, \ldots, n,$$

das heißt $x^{(k+1)}$ ergibt sich durch Auflösen des Dreiecksystems

$$(D - L) x^{(k+1)} = b + R x^{(k)}. \tag{8.4}$$

Wir haben also wieder eine Fixpunktiteration wie in Algorithmus 7.1, diesmal mit $M = D - L$ und $N = R$; $\mathcal{L} = (D - L)^{-1} R$ ist der *Einzelschrittoperator*.

Eine Anwendung der allgemeinen Theorie des vorherigen Abschnitts liefert das folgende Konvergenzkriterium:

Satz 8.1. *Ist A strikt diagonaldominant, dann konvergieren Gesamt- und Einzelschrittverfahren für jeden Startvektor $x^{(0)} \in \mathbb{K}^n$ gegen die eindeutige Lösung von $Ax = b$.*

Beweis. Aufgrund der strikten Diagonaldominanz sind alle Diagonaleinträge von A von Null verschieden und die beiden Iterationsverfahren wohldefiniert. Für den Konvergenzbeweis wollen wir Satz 7.2 anwenden. Zunächst betrachten wir das Gesamtschrittverfahren. Aus der strikten Diagonaldominanz von A folgt unmittelbar

$$\|\mathcal{J}\|_\infty = \|D^{-1}(L + R)\|_\infty = \max_{i=1}^{n} \sum_{\substack{j=1 \\ j \neq i}}^{n} \frac{|a_{ij}|}{|a_{ii}|} =: q < 1,$$

d. h. die Voraussetzung von Satz 7.2 ist für die Zeilensummennorm erfüllt.

Für das Einzelschrittverfahren ist der Beweis etwas komplizierter. Wieder greifen wir auf die Zeilensummennorm zurück und wollen nachweisen, vgl. Aufgabe I.6, daß

$$\|\mathcal{L}\|_\infty \;=\; \max_{\|x\|_\infty=1} \|\mathcal{L}x\|_\infty \;<\; 1\,.$$

Sei also $\|x\|_\infty = 1$ und q wie zuvor definiert. Die einzelnen Komponenten y_i von $y = \mathcal{L}x$ ergeben sich aus Algorithmus 8.2 mit $b = 0$, $x^{(k)} = x$ und $y = x^{(k+1)}$:

$$y_i = \frac{1}{a_{ii}} \big(-\sum_{j<i} a_{ij} y_j - \sum_{j>i} a_{ij} x_j\big)\,. \tag{8.5}$$

Wir zeigen nun induktiv, daß $|y_i| \le q < 1$ für alle $i = 1, \ldots, n$ gilt: Hierzu schätzen wir in (8.5) $|y_i|$ mit der Dreiecksungleichung und der Induktionsannahme wie folgt ab:

$$\begin{aligned} |y_i| \;&\le\; \frac{1}{|a_{ii}|} \big(\sum_{j<i} |a_{ij}||y_j| + \sum_{j>i} |a_{ij}||x_j|\big) \;\le\; \frac{1}{|a_{ii}|} \big(\sum_{j<i} |a_{ij}| q + \sum_{j>i} |a_{ij}|\, \|x\|_\infty\big) \\ &\le\; \frac{1}{|a_{ii}|} \big(\sum_{j<i} |a_{ij}| + \sum_{j>i} |a_{ij}|\big) \;\le\; q\,. \end{aligned}$$

Hieraus folgt $\|y\|_\infty \le q$ und somit ist $\|\mathcal{L}\|_\infty \le q$. □

Beispiele. Gegeben sei das lineare Gleichungssystem $Ax = b$ mit

$$A = \begin{bmatrix} 2 & 0 & 1 \\ 1 & -4 & 1 \\ 0 & -1 & 2 \end{bmatrix}, \quad b = \begin{bmatrix} 1 \\ 4 \\ -1 \end{bmatrix}; \quad \text{die Lösung lautet} \quad \widehat{x} = \begin{bmatrix} 1 \\ -1 \\ -1 \end{bmatrix}.$$

Man beachte, daß A strikt diagonaldominant ist. Aus dem Beweis von Satz 8.1 ergibt sich der Kontraktionsfaktor $q = 1/2$ bezüglich der Zeilensummennorm. Für den Startvektor $x^{(0)} = [1, 1, 1]^T$ ist $\|x^{(0)} - \widehat{x}\|_\infty = 2$, und nach einer Iteration haben wir bei dem

- Gesamtschrittverfahren: $x_{\mathcal{J}}^{(1)} = [0, -1/2, 0]^T$ mit Fehler $\|x_{\mathcal{J}}^{(1)} - \widehat{x}\|_\infty = 1$;
- Einzelschrittverfahren: $x_{\mathcal{L}}^{(1)} = [0, -3/4, -7/8]^T$ mit Fehler $\|x_{\mathcal{L}}^{(1)} - \widehat{x}\|_\infty = 1$.

Bezüglich der Maximumnorm wird der Fehler also tatsächlich in beiden Fällen genau um den Faktor q reduziert. Anhand der einzelnen Komponenten erkennt man aber auch, daß die Iterierte des Einzelschrittverfahrens geringfügig besser ist.

Als zweites Beispiel betrachten wir das 16-dimensionale Gleichungssystem (3.5) aus Abschnitt 3. Für dieses Gleichungssystem divergiert das Jacobi-Verfahren,

denn die Matrix $\mathcal{J}$ besitzt einen Eigenwert $\lambda \approx -1.1082$ außerhalb des Einheitskreises. Das Gauß-Seidel-Verfahren hingegen konvergiert: Alle Eigenwerte des Einzelschrittoperators $\mathcal{L}$ liegen innerhalb des Einheitskreises und der Spektralradius $\varrho(\mathcal{L})$ liegt ungefähr bei 0.9520. Dies deutet jedoch auf eine langsame Konvergenz hin (vgl. Satz 7.4). Tatsächlich benötigt das Verfahren 94 Iterationen, um den relativen Fehler $\|\widehat{x} - x\|_2 / \|\widehat{x}\|_2$ unter 10^{-2} zu drücken. ◊

Obwohl sich Beispiele konstruieren lassen, für die das Gesamtschrittverfahren überlegen ist, konvergiert das Einzelschrittverfahren häufig schneller als das Gesamtschrittverfahren. Für spezielle Matrizen A lassen sich derartige Vergleiche präzisieren. Im folgenden soll ein solcher Vergleich für Matrizen der Form

$$A = \begin{bmatrix} I & -B^* \\ -B & I \end{bmatrix} \in \mathbb{K}^{n\times n} \tag{8.6}$$

mit $B \in \mathbb{K}^{p\times q}$, $0 < p, q < n$, $p + q = n$, exemplarisch vorgeführt werden. Im vorliegenden Fall ist $D = I$ und

$$L = \begin{bmatrix} 0 & 0 \\ B & 0 \end{bmatrix}, \qquad R = \begin{bmatrix} 0 & B^* \\ 0 & 0 \end{bmatrix}.$$

Daher ergeben sich für Einzel- und Gesamtschrittverfahren die Iterationsmatrizen

$$\mathcal{L} = \begin{bmatrix} I & 0 \\ -B & I \end{bmatrix}^{-1} \begin{bmatrix} 0 & B^* \\ 0 & 0 \end{bmatrix} = \begin{bmatrix} I & 0 \\ B & I \end{bmatrix} \begin{bmatrix} 0 & B^* \\ 0 & 0 \end{bmatrix} = \begin{bmatrix} 0 & B^* \\ 0 & BB^* \end{bmatrix} \tag{8.7}$$

sowie

$$\mathcal{J} = \begin{bmatrix} 0 & B^* \\ B & 0 \end{bmatrix} \qquad \text{und daher} \qquad \mathcal{J}^2 = \begin{bmatrix} B^*B & 0 \\ 0 & BB^* \end{bmatrix}. \tag{8.8}$$

Lemma 8.2. *Es sei $X \in \mathbb{K}^{p\times q}$, $Y \in \mathbb{K}^{q\times p}$ und $Z \in \mathbb{K}^{n\times n}$ mit $p, q, n \in \mathbb{N}$. Dann gilt:*

(a) $\sigma(XY) \setminus \{0\} = \sigma(YX) \setminus \{0\}$
(b) $\sigma(Z^2) = \{\lambda^2 : \lambda \in \sigma(Z)\}$.

Beweis. (a): Ist $\lambda \in \sigma(XY) \setminus \{0\}$, dann existiert ein Eigenvektor $u \neq 0$ mit $XYu = \lambda u \neq 0$. Daher ist auch $v = Yu \neq 0$ und es gilt

$$YXv = Y(XYu) = Y(\lambda u) = \lambda Yu = \lambda v;$$

folglich ist $\lambda \in \sigma(YX)$ und $\sigma(XY)\setminus\{0\} \subset \sigma(YX)$. Mit dem gleichen Argument sieht man auch, daß $\sigma(YX) \setminus \{0\} \subset \sigma(XY)$.

(b): Ist $\lambda \in \sigma(Z)$, dann existiert $x \neq 0$ mit $Zx = \lambda x$, und damit ergibt sich $Z^2x = Z(\lambda x) = \lambda\, Zx = \lambda^2 x$. Also ist $\lambda^2 \in \sigma(Z^2)$. Ist umgekehrt $\mu \in \sigma(Z^2)$ und sind $\pm\lambda$ die beiden (ggf. komplexen) Wurzeln von μ, dann gilt

$$0 \;=\; \det(Z^2 - \mu I) \;=\; \det\big((Z - \lambda I)(Z + \lambda I)\big) \;=\; \det(Z - \lambda I)\det(Z + \lambda I)\,.$$

Daher ist entweder λ oder $-\lambda$ im Spektrum von Z, was zu zeigen war. □

Nun können wir Gesamt- und Einzelschrittverfahren für Matrizen der Form (8.6) vergleichen:

Satz 8.3. *Hat A die Gestalt* (8.6)*, dann gilt $\varrho(\mathcal{L}) = \varrho(\mathcal{J})^2$.*

Beweis. Nach (8.7) gilt

$$\begin{aligned}\det(\mathcal{L} - \lambda I) \;&=\; \det\begin{bmatrix} -\lambda I & B^* \\ 0 & BB^* - \lambda I \end{bmatrix} \\ &=\; \det(-\lambda I)\,\det(BB^* - \lambda I) \;=\; (-\lambda)^q\,\det(BB^* - \lambda I)\,.\end{aligned}$$

Also ist $\sigma(\mathcal{L}) = \{0\} \cup \sigma(BB^*)$ und $\varrho(\mathcal{L}) = \varrho(BB^*)$.

Andererseits ist wegen (8.8) und Lemma 8.2 (a)

$$\sigma(\mathcal{J}^2) = \sigma(BB^*) \qquad \text{ggf. zuzüglich des Eigenwerts } 0\,,$$

so daß nach Lemma 8.2 (b) die Gleichungskette $\varrho(\mathcal{J})^2 = \varrho(\mathcal{J}^2) = \varrho(BB^*) = \varrho(\mathcal{L})$ gültig ist. □

Mit anderen Worten (vgl. Korollar 7.3 und Satz 7.4): Für Matrizen der Form (8.6) ist entweder $\varrho(\mathcal{J}) < 1$ und sowohl das Gesamt- als auch das Einzelschrittverfahren konvergieren oder es ist $\varrho(\mathcal{J}) \geq 1$ und beide Verfahren divergieren; im konvergenten Fall braucht das Einzelschrittverfahren für eine vorgegebene Genauigkeit – grob gesprochen – nur halb so viele Iterationen wie das Gesamtschrittverfahren.

Wir erwähnen schließlich noch ein drittes Verfahren, das *symmetrische Gauß-Seidel-Verfahren* für hermitesche Matrizen. Bei dieser Variante des Einzelschrittverfahrens werden die einzelnen Gleichungen des Systems $Ax = b$ zunächst von oben nach unten und dann wieder von unten nach oben durchlaufen. Bei jedem Durchlaufen der i-ten Gleichung wird die Komponente x_i der Näherungslösung entsprechend aktualisiert:

$$\begin{aligned} x_i^{(k+1/2)} \;&=\; \frac{1}{a_{ii}}\,(b_i - \sum_{j<i} a_{ij}x_j^{(k+1/2)} - \sum_{j>i} a_{ij}x_j^{(k)})\,, \qquad i = 1,\ldots,n\,, \\ x_i^{(k+1)} \;&=\; \frac{1}{a_{ii}}\,(b_i - \sum_{j<i} a_{ij}x_j^{(k+1/2)} - \sum_{j>i} a_{ij}x_j^{(k+1)})\,, \qquad i = n,\ldots,1\,. \end{aligned}$$

Da A hermitesch sein soll, ist $L = R^*$, und entsprechend zu (8.4) ergibt sich die Matrixnotation

$$(D - R^*)x^{(k+1/2)} = b + Rx^{(k)}\,, \tag{8.9a}$$

$$(D - R)x^{(k+1)} = b + R^*x^{(k+1/2)} \tag{8.9b}$$

dieses Verfahrens. Durch Multiplikation von (8.9b) mit $(D - R^*)D^{-1}$ ergibt sich

$$\begin{aligned}(D - R^*)D^{-1}(D - R)x^{(k+1)} &= (D - R^*)D^{-1}b + (D - R^*)D^{-1}R^*x^{(k+1/2)}\\ &= (D - R^*)D^{-1}b + (D - R^*)D^{-1}(R^* - D)x^{(k+1/2)} + (D - R^*)x^{(k+1/2)}\\ &\overset{(8.9a)}{=} (D - R^*)D^{-1}b - (D - R^*)D^{-1}b - (D - R^*)D^{-1}Rx^{(k)} + b + Rx^{(k)}\\ &= b + R^*D^{-1}Rx^{(k)}\,.\end{aligned}$$

Mit

$$M = (D - R^*)D^{-1}(D - R) \qquad \text{und} \qquad N = R^*D^{-1}R \tag{8.10}$$

erhalten wir somit aus (8.9) die Iterationsvorschrift

$$Mx^{(k+1)} = Nx^{(k)} + b\,,$$

wobei M und N die Bedingung

$$M - N \;=\; DD^{-1}(D - R) - R^*D^{-1}D \;=\; D - R - R^* \;=\; A$$

erfüllen. Wenn die Diagonaleinträge von A allesamt von Null verschieden sind, sind die Diagonalmatrix D, die Dreiecksmatrizen $D - R^*$ und $D - R$ und somit auch M invertierbar. Das symmetrische Gauß-Seidel-Verfahren paßt also ebenfalls in das allgemeine Schema des Algorithmus 7.1.

Aufwand. Niethammer [76] bemerkte, daß das symmetrische Gauß-Seidel-Verfahren so implementiert werden kann, daß der Aufwand der gleiche ist wie für das Gesamtschrittverfahren und das Einzelschrittverfahren und im wesentlichen pro Iteration einer Matrix-Vektor-Multiplikation mit der Matrix A entspricht. Hierfür müssen lediglich die Vektoren $R^*x^{(k+1/2)}$ und $Rx^{(k+1)}$ zwischengespeichert werden, da sie in dem jeweils darauffolgenden Iterationshalbschritt von (8.9) benötigt werden, vgl. Algorithmus 8.3. ◊

Satz 8.4. *Sei $A = D - R - R^* \in \mathbb{K}^{n\times n}$ hermitesch und positiv definit. Dann konvergiert das symmetrische Gauß-Seidel-Verfahren* (8.9).

Initialisierung: Gegeben sei das Gleichungssystem $Ax = b$ mit $a_{ii} \neq 0$, $i = 1, \ldots, n$

wähle beliebiges $x^{(0)} \in \mathbb{K}^n$
`for` $i = 1, \ldots, n$ `do`
$v_i = \sum_{j>i} a_{ij} x_j^{(0)}$
`end for`
% nun gilt $v = -Rx^{(0)}$
`for` $k = 0, 1, \ldots$ `do` % k: Iterationsindex
`for` $i = 1, 2, \ldots, n$ `do`
$w_i = \sum_{j<i} a_{ij} x_j^{(k+1/2)}$
$x_i^{(k+1/2)} = (b_i - v_i - w_i)/a_{ii}$
`end for`
% nun gilt $w = -R^* x^{(k+1/2)}$
`for` $i = n, n-1, \ldots, 1$ `do`
$v_i = \sum_{j>i} a_{ij} x_j^{(k+1)}$
$x_i^{(k+1)} = (b_i - v_i - w_i)/a_{ii}$
`end for`
% nun gilt $v = -Rx^{(k+1)}$
`until` stop % end for k

Algorithmus 8.3: Symmetrisches Gauß-Seidel-Verfahren

Beweis. Da A positiv definit sein soll, sind alle Diagonaleinträge von A positiv und das symmetrische Gauß-Seidel-Verfahren ist wohldefiniert. Der Iterationsoperator des symmetrischen Gauß-Seidel-Verfahrens ist durch $\mathcal{S} = M^{-1}N$ mit M und N aus (8.10) gegeben. Wir definieren nun die Diagonalmatrix $D^{-1/2}$, deren Diagonaleinträge gerade die Quadratwurzeln der entsprechenden Einträge von D^{-1} sind, das heißt es gilt

$$D^{-1/2} D^{-1/2} = D^{-1} \qquad \text{und} \qquad D^{-1/2^*} = D^{-1/2}\,.$$

Aus (8.10) folgt hiermit die Cholesky-Zerlegung von M:

$$M = U^* U \qquad \text{mit} \quad U = D^{-1/2}(D - R)\,. \tag{8.11}$$

Wir betrachten nun die Matrix

$$U\mathcal{S}U^{-1} \;=\; UM^{-1}NU^{-1} \;=\; UU^{-1}U^{-*}NU^{-1} \;=\; U^{-*}NU^{-1}\,, \tag{8.12}$$

die sich durch eine Ähnlichkeitstransformation von $\mathcal{S}$ ergibt und somit die gleichen Eigenwerte wie $\mathcal{S}$ besitzt. Wegen (8.12) ist $U\mathcal{S}U^{-1}$ hermitesch und

zudem positiv semidefinit, da nach (8.12) und (8.10)

$$x^* U \mathcal{S} U^{-1} x \;=\; x^* U^{-*} R^* D^{-1} R U^{-1} x \;=\; \|D^{-1/2} R U^{-1} x\|_2^2 \;\geq\; 0$$

für jedes $x \in \mathbb{K}^n$. Sei nun x mit $\|x\|_2 = 1$ ein Eigenvektor von $U\mathcal{S}U^{-1}$ zum Eigenwert λ. Dann ist $\lambda \geq 0$ und

$$\begin{aligned} \lambda \;&=\; x^* U \mathcal{S} U^{-1} x \;=\; x^* U^{-*} N U^{-1} x \;=\; x^* U^{-*} (M - A) U^{-1} x \\ &=\; 1 - (U^{-1}x)^* A U^{-1} x \,, \end{aligned}$$

und da A positiv definit ist, muß λ echt kleiner als Eins sein. Somit ist

$$\sigma(\mathcal{S}) \;=\; \sigma(U\mathcal{S}U^{-1}) \;\subset\; [0,1) \,,$$

und die Behauptung folgt aus Korollar 7.3. □

Beispiel. Da die Steifigkeitsmatrix im Gleichungssystem (3.5) symmetrisch und positiv definit ist (vgl. Aufgabe 1), kann das symmetrische Gauß-Seidel-Verfahren zur Lösung dieses Gleichungssystems angewendet werden. Der Spektralradius $\varrho(\mathcal{S}) \approx 0.9533$ der Iterationsmatrix ist geringfügig schlechter als der des Einzelschrittoperators; für zwei Dezimalstellen Genauigkeit braucht das symmetrische Gauß-Seidel-Verfahren entsprechend etwa zwei Iterationen mehr. ◊

9 Das Verfahren der konjugierten Gradienten

Zum Abschluß dieses Kapitels behandeln wir das vermutlich effizienteste Iterationsverfahren für lineare Gleichungssysteme $Ax = b$, deren Koeffizientenmatrix $A \in \mathbb{K}^{n\times n}$ hermitesch und positiv definit ist; der Einfachheit halber beschränken wir uns dabei auf reelle (symmetrische) Matrizen.

Das besagte Verfahren läßt sich nicht in das allgemeine Schema aus Abschnitt 7 einordnen. Statt dessen betrachten wir das quadratische Funktional

$$\Phi(x) \;=\; \frac{1}{2}\, x^* A x \;-\; x^* b \,, \qquad x \in \mathbb{R}^n \,.$$

Analog zu Beispiel 6.1 in Abschnitt 6 folgt für $\widehat{x} = A^{-1}b$, daß

$$\begin{aligned} \Phi(x) - \Phi(\widehat{x}) \;&=\; \frac{1}{2}\, x^* A x - x^* b \;-\; \frac{1}{2}\, \widehat{x}^* A \widehat{x} + \widehat{x}^* b \\ &=\; \frac{1}{2}\, (x - \widehat{x})^* A (x - \widehat{x}) + x^* A \widehat{x} - \widehat{x}^* A \widehat{x} - x^* b + \widehat{x}^* b \\ &=\; \frac{1}{2}\, (x - \widehat{x})^* A (x - \widehat{x}) \;\geq\; 0 \,. \end{aligned} \tag{9.1}$$

Da A positiv definit sein soll, hat das Funktional Φ ein eindeutiges Minimum an der Stelle $x = \widehat{x}$.

Definition 9.1. Ist $A \in \mathbb{R}^{n\times n}$ hermitesch und positiv definit, dann wird durch

$$\|x\|_A = \sqrt{x^* A x}\,, \qquad x \in \mathbb{R}^n\,,$$

eine Norm in $\mathbb{R}^n$ definiert, die sogenannte *Energienorm*. Zu der Energienorm gehört ein *Innenprodukt* (vgl. Abschnitt 31), nämlich

$$\langle x, y\rangle_A \;=\; x^* A y\,, \qquad x, y \in \mathbb{R}^n\,.$$

Beispiel. Den Begriff der Energienorm und die Bedeutung des Funktionals Φ erläutern wir anhand des Mechanikbeispiels aus Abschnitt 3. Um einen Stab der Länge l aus diesem Tragwerk auf die Länge $l+d$ zu strecken (bzw. zu stauchen, falls $d < 0$ ist), wird Energie benötigt: Anteilig muß für die infinitesimale Dehnung des Stabs von der Länge $l+s$ auf $l+s+ds$ die Arbeit (definiert als Kraft $\cdot$ Weg)

$$dW \;=\; f\,ds$$

geleistet werden. Hierbei ist $f = \eta s/l$ durch das Hookesche Gesetz (3.2) gegeben. Integration über s ergibt dann die Arbeit für die vollständige Längenänderung:

$$W \;=\; \int_0^d dW \;=\; \int_0^d f\,ds \;=\; \int_0^d \eta\,\frac{s}{l}\,ds \;=\; \frac{\eta}{2l}\,s^2\Big|_0^d \;=\; \frac{\eta}{2l}\,d^2\,.$$

Durch Summation über alle Stäbe des Tragwerks erhalten wir aus (3.4) mit der Notation aus Abschnitt 3 die Gesamtarbeit

$$P_1 \;=\; \frac{\eta}{2}\sum_{k=1}^{18} l_k^{-1} d_k^2 \;=\; \frac{\eta}{2}\,d^* L^{-1} d \;=\; \frac{\eta}{2}\,x^* E L^{-1} E^* x \;=\; \frac{1}{2}\,x^* A x\,,$$

die für eine Verschiebung x des Tragwerks benötigt wird. Diese Arbeit ist danach als *potentielle Energie* in dem deformierten Tragwerk gespeichert.

In Gegenwart äußerer Kräfte ist ein weiterer Term in die Energiebilanz aufzunehmen: Wirkt nämlich auf das Gelenk z_i die äußere Kraft p_i, so wird bei der Verschiebung des Gelenks um ein infinitesimales Wegstück dx_i die Energie $p_i^* dx_i$ frei, beziehungsweise ist die Arbeit $-p_i^* dx_i$ aufzuwenden. Integration und anschließende Summation über alle Gelenke führen auf die Arbeit

$$P_2 \;=\; -\sum_{i=1}^{8} p_i^* x_i \;=\; -p^* x\,.$$

Hieraus resultiert insgesamt der Zuwachs

$$P(x) \;=\; P_1 + P_2 \;=\; \frac{1}{2}\, x^* A x - p^* x$$

der potentiellen Energie des Tragwerks bei einer Verschiebung x. Dies entspricht gerade dem Wert des Funktionals Φ.

Wie wir eingangs gesehen haben, minimiert die Lösung $\widehat{x}$ des Gleichungssystems $Ax = p$ das Funktional Φ. Nach Abschnitt 3 ist $\widehat{x}$ der Verschiebungsvektor des Tragwerks, der sich unter dem Einfluß der äußeren Kraft p einstellt. Physikalisch bedeutet dies nichts anderes, als daß diejenige Verschiebung resultiert, die die totale potentielle Energie des Tragwerks minimiert. Der Minimalwert dieser Energie ist übrigens

$$P(\widehat{x}) \;=\; \frac{1}{2}\, \widehat{x}^* A \widehat{x} - p^* \widehat{x} \;=\; \frac{1}{2}\, \widehat{x}^* A \widehat{x} - (A\widehat{x})^* \widehat{x} \;=\; \frac{1}{2}\, \widehat{x}^* A \widehat{x} - \widehat{x}^* A \widehat{x} \;=\; -\frac{1}{2}\, \widehat{x}^* A \widehat{x}\,,$$

also negativ. Aus diesem Grund verändert das Gerüst bei Einfluß der Schwerkraft „freiwillig“ seine Form. ◊

Aufgrund unserer Überlegungen ist die Abweichung (9.1) des Funktionals Φ von seinem Minimum,

$$\Phi(x) - \Phi(\widehat{x}) \;=\; \frac{1}{2}\, (x - \widehat{x})^* A (x - \widehat{x}) \;=\; \frac{1}{2}\, \|x - \widehat{x}\|_A^2\,, \tag{9.2}$$

ein gut geeignetes Fehlermaß für den Abstand zwischen x und $\widehat{x}$. Geometrisch bedeutet (9.2), daß der Graph der Funktion Φ bezüglich der Energienorm ein kreisförmiges Paraboloid ist, dessen Mittelpunkt über $\widehat{x}$ liegt.

Es liegt nun nahe, iterative Verfahren zur Approximation von $\widehat{x}$ so zu konstruieren, daß das Funktional Φ sukzessive minimiert wird. Konkret gehen wir dabei folgendermaßen vor: Zu der aktuellen Iterierten $x^{(k)}$ bestimmen wir eine „Suchrichtung“ $d^{(k)} \neq 0$ und wählen im nächsten Schritt die neue Iterierte $x^{(k+1)}$ über den Ansatz

$$x^{(k+1)} = x^{(k)} + \alpha d^{(k)}\,. \tag{9.3}$$

In Abhängigkeit von α nimmt dann das Funktional den Wert

$$\Phi(x^{(k)} + \alpha d^{(k)}) \;=\; \Phi(x^{(k)}) + \alpha\, d^{(k)*} A x^{(k)} + \frac{1}{2}\, \alpha^2 d^{(k)*} A d^{(k)} - \alpha\, d^{(k)*} b \tag{9.4}$$

an, und durch Differentiation nach α erhält man die Schrittweite α_k, für die dieser Wert minimal wird, nämlich

$$\alpha_k = \frac{r^{(k)*} d^{(k)}}{d^{(k)*} A d^{(k)}}\,, \qquad r^{(k)} = b - A x^{(k)}\,. \tag{9.5}$$

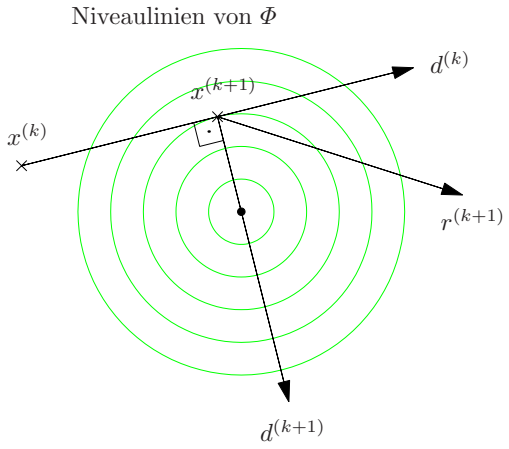


Abb. 9.1:
Skizze in der $\|\cdot\|_A$-Geometrie: $d^{(k+1)}$ ist die optimale Suchrichtung

Dabei ist der Nenner ungleich Null, da A positiv definit sein soll.

Offen in diesem Schema ist noch die Wahl der jeweiligen Suchrichtung. Aus der Darstellung (9.4) von $\Phi(x^{(k)}+\alpha d^{(k)})$ errechnet sich die Richtungsableitung

$$\begin{aligned}\frac{\partial}{\partial d^{(k)}}\Phi(x^{(k)}) &= \operatorname{grad}\Phi(x^{(k)})^* d^{(k)} = \lim_{\alpha\to 0}\frac{\Phi(x^{(k)}+\alpha d^{(k)})-\Phi(x^{(k)})}{\alpha}\\ &= d^{(k)*}(Ax^{(k)}-b)\,.\end{aligned}$$

Demnach ist

$$\operatorname{grad}\Phi(x) = Ax-b$$

der Gradient von Φ im Punkt x. Der negative Gradient, der die Richtung des steilsten Abfalls von Φ angibt, stimmt mit dem Residuum überein. Das Residuum muß jedoch nicht unbedingt die bestmögliche Wahl für die Suchrichtung sein.

Um dies zu erläutern, sei auf Abbildung 9.1 verwiesen, in der die Höhenlinien von Φ in der durch $d^{(k)}$ und $r^{(k+1)}$ aufgespannten Ebene dargestellt sind. Diese Skizze bezieht sich auf die Geometrie, die durch die Energienorm erzeugt wird. In dieser Geometrie sind die Niveauflächen von Φ gerade Kugeloberflächen mit Zentrum $\widehat{x}$, vgl. (9.2). Die Niveaulinien in der dargestellten Ebene sind also konzentrische Kreise und der gemeinsame Mittelpunkt dieser Kreise ist die Minimalstelle von Φ über dieser Ebene. Als neue Suchrichtung bietet sich daher der Vektor $d^{(k+1)}$ an, der in dieser Ebene liegt und (bezüglich des Energie-Innenprodukts) senkrecht zu $d^{(k)}$ ist. Man beachte hierbei, daß die Gerade $x^{(k)}+\alpha d^{(k)}$, $\alpha\in\mathbb{R}$, im Punkt $x^{(k+1)}$ eine Höhenlinie berührt; dies folgt aus der Minimaleigenschaft des entsprechenden Parameters α_k. Die Wahl der nächsten Schrittweite α_{k+1} gemäß (9.5) führt dann automatisch im nächsten Schritt in den Kreismittelpunkt.

Dies ergibt den Ansatz

$$d^{(k+1)} = r^{(k+1)} + \beta_k d^{(k)} \qquad \text{mit} \qquad \langle d^{(k+1)}, d^{(k)} \rangle_A \stackrel{!}{=} 0\,. \tag{9.6}$$

Die resultierende Bedingung für β_k lautet

$$\beta_k = -\frac{r^{(k+1)*} A d^{(k)}}{d^{(k)*} A d^{(k)}}\,. \tag{9.7}$$

Die Definitionen (9.5) und (9.7) sind nur wohldefiniert, wenn $d^{(k)}$ ungleich Null ist. Aus (9.6) sieht man aber, daß $d^{(k)}$ nur Null werden kann, wenn $r^{(k)}$ und $d^{(k-1)}$ linear abhängig sind. Da $d^{(k-1)}$ tangential zur Niveaufläche von Φ, also (bezüglich des euklidischen Innenprodukts) orthogonal zum Gradienten $r^{(k)}$ ist, kann dies nur dann der Fall sein, wenn $r^{(k)} = 0$ ist, also mit $x^{(k)} = \widehat{x}$ die Lösung erreicht wurde. Solange also $x^{(k)} \neq \widehat{x}$ ist, ist der durch die Anweisungen (9.3) bis (9.7) definierte Algorithmus wohldefiniert.

Wegen der speziellen Orthogonalitätsbedingung $\langle d^{(k+1)}, d^{(k)} \rangle_A = 0$ aus (9.6) nennt man die Suchrichtungen auch *zueinander A-konjugiert* und spricht aus diesem Grund vom *Verfahren der konjugierten Gradienten* oder kurz *CG-Verfahren* (engl.: ***c****onjugate* ***g****radients*).

Entscheidend an diesem Verfahren ist eine Optimalitätseigenschaft (Satz 9.5), die wir im weiteren herleiten.

Lemma 9.2. *Sei $x^{(0)}$ ein beliebiger Startvektor und $d^{(0)} = r^{(0)} = b - Ax^{(0)}$. Wenn $x^{(k)} \neq \widehat{x}$ für $k = 0, \ldots, m$, dann gilt*

$$\begin{aligned}
&(a) \qquad r^{(m)*} d^{(j)} = 0 \quad \text{für alle } \ 0 \leq j < m\,,\\
&(b) \qquad r^{(m)*} r^{(j)} = 0 \quad \text{für alle } \ 0 \leq j < m\,,\\
&(c) \quad \langle d^{(m)}, d^{(j)} \rangle_A = 0 \quad \text{für alle } \ 0 \leq j < m\,.
\end{aligned}$$

Beweis. Für $k \geq 0$ gilt $Ax^{(k+1)} = Ax^{(k)} + \alpha_k A d^{(k)}$ und somit ist

$$r^{(k+1)} = r^{(k)} - \alpha_k A d^{(k)}\,, \qquad k \geq 0\,. \tag{9.8}$$

Daher bewirkt die Wahl (9.5) für α_k, daß

$$r^{(k+1)*} d^{(k)} \;=\; (r^{(k)} - \alpha_k A d^{(k)})^* d^{(k)} \;=\; r^{(k)*} d^{(k)} - \alpha_k d^{(k)*} A d^{(k)} \;=\; 0\,. \tag{9.9}$$

Nach dieser einleitenden Beobachtung führen wir den Rest des Beweises durch Induktion über m.

<u>$m = 1$:</u>
Setzt man $k = 0$ in (9.9), dann entspricht dies der Behauptung (a) für $m = 1$ und wegen $r^{(0)} = d^{(0)}$ zudem der Behauptung (b). Für $m = 1$ folgt schließlich auch Behauptung (c) aus dem Konstruktionsprinzip (9.6) mit $k = 0$.

$\underline{\overline{m} \to \overline{m}+1:}$
Im Induktionsschritt nehmen wir an, daß *alle drei* Aussagen in der genannten Form für alle $m \leq \overline{m}$ richtig sind und beweisen nun ihre Gültigkeit für $m = \overline{m}+1$. Dann folgt zunächst $r^{(\overline{m}+1)*}d^{(\overline{m})} = 0$ aus (9.9). Außerdem ergibt (9.8) zusammen mit den beiden Induktionsannahmen (a) und (c)

$$r^{(\overline{m}+1)*}d^{(j)} \;=\; r^{(\overline{m})*}d^{(j)} - \alpha_{\overline{m}}\,\langle d^{(\overline{m})}, d^{(j)}\rangle_A \;=\; 0\,, \qquad 0 \leq j < \overline{m}\,.$$

Folglich gilt (a) für $\overline{m}+1$ anstelle von m.

Wegen (9.6) ist $r^{(j)} = d^{(j)} - \beta_{j-1}d^{(j-1)}$ für $1 \leq j \leq \overline{m}$ und $r^{(0)} = d^{(0)}$; daher folgt Behauptung (b) aus (a).

Die Aussage (c) mit $j = \overline{m}$ und $m = \overline{m}+1$ folgt unmittelbar aus der Konstruktion von $d^{(k+1)}$, vgl. (9.6). Für $j < \overline{m}$ ergibt sich aus (9.6) und der Induktionsannahme die Darstellung

$$\langle d^{(\overline{m}+1)}, d^{(j)}\rangle_A \;=\; \langle r^{(\overline{m}+1)}, d^{(j)}\rangle_A + \beta_{\overline{m}}\,\langle d^{(\overline{m})}, d^{(j)}\rangle_A \;=\; r^{(\overline{m}+1)*}Ad^{(j)}\,.$$

Ersetzt man hier $Ad^{(j)}$ mit Hilfe von (9.8), dann folgt

$$\alpha_j\langle d^{(\overline{m}+1)}, d^{(j)}\rangle_A \;=\; r^{(\overline{m}+1)*}r^{(j)} - r^{(\overline{m}+1)*}r^{(j+1)}\,, \qquad 0 \leq j < \overline{m}\,,$$

und die rechte Seite ist Null aufgrund der bereits bewiesenen Behauptung (b). Damit bleibt für den Nachweis von (c) lediglich noch zu zeigen, daß $\alpha_j \neq 0$ ist. Nehmen wir an, α_j wäre Null: Wegen (9.5) ist dies gleichbedeutend mit $r^{(j)*}d^{(j)} = 0$ und aus (9.6) folgt dann mit der Induktionsannahme

$$0 \;=\; r^{(j)*}(r^{(j)} + \beta_{j-1}d^{(j-1)}) \;=\; r^{(j)*}r^{(j)} + \beta_{j-1}r^{(j)*}d^{(j-1)} \;=\; \|r^{(j)}\|_2^2$$

für $0 < j < \overline{m}$ beziehungsweise

$$0 \;=\; r^{(0)*}d^{(0)} \;=\; r^{(0)*}r^{(0)} \;=\; \|r^{(0)}\|_2^2$$

für $j = 0$. In jedem Fall ergibt dies also $r^{(j)} = 0$ im Widerspruch zu $x^{(j)} \neq \widehat{x}$. Somit ist $\alpha_j \neq 0$ und $\langle d^{(m+1)}, d^{(j)}\rangle_A = 0$ für alle $0 \leq j < \overline{m}+1$. Damit ist der Induktionsschluß vollständig bewiesen. □

Gemäß Lemma 9.2 (c) sind also *alle* Suchrichtungen paarweise A-konjugiert. Ferner sind nach Lemma 9.2 (b) alle Residuen linear unabhängig (alle Orthogonalsysteme sind linear unabhängig) und daher ergibt sich nach spätestens n Schritten $r^{(n)} = 0$, also $x^{(n)} = \widehat{x}$.

Korollar 9.3. *Für $A \in \mathbb{R}^{n\times n}$ hermitesch und positiv definit findet das CG-Verfahren nach höchstens n Schritten die exakte Lösung $x^{(n)} = \widehat{x}$.*

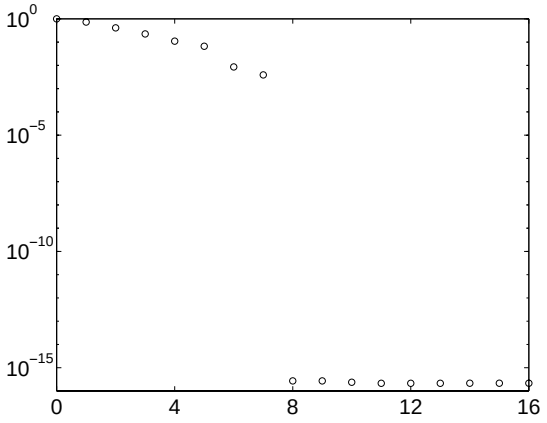


Abb. 9.2: Iterationsfehler des CG-Verfahrens bei dem Mechanikbeispiel

Beispiel. Wendet man das CG-Verfahren auf das Gleichungssystem (3.5) für die Verschiebungen des Brückentragwerks an, so beobachtet man eine sehr schnelle Konvergenz. In Abbildung 9.2 ist der relative Fehler $\|\widehat{x} - x^{(k)}\|_2 / \|\widehat{x}\|_2$ über dem Iterationsindex k aufgetragen. Korollar 9.3 besagt, daß das Verfahren nach maximal 16 Iterationen die exakte Lösung berechnet hat. Tatsächlich ist bereits nach nur acht Iterationen der Verschiebungsvektor $\widehat{x}$ bis auf Maschinengenauigkeit berechnet. Zum Vergleich: Das Gauß-Seidel-Verfahren benötigt 94 Iterationen für einen relativen Fehler von lediglich 10^{-2}, das Jacobi-Verfahren konvergiert gar nicht. ◊

Wie dieses Beispiel zeigt, ist das Ergebnis von Korollar 9.3 in der Praxis nur von eingeschränkter Bedeutung, da das CG-Verfahren in erster Linie iterativ eingesetzt wird und hierbei oftmals nur mit wesentlich weniger als n Iterationsschritten effizient ist. Zudem gehen die Orthogonalitätseigenschaften aus Lemma 9.2 mit zunehmender Iterationsdauer aufgrund von Rundungsfehlern verloren, so daß Korollar 9.3 für die Praxis nicht mehr relevant ist.

Für die Interpretation des CG-Verfahrens als iteratives Verfahren ist die folgende Optimalitätseigenschaft von größerer Bedeutung.

Definition 9.4. Sei $A \in \mathbb{K}^{n\times n}$ und $y \in \mathbb{K}^n$. Dann heißt der Untervektorraum

$$\mathcal{K}_k(A, y) = \operatorname{span}\{y, Ay, \ldots, A^{k-1}y\}$$

Krylov-Raum der Dimension k von A bezüglich y.

Satz 9.5. *Sei $A \in \mathbb{R}^{n\times n}$ hermitesch und positiv definit, $d^{(0)} = r^{(0)}$ und $x^{(k)}$ / $\widehat{x}$ die k-te Iterierte des CG-Verfahrens. Dann gilt*

$$x^{(k)} \in x^{(0)} + \mathcal{K}_k(A, r^{(0)}) \tag{9.10}$$

und $x^{(k)}$ ist in diesem affinen Raum die eindeutige Minimalstelle der Zielfunktion Φ.

Beweis. Wir beweisen zunächst induktiv, daß

$$d^{(j)} \in \operatorname{span}\{r^{(0)}, \ldots, r^{(j)}\}\,, \qquad j = 0, \ldots, k-1\,. \tag{9.11}$$

Für $j = 0$ ist dies offensichtlich erfüllt und der Induktionsschluß ergibt sich sofort aus (9.6). Somit ist $\operatorname{span}\{d^{(0)}, \ldots, d^{(k-1)}\} \subset \operatorname{span}\{r^{(0)}, \ldots, r^{(k-1)}\}$ und solange $x^{(k)} \;/\!\!=\!\!\!\!\widehat{\;\;}\; \widehat{x}$ ist, folgt aus Lemma 9.2, daß beide Systeme $\{d^{(j)}\}_{j=0}^{k-1}$ und $\{r^{(j)}\}_{j=0}^{k-1}$ linear unabhängig sind. Demnach ist

$$\operatorname{span}\{d^{(0)}, \ldots, d^{(k-1)}\} \;=\; \operatorname{span}\{r^{(0)}, \ldots, r^{(k-1)}\}\,. \tag{9.12}$$

Wegen (9.3) gilt ferner

$$x^{(k)} \;=\; x^{(0)} + \sum_{j=0}^{k-1} \alpha_j d^{(j)} \;\in\; x^{(0)} + \operatorname{span}\{r^{(0)}, \ldots, r^{(k-1)}\}\,.$$

Nun zeigen wir induktiv, daß

$$r^{(j)} \in \operatorname{span}\{r^{(0)}, \ldots, A^j r^{(0)}\}\,, \qquad j = 0, \ldots, k-1\,.$$

Für $j = 0$ ist dies sicher richtig. Beim Induktionsschluß „$j-1 \to j$“ gilt zunächst wegen (9.11) und der Induktionsannahme die Beziehung

$$d^{(j-1)} \;\in\; \operatorname{span}\{r^{(0)}, \ldots, r^{(j-1)}\} \;\subset\; \operatorname{span}\{r^{(0)}, \ldots, A^{j-1} r^{(0)}\}\,,$$

und aus (9.8) ergibt sich wiederum mit der Induktionsannahme die gewünschte Inklusion

$$r^{(j)} \;=\; r^{(j-1)} - \alpha_{j-1} A d^{(j-1)} \;\in\; \operatorname{span}\{r^{(0)}, Ar^{(0)}, \ldots, A^j r^{(0)}\}\,.$$

Demnach ist

$$\operatorname{span}\{r^{(0)}, \ldots, r^{(k-1)}\} \;\subset\; \operatorname{span}\{r^{(0)}, Ar^{(0)}, \ldots, A^{k-1} r^{(0)}\}\,,$$

und da die aufspannenden Vektoren $\{r^{(j)}\}_{j=0}^{k-1}$ ein Orthogonalsystem bilden, hat die Menge auf der linken Seite die maximal mögliche Dimension k. Also stimmen die beiden Mengen überein und wegen (9.12) haben wir somit

$$\operatorname{span}\{d^{(0)}, \ldots, d^{(k-1)}\} \;=\; \operatorname{span}\{r^{(0)}, \ldots, r^{(k-1)}\} \;=\; \mathcal{K}_k(A, r^{(0)})\,. \tag{9.13}$$

Aus Korollar 9.3 folgt schließlich die Existenz eines Iterationsindex $m \leq n$, für den

$$\widehat{x} \;=\; x^{(m)} \;=\; x^{(0)} + \sum_{j=0}^{m-1} \alpha_j d^{(j)}$$

gilt (m muß nicht unbedingt mit n übereinstimmen). Demnach ist

$$\widehat{x} - x^{(k)} = \sum_{j=k}^{m-1} \alpha_j d^{(j)}$$

und für ein beliebiges anderes Element $x \in x^{(0)} + \mathcal{K}_k(A, r^{(0)})$ gilt wegen (9.13)

$$\begin{aligned} \widehat{x} - x \;&=\; \widehat{x} - x^{(k)} + x^{(k)} - x \;=\; \widehat{x} - x^{(k)} + \sum_{j=0}^{k-1} \delta_j d^{(j)} \\ &= \sum_{j=0}^{k-1} \delta_j d^{(j)} + \sum_{j=k}^{m-1} \alpha_j d^{(j)} \end{aligned}$$

für gewisse $\delta_j \in \mathbb{R}$. Da die Suchrichtungen nach Lemma 9.2 (c) A-konjugiert sind, folgt daher aus dem Satz von Pythagoras (vgl. ggf. Satz 31.6)

$$\begin{aligned} \Phi(x) - \Phi(\widehat{x}) \;&=\; \frac{1}{2}\,\|x - \widehat{x}\|_A^2 \;=\; \frac{1}{2}\,\|x^{(k)} - \widehat{x}\|_A^2 \;+\; \frac{1}{2}\,\Big\|\sum_{j=0}^{k-1} \delta_j d^{(j)}\Big\|_A^2 \\ &=\; \Phi(x^{(k)}) - \Phi(\widehat{x}) \;+\; \frac{1}{2}\,\Big\|\sum_{j=0}^{k-1} \delta_j d^{(j)}\Big\|_A^2 . \end{aligned}$$

Demnach ist $\Phi(x) \geq \Phi(x^{(k)})$ mit Gleichheit genau für $x = x^{(k)}$. □

Für eine Implementierung des CG-Verfahrens sollte man nicht die oben bestimmten Gleichungen (9.5) und (9.7) für α_k und β_k verwenden, sondern die folgenden Darstellungen (9.5′) und (9.7′), die etwas stabiler sind.

Für die Herleitung dieser alternativen Formeln beachte man, daß aufgrund von Lemma 9.2 (a) und (9.6)

$$r^{(k)*} d^{(k)} = r^{(k)*} r^{(k)} + \beta_{k-1} r^{(k)*} d^{(k-1)} = r^{(k)*} r^{(k)}$$

gilt. In (9.5) eingesetzt, ergibt sich somit

$$\alpha_k = \frac{\|r^{(k)}\|_2^2}{d^{(k)*} A d^{(k)}}\,. \tag{9.5′}$$

Initialisierung: $A \in \mathbb{R}^{n\times n}$ sei hermitesch und positiv definit

wähle beliebiges $x^{(0)} \in \mathbb{R}^n$

$r^{(0)} = b - Ax^{(0)}, \quad d^{(0)} = r^{(0)}$

`for` $k = 0, 1, 2, \ldots$ `do`

$\alpha_k = \|r^{(k)}\|_2^2 \;/\; d^{(k)*}Ad^{(k)}$ % $Ad^{(k)}$ für später abspeichern

$x^{(k+1)} = x^{(k)} + \alpha_k d^{(k)}$

$r^{(k+1)} = r^{(k)} - \alpha_k Ad^{(k)}$

$\beta_k = \|r^{(k+1)}\|_2^2 \;/\; \|r^{(k)}\|_2^2$ % $\|r^{(k+1)}\|_2^2$ für später abspeichern

$d^{(k+1)} = r^{(k+1)} + \beta_k d^{(k)}$

`until` stop % end for

Ergebnis: $x^{(k)}$ ist die Approximation von $A^{-1}b$, $r^{(k)} = b - Ax^{(k)}$ das zugehörige Residuum

Algorithmus 9.1: Verfahren der konjugierten Gradienten (CG-Verfahren)

Entsprechend ist wegen (9.8) und Lemma 9.2 (b) sowie (9.5′)

$$\begin{aligned} r^{(k+1)*}Ad^{(k)} &= \frac{1}{\alpha_k}\left(r^{(k+1)*}r^{(k)} - r^{(k+1)*}r^{(k+1)}\right) = -\frac{1}{\alpha_k}\,\|r^{(k+1)}\|_2^2 \\ &= -\frac{\|r^{(k+1)}\|_2^2}{\|r^{(k)}\|_2^2}\; d^{(k)*}Ad^{(k)}\,; \end{aligned}$$

anstelle von (9.7) verwendet man daher die Formel

$$\beta_k = \frac{\|r^{(k+1)}\|_2^2}{\|r^{(k)}\|_2^2}\,. \tag{9.7′}$$

In Algorithmus 9.1 sind diese Ergebnisse zusammengefaßt.

Aufwand. Abgesehen von den beiden zu berechnenden Innenprodukten mit Aufwand $O(n)$ wird lediglich eine Matrix-Vektor-Multiplikation ($Ad^{(k)}$) in jedem Iterationsschritt benötigt. Unter der Voraussetzung, daß A deutlich mehr als n von Null verschiedene Einträge besitzt, ist der Aufwand des CG-Verfahrens daher im wesentlichen der gleiche wie für Gesamt- und Einzelschrittverfahren. ◊

Bemerkung 9.6. Die Eigenschaften des CG-Verfahrens übertragen sich ausnahmslos auch auf komplexe (hermitesche) Matrizen, also $\mathbb{K} = \mathbb{C}$. In diesem Fall wird das Funktional

$$\Phi(x) \;=\; \frac{1}{2}\,x^*Ax \;-\; \mathrm{Re}\,x^*b \;=\; \frac{1}{2}\,\|x - \widehat{x}\|_A^2 \;-\; \frac{1}{2}\,\widehat{x}^*A\widehat{x} \tag{9.14}$$

für $x \in \mathbb{C}^n$ minimiert (vgl. Beispiel 6.1 für eine entsprechende Vorgehensweise).

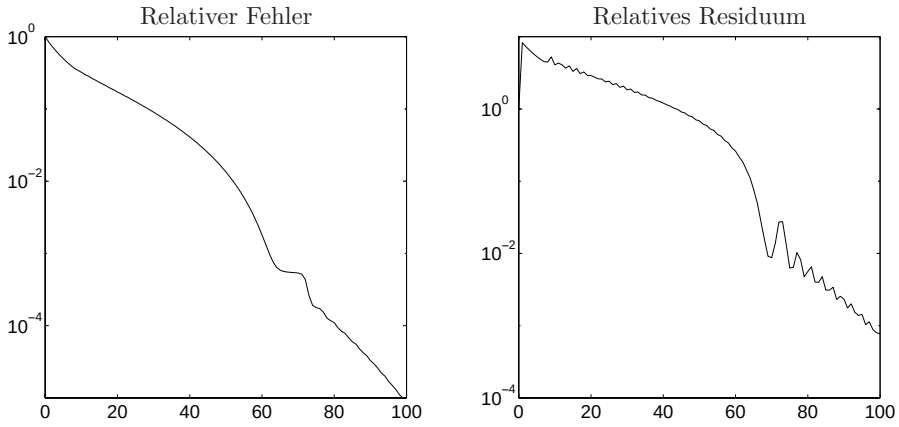


Abb. 9.3: Konvergenzverlauf des CG-Verfahrens

Leider gibt es bislang kein Verfahren für Gleichungssysteme mit nicht hermiteschen Matrizen, das ähnliche Konvergenzeigenschaften wie das CG-Verfahren besitzt und genauso effizient implementiert werden kann. Mindestens eine dieser beiden Eigenschaften geht bei den bislang bekannten Verfahren verloren. Beispielhaft werden wir in Abschnitt 16 das GMRES-Verfahren vorstellen: Das GMRES-Verfahren besitzt ähnliche Konvergenzeigenschaften wie das CG-Verfahren, ist aber wesentlich aufwendiger zu implementieren. ◊

Beispiel 9.7. Anwendungsbeispiele für das CG-Verfahren ergeben sich bei der numerischen Lösung elliptischer partieller Differentialgleichungen, vgl. Kapitel XVI. Als typisches Modellproblem für derartige Gleichungssysteme soll hier das System $Ax = b$ aus (93.11) herangezogen werden, bei dem A die spezielle (Block-)Form

$$A = \begin{bmatrix} C & -I & & \\ -I & C & \ddots & \\ & \ddots & \ddots & -I \\ & & -I & C \end{bmatrix} \quad \text{mit} \quad C = \begin{bmatrix} 4 & -1 & & \\ -1 & 4 & \ddots & \\ & \ddots & \ddots & -1 \\ & & -1 & 4 \end{bmatrix}$$

besitzt. Es sei erwähnt, daß die Komponenten des Lösungsvektors x Näherungen für die Funktionswerte der Lösung $u : [0,1]^2 \to \mathbb{R}$ der Differentialgleichung sind. Für eine gute Approximation der Lösung muß das Problem hinreichend fein diskretisiert werden. Entsprechend groß wird die Matrix A: Im vorliegenden Fall werden 99×99 Funktionswerte von u gesucht, so daß A die Dimension 9801×9801 besitzt.

Abbildung 9.3 zeigt den Konvergenzverlauf des CG-Verfahrens, angewandt auf

ein Gleichungssystem mit dieser Matrix und der vorgegebenen Lösung $\widehat{x}$ mit den entsprechenden Funktionswerten von

$$u(\xi, \eta) = \xi(1-\xi)\eta(1-\eta)\,.$$

Da die Lösung bekannt ist, kann neben der Konvergenz der Euklidnorm des Residuums $b - Ax^{(k)}$ auch die des Fehlers $x^{(k)} - \widehat{x}$ dargestellt werden. Zum Vergleich seien auch die Ergebnisse der drei Verfahren des vorigen Abschnitts nach 100 Iterationen angeführt (Startvektor ist jeweils $x^{(0)} = 0$):

100 Iterationen	rel. Fehler	rel. Residuum
Jacobi	0.9514	0.9207
Gauß-Seidel	0.9053	0.8679
symm. Gauß-Seidel	0.8216	0.7807

Wie man sieht, sind diese Verfahren bedeutend langsamer, wobei das symmetrische Gauß-Seidel-Verfahren noch am besten abschneidet. Für das Jacobi- und das Gauß-Seidel-Verfahren können die asymptotischen Konvergenzfaktoren $q \approx 0.9995$ bzw. $q \approx 0.9990$ in diesem Beispiel explizit angegeben werden. Der Konvergenzfaktor des symmetrischen Gauß-Seidel-Verfahrens ist etwas geringer, nämlich $q \approx 0.9980$. Zum Vergleich: Nach 100 Iterationen des CG-Verfahrens ergibt sich ein „mittlerer Konvergenzfaktor“ $\bar{q} \approx 0.8897$, der durch die Gleichung

$$\|x^{(100)} - \widehat{x}\|_2 \;=\; \bar{q}^{100}\,\|x^{(0)} - \widehat{x}\|_2$$

definiert wird. ◊

10 Präkonditionierung

In Abschnitt 35.3 (Satz 35.7 und der daran anschließende Absatz) werden wir sehen, daß der Iterationsfehler des CG-Verfahrens durch eine obere Schranke $O(q^k)$ mit

$$q \;\approx\; 1 - 2\,\mathrm{cond}_2^{-1/2}(A) \tag{10.1}$$

abgeschätzt werden kann. Für die Konvergenzgeschwindigkeit des CG-Verfahrens ist die Kondition der Koeffizientenmatrix A also ein entscheidender Parameter: In der Regel ist die Konvergenz um so langsamer, je schlechter A konditioniert ist.

Während dies für die Praxis durchaus einen brauchbaren Anhaltspunkt liefert, ist die Abschätzung (10.1) nicht immer scharf: Abbildung 9.3 zu Beispiel 9.7

weist beispielsweise auf einen mittleren Konvergenzfaktor $\bar{q}$ zwischen 0.89 und 0.93 hin, obwohl der Schätzwert (10.1) den deutlich schlechteren Konvergenzfaktor $q \approx 0.9686$ ergibt (mit q anstelle von $\bar{q}$ wären mehr als doppelt so viele Iterationen für die gleiche Genauigkeit nötig). In diesem speziellen Fall hängt die unerwartet schnelle Konvergenz mit der speziellen rechten Seite b des Gleichungssystems zusammen, vgl. Aufgabe 16.

Bei Werten von q nahe bei Eins kann die Konvergenz unter Umständen durch eine sogenannte *Präkonditionierung* beschleunigt werden. Hierzu transformiert man das Gleichungssystem $Ax = b$ zunächst in ein geeignetes äquivalentes System

$$M^{-1}Ax = M^{-1}b\,, \tag{10.2}$$

wobei die Matrix $M \in \mathbb{R}^{n\times n}$ hermitesch und positiv definit sein muß. Im allgemeinen ist zwar $M^{-1}A$ nicht hermitesch, aber mit der entsprechenden Cholesky-Zerlegung $M = LL^*$ ist das Gleichungssystem (10.2) seinerseits zu dem System

$$L^{-1}AL^{-*}z = L^{-1}b\,, \qquad x = L^{-*}z\,, \tag{10.3}$$

äquivalent. Die Koeffizientenmatrix $L^{-1}AL^{-*}$ aus (10.3) ist hermitesch und positiv definit, denn für ein beliebiges $z \in \mathbb{R}^n$ und $x = L^{-*}z$ gilt

$$z^*L^{-1}AL^{-*}z = x^*Ax \geq 0$$

mit Gleichheit genau für $x = z = 0$. Folglich kann das CG-Verfahren zur Lösung des Gleichungssystems (10.3) angewendet werden. Wird die Matrix M so konstruiert, daß die Konditionszahl von $L^{-1}AL^{-*}$ kleiner ist als die Konditionszahl von A, so wird man für die entsprechenden Iterierten $z^{(k)}$ und die resultierenden Näherungen $x^{(k)} = L^{-*}z^{(k)}$ für $A^{-1}b$ eine schnellere Konvergenz erwarten als für die Iterierten des CG-Verfahrens angewandt auf $Ax = b$. Die Matrix M wird daher *Präkonditionierungsmatrix* genannt.

Entscheidend ist, daß die Faktorisierung $M = LL^*$ nicht explizit berechnet werden muß, da die in (10.3) „künstlich eingeführte Variable“ z in der Implementierung wieder durch den zugehörigen Vektor x geeignet substituiert werden kann. Lediglich für die Berechnung der Koeffizienten β_k und die dafür benötigten Normen $\|L^{-1}b - L^{-1}AL^{-*}z^{(k)}\|_2$ wird neben $r^{(k)} = b - Ax^{(k)}$ ein zusätzlicher Hilfsvektor

$$s^{(k)} \;=\; M^{-1}r^{(k)} \;=\; M^{-1}b - M^{-1}Ax^{(k)}$$

benötigt. Dann gilt nämlich

$$\|L^{-1}b - L^{-1}AL^*z^{(k)}\|_2^2 \;=\; \|L^{-1}(b - Ax^{(k)})\|_2^2 \;=\; r^{(k)*}L^{-*}L^{-1}r^{(k)} \;=\; r^{(k)*}s^{(k)}.$$

Initialisierung: A und $M \in \mathbb{R}^{n\times n}$ seien hermitesch und positiv definit

wähle beliebiges $x^{(0)} \in \mathbb{K}^n$
$r^{(0)} = b - Ax^{(0)}$,
löse $Ms^{(0)} = r^{(0)}$
$d^{(0)} = s^{(0)}$
`for` $k = 0, 1, 2, \ldots$ `do`
  $\alpha_k = r^{(k)*}s^{(k)} \;/\; d^{(k)*}Ad^{(k)}$ % $Ad^{(k)}$ für später abspeichern
  $x^{(k+1)} = x^{(k)} + \alpha_k d^{(k)}$
  $r^{(k+1)} = r^{(k)} - \alpha_k Ad^{(k)}$
  löse $Ms^{(k+1)} = r^{(k+1)}$
  $\beta_k = r^{(k+1)*}s^{(k+1)} \;/\; r^{(k)*}s^{(k)}$ % $r^{(k+1)*}s^{(k+1)}$ für später abspeichern
  $d^{(k+1)} = s^{(k+1)} + \beta_k d^{(k)}$
`until` stop % end for

Ergebnis: $x^{(k)}$ ist die Approximation von $A^{-1}b$, $r^{(k)} = b - Ax^{(k)}$ das zugehörige Residuum und $s^{(k)}$ das Residuum von (10.2)

Algorithmus 10.1: Präkonditioniertes CG-Verfahren (PCG-Verfahren)

Die entsprechende Transformation von Algorithmus 9.1 ergibt das *präkonditionierte CG-Verfahren* (PCG-Verfahren, Algorithmus 10.1).

Aufwand. Verglichen mit dem CG-Verfahren erhöht sich in der Regel der Aufwand des PCG-Verfahrens pro Iteration um die Lösung eines linearen Gleichungssystems $Ms = r$. Die (erhoffte) Reduktion der Iterationsanzahl aufgrund der Präkonditionierung macht sich also nur dann bezahlt, wenn derartige Gleichungssysteme entsprechend billig gelöst werden können. Dabei ist zu beachten, daß Iterationsverfahren in der Regel nur dann eingesetzt werden, wenn die Matrix A dünn besetzt ist; daher dominieren die Kosten für die Gleichungssysteme mit M leicht die Gesamtkosten des PCG-Verfahrens. ◊

Auf die Wahl von M werden wir gleich etwas detaillierter eingehen. Zunächst soll jedoch ein Analogon von Satz 9.5 formuliert werden.

Satz 10.1. *Die k-te Iterierte $x^{(k)}$ von Algorithmus* 10.1 *liegt in dem affin verschobenen Krylov-Raum*

$$x^{(0)} + \mathcal{K}_k(M^{-1}A, M^{-1}r^{(0)})$$

und ist in dieser Menge die eindeutig bestimmte Minimalstelle des Funktionals

$$\Phi(x) = \frac{1}{2}\, x^*Ax - x^*b\,.$$

Beweis. Aufgrund der Herleitung entspricht Algorithmus 10.1 dem CG-Verfahren, angewandt auf das Gleichungssystem (10.3). Nach Satz 9.5 liegt die entsprechende Iterierte $z^{(k)} = L^* x^{(k)}$ in dem affin verschobenen Krylov-Raum

$$z^{(0)} + \mathcal{K}_k(L^{-1}AL^{-*}, L^{-1}b - L^{-1}AL^{-*}z^{(0)}), \qquad z^{(0)} = L^* x^{(0)},$$

und minimiert in dieser Menge das Fehlerfunktional

$$\Psi(z) = \frac{1}{2} z^* L^{-1}AL^{-*}z - z^* L^{-1}b.$$

Durch die Transformation $x = L^{-*}z$ werden die Iterierten und die genannten Krylov-Räume aufeinander abgebildet und es gilt

$$\Psi(z) = \frac{1}{2} x^* Ax - x^* b = \Phi(x).$$

Damit ist der Satz bewiesen. □

Die Konstruktion geeigneter Präkonditionierungsmatrizen M ist eine schwierige Aufgabe, der in der gegenwärtigen Forschung viel Raum gewidmet wird. Es gibt eine ganze Reihe recht allgemeiner Ansätze, deren (zum Teil beeindruckende) Konvergenzbeschleunigung aber meist nur in Ausnahmefällen theoretisch untermauert werden kann. Grundsätzlich macht es sich bezahlt, bei der Auswahl des Präkonditionierers die Struktur der Koeffizientenmatrix A zu berücksichtigen. Ein besonders treffendes Beispiel hierfür sind Toeplitz-Matrizen, wie wir später noch in Abschnitt 54 sehen werden.

Im Rahmen dieses Buches ist es unmöglich, auch nur einen halbwegs vollständigen Überblick über die Vielzahl der möglichen Präkonditionierungsstrategien zu geben. Hierfür muß auf die jeweilige Spezialliteratur verwiesen werden, man vergleiche etwa die Bücher von Greenbaum [36] oder Golub und Van Loan [34, Abschnitt 10.3] und die dort angegebenen Literaturverweise. Lediglich auf die symmetrische Gauß-Seidel-Präkonditionierung soll hier etwas genauer eingegangen werden, da sie in vielen Fällen zu einer Konvergenzbeschleunigung führt und keinen wesentlichen Mehraufwand bei der Implementierung des PCG-Verfahrens erfordert.

Bei der *symmetrischen Gauß-Seidel-Präkonditionierung* wird in (10.2) die Matrix M aus der Zerlegung $A = M - N$ des symmetrischen Gauß-Seidel-Verfahrens verwendet. Schreiben wir wieder $A = D - R - R^*$, wobei D den Diagonalanteil und R die rechte obere Dreiecksmatrix von A bezeichnet, so hat M nach (8.11) den Cholesky-Faktor

$$L = (D - R^*)D^{-1/2}.$$

Um zumindest ansatzweise zu rechtfertigen, daß diese Präkonditionierung vorteilhaft ist, betrachten wir die bereits in Abschnitt 8 untersuchten Beispielmatrizen der Gestalt

$$A = \begin{bmatrix} I & -B^* \\ -B & I \end{bmatrix} \in \mathbb{K}^{n\times n} \tag{10.4}$$

mit $B \in \mathbb{K}^{p\times q}$, $0 < p, q < n$, $p + q = n$. Die Matrix L hat in diesem Fall die Form

$$L = \begin{bmatrix} I & 0 \\ -B & I \end{bmatrix}. \tag{10.5}$$

Proposition 10.2. *Die Matrix A aus* (10.4) *ist genau dann positiv definit, wenn $\|B\|_2 < 1$. In diesem Fall ergibt sich für L aus* (10.5)*, daß*

$$\operatorname{cond}_2(L^{-1}AL^{-*}) \le \operatorname{cond}_2(A)$$

mit Gleichheit nur für $A = L = I$.

Beweis. Nach Satz 2.7 stimmen für eine hermitesche Matrix H Spektralradius und Spektralnorm überein. Ferner ist $\|H^{-1}\|_2$ der Kehrwert des betragskleinsten Eigenwerts von H. Daher berechnen wir als nächstes die Eigenwerte von A und $L^{-1}AL^{-*}$. Unter den genannten Voraussetzungen ist

$$\begin{aligned} L^{-1}AL^{-*} &= \begin{bmatrix} I & 0 \\ B & I \end{bmatrix} \begin{bmatrix} I & -B^* \\ -B & I \end{bmatrix} \begin{bmatrix} I & B^* \\ 0 & I \end{bmatrix} \\ &= \begin{bmatrix} I & 0 \\ B & I \end{bmatrix} \begin{bmatrix} I & 0 \\ -B & I - BB^* \end{bmatrix} = \begin{bmatrix} I & 0 \\ 0 & I - BB^* \end{bmatrix}. \end{aligned}$$

An dieser Darstellung können die Eigenwerte von $L^{-1}AL^{-*}$ abgelesen werden:

$$\sigma(L^{-1}AL^{-*}) = \{1\} \cup \{\, 1 - \lambda : \lambda \in \sigma(BB^*) \,\}. \tag{10.6}$$

Die Eigenwerte von A können ebenfalls über die Eigenwerte von BB^* ausgedrückt werden. Dazu bestimmen wir die Block-LR-Zerlegung von $A - \mu I$ wie in (5.1):

$$A - \mu I = \begin{bmatrix} (1-\mu)I & -B^* \\ -B & (1-\mu)I \end{bmatrix} = \begin{bmatrix} I & 0 \\ -B/(1-\mu) & I \end{bmatrix} \begin{bmatrix} (1-\mu)I & -B^* \\ 0 & S \end{bmatrix}$$

mit dem Schur-Komplement

$$S = (1-\mu)I - BB^*/(1-\mu) \in \mathbb{K}^{p\times p}.$$

Aus den Rechenregeln für Determinanten ergibt sich daher

$$\det(A - \mu I) \;=\; (1-\mu)^q \det S \;=\; (1-\mu)^{q-p} \det\left((1-\mu)^2 I - BB^*\right).$$

Folglich ist das Spektrum von A durch die Menge $\{1 \pm \sqrt{\lambda} \,:\, \lambda \in \sigma(BB^*)\}$ gegeben, eventuell zuzüglich des Eigenwerts $\mu = 1$. Hieraus folgt, daß A genau dann positiv definit ist, wenn $\|B\|_2$ kleiner als Eins ist.

Aus den Spektren von A und $L^{-1}AL^{-*}$, vgl. (10.6), erhalten wir somit

$$\mathrm{cond}_2(A) \;=\; \frac{1+\|B\|_2}{1-\|B\|_2} \;=\; \frac{(1+\|B\|_2)^2}{1-\|B\|_2^2}\,, \qquad \mathrm{cond}_2(L^{-1}AL^{-*}) \;=\; \frac{1}{1-\|B\|_2^2}\,,$$

und die Kondition von A ist daher mindestens so groß wie die von $L^{-1}AL^{-*}$. Gleichheit tritt dabei offensichtlich nur für $B = 0$ bzw. $A = I$ ein. □

Aufgrund unserer Vorüberlegungen führt die symmetrische Gauß-Seidel-Präkonditionierung für Matrizen der Form (10.4) also in der Regel zu einer Konvergenzbeschleunigung.

Aufwand. Wenn das PCG-Verfahren mit dem symmetrischen Gauß-Seidel-Präkonditionierer wie in Algorithmus 10.1 implementiert wird, muß in jedem Schritt eine Matrix-Vektor-Multiplikation mit A und ein Gleichungssystem mit der Matrix M aus (8.11) gelöst werden. Verwendet man hierfür die Cholesky-Zerlegung von M, so ist letzteres etwa genauso aufwendig wie eine Matrix-Vektor-Multiplikation mit A. Wie bei dem symmetrischen Gauß-Seidel-Verfahren kann man aber auch hier den Aufwand ungefähr halbieren, wenn man die geschicktere Implementierung in Algorithmus 10.2 wählt, die auf dem sogenannten *Eisenstat-Trick* [25] beruht. Wir verzichten hier auf eine detaillierte Herleitung und verweisen statt dessen auf Aufgabe 17. Wie man sieht, liegt der Hauptaufwand von Algorithmus 10.2 in der Lösung der beiden Dreiecksysteme zur Berechnung der Vektoren $d^{(k)}$ und $g^{(k)}$ (die Multiplikationen mit der Diagonalmatrix D sind demgegenüber vernachlässigbar). Da $A = D - R - R^*$ ist, entspricht dieser Aufwand in etwa einer Matrix-Vektor-Multiplikation mit A. ◊

Beispiel 10.3. Wir veranschaulichen die Effizienz der symmetrischen Gauß-Seidel-Präkonditionierung schließlich noch anhand eines numerischen Beispiels. Abbildung 10.1 vergleicht die Konvergenz des zugehörigen PCG-Verfahrens und des CG-Verfahrens anhand des Beispiels 9.7 aus dem vorangegangenen Abschnitt. Anstelle der 82 Iterationen des CG-Verfahrens benötigt das PCG-Verfahren lediglich 34 Iterationen, um den relativen Fehler unter die Schranke 10^{-4} zu reduzieren. In diesem Beispiel führt die symmetrische Gauß-Seidel-Präkonditionierung also zu einer klaren Reduktion des Gesamtaufwands. ◊

Initialisierung: $A = D - R - R^* \in \mathbb{R}^{n\times n}$ sei hermitesch und positiv definit

wähle beliebiges $x^{(0)} \in \mathbb{R}^n$
löse $(D - R^*)s^{(0)} = b - Ax^{(0)}$,
$w^{(0)} = s^{(0)}$
`for` $k = 0, 1, 2, \ldots$ `do`
  löse $(D - R)d^{(k)} = Dw^{(k)}$ % $Dw^{(k)}$ für später abspeichern
  löse $(D - R^*)g^{(k)} = Dw^{(k)} - Dd^{(k)}$
  $v^{(k)} = d^{(k)} + g^{(k)}$
  $\alpha_k = s^{(k)*}Ds^{(k)} \;/\; v^{(k)*}Dw^{(k)}$
  $x^{(k+1)} = x^{(k)} + \alpha_k d^{(k)}$
  $s^{(k+1)} = s^{(k)} - \alpha_k v^{(k)}$
  $\beta_k = s^{(k+1)*}Ds^{(k+1)} \;/\; s^{(k)*}Ds^{(k)}$ % $s^{(k+1)*}Ds^{(k+1)}$ für später abspeichern
  $w^{(k+1)} = s^{(k+1)} + \beta_k w^{(k)}$
`until` stop % end for

Ergebnis: $x^{(k)}$ ist die Approximation von $A^{-1}b$, $s^{(k)} = (D - R^*)^{-1}(b - Ax^{(k)})$ ein verallgemeinertes Residuum

Algorithmus 10.2: CG-Verfahren mit symmetrischer Gauß-Seidel-Präkonditionierung

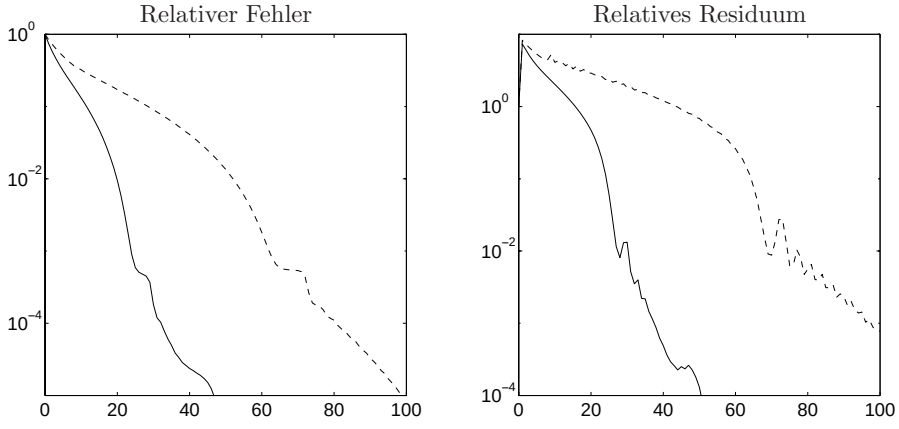


Abb. 10.1: Konvergenzverlauf von CG- und PCG-Verfahren

Aufgaben

1. Zeigen Sie, daß die Gleichgewichtsmatrix E aus Abschnitt 3 vollen Rang besitzt. Folgern Sie hieraus, daß die Steifigkeitsmatrix A positiv definit und das lineare Gleichungssystem (3.5) eindeutig lösbar ist.

2. Sei E die Gleichgewichtsmatrix eines beliebigen Tragwerks, vgl. Abschnitt 3. Weisen Sie nach, daß ein Element a_{ij} der Steifigkeitsmatrix $A = EL^{-1}E^*$ Null ist, falls die zugehörigen Gelenke z_i und z_j in dem Tragwerk nicht durch einen Stab verbunden sind. Hieraus folgt, daß die Steifigkeitsmatrix in der Regel dünn besetzt ist.

3. $A = [a_{ij}]$ sei eine strikt diagonaldominante Bandmatrix mit Bandbreite l, d. h. es gilt

$$a_{ij} = 0 \qquad \text{für } |i-j| > l\,.$$

Weisen Sie nach, daß die Faktoren L und R der LR-Zerlegung von A ebenfalls Bandbreite l haben und daß sich der Aufwand zur Berechnung der LR-Zerlegung in diesem Fall auf lediglich $l^2 n$ Multiplikationen bzw. Divisionen reduziert.

4. Seien $A \in \mathbb{K}^{n\times n}$ hermitesch und positiv definit, $A_{11} \in \mathbb{K}^{p\times p}$, $1 \le p < n$, die linke obere Submatrix von A und S das zugehörige Schur-Komplement.
(a) Zeigen Sie, daß

$$y^*Sy \;=\; \inf_{0\neq x\in\mathbb{K}^p} \begin{bmatrix} x \\ y \end{bmatrix}^* A \begin{bmatrix} x \\ y \end{bmatrix}, \qquad y \in \mathbb{K}^{n-p}\,.$$

(b) Verwenden Sie (a), um zu beweisen, daß

$$\operatorname{cond}_2(S) \le \operatorname{cond}_2(A)\,.$$

5. Sei $T \in \mathbb{K}^{n\times n}$ eine beliebige Toeplitz-Matrix und

$$S_\alpha \;=\; \begin{bmatrix} 0 & & & \alpha \\ 1 & \ddots & & \\ & \ddots & \ddots & \\ & & 1 & 0 \end{bmatrix}, \qquad \alpha \in \mathbb{C}\,.$$

(a) Zeigen Sie, daß für jedes $\alpha, \beta \in \mathbb{C}$ der Rang von $S_\alpha T - TS_\beta$ höchstens zwei ist.
(b) Für welche Toeplitz-Matrizen ist $\operatorname{Rang}(S_1 T - TS_1) < 2$?
(c) Zeigen Sie, daß für invertierbare Toeplitz-Matrizen der Rang von $S_\alpha T^{-1} - T^{-1}S_\beta$ ebenfalls höchstens zwei ist.
Aufgrund dieser Eigenschaften sagt man, Toeplitz-Matrizen haben einen *Displacement-Rang* $p = 2$.

6. Eine Matrix $C = [c_{ij}] \in \mathbb{K}^{n\times n}$ heißt *verallgemeinerte Cauchy-Matrix*, falls Vektoren $z_i, y_i \in \mathbb{C}^p$, $p \in \mathbb{N}$, und komplexe Zahlen s_i, t_i existieren $(i = 1, \ldots, n)$ mit

$$c_{ij} \;=\; \frac{z_i^* y_j}{s_i - t_j}\,, \qquad i, j = 1, \ldots, n\,.$$

Analog zu Toeplitz-Matrizen (vgl. Aufgabe 5) nennt man p den *Displacement-Rang* von C. Zeigen Sie:

(a) Ist C eine verallgemeinerte Cauchy-Matrix mit Displacement-Rang p, dann ist der Rang von $SC - CT$ höchstens gleich p; hierbei seien S und T die Diagonalmatrizen mit den Einträgen s_i bzw. t_i, $i = 1, \ldots, n$, auf den Diagonalen.
(b) Sind $S, T \in \mathbb{K}^{n\times n}$ Diagonalmatrizen, wobei alle Diagonaleinträge von S von allen Diagonaleinträgen von T verschieden sind, und ist der Rang von $SC - CT$ gleich p für eine Matrix $C \in \mathbb{K}^{n\times n}$, dann ist C eine verallgemeinerte Cauchy-Matrix mit Displacement-Rang p.
(c) Ist C eine invertierbare verallgemeinerte Cauchy-Matrix, dann ist auch C^{-1} eine verallgemeinerte Cauchy-Matrix.

7. Die Struktur verallgemeinerter Cauchy-Matrizen (vgl. Aufgabe 6) ist von Bedeutung, da sie bei der Durchführung des Gauß-Algorithmus erhalten bleibt: Beweisen Sie, daß jeweils der rechte untere quadratische Block $[a^{(k+1)}_{k+i,k+j}]^{n-k}_{i,j=1}$ der in Abschnitt 4 eingeführten Matrizen A_{k+1}, $k = 1, \ldots, n-1$, eine verallgemeinerte Cauchy-Matrix ist, falls $A = A_1 \in \mathbb{K}^{n\times n}$ eine verallgemeinerte Cauchy-Matrix ist.

8. Wenden Sie den Levinson-Algorithmus zur Lösung des n-dimensionalen linearen Gleichungssystems

$$\begin{bmatrix} 2 & -1 & & \\ -1 & 2 & \ddots & \\ & \ddots & \ddots & -1 \\ & & -1 & 2 \end{bmatrix} x = \begin{bmatrix} 1 \\ 0 \\ \vdots \\ 0 \end{bmatrix}$$

an. Bestimmen Sie anschließend mit dem Algorithmus von Trench die Inverse dieser tridiagonalen Toeplitz-Matrix.

9. Die Toeplitz-Matrix T aus (6.4) sei reell, symmetrisch und positiv definit mit $t_0 = 1$, und $y^{(k)} \in \mathbb{R}^k$, $k = 1, \ldots, n-1$, seien die Lösungen der Yule-Walker-Gleichungen (6.11). Diese Vektoren definieren eine linke untere Dreiecksmatrix

$$L = \begin{bmatrix} 1 & 0 & 0 & \cdots & 0 \\ & 1 & 0 & & 0 \\ -y^{(n-1)} & & 1 & & \vdots \\ & -y^{(n-2)} & & \ddots & 0 \\ & & -y^{(n-3)} & & 1 \end{bmatrix} \in \mathbb{R}^{n\times n}.$$

Zeigen Sie, daß $T^{-1} = LDL^*$ mit einer Diagonalmatrix D gilt, und implementieren Sie diese Faktorisierung mit lediglich $n^2 + O(n)$ Multiplikationen.

10. Untersuchen Sie die Variante des Einzelschrittverfahrens, bei der die Gleichungen der Matrix *rückwärts* durchlaufen werden, d. h. der innere Schleifenindex i in Algorithmus 8.2 läuft rückwärts von $i = n$ bis $i = 1$.
(a) Bestimmen Sie die Iterationsmatrix $\mathcal{U}$ dieses Verfahrens.
(b) Zeigen Sie, daß für eine Tridiagonalmatrix $A = D - L - R$ und beliebiges $\alpha \in \mathbb{R} \setminus \{0\}$ die Matrizen $D - L - \alpha R$ und $D - R - \alpha L$ zueinander ähnlich sind.
(c) Folgern Sie aus (b), daß $\mathcal{U}$ und die Iterationsmatrix $\mathcal{L}$ des Einzelschrittverfahrens bei Tridiagonalmatrizen A mit nichtsingulärem Diagonalanteil die gleichen Eigenwerte haben.

(d) Implementieren Sie beide Varianten des Gauß-Seidel-Verfahrens und wenden Sie sie auf das 19×19-dimensionale lineare Gleichungssystem

$$\begin{bmatrix} 2\varepsilon+h & -\varepsilon-h & & \\ -\varepsilon & 2\varepsilon+h & \ddots & \\ & \ddots & \ddots & -\varepsilon-h \\ & & -\varepsilon & 2\varepsilon+h \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_{19} \end{bmatrix} = \begin{bmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{bmatrix}$$

mit $\varepsilon = 10^{-6}$ und $h = 0.05$ an. Verwenden Sie in beiden Fällen den gleichen Startvektor $x^{(0)} = 0$. Interpretieren Sie die Ergebnisse und vergleichen Sie mit (c). Gleichungssysteme dieser Bauart treten bei Randwertaufgaben für singulär gestörte Differentialgleichungen auf, vgl. Abschnitt 85.

11. Beim SOR-Verfahren (engl.: ***s**uccessive **o**ver**r**elaxation method*) wählt man ein $\omega \in \mathbb{R}$ und ersetzt in Algorithmus 8.2 die Rechenvorschrift für $x_i^{(k+1)}$ durch

$$x_i^{(k+1)} \;=\; (1-\omega)x_i^{(k)} \;+\; \frac{\omega}{a_{ii}} \left(b_i - \sum_{j<i} a_{ij} x_j^{(k+1)} - \sum_{j>i} a_{ij} x_j^{(k)} \right).$$

(a) Zeigen Sie, daß das SOR-Verfahren in das allgemeine Schema aus Algorithmus 7.1 paßt, und zwar für $M_\omega = \frac{1}{\omega} D - L$, falls die Koeffizientenmatrix A wie in (8.2) zerlegt wird.
(b) Die Iterationsmatrix des SOR-Verfahrens wird üblicherweise mit $\mathcal{L}_\omega$ bezeichnet. Rechnen Sie nach, daß

$$\det(\mathcal{L}_\omega) \;=\; (1-\omega)^n \,.$$

Folgern Sie hieraus, daß das SOR-Verfahren allenfalls für $\omega \in (0,2)$ konvergieren kann.
(c) Berechnen Sie $\mathcal{L}_\omega$ für Matrizen A der Gestalt (8.6). Weisen Sie für diesen Fall die folgende Identität nach:

$$\big(\mathcal{L}_\omega + (\omega-1)I\big)^2 = \omega^2 \mathcal{J}^2 \mathcal{L}_\omega \,.$$

12. Sei $A \in \mathbb{K}^{n\times n}$ hermitesch und positiv definit. Welche Basen des $\mathbb{K}^n$ sind sowohl im euklidischen Innenprodukt als auch im Energie-Innenprodukt (vgl. Definition 9.1) orthogonal?

13. $x^{(k)}$ bezeichne die k-te Iterierte des CG-Verfahrens angewandt auf $Ax = b$. Zeigen Sie, daß die Iteration von der Wahl des zugrundeliegenden orthonormalen Koordinatensystems unabhängig ist: Ist $V \in \mathbb{K}^{n\times n}$ eine unitäre Matrix,

$$\widetilde{A} = VAV^* \,, \qquad \widetilde{b} = Vb \,, \qquad \widetilde{x} = Vx \,,$$

und wird das CG-Verfahren auf das Gleichungssystem $\widetilde{A}\widetilde{x} = \widetilde{b}$ mit Startvektor $\widetilde{x}^{(0)} = Vx^{(0)}$ angewendet, so ergibt sich die k-te Iterierte $\widetilde{x}^{(k)} = Vx^{(k)}$.

14. $A \in \mathbb{R}^{3\times 3}$ sei symmetrisch und positiv definit und besitze lediglich zwei verschiedene Eigenwerte. Zeigen Sie, daß das CG-Verfahren nach maximal zwei Iterationen die exakte Lösung berechnet hat. Unter welcher Zusatzvoraussetzung liegt Konvergenz im ersten Schritt vor, falls $x^{(0)} = 0$ gewählt wird?
Hinweis: Verwenden Sie ein Koordinatensystem, in dem die Basisvektoren die Eigenvektoren von A sind (dies ist nach Aufgabe 13 zulässig).

15. (a) Zeigen Sie, daß die Euklidnorm des Fehlers $x^{(k)} - \widehat{x}$ für das CG-Verfahren in Abhängigkeit von k monoton fallend ist.
(b) Implementieren Sie das CG-Verfahren, und testen Sie Ihr Programm an dem Gleichungssystem $Ax = b$ mit $b = [1, \ldots, 1]^T$ und $A = LL^T$, wobei

$$L = \begin{bmatrix} l_1 & & & & \\ & l_2 & & & \\ & & \ddots & & \\ & & & l_n & \\ \epsilon & \epsilon & \cdots & \epsilon & \epsilon \end{bmatrix} \in \mathbb{R}^{(n+1)\times(n+1)} \qquad \text{mit} \quad \epsilon = 10^{-4} \text{ und } l_i = (i+1)/i\,.$$

Wählen Sie $n = 24$, und berechnen Sie die exakte Lösung $\widehat{x}$ über die gegebene Cholesky-Zerlegung $A = LL^T$. Verwenden Sie für das CG-Verfahren

$$x^{(0)} = \widehat{x} - 10^4 \mathsf{eps}\, b$$

als Startvektor, wobei eps wie in Abschnitt 1 die Maschinengenauigkeit bezeichnet. Lassen Sie sich den Fehler $\|x^{(k)} - \widehat{x}\|_2$ der ersten zehn Iterationen ausgeben, und vergleichen Sie die Werte mit dem theoretischen Ergebnis aus Teilaufgabe (a).

16. Implementieren Sie das CG-Verfahren für Beispiel 9.7, und experimentieren Sie mit verschiedenen rechten Seiten b. Wie ändert sich die Konvergenzgeschwindigkeit? Vergleichen Sie Ihre Ergebnisse mit dem Schätzwert $q \approx 0.9686$ aus (10.1) für die mittlere Konvergenzrate.

17. Sei $A = D - R - R^* \in \mathbb{R}^{n\times n}$ eine hermitesche und positiv definite Matrix mit Diagonalanteil D und oberem Dreiecksanteil R. Mit $D^{\pm 1/2}$ werden die beiden Diagonalmatrizen bezeichnet, deren Diagonaleinträge die Quadratwurzeln der entsprechenden Einträge von $D^{\pm 1}$ sind. Schließlich sei $L = (D - R^*)D^{-1/2}$.
(a) Zeigen Sie: Ist $p \in \mathbb{R}^n$, $w = D^{-1/2}p$ und $d = (D - R)^{-1}Dw$, so gilt

$$L^{-1}AL^{-*}p = D^{1/2}\big(d + (D - R^*)^{-1}D(w - d)\big)\,.$$

(b) Übertragen Sie Algorithmus 9.1 auf das Gleichungssystem $L^{-1}AL^{-*}z = L^{-1}b$; bezeichnen Sie die Iterierten mit $z^{(k)}$ und verwenden Sie den Namen $p^{(k)}$ anstelle von $d^{(k)}$ für die Suchrichtungen.
(c) Ersetzen Sie in diesem Algorithmus das Matrix-Vektor-Produkt $L^{-1}AL^{-*}p^{(k)}$ durch die Darstellung aus Aufgabenteil (a) und speichern Sie das Ergebnis in $q^{(k)}$. Führen Sie die hierzu notwendigen Hilfsvariablen $w^{(k)}$ und $d^{(k)}$ entsprechend ein.
(d) Sie erhalten nun Algorithmus 10.2, wenn Sie die Variablen z, r und q durch neue Variablen $x = L^{-*}z$, $s = D^{-1/2}r$ und $v = D^{-1/2}q$ ersetzen. Die Variable p kann durch die bereits vorhandene Variable $w = D^{-1/2}p$ ersetzt werden. Machen Sie sich klar, daß jeweils

$$s^{(k)} = (D - R^*)^{-1}(b - Ax^{(k)})$$

gilt.

III Lineare Ausgleichsrechnung

Übersteigt die Anzahl der Gleichungen die der Unbekannten, so ist das Gleichungssystem *überbestimmt* und in der Regel nicht lösbar. Überbestimmte Gleichungssysteme $Ax = b$ mit $A \in \mathbb{K}^{m\times n}$, $m > n$, treten jedoch vielfach in Anwendungen auf, etwa in dem Tomographiebeispiel aus der Einleitung. Oft ist es dann sinnvoll, die Lösung $\widehat{x}$ des sogenannten *linearen Ausgleichsproblems*

$$\text{minimiere} \quad \|b - Ax\|_2 \quad \text{über } x \in \mathbb{K}^n$$

zu suchen, etwa wenn der Vektor b aus fehlerbehafteten Meßdaten besteht und die Meßfehler geeignete statistische Eigenschaften aufweisen.

Die so definierten verallgemeinerten Lösungen des Gleichungssystems heißen *Kleinste-Quadrate-Lösungen*. Im folgenden untersuchen wir diesen neuen Lösungsbegriff genauer und leiten numerische Verfahren zur Berechnung Kleinster-Quadrate-Lösungen her. Eine unerschöpfliche Quelle für weitergehende Fragestellungen ist das Buch von Björck [8].

11 Die Gaußschen Normalengleichungen

Im weiteren sei also $A \in \mathbb{K}^{m\times n}$ und $b \in \mathbb{K}^m$, wobei die Einschränkung $m > n$ erst später relevant wird. Das Hauptresultat dieses Abschnitts führt den neuen Lösungsbegriff der Kleinsten-Quadrate-Lösung auf die Lösung eines quadratischen linearen Gleichungssystems zurück, die sogenannten *Gaußschen Normalengleichungen*

$$A^*Ax = A^*b\,. \tag{11.1}$$

Satz 11.1. *Sei $A \in \mathbb{K}^{m\times n}$ und $b \in \mathbb{K}^m$. Jede Lösung des linearen Ausgleichsproblems*

$$\textit{minimiere} \quad \|b - Ax\|_2 \tag{11.2}$$

ist eine Lösung der Gaußschen Normalengleichungen (11.1) *und umgekehrt.*

Bevor wir diesen Satz beweisen, untersuchen wir zunächst die Koeffizientenmatrix A^*A aus (11.1).

Lemma 11.2. *Die Matrix A^*A ist hermitesch und positiv semidefinit. Darüber hinaus ist A^*A genau dann positiv definit, wenn der Kern von A trivial ist, d. h. wenn $\mathcal{N}(A) = \{0\}$ ist. Ferner ist in jedem Fall*

$$\mathcal{N}(A^*A) = \mathcal{N}(A) \qquad \textit{und} \qquad \mathcal{R}(A^*A) = \mathcal{R}(A^*) = \mathcal{N}(A)^\perp .$$

Beweis. Offensichtlich ist A^*A hermitesch. Ferner ist

$$x^*A^*Ax \;=\; \|Ax\|_2^2 \;\geq\; 0 \qquad \text{für alle } x \in \mathbb{K}^n ,$$

d. h. A^*A ist positiv semidefinit und es gilt $\mathcal{N}(A^*A) \subset \mathcal{N}(A)$. Die umgekehrte Inklusion $\mathcal{N}(A) \subset \mathcal{N}(A^*A)$ ist offensichtlich und somit ist $\mathcal{N}(A) = \mathcal{N}(A^*A)$. Weiterhin gilt trivialerweise $\mathcal{R}(A^*A) \subset \mathcal{R}(A^*)$ und aufgrund der bereits nachgewiesenen Identität für $\mathcal{N}(A^*A)$ folgt

$$\begin{aligned} \dim \mathcal{R}(A^*A) \;&=\; n - \dim \mathcal{N}(A^*A) \;=\; n - \dim \mathcal{N}(A) \\ &=\; \operatorname{Rang} A \;=\; \dim \mathcal{R}(A^*) . \end{aligned} \tag{11.3}$$

Damit ist $\mathcal{R}(A^*A) = \mathcal{R}(A^*)$, und es verbleibt noch zu zeigen, daß $\mathcal{R}(A^*)$ und $\mathcal{N}(A)$ orthogonale Komplementärräume sind. Seien also $z \in \mathcal{R}(A^*)$ und $x \in \mathcal{N}(A)$ beliebig gewählt: Dann existiert ein $y \in \mathbb{K}^m$ mit $z = A^*y$ und es folgt

$$x^*z \;=\; x^*A^*y \;=\; (Ax)^*y \;=\; 0^*y \;=\; 0 .$$

Daher sind $\mathcal{N}(A)$ und $\mathcal{R}(A^*)$ zueinander orthogonal und wegen

$$\dim \mathcal{N}(A) + \dim \mathcal{R}(A^*) = n ,$$

vgl. (11.3), folgt $\mathcal{R}(A^*) = \mathcal{N}(A)^\perp$. □

Nun können wir Satz 11.1 beweisen:

Beweis von Satz 11.1. Für $x \in \mathbb{K}^n$ sei

$$\begin{aligned} \Phi(x) \;&=\; \frac{1}{2}\,\|b - Ax\|_2^2 \;=\; \frac{1}{2}\,(b - Ax)^*(b - Ax) \\ &=\; \frac{1}{2}\,x^*A^*Ax - \operatorname{Re} x^*A^*b + \frac{1}{2}\,\|b\|_2^2 . \end{aligned}$$

Wie in Abschnitt 9 erhalten wir für jede Lösung $\widehat{x}$ der Normalengleichungen durch quadratische Ergänzung

$$\begin{aligned} \Phi(x) - \Phi(\widehat{x}) \;&=\; \frac{1}{2}\,(x - \widehat{x})^*A^*A(x - \widehat{x}) \;-\; \widehat{x}^*A^*b \;+\; \operatorname{Re}\widehat{x}^*A^*b \\ &=\; \frac{1}{2}\,(x - \widehat{x})^*A^*A(x - \widehat{x}) , \end{aligned}$$

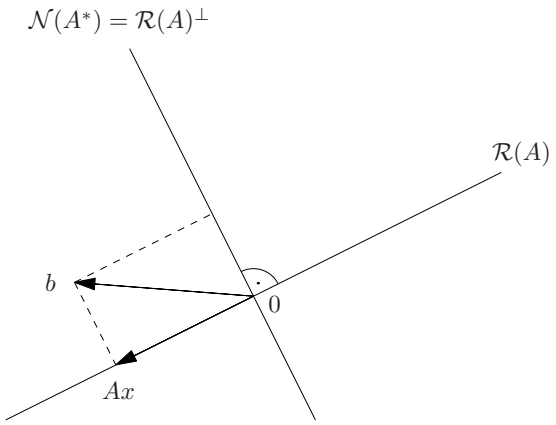


Abb. 11.1: Kleinste-Quadrate-Lösung

da $\widehat{x}^* A^* b = \widehat{x}^* A^* A\widehat{x}$ reell ist. Weil A^*A nach Lemma 11.2 zudem positiv semidefinit ist, ergibt sich $\Phi(x) \geq \Phi(\widehat{x})$, und folglich wird das globale Minimum von Φ an den Lösungen von (11.1) und nur dort angenommen. Derartige Lösungen existieren nach Lemma 11.2, da A^*b zu $\mathcal{R}(A^*) = \mathcal{R}(A^*A)$ gehört. Zu beachten ist allerdings, daß die Matrix A^*A singulär sein kann; in diesem Fall haben die Gaußschen Normalengleichungen mehrere Lösungen und das quadratische Funktional Φ wird an allen Lösungen minimal. □

Den Normalengleichungen (11.1) läßt sich entnehmen, daß das Residuum $r = b - Ax$ zu $\mathcal{N}(A^*)$ gehört. Ersetzt man in Lemma 11.2 A durch A^*, ergibt sich

$$\mathcal{R}(A) = \mathcal{N}(A^*)^\perp \qquad \text{bzw.} \qquad \mathcal{N}(A^*) = \mathcal{R}(A)^\perp . \tag{11.4}$$

Somit steht das Residuum senkrecht zum Bild von A, vgl. Abbildung 11.1.

Bemerkung 11.3. Im reellen Fall $\mathbb{K} = \mathbb{R}$ kann ein alternativer Beweis von Satz 11.1 mit den Methoden der Differentialrechnung geführt werden. Dazu bestimmt man wie in Abschnitt 9 die Richtungsableitung von Φ in Richtung $d \in \mathbb{R}^n$,

$$\frac{\partial}{\partial d}\Phi(x) \;=\; \lim_{t\to 0} \frac{\Phi(x+td) - \Phi(x)}{t} .$$

Wegen

$$\begin{aligned} &\Phi(x+td) - \Phi(x) \\ &\quad = \frac{1}{2}\,(b - Ax - tAd)^*(b - Ax - tAd) - \frac{1}{2}\,(b-Ax)^*(b-Ax) \\ &\quad = t(Ad)^*(Ax - b) + \frac{1}{2}\,t^2 (Ad)^* Ad \end{aligned}$$

ergibt dies

$$\frac{\partial}{\partial d}\,\Phi(x) \;=\; (Ax-b)^*Ad\,,$$

und somit ist

$$\operatorname{grad}\Phi(x) \;=\; A^*(Ax-b) \tag{11.5}$$

das negative Residuum der Normalengleichungen. Da der Gradient an der Minimalstelle von Φ verschwindet, lösen die Minimalstellen von Φ somit notwendigerweise die Gaußschen Normalengleichungen. ◊

Beispiel 11.4. Gesucht ist die Gerade $y = \alpha + \beta x$, deren y-Werte kleinsten Quadratsummen-Abstand von den vorgegebenen Daten $\tilde{y} = 1, 2, 6, 4$ an den entsprechenden Abszissen $x = 0, 3, 4, 7$ haben, die sogenannte *Ausgleichsgerade*. Dies führt auf das Gleichungssystem

$$A\begin{bmatrix}\alpha\\ \beta\end{bmatrix} = b \qquad \text{mit} \quad A = \begin{bmatrix}1 & 0\\ 1 & 3\\ 1 & 4\\ 1 & 7\end{bmatrix}, \quad b = \begin{bmatrix}1\\ 2\\ 6\\ 4\end{bmatrix},$$

das im Kleinste-Quadrate-Sinn zu lösen ist. Wegen

$$A^*A \;=\; \begin{bmatrix}1 & 1 & 1 & 1\\ 0 & 3 & 4 & 7\end{bmatrix}\begin{bmatrix}1 & 0\\ 1 & 3\\ 1 & 4\\ 1 & 7\end{bmatrix} \;=\; \begin{bmatrix}4 & 14\\ 14 & 74\end{bmatrix},$$

$$A^*b \;=\; \begin{bmatrix}1 & 1 & 1 & 1\\ 0 & 3 & 4 & 7\end{bmatrix}\begin{bmatrix}1\\ 2\\ 6\\ 4\end{bmatrix} \;=\; \begin{bmatrix}13\\ 58\end{bmatrix},$$

lauten die zugehörigen Normalengleichungen

$$\begin{bmatrix}4 & 14\\ 14 & 74\end{bmatrix}\begin{bmatrix}\alpha\\ \beta\end{bmatrix} \;=\; \begin{bmatrix}13\\ 58\end{bmatrix}.$$

Die Determinante der Matrix A^*A ist von Null verschieden, also haben die Normalengleichungen eine eindeutige Lösung, nämlich $[\alpha, \beta]^T = [3/2, 1/2]^T$. Zur Probe überzeugt man sich, daß

$$b - A\widehat{x} \;=\; \begin{bmatrix}1\\ 2\\ 6\\ 4\end{bmatrix} - \begin{bmatrix}1 & 0\\ 1 & 3\\ 1 & 4\\ 1 & 7\end{bmatrix}\begin{bmatrix}3/2\\ 1/2\end{bmatrix} \;=\; \begin{bmatrix}-1/2\\ -1\\ 5/2\\ -1\end{bmatrix}$$

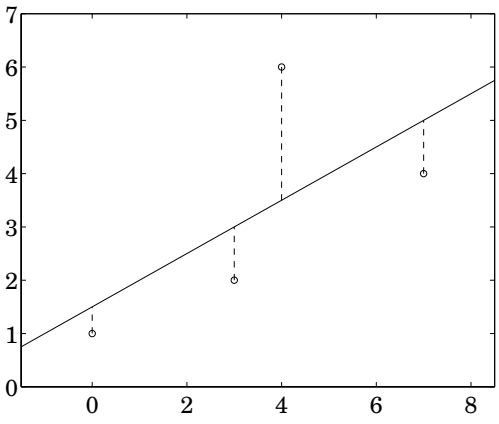


Abb. 11.2: Ausgleichsgerade

senkrecht auf $\mathcal{R}(A)$ steht.

Die zugehörige Gerade ist in Abbildung 11.2 dargestellt. Die Quadratsumme der Längen aller gebrochenen Linien wird durch diese Gerade minimiert. ◊

12 Singulärwertzerlegung und Pseudoinverse

Offensichtlich spielt die Matrix A^*A eine große Rolle beim linearen Ausgleichsproblem. Im folgenden sei p der Rang von A, $\lambda_1, \ldots, \lambda_n$ bezeichnen die absteigend sortierten Eigenwerte von A^*A,

$$\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_p > \lambda_{p+1} = \ldots = \lambda_n = 0\,,$$

sowie $v_1, \ldots, v_n \in \mathbb{K}^n$ eine zugehörige Basis aus orthonormalen Eigenvektoren. Eine solche Spektralzerlegung existiert, da A^*A hermitesch und positiv semidefinit ist, vgl. Kapitel V. Schließlich führen wir Vektoren $u_i \in \mathbb{K}^m$ ein durch

$$u_i = \frac{1}{\sqrt{\lambda_i}}\, Av_i\,, \qquad i = 1, \ldots, p\,. \tag{12.1}$$

Für $1 \leq i, j \leq p$ folgt dann,

$$u_i^* u_j = \frac{1}{\sqrt{\lambda_i}} \frac{1}{\sqrt{\lambda_j}}\, (Av_i)^*(Av_j) = \frac{1}{\sqrt{\lambda_i \lambda_j}}\, v_i^*(A^*Av_j) = \frac{\lambda_j}{\sqrt{\lambda_i \lambda_j}}\, v_i^* v_j = \delta_{ij}\,.$$

Die Vektoren $\{u_i : i = 1, \dots, p\}$ bilden also eine Orthonormalbasis des gesamten Bildraums $\mathcal{R}(A)$, denn

$$\dim \mathcal{R}(A) = n - \dim \mathcal{N}(A) = n - \dim \mathcal{N}(A^*A) = n - (n-p) = p\,.$$

Dieses Orthonormalsystem kann durch weitere $m - p$ Vektoren $u_{p+1}, \dots, u_m$ zu einer Orthonormalbasis des $\mathbb{K}^m$ ergänzt werden, wobei diese zusätzlichen Vektoren den Raum $\mathcal{R}(A)^\perp = \mathcal{N}(A^*)$ aufspannen, vgl. (11.4).

Wegen (12.1) gilt

$$\begin{aligned} A^*u_i &= \frac{1}{\sqrt{\lambda_i}}\, A^*Av_i = \sqrt{\lambda_i}\, v_i\,, && i = 1, \dots, p\,, \\ A^*u_i &= 0\,, && i = p+1, \dots, m\,. \end{aligned}$$

Wir fassen zusammen:

Definition und Satz 12.1. *Jede Matrix $A \in \mathbb{K}^{m\times n}$ mit Rang p besitzt eine* Singulärwertzerlegung*, d. h. ein System*

$$\{\, \sigma_i, u_j, v_k \,:\, i = 1, \dots, p,\ j = 1, \dots, m,\ k = 1, \dots, n \,\}$$

mit $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_p > 0$ und Orthonormalbasen $\{u_j\}_{j=1}^m$ und $\{v_k\}_{k=1}^n$ des $\mathbb{K}^m$ bzw. $\mathbb{K}^n$, wobei

$$\begin{aligned} &Av_i = \sigma_i u_i\,, && A^*u_i = \sigma_i v_i\,, && i = 1, \dots, p\,, \\ &Av_k = 0\,, && A^*u_j = 0\,, && j, k > p\,. \end{aligned}$$

Die σ_i sind die sogenannten Singulärwerte *von A. Ihre Quadrate σ_i^2 sind (entsprechend ihrer Vielfachheit) genau die von Null verschiedenen Eigenwerte von A^*A.*

In Matrixnotation läßt sich dieses Ergebnis kürzer formulieren. Dazu führen wir die Matrizen

$$U = \big[u_1, \dots, u_m\big] \in \mathbb{K}^{m\times m} \qquad \text{und} \qquad V = \big[v_1, \dots, v_n\big] \in \mathbb{K}^{n\times n}$$

ein. Da deren Spalten Orthonormalbasen bilden, sind U und V unitär, d. h.

$$U^*U = I \in \mathbb{K}^{m\times m} \qquad \text{und} \qquad V^*V = I \in \mathbb{K}^{n\times n}\,.$$

Mit

$$\Sigma = \left[\begin{array}{ccc:c} \sigma_1 & & 0 & 0 \\ & \ddots & & \vdots \\ 0 & & \sigma_p & 0 \\ \hdashline 0 & \cdots & 0 & 0 \\ & & & \end{array}\right] \in \mathbb{K}^{m\times n} \tag{12.2}$$

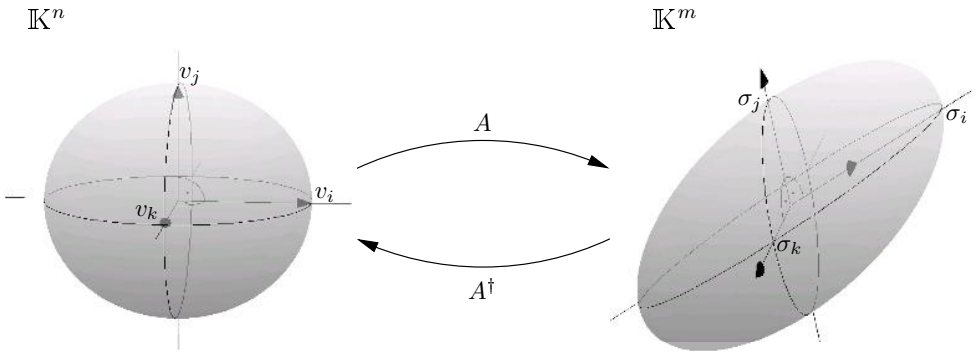


Abb. 12.1: Abbildungseigenschaften von A und $A^\dagger$

ergibt sich dann aus Satz 12.1 $AV = U\Sigma$ und daher ist

$$A = U\Sigma V^*, \qquad A^* = V\Sigma^* U^*. \tag{12.3}$$

Alternativ kann die Summendarstellung

$$A = \sum_{i=1}^{p} \sigma_i u_i v_i^* \,, \qquad A^* = \sum_{i=1}^{p} \sigma_i v_i u_i^* \,, \tag{12.4}$$

verwendet werden, deren Gültigkeit man ebenfalls leicht anhand von Satz 12.1 überprüfen kann.

Abbildung 12.1 illustriert die Abbildungseigenschaften von A. Wenn die Vektoren $x = \xi_i v_i + \xi_j v_j + \xi_k v_k$ die Einheitssphäre des Unterraums $\operatorname{span}\{v_i, v_j, v_k\}$ des $\mathbb{K}^n$ durchlaufen (d. h. $\xi_i^2 + \xi_j^2 + \xi_k^2 = \|x\|_2^2 = 1$), dann durchlaufen ihre Bilder

$$Ax \;=\; \sigma_i \xi_i u_i + \sigma_j \xi_j u_j + \sigma_k \xi_k u_k \;=\; \eta_i u_i + \eta_j u_j + \eta_k u_k$$

ein Ellipsoid in dem durch u_i, u_j und u_k aufgespannten Teilraum des $\mathbb{K}^m$: Es gilt nämlich

$$\frac{1}{\sigma_i^2}\,\eta_i^2 \;+\; \frac{1}{\sigma_j^2}\,\eta_j^2 \;+\; \frac{1}{\sigma_k^2}\,\eta_k^2 \;=\; \xi_i^2 + \xi_j^2 + \xi_k^2 = 1\,,$$

und dies ist die Gleichung eines Ellipsoids mit Scheitelpunkten $(\pm\sigma_i, 0, 0)$, $(0, \pm\sigma_j, 0)$ und $(0, 0, \pm\sigma_k)$ in den zu (u_i, u_j, u_k) gehörenden Koordinaten.

Definition 12.2. Sei $U\Sigma V^*$ die Singulärwertzerlegung (12.3) der Matrix A

aus $\mathbb{K}^{m\times n}$ mit Σ aus (12.2). Dann heißt die Matrix $A^\dagger = V\Sigma^\dagger U^* \in \mathbb{K}^{n\times m}$ mit

$$\Sigma^\dagger = \left[\begin{array}{ccc:c} \sigma_1^{-1} & & 0 & 0 \\ & \ddots & & \vdots \\ 0 & & \sigma_p^{-1} & 0 \\ \hdashline 0 & \cdots & 0 & 0 \end{array}\right] \in \mathbb{K}^{n\times m}$$

Pseudoinverse oder *Moore-Penrose-Inverse* von A (insbesondere ist $\Sigma^\dagger$ die Pseudoinverse von Σ).

Entsprechend zu (12.4) gilt die Darstellung

$$A^\dagger = \sum_{i=1}^{p} \sigma_i^{-1} v_i u_i^* \,, \tag{12.5}$$

aus der unmittelbar folgt, daß

$$\mathcal{N}(A^\dagger) = \mathcal{N}(A^*) = \mathcal{R}(A)^\perp, \qquad \mathcal{R}(A^\dagger) = \mathcal{R}(A^*) = \mathcal{N}(A)^\perp \,. \tag{12.6}$$

Beispiel 12.3. Für die Matrix

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

spannt der Vektor $[1,1,1,1]^T$ offensichtlich den Bildraum $\mathcal{R}(A)$ auf und der Vektor $[1,-1]^T$ den Kern $\mathcal{N}(A)$. Daher wird $\mathcal{R}(A^\dagger) = \mathcal{N}(A)^\perp$ durch den Vektor $[1,1]^T$ aufgespannt und $A^\dagger$ hat folglich die Form

$$A^\dagger = \begin{bmatrix} \alpha & \beta & \gamma & \delta \\ \alpha & \beta & \gamma & \delta \end{bmatrix}$$

mit geeigneten Koeffizienten α, β, γ und δ. Wegen $\mathcal{N}(A^\dagger) = \mathcal{R}(A)^\perp$ ist $A^\dagger u = 0$ für jedes $u \in \mathbb{R}^4$, das senkrecht auf dem Vektor $[1,1,1,1]^T$ steht. Somit muß zwangsläufig jede Zeile von $A^\dagger$ ein skalares Vielfaches des Zeilenvektors $[1,1,1,1]$ sein, und es verbleibt lediglich noch der Parameter α in

$$A^\dagger = \alpha \begin{bmatrix} 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 \end{bmatrix}$$

zu berechnen. Zuvor bestimmen wir die relevanten Teile der Singulärwertzerlegung von A: Offensichtlich ist $p = 1$,

$$v_1 = \frac{1}{\sqrt{2}} \begin{bmatrix} 1 \\ 1 \end{bmatrix}, \quad u_1 = \frac{1}{2} \begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \end{bmatrix} \qquad \text{und} \qquad Av_1 = \sqrt{2} \begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \end{bmatrix} = 2\sqrt{2}\, u_1 ;$$

also ist $\sigma_1 = 2\sqrt{2}$. Hieraus läßt sich schließlich α rekonstruieren: Wegen

$$\frac{\alpha}{2} \begin{bmatrix} 4 \\ 4 \end{bmatrix} = A^\dagger u_1 \overset{!}{=} \frac{1}{2\sqrt{2}}\, v_1 = \frac{1}{4} \begin{bmatrix} 1 \\ 1 \end{bmatrix}$$

ergibt sich $\alpha = 1/8$. ◊

Der Name „Pseudoinverse“ beruht auf dem folgenden Resultat:

Satz 12.4. *Die Pseudoinverse $A^\dagger$ von $A \in \mathbb{K}^{m\times n}$ ist die eindeutige Lösung der vier Gleichungen*

$$\begin{aligned} &(i) \quad AXA = A\,, && (ii) \quad XAX = X\,, \\ &(iii) \quad (AX)^* = AX\,, && (iv) \quad (XA)^* = XA\,. \end{aligned}$$

Beweis. Zunächst weisen wir nach, daß $X = A^\dagger$ die vier Gleichungen erfüllt. Wegen

$$\Sigma\Sigma^\dagger = \left[\begin{array}{ccc:c} \sigma_1 & & 0 & 0 \\ & \ddots & & \vdots \\ 0 & & \sigma_p & 0 \\ \hdashline 0 & \cdots & 0 & 0 \end{array}\right] \left[\begin{array}{ccc:c} \sigma_1^{-1} & & 0 & 0 \\ & \ddots & & \vdots \\ 0 & & \sigma_p^{-1} & 0 \\ \hdashline 0 & \cdots & 0 & 0 \end{array}\right] = \left[\begin{array}{c:c} I & 0 \\ \hdashline 0 & 0 \end{array}\right] \tag{12.7}$$

($\Sigma\Sigma^\dagger$ gehört zu $\mathbb{K}^{m\times m}$) ergibt sich zunächst

$$\Sigma\Sigma^\dagger\Sigma = \Sigma \qquad \text{und} \qquad \Sigma^\dagger\Sigma\Sigma^\dagger = \Sigma^\dagger\,.$$

Daraus folgt

$$\begin{aligned} AA^\dagger A &= U\Sigma V^* V\Sigma^\dagger U^* U\Sigma V^* = U\Sigma\Sigma^\dagger\Sigma V^* = U\Sigma V^* = A\,, \\ A^\dagger AA^\dagger &= V\Sigma^\dagger U^* U\Sigma V^* V\Sigma^\dagger U^* = V\Sigma^\dagger\Sigma\Sigma^\dagger U^* = V\Sigma^\dagger U^* = A^\dagger\,. \end{aligned}$$

Ferner ist $AA^\dagger = U\Sigma\Sigma^\dagger U^*$ und somit wegen (12.7) hermitesch. Entsprechend sieht man, daß $A^\dagger A$ hermitesch ist. Also erfüllt $X = A^\dagger$ die Gleichungen (i) bis (iv).

Es verbleibt noch zu zeigen, daß die vier Gleichungen nur die eine Lösung haben. Sei also X eine Lösung der Gleichungen (i)-(iv). Aus (i) folgt dann

$$0 \;=\; AXAv_i - Av_i \;=\; A(XAv_i - v_i)\,, \qquad i = 1, \ldots, p\,.$$

Wegen $Av_i = \sigma_i u_i$ bedeutet dies, daß

$$Xu_i \;=\; \frac{1}{\sigma_i}\, v_i + w_i \qquad \text{für ein } \; w_i \in \mathcal{N}(A)\,.$$

Nach (iv) und (11.4) ist jedoch $\mathcal{R}(XA) = \mathcal{N}(XA)^\perp$ und wegen $w_i \in \mathcal{N}(A) \subset \mathcal{N}(XA)$ folgt für jedes $i = 1, \ldots, p$

$$0 \;=\; w_i^* XA(\frac{1}{\sigma_i} v_i) \;=\; w_i^* Xu_i \;=\; w_i^*(\frac{1}{\sigma_i} v_i + w_i) \;=\; \frac{1}{\sigma_i}\, w_i^* v_i + \|w_i\|_2^2\,.$$

Wegen $w_i \in \mathcal{N}(A)$ und $v_i \in \mathcal{N}(A)^\perp$ verschwindet der erste Term auf der rechten Seite und es ergibt sich $w_i = 0$. Somit ist

$$Xu_i \;=\; \frac{1}{\sigma_i}\, v_i\,, \qquad i = 1, \ldots, p\,. \tag{12.8}$$

Hieraus folgt die Inklusion

$$\begin{aligned} \mathcal{R}(AX) \;&\supset\; \operatorname{span}\{AXu_i \,:\, i = 1, \ldots, p\} \;=\; \operatorname{span}\{Av_i \,:\, i = 1, \ldots, p\} \\ &=\; \operatorname{span}\{u_i \,:\, i = 1, \ldots, p\} \;=\; \mathcal{R}(A)\,. \end{aligned}$$

Da andererseits trivialerweise $\mathcal{R}(AX) \subset \mathcal{R}(A)$ ist, ergibt dies

$$\mathcal{R}(AX) = \mathcal{R}(A). \tag{12.9}$$

Aus (iii) und (11.4) folgt weiterhin

$$\mathcal{N}(AX) \;=\; \mathcal{R}(AX)^\perp \;=\; \mathcal{R}(A)^\perp \;=\; \operatorname{span}\{u_i \,:\, i = p+1, \ldots, m\}\,.$$

Demnach ist

$$AXu_i = 0 \quad \text{bzw.} \quad Xu_i = \tilde{w}_i \in \mathcal{N}(A)\,, \qquad i = p+1, \ldots, m\,.$$

Wegen (ii) muß allerdings $\tilde{w}_i = 0$ sein, denn

$$\tilde{w}_i \;=\; Xu_i \;=\; X(AXu_i) \;=\; 0\,, \qquad i = p+1, \ldots, m\,. \tag{12.10}$$

Ein Vergleich von (12.8) und (12.10) mit (12.5) zeigt, daß X und $A^\dagger$ übereinstimmen. Also ist $A^\dagger$ die einzige Lösung der vier Gleichungen (i) bis (iv).

□

Für invertierbare Matrizen $A \in \mathbb{K}^{n\times n}$ ist $A^\dagger = A^{-1}$ aufgrund der Gleichungen (i) oder (ii). Demnach ist die Pseudoinverse eine Verallgemeinerung der klassischen Inversen für singuläre oder nicht quadratische Matrizen. Dabei ist vor allem Bedingung (i) interessant, denn sie garantiert, daß $A^\dagger b$ für jedes $b \in \mathcal{R}(A)$ ein Urbild von b ist: gilt nämlich $b = Ax$ für ein $x \in \mathbb{K}^n$, so folgt

$$A(A^\dagger b) = AA^\dagger Ax = Ax = b\,.$$

Andererseits ergibt sich aus (12.10) $A^\dagger b = 0$ für $b \in \mathcal{R}(A)^\perp$.

Interessant ist auch die folgende Verallgemeinerung der Eigenschaften $AA^{-1} = I = A^{-1}A$ der klassischen Inversen:

Korollar 12.5. *$AA^\dagger$ ist der Orthogonalprojektor auf $\mathcal{R}(A)$ und $A^\dagger A$ der Orthogonalprojektor auf $\mathcal{N}(A)^\perp$.*

Beweis. Wir beschränken uns darauf, die Eigenschaften von $P = AA^\dagger$ nachzuweisen. P ist genau dann ein Orthogonalprojektor, falls $P^2 = P$ und $P = P^*$ gilt, und diese beiden Eigenschaften folgen unmittelbar aus (i) – nach Multiplikation mit $X = A^\dagger$ von rechts – und aus (iii). Somit ist $AA^\dagger$ ein Orthogonalprojektor und nach (12.9) gilt $\mathcal{R}(AA^\dagger) = \mathcal{R}(A)$. □

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

Satz 12.6. *Der Vektor $A^\dagger b$ ist die eindeutig bestimmte Lösung des linearen Ausgleichsproblems* (11.2) *mit minimaler Euklidnorm.*

Beweis. Nach Satz 12.4 (ii) ist

$$AA^\dagger b - b \,\in\, \mathcal{N}(A^\dagger) \overset{(12.6)}{=} \mathcal{R}(A)^\perp \,=\, \mathcal{N}(A^*)\,.$$

Also erfüllt $A^\dagger b$ die Normalengleichung (11.1) und ist daher eine Lösung des linearen Ausgleichsproblems.

Ist z eine zweite Lösung von (11.1), dann ist

$$w \,=\, A^\dagger b - z \,\in\, \mathcal{N}(A^*A) \,= \mathcal{N}(A)\,.$$

Andererseits ist $A^\dagger b \in \mathcal{R}(A^\dagger)$ und dieser Bildraum stimmt nach (12.6) mit $\mathcal{N}(A)^\perp$ überein. Folglich haben wir eine orthogonale Zerlegung von $z = A^\dagger b - w$, und nach dem Satz von Pythagoras gilt

$$\|z\|_2^2 \,=\, \|A^\dagger b\|_2^2 + \|w\|_2^2 \,\geq\, \|A^\dagger b\|_2^2$$

mit Gleichheit genau für $w = 0$, d. h. für $z = A^\dagger b$. □

Für Matrizen $A \in \mathbb{K}^{m\times n}$ mit Rang n bedeutet dies:

Korollar 12.7. *Ist* $\mathcal{N}(A) = \{0\}$*, dann gilt* $A^\dagger = (A^*A)^{-1}A^*$.

Beispiel. Zu lösen sei das lineare Ausgleichsproblem mit

$$A = \begin{bmatrix} 1 & 1 \\ 1 & 1 \\ 1 & 1 \\ 1 & 1 \end{bmatrix} \qquad \text{und} \qquad b = \begin{bmatrix} 2 \\ 0 \\ 0 \\ -1 \end{bmatrix}.$$

Die Lösung $\widehat{x}$ mit minimaler Euklidnorm lautet $\widehat{x} = A^\dagger b$, wobei die Pseudoinverse $A^\dagger$ von A in Beispiel 12.3 bestimmt wurde:

$$\widehat{x} \;=\; A^\dagger \begin{bmatrix} 2 \\ 0 \\ 0 \\ -1 \end{bmatrix} \;=\; \frac{1}{8}\begin{bmatrix} 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 \end{bmatrix} \begin{bmatrix} 2 \\ 0 \\ 0 \\ -1 \end{bmatrix} \;=\; \frac{1}{8}\begin{bmatrix} 1 \\ 1 \end{bmatrix}.$$

Damit ist $A\widehat{x} = AA^\dagger b$ die Projektion von b auf $\mathcal{R}(A)$,

$$A\widehat{x} \;=\; \frac{1}{8}\begin{bmatrix} 2 \\ 2 \\ 2 \\ 2 \end{bmatrix} \;=\; \begin{bmatrix} 1/4 \\ 1/4 \\ 1/4 \\ 1/4 \end{bmatrix}.$$

Alle anderen Lösungen des linearen Ausgleichsproblems unterscheiden sich von $\widehat{x}$ durch ein additives Element aus $\mathcal{N}(A^*A) = \mathcal{N}(A)$, d. h. jede Lösung x von (11.2) hat die Gestalt

$$x \;=\; \frac{1}{8}\begin{bmatrix} 1 \\ 1 \end{bmatrix} + \xi \begin{bmatrix} 1 \\ -1 \end{bmatrix}, \qquad \xi \in \mathbb{R}\,,$$

so daß

$$\|x\|_2^2 \;=\; \|\widehat{x}\|_2^2 + 2\xi^2\,.$$

Wie man hieran sieht, ist $\widehat{x}$ die Lösung des linearen Ausgleichsproblems mit minimaler euklidischer Norm. $\lozenge$

13 Die QR-Zerlegung

Prinzipiell könnten die Ergebnisse aus den Abschnitten 11 und 12 auch in numerische Algorithmen zur Lösung des linearen Ausgleichsproblems umgesetzt werden. Beispielsweise könnte die Singulärwertzerlegung $A = U\Sigma V^*$ numerisch berechnet werden, um dann die Lösung $A^\dagger b$ durch Auswertung der Summendarstellung

$$A^\dagger b \;=\; \sum_{i=1}^{p} \frac{u_i^* b}{\sigma_i}\, v_i \,,$$

vgl. (12.5), zu berechnen. Dies ist allerdings sehr aufwendig, so daß die Singulärwertzerlegung hauptsächlich für theoretische Untersuchungen herangezogen wird.

Alternativ könnten die Normalengleichungen über eine Cholesky-Zerlegung der Koeffizientenmatrix A^*A gelöst werden, was jedoch ein klassisches Beispiel für einen numerischen Algorithmus ist, der schlechter konditioniert ist als das eigentlich zu lösende Problem. Man sieht das am einfachsten in dem Fall, in dem $A \in \mathbb{K}^{n\times n}$ invertierbar ist. Dann ergibt sich nämlich für die Kondition des Gleichungssystems $Ax = b$ in der Euklidnorm

$$\mathrm{cond}_2(A) \;=\; \|A\|_2 \|A^{-1}\|_2 \;=\; (\lambda_1/\lambda_n)^{1/2},$$

während die Kondition der Normalengleichungen durch

$$\mathrm{cond}_2(A^*A) \;=\; \|A^*A\|_2 \|(A^*A)^{-1}\|_2 \;=\; \lambda_1/\lambda_n \;=\; \big(\mathrm{cond}_2(A)\big)^2$$

gegeben ist; $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_n$ sind dabei wieder die absteigend sortierten Eigenwerte von A^*A. Da λ_1/λ_n größer als Eins ist, sind die Normalengleichungen also deutlich schlechter konditioniert als das lineare Ausgleichsproblem an sich.

Für rechteckige oder singuläre Matrizen A verallgemeinert man die (euklidische) Kondition durch

$$\mathrm{cond}_2(A) \;=\; \|A\|_2 \|A^\dagger\|_2 \;=\; \sigma_1/\sigma_p \;=\; (\lambda_1/\lambda_p)^{1/2}\,,$$

wobei λ_p wieder den kleinsten von Null verschiedenen Eigenwert von A^*A bezeichnet.

Wir beschreiben nun einen Algorithmus zur Lösung des linearen Ausgleichsproblems, der die Normalengleichungen vermeidet, beschränken uns allerdings auf den Fall $\mathrm{Rang}\, A = n \leq m$.

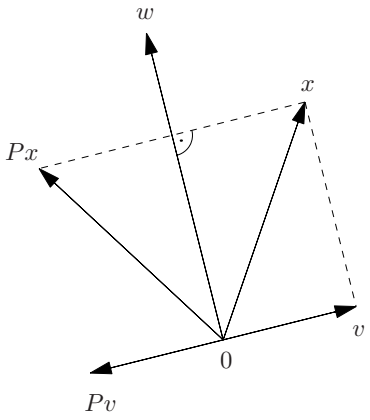


Abb. 13.1:
Householder-Transformationen sind Spiegelungen

Definition 13.1. Eine Matrix der Gestalt

$$P = I - \frac{2}{v^*v}\, vv^* \in \mathbb{K}^{r\times r} \qquad \text{mit } v \in \mathbb{K}^r \setminus \{0\}$$

wird *Householder-Transformation* genannt.

Lemma 13.2. *Die Householder-Transformation P aus Definition* 13.1 *ist eine hermitesche, unitäre Matrix mit*

$$Pv = -v \qquad \textit{und} \qquad Pw = w \quad \textit{für alle } w \in \{v\}^\perp .$$

Beweis. Aus der Definition von P folgt unmittelbar, daß P hermitesch ist. Außerdem ist P unitär, denn

$$P^*P \;=\; P^2 \;=\; I - \frac{4}{v^*v}\, vv^* + \frac{4}{(v^*v)^2}\, v(v^*v)v^* \;=\; I - \frac{4}{v^*v}\, vv^* + \frac{4}{v^*v}\, vv^* \;=\; I\,.$$

Schließlich ergibt sich für den Vektor v aus der Definition von P und für ein beliebiges $w \perp v$

$$\begin{aligned} Pv &= Iv - \frac{2}{v^*v}\, v(v^*v) \;=\; v - 2v \;=\; -v\,, \\ Pw &= Iw - \frac{2}{v^*v}\, v(v^*w) \;=\; w - 0 \;=\; w\,. \end{aligned}$$

□

Abbildung 13.1 illustriert die Aussage von Lemma 13.2: Demnach ist eine Householder-Transformation nichts anderes als die Abbildungsmatrix einer geometrischen Spiegelung. Man beachte, daß diese Spiegelungen die Euklidnorm invariant lassen, denn es ist

$$\|Px\|_2^2 \;=\; (Px)^*Px \;=\; x^* \underbrace{P^*P}_{I} x \;=\; x^*x \;=\; \|x\|_2^2\,. \tag{13.1}$$

Wir verwenden nun Householder-Transformationen, um – ähnlich zu den Eliminationsmatrizen L_k in Abschnitt 4 – die Matrix A auf „obere Dreiecksgestalt“ zu transformieren (wie das bei einer rechteckigen Matrix zu verstehen ist, werden wir gleich erläutern).

Um dies zu bewerkstelligen, konstruieren wir zunächst eine Householder-Transformation P, die einen beliebig vorgegebenen Vektor $x \in \mathbb{K}^r \setminus \{0\}$ auf ein Vielfaches von $e_1 \in \mathbb{K}^r$ spiegelt, d. h.

$$Px \;=\; x - \frac{2}{v^*v}\, v(v^*x) \;\stackrel{!}{=}\; \xi e_1\,, \qquad |\xi| \;=\; \|x\|_2\,,$$

vgl. (13.1). Aus Abbildung 13.2 wird deutlich, daß v ein Vielfaches von $x - \xi e_1$ sein muß; damit bei dieser Subtraktion keine Auslöschung auftritt, wählen wir

$$\xi \;=\; \begin{cases} -\dfrac{x_1}{|x_1|}\,\|x\|_2 & \text{für } x_1 \neq 0\,, \\ -\|x\|_2 & \text{für } x_1 = 0\,, \end{cases}$$

wobei x_1 die erste Komponente von x bezeichnet. Mit geeigneter Normierung (P ist unabhängig von der Norm von v) ergibt dies

$$v \;=\; \frac{1}{\|x\|_2}\left(x + \frac{x_1\|x\|_2}{|x_1|}\,e_1\right) \;=\; \frac{1}{|x_1|\,\|x\|_2}\left(|x_1|x + x_1\|x\|_2 e_1\right) \tag{13.2a}$$

für $x_1 \neq 0$ beziehungsweise

$$v = x/\|x\|_2 + e_1 \qquad \text{für } x_1 = 0\,. \tag{13.2b}$$

In jedem Fall ist $v^*x = \|x\|_2 + |x_1|$ und

$$v^*v \;=\; 2 + 2\,|x_1|/\|x\|_2\,. \tag{13.3}$$

Mit diesem v ergibt sich in der Tat für $x_1 \neq 0$

$$\begin{aligned} Px \;&=\; x - \frac{2}{v^*v}\,v(v^*x) \;=\; x - 2\,\frac{\|x\|_2 + |x_1|}{2 + 2\,|x_1|/\|x\|_2}\,\frac{|x_1|\,x + x_1\|x\|_2 e_1}{|x_1|\,\|x\|_2} \\ &=\; x - x - \frac{x_1}{|x_1|}\,\|x\|_2 e_1 \;=\; -\frac{x_1}{|x_1|}\,\|x\|_2 e_1 \end{aligned}$$

und entsprechend $Px = x - \|x\|_2 v = -\|x\|_2 e_1$ für $x_1 = 0$.

Definition und Satz 13.3. *Sei $A \in \mathbb{K}^{m\times n}$ mit $m \geq n$ und* $\operatorname{Rang} A = n$. *Dann existiert eine unitäre Matrix $Q \in \mathbb{K}^{m\times m}$ und eine „rechte obere Drei-*

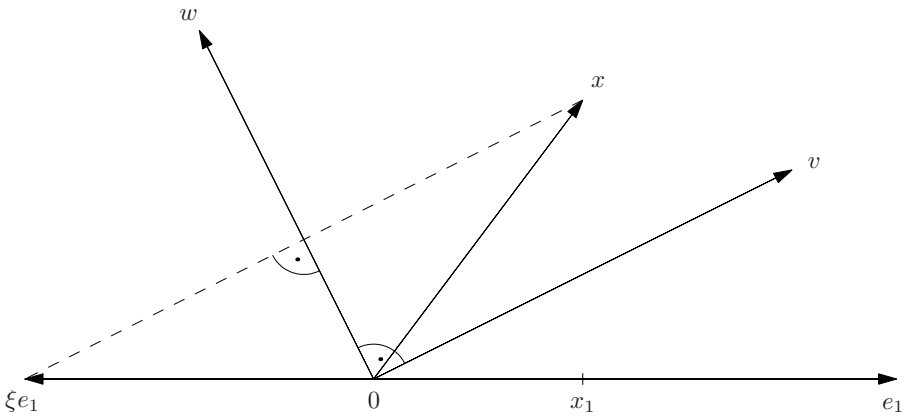


Abb. 13.2: Spiegelung von x auf ein geeignetes Vielfaches von e_1

ecksmatrix“

$$R = \left[\begin{array}{ccc} r_{11} & \cdots & r_{1n} \\ & \ddots & \vdots \\ 0 & & r_{nn} \\ \hdashline 0 & \dots & 0 \\ & & \end{array}\right] \in \mathbb{K}^{m\times n}$$

*mit $A = QR$. Dabei sind $r_{11}, \dots, r_{nn}$ jeweils von Null verschieden. Eine solche Faktorisierung von A wird QR-*Zerlegung *genannt.*

Beweis. Wir bestimmen die gesuchte Faktorisierung, indem wir A in jedem Schritt von links mit einer Householder-Transformation multiplizieren, um sukzessive die Spalten von R zu erhalten. Dies ergibt dann eine Darstellung

$$P_n \cdots P_1 A = R \tag{13.4}$$

mit Householder-Transformationen P_i, und hieraus folgt die QR-Faktorisierung

$$A = QR \qquad \text{mit} \qquad Q = P_1^* \cdots P_n^* = P_1 \cdots P_n \,.$$

Im ersten Schritt setzen wir $A_1 = A$ und für x die erste Spalte a_1 von A_1 und bestimmen die Householder-Transformation $P_1 \in \mathbb{K}^{m\times m}$ mit v aus (13.2). Es folgt

$$P_1 a_1 = r_{11} e_1 \qquad \text{mit} \qquad |r_{11}| = \|a_1\|_2 \neq 0$$

beziehungsweise

$$P_1 A = \left[\begin{array}{c|c} r_{11} & \cdots \\ \hline 0 & A_2 \end{array}\right] \qquad \text{mit } A_2 \in \mathbb{K}^{(m-1)\times(n-1)}.$$

Nehmen wir nun an, daß wir nach i Schritten Householder-Transformationen $P_1, \ldots, P_i$ konstruiert haben mit

$$P_i \cdots P_1 A = \left[\begin{array}{ccc|c} r_{11} & \cdots & r_{1i} & \\ & \ddots & \vdots & R_i' \\ 0 & & r_{ii} & \\ \hline 0 & \cdots & 0 & A_{i+1} \\ & & & \end{array}\right] \tag{13.5}$$

mit $R_i' \in \mathbb{K}^{i\times(n-i)}$ und $A_{i+1} \in \mathbb{K}^{(m-i)\times(n-i)}$. Da A nach Voraussetzung vollen Spaltenrang hat, trifft dies wegen der gegebenen Blockstruktur in (13.5) auch auf A_{i+1} zu. Im nächsten Schritt können wir daher für $x \in \mathbb{K}^{m-i}$ die erste Spalte $a_{i+1} \neq 0$ von A_{i+1} wählen und konstruieren die Householder-Transformation $P_i' \in \mathbb{K}^{(m-i)\times(m-i)}$ mit dem Vektor $v' \in \mathbb{K}^{m-i}$ aus (13.2). Auf diese Weise ergibt sich

$$P_{i+1}' A_{i+1} = \left[\begin{array}{c|c} r_{i+1,i+1} & \cdots \\ \hline 0 & A_{i+2} \end{array}\right]$$

mit $|r_{i+1,i+1}| = \|a_{i+1}\|_2 \neq 0$ und $A_{i+2} \in \mathbb{K}^{(m-i-1)\times(n-i-1)}$, und es folgt

$$\underbrace{\left[\begin{array}{c|c} I & 0 \\ \hline 0 & P_{i+1}' \end{array}\right]}_{P_{i+1}} P_i \cdots P_1 A = \left[\begin{array}{ccc|cc} r_{11} & \cdots & r_{1i} & & \\ & \ddots & \vdots & & R_i' \\ 0 & & r_{ii} & & \\ \hline 0 & \cdots & 0 & r_{i+1,i+1} & \cdots \\ \hline 0 & \cdots & 0 & 0 & A_{i+2} \end{array}\right].$$

Man beachte, daß sich in diesem Schritt die ersten i Zeilen *nicht* verändern. P_{i+1} kann selbst wieder als Householder-Transformation mit einem Vektor $v \in \mathbb{K}^m$ der Form $v^T = [0, v'^T]$ aufgefaßt werden. Durch vollständige Induktion erhalten wir nun die gewünschte Zerlegung (13.4). □

Initialisierung: $A_1 = A \in \mathbb{K}^{m\times n}$ habe vollen Rang n

`for` $i = 1, \ldots, n$ `do` % erzeuge Zerlegung (13.5)

% bezeichne mit a_i die erste Spalte von A_i und mit a_{i1} deren erste Komponente

$v = a_i/\|a_i\|_2 + a_{i1}\, e_1/|a_{i1}|$ % vgl. (13.2); $a_{i1}/|a_{i1}| := 1$ für $a_{i1} = 0$

$\beta = 2/v^*v = \left(1 + |a_{i1}|/\|a_i\|_2\right)^{-1}$ % vgl. (13.3)

$w = A_i^* v$

$A_i = A_i - \beta v w^*$

% falls der Householder-Vektor v abgespeichert werden soll:

$v = v/v_1$ % skaliere v so, daß die erste Komponente Eins ist

überschreibe untere $m - i$ Komponenten von a_i mit v

% A_{i+1} bezeichnet im folgenden Schleifendurchlauf den rechten unteren

% $(m - i) \times (n - i)$-Block von A_i, vgl. (13.5)

`end for`

Ergebnis: Der obere Dreiecksanteil der $m \times n$-Matrix enthält R; unterhalb von R stehen die hinteren Komponenten der einzelnen Householder-Vektoren

Algorithmus 13.1: QR-Zerlegung

Bei der Implementierung ist darauf zu achten, daß die Householder-Transformationsmatrizen *niemals* explizit gebildet werden, denn eine Matrix-Matrix-Multiplikation PA mit einer Householder-Transformation P kostet $O(m^2 n)$ Multiplikationen. Statt dessen kann PA mit nur $2mn$ Multiplikationen über die Darstellung

$$PA \;=\; A - \frac{2}{v^*v}\, vv^*A \;=\; A - \frac{2}{v^*v} vw^*\,, \qquad w = A^*v\,,$$

ausgerechnet werden. Um später auf P wieder zugreifen zu können, empfiehlt es sich, den Vektor v abzuspeichern. Hierzu können die neu erzeugten Nulleinträge von PA genutzt werden: Dazu muß allerdings v zunächst so umskaliert werden, daß die erste Komponente eine Eins (und somit redundant) ist; die restlichen Einträge werden dann in der entsprechenden Spalte von PA abgespeichert.

Aufwand. Algorithmus 13.1 faßt die einzelnen Schritte der QR-Zerlegung zusammen. Es erweist sich dabei als hilfreich, den Faktor $\beta = 2/v^*v$ als Zwischenresultat abzuspeichern. Beim i-ten Schleifendurchlauf des Algorithmus schlagen sich dann die einzelnen Teilschritte wie folgt zu Buche:

(13.2):	$2(m-i+1)$	Multiplikationen/Divisionen
β:	2	Multiplikationen/Divisionen
w:	$(n-i+1)(m-i+1)$	Multiplikationen/Divisionen
A_{i+1}:	$(n-i)(m-i)$	Multiplikationen/Divisionen
$v = v/v_1$:	$m-i$	Multiplikationen/Divisionen
$\sum \approx$	$2(n-i)(m-i)$	Multiplikationen/Divisionen

Initialisierung: $A \in \mathbb{K}^{m\times n}$ habe vollen Rang n und $b \in \mathbb{K}^m$ sei der Vektor in (11.2)
faktorisiere $A = QR \ldots$ % mit Algorithmus 13.1
$\ldots$ und berechne gleichzeitig $c = Q^*b = P_n \cdots P_1 b$
löse $R_1 x = c_1$ durch Rückwärtssubstitution % R_1 und c_1 wie in (13.6)
Ergebnis: x löst das lineare Ausgleichsproblem (11.2)

Algorithmus 13.2: Lösung des linearen Ausgleichsproblems

Damit ergibt sich insgesamt im wesentlichen ein Aufwand von

$$2\sum_{i=1}^{n-1}(n-i)(m-i) \;=\; 2\sum_{i=1}^{n-1} i(m-n+i) \;=\; mn^2 - \frac{1}{3}n^3 + O(mn)$$

Multiplikationen/Divisionen. ◊

Die QR-Zerlegung $A = QR$ kann natürlich auch alternativ zur LR-Zerlegung aus Abschnitt 4 zur Lösung eines nichtsingulären Gleichungssystems $Ax = b$ verwendet werden: In diesem Fall löst man $QRx = b$ durch Rückwärtssubstitution aus der Gleichung $Rx = Q^*b$. (Die rechte Seite Q^*b kann wieder mit $O(n^2)$ Multiplikationen berechnet werden, indem man jede einzelne Householder-Transformation P_i an den Vektor b heranmultipliziert). Dieses Verfahren ist allerdings etwa doppelt so teuer wie die Gauß-Elimination.

Entscheidender ist die Anwendbarkeit der QR-Zerlegung auf das lineare Ausgleichsproblem Durch Einsetzen von $A = QR$ ergibt sich nämlich

$$\|b - Ax\|_2 \overset{(13.1)}{=} \|Q^*(b - Ax)\|_2 \;=\; \|Q^*b - Q^*QRx\|_2 \;=\; \|c - Rx\|_2$$

mit $c = Q^*b$. Zerlegen wir R und c konform in

$$R = \begin{bmatrix} R_1 \\ 0 \end{bmatrix}, \quad c = \begin{bmatrix} c_1 \\ c_2 \end{bmatrix}, \qquad R_1 \in \mathbb{K}^{n\times n},\ c_1 \in \mathbb{K}^n, \tag{13.6}$$

dann ist nach dem Satz von Pythagoras

$$\|b - Ax\|_2^2 \;=\; \|c - Rx\|_2^2 \;=\; \|c_1 - R_1x\|_2^2 + \|c_2\|_2^2 \;\geq\; \|c_2\|_2^2$$

mit Gleichheit genau für $x = R_1^{-1}c_1$ (R_1 ist eine obere Dreiecksmatrix mit nichtverschwindenden Diagonalelementen, also ist R_1 invertierbar). Damit ist gezeigt, daß $\widehat{x} = R_1^{-1}c_1$ die Lösung des linearen Ausgleichsproblems ist, vgl. Algorithmus 13.2.

Beispiel 13.4. Wir berechnen die Ausgleichsgerade aus Beispiel 11.4 mit Hilfe der QR-Zerlegung, d. h. wir lösen das lineare Ausgleichsproblem mit

$$A = \begin{bmatrix} 1 & 0 \\ 1 & 3 \\ 1 & 4 \\ 1 & 7 \end{bmatrix} \qquad \text{und} \qquad b = \begin{bmatrix} 1 \\ 2 \\ 6 \\ 4 \end{bmatrix}.$$

Im ersten Schritt von Algorithmus 13.1 ist $a_1 = [1,1,1,1]^T$, also

$$\|a_1\|_2 = 2\,, \qquad v = [3/2, 1/2, 1/2, 1/2]^T \qquad \text{und} \qquad \beta = 2/3\,.$$

Ferner ist

$$w = A^* v = \begin{bmatrix} 1 & 1 & 1 & 1 \\ 0 & 3 & 4 & 7 \end{bmatrix} \begin{bmatrix} 3/2 \\ 1/2 \\ 1/2 \\ 1/2 \end{bmatrix} = \begin{bmatrix} 3 \\ 7 \end{bmatrix},$$

und daher ergibt sich

$$P_1 A = A - \beta v w^* = \begin{bmatrix} 1 & 0 \\ 1 & 3 \\ 1 & 4 \\ 1 & 7 \end{bmatrix} - \frac{2}{3} \begin{bmatrix} 3/2 \\ 1/2 \\ 1/2 \\ 1/2 \end{bmatrix} \begin{bmatrix} 3 & 7 \end{bmatrix} = \begin{bmatrix} -2 & -7 \\ 0 & 2/3 \\ 0 & 5/3 \\ 0 & 14/3 \end{bmatrix}.$$

Daneben berechnen wir

$$v^* b \;=\; \begin{bmatrix} 3/2 & 1/2 & 1/2 & 1/2 \end{bmatrix} \begin{bmatrix} 1 \\ 2 \\ 6 \\ 4 \end{bmatrix} \;=\; 15/2$$

und

$$P_1 b = b - \beta (v^* b) v = b - 5v = \begin{bmatrix} 1 \\ 2 \\ 6 \\ 4 \end{bmatrix} - 5 \begin{bmatrix} 3/2 \\ 1/2 \\ 1/2 \\ 1/2 \end{bmatrix} = \begin{bmatrix} -13/2 \\ -1/2 \\ 7/2 \\ 3/2 \end{bmatrix}.$$

Somit wird der zweite Schritt von Algorithmus 13.1 auf die Restmatrix $A_2 = a_2 = [2/3, 5/3, 14/3]^T$ und die unteren drei Einträge $b_2 = [-1/2, 7/2, 3/2]^T$ von $P_1 b$ angewandt. Wegen $\|a_2\|_2 = 5$ führt dies auf

$$v = \begin{bmatrix} 17/15 \\ 1/3 \\ 14/15 \end{bmatrix}, \qquad \beta = 15/17\,,$$

und

$$w = A_2^* v = \begin{bmatrix} 2/3 & 5/3 & 14/3 \end{bmatrix} \begin{bmatrix} 17/15 \\ 1/3 \\ 14/15 \end{bmatrix} = \frac{17}{3},$$

$$\beta v^* b_2 = \frac{15}{17} \begin{bmatrix} 17/15 & 1/3 & 14/15 \end{bmatrix} \begin{bmatrix} -1/2 \\ 7/2 \\ 3/2 \end{bmatrix} = \frac{30}{17}.$$

Demnach ist

$$P_2 A_2 = A_2 - \beta v w^* = \begin{bmatrix} 2/3 \\ 5/3 \\ 14/3 \end{bmatrix} - 5 \begin{bmatrix} 17/15 \\ 1/3 \\ 14/15 \end{bmatrix} = \begin{bmatrix} -5 \\ 0 \\ 0 \end{bmatrix},$$

$$P_2 b_2 = b_2 - \beta (v^* b_2) v = \begin{bmatrix} -1/2 \\ 7/2 \\ 3/2 \end{bmatrix} - \frac{30}{17} \begin{bmatrix} 17/15 \\ 1/3 \\ 14/15 \end{bmatrix} = \begin{bmatrix} -5/2 \\ 99/34 \\ -5/34 \end{bmatrix}.$$

Damit ist die QR-Zerlegung von A abgeschlossen. Das Ergebnis

$$\begin{bmatrix} -2 & -7 \\ 1/3 & -5 \\ 1/3 & 5/17 \\ 1/3 & 14/17 \end{bmatrix}$$

enthält im oberen rechten Dreiecksteil die von Null verschiedenen Einträge der Matrix R_1 aus (13.6) und im dunkel hinterlegten Rest die hinteren Komponenten der umskalierten Householder-Vektoren aus den einzelnen Reduktionsschritten. Ferner haben wir

$$c = Q^* b = \left[\begin{array}{c} -13/2 \\ -5/2 \\ \hdashline 99/34 \\ -5/34 \end{array}\right] = \left[\begin{array}{c} c_1 \\ \hdashline c_2 \end{array}\right].$$

Die Lösung $\widehat{x} = [x_1, x_2]^T$ des linearen Ausgleichsproblems ist die Lösung des oberen Dreieckssystems

$$\begin{bmatrix} -2 & -7 \\ 0 & -5 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} -13/2 \\ -5/2 \end{bmatrix},$$

also $\widehat{x} = [3/2, 1/2]^T$. ◊

Bemerkung 13.5. Wenn A keinen vollen Spaltenrang hat, kann Algorithmus 13.1 mit $a_i = 0$ in einem Schleifendurchlauf vorzeitig zusammenbrechen. Um dies zu vermeiden, müssen bei der QR-Zerlegung vor jedem Teilschritt die Spalten der Submatrix A_{i+1} aus (13.5) derart permutiert werden (ähnlich zur Pivotsuche bei der Gauß-Elimination), daß die Euklidnorm der jeweils ersten Spalte maximal wird. In Analogie zur Gauß-Elimination wird dies *Spaltenpivotsuche* genannt (vgl. Abschnitt 4.1). Mit dieser Spaltenpivotsuche bricht Algorithmus 13.1 erst dann zusammen, wenn die gesamte Restmatrix A_{i+1} aus (13.5) die Nullmatrix ist. In dem Moment hat man eine Faktorisierung

$$Q^* A \Pi = \begin{bmatrix} R_1 & R_2 \\ 0 & 0 \end{bmatrix} \tag{13.7}$$

berechnet, wobei $\Pi \in \mathbb{K}^{n\times n}$ eine Permutationsmatrix, $R_1 \in \mathbb{K}^{p\times p}$ eine rechte obere Dreiecksmatrix und R_2 eine (in der Regel voll besetzte) $p \times (n-p)$-Matrix ist. Diese Faktorisierung reicht allerdings nur aus, um *eine* Lösung des linearen Ausgleichsproblems zu bestimmen, im allgemeinen ist dies nicht die Lösung minimaler Norm. ◊

Die QR-Zerlegung gehört zu den stabilsten Algorithmen in der numerischen linearen Algebra. Der Grund liegt darin, daß unitäre Transformationen wegen $\mathrm{cond}_2(Q) = 1$ keinerlei Fehlerverstärkung hervorrufen. Die abschließende Rückwärtssubstitution hat die gleiche Kondition wie das Ausgangsproblem, denn es gilt

$$A^*A \;=\; (QR)^*QR \;=\; R^*Q^*QR \;=\; R^*R\,, \tag{13.8}$$

und folglich ist

$$\mathrm{cond}_2(R) = \mathrm{cond}_2(A)\,.$$

Wegen (13.8) ist R^* zudem ein „Cholesky-artiger“ Faktor von A^*A, im allgemeinen jedoch nicht der Cholesky-Faktor selbst, da die Diagonalelemente von R nicht positiv sein müssen. Man könnte jedoch eine QR-Zerlegung von A so bestimmen, daß R positive Diagonalelemente hat (man überlege sich im Beweis von Satz 13.3 die notwendigen Änderungen; vgl. auch Aufgabe 8).

14 Givens-Rotationen

Für manche Matrizen $A \in \mathbb{K}^{m\times n}$ kann die QR-Zerlegung etwas billiger berechnet werden, wenn anstelle von Householder-Transformationen andere unitäre Transformationen verwendet werden. Wir beschreiben in diesem Abschnitt eine Möglichkeit für sogenannte Hessenberg-Matrizen.

Definition 14.1. Eine Matrix $H = [h_{ij}] \in \mathbb{K}^{m\times n}$ mit $m \geq n$ hat *obere Hessenberg-Form*, wenn H die Gestalt

$$H = \left[\begin{array}{ccccc} h_{11} & h_{12} & h_{13} & \cdots & h_{1n} \\ h_{21} & h_{22} & h_{23} & \cdots & h_{2n} \\ 0 & h_{32} & h_{33} & & h_{3n} \\ & & \ddots & \ddots & \vdots \\ \vdots & & & h_{n,n-1} & h_{nn} \\ 0 & \cdots & & 0 & h_{n+1,n} \\ \hdashline & & & & \\ & & 0 & & \\ & & & & \end{array}\right]$$

besitzt, d. h. wenn $h_{ij} = 0$ ist für alle Indizes i und j mit $j < i - 1$.

Hessenberg-Matrizen und ihre QR-Zerlegung werden uns im Verlauf dieses Buches noch zweimal begegnen (in den Abschnitten 16 und 27). Hier betrachten wir eine erste Anwendung.

Beispiel 14.2. Ein lineares Ausgleichsproblem (11.2) soll um eine neue Gleichung $a^*x = \beta$ mit $a \in \mathbb{K}^n$ und $\beta \in \mathbb{K}$ erweitert werden. Dieses erweiterte Ausgleichsproblem kann mit geringem Aufwand gelöst werden, wenn die Lösung des ursprünglichen Problems über eine QR-Zerlegung $A = QR$ berechnet wurde. Mit der Notation aus (13.6) ergibt sich nämlich

$$\left[\begin{array}{c:c} I & \\ \hdashline & Q^* \end{array}\right] \left[\begin{array}{c} a^* \\ \hdashline A \end{array}\right] = \left[\begin{array}{c} a^* \\ \hdashline R_1 \\ 0 \end{array}\right], \qquad \left[\begin{array}{c:c} I & \\ \hdashline & Q^* \end{array}\right] \left[\begin{array}{c} \beta \\ \hdashline b \end{array}\right] = \left[\begin{array}{c} \beta \\ \hdashline c_1 \\ c_2 \end{array}\right],$$

und aufgrund der Invarianz der Euklidnorm unter unitären Transformationen ist die Lösung x^+ des erweiterten $(m + 1) \times n$-dimensionalen linearen Ausgleichsproblems daher auch eine Lösung des $(n + 1) \times n$-dimensionalen Ausgleichsproblems

$$\text{minimiere} \quad \left\| \begin{bmatrix} \beta \\ c_1 \end{bmatrix} - \begin{bmatrix} a^* \\ R_1 \end{bmatrix} x \right\|_2, \tag{14.1}$$

dessen Koeffizientenmatrix obere Hessenberg-Form besitzt. ◊

Für die QR-Zerlegung einer Hessenberg-Matrix werden sogenannte Givens-Rotationen verwendet.

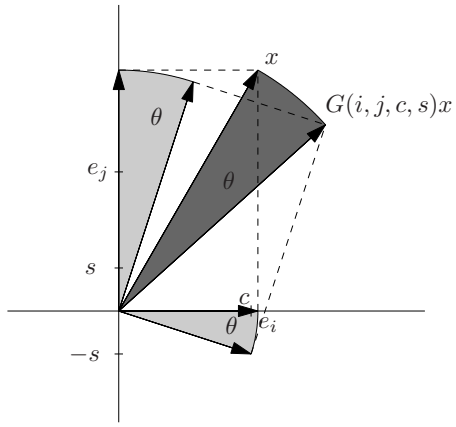


Abb. 14.1:
Geometrische Interpretation einer reellen Givens-Rotation

Definition 14.3. Sei $\theta \in (-\pi, \pi]$, $\alpha, \beta \in [0, 2\pi)$ und $c = e^{\mathrm{i}\alpha} \cos\theta$, $s = e^{\mathrm{i}\beta} \sin\theta$. Die Matrix

$$G(i,j,c,s) = \begin{bmatrix} 1 & & & & & & \\ & \ddots & & & & & \\ & & 1 & & & & \\ & & & \overline{c} & & \overline{s} & \\ & & & & 1 & & \\ & & & & & \ddots & \\ & & & & & & 1 \\ & & & -s & & c & \\ & & & & & & & 1 \\ & & & & & & & & \ddots \\ & & & & & & & & & 1 \end{bmatrix} \in \mathbb{K}^{m \times m} \qquad (14.2)$$

wird (komplexe) *Givens-Rotation* genannt (die gebrochenen Linien grenzen den Bereich von der i-ten bis zur j-ten Zeile bzw. Spalte ab).

Givens-Rotationen sind unitäre Matrizen, wobei vornehmlich der reelle Fall mit $\alpha = \beta = 0$ von Bedeutung ist. In diesem Fall entspricht die Abbildung (14.2) einer Rotation um den Winkel θ in der von e_i und e_j aufgespannten kartesischen Ebene, vgl. Abbildung 14.1.

Bemerkung 14.4. Bei der Operation $A \mapsto GA$ mit $G = G(i,j,c,s)$ aus (14.2) werden die Zeilen i und j von A (a_i^* bzw. a_j^*) durch Linearkombinationen $\bar{c}a_i^* + \bar{s}a_j^*$ bzw. $-sa_i^* + ca_j^*$ ersetzt, während bei einer Multiplikation von rechts ($A \mapsto AG$) die Spalten i und j von A durch entsprechende Linearkombinationen ersetzt werden. ◊

Die Givens-Rotation G kann so konstruiert werden, daß nach der Transformation $A \mapsto GA$ das (j,k)-Element von GA Null ist. Dies entspricht einer Rotation

des Anteils der k-ten Spalte von A in der (i,j)-Ebene, bei der die j-te Komponente Null wird. Algebraisch führt dies auf die Forderung $-sa_{ik} + ca_{jk} = 0$ unter der Nebenbedingung $|c|^2 + |s|^2 = 1$. Dieses Problem hat zwei Lösungen, die sich nur durch das Vorzeichen unterscheiden; eine Lösung lautet

$$c = \frac{a_{ik}}{(|a_{ik}|^2 + |a_{jk}|^2)^{1/2}}\,, \qquad s = \frac{a_{jk}}{(|a_{ik}|^2 + |a_{jk}|^2)^{1/2}}\,.$$

Um bei der Berechnung von s und c Overflow zu vermeiden, werden üblicherweise die mathematisch äquivalenten Darstellungen

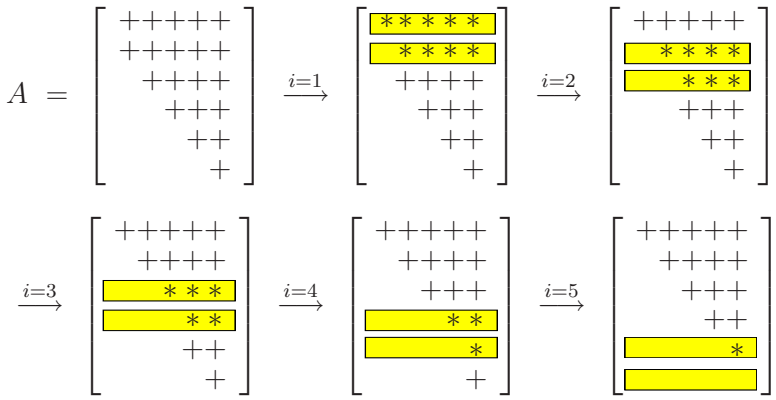
$$\begin{aligned} c &= \frac{a_{ik}/|a_{ik}|}{(1+|t|^2)^{1/2}}\,, \quad s = \frac{t}{(1+|t|^2)^{1/2}}\,, \quad t = a_{jk}/|a_{ik}|\,, \qquad \text{für } |a_{ik}| \geq |a_{jk}|\,, \\ c &= \frac{t}{(1+|t|^2)^{1/2}}\,, \quad s = \frac{a_{jk}/|a_{jk}|}{(1+|t|^2)^{1/2}}\,, \quad t = a_{ik}/|a_{jk}|\,, \qquad \text{für } |a_{ik}| < |a_{jk}|\,, \end{aligned}$$

verwendet.

Für die QR-Zerlegung einer Matrix A in oberer Hessenberg-Form werden nun sukzessive Givens-Rotationen $G(i, i+1, c_i, s_i)$, $i = 1, \ldots, n$, von links an A heranmultipliziert, um jeweils das $(i+1, i)$-Element auf Null zu rotieren:

$$\begin{aligned} A \;\mapsto\; R &= G(n, n+1, c_n, s_n) \cdots G(1, 2, c_1, s_1)\, A\,, \\ Q^* &= G(n, n+1, c_n, s_n) \cdots G(1, 2, c_1, s_1)\,. \end{aligned} \tag{14.3}$$

Das folgende Schema illustriert – für $n = 5$ und $m = 6$ – die dabei notwendigen Transformationen:



Dabei deuten die grau hinterlegten Flächen jeweils an, welche Zeilen der Matrix durch die Givens-Rotationen kombiniert werden, und die Sterne kennzeichnen neu berechnete Einträge.

Aufwand. Die Anzahl der Multiplikationen der n Givens-Rotationen summiert sich ungefähr zu

$$\sum_{i=1}^{n} 4(n-i+1) \;=\; \sum_{i=1}^{n} 4i \;\sim\; 2n^2\,.$$

Dies ist etwa um den Faktor $n/3$ geringer als der Aufwand einer herkömmlichen QR-Zerlegung. ◊

Beispiel. Für die Ausgleichsgerade aus Beispiel 11.4 wurde in Beispiel 13.4 die QR-Zerlegung der Koeffizientenmatrix und die entsprechende rechte Seite bestimmt:

$$A = \begin{bmatrix} 1 & 0 \\ 1 & 3 \\ 1 & 4 \\ 1 & 7 \end{bmatrix} = Q \begin{bmatrix} -2 & -7 \\ 0 & -5 \\ 0 & 0 \\ 0 & 0 \end{bmatrix}, \qquad c = Q^* b = \begin{bmatrix} -13/2 \\ -5/2 \\ 99/34 \\ -5/34 \end{bmatrix}.$$

Wir wollen nun wie in Beispiel 14.2 annehmen, daß ein fünfter Meßpunkt $(x, \tilde{y}) = (1, 1)$ neu hinzukommt und die aktualisierte Ausgleichsgerade $y = \alpha^+ + \beta^+ x$ bezüglich der erweiterten Datenmenge berechnen. Das zugehörige Problem (14.1) lautet dann

$$\text{minimiere} \quad \left\| \begin{bmatrix} 1 \\ -13/2 \\ -5/2 \end{bmatrix} - \begin{bmatrix} 1 & 1 \\ -2 & -7 \\ 0 & -5 \end{bmatrix} \begin{bmatrix} \alpha^+ \\ \beta^+ \end{bmatrix} \right\|_2 .$$

Die QR-Zerlegung dieser Koeffizientenmatrix in Hessenberg-Form erfolgt durch zwei Givens-Rotationen. Die erste Rotation kombiniert die ersten beiden Zeilen derart, daß der $(2,1)$-Eintrag Null wird. Die entsprechenden Parameter von $G(1, 2, c_1, s_1)$ sind durch

$$t_1 = 1/2\,, \qquad c_1 = 1/\sqrt{5}\,, \qquad s_1 = -2/\sqrt{5}$$

gegeben, und die zugehörige Transformation (inklusive der rechten Seite) lautet

$$\left[\begin{array}{cc|c} 1 & 1 & 1 \\ -2 & -7 & -13/2 \\ 0 & -5 & -5/2 \end{array}\right] \xrightarrow{i=1} \left[\begin{array}{cc|c} \sqrt{5} & 3\sqrt{5} & 14/\sqrt{5} \\ 0 & -\sqrt{5} & -4.5/\sqrt{5} \\ 0 & -5 & -5/2 \end{array}\right].$$

Die zweite Givens-Rotation mit den Koeffizienten

$$t_2 = -1/\sqrt{5}\,, \qquad c_2 = -1/\sqrt{6}\,, \qquad s_2 = -\sqrt{5}/\sqrt{6}$$

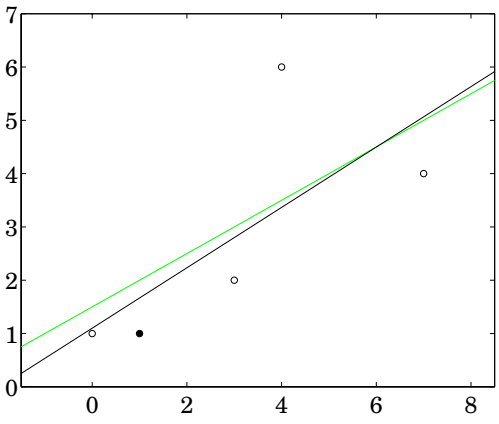


Abb. 14.2: Aktualisierte Ausgleichsgerade

transformiert die zweite und die dritte Zeile derart, daß der $(3,2)$-Eintrag Null wird:

$$\left[\begin{array}{cc|c} \sqrt{5} & 3\sqrt{5} & 14/\sqrt{5} \\ 0 & -\sqrt{5} & -4.5/\sqrt{5} \\ 0 & -5 & -5/2 \end{array}\right] \xrightarrow{i=2} \left[\begin{array}{cc|c} \sqrt{5} & 3\sqrt{5} & 14/\sqrt{5} \\ 0 & \sqrt{30} & 17/\sqrt{30} \\ 0 & 0 & -2/\sqrt{6} \end{array}\right].$$

Die Parameter α^+ und β^+ der neuen Ausgleichsgeraden lösen somit das Gleichungssystem

$$\begin{bmatrix} \sqrt{5} & 3\sqrt{5} \\ 0 & \sqrt{30} \end{bmatrix} \begin{bmatrix} \alpha^+ \\ \beta^+ \end{bmatrix} = \begin{bmatrix} 14/\sqrt{5} \\ 17/\sqrt{30} \end{bmatrix}.$$

Die resultierende Ausgleichsgerade lautet $y = 11/10 + 17/30\, x$. Abbildung 14.2 zeigt die fünf Datenpunkte sowie die neue und die alte Ausgleichsgerade; die gepunktete Linie ist die alte Gerade. Der neue Punkt $(1,1)$ ist durch einen ausgefüllten Kreis hervorgehoben. ◊

15 Ein CG-Verfahren für das Ausgleichsproblem

Wenn die Matrix A groß und dünn besetzt ist, wird man zur Lösung des linearen Ausgleichsproblems der QR-Zerlegung unter Umständen ein Iterationsverfahren vorziehen. Da die Koeffizientenmatrix A^*A der Gaußschen Normalengleichungen hermitesch und – falls A vollen Spaltenrang hat – positiv definit ist, bietet sich das CG-Verfahren an, angewandt auf die Normalengleichungen

$$A^*Ax = A^*b\,.$$

Allerdings haben wir bereits früher gesehen, daß die Kondition der Matrix A^*A wesentlich schlechter ist als die Kondition der Matrix A. Daher sollte nach Möglichkeit das explizite Ausmultiplizieren von A^*A vermieden werden. Zudem ist die Matrix A^*A in der Regel nicht mehr dünn besetzt, und damit wäre ein Hauptargument für die Verwendung iterativer Verfahren zunichte.

Durch eine geschickte Umformulierung des CG-Verfahrens können beide Probleme umgangen werden. Wir bezeichnen im weiteren mit $x^{(k)}$ die Iterierten des auf die Normalengleichungen angewendeten CG-Verfahrens und führen die beiden Residuen

$$r^{(k)} \;=\; b - Ax^{(k)} \qquad \text{und} \qquad s^{(k)} \;=\; A^*b - A^*Ax^{(k)} \;=\; A^*r^{(k)}$$

ein: $r^{(k)}$ ist wie bisher üblich das Residuum des Ausgleichsproblems während $s^{(k)}$ das Residuum der Normalengleichungen angibt. Da das CG-Verfahren auf die Normalengleichungen angewendet wird, übernimmt also $s^{(k)}$ die Rolle von $r^{(k)}$ in Algorithmus 9.1.

Die Koeffizientenmatrix A^*A geht in Algorithmus 9.1 an zwei Stellen ein: bei der Berechnung von α_k und bei der Aktualisierung von $s^{(k)}$. Bei der Definition von α_k kann wegen des Innenprodukts leicht auf das Ausmultiplizieren der Koeffizientenmatrix A^*A verzichtet werden, denn es gilt

$$\alpha_k \;=\; \|s^{(k)}\|_2^2 \,/\, \|Ad^{(k)}\|_2^2 \,. \tag{15.1}$$

Die Berechnung von $s^{(k+1)}$ hingegen erfolgt am einfachsten in zwei Schritten: Zunächst aktualisiert man das Residuum des Ausgleichsproblems,

$$r^{(k+1)} \;=\; r^{(k)} - \alpha_k Ad^{(k)} \,, \tag{15.2}$$

und berechnet anschließend

$$s^{(k+1)} \;=\; A^*r^{(k+1)} \,. \tag{15.3}$$

Diese Umformungen liefern das Verfahren CGLS (engl.: ***c**onjugate **g**radient method for linear **l**east- **s**quares problems*), das bei exakter Arithmetik äquivalent zu Algorithmus 9.1, angewandt auf die Normalengleichungen, ist.

Aufwand. Wenn $Ad^{(k)}$ zwischengespeichert wird, benötigt das CGLS-Verfahren (Algorithmus 15.1) neben zwei Innenprodukten lediglich je eine Multiplikation mit A und A^* pro Iterationsschritt. ◊

Die wichtigsten Eigenschaften des CGLS-Verfahrens ergeben sich als unmittelbare Konsequenz aus den Resultaten von Abschnitt 9.

Initialisierung: $A \in \mathbb{K}^{m\times n}$ und $b \in \mathbb{K}^m$ seien gegeben

wähle beliebiges $x^{(0)} \in \mathbb{K}^n$
$r^{(0)} = b - Ax^{(0)}$
$s^{(0)} = A^* r^{(0)}$
$d^{(0)} = s^{(0)}$
`for` $k = 0, 1, 2, \ldots$ `do`
  $\alpha_k = \|s^{(k)}\|_2^2 \;/\; \|Ad^{(k)}\|_2^2$ % $Ad^{(k)}$ für später abspeichern
  $x^{(k+1)} = x^{(k)} + \alpha_k d^{(k)}$
  $r^{(k+1)} = r^{(k)} - \alpha_k Ad^{(k)}$
  $s^{(k+1)} = A^* r^{(k+1)}$
  $\beta_k = \|s^{(k+1)}\|_2^2 \;/\; \|s^{(k)}\|_2^2$
  $d^{(k+1)} = s^{(k+1)} + \beta_k d^{(k)}$
`until` stop % end for

Ergebnis: $x^{(k)}$ ist die Approximation der Lösung des Ausgleichsproblems (11.2), $r^{(k)} = b - Ax^{(k)}$ das Residuum und $s^{(k)}$ das Residuum der Normalengleichungen

Algorithmus 15.1: CGLS-Verfahren

Satz 15.1. *Die k-te Iterierte $x^{(k)}$ des CGLS-Verfahrens liegt in dem verschobenen Krylov-Rraum*

$$x^{(0)} + \mathcal{K}_k(A^*A, A^*r^{(0)}) = x^{(0)} + \operatorname{span}\{A^*r^{(0)}, (A^*A)A^*r^{(0)}, \ldots, (A^*A)^{k-1}A^*r^{(0)}\}\,.$$

Unter allen Elementen x dieses affinen Raums minimiert $x^{(k)}$ die Residuennorm $\|b - Ax\|_2$.

Beweis. Die erste Behauptung, $x^{(k)} \in x^{(0)} + \mathcal{K}_k(A^*A, A^*r^{(0)})$, folgt sofort aus der Aussage (9.10) von Satz 9.5. Zudem minimiert $x^{(k)}$ nach demselben Satz die Zielfunktion Φ aus Abschnitt 9 unter allen Elementen dieses Krylov-Raums; im hiesigen Kontext hat Φ die Gestalt (vgl. Bemerkung 9.6)

$$\Phi(x) = \frac{1}{2}\, x^*(A^*A)x - \operatorname{Re} x^*(A^*b)\,.$$

Eine einfache Rechnung ergibt

$$\frac{1}{2}\,\|b - Ax\|_2^2 = \frac{1}{2}\,\|b\|_2^2 - \operatorname{Re} b^*Ax + \frac{1}{2}\,\|Ax\|_2^2 = \Phi(x) + \frac{1}{2}\,\|b\|_2^2$$

und folglich stimmen $\Phi(x)$ und $\frac{1}{2}\|b - Ax\|_2^2$ bis auf eine additive Konstante überein. Damit ist der Beweis vollständig. □

Wegen dieser Minimierungseigenschaft ist das CGLS-Verfahren also direkt auf das lineare Ausgleichsproblem zugeschnitten.

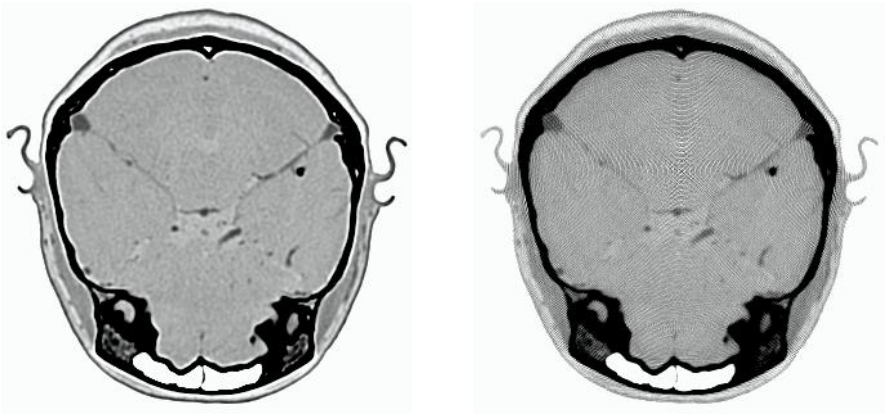


Abb. 15.1: Röntgentomographie: Siemens-Rekonstruktion (li.) und CGLS-Iterierte (re.)

Bemerkung 15.2. Das CGLS-Verfahren kann ohne Modifikationen auch auf Matrizen A ohne vollen Spaltenrang angewendet werden. Auch Satz 15.1 bleibt unverändert gültig, und die Iterierten $x^{(k)}$ konvergieren gegen eine Lösung des linearen Ausgleichsproblems. Bei Matrizen ohne vollen Spaltenrang ist das lineare Ausgleichsproblem allerdings nicht mehr eindeutig lösbar. Der Grenzwert $\widehat{x}$ der CGLS-Iterierten ist jedoch durch den folgenden Zusatz *eindeutig* festgelegt:

> Unter allen Lösungen des linearen Ausgleichsproblems minimiert $\widehat{x}$ den euklidischen Abstand $\|\widehat{x} - x^{(0)}\|_2$ zu $x^{(0)}$.

Die Wahl von $x^{(0)}$ dient hier also als eine Art Auswahlkriterium des Grenzwerts und kann dazu verwendet werden, A-priori-Informationen über die Lösung in das Verfahren einfließen zu lassen. Für einen Beweis dieses Zusatzes vgl. Aufgabe 11. ◊

Beispiel 15.3. Algorithmus 15.1 bietet eine Möglichkeit zur Lösung des riesigen überbestimmten linearen Gleichungssystems $Ax = b$, das in der Einleitung für die Computertomographie hergeleitet wurde. Für dieses Gleichungssystem der Dimension 389760×262144 benötigt das Verfahren nur 21 Iterationen, um den Kopfquerschnitt relativ gut zu approximieren, vgl. Abbildung 15.1. Genauere Rekonstruktionen sind leider wegen der Modellierungsfehler mit diesem einfachen Ansatz nicht möglich, da der tatsächliche Kopfquerschnitt $\widehat{x}$ das Modell nur bis auf einen relativen Fehler von rund 1.7% erfüllt, d. h.

$$\frac{\|b - A\widehat{x}\|_2}{\|b\|_2} \approx 0.0172\,.$$

Dieser recht hohe Datenfehler und die schlechte Kondition der Matrix A sind

die Ursache dafür, daß weitere Iterationen das Bild nicht mehr verbessern sondern, im Gegenteil, verschlechtern. Dieses verwunderliche Phänomen der *Semikonvergenz* läßt sich mit einer detaillierteren Konvergenzanalyse erklären (siehe [26] oder [75]). ◊

16 Nachtrag: Das GMRES-Verfahren

Zum Abschluß dieses Kapitels stellen wir noch das GMRES-Verfahren vor, ein CG-artiges Iterationsverfahren für *invertierbare* Matrizen $A \in \mathbb{K}^{n\times n}$, die allerdings nicht wie in Abschnitt 9 hermitesch und positiv definit sein müssen. Obwohl zu dem aktuellen Kapitel kein unmittelbarer thematischer Bezug zu bestehen scheint, haben wir die Herleitung dieses Verfahrens bis zum jetzigen Zeitpunkt zurückgestellt, da bei der Konstruktion Methoden der linearen Ausgleichsrechnung Verwendung finden.

Die Zielsetzung des GMRES-Verfahrens ist ähnlich zu der des CG-Verfahrens, vgl. Satz 9.5: Gesucht wird das Element

$$x^{(k)} \in x^{(0)} + \mathcal{K}_k(A, r^{(0)})\,, \qquad r^{(0)} = b - Ax^{(0)}\,, \tag{16.1}$$

welches innerhalb dieses affin verschobenen Krylov-Raums die Euklidnorm des Residuums

$$\Phi(x) = \|b - Ax\|_2$$

minimiert. Die Zielfunktion Φ ersetzt dabei die Energienorm aus Abschnitt 9, da diese für allgemeine Matrizen A keine Norm darstellt.

Für die Lösung dieses Minimierungsproblems konstruieren wir zunächst eine Orthogonalbasis von $\mathcal{K}_k(A, r^{(0)})$; beim CG-Verfahren wird diese Orthogonalbasis durch die Residuenvektoren erzeugt, vgl. Lemma 9.2. Im allgemeinen Fall ist die Berechnung einer solchen Orthogonalbasis erheblich aufwendiger. Wir verwenden zu diesem Zweck den *Arnoldi-Prozeß*, der in den ersten $j = 1, 2, \ldots, n$ Iterationsschritten sukzessive den Startvektor

$$v_1 = r^{(0)} / \|r^{(0)}\|_2 \tag{16.2}$$

zu einer Orthonormalbasis $\{v_1, \ldots, v_j\}$ von $\mathcal{K}_j(A, r^{(0)})$ ergänzt.. Diese Basisvektoren werden für das weitere in Matrizen

$$V_j = [v_1, \ldots, v_j] \in \mathbb{K}^{n\times j}\,, \qquad j = 1, 2, \ldots, n\,,$$

gesammelt. Die zugehörigen Produkte $P_j = V_j V_j^*$, $j = 1, \ldots, n$, bilden dann Orthogonalprojektoren auf $\mathcal{K}_j(A, r^{(0)})$.

Um V_j zu V_{j+1} zu ergänzen, setzen wir

$$v_{j+1} \;=\; q_{j+1}/\,\|q_{j+1}\|_2\,, \qquad q_{j+1} = (I - P_j)Av_j\,. \tag{16.3}$$

Offensichtlich gehört Av_j zu $\mathcal{K}_{j+1}(A, r^{(0)})$ und aufgrund der Konstruktion gilt $v_{j+1} \perp \mathcal{K}_j(A, r^{(0)})$. Also ist v_{j+1} die (bis auf das Vorzeichen eindeutige) Ergänzung der Orthonormalbasis von $\mathcal{K}_j(A, r^{(0)})$ zu einer Orthonormalbasis von $\mathcal{K}_{j+1}(A, r^{(0)})$. Eine Ausnahmesituation tritt ein, wenn q_{j+1} der Nullvektor ist. Dann gehört Av_j wegen (16.3) selbst zum Krylov-Raum $\mathcal{K}_j(A, r^{(0)})$ und es folgt

$$\mathcal{K}_j(A, r^{(0)}) = \mathcal{K}_{j+1}(A, r^{(0)})\,,$$

d. h. A ist eine Selbstabbildung des Krylov-Raums $\mathcal{K}_j(A, r^{(0)})$. In diesem Fall ist die Menge $\{v_1, \ldots, v_j\}$ eine vollständige Basis *aller* Krylov-Räume $\mathcal{K}_i(A, r^{(0)})$ mit $i \geq j$, und die GMRES-Iterierte $x^{(j)}$ ist die exakte Lösung des linearen Gleichungssystems $Ax = b$, vgl. Aufgabe 14.

Aus (16.3) erhalten wir mit $P_j = V_j V_j^*$

$$Av_j \;=\; V_j V_j^* Av_j + \|q_{j+1}\|_2 v_{j+1} \;=\; \sum_{i=1}^{j} (v_i^* Av_j) v_i \;+\; \|q_{j+1}\|_2 v_{j+1}\,. \tag{16.4}$$

Läuft j von 1 bis k, so ergibt (16.4) die Matrix-Gleichung

$$AV_k \;=\; V_{k+1} H_k\,, \tag{16.5}$$

wobei

$$H_k = \begin{bmatrix} h_{11} & h_{12} & h_{13} & \cdots & h_{1k} \\ h_{21} & h_{22} & h_{23} & \cdots & h_{2k} \\ 0 & h_{32} & h_{33} & & h_{3k} \\ & \ddots & \ddots & \ddots & \vdots \\ \vdots & & 0 & h_{k,k-1} & h_{kk} \\ 0 & \ldots & & 0 & h_{k+1,k} \end{bmatrix} \in \mathbb{K}^{(k+1)\times k} \tag{16.6}$$

eine obere Hessenberg-Matrix mit den entsprechenden Einträgen

$$h_{ij} \;=\; \begin{cases} v_i^* Av_j\,, & i \leq j\,, \\ \|q_i\|_2\,, & i = j+1\,, \\ 0\,, & i > j+1\,, \end{cases}$$

ist.

Initialisierung: $A \in \mathbb{K}^{n\times n}$ und $v_1 \in \mathbb{K}^n$ mit $\|v_1\|_2 = 1$ seien gegeben

```
for j = 1, 2, ... do
  ṽ_{j+1} = A v_j
  for i = 1, 2, ..., j do
    h_ij = v_i^* ṽ_{j+1}
    ṽ_{j+1} = ṽ_{j+1} − h_ij v_i
  end for
  h_{j+1,j} = ‖ṽ_{j+1}‖_2
  if h_{j+1,j} ≠ 0 then
    v_{j+1} = ṽ_{j+1} / h_{j+1,j}
  end if
until h_{j+1,j} = 0        % end for
k = j        % letzter Wert des Lauf-Indizes der for-Schleife
```

Ergebnis: Die Werte h_{ij} sind die Einträge der Hessenberg-Matrix H_k aus (16.6), $\{v_j\}_{j=1}^k$ bildet eine Orthonormalbasis des Krylov-Raums $\mathcal{K}_k(A, v_1)$

Algorithmus 16.1: Arnoldi-Prozeß

Die Gleichung (16.4) besagt, daß die Summe auf der rechten Seite von Av_j abgezogen werden muß, um die orthogonale Ergänzung v_{j+1} der Orthonormalbasis $\{v_1, \ldots, v_j\}$ zu bestimmen. Diese Vorgehensweise, bei der der neue Basisvektor sukzessive gegen alle alten Basisvektoren orthogonalisiert wird, beruht auf dem *Gram-Schmidt-Orthogonalisierungsverfahren*. Aufgrund von Rundungsfehlern ergibt sich jedoch selten wirklich die Identität $V_j^* V_j = I$, d. h. die Orthogonalität der Basisvektoren geht mit wachsendem j zunehmend verloren. Als etwas stabiler erweist sich das *modifizierte Gram-Schmidt-Verfahren*, das in Algorithmus 16.1 verwendet wird und das bei exakter Arithmetik mathematisch äquivalent zu dem Gram-Schmidt-Verfahren ist, vgl. etwa Björck [8] oder Higham [55].

Wir wenden uns nun wieder dem GMRES-Verfahren zu. Mit Hilfe der Orthonormalbasis aus dem Arnoldi-Prozeß kann der gesuchte Vektor $x^{(k)}$ in der Form

$$x^{(k)} = x^{(0)} + V_k z^{(k)}$$

mit einem geeigneten, aber noch zu bestimmenden $z^{(k)}$ geschrieben werden. Wird diese Darstellung in die Zielfunktion Φ eingesetzt, so ergibt sich für $z^{(k)}$ das lineare Ausgleichsproblem

$$\text{minimiere} \quad \|r^{(0)} - AV_k z^{(k)}\|_2 \,. \tag{16.7}$$

Wegen der Wahl (16.2) des Startvektors für den Arnoldi-Prozeß haben wir

$r^{(0)} = \rho_0 V_{k+1} e_1$ mit $\rho_0 = \|r^{(0)}\|_2$, ferner können wir AV_k mit Hilfe der Faktorisierung (16.5) ersetzen. Somit ist das Ausgleichsproblem (16.7) für $z^{(k)}$ äquivalent zu

$$\text{minimiere} \quad \|\rho_0 V_{k+1} e_1 - V_{k+1} H_k z^{(k)}\|_2 \;=\; \|\rho_0 e_1 - H_k z^{(k)}\|_2 \,.$$

Zur Berechnung von $z^{(k)}$ muß nun ein $(k+1) \times k$-dimensionales lineares Ausgleichsproblem mit einer oberen Hessenberg-Matrix gelöst werden.

Eine QR-Zerlegung von H_k kann wie in Abschnitt 14 effizient mit Givens-Rotationen implementiert werden. Da die ersten k Zeilen und die ersten $k-1$ Spalten von H_k mit H_{k-1} übereinstimmen,

$$H_k \;=\; \left[\begin{array}{c:c} H_{k-1} & h_k \\ \hdashline 0 & h_{k+1,k} \end{array}\right],$$

kann dabei die QR-Zerlegung von H_{k-1} aus dem vorangegangenen Iterationsschritt ausgenutzt werden. Ist

$$G_{k-1} \cdots G_2 G_1 H_{k-1} \;=\; \begin{bmatrix} R_{k-1} \\ 0 \end{bmatrix} \;=\; \begin{bmatrix} 0 & & \\ \vdots & \ddots & \\ 0 & \cdots & 0 \end{bmatrix}$$

mit den Givens-Rotationen $G_j = G(j, j+1, c_j, s_j) \in \mathbb{K}^{k \times k}$ und der oberen Dreiecksmatrix $R_{k-1} \in \mathbb{K}^{(k-1)\times(k-1)}$ die QR-Zerlegung von H_{k-1}, so ergibt sich entsprechend mit den in $\mathbb{K}^{(k+1)\times(k+1)}$ eingebetteten Givens-Rotationen $\tilde{G}_j$ die Faktorisierung von H_k:

$$\tilde{G}_{k-1} \cdots \tilde{G}_2 \tilde{G}_1 H_k \;=\; \left[\begin{array}{ccc:c} 0 & & & \\ \vdots & \ddots & & \tilde{h}_k \\ 0 & \cdots & 0 & \\ \hdashline 0 & \cdots & 0 & h_{k+1,k} \end{array}\right], \quad \tilde{h}_k = G_{k-1} \cdots G_2 G_1 h_k \;\in \mathbb{K}^k \,.$$

Offensichtlich müssen die Givens-Rotationen $G_1, \ldots, G_{k-1}$ lediglich auf die letzte Spalte h_k von H_k angewendet werden, um $\tilde{h}_k$ zu berechnen. Danach muß noch eine neue Givens-Rotation $G_k = G(k, k+1, c_k, s_k) \in \mathbb{K}^{(k+1)\times(k+1)}$ bestimmt werden, die das $(k+1, k)$-Element auf Null transformiert.

Algorithmus 16.2 faßt alle Schritte des GMRES-Verfahrens zusammen (der Übersichtlichkeit halber für reelle Matrizen), wobei der Arnoldi-Prozeß gleich in der oben beschriebenen Weise mit der Reduktion von H_k auf obere Dreiecksgestalt verzahnt ist.

Initialisierung: $A \in \mathbb{K}^{n\times n}$, $b \in \mathbb{K}^n$ und $x^{(0)} \in \mathbb{K}^n$ seien gegeben

$r^{(0)} = b - Ax^{(0)}$
$d = [d_0] = \left[\, \|r^{(0)}\|_2 \,\right] \in \mathbb{R}^1$
$v_1 = r^{(0)}/d_0$
`for` $k = 1, 2, \ldots$ `do` % Iteration von GMRES
  $\tilde{v}_{k+1} = Av_k$
  `for` $i = 1, 2, \ldots, k$ `do` % Arnoldi-Prozeß
    $h_{ik} = v_i^* \tilde{v}_{k+1}$
    $\tilde{v}_{k+1} = \tilde{v}_{k+1} - h_{ik} v_i$
  `end for`
  $\omega = \|\tilde{v}_{k+1}\|_2$ % ω entspricht $h_{k+1,k}$; wird aber nicht in Matrix H abgespeichert
  `for` $i = 1, 2, \ldots, k-1$ `do` % wende die alten Givens-Rotationen auf h_k an
    $\tilde{h} = c_i h_{ik} + s_i h_{i+1,k}$
    $h_{i+1,k} = -s_i h_{ik} + c_i h_{i+1,k}$
    $h_{ik} = \tilde{h}$
  `end for`
  % bestimme die neue Givens-Rotation G_k
  `if` $\omega \le |h_{kk}|$ `then`
    $t_k = \omega/|h_{kk}|$
    $c_k = h_{kk}/\big(|h_{kk}|(1+t_k^2)^{1/2}\big)$
    $s_k = t_k/(1+t_k^2)^{1/2}$
  `else`
    $t_k = h_{kk}/\omega$
    $c_k = t_k/(1+t_k^2)^{1/2}$
    $s_k = 1/(1+t_k^2)^{1/2}$
  `end if`
  % wende die neue Givens-Rotation auf die letzte Spalte von H und die rechte Seite an
  $h_{kk} = c_k h_{kk} + s_k \omega$ % $H = [h_{ij}]$ ist nun die reduzierte $k \times k$ obere Dreiecksmatrix
  % ergänze $d = [d_j]_0^{k-1}$ zu einem Vektor in $\mathbb{K}^{k+1}$
  $d_k = -s_k d_{k-1}$ % $|d_k|$ ist gleichzeitig die aktuelle Residuennorm $\|b - Ax^{(k)}\|_2$
  $d_{k-1} = c_k d_{k-1}$
`until` $|d_k|$ ist hinreichend klein % end for
% Berechnung von $x^{(k)}$ nur am Schluß
$Hz^{(k)} = d$ % bestimme Entwicklungskoeffizienten durch Lösen des Dreiecksystems
$x^{(k)} = x^{(0)} + V_k z^{(k)}$ % $V_k = [v_1, \ldots, v_k]$

Ergebnis: $x^{(k)}$ ist die letzte Iterierte, $|d_k|$ die zugehörige Residuennorm $\|b - Ax^{(k)}\|_2$

Algorithmus 16.2: (Reelles) GMRES-Verfahren

Aufwand. Die Kosten des GMRES-Verfahrens werden von denen des Arnoldi-Prozesses dominiert. Neben der Matrix-Vektor-Multiplikation mit A besteht der Hauptaufwand der k-ten Iteration aus den $k+1$ Innenprodukten zur Berechnung der letzten Spalte von H_k. Werden die Kosten einer Matrix-Vektor-Multiplikation bei einer dünn besetzten Matrix mit $O(n)$ Operationen angesetzt, so ergibt sich ein Gesamtaufwand von $k^2n/2 + O(kn)$ Multiplikationen für die ersten k Iterationen des Verfahrens.

Von Bedeutung ist auch der Speicheraufwand des Verfahrens, da alle Basisvektoren v_j, $j = 1, \ldots, k$, abgespeichert werden müssen, um am Schluß die Näherungslösung $x^{(k)}$ berechnen zu können. k Iterationen benötigen etwa kn Speicherplätze, was bei vielen Anwendungen bereits für moderate k den Speicherbedarf der Koeffizientenmatrix übersteigt. ◊

Der Rechenaufwand ist also um den Faktor $k/2$ höher als der des CG-Verfahrens im hermitesch positiv definiten Fall. Der anwachsende Speicheraufwand ist jedoch häufig das größere Problem. Für die Effizienz des GMRES-Verfahrens ist daher in besonderem Maße die Konvergenzgeschwindigkeit entscheidend.

Leider fehlt – im Gegensatz zu dem CG-Verfahren für den positiv definiten Fall – eine befriedigende Konvergenztheorie für das GMRES-Verfahren. Klar ist aufgrund der Minimierungsvorgabe, daß das Verfahren (bei exakter Arithmetik) nach spätestens n Iterationen die exakte Lösung des Gleichungssystems gefunden hat. Andererseits kommen so viele Iterationen aufgrund der Aufwandsabschätzung nicht ernsthaft in Betracht.

In der Praxis beobachtet man in der Regel verschiedene Phasen der Iteration, bei denen die Fehlerkurve abwechselnd zu stagnieren scheint, dann linear oder gar superlinear abfällt. Um Stagnationen zu vermeiden, ist wie bei dem CG-Verfahren ein guter Präkonditionierer oftmals unverzichtbar. Bei der Konstruktion des Präkonditionierers (der in diesem Kontext nicht hermitesch positiv definit zu sein braucht) werden in der Regel die gleichen Kriterien angelegt, die sich beim CG-Verfahren bewährt haben: Demnach soll der Präkonditionierer entweder die Kondition der Koeffizientenmatrix reduzieren oder deren Spektrum in gewissen Punkten $\lambda \neq 0$ konzentrieren. Allerdings kann man auch Gleichungssysteme konstruieren, bei denen eine solche Strategie fehlschlägt; entsprechende Beispiele finden sich etwa in dem Buch von Greenbaum [36].

Um den Speicheraufwand in den Griff zu bekommen, kann man das GMRES-Verfahren nach jeweils ℓ Iterationen abbrechen und dann die Iteration mit der aktuellen Iterierten als Startvektor neu beginnen. Die Iterierten dieser Variante GMRES(ℓ) liegen jeweils in den gleichen Krylov-Räumen wie die entsprechenden Iterierten des GMRES-Verfahrens, minimieren aber ab dem ersten Neustart nicht mehr die Euklidnorm des Residuums, da bei GMRES(ℓ) zur Lösung des Minimierungsproblems (16.7) nur noch diejenigen Arnoldi-Vektoren

herangezogen werden, die seit dem letzten Neustart konstruiert wurden. Insbesondere findet diese Variante in der Regel nicht mehr die exakte Lösung des Gleichungssystems in endlicher Zeit. Dennoch kann für eine relativ große Klasse von Matrizen Konvergenz bewiesen werden.

Satz 16.1. *Sei $A \in \mathbb{K}^{n\times n}$ mit $\operatorname{Re} x^*Ax \geq \alpha \|x\|_2^2$ für ein $\alpha > 0$ und alle $x \in \mathbb{K}^n$. Dann konvergiert das GMRES(ℓ)-Verfahren gegen die Lösung des Gleichungssystems $Ax = b$.*

Bemerkung. Die Bedeutung der hier geforderten Voraussetzung an A wird in anderem Kontext in Abschnitt 23 untersucht: Beispielsweise folgt aus dem Satz von Bendixson (Satz 23.6), daß unter dieser Voraussetzung das Spektrum von A notwendigerweise in der rechten Halbebene der komplexen Ebene liegt. ◊

Beweis von Satz 16.1. Für beliebiges $\omega \in \mathbb{R}$ und $x \in \mathbb{K}^n$ ergibt sich aufgrund der Voraussetzung an A die Ungleichung

$$\begin{aligned}\|(I-\omega A)x\|_2^2 &= \|x\|_2^2 - 2\omega \operatorname{Re} x^*Ax + \omega^2 \|Ax\|_2^2 \\ &\leq (1 - 2\omega\alpha + \omega^2 \|A\|_2^2)\, \|x\|_2^2 .\end{aligned}$$

Für $0 < \omega < 2\alpha/\|A\|_2^2$ ist die rechte Seite kleiner als $\|x\|_2^2$ und somit $I - \omega A$ eine Kontraktion bezüglich der Euklidnorm. Wir fixieren ein entsprechendes ω und bezeichnen den Kontraktionsfaktor mit q, $0 \leq q < 1$. Für $1 \leq k \leq \ell$ stimmen die k-te Iterierte $x^{(k)}$ von GMRES(ℓ) und die entsprechende GMRES-Iterierte überein. Aufgrund der Minimaleigenschaft der GMRES-Iterierten können wir daher das zugehörige Residuum mit dem Residuum von

$$\widetilde{x}^{(k)} = x^{(0)} + \omega \sum_{j=0}^{k-1} (I-\omega A)^j r^{(0)} \in x^{(0)} + \mathcal{K}_k(A, r^{(0)})$$

vergleichen: Aus

$$\begin{aligned} b - A\widetilde{x}^{(k)} &= r^{(0)} - \omega A \sum_{j=0}^{k-1} (I-\omega A)^j r^{(0)} \\ &= r^{(0)} - \sum_{j=0}^{k-1} (I-\omega A)^j r^{(0)} + \sum_{j=0}^{k-1} (I-\omega A)^{j+1} r^{(0)} \\ &= r^{(0)} - r^{(0)} + (I-\omega A)^k r^{(0)} = (I-\omega A)^k r^{(0)} \end{aligned}$$

folgt

$$\|b - Ax^{(k)}\|_2 \leq \|b - A\widetilde{x}^{(k)}\|_2 = \|(I-\omega A)^k r^{(0)}\|_2 \leq q^k \|r^{(0)}\|_2 .$$

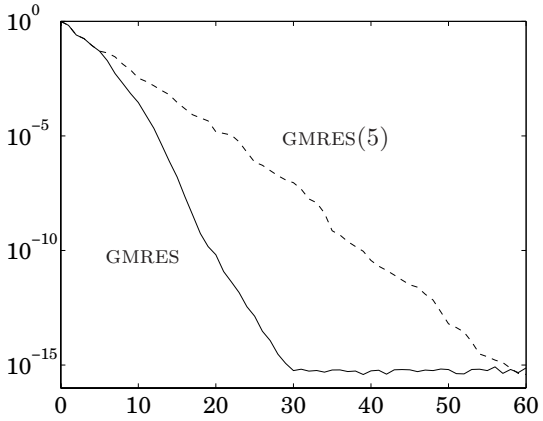


Abb. 16.1: Konvergenzverlauf von GMRES und GMRES(5)

Entsprechend ergibt sich nach dem ersten Neustart, also nach k Iterationen mit $\ell < k \leq 2\ell$ die Abschätzung

$$\|b - Ax^{(k)}\|_2 \;\leq\; q^{k-\ell}\|b - Ax^{(\ell)}\|_2 \;\leq\; q^{k-\ell}q^{\ell}\|r^{(0)}\|_2 \;=\; q^k\|r^{(0)}\|_2\,.$$

So fortfahrend erhalten wir dieselbe Abschätzung für jedes $k > 0$, so daß sich der Iterationsfehler $A^{-1}b - x^{(k)} = A^{-1}(b - Ax^{(k)})$ durch

$$\|A^{-1}b - x^{(k)}\|_2 \;\leq\; q^k\|A^{-1}\|_2\|r^{(0)}\|_2\,, \qquad k = 0, 1, 2, \ldots, \tag{16.8}$$

abschätzen läßt. Da die rechte Seite für $k \to \infty$ gegen Null konvergiert, folgt hieraus die Behauptung. □

Beispiel 16.2. Die Kurven in Abbildung 16.1 zeigen die Entwicklung des relativen Fehlers für GMRES und GMRES(5) bei einem Gleichungssystem, dessen Koeffizientenmatrix nicht hermitesch ist. Es handelt sich dabei um das Gleichungssystem aus Abschnitt 59 mit der Dimension $n = 256$. Dies ist ein Beispiel, in dem GMRES sehr gut konvergiert: Man sieht fast durchweg ein lineares Konvergenzverhalten, das nach den ersten fünf Iterationen etwas schneller wird. Erwartungsgemäß ist die Konvergenz von GMRES(5) nicht ganz so gut. Es sei angemerkt, daß in diesem Beispiel die Voraussetzung von Satz 16.1 erfüllt ist. ◊

Die Konstruktion effizienter Iterationsverfahren zur Lösung nichthermitescher Gleichungssysteme ist gegenwärtig ein sehr aktives Forschungsgebiet. Neben dem GMRES-Verfahren werden auch Verfahren untersucht, bei denen anstelle der Orthogonalbasis aus dem Arnoldi-Prozeß (mit dem hohen Speicherbedarf

aufgrund der langen Rekursionen) andere Basen des Krylov-Raums mit kurzen Rekursionen erzeugt werden. Während diese Verfahren mit geringem Speicherplatz und oft auch dem Rechenaufwand des klassischen CG-Verfahrens auskommen, minimieren die Iterierten in der Regel keine der üblichen Zielfunktionen wie etwa die Residuennorm. Entsprechend schwierig ist ihre Konvergenzuntersuchung. Für eine Übersicht dieser Verfahren sei wieder auf das Buch von Greenbaum [36] verwiesen.

Aufgaben

1. Gegeben sei die Wertetabelle

x	0	h	$2h$	$\cdots$	nh
y	y_0	y_1	y_2	$\cdots$	y_n

mit $h \in \mathbb{R}^+$ und $n \in \mathbb{N}$ für die Funktion $y = y(x)$. Rechnen Sie nach, daß

$$\beta \;=\; \frac{6}{n(n+1)(n+2)\,h} \sum_{i=0}^{n} (2i-n)\, y_i$$

die Steigung der besten Ausgleichsgerade $G = \{(x,y) : y = \alpha + \beta x\}$ im Sinne von Beispiel 11.4 ist.

2. Der Abstand eines Punktes (x_i, y_i) von einer Geraden $G = \{(x,y) : y = \alpha + \beta x\}$ ist definiert als

$$d_i \;=\; \min_{(x,y)\in G} \big((x_i - x)^2 + (y_i - y)^2\big)^{1/2}.$$

(a) Weisen Sie die Identität

$$d_i \;=\; \frac{1}{1+\beta^2} \left\| \begin{bmatrix} \beta^2 & -\beta \\ -\beta & 1 \end{bmatrix} \begin{bmatrix} x_i \\ y_i - \alpha \end{bmatrix} \right\|_2$$

nach.
(b) Sind n Punkte $\{(x_i, y_i)\}_{i=1}^n$ gegeben, dann ist $d = (\sum_{i=1}^n d_i^2)^{1/2}$ der Abstand aller Punkte von der Geraden G. Bestimmen Sie den Abstand der vier Punkte von der Ausgleichsgeraden aus Beispiel 11.4 und überlegen Sie sich eine andere Wahl von α und β, für die dieser Abstand kleiner wird.

3. Es seien $A \in \mathbb{R}^{m\times n}$ und $C \in \mathbb{R}^{p\times n}$ mit $p < n < m + p$ und beide Matrizen haben vollen Rang. Ferner sei $b \in \mathbb{R}^m$ und $d \in \mathbb{R}^p$. Betrachten Sie das restringierte lineare Ausgleichsproblem $\frac{1}{2}\|b - Ax\|_2^2 \longrightarrow \min$ unter der *linearen Nebenbedingung* $Cx = d$. Zeigen Sie, daß die Lösung $\widehat{x}$ und $\widehat{r} = b - A\widehat{x}$ die entsprechenden (Block-)Komponenten der Lösung des linearen Gleichungssystems

$$\begin{bmatrix} I & 0 & A \\ 0 & 0 & C \\ A^* & C^* & 0 \end{bmatrix} \begin{bmatrix} r \\ \lambda \\ x \end{bmatrix} = \begin{bmatrix} b \\ d \\ 0 \end{bmatrix}$$

sind.
Hinweis: Verwenden Sie die Lagrange-Multiplikatorenregel.

4. Bei dem mechanischen Beispiel aus Abschnitt 3 stellen sich innere Kräfte f_k ein, die die äußeren Kräfte p_i an den Knoten kompensieren. Dieser Gleichgewichtszustand wird durch das lineare Gleichungssystem $p = Ef$ beschrieben, vgl. (3.1).
(a) Zeigen Sie unter der Voraussetzung, daß die Diagonalmatrix L in Abschnitt 3 die Einheitsmatrix ist, daß der Vektor f das Minimierungsproblem löst:

Minimiere $\|f\|_2^2$ unter der Nebenbedingung $Ef = p$.

(b) Betrachten Sie nun den Fall einer allgemeinen Diagonalmatrix L: Zeigen Sie, daß f dann das Minimierungsproblem

$$\text{minimiere} \quad \sum_{k=1}^{18} l_k f_k^2 \quad \text{unter der Nebenbedingung } Ef = p\,.$$

löst.

5. Seien $A \in \mathbb{R}^{m\times n}$ und $b \in \mathbb{R}^m$. Bei der *Tikhonov-Regularisierung* wird für $\alpha > 0$ die folgende Minimierungsaufgabe für $x \in \mathbb{R}^n$ gelöst:

$$\text{minimiere} \quad T_\alpha(x) \;=\; \|b - Ax\|_2^2 + \alpha\|x\|_2^2\,.$$

Zeigen Sie:
(a) Für jedes $\alpha > 0$ gibt es ein eindeutig bestimmtes $x_\alpha \in \mathbb{R}^n$, so daß

$$T_\alpha(x_\alpha) \le T_\alpha(x) \quad \text{für alle } x \in \mathbb{R}^n\,.$$

(b) x_α erfüllt das lineare Gleichungssystem

$$(A^TA + \alpha I)x_\alpha = A^Tb\,.$$

(c) Ist $A^\dagger$ die Pseudoinverse von A, dann gilt

$$x_\alpha \to A^\dagger b \quad \text{für } \alpha \to 0\,.$$

6. Sei $A \in \mathbb{K}^{n\times n}$. Zeigen Sie, daß gilt

$$\min\{\,\|A - Q\|_F^2 \,:\, Q \text{ unitär}\,\} \;=\; \sum_{i=1}^{n}(\sigma_i - 1)^2\,,$$

und geben Sie eine unitäre Matrix Q an, für die dieses Minimum angenommen wird.
Hinweis: Verwenden Sie $\|A - Q\|_F = \mathrm{Spur}\big((A-Q)(A^* - Q^*)\big)$.

7. Sei $A \in \mathbb{R}^{m\times n}$, $m \ge n$. Zeigen Sie, daß A eine Zerlegung der Gestalt $A = U^*BV$ besitzt, wobei $U \in \mathbb{R}^{m\times m}$ sowie $V \in \mathbb{R}^{n\times n}$ orthogonal sind und $B \in \mathbb{R}^{m\times n}$ eine untere Bidiagonalmatrix ist, d. h.

$$B = \begin{bmatrix} + & & & 0 \\ + & + & & \\ & + & \ddots & \\ & & \ddots & + \\ 0 & & & + \\ & & 0 & \end{bmatrix}.$$

8. $Q_1R_1 = Q_2R_2$ seien zwei QR-Faktorisierungen einer Matrix $A \in \mathbb{K}^{m\times n}$ mit $m \ge n$, d. h. Q_1 und Q_2 sind unitäre Matrizen und R_1 und R_2 obere Dreiecksmatrizen. Zeigen Sie, daß eine unitäre Diagonalmatrix $S \in \mathbb{K}^{m\times m}$ existiert, für die gilt:

$$Q_1 = Q_2S^*\,, \qquad R_1 = SR_2\,.$$

9. Diese Aufgabe beschäftigt sich mit der QR-Zerlegung mit Spaltenpivotsuche für Matrizen A ohne vollen Spaltenrang, vgl. Bemerkung 13.5.
(a) Überlegen Sie sich die Details der Spaltenpivotsuche, und leiten Sie die resultierende Faktorisierung (13.7) her.
(b) Wie kann man aus (13.7) eine Lösung des linearen Ausgleichsproblems bestimmen? Zeigen Sie an einem Beispiel, daß Ihre Lösung im allgemeinen keine minimale Norm besitzt.
(c) Implementieren Sie diesen Algorithmus und vergleichen Sie die von Ihnen berechnete Lösung des Ausgleichsproblems mit der Lösung mit minimaler Norm.

10. Überlegen Sie sich den genauen Rechenaufwand, wenn die QR-Zerlegung einer Matrix in oberer Hessenberg-Form mit Householder-Transformationen berechnet wird. Vergleichen Sie das Ergebnis mit dem Aufwand des Algorithmus aus Abschnitt 14.

11. Beweisen Sie die Aussagen aus Bemerkung 15.2.

12. $A \in \mathbb{R}^{m\times n}$, $m \geq n$, besitze vollen Spaltenrang, und es sei $b \in \mathcal{R}(A)$. Gesucht ist die Lösung $x^\dagger$ des linearen Ausgleichsproblems

$$\text{minimiere} \quad \|b - Ax\|_2^2 .$$

(a) Zeigen Sie, daß $x^\dagger = A^T z$, wobei z eine Lösung von $AA^T z = b$ ist.
(b) Wenden Sie das CG-Verfahren auf das Gleichungssystem für z an (vgl. Bemerkung 15.2). Substituieren Sie in Ihrem Algorithmus $x^{(k)} = A^T z^{(k)}$ sowie $s^{(k)} = A^T d^{(k)}$.
(c) Zeigen Sie, daß $x^{(k)}$ die Fehlernorm $\|x^{(k)} - x^\dagger\|_2^2$ in $\mathcal{K}_k(AA^T, b - Ax^{(0)})$ minimiert.

13. Zeigen Sie, daß das modifizierte Gram-Schmidt-Orthogonalisierungsverfahren aus Algorithmus 16.1 äquivalent zu dem klassischen Gram-Schmidt-Verfahren (16.4) ist.

14. Sei $A \in \mathbb{K}^{n\times n}$ nichtsingulär, $r^{(0)} = b - Ax^{(0)}$ und $\mathcal{K}_m(A, r^{(0)}) = \mathcal{K}_{m+1}(A, r^{(0)})$ für ein $m \leq n$. Beweisen Sie, daß das GMRES-Verfahren in diesem Fall nach höchstens m Schritten die exakte Lösung $x^{(m)} = A^{-1}b$ berechnet hat.

15. Zeigen Sie, daß die oberen Hessenberg-Matrizen H_k aus (16.5) rechteckige Tridiagonalmatrizen sind, falls A eine hermitesche Matrix ist. Vergleichen Sie für diesen Fall den Aufwand des GMRES-Verfahrens mit der allgemeinen Aufwandsabschätzung.

16. Das GMRES-Verfahren werde auf ein Gleichungssystem $Ax = b$ mit Koeffizientenmatrix

$$A = \begin{bmatrix} 0 & & & 0 & -a_0 \\ 1 & 0 & & 0 & -a_1 \\ & 1 & \ddots & \vdots & \vdots \\ & & \ddots & 0 & -a_{n-2} \\ 0 & & & 1 & -a_{n-1} \end{bmatrix}$$

angewendet. Zeigen Sie, daß die Residuen $r^{(i)}$, $i = 0, \ldots, n-1$, der Iterierten alle die gleiche Norm haben, falls der Startvektor $x^{(0)}$ die Gleichungen $x_i^{(0)} = b_{i+1} + a_i x_n^{(0)}$, $i = 1, \ldots, n-1$, erfüllt.

IV Nichtlineare Gleichungen

Nach den linearen Gleichungssystemen wenden wir uns nun nichtlinearen Gleichungen in einer und mehreren Variablen zu. Nichtlineare Gleichungen werden zumeist als Nullstellenaufgabe formuliert, d. h. gesucht wird die Nullstelle einer Abbildung

$$F : \mathcal{D}(F) \subset \mathbb{K}^n \to \mathbb{K}^n$$

oder das Minimum von $\|F(x)\|_2$ über $\mathcal{D}(F)$. Hier und im folgenden bezeichnet $\mathcal{D}(F)$ den Definitionsbereich von F, der im weiteren als offen und zusammenhängend vorausgesetzt wird. Durch die Transformation $F(x) = G(x) - y$ kann jede nichtlineare Gleichung $G(x) = y$ unmittelbar in eine solche Nullstellenaufgabe überführt werden.

Da nichtlineare Gleichungen in der Regel nicht geschlossen gelöst werden können, ihre Lösungen also nicht in endlich vielen Schritten berechenbar sind, kommen fast ausschließlich Iterationsverfahren zur Approximation der Lösung zur Anwendung.[1]

Ein Standardwerk zu diesem Thema ist das Buch von Ortega und Rheinboldt [78]. Der wichtige eindimensionale Fall wird sehr ausführlich in [9] behandelt. Aus der weiterführenden Literatur zum nichtlinearen Ausgleichsproblem sind die Bücher [32, 74] empfehlenswert.

17 Konvergenzbegriffe

Zur Erläuterung einiger grundlegender Aussagen über Iterationsverfahren zur Lösung nichtlinearer Gleichungen betrachten wir als Einführung das *Heron-*

[1] Für Iterationsfolgen im $\mathbb{K}^n$ verwenden wir wie in den vorangegangenen Kapiteln die Notation $\{x^{(k)}\}$ mit hochgestelltem und geklammertem Iterationsindex; ein tiefgestellter Index wie in $x_i^{(k)}$ bezeichnet dann die entsprechende Komponente der k-ten Iterierten. Lediglich im Eindimensionalen werden wir hiervon abweichen und den Iterationsindex tiefstellen (ohne Klammern). Da die Dimension des Grundraums immer offensichtlich ist, dürften Mißverständnisse ausgeschlossen sein.

Verfahren zur Berechnung der ν-ten Wurzel einer Zahl $a \in \mathbb{C}\setminus\{0\}$. Ausgehend von einer Startnäherung $x_0 \in \mathbb{C}$ sind die weiteren Iterierten dieses Verfahrens durch die Rekursion

$$x_{k+1} \;=\; \frac{1}{\nu}\left((\nu-1)x_k + a/x_k^{\nu-1}\right), \qquad k = 0, 1, \dots, \tag{17.1}$$

definiert. Wenn die Folge konvergiert, erfüllt der Grenzwert x die Fixpunktgleichung

$$\nu x \;=\; (\nu-1)x + a/x^{\nu-1}$$

und somit ist $x^\nu = a$. Unklar ist jedoch zunächst, ob und gegen welche ν-te Wurzel die Folge konvergiert und wie schnell die Konvergenz gegebenenfalls ist.

Im klassischen Heron-Verfahren – dies entspricht dem Fall $\nu = 2$ – fällt die Beantwortung dieser Fragen noch relativ leicht: Die Transformation

$$z_k \;=\; \frac{x_k - \sqrt{a}}{x_k + \sqrt{a}} \tag{17.2}$$

mit einer beliebig fixierten Wurzel $\sqrt{a}$ führt auf die Rekursion

$$\begin{aligned} z_{k+1} \;&=\; \frac{x_{k+1} - \sqrt{a}}{x_{k+1} + \sqrt{a}} \;=\; \frac{x_k + a/x_k - 2\sqrt{a}}{x_k + a/x_k + 2\sqrt{a}} \;=\; \frac{x_k^2 + a - 2\sqrt{a}\,x_k}{x_k^2 + a + 2\sqrt{a}\,x_k} \\ &=\; \left(\frac{x_k - \sqrt{a}}{x_k + \sqrt{a}}\right)^2 \;=\; z_k^2 . \end{aligned} \tag{17.3}$$

Folglich ist

$$\lim_{k\to\infty} |z_k| \;=\; \begin{cases} 0 & \text{für } |z_0| < 1\,, \\ 1 & \text{für } |z_0| = 1\,, \\ \infty & \text{für } |z_0| > 1\,, \end{cases}$$

d. h. die Folge $\{x_k\}$ konvergiert gegen $\sqrt{a}$, falls $|z_0| < 1$ ist, und gegen $-\sqrt{a}$, falls $|z_0| > 1$ ist; in allen anderen Fällen liegt keine Konvergenz vor.

Die Bedingung

$$|z_0| \;=\; \left|\frac{x_0 - \sqrt{a}}{x_0 + \sqrt{a}}\right| \;\leqslant\; 1$$

für die komplexe Startnäherung x_0 ist äquivalent zu

$$\begin{aligned} |x_0|^2 - 2\,\mathrm{Re}\,\overline{x_0}\sqrt{a} + |a| \;&=\; |x_0 - \sqrt{a}|^2 \\ &\leqslant\; |x_0 + \sqrt{a}|^2 \;=\; |x_0|^2 + 2\,\mathrm{Re}\,\overline{x_0}\sqrt{a} + |a|\,, \end{aligned}$$

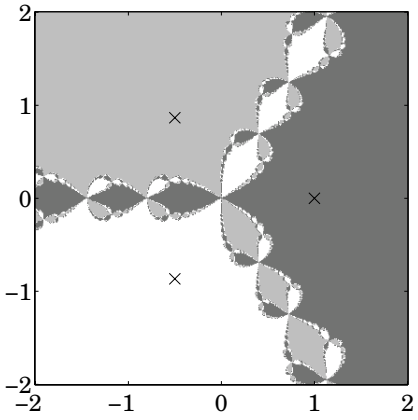


Abb. 17.1:
Konvergenzgebiete des Heron-Verfahrens

also zu $\operatorname{Re}\overline{x_0}\sqrt{a} \gtrless 0$. Für ein positives a ergibt sich somit Konvergenz gegen diejenige Wurzel von a, die dasselbe Vorzeichen hat wie der Realteil von x_0.

Bereits für den Fall $\nu = 3$, also das Heron-Verfahren zur Berechnung der dritten Wurzel einer Zahl a, ist die Konvergenzdiskussion wesentlich komplizierter. In diesem Fall kommen drei Grenzwerte in Frage und es ergibt sich eine scheinbar chaotische Abhängigkeit des Grenzwerts vom Startwert x_0. Abbildung 17.1 zeigt in verschiedenen Graustufen die drei Teilmengen der komplexen Ebene, in denen der Startwert x_0 liegen muß, damit das Verfahren gegen die in der jeweiligen Menge liegende dritte Wurzel der Zahl $a = 1$ konvergiert (die möglichen Grenzwerte sind mit Kreuzen markiert).

17.1 Lokale Konvergenz

Offensichtlich konvergieren Iterationsverfahren für nichtlineare Gleichungen in der Regel nicht mit jedem Startwert gegen die gesuchte Lösung. Andererseits sehen wir zumindest beim Heron-Verfahren, daß das Verfahren gegen eine Lösung $\widehat{x}$ konvergiert, falls der Startwert nur hinreichend nahe bei $\widehat{x}$ gewählt wird. Dies legt die folgende Definition nahe:

Definition 17.1. Ein Iterationsverfahren $x^{(k+1)} = \Phi(x^{(k)})$ mit einer Funktion $\Phi : \mathcal{D}(\Phi) \subset \mathbb{K}^n \to \mathbb{K}^n$ heißt *lokal konvergent* gegen $\widehat{x} \in \mathbb{K}^n$, falls eine Umgebung $\mathcal{U} \subset \mathcal{D}(\Phi)$ um $\widehat{x} \in \mathcal{U}$ existiert, so daß für alle Startvektoren $x^{(0)} \in \mathcal{U}$ die resultierende Folge $\{x^{(k)}\}$ gegen $\widehat{x}$ konvergiert. In diesem Fall spricht man von einem *anziehenden Fixpunkt* $\widehat{x}$ von Φ. Das Iterationsverfahren heißt *global konvergent*, wenn $\mathcal{U}$ der gesamte Raum $\mathbb{K}^n$ ist.

Das Heron-Verfahren (zumindest in den Fällen $\nu = 2$ und $\nu = 3$) ist also für jede ν-te Wurzel von $a \in \mathbb{C}$ lokal konvergent. Eine hinreichende Bedin-

gung für die lokale Konvergenz eines allgemeinen Iterationsverfahrens liefert der folgende Satz.

Satz 17.2. *Die Funktion $\Phi : \mathcal{D}(\Phi) \subset \mathbb{K}^n \to \mathbb{K}^n$ sei stetig differenzierbar und habe einen Fixpunkt $\widehat{x}$ in $\mathcal{D}(\Phi)$. Ferner sei $\|\cdot\|$ eine Norm in $\mathbb{K}^n$ und $|||\cdot|||$ eine damit verträgliche Norm in $\mathbb{K}^{n\times n}$ mit $|||\Phi'(\widehat{x})||| < 1$. Dann ist Φ in einer Umgebung $\mathcal{U}$ von $\widehat{x}$ eine Kontraktion und die Fixpunktiteration*

$$x^{(k+1)} = \Phi(x^{(k)})\,, \qquad k = 0, 1, 2, \ldots$$

lokal konvergent gegen $\widehat{x}$.

Beweis. Wegen der Stetigkeit von Φ' existiert bezüglich der genannten Norm in $\mathbb{K}^n$ eine abgeschlossene Kugel $\mathcal{U} \subset \mathcal{D}(\Phi)$ um $\widehat{x}$ mit Radius $\rho > 0$, so daß

$$|||\Phi'(x)||| \le q < 1 \qquad \text{für alle } x \in \mathcal{U}\,.$$

Aus dem Mittelwertsatz der Differentialrechnung im $\mathbb{K}^n$ (vgl. Heuser [53, Satz 176.4]),

$$\Phi(y) - \Phi(x) = \int_0^1 \Phi'\big(x + t(y-x)\big)(y-x)\,dt\,,$$

folgt

$$\|\Phi(y) - \Phi(x)\| \;\le\; \int_0^1 |||\Phi'\big(x + t(y-x)\big)|||\;\|y-x\|\,dt \;\le\; q\,\|y-x\| \tag{17.4}$$

für alle $x, y \in \mathcal{U}$. Speziell für $y = \widehat{x}$ ergibt dies

$$\|\Phi(x) - \widehat{x}\| \le q\,\|x - \widehat{x}\| \le q\rho < \rho\,,$$

d. h. Φ ist eine kontrahierende Selbstabbildung von $\mathcal{U}$. Die Behauptung folgt somit aus dem Banachschen Fixpunktsatz 7.1. □

Beispiel 17.3. Abbildung 17.2 zeigt eine Aufhängevorrichtung für ein Gewicht der Masse m, etwa ein Wirtshausschild, das am Punkt C befestigt werden soll: Die beiden Stäbe der Länge l_1 bzw. l_2 seien durch ein Gelenk B miteinander verbunden und mit einem Lager A an der Wand befestigt. Wir nehmen an, daß die Stäbe starr sind und eine vernachlässigbare Masse besitzen, so daß äußere Kräfte allenfalls die Winkel der Gelenke verändern können. Wir modellieren diesbezüglich die Steifigkeit der Gelenke durch Drehfedern, die einer Winkeländerung aus der in Abbildung 17.2 links dargestellten Ruhelage entgegenwirken.

Mit dem angehängten Gewicht stellt sich die neue Gleichgewichtslage aus Abbildung 17.2 rechts ein, die durch die beiden eingezeichneten Winkel θ_1 und θ_2 beschrieben wird: θ_1 ist die Auslenkung des ersten Stabs aus der Horizontalen

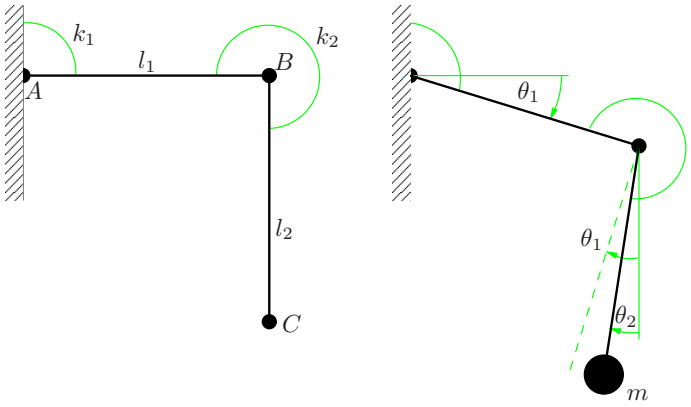


Abb. 17.2: Aufhängevorrichtung mit und ohne Gewicht

und θ_2 die Auslenkung des zweiten Stabs aus der Vertikalen, und zwar jeweils im Uhrzeigersinn. Diese beiden Winkel sollen im folgenden berechnet werden.

Eine äußere Kraft übt auf jedes der beiden Gelenke ein Drehmoment aus, das sich aus der Formel

$$\text{Moment} \;=\; \text{Hebelarm} \;\times\; \text{Kraft} \tag{17.5}$$

errechnet.[2] In unserem Fall ist die einzige äußere Kraft die Gewichtskraft im Punkt C, und die Stärke des Drehmoments ist das Produkt aus der Gewichtskraft mg und dem horizontalen Abstand des Punkts C von dem jeweils betrachteten Gelenk. Das Vorzeichen ist positiv, wenn das Drehmoment im mathematisch positiven Sinn (gegen den Uhrzeigersinn) wirkt.

Im Gleichgewichtszustand ist die Summe aller Drehmomente in jedem Gelenk gleich Null. Da das Gerüst ansonsten starr ist, errechnen sich hieraus die Rückstellmomente der einzelnen Federn, die proportional zu den Winkelauslenkungen sind; die Stärken k_1 und k_2 der beiden Drehfedern sind die jeweiligen Proportionalitätskonstanten.

Für das Lager A ist der horizontale Abstand zum Punkt C durch $l_1 \cos\theta_1 - l_2 \sin\theta_2$ gegeben und somit muß gelten

$$k_1\theta_1 \;-\; mg(l_1 \cos\theta_1 - l_2 \sin\theta_2) \;=\; 0\,.$$

Der Winkel von Gelenk B verringert sich in dem Gleichgewichtszustand um den Winkel $\theta_1 - \theta_2$ gegenüber dem unbelasteten Zustand, woraus die zweite

[2] In der Mechanik werden alle auftretenden Größen in (17.5) als Vektoren interpretiert, und das Drehmoment bezüglich eines Punkts wird durch das *Kreuzprodukt* der beiden Vektoren auf der rechten Seite von (17.5) repräsentiert. Der *Hebelarm* ist dann durch den Vektor vom Bezugspunkt zum Angriffspunkt des Kraftvektors zu ersetzen.

Gleichung

$$-k_2(\theta_1 - \theta_2) \;+\; mgl_2 \sin\theta_2 \;=\; 0$$

folgt. Eine einfache Umformung führt daher auf das folgende nichtlineare Gleichungssystem für die gesuchten Winkel θ_1 und θ_2:

$$\begin{aligned} \theta_1 \;&=\; mg\big(\frac{l_1}{k_1}\cos\theta_1 - \frac{l_2}{k_1}\sin\theta_2\big)\,, \\ \theta_2 \;&=\; mg\big(\frac{l_1}{k_1}\cos\theta_1 - (\frac{l_2}{k_1} + \frac{l_2}{k_2})\sin\theta_2\big)\,. \end{aligned}$$

Für das numerische Beispiel nehmen wir wie in der Skizze an, daß die beiden Stäbe die gleiche Länge haben ($l_1 = l_2 = l$) und die beiden Drehfedern gleich stark sind ($k_1 = k_2 = k$). Dann sind θ_1 und θ_2 Fixpunkte der Abbildung $\Phi : \mathbb{R}^2 \to \mathbb{R}^2$ mit

$$\Phi(\theta_1, \theta_2) \;=\; c \begin{bmatrix} \cos\theta_1 - \sin\theta_2 \\ \cos\theta_1 - 2\sin\theta_2 \end{bmatrix}, \qquad c = mgl/k\,.$$

Die Jacobi-Matrix $\Phi'(\cdot)$ von Φ lautet

$$\Phi'(\theta_1, \theta_2) \;=\; -c \begin{bmatrix} \sin\theta_1 & \cos\theta_2 \\ \sin\theta_1 & 2\cos\theta_2 \end{bmatrix}. \tag{17.6}$$

Im weiteren fassen wir die beiden Winkel in einem Vektor $x = [\theta_1, \theta_2]^T \in \mathbb{R}^2$ zusammen. Aus (17.6) folgt dann für jedes $x \in \mathbb{R}^2$

$$\|\,\Phi'(x)\,\|_\infty \;\le\; 3c\,,$$

und da die Zeilensummennorm mit der Maximumnorm verträglich ist, sieht man wie in (17.4), daß

$$\|\,\Phi(x) - \Phi(y)\,\|_\infty \;\le\; 3c\|x - y\|_\infty \qquad \text{für alle } x, y \in \mathbb{R}^2\,.$$

Folglich ist Φ für $c < 1/3$, also für hinreichend große Federkonstanten, eine Kontraktion des $\mathbb{R}^2$ und besitzt nach dem Banachschen Fixpunktsatz in $\mathbb{R}^2$ genau einen Fixpunkt $\widehat{x} = [\widehat{\theta}_1, \widehat{\theta}_2]^T$. Zudem konvergiert dann die Folge

$$x^{(k+1)} \;=\; \Phi(x^{(k)})$$

für jeden Startvektor $x^{(0)} \in \mathbb{R}^2$ gegen $\widehat{x}$. Für $x^{(0)} = [0, 0]^T$ und $c = 1/4$ fragen wir uns nun, nach wievielen Iterationen der Fehler $\|x^{(k)} - \widehat{x}\|_\infty$ kleiner als

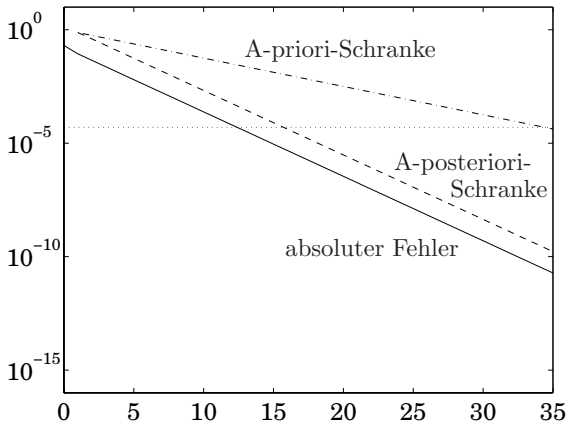


Abb. 17.3: Konvergenzverlauf der Fixpunktiteration

$5 \cdot 10^{-5}$ ist. Mit $q = 3c = 3/4$ und der A-priori-Fehlerabschätzung aus Satz 7.1 erhalten wir die Bedingung

$$\|x^{(k)} - \widehat{x}\|_\infty \;\leq\; \frac{q^k}{1-q}\,\|x^{(1)} - x^{(0)}\|_\infty \;=\; c\,\frac{q^k}{1-q} \;\overset{!}{<}\; 5 \cdot 10^{-5}\,, \tag{17.7}$$

die für $k \geq 35$ erfüllt ist. Abbildung 17.3 zeigt neben der Maximumnorm des Fehlers $x^{(k)} - \widehat{x}$ (durchgezogene Linie) und der obigen A-priori-Abschätzung (Strichpunktlinie) noch die Fehlerschranke

$$\varepsilon_k \;=\; \frac{q}{1-q}\,\|\,x^{(k)} - x^{(k-1)}\,\|_\infty$$

aus der A-posteriori-Abschätzung aus Satz 7.1 (gebrochene Linie).

Offensichtlich wird die Fehlerschranke $5{\cdot}10^{-5}$ bereits nach 13 Iterationen unterschritten. Auf der Grundlage der A-posteriori-Abschätzung würde man nach 16 Iterationen stoppen. Die Diskrepanz zwischen den Steigungen der beiden Fehlerabschätzungen liegt unter anderem daran, daß für q die globale obere Schranke von $\|\Phi'(x)\|_\infty$ für *alle* $x \in \mathbb{R}^2$ eingesetzt wurde. Tatsächlich hängt die asymptotische Konvergenzgeschwindigkeit der Fixpunktiteration nur von der Ableitung Φ' in der Nähe des Fixpunkts ab, vgl. Aufgabe 2. ◊

17.2 Konvergenzordnung

Als nächstes analysieren wir die Konvergenzgeschwindigkeit einer Folge.

Definition 17.4. Für eine reelle nichtnegative Nullfolge $\{\varepsilon_k\}_{k\in\mathbb{N}}$ wird

$$\kappa = \limsup_{k\to\infty} \varepsilon_k^{1/k} \tag{17.8}$$

als *asymptotischer Konvergenzfaktor* bezeichnet. Die Folge $\{\varepsilon_k\}$ heißt *sublinear*, *linear* bzw. *superlinear* konvergent, je nachdem ob $\kappa = 1$, $0 < \kappa < 1$ oder $\kappa = 0$ ist. Gilt im superlinear konvergenten Fall zudem

$$\varepsilon_{k+1} \le C\varepsilon_k^p \qquad \text{für ein } p > 1,\ C > 0 \text{ und fast alle } k \in \mathbb{N}\,, \tag{17.9}$$

dann hat die Folge die *Konvergenzordnung* p. Entsprechend wird die Terminologie für konvergente Folgen $\{x^{(k)}\} \subset \mathbb{K}^n$ mit Grenzwert $\widehat{x}$ über die Fehlerfolge $\varepsilon_k = \|x^{(k)} - \widehat{x}\|$ eingeführt (dabei spielt es keine Rolle, welche Norm im $\mathbb{K}^n$ verwendet wird).

Generell gilt (zumindest asymptotisch): Superlinear konvergente Folgen konvergieren schneller als linear konvergente, und die Konvergenz ist um so schneller, je höher die Konvergenzordnung ist. Man macht sich leicht klar, daß jede Nullfolge $\{\varepsilon_k\}$, die die Bedingung (17.9) erfüllt, superlinear konvergiert.

Bemerkung 17.5. Als Faustregel erwartet man bei einem Iterationsverfahren mit Konvergenzordnung p, daß sich die Anzahl der korrekten Dezimalstellen bei jeder Iteration „ver-p-facht“. ◊

Beispiele. 1. Lineare Konvergenz tritt in der Praxis sehr häufig auf, z. B. im Zusammenhang mit dem Banachschen Fixpunktsatz, vgl. Satz 7.1 (a). Entsprechend ist die Konvergenz in Satz 17.2 ebenfalls (mindestens) lokal linear. Lineare Konvergenz ist uns auch in Abschnitt 8 begegnet: Nach Satz 7.4 konvergieren das Gesamtschrittverfahren, das Einzelschrittverfahren und das symmetrische Gauß-Seidel-Verfahren linear (und global), falls die Spektralradien der jeweiligen Iterationsmatrizen kleiner als Eins sind. Für dieses Beispiel stimmt die Definition des asymptotischen Konvergenzfaktors mit der aus Definition 7.5 überein. Auch das GMRES(ℓ)-Verfahren konvergiert unter den Voraussetzungen von Satz 16.1 mindestens linear, vgl. (16.8).

2. Bei Konvergenzordnung $p = 2$ spricht man von *quadratischer Konvergenz*. Ein Beispiel hierfür ist das Heron-Verfahren mit $\nu = 2$ und $a > 0$: Für die Folge $\{z_k\}$ aus (17.2) ist das wegen (17.3) offensichtlich, für die eigentliche Iterationsfolge $\{x_k\}$ des Heron-Verfahrens ergibt sich unter der Voraussetzung $\sqrt{a}/2 \le x_k \le 2\sqrt{a}$, daß auch $x_{k+1} \in [\sqrt{a}/2, 2\sqrt{a}]$, und in diesem Fall folgt aus (17.3)

$$|x_{k+1} - \sqrt{a}| \;=\; \frac{x_{k+1} + \sqrt{a}}{|x_k + \sqrt{a}|^2}\,|x_k - \sqrt{a}|^2 \;\le\; \frac{3\sqrt{a}}{9/4\,a}\,|x_k - \sqrt{a}|^2\,.$$

Also konvergieren die Iterierten des Heron-Verfahrens lokal quadratisch gegen $\sqrt{a}$. Quadratische Konvergenz wird uns im nächsten Abschnitt im Zusammenhang mit dem *Newton-Verfahren* erneut begegnen.

3. Die *Rayleigh-Quotienten-Iteration* (Algorithmus 25.2) zur Berechnung des kleinsten Eigenwerts einer hermiteschen Matrix ist *kubisch konvergent*, d. h. die Konvergenzordnung ist $p = 3$.

4. Die Konvergenzordnung braucht nicht unbedingt eine ganze Zahl zu sein; ein entsprechendes Beispiel ist das *Sekantenverfahren* aus Abschnitt 18.2.

5. Die Nullfolge $\varepsilon_k = k^{-\nu}$ mit $\nu \in \mathbb{R}^+$ ist ein Beispiel für eine sublinear konvergente Folge. Die Folge $\varepsilon_k = 1/k!$ konvergiert hingegen superlinear, denn

$$\varepsilon_k^{1/k} \;=\; \frac{1}{(1 \cdot 2 \cdots k)^{1/k}} \;\le\; \frac{1}{(k^{k/2})^{1/k}} \;=\; 1/\sqrt{k} \;\longrightarrow\; 0, \qquad k \to \infty\,.$$

Bei dieser Folge divergieren jedoch für jedes $p > 1$ die Quotienten

$$\frac{\varepsilon_{k+1}}{\varepsilon_k^p} \;=\; \frac{1}{k+1}\,\varepsilon_k^{1-p} \;\ge\; \frac{k^{k(p-1)/2}}{k+1} \;\longrightarrow\; \infty\,, \qquad k \to \infty\,,$$

so daß die Bedingung (17.9) für kein $p > 1$ erfüllt ist. ◊

Die Konvergenzordnung einer superlinear konvergenten skalaren Iterationsfolge kann in vielen Fällen mit der folgenden Technik bestimmt werden.

Satz 17.6. *Die Funktion $\Phi : \mathcal{D}(\Phi) \subset \mathbb{K} \to \mathbb{K}$ sei $(p+1)$-mal stetig differenzierbar und habe einen Fixpunkt $\widehat{x} \in \mathcal{D}(\Phi)$. Ferner sei $p \ge 2$ und*

$$0 = \Phi'(\widehat{x}) = \ldots = \Phi^{(p-1)}(\widehat{x}) \qquad \textit{und} \qquad \Phi^{(p)}(\widehat{x}) \neq 0\,. \tag{17.10}$$

Dann ist die Fixpunktiteration $x_{k+1} = \Phi(x_k)$ lokal superlinear konvergent gegen $\widehat{x}$ und die Konvergenzordnung ist genau p.

Beweis. Die lokale Konvergenz folgt unmittelbar aus Satz 17.2, da $\Phi'(\widehat{x}) = 0$ vorausgesetzt ist. Für den Nachweis der Konvergenzordnung entwickeln wir Φ um $\widehat{x}$ in ein Taylorpolynom,

$$\Phi(x_k) \;=\; \Phi(\widehat{x}) + \sum_{i=1}^{p} \frac{\Phi^{(i)}(\widehat{x})}{i!}(x_k - \widehat{x})^i + O\big(|x_k - \widehat{x}|^{p+1}\big)\,,$$

und durch Einsetzen von (17.10) ergibt dies

$$x_{k+1} \;=\; \Phi(x_k) \;=\; \widehat{x} + \frac{\Phi^{(p)}(\widehat{x})}{p!}(x_k - \widehat{x})^p + \xi_k(x_k - \widehat{x})^p$$

mit einem $\xi_k = \xi_k(x_k) = O(|x_k - \widehat{x}|)$. Da $\Phi^{(p)}(\widehat{x}) \neq 0$ vorausgesetzt ist, existiert eine abgeschlossene Kugel $\mathcal{U}$ um $\widehat{x}$, so daß

$$|\xi_k| \;\leq\; \frac{1}{2}\left|\frac{\Phi^{(p)}(\widehat{x})}{p!}\right|, \qquad x_k \in \mathcal{U}\,.$$

Folglich ist

$$\frac{1}{2}\,\frac{|\Phi^{(p)}(\widehat{x})|}{p!}\,|x_k - \widehat{x}|^p \;\leq\; |x_{k+1} - \widehat{x}| \;\leq\; \frac{3}{2}\,\frac{|\Phi^{(p)}(\widehat{x})|}{p!}\,|x_k - \widehat{x}|^p$$

für alle $x_k \in \mathcal{U}$ und die Konvergenzordnung ist demnach genau p. □

Bemerkung 17.7. Die gleiche Beweismethode ergibt: Ist Φ in einer Umgebung des Fixpunkts $\widehat{x}$ p-mal stetig differenzierbar mit $p \geq 2$ und gilt (17.10), wobei auch $\Phi^{(p)}(\widehat{x}) = 0$ zugelassen ist, dann konvergiert $x_{k+1} = \Phi(x_k)$ (lokal) *mindestens* mit Ordnung p gegen $\widehat{x}$. ◊

Beispiel. Wir wenden Satz 17.6 auf das allgemeine Heron-Verfahren (17.1) an. Hier ist $\Phi(x) = ((\nu-1)x + a/x^{\nu-1})/\nu$ mit den Ableitungen

$$\Phi'(x) \;=\; \frac{\nu-1}{\nu}\,(1 - ax^{-\nu}) \qquad \text{und} \qquad \Phi''(x) \;=\; (\nu-1)ax^{-\nu-1}\,.$$

Wegen $\widehat{x}^\nu = a$ ist $\Phi'(\widehat{x}) = 0$ und $\Phi''(\widehat{x}) = (\nu-1)/\widehat{x} \neq 0$. Somit ist die Konvergenz des allgemeinen Heron-Verfahrens (genau) lokal quadratisch. ◊

18 Nullstellenbestimmung reeller Funktionen

Wir suchen nun allgemeine Schemata zur Konstruktion superlinear konvergenter Iterationsverfahren zur Berechnung von Nullstellen nichtlinearer Abbildungen. Dabei beschränken wir uns zunächst auf den Fall einer skalaren Funktion; der mehrdimensionale Fall ist dann das Thema von Abschnitt 19.

18.1 Das Newton-Verfahren

Um die Ergebnisse des vorangegangenen Abschnitts, speziell Satz 17.6, nutzen zu können, bringen wir zunächst die Nullstellenaufgabe $f(x) = 0$ in Fixpunktform. Denkbar ist etwa eine Gleichung der Form

$$x = x + g(x)f(x) =: \Phi(x)$$

mit einer glatten Funktion g, von der wir zunächst nur voraussetzen, daß sie in einer Umgebung der Nullstelle $\widehat{x}$ von f von Null verschieden ist. Im Hinblick

auf Satz 17.6 fordern wir als nächstes, daß $\Phi'(\widehat{x})$ verschwindet. Wegen

$$\Phi'(\widehat{x}) \;=\; 1 + g'(\widehat{x}) \underbrace{f(\widehat{x})}_{=0} + g(\widehat{x}) f'(\widehat{x}) \;\stackrel{!}{=}\; 0$$

führt dies auf die Bedingung

$$g(\widehat{x}) = -1/f'(\widehat{x})\,,$$

was natürlich nur für $f'(\widehat{x}) \neq 0$ möglich ist. Wählen wir speziell $g = -1/f'$, so erhalten wir das *Newton-Verfahren*

$$x_{k+1} = x_k - \frac{f(x_k)}{f'(x_k)}\,, \qquad k = 0, 1, \ldots \tag{18.1}$$

Die rechte Seite von (18.1) ist die Schnittstelle der Tangente

$$y \;=\; f(x_k) + f'(x_k)(x - x_k)$$

an den Graph von f im Punkt $(x_k, f(x_k))$ mit der x-Achse, vgl. Abbildung 18.1.

Beispiel. Die Nullstellen der Funktion

$$f(x) = x^\nu - a\,, \qquad \nu \in \mathbb{N} \setminus \{1\}\,,\; a \in \mathbb{R}^+\,,$$

sind die ν-ten Wurzeln der Zahl a. Das Newton-Verfahren lautet in diesem Fall

$$x_{k+1} \;=\; x_k - \frac{x_k^\nu - a}{\nu x_k^{\nu-1}} \;=\; \frac{\nu - 1}{\nu}\, x_k \;+\; \frac{a}{\nu}\, x_k^{1-\nu}\,, \qquad k = 0, 1, \ldots\,,$$

und entspricht somit dem Heron-Verfahren. ◊

Satz 18.1. *Sei $f \in C^3[a,b]$ und $\widehat{x} \in (a,b)$ mit $f(\widehat{x}) = 0$ und $f'(\widehat{x}) \neq 0$. Dann konvergiert das Newton-Verfahren (mindestens) lokal quadratisch gegen $\widehat{x}$.*

Beweis. Die Behauptung folgt aufgrund der Konstruktion des Verfahrens sofort aus Satz 17.6 und Bemerkung 17.7. □

Bemerkungen. Offensichtlich bleibt die Aussage des Satzes auch für komplexwertige Funktionen f gültig; in diesem Fall kann das Intervall $[a,b]$ durch eine Umgebung $\mathcal{U} \subset \mathbb{C}$ von $\widehat{x}$ ersetzt werden.

Die Differenzierbarkeitsanforderungen an f aus Satz 18.1 können erheblich abgeschwächt werden: Für die Aussage des Satzes ist es bereits hinreichend, daß f' in einer Umgebung von $\widehat{x}$ Lipschitz-stetig ist. Für einen Beweis dieses stärkeren Resultats verweisen wir auf das allgemeine Konvergenzresultat für das Newton-Verfahren im $\mathbb{R}^n$ (Satz 19.1). Die Lipschitz-Stetigkeit von f' ist

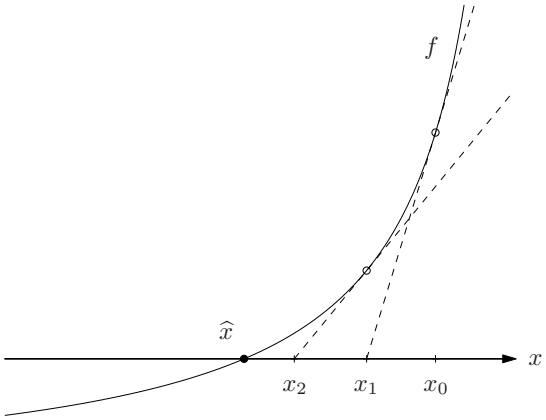


Abb. 18.1: Geometrische Interpretation des Newton-Verfahrens

allerdings eine Minimalforderung, auf die nicht verzichtet werden kann, wie das folgende Beispiel demonstriert: Für

$$f(x) = x + x^\alpha\,, \qquad \alpha > 1\,,$$

mit Nullstelle $\widehat{x} = 0$ ist die Ableitung $f'(x) = 1+\alpha x^{\alpha-1}$ für $\alpha < 2$ im Nullpunkt nicht Lipschitz-stetig. Die Iterationsvorschrift des Newton-Verfahrens lautet für dieses Beispiel

$$x_{k+1} \;=\; x_k - \frac{x_k + x_k^\alpha}{1 + \alpha x_k^{\alpha-1}} \;=\; (\alpha - 1)\,\frac{x_k^\alpha}{1 + \alpha x_k^{\alpha-1}}\,.$$

Für $x_k \in (0,1)$ und $1 < \alpha < 2$ ergibt sich somit

$$0 \;<\; \frac{\alpha-1}{3}\,x_k^\alpha \;\le\; x_{k+1} \;\le\; (\alpha-1)\,x_k^\alpha \;<\; 1\,,$$

das heißt, das Newton-Verfahren konvergiert in diesem Fall genau mit Ordnung $\alpha \in (1,2)$. ◊

Wenn in Satz 18.1 die Voraussetzung $f'(\widehat{x}) \neq 0$ verletzt ist, dann ist die Konvergenzordnung ebenfalls nicht mehr lokal quadratisch. Wir setzen

$$\Phi(x) = x - \frac{f(x)}{f'(x)}\,, \qquad x \neq \widehat{x}\,,$$

und nehmen an, daß $f \in C^{p+1}[a,b]$, $p \in \mathbb{N} \setminus \{1\}$, eine p-fache Nullstelle in $\widehat{x}$ hat, also daß

$$f(\widehat{x}) = f'(\widehat{x}) = \ldots = f^{(p-1)}(\widehat{x}) = 0\,, \qquad f^{(p)}(\widehat{x}) \neq 0\,.$$

Dann ergibt sich

$$f(x) \;=\; \underbrace{f(\widehat{x})}_{=0} + \frac{f^{(p)}(\widehat{x})}{p!}\,(x-\widehat{x})^p \;+\; O(\,|x-\widehat{x}|^{p+1}\,)\,,$$

$$f'(x) \;=\; \underbrace{f'(\widehat{x})}_{=0} + \frac{f^{(p)}(\widehat{x})}{(p-1)!}\,(x-\widehat{x})^{p-1} \;+\; O(\,|x-\widehat{x}|^{p}\,)\,,$$

so daß in einer Umgebung von $\widehat{x}$ die Approximation

$$\Phi(x) = x - \frac{1}{p}\,(x-\widehat{x}) + O(\,|x-\widehat{x}|^2\,)$$

gültig ist. Für das Newton-Verfahren $x_{k+1} = \Phi(x_k)$ erhalten wir daher

$$|x_{k+1}-\widehat{x}| \;=\; |\Phi(x_k)-\widehat{x}| \;=\; (1-1/p)\,|x_k-\widehat{x}| + O(\,|x_k-\widehat{x}|^2\,)\,.$$

Also ist die Konvergenz genau linear, und sie ist um so langsamer, je größer p ist.

Beispiel 18.2. Wir wollen den Zeitpunkt $\widehat{t}$ bestimmen, an dem die Weltbevölkerung die Neun-Milliarden-Grenze überschreitet. Grundlage sei das Verhulst-Modell aus Abschnitt 60, wonach die Funktion

$$f(t) \;=\; a/(1-ce^{-dt})$$

mit den Parametern

$$a = 9.8606\,, \qquad c = -1.1085\cdot 10^{25}\,, \qquad d = 0.029\,,$$

die Bevölkerung (in Milliarden Menschen) zum Zeitpunkt t (in Jahren) angibt, vgl. (60.4), (60.5).

Gesucht ist also eine Nullstelle der Funktion $f(t)-9$. Berücksichtigt man die Identität $f'(t) = f(t)(a-f(t))d/a$, so führt die Newton-Vorschrift auf die Iteration

$$t_{k+1} \;=\; t_k \;+\; \frac{a}{d}\,\frac{f(t_k)-9}{f(t_k)\big(f(t_k)-a\big)}\,, \qquad k \ge 0\,.$$

Wählen wir als Startwert das Jahr $t_0 = 1961$, so ergeben sich die Iterierten t_k aus der folgenden Tabelle:

k	Newton–Verfahren
0	1961
1	2058.05620301193
2	2068.11470840815
3	2069.45919347077
4	2069.48118224579
5	2069.48118803443

Man kann recht gut erkennen, daß sich die Anzahl der korrekten Stellen (die dunklen Ziffern in der Tabelle) in jeder Iteration etwa verdoppelt – dies bestätigt die in Bemerkung 17.5 angeführte Faustregel. Bereits die fünfte Newton-Iterierte liefert das Ergebnis im Rahmen der Maschinengenauigkeit.

Zum Vergleich ziehen wir das einfache *Intervallhalbierungsverfahren* heran: Bei diesem einfachen Verfahren startet man mit einem Intervall, das die Lösung enthält (etwa $[1961, 2200]$), halbiert das Intervall in jeder Iteration und betrachtet im weiteren nur diejenige Intervallhälfte, in der aufgrund des Vorzeichenverhaltens von $f-9$ die gesuchte Lösung liegen muß. Mit diesem Verfahren ergibt sich nach neun Iterationen der Einschluß

$$\widehat{t} \in [\,\mathsf{2069}.2968\ldots,\, \mathsf{2069}.7636\ldots\,]\,,$$

der wenigstens das korrekte Jahr 2069 als Antwort festlegt. ◊

Man beachte, daß in diesem Beispiel die Funktion f nur ein grobes Modell darstellt. Es ist daher eigentlich sinnlos, die Lösung der nichtlinearen Gleichung auf volle Genauigkeit zu bestimmen. Statt dessen würde es beispielsweise ausreichen, lediglich das ungefähre Jahr zu bestimmen, in dem die Neun-Milliarden-Marke überschritten wird (also das Jahr 2069). Hierfür würden bereits zwei oder drei Newton-Iterationen ausreichen.

Damit stellt sich die Frage nach einer effektiven A-posteriori-Abschätzung des Iterationsfehlers. Scharfe Abschätzungen dieser Art sind in der Regel heuristisch begründet: So kann man beispielsweise mutmaßen, daß aufgrund der sehr schnellen Konvergenz

$$|x_k - \widehat{x}| \lesssim |x_k - x_{k+1}| \tag{18.2}$$

eine gute Abschätzung für den k-ten Iterationsfehler ist. In Beispiel 18.2 würde man auf diese Weise nach drei Iterationen vermuten, daß die zweite Iterierte die exakte Zeit auf ein bis zwei Jahre genau angibt.

Die Konvergenz des Newton-Verfahrens ist in der Regel nur lokal. Nur in Ausnahmefällen kann globale Konvergenz garantiert werden; eine solche Ausnahme ist der Fall einer konvexen Funktion f.

Satz 18.3. *Sei $\mathcal{I} \subset \mathbb{R}$ ein Intervall und $f : \mathcal{I} \to \mathbb{R}$ differenzierbar, streng monoton wachsend und konvex mit (eindeutiger) Nullstelle $\widehat{x} \in \mathcal{I}$. Dann konvergiert das Newton-Verfahren für alle $x_0 \in \mathcal{I}$ mit $x_0 \geq \widehat{x}$ monoton gegen $\widehat{x}$.*

Beweis. Wir nehmen an, daß für ein $k \geq 0$ die aktuelle Iterierte x_k größer als $\widehat{x}$ ist, und beweisen die Induktionsbehauptung

$$\widehat{x} \leq x_{k+1} \leq x_k\,. \tag{18.3}$$

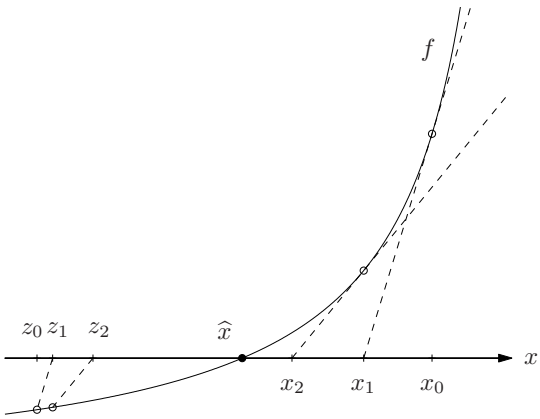


Abb. 18.2: Veranschaulichung von Satz 18.3

Da f differenzierbar ist, folgt aus der Konvexität, daß der Graph von f oberhalb der Tangente an f im Punkt x_k verläuft, also daß

$$f(x) \geq f'(x_k)(x - x_k) + f(x_k)\,.$$

Speziell für $x = x_{k+1}$ wird die rechte Seite dieser Ungleichung Null aufgrund der Newton-Vorschrift (18.1). Also ist $f(x_{k+1})$ nichtnegativ und wegen der Monotonie von f folglich $x_{k+1} \geq \widehat{x}$. Andererseits ist wegen (18.1) $x_{k+1} \leq x_k$, da nach Voraussetzung bzw. Induktionsvoraussetzung sowohl $f(x_k)$ als auch $f'(x_k)$ nichtnegativ sind. Die Konvergenz folgt nun aus der Monotonie und der Beschränktheit der Iterierten. Wegen (18.1) muß der Grenzwert eine Nullstelle von f sein. Damit ist die Behauptung (18.3) vollständig bewiesen. □

Bemerkung. Entsprechende Resultate gelten für konkave und für monoton fallende Funktionen. Beispielsweise konvergiert das Newton-Verfahren für jede konkave monoton wachsende Funktion f mit Nullstelle $\widehat{x}$, wenn $x_0 \leq \widehat{x}$ gewählt wird. Zum Beweis wendet man Satz 18.3 auf die Gleichung $-f(-x) = 0$ an. ◊

Die Voraussetzungen von Satz 18.3 sind insbesondere bei der Berechnung der größten Nullstelle $\widehat{x}$ eines reellen Polynoms $f(x) = x^n + a_{n-1}x^{n-1} + \ldots + a_0$ mit n reellen Nullstellen erfüllt. Für $\mathcal{I}$ kann hierbei etwa das Intervall $[\widehat{x}, \infty)$ gewählt werden. Eine weitere Anwendung dieses Satzes folgt in Abschnitt 18.3.

Abbildung 18.2 veranschaulicht das Resultat aus Satz 18.3. Man macht sich leicht klar, daß unter der Voraussetzung des Satzes eine Startnäherung $x_0 < \widehat{x}$, $x_0 \in \mathcal{I}$, auf $x_1 > \widehat{x}$ abgebildet wird. Falls x_1 weiterhin in $\mathcal{I}$ liegt, bilden schließlich die Iterierten x_k, $k \geq 1$, obere Schranken für die Nullstelle $\widehat{x}$. Man

kann darüber hinaus in einfacher Weise eine Folge monoton wachsender unterer Schranken definieren, vgl. Aufgabe 5: Dazu sucht man eine Näherung $z_0 < \widehat{x}$, $z_0 \in \mathcal{I}$, und konstruiert die Folge $\{z_k\}$ durch

$$z_{k+1} = z_k - \frac{f(z_k)}{f'(x_k)}\,. \tag{18.4}$$

Dabei wird in (18.4) im Nenner wieder die Newton-Iterierte x_k bei der Auswertung der Ableitung eingesetzt. Die ersten Iterierten der Folge $\{z_k\}$ sind ebenfalls in Abbildung 18.2 eingezeichnet.

18.2 Das Sekantenverfahren

Der Aufwand bei der Implementierung des Newton-Verfahrens (18.1) steckt in der Auswertung von f und von f'. In der Praxis ist die Funktion f' oft nicht explizit bekannt oder um ein Vielfaches komplizierter als die Funktion f. Daher ersetzt man gelegentlich die Ableitung $f'(x_k)$ in (18.1) durch einen Differenzenquotienten, etwa

$$f'(x_k) \;\approx\; \frac{f(x_k) - f(x_{k-1})}{x_k - x_{k-1}}\,.$$

Ausgehend von zwei Startnäherungen x_0 und x_1 erhält man so für $k \geq 1$ die Iterationsvorschrift des *Sekantenverfahrens*:

$$x_{k+1} \;=\; x_k - \frac{x_k - x_{k-1}}{f(x_k) - f(x_{k-1})}\, f(x_k) \;=\; \frac{x_{k-1} f(x_k) - x_k f(x_{k-1})}{f(x_k) - f(x_{k-1})}\,. \tag{18.5}$$

Der Name „Sekantenverfahren“ beruht auf der geometrischen Interpretation in Abbildung 18.3:

$$y \;=\; f(x_k) \;+\; \frac{f(x_k) - f(x_{k-1})}{x_k - x_{k-1}}\,(x - x_k)$$

ist die Gleichung der Sekante an den Graph der Funktion f durch die Punkte $(x_{k-1}, f(x_{k-1}))$ und $(x_k, f(x_k))$. Die Nullstelle dieser Sekante ist die neue Iterierte x_{k+1} aus (18.5).

Satz 18.4. *f sei zweimal stetig differenzierbar in $[a, b]$ und habe eine Nullstelle $\widehat{x} \in (a, b)$ mit $f'(\widehat{x}) \neq 0$ und $f''(\widehat{x}) \neq 0$. Dann konvergiert das Sekantenverfahren lokal gegen $\widehat{x}$ mit der genauen Konvergenzordnung*

$$p = \frac{1}{2}(1 + \sqrt{5}) = 1.61803\ldots$$

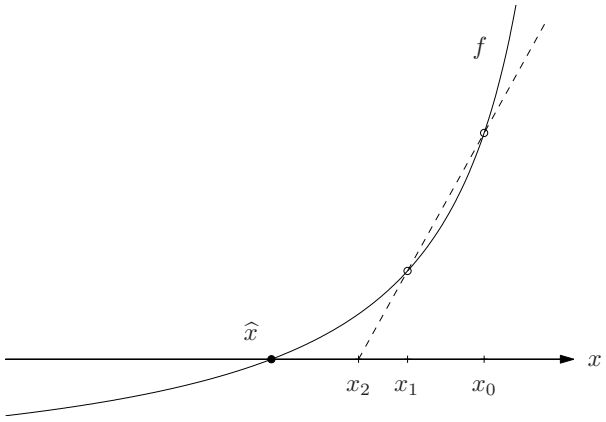


Abb. 18.3: Geometrische Interpretation des Sekantenverfahrens

Beweis. Aufgrund der Voraussetzung an $f'(\widehat{x})$ ist f' in einer Umgebung der Nullstelle von Null verschieden, also f dort injektiv und das Sekantenverfahren wohldefiniert. Den Konvergenzbeweis unterteilen wir in drei Schritte.

1. Aus (18.5) folgt für den Fehler

$$e_k = \widehat{x} - x_k$$

für $k \geq 1$ die Rekursion

$$e_{k+1} \;=\; e_k - \frac{e_k - e_{k-1}}{f(x_k) - f(x_{k-1})}\, f(x_k) \;=\; \frac{e_{k-1} f(x_k) - e_k f(x_{k-1})}{f(x_k) - f(x_{k-1})}\,,$$

also

$$\frac{e_{k+1}}{e_k e_{k-1}} \;=\; \frac{1}{f(x_k) - f(x_{k-1})}\Big(\frac{f(x_{k-1})}{x_{k-1} - \widehat{x}} - \frac{f(x_k)}{x_k - \widehat{x}}\Big) \;=\; \frac{g(x_k) - g(x_{k-1})}{f(x_k) - f(x_{k-1})}$$

mit

$$g(x) = -\frac{f(x)}{x - \widehat{x}} \qquad \text{und} \qquad g'(x) = \frac{-f'(x)(x - \widehat{x}) + f(x)}{(x - \widehat{x})^2}\,.$$

(Die Funktion g kann durch $g(\widehat{x}) = -f'(\widehat{x})$ und $g'(\widehat{x}) = -\frac{1}{2} f''(\widehat{x})$ stetig differenzierbar in den Punkt $x = \widehat{x}$ fortgesetzt werden.) Aus dem (verallgemeinerten) Mittelwertsatz (vgl. Heuser [53, Abschnitt 49.9]) ergibt sich die Existenz eines ξ_k zwischen x_k und x_{k-1}, so daß

$$\frac{e_{k+1}}{e_k e_{k-1}} \;=\; \frac{g'(\xi_k)}{f'(\xi_k)} \;=\; \frac{1}{f'(\xi_k)}\, \frac{f(\xi_k) + f'(\xi_k)(\widehat{x} - \xi_k)}{(\xi_k - \widehat{x})^2}\,.$$

Der Zähler läßt sich dabei als Taylor-Restglied interpretieren, so daß wir schließlich die Darstellung

$$\frac{e_{k+1}}{e_k e_{k-1}} \;=\; \frac{1}{f'(\xi_k)}\,\frac{f(\xi_k)+f'(\xi_k)(\widehat{x}-\xi_k)}{(\xi_k-\widehat{x})^2} \;=\; -\frac{1}{2}\,\frac{f''(\zeta_k)}{f'(\xi_k)} \tag{18.6}$$

mit einem geeigneten ζ_k zwischen $\widehat{x}$ und ξ_k erhalten. In einem hinreichend kleinen Intervall um $\widehat{x}$ ist die rechte Seite von (18.6) betragsmäßig beschränkt durch $C > 0$ und daher

$$|e_{k+1}| \;\le\; \big(\,C|e_{k-1}|\,\big)\;|e_k|\,, \qquad k \ge 1\,.$$

Hieraus folgt, daß der Fehler des Sekantenverfahrens monoton und mindestens linear gegen Null konvergiert, wenn x_0 und x_1 nur hinreichend nahe bei $\widehat{x}$ liegen (nämlich so nahe, daß $|e_0|$ und $|e_1|$ kleiner als $1/C$ sind).

2. Zum Nachweis der Konvergenzordnung setzen wir für $k \in \mathbb{N}$

$$\varepsilon_k = \frac{|e_k|}{|e_{k-1}|^p} \qquad \text{mit} \qquad p = \frac{1}{2}\,(1+\sqrt{5}) \tag{18.7}$$

und leiten eine Rekursion für $\gamma_k = \log \varepsilon_k$ her. Wegen

$$\frac{1}{p} \;=\; \frac{2}{\sqrt{5}+1} \;=\; \frac{1}{2}\,(\sqrt{5}-1) \;=\; p-1$$

ergibt sich aus (18.6)

$$\varepsilon_{k+1} \;=\; \frac{|e_{k+1}|}{|e_k|^p} \;=\; \frac{|e_{k+1}|}{|e_k|}\,|e_k|^{1-p} \;=\; \alpha_k\,|e_{k-1}|\,|e_k|^{-1/p} \;=\; \alpha_k \varepsilon_k^{-1/p}$$

mit

$$\alpha_k \;=\; |f''(\zeta_k)|\,/\,|2f'(\xi_k)|\,.$$

Dies ist gleichbedeutend mit

$$\gamma_{k+1} = \log \alpha_k - \gamma_k/p\,, \qquad \gamma_k = \log \varepsilon_k\,, \quad k \in \mathbb{N}\,. \tag{18.8}$$

3. Die Rekursion (18.8) läßt sich auflösen und ergibt

$$\begin{aligned}\gamma_{k+1} \;&=\; \log \alpha_k - \gamma_k/p \;=\; \log \alpha_k - \big(\log \alpha_{k-1}\big)/p + \gamma_{k-1}/p^2 \\ &=\; \ldots \;=\; \sum_{j=1}^{k} \big(-\frac{1}{p}\big)^{k-j} \log \alpha_j \;+\; \big(-\frac{1}{p}\big)^k \gamma_1\,.\end{aligned}$$

Nach dem ersten Beweisschritt existiert ein Intervall um $\widehat{x}$, in dem die Konvergenz monoton ist und in dem f' und f'' strikt positiv und beschränkt sind. Sofern x_0 und x_1 in diesem Intervall liegen, existiert ein $a > 0$ mit

$$|\log \alpha_j| \le a < \infty \qquad \text{für alle } j \in \mathbb{N}\,,$$

und wegen $1/p < 1$ folgt

$$|\gamma_{k+1}| \;<\; |\gamma_1| + a \sum_{j=1}^{\infty} \frac{1}{p^j} \;=:\; c \;<\; \infty \tag{18.9}$$

für alle $k \in \mathbb{N}_0$. Somit ist $\varepsilon_k = e^{\gamma_k} \in (e^{-c}, e^c)$ und wir erhalten aus (18.7)

$$e^{-c}\,|e_{k-1}|^p \;\le\; |e_k| \;\le\; e^c\,|e_{k-1}|^p\,, \qquad k \in \mathbb{N}\,.$$

Mit anderen Worten: Das Sekantenverfahren hat genau Konvergenzordnung p.

□

Gemäß der Faustregel aus Bemerkung 17.5 benötigt das Newton-Verfahren bei einem vernünftigen Startwert etwa vier Iterationen, um den Grenzwert auf sechzehn Stellen genau zu berechnen. Entsprechend würde man bei dem Sekantenverfahren etwa sechs Iterationen erwarten ($\log 16/\log p \approx 5.76$). Dazu müssen jedoch je vier Funktionswerte von f und f' beim Newton-Verfahren gegenüber sieben Funktionswerten von f beim Sekantenverfahren berechnet werden. Da zudem die Auswertung von f' häufig aufwendiger ist als die von f, erweist sich das Sekantenverfahren in der Praxis oft als konkurrenzfähig zum Newton-Verfahren.

Allerdings ist das Sekantenverfahren nicht so stabil wie das Newton-Verfahren, da bei der Auswertung von (18.5) die Gefahr der Auslöschung im Nenner besteht.

Beispiel. Das Sekantenverfahren liefert in Beispiel 18.2 bei Startwerten $t_0 = 1961$ und $t_1 = 2200$ nach zehn Iterationen die Lösung im Rahmen der Maschinengenauigkeit.

k	Newton-Verfahren	Sekantenverfahren
0	1961	1961
1	2058.05620301193	2200
2	2068.11470840815	2170.41324960649
3	2069.45919347077	1340.09208796042
4	2069.48118224579	2101.82728726442
5	2069.48118803443	2061.36658860449
6		2072.99776293826
7		2069.81288232448
8		2069.46691136301
9		2069.48124481884
10		2069.48118804413
11		2069.48118803443

Dabei setzt die eigentliche schnelle Konvergenz erst ab etwa der fünften Iteration ein, da sich der zweite Startwert t_1 als sehr ungünstig erweist. ◊

18.3 Spezialfälle

In späteren Abschnitten stellt sich wiederholt die Aufgabe, spezielle rationale nichtlineare Gleichungen zu lösen. Beispielhaft sei in diesem Abschnitt die Gleichung

$$r(x) \;=\; \sum_{i=1}^{n} \frac{z_i^2}{(d_i + x)^2} \;\stackrel{!}{=}\; \rho \tag{18.10}$$

angeführt, wobei z_i, d_i, $i = 1, \ldots, n$, und ρ positiv sein sollen. Ferner sei angenommen, daß die d_i streng monoton fallend angeordnet sind, d. h. es ist $d_1 > d_2 > \ldots > d_n$. Offensichtlich hat r an jeder der n Abszissen $x = -d_i < 0$ einen Pol und konvergiert für $x \to \pm\infty$ gegen Null. Für $x > -d_n$ sind die einzelnen Summanden in der Definition (18.10) von r streng monoton fallend und der Wertebereich von r umfaßt alle positiven Zahlen. Daher gibt es genau eine Lösung $\widehat{x}$ von (18.10) im Intervall $(-d_n, \infty)$.

Gleichungen der Form (18.10) müssen in jedem Iterationsschritt des *Levenberg-Marquardt-Verfahrens* gelöst werden, vgl. Abschnitt 21. Dort ist die Lösung $\widehat{x}$ nur von Bedeutung, sofern sie positiv ist. Wegen der Monotonie von r hat (18.10) genau dann eine positive Lösung, wenn die Bedingung

$$r(0) > \rho \tag{18.11}$$

erfüllt ist.

Man rechnet leicht nach (vgl. auch Abbildung 18.4), daß die Funktion $r - \rho$ in dem fraglichen Bereich $x > -d_n$ nicht nur streng monoton fallend sondern auch noch konvex ist. Nach Satz 18.3 und der nachfolgenden Bemerkung konvergieren daher die Iterierten des Newton-Verfahrens, angewandt auf die Funktion $f(x) = r(x) - \rho$ mit Startwert $x_0 = 0$, monoton und lokal quadratisch gegen die gesuchte Lösung von (18.10).

Beispiel 18.5. Abbildung 18.4 zeigt links den Graph der Funktion r aus (18.10) für die Parameter

$$\begin{aligned} d_1 &= 5\,, & d_2 &= 1\,, & d_3 &= 0.5\,, & d_4 &= 0.1\,, \\ z_1 &= 1\,, & z_2 &= 0.1\,, & z_3 &= 2\,, & z_4 &= 0.1\,. \end{aligned}$$

Die Singularitäten sind durch gepunktete Asymptoten dargestellt; die Singularität „am weitesten rechts“ gehört zu $x = -d_4 = -0.1$. Im rechten Teil der Abbildung erkennt man die einzelnen Newton-Schritte mit den jeweiligen Iterierten und die zugehörigen Tangenten an den Graph von $r - \rho$. Für die rechte Seite von (18.10) wurde dabei $\rho = 0.01$ verwendet. Die Newton-Iteration zeigt

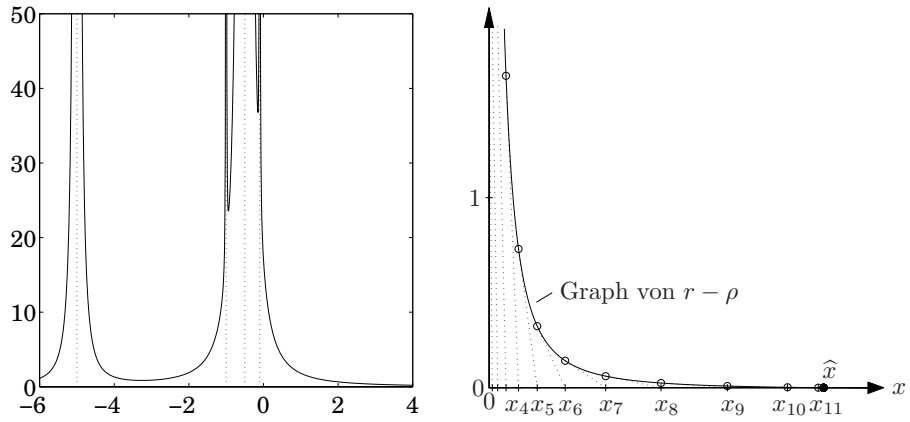


Abb. 18.4: Die Funktion r und die Iterierten des Newton-Verfahrens für $r(x) - \rho = 0$

nicht die erwartet schnelle Konvergenz, denn wegen der sehr steilen Tangenten für x nahe bei Null nähern sich die Iterierten nur sehr langsam der Lösung $\widehat{x}$ und dem Einzugsbereich des quadratischen Konvergenzverlaufs. ◊

Die langsame Konvergenz des Newton-Verfahrens läßt sich dadurch erklären, daß die Funktion r nur in sehr kleinen Intervallen gut durch ihre Tangenten angenähert wird. Um das Verfahren zu verbessern, nutzen wir das asymptotische Verhalten

$$r(x) \sim \|z\|_2^2 / x^2 , \qquad x \to \infty ,$$

aus, wobei z hier für den Vektor $[z_1, \ldots, z_n]^T \in \mathbb{R}^n$ steht. Für große x verhält sich demnach der Graph der Funktion $1/\sqrt{r}$ im wesentlichen wie eine Gerade. Wie schnell sich diese Asymptotik durchsetzt, hängt von den z_i und der relativen Lage der d_i untereinander ab. Für das spezielle r aus Beispiel 18.5 ist der Graph von $1/\sqrt{r}$ auf der linken Seite von Abbildung 18.5 dargestellt.

Es bietet sich daher an, das Newton-Verfahren auf die Nullstellengleichung

$$g(x) = \frac{1}{\sqrt{r(x)}} - \frac{1}{\sqrt{\rho}} \stackrel{!}{=} 0$$

anzuwenden. Dies führt auf das *Verfahren von Hebden*, vgl. (18.13) weiter unten, für das mit dem Startwert $x_0 = 0$ nun noch globale und lokal quadratische Konvergenz nachgewiesen werden soll. Da r über der positiven Halbachse streng monoton fällt, ist g in diesem Bereich streng monoton wachsend. Die ersten beiden Ableitungen von g lauten

$$g'(x) = -\frac{r'(x)}{2r^{3/2}(x)} \qquad \text{und} \qquad g''(x) = \frac{3(r'(x))^2 - 2r(x)r''(x)}{4r^{5/2}(x)} .$$

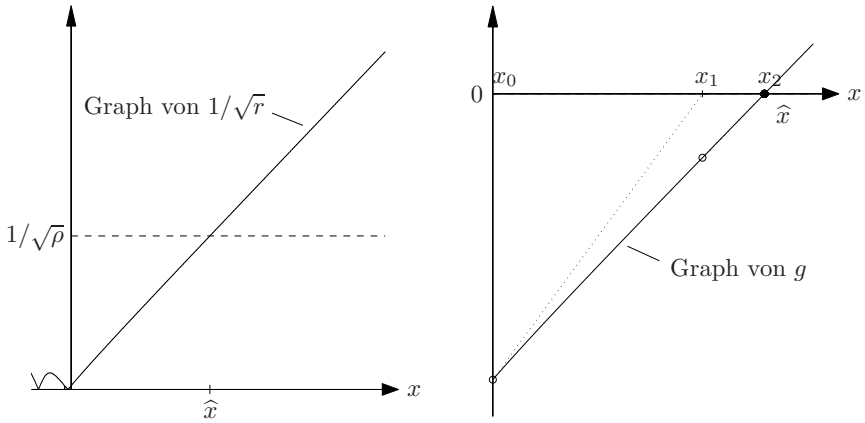


Abb. 18.5: Graph von $1/\sqrt{r}$ (links) und die Hebden-Iterierten (rechts)

Für die Vorzeichendiskussion der zweiten Ableitung benötigen wir die Ungleichung

$$\begin{aligned}\sum_{i=1}^{n} \frac{z_i^2}{(d_i+x)^3} &= \sum_{i=1}^{n} \frac{z_i}{d_i+x}\,\frac{z_i}{(d_i+x)^2} \\ &\le \Big(\sum_{i=1}^{n} \frac{z_i^2}{(d_i+x)^2}\Big)^{1/2}\Big(\sum_{i=1}^{n} \frac{z_i^2}{(d_i+x)^4}\Big)^{1/2},\end{aligned} \tag{18.12}$$

die aus der Cauchy-Schwarz-Ungleichung im $\mathbb{R}^n$ folgt. Damit ergibt sich

$$\begin{aligned}4r^{5/2}(x)g''(x) &= 3\,(r'(x))^2 - 2r(x)r''(x) \\ &= 12\Big(\sum_{i=1}^{n} \frac{z_i^2}{(d_i+x)^3}\Big)^2 - 12\Big(\sum_{i=1}^{n} \frac{z_i^2}{(d_i+x)^2}\Big)\Big(\sum_{i=1}^{n} \frac{z_i^2}{(d_i+x)^4}\Big) \le 0\,.\end{aligned}$$

Mit anderen Worten: Die Funktion g ist über $\mathbb{R}^+$ streng monoton wachsend und konkav. Unter der Voraussetzung (18.11) folgt somit aus Satz 18.3 bzw. der daran anschließenden Bemerkung die Konvergenz des Hebden-Verfahrens mit dem Startwert $x_0 = 0$.

Für die Iterierten des Hebden-Verfahrens ergibt sich die Rekursion

$$\begin{aligned}x_{k+1} &= x_k - \frac{g(x_k)}{g'(x_k)} = x_k + \Big(\frac{1}{\sqrt{r(x_k)}} - \frac{1}{\sqrt{\rho}}\Big)\frac{2r^{3/2}(x_k)}{r'(x_k)} \\ &= x_k + \frac{2r(x_k)\big(1-\sqrt{r(x_k)/\rho}\big)}{r'(x_k)}\,, \qquad k = 0,1,2,\ldots\end{aligned} \tag{18.13}$$

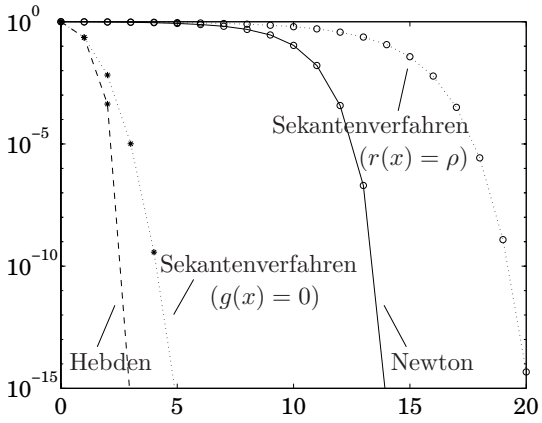


Abb. 18.6: Relative Fehler der verschiedenen Verfahren

Die berechneten Näherungen für Beispiel 18.5 sind in Abbildung 18.5 dargestellt. Für einen Vergleich mit den Näherungen des Newton-Verfahrens aus Beispiel 18.5 beachte man Abbildung 18.6: Die durchgezogene Kurve gehört zum Newton-Verfahren, die gebrochene Kurve zum Hebden-Verfahren. Bei dem bereits angesprochenen Levenberg-Marquardt-Verfahren kommt es nicht auf eine sehr genaue Näherung von $\widehat{x}$ an; ein bis zwei Dezimalziffern Genauigkeit sind völlig ausreichend. Das Hebden-Verfahren erreicht diese Genauigkeit mit nur zwei Iterationen, das Newton-Verfahren benötigt im Vergleich hierzu über zehn Iterationsschritte. Der Aufwand je Iteration wird bei beiden Verfahren durch die $O(n)$ Operationen zur Berechnung von r und r' dominiert und ist daher im wesentlichen gleich.

Anstelle des Newton-Verfahrens kann natürlich auch das Sekantenverfahren zur Lösung von $g(x) = 0$ verwendet werden. Steht außer x_0 noch eine vernünftige Approximation x_1 von $\widehat{x}$ zur Verfügung, wird man anhand des Graphs von $1/\sqrt{r}$ in Abbildung 18.5 ebenfalls sehr schnelle Konvergenz erwarten. Zur Illustration zeigen die gepunkteten Kurven in Abbildung 18.6 das Verhalten des Sekantenverfahrens für obiges Beispiel, wenn für x_1 die erste Newton-Iterierte gewählt wird.

Bemerkung. Das Hebden-Verfahren kann auch so interpretiert werden, daß die Funktion r aus (18.10) in jeder Iteration durch eine rationale Funktion der Form

$$h(x) \;=\; \frac{\zeta}{(\delta + x)^2}$$

approximiert wird; ζ und δ werden dabei so bestimmt, daß sich r und h in x_k

berühren, also daß gilt

$$h(x_k) \,=\, r(x_k)\,, \qquad h'(x_k) \,=\, r'(x_k)\,.$$

Die nächste Iterierte x_{k+1} löst dann die Gleichung $h(x) = \rho$, vgl. Aufgabe 8. Dies ist eine weitere Verbindung zwischen Newton-Verfahren und Hebden-Verfahren: Beim Newton-Verfahren approximiert man die Funktion r durch eine *Gerade*, die die Funktion r im Punkt x_k berührt, beim Hebden-Verfahren wählt man anstelle der Geraden eine einfache rationale Funktion. ◊

Beispiel 18.6. Einen ganz ähnlichen Trick wendet man an, um die sogenannte *Säkulargleichung* oder *charakteristische Gleichung*

$$r(x) \,=\, 1 \,+\, \sum_{i=1}^{n} \frac{z_i^2}{d_i - x} \,\stackrel{!}{=}\, 0 \tag{18.14}$$

zu lösen, die bei gewissen Eigenwertproblemen eine Rolle spielt, siehe etwa Abschnitt 29. Hierbei sind z_i und d_i, $i = 1, \ldots, n$, vorgegebene reelle Zahlen, wobei die d_i paarweise verschieden und streng monoton fallend angeordnet seien.

Die Nullstellenaufgabe (18.14) ähnelt sehr dem oben ausführlich diskutierten Problem, allerdings ergeben sich zusätzliche Schwierigkeiten, wenn eine Nullstelle *zwischen* zwei Polstellen gesucht ist. Dann konvergiert das entsprechend modifizierte Newton-Verfahren in der Regel nicht mehr für jeden Startwert. Für eine ausführliche Behandlung dieses Problems verweisen wir auf das Buch von Demmel [22]. ◊

19 Das Newton-Verfahren im $\mathbb{R}^n$

Das Newton-Verfahren (18.1) läßt sich formal unmittelbar auf die Nullstellenaufgabe für eine Funktion $F : \mathcal{D}(F) \subset \mathbb{R}^n \to \mathbb{R}^n$ verallgemeinern:

$$x^{(k+1)} = x^{(k)} - F'(x^{(k)})^{-1} F(x^{(k)})\,. \tag{19.1}$$

Hierbei ist $F'(x)$ die Jacobi-Matrix

$$F'(x) \,=\, \Big[\, \frac{\partial F_i}{\partial x_j}\,(x) \Big]_{ij} \,\in\, \mathbb{R}^{n\times n}\,.$$

Die Bedingung $f'(\widehat{x}) \neq 0$ aus Abschnitt 18 muß durch die entsprechende Bedingung, daß $F'(\widehat{x})$ invertierbar ist, ersetzt werden.

Initialisierung: $x^{(0)} \in \mathcal{D}(F)$ sei eine Approximation einer Nullstelle von F

```
for k = 0, 1, 2, ... do
  F'(x^(k)) h^(k) = -F(x^(k))      % löse lineares Gleichungssystem, vgl. Kapitel II
  x^(k+1) = x^(k) + h^(k)
  if x^(k+1) ∉ D(F) then
    error     % „Overflow“
  end if
  überprüfe Konvergenz, vgl. Bemerkung 19.2, ggf. Abbruch wegen Divergenz
until stop      % end for
```

Algorithmus 19.1: Newton-Verfahren im $\mathbb{R}^n$

Für die Implementierung des Newton-Verfahrens verwendet man anstelle von (19.1) zumeist die äquivalente Formulierung

$$F(x^{(k)}) + F'(x^{(k)})(x^{(k+1)} - x^{(k)}) \;=\; 0\,, \tag{19.2}$$

bei der ein lineares Gleichungssystem mit der Jacobi-Matrix $F'(x^{(k)})$ zu lösen ist, vgl. Algorithmus 19.1. An (19.2) erkennt man, daß $x^{(k+1)}$ eine Nullstelle des ersten Taylorpolynoms von F um $x^{(k)}$ darstellt. Durch (19.2) wird also die ursprüngliche *nichtlineare* Gleichung $F(x) = 0$ durch die *lokale Linearisierung*

$$F(x^{(k)}) + F'(x^{(k)})(x - x^{(k)}) \;=\; 0$$

ersetzt. Dieses Konzept der Linearisierung ist exemplarisch für die numerische Behandlung vieler nichtlinearer Gleichungen und wird uns immer wieder begegnen.

Wir beweisen nun den folgenden Konvergenzsatz für Algorithmus 19.1.

Satz 19.1. *$\|\cdot\|$ und $|\!|\!|\cdot|\!|\!|$ seien verträgliche Normen in $\mathbb{K}^n$ bzw. $\mathbb{K}^{n\times n}$, die Funktion $F : \mathcal{D}(F) \subset \mathbb{R}^n \to \mathbb{R}^n$ sei stetig differenzierbar und habe eine Nullstelle $\widehat{x}$ im Innern von $\mathcal{D}(F)$. Ferner sei $F'(x)$ invertierbar für alle x aus einer Kugel $\mathcal{U} \subset \mathcal{D}(F)$ um $\widehat{x}$ und es gelte*

$$|\!|\!|F'(x)^{-1}\big(F'(y) - F'(x)\big)|\!|\!| \le L\,\|y - x\| \tag{19.3}$$

für alle $x, y \in \mathcal{U}$ mit einem festen $L > 0$. Dann konvergiert das Newton-Verfahren (mindestens) lokal quadratisch gegen $\widehat{x}$.

Beweis. Wegen $F(\widehat{x}) = 0$ gilt

$$\begin{aligned} x^{(k+1)} - \widehat{x} \;&=\; x^{(k)} - F'(x^{(k)})^{-1}F(x^{(k)}) - \widehat{x} \\ &=\; x^{(k)} - \widehat{x} - F'(x^{(k)})^{-1}\big(F(x^{(k)}) - F(\widehat{x})\big) \\ &=\; F'(x^{(k)})^{-1}\big(F(\widehat{x}) - F(x^{(k)}) - F'(x^{(k)})(\widehat{x} - x^{(k)})\big)\,. \end{aligned}$$

Mit dem Mittelwertsatz folgt für $x^{(k)} \in \mathcal{U}$ mit $h = \widehat{x} - x^{(k)}$ somit

$$\begin{aligned} x^{(k+1)} - \widehat{x} \;&=\; F'(x^{(k)})^{-1}\Big(\int_0^1 F'(x^{(k)} + th)h\,dt \;-\; F'(x^{(k)})h\Big) \\ &=\; \int_0^1 F'(x^{(k)})^{-1}\big(F'(x^{(k)} + th) - F'(x^{(k)})\big)h\,dt\,. \end{aligned}$$

Folglich gilt

$$\begin{aligned} \|x^{(k+1)} - \widehat{x}\| \;&\le\; \int_0^1 \|\!| F'(x^{(k)})^{-1}\big(F'(x^{(k)} + th) - F'(x^{(k)})\big) |\!\| \, \|h\|\,dt \\ &\le\; L\,\|h\|^2 \int_0^1 t\,dt \;=\; \frac{L}{2}\,\|x^{(k)} - \widehat{x}\|^2\,, \end{aligned}$$

also

$$\|x^{(k+1)} - \widehat{x}\| \;\le\; \frac{L}{2}\,\|x^{(k)} - \widehat{x}\|^2\,. \tag{19.4}$$

Speziell für $\|x^{(k)} - \widehat{x}\| \le \rho < 2/L$ folgt

$$\|x^{(k+1)} - \widehat{x}\| \;\le\; \big(\,\frac{L}{2}\,\|x^{(k)} - \widehat{x}\|\,\big)\,\|x^{(k)} - \widehat{x}\| \;\le\; \rho\,\frac{L}{2}\,\|x^{(k)} - \widehat{x}\|\,,$$

d. h. die Newton-Iteration (19.1) ist eine kontrahierende Selbstabbildung jeder Kugel in $\mathcal{D}(F)$ um $\widehat{x}$ mit Radius kleiner als $2/L$ und somit nach Satz 7.1 bei entsprechender Wahl des Startvektors konvergent. Die quadratische Konvergenz folgt aus (19.4). □

Bemerkung. (19.3) ist eine Lipschitz-Bedingung an $F'(\,\cdot\,)$. Die Konstante L ist dabei unabhängig von möglichen linearen Transformationen

$$\widetilde{F}(x) = AF(x) \qquad \text{mit nichtsingulärem } A \in \mathbb{R}^{n\times n}\,.$$

Lipschitz-stetig ist. Dies ist eine deutlich schwächere Voraussetzung als in Satz 18.1. ◊

Beispiel. Wir greifen noch einmal Beispiel 17.3 aus Abschnitt 17 auf. Dort haben wir eine Lösung $[\widehat{x}_1, \widehat{x}_2]^T$ des nichtlinearen Gleichungssystems

$$\begin{aligned} x_1 \;&=\; (\cos x_1 - \sin x_2)/4\,, \\ x_2 \;&=\; (\cos x_1 - 2\sin x_2)/4\,, \end{aligned}$$

mit einer Fixpunktiteration approximiert. Zum Vergleich soll hier das Newton-Verfahren verwendet werden. Dazu muß zunächst eine Funktion F konstruiert werden, die $\widehat{x}$ als Nullstelle besitzt, etwa

$$F(x_1, x_2) \;=\; \begin{bmatrix} x_1 - 0.25\cos x_1 + 0.25\sin x_2 \\ x_2 - 0.25\cos x_1 + 0.5\ \ \sin x_2 \end{bmatrix}\,.$$

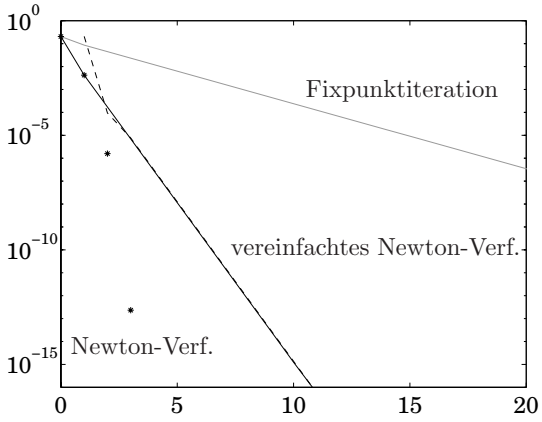


Abb. 19.1: Konvergenzverlauf der Newton-Iteration

Die Ableitung F' lautet

$$F'(x_1, x_2) \;=\; \begin{bmatrix} 1 + 0.25 \sin x_1 & 0.25 \cos x_2 \\ 0.25 \sin x_1 & 1 + 0.5 \cos x_2 \end{bmatrix}$$

mit Determinante $\det F' = 1 + 0.25 \sin x_1 + 0.5 \cos x_2 + (1/16) \sin x_1 \cos x_2$. Da diese Determinante ungleich Null ist, sind die Iterationen (19.1) des Newton-Verfahrens wohldefiniert mit

$$F'(x_1, x_2)^{-1} \;=\; \frac{1}{\det F'} \begin{bmatrix} 1 + 0.5 \cos x_2 & -0.25 \cos x_2 \\ -0.25 \sin x_1 & 1 + 0.25 \sin x_1 \end{bmatrix} .$$

Als Startvektor wählen wir wie in Beispiel 17.3 $x^{(0)} = 0$.

Abbildung 19.1 demonstriert die typische Überlegenheit des quadratisch konvergenten Newton-Verfahrens gegenüber der linear konvergenten Fixpunktiteration aus Beispiel 17.3: Die schwach eingezeichnete Linie zeigt den Fehlerverlauf der Fixpunktiteration aus Abbildung 17.3, die Sterne repräsentieren die absoluten Fehler der Newton-Iterierten (jeweils bezüglich der Maximumnorm). Die anderen beiden Kurven werden auf Seite 177 erläutert. In der nachfolgenden Tabelle sind die signifikanten Ziffern der Newton-Iterierten dunkel gedruckt.

k	$x_1^{(k)}$	$x_2^{(k)}$
0	0	0
1	0.20833333333333	0.16666666666667
2	0.20413062486814	0.16344969131265
3	0.20412903125185	0.16344858405833
4	0.20412903125162	0.16344858405816

$\Diamond$

Initialisierung: $x^{(0)} \in \mathcal{D}(F)$ sei eine Approximation einer Nullstelle $\widehat{x}$ von F,
A eine Approximation von $F'(\widehat{x})$

`for` $k = 0, 1, 2, \ldots$ `do`
  $Ah^{(k)} = -F(x^{(k)})$ % löse lineares Gleichungssystem
  $x^{(k+1)} = x^{(k)} + h^{(k)}$
  `if` $x^{(k+1)} \notin \mathcal{D}(F)$ `then`
    `error` % „Overflow“
  `end if`
  überprüfe Konvergenz, vgl. Bemerkung 19.2, ggf. Abbruch wegen Divergenz
`until` stop % end for

Algorithmus 19.2: Vereinfachtes Newton-Verfahren

Bei höherdimensionalen Problemen wird das Newton-Verfahren sehr aufwendig, da in jedem Schritt die neue Ableitungsmatrix ausgewertet und zur Lösung des Gleichungssystems (19.2) faktorisiert werden muß. In vielen Fällen kann man sich daher mit dem sogenannten *vereinfachten Newton-Verfahren* (Algorithmus 19.2) behelfen, bei dem die Ableitungsmatrix in den Gleichungssystemen durch eine Approximation $A \approx F'(x^{(k)})$ ersetzt wird; denkbar ist etwa die Wahl $A = F'(x^{(0)})$.

Die vereinfachte Newton-Iteration kann als Fixpunktiteration $x^{(k+1)} = \Phi(x^{(k)})$ mit Fixpunktoperator

$$\Phi(x) \;=\; x - A^{-1}F(x)$$

interpretiert werden. Nach Satz 17.2 konvergiert daher Algorithmus 19.2 lokal gegen $\widehat{x}$, falls eine geeignete Norm von

$$\Phi'(\widehat{x}) \;=\; I - A^{-1}F'(\widehat{x}) \;=\; A^{-1}\big(A - F'(\widehat{x})\big)$$

kleiner als Eins ist, also falls A eine hinreichend gute Approximation an $F'(\widehat{x})$ ist. In diesem Fall ist das vereinfachte Newton-Verfahren eine Kontraktion in einer Umgebung der Nullstelle $\widehat{x}$ und dort linear konvergent. Eine höhere Konvergenzordnung liegt in der Regel nicht vor.

Bemerkung 19.2. Da diese Konvergenzaussage nur lokal ist, stellt sich die Frage, wie bei einer konkreten Iteration entschieden werden kann, ob das Verfahren mit dem verwendeten $x^{(0)}$ und der jeweiligen Matrix A konvergiert oder nicht. Hierzu können die Resultate des Banachschen Fixpunktsatzes 7.1 herangezogen werden. Um zu klären, ob die Iterierten aus Algorithmus 19.2 im Konvergenzbereich liegen, kann beispielsweise der Kontraktionsfaktor q des

Fixpunktoperators Φ, vgl. (7.1), aus dem Quotienten

$$q_k = \frac{\|\Phi(x^{(k)}) - \Phi(x^{(k-1)})\|}{\|x^{(k)} - x^{(k-1)}\|} = \frac{\|x^{(k+1)} - x^{(k)}\|}{\|x^{(k)} - x^{(k-1)}\|} = \frac{\|h^{(k)}\|}{\|h^{(k-1)}\|}\,, \quad k \geq 1\,, \tag{19.5}$$

geschätzt werden. Die Wahl der Norm ergibt sich hierbei aus der jeweiligen Anwendung. Ein Schätzwert $q_k \ll 1$ (etwa $q_k \leq 1/2$) deutet auf Konvergenz hin, liegt hingegen der Schätzwert q_k für mehrere aufeinanderfolgende Iterationen in der Nähe von Eins oder gar darüber, so ist zu befürchten, daß die vereinfachte Newton-Iteration nicht konvergiert; in diesem Fall verbleibt nur die Möglichkeit, das Verfahren mit einer besseren Startnäherung oder einer besseren Näherungsmatrix A neu zu starten. Gleiches gilt, wenn eine Iterierte $x^{(k+1)}$ außerhalb des Definitionsbereichs $\mathcal{D}(F)$ liegt.

Ist der Schätzwert q_k aus (19.5) kleiner als Eins, ergibt sich zudem die Möglichkeit, den Iterationsfehler mit der A-posteriori-Schranke des Banachschen Fixpunktsatzes zu schätzen:

$$\|x^{(k+1)} - \widehat{x}\| \lesssim \frac{q_k}{1 - q_k}\,\|x^{(k+1)} - x^{(k)}\| = \frac{\|h^{(k)}\|^2}{\|h^{(k-1)}\| - \|h^{(k)}\|}\,. \tag{19.6}$$

Dies ermöglicht ein effektives Abbruchkriterium für Algorithmus 19.2. Beim klassischen Newton-Verfahren kann entsprechend vorgegangen werden. ◊

Beispiel. In dem obigen Beispiel ergibt sich für das vereinfachte Newton-Verfahren mit $A = F'(0)$ die dritte Fehlerkurve aus Abbildung 19.1 (durchgezogene Linie). Die gebrochene Linie, die weitgehend mit dieser Fehlerkurve übereinstimmt, entspricht dem zugehörigen Fehlerschätzer (19.6). Offensichtlich konvergiert das Verfahren deutlich schneller als die Fixpunktiteration aus Abschnitt 17.1, dennoch ist die Konvergenzordnung lediglich linear. ◊

20 Das nichtlineare Ausgleichsproblem

Nichtlineare Gleichungssysteme haben häufig mehr Gleichungen als Unbekannte und sind dann nicht unbedingt lösbar, analog zu überbestimmten linearen Gleichungssystemen. Formulieren wir ein solches Problem wieder als Nullstellenaufgabe für eine Funktion $F : \mathcal{D}(F) \subset \mathbb{R}^n \to \mathbb{R}^m$, so landen wir in natürlicher Weise bei einem ($m \times n$-dimensionalen) *nichtlinearen Ausgleichsproblem*:

$$\text{minimiere} \quad \Phi(x) = \frac{1}{2}\,\|F(x)\|_2^2\,. \tag{20.1}$$

Im folgenden wird generell vorausgesetzt, daß $m \geq n$ und F hinreichend glatt ist.

Wir erinnern zunächst daran, daß dann die Bedingungen

$$\operatorname{grad}\Phi(\widehat{x}) = 0 \qquad \text{und} \qquad \Phi''(\widehat{x}) \text{ positiv definit} \tag{20.2}$$

notwendig und hinreichend dafür sind, daß in $\widehat{x}$ ein lokales Minimum von Φ vorliegt. Dabei ist $\operatorname{grad}\Phi(x) \in \mathbb{R}^n$ der *Gradient* von Φ und

$$\Phi''(x) \;=\; \left[\frac{\partial^2}{\partial x_i \partial x_j}\,\Phi(x)\right]_{ij} \in \mathbb{R}^{n\times n}$$

die (symmetrische) *Hesse-Matrix*. Ist zumindest die erste der beiden Gleichungen in (20.2) erfüllt, so nennen wir $\widehat{x}$ einen stationären Punkt.

Zur Lösung von (20.1) gibt es eine Vielzahl iterativer Algorithmen, die hier nur exemplarisch vorgestellt werden können. Grob gesprochen kann man zwei Verfahrensklassen unterscheiden: *Gradientenverfahren* (oder *Abstiegsverfahren*), die in jedem Iterationsschritt das Funktional Φ in einem eindimensionalen affinen Raum minimieren, und *Newton-artige Verfahren*, bei denen (wie in Abschnitt 19) Φ oder F in (20.1) durch eine lokale Linearisierung ersetzt wird. Im folgenden diskutieren wir am Beispiel des Verfahrens des steilsten Abstiegs die typischen Problemstellungen bei Gradientenverfahren; im nächsten Abschnitt stellen wir dann das Levenberg-Marquardt-Verfahren als Vertreter der Newton-artigen Methoden vor.

Gradientenverfahren approximieren das Minimum $\widehat{x}$ von (20.1) durch eine Iterationsfolge $\{x^{(k)}\}$, bei der sich $x^{(k+1)}$ aus $x^{(k)}$ durch die Wahl einer „Suchrichtung“ $d^{(k)}$ und einer Schrittweite $\alpha_k > 0$ ergibt:

$$x^{(k+1)} = x^{(k)} + \alpha_k\, d^{(k)}\,. \tag{20.3}$$

Suchrichtung und Schrittweite werden dabei so bestimmt, daß eine Abstiegsbedingung $\Phi(x^{(k+1)}) < \Phi(x^{(k)})$ erfüllt ist (für Konvergenzaussagen muß die Abstiegsbedingung wie in (20.8) weiter verschärft werden).

Ein Vektor $d^{(k)}$ wird in diesem Zusammenhang *Abstiegsrichtung* genannt, wenn die Richtungsableitung von Φ in Richtung $d^{(k)}$, gegeben durch

$$\frac{\partial\Phi}{\partial d^{(k)}}(x^{(k)}) \;=\; \operatorname{grad}\Phi(x^{(k)})^* d^{(k)}$$

negativ ist, also wenn

$$d^{(k)} \;=\; -\operatorname{grad}\Phi(x^{(k)}) + p^{(k)} \qquad \text{für ein} \qquad p^{(k)} \perp \operatorname{grad}\Phi(x^{(k)})\,. \tag{20.4}$$

In gewissem Sinn ist dabei die Wahl $d^{(k)} = -\operatorname{grad}\Phi(x^{(k)})$ optimal, denn nach der Cauchy-Schwarz-Ungleichung gilt für jede Richtung $d^{(k)}$

$$\frac{\partial\Phi}{\partial d^{(k)}}(x^{(k)}) \;=\; \operatorname{grad}\Phi(x^{(k)})^* d^{(k)} \;\geq\; -\,\|\operatorname{grad}\Phi(x^{(k)})\|_2\,\|d^{(k)}\|_2$$

mit Gleichheit genau dann, wenn $d^{(k)}$ in die Richtung des negativen Gradienten zeigt.

Für $d^{(k)} = -\operatorname{grad}\Phi(x^{(k)})$ nennt man das resultierende Verfahren (20.3) daher auch die *Methode des steilsten Abstiegs*. Bezeichnen wir mit $F_i(x) : \mathcal{D}(F) \to \mathbb{R}$, $i = 1, \ldots, m$, die einzelnen Koeffizientenfunktionen von F, dann ist

$$F(x) \;=\; \begin{bmatrix} F_1(x) \\ \vdots \\ F_m(x) \end{bmatrix}, \qquad \Phi(x) \;=\; \frac{1}{2}\, F(x)^* F(x) \;=\; \frac{1}{2} \sum_{i=1}^{m} F_i(x)^2 \,,$$

und der Gradient von Φ lautet

$$\operatorname{grad}\Phi(x) \;=\; \sum_{i=1}^{m} F_i(x) \operatorname{grad} F_i(x) \;=\; F'(x)^* F(x) \,, \tag{20.5}$$

wobei F' wieder die Jacobi-Matrix von F bezeichnet. Das Verfahren des steilsten Abstiegs genügt somit der Rekursion

$$x^{(k+1)} = x^{(k)} - \alpha_k F'(x^{(k)})^* F(x^{(k)}) \,, \tag{20.6}$$

in der noch die Schrittweite α_k geeignet zu bestimmen ist.

Während beim Verfahren des steilsten Abstiegs die Suchrichtung $d^{(k)}$ von Schritt zu Schritt optimal ist, braucht sie auf lange Sicht nicht optimal zu sein. In der Praxis werden daher alternative Suchrichtungen verwendet, die jedoch nicht „zu weit“ vom Gradienten abweichen sollten. Üblich sind die folgenden beiden Einschränkungen an $d^{(k)}$:

$$c \,\|\operatorname{grad}\Phi(x^{(k)})\|_2 \;\le\; \|d^{(k)}\|_2 \;\le\; C \,\|\operatorname{grad}\Phi(x^{(k)})\|_2 \tag{20.7a}$$

für feste Konstanten $c, C > 0$ und

$$\cos\sphericalangle\big(\operatorname{grad}\Phi(x^{(k)}), d^{(k)}\big) \;=\; \frac{\operatorname{grad}\Phi(x^{(k)})^* d^{(k)}}{\|\operatorname{grad}\Phi(x^{(k)})\|_2 \|d^{(k)}\|_2} \;\le\; -\delta \tag{20.7b}$$

für ein festes $\delta \in (0, 1]$. Demnach soll $d^{(k)}$ von derselben Größenordnung sein wie der Gradient von Φ und der eingeschlossene Winkel soll strikt größer als $\pi/2$ und kleiner als $3\pi/2$ bleiben. Die Bedingung (20.7b) ist in Abbildung 20.1 veranschaulicht: Die Suchrichtung darf nur in den grau eingezeichneten Bereich zeigen.

Damit verbleibt schließlich noch, die *Schrittweite* festzulegen, also die Wahl von α_k. In Algorithmus 20.1 wird die Schrittweite durch die `while`-Schleife gesteuert: Die Abbruchbedingung garantiert nicht nur die Abstiegsbedingung

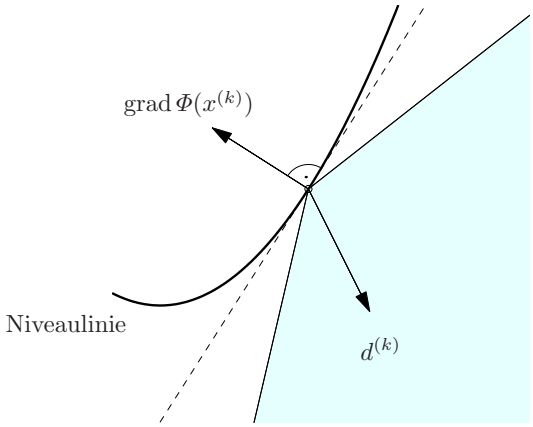


Abb. 20.1: Abstiegsrichtung

$\Phi(x^{(k+1)}) < \Phi(x^{(k)})$, sondern sogar die etwas stärkere *Armijo-Goldstein-Bedingung*

$$\Phi(x^{(k)} + \alpha_k d^{(k)}) \;\le\; \Phi(x^{(k)}) \;-\; \mu\alpha_k \,|\operatorname{grad}\Phi(x^{(k)})^* d^{(k)}|\,. \tag{20.8}$$

Abbildung 20.2 illustriert die Armijo-Goldstein-Bedingung: Die Schrittweite α ist so einzustellen, daß sich der neue Funktionswert von Φ im grauen Bereich befindet, d. h. α muß in dem eingezeichneten Intervall $(0, \alpha_*]$ liegen.

Bemerkung. Damit Algorithmus 20.1 wohldefiniert ist, muß die `while`-Schleife zur Schrittweitensteuerung terminieren. Dies folgt aus einer Taylorentwicklung von Φ, denn unter der Voraussetzung (20.7b) ist $\operatorname{grad}\Phi(x^{(k)})^* d^{(k)}$ negativ, also

$$\Phi(x^{(k)} + \alpha d^{(k)}) \;=\; \Phi(x^{(k)}) - \alpha\,|\operatorname{grad}\Phi(x^{(k)})^* d^{(k)}| + O(\alpha^2)\,, \qquad \alpha \to 0\,,$$

Initialisierung: $\mu \in (0,1)$ sei gegeben, etwa $\mu = 0.5$
  wähle $x^{(0)}$ und Suchrichtung $d^{(0)}$
  `for` $k = 0, 1, 2, \ldots$ `do` % Iterationsindex
    $\alpha = 1$ % Schrittweite initialisieren
    `while` $\Phi(x^{(k)} + \alpha d^{(k)}) > \Phi(x^{(k)}) - \mu\alpha|\operatorname{grad}\Phi(x^{(k)})^* d^{(k)}|$ `do`
      $\alpha = \alpha/2$
    `end while`
    $x^{(k+1)} = x^{(k)} + \alpha d^{(k)}$
    wähle neue Suchrichtung $d^{(k+1)}$ unter Beachtung der Einschränkungen (20.7)
  `until` stop % end for

Algorithmus 20.1: Allgemeines Abstiegsverfahren

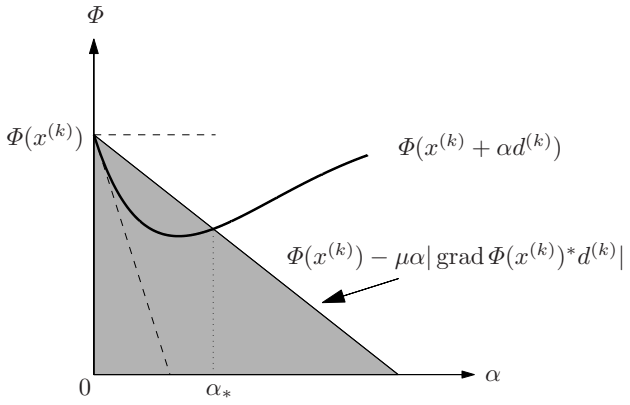


Abb. 20.2: Armijo-Goldstein-Bedingung

und für hinreichend kleine $\alpha = \alpha_k$ ist die rechte Seite dieser Taylorentwicklung kleiner als die rechte Seite von (20.8). ◊

Wir beweisen nun ein Konvergenzresultat für diese allgemeine Klasse von Abstiegsverfahren.

Satz 20.1. *Die Funktion F sei in einer offenen Menge $\mathcal{U} \subset \mathcal{D}(F)$ stetig differenzierbar mit Lipschitz-stetiger Ableitung F'. Ferner enthalte $\mathcal{U}$ den Startvektor $x^{(0)}$ sowie die gesamte Menge*

$$\mathcal{M}(x_0) \;=\; \{x \in \mathcal{D}(F) \;:\; \Phi(x) \le \Phi(x^{(0)})\}\,. \tag{20.9}$$

Falls die Suchrichtung $d^{(k)}$ in jedem Iterationsschritt die Bedingungen (20.7) *erfüllt, dann gilt für die Iterierten von Algorithmus* 20.1*, daß*

$$\operatorname{grad}\Phi(x^{(k)}) \longrightarrow 0\,, \qquad k \to \infty\,.$$

Beweis. Um die folgenden Argumente nicht unnötig kompliziert gestalten zu müssen, beschränken wir uns für den Beweis auf die Methode des steilsten Abstiegs, also den Fall

$$d^{(k)} \;=\; -\operatorname{grad}\Phi(x^{(k)})\,.$$

In diesem Fall ergibt sich für die Schrittweite α_k in (20.3) aus der Armijo-Goldstein-Bedingung (20.8) die Ungleichung

$$\Phi(x^{(k+1)}) \;\le\; \Phi(x^{(k)}) \;-\; \mu\alpha_k \,\|\operatorname{grad}\Phi(x^{(k)})\|_2^2\,. \tag{20.10}$$

Ferner können wir annehmen, daß $\operatorname{grad}\Phi(x^{(k)})$ für alle Iterierten von Null verschieden ist, denn ansonsten findet die Iteration „zufällig“ einen stationären Punkt und terminiert.

1. Aufgrund von (20.10) ist die Folge $\{\Phi(x^{(k)})\}$ monoton fallend und nach unten durch Null beschränkt. Daraus folgt insbesondere, daß alle Iterierten zu $\mathcal{U} \subset \mathcal{D}(F)$ gehören (dies ist der Grund für die Voraussetzung, daß $\mathcal{U}$ die gesamte Menge $\mathcal{M}(x_0)$ aus (20.9) enthalten soll). Aus (20.10) ergibt sich induktiv die Abschätzung

$$\begin{aligned}\Phi(x^{(0)}) &\geq \Phi(x^{(1)}) + \mu\alpha_0 \,\|\operatorname{grad}\Phi(x^{(0)})\|_2^2 \;\geq\; \dots \\ &\geq \Phi(x^{(k+1)}) + \mu \sum_{j=0}^{k} \alpha_j \,\|\operatorname{grad}\Phi(x^{(j)})\|_2^2 \;\geq\; \mu \sum_{j=0}^{k} \alpha_j \,\|\operatorname{grad}\Phi(x^{(j)})\|_2^2 \,.\end{aligned}$$

Demnach konvergiert die letzte Summe für $k \to \infty$, d. h. die einzelnen Summanden müssen für $k \to \infty$ gegen Null konvergieren:

$$\alpha_k \|\operatorname{grad}\Phi(x^{(k)})\|_2^2 \;\longrightarrow\; 0\,, \qquad k \to \infty\,. \tag{20.11}$$

2. Nun ist noch zu zeigen, daß ein $\varepsilon > 0$ existiert mit $\alpha_k \geq \varepsilon$ für alle $k \in \mathbb{N}$. Nehmen wir also an, daß $\alpha_k < 1$ ist für ein $k \in \mathbb{N}$ (ansonsten kann $\varepsilon = 1$ gewählt werden und der Beweis ist fertig). In diesem Fall muß in der k-ten Iteration die `while`-Schleife aus Algorithmus 20.1 mindestens einmal durchlaufen worden sein, d. h. die Abbruchbedingung der Schleife war für $2\alpha_k$ *nicht* erfüllt. Wegen der speziellen Gestalt (20.10) der Abbruchbedingung für das Verfahren des steilsten Abstiegs folgt hieraus

$$2\mu\alpha_k \,\|\operatorname{grad}\Phi(x^{(k)})\|_2^2 \;>\; \Phi(x^{(k)}) \;-\; \Phi(x^{(k)} + 2\alpha_k d^{(k)})\,.$$

Durch Taylorentwicklung kann die rechte Seite weiter abgeschätzt werden: Wegen der Lipschitz-Stetigkeit von F' ist auch $\operatorname{grad}\Phi$ Lipschitz-stetig, und somit existiert eine Konstante $\gamma > 0$ mit

$$\begin{aligned}2\mu\alpha_k \,\|\operatorname{grad}\Phi(x^{(k)})\|_2^2 \;&>\; -\,2\alpha_k \operatorname{grad}\Phi(x^{(k)})^* d^{(k)} \;-\; \gamma\,\alpha_k^2 \,\|d^{(k)}\|_2^2 \\ &=\; 2\alpha_k \|\operatorname{grad}\Phi(x^{(k)})\|_2^2 \;-\; \gamma\,\alpha_k^2 \|\operatorname{grad}\Phi(x^{(k)})\|_2^2\,.\end{aligned}$$

Aufgrund unserer Annahme $\operatorname{grad}\Phi(x^{(k)}) \;/= 0$ können die Gradientennormen herausgekürzt werden und wir erhalten die Ungleichung

$$\gamma\alpha_k^2 \;>\; 2(1-\mu)\alpha_k\,.$$

Wegen $\mu < 1$ und $\alpha_k \;/= 0$ folgt hieraus unmittelbar die untere Schranke $\varepsilon = 2(1-\mu)/\gamma > 0$ für α_k. Zusammen mit (20.11) folgt schließlich die Aussage des Satzes. □

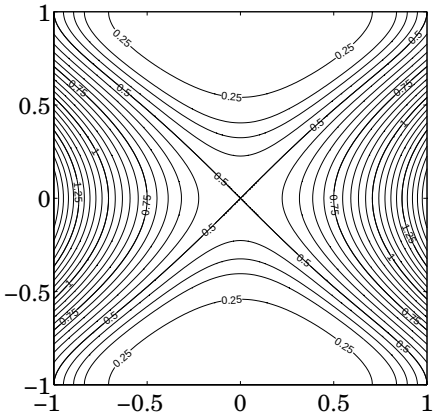


Abb. 20.3:
Höhenlinien von Φ aus (20.12)

Bemerkung. Satz 20.1 besagt *nicht*, daß die Folge $\{x^{(k)}\}$ konvergent ist. Selbst wenn sie konvergiert, braucht der Grenzwert kein Minimum von Φ zu sein. Im allgemeinen konvergieren die Iterierten lediglich gegen einen stationären Punkt von Φ. Hat hingegen Φ nur einen stationären Punkt $\widehat{x}$ und ist die Menge $\mathcal{M}(x^{(0)})$ zudem beschränkt und F hinreichend glatt, dann konvergiert die Folge $\{x^{(k)}\}$ gegen $\widehat{x}$. In diesem Fall ist $\widehat{x}$ zwangsläufig ein Minimum von Φ. ◊

Beispiel. Gegeben sei die Funktion $F(x_1, x_2) = [x_1, x_2^2 - 1, x_1(x_2^2 - 1)]^T$ für das nichtlineare Ausgleichsproblem (20.1), so daß

$$\Phi(x_1, x_2) \;=\; \frac{1}{2}\,\|F(x_1, x_2)\|_2^2 \;=\; \frac{1}{2}\left(x_1^2 + (x_2^2 - 1)^2 + x_1^2(x_2^2 - 1)^2\right). \qquad (20.12)$$

Das Minimum von Φ ist Null und wird für $x_1 = 0$ und $x_2 = \pm 1$ angenommen. Hat eine Iterierte $x^{(k)}$ von Algorithmus 20.1 die Gestalt $x^{(k)} = [x_1, 0]^T$, dann lautet die Suchrichtung für das Verfahren des steilsten Abstiegs

$$d^{(k)} \;=\; -\operatorname{grad}\Phi(x^{(k)}) \;=\; -\,[2x_1, 0]^T \;=\; -2x^{(k)}\,.$$

Nach Satz 20.1 konvergiert daher $x^{(k)} = \operatorname{grad}\Phi(x^{(k)})/2$ gegen Null für $k \to \infty$. Der Nullpunkt ist jedoch lediglich ein Sattelpunkt von Φ, da die (eindimensionale) Funktion $\Phi(0, x_2)$ für $x_2 = 0$ ein lokales Maximum aufweist, vgl. die Höhenlinien der Funktion Φ in Abbildung 20.3. ◊

Zur numerischen Illustration betrachten wir ein beliebtes Testproblem für nichtlineare Optimierungsalgorithmen:

Beispiel 20.2. Gegeben sei das Ausgleichsproblem (20.1) mit

$$F : \mathbb{R}^2 \to \mathbb{R}^2\,, \qquad F(x_1, x_2) \;=\; \begin{bmatrix} 10(x_2 - x_1^2) \\ x_1 - 1 \end{bmatrix}.$$

Offensichtlich hat F genau eine Nullstelle für $x_1 = x_2 = 1$, d. h. $\widehat{x} = [1,1]^T$ ist die gesuchte Lösung des Ausgleichsproblems.

Die Niveaulinien von Φ zum Niveau c ergeben sich als Lösungen der Gleichung

$$100(x_2 - x_1^2)^2 \;=\; 2c - (1 - x_1)^2 \,.$$

Daraus erhalten wir notwendigerweise $-\sqrt{2c} < x_1 - 1 < \sqrt{2c}$ und schließlich

$$x_2 = x_1^2 \pm \frac{1}{10}\sqrt{2c - (1-x_1)^2}\,, \qquad 1 - \sqrt{2c} < x_1 < 1 + \sqrt{2c}\,.$$

Zur Minimierung von Φ betrachten wir die Methode des steilsten Abstiegs. Wegen

$$F'(x_1, x_2) \;=\; \begin{bmatrix} -20x_1 & 10 \\ 1 & 0 \end{bmatrix} \tag{20.13}$$

ergibt sich dabei jeweils in $x^{(k)} = [x_1, x_2]^T$ die Suchrichtung

$$d^{(k)} \;=\; -F'(x^{(k)})^* F(x^{(k)}) \;=\; \begin{bmatrix} -200x_1(x_2 - x_1)^2 + x_1 - 1 \\ 100(x_2 - x_1)^2 \end{bmatrix}.$$

Der linke Teil von Abbildung 20.4 zeigt einige ausgewählte Niveaulinien der Funktion Φ und die Iterierten des Verfahrens in der (x_1, x_2)-Ebene (durch Kreise gekennzeichnet). Man beachte das zentrale langgestreckte Tal in Form einer Banane sowie die relativ steil aufragenden Talwände (die nach außen hin immer steiler werden).

Startpunkt für die Iteration ist $x^{(0)} = [-0.5, -0.4]^T$, links unten in Abbildung 20.4; die Lösung $\widehat{x} = [1,1]^T$ ist in der Abbildung rechts oben durch einen dickeren Punkt markiert. Die Schrittweite α_k wird wie in Algorithmus 20.1 durch die Armijo-Goldstein-Bedingung (mit Parameter $\mu = 0.5$) gesteuert. Wie man sieht, ist die Konvergenz sehr langsam; selbst nach tausend Iterationsschritten haben die Iterierten den Grenzwert noch nicht erreicht. Gleichwohl konvergiert die Iteration letztendlich, denn Φ hat nur den einen stationären Punkt $\widehat{x}$ und die Menge $\mathcal{M}(x^{(0)})$ ist beschränkt.

Im rechten Teil derselben Abbildung sieht man die Iterierten des Levenberg-Marquardt-Verfahrens, das wesentlich schneller konvergiert. Dieses Verfahren ist Gegenstand des folgenden Abschnitts. ◊

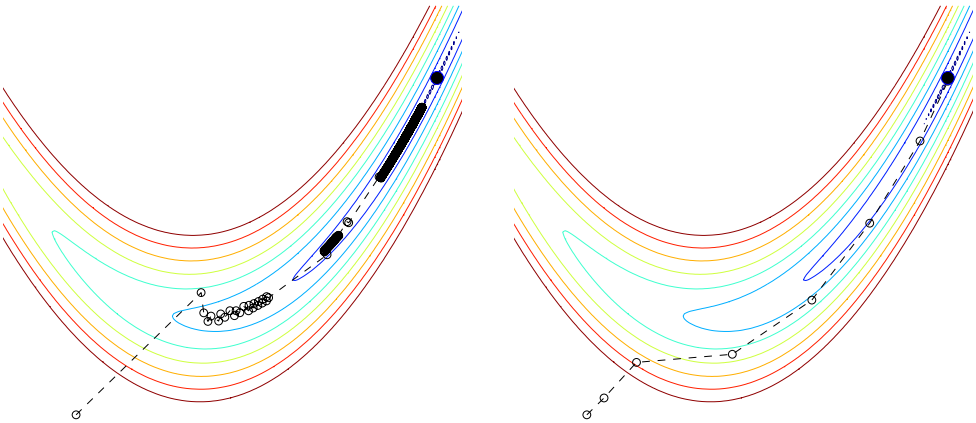


Abb. 20.4: Das Verfahren des steilsten Abstiegs (links) und die Levenberg-Marquardt-Iteration (rechts)

21 Das Levenberg-Marquardt-Verfahren

Für die Konstruktion schnellerer Verfahren bietet es sich an, die Funktion F in der Definition von Φ wie beim Newton-Verfahren durch das lineare Taylorpolynom um die aktuelle Iterierte zu ersetzen,

$$\Phi(x) \;=\; \frac{1}{2}\,\|F(x)\|_2^2 \;\approx\; \frac{1}{2}\,\|F(x^{(k)}) + F'(x^{(k)})(x - x^{(k)})\|_2^2\,. \tag{21.1}$$

Dies führt auf das linearisierte Problem

$$\text{minimiere} \quad \frac{1}{2}\,\|F(x^{(k)}) + F'(x^{(k)})(x - x^{(k)})\|_2^2\,, \tag{21.2}$$

dessen Lösung $x = x^{(k+1)}$ dann die nächste Iterierte ergibt. Dies ist ein $m \times n$-dimensionales lineares Ausgleichsproblem, dessen Lösung in der Regel über die QR-Zerlegung der Koeffizientenmatrix $F'(x^{(k)}) \in \mathbb{R}^{m\times n}$ erfolgt, vgl. Abschnitt 13. Die Lösung $x^{(k+1)}$ kann über die Pseudoinverse von $F'(x^{(k)})$ ausgedrückt werden,

$$x^{(k+1)} = x^{(k)} - F'(x^{(k)})^{\dagger} F(x^{(k)})\,. \tag{21.3}$$

Hat $F'(x^{(k)})$ keinen vollen Spaltenrang, dann ist die Lösung von (21.2) nicht eindeutig bestimmt. In diesem Fall ist $x^{(k+1)}$ aus (21.3) die Lösung von (21.2) mit dem kleinsten euklidischen Abstand zu $x^{(k)}$. Das resultierende Iterationsverfahren zur Lösung des nichtlinearen Ausgleichsproblems wird *Gauß-Newton-Verfahren* genannt.

Leider lassen sich Beispiele konstruieren, bei denen das Gauß-Newton-Verfahren noch nicht einmal lokal gegen die Lösung des nichtlinearen Ausgleichsproblems konvergiert, vgl. Aufgabe 12. Der Grund liegt darin, daß bei einer schlechten Kondition von $F'(x^{(k)})$ der Vektor $x^{(k+1)} - x^{(k)}$ sehr groß sein kann, während die Linearisierung nur für „kleine“ $x^{(k+1)} - x^{(k)}$ sinnvoll ist. Daher wird bei den meisten praktischen Anwendungen das Gauß-Newton-Verfahren durch eine sogenannte *Trust-Region-Strategie* modifiziert, was auf das *Levenberg-Marquardt-Verfahren* führt.

Die Grundidee dieser Strategie besteht darin, daß der Approximation (21.1) nur in einer Kugel (*trust region*)

$$\mathcal{R}_k \;=\; \{\, x \in \mathbb{R}^n \;:\; \|x - x^{(k)}\|_2 \le \rho_k \,\}$$

um die aktuelle Iterierte $x^{(k)}$ „vertraut“ (engl.: *to trust*) werden kann. Aus diesem Grund wird das linearisierte Problem (21.2) folgendermaßen abgewandelt:

$$\begin{aligned} &\text{minimiere} \quad \frac{1}{2}\,\|F(x^{(k)}) + F'(x^{(k)})(x - x^{(k)})\|_2^2 \\ &\text{unter der Nebenbedingung} \quad \|x - x^{(k)}\|_2 \le \rho_k\,. \end{aligned} \tag{21.4}$$

Die Radien ρ_k werden dabei dem jeweiligen Verlauf der Iteration angepaßt, siehe unten. ie Vorschrift (21.4) erfordert die Minimierung einer stetigen reellwertigen Funktion über einer kompakten Teilmenge $\mathcal{R}_k \subset \mathbb{R}^n$; dieses *restringierte Ausgleichsproblem* hat bekanntermaßen (mindestens) eine Lösung.

Für die weiteren Untersuchungen überführen wir (21.4) in eine uns vertrautere Notation. Mit

$$A_k = F'(x^{(k)})\,, \qquad h = x - x^{(k)}\,, \qquad b^{(k)} = -F(x^{(k)})\,,$$

ergibt (21.4) das folgende Minimierungsproblem für h:

$$\begin{aligned} &\text{minimiere} \quad \Psi(h) \;=\; \frac{1}{2}\,\|A_k h - b^{(k)}\|_2^2 \\ &\text{unter der Nebenbedingung} \quad \|h\|_2 \le \rho_k\,. \end{aligned} \tag{21.5}$$

Ist eine Lösung $h^{(k)}$ dieses Problems gefunden, dann ergibt sich die nächste Iterierte als

$$x^{(k+1)} = x^{(k)} + h^{(k)}\,.$$

Um $h^{(k)}$ numerisch berechnen zu können, ist die Beschreibung über ein restringiertes Minimierungsproblem allerdings wenig hilfreich. Statt dessen unterscheiden wir im weiteren die folgenden beiden Fälle:

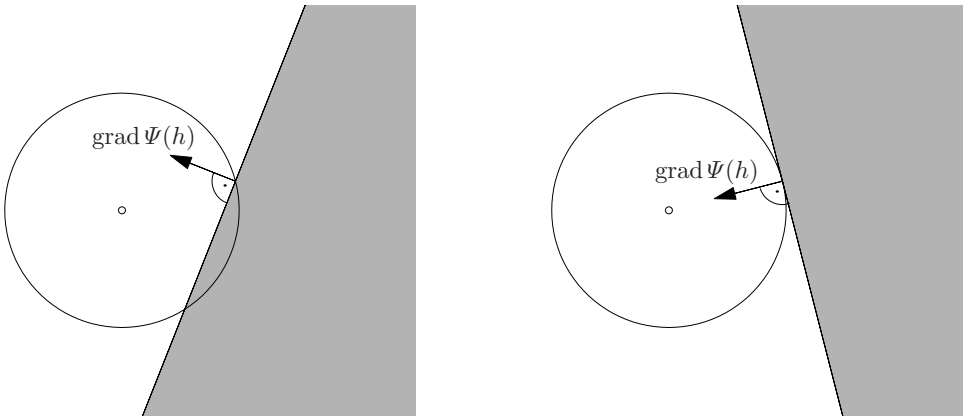


Abb. 21.1: Mögliche Abstiegsrichtungen

(*a*) $\|h^{(k)}\|_2 < \rho_k$:

Dann ist $h^{(k)}$ ein stationärer Punkt von Ψ, d. h. es ist $\operatorname{grad}\Psi(h^{(k)}) = 0$. Den Gradienten von Ψ haben wir bereits in Bemerkung 11.3 berechnet:

$$\operatorname{grad}\Psi(h) \;=\; A_k^*(A_k h - b^{(k)})\,.$$

Folglich löst $h^{(k)}$ die Gaußschen Normalengleichungen

$$A_k^* A_k h^{(k)} \;=\; A_k^* b^{(k)}\,. \tag{21.6a}$$

(*b*) $\|h^{(k)}\|_2 = \rho_k$:

In diesem Fall muß $\operatorname{grad}\Psi(h^{(k)})$ auf den Mittelpunkt der Kugel $\{\, h \,:\, \|h\|_2 \le \rho_k \,\}$ zeigen, da ansonsten eine Richtungsableitung von Ψ ins Kreisinnere negativ wäre. Dies ist in Abbildung 21.1 illustriert: Beide Skizzen zeigen mögliche Richtungen des Gradienten in einem Randpunkt h des Kreises, die graue Fläche gibt alle zugehörigen Abstiegsmöglichkeiten an, vgl. auch (20.4). Damit der Punkt h eine globale Minimalstelle der Funktion im Kreis sein kann, darf keine Abstiegsrichtung ins Innere des Kreises zeigen. Im linken Bild zeigt der Gradient nicht auf den Kreismittelpunkt und daher existieren Abstiegsmöglichkeiten ins Kreisinnere. Im rechten Bild zeigt der Gradient genau auf den Kreismittelpunkt und es gibt keine Abstiegsrichtung in das Innere des Kreises. Folglich existiert eine Darstellung

$$\operatorname{grad}\Psi(h^{(k)}) \;=\; A_k^* A_k h^{(k)} - A_k^* b^{(k)} \;=\; -\lambda_k h^{(k)} \qquad \text{für ein } \;\lambda_k > 0$$

beziehungsweise

$$\big(A_k^* A_k + \lambda_k I\big) h^{(k)} \;=\; A_k^* b^{(k)}\,. \tag{21.6b}$$

Offensichtlich kann (21.6a) als Grenzfall $\lambda_k = 0$ von (21.6b) aufgefaßt werden. Somit ist das folgende Resultat bewiesen.

Satz 21.1. *Die Lösung $h = h^{(k)}$ von* (21.5) *genügt der Gleichung*

$$(A_k^* A_k + \lambda_k I)\, h \;=\; A_k^* b^{(k)} \tag{21.7}$$

für ein $\lambda_k \geq 0$. Dabei ist λ_k genau dann positiv, wenn $\|h^{(k)}\|_2 = \rho_k > 0$ gilt.

Bemerkung 21.2. Es läßt sich zeigen, daß der Parameter λ_k der *Lagrange-Parameter* zu dem restringierten Minimierungsproblem (21.5) ist und daß die Lösung $h^{(k)}$ von (21.7) das Lagrange-Funktional

$$\frac{1}{2}\,\|A_k h - b^{(k)}\|_2^2 \;+\; \lambda_k \frac{1}{2}\,(\,\|\,h^{(k)}\|_2^2 - \rho_k)$$

minimiert, vgl. Heuser [53, Abschnitt 174]. ◊

Die Berechnung von $h^{(k)}$ kann auf das Problem in Abschnitt 18.3 zurückgeführt werden. Dazu faßt man (21.7) als eine einparametrige Schar linearer Gleichungen mit Parameter λ und Lösungen $h = h_\lambda$ auf. Gesucht ist nun der Parameter $\lambda = \lambda_k$ und die zugehörige Lösung $h^{(k)} = h_\lambda$, für die die Nebenbedingung

$$\|h_\lambda\|_2 \;=\; \|(A_k^* A_k + \lambda I)^{-1} A_k^* b^{(k)}\|_2 \;=\; \rho_k \tag{21.8}$$

erfüllt ist (sofern überhaupt ein entsprechender positiver Parameter λ_k existiert). Sind u_i die orthonormierten Eigenvektoren der symmetrischen und positiv semidefiniten Matrix $A_k^* A_k$ und d_i, $i = 1, \ldots, n$, die zugehörigen nichtnegativen Eigenwerte, dann kann $A_k^* b^{(k)}$ in dieser Eigenbasis entwickelt werden,

$$A_k^* b^{(k)} = \sum_{i=1}^{n} z_i u_i\,, \qquad z_i \in \mathbb{R}\,.$$

Mit dieser Darstellung ergibt sich

$$h_\lambda = (A_k^* A_k + \lambda I)^{-1} A_k^* b^{(k)} = \sum_{i=1}^{n} \frac{z_i}{d_i + \lambda}\, u_i$$

und aus (21.8) erhalten wir die rationale Gleichung

$$r(\lambda) \;=\; \sum_{i=1}^{n} \frac{z_i^2}{(d_i + \lambda)^2} \stackrel{!}{=} \rho_k^2$$

für den gesuchten Lagrange-Parameter λ, die mit dem Verfahren von Hebden aus Abschnitt 18.3 numerisch gelöst werden kann. Entscheidend ist hierbei, daß

die Spektralzerlegung von $A_k^* A_k$ nicht explizit berechnet werden muß, denn die Auswertung von $r(\lambda)$ erfolgt über $r(\lambda) = \|h_\lambda\|_2^2$, und entsprechend ergibt sich

$$r'(\lambda) \;=\; -2\, b^{(k)*} A_k \big(A_k^* A_k + \lambda I\big)^{-3} A_k^* b^{(k)} \;=\; -2 h_\lambda^* g_\lambda$$
$$\text{mit} \quad (A_k^* A_k + \lambda I) g_\lambda = h_\lambda \,.$$

Für jede Hebden-Iterierte λ sind also zur Berechnung von r und r' zwei Gleichungssysteme mit derselben Matrix $A_k^* A_k + \lambda I$ zu lösen (zur Bestimmung von h_λ und g_λ). Für eine effiziente Implementierung sei auf Aufgabe 11 verwiesen. Wir wenden uns nun der Anpassung des Trust-Region-Radius ρ_k zu. Dieser Radius sollte so groß wie möglich gewählt werden, denn wegen $\|x^{(k+1)} - x^{(k)}\|_2 = \|h^{(k)}\|_2 \le \rho_k$ schränkt ein allzu kleiner Radius ρ_k ein zügiges Voranschreiten der Iterierten ein. Andererseits darf ρ_k nur so groß gewählt werden, daß der Fehler durch die Linearisierung in einem angemessenen Rahmen bleibt.

Daher benötigen wir ein Kriterium, um die Güte der Linearisierung zu kontrollieren. Nach Möglichkeit soll dieses Kriterium nur auf leicht berechenbaren Größen beruhen, um den Rechenaufwand nicht wesentlich zu vergrößern. Ein erster Ansatz für ein solches Kriterium liegt auf der Hand: Der Wert der eigentlich zu minimierenden Zielfunktion Φ soll durch die neue Näherung $x^{(k+1)}$ verkleinert werden. Dies allein ist allerdings nicht unbedingt ausreichend. Deshalb greifen wir wieder auf das schärfere Armijo-Goldstein-Kriterium (20.8) zurück: Demnach ist eine neue Näherung $x^{(k+1)} = x^{(k)} + h^{(k)}$ akzeptabel, wenn

$$\frac{\Phi(x^{(k)}) - \Phi(x^{(k)} + h^{(k)})}{|\operatorname{grad}\Phi(x^{(k)})^* h^{(k)}|} \;\ge\; \mu$$

für einen Parameter $\mu \in (0,1)$ erfüllt ist. Der Iterationsschritt ist um so erfolgreicher, je größer μ gewählt werden kann. Wegen $\operatorname{grad}\Phi = F'^* F$, vgl. (20.5), kann der obige Bruch über die Darstellung

$$\mu_k \;=\; \frac{1}{2}\, \frac{\|F(x^{(k)})\|_2^2 \;-\; \|F(x^{(k)} + h^{(k)})\|_2^2}{-h^{(k)*} F'(x^{(k)})^* F(x^{(k)})} \tag{21.9}$$

ausgewertet werden. Dabei ist zu beachten, daß der Nenner ν_k dieses Bruchs wegen (21.7) durch

$$\begin{aligned} \nu_k \;&=\; -h^{(k)*} F'(x^{(k)})^* F(x^{(k)}) \;=\; h^{(k)*} A_k^* b^{(k)} \\ &=\; (A_k^* b^{(k)})^* (A_k^* A_k + \lambda_k I)^{-1} (A_k^* b^{(k)}) \end{aligned} \tag{21.10}$$

gegeben ist und somit, wie gewünscht, positiv ist.

Für die Anwendung des Armijo-Goldstein-Kriteriums verwenden wir zwei Toleranzschranken μ_- und μ_+ mit $0 < \mu_- < \mu_+ < 1$. Dabei wird die potentielle

Initialisierung: Parameter μ_-, μ_+ mit $0 < \mu_- < \mu_+ < 1$ seien gegeben, etwa $\mu_- = 1/4$ und $\mu_+ = 1/2$
wähle $x^{(0)}$ und Trust-Region-Radius ρ_0
`for` $k = 0, 1, 2, \ldots$ `do`
  $A_k = F'(x^{(k)})$
  $b^{(k)} = -F(x^{(k)})$
  berechne die Lösung $h^{(k)}$ von (21.5) mit dem Hebden-Verfahren, vgl. Abschnitt 18.3
  $\mu_k = \dfrac{1}{2} \dfrac{\|F(x^{(k)})\|_2^2 - \|F(x^{(k)} + h^{(k)})\|_2^2}{-h^{(k)*} F'(x^{(k)})^* F(x^{(k)})}$
  `if` $\mu_k < \mu_-$ `then` % Schritt nicht erfolgreich, verkleinere Trust-Region
    $x^{(k+1)} = x^{(k)}$
    $\rho_{k+1} = \rho_k / 2$
  `else` % Schritt erfolgreich
    $x^{(k+1)} = x^{(k)} + h^{(k)}$
    `if` $\mu_k > \mu_+$ `then` % vergrößere Trust-Region
      $\rho_{k+1} = 2\rho_k$
    `end if`
  `end if`
`until` stop % end for

Algorithmus 21.1: Levenberg-Marquardt-Verfahren

neue Näherung $x^{(k)} + h^{(k)}$ akzeptiert, falls $\mu_k \geq \mu_-$ ist. Wir sprechen in diesem Fall von einem *erfolgreichen Iterationsschritt*. Offensichtlich war in diesem Fall der Trust-Region-Radius ρ_k sinnvoll und bietet sich daher auch als Radius für die nächste Iteration an; gegebenenfalls könnte man die Trust-Region sogar vergrößern. Letzteres empfiehlt sich, wenn μ_k nahe bei Eins liegt, also für $\mu_k > \mu_+$, denn dies kann als Hinweis dafür interpretiert werden, daß sich die Funktion Φ in der Umgebung der aktuellen Iterierten nahezu linear verhält. Ist hingegen $\mu_k < \mu_-$, dann ist die Armijo-Goldstein-Bedingung verletzt und $x^{(k)} + h^{(k)}$ wird nicht als neue Iterierte akzeptiert. Der Iterationsschritt war also *nicht erfolgreich*. In diesem Fall war die Trust-Region offensichtlich zu groß und der Iterationsschritt muß mit einem kleineren Radius ρ_k wiederholt werden.

Das Levenberg-Marquardt-Verfahren wird in Algorithmus 21.1 zusammengefaßt. Wir beweisen im folgenden für diesen Algorithmus ein Analogon des Konvergenzsatzes 20.1 für allgemeine Abstiegsverfahren. Der Beweis ist hier allerdings deutlich schwieriger.

Satz 21.3. *Die Funktion F sei in einer kompakten Menge $\mathcal{U} \subset \mathcal{D}(F)$ Lipschitz-stetig differenzierbar. Ferner enthalte $\mathcal{U}$ den Startpunkt $x^{(0)}$ sowie die*

Menge $\mathcal{M}(x^{(0)})$ *aus* (20.9). *Dann gilt für die Iterierten* $x^{(k)}$ *von Algorithmus* 21.1, *daß*

$$\operatorname{grad}\Phi(x^{(k)}) \;\longrightarrow\; 0\,, \qquad k \to \infty\,.$$

Beweis. 1. Zunächst beweisen wir eine obere Schranke für den Lagrange-Parameter λ_k aus (21.7). Dazu nehmen wir ohne Einschränkung an, daß λ_k positiv (also ungleich Null) ist. Aus Satz 21.1 folgt durch Multiplikation von (21.7) mit $h^{(k)}$ und der Cauchy-Schwarz-Ungleichung

$$h^{(k)*}(A_k^*A_k + \lambda_k I)h^{(k)} \;=\; h^{(k)*}(A_k^*b^{(k)}) \;\le\; \|h^{(k)}\|_2\,\|A_k^*b^{(k)}\|_2\,.$$

Da $A_k^*A_k$ positiv semidefinit ist, kann die linke Seite nach unten durch $\lambda_k\|h^{(k)}\|_2^2$ abgeschätzt werden. Ferner gilt in dem betrachteten Fall $\|h^{(k)}\|_2 = \rho_k$ und daher ergibt dies

$$\lambda_k \;\le\; \|A_k^*b^{(k)}\|_2 \,/\, \rho_k\,. \tag{21.11}$$

2. Als nächstes verwenden wir diese Abschätzung, um eine untere Schranke für den Nenner ν_k des Armijo-Goldstein-Kriteriums (21.9) herzuleiten. Wir beschränken uns zunächst auf den Fall $\lambda_k > 0$. Dann ist $A_k^*A_k + \lambda_k I$ positiv definit und folglich existiert eine Cholesky-Faktorisierung

$$A_k^*A_k + \lambda_k I \;=\; LL^*$$

mit invertierbarem $L \in \mathbb{R}^{n\times n}$. Mit dieser Faktorisierung kann der Ausdruck (21.10) für ν_k weiter umgeformt werden und man erhält mit $w = L^{-1}A_k^*b^{(k)}$

$$\begin{aligned}\nu_k \;&=\; (A_k^*b^{(k)})^*(LL^*)^{-1}(A_k^*b^{(k)}) \;=\; w^*w\,\frac{\|A_k^*b^{(k)}\|_2^2}{(A_k^*b^{(k)})^*(A_k^*b^{(k)})} \\ &=\; \frac{\|w\|_2^2\,\|A_k^*b^{(k)}\|_2^2}{w^*L^*Lw}\,.\end{aligned} \tag{21.12}$$

Dabei gilt

$$\|L^*L\|_2 \;=\; \|LL^*\|_2 \;=\; \|A_k^*A_k + \lambda_k I\|_2 \;=\; \|A_k\|_2^2 + \lambda_k\,,$$

und nach Voraussetzung ist $\|A_k\|_2 = \|F'(x^{(k)})\|_2$ für alle $x^{(k)} \in \mathcal{U}$ durch eine Konstante $c > 0$ beschränkt. Zusammen mit (21.11) folgt somit

$$w^*L^*Lw \;\le\; \big(c^2 \,+\, \|A_k^*b^{(k)}\|_2/\rho_k\big)\,\|w\|_2^2\,.$$

Eingesetzt in (21.12) erhalten wir hieraus eine erste untere Schranke für ν_k:

$$\nu_k \;\ge\; \frac{1}{c^2 + \|A_k^*b^{(k)}\|_2/\rho_k}\,\|A_k^*b^{(k)}\|_2^2 \;=\; \frac{\rho_k}{c^2\rho_k + \|A_k^*b^{(k)}\|_2}\,\|A_k^*b^{(k)}\|_2^2\,.$$

Es gilt nun, die beiden Fälle $\|A_k^* b^{(k)}\|_2$ größer oder kleiner als ρ_k zu unterscheiden. Dies führt auf

$$\nu_k \;\geq\; \begin{cases} \frac{1}{c^2+1}\,\|A_k^* b^{(k)}\|_2^2 & \text{für } \|A_k^* b^{(k)}\|_2 \;\leq\; \rho_k\,, \\ \frac{1}{c^2+1}\,\rho_k\,\|A_k^* b^{(k)}\|_2 & \text{für } \|A_k^* b^{(k)}\|_2 \;>\; \rho_k\,, \end{cases}$$

so daß insgesamt

$$\nu_k \;\geq\; \tfrac{1}{c^2+1}\,\|A_k^* b^{(k)}\|_2\,\min\{\rho_k,\,\|A_k^* b^{(k)}\|_2\} \tag{21.13}$$

gilt.

Diese Abschätzung läßt sich für $\lambda_k = 0$ direkt herleiten: In diesem Fall ergibt sich nämlich aus (21.10)

$$\nu_k \;=\; (A_k^* b^{(k)})^* A_k^\dagger b^{(k)} \;=\; b^{(k)*} A_k A_k^\dagger b^{(k)} \;=\; \|P_k b^{(k)}\|_2^2\,,$$

wobei P_k den Orthogonalprojektor auf $\mathcal{R}(A_k)$ bezeichnet, vgl. Korollar 12.5. Wegen $\mathcal{N}(A_k^*) = \mathcal{R}(A_k)^\perp$ folgt weiterhin

$$\|A_k^* b^{(k)}\|_2^2 \;=\; \|A_k^* P_k b^{(k)}\|_2^2 \;\leq\; \|A_k\|_2^2 \|P_k b^{(k)}\|_2^2 \;\leq\; c^2 \nu_k\,,$$

und daher ist (21.13) auch im Fall $\lambda_k = 0$ gültig.

3. Wir beweisen nun, daß die rechte Seite von (21.13) gegen Null konvergiert. Bei einem erfolgreichen Iterationsschritt ist $\mu_k \geq \mu_-$ und aus (21.9) und (21.13) folgt

$$\begin{aligned} \|F(x^{(k)})\|_2^2 \;-\; \|F(x^{(k+1)})\|_2^2 \;&\geq\; 2\mu_- \nu_k \\ &\geq\; \frac{2\mu_-}{c^2+1}\,\|A_k^* b^{(k)}\|_2\,\min\bigl\{\rho_k,\,\|A_k^* b^{(k)}\|_2\bigr\}\,. \end{aligned} \tag{21.14}$$

Da $\{\,\|F(x^{(k)})\|_2\}_k$ offensichtlich eine monoton fallende Folge und gleichzeitig nach unten durch Null beschränkt ist, konvergiert die linke Seite von (21.14) für $k \to \infty$ gegen Null. Daher gilt zwangsläufig

$$\min\bigl\{\rho_k,\,\|A_k^* b^{(k)}\|_2\bigr\} \;\longrightarrow\; 0\,, \qquad k \to \infty\,. \tag{21.15}$$

Zunächst ist hierbei k auf die Indexmenge $\mathcal{K}$ der erfolgreichen Iterationsschritte einzuschränken. Da aber ρ_k bei erfolglosen Iterationsschritten halbiert wird, während A_k und $b^{(k)}$ unverändert bleiben, gilt (21.15) für alle $k \to \infty$. Dieselbe Argumentation gilt für den Fall, daß nur endlich viele Iterationen erfolgreich sind, da ρ_k immer nur weiter halbiert wird. 4. Als nächstes wird bewiesen, daß

$\|\operatorname{grad}\Phi(x^{(k)})\|_2$ für eine Teilfolge $\{k_m\}$ mit $k_m \to \infty$ gegen Null konvergiert. Nach (20.5) ist

$$\operatorname{grad}\Phi(x^{(k)}) = F'(x^{(k)})^* F(x^{(k)}) = -A_k^* b^{(k)}\,,$$

und falls die Behauptung falsch ist, existieren $\varepsilon > 0$ und $k(\varepsilon) \in \mathbb{N}$ mit

$$\|\operatorname{grad}\Phi(x^{(k)})\|_2 \;=\; \|A_k^* b^{(k)}\|_2 \;\geq\; \varepsilon \;>\; 0\,, \qquad k \geq k(\varepsilon)\,. \tag{21.16}$$

Aus (21.15) ergibt sich dann unmittelbar

$$\rho_k \to 0\,, \qquad k \to \infty\,, \tag{21.17}$$

womit wir aber aus der Definition (21.9) von μ_k durch Taylorentwicklung erhalten, daß

$$\mu_k \;=\; \frac{h^{(k)*}F'(x^{(k)})^*F(x^{(k)}) + O(\|h^{(k)}\|_2^2)}{h^{(k)*}F'(x^{(k)})^*F(x^{(k)})} \;=\; 1 \,+\, O(\rho_k^2/\nu_k)$$

$$\overset{(21.13)}{=} 1 \,+\, O(\rho_k/\|A_k^* b^{(k)}\|_2) \overset{(21.16)}{=} 1 \,+\, O(\rho_k)\,, \qquad k \to \infty\,.$$

Also existiert ein $k_0 \in \mathbb{N}$ mit $k_0 \geq k(\varepsilon)$ und $\mu_k > \mu_+$ für alle $k \geq k_0$, d. h. ab dem k_0-ten Iterationsschritt würde ρ_k aufgrund von Algorithmus 21.1 in jedem Schritt verdoppelt – im Widerspruch zur obigen Folgerung (21.17).

5. Wir beweisen nun die Aussage des Satzes, wiederum durch indirekte Beweisführung, und nehmen dazu an, daß eine Teilfolge von $\{A_k^* b^{(k)}\}$ nicht gegen Null konvergiert. Nach der im 4. Schritt bewiesenen Aussage existiert dann ein $\varepsilon > 0$ und zwei Indizes m und l mit

$$\begin{aligned} &\|A_l^* b^{(l)}\|_2 \geq 2\varepsilon\,, \qquad \|A_m^* b^{(m)}\|_2 \leq \varepsilon\,, \\ &\|A_k^* b^{(k)}\|_2 > \varepsilon\,, \qquad k = l+1, \ldots, m-1\,. \end{aligned} \tag{21.18}$$

Da $\{\, \|F(x^{(k)})\| \,\}$ eine Cauchy-Folge ist, kann l dabei so groß gewählt werden, daß

$$\|F(x^{(l)})\|_2^2 - \|F(x^{(m)})\|_2^2 \;<\; \frac{2\mu_-}{c^2+1}\,\frac{\varepsilon^2}{G}\,, \tag{21.19}$$

wobei G eine Lipschitz-Konstante von $F'(x)^*F(x)$ in $\mathcal{U}$ bezeichne; ohne Beschränkung der Allgemeinheit wählen wir $G > 1$. Wegen $\|x^{(k+1)} - x^{(k)}\|_2 \leq \rho_k$ folgt aus (21.14) und (21.18) für erfolgreiche Schritte $k \in \{l, \ldots, m-1\} \cap \mathcal{K}$ die Ungleichung

$$\|F(x^{(k)})\|_2^2 - \|F(x^{(k+1)})\|_2^2 \;\geq\; \frac{2\mu_-}{c^2+1}\,\varepsilon\,\min\{\,\|x^{(k+1)} - x^{(k)}\|_2, \varepsilon\}\,.$$

Dies gilt trivialerweise auch für erfolglose Iterationsschritte, da dann $x^{(k)} = x^{(k+1)}$ ist. Durch Summation von $k = l$ bis $m-1$ erhalten wir

$$\frac{2\mu_-}{c^2+1}\,\varepsilon \sum_{k=l}^{m-1} \min\{\,\|x^{(k+1)} - x^{(k)}\|_2, \varepsilon\} \\ \leq\ \|F(x^{(l)})\|_2^2 - \|F(x^{(m)})\|_2^2 \overset{(21.19)}{<} \frac{2\mu_-}{c^2+1}\,\frac{\varepsilon^2}{G}\,,$$

und dies kann wegen $G > 1$ offensichtlich nur dann erfüllt sein, wenn in jedem einzelnen Summanden

$$\min\{\,\|x^{(k+1)} - x^{(k)}\|_2, \varepsilon\} \ =\ \|x^{(k+1)} - x^{(k)}\|_2$$

und insgesamt

$$\sum_{k=l}^{m-1} \|x^{(k+1)} - x^{(k)}\|_2 \ <\ \varepsilon/G$$

gilt. Die linke Seite dieser Ungleichung kann mit der Dreiecksungleichung nach unten durch $\|x^{(m)} - x^{(l)}\|_2$ abgeschätzt werden und daher ist $\|x^{(m)} - x^{(l)}\|_2 < \varepsilon/G$. Wegen der Lipschitz-Stetigkeit von $\operatorname{grad}\Phi = F'^{*}F$ ergibt sich somit

$$\|A_m^* b^{(m)} - A_l^* b^{(l)}\|_2 \ =\ \|\operatorname{grad}\Phi(x^{(m)}) - \operatorname{grad}\Phi(x^{(l)})\|_2 \\ \leq\ G\|x^{(m)} - x^{(l)}\|_2 \ <\ \varepsilon$$

im Widerspruch zur Annahme (21.18). Damit ist der Satz bewiesen. □

Die Bemerkung im Anschluß an Satz 20.1 gilt sinngemäß auch für das Levenberg-Marquardt-Verfahren.

Beispiel. Die Iterierten des Levenberg-Marquardt-Verfahrens für Beispiel 20.2 haben wir bereits im vorigen Abschnitt im rechten Teil der Abbildung 20.4 gesehen. Für dieses Beispiel wurde in Algorithmus 21.1 der gleiche Startwert wie bei dem Verfahren des steilsten Abstiegs gewählt sowie die Parameter $\mu_- = 1/4$ und $\mu_+ = 1/2$. Das gleiche Argument wie in Abschnitt 20 kann auch hier verwendet werden, um die Konvergenz der Levenberg-Marquardt-Iterierten gegen den einzigen stationären Punkt $\widehat{x} = [1,1]^T$ nachzuweisen. Abbildung 21.2 demonstriert in anderer Form die Überlegenheit des Levenberg-Marquardt-Verfahrens: Der Grenzwert wird bereits nach neun Schritten mit hinreichender Genauigkeit erreicht. Auf der anderen Seite ist ein einzelner Iterationsschritt wesentlich teurer als bei der Methode des steilsten Abstiegs, da lineare Gleichungssysteme zu lösen sind. Anhand der Abbildung wird man

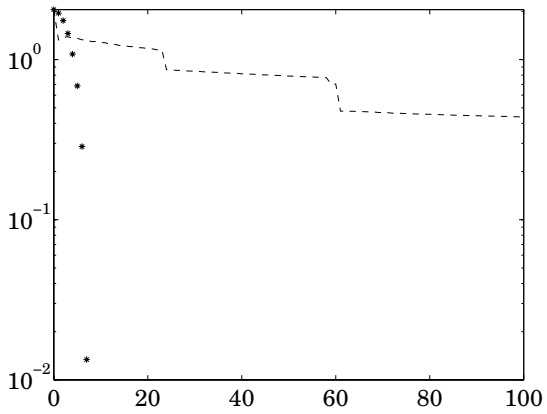


Abb. 21.2: Konvergenzverlauf des Verfahrens des steilsten Abstiegs und des Levenberg-Marquardt-Verfahrens (Sterne)

für das Levenberg-Marquardt-Verfahren superlineare Konvergenz vermuten. In diesem Fall ist $F(\widehat{x}) = 0$ und $F'(\widehat{x})$ nicht singulär und das Levenberg-Marquardt-Verfahren geht in einer Umgebung von $\widehat{x}$ in das Newton-Verfahren über. Daher ist die Konvergenz tatsächlich lokal quadratisch. $\Diamond$

Aufgaben

1. Zeigen Sie, daß das allgemeine Heron-Verfahren mit $\nu > 2$ für jedes $a > 0$ und jeden positiven Startwert x_0 konvergiert. Betrachten Sie im Fall $\nu = 3$ mit positivem a auch negative Startwerte. Was können Sie hier über die Konvergenz sagen?

2. Die stetig differenzierbare Funktion $\Phi : \mathcal{D}(\Phi) \to \mathbb{R}^n$ mit Fixpunkt $\widehat{x}$ erfülle die Voraussetzungen des Banachschen Fixpunktsatzes. Zeigen Sie, daß der asymptotische Konvergenzfaktor

$$q = \max_{x^{(0)} \in \mathcal{D}(\Phi)} \limsup_{k \to \infty} \|\widehat{x} - x^{(k)}\|^{1/k}$$

durch $q \le \varrho(\Phi'(\widehat{x}))$ abgeschätzt werden kann. Ist diese Abschätzung scharf?

3. Formulieren Sie hinreichende Bedingungen an die Funktion f, damit das Newton-Verfahren für die Nullstellengleichung $f(x) = 0$ Konvergenzordnung $p \ge 3$ besitzt. Interpretieren Sie Ihre Bedingungen geometrisch.

4. Die $p+1$-mal stetig differenzierbare Funktion f habe in $\widehat{x}$ eine p-fache Nullstelle mit $p \in \mathbb{N} \setminus \{1\}$. Zur Approximation von $\widehat{x}$ kann das folgende Verfahren verwendet werden:

$$x_{k+1} = x_k - p\,\frac{f(x_k)}{f'(x_k)}\,, \qquad k = 0, 1, 2, \ldots\,.$$

(a) Bestimmen Sie die (lokale) Konvergenzordnung dieser Iteration.
(b) Zeigen Sie, daß die Iteration mit dem Newton-Verfahren, angewandt auf die Funktion $g = f^{1/p}$ übereinstimmt.
(c) Wie lautet das Newton-Verfahren, angewandt auf die Funktion $h = f/f'$, und welche Konvergenzordnung hat es?

5. $f : \mathbb{R} \to \mathbb{R}$ sei differenzierbar, streng monoton wachsend und konvex. Ferner habe f eine Nullstelle $\widehat{x}$ und für gegebene $z_0 < \widehat{x} < x_0$ bezeichne $\{x_k\}$ die Folge der Newton-Iterierten und $\{z_k\}$ die Folge der Iterierten aus (18.4). Zeigen Sie, daß die Folge $\{z_k\}$ monoton wachsend gegen $\widehat{x}$ konvergiert.

6. Diskutieren Sie die Konvergenz des Newton-Verfahrens für die Funktion

$$f(x) = xe^{-x}$$

für alle (zulässigen) positiven Startwerte.

7. Implementieren Sie das eindimensionale Newton-Verfahren.
(a) Wenden Sie Ihr Programm auf die Testfunktion

$$f(x) = e^{1-x} - 1$$

an. Verwenden Sie $x_0 = 10$ als Startnäherung;
(b) Verwenden Sie

$$f(x) = \frac{11}{91}x^5 - \frac{38}{91}x^3 + x$$

als Testfunktion und $x_0 = 1.01$ als Startnäherung.

Interpretieren Sie die Ergebnisse. Überlegen Sie sich eine Schrittweitensteuerung, um ein etwaiges Fehlverhalten zu verhindern. Vergleichen Sie die Resultate mit und ohne Schrittweitensteuerung.

8. Sei f die Funktion aus (18.10). Zeigen Sie, daß die Iterierte x_{k+1} des Hebden-Verfahrens die Gleichung

$$h(x) := \frac{\zeta}{(\delta + x)^2} = \rho$$

löst, wobei ζ und δ so gewählt sind, daß sich h und f im Punkt x_k berühren, d. h.

$$h(x_k) = f(x_k)\,, \qquad h'(x_k) = f'(x_k)\,.$$

9. Gegeben sei das nichtlineare Gleichungssystem

$$x_1 x_2 = 0\,, \qquad x_1 x_2^2 + x_1 - x_2 = 0\,,$$

mit der eindeutigen Lösung $\widehat{x} = 0$.
(a) Überprüfen Sie die Voraussetzungen von Satz 19.1. Ist das Newton-Verfahren für dieses Beispiel lokal konvergent?
(b) Zeigen Sie, daß das Newton-Verfahren für jeden Startwert $x^{(0)} \in [0,1]^2$ gegen $\widehat{x}$ konvergiert.
(c) Bestimmen Sie für diesen Fall die Konvergenzordnung.
(d) Implementieren Sie das Newton-Verfahren und plotten Sie die Iterierten für verschiedene Startwerte in das Einheitsquadrat. Diskutieren Sie das Ergebnis.

10. Führen Sie den Beweis von Satz 20.1 unter den dort angegebenen allgemeinen Voraussetzungen an die Suchrichtungen $d^{(k)}$.

11. $b \in \mathbb{R}^m$ und $\lambda > 0$. Ferner sei h die Lösung von $(A^*A + \lambda I)h = A^*b$ und g die Lösung von $(A^*A + \lambda I)g = h$.
(a) Zeigen Sie, daß h und g die linearen Ausgleichsprobleme

$$\text{minimiere} \quad \left\| \begin{bmatrix} A \\ \sqrt{\lambda} I \end{bmatrix} h - \begin{bmatrix} b \\ 0 \end{bmatrix} \right\|_2 \quad \text{bzw.} \quad \text{minimiere} \quad \left\| \begin{bmatrix} A \\ \sqrt{\lambda} I \end{bmatrix} g - \begin{bmatrix} 0 \\ h/\sqrt{\lambda} \end{bmatrix} \right\|_2$$

lösen.
(b) In Aufgabe III.7 wurde ein Algorithmus beschrieben, um eine $m \times n$-Matrix mit orthogonalen Transformationen auf Bidiagonalgestalt B zu bringen. Überlegen Sie sich, wie Sie mit den Techniken aus Abschnitt 14 (Givens-Rotationen) eine QR-Faktorisierung der resultierenden Blockmatrix

$$\begin{bmatrix} B \\ \sqrt{\lambda} I \end{bmatrix} = QR$$

bestimmen können. Zeigen Sie, daß für eine solche Faktorisierung $2n$ Givens-Rotationen benötigt werden.
(c) Fügen Sie diese Unterprogramme zu einem Algorithmus zur Berechnung von h und g zusammen. Was kann wiederverwendet werden, und was muß neu berechnet werden, wenn sich λ ändert?

12. Gegeben sei die Funktion

$$F : \mathbb{R} \to \mathbb{R}^2, \qquad F(x) = \begin{bmatrix} x+1 \\ \lambda x^2 + x - 1 \end{bmatrix},$$

wobei λ ein reeller Parameter sei.
(a) Zeigen Sie, daß das nichtlineare Ausgleichsproblem

$$\text{minimiere} \quad \Phi(x) = \frac{1}{2}\|F(x)\|_2^2$$

für $\lambda < 1$ in $x = 0$ ein lokales Minimum besitzt. Zeigen Sie ferner, daß dies für $\lambda < 7/16$ das einzige lokale Minimum ist.
(b) Weisen Sie nach, daß $x = 0$ für $\lambda < -1$ ein *abstoßender Fixpunkt* des Gauß-Newton-Verfahrens (21.3) ist, d. h. es gibt ein $\delta > 0$, so daß

$$|x_{k+1} - 0| > |x_k - 0| \qquad \text{für alle } x_k \text{ mit } \ 0 < |x_k - 0| < \delta\,.$$

(c) Zeigen Sie, daß andererseits die Iterierten des Levenberg-Marquardt-Verfahrens für $\lambda < 7/16$ global gegen $x = 0$ konvergieren.
(d) Implementieren Sie das Levenberg-Marquardt-Verfahren für dieses Beispiel. Verwenden Sie den Startwert $x_0 = 10$ und experimentieren Sie mit verschiedenen Parametern $\lambda < 7/16$. Welche Konvergenzgeschwindigkeit läßt sich beobachten?

13. Implementieren Sie das Verfahren des steilsten Abstiegs und das Levenberg-Marquardt-Verfahren, und wenden Sie beide auf das nichtlineare Ausgleichsgeradenproblem aus Aufgabe III.2 an.

V Eigenwerte

Für die Entwicklung effizienter Algorithmen ist die Struktur des zugrundeliegenden Problems von entscheidender Bedeutung. Ein gutes Beispiel hierfür sind Eigenwertprobleme: Das Eigenwertproblem für eine Matrix $A \in \mathbb{K}^{n\times n}$,

$$Ax = \lambda x\,, \qquad x \in \mathbb{C}^n \setminus \{0\}\,,\ \lambda \in \mathbb{C}\,,$$

ist *nichtlinear*, denn die Unbekannten λ und x treten im Produkt auf. Trotzdem verwendet die gängige Software zur Berechnung von λ und/oder x keines der Verfahren aus dem vorangegangenen Kapitel.

Algorithmen für die Lösung des Eigenwertproblems und für die Lösung nichtlinearer Gleichungen haben jedoch wesentliche Eigenschaften gemeinsam: Die Methoden sind fast ausschließlich iterativ und die besseren Verfahren konvergieren lokal quadratisch oder gar kubisch. Nicht immer ist die Konvergenz jedoch global, so daß auch die Suche nach guten Startnäherungen und die Stabilität der Lösung angesprochen werden muß.

Neben dem bereits in Kapitel II genannten Buch von Golub und Van Loan [34] sei hier noch das Buch von Demmel [22] hervorgehoben, das auch einen guten Überblick über die zur Verfügung stehende Software bietet.

22 Wozu werden Eigenwerte berechnet ?

Das Programmpaket MATLAB enthält ein Animationsprogramm `truss`, das die natürlichen Eigenschwingungen einer zweidimensionalen Brücke vorführt. Zur Berechnung dieser Schwingungen müssen die Eigenvektoren und Eigenwerte einer Matrix bestimmt werden, was im folgenden erläutert werden soll. Nicht zuletzt aus diesem Grund haben wir bereits für das Beispiel in Abschnitt 3 jene Brücke ausgewählt.

Während in Abschnitt 3 nur das statische Gleichgewicht in Gegenwart äußerer Kräfte untersucht wurde, betrachten wir nun das dynamische Verhalten des Tragwerks. Dazu betrachten wir die Positionen z_i, $i = 1, \ldots, 8$, der Gelenke als

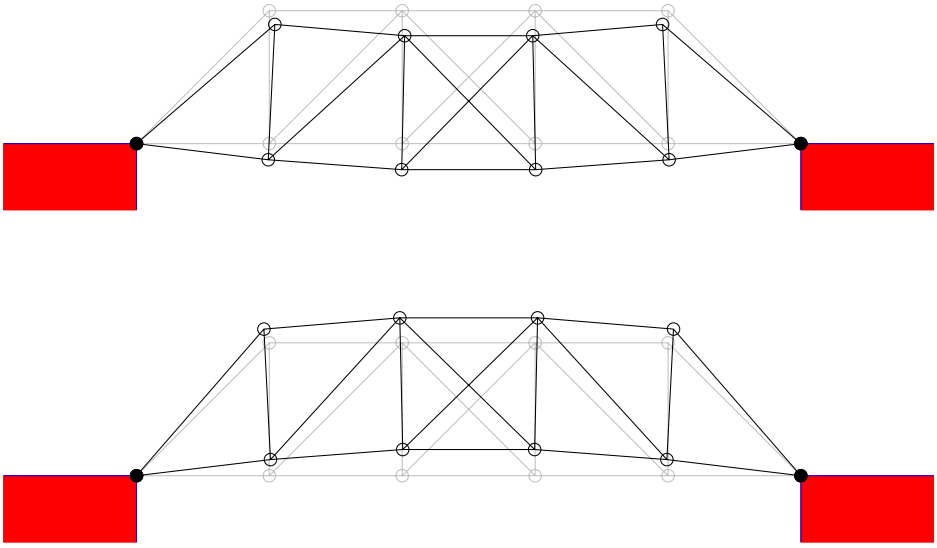


Abb. 22.1: Langsamste Eigenschwingung der Brücke

Funktionen der Zeit, $z_i = z_i(t)$, die jeweils eine Kurve in $\mathbb{R}^2$ beschreiben. Die Ableitungen $z_i'(t)$ geben die zugehörigen Geschwindigkeitsvektoren, die zweiten Ableitungen $z_i''(t)$ die Beschleunigungsvektoren zum Zeitpunkt t an. Im folgenden werden alle acht Gelenke gemeinsam betrachtet und ihre Koordinaten wie zuvor in einem Vektor $z = z(t) \in \mathbb{R}^{16}$ zusammengefaßt.

Bei einer Auslenkung $x(t) = z(t) - z^{(0)}$ der Gelenke aus dem statischen Gleichgewichtszustand $z^{(0)}$ ergibt sich eine Rückstellkraft $-Ax(t)$, vgl. Abschnitt 3, die nach dem Newtonschen Gesetz

$$\text{Kraft} \;=\; \text{Masse} \cdot \text{Beschleunigung} \tag{22.1}$$

eine Beschleunigung der Gelenke

$$mx''(t) \;=\; -Ax(t) \tag{22.2}$$

nach sich zieht; m ist wieder die Masse der einzelnen Gelenke. (22.2) ist ein Differentialgleichungssystem zweiter Ordnung für die Verschiebungsvektoren $x(t)$.

Differentialgleichungen werden erst später in diesem Buch behandelt, daher soll an dieser Stelle auf die allgemeine Lösungstheorie nicht weiter eingegangen

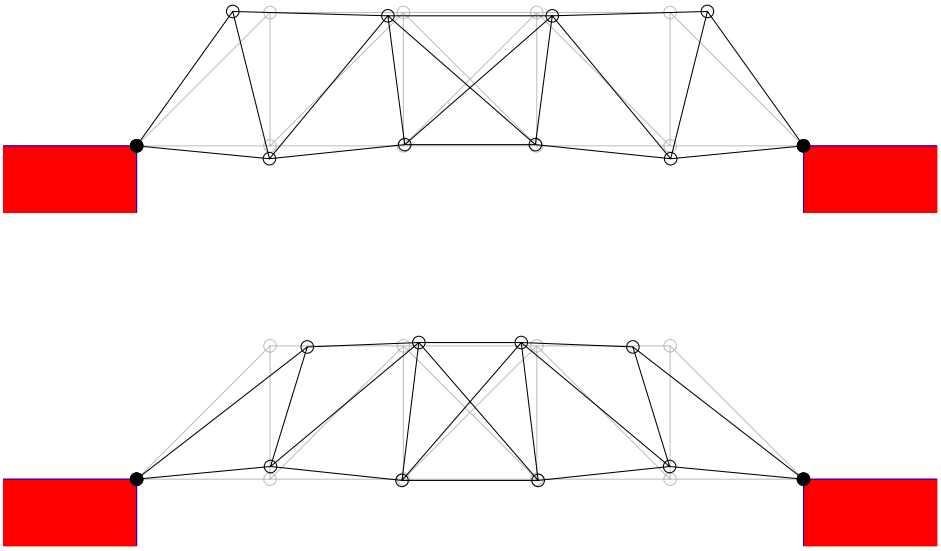


Abb. 22.2: Eine andere Eigenschwingung der Brücke

werden. Für unsere Zwecke reicht es aus, spezielle Lösungen dieser Gleichung zu betrachten. Hierzu benötigen wir die Eigenwerte und Eigenvektoren von A. Da A symmetrisch und positiv definit ist, existiert eine Orthogonalbasis des $\mathbb{R}^{16}$ aus Eigenvektoren von A und die zugehörigen Eigenwerte sind positiv. Sei v ein solcher Eigenvektor und λ der zugehörige Eigenwert. Normieren wir der Einfachheit halber die Masse m auf Eins, dann ist

$$x(t) \;=\; \cos(\sqrt{\lambda}\,t)\,v$$

offensichtlich eine Lösung von (22.2): In diesem Fall ist

$$x'(t) \;=\; -\sqrt{\lambda}\,\sin(\sqrt{\lambda}\,t)\,v\,, \qquad x''(t) \;=\; -\lambda\cos(\sqrt{\lambda}\,t)\,v\,,$$

und da v ein Eigenvektor von A ist, ergibt $-Ax(t)$ ebenfalls $-\lambda\cos(\sqrt{\lambda}\,t)\,v$. Die zugehörige Funktion

$$z(t) \;=\; z^{(0)} + \cos(\sqrt{\lambda}\,t)\,v \tag{22.3}$$

beschreibt eine kosinusförmige Oszillation der Brücke um den Gleichgewichtszustand $z^{(0)}$ mit Periodendauer $T = 2\pi/\sqrt{\lambda}$. Je größer der Eigenwert λ ist, desto schneller schwingt die Brücke. Die Abbildungen 22.1 und 22.2 stellen die

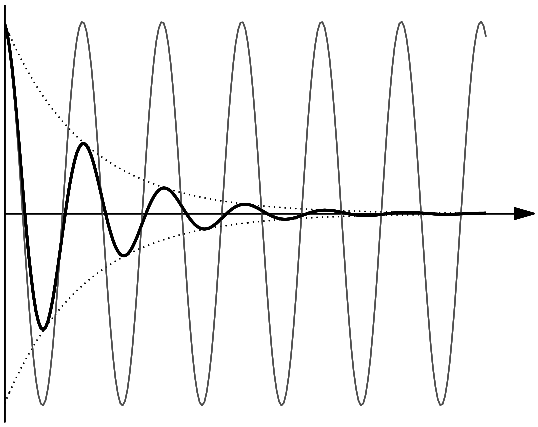


Abb. 22.3: Zeitlicher Verlauf einer Eigenschwingung mit und ohne Dämpfung

beiden extremalen Positionen (für $t = k\pi/\sqrt{\lambda}$, $k \in \mathbb{Z}$) der Brücke für zwei der insgesamt sechzehn möglichen Eigenschwingungen dar. Die zugehörigen Verschiebungen werden durch die jeweiligen Eigenvektoren beschrieben und mit positivem beziehungsweise negativem Vorzeichen zum Gleichgewichtszustand $z^{(0)}$ addiert. Das MATLAB-Programm `truss` animiert diese Oszillationen.

Gemäß (22.3) würde eine einmal angeregte Brücke für alle Zeiten gleichbleibend auf und ab schwingen – ein wenig realistisches Modell. Tatsächlich erwartet man in der Praxis aufgrund von Reibungswiderständen eine abklingende Oszillation. Derartige Widerstandskräfte wirken entgegengesetzt zu der Bewegung und sind im einfachsten Fall proportional zur vorliegenden Geschwindigkeit. Zu der Rückstellkraft $-Ax(t)$ kommt also noch eine weitere Kraft $-dx'(t)$ mit Dämpfungskonstante $d > 0$ hinzu. Das Newtonsche Gesetz (22.1) führt bei diesem erweiterten Modell auf die Differentialgleichung

$$mx''(t) \;=\; -dx'(t) - Ax(t)\,. \tag{22.4}$$

Wieder setzen wir $m = 1$ und nehmen an, daß d hinreichend klein ist. In diesem Fall lautet eine typische Lösung der Differentialgleichung (22.4)

$$x(t) \;=\; e^{-d\,t/2}\cos(\omega t)\,v \qquad \text{mit} \quad \omega \;=\; \sqrt{\lambda - d^2/4} \;>\; 0\,, \tag{22.5}$$

wie man mit etwas mehr Aufwand als zuvor nachrechnet. Erneut sind es die Eigenvektoren, die die Form der Eigenschwingungen bestimmen. Allerdings sind die Oszillationen nun etwas langsamer ($\omega < \sqrt{\lambda}$) und vor allem werden die Amplituden mit wachsender Zeit immer stärker gedämpft, vgl. die dickere Kurve in Abbildung 22.3. Die Amplitudenfunktion $e^{-dt/2}$ ist zur Veranschaulichung

gepunktet dargestellt. Die dünnere Kurve im Hintergrund zeigt schließlich die entsprechende Lösung für $d = 0$.

Reibungswiderstände führen also dazu, daß einmalige Störungen des Gleichgewichtszustands nur kurze Zeit meßbare Auswirkungen haben. Die Situation ist völlig anders, wenn zeitabhängige äußere Kräfte das statische Gleichgewicht der Brücke stören. Beispielhaft betrachten wir eine periodische Erregung mit Frequenz ω_0 in einer Eigenrichtung v von A die etwa durch den Verkehr über die Brücke verursacht werden könnte. Das zugehörige vollständige Differentialgleichungsmodell lautet dann

$$x''(t) \;=\; -dx'(t) - Ax(t) + \cos(\omega_0 t)\, v\,. \tag{22.6}$$

Solche Kräfte führen selbst im gedämpften Fall auf periodische, nicht abklingende Lösungen $x(t)$, und zwar mit der gleichen Frequenz ω_0, vgl. Aufgabe 1. Ist λ der zu v gehörende Eigenwert von A, dann ist für $\omega_0 \approx \sqrt{\lambda}$ die Amplitude der resultierenden Lösung besonders groß: man spricht dann von *Resonanz*, $\sqrt{\lambda}$ wird als *Eigenfrequenz* der Brücke bezeichnet. Für $\omega_0 = \sqrt{\lambda}$ wird die Amplitude am größten, und in diesem Fall verhalten sich alle Lösungen der Differentialgleichung (22.6) für große Zeiten t wie die spezielle Lösung

$$x(t) \;=\; \frac{1}{d\sqrt{\lambda}}\, \sin(\sqrt{\lambda}\, t)\, v\,. \tag{22.7}$$

In Abhängigkeit von dem Dämpfungsfaktor d können diese Resonanzen beliebig groß werden.

Die Lösung (22.7) bezieht sich auf den gedämpften Fall ($d \neq 0$). Sofern die Störfrequenz ω_0 von $\sqrt{\lambda}$ verschieden ist, ergibt sich auch im ungedämpften Fall eine periodische Lösung mit Frequenz ω_0. Sobald jedoch Resonanz eintritt, also wenn $\omega_0 = \sqrt{\lambda}$ ist, schaukeln sich die Schwingungen auf und die Brücke wird instabil. Eine solche Lösung lautet

$$x(t) \;=\; \frac{1}{2\sqrt{\lambda}}\, t\, \sin(\sqrt{\lambda}\, t)\, v\,. \tag{22.8}$$

Abbildung 22.4 illustriert das Schwingungsverhalten im Resonanzfall. Die dünnere Kurve im Hintergrund zeigt die sich aufschaukelnde Lösung (22.8) für den reibungsfreien Fall ($d = 0$), die dickere Kurve zeigt die entsprechende Lösung im Fall $d \neq 0$, die asymptotisch in die periodische Lösung (22.7) einschwingt.

Aufgrund der großen Amplitude in (22.7) können Resonanzphänomene selbst im gedämpften Fall zu Instabilitäten führen. Besonders fragwürdige Popularität erwarb sich auf diese Weise die Tacoma-Brücke (USA), die 1940 durch

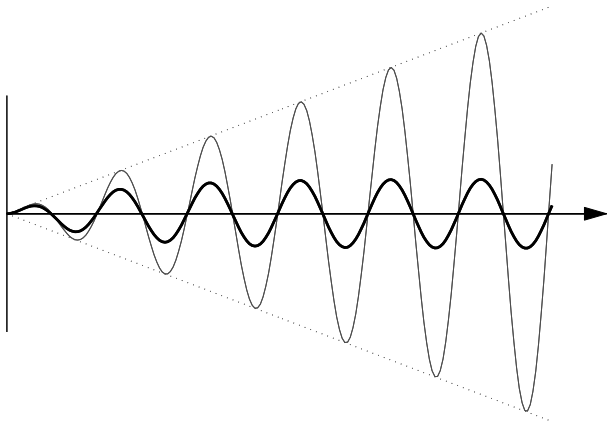


Abb. 22.4: Resonanzverhalten mit und ohne Dämpfung

Windturbulenzen zunächst in starke Schwingungen versetzt wurde und schließlich einstürzte;[1] die Tacoma-Brücke war allerdings eine Hängebrücke, die durch das obige Modell nicht beschrieben wird, vgl. statt dessen etwa McKenna [73].

Aus den genannten Gründen gehört bei der Konstruktion eines Bauwerks die Spektralanalyse der zugehörigen Steifigkeitsmatrix zu den zentralen Aufgaben der verantwortlichen Ingenieure. Dabei muß beachtet werden, daß mögliche Erregerschwingungen außerhalb der Resonanzbereiche der Steifigkeitsmatrix liegen. Vor allem die kleinen Eigenwerte sind dabei kritisch, da diese reziprok zu ihrer Wurzel in die Amplitudenfaktoren in (22.7) und (22.8) eingehen.

23 Eigenwerteinschließungen

Nach dieser Einführung stellen wir nun eine Reihe von Verfahren vor, mit denen das Spektrum einer Matrix $A \in \mathbb{K}^{n\times n}$ grob eingegrenzt werden kann. Solche Abschätzungen können später zur Wahl geeigneter Startwerte für lokal konvergente Iterationsverfahren ausgenutzt werden.

Zuvor jedoch stellen wir ohne Beweis einige wichtige theoretische Resultate über Eigenwerte und Eigenvektoren zusammen. Das *charakteristische Polynom* $p(\lambda) = \det(A - \lambda I)$ ist ein (komplexwertiges) Polynom über $\mathbb{C}$ vom Grad n. Jede der n (komplexen) Nullstellen von p ist ein *Eigenwert* von A, d. h. zu einer solchen Nullstelle λ gibt es einen *Eigenvektor* $x \in \mathbb{C}^n \setminus \{0\}$ mit $Ax = \lambda x$;

[1] Unter `http://www.gallopinggertie.com/` findet sich im Internet ein Video dieses Einsturzes.

umgekehrt ist auch jeder Eigenwert eine Nullstelle von p.

Ist $\lambda \in \sigma(A)$, dann ist $\overline{\lambda} \in \sigma(A^*)$. Folglich gibt es einen Vektor $y \neq 0$ mit $A^*y = \overline{\lambda}y$ und es gilt

$$\lambda y^* = (\overline{\lambda}y)^* = (A^*y)^* = y^*A\,.$$

Daher heißt y auch *linker Eigenvektor* von A zu λ. Aus $y \in \mathcal{N}((A-\lambda I)^*)$ und Lemma 11.2 folgt, daß y senkrecht auf $\mathcal{R}(A-\lambda I)$ steht. Insbesondere ist also y orthogonal zu allen Eigenvektoren von A mit Eigenwerten $\tilde{\lambda} \neq \lambda$. Ist hingegen λ ein einfacher Eigenwert von A und $x \neq 0$ ein zugehöriger Eigenvektor, dann ist $y^*x \neq 0$.

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

$$A\overline{x} = \overline{Ax} = \overline{\lambda x} = \overline{\lambda}\overline{x}\,.$$

Zuvor haben wir bereits erwähnt, daß die Eigenwerte einer hermiteschen Matrix $A = A^*$ allesamt reell sind. Die zugehörigen Eigenvektoren bilden eine Orthogonalbasis des $\mathbb{K}^n$. Sind A und B hermitesch, dann hat auch das Produkt AB ausschließlich reelle Eigenwerte, vgl. Aufgabe 2.

Als ein erstes Einschließungsresultat formulieren wir den wichtigen Satz von Gerschgorin.

Satz 23.1 (Satz von Gerschgorin). *Sei $A = [a_{ij}] \in \mathbb{K}^{n\times n}$ und λ ein beliebiger Eigenwert von A. Dann gilt*

$$\lambda \;\in\; \bigcup_{i=1}^{n} \mathcal{K}_i \;=\; \bigcup_{i=1}^{n} \Big\{\, \zeta \;:\; |\zeta - a_{ii}| \le \sum_{\substack{j=1\\ j\neq i}}^{n} |a_{ij}| \,\Big\}\,. \tag{23.1}$$

Beweis. Sei $Ax = \lambda x$ mit $x = [x_i] \neq 0$. Dann existiert ein x_i mit $|x_j| \le |x_i|$ für alle $j \neq i$. Bezeichnet $(Ax)_i$ die i-te Komponente von Ax, dann ist

$$\lambda x_i = (Ax)_i = \sum_{j=1}^{n} a_{ij}x_j\,,$$

und somit folgt

$$|\lambda - a_{ii}| \;=\; \Big|\sum_{\substack{j=1\\ j\neq i}}^{n} a_{ij}\,\frac{x_j}{x_i}\Big| \;\le\; \sum_{\substack{j=1\\ j\neq i}}^{n} |a_{ij}|\,.$$

Also ist $\lambda \in \mathcal{K}_i \subset \bigcup_{j=1}^{n} \mathcal{K}_j$. □

Für $\overline{\lambda} \in \sigma(A^*)$ gilt der Satz von Gerschgorin entsprechend, nämlich

$$\overline{\lambda} \in \bigcup_{i=1}^{n} \{ \zeta : |\zeta - \overline{a}_{ii}| \le \sum_{\substack{j=1 \\ j\neq i}}^{n} |a_{ji}| \}$$

oder äquivalent

$$\lambda \in \bigcup_{i=1}^{n} \mathcal{K}_i^* := \bigcup_{i=1}^{n} \{ \zeta : |\zeta - a_{ii}| \le \sum_{\substack{j=1 \\ j\neq i}}^{n} |a_{ji}| \} . \tag{23.2}$$

Die Kreise $\mathcal{K}_i$ und $\mathcal{K}_i^*$ werden *Gerschgorin-Kreise* genannt.

Die Aussage von Satz 4.6 beinhaltet, daß strikt diagonaldominante Matrizen nicht singulär sind. Dies kann auch unmittelbar als Folgerung aus dem Satz von Gerschgorin geschlossen werden: Da für solche Matrizen in jeder Zeile die Betragssumme der Nebendiagonalelemente kleiner als der Betrag des Diagonalelements ist, liegt $\zeta = 0$ in keinem der Gerschgorin-Kreise $\mathcal{K}_i$. Etwas interessanter ist der folgende Fall, in dem $\zeta = 0$ Randpunkt eines der Gerschgorin-Kreise ist. Derartige Matrizen treten zum Beispiel bei Randwertaufgaben auf, vgl. Abschnitt 83.

Definition 23.2. Eine Tridiagonalmatrix $A = [a_{ij}] \in \mathbb{K}^{n\times n}$ heißt *irreduzibel*, falls alle Elemente auf der oberen und der unteren Nebendiagonalen von Null verschieden sind und *irreduzibel diagonaldominant*, falls A irreduzibel ist,

$$|a_{ii}| \ge \sum_{j\neq i} |a_{ij}| \qquad \text{für alle } i = 1, \ldots, n \tag{23.3}$$

gilt und dabei für mindestens ein i die echte Ungleichung in (23.3) erfüllt ist.

Im Gegensatz zu stark diagonaldominanten Matrizen kann bei irreduzibel diagonaldominanten Tridiagonalmatrizen der Wert $\zeta = 0$ (als Randpunkt) zu mehreren Gerschgorin-Kreisen gehören, allerdings nicht zu allen. Dennoch kann $\zeta = 0$ kein Eigenwert sein, wie der folgende Satz zeigt.

Satz 23.3. *Jede irreduzibel diagonaldominante Tridiagonalmatrix ist nichtsingulär.*

Beweis. Der Beweis geht analog zu dem Beweis von Satz 23.1. Wir nehmen an, daß A eine irreduzibel diagonaldominante Tridiagonalmatrix und x ein Vektor aus $\mathcal{N}(A) \setminus \{0\}$ ist. Ferner sei x_i eine Komponente von x mit maximalem Betrag, $|x_i| = \|x\|_\infty \neq 0$. Um die Beweisführung zu vereinfachen, gehen wir davon aus, daß x_i weder die erste noch die letzte Komponente ist. Die i-te Zeile von $Ax = 0$ ergibt dann

$$a_{ii}x_i = -a_{i,i-1}x_{i-1} - a_{i,i+1}x_{i+1} ,$$

und daraus folgt aufgrund von (23.3)

$$\begin{aligned} |a_{ii}|\,\|x\|_\infty \;=\; |a_{ii}|\,|x_i| \;&\le\; |a_{i,i-1}|\,|x_{i-1}| \,+\, |a_{i,i+1}|\,|x_{i+1}| \\ &\le\; \|x\|_\infty\big(|a_{i,i-1}| + |a_{i,i+1}|\big) \;\le\; \|x\|_\infty |a_{ii}|\,. \end{aligned}$$

Demnach gilt überall das Gleichheitszeichen, das heißt

$$|x_{i-1}| = |x_{i+1}| = |x_i| = \|x\|_\infty$$

und

$$|a_{i,i-1}| + |a_{i,i+1}| \;=\; |a_{ii}|\,. \tag{23.4}$$

Da für diese Herleitung ein beliebiger Index i mit $|x_i| = \|x\|_\infty$ gewählt werden kann, folgt induktiv, daß alle Einträge von x den gleichen Betrag aufweisen. Da jedoch für mindestens einen Index i in (23.3) kein Gleichheitszeichen steht, ergibt sich mit (23.4) ein Widerspruch, d. h. A ist nicht singulär.

Für den Fall, daß etwa die erste Komponente von x maximalen Betrag aufweist, folgt aus der ersten Zeile von $Ax = 0$ entsprechend

$$|a_{11}|\|x\|_\infty \;=\; |a_{11}||x_1| \;=\; |a_{12}||x_2| \;\le\; |a_{12}|\|x\|_\infty \overset{(23.3)}{\le} |a_{11}|\|x\|_\infty\,,$$

und wieder muß überall das Gleichheitszeichen gelten. Insbesondere ist also $|x_2| = \|x\|_\infty$ und nach dem bereits behandelten Fall steht dies im Widerspruch zu den Voraussetzungen. □

Weitere Einschließungsresultate für das Spektrum $\sigma(A)$ beruhen auf dem Konzept des Wertebereichs einer Matrix.

Definition 23.4. Unter dem *Wertebereich* einer Matrix $A \in \mathbb{K}^{n\times n}$ versteht man die Menge aller *Rayleigh-Quotienten* x^*Ax/x^*x mit $x \in \mathbb{C}^n \setminus \{0\}$,

$$\begin{aligned} \mathcal{W}(A) \;&=\; \Big\{\, \zeta = \frac{x^*Ax}{x^*x} \;:\; x \in \mathbb{C}^n \setminus \{0\} \,\Big\} \\ &=\; \Big\{\, \zeta = x^*Ax \;:\; x \in \mathbb{C}^n\,,\, \|x\|_2 = 1 \,\Big\} \;\subset\; \mathbb{C}\,. \end{aligned}$$

In dieser Definition ist es wesentlich, daß x alle *komplexen* Vektoren durchläuft, selbst dann, wenn A eine reelle Matrix ist, vgl. Aufgabe 4. Die Bedeutung des Wertebereichs beruht darauf, daß er insbesondere die Eigenwerte der Matrix enthält. Ist nämlich $x \in \mathbb{C}^n \setminus \{0\}$ ein Eigenvektor von A, so ist der entsprechende Rayleigh-Quotient der zugehörige Eigenwert. Weitere wichtige Eigenschaften des Wertebereichs sind in dem folgenden Lemma zusammengefaßt.

Lemma 23.5.

(*a*) *$\mathcal{W}(A)$ ist zusammenhängend.*
(*b*) *Ist $A \in \mathbb{K}^{n\times n}$ hermitesch, dann ist $\mathcal{W}(A)$ das reelle Intervall $[\lambda_n, \lambda_1]$, wobei λ_1 den größten und λ_n den kleinsten Eigenwert von A bezeichnet.*
(*c*) *Ist A* schiefhermitesch*, d. h. $A^* = -A$, dann ist $\mathcal{W}(A)$ ein rein imaginäres Intervall, nämlich die konvexe Hülle aller Eigenwerte von A.*

Beweis. (*a*) Liegen ζ_0 und $\zeta_1 \neq \zeta_0$ im Wertebereich $\mathcal{W}(A)$, dann existieren gewisse $x_0, x_1 \in \mathbb{C}^n \setminus \{0\}$ mit

$$\zeta_0 \,=\, x_0^* A x_0 / x_0^* x_0\,, \qquad \zeta_1 \,=\, x_1^* A x_1 / x_1^* x_1.$$

Wegen $\zeta_1 \neq \zeta_0$ sind x_0 und x_1 offensichtlich linear unabhängig, so daß die Verbindungsgerade

$$[x_0, x_1] \,=\, \{\, x_t = x_0 + t(x_1 - x_0) \,:\, t \in [0,1] \,\}$$

nicht den Nullpunkt enthält. Damit ist die Stetigkeit der Abbildung $t \mapsto x_t^* A x_t / x_t^* x_t$ gewährleistet, so daß

$$\zeta_t = x_t^* A x_t / x_t^* x_t\,, \qquad 0 \le t \le 1\,,$$

eine stetige Kurve in $\mathcal{W}(A)$ bildet, die ζ_0 mit ζ_1 verbindet.

(*b*) Aufgrund der Definition ist $\mathcal{W}(A)$ das Bild der kompakten Einheitssphäre unter einer stetigen Abbildung und somit eine kompakte Menge. Nach Teil (*a*) des Lemmas ist $\mathcal{W}(A)$ zusammenhängend und außerdem reell, wenn A hermitesch ist. Also ist $\mathcal{W}(A)$ ein kompaktes reelles Intervall. Wir wählen nun ein $\alpha > 0$ derart, daß $A + \alpha I$ positiv definit ist. Für dieses α existiert somit die Cholesky-Zerlegung $A + \alpha I = LL^*$, und für $x \in \mathbb{C}^n$ mit $\|x\|_2 = 1$ gilt

$$x^* A x \,=\, x^*(A + \alpha I)x - \alpha \,=\, x^* L L^* x - \alpha \,=\, \|L^* x\|_2^2 - \alpha\,.$$

Aus Satz 2.7 folgt somit

$$\max \mathcal{W}(A) \,=\, \|L^*\|_2^2 - \alpha \,=\, \varrho(LL^*) - \alpha \,=\, \varrho(A + \alpha I) - \alpha \,=\, \lambda_1\,.$$

Durch Übergang zu $\mathcal{W}(-A)$ beweist man entsprechend, daß der linke Endpunkt von $\mathcal{W}(A)$ der kleinste Eigenwert λ_n von A ist.

(*c*) Wegen $A^* = -A$ ist $\mathrm{i}A$ hermitesch, denn

$$(\mathrm{i}A)^* \,=\, \bar{\mathrm{i}} A^* \,=\, -\mathrm{i}A^* \,=\, \mathrm{i}A\,.$$

Ferner ist $\mathcal{W}(\mathrm{i}A) = \mathrm{i}\,\mathcal{W}(A)$ und $\sigma(\mathrm{i}A) = \mathrm{i}\,\sigma(A)$. Also folgt die Behauptung aus Teil (*b*). □

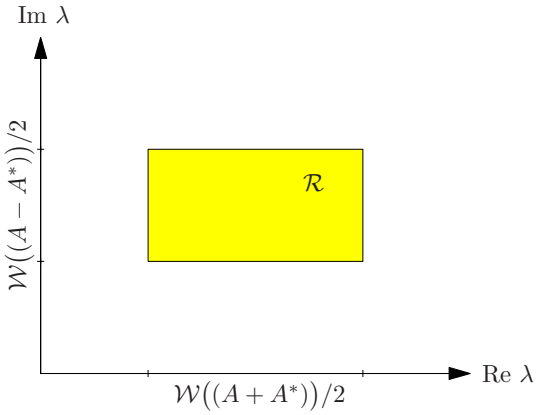


Abb. 23.1: Satz von Bendixson.

Für jede beliebige Matrix $A \in \mathbb{K}^{n\times n}$ ist

$$A \;=\; \frac{A+A^*}{2} + \frac{A-A^*}{2}$$

eine Zerlegung in eine hermitesche Matrix $(A+A^*)/2$ und eine schiefhermitesche Matrix $(A-A^*)/2$. Dies ist die Grundlage des folgenden Einschließungssatzes.

Satz 23.6 (Satz von Bendixson). *Das Spektrum von $A \in \mathbb{K}^{n\times n}$ ist in dem Rechteck*

$$\mathcal{R} \;=\; \mathcal{W}\big(\frac{A+A^*}{2}\big) + \mathcal{W}\big(\frac{A-A^*}{2}\big) \tag{23.5}$$

enthalten.[2]

Beweis. Wir beweisen die stärkere Aussage $\mathcal{W}(A) \subset \mathcal{R}$: Für $x \in \mathbb{C}^n$ mit $\|x\|_2 = 1$ gilt nämlich

$$\begin{aligned} x^*Ax \;&=\; x^*\big(\frac{A+A^*}{2} + \frac{A-A^*}{2}\big)x \;=\; x^*\,\frac{A+A^*}{2}\,x \;+\; x^*\,\frac{A-A^*}{2}\,x \\ &\in\; \mathcal{W}\big(\frac{A+A^*}{2}\big) + \mathcal{W}\big(\frac{A-A^*}{2}\big)\,. \end{aligned}$$

Aus Lemma 23.5 folgt, daß diese Einschlußmenge $\mathcal{R}$ ein Rechteck ist, vgl. Abbildung 23.1. □

[2]Sind $\mathcal{A}, \mathcal{B} \subset \mathbb{K}^n$, dann bezeichnet $\mathcal{A}+\mathcal{B}$ die Menge $\{\, a+b \,:\, a \in \mathcal{A},\ b \in \mathcal{B} \,\}$.

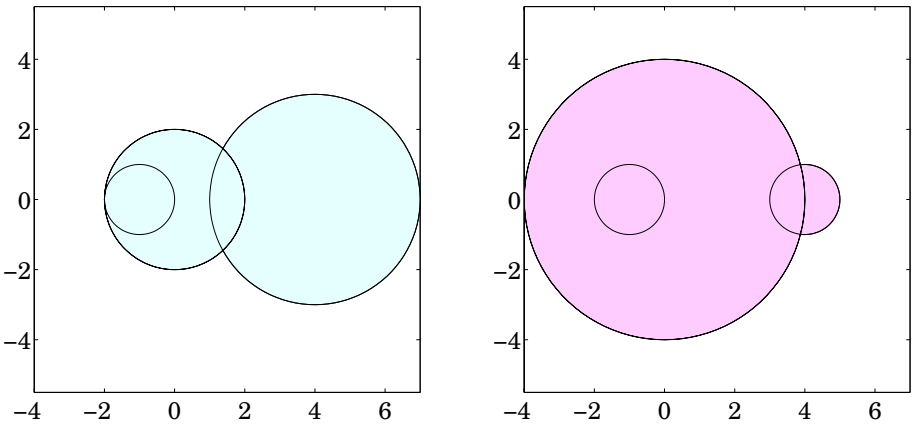


Abb. 23.2: Gerschgorin-Kreise für A (links) und A^* (rechts)

Beispiel. Wir wenden die Resultate aus den Sätzen von Gerschgorin und Bendixson auf die Matrix

$$A = \begin{bmatrix} 4 & 0 & -3 \\ 0 & -1 & 1 \\ -1 & 1 & 0 \end{bmatrix}$$

an. Abbildung 23.2 zeigt die Gerschgorin-Kreise (23.1) für A und A^*, vgl. (23.2).

Für den Satz von Bendixson berechnen wir ferner den symmetrischen und den schiefsymmetrischen Anteil von A,

$$H = \frac{A + A^*}{2} = \begin{bmatrix} 4 & 0 & -2 \\ 0 & -1 & 1 \\ -2 & 1 & 0 \end{bmatrix}, \qquad S = \frac{A - A^*}{2} = \begin{bmatrix} 0 & 0 & -1 \\ 0 & 0 & 0 \\ 1 & 0 & 0 \end{bmatrix}.$$

Die Spektren von H und S können ihrerseits wieder mit Hilfe des Satzes von Gerschgorin eingeschlossen werden: Auf diese Weise erhält man das etwas gröbere Rechteck

$$\tilde{\mathcal{R}} \;=\; [-3, 6] + [-\mathrm{i}, \mathrm{i}] \;\supset\; \mathcal{R} \;\supset \sigma(A)\,.$$

Folglich muß das Spektrum von A im Schnitt *aller* drei Einschlußmengen liegen. Dies ergibt die am dunkelsten eingefärbte Menge in Abbildung 23.3. Tatsächlich ist das Spektrum durch die eingezeichneten Kreuze gegeben:

$$\sigma(A) \;=\; \{\, -1.7878...\,,\; 0.1198...\,,\; 4.6679...\,\}\,.$$

$\Diamond$

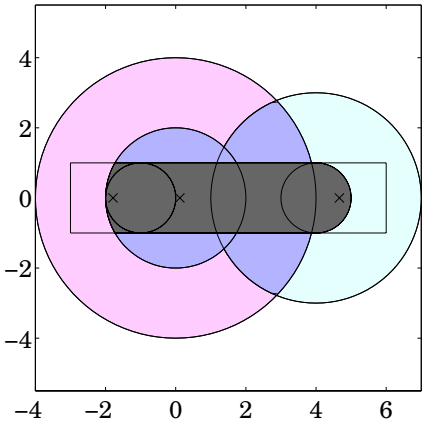


Abb. 23.3:
Alle Einschließungssätze gemeinsam

Für hermitesche Matrizen ist noch das folgende Resultat von großer theoretischer Bedeutung.

Satz 23.7 (Satz von Courant-Fischer). *Sei $A \in \mathbb{K}^{n\times n}$ hermitesch und $\{z_1,\ldots,z_n\} \subset \mathbb{K}^n$ ein orthonormales System. Ferner seien $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_n$ die absteigend sortierten Eigenwerte von A mit orthonormierten Eigenvektoren x_i, $i = 1,\ldots,n$. Dann gilt für alle $k = 1,\ldots,n$*

$$\min_{0\neq x\in\mathcal{Z}_k} \frac{x^*Ax}{x^*x} \leq \lambda_k\,, \qquad \mathcal{Z}_k = \operatorname{span}\{z_1,\ldots,z_k\}\,, \tag{23.6}$$

mit Gleichheit für $\mathcal{Z}_k = \operatorname{span}\{x_1,\ldots,x_k\}$.

Beweis. Für $k = 1$ folgt die Behauptung aus Lemma 23.5 (b). Für $k > 1$ konstruieren wir zunächst einen nichttrivialen Vektor $x = \sum_{i=1}^{k} \zeta_i z_i \in \mathcal{Z}_k$, der senkrecht auf den Eigenvektoren $x_1,\ldots,x_{k-1}$ von A steht. Dann erfüllen die Koeffizienten $\zeta_1,\ldots,\zeta_k$ das homogene lineare Gleichungssystem

$$x_j^*x = \sum_{i=1}^{k} (x_j^*z_i)\,\zeta_i = 0\,, \qquad j = 1,\ldots,k-1\,. \tag{23.7}$$

Da dieses Gleichungssystem unterbestimmt ist ($k-1$ Gleichungen für k Koeffizienten), hat es eine nichttriviale Lösung $[\zeta_1,\ldots,\zeta_k]^T \neq 0$. Wird der zugehörige Vektor $x \neq 0$ in die Eigenbasis von A umentwickelt, $x = \sum_{i=1}^{n} \xi_i x_i$, dann folgt

$$x^*x = \sum_{i,j=1}^{n} \xi_i\overline{\xi_j}\,\underbrace{x_j^*x_i}_{\delta_{ij}} = \sum_{i=1}^{n} |\xi_i|^2$$

und

$$x^*Ax \;=\; x^*\sum_{i=1}^{n}\xi_i Ax_i \;=\; x^*\sum_{i=1}^{n}\xi_i\lambda_i x_i \overset{(23.7)}{=} \sum_{i=k}^{n}\lambda_i\xi_i x^*x_i$$
$$= \sum_{i=k}^{n}\lambda_i\xi_i\sum_{j=1}^{n}\overline{\xi_j}\,\underbrace{x_j^*x_i}_{\delta_{ij}} \;=\; \sum_{i=k}^{n}\lambda_i|\xi_i|^2 \;\le\; \lambda_k\sum_{i=k}^{n}|\xi_i|^2\,.$$

Dies entspricht der Ungleichung $x^*Ax \le \lambda_k x^*x$ und somit ist die Behauptung (23.6) bewiesen.

Für $\mathcal{Z}_k = \operatorname{span}\{x_1,\ldots,x_k\}$ kann jedes $x \in \mathcal{Z}_k$ in der Form $x = \sum_{i=1}^{k}\xi_i x_i$ geschrieben werden, und es folgt entsprechend

$$x^*Ax \;=\; \sum_{i=1}^{k}\lambda_i\xi_i\sum_{j=1}^{k}\overline{\xi_j}\,x_j^*x_i \;=\; \sum_{i=1}^{k}\lambda_i|\xi_i|^2 \;\ge\; \lambda_k\sum_{i=1}^{k}|\xi_i|^2 \;=\; \lambda_k\,x^*x\,.$$

Somit ist

$$\min_{0\neq x\in\mathcal{Z}_k}\frac{x^*Ax}{x^*x} \;\ge\; \lambda_k$$

und demnach gilt in diesem Fall Gleichheit in (23.6). □

Die Aussage von Satz 23.7 wird gelegentlich als *Maxmin-Prinzip* bezeichnet, da der Eigenwert λ_k auf diese Weise als das Maximum der betrachteten Minima charakterisiert wird. Unter denselben Voraussetzungen gilt ein entsprechendes *Minmax-Prinzip*:

$$\max_{0\neq x\perp\mathcal{Z}_k}\frac{x^*Ax}{x^*x} \;\ge\; \lambda_{k+1} \tag{23.8}$$

mit Gleichheit für $\mathcal{Z}_k = \operatorname{span}\{x_1,\ldots,x_k\}$, vgl. Aufgabe 8.

24 Kondition des Eigenwertproblems

Wir untersuchen in diesem Abschnitt die Auswirkung einer Störung der Matrix A auf ihr Spektrum $\sigma(A)$. Wir beginnen hierzu mit der Matrix

$$A = \begin{bmatrix} 0 & \cdots & 0 & -a_0 \\ 1 & \ddots & \vdots & \vdots \\ & \ddots & 0 & -a_{n-2} \\ 0 & & 1 & -a_{n-1} \end{bmatrix} \tag{24.1}$$

und entwickeln die Determinante von $A-\lambda I$ nach der letzten Spalte. Auf diese Weise erhalten wir das charakteristische Polynom von A:

$$p(\lambda) \;=\; (-1)^n \det(A-\lambda I) \;=\; \lambda^n + a_{n-1}\lambda^{n-1} + \ldots + a_1\lambda + a_0\,.$$

Ist umgekehrt p ein beliebig vorgegebenes monisches Polynom vom Grad n mit Koeffizienten $a_0,\ldots,a_{n-1}$, dann nennt man die Matrix A aus (24.1) die *Frobenius-Begleitmatrix* von p.

Das spezielle monische Polynom $p_0(\lambda) = (\lambda-a)^n$ mit $a \neq 0$ hat eine n-fache Nullstelle $\widehat{\lambda} = a$, während $p_\varepsilon(\lambda) = (\lambda-a)^n - \varepsilon$ (mit $\varepsilon > 0$) die Nullstellen

$$\lambda_k = a + \varepsilon^{1/n} e^{\mathrm{i}2\pi k/n}\,, \qquad k = 1,\ldots,n\,,$$

besitzt. Das „gestörte“ Polynom p_ε unterscheidet sich von p_0 lediglich in dem Koeffizienten vor λ^0 (und zwar gerade um ε). Bezeichnen wir die entsprechenden Frobenius-Begleitmatrizen mit A_ε und mit $\|\cdot\|$ die Zeilensummen-, Spaltensummen-, Spektral- oder Frobeniusnorm, so hat die „Störungsmatrix“

$$\triangle A \;=\; A_\varepsilon - A \;=\; \begin{bmatrix} 0 & \cdots & 0 & \varepsilon \\ 0 & & & 0 \\ \vdots & & & \vdots \\ 0 & & \cdots & 0 \end{bmatrix}$$

jeweils die Norm $\|\triangle A\| = \varepsilon$, während die Eigenwerte der beiden Matrizen den Abstand

$$|\triangle\lambda| \;=\; |\lambda_k - \widehat{\lambda}| \;=\; \varepsilon^{1/n}\,, \qquad k = 1,\ldots,n\,, \tag{24.2}$$

haben.

Für $a \neq 0$ ist daher

$$\frac{|\triangle\lambda|}{|\widehat{\lambda}|} \;=\; \frac{\varepsilon^{1/n}}{|a|} \;=\; c_\varepsilon \frac{\|\triangle A\|}{\|A\|} \qquad \text{mit} \qquad c_\varepsilon \;=\; \frac{\|A\|}{|a|}\,\frac{\varepsilon^{1/n}}{\varepsilon}\,.$$

Da c_ε für $\varepsilon \to 0$ beliebig groß wird, kann die relative Konditionszahl des Eigenwertproblems ohne Zusatzvoraussetzungen an die Matrix also beliebig groß sein. Auf der anderen Seite hängen die Eigenwerte stetig von den Einträgen der Matrix ab und der gefundene Exponent $1/n$ in (24.2) ist schlimmstmöglich, vgl. Aufgabe 11. Ferner werden wir gleich in Satz 24.1 sehen, daß für einige wichtige Matrizen das Eigenwertproblem deutlich besser konditioniert ist.

Zunächst erinnern wir jedoch daran, daß eine Matrix $A \in \mathbb{K}^{n\times n}$ *diagonalisierbar* heißt, falls eine Basis $\{x_i\}_{i=1}^n$ des $\mathbb{K}^n$ aus Eigenvektoren existiert. Mit $X = [x_1,\ldots,x_n] \in \mathbb{K}^{n\times n}$ gilt dann

$$A = X\Lambda X^{-1}\,, \tag{24.3}$$

wobei in der Diagonalmatrix Λ die Eigenwerte $\lambda_1, \ldots, \lambda_n$ auf der Diagonalen stehen. Die Matrix X wird *Eigenvektormatrix* genannt. Falls X unitär gewählt werden kann, heißt A *normal*. Normale Matrizen lassen sich auch durch die Gleichung $AA^* = A^*A$ charakterisieren. Insbesondere sind also hermitesche Matrizen ein Spezialfall der normalen Matrizen.

Man beachte, daß eine Matrix im allgemeinen *nicht* diagonalisierbar ist, statt dessen treten *Hauptvektoren* und „Jordankästchen“ in der Jordan-Normalform auf. Ein typisches Beispiel ist die Matrix

$$A = \begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix},$$

deren charakteristisches Polynom $p(\lambda) = \lambda^2$ eine doppelte Nullstelle in $\lambda = 0$ hat. A hat jedoch nur einen eindimensionalen Eigenraum, der durch den Vektor $[1, 0]^T$ aufgespannt wird, während $[0, 1]^T$ ein Hauptvektor zu diesem Eigenvektor ist.

Satz 24.1 (Satz von Bauer und Fike). *$A \in \mathbb{K}^{n\times n}$ sei diagonalisierbar mit entsprechender Faktorisierung $A = X\Lambda X^{-1}$ wie in (24.3). Ferner sei $E \in \mathbb{K}^{n\times n}$ und λ ein Eigenwert von $A + E$. Dann existiert ein Eigenwert $\widehat{\lambda}$ von A mit*

$$|\lambda - \widehat{\lambda}| \leq \operatorname{cond}(X)\,\|E\|\,.$$

Hierbei bezeichnet $\|\cdot\|$ wahlweise die Zeilensummen-, Spaltensummen- oder Spektralnorm und $\operatorname{cond}(\cdot)$ die entsprechende Kondition.

Beweis. Falls λ im Spektrum von A liegt, ist die Behauptung trivial. Andernfalls existiert $(\lambda I - A)^{-1}$, und für einen Eigenvektor $x \neq 0$ von $A + E$ zum Eigenwert λ gilt

$$Ex = (A + E - A)x = (\lambda I - A)x\,,$$

also

$$(\lambda I - A)^{-1}Ex = x\,.$$

Folglich ist

$$\begin{aligned} 1 &\leq \|(\lambda I - A)^{-1}E\| = \|X(\lambda I - \Lambda)^{-1}X^{-1}E\| \\ &\leq \|X\|\,\|X^{-1}\|\,\|E\|\,\|(\lambda I - \Lambda)^{-1}\| \\ &= \operatorname{cond}(X)\,\|E\| \max_{\widehat{\lambda}\in\sigma(A)} |\lambda - \widehat{\lambda}|^{-1}\,. \end{aligned}$$

□

Korollar 24.2. *Ist $A \in \mathbb{K}^{n\times n}$ normal (z. B. hermitesch) und λ ein Eigenwert von $A + E \in \mathbb{K}^{n\times n}$, dann existiert ein $\widehat{\lambda} \in \sigma(A)$ mit*

$$|\lambda - \widehat{\lambda}| \le \|E\|_2 .$$

Die Matrix E braucht hierbei selbst nicht normal zu sein.

Beweis. Im betrachteten Fall ist die Eigenvektormatrix X eine unitäre Matrix. Folglich ist $\|X\|_2 = 1$ und $\|X^{-1}\|_2 = \|X^*\|_2 = 1$. □

Satz 24.1 gibt der Kondition $\mathrm{cond}(X)$ der Eigenvektormatrix X einer diagonalisierbaren Matrix A die Bedeutung einer absoluten Konditionszahl für die Berechnung der Eigenwerte von A, ähnlich wie in Definition 2.9 die Kondition $\mathrm{cond}(A)$ als relative Konditionszahl für die Lösung eines linearen Gleichungssystems mit Koeffizientenmatrix A hergeleitet wurde.

Prinzipiell ist der Satz 24.1 von Bauer und Fike auch anwendbar, um Eigenwerte von A einzuschließen. Allerdings kann in der Praxis nicht immer eine Näherungsmatrix $A+E$ angegeben werden, deren Eigenwerte explizit bekannt sind. Häufig läßt sich hingegen ein Vektor bestimmen, der „beinahe“ ein Eigenvektor ist, was folgendermaßen zu einer A-posteriori-Abschätzung einer Eigenwertnäherung herangezogen werden kann.

Satz 24.3. *$A \in \mathbb{K}^{n\times n}$ sei eine hermitesche Matrix, $\lambda \in \mathbb{R}$ und $x \in \mathbb{K}^n$ ein Vektor mit*

$$\|Ax - \lambda x\|_2 = \varepsilon \|x\|_2 , \qquad \varepsilon > 0 .$$

Dann besitzt A einen Eigenwert $\widehat{\lambda}$ mit

$$|\lambda - \widehat{\lambda}| = \min_{\mu\in\sigma(A)} |\lambda - \mu| \le \varepsilon .$$

Ist ferner $\widehat{x}$ ein Eigenvektor zum Eigenwert $\widehat{\lambda}$ und γ der Abstand von λ zu dem nächstgelegenen von $\widehat{\lambda}$ verschiedenen Eigenwert von A, dann gilt für den eingeschlossenen Winkel θ zwischen x und $\widehat{x}$ die Abschätzung

$$|\sin\theta| \le \varepsilon/\gamma .$$

Beweis. Falls λ im Spektrum von A liegt, kann $\widehat{\lambda} = \lambda$ gewählt werden. Ansonsten ist $A - \lambda I$ invertierbar und ebenfalls hermitesch, und es ist

$$\max_{\mu\in\sigma(A)} \frac{1}{|\lambda - \mu|} = \|(A - \lambda I)^{-1}\|_2 = \sup_{z\neq 0} \frac{\|(A - \lambda I)^{-1} z\|_2}{\|z\|_2} .$$

Speziell für $z = Ax - \lambda x$ ergibt sich somit nach Voraussetzung die untere Schranke

$$\max_{\mu\in\sigma(A)} \frac{1}{|\lambda - \mu|} \ge \frac{\|(A - \lambda I)^{-1} z\|_2}{\|z\|_2} = \frac{\|x\|_2}{\|Ax - \lambda x\|_2} = 1/\varepsilon .$$

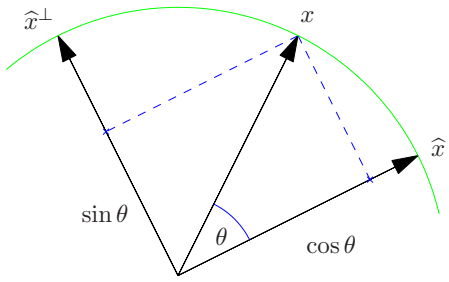


Abb. 24.1:
Beweisskizze

Also existiert ein $\widehat{\lambda} \in \sigma(A)$ mit

$$1/|\lambda - \widehat{\lambda}| \;\geq\; 1/\varepsilon\,,$$

und dies ist gerade die erste Behauptung.

Für die Abschätzung des Eigenvektors können wir ohne Einschränkung annehmen, daß x selbst kein Eigenvektor zum Eigenwert $\widehat{\lambda}$ von A ist. Dann wählen wir einen Eigenvektor $\widehat{x}$ zum Eigenwert $\widehat{\lambda}$ mit $x^*\widehat{x} \geq 0$ und $\|\widehat{x}\|_2 = \|x\|_2$, und erhalten für x eine Darstellung

$$x \;=\; c\widehat{x} \;+\; s\widehat{x}^\perp \qquad \text{mit} \quad c = \cos\theta\,,\; s = \sin\theta\,,\; 0 \leq \theta \leq \pi/2\,,$$

und einem Vektor $\widehat{x}^\perp$ mit $\widehat{x}^*\widehat{x}^\perp = 0$ und $\|\widehat{x}^\perp\|_2 = \|\widehat{x}\|_2$, vgl. Abbildung 24.1. Daraus folgt

$$(A - \lambda I)x \;=\; c(\widehat{\lambda} - \lambda)\widehat{x} \;+\; s(A - \lambda I)\widehat{x}^\perp\,,$$

und wegen $\widehat{x}^*(A - \lambda I) = (\widehat{\lambda} - \lambda)\widehat{x}^*$ ergibt sich

$$\begin{aligned}
&\|Ax - \lambda x\|_2^2 \\
&= c^2|\lambda - \widehat{\lambda}|^2\,\|\widehat{x}\|_2^2 + 2sc(\widehat{\lambda} - \lambda)\mathrm{Re}\,\widehat{x}^*(A - \lambda I)\widehat{x}^\perp + s^2\,\|(A - \lambda I)\widehat{x}^\perp\|_2^2 \\
&= c^2|\lambda - \widehat{\lambda}|^2\,\|\widehat{x}\|_2^2 + 2sc(\widehat{\lambda} - \lambda)^2\mathrm{Re}\,\widehat{x}^*\widehat{x}^\perp + s^2\,\|(A - \lambda I)\widehat{x}^\perp\|_2^2 \\
&= c^2|\lambda - \widehat{\lambda}|^2\,\|\widehat{x}\|_2^2 + s^2\,\|(A - \lambda I)\widehat{x}^\perp\|_2^2\,.
\end{aligned}$$

Aufgrund der Voraussetzung erhalten wir hiermit die Ungleichung

$$\varepsilon^2\,\|x\|_2^2 \;\geq\; s^2\,\|(A - \lambda I)\widehat{x}^\perp\|_2^2\,,$$

und mit $z = (A - \lambda I)\widehat{x}^{\perp}$ folgt hieraus wegen $\|\widehat{x}^{\perp}\|_2 = \|x\|_2$

$$s^2 \;\leq\; \varepsilon^2 \,\frac{\widehat{x}^{\perp *}\widehat{x}^{\perp}}{\widehat{x}^{\perp *}(A-\lambda I)^2\widehat{x}^{\perp}} \;=\; \varepsilon^2\,\frac{z^*(A-\lambda I)^{-2}z}{z^*z}\,.$$

Ferner ist $z^*\widehat{x} = \widehat{x}^{\perp *}(A-\lambda I)\widehat{x} = (\widehat{\lambda}-\lambda)\widehat{x}^{\perp *}\widehat{x} = 0$, d. h. z steht senkrecht auf dem Eigenvektor $\widehat{x}$ des dominanten Eigenwerts $(\widehat{\lambda}-\lambda)^{-2}$ von $(A-\lambda I)^{-2}$. Aus dem Minmaxprinzip (23.8) von Courant-Fischer ergibt sich daher die obere Schranke

$$s^2 \;\leq\; \varepsilon^2 \max_{0\neq z\perp\widehat{x}} \frac{z^*(A-\lambda I)^{-2}z}{z^*z} \;=\; \varepsilon^2 \min_{\substack{\mu\in\sigma(A)\\ \mu\neq\widehat{\lambda}}} |\mu-\lambda|^{-2} \;=\; (\varepsilon/\gamma)^2\,.$$

Damit ist auch die zweite Behauptung bewiesen. □

Eigenvektoren reagieren also unter Umständen wesentlich sensibler als Eigenwerte auf Störungen in der Matrix – zumindest bei hermiteschen Matrizen. Je näher zwei Eigenwerte einer Matrix beieinander liegen, um so schwieriger ist es, den Eigenvektor stabil zu approximieren.

Zum Abschluß dieses Abschnitts noch ein Analogon zu Korollar 24.2 für die Frobeniusnorm.

Satz 24.4 (Satz von Wielandt-Hoffman). *Die $n \times n$ Matrizen A und E seien hermitesch, $\widehat{\lambda}_1 \geq \ldots \geq \widehat{\lambda}_n$ und $\lambda_1 \geq \ldots \geq \lambda_n$ seien die Eigenwerte von A bzw. $A+E$. Dann ist*

$$\sum_{i=1}^{n}(\lambda_i - \widehat{\lambda}_i)^2 \;\leq\; \|E\|_F^2\,. \tag{24.4}$$

Beweis. Wir beweisen hier lediglich ein etwas schwächeres Resultat, nämlich die Existenz von n Eigenwerten $\widehat{\lambda}_i$ von A, die die Ungleichung (24.4) erfüllen; mit diesem Beweis können wir aber nicht garantieren, daß jeder Eigenwert von A genau einmal in der Summe auftritt.

1. Von unabhängigem Interesse ist zunächst das folgende Hilfsresultat: Ist $Q \in \mathbb{K}^{n\times n}$ unitär und $A \in \mathbb{K}^{n\times n}$, dann ist

$$\|QA\|_F = \|AQ\|_F = \|A\|_F\,. \tag{24.5}$$

Wegen $\|AQ\|_F = \|Q^*A^*\|_F$ ist es hinreichend, die erste Gleichung in (24.5) nachzuweisen. Dazu seien die Spalten von A mit $a_1, \ldots, a_n$ bezeichnet. Dann ist $QA = \big[\,Qa_1, \ldots, Qa_n\,\big]$ und die Teilbehauptung ergibt sich wie folgt:

$$\|QA\|_F^2 \;=\; \sum_{i=1}^{n}\|Qa_i\|_2^2 \overset{(13.1)}{=} \sum_{i=1}^{n}\|a_i\|_2^2 \;=\; \|A\|_F^2\,.$$

2. Im zweiten Beweisschritt betrachten wir die Faktorisierung $A+E = Q\Lambda Q^*$ von $A+E$ in die Diagonalmatrix Λ mit den Eigenwerten λ_i, $i = 1, \ldots, n$, und eine zugehörige (unitäre) Eigenvektormatrix Q. Aufgrund von (24.5) gilt dann

$$\|E\|_F^2 \;=\; \|A+E-A\|_F^2 \;=\; \|\Lambda - Q^*AQ\|_F^2 \;=\; \sum_{i=1}^{n} \|Q^*AQe_i - \lambda_i e_i\|_2^2$$

und die rechte Seite läßt sich nach Satz 24.3 durch

$$\sum_{i=1}^{n} \|Q^*AQe_i - \lambda_i e_i\|_2^2 \;=\; \sum_{i=1}^{n} \frac{\|Q^*AQe_i - \lambda_i e_i\|_2^2}{\|e_i\|_2^2} \;\geq\; \sum_{i=1}^{n} |\lambda_i - \widehat{\lambda}_{j_i}|^2$$

für geeignete $\widehat{\lambda}_{j_i} \in \sigma(Q^*AQ)$ abschätzen. Da Q^*AQ und A zueinander ähnlich sind, folgt hieraus das Resultat

$$\sum_{i=1}^{n} |\lambda_i - \widehat{\lambda}_{j_i}|^2 \;\leq\; \|E\|_F^2 \qquad \text{für gewisse } \widehat{\lambda}_{j_i} \in \sigma(A)\,.$$

Der Beweis des vollständigen Satzes (vgl. etwa das Buch von Wilkinson [108, S. 108ff.]) verwendet andere Techniken und ist wesentlich komplizierter. □

25 Die Potenzmethode

Als erstes konstruktives Verfahren zur Berechnung einzelner Eigenwerte und Eigenvektoren betrachten wir die *Potenzmethode nach von Mises*.

Um die Grundidee dieses Verfahrens herauszustellen, gehen wir zunächst von einer reellen diagonalisierbaren $n \times n$ Matrix A mit n betragsmäßig verschiedenen (und daher reellen) Eigenwerten λ_i aus, die wie folgt angeordnet seien:

$$|\lambda_1| > |\lambda_2| > \ldots > |\lambda_n| \geq 0\,.$$

Ist $\|\cdot\|$ eine Vektornorm und sind v_i, $i = 1, \ldots, n$, mit $\|v_i\| = 1$ jeweils zu λ_i gehörende Eigenvektoren von A, dann kann jeder Vektor $x \in \mathbb{K}^n$ in diese Eigenvektorbasis entwickelt werden,

$$x = \sum_{i=1}^{n} \xi_i v_i\,. \tag{25.1}$$

Hieraus erhalten wir

$$A^k x = \sum_{i=1}^{n} \lambda_i^k \xi_i v_i \tag{25.2}$$

und stellen fest, daß sich für große k auf der rechten Seite von (25.2) der Summand mit dem dominanten Eigenwert durchsetzt, falls ξ_1 von Null verschieden ist, d. h.

$$A^k x \approx \lambda_1^k \xi_1 v_1 \,.$$

Aus den Vektoren $A^k x$ kann also prinzipiell auf den betragsgrößten Eigenwert und einen zugehörigen Eigenvektor geschlossen werden. Dies ist die Grundidee des von Mises-Verfahrens. In der Praxis wird nach jeder Multiplikation mit A der Vektor neu normiert, um möglichen Overflow bzw. Underflow zu vermeiden: Ausgehend von einem Startvektor $z^{(0)}$ mit $\|z^{(0)}\| = 1$ berechnet man die Iterationsfolge

$$\tilde{z}^{(k)} \;=\; A z^{(k-1)} \quad \text{und} \quad z^{(k)} = \tilde{z}^{(k)} / \|\tilde{z}^{(k)}\|, \qquad k = 1, 2, \ldots \,, \tag{25.3}$$

und durch vollständige Induktion sieht man unmittelbar, daß

$$z^{(k)} \;=\; A^k z^{(0)} / \|A^k z^{(0)}\| \,, \qquad k = 0, 1, \ldots \tag{25.4}$$

Wir untersuchen nun dieses Verfahren unter etwas allgemeineren Voraussetzungen.

Satz 25.1. *Die (nicht notwendigerweise diagonalisierbare) Matrix $A \in \mathbb{K}^{n\times n}$ habe die Eigenwerte*

$$|\lambda_1| > |\lambda_2| \geq \ldots \geq |\lambda_n| \,,$$

ferner sei y ein linker Eigenvektor zum Eigenwert λ_1 mit $\|y\| = 1$. Ist $z^{(0)} \in \mathbb{K}^n$ ein Startvektor mit $\|z^{(0)}\| = 1$ und $y^* z^{(0)} \;/=, 0$ dann gilt

$$\limsup_{k\to\infty} \big|\, \|\tilde{z}^{(k)}\| - |\lambda_1| \,\big|^{1/k} \;\leq\; |\lambda_2/\lambda_1| \,, \tag{25.5}$$

und es existiert ein Eigenvektor v von A zum Eigenwert λ_1 mit $\|v\| = 1$ und

$$\limsup_{k\to\infty} \|z^{(k)} - \operatorname{sign}(\lambda_1^k)\, v\|^{1/k} \;\leq\; |\lambda_2/\lambda_1| \,.$$

Dabei ist die Vorzeichenfunktion sign *für alle $\lambda \in \mathbb{C} \setminus \{0\}$ durch*

$$\operatorname{sign}(\lambda) \;=\; \lambda/|\lambda|$$

definiert.

Beweis. Da λ_1 unter den getroffenen Annahmen ein einfacher Eigenwert von A ist, existiert genau ein zugehöriger Eigenvektor v mit $\|v\| = 1$ und $\operatorname{sign}(y^*v) = \operatorname{sign}(y^*z^{(0)})$, mit dem wir die Matrix

$$E \;=\; A - \lambda_1 \frac{vy^*}{y^*v} \tag{25.6}$$

definieren. Es gilt dann

$$E^*y \;=\; A^*y - \overline{\lambda}_1 \frac{yv^*}{v^*y}\, y \;=\; \overline{\lambda}_1 y - \overline{\lambda}_1 \frac{v^*y}{v^*y}\, y \;=\; 0$$

und nach Lemma 11.2 gehört y zu $\mathcal{N}(E^*) = \mathcal{R}(E)^\perp$. Für jeden Eigenvektor $z \neq 0$ von E mit Eigenwert $\mu \neq 0$ folgt demnach

$$y^*z \;=\; \frac{1}{\mu}\, y^*Ez \;=\; 0\,. \tag{25.7}$$

Aus (25.6) und (25.7) erhalten wir daher

$$Az \;=\; Ez + \lambda_1 \frac{y^*z}{y^*v}\, v \;=\; \mu z\,,$$

d. h. μ ist auch ein Eigenwert von A. Dieser Eigenwert muß allerdings von λ_1 verschieden sein, da y und v nicht zueinander orthogonal sind. Somit gilt $\sigma(E) \setminus \{0\} \subset \sigma(A) \setminus \{\lambda_1\}$ beziehungsweise

$$\varrho(E) \;\leq\; |\lambda_2|\,. \tag{25.8}$$

Wir zeigen als nächstes, daß

$$A^k z^{(0)} \;=\; \lambda_1^k \xi v + E^k z^{(0)}\,, \qquad \text{wobei} \;\; \xi = y^*z^{(0)}/y^*v > 0\,. \tag{25.9}$$

Offensichtlich gilt (25.9) für $k = 1$, denn aus (25.6) folgt

$$Az^{(0)} \;=\; \lambda_1 \xi v + Ez^{(0)}\,.$$

Für $k \geq 1$ schließt man nun induktiv aus (25.9) und (25.6), daß

$$\begin{aligned} A^{k+1}z^{(0)} \;&=\; A(A^k z^{(0)}) \;=\; \lambda_1^k \xi Av + AE^k z^{(0)} \\ &=\; \lambda_1^{k+1}\xi v + E^{k+1}z^{(0)} + \lambda_1 \frac{y^*E^k z^{(0)}}{y^*v}\, v\,. \end{aligned}$$

Wegen $y^*E^k z^{(0)} = (E^*y)^*E^{k-1}z^{(0)} = 0$ folgt hieraus (25.9) für $k{+}1$. Im weiteren schreiben wir (25.9) in der Form

$$A^k z^{(0)} \;=\; \lambda_1^k \xi (v + w^{(k)}) \qquad \text{mit} \qquad w^{(k)} \;=\; \frac{E^k z^{(0)}}{\lambda_1^k \xi}\,. \tag{25.10}$$

Nach Aufgabe I.10 können wir nun für ein gegebenes $\varepsilon > 0$ eine Norm $\|\cdot\|_\varepsilon$ in $\mathbb{K}^n$ und eine dadurch induzierte Norm $|\!|\!|\cdot|\!|\!|_\varepsilon$ in $\mathbb{K}^{n\times n}$ wählen mit $|\!|\!|E|\!|\!|_\varepsilon \le \varrho(E) + \varepsilon$. Mit dieser Norm gilt wegen (25.8)

$$\|w^{(k)}\|_\varepsilon \le \frac{|\!|\!|E|\!|\!|_\varepsilon^k}{|\lambda_1|^k} \frac{\|z^{(0)}\|_\varepsilon}{\xi} \le \Big(\frac{|\lambda_2| + \varepsilon}{|\lambda_1|}\Big)^k \frac{\|z^{(0)}\|_\varepsilon}{\xi}\,.$$

Andererseits ist diese Norm $\|\cdot\|_\varepsilon$ äquivalent zu der vorgegebenen Norm $\|\cdot\|$, d. h. es existieren positive Konstanten c_ε und C_ε mit

$$c_\varepsilon \|x\|_\varepsilon \le \|x\| \le C_\varepsilon \|x\|_\varepsilon \qquad \text{für alle } x \in \mathbb{K}^n\,.$$

Folglich ist

$$\|w^{(k)}\| \le C_\varepsilon \|w^{(k)}\|_\varepsilon \le \frac{C_\varepsilon \|z^{(0)}\|_\varepsilon}{\xi} \Big(\frac{|\lambda_2| + \varepsilon}{|\lambda_1|}\Big)^k \le \frac{C_\varepsilon}{c_\varepsilon \xi}\, q_\varepsilon^k \tag{25.11}$$

mit

$$q_\varepsilon = \frac{|\lambda_2| + \varepsilon}{|\lambda_1|}\,. \tag{25.12}$$

Aus (25.4) und (25.10) erhalten wir somit

$$z^{(k)} = \operatorname{sign}(\lambda_1^k \xi)\, \frac{v + w^{(k)}}{\|v + w^{(k)}\|} = \operatorname{sign}(\lambda_1^k)\, v + e^{(k)} \tag{25.13}$$

mit

$$e^{(k)} = \frac{\operatorname{sign}(\lambda_1^k)}{\|v + w^{(k)}\|}\big(w^{(k)} + (1 - \|v + w^{(k)}\|)\, v\big)\,, \tag{25.14}$$

da ξ nach (25.9) positiv ist. Wegen $\|v\| - \|w^{(k)}\| \le \|v + w^{(k)}\| \le \|v\| + \|w^{(k)}\|$ und $\|v\| = 1$ folgt aus (25.11) die Abschätzung

$$\big|\, 1 - \|v + w^{(k)}\| \,\big| \le \|w^{(k)}\| = O(q_\varepsilon^k)\,, \qquad k \to \infty\,. \tag{25.15}$$

Demnach gilt $\|e^{(k)}\| = O(q_\varepsilon^k)$ für $k \to \infty$, und $z^{(k)}$ und $\tilde{z}^{(k+1)}$ verhalten sich für $k \to \infty$ wie

$$z^{(k)} = \operatorname{sign}(\lambda_1^k)\, v + O(q_\varepsilon^k) \quad \text{und} \quad \tilde{z}^{(k+1)} = \lambda_1 \operatorname{sign}(\lambda_1^k)\, v + O(q_\varepsilon^k)\,.$$

Hieraus folgen unmittelbar die Behauptungen mit der oberen Schranke q_ε aus (25.12) anstelle von $|\lambda_2/\lambda_1|$. Da ε jedoch beliebig klein gewählt werden kann, gelten die Behauptungen auch wie in Satz 25.1 formuliert. □

Initialisierung: Startvektor $z^{(0)}$ mit $\|z^{(0)}\|_2 = 1$ sei gegeben

```
for k = 1, 2, ... do
```
$\tilde{z}^{(k)} = Az^{(k-1)}$
$z^{(k)} = \tilde{z}^{(k)} / \|\tilde{z}^{(k)}\|_2$
$\mu^{(k)} = z^{(k-1)*} \tilde{z}^{(k)}$
```
until stop      % end for
```
Ergebnis: $\mu^{(k)}$ ist eine Näherung des betragsgrößten Eigenwerts von A und $z^{(k)}$ approximiert einen zugehörigen Eigenvektor

Algorithmus 25.1: Von Mises-Potenzmethode (bzgl. der Euklidnorm)

Bei der von Mises-Iteration konvergiert also $z^{(k)}$ gegen eine Eigenrichtung von A und die Norm von $\tilde{z}^{(k)}$ gegen den Spektralradius von A. Die Konvergenz ist jeweils linear und $q = |\lambda_2/\lambda_1|$ ist der asymptotische Konvergenzfaktor. Bei einer Implementierung der Potenzmethode bestimmt man zunächst $|\lambda_1|$ aus (25.5). Anhand des Vorzeichenverhaltens von $z^{(k)}$ kann man dann auf das Vorzeichen von λ_1 schließen. Im Reellen gilt beispielsweise: Alternieren die Vorzeichen von $z^{(k)}$, dann ist $\lambda_1 < 0$, ansonsten ist $\lambda_1 > 0$.

Bemerkung 25.2. Da die Eigenvektoren nicht bekannt sind, kann die Voraussetzung $y^* z^{(0)} \neq 0$ aus Satz 25.1 nicht a priori überprüft werden. Diese Voraussetzung erweist sich jedoch in der Praxis als wenig problematisch, da in der Regel im Verlauf der Iteration aufgrund von Rundungsfehlern eine Komponente von $z^{(k)}$ längs v „eingeschleppt“ wird. Die Asymptotik (25.5) ist in diesem Fall allerdings erst für größere k feststellbar. Ähnlich ist die Situation, wenn in (25.1) $|\xi_1| \ll \|x\|$ gilt. ◊

Verwendet man die Euklidnorm, so kann der Eigenwert auch wie in Algorithmus 25.1 durch den Rayleigh-Quotienten von $z^{(k-1)}$ approximiert werden,

$$\lambda_1 \approx \mu^{(k)} = z^{(k-1)*} A z^{(k-1)} = z^{(k-1)*} \tilde{z}^{(k)} . \tag{25.16}$$

Diese Eigenwertnäherung konvergiert für hermitesche Matrizen schneller.

Korollar 25.3. *Ist A neben den Voraussetzungen aus Satz* 25.1 *auch noch hermitesch und ist $\|\cdot\| = \|\cdot\|_2$, dann gilt*

$$|\lambda_1 - \mu^{(k)}| = O(|\lambda_2/\lambda_1|^{2k}), \qquad k \to \infty .$$

Beweis. Für hermitesche Matrizen stimmen die beiden Eigenvektoren y und v aus dem Beweis von Satz 25.1 bis auf das Vorzeichen überein. Nach (25.10) gehört $w^{(k)}$ zu $\mathcal{R}(E)$, und wegen $v = y \in \mathcal{R}(E)^\perp$ sind $v + w^{(k)}$ in (25.13) sowie die Darstellung (25.14) von $e^{(k)}$ orthogonale Summen. Aus dem Satz von

Pythagoras folgt somit

$$\|v + w^{(k)}\|_2 \geq \|v\|_2 = 1$$

und, zusammen mit (25.15),

$$\|e^{(k)}\|_2^2 \leq \|w^{(k)}\|_2^2 + \left(1 - \|v + w^{(k)}\|_2\right)^2 \leq 2\,\|w^{(k)}\|_2^2 . \tag{25.17}$$

Aus der zweiten Zerlegung von $z^{(k)}$ in (25.13) erhalten wir ferner

$$(A - \lambda_1 I)z^{(k)} = (A - \lambda_1 I)e^{(k)} ,$$

und hieraus folgt die Fehlerdarstellung

$$\begin{aligned} \lambda_1 - \mu^{(k+1)} &= \lambda_1 z^{(k)*} z^{(k)} - z^{(k)*} A z^{(k)} = z^{(k)*}(\lambda_1 I - A)z^{(k)} \\ &= z^{(k)*}(\lambda_1 I - A)e^{(k)} . \end{aligned}$$

Da A hermitesch ist, können wir dies weiter umformen,

$$\lambda_1 - \mu^{(k+1)} = z^{(k)*}(\lambda_1 I - A)e^{(k)} = e^{(k)*}(\lambda_1 I - A)z^{(k)} = e^{(k)*}(\lambda_1 I - A)e^{(k)} ,$$

und zusammen mit (25.17) folgt dann unmittelbar die Fehlerabschätzung

$$|\lambda_1 - \mu^{(k+1)}| \leq \left(|\lambda_1| + \|A\|_2\right)\|e^{(k)}\|_2^2 \leq 4\|A\|_2\|w^{(k)}\|_2^2 . \tag{25.18}$$

Mit A ist auch die Matrix E hermitesch. Daher ist $\|E\|_2 = \varrho(E)$ und aus (25.10) und (25.8) folgt in diesem Fall

$$\|w^{(k)}\|_2 \leq \frac{1}{|\xi|}\,|\lambda_2/\lambda_1|^k .$$

Eingesetzt in (25.18) erhalten wir also

$$|\lambda_1 - \mu^{(k+1)}| \leq \frac{4\|A\|_2}{|\xi|^2}\,|\lambda_2/\lambda_1|^{2k} ,$$

was zu zeigen war. □

Die Potenzmethode kann in der bislang betrachteten Form nur verwendet werden, um den betragsgrößten Eigenwert und einen zugehörigen Eigenvektor zu bestimmen. Zur Berechnung anderer Eigenwerte und Eigenvektoren kann die Matrix jedoch geeignet transformiert werden:

(*a*) Sofern A invertierbar ist, ersetzt man bei der *inversen Iteration* A durch A^{-1} in Algorithmus 25.1, um den betragskleinsten Eigenwert λ_n und einen zugehörigen Eigenvektor zu approximieren. Da A^{-1} die Eigenwerte $\lambda_i^{-1}, i = 1, \ldots, n$, mit den gleichen Eigenvektoren wie A besitzt, konvergiert die inverse Iteration gegen den dominanten Eigenwert λ_n^{-1} von A^{-1}.

Initialisierung: $(\mu_0, z^{(0)})$ mit $\|z^{(0)}\|_2 = 1$ sei Approximation eines Eigenpaars von A

`for` $k = 1, 2, \ldots$ `do`
$(A - \mu_{k-1} I)\tilde{z}^{(k)} = z^{(k-1)}$
$z^{(k)} = \tilde{z}^{(k)} / \|\tilde{z}^{(k)}\|_2$
$\mu_k = z^{(k)*} A z^{(k)}$
`until` stop % end for

Ergebnis: $z^{(k)}$ approximiert einen Eigenvektor von A und μ_k den zugehörigen Eigenwert

Algorithmus 25.2: Rayleigh-Quotienten-Iteration

(*b*) Ist λ eine Näherung an einen Eigenwert von A, liegt aber selbst nicht im Spektrum $\sigma(A)$, dann ergibt Algorithmus 25.1 mit $(A - \lambda I)^{-1}$ anstelle von A die *gebrochene Iteration von Wielandt*. $(A - \lambda I)^{-1}$ besitzt die Eigenwerte $(\lambda_i - \lambda)^{-1}$, $i = 1, \ldots, n$, und die gebrochene Iteration approximiert daher einen Eigenvektor zu dem Eigenwert λ_i von A, der am nächsten an λ liegt.

Die gebrochene Iteration wird in der Praxis vorrangig eingesetzt, um die Eigenvektoren zu bereits berechneten Eigenwerten zu bestimmen, vgl. auch Abschnitt 29.2 und insbesondere Beispiel 29.5. Dabei muß in jedem Iterationsschritt ein lineares Gleichungssystem mit derselben Matrix $A - \lambda I$ gelöst werden. Man verwendet daher in der Implementierung üblicherweise die LR-Zerlegung oder die QR-Zerlegung von $A - \lambda I$. Entsprechendes gilt für die inverse Iteration.

Ist λ_i der gesuchte Eigenwert, dann konvergiert die gebrochene Iteration um so schneller, je näher λ an λ_i liegt, da dann der Konvergenzfaktor

$$q = \max_{j \neq i} \frac{|\lambda_j - \lambda|^{-1}}{|\lambda_i - \lambda|^{-1}} = \max_{j \neq i} \frac{|\lambda_i - \lambda|}{|\lambda_j - \lambda|}$$

entsprechend klein wird. Das Verfahren kann daher beschleunigt werden, indem der Schätzwert λ während der Iteration ständig verbessert wird. Dies führt uns auf die *Rayleigh-Quotienten-Iteration* (Algorithmus 25.2), bei der in jedem Iterationsschritt eine neue Näherung $\lambda = \mu_k$ aus dem aktuellen Rayleigh-Quotienten bestimmt wird.

Aufgrund dieser Konstruktion mag man vermuten, daß die Rayleigh-Quotienten-Iteration superlinear konvergiert. Im allgemeinen ist dies auch tatsächlich der Fall, für $A = A^*$ ist die Konvergenz sogar *lokal kubisch*.

Satz 25.4. *Es sei $A = A^* \in \mathbb{K}^{n \times n}$ und die Folge $\{z^{(k)}\}$ aus Algorithmus* 25.2 *konvergiere gegen einen Eigenvektor $\widehat{x}$ von A. Dann konvergieren die Näherungen μ_k aus Algorithmus* 25.2 *lokal kubisch gegen den zugehörigen Eigenwert $\widehat{\lambda}$.*

Beweisskizze. Wir verzichten hier auf den vollständigen Beweis, der einige subtile technische Fallunterscheidungen benötigt (dazu verweisen wir auf das Buch von Parlett [81]). Statt dessen skizzieren wir lediglich die Beweisidee für den wichtigsten Fall, in dem die Iterierten asymptotisch durch eine Linearkombination zweier Eigenvektoren darstellbar sind.

Sei (λ, x) das Eigenpaar mit $\lambda \neq \widehat{\lambda}$, dessen Eigenwert λ am nächsten an $\widehat{\lambda}$ liegt, ferner sei $\|x\|_2 = \|\widehat{x}\|_2 = 1$. Wir wollen für diese Beweisskizze annehmen, daß sich $z^{(k)}$ darstellen läßt als

$$\begin{aligned} & z^{(k)} = \xi_k \widehat{x} + \zeta_k x + y^{(k)} \\ & \text{mit } \; y^{(k)} \perp \operatorname{span}\{\widehat{x}, x\} \; \text{ und } \; \|y^{(k)}\|_2 = o(|\zeta_k|) \text{ für } k \to \infty \,. \end{aligned} \tag{25.19}$$

Wegen der angenommenen Konvergenz der $z^{(k)}$ gilt

$$\xi_k \to 1 \,, \; \zeta_k \to 0 \,, \qquad k \to \infty \,.$$

Daher ergibt sich aus der Orthogonalität der Eigenvektoren von A und aus (25.19)

$$\begin{aligned} |\widehat{\lambda} - \mu_k| &= |\widehat{\lambda} - z^{(k)*} A z^{(k)}| = |z^{(k)*}(\widehat{\lambda} I - A) z^{(k)}| \\ &= \big| \, |\xi_k|^2 \, \widehat{x}^*(\widehat{\lambda} I - A)\widehat{x} + |\zeta_k|^2 \, x^*(\widehat{\lambda} I - A)x + o(|\zeta_k|^2) \, \big| \\ &= |\zeta_k|^2 \, |\widehat{\lambda} - \lambda| + o(|\zeta_k|^2) \,, \qquad k \to \infty \,. \end{aligned} \tag{25.20}$$

Aufgrund der Iterationsvorschrift und wegen (25.19) ist

$$z^{(k)} = \frac{\tilde{z}^{(k)}}{\|\tilde{z}^{(k)}\|_2} \qquad \text{mit} \qquad \tilde{z}^{(k)} = \frac{\xi_{k-1}}{\widehat{\lambda} - \mu_{k-1}} \, \widehat{x} + \frac{\zeta_{k-1}}{\lambda - \mu_{k-1}} \, x + \tilde{y}^{(k)}$$

mit einem $\tilde{y}^{(k)} \perp \operatorname{span}\{\widehat{x}, x\}$, so daß aus dem Satz von Pythagoras $\|\tilde{z}^{(k)}\|_2 \geq |\xi_{k-1}/(\widehat{\lambda} - \mu_{k-1})|$ folgt. Außerdem gilt

$$|\zeta_k| = \frac{1}{\|\tilde{z}^{(k)}\|_2} \left| \frac{\zeta_{k-1}}{\lambda - \mu_{k-1}} \right| \leq \left| \frac{\zeta_{k-1}(\widehat{\lambda} - \mu_{k-1})}{\xi_{k-1}(\lambda - \mu_{k-1})} \right| .$$

Eingesetzt in (25.20) ergibt dies

$$\begin{aligned} |\widehat{\lambda} - \mu_k| &= |\zeta_k|^2 \, |\widehat{\lambda} - \lambda| \big(1 + o(1)\big) \\ &\leq |\zeta_{k-1}|^2 \, |\widehat{\lambda} - \lambda| \, \frac{|\widehat{\lambda} - \mu_{k-1}|^2}{|\xi_{k-1}|^2 \, |\lambda - \mu_{k-1}|^2} \big(1 + o(1)\big) \\ &\overset{(25.20)}{=} |\widehat{\lambda} - \mu_{k-1}| \, \frac{|\widehat{\lambda} - \mu_{k-1}|^2}{|\xi_{k-1}|^2 |\lambda - \mu_{k-1}|^2} \big(1 + o(1)\big) \,, \end{aligned}$$

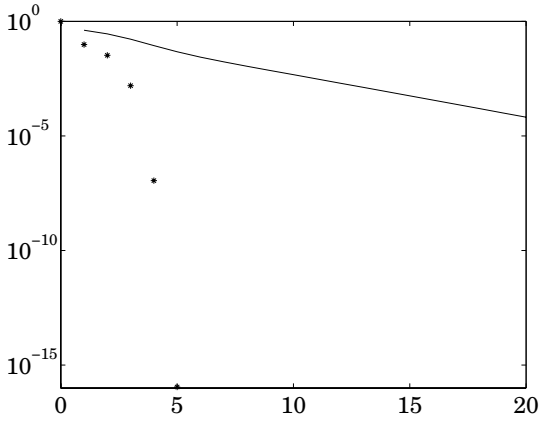


Abb. 25.1: Approximationsfehler bei Potenzmethode und Rayleigh-Quotienten-Iteration

und wegen $\xi_k \to 1$ existiert somit eine Konstante $C > 0$ mit

$$|\widehat{\lambda} - \mu_k| \;\le\; \frac{C}{|\lambda - \mu_{k-1}|^2}\,|\widehat{\lambda} - \mu_{k-1}|^3\,, \qquad k \to \infty\,.$$

Unter den genannten Vereinfachungen haben wir daher kubische Konvergenz nachgewiesen. □

Beispiel 25.5. Gesucht sei der größte Eigenwert der symmetrischen Tridiagonalmatrix

$$A = \begin{bmatrix} 1/2 & \beta_1 & & & 0 \\ \beta_1 & 1/2 & \beta_2 & & \\ & \beta_2 & 1/2 & \beta_3 & \\ & & \beta_3 & 1/2 & \beta_4 \\ 0 & & & \beta_4 & 1/2 \end{bmatrix} \qquad \text{mit} \qquad \beta_i = \frac{i}{2\sqrt{4i^2-1}}\,.$$

Nach dem Satz von Gerschgorin ist der größte Eigenwert von A kleiner als $\mu_0 = 1/2+\beta_1+\beta_2 = 1.046874\ldots$. Diese obere Schranke wählen wir als Startwert für die Rayleigh-Quotienten-Iteration. Für $z^{(0)}$ verwenden wir einen Zufallsvektor. Abbildung 25.1 veranschaulicht durch eingezeichnete Sterne die sehr schnelle Konvergenz dieses Verfahrens, die durchgezogene (linear abfallende) Kurve zeigt zum Vergleich die Fehler der Potenzmethode (Algorithmus 25.1) für dieses Beispiel. Für die Potenzmethode ergeben sich nach 20 Iterationen etwa vier signifikante Dezimalstellen des größten Eigenwerts; der Konvergenzfaktor liegt in diesem Beispiel nach Korollar 25.3 bei $(\lambda_2/\lambda_1)^2 \approx 0.6514$. Die kubisch konvergente Rayleigh-Quotienten-Iteration braucht hingegen nur fünf

Iterationen, um λ_1 auf Maschinengenauigkeit zu berechnen; die Näherungswerte lauten

k	μ_k
0	1.04687402434197
1	0.85630945419961
2	0.92010158422290
3	0.95153127920772
4	0.95308981076350
5	0.95308992296933

Die dunklen Ziffern sind jeweils korrekt. ◊

Die Rayleigh-Quotienten-Iteration kann auch bei beliebigen Matrizen $A \in \mathbb{K}^{n\times n}$ eingesetzt werden. Im allgemeinen ist die Konvergenz aber nur lokal quadratisch, vgl. Stewart [98, S. 345ff].

26 Das QR-Verfahren

Wir stellen im weiteren das QR-Verfahren vor, das in der Praxis am häufigsten zur Berechnung aller Eigenwerte einer beliebigen quadratischen Matrix eingesetzt wird. Das Verfahren vereinigt in sich die globalen Konvergenzeigenschaften der Potenzmethode und die schnelle lokale Konvergenz der Rayleigh-Quotienten-Iteration. Wir leiten in diesem Abschnitt zunächst das Verfahren und seine theoretischen Eigenschaften her, die effiziente numerische Implementierung ist dann Gegenstand des nächsten Abschnitts.

Das QR-Verfahren an sich läßt sich sehr einfach formulieren: Sei $A_0 = A$ und $\{\mu_k\}_{k\geq 0}$ eine Folge komplexer Zahlen, sogenannte „Shifts“. Dann berechnet man im k-ten Iterationsschritt mit Hilfe der QR-Zerlegung aus Abschnitt 13

$$A_k - \mu_k I = Q_k R_k \,, \tag{26.1a}$$

$$A_{k+1} = R_k Q_k + \mu_k I \,. \tag{26.1b}$$

Hierbei ist jeweils Q_k unitär und R_k eine rechte obere Dreiecksmatrix.

Lemma 26.1. *Seien μ_k, A_k, Q_k und R_k wie in* (26.1) *definiert. Dann gelten die folgenden Identitäten:*

(a) $A_{k+1} = Q_k^* A_k Q_k \,,$

(b) $A_{k+1} = (Q_0 Q_1 \cdots Q_k)^* A (Q_0 Q_1 \cdots Q_k) \,,$

(c) $\displaystyle\prod_{j=0}^{k} (A - \mu_j I) = (Q_0 Q_1 \cdots Q_k)(R_k R_{k-1} \cdots R_0) \,.$

Beweis. (a) Aufgrund von (26.1) ist

$$\begin{aligned} A_{k+1} &= R_k Q_k + \mu_k I = Q_k^* Q_k R_k Q_k + \mu_k Q_k^* Q_k \\ &= Q_k^* (Q_k R_k + \mu_k I) Q_k = Q_k^* A_k Q_k \,. \end{aligned}$$

(b) folgt sofort aus (a) durch vollständige Induktion.

(c) wird ebenfalls durch Induktion über k bewiesen. Für $k = 0$ ergibt sich die Behauptung aus der QR-Zerlegung (26.1a) von $A - \mu_0 I = A_0 - \mu_0 I$. Ferner ist wegen (b)

$$\begin{aligned} Q_{k+1} R_{k+1} &= A_{k+1} - \mu_{k+1} I \\ &= (Q_0 \cdots Q_k)^* A (Q_0 \cdots Q_k) - \mu_{k+1} (Q_0 \cdots Q_k)^* (Q_0 \cdots Q_k) \\ &= (Q_0 \cdots Q_k)^* (A - \mu_{k+1} I)(Q_0 \cdots Q_k) \,. \end{aligned}$$

Daraus folgt $Q_0 \cdots Q_k Q_{k+1} R_{k+1} = (A - \mu_{k+1} I)(Q_0 \cdots Q_k)$, und unter Verwendung der Induktionsannahme (c) erhalten wir schließlich

$$\begin{aligned} (Q_0 \cdots Q_k Q_{k+1})(R_{k+1} R_k) \cdots R_0 &= (A - \mu_{k+1} I)(Q_0 \cdots Q_k)(R_k \cdots R_0) \\ &= (A - \mu_{k+1} I) \prod_{j=0}^{k} (A - \mu_j I) \,, \end{aligned}$$

was zu beweisen war. □

Nach Lemma 26.1 (a) haben also alle Matrizen A_k, $k \in \mathbb{N}_0$, die gleichen Eigenwerte, da sie durch Ähnlichkeitstransformationen ineinander überführt werden können.

Wir wollen nun motivieren, daß die Matrizen A_k im Verlauf der Iteration allmählich obere Dreiecksmatrizen werden, von deren Diagonalen schließlich die Eigenwerte von A abgelesen werden können. Die Argumente hierfür beruhen im wesentlichen auf überraschenden Zusammenhängen zwischen dem QR-Verfahren und der Potenzmethode beziehungsweise der gebrochenen Iteration.

1. Zunächst beschränken wir uns auf den Fall ohne Shifts (d. h. wir setzen $\mu_k = 0$ für alle $k \in \mathbb{N}_0$). Nach Lemma 26.1 (c) ist dann

$$A^{k+1} = (Q_0 \cdots Q_k)(R_k \cdots R_0) = \boldsymbol{Q}_k \boldsymbol{R}_k \tag{26.2}$$

eine QR-Zerlegung von A^{k+1}. Vergleicht man speziell die erste Spalte dieser Matrixgleichung, so ergibt sich

$$A^{k+1} e_1 = \boldsymbol{Q}_k \boldsymbol{r}_{11}^{(k)} e_1 = \boldsymbol{r}_{11}^{(k)} \boldsymbol{q}_1^{(k)} \,,$$

wobei $\boldsymbol{r}_{11}^{(k)}$ das $(1,1)$-Element von $\boldsymbol{R}_k$ und $\boldsymbol{q}_1^{(k)}$ die erste Spalte von $\boldsymbol{Q}_k$ ist. Aufgrund der Erkenntnisse aus Abschnitt 25 über die Potenzmethode kann man daher erwarten, daß $\boldsymbol{q}_1^{(k)}$ für hinreichend große k eine gute Näherung an einen Eigenvektor zum dominanten Eigenwert λ_1 von A ist.
Nach Lemma 26.1 (b) ist $A_{k+1} = \boldsymbol{Q}_k^* A \boldsymbol{Q}_k$. Folglich gilt

$$A_{k+1} e_1 \;=\; \boldsymbol{Q}_k^* A \boldsymbol{q}_1^{(k)} \;\approx\; \lambda_1 \boldsymbol{Q}_k^* \boldsymbol{q}_1^{(k)} \;=\; \lambda_1 e_1 \,,$$

und somit hat A_{k+1} in etwa die Gestalt

$$A_{k+1} \;\approx\; \left[\begin{array}{c:c} \lambda_1 & \cdots \\ 0 & \\ \vdots & \\ 0 & \cdots \end{array}\right] .$$

2. Aus (26.2) folgt für eine invertierbare Matrix A wegen der Orthogonalität der Matrix $\boldsymbol{Q}_k$ die Gleichung $\boldsymbol{Q}_k^* = \boldsymbol{R}_k A^{-(k+1)}$ und Multiplikation mit e_n^* von links ergibt

$$\boldsymbol{q}_n^{(k)*} \;=\; e_n^* \boldsymbol{Q}_k^* \;=\; e_n^* \boldsymbol{R}_k A^{-(k+1)} \;=\; \boldsymbol{r}_{nn}^{(k)} e_n^* A^{-(k+1)} \,.$$

Der Vektor $\boldsymbol{q}_n^{(k)}$ – die letzte Spalte von $\boldsymbol{Q}_k$ – ist also nichts anderes als das Resultat von $k+1$ Schritten der inversen Iteration mit A^* und somit eine Näherung für einen linken Eigenvektor zu dem betragskleinsten Eigenwert λ_n von A.
Aus Lemma 26.1 (b) folgt daher

$$e_n^* A_{k+1} \;=\; e_n^* \boldsymbol{Q}_k^* A \boldsymbol{Q}_k \;=\; \boldsymbol{q}_n^{(k)*} A \boldsymbol{Q}_k \;\approx\; \lambda_n \boldsymbol{q}_n^{(k)*} \boldsymbol{Q}_k \;=\; \lambda_n e_n^* \,. \qquad (26.3)$$

Also ist die letzte Zeile von A_{k+1} näherungsweise ein Vielfaches von e_n^*, und zusammen mit der ersten Beobachtung zuvor ergibt sich für A_{k+1} näherungsweise die Gestalt

$$A_{k+1} \;\approx\; \left[\begin{array}{c:ccc} \lambda_1 & & \cdots & \\ 0 & & & \\ \vdots & & \cdots & \\ 0 & & & \\ \hdashline 0 & 0 & \cdots\; 0 & \lambda_n \end{array}\right] .$$

Damit soll nun ausreichend motiviert sein, daß die Matrizen A_k für $k \to \infty$ die Form einer rechten oberen Dreiecksmatrix annehmen.

3. Stellen wir uns nun vor, wir wollten die Konvergenz der in 2. beobachteten inversen Iteration zum kleinsten Eigenwert λ_n von A (und damit auch von A_k) beschleunigen. Als linken Näherungs-Eigenvektor für A_k haben wir in (26.3) den n-ten kartesischen Einheitsvektor identifiziert. Wegen der lokal schnellen Konvergenz der Rayleigh-Quotienten-Iteration liegt es nahe, Algorithmus 25.2 anzuwenden, also den Rayleigh-Quotienten $\mu_k = e_n^* A_k e_n$ zu bilden (das ist gerade das rechte untere Eckelement von A_k) und dann einen Schritt der inversen Iteration bezüglich des linken Eigenvektors auszuführen: Wegen (26.1a) ergibt dies

$$e_n^*(A_k - \mu_k I)^{-1} = e_n^* R_k^{-1} Q_k^* = \frac{1}{r_{nn}^{(k)}} e_n^* Q_k^* = \frac{1}{r_{nn}^{(k)}} q_n^{(k)*},$$

wobei $r_{nn}^{(k)}$ das rechte untere Eckelement von R_k und $q_n^{(k)}$ die hinterste Spalte von Q_k bezeichnet; hierzu muß man lediglich beachten, daß R_k^{-1} wieder eine obere Dreiecksmatrix ist, deren Diagonaleinträge gerade die Kehrwerte der entsprechenden Diagonaleinträge von R_k sind.
Mit anderen Worten: Ein Schritt der Rayleigh-Quotienten-Iteration ergibt gerade die hinterste Spalte von Q_k als neue Näherung an den linken Eigenvektor von A_k zu λ_n. Darüberhinaus sieht man mit Hilfe von Lemma 26.1 (a) sofort, daß das rechte untere Eckelement von A_{k+1} gerade der zugehörige Rayleigh-Quotient, also der nächste Shift μ_{k+1} aus Algorithmus 25.2 ist:

$$\mu_{k+1} \;=\; q_n^{(k)*} A_k q_n^{(k)} \;=\; e_n^* Q_k^* A_k Q_k e_n \;=\; e_n^* A_{k+1} e_n \,.$$

Wählt man also als Shift μ_k in (26.1) jeweils das (n,n)-Element von A_k, dann darf man wie bei der Rayleigh-Quotienten-Iteration sehr schnelle (quadratische oder gar kubische) Konvergenz dieser Eckelemente gegen den kleinsten Eigenwert λ_n von A erwarten.
Shifts in (26.1) dienen also der Konvergenzbeschleunigung des QR-Verfahrens.

Dies mag als Motivation des QR-Verfahrens genügen. Der Vollständigkeit halber beweisen wir nun die lineare Konvergenz des Verfahrens ohne Shifts für einen Spezialfall:

Satz 26.2. *$A \in \mathbb{K}^{n\times n}$ sei diagonalisierbar mit paarweise verschiedenen Eigenwerten $|\lambda_1| > |\lambda_2| > \ldots > |\lambda_n| > 0$, $\Lambda \in \mathbb{K}^{n\times n}$ die entsprechende Diagonalmatrix mit den Diagonaleinträgen λ_i, $i = 1,\ldots,n$ und $X = [x_1,\ldots,x_n]$ die zugehörige Eigenvektormatrix, d. h. es gilt $A = X\Lambda X^{-1}$. Existiert die LR-Zerlegung von X^{-1}, $X^{-1} = LU$, dann sind die Matrizen A_k des QR-Verfahrens ohne Shifts (d. h. mit $\mu_k = 0$ in (26.1) für alle $k \in \mathbb{N}_0$) asymptotisch obere Dreiecksmatrizen und ihr Diagonalanteil* $\operatorname{diag}(A_k)$ *konvergiert für $k \to \infty$ mindestens linear gegen Λ.*[3]

[3]Für $A \in \mathbb{K}^{n\times n}$ bezeichnet $\operatorname{diag}(A) \in \mathbb{K}^{n\times n}$ diejenige Diagonalmatrix, deren Hauptdiagonale mit der von A übereinstimmt.

Beweis. Wegen $\mu_k = 0$ ergibt sich aus Lemma 26.1 (c) die QR-Zerlegung

$$A^k \;=\; \boldsymbol{Q}_{k-1}\boldsymbol{R}_{k-1} \;=\; (Q_0 \cdots Q_{k-1})(R_{k-1} \cdots R_0) \tag{26.4}$$

von A^k. Andererseits ist wegen der Existenz der LR-Zerlegung von X^{-1}

$$A^k \;=\; (X\Lambda X^{-1})^k \;=\; X\Lambda^k X^{-1} \;=\; X\Lambda^k LU \;=\; (X\Lambda^k L\Lambda^{-k})\Lambda^k U \;=\; X_k\Lambda^k U\,,$$

wobei $X_k = X\Lambda^k L\Lambda^{-k}$ gesetzt wurde. X_k hat eine QR-Zerlegung $X_k = P_k U_k$ mit invertierbarer oberer Dreiecksmatrix U_k (da auch X_k invertierbar ist), und

$$A^k \;=\; P_k(U_k\Lambda^k U)$$

ist somit eine weitere QR-Zerlegung von A^k. Durch einen Vergleich mit (26.4) ergibt sich daher, daß P_k und $\boldsymbol{Q}_{k-1}$ sowie $U_k\Lambda^k U$ und $\boldsymbol{R}_{k-1}$ bis auf eine unitäre Diagonalmatrix S_k übereinstimmen, vgl. Aufgabe III.8:

$$\boldsymbol{Q}_{k-1} = P_k S_k^*\,, \qquad \boldsymbol{R}_{k-1} = S_k U_k \Lambda^k U\,. \tag{26.5}$$

Wegen

$$Q_k \;=\; (Q_0 \cdots Q_{k-1})^{-1}(Q_0 \cdots Q_{k-1}Q_k) \;=\; \boldsymbol{Q}_{k-1}^{-1}\boldsymbol{Q}_k \;=\; S_k P_k^{-1} P_{k+1} S_{k+1}^*$$

und

$$\begin{aligned} R_k \;&=\; (R_k R_{k-1} \cdots R_0)(R_{k-1} \cdots R_0)^{-1} \;=\; \boldsymbol{R}_k \boldsymbol{R}_{k-1}^{-1} \\ &=\; S_{k+1}U_{k+1}\Lambda^{k+1}UU^{-1}\Lambda^{-k}U_k^{-1}S_k^* \;=\; S_{k+1}U_{k+1}\Lambda U_k^{-1}S_k^* \end{aligned}$$

folgt daher aus (26.1a)

$$\begin{aligned} A_k \;&=\; Q_k R_k \;=\; S_k P_k^{-1} P_{k+1} S_{k+1}^* S_{k+1} U_{k+1} \Lambda U_k^{-1} S_k^* \\ &=\; S_k U_k U_k^{-1} P_k^{-1} P_{k+1} U_{k+1} \Lambda U_k^{-1} S_k^* \\ &=\; S_k U_k X_k^{-1} X_{k+1} \Lambda U_k^{-1} S_k^*\,. \end{aligned} \tag{26.6}$$

Bezeichnen wir mit l_{ij} die Einträge von L, dann ist insbesondere $l_{ii} = 1$ und aus der Anordnung der Eigenwerte λ_i in Λ ergeben sich die Einträge von $\Lambda^k L\Lambda^{-k}$ zu

$$(\Lambda^k L\Lambda^{-k})_{ij} \;=\; \lambda_i^k l_{ij} \lambda_j^{-k} \;=\; \begin{cases} 0 & \text{für } i < j\,, \\ 1 & \text{für } i = j\,, \\ O(q^k) & \text{für } i > j\,, \end{cases}$$

und für ein q mit $0 < q < 1$. Somit ist

$$X_k = X\Lambda^k L\Lambda^{-k} = X + E_k \qquad \text{mit} \quad \|E_k\|_2 = O(q^k)\,, \quad k \to \infty\,.$$

Demzufolge erhalten wir

$$X_k^{-1} X_{k+1} \;=\; (X + E_k)^{-1}(X + E_{k+1}) \;=\; I + F_k$$

mit $\|F_k\|_2 = O(q^k)$, und eingesetzt in (26.6) ergibt sich

$$A_k \;=\; S_k U_k \Lambda U_k^{-1} S_k^* \;+\; S_k U_k F_k \Lambda U_k^{-1} S_k^* \,. \tag{26.7}$$

Da P_k und S_k unitäre Matrizen sind, ist $\|U_k\|_2 = \|P_k^* X_k\|_2 = \|X_k\|_2$ und $\|U_k^{-1}\|_2 = \|X_k^{-1}\|_2$, so daß der zweite Term in (26.7) wegen der Konvergenz $X_k \to X$ für $k \to \infty$ durch

$$\|S_k U_k F_k \Lambda U_k^{-1} S_k^*\|_2 \;\le\; \mathrm{cond}_2(X_k)\, |\lambda_1|\, \|F_k\|_2 \;=\; O(q^k) \tag{26.8}$$

abgeschätzt werden kann. Daher ergibt sich asymptotisch

$$A_k \;\sim\; S_k U_k \Lambda U_k^{-1} S_k^* \,, \qquad k \to \infty \,,$$

und dies ist als Produkt oberer Dreiecksmatrizen selbst eine obere Dreiecksmatrix mit

$$\mathrm{diag}(A_k) \;\sim\; S_k \,\mathrm{diag}(U_k) \Lambda \,\mathrm{diag}(U_k)^{-1} S_k^* \;=\; \Lambda \,, \qquad k \to \infty \,.$$

Nach (26.7) und (26.8) konvergiert der Fehler

$$\|A_k - S_k U_k \Lambda U_k^{-1} S_k\|_2 \;=\; \|S_k U_k F_k \Lambda U_k^{-1} S_k^*\|_2$$

linear gegen Null. □

27 Implementierung des QR-Verfahrens

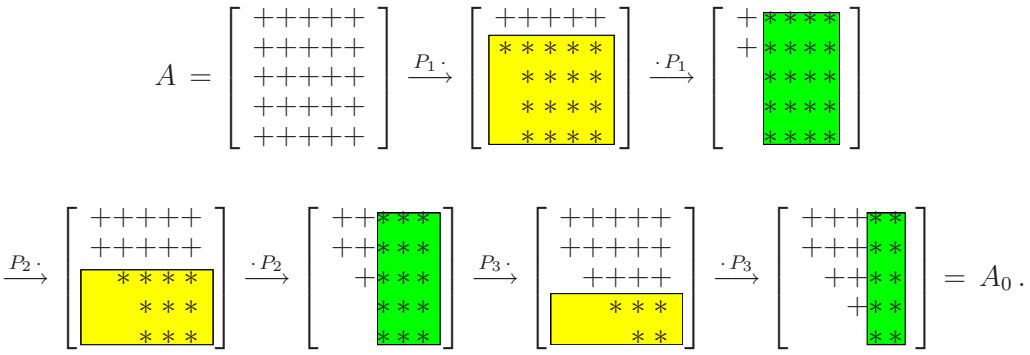
Im folgenden gehen wir auf einige praktische Aspekte bei der Implementierung des QR-Verfahrens ein.

27.1 Reduktion auf Hessenberg-Form

Für beliebige Matrizen $A \in \mathbb{K}^{n\times n}$ ist das QR-Verfahren sehr aufwendig, denn jede Iteration benötigt etwa $O(n^3)$ Operationen. Um diesen Aufwand zu reduzieren, wird die Matrix A zunächst durch Ähnlichkeitstransformationen auf obere Hessenberg-Form transformiert (vgl. Definition 14.1). Die Transformation wird ähnlich wie in Abschnitt 13 durchgeführt: Man konstruiert geeignete Householder-Matrizen $P_1, \ldots, P_{n-2}$ und transformiert A sukzessive:

$$A \;\longmapsto A_0 = P^* A P \,, \qquad P = P_1 \cdots P_{n-2} \,. \tag{27.1}$$

Das Konstruktionsprinzip ist in dem folgenden Schema für $n = 5$ verdeutlicht:



Die hellgrau hinterlegten Flächen heben diejenigen Zeilen hervor, die bei der Multiplikation mit den Householder-Transformationen von links verändert werden. Die Householder-Transformation wird wie in Abschnitt 13 jeweils so bestimmt, daß die erste von Null verschiedene Spalte in dem hellgrauen Bereich ein Vielfaches von e_1 ist. Bei der anschließenden Multiplikation dieser Householder-Transformation von rechts werden die dunkelgrau hinterlegten Einträge der Matrix verändert. Die Sterne geben die jeweils neu zu berechnenden Matrixeinträge an.

Aufwand. Durch Vergleich mit der Aufwandsabschätzung aus Abschnitt 13 ergibt sich unmittelbar, daß für jeden Stern in der obigen Umformung etwa zwei Multiplikationen benötigt werden. Daher umfaßt der Aufwand für die Transformation in Hessenberg-Form ungefähr

$$\sum_{k=2}^{n-1} 2(k^2 + nk) \approx \frac{2}{3}\, n^3 + n^3 = \frac{5}{3}\, n^3 \quad \text{Multiplikationen}\,.$$

$\Diamond$

27.2 Ein Iterationsschritt des QR-Verfahrens

Für die QR-Zerlegung der Hessenberg-Matrix $A_k - \mu_k I$ in Teilschritt (26.1a) der QR-Iteration empfiehlt sich der effiziente Algorithmus aus Abschnitt 14. Mit Hilfe von Givens-Rotationen $G(j, j+1, c_j, s_j)$, $j = 1, \ldots, n-1$, erhält man die Transformation

$$\begin{aligned} A_k - \mu_k I \;\longmapsto\; R_k &= Q_k^*(A_k - \mu_k I) \qquad \text{mit} \\ Q_k^* &= G(n-1, n, c_{n-1}, s_{n-1}) \cdots G(1, 2, c_1, s_1)\,, \end{aligned}$$

die im folgenden Schema wieder für $n = 5$ illustriert wird:

$$A_k - \mu_k I \;=\; \begin{bmatrix} + & + & + & + & + & \\ + & + & + & + & + & \\ & + & + & + & + & \\ & & + & + & + & \\ & & & + & + & \end{bmatrix} \xrightarrow{i=1} \begin{bmatrix} * & * & * & * & * & \\ * & * & * & * & & \\ & + & + & + & + & \\ & & + & + & + & \\ & & & + & + & \end{bmatrix}$$

$$\xrightarrow{i=2} \begin{bmatrix} + & + & + & + & + & \\ * & * & * & * & & \\ & * & * & * & & \\ & & + & + & + & \\ & & & + & + & \end{bmatrix} \xrightarrow{i=3} \begin{bmatrix} + & + & + & + & + & \\ & + & + & + & + & \\ & & * & * & * & \\ & & * & * & & \\ & & & + & + & \end{bmatrix} \xrightarrow{i=4} \begin{bmatrix} + & + & + & + & + & \\ & + & + & + & + & \\ & & + & + & + & \\ & & & * & * & \\ & & & & * & \end{bmatrix} \;=\; R_k\,.$$

An den grau hinterlegten Flächen erkennt man wieder, welche Zeilen der Matrix durch die Givens-Rotationen kombiniert werden, die Sterne kennzeichnen neu berechnete Einträge.

Im Teilschritt (26.1b) der Iteration muß das Produkt der Givens-Rotationen von rechts an R_k heranmultipliziert werden:

$$R_k \;\mapsto A_{k+1} \;=\; R_k\, G(1,2,c_1,s_1)^* \cdots G(n-1,n,c_{n-1},s_{n-1})^* \;+\; \mu_k I.$$

Dabei kombiniert eine Multiplikation von rechts mit $G(i,i{+}1,c_i,s_i)^*$ jeweils die Spalten i und $i{+}1$ des Zwischenergebnisses. Daher hat A_{k+1} wieder Hessenberg-Form:

$$R_k \;=\; \begin{bmatrix} + & + & + & + & + & \\ & + & + & + & + & \\ & & + & + & + & \\ & & & + & + & \\ & & & & + & \end{bmatrix} \xrightarrow{i=1} \begin{bmatrix} * & * & + & + & + & \\ * & * & + & + & + & \\ & & + & + & + & \\ & & & + & + & \\ & & & & + & \end{bmatrix}$$

$$\xrightarrow{i=2} \begin{bmatrix} + & * & * & + & + & \\ + & * & * & + & + & \\ & * & * & + & + & \\ & & & + & + & \\ & & & & + & \end{bmatrix} \xrightarrow{i=3} \begin{bmatrix} + & + & * & * & + & \\ + & + & * & * & + & \\ & + & * & * & + & \\ & & * & * & + & \\ & & & & + & \end{bmatrix} \xrightarrow{i=4} \begin{bmatrix} + & + & + & * & * & \\ + & + & + & * & * & \\ & + & + & * & * & \\ & & + & * & * & \\ & & & * & * & \end{bmatrix} \;=\; A_{k+1} - \mu_k I\,.$$

Aufwand. Die beiden Teilschritte (26.1) benötigen somit

$$2\sum_{i=1}^{n-1} 4(n-i+1) \;=\; 2\sum_{i=2}^{n} 4i \;\sim\; 4n^2$$

Multiplikationen. Bei geeigneter Wahl der Shift-Parameter μ_k (vgl. Abschnitt 27.3) konvergiert das QR-Verfahren in der Regel quadratisch und man kann

daher davon ausgehen, daß eine konstante Anzahl Iterationen ausreicht, um einen einzelnen Eigenwert hinreichend genau zu bestimmen. Insgesamt werden somit $O(n)$ Iterationen für die Berechnung aller Eigenwerte benötigt. Auf der Grundlage dieser Abschätzungen ergibt sich ungefähr ein Gesamtaufwand $O(n^3)$ für das vollständige QR-Verfahren (inklusive der Transformation in obere Hessenberg-Form, die in Abschnitt 27.1 beschrieben wurde). $\Diamond$

27.3 Bestimmung der Shifts und Deflation

Gemäß der Vorüberlegungen in Abschnitt 26 bietet sich für die Konvergenzbeschleunigung des QR-Verfahrens als Shift μ_k das (n,n)-Eckelement $a_{nn}^{(k)}$ von A_k an. Als noch erfolgreicher erweist sich eine andere Strategie, bei der μ_k aus dem rechten unteren (2×2)-Eckblock

$$A_k^{(2\times 2)} = \begin{bmatrix} a_{n-1,n-1}^{(k)} & a_{n-1,n}^{(k)} \\ a_{n,n-1}^{(k)} & a_{nn}^{(k)} \end{bmatrix} \tag{27.2}$$

von A_k wie folgt bestimmt wird:

$$\begin{aligned} &\mu_k \text{ sei der Eigenwert von } A_k^{(2\times 2)} \text{ aus (27.2),} \\ &\text{der am nächsten an } a_{nn}^{(k)} \text{ liegt.} \end{aligned} \tag{27.3}$$

In beiden Fällen konvergiert das (n,n)-Element von A_k sehr schnell gegen den exakten Eigenwert und das $(n,n-1)$-Element gegen Null, also

$$A_k \longrightarrow \left[\begin{array}{cccc:c} * & \cdots & \cdots & * & * \\ * & \ddots & & \vdots & \vdots \\ & \ddots & \ddots & \vdots & \vdots \\ 0 & & * & * & * \\ \hdashline 0 & \cdots & \cdots & 0 & \lambda_n \end{array}\right] = \left[\begin{array}{ccc:c} & & & * \\ & B_{k-1} & & \vdots \\ & & & * \\ \hdashline 0 & \cdots & 0 & \lambda_n \end{array}\right].$$

Ab diesem Zeitpunkt reicht es aus, nur noch das kleinere Teilproblem mit der Hessenberg-Matrix B_{k-1} zu betrachten; diese Reduktion wird *Deflation* genannt. Das Problem zerfällt auch ansonsten gelegentlich in Teilprobleme, wenn ein Nebendiagonalelement $a_{i+1,i}^{(k)}$, $i = 1, \ldots, n-2$, Null wird oder im Bereich der Maschinengenauigkeit liegt.

27.4 Komplexe Eigenwerte

Ist die Matrix A reell, so sind auch die Matrizen Q_k und R_k aus (26.1) reell und das gesamte Verfahren kann reell formuliert werden. Auf diese Weise lassen sich

jedoch keine komplex konjugierten Eigenwerte von A bestimmen. Statt dessen entwickeln sich in diesem Fall rechts unten in A_k 2×2-Blöcke wie in (27.2), deren Eigenwerte mit den beiden gesuchten komplex konjugierten Eigenwerten übereinstimmen – ähnlich wie bei der reellen Jordan-Normalform. Um diese Eigenwerte zu bestimmen, müßte eigentlich mit komplexen $\mu_k \in \mathbb{C}$ geshiftet werden. Komplexe Shifts und die damit verbundene komplexe Arithmetik lassen sich jedoch umgehen, wenn zwei komplexe Shifts (mit komplex konjugierten μ_k und $\mu_{k+1} = \overline{\mu_k}$) zu zwei reellen Schritten geeignet zusammengefaßt werden; für genauere Details sei auf das Buch von Golub und Van Loan [34] verwiesen.

27.5 Hermitesche Matrizen

Ist A hermitesch, dann ist die Konvergenz lokal kubisch, entsprechend zur Konvergenz der Rayleigh-Quotienten-Iteration, vgl. Satz 25.4 und die Motivation aus Abschnitt 26 für die Verwendung der Shifts μ_k. Zudem sind unter dieser Zusatzvoraussetzung alle Matrizen A_k nach Lemma 26.1 (a) hermitesch. A_k ist also eine hermitesche Hessenberg-Matrix und damit zwangsläufig eine Tridiagonalmatrix. Dadurch werden die einzelnen Iterationsschritte billiger. Sie erfordern nur $O(n)$ Operationen, so daß nach etwa $O(n^2)$ Operationen alle Eigenwerte von A berechnet sind. Nach Satz 26.2 konvergieren die Matrizen A_k in diesem Fall gegen eine Diagonalmatrix.

Für hermitesche Tridiagonalmatrizen gibt es aber inzwischen neuere Verfahren zur Bestimmung aller (oder einzelner) Eigenwerte und Eigenvektoren, deren Aufwand ähnlich oder gar geringer ist wie der Aufwand des QR-Verfahrens, vgl. Abschnitt 29.

27.6 Bestimmung von Eigenvektoren

Prinzipiell gibt es zwei Möglichkeiten zur Berechnung der zugehörigen Eigenvektoren: Faßt man alle orthogonalen Transformationen in einem Produkt $\boldsymbol{Q} = Q_0 Q_1 Q_2 \cdots$ zusammen, dann ergibt sich im Grenzübergang die Faktorisierung $R = \boldsymbol{Q}^* A \boldsymbol{Q}$ der oberen Dreiecksmatrix R. Die Eigenwerte von A sind gerade die Diagonalelemente r_{ii} von R und aus den zugehörigen Eigenvektoren z_i von R ergeben sich die Eigenvektoren $\boldsymbol{Q} z_i$ von A. Der Eigenvektor $z_i = [\zeta_j]_{j=1}^n \in \mathcal{N}(R - r_{ii} I)$ kann beispielsweise bestimmt werden, indem man $\zeta_i = 1$ und $\zeta_j = 0$ für $j > i$ initialisiert und dann die verbliebenen Komponenten $\zeta_1, \ldots, \zeta_{i-1}$ durch Rückwärtssubstitution aus dem Gleichungssystem $(R - r_{ii} I) z_i = 0$ bestimmt. Dies ist jedoch nicht unbedingt stabil, und wegen der notwendigen expliziten Berechnung von $\boldsymbol{Q}$ ist diese Vorgehensweise

darüber hinaus recht teuer.

Alternativ kann zunächst die gebrochene Iteration auf die Hessenberg-Matrix $A_0 = P^* A P$ zur Approximation der Eigenvektoren von A_0 angewendet werden. Die dabei auftretenden Gleichungssysteme mit Hessenberg-Matrizen $A_0 - \lambda I$ können sehr effizient mit einer QR-Zerlegung gelöst werden, die analog zu Abschnitt 27.2 implementiert werden kann. Als Shift λ verwendet man hierbei die berechnete Eigenwertnäherung des QR-Verfahrens. Da diese Approximation sehr gut ist, reicht meist ein einziger Schritt der gebrochenen Iteration aus, um den Eigenvektor hinreichend genau zu bestimmen. Allerdings hängt die Genauigkeit der Näherung unter Umständen stark von dem verwendeten Startvektor ab, vgl. Abschnitt 29.2.

27.7 Ein abschließendes Beispiel

Wir wenden das QR-Verfahren auf die reelle symmetrische Tridiagonalmatrix $A_0 = A$ aus Beispiel 25.5 an. Die Shifts μ_k werden dabei nach der Strategie (27.3) bestimmt. Nachfolgend werden für jedes k die rechten unteren Eckelemente sowie die Shifts μ_k angegeben:[4]

k	$a_{54}^{(k)}$	$a_{55}^{(k)}$	μ_k
0	0.25197631533948	0.50000000000000	0.75197631533948
1	0.03040951681439	0.76711805522064	0.76964359979529
2	−0.00004026652781	0.76923465228119	0.76923465966410
3	0.00000000000073	0.76923465505284	0.76923465505284
4	0.00000000000000	0.76923465505284	

Nach vier Schritten wird also $\lambda_5 = 0.76923465505284$ erkannt.

Wie in Abschnitt 27.3 erläutert, kann die Dimension des Problems nun um Eins erniedrigt werden. Die Weiterbehandlung der verbliebenen 4×4-Matrix ergibt dann nach weiteren drei Schritten $\lambda_4 = 0.95308992296933$:

k	$a_{43}^{(k)}$	$a_{44}^{(k)}$	μ_k
4	−0.22702435034463	0.75365545870499	0.95085368678850
5	−0.00198108984877	0.95308129563224	0.95308990180718
6	−0.00000000009223	0.95308992296933	0.95308992296933
7	−0.00000000000000	0.95308992296933	

[4] Die MATLAB-Routine `eigmovie` bietet eine entsprechende Möglichkeit, die Zwischenergebnisse des QR-Verfahrens mitzuverfolgen.

Nach erneuter Deflation erhalten wir

k	$a_{32}^{(k)}$	$a_{33}^{(k)}$	μ_k
7	0.01101753998910	0.49955690005666	0.49999463431359
8	0.00000021661453	0.49999999999983	0.50000000000000
9	0.00000000000000	0.50000000000000	

und $\lambda_3 = 0.5$ bis auf Maschinengenauigkeit. Die verbliebene 2×2–Matrix hat die Eigenwerte

$$\lambda_1 = 0.04691007703067 \qquad \text{und} \qquad \lambda_2 = 0.23076534494716\,.$$

Wie man sieht, werden aufgrund der überaus schnellen Konvergenz für keinen Eigenwert mehr als vier Iterationen des QR-Verfahrens benötigt, was die Aufwandsabschätzung aus Abschnitt 27.2 bestätigt.

Zum Vergleich noch das Ergebnis des *QR-Verfahrens ohne Shift* nach neun Iterationen. Die Diagonal- und Nebendiagonalelemente der (Tridiagonal-)Matrix $A_9 \in \mathbb{R}^{5\times 5}$ lauten

i	$a_{i,i-1}^{(9)}$	$a_{ii}^{(9)}$
1		0.94555805304325
2	0.03647724695393	0.77591442936831
3	0.01524997438615	0.50085001330370
4	0.00074933824927	0.23076742725381
5	−0.00000021714632	0.04691007703092

Man beachte, daß die Näherungen für die Eigenwerte von A in der Diagonalen stehen, also in der letzten Spalte der Tabelle zu suchen sind; fehlerhafte Dezimalstellen sind dort heller dargestellt. Anhand der Nebendiagonalelemente erkennt man, daß kein Eigenwert im Rahmen der Maschinengenauigkeit ist, lediglich a_{55} stellt eine vernünftige Näherung an λ_5 dar.

Für die Berechnung der zugehörigen Eigenvektoren mit der inversen Iteration verweisen wir an dieser Stelle auf das spätere Beispiel 29.5.

28 Das Jacobi-Verfahren

In den verbleibenden Abschnitten dieses Kapitels beschränken wir uns auf das numerische Eigenwertproblem für hermitesche Matrizen $A = A^* \in \mathbb{K}^{n\times n}$.

Zunächst soll das *Jacobi-Verfahren* vorgestellt werden. Dazu schreiben wir

$$A = D - R - R^*\,,$$

wobei $D = \operatorname{diag}(A)$ eine Diagonalmatrix und R eine echte obere Dreiecksmatrix ist. Dann ist

$$S(A) \;=\; \|R + R^*\|_F^2 \;=\; \sum_{\substack{i,j=1 \\ i \neq j}}^{n} |a_{ij}|^2 \tag{28.1}$$

ein Maß dafür, wie gut die Diagonalelemente von A die Eigenwerte von A approximieren:

Proposition 28.1. *Ist d_{ii} ein beliebiges Diagonalelement von A, dann existiert ein $\lambda \in \sigma(A)$ mit*

$$|d_{ii} - \lambda| \;\leq\; \sqrt{S(A)}\,.$$

Beweis. Die Aussage folgt sofort aus dem Satz 24.4 von Wielandt-Hoffman oder aus Aufgabe I.11 und Korollar 24.2 zu dem Satz von Bauer-Fike: Demnach existiert ein $\lambda \in \sigma(A)$ mit

$$|d_{ii} - \lambda| \;\leq\; \|R + R^*\|_2 \;\leq\; \|R + R^*\|_F \;=\; \sqrt{S(A)}\,. \qquad \square$$

Beim Jacobi-Verfahren soll das Maß $S(A)$ aus (28.1) mit unitären Ähnlichkeitstransformationen

$$A_{k+1} = Q_k^* A_k Q_k\,, \quad k = 0, 1, 2, \ldots\,, \qquad A_0 = A\,, \tag{28.2}$$

sukzessive verkleinert werden. Als Transformationsmatrizen Q_k werden Givens-Rotationen verwendet. Um die folgende Darstellung möglichst einfach zu gestalten, beschränken wir uns auf reelle (symmetrische) Matrizen $A \in \mathbb{R}^{n\times n}$.

Lemma 28.2. *$A = [a_{ij}] \in \mathbb{R}^{n\times n}$ sei eine symmetrische Matrix und $i, j \in \{1, \ldots, n\}$ mit $i < j$ seien fest gewählte Indizes. Ferner sei $G = G(i, j, c, s)$ mit $c = \cos\theta$ und $s = \sin\theta$ die reelle Givens-Rotation mit dem (eindeutig bestimmten) Winkel θ, der die Gleichung*

$$\tan(2\theta) \;=\; \frac{2a_{ij}}{a_{jj} - a_{ii}}\,, \qquad |\theta| \leq \pi/4\,, \tag{28.3}$$

erfüllt (für $a_{ii} = a_{jj}$ sei $\theta = \pi/4$). Dann ist das (i, j)-Element b_{ij} der Matrix $B = G^*AG$ Null.

Beweis. Sei $B = G^*AG = [b_{kl}]$. Nach Bemerkung 14.4 hängen die Elemente b_{ii}, b_{ij}, b_{ji} und b_{jj} lediglich von den Einträgen a_{ii}, a_{ij}, a_{ji} und a_{jj} von A ab: Genauer gilt, vgl. (14.2),

$$\begin{bmatrix} b_{ii} & b_{ij} \\ b_{ji} & b_{jj} \end{bmatrix} \;=\; \begin{bmatrix} c & -s \\ s & c \end{bmatrix} \begin{bmatrix} a_{ii} & a_{ij} \\ a_{ji} & a_{jj} \end{bmatrix} \begin{bmatrix} c & s \\ -s & c \end{bmatrix}\,. \tag{28.4}$$

Insbesondere ist also

$$\begin{aligned} b_{ij} &= \begin{bmatrix} ca_{ii} - sa_{ji},\ ca_{ij} - sa_{jj} \end{bmatrix} \begin{bmatrix} s \\ c \end{bmatrix} = sca_{ii} - s^2 a_{ji} + c^2 a_{ij} - sca_{jj} \\ &= sc(a_{ii} - a_{jj}) + (c^2 - s^2) a_{ij}\,, \end{aligned}$$

da A symmetrisch ist. Für $a_{ii} = a_{jj}$ und $s = c = 1/\sqrt{2}$, also für $\theta = \pi/4$, ergibt dies $b_{ij} = 0$. Für $a_{ii} \neq a_{jj}$ ergibt die spezielle Wahl (28.3) von θ die Identität

$$\frac{2a_{ij}}{a_{jj} - a_{ii}} = \tan(2\theta) = \frac{\sin(2\theta)}{\cos(2\theta)} = \frac{2sc}{c^2 - s^2}\,, \tag{28.5}$$

und damit ist auch in diesem Fall $b_{ij} = 0$. □

Sei nun $B = G^* A G$ die Transformation aus Lemma 28.2. Dann ist $\|B\|_F = \|A\|_F$, vgl. (24.5), und wegen

$$S(A) = \|A\|_F^2 - \sum_{\nu=1}^{n} a_{\nu\nu}^2$$

ergibt sich

$$S(B) = \|B\|_F^2 - \sum_{\nu=1}^{n} b_{\nu\nu}^2 = \|A\|_F^2 - \sum_{\nu=1}^{n} b_{\nu\nu}^2 = S(A) + \sum_{\nu=1}^{n} \left(a_{\nu\nu}^2 - b_{\nu\nu}^2\right).$$

Da die Transformation $B = G^* A G$ außer den Elementen a_{ii} und a_{jj} alle Diagonalelemente von A invariant läßt, erhalten wir

$$S(B) = S(A) + (a_{ii}^2 + a_{jj}^2 - b_{ii}^2 - b_{jj}^2)\,. \tag{28.6}$$

Die gleiche Überlegung kann auf die 2×2-Transformation (28.4) angewendet werden: Dann ergibt sich aufgrund der Symmetrie von A und B entsprechend

$$2b_{ij}^2 = 2a_{ij}^2 + (a_{ii}^2 + a_{jj}^2 - b_{ii}^2 - b_{jj}^2)\,.$$

Aus (28.6) und $b_{ij} = 0$ folgt somit

$$S(B) = S(A) + 2b_{ij}^2 - 2a_{ij}^2 = S(A) - 2a_{ij}^2\,. \tag{28.7}$$

Bei dem Jacobi-Verfahren geht man nun wie folgt vor: Für den Iterationsschritt (28.2) wählt man ein Nebendiagonalelement von A_k aus und bestimmt die Givens-Rotation Q_k, die dieses Element gemäß Lemma 28.2 auf Null transformiert. Aus (28.7) folgt, daß auf diese Weise das Maß $S(A_k)$ reduziert wird. Für die Auswahl des entsprechenden Matrixelements werden hauptsächlich die folgenden beiden Varianten verwendet:

1. Beim *klassischen Jacobi-Verfahren* wird - im Hinblick auf (28.7) - ein betragsgrößtes Nebendiagonalelement ausgewählt. Die Maximumssuche erfordert jedoch $O(n^2)$ Operationen und ist damit relativ aufwendig.
2. Billiger zu implementieren ist das *zyklische Jacobi-Verfahren*, bei dem die Nebendiagonalelemente zyklisch (zeilenweise) durchlaufen werden:

$$\begin{aligned}(i,j) \;=\; & (1,2),(1,3),\ldots,(1,n),(2,3),\ldots,(2,n),\ldots,(n-1,n),\\ & (1,2),(1,3),\ldots\end{aligned}$$

In beiden Fällen konvergiert $S(A_k)$ gegen Null für $k \to \infty$. Wir beweisen das im folgenden für die klassische Variante; für das zyklische Verfahren ist der Beweis erheblich schwieriger, vgl. Forsythe und Henrici [28].

Satz 28.3. *Beim klassischen Jacobi-Verfahren gilt $S(A_k) \to 0$ für $k \to \infty$, d. h. die Diagonaleinträge von A_k konvergieren gegen die Eigenwerte von A.*

Beweis. Sei $a_{ij}^{(k)}$ ein betragsgrößtes Nebendiagonalelement von A_k. Dann gilt

$$S(A_k) \;\le\; (n-1)n|a_{ij}^{(k)}|^2\,,$$

und eingesetzt in (28.7) (mit $B = A_{k+1}$ und $A = A_k$) folgt

$$S(A_{k+1}) \le S(A_k) - \frac{2}{(n-1)n}S(A_k) = \big(1 - \frac{2}{(n-1)n}\big)S(A_k)\,. \quad (28.8)$$

Also konvergiert $S(A_k)$ mindestens linear gegen Null und nach Proposition 28.1 konvergieren die Diagonaleinträge von A_k gegen die Eigenwerte von A. □

Aufwand. Pro Iterationsschritt sind $8n$ Multiplikationen notwendig. Hinreichend für $S(A_k) < \varepsilon$ ist nach (28.8) die Abschätzung

$$S(A_k) \;\le\; \big(1 - \frac{2}{(n-1)n}\big)^k S(A_0) \;\overset{!}{<}\; \varepsilon\,.$$

Wegen

$$\big(1 - \frac{2}{(n-1)n}\big)^k \;\approx\; 1 - \frac{2k}{(n-1)n}\,, \qquad k \ll n^2\,,$$

wird man bis zum Erreichen der vorgegebenen Schranke ε mindestens $k \sim n^2$ Iterationen erwarten. Dies entspricht einem Gesamtaufwand $O(n^3)$. ◊

Tatsächlich ist die Konvergenz der beiden angeführten Varianten des Jacobi-Verfahrens wesentlich schneller als (28.8) suggeriert:

Satz 28.4. *Die hermitesche Matrix $A \in \mathbb{K}^{n\times n}$ habe paarweise verschiedene Eigenwerte. Dann gilt sowohl bei der klassischen als auch bei der zyklischen Variante des Jacobi-Verfahrens, daß*

$$S(A_{k+N}) \le CS(A_k)^2$$

für $N = n(n-1)/2$ und hinreichend große k.

Beweis. Nach Voraussetzung existiert ein $\delta > 0$ mit

$$|\lambda_i - \lambda_j| > 2\delta \qquad \text{für} \quad \lambda_i \neq \lambda_j,\ \lambda_i, \lambda_j \in \sigma(A)\,,$$

und wir wollen im folgenden annehmen, daß $S(A_k)$ bereits kleiner als $\delta^2/4$ ist. Demnach ist $S(A_k) = \varepsilon^2\delta^2$ für ein $\varepsilon \in (0, 1/2)$ und die Einträge $a_{ij}^{(k)}$ von A_k sind durch

$$|a_{ij}^{(k)}| \le \varepsilon\delta\,, \qquad i \neq j\,, \tag{28.9}$$

beschränkt. Aus Proposition 28.1 folgt für beliebige $i \neq j$ und entsprechende Eigenwerte λ_i, λ_j, daß

$$\begin{aligned} |a_{ii}^{(k)} - a_{jj}^{(k)}| &= |a_{ii}^{(k)} - \lambda_i + \lambda_i - \lambda_j + \lambda_j - a_{jj}^{(k)}| \\ &\ge |\lambda_i - \lambda_j| - |\lambda_i - a_{ii}^{(k)}| - |\lambda_j - a_{jj}^{(k)}| \\ &> 2\delta - \delta/2 - \delta/2 \;=\; \delta\,. \end{aligned} \tag{28.10}$$

Betrachten wir nun die Givens-Rotation in der $(k+1)$-ten Iteration, so gilt wegen (28.3)

$$|s| \;=\; |\sin\theta_k| \;\le\; |\theta_k| \;=\; \frac{1}{2}\,|2\theta_k| \;\le\; \frac{1}{2}\,|\tan(2\theta_k)| \;=\; \Big|\,\frac{a_{ij}^{(k)}}{a_{jj}^{(k)} - a_{ii}^{(k)}}\,\Big|\,,$$

und mit (28.9) und (28.10) führt dies auf die Abschätzungen

$$|s| \le \varepsilon\,, \qquad 1 \ge c \ge (1-\varepsilon^2)^{1/2} \ge 1 - \varepsilon\,; \tag{28.11}$$

man beachte, daß $|\theta_k| \le \pi/4$, also c positiv ist.

Bei der Transformation $A_{k+1} = Q_k^* A_k Q_k$ sind nun für $i \neq j$ drei Fälle zu unterscheiden:

(a) $a_{ij}^{(k)} \;\mapsto\; a_{ij}^{(k+1)} = 0\,,$
$(b1)$ $a_{ij}^{(k)} \;\mapsto\; a_{ij}^{(k+1)} = a_{ij}^{(k)}\,,$

$(b2)$ $a_{ij}^{(k)}$ wird durch genau eine der beiden Givens-Rotationen mit einem anderen Nebendiagonalelement $a_{\mu\nu}^{(k)}$ verknüpft, vgl. Bemerkung 14.4: In diesem Fall ergibt sich

$$a_{ij}^{(k)} \;\mapsto a_{ij}^{(k+1)} \;=\; ca_{ij}^{(k)} \pm sa_{\mu\nu}^{(k)} \;=\; a_{ij}^{(k)} \;+\; \big((c-1)a_{ij}^{(k)} \pm sa_{\mu\nu}^{(k)}\big)\,.$$

In den letzten beiden Fällen $(b1)$ und $(b2)$ ergibt sich daher wegen (28.10) und (28.11) eine Änderung

$$|a_{ij}^{(k+1)} - a_{ij}^{(k)}| \;\le\; (|s| + |1-c|) \max_{\mu\neq\nu} |a_{\mu\nu}^{(k)}| \;\le\; 2\delta\varepsilon^2\,. \tag{28.12}$$

Sofern also das Nebendiagonalelement $a_{ij}^{(k_0)}$, $k_0 \ge k$, während der N Iterationsschritte

$$A_k \mapsto A_{k+1} \mapsto \ldots \mapsto A_{k+N}$$

auf Null transformiert wird, bleibt dessen Betrag in den nachfolgenden Iterationen wegen (28.12) unterhalb der Schranke $2N\delta\varepsilon^2$:

$$|a_{ij}^{(l)}| \;\le\; |a_{ij}^{(l)} - a_{ij}^{(l-1)}| \;+\; \ldots \;+\; |a_{ij}^{(k_0+1)} - a_{ij}^{(k_0)}| \;\le\; N2\delta\varepsilon^2\,, \quad k_0 \le l \le k+N\,.$$

In jedem Iterationsschritt werden nun (wegen der Symmetrie) zwei Nebendiagonalelemente gemäß (a) auf Null transformiert. Da es genau $2N$ Nebendiagonalelemente gibt, umfaßt die Anzahl der Nebendiagonalelemente, die größer als $2N\delta\varepsilon^2$ sind, nach der $(k+1)$-ten Iteration allenfalls $2N-2$ Elemente, nach der $(k+2)$-ten Iteration allenfalls $2N-4$ Elemente und so fort. Nach $k+N$ Iterationen sind alle Nebendiagonalelemente kleiner als $2N\delta\varepsilon^2$, d. h.

$$S(A_{k+N}) \;\le\; N(4N^2\delta^2\varepsilon^4) \;=\; \frac{4N^3}{\delta^2}\,S(A_k)^2\,. \qquad \square$$

Satz 28.4 besagt, daß unter den genannten Voraussetzungen bei Zyklen von je N Jacobi-Iterationen das Fehlermaß $S(A_{kN})$ quadratisch gegen Null konvergiert.

Bei der Implementierung der Givens-Rotationen müssen zunächst die Größen $c = \cos\theta$ und $s = \sin\theta$ zu dem Winkel θ aus (28.3) berechnet werden. Dazu beachte man, daß aus (28.5) mit $t = \tan(2\theta)$ zunächst

$$4s^2c^2 = t^2(c^2 - s^2)^2 \tag{28.13}$$

und dann mit $c^2 + s^2 = 1$

$$4c^2 - 4c^4 \;=\; t^2(4c^4 - 4c^2 + 1) \quad \text{und} \quad 4s^2 - 4s^4 \;=\; t^2(4s^4 - 4s^2 + 1)$$

folgt. Auflösen dieser (in c^2 und s^2) quadratischen Gleichungen ergibt

$$c^2 \;=\; \frac{1}{2} \,\pm\, \frac{1}{2}\Big(\frac{1}{t^2+1}\Big)^{1/2}, \qquad s^2 \;=\; \frac{1}{2} \,\pm\, \frac{1}{2}\Big(\frac{1}{t^2+1}\Big)^{1/2}.$$

Wegen $|\theta| \le \pi/4$ ist für c^2 das positive und für s^2 das negative Vorzeichen maßgeblich. Um Auslöschung zu vermeiden, berechnet man s^2 besser aus (28.13):

$$s^2 \;=\; \frac{t^2(2c^2-1)^2}{4c^2} \;=\; \frac{t^2}{4c^2(t^2+1)}\,.$$

Setzt man noch

$$r \;=\; \frac{1}{t} \overset{(28.5)}{=} \frac{a_{jj}-a_{ii}}{2a_{ij}}\,,$$

so erhält man schließlich die Formeln

$$c \;=\; \Big(\frac{1}{2} \,+\, \frac{1}{2}\Big(\frac{r^2}{1+r^2}\Big)^{1/2}\Big)^{1/2} \qquad \text{und} \qquad s \;=\; \frac{\operatorname{sign}(r)}{2c(1+r^2)^{1/2}}\,.$$

Ein praktischer Vergleich zwischen dem Aufwand des Jacobi- und des QR-Verfahren ergibt typischerweise eine Überlegenheit des QR-Verfahrens um einen Faktor zwischen 4 und 10. Dafür ist das Jacobi-Verfahren einfacher zu programmieren und zu parallelisieren.

Beispiel 28.5. Zur Illustration des klassischen Jacobi-Verfahrens betrachten wir die symmetrische Matrix

$$A \;=\; \big[\, a_{ij} \,\big] \;\in \mathbb{R}^{n\times n} \qquad \text{mit} \qquad a_{ij} = \pi^2\,\frac{i(n-j+1)}{(n+1)^3}\;, \quad i \le j\,,$$

deren Eigenwerte explizit bekannt sind:

$$\lambda_k \;=\; \frac{\pi^2}{4(n+1)^2\sin^2\big(k\pi/(2n+2)\big)}\,, \qquad k = 1,\ldots,n\,.$$

Abbildung 28.1 zeigt für dieses Beispiel mit $n = 100$ die Entwicklung des Konvergenzmaßes $\sqrt{S(A_k)}$ im Verlauf der Iteration (durchgeführt wurden $2N =$ 19 800 Iterationen). Nach diesen $2N$ Iterationen erreicht $\sqrt{S(A_k)}$ den Wert $1.1 \cdot 10^{-13}$ und der relative Fehler der berechneten Eigenwertnäherungen auf der Diagonalen von A_{2N} liegt durchweg zwischen $1.4 \cdot 10^{-16}$ und $2.4 \cdot 10^{-14}$.

Offensichtlich ist die Konvergenz über weite Strecken der Iteration lediglich linear. Erst ab etwa $k = 3N/2$ beziehungsweise $\sqrt{S(A_k)} \approx 10^{-7}$ macht sich

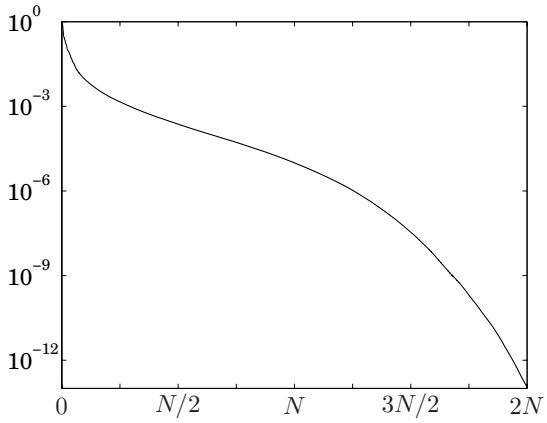


Abb. 28.1: Konvergenzverlauf von $\sqrt{S(A_k)}$

das Einsetzen quadratischer Konvergenz bemerkbar. Dies liegt an dem kleinen Wert $\delta \approx 8.783 \cdot 10^{-8}$ aus (28.10), denn die Eigenwerte λ_k von A liegen für große k und n sehr nah beieinander.

Interessant ist auch die Entwicklung der Nebendiagonalelemente, die in Abbildung 28.2 nachvollzogen werden kann. Dort sind die Matrizen A_k nach je $N/4$ Iterationen in logarithmischen Grauwertstufen dargestellt, die mit kleiner werdenden Einträgen immer heller werden. ◊

29 Spezielle Verfahren für hermitesche Tridiagonalmatrizen

Als nächstes betrachten wir hermitesche Tridiagonalmatrizen und stellen für die Berechnung aller Eigenwerte und Eigenvektoren solcher Matrizen eine Alternative zu dem QR-Verfahren vor. Hierzu kann ohne Beschränkung der Allgemeinheit angenommen werden, daß die Nebendiagonalelemente der Matrix reell und nichtnegativ sind; dies ergibt sich aus Aufgabe 16. Daher setzen wir im weiteren voraus, daß A die Gestalt

$$A = \begin{bmatrix} \alpha_1 & \beta_1 & & 0 \\ \beta_1 & \alpha_2 & \ddots & \\ & \ddots & \ddots & \beta_{n-1} \\ 0 & & \beta_{n-1} & \alpha_n \end{bmatrix} \tag{29.1}$$

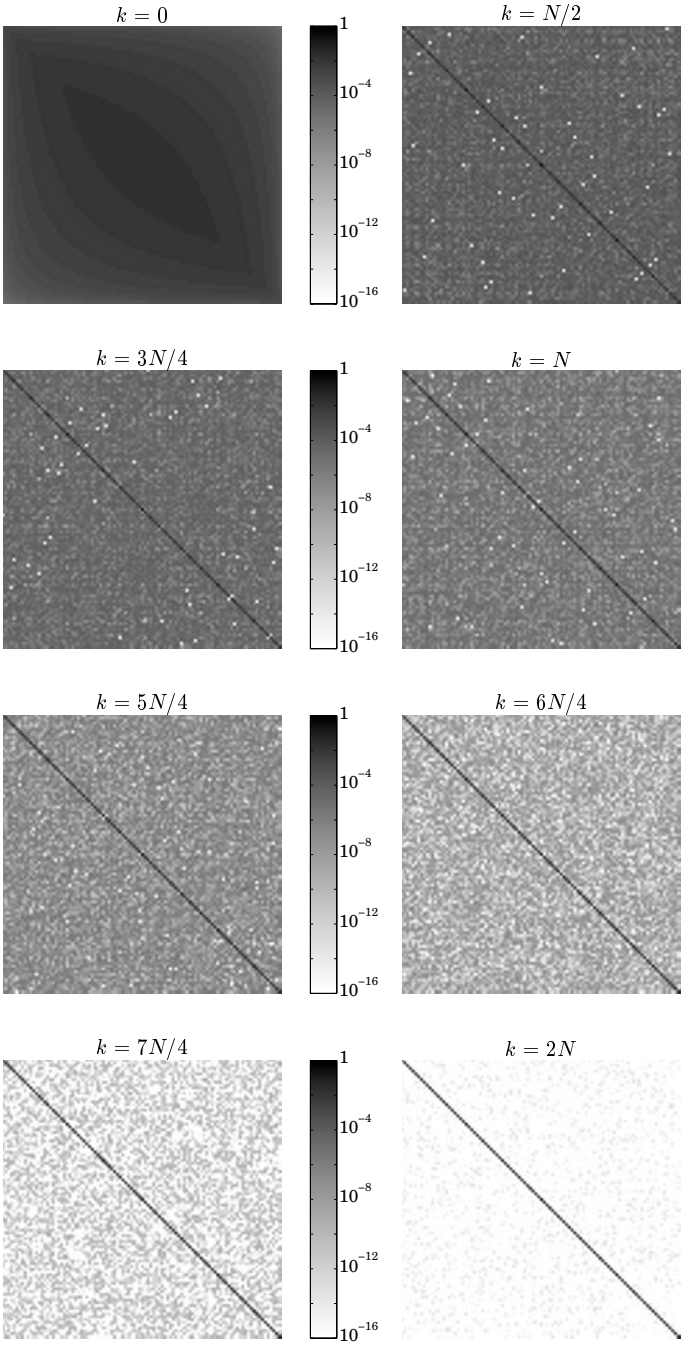


Abb. 28.2: Iterierte A_k des Jacobi-Verfahrens

besitzt, wobei $\alpha_i \in \mathbb{R}$, $i = 1, \ldots, n$, und $\beta_i \geq 0$, $i = 1, \ldots, n-1$.

Wie beim QR-Verfahren kann der Aufwand dieses Verfahrens mit $O(n^2)$ Operationen abgeschätzt werden; laut Demmel [22] ist der Aufwand in der Praxis bereits für moderate n dem QR-Verfahren überlegen. Anschließend wird in Abschnitt 29.2 noch auf eine stabile Berechnung der zugehörigen Eigenvektoren eingegangen.

An dieser Stelle sei auch noch auf das Bisektionsverfahren in Abschnitt 35.2 verwiesen: Das Bisektionsverfahren ist das effizienteste Verfahren für hermitesche Tridiagonalmatrizen, wenn nur die Eigenwerte aus einem bestimmten Teilintervall gesucht werden.

29.1 Ein Divide-and-Conquer-Verfahren

Unter Divide-and-Conquer-Verfahren versteht man Algorithmen, die ein gegebenes Problem in gleichartige Teilprobleme kleinerer Dimension zerlegen (die dann wiederum zerlegt werden können etc.). Im vorliegenden Fall wird ein Index $m \in \{2, \ldots, n\}$ gewählt und A in

$$A = \left[\begin{array}{ccccc:ccccc} \alpha_1 & \beta_1 & & & & & & & & \\ \beta_1 & \ddots & \ddots & & & & & & & \\ & \ddots & \ddots & \beta_{m-1} & & & & & & \\ & & \beta_{m-1} & \alpha_m & \beta_m & & & & & \\ \hdashline & & & \beta_m & \alpha_{m+1} & \beta_{m+1} & & & & \\ & & & & \beta_{m+1} & \ddots & \ddots & & & \\ & & & & & \ddots & \ddots & \beta_{n-1} & & \\ & & & & & & \beta_{n-1} & \alpha_n \end{array}\right] = \left[\begin{array}{ccccc:ccccc} A_1 & & & & & & & & & \\ & & & & & & & & & \\ & & & & & & & & & \\ & & & & & & & & & \\ \hdashline & & & & & & & & & \\ & & & & & & A_2 & & & \\ & & & & & & & & & \\ & & & & & & & & & \\ & & & & & & & & & \end{array}\right] + \left[\begin{array}{ccccc:ccccc} & & & & & & & & & \\ & & & & & & & & & \\ & & & & & & & & & \\ & & & & & \beta_m & \beta_m & & & \\ \hdashline & & & & & \beta_m & \beta_m & & & \\ & & & & & & & & & \\ & & & & & & & & & \\ & & & & & & & & & \end{array}\right]$$

additiv zerlegt mit

$$A_1 = \begin{bmatrix} \alpha_1 & \beta_1 & & \\ \beta_1 & \alpha_2 & \ddots & \\ & \ddots & \ddots & \beta_{m-1} \\ & & \beta_{m-1} & \gamma_{m-1} \end{bmatrix} \quad \text{und} \quad A_2 = \begin{bmatrix} \gamma_m & \beta_{m+1} & & \\ \beta_{m+1} & \alpha_{m+2} & \ddots & \\ & \ddots & \ddots & \beta_{n-1} \\ & & \beta_{n-1} & \alpha_n \end{bmatrix},$$

wobei $\gamma_{m-1} = \alpha_m - \beta_m$ und $\gamma_m = \alpha_{m+1} - \beta_m$ ist. Mit Hilfe des Vektors $w = e_m + e_{m+1} \in \mathbb{R}^n$ kann diese Matrixzerlegung auch in der Form

$$A = \begin{bmatrix} A_1 & 0 \\ 0 & A_2 \end{bmatrix} + \beta_m w w^* \tag{29.2}$$

geschrieben werden.

Der Divide-and-Conquer-Philosophie entsprechend wird im weiteren angenommen, daß alle Eigenwerte und Eigenvektoren von A_1 und A_2 bereits berechnet sind: Es seien also Faktorisierungen

$$A_1 = V_1 \Lambda_1 V_1^* , \qquad A_2 = V_2 \Lambda_2 V_2^*$$

von A_1 und A_2 gegeben, wobei die Diagonalmatrizen Λ_1 und Λ_2 die jeweiligen Eigenwerte und die Spalten von V_1 und V_2 die zugehörigen orthonormierten Eigenvektoren enthalten. Dann ist

$$\begin{bmatrix} A_1 & 0 \\ 0 & A_2 \end{bmatrix} = V \Lambda V^* \qquad \text{mit} \quad V = \begin{bmatrix} V_1 & 0 \\ 0 & V_2 \end{bmatrix} , \quad \Lambda = \begin{bmatrix} \Lambda_1 & 0 \\ 0 & \Lambda_2 \end{bmatrix}$$

die Spektralzerlegung der Blockdiagonalmatrix und aus (29.2) folgt die Gleichung

$$V^* A V = \Lambda + \beta_m z z^* , \qquad z = V^* w = \begin{bmatrix} v_1 \\ v_2 \end{bmatrix} \tag{29.3}$$

mit $v_1 \in \mathbb{R}^n$ und $v_2 \in \mathbb{R}^{n-m}$. Wegen $V^* w = V^* e_m + V^* e_{m+1}$ ist dabei v_1 die hinterste Spalte von V_1^* und v_2 die erste Spalte von V_2^*. Offensichtlich ist $V^* A V$ eine Ähnlichkeitstransformation von A und somit ist das gesuchte Spektrum von A identisch mit dem Spektrum der Matrix $\Lambda + \beta_m z z^*$ in (29.3).

Die Berechnung der Eigenwerte von $\Lambda + \beta_m z z^*$ beruht schließlich auf der folgenden Proposition.

Proposition 29.1. *Sei $D = [d_{ij}] \in \mathbb{R}^{n \times n}$ eine Diagonalmatrix mit paarweise verschiedenen Einträgen $d_i = d_{ii}$ und $z = [z_1, \ldots, z_n]^T \in \mathbb{R}^n$ ein Vektor, dessen Einträge alle von Null verschieden seien. Dann sind die Eigenwerte von $D + \beta z z^*$, $\beta \neq 0$, gerade die n Nullstellen der rationalen Funktion*

$$f(\lambda) \;=\; 1 + \beta \sum_{i=1}^{n} \frac{z_i^2}{d_i - \lambda} \; . \tag{29.4}$$

Beweis. Sei $\beta \neq 0$ und $x \neq 0$ ein Eigenvektor von $D + \beta z z^*$ zum Eigenwert λ. Dann folgt

$$(D - \lambda I)x \;=\; (D + \beta z z^*)x - \beta z z^* x - \lambda x \;=\; -\beta (z^* x) z \, . \tag{29.5}$$

Wir zeigen zunächst, daß λ kein Eigenwert von D ist. Da D eine Diagonalmatrix ist, wäre ansonsten einer der kartesischen Basisvektoren Eigenvektor, etwa e_j, und durch Multiplikation von (29.5) mit e_j von links folgt

$$0 \;=\; -\beta (z^* x)(e_j^* z) \, .$$

Nach Voraussetzung sind β und $e_j^* z = z_j$ von Null verschieden, so daß in diesem Fall $z^* x = 0$ gelten muß. Eingesetzt in (29.5) ergeben sich hieraus die beiden Gleichungen

$$(D - \lambda I)x \;=\; 0 \qquad \text{und} \qquad z^* x \;=\; 0\,.$$

Aus der ersten Gleichung folgt, daß alle Einträge x_i von x mit $i \neq j$ gleich Null sein müssen, da die entsprechenden Diagonaleinträge von $D - \lambda I$ von Null verschieden sind (sonst wären zwei Diagonaleinträge von D gleich, was durch die Voraussetzungen ausgeschlossen ist). Damit ist aber $x \neq 0$ ein Vielfaches von e_j und die zweite Gleichung $z^* x = 0$ steht im Widerspruch zu der Voraussetzung $z_j \neq 0$.

Die Annahme $\lambda \in \sigma(D)$ ist folglich zum Widerspruch geführt worden, d. h. $D - \lambda I$ ist invertierbar. Aus (29.5) ergibt sich somit

$$x = -\beta(z^* x)(D - \lambda I)^{-1} z\,, \tag{29.6}$$

und durch Multiplikation von links mit z erhalten wir

$$z^* x \;=\; -\beta(z^* x) z^* (D - \lambda I)^{-1} z\,.$$

Aus (29.5) folgt weiterhin $z^* x \neq 0$, denn ansonsten wäre $(D - \lambda I)x = 0$ und λ doch ein Eigenwert von D. Also kann abschließend durch $z^* x$ gekürzt werden, und die Behauptung ist bewiesen. □

Bemerkungen. 1. Proposition 29.1 läßt sich relativ leicht verallgemeinern, wenn eine der getroffenen Voraussetzungen an d_i und z_i, $i = 1, \ldots, n$, nicht erfüllt ist. Diese Verallgemeinerungen laufen darauf hinaus, daß entsprechende Eigenwerte von D auch Eigenwerte von $D + \beta z z^*$ sind: Für $z_i = 0$ ist d_i ein Eigenwert von $D + \beta z z^*$ mit Eigenvektor e_i; auch wenn $d_i = d_j$ für ein $j \neq i$ gilt, ist d_i ein Eigenwert, diesmal mit Eigenvektor $x = z_j e_i - z_i e_j$. Die restlichen Eigenwerte ergeben sich nach wie vor als Nullstellen der Funktion f aus (29.4).

2. Aus dem Beweis von Proposition 29.1 können auch die Eigenvektoren von $D + \beta z z^*$ abgelesen werden: Ist λ ein Eigenwert von $D + \beta z z^*$, dann ist $x = (D - \lambda I)^{-1} z$ ein zugehöriger Eigenvektor, vgl. (29.6). Nach (29.3) ist schließlich Vx ein entsprechender Eigenvektor von A. Diese Vorgehensweise zur Berechnung der Eigenvektoren erweist sich allerdings als nicht ausreichend stabil. ◊

Beispiel. Gegeben sei die Tridiagonalmatrix

$$A \;=\; \begin{bmatrix} 1 & 3 & & \\ 3 & -6 & 1 & \\ & 1 & 3 & 1 \\ & & 1 & 2 \end{bmatrix} \;=\; \left[\begin{array}{cc:cc} 1 & 3 & & \\ 3 & -7 & & \\ \hdashline & & 2 & 1 \\ & & 1 & 2 \end{array}\right] + \left[\begin{array}{cc:cc} & & & \\ & 1 & 1 & \\ \hdashline & 1 & 1 & \\ & & & \end{array}\right].$$

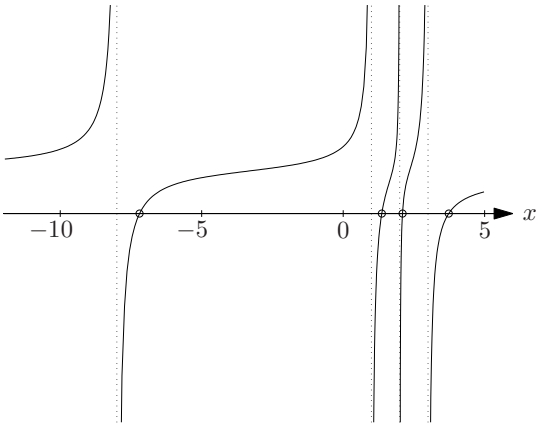


Abb. 29.1: Die rationale Funktion f

Die Spektralzerlegungen von A_1 und A_2 sind einfach zu berechnen: A_1 hat die Eigenwerte $d_1 = -8$ und $d_2 = 2$ mit Eigenvektoren $[1,-3]^T$ und $[3,1]^T$, A_2 die Eigenwerte $d_3 = 1$ und $d_4 = 3$ mit Eigenvektoren $[1,-1]^T$ und $[1,1]^T$. Durch Normierung führt dies auf die Matrizen

$$V_1 = \frac{1}{\sqrt{10}} \begin{bmatrix} 1 & 3 \\ -3 & 1 \end{bmatrix} \qquad \text{und} \qquad V_2 = \frac{1}{\sqrt{2}} \begin{bmatrix} 1 & 1 \\ -1 & 1 \end{bmatrix}.$$

Aus (29.3) ergibt sich dann $z = [-3, 1, \sqrt{5}, \sqrt{5}]^T/\sqrt{10}$ und die rationale Funktion f aus (29.4) lautet

$$f(\lambda) = 1 - \frac{9/10}{\lambda+8} - \frac{1/2}{\lambda-1} - \frac{1/10}{\lambda-2} - \frac{1/2}{\lambda-3}.$$

Diese Funktion ist in Abbildung 29.1 geplottet. Ihre Nullstellen sind durch Kreise markiert: Dies sind die Eigenwerte von A. ◊

Für effiziente numerische Verfahren zur Lösung der *charakteristischen Gleichung* $f(\lambda) = 0$ sei an dieser Stelle lediglich auf den Hinweis in Beispiel 18.6 verwiesen. Algorithmus 29.1 beschreibt eine Implementierung des oben beschriebenen Verfahrens für den Fall, daß n eine Zweierpotenz ist. Diese Voraussetzung ist allerdings nicht wesentlich, sie erleichtert lediglich die Darstellung, da dann in jedem Schritt A_1 und A_2 gleich groß gewählt werden können.

Aufwand. Für die Aufwandsabschätzung dieses Algorithmus wird verwendet, daß jede einzelne Nullstelle der rationalen Funktion f in einer geringen Zahl von Iterationsschritten auf maximale Genauigkeit berechnet werden kann. In jedem Iterationsschritt müssen dabei (im wesentlichen) die Funktion f und ihre

```
function [V, λ] = divide_and_conquer (A, n)
  % berechnet Spektralzerlegung von A: V enthält die Eigenvektoren,
  % λ die Eigenwerte in entsprechender Reihenfolge
  if n = 1 then
    λ = A
    V = 1
  else
    m = n/2
    zerlege A wie in (29.2)
    [V1, λ1] = divide_and_conquer (A1, m)
    [V2, λ2] = divide_and_conquer (A2, m)
    % v1 sei die hinterste Spalte von V1*, v2 sei die vorderste Spalte von V2*
    d = [λ1; λ2],   z = [v1; v2]
    % f sei die rationale Funktion aus (29.4)
    λ = roots (f)      % berechnet alle n Nullstellen von f, vgl. Beispiel 18.6
    ggf. Sonderbehandlung, falls die Voraussetzungen von Proposition 29.1 verletzt sind
    bestimme V durch gebrochene Iteration, vgl. Algorithmus 29.2
  end if
end      % divide_and_conquer

Initialisierung: A sei n × n-Tridiagonalmatrix mit n = 2^p
  [V, λ] = divide_and_conquer (A, n)
```

Algorithmus 29.1: Divide-and-Conquer-Verfahren

Ableitung ausgewertet werden; dies benötigt $O(n)$ Multiplikationen und Divisionen. Hinzu kommen weitere $O(n)$ Multiplikationen für die Bestimmung eines zugehörigen Eigenvektors (vgl. etwa Algorithmus 29.2 im folgenden Abschnitt). Die Berechnung aller $n = 2^p$ Eigenwerte von A – bei Kenntniss der Spektralzerlegungen von A_1 und A_2 – benötigt daher einen Aufwand $b_p = O(n^2) = O(4^p)$.

Geht man bei der Implementierung rekursiv vor, berechnet also auch die Spektralzerlegungen von A_1 und A_2 mit dem Divide-and-Conquer-Verfahren, dann errechnet sich eine obere Schranke a_p für den Gesamtaufwand zur Berechnung aller Eigenwerte und Eigenvektoren einer Matrix $A \in \mathbb{R}^{n\times n}$ mit $n = 2^p$ aus der Rekursion

$$a_p = 2a_{p-1} + b_p = 2a_{p-1} + O(4^p)\,.$$

Somit existiert ein $c > 0$, so daß $a_p \le 2a_{p-1} + c4^p$ ist, und durch Auflösen der

Rekursion ergibt sich

$$a_p \leq c4^p + 2c4^{p-1} + 4c4^{p-2} + \ldots = c4^p \sum_{j=0}^{p-1} 2^{-j} = O(n^2)\,.$$

$\Diamond$

29.2 Berechnung der Eigenvektoren

Im weiteren sei λ die Approximation eines Eigenwerts der symmetrischen Tridiagonalmatrix A aus (29.1) und $x = [x_1, \ldots, x_n]^T$ der gesuchte Eigenvektor. Zur Vereinfachung der Darstellung wollen wir annehmen, daß A irreduzibel ist, ansonsten tritt Deflation auf und die Eigenvektoren von A können unmittelbar aus den Eigenvektoren der entsprechenden Teilmatrizen A_1 und A_2 zusammengesetzt werden, vgl. (29.3). Wir beginnen mit dem folgenden Lemma:

Lemma 29.2. *Sei $A \in \mathbb{R}^{n\times n}$ eine symmetrische irreduzible Tridiagonalmatrix der Gestalt* (29.1)*. Ist x ein zugehöriger Eigenvektor, dann sind dessen erste und letzte Komponente von Null verschieden.*

Beweis. Wir schreiben die ersten $n-1$ Zeilen der Eigenwertgleichung $Ax = \lambda x$ mit $x = [x_1, \ldots, x_n]^T \neq 0$ in der Gestalt

$$\begin{aligned}(\alpha_1 - \lambda)x_1 + \beta_1 x_2 &= 0\,,\\ \beta_1 x_1 + (\alpha_2 - \lambda)x_2 + \beta_2 x_3 &= 0\,,\\ &\vdots\\ \beta_{n-2}x_{n-2} + (\alpha_{n-1} - \lambda)x_{n-1} + \beta_{n-1}x_n &= 0\,.\end{aligned}$$

Wäre nun $x_1 = 0$, so ergibt sich durch Vorwärtssubstitution zunächst aus der ersten Gleichung wegen $\beta_1 \neq 0$ unmittelbar $x_2 = 0$. Aus der zweiten Gleichung folgt dann $x_3 = 0$, und so fortfahrend ergibt sich $x_4 = \ldots = x_n = 0$ im Widerspruch zu der Voraussetzung $x \neq 0$. Durch eine entsprechende Argumentation mit Rückwärtssubstitution von x aus den letzten $n-1$ Gleichungen sieht man, daß $x_n \neq 0$ gilt. $\square$

Wir können also im weiteren davon ausgehen, daß $x_1 \neq 0$ ist. Da multiplikative Faktoren bei der Bestimmung eines Eigenvektors keine Rolle spielen, bietet es sich daher an, $x_1 = 1$ zu setzen und dann, ähnlich zu dem obigen Beweis die Zeilen $i = 1$ bis $i = n-1$ des Gleichungssystems $Ax = \lambda x$ zu durchlaufen, um die restlichen Einträge von x durch Vorwärtssubstitution zu berechnen:

$x_1 = 1$
$x_2 = (\lambda - \alpha_1) \,/\, \beta_1$
`for` $i = 2, 3, \ldots, n-1$ `do`
  $x_{i+1} = \big((\lambda - \alpha_i)x_i - \beta_{i-1}x_{i-1}\big) \,/\, \beta_i$
`end for`
$x^+ = [x_1, \ldots, x_n]^T$

Tatsächlich ist diese Rekursion instabil und für die Praxis nicht zu empfehlen. Wir wollen jedoch hierauf (zunächst) nicht eingehen und nehmen an, der obige Algorithmus würde in exakter Arithmetik ausgeführt. Selbst dann kann der berechnete Vektor unter Umständen ein völlig unbrauchbares Ergebnis darstellen:

Beispiel 29.3. Gegeben sei die 3×3-Tridiagonalmatrix

$$A = \begin{bmatrix} 0 & \varepsilon & 0 \\ \varepsilon & 1 & \varepsilon^2 \\ 0 & \varepsilon^2 & \varepsilon \end{bmatrix},$$

wobei $\varepsilon > 0$ ein kleiner Parameter sei. Für $\varepsilon \to 0$ ist

$$\widehat{\lambda} = 1 + \varepsilon^2 + O(\varepsilon^3), \qquad \widehat{x} = \begin{bmatrix} \varepsilon \\ 1 \\ \varepsilon^2 \end{bmatrix} + O(\varepsilon^3),$$

ein Eigenpaar von A. (Die anderen Eigenwerte sind von der Größenordnung $O(\varepsilon)$ und im weiteren ohne Bedeutung.) Mit dem obigen Algorithmus ergibt sich für die Näherung $\lambda = 1$ mit exakter Arithmetik der Näherungsvektor

$$x^+ = \begin{bmatrix} 1 \\ 1/\varepsilon \\ -1/\varepsilon \end{bmatrix} = 1/\varepsilon \begin{bmatrix} \varepsilon \\ 1 \\ -1 \end{bmatrix},$$

dessen letzte Komponente offensichtlich völlig falsch ist. ◊

Um die Ursache für diese schlechte Approximation zu verstehen, muß man beachten, daß aufgrund der Schleifenkonstruktion in dem Algorithmus der berechnete Vektor x^+ nur die *oberen* $n-1$ Gleichungen des linearen Gleichungssystems $Ax = \lambda x$ erfüllt. Die letzte Gleichung ist in der Regel hingegen nicht erfüllt – sie ist lediglich dann erfüllt, wenn λ ein exakter Eigenwert von A ist. Mit anderen Worten: im allgemeinen ist

$$Ax^+ - \lambda x^+ = \gamma_n e_n \tag{29.7}$$

mit $\gamma_n \neq 0$.

Man könnte nun aber auch $x_n = 1$ setzen, und dann die Gleichungen $Ax = \lambda x$ von unten nach oben durchlaufen, um die restlichen Komponenten von x durch Rückwärtssubstitution zu bestimmen:

$x_n = 1$
$x_{n-1} = (\lambda - \alpha_n) \,/\, \beta_{n-1}$
`for` $i = n-1, n-2, \ldots, 2$ `do`
  $x_{i-1} = \big((\lambda - \alpha_i) x_i - \beta_i x_{i+1}\big) \,/\, \beta_{i-1}$
`end for`
$x^- = [x_1, \ldots, x_n]^T$

In exakter Arithmetik (dieser Algorithmus ist ansonsten ebenfalls nicht zu empfehlen) ergibt dies einen Vektor x^-, der die *unteren* $n-1$ Gleichungen des Gleichungssystems $Ax = \lambda x$ erfüllt. Die erste Gleichung ist nur dann erfüllt, wenn λ ein Eigenwert von A ist. Folglich ist

$$Ax^- - \lambda x^- = \gamma_1 e_1 \,, \tag{29.8}$$

wobei γ_1 in der Regel von Null verschieden ist.

Beispiel. In Beispiel 29.3 ergibt sich mit $\lambda = 1$ bei dieser alternativen Vorgehensweise der Vektor

$$x^- \;=\; \begin{bmatrix} -\varepsilon \\ 1/\varepsilon^2 - 1/\varepsilon \\ 1 \end{bmatrix} \;=\; 1/\varepsilon^2 \begin{bmatrix} -\varepsilon^3 \\ 1-\varepsilon \\ \varepsilon^2 \end{bmatrix} .$$

Somit ist zwar $\varepsilon^2 x^- = \widehat{x} + O(\varepsilon)$, also keine völlig unbrauchbare Näherung an den gesuchten Eigenvektor, aber das Vorzeichen und die Größenordnung der ersten Komponente von x^- ist falsch. Bemerkenswerterweise sind γ_1 und γ_3 gleich, nämlich

$$\gamma_1 = \gamma_3 = 1/\varepsilon - 1 + \varepsilon \,.$$

Letzteres ist bei dieser Vorgehensweise immer der Fall, vgl. Aufgabe 17. ◊

Behalten wir die Annahme bei, daß λ selbst kein Eigenwert von A sondern nur eine gute Näherung an einen solchen ist, so können die beiden Näherungen x^+ und x^- aus (29.7) und (29.8) als das Ergebnis der gebrochenen Iteration von Wielandt mit Startvektor e_n bzw. e_1 interpretiert werden. Die Güte dieser Näherungen hängt davon ab, inwieweit der jeweilige Startvektor eine relevante Komponente in Richtung des gesuchten Eigenvektors besitzt, vgl. Bemerkung 25.2. In Beispiel 29.3 liegt weder für den Startvektor e_1 noch für e_3 eine signifikante Komponente vor, allerdings ist aus dieser Überlegung heraus e_1 zumindest noch etwas besser geeignet als e_3, da die erste Komponente von $\widehat{x}$ deutlich größer als die dritte ist.

Satz 24.3 liefert eine andere Möglichkeit zur Interpretation von (29.7) und (29.8): Demnach sind nicht $|\gamma_1|$ und $|\gamma_n|$ die relevanten Gütemaße für die Näherungsvektoren sondern die entsprechenden *relativen Größen* (in Satz 24.3 mit ε bezeichnet), die sich in Beispiel 29.3 folgendermaßen verhalten:

$$\frac{|\gamma_3|}{\|x^+\|_2} \;\sim\; 1/\sqrt{2} \,, \quad \frac{|\gamma_1|}{\|x^-\|_2} \;\sim\; \varepsilon \,, \qquad \varepsilon \to 0 \,.$$

Dies bestätigt wiederum die Überlegenheit von x^- in diesem Beispiel.

Im dem betrachteten Beispiel sind weder x^+ noch x^- zufriedenstellende Näherungen, da weder e_1 noch e_3 einen signifikanten Anteil in die Richtung des gesuchten Eigenvektors besitzen. Wie wir gleich sehen werden, ist es jedoch ohne allzu großen Zusatzaufwand möglich, aus der Kenntnis von x^+ und x^- für *jeden* kartesischen Startvektor e_i die zugehörige Näherung $x^{(i)}$ der gebrochenen Iteration von Wielandt zu bestimmen, also die Lösungen $x^{(i)}$ von

$$(A - \lambda I)x^{(i)} = e_i\,, \qquad i = 1, \ldots, n\,. \tag{29.9}$$

Nach Satz 24.3 ist dann diejenige Näherung $x^{(i)}$ am verläßlichsten, für die $\|x^{(i)}\|_2$ maximal ist.

Satz 29.4. *$T \in \mathbb{R}^{n\times n}$ sei eine symmetrische, invertierbare und irreduzible Tridiagonalmatrix. Dann existieren Vektoren $u = [u_i]$, $v = [v_i] \in \mathbb{R}^n$, so daß T^{-1} die Gestalt*

$$T^{-1} = \begin{bmatrix} u_1v_1 & u_2v_1 & u_3v_1 & \ldots & u_nv_1 \\ u_2v_1 & u_2v_2 & u_3v_2 & \ldots & u_nv_2 \\ u_3v_1 & u_3v_2 & u_3v_3 & & u_nv_3 \\ \vdots & \vdots & & \ddots & \vdots \\ u_nv_1 & u_nv_2 & u_nv_3 & \ldots & u_nv_n \end{bmatrix}$$

besitzt.

Beweis. Wir setzen zunächst $u = T^{-1}e_1 = [u_i]_{i=1}^n$. Nach Lemma 29.2 ist dann $u_n \neq 0$ und wir setzen $v = T^{-1}e_n/u_n$. u und u_nv sind dann gerade die erste bzw. die letzte Spalte von T^{-1}. Wegen der Symmetrie von T ist auch T^{-1} symmetrisch und somit gilt $u_n = u_nv_1$. Wegen $u_n \neq 0$ folgt hieraus $v_1 = 1$. Damit haben die erste und die letzte Spalte von T^{-1} die gewünschte Form.

Für $i \in \{2, \ldots, n-1\}$ betrachten wir nun

$$w = [u_iv_1, \ldots, u_iv_i, u_{i+1}v_i, \ldots, u_nv_i]^T \in \mathbb{R}^n\,.$$

Offensichtlich stimmen die ersten i Komponenten von w und u_iv sowie die letzten $n - i + 1$ Komponenten von w und v_iu überein. Da T eine Tridiagonalmatrix ist, sind somit die ersten $i - 1$ Komponenten von Tw und von u_iTv sowie die letzten $n - i$ Komponenten von Tw und v_iTu jeweils gleich. Wegen $Tv = e_n/u_n$ und $Tu = e_1$ sind diese Komponenten allesamt Null. Damit ist höchstens die i-te Komponente von Tw von Null verschieden, d. h. es existiert ein $\omega \in \mathbb{R}$ mit

$$Tw = \omega\, e_i\,.$$

Falls $\omega \neq 0$ ist, ergibt sich hieraus die i-te Spalte von T^{-1} als w/ω. Wegen der Symmetrie von T^{-1} ist die erste Komponente dieser Spalte gleich u_i, und aus der Definition von w ergibt sich daher die Bedingung

$$u_i \stackrel{!}{=} u_i v_1/\omega = u_i/\omega\,.$$

Somit ist $\omega = 1$ und die i-te Spalte von T^{-1} stimmt wie behauptet mit w überein.

Also bleibt nur noch der Fall $\omega = 0$ zu untersuchen. In diesem Fall ist $Tw = 0$ und da T invertierbar ist, muß $w = 0$ sein. Wegen $v_1 = 1$ und $u_n \neq 0$ ergibt sich daraus $u_i = v_i = 0$. Folglich sind die beiden i-ten Komponenten der ersten und letzten Spalte von T^{-1} gleich Null. Wegen der Symmetrie von T^{-1} überträgt sich das auf die erste und die letzte Komponente des i-ten Spaltenvektors x von T^{-1}, also der Lösung des Gleichungssystems

$$Tx = e_i\,.$$

Durch Vorwärtssubstitution aus den ersten $i-1$ Gleichungen dieses Systems und durch Rückwärtssubstitution aus den letzten $n-i$ Gleichungen ergibt sich dann aber, daß $x = 0$ ist, also ein Widerspruch. Somit kann der Fall $\omega = 0$ nicht auftreten, und der Beweis ist abgeschlossen. □

Die gesuchten Vektoren $x^{(i)}$, $i = 1, \ldots, n$, sind nach (29.9) gerade die Spalten der Inversen von $A - \lambda I$. Nach Satz 29.4 können alle Spalten aus den beiden Vektoren

$$u = x^{(1)} \qquad \text{und} \qquad v = x^{(n)}/u_n$$

konstruiert werden. Anstelle der instabilen Rekursion zur Berechnung der Vektoren $x^{\pm}$, sollte jedoch für $x^{(1)}$ und $x^{(n)}$ ein stabiler Algorithmus zur Lösung der linearen Gleichungssysteme (29.9) verwendet werden, etwa mit Hilfe einer QR-Zerlegung von A durch Givens-Rotationen, vgl. Abschnitt 27.2.

Aufgrund von Satz 29.4 ist es dann möglich, mit nur $O(n)$ Operationen alle Normen $\|x^{(i)}\|_2$ zu berechnen und eine Spalte mit maximaler Norm auszuwählen. Es ist nämlich

$$\|x^{(i)}\|_2^2 \;=\; \sum_{j=1}^{i}(u_i v_j)^2 + \sum_{j=i+1}^{n}(u_j v_i)^2 \;=\; (u_i v_i)^2\Big(\sum_{j=1}^{i}\big(\frac{v_j}{v_i}\big)^2 + \sum_{j=i+1}^{n}\big(\frac{u_j}{u_i}\big)^2\Big)\,,$$

also

$$\|x^{(i)}\|_2^2 \;=\; (u_i v_i)^2(a_i + b_i)\,, \qquad i = 1, \ldots, n\,, \tag{29.10}$$

Initialisierung: $\lambda \notin \sigma(A)$ approximiere einen Eigenwert von A aus (29.1)
  faktorisiere $A - \lambda I = QR$ % QR-Faktorisierung, vgl. Abschnitt 27.2
  löse $Ru = Q^* e_1$ % Rückwärtssubstitution; $u = [u_1, \ldots, u_n]^T$
  löse $Rv = Q^* e_n / u_n$ % Rückwärtssubstitution; $v = [v_1, \ldots, v_n]^T$
  $a_1 = 1, \quad b_n = 0$
  for $i = 1, \ldots, n-1$ **do** % Rekursion für (29.11)
    $a_{i+1} = 1 + (v_i / v_{i+1})^2 a_i$
    $b_{n-i} = (u_{n-i+1} / u_{n-i})^2 (1 + b_{n-i+1})$
  end for
  % k sei ein Index, für den $(u_i v_i)^2 (a_i + b_i)$ maximal ist, vgl. (29.10)
  $x^{(k)} = [u_k v_1, \ldots, u_k v_k, u_{k+1} v_k, \ldots, u_n v_k]^T$
Ergebnis: $x^{(k)}$ ist eine Approximation an den gesuchten Eigenvektor

Algorithmus 29.2: Approximation eines Eigenvektors

mit den Hilfsgrößen

$$a_i = \sum_{j=1}^{i} \left(\frac{v_j}{v_i}\right)^2 \qquad \text{und} \qquad b_i = \sum_{j=i+1}^{n} \left(\frac{u_j}{u_i}\right)^2 . \tag{29.11}$$

Diese Hilfsgrößen können rekursiv aufsummiert werden, vgl. Algorithmus 29.2.

Aufwand. Die QR-Faktorisierung von $A - \lambda I$ kostet etwa $13n$ multiplikative Operationen, vgl. Aufgabe 14. Die resultierende obere Dreiecksmatrix R besitzt drei von Null verschiedene Diagonalen, d. h. der Aufwand zur Lösung eines Gleichungssystems mit R beträgt etwa $2n$ Multiplikationen und n Divisionen. Für die zwei Gleichungssysteme müssen zuvor die rechten Seiten mit den Givens-Rotationen multipliziert werden. Bei der rechten Seite e_n werden nur zwei Multiplikationen benötigt, bei e_1 hingegen weitere $2n$ Multiplikationen. Die Berechnung aller Hilfsgrößen a_i und b_i, $i = 1, \ldots, n-1$, kostet $4n$ Multiplikationen sowie $2n$ Divisionen, und weitere $2n$ Multiplikationen sind für die Suche nach der Spalte $x^{(k)}$ mit maximaler Norm nötig, wenn die Quadrate u_i^2, v_i^2, $i = 1, \ldots, n$, zwischengespeichert wurden. Der Aufwand für die explizite Berechnung von $x^{(k)}$ beträgt abschließend noch einmal n Multiplikationen. Insgesamt ergeben sich somit etwa $30n$ Multiplikationen/Divisionen, um den gewünschten Eigenvektor auszurechnen. ◊

Dieser Aufwand kann noch reduziert werden, wenn anstelle der QR-Zerlegung eine sogenannte LDL^*-Faktorisierung (ähnlich der LR-Zerlegung) von $A - \lambda I$ berechnet wird. Details finden sich etwa in der Arbeit von Parlett und Dhillon [82] oder in dem Buch von Demmel [22]. Der Aufwand für diese Faktorisierung und die Lösung der beiden Gleichungssysteme umfaßt dann nur noch

etwa $6n$ Multiplikationen/Divisionen.

Eine weitere Möglichkeit zur Reduktion des Aufwands besteht darin, die Eigenvektornäherung nicht anhand der Norm des relativen Residuums

$$\|Ax^{(i)} - \lambda x^{(i)}\|_2 / \|x^{(i)}\|_2 \;=\; 1 \,/\, \|x^{(i)}\|_2$$

sondern anhand der Güte von e_i als Startvektor für die Wielandt-Iteration zu messen. Letzteres könnte etwa über die Größe des Rayleigh-Quotienten

$$\rho_i \;=\; \frac{e_i^*(A-\lambda I)^{-1}e_i}{e_i^* e_i} \;=\; e_i^*(A-\lambda I)^{-1}e_i \tag{29.12}$$

überprüft werden. Offensichtlich ist ρ_i das (i,i)-Element von $(A-\lambda I)^{-1}$, d. h. $\rho_i = u_i v_i$ in der Notation von Satz 29.4. Die Berechnung aller ρ_i, $i = 1, \ldots, n$, kostet somit nur n Multiplikationen im Vergleich zu den $8n$ Multiplikationen/Divisionen, die zur Optimierung von $\|x^{(i)}\|_2$ benötigt wurden. Als Näherungseigenvektor wird dann die Spalte $x^{(k)}$ von $(A-\lambda I)^{-1}$ ausgewählt, für die $|\rho_k|$ maximal ist.

Verwendet man das Kriterium (29.12) und die LDL^*-Faktorisierung zur Lösung der beiden Gleichungssysteme, so kostet die Berechnung einer geeigneten Eigenvektornäherung nur noch $8n$ statt $30n$ multiplikativen Operationen.

Beispiel. Wenden wir Algorithmus 29.2 und die Variante mit Kriterium (29.12) auf das Beispiel 29.3 an, so ist

$$u \;=\; \frac{x^-}{\gamma_1} \;=\; \frac{1}{\gamma_1}\begin{bmatrix} -\varepsilon \\ 1/\varepsilon^2 - 1/\varepsilon \\ 1 \end{bmatrix}, \qquad v \;=\; \frac{x^+}{\gamma_3 u_3} \;=\; \begin{bmatrix} 1 \\ 1/\varepsilon \\ -1/\varepsilon \end{bmatrix}$$

(x^- und x^+ haben wir bereits oben berechnet). Daraus ergeben sich die Koeffizienten ρ_i aus (29.12) zu

$$\rho_1 \sim \varepsilon^2\,, \quad \rho_2 \sim \varepsilon^{-2}\,, \quad \rho_3 \sim 1\,, \qquad \varepsilon \to 0\,.$$

Offensichtlich ist ρ_2 am größten. Zum Vergleich: Die Parameter a_i und b_i aus (29.11) verhalten sich wie

$$\begin{aligned} &a_1 = 1\,, \quad a_2 \sim 1\,, \quad a_3 \sim 2\,, \\ &b_3 = 0\,, \quad b_2 \sim \varepsilon^4\,, \quad b_1 \sim \varepsilon^{-6}\,, \end{aligned} \qquad \varepsilon \to 0\,,$$

und (29.10) liefert daher

$$\|x^{(1)}\|_2 \sim \varepsilon^{-1}\,, \quad \|x^{(2)}\|_2 \sim \varepsilon^{-2}\,, \quad \|x^{(3)}\|_2 \sim \sqrt{2}\,, \qquad \varepsilon \to 0\,,$$

das heißt die Norm von $x^{(2)}$ ist für kleine ε am größten.

Beide Auswahlstrategien wählen daher $x^{(2)}$ als Näherungseigenvektor, der sich aus den Vektoren u und v berechnet:

$$x^{(2)} = \frac{1}{\gamma_1}\begin{bmatrix} 1/\varepsilon^2 - 1/\varepsilon \\ 1/\varepsilon^3 - 1/\varepsilon^2 \\ 1/\varepsilon \end{bmatrix} = \frac{1-\varepsilon}{\varepsilon^2 - \varepsilon^3 + \varepsilon^4}\left(\begin{bmatrix} \varepsilon \\ 1 \\ \varepsilon^2 \end{bmatrix} + O(\varepsilon^3)\right).$$

$x^{(2)}$ weist eine gute Übereinstimmung mit dem gesuchten Eigenvektor $\widehat{x}$ auf. ◊

Beispiel 29.5. Zum Abschluß dieses Abschnitts rechnen wir noch ein numerisches Beispiel. In Beispiel 25.5 (und in Abschnitt 27.7) hatten wir den größten Eigenwert $\lambda = 0.95308992296933$ der dort angegebenen symmetrischen Tridiagonalmatrix A bestimmt. Wenden wir Algorithmus 29.2 auf dieses Problem an, so wird der dritte Testvektor $x^{(3)}$ als Näherung des gesuchten Eigenvektors ausgewählt. Die folgende Tabelle zeigt die drei Vektoren u, v und $x^{(3)}$, jeweils so normiert, daß ihr Maximaleintrag Eins ist (die dritte Komponente).

u	v	$x^{(3)}$
−0.611162218274239	−0.611162218274235	−0.611162218274235
0.959249374866718	0.959249374866716	0.959249374866716
1	1	1
0.810159006433173	0.810159006433175	0.810159006433173
−0.450552684867252	−0.450552684867256	−0.450552684867252

Wie man sieht, stimmen fast alle signifikanten Ziffern von u und v mit $x^{(3)}$ überein (die fehlerhaften Ziffern sind heller dargestellt), aber weder u noch v haben die maximale Genauigkeit. ◊

30 Das Lanczos-Verfahren

Fast alle bislang betrachteten numerischen Verfahren für das Eigenwertproblem mit einer (allgemeinen) hermiteschen Matrix $A \in \mathbb{K}^{n\times n}$ benötigen einen kubisch mit der Dimension anwachsenden Aufwand und sind daher nicht praktikabel, wenn n sehr groß ist. Falls die Matrix A dünn besetzt ist, bieten eventuell andere Verfahren wie etwa die Potenzmethode einen Ausweg, die nur Multiplikationen mit A zur Lösung des Eigenwertproblems einsetzen. Allerdings approximiert die Potenzmethode nur den betragsgrößten Eigenwert von A. Die gebrochene Iteration von Wielandt zur Approximation anderer

Eigenwerte ist für große n wieder zu aufwendig, da in jeder Iteration ein teures lineares Gleichungssystem zu lösen ist. Im folgenden soll ein Verfahren hergeleitet werden, mit dem mehrere (im Extremfall alle) Eigenwerte von A gleichzeitig approximiert werden und das die Matrix A lediglich in Form von Matrix-Vektor-Multiplikationen einsetzt.

Wie zuvor sei im weiteren vorausgesetzt, daß $A \in \mathbb{K}^{n\times n}$ hermitesch ist und die Eigenwerte $\{\lambda_i\}$ von A absteigend sortiert seien,

$$\lambda_1 \;\geq\; \lambda_2 \geq \ldots \geq\; \lambda_n\,.$$

Ferner nehmen wir ohne Einschränkung an, daß λ_1 der betragsgrößte Eigenwert ist und bezeichnen mit $z^{(k)}$, $k = 0, 1, 2, \ldots$, wie in Abschnitt 25 die Iterierten der Potenzmethode bei vorgegebenem Startvektor $z^{(0)} \in \mathbb{K}^n$. Bei hermiteschen Matrizen bieten sich nach Korollar 25.3 die Rayleigh-Quotienten

$$\mu^{(k)} \;=\; z^{(k-1)*} A z^{(k-1)} \;=\; \frac{z^{(k-1)*} A z^{(k-1)}}{z^{(k-1)*} z^{(k-1)}}$$

als Approximationen des betragsgrößten Eigenwerts von A an. Nach Lemma 23.5 kann $\mu^{(k)}$ folgendermaßen abgeschätzt werden:

$$\mu^{(k)} \;\leq\; \mu_1^{(k)} \;:=\; \max_{0\neq z\,\in\,\mathcal{Z}_k} \frac{z^* A z}{z^* z} \;\leq\; \lambda_1\,, \tag{30.1}$$

wobei $\mathcal{Z}_k = \operatorname{span}\{z^{(0)}, \ldots, z^{(k-1)}\}$. $\mu_1^{(k)}$ ist also eine mindestens genauso gute Approximation an λ_1 wie $\mu^{(k)}$.

Wir erläutern nun, wie $\mu_1^{(k)}$ verhältnismäßig einfach berechnet werden kann. Dazu setzen wir zunächst voraus, daß eine Orthonormalbasis $\{v_1, \ldots, v_k\}$ des Krylov-Raums

$$\mathcal{K}_k(A, z^{(0)}) \;=\; \operatorname{span}\{z^{(0)}, Az^{(0)}, \ldots, A^{(k-1)} z^{(0)}\} \;=\; \mathcal{Z}_k \tag{30.2}$$

gegeben sei. Die Berechnung einer solchen Orthonormalbasis kann mit dem weiter unten folgenden Lanczos-Prozeß erfolgen (vgl. Satz 30.2), der etwa den gleichen Aufwand wie die Potenzmethode besitzt. Ist

$$V_k \;=\; \big[\, v_1, \ldots, v_k \,\big] \;\in \mathbb{K}^{n\times k}$$

die Matrix mit den orthonormalen Basisvektoren in den einzelnen Spalten, so ergibt die Entwicklung eines Vektor $z \in \mathcal{K}_k(A, z^{(0)})$ in dieser Basis einen Koeffizientenvektor $y \in \mathbb{K}^k$ mit $z = V_k y$. Den Rayleigh-Quotienten können wir dann wie folgt umgeschreiben:

$$\frac{z^* A z}{z^* z} \;=\; \frac{y^* V_k^* A V_k y}{y^* V_k^* V_k y} \;=\; \frac{y^* V_k^* A V_k y}{y^* y}\,.$$

Eingesetzt in (30.1) erhalten wir somit

$$\mu_1^{(k)} \;=\; \max_{0\neq y\,\in\,\mathbb{K}^k} \frac{y^* V_k^* A V_k y}{y^* y}\,,$$

d. h. $\mu_1^{(k)}$ ist der größte Eigenwert der $k \times k$-dimensionalen Matrix $V_k^* A V_k$.

Daneben können aber auch die anderen Eigenwerte von $V_k^* A V_k$ als Näherungen an Eigenwerte von A herangezogen werden. Dies läßt sich wie folgt motivieren: $P_k = V_k V_k^*$ ist ein Orthogonalprojektor mit $\mathcal{V}_k = \mathcal{R}(P_k) = \mathcal{R}(V_k) = \mathcal{K}_k(A, z^{(0)})$ und die Matrix $V_k^* A V_k$ kann als Repräsentant der *Orthogonalprojektion* A_k von A auf $\mathcal{V}_k$ verstanden werden,

$$A_k : \mathcal{V}_k \to \mathcal{V}_k\,, \qquad A_k = P_k A\big|_{\mathcal{V}_k}\,,$$

denn für $j = 1, \ldots, k$ gilt

$$A_k v_j \;=\; \sum_{i=1}^{k} \xi_i v_i \qquad \text{genau dann, wenn} \qquad (V_k^* A V_k) e_j \;=\; \sum_{i=1}^{k} \xi_i e_i\,.$$

Je größer der Unterraum $\mathcal{V}_k$ ist, um so besser wird A_k die Abbildung A approximieren – und um so näher werden die Eigenwerte von A_k an den Eigenwerten von A liegen. Diese Idee der *Projektionsverfahren* oder *Ritz-Verfahren* ist nicht auf Krylov-Räume $\mathcal{V}_k$ beschränkt. Wie wir noch sehen werden, ist aber in diesem Fall die numerische Implementierung um vieles einfacher als im allgemeinen Fall. Die Eigenwerte von A_k nennt man übrigens *Ritzwerte*.

Satz 30.1. *$\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_n$ und $\mu_1^{(k)} \geq \mu_2^{(k)} \geq \ldots \geq \mu_k^{(k)}$ seien die Eigenwerte der hermiteschen Matrix $A \in \mathbb{K}^{n\times n}$ bzw. von $V_k^* A V_k$, $k = 1, 2, \ldots$ Dann gelten für $1 \leq j \leq k$ die Ungleichungsketten*

$$\lambda_{n-j+1} \;\leq\; \mu_{k+1-j+1}^{(k+1)} \;\leq\; \mu_{k-j+1}^{(k)} \quad \textit{und} \quad \mu_j^{(k)} \;\leq\; \mu_j^{(k+1)} \;\leq\; \lambda_j\,.$$

Mit anderen Worten, mit wachsendem k fällt der j-kleinste Eigenwert von $V_k^* A V_k$ monoton von oben gegen den j-kleinsten Eigenwert von A, während der j-größte Eigenwert von $V_k^* A V_k$ von unten monoton gegen den j-größten Eigenwert von A wächst.

In Abbildung 30.1 ist dieses Monotonieverhalten für ein Beispiel mit einer Matrix $A \in \mathbb{R}^{100\times 100}$ dargestellt: In dieser Abbildung geben die Kreise in der k-ten Zeile die k Eigenwerte von $V_k^* A V_k$ an, $k = 1, \ldots, 20$. In der untersten Zeile sind außer den Ritzwerten für $k = 20$ auch die hundert Eigenwerte von A mit Strichen auf der λ-Achse markiert. Schließlich illustrieren die gebrochenen bzw. gepunkteten Verbindungslinien die Monotonieaussagen des obigen Satzes: Die gebrochenen Linien verbinden jeweils die j-kleinsten Ritzwerte, die gepunkteten Linien entsprechend die j-größten Ritzwerte.

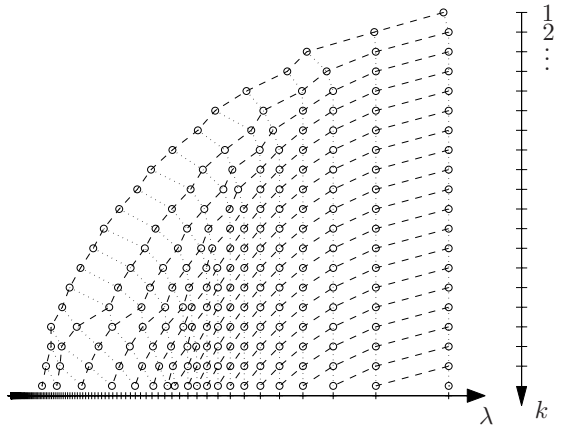


Abb. 30.1: Monotonie-Eigenschaften der Ritzwerte

Man beachte die guten Approximationen der etwa zehn größten Ritzwerte in der untersten Zeile. Auf der anderen Seite sieht man aber auch, daß die kleinen Ritzwerte keine guten Approximationen an die kleinsten Eigenwerte von A sind.

Der folgende Beweis von Satz 30.1 ist im wesentlichen eine Anwendung des Minmax-Prinzips von Courant-Fischer (für einen alternativen Beweis sei auf Abschnitt 35.1 verwiesen).

Beweis von Satz 30.1. Sei $k < n$ fest gewählt, und seien $\{y_i\}$ die orthonormierten Eigenvektoren von $V_k^* A V_k$ zu den Eigenwerten $\mu_i^{(k)}$, $i = 1, \ldots, k$. Ferner sei $z_i = V_k y_i$, $i = 1, \ldots, k$. Man beachte, daß für beliebiges $j \in \{1, \ldots, k\}$ die Vektoren $\{z_i\}_{i=1}^{j}$ wegen der Isometrie-Eigenschaft der Abbildung $V_k : \mathbb{K}^k \to \mathbb{K}^n$ eine Orthonormalbasis der linearen Hülle $\operatorname{span}\{z_1, \ldots, z_j\}$ bilden.

1. Zunächst beweisen wir die Ungleichung

$$\mu_j^{(k)} \;\le\; \lambda_j\,, \qquad j = 1, \ldots, k\,.$$

Aufgrund der Diskussion des Gleichheitszeichens beim Minmax-Prinzip von Courant-Fischer (Satz 23.7) gilt

$$\mu_j^{(k)} \;=\; \min_{0 \neq y \in \mathcal{Y}_j} \frac{y^* V_k^* A V_k y}{y^* y} \qquad \text{mit} \qquad \mathcal{Y}_j \;=\; \operatorname{span}\{y_1, \ldots, y_j\}\,. \quad (30.3)$$

Setzen wir $z = V_k y$, dann ist $z^* z = y^* y$ und z durchläuft den j-dimensionalen Unterraum $\mathcal{Z}_j \;=\; \operatorname{span}\{z_1, \ldots z_j\}$. Aus Satz 23.7 folgt somit

$$\mu_j^{(k)} \;=\; \min_{0 \neq y \in \mathcal{Y}_j} \frac{y^* V_k^* A V_k y}{y^* y} \;=\; \min_{0 \neq z \in \mathcal{Z}_j} \frac{z^* A z}{z^* z} \;\le\; \lambda_j\,,$$

was zu zeigen war.

2. Nun zeigen wir die Monotonie der Eigenwertapproximationen bei wachsendem k,

$$\mu_j^{(k)} \leq \mu_j^{(k+1)}, \qquad j = 1, \ldots, k\,.$$

Wegen der speziellen Form der Matrizen V_k, $k = 1, 2, \ldots$, gilt für jedes $y \in \mathbb{K}^k$

$$V_k y = V_{k+1}\widehat{y} \qquad \text{mit} \qquad \widehat{y} = \begin{bmatrix} y \\ 0 \end{bmatrix} \in \mathbb{K}^{k+1}\,.$$

Konstruieren wir entsprechend die Vektoren $\widehat{y}_1, \ldots, \widehat{y}_k$, so folgt aus (30.3) für ein beliebiges $j \in \{1, \ldots, k\}$

$$\mu_j^{(k)} = \min_{0 \neq y \in \mathcal{Y}_j} \frac{y^* V_k^* A V_k y}{y^* y} = \min_{0 \neq \widehat{y} \in \widehat{\mathcal{Y}}_j} \frac{\widehat{y}^* V_{k+1}^* A V_{k+1} \widehat{y}}{\widehat{y}^* \widehat{y}}$$

mit $\widehat{\mathcal{Y}}_j = \operatorname{span}\{\widehat{y}_1, \ldots, \widehat{y}_j\}$. Dabei ist die rechte Seite, wiederum nach Satz 23.7, kleiner gleich $\mu_j^{(k+1)}$, und dieser Teil der Behauptung ist ebenfalls bewiesen.

3. Wir wenden abschließend die bereits bewiesenen Aussagen auf die Matrix $-A$ an. $-A$ und $V_k^*(-A)V_k$ haben die Eigenwerte

$$-\lambda_n \geq \ldots \geq -\lambda_1 \qquad \text{bzw.} \qquad -\mu_k^{(k)} \geq \ldots \geq -\mu_1^{(k)}\,.$$

Nach den bereits bewiesenen Aussagen wachsen die j-größten Eigenwerte von $V_k^*(-A)V_k$ monoton gegen den j-größten Eigenwert von $-A$,

$$-\mu_{k-j+1}^{(k)} \leq -\mu_{k+1-j+1}^{(k+1)} \leq -\lambda_{n-j+1}\,,$$

und der Beweis ist vollständig. □

Satz 30.1 macht keine Aussage über die Güte der Eigenwertnäherungen $\mu_j^{(k)}$. Klar ist nur, daß $\mu_1^{(k)}$ wegen (30.1) eine bessere Approximation an λ_1 ist als die entsprechende Näherung der Potenzmethode. Wir werden weiter unten (in Proposition 30.4) eine A-posteriori-Abschätzung angeben, mit der man die Genauigkeit aller Näherungen aus Satz 30.1 überprüfen kann.

Offen ist bislang auch noch die Frage, wie die Orthonormalbasis $\{v_i\}_{i=1}^k$ des Krylov-Raums $\mathcal{K}_k(A, z^{(0)})$ effizient bestimmt werden kann. Eine Möglichkeit hierfür ist der *Lanczos-Prozeß*:

Satz 30.2. *Sei $A \in \mathbb{K}^{n\times n}$ hermitesch und $v_1 = z^{(0)}$ ein beliebiger Vektor mit $\|z^{(0)}\|_2 = 1$. Ferner sei v_0 der Nullvektor in $\mathbb{K}^n$ und $\beta_0 = 0$ gesetzt. Dann bilden die Vektoren $\{v_i\}_{i=1}^k$ aus der dreistufigen Rekursionsformel*

$$r_{i+1} = (A - \alpha_i I)v_i - \beta_{i-1}v_{i-1}\,, \qquad v_{i+1} = r_{i+1}/\beta_i\,, \tag{30.4}$$

$i = 1, \ldots, k-1$, mit $\alpha_i = v_i^* A v_i$ und $\beta_i = \|r_{i+1}\|_2$ eine Orthonormalbasis von $\mathcal{K}_k(A, z^{(0)})$, falls alle β_i, $i = 1, \ldots, k-1$, von Null verschieden sind.

Beweis. Der Beweis geht induktiv, wobei die Aussage für $k = 1$ trivial ist. Im Induktionsschritt $k \to k+1$, $k \geq 1$, sei angenommen, daß $\{v_1, \ldots, v_k\}$ eine Orthonormalbasis von $\mathcal{K}_k(A, z^{(0)})$ und alle β_j, $1 \leq j \leq k$, von Null veschieden sind. Dann folgt aus (30.4) die Darstellung

$$r_{k+1} = Av_k + w_k \tag{30.5}$$

mit $w_k \in \mathcal{K}_k(A, z^{(0)})$ und $Av_k \in \mathcal{K}_{k+1}(A, z^{(0)})$. Insbesondere gehört also v_{k+1} zu $\mathcal{K}_{k+1}(A, z^{(0)})$. Zum Nachweis der Orthogonalität sei zunächst $i \in \{1, \ldots, k-2\}$: Dann folgt wegen der Induktionsvoraussetzung aus (30.4), daß

$$v_i^* r_{k+1} \;=\; v_i^* A v_k - \alpha_k \underbrace{v_i^* v_k}_{=0} - \beta_{k-1} \underbrace{v_i^* v_{k-1}}_{=0} \;=\; (Av_i)^* v_k\,.$$

Da $1 \leq i \leq k-2$ ist, gehört Av_i zu $\mathcal{K}_{k-1}(A, z^{(0)})$ und ist daher orthogonal zu v_k. Demnach gilt

$$v_i^* r_{k+1} \;=\; v_i^* v_{k+1} \;=\; 0\,, \qquad i = 1, \ldots, k-2\,. \tag{30.6}$$

Es verbleibt somit noch der Nachweis von (30.6) für $i = k-1$, falls $k > 1$ ist, und für $i = k$. Für $i = k-1$ kann man zunächst wie vorher argumentieren:

$$v_{k-1}^* r_{k+1} \;=\; v_{k-1}^* A v_k - \alpha_k v_{k-1}^* v_k - \beta_{k-1} v_{k-1}^* v_{k-1} \;=\; (Av_{k-1})^* v_k - \beta_{k-1}\,.$$

Ersetzt man nun Av_{k-1} gemäß (30.5), so ergibt sich hieraus

$$\begin{aligned} v_{k-1}^* r_{k+1} \;&=\; (r_k - w_{k-1})^* v_k - \beta_{k-1} \;=\; r_k^* v_k - w_{k-1}^* v_k - \beta_{k-1} \\ &=\; \|r_k\|_2 - w_{k-1}^* v_k - \|r_k\|_2 \;=\; -w_{k-1}^* v_k\,. \end{aligned}$$

Da $w_{k-1} \in \mathcal{K}_{k-1}(A, z^{(0)})$ ist, folgt aus der Induktionsvoraussetzung $w_{k-1}^* v_k = 0$, und daher gilt (30.6) auch für $i = k-1$.

Für $i = k$ ist schließlich

$$v_k^* r_{k+1} \;=\; v_k^* A v_k - \alpha_k \underbrace{v_k^* v_k}_{=1} - \beta_{k-1} \underbrace{v_k^* v_{k-1}}_{=0} \;=\; v_k^* A v_k - \alpha_k\,,$$

und letzteres verschwindet aufgrund der Definition von α_k. Somit steht r_{k+1} aus $\mathcal{K}_{k+1}(A, z^{(0)})$ senkrecht auf $\mathcal{K}_k(A, z^{(0)})$. Da $\beta_k = \|r_{k+1}\|_2 \ / = 0$ angenommen wurde, ist $v_{k+1} = r_{k+1}/\|r_{k+1}\|_2$ also wohldefiniert und ergänzt $\{v_1, \ldots, v_k\}$ zu einer Orthonormalbasis von $\mathcal{K}_{k+1}(A, z^{(0)})$. □

Als nächstes untersuchen wir die Matrix $V_k^* A V_k$, die sich aus der Orthonormalbasis aus Satz 30.2 ergibt.

Proposition 30.3. *Sei $V_k = [v_1, \ldots, v_k]$ mit den Vektoren v_i, $i = 1, \ldots, k$, aus Satz* 30.2. *Dann gilt*

$$T_k \ = \ V_k^* A V_k \ = \ \begin{bmatrix} \alpha_1 & \beta_1 & & 0 \\ \beta_1 & \alpha_2 & \ddots & \\ & \ddots & \ddots & \beta_{k-1} \\ 0 & & \beta_{k-1} & \alpha_k \end{bmatrix} \tag{30.7}$$

mit den Koeffizienten $\{\alpha_i\}$ und $\{\beta_i\}$ aus Satz 30.2.

Beweis. Der (i,j)-Eintrag t_{ij} von $V_k^* A V_k$ ist gegeben durch $v_i^* A v_j$. Also ist $V_k^* A V_k$ hermitesch und aufgrund der Definition der Koeffizienten in Satz 30.2 ist $t_{ii} = \alpha_i$, $i = 1, \ldots, k$. Ferner ist $t_{ij} = 0$ für $i - j > 1$, denn Av_j gehört zu $\mathcal{K}_{j+1}(A, z^{(0)})$ und dieser Unterraum ist orthogonal zu v_i. Für $i = j + 1$, $j = 1, \ldots, k-1$, ergibt sich schließlich

$$t_{j+1,j} \ = \ v_{j+1}^* A v_j \ \overset{(30.5)}{=} \ v_{j+1}^* (r_{j+1} - w_j) \ = \ v_{j+1}^* r_{j+1} \ = \ \beta_j \, .$$

□

Die Eigenwerte von T_k können beispielsweise mit dem QR-Verfahren sehr effizient ermittelt werden, zumal in der Regel k wesentlich kleiner als n ist. Anhand der folgenden A-posteriori-Abschätzung läßt sich überprüfen, wie gut diese Eigenwerte diejenigen von A approximieren.

Proposition 30.4. *Sei (μ, w) ein Eigenpaar von $T_k = V_k^* A V_k$ mit $\|w\|_2 = 1$ und ω_k die letzte Komponente von w. Dann besitzt A einen Eigenwert λ mit*

$$|\lambda - \mu| \ \leq \ \beta_k |\omega_k| \, , \tag{30.8}$$

wobei β_k wie in Satz 30.2 *definiert ist.*

Beweis. Der Beweis beruht auf der Darstellung

$$A V_k - V_k T_k \ = \ \beta_k v_{k+1} e_k^* \, . \tag{30.9}$$

Zum Nachweis dieser Identität multiplizieren wir die linke Seite von (30.9) mit e_i und erhalten für $1 < i < k$ aus (30.4)

$$\begin{aligned} (AV_k - V_k T_k) e_i \ &= \ A v_i - V_k (\beta_{i-1} e_{i-1} + \alpha_i e_i + \beta_i e_{i+1}) \\ &= \ A v_i - \beta_{i-1} v_{i-1} - \alpha_i v_i - \beta_i v_{i+1} \ = \ r_{i+1} - \beta_i v_{i+1} \ = \ 0 \end{aligned}$$

Initialisierung: $A \in \mathbb{K}^{n\times n}$ sei hermitesch, groß und dünn besetzt
wähle Startvektor $z^{(0)} \in \mathbb{K}^n$
$T_0 = [\,]$ % die leere Matrix
$v_0 = 0, \quad r_1 = z^{(0)}$
`for` $k = 1, 2, \ldots$ `do`
$\beta_{k-1} = \|r_k\|_2$
$v_k = r_k/\beta_{k-1}$
$\alpha_k = v_k^* A v_k$
$r_{k+1} = (A - \alpha_k I)v_k - \beta_{k-1}v_{k-1}$
$T_k = \begin{bmatrix} T_{k-1} & \beta_{k-1}e_{k-1} \\ \beta_{k-1}e_{k-1}^* & \alpha_k \end{bmatrix}$ % $T_1 = [\alpha_1]$ für $k = 1$
bestimme alle Eigenwerte von T_k ... % z. B. mit dem QR-Verfahren
... und die zugehörigen Eigenvektoren % mit Algorithmus 29.2
`until` hinreichend viele Eigenwerte von A approximiert werden, vgl. Proposition 30.4

Algorithmus 30.1: Lanczos-Verfahren

aufgrund der Definition von β_i. Wegen $v_0 = 0$ bleibt diese Umformung auch für $i = 1$ korrekt, sofern $e_0 = 0$ gesetzt wird. Schließlich ergibt sich für $i = k$

$$\begin{aligned}(AV_k - V_kT_k)e_k &= Av_k - V_k(\beta_{k-1}e_{k-1} + \alpha_k e_k)\\ &= Av_k - \beta_{k-1}v_{k-1} - \alpha_k v_k = r_{k+1} = \beta_k v_{k+1}\,,\end{aligned}$$

und somit ist (30.9) spaltenweise nachgewiesen. Für $x = V_k w$ mit $\|x\|_2 = \|w\|_2 = 1$ folgt aus (30.9)

$$Ax - \mu x = AV_k w - V_k(\mu w) = (AV_k - V_kT_k)w = \beta_k v_{k+1}e_k^* w = \beta_k \omega_k v_{k+1}\,.$$

Wegen $\beta_k > 0$ ergibt dies

$$\|Ax - \mu x\|_2 = \beta_k\,|\omega_k| = \beta_k\,|\omega_k|\,\|x\|_2\,,$$

und damit folgt die Behauptung aus Satz 24.3. □

Bemerkung. Prinzipiell ist es möglich, daß der Lanczos-Prozeß (30.4) vorzeitig terminiert, weil ein r_{k+1} mit $k + 1 < n$ Null wird. In diesem Fall ist

$$Av_k = \alpha_k v_k + \beta_{k-1}v_{k-1} \in \mathcal{K}_k(A, z^{(0)})$$

und somit $\mathcal{K}_{k+1}(A, z^{(0)}) = \mathcal{K}_k(A, z^{(0)})$. Mit anderen Worten, $\mathcal{K}_k(A, z^{(0)})$ ist ein invarianter Unterraum von A und die Eigenwerte und Eigenvektoren der Orthogonalprojektion A_k stimmen mit Eigenwerten und Eigenvektoren von A

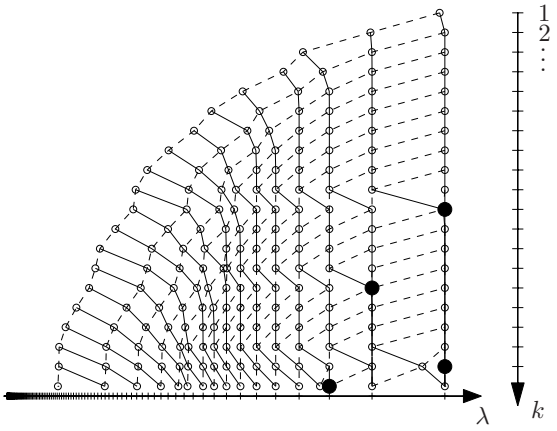


Abb. 30.2: Ritzwerte inklusive Rundungsfehler

überein. Wegen $\beta_k = 0$ kann man dies auch an Proposition 30.4 ablesen: Alle Eigenwerte von T_k sind auch Eigenwerte von A; ist w ein Eigenvektor von T_k, dann ist $V_k w$ der entsprechende Eigenvektor von A. Um weitere Eigenwerte von A zu bestimmen, muß das Verfahren mit einem anderen Startvektor $\tilde{z}^{(0)} \notin \mathcal{K}_k(A, z^{(0)})$ neu gestartet werden. ◊

Beispiel. Zwanzig Schritte des Lanczos-Verfahrens ergeben bei der Matrix A aus Beispiel 28.5 mit exakter Arithmetik die in Abbildung 30.1 dargestellten Ritzwerte. In jeder der 20 Zeilen dieser Abbildung sieht man die Ritzwerte nach dem k-ten Iterationsschritt, $k = 1, \ldots, 20$. In der Praxis sieht das Ergebnis in der Regel nicht ganz so gut aus, denn aufgrund von Rundungsfehlern geht die Orthogonalität der Basisvektoren v_i, $i = 1, 2, \ldots$, recht bald verloren. Für die Berechnung der Ritzwerte aus Abbildung 30.1 wurden daher die v_i *reorthogonalisiert*; verschiedene Strategien hierzu findet man in dem Buch von Golub und Van Loan [34]. Ohne eine solche Reorthogonalisierung ergeben sich die Ritzwerte aus Abbildung 30.2.

Vergleicht man die beiden Abbildungen, so erkennt man gewisse „Spaltensprünge“ in Abbildung 30.2: An den vier durch etwas dickere ausgefüllte Kreise hervorgehobenen Stellen werden Ritzwerte „verdoppelt“, d. h. ab diesem Index k approximieren zwei Ritzwerte denselben Eigenwert von A; man spricht von sogenannten *Geistern*. Die obige Darstellungsweise bietet eine Möglichkeit, solchen Geistern auf die Spur zu kommen.

Glücklicherweise können Geister in der Regel ohne größeren Genauigkeitsverlust bei der weiteren Rechnung ignoriert werden. Dies wird in der Tabelle in Abbildung 30.3 verdeutlicht, die den größten Ritzwert mit der entsprechenden Näherung der Potenzmethode bei gleichem Startvektor vergleicht. Trotz des in

Konvergenz gegen den größten Eigenwert:

k	Lanczos-Verfahren	Potenzmethode
1	0.9047943877776	0.9047943877776
2	0.9998925431301	0.9935201060703
3	1.0000804817493	0.9996682526178
4	1.0000806298220	1.0000548418775
5	1.0000806300188	1.0000790175509
6	1.0000806300189	1.0000805291921
7	1.0000806300189	1.0000806237142
8	1.0000806300189	1.0000806296246
9	1.0000806300189	1.0000806299942
10	1.0000806300189	1.0000806300173
11	1.0000806300189	1.0000806300189

Approximationen nach 20 Iterationen:

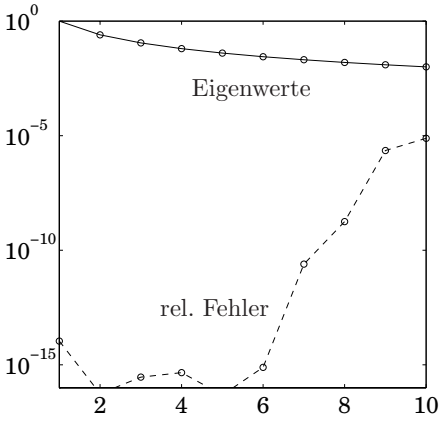


Abb. 30.3: Konvergenzgeschwindigkeit und relative Genauigkeit des Lanczos-Verfahrens

diesem Fall sehr günstigen Konvergenzfaktors $q = \lambda_2/\lambda_1 \approx 0.25$ der Potenzmethode benötigt das Lanczos-Verfahren nur etwa halb so viele Iterationen wie die Potenzmethode, um alle signifikanten Stellen von λ_1 zu bestimmen.

In den zwanzig Iterationen des Lanczos-Verfahrens werden neben dem größten Eigenwert von A trotz der Rundungsfehler und der Geister noch knapp zehn weitere Eigenwerte von A auf mehrere Stellen genau berechnet. Die Graphik in Abbildung 30.3 zeigt die zehn größten Eigenwerte von A sowie den relativen Fehler der approximierenden Ritzwerte: Die sechs größten Eigenwerte sind demnach auf rund 14 Stellen genau berechnet worden, bei den anderen vier Näherungen sind immerhin noch drei bis neun signifikante Dezimalstellen korrekt. ◊

Aufgaben

1. Gegeben sei die Differentialgleichung

$$x''(t) + dx'(t) + Ax(t) \;=\; \cos(\omega_0 t)\, v$$

mit $\omega_0 > 0$, wobei $A \in \mathbb{R}^{n\times n}$ symmetrisch positiv definit und $v \in \mathbb{R}^n$ ein Eigenvektor von A zum Eigenwert λ sei. Zeigen Sie, daß die Funktion

$$x(t) \;=\; a \sin(\omega_0 t - \theta)\, v$$

eine Lösung dieser Differentialgleichung ist, wenn a und θ wie folgt gegeben sind:

$$a \;=\; \big((\omega_0^2 - \lambda)^2 + d^2\omega_0^2\big)^{-1/2} \quad \text{und} \quad \theta \;=\; \arctan\frac{\omega_0^2 - \lambda}{d\,\omega_0} \qquad \text{für } d > 0$$

beziehungsweise

$$a \;=\; 1/(\omega_0^2 - \lambda) \quad \text{und} \quad \theta \;=\; \pi/2 \qquad \text{für } d = 0 \text{ und } \omega_0 \neq \sqrt{\lambda}\,.$$

2. Es seien $A, B \in \mathbb{K}^{n\times n}$ hermitesch. Zeigen Sie:
(a) Alle Eigenwerte von AB sind reell.
(b) Ist A zudem positiv definit, dann ist AB diagonalisierbar.
Geben Sie hermitesche Matrizen A und B an, deren Produkt AB nicht diagonalisierbar ist.

3. Sei A eine $n \times n$-Matrix mit paarweise disjunkten Gerschgorin-Kreisen. Zeigen Sie, daß jeder Gerschgorin-Kreis genau einen Eigenwert enthält und daß im Fall einer reellen Matrix A unter dieser Voraussetzung alle Eigenwerte reell sind.
Hinweis: Zerlegen Sie $A = D + N$ in den Diagonal- und Nebendiagonalanteil und betrachten Sie die Eigenwerte der Matrizenschar $A_\tau = D + \tau N$, $0 \le \tau \le 1$.

4. Die Matrix $S \in \mathbb{R}^{n\times n}$ sei schiefhermitesch und es sei $A = \alpha I + S$ mit $\alpha \in \mathbb{R}$. Zeigen Sie, daß der reelle Wertebereich $\mathcal{W}_{\mathbb{R}}(A)$,

$$\mathcal{W}_{\mathbb{R}}(A) \;=\; \big\{\, \zeta = \frac{x^*Ax}{x^*x} \;:\; x \in \mathbb{R}^n \setminus \{0\} \,\big\},$$

aus genau einem Punkt besteht.

5. Gegeben sei $A \in \mathbb{R}^{n\times n}$ hermitesch. Zeigen Sie, daß der Gradient der Funktion $r(x) = (x^*Ax)/(x^*x)$ für $x \in \mathbb{R}^n \setminus \{0\}$ genau dann verschwindet, wenn x ein Eigenvektor von A ist.

6. Zeigen Sie, daß der Wertebereich der Matrix

$$A = \begin{bmatrix} d_1 & a \\ 0 & d_2 \end{bmatrix}$$

eine Ellipse mit Brennpunkten d_1 und d_2 und kleiner Halbachse $|a|/2$ ist.

7. Sei λ ein Eigenwert von A auf dem Rand des Wertebereichs $\mathcal{W}(A)$. Zeigen Sie, daß jeder zu λ gehörende Eigenvektor senkrecht auf allen Eigenvektoren zu anderen Eigenwerten von A steht.
Hinweis: Betrachten Sie $\mathcal{W}(A|_{\mathrm{span}\{v,w\}})$ mit $Av = \lambda v$ und einem weiteren Eigenvektor w von A.

8. Beweisen Sie das Minmax-Prinzip (23.8) von Courant-Fischer.

9. Zu gegebenem $A, B \in \mathbb{K}^{n\times n}$ mit Rang $B = n$ und $A = A^*$ sei

$$M = \begin{bmatrix} A & B \\ B^* & 0 \end{bmatrix} \in \mathbb{K}^{2n\times 2n}.$$

(a) Zeigen Sie mit Hilfe des Minmax-Prinzips von Courant-Fischer, daß M jeweils n positive und negative Eigenwerte hat.
(b) Drücken Sie für die beiden Spezialfälle $A = 0$ und $A = I$ die Eigenwerte und Eigenvektoren von M durch Singulärwerte und Singulärvektoren von B aus.

10. Betrachten Sie die Matrix

$$A_\varepsilon = \begin{bmatrix} 1+\varepsilon\cos(2/\varepsilon) & -\varepsilon\sin(2/\varepsilon) \\ -\varepsilon\sin(2/\varepsilon) & 1-\varepsilon\cos(2/\varepsilon) \end{bmatrix},$$

und bestimmen Sie die Eigenwerte und Eigenvektoren von A_ε. Untersuchen Sie den Grenzfall $\varepsilon \to 0$.

11. Gegeben sei $A \in \mathbb{K}^{n\times n}$ mit

$$S^{-1}AS = J = \begin{bmatrix} \lambda_1 & 1 & & 0 \\ & \lambda_1 & \ddots & \\ & & \ddots & 1 \\ 0 & & & \lambda_1 \end{bmatrix}$$

und eine Approximation (λ, x), $\|x\|_2 = 1$, an das Eigenpaar von A.
Beweisen Sie für $u = Ax - \lambda x$ die Abschätzung

$$\frac{1-|\lambda_1-\lambda|}{1-|\lambda_1-\lambda|^n}\,|\lambda_1-\lambda|^n \le \operatorname{cond}_2(S)\,\|u\|_2\,.$$

Hinweis: Verwenden Sie, daß $1 = \|x\|_2 = \|S(J-\lambda I)^{-1}S^{-1}u\|_2$ gilt, und schätzen Sie $\|(J-\lambda I)^{-1}\|_2$ mit Hilfe von Satz 2.8 nach oben ab.

12. Betrachten Sie die Potenzmethode unter den Voraussetzungen von Satz 25.1, wobei darüber hinaus alle Eigenwerte von A nichtnegativ seien. Ersetzen Sie in Algorithmus 25.1 die Matrix A durch die geshiftete Matrix $A - \mu I$. Für welche Werte $\mu \in \mathbb{R}$ läßt sich mit diesem Verfahren eine Näherung an den Spektralradius $\varrho(A)$ berechnen? Bestimmen Sie den Wert $\widehat{\mu}$, für den die Konvergenzgeschwindigkeit maximal wird.

13. Gegeben sei die *zirkulante Shiftmatrix*

$$S = \begin{bmatrix} 0 & 1 & & 0 \\ & 0 & \ddots & \\ & & \ddots & 1 \\ 1 & & & 0 \end{bmatrix} \in \mathbb{R}^{n\times n}.$$

(a) Überlegen Sie sich, daß alle Eigenwerte von S im abgeschlossenen Einheitskreis liegen.
(b) Wenden Sie die Potenzmethode auf den Startvektor $z^{(0)} = e_1$ an. Wie lautet die k-te Iterierte?
(c) Offensichtlich konvergiert die Potenzmethode mit diesem Startvektor nicht. Wieso steht das nicht im Widerspruch zu Satz 25.1?

14. Zeigen Sie, daß die QR-Zerlegung einer $n \times n$ Tridiagonalmatrix mit Givens-Rotationen wie in Abschnitt 27.2 etwa $13n$ Multiplikationen, Divisionen und Quadratwurzeln erfordert.

15. Sei A eine symmetrische Tridiagonalmatrix mit $\operatorname{diag}(A) = 0$.
(a) Machen Sie sich klar, daß jeder Iterationsschritt des QR-Verfahrens ohne Shift (vgl. Abschnitt 27.2) diese Eigenschaft invariant läßt, d. h. für alle Iterierten ist $\operatorname{diag}(A_k) = 0$.
(b) Wie vereinbart sich das mit Satz 26.2?

16. Seien $\alpha_i \in \mathbb{R}$, $i = 1, \ldots, n$, und $\beta_i \in \mathbb{C}$, $i = 1, \ldots, n-1$. Zeigen Sie, daß die hermiteschen Tridiagonalmatrizen

$$\begin{bmatrix} \alpha_1 & \beta_1 & & \\ \overline{\beta_1} & \alpha_2 & \ddots & \\ & \ddots & \ddots & \beta_{n-1} \\ & & \overline{\beta_{n-1}} & \alpha_n \end{bmatrix} \quad \text{und} \quad \begin{bmatrix} \alpha_1 & |\beta_1| & & \\ |\beta_1| & \alpha_2 & \ddots & \\ & \ddots & \ddots & |\beta_{n-1}| \\ & & |\beta_{n-1}| & \alpha_n \end{bmatrix}$$

die gleichen Eigenwerte haben. Wie ergeben sich die Eigenvektoren der ersten Matrix aus den entsprechenden Eigenvektoren der zweiten Matrix?

17. Überlegen Sie sich, daß die Faktoren γ_1 und γ_n aus (29.7) und (29.8) immer gleich sind, wenn A den Voraussetzungen von Lemma 29.2 genügt.

18. Die symmetrische Tridiagonalmatrix $A \in \mathbb{R}^{n \times n}$ aus (29.1) mit

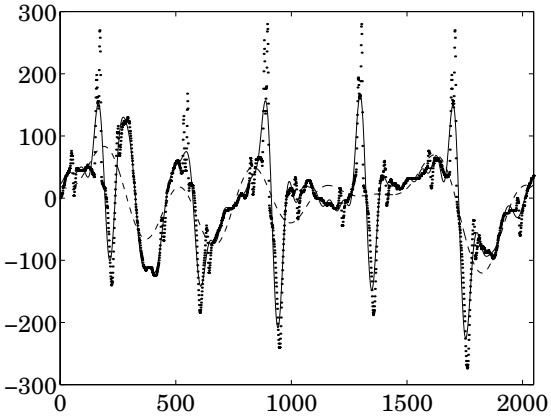
$$\alpha_1 = \ldots = \alpha_n = \frac{n-1}{2} \qquad \text{und} \qquad \beta_k = \frac{k}{2}\Big(\frac{n^2 - k^2}{4k^2 - 1}\Big)^{1/2}, \qquad k = 1, \ldots, n-1,$$

hat die Eigenwerte $0, \ldots, n-1$, vgl. Aufgabe VI.16. Implementieren Sie Algorithmus 29.2 und berechnen Sie damit für $n = 100$ die Näherungseigenvektoren $x^{(k)}$, $k = 1, \ldots, n$, zum Eigenwert $\lambda = 0$. Plotten Sie sowohl die Normen der $x^{(k)}$ als auch der Rayleigh-Quotienten ρ_k aus (29.12) über k. Vergleichen Sie die Ergebnisse für große k mit denen für kleine k.

19. Zeigen Sie, daß für hermitesche Matrizen $A \in \mathbb{K}^{n \times n}$ das Arnoldi-Verfahren mathematisch äquivalent zum Lanczos-Verfahren ist.

20. Implementieren Sie das Lanczos-Verfahren und bestimmen Sie hiermit den Spektralradius der Iterationsmatrix des symmetrischen Gauß-Seidel-Verfahrens für das Modellproblem aus Beispiel 9.7. Beachten Sie, daß die Iterationsmatrix dieses Verfahrens nicht symmetrisch ist.

Interpolation und Approximation



VI Orthogonalpolynome

Effiziente Darstellungen bzw. Methoden zur Approximation von Funktionen einer reellen Variablen stehen im Mittelpunkt dieses zweiten Buchteils. Am einfachsten ist es, lediglich endlich viele Funktionswerte an gewissen *Knoten* abzuspeichern. Werden Funktionswerte zwischen den Knoten benötigt, müssen diese Werte *interpoliert* werden. Alternativ kann die Funktion durch ein Element eines endlichdimensionalen Funktionenraums $\mathcal{F}$ *approximiert* werden, repräsentiert durch eine Linearkombination geeigneter Basisfunktionen.

Für den ersten Ansatz verweisen wir auf die Kapitel VIII und IX. In dem nun folgenden Kapitel über Orthogonalpolynome wird der zweite Zugang und für $\mathcal{F}$ der Raum Π_n aller Polynome mit Grad kleiner oder gleich n gewählt. Zur Approximation einer Funktion f wird dasjenige Polynom $p \in \Pi_n$ bestimmt, das bezüglich einer geeigneten Innenproduktnorm am nächsten an f liegt. Die Berechnung dieses Polynoms erfolgt mit Hilfe *orthogonaler Polynome*.

Die reichhaltige (analytische) Theorie der Orthogonalpolynome wird umfassend in dem Standardwerk von Szegö [100] dargestellt. Ergänzend sei auch auf das Buch von Chihara [15] verwiesen.

31 Innenprodukträume, Orthogonalbasen und Gramsche Matrizen

Der Begriff der *Orthogonalität* ist aus dem $\mathbb{K}^n$ bereits vertraut: Zwei Vektoren $x, y \in \mathbb{K}^n$ heißen zueinander orthogonal, falls

$$x^* y = \sum_{i=1}^{n} \overline{x_i} y_i = 0\,.$$

Geometrisch bedeutet dies, daß zwischen x und y ein rechter Winkel ist. In beliebigen Vektorräumen werden Orthogonalität und Winkel mit Hilfe eines Innenprodukts definiert.

Definition 31.1. Eine Abbildung $\langle \cdot, \cdot \rangle : X \times X \to \mathbb{K}$ in einem Vektorraum X über $\mathbb{K}$ heißt *Innenprodukt* oder *Skalarprodukt*, falls

$$\begin{aligned}
&(i) && \langle f, g \rangle = \overline{\langle g, f \rangle} && \text{für alle } f, g \in X, \\
&(ii) && \langle f, \alpha g + \beta h \rangle = \alpha \langle f, g \rangle + \beta \langle f, h \rangle && \text{für alle } f, g, h \in X,\ \alpha, \beta \in \mathbb{K}, \\
&(iii) && \langle f, f \rangle > 0 && \text{für alle } f \in X \setminus \{0\}.
\end{aligned}$$

Zwei Elemente $f, g \in X$ sind zueinander *orthogonal*, falls $\langle f, g \rangle = 0$. Wegen (i) gilt eine ähnliche Linearitätsbedingung (ii) auch für das erste Argument, die Koeffizienten treten hierbei jedoch komplex konjugiert auf:

$$\langle \alpha g + \beta h, f \rangle = \overline{\alpha} \langle g, f \rangle + \overline{\beta} \langle h, f \rangle .$$

Eine Abbildung $\langle \cdot, \cdot \rangle$, die (i) und (ii) erfüllt, wird daher auch *hermitesche Bilinearform* genannt. Für $\beta = 0$ und $f = \alpha g$ ergibt sich aus (ii) die Gleichung $\langle \alpha g, \alpha g \rangle = |\alpha|^2 \langle g, g \rangle$, speziell für $\alpha = 0$ folgt $\langle 0, 0 \rangle = 0$.

Proposition 31.2. *Ist $\langle \cdot, \cdot \rangle$ ein Innenprodukt in einem Vektorraum X über $\mathbb{K}$, so gilt die* Cauchy-Schwarz-Ungleichung

$$|\langle g, h \rangle|^2 \le \langle g, g \rangle \langle h, h \rangle \qquad \textit{für alle } g, h \in X ,$$

und Gleichheit gilt genau dann, wenn g und h linear abhängig sind.

Beweis. Zunächst macht man sich leicht klar, daß die Ungleichung erfüllt ist, wenn g und h zueinander orthogonal sind und daß das Gleichheitszeichen gilt, falls g und h linear abhängig sind. Im weiteren können wir uns also auf den Fall beschränken, daß g und h linear unabhängig und nicht zueinander orthogonal sind. Aus (iii) und (ii) folgt dann für $f = \alpha g + \beta h$

$$0 < \langle f, f \rangle = |\alpha|^2 \langle g, g \rangle + \overline{\alpha}\beta \langle g, h \rangle + \alpha\overline{\beta} \langle h, g \rangle + |\beta|^2 \langle h, h \rangle .$$

Somit ist

$$-2 \operatorname{Re} \left(\overline{\alpha}\beta \langle g, h \rangle \right) < |\alpha|^2 \langle g, g \rangle + |\beta|^2 \langle h, h \rangle ,$$

und speziell für

$$\alpha = \langle g, h \rangle \langle h, h \rangle^{1/2} \qquad \text{und} \qquad \beta = -|\langle g, h \rangle| \langle g, g \rangle^{1/2}$$

folgt

$$2 |\langle g, h \rangle|^3 \langle g, g \rangle^{1/2} \langle h, h \rangle^{1/2} < 2 |\langle g, h \rangle|^2 \langle g, g \rangle \langle h, h \rangle$$

und damit die Behauptung. □

In reellen Innenprodukträumen ist somit

$$\frac{\langle x, y \rangle}{\|x\| \, \|y\|} \in [-1, 1] \qquad \text{für } x, y \neq 0 \,,$$

und dieser Bruch ist genau dann gleich ± 1, wenn x und y linear abhängig sind. In diesem Fall gibt das Vorzeichen darüber Aufschluß, ob x und y in die gleiche oder in entgegengesetzte Richtungen zeigen. Der Bruch ist Null, wenn x und y zueinander orthogonal sind. Allgemein definiert

$$\cos\theta \;=\; \frac{\langle x, y \rangle}{\|x\| \, \|y\|}$$

in einem reellen Innenproduktraum X den eingeschlossenen *Winkel* $\theta = \sphericalangle(x, y)$ zwischen zwei von Null verschiedenen Elementen $x, y \in X$. Sind $U, V \subset X$ zwei Unterräume des reellen Innenproduktraums X, so wird durch

$$\cos\theta \;=\; \sup_{\substack{0 \neq u \in U \\ 0 \neq v \in V}} \frac{\langle u, v \rangle}{\|u\| \, \|v\|}$$

der Winkel $\theta = \sphericalangle(U, V) \in [0, \pi/2]$ zwischen den beiden Unterräumen definiert. Diese Definition des Winkels läßt sich auf komplexe Innenprodukträume verallgemeinern, indem das Innenprodukt durch dessen Realteil ersetzt wird. Der Winkel zwischen zwei Unterräumen ist grundsätzlich kleiner oder gleich $\pi/2$.

Definition und Satz 31.3. *Zu jedem Innenprodukt $\langle \cdot, \cdot \rangle : X \times X \to \mathbb{K}$ gehört die Norm $\|f\| = \langle f, f \rangle^{1/2}$ in X.*

Beweis. Zum Beweis sind die drei Bedingungen aus Definition 2.1 zu überprüfen. Die Definitheit entspricht Bedingung (*iii*) aus Definition 31.1, während die Homogenität aus der Linearitätsbedingung (*ii*) bzw. ihrem Analogon für das erste Argument folgt: Ist $f \in X$ und $\alpha \in \mathbb{K}$, so gilt

$$\|\alpha f\|^2 \;=\; \langle \alpha f, \alpha f \rangle \;=\; \overline{\alpha}\alpha \langle f, f \rangle \;=\; |\alpha|^2 \|f\|^2 \,.$$

Die Dreiecksungleichung für $\|f + g\|$ mit $f, g \in X$ ergibt sich schließlich aus der Cauchy-Schwarz-Ungleichung:

$$\begin{aligned} \|f + g\|^2 \;&=\; \langle f + g, f + g \rangle \;=\; \|f\|^2 + \langle f, g \rangle + \langle g, f \rangle + \|g\|^2 \\ &=\; \|f\|^2 + 2 \operatorname{Re} \langle f, g \rangle + \|g\|^2 \;\leq\; \|f\|^2 + 2\|f\| \, \|g\| + \|g\|^2 \\ &=\; \big(\|f\| + \|g\|\big)^2 \,. \end{aligned}$$

□

Beispiele. Die folgenden Beispiele spielen in der Numerik eine wichtige Rolle:

1. $X = \mathbb{C}^{m\times n} \quad (m,n \in \mathbb{N})$:

Im Raum X der komplexwertigen $m \times n$-Matrizen wird für $A = [a_{ij}]$ und $B = [b_{ij}]$ ein Innenprodukt durch

$$\langle\!\langle\, A, B \,\rangle\!\rangle \;=\; \mathrm{Spur}(A^*B) \;=\; \sum_{i=1}^{m}\sum_{j=1}^{n} \overline{a_{ij}} b_{ij}$$

definiert. Wegen $\langle\!\langle\, A, A \,\rangle\!\rangle \;=\; \|A\|_F^2$ ist die Frobeniusnorm die zugehörige Norm.

2. $X = \mathcal{L}^2(\Omega) \quad (\Omega \subset \mathbb{R}^d)$:

Sei X der Raum aller komplexwertigen Funktionen $f : \Omega \to \mathbb{C}$, deren Betrag quadratisch integrierbar ist. In X wird das folgende Innenprodukt mit der entsprechenden Norm definiert:

$$\langle\, f, g \,\rangle_{\mathcal{L}^2(\Omega)} \;=\; \int_\Omega \overline{f(x)} g(x)\, dx\,, \qquad \|f\|_{\mathcal{L}^2(\Omega)} \;=\; \Big(\int_\Omega |f(x)|^2\, dt\Big)^{1/2}.$$

Der Raum X, versehen mit diesem Innenprodukt, heißt $\mathcal{L}^2(\Omega)$.

3. $X = \Pi_n$:

Sind x_i, $i = 0,\ldots,n$, paarweise verschiedene reelle und ω_i beliebige positive Gewichte, dann wird durch

$$\langle\, f, g \,\rangle \;=\; \sum_{i=0}^{n} \omega_i \overline{f(x_i)} g(x_i)$$

ein (diskretes) Innenprodukt im Raum Π_n aller (komplexen) Polynome vom Grad kleiner oder gleich n definiert. ◊

Die $\mathcal{L}^2$-Norm mißt nur die „absolute Größe“ einer Funktion, aber nicht ihre „Glattheit“. Zu diesem Zweck führen wir noch eine andere Norm ein:

Beispiel 31.4. Ist $\mathcal{I} = (a,b) \subset \mathbb{R}$ ein endliches Intervall, dann bezeichnet $H^1(\mathcal{I})$ den Raum aller komplexwertigen Funktionen $F \in \mathcal{L}^2(\mathcal{I})$ mit der folgenden Eigenschaft:

$$\text{Es gibt ein } c \in \mathbb{C} \text{ und ein } f \in \mathcal{L}^2(\mathcal{I}) \text{ mit } F(x) = c + \int_a^x f(t)\, dt\,. \tag{31.1}$$

Eine Funktion $F \in H^1(\mathcal{I})$ ist stetig in $\mathcal{I}$ (sogar absolutstetig) und kann stetig auf das abgeschlossene Intervall fortgesetzt werden. Dies liegt daran, daß die zugehörige Funktion f aus (31.1) zu $\mathcal{L}^2(\mathcal{I})$ gehört, also insbesondere meßbar

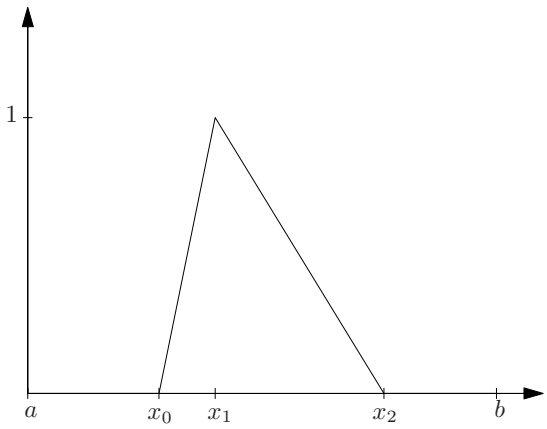


Abb. 31.1: Hutfunktion Λ

ist. Somit hängt das Integral $\int_a^x f(t)\,dt$ stetig von der oberen Integrationsgrenze x ab.[1]

Ist $f \in \mathcal{L}^2(\mathcal{I})$ wie in (31.1) mit $F \in H^1(\mathcal{I})$ assoziiert und $g \in \mathcal{L}^2(\mathcal{I})$ in entsprechender Weise mit $G \in H^1(\mathcal{I})$, dann sind

$$\begin{aligned} \langle F, G \rangle_{H^1(\mathcal{I})} &= \int_{\mathcal{I}} \overline{F(t)} G(t)\,dt + \int_{\mathcal{I}} \overline{f(t)} g(t)\,dt\,, \\ \|F\|_{H^1(\mathcal{I})} &= \Big(\int_{\mathcal{I}} |F(t)|^2\,dt + \int_{\mathcal{I}} |f(t)|^2\,dt \Big)^{1/2}, \end{aligned}$$

das zum Raum $H^1(\mathcal{I})$ gehörige Innenprodukt und die entsprechende Norm. Die Funktion f wird *schwache Ableitung* von F genannt und mit F' bezeichnet. Für differenzierbare Funktionen stimmt die schwache Ableitung nach dem Hauptsatz der Differential- und Integralrechnung mit der klassischen Ableitung überein. Der Raum $H^1(\mathcal{I})$ wird *Sobolevraum* genannt.

Als Beispiel einer nicht überall differenzierbaren H^1-Funktion betrachten wir die *Hutfunktion* $\Lambda \in H^1(a,b)$ aus Abbildung 31.1,

$$\Lambda(x) = \begin{cases} (x - x_0)/(x_1 - x_0)\,, & x_0 \le x < x_1\,, \\ (x - x_2)/(x_1 - x_2)\,, & x_1 \le x < x_2\,, \\ 0\,, & \text{sonst}\,, \end{cases}$$

mit $a \le x_0 < x_1 < x_2 \le b$. Man rechnet leicht anhand der Definition nach,

[1] Man kann zeigen, vgl. etwa Rudin [93, Satz 7.11], daß F fast überall differenzierbar ist und dort $F' = f$ gilt.

daß die schwache Ableitung von Λ durch die Funktion λ mit

$$\lambda(x) \;=\; \begin{cases} 1/(x_1-x_0)\,, & x_0 < x < x_1\,, \\ -1/(x_2-x_1)\,, & x_1 < x < x_2\,, \\ 0\,, & x < x_0 \quad \text{oder} \quad x > x_2\,. \end{cases}$$

gegeben ist. Die Werte $\lambda(x_i)$, $i = 0, 1, 2$, können beliebig festgelegt werden.

Man beachte, daß für die hermitesche Bilinearform

$$\langle\, F, G \,\rangle_{H^1_0(a,b)} = \int_a^b \overline{F'(t)} G'(t)\, dt\,, \qquad F, G \in H^1(a,b)\,, \tag{31.2}$$

die Eigenschaft (*iii*) aus Definition 31.1 für die konstanten Funktionen (und nur für diese) verletzt ist. Somit definiert diese hermitesche Bilinearform zwar kein Innenprodukt auf $H^1(a,b)$, aber auf den beiden Teilräumen

$$\begin{aligned} H^1_0(a,b) \;&=\; \{F \in H^1(a,b) \,:\, F(a) = F(b) = 0\} \qquad \text{und} \\ H^1_\diamond(a,b) \;&=\; \{F \in H^1(a,b) \,:\, \int_a^b F(x)\, dx = 0\}\,. \end{aligned}$$

In $H^1(a,b)$ induziert diese Bilinearform die *Halbnorm*

$$|F|_{H^1(a,b)} \;=\; \Big(\int_a^b |F|^2(x)\, dx \Big)^{1/2}$$

$\Diamond$

Ist $\langle\, \cdot\,, \cdot \,\rangle$ ein Innenprodukt in X, dann spricht man wie im $\mathbb{K}^n$ von einer *Orthonormalbasis* $\{\phi_i\}_{i=1}^n$ eines Teilraums $X_n \subset X$, falls

$$\langle\, \phi_i, \phi_j \,\rangle \;=\; \delta_{ij}\,, \qquad i, j = 1, \ldots, n\,.$$

Beispiel 31.5. Die Menge $\{1/\sqrt{2\pi}, (\sin kx)/\sqrt{\pi}, (\cos kx)/\sqrt{\pi} \;:\; 1 \le k \le n\}$ bildet eine Orthonormalbasis der sogenannten *reellen trigonometrischen Polynome* vom Grad n bezüglich $\mathcal{L}^2(-\pi, \pi)$, denn für $j \neq 0$ erhält man durch partielle Integration

$$\begin{aligned} \int_{-\pi}^{\pi} \cos jx \cos kx\, dx \;&=\; \frac{1}{j} \sin jx \cos kx \Big|_{-\pi}^{\pi} \;+\; \frac{k}{j} \int_{-\pi}^{\pi} \sin jx \sin kx\, dx \\ &=\; -\,\frac{k}{j^2} \cos jx \sin kx \Big|_{-\pi}^{\pi} \;+\; \frac{k^2}{j^2} \int_{-\pi}^{\pi} \cos jx \cos kx\, dx\,. \end{aligned}$$

Auf eine Seite gebracht ergibt sich also

$$\frac{j^2 - k^2}{j^2} \int_{-\pi}^{\pi} \cos jx \cos kx\, dx \;=\; 0\,,$$

und für $j \neq k$ liefert dies die Orthogonalität der Funktionen $\cos jx$ und $\cos kx$. Für den Fall $j = k$ bricht man die obigen Umformungen nach der ersten Zeile ab und erhält statt dessen

$$\int_{-\pi}^{\pi} \cos^2 jx\, dx \;=\; \int_{-\pi}^{\pi} \sin^2 jx\, dx \;=\; \int_{-\pi}^{\pi} (1 - \cos^2 jx)\, dx\,,$$

also $\|(\cos jx)/\sqrt{\pi}\|_{\mathcal{L}^2(-\pi,\pi)} = 1$. Die paarweise Orthogonalität der Sinusfunktionen und die Orthogonalität zwischen Sinus- und Kosinusfunktionen wird entsprechend bewiesen. ◊

Hat man erst einmal eine Orthonormalbasis eines endlichdimensionalen Teilraums $X_n \subset X$ zur Verfügung, dann gelten folgende Resultate:

Satz 31.6. *Sei $\{\phi_i\}_{i=1}^n$ eine Orthonormalbasis von $X_n \subset X$. Dann gilt:*

(*a*) *für $f \in X_n$ ist* $f = \sum_{i=1}^{n} \langle\, \phi_i, f \,\rangle\, \phi_i\,,$

(*b*) *für $f \in X_n$ ist* $\|f\|^2 = \sum_{i=1}^{n} |\,\langle\, \phi_i, f \,\rangle\,|^2,$

(*c*) *für $f \notin X_n$ ist* $f_n = \sum_{i=1}^{n} \langle\, \phi_i, f \,\rangle\, \phi_i$ *die Bestapproximation an f aus X_n:*

$$\|f - f_n\| \;<\; \|f - g\| \qquad \text{für alle } g \in X_n \setminus \{f_n\}\,,$$

(*d*) *für alle $f \in X$ ist* $\sum_{i=1}^{n} |\,\langle\, \phi_i, f \,\rangle\,|^2 \;\leq\; \|f\|^2\,.$

Bemerkung. Eigenschaft (*b*) entspricht dem *Satz von Pythagoras*; (*d*) wird *Besselsche Ungleichung* genannt. ◊

Beweis. (*a*) Nach Voraussetzung gilt $f = \sum_{i=1}^{n} \alpha_i \phi_i$ für gewisse $\alpha_i \in \mathbb{C}$. Also folgt

$$\langle\, \phi_j, f \,\rangle \;=\; \langle\, \phi_j, \sum_{i=1}^{n} \alpha_i \phi_i \,\rangle \;=\; \sum_{i=1}^{n} \alpha_i \underbrace{\langle\, \phi_j, \phi_i \,\rangle}_{\delta_{ij}} \;=\; \alpha_j\,. \tag{31.3}$$

(*b*) Aus (*a*) folgt

$$\begin{aligned} \|f\|^2 \;&=\; \langle\, f, f \,\rangle \;=\; \langle\, \sum_{i=1}^{n} \alpha_i \phi_i, \sum_{j=1}^{n} \alpha_j \phi_j \,\rangle \;=\; \sum_{i=1}^{n} \sum_{j=1}^{n} \overline{\alpha_i} \alpha_j \,\langle\, \phi_i, \phi_j \,\rangle \\ &=\; \sum_{i=1}^{n} |\alpha_i|^2 \;=\; \sum_{i=1}^{n} |\,\langle\, \phi_i, f \,\rangle\,|^2\,. \end{aligned}$$

(*c*) Wir definieren α_j, $j = 1, \ldots, n$, durch (31.3) und rechnen nach, daß

$$\begin{aligned}\|f - \sum_{i=1}^{n} \tilde{\alpha}_i \phi_i\|^2 &= \langle\, f - \sum_{i=1}^{n} \tilde{\alpha}_i \phi_i, f - \sum_{j=1}^{n} \tilde{\alpha}_j \phi_j \,\rangle \\ &= \|f\|^2 - \sum_{j=1}^{n} 2\,\mathrm{Re}\,(\tilde{\alpha}_j \langle\, f, \phi_j \,\rangle) + \sum_{i=1}^{n}\sum_{j=1}^{n} \overline{\tilde{\alpha}_i}\tilde{\alpha}_j \langle\, \phi_i, \phi_j \,\rangle \\ &\overset{(31.3)}{=} \|f\|^2 - 2\sum_{j=1}^{n} \mathrm{Re}\,(\tilde{\alpha}_j \overline{\alpha_j}) + \sum_{i=1}^{n} |\tilde{\alpha}_i|^2 \,.\end{aligned}$$

Durch quadratische Ergänzung ergibt sich daher

$$\|f - \sum_{i=1}^{n} \tilde{\alpha}_i \phi_i\|^2 = \|f\|^2 - \sum_{i=1}^{n} |\alpha_i|^2 + \sum_{i=1}^{n} |\tilde{\alpha}_i - \alpha_i|^2 \,, \tag{31.4}$$

und die rechte Seite von (31.4) wird genau dann minimal, wenn $\tilde{\alpha}_i = \alpha_i$ für $i = 1, \ldots, n$.

(*d*) Für $f \in X_n$ ist die Behauptung nach Teil (*b*) richtig. Für $f \in X \setminus X_n$ gilt nach (31.4) mit $\alpha_i = \langle\, \phi_i, f \,\rangle$

$$0 < \|f - \sum_{i=1}^{n} \alpha_i \phi_i\|^2 = \|f\|^2 - \sum_{i=1}^{n} |\alpha_i|^2 = \|f\|^2 - \sum_{i=1}^{n} |\langle\, \phi_i, f \,\rangle|^2 \,.$$

□

Dieser Satz macht klar, warum Orthogonalität in der Numerik so wichtig ist: Man hat einerseits einfache Basisdarstellungen und kann andererseits unmittelbar die Bestapproximation an eine vorgegebene Funktion berechnen. Sofern die vorliegende Basis von $X_n \subset X$ keine Orthonormalbasis ist, führt die Bestimmung der Bestapproximation auf die Lösung eines linearen Gleichungssystems:

Definition 31.7. Sei $\{\phi_i\}_{i=1}^n$ ein beliebiges Funktionensystem von $X_n \subset X$. Dann ist

$$G = \left[\, \langle\, \phi_i, \phi_j \,\rangle \,\right]_{i,j=1}^{n} \in \mathbb{K}^{n\times n}$$

die sogenannte *Gramsche Matrix* dieses Funktionensystems.

Proposition 31.8. *Eine Gramsche Matrix ist hermitesch und positiv semidefinit. Sie ist genau dann positiv definit, wenn das zugehörige Funktionensystem linear unabhängig ist.*

Beweis. Sei $\{\phi_i\}_{i=1}^n$ ein Funktionensystem und G die zugehörige Gramsche Matrix. Aus den definierenden Eigenschaften eines Innenprodukts folgt unmittelbar, daß G hermitesch ist. Für einen Vektor $x = [x_1, \ldots, x_n]^T \in \mathbb{K}^n$ gilt

ferner

$$\begin{aligned} x^*Gx &= [\overline{x_1},\dots,\overline{x_n}]\, G \begin{bmatrix} x_1 \\ \vdots \\ x_n \end{bmatrix} = [\overline{x_1},\dots,\overline{x_n}] \begin{bmatrix} \langle\, \phi_1, \sum_j x_j\phi_j \,\rangle \\ \vdots \\ \langle\, \phi_n, \sum_j x_j\phi_j \,\rangle \end{bmatrix} \\ &= \langle \sum_{i=1}^n x_i\phi_i \,,\, \sum_{j=1}^n x_j\phi_j \,\rangle = \| \sum_{i=1}^n x_i\phi_i \|^2 . \end{aligned}$$

Offensichtlich ist also $x^*Gx \geq 0$ für alle $x \in \mathbb{K}^n$ und Gleichheit gilt genau für das Nullelement $\sum_{i=1}^n x_i\phi_i = 0$. Falls $\{\phi_i\}_{i=1}^n$ ein linear unabhängiges Funktionensystem ist, so folgt aus letzterem $x_i = 0$ für alle $i = 1,\dots,n$. Ist umgekehrt $x^*Gx > 0$ für alle $x \neq 0$, dann ist die Menge $\{\phi_i\}$ linear unabhängig. □

Bemerkung 31.9. Allgemeiner gilt

$$x^*Gy = \langle \sum_{i=1}^n x_i\phi_i \,,\, \sum_{j=1}^n y_j\phi_j \,\rangle \qquad \text{für alle } x, y \in \mathbb{K}^n . \tag{31.5}$$

Sind also f und g durch ihre Entwicklungskoeffizenten $\{x_i\}$ bzw. $\{y_i\}$ bezüglich einer geeigneten Basis $\{\phi_i\}_{i=1}^n$ im Rechner dargestellt, dann kann das Innenprodukt $\langle\, f, g \,\rangle$ über die Gramsche Matrix mittels (31.5) ausgerechnet werden. ◊

Wir bestimmen nun die Bestapproximation aus $X_n \subset X$ an eine Funktion $f \in X$.

Satz 31.10. *Ist $f \in X$, $\{\phi_i\}_{i=1}^n$ eine Basis von $X_n \subset X$ und $G = \big[\langle\, \phi_i, \phi_j \,\rangle\big]_{ij}$ die zugehörige Gramsche Matrix. Dann ist $f_n = \sum_{i=1}^n x_i\phi_i$ genau dann die Bestapproximation an f aus X_n, wenn*

$$Gx = b \qquad \textit{für} \quad x = \begin{bmatrix} x_1 \\ \vdots \\ x_n \end{bmatrix} \;\; \textit{und} \;\; b = \begin{bmatrix} \langle\, \phi_1, f \,\rangle \\ \vdots \\ \langle\, \phi_n, f \,\rangle \end{bmatrix} .$$

Beweis. Unter Verwendung der angegebenen Vektoren x und b ergibt sich

$$\begin{aligned} \frac{1}{2}\| f - \sum_{i=1}^n x_i\phi_i \|^2 &= \frac{1}{2}\|f\|^2 - \operatorname{Re}\langle \sum_{i=1}^n x_i\phi_i, f \,\rangle + \frac{1}{2}\| \sum_{i=1}^n x_i\phi_i\|^2 \\ &= \frac{1}{2}\|f\|^2 - \operatorname{Re} x^*b + \frac{1}{2}x^*Gx . \end{aligned}$$

Nach Bemerkung 9.6 nimmt dieses Funktional sein Minimum für $x = G^{-1}b$ an. (Der Wert von $\|f\|$ beeinflußt nicht die Minimalstelle sondern nur den Wert des Minimums.) □

32 Tschebyscheff-Polynome

Die *Tschebyscheff-Polynome* sind gegeben durch

$$T_n(x) = \cos(n \arccos x)\,, \qquad -1 \le x \le 1\,. \tag{32.1}$$

Für $n = 0$ und $n = 1$ ergibt dies

$$T_0(x) = 1\,, \qquad T_1(x) = x\,,$$

also Polynome vom Grad 0 bzw. 1. Für beliebiges $n \in \mathbb{N}$ ist T_n jedoch nicht unmittelbar als Polynom identifizierbar. Um die Zugehörigkeit zum Raum der Polynome zu erkennen, verwendet man trigonometrische Identitäten und die Substitution $t = \arccos x$ und erhält

$$\begin{aligned} T_{n-1}(x) + T_{n+1}(x) \;&=\; \cos\big((n-1)t\big) \,+\, \cos\big((n+1)t\big) \\ &=\; \cos nt\, \cos t \,-\, \sin nt\, \sin(-t) \,+\, \cos nt\, \cos t \,-\, \sin nt\, \sin t \\ &=\; 2\cos(\arccos x)\cos(n \arccos x) \;=\; 2x\, T_n(x)\,. \end{aligned}$$

Folglich gilt die Rekursion

$$\begin{aligned} T_{n+1}(x) \;&=\; 2x\, T_n(x) - T_{n-1}(x)\,, \quad n = 1, 2, \dots\,, \\ T_0(x) \;&=\; 1\,, \qquad T_1(x) \;=\; x\,, \end{aligned} \tag{32.2}$$

und somit ist T_n ein Polynom vom Grad n, das als solches für alle $x \in \mathbb{R}$ bzw. $x \in \mathbb{C}$ wohldefiniert ist. Zudem ist T_n ein gerades Polynom, falls n gerade ist, und ein ungerades Polynom, falls n ungerade ist.

Wie wir nun sehen werden, bilden die Polynome $\{\, T_i \,\}_{i=0}^n$ bezüglich eines speziell gewichteten reellen $\mathcal{L}^2$-Innenprodukts eine Orthogonalbasis des Polynomraums Π_n.

Satz 32.1. *Es gilt*

$$\langle\, T_i, T_j \,\rangle := \int_{-1}^{1} T_i(x) T_j(x)\, \frac{1}{\sqrt{1-x^2}}\, dx \;=\; \begin{cases} 0\,, & i \ne j\,, \\ \pi\,, & i = j = 0\,, \\ \pi/2\,, & i = j \ne 0\,. \end{cases}$$

Beweis. Nach (32.1) ist

$$\langle\, T_i, T_j \,\rangle \;=\; \int_{-1}^{1} \cos(i \arccos x)\, \cos(j \arccos x)\, \frac{1}{\sqrt{1-x^2}}\, dx\,,$$

und mit der Substitution

$$x = \cos t\,, \quad dx = -\sin t\, dt = -\sqrt{1-x^2}\, dt\,, \qquad 0 \le t \le \pi\,,$$

ergibt sich aus der Symmetrie der Kosinusfunktionen und aus Beispiel 31.5

$$\langle T_i, T_j \rangle = \int_0^\pi \cos it \cos jt\, dt = \frac{1}{2} \int_{-\pi}^{\pi} \cos it \cos jt\, dt$$

$$= \begin{cases} 0\,, & i \neq j\,, \\ \pi\,, & i = j = 0\,, \\ \pi/2\,, & i = j \neq 0\,. \end{cases}$$

Damit ist die Behauptung bewiesen. □

Aus (32.1) folgt unmittelbar die Abschätzung

$$|T_n(x)| \leq 1\,, \qquad -1 \leq x \leq 1\,, \tag{32.3}$$

und aus der Rekursion (32.2) errechnet sich der Höchstkoeffizient:

$$T_n(x) = 2^{n-1}x^n + \ldots\,. \tag{32.4}$$

Zudem überprüft man leicht, daß alle n Nullstellen von T_n im Intervall $(-1, 1)$ liegen, vgl. Abbildung 33.1 und Aufgabe 6 (c).

Die Tschebyscheff-Polynome haben eine Reihe extremaler Eigenschaften, die sie für viele Approximationsprobleme in der Numerik interessant machen. Eine dieser Eigenschaften soll im folgenden exemplarisch vorgeführt werden.

Satz 32.2. *Sei* $\xi \notin [-1, 1]$: *Unter allen Polynomen* p_n *vom Grad* n *mit* $p_n(\xi) = 1$ *minimiert* $p_n = T_n/T_n(\xi)$ *die Maximumnorm* $\|p_n\|_{[-1,1]}$ *über* $[-1, 1]$.

Beweis. Wir nehmen an, daß ein Polynom $p_n \in \Pi_n$ mit $p_n(\xi) = 1$ existiert, welches eine kleinere Maximumnorm $\|p_n\|_{[-1,1]}$ hat als $t_n = T_n/T_n(\xi)$. Da t_n nach (32.1) an $n + 1$ aufeinanderfolgenden Stellen $x_k = \cos(k\pi/n) \in [-1, 1]$, $k = 0, \ldots, n$, die Extremalwerte $\pm 1/T_n(\xi)$ alternierend annimmt, während p_n aufgrund der getroffenen Annahme an all diesen Stellen betragsmäßig kleiner ist, hat $t_n - p_n$ mindestens n Vorzeichenwechsel, also n Nullstellen (vgl. Abbildung 32.1).

Wegen der Normierung hat $t_n - p_n$ darüber hinaus noch eine weitere Nullstelle im Punkt ξ. Da $t_n - p_n$ ein Polynom mit Grad kleiner gleich n ist, müssen t_n und p_n folglich identisch sein. Dies steht im Widerspuch zu der Annahme, daß die Maximumnorm von p_n kleiner als die von t_n ist. Damit ist der Satz bewiesen. □

Bemerkung. Satz 32.2 läßt sich auch folgendermaßen interpretieren: Unter allen Polynomen p_n vom Grad n mit $\|p_n\|_{[-1,1]} \leq 1$ wächst T_n außerhalb des Intervalls $[-1, 1]$ am schnellsten. Dies folgt für jedes beliebige $\xi \notin [-1, 1]$ durch

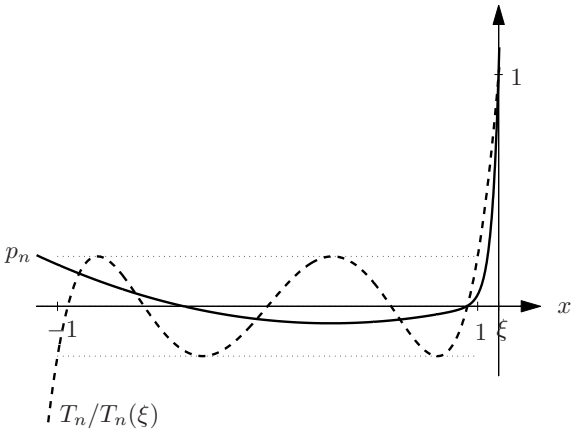


Abb. 32.1: Skizze zum Beweis von Satz 32.2.

Vergleich von $T_n/T_n(\xi)$ und $p_n/p_n(\xi)$ mit Hilfe von Satz 32.2: Aus ihm folgt nämlich

$$\frac{1}{|p_n(\xi)|} \;\geq\; \Big\|\frac{p_n}{p_n(\xi)}\Big\|_{[-1,1]} \;\geq\; \Big\|\frac{T_n}{T_n(\xi)}\Big\|_{[-1,1]} \;=\; \frac{1}{|T_n(\xi)|}\,,$$

und somit die gewünschte Behauptung. ◊

Die Orthogonalität der Tschebyscheff-Polynome und ihre Beschränkung durch Eins im Interval $[-1,1]$ machen sie zu einer günstigen Basis aller Polynome, besser etwa als die herkömmliche Basis der Monome x^k, $k = 0,\ldots,n$. Beachtet man, daß die orthonormierten Tschebyscheff-Polynome nach Satz 32.1 durch

$$u_0(x) = 1/\sqrt{\pi}\,, \qquad u_n(x) = \sqrt{2/\pi}\;T_n(x)\,, \quad n \in \mathbb{N}\,, \tag{32.5}$$

gegeben sind, so kann jedes Polynom $p_n \in \Pi_n$ nach Satz 31.6 unmittelbar in der Form

$$p_n(x) \;=\; \sum_{i=0}^{n} \langle\, u_i, p_n \,\rangle u_i(x) = \sum_{i=0}^{n} \alpha_i T_i(x) \tag{32.6}$$

geschrieben werden, wobei

$$\begin{aligned}
\alpha_0 \;&=\; \int_{-1}^{1} p_n(x)\frac{1}{\pi}\frac{1}{\sqrt{1-x^2}}\;dx\,,\\
\alpha_i \;&=\; \int_{-1}^{1} p_n(x)\frac{2}{\pi}\;T_i(x)\frac{1}{\sqrt{1-x^2}}\;dx\,, \quad i = 1,2,\ldots
\end{aligned}$$

Initialisierung: $p_n \in \Pi_n$ sei durch die Tschebyscheff-Entwicklung (32.6) gegeben

$y = 2x$
$\beta_n = \alpha_n$
$\beta_{n-1} = \alpha_{n-1} + y\beta_n$
`for` $i = n-2, n-3, \ldots, 1$ `do`
  $\beta_i = \alpha_i + y\beta_{i+1} - \beta_{i+2}$
`end for`

Ergebnis: $p_n(x) = \alpha_0 + x\beta_1 - \beta_2$.

Algorithmus 32.1: Clenshaw-Algorithmus

Diese Entwicklung hat den Vorteil, daß alle Entwicklungskoeffizienten gleichmäßig beschränkt sind:

$$|\alpha_0| \;=\; \Big|\int_{-1}^{1} p_n(x)\frac{1}{\pi}\frac{1}{\sqrt{1-x^2}}\,dx\Big| \;=\; \frac{1}{\pi}\,\Big|\int_0^{\pi} p_n(\cos t)\,dt\Big| \;\le\; \|p_n\|_{\,[-1,1]}\,,$$

$$\begin{aligned}|\alpha_i| \;&=\; \Big|\int_{-1}^{1} p_n(x)\frac{2}{\pi}\cos(i\arccos x)\frac{1}{\sqrt{1-x^2}}\,dx\Big| \\ &=\; \frac{2}{\pi}\,\Big|\int_0^{\pi} p_n(\cos t)\cos\,it\,dt\,\Big| \;\le\; \frac{4}{\pi}\,\|p_n\|_{\,[-1,1]}\,, \qquad 1 \le i \le n\,.\end{aligned}$$

Die Tschebyscheff-Entwicklung (32.6) läßt sich zudem sehr effizient mit Hilfe der Rekursionsformel (32.2) auswerten:

$$\begin{aligned}p_n(x) \;&=\; \sum_{i=0}^{n-1}\alpha_i T_i(x) + \underbrace{\alpha_n}_{=:\,\beta_n} T_n(x) \\ &\overset{(32.2)}{=} \sum_{i=0}^{n-3}\alpha_i T_i(x) + (\alpha_{n-2}-\beta_n)T_{n-2}(x) + \underbrace{(\alpha_{n-1}+2x\,\beta_n)}_{=:\,\beta_{n-1}} T_{n-1}(x) \\ &\overset{(32.2)}{=} \sum_{i=0}^{n-4}\alpha_i T_i(x) + (\alpha_{n-3}-\beta_{n-1})T_{n-3}(x) + \underbrace{(\alpha_{n-2}+2x\,\beta_{n-1}-\beta_n)}_{=:\,\beta_{n-2}} T_{n-2}(x) \\ &=\; \ldots \;=\; (\alpha_0-\beta_2)T_0(x) + \beta_1 T_1(x) \;=\; \beta_1 x + \alpha_0 - \beta_2\,.\end{aligned}$$

Dies ergibt den sogenannten *Clenshaw-Algorithmus* (Algorithmus 32.1) zur Auswertung von $p_n(x)$.

Aufwand. Der Clenshaw-Algorithmus ist mit $n+1$ Multiplikationen billiger als die Berechnung der einzelnen Funktionswerte $T_n(x)$ über ihre Rekursionsformel

und anschließende Summierung der Reihe (32.6). Algorithmus 32.1 ist zudem stabiler. ◊

33 Allgemeine Orthogonalpolynome

Eine *Gewichtsfunktion* w über einem Intervall $\mathcal{I} \subset \mathbb{R}$ ist eine positive und integrierbare Funktion über $\mathcal{I}$. Zu einer gegebenen Gewichtsfunktion kann das reelle Innenprodukt

$$\langle\, p, q \,\rangle \;=\; \int_{\mathcal{I}} p(x)q(x)w(x)\,dx \tag{33.1}$$

und die zugehörige gewichtete $\mathcal{L}^2$-Norm

$$\|p\|_{\,w} \;=\; \Big(\int_{\mathcal{I}} |p(x)|^2 w(x)\,dx\Big)^{1/2} \tag{33.2}$$

über dem Raum $\mathcal{L}^2(\mathcal{I})$ der reellwertigen quadratisch integrierbaren Funktionen definiert werden. Mit $\Pi_n \subset \mathcal{L}^2(\mathcal{I})$ wird weiterhin der Teilraum aller reellen Polynome vom Grad kleiner oder gleich n bezeichnet.

Satz 33.1. *Zu jeder Gewichtsfunktion w und zugehörigem Innenprodukt* (33.1) *existiert eine eindeutig bestimmte Folge $\{u_n\}_{n=0}^{\infty}$ mit*

$$u_n(x) \;=\; \gamma_n x^n + \ldots \;\in \Pi_n\,, \qquad \gamma_n > 0\,,$$

und $\langle\, u_n, u_m \,\rangle = \delta_{mn}$. Insbesondere ist

$$u_0 \;=\; \gamma_0 \;=\; \Big(\int_{\mathcal{I}} w(x)\,dx\Big)^{-1/2}.$$

Setzt man $u_{-1} = 0$, so gilt für $n \geq 0$ die dreistufige Rekursionsformel

$$\beta_{n+1}u_{n+1}(x) \;=\; xu_n(x) - \alpha_{n+1}u_n(x) - \beta_n u_{n-1}(x)\,, \tag{33.3}$$

wobei $\alpha_{n+1} = \langle\, u_n, xu_n \,\rangle$ und $\beta_{n+1} = \gamma_n/\gamma_{n+1}$ ist. Für β_0 sei der Wert Null vereinbart.

Beweis. Die Darstellung von u_0 ist offensichtlich. Für die höheren Polynomgrade gehen wir induktiv vor und nehmen an, $\{u_k\}_{k=0}^{n}$ sei eine Orthonormalbasis von Π_n mit den gewünschten Eigenschaften. Dann definieren wir

$$r_{n+1}(x) \;=\; xu_n(x) - \alpha_{n+1}u_n(x) - \beta_n u_{n-1}(x) \;=\; \gamma_n x^{n+1} + \ldots \;\in \Pi_{n+1} \tag{33.4}$$

und müssen zunächst zeigen, daß r_{n+1} orthogonal zu Π_n ist; wir weisen dies für jede Basisfunktion u_k, $k = 0, \ldots, n$, nach. Für $k = 0, \ldots, n-2$ ist

$$\begin{aligned} \langle\, u_k, r_{n+1} \,\rangle &= \langle\, u_k, xu_n \,\rangle - \alpha_{n+1} \underbrace{\langle\, u_k, u_n \,\rangle}_{=0} - \beta_n \underbrace{\langle\, u_k, u_{n-1} \,\rangle}_{=0} \\ &= \int_{\mathcal{I}} u_k(x) x u_n(x) w(x)\, dx \;=\; \langle\, xu_k, u_n \,\rangle\,, \end{aligned}$$

und da aufgrund der Einschränkung an k das Polynom $xu_k(x)$ zu Π_{n-1} gehört, verschwindet auch das letzte Innenprodukt.

Es verbleibt somit noch der Nachweis der Orthogonalität von r_{n+1} zu u_n und u_{n-1}, letzteres allerdings nur für $n \geq 1$. In diesem letzteren Fall ergibt sich zunächst wie zuvor

$$\begin{aligned} \langle\, u_{n-1}, r_{n+1} \,\rangle &= \langle\, u_{n-1}, xu_n \,\rangle - \alpha_{n+1} \underbrace{\langle\, u_{n-1}, u_n \,\rangle}_{=0} - \beta_n \underbrace{\langle\, u_{n-1}, u_{n-1} \,\rangle}_{=1} \\ &= \int_{\mathcal{I}} x u_{n-1}(x) u_n(x) w(x)\, dx \;-\; \beta_n\,. \end{aligned}$$

Nach (33.3) ist $xu_{n-1} = \beta_n u_n + q_{n-1}$ mit einem Polynom $q_{n-1} \in \Pi_{n-1}$ und es folgt

$$\langle\, u_{n-1}, r_{n+1} \,\rangle \;=\; \beta_n \underbrace{\langle\, u_n, u_n \,\rangle}_{=1} + \underbrace{\langle\, q_{n-1}, u_n \,\rangle}_{=0} - \beta_n \;=\; 0\,.$$

Für u_n ergibt sich schließlich wegen der Definition von α_{n+1}

$$\langle\, u_n, r_{n+1} \,\rangle = \underbrace{\langle\, u_n, xu_n \,\rangle}_{=\alpha_{n+1}} - \alpha_{n+1} \underbrace{\langle\, u_n, u_n \,\rangle}_{=1} - \beta_n \underbrace{\langle\, u_n, u_{n-1} \,\rangle}_{=0} = \alpha_{n+1} - \alpha_{n+1} = 0\,.$$

Somit steht $r_{n+1} \in \Pi_{n+1}$ senkrecht auf Π_n. Da der Höchstkoeffizient γ_n von r_{n+1} positiv ist, können wir $u_{n+1} = r_{n+1} / \|r_{n+1}\|_w \in \Pi_{n+1}$ setzen und auf diese Weise die Menge $\{u_k\}_{k=0}^n$ zu einer Orthonormalbasis von Π_{n+1} ergänzen. Bekanntlich ist diese Ergänzung bis auf das Vorzeichen eindeutig bestimmt. Nach (33.4) hat u_{n+1} den exakten Polynomgrad $n+1$ mit Höchstkoeffizienten

$$\gamma_{n+1} \;=\; \gamma_n \,/\, \|r_{n+1}\|_w \;>\; 0\,.$$

Damit ist u_{n+1} eindeutig festgelegt und wegen

$$r_{n+1} \;=\; \|r_{n+1}\|_w u_{n+1} \;=\; \frac{\gamma_n}{\gamma_{n+1}}\, u_{n+1} \;=\; \beta_{n+1} u_{n+1}$$

folgt aus (33.4) die gewünschte Rekursion (33.3). □

Bemerkung. Es sei darauf hingewiesen, daß sich die Beweise der Sätze 33.1 und 30.2 sehr ähneln. Dies ist kein Zufall, denn es gibt einen engen Bezug zwischen dem Lanczos-Prozeß und Orthogonalpolynomen; wir werden auf diesen Zusammenhang in Abschnitt 35.1 genauer eingehen. ◊

Beispiel. Aus (32.4) und (32.5) errechnen sich die Höchstkoeffizienten der orthonormierten Tschebyscheff-Polynome u_n zu

$$\gamma_0 = 1/\sqrt{\pi}\,, \qquad \gamma_n = 2^n/\sqrt{2\pi}\,, \quad n \in \mathbb{N}\,. \tag{33.5}$$

Damit ergeben sich aus Satz 33.1 die folgenden Koeffizienten für die Rekursionsformel der orthonormierten Tschebyscheff-Polynome:

$$\beta_1 = 1/\sqrt{2}\,, \qquad \beta_n = 1/2\,, \quad n \geq 2\,. \tag{33.6}$$

Da die Polynome $xu_n^2(x)$ für alle $n \in \mathbb{N}$ ungerade Funktionen sind, ist $\alpha_n = 0$ für alle $n \in \mathbb{N}$. ◊

Für viele Anwendungen ist die Gewichtsfunktion $w = 1$ über einem beschränkten Intervall von besonderem Interesse, da die zugehörige Norm mit der $\mathcal{L}^2$-Norm übereinstimmt. Für $\mathcal{I} = (-1, 1)$ führt dies auf die *Legendre-Polynome.*

Beispiel 33.2 (Legendre-Polynome). Die Legendre-Polynome $P_n \in \Pi_n$ sind durch

$$P_n(x) \;=\; \frac{1}{2^n n!}\,\frac{d^n}{dx^n}\,(x^2-1)^n \tag{33.7}$$

definiert. Ihre Orthogonalität über $(-1, 1)$ bezüglich der Gewichtsfunktion $w = 1$ wird folgendermaßen bewiesen: Für $n \geq m$ ist

$$\begin{aligned} 2^n n!\, 2^m m!\, \langle\, P_n, P_m \,\rangle \;&=\; \int_{-1}^{1} \frac{d^n}{dx^n}\,(x^2-1)^n\,\frac{d^m}{dx^m}\,(x^2-1)^m\,dx \\ &=\; \frac{d^{n-1}}{dx^{n-1}}\,(x^2-1)^n\,\frac{d^m}{dx^m}\,(x^2-1)^m\,\Big|_{-1}^{1} \\ &\qquad - \int_{-1}^{1} \frac{d^{n-1}}{dx^{n-1}}\,(x^2-1)^n\,\frac{d^{m+1}}{dx^{m+1}}\,(x^2-1)^m\,dx\,. \end{aligned}$$

Da $(x^2-1)^n$ jeweils n-fache Nullstellen in ± 1 besitzt, verschwinden die beiden Randterme und es folgt

$$2^n n!\, 2^m m!\, \langle\, P_n, P_m \,\rangle \;=\; -\int_{-1}^{1} \frac{d^{n-1}}{dx^{n-1}}\,(x^2-1)^n\,\frac{d^{m+1}}{dx^{m+1}}\,(x^2-1)^m\,dx\,.$$

Durch wiederholte partielle Integration ergibt sich schließlich in derselben Weise

$$2^n n!\, 2^m m!\, \langle\, P_n, P_m \,\rangle \;=\; (-1)^n \int_{-1}^{1} (x^2-1)^n \, \frac{d^{m+n}}{dx^{m+n}} \, (x^2-1)^m \, dx \,.$$

Für $n > m$ ist der Integrand dieses letzten Integrals identisch Null, da $(x^2-1)^m$ ein Polynom vom Grad $2m$ ist und dieses Polynom $m+n > 2m$ mal abgeleitet wird. Daher sind die Polynome P_n aus (33.7) paarweise orthogonal zueinander. Für $m = n$ ist die $2n$-te Ableitung von $(x^2-1)^n = x^{2n} + \dots$ die konstante Funktion mit Wert $(2n)!$ und durch n-malige partielle Integration erhält man

$$2^n n!\, 2^n n!\, \langle\, P_n, P_n \,\rangle \;=\; (-1)^n (2n)! \int_{-1}^{1} (x-1)^n (x+1)^n \, dx \;=\; (n!)^2 \frac{2^{2n+1}}{2n+1} \,.$$

Folglich ist $\langle\, P_n, P_n \,\rangle = 2/(2n+1)$, $n \in \mathbb{N}_0$, und

$$u_n \;=\; \big(\frac{2n+1}{2}\big)^{1/2} P_n \;=\; \big(\frac{2n+1}{2}\big)^{1/2} \frac{(2n)!}{2^n (n!)^2} \; x^n + \dots$$

ist das entsprechende orthonormierte Legendre-Polynom mit Höchstkoeffizient

$$\gamma_n \;=\; \big(\frac{2n+1}{2}\big)^{1/2} \frac{(2n)!}{2^n (n!)^2} \,. \tag{33.8}$$

Für die Koeffizienten β_n der Rekursionsformel (33.3) aus Satz 33.1 ergibt sich dann

$$\begin{aligned} \beta_n \;&=\; \frac{\gamma_{n-1}}{\gamma_n} \;=\; \frac{2^n (n!)^2}{(2n)!} \big(\frac{2}{2n+1}\big)^{1/2} \big(\frac{2n-1}{2}\big)^{1/2} \frac{(2n-2)!}{2^{n-1}\big((n-1)!\big)^2} \\ &=\; \big(\frac{2n-1}{2n+1}\big)^{1/2} \frac{2n^2}{2n(2n-1)} \;=\; \frac{n}{(4n^2-1)^{1/2}} \,, \qquad n \in \mathbb{N} \,. \end{aligned}$$

Wegen (33.7) ist P_n für gerade n ein gerades Polynom und für ungerade n ein ungerades Polynom; also ist wie bei den Tschebyscheff-Polynomen α_n immer gleich Null. Damit lautet die Rekursionsformel

$$\frac{n+1}{\big(4(n+1)^2-1\big)^{1/2}} \, u_{n+1}(x) \;=\; x u_n(x) - \frac{n}{(4n^2-1)^{1/2}} \, u_{n-1}(x) \tag{33.9}$$

für $n \in \mathbb{N}_0$ mit $u_{-1} = 0$ und $u_0 = 1/\sqrt{2}$. ◊

Für einen Vergleich der Legendre-Polynome mit den Tschebyscheff-Polynomen sei auf Abbildung 33.1 verwiesen.

In der Theorie allgemeiner Orthogonalpolynome, insbesondere bei der Untersuchung ihrer Nullstellen (vgl. den nachfolgenden Abschnitt), spielen noch zwei weitere Funktionenfolgen eine wichtige Rolle.

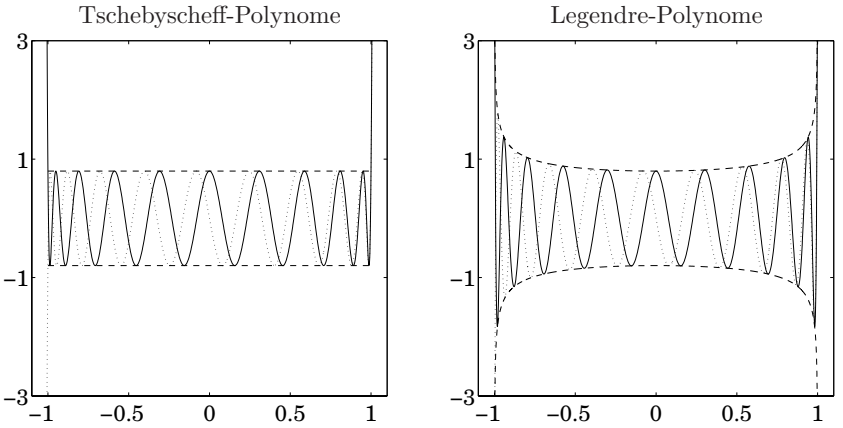


Abb. 33.1: Orthonormalpolynome vom Grad 20 (gepunktet: Grad 19).

Definition 33.3. Seien w eine Gewichtsfunktion und $\{u_n\}_{n=0}^{\infty}$ die zugehörigen Orthonormalpolynome aus Satz 33.1. Dann ist

$$K_n(\xi, x) \;=\; \sum_{i=0}^{n} u_i(\xi) u_i(x)$$

für festes $\xi \in \mathbb{R}$ und $n \in \mathbb{N}_0$ ein Polynom in x vom Grad n, das sogenannte *Kernpolynom* zu dem Innenprodukt (33.1). Die zugehörige rationale Funktion

$$\Lambda_n(x) \;=\; \frac{1}{K_{n-1}(x,x)} \;=\; \Big(\sum_{i=0}^{n-1} |u_i(x)|^2 \Big)^{-1}, \qquad n \geq 1\,,$$

heißt *Christoffel-Funktion*. Wegen $K_{n-1}(x,x) \geq |u_0(x)|^2 > 0$ ist Λ_n wohldefiniert.

Falls ξ nicht zu $\mathcal{I}$ gehört, bilden die Kernpolynome ebenfalls ein orthogonales System, und zwar zu der Gewichtsfunktion $\widetilde{w}(x) = |x - \xi| w(x)$ über $\mathcal{I}$. Dies wird im nächsten Abschnitt bewiesen (vgl. Satz 34.4). Die Kernpolynome haben darüber hinaus eine Extremaleigenschaft bezüglich der gewichteten $\mathcal{L}^2$-Norm (33.2), ganz ähnlich zu dem Resultat aus Satz 32.2 für die Tschebyscheff-Polynome in der Maximumnorm.

Satz 33.4. *Sei $\xi \in \mathbb{R}$. Das Minimum von $\|p_n\|_w$ unter allen Polynomen p_n vom Grad n mit $p_n(\xi) = 1$ hat den Wert $\Lambda_{n+1}^{1/2}(\xi)$ und dieser wird nur für $p_n = K_n(\xi, \cdot)/K_n(\xi,\xi)$ angenommen.*

Beweis. Sei $p_n \in \Pi_n$ mit $p_n(\xi) = 1$. Nach Satz 31.6 ist

$$p_n \;=\; \sum_{i=0}^{n} \langle\, u_i, p_n \,\rangle\, u_i\,, \qquad \|p_n\|_w^2 \;=\; \sum_{i=0}^{n} |\,\langle\, u_i, p_n \,\rangle\,|^2\,,$$

und aus der Cauchy-Schwarz-Ungleichung im $\mathbb{R}^{n+1}$ ergibt sich

$$\begin{aligned} 1 \;=\; p_n^2(\xi) \;&=\; \Big(\sum_{i=0}^{n} \langle\, u_i, p_n \,\rangle u_i(\xi)\Big)^2 \\ &\le\; \Big(\sum_{i=0}^{n} |\,\langle\, u_i, p_n \,\rangle\,|^2\Big)\Big(\sum_{i=0}^{n} |u_i(\xi)|^2\Big) \;=\; \|p_n\|_w^2 \Lambda_{n+1}^{-1}(\xi)\,. \end{aligned}$$

Dabei gilt das Gleichheitszeichen genau dann, wenn ein $c \in \mathbb{R}$ existiert mit $\langle\, u_i, p_n \,\rangle = c u_i(\xi)$, $i = 0, \ldots, n$. In diesem Fall ist also

$$p_n(x) \;=\; c \sum_{i=0}^{n} u_i(\xi) u_i(x) \;=\; c K_n(\xi, x)$$

und wegen $p_n(\xi) = 1$ ergibt sich $c = K_n^{-1}(\xi, \xi)$. Folglich ist $\|p_n\|_w^2 \ge \Lambda_{n+1}(\xi)$ und diese Schranke wird nur für $p_n = K_n(\xi, \cdot\,)/K_n(\xi, \xi)$ angenommen. □

Bemerkung. Satz 33.4 und sein Beweis bleiben auch für komplexe Polynome gültig, allerdings muß $u_i(\xi)$ in der Definition des Kernpolynoms komplex adjungiert werden. Hierauf werden wir später in den Anwendungen aus Abschnitt 35 zurück kommen. ◊

34 Nullstellen von Orthogonalpolynomen

Die Nullstellen der Orthogonalpolynome $\{u_n\}$ bezüglich einer Gewichtsfunktion w über $\mathcal{I}$ haben einige wichtige Eigenschaften: Sie sind allesamt reell, liegen im Innern von $\mathcal{I}$ und die Nullstellen aufeinanderfolgender Orthogonalpolynome trennen sich gegenseitig. Dies wollen wir im folgenden beweisen.

Satz 34.1. *Die Nullstellen der Orthogonalpolynome $\{u_n\}$ sind alle einfach und liegen im Innern von $\mathcal{I}$.*

Beweis. Wir nehmen an, die Aussage sei falsch. Hat u_n etwa eine Nullstelle z auf dem Rand von $\mathcal{I}$ oder in $\mathbb{R} \setminus \mathcal{I}$, dann ist

$$p_{n-1}(x) \;=\; \frac{u_n(x)}{x - z}$$

ein Polynom vom Grad $n-1$ und demnach ist

$$0 \;=\; \langle\, p_{n-1}, u_n \,\rangle \;=\; \int_{\mathcal{I}} \frac{u_n^2(x)}{x-z}\, w(x)\, dx\,.$$

Da jedoch $w(x)/(x-z)$ in $\mathcal{I}$ keinen Vorzeichenwechsel hat und $u_n^2(x)$ nichtnegativ und nicht identisch Null ist, ergibt dies einen Widerspruch.

Ist hingegen $z \in \mathcal{I}$ eine mehrfache Nullstelle von u_n oder liegt $z \;/\!\in \mathbb{R}$, dann wenden wir das entsprechende Argument auf

$$p_{n-2}(x) \;=\; \frac{u_n(x)}{(x-z)(x-\overline{z})} \;=\; \frac{u_n(x)}{|x-z|^2} \;\in \Pi_{n-2}$$

an. Dazu beachte man, daß nach Satz 33.1 u_n ein reelles Polynom ist, also mit $z \;/\!\in \mathbb{R}$ auch $\overline{z}$ eine Nullstelle von u_n ist; folglich ist p_{n-2} ein reelles Polynom vom Grad $n-2$. □

Bevor weitere Eigenschaften dieser Nullstellen hergeleitet werden, wenden wir uns noch einmal den Kernpolynomen zu.

Lemma 34.2 (Christoffel-Darboux-Identität). *Sind $\xi, x \in \mathbb{R}$ mit $\xi \neq x$, so gilt*

$$K_n(\xi, x) \;=\; \beta_{n+1}\, \frac{u_{n+1}(x)u_n(\xi) - u_n(x)u_{n+1}(\xi)}{x-\xi}\;. \tag{34.1}$$

Hierbei hat $\beta_{n+1} = \gamma_n/\gamma_{n+1} > 0$ die gleiche Bedeutung wie in der Rekursionsformel (33.3).

Beweis. Eine Anwendung der dreistufigen Rekursionsformel (33.3) auf $u_{n+1}(x)$ und $u_{n+1}(\xi)$ ergibt

$$\begin{aligned}
&\beta_{n+1}\big(u_{n+1}(x)u_n(\xi) - u_n(x)u_{n+1}(\xi)\big) \\
&\quad= \; u_n(\xi)\big(xu_n(x) - \alpha_{n+1}u_n(x) - \beta_n u_{n-1}(x)\big) \\
&\qquad\quad - u_n(x)\big(\xi u_n(\xi) - \alpha_{n+1}u_n(\xi) - \beta_n u_{n-1}(\xi)\big) \\
&\quad= \; (x-\xi)u_n(x)u_n(\xi) \;+\; \beta_n\big(u_n(x)u_{n-1}(\xi) - u_{n-1}(x)u_n(\xi)\big)\,.
\end{aligned}$$

Der zweite Summand kann in der gleichen Weise ersetzt werden und somit ergibt sich nach n entsprechenden Rekursionsschritten

$$\begin{aligned}
&\beta_{n+1}\big(u_{n+1}(x)u_n(\xi) - u_n(x)u_{n+1}(\xi)\big) \\
&\quad= \; \sum_{i=0}^{n} (x-\xi)u_i(x)u_i(\xi) \;+\; \beta_0\big(u_0(x)\,\underbrace{u_{-1}(\xi)}_{=0} - \underbrace{u_{-1}(x)}_{=0}\,u_0(\xi)\big) \\
&\quad= \; (x-\xi)K_n(\xi, x)\,.
\end{aligned}$$

□

Korollar 34.3. *Für alle $x \in \mathbb{R}$ ist*

$$K_n(x,x) \;=\; \beta_{n+1}\big(u'_{n+1}(x)u_n(x) - u_{n+1}(x)u'_n(x)\big)\,. \tag{34.2}$$

Beweis. Zum Beweis ersetzen wir in (34.1) den Zähler durch

$$\begin{aligned} &u_{n+1}(x)u_n(\xi) - u_n(x)u_{n+1}(\xi) \\ &\qquad = \big(u_{n+1}(x) - u_{n+1}(\xi)\big)u_n(\xi) \;-\; u_{n+1}(\xi)\big(u_n(x) - u_n(\xi)\big) \end{aligned}$$

und führen anschließend den Grenzübergang $\xi \to x$ durch. □

Nun können wir die bereits angesprochene Orthogonalität der Kernpolynome beweisen:

Satz 34.4. *Sei $\xi \in \mathbb{R} \setminus \mathcal{I}$. Dann sind die Kernpolynome $K_n(\xi,\,\cdot\,)$, $n \in \mathbb{N}_0$, paarweise orthogonal bezüglich $|x-\xi|w(x)$, d. h.*

$$\int_{\mathcal{I}} K_n(\xi,x)K_m(\xi,x)\,|x-\xi|w(x)\,dx \;=\; 0\,, \qquad n \neq m\,.$$

Beweis. Sei $n > m$. Da $K_m(\xi,\,\cdot\,)$ nach Definition 33.3 ein Polynom vom Grad $m < n$ ist, folgt aus der Identität (34.1) von Christoffel-Darboux, daß

$$\begin{aligned} &\int_{\mathcal{I}} K_n(\xi,x)K_m(\xi,x)|x-\xi|w(x)\,dx \\ &\qquad = \;\pm\beta_{n+1}\int_{\mathcal{I}}\big(u_{n+1}(x)u_n(\xi) - u_n(x)u_{n+1}(\xi)\big)K_m(\xi,x)w(x)\,dx \\ &\qquad = \;\pm\beta_{n+1}\Big(u_n(\xi)\underbrace{\langle\, u_{n+1}, K_m(\xi,\,\cdot\,)\,\rangle}_{=0} - u_{n+1}(\xi)\underbrace{\langle\, u_n, K_m(\xi,\,\cdot\,)\,\rangle}_{=0}\Big) \\ &\qquad = \;0\,. \end{aligned}$$

Dabei hängt das Vorzeichen davon ab, ob ξ rechts oder links des Intervalls $\mathcal{I}$ liegt. □

Nun zurück zu den Nullstellen von u_n und der sogenannten *Trennungseigenschaft*:

Satz 34.5. *Zwischen je zwei Nullstellen von u_{n+1} befindet sich genau eine Nullstelle von u_n.*

Beweis. Da β_{n+1} und $K_n(x,x)$ positiv sind (vgl. Definition 33.3), folgt aus (34.2), daß

$$u'_{n+1}(x)u_n(x) \;>\; u_{n+1}(x)u'_n(x) \qquad \text{für alle } x \in \mathbb{R}\,.$$

Insbesondere gilt dies für alle $n+1$ Nullstellen von u_{n+1}, die nach Satz 34.1 reell und einfach sind. An einer solchen Nullstelle ist also $u'_{n+1}(x)u_n(x)$ positiv, und da u'_{n+1} das Vorzeichen zwischen zwei aufeinanderfolgenden Nullstellen von u_{n+1} genau einmal wechselt, hat somit auch u_n mindestens einen Vorzeichenwechsel (also eine Nullstelle) zwischen je zwei aufeinanderfolgenden Nullstellen von u_{n+1}. Auf diese Weise finden sich alle n Nullstellen von u_n. □

Zur Illustration dieses Resultats sei noch einmal auf Abbildung 33.1 verwiesen.

Sind die Rekursionskoeffizienten der dreistufigen Rekursionsformel (33.3) gegeben, so lassen sich die Nullstellen der Orthogonalpolynome am stabilsten über die sogenannten *Jacobi-Matrizen*

$$J_n \;=\; \begin{bmatrix} \alpha_1 & \beta_1 & & 0 \\ \beta_1 & \alpha_2 & \ddots & \\ & \ddots & \ddots & \beta_{n-1} \\ 0 & & \beta_{n-1} & \alpha_n \end{bmatrix} \in \mathbb{R}^{n\times n} \tag{34.3}$$

berechnen, denn die Eigenwerte von J_n sind gerade die Nullstellen des n-ten Orthogonalpolynoms u_n:

Satz 34.6. *Die Nullstellen von u_n stimmen mit den Eigenwerten von J_n überein. Ist λ ein Eigenwert von J_n und $w = [w_1, \ldots, w_n]^T$ ein zugehöriger Eigenvektor mit $\|w\|_2 = 1$ und $w_1 \geq 0$, dann ist $w_k = \gamma u_{k-1}(\lambda)$, $k = 1, \ldots, n$, mit $\gamma = \Lambda_n^{1/2}(\lambda)$.*

Beweis. Wir wählen eine Nullstelle λ von u_n und setzen $w_k = \gamma u_{k-1}(\lambda)$ an, $k = 1, \ldots, n$, mit einer beliebigen Konstanten $\gamma \;/= 0$. Dann folgt aus der Rekursionsformel (33.3), daß die $k+1$-te Komponente, $k = 1, \ldots, n-2$, von $J_n w$ durch

$$\begin{aligned} (J_n w)_{k+1} &= \beta_k w_k + \alpha_{k+1} w_{k+1} + \beta_{k+1} w_{k+2} \\ &= \gamma\big(\, \beta_k u_{k-1}(\lambda) + \alpha_{k+1} u_k(\lambda) + \beta_{k+1} u_{k+1}(\lambda) \,\big) \\ &= \gamma\lambda u_k(\lambda) \;=\; \lambda w_{k+1} \end{aligned}$$

gegeben ist. Die gleiche Umformung bleibt aber auch für die erste ($k = 0$) und letzte ($k = n-1$) Komponente von $J_n w$ richtig, wenn man formal $w_0 = u_{-1}(\lambda) = 0$ und $w_{n+1} = u_n(\lambda) = 0$ setzt. Folglich ist λ ein Eigenwert von J_n und $w = [w_1, \ldots, w_n]^T$ ein zugehöriger Eigenvektor. Insbesondere ist $w_1 > 0$. Für den normierten Eigenvektor gilt

$$1 \;=\; \|w\|_2^2 \;=\; \gamma^2 \sum_{k=0}^{n-1} u_k^2(\lambda) \;=\; \gamma^2 / \Lambda_n(\lambda)$$

und somit ist $\gamma = \Lambda_n^{1/2}(\lambda)$. Da u_n n paarweise verschiedene Nullstellen besitzt, ergeben sich auf diese Weise alle Eigenwerte von J_n. □

Beispiel 34.7. Nach (33.6) ist

$$J_n = \begin{bmatrix} 0 & 1/\sqrt{2} & & & \\ 1/\sqrt{2} & 0 & 1/2 & & \\ & 1/2 & 0 & \ddots & \\ & & \ddots & \ddots & 1/2 \\ & & & 1/2 & 0 \end{bmatrix} \in \mathbb{R}^{n\times n} \tag{34.4}$$

die Jacobi-Matrix der orthonormierten Tschebyscheff-Polynome. Die Eigenwerte von J_n sind die Nullstellen von T_n, d. h.

$$\lambda_k = \cos\theta_k\,, \quad \theta_k = \frac{(2k-1)\pi}{2n}\,, \qquad k = 1,\dots,n\,,$$

vgl. Aufgabe 6. ◊

35 Anwendungen in der numerischen linearen Algebra

Die Resultate der vorangegangenen Abschnitte treffen auch auf *diskrete Innenprodukte* der Form

$$\langle\, p, q\,\rangle = \sum_{i=1}^{N} w_i p(\lambda_i) q(\lambda_i)\,, \qquad w_i > 0\,, \tag{35.1}$$

zu, wobei $N = \infty$ oder $N < \infty$ sein kann. Lediglich im endlichen Fall ($N < \infty$) muß bei den Orthonormalpolynomen u_n die Einschränkung $n < N$ vorgenommen werden. In diesem Fall liefert die rechte Seite der Rekursionsformel (33.4) für $n = N-1$ gerade das Polynom

$$r_N(x) = \gamma_{N-1} \prod_{i=1}^{N} (x - \lambda_i)\,,$$

vgl. Aufgabe 15. Im Hinblick auf (33.3) setzen wir in diesem Fall $\beta_N = 1$ und $u_N = r_N$.

Von den Ergebnissen aus Abschnitt 34 muß lediglich Satz 34.1 für diskrete Innenprodukte abgewandelt werden: Die Nullstellen der entsprechenden Orthonormalpolynome sind einfach und liegen in dem Intervall

$$\mathcal{I} = \operatorname{conv}\{\lambda_i : 1 \le i \le N\}\,, \tag{35.2}$$

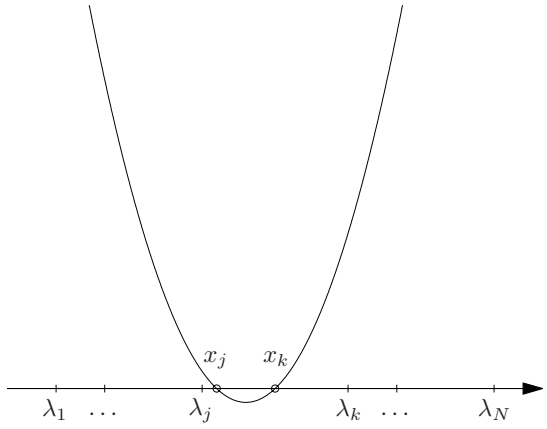


Abb. 35.1: Die Parabel $(\lambda - x_j)(\lambda - x_k)$

wobei $\mathrm{conv}(\mathcal{M})$ die konvexe Hülle der Menge $\mathcal{M}$ bezeichnet. Alle anderen Resultate bleiben unverändert gültig. Insbesondere trennen sich die Nullstellen zweier aufeinanderfolgender Orthogonalpolynome und die Kernpolynome $K_k(\xi, \cdot)$, $0 \le k \le N$, sind für $\xi \in \mathbb{R} \setminus \mathcal{I}$ ihrerseits Orthogonalpolynome bezüglich des diskreten Innenprodukts

$$\langle\, p, q \,\rangle' = \sum_{i=1}^{N} w_i |\lambda_i - \xi| p(\lambda_i) q(\lambda_i)\,. \tag{35.3}$$

Gegenüber dem in Abschnitt 34 behandelten Fall läßt sich für Innenprodukte der Form (35.1) die Lage der Nullstellen von u_n wie folgt präzisieren:

Satz 35.1. *Sind $\{u_i\}_{i=0}^N$ die Orthonormalpolynome für das Innenprodukt* (35.1) *und $n < N$, dann hat u_n zwischen je zwei aufeinanderfolgenden Knoten λ_j und λ_k höchstens eine Nullstelle.*

Beweis. Seien $\lambda_j < \lambda_k$ zwei aufeinanderfolgende Knoten von (35.1) und es sei $u_n(x_j) = u_n(x_k) = 0$ mit $\lambda_j < x_j < x_k < \lambda_k$. Dann ist

$$p_{n-2}(x) \;=\; \frac{u_n(x)}{(x - x_j)(x - x_k)}$$

ein Polynom vom Grad $n-2$, und an allen Knoten λ_i von (35.1) gilt

$$\frac{u_n^2(\lambda_i)}{(\lambda_i - x_j)(\lambda_i - x_k)} \;\ge\; 0\,, \qquad 1 \le i \le N\,,$$

da die Parabel $(\lambda - x_j)(\lambda - x_k)$ nur im Intervall (x_j, x_k) negativ ist (vgl. Abbildung 35.1). Dies ergibt jedoch einen Widerspruch, denn somit ist

$$0 \;=\; \langle\, p_{n-2}, u_n \,\rangle \;=\; \sum_{i=1}^{N} w_i \, \frac{u_n^2(\lambda_i)}{(\lambda_i - x_j)(\lambda_i - x_k)} \;\geq\; 0\,,$$

und Gleichheit kann in der letzten Ungleichung nur dann auftreten, wenn u_n an allen N Knoten verschwindet; dies ist jedoch wegen des Polynomgrads $n < N$ von u_n nicht möglich. □

Ist $N < \infty$, so hat u_{N-1} also zwischen je zwei aufeinanderfolgenden Knoten λ_i aus (35.1) genau eine Nullstelle. Aufgrund der Festlegung $u_N = \gamma_{N-1}\omega$ hat u_N genau die Knoten λ_i als Nullstellen.

Wir illustrieren nun die Bedeutung der in diesem Kapitel entwickelten Theorie für die Numerische Mathematik anhand dreier Beispiele aus der numerischen linearen Algebra. Die vielleicht wichtigste Anwendung, die Gauß-Quadratur, folgt am Ende des nächsten Kapitels.

35.1 Das Lanczos-Verfahren

Sei $A = A^* \in \mathbb{K}^{N\times N}$. Bei dem Lanczos-Prozeß aus Abschnitt 30 wird – ausgehend von einem Startvektor $z \in \mathbb{K}^n$ – durch

$$\begin{aligned} r_{i+1} \;&=\; (A - \alpha_i I)v_i - \beta_{i-1}v_{i-1}\,, \quad v_{i+1} = r_{i+1}/\,\|r_{i+1}\|_2\,, \\ &\qquad\qquad v_1 = z/\,\|z\|_2\,, \qquad v_0 = 0\,, \end{aligned} \tag{35.4}$$

$i = 1, \ldots, n-1$, mit

$$\alpha_i = v_i^* A v_i \qquad \text{und} \qquad \beta_i = \|r_{i+1}\|_2$$

zunächst eine Orthonormalbasis $\{v_1, \ldots, v_n\}$ des n-ten Krylovraums

$$\mathcal{K}_n(A, z) \;=\; \operatorname{span}\{z, Az, \ldots, A^{n-1}z\} \tag{35.5}$$

erzeugt.

Jedes Element $v \in \mathcal{K}_n(A, z)$ läßt sich mit geeigneten Koeffizienten $\pi_i \in \mathbb{K}$, $i = 0, \ldots, n-1$, entwickeln in

$$v \;=\; \sum_{i=0}^{n-1} \pi_i A^i z = \Big(\sum_{i=0}^{n-1} \pi_i A^i\Big)\, z\,. \tag{35.6}$$

Wir führen nun das Polynom

$$p(\lambda) \;=\; \sum_{i=0}^{n-1} \pi_i \lambda^i \tag{35.7}$$

vom Grad $n-1$ ein und schreiben $p(A)$ für die $N \times N$-Matrix, die sich ergibt, wenn anstelle von λ in (35.7) jeweils A eingesetzt wird; damit führt (35.6) auf die Darstellung

$$v = p(A)z\,.$$

Solange die Entwicklung (35.6) eindeutig ist, also solange $\dim \mathcal{K}_n(A,z) = n$ gilt, wird auf diese Weise jedem Element $v \in \mathcal{K}_n(A,z)$ in eindeutiger Weise ein Polynom p vom Grad $n-1$ zugeordnet und umgekehrt. Unter dieser Bedingung definiert

$$\langle\, p, q \,\rangle \;=\; \big(p(A)z\big)^* \big(q(A)z\big) \tag{35.8}$$

ein Innenprodukt in Π_{n-1}, denn dann ist

$$\langle\, p, p \,\rangle \;=\; \big(p(A)z\big)^* \big(p(A)z\big) \;=\; \|p(A)z\|_2^2 \tag{35.9}$$

für jedes $p \in \Pi_{n-1} \setminus \{0\}$ positiv.

Im weiteren wird der Einfachheit halber angenommen, daß alle Eigenwerte $\lambda_1 > \lambda_2 > \ldots > \lambda_N$ von A paarweise verschieden sind und $x_1, \ldots, x_N$ die zugehörigen orthonormierten Eigenvektoren von A bezeichnen. Dann kann der Startvektor in diese Basis entwickelt werden,

$$z \;=\; \sum_{i=1}^{N} \xi_i x_i \qquad \text{für gewisse } \xi_i \in \mathbb{R}\,,$$

und aus (35.8) folgt

$$\begin{aligned}\langle\, p, q \,\rangle &= \Big(p(A)\sum_{i=1}^{N} \xi_i x_i\Big)^* \Big(q(A)\sum_{j=1}^{N} \xi_j x_j\Big) = \Big(\sum_{i=1}^{N} p(\lambda_i)\xi_i x_i\Big)^* \Big(\sum_{j=1}^{N} q(\lambda_j)\xi_j x_j\Big) \\ &= \sum_{i,j=1}^{N} \overline{p(\lambda_i)}\, q(\lambda_j)\, \xi_i \xi_j\, x_i^* x_j \;=\; \sum_{i=1}^{N} |\xi_i|^2\, \overline{p(\lambda_i)}\, q(\lambda_i)\,.\end{aligned}$$

Mit anderen Worten: Das Innenprodukt (35.8) ist ein diskretes Innenprodukt der Form (35.1), die Eigenwerte von A sind die Knoten λ_i und die Gewichte $w_i = |\xi_i|^2$ sind die Eigenanteile des Startvektors.

Die Vektoren v_i, $i = 1, \ldots, n$, aus (35.4) können über (35.6), (35.7) mit Polynomen u_{i-1} vom Grad $i-1$ identifiziert werden. Aus der Orthonormalität der v_i folgt dann

$$\langle\, u_i, u_j \,\rangle \;=\; \big(u_i(A)z\big)^* \big(u_j(A)z\big) \;=\; v_{i+1}^* v_{j+1} \;=\; \delta_{ij}\,, \qquad i,j = 0, \ldots, n-1\,,$$

d. h. die Polynome $\{u_i\}_{i=0}^{n-1}$ sind gerade die Orthonormalpolynome für das diskrete Innenprodukt (35.8). Eingesetzt in (35.4) ergibt sich unmittelbar die Rekursion

$$\beta_i u_i(A)z \;=\; (A-\alpha_i I)u_{i-1}(A)z - \beta_{i-1}u_{i-2}(A)z\,,$$

d. h. die dreistufige Rekursionsformel dieser Orthonormalpolynome lautet

$$\beta_i u_i(x) \;=\; xu_{i-1}(x) - \alpha_i u_{i-1}(x) - \beta_{i-1}u_{i-2}(x)\,. \tag{35.10}$$

Insbesondere sind die u_i reellwertige Polynome, denn die Koeffizienten aus (35.10) sind reelle Zahlen. Aus (35.10) folgt schließlich, daß die Tridiagonalmatrix

$$T_n = \begin{bmatrix} \alpha_1 & \beta_1 & & 0 \\ \beta_1 & \alpha_2 & \ddots & \\ & \ddots & \ddots & \beta_{n-1} \\ 0 & & \beta_{n-1} & \alpha_n \end{bmatrix}$$

aus (30.7) gerade die Jacobi-Matrix zu der Rekursion (35.10) ist.

Das Lanczos-Verfahren bestimmt die Eigenwerte dieser Jacobi-Matrix als Näherungen an die Eigenwerte von A. Nach Satz 34.6 sind diese Eigenwerte gerade die Nullstellen von u_n. Wie wir gesehen haben, liegen diese Nullstellen in der konvexen Hülle der Knoten $\{\lambda_i\}_{i=1}^{N}$, also in der konvexen Hülle des Spektrums von A, und die Nullstellen von u_n und u_{n+1} (also die Eigenwertnäherungen des Lanczos-Verfahrens nach n und $n+1$ Iterationsschritten) trennen sich. Diese Ergebnisse entsprechen dem Hauptresultat aus Abschnitt 30 (Satz 30.1), das dort mit dem Satz von Courant und Fischer bewiesen wurde.

Darüber hinaus wissen wir jetzt aber noch etwas mehr:

(*i*) das Lanczos-Verfahren liefert nach n Iterationen n paarweise verschiedene Eigenwerte (Satz 34.1);
(*ii*) zwischen je zwei Eigenwerten von A liegt höchstens ein Eigenwert der Jacobi-Matrix T_n (Satz 35.1).

35.2 Das Bisektionsverfahren

Das Bisektionsverfahren bietet eine Möglichkeit, gezielt spezielle Eigenwerte einer Tridiagonalmatrix zu approximieren. Um dieses Verfahren herzuleiten, betrachten wir zunächst das zu dem vorigen Abschnitt inverse Problem:

Problem 35.2. *Gegeben sei die symmetrische Tridiagonalmatrix*

$$T = \begin{bmatrix} \alpha_1 & \beta_1 & & 0 \\ \beta_1 & \alpha_2 & \ddots & \\ & \ddots & \ddots & \beta_{N-1} \\ 0 & & \beta_{N-1} & \alpha_N \end{bmatrix} \in \mathbb{K}^{N\times N} \tag{35.11}$$

mit positiven Nebendiagonaleinträgen $\beta_1, \dots, \beta_{N-1}$. Gibt es dann ein Innenprodukt in Π_{N-1}, für das T die zugehörige Jacobi-Matrix ist?

Die Antwort auf diese Frage ist *„Ja“*. Zur Konstruktion dieses Innenprodukts benötigen wir die Spektralzerlegung $W\Lambda W^*$ von T, wobei W eine unitäre Matrix und Λ eine Diagonalmatrix ist, die die Eigenwerte λ_i, $i = 1, \dots, N$, von T auf der Diagonalen enthält. Die orthonormierten Eigenvektoren in den Spalten von $W = [w_{ij}]$ seien ferner so gewählt, daß die jeweils erste Komponente w_{1i}, $i = 1, \dots, N$, positiv ist. (Nach Lemma 29.2 ist die erste Eigenvektorkomponente einer irreduziblen symmetrischen Tridiagonalmatrix (35.11) immer von Null verschieden.)

Satz 35.3. *Seien α_i, $i = 1, \dots, N$, und β_i, $i = 1, \dots, N-1$, die Einträge der Tridiagonalmatrix T aus* (35.11)*. Ferner sei $\beta_N = 1$ gesetzt. Dann bilden die für $i = 1, \dots, N$ durch*

$$\beta_i u_i(x) \;=\; (x - \alpha_i)u_{i-1}(x) - \beta_{i-1}u_{i-2}(x) \;\in \Pi_i \tag{35.12}$$

definierten Polynome $\{u_i\}_{i=0}^{N-1}$ mit $u_{-1} = 0$ und $u_0 = 1$ eine Orthonormalbasis von Π_{N-1} bezüglich des Innenprodukts

$$\langle\, p, q \,\rangle \;=\; \sum_{i=1}^{N} w_{1i}^2 \, p(\lambda_i) q(\lambda_i)\,. \tag{35.13}$$

Das letzte Polynom $u_N \in \Pi_N$ hat Nullstellen in allen N Eigenwerten von T, die paarweise verschieden sind.

Beweis. Wir verwenden die oben eingeführte Spektralzerlegung von T. Dann erhält die j-te Spalte $w = [w_{1j}, \dots, w_{Nj}]^T$ von W einen (reellen) normierten Eigenvektor von T zum Eigenwert λ_j, und die einzelnen Zeilen der entsprechenden Gleichung $Tw = \lambda_j w$ lauten

$$\alpha_1 w_{1j} + \beta_1 w_{2j} \;=\; \lambda_j w_{1j}\,, \tag{35.14a}$$

$$\beta_{i-1} w_{i-1,j} + \alpha_i w_{ij} + \beta_i w_{i+1,j} \;=\; \lambda_j w_{ij}\,, \qquad 2 \le i \le N-1\,, \tag{35.14b}$$

$$\beta_{N-1} w_{N-1,j} + \alpha_N w_{N,j} \;=\; \lambda_j w_{Nj}\,. \tag{35.14c}$$

Wir zeigen nun induktiv, daß die Komponenten dieses Eigenvektors mit den Polynomen $\{u_i\}_{i=0}^{N-1}$ in folgendem Zusammenhang stehen:

$$w_{kj} = w_{1j} u_{k-1}(\lambda_j)\,, \qquad k = 1, \ldots, N\,. \tag{35.15}$$

zunächst ist die Aussage für $k = 1$ wegen $u_0 = 1$ offensichtlich. Hieraus folgt ferner mit (35.14a) und (35.12) für $i = 1$, daß

$$\beta_1 w_{2j} = (\lambda_j - \alpha_1) w_{1j} = (\lambda_j - \alpha_1) w_{1j} u_0(\lambda_j) = \beta_1 w_{1j} u_1(\lambda_j)\,,$$

da $u_{-1} = 0$ gesetzt wurde. Wegen $\beta_1 \neq 0$ gilt somit (35.15) auch für $k = 2$. Nehmen wir nun an, daß (35.15) für alle $k \leq i$ mit einem $i \in \{2, \ldots, N-1\}$ erfüllt ist. Aus (35.14b) erhalten wir dann

$$\begin{aligned} \beta_i w_{i+1,j} &= (\lambda_j - \alpha_i) w_{ij} - \beta_{i-1} w_{i-1,j} \\ &= (\lambda_j - \alpha_i) w_{1j} u_{i-1}(\lambda_j) - \beta_{i-1} w_{1j} u_{i-2}(\lambda_j) \\ &= w_{1j}\big((\lambda_j - \alpha_i) u_{i-1}(\lambda_j) - \beta_{i-1} u_{i-2}(\lambda_j)\big) \end{aligned}$$

und ein Vergleich mit (35.12) ergibt

$$\beta_i w_{i+1,j} = w_{1j} \beta_i u_i(\lambda_j).$$

Wie zuvor folgt nun die Induktionsbehauptung (35.15) mit $k = i+1$, da $\beta_i \neq 0$ ist.

Hieraus ergibt sich insbesondere, daß alle Eigenwerte von T paarweise verschieden sein müssen, da ansonsten zwei Eigenvektoren linear abhängig wären. Wegen $\beta_N = 1$ ist ferner

$$\begin{aligned} u_N(\lambda_j) &= \beta_N u_N(\lambda_j) \overset{(35.12)}{=} (\lambda_j - \alpha_N) u_{N-1}(\lambda_j) - \beta_{N-1} u_{N-2}(\lambda_j) \\ &\overset{(35.15)}{=} \frac{1}{w_{1j}} \left(\lambda_j w_{Nj} - \alpha_N w_{Nj} - \beta_{N-1} w_{N-1,j}\right) \overset{(35.14c)}{=} 0\,, \end{aligned}$$

das heißt, die Eigenwerte von T sind die Nullstellen von u_N.

Da W eine unitäre Matrix ist, sind die Zeilen von W paarweise orthogonal und es folgt

$$\begin{aligned} \delta_{jk} &= [w_{j1}, \ldots, w_{jN}] \begin{bmatrix} w_{k1} \\ \vdots \\ w_{kn} \end{bmatrix} \overset{(35.15)}{=} \sum_{i=1}^{N} w_{1i}^2 u_{j-1}(\lambda_i) u_{k-1}(\lambda_i) \\ &= \langle\, u_{j-1}, u_{k-1} \,\rangle\,. \end{aligned}$$

Die Polynome $\{u_i\}_{i=0}^{N-1}$ bilden also eine Orthonormalbasis von Π_{N-1} für das Innenprodukt (35.13). □

Bemerkung. Die Voraussetzung, daß die Nebendiagonaleinträge β_i von T positiv sind, ist nicht wesentlich; wichtig ist lediglich, daß sie von Null verschieden sind. Bei beliebigen Vorzeichen der β_i gilt Satz 35.3 weiterhin, lediglich die Höchstkoeffizienten von u_i, $i = 0, \ldots, N-1$, sind nicht länger allesamt positiv.

$\Diamond$

Wir können nun die Resultate über die Nullstellen orthogonaler Polynome ausnutzen, um die Eigenwerte von T zu lokalisieren. Dabei sei wieder vorausgesetzt, daß alle Nebendiagonalelemente von T positiv sind. Ferner seien

$$\lambda_n^{(n)} < \lambda_{n-1}^{(n)} < \ldots < \lambda_2^{(n)} < \lambda_1^{(n)}$$

die Nullstellen von u_n, $1 \le n \le N$. Zur Vereinfachung der weiteren Darstellung setzen wir noch $\lambda_{n+1}^{(n)} = -\infty$ und $\lambda_0^{(n)} = +\infty$.

Da alle Nullstellen von u_n einfach sind, wechselt u_n an jeder Nullstelle das Vorzeichen. Zudem konvergiert $u_n(x) \to +\infty$ für $x \to +\infty$, da der Höchstkoeffizient γ_n positiv ist; daher ist $u_n(x) > 0$ für $x > \lambda_1^{(n)}$. Daraus folgt

$$\operatorname{sign}\big(u_n(x)\big) = (-1)^{k-1}\,, \qquad \lambda_k^{(n)} < x < \lambda_{k-1}^{(n)}\,, \tag{35.16}$$

$1 \le k \le n+1$ und $n \le N$.

Satz 35.4 (Satz von Sturm). *Sei $x \in \mathbb{R}$ beliebig gewählt. Dann entspricht die Anzahl der Vorzeichenwechsel in der Folge $\{u_i(x)\}_{i=0}^N$ der Anzahl der Eigenwerte von T, die größer oder gleich x sind. Das Auftreten einer Null gilt als Vorzeichenwechsel; ein Übergang $\ldots, \pm, 0, \mp, \ldots$ zählt als* ein *Vorzeichenwechsel.*

Beweis. Der Beweis geht induktiv über die Dimension der Matrix T. Für $N = 1$ hat T lediglich einen Eigenwert, nämlich die Nullstelle $\lambda_1^{(1)}$ von u_1. Die Folge $\{u_0(x), u_1(x)\}$ besteht aus zwei Elementen: $u_0(x)$ ist aufgrund der Definition immer positiv; $u_1(x)$ ist positiv, falls $x > \lambda_1^{(1)}$ ist und nichtpositiv sonst. Im ersten Fall hat die Folge keinen Vorzeichenwechsel, im letzteren Fall genau einen. Also ist die Behauptung für $N = 1$ richtig.

Nehmen wir nun an, daß die Induktionsannahme für die Dimension $n = N-1$ richtig ist und betrachten für ein beliebiges k mit $1 \le k \le N$ ein x aus dem Intervall $(\lambda_k^{(n)}, \lambda_{k-1}^{(n)}]$. Da die linke obere $n \times n$-Submatrix von T nach Satz 35.3 die Eigenwerte $\lambda_i^{(n)}$, $1 \le i \le n$, besitzt, besagt die Induktionsannahme, daß die Folge $\{u_i(x)\}_{i=1}^n$ $k-1$ Vorzeichenwechsel hat. Wegen der Trennungseigenschaft der Nullstellen der Orthogonalpolynome (Satz 34.5) gilt

$$\lambda_{k+1}^{(N)} < \lambda_k^{(n)} < \lambda_k^{(N)} < \lambda_{k-1}^{(n)} < \lambda_{k-1}^{(N)}\,,$$

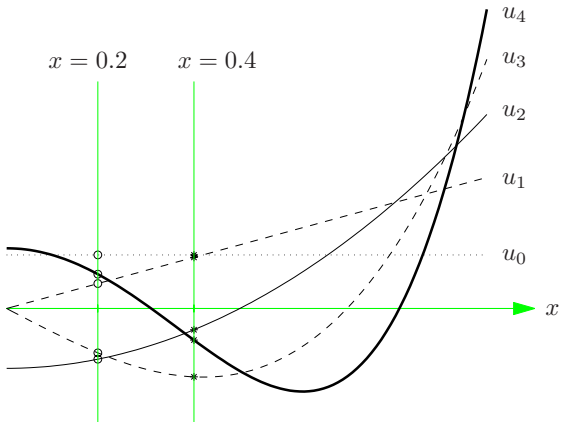


Abb. 35.2: Die Legendre-Polynome und der Satz von Sturm

so daß nur zwei Möglichkeiten für die Position von x denkbar sind:

$$(a) \quad \lambda_k^{(n)} < x \leq \lambda_k^{(N)},$$

$$(b) \quad \lambda_k^{(N)} < x \leq \lambda_{k-1}^{(n)}.$$

Im Fall (a) ist $\operatorname{sign}\big(u_N(x)\big) = (-1)^k = -\operatorname{sign}\big(u_n(x)\big)$ oder gleich Null, d. h. die Folge $\{u_i(x)\}_{i=0}^N$ besitzt k Vorzeichenwechsel; im zweiten Fall ist $\operatorname{sign}\big(u_N(x)\big) = (-1)^{k-1} = \operatorname{sign}\big(u_n(x)\big)$ und die erweiterte Folge hat weiterhin $k-1$ Vorzeichenwechsel. □

Beispiel. Abbildung 35.2 zeigt die Legendre-Polynome $u_0, \ldots, u_4$. An den Stellen $x = 0.2$ (durch Kreise gekennzeichnet) und $x = 0.4$ (durch Sterne gekennzeichnet) ergeben sich die folgenden Vorzeichen:

x	$u_0(x)$	$u_1(x)$	$u_2(x)$	$u_3(x)$	$u_4(x)$
0.2	+	+	−	−	+
0.4	+	+	−	−	−

Folglich hat die Folge $\{u_i(0.2)\}_{i=0}^4$ zwei Vorzeichenwechsel, während die Folge $\{u_i(0.4)\}_{i=0}^4$ nur einen Vorzeichenwechsel aufweist. Daher liegt die zweitgrößte Nullstelle von u_4 zwischen 0.2 und 0.4. ◊

Prinzipiell könnte die Folge $\{u_i(x)\}_{i=0}^N$ für $x \in \mathbb{R}$ über die Rekursionsformel (35.12) ausgewertet werden, um dann anhand der Vorzeichenwechsel zu entscheiden, ob x eine obere oder untere Schranke für den k-größten Eigenwert von T ist. Leider ist diese Rekursion sehr instabil. In der Praxis geht man

Initialisierung: Das Intervall (x_u, x_o) enthalte einen Eigenwert der Matrix T aus (35.11)

```
repeat
  λ = (x_u + x_o)/2
  q_1(λ) = α_1 − λ
  if q_1(λ) ≥ 0 then      % Vorzeichenwechsel
    j = 1
  else
    j = 0
  end if
  for i = 1, 2, ..., N − 1 do
    q_{i+1}(λ) = α_{i+1} − λ − β_i^2/q_i(λ)
    if q_{i+1}(λ) ≥ 0 then      % Vorzeichenwechsel
      j = j + 1
    end if
  end for      % j ist die Anzahl der Vorzeichenwechsel
  if j ≥ k then
    x_u = λ
  else
    x_o = λ
  end if
until x_o − x_u hinreichend klein.
```

Ergebnis: (x_u, x_o) enthält nach wie vor einen Eigenwert von T

Algorithmus 35.1: Bisektionsverfahren

daher anders vor: Für die Quotienten

$$q_i(x) \;=\; -\beta_i u_i(x)/u_{i-1}(x)\,, \qquad i = 1, 2, \ldots\,, \tag{35.17}$$

ergibt sich entsprechend aus (35.12) $q_1(x) = -\beta_1 u_1(x) = \alpha_1 - x$, und für $i \geq 1$ folgt die Rekursion

$$\begin{aligned} q_{i+1}(x) &= -\beta_{i+1}u_{i+1}(x)/u_i(x) \;=\; \alpha_{i+1} - x + \beta_i u_{i-1}(x)/u_i(x) \\ &= \alpha_{i+1} - x - \beta_i^2/q_i(x)\,. \end{aligned} \tag{35.18}$$

Interessanterweise kann die Folge $\{q_i(x)\}_{i=1}^N$ durch Auswertung der Rekursion (35.18) stabil berechnet werden, obwohl die Nenner $q_i(x)$ auf der rechten Seite im Verlauf der Rechnung beliebig klein werden können. Die Rechnerarithmetik muß lediglich Underflow und Overflow bzw. Division durch Null abfangen können. Dieses Stabilitätsresultats wird in dem Buch von Demmel [22] erläutert.

Da alle β_i positiv sind, entspricht ein Vorzeichenwechsel von $u_{i-1}(x)$ nach $u_i(x)$ einem Wert $q_i(x) \in \mathbb{R}_0^+$, während kein Vorzeichenwechsel einem Wert $q_i(x) \in$

$\mathbb{R}^- \cup \{\pm\infty\}$ entspricht. Auf dem Satz von Sturm beruht Algorithmus 35.1 (das sogenannte *Bisektionsverfahren*) zur Bestimmung des k-größten Eigenwerts λ_k von T.

Aufwand. Die Konvergenz von Algorithmus 35.1 ist linear, der Konvergenzfaktor ist $q = 1/2$. Jeder Iterationsschritt kann mit N Divisionen realisiert werden, wenn die Quadrate β_i^2 im voraus berechnet werden. $\Diamond$

35.3 Das CG-Verfahren

Als letztes Anwendungsbeispiel der Orthogonalpolynomtheorie greifen wir noch einmal das CG-Verfahren aus Abschnitt 9 zur Lösung eines linearen Gleichungssystems $Ax = b$ auf, wobei $b \in \mathbb{K}^N$ und $A \in \mathbb{K}^{N\times N}$ hermitesch und positiv definit sei. Die Lösung dieses Gleichungssystems wird wieder mit $\widehat{x}$ bezeichnet.

Wir können uns dabei auf den Fall beschränken, daß als Startvektor $x^{(0)} = 0$ gewählt wird. Wie wir in Satz 9.5 gesehen haben, bestimmt das CG-Verfahren in diesem Fall nach k Iterationsschritten eine Näherung $x^{(k)}$ in dem Krylovraum

$$\mathcal{K}_k(A, b) \;=\; \operatorname{span}\{b, Ab, \ldots, A^{k-1}b\}\,,$$

wobei $x^{(k)}$ unter allen Elementen dieser Menge ein Zielfunktional Φ minimiert, das bis auf eine additive Konstante durch

$$\Phi(x) \;=\; \|x - \widehat{x}\|_A^2 \;=\; (x - \widehat{x})^* A (x - \widehat{x})$$

gegeben ist, vgl. (9.14).

Ein gegebenes Element $x \in \mathcal{K}_k(A, b)$ läßt sich wie in (35.6) und (35.7) schreiben als

$$x \;=\; q(A)b \tag{35.19}$$

mit einem (ggf. komplexen) Polynom q vom Grad $k - 1$. Von Bedeutung ist auch das zugehörige Polynom

$$p(\lambda) \;=\; 1 - \lambda q(\lambda) \tag{35.20}$$

vom Grad k, denn einerseits ist

$$\widehat{x} - x \;=\; \widehat{x} - q(A)b \;=\; \big(I - q(A)A\big)\widehat{x} \;=\; p(A)\widehat{x} \tag{35.21}$$

und andererseits ist

$$b - Ax = b - Aq(A)b = (I - Aq(a))b = p(A)b\,. \tag{35.22}$$

Man beachte, daß aufgrund der Definition (35.20) immer

$$p(0) = 1 \tag{35.23}$$

gilt. Entwickeln wir schließlich $\widehat{x}$ in die Eigenbasis von A,

$$\widehat{x} \;=\; \sum_{i=1}^{N} \xi_i v_i \,, \qquad Av_i \;=\; \lambda_i v_i \,,\; v_i^* v_j \;=\; \delta_{ij} \,,\; i,j = 1,\dots,N \,,$$

dann folgt aus (35.21), daß

$$\begin{aligned} \Phi(x) \;&=\; (x-\widehat{x})^* A (x-\widehat{x}) \;=\; \big(p(A)\widehat{x}\big)^* \big(Ap(A)\widehat{x}\big) \\ &=\; \sum_{i,j=1}^{N} \overline{p(\lambda_i)\xi_i}\, \lambda_j p(\lambda_j) \xi_j \, v_i^* v_j \;=\; \sum_{i=1}^{N} \lambda_i |\xi_i|^2 \, |p(\lambda_i)|^2 \,. \end{aligned} \tag{35.24}$$

Dies ist das Quadrat einer diskreten Norm von p und nach Satz 33.4 wird dieser Ausdruck unter der Nebenbedingung (35.23) durch das entsprechende Kernpolynom

$$p_k \;=\; \frac{K_k(0,\,\cdot\,)}{K_k(0,0)} \;\in \Pi_k$$

minimiert.

Da auch das zu der CG-Iterierten gehörende Polynom $p \in \Pi_k$ den Ausdruck (35.24) minimiert und das minimierende Polynom eindeutig bestimmt ist (vgl. Satz 33.4) müssen p und p_k übereinstimmen. Ferner wissen wir aus (35.3), daß die Kernpolynome $K_k(0,\,\cdot\,)$, $k \in N_0$, paarweise orthogonal sind, und zwar bezüglich des diskreten Innenprodukts

$$\langle\, p, q \,\rangle \;=\; \sum_{i=1}^{N} \lambda_i^2 |\xi_i|^2 \, \overline{p(\lambda_i)} q(\lambda_i) \;=\; \big(p(A)b\big)^* \big(q(A)b\big) \,. \tag{35.25}$$

Damit haben wir das folgende Resultat bewiesen:

Proposition 35.5. *Die Iterierten des CG-Verfahrens mit Startvektor $x^{(0)} = 0$ können in der Form*

$$x^{(k)} \;=\; q_{k-1}(A)b \,, \qquad k = 0,\dots,N \,,$$

mit $q_{k-1} \in \Pi_{k-1}$ geschrieben werden. Die zugehörigen Polynome $p_k(\lambda) = 1 - \lambda q_{k-1}(\lambda)$, $k = 0,\dots,N$, bilden die durch (35.23) *normierten Orthogonalpolynome zu dem Innenprodukt* (35.25) *und minimieren das quadratische Funktional* (35.24). *Wegen* (35.22) *werden diese Polynome* Residuenpolynome *genannt.*

Bemerkung 35.6. Zwischen dem CG-Verfahren und dem Lanczos-Verfahren besteht ein enger Zusammenhang. Wird nämlich für das Lanczos-Verfahren der Startvektor $z = b$ gewählt, dann stimmen die Innenprodukte (35.8) und (35.25) überein. Somit stimmen auch die Iterierten $v_{k+1} = u_k(A)b$ aus dem Lanczos-Verfahren und die Residuenvektoren $b - Ax^{(k)} = p_k(A)b$ des CG-Verfahrens bis auf ein Vielfaches miteinander überein. Anders ausgedrückt: Die Nullstellen des Polynoms p_k aus dem CG-Verfahren sind die Näherungseigenwerte nach k Schritten des Lanczos-Verfahrens. Dieser Zusammenhang kann auch numerisch genutzt werden: Die Implementierung des Lanczos-Prozesses kann so erweitert werden, daß gleichzeitig die CG-Iterierten mitberechnet werden. Der resultierende Algorithmus SYMMLQ stammt von Paige und Saunders [80]. Der Aufwand von SYMMLQ ist nur unwesentlich größer als der des CG-Verfahrens.

◊

Über die polynomiale Interpretation der Minimierungseigenschaft des CG-Verfahrens kann nun auch eine Abschätzung der Konvergenzgeschwindigkeit hergeleitet werden.

Satz 35.7. *Sei $A \in \mathbb{K}^{N\times N}$ hermitesch und positiv definit mit $\sigma(A) \subset [a,d] \cup \{\lambda_1, \ldots, \lambda_m\}$. Hierbei sei $m < N$ und $0 < a < d < \lambda_m \le \ldots \le \lambda_1$. Dann gilt für den Fehler nach $k \ge m$ Schritten des CG-Verfahrens die Abschätzung*

$$\|\widehat{x} - x^{(k)}\|_2 \;\le\; 2\sqrt{d/a}\, q^{k-m}\, \|\widehat{x}\|_2 \quad \textit{mit} \quad q \;=\; \frac{\sqrt{d/a}-1}{\sqrt{d/a}+1}\,. \tag{35.26}$$

Beweis. Die affin-lineare Transformation

$$z(\lambda) \;=\; \frac{d+a-2\lambda}{d-a}\,, \qquad \lambda \in \mathbb{R}\,,$$

bildet das Intervall $[a,d]$ auf das Intervall $[-1,1]$ ab. Wir setzen nun

$$\omega(\lambda) \;=\; \prod_{j=1}^{m} \frac{\lambda_j - \lambda}{\lambda_j} \qquad \text{und} \qquad p(\lambda) \;=\; \frac{T_{k-m}\big(z(\lambda)\big)}{T_{k-m}\big(z(0)\big)}\,\omega(\lambda)$$

wobei T_{k-m} das $(k-m)$-te Tschebyscheff-Polynom bezeichnet. Wegen $z(0) = (d+a)/(d-a) > 1$ ist der Nenner ungleich Null und somit p ein Polynom vom Grad k mit $p(0) = 1$. Der hintere Faktor ω verschwindet an den Eigenwerten $\lambda_1, \ldots, \lambda_m$ von A und aufgrund der Anordnung dieser Eigenwerte liegt $\omega(\lambda)$ zwischen Null und Eins für $\lambda \in [a,d]$. Daraus folgt die Abschätzung

$$\max_{\lambda\in\sigma(A)} |p(\lambda)| \;\le\; \max_{\lambda\in[a,d]} |p(\lambda)| \;\le\; \Big|\, T_{k-m}\big(\frac{d+a}{d-a}\big)\,\Big|^{-1} \max_{-1\le z\le 1} |T_{k-m}(z)|\,,$$

wobei das letzte Maximum Eins ist, vgl. (32.3). Nun verwenden wir die Minimalitätseigenschaft der Residuenpolynome $\{p_k\}$ des CG-Verfahrens bezüglich des Funktionals (35.24): Wegen $p \in \Pi_k$ mit $p(0) = 1$ und $p(\lambda_j) = \omega(\lambda_j) = 0$ für $j = 1, \ldots, m$ folgt daraus

$$\begin{aligned}\sum_{i=1}^{N} \lambda_i |\xi_i|^2 \, |p_k(\lambda_i)|^2 \;&\le\; \sum_{i=1}^{N} \lambda_i |\xi_i|^2 \, |p(\lambda_i)|^2 \;=\; \sum_{i=m+1}^{N} \lambda_i |\xi_i|^2 \, |p(\lambda_i)|^2 \\ &\le\; d \, \big|T_{k-m}\big(\frac{d+a}{d-a}\big)\big|^{-2} \sum_{i=m+1}^{N} |\xi_i|^2 \;\le\; d \, \big|T_{k-m}\big(\frac{d+a}{d-a}\big)\big|^{-2} \, \|\widehat{x}\|_2^2 \,.\end{aligned}$$

Dabei haben wir verwendet, daß die restlichen Eigenwerte $\lambda_{m+1}, \ldots, \lambda_N$ von A im Intervall $[a, d]$ liegen. Hieraus folgt nun mit (35.21)

$$\begin{aligned}\|\widehat{x} - x^{(k)}\|_2^2 \;&=\; \big(p_k(A)\widehat{x}\big)^* \big(p_k(A)\widehat{x}\big) \;=\; \sum_{i,j=1}^{N} \overline{\xi_i} \xi_j \, \overline{p_k(\lambda_i)} p_k(\lambda_j) \, v_i^* v_j \\ &=\; \sum_{i=1}^{N} |\xi_i|^2 \, |p_k(\lambda_i)|^2 \;\le\; \frac{1}{a} \sum_{i=1}^{N} \lambda_i |\xi_i|^2 \, |p_k(\lambda_i)|^2 \\ &\le\; \frac{d}{a} \, \big|T_{k-m}\big(\frac{d+a}{d-a}\big)\big|^{-2} \, \|\widehat{x}\|_2^2 \,.\end{aligned}$$

Für den Wert des Tschebyscheff-Polynoms haben wir schließlich nach Aufgabe 6 die Abschätzung

$$T_{k-m}\big(\frac{d+a}{d-a}\big) \;\ge\; \frac{1}{2} \, q^{m-k}$$

mit q aus (35.26) und daraus folgt dann unmittelbar die Behauptung. □

Wir interpretieren Satz 35.7 zunächst für den Fall $m = 0$, d. h. für $[a, d] = [\lambda_N, \lambda_1]$. In diesem Fall bestimmt im wesentlichen der Quotient λ_1/λ_N (also die Kondition $\mathrm{cond}_2(A)$ der Matrix, vgl. Aufgabe I.14 (b)) die Konvergenzgeschwindigkeit des CG-Verfahrens: Die Konvergenz ist demnach mindestens linear mit Konvergenzfaktor

$$q \;=\; 1 - \frac{2}{1 + \mathrm{cond}_2^{1/2}(A)} \;\approx\; 1 - 2 \, \mathrm{cond}_2^{-1/2}(A) \,.$$

Diese Abschätzung ist relativ scharf, wenn das Spektrum von A das Intervall $[\lambda_N, \lambda_1]$ „dicht ausfüllt“ (vgl. Beispiel 9.7 und die zugehörige Diskussion zu Beginn von Abschnitt 10).

Der Fall $m > 0$ bezieht sich auf die Konstellation, in der lediglich die kleineren Eigenwerte $\lambda_N, \ldots, \lambda_{m+1}$ ein Intervall dicht ausfüllen, während die m größten

Eigenwerte weiter gestreut sind. In diesem Fall wird nach Satz 35.7 die Konvergenz durch eine „effektive Kondition“ $\lambda_{m+1}/\lambda_N < \mathrm{cond}_2(A)$ beschrieben; die „Ausreißer“ $\lambda_1, \ldots, \lambda_m$ bewirken lediglich eine Verzögerung zu Beginn der Iteration: Die schnellere Konvergenz setzt mit etwa m Schritten Verzögerung ein.

Man kann Satz 35.7 auch so interpretieren: Ab dem m-ten Schritt ($1 \leq m \leq N$) ist die Konvergenz mindestens linear mit Konvergenzfaktor

$$q_m = \frac{\sqrt{\lambda_m/\lambda_N} - 1}{\sqrt{\lambda_m/\lambda_N} + 1}.$$

Da q_m mit wachsendem m monoton fällt, ist die Konvergenz des CG-Verfahrens praktisch *superlinear*. Dieser superlineare Konvergenzeffekt läßt sich in Beispiel 9.7 erkennen (vgl. Abbildung 9.3), in dem die Konvergenz in den zweiten fünfzig Iterationen deutlich schneller ist als in den ersten fünfzig Iterationen.

Eine weitere Anwendung von Satz 35.7 wird in Abschnitt 54 präsentiert im Zusammenhang mit der Präkonditionierung linearer Gleichungssysteme mit Toeplitz-Struktur, vgl. Beispiel 54.3.

Aufgaben

1. Gegeben sei eine Basis $\{\varphi_k\}$ eines n-dimensionalen (komplexen) Vektorraums X sowie n weitere Funktionen

$$\phi_j \;=\; \sum_{k=1}^{n} \overline{a}_{jk}\varphi_k\,, \qquad a_{jk} \in \mathbb{K}\,, \quad j = 1,\ldots,n\,.$$

Ferner sei $f = \sum_{j=1}^{n} \zeta_j\phi_j \in X$ gegeben und $A = [a_{jk}] \in \mathbb{K}^{n\times n}$. Zeigen Sie: Ist $z = [\zeta_1,\ldots,\zeta_n]^T$ und $A^*z = [\xi_1,\ldots,\xi_n]^T$, dann ist

$$f = \sum_{j=1}^{n} \xi_j\varphi_j\,.$$

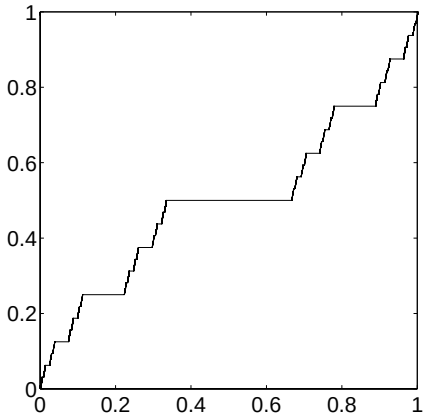
2. Sei X ein linearer Raum mit Innenprodukt $\langle\,\cdot\,,\,\cdot\,\rangle$ und $\{\phi_i : i = 1,\ldots,n\}$ eine Basis von X sowie $G = LL^*$ eine Zerlegung der zur Basis gehörigen Gramschen Matrix mit $L^{-*} = [l_{ij}]$. Zeigen Sie, daß die Funktionen

$$\psi_j \;=\; \sum_{i=1}^{n} l_{ij}\phi_i\,, \qquad j = 1,\ldots,n\,,$$

ein Orthonormalsystem in X bilden.

3. Die abgebildete *Cantor-Funktion* ist ein beliebtes Beispiel einer monotonen stetigen Funktion, die fast überall differenzierbar ist und deren Ableitung dort jeweils Null ist. Sie kann formal durch den Aufruf `cantor(0,1,0,1)` des folgenden Algorithmus definiert werden:

```
function  f = cantor (f1, f2, a, b)
  %  berechnet f im Intervall [a, b];
  %  f1 = min f|[a,b], f2 = max f|[a,b]
  l = (b − a)/3
  I1 = [a, a + l]
  I2 = [a + l, a + 2l]
  I3 = [a + 2l, b]
  m = (f1 + f2)/2
  f|I1 = cantor (f1, m, a, a + l)
  f|I1 = m
  f|I3 = cantor (m, f2, a + 2l, b)
end       %  cantor
```



Überlegen Sie sich, ob die Cantor-Funktion zu $H^1(0,1)$ gehört.

4. Zeigen Sie, daß Funktionen $u, v \in H^1(a,b)$ partiell integriert werden können:

$$\int_a^b uv'\,dx \;=\; uv\Big|_a^b - \int_a^b u'v\,dx\,.$$

5. Betrachtet wird das $H^1(-1,1)$-Innenprodukt im Raum der Polynome. Zeigen Sie:
(a) Es existiert eine eindeutig bestimmte Orthonormalbasis $\{u_n\}$ mit positiven Höchstkoeffizienten.
(b) Für jedes $n \in \mathbb{N}_0$ ist u_{2n} ein gerades und u_{2n+1} ein ungerades Polynom.
(c) u_0, u_1 und u_2 stimmen bis auf Normierungen mit den Legendre-Polynomen P_0, P_1 und P_2 überein. Stimmt dies auch noch für u_3?
(d) Schreiben Sie ein Programm, daß die Polynome $u_0, \ldots, u_n$ berechnet und plottet. Bestimmen Sie dazu zunächst die Gramsche Matrix der ersten n Legendre-Polynome in diesem Innenprodukt (beachten Sie Aufgabe 9) und verwenden Sie dann Aufgabe 2.

6. (a) Zeigen Sie, daß die Tschebyscheff-Polynome für $x \geq 1$ durch

$$T_n(x) \;=\; \cosh(n\operatorname{Arcosh} x)$$

definiert sind.
(b) Sei $x > 1$. Beweisen Sie die Ungleichung

$$T_n\big(\frac{x+1}{x-1}\big) \;\geq\; \frac{1}{2}\Big(\frac{\sqrt{x}+1}{\sqrt{x}-1}\Big)^n .$$

(c) Geben Sie die Nullstellen von T_n an. Überprüfen Sie, daß alle Nullstellen im Intervall $(-1,1)$ liegen und symmetrisch zum Nullpunkt angeordnet sind.
(d) Zeigen Sie, daß die größte Nullstelle $\lambda_1^{(n)}$ von T_n das asymptotische Verhalten

$$\lambda_1^{(n)} = 1 \;-\; \frac{\pi^2}{4} n^{-2} \;+\; O(n^{-4}) , \qquad n \to \infty ,$$

besitzt.

7. Für $n \in \mathbb{N}_0$ werden die *Tschebyscheff-Polynome 2. Art* durch

$$U_n(x) \;=\; \frac{1}{n+1}\, T'_{n+1}(x) \;=\; \frac{\sin\big((n+1)\arccos x\big)}{\sin\arccos x} , \qquad -1 < x < 1 ,$$

definiert.
(a) Zeigen Sie, daß die Polynome $\sqrt{2/\pi}\, U_n$ die Orthonormalpolynome zu der Gewichtsfunktion $w(x) \;=\; \sqrt{1-x^2}$ über $(-1,1)$ sind.
(b) Ermitteln Sie die Rekursionsformel der orthonormierten Tschebyscheff-Polynome 2. Art.
(c) Leiten Sie eine Darstellung analog zu Aufgabe 6 (a) für $U_n(x)$ mit $x > 1$ her und zeigen Sie, daß

$$U_n\big(\frac{x+1}{x-1}\big) \;\geq\; \Big(\frac{\sqrt{x}+1}{\sqrt{x}-1}\Big)^n , \qquad x > 1 .$$

8. Die Polynome $V_0 = 1$ sowie $V_n = U_n + U_{n-1} \in \Pi_n$, $n \in \mathbb{N}$, seien definiert als die Summe zweier aufeinander folgender Tschebyscheff-Polynome 2. Art.
(a) Leiten Sie für $n \in \mathbb{N}_0$ die Darstellung

$$V_n(x) \;=\; \frac{\sin((n+1/2)\arccos x)}{\sin(1/2\arccos x)} , \qquad -1 < x < 1 ,$$

her.

(b) Zeigen Sie, daß die V_n Orthogonalpolynome zu der Gewichtsfunktion

$$w(x) \,=\, \frac{\sqrt{1-x}}{\sqrt{1+x}} \qquad \text{über } (-1,1)$$

sind.
(c) Bestimmen Sie die zugehörige Jacobi-Matrix.

9. Es seien P_n die Legendre-Polynome. Rechnen Sie nach, daß

$$P_n(1) \,=\, 1 \quad \text{und} \quad P_n'(1) \,=\, \frac{1}{2}\, n(n+1)\,, \qquad n \geq 0\,.$$

10. Für $n \in \mathbb{N}_0$ werden die *Hermite-Polynome* durch

$$H_n(x) \,=\, (-1)^n e^{x^2}\, \frac{d^n}{dx^n} e^{-x^2}\,, \qquad x \in \mathbb{R}\,,$$

definiert. Zeigen Sie:
(a) Die Hermite-Polynome genügen für $n \in \mathbb{N}$ der Rekursionsformel

$$H_{n+1}(x) \,=\, 2xH_n(x) \,-\, 2nH_{n-1}(x)$$

mit $H_0(x) = 1$ und $H_1(x) = 2x$. (Damit ist klar, daß H_n ein Polynom vom Grad n ist.)
(b) Für die Ableitung von H_n gilt $H_n'(x) = 2nH_{n-1}(x)$.
(c) Die Funktionen

$$\frac{1}{\sqrt{2^n n! \sqrt{\pi}}}\, H_n(x)\,, \qquad n \in \mathbb{N}_0\,,$$

sind die Orthonormalpolynome zum Gewicht $w(x) = e^{-x^2}$ auf dem Intervall $(-\infty, \infty)$.
Hinweis: Verwenden Sie in (a) die Leibnizsche Produktregel für die n-te Ableitung und beachten Sie in (c), daß $\int_{-\infty}^{\infty} e^{-x^2}\, dx \,=\, \sqrt{\pi}$.

11. (a) Sei $T \subset \mathbb{C}$ der Rand des Einheitskreises. Dann definiert

$$\langle\, p, q \,\rangle \,=\, \frac{1}{2\pi \mathrm{i}} \int_T \overline{p(z)} q(z)\, \frac{dz}{z}$$

ein Innenprodukt im Raum aller Polynome. Rechnen Sie nach, daß die Monome $p(z) = z^n$, $n \in \mathbb{N}_0$, für dieses Innenprodukt eine Orthonormalbasis bilden.
(b) Gibt es eine Gewichtsfunktion w über $\mathbb{R}$, so daß die Monome orthogonal bezüglich des zugehörigen Innenprodukts (33.1) sind?

12. Sei Π_n' die Menge aller monischen Polynome (d. h. die Menge aller Polynome, deren Höchstkoeffizient Eins ist) mit exaktem Grad n.
(a) Sei $u_n = \gamma_n x^n + \ldots$ das Orthonormalpolynom aus Satz 33.1 Beweisen Sie, daß $p_n = u_n/\gamma_n$ die gewichtete $\mathcal{L}^2$-Norm (33.2) unter allen monischen Polynomen $p_n \in \Pi_n'$ minimiert.
(b) Zeigen Sie, daß die monischen Tschebyscheff-Polynome $2^{1-n}T_n$ auch das folgende Minimierungsproblem lösen:

$$\text{minimiere} \quad \|p_n\|_{[-1,1]} \qquad \text{unter allen } p_n \in \Pi_n'\,.$$

(c) Berechnen Sie

$$\min_{p_n \in \Pi'_n} \|p_n\|_{[-1,1]} \qquad \text{und} \qquad \min_{p_n \in \Pi'_n} \|p_n\|_{\mathcal{L}^2(-1,1)}$$

und vergleichen Sie die Ergebnisse. Zeigen Sie mit Hilfe der Stirling-Formel (41.5), daß

$$\min_{p_n \in \Pi'_n} \|p_n\|_{\mathcal{L}^2([-1,1])} \;\approx\; \frac{1}{2}\sqrt{\pi} \min_{p_n \in \Pi'_n} \|p_n\|_{[-1,1]} .$$

13. Gegeben sei die $n \times n$-Tridiagonalmatrix

$$A_n \;=\; \begin{bmatrix} 2 & -1 & & \\ -1 & 2 & \ddots & \\ & \ddots & \ddots & -1 \\ & & -1 & 2 \end{bmatrix} .$$

Zeigen Sie mit Hilfe von Aufgabe 7, daß A_n die Eigenwerte $\lambda_k = 2 - 2\cos\theta_k$ mit $\theta_k = k\pi/(n+1)$, $k = 1, \ldots, n$, besitzt, und daß

$$x_k \;=\; \big[\, \sin j\theta_k \,\big]_{j=1}^{n}$$

ein zu λ_k gehöriger Eigenvektor ist $(k = 1, \ldots, n)$.

14. Berechnen Sie die Normen $\|L\|_2$ und $\|L^{-1}\|_2$, wobei

$$L \;=\; \begin{bmatrix} 1 & & & \\ 1 & 1 & & \\ & \ddots & \ddots & \\ & & 1 & 1 \end{bmatrix} \in \mathbb{R}^{n\times n} .$$

Hinweis: Bilden Sie LL^* und vergleichen Sie mit Aufgabe 8.

15. Gegeben sei ein diskretes Innenprodukt (35.1) mit $0 < N < \infty$. Zeigen Sie, daß in diesem Fall die rechte Seite der Rekursionsformel (33.3) für $n = N - 1$ das Polynom $r_N = \gamma_{N-1}\omega$ mit $\omega(x) = \prod_{i=1}^{N}(x - \lambda_i)$ ergibt.

16. Die Orthonormalpolynome zu dem diskreten Innenprodukt

$$\langle\, f, g \,\rangle \;=\; \sum_{k=0}^{N-1} f(k)g(k)$$

heißen *diskrete Tschebyscheff-Polynome* und können explizit angegeben werden, vgl. Jordan [58, §§139,140]. Sie genügen der Rekursionsformel (33.3) mit

$$\alpha_n = \frac{N-1}{2} \quad \text{und} \quad \beta_n = \frac{n}{2}\Big(\frac{N^2 - n^2}{4n^2 - 1}\Big)^{1/2}, \qquad n < N .$$

(a) Berechnen Sie die Werte $u_n(N-1)$, $n = 1, \ldots, N-1$, über die dreistufige Rekursionsformel (33.3). Kann Ihr Ergebnis korrekt sein?

(b) Die Auswertung der dreistufigen Rekursionsformel kann als Vorwärtssubstitution zur Lösung eines linearen Gleichungssystems (für die Polynomwerte $u_n(N-1)$, $n = 1, \ldots, N$) in Dreiecksform interpretiert werden. Stellen Sie dieses Gleichungssystem auf und lösen Sie es mit der QR-Zerlegung. Ist dieses Ergebnis besser?
(c) Eine stabilere Möglichkeit zur Berechnung der fraglichen Polynomwerte besteht darin, die Gleichungen (33.3) für $n = 1, \ldots, N-1$ als ein symmetrisches Gleichungssystem für die Unbekannten $u_n(N-1)$, $n = 1, \ldots, N-1$, aufzufassen. Wie lautet die rechte Seite dieses Gleichungssystems? Welche Lösung erhalten Sie mit diesem Verfahren?
Vergleichen Sie alle berechneten Polynomwerte in einer halblogarithmischen Darstellung (siehe auch Aufgabe V.18).

17. Gegeben sei das diskrete Innenprodukt

$$\langle\, p, q \,\rangle \;=\; \sum_{i=1}^{N} w_i p(\lambda_i) q(\lambda_i)\,, \qquad w_i > 0\,,$$

mit Knoten $\lambda_1 < \lambda_2 < \ldots < \lambda_N$ und positiven Gewichten $w_i > 0$ $(i = 0, \ldots, N)$. Mit $u_j(\cdot)$ werden die zugehörigen Orthonormalpolynome und mit $K_j(\,\xi\,,\,\cdot\,)$ für $\xi \in \mathbb{R}$ die jeweiligen Kernpolynome bezeichnet $(j = 0, \ldots, N)$. Beweisen Sie, daß die Kernpolynome $K_j(\,\xi\,,\,\cdot\,)$ für $\xi \,/\!\in [\lambda_1, \lambda_N]$ bezüglich des Innenprodukts

$$\langle\!\langle\, p, q \,\rangle\!\rangle \;=\; \sum_{i=1}^{N} w_i |\lambda_i - \xi| p(\lambda_i) q(\lambda_i)$$

orthogonal sind.

18. Implementieren Sie das Bisektionsverfahren (Algorithmus 35.1) und berechnen Sie damit Näherungen für den k-ten Eigenwert der Jacobi-Matrix J_{20} der orthonormierten Tschebyscheff-Polynome (vgl. Beispiel 34.7).

VII Numerische Quadratur

Gegenstand dieses Kapitels ist die numerische Approximation bestimmter Integrale

$$I[f] \;=\; \int_a^b f(x)\,dx\,, \qquad -\infty \le a < b \le \infty\,,$$

die nicht in geschlossener Form ausgewertet werden können. Zur Approximation werden geeignete *Quadraturformeln* verwendet, die wenige Funktionswerte von f zu einer Integralnäherung mitteln. Durch Anwendung einer solchen Quadraturformel auf einzelne Teilintervalle von $[a,b]$ der Länge h ergibt sich ein zusammengesetztes *Quadraturverfahren*, das für $h \to 0$ gegen $I[f]$ konvergiert.

Dieser Stoff wird in jedem Numerik-Lehrbuch ausführlich behandelt, hier sei vor allem auf die neueren Bücher von Gautschi [31] und Kreß [63] verwiesen, die zum Teil von der hier gewählten Darstellung abweichen. Für weiterführende Resultate empfiehlt sich neben dem Standardwerk von Davis und Rabinowitz [21] noch das Buch von Krommer und Überhuber [65], in dem auch auf aktuelle Software eingegangen wird.

36 Die Trapezformel

Die einfachsten Quadraturformeln sind die sogenannte *Mittelpunktformel*

$$\int_a^b f(x)\,dx \;=\; (b-a)\,f\big(\frac{a+b}{2}\big) \;=:\; M[f] \tag{36.1}$$

und die *Trapezformel*

$$\int_a^b f(x)\,dx \;=\; \frac{b-a}{2}\,f(a) + \frac{b-a}{2}\,f(b) \;=:\; T[f]\,. \tag{36.2}$$

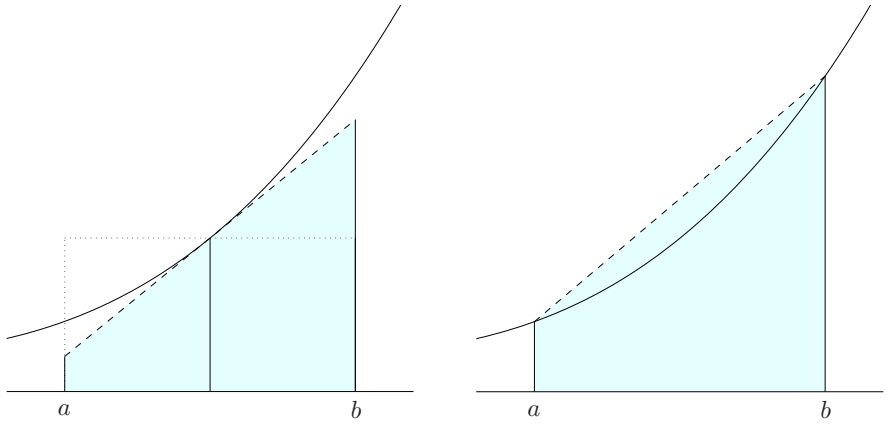


Abb. 36.1: Geometrische Interpretation der Mittelpunktformel und der Trapezformel.

Natürlich gilt im allgemeinen weder in (36.1) noch in (36.2) das Gleichheitszeichen. Die Formel ist lediglich eine Näherung, vgl. Abbildung 36.1. Um diese Näherung zu verbessern, kann man jedoch das Intervall $[a, b]$ in n gleich große Teilintervalle zerlegen und die entsprechende Formel auf jedes Teilintervall anwenden. Bei der Mittelpunktformel ergibt sich so eine spezielle Riemannsche Zwischensumme, die Trapezformel führt auf die sogenannte *Trapezsumme*. Bezeichnen wir die Randpunkte der n Teilintervalle mit $x_i = a + ih$, $i = 0, \ldots, n$, wobei $h = (b-a)/n$ die Länge der einzelnen Teilinvertvalle ist, so lautet beispielsweise die Trapezsumme

$$\begin{aligned} T_n[f] &= \sum_{i=1}^{n} \frac{x_i - x_{i-1}}{2} \big(f(x_i) + f(x_{i-1})\big) \\ &= \frac{h}{2} f(a) + h \sum_{i=1}^{n-1} f(x_i) + \frac{h}{2} f(b)\,. \end{aligned} \tag{36.3}$$

Man macht sich leicht klar, daß die beiden zusammengesetzten Quadraturverfahren für $n \to \infty$ gegen $I[f]$ konvergieren, falls f über $[a, b]$ Riemannintegrierbar ist. Unter zusätzlichen Voraussetzungen an f kann darüber hinaus die folgende Fehlerabschätzung bewiesen werden (vgl. Aufgabe 2 für das entsprechende Resultat bei der Mittelpunktformel):

Satz 36.1. *Sei $f \in C^2[a, b]$ und $h = (b-a)/n$ für ein $n \in \mathbb{N}$. Dann gilt*

$$\big|\, I[f] - T_n[f] \,\big| \;\le\; \frac{b-a}{12}\, \|f''\|_{[a,b]}\, h^2\,.$$

Dabei bezeichnet $\|\cdot\|_{[a,b]}$ die Maximumnorm über dem Intervall $[a, b]$.

Beweis. Wir betrachten zunächst den Fall $n = 1$, also die Trapezformel. Deren Fehler läßt sich mit einer Stammfunktion F von f in der Form

$$I[f] - T_1[f] \;=\; F(b) - F(a) - \frac{b-a}{2}\big(f(a) + f(b)\big)$$

schreiben. Eine Taylorentwicklung von F um $x = a$ ergibt dann

$$\begin{aligned} I[f] - T_1[f] \;&=\; F'(a)(b-a) \;+\; \int_a^b F''(x)(b-x)\,dx \;-\; \frac{b-a}{2}\big(f(a) + f(b)\big) \\ &=\; \frac{b-a}{2}\big(f(a) - f(b)\big) \;+\; \int_a^b f'(x)(b-x)\,dx\,, \end{aligned}$$

und aus dem Hauptsatz der Integralrechnung folgt schließlich

$$I[f] - T_1[f] \;=\; -\int_a^b f'(x)\big(x - \frac{a+b}{2}\big)\,dx\,.$$

Man beachte, daß der zweite Faktor unter dem Integral die Ableitung der Parabel

$$K(x) \;=\; \frac{1}{2}(x-a)(x-b) \tag{36.4}$$

ist, die in beiden Randpunkten des Intervalls verschwindet. Durch partielle Integration ergibt sich daher

$$\begin{aligned} I[f] - T_1[f] \;&=\; -f'(x)K(x)\Big|_a^b \;+\; \int_a^b f''(x)K(x)\,dx \\ &=\; \int_a^b f''(x)K(x)\,dx\,. \end{aligned} \tag{36.5}$$

Die Fehlerdarstellung (36.5) geht auf Peano zurück, die Funktion K aus (36.4) wird aus diesem Grund *Peano-Kernfunktion* genannt.

Aus der Dreiecksungleichung folgt nun unmittelbar mit der Substitution $u = (2x - a - b)/(b - a)$

$$\begin{aligned} \big|\, I[f] - T_1[f] \,\big| \;&\le\; \int_a^b |f''(x)K(x)|\,dx \;\le\; \frac{1}{2}\,\|f''\|_{[a,b]} \int_a^b (x-a)(b-x)\,dx \\ &=\; \frac{1}{2}\,\|f''\|_{[a,b]} \int_{-1}^1 \frac{(b-a)^2}{4}\,(1-u^2)\,\frac{b-a}{2}\,du \\ &=\; \frac{1}{16}\,\|f''\|_{[a,b]}(b-a)^3(u - \frac{1}{3}\,u^3)\Big|_{-1}^1 \;=\; \frac{1}{12}\,\|f''\|_{[a,b]}(b-a)^3\,. \end{aligned}$$

Ersetzen wir in diesem Zwischenergebnis a durch x_i, b durch $x_{i+1} = x_i + h$ und wenden das Ergebnis auf (36.3) an, so folgt

$$\begin{aligned} \big|\, I[f] - T_n[f] \,\big| \;&\le\; \sum_{i=1}^{n} \Big| \int_{x_{i-1}}^{x_i} f(x)\,dx \;-\; \frac{h}{2}\,\big(f(x_{i-1}) + f(x_i)\big) \,\Big| \\ &\le\; \sum_{i=1}^{n} \frac{1}{12}\,\|f''\|_{[a,b]}\,h^3 \;=\; \frac{n}{12}\,\|f''\|_{[a,b]}\,h^3 \;=\; \frac{b-a}{12}\,\|f''\|_{[a,b]}\,h^2. \quad \square \end{aligned}$$

Bei vielen Anwendungen erfordert die obere Schranke $O(h^2)$ aus Satz 36.1 eine sehr kleine Wahl von h (und damit sehr viele Funktionsauswertungen), um eine vorgeschriebene Genauigkeit garantieren zu können. Man mag sich daher fragen, ob es nicht bessere Verfahren als die Trapezsumme gibt. Im folgenden entwickeln wir solche Verfahren und betrachten dabei das (geringfügig allgemeinere) Integral

$$I[f;w] \;=\; \int_a^b f(x)w(x)\,dx\,.$$

Dabei ist $w(x)$ wie in Kapitel VI eine Gewichtsfunktion, also eine positive integrierbare Funktion über dem Intervall $[a,b]$ mit

$$\int_a^b w(x)\,dx < \infty\,.$$

Zur Approximation von $I[f;w]$ betrachten wir Ausdrücke der Form

$$Q[f] \;=\; \sum_{i=0}^{m} w_i f(x_i) \tag{36.6}$$

mit *Knoten* x_i und *Gewichten* w_i, $i = 0,\ldots,m$. Dabei sei ausdrücklich betont, daß zwischen w_i und $w(x_i)$ *kein* unmittelbarer Zusammenhang zu bestehen braucht. Eine Näherung (36.6) nennen wir eine *Quadraturformel*, wenn die Wahl von m, $\{x_i\}$ und $\{w_i\}$ fest ist, wie zum Beispiel bei der Trapezformel (36.2). Unter dem dazugehörigen (*zusammengesetzten*) *Quadraturverfahren* verstehen wir dann die Unterteilung von $[a,b]$ in n gleich große Teilintervalle mit Intervallänge $h = (b-a)/n$, in denen jeweils die Quadraturformel angewandt wird. Bei n Teilintervallen ($n > 1$) bezeichnen wir die entsprechende Näherung des Quadraturverfahrens mit $Q_n[f]$.

Um die qualitativen Merkmale einer Quadraturformel beziehungsweise eines Quadraturverfahrens beschreiben zu können, führen wir noch zwei Definitionen an.

Definition 36.2. (*a*) Eine Quadraturformel $Q[\,\cdot\,]$ für das Integral $I[\,\cdot\,;w]$ hat *Exaktheitsgrad* q, falls

$$Q[p] = I[p;w] \qquad \text{für alle } p \in \Pi_q\,.$$

Π_q ist dabei wieder der Raum aller Polynome mit Grad höchstens q.
(*b*) Ein Quadraturverfahren $Q_n[\,\cdot\,]$ hat die *Konsistenzordnung* s, falls

$$\big|\,Q_n[f] - I[f;w]\,\big| \;=\; O(h^s)\,, \qquad n \to \infty\,,$$

für hinreichend glatte Funktionen f.

Beispiel. Für $w = 1$ hat die Trapezformel den Exaktheitsgrad $q = 1$, und die Trapezsumme hat Konsistenzordnung $s = 2$ für alle $f \in C^2[a,b]$, vgl. Satz 36.1.

◊

Bemerkung. Für die Bestimmung des Exaktheitsgrades einer Quadraturformel ist es hinreichend, eine Basis von Π_q zu untersuchen, da sowohl $Q[\,\cdot\,]$ als auch $I[\,\cdot\,;w]$ lineare Abbildungen sind. ◊

37 Polynominterpolation

Um weitere Quadraturformeln und –verfahren einführen zu können, benötigen wir ein Hilfsmittel: die Polynominterpolation. Hierzu benötigen wir ein *Gitter*

$$\triangle \;=\; \{x_0 < x_1 < \ldots < x_m\} \;\subset\; \mathbb{R} \tag{37.1}$$

aus $m+1$ ansteigend angeordneten *Knoten*. Mit $h_i = x_i - x_{i-1}$, $i = 1,\ldots,m$, bezeichnen wir die Länge der einzelnen Teilintervalle,

$$h \;=\; \max_{i=1,\ldots,m} h_i \tag{37.2}$$

ist die *Gitterweite* des Gitters. Ein Gitter heißt *äquidistant*, wenn alle Gitterintervalle gleich lang sind; in diesem Fall ist $h = (b-a)/m$.

Problem 37.1 (Interpolationsaufgabe). *Gegeben seien ein Gitter* (37.1) *sowie* $m+1$ *Werte* $y_0,\ldots,y_m$. *Gesucht ist ein Polynom* $p \in \Pi_m$ *mit*

$$p(x_i) = y_i\,, \qquad i = 0,\ldots,m\,. \tag{37.3}$$

Bevor wir die Lösbarkeit dieser Interpolationsaufgabe diskutieren, müssen noch einige weitere zentrale Begriffe eingeführt werden.

Definition 37.2. Wir bezeichnen mit

$$\omega(x) \;=\; \prod_{i=0}^{m}(x - x_i) \;\in \Pi_{m+1}$$

das zu dem Gitter $\triangle$ gehörende *Knotenpolynom*. Die Polynome

$$l_i(x) \;=\; \frac{\omega(x)}{(x - x_i)\omega'(x_i)} \;=\; \prod_{\substack{j=0\\ j\neq i}}^{m} \frac{x - x_j}{x_i - x_j} \;\in \Pi_m$$

werden *Lagrange-Grundpolynome* genannt.

Von entscheidender Bedeutung ist die folgende leicht nachzurechnende Eigenschaft der Lagrange-Grundpolynome,

$$l_i(x_j) = \delta_{ij}\,, \qquad i, j = 0, \ldots, m\,, \tag{37.4}$$

denn aus ihr folgt unmittelbar, daß

$$p = \sum_{i=0}^{m} y_i\, l_i \;\in \Pi_m \tag{37.5}$$

die Interpolationsaufgabe löst.

Satz 37.3. *Das Problem* 37.1 *hat genau eine Lösung* $p \in \Pi_m$. *Dieses Interpolationspolynom kann in der Form* (37.5) *dargestellt werden.*

Beweis. Es verbleibt noch der Nachweis der Eindeutigkeit. Sind p und q aus Π_m zwei Lösungen von Problem 37.1, dann folgt aus der Interpolationseigenschaft (37.3)

$$(p - q)(x_i) = 0\,, \qquad i = 0, \ldots, m\,.$$

Das Polynom $p - q \in \Pi_m$ hat somit $m + 1$ Nullstellen, d. h. es muß $p = q$ sein. □

Beispiel. Eine Funktion f wird durch das Polynom

$$p(x) \;=\; f(a)\,\frac{x - b}{a - b} \;+\; f(b)\,\frac{x - a}{b - a} \;=\; f(a) \;+\; \frac{f(b) - f(a)}{b - a}\,(x - a)$$

in den Punkten $\big(a, f(a)\big)$ und $\big(b, f(b)\big)$ interpoliert ($p \in \Pi_1$ ist die Sekante an f durch diese Punkte). ◊

Satz 37.4. *Sei* $f \in C^{m+1}[a, b]$ *und* $p \in \Pi_m$ *das Interpolationspolynom zu dem Gitter* $\triangle \subset [a, b]$ *aus* (37.1) *und den Werten* $y_i = f(x_i)$, $i = 0, \ldots, m$. *Dann*

gibt es zu jedem $x \in [a,b]$ *ein* ξ *aus der konvexen Hülle* $\mathcal{I} = \operatorname{conv}\{x, x_0, \ldots, x_m\}$ *mit*

$$f(x) - p(x) \;=\; \frac{f^{(m+1)}(\xi)}{(m+1)\,!}\,\omega(x)\,. \tag{37.6}$$

Beweis. Offensichtlich ist (37.6) für $x \in \triangle$ richtig. Sei nun ω das Knotenpolynom zu dem Gitter und $x \notin \triangle$ fest gewählt. Dann hat die Funktion

$$h(t) \;=\; f(t) - p(t) - \frac{\omega(t)}{\omega(x)}\,\big(f(x) - p(x)\big)\,, \qquad t \in \mathbb{R}\,, \tag{37.7}$$

Nullstellen in jedem der $m+1$ Gitterpunkte sowie in $t = x$. Jedes der $m+1$ Teilintervalle zwischen diesen Nullstellen enthält nach dem Satz von Rolle eine Nullstelle τ_i' von h', $i = 1, \ldots, m+1$. So fortfahrend erhält man $m+2-k$ Nullstellen $\tau_i^{(k)}$ von $h^{(k)}$, $k = 1, 2, \ldots$, in dem Intervall $\mathcal{I}$. Insbesondere ergibt sich *eine* Nullstelle $\xi = \tau_1^{(m+1)}$ von $h^{(m+1)}$ in $\mathcal{I}$. Wegen $p \in \Pi_m$ ist die $(m+1)$-te Ableitung von p identisch Null. Die $(m+1)$-te Ableitung von ω ist ebenfalls konstant, nämlich

$$\omega^{(m+1)}(t) \;=\; \frac{d^{m+1}}{dt^{m+1}}\,(t^{m+1} + \ldots) \;=\; (m+1)!\,.$$

Hieraus folgt mit (37.7) für die Nullstelle ξ von $h^{(m+1)}$, daß

$$0 \;=\; h^{(m+1)}(\xi) \;=\; f^{(m+1)}(\xi) - \frac{(m+1)\,!}{\omega(x)}\,\big(f(x) - p(x)\big)$$

beziehungsweise

$$f(x) - p(x) \;=\; \frac{f^{(m+1)}(\xi)}{(m+1)\,!}\,\omega(x)\,. \qquad \square$$

Als Verfahren der numerischen Approximation ist die Polynominterpolation leider nur sehr eingeschränkt brauchbar, nämlich nur bei wenigen Interpolationspunkten, d. h. bei kleinem Polynomgrad. Andernfalls ergeben sich unter Umständen starke Oszillationen zwischen den Interpolationspunkten. Das folgende Beispiel mag als Warnung genügen:

Beispiel 37.5. Gegeben seien die Werte $y_i = f(x_i)$ der Funktion

$$f(x) \;=\; \frac{1}{1 + 25x^2}$$

über dem äquidistanten Gitter $\triangle_m = \{x_i = i/m \,:\, -m \le i \le m\} \subset [-1,1]$. Runge hat im Jahr 1901 dieses Beispiel angegeben und bewiesen, daß die zugehörigen Interpolationspolynome $p_{2m} \in \Pi_{2m}$ für $m \to \infty$ in $[-1,1]$ nicht

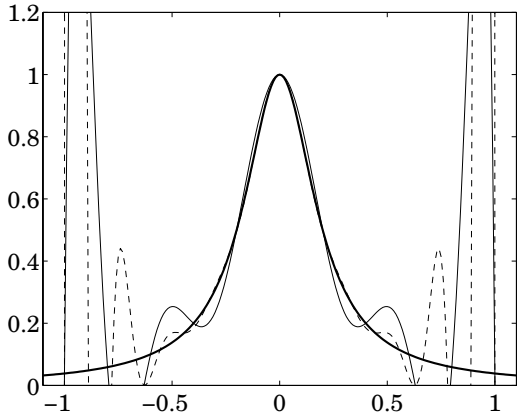


Abb. 37.1: Das Beispiel von Runge

punktweise gegen f konvergieren. Abbildung 37.1 zeigt die Funktion f (fette Kurve) und die Interpolationspolynome p_{10} (dünne Kurve) und p_{18} (gebrochene Kurve). ◊

Bemerkung 37.6. Bei der *Hermite-Interpolation* werden einzelne Gitterknoten mehrfach zugelassen. Tritt beispielsweise der Knoten x_i k mal auf, so werden neben dem Funktionswert y_i an der Stelle $x = x_i$ noch die ersten $k-1$ Ableitungen des Interpolationspolynoms vorgeschrieben:

$$p(x_i) = y_i\,, \quad p'(x_i) = y_i'\,, \quad \ldots\,, \quad p^{(k-1)}(x_i) = y_i^{(k-1)}\,.$$

Auch diese Interpolationsaufgabe ist eindeutig lösbar und Satz 37.4 gilt entsprechend: Im Knotenpolynom ω treten mehrfache Knoten x_i dann entsprechend ihrer Vielfachheit auf, vgl. Aufgabe 4. ◊

38 Newton-Cotes-Formeln

Über die Polynominterpolation lassen sich leicht Quadraturformeln für $I[\,\cdot\,;w]$ mit beliebigem Exaktheitsgrad q angeben. Ist $\triangle = \{x_0 < \ldots < x_m\} \subset [a,b]$ ein Gitter und

$$w_i \;=\; \int_a^b l_i(x) w(x)\,dx\,, \qquad i = 0, \ldots, m\,, \tag{38.1}$$

das Integral des i-ten zugehörigen Lagrange-Grundpolynoms, dann gilt das folgende Resultat.

Proposition 38.1. *$Q[\,\cdot\,]$ aus* (36.6) *sei eine Quadraturformel für das Integral $I[\,\cdot\,;w]$ über dem Gitter $\triangle$. Dann hat Q genau dann einen Exaktheitsgrad $q \geq m$, wenn die zugehörigen Gewichte w_i durch* (38.1) *gegeben sind.*

Beweis. Sei zunächst (38.1) erfüllt für alle $i = 0, \ldots, m$ und $p \in \Pi_m$ beliebig gewählt. Offensichtlich interpoliert p sich selbst in den Gitterpunkten von $\triangle$. Wegen der Eindeutigkeit des Interpolationspolynoms gilt daher

$$p(x) \;=\; \sum_{i=0}^{m} p(x_i)\, l_i(x)\,,$$

vgl. Satz 37.3. Daraus folgt

$$\begin{aligned} I[p;w] \;&=\; \int_a^b p(x)w(x)\,dx \;=\; \int_a^b \sum_{i=0}^{m} p(x_i) l_i(x) w(x)\,dx \\ &=\; \sum_{i=0}^{m} p(x_i) \int_a^b l_i(x)w(x)\,dx \overset{(38.1)}{=} \sum_{i=0}^{m} w_i\, p(x_i)\,, \end{aligned}$$

das heißt Q hat mindestens Exaktheitsgrad $q = m$.

Für die umgekehrte Beweisrichtung sei angenommen, daß $q \geq m$ ist. Dann ist insbesondere

$$\int_a^b l_i(x)w(x)\,dx \;=\; I[l_i;w] \;=\; Q[l_i] \;=\; \sum_{j=0}^{m} w_j \underbrace{l_i(x_j)}_{=\delta_{ij}} \;=\; w_i\,,$$

also gilt (38.1). □

Aus dem Interpolationsfehler erhalten wir außerdem unmittelbar eine Abschätzung für den Quadraturfehler.

Proposition 38.2. *Sei $f \in C^{m+1}[a,b]$. Dann gilt die folgende Fehlerabschätzung für die Quadraturformel $Q[f]$ aus Proposition* 38.1*:*

$$\big|\, I[f;w] - Q[f] \,\big| \;\leq\; \frac{\|f^{(m+1)}\|_{[a,b]}}{(m+1)\,!} \int_a^b |\,\omega(x)|\, w(x)\,dx\,. \tag{38.2}$$

Hierbei ist $\omega(x)$ das Knotenpolynom aus Definition 37.2*.*

Beweis. Sei $p \in \Pi_m$ das Interpolationspolynom zu f über dem Gitter $\triangle$. Dann gilt offensichtlich $Q[f] = Q[p]$ und wegen des Exaktheitsgrads der Quadraturformel $Q[\,\cdot\,]$ folgt

$$I[f;w] - Q[f] \;=\; I[f;w] - I[p;w] \;=\; \int_a^b (f-p)(x)w(x)\,dx\,.$$

Damit folgt die gewünschte Aussage unmittelbar aus Satz 37.4. □

Beispiele. Im weiteren beschränken wir uns auf den Fall äquidistanter Knoten $a = x_0 < x_1 < \ldots < x_m = b$ und $w = 1$. In diesem Fall spricht man von (*abgeschlossenen*) *Newton-Cotes-Formeln.* Für $m = 1$ ergibt sich etwa die Trapezformel (36.2). Man überzeugt sich relativ leicht davon, daß die *Simpson-Formel*

$$\int_a^b f(x)\,dx \;\approx\; S[f] \;=\; \frac{b-a}{6}\,\big(\,f(a) + 4f\big(\frac{a+b}{2}\big) + f(b)\,\big) \tag{38.3}$$

mindestens Exaktheitsgrad $q = 2$ besitzt. Nach Proposition 38.1 muß es sich daher bei (38.3) um die zweite Newton-Cotes-Formel handeln. ◊

Da die konstanten Funktionen in allen Polynomräumen Π_q mit $q \geq 0$ liegen, gilt für jede Quadraturformel Q mit Exaktheitsgrad $q \geq 0$, daß

$$\int_a^b 1\,dx \;=\; I[1] \;=\; Q[1] \;=\; \sum_{i=0}^{m} w_i \cdot 1 \;=\; \sum_{i=0}^{m} w_i\,.$$

Daher ist

$$\sum_{i=0}^{m} w_i = b - a \tag{38.4}$$

bei jeder Quadraturformel für $I[\,\cdot\,]$ mit Exaktheitsgrad $q \geq 0$ (insbesondere also für alle Newton-Cotes-Formeln).

Weil die Polynominterpolation nur bedingt gute Approximationen an f liefert (vgl. Beispiel 37.5), macht es im allgemeinen wenig Sinn, die Anzahl m der Gitterknoten hochzuschrauben. Zudem treten für größere m negative Gewichte und dadurch unter Umständen Stabilitätsverluste selbst bei positiven Integranden auf. Für $m \leq 10$ sind die Gewichte etwa in der Formelsammlung von Abramowitz und Stegun [1, Formeln 25.4.13–25.4.20] tabelliert.

Um genauere Approximationen an den Integralwert zu erhalten, wird in der Praxis statt dessen wie in Abschnitt 36 das Integrationsintervall unterteilt und das entsprechende zusammengesetzte Newton-Cotes-Verfahren verwendet. Im Fall $m = 2$ erhält man so das *zusammengesetzte Simpson-Verfahren*

$$\begin{aligned}\int_a^b f(x)\,dx \approx\; & \frac{h}{3}\,\big(f(a) + 4f(x_1) + 2f(x_2) + 4f(x_3) + \ldots + 4f(x_{2n-1}) + f(b)\big) \\ =\; & S_n[f]\end{aligned}$$

mit $h = (b-a)/(2n)$ und $x_i = a + ih$.

Das folgende Hilfsresultat wird zur Abschätzung des Quadraturfehlers des zusammengesetzten Simpson-Verfahrens benötigt.

Lemma 38.3. *Sei $Q[f]$ eine Quadraturformel für $I[f] = \int_{-1}^{1} f(x)\,dx$ mit zum Nullpunkt symmetrischen Knoten und Gewichten. Gilt $Q[p] = I[p]$ für alle Polynome $p \in \Pi_{2q}$, dann hat $Q[\,\cdot\,]$ (mindestens) den Exaktheitsgrad $2q+1$.*

Beweis. Wir betrachten die Basis $\{x^0, x^1, \ldots, x^{2q}, x^{2q+1}\}$ von Π_{2q+1}. Nach Voraussetzung ist $Q[x^j] = I[x^j]$ für $j = 0, \ldots, 2q$. Wegen der Punktsymmetrie von x^{2q+1} ist ferner sowohl $I[x^{2q+1}]$ als auch $Q[x^{2q+1}]$ gleich Null. Letzteres folgt dabei aus der Symmetrie der Knoten und Gewichte. Somit stimmen $Q[\,\cdot\,]$ und $I[\,\cdot\,]$ auf einer Basis von Π_{2q+1} überein und damit automatisch auf ganz Π_{2q+1}. □

Quadraturformeln, die die Voraussetzungen von Lemma 38.3 erfüllen, werden *symmetrisch* genannt. Durch eine affin-lineare Transformation $[a,b] \mapsto [-1,1]$ ändern sich zwar in entsprechender Weise die Knoten und Gewichte einer Newton-Cotes-Formel (die Knoten werden mittransformiert, die Gewichte werden mit $2/(b-a)$ multipliziert), jedoch nicht ihr Exaktheitsgrad, da affin-lineare Transformationen nicht den Polynomgrad des Integranden ändern. Die auf das Intervall $[-1,1]$ transformierte Simpson-Formel hat die Knoten -1, 0 und 1 und die Gewichte 1/3, 4/3, 1/3. Die Simpson-Formel ist daher symmetrisch und hat folglich mindestens den Exaktheitsgrad $q = 3$ (und nicht lediglich den Exaktheitsgrad $q = m = 2$, wie er durch Proposition 38.1 garantiert ist).

Für das zusammengesetzte Simpson-Verfahren ergibt sich damit die folgende Fehlerabschätzung.

Satz 38.4. *Sei $f \in C^4[a,b]$ und $n \in \mathbb{N}$. Dann gilt mit $h = (b-a)/(2n)$*

$$\big|\, I[f] - S_n[f] \,\big| \;\le\; \frac{b-a}{180}\, \|f^{(4)}\|_{[a,b]}\, h^4\,,$$

d. h. das zusammengesetzte Simpson-Verfahren hat die Ordnung $s = 4$.

Beweis. Sei $[c,d] = [x_{2i-2}, x_{2i}]$, $i = 1, \ldots, n$, eines der n Teilintervalle von $[a,b]$, auf das die Simpson-Formel angewandt wird. Wir interpolieren f in $[c,d]$ durch ein Polynom p dritten Grades mit Stützstellen in c, d und $(c+d)/2$. Die Stützstelle in der Mitte des Intervalls wird dabei doppelt vorgegeben im Sinne der Hermite-Interpolation, vgl. Bemerkung 37.6. Wie im Beweis von Proposition 38.2 ergibt sich der Quadraturfehler im Intervall $[c,d]$ somit als das Integral über $f-p$. Um dieses abzuschätzen, verwenden wir die Fehlerdarstellung (37.6) für den Interpolationsfehler mit dem entsprechenden Knotenpolynom

$$\omega(x) \;=\; (x-c)(x-d)\big(x-(c+d)/2\big)^2$$

(vgl. erneut Bemerkung 37.6). Durch Integration ergibt dies mit der Dreiecks-

ungleichung und der Substitution $t = (2x - c - d)/(d - c)$

$$\Big| \int_c^d \big(f(x) - p(x)\big)\, dx \Big| \;\le\; \frac{\|f^{(4)}\|_{[a,b]}}{4!} \big(\frac{d-c}{2}\big)^5 \int_{-1}^{1} t^2(1-t^2)\, dt$$

$$= \;\frac{\|f^{(4)}\|_{[a,b]}}{4!} \big(\frac{d-c}{2}\big)^5 \big(\frac{1}{3}t^3 - \frac{1}{5}t^5\big)\Big|_{-1}^{1} \;=\; \frac{d-c}{180}\,\big(\frac{b-a}{2n}\big)^4 \|f^{(4)}\|_{[a,b]}$$

und durch Summation dieser jeweiligen Fehler aus den einzelnen Teilintervallen folgt die Behauptung. □

Beispiel. Die Simpson-Formel liefert für das Integral $I = \int_{-1}^{1} x^4 dx = 2/5$ den Wert $S = 2/3 \neq I$, so daß der Exaktheitsgrad der Simpson-Formel genau $q = 3$ ist. Die Fehlerschranke von Satz 38.4 liefert für dieses Beispiel den exakten Fehler $|I - S| = 4/15$, diese Schranke ist also scharf (vgl. auch Aufgabe 6). ◊

Man beachte, daß der Wert für h in Satz 38.4 bereits berücksichtigt, daß für das zusammengesetzte Simpson-Verfahren doppelt so viele Funktionswerte benötigt werden wie für die Trapezsumme: Bei der gleichen Anzahl von Funktionsauswertungen für die Trapezsumme und das zusammengesetzte Simpson-Verfahren stimmen auch die Werte von h in Satz 36.1 und Satz 38.4 überein.

Beispiel 38.5. Wir vergleichen die Ergebnisse der Trapezsumme mit dem zusammengesetzten Simpson-Verfahren. Abbildung 38.1 zeigt die relativen Fehler der entsprechenden Näherungen an das Integral

$$\int_0^1 \frac{1}{x+1}\, dx \;=\; \log 2$$

für $n = 1/h = 1, 2, \ldots, 2000$. Dabei ist $n + 1$ die Anzahl der Funktionsauswertungen des Integranden. An der Abbildung ist deutlich die sublineare Konvergenzgeschwindigkeit (n^{-2} bzw. n^{-4}) zu erkennen. Zudem ist offensichtlich der relative Fehler des Simpson-Verfahrens ziemlich genau das Quadrat des entsprechenden Fehlers der Trapezsumme. ◊

39 Das Romberg-Verfahren

In der Fehlerabschätzung der Trapezsumme und des zusammengesetzten Simpson-Verfahrens ist der Faktor h^s mit festem s der entscheidende Term (für die Trapezsumme ist $s = 2$, für das Simpson-Verfahren ist $s = 4$). Eine Halbierung von h führt demnach zu einer Verbesserung des Fehlers um einen Faktor von etwa 2^{-s}. In diesem Abschnitt stellen wir eine Methode vor, bei der neben einer Reduktion von h gleichzeitig auch der Exponent s erhöht wird.

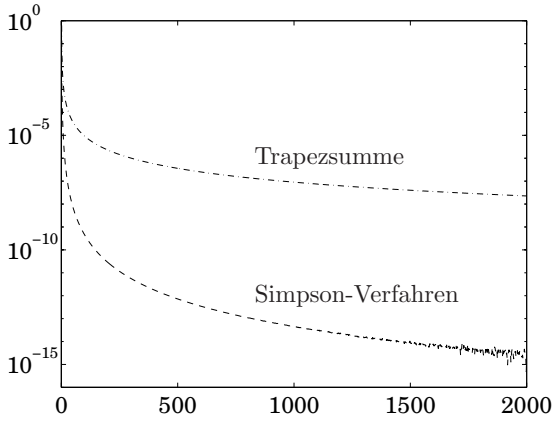


Abb. 38.1: Relative Fehler bei Trapezsumme und zusammengesetztem Simpson-Verfahren

Beispiel 39.1. Zur Motivation betrachten wir die Trapezformel und nehmen an, daß die Fehlerabschätzung aus Satz 36.1 scharf ist. Dann gilt

$$\begin{aligned} T_1[f] &\approx I[f] + \varepsilon \qquad \text{und} \\ T_2[f] &\approx I[f] + \varepsilon/4 \end{aligned}$$

mit $\varepsilon = (b-a)^3 \|f''\|_{[a,b]}/12$. Wir können also erhoffen, daß

$$T_{(1,2)}[f] \;=\; T_2[f] + \frac{1}{3}\big(\,T_2[f] - T_1[f]\,\big) \;\approx\; I[f] + 0 \cdot \varepsilon \tag{39.1}$$

eine noch bessere Näherung an $I[f]$ als $T_2[f]$ ist – und das bei im wesentlichen demselben Rechenaufwand. Tatsächlich erhält man auf diese Weise

$$\begin{aligned} T_{(1,2)}[f] &= \frac{b-a}{4}\big(f(a) + 2f(\tfrac{a+b}{2}) + f(b)\big) \\ &\qquad + \frac{b-a}{12}\big(f(a) + 2f(\tfrac{a+b}{2}) + f(b) - 2f(a) - 2f(b)\big) \\ &= \frac{b-a}{6}\big(f(a) + 4f(\tfrac{a+b}{2}) + f(b)\big) \end{aligned}$$

die Simpson-Formel $S[f]$, die einen höheren Exaktheitsgrad besitzt. ◊

Diese Idee, verschiedene Trapezsummen zu mitteln, kann rekursiv fortgeführt werden, wie im folgenden skizziert werden soll. Dazu nehmen wir zunächst an, daß f um jeden Punkt $x \in [a,b]$ in eine global konvergente Potenzreihe entwickelt werden kann, also daß

$$f(x+h) \;=\; f(x) + \frac{f'(x)}{1!}\,h + \,\ldots\, + \frac{f^{(k)}(x)}{k!}\,h^k + \,\ldots$$

für alle $h > 0$. Dann definieren wir

$$g(x,h) \;=\; \int_x^{x+h} f(t)\,dt\,. \tag{39.2}$$

Bezeichnet F eine Stammfunktion von f, so ist $g(x,h) = F(x+h) - F(x)$, und aus der Potenzreihenentwicklung von f folgt

$$g(x,h) \;=\; f(x)\,h + \frac{f'(x)}{2!}\,h^2 + \,\ldots\, + \frac{f^{(k)}(x)}{(k+1)!}\,h^{k+1} + \,\ldots$$

Wir halten nun h fest und betrachten geeignete Vielfache der partiellen Ableitungen von g nach x, die durch gliedweise Differentiation nach x berechnet werden können:

$$\begin{aligned}
g(x,h) \;&= f(x)\,h \,+\, \frac{1}{2}f'(x)\,h^2 + \;\frac{1}{6}\,f''(x)\,h^3 \,+\,\ldots+\quad \frac{1}{(k+1)!}\,f^{(k)}(x)\,h^{k+1} \,+\,\ldots\,,\\
-\frac{1}{2}\,h\,g_x(x,h) \;&= \qquad -\frac{1}{2}\,f'(x)\,h^2 - \;\frac{1}{4}\,f''(x)\,h^3 \,-\,\ldots-\qquad \frac{1}{2k!}\,f^{(k)}(x)\,h^{k+1} \,-\,\ldots\,,\\
\frac{1}{12}\,h^2 g_{xx}(x,h) \;&= \qquad\qquad +\frac{1}{12}\,f''(x)\,h^3 \,+\,\ldots+\frac{1}{12(k-1)!}\,f^{(k)}(x)\,h^{k+1} \,+\,\ldots\,,\\
&\;\;\vdots \qquad\qquad\qquad\qquad\qquad\qquad \ddots \qquad\qquad \vdots
\end{aligned}$$

Die Vorfaktoren sind so gewählt, daß sich bei einer Summation dieser Reihen die höheren Ableitungen von f wegheben,

$$\begin{aligned}
f(x)\,h \;&=\; g(x,h) \,-\, \frac{1}{2}\,h\,g_x(x,h) \,+\, \frac{1}{12}\,h^2\,g_{xx}(x,h) \,-\,\ldots\\
&=\; \sum_{k=0}^{\infty} \frac{B_k}{k!}\,h^k\,\frac{\partial^k}{\partial x^k}\,g(x,h)\,.
\end{aligned} \tag{39.3}$$

Die Koeffizenten B_k, $k = 0, 1, \ldots$, die in dieser formalen Reihenentwicklung auftreten, sind die sogenannten *Bernoulli-Zahlen*. Sie lauten

$$\begin{aligned}
&B_0 = 1\,, \quad B_1 = -1/2\,, \quad B_2 = 1/6\,,\\
&B_3 = B_5 = B_7 = \ldots = B_{2k+1} = \ldots = 0\,,\\
&B_4 = -1/30\,, \quad B_6 = 1/42\,, \quad \ldots\;;
\end{aligned}$$

die Bernoulli-Zahlen mit geradem Index wachsen sehr schnell an und verhalten sich für große k wie

$$|B_{2k}| \;\sim\; 2\,(2k)!\,(2\pi)^{-2k}\,, \qquad k \to \infty\,.$$

Aus (39.2) folgt für $k \geq 1$

$$\frac{\partial^k}{\partial x^k}\,g(x,h) \;=\; f^{(k-1)}(x+h) - f^{(k-1)}(x)\,.$$

Eingesetzt in (39.3) erhält man wegen $B_{2k+1} = 0$ für $k = 1, 2, \ldots$, somit

$$f(x)\,h \;=\; \int_x^{x+h} f(t)\,dt \;-\; \frac{1}{2}\,h\,\big(\,f(x+h) - f(x)\,\big) + \sum_{k=1}^{\infty} \frac{B_{2k}}{(2k)!}\,h^{2k}\,\big(\,f^{(2k-1)}(x+h) - f^{(2k-1)}(x)\,\big)$$

beziehungsweise

$$\frac{h}{2}\,\big(\,f(x+h) + f(x)\,\big) \;=\; \int_x^{x+h} f(t)\,dt \;+\; \sum_{k=1}^{\infty} \frac{B_{2k}}{(2k)!}\,\big(\,f^{(2k-1)}(x+h) - f^{(2k-1)}(x)\,\big)\,h^{2k}\,.$$

Ist $x = x_{i-1}$, $i = 1, \ldots, n$, ein Gitterpunkt der Trapezsumme und $h = (b-a)/n$, dann ist die linke Seite gerade die Näherung der Trapezformel für das Integral $\int_{x_{i-1}}^{x_i} f(t)\,dt$. Durch Summation von $i = 1$ bis n erhalten wir daher

$$T_n[f] \;=\; I[f] \;+\; \sum_{k=1}^{\infty} \underbrace{\frac{B_{2k}}{(2k)!}\,\big(\,f^{(2k-1)}(b) - f^{(2k-1)}(a)\,\big)}_{=:\tau_{2k}}\,h^{2k}\,. \tag{39.4}$$

Dies ist die sogenannte *Euler-Maclaurin-Summenformel*. Falls f weniger glatt ist oder nicht in eine konvergente Potenzreihe entwickelt werden kann, läßt sich zumindest das folgende Resultat beweisen.

Satz 39.2. *Sei $f \in C^{2m+2}[a,b]$, $h = (b-a)/n$ und τ_{2k} wie in* (39.4) *definiert. Dann gilt*

$$T_n[f] = I[f] + \sum_{k=1}^{m} \tau_{2k} h^{2k} + O(h^{2m+2})\,. \tag{39.5}$$

Dabei geht die Maximumnorm von $f^{(2m+2)}$ in den $O(\,\cdot\,)$-Fehlerterm ein.

Für einen Beweis dieses Satzes sei auf das Buch von Kreß [63] verwiesen, man vergleiche auch Aufgabe IX.9. Die Gleichung (39.5) liefert eine präzise asymptotische Entwicklung des Quadraturfehlers der Trapezsumme. Man beachte, daß für eine $(b-a)$-periodische Funktion f über $\mathbb{R}$ alle Koeffizienten τ_{2k} aus (39.4) verschwinden. Hat eine solche Funktion eine Potenzreihenentwicklung, die über das Intervall $[a,b]$ hinaus konvergiert, so ist die Konvergenz der Trapezsumme sogar linear, vgl. etwa [63].

Beispiel. Wir greifen noch einmal das einführende Beispiel 39.1 auf. Nach Satz 39.2 gilt für $f \in C^4[a,b]$

$$\begin{aligned} T_n[f] &= I[f] + \quad \tau_2\, h^2 + O(h^4)\,, \\ T_{2n}[f] &= I[f] + \tau_2/4\, h^2 + O(h^4)\,, \end{aligned}$$

jeweils mit $h = (b-a)/n$ und $\tau_2 = (f'(b) - f'(a))/12$. Somit ist

$$T_n[f] - T_{2n}[f] \;=\; \frac{3}{4}\,\tau_2\, h^2 + O(h^4)$$

beziehungsweise

$$\tau_2 h^2 \;=\; \frac{4}{3}\,\big(\,T_n[f] - T_{2n}[f]\,\big) \;+\; O(h^4)\,.$$

Damit folgt

$$T_{2n}[f] \;-\; \frac{1}{3}\,\big(\,T_n[f] - T_{2n}[f]\,\big) \;=\; I[f] + O(h^4)\,.$$

Die linke Seite ist gerade wieder das zusammengesetzte Simpson-Verfahren $S_n[f]$, vgl. Aufgabe 5. Auf diese Weise ergibt sich also ein weiterer Beweis für die höhere Konvergenzordnung des zusammengesetzten Simpson-Verfahrens gemäß Satz 38.4. ◊

Das *Romberg-Verfahren* perfektioniert die Vorgehensweise aus diesem Beispiel. Wir gehen davon aus, daß für eine gegebene Funktion $f \in C^{2m+2}[a,b]$ die Trapezsumme für m verschiedene Gitterweiten $h_i = (b-a)/n_i$, $i = 1,\ldots,m$, mit $h_1 > h_2 > \ldots > h_m$ berechnet worden ist. Setzen wir $\varphi(h^2)$ für die rechte Seite von (39.5), so läßt sich aus den berechneten Integralnäherungen das Polynom $p_{m-1} \in \Pi_{m-1}$ bestimmen, das die Funktion φ an den Abszissen h_i^2 interpoliert,

$$p_{m-1}(h_i^2) \;=\; \varphi(h_i^2) \;=\; T_{n_i}[f]\,, \qquad i = 1,\ldots,m\,.$$

Dies erlaubt dann die Approximation von $I[f] = \varphi(0)$ durch $p_{m-1}(0)$. Man nennt diese Vorgehensweise *Extrapolation*, da aus den Werten von φ an den Stellen $h_1^2,\ldots,h_m^2$ auf den Wert $\varphi(0)$ geschlossen werden soll und die Null nicht in der konvexen Hülle der h_i^2 liegt. Der Extrapolationsfehler $|p_{m-1}(0) - I[f]|$ wird in einer Arbeit von Bulirsch [12] genauer untersucht.

Für die Knotenzahlen n_i der Trapezsummen wählt man in der Praxis meist die *klassische Romberg-Folge*

$$n_i = 2^{i-1}\,, \qquad i = 1,2,\ldots\,,$$

oder die sogenannte *Bulirsch-Folge*

$$n_1 = 1\,, \quad n_{2i} = 2^i\,, \quad n_{2i+1} = 3\cdot 2^{i-1}\,, \qquad i = 1,2,\ldots$$

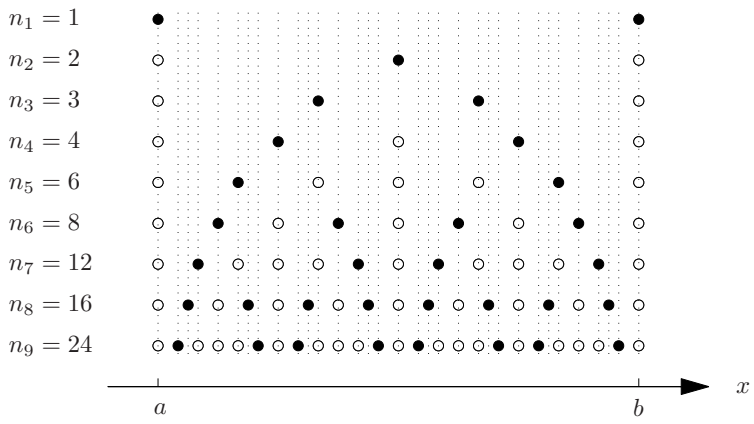


Abb. 39.1: Die zu berechnenden Funktionswerte bei der Bulirsch-Folge

Während die klassische Romberg-Folge zu einem einfacheren Rechenschema führt, wird in der Praxis zumeist die Bulirsch-Folge implementiert, da sie mit weniger Funktionsauswertungen auskommt. Für $m = 9$ werden beispielsweise bei der Bulirsch-Folge lediglich 33 Funktionsauswertungen benötigt (statt 257 Funktionsauswertungen bei der klassischen Romberg-Folge). Abbildung 39.1 soll dies veranschaulichen: Die Kreise kennzeichnen die Knoten der jeweiligen Gitter für die zusammengesetzte Trapezsumme mit der Bulirsch-Folge; ein weißer Kreis zeigt dabei an, daß der entsprechende Funktionswert $f(x_i)$ bereits für eine vorangegangene Trapeznäherung T_{n_i} berechnet wurde.

Das Interpolationspolynom p_{m-1} kann in der folgenden Weise rekursiv berechnet werden.

Lemma 39.3 (Aitken-Neville). *Für $1 \le i \le j \le m$ bezeichne $p_{(i,j)} \in \Pi_{j-i}$ das (eindeutig bestimmte) Polynom, das die Interpolationsaufgabe*

$$p_{(i,j)}(h_k^2) \;=\; T_{n_k}[f]\,, \qquad k = i, \ldots, j\,, \tag{39.6}$$

erfüllt. Dann gilt für $1 \le i \le j < m$:

$$p_{(i,j+1)}(x) \;=\; \frac{x - h_i^2}{h_{j+1}^2 - h_i^2}\, p_{(i+1,j+1)}(x) \;-\; \frac{x - h_{j+1}^2}{h_{j+1}^2 - h_i^2}\, p_{(i,j)}(x)\,. \tag{39.7}$$

Beweis. Um die Rekursion (39.7) für feste i und j mit $1 \le i \le j < m$ nachzuweisen, zeigen wir, daß das durch die rechte Seite von (39.7) definierte Polynom

$$p(x) \;=\; \frac{x - h_i^2}{h_{j+1}^2 - h_i^2}\, p_{(i+1,j+1)}(x) \;-\; \frac{x - h_{j+1}^2}{h_{j+1}^2 - h_i^2}\, p_{(i,j)}(x)$$

die entsprechenden Interpolationsbedingungen (39.6) erfüllt. Zunächst ist

$$p(h_i^2) \,=\, p_{(i,j)}(h_i^2) \,=\, T_{n_i}[f]\,, \qquad p(h_{j+1}^2) \,=\, p_{(i+1,j+1)}(h_{j+1}^2) \,=\, T_{n_{j+1}}[f]\,,$$

und für $i+1 \le k \le j$ ergibt sich

$$\begin{aligned} p(h_k^2) \,&=\, \frac{h_k^2 - h_i^2}{h_{j+1}^2 - h_i^2}\, \underbrace{p_{(i+1,j+1)}(h_k^2)}_{T_{n_k}[f]} - \frac{h_k^2 - h_{j+1}^2}{h_{j+1}^2 - h_i^2}\, \underbrace{p_{(i,j)}(h_k^2)}_{T_{n_k}[f]} \\ &=\, \frac{h_k^2 - h_i^2 - h_k^2 + h_{j+1}^2}{h_{j+1}^2 - h_i^2}\, T_{n_k}[f] \,=\, T_{n_k}[f]\,. \end{aligned}$$

Die Polynome p und $p_{(i,j+1)}$ erfüllen also die gleichen $j+2-i$ Interpolationsbedingungen und liegen beide in Π_{j-i+1}. Nach Satz 37.3 müssen sie daher übereinstimmen. Damit ist (39.7) bewiesen. □

Die Rekursion (39.7) kann nun verwendet werden, um die Integralnäherung

$$T_{(1,m)}[f] \,=\, p_{(1,m)}(0) \,=\, p_{m-1}(0) \,\approx\, \varphi(0) \,=\, I[f]$$

über die Zwischenergebnisse

$$T_{(j,j+k)}[f] \,=\, p_{(j,j+k)}(0)\,, \qquad k = 0, \ldots, m-1,\ j = 1, \ldots, m-k\,,$$

rekursiv zu berechnen. Aus (39.7) folgt nämlich

$$\begin{aligned} T_{(j,j+k)}[f] \,&=\, -\frac{h_j^2}{h_{j+k}^2 - h_j^2}\, T_{(j+1,j+k)}[f] \,+\, \frac{h_{j+k}^2}{h_{j+k}^2 - h_j^2}\, T_{(j,j+k-1)}[f] \\ &=\, T_{(j+1,j+k)}[f] \,+\, \frac{h_{j+k}^2}{h_j^2 - h_{j+k}^2} \left(T_{(j+1,j+k)}[f] - T_{(j,j+k-1)}[f] \right), \end{aligned}$$

und die Rechnung kann über zwei geschachtelte Schleifen wie in Algorithmus 39.1 organisiert werden.

Beispiele. Bei der klassischen Romberg-Folge ist $n_j = 2^{j-1}$ und somit $\nu_j^{(k)} = (n_{j+k}/n_j)^2 = 4^k$. Daher vereinfacht sich die Rekursion in Algorithmus 39.1 zu

$$\begin{aligned} T_{(j,j+k)}[f] \,&=\, \frac{4^k T_{(j+1,j+k)}[f] - T_{(j,j+k-1)}[f]}{4^k - 1} \\ &=\, T_{(j+1,j+k)}[f] \,+\, \frac{T_{(j+1,j+k)}[f] \,-\, T_{(j,j+k-1)}[f]}{4^k - 1}\,. \end{aligned} \tag{39.8}$$

Initialisierung: n_j, $j = 1, \ldots, m$, seien aufsteigend angeordnete natürliche Zahlen, etwa die Romberg-Folge oder die Bulirsch-Folge

berechne $T_{(j,j)}[f] = T_{n_j}[f]$, $j = 1, \ldots, m$
`for` $k = 1, \ldots, m-1$ `do`
  `for` $j = 1, \ldots, m-k$ `do`

$$\nu_j^{(k)} = \Big(\frac{h_j}{h_{j+k}}\Big)^2 = \Big(\frac{n_{j+k}}{n_j}\Big)^2$$

$$T_{(j,j+k)}[f] \;=\; T_{(j+1,j+k)}[f] \;+\; \frac{T_{(j+1,j+k)}[f] \;-\; T_{(j,j+k-1)}[f]}{\nu_j^{(k)} - 1}\,.$$

  `end for` % j-Schleife
`end for` % k-Schleife

Ergebnis: $I[f] \approx T_{(1,m)}[f]$

Algorithmus 39.1: Romberg-Verfahren

Für das Integral $\dfrac{1}{\pi}\displaystyle\int_0^\pi \sin^4 x\,dx$ lauten die ersten drei Trapeznäherungen

$$\begin{aligned} T_1 &= \frac{1}{2}\left(\sin^4 0 + \sin^4 \pi\right) \;=\; 0\,, \\ T_2 &= \frac{1}{4}\left(\sin^4 0 + 2\sin^4 \tfrac{\pi}{2} + \sin^4 \pi\right) \;=\; 1/2\,, \\ T_4 &= \frac{1}{8}\left(\sin^4 0 + 2\sin^4 \tfrac{\pi}{4} + 2\sin^4 \tfrac{\pi}{2} + 2\sin^4 \tfrac{3\pi}{4} + \sin^4 \pi\right) \;=\; 3/8\,. \end{aligned}$$

Das Romberg-Verfahren kann nun gemäß Algorithmus 39.1 bzw. (39.8) mit dem folgenden *Romberg-Tableau* ausgewertet werden:

$$\begin{array}{lcccc} T_{(1,1)} = T_1 = 0 & & & & \\ & \searrow^{-1/3} & & & \\ T_{(2,2)} = T_2 = 1/2 & \xrightarrow{4/3} & T_{(1,2)} = 2/3 & & \\ & \searrow^{-1/3} & & \searrow^{-1/15} & \\ T_{(3,3)} = T_4 = 3/8 & \xrightarrow{4/3} & T_{(2,3)} = 1/3 & \xrightarrow{16/15} & T_{(1,3)} = 14/45 \end{array}$$

Dabei werden jeweils zwei Näherungen gemäß der Verbindungslinien des Tableaus und den daran stehenden Faktoren linear kombiniert. Die gesuchte Approximation für den Integralwert ergibt sich rechts unten: $T_{(1,3)} = 14/45$.

Als ein zweites numerisches Beispiel wenden wir das Romberg-Verfahren auf das Integral aus Beispiel 38.5 an. Die Dreiecke in Abbildung 39.2 entsprechen den Näherungen $T_{(1,j)}$ auf der Diagonalen des Romberg-Tableaus; Dreiecke nach oben kennzeichnen Näherungen für die klassische Romberg-Folge,

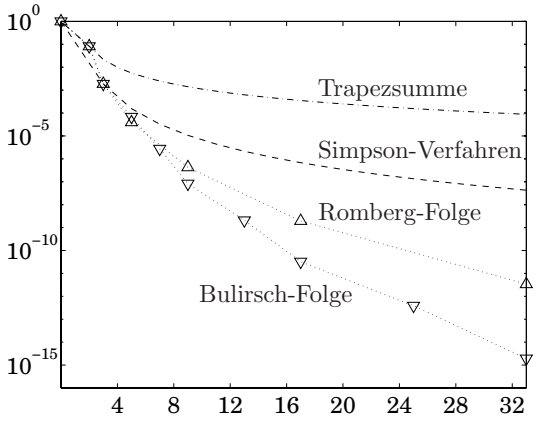


Abb. 39.2: Relative Fehler des Rombergverfahrens

Dreiecke nach unten gehören zu den Näherungen mit der Bulirsch-Folge. Die Abszisse einer Dreiecksposition gibt die Zahl der nötigen Auswertungen des Integranden an. Für $n_9 = 24$ werden bei der Bulirsch-Folge beispielsweise 33 Funktionsauswertungen benötigt und die resultierende Näherung des Romberg-Verfahrens ist fast im Bereich der Maschinengenauigkeit eps. Die gleiche Anzahl Funktionsauswertungen ergibt bei der klassischen Romberg-Folge lediglich eine Genauigkeit zwischen 10^{-11} und 10^{-12}. Zum Vergleich: Der relative Fehler des Simpson-Verfahrens (gebrochene Kurve) liegt bei 33 Funktionsauswertungen knapp unter 10^{-7}, der Fehler der Trapezsumme (Strichpunkte) liegt bei 10^{-4}. ◊

40 Gauß-Quadratur

Wir beginnen diesen Abschnitt mit dem folgenden

Problem 40.1. *Gegeben seien m paarweise verschiedene Knoten $x_1, \ldots, x_m$ und m Gewichte $w_1, \ldots, w_m$. Wie groß ist maximal der Exakheitsgrad der Quadraturformel*

$$Q[f] = \sum_{i=1}^{m} w_i f(x_i) \approx I[f; w] \; ? \tag{40.1}$$

Dabei ist w wie zuvor eine fest vorgegebene Gewichtsfunktion über dem Integrationsintervall $\mathcal{I} = [a, b]$.

Zu beachten ist die von den vorhergehenden Abschnitten abweichende Indizierung der Knoten und Gewichte der Quadraturformel $Q[\,\cdot\,]$. Während bisher der Index immer von $i = 0$ bis m lief, läuft er hier und im folgenden lediglich von $i = 1$ bis m. Unter Berücksichtigung dieser Änderung kann aus Proposition 38.1 unmittelbar geschlossen werden, daß der maximale Exaktheitsgrad mindestens $m - 1$ ist, falls die Gewichte w_i gemäß (38.1) durch Integration der Lagrange-Grundpolynome bestimmt werden. Das folgende Resultat liefert eine Oberschranke für den größtmöglichen Exaktheitsgrad.

Proposition 40.2. *Sei $w > 0$ in (a, b). Dann ist der Exaktheitsgrad der Quadraturformel $Q[\,\cdot\,]$ aus (40.1) höchstens $q = 2m - 1$.*

Beweis. Wir betrachten das quadrierte Knotenpolynom

$$p(x) \;=\; \omega^2(x) \;=\; \prod_{i=1}^{m} (x - x_i)^2 \in \Pi_{2m}\,.$$

Offensichtlich ist $Q[p] = 0$ und

$$I[p; w] \;=\; \int_a^b \prod_{i=1}^{m} (x - x_i)^2 w(x)\, dx \;>\; 0\,,$$

da alle Faktoren nichtnegativ und auf einer offenen Teilmenge positiv sind. Daher ist $Q[p] \neq I[p; w]$ und der Exaktheitsgrad der Quadraturformel ist höchstens $2m - 1$. □

Um zu untersuchen, ob Proposition 40.2 scharf ist, leiten wir im weiteren notwendige Bedingungen an eine Quadraturformel mit Exaktheitsgrad $2m-1$ her. Motiviert durch den Beweis von Proposition 40.2 betrachten wir weiterhin das Knotenpolynom

$$\omega(x) \;=\; \prod_{i=1}^{m} (x - x_i) \in \Pi_m$$

und ein beliebiges Polynom $p \in \Pi_{m-1}$. Dann gehört das Produkt ωp zu Π_{2m-1} und es ist $(\omega p)(x_i) = 0$, $i = 1, \ldots, m$. Nehmen wir nun an, die Quadraturformel $Q[\,\cdot\,]$ habe den Exaktheitsgrad $q = 2m - 1$. Dann folgt

$$0 \;=\; Q[\omega p] \;=\; I[\omega p; w] \;=\; \int_a^b \omega(x) p(x) w(x)\, dx\,; \tag{40.2}$$

mit anderen Worten: ω steht *orthogonal* auf dem gesamten Unterraum Π_{m-1} bezüglich des Innenprodukts

$$\langle \varphi, \psi \rangle \;=\; \int_a^b \varphi(x) \psi(x) w(x)\, dx\,. \tag{40.3}$$

Nach Satz 33.1 muß daher das Knotenpolynom ω ein skalares Vielfaches des m-ten Orthonormalpolynoms u_m zu der Gewichtsfunktion w sein. Insbesondere müssen die Nullstellen von ω und u_m übereinstimmen. Die Knoten x_i, $i = 1, \ldots, m$, der Quadraturformel (40.1) müssen also die nach Satz 34.1 paarweise verschiedenen und im offenen Intervall (a, b) gelegenen Nullstellen des eindeutig bestimmten Orthonormalpolynoms u_m sein. Dies ist eine erste notwendige Bedingung an eine Quadraturformel $Q[\,\cdot\,]$ mit Exaktheitsgrad $q = 2m - 1$.

Ferner sind nach Proposition 38.1 die Gewichte $w_i, i = 1, \ldots, m$, der Quadraturformel $Q[\,\cdot\,]$ durch die Integrale der Lagrange-Grundpolynome l_i eindeutig festgelegt,

$$w_i \;=\; \int_a^b l_i(x) w(x)\, dx\,, \qquad i = 1, \ldots, m\,, \tag{40.4}$$

vgl. (38.1). Es bleibt also eine einzige Quadraturformel als möglicher Kandidat für den maximal erreichbaren Exaktheitsgrad $q = 2m - 1$ übrig. Dies ist die sogenannte *m-stufige Gauß-Formel* $G_m[\,\cdot\,; w]$ zu der Gewichtsfunktion w.

Satz 40.3. *Die m-stufige Gauß-Formel mit den Nullstellen x_i, $i = 1, \ldots, m$, des m-ten Orthonormalpolynoms bezüglich des Innenprodukts* (40.3) *als Knoten und den Gewichten w_i aus* (40.4) *hat den maximal möglichen Exaktheitsgrad $q = 2m - 1$. Die Gewichte sind dabei allesamt positiv und hängen über*

$$w_i \;=\; \Lambda_m(x_i)\,, \qquad i = 1, \ldots, m\,, \tag{40.5}$$

mit den Christoffel-Funktionen aus Definition 33.3 *zusammen.*

Beweis. Um den Exaktheitsgrad $q = 2m - 1$ nachzuweisen, betrachten wir die Basis $\big\{\, 1, x, \ldots, x^{m-1}, \omega(x), x\,\omega(x), \ldots, x^{m-1}\omega(x) \,\big\}$ von Π_{2m-1}. Wegen der Wahl der Gewichte (und unter Berücksichtigung der modifizierten Indizierung) ist die Gauß-Formel nach Proposition 38.1 exakt für die Monome $1, x, \ldots, x^{m-1}$. Für die Basisfunktionen $x^j\omega(x)$, $j = 0, \ldots, m-1$, folgt die Exaktheit aus (40.2), denn die rechte Seite ist wegen der Orthogonalität von ω und Π_{m-1} gleich Null. Also hat $G_m[\,\cdot\,; w]$ den Exaktheitsgrad $q = 2m - 1$.

Für den Nachweis von (40.5) entwickeln wir das j-te Lagrange-Grundpolynom $l_j \in \Pi_{m-1}$ in der Orthonormalpolynombasis $\{u_k\}_{k=0}^{m-1}$ und erhalten

$$l_j = \sum_{k=0}^{m-1} \alpha_k u_k$$

mit

$$\alpha_k \;=\; \langle\, u_k, l_j \,\rangle \;=\; \int_a^b u_k(x) l_j(x) w(x)\, dx \;=\; I[u_k l_j; w]\,,$$

vgl. Satz 31.6. Aus $k \le m-1$ folgt $u_k l_j \in \Pi_{2m-2}$ und somit ist $I[u_k l_j; w] = G_m[u_k l_j; w]$ und

$$\alpha_k \;=\; G_m[u_k l_j; w] \;=\; \sum_{i=1}^{m} w_i u_k(x_i) \underbrace{l_j(x_i)}_{=\delta_{ij}} \;=\; w_j u_k(x_j)\,.$$

Für die entsprechende Norm von l_j folgt somit aus dem Satz von Pythagoras

$$\|l_j\|_w^2 \;=\; \langle\, l_j, l_j \,\rangle \;=\; \sum_{k=0}^{m-1} \alpha_k^2 \;=\; w_j^2 \sum_{k=0}^{m-1} u_k^2(x_j)\,. \tag{40.6}$$

Andererseits ist – wiederum wegen des Exaktheitsgrads von $G_m[\,\cdot\,; w]$ –

$$\|l_j\|_w^2 \;=\; I[\,l_j^2; w] \;=\; G_m[\,l_j^2; w] \;=\; \sum_{i=1}^{m} w_i l_j^2(x_i) \;=\; w_j\,.$$

Hieraus folgt zunächst die Positivität von w_j und durch einen Vergleich mit (40.6)

$$w_j \;=\; w_j^2 \sum_{k=0}^{m-1} u_k^2(x_j) \;=\; w_j^2 \Lambda_m^{-1}(x_j)\,.$$

Damit ist der Satz vollständig bewiesen. □

Bemerkung. Nach Satz 33.4 ist $\Lambda_m(x_i)$ das Minimum von $\|p\|_w^2$, wobei p alle Polynome aus Π_{m-1} mit $p(x_i) = 1$ durchläuft. Ferner ist das minimierende Polynom ein Vielfaches des Kernpolynoms aus Definition 33.3 vom Grad $m-1$ mit $\xi = x_i$. Wegen (40.6) und (40.5) ist das minimierende Polynom offensichtlich gerade das entsprechende Lagrange-Grundpolynom l_i. ◊

Beispiel 40.4. Für die Tschebyscheff-Gewichtsfunktion $w(x) = (1-x^2)^{-1/2}$ im Intervall $(-1, 1)$ sind die Knoten der m-ten Gauß-Formel die Nullstellen des m-ten Tschebyscheff-Polynoms T_m,

$$x_i = \cos\,\frac{2i-1}{2m}\,\pi, \qquad i = 1, \ldots, m\,,$$

vgl. Beispiel 34.7. Nach (32.5) sind die orthonormierten Tschebyscheff-Polynome durch $u_0 = 1/\sqrt{\pi}$ und $u_k = \sqrt{2/\pi}\,T_k$ für $k \ge 1$ gegeben. Aus (40.5) und der expliziten Darstellung (32.1) der Tschebyscheff-Polynome erhalten wir somit

$$w_i \;=\; \Big(\sum_{k=0}^{m-1} u_k^2(x_i)\Big)^{-1} \;=\; \Big(\frac{1}{\pi} + \frac{2}{\pi}\sum_{k=1}^{m-1} \cos^2\big(k(2i-1)\frac{\pi}{2m}\big)\Big)^{-1}.$$

Mit der Kosinus-Halbwinkelformel folgt hieraus

$$w_i = \Big(\frac{1}{\pi} + \frac{1}{\pi}\sum_{k=1}^{m-1}\Big(1+\cos\big(k(2i-1)\frac{\pi}{m}\big)\Big)\Big)^{-1}$$
$$= \Big(\frac{m}{\pi} + \frac{1}{\pi}\sum_{k=1}^{m-1}\cos\big(k(2i-1)\frac{\pi}{m}\big)\Big)^{-1}.$$

In der letzten Summe treten jeweils für $k = l$ und $k = m - l$ die gleichen Summanden mit gegengesetzten Vorzeichen auf. Folglich hat diese Summe den Wert Null und es ergibt sich $w_i = \pi/m$. Damit lautet die m-stufige *Gauß-Tschebyscheff-Formel*

$$\int_{-1}^{1} \frac{f(x)}{\sqrt{1-x^2}}\,dx \;\approx\; \frac{\pi}{m}\sum_{i=1}^{m} f\big(\cos\frac{2i-1}{2m}\pi\big)\,. \tag{40.7}$$

◊

Zum Abschluß dieses Abschnitts schätzen wir noch den Quadraturfehler der Gauß-Formeln ab.

Satz 40.5. *Sei $f \in C^{2m}[a,b]$ und $G_m[\,\cdot\,;w]$ die m-stufige Gauß-Formel für $I[\,\cdot\,;w]$. Dann gilt*

$$\big|\, I[f;w] - G_m[f;w]\,\big| \;\le\; \frac{\|f^{(2m)}\|_{[a,b]}}{(2m)!\,\gamma_m^2}\,,$$

wobei γ_m wie in Satz 33.1 *den Höchstkoeffizienten des Orthonormalpolynoms u_m bezeichnet.*

Beweis. Wie bereits bei der Simpson-Formel greifen wir auf die in Bemerkung 37.6 angesprochene Hermite-Interpolation zurück und interpolieren die Funktionswerte $f(x_i)$ zusammen mit den Ableitungen $f'(x_i)$ in *allen* Knoten x_i der Gauß-Formel durch ein Polynom $p_{2m-1} \in \Pi_{2m-1}$. Die entsprechende Verallgemeinerung von Satz 37.4 liefert für jedes $x \in [a,b]$ ein $\xi \in (a,b)$ mit

$$f(x) - p_{2m-1}(x) \;=\; \frac{f^{(2m)}(\xi)}{(2m)!}\prod_{i=1}^{m}(x-x_i)^2\,.$$

Da die Knoten x_i die Nullstellen des m-ten Orthonormalpolynoms sind, gilt also

$$f(x) - p_{2m-1}(x) \;=\; \frac{f^{(2m)}(\xi)}{(2m)!}\,\frac{u_m^2(x)}{\gamma_m^2}\,,$$

und wegen des Exaktheitsgrads von $G_m[\,\cdot\,;w]$ folgt

$$\begin{aligned} \big|\, I[f;w] - G_m[f;w] \,\big| \;&=\; \big|\, I[f;w] - G_m[p_{2m-1};w] \,\big| \\ &=\; \Big|\, \int_a^b (f-p_{2m-1})(x)w(x)\,dx \,\Big| \;\le\; \frac{\|f^{(2m)}\|_{[a,b]}}{(2m)!\,\gamma_m^2} \underbrace{\int_a^b u_m^2(x)w(x)\,dx}_{=1} \\ &=\; \frac{\|f^{(2m)}\|_{[a,b]}}{(2m)!\,\gamma_m^2}\,. \end{aligned}$$

□

Bemerkung. Speziell für die Gauß-Tschebyscheff-Formel (40.7) ist

$$\gamma_0^2 = 1/\pi\,, \qquad \gamma_m^2 = 4^m/(2\pi)\,, \quad m \ge 1\,,$$

vgl. (33.5). ◊

41 Gauß-Legendre-Formeln

Leider lassen sich nur die wenigsten Gauß-Formeln so wie die Gauß-Tschebyscheff-Formeln explizit angeben. Insbesondere für den wichtigen Spezialfall $w = 1$, der zu der klassischen $\mathcal{L}^2$-Norm gehört, können die Knoten der Gauß-Formeln nur für kleine m explizit angegeben werden. Im Intervall $[-1,1]$ sind dies die Nullstellen der Legendre-Polynome aus Beispiel 33.2. Die zugehörige Gauß-Formel heißt daher *Gauß-Legendre-Formel* $G_m[\,\cdot\,;w=1]$ oder kurz $G_m[\,\cdot\,]$.

Beispiel 41.1. Die erste Gauß-Legendre-Formel ist die Mittelpunktformel, $G_1[f] = M[f]$, denn sie hat nach Lemma 38.3 den Exaktheitsgrad $q = 1$ und ist, wie wir im vorigen Abschnitt gesehen haben, dadurch eindeutig als Gauß-Formel festgelegt. Nach Beispiel 33.2 ist

$$P_2(x) \;=\; \frac{3}{2}\,x^2 - \frac{1}{2}$$

das Legendre-Polynome vom Grad 2. Folglich sind die Nullstellen $x_{1/2} = \pm 1/\sqrt{3}$ von P_2 die Knoten von $G_2[\,\cdot\,]$. Damit $G_2[f]$ für $f(x) = x$ den Integralwert Null ergibt, müssen die beiden Gewichte w_1 und w_2 gleich sein. Somit ist

$$G_2[f] \;=\; f(-1/\sqrt{3}) + f(1/\sqrt{3})$$

die zweistufige Gauß-Legendre-Formel. Die dritte Gauß-Legendre-Formel läßt sich ebenfalls noch explizit angeben:

$$G_3[f] \;=\; \frac{5}{9}\,f(-\sqrt{15}/5) \;+\; \frac{8}{9}\,f(0) \;+\; \frac{5}{9}\,f(\sqrt{15}/5)\,.$$

◊

Die Gauß-Legendre-Formeln sind in der Formelsammlung [1, Table 25.4] für verschiedene m tabelliert. Sie können mit dem folgenden Algorithmus von Golub und Welsch [35] numerisch bestimmt werden. Dieses Verfahren beruht auf der bekannten Rekursionsformel der Legendre-Polynome und den Resultaten aus den Abschnitten 34 und 35.2.

Wir erinnern zunächst an Satz 34.6, nach dem die Nullstellen x_i, $i = 1, \dots, m$, des m-ten Legendre-Polynoms, also die Knoten der m-stufigen Gauß-Legendre-Formel, gerade die Eigenwerte der m-ten Jacobi-Matrix dieser Orthonormalpolynome sind. Nach Beispiel 33.2 genügen die orthonormierten Legendre-Polynome der Rekursionsformel (33.9),

$$\beta_{n+1} u_{n+1}(x) \;=\; x u_n(x) - \beta_n u_{n-1}(x)\,, \qquad n \in \mathbb{N}_0\,, \tag{41.1}$$

mit $u_{-1} = 0$, $u_0 = 1/\sqrt{2}$ und

$$\beta_n \;=\; \frac{n}{(4n^2-1)^{1/2}}\,, \qquad n \in \mathbb{N}\,. \tag{41.2}$$

Die zugehörige $m \times m$-Jacobi-Matrix lautet

$$J_m \;=\; \begin{bmatrix} 0 & \beta_1 & & & 0 \\ \beta_1 & 0 & \ddots & & \\ & \ddots & \ddots & \beta_{m-2} & \\ & & \beta_{m-2} & 0 & \beta_{m-1} \\ 0 & & & \beta_{m-1} & 0 \end{bmatrix} \in \mathbb{R}^{m\times m}, \tag{41.3}$$

vgl. (34.3).

Die Gewichte w_i, $i = 1, \dots, m$, der Gauß-Formel ergeben sich nach Satz 40.3 aus den Christoffel-Funktionen, $w_i = \Lambda_m(x_i)$, und demzufolge besteht nach Satz 34.6 der folgende Zusammenhang zwischen dem Gewicht w_i und der ersten Komponente v_{1i} eines Eigenvektors $v_i \in \mathbb{R}^m$ von J_m zum Eigenwert x_i:

$$v_{1i}^2 / \|v_i\|_2^2 \;=\; \Lambda_m(x_i) u_0^2(x_i) \;=\; w_i u_0^2(x_i) \;=\; w_i/2\,.$$

Wir fassen diese Ergebnisse in dem folgenden Resultat zusammen:

Proposition 41.2. *Sei J_m aus* (41.3) *die $m \times m$-Jacobi-Matrix der Legendre-Polynome mit den Eigenwerten λ_i und zugehörigen Eigenvektoren $v_i = [v_{ji}]_{j=1}^m$, $i = 1, \dots, m$. Dann lautet die m-te Gauß-Legendre-Formel*

$$G_m[f] \;=\; \sum_{i=1}^{m} \frac{2 v_{1i}^2}{\|v_i\|_2^2}\, f(\lambda_i)\,. \tag{41.4}$$

Aufwand. Da J_m eine Tridiagonalmatrix ist, können ihre Eigenwerte und Eigenvektoren mit etwa $O(m^2)$ Operationen berechnet werden, vgl. Abschnitt 29. ◊

Beispiel 41.3. Als Beispiel für den Algorithmus von Golub und Welsh greifen wir auf das Beispiel aus Abschnitt 27.7 zurück. Dort wurden die Eigenwerte der Matrix $(J_5 + I)/2$ berechnet, also die Stützstellen der fünften Gauß-Legendre-Formel für das Intervall $[0, 1]$. In Beispiel 29.5 wurde ferner anhand des größten Eigenwerts $\lambda_5 = 0.95308992296933...$ demonstriert, wie die zugehörigen Eigenvektoren effizient und stabil berechnet werden. Aus dem dort angegebenen Ergebnis errechnet sich das zugehörige Gewicht w_5 der Quadraturformel zu

$$w_5 \;=\; \frac{v_{15}^2}{\|v_5\|_2^2} \;=\; 0.11846344252809...$$

Da hier $\mathcal{I} = [0, 1]$ das Integrationsintervall ist, fehlt der Faktor 2 aus (41.4). ◊

Aus der allgemeinen Theorie zur Gauß-Quadratur aus dem vorangegangenen Abschnitt ergibt sich die folgende Fehlerabschätzung für die m-stufige Gauß-Legendre-Formel.

Satz 41.4. *Für $f \in C^{2m}[-1, 1]$ ist*

$$\big|\, I[f] - G_m[f] \,\big| \;\le\; \varepsilon_m \, \|f^{(2m)}\|_{[-1,1]}$$

mit

$$\varepsilon_m \;=\; \frac{2}{2m+1}\,\frac{4^m (m!)^4}{\big((2m)!\big)^3} \;=\; \frac{\sqrt{\pi}}{2\sqrt{m}}\,(4em)^{-2m}\big(1 + O(\tfrac{1}{m})\big)\,, \quad m \to \infty\,.$$

Beweis. In Beispiel 33.2 haben wir den Höchstkoeffizienten γ_m des m-ten orthonormierten Legendre-Polynoms ausgerechnet. Demnach ist

$$\gamma_m^2 \;=\; \frac{2m+1}{2}\,\frac{\big((2m)!\big)^2}{4^m (m!)^4}\,,$$

vgl. (33.8), und somit ergibt sich die gewünschte Fehlerabschätzung mit dem exakten Wert ε_m aus dem allgemeinen Satz 40.5. Das asymptotische Verhalten von ε_m für $m \to \infty$ folgt aus der *Stirling-Formel*

$$m! \;=\; \sqrt{2\pi} e^{-m-1} (m+1)^{m+1/2} \big(1 + O(\tfrac{1}{m})\big)\,, \qquad m \to \infty\,. \tag{41.5}$$

□

Schließlich konstruieren wir noch die *Radau-Legendre-Formeln*, die unter anderem bei der numerischen Lösung von Anfangswertaufgaben bei gewöhnlichen

Differentialgleichungen eine wichtige Rolle spielen, vgl. Abschnitt 79. Die m-stufige Radau-Legendre-Formel $R_m[\,\cdot\,]$ ist eine Quadraturformel für $I[\,\cdot\,]$, bei der ein Knoten von vornherein festgelegt wird, nämlich $\tilde{x}_m = 1$:

$$R_m[f] \;=\; \tilde{w}_m f(1) + \sum_{i=1}^{m-1} \tilde{w}_i f(\tilde{x}_i) \;\approx\; \int_{-1}^{1} f(x)\,dx\,. \tag{41.6}$$

Dabei soll R_m (unter der Nebenbedingung $\tilde{x}_m = 1$) den maximal möglichen Exaktheitsgrad haben. Da $\tilde{x}_m = 1$ ein Randpunkt des Intervalls $[-1,1]$ ist und die Nullstellen der Legendre-Polynome allesamt im Innern dieses Intervalls liegen, sind die m-stufige Radau-Legendre-Formel und die entsprechende Gauß-Legendre-Formel niemals identisch, so daß der Exaktheitsgrad von $R_m[\,\cdot\,]$ maximal $2m-2$ sein kann. Dieser Exaktheitsgrad wird auch tatsächlich erreicht, wie im weiteren gezeigt werden soll.

Dazu erinnern wir zunächst an das inverse Problem 35.2 aus Abschnitt 35.2: Modifizieren wir das rechte untere Eckelement der Jacobi-Matrix J_m aus (41.3) mit den Koeffizienten β_n aus (41.2) gemäß

$$\tilde{J}_m \;=\; \begin{bmatrix} 0 & \beta_1 & & & 0 \\ \beta_1 & 0 & \ddots & & \\ & \ddots & \ddots & \beta_{m-2} & \\ & & \beta_{m-2} & 0 & \beta_{m-1} \\ 0 & & & \beta_{m-1} & \alpha_m \end{bmatrix} \in \mathbb{R}^{m\times m}, \qquad \alpha_m \in \mathbb{R}\,, \tag{41.7}$$

so existiert nach Satz 35.3 ein diskretes Innenprodukt

$$\langle\!\langle\, p, q \,\rangle\!\rangle \;=\; \sum_{i=1}^{m} \tilde{v}_{1i}^2\, p(\tilde{x}_i) q(\tilde{x}_i) \tag{41.8}$$

im Raum Π_{m-1}, für das die Polynome $\{\tilde{u}_n\}_{n=0}^{m-1}$ aus der Rekursion

$$\beta_{n+1}\tilde{u}_{n+1}(x) \;=\; x\tilde{u}_n(x) - \beta_n \tilde{u}_{n-1}(x) \;\in \Pi_{n+1}\,, \tag{41.9}$$

$n = 0, 1, \ldots, m-2$, mit $\tilde{u}_{-1} = 0$ und $\tilde{u}_0 = 1$ die zugehörigen Orthonormalpolynome sind. Dabei sind die Stützstellen $\tilde{x}_i$ von (41.8) die Eigenwerte von $\tilde{J}_m$ und die zugehörigen Gewichte die jeweils erste Komponente $\tilde{v}_{1i}$ des bezüglich der Euklidnorm zu Eins normierten entsprechenden Eigenvektors von $\tilde{J}_m$. Man beachte die Analogie zur Gauß-Legendre-Formel.

Offensichtlich stimmt die Rekursion (41.9) (bis auf die Initialisierung von $\tilde{u}_0$) mit der Rekursion (41.1) der orthonormierten Legendre-Polynome $\{u_n\}$ überein. Daher ergibt sich unmittelbar

$$\tilde{u}_n(x) \;=\; \sqrt{2}\,u_n(x)\,, \qquad n = 0, 1, \ldots, m-1\,. \tag{41.10}$$

Proposition 41.5. *Sei $\alpha_m > 0$ in (41.7) eine beliebige reelle Zahl. Dann hat die Quadraturformel*

$$\sum_{i=1}^{m} \tilde{w}_i \, f(\tilde{x}_i) \;\approx\; \int_{-1}^{1} f(x)\, dx \tag{41.11}$$

mit den oben spezifizierten Knoten $\tilde{x}_i$ und den positiven Gewichten $\tilde{w}_i = 2\tilde{v}_{1i}^2$, $i = 1, \ldots, m$, mindestens den Exaktheitsgrad $2m-2$.

Beweis. Nach Satz 35.3 sind die Faktoren $\tilde{v}_{1i}^2$ aus (41.8) und damit die Gewichte $\tilde{w}_i$ für $i = 1, \ldots, m$ positiv. Für die Abschätzung des Exaktheitsgrades wählen wir ein beliebiges $n \le 2m-2$ und faktorisieren das Monom $f(x) = x^n$ in zwei Monome $p(x) = x^k$ und $q(x) = x^l$ mit $0 \le k, l \le m-1$, $k + l = n$. Entwickeln wir p und q in die Legendre-Polynome,

$$p(x) \;=\; \sum_{i=0}^{m-1} a_i u_i(x)\,, \qquad q(x) \;=\; \sum_{i=0}^{m-1} b_i u_i(x)\,, \tag{41.12}$$

so gilt mit der Notation $\langle \cdot\,, \cdot \rangle$ für das $\mathcal{L}^2$-Innenprodukt über $(-1, 1)$

$$\int_{-1}^{1} x^n\, dx \;=\; \langle\, p, q\,\rangle \;=\; \sum_{i,j=0}^{m-1} a_i b_j \langle\, u_i, u_j\,\rangle \;=\; \sum_{i=0}^{m-1} a_i b_i\,. \tag{41.13}$$

Auf der anderen Seite folgt aus (41.12) und (41.10)

$$p(x) \;=\; \frac{1}{\sqrt{2}} \sum_{i=0}^{m-1} a_i \tilde{u}_i(x)\,, \qquad q(x) \;=\; \frac{1}{\sqrt{2}} \sum_{i=0}^{m-1} b_i \tilde{u}_i(x)\,,$$

also die Entwicklung von p und q in die Orthonormalpolynome $\{\tilde{u}_n\}$ zu dem Innenprodukt (41.8). Folglich ist

$$\langle\!\langle\, p, q\,\rangle\!\rangle \;=\; \frac{1}{2} \sum_{i,j=0}^{m-1} a_i b_j \langle\!\langle\, \tilde{u}_i, \tilde{u}_j\,\rangle\!\rangle \;=\; \frac{1}{2} \sum_{i=0}^{m-1} a_i b_i\,,$$

und ein Vergleich mit (41.13) ergibt

$$\int_{-1}^{1} x^n\, dx \;=\; 2\langle\!\langle\, p, q\,\rangle\!\rangle \overset{(41.8)}{=} 2 \sum_{i=1}^{m} \tilde{v}_{1i}^2 \tilde{x}_i^k \tilde{x}_i^l \;=\; \sum_{i=1}^{m} \tilde{w}_i \tilde{x}_i^n\,.$$

Demnach ist die Quadraturformel (41.11) für alle Polynome vom Grad kleiner oder gleich $2m-2$ exakt. □

Die Aussage von Proposition 41.5 ist unabhängig von der Wahl des bislang freien Parameters α_m in der Matrix $\tilde{J}_m$. Für die Konstruktion der Radau-Legendre-Formel wählt man α_m derart, daß der größte Eigenwert $\tilde{x}_m$ von $\tilde{J}_m$ gleich Eins ist.

Lemma 41.6. *Es gibt genau ein $\alpha_m \in \mathbb{R}$, nämlich*

$$\alpha_m \;=\; 1/e_m^*(I - J_m)^{-1}e_m \;=\; \frac{m}{2m-1}\,,$$

für das die Matrix $\tilde{J}_m$ aus (41.7) *den Eigenwert $\lambda = 1$ besitzt.*

Beweis. Wir bezeichnen wieder mit $e_m \in \mathbb{R}^m$ den m-ten kartesischen Einheitsvektor. Dann gilt $\tilde{J}_m = J_m + \alpha_m e_m e_m^*$.

Sei $v \in \mathbb{R}^m$ ein Eigenvektor von $J_m + \alpha e_m e_m^*$ zum Eigenwert $\lambda = 1$. Dann gilt

$$v \;=\; (J_m + \alpha e_m e_m^*)v \;=\; J_m v + \alpha(e_m^* v)e_m\,, \tag{41.14}$$

d. h. $(J_m - I)v$ ist ein Vielfaches von e_m, und zwar ein von Null verschiedenes Vielfaches. Letzteres liegt daran, daß die Eigenwerte von J_m die Nullstellen des m-ten Legendre-Polynoms sind, die allesamt in $(-1, 1)$ liegen, so daß $J_m - I$ invertierbar ist. Normieren wir v derart, daß

$$(I - J_m)v \;=\; e_m \tag{41.15}$$

erfüllt ist, so folgt aus (41.14) unmittelbar

$$\alpha(e_m^* v) = 1$$

und somit gilt

$$\alpha \;=\; 1/e_m^* v \;=\; 1/e_m^*(I - J_m)^{-1}e_m \;=\; \alpha_m\,.$$

Umgekehrt sieht man, daß für v aus (41.15) und $\alpha = \alpha_m$ die Eigenwertgleichung (41.14) erfüllt ist.

Für die zweite Darstellung von α_m verwenden wir wieder die Aussage aus Satz 35.3, wonach das Polynom

$$\tilde{u}_m(x) \;=\; (x - \alpha_m)\tilde{u}_{m-1}(x) - \beta_{m-1}\tilde{u}_{m-2}(x)$$

in $x = 1$ eine Nullstelle besitzt, sofern $\lambda = 1$ ein Eigenwert von $\tilde{J}_m$ ist. Dies ergibt die Gleichung

$$(1 - \alpha_m)\tilde{u}_{m-1}(1) \;=\; \beta_{m-1}\tilde{u}_{m-2}(1)\,, \tag{41.16}$$

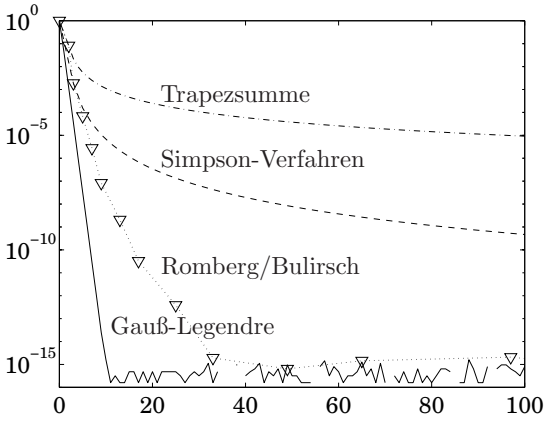


Abb. 41.1: Relative Fehler der Quadraturverfahren

zu deren Lösung die Funktionswerte von $\tilde{u}_n$ an der Stelle $x = 1$ benötigt werden. Letztere ergeben sich unmittelbar aus (41.9) durch Induktion über n:

$$\tilde{u}_n(1) \;=\; \sqrt{2n+1}\,, \qquad n = 0, \ldots, m-1\,.$$

Eingesetzt in (41.16) erhalten wir somit

$$(1-\alpha_m)\sqrt{2m-1} \;=\; \frac{m-1}{\sqrt{(2m-1)(2m-3)}}\sqrt{2m-3}\,,$$

woraus die Behauptung folgt. □

Die m-te Radau-Legendre-Formel ist also die Quadraturformel (41.11), die sich für $\alpha_m = m/(2m-1)$ ergibt. Der Aufwand zur Berechnung dieser Formel ist der gleiche wie zur Berechnung der Gauß-Legendre-Formel.

Radau-Formeln lassen sich wie Gauß-Formeln auch für allgemeinere Gewichtsfunktionen herleiten. Im allgemeinen muß dann der Parameter α_m aus der ersten Identität von Lemma 41.6 durch Lösen des Gleichungssystems (41.15) bestimmt werden.

Beispiel. Wir vergleichen die verschiedenen Quadraturverfahren noch einmal anhand des Integrals

$$\int_0^1 \frac{1}{x+1}\; dx$$

aus Beispiel 38.5. Abbildung 41.1 zeigt neben den bereits bekannten Fehlerkurven für die Trapezsumme, das zusammengesetzte Simpson-Verfahren so-

wie das Romberg-Verfahren (mit der Bulirsch-Folge) die Ergebnisse der Gauß-Legendre-Quadratur bei jeweils gleicher Knotenzahl. Da hier $[0,1]$ das Integrationsintervall ist, müssen die oben berechneten Knoten x_i der Gauß-Legendre-Formeln wie in Beispiel 41.3 linear auf das Intervall $[0,1]$ transformiert werden und die Gewichte w_i entsprechend halbiert werden. Die Konvergenzkurve der Gauß-Legendre-Quadraturformeln zeigt eine im wesentlichen lineare und äußerst schnelle Konvergenz. Um alle signifikanten Stellen des Integralwerts zu berechnen, werden nur etwa 12 Funktionsauswertungen benötigt. Das ist um einen Faktor drei bis vier besser als beim Romberg-Verfahren. Abbildung 41.1 enthält keine Näherungen der Radau-Legendre-Formeln, da diese mit den Näherungen der jeweiligen Gauß-Legendre-Formeln nahezu übereinstimmen. ◊

Man beachte allerdings, daß die Konvergenzgeschwindigkeit bei allen Quadraturformeln entscheidend von dem Integranden abhängt. Die Resultate aus Abbildung 41.1 sind daher nur eingeschränkt auf andere Integranden übertragbar.

42 Ein adaptives Quadraturverfahren

Bei der Approximation eines konkreten Integrals sind die A-priori-Fehlerschranken der vorangegangenen Abschnitte oft nicht anwendbar, da sie eine scharfe Abschätzung einer höheren Ableitung des Integranden benötigen und den Fehler dennoch oft nur grob einschließen.

Zudem gehen diese Abschätzungen davon aus, daß das Integrationsintervall gleichmäßig unterteilt ist – speziell bei Integranden mit Singularitäten ist es hingegen sinnvoll, in der Nähe der Singularität ein feineres Gitter zu wählen, um die Anzahl der nötigen Funktionsauswertungen zu minimieren. In der Praxis wird das Gitter zumeist adaptiv verfeinert. Diese Verfeinerung erfolgt anhand von Fehlerindikatoren, die den lokalen Integrationsfehler schätzen.

Wir beschreiben nun eine von mehreren Möglichkeiten, einen solchen Algorithmus zu implementieren. Dazu betrachten wir ein einzelnes Teilintervall $[a,b]$ des Gitters und bezeichnen mit $I[f]$ das zu berechnende Integral über diesem Intervall. Dieses Integral kann durch die zusammengesetzte Simpson-Formel $S_2[f]$ approximiert werden, die fünf Funktionswerte an äquidistanten Gitterpunkten des Intervalls $[a,b]$ verwendet.

Um die Genauigkeit dieser Näherung zu schätzen, verwenden wir die Idee des Romberg-Verfahrens. Die Simpson-Näherung $S_2[f]$ entspricht der Approximation $T_{(2,3)}[f]$ im Romberg-Tableau, vgl. Aufgabe 5. Daher kann *ohne zusätzliche Funktionsauswertungen* auch ein Kontrollverfahren, nämlich die Romberg-Näherung $T_{(1,3)}[f]$ berechnet werden, die in der Regel eine deutlich höhere

```
function I = adaptive_quadratur (a, b, f0, f2, f4, ε)
  % berechnet das Integral von f über [a, b]. Eingabe sind die Funktionswerte an den
  % beiden Randpunkten (f0 und f4) und der Funktionswert f2 in der Integralmitte
  % ε ist die vorgegebene absolute Fehlertoleranz
  h = (b − a)/2
  m = (a + b)/2        % Intervallmitte
  S1 = (f0 + 4f2 + f4)h/3        % Simpson-Formel S1[f] = T(1,2)[f]
  f1 = f(a + h/2)
  f3 = f(b − h/2)
  S2 = (f0 + 4f1 + 2f2 + 4f3 + f4)h/6        % entspricht T(2,3)[f]
  T(1,3) = (16S2 − S1)/15
  δ = | S2 − T(1,3) |        % Fehlerschätzer
  if δ < ε then
    I = T(1,3)
  else      % rekursiver Aufruf
    I1 = adaptive_quadratur(a, m, f0, f1, f2, ε)
    I2 = adaptive_quadratur(m, b, f2, f3, f4, ε)
    I = I1 + I2
  end if
end      % adaptive_quadratur
```

Algorithmus 42.1: Adaptive Simpson-Quadratur

Genauigkeit aufweist. Dies eröffnet die Möglichkeit, den Quadraturfehler der Simpson-Formel durch folgende Heuristik zu schätzen:

$$\big|\, I[f] - S_2[f] \,\big| \;\approx\; \big|\, T_{(1,3)}[f] - S_2[f] \,\big| \;=:\; \delta\,.$$

Liegt der Fehlerschätzer δ unter einem vom Benutzer vorzugebenden Schwellwert ϵ, so wird $S_2[f]$ als Integralnäherung akzeptiert. Wahlweise kann $S_2[f]$ durch $T_{(1,3)}[f]$ ersetzt werden. Ist der Wert des Fehlerschätzers hingegen zu groß, so wird das Gitter über $[a, b]$ verfeinert und die Näherung $S_4[f]$ berechnet. Dabei können alle bereits berechneten Funktionswerte weiter verwendet werden. Da $S_4[f]$ der Summe der S_2-Näherungen für die beiden Integrale über den neuen Gitterintervallen $[a, (a + b)/2]$ und $[(a + b)/2, b]$ entspricht, ist auf diese Weise die Grundlage für ein rekursives Fortschreiten geschaffen worden.

Ein Implementierungsvorschlag für dieses Verfahren ist in Algorithmus 42.1 enthalten. Für eine robuste Implementierung muß allerdings die `if`-Bedingung in Algorithmus 42.1 sorgfältiger formuliert werden, damit eine endliche Terminierung des Algorithmus garantiert ist (man vergleiche die Diskussion bei Gander und Gautschi [30]). Dies kann etwa dadurch erreicht werden, daß die

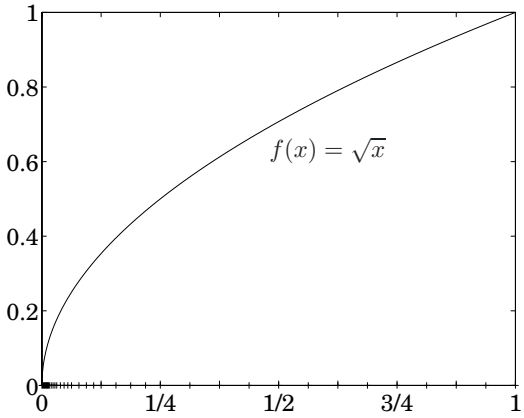


Abb. 42.1: Adaptives Gitter für die Integration von $\sqrt{x}$.

Funktion verlassen wird, sobald eine gewisse Rekursionstiefe erreicht ist oder wenn zwischen a und b keine weitere Maschinenzahl liegt.

Beispiel. Um die Effizienz dieses Verfahrens zu demonstrieren, betrachten wir das Integral

$$\int_0^1 \sqrt{x}\, dx \;=\; 2/3\,,$$

für das die Fehlerabschätzungen der vorangegangenen Abschnitte allesamt nicht anwendbar sind, da die Ableitungen des Integranden unbeschränkt sind. Bei Vorgabe einer Genauigkeit $\epsilon = 10^{-6}$ liefert Algorithmus 42.1 das Ergebnis $I = 0.666660323...$ mit einem relativen Fehler $9.515 \cdot 10^{-6}$. Hierfür werden die 37 in Abbildung 42.1 dargestellten Gitterpunkte ausgewählt. Offensichtlich erkennt der Algorithmus die Notwendigkeit, in der Nähe der Wurzelsingularität das Gitter zu verfeinern.

Für $\epsilon = 10^{-8}$ wählt der Algorithmus ein Gitter mit 97 Gitterpunkten. Der resultierende relative Fehler $5.3716 \cdot 10^{-8}$ ist erwartungsgemäß zwei Fehlerpotenzen kleiner als zuvor. ◊

Um die gleiche Genauigkeit mit weniger Funktionsauswertungen zu erzielen, muß die Simpson-Formel in Algorithmus 42.1 durch eine Gauß-Formel ersetzt werden. Für den Fehlerschätzer werden dann sogenannte *Gauß-Kronrod-Formeln* verwendet. Details können in [30] nachgelesen werden.

Aufgaben

1. Sei $f : \mathcal{I} \to \mathbb{R}$ konvex und differenzierbar. Zeigen Sie, daß die Trapezsumme eine obere Schranke und das zusammengesetzte Mittelpunktverfahren eine untere Schranke für $\int_{\mathcal{I}} f(x)\,dx$ liefert.

2. Das zusammengesetzte Mittelpunktverfahren für das Integral $I[f] = \int_a^b f(t)\,dt$ lautet

$$M_n[f] \;=\; \frac{b-a}{n} \sum_{i=1}^{n} f\big(a + \frac{2i-1}{2n}\,(b-a)\big)\,.$$

Beweisen Sie für $f \in C^2[a,b]$ die Fehlerabschätzung

$$\big|\, I[f] - M_n[f] \,\big| \;\le\; \frac{b-a}{24}\,\|f''\|_{\,[a,b]}\,h^2\,.$$

Vergleichen Sie das Resultat mit dem Ergebnis von Satz 36.1 für die Trapezsumme.

3. Beweisen Sie die folgende Fehlerabschätzung für die Trapezsumme: Gehört f'' zu $\mathcal{L}^2(a,b)$, dann gilt

$$\big|\, I[f] - T_n[f] \,\big| \;\le\; \frac{\sqrt{b-a}}{\sqrt{120}}\,\|f''\|_{\,\mathcal{L}^2(a,b)}\,h^2\,.$$

4. Gegeben seien ein Gitter $\triangle = \{x_0, \ldots, x_m\}$ und $2m+2$ Werte $y_0, \ldots, y_m \in \mathbb{R}$ und $y_0', \ldots, y_m' \in \mathbb{R}$. Zeigen Sie, daß es genau ein Polynom $p_{2m+1} \in \mathit{\Pi}_{2m+1}$ gibt mit der Eigenschaft

$$p_{2m+1}(x_j) \;=\; y_j\,, \quad p'_{2m+1}(x_j) = y_j'\,, \qquad j = 0, \ldots, m\,,$$

Leiten Sie für dieses *Hermite-Interpolationspolynom* die Darstellung

$$p_{2m+1} \;=\; \sum_{i=0}^{m} (y_i L_{0i} + y_i' L_{1i})$$

her, wobei die Polynome L_{0i} und L_{1i} mittels der Lagrange-Grundpolynome durch

$$L_{0i}(x) \;=\; \big(1 \,-\, 2l_i'(x_i)(x-x_i)\big) l_i^2(x)\,, \qquad L_{1i}(x) \;=\; (x-x_i) l_i^2(x)\,,$$

definiert sind.

5. Zeigen Sie, daß zwischen dem zusammengesetzten Simpson-Verfahren und der Trapezsumme der folgende Zusammenhang besteht:

$$S_n[f] \;=\; T_{2n}[f] - \big(\,T_n[f] - T_{2n}[f]\,\big)/3\,.$$

6. Geben Sie eine Funktion an, für die die Fehlerabschätzung aus Satz 38.4 mit dem tatsächlichen Fehler des zusammengesetzten Simpson-Verfahrens übereinstimmt.

7. (a) Weisen Sie nach, daß die Quadraturformel $T_{(1,2)}[\,\cdot\,]$ des Romberg-Verfahrens mit den Schrittweiten $h_1 = b-a$ und $h_2 = (b-a)/3$ die Newton-Cotes-Formel für $n = 3$ liefert (die sogenannte *Keplersche Faßregel*).

(b) Zeigen Sie, daß die Quadraturformel $T_{(1,3)}[\,\cdot\,]$ des Romberg-Verfahrens bei Verwendung der klassischen Rombergfolge gerade die Newton-Cotes-Formel für $n = 4$ ist (die *Milne-Formel*).

8. Sei $f \in C^{2m+2}[a,b]$, $m \geq 0$. Zeigen Sie, daß für das Romberg-Verfahren bei Verwendung der klassischen Rombergfolge $n_i = 2^{i-1}$, $i = 1, 2, \ldots$, die folgende Fehlerabschätzung gilt:

$$\Big| \int_a^b f(x)\,dx - T_{(j,j+m)}[f] \Big| \;\leq\; C_m \, \|f^{(2m+2)}\|_{[a,b]} \Big(\frac{b-a}{2^{j-1}} \Big)^{2m+2}, \quad j = 1, 2, \ldots$$

Hinweis: Verwenden Sie die Fehlerabschätzung aus Satz 39.2.

9. Bestimmen Sie die Knoten und die Gewichte der ersten zwei Gauß-Formeln zur näherungsweisen Berechnung von

$$I[f; 1-x] \;=\; \int_0^1 (1-x) f(x)\,dx\,.$$

10. Gegeben sei das uneigentliche Integral

$$I[f; e^{-x^2}] \;=\; \int_{-\infty}^{\infty} f(x) e^{-x^2}\,dx\,,$$

das durch eine m-punktige Gauß-Formel approximiert werden soll. Die Orthogonalpolynome zur Gewichtsfunktion $w(x) = e^{-x^2}$ über $\mathbb{R}$ sind die Hermite-Polynome (vgl. Aufgabe VI.10), daher heißt die entsprechende Quadraturformel *Hermite-Gauß-Formel*. Bestimmen Sie die ein- und die zweipunktige Hermite-Gauß-Formel.

11. (a) Beweisen Sie die lineare Konvergenz der Gauß-Legendre-Formeln für das Integral aus Beispiel 38.5 (vgl. Abbildung 41.1). Betrachten Sie dazu die Quotienten aus den Fehlerabschätzungen von Satz 41.4 für zwei aufeinanderfolgende Quadraturformeln. Wie groß ist der Konvergenzfaktor?
(b) Was läßt sich über die Konvergenz sagen, wenn die Gauß-Legendre-Formel auf das Integral

$$\int_0^1 \frac{1}{ax+1}\,dx\,, \qquad a > 0\,,$$

angewendet wird?
(c) Implementieren Sie die Gauß-Tschebyscheff-Formel für dieses Beispiel (mit $a = 1$) und vergleichen Sie ihre Ergebnisse mit Abbildung 41.1. Beachten Sie die Integrationsgrenzen.

12. Sei J_m die Jacobi-Matrix (41.3) der Legendre-Polynome. Zeigen Sie, daß der Wert der Gauß-Legendre-Formel für das Integral $\int_{-1}^1 1/(x-a)\,dx$ durch

$$G_m[\,\frac{1}{x-a}\,] \;=\; 2e_1^T (J_m - aI)^{-1} e_1$$

gegeben ist.

13. Leiten Sie die zweite Radau-Legendre-Formel über dem Intervall $[0,1]$ her:

$$R_2[f] \,=\, 3/4\, f(1/3) \,+\, 1/4\, f(1) \,\approx\, \int_0^1 f(x)\, dx\,.$$

14. (a) Zeigen Sie, daß die Gewichte $\tilde{w}_i$, $i = 1, \ldots, m$, der Radau-Legendre-Formel (41.6) über

$$\tilde{w}_i = \Lambda_m(\tilde{x}_i)\,, \qquad i = 1, \ldots, m\,,$$

mit den Christoffel-Funktionen aus Definition 33.3 zusammenhängen.
(b) Beweisen Sie, daß das Gewicht $\tilde{w}_m$ der m-ten Radau-Legendre-Formel durch

$$\tilde{w}_m = \frac{2}{m^2}$$

gegeben ist.

15. (a) Beweisen Sie, daß das Knotenpolynom der m-stufigen Radau-Legendre-Formel bezüglich des $\mathcal{L}^2$-Innenprodukts über $(-1,1)$ senkrecht auf Π_{m-2} steht.
(b) Zeigen Sie, daß die Stützstellen $\tilde{x}_1, \ldots, \tilde{x}_{m-1}$ dieser Formel die Nullstellen des $(m-1)$-ten Orthogonalpolynoms zur Gewichtsfunktion $w(x) = 1 - x$ über $(-1,1)$ sind.

16. Bei der m-stufigen Quadraturformel

$$L_m[f; w] \,=\, w_1 f(-1) \,+\, \sum_{i=2}^{m-1} w_i f(x_i) \,+\, w_m f(1) \,\approx\, \int_{-1}^1 f(x) w(x)\, dx$$

seien die Knoten $x_1 = -1$ und $x_m = 1$ fest vorgeschrieben. Die inneren Knoten $x_2, \ldots, x_{m-1}$ und die Gewichte $w_1, \ldots, w_m$ sollen so bestimmt werden, daß die Quadraturformel größtmöglichen Exaktheitsgrad besitzt. Die so erhaltenen Quadraturformeln heißen *Lobatto-Formeln*.
(a) Weisen Sie nach, daß der Exaktheitsgrad der m-stufigen Lobatto-Formel maximal $2m-3$ ist.
(b) Zeigen Sie, daß der Exaktheitsgrad $2m-3$ genau dann erreicht wird, wenn die beiden folgenden Bedingungen erfüllt sind:

(i) Der Exaktheitsgrad ist mindestens $m-1$;

(ii) $\int\limits_{-1}^{1} p(x)\omega(x)(1-x^2)w(x)\, dx \,=\, 0$ für alle $p \in \Pi_{m-3}$.

Dabei bezeichnet ω das Knotenpolynom zu den Knoten $x_2, \ldots, x_{m-1}$.
(c) Berechnen Sie die Knoten und die Gewichte von $L_3[\,\cdot\,; 1]$.

17. Zur Approximation des Integrals

$$\int_{\triangle} f(x,y)\, dy\, dx$$

einer Funktion f über dem Dreieck $\triangle = \{\,(x,y) \,:\, 0 \le x, y \le 1, x+y \le 1\,\}$ sollen Quadraturformeln unter Verwendung von 1- bzw. 2-punktigen Gauß-Quadraturformeln konstruiert werden. Verwenden Sie dazu den Satz von Fubini und approximieren Sie anschließend zunächst das innere und dann das äußere Integral durch geeignete Gauß-Formeln. Verwenden Sie Aufgabe 9.

18. Implementieren Sie Algorithmus 42.1 und berechnen Sie damit das uneigentliche Integral

$$\int_0^1 \frac{1}{\sqrt{x}}\, dx \;=\; 2\,.$$

Setzen Sie hierfür die Funktion $f(x) = 1/\sqrt{x}$, $x > 0$, durch $f(0) = 0$ fort. Modifizieren Sie den Algorithmus so, daß auch das verwendete Gitter ausgegeben wird. Überzeugen Sie sich von der Korrektheit des Algorithmus, indem Sie verschiedene Werte von ϵ vorgeben.

VIII Splines

Historisch wurde zunächst die Polynominterpolation zur Approximation skalarer Funktionen verwendet. Die interpolierenden Polynome weisen jedoch in der Regel bei feineren Gittern starke Oszillationen auf und nur eine geringe qualitative Übereinstimmungen mit der gesuchten Funktion. Daher ist diese Art der Interpolation lediglich für sehr kleine Polynomgrade beziehungsweise spezielle Interpolationsgitter sinnvoll.

Statt dessen bietet es sich an (wie beim Übergang von Quadraturformeln zu zusammengesetzten Quadraturverfahren), die gesuchte Funktion durch stückweise zusammengesetzte Polynome niedrigen Grades zu interpolieren. Dies führt auf die sogenannten *Splines*, denen wir uns in diesem Kapitel zuwenden.

Die hier gewählte Darstellung konzentriert sich auf Fehlerabschätzungen in der $\mathcal{L}^2$-Norm. Andere Fehlerabschätzungen und viele weitergehende Fragen werden in dem Buch von Schumaker [95] behandelt. Schließlich seien noch die Lehrbücher von Hämmerlin und Hoffmann [46] und von Kreß [63] genannt, in denen ein alternativer Zugang über B-Splines im Vordergrund steht.

43 Treppenfunktionen

Gegeben sei ein reelles Intervall $[a, b]$ und ein Gitter

$$\triangle \;=\; \{a = x_0 < x_1 < \ldots < x_l = b\} \tag{43.1}$$

aus $l + 1$ streng monoton wachsend angeordneten Knoten. Wie zuvor seien h_i die Längen der einzelnen Gitterintervalle und $h = \max_{i=1,\ldots,l} h_i$ die Gitterweite von $\triangle$. Unter einer *Treppenfunktion* s verstehen wir eine rechtsseitig stetige Funktion, die in jedem der halboffenen Teilintervalle $[x_{i-1}, x_i)$ konstant ist,

$$s(x) \;=\; s_i\,, \quad x_{i-1} \le x < x_i\,, \qquad i = 1, \ldots, l\,.$$

Die Treppenfunktionen über $\triangle$ bilden offensichtlich einen linearen Raum $S_{0,\triangle}$ der Dimension l. Als Basisfunktionen bieten sich die charakteristischen Funktionen χ_i der l Teilintervalle an (die Funktion χ_i hat somit den Wert Eins

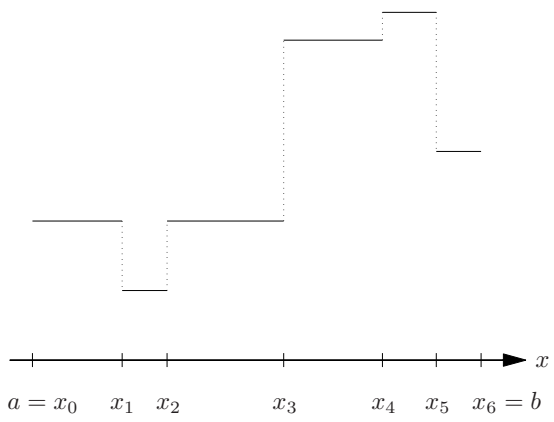


Abb. 43.1: Treppenfunktion

im Intervall $[x_{i-1}, x_i)$ und sonst den Wert Null), denn mit ihnen ergibt sich unmittelbar die Basisdarstellung

$$s \;=\; \sum_{i=1}^{l} s_i \chi_i \,.$$

$S_{0,\triangle}$ ist ein Teilraum von $\mathcal{L}^2(a,b)$, die Basisfunktionen $\chi_i/\sqrt{h_i}$, $i = 1, \ldots, l$, bilden eine zugehörige Orthonormalbasis von $S_{0,\triangle}$. Daher können wir Satz 31.6 anwenden, um die Bestapproximation aus $S_{0,\triangle}$ an eine vorgegebene Funktion $f \in \mathcal{L}^2(a,b)$ zu bestimmen:

Satz 43.1. *Sei $f \in \mathcal{L}^2(a,b)$. Dann ist $s = \sum_{i=1}^{l} s_i \chi_i$ mit*

$$s_i \;=\; \frac{1}{h_i} \int_{x_{i-1}}^{x_i} f(x)\, dx \,, \qquad i = 1, \ldots, l \,, \tag{43.2}$$

bezüglich der $\mathcal{L}^2$-Norm die Bestapproximation aus $S_{0,\triangle}$ an f. Ist $f \in H^1(a,b)$, dann gilt

$$\|f - s\|_{\mathcal{L}^2(a,b)} \;\le\; h\, \|f'\|_{\mathcal{L}^2(a,b)} \,.$$

Beweis. Nach Satz 31.6 (c) hat die Bestapproximation die Darstellung

$$s \;=\; \sum_{i=1}^{l} \Big\langle \frac{\chi_i}{\sqrt{h_i}} \,,\, f \Big\rangle_{\mathcal{L}^2(a,b)} \frac{\chi_i}{\sqrt{h_i}} \;=\; \sum_{i=1}^{l} \frac{1}{h_i} \Big(\int_{x_{i-1}}^{x_i} f(x)\, dx \Big) \chi_i \;=\; \sum_{i=1}^{l} s_i \chi_i \,.$$

Ferner gilt die Fehlerdarstellung

$$\|f-s\|^2_{\mathcal{L}^2(a,b)} \;=\; \int_a^b \big|f(x)-s(x)\big|^2\,dx \;=\; \sum_{i=1}^{l}\int_{x_{i-1}}^{x_i} \big|f(x)-s_i\big|^2\,dx\,. \tag{43.3}$$

Ist $f \in H^1(a,b)$, dann existiert nach dem Mittelwertsatz der Integralrechnung (Funktionen aus $H^1(a,b)$ sind nach Beispiel 31.4 stetig) in jedem Teilintervall $[x_{i-1},x_i)$ eine Zwischenstelle ξ_i mit $f(\xi_i) = s_i$, und für $x \in [x_{i-1},x_i)$ folgt aus der Cauchy-Schwarz-Ungleichung

$$\begin{aligned}\big|f(x)-s_i\big|^2 \;&=\; \big|f(x)-f(\xi_i)\big|^2 \;=\; \Big|\int_{\xi_i}^{x} f'(t)\,dt\Big|^2 \\ &\le\; \Big(\int_{x_{i-1}}^{x_i} |f'(t)|\,dt\Big)^2 \le \int_{x_{i-1}}^{x_i} |f'(t)|^2\,dt \int_{x_{i-1}}^{x_i} dt \;=\; h_i \int_{x_{i-1}}^{x_i} |f'(t)|^2\,dt\,.\end{aligned}$$

Eingesetzt in (43.3) folgt somit die behauptete Fehlerabschätzung

$$\begin{aligned}\|f-s\|^2_{\mathcal{L}^2(a,b)} \;&\le\; \sum_{i=1}^{l} h_i \int_{x_{i-1}}^{x_i}\int_{x_{i-1}}^{x_i} |f'(t)|^2\,dt\,dx \\ &=\; \sum_{i=1}^{l} h_i^2 \int_{x_{i-1}}^{x_i} |f'(t)|^2\,dt \;\le\; \max_{i=1,\dots,l} h_i^2 \int_a^b |f'(t)|^2\,dt\,.\end{aligned}$$

□

Ersetzt man in der Darstellung (43.2) das Integral durch die zusammengesetzte Mittelpunktformel, dann ergibt sich die Näherung

$$s \;\approx\; s_I \;=\; \sum_{i=1}^{l} f\big(\,\frac{x_{i-1}+x_i}{2}\,\big)\,\chi_i\,. \tag{43.4}$$

s_I ist die eindeutig bestimmte Treppenfunktion, die die Funktion f in den Gitterintervallmittelpunkten interpoliert.

Bemerkung. Die Fehlerabschätzung aus Satz 43.1 bleibt auch für die interpolierende Treppenfunktion s_I aus (43.4) richtig; im Beweis muß lediglich die Zwischenstelle ξ_i durch $(x_{i-1}+x_i)/2$ ersetzt werden. ◊

44 Lineare Splines

Man macht sich leicht durch Taylorentwicklung klar, daß die Größenordnung $O(h)$ des Fehlers bei der Approximation durch Treppenfunktionen bestmöglich

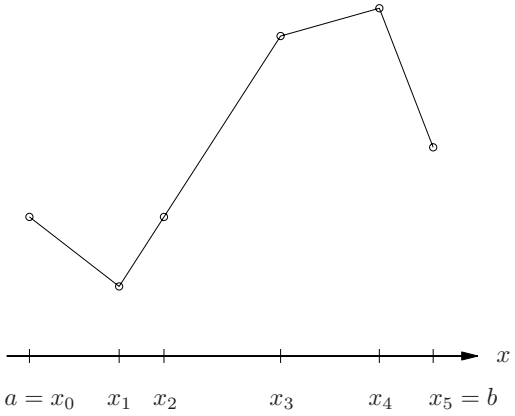


Abb. 44.1: Linearer Spline

ist. Um höhere h-Potenzen zu erhalten, muß man (wie bei der Quadratur) von stückweise konstanten Funktionen zu stückweise definierten Polynomen höheren Grades übergehen.

Definition 44.1. Sei $\triangle$ wie in (43.1) ein Gitter aus $l+1$ Knoten x_i, $i = 0, \ldots, l$. Ein *Spline* vom Grad $n \in \mathbb{N}$ ist eine Funktion $s \in C^{n-1}[a,b]$, die auf jedem Intervall $[x_{i-1}, x_i)$, $i = 1, \ldots, l$, mit einem Polynom $p_i \in \Pi_n$ übereinstimmt. Für den Raum der Splines vom Grad n bezüglich $\triangle$ schreiben wir $S_{n,\triangle}$. Treppenfunktionen bezeichnen wir auch als Splines vom Grad $n = 0$.

Neben den Treppenfunktionen sind die stückweise linearen Funktionen, die sogenannten *linearen Splines* (vgl. Abbildung 44.1) für $n = 1$ und die *kubischen Splines* ($n = 3$, siehe Abschnitt 46) die wichtigsten Spezialfälle.

Der Raum $S_{n,\triangle}$ ist für jedes $n \in \mathbb{N}$ ein linearer Vektorraum mit

$$\Pi_n \subset S_{n,\triangle}\,. \tag{44.1}$$

Ist $s \in S_{n,\triangle}$ dann gehört $s^{(k)}$ zu $S_{n-k,\triangle}$ für $0 \le k < n$. Ferner existiert im Innern eines jeden Gitterintervalls die n-te Ableitung $s^{(n)}$ und ist dort jeweils konstant. Lediglich in den Gitterpunkten von $\triangle$ ist $s^{(n)}$ nicht definiert. Durch geeignete Fortsetzung in die Gitterpunkte kann $s^{(n)}$ jedoch mit einer entsprechenden Treppenfunktion in $S_{0,\triangle}$ identifiziert werden. Umgekehrt gehört die n-te Stammfunktion einer beliebigen Treppenfunktion zu $S_{n,\triangle}$, da sie $n-1$ mal stetig differenzierbar ist und in jedem Gitterintervall mit einem Polynom n-ten Grades übereinstimmt.

Proposition 44.2. *$S_{n,\triangle}$ ist ein $(n+l)$-dimensionaler Unterraum von $\mathcal{L}^2(a,b)$.*

Beweis. Für jedes $i = 1, \ldots, l$ bezeichne Ξ_i eine beliebige n-te Stammfunktion der charakteristischen Funktion χ_i des Gitterintervalls $[x_{i-1}, x_i)$. Die Funktionen Ξ_i, $i = 1, \ldots, l$, gehören allesamt zu $S_{n,\triangle}$, genau wie die Monome x^j, $j = 0, 1, \ldots, n-1$, vgl. (44.1). Wir beweisen nun, daß die Menge

$$\{\Xi_i : i = 1, \ldots, l\} \cup \{x^j : j = 0, \ldots, n-1\} \tag{44.2}$$

eine Basis von $S_{n,\triangle}$ bildet.

Ist $s \in S_{n,\triangle}$ beliebig gewählt, so gehört $s^{(n)}$ zu $S_{0,\triangle}$ und es gilt

$$s^{(n)} = \sum_{i=1}^{l} s_i \chi_i$$

für gewisse $s_1, \ldots, s_l \in \mathbb{R}$. Folglich stimmen s und $\sum_{i=1}^{l} s_i \Xi_i$ bis auf ein Polynom vom Grad $n-1$ überein. Dies beweist, daß jeder Spline $s \in S_{n,\triangle}$ als Linearkombination des Funktionensystems (44.2) darstellbar ist.

Zum Nachweis der linearen Unabhängigkeit nehmen wir an, es gäbe ein Polynom $p \in \Pi_{n-1}$ und Zahlen $s_1, \ldots, s_l$ mit

$$s = p + \sum_{i=1}^{l} s_i \Xi_i = 0 \qquad \text{in } [a, b]\,.$$

Hieraus folgt unmittelbar

$$s^{(n)} = \sum_{i=1}^{l} s_i \chi_i = 0\,,$$

und da die χ_i, $i = 1, \ldots, l$, eine Basis von $S_{0,\triangle}$ bilden, ergibt sich $s_1 = \ldots = s_l = 0$. Folglich ist $s = p = 0$, d. h. die Elemente von (44.2) sind linear unabhängig und bilden eine Basis von $S_{n,\triangle}$. □

Für *lineare Splines* über $\triangle$ ergeben sich somit genau $l+1$ Freiheitsgrade. Anstelle der im Beweis konstruierten Basis (44.2) verwendet man in der Numerik üblicherweise die sogenannte *nodale Basis* der *Hutfunktionen* Λ_i, $i = 0, \ldots, l$, aus Abbildung 44.2.

Bemerkung. In Beispiel 31.4 haben wir gesehen, daß die Hutfunktionen Λ_i zu $H^1(a, b)$ gehören und haben ihre schwache Ableitung ausgerechnet. Somit ist $S_{1,\triangle}$ ein linearer Unterraum von $H^1(a, b)$. Die schwache Ableitung s' eines linearen Splines $s \in S_{1,\triangle}$ ist mit der im Beweis von Proposition 44.2 berechneten Treppenfunktion identisch. ◊

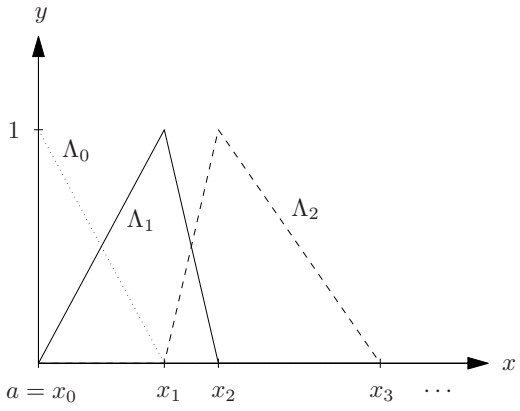


Abb. 44.2: Hutfunktionen Λ_0, Λ_1 und Λ_2

Ähnlich dem Lagrange-Grundpolynom nimmt die Hutfunktion Λ_i an den Knoten x_j des Gitters die Werte

$$\Lambda_i(x_j) = \delta_{ij}\,, \qquad i,j = 0,\ldots,l\,, \tag{44.3}$$

an. Damit können wir das folgende Interpolationsresultat formulieren:

Satz 44.3. *Seien $\triangle$ aus* (43.1) *ein Gitter über $[a,b]$ und $y_0,\ldots,y_l$ vorgegebene Daten. Dann ist $s = \sum_{i=0}^{l} y_i\Lambda_i \in S_{1,\triangle}$ der eindeutig bestimmte lineare Spline mit*

$$s(x_i) = y_i\,, \qquad i = 0,\ldots,l\,.$$

Beweis. Wegen (44.3) erfüllt s offensichtlich die Interpolationsbedingung. Die Eindeutigkeit folgt aus der Tatsache, daß die lineare Teilfunktion im Gitterintervall $[x_{i-1},x_i)$ durch die beiden Wertepaare (x_{i-1},y_{i-1}) und (x_i,y_i) eindeutig festgelegt ist. □

45 Fehlerabschätzungen für lineare Splines

Bevor wir den Approximationsfehler des interpolierenden linearen Splines abschätzen, beweisen wir den folgenden Hilfssatz über den Fehler einer beliebigen interpolierenden Funktion.

Lemma 45.1. *Zu gegebenem $f \in H^1(a,b)$ sei $\varphi \in H^1(a,b)$ eine Funktion, die f über dem Gitter $\triangle \subset [a,b]$ aus* (43.1) *mit Gitterweite h interpoliert. Dann*

gilt

$$\|f-\varphi\|_{\mathcal{L}^2(a,b)} \le \frac{h}{\sqrt{2}}\,\|f'-\varphi'\|_{\mathcal{L}^2(a,b)}\,.$$

Beweis. Sei $x \in [x_{i-1}, x_i)$ für ein $i \in \{1,\dots,l\}$. Aufgrund der Interpolationsbedingung gilt

$$f(x)-\varphi(x) = \int_{x_{i-1}}^{x} \big(f'(t)-\varphi'(t)\big)\,dt$$

und aus der Cauchy-Schwarz-Ungleichung folgt

$$\begin{aligned}
\int_{x_{i-1}}^{x_i} |f(x)-\varphi(x)|^2\,dx &= \int_{x_{i-1}}^{x_i} \Big|\int_{x_{i-1}}^{x} \big(f'(t)-\varphi'(t)\big)\,dt\Big|^2\,dx \\
&\le \int_{x_{i-1}}^{x_i} \Big(\int_{x_{i-1}}^{x} |f'(t)-\varphi'(t)|^2\,dt \int_{x_{i-1}}^{x} dt\Big)\,dx \\
&= \int_{x_{i-1}}^{x_i} (x-x_{i-1}) \int_{x_{i-1}}^{x} |f'(t)-\varphi'(t)|^2\,dt\,dx \\
&\le \int_{x_{i-1}}^{x_i} |f'(t)-\varphi'(t)|^2\,dt \int_{x_{i-1}}^{x_i} (x-x_{i-1})\,dx \\
&= \frac{1}{2}\,h_i^2 \int_{x_{i-1}}^{x_i} |f'(t)-\varphi'(t)|^2\,dt\,.
\end{aligned}$$

Summation von $i=1,\dots,l$ liefert daher die gewünschte Ungleichung. □

Bemerkung. Die Konstante $1/\sqrt{2}$ in der Abschätzung aus Lemma 45.1 ist nicht bestmöglich. Die optimale Konstante lautet $1/\pi$ (vgl. Aufgabe IX.8). ◊

Mit Hilfe von Lemma 45.1 können wir nun den Fehler des interpolierenden linearen Splines abschätzen:

Satz 45.2. *Sei $f \in H^1(a,b)$ und s der interpolierende lineare Spline zu f über einem Gitter $\triangle \subset [a,b]$ mit Gitterweite h. Dann ist*

$$\|f-s\|_{\mathcal{L}^2(a,b)} \le \frac{h}{\sqrt{2}}\,\|f'\|_{\mathcal{L}^2(a,b)}\,.$$

Beweis. Im Hinblick auf Lemma 45.1 (der lineare Spline s gehört zu $H^1(a,b)$) ist es erforderlich, den $\mathcal{L}^2$-Abstand zwischen f' und s' abzuschätzen. Dazu beachten wir, daß

$$\begin{aligned}
\|f'-s'\|^2_{\mathcal{L}^2(a,b)} &= \|f'\|^2_{\mathcal{L}^2(a,b)} - \langle\, 2f'-s', s'\,\rangle_{\mathcal{L}^2(a,b)} \\
&= \|f'\|^2_{\mathcal{L}^2(a,b)} - \int_a^b (2f'-s')(x)s'(x)\,dx\,.
\end{aligned} \tag{45.1}$$

Wir zeigen nun, daß das hintere Integral positiv ist. Da s' stückweise konstant ist mit

$$s'(x) \;=\; \frac{f(x_i) - f(x_{i-1})}{x_i - x_{i-1}}\,, \qquad x \in (x_{i-1}, x_i)\,,$$

folgt

$$\begin{aligned}\int_a^b (2f' - s')(x)s'(x)\,dx \;&=\; \sum_{i=1}^{l} \frac{f(x_i) - f(x_{i-1})}{x_i - x_{i-1}} \int_{x_{i-1}}^{x_i} (2f' - s')(x)\,dx \\ &= \sum_{i=1}^{l} \frac{f(x_i) - f(x_{i-1})}{x_i - x_{i-1}} \big(2f(x_i) - s(x_i) - 2f(x_{i-1}) + s(x_{i-1})\big)\,.\end{aligned}$$

Wegen der Interpolationseigenschaft ist dies äquivalent zu

$$\begin{aligned}\int_a^b (2f' - s')(x)s'(x)\,dx \;&=\; \sum_{i=1}^{l} \frac{f(x_i) - f(x_{i-1})}{x_i - x_{i-1}} \big(f(x_i) - f(x_{i-1})\big) \\ &=\; \sum_{i=1}^{l} \frac{\big(f(x_i) - f(x_{i-1})\big)^2}{x_i - x_{i-1}} \;\geq\; 0\,.\end{aligned}$$

In (45.1) eingesetzt, folgt somit

$$\|f' - s'\|^2_{\mathcal{L}^2(a,b)} \;\leq\; \|f'\|^2_{\mathcal{L}^2(a,b)}$$

und damit ergibt sich die Behauptung aus Lemma 45.1. □

Die Abschätzung aus Satz 45.2 stimmt bis auf einen konstanten Faktor mit derjenigen aus Satz 43.1 für Treppenfunktionen überein. Allerdings lassen sich unter zusätzlichen Glattheitsannahmen an f bessere Fehlerabschätzungen beweisen.

Definition 45.3. Mit $H^2(a,b)$ bezeichnen wir den Sobolevraum derjenigen Funktionen $f \in H^1(a,b)$, deren (schwache) Ableitung f' ebenfalls zu $H^1(a,b)$ gehört. Die Ableitung von f' bezeichnen wir mit f''.

Satz 45.4. *Sei $f \in H^2(a,b)$ und s der interpolierende lineare Spline zu einem Gitter $\triangle \subset [a,b]$. Dann gilt*

$$\|f - s\|_{\mathcal{L}^2(a,b)} \;\leq\; \frac{h^2}{2}\,\|f''\|_{\mathcal{L}^2(a,b)}\,, \tag{45.2}$$

$$\|f' - s'\|_{\mathcal{L}^2(a,b)} \;\leq\; \frac{h}{\sqrt{2}}\,\|f''\|_{\mathcal{L}^2(a,b)}\,. \tag{45.3}$$

Beweis. Durch partielle Integration (vgl. Aufgabe VI.4) auf jedem Teilintervall (x_{i-1}, x_i) ergibt sich aus der Interpolationseigenschaft

$$\begin{aligned}\|f' - s'\|^2_{\mathcal{L}^2(a,b)} &= \sum_{i=1}^{l} \Big((f-s)(x)(f'-s')(x)\Big|_{x_{i-1}}^{x_i} \\ &\qquad - \int_{x_{i-1}}^{x_i} (f-s)(x)(f-s)''(x)\,dx \Big) \\ &= -\int_a^b (f-s)(x) f''(x)\,dx\,.\end{aligned}$$

Aus der Cauchy-Schwarz-Ungleichung und Lemma 45.1 folgt somit

$$\begin{aligned}\|f' - s'\|^2_{\mathcal{L}^2(a,b)} &\le \|f-s\|_{\mathcal{L}^2(a,b)} \|f''\|_{\mathcal{L}^2(a,b)} \\ &\le \frac{h}{\sqrt{2}} \|f'-s'\|_{\mathcal{L}^2(a,b)} \|f''\|_{\mathcal{L}^2(a,b)}\,.\end{aligned}$$

Division durch $\|f' - s'\|_{\mathcal{L}^2(a,b)}$ ergibt somit die Behauptung (45.3). Damit folgt aber sofort auch (45.2) aus Lemma 45.1 und (45.3). □

Ein Vergleich mit Satz 43.1 ergibt, daß die Erhöhung des Splinegrads von $n = 0$ auf $n = 1$ eine entsprechende Erhöhung der h-Potenz in der Abschätzung für den Interpolationsfehler in der $\mathcal{L}^2$-Norm nach sich zieht (vorausgesetzt, die zu interpolierende Funktion f ist hinreichend glatt).

Die Fehlerabschätzung für den interpolierenden Spline ist natürlich gleichzeitig eine Fehlerabschätzung für die Bestapproximation aus $S_{1,\triangle}$ an f bezüglich der $\mathcal{L}^2$-Norm. Während wir in Satz 43.1 die bestapproximierende Treppenfunktion noch explizit angeben konnten, ist dies für lineare Splines jedoch nicht mehr möglich. Allerdings haben wir in Satz 31.10 gesehen, daß eine numerische Berechnung der Bestapproximation leicht implementiert werden kann, wenn die Gramsche Matrix $G = [\,\langle \Lambda_i, \Lambda_j \rangle_{\mathcal{L}^2(a,b)}\,]_{ij}$ der Hutfunktionen bekannt ist. Der Vollständigkeit halber berechnen wir daher nun noch die Gramsche Matrix der Hutfunktionen.

Beispiel 45.5. Seien wie zuvor $\triangle$ ein Gitter der Form (43.1) über $[a, b]$ und Λ_i, $i = 0, \ldots, l$, die zugehörigen Hutfunktionen. Wir bezeichnen mit

$$g_{ij} = \langle\, \Lambda_i, \Lambda_j \,\rangle_{\mathcal{L}^2(a,b)}$$

den (i, j)-Eintrag von G. Man beachte, daß i und j von 0 bis l laufen, G also eine $(l + 1) \times (l + 1)$-Matrix ist. Offensichtlich ist das Produkt $\Lambda_i(x)\Lambda_j(x)$ identisch Null, wenn $|i - j| \ge 2$ ist, d. h. G ist eine Tridiagonalmatrix.

Für $i = j \in \{1, \dots, l\}$ ergibt sich

$$\int_{x_{i-1}}^{x_i} \Lambda_i^2(x)\, dx \;=\; \int_0^{h_i} \frac{t^2}{h_i^2}\, dt \;=\; h_i/3\,.$$

Entsprechend ist

$$\int_{x_i}^{x_{i+1}} \Lambda_i^2(x)\, dx \;=\; h_{i+1}/3\,, \qquad i = 0, \dots, l-1\,,$$

und daher berechnet sich die Diagonale von G zu

$$g_{ii} \;=\; \begin{cases} h_1/3\,, & i = 0\,, \\ (h_i + h_{i+1})/3\,, & 1 \le i \le l-1\,, \\ h_l/3\,, & i = l\,. \end{cases}$$

Wegen der Symmetrie von G reicht es nun aus, noch die unteren Nebendiagonalelemente von G zu bestimmen:

$$\begin{aligned} g_{i,i-1} \;&=\; \int_{x_{i-1}}^{x_i} \Lambda_i(x)\Lambda_{i-1}(x)\, dx \;=\; \int_0^{h_i} \frac{t}{h_i}\,\frac{h_i - t}{h_i}\, dt \\ &=\; \frac{1}{h_i^2}\,(h_i^3/2 - h_i^3/3) \;=\; h_i/6\,, \end{aligned}$$

$i = 1, \dots, l$. Damit ergibt sich

$$G \;=\; \frac{1}{6} \begin{bmatrix} 2h_1 & h_1 & & & 0 \\ h_1 & 2(h_1+h_2) & h_2 & & \\ & h_2 & \ddots & \ddots & \\ & & \ddots & 2(h_{l-1}+h_l) & h_l \\ 0 & & & h_l & 2h_l \end{bmatrix} \tag{45.4}$$

als Gramsche Matrix der Hutfunktionen $\{\Lambda_0, \dots, \Lambda_l\}$. ◊

46 Kubische Splines

Kubische Splines werden vor allem in der graphischen Datenverarbeitung eingesetzt, denn $S_{3,\triangle}$ ist ein Unterraum von C^2 und C^2-Kurven werden vom menschlichen Auge als „glatt“ empfunden. Dies ist einer der Gründe, warum quadratische Splines nur von untergeordneter Bedeutung sind.

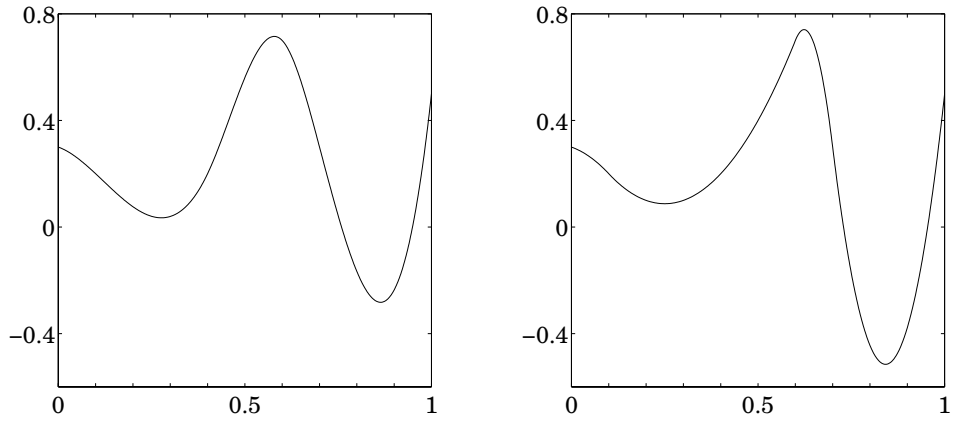


Abb. 46.1: Quadratischer und kubischer Spline (welcher ist welcher?)

Beispiel 46.1. Abbildung 46.1 enthält einen quadratischen und einen kubischen Spline, die die gleichen sechs Punkte (Randpunkte eingeschlossen) im Intervall $[0, 1]$ interpolieren und an den Randpunkten die gleichen Ableitungen haben. Welche Kurve ist glatter? Versuchen Sie mit dem Auge die „Unglattheitsstellen“ des quadratischen Splines zu finden (also die Sprünge in den zweiten Ableitungen). Abbildung 46.2 auf der nächsten Seite zeigt die Lösung: Die helle Kurve dort ist der quadratische Spline, die kleinen Kreise markieren die Interpolationspunkte, also die Unstetigkeitsstellen der zweiten Ableitung des quadratischen Splines. ◊

Nach Proposition 44.2 ist $S_{3,\triangle}$ ein $(l+3)$-dimensionaler Vektorraum, falls $\triangle$ wie zuvor ein Gitter mit $l+1$ Gitterpunkten $x_0, \ldots, x_l$ ist. Damit ist (formal) die Interpolationsaufgabe für kubische Splines unterbestimmt und man kann weitere Zusatzbedingungen an den interpolierenden kubischen Spline stellen.

Die Untersuchung von Existenz und Eindeutigkeit (bei entsprechenden Zusatzbedingungen) interpolierender kubischer Splines führen wir mit Hilfe der zweiten Ableitung s'' eines allgemeinen kubischen Splines $s \in S_{3,\triangle}$ durch. Wie wir in Abschnitt 44 gesehen haben, gehört s'' zu $S_{1,\triangle}$, das heißt

$$s'' = \sum_{i=0}^{l} \gamma_i \Lambda_i \qquad \text{mit} \qquad \gamma_i = s''(x_i)\,, \quad i = 0, \ldots, l\,. \tag{46.1}$$

Die Koeffizienten γ_i (die sogenannten *Momente* des kubischen Splines) spielen im folgenden eine wesentliche Rolle. Mit den Abkürzungen $s_i = s(x_i)$ und $s_i' = s_i'(x_i)$, $i = 0, \ldots, l$, für die Funktionswerte und die Werte der ersten

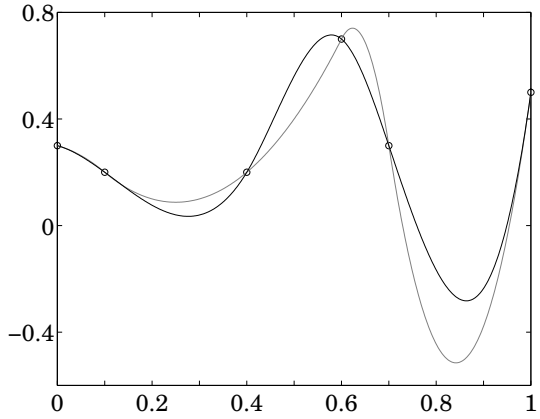


Abb. 46.2: Kubischer Spline (dunkel) und quadratischer Spline (hell)

Ableitung des Splines gilt nach dem Satz von Taylor

$$s(x) \;=\; s_i + s_i'(x - x_i) + \int_{x_i}^{x} (x-t)s''(t)\,dt\,, \qquad x \in [x_{i-1}, x_i]\,,$$

und da sich (46.1) in diesem Teilintervall zu $s'' = \gamma_{i-1}\Lambda_{i-1} + \gamma_i\Lambda_i$ vereinfacht, ergibt sich hieraus die Darstellung

$$\begin{gathered} s(x) \;=\; s_i + s_i'\,(x - x_i) + \gamma_i\,\frac{(x-x_i)^2}{2} + \frac{\gamma_i - \gamma_{i-1}}{h_i}\,\frac{(x-x_i)^3}{6} \\ \text{für} \quad x \in [x_{i-1}, x_i]\,, \quad i = 1, \ldots, l\,. \end{gathered} \tag{46.2}$$

Insbesondere haben wir daher am linken Randpunkt die Funktionswerte

$$s(x_{i-1}) \;=\; s_i - s_i' h_i + \gamma_{i-1}h_i^2/6 + \gamma_i h_i^2/3\,, \tag{46.3a}$$

$$s'(x_{i-1}) \;=\; s_i' - \gamma_{i-1}h_i/2 - \gamma_i h_i/2\,, \qquad i = 1, \ldots, l\,. \tag{46.3b}$$

Für $i = 2, \ldots, l$, ist $s'(x_{i-1}) = s_{i-1}'$ wegen der Stetigkeit von s' und daher folgt aus (46.3b), daß

$$s_i' - s_{i-1}' \;=\; \frac{h_i}{2}\,(\gamma_{i-1} + \gamma_i)\,, \qquad i = 2, \ldots, l\,. \tag{46.4}$$

Durch Kombination der Gleichungen (46.3a) mit Index i und $i+1$ ergibt sich weiterhin

$$\frac{s_{i+1} - s_i}{h_{i+1}} \;-\; \frac{s_i - s_{i-1}}{h_i} \;=\; s_{i+1}' - \gamma_i\frac{h_{i+1}}{6} - \gamma_{i+1}\frac{h_{i+1}}{3} - s_i' + \gamma_{i-1}\frac{h_i}{6} + \gamma_i\frac{h_i}{3}\,,$$

$i = 1, \dots, l-1$. Die hier auftretende Differenz der ersten Ableitungen kann mit Hilfe von (46.4) ersetzt werden. Damit erhalten wir schließlich die Gleichungen

$$\frac{s_{i+1} - s_i}{h_{i+1}} - \frac{s_i - s_{i-1}}{h_i} = \gamma_{i-1}\frac{h_i}{6} + \gamma_i\frac{h_i + h_{i+1}}{3} + \gamma_{i+1}\frac{h_{i+1}}{6}$$

für $i = 1, \dots, l-1$ zwischen den Funktionswerten und den Momenten eines beliebigen kubischen Splines $s \in S_{3,\triangle}$. In Matrixnotation lautet dieses System

$$\begin{aligned} \frac{1}{6}\begin{bmatrix} h_1 & 2(h_1+h_2) & h_2 & & & 0 \\ & h_2 & 2(h_2+h_3) & \ddots & & \\ & & \ddots & \ddots & h_{l-1} & \\ 0 & & & h_{l-1} & 2(h_{l-1}+h_l) & h_l \end{bmatrix} \begin{bmatrix} \gamma_0 \\ \gamma_1 \\ \vdots \\ \gamma_{l-1} \\ \gamma_l \end{bmatrix} \\ = -\begin{bmatrix} -h_1^{-1} & h_1^{-1}+h_2^{-1} & -h_2^{-1} & & & 0 \\ & -h_2^{-1} & h_2^{-1}+h_3^{-1} & \ddots & & \\ & & \ddots & \ddots & -h_{l-1}^{-1} & \\ 0 & & & -h_{l-1}^{-1} & h_{l-1}^{-1}+h_l^{-1} & -h_l^{-1} \end{bmatrix} \begin{bmatrix} s_0 \\ s_1 \\ \vdots \\ s_{l-1} \\ s_l \end{bmatrix}. \end{aligned} \tag{46.5}$$

Erfüllen umgekehrt die Momente γ_i und die Funktionswerte $s_i = s(x_i)$ die Gleichungen (46.5), dann definieren diese Werte genau einen interpolierenden kubischen Spline (46.2), dessen Ableitungen $s_i' = s'(x_i)$ durch (46.3a) gegeben sind, also

$$s_i' = \frac{s_i - s_{i-1}}{h_i} + \gamma_{i-1}\frac{h_i}{6} + \gamma_i\frac{h_i}{3}, \qquad i = 1, \dots, l\,. \tag{46.6}$$

Die beiden Matrizen in (46.5) haben jeweils die Dimension $(l-1) \times (l+1)$. Für die übliche Interpolationsaufgabe bedeutet dies, daß der Momentenvektor selbst bei Vorgabe aller Funktionswerte s_i, $i = 0, \dots, l$, nicht eindeutig bestimmt ist; es werden noch zwei zusätzliche Gleichungen benötigt. Wir wollen im folgenden zwei Möglichkeiten hierzu vorstellen.

Als erstes betrachten wir den Teilraum der sogenannten *natürlichen kubischen Splines* $s \in S_{3,\triangle}$ mit

$$s''(a) = s''(b) = 0\,.$$

Wegen $\gamma_0 = s''(a)$ und $\gamma_l = s''(b)$ vereinfacht sich das Gleichungssystem (46.5) für natürliche kubische Splines zu

$$\frac{1}{6}\begin{bmatrix} 2(h_1+h_2) & h_2 & & 0 \\ h_2 & 2(h_2+h_3) & \ddots & \\ & \ddots & \ddots & h_{l-1} \\ 0 & & h_{l-1} & 2(h_{l-1}+h_l) \end{bmatrix} \begin{bmatrix} \gamma_1 \\ \gamma_2 \\ \vdots \\ \gamma_{l-1} \end{bmatrix} = \begin{bmatrix} d_1 \\ d_2 \\ \vdots \\ d_{l-1} \end{bmatrix} \tag{46.7}$$

mit

$$d_i = \frac{s_{i+1} - s_i}{h_{i+1}} - \frac{s_i - s_{i-1}}{h_i}, \qquad i = 1, \ldots, l-1. \tag{46.8}$$

Die resultierende $(l-1) \times (l-1)$-Matrix auf der linken Seite von (46.7) ist die Gramsche Matrix der relevanten Hutfunktionen $\{\Lambda_1, \ldots, \Lambda_{l-1}\}$, vgl. Beispiel 45.5. Sie ist somit positiv definit und das Gleichungssystem (46.7) hat für jede Interpolationsvorgabe $s_i = y_i$, $i = 0, \ldots, l$, einen eindeutig bestimmten Lösungsvektor. Wir fassen zusammen:

Satz 46.2. *Seien $\triangle$ ein Gitter über $[a, b]$ und $y_0, \ldots, y_l$ vorgegebene Daten. Dann gibt es genau einen natürlichen kubischen Spline $s \in S_{3,\triangle}$ mit*

$$s(x_i) = y_i, \qquad i = 0, \ldots, l.$$

Beispiele. Wir bestimmen den natürlichen kubischen Spline s, der die Lagrange-Interpolationsbedingungen

$$s(x_i) = \delta_{ij}, \qquad i = 0, \ldots, l,$$

für ein festes $j \in \{0, \ldots, l\}$ erfüllt. Für ein äquidistantes Gitter $\triangle$ mit Gitterweite h hat das zugehörige Gleichungssystem (46.7) die Form

$$\begin{bmatrix} 4 & 1 & & 0 \\ 1 & 4 & \ddots & \\ & \ddots & \ddots & 1 \\ 0 & & 1 & 4 \end{bmatrix} \begin{bmatrix} \gamma_1 \\ \gamma_2 \\ \vdots \\ \gamma_{l-1} \end{bmatrix} = \frac{6}{h^2} \left(e_{j+1} - 2e_j + e_{j-1}\right),$$

wobei die Vektoren e_j, $j = 1, \ldots, l-1$, die kartesischen Basisvektoren im $\mathbb{R}^{l-1}$ bezeichnen und $e_0 = e_l = 0$ gesetzt seien. Links in Abbildung 46.3 sieht man einen solchen Spline, die Interpolationsvorgaben über dem Gitter $\triangle$ sind hierbei durch Kreise markiert. Dabei zeigt sich ein wichtiger Nachteil dieser für die Interpolation wesentlichen Basissplines: Sie sind auf jedem Gitterintervall von Null verschieden. Dies bedeutet, daß die Modifikation der Interpolationsvorgabe in einem einzigen Gitterpunkt den Spline in dem *gesamten* Intervall $[a, b]$ beeinflußt.

Das rechte Bild derselben Abbildung zeigt einen kubischen Spline, der auf einem größtmöglichen Bereich des zugrundeliegenden Intervalls verschwindet. Dieser sogenannte *B-Spline* s muß in dem verbliebenen Restintervall $\mathcal{I}'$ mindestens zwei Wendepunkte haben, d. h. s'' ist ein linearer Spline mit mindestens zwei Nullstellen im Innern von $\mathcal{I}'$. Daher muß $\mathcal{I}'$ aus mindestens vier aufeinanderfolgenden Gitterintervallen von $\triangle$ bestehen. $\Diamond$

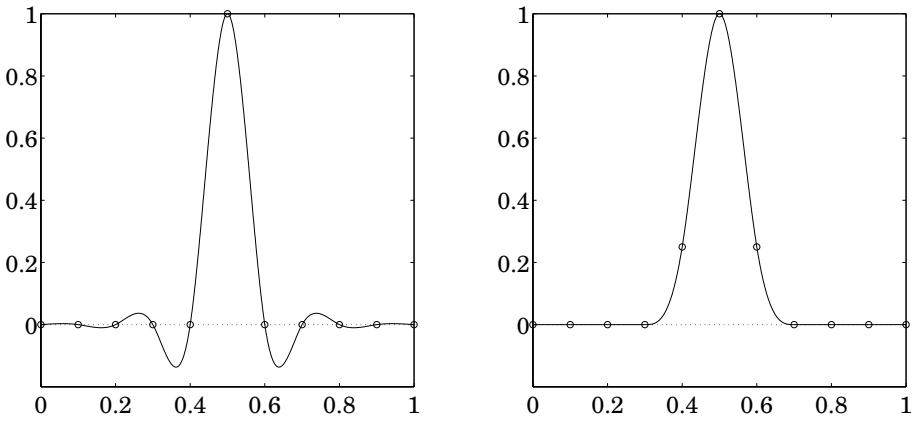


Abb. 46.3: Natürliche kubische Splines

Läßt man die Einschränkung auf natürliche Splines fallen, können Zusatzbedingungen an die interpolierenden kubischen Splines gestellt werden. In der Regel sind dies zusätzliche Randbedingungen, etwa Vorgaben an die erste Ableitung des interpolierenden Splines am Rand. Man spricht in diesem Fall von der *vollständigen Interpolationsaufgabe*

Satz 46.3. *Es gibt genau einen kubischen Spline $s \in S_{3,\triangle}$, der die vollständige Interpolationsaufgabe*

$$s(x_i) = y_i\,, \qquad i = 0, \ldots, l\,,$$

mit den Randbedingungen

$$s'(x_0) = y'_0\,, \quad s'(x_l) = y'_l\,, \tag{46.9}$$

löst. Dabei seien $y_0, \ldots, y_l$, y'_0 und y'_l beliebig vorgegebene Werte.

Beweis. Aus (46.3b) mit $i = 1$ folgt

$$s'(a) \;=\; s'_1 - \gamma_0 h_1/2 - \gamma_1 h_1/2\,.$$

Wegen (46.6) ist die zusätzliche Interpolationsbedingung am linken Rand somit äquivalent zu

$$\frac{h_1}{3}\,\gamma_0 \;+\; \frac{h_1}{6}\,\gamma_1 \;=\; \widetilde{d}_0 \;=\; \frac{y_1 - y_0}{h_1} \;- y'_0\,. \tag{46.10a}$$

Die Randbedingung am rechten Rand erfordert, daß die durch (46.6) gegebene Ableitung s'_l mit der Vorgabe y'_l übereinstimmt. Hieraus ergibt sich die zweite

Gleichung

$$\frac{h_l}{6}\,\gamma_{l-1} + \frac{h_l}{3}\,\gamma_l = \widetilde{d}_l = y_l' - \frac{y_l - y_{l-1}}{h_l}\,. \tag{46.10b}$$

Zusammen mit den Bedingungen (46.5) ergibt dies das Gleichungssystem

$$\frac{1}{6}\begin{bmatrix} 2h_1 & h_1 & & & & 0 \\ h_1 & 2(h_1+h_2) & h_2 & & & \\ & h_2 & \ddots & \ddots & & \\ & & \ddots & 2(h_{l-1}+h_l) & h_l & \\ 0 & & & h_l & 2h_l \end{bmatrix} \begin{bmatrix} \gamma_0 \\ \gamma_1 \\ \vdots \\ \gamma_{l-1} \\ \gamma_l \end{bmatrix} = \begin{bmatrix} \widetilde{d}_0 \\ d_1 \\ \vdots \\ d_{l-1} \\ \widetilde{d}_l \end{bmatrix} \tag{46.11}$$

für die Momente des Splines. Die Koeffizientenmatrix aus (46.11) ist auch hier die Gramsche Matrix der relevanten Hutfunktionen $\{\Lambda_0, \ldots, \Lambda_l\}$, vgl. Beispiel 45.5, und somit ist auch dieses Gleichungssystem eindeutig lösbar. Folglich gibt es genau einen kubischen Spline, der die vollständige Interpolationsaufgabe löst. □

Die Berechung des kubischen Splines für die vollständige Interpolationsaufgabe ist in Algorithmus 46.1 zusammengefaßt. Mit entsprechenden Modifikationen kann der Algorithmus leicht auf die Berechnung des interpolierenden natürlichen Splines übertragen werden.

Aufwand. Da die Koeffizientenmatrix von (46.11) eine Tridiagonalmatrix ist, erfordert das Aufstellen des linearen Gleichungssystems und dessen Lösung nach Aufgabe II.3 nur $O(l)$ Operationen. ◊

Bemerkung 46.4. Sind $y_i = f(x_i)$ die Funktionswerte einer zu interpolierenden Funktion, so liegt es nahe, für y_0' und y_l' die Werte der Ableitung von f an den Intervallrändern a und b einzusetzen. In der Praxis liegen diese Werte oft nicht vor, können jedoch für die vollständige Interpolationsaufgabe aus den bekannten Funktionswerten in der Nähe der Randpunkte approximiert werden. Für ein äquidistantes Gitter $\triangle$ mit Gitterweite h empfehlen sich die Differenzenquotienten

$$\begin{aligned} y_0' &= (-11y_0 + 18y_1 - 9y_2 + 2y_3)/(6h)\,, \\ y_l' &= (11y_l - 18y_{l-1} + 9y_{l-2} - 2y_{l-3})/(6h)\,, \end{aligned} \tag{46.12}$$

die eine Genauigkeit

$$y_0' = f'(a) + O(h^3)\,, \qquad y_l' = f'(b) + O(h^3)\,,$$

aufweisen, sofern f hinreichend glatt ist (vgl. Aufgabe XV.3). ◊

Initialisierung: Sei $x_0 < x_1 < \ldots < x_l$ ein Gitter. Gesucht ist der kubische Spline s mit $s(x_i) = y_i$, $i = 0, \ldots, l$, und $s'(x_0) = y'_0$, $s'(x_l) = y'_l$

% berechne zunächst die rechte Seite $d = [\widetilde{d}_0, d_1, \ldots, d_{l-1}, \widetilde{d}_l]^T$ von (46.11)
$h_1 = x_1 - x_0$
$\widetilde{d}_0 = \dfrac{y_1 - y_0}{h_1} - y'_0$
`for` $i = 1, \ldots, l-1$ `do`
  $h_{i+1} = x_{i+1} - x_i$
  $d_i = \dfrac{y_{i+1} - y_i}{h_{i+1}} - \dfrac{y_i - y_{i-1}}{h_i}$
`end for`
$\widetilde{d}_l = y'_l - \dfrac{y_l - y_{l-1}}{h_l}$
% G bezeichne die Koeffizientenmatrix aus (46.11)
löse $Gc = d$ % ... mit der Cholesky-Zerlegung von G
% sei $c = [\gamma_0, \ldots, \gamma_l]^T$
`for` $i = 1, \ldots, l$ `do`
  $s'_i = \dfrac{y_i - y_{i-1}}{h_i} + \gamma_{i-1} h_i/6 + \gamma_i h_i/3$
`end for` % $s'_i = s'(x_i)$

Ergebnis: für $x \in [x_{i-1}, x_i]$ gilt $s(x) = y_i + s'_i(x - x_i) + \gamma_i \dfrac{(x - x_i)^2}{2} + \dfrac{\gamma_i - \gamma_{i-1}}{h_i} \dfrac{(x - x_i)^3}{6}$

Algorithmus 46.1: Vollständige kubische Splineinterpolation

Beispiel. Als Anwendung der Splineinterpolation betrachten wir folgende Aufgabe: Ein Roboterarm soll zu vorgegebenen Zeitpunkten t_i gewisse Punkte $y_i \in \mathbb{R}^3$, $i = 1, \ldots, l-1$, ansteuern und dann zu seinem Ausgangspunkt y_0 zurückkehren. Der Roboterarm soll sich also auf einer Kurve $s(t)$ im $\mathbb{R}^3$ bewegen, die die Interpolationsbedingungen

$$s(t_i) = y_i\,, \qquad i = 0, \ldots, l\,,$$

mit $y_l = y_0$ erfüllt. Als Nebenbedingung wird gefordert, dass der Roboter zu Beginn und nach Abschluß der Aufgabe im Ruhezustand ist. Die Geschwindigkeit des Roboterarms ist die (komponentenweise gebildete) Zeitableitung $s'(t) \in \mathbb{R}^3$. Die Nebenbedingung führt also auf die vollständige Interpolationsbedingung

$$s'(t_0) = s'(t_l) = 0\,.$$

Nach Satz 46.3 gibt es für jede Ortskoordinate genau einen kubischen Spline, der die entsprechende vollständige Interpolationsaufgabe löst. Der Vektor, der aus diesen drei Splinefunktionen besteht, liefert also eine zulässige Roboterbahn.

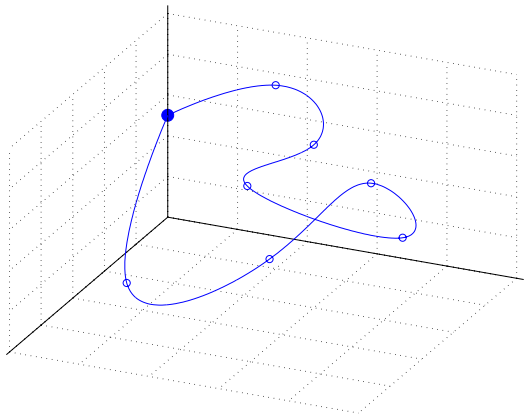


Abb. 46.4: Dreidimensionale Splinekurve

Man beachte, daß die zweite Ableitung $s''(t) \in \mathbb{R}^3$ der Beschleunigung des Roboters entspricht. Für die berechnete Splinekurve verändert sich daher der Beschleunigungsvektor stetig mit der Zeit. Dies vermeidet Verschleißerscheinungen an den Robotergelenken, die durch abrupte Richtungsänderungen hervorgerufen werden können.

Abbildung 46.4 zeigt eine solche Roboterbahn, bei der sieben Punkte nacheinander in gleichen Zeitabständen angefahren werden. Ausgangspunkt und Ziel der Bewegung sind die Ruhelage $(0, 0, 0.5)$, die durch einen fetten Punkt markiert ist. ◊

47 Fehlerabschätzung für kubische Splines

Interpolierende kubische Splines haben eine wichtige Optimalitätseigenschaft. Für eine Funktion $y : [a, b] \to \mathbb{R}$ mißt

$$\frac{y''(x)}{\big(1 + y'(x)^2\big)^{3/2}}$$

die *Krümmung* des Graphs von y im Punkt $\big(x, y(x)\big)$. Beschreibt $y = y(x)$ etwa die Form einer Holzlatte, so bezeichnet

$$E = \frac{1}{2} \int_a^b \Big(\frac{y''(x)}{\big(1 + y'(x)^2\big)^{3/2}} \Big)^2 dx$$

die *Biegeenergie* der Latte.

Ohne Auswirkung äußerer Kräfte nimmt die Latte einen Zustand minimaler Biegeenergie ein, soweit es die örtlichen Gegebenheiten zulassen. Wird beispielsweise durch Pflöcke erzwungen, daß die Latte durch gewisse Punkte (x_i, y_i), $i = 0, \ldots, l$, verläuft, so wird die Latte eine Form einnehmen, die die Biegeenergie E unter den Nebenbedingungen $y(x_i) = y_i$, $i = 0, \ldots, l$, und unter geeigneten Randbedingungen für $x = a$ und $x = b$ minimiert.

Für kleine Auslenkungen $y'(t)$ kann die Biegeenergie durch das quadratische Funktional

$$E \approx \frac{1}{2} \int_a^b \left|y''(x)\right|^2 dx = \frac{1}{2} \|y''\|^2_{\mathcal{L}^2(a,b)}$$

approximiert werden. Wie wir gleich sehen werden, wird dieses Funktional durch den interpolierenden kubischen Spline minimiert. Der Graph des kubischen Splines hat somit näherungsweise die Form einer an den Punkten (x_i, y_i), $i = 0, \ldots, l$, eingespannten Holzlatte.

Satz 47.1. *Der kubische Spline s interpoliere die Punkte (x_i, y_i), $i = 0, \ldots, l$, und $g \in H^2(a, b)$ sei eine beliebige weitere interpolierende Funktion mit*

$$g'(a) = s'(a) \qquad \text{und} \qquad g'(b) = s'(b)\,.$$

Dann gilt $\|s''\|_{\mathcal{L}^2(a,b)} \le \|g''\|_{\mathcal{L}^2(a,b)}$.

Beweis. Es gilt

$$\begin{aligned} \|g''\|^2_{\mathcal{L}^2(a,b)} &= \|s'' + (g'' - s'')\|^2_{\mathcal{L}^2(a,b)} \\ &= \|s''\|^2_{\mathcal{L}^2(a,b)} + \|g'' - s''\|^2_{\mathcal{L}^2(a,b)} + 2\int_a^b s''(g'' - s'')\,dx\,. \end{aligned}$$

Wir zeigen nun, daß der letzte Term verschwindet, also daß

$$\|g''\|^2_{\mathcal{L}^2(a,b)} = \|s''\|^2_{\mathcal{L}^2(a,b)} + \|g'' - s''\|^2_{\mathcal{L}^2(a,b)} \tag{47.1}$$

ist, woraus dann sofort die Behauptung folgt. Dazu verwenden wir partielle Integration, vgl. Aufgabe VI.4:

$$\int_a^b s''(g'' - s'')\,dx = s''(g' - s')\Big|_a^b - \int_a^b s'''(g' - s')\,dx\,.$$

Der erste Term auf der rechten Seite verschwindet aufgrund der Randbedingungen an g; da s''' im Innern von $[x_{i-1}, x_i]$ konstant ist, etwa gleich σ_i, folgt somit

$$\int_a^b s''(g'' - s'')\,dx = -\sum_{i=1}^{l} \sigma_i \int_{x_{i-1}}^{x_i} (g' - s')\,dx = -\sum_{i=1}^{l} \sigma_i (g - s)\Big|_{x_{i-1}}^{x_i} = 0\,,$$

da g und s die gleichen Daten interpolieren. Also gilt (47.1) wie behauptet. □

Bemerkung 47.2. Ist s ein natürlicher Spline, dann gilt die Identität (47.1) und somit die Aussage von Satz 47.1 für *jede* Funktion $g \in H^2(a,b)$, die dieselben Punkte (x_i, y_i), $i = 0, \ldots, l$, interpoliert. Dies folgt unmittelbar aus dem obigen Beweis. ◊

Korollar 47.3. *Sei $f \in H^2(a,b)$ und s der interpolierende kubische Spline zu f bezüglich des Gitters $\triangle$ mit den Randbedingungen $s'(a) = f'(a)$ und $s'(b) = f'(b)$. Dann gilt*

$$\|f'' - s''\|_{\mathcal{L}^2(a,b)} \leq \|f'' - \tilde{s}\|_{\mathcal{L}^2(a,b)}$$

für alle linearen *Splines $\tilde{s} \in S_{1,\triangle}$.*

Beweis. Zu $\tilde{s} \in S_{1,\triangle}$ wählen wir $\tilde{S} \in S_{3,\triangle}$ mit $\tilde{S}'' = \tilde{s}$. Wir wenden (47.1) auf die Funktion $g = f - \tilde{S}$ an und berücksichtigen, daß $s - \tilde{S}$ der vollständig interpolierende kubische Spline zu g ist. Es folgt also

$$\begin{aligned} \|(f - \tilde{S})''\|^2_{\mathcal{L}^2(a,b)} &= \|g''\|^2_{\mathcal{L}^2(a,b)} \geq \|g'' - (s - \tilde{S})''\|^2_{\mathcal{L}^2(a,b)} \\ &= \|(f - \tilde{S})'' - (s - \tilde{S})''\|^2_{\mathcal{L}^2(a,b)} = \|f'' - s''\|^2_{\mathcal{L}^2(a,b)}\,, \end{aligned}$$

was zu beweisen war. □

Bemerkung 47.4. Korollar 47.3 besagt, daß die zweite Ableitung des vollständig interpolierenden kubischen Splines zu f die *Bestapproximation* (bezüglich $\mathcal{L}^2$) an die Funktion $g = f''$ aus dem Raum der *linearen Splines* ist. Dieses Ergebnis hätte man auch aus dem linearen Gleichungssystem (46.11) für die Momente ablesen können: Da die Koeffizientenmatrix von (46.11) die Gramsche Matrix (45.4) der Hutfunktionen ist, ist die Funktion s'' nach Satz 31.10 die Bestapproximation an jene Funktion g, für die die rechte Seite von (46.11) mit den entsprechenden Innenprodukten übereinstimmt:

$$\langle\, \Lambda_0, g \,\rangle = \tilde{d}_0\,, \quad \langle\, \Lambda_i, g \,\rangle = d_i\,,\ i = 1, \ldots, l-1\,, \quad \langle\, \Lambda_l, g \,\rangle = \tilde{d}_l\,.$$

Es ist nicht schwer nachzurechnen, daß dies für $g = f''$ erfüllt ist. ◊

Analog zu dem Raum $H^2(a,b)$ führen wir schließlich noch den Sobolevraum $H^4(a,b)$ derjenigen Funktionen $f \in H^2(a,b)$ ein, deren zweite Ableitung f'' wiederum zu $H^2(a,b)$ gehört. Eine solche Funktion gehört also zu $C^3[a,b]$, die dritte Ableitung f''' liegt sogar in $H^1(a,b)$, und es gibt darüber hinaus eine schwache vierte Ableitung $f^{(4)} \in \mathcal{L}^2(a,b)$ von f.

Satz 47.5. *Sei $f \in H^4(a,b)$ und $\triangle$ ein Gitter über $[a,b]$ mit Gitterweite h. Ist s der interpolierende kubische Spline zu f mit $s'(a) = f'(a)$ und $s'(b) = f'(b)$, dann ist*

$$\|f - s\|_{\mathcal{L}^2(a,b)} \leq \frac{h^4}{4}\, \|f^{(4)}\|_{\mathcal{L}^2(a,b)}\,. \tag{47.2}$$

Beweis. Wir wenden Korollar 47.3 an und vergleichen den Abstand von f'' und s'' in der $\mathcal{L}^2$-Norm mit dem Approximationsfehler des f'' linear interpolierenden Splines $\tilde{s} \in S_{1,\triangle}$. Da nach Voraussetzung f'' zu $H^2(a,b)$ gehört, gilt nach Satz 45.4

$$\|f'' - \tilde{s}\|_{\mathcal{L}^2(a,b)} \leq \frac{h^2}{2} \|f^{(4)}\|_{\mathcal{L}^2(a,b)}$$

und aus Korollar 47.3 folgt somit

$$\|f'' - s''\|_{\mathcal{L}^2(a,b)} \leq \frac{h^2}{2} \|f^{(4)}\|_{\mathcal{L}^2(a,b)} . \tag{47.3}$$

Aufgrund der Interpolationseigenschaft des kubischen Splines s verschwindet die Funktion $f - s$ an allen Gitterpunkten x_i, $i = 0, \ldots, l$, von $\triangle$. Folglich ist der (eindeutig bestimmte) lineare Spline φ, der $f - s$ in $\triangle$ interpoliert, die Nullfunktion $\varphi = 0$. Wegen $f - s \in H^2(a,b)$ kann der Interpolationsfehler von φ wieder mit Hilfe von Satz 45.4 abgeschätzt werden und somit gilt

$$\|f - s\|_{\mathcal{L}^2(a,b)} = \|f - s - \varphi\|_{\mathcal{L}^2(a,b)} \leq \frac{h^2}{2} \|f'' - s''\|_{\mathcal{L}^2(a,b)} .$$

Die Behauptung folgt schließlich durch Einsetzen von (47.3). □

Bemerkung. Für äquidistante Gitter lassen sich entsprechende Abschätzungen beweisen, falls die Randapproximationen (46.12) aus Bemerkung 46.4 für die vollständige Interpolation verwendet werden, vgl. Swartz und Varga [99]. ◊

48 Geglättete kubische Splines

Im folgenden betrachten wir den Fall, daß die zu interpolierenden Werte y_i nur im Rahmen einer möglichen Ungenauigkeit δ mit den Funktionswerten der zugrunde liegenden Funktion y übereinstimmen. Dabei sei vorausgesetzt, daß y zu $H^2(a,b)$ gehört und homogene Randwerte $y(a) = y(b) = 0$ besitzt. Aufgrund der Unsicherheit in den Daten,

$$|y_i - y(x_i)| \leq \delta , \qquad i = 1, \ldots, l-1 ,$$

bestimmen wir in diesem Fall nicht den interpolierenden Spline sondern suchen statt dessen die „glatteste“ Funktion f, welche die Interpolationsaufgabe in einem Kleinste-Quadrate-Sinn löst:

Problem 48.1. *Minimiere* $\|f''\|_{\mathcal{L}^2(a,b)}$ *unter allen Funktionen* $f \in H^2(a,b)$ *mit* $f(a) = f(b) = 0$ *und unter der Nebenbedingung*

$$\frac{1}{l-1} \sum_{i=1}^{l-1} |y_i - f(x_i)|^2 \;\le\; \delta^2 \,. \tag{48.1}$$

Wie wir gleich sehen werden, ist die Lösung von Problem 48.1 wieder ein kubischer Spline über dem Gitter $\triangle$.

Satz 48.2. f_* *sei ein natürlicher kubischer Spline über dem Gitter* $\triangle$ *mit* $f_*(a) = f_*(b) = 0$*, der die Nebenbedingung* (48.1) *mit Gleichheit erfüllt,*

$$\sum_{i=1}^{l-1} |y_i - f_*(x_i)|^2 \;=\; (l-1)\,\delta^2 \,,$$

und dessen dritte Ableitung an den Gitterknoten für ein $\lambda > 0$ *die folgenden Sprünge aufweist:*

$$\big[\, f_*''' \,\big]_{x_i} \;:=\; f_*'''(x_i+) - f_*'''(x_i-) \;=\; \lambda\big(y_i - f_*(x_i)\big) \,, \tag{48.2}$$

$i = 1, \dots, l-1$*. Dann ist* f_* *die eindeutig bestimmte Lösung von Problem* 48.1*.*

Beweis. Wir wählen eine beliebige Funktion $g \in H^2(a,b)$ mit $g(a) = g(b) = 0$. Da f_* ein natürlicher kubischer Spline sein soll, ergibt sich durch partielle Integration

$$\begin{aligned} \int_a^b g''(x) f_*''(x)\,dx \;&=\; g'(b) f_*''(b) - g'(a) f_*''(a) \;-\; \int_a^b g'(x) f_*'''(x)\,dx \\ &=\; -\sum_{i=1}^{l} \int_{x_{i-1}}^{x_i} g'(x) f_*'''(x)\,dx \;=\; -\sum_{i=1}^{l} f_*'''\big|_{[x_{i-1},x_i]} g(x)\Big|_{x_{i-1}}^{x_i} . \end{aligned}$$

Da g an den Randpunkten $x_0 = a$ und $x_l = b$ verschwindet, kann diese Summe wie folgt umgeordnet werden:

$$\int_a^b g''(x) f_*''(x)\,dx \;=\sum_{i=1}^{l-1} g(x_i) \,\big[\, f_*''' \,\big]_{x_i} . \tag{48.3}$$

Sei nun $f \in H^2(a,b)$ eine beliebige Funktion mit $f(a) = f(b) = 0$, die die Nebenbedingung (48.1) erfüllt. Wenden wir (48.3) auf die Funktion $g = f - f_*$

an, dann erhalten wir

$$\begin{aligned}
&\|f''\|^2_{\mathcal{L}^2(a,b)} - \|f_*''\|^2_{\mathcal{L}^2(a,b)} \\
&\quad = \quad \|f'' - f_*''\|^2_{\mathcal{L}^2(a,b)} + 2\int_a^b (f - f_*)''(x) f_*''(x)\,dx \\
&\quad \overset{(48.3)}{=} \quad \|f'' - f_*''\|^2_{\mathcal{L}^2(a,b)} + 2\sum_{i=1}^{l-1} \big(f(x_i) - f_*(x_i)\big)\,\big[\,f_*'''\,\big]_{x_i} \\
&\quad \overset{(48.2)}{=} \quad \|f'' - f_*''\|^2_{\mathcal{L}^2(a,b)} + 2\lambda\sum_{i=1}^{l-1} \big(f(x_i) - f_*(x_i)\big)\big(y_i - f_*(x_i)\big)\,.
\end{aligned}$$

Bezeichnen $r, r_* \in \mathbb{R}^{l-1}$ die Vektoren

$$r \;=\; \big[y_i - f(x_i)\big]_{i=1}^{l-1}, \qquad r_* \;=\; \big[y_i - f_*(x_i)\big]_{i=1}^{l-1},$$

dann kann dies folgendermaßen geschrieben werden:

$$\begin{aligned}
\|f''\|^2_{\mathcal{L}^2(a,b)} - \|f_*''\|^2_{\mathcal{L}^2(a,b)} \;&=\; \|f'' - f_*''\|^2_{\mathcal{L}^2(a,b)} + 2\lambda(r_* - r)^* r_* \\
&=\; \|f'' - f_*''\|^2_{\mathcal{L}^2(a,b)} + \lambda\big(\|r_* - r\|_2^2 + \|r_*\|_2^2 - \|r\|_2^2\big)\,.
\end{aligned}$$

Nach Voraussetzung ist $\|r_*\|_2^2 = (l-1)\delta^2$, während $\|r\|_2^2$ aufgrund der Nebenbedingung (48.1) höchstens so groß wie $(l-1)\delta^2$ ist. Folglich ist

$$\|f''\|^2_{\mathcal{L}^2(a,b)} - \|f_*''\|^2_{\mathcal{L}^2(a,b)} \;\geq\; \|f'' - f_*''\|^2_{\mathcal{L}^2(a,b)} + \lambda\,\|r_* - r\|_2^2 \;\geq\; 0\,, \tag{48.4}$$

so daß f_* (sofern es existiert) eine Lösung von Problem 48.1 ist.

Für den Nachweis der Eindeutigkeit muß man in (48.4) das letzte Gleichheitszeichen diskutieren. Offensichtlich liegt Gleichheit höchstens dann vor, wenn $f'' - f_*''$ verschwindet, also wenn $f - f_*$ eine lineare Funktion ist. Da f und f_* beide an den Randpunkten $x_0 = a$ und $x_l = b$ Nullstellen besitzen, ergibt dies die Gleichheit von f und f_*. Also ist das Minimum unter den gemachten Annahmen eindeutig bestimmt. □

Satz 48.2 macht keine Aussage darüber, ob die gestellten Anforderungen an den Spline f_* erfüllbar sind. Diese Frage beantwortet der folgende Satz, in dem f_* explizit konstruiert wird.

Satz 48.3. *Unter der Voraussetzung*

$$\frac{1}{l-1}\sum_{i=1}^{l-1} |y_i|^2 > \delta^2 \tag{48.5}$$

existiert ein Spline $s = f_*$, der alle Bedingungen aus Satz 48.2 *erfüllt.*

Beweis. Sei $c \in \mathbb{R}^{l-1}$ der Vektor der Momente $\gamma_i = s''(x_i)$, $i = 1, \dots, l-1$, eines natürlichen kubischen Splines s. Nach Gleichung (46.7) besteht zwischen diesem Momentenvektor und den Funktionswerten $\boldsymbol{s} = [s(x_1), \dots, s(x_{l-1})]^T$ von s der Zusammenhang

$$Gc = T\boldsymbol{s}\,, \tag{48.6}$$

falls $s(x_0) = s(x_l) = 0$ ist. Dabei sind G und T durch

$$G = \frac{1}{6}\begin{bmatrix} 2(h_1+h_2) & h_2 & & 0 \\ h_2 & 2(h_2+h_3) & \ddots & \\ & \ddots & \ddots & h_{l-1} \\ 0 & & h_{l-1} & 2(h_{l-1}+h_l) \end{bmatrix}$$

und

$$T = -\begin{bmatrix} h_1^{-1}+h_2^{-1} & -h_2^{-1} & & 0 \\ -h_2^{-1} & h_2^{-1}+h_3^{-1} & \ddots & \\ & \ddots & \ddots & -h_{l-1}^{-1} \\ 0 & & -h_{l-1}^{-1} & h_{l-1}^{-1}+h_l^{-1} \end{bmatrix}$$

gegeben, vgl. auch (46.5). Die Matrix T tritt auch bei der Berechnung der Sprünge der dritten Ableitung von s an den Gitterknoten auf: Wegen

$$\left[\, s''' \,\right]_{x_i} \;=\; s'''(x_i+) - s'''(x_i-) \;=\; \frac{\gamma_{i+1}-\gamma_i}{h_{i+1}} \;-\; \frac{\gamma_i - \gamma_{i-1}}{h_i}$$

für $i = 1, \dots, l-1$, entspricht die Bedingung (48.2) aus Satz 48.2 nämlich dem linearen Gleichungssystem

$$Tc \;=\; \lambda\,(\boldsymbol{y} - \boldsymbol{s}) \tag{48.7}$$

mit $\boldsymbol{y} = [y_i] \in \mathbb{R}^{l-1}$. Da T irreduzibel diagonaldominant und somit nach Satz 23.3 nichtsingulär ist, führt eine Multiplikation mit T von links und Einsetzen von (48.6) auf das äquivalente Gleichungssystem

$$T^2c \;=\; \lambda\,(T\boldsymbol{y} - T\boldsymbol{s}) \;=\; \lambda\,T\boldsymbol{y} - \lambda\,Gc$$

beziehungsweise

$$(G + \alpha T^2)c = T\boldsymbol{y}\,, \qquad \alpha = 1/\lambda\,. \tag{48.8}$$

G und T^2 sind beide symmetrisch und positiv definit. Folglich ist $G + \alpha T^2$ für jedes positive α invertierbar, das heißt für jedes $\alpha = 1/\lambda > 0$ existiert

ein natürlicher kubischer Spline mit homogenen Randwerten, der die Bedingung (48.2) erfüllt.

Es bleibt daher nur noch zu zeigen, daß der positive Parameter λ so gewählt werden kann, daß s auch die Nebenbedingung (48.1) mit Gleichheit, also

$$\|\boldsymbol{y} - \boldsymbol{s}\|_2^2 \;=\; (l-1)\,\delta^2 \tag{48.9}$$

erfüllt. Wegen (48.7) und (48.8) führt dies auf die Forderung

$$\begin{aligned}(l-1)\delta^2 \;&\stackrel{!}{=}\; \frac{1}{\lambda^2}\,\|Tc\|_2^2 \;=\; \frac{1}{\lambda^2}\,\|T(G+\frac{1}{\lambda}T^2)^{-1}T\boldsymbol{y}\|_2^2 \\ &=\; \|(I+\lambda T^{-1}GT^{-1})^{-1}\boldsymbol{y}\|_2^2 =: r(\lambda)\,.\end{aligned} \tag{48.10}$$

Nach Aufgabe 13 hat diese Funktion die Form der Funktion r aus (18.10) in Abschnitt 18.3. Somit existiert nach (18.11) genau dann eine eindeutige positive Lösung λ der Gleichung $r(\lambda) = (l-1)\delta^2$, wenn

$$r(0) \;=\; \|\boldsymbol{y}\|_2^2 \;>\; (l-1)\delta^2\,.$$

Dies entspricht aber gerade der Voraussetzung (48.5). □

Wir fassen zusammen: Falls die Bedingung (48.5) erfüllt ist, also falls der Datenfehler relativ klein ist, ergibt sich die Lösung f_* von Problem 48.1 als natürlicher kubischer Spline über $\triangle$, der über seinen Momentenvektor c als Lösung von (48.8) definiert ist. Der Parameter λ ist dabei so einzustellen, daß die Säkulargleichung (48.10) erfüllt ist. Ist hingegen die Bedingung (48.5) nicht erfüllt, dann hat Problem 48.1 offensichtlich die triviale Lösung $f = 0$.

Bemerkungen. Der Parameter λ ist der *Lagrange-Parameter* zu dem Minimierungsproblem 48.1 unter der Nebenbedingung

$$\sum_{i=1}^{l-1} |y_i - f(x_i)|^2 \;=\; (l-1)\,\delta^2\,,$$

vgl. auch Bemerkung 21.2. Tatsächlich erkennt man aus dem Beweis von Satz 48.2, daß der Spline f_* das zugehörige Lagrange-Funktional

$$\|f''\|_{\mathcal{L}^2(a,b)}^2 + \lambda \sum_{i=1}^{l-1} |y_i - f(x_i)|^2$$

mit dem Parameter λ aus (48.2) minimiert, vgl. Aufgabe 11. Dividiert man dieses Funktional durch λ und führt wieder $\alpha = 1/\lambda$ ein, so ergibt sich das Minimierungsproblem

$$\text{minimiere} \quad \sum_{i=1}^{l-1} |y_i - f(x_i)|^2 + \alpha\,\|f''\|_{\mathcal{L}^2(a,b)}^2\,,$$

das einer *Tikhonov-Regularisierung* der Interpolationsaufgabe entspricht, vgl. Aufgabe III.5. Für $\alpha = 0$ werden die Daten (sofern sie widerspruchsfrei sind) exakt interpoliert und das Minimum ist Null. Bei großen Datenfehlern würde dies zu starken Oszillationen der interpolierenden Funktionen führen, was durch Hinzunahme des *Strafterms* $\|f''\|^2_{\mathcal{L}^2(a,b)}$ unterdrückt wird. ◊

Aufwand. Für jeden Wert von α (respektive $\lambda = 1/\alpha$) kann der zugehörige kubische Spline mit $O(l)$ Operationen berechnet werden. Für die Lösung des Problems 48.1 muß der Parameter λ jedoch die Gleichung (48.10) erfüllen. Wie im Beweis von Satz 48.3 bereits angedeutet wurde, kann diese nichtlineare Gleichung effizient mit dem Hebden-Verfahren aus Abschnitt 18.3 (etwa unter Verwendung des Sekantenverfahrens) gelöst werden. Dabei ist in jedem Iterationsschritt ein kubischer Spline zu berechnen. Aufgrund der schnellen Konvergenz des Sekantenverfahrens ist der Gesamtaufwand für die Lösung von Problem 48.1 ebenfalls von der Größenordnung $O(l)$. ◊

49 Numerische Differentiation

Der geglättete kubische Spline aus dem vorangegangenen Abschnitt wird in der Praxis häufig eingesetzt, um die Ableitung einer Funktion y zu approximieren, von der nur ungenaue Funktionswerte $y_i \approx y(x_i)$ über einem Gitter $\triangle$ bekannt sind.

Beispiel. Die zum Teil fatalen Auswirkungen solcher Datenfehler werden anhand des folgenden Rechenbeispiels klar: Sind

$$y_{i-1} = y(x_{i-1}) - \delta \qquad \text{und} \qquad y_i = y(x_i) + \delta$$

die gegebenen Daten für zwei benachbarte Gitterpunkte, so ergibt die übliche Approximation der Ableitung von y durch einen Differenzenquotienten

$$\frac{y_i - y_{i-1}}{h_i} \;=\; \frac{y(x_i) - y(x_{i-1})}{x_i - x_{i-1}} \;+\; 2\frac{\delta}{h_i} \;=\; y'(x_i) \;+\; \varepsilon_i \;+\; 2\frac{\delta}{h_i}\,.$$

Um den Approximationsfehler ε_i abzuschätzen, verwenden wir den Mittelwertsatz: Für eine entsprechende Zwischenstelle $\xi \in (x_{i-1}, x_i)$ ergibt dies

$$\varepsilon_i \;=\; y'(\xi) - y'(x_i) \;=\; O(h_i)\,.$$

Damit hat diese Approximation der Ableitung den Fehler

$$\Big|\, \frac{y_i - y_{i-1}}{h_i} - y'(x_i) \,\Big| \;=\; 2\frac{\delta}{h_i} \;+\; O(h_i)\,. \tag{49.1}$$

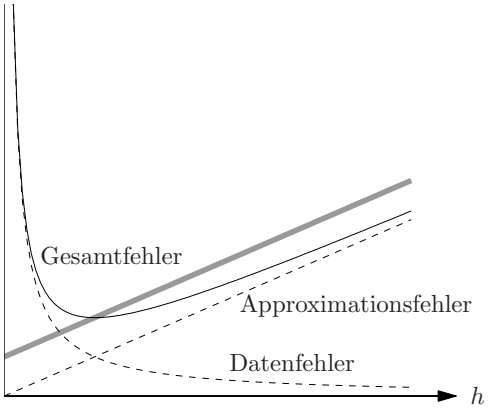


Abb. 49.1:
Approximationsfehler und fortgepflanzter Datenfehler

Für $\delta > 0$ ergeben sich somit bei kleinen Gitterweiten sehr große Fehler, obwohl der eigentliche Approximationsfehler ε_i in dem Fall klein ist. Abbildung 49.1 illustriert die beiden gegenläufigen Fehleranteile auf der rechten Seite von (49.1) (die fettere Linie wird weiter unten erläutert). Man beachte, daß der Gesamtfehler in der Regel mindestens die Größenordnung $\sqrt{\delta}$ hat. ◊

Wir beweisen nun eine Fehlerabschätzung aus [48] für die Ableitung des geglätteten kubischen Splines, falls das verwendete Gitter $\triangle$ äquidistant ist. Wie immer setzen wir dafür eine gewisse Glattheit der Ausgangsfunktion y voraus. Aufgrund der Problemstellung 48.1 erscheint hier die Forderung $y \in H^2(a,b)$ angemessen.

Im Beweis der Fehlerabschätzung wird der geglättete Spline mit dem (natürlichen) interpolierenden kubischen Spline der exakten Funktionswerte von y über dem Gitter $\triangle$ verglichen. Daher beweisen wir zunächst das folgende Hilfsresultat.

Lemma 49.1. *Sei $y \in H^2(a,b)$ und s der zugehörige interpolierende natürliche kubische Spline über dem äquidistanten Gitter $\triangle$ mit Gitterweite h. Dann gilt*

$$\|s' - y'\|_{\mathcal{L}^2(a,b)} \leq \frac{h}{\sqrt{2}} \|y''\|_{\mathcal{L}^2(a,b)} .$$

Beweis. Wir interpolieren $s - y$ über $\triangle$ mit dem linearen Spline φ. Nach Voraussetzung ist $s(x_i) = y(x_i)$, $i = 1, \ldots, l$, und daher φ die Nullfunktion. Die Abschätzung aus Satz 45.4 für den Interpolationsfehler von φ liefert somit in diesem speziellen Fall

$$\|s' - y'\|_{\mathcal{L}^2(a,b)} = \|s' - y' - \varphi'\|_{\mathcal{L}^2(a,b)} \leq \frac{h}{\sqrt{2}} \|s'' - y''\|_{\mathcal{L}^2(a,b)} .$$

Nach Bemerkung 47.2 gilt (47.1) mit y anstelle von g, und daher ist die $\mathcal{L}^2$-Norm von $s'' - y''$ höchstens so groß wie die von y''. Oben eingesetzt, ergibt dies die Behauptung. □

Nun zu dem eigentlichen Hauptresultat dieses Abschnitts.

Satz 49.2. *Sei $y \in H^2(a,b)$, $y(a) = y(b) = 0$, und $|y_i - y(x_i)| \le \delta$ für $i = 1, \dots, l-1$. Dann gilt für die Ableitung f_*' der Lösung f_* von Problem* 48.1 *die Abschätzung*

$$\|f_*' - y'\|_{\mathcal{L}^2(a,b)} \le \sqrt{8}\left(h\,\|y''\|_{\mathcal{L}^2(a,b)} + \left(\sqrt{b-a}\,\delta\,\|y''\|_{\mathcal{L}^2(a,b)}\right)^{1/2}\right).$$

Beweis. Wir schreiben $f_*' - y' = (f_*' - s') + (s' - y')$, wobei s den natürlichen kubischen Spline bezeichnet, der y in den Gitterpunkten von $\triangle$ interpoliert. Lemma 49.1 liefert eine Ungleichung für den Fehler $s' - y'$, es verbleibt daher noch die Abschätzung der Ableitung von $e = f_* - s$.

Sei σ die (bezüglich $\mathcal{L}^2$) bestapproximierende Treppenfunktion über $\triangle$ an e',

$$\sigma = \sum_{i=1}^{l} \alpha_i \chi_i\,, \tag{49.2}$$

wobei $\chi_i = \chi_{[x_{i-1},x_i]}$, $i = 1, \dots, l$, die charakteristischen Funktionen der einzelnen Teilintervalle sind und

$$\alpha_i = \frac{1}{h}\int_{x_{i-1}}^{x_i} e'(x)\,dx = \frac{e(x_i) - e(x_{i-1})}{h}\,, \qquad i = 1, \dots, l\,,$$

vgl. Satz 43.1. Zur Abschätzung von $\|e'\|_{\mathcal{L}^2(a,b)}$ zerlegen wir

$$\|e'\|^2_{\mathcal{L}^2(a,b)} = \int_a^b e'(x)\big(e'(x) - \sigma(x)\big)\,dx + \int_a^b e'(x)\sigma(x)\,dx \tag{49.3}$$

und untersuchen die beiden Summanden separat.

Der erste Term kann mit der Cauchy-Schwarz-Ungleichung und Satz 43.1 wie folgt abgeschätzt werden:

$$\begin{aligned}\int_a^b e'(x)\big(e'(x) - \sigma(x)\big)\,dx &\le \|e'\|_{\mathcal{L}^2(a,b)}\,\|e' - \sigma\|_{\mathcal{L}^2(a,b)} \\ &\le h\,\|e'\|_{\mathcal{L}^2(a,b)}\,\|e''\|_{\mathcal{L}^2(a,b)}\,.\end{aligned}$$

Mit der Dreiecksungleichung ergibt sich weiterhin

$$\|e''\|_{\mathcal{L}^2(a,b)} \le \|f_*''\|_{\mathcal{L}^2(a,b)} + \|s''\|_{\mathcal{L}^2(a,b)} \le 2\|y''\|_{\mathcal{L}^2(a,b)}\,, \tag{49.4}$$

wobei letzteres daraus folgt, daß sowohl die $\mathcal{L}^2$-Norm von s'' als auch die von f_*'' durch $\|y''\|_{\mathcal{L}^2(a,b)}$ beschränkt sind: Die Ungleichung $\|s''\|_{\mathcal{L}^2(a,b)} \le \|y''\|_{\mathcal{L}^2(a,b)}$ folgt aus (47.1) (vgl. Bemerkung 47.2) und die Abschätzung $\|f_*''\|_{\mathcal{L}^2(a,b)} \le \|y''\|_{\mathcal{L}^2(a,b)}$ ergibt sich aus dem Minimierungsproblem 48.1, da sowohl f_* als auch y die dortigen Voraussetzungen erfüllen. Also haben wir gezeigt, daß

$$\int_a^b e'(x)\big(e'(x)-\sigma(x)\big)\,dx \;\le\; 2h\,\|e'\|_{\mathcal{L}^2(a,b)}\,\|y''\|_{\mathcal{L}^2(a,b)}\,. \tag{49.5}$$

Zur Abschätzung des zweiten Terms verwenden wir die Darstellung von σ aus (49.2). Damit folgt

$$\begin{aligned}\int_a^b e'(x)\sigma(x)\,dx &= \sum_{i=1}^{l}\alpha_i\int_{x_{i-1}}^{x_i} e'(x)\,dx \;=\; \sum_{i=1}^{l}\alpha_i\,\big(e(x_i)-e(x_{i-1})\big)\\ &= \sum_{i=1}^{l-1} e(x_i)\,(\alpha_i-\alpha_{i+1}) \;+\; e(b)\alpha_l - e(a)\alpha_1\,.\end{aligned}$$

Da sowohl f_* als auch s die Randwerte $y_0 = y_l = 0$ exakt interpolieren, ist $e(a) = e(b) = 0$. Eine Anwendung der Cauchy-Schwarz-Ungleichung in $\mathbb{R}^{l-1}$ ergibt somit

$$\begin{aligned}\Big(\int_a^b e'(x)\sigma(x)\,dx\Big)^2 &\le \sum_{i=1}^{l-1} e^2(x_i)\,\sum_{i=1}^{l-1}\big(\alpha_i-\alpha_{i+1}\big)^2\\ &= \sum_{i=1}^{l-1} e^2(x_i)\,\sum_{i=1}^{l-1}\frac{1}{h^2}\,\Big(\int_{x_{i-1}}^{x_i}\big(e'(x)-e'(x+h)\big)\,dx\Big)^2\,.\end{aligned}$$

Hierbei ist

$$\begin{aligned}\sum_{i=1}^{l-1} e^2(x_i) &= \sum_{i=1}^{l-1}\big(f_*(x_i)-y_i+y_i-y(x_i)\big)^2\\ &\le \sum_{i=1}^{l-1} 2\,\Big(|f_*(x_i)-y_i|^2+|y_i-y(x_i)|^2\Big) \;\le\; 4(l-1)\delta^2\,,\end{aligned}$$

während die einzelnen Summanden des zweiten Faktors durch

$$\begin{aligned}&\Big|\int_{x_{i-1}}^{x_i}\big(e'(x)-e'(x+h)\big)\,dx\Big|\\ &\qquad\le \int_{x_{i-1}}^{x_i}\int_{x}^{x+h}|e''(t)|\,dt\,dx \;\le\; \int_{x_{i-1}}^{x_i}\int_{x_{i-1}}^{x_{i+1}}|e''(t)|\,dt\,dx\\ &\qquad= h\int_{x_{i-1}}^{x_{i+1}}|e''(t)|\,dt \;\le\; h\sqrt{2h}\,\Big(\int_{x_{i-1}}^{x_{i+1}}|e''(t)|^2\,dt\Big)^{1/2}\end{aligned}$$

abgeschätzt werden können. Insgesamt folgt somit wegen $h = (b-a)/l$ und (49.4)

$$\Big(\int_a^b e'(x)\sigma(x)\,dx\Big)^2 \;\le\; 4(l-1)\,\delta^2 \sum_{i=1}^{l-1} 2h \int_{x_{i-1}}^{x_{i+1}} |e''(t)|^2\,dt$$
$$\le\; 8(b-a)\,\delta^2 \Big(2\int_a^b |e''(t)|^2\,dt\Big) \;\le\; 64(b-a)\,\delta^2\,\|y''\|^2_{\mathcal{L}^2(a,b)}\,.$$

Zusammen mit (49.5) setzen wir dies nun in (49.3) ein und erhalten

$$\|e'\|^2_{\mathcal{L}^2(a,b)} \;\le\; 2h\|e'\|_{\mathcal{L}^2(a,b)}\|y''\|_{\mathcal{L}^2(a,b)} \;+\; 8\sqrt{b-a}\,\delta\,\|y''\|_{\mathcal{L}^2(a,b)}\,.$$

Durch quadratische Ergänzung folgt hieraus

$$\Big(\|e'\|_{\mathcal{L}^2(a,b)} - h\|y''\|_{\mathcal{L}^2(a,b)}\Big)^2 \;\le\; \Big(h\|y''\|_{\mathcal{L}^2(a,b)} + \big(8\sqrt{b-a}\,\delta\big)^{1/2}\,\|y''\|^{1/2}_{\mathcal{L}^2(a,b)}\Big)^2$$

und somit ist

$$\|e'\|_{\mathcal{L}^2(a,b)} \;\le\; 2h\|y''\|_{\mathcal{L}^2(a,b)} + \big(8\sqrt{b-a}\,\delta\big)^{1/2}\,\|y''\|^{1/2}_{\mathcal{L}^2(a,b)}\,.$$

Zusammen mit Lemma 49.1 ergibt dies

$$\|f_*' - y'\|_{\mathcal{L}^2(a,b)} \;\le\; \|e'\|_{\mathcal{L}^2(a,b)} \;+\; \|s' - y'\|_{\mathcal{L}^2(a,b)}$$
$$\le\; 2h\|y''\|_{\mathcal{L}^2(a,b)} + \big(8\sqrt{b-a}\,\delta\big)^{1/2}\,\|y''\|^{1/2}_{\mathcal{L}^2(a,b)} + \frac{h}{\sqrt{2}}\,\|y''\|_{\mathcal{L}^2(a,b)}\,.$$

Da $2 + 1/\sqrt{2}$ kleiner als $\sqrt{8}$ ist, ist der Satz somit vollständig bewiesen. □

Die fettere Linie in Abbildung 49.1 gehört zu der Fehlerabschätzung des geglätteten kubischen Splines. Dieser Fehler setzt sich nach Satz 49.2 aus zwei Komponenten zusammen. Der erste, von δ unabhängige Term, ergibt sich aus dem Interpolationsfehler bei exakten Daten, vgl. Lemma 49.1. Dieser Approximationsfehler ist für allgemeine Funktionen $y \in H^2(a,b)$ unvermeidlich. Der zweite Term beschreibt den Einfluß des Datenfehlers, der mit $O(\sqrt{\delta})$ in die Fehlerschranke eingeht – unabhängig von der Größe von h. Für $h \ll \sqrt{\delta}$ ist dieser Fehleranteil dominierend. Bei der Approximation von y' mit Differenzenquotienten wächst der Datenfehler nach (49.1) für $h \to 0$ hingegen unbeschränkt an. Ein entscheidender Vorteil der glättenden kubischen Splines gegenüber den Differenzenquotienten besteht jedoch darin, daß f_*' eine *glatte* Approximation an y' liefert, nämlich einen quadratischen Spline.

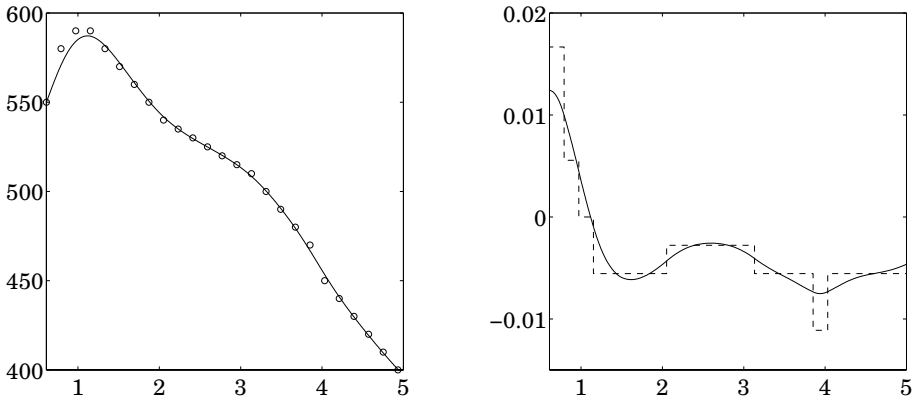


Abb. 49.2: Geglätteter kubischer Spline und numerische Ableitungen

Beispiel. Als Anwendung sei ein Beispiel angeführt, bei dem es darum geht, die effektive Wärmeleitfähigkeit einer Gießform für die Stahlschmelze zu bestimmen. Bei dem in [47] implementierten Verfahren zur Lösung dieses Problems ist es erforderlich, gemessene Temperaturwerte im Innern der Gießform numerisch zu differenzieren. Die Meßdaten sind als Kreise in dem linken Bild aus Abbildung 49.2 in ℃ über der Zeit [1000 s] eingezeichnet.

Die zur Verfügung gestellten Meßdaten erweisen sich allesamt als Vielfache von 5 ℃, die Daten liegen also allenfalls im Rahmen dieser Genauigkeit. Daher bietet sich für die Meßungenauigkeit der Wert $\delta = 2.5$ an. Abbildung 49.2 (links) zeigt den resultierenden geglätteten kubischen Spline. Für dessen Berechnung werden die Daten zunächst durch Subtraktion der linearen Funktion

$$\ell(x) = y_0 + \frac{y_l - y_0}{x_l - x_0}(x - x_0)$$

auf homogene Randwerte transformiert.

Das rechte Bild der Abbildung zeigt zwei Approximationen der Ableitung: die Treppenfunktion mit den Werten der Differenzenquotienten (dies entspricht der Ableitung des interpolierenden linearen Splines) und die Ableitung des geglätteten kubischen Splines. In diesem Beispiel ist die Gitterweite h bereits so klein, daß an manchen Stellen eine deutliche Fehlerverstärkung in der numerischen Ableitung des linearen Splines zu erkennen ist. Für die spezielle Anwendung ist die Treppenfunktion jedoch vor allem aufgrund ihrer fehlenden Glattheit unbrauchbar. ◊

Aufgaben

1. Seien $y_i = f(x_i)$, $i = 0, \ldots, l$, mit $f \in C^2[a,b]$ die Interpolationsvorgaben über einem Gitter $\triangle \subset [a,b]$ mit Gitterweite h. Zeigen Sie, daß der zugehörige lineare interpolierende Spline s in jedem Punkt $x \in [a,b]$ die Abschätzung

$$|f(x) - s(x)| \;\le\; \frac{h^2}{8}\, \|f''\|_{[a,b]}$$

erfüllt. Vergleichen Sie mit Satz 45.4.

2. Sei $\triangle \subset [a,b]$ ein äquidistantes Gitter und $s \in S_{1,\triangle}$ die Bestapproximation an eine Funktion $f \in C[a,b]$ bezüglich der $\mathcal{L}^2$-Norm. Zeigen Sie, daß s dann auch eine gute Approximation bezüglich der Maximumnorm ist, genauer daß

$$\|f - s\|_{[a,b]} \;\le\; 4\, \|f - \tilde{s}\|_{[a,b]}$$

für alle $\tilde{s} \in S_{1,\triangle}$.
Hinweis: Zeigen Sie zunächst mit Hilfe von Satz 31.10, daß die Bestapproximation $s_{\mathcal{L}} \in S_{1,\triangle}$ bezüglich $\mathcal{L}^2$ an eine Funktion $g \in C[a,b]$ die Ungleichung $\|s_{\mathcal{L}}\|_{[a,b]} \le 3\|g\|_{[a,b]}$ erfüllt. Wenden Sie dieses Hilfsresultat dann auf die Funktion $g = f - \tilde{s}$ an.

3. Entwickeln Sie mit den Methoden aus Abschnitt 46 einen Algorithmus, um vorgebene Daten y_i, $i = 0, \ldots, l$, über einem Gitter $\triangle = \{x_0 < x_1 < \ldots < x_l\}$ mit einem quadratischen Spline $s \in S_{2,\triangle}$ zu interpolieren. Geben Sie gegebenenfalls noch zusätzliche Randbedingungen an die Ableitung vor, damit s eindeutig bestimmt ist.
Implementieren Sie ihren Algorithmus und vergleichen Sie den quadratischen Spline mit dem interpolierenden kubischen Spline, der die gleichen Randableitungen besitzt.

4. Bestimmen Sie den Kern der Matrix auf der rechten Seite von (46.5). Interpretieren Sie das Resultat.

5. Berechnen Sie zu einem Gitter $\triangle \subset [a,b]$ eine „Lagrange-artige“ Basis von $S_{3,\triangle}$ bezüglich der vollständigen Interpolationsaufgabe. Wie sind die Randbedingungen in diesem Fall einzuarbeiten? Plotten Sie die berechneten Basisfunktionen und vergleichen Sie mit Abbildung 46.3.

6. Sei $\triangle = \{a = x_0 < \ldots < x_l = b\}$ ein äquidistantes Gitter und x_i ein Punkt dieses Gitters mit $2 \le i \le l-2$. Weisen Sie nach, daß es genau einen kubischen Spline s mit $s(x_i) = 1$ gibt, der in $[a,b] \setminus [x_{i-2}, x_{i+2}]$ verschwindet. Zeigen Sie, daß s in (x_{i-2}, x_{i+2}) positiv ist.

7. Sei $B_0 = \chi_{[-1/2,1/2]}$ die charakteristische Funktion des Intervalls $[-1/2, 1/2]$ und B_m, $m \in \mathbb{N}$, rekursiv definiert durch

$$B_m(x) \;=\; \int_{\mathbb{R}} \chi_{[-1/2,1/2]}(x-y) B_{m-1}(y)\, dy \;=\; \int_{x-1/2}^{x+1/2} B_{m-1}(y)\, dy.$$

(a) Zeigen Sie, daß B_m, $m \in \mathbb{N}$, nichtnegativ und $(m-1)$-mal stetig differenzierbar ist. Beweisen Sie außerdem, daß B_m außerhalb des Intervalls $[-(m+1)/2, (m+1)/2]$ verschwindet und im Fall m gerade auf jedem Teilintervall $[k-1/2, k+1/2]$ bzw. im Fall m ungerade auf jedem Teilintervall $[k, k+1]$, $k \in \mathbb{Z}$, ein Polynom vom Grad m ist.

(b) Berechnen und plotten Sie die Funktionen B_1, B_2 und B_3.
(c) Zeigen Sie, daß die verschobenen Funktionen $B_m(\cdot - k)$, $k = 0, \ldots, m$, in dem Intervall $[(m-1)/2, (m+1)/2]$ linear unabhängig sind.
(d) Sei $\triangle = \{c = x_0 < \ldots < x_n = d\}$ ein äquidistantes Gitter über $[c, d]$ mit Gitterweite h und sei m ungerade, $m = 2l - 1$ für ein $l \in \mathbb{N}$. Zeigen Sie, daß die *B-Splines*

$$B_{m,k}(x) \;=\; B_m\Big(\frac{x - c - hk}{h}\Big)\,, \qquad x \in [c, d]\,,$$

für $k = -l + 1, \ldots, n + l - 1$, eine Basis von $S_{m,\triangle}$ bilden.

8. Die Funktion $f \in C^2(\mathbb{R})$ sei $(b-a)$-periodisch und $\triangle$ ein Gitter über $[a, b]$.
(a) Zeigen Sie, daß es genau einen kubischen Spline $s \in S_{3,\triangle}$ gibt, der die Funktion f über $\triangle$ interpoliert und zu einer zweimal stetig differenzierbaren $(b-a)$-periodischen Funktion über $\mathbb{R}$ fortgesetzt werden kann. Geben Sie einen Algorithmus mit Aufwand $O(l)$ zur Berechnung der Koeffizienten dieses Splines an.
(b) Beweisen Sie für diesen Spline ein Analogon von Korollar 47.3, nämlich

$$\|f'' - s''\|_{\mathcal{L}^2(a,b)} \;\le\; \|f'' - \tilde{s}\|_{\mathcal{L}^2(a,b)}$$

für alle linearen Splines $\tilde{s} \in S_{1,\triangle}$ mit $\tilde{s}(a) = \tilde{s}(b)$.
(c) Zeigen Sie, daß für diesen Spline ebenfalls die Fehlerabschätzung aus Satz 47.5 gilt.

9. (a) Überlegen Sie sich, warum Satz 47.5 nicht für den interpolierenden natürlichen kubischen Spline an eine beliebig vorgegebene Funktion $f \in H^4(a, b)$ gilt.
(b) Implementieren Sie den Algorithmus zur Berechnung des interpolierenden natürlichen kubischen Splines an eine Funktion f über einem äquidistanten Gitter $\triangle \subset [a, b]$ mit Gitterweite h. Betrachten Sie speziell die Funktion $f(x) = e^x$ auf dem Intervall $[-1, 1]$ und plotten Sie für $h = 2^{-j}$, $j = 1, 2, \ldots$, den Interpolationsfehler in der Maximumnorm.

10. Sei $f \in C^4[a, b]$, $\triangle \subset [a, b]$ ein äquidistantes Gitter mit Gitterweite h und s die Lösung der vollständigen Interpolationsaufgabe. Beweisen Sie die Fehlerabschätzung

$$\|f - s\|_{[a,b]} \;\le\; \frac{h^4}{16}\,\|f^{(4)}\|_{[a,b]}$$

bezüglich der Maximumnorm.
(a) Zeigen Sie zunächst mit Hilfe von Aufgabe 2, daß

$$\|f'' - s''\|_{[a,b]} \;\le\; \frac{h^2}{2}\,\|f^{(4)}\|_{[a,b]}\,.$$

(b) Die gesuchte Abschätzung folgt dann aus Aufgabe 1. Wie?

11. Beweisen Sie, daß der kubische Spline f_* aus Satz 48.2 das Lagrange-Funktional

$$\|f''\|^2_{\mathcal{L}^2(a,b)} \;+\; \lambda \sum_{i=1}^{l-1} |y_i - f(x_i)|^2$$

unter allen Funktionen $f \in H^2(a, b)$ minimiert.

12. Sei $G = LL^*$ die Cholesky-Zerlegung der Gramschen Matrix G in (48.6). Zeigen Sie, daß der Momentenvektor c aus (48.8) das lineare Ausgleichsproblem

$$\text{minimiere} \quad \left\| \begin{bmatrix} L^* \\ \sqrt{\alpha}\, T \end{bmatrix} c \; - \; \begin{bmatrix} 0 \\ y/\sqrt{\alpha} \end{bmatrix} \right\|_2$$

löst.

13. Überführen Sie die Funktion r aus (48.10) in die Gestalt (18.10) der in Abschnitt 18.3 betrachteten Funktionenklasse. Welche Bedeutung haben dabei die Parameter n, sowie die Koeffizienten d_i und z_i, $i = 1, \ldots, n$?

IX Fourierreihen

Nach der Interpolation durch Splines wenden wir uns nun der Interpolation und Approximation einer (komplexwertigen) Funktion einer Veränderlichen durch *trigonometrische Polynome* zu. In den Anwendungen werden trigonometrische Polynome häufig verwendet, da die zugehörigen Entwicklungskoeffizienten mit der schnellen Fouriertransformation (FFT) sehr effizient berechnet werden können. Für die zugehörigen Fehlerabschätzungen führen wir eine Skala periodischer *Sobolevräume* über einem reellen Intervall ein.

Dieses Kapitel reicht in ein sehr weites Gebiet der Analysis hinein, in welches das Buch von Zygmund [110] einen umfassenden Einblick gibt. Eine selektivere Darstellung aus dem Blickwinkel der numerischen Mathematik findet sich etwa in dem Buch von Henrici [50, Kapitel 13]. Für die hier vorgestellten periodischen Sobolevräume sei schließlich noch auf das Buch von Kreß [64] verwiesen.

50 Trigonometrische Polynome

Wir beginnen mit der Definition eines trigonometrischen Polynoms.

Definition 50.1. Ein (komplexes) *trigonometrisches Polynom* vom Grad n ist eine Funktion der Form

$$t(\theta) = \sum_{k=-n}^{n} \alpha_k e^{\mathrm{i}k\theta}\,, \qquad \alpha_k \in \mathbb{C}\,. \tag{50.1}$$

Die Menge aller trigonometrischen Polynome vom Grad n wird im folgenden mit $\mathcal{T}_n$ bezeichnet.

Wegen ihrer Periodizität verwendet man trigonometrische Polynome vor allem zur Approximation *periodischer* Funktionen. Periodische Funktionen treten bei Anwendungen zum Beispiel auf, wenn Funktionen über einer geschlossenen Kurve im $\mathbb{R}^n$ als Funktion der Bogenlänge parametrisiert werden.

Bemerkung 50.2. Wenn $\alpha_k = \overline{\alpha_{-k}}$ für alle $k = 0, \dots, n$, dann nimmt t über $\mathbb{R}$ nur reelle Werte an. In diesem Fall kann t auch als reelles trigonometrisches Polynom dargestellt werden, nämlich

$$t(\theta) \;=\; \frac{a_0}{2} \;+\; \sum_{k=1}^{n} \bigl(a_k \cos k\theta + b_k \sin k\theta\bigr)$$

mit $a_0 = 2\alpha_0$ und

$$a_k = 2\,\mathrm{Re}\,\alpha_k\,, \quad b_k = -2\,\mathrm{Im}\,\alpha_k\,, \qquad k = 1, \dots, n\,. \qquad \Diamond$$

Wir wollen nun eine gegebene Funktion f über $[0, 2\pi]$ durch trigonometrische Polynome approximieren. Aufgrund der Ergebnisse aus Abschnitt 31 und wegen des folgenden Resultats bietet sich dabei die $\mathcal{L}^2$-Norm als „Gütemaß“ an:

Proposition 50.3. *Die Funktionen*

$$\frac{1}{\sqrt{2\pi}}\, e^{\mathrm{i}k\theta}\,, \qquad k = -n, \dots, n\,, \tag{50.2}$$

bilden eine Orthonormalbasis von $\mathcal{T}_n$ bezüglich $\mathcal{L}^2(0, 2\pi)$ und die $\mathcal{L}^2(0, 2\pi)$-Bestapproximation aus $\mathcal{T}_n$ an f hat die Form (50.1) *mit*

$$\alpha_k = \frac{1}{2\pi} \int_0^{2\pi} f(\theta) e^{-\mathrm{i}k\theta}\, d\theta\,, \qquad k = -n, \dots, n\,. \tag{50.3}$$

Beweis. Für $-n \le j, k \le n$ gilt

$$\Bigl\langle \frac{1}{\sqrt{2\pi}}\, e^{\mathrm{i}k\theta}, \frac{1}{\sqrt{2\pi}}\, e^{\mathrm{i}j\theta} \Bigr\rangle \;=\; \frac{1}{2\pi} \int_0^{2\pi} e^{-\mathrm{i}k\theta} e^{\mathrm{i}j\theta}\, d\theta$$

$$= \begin{cases} \dfrac{1}{2\pi} \displaystyle\int_0^{2\pi} d\theta \;=\; 1\,, & k = j\,, \\[2ex] \dfrac{1}{2\pi}\, \dfrac{1}{\mathrm{i}(j-k)}\, e^{\mathrm{i}(j-k)\theta}\Big|_0^{2\pi} \;=\; 0\,, & k \ne j\,. \end{cases}$$

Die Koeffizienten der Bestapproximation ergeben sich daher aus Satz 31.6 (c). □

Man beachte, daß die sogenannten *Fourierkoeffizienten* α_k nicht von n abhängen. Daher drängt sich die Frage auf, ob für $n \to \infty$ in einem geeigneten Sinn

$$f(\theta) \sim \sum_{k=-\infty}^{\infty} \alpha_k e^{\mathrm{i}k\theta} \tag{50.4}$$

Gültigkeit besitzt. Wir verwenden die Notation $\sim$, da zunächst weder klar ist, ob die rechte Seite konvergiert noch ob ihr Wert im Konvergenzfall mit $f(\theta)$ übereinstimmt. Daher wird die rechte Seite von (50.4) als *formale Fourierreihe* von f bezeichnet. Die Untersuchung ihrer Konvergenzeigenschaften ist Gegenstand der *Fourieranalyse*. Hier wollen wir zunächst nur einige wichtige Resultate zitieren:

Satz 50.4. (*a*) *Ist $f \in \mathcal{L}^2(0,2\pi)$ in einer Umgebung von $\theta \in (0,2\pi)$ von beschränkter Variation, dann konvergiert die Fourierreihe an der Stelle θ gegen $\big(f(\theta+) + f(\theta-)\big)/2$. Für $\theta = 0$ und $\theta = 2\pi$ gilt ein entsprechendes Resultat mit Grenzwert $\big(f(0+) + f(2\pi-)\big)/2$.*

(*b*) *Ist f zudem stetig und stückweise stetig differenzierbar mit $f(0) = f(2\pi)$, dann konvergiert die Fourierreihe gleichmäßig gegen f.*

(*c*) *Konvergiert umgekehrt die Fourierreihe von f gleichmäßig, so stimmt die Grenzfunktion bis auf eine Nullmenge mit f überein.*

Für die ersten beiden Aussagen vergleiche man die Sätze 136.1 und 137.2 aus dem Buch von Heuser [53]. Die letzte Aussage folgt aus der allgemeineren $\mathcal{L}^2$-Konvergenztheorie der Fourierreihen, vgl. [53, Abschnitt 141]. Wir beweisen nun den folgenden Zusammenhang zwischen der $\mathcal{L}^2$-Norm und den Fourierkoeffizenten von f:

Proposition 50.5. *Die Funktion $f \in \mathcal{L}^2(0,2\pi)$ besitze die formale Fourierreihe* (50.4).

(*a*) *Dann ist*

$$\sum_{k=-\infty}^{\infty} |\alpha_k|^2 \;\le\; \frac{1}{2\pi}\,\|f\|^2_{\mathcal{L}^2(0,2\pi)}\,. \tag{50.5}$$

(*b*) *Konvergiert die Fourierreihe von f gleichmäßig in $[0,2\pi]$, dann gilt Gleichheit in* (50.5).

Beweis. (*a*) Die Entwicklungskoeffizenten von f bezüglich der Orthonormalbasis (50.2) sind durch $\sqrt{2\pi}\,\alpha_k$ für $-n \le k \le n$ gegeben. Nach der Besselschen Ungleichung, Satz 31.6 (*d*), gilt daher

$$2\pi \sum_{k=-n}^{n} |\alpha_k|^2 \;\le\; \|f\|^2_{\mathcal{L}^2(0,2\pi)}\,.$$

Da n beliebig gewählt war und die Fourierkoeffizenten nicht von n abhängen, folgt (50.5).

(*b*) Im Falle gleichmäßiger Konvergenz der Fourierreihe gilt

$$\int_0^{2\pi} |f(\theta)|^2 \, d\theta \;=\; \lim_{n\to\infty} \int_0^{2\pi} \Big| \sum_{k=-n}^{n} \alpha_k e^{\mathrm{i}k\theta} \Big|^2 \, d\theta \,.$$

Wegen der Orthogonalität der Basisfunktionen von $\mathcal{T}_n$ ergibt das Integral auf der rechten Seite $\sum_{k=-n}^{n} 2\pi\,|\alpha_k|^2$, und wegen (50.5) konvergiert dies für $n \to \infty$ gegen die unendliche Reihe, also gilt

$$\|f\|^2_{\mathcal{L}^2(0,2\pi)} \;=\; 2\pi \sum_{k=-\infty}^{\infty} |\alpha_k|^2 \,. \qquad \square$$

Bemerkung 50.6. Es sei an dieser Stelle festgehalten, vgl. etwa [64, S. 126], daß für *jede* Funktion $f \in \mathcal{L}^2(0, 2\pi)$ Gleichheit in (50.5) gilt und $\mathcal{L}^2$-Funktionen gerade dadurch charakterisiert werden können, daß die Reihe auf der linken Seite von (50.5) konvergiert. $\Diamond$

Beispiel 50.7. Als Beispiel betrachten wir die charakteristische Funktion $\chi_{[a,b]}$ eines Intervalls $[a, b] \subsetneq [0, 2\pi]$. Wir setzen $c = (a+b)/2$ und $d = (b-a)/2 < \pi$. Gemäß (50.3) gilt für die Fourierkoeffizienten die Formel

$$\alpha_k \;=\; \frac{1}{2\pi} \int_a^b e^{-\mathrm{i}k\theta} \, d\theta \,, \qquad k \in \mathbb{Z} \,,$$

d. h. es ist $\alpha_0 = (b-a)/(2\pi) = d/\pi$ und für $k \neq 0$ ergibt sich

$$\begin{aligned} \alpha_k \;&=\; \frac{1}{2\pi} \, \frac{1}{-\mathrm{i}k} \, e^{-\mathrm{i}k\theta} \Big|_a^b \;=\; -\frac{1}{2k\pi\mathrm{i}} \, e^{-\mathrm{i}kc} \big(e^{-\mathrm{i}kd} - e^{\mathrm{i}kd} \big) \\ &=\; -\frac{1}{2k\pi\mathrm{i}} \, e^{-\mathrm{i}kc} \big(\cos kd - \mathrm{i}\sin kd - \cos kd - \mathrm{i}\sin kd \big) \\ &=\; \frac{1}{\pi} \, e^{-\mathrm{i}kc} \, \frac{\sin kd}{k} \,. \end{aligned}$$

Somit hat die formale Fourierreihe von $\chi_{[a,b]}$ die Gestalt

$$\chi_{[a,b]}(\theta) \;\sim\; \frac{d}{\pi} \;+\; \frac{1}{\pi} \sum_{|k|=1}^{\infty} e^{-\mathrm{i}kc} \, \frac{\sin kd}{k} \, e^{\mathrm{i}k\theta} \,.$$

Werden die Terme für $\pm k$ zusammengefaßt, ergibt sich die rein reelle Darstellung

$$\begin{aligned} \chi_{[a,b]}(\theta) \;&\sim\; \frac{d}{\pi} \;+\; \frac{1}{\pi} \sum_{k=1}^{\infty} \frac{\sin kd}{k} \, \big(e^{\mathrm{i}k(\theta-c)} + e^{-\mathrm{i}k(\theta-c)} \big) \\ &=\; \frac{d}{\pi} \;+\; \frac{2}{\pi} \sum_{k=1}^{\infty} \frac{1}{k} \, \sin kd \, \cos k(\theta - c) \,. \end{aligned}$$

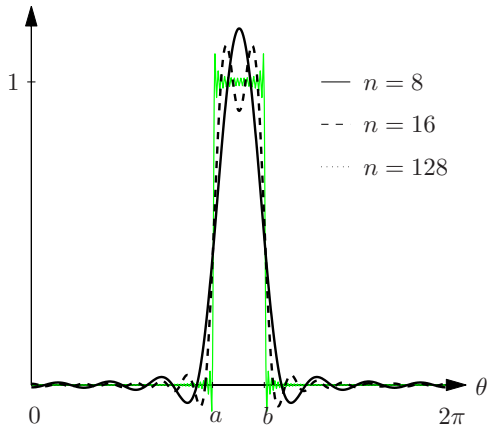


Abb. 50.1: Konvergenz der Fourierreihe einer charakteristischen Funktion

Bezeichnen wir mit t_n die Bestapproximation aus $\mathcal{T}_n$ an $\chi_{[a,b]}$, dann gilt für den Fehler $\chi_{[a,b]} - t_n$ gemäß Proposition 50.5

$$\|\chi_{[a,b]} - t_n\|^2_{\mathcal{L}^2(0,2\pi)} \;\geq\; 2\pi \sum_{|k|=n+1}^{\infty} |\alpha_k|^2 \;=\; \frac{2}{\pi} \sum_{|k|=n+1}^{\infty} \frac{\sin^2 kd}{k^2} \,.$$

Da $d < \pi$ ist, verhält sich die Summe auf der rechten Seite wie

$$\sum_{|k|=n+1}^{\infty} k^{-2} \;\approx\; 2 \int_n^{\infty} t^{-2}\, dt \;=\; 2n^{-1} \,,$$

das heißt es existiert ein $\varepsilon > 0$ mit

$$\|\chi_{[a,b]} - t_n\|_{\mathcal{L}^2(0,2\pi)} \;\geq\; \varepsilon n^{-1/2} \,, \qquad n \in \mathbb{N} \,. \tag{50.6}$$

Im Hinblick auf Bemerkung 50.6 verhält sich der Approximationsfehler asymptotisch genau wie $n^{-1/2}$.

Abbildung 50.1 illustriert exemplarisch das Konvergenzverhalten dieser Fourierreihe. Dargestellt sind die Polynome t_n mit $n = 8$, 16 und 128. ◊

51 Sobolevräume

In Abschnitt 31 hatten wir den Raum $H^1(0, 2\pi)$ eingeführt. Wir wollen uns im folgenden auf die Teilmenge $H^1_\pi(0, 2\pi)$ der 2π-periodischen Funktionen in $H^1(0, 2\pi)$ beschränken, also auf Funktionen $f \in H^1(0, 2\pi)$ mit $f(0) = f(2\pi)$.

Eine Funktion $f \in H^1_\pi(0,2\pi)$ besitzt nach Voraussetzung eine schwache Ableitung

$$f'(\theta) \sim \sum_{k=-\infty}^{\infty} \beta_k e^{\mathrm{i}k\theta} \qquad \text{mit} \qquad f(\theta) = f(0) + \int_0^\theta f'(t)\,dt\,. \tag{51.1}$$

Wegen der 2π-Periodizität von f kommen dabei nicht alle möglichen $\mathcal{L}^2(0,2\pi)$-Funktionen f' in Frage, denn es muß

$$f(0) = f(2\pi) = f(0) + \int_0^{2\pi} f'(t)\,dt$$

gelten. Folglich ist

$$\beta_0 = \frac{1}{2\pi}\int_0^{2\pi} f'(\theta)\,d\theta = 0\,. \tag{51.2}$$

Ausgehend von (51.1) bietet es sich an, die Fourierreihe von f durch gliedweise Integration der Fourierreihe von f' zu bestimmen. Zunächst ist diese Vorgehensweise jedoch keineswegs gerechtfertigt, da die Konvergenzeigenschaften der beiden Fourierreihen nicht geklärt sind. Dennoch führt dieser Ansatz zum Ziel, wie das folgende Resultat bestätigt.

Lemma 51.1. *Sei $f \in H^1_\pi(0,2\pi)$ und f' durch* (51.1), (51.2) *definiert. Dann sind für $k \neq 0$ die Fourierkoeffizienten von f durch $\alpha_k = \beta_k/(\mathrm{i}k)$ gegeben.*

Beweis. Nach Proposition 50.3 berechnen sich die Fourierkoeffizienten von f aus der Formel

$$\alpha_k = \frac{1}{2\pi}\int_0^{2\pi} f(\theta) e^{-\mathrm{i}k\theta}\,d\theta\,.$$

Durch Einsetzen von (51.1) ergibt sich für $k \neq 0$

$$\begin{aligned}
\alpha_k &= \frac{1}{2\pi}\int_0^{2\pi} \Big(f(0) + \int_0^\theta f'(t)\,dt\Big) e^{-\mathrm{i}k\theta}\,d\theta \\
&= \frac{f(0)}{2\pi}\int_0^{2\pi} e^{-\mathrm{i}k\theta}\,d\theta \;+\; \frac{1}{2\pi}\int_0^{2\pi} f'(t)\int_t^{2\pi} e^{-\mathrm{i}k\theta}\,d\theta\,dt \\
&= \frac{1}{2\pi}\int_0^{2\pi} f'(t)\,\frac{1}{-\mathrm{i}k}\,e^{-\mathrm{i}k\theta}\Big|_t^{2\pi} dt \\
&= -\frac{1}{\mathrm{i}k}\,\frac{1}{2\pi}\Big(\int_0^{2\pi} f'(t)\,dt \;-\; \int_0^{2\pi} f'(t) e^{-\mathrm{i}kt}\,dt\Big) \\
&= -\frac{1}{\mathrm{i}k}\,(\beta_0 - \beta_k) \overset{(51.2)}{=} \frac{1}{\mathrm{i}k}\,\beta_k\,.
\end{aligned}$$

□

Als Konsequenz aus diesem Hilfsatz erhalten wir

Satz 51.2. *Eine Funktion $f \in \mathcal{L}^2(0,2\pi)$ mit $f(\theta) \sim \sum_{k=-\infty}^{\infty} \alpha_k e^{\mathrm{i}k\theta}$ gehört genau dann zu $H^1_\pi(0,2\pi)$, falls*

$$\sum_{k=-\infty}^{\infty} k^2 |\alpha_k|^2 \,<\, \infty\,.$$

In diesem Fall ist

$$\sum_{k=-\infty}^{\infty} (k^2+1)|\alpha_k|^2 \;=\; \frac{1}{2\pi}\,\|f\|^2_{H^1(0,2\pi)}\;.$$

Beweis. Die eine Beweisrichtung folgt unmittelbar aus dem vorangegangenen Lemma 51.1: Für eine Funktion $f \in H^1_\pi(0,2\pi)$ sind $\mathrm{i}k\alpha_k$ die Fourierkoeffizienten der schwachen Ableitung $f' \in \mathcal{L}^2(0,2\pi)$. Also konvergiert nach Proposition 50.5 die unendliche Reihe $\sum k^2|\alpha_k|^2$ und wegen Bemerkung 50.6 ist

$$\begin{aligned}\sum_{k=-\infty}^{\infty} (k^2+1)|\alpha_k|^2 \;&=\; \sum_{k=-\infty}^{\infty} k^2|\alpha_k|^2 \,+\, \sum_{k=-\infty}^{\infty} |\alpha_k|^2 \\ &=\; \frac{1}{2\pi}\Big(\|f'\|^2_{\mathcal{L}^2(0,2\pi)} \,+\, \|f\|^2_{\mathcal{L}^2(0,2\pi)}\Big)\;.\end{aligned}$$

Konvergiert umgekehrt die Reihe $\sum_{k=-\infty}^{\infty} k^2|\alpha_k|^2$, dann definiert nach Bemerkung 50.6 die formale Fourierreihe $\sum_{k=-\infty}^{\infty} \mathrm{i}k\alpha_k e^{\mathrm{i}k\theta}$ eine Funktion g aus $\mathcal{L}^2(0,2\pi)$. Zu g existiert eine Stammfunktion $G \in H^1_\pi(0,2\pi)$, die wie in (31.1) definiert ist, wobei die Konstante c so zu wählen ist, daß $\int_0^{2\pi} G(\theta)d\theta = 2\pi\alpha_0$ gilt. Nach Lemma 51.1 haben G und f dann die gleichen Fourierkoeffizienten, und daher ist $f = G \in H^1_\pi(0,2\pi)$. □

Beispiel 51.3. In Beispiel 50.7 haben wir die formale Fourierreihe der charakteristischen Funktion eines Intervalls $[a,b] \subsetneq [0,2\pi]$ bestimmt. Demnach ist für $0 \le x_0 < x_1 < x_2 \le 2\pi$

$$\begin{aligned}f(\theta) \;&:=\; \frac{1}{x_1-x_0}\,\chi_{[x_0,x_1]} \,-\, \frac{1}{x_2-x_1}\,\chi_{[x_1,x_2]} \\ &\sim\; \sum_{|k|=1}^{\infty} \Big(\frac{1}{2d_1\pi}\,e^{-\mathrm{i}kc_1}\,\frac{\sin kd_1}{k} \,-\, \frac{1}{2d_2\pi}\,e^{-\mathrm{i}kc_2}\,\frac{\sin kd_2}{k}\Big)\,e^{\mathrm{i}k\theta}\,,\end{aligned}$$

wobei $c_1 = (x_0+x_1)/2$, $c_2 = (x_1+x_2)/2$, $d_1 = (x_1-x_0)/2$ und $d_2 = (x_2-x_1)/2$ die jeweiligen Intervallmittelpunkte bzw. -radien sind. Nach Beispiel 31.4 ist f die schwache Ableitung der Hutfunktion $\Lambda \in H^1_\pi(0,2\pi)$ mit

$$\Lambda(\theta) \;=\; \begin{cases} (\theta - x_0)/(x_1 - x_0)\,, & x_0 \le \theta < x_1\,, \\ (\theta - x_2)/(x_1 - x_2)\,, & x_1 \le \theta < x_2\,, \\ 0\,, & \text{sonst}\,. \end{cases}$$

Folglich hat Λ nach Lemma 51.1 die (gleichmäßig konvergente) Fourierreihe

$$\Lambda(\theta) \;=\; \alpha_0 + \sum_{|k|=1}^{\infty} \Big(\frac{1}{2d_1\pi}\, e^{-\mathrm{i}kc_1} \sin kd_1 - \frac{1}{2d_2\pi}\, e^{-\mathrm{i}kc_2} \sin kd_2 \Big) \frac{1}{\mathrm{i}k^2}\, e^{\mathrm{i}k\theta}\,. \qquad (51.3)$$

Dabei ist $\alpha_0 = \frac{1}{2\pi}\int_0^{2\pi} \Lambda(\theta)\,d\theta = (d_1 + d_2)/(2\pi)$. ◊

Ausgehend von Satz 51.2 können wir nun eine sogenannte *Skala* von Funktionenräumen einführen.

Definition und Satz 51.4. *Für $s > 0$ ist*

$$H^s_\pi(0,2\pi) \;=\; \{\, f(\theta) \sim \sum_{k=-\infty}^{\infty} \alpha_k e^{\mathrm{i}k\theta} \;:\; \sum_{k=-\infty}^{\infty} |k|^{2s} |\alpha_k|^2 < \infty \,\}$$

der (periodische) Sobolevraum *der Ordnung s. Das zugehörige Innenprodukt für $f \sim \sum \alpha_k e^{\mathrm{i}k\theta}$ und $g \sim \sum \beta_k e^{\mathrm{i}k\theta}$ aus $H^s_\pi(0,2\pi)$ ist definiert durch*

$$\langle\, f, g \,\rangle_{H^s_\pi(0,2\pi)} \;=\; 2\pi \sum_{k=-\infty}^{\infty} (|k|^{2s} + 1)\, \overline{\alpha}_k \beta_k\,.$$

Wir verzichten auf den Nachweis, daß $\langle\, \cdot\,, \cdot \,\rangle_{H^s_\pi(0,2\pi)}$ tatsächlich ein Innenprodukt definiert. Statt dessen halten wir die folgenden Eigenschaften dieser (periodischen) Sobolevräume fest:

• Für $r > s > 0$ gilt $H^r_\pi(0,2\pi) \subset H^s_\pi(0,2\pi)$. Dies folgt unmittelbar aus der Definition 51.4 mit dem Majorantenkriterium.

• Für $s > 1$ ist $H^s_\pi(0,2\pi) \subset H^1_\pi(0,2\pi)$ und daher hat jede Funktion $f \in H^s_\pi(0,2\pi)$ mit $s > 1$ eine schwache Ableitung f'. Aus Lemma 51.1 folgt zudem unmittelbar $f' \in H^{s-1}_\pi(0,2\pi)$. Insbesondere hat also eine Funktion $f \in H^s_\pi(0,2\pi)$, $s \in \mathbb{N}$, s schwache Ableitungen $f', f'', \ldots, f^{(s)}$, wobei $f, \ldots, f^{(s-1)}$ stetig und 2π-periodisch sind und $f^{(s)}$ noch zu $\mathcal{L}^2(0,2\pi)$ gehört.

Sobolevräume sind aus der modernen angewandten und numerischen Mathematik kaum mehr wegzudenken und haben inzwischen vielfach die Rolle der Funktionenräume C^s (mit $s \in \mathbb{N}$) übernommen. In beiden Fällen charakterisiert der Index s eine gewisse „Glattheit“ entsprechender Funktionen $f \in H^s_\pi$

bzw. $f \in C^s$, die allerdings für die beiden Funktionenräume nicht äquivalent ist. So gehören 2π-periodische Funktionen $f \in C^1(\mathbb{R})$ immer auch zu $H^1_\pi(0,2\pi)$ und entsprechend gehören 2π-periodische Funktionen $f \in C^s(\mathbb{R})$ mit $s \in \mathbb{N}$ immer auch zu $H^s_\pi(0,2\pi)$. Die Umkehrung ist allerdings falsch, wie das Beispiel der Hutfunktion belegt: Die Funktion Λ aus Beispiel 51.3 ist nicht stetig differenzierbar, gehört aber zu $H^1_\pi(0,2\pi)$. Sie gehört sogar zu $H^s_\pi(0,2\pi)$ für jedes $s < 3/2$, denn mit den Fourierkoeffizienten aus (51.3) gilt

$$\sum_{k=-\infty}^{\infty} |k|^{2s} |\alpha_k|^2 \;\le\; \frac{d_1+d_2}{d_1 d_2 \pi} \sum_{k=1}^{\infty} |k|^{2s-4} \;<\; \infty\,.$$

für $s < 3/2$.

Um ein Gefühl für die „Glattheit“ einer Funktion $f \in H^s_\pi(0,2\pi)$ zu bekommen, beweisen wir abschließend noch die folgenden beiden Aussagen:

Satz 51.5. *Für $s > 1/2$ konvergiert die Fourierreihe einer Funktion $f \in H^s_\pi(0,2\pi)$ gleichmäßig in $[0,2\pi]$, d. h. alle Funktionen $f \in H^s_\pi(0,2\pi)$ sind stetig und 2π-periodisch. Speziell für $s = 1$ sind die Funktionen $f \in H^1_\pi(0,2\pi)$ sogar Hölder-stetig mit Hölder-Exponenten $\alpha = 1/2$. Genauer gilt*

$$|f(\theta) - f(t)| \;\le\; \|f\|_{H^1_\pi(0,2\pi)}\, |\theta - t|^{1/2}$$

für alle $\theta, t \in [0,2\pi]$ und alle $f \in H^1_\pi(0,2\pi)$.

Beweis. Die gleichmäßige Konvergenz der Fourierreihe für eine Funktion $f \in H^s_\pi(0,2\pi)$ mit $s > 1/2$ folgt unmittelbar aus der Cauchy-Schwarz-Ungleichung: Für $m > n$ gilt nämlich

$$\begin{aligned}\sum_{|k|=n}^{m} |\alpha_k e^{\mathrm{i}k\theta}| \;&=\; \sum_{|k|=n}^{m} |k|^{-s}\big(|k|^s|\alpha_k|\big) \le \Big(\sum_{|k|=n}^{m} |k|^{-2s} \sum_{|k|=n}^{m} |k|^{2s}|\alpha_k|^2\Big)^{1/2} \\ &\le\; \frac{1}{\sqrt{\pi}}\, \|f\|_{H^s_\pi(0,2\pi)} \Big(\sum_{k=n}^{\infty} |k|^{-2s}\Big)^{1/2},\end{aligned}$$

und somit folgt die gleichmäßige Konvergenz aus dem entsprechenden Cauchy-Kriterium.

Für $s = 1$ ergibt sich für $0 \le t < \theta \le 2\pi$ (wiederum mit der Cauchy-Schwarz-Ungleichung, diesmal in $\mathcal{L}^2$) aus der Definition einer Funktion $f \in H^1(0,2\pi)$, daß

$$\begin{aligned}|f(\theta) - f(t)| \;&=\; \Big|\int_t^\theta f'(\tau)\,d\tau\Big| \;\le\; \Big(\int_t^\theta d\tau\Big)^{1/2} \Big(\int_t^\theta |f'(\tau)|^2\,d\tau\Big)^{1/2} \\ &\le\; |\theta - t|^{1/2}\, \|f\|_{H^1_\pi(0,2\pi)}\,.\end{aligned}$$

□

Die Schärfe dieses Satzes macht man sich unmittelbar an den folgenden beiden Beispielen klar.

Beispiele. Die charakteristische Funktion χ eines Intervalls $[a,b] \subsetneq [0,2\pi]$ ist nicht stetig, gehört aber nach Beispiel 50.7 und Definition 51.4 zu allen $H^s_\pi(0,2\pi)$-Räumen mit $s < 1/2$.

Die $\mathcal{L}^2$-Funktionen $f_\alpha(\theta) = \big(\theta(2\pi-\theta)\big)^\alpha$ mit $0 < \alpha < 1$ haben in $(0,2\pi)$ die klassischen Ableitungen

$$f'_\alpha(\theta) = 2\alpha\big(\theta(2\pi-\theta)\big)^{\alpha-1}(\pi-\theta)\,.$$

Daher gilt

$$\int_0^{2\pi} |f'_\alpha(\theta)|^2\,d\theta = 4\alpha^2 \int_0^{2\pi} \theta^{2\alpha-2}(2\pi-\theta)^{2\alpha-2}(\pi-\theta)^2\,d\theta\,,$$

und dieses uneigentliche Integral existiert lediglich für $\alpha > 1/2$. Mit anderen Worten: f_α gehört genau dann zu $H^1_\pi(0,2\pi)$, wenn $\alpha > 1/2$ ist. Man beachte, daß die Funktion f_α wegen ihres Verhaltens an der Stelle $\theta = 0$ lediglich Hölder-stetig mit Hölder-Exponenten α ist. Somit kann für eine $H^1_\pi(0,2\pi)$-Funktion keine Hölder-Stetigkeit mit einem Hölder-Exponenten größer als $\alpha = 1/2$ garantiert werden. ◊

Entsprechende Resultate gelten dann natürlich auch für die schwachen Ableitungen von Funktionen aus $H^s_\pi(0,2\pi)$ mit größeren s. Beispielsweise ist eine Funktion $f \in H^2_\pi(0,2\pi)$ grundsätzlich stetig differenzierbar; ihre (klassische) Ableitung ist zudem 2π-periodisch und Hölder-stetig mit Exponenten $\alpha = 1/2$.

52 Trigonometrische Interpolation

Als nächstes wenden wir uns der numerischen Approximation einer Funktion f durch trigonometrische Polynome zu. Am naheliegendsten ist dabei der Zugang über die Bestapproximation bezüglich $\mathcal{L}^2$, also die Berechnung des Polynoms aus Proposition 50.3. Leider können dessen Fourierkoeffizienten nur in den seltensten Fällen analytisch berechnet werden. Zur numerischen Approximation von (50.3) bietet sich statt dessen die Trapezsumme an: Dazu verwenden wir im weiteren die Funktionswerte von f über einem äquidistanten Gitter

$$\triangle \;=\; \{\,\theta_j = 2j\pi/N \,:\, 0 \le j \le N\,\}\,.$$

Wegen der Periodizität von f ergibt die Trapezsumme die Approximation

$$\alpha_k \;\approx\; \hat\alpha_k \;=\; \frac{1}{N}\sum_{j=0}^{N-1} f(\theta_j)e^{-\mathrm{i}k\theta_j}\,. \tag{52.1}$$

Die Näherungen $\hat{\alpha}_k$ werden diskrete Fourierkoeffizienten genannt. Für die Trapezsumme *periodischer* Funktionen liefert der folgende Satz eine besondere Fehlerabschätzung, die wesentlich besser als die übliche Abschätzung ist. Ihren Beweis werden wir am Ende dieses Abschnitts nachreichen.

Satz 52.1. *Für jede Funktion $g \in H^s_\pi(0,2\pi)$ mit $s > 1/2$ gilt*

$$\Big| \int_0^{2\pi} g(\theta)\, d\theta \,-\, \frac{2\pi}{N} \sum_{j=0}^{N-1} g(\theta_j) \Big| \;\le\; C_s \, \|g\|_{H^s_\pi(0,2\pi)} \, h^s \tag{52.2}$$

mit $h = 2\pi/N$ und einer positiven Konstanten C_s, die nur von s abhängt.

Bemerkung. Im Rahmen der trigonometrischen Interpolation beschränken wir uns im folgenden wie bereits in Satz 52.1 auf Funktionen $f \in H^s_\pi(0,2\pi)$ mit $s > 1/2$, denn nur für solche Funktionen können wir mit Hilfe von Satz 51.5 Stetigkeit und wohldefinierte Funktionswerte über dem Gitter $\triangle$ garantieren.

$\Diamond$

Besonders für glatte periodische Funktionen sind die Koeffizenten $\hat{\alpha}_k$ aus (52.1) also sehr gute Näherungen für α_k, zumindest solange der zweite Faktor $e^{-\mathrm{i}k\theta}$ in (50.3) nicht zu stark oszilliert, also solange k nicht zu groß ist. Wir studieren daher im weiteren die Näherungspolynome $t_n \in \mathcal{T}_n$ mit $n \le N/2$, gegeben durch

$$t_n(\theta) \;=\; \sum_{k=-n}^{n} \hat{\alpha}_k e^{\mathrm{i}k\theta}\,, \qquad n < N/2\,, \tag{52.3a}$$

$$\text{bzw.} \quad t_n(\theta) \;=\; \sum_{k=1-n}^{n} \hat{\alpha}_k e^{\mathrm{i}k\theta}\,, \qquad n = N/2,\ N \text{ gerade}\,. \tag{52.3b}$$

Lemma 52.2. *Seien $\phi, \psi \in H^s_\pi(0,2\pi)$ für ein $s > 1/2$. Dann definiert*

$$\langle\!\langle\, \phi, \psi \,\rangle\!\rangle \;=\; \sum_{\nu=0}^{N-1} \overline{\phi(\theta_\nu)}\psi(\theta_\nu)\,, \qquad \theta_\nu = \frac{2\pi\nu}{N}\,, \tag{52.4}$$

eine (diskrete) hermitesche Bilinearform mit

$$\langle\!\langle\, e^{\mathrm{i}j\theta}, e^{\mathrm{i}k\theta} \,\rangle\!\rangle \;=\; \begin{cases} N & \text{für } j,k \in \mathbb{Z} \ \text{ und } \ k-j = lN \ \text{ für ein } l \in \mathbb{Z}\,, \\ 0 & \text{sonst}\,. \end{cases}$$

Folglich ist (52.4) ein (diskretes) Innenprodukt in dem Teilraum $\operatorname{span}\{e^{\mathrm{i}k\theta} : -N/2 < k \le N/2\} \subset \mathcal{T}_{\lceil N/2 \rceil}$ *und die Funktionen $e^{\mathrm{i}k\theta}/\sqrt{N}$, $-N/2 < k \le N/2$, bilden ein Orthonormalsystem bezüglich dieses Innenprodukts.*

Beweis. Seien $j, k \in \mathbb{Z}$ beliebig gewählt. Dann ist

$$\langle\!\langle\, e^{\mathrm{i}j\theta}, e^{\mathrm{i}k\theta} \,\rangle\!\rangle \;=\; \sum_{\nu=0}^{N-1} e^{-\mathrm{i}j\theta_\nu} e^{\mathrm{i}k\theta_\nu} \;=\; \sum_{\nu=0}^{N-1} \bigl(e^{\mathrm{i}(k-j)2\pi/N}\bigr)^\nu$$

$$= \begin{cases} \displaystyle\sum_{\nu=0}^{N-1} \bigl(e^{\mathrm{i}2l\pi}\bigr)^\nu \;=\; N\,, & \text{falls } k-j = lN \;\text{ für ein } l \in \mathbb{Z}\,, \\[2ex] \dfrac{1-e^{\mathrm{i}(k-j)2\pi}}{1-e^{\mathrm{i}(k-j)2\pi/N}} \;=\; 0\,, & \text{sonst}\,. \end{cases}$$

Hieraus folgt unmittelbar auch die zweite Behauptung. □

Es ist zu beachten, daß das in Lemma 52.2 genannte Orthonormalsystem immer genau N Basisfunktionen enthält, unabhängig davon, ob N gerade oder ungerade ist. Der ungewöhnliche Indexbereich für k in (52.3b) ist dadurch begründet, daß die Funktionen $e^{\mathrm{i}n\theta}$ und $e^{-\mathrm{i}n\theta}$ für $n = N/2$ mit geradem N an den diskreten Punkten θ_j übereinstimmen und daher bezüglich des diskreten Innenprodukts (52.4) nicht „unterscheidbar“ sind. Mit anderen Worten: Für gerades N und $n = N/2$ ist (52.4) kein Innenprodukt über dem gesamten Raum $\mathcal{T}_n$.

Aus Lemma 52.2 erhalten wir nun sofort

Satz 52.3. *Sei $f \in H^s_\pi(0, 2\pi)$ mit $s > 1/2$ und t_n für $n < N/2$ wie in* (52.3a) *mit $\hat\alpha_k$ aus* (52.1) *definiert. Dann gilt für jedes $t \in \mathcal{T}_n \setminus \{t_n\}$:*

$$\sum_{\nu=0}^{N-1} \bigl|t_n(\theta_\nu) - f(\theta_\nu)\bigr|^2 \;<\; \sum_{\nu=0}^{N-1} \bigl|t(\theta_\nu) - f(\theta_\nu)\bigr|^2\,. \tag{52.5}$$

Beweis. Da $n < N/2$ ist, bilden die Funktionen $e^{\mathrm{i}k\theta}/\sqrt{N}$, $k = -n, \ldots, n$, eine Orthonormalbasis von $\mathcal{T}_n$ bezüglich (52.4). Wegen

$$\sum_{\nu=0}^{N-1} |t(\theta_\nu) - f(\theta_\nu)|^2 \;=\; \langle\!\langle\, t-f, t-f \,\rangle\!\rangle$$

und der Definition $\hat\alpha_k = \langle\!\langle\, e^{\mathrm{i}k\theta}/\sqrt{N}, f \,\rangle\!\rangle/\sqrt{N}$ aus (52.1) folgt die Aussage daher mit dem gleichen Argument wie im Beweis von Satz 31.6 (*c*). □

Das trigonometrische Polynom t_n ist also die Bestapproximation aus $\mathcal{T}_n$ an die Funktion f über dem Gitter $\triangle$ im Kleinste-Quadrate-Sinn. Besonders interessant ist der Fall $N = 2n$:

Satz 52.4. *Für $N = 2n$ interpoliert t_n aus* (52.3b) *die Funktion f in allen Knoten des Gitters $\triangle$, d. h. $\langle\!\langle\, f - t_n, f - t_n \,\rangle\!\rangle = 0$.*

Beweis. Nach Lemma 52.2 bilden die Funktionen

$$\tau_k(\theta) \;=\; \frac{1}{\sqrt{N}}\, e^{\mathrm{i}k\theta}\,, \qquad 1-n \le k \le n\,,$$

ein Orthonormalsystem bezüglich $\langle\!\langle\, \cdot, \cdot \,\rangle\!\rangle$, die Vektoren $y_k = \big[\tau_k(\theta_j)\big]_{j=0}^{N-1} \in \mathbb{C}^N$, $1-n \le k \le n$, also eine Orthonormalbasis des $\mathbb{C}^N$. Folglich gibt es genau eine Linearkombination

$$y \;=\; \sum_{k=1-n}^{n} \tilde{\alpha}_k y_k \;=\; \big[f(\theta_j)\big]_{j=0}^{N-1} \;\in \mathbb{C}^N\,,$$

also ein zugehöriges trigonometrisches Polynom $\tilde{t}(\theta) = \sum_{k=1-n}^{n} \tilde{\alpha}_k e^{\mathrm{i}k\theta}$ mit $\langle\!\langle\, \tilde{t} - f, \tilde{t} - f \,\rangle\!\rangle = 0$. Wie in Satz 52.3 sieht man, daß t_n aus (52.3b) die Bestapproximation von f aus $\operatorname{span}\{e^{\mathrm{i}k\theta} : -n < k \le n\} \subset \mathcal{T}_n$ bezüglich (52.5) ist. Daher muß $\tilde{t} = t_n$ sein. □

Das Polynom aus Satz 52.4 ist das *trigonometrische Interpolationspolynom*, für das wir im weiteren eine Fehlerdarstellung herleiten wollen. Zunächst beweisen wir das folgende Hilfsresultat (bei dem N auch ungerade sein darf).

Lemma 52.5. *Sei $f \in H^s_\pi(0, 2\pi)$ für ein $s > 1/2$. Dann gilt*

$$\hat{\alpha}_k = \sum_{l=-\infty}^{\infty} \alpha_{k+lN}\,, \qquad -N/2 < k \le N/2\,. \tag{52.6}$$

Beweis. Nach Satz 51.5 konvergiert die Fourierreihe von f aufgrund der getroffenen Voraussetzungen gleichmäßig gegen f. Daher gilt nach (52.1) und Lemma 52.2:

$$\begin{aligned}\hat{\alpha}_k \;&=\; \frac{1}{N}\sum_{j=0}^{N-1}\Big(\sum_{\nu=-\infty}^{\infty} \alpha_\nu e^{\mathrm{i}\nu\theta_j}\Big) e^{-\mathrm{i}k\theta_j} \;=\; \frac{1}{N}\sum_{\nu=-\infty}^{\infty} \alpha_\nu \langle\!\langle\, e^{\mathrm{i}k\theta}, e^{\mathrm{i}\nu\theta} \,\rangle\!\rangle \\ &=\; \sum_{l=-\infty}^{\infty} \alpha_{k+lN}\,.\end{aligned}$$

□

Lemma 52.5 beschreibt ein Phänomen, das als *Aliasing* bekannt ist, und beispielsweise vom Fernsehen vertraut ist: Beobachtet man dort etwa das Wagenrad einer anfahrenden Kutsche, dann scheinen bei einer gewissen Geschwindigkeit der Kutsche die Speichen des Rads stillzustehen, bevor sie sich langsam

rückwärts zu drehen beginnen – obwohl die Kutsche weiter an Geschwindigkeit zulegt. Der Film zeigt uns Bilder mit einem gewissen zeitlichen Abstand, die vom Gehirn durch eine kontinuierliche Bildsequenz interpoliert werden. Das Auge nimmt lediglich wahr, daß sich das Wagenrad von einem Bild zum nächsten um den Winkel θ etwa gegen die Fahrtrichtung gedreht hat. Für das Auge ist jedoch nicht unterscheidbar, ob sich das Wagenrad zwischen den beiden Bildern wirklich um den Winkel θ gegen die Fahrtrichtung, oder vielmehr um einen Winkel $2\pi-\theta$, $4\pi-\theta$, etc. *in Fahrtrichtung* bewegt hat. Die zugehörigen Frequenzen fallen bei der interpolierten Funktion übereinander, entsprechend der Aussage von Lemma 52.5.

Man kann Lemma 52.5 auch folgendermaßen interpretieren: Ein trigonometrisches Polynom vom Grad n kann nur eine bestimmte *Bandbreite* an Frequenzen auflösen, nämlich Frequenzen bis hin zu $n/(2\pi)$. Falls die zugrundeliegende Funktion wesentliche Frequenzen oberhalb dieser Schranke besitzt, muß der Grad der trigonometrischen Approximation erhöht werden, damit diese Frequenzen als solche erkannt werden.

Satz 52.6. *Sei $f \in H^s_\pi(0,2\pi)$ für ein $s > 1/2$ und t_n, $n \le N/2$, die trigonometrische Approximation an f aus* (52.3a) *oder das trigonometrische Interpolationspolynom* (52.3b) *zu f über dem Gitter $\triangle$ bei geradem N. Dann gilt*

$$\|f - t_n\|_{\mathcal{L}^2(0,2\pi)} \le \sqrt{1+c_s}\, n^{-s} \|f\|_{H^s_\pi(0,2\pi)}$$

mit $c_s = 2\sum_{l=1}^{\infty}(2l-1)^{-2s}$.

Beweis. Wir beschränken uns auf den Beweis der Fehlerabschätzung für das Interpolationspolynom. In diesem Fall ist $N = 2n$ und aus Lemma 52.5 erhalten wir

$$\hat\alpha_k - \alpha_k = \sum_{|l|=1}^{\infty} \alpha_{k+2ln}\,, \qquad 1-n \le k \le n\,.$$

Eine Anwendung der Cauchy-Schwarz-Ungleichung ergibt daher

$$\begin{aligned} |\alpha_k - \hat\alpha_k|^2 &\le \sum_{|l|=1}^{\infty} |k+2ln|^{-2s} \sum_{|l|=1}^{\infty} |k+2ln|^{2s} |\alpha_{k+2ln}|^2 \\ &\le \sum_{|l|=1}^{\infty} (2|l|n-n)^{-2s} \sum_{|l|=1}^{\infty} |k+2ln|^{2s} |\alpha_{k+2ln}|^2\,, \end{aligned}$$

also

$$|\alpha_k - \hat\alpha_k|^2 \;\le\; c_s n^{-2s} \sum_{|l|=1}^{\infty} |k+2ln|^{2s} |\alpha_{k+2ln}|^2 \,. \tag{52.7}$$

Für den Interpolationsfehler

$$f - t_n = \sum_{k=-\infty}^{\infty} \beta_k e^{\mathrm{i}k\theta} \qquad \text{mit} \qquad \beta_k \;=\; \begin{cases} \alpha_k - \hat\alpha_k \,, & 1-n \le k \le n \,, \\ \alpha_k \,, & \text{sonst} \,, \end{cases}$$

ist die $\mathcal{L}^2$-Norm durch

$$\frac{1}{2\pi} \, \|f - t_n\|^2_{\mathcal{L}^2(0,2\pi)} \;=\; \sum_{k=-\infty}^{\infty} |\beta_k|^2 \;\le\; \sum_{k=1-n}^{n} |\alpha_k - \hat\alpha_k|^2 \;+\; \sum_{|k|=n}^{\infty} |\alpha_k|^2$$

beschränkt. Die beiden Teilsummen können dabei separat weiter abgeschätzt werden: Aus (52.7) erhalten wir

$$\begin{aligned} \sum_{k=1-n}^{n} |\alpha_k - \hat\alpha_k|^2 \;&\le\; c_s n^{-2s} \sum_{k=1-n}^{n} \sum_{|l|=1}^{\infty} |k+2ln|^{2s} |\alpha_{k+2ln}|^2 \\ &\le\; c_s n^{-2s} \sum_{\nu=-\infty}^{\infty} |\nu|^{2s} |\alpha_\nu|^2 \;\le\; \frac{c_s}{2\pi} \, n^{-2s} \, \|f\|^2_{H^s_\pi(0,2\pi)} \,. \end{aligned}$$

Die andere Teilsumme wird wie folgt abgeschätzt:

$$\sum_{|k|=n}^{\infty} |\alpha_k|^2 \;\le\; \sum_{|k|=n}^{\infty} (|k|/n)^{2s} |\alpha_k|^2 \;\le\; n^{-2s} \sum_{k=-\infty}^{\infty} |k|^{2s} |\alpha_k|^2 \;=\; \frac{n^{-2s}}{2\pi} \, \|f\|^2_{H^s_\pi(0,2\pi)} \,.$$

Zusammen ergibt dies die Behauptung. Der Beweis für das Polynom (52.3a) geht entsprechend. □

Ganz analog zu dem Beweis von Satz 52.6 kann auch die eingangs erwähnte Fehlerabschätzung für die Trapezsumme bewiesen werden:

Beweis von Satz 52.1. Wir betrachten die Koeffizienten $\hat\alpha_k$ der trigonometrischen Polynome t_n aus (52.3) zu einer Funktion $g \in H^s_\pi(0,2\pi)$ mit der gleichmäßig konvergenten Fourierreihe $\sum \alpha_k e^{\mathrm{i}k\theta}$. Nach (52.1) und (50.3) gilt

$$\frac{1}{N} \sum_{j=0}^{N-1} g(\theta_j) \;=\; \hat\alpha_0 \,, \qquad \int_0^{2\pi} g(\theta) \, d\theta \;=\; 2\pi\alpha_0 \,,$$

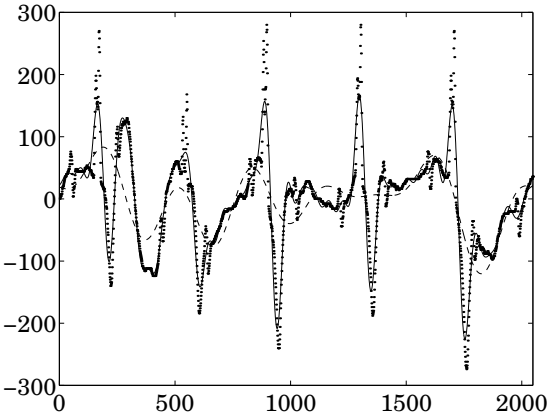


Abb. 52.1: EKG-Signal mit trigonometrischen Bestapproximationen

und daher folgt wie in (52.7)

$$\Big| \int_0^{2\pi} g(\theta)\, d\theta \,-\, \frac{2\pi}{N} \sum_{j=0}^{N-1} g(\theta_j) \Big|^2 \;=\; 4\pi^2 \, |\alpha_0 - \hat{\alpha}_0|^2$$

$$\leq\; 4\pi^2 c_s 4^s N^{-2s} \sum_{|l|=1}^{\infty} |lN|^{2s} |\alpha_{lN}|^2 \;\leq\; C_s^2 h^{2s} \, \|g\|^2_{H^s_\pi(0,2\pi)}$$

mit $C_s = \pi^{-s}\sqrt{2\pi c_s}$ und c_s wie in Satz 52.6. Damit ist Satz 52.1 bewiesen. □

Beispiel 52.7. Trigonometrische Approximationen bieten die Möglichkeit, ein gegebenes Signal auf die wesentlichen auftretenden Frequenzen zu reduzieren. Abbildung 52.1 zeigt beispielsweise ein EKG-Signal[1] mit $N = 2048$ Meßwerten sowie die in Satz 52.3 charakterisierten Bestapproximationen aus $\mathcal{T}_8$ (gebrochene Linie) und $\mathcal{T}_{32}$ (durchgezogene Linie). Die Koeffizienten $\hat{\alpha}_k$ dieser Polynome sind dabei die gleichen wie für das trigonometrische Interpolationspolynom (52.3b). Man sieht sehr deutlich, wie die scharfen Peaks in den Meßwerten mit abnehmendem Polynomgrad ausgeglättet werden und nur die wichtigsten Informationen übrig bleiben. Der Polynomgrad $n = 8$ ist für dieses Signal allerdings so klein, daß nicht mehr alle relevanten Details erfaßt werden: so wird zum Beispiel das Minimum am ca. 200. Datenpunkt von t_8 „übersehen“, für t_8 ist das Signal an dieser Stelle zu hochfrequent. ◊

[1]Die Daten wurden freundlicherweise von Prof. Dr. P. Maaß (Universität Bremen) zur Verfügung gestellt.

53 Schnelle Fouriertransformation

In diesem Abschnitt wollen wir voraussetzen, daß $N = 2n = 2^p$ eine Zweierpotenz ist. Ferner bezeichnen wir mit $\omega = e^{-\mathrm{i}2\pi/N}$ die N-te Einheitswurzel.

Die Abbildung, die den Funktionswerten $y_j = f(\theta_j)$ einer Funktion f an den Gitterpunkten $\theta_j = 2j\pi/N$, $j = 0, \ldots, N-1$, die Koeffizienten $N\hat{\alpha}_k$, $k = 1-n, \ldots, n$, aus (52.1) zuordnet, wird *diskrete Fouriertransformation* genannt. Um das trigonometrische Interpolationspolynom beziehungsweise eine Bestapproximation niedrigeren Grades zu berechnen, muß die diskrete Fouriertransformation möglichst effizient implementiert werden.

Die diskrete Fouriertransformation kann durch die folgende Matrix-Vektor-Multiplikation dargestellt werden:

$$\begin{bmatrix} c_0 \\ \vdots \\ c_n \\ c_{n+1} \\ \vdots \\ c_{N-1} \end{bmatrix} = N \begin{bmatrix} \hat{\alpha}_0 \\ \vdots \\ \hat{\alpha}_n \\ \hat{\alpha}_{1-n} \\ \vdots \\ \hat{\alpha}_{-1} \end{bmatrix} = \begin{bmatrix} \omega^0 & \omega^0 & \cdots & \omega^0 \\ \omega^0 & \omega^1 & \cdots & \omega^{N-1} \\ \omega^0 & \omega^2 & \cdots & \omega^{2(N-1)} \\ \vdots & \vdots & & \vdots \\ \omega^0 & \omega^{N-1} & \cdots & \omega^{(N-1)^2} \end{bmatrix} \begin{bmatrix} y_0 \\ \vdots \\ y_n \\ y_{n+1} \\ \vdots \\ y_{N-1} \end{bmatrix}$$

beziehungsweise

$$c = F\,y\,. \tag{53.1}$$

Die komplexe symmetrische Matrix F heißt *Fouriermatrix*. Aus Lemma 52.2 folgt $F^*F/N = I$, d. h. $F/\sqrt{N}$ ist eine unitäre Matrix und

$$F^{-1} = \frac{1}{N}\,F^*\,. \tag{53.2}$$

Mit einer herkömmlichen Matrix-Vektor-Multiplikation in (53.1) würde die Berechnung der diskreten Fourierkoeffizienten N^2 Multiplikationen kosten. Im folgenden stellen wir einen Algorithmus vor, der nur mit $O(N \log N)$ Operationen auskommt, die sogenannte *schnelle Fouriertransformation* (FFT).

Dank (53.2) kann die schnelle Fouriertransformation auch ausgenutzt werden, um die Werte $y_j = t(\theta_j)$ eines trigonometrischen Polynoms mit Koeffizienten $\hat{\alpha}_k$ zu berechnen: Mit denselben Vektoren c und y wie in (53.1) ergibt sich nämlich wegen der Symmetrie von F

$$y = F^{-1}c = \frac{1}{N}\,F^*c = \frac{1}{N}\,\overline{F\overline{c}}\,.$$

Diese Formel, von deren Gültigkeit man sich übrigens auch durch einen Vergleich von (52.1) und (52.3) vergewissern kann, ist die Grundlage für die *schnelle inverse Fouriertransformation* (IFFT).

Beispiel. Um die trigonometrischen Bestapproximationen des EKG-Signals aus Beispiel 52.7 zu berechnen, führt man zunächst eine FFT der $N = 2048$ Meßwerte durch. Für die Bestapproximation t_m vom Grad m im Sinne von Satz 52.3 werden anschließend alle berechneten Koeffizienten $\hat{\alpha}_k$ mit $|k| > m$ durch Null ersetzt. Die in Abbildung 52.1 dargestellten Funktionswerte $t_m(\theta_j)$ ergeben sich schließlich durch eine IFFT dieser Entwicklungskoeffizienten. Dabei ist allerdings zu beachten, daß die berechneten Werte $t_m(\theta_j)$ aufgrund von Rundungsfehlern nicht mehr rein reell sind. Da die Imaginärteile jedoch im Bereich der Maschinengenauigkeit liegen, können sie getrost vernachlässigt werden. ◊

Ausgangspunkt für die Herleitung der FFT ist die folgende Beobachtung:

Lemma 53.1. *Sei $M = 2m$ gerade, $\omega_M = e^{-\mathrm{i}2\pi/M}$ und $\gamma_j = \sum_{\nu=0}^{M-1} \eta_\nu \omega_M^{\nu j}$, $j = 0, \ldots, M-1$. Dann gilt*

$$\begin{aligned} \gamma_{2l} &= \sum_{\nu=0}^{m-1} \eta_\nu^{(+)} \omega_m^{\nu l}\,, & \eta_\nu^{(+)} &= \eta_\nu + \eta_{\nu+m}\,, \\ \gamma_{2l+1} &= \sum_{\nu=0}^{m-1} \eta_\nu^{(-)} \omega_m^{\nu l}\,, & \eta_\nu^{(-)} &= (\eta_\nu - \eta_{\nu+m})\omega_M^\nu\,, \end{aligned} \tag{53.3}$$

für $l = 0, \ldots, m-1$, wobei $\omega_m = \omega_M^2$ die entsprechende m-te Einheitswurzel ist.

Beweis. Für gerade Indizes ergibt sich

$$\gamma_{2l} = \sum_{\nu=0}^{M-1} \eta_\nu \omega_M^{\nu 2l} = \sum_{\nu=0}^{m-1} \big(\eta_\nu \omega_M^{2\nu l} + \eta_{\nu+m} \underbrace{\omega_M^{2\nu l + 2lm}}_{=\omega_M^{2\nu l}}\big) = \sum_{\nu=0}^{m-1} (\eta_\nu + \eta_{\nu+m}) \omega_m^{\nu l}\,.$$

Entsprechend erhält man für ungerade Indizes

$$\begin{aligned} \gamma_{2l+1} &= \sum_{\nu=0}^{m-1} \big(\eta_\nu \omega_M^{\nu(2l+1)} + \eta_{\nu+m} \omega_M^{(\nu+m)(2l+1)}\big) \\ &= \sum_{\nu=0}^{m-1} \big(\eta_\nu + \eta_{\nu+m} \underbrace{\omega_M^{(2l+1)m}}_{=-1}\big) \omega_M^{(2l+1)\nu} = \sum_{\nu=0}^{m-1} (\eta_\nu - \eta_{\nu+m}) \omega_M^\nu \omega_m^{\nu l}. \end{aligned}$$

□

function $\gamma = \mathrm{DFT}\,(\eta, \omega, M)$

% berechnet diskrete Fouriertransformation des Vektors $\eta \in \mathbb{C}^M$ mit geradem M, die
% Indizierung der Vektoren beginnt bei Null; $\omega = e^{-\mathrm{i}2\pi/M}$

$m = M/2$

$$\eta^{(+)} = \begin{bmatrix} \eta_0 \\ \eta_1 \\ \vdots \\ \eta_{m-1} \end{bmatrix} + \begin{bmatrix} \eta_m \\ \eta_{m+1} \\ \vdots \\ \eta_{M-1} \end{bmatrix}$$

$$\eta^{(-)} = \left(\begin{bmatrix} \eta_0 \\ \eta_1 \\ \vdots \\ \eta_{m-1} \end{bmatrix} - \begin{bmatrix} \eta_m \\ \eta_{m+1} \\ \vdots \\ \eta_{M-1} \end{bmatrix} \right) \bullet \begin{bmatrix} 1 \\ \omega \\ \vdots \\ \omega^{m-1} \end{bmatrix}$$

% $\bullet$ steht für die komponentenweise Multiplikation der beiden Vektoren

if $M = 2$ **then**

$$\gamma = \begin{bmatrix} \eta^{(+)} \\ \eta^{(-)} \end{bmatrix}$$

else % rekursiver Aufruf

$\gamma^{(+)} = \mathrm{DFT}\,(\eta^{(+)}, \omega^2, m)$

$\gamma^{(-)} = \mathrm{DFT}\,(\eta^{(-)}, \omega^2, m)$

$\gamma = [\gamma_0^{(+)}, \gamma_0^{(-)}, \gamma_1^{(+)}, \gamma_1^{(-)}, \ldots, \gamma_{m-1}^{(+)}, \gamma_{m-1}^{(-)}]^T$

end if

end % DFT

function $c = \mathrm{FFT}\,(y, N)$

% schnelle Fouriertransformation des Vektors $y \in \mathbb{C}^N$

$\omega = e^{-\mathrm{i}2\pi/N}$

$c = \mathrm{DFT}\,(y, \omega, N)$

end % FFT

function $y = \mathrm{IFFT}\,(c, N)$

% inverse schnelle Fouriertransformation des Vektors $c \in \mathbb{C}^N$

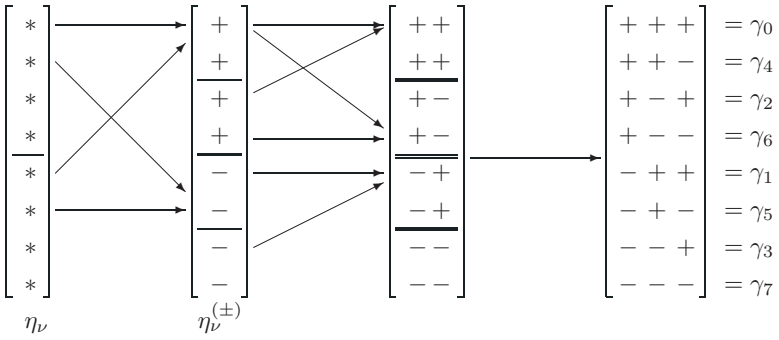
$y = \overline{\mathrm{FFT}\,(\overline{c}, N)} \,/\, N$

end % IFFT

Algorithmus 53.1: Schnelle Fouriertransformation

Da die Summen in (53.3) wieder die gleiche Gestalt wie die zu berechnende Ausgangsumme haben (allerdings nur mit halb so vielen Summanden), kann ihre Berechnung in der gleichen Weise erfolgen. Dies ergibt den rekursiven Algorithmus 53.1.

Für eine effiziente Implementierung vermeidet man rekursive Algorithmen, da Funktionsaufrufe relativ langsam sind. Statt dessen werden auf jeder Rekursionsstufe die alten Größen η_ν (die nicht weiter gebraucht werden) durch die neuen Größen $\eta_\nu^{(\pm)}$ gemäß dem folgenden Schema (dargestellt für $N = 8$) überschrieben:



Man beachte, daß der Zielvektor nicht die richtige Reihenfolge hat. Ersetzt man hingegen in dem schematisch dargestellten Zielvektor jeweils „+“ durch „0“ und „−“ durch „1“ und schreibt diese Ziffern von hinten nach vorne auf, dann ergeben sich gerade die Binärdarstellungen der entsprechenden γ-Indizes:

$$
\begin{array}{lll}
\gamma_0: & 000 & \longrightarrow \quad 000 = 0 \\
\gamma_4: & 001 & \longrightarrow \quad 100 = 4 \\
\gamma_2: & 010 & \longrightarrow \quad 010 = 2 \\
\gamma_6: & 011 & \longrightarrow \quad 110 = 6 \\
\gamma_1: & 100 & \longrightarrow \quad 001 = 1 \\
\gamma_5: & 101 & \longrightarrow \quad 101 = 5 \\
\gamma_3: & 110 & \longrightarrow \quad 011 = 3 \\
\gamma_7: & 111 & \longrightarrow \quad 111 = 7
\end{array}
$$

Diese „bit-reversal-Methode“ liefert die korrekte Zuordnung zwischen Speicherplatz und γ-Index.

Tatsächlich kann die schnelle Fouriertransformation mit verschiedenen Algorithmen realisiert werden. Wir verweisen diesbezüglich auf das detaillierte Buch von Van Loan [104].

Aufwand. Sei $N = 2^p$, also $p = \log_2 N$. Werden alle Potenzen $\omega^0, \ldots, \omega^{N-1}$ im Vorfeld berechnet, dann müssen in jedem Rekursionsschritt N komplexe

Additionen und $N/2$ komplexe Multiplikationen ausgeführt werden. Da p Rekursionsschritte durchzuführen sind, ergibt dies den Gesamtaufwand

$$N \log_2 N \text{ kompl. Additionen}\,, \quad \frac{N}{2} \log_2 N \text{ kompl. Multiplikationen}\,.$$

Rechnet man vier reelle Multiplikationen für eine komplexe Multiplikation und je zwei reelle Additionen für jede komplexe Multiplikation/Addition, dann ergeben sich somit

$$3N \log_2 N \text{ Additionen}\,, \quad 2N \log_2 N \text{ Multiplikationen}\,. \qquad \Diamond$$

Beispiel 53.2. Für eine typische Anwendung der FFT in den Ingenieurwissenschaften greifen wir die in Abschnitt 22 behandelte Schwingung eines Tragwerks auf. Dort hatten wir ohne Berücksichtigung der Reibung in (22.2) die Differentialgleichung

$$mx''(\theta) = -Ax(\theta)$$

als einfaches mathematisches Modell für die Bewegung der Gelenkkoordinaten hergeleitet. $A \in \mathbb{R}^{16\times 16}$ ist dabei die Steifigkeitsmatrix aus (3.5), die Masse m sei im folgenden gleich Eins und die Variable θ repräsentiert die Zeit.

Der Einfachheit halber nehmen wir im weiteren an, daß die Brücke zum Zeitpunkt $\theta = 0$ um eine Verschiebung $x(0)$ aus ihrem Gleichgewichtszustand ausgelenkt sei und in der Folge ohne weitere Einwirkung äußerer Kräfte schwingt. Entwickeln wir $x(0)$ in die 16 Eigenvektoren v_j der Matrix A,

$$x(0) \;=\; \sum_{j=1}^{16} \gamma_j v_j\,,$$

und bezeichnen mit λ_j, $j = 1, \ldots, 16$, die zu den Eigenvektoren gehörenden positiven Eigenwerte von A, so ergibt sich als Lösung der Differentialgleichung

$$x(\theta) \;=\; \sum_{j=1}^{16} \gamma_j \cos(\sqrt{\lambda_j}\,\theta) v_j\,, \tag{53.4}$$

wie man leicht durch Differentiation überprüft.

Abbildung 53.1 zeigt die entsprechende Oszillation des linken oberen Gelenks in diesem Tragwerk für den Fall, daß $x(0)$ mit der zu (3.6) gehörenden Verschiebung übereinstimmt: Dargestellt sind die vertikale η-Koordinate (durchgezogene Linie) und die horizontale ξ-Koordinate (gebrochene Linie) der Verschiebung des Gelenks als Funktion der Zeit. Aufgrund der speziellen Form der

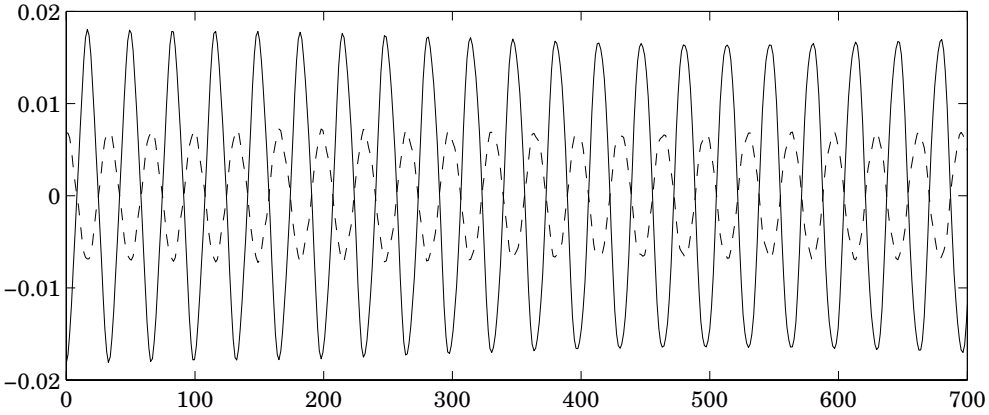


Abb. 53.1: Vertikale und horizontale Schwingung des ersten Gelenks der Brücke

auslenkenden Kraft aus (3.6) ist die vertikale Bewegung des Gelenks stärker als die horizontale Bewegung.

Gemäß der Übereinkunft aus Abschnitt 3 enthält die zweite Komponente x_2 des Lösungsvektors x aus (53.4) die vertikale Bewegung dieses ersten Gelenks. Bezeichnen wir noch mit v_{2j}, $j = 1, \ldots, 16$, die zweiten Komponenten der Eigenvektoren v_j von A, so können wir die durchgezogene Linie in Abbildung 53.1 als den Graph der Funktion

$$f(\theta) \;=\; x_2(\theta) \;=\; \sum_{j=1}^{16} \gamma_j \cos(\sqrt{\lambda_j}\,\theta) v_{2j}$$

identifizieren. Man beachte, daß f nur dann eine periodische Funktion ist, wenn alle Quotienten $\sqrt{\lambda_j/\lambda_k}$, $j, k = 1, \ldots, 16$, rationale Zahlen sind. Ansonsten heißt f *fast periodisch*.

Wir wollen nun annehmen, daß die Steifigkeitsmatrix A nicht bekannt ist und statt dessen die tatsächliche Schwingung des Tragwerks beobachtet wird. In einem solchen Fall werden in den Natur- und Ingenieurwissenschaften mit Hilfe des sogenannten *Energiespektrums* die Eigenfrequenzen des zugrunde liegenden Systems bestimmt. Dazu wird die Funktion f aus Abbildung 53.1 über einem äquidistanten Gitter

$$\triangle_\tau \;=\; \{\theta_j = j\tau \;:\; j = 0, \ldots, N-1\} \;\subset\; [0, T)$$

mit $T = N\tau$ und *Abtastrate* τ abgetastet[2]. Zu diesen gemessenen Funktionswerten $f_j = f(\theta_j)$ wird das T-periodische trigonometrische Interpolationspo-

[2]Der Begriff des „Abtastens“ (engl.: *to sample*) aus der Signalverarbeitung steht für die Messung der genannten Funktionswerte.

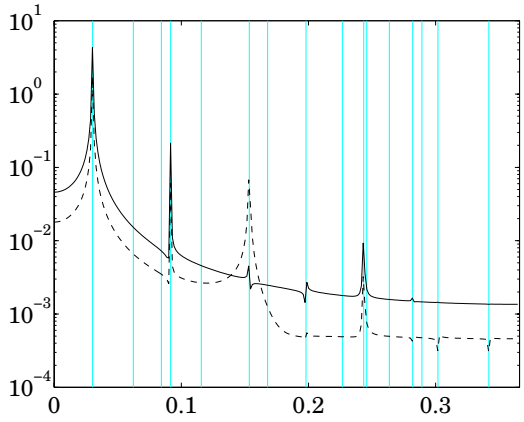


Abb. 53.2: Energiespektren der vertikalen und der horizontalen Schwingung des ersten Brückengelenks

lynom

$$t_n(\theta) \;=\; \sum_{k=1-n}^{n} \hat{\alpha}_k e^{\mathrm{i}k2\pi\theta/T}\,, \qquad n = N/2\,,$$

bestimmt (wir nehmen weiterhin an, daß N gerade ist). Unter dem Energiespektrum (engl.: *power spectrum*) der Funktion f (Abbildung 53.2) versteht man die Zuordnung des Betragsquadrats $|\hat{\alpha}_k|^2$ der Entwicklungskoeffizienten $\hat{\alpha}_k$ zu der entsprechenden Frequenz $\omega_k = k/T \in \big(-1/(2\tau), 1/(2\tau)\big]$. Wenn die abgetasteten Funktionswerte wie in diesem Beispiel reell sind, stimmen die Beträge der Koeffizienten $\hat{\alpha}_k$ und $\hat{\alpha}_{-k}$ überein, und es reicht, den positiven Frequenzbereich darzustellen.

Für Abbildung 53.2 wurde die vertikale und die horizontale Schwingung des ersten Brückengelenks im Intervall $[0, 700)$ mit Abtastrate $\tau = \pi/2.3$ abgetastet. Dies entspricht $N = 512$ Funktionswerten und einem Frequenzbereich $\omega \in [0, 0.3661)$. Die als Funktion der Frequenz dargestellten Entwicklungskoeffizienten der beiden trigonometrischen Interpolationspolynome ergeben sich gemäß (53.1) durch je eine FFT der 512 abgetasteten Funktionswerte. Die durchgezogene Linie entspricht wieder der vertikalen Schwingung des Gelenks, die gebrochene Linie gehört zu der horizontalen Schwingung. Die im Hintergrund heller eingezeichneten senkrechten Linien zeigen zum Vergleich die tatsächlichen Eigenfrequenzen $\sqrt{\lambda_j}/(2\pi)$, $j = 1, \ldots, 16$, dieses Tragwerks: Wie man sieht, weist das Energiespektrum signifikante Ausschläge über den gesuchten Frequenzen auf, allerdings sind nicht alle Eigenfrequenzen in (53.4) gleich stark vertreten. Dies liegt daran, daß die Initialauslenkung $x(0)$ nur wenige wesentliche Eigenanteile aufweist, wie Tabelle 53.1 zeigt. ◊

Tab. 53.1: Initialauslenkung je Eigenvektor

$\sqrt{\lambda_j}/2\pi$	γ_j	$\sqrt{\lambda_j}/2\pi$	γ_j	$\sqrt{\lambda_j}/2\pi$	γ_j	$\sqrt{\lambda_j}/2\pi$	γ_j
0.0302	−7.5165	0.1157	−0.0000	0.2267	−0.0000	0.2820	0.0012
0.0622	0.0000	0.1534	−0.0493	0.2434	−0.0084	0.2892	−0.0000
0.0843	−0.0000	0.1679	−0.0000	0.2459	−0.0000	0.3019	0.0011
0.0916	−0.1979	0.1981	−0.0033	0.2636	−0.0000	0.3419	−0.0005

54 Zirkulante Matrizen

Bisher haben wir die diskrete Fouriertransformation hauptsächlich im Kontext der Approximation 2π-periodischer Funktionen kennengelernt. Daneben hat sie aber auch ihre Bedeutung in der numerischen linearen Algebra. Dies liegt letztendlich daran, daß die Vektoren y_k, die uns im Beweis von Satz 52.4 begegnet sind, eine Orthonormalbasis im $\mathbb{C}^N$ bilden, die für manche Anwendungen besser geeignet ist als die herkömmliche kartesische Basis. Wir beschränken uns bei der weiteren Darstellung wie im vorangegangenen Abschnitt auf Zweierpotenzen $N = 2^p$, diese Einschränkung ist aber nicht wesentlich.

Definition 54.1. Eine *zirkulante Matrix* $C \in \mathbb{C}^{N\times N}$ ist eine Toeplitz-Matrix $C = [\, c_{j-i} \,]_{ij}$ mit $c_k = c_{N+k}$, $1 - N \le k < 0$, das heißt

$$C = \begin{bmatrix} c_0 & c_1 & \cdots & c_{N-2} & c_{N-1} \\ c_{N-1} & c_0 & c_1 & & c_{N-2} \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ c_2 & & c_{N-1} & c_0 & c_1 \\ c_1 & c_2 & \dots & c_{N-1} & c_0 \end{bmatrix}.$$

Zirkulante Matrizen sind dadurch ausgezeichnet, daß ihre Eigenvektoren die Spalten der Fouriermatrix sind:

Satz 54.2. *Ist $C \in \mathbb{C}^{N\times N}$ eine zirkulante Matrix und F die N-dimensionale Fouriermatrix* (53.1)*, dann gilt*

$$CF^* = F^*D\,, \tag{54.1}$$

wobei die Diagonalmatrix D die Eigenwerte λ_k, $k = 0, \dots, N-1$, von C enthält.

Beweis. Mit $\omega = e^{-\mathrm{i}2\pi/N}$ ist die k-te Spalte (in der ungewöhnlichen Zählweise $k = 0, \dots, N-1$) der Matrix F^* durch

$$v_k = [\, 1, \omega^k, \omega^{2k}, \dots, \omega^{(N-1)k} \,]^*$$

gegeben. Die j-te Komponente von Cv_k, $j = 0, \ldots, N-1$, hat daher die Form

$$[\,Cv_k\,]_j \;=\; \sum_{\nu=0}^{N-1} c_{\nu-j}\omega^{-k\nu} \;=\; \omega^{-jk} \sum_{\nu=0}^{N-1} c_{\nu-j}\omega^{(j-\nu)k} \;=\; \omega^{-jk} \sum_{\mu=j+1-N}^{j} c_{-\mu}\omega^{\mu k}\,.$$

Nach Voraussetzung ist $c_{-\mu}\omega^{\mu k} = c_{N-\mu}\omega^{(N+\mu)k}$ und somit erhalten wir

$$\begin{aligned}
[\,Cv_k\,]_j \;&=\; \omega^{-jk}\Big(\sum_{\mu=j+1-N}^{-1} c_{-\mu}\omega^{\mu k} + c_0 + \sum_{\mu=1}^{j} c_{-\mu}\omega^{\mu k}\Big) \\
&=\; \omega^{-jk}\Big(\sum_{\mu=j+1}^{N-1} c_{-\mu}\omega^{\mu k} + c_0 + \sum_{\mu=1}^{j} c_{-\mu}\omega^{\mu k}\Big) \\
&=\; \lambda_k\omega^{-jk}
\end{aligned}$$

mit

$$\lambda_k \;=\; c_0 + \sum_{\nu=1}^{N-1} c_{-\nu}\omega^{\nu k} \;=\; c_0 + \sum_{\nu=1}^{N-1} c_{N-\nu}\omega^{\nu k} \tag{54.2}$$

Damit ist gezeigt, daß v_k ein Eigenvektor von C mit Eigenwert λ_k ist, $k = 0, \ldots, N-1$, und folglich gilt $CF^* = F^*D$ mit

$$D \;=\; \begin{bmatrix} \lambda_0 & & \\ & \ddots & \\ & & \lambda_{N-1} \end{bmatrix}\,.$$

□

Die Eigenwerte λ_k, $k = 0, \ldots, N-1$, einer zirkulanten Matrix sind also durch (54.2) gegeben, bzw. durch

$$\begin{bmatrix} \lambda_0 \\ \lambda_1 \\ \vdots \\ \lambda_{N-1} \end{bmatrix} \;=\; F \begin{bmatrix} c_0 \\ c_{N-1} \\ \vdots \\ c_1 \end{bmatrix}\,. \tag{54.3}$$

Daher können alle Eigenwerte einer zirkulanten Matrix durch eine FFT ihrer ersten Spalte berechnet werden. Bei einer $N \times N$-Matrix kostet dies nur $O(N \log N)$ Operationen.

Auch die Matrix-Vektor-Multiplikation mit einer zirkulanten $N \times N$-Matrix kann in $O(N \log N)$ anstelle von N^2 Operationen erfolgen. Aus (54.1) und (53.2) folgt nämlich die Darstellung

$$C \;=\; F^*DF^{-*} \;=\; F^{-1}DF\,,$$

Initialisierung: $C \in \mathbb{C}^{N\times N}$ sei zirkulant, c die erste Spalte von C, $x \in \mathbb{C}^N$

$\lambda = \text{FFT}(c)$
$\widehat{x} = \text{FFT}(x)$
$\widehat{y} = \widehat{x} \bullet \lambda$ % komponentenweise Multiplikation
$y = \text{IFFT}(\widehat{y})$

Ergebnis: $y = Cx$

Algorithmus 54.1: Matrix-Vektor-Multiplikation mit zirkulanter Matrix

die mit den Ergebnissen aus Abschnitt 53 leicht in (I)FFTs übersetzt werden kann (vgl. Algorithmus 54.1).

Entsprechend können Gleichungssysteme $Cz = x$ gelöst werden, indem die komponentenweise Multiplikation in Algorithmus 54.1 durch eine komponentenweise Division der Vektoren ersetzt wird.

Aufwand. Algorithmus 54.1 benötigt drei (I)FFTs und hat daher einen Aufwand von rund $6N \log_2 N$ Multiplikationen. ◊

Algorithmus 54.1 ist auch für Toeplitz-Matrizen relevant. Jede Toeplitz-Matrix $T = [\, t_{j-i} \,]_{ij} \in \mathbb{C}^{n\times n}$ kann in eine zirkulante Matrix C der Dimension $N = 2n$ eingebettet werden:

$$C = \begin{bmatrix} T & E \\ E & T \end{bmatrix} \quad \text{mit} \quad E = \begin{bmatrix} 0 & t_{1-n} & \dots & t_{-1} \\ t_{n-1} & 0 & \ddots & \vdots \\ \vdots & \ddots & \ddots & t_{1-n} \\ t_1 & \dots & t_{n-1} & 0 \end{bmatrix}. \tag{54.4}$$

Mit Hilfe dieser Einbettung läßt sich Tx für $x \in \mathbb{C}^n$ aus dem Matrix-Vektor-Produkt

$$C \begin{bmatrix} x \\ 0 \end{bmatrix} = \begin{bmatrix} Tx \\ Ex \end{bmatrix}$$

mit der zirkulanten Matrix C ablesen. Eine Implementierung dieser Idee mit (I)FFTs findet sich in Algorithmus 54.2.

Aufwand. Da die Fouriertransformationen in Algorithmus 54.2 die doppelte Größe $N = 2n$ haben, sind die Kosten entsprechend höher als für Algorithmus 54.1, nämlich etwa $12n \log_2 n$ Multiplikationen. ◊

Zirkulante Matrizen können auch als Präkonditionierungsmatrizen eingesetzt werden, um *Toeplitz-Gleichungssysteme*

$$Tx = b \tag{54.5}$$

Initialisierung: Sei T die Toeplitz-Matrix $[t_{i-j}]_{ij} \in \mathbb{C}^{n\times n}$ und $x = [x_i] \in \mathbb{C}_n$
$c = [\, t_0, t_{-1}, \ldots, t_{1-n}, 0, t_{n-1}, \ldots, t_1 \,]^T$
$\lambda = \text{FFT}\,(c)$
$z = [\, x_1, \ldots, x_n, 0, \ldots, 0 \,]^T \in \mathbb{C}^{2n}$
$\widehat{z} = \text{FFT}\,(z)$
$\widehat{y} = \lambda \bullet z$ % komponentenweise Multiplikation
$\widetilde{y} = \text{IFFT}\,(\widehat{y})$
Ergebnis: Die ersten n Komponenten von $\widetilde{y}$ enthalten das Resultat $y = Tx$.

Algorithmus 54.2: Matrix-Vektor-Multiplikation mit Toeplitz-Matrix

mit dem präkonditionierten CG-Verfahren zu lösen, vgl. Abschnitt 10. Wir beschränken uns wieder auf den Fall, in dem die Toeplitz-Matrix

$$T = \begin{bmatrix} t_0 & t_1 & \cdots & t_{n-1} \\ t_1 & t_0 & \ddots & \vdots \\ \vdots & \ddots & \ddots & t_1 \\ t_{n-1} & \cdots & t_1 & t_0 \end{bmatrix}$$

geradzahlige Dimension aufweist (d. h. n ist gerade), reell, symmetrisch und positiv definit ist. Man macht sich dann leicht klar, daß in diesem Fall der zirkulante *Strang-Präkonditionierer*

$$M = \begin{bmatrix} c_0 & c_1 & \cdots & c_{n-1} \\ c_{n-1} & c_0 & \ddots & \vdots \\ \vdots & \ddots & \ddots & c_1 \\ c_1 & \cdots & c_{n-1} & c_0 \end{bmatrix} \tag{54.6}$$

mit

$$c_j = \begin{cases} t_j\,, & 0 \le j \le n/2 - 1\,, \\ 0\,, & j = n/2\,, \\ t_{n-j}\,, & n/2 + 1 \le j \le n - 1\,, \end{cases} \tag{54.7}$$

ebenfalls symmetrisch ist. Ferner kann man zeigen, daß fast alle Eigenwerte der zugehörigen präkonditionierten Matrix $M^{-1}T$ in unmittelbarer Nachbarschaft von $\lambda = 1$ liegen, falls die Nebendiagonaleinträge t_i von T mit wachsendem i hinreichend schnell abklingen, vgl. etwa Chan und Ng [14]. Aufgrund der Abschätzung aus Satz 35.7 benötigt das PCG-Verfahren dann nur wenige Iterationen, um die Lösung des Gleichungssystems hinreichend genau zu approximieren.

Da in jedem Schritt des PCG-Verfahrens lediglich Matrix-Vektor-Multiplikationen mit T und M^{-1} ausgeführt werden müssen und dies mit Hilfe der beiden Algorithmen 54.1 und 54.2 mit jeweils $O(n \log n)$ Multiplikationen geschehen kann, werden auch insgesamt lediglich $O(n \log n)$ Operationen benötigt, um das Toeplitz-Gleichungssystems (54.5) im Rahmen einer vorgegebenen Genauigkeit zu lösen.

Beispiel 54.3. Zur Illustration verwenden wir das präkonditionierte CG-Verfahren zur Berechnung des Wiener-Filters

$$\tilde{y}_i \;=\; \sum_{k=1}^{n} \xi_k y_{i-k}$$

zu einem reellen stationären Prozeß $\{y_i\}$, dessen Kovarianzmatrix bekannt sei. Nach Beispiel 6.1 löst der Koeffizientenvektor $x = [\xi_k]$ dieses Filters das Toeplitz-System (54.5) mit den Matrixeinträgen

$$t_k \;=\; \mathcal{E}(y_i y_{i-k})\,, \qquad k = 0, \ldots, n-1\,,$$

und der rechten Seite

$$b = [t_1, \ldots, t_n]^*\,,$$

wobei t_n entsprechend definiert ist. Ist das gegebene Signal beispielsweise ein *AR(1)-Prozeß*, d. h. gilt

$$y_i = \rho y_{i-1} + z_i\,, \qquad i \in \mathbb{Z}\,, \tag{54.8}$$

mit $\rho \in (-1, 1)$ und unabhängigen normalverteilten Zufallsvariablen $z_i \in \mathbb{R}$ mit Mittelwert Null und Varianz η^2 (entsprechende Modelle werden etwa in der Finanzwirtschaft genutzt), so ergibt sich durch Auflösen der Rekursion (54.8) die Darstellung

$$y_i \;=\; \sum_{\nu=0}^{\infty} \rho^\nu z_{i-\nu}\,,$$

und aus der Unabhängigkeit der Zufallsvariablen folgt

$$t_k \;=\; \mathcal{E}(y_i y_{i-k}) \;=\; \sum_{\mu,\nu=0}^{\infty} \rho^\nu \rho^\mu \mathcal{E}(z_{i-\nu} z_{i-k-\mu}) \;=\; \sum_{\nu=0}^{\infty} \rho^\nu \rho^{\nu-k} \eta^2 \;=\; \frac{\eta^2 \rho^{-k}}{1-\rho^2}\,.$$

Entsprechend rechnet man nach, daß ein AR(1)-Prozeß stationär ist.

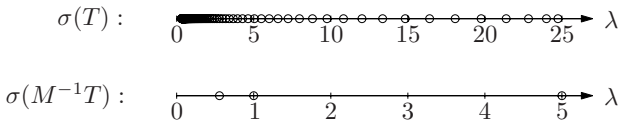


Abb. 54.1: Eigenwerte (durch Kreise markiert) vor und nach der Präkonditionierung

Dieses einfache Beispiel hat den Vorteil, daß die exakte Lösung $x = [\rho, 0, \ldots, 0]^T$ unmittelbar angegeben werden kann. Dies ermöglicht eine genauere Untersuchung des Konvergenzverhaltens des CG-Verfahrens. Dazu wählen wir M wie in (54.6), (54.7), und betrachten zunächst die Spektren von T und $M^{-1}T$ in Abbildung 54.1 (für die Parameter $\rho = 0.8$, $\eta = 1$ und die Dimension $n = 128$): Während das Spektrum von T den Wertebereich der Matrix ohne größere Lücken ausfüllt (dies ist eine ungünstige Eigenwertverteilung für das CG-Verfahren), liegen bei der präkonditionierten Matrix lediglich zwei Eigenwerte außerhalb eines Intervalls $(1 - \varepsilon, 1 + \varepsilon)$ mit $\varepsilon \approx 7 \cdot 10^{-6}$.

Aufgrund von Satz 35.7 wird man daher davon ausgehen, daß das PCG-Verfahren nur etwa zwei Iterationen benötigt, um die Eigenvektorkomponenten der beiden „Ausreißer-Eigenwerte“ in x zu rekonstruieren, bevor eine sehr schnelle Konvergenz gegen die restlichen Anteile der Lösung einsetzt. Dies wird durch die numerischen Resultate belegt: Abbildung 54.2 zeigt die Entwicklung des relativen Fehlers $\|x - x_k\|_2 / \|x\|_2$ für die Iterierten x_k der beiden Varianten des CG-Verfahrens mit und ohne Präkonditionierung: die durchgezogene Kurve zeigt den Konvergenzverlauf ohne Präkonditionierung, die Sterne geben das Ergebnis mit Präkonditionierung wieder.

Mit Präkonditionierung benötigt das Verfahren lediglich fünf Iterationen, um die Lösung auf Maschinengenauigkeit zu bestimmen. Für eine Genauigkeit von 10^{-6} sind nur drei Iterationen hinreichend, während die Variante ohne Präkonditionierung hierfür bereits über vierzig Iterationen benötigt. Beachtet man noch, daß Matrix-Vektor-Produkte mit M^{-1} billiger sind als jene mit T (da nach Algorithmus 54.2 für die Multiplikation mit T FFTs doppelter Länge notwendig sind), ergibt sich eine Reduktion der ursprünglichen Rechenzeit um fast neunzig Prozent. ◊

55 Symmetrische Transformationen

Im folgenden leiten wir aus der Fouriertransformation spezielle Darstellungen reeller Funktionen als reelle trigonometrische Reihen ab.

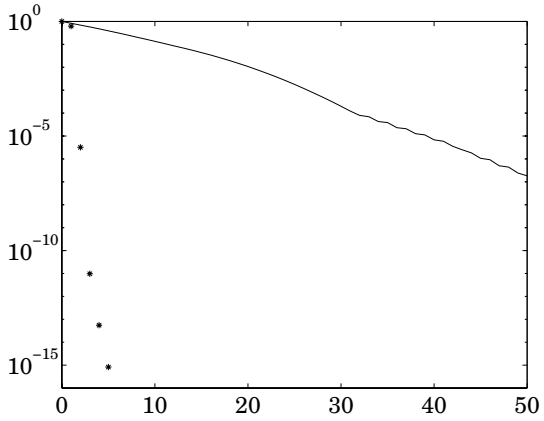


Abb. 54.2: Die relativen Fehler beim (P)CG-Verfahren

55.1 Die Sinustransformation

In mathematischen Anwendungen treten häufig relle Funktionen über einem Intervall auf, die an beiden Intervallrändern verschwinden. Erfüllt die Funktion f diese Voraussetzung im Intervall $[0, \pi]$, so kann f durch

$$f(\theta) = -f(2\pi - \theta)\,, \qquad \pi < \theta \le 2\pi\,,$$

zu einer bezüglich $\theta = \pi$ punktsymmetrischen Funktion über $[0, 2\pi]$ fortgesetzt werden. Die Fourierkoeffizienten α_k der so fortgesetzten Funktion sind rein imaginär, denn

$$\begin{aligned}\alpha_k &= \frac{1}{2\pi}\int_0^{2\pi} f(\theta)e^{-\mathrm{i}k\theta}\,d\theta = \frac{1}{2\pi}\int_0^{\pi} f(\theta)\left(e^{-\mathrm{i}k\theta} - e^{-\mathrm{i}k(2\pi-\theta)}\right)d\theta \\ &= -\frac{\mathrm{i}}{\pi}\int_0^{\pi} f(\theta)\sin k\theta\,d\theta\,.\end{aligned}$$

Mit der Abkürzung

$$b_k = -2\,\mathrm{Im}\,\alpha_k = \frac{2}{\pi}\int_0^{\pi} f(\theta)\sin k\theta\,d\theta$$

ergibt sich somit wie in Bemerkung 50.2 für f eine rein reelle Darstellung der formalen Fourierreihe:

$$f(\theta) \sim \sum_{k=1}^{\infty} b_k \sin k\theta\,, \qquad 0 \le \theta \le \pi\,. \tag{55.1}$$

Damit diese Reihenentwicklung im Sinne der Bemerkung 50.6 sinnvoll ist, muß f lediglich quadratisch integrierbar sein. Die eingangs genannten Randbedingungen $f(0) = f(\pi) = 0$ müssen hingegen nicht unbedingt erfüllt sein.

Beispiel 55.1. Wir betrachten nochmals die Hutfunktion Λ aus Beispiel 51.3, schränken aber x_2 auf $x_2 \le \pi$ ein, so daß $\Lambda(0) = \Lambda(\pi) = 0$ ist. Mit den anderen Bezeichnungen dieses Beispiels ergibt sich aus der Fourierreihe (51.3) von Λ die Darstellung

$$\begin{aligned}
&\Lambda(\theta) - \Lambda(2\pi - \theta) \\
&= \sum_{|k|=1}^{\infty} \Big(\frac{1}{2d_1\pi}\, e^{-\mathrm{i}kc_1} \sin kd_1 - \frac{1}{2d_2\pi}\, e^{-\mathrm{i}kc_2} \sin kd_2 \Big) \frac{1}{\mathrm{i}k^2} \left(e^{\mathrm{i}k\theta} - e^{-\mathrm{i}k\theta} \right) \\
&= \sum_{|k|=1}^{\infty} \Big(\frac{1}{2d_1\pi}\, e^{-\mathrm{i}kc_1} \sin kd_1 - \frac{1}{2d_2\pi}\, e^{-\mathrm{i}kc_2} \sin kd_2 \Big) \frac{2}{k^2}\, \sin k\theta
\end{aligned}$$

Wenn wir nun noch die Terme mit Index k und $-k$ jeweils zusammenfassen, erhalten wir in dem Intervall $[0, \pi]$, in dem $\Lambda(2\pi - \,\cdot\,)$ verschwindet, schließlich die Sinusreihe

$$\Lambda(\theta) \;=\; \sum_{k=1}^{\infty} \Big(\frac{1}{d_1\pi}\, \cos kc_1 \sin kd_1 - \frac{1}{d_2\pi}\, \cos kc_2 \sin kd_2 \Big) \frac{2}{k^2}\, \sin k\theta\,. \qquad \lozenge$$

Wir betrachten nun die punktsymmetrische Fortsetzung einer Funktion $f \in H_0^1(0, \pi)$ und berechnen wie in den vorangegangenen Kapiteln das trigonometrische Interpolationspolynom über dem Gitter $\triangle \subset [0, 2\pi]$ mit $N = 2n$. Aufgrund der Darstellung (52.6) sind dessen Koeffizienten $\hat\alpha_k$, $-n < k \le n$, ebenfalls rein imaginär, und darüber hinaus ergibt sich für den letzten Koeffizienten

$$\begin{aligned}
\hat\alpha_n \;&=\; \frac{1}{2n} \sum_{j=1}^{2n-1} e^{-\mathrm{i}j\pi} f(j\pi/n) \;=\; \frac{1}{2n} \sum_{j=1}^{2n-1} (-1)^j f(j\pi/n) \\
&=\; \frac{1}{2n} \sum_{j=1}^{n-1} \big((-1)^j - (-1)^{2n-j} \big) f(j\pi/n) \;=\; 0\,.
\end{aligned}$$

Das resultierende Polynom hat also ebenfalls eine Sinusentwicklung.

Satz 55.2. *Jede Funktion $f \in H_0^1(0, \pi)$ kann über $[0, \pi]$ in eine gleichmäßig konvergente Sinusreihe der Form* (55.1) *entwickelt werden. Das reelle trigonometrische Polynom*

$$t(\theta) \;=\; \sum_{k=1}^{n-1} \hat b_k \sin k\theta\,, \qquad 0 \le \theta \le \pi\,,$$

mit

$$\hat{b}_k \;=\; \frac{2}{n}\sum_{j=1}^{n-1} f(j\pi/n)\,\sin jk\pi/n\,, \qquad k=1,\dots,n-1\,, \tag{55.2}$$

interpoliert die Funktion f über dem Gitter $\widetilde{\triangle} = \{\,\theta_j = j\pi/n \,:\, 0 \le j \le n\,\}$.

Beweis. Für den Beweis muß lediglich noch die Darstellung (55.2) der Entwicklungskoeffizienten nachgereicht werden. Wie bei der Funktion f ergibt sich auch für das Polynom t die Gleichung $\hat{b}_k = -2\,\mathrm{Im}\,\hat{\alpha}_k$ und damit folgt aus (52.1) unmittelbar die Behauptung:

$$\hat{b}_k \;=\; \frac{1}{n}\sum_{j=1}^{2n-1} f(j\pi/n)\sin jk\pi/n \;=\; \frac{2}{n}\sum_{j=1}^{n-1} f(j\pi/n)\sin jk\pi/n\,.$$

□

Wie in (53.1) sammeln wir die Entwicklungskoeffizienten des in Satz 55.2 bestimmten Interpolationspolynoms und die Funktionswerte von f über dem Gitter $\widetilde{\triangle}$ in Vektoren

$$b = \big[\,\hat{b}_k\,\big]_{k=1}^{n-1} \qquad \text{und} \qquad y = \big[\,f(j\pi/n)\,\big]_{j=1}^{n-1}\,.$$

Dann ergibt sich b gemäß (55.2) aus dem Matrix-Vektor-Produkt

$$b \;=\; \frac{2}{n}\,Sy \qquad \text{mit} \quad S = \big[\,\sin jk\pi/n\,\big]_{j,k=1}^{n-1}\,. \tag{55.3}$$

Da aufgrund der Interpolationseigenschaft zudem

$$f(j\pi/n) \;=\; t(j\pi/n) \;=\; \sum_{k=1}^{n} \hat{b}_k \sin kj\pi/n\,, \qquad j=1,\dots,n-1\,,$$

gilt, haben wir außerdem die Beziehung

$$y = Sb\,. \tag{55.4}$$

Proposition 55.3. *Die* Sinusmatrix S *aus* (55.3) *ist reell und symmetrisch, die skalierte Matrix $\sqrt{2/n}\,S$ ist unitär.*

Beweis. Die Symmetrie von S ist offensichtlich und die Identität $S^2 = (n/2)I$ folgt unmittelbar aus (55.3) und (55.4). □

Die Matrix-Vektor-Produkte mit S in (55.3) und (55.4) können mit Hilfe der schnellen Fouriertransformation mit lediglich $O(n\log n)$ statt n^2 Operationen ausgewertet werden (*schnelle Sinustransformation*), wenn man die Herleitung von Satz 55.2 in einen entsprechenden Algorithmus umsetzt. Dies ergibt die Routine DST aus Algorithmus 55.1. Allerdings sei an dieser Stelle betont, daß es für die schnelle Sinustransformation bessere Algorithmen gibt, die nur etwa halb so viele Operationen benötigen, vgl. [104].

```
function  z = DST(y, n)
  % berechnet z = Sy für y ∈ R^{n-1} (schnelle Sinustransformation)
  η = [0, y_1, ..., y_{n-1}, 0, -y_{n-1}, ..., -y_1]^T ∈ R^{2n}
  α = FFT(η, 2n)        % α = [α_0, α_1, ..., α_{2n-1}]^T
  z = i [α_1, α_2, ..., α_{n-1}]^T / 2
end      % DST
```

Algorithmus 55.1: Schnelle Sinustransformation

55.2 Die Kosinustransformation

Während die Sinustransformation vor allem innerhalb der Mathematik von Bedeutung ist (vgl. etwa Kapitel XVII), hat die *Kosinustransformation* auch außerhalb der Mathematik Anwendungen, da sie keine Einschränkungen an die Randwerte $f(0)$ und $f(\pi)$ vorgibt. Bei der Kosinustransformation wird die reellwertige Funktion $f \in \mathcal{L}^2(0,\pi)$ zunächst durch $f(\theta) = f(2\pi - \theta)$ spiegelsymmetrisch zu einer Funktion über $[0, 2\pi]$ fortgesetzt wird. Letztere hat rein reelle Fourierkoeffizienten

$$\alpha_k \;=\; \frac{1}{2\pi}\int_0^{2\pi} f(\theta) e^{-\mathrm{i}k\theta}\,d\theta \;=\; \frac{1}{\pi}\int_0^{\pi} f(\theta)\cos k\theta\,d\theta$$

und somit folgt aus Bemerkung 50.2 die formale Entwicklung

$$f(\theta) \;\sim\; \frac{a_0}{2} \;+\; \sum_{k=1}^{\infty} a_k \cos k\theta \tag{55.5}$$

in eine Kosinusreihe mit Entwicklungskoeffizienten

$$a_k \;=\; 2\,\mathrm{Re}\,\alpha_k \;=\; \frac{2}{\pi}\int_0^{\pi} f(\theta)\cos k\theta\,d\theta\,.$$

Beispiel. Um die entsprechende Kosinusreihe der Hutfunktion aus Beispiel 55.1 zu berechnen, verwenden wir die Fourierreihe von $\Lambda(\theta)+\Lambda(2\pi-\theta)$: Mit (51.3) folgt

$$\begin{aligned}
&\Lambda(\theta) + \Lambda(2\pi-\theta) - \frac{x_2 - x_0}{2\pi} \\
&\quad= \sum_{|k|=1}^{\infty} \Big(\frac{1}{2d_1\pi}\, e^{-\mathrm{i}kc_1} \sin kd_1 \,-\, \frac{1}{2d_2\pi}\, e^{-\mathrm{i}kc_2} \sin kd_2 \Big) \frac{1}{\mathrm{i}k^2} \left(e^{\mathrm{i}k\theta} + e^{-\mathrm{i}k\theta} \right) \\
&\quad= \sum_{k=1}^{\infty} \Big(\frac{1}{d_2\pi}\, \sin kc_2 \sin kd_2 \,-\, \frac{1}{d_1\pi}\, \sin kc_1 \sin kd_1 \Big) \frac{2}{k^2}\, \cos k\theta\,,
\end{aligned}$$

und hieraus ergibt sich wegen $\Lambda(2\pi - \,\cdot\,) = 0$ über $[0,\pi]$ die Kosinusreihe

$$\Lambda(\theta) \;=\; \frac{x_2 - x_0}{2\pi} \;+\; \sum_{k=1}^{\infty} \Big(\frac{1}{d_2\pi}\, \sin kc_2 \sin kd_2 \,-\, \frac{1}{d_1\pi}\, \sin kc_1 \sin kd_1 \Big) \frac{2}{k^2}\, \cos k\theta$$

von Λ über $[0,\pi]$. $\lozenge$

Im Kontext der Kosinustransformation ist es von Vorteil, die Funktion f über dem *verschobenen Gitter*

$$\triangle' \;=\; \{\, \tfrac{2j+1}{2n}\,\pi \;:\; j = 0, 1, \ldots, n-1 \,\} \;\subset\; (0,\pi) \tag{55.6}$$

zu interpolieren. Um dennoch auf die Resultate dieses Kapitels zurückgreifen zu können, betrachten wir die Funktion $g(\theta) = f(\theta + \frac{\pi}{2n})$ über $[0, 2\pi]$. Dazu denken wir uns f zunächst spiegelsymmetrisch auf $[\pi, 2\pi]$ und dann 2π-periodisch auf $\mathbb{R}$ fortgesetzt. Nach Satz 52.4 wird g durch das trigonometrische Polynom

$$t_g(\theta) \;=\; \sum_{k=1-n}^{n} \hat\alpha_k e^{\mathrm{i}k\theta}$$

mit

$$\begin{aligned}
\hat\alpha_k \;&=\; \frac{1}{2n} \sum_{j=0}^{2n-1} g(\tfrac{j}{n}\pi) e^{-\mathrm{i}k\frac{j}{n}\pi} \;=\; \frac{1}{2n} \sum_{j=0}^{2n-1} f\big(\tfrac{2j+1}{2n}\pi\big) e^{-\mathrm{i}k\frac{j}{n}\pi} \\
&=\; \frac{e^{\mathrm{i}k\pi/(2n)}}{2n} \sum_{j=0}^{2n-1} f(\tfrac{2j+1}{2n}\pi) e^{-\mathrm{i}k\frac{2j+1}{2n}\pi} \\
&=\; \frac{e^{\mathrm{i}k\pi/(2n)}}{n} \sum_{j=0}^{n-1} f\big(\tfrac{2j+1}{2n}\pi\big) \cos k\,\tfrac{2j+1}{2n}\pi\,, \qquad k = 1-n, \ldots, n\,,
\end{aligned}$$

über dem Gitter $\triangle$ interpoliert, d. h.

$$t_g(\tfrac{j}{n}\pi) = g(\tfrac{j}{n}\pi) = f(\tfrac{2j+1}{2n}\pi), \qquad j = 0, \ldots, N\,. \tag{55.7}$$

Dabei gilt $\hat\alpha_n = 0$, da der Kosinus für $k = n$ nur an seinen Nullstellen ausgewertet wird. Aus (55.7) folgt unmittelbar, daß das trigonometrische Polynom

$$t_f(\theta) \;=\; t_g(\theta - \tfrac{\pi}{2n}) \;=\; \sum_{k=1-n}^{n-1} \hat\alpha_k e^{-\mathrm{i}k\pi/(2n)} e^{\mathrm{i}k\theta} \;=\; \frac{1}{\sqrt{2}}\,\hat a_0 + \frac{1}{2} \sum_{|k|=1}^{n-1} \hat a_k e^{\mathrm{i}k\theta}$$

mit

$$\hat a_k \;=\; \begin{cases} \sqrt{2}\,\hat\alpha_k & \text{für } k = 0\,, \\ 2e^{-\mathrm{i}k\pi/(2n)}\hat\alpha_k & \text{für } 1 \le |k| \le n-1\,, \end{cases}$$

die Funktion f über dem verschobenen Gitter $\triangle'$ interpoliert. Die Koeffizienten $\hat{a}_k$ können über die unten stehende Summenformel (55.8) berechnet werden. Insbesondere ist $\hat{a}_k = \hat{a}_{-k}$. Damit haben wir das folgende Resultat bewiesen:

Satz 55.4. *Jede Funktion $f \in H^1(0,\pi)$ kann über $[0,\pi]$ in eine gleichmäßig konvergente Kosinusreihe der Form* (55.5) *entwickelt werden. Das reelle trigonometrische Polynom*

$$t(\theta) \;=\; \frac{1}{\sqrt{2}}\,\hat{a}_0 \;+\; \sum_{k=1}^{n-1} \hat{a}_k \cos k\theta\,, \qquad 0 \le \theta \le \pi\,,$$

mit

$$\hat{a}_k \;=\; \begin{cases} \dfrac{\sqrt{2}}{n} \displaystyle\sum_{j=0}^{n-1} f(\tfrac{2j+1}{2n}\pi) & \text{für } k = 0\,, \\[2ex] \dfrac{2}{n} \displaystyle\sum_{j=0}^{n-1} f(\tfrac{2j+1}{2n}\pi)\, \cos k\,\tfrac{2j+1}{2n}\pi & \text{für } k = 1,\ldots,n-1\,, \end{cases} \tag{55.8}$$

interpoliert die Funktion f über dem verschobenen Gitter $\triangle'$ aus (55.6).

Für die Vektoren

$$a = \big[\, \hat{a}_k \,\big]_{k=0}^{n-1} \qquad \text{und} \qquad y = \big[\, f(\tfrac{2j+1}{2n}\pi) \,\big]_{j=0}^{n-1}$$

erhalten wir aus (55.8) und der Interpolationseigenschaft die beiden Gleichungen

$$a \;=\; \frac{2}{n}\, Cy \qquad \text{und} \qquad y \;=\; C^* a \tag{55.9}$$

mit der Kosinusmatrix

$$C = \big[\, c_{kj} \,\big]_{k,j=0}^{n-1} \qquad \text{mit} \quad c_{kj} = \begin{cases} 1/\sqrt{2}\,, & k = 0\,, \\ \cos k\tfrac{2j+1}{2n}\pi\,, & k \ne 0\,. \end{cases} \tag{55.10}$$

Die Matrix C ist im Gegensatz zu S nicht symmetrisch, aber wie zuvor ist $\sqrt{2/n}\,C$ unitär. Ferner können Matrix-Vektor-Produkte mit C und C^* ebenfalls mit $O(n \log n)$ Operationen berechnet werden. Eine entsprechende Implementierung für die *schnelle* (*inverse*) *Kosinustransformation* findet sich in Algorithmus 55.2.

Ein Vorteil der Kosinustransformation besteht darin, daß eine Funktion f aus $H^1(0,\pi)$ durch die symmetrische Definition $f(\theta) = f(2\pi - \theta)$ zu einer Funktion

```
function  z = DCT(y, n)
  % berechnet z = Cy für y = [y_1, ..., y_n]^T ∈ R^n (schnelle Kosinustransformation)
  η = [y_1, ..., y_n, y_n, ..., y_1]^T ∈ R^{2n}
  α = FFT(η, 2n)        % α = [α_0, α_1, ..., α_{2n-1}]^T ∈ R^{2n}
  z_0 = α_0/(2√2)
  for k = 1, ..., n − 1 do
    z_k = α_k e^{−ikπ/(2n)}/2
  end for
  % z = [z_0, z_1, ..., z_{n-1}]^T
end       % DCT

function  y = IDCT(z, n)
  % berechnet y = C*z für z = [z_0, ..., z_{n-1}]^T ∈ R^n (inverse Kosinustransformation)
  α_0 = 2√2 z_0
  for k = 1, ..., n − 1 do
    α_k = 2e^{ikπ/(2n)} z_k
    α_{2n−k} = conj(α_k)
  end for
  α_n = 0        % α = [α_0, α_1, ..., α_{2n-1}]^T ∈ R^{2n}
  η = IFFT(α, 2n)        % η = [η_0, η_1, ..., η_{2n-1}]^T ∈ R^{2n}
  y = n [η_0, η_1, ..., η_{n-1}]^T/2
end       % IDCT
```

Algorithmus 55.2: Schnelle Kosinustransformation

$f \in H^1_\pi(0, 2\pi)$ fortgesetzt wird. Die π-periodische Fortsetzung, die der Fouriertransformation zugrunde liegt oder die punktsymmetrische Fortsetzung durch die Sinustransformation wären hingegen beide im Punkt $\theta = \pi$ in der Regel nicht mehr stetig und lägen somit in keinem Raum $H^s_\pi(0, 2\pi)$ mit $s > 1/2$. Die Koeffizienten der Kosinustransformation fallen somit stärker ab als die entsprechenden Fourierkoeffizienten. Deswegen bietet sich die Kosinustransformation für *Datenkompressionen* an.

Beispiel 55.5. Seit Beginn der 90er Jahre wird die Kosinustransformation zur Bildkompression im sogenannten *JPEG-Standard* verwendet. Abbildung 55.1 zeigt ein bekanntes Testbild[3] und dessen zweidimensionale Kosinustransformierte. Die zweidimensionale Kosinustransformation der gegebenen Daten y_{jk}, $j, k = 0, \ldots, n-1$, erfolgt in Verallgemeinerung von (55.9) über die Doppel-

[3]Der Abdruck erfolgt mit freundlicher Genehmigung des Massachusetts Institute of Technology. Das Bild ist in der MATLAB-Programmierumgebung als Datei `cameraman.tif` enthalten

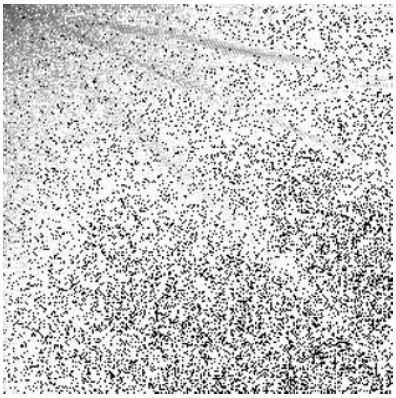


Abb. 55.1: Originalbild und dessen Kosinustransformation

summe

$$\hat{a}_{jk} \;=\; \frac{4}{n^2} \sum_{\mu=0}^{n-1} \sum_{\nu=0}^{n-1} c_{j\nu} c_{k\mu} y_{\nu\mu} \;=\; \frac{4}{n^2} \sum_{\mu=0}^{n-1} c_{k\mu} \sum_{\nu=0}^{n-1} c_{j\nu} y_{\nu\mu}$$

mit $c_{\nu\mu}$ aus (55.10). Dies entspricht einer vertikalen Kosinustransformation der einzelnen Spalten der Matrix $Y = [y_{jk}]$, gefolgt von einer zweiten (horizontalen) Kosinustransformation, angewendet auf die Zeilen des Resultats. Diese Summe kann mit Hilfe der Kosinusmatrix kurz in der Form

$$A = CYC^* = \big(C(CY)^*\big)^*$$

mit $A = [\hat{a}_{jk}]$ geschrieben werden. Die Berechnung von A kann somit die Funktion DCT aus Algorithmus 55.2 verwenden; dies ergibt die *schnelle zweidimensionale Kosinustransformation.*

In unserem Beispiel besteht das Bild aus 256×256 Bildpixeln, die in entsprechender Anordnung die Matrix $Y \in \mathbb{R}^{256\times256}$ bilden. Das rechte Bild aus Abbildung 55.1 zeigt die Kosinuskoeffizienten als Grauskalenbild der Matrix A; die Einträge von A umspannen acht (!) Zehnerpotenzen und sind in logarithmischer Darstellung aufgetragen. Die dunklen Werte gehören zu großen Einträgen. Sie sind zu einem Großteil in der linken oberen Ecke konzentriert, die zu den niedrigen Kosinusfrequenzen gehört: die horizontalen Frequenzen werden von links nach rechts größer, die vertikalen Frequenzen sind von oben nach unten aufsteigend angeordnet. Durch die scharfen Kontraste des Bildes ergeben sich auch große Koeffizienten bei irregulär verteilten hohen Frequenzen.

Abb. 55.2: Originalbild und JPEG-Kompression auf etwa 10% Speicherbedarf

Zur Bildkompression werden die Werte der Kosinuskoeffizienten auf äquidistante Gitter gerundet. Kleine Koeffizienten werden auf diese Weise zu Null gerundet und müssen nicht mehr abgespeichert werden. Die entsprechende Gitterweite wird üblicherweise in Abhängigkeit von der Frequenz gewählt: Im allgemeinen wächst die Gitterweite mit der Frequenz, um so die speziellen kognitiven Fähigkeiten des Gehirns zu berücksichtigen. Grob gesprochen ergibt sich so eine stärkere Datenkompression in den hohen Frequenzen. Im JPEG-Standard wird die Kompression schließlich nicht auf das Bild als Ganzes angewendet, sondern auf Teile von jeweils 8×8 oder 16×16 Pixeln; für genauere Details sei auf das Buch von Pennebaker und Mitchell [83] verwiesen.

Abbildung 55.2 zeigt eine JPEG-Kompression, bei der das komprimierte Bild nur einen Speicherbedarf von rund 10% des Originalbilds benötigt. ◊

Aufgaben

1. Bestimmen Sie die Fourierreihe der über $[0, 2\pi]$ definierten Funktionen

$$\text{(a)} \quad f(\theta) = \theta\,, \qquad\qquad \text{(b)} \quad g(\theta) = \theta(2\pi - \theta)\,.$$

In welchen Sobolevräumen $H^s_\pi(0, 2\pi)$ liegen f und g? Warum steht das nicht im Widerspruch zu der Tatsache, daß f und g unendlich oft differenzierbar sind?

2. (a) Für zwei 2π-periodische Funktionen $f, g \in \mathcal{L}^2(0, 2\pi)$ wird die Faltung $f * g$ von f und g definiert durch

$$(f * g)(x) \;=\; \int_0^{2\pi} f(x - y)g(y)\,dy\,.$$

Zeigen Sie, daß zwischen den Fourierkoeffizienten γ_k von $f * g$ und den Fourierkoeffizienten α_k, β_k von f und g folgender Zusammenhang besteht:

$$\gamma_k \;=\; 2\pi\alpha_k\beta_k\,, \qquad k \in \mathbb{Z}\,.$$

(b) Sei $f \in \mathcal{L}^2(0, 2\pi)$ eine 2π-periodische Funktion mit den Fourierkoeffizienten α_k. Zeigen Sie, daß die Funktion $f(\,\cdot - s)$ für festes $s \in \mathbb{R}$ die Fourierkoeffizienten $\beta_k = e^{-iks}\alpha_k$ hat.

3. Die ansonsten stetige, 2π-periodische Funktion f habe eine Sprungstelle bei $\theta_0 \in [0, 2\pi)$ mit $f(\theta_0+) < f(\theta_0-)$. Das *Gibbs-Phänomen* besagt, daß die Bestapproximation t_n von f aus $\mathcal{T}_n$ (bezüglich $\mathcal{L}^2$) für große n das tatsächliche Sprungintervall $[\,f(\theta_0+), f(\theta_0-)\,]$ um etwa 9% in beide Richtungen überschätzt (vgl. Abbildung 50.1).
Weisen Sie das Gibbs-Phänomen exemplarisch für die Funktion $f(\theta) = \theta$ aus Aufgabe 1 (mit Sprungintervall $[0, 2\pi]$ bei $\theta_0 = 0$) nach: Zeigen Sie, daß

$$\lim_{n\to\infty} \inf_{\theta\in(0,2\pi)} t_n(\theta) \;=\; \pi - 2\int_0^\pi \frac{\sin\tau}{\tau}\,d\tau \;=\; -0.56228... \;=\; -\gamma\,2\pi$$

mit $\gamma = 0.08948...$
Hinweis: Betrachten Sie $t_n(\rho/n)$ für festes $\rho > 0$.

4. Betrachten Sie die charakteristische Funktion χ_h des Intervalls $[0, h] \subset [0, 2\pi)$. Zeigen Sie, daß für $s < 1/2$ Konstanten $0 < c_s < C_s$ existieren mit

$$c_s\,h^{-s}\,\|\chi_h\|_{\mathcal{L}^2(0,2\pi)} \;\le\; \|\chi_h\|_{H^s_\pi(0,2\pi)} \;\le\; C_s\,h^{-s}\,\|\chi_h\|_{\mathcal{L}^2(0,2\pi)}\,, \qquad 0 \le h < 2\pi\,.$$

5. Sei $s > 0$ und $f \in H^s_\pi(0, 2\pi)$. Beweisen Sie die *Interpolationsungleichung*

$$\|f\|_{H^\sigma_\pi(0,2\pi)} \;\le\; c\,\|f\|^{1-\sigma/s}_{\mathcal{L}^2(0,2\pi)}\,\|f\|^{\sigma/s}_{H^s_\pi(0,2\pi)}\,, \qquad 0 \le \sigma \le s\,.$$

6. Sei $f \in H^s_\pi(0, 2\pi)$ für ein $s > 0$ und $t_n \in \mathcal{T}_n$ die Bestapproximation an f bezüglich der $\mathcal{L}^2$-Norm. Zeigen Sie, daß

$$\|f - t_n\|_{\mathcal{L}^2(0,2\pi)} \;\le\; n^{-s}\,\|f\|_{H^s_\pi(0,2\pi)}$$

und

$$\|f - t_n\|_{H^\sigma_\pi(0,2\pi)} \;\le\; n^{\sigma-s}\,\|f\|_{H^s_\pi(0,2\pi)}$$

für $0 < \sigma < s$.

7. Sei $f \in H^s_\pi(0,2\pi)$ für ein $s \in (0,1)$ und V_k der Raum der Treppenfunktionen mit Gitterweite $2\pi/2^k$. Zeigen Sie, daß

$$\inf_{\varphi \in V_k} \|f - \varphi\|_{\mathcal{L}^2(0,2\pi)} \;\le\; c\, 2^{-ks}\, \|f\|_{H^s_\pi(0,2\pi)}$$

mit einer von k und s unabhängigen Konstanten $c > 0$.
Hinweis: Beweisen Sie die Behauptung zunächst für trigonometrische Polynome $f = t_n \in \mathcal{T}_n$, $n = 2^k$, mit Hilfe von Satz 43.1.

8. (a) Sei $y \in H^1_\pi(0,\pi)$ mit $y(0) = y(\pi) = 0$. Beweisen Sie die Ungleichung

$$\|y\|_{\mathcal{L}^2(0,\pi)} \;\le\; \|y'\|_{\mathcal{L}^2(0,\pi)}\,.$$

(b) $s \in H^1(a,b)$ interpoliere eine Funktion $f \in H^1(a,b)$ über einem Gitter $\triangle \subset [a,b]$ mit Gitterweite h. Verwenden Sie die Ungleichung aus (a) für die Fehlerabschätzung

$$\|f - s\|_{\mathcal{L}^2(a,b)} \;\le\; \frac{h}{\pi}\, \|f' - s'\|_{\mathcal{L}^2(a,b)}\,.$$

(c) Zeigen Sie anhand von Beispielen, daß die Konstanten Eins bzw. $1/\pi$ in den beiden Abschätzungen scharf sind.

9. Wenden Sie Satz 52.1 zu einem Beweis der asymptotischen Fehlerentwicklung der Trapezsumme aus Satz 39.2 für 2π-periodische Funktionen $f \in C^{2m+2}[0,2\pi]$ an, deren Ableitungen ebenfalls alle 2π-periodisch sind.

10. Sei $N = 2n$ und $0 \le k < N$. Zeigen Sie, daß

$$l_j(\theta) \;=\; \frac{\sin n(\theta - \theta_j)}{N}\, \cot \frac{\theta - \theta_j}{2}$$

ein trigonometrisches Polynom vom Grad n ist und daß

$$l_j(\theta_k) = \delta_{jk}\,, \qquad \theta_j = 2j\pi/N,\ j = 0, \dots, N-1\,.$$

l_j übernimmt also bei der trigonometrischen Interpolation die Rolle des *Lagrange-Grundpolynoms* aus Abschnitt 37.

11. Beweisen Sie für $f \in H^s_\pi(0,2\pi)$ mit $s > 1/2$ und das zugehörige trigonometrische Interpolationspolynom $t_n \in \mathcal{T}_n$ aus (52.3b) die Fehlerabschätzung

$$\|f - t_n\|_{[0,2\pi]} \;\le\; \tilde{c}_s n^{1/2-s}\, \|f\|_{H^s_\pi(0,2\pi)}$$

mit einer geeigneten Konstanten $\tilde{c}_s > 0$.
Hinweis: Verwenden Sie die Ungleichung $\sum_{\nu=n}^{\infty} \nu^{-2s} \;\le\; n^{-2s} + \int_n^\infty t^{-2s}\, dt$.

12. Schreiben Sie ein Programm, welches die n-te Partialsumme $\sum_{k=-n}^{n} \alpha_k e^{ik\theta}$ einer Fourierreihe an den N äquidistanten Stützstellen $2j\pi/N$, $j = 0, \dots, N-1$, plottet. Wählen Sie $n < N = 2^p$ und verwenden Sie die (I)FFT. Approximieren Sie mit Ihrem Programm die folgenden (formalen) Fourierreihen:

(a) die sogenannte *Delta-Distribution* $\delta(\cdot - \pi)$, konzentriert im Punkt $\theta = \pi$, mit der formalen Fourierreihe

$$\delta(\theta - \pi) \sim \sum_{k=-\infty}^{\infty} \frac{(-1)^k}{2\pi} e^{ik\theta};$$

(b) die Fourierreihe

$$f(\theta) \sim \sum_{|k|=2}^{\infty} \frac{1}{|k| \log |k|} e^{ik\theta}.$$

Beachten Sie, daß die Funktion f nach Definition 51.4 zu dem Sobolevraum $H_\pi^{1/2}(0, 2\pi)$ gehört, daß aber die Partialsummen der Fourierreihe für $\theta = 0$ gegen $+\infty$ divergieren (vgl. Heuser [53, Satz 33.3]). Die Funktion f ist daher nicht stetig periodisch fortsetzbar. Warum nicht?

13. Einer Funktion $f \in \mathcal{L}^2((0, 2\pi)^2)$ kann man entsprechend zu Abschnitt 50 eine formale zweidimensionale Fourierreihe

$$f(\theta, \phi) \sim \sum_{j,k=-\infty}^{\infty} \alpha_{jk} e^{i(j\theta + k\phi)}$$

zuordnen, wobei die Koeffizienten α_{jk}, $j, k \in \mathbb{Z}$, durch

$$\alpha_{jk} = \frac{1}{(2\pi)^2} \int_{[0,2\pi]^2} f(\theta, \phi) e^{-i(j\theta + k\phi)} \, d(\theta, \phi)$$

gegeben sind.
(a) Bestimmen Sie die Koeffizienten des (zweidimensionalen) trigonometrischen Polynoms

$$t_n(\theta, \phi) = \sum_{j,k=1-n}^{n} \hat{\alpha}_{jk} e^{i(j\theta + k\phi)},$$

das die Interpolationsaufgabe

$$t_n(\theta_\nu, \theta_\mu) = f(\theta_\nu, \theta_\mu), \qquad \nu, \mu = 0, \ldots, 2n-1,$$

mit den Abszissen $\theta_\nu = \pi\nu/n$ löst.
(b) Sei $N = 2n$ und $Y = [\, f(\theta_\nu, \theta_\mu) \,]_{\nu,\mu=0}^{N-1}$ die $N \times N$-Matrix der zu interpolierenden Funktionswerte von f. Zeigen Sie, daß eine geeignete Anordnung der Koeffizienten $\hat{\alpha}_{jk}$ aus (a) in einer $N \times N$-Matrix C existiert, so daß

$$N^2 C = FYF;$$

hierbei ist F die Fouriermatrix aus (53.1).
(c) Schreiben Sie Routinen

$$C = \texttt{fft2D}\,(Y, N), \qquad Y = \texttt{ifft2D}\,(C, N),$$

die die entsprechenden Transformationen mit $O(N^2 \log N)$ Operationen durchführen (zweidimensionale FFT). Verwenden Sie hierzu die Routinen `fft` und `ifft` aus Abschnitt 53.

14. In Aufgabe VIII.8 soll ein Algorithmus zur Berechnung des interpolierenden periodischen kubischen Splines bestimmt werden. Bei äquidistanten Knoten muß dazu ein lineares Gleichungssystem gelöst werden, dessen Koeffizientenmatrix A die folgende Gestalt hat:

$$A = \begin{bmatrix} 2 & 1/2 & & 0 & & 1/2 \\ 1/2 & 2 & 1/2 & & \ddots & \\ & 1/2 & 2 & \ddots & & 0 \\ 0 & & \ddots & \ddots & & \\ & \ddots & & & \ddots & 1/2 \\ 1/2 & & 0 & & 1/2 & 2 \end{bmatrix}.$$

(a) Bestimmen Sie alle Eigenwerte von A.
(b) Welchen Aufwand erfordert die Lösung des linearen Gleichungssystems mit den Techniken aus Abschnitt 54? Vergleichen Sie dies mit der Gauß-Elimination. Was ist effizienter?

15. Verallgemeinern Sie Satz 54.2 auf die Singulärwertzerlegung einer rechteckigen Matrix der Gestalt

$$C = \begin{bmatrix} c_0 & c_1 & c_2 & c_3 & \dots & c_{2n-1} \\ c_{2n-2} & c_{2n-1} & c_0 & c_1 & & c_{2n-3} \\ c_{2n-4} & c_{2n-3} & c_{2n-2} & c_{2n-1} & & c_{2n-5} \\ \vdots & & \ddots & \ddots & & \vdots \\ c_2 & c_3 & & & \dots & c_1 \end{bmatrix}.$$

Dazu bezeichne F_m die $m-$dimensionale Fouriermatrix und $x_k^{(m)}$, $k = 1, \dots, m$, seien die Spaltenvektoren von F_m^*. Gehen Sie dann wie folgt vor:
(a) Zeigen Sie, daß für $k = 1, \dots, n$

$$Cx_k^{(2n)} = \mu_k x_k^{(n)} \qquad \text{und} \qquad Cx_{n+k}^{(2n)} = \mu_{n+k} x_k^{(n)}$$

mit geeigneten μ_k und μ_{n+k} gilt.
(b) Setzen Sie für $k = 1, \dots, n$ die Singulärvektoren von C als Linearkombinationen

$$v_k = \alpha_k x_k^{(2n)} + \beta_k x_{n+k}^{(2n)} \qquad \text{und} \qquad v_{n+k} = \alpha_{n+k} x_k^{(2n)} + \beta_{n+k} x_{n+k}^{(2n)}$$

an.

16. Sei $T \in \mathbb{K}^{n \times n}$ eine Toeplitz-Matrix.
(a) Bestimmen Sie die zirkulante Matrix $C \in \mathbb{K}^{n \times n}$, die das Minimierungsproblem

$$\text{minimiere} \quad \|T - C\|_F$$

löst.
(b) Beweisen Sie, daß mit T auch die Lösung C hermitesch und positiv definit ist.
Hinweis: Verwenden Sie, daß die Frobenius-Norm invariant unter Orthogonaltransformationen ist.

17. $F \in \mathbb{C}^{N\times N}$ sei die Fouriermatrix und T eine beliebige Toeplitz-Matrix der Dimension N. Ferner sei

$$D \;=\; \begin{bmatrix} 1 & & & \\ & e^{-\mathrm{i}\pi/N} & & \\ & & \ddots & \\ & & & e^{-\mathrm{i}(N-1)\pi/N} \end{bmatrix}.$$

Zeigen Sie, daß $\frac{1}{N}F^*TDF$ eine verallgemeinerte Cauchy-Matrix ist.
Hinweis: Verwenden Sie Aufgabe II.5 (a) mit $\alpha = 1$ und $\beta = -1$ sowie Aufgabe II.6 (b).

18. Sei (55.1) die gleichmäßig konvergente Sinusreihe von $f \in H_0^1(0,\pi)$. Zeigen Sie, daß

$$\|f\|^2_{\mathcal{L}^2(0,\pi)} \;=\; \frac{\pi}{2}\sum_{j=1}^{\infty} b_j^2\,.$$

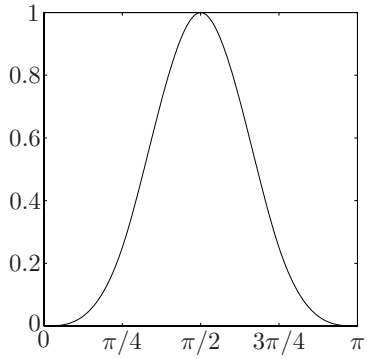
19. Rechnen Sie nach, daß

$$B(x) \;=\; \sum_{j=1}^{\infty} b_{2j-1}\sin(2j-1)x$$

mit

$$b_{2j-1} \;=\; \frac{3\cdot 2^{10}}{\pi^4}(-1)^{j+1}\Big(\frac{\sin\big((2j-1)\pi/8\big)}{2j-1}\Big)^4$$

die Sinusreihe für den kubischen B-Spline zu dem Gitter $\triangle = \{0, \pi/4, \pi/2, 3\pi/4, \pi\}$ ist (vgl. Abbildung).



20. Sei S die Sinusmatrix aus (55.3) und $A \in \mathbb{R}^{(n-1)\times(n-1)}$. Zeigen Sie, daß SAS^{-1} genau dann eine reelle Diagonalmatrix ist, wenn A die Gestalt

$$A \;=\; \begin{bmatrix} a_0 & a_1 & & \dots & a_{n-3} & a_{n-2} \\ a_1 & a_0 & a_1 & & & a_{n-3} \\ \vdots & a_1 & a_0 & \ddots & & \vdots \\ \vdots & & \ddots & \ddots & \ddots & a_2 \\ a_{n-3} & & & \ddots & a_0 & a_1 \\ a_{n-2} & a_{n-3} & \dots & a_2 & a_1 & a_0 \end{bmatrix} - \begin{bmatrix} a_2 & a_3 & \dots & a_{n-2} & 0 & 0 \\ a_3 & & \iddots & 0 & 0 & 0 \\ \vdots & \iddots & \iddots & \iddots & 0 & a_{n-2} \\ a_{n-2} & 0 & \iddots & \iddots & \iddots & \vdots \\ 0 & 0 & 0 & \iddots & & a_3 \\ 0 & 0 & a_{n-2} & \dots & a_3 & a_2 \end{bmatrix}$$

besitzt.
Hinweis: Bestimmen Sie eine Basis für den Unterraum der Matrizen mit obiger Gestalt und beweisen Sie die gewünschte Eigenschaft für die Basiselemente.

X Multiskalenbasen

Die Approximation mit trigonometrischen Polynomen hat den Vorteil, daß eine vorgegebene Funktion elegant in einen niederfrequenten („glatten“) und einen hochfrequenten Anteil zerlegt werden kann. Zudem können die zugehörigen Entwicklungskoeffizienten effizient ausgerechnet werden. Ein Nachteil der trigonometrischen Polynome ist hingegen ihre schlechte *Lokalisierungseigenschaft*, die dazu führt, daß zur Approximation von Sprungfunktionen Polynome hohen Grades benötigt werden.

In den vergangenen Jahren wurden aus diesem Grund alternative Funktionenbasen vorgeschlagen (sogenannte *Multiskalenbasen*), die sich wie die trigonometrischen Polynome gut zur Frequenzanalyse eignen, aber eine verbesserte örtliche Lokalisierung aufweisen. In der Numerik sind derartige Basen interessant, wenn in einer Anwendung verschiedene „Längenskalen“ einer Variablen von Bedeutung sind.

In der Literatur werden Multiskalenbasen in der Regel unter dem Stichwort *Wavelets* behandelt. Aus der rasch anwachsenden Zahl von Monographien zu diesem Thema seien hier die Bücher von Louis, Maaß und Rieder [70] und von Jensen und la Cour-Harbo [57] angeführt. Die hier behandelten linearen Spline-Wavelets kommen in den meisten Büchern nicht in dieser Form vor. Lesenswerte Ausnahmen sind das Buch von Chui [17] und der Übersichtsartikel von Cohen [20].

56 Das Haar-Wavelet

Beispiel. Aus Beispiel 50.7 erhalten wir durch eine lineare Transformation des Intervalls $(0, 2\pi)$ auf das Intervall $(0, 1)$ die 1-periodische Fourierreihe der charakteristischen Funktion $f = \chi_{[a,b]}$ eines Intervalls $[a, b] \subset (0, 1)$:

$$f(x) \sim b - a + \frac{1}{\pi} \sum_{|k|=1}^{\infty} e^{-\mathrm{i}kc} \frac{\sin kd}{k} e^{\mathrm{i}2k\pi x}, \qquad 0 \le x \le 1,$$

mit $c = \pi(a+b)$ und $d = \pi(b-a)$. Die Entwicklungskoeffizienten verhalten sich (bis auf die durch den Sinusterm bedingte Oszillation) im wesentlichen wie $1/k$, fallen also nur relativ langsam ab. Bezeichnet t_n die nach n Termen abgebrochene Reihe, dann ergibt sich ein Fehler $\|f - t_n\|_{\mathcal{L}^2(0,1)} \sim n^{-1/2}$, vgl. (50.6). Für eine gute Approximation der charakteristischen Funktion sind also viele Entwicklungsterme der Fourierreihe notwendig. $\Diamond$

Das gleiche Phänomen beobachtet man bei anderen $\mathcal{L}^2$-Funktionen mit Sprungunstetigkeiten. Man sagt daher, daß rapide Änderungen im „Ortsbereich" einer Funktion (also bezüglich der x-Variablen) durch trigonometrische Polynome nur schlecht approximiert werden können; sie haben eine schlechte *Lokalisierungseigenschaft* bezüglich der Ortsvariablen.

Im folgenden konstruieren wir exemplarisch Funktionenbasen, die ebenfalls eine Aufspaltung in unterschiedliche Frequenzbereiche erlauben, andererseits aber auch gute Lokalisierungseigenschaften aufweisen. Wir greifen für unsere Konstruktion auf Splineräume zurück, denen in der Numerik ohnehin eine sehr große Bedeutung zukommt.

Für $k = 0, 1, 2, \ldots, p$ bezeichne

$$\triangle_k \;=\; \{\, jh_k \;:\; j = 0, 1, \ldots, 2^k,\; h_k = 2^{-k} \,\} \;\subset\; [0,1] \tag{56.1}$$

eine Familie äquidistanter Gitter über $[0,1]$. Dabei entsteht $\triangle_{k+1}$ durch eine einmalige Verfeinerung von $\triangle_k$, d. h. einer Halbierung aller Teilintervalle von $\triangle_k$. Das feinste Gitter $\triangle_p$ hat die Gitterweite $h_p = 2^{-p}$, während das gröbste Gitter $\triangle_0$ lediglich aus einem einzigen Intervall besteht.

Mit V_k wird zunächst der Raum der Treppenfunktionen über $\triangle_k$ bezeichnet, also die Menge aller Funktionen $f \in \mathcal{L}^2(0,1)$, die im Innern aller Teilintervalle $\mathcal{I}_{kj} = \big[jh_k, (j+1)h_k\big]$, $j = 0, \ldots, 2^k - 1$, jeweils konstant sind. Da $\triangle_{k+1}$ eine Verfeinerung von $\triangle_k$ ist, ergibt dies eine ganze *Skala* von Funktionenräumen

$$V_0 \subset V_1 \subset \cdots \subset V_{p-1} \subset V_p\,.$$

Als Basis für V_k verwenden wir wie in Abschnitt 43 die charakteristischen Funktionen χ_{kj} der Teilintervalle $\mathcal{I}_{kj}$, $j = 0, \ldots, 2^k - 1$. Im Gegensatz zu Abschnitt 43 werden die Basisfunktionen im folgenden allerdings so skaliert, daß ihre $\mathcal{L}^2$-Norm Eins ist:

$$\chi_{kj}(x) = 2^{k/2}\chi(2^k x - j)\,, \qquad \chi = \chi_{[0,1]}\,.$$

Etwas eingänglicher ist die äquivalente Darstellung

$$\begin{aligned} \chi_{k0}(x) = 2^{k/2}\chi(2^k x)\,, \qquad \chi_{kj}(x + jh_k) = \chi_{k0}(x)\,, \\ k = 0, \ldots, p\,, \quad j = 1, \ldots, 2^k - 1\,. \end{aligned} \tag{56.2}$$

Die Transformation $x \mapsto 2^k x$ in (56.2) entspricht einer Stauchung der Ausgangsfunktion χ, während das Argument $x + jh_k$ von χ_{kj} eine Verschiebung (Translation) der Grundfunktion χ_{k0} um jh_k *nach rechts* bewirkt.

Für festes k können in dem Raum V_k nur bestimmte Frequenzen einer Funktion f dargestellt werden (ähnlich dem Raum $\mathcal{T}_n$ der trigonometrischen Polynome vom Grad kleiner gleich $n = 2^k$, der die gleiche Dimension aufweist). Zur Darstellung höherer Frequenzen muß das Gitter $\triangle_k$ verfeinert werden. Dies führt auf das Gitter $\triangle_{k+1}$ und den zugehörigen Funktionenraum V_{k+1}. Da V_k ein Unterraum von V_{k+1} ist, wäre es in diesem Kontext wünschenswert, die Basis von V_k geeignet zu einer Basis von V_{k+1} zu ergänzen. Hierzu benötigen wir eine Zerlegung

$$V_{k+1} = V_k \oplus W_k \tag{56.3}$$

in den Unterraum V_k der (relativ) niederfrequenten Funktionen und einen Komplementärraum W_k der (relativ zu V_k) hochfrequenten Funktionen sowie eine geeignete Basis von W_k.

In diesem Abschnitt beschränken wir uns auf den Fall, daß W_k das Orthogonalkomplement von V_k ist und suchen – in Analogie zu (56.2) – Basisfunktionen ψ_{kj} von W_k, die aus einer einzigen Funktion ψ durch

$$\psi_{kj}(x) = 2^{k/2}\psi(2^k x - j)$$

beziehungsweise

$$\begin{gathered} \psi_{k0}(x) = 2^{k/2}\psi(2^k x)\,, \qquad \psi_{kj}(x + jh_k) = \psi_{k0}(x)\,, \\ k = 0, \ldots, p-1\,, \quad j = 1, \ldots, 2^k - 1\,, \end{gathered} \tag{56.4}$$

erzeugt werden können. Wegen der Einschränkung $W_k \subset V_{k+1}$ muß ψ_{k0} eine Treppenfunktion über $\triangle_{k+1}$ sein. Aus (56.4) folgt daher, daß ψ eine Treppenfunktion über dem *Referenzgitter*

$$\triangle_* = \{j/2 : j \in \mathbb{Z}\} \tag{56.5}$$

ist. Als geeignet erweist sich die Funktion

$$\psi(x) = \begin{cases} 1\,, & 0 < x \le 1/2\,, \\ -1\,, & 1/2 < x \le 1\,, \\ 0\,, & \text{sonst}\,, \end{cases} \tag{56.6}$$

das sogenannte *Haar-Wavelet*[1]. Abbildung 56.1 zeigt die zugehörigen Funktionen $\psi_{3,2}$ und $\psi_{4,12}$ aus (56.4). Der englische Begriff *Wavelet* läßt sich mit „kleine Welle“ übersetzen.

[1]Die Funktion ψ aus (56.6) wurde erstmals von Haar [41] im Jahr 1910 eingeführt. Die Wiederentdeckung in der Wavelet-Theorie erfolgte in den 80er Jahren.

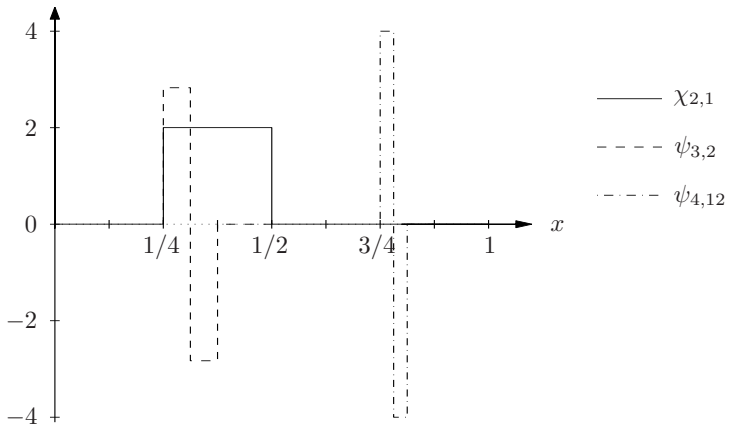


Abb. 56.1: Verschiedene Basisfunktionen für den Raum der Treppenfunktionen

Im weiteren zeigen wir, daß das Funktionensystem $\{\psi_{kj} : j = 0, \ldots, 2^k - 1\}$ aus (56.4) das Orthogonalkomplement W_k von V_k in V_{k+1} aufspannt. Bevor wir dies beweisen, führen wir noch den folgenden Begriff ein:

Definition 56.1. Unter dem *Träger* $\operatorname{supp}(f)$ einer stückweise stetigen Funktion f über einem Intervall $\mathcal{I} \subset \mathbb{R}$ wird der Abschluß aller $x \in \mathcal{I}$ mit $f(x) \mathrel{/}= 0$ verstanden (engl. *support*).

Der Träger von χ_{kj} ist also gerade das Intervall $\mathcal{I}_{kj}$.

Proposition 56.2. *Sei $\chi = \chi_{[0,1]}$ und ψ wie in* (56.6) *definiert. Dann bilden $\{\chi_{kj} : j = 0, \ldots, 2^k - 1\}$ und $\{\psi_{kj} : j = 0, \ldots, 2^k - 1\}$ Orthonormalbasen (bezüglich $\mathcal{L}^2$) der orthogonalen Komplementärräume V_k und W_k von V_{k+1}.*

Beweis. Sowohl χ_{kj} als auch ψ_{kj} haben denselben Träger $\mathcal{I}_{kj}$. Da sich diese Intervalle für verschiedene j und festes k in maximal einem gemeinsamen Punkt berühren, verschwinden alle Integrale

$$\int_0^1 \chi_{kj}\chi_{kj'}\,dx\,, \quad \int_0^1 \psi_{kj}\psi_{kj'}\,dx \quad \text{und} \quad \int_0^1 \chi_{kj}\psi_{kj'}\,dx$$

für $j \mathrel{/}\neq j'$, $j, j' \in \{0, \ldots, 2^k - 1\}$. Zudem gilt für $j = j'$

$$\int_0^1 \chi_{kj}^2(x)\,dx \;=\; \int_{j2^{-k}}^{(j+1)2^{-k}} 2^k\chi^2(2^kx - j)\,dx \;=\; \int_0^1 \chi^2(t)\,dt \;=\; 1\,,$$

$$\int_0^1 \psi_{kj}^2(x)\,dx \;=\; \int_{j2^{-k}}^{(j+1)2^{-k}} 2^k\psi^2(2^kx - j)\,dx \;=\; \int_0^1 \psi^2(t)\,dt \;=\; 1\,,$$

und

$$\int_0^1 \chi_{kj}(x)\psi_{kj}(x)\,dx \;=\; \int_{j2^{-k}}^{(j+1)2^{-k}} 2^k\chi(2^kx-j)\psi(2^kx-j)\,dx$$
$$= \int_0^1 \chi(t)\psi(t)\,dt \;=\; \int_0^{1/2} dt \;-\; \int_{1/2}^1 dt \;=\; 1/2-1/2 \;=\; 0\,.$$

Folglich bilden die χ_{kj} eine Orthonormalbasis von V_k und die ψ_{kj} ein orthonormales System in V_{k+1}, das senkrecht auf V_k steht. Da die Anzahl der betrachteten Funktionen mit der Dimension 2^{k+1} von V_{k+1} übereinstimmt, bilden die ψ_{kj} somit eine Orthonormalbasis von W_k. □

Neben der eingangs genannten Basis $\{\chi_{k+1,j}\}_j$ bildet das Funktionensystem

$$\{\chi_{kj}, \psi_{kj} \,:\, j = 0, \ldots, 2^k-1\}$$

somit eine weitere Orthonormalbasis von V_{k+1} entsprechend der orthogonalen Zerlegung (56.3). Diese Basis wird *Zweiskalenbasis* von V_{k+1} genannt. Für eine Funktion $f \in V_{k+1}$ bezeichnet man ihren Anteil in V_k als *Trend*, ihren Anteil in W_k als *Fluktuation* bezüglich des Gitters $\triangle_k$.

Aus der Zweiskalenbasis ergibt sich in entsprechender Weise eine *Multiskalenbasis*, wenn die Zerlegung (56.3) für jedes Gitter rekursiv vorgenommen wird, also

$$\begin{aligned} V_p &= V_{p-1} \oplus W_{p-1} \;=\; V_{p-2} \oplus W_{p-2} \oplus W_{p-1} \;=\; \ldots \\ &= V_0 \oplus W_0 \oplus W_1 \oplus \cdots \oplus W_{p-1} \end{aligned} \tag{56.7}$$

mit den entsprechenden Basisfunktionen von $V_0, W_0, \ldots, W_{p-2}$ und W_{p-1}. Da das Haar-Wavelet der Grundbaustein dieser Multiskalenbasis ist, spricht man auch von der *Haar-Basis* aller Treppenfunktionen über $\triangle_p$.

Beispiel. Zur Illustration betrachten wir wieder das einführende Beispiel und approximieren die charakteristische Funktion $f = \chi_{[a,b]}$ eines Teilintervalls $[a,b] \subset (0,1)$. Wir haben bereits gesehen, daß zur Approximation von f sehr viele trigonometrische Basisfunktionen benötigt werden und mit n Koeffizienten im wesentlichen eine Genauigkeit der Größenordnung $n^{-1/2}$ erreicht wird.

Mit der Haar-Basis geht das wesentlich besser. Wir definieren

$$a_k = \sup\{x \in \triangle_k \,:\, x \le a\}\,, \qquad b_k = \inf\{x \in \triangle_k \,:\, x \ge b\}\,,$$

und setzen $\varphi_k = \chi_{[a_k,b_k]}$. Aufgrund der Konstruktion ist $\varphi_k \in V_k$ und

$$\inf_{\varphi\in V_k} \|f-\varphi\|^2_{\mathcal{L}^2(0,1)} \le \|f-\varphi_k\|^2_{\mathcal{L}^2(0,1)} \le |a-a_k| + |b-b_k| \le 2\cdot 2^{-k}\,.$$

Mit $n = 2^k$ Ansatzfunktionen erhalten wir also in etwa die gleiche Approximationsgüte wie mit n Basisfunktionen der trigonometrischen Basis. Bei der Bestapproximation an f aus V_k treten allerdings weit weniger als n Terme in der Multiskalenbasisentwicklung auf. Nach Satz 31.6 (c) werden hierfür nur diejenigen Basisfunktionen benötigt, die nicht orthogonal zu f sind. Neben χ_{00} sind das lediglich diejenigen ψ_{kj}, deren Träger den Punkt a oder den Punkt b im Innern enthalten, also höchstens zwei Basisfunktionen auf jedem Gitter. Somit müssen für die Approximationsgüte $O(n^{-1/2})$ bei der Haar-Multiskalenbasis lediglich $O(\log n)$ Koeffizienten abgespeichert werden. ◊

Es bleibt noch die Frage zu klären, wie aus der konventionellen Darstellung einer Funktion $f \in V_p$ wie in Abschnitt 43 (d. h. einer Entwicklung bzgl. der $\{\chi_{pj}\}_j$) die Entwicklung in die Multiskalenbasis berechnet werden kann. Eine entsprechende Basistransformation ist auf der Grundlage von (56.7) möglich. Hierfür werden die Identitäten

$$\begin{aligned} \sqrt{2}\,\chi_{kj} &= \chi_{k+1,2j} + \chi_{k+1,2j+1}\,, \\ \sqrt{2}\,\psi_{kj} &= \chi_{k+1,2j} - \chi_{k+1,2j+1}\,, \end{aligned} \qquad j = 0, \dots, 2^k - 1\,, \tag{56.8}$$

beziehungweise

$$\begin{aligned} \chi_{k+1,2j} &= \chi_{kj}/\sqrt{2} + \psi_{kj}/\sqrt{2}\,, \\ \chi_{k+1,2j+1} &= \chi_{kj}/\sqrt{2} - \psi_{kj}/\sqrt{2}\,, \end{aligned} \qquad j = 0, \dots, 2^k - 1\,, \tag{56.9}$$

benötigt. Diese führen leicht auf eine entsprechende Matrixformulierung der Basistransformation: Ist $f = \sum_{j=0}^{N-1} \xi_{k+1,j}\chi_{k+1,j} \in V_{k+1}$ gegeben, $n = 2^k$ und $N = 2n$, dann folgt aus Aufgabe VI.1 unmittelbar die Entwicklung

$$f = \sum_{j=0}^{n-1} \big(\xi_{kj}\chi_{kj} + \eta_{kj}\psi_{kj}\big) \qquad \text{mit} \qquad \begin{aligned} \xi_{kj} &= \tfrac{1}{\sqrt{2}}\big(\xi_{k+1,2j} + \xi_{k+1,2j+1}\big), \\ \eta_{kj} &= \tfrac{1}{\sqrt{2}}\big(\xi_{k+1,2j} - \xi_{k+1,2j+1}\big), \end{aligned} \tag{56.10}$$

bezüglich der Zweiskalenbasis. Algorithmus 56.1 zeigt, wie hieraus durch rekursives Fortschreiten die Transformation in die Multiskalenbasis berechnet werden kann. Die Entwicklungskoeffizienten werden dabei in dem Vektor

$$w = [\,\xi_{0,0}, \eta_{0,0}, \eta_{1,0}, \eta_{1,1}, \eta_{2,0}, \dots, \eta_{p,n-1}\,]^T$$

ausgegeben.

Aufwand. Die Transformation (56.10) kostet $2 \cdot 2^k = N$ Multiplikationen und genauso viele Additionen. Die Multiskalentransformation einer Funktion $f \in V_p$ benötigt daher insgesamt jeweils $\sum_{k=0}^{p-1} 2^{k+1} \approx 2^{p+1} = 2 \dim V_p$ Multiplikationen und Additionen und ist somit noch billiger zu implementieren als die FFT. ◊

```
function  w = fhwt(x, N)        % Wavelettransformation
  % N sei Zweierpotenz, n = N/2; x = [x_j]_{j=0}^{N-1} enthalte die Koeffizienten ξ_{k+1,j},
  % die Hilfsvektoren ξ = [ξ_j]_{j=0}^{n-1} und η = [η_j]_{j=0}^{n-1} enthalten die Koeffizienten ξ_{kj} bzw. η_{kj}
  if N = 1 then
    w = x
  else
```

$$\xi = \frac{1}{\sqrt{2}} \begin{bmatrix} x_0 + x_1 \\ x_2 + x_3 \\ \vdots \\ x_{N-2} + x_{N-1} \end{bmatrix}$$

$$\eta = \frac{1}{\sqrt{2}} \begin{bmatrix} x_0 - x_1 \\ x_2 - x_3 \\ \vdots \\ x_{N-2} - x_{N-1} \end{bmatrix}$$

$$w = \begin{bmatrix} \text{fhwt}\,(\xi, n) \\ \eta \end{bmatrix}$$

```
  end if
end       % fhwt

function  x = ifhwt (w, N)      % inverse Wavelettransformation
  % N Zweierpotenz; x, ξ und η wie oben; w = [w_j]_{j=0}^{N-1} ∈ ℝ^N
  n = N/2
  v = [w_0, ..., w_{n-1}]^T
  η = [w_n, ..., w_{N-1}]^T
  if n = 1 then
    ξ = v
  else
    ξ = ifhwt (v, n)
  end if
```

$$x = \frac{1}{\sqrt{2}} \begin{bmatrix} \xi_0 + \eta_0 \\ \xi_0 - \eta_0 \\ \xi_1 + \eta_1 \\ \xi_1 - \eta_1 \\ \vdots \\ \xi_{n-1} + \eta_{n-1} \\ \xi_{n-1} - \eta_{n-1} \end{bmatrix}$$

```
end       % ifhwt
```

Algorithmus 56.1: Schnelle Haar-Wavelet-Transformation

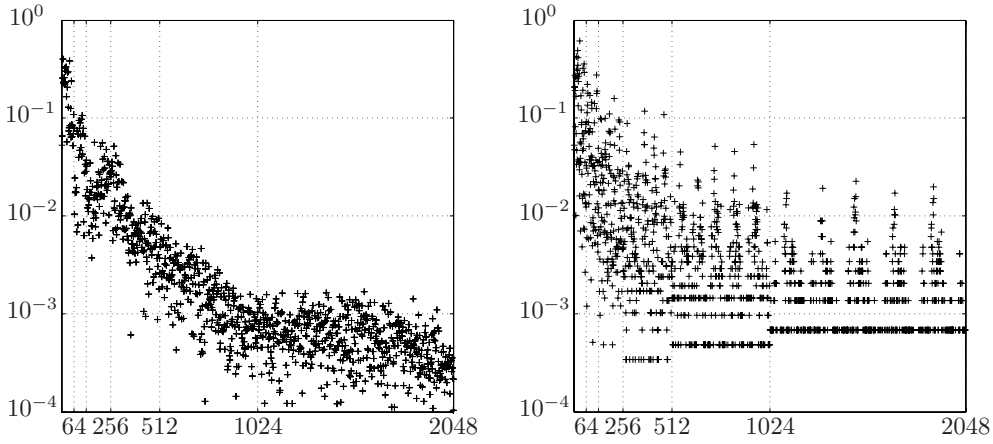


Abb. 56.2: Fouriertransformation von f (links) und Wavelettransformation (rechts)

Beispiel 56.3. Zum Vergleich verwenden wir einmal trigonometrische Polynome und einmal die Waveletbasis, um das EKG-Signal $f \in V_{11}$ aus Abbildung 52.1 in hoch- und niederfrequente Strukturen zu zerlegen. Abbildung 56.2 zeigt in einer logarithmischen Skala die Absolutbeträge der jeweiligen Entwicklungskoeffizienten. Diese sind in beiden Bildern der Übersicht halber nach zunehmenden Frequenzen der entsprechenden Basisfunktionen sortiert. Die vertikalen Unterteilungen deuten die Unterräume V_k an.

In beiden Fällen werden die Entwicklungskoeffizienten von f mit zunehmender Frequenz kleiner. Dieser Abfall ist bei den Fourierkoeffizienten etwas ausgeprägter und vor allem gleichmäßiger, was daran liegt, daß die Waveletdarstellung das *lokale* Frequenzverhalten analysiert. Solche ortsabhängigen Charakteristika der Funktion f erkennt man nur an den Waveletkoeffizienten aber nicht an den Fourierkoeffizienten.

Beide Bilder in Abbildung 56.2 zeigen lediglich diejenigen Entwicklungskoeffizienten, deren Betrag größer als 10^{-4} ist. Für mögliche Anwendungen in der Datenkompression sei jedoch angemerkt, daß von den 2048 Waveletkoeffizienten über ein Viertel betragsmäßig kleiner als 10^{-4} sind. Sie können gegebenenfalls vernachlässigt werden. Hingegen liegen lediglich 24 Fourierkoeffizienten, also weniger als 1.2% unterhalb dieser Toleranzschwelle.

In Abbildung 56.3 wird die rekursive Waveletzerlegung an dem EKG-Signal f demonstriert. In der obersten Zeile ist der Anteil von f in V_8 dargestellt. Die anderen Zeilen zeigen die Anteile von f in V_k (links) und W_k (rechts), also den Trend und die jeweilige Fluktuation bezüglich der Gitter $\triangle_k$ mit $k = 7$, $k = 6$, usw. bis $k = 2$ ganz unten. Man kann gut erkennen, wie das Ausgangssignal aus diesen einzelnen Komponenten zusammengesetzt ist. $\Diamond$

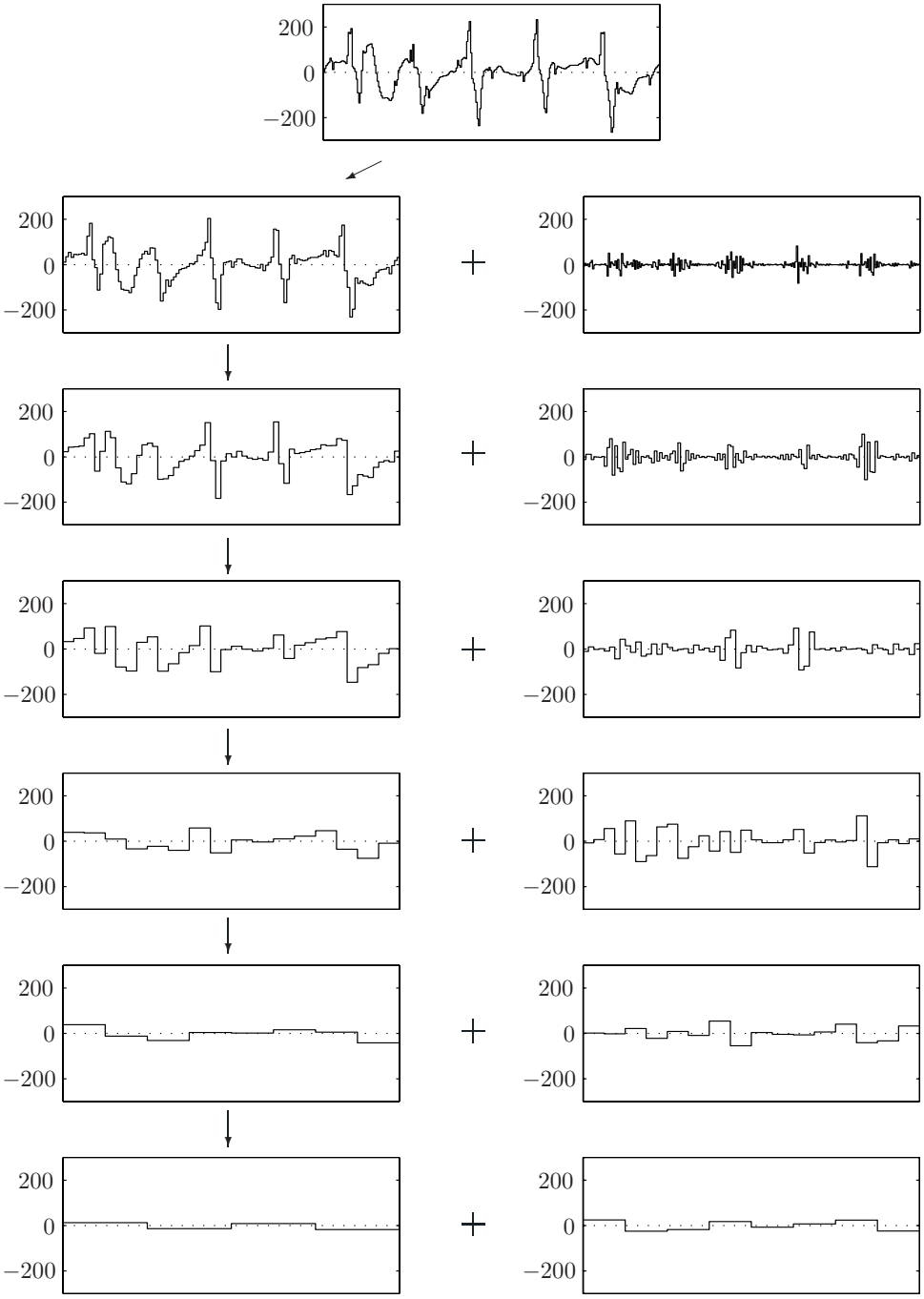


Abb. 56.3: Waveletzerlegung des EKG-Signals

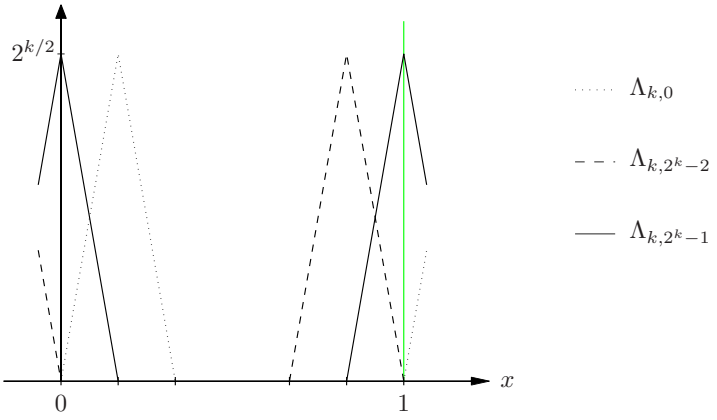


Abb. 57.1: 1-periodische lineare Splines

57 Semiorthogonale Spline-Wavelets

Wegen der schlechten Approximationsordnung der Treppenfunktionen wird das Haar-Wavelet nur selten in der Praxis eingesetzt. Da in der Numerischen Mathematik lineare oder kubische Splines bevorzugt werden, stellen wir nun Multiskalenbasen für lineare Splines vor.

Bei linearen Splines übernimmt die Hutfunktion

$$\Lambda(x) = \begin{cases} x\,, & 0 < x \le 1\,, \\ 2-x\,, & 1 < x \le 2\,, \\ 0\,, & \text{sonst}\,, \end{cases}$$

in natürlicher Weise die Rolle der charakteristischen Funktion aus dem vorigen Abschnitt. Dabei ist lediglich zu beachten, daß ihr Träger $\operatorname{supp}(\Lambda) = [0,2]$ zwei Gitterintervalle überdeckt. Zudem ergeben sich gewisse Probleme am Rand des Splinegitters, da die „abgeschnittenen“ Randfunktionen aus Abschnitt 44 nicht die Form der anderen Hutfunktionen haben. Wir beschränken uns hier daher auf periodische lineare Splines.

Im weiteren bezeichne also V_k, $k \ge 1$, den Raum der linearen Splines über

$$\widetilde{\triangle}_k = \{jh_k \,:\, j \in \mathbb{Z}, h_k = 2^{-k}\}$$

mit Periode 1, vgl. Abbildung 57.1. Analog zu dem vorigen Abschnitt wird die nodale Basis $\{\Lambda_{kj}, j = 0, \ldots, 2^k - 1\}$ von V_k für $x \in [0,1]$ durch

$$\begin{aligned} \Lambda_{k0}(x) = 2^{k/2}\Lambda(2^k x)\,, \qquad \Lambda_{kj}(x + jh_k) = \Lambda_{k0}(x)\,, \\ k = 0, \ldots, p\,, \quad j = 1, \ldots, 2^k - 1\,, \end{aligned} \tag{57.1}$$

definiert und anschließend 1-periodisch fortgesetzt. Besondere Aufmerksamkeit verdient dabei die Basisfunktion $\Lambda_{k,2^k-1}$, denn

$$\operatorname{supp}(\Lambda_{k,2^k-1}) \cap [0,1] \;=\; \mathcal{I}_{k0} \cup \mathcal{I}_{k,2^k-1}\,,$$

vgl. Abbildung 57.1. An dieser Stelle sei festgehalten, daß die nodale Basis keine Orthogonalbasis ist. Tatsächlich muß bei Waveletbasen für lineare Splines weitgehend auf Orthogonalitätseigenschaften verzichtet werden.[2]

Wie im vorigen Abschnitt suchen wir nun einen Komplementärraum W_k zu V_k mit

$$V_{k+1} = V_k \oplus W_k\,.$$

Dieser Komplementärraum soll durch eine Basis $\{\psi_{k,j},\, j = 0,\ldots,2^k-1\}$ aus 1-periodischen linearen Splines über $\widetilde{\triangle}_{k+1}$ aufgespannt werden, die durch eine geeignete Funktion ψ über $\mathbb{R}$ erzeugt werden:

$$\begin{aligned} &\psi_{k0}(x) = 2^{k/2}\psi(2^k x)\,, \qquad \psi_{kj}(x + jh_k) = \psi_{k0}(x)\,, \\ &0 \le x \le 1\,, \quad k = 0,\ldots,p-1\,, \quad j = 1,\ldots,2^k-1\,. \end{aligned} \tag{57.2}$$

Folglich muß ψ ein linearer Spline über dem Referenzgitter $\triangle_*$ aus (56.5) sein. Wir beschränken uns im weiteren auf den Fall, daß ψ kompakten Träger im Intervall $[0,3]$ hat und $k \ge 2$ ist. Damit ist garantiert, daß die Einschränkung von ψ_{k0} auf das Intervall $[0,1]$ eine vollständige gestauchte Kopie von ψ darstellt.

Definition 57.1. Ein linearer Spline ψ über $\triangle_*$ mit $\operatorname{supp}(\psi) \subset [0,3]$ heißt *Wavelet*, falls ein $\alpha > 0$ existiert, so daß der Winkel zwischen den Teilräumen

$$W_k = \operatorname{span}\{\psi_{k,j} \,:\, j = 0,\ldots,2^k-1\} \tag{57.3}$$

und V_k von V_{k+1} für jedes $k \ge 2$ größer als α ist. Die den Räumen V_k zugrundeliegende Funktion Λ heißt *Skalierungsfunktion* (manchmal spricht man auch von *Mutterwavelet* ψ und *Vaterwavelet* Λ). Ist ψ ein Wavelet, dann existiert eine Multiskalenzerlegung $V_p = V_2 \oplus W_2 \oplus W_3 \oplus \cdots \oplus W_{p-1}$. Die zugehörige Basis heißt *Waveletbasis*.

Unter den getroffenen Annahmen ist die Funktion ψ durch fünf Parameter $\psi_1,\ldots,\psi_5$ festgelegt, nämlich die Funktionswerte

$$\psi_j = \psi(j/2)\,, \qquad j = 1,\ldots,5\,.$$

[2]Geht man anstelle der periodischen linearen Splines zu linearen Splines auf unbeschränkten Gittern über, so läßt sich hingegen eine orthogonale Waveletbasis konstruieren, vgl. Chui [17]. Das zugehörige *Battle-Lemarié-Wavelet* hat allerdings keinen kompakten Träger.

Ferner gilt die Darstellung

$$\begin{aligned}\psi(x) \;=\; &\psi_1\Lambda(2x) + \psi_2\Lambda(2x-1) + \psi_3\Lambda(2x-2)\\ &+ \psi_4\Lambda(2x-3) + \psi_5\Lambda(2x-4)\,.\end{aligned} \tag{57.4}$$

Die Funktionen $\{\Lambda_{kj}\}$ und $\{\psi_{kj}\}$ über dem k-ten Gitter können wie in (56.8) durch die nodale Basis des $(k+1)$-ten Gitters ausgedrückt werden, nämlich[3]

$$\begin{aligned}\sqrt{2}\,\Lambda_{kj} \;&=\; \tfrac{1}{2}\,\Lambda_{k+1,2j} + \Lambda_{k+1,2j+1} + \tfrac{1}{2}\,\Lambda_{k+1,2j+2}\,,\\ \sqrt{2}\,\psi_{kj} \;&=\; \psi_1\,\Lambda_{k+1,2j} + \psi_2\,\Lambda_{k+1,2j+1} + \psi_3\,\Lambda_{k+1,2j+2}\\ &\qquad + \psi_4\,\Lambda_{k+1,2j+3} + \psi_5\,\Lambda_{k+1,2j+4}\,.\end{aligned} \tag{57.5}$$

Um diese Darstellung zu erhalten, muß lediglich (57.4) in (57.2) eingesetzt werden.

Aus (57.5) ergibt sich die Basistransformation von der Zweiskalenbasis in die nodale Basis von V_{k+1}. Ist nämlich

$$f \;=\; \sum_{j=0}^{n-1}\Big(\xi_{kj}\Lambda_{kj} + \eta_{kj}\psi_{kj}\Big) \;=\; \sum_{j=0}^{N-1}\xi_{k+1,j}\Lambda_{k+1,j}\,, \qquad n=2^k,\ N=2n\,,$$

dann erfüllen die Entwicklungskoeffizienten nach Aufgabe VI.1 das Gleichungssystem

$$\begin{bmatrix}\xi_{k+1,0}\\ \xi_{k+1,1}\\ \xi_{k+1,2}\\ \xi_{k+1,3}\\ \xi_{k+1,4}\\ \vdots\\ \xi_{k+1,N-3}\\ \xi_{k+1,N-2}\\ \xi_{k+1,N-1}\end{bmatrix} = \frac{1}{\sqrt{2}}\left[\begin{array}{cccccc:cccccc}
1/2 & & & & & 1/2 & \psi_1 & & & & \psi_5 & \psi_3\\
1 & & & & & & \psi_2 & & & & & \psi_4\\
1/2 & 1/2 & & & & & \psi_3 & \psi_1 & & & & \psi_5\\
 & 1 & & & & & \psi_4 & \psi_2 & & & & \\
 & 1/2 & 1/2 & & & & \psi_5 & \psi_3 & \psi_1 & & & \\
 & & \vdots & \ddots & \vdots & & & \vdots & \vdots & \ddots & \vdots & \\
 & & & & 1 & & & & & & \psi_4 & \psi_2\\
 & & & & 1/2 & 1/2 & & & & & \psi_5 & \psi_3\\
 & & & & & 1 & & & & & \psi_4 & \psi_2
\end{array}\right]\begin{bmatrix}\xi_{k,0}\\ \xi_{k,1}\\ \vdots\\ \xi_{k,n-1}\\ \eta_{k,0}\\ \eta_{k,1}\\ \vdots\\ \eta_{k,n-1}\end{bmatrix}. \tag{57.6}$$

Ohne Zusatzforderungen ist der Raum W_k aus (57.3) im allgemeinen kein Komplementärraum von V_k, also ψ kein Wavelet. Um zu erkennen, ob ψ ein Wavelet ist, bilden wir für jedes gerade $j = 0, 2, 4, \ldots, N-2$ die Summen $2\xi_{k+1,j} - \xi_{k+1,j-1} - \xi_{k+1,j+1}$ und reduzieren (57.6) zu dem $n \times n$ linearen Gleichungssystem

$$\begin{bmatrix}2\xi_{k+1,2} - \xi_{k+1,1} - \xi_{k+1,3}\\ 2\xi_{k+1,4} - \xi_{k+1,3} - \xi_{k+1,5}\\ \vdots\\ 2\xi_{k+1,N-2} - \xi_{k+1,N-3} - \xi_{k+1,N-1}\\ 2\xi_{k+1,0} - \xi_{k+1,N-1} - \xi_{k+1,1}\end{bmatrix} = \sqrt{2}\begin{bmatrix}c_0 & c_1 & & & c_{-1}\\ c_{-1} & c_0 & c_1 & & \\ & \ddots & \ddots & \ddots & \\ & & c_{-1} & c_0 & c_1\\ c_1 & & & c_{-1} & c_0\end{bmatrix}\begin{bmatrix}\eta_{k,0}\\ \eta_{k,1}\\ \vdots\\ \eta_{k,n-2}\\ \eta_{k,n-1}\end{bmatrix} \tag{57.7}$$

[3]Im weiteren soll der zweite Index j bei einer Basisentwicklung in $\{\Lambda_{kj}\}_j$ bzw. $\{\psi_{kj}\}_j$ immer modulo 2^k verstanden werden. Beispielsweise ist $\xi_{k,-1} = \xi_{k,n-1}$ für $n = 2^k$.

mit

$$c_{-1} = \psi_5 - \psi_4/2, \quad c_0 = \psi_3 - \psi_4/2 - \psi_2/2, \quad c_1 = \psi_1 - \psi_2/2\,. \tag{57.8}$$

Der folgende Satz gibt nun ein einfaches Kriterium dafür, daß ψ ein Wavelet ist.

Satz 57.2. *Seien $\psi_1, \ldots, \psi_5 \in \mathbb{R}$ gegeben und c_{-1}, c_0 und c_1 durch* (57.8) *definiert. Falls das trigonometrische Polynom*

$$t_\psi(\theta) \;=\; c_{-1}e^{-\mathrm{i}\theta} + c_0 + c_1 e^{\mathrm{i}\theta} \tag{57.9}$$

keine rellen Nullstellen besitzt, dann ist der lineare Spline ψ aus (57.4) *ein Wavelet.*

Beweis. Wir müssen zeigen, daß ein $\delta > 0$ existiert, so daß für jedes $k \geq 2$ und zwei beliebige von Null verschiedene Funktionen $v \in V_k$ und $w \in W_k$ die Ungleichung

$$\frac{\langle\, v, w \,\rangle_{\mathcal{L}^2(0,1)}}{\|v\|_{\mathcal{L}^2(0,1)}\,\|w\|_{\mathcal{L}^2(0,1)}} \;\leq\; 1 - \delta \tag{57.10}$$

erfüllt ist, denn dann ist auch der Kosinus des Winkels zwischen V_k und W_k kleiner als $1 - \delta$. Dabei können wir uns auf Funktionen $v \in V_k$ und $w \in W_k$ mit $\|v\|_{\mathcal{L}^2(0,1)} = \|w\|_{\mathcal{L}^2(0,1)} = 1$ beschränken. Wir entwickeln

$$v \;=\; \sum_{j=0}^{n-1} \xi_{kj}\Lambda_{kj}\,, \quad w \;=\; -\sum_{j=0}^{n-1} \eta_{kj}\psi_{kj}\,, \qquad n = 2^k,\; N = 2n\,,$$

und definieren

$$u \;=\; v - w \;=\; \sum_{j=0}^{N-1} \xi_{k+1,j}\Lambda_{k+1,j} \;\in\; V_{k+1}\,.$$

$\boldsymbol{u} = [\xi_{k+1,0}, \ldots, \xi_{k+1,N-1}]^T \in \mathbb{R}^N$ und $\boldsymbol{w} = [\eta_{k0}, \ldots, \eta_{k,n-1}]^T \in \mathbb{R}^n$. Angesichts der Normierung der $\Lambda_{k+1,j}$ folgt mit der zugehörigen Gramschen Matrix G, daß

$$\|u\|^2_{\mathcal{L}^2(0,1)} \;=\; \boldsymbol{u}^T G \boldsymbol{u} \;\geq\; \frac{1}{3}\|\boldsymbol{u}\|_2^2\,, \tag{57.11}$$

denn das Spektrum dieser (symmetrischen) Matrix liegt nach dem Satz 23.1 von Gerschgorin in dem Intervall $[1/3, 1]$, vgl. Aufgabe 2.

Die Matrix C aus (57.7) ist eine zirkulante Matrix, deren Eigenwerte λ_j sich nach Satz 54.2 und (54.3) durch eine diskrete Fouriertransformation der ersten Spalte ergeben:

$$\lambda_j \;=\; c_0 + c_{-1}e^{-\mathrm{i}j2\pi/n} + c_1 e^{\mathrm{i}j2\pi/n} \;=\; t_\psi(j2\pi/n)\,, \qquad j = 0,\dots,n-1\,.$$

Nach Voraussetzung ist $|t_\psi(\theta)| \geq \varepsilon$ für ein $\varepsilon > 0$ und alle $\theta \in \mathbb{R}$, d. h. C ist invertierbar. Genauer ist

$$\|C^{-1}\|_2 \;=\; \max_{j=0,\dots,n-1} |\lambda_j|^{-1} \;\leq\; 1/\varepsilon\,,$$

denn $|\lambda_j|$, $j = 0,\dots,n-1$, sind gerade die Singulärwerte von C. Um dies zu sehen, beachten wir, daß nach Satz 54.2

$$C^*C \;=\; \frac{1}{n^2}F^*D^*FF^*DF \;=\; \frac{1}{n}F^*\Sigma F\,, \qquad \Sigma = D^*D\,,$$

wobei F die (bis auf den Faktor $\sqrt{n}$ unitäre) Fourier-Matrix ist. Somit können die Quadrate der Singulärwerte von C von der Diagonalmatrix Σ abgelesen werden. Aus (57.7) und (57.11) folgt somit

$$\|\boldsymbol{w}\|_2 \;\leq\; \frac{4}{\sqrt{2}\,\varepsilon}\,\|\boldsymbol{u}\|_2 \;\leq\; \frac{2\sqrt{6}}{\varepsilon}\,\|u\|_{\mathcal{L}^2(0,1)}\,. \tag{57.12}$$

Betrachten wir nun die Gramsche Matrix $W = [w_{ij}]$ der Funktionen $\{\psi_{kj}\}_j$. Wegen $\mathrm{supp}(\psi) \subset [0,3]$ ist W eine zirkulante Matrix mit maximal fünf von Null verschiedenen Diagonalen, deren Einträge nach der Cauchy-Schwarz-Ungleichung allesamt durch

$$|w_{ij}| \;=\; |\langle\, \psi_{ki}, \psi_{kj}\,\rangle_{\mathcal{L}^2(0,1)}| \;\leq\; \|\psi_{ki}\|_{\mathcal{L}^2(0,1)}\|\psi_{kj}\|_{\mathcal{L}^2(0,1)} \;=\; \|\psi\|^2_{\mathcal{L}^2(0,3)}$$

beschränkt sind. Nach dem Satz von Gerschgorin ist somit $\|W\|_2 \leq 5\|\psi\|^2_{\mathcal{L}^2(0,3)}$ und es folgt

$$1 \;=\; \|w\|^2_{\mathcal{L}^2(0,1)} \;=\; \boldsymbol{w}^T W\boldsymbol{w} \;\leq\; 5\|\psi\|^2_{\mathcal{L}^2(0,3)}\|\boldsymbol{w}\|^2_2\,.$$

Zusammen mit (57.12) ergibt sich hieraus die Ungleichung

$$\|u\|^2_{\mathcal{L}^2(0,1)} \;\geq\; \frac{\varepsilon^2}{120}\,\|\psi\|^{-2}_{\mathcal{L}^2(0,3)} \;=:\; 2\delta > 0\,.$$

Damit gilt

$$2\langle\, v, w\,\rangle_{\mathcal{L}^2(0,1)} \;=\; \|v\|^2_{\mathcal{L}^2(0,1)} + \|w\|^2_{\mathcal{L}^2(0,1)} - \|u\|^2_{\mathcal{L}^2(0,1)} \leq\; 2 - 2\delta$$

und die Behauptung (57.10) ist nachgewiesen. □

Wie wir im Beweis dieses Satzes gesehen haben, ist die Matrix C des Gleichungssystems (57.7) unter der genannten Vorraussetzung invertierbar und das Gleichungssystem kann zur Berechnung der $\{\eta_{kj}\}_j$ genutzt werden. Aus den Gleichungen (57.6) für $\xi_{k+1,2j+1}$ können die fehlenden Koeffizienten $\{\xi_{kj}\}_j$ durch Rücksubstitution berechnet werden:

$$\xi_{kj} \;=\; \sqrt{2}\,\xi_{k+1,2j+1} - \psi_4\eta_{k,j-1} - \psi_2\eta_{kj}\,, \qquad j = 0,\dots,2^k-1\,. \quad (57.13)$$

Aufwand. Die Lösung des linearen Gleichungssystems (57.7) kann mit der Gauß-Elimination erfolgen. Das Aufstellen dieses linearen Gleichungssystems erfordert n Multiplikationen und $2n$ Additionen. Für die Gauß-Elimination sind etwa $8n+3n$ multiplikative Operationen notwendig. Die Berechnung der $\{\xi_{kj}\}_j$ gemäß (57.13) kostet schließlich weitere $3n$ Multiplikationen. Zusammen ergibt das ungefähr $15n$ Multiplikationen. Die Rücktransformation in die nodale Basis geschieht hingegen einfach via (57.6). Da die entsprechende Matrix höchstens $8n$ von Null verschiedene Einträge hat, kostet diese Transformation maximal $8n$ Multiplikationen.

Die vollständige Transformation von der nodalen Basis in die Multiskalenbasis läßt sich somit durch $\sum_{k=2}^{p-1} 15\cdot 2^k \;\approx\; 15\cdot 2^p$ multiplikative Operationen bewerkstelligen. Die Rücktransformation kostet etwa $8\cdot 2^p$ Multiplikationen. $\lozenge$

Welche Werte $\psi_1,\dots,\psi_5$ ergeben nun ein sinnvolles Wavelet? Als erstes Beispiel wählen wir die Parameter derart, daß die Teilräume V_k und W_k zueinander senkrecht sind. Damit Λ_{kj} und $\psi_{kj'}$ bei festem k für alle j und j' paarweise orthogonal sind, müssen die vier Orthogonalitätsbedingungen

$$\Lambda \perp \psi(\,\cdot - j)\,, \qquad j = -2,-1,0,1\,,$$

erfüllt sein. Für alle anderen Werte von j haben die genannten Funktionen disjunkte Träger und sind daher trivialerweise orthogonal. Für diese vier Innenprodukte verwenden wir Bemerkung 31.9, nach der das Innenprodukt zweier linearer Splines ϕ und φ über dem Referenzgitter $\triangle_*$ aus (56.5) mit $\operatorname{supp}(\phi\varphi) = [0,2]$ durch die Formel

$$\int_0^2 \phi(x)\varphi(x)\,dx \;=\; \frac{1}{12}\,\begin{bmatrix}\phi_0 & \phi_1 & \dots & \phi_4\end{bmatrix}\begin{bmatrix}2 & 1 & & 0\\ 1 & 4 & \ddots & \\ & \ddots & 4 & 1\\ 0 & & 1 & 2\end{bmatrix}\begin{bmatrix}\varphi_0\\ \varphi_1\\ \vdots\\ \varphi_4\end{bmatrix} \quad (57.14)$$

gegeben ist, wobei $\phi_i = \phi(i/2)$ und $\varphi_i = \varphi(i/2)$, $i = 0,\dots,4$. Die vier Gleichungen $\langle\,\Lambda, \psi(\,\cdot - j)\,\rangle = 0$ für $j = -2,-1,0,1$, führen somit auf das lineare

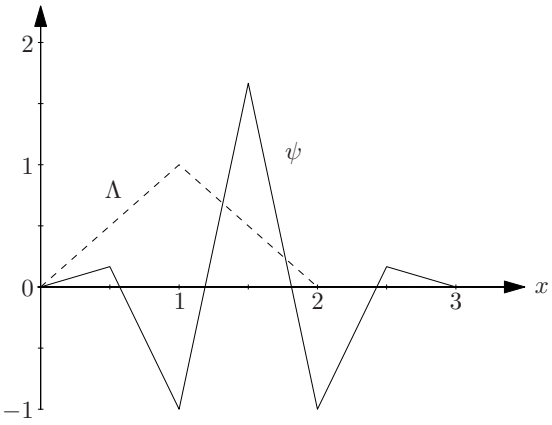


Abb. 57.2: Skalierungsfunktion Λ und semiorthogonales Wavelet ψ

Gleichungssystem

$$\begin{array}{rcrcrcrcrcl}
3\,\psi_1 & + & 1/2\,\psi_2 & & & & & & & = & 0\,, \\
3\,\psi_1 & + & 5\,\psi_2 & + & 3\,\psi_3 & + & 1/2\,\psi_4 & & & = & 0\,, \\
 & & 1/2\,\psi_2 & + & 3\,\psi_3 & + & 5\,\psi_4 & + & 3\,\psi_5 & = & 0\,, \\
 & & & & & & 1/2\,\psi_4 & + & 3\,\psi_5 & = & 0\,,
\end{array}$$

für die Koeffizienten $\psi_j = \psi(j/2)$, $j = 1, \ldots, 5$. Es folgt $\psi_2 = -6\psi_1$, $\psi_4 = -6\psi_5$ und die verbliebenen beiden Gleichungen vereinfachen sich zu

$$\begin{array}{rcrcrcl}
-27\,\psi_1 & + & 3\,\psi_3 & - & 3\,\psi_5 & = & 0\,, \\
-3\,\psi_1 & + & 3\,\psi_3 & - & 27\,\psi_5 & = & 0\,.
\end{array}$$

Durch Elimination von ψ_3 ergibt sich unmittelbar $\psi_1 = \psi_5$ und damit die (bis auf einen multiplikativen Faktor eindeutige) Lösung

$$\psi_1 = 1/6, \quad \psi_2 = -1, \quad \psi_3 = 5/3, \quad \psi_4 = -1, \quad \psi_5 = 1/6\,.$$

Die zugehörige Funktion ψ ist in Abbildung 57.2 dargestellt. Man beachte wieder den Wellencharakter dieser Funktion.

Nach Satz 57.2 ist ψ ein Wavelet, da das trigonometrische Polynom

$$t_\psi(\theta) \;=\; 8/3 + 4/3\cos\theta$$

aus (57.9) für alle $\theta \in \mathbb{R}$ nach unten durch $4/3$ beschränkt ist. Die Unterräume W_k und V_k sind zueinander orthogonal, die Basisfunktionen ψ_{kj} von W_k jedoch nicht paarweise orthogonal; daher nennt man ψ *semiorthogonales Wavelet*.

58 Biorthogonale Spline-Wavelets

Die durch das semiorthogonale Wavelet ψ aus Abbildung 57.2 erzeugte Waveletbasis hat einen Nachteil, wie wir später sehen werden: Zur Darstellung eines linearen Splines mit kleinem Träger (etwa einer Hutfunktion $\Lambda_{pj} \in V_p$) sind in der Regel alle Basiselemente dieser Waveletbasis notwendig. Dies hängt damit zusammen, daß die Inverse der entsprechenden zirkulanten Matrix in (57.7) voll besetzt ist. Die örtliche Lokalisierungseigenschaft des semiorthogonalen Wavelets ist also nicht so gut wie die des Haar-Wavelets.

Definition 58.1. Wir sagen, daß ein lineares Spline-Wavelet ψ die *Lokalisierungseigenschaft* besitzt, falls eine Konstante $l \in \mathbb{N}_0$ und reelle Koeffizienten α_j, β_j, $\tilde{\alpha}_j$, $\tilde{\beta}_j$, $-l \le j \le l$, existieren mit

$$\begin{aligned} \Lambda(2x) &= \sum_{j=-l}^{l} \big(\alpha_j \Lambda(x-j) + \beta_j \psi(x-j)\big)\,, \\ \Lambda(2x-1) &= \sum_{j=-l}^{l} \big(\tilde{\alpha}_j \Lambda(x-j) + \tilde{\beta}_j \psi(x-j)\big)\,. \end{aligned} \tag{58.1}$$

Besitzt ein Wavelet die Lokalisierungseigenschaft, so folgen aus (57.1) und (57.2) unmittelbar entsprechende Darstellungen von $\Lambda_{k+1,0}$ und $\Lambda_{k+1,1}$ in der Zweiskalenbasis von V_{k+1} (vorausgesetzt, daß 2^k größer als $2l+2$ ist):

$$\Lambda_{k+1,0} = \sqrt{2} \sum_{j=-l}^{l} \big(\alpha_j \Lambda_{kj} + \beta_j \psi_{kj}\big)\,, \quad \Lambda_{k+1,1} = \sqrt{2} \sum_{j=-l}^{l} \big(\tilde{\alpha}_j \Lambda_{kj} + \tilde{\beta}_j \psi_{kj}\big)\,.$$

Da sich die anderen Basisfunktionen $\Lambda_{k+1,j}$ mit $j \ge 2$ durch geeignete Verschiebungen dieser beiden Funktionen ergeben, vgl. (57.1), gelten entsprechende Darstellungen für alle Basisfunktionen der nodalen Basis von V_{k+1}, d. h. jedes Element der nodalen Basis von V_{k+1} kann durch eine feste Anzahl von Basisfunktionen der entsprechenden Zweiskalenbasis ausgedrückt werden. Aus diesem Grund beeinflußt eine Störung einer Funktion $f \in V_p$ in einem einzigen Gitterpunkt $x \in \triangle_p$ lediglich die Koeffizienten jener Basisfunktionen der Waveletbasis, deren Träger in der Nachbarschaft von x liegt. Wavelets, die die Lokalisierungseigenschaft besitzen, lassen sich wie folgt charakterisieren:

Satz 58.2. *Ein lineares Spline-Wavelet ψ über $\triangle_*$ mit* $\operatorname{supp}\psi \subset [0,3]$ *und Werten $\psi_j = \psi(j/2)$, $j = 1,\dots,5$, hat genau dann die Lokalisierungseigenschaft, falls zwei der Parameter c_{-1}, c_0 und c_1 aus* (57.8) *gleich Null sind.*

Beweis. Der Beweis beruht auf der Beobachtung, daß die zirkulante Matrix aus (57.7) genau dann eine Bandmatrix als Inverse besitzt, wenn sie ein Vielfaches der Einheits- oder der (zirkulanten) Shiftmatrix aus Aufgabe V.13 ist.

Wir nehmen zunächst an, daß zwei der drei Parameter gleich Null sind und zeigen, daß ψ dann die Lokalisierungseigenschaft besitzt. Dabei beschränken wir uns auf den Fall, daß $c_{-1} = c_1 = 0$ und c_0 von Null verschieden ist. Die anderen beiden Fälle können analog behandelt werden. Gemäß (57.8) gilt dann

$$\psi_2 = 2\psi_1 \quad \text{und} \quad \psi_4 = 2\psi_5 \tag{58.2}$$

und aus (57.7) folgt unmittelbar

$$\eta_{kj} \;=\; \frac{1}{c_0\sqrt{2}} \left(2\,\xi_{k+1,2(j+1)} - \xi_{k+1,2(j+1)-1} - \xi_{k+1,2(j+1)+1}\right) \tag{58.3}$$

für $j = 0, \ldots, 2^k - 1$. Speziell für die Entwicklungskoeffizienten $\xi_{k+1,j} = \delta_{j0}$ von $\Lambda_{k+1,0}$ bezüglich der nodalen Basis von V_{k+1} ergibt dies mit $n = 2^k$

$$\eta_{k,n-1} = \sqrt{2}/c_0 \qquad \text{und} \qquad \eta_{kj} = 0 \;\; \text{für} \;\; j \neq n-1$$

und aus (57.13) folgt ferner

$$\xi_{k0} = -\,\frac{\psi_4\sqrt{2}}{c_0}\,, \quad \xi_{k,n-1} = -\,\frac{\psi_2\sqrt{2}}{c_0}\,, \quad \xi_{kj} = 0 \;\; \text{für} \;\; j \notin \{0, n-1\}\,.$$

Damit hat $\Lambda_{k+1,0}$ die Basisentwicklung

$$\Lambda_{k+1,0} \;=\; -\,\frac{\sqrt{2}}{c_0} \left(\psi_2\Lambda_{k,n-1} + \psi_4\Lambda_{k0} - \psi_{k,n-1}\right)$$

und die erste Gleichung aus (58.1) gilt mit $l = 1$ und $\alpha_{-1} = -\psi_2/c_0$, $\alpha_0 = -\psi_4/c_0$, $\beta_{-1} = 1/c_0$, $\beta_0 = \beta_1 = \alpha_1 = 0$. Entsprechend ergibt sich

$$\Lambda_{k+1,1} \;=\; \frac{1}{c_0\sqrt{2}} \Big(\psi_2\Lambda_{k,n-1} + (2c_0 + \psi_4 + \psi_2)\Lambda_{k0} + \psi_4\Lambda_{k1} - \psi_{k,n-1} - \psi_{k0}\Big)$$

und die zweite Gleichung von (58.1), wiederum mit $l = 1$. Mit anderen Worten, das Wavelet ψ erfüllt die Lokalisierungseigenschaft.

Hat umgekehrt das Wavelet ψ die Lokalisierungseigenschaft, dann existieren nach Definition 58.1 ein $l \in \mathbb{N}_0$ und Koeffizienten α_j und β_j mit

$$\Lambda_{k+1,0} \;=\; \sqrt{2} \sum_{j=-l}^{l} \big(\alpha_j\Lambda_{kj} \;+\; \beta_j\psi_{kj}\big)$$

für hinreichend große k, etwa für $2^k > 2l+2$. Gemäß (57.7) erfüllen die β_j dann das folgende lineare Gleichungssystem:

$$\begin{bmatrix} 0 \\ 0 \\ 0 \\ \vdots \\ 0 \\ 2 \end{bmatrix} = 2 \begin{bmatrix} c_0 & c_1 & & & & c_{-1} \\ c_{-1} & c_0 & c_1 & & & \\ & c_{-1} & c_0 & c_1 & & \\ & & \ddots & \ddots & \ddots & \\ & & & c_{-1} & c_0 & c_1 \\ c_1 & & & & c_{-1} & c_0 \end{bmatrix} \begin{bmatrix} \beta_0 \\ \vdots \\ \beta_l \\ 0 \\ \vdots \\ 0 \\ \beta_{-l} \\ \vdots \\ \beta_{-1} \end{bmatrix} . \tag{58.4}$$

Ohne Einschränkung sei angenommen, daß β_l und β_{-l} nicht beide verschwinden.

Wir unterscheiden im folgenden zwei Fälle. Falls $\beta_l \neq 0$ ist, lautet die $(l+2)$-te Gleichung des Gleichungssystems (58.4)

$$0 = 2(c_{-1}\beta_l + c_0 \cdot 0 + c_1 \cdot 0)$$

und es folgt $c_{-1} = 0$; die $(l+1)$-te Gleichung ergibt dann

$$0 = 2(0 \cdot \beta_{l-1} + c_0\beta_l + c_1 \cdot 0)$$

und daher ist auch $c_0 = 0$. In diesem Fall ist also lediglich c_1 von Null verschieden und aufgrund der letzten Gleichung von (58.4) ist $l = 0$ und $\beta_0 = 1/c_1$.

Im zweiten Fall ist $\beta_l = 0$ und $\beta_{-l} \neq 0$. Insbesondere ist daher $l > 0$ und die $(n-l)$-te Gleichung lautet

$$0 = 2(c_{-1} \cdot 0 + c_0 \cdot 0 + c_1\beta_{-l}) ,$$

d. h. $c_1 = 0$. An dieser Stelle ist eine weitere Fallunterscheidung nötig. Falls $l \geq 2$ ist, ergibt die $(n-l+1)$-te Gleichung

$$0 = 2(c_{-1} \cdot 0 + c_0\beta_{-l} + 0 \cdot \beta_{1-l}) .$$

Demnach ist $c_0 = 0$ und $c_{-1} \neq 0$ und aus der letzten Gleichung ergibt sich zwangsläufig $l = 2$ und $\beta_{-2} = 1/c_{-1}$. Im anderen Fall ist $l = 1$ und die n-te Gleichung führt auf die Bedingung

$$2 = 2(0 \cdot \beta_0 + c_{-1} \cdot 0 + c_0\beta_{-1}) .$$

In diesem Fall ist also $\beta_{-1} = 1/c_0$, d. h. c_0 muß von Null verschieden sein. Wäre nun $c_{-1} \neq 0$, dann folgt aus der zweiten Gleichung wegen $\beta_1 = 0$

$$0 = 2(c_{-1}\beta_0 + c_0 \cdot 0 + 0)$$

und somit $\beta_0 = 0$, d. h. die erste Zeile ergibt einen Widerspruch:

$$0 \;=\; 2(c_0 \cdot 0 + 0 + c_{-1}\beta_{-1}) \;\; /= .0$$

Somit ist in diesem letzten Fall $c_{-1} = 0$ und lediglich c_0 von Null verschieden. □

Für das semiorthogonale Wavelet aus Abbildung 57.2 sind alle drei Parameter c_{-1}, c_0 und c_1 von Null verschieden. Daher hat das semiorthogonale Wavelet *nicht* die Lokalisierungseigenschaft.

Die Bedingung $c_{-1} = c_1 = 0$ ist äquivalent zu den beiden Gleichungen (58.2) und diese stellen zwei lineare Bedingungen an die fünf gesuchten Koeffizienten ψ_j, $j = 1, \ldots, 5$, von ψ dar. Da ψ ohnehin höchstens bis auf (multiplikative) Normierung eindeutig ist, können noch zwei weitere Nebenbedingungen an ψ gestellt werden, zum Beispiel Orthogonalitätsbedingungen. Dabei ist aber zu beachten, daß diese Bedingungen nicht auf $c_0 = 0$ führen dürfen.

Zwei weitere Nebenbedingungen an ψ sind allerdings zu wenig, um Orthogonalität von ψ zu allen Hutfunktionen über dem Referenzgitter $\triangle_*$ zu erzwingen; dazu wären mindestens drei Nebenbedingungen notwendig. Daher gibt man andere Orthogonalitätsbedingungen vor, etwa $\psi \perp \Pi_1$. (Man beachte, daß Π_1 ein Teilraum der linearen Splines ist, aber kein Teilraum der periodischen linearen Splines. Dennoch hat diese Orthogonalitätsforderung in der Waveletliteratur Tradition, vgl. [70].) Π_1 wird von der Konstanten $y = 1$ und der Geraden $y = x$ erzeugt. Die beiden Orthogonalitätsbedingungen lauten also

$$\int_0^3 \psi(x)\,dx = 0 \qquad \text{und} \qquad \int_0^3 x\psi(x)\,dx = 0 \tag{58.5}$$

und führen analog zu (57.14) auf die beiden Bedingungsgleichungen

$$\begin{aligned} \psi_1 + \;\; \psi_2 + \;\; \psi_3 + \;\; \psi_4 + \;\; \psi_5 \;&=\; 0\,, \\ \psi_1 + 2\,\psi_2 + 3\,\psi_3 + 4\,\psi_4 + 5\,\psi_5 \;&=\; 0\,, \end{aligned}$$

vgl. Aufgabe 9. Zusammen mit (58.2) ergibt dies die (bis auf Vielfache) eindeutig bestimmte Lösung

$$\psi_1 = -1/6\,, \quad \psi_2 = -1/3\,, \quad \psi_3 = 1\,, \quad \psi_4 = -1/3\,, \quad \psi_5 = -1/6\,.$$

Damit ist das trigonometrische Polynom t_ψ aus (57.9) konstant gleich $c_0 = 4/3$, und der lineare Spline ψ mit den obigen Koeffizienten nach Satz 57.2 ein Wavelet, das sogenannte *biorthogonale Wavelet* (vgl. Abbildung 58.1).

Aufwand. Die Transformation von der nodalen Basis in die Zweiskalenbasis via (58.3) und (57.13) kann mit $4n$ multiplikativen Operationen implementiert werden. Die Rücktransformation über (57.6) kostet $5n$ Multiplikationen. Der Aufwand ist also geringer als für das semiorthogonale Wavelet. ◊

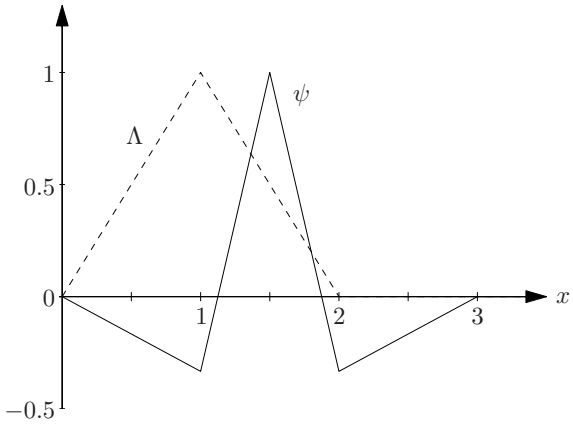


Abb. 58.1: Skalierungsfunktion Λ und biorthogonales Wavelet ψ

59 Ein Anwendungsbeispiel

Eine Punktladung in einem Punkt $y_* \in \mathbb{R}^3$ erzeugt ein elektrisches Potential[4]

$$u_*(x) \;=\; G(x, y_*) \;=\; \frac{1}{4\pi}\,\frac{1}{|x - y_*|}\,, \qquad x \in \mathbb{R}^3 \setminus \{y_*\}\,, \tag{59.1}$$

und das zugehörige elektrische Feld

$$E_*(x) \;=\; -\operatorname{grad} u_*(x) \;=\; \frac{1}{4\pi}\,\frac{x - y_*}{|x - y_*|^3}\,.$$

Wird in dieses elektrische Feld ein elektrischer Isolator $\Omega \subset \mathbb{R}^3$ platziert, daß den Punkt y_* nicht enthält, so ergibt sich ein neues Potential u_0, dessen Höhenlinien (die Feldlinien des zugehörigen elektrischen Felds) senkrecht in den Rand Γ von Ω münden, vgl. Beispiel 70.1. Bezeichnet ν den äußeren Normalenvektor auf Γ, so entspricht diese geometrische Beschreibung der Feldlinien der Gleichung

$$\frac{\partial u_0}{\partial \nu}(x) \;=\; 0\,, \qquad x \in \Gamma\,, \tag{59.2}$$

für die Richtungsableitung von u_0 senkrecht zum Rand, die sogenannte *Normalenableitung*.

[4]In diesem Abschnitt bezeichnet $|x|$ für $x = (x_1, x_2, x_3) \in \mathbb{R}^3$ immer die Euklidnorm im $\mathbb{R}^3$, d. h. $|x|^2 = x_1^2 + x_2^2 + x_3^2$.

In Beispiel 70.1 werden wir herleiten, daß jedes elektrische Potential w in einem homogenen Medium die *Laplace-Gleichung*

$$\Delta w = \frac{\partial^2}{\partial x_1^2} w + \frac{\partial^2}{\partial x_2^2} w + \frac{\partial^2}{\partial x_3^2} w = 0 ,$$

löst; Δ ist der *Laplace-Operator*. Diese Behauptung kann für die sogenannte *Grundlösung* $G(\,\cdot\,, y_*)$ aus (59.1) in $\mathbb{R}^3 \setminus \{y_*\}$ leicht nachgerechnet werden. Die Differenz $u = u_0 - u_*$ löst wegen $\Delta u = \Delta u_* - \Delta u_0$ ebenfalls die Laplace-Gleichung: aus (59.2) folgt

$$\begin{aligned} &\Delta u = 0 \quad \text{in } \mathbb{R}^3 \setminus \overline{\Omega} , \\ &\frac{\partial u}{\partial \nu} = g := -\frac{\partial u_*}{\partial \nu} \quad \text{auf } \Gamma = \partial\Omega . \end{aligned} \tag{59.3}$$

(59.3) ist eine (elliptische) partielle Differentialgleichung; die Randbedingung für die Richtungsableitung von u wird *Neumann-Bedingung* genannt. In Kapitel XVI behandeln wir numerische Methoden zur Lösung derartiger Probleme, falls die Differentialgleichung in einem beschränkten Gebiet zu lösen ist. Hier ist die Lösung in einem *unbeschränkten Gebiet* gesucht; es handelt sich um ein sogenanntes *Außenraumproblem*. Ist g eine stetige Funktion mit $\int_\Gamma g\,ds = 0$ (dies kann man für das obige Beispiel zeigen), so hat das Problem (59.3) genau eine physikalisch sinnvolle Lösung u, d. h. eine Lösung u, die für $|x| \to \infty$ verschwindet.

Diese Lösung kann mit der sogenannten *Randintegralmethode* bestimmt werden. Dazu macht man einen Lösungsansatz

$$u(x) = \int_\Gamma G(x,y)\varphi(y)\,ds(y) , \qquad x \in \mathbb{R}^3 \setminus \overline{\Omega} , \tag{59.4}$$

in Form eines Kurvenintegrals (bezüglich der y-Variablen) über den Rand des Gebiets. Man kann (59.4) als eine Überlagerung von Punktladungen auf dem Rand von Γ interpretieren, einem sogenannten *Einfachschichtpotential*; φ ist die zugehörige Ladungsdichte. Da $G(\,\cdot\,, y)$ die Laplace-Gleichung für $x \neq y$ löst, ergibt sich aus (59.4) durch Vertauschung von Differentiation und Integration, daß u wie gewünscht die Laplace-Gleichung in $\mathbb{R}^3 \setminus \overline{\Omega}$ erfüllt.

Die Ladungsdichte φ ist so zu wählen, daß u auch die Neumann-Randbedingung aus (59.3) erfüllt. Will man jedoch durch Differentiation von (59.4) die entsprechende Richtungsableitung von u bestimmen, so treten wegen der Singularität des Integranden für $x = y$ Schwierigkeiten auf, die im Rahmen dieses Buches nicht näher erläutert werden können. Es ergibt sich das folgende Resultat, für dessen Beweis auf das Buch von Kreß [64, Theorem 6.27] verwiesen wird:

Satz 59.1. *Die Funktion u aus* (59.4) *löst das Außenraumproblem* (59.3)*, falls φ eine stetige Lösung der* Integralgleichung

$$-\frac{1}{2}\,\varphi(x) \,+\, \int_\Gamma \frac{\partial G(x,y)}{\partial \nu(x)}\,\varphi(y)\,ds(y) \;=\; g(x)\,, \qquad x \in \Gamma\,, \tag{59.5}$$

ist. Dabei ist $\partial G/\partial\nu(x)$ die Normalenableitung von G bezüglich der x-Variablen im Punkt $x \in \Gamma$.

Für die Anwendung der Wavelet-Basen betrachten wir das entsprechende Problem im $\mathbb{R}^2$. In diesem Fall lautet die Grundlösung

$$G(x,y) \;=\; \frac{1}{2\pi}\,\log\,\frac{1}{|x-y|}$$

mit dem zugehörigen Feld

$$E(x) \;=\; \frac{1}{2\pi}\,\frac{x-y}{|x-y|^2}\,.$$

Für Ω wählen wir eine Ellipse um den Nullpunkt mit Halbachsen der Länge α und β. Mit der Koordinatentransformation

$$\begin{aligned} x_1 &= \alpha\cos\theta\,, \qquad y_1 = \alpha\cos\tau\,, \\ x_2 &= \beta\sin\theta\,, \qquad y_2 = \beta\sin\tau\,, \end{aligned} \qquad 0 \le \tau, \theta < 2\pi\,,$$

für den Ellipsenrand Γ wird aus (59.5) die Integralgleichung

$$-\frac{1}{2}\,\varphi(\theta) \,+\, \int_0^{2\pi} k(\theta,\tau)\varphi(\tau)\,d\tau \;=\; g(\theta) \tag{59.6}$$

über $[0, 2\pi)$ mit *Kernfunktion*

$$k(\theta,\tau) \;=\; -\,\frac{1}{2\pi}\,\Big(\frac{\beta^2\cos^2\tau + \alpha^2\sin^2\tau}{\beta^2\cos^2\theta + \alpha^2\sin^2\theta}\Big)^{1/2}\,\frac{\alpha\beta}{\alpha^2+\beta^2-(\alpha^2-\beta^2)\cos(\tau+\theta)}\,,$$

vgl. Aufgabe 12. Um die Notation nicht unnötig kompliziert zu gestalten, haben wir in (59.6) die Funktionen φ und g auf dem Rand Γ mit den entsprechend substituierten Funktionen über $[0, 2\pi)$ identifiziert.

Wir suchen nun eine Näherung $\varphi_n \in V_p$ an die (periodische) Lösung φ von (59.6), wobei der Ansatzraum V_p hier den Raum der 2π-periodischen linearen Splines über $[0, 2\pi]$ mit $n = 2^p$ äquidistanten Stützstellen bezeichnen soll. Ist $\{\phi_1, \phi_2, \ldots, \phi_n\}$ eine Basis von V_p, dann können wir $\varphi_n \in V_p$ bezüglich dieser Basis entwickeln,

$$\varphi_n = \sum_{i=1}^{n} \zeta_i \phi_i\,. \tag{59.7}$$

Neben der klassischen nodalen Basis verwenden wir im folgenden auch die semiorthogonale und die biorthogonale Waveletbasis.

In jedem Fall werden n Gleichungen benötigt, um die n unbekannten Koeffizienten ζ_i in (59.7) zu bestimmen. Beim *Galerkin-Verfahren* bildet man dazu Innenprodukte der Gleichung (59.6) mit den Basisfunktionen ϕ_i von V_p. Das ergibt das lineare Gleichungssystem

$$Az = b\,, \tag{59.8}$$

mit $z = [\zeta_1, \ldots, \zeta_n]^T$, $b = [\,\langle\, \phi_i, g \,\rangle_{\mathcal{L}^2(0,2\pi)}\,]_i$ und der Matrix

$$A = \big[\, -\frac{1}{2}\,\langle\, \phi_i, \phi_j \,\rangle_{\mathcal{L}^2(0,2\pi)} + \langle\, \phi_i, K\phi_j \,\rangle_{\mathcal{L}^2(0,2\pi)}\,\big]_{i,j}\,, \tag{59.9}$$

wobei K den *Integraloperator*

$$K : \phi \;\mapsto K\phi \;=\; \int_0^{2\pi} k(\tau,\theta)\phi(\tau)\,d\tau \tag{59.10}$$

mit k aus (59.6) bezeichnet.

Die Graustufendarstellungen in Abbildung 59.1 veranschaulichen für die drei genannten Basen von V_8 die Größenordnungen der jeweiligen Einträge der Koeffizientenmatrix $A \in \mathbb{R}^{256\times 256}$ in einer logarithmischen Skala. Die Matrizen gehören zu einer Ellipse Ω mit Halbachsenlängen $\alpha = 10$ und $\beta = 1$. Dabei ist zu beachten, daß in der oberen Abbildung eine andere Graustufeneinteilung verwendet wird als in den unteren beiden (man vergleiche die abgebildeten Skalen).

Generell werden die Matrixeinträge für die Waveletbasen bei den feineren Skalen sehr schnell sehr klein, da die Kernfunktion sehr glatt ist und nur eine kleine Fluktuation auf den feinen Gittern aufweist. Interessant sind darüber hinaus die fingerähnlichen Strukturen bei der Koeffizientenmatrix für die biorthogonale Waveletbasis. Sie beruhen auf den Innenprodukten $\langle\, \phi_i, \phi_j \,\rangle_{\mathcal{L}^2(0,2\pi)}$ der jeweiligen Basisfunktionen in der Matrix A, vgl. (59.9). Dies erkennt man daran, daß diese „Finger“ bei der semiorthogonalen Waveletbasis fehlen, da die Wavelets auf unterschiedlichen Skalen zueinander orthogonal sind.

Abbildung 59.1 läßt vermuten, daß für beide Waveletbasen in jeder Zeile der Koeffizientenmatrix A alle bis auf etwa $O(\log n)$ Einträge ohne merklichen Genauigkeitsverlust durch Null ersetzt werden können. Im Gegensatz dazu kann bei der konventionellen Basisdarstellung praktisch kein Matrixelement vernachlässigt werden. Eine solche Datenkompression spart Speicherplatz und verbilligt den Einsatz iterativer Methoden zur Lösung des Gleichungssystems, da Matrix-Vektor-Produkte mit dünn besetzten Matrizen sehr viel billiger sind.

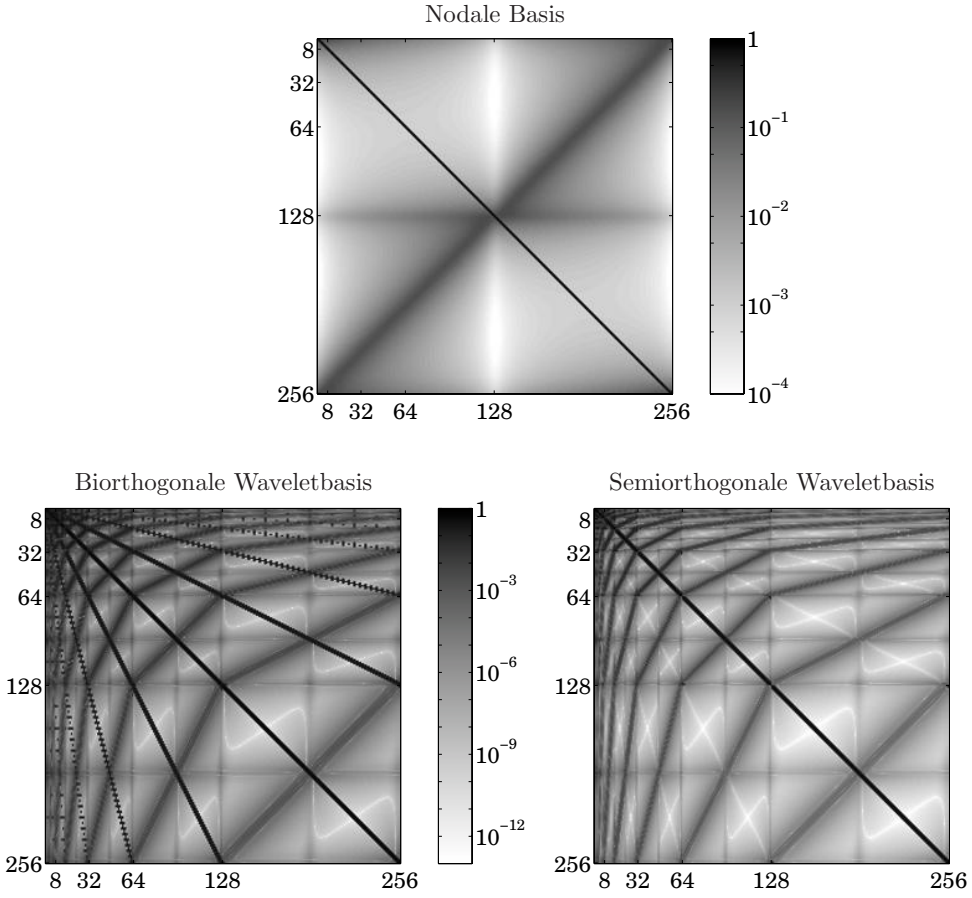


Abb. 59.1: Größe der Einträge der Koeffizientenmatrix A

In dem konkreten Beispiel ist $h = 2\pi/256$ die Gitterweite des zugrundeliegenden Splinegitters. Wegen des Approximationsfehlers $O(h^2)$ linearer Splines ist es unsinnig, das lineare Gleichungssystem (59.8) auf mehr als drei oder vier Stellen Genauigkeit zu lösen. Berücksichtigt man, daß die Kondition $\mathrm{cond}_2(A)$ von A etwas größer als zehn ist, so garantiert die Abbruchbedingung

$$\frac{\|b - Ax^{(k)}\|_2}{\|b\|_2} \lesssim \frac{h^2}{\mathrm{cond}_2(A)} \approx 5 \cdot 10^{-5} \tag{59.11}$$

aufgrund der Abschätzung (2.5) eine Genauigkeit der letzten Iterierten $x^{(k)}$ im Bereich der Approximationsordnung. Die Abschätzung (59.11) legt aber auch nahe, daß entsprechend große Störungen in der Matrix A ohne Genauigkeitsverlust vorgenommen werden können.

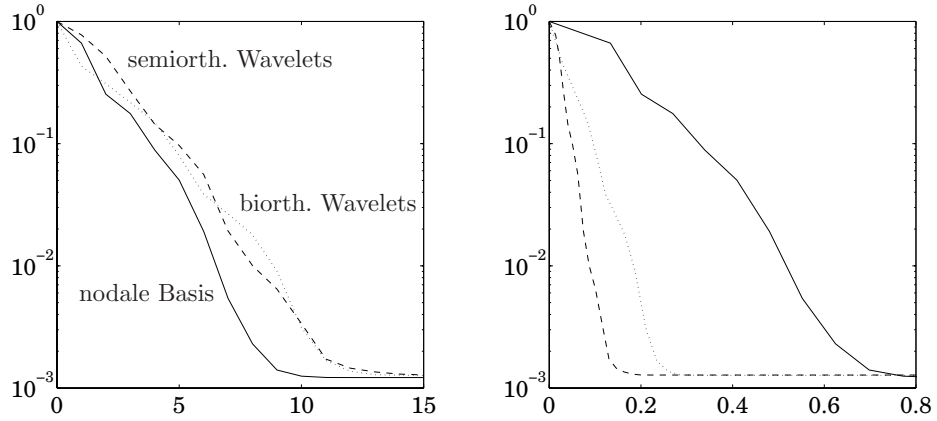


Abb. 59.2: Konvergenzverhalten von GMRES in Abhängigkeit vom Iterationsindex (links) bzw. von der Anzahl Multiplikationen (in Mio., rechts)

Wir ersetzen daher bei den Matrizen alle Elemente, die betragsmäßig kleiner als $5 \cdot 10^{-5}$ sind (relativ zu dem betragsgrößten Element) durch Null. Für die biorthogonale Waveletbasis verbleiben dann lediglich 22% der Matrixelemente ungleich Null, für die semiorthogonale Waveletbasis sogar nur 9%. Für die nodale Basis liegen alle Matrixeinträge *oberhalb* dieser Schranke.

Nun wird das GMRES-Verfahren auf die drei Gleichungssysteme angewendet. Als rechte Seite wählen wir in (59.3) die Funktion g, die zu einer Punktladung im Punkt $y_* = (11, 1)$ gehört. Abbildung 59.2 zeigt die jeweilige Fehlerentwicklung; aufgetragen ist der relative Fehler in der $\mathcal{L}^2$-Norm. Das linke Bild trägt den Fehler in der herkömmlichen Weise über dem Iterationsindex auf, das rechte Bild illustriert die Fehlerentwicklung in Abhängigkeit von der Anzahl der benötigten Multiplikationen. Erwartungsgemäß stagniert der Iterationsfehler für alle drei Matrizen im Bereich des Diskretisierungsfehlers, also bei einem Wert knapp oberhalb von 10^{-3}. Entscheidend ist aber die Tatsache, daß eine Iteration mit den „ausgedünnten“ Matrizen erheblich billiger ist und daher für die Waveletbasen zwar etwas mehr Iterationen, aber wesentlich weniger Multiplikationen benötigt werden, um diesen Fehler zu erzielen. Die Verwendung der semiorthogonalen Waveletbasis reduziert beispielsweise den Aufwand für die Lösung des Gleichungssystems um mehr als 75%.

Aufgaben

1. Auf dem Teilraum $H_0^1(0,1)$ bildet die Bilinearform

$$\langle \varphi, \psi \rangle = \int_0^1 \varphi'(x)\psi'(x)\,dx$$

ein Innenprodukt.
(a) Konstruieren Sie den linearen Spline ψ über dem Referenzgitter $\triangle_*$ mit kleinstmöglichem Träger, der bezüglich dieses Innenprodukts orthogonal zu Λ ist und der $\langle \psi, \psi \rangle = 1$ erfüllt. Wie hängt ψ mit dem Haar-Wavelet zusammen?
(b) Zeigen Sie, daß ψ ein Wavelet bezüglich des $\mathcal{L}^2(0,1)$-Innenprodukts ist und geben Sie explizite Formeln für die entsprechenden Basistransformationen von der Zweiskalenbasis in die nodale Basis und zurück an.
Die durch dieses Wavelet erzeugte Mehrskalenbasis im Raum der linearen Splines ist die sogenannte *hierarchische Basis*.

2. Sei V_k der Raum der 1-periodischen linearen Splines über dem Gitter $\widetilde{\triangle}_k$ und

$$u = \sum_{j=0}^{n-1} \xi_{k+1,j}\Lambda_{k+1,j} \in V_k\,, \qquad n = 2^k\,.$$

Der Vektor $\boldsymbol{u} = [\xi_{k,0}, \ldots, \xi_{k,n-1}]^T \in \mathbb{R}^n$ sei der entsprechende Koeffizientenvektor.
(a) Berechnen Sie die zur Basis $\{\Lambda_{k,j} : j = 0, \ldots, 2^k - 1\}$ gehörige Gramsche Matrix G.
(b) Zeigen Sie mit dem Satz 23.1 von Gerschgorin, daß

$$\|u\|^2_{\mathcal{L}^2(0,1)} = \boldsymbol{u}^* G \boldsymbol{u} \geq \frac{1}{3}\|\boldsymbol{u}\|_2^2\,.$$

3. Die folgende Wertetabelle definiert fünf lineare Splines $\psi^{(j)}$ über dem Referenzgitter $\triangle_*$ mit $\operatorname{supp}\psi^{(j)} \subset [0,3]$. Skizzieren Sie diese Splines und untersuchen Sie, welche davon Wavelets sind.

x	0	0.5	1	1.5	2	2.5	3
$\psi^{(1)}$	0	1	−1	0	0	0	0
$\psi^{(2)}$	0	1	0	−1	0	0	0
$\psi^{(3)}$	0	−1	2	−1	0	0	0
$\psi^{(4)}$	0	0	1	−1	0	0	0
$\psi^{(5)}$	0	0	−1	2	−1	0	0

4. Der in Aufgabe 3 definierte Spline $\psi^{(3)}$ ist kein Wavelet, hat aber dennoch interessante Eigenschaften. Wie üblich seien die Funktionen $\psi_{kj} = 2^{k/2}\psi^{(3)}(2^k \cdot -j) \in V_{k+1}$ durch Stauchung und Verschiebung von $\psi^{(3)}$ definiert.
(a) Zeigen Sie, daß $\psi^{(3)}$ die Orthogonalitätsbedingungen (58.5) erfüllt und daß die Funktionen ψ_{kj} und $\psi_{kj'}$ für $j \neq j'$ zueinander orthogonal sind.
(b) Betrachten Sie die lineare Hülle

$$\widetilde{V}_{k+1} = \operatorname{span}\{\,\Lambda_{kj}, \psi_{kj} : j = 0, \ldots, 2^k - 1\,\} \subset V_{k+1}$$

und bestimmen Sie eine Funktion $\varphi \in V_{k+1} \setminus \{0\}$, die senkrecht auf $\widetilde{V}_{k+1}$ steht.

5. Es sei V_k der Raum der 1-periodischen linearen Splines über dem Gitter $\widetilde{\triangle}_k$ und ψ das semiorthogonale Wavelet aus Abschnitt 57.
(a) Berechnen Sie die Gramsche Matrix G der zugehörigen Multiskalenbasis für das $\mathcal{L}^2$-Innenprodukt und bestimmen Sie von p unabhängige Konstanten $c, C > 0$ mit

$$c\,\|z\|_2^2 \;\le\; z^* G z \;\le\; C\,\|z\|_2^2 \qquad \text{für alle } z \in \mathbb{R}^{2^k}.$$

(b) Über die Multiskalenbasisdarstellung

$$f \;=\; \sum_{j=0}^{3} \xi_{2j} \Lambda_{2j} \;+\; \sum_{k=2}^{p-1} \sum_{j=0}^{2^k-1} \eta_{kj} \psi_{kj}$$

einer Funktion $f \in V_p = V_2 \oplus W_2 \oplus \cdots \oplus W_{p-1}$ wird eine Norm $\|\cdot\|_{V_p}$ durch

$$\|f\|_{V_p}^2 \;=\; \sum_{j=0}^{3} |\xi_{2j}|^2 \;+\; \sum_{k=2}^{p-1} \sum_{j=0}^{2^k-1} |\eta_{kj}|^2$$

definiert. Bestimmen Sie für diese Norm von p unabhängige Konstanten $\tilde{c}, \tilde{C} > 0$ mit

$$\tilde{c}\,\|f\|_{V_p} \;\le\; \|f\|_{\mathcal{L}^2(0,1)} \;\le\; \tilde{C}\,\|f\|_{V_p} \qquad \text{für alle } f \in V_p\,.$$

6. Sei ψ ein Wavelet, das die Lokalisierungseigenschaft besitzt. Zeigen Sie, daß jede Hutfunktion aus V_{k+1} durch eine feste Anzahl von Basisfunktionen der zugehörigen Zweiskalenbasis ausgedrückt werden kann.

7. Vervollständigen Sie den Beweis von Satz 58.2 und konstruieren Sie das Wavelet ψ mit $\operatorname{supp}\psi \subset [0,3]$, das die Orthogonalitätsbedingungen (58.5) erfüllt und für das die Koeffizienten c_{-1} und c_0 aus (57.8) verschwinden. Geben Sie explizite Formeln für die Basistransformationen von der nodalen Basis in die Zweiskalenbasis an.

8. In dieser Aufgabe bezeichne V_k den Raum der stückweise konstanten Splines über dem Gitter $\triangle_k$ und χ sei wie in Abschnitt 56 die zugehörige Skalierungsfunktion.
(a) Bestimmen Sie alle stückweise konstanten Spline-Wavelets ψ über dem Referenzgitter $\triangle_*$ mit Träger $\operatorname{supp}\psi \subset [0,3]$, so daß W_k und V_k jeweils orthogonal zueinander sind und ψ die Orthogonalitätsbedingungen (58.5) erfüllt.
(b) Zeigen Sie, daß genau zwei (welche?) dieser Wavelets punktsymmetrisch zum Mittelpunkt $x = 1.5$ des Trägerintervalls und durch $\int_0^3 \psi^2(x)dx = 1$ normiert sind.
(c) Zeigen Sie, daß keines der Spline-Wavelets aus (a) die Lokalisierungseigenschaft besitzt.
Hinweis: Für die entsprechende Definition der Lokalisierungseigenschaft ist in Definition 58.1 die Skalierungsfunktion Λ durch χ zu ersetzen.

9. Zeigen Sie, daß die beiden Orthogonalitätsbedingungen (58.5) zu den beiden Bedingungsgleichungen

$$\begin{aligned} \psi_1 + \;\; \psi_2 + \;\; \psi_3 + \;\; \psi_4 + \;\; \psi_5 \;&=\; 0\,, \\ \psi_1 + 2\,\psi_2 + 3\,\psi_3 + 4\,\psi_4 + 5\,\psi_5 \;&=\; 0\,, \end{aligned}$$

führen.

10. Gegeben sei ein linearer Spline $f \in V_{k+1}$ mit

$$f \;=\; \sum_{j=0}^{N-1} \xi_{k+1,j}\Lambda_{k+1,j} \;=\; \sum_{j=0}^{n-1} \big(\xi_{kj}\Lambda_{kj} + \eta_{kj}\psi_{kj}\big)\,, \qquad n = 2^k - 1,\ N = 2n\,,$$

wobei die ψ_{kj} aus dem biorthogonalen Wavelet aus Abschnitt 58 abgeleitet sind. Mit

$$\begin{aligned} \boldsymbol{x} &= [\xi_{k0}, \xi_{k1}, \ldots, \xi_{k,n-1}]^T \in \mathbb{R}^n\,,\\ \boldsymbol{y} &= [\eta_{k0}, \eta_{k1}, \ldots, \eta_{k,n-1}]^T \in \mathbb{R}^n\,,\\ \boldsymbol{z} &= [\xi_{k+1,0}, \xi_{k+1,1}, \ldots, \xi_{k+1,N-1}]^T \in \mathbb{R}^N\,, \end{aligned}$$

seien die zugehörigen Koeffizientenvektoren bezeichnet.
(a) Zeigen Sie, daß

$$\boldsymbol{x} = \frac{1}{4\sqrt{2}} \begin{bmatrix} 2 & 6 & 2 & -1 & & & -1 \\ & -1 & 2 & 6 & 2 & -1 & \\ & & & \ddots & & \ddots & \\ 2 & -1 & & & -1 & 2 & 6 \end{bmatrix} \boldsymbol{z} \qquad \text{und} \qquad \boldsymbol{y} = \frac{3}{4\sqrt{2}} \begin{bmatrix} & -1 & 2 & -1 & & & \\ & & & -1 & 2 & -1 & \\ & & & & & \ddots & \\ 2 & -1 & & & & & -1 \end{bmatrix} \boldsymbol{z}\,.$$

(b) Betrachten Sie die gestörten Koeffizienten

$$\tilde{\boldsymbol{z}} \;=\; \boldsymbol{z} \,+\, \delta\big[\,1, e^{\mathrm{i}2\pi/N}, \ldots, e^{\mathrm{i}(N-1)2\pi/N}\,\big]^T.$$

Wie wirkt sich diese Störung auf $\boldsymbol{x}$ aus? Zeigen Sie, daß die fortgepflanzte Störung $\tilde{\boldsymbol{x}} - \boldsymbol{x}$ der Abschätzung

$$\frac{\|\tilde{\boldsymbol{x}} - \boldsymbol{x}\|_2}{\|\tilde{\boldsymbol{z}} - \boldsymbol{z}\|_2} \;\le\; 9/8$$

genügt.
Hinweis: Verwenden Sie Aufgabe IX.15.

11. Bestimmen Sie die Bestapproximation s_p der charakteristischen Funktion $\chi_{[0.4,0.6]}$ des Intervalls $[0.4, 0.6]$ aus dem Raum V_k der periodischen linearen Splines über dem Gitter $\widetilde{\triangle}_p$ mit Gitterweite $h = 2^{-p}$ und plotten Sie sowohl für die semiorthogonale als auch die biorthogonale Multiskalenbasis den Anteil von s_p in V_k für $k = n/2$.

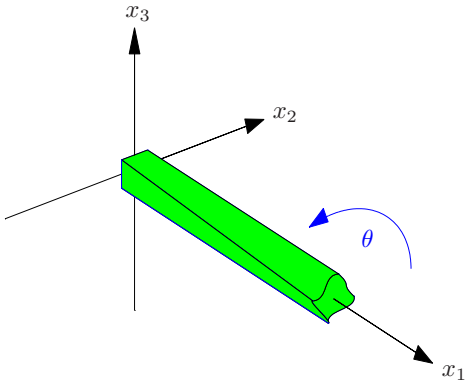
12. Leiten Sie die Kernfunktion $k(\theta, \tau)$ der Integralgleichung (59.5), (59.6) für den Fall einer Ellipse Ω mit Halbachsen α und β her.

13. Sei k eine doppelt 2π-periodische, viermal stetig differenzierbare Kernfunktion und K der zugehörige Integraloperator (59.10). Ferner sei ψ ein Spline-Wavelet (stückweise konstant oder stückweise linear) mit $\operatorname{supp}\psi = [0, l]$, welches die Orthogonalitätsbedingungen (58.5) erfüllt. Zeigen Sie, daß unter diesen Voraussetzungen eine positive Konstante C existiert, so daß die Abschätzung

$$\big|\,\langle\,\psi_{\nu j}, K\psi_{\mu j}\,\rangle_{\mathcal{L}^2(0,2\pi)}\,\big| \;\le\; C\,2^{-(\nu+\mu)5/2}\,, \qquad \nu, \mu > 0\,,$$

gültig ist. Interpretieren Sie dieses Ergebnis.
Hinweis: Beachten Sie die Darstellung (59.9) für die Matrixelemente der Galerkin-Matrizen.

Mathematische Modellierung



XI Dynamik

Die Modellierung technisch-naturwissenschaftlicher Vorgänge ist eine zentrale Aufgabe des wissenschaftlichen Rechnens. Das entscheidende Problem besteht darin, die Realität so genau abzubilden, wie es für die jeweilige Anwendung erforderlich ist, ohne dabei die numerische Umsetzbarkeit aus den Augen zu verlieren.

Mathematisches Modellieren ist eine Frage der Erfahrung und des Abstraktionsvermögens. Viele Modellgleichungen treten mit nur leichten Variationen in völlig unterschiedlichen Anwendungen auf. Dies soll in den folgenden drei Kapiteln anhand möglichst einfacher Beispiele illustriert werden. Andere Beispiele finden sich in den Modellierungsbüchern [5, 29, 42, 69] sowie in der Sammlung [60] von Modellierungsaufgaben aus der Zeitschrift *SIAM Review*.

Zunächst betrachten wir zeitabhängige Prozesse, die das dynamische Verhalten einer oder mehrerer gekoppelter Größen beschreiben und bei der Modellierung zumeist auf Anfangswertprobleme für Differentialgleichungen oder differential-algebraische Gleichungen führen.

60 Populationsmodelle

Populationsmodelle, die Wechselwirkungen zwischen verschiedenen Spezies eines ökologischen oder soziologischen Systems beschreiben, sind ein typisches Beispiel für dynamische Prozesse. Derartige Modelle beruhen selten auf physikalischen Naturgesetzen sondern häufig auf Plausibilitätsüberlegungen. Es ist daher wichtig, numerische Simulationen mit konkreten experimentellen Daten zu vergleichen.

Betrachten wir zunächst eine von der Umwelt völlig unabhängige Spezies mit Kopfzahl $x = x(t)$ zur Zeit t. Unter der Annahme, daß im wesentlichen konstante Geburts- und Sterberaten $g > 0$ bzw. $s > 0$ vorliegen, ergibt sich in einem kleinen Zeitintervall dt ein relativer Zuwachs

$$\frac{dx}{x} = (g - s)dt$$

der Population. Division durch dt und Grenzübergang $dt \to 0$ ergibt die Differentialgleichung

$$x' \;=\; (g-s)\,x \tag{60.1}$$

mit der allgemeinen Lösung $x(t) = ce^{(g-s)(t-t_0)}$; der Wert von c entspricht der Größe der Population zum Zeitpunkt $t = t_0$, dem sogenannten Anfangswert für die Differentialgleichung (60.1). Als Beispiel sei die Weltbevölkerung angeführt, die 1961 bei etwa 3.06 Milliarden Menschen lag und damals pro Jahr um etwa 2% zunahm. Dies entspricht einem Faktor $g - s \approx 0.02$ in (60.1), wenn wir die Zeit in Jahren angeben. Die physikalische Einheit der Parameter g und s ist 1/Jahr.

Allerdings hat dieses Modell Defizite, wenn die Geburtenrate höher als die Sterberate ist, da eine über alle Grenzen anwachsende Population angesichts beschränkter Ressourcen unrealistisch ist. Für das oben genannte Zahlenbeispiel ergibt sich beispielsweise

$$x(t+35) \;=\; ce^{0.02(t-t_0+35)} \;=\; e^{0.02\cdot 35}x(t) \;\approx\; 2.01\cdot x(t)\,,$$

d. h. nach diesem Modell würde sich die Weltbevölkerung alle 35 Jahre in etwa verdoppeln.

Dieser Schwachpunkt beruht auf der fehlenden Berücksichtigung ökologischer und sozialer Probleme, die eine große Population mit sich bringt. Derartige Probleme, etwa Kriminalität, Abfallbelastung, etc., erhöhen die Sterberate. Auf Verhulst geht daher der Vorschlag zurück, diese Auswirkungen durch einen zusätzlichen Term in der Sterberate $s = s(x)$ zu modellieren; demnach setzt sich die Sterberate

$$s = s_0 + ax\,, \qquad a > 0\,, \tag{60.2}$$

aus einer „natürlichen“ Sterberate s_0 und einem zweiten Anteil ax proportional zur Größe der Population zusammen.

Dieses Verhulst-Modell führt auf die *logistische Differentialgleichung*

$$x' = x(d-ax)\,, \tag{60.3}$$

wobei $d = g - s_0$ gesetzt wurde. Für kleine Werte von x sind die Differentialgleichungen (60.1) und (60.3) fast gleich; erst wenn die Population groß ist, ergeben sich wesentliche Unterschiede. Abbildung 60.1 illustriert diese Unterschiede für das Beispiel der Weltbevölkerung. Die gebrochene Linie zeigt die Entwicklung für das einfachere Modell mit exponentiellem Wachstum. Für das

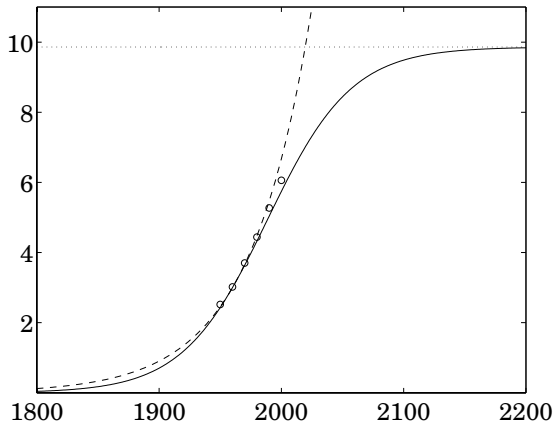


Abb. 60.1: Evolution der Weltbevölkerung nach Verhulst (in Milliarden Menschen)

Modell von Verhulst (die durchgezogene Kurve) finden sich in dem Buch von Braun [11, S. 39] die Parameterwerte

$$d = 0.029 \qquad \text{und} \qquad a = 2.941 \cdot 10^{-3} \tag{60.4}$$

auf der Grundlage der Bevölkerungsentwicklung bis ins Jahr 1961. Diese Werte beziehen sich auf Angaben von t in Jahren und x in Milliarden Menschen, d. h. a hat die Einheit $1/\text{Jahr} \cdot \text{Milliarden Menschen}$. Seit damals sind einige Jahre vergangen und die Kreise in Abbildung 60.1 zeigen die tatsächliche Entwicklung der Weltbevölkerung.[1] Diese neu erhobenen statistischen Daten liegen etwa zwischen den beiden berechneten Kurven (vgl. auch Beispiel 88.1 in Kapitel XV).

Die Differentialgleichung (60.3) läßt sich durch Trennung der Veränderlichen (vgl. etwa [52, Abschnitt 8]) exakt integrieren und besitzt die Lösungsschar

$$x(t) \;=\; \frac{d/a}{1 - ce^{-d(t-t_0)}}\,, \tag{60.5}$$

wobei der Parameter c dieser Schar wieder aus dem Anfangswert zu einer Zeit $t = t_0$ errechnet werden muß. Da $x(t_0)$ positiv sein soll, ist $c < 1$ für positive

[1] Die folgende Tabelle aus *United Nations, World Population Prospects: The 1998 Revision* bzw. *United Nations, World Population Prospects: The 2000 Revision* (vgl. auch `http://www.un.org/popin`) illustriert die Entwicklung der Weltbevölkerung (in Milliarden Menschen) in den Jahren von 1950 bis 2000:

1950 :	2.52	1970 :	3.70	1990 :	5.27
1960 :	3.02	1980 :	4.44	2000 :	6.06

Am 12.10.1999 wurde in den Medien offiziell die Geburt des 6 Milliardsten Menschen gefeiert.

d und $c > 1$ für negative d. Im ersten Fall (d positiv) wächst die Population nach diesem Modell monoton gegen den Grenzwert

$$\lim_{t\to\infty} x(t) \;=\; d/a\,. \tag{60.6}$$

Im zweiten Fall (d negativ) stirbt die Population mit exponentieller Geschwindigkeit aus. Für die obigen Parameter (60.4) sagt das Modell von Verhulst also eine maximale Bevölkerung von knapp 10 Milliarden Menschen auf der Erde voraus (vergleiche die gepunktete Asymptote in Abbildung 60.1); die bereits heute erkennbare Diskrepanz zu den tatsächlichen Bevölkerungszahlen läßt vermuten, daß sich dieser Grenzwert als zu gering erweisen wird.

Mathematisch erheblich interessanter werden derartige Populationsmodelle, wenn das ökologische System aus mehreren Spezies besteht. Am bekanntesten ist hier die sogenannte *Räuber-Beute-Gleichung*

$$\begin{aligned} x_1' &= x_1(d_1 - a_1x_1 - rx_2)\,, \\ x_2' &= x_2(-d_2 + bx_1 - a_2x_2)\,, \end{aligned} \tag{60.7}$$

mit nichtnegativen Parametern a_1, a_2, d_1, d_2, b und r. Dabei bezeichnet x_2 eine Raubtierpopulation und x_1 dessen Beute. Ohne Raubtiere ($x_2 = 0$) ergibt sich für x_1 das bisherige Modell (60.3). Das gleiche gilt für die Raubtierspezies bei nicht vorhandener Beute ($x_1 = 0$): In diesem Fall ist allerdings $d = -d_2 < 0$ in (60.3) und nach den obigen Erkenntnissen stirbt die Raubtierspezies aus. Sind sowohl Raubtiere wie Beutetiere vorhanden, dann ergibt sich für die Beutetiere eine höhere Sterberate

$$s_1 = s_0 + ax_1 + rx_2\,,$$

wobei der gegenüber (60.2) neue Anteil rx_2 besagt, daß die Rate der gefressenen Beutetiere proportional zur Anzahl der Raubtiere ist. Entsprechend hat die Anzahl der Beutetiere einen positiven Einfluß auf die Vermehrung der Raubtiere.

Wenn die Koeffizienten a_1 und a_2 vernachlässigbar sind, ergibt sich durch Division die Differentialgleichung

$$\frac{dx_2}{dx_1} \;=\; \frac{bx_1 - d_2}{x_1}\,\frac{x_2}{d_1 - rx_2}\,,$$

die durch Trennung der Veränderlichen in die implizite Lösungsdarstellung

$$x_1^{d_2}x_2^{d_1} \;=\; c\,e^{bx_1}e^{rx_2} \tag{60.8}$$

überführt werden kann; die Konstante $c > 0$ wird wieder aus Anfangswerten zu einem Zeitpunkt t_0 bestimmt. Man kann zeigen, daß die Lösungsmenge dieser

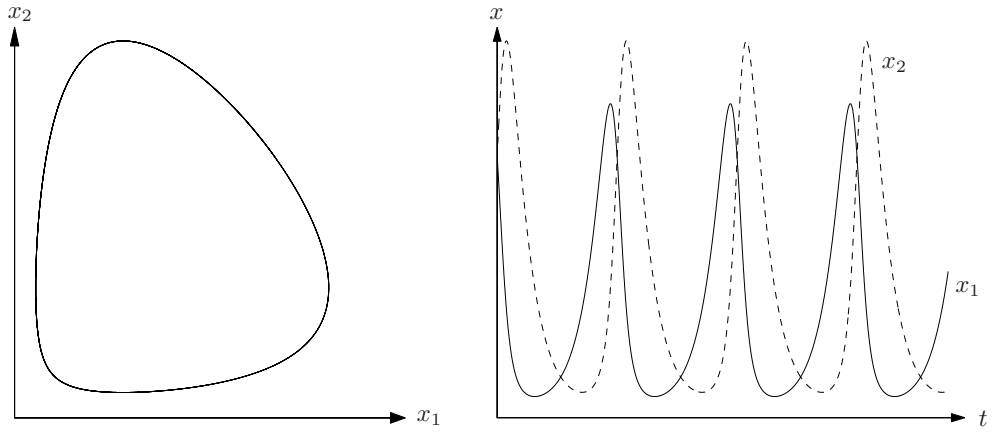


Abb. 60.2: Räuber-Beute-Modelle: Der periodische Fall

Gleichung eine geschlossene Kurve in der (x_1, x_2)-Ebene darstellt, vergleiche Aufgabe 1 und Abbildung 60.2 (links). In Abhängigkeit von t ergibt sich für diesen Fall das periodische Verhalten aus dem rechten Bild. (Die durchgezogene Kurve kennzeichnet die Beute, die gebrochene Linie die Population der Räuber.)

Im allgemeinen Fall (60.7) mit $a_1 a_2 \neq 0$ läßt sich hingegen zeigen (vgl. [11, Abschnitt 4.8]), daß $x_2(t)$ für $t \to \infty$ gegen Null konvergiert, falls $d_2/b > d_1/a_1$ ist. Da d_1/a_1 in dem Verhulst-Modell die asymptotische Zahl der Beutetiere bei Abwesenheit von Raubtieren angibt, vgl. (60.6), bedeutet diese Ungleichung, daß die Sterberate der Raubtiere zu groß ist, um durch das vorhandene Nahrungsangebot kompensiert zu werden: Die Raubtiere sterben in diesem Fall aus.

Ist hingegen $d_2/b < d_1/a_1$, dann hat das Gleichungssystem

$$a_1 x_1 + r x_2 \;=\; d_1\,, \qquad b x_1 - a_2 x_2 \;=\; d_2\,,$$

eine positive Lösung $x_1, x_2 > 0$, nämlich

$$\begin{bmatrix} x_1 \\ x_2 \end{bmatrix} \;=\; \begin{bmatrix} a_1 & r \\ b & -a_2 \end{bmatrix}^{-1} \begin{bmatrix} d_1 \\ d_2 \end{bmatrix} \;=\; \frac{1}{a_1 a_2 + b\,r} \begin{bmatrix} a_2 d_1 + r d_2 \\ b d_1 - a_1 d_2 \end{bmatrix}. \tag{60.9}$$

Man überprüft unmittelbar durch Einsetzen in die Differentialgleichung (60.7), daß diese Lösung gleichzeitig eine konstante (*stationäre*) Lösungsfunktion der Räuber-Beute-Gleichung ist. Die Populationen $x_1(t)$ und $x_2(t)$ konvergieren bei positiven Anfangswerten für $t \to \infty$ gegen diese Lösung. Dies ist in Abbildung 60.3 an einem Beispiel illustriert: Links sieht man wieder die Kurve

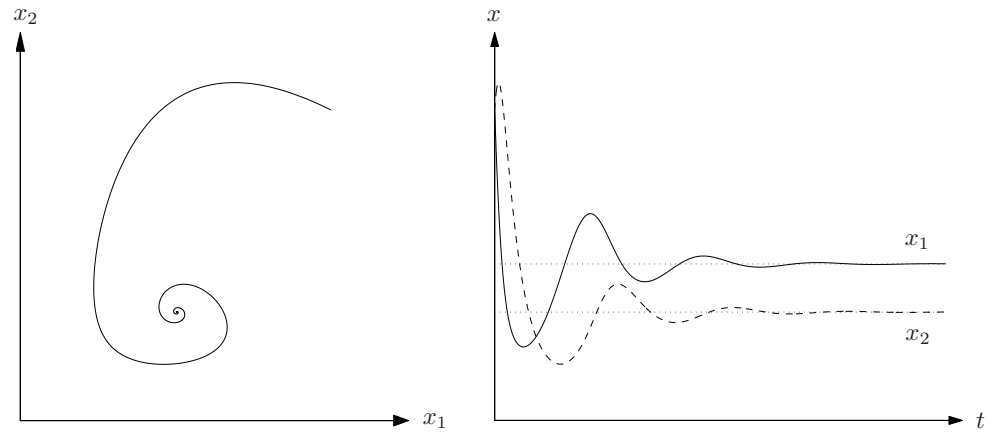


Abb. 60.3: Räuber-Beute-Modelle: Der konvergente Fall

$(x_1(t), x_2(t))$ in der (x_1, x_2)-Ebene, rechts sind die beiden Populationen über der Zeit aufgetragen; die gepunkteten Linien bzw. das Zentrum der „Spirale“ im linken Bild kennzeichnen den Grenzwert (60.9).

Mit ähnlichen Modellen können auch *Epidemien* innerhalb einer Population beschrieben werden. In diesem Fall entsprechen x_1 und x_2 den gesunden bzw. infizierten Individuen der Population, denn letztere reduzieren den gesunden Bestand durch Infektion. Die Infektionswahrscheinlichkeit ist wie zuvor proportional dazu, daß gesunde und infizierte Individuen aufeinandertreffen; r sei die entsprechende Infektionsrate. Daneben unterliegen beide Populationen einer natürlichen Sterblichkeitsrate ($s_1 = s$ bzw. $s_2 = s + v$), wobei die Sterblichkeit der infizierten Individuen in der Regel höher ist, daher der zusätzliche Term $v > 0$. Wir nehmen an, daß die Geburtenrate g beider Populationen gleich sei, aber alle Nachkommen gesund zur Welt kommen. Daneben gibt es einen zweiten Zuwachs für die gesunde Population aufgrund der Genesung vormals infizierter Individuen. Dieser Zuwachs wird proportional zu x_2 angenommen (mit Rate $n > 0$) und reduziert natürlich in derselben Weise den infizierten Bestand, wie er dem gesunden Anteil zugute kommt. Insgesamt erhalten wir somit das folgende Modell:

$$\begin{aligned} x_1' &= x_1(g - s - ax_1 - ax_2 - rx_2) + x_2(g + n)\,, \\ x_2' &= x_2(-s - v - n + rx_1 - ax_1 - ax_2)\,. \end{aligned} \tag{60.10}$$

Um diese Differentialgleichung zu lösen, werden wie zuvor Anfangswerte $x_1(t_0)$ und $x_2(t_0)$ zu einem Zeitpunkt t_0 benötigt. In diesem Beispiel ist die Situation interessant, in dem ein infiziertes Individuum zum Zeitpunkt $t = t_0$ zu einer bislang infektionsfreien Population stößt. Das infizierte Individuum, repräsen-

tiert durch den Wert $x_2(t_0) = 1$, verursacht im anschließenden Zeitintervall dt ungefähr

$$dx_2 \approx \big(-s - v - n + (r - a)x_1\big)dt$$

Neuerkrankungen, und wenn die rechte Seite positiv ist, kann sich die Epidemie ausbreiten, also wenn die Größe x_1 der Population oberhalb des Schwellenwerts

$$(s + v + n)/(r - a)$$

liegt. Andernfalls wird der Infektionsherd gesund (oder er geht zugrunde), bevor es zu weiteren Ansteckungen kommen kann. Diese Ungleichung legt verschiedene Maßnahmen nahe, um den Ausbruch einer Epidemie zu verhindern:

- Erhöhung der allgemeinen Sterberate s (beispielsweise durch aufgestellte Fallen in Tollwutgebieten);
- gezielte Tötung kranker Individuen, um die zusätzliche Sterberate v zu erhöhen;
- medizinische Heilbehandlung, um die Genesungsrate n zu verbessern;
- Impfungen, um die Infektionsrate r zu reduzieren;
- Isolation infizierter Individuen (Quarantäne), um die Anzahl der Kontakte zu reduzieren.

Für eine Vielzahl weitergehender Fragestellungen und Anwendungen sei auf die Bücher von Haberman [42] und von Yeargers, Shonkwiler und Herod [109] verwiesen.

61 Ein Modell für Aids

Entsprechende Modelle werden gegenwärtig auch für den Verlauf der Immunschwächekrankheit Aids entwickelt, man vergleiche etwa [109] oder das Buch von Nowak und May [77]. Von Interesse ist dabei nicht nur die Ausbreitung der Krankheit innerhalb der Bevölkerung sondern auch die Simulation der körpereigenen Abwehrreaktionen nach einer Infektion durch den HIV-Virus. Letzteres ist unter dem Vorbehalt zu sehen, daß die verschiedenen Mechanismen der Immunreaktionen im Körper derzeit noch nicht vollständig verstanden werden. Es sei aber vermerkt, daß die bisherige Zusammenarbeit von Mathematikern und Medizinern zu neuen Strategien für den Einsatz von Medikamenten bei Aids-Patienten geführt hat.

Der menschliche Organismus verfügt über ein hochentwickeltes Immunsystem zur Bekämpfung eingedrungener Fremdkörper und Viren. Ein zentraler Bestandteil dieses Immunsystems sind die sogenannten T-Zellen, die fortlaufend

von der Thymusdrüse abgeschieden werden und dann nach endlicher Zeit einen programmierten Zelltod, die sogenannte Apoptosis, erleiden. Wenn eine Immunreaktion aktiviert wird, vermehren sich die T-Zellen durch Zellteilung, um die Eindringlinge abzuwehren. Eine Proteinmarkierung auf der Oberfläche der Zelle unterscheidet die sogenannten $CD4^+$ und $CD8^+$ T-Zellen; letztere werden auch Killerzellen genannt: Sie machen die eingedrungenen Viren unschädlich.

Wie alle Viren sucht sich das HIV-Virus eine sogenannte Wirtszelle, dringt in diese Zelle ein und wird dann innerhalb der Zelle reproduziert. Das Besondere an dem HIV-Virus ist, daß gerade die $CD4^+$ T-Zellen als Wirtszellen ausgebeutet werden und somit in unmittelbarer Weise das Immunsystem des Körpers gestört wird; $CD8^+$ T-Zellen werden vom Virus nicht befallen. Bei dem Versuch sich zu teilen, platzen die infizierten $CD4^+$ T-Zellen und alle zwischenzeitlich produzierten Viren gelangen auf diese Weise in den Blutkreislauf.

Im folgenden wird ein mögliches mathematisches Modell zur Simulation des Krankheitsverlaufs beschrieben; wir beziehen uns dabei auf eine Arbeit von Kirschner [59]. $CD4^+$ und $CD8^+$ T-Zellen werden in diesem Modell nicht unterschieden sondern nur die Anzahlen x und y der gesunden und der infizierten T-Zellen; v ist die Anzahl der HIV-Viren:

$$x' = p - s_x x - r_x\, xv + g(v)x\,, \tag{61.1a}$$

$$y' = r_x\, xv - s_y y - g(v)y\,, \tag{61.1b}$$

$$v' = ng(v)y + f(v) - s_v v - (r_x + r_v)\, xv\,. \tag{61.1c}$$

Die Terme lassen sich dabei im einzelnen wie folgt erläutern. Die im wesentlichen konstante Produktion neuer (gesunder) T-Zellen in der Thymusdrüse wird durch den Zuwachs p auf der rechten Seite von (61.1a) beschrieben. Zu der programmierten Sterberate s_x der gesunden T-Zellen kommt noch ein durch den Virus bedingter Faktor $r_x v$ für die Infektion gesunder T-Zellen durch den HIV-Virus (entsprechend dem Infektionsterm rx_2 in (60.10)).

Die Funktion $g(v)$ modelliert die *Zellteilungsrate* der T-Zellen zur Bekämpfung der in den Körper eingedrungenen Viren. Gesunde T-Zellen verdoppeln sich bei der Zellteilung, daher der Zuwachs $g(v)x$ in (61.1a). Infizierte T-Zellen werden bei dieser Aktion hingegen zerstört und führen im Mittel n reproduzierte Viren dem Blutkreislauf zu; es ergibt sich daher eine Reduktion $-g(v)y$ in (61.1b) und ein Zuwachs $ng(v)y$ in (61.1c). Die Zellteilungsrate ist ein freier Parameter des Modells. Natürliche Bedingungen an g sind im einzelnen

- $g(0) = 0$: es findet keine Zellteilung statt, solange keine Viren in den Kreislauf eingedrungen sind;
- g ist monoton wachsend: je mehr Viren vorhanden sind, um so stärker ist die Stimulanz;

• $g(v) \leq g_\infty$: die Zellteilungsrate kann eine bestimmte Rate g_∞ nicht übersteigen.

Eine einfache Funktion, die diese Anforderungen erfüllt, lautet

$$g(v) \;=\; \frac{v}{v_g + v}\, g_\infty \,,$$

wobei $v_g > 0$ die Virenzahl angibt, für die die Zellteilungsrate halb so groß wie die Maximalrate ist.

Die Gleichung (61.1b) enthält neben dem bereits diskutierten Term $g(v)y$ den Zuwachs $r_x xv$ neu infizierter T-Zellen sowie die Abnahme $s_y y$ durch Zelltod. Dabei ist die Sterberate s_y der infizierten Zellen um eine Größenordnung höher als die Sterberate s_x der gesunden T-Zellen.

Die letzte Gleichung (61.1c) beschreibt die Änderung der Virenpopulation. Demnach werden aus jeder infizierten T-Zelle im Mittel n Viren freigesetzt, wenn sie bei dem Versuch einer Zellteilung platzt. Darüber hinaus werden HIV-Viren aber auch von anderen infizierten Zellen (etwa von Thymocyten und Macrophagen) reproduziert. Daher kommt der zusätzliche positive Term $f(v)$, auf den wir hier nicht näher eingehen wollen. Die verbleibenden Terme in (61.1c) quantifizieren die Abnahme der Virenpopulation, wie zum Beispiel durch natürlichen Virentod mit Rate s_v; daneben gehen jene $r_x xv$ HIV-Viren verloren, die in die gesunden $CD4^+$ T-Zellen eindringen, während der hinterste Term $r_v xv$ die Virenabnahme aufgrund der körpereigenen Abwehrkräfte modelliert.

Es sei erneut betont, daß dieses Modell die komplexen Immunreaktionen im Körper sehr vereinfacht repräsentiert. Zudem sind einige der beschriebenen Mechanismen unter Biochemikern strittig: So herrscht beispielsweise in der Literatur keine Einigkeit darüber, ob die Killerzellen tatsächlich direkt wie oben angenommen HIV-Viren vernichten oder ob sie vielmehr die Virenreproduktion hemmen, indem sie infizierte Zellen binden und vor dem Zelltod aus dem Blutkreislauf entfernen. In letzterem Fall müßte anstelle des Terms $-r_v xv$ in (61.1c) eher ein Term $-r_y xy$ in (61.1b) auftreten.

Eine andere Schwäche dieses Modells mag die gemeinsame Modellierung der $CD8^+$ und der gesunden $CD4^+$ T-Zellen in einer Kopfzahl x sein. Der Term $-r_v xv$ in (61.1c) ist beispielsweise nur dann plausibel, wenn die $CD8^+$ T-Zellen einen festen Anteil aller gesunden T-Zellen ausmachen. Auf der anderen Seite dringen Viren nur in $CD4^+$ T-Zellen ein; der Term $r_x xv$ in (61.1a) und (61.1b) sollte also die $CD8^+$ T-Zellen nicht in der gleichen Weise reduzieren wie die $CD4^+$ T-Zellen. Hier ist die Modellierung nicht ganz konsistent.

Das Beeindruckende an dem obigen Modell ist jedoch die frappierend gute Übereinstimmung der numerischen Simulationen mit dem in der Realität be-

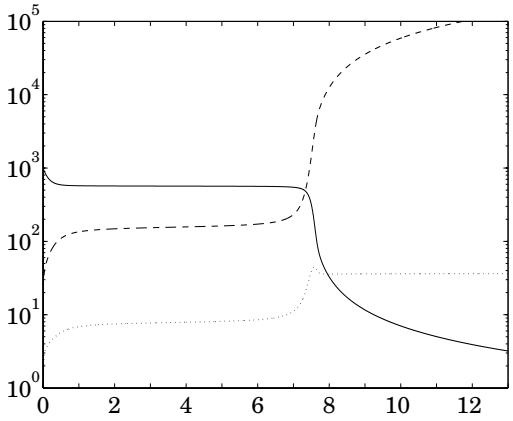


Abb. 61.1: Simulierter Verlauf der ersten 13 Jahre einer Aids-Infektion

obachteten Krankheitsverlauf, zumindest bei geeigneter Wahl der zahlreichen Parameter in (61.1). Der HIV-Virus verharrt nach der Infektion über Jahre hinweg in einer Art „Lauerzustand“ im Körper des Patienten (der Patient ist in diesem Zeitraum HIV-positiv), bevor schließlich nach 5–10 Jahren die eigentliche Immunschwächekrankheit Aids „ausbricht“. Danach schwinden die körpereigenen Abwehrkräfte zusehends und der Patient stirbt letztendlich oft an den Folgen einer an und für sich harmlosen Erkältung.

Zum Vergleich nun in Abbildung 61.1 die Lösung des Differentialgleichungsmodells (61.1) für die Parameterwerte

$$\begin{aligned} s_x &= 0.02\,, & p &= 10\,, & n &= 1000\,,\\ s_y &= 0.265\,, & r_x &= 2.4\cdot 10^{-5}\,, & g_\infty &= 0.01\,,\\ s_v &= 0\,, & r_v &= 7.4\cdot 10^{-4}\,, & v_g &= 100 \end{aligned}$$

und die Funktion $f(v) = 20v/(v+1)$. Dabei wird die Zeit in Tagen gemessen und die Variablen x, y und v geben die jeweilige Zellenzahl pro Kubikmillimeter an.[2] In dieser Simulation wird ein gesunder Körper (mit anfänglich 1000 T-Zellen) ab dem Zeitpunkt der Infektion betrachtet, an dem ein HIV-Virus in den Körper eindringt. Aufgetragen ist die Dynamik der gesunden T-Zellenpopulation x inklusive der $CD8^+$ Killerzellen (die durchgezogene Kurve) und die HIV-Virenpopulation v (die gebrochene Kurve) über der Zeit. Die

[2]Diese Parameter sind (mit geringfügigen Modifikationen) der bereits zitierten Quelle [59] entnommen. Dabei ist allerdings kritisch anzumerken, daß die Vernachlässigung der natürlichen Sterberate s_v der Viren eigentlich nicht zulässig ist, da Messungen im Labor eher auf eine recht hohe Sterberate s_v zwischen zwei und fünf pro Tag hinweisen, vgl. etwa Perelson, Kirschner und De Boer [84].

gepunktete Linie gibt die Zahl y der infizierten T-Zellen an. Wie man sieht, scheint das System fast sieben Jahre lang stabil zu sein, bevor ein unerklärlicher Anstieg der infizierten T-Zellen dieses scheinbare Gleichgewicht zum Einsturz bringt. Danach nimmt die Zahl der Viren im Körper stark zu, während der gesunde T-Zellenbestand rapide abnimmt.

62 Chemische Reaktionskinetik

Auch der Ablauf chemischer Reaktionen kann durch ein System gewöhnlicher Differentialgleichungen beschrieben werden, vgl. Aris [5, Kapitel 8]. Betrachten wir etwa das Reaktionsschema

$$A + B \longrightarrow 2C + D\,,$$

bei dem je ein Molekül der beiden Gase A und B zu zwei Molekülen von C und einem Molekül D reagieren. Nach dem *Massenwirkungsgesetz* ist die Reaktionsgeschwindigkeit bei konstantem Druck, Volumen und Temperatur proportional zu der Wahrscheinlichkeit, daß zwei Moleküle der beteiligten Gase aufeinander treffen, also proportional zu dem Produkt der Konzentrationen von A und B.

Bezeichnen $c_A, \ldots, c_D$ die Konzentrationen der Gase A bis D (Einheit: mol/l), so ergeben sich aus dieser Proportionalitätsannahme die Differentialgleichungen

$$c_A' = c_B' = -\kappa c_A c_B\,, \qquad c_C' = 2\kappa c_A c_B\,, \qquad c_D' = \kappa c_A c_B\,, \tag{62.1}$$

wobei κ die entsprechende Proportionalitätskonstante ist; je größer κ ist, desto größer ist die Reaktionsgeschwindigkeit. Ohne äußere Einflüsse beschreibt (62.1) das chemische System für alle Zeiten $t > 0$ und muß lediglich noch mit Konzentrationen der Gase zum Zeitpunkt $t = 0$ versehen werden. Finden davon unabhängig weitere Reaktionen unter den betroffenen Gasen statt, so müssen entsprechende Terme hinzugefügt werden.

Beispiel 62.1. Gegeben sei das chemische Reaktionsschema

$$\begin{aligned} A &\xrightarrow{0.04} B\,,\\ B + B &\xrightarrow{3\cdot 10^7} C + B\,,\\ B + C &\xrightarrow{10^4} A + C\,, \end{aligned}$$

zwischen den Gasen A, B und C, wobei die Reaktionskoeffizienten über den Reaktionspfeilen vermerkt sind. Bei diesem System handelt es sich um ein

konstruiertes Beispiel mit dimensionslosen Größen aus dem Buch von Hairer und Wanner [45, Section IV.1], das ein beliebtes Testbeispiel für numerische Algorithmen ist. In Abschnitt 77 werden wir auf dieses Beispiel zurückkommen.

Anhand des Reaktionsschemas macht man sich unmittelbar klar, daß alle drei Gase A, B und C die gleiche atomare Zusammensetzung haben müssen; lediglich die innermolekularen Bindungen sind unterschiedlich. Dabei sind A und B instabile Anordnungen des Moleküls, C ist ein stabiles Endprodukt. Während A jedoch nur relativ langsam in die Zwischenform B übergeht, wirkt die Existenz von B katalytisch für eine weitere, sehr schnelle Transformation von B nach C. (Man spricht bei dieser zweiten Reaktion von einer *autokatalytischen Reaktion*, da der Katalysator und die reagierende Substanz identisch sind.) Daneben findet in Anwesenheit von C auch eine Rücktransformation von B nach A statt. Das Gas C hat hier also ebenfalls eine katalytische Wirkung. Es ist zu erwarten, daß die Substanz B nur einen kleinen Anteil des Gasgemischs ausmacht; dieser ist jedoch entscheidend, um die Weiterreaktion zu C in Gang zu halten.

Da C ein stabiles Endprodukt ist und allenfalls katalytische Wirkung hat, nimmt sein Anteil c_C im Gasgemisch durchweg zu. Ferner überlegt man sich recht schnell, daß die Summe der Konzentrationen von A, B und C immer gleich bleiben muß, da in jeder Reaktion genau ein Molekül in genau ein anderes Molekül umgewandelt wird. Dies ist das Gesetz der *Massenerhaltung*.

Das zu diesem Reaktionsschema gehörende Differentialgleichungssystem lautet

$$\begin{aligned} c_A' &= -0.04\, c_A + 10^4\, c_B c_C\,, \\ c_B' &= 0.04\, c_A - 10^4 c_B c_C - 3 \cdot 10^7 c_B^2\,, \\ c_C' &= 3 \cdot 10^7 c_B^2\,. \end{aligned} \tag{62.2}$$

Unter der Annahme, daß zu Beginn der Reaktion lediglich Gas A vorhanden ist, wählen wir

$$c_A(0) = 1 \qquad \text{und} \qquad c_B(0) = c_C(0) = 0$$

als Anfangswerte.

Auch aus (62.2) kann man ablesen, daß c_C während der Reaktion monoton zunimmt, da die rechte Seite der Differentialgleichung für c_C nichtnegativ ist; darüber hinaus sieht man durch Addition der drei Gleichungen, daß die Summe aller Konzentrationen konstant ist, denn es ergibt sich

$$(c_A + c_B + c_C)' = 0\,.$$

Folglich ist $c_A + c_B + c_C = 1$ für alle $t \geq 0$. Weniger offensichtlich ist, daß alle drei Konzentrationen für alle Zeiten nichtnegativ bleiben.

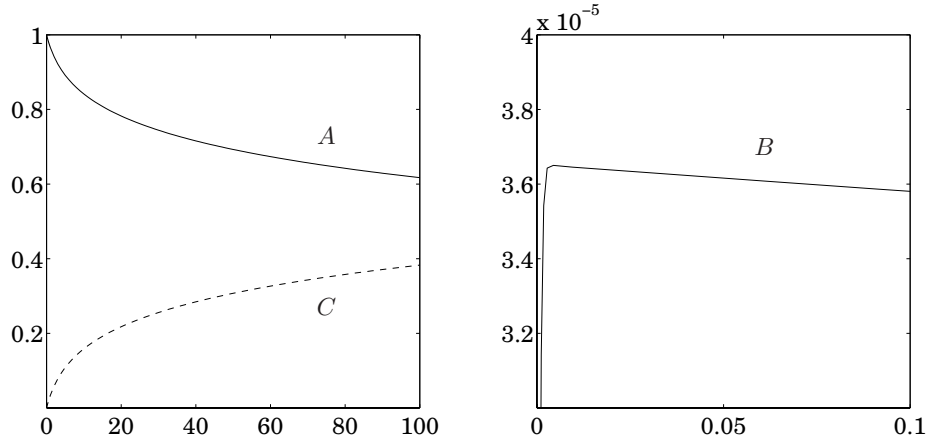


Abb. 62.1: Die einzelnen Konzentrationen als Funktionen der Zeit

Der zeitliche Ablauf der Reaktionen ist in Abbildung 62.1 illustriert. Im linken Bild sind die Konzentrationen von A (durchgezogene Kurve) und C (gebrochene Kurve) über der Zeit aufgetragen. Das Verhalten von c_B wäre in dieser Abbildung nicht zu erkennen, denn die Konzentration von B ist durchweg um vier bis fünf Zehnerpotenzen kleiner als die von A und C. Trotzdem ist die Präsenz von B in dem chemischen System und die korrekte Berechnung von c_B für eine genaue Simulation des Reaktionsverlaufs entscheidend. Aus diesem Grund diskutieren wir die Funktion c_B (im rechten Bild von Abbildung 62.1) nun noch etwas detaillierter. Die Konzentration von B wächst zunächst schnell an und nimmt dann sehr allmählich wieder etwas ab (man beachte den Unterschied der beiden Zeitskalen in Abbildung 62.1). Dieser Umschlagpunkt wird erreicht, wenn in der zweiten Gleichung von (62.2)

$$0.04\, c_A \;=\; 10^4 c_B c_C + 3 \cdot 10^7 c_B^2$$

beziehungsweise

$$c_B \;=\; \tfrac{1}{6} 10^{-3} c_C \;+\; (\tfrac{1}{36} 10^{-6} c_C^2 + \tfrac{4}{3} 10^{-9} c_A)^{1/2} \tag{62.3}$$

ist. Da in diesem Anfangsstadium die Konzentrationen von A und C näherungsweise durch

$$c_A(t) \approx e^{-0.04t} \qquad \text{und} \qquad c_C(t) \approx 1 - c_A(t) \approx 0.04\, t$$

gegeben sind, liegt die Konzentration von A für $t \leq 0.1$ im Bereich von Eins, während c_C die Größenordnung 10^{-3} bis 10^{-2} hat. Somit dominiert der c_A-Term die rechte Seite von (62.3) und für den Maximalwert von c_B ergibt sich die Näherung

$$c_B \;\approx\; (\tfrac{4}{3} 10^{-9})^{1/2} \;\approx\; 3.65 \cdot 10^{-5}\,.$$

$\Diamond$

Eine Vielzahl weiterer Beispiele mit MATLAB-Realisierungen findet sich in dem Buch von Löwe [71].

63 Mehrkörpersysteme

Wir betrachten n Körper (Massepunkte) mit Massen m_i, $i = 1, \ldots, n$, die sich entlang gewisser Bahnen $x_i(t) \in \mathbb{R}^3$, $t \in \mathcal{I} \subset \mathbb{R}$, im Raum bewegen. Die Ableitungen $x_i'(t)$ und $x_i''(t)$ geben die jeweilige Geschwindigkeit und die Beschleunigung der Körper an. Die Bewegung der Körper wird durch das Newtonsche Gesetz (22.1) beschrieben,

$$m_i x_i''(t) = F_i\,, \qquad i = 1, \ldots, n\,, \tag{63.1}$$

wobei $F_i \in \mathbb{R}^3$ die auf den i-ten Körper wirkende Kraft bezeichnet.

Im Gegensatz zu den Gleichungen der vorangegangenen Abschnitte handelt es sich bei (63.1) um Differentialgleichungen *zweiter Ordnung* für die unbekannten Funktionen x_i, da deren zweite Ableitungen in den Gleichungen auftreten. Daher werden sowohl Anfangsbedingungen für die Ortskoordinaten $x_i(0)$ der Körper zur Zeit $t = 0$ als auch für die Geschwindigkeiten $x_i'(0)$ benötigt. Alternativ können Anfangs- und Endbedingungen vorgegeben werden (man spricht dann von einem *Randwertproblem*), etwa die Positionen der Körper zur Anfangszeit $t = 0$ und zur Zeit $t = T > 0$.

63.1 Das Zweikörperproblem

Ein erstes Beispiel für ein Mehrkörperproblem ist das sogenannte Zweikörperproblem, in dem die Bewegung zweier Himmelskörper unter dem Einfluß ihrer Gravitationskraft untersucht wird. Die Stärke der von Newton bestimmten Gravitationskraft ist proportional zu den beiden Massen und antiproportional zu dem Quadrat des Abstands[3] zwischen den beiden Körpern:

$$f(r) \;=\; \gamma\,\frac{m_1 m_2}{r^2}\,, \qquad r = |x_2 - x_1|\,.$$

Der Proportionalitätsfaktor $\gamma = 6.673 \cdot 10^{-11}$ $\mathrm{m}^3/\mathrm{s}^2\mathrm{kg}$ heißt *Gravitationskonstante*. Auf die beiden Körper wirken dann die Kräfte

$$F_1(x_1, x_2) \;=\; f(r)\,\frac{x_2 - x_1}{r} \qquad \text{und} \qquad F_2(x_1, x_2) \;=\; f(r)\,\frac{x_1 - x_2}{r}$$

[3]Für zwei Ortsvektoren x_1 und x_2 verwenden wir im weiteren die Notationen $|x_1 - x_2|$ für den euklidischen Abstand (die Euklidnorm) und $x_1 \cdot x_2$ für das Innenprodukt.

und die Bewegungsgleichungen (63.1) haben die Form

$$x_1'' = \frac{\gamma\, m_2}{|x_1 - x_2|^3}\,(x_2 - x_1)\,, \qquad x_2'' = \frac{\gamma\, m_1}{|x_1 - x_2|^3}\,(x_1 - x_2)\,.$$

Abhängig von den Anfangsbedingungen kann die Lösungskurve durch eine Ellipse, eine Parabel oder eine Hyberbel beschrieben werden, vgl. Goldstein [33].

Ein sehr einfacher Spezialfall des Zweikörperproblems ergibt sich, wenn wir die Bahn $x_1 = x_1(t)$ einer (antriebslosen) Rakete in der Nähe der Erde betrachten. Dazu nehmen wir der Einfachheit halber an, die Masse der Erde sei in ihrem Mittelpunkt x_2 konzentriert und bestimmen zunächst den gemeinsamen Schwerpunkt

$$s \;=\; \frac{m_1 x_1 + m_2 x_2}{m_1 + m_2} \;=\; x_2 + \frac{m_1}{m_1 + m_2}\,(x_1 - x_2)$$

der beiden Körper. Falls die Masse m_1 der Rakete gegenüber der Erdmasse vernachlässigbar ist, fallen der Schwerpunkt und der Erdmittelpunkt praktisch zusammen. Davon unabhängig ist

$$s'' \;=\; \frac{m_1}{m_1 + m_2}\,\frac{\gamma m_2}{r^3}\,(x_2 - x_1) \;+\; \frac{m_2}{m_1 + m_2}\,\frac{\gamma m_1}{r^3}\,(x_1 - x_2) \;=\; 0\,,$$

das heißt wir können den Schwerpunkt, beziehungsweise in erster Näherung auch den Erdmittelpunkt als Ursprung eines unbeschleunigten Koordinatensystems wählen.

Von dem Differentialgleichungssystem verbleibt also nur die Gleichung

$$x_1'' \;=\; -\,\frac{a}{|x_1|^3}\,x_1$$

mit einem geeigneten $a > 0$. Nehmen wir zudem an, daß die Bewegung der Rakete senkrecht zur Erdoberfläche verläuft, so kann das Problem auf eine Dimension reduziert werden, wenn mit $x = x(t)$ der Abstand der Rakete von der Erde bezeichnet wird. Dieser Abstand x genügt dann der Differentialgleichung

$$x'' = -a/x^2\,, \tag{63.2}$$

die wir im weiteren in dimensionslosen Koordinaten betrachten wollen. Für die Anfangswerte $x(0) = 1$ und $x'(0) = -2/3$ prüft man leicht nach, daß die Funktion

$$x_-(t) = (1 - t)^{2/3}\,, \qquad 0 \le t \le 1\,,$$

die Differentialgleichung (63.2) für $a = 2/9$ löst. Offensichtlich prallt in diesem Fall die Rakete zur Zeit $t = 1$ auf die Erde auf. Ihre Geschwindigkeit ist zu diesem Zeitpunkt $x'_-(1) = -\infty$.

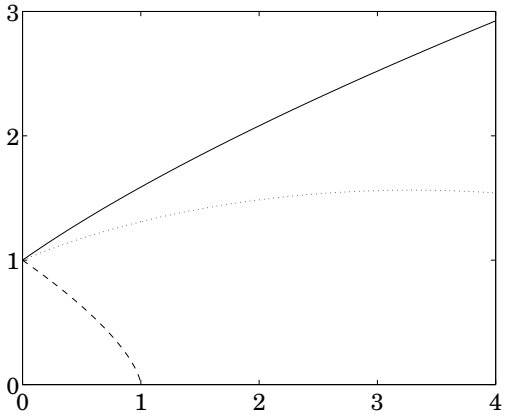


Abb. 63.1: Entfernung der Rakete als Funktion der Zeit bei verschiedenen Startgeschwindigkeiten

Auf der anderen Seite ergibt sich für die Anfangsgeschwindigkeit $x'(0) = 2/3$ bei gleichem a und gleicher Startposition $x(0)$ die Lösung

$$x_+(t) = (1+t)^{2/3}\,, \qquad t \geq 0\,.$$

Diese Lösung existiert für alle nichtnegativen t mit $\lim_{t\to\infty} x_+(t) = \infty$ und $\lim_{t\to\infty} x'_+(t) = 0$. Mit dieser Anfangsgeschwindigkeit schafft die Rakete also den „Absprung“ von der Erde und zudem ist $x'_+(0) = 2/3$ die hierfür minimale Anfangsgeschwindigkeit (Fluchtgeschwindigkeit), vgl. Aufgabe 5. Abbildung 63.1 zeigt die beiden Lösungen x_- und x_+ sowie eine weitere Lösungskurve, bei der die Anfangsgeschwindigkeit $x'(0)$ etwas kleiner als $x'_+(0)$ ist; man kann erkennen, daß die Rakete bei dieser kleineren Startgeschwindigkeit auf die Erde zurückstürzen wird.

Bei mehr als zwei Himmelskörpern überlagern sich die einzelnen Gravitationskräfte und die zugehörigen Bewegungsgleichungen sind nur noch in Ausnahmefällen analytisch lösbar. Die Lösung kann jedoch numerisch bestimmt werden; für die zwölf Planeten unseres Sonnensystems ist das mit erträglichem Aufwand machbar.

63.2 Partikelmethoden

Partikelmethoden können zur Simulation eines *Fluids* (also eines Gases oder einer Flüssigkeit) verwendet werden. Dabei modelliert man das Fluid durch ein System endlich vieler Partikel, die sich ähnlich wie Massepunkte verhalten. Je nach Modell kann man sich unter einem solchen Partikel ein einzelnes Molekül oder ein kleines Fluidvolumen aus vielen Molekülen vorstellen.

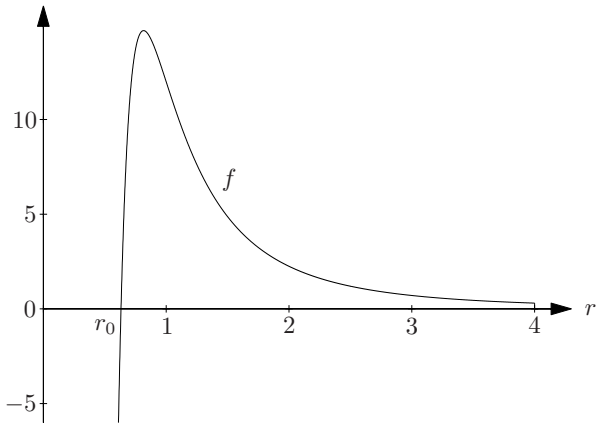


Abb. 63.2: Modell für die innermolekularen Kräfte in einem Wassertropfen

Um einen Eindruck von dieser Technik zu vermitteln, greifen wir auf ein Beispiel aus dem Buch von Greenspan [37] zurück und verwenden im folgenden eine Partikelmethode zur numerischen Simulation eines sich von der Decke ablösenden Wassertropfens. Dazu fassen wir die einzelnen Partikel als Wassermoleküle auf und modellieren zunächst deren Wechselwirkungen untereinander. Zwischen zwei Molekülen wirken sowohl anziehende als auch abstoßende Kräfte, jedoch mit unterschiedlicher Reichweite. Die resultierende Kraft kann – ähnlich wie im vorigen Abschnitt – durch eine Funktion $f = f(r)$ beschrieben werden, die nur vom Abstand r zwischen den beiden Molekülen abhängt. Unterschreitet r einen gewissen Mindestabstand r_0, so überwiegen die abstoßenden Kräfte und f wird negativ; für $r \to 0$ strebt $f(r) \to -\infty$. Für $r > r_0$ dominieren hingegen die anziehenden Kräfte, die allerdings für große r immer schwächer werden. Abbildung 63.2 zeigt den typischen Graph einer solchen Funktion, in diesem Beispiel gegeben durch

$$f(r) \;=\; \frac{20}{r^3} \;-\; \frac{8}{r^5}\,. \tag{63.3}$$

In dem Ensemble aller Moleküle wird somit auf das i-te Partikel im Ort x_i durch jedes andere Molekül in $x_j \neq x_i$ eine entsprechende Kraft

$$F_{ij} \;=\; f(r_{ij})\,\frac{x_j - x_i}{r_{ij}}\,, \qquad r_{ij} \;=\; |x_i - x_j|\,,$$

ausgeübt. Die Summe $\sum_{j\neq i} F_{ij}$ ist die Gesamtkraft, die auf das i-te Partikel wirkt. Diese Kräfte führen zu einem Gleichgewichtszustand mit einer gitterförmigen Anordnung der einzelnen Teilchen. Abbildung 63.3 zeigt links eine

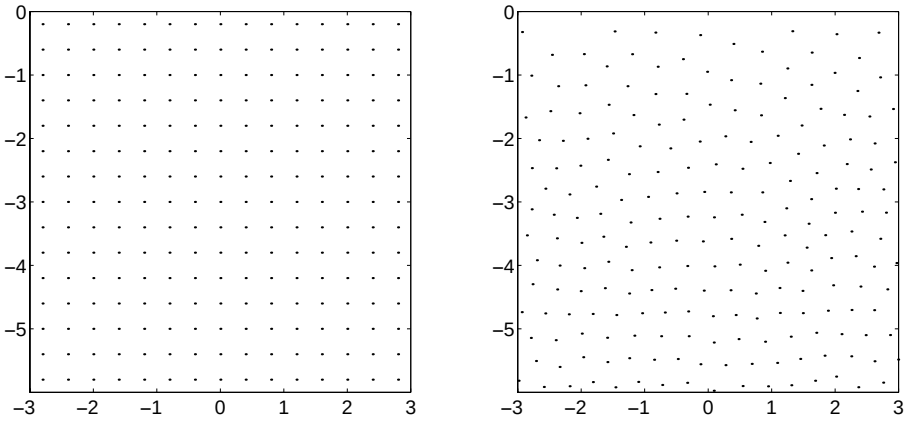


Abb. 63.3: Gitterartige Strukturen zwischen den einzelnen Molekülen

künstlich als Startzustand vorgegebene viereckige Gitterstruktur; rechts sieht man sehr schön im Übergang von oben nach unten, wie sich im Verlauf der Simulation allmählich eine mehr dreieckige Gitterstruktur einstellt.

Zur Simulation eines sich von der Decke loslösenden Wassertropfens verwenden wir nun dieses Modell und verteilen in einem Halbkreis

$$\{x = (\xi, \eta) \,:\, |x| < 14\,,\ \eta < 0\}$$

knapp 2000 Partikel in einer viereckigen Gitteranordnung (vgl. Abbildung 63.3 links) mit Gitterweite $h = 0.4$; die Achse $\eta = 0$ entspricht der Decke und die Halbebene $\eta < 0$ dem Raum, in den der Tropfen fallen soll. Zu Beginn wird die Geschwindigkeit aller Partikel mit Null initialisiert.[4]

Neben den oben beschriebenen Wechselwirkungen der einzelnen Partikel untereinander wirkt auf jedes Molekül noch die Schwerkraft, also eine Kraft $G = -mge_2$ mit der Erdbeschleunigung g und dem nach oben weisenden kartesischen Koordinatenvektor e_2. Um die Adhäsionskräfte an der Decke zu modellieren, ergänzen wir wie in [37] die bisher beschriebene Ausgangsstruktur durch eine zusätzliche (oberste) Gitterreihe mit unbeweglichen (Decken-)Partikeln gleicher Masse, die auf die eigentlichen Partikel ähnliche Kräfte wie die Wassermoleküle untereinander ausüben, vgl. (63.3), lediglich um den Faktor 5/4 verstärkt. Dadurch sind die Kräfte und die Anfangsbedingungen an die Wasserpartikel für das Differentialgleichungssystem festgelegt.

Erstaunlicherweise führt diese relativ grobe Modellierung zu realistischen Ergebnissen, vgl. Abbildung 63.4. Man erkennt zunächst einen Übergang zwi-

[4] Alle Zahlenangaben in diesem Abschnitt beziehen sich auf ein dimensionsloses Koordinatensystem, da lediglich das qualitative Verhalten von Interesse ist.

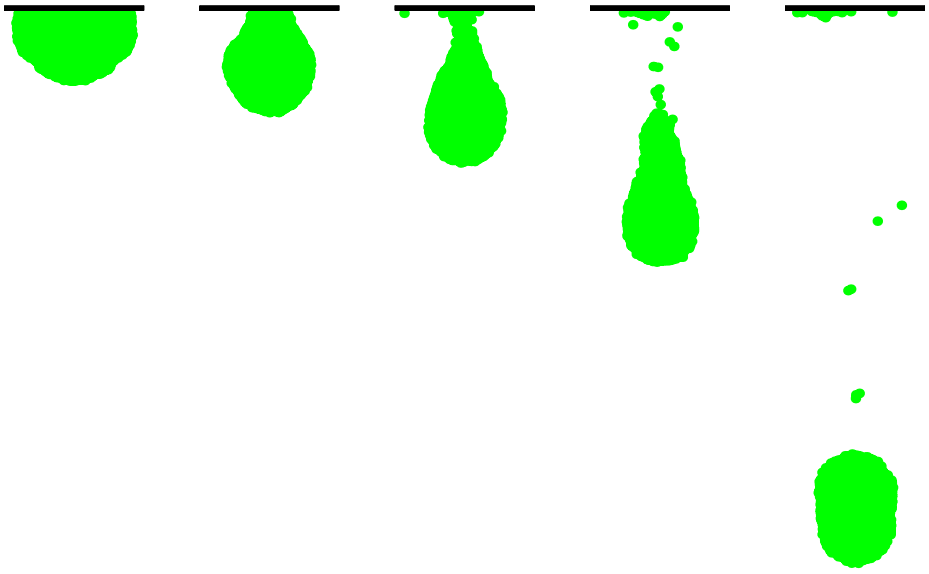


Abb. 63.4: Simulation eines sich ablösenden Tropfens

schen dem wenigen an der Decke verbleibenden Wasser und dem sich loslösenden Tropfen, der immer länger gestreckt wird bis er schließlich abreißt. Die längliche Form des Tropfens wird später im Fall wieder kreisförmig.

Simulationen mit derart vielen Partikeln erreichen leicht die Grenzen heutiger Rechenkapazität, wenn alle Wechselwirkungen zwischen den einzelnen Molekülen ausgewertet werden müssen (bei einer Menge von n Partikeln ergibt dies einen Aufwand $O(n^2)$). Moderne numerische Verfahren versuchen daher, die relativ geringen Wechselwirkungen zwischen weit entfernten Partikeln geschickt zu approximieren. In dem hier gerechneten Beispiel wurde der Aufwand dadurch reduziert, daß Wechselwirkungen zwischen Partikeln mit $r_{ij} > 4$ vernachlässigt wurden.

Auf die Strömung eines Fluids werden wir in Abschnitt 67 im Rahmen eines anderen Modells zurückkommen.

63.3 Restringierte Mehrkörpersysteme

Bei mechanischen Mehrkörpersystemen (etwa bei der Modellierung von Autofahrwerken oder Schienenfahrzeugen, vgl. die Beispiele in dem Buch von Eich-Soellner und Führer [24]) sind einzelne Komponenten oftmals über starre oder flexible Verbindungen aneinander gekoppelt. Dies führt zu Nebenbedingungen

an die Ortskoordinaten der einzelnen Körper, die in die mathematischen Gleichungen aufgenommen werden müssen; man spricht dann von *restringierten Mehrkörperproblemen*.

Für die folgende Darstellung sammeln wir die Ortskoordinaten x_i und die Kraftvektoren F_i, $i = 1, \ldots, n$, der n Körper in Vektoren $x \in \mathbb{R}^{3n}$ und $F \in \mathbb{R}^{3n}$; bei einem mechanischen Mehrkörpersystem hängt F in der Regel von der Zeit, den Positionen der einzelnen Körper und ihren Geschwindigkeiten ab, $F = F(t, x, x')$. Die Bewegungsgleichungen (63.1) können dann in der Form

$$Mx'' = F(t, x, x') \tag{63.4}$$

mit zugehöriger Massematrix

$$M = \begin{bmatrix} m_1 I & & & \\ & m_2 I & & \\ & & \ddots & \\ & & & m_n I \end{bmatrix} \in \mathbb{R}^{3n \times 3n}$$

geschrieben werden.

Die Nebenbedingungen an die Körper seien durch die algebraische Gleichung

$$g(x) = 0 \tag{63.5}$$

gegeben, wobei die Funktion $g : \mathbb{R}^{3n} \to \mathbb{R}^p$ mit $p < 3n$ hinreichend glatt sein und $g'(x)$ für jedes x vollen Zeilenrang besitzen soll. Denkbar ist auch eine allgemeinere Situation, in der g zusätzlich von der Zeit und der Geschwindigkeit der einzelnen Körper abhängt.

Beispiel 63.1. Als einfaches Beispiel stellen wir uns eine Punktmasse vor, die unter dem Einfluß der Schwerkraft längs einer Kurve $x(t)$ auf einer zusammenhängenden Fläche im $\mathbb{R}^3$ gleitet. Die Fläche werde durch die implizite Gleichung (63.5) mit einer skalaren Funktion $g : \mathbb{R}^3 \to \mathbb{R}$ repräsentiert. Die Voraussetzung, daß g' vollen Zeilenrang hat, impliziert, daß die Gleichung $z \cdot \operatorname{grad} g(x) = 0$ in jedem Punkt x der Fläche zwei linear unabhängige Lösungen $z \in \mathbb{R}^3$ besitzt. Dies bedeutet, daß in jedem Punkt der Fläche eine wohldefinierte Tangentialebene existiert, auf der der Vektor $\operatorname{grad} g(x)$ senkrecht steht. ◊

Die für die Bewegung nötigen Anfangsbedingungen an x und x' zu einem Zeitpunkt $t = 0$ sind für das restringierte Mehrkörperproblem nicht mehr völlig frei wählbar. Offensichtlich muß die Anfangsvorgabe $x(0)$ die Nebenbedingung $g(x(0)) = 0$ erfüllen. Aber auch die Geschwindigkeit $x'(0)$ ist eingeschränkt,

denn durch Differentiation der Nebenbedingung (63.5) nach der Zeit ergibt sich eine *versteckte Nebenbedingung*

$$g'(x)x' \;=\; 0\,. \tag{63.6}$$

Für Beispiel 63.1 besagt diese zweite Nebenbedingung, daß die Geschwindigkeit des Körpers tangential zur Fläche ist. Man beachte, vgl. Aufgabe 8, daß unter der Voraussetzung $g(x(0)) = 0$ die Nebenbedingungen (63.5) und (63.6) zueinander äquivalent sind.

Die Lösung der herkömmlichen Bewegungsgleichung (63.4) für den unrestringierten Fall wird selbst bei zulässigen Anfangswerten $x(0)$ und $x'(0)$ in der Regel nicht die beiden Nebenbedingungen (63.5) und (63.6) erfüllen. Statt dessen müssen die äußeren Kräfte F in (63.4) um eine *Zwangskraft* Z ergänzt werden, die die Nebenbedingung ersetzt (analog zu dem Schnittprinzip in Abschnitt 3). Die Zwangskraft steht lokal immer senkrecht zum Tangentialraum an die Menge $\{x \,:\, g(x) = 0\}$, wirkt also in Tangentialrichtung nicht beschleunigend. Da der Tangentialraum durch $\mathcal{N}(g'(x)) \;=\; \mathcal{R}(g'(x)^*)^\perp$ gegeben ist, folgt hieraus die Darstellung

$$Z = g'(x)^*\lambda \qquad \text{für ein } \lambda \in \mathbb{R}^p\,. \tag{63.7}$$

Somit ist die *differential-algebraische Gleichung*

$$\begin{aligned} Mx'' \;&=\; F(t,x,x') + g'(x)^*\lambda\,,\\ 0 \;&=\; g(x)\,, \end{aligned} \tag{63.8}$$

die korrekte Bewegungsgleichung (die sogenannte *Euler-Lagrange-Gleichung*) für das restringierte Mehrkörpersystem (vgl. auch die Darstellung in Rabier und Rheinboldt [88, Chapter 2]).

Der Vektor λ ist durch das Gleichungssystem (63.8) eindeutig festgelegt. Leiten wir nämlich die versteckte Nebenbedingung (63.6) ein weiteres Mal nach der Zeit ab,

$$g'(x)x'' \;+\; g''(x)(x',x') \;=\; 0\,, \tag{63.9}$$

so erhalten wir mit (63.8) das positiv definite lineare Gleichungssystem

$$g'(x)M^{-1}g'(x)^*\lambda \;=\; -g'(x)M^{-1}F(t,x,x') - g''(x)(x',x') \tag{63.10}$$

für λ. Eingesetzt in (63.8) ergibt sich hieraus für x die gewöhnliche Differentialgleichung zweiter Ordnung

$$Mx'' \;=\; F - G^*(GM^{-1}G^*)^{-1}\big(GM^{-1}F + H\big) \tag{63.11}$$

mit $F = F(t, x, x')$, $G = g'(x)$ und $H = g''(x)(x', x')$, die für konsistente Anfangsbedingungen zu dem differential-algebraischen System (63.8) äquivalent ist. Für die Numerik ist diese Formulierung allerdings nicht von Vorteil, da die resultierenden Näherungslösungen wegen des Diskretisierungsfehlers im allgemeinen nicht die Nebenbedingung einhalten.

Beispiel. Bei dem Spezialfall aus Beispiel 63.1 ist λ ein Skalar und die Zwangskraft $Z = \lambda \operatorname{grad} g(x)$ steht senkrecht zur Tangentialebene in dem Auflagepunkt des Körpers. Sie entspricht der Kraft, die dem Körper von der Fläche entgegengesetzt wird. Diese Zwangskraft ist auch für die Modellierung der Reibungskraft von Bedeutung. Die Gleitreibungskraft ist nach dem *Coulombschen Reibungsgesetz* nämlich proportional zu der Stärke der Zwangskraft und ihre Richtung ist der Geschwindigkeit entgegengesetzt. Ein vollständigeres Modell für die Bewegung des Körpers auf der Fläche wird somit durch die differential-algebraische Gleichung

$$\begin{aligned} mx'' &= F - \mu|\lambda| \frac{|\operatorname{grad} g(x)|}{|x'|} x' + \lambda \operatorname{grad} g(x)\,, \\ 0 &= g(x)\,, \end{aligned} \tag{63.12}$$

beschrieben, bei der $\mu > 0$ der Gleitreibungskoeffizient und m die Masse des Teilchens ist.

Das Vorzeichen des Lagrange-Parameters λ gibt Aufschluß über die Richtung der Zwangskraft. Um dies zu veranschaulichen, betrachten wir ein anderes Beispiel, nämlich eine Eisenbahn, die sich auf Schienen in der Ebene (im $\mathbb{R}^2$) bewegt. Die Schienenstrecke werde durch die Menge $g(x) = 0$ repräsentiert, wobei $g : \mathbb{R}^2 \to \mathbb{R}$ links der Bahnlinie negativ und rechts der Bahnlinie positiv sein soll. In diesem Beispiel ist der Lagrange-Parameter λ aus (63.8) positiv, falls die Räder der Bahn von rechts gegen die Schienen drücken, und negativ, falls die Räder von links gegen die Schienen drücken.

Zum Abschluß betrachten wir eine Punktmasse auf der Fläche, die durch den Graph der Funktion

$$\Phi(x_1, x_2) = \big(100(x_2 - x_1^2)^2 + (x_1 - 1)^2\big)/2$$

aus Beispiel 20.2 gegeben ist, d. h. $g(x_1, x_2, x_3) = x_3 - \Phi(x_1, x_2)$. Wir wollen die Bahn dieser Punktmasse unter dem Einfluß der Schwerkraft berechnen und vernachlässigen dabei den Reibungsterm. Statt dessen sei zugelassen, daß das Teilchen bei zu hoher Geschwindigkeit von der Fläche abhebt. Im Gegensatz zu (63.5) ist die korrekte Nebenbedingung für dieses Problem die *Ungleichung* $g(x) \geq 0$ und die Zwangskraft tritt nur dann auf, wenn die Punktmasse gegen die Fläche drückt, also wenn λ positiv ist. Sobald λ negativ wird, verliert der

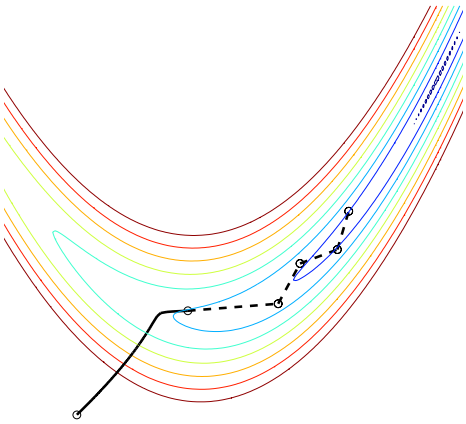


Abb. 63.5:
Bewegung eines Körpers auf einer Fläche

Körper den Kontakt zur Fläche. Ab diesem Moment gilt die unrestringierte Bewegungsgleichung (63.4) und der Körper beschreibt wie beim schiefen Wurf einen parabelförmigen Bogen, bis er wieder auf die Fläche auftrifft.

Abbildung 63.5 zeigt die entsprechende Bahn eines Teilchens, das mit Geschwindigkeit $v = 0$ im Punkt $(-0.5, -0.4, \Phi(-0.5, -0.4))$ auf der Fläche startet; die Flugphasen sind als gebrochene Linien eingezeichnet. Wie man sieht, gleitet das Masseteilchen zunächst in das bananenförmige Tal der Fläche; aufgrund der Neigung des Talbodens beschreibt das Teilchen dann eine Rechtskurve und kann bald darauf nicht mehr der entgegengesetzten Talkrümmung folgen. In diesem Moment verliert das Teilchen den Kontakt zur Fläche und prallt in der Folge zwischen den Talwänden hin und her. Wir gehen dabei davon aus, daß das Masseteilchen vollelastisch ist: Daher wird die Geschwindigkeit nach dem Aufprall durch eine Spiegelung der alten Geschwindigkeit an der Flächennormalen bestimmt („Einfallswinkel = Ausfallswinkel“). ◊

64 Elektrische Schaltkreise

Die Simulation umfangreicher elektrischer Schaltkreise ist eine Standardaufgabe im Chipdesign. Die individuellen Eigenschaften der einzelnen Bauteile eines Schaltkreises (Widerstände, Spulen, Kondensatoren, etc.) führen unter Zuhilfenahme des *Kirchhoffschen Gesetzes* der Stromerhaltung auf ein System von Differentialgleichungen und algebraischen Gleichungen, dessen Lösung das zeitliche Verhalten der elektrischen Spannungen innerhalb des Schaltkreises beschreibt, vgl. etwa das Buch von Vlach und Singhal [105]. Als Beispiel betrachten wir die sogenannte *Emitterschaltung* aus Abbildung 64.1, eine gängige Verstärkerstufe für eine Eingangs-Wechselspannung u_S mit Hilfe eines

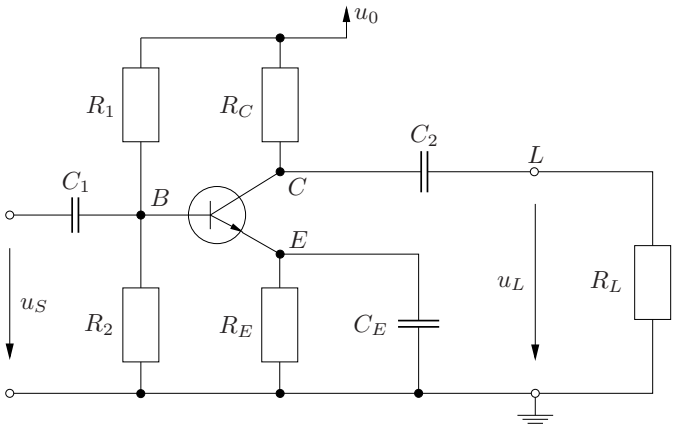


Abb. 64.1: Verstärker-Grundschaltung

npn-Transistors.

Der Schaltkreis enthält Widerstände $R_1, R_2, \ldots$, Kondensatoren $C_1, C_2, \ldots$, sowie einen Transistor im Zentrum der Schaltung. Der Widerstand R_L hängt am Ausgang L der Schaltung und repräsentiert eine Last. An den Widerständen genügen die Spannung u und der elektrische Strom[5] ι dem *Ohmschen Gesetz*

$$u = r\iota\,,$$

wobei wir im weiteren mit kleinen Buchstaben $r = r_1, r_2, \ldots$, den jeweiligen Ohmschen Widerstand bezeichnen. An den Kondensatoren ist der elektrische Strom proportional zu Spannungsänderungen,

$$\iota = cu'\,.$$

Die Proportionalitätskonstanten $c = c_1, c_2, \ldots$, sind die *elektrischen Kapazitäten* der Kondensatoren. Ist u die Überlagerung einer Gleichspannung u_0 und einer sinusförmigen Wechselspannung mit Kreisfrequenz ω und Amplitude u_1,

$$u(t) = u_0 + u_1 \sin \omega t\,,$$

so blockiert der Kondensator den Gleichstromanteil und verhält sich bezüglich des Wechselstromanteils – abgesehen von einer Phasenverschiebung – wie ein Widerstand mit $r = 1/(\omega c)$.

[5] Wir folgen der Konvention, daß der Strom „von Plus nach Minus“ fließt, also in die Richtung des Spannungsabfalls.

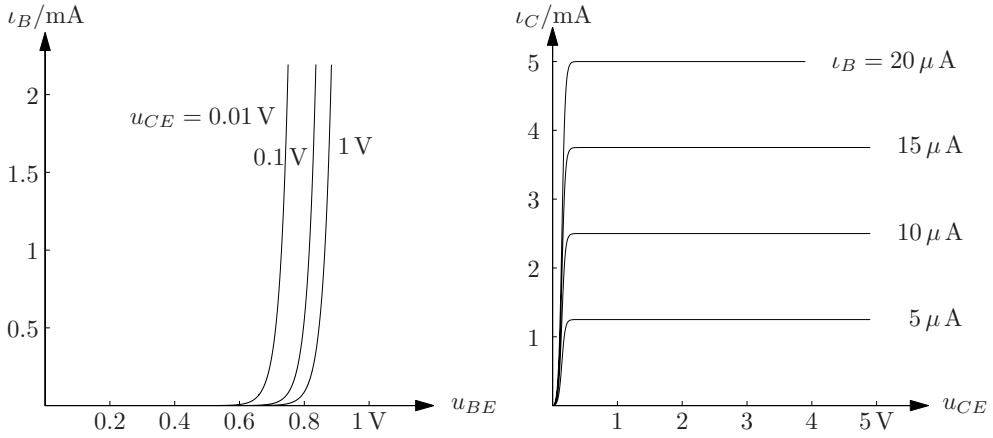


Abb. 64.2: Kennlinien eines Transistors

Das entscheidende Bauteil der Verstärkerschaltung ist der Transistor, der aus drei Zonen besteht, dem Kollektor C, der Basis B in der Mitte und dem Emitter E, an die unterschiedliche Spannungen angelegt werden können. Durch die Spannungen sind auch die elektrischen Ströme zwischen den drei Zonen festgelegt, allerdings sind die Abhängigkeiten komplizierter als bei einem Widerstand oder einem Kondensator. Hierzu kommen wir als nächstes.

Dazu bezeichnen wir mit ι_C, ι_B und ι_E die drei Ströme, die an den Anschlüssen C, B und E in den Transistor fließen sowie mit u_C, u_B und u_E die jeweiligen Potentiale an den Anschlüssen. Nach dem Kirchhoffschen Gesetz ist einer dieser Ströme redundant,

$$\iota_E \;=\; -\iota_B - \iota_C\,, \tag{64.1}$$

die anderen beiden hängen von den angelegten Spannungen (Potentialdifferenzen) ab und werden durch sogenannte *Transistor-Kennlinien* angegeben, vgl. Abbildung 64.2. Die *Eingangskennlinien* (links im Bild) stellen die funktionale Abhängigkeit

$$\iota_B \;=\; \varphi(u_{BE}, u_{CE})\,, \qquad \begin{aligned} u_{BE} &= u_B - u_E\,, \\ u_{CE} &= u_C - u_E\,, \end{aligned} \tag{64.2a}$$

für verschiedene Parameter u_{CE} dar; Kurven für $u_{CE} \geq 1\,\mathrm{V}$ sind vom Auge nicht zu unterscheiden. Die *Ausgangskennlinien* (rechts im Bild) geben die Zuordnung

$$\iota_C \;=\; \psi(u_{CE}, \iota_B) \tag{64.2b}$$

für verschiedene Parameterwerte ι_B wieder.

Für größere Werte von u_{CE} liegen die Eingangskennlinien praktisch übereinander und die Ausgangskennlinien sind fast konstant; dies ist der Arbeitsbereich des Transistors. Um in den Arbeitsbereich zu kommen, wird die Wechselspannung u_S mit Hilfe einer positiven Spannung u_0 (z. B. aus einer Batterie) verschoben und zwischen Kollektor und Emitter geführt. Als Folge wird der an der Basis fließende Strom ι_B nahezu ausschließlich durch u_{BE} gesteuert und im Kollektorstrom (64.2b) etwa auf das 250fache verstärkt (vgl. Abbildung 64.2). Nach (64.1) haben dann Kollektorstrom und Emitterstrom fast den gleichen Betrag, die Spannungen an den Widerständen R_E und R_C stehen also ungefähr im Verhältnis r_C/r_E. Dies sorgt für die gewünschte Spannungsverstärkung. Die beiden Kondensatoren C_1 und C_2 am Eingang beziehungsweise Ausgang der Schaltung dienen als Sperre für die Gleichspannung. Außerdem bewirken sie eine Phasenverschiebung von 180° zwischen u_S und der Ausgangsspannung u_L.

Wir wollen nun die elektrischen Potentiale an den Transistoreingängen sowie die Ausgangsspannung an der Last bestimmen. Mit dem Kirchhoffschen Gesetz erhalten wir ein entsprechendes Gleichungssystem, indem wir die Summe aller eingehenden Ströme an den vier Knoten B, C, E und L jeweils gleich Null setzen. Unter Berücksichtigung von (64.1) liefert dies die vier Gleichungen

$$\begin{aligned}
0 &= -\iota_B + (u_0 - u_B)/r_1 + (u_S - u_B)'c_1 - u_B/r_2\,,\\
0 &= (u_L - u_C)'c_2 + (u_0 - u_C)/r_C - \iota_C\,,\\
0 &= -u_E'c_E + \iota_B + \iota_C - u_E/r_E\,,\\
0 &= -u_L/r_L + (u_C - u_L)'c_2\,,
\end{aligned}$$

die wir in Form eines Differentialgleichungssystems

$$Mx' = \phi(x) \tag{64.3}$$

für die unbekannten Potentiale $x = [u_B, u_C, u_E, u_L]^T$ mit Koeffizientenmatrix

$$M = \begin{bmatrix} c_1 & & & \\ & c_2 & & -c_2 \\ & & c_E & \\ & -c_2 & & c_2 \end{bmatrix} \tag{64.4}$$

und entsprechender rechter Seite ϕ schreiben. Die Matrix M ist aufgrund der algebraischen Gleichungen des Ohmschen Gesetzes singulär. Daher ist (64.3) keine gewöhnliche Differentialgleichung, sondern ein differential-algebraisches Gleichungssystem.

Für numerische Simulationen kann das *Ebers-Moll-Modell* zur quantitativen Approximation der Kennlinien aus Abbildung 64.2 herangezogen werden (vgl.

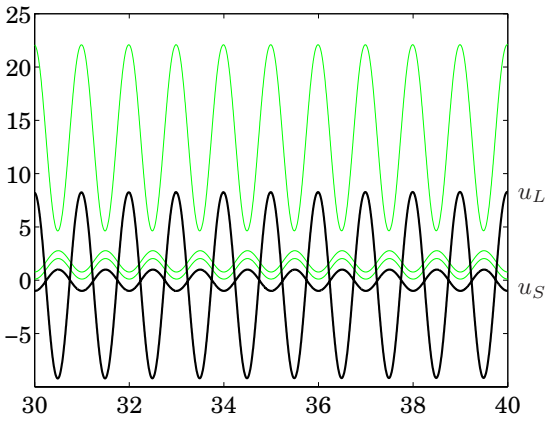


Abb. 64.3: u_S und u_L über der Zeit

Vlach und Singhal [105]): Nach diesem Modell ist

$$\begin{aligned} \varphi(u,v) &= \alpha(e^{u/u_*}-1) + \beta(e^{(u-v)/u_*}-1)\,, \\ \psi(v,\iota) &= (\iota+\alpha+\beta)\frac{\gamma e^{v/u_*}-\delta}{\alpha e^{v/u_*}+\beta} - \gamma + \delta\,, \end{aligned} \tag{64.5}$$

für gewisse positive Parameter α, β, γ, δ und u_* (u und v stehen hier für u_{BE} bzw. u_{CE}, ι ist durch ι_B zu ersetzen). Für $u_{CE} \gg u_*$ ist der Transistor im Arbeitsbereich: Änderungen von ι_B werden dann im Kollektorstrom ungefähr um den Faktor γ/α verstärkt. In Analogie zu (64.2a) sei der Vollständigkeit halber noch eine alternative Darstellung von ι_C angeführt,

$$\iota_C = \gamma(e^{u_{BE}/u_*}-1) - \delta(e^{(u_{BE}-u_{CE})/u_*}-1)\,, \tag{64.6}$$

die mit elementaren Umformungen aus (64.2) und (64.5) folgt.

Abbildung 64.3 zeigt das Resultat einer numerischen Simulation für die Parameter

$$\begin{aligned} &r_1 = 240\,, \quad r_2 = 20\,, \quad r_C = 10\,, \quad r_E = 1\,, \quad r_L = 100\,, \\ &c_1 = 10^{-3}\,, \quad c_2 = 10^{-3}\,, \quad c_E = 10^{-5}\,, \quad u_0 = 24\,, \end{aligned}$$

und Eingangs-Wechselspannung $u_S(t) = -\cos(2000\pi t)$; die Ohmschen Widerstände sind in kΩ, die Spannungen in V und die elektrischen Kapazitäten in mF=s/kΩ angegeben. Der Transistor wird duch die Ebers-Moll-Parameter

$$\alpha = 3.8\cdot 10^{-15}\,, \quad \beta = 9.5\cdot 10^{-13}\,, \quad \gamma = 9.5\cdot 10^{-13}\,, \quad \delta = 1.9\cdot 10^{-12}$$

(jeweils mit Einheit mA=V/kΩ) und $u_* = 0.026\,\mathrm{V}$ modelliert.

Nach einer kurzen Einschwingzeit ergibt sich die gleichmäßige Schwingung aus Abbildung 64.3 (dort ist eine Zeitperiode von 10 ms dargestellt). Die beiden dunklen Linien in der Abbildung zeigen die Eingangsspannung u_S und die Ausgangsspannung u_L. Wie man sieht, ist die Phase von u_L verschoben und die Amplitude etwa um den Faktor neun verstärkt (zum Vergleich: $r_C/r_E = 10$). Die drei helleren Kurven geben (in der Reihenfolge von unten nach oben) die Potentiale an Emitter, Basis und Kollektor wieder.

Aufgaben

1. Zeigen Sie, daß die Gleichung (60.8) eine geschlossene Kurve definiert. Gehen Sie dazu wie folgt vor:

(a) Skizzieren Sie für $b, d > 0$ die Funktion $f(x) = x^d e^{-bx}$ über $\mathbb{R}_0^+$ und folgern Sie, daß die Gleichung (60.8) bei gegebenen $c, x_2 > 0$

– genau eine Lösung x_1 hat, falls

$$\frac{ce^{rx_2}}{x_2^{d_1}} = \left(\frac{d_2}{be}\right)^{d_2};$$

– genau zwei (verschiedene) positive Lösungen x_1 besitzt, wenn

$$\frac{ce^{rx_2}}{x_2^{d_1}} < \left(\frac{d_2}{be}\right)^{d_2}.$$

(b) Schließen Sie daraus, daß für $c < (d_1/er)^{d_1}(d_2/eb)^{d_2}$ ein Intervall $[\alpha, \beta] \subset \mathbb{R}^+$ existiert, so daß die Gleichung (60.8)

– keine Lösung x_1 hat, falls $x_2 \notin [\alpha, \beta]$;
– genau eine Lösung $x_1 = b/d_2$ besitzt, falls $x_2 \in \{\alpha, \beta\}$;
– genau zwei (verschiedene) positive Lösungen x_1 besitzt, falls $x_2 \in (\alpha, \beta)$.

Skizzieren Sie alle Lösungen in der (x_1, x_2)-Ebene.

2. Einer Firma gelingt ein entscheidender technologischer Durchbruch und sie bringt mit großem Werbeetat ein neues Gerät auf den Markt, daß der gesamten Konkurrenz deutlich überlegen ist. Neben der Werbung wirkt auch die Mund-zu-Mund-Propaganda der zufriedenen Käufer verkausfördernd.
Entwickeln Sie ein mathematisches Modell für die Anzahl $x = x(t)$ der Besitzer eines solchen Geräts. Gehen Sie von einer festen Schranke X für die Zahl aller potentiellen Kunden in der Bevölkerung aus. Bestimmen Sie die Lösung ihrer Modellgleichung.

3. Die chemische Reaktion $A + B \longrightarrow C$ gehorche dem Massenwirkungsgesetz mit Reaktionskoeffizient k. Bestimmen Sie die Konzentration $y = y(t)$ von C unter der Annahme, daß die Konzentrationen von A, B und C zur Zeit $t = 0$ durch α, β und $y(0) = 0$ gegeben sind.

4. Betrachtet werden n Himmelskörper, die sich unter dem Einfluß ihrer Gravitationskräfte bewegen. Zeigen Sie, daß sich der Schwerpunkt dieser Himmelskörper mit konstanter Geschwindigkeit bewegt.

5. Gegeben sei das Anfangswertproblem aus (63.2):

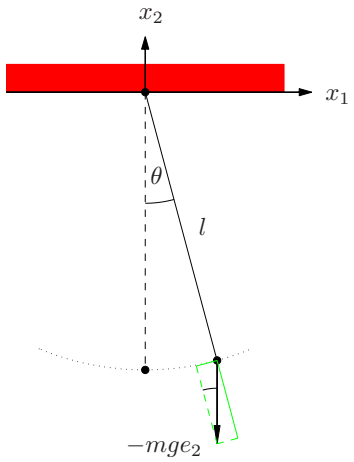
$$x'' = -\frac{2}{9x^2}, \qquad x(0) = 1, \; x'(0) = \alpha > 0.$$

(a) Beweisen Sie, daß die Lösung für $\alpha \geq 2/3$ für alle $t > 0$ existiert mit $x(t) \to \infty$ für $t \to \infty$. Bestimmen Sie in diesem Fall $\lim_{t\to\infty} x'(t)$.
(b) Argumentieren Sie, warum die Lösung für $\alpha < 2/3$ nur für endliche Zeit existiert.
Hinweis: Multiplizieren Sie die Differentialgleichung zunächst mit x'.

6. Leiten Sie die Differentialgleichung

$$l\,\theta'' = -g\sin\theta$$

für die Auslenkung $\theta(t)$ eines mathematischen Pendels der Masse m und Länge l zur Zeit t her, vgl. Skizze. Hierbei bezeichnet g die Erdbeschleunigung. Schneiden Sie dazu den Massepunkt frei (vgl. Abschnitt 3 und insbesondere Abbildung 3.2) und stellen Sie das Kräftegleichgewicht tangential zur Bahnkurve des Massepunkts auf.



7. Das mathematische Pendel aus der vorigen Aufgabe kann auch in den Ortskoordinaten (x_1, x_2) beschrieben werden. Führen Sie hierzu wie in Beispiel 63.1 die Zwangskraft ein, die auf das Pendel wirkt, und leiten Sie das zugehörgie differential-algebraische Gleichungssystem für die Orts- und Geschwindigkeitsvariablen her.

8. (a) Beweisen Sie, daß die differential-algebraische Gleichung

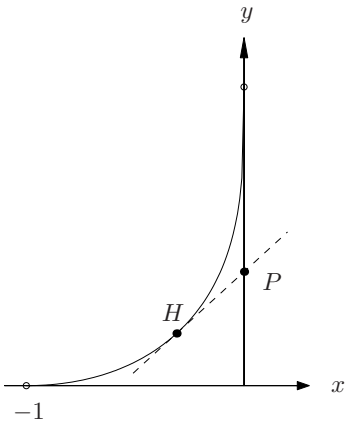
$$\begin{aligned} Mx'' &= F(t,x,x') + g'(x)^*\lambda\,, & x(0) &= x_0\,,\\ 0 &= g(x)\,, & x'(0) &= x_0'\,, \end{aligned}$$

mit konsistenten Anfangsbedingungen $g(x_0) = 0$ und $g'(x_0)x_0'$ äquivalent zu (63.8) ist.
(b) Zeigen Sie unter der gleichen Voraussetzung, daß auch die gewöhnliche Differentialgleichung (63.11) zu dem System (63.8) äquivalent ist.

9. Eine Person P befindet sich zur Zeit $t = 0$ im Nullpunkt und wandert mit konstanter Geschwindigkeit a in Richtung der positiven y–Achse. Ein Hund H läuft ausgehend vom Punkt $(-1, 0)$ mit gleichbleibender Geschwindigkeit c immer auf die Person zu. Die Position des Hundes zur Zeit t sei $(x(t), y(t))$, vgl. Skizze.
(a) Zeigen Sie, daß die Bahnkurve des Hundes in der xy–Ebene durch die Differentialgleichung

$$xy'' = -\frac{a}{c}\sqrt{1+y'^2}$$

beschrieben wird.
(b) Bestimmen Sie die Anfangswerte für die Differentialgleichung und lösen Sie das Anfangswertproblem. Unterscheiden Sie die Fälle $c > a$ (Hund ist schneller), $c < a$ (Person ist schneller) und $c = a$ (beide sind gleich schnell). Diskutieren Sie das Ergebnis.



Hinweis zu (a): Substituieren Sie zunächst $w(x) = y'(x)$.

XII Erhaltungsgleichungen

Im Gegensatz zu den Partikelmodellen aus Abschnitt 63.2 werden bei der *makroskopischen Sichtweise* alle wesentlichen physikalischen Größen als kontinuierlich angenommen.[1] Räumliche und zeitliche Veränderungen dieser Größen genügen häufig Erhaltungsgesetzen, die unter hinreichenden Glattheitsvoraussetzungen zu partiellen Differentialgleichungen äquivalent sind. Der nachfolgende Abschnitt gibt eine Einführung in dieses Prinzip. Für eine umfassendere und rigorosere Behandlung der physikalischen und mathematischen Grundlagen sei etwa auf das Buch von Rubinstein und Rubinstein [92] verwiesen.

65 Integrale und differentielle Erhaltungsform

Im folgenden bezeichnet $\Omega \subset \mathbb{R}^d$ ein Gebiet, $x \in \Omega$ die Ortsvariable und $t \geq 0$ die Zeit. Mit $|x|$ wird die Euklidnorm der Ortsvariablen und mit $x \cdot y$ das Innenprodukt zweier Ortsvektoren $x, y \in \mathbb{R}^d$ bezeichnet.

Sind $u(x,t)$ die *Dichte* und $v = v(x,t)$ die *Geschwindigkeit* einer zu betrachtenden Erhaltungsgröße, so gibt das Vektorfeld $J = uv \in \mathbb{R}^d$ den *Fluß* dieser Größe an: Ist G ein beliebiges glatt berandetes und beschränktes Teilgebiet von Ω und ν die äußere Normale im Punkt $x \in \partial G$, so ist $\nu \cdot J$ die Menge der Erhaltungsgröße, die das Gebiet G pro Zeiteinheit durch ein infinitesimales Oberflächenelement ds im Punkt $x \in \partial G$ verläßt.

Aufgrund der Erhaltungseigenschaft muß der gesamte Abfluß durch den Rand ∂G mit der zeitlichen Abnahme der Erhaltungsgröße in G übereinstimmen:

$$\int_{\partial G} \nu \cdot J \, ds \;=\; -\frac{d}{dt} \int_G u \, dx \,.$$

[1] Ein Bezug zwischen der makroskopischen Betrachtungsweise auf der einen und Partikelmodellen auf der anderen Seite läßt sich herstellen, indem man jeden Punkt im Ort mit einem Volumenelement identifiziert, das eine sehr große Zahl von Partikeln enthält. Dann können die kontinuierlichen Größen des makroskopischen Modells als Mittelwerte der entsprechenden Eigenschaften aller Partikel in dem Volumenelement interpretiert werden.

Sind in G zudem noch „Quellen“ oder „Senken“, d. h. Zu- oder Abflüsse zu berücksichtigen, dann muß die Integralidentität durch einen Term mit der *Quelldichte* f korrigiert werden,

$$\frac{d}{dt}\int_G u\,dx \;=\; -\int_{\partial G} \nu\cdot J\,ds \;+\; \int_G f\,dx\,. \tag{65.1}$$

Unter der Annahme, daß Differentiation und Integration vertauscht werden können, liefert der Gaußsche Satz (siehe etwa Heuser [53, Satz 210.1])[2]

$$\int_G \big(u_t + \operatorname{div} J\big)\,dx \;=\; \int_G f\,dx\,.$$

Da $G \subset \Omega$ beliebig klein gewählt werden kann, ist diese Gleichung nur dann für alle Gebiete G erfüllt, wenn die beiden Integranden überall übereinstimmen, also wenn gilt

$$u_t + \operatorname{div} J = f \qquad \text{in } \Omega\,. \tag{65.2}$$

Die Darstellungen (65.1) und (65.2) sind unterschiedliche Formen desselben Erhaltungsgesetzes, einmal als Integralidentität und einmal als partielle Differentialgleichung. Hängt J nur von u, aber nicht von den partiellen Ableitungen von u ab, so spricht man von einer *hyperbolischen Differentialgleichung* erster Ordnung Wir werden später wiederholt sehen, daß der Gebrauch der Differentialgleichung mit Vorsicht erfolgen muß, wenn die Lösung der Erhaltungsgleichung nicht die nötigen Glattheitseigenschaften besitzt.

Beispiel 65.1. Als einfaches Beispiel betrachten wir die Ausbreitung einer Wasserverschmutzung in einem dünnen Leitungsrohr der Länge L. Wir modellieren das Rohr als eindimensional, d. h. wir wählen $\Omega = (0, L)$ und bezeichnen mit $u = u(x,t)$ die Dichte der Verunreinigung im Ort $x \in \Omega$ zur Zeit t. Aufgrund der Bewegung des Wassers wird die Verschmutzung durch das Rohr transportiert. Solange Diffusionsprozesse ignoriert werden können (vgl. Beispiel 69.1), ist der Fluß gleich $J = au$, wobei a die Fließgeschwindigkeit des Wassers ist. Abhängig von dem Vorzeichen von a weist der Fluß entweder in die positive oder negative Richtung. Die Differentialgleichung (65.2) für dieses Beispiel heißt *Transportgleichung*:

$$u_t + au_x = 0\,, \qquad x \in \Omega, \quad t > 0\,. \tag{65.3}$$

[2] Ist $J = [J_1, \dots, J_d]^T \in \mathbb{R}^d$ mit $J_i = J_i(x_1, \dots, x_d)$ ein differenzierbares Vektorfeld über Ω, so ist die *Divergenz* von J gegeben durch

$$\operatorname{div} J \;=\; \frac{\partial J_1}{\partial x_1} \;+\; \dots \;+\; \frac{\partial J_d}{\partial x_d}\,.$$

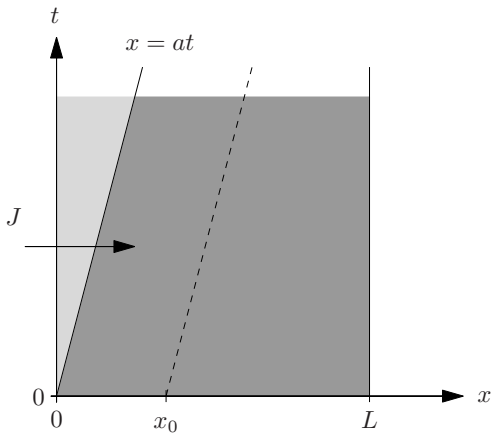


Abb. 65.1:
Die Gerade $x = at$ und die Abhängigkeit der Lösung von Rand- und Anfangswerten

Man sieht sofort, daß $u(x,t) = h(x-at)$ für jede differenzierbare Funktion h die Transportgleichung löst, denn es ist

$$u_t(x,t) = h'(x-at)\cdot(-a) \qquad \text{und} \qquad u_x(x,t) = h'(x-at)\,.$$

Sofern h wenigstens stetig ist, erfüllt u zumindest die integrale Form der Erhaltungsgleichung. Die Graphen der Funktionen $u(\,\cdot\,,t)$ „sehen für alle Zeiten $t > 0$ gleich aus“, sind lediglich um at gegenüber der Funktion h nach rechts (für $a > 0$) bzw. links (für $a < 0$) verschoben.

Nehmen wir für den Moment an, daß das Rohr in beide Richtungen unendlich ausgedehnt ist. In diesem Fall ist $\Omega = \mathbb{R}$ und die Lösung von (65.3) wird durch das Anfangswertproblem

$$u(x,0) = u^\circ(x)\,, \qquad x \in \Omega\,, \tag{65.4}$$

eindeutig bestimmt: sie hat die obige Form mit $h = u^\circ$. Bei einem beschränkten Intervall $\Omega = (0,L)$ ist die Anfangsvorgabe (65.4) hingegen nicht ausreichend: Für $a > 0$ ist beispielsweise der Verlauf der Lösung für $x < at$ unklar, vgl. den etwas heller eingezeichneten Bereich des Halbstreifens $(0,L)\times\mathbb{R}^+$ in Abbildung 65.1. Um u auch in diesem Bereich eindeutig festlegen zu können, bedarf es noch einer Randbedingung am linken Rand des Intervalls. Dies ist physikalisch sinnvoll, da dies der *Einflußrand* der Strömung ist, d. h. hier wird zusätzliche Materie in das Gebiet hineintransportiert. Entsprechend wird der gegenüberliegende Rand *Ausflußrand* genannt.

Etwas komplizierter wird das Modell, wenn sich ein fester Anteil der Verschmutzung jeweils am Boden des Rohrs absetzt. In diesem Fall liegt eine Senke f in (65.2) vor, deren Größe proportional zur Verschmutzung u ist, etwa $f = -cu$ mit einem festen $c > 0$. Die zugehörige Differentialgleichung lautet

$$u_t + a u_x \;=\; -cu$$

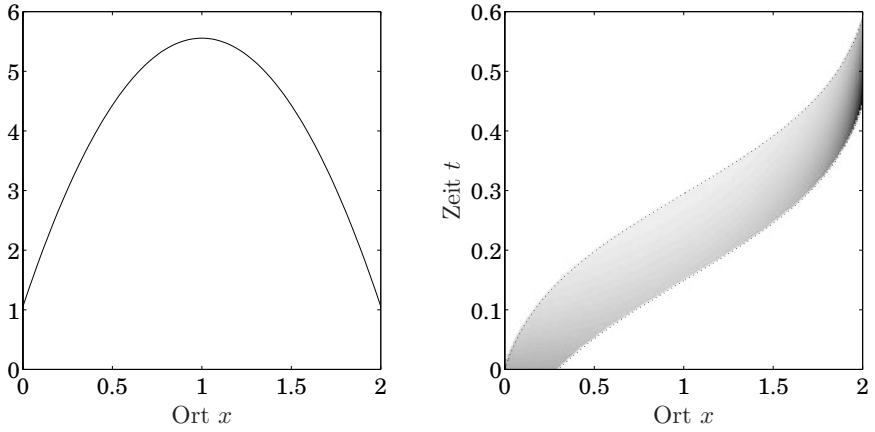


Abb. 65.2: Geschwindigkeitsprofil $a(x)$ (links) und Lösung u der Advektionsgleichung (rechts)

und hat die Lösung

$$u(x,t) \;=\; e^{-ct}u^\circ(x-at)\,.$$

Die Funktion u ist nun nicht mehr konstant gleich $u^\circ(x_0)$ längs der in Abbildung 65.1 gebrochen eingezeichneten Geraden $x(t) = x_0 + at$, sondern die Anfangsvorgabe $u^\circ(x_0)$ klingt entlang dieser Geraden exponentiell ab.

Während die Lösung bislang noch unmittelbar hingeschrieben werden konnte, ist dies nicht mehr möglich, wenn sich die Strömungsgeschwindigkeit innerhalb des Rohrs oder eines Rohrsystems verändert, etwa weil sich das Rohr verengt. Dieser Fall läßt sich durch eine ortsabhängige Fließgeschwindigkeit $a = a(x)$ modellieren und wir erhalten aus (65.2) die sogenannte *Advektionsgleichung*

$$u_t + (au)_x \;=\; u_t + au_x + a'u \;=\; 0 \qquad \text{mit} \qquad a' \;=\; \frac{d}{dx}\,a\,.$$

Für das parabolische Geschwindigkeitsprofil a aus Abbildung 65.2 links gibt das Graustufenbild im rechten Teil der Abbildung einen Eindruck der zugehörigen Lösung über der (x,t)-Ebene: je dunkler die Farbe, desto größer ist der Wert von u. Außerhalb der beiden gepunktet eingezeichneten Randkurven verschwindet die Lösung. Als Anfangsbedingung u° wurde hier die charakteristische Funktion des Intervalls $[0, 0.3]$ gewählt, das heißt

$$u^\circ(x) \;=\; \begin{cases} 0\,, & x < 0\,, \\ 1\,, & 0 \le x \le 0.3\,, \\ 0\,, & x > 0.3\,. \end{cases}$$

Man kann gut erkennen, wie die schnellere Fließgeschwindigkeit in der Mitte des Intervalls die Verschmutzung auseinanderzieht; umgekehrt staut sich die Verunreinigung am rechten Ende des x-Intervalls auf, wo die Geschwindigkeit wieder abnimmt, bevor die Verunreinigung schließlich aus dem Rohr austritt. Am rechten Rand treten sogar Konzentrationen auf, die oberhalb der Eingangskonzentration $u = 1$ liegen. ◊

66 Chromatographie

Die Chromatographie ist ein Verfahren zur Auftrennung eines Gemischs verschiedener Substanzen, vgl. Rhee, Aris und Amundson [89]. Zu diesem Zweck wird das Gemisch in einem Trägermedium gelöst, das mit konstanter Geschwindigkeit durch eine mit einem granularen Feststoff gefüllte chromatographische Säule strömt. Dabei werden einzelne Moleküle der gelösten Substanzen vorübergehend an der Oberfläche des Füllmaterials *adsorbiert*: Die Teilchen lagern sich an dem Füllmaterial an, wechseln also aus der sogenannten mobilen Phase der Trägersubstanz (und der gelösten Stoffe) in die stationäre Phase des Feststoffs und kehren zu einem späteren Zeitpunkt wieder in die mobile Phase zurück (Desorption).

Die Adsorptionsrate, also die Wahrscheinlichkeit, daß ein Teilchen an dem Feststoff adsorbiert, hängt von verschiedenen äußeren Umständen ab, wie zum Beispiel von der Temperatur und dem Druck. Wir beschränken uns im weiteren darauf, daß diese äußeren Parameter konstant bleiben. Unter dieser Voraussetzung hängen die einzelnen Adsorptionsraten nur noch von den beteiligten chemischen Substanzen ab. Verschiedene Stoffe weisen in der Regel unterschiedliche Adsorptionsraten auf und verweilen daher beim Durchfluß der Säule unterschiedlich lange in der stationären Phase. Dies hat zur Folge, daß sie zu unterschiedlichen Zeitpunkten wieder mit der Trägersubstanz am Ende der Säule austreten und auf diese Weise voneinander getrennt werden.

Im folgenden bezeichnen wir mit $u_i = u_i(x,t)$, $i = 1, \dots, m$, die Konzentrationen der m Substanzen des Gemischs in der Säule. (Die Konzentrationen werden in mol/l angegeben.) Der Wert von u_i setzt sich aus den beiden Anteilen v_i und w_i der entsprechenden Substanz in der mobilen beziehungsweise stationären Phase zusammen,

$$u_i = v_i + w_i \,. \tag{66.1}$$

Wir wollen für unser Modell davon ausgehen, daß die Anzahl der „Adsorptionsplätze“ allein proportional zu dem Volumen der Säule ist (und nicht von

der Form der Teilchen abhängt, die adsorbiert werden). Ist ρ die positive Proportionalitätskonstante, so können in einem infinitesimalen Kontrollvolumen dx der Säule maximal $\rho\,dx$ Teilchen adsorbiert werden, das heißt

$$\sum_{i=1}^{m} w_i \;\le\; \rho\,.$$

Wir betrachten nun zunächst den eigentlichen Adsorptionsvorgang isoliert vom restlichen Chromatographieprozeß und gehen dazu von konstanten Konzentrationen u_i, $i = 1, \dots, m$, in einem festen Kontrollvolumen dx der Säule aus. Einem Modell von *Langmuir* folgend setzen wir die Adsorptionsrate einer Substanz proportional zu der in Lösung befindlichen Konzentration dieses Stoffes und den freien Adsorptionsplätzen an, also proportional zu der Wahrscheinlichkeit, daß ein Teilchen einen freien Adsorptionsplatz findet. (Dies setzt die Annahme voraus, daß alle Adsorptionsplätze gleichberechtigt sind, und damit insbesondere, daß sich „benachbarte“ adsorbierte Teilchen nicht gegenseitig beeinflussen.) Die gleichzeitig stattfindende Desorption ist allein proportional zu der Anzahl adsorbierter Teilchen. Der Adsorptions-/Desorptionsprozeß im Punkt x kann somit durch das System von $2m$ gewöhnlichen Differentialgleichungen

$$\frac{dw_i}{d\tau} \;=\; a_i v_i(\rho - \sum_{j=1}^{m} w_j) \;-\; d_i w_i\,, \qquad \frac{dv_i}{d\tau} \;=\; -\,\frac{dw_i}{d\tau}\,, \tag{66.2}$$

bezüglich der Zeit τ modelliert werden; a_i und d_i sind die stoffspezifischen (positiven) Proportionalitätskonstanten. Diese Differentialgleichung ist ähnlich zu den Gleichungen, die wir in Kapitel XI betrachtet haben und die Lösung konvergiert für $\tau \to \infty$ gegen einen stationären Zustand, das sogenannte *Adsorptionsgleichgewicht*. In diesem Zustand sind w_i und v_i konstant, d. h. die linken Seiten von (66.2) sind jeweils Null. Hieraus ergibt sich das Gleichungssystem

$$w_i \;=\; \kappa_i v_i(\rho - \sum_{j=1}^{m} w_j)\,, \qquad i = 1, \dots, m\,, \tag{66.3}$$

mit $\kappa_i = a_i/d_i > 0$ für die Konzentrationen in der mobilen und der stationären Phase im Gleichgewichtszustand; durch Summation von (66.3) über $i = 1, \dots, m$ folgt die geforderte Ungleichung

$$\sum_{j=1}^{m} w_j \;=\; \rho\,\frac{\sum_{i=1}^{m} \kappa_i v_i}{1 + \sum_{i=1}^{m} \kappa_i v_i} \;<\; \rho\,.$$

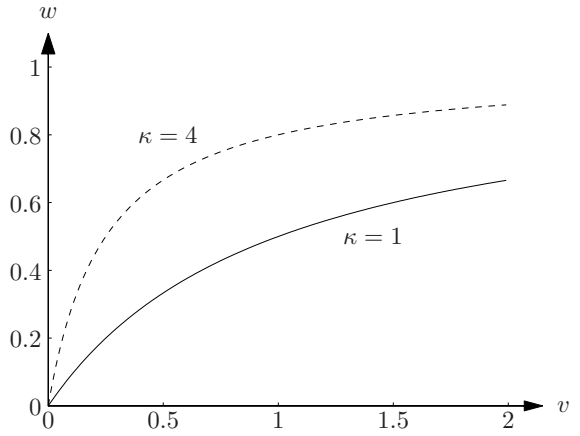


Abb. 66.1: Langmuir-Isothermen $w = w(v)$ für $m = 1$ und zwei verschiedene Werte von κ

Wieder in (66.3) eingesetzt, erhalten wir schließlich die explizite Darstellung

$$w_i \;=\; h_i(v_1, \ldots, v_m) \;=\; \rho\, \frac{\kappa_i v_i}{1 + \sum_{j=1}^{m} \kappa_j v_j}\,, \qquad i = 1, \ldots, m\,, \qquad (66.4)$$

für die Konzentrationen in der stationären Phase in Abhängigkeit von den Konzentrationen in der mobilen Phase. Diese Funktionen h_i werden *Langmuir-Adsorptionsisothermen* genannt, wobei der Begriff Isotherme daher rührt, daß die Parameter κ_i, $i = 1, \ldots, m$, nur bei festen äußeren Bedingungen (wie etwa der Temperatur) als konstant angesehen werden können. Abbildung 66.1 zeigt zwei Langmuir-Isothermen bei einer einzigen Substanz ($m = 1$): In beiden Fällen kann die stationäre Phase die gleiche Stoffkonzentration aufnehmen ($\rho = 1$ mol/l) und bei einer Konzentration von $v = 2$ mol/l in der mobilen Phase werden zwei Drittel beziehungsweise $8/9$ aller Adsorptionsplätze belegt.

Nun wenden wir uns dem eigentlichen Fließvorgang in der Säule während der Chromatographie zu. Dabei stellen wir uns der Einfachheit halber die chromatographische Säule als senkrecht ausgerichtetes Rohr der Länge L mit konstantem Querschnitt vor, das von der Trägersubstanz mit konstanter Geschwindigkeit mit Betrag a durchflossen wird. Dies erlaubt eine eindimensionale Modellierung der Säule mit einer Ortskoordinate $x \in \Omega = (0, L)$, bei der die Stelle $x = 0$ den Einspritzpunkt und $x = L$ den Ausflußpunkt der Säule angibt. In der physikalischen Chemie wird üblicherweise angenommen, daß der bereits skizzierte Adsorptionsprozeß in einer wesentlich kürzeren Zeitspanne abläuft als der Durchfluß der mobilen Phase, also in einer wesentlich feineren Zeitskala. Wir wollen daher unser Modell weiter vereinfachen und davon ausgehen, daß sich während der Chromatographie in jedem Zeitpunkt und an jedem Ort *unmittelbar* das Adsorptionsgleichgewicht einstellt.

Die Konzentrationen u_i der m eingeleiteten Substanzen sind Erhaltungsgrößen im Sinne von Abschnitt 65, für die wir nun das Erhaltungsgesetz herleiten wollen. Für u_i ergibt sich der Fluß $J_i = av_i$, da nur der Anteil in der mobilen Phase transportiert wird. Quellen und Senken für die m Stoffe im Innern der Säule gibt es keine, so daß wir für jede Substanz die differentielle Form

$$\frac{\partial}{\partial t}\, u_i \,+\, a\, \frac{\partial}{\partial x}\, v_i \,=\, 0\,, \qquad i = 1, \ldots, m\,, \tag{66.5}$$

der Erhaltungsgleichung bekommen, beziehungsweise

$$u_t + a v_x = 0\,,$$

wenn wir die Funktionen u_i und v_i in Vektoren $u = [u_1, \ldots, u_m]^T$ und $v = [v_1, \ldots, v_m]^T$ zusammenfassen.

Für numerische Simulationen ist die Differentialgleichung (66.5) nicht sehr gut geeignet, da die Abhängigkeit $v = v(u)$ aufgrund der Form der Isothermen im allgemeinen nicht explizit ausgedrückt werden kann. Lediglich für $m = 1$, also wenn nur ein einziger Stoff in die Säule eingeleitet wird, ergibt sich eine explizite Darstellung,

$$v(u) \,=\, \frac{u - \rho - 1/\kappa}{2} \,+\, \frac{1}{2}\Big(\, \frac{4u}{\kappa} + (u - \rho - 1/\kappa)^2 \Big)^{1/2}\,. \tag{66.6}$$

Auf der anderen Seite ist die Umkehrfunktion $u = u(v)$ mittels

$$u_i \,=\, v_i + w_i \,=\, v_i + h_i(v_1, \ldots, v_m)\,, \qquad i = 1, \ldots, m\,,$$

für jedes m explizit gegeben, nämlich

$$u_i \,=\, H_i(v_1, \ldots, v_m) \,=\, v_i + \rho\, \frac{\kappa_i v_i}{1 + \sum_{j=1}^{m} \kappa_j v_j}\,. \tag{66.7}$$

Aus (66.5) erhalten wir somit für unser Modell die alternative Differentialgleichung

$$v_x + \frac{1}{a}\, \frac{d}{dt} H(v) \,=\, 0\,, \qquad H = [H_1, \ldots, H_m]^T\,, \tag{66.8}$$

die bis auf eine Vertauschung der Rollen von t und x wieder die differentielle Erhaltungsform (65.2) aufweist.

Beispiel 66.1. Das linke Bild in Abbildung 66.2 zeigt die Lösung der Erhaltungsgleichung (66.5) für eine Substanz ($m = 1$) und die Isotherme aus Abbildung 66.1 mit $\kappa = 1$: Die Säule (also die Ortsvariable x) bildet die senkrechte Achse, der Einspritzpunkt ist oben und der Ausfluß unten bei $L = 10$;

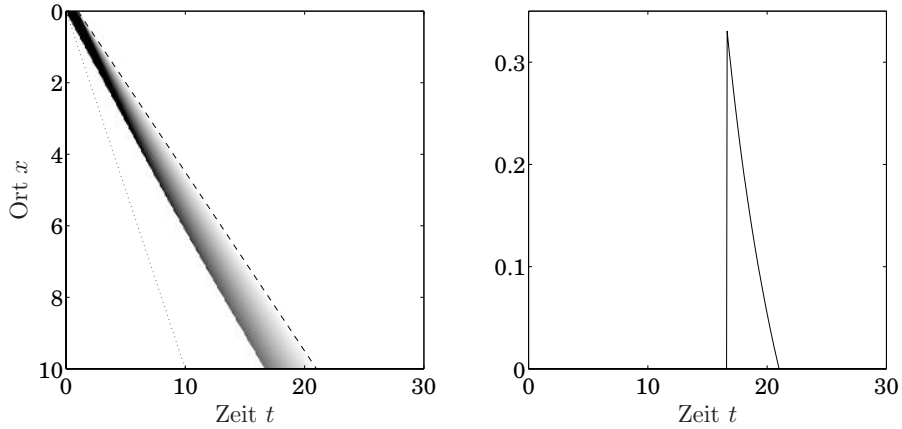


Abb. 66.2: Konzentration $u(x,t)$ in der Säule (li.) und Ausfluß $av(L,\cdot)$ (re.) über der Zeit

auf der horizontalen Achse ist von links nach rechts die Zeit aufgetragen; die Grautöne in der Abbildung veranschaulichen die Gesamtkonzentration $u(x,t)$ in der Säule.

Simuliert wird eine Chromatographie, bei der zwischen $t = 0$ und $t = 1$ eine Substanz mit konstanter Konzentration u_* eingeleitet wird. Diese Substanz fließt mit einer scharf ausgeprägten Front durch die Säule. Die eingezeichnete gebrochene Linie zeigt an, wann die Substanz einen Punkt der Säule vollständig durchflossen hat, die gepunktete Linie veranschaulicht die Fließgeschwindigkeit der mobilen Phase. Ohne Adsorption würde die Front auf dieser Linie liegen. Tatsächlich folgt die Front zunächst einer weniger steilen Geraden, bevor sie allmählich langsamer wird und „abzubröckeln“ beginnt; in diesem späteren Verlauf folgt die Front einer Parabelkurve, vgl. auch Aufgabe XVIII.4.

In der Chemie ist der Ausfluß $av(L,\cdot)$ am unteren Ende der Säule der wesentliche Aspekt bei der Chromatographie. Im rechten Bild von Abbildung 66.2 erkennt man den scharfen „Peak“, mit dem die Lösungsfront aus der Säule austritt. Im Anschluß an den Peak folgt das sogenannte „Tailing“, verursacht durch die länger in der Säule verbliebenen Moleküle. Das Integral unter der Kurve entspricht aufgrund der Massenerhaltung der eingeleiteten Stoffmenge und erlaubt deren quantitative Bestimmung. ◊

Beispiel 66.2. Abbildung 66.3 zeigt die Lösung von (66.5) für zwei Substanzen S_1 und S_2 mit $\kappa_1 = 1$ und $\kappa_2 = 4$, die zwischen $t = 0$ und $t = 1$ gleichzeitig in die Säule eingeleitet werden. Wegen $\kappa_1 < \kappa_2$ durchströmt S_1 die Säule schneller als S_2 und die beiden Stoffe beginnen sich zu trennen; dabei entstehen interessante Konzentrationsverteilungen (vgl. das Bild links). Nach einer

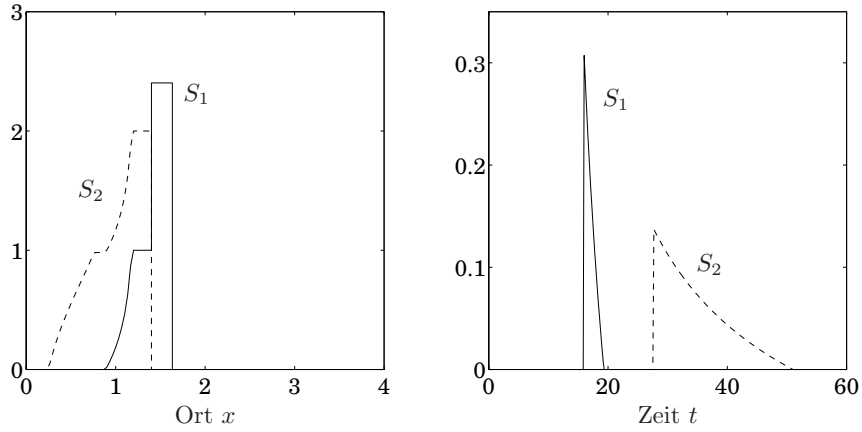


Abb. 66.3: Konzentrationen $u_i(x,t)$ zum Zeitpunkt $t = 2.24$ (links) und Ausflüsse $av_i(L,\cdot)$ (rechts) bei zwei Substanzen S_i, $i = 1, 2$

gewissen Strecke sind die Substanzen vollständig getrennt und fließen schließlich nacheinander aus der Säule heraus (Bild rechts). Aufgrund der stärkeren Adsorption von S_2 ist bei dem zweiten Stoff das Tailing stärker. Ein Vergleich mit Abbildung 66.2 zeigt, daß S_1 die Säule etwas schneller durchfließt als in Beispiel 66.1, da S_2 zu Beginn die Mehrzahl der Adsorptionsplätze blockiert und S_1 in diesem Bereich ungehinderter fließen kann. ◊

67 Strömungsmechanik

Wir beschreiben nun in relativ allgemeiner Form die Strömung eines Fluids in einem Gebiet $\Omega \subset \mathbb{R}^3$. Zentrale Eigenschaften einer solchen Strömung sind die Erhaltungsprinzipien für die Zustandsgrößen *Masse*, *Impuls* und *Energie* mit den jeweiligen Dichten ρ, m und E. In Anlehnung an die allgemeinen Überlegungen in Abschnitt 65 wollen wir im folgenden die entsprechenden Flußvektoren der drei Erhaltungsgrößen herleiten. Als ergänzende Literatur seien die Bücher von Chorin und Marsden [16] und von Temam und Miranville [101] empfohlen.

67.1 Massenerhaltung

Das Gesetz der Massenerhaltung ergibt sich wie in Beispiel 65.1 mit dem Fluß

$$J_\rho = \rho v\,,$$

wobei $v \in \mathbb{R}^3$ die Geschwindigkeit des Fluids bezeichnet. Ohne zusätzliche Quellen und Senken ist somit

$$\rho_t + \operatorname{div}(\rho v) \;=\; 0 \tag{67.1}$$

die (differentielle) Erhaltungsgleichung für die Masse. Diese Gleichung wird *Kontinuitätsgleichung* genannt.

67.2 Impulserhaltung

Für die Impulserhaltung müssen wir zunächst erläutern, welche inneren Kräfte innerhalb eines (zunächst beliebigen) Kontinuums wirken. Dazu bedienen wir uns wie in Abschnitt 3 des *Schnittprinzips* und schneiden ein Teilgebiet $G \subset \Omega$ frei. Uns interessiert die Oberflächenkraft, die das umgebende Kontinuum auf den Rand von G ausübt, beziehungsweise der zugehörige *Spannungsvektor*, das ist der Anteil der Kraft pro infinitesimalem Flächenelement. Im einfachsten Fall drückt die Umgebung immer senkrecht auf die Oberfläche, d. h. der Spannungsvektor ist in jedem Punkt $x \in \partial G$ dem Normalenvektor ν der Oberfläche entgegengesetzt. Ist das überall der Fall, so nennt man das Kontinuum *reibungsfrei* und spricht von einem *idealen Fluid*. Für ein ideales Fluid kann der Spannungsvektor also durch eine skalare Größe (seine Länge) repräsentiert werden, den *Druck* p. Die auf ∂G wirkende Kraft S ist dann das Integral

$$S \;=\; -\int_{\partial G} p\nu \, ds\,. \tag{67.2}$$

Unter dem *Impuls* eines Teilchens versteht man das Produkt aus seiner Masse und seiner Geschwindigkeit. Entsprechend ist die Impulsdichte $m(x,t)$ des Fluids das Produkt der Massendichte mit der Geschwindigkeit, $m = \rho v$. Man beachte, daß der Impuls ein Vektor mit drei Komponenten m_1, m_2 und m_3 ist. Für die folgende Herleitung des Impulserhaltungsgesetzes greifen wir noch einmal die Argumentation aus Abschnitt 65 auf und stellen für jede Impulskomponente m_i und die obige Teilmenge $G \subset \Omega$ die integrale Erhaltungsgleichung auf. Auf der einen Seite ergibt sich eine Impulsänderung analog zur Massenänderung aufgrund der Bewegung (der Geschwindigkeit) des Fluids: der zugehörige Anteil des Impulsflusses ist durch $m_i v$ gegeben. Auf der anderen Seite bewirkt auch die Oberflächenkraft (67.2) nach dem Newtonschen Beschleunigungsgesetz (22.1) eine Impulsänderung: die i-te Komponente des Impulses wird in diesem Fall um die entsprechende Kraftkomponente S_i mo-

difiziert. Somit gilt

$$\begin{aligned} -\frac{d}{dt}\int_G m_i\,dx \;&=\; \int_{\partial G} \nu\cdot(m_i v)\,ds \;-\; S_i \\ &=\; \int_{\partial G}(\nu\cdot v)m_i\,ds \;+\; \int_{\partial G} p\nu_i\,ds \end{aligned} \tag{67.3}$$

für $i = 1, 2, 3$, beziehungsweise in vektorieller Schreibweise

$$-\frac{d}{dt}\int_G m\,dx \;=\; \int_{\partial G}(\nu\cdot v)m\,ds \;+\; \int_{\partial G} p\nu\,ds \;=\; \int_{\partial G}(mv^* + pI)\nu\,ds\,.$$

Ein Vergleich mit (65.1) legt somit nahe, den *Impulsfluß* als Matrix

$$\boldsymbol{J}_m \;=\; mv^* + pI \;=\; \rho vv^* + pI$$

zu definieren, deren *Zeilen* die (transponierten) Flußvektoren $J_{m_i} = m_i v + pe_i$ der einzelnen Impulskomponenten sind.

Eine Anwendung des Gaußschen Satzes auf (67.3) ergibt

$$\int_G \Big(\frac{\partial m_i}{\partial t} + \operatorname{div}(m_i v) + \frac{\partial p}{\partial x_i}\Big)\,dx \;=\; 0$$

und damit bei hinreichender Glattheit die differentielle Erhaltungsform

$$\frac{\partial m_i}{\partial t} + \operatorname{div}(m_i v) + \frac{\partial p}{\partial x_i} \;=\; 0\,, \qquad i = 1, \ldots, d\,.$$

Die übliche Vektorschreibweise hierfür lautet

$$m_t + \mathbf{div}(mv^*) + \operatorname{grad} p \;=\; 0\,, \tag{67.4}$$

wobei die *Vektordivergenz* **div** durch eine Anwendung des Divergenzoperators auf die einzelnen Zeilen der Matrix im Argument definiert ist.

Zusammen mit der Kontinuitätsgleichung kann die Impulsgleichung ein wenig vereinfacht werden. Dazu ersetzen wir zunächst $m = \rho v$ und wenden die Produktregel an:

$$\begin{aligned} 0 \;&=\; v\rho_t \;+\; \rho v_t \;+\; m\operatorname{div} v \;+\; \frac{\partial m}{\partial x}\,v \;+\; \operatorname{grad} p \\ &=\; v\rho_t \;+\; \rho v_t \;+\; \rho v \operatorname{div} v \;+\; \rho\,\frac{\partial v}{\partial x}\,v \;+\; v(v\cdot\operatorname{grad}\rho) \;+\; \operatorname{grad} p\,. \end{aligned}$$

Die Zeitableitung der Dichte kann nun mittels (67.1) eliminiert werden: Dies führt auf die Gleichung

$$\rho\, v_t \;+\; \rho\,\frac{\partial v}{\partial x}\,v + \operatorname{grad} p \;=\; 0\,. \tag{67.5}$$

67.3 Energiebilanz

Neben der Massen- und Impulserhaltung ist eine ausgeglichene Energiebilanz die dritte zentrale Bedingung an eine Strömung. Der Vollständigkeit halber stellen wir daher noch die Erhaltungsgleichung für die Energie des Fluids auf, obwohl wir sie später nicht mehr benötigen. Für die Herleitung betrachten wir wieder eine beliebige Teilmenge $G \subset \Omega$, in der sich zunächst wegen des Transports der Materie ein Anteil Ev für den Energiefluß ergibt. Auf der anderen Seite leisten die Druckkräfte Arbeit, wenn das Fluid die Oberfläche ∂G durchdringt. Diese *Leistung* (Arbeit dW pro infinitesimalem Zeitintervall dt) ist definiert durch Kraft · Geschwindigkeit; aus (67.2) folgt daher

$$\frac{dW}{dt} \;=\; \int_{\partial G} (-p\nu) \cdot v \, ds \,.$$

Nach dem *Arbeitssatz* bewirkt die Arbeit dW eine Energiezunahme, so daß wir insgesamt die folgende Energiebilanz erhalten:

$$\begin{aligned} -\frac{d}{dt} \int_G E \, dx \;&=\; \int_{\partial G} \nu \cdot (Ev) \, ds \;+\; \int_{\partial G} p\nu \cdot v \, ds \;=\; \int_{\partial G} \nu \cdot \big((E+p)v\big) \, ds \\ &=\; \int_G \operatorname{div}\big((E+p)v\big) \, dx \,. \end{aligned}$$

Die differentielle Form der Energiebilanz lautet somit

$$E_t \;+\; \operatorname{div}\big((E+p)v\big) \;=\; 0 \,. \tag{67.6}$$

Die drei Erhaltungsgleichungen (67.1), (67.5) und (67.6) bilden die sogenannten *Euler-Gleichungen*, fünf skalare Differentialgleichungen für sechs skalare Größen: die Dichte ρ, den Druck p, die Energie E sowie die drei Komponenten der Geschwindigkeit v. Um diese Unterbestimmtheit zu beheben, müssen zusätzliche Annahmen an das Fluid gestellt werden, auf die wir nun beispielhaft eingehen wollen.

67.4 Potentialströmungen

Wir betrachten die von einem Punkt x_0 ausgehende Trajektorie $x(t)$, gegeben durch

$$\frac{d}{dt} \, x(t) \;=\; v(x(t), t) \,, \qquad x(0) = x_0 \,.$$

Das Fluid wird *inkompressibel* genannt, wenn die Dichte entlang jeder solchen Trajektorie konstant ist, das heißt wenn

$$0 \;=\; \frac{d}{dt}\,\rho(x(t),t) \;=\; v\cdot\operatorname{grad}\rho + \rho_t\,.$$

Ein Vergleich mit der Kontinuitätsgleichung (67.1) ergibt, daß dies zu

$$\operatorname{div} v = 0 \tag{67.7}$$

äquivalent ist. Es sei betont, daß allein aus der Inkompressibilität *nicht* gefolgert werden kann, daß die Dichte ρ im gesamten Fluid konstant ist.

Als Beispiel beschränken wir uns nun jedoch auf eine ebene stationäre Strömung in den zwei Variablen (x_1, x_2), in der $\rho = \rho_0$ konstant ist.[3] In diesem Fall vereinfacht sich (67.5) zu den beiden Gleichungen

$$\begin{aligned}
\rho_0 v_1 \frac{\partial v_1}{\partial x_1} \;+\; \rho_0 v_2 \frac{\partial v_1}{\partial x_2} \;+\; \frac{\partial p}{\partial x_1} \;&=\; 0\,,\\
\rho_0 v_1 \frac{\partial v_2}{\partial x_1} \;+\; \rho_0 v_2 \frac{\partial v_2}{\partial x_2} \;+\; \frac{\partial p}{\partial x_2} \;&=\; 0\,.
\end{aligned}$$

Durch Differentiation der ersten Gleichung nach x_2 und der zweiten Gleichung nach x_1 sowie anschließende Subtraktion folgt mit (67.7) nach einigen Umformungen, daß

$$v_1 \frac{\partial}{\partial x_1}\Big(\frac{\partial v_1}{\partial x_2} - \frac{\partial v_2}{\partial x_1}\Big) \;+\; v_2 \frac{\partial}{\partial x_2}\Big(\frac{\partial v_1}{\partial x_2} - \frac{\partial v_2}{\partial x_1}\Big) \;=\; 0\,.$$

Dies besagt, daß die *Wirbeldichte*

$$\operatorname{rot} v \;=\; \frac{\partial v_2}{\partial x_1} - \frac{\partial v_1}{\partial x_2} \tag{67.8}$$

ebenfalls entlang aller Trajektorien konstant ist:

$$\frac{d}{dt}\operatorname{rot} v(x(t),t) \;=\; v\cdot\operatorname{grad}(\operatorname{rot} v) \;+\; \operatorname{rot} v_t \;=\; 0\,.$$

Wir leiten nun spezielle ebene inkompressible und stationäre Strömungen her, in denen die Dichte konstant ist und die Wirbeldichte verschwindet, d. h. es ist

$$\operatorname{div} v = 0 \qquad \text{und} \qquad \operatorname{rot} v = 0 \tag{67.9}$$

[3] Bei einer stationären Strömung sind alle makroskopischen Größen von der Zeit unabhängig; eine stationäre Strömung ist jedoch nicht mit einer statischen (ruhenden) Strömung zu verwechseln, in der $v = 0$ ist.

in einem Gebiet $\Omega \subset \mathbb{R}^2$. Solche Strömungen heißen *Potentialströmungen*. Es sei daran erinnert, daß unter den genannten Voraussetzungen die Kontinuitätsgleichung erfüllt ist. Aus der Identität

$$\frac{1}{2}\,\operatorname{grad}|v|^2 \,-\, \frac{\partial v}{\partial x}\,v \;=\; \operatorname{rot} v \begin{bmatrix} v_2 \\ -v_1 \end{bmatrix} \;=\; 0$$

folgt zudem, daß auch die Impulsbilanz beziehungsweise die äquivalente Gleichung (67.5) durch den Druck

$$p \;=\; p_0 - \frac{\rho_0}{2}\,|v|^2 \tag{67.10}$$

mit konstantem p_0 (dem sogenannten *Staudruck*) eingehalten wird. Mit anderen Worten: ist $v : \Omega \to \mathbb{R}^2$ ein Geschwindigkeitsfeld, das die beiden Gleichungen (67.9) erfüllt, so lösen $\rho = \rho_0$, p aus (67.10) und dieses Geschwindigkeitsfeld v die Euler-Gleichungen (67.1) und (67.5); die *kinetische Energie*

$$E \;=\; \frac{\rho_0}{2}\,|v|^2$$

löst die verbleibende Gleichung (67.6).

Lösungen von (67.9) lassen sich mit Hilfe der Funktionentheorie charakterisieren: Die beiden Differentialgleichungen sind nämlich die *Cauchy-Riemann-Differentialgleichungen* (vgl. Ahlfors [2]) der Funktionen $(v_1, -v_2)$ und daher ist (67.9) äquivalent dazu, daß die komplexwertige Funktion

$$f(x_1 + \mathrm{i}x_2) \;=\; v_1 - \mathrm{i}v_2\,, \qquad (x_1, x_2) \in \Omega\,,$$

in dem entsprechenden Gebiet der komplexen Ebene holomorph ist.

Beispiel 67.1. Als Beispiel betrachten wir die Funktion

$$f(z) \;=\; 1 - z^{-2}\,, \qquad z = x_1 + \mathrm{i}x_2\,,$$

außerhalb des Einheitskreises. Sie führt auf das Geschwindigkeitsfeld

$$v_1(x) \;=\; \frac{x_1^4 + 2x_1^2x_2^2 + x_2^4 - x_1^2 + x_2^2}{(x_1^2 + x_2^2)^2}\,, \qquad v_2(x) \;=\; -\,\frac{2x_1x_2}{(x_1^2 + x_2^2)^2}\,,$$

dessen Normalkomponente $\nu \cdot v$ am Einheitskreisrand verschwindet:

$$\begin{aligned} \nu \cdot v \;&=\; x_1\big((x_1^2 + x_2^2)^2 - x_1^2 + x_2^2\big) - 2x_1x_2^2 \;=\; x_1 - x_1^3 - x_1x_2^2 \\ &=\; x_1(1 - x_1^2 - x_2^2) \;=\; 0\,. \end{aligned}$$

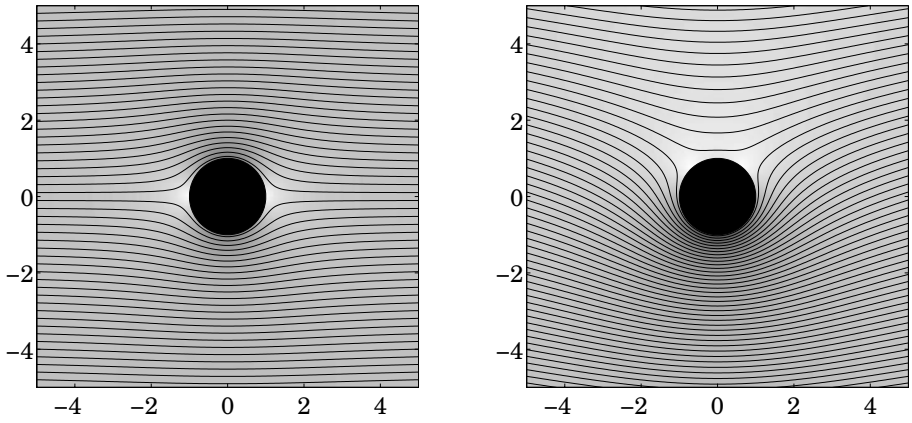


Abb. 67.1: Umströmungen eines Hindernisses im Einheitskreis

Abbildung 67.1 zeigt links einige Trajektorien dieser Strömung. Die Graustufen visualisieren den Betrag der Geschwindigkeit, der über (67.10) an den Druck und die kinetische Energie des Fluids gekoppelt ist. An den Punkten $(\pm 1, 0)$ verschwindet die Geschwindigkeit; hier liegen *Staupunkte* vor, an denen der Druck seinen Maximalwert (den Staudruck) p_0 annimmt.

Als zweites Beispiel betrachten wir die Funktion

$$\tilde{f}(z) \;=\; 1 + \frac{\gamma}{2\pi \mathrm{i}} z^{-1} - z^{-2} \;=\; f(z) + \frac{\gamma}{2\pi \mathrm{i}} z^{-1}\,, \qquad z = x_1 + \mathrm{i} x_2\,,$$

mit einem reellen Parameter γ, zu der das Geschwindigkeitsfeld $\tilde{v}$ mit

$$\tilde{v}_1 \;=\; v_1 - \frac{\gamma}{2\pi} \frac{x_2}{x_1^2 + x_2^2}\,, \qquad \tilde{v}_2 \;=\; v_2 + \frac{\gamma}{2\pi} \frac{x_1}{x_1^2 + x_2^2}\,,$$

gehört. Auch dieses Geschwindigkeitsfeld ist tangential zum Rand des Einheitskreises; die Trajektorien sind in Abbildung 67.1 (rechts) abgebildet.

Beide Lösungen beschreiben also die Umströmung desselben Hindernisses; das zweite Beispiel entspricht einer Überlagerung der ersten Lösung mit einem Wirbel mit Zirkulation γ. Während bei der ersten Lösung die Druckverteilung auf den Körper im Einheitskreis zwei Symmetrieachsen aufweist und die Oberflächenkräfte zueinander im Gleichgewicht stehen, bewirken bei der zweiten Lösung die Oberflächenkräfte auf das Hindernis eine nach oben resultierende *Auftriebskraft*, die von ihm aufgefangen werden muß. ◊

68 Schallwellen

Unter den Gasen sind *isentrope Gase* von vorrangiger Bedeutung, bei denen der Druck nur von der Dichte abhängt, $p = p(\rho)$, und mit zunehmender Dichte größer wird. Falls ein solcher funktionaler Zusammenhang bekannt ist, bilden die Erhaltungsgleichungen (67.1) und (67.5) für die Masse und den Impuls aus dem vorangegangenen Abschnitt ein in sich geschlossenes Gleichungssystem:

$$\rho_t + \operatorname{div}(\rho v) \,=\, 0\,, \qquad v_t + \frac{\partial v}{\partial x}\, v \,+\, p'(\rho)\,\frac{\operatorname{grad}\rho}{\rho} \,=\, 0\,, \tag{68.1}$$

mit $p' = dp/d\rho > 0$.

Eine Schallwelle in einem ruhenden isentropen Gas wird durch kleine Druckänderungen verursacht, vgl. Temam und Miranville [101]. Zur Untersuchung verwenden wir den Ansatz

$$v \,=\, \tilde{v}\,, \qquad \rho \,=\, \rho_0 + \tilde{\rho}\,,$$

mit konstantem ρ_0 und Störungen $\tilde{v}$ und $\tilde{\rho}$, die so klein sind, daß wir im weiteren quadratische Terme in $\tilde{v}$ und $\tilde{\rho}$ vernachlässigen können. In diesem Rahmen ist

$$p \,\approx\, p_0 + c_0^2\,\tilde{\rho}\,, \qquad p_0 = p(\rho_0)\,,\; c_0 = \sqrt{p'(\rho_0)}\,,$$

und in (68.1) eingesetzt ergeben sich (wiederum nach Vernachlässigung quadratischer Terme in $\tilde{\rho}$ und $\tilde{v}$) die Gleichungen

$$\tilde{\rho}_t + \rho_0 \operatorname{div}\tilde{v} \,=\, 0\,, \qquad \tilde{v}_t \,+\, \frac{c_0^2}{\rho_0}\,\operatorname{grad}\tilde{\rho} \,=\, 0\,. \tag{68.2}$$

Die Geschwindigkeit $\tilde{v}$ kann aus diesem System eliminiert werden, indem man die erste Gleichung nach der Zeit und die zweite Gleichung nach dem Ort differenziert:

$$\tilde{\rho}_{tt} + \rho_0 \operatorname{div}\tilde{v}_t \,=\, 0\,, \qquad \operatorname{div}\tilde{v}_t \,+\, \frac{c_0^2}{\rho_0}\,\operatorname{div}(\operatorname{grad}\tilde{\rho}) \,=\, 0\,.$$

Auf diese Weise ergibt sich die *Wellengleichung*

$$\tilde{\rho}_{tt} \,=\, c_0^2 \Delta\tilde{\rho}\,, \tag{68.3}$$

der Prototyp einer hyperbolischen Differentialgleichung zweiter Ordnung. Δ bezeichnet dabei wieder den Laplace-Operator.

Wir beschränken uns im weiteren auf die Schallausbreitung in einer Raumdimension. Durch Einführung des Vektors

$$u = \begin{bmatrix} u^+ \\ u^- \end{bmatrix} \in \mathbb{R}^2\,, \qquad u^\pm = \frac{1}{2}\,(c_0\tilde{\rho}/\rho_0 \pm \tilde{v})\,,$$

wird das System (68.2) in das äquivalente System erster Ordnung

$$u_t + Au_x = 0\,, \qquad A = \begin{bmatrix} c_0 & 0 \\ 0 & -c_0 \end{bmatrix},$$

überführt. Da A eine Diagonalmatrix ist, sind die beiden Gleichungen entkoppelte Transportgleichungen,

$$u_t^+ + c_0 u_x^+ = 0\,, \qquad u_t^- - c_0 u_x^- = 0\,,$$

die separat wie in Abschnitt 65 gelöst werden können. Somit ist

$$u^+(x,t) = h^+(x - c_0 t)\,, \qquad u^-(x,t) = h^-(x + c_0 t)\,,$$

für gewisse Funktionen $h^\pm : \mathbb{R} \to \mathbb{R}$, die aus Anfangswerten zur Zeit $t = 0$ bestimmt werden können. Nach Rücktransformation ergibt dies die Lösungen

$$\begin{aligned} \tilde{v}(x,t) &= h^+(x - c_0 t) - h^-(x + c_0 t)\,, \\ \tilde{\rho}(x,t) &= \frac{\rho_0}{c_0}\left(h^+(x - c_0 t) + h^-(x + c_0 t)\right), \end{aligned} \tag{68.4}$$

für die Ausbreitung des Schalls.

An (68.4) erkennt man, daß sich die Schallwelle in *beide* Richtungen mit Geschwindigkeit c_0 fortbewegt; $c_0 = \sqrt{p'(\rho_0)}$ wird daher *Schallgeschwindigkeit* genannt. Dabei handelt es sich um eine Konstante, die nur von dem betrachteten Gas und dem Referenzwert der Dichte abhängt.

Für konkrete Anfangsvorgaben

$$\tilde{\rho}(x,0) = r(x)\,, \quad \tilde{\rho}_t(x,0) = s(x)\,, \qquad x \in \mathbb{R}\,, \tag{68.5}$$

der Wellengleichung ergeben sich aus (68.4) mit $t = 0$ die Bedingungen

$$r = \frac{\rho_0}{c_0}\,(h^+ + h^-)\,, \qquad s = \rho_0 (h^- - h^+)'\,.$$

Durch Integration der zweiten Gleichung und anschließende Kombination mit der ersten Gleichungen erhalten wir

$$h^+ = \frac{c_0}{2\rho_0}\,r - \frac{1}{2\rho_0}\,S\,, \qquad h^- = \frac{c_0}{2\rho_0}\,r + \frac{1}{2\rho_0}\,S\,,$$

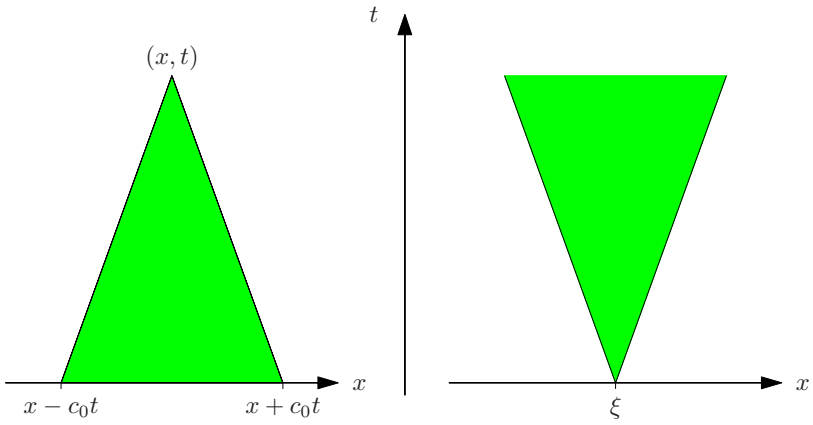


Abb. 68.1: Abhängigkeitsbereich (links) und Einflußbereich (rechts)

wobei S eine geeignet zu wählende Stammfunktion von s bezeichnet. Eingesetzt in (68.4) ergibt dies die *d'Alembertsche Lösungsformel*

$$\tilde{\rho}(x,t) \;=\; \frac{1}{2}\big(r(x-c_0t)+r(x+c_0t)\big) \;+\; \frac{1}{2c_0}\int_{x-c_0t}^{x+c_0t} s(\xi)\,d\xi\,.$$

Aus dieser Lösungsdarstellung können wir ablesen, daß die Dichte zu einem festen Zeitpunkt t an einem festen Ort x von allen Werten der Funktion s über dem Intervall $[x-c_0t, x+c_0t]$ und den Werten von r auf dem Rand dieses Intervalls abhängt; dieses Intervall bezeichnet man daher als *Abhängigkeitsbereich* von (x,t). Umgekehrt beeinflußt eine Anfangsstörung im Ort ξ alle Punkte (t,x) mit $\xi - c_0t \le x \le \xi + c_0t$; dies ist der sogenannte *Einflußbereich* des Punktes ξ, vgl. Abbildung 68.1.

Aufgaben

1. Bestimmen Sie die Lösung $u = u(x,t)$ der zeitabhängigen Transportgleichung

$$\begin{aligned} u_t + a(t)u_x &= 0\,, & x \in \mathbb{R}\,,\ t > 0\,,\\ u(x,0) &= u^\circ(x)\,, & x \in \mathbb{R}\,, \end{aligned}$$

wobei $u^\circ : \mathbb{R} \to \mathbb{R}$ differenzierbar und $a : \mathbb{R}^+ \to \mathbb{R}$ stetig sei.

2. Heißes Wasser mit Temperatur $u = u(x,t)$ fließt mit konstanter Geschwindigkeit $v > 0$ von links nach rechts in einem (eindimensional zu modellierenden) Rohr der Länge 1 durch einen Kühlraum mit konstanter Temperatur ω. Am Eintrittspunkt $x = 0$ hat das Wasser die Temperatur $u_0 > \omega$. Nach dem *Newtonschen Abkühlungsgesetz* kühlt sich das Fluid an jedem Punkt $x > 0$ im Zeitintervall dt um

$$-du = \gamma(u - \omega)\,dt$$

ab; $\gamma > 0$ ist eine geeignete Proportionalitätskonstante.
(a) Stellen Sie eine Differentialgleichung für die Temperaturverteilung u in dem Rohr auf. (Vernachlässigen Sie Diffusionseffekte.) Welche Anfangs- und Randbedingungen werden benötigt? Lösen Sie die Differentialgleichung.
(b) Falls das Rohr nicht in einem Kühlraum sondern durch einen dünnen, in die Gegenrichtung fließenden Wasserfilm mit Temperatur $w = w(x,t)$ gekühlt wird, so wird der Wasserfilm auf dem Weg nach links immer wärmer. Wie lautet das entsprechende Differentialgleichungssystem für diesen Fall. Ignorieren Sie die Temperaturaufnahme des Rohrs.
(c) Gehen Sie davon aus, daß der Wasserfilm ebenfalls mit konstanter Geschwindigkeit fließt und an der Stelle $x = L$ mit konstanter Temperatur ω eingespeist wird. Welche Temperaturverteilung ergibt sich in diesem Fall? Plotten Sie die Lösungen für verschiedene Parameter $\gamma > 0$.

3. Diese Aufgabe modelliert den Verkehr auf einer einspurigen (unendlich langen) Straße. Dabei wird die Straße durch die x-Achse repräsentiert, und $u = u(x,t)$ gibt die „Dichte“ der Autos auf der Straße zur Zeit t an (gegeben durch die Anzahl Autos pro km Straße). Es sei vorausgesetzt, daß u nach oben durch eine Konstante u_∞ beschränkt ist, bei der sich die Stoßstangen aller Autos berühren und der Verkehr zusammenbricht.
Nehmen Sie an, daß die Fahrer ihre Geschwindigkeit v der Autodichte anpassen (je weniger Autos auf der Straße sind, desto schneller kann man fahren), dabei aber die erlaubte Höchstgeschwindigkeit v_∞ respektieren.
(a) Überlegen Sie sich eine plausible Abhängigkeit $v = v(u)$.
(b) Stellen Sie für u eine Erhaltungsgleichung auf. Wie modellieren Sie Auf- und Abfahrten der Straße?

4. (a) Verallgemeinern Sie die Impulserhaltungsgleichung (67.5) auf die Strömung eines idealen Fluids in einem äußeren Kraftfeld $f(x,t)$ (pro Masse).
(b) Betrachten Sie eine stationäre Strömung $v = \operatorname{grad}\phi$ in $\Omega \subset \mathbb{R}^3$ mit einem glatten Geschwindigkeitspotential $\phi : \Omega \to \mathbb{R}$ und konstanter Dichte ρ_0. Zeigen Sie, daß unter dem Einfluß der (entgegengesetzt zur x_3-Achse wirkenden) Schwerkraft der Ausdruck

$$\frac{\rho_0}{2}\,|v|^2 + p + \rho_0 g x_3$$

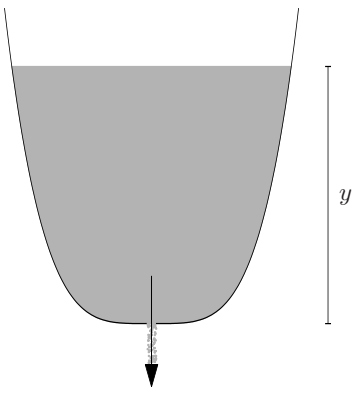
im ganzen Gebiet Ω konstant ist. Dies ist ein Spezialfall der *Bernoulli-Gleichung*.

5. Für ein ideales Gas in einem Volumen V ist der Druck p durch die Zustandsgleichung

$$pV = nkT$$

gegeben, wobei T die Temperatur, n die Anzahl der Gasmoleküle in dem Volumen und $k = 1.3806 \cdot 10^{-23}$ J/K die Boltzmann-Konstante bezeichnen. Interpretieren Sie dieses Gesetz und argumentieren Sie, warum der Druck nur von der Anzahl der Moleküle und nicht von ihrem Gewicht abhängt.

6. Aus einem rotationssymmetrischen Behälter strömt Wasser (wie in der Abbildung skizziert) aus einem Loch mit Querschnitt $a > 0$.



(a) Falls der Wasserpegel durch Wasserzufuhr auf einem konstanten Niveau y gehalten wird, so kann die Geschwindigkeit des Wassers an der Oberfläche vernachlässigt werden. Leiten Sie unter dieser vereinfachenden Annahme für die Ausflußgeschwindigkeit v die *Torricelli-Formel*

$$|v| = \sqrt{2gy}$$

her, in der g die Erdbeschleunigung bezeichnet.

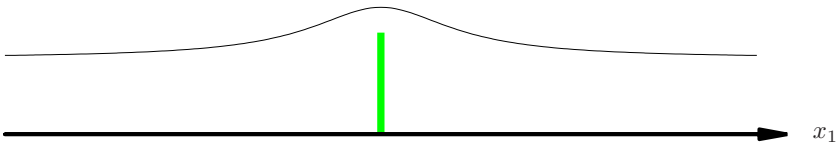
(b) Die Torricelli-Formel ist auch dann noch eine gute Approximation, falls kein Wasser in den Behälter nachströmt und der Wasserpegel $y = y(t)$ absinkt. Bestimmen Sie im Rahmen dieses Modells eine Behälterform, für die der Wasserpegel mit konstanter Geschwindigkeit sinkt.

Hinweis zu (a): Aufgabe 4.

7. Die Stromlinien einer Potentialströmung (wie in Abbildung 67.1) lassen sich in MATLAB einfach als Höhenlinien einer geeigneten Funktion w darstellen (mit dem Kommando `contour`). Bestimmen Sie entsprechende Funktionen w bzw. $\tilde{w}$ für die beiden Strömungen aus Beispiel 67.1.

8. Berechnen Sie in Beispiel 67.1 die Auftriebskraft, die bei den beiden Strömungen auf das Hindernis wirkt.

9. Bestimmen und visualisieren Sie diejenige Potentialströmung $v = v(x)$ über eine Mauer (vgl. Abbildung), für die $v(x) \to e_1$ für $|x| \to \infty$.



Hinweis: Nutzen Sie aus, daß die *Joukowski-Transformation* $\Psi(w) = (w - 1/w)/2$ das Äußere des komplexen Einheitskreises holomorph und bijektiv auf das Gebiet $\mathbb{C} \setminus [-\mathrm{i}, \mathrm{i}]$ abbildet.

10. (a) Sei $x_0 \in \mathbb{R}$ und $0 < t \leq t_0$. Rechnen Sie nach, daß für jede Lösung u der Wellengleichung (68.3) die Beziehung

$$u(x_0 - c_0 t, t_0) + u(x_0 + c_0 t, t_0) \;=\; u(x_0, t_0 + t) + u(x_0, t_0 - t)$$

gilt.
(b) Betrachten Sie das Anfangswertproblem $u(x,0) = r(x)$, $u_t(x,0) = s(x)$ für die Wellengleichung (68.3), wobei r und s beide beschränkten Träger haben sollen. Zeigen Sie, daß unter dieser Voraussetzung

$$\int_{\mathbb{R}} u_t^2(x,t)\,dx + c_0^2 \int_{\mathbb{R}} u_x^2(x,t)\,dx \;=\; \int_{\mathbb{R}} s^2(x)\,dx + c_0^2 \int_{\mathbb{R}} r_x^2(x)\,dx \quad \text{für alle } t \geq 0.$$

XIII Diffusionsprozesse

Zum Abschluß dieses Modellierungsteils wenden wir uns noch Diffusionsprozessen zu, die in einer Vielzahl unterschiedlicher Anwendungen auftreten. Diffusionsprozesse haben ausgleichenden Charakter: In vielen Fällen gibt es stationäre (von der Zeit unabhängige) Lösungen, die als Gleichgewichtszustände interpretiert werden können. Eine davon abweichende Anfangsvorgabe zur Zeit $t = 0$ führt zu einer zeitabhängigen Lösung (einer parabolischen partiellen Differentialgleichung), die im Grenzübergang $t \to \infty$ wieder gegen diesen Gleichgewichtszustand konvergiert.

69 Brownsche Bewegung und Diffusion

Zunächst soll der Begriff der Diffusion veranschaulicht werden. Man vergleiche hierzu auch die Darstellung in dem bereits zitierten Buch von Lin und Segel [69, Chapter 3]. Dazu betrachten wir ein Teilchen an der Position $x_0 = 0$ zur Zeit $t = 0$. Wir nehmen an, das Teilchen bewegt sich in einer Zeitspanne $\tau > 0$ aufgrund von Kollisionen oder Stößen mit gleicher Wahrscheinlichkeit nach links oder rechts um eine Weglänge $\triangle x = \sqrt{2\sigma\tau}$, wobei $\sigma > 0$ ein fest gewählter Parameter sei. Befindet sich das Teilchen nach n Zeitschritten am Ort x_n, so ergibt sich zur Zeit $t_{n+1} = (n+1)\tau$ die Position

$$x_{n+1} = x_n + \triangle x_n \,,$$

wobei die einzelnen Weglängen $\triangle x_n$, $n \in \mathbb{N}_0$, voneinander unabhängige zweiwertige Zufallsvariablen sind mit Eintrittswahrscheinlichkeiten

$$\mathcal{P}(\triangle x_n = \sqrt{2\sigma\tau}) \;=\; \mathcal{P}(\triangle x_n = -\sqrt{2\sigma\tau}) \;=\; 1/2\,.$$

Die so spezifizierte Bewegung des Teilchens wird stochastische Irrfahrt (*random walk*) genannt; eine Realisierung einer stochastischen Irrfahrt ist in Abbildung 69.1 dargestellt.

Wir fixieren nun einen Zeitpunkt $t > 0$, setzen $\tau = t/n$ und betrachten die Wahrscheinlichkeit $\mathcal{P}(x_n \le x)$ dafür, daß sich das Teilchen zum Zeitpunkt

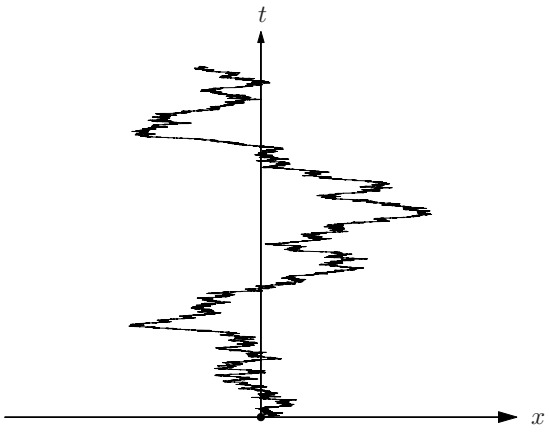


Abb. 69.1: Stochastische Irrfahrt: Approximation einer Brownschen Bewegung

$t = t_n$ an einem Ort $x_n \le x$ aufhält. Aus dem *zentralen Grenzwertsatz* der Stochastik folgt dann, daß diese Wahrscheinlichkeit für große n (also für kleine Zeitschritte) näherungsweise durch

$$\mathcal{P}(x_n \le x) \;\approx\; \frac{1}{\sqrt{2\sigma t_n}} \int_{-\infty}^{x} \varphi\big(\frac{\xi}{\sqrt{2\sigma t_n}}\big)\, d\xi \;=\; \int_{-\infty}^{x} \frac{1}{\sqrt{4\pi\sigma t_n}}\, \exp\big(-\frac{\xi^2}{4\sigma t_n}\big)\, d\xi$$

gegeben ist; φ ist dabei die Dichte der Standard-Normalverteilung. Die vorgenommene Normierung der Weglänge $\triangle x$ proportional zu $\sqrt{\tau}$ führt also im Grenzübergang $\tau \to 0$ zu einem zeitkontinuierlichen stochastischen Prozeß $x = x(t)$, der normalverteilt ist mit

$$\mathcal{E}\big(x(t)\big) = 0 \qquad \text{und} \qquad \mathcal{V}\big(x(t)\big) = 2\sigma t\,.$$

Dieser Prozeß wird *Brownsche* (*Molekular-*)*Bewegung* genannt.

Sind mehrere Teilchen (etwa u_0 Stück) zum Zeitpunkt $t = 0$ im Punkt $x = 0$ konzentriert und führen alle unabhängig voneinander eine Irrfahrt durch, so ergeben sich unterschiedliche Realisierungen derselben Zufallsvariablen und in der Regel werden sich nicht alle Teilchen nach n Teilschritten am gleichen Ort befinden. Statt dessen werden sich manche mehr nach rechts, andere mehr nach links bewegen: Die Teilchen *diffundieren* auseinander; der Parameter σ steuert die Stärke der Diffusion. Für kleine Zeitschritte approximiert die Funktion

$$u(x,t) \;=\; \frac{u_0}{\sqrt{4\sigma\pi t}}\, \exp\big(-\frac{x^2}{4\sigma t}\big)\,, \qquad x \in \mathbb{R}\,,$$

die Dichte der Teilchen über $\mathbb{R}$ zum Zeitpunkt t. Da das Integral über die Gaußsche Glockenkurve den Wert Eins ergibt, bleibt die Gesamtzahl aller Teilchen zu jedem Zeitpunkt konstant gleich u_0 (Massenerhaltung).

In der Regel beginnt die Diffusion nicht mit einer endlichen Zahl u_0 von Teilchen im Punkt $x = 0$, sondern mit einer Teilchendichte $u^\circ(x)$ über der gesamten reellen Achse ($x \in \mathbb{R}$), wobei sich alle Teilchen voneinander unabhängig bewegen. Durch Überlagerung der einzelnen Bewegungen ergibt dies die Dichte

$$u(x,t) \;=\; \frac{1}{\sqrt{4\sigma\pi t}} \int_{-\infty}^{\infty} \exp\Big(-\frac{(x-\xi)^2}{4\sigma t}\Big)\, u^\circ(\xi)\, d\xi\,, \quad x \in \mathbb{R}\,, \tag{69.1}$$

der Teilchen zum Zeitpunkt $t > 0$.

Das Integral (69.1) ist für jede beschränkte Anfangsvorgabe u° absolut konvergent. Das gleiche gilt für die Integrale, bei denen anstelle des Exponentialausdrucks in (69.1) dessen partielle Ableitungen nach x oder t stehen. Die partiellen Ableitungen u_t und u_{xx} von u existieren daher und können durch Vertauschung von Integration und Differentiation berechnet werden. Auf diese Weise sieht man, daß u die partielle Differentialgleichung

$$u_t \;=\; \sigma u_{xx}\,, \qquad x \in \mathbb{R}\,, \quad t > 0\,, \tag{69.2}$$

erfüllt. Etwas technischer ist der Nachweis, daß bei stetiger Anfangsfunktion u° auch der Grenzwert von $u(x,t)$ für $t \to 0$ existiert. Wie man erwarten wird, ist u aus (69.1) in diesem Fall durch $u(x,0) = u^\circ(x)$ stetig fortsetzbar (siehe Cannon [13, Lemma 3.4.1]).

Die Differentialgleichung (69.2) ist eine sogenannte *Diffusionsgleichung*. Sie kann in das allgemeine Erhaltungskonzept aus Kapitel XII eingeordnet werden mit der Erhaltungsgröße u und ihrem Fluß

$$J \;=\; -\sigma u_x\,, \tag{69.3}$$

vgl. (65.2). Der Fluß ist also proportional zu dem Gradienten der Dichte, er verschwindet, wenn die Dichte konstant ist. Dies kann man sich auch unmittelbar anhand der Definition der stochastischen Irrfahrt klarmachen. Befinden sich zwei Teilchen zu einem Zeitpunkt t an zwei benachbarten Orten x und $x + \triangle x$, so ist die Wahrscheinlichkeit, daß das eine Teilchen im nächsten Zeitschritt von x nach $x + \triangle x$ gestoßen wird genauso groß wie die Wahrscheinlichkeit für die umgekehrte Bewegung, bei der das andere Teilchen von $x + \triangle x$ nach x kommt. Befinden sich hingegen nur wenige Teilchen in x (etwa u Stück) und viele $(u + \triangle u)$ in $x + \triangle x$, so werden im Mittel mehr Teilchen von $x + \triangle x$ nach x wandern als in die umgekehrte Richtung; der Zuwachs im Punkt x ist proportional zu der Differenz $(u + \triangle u) - u = \triangle u$.

Außer den konstanten Funktionen sind auch die bezüglich des Ortes linearen Funktionen

$$u(x,t) \;=\; a + bx\,, \qquad a, b \in \mathbb{R}\,,$$

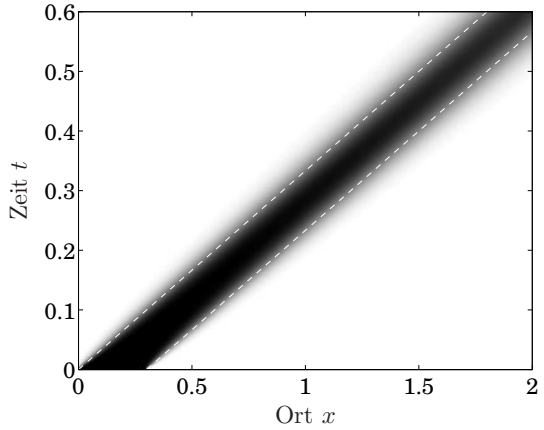


Abb. 69.2: Transport einer Verunreinigung mit Diffusion

mit $b \neq 0$ stationäre Lösungen der Differentialgleichung (69.2). In diesem Fall ist der Fluß J nicht Null sondern konstant.

Beispiel 69.1. In Beispiel 65.1 haben wir die Ausbreitung einer Verunreinigung in einem fließenden Gewässer betrachtet und dabei die Diffusion vernachlässigt. Wenn die Diffusion nicht vernachlässigbar ist, ergibt sich der Gesamtfluß der einzelnen Schmutzpartikel aus einer Überlagerung der Transportbewegung und der Diffusionsbewegung,

$$J(x,t) \;=\; au - \varepsilon u_x \,,$$

wobei a die Fließgeschwindigkeit des Gewässers und $\varepsilon > 0$ der Diffusionsparameter ist. Die zugehörige Differentialgleichung lautet somit

$$u_t \;=\; -au_x + \varepsilon u_{xx} \,, \qquad x \in \mathbb{R} \,, \quad t > 0 \,. \tag{69.4}$$

Abbildung 69.2 zeigt die Lösung in der (x,t)-Ebene für die Parameter $a = 3$ und $\varepsilon = 0.01$ bei der gleichen Anfangsvorgabe wie in Beispiel 65.1,

$$u^\circ(x) \;=\; \begin{cases} 0 \,, & x < 0 \,, \\ 1 \,, & 0 \le x \le 0.3 \,, \\ 0 \,, & x > 0.3 \,. \end{cases}$$

Durch gebrochene Linien ist dabei der Transport der Verschmutzung eingezeichnet, wie er sich ohne Diffusion ergeben würde. Man erkennt, daß ein gewisser Anteil der Schmutzpartikel aus diesen Grenzen herausdiffundiert. $\lozenge$

Beispiel 69.2. Als zweites Beispiel betrachten wir die Temperatur $u = u(x,t)$ in einem Gebiet $\Omega \subset \mathbb{R}^d$ zum Zeitpunkt $t \geq 0$. Bei inhomogener Temperatur ergibt sich ein Energiefluß $J = J(x,t)$ mit dem Bestreben, die Temperatur innerhalb des Gebiets auszugleichen, vgl. etwa Lin und Segel [69, Chapter 4]. Nach dem *Fourierschen Gesetz* des Wärmeausgleichs ist dieser Energiefluß proportional zu dem Gradienten $\operatorname{grad} u$ der Temperatur,

$$J \;=\; -\sigma \operatorname{grad} u\,, \tag{69.5}$$

die positive Proportionalitätskonstante $\sigma = \sigma(x)$ beschreibt die *Wärmeleitfähigkeit* des jeweiligen Materials.

In diesem Beispiel ist die Energie die Erhaltungsgröße. Für das Erhaltungsgesetz muß noch die in einem Teilgebiet $G \subset \Omega$ gespeicherte Energie bestimmt werden, die über die *Wärmekapazität* $\gamma = \gamma(x)$ an die Temperatur gekoppelt ist. Diese Proportionalitätskonstante gibt an, wieviel Energie eine bestimmte Masse des Materials aufnehmen muß, um sich um ein Kelvin zu erwärmen. Um auf die Energiedichte zu kommen, muß zusätzlich die Dichte $\rho = \rho(x)$ des Körpers berücksichtigt werden. Die integrale Erhaltungsform aus Abschnitt 65 für die Energie hat dann die Gestalt

$$\frac{d}{dt}\int_G \gamma\rho u\,dx \;=\; -\int_{\partial G} \nu \cdot J\,ds \;=\; \int_{\partial G} \sigma\nu \cdot \operatorname{grad} u\,ds\,,$$

die zugehörige Differentialgleichung lautet

$$\gamma\rho\, u_t \;=\; \operatorname{div}(\sigma \operatorname{grad} u)\,, \qquad x \in \Omega\,, \quad t > 0\,. \tag{69.6}$$

Bei einem homogenen Körper sind ρ, γ und σ konstant und die Gleichung (69.6) vereinfacht sich zu

$$u_t \;=\; \kappa\,\Delta u\,, \qquad \kappa = \frac{\sigma}{\gamma\rho}\,,$$

die sogenannte *Wärmeleitungsgleichung*, die im Eindimensionalen bis auf den Faktor κ mit der Gleichung (69.2) übereinstimmt. Für die weitere Diskussion sehen wir von physikalischen Einheiten ab und setzen $\kappa = 1$. Außerdem beschränken wir uns auf das eindimensionale Intervall $\Omega = (0,\pi)$, in dem u etwa die Temperaturverteilung eines homogenen Stabs beschreibt. Da dieses Intervall endlich ist, müssen neben der Anfangstemperatur $u(x,0) = u^\circ(x)$ des Stabs zum Zeitpunkt $t = 0$ noch Randbedingungen auf $\partial\Omega$ vorgegeben werden. Man könnte sich beispielsweise vorstellen, daß der Stab am linken und am rechten Ende bei konstanter Temperatur $u = 0$ gehalten wird. Dann haben wir ein für Diffusionsgleichungen typisches *Anfangsrandwertproblem*

$$u_t = u_{xx}\,, \qquad u(x,0) = u^\circ(x)\,, \quad u(0,t) = u(\pi,t) = 0\,, \tag{69.7}$$

für $0 < x < \pi$ und $t \geq 0$. Damit dieses Problem eine stetige Lösung besitzen kann, muß $u^\circ(0) = u^\circ(\pi) = 0$ sein. Die örtlichen Randbedingungen an die Werte der Funktion u werden *Dirichlet-Randbedingungen* genannt.

Das Problem (69.7) kann zu einem gewissen Grad „explizit“ gelöst werden. Dazu beachten wir, daß die Funktionen

$$v_n(x) = \sin nx\,, \qquad n \in \mathbb{N}\,,$$

Eigenfunktionen des Differentialoperators $L[v] = v_{xx}$ mit $v(0) = v(\pi) = 0$ sind: es gilt nämlich

$$L[v_n] \;=\; \frac{d^2}{dx^2} \sin nx \;=\; \lambda_n v_n\,, \qquad \lambda_n = -n^2\,. \tag{69.8}$$

Entwickeln wir die Anfangsvorgabe u° in eine (formale) Sinusreihe

$$u^\circ(x) \;=\; \sum_{n=1}^{\infty} b_n \sin nx\,,$$

(nach Abschnitt 55.1 ist dies für jede Funktion $u^\circ \in \mathcal{L}^2(0,\pi)$ möglich und für $u \in H^1(0,\pi)$ mit $u(0) = u(\pi) = 0$ ist diese Reihe in $[0,\pi]$ gleichmäßig konvergent), so führt der *Separationsansatz*

$$u(x,t) \;=\; \sum_{n=1}^{\infty} \eta_n(t) \sin nx$$

für die Lösung von (69.7) zum Erfolg. Durch partielle Differentiation nach t erhalten wir nämlich formal

$$u_t(x,t) \;=\; \sum_{n=1}^{\infty} \eta_n'(t) \sin nx\,,$$

während eine zweimalige Differentiation nach x entsprechend

$$u_{xx}(x,t) \;=\; -\sum_{n=1}^{\infty} n^2 \eta_n(t) \sin nx$$

ergibt. Aus (69.7) lassen sich daher die unbekannten Funktionen η_n durch einen Koeffizientenvergleich bestimmen, und zwar als Lösungen des entkoppelten Systems gewöhnlicher Differentialgleichungen

$$\eta_n' = -n^2\,\eta_n\,, \quad \eta_n(0) = b_n\,, \qquad n \in \mathbb{N}\,,$$

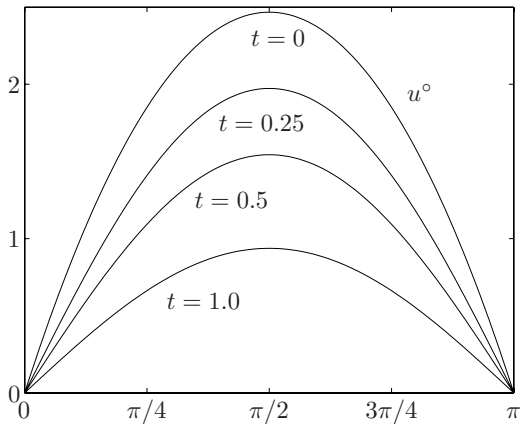


Abb. 69.3: Wärmeverteilung in einem Stab zu verschiedenen Zeitpunkten

also

$$\eta_n(t) \;=\; b_n e^{-n^2 t}\,, \qquad n \in \mathbb{N}\,. \tag{69.9}$$

Es sei noch einmal betont, daß der Schlüssel zu dieser erfolgreichen Separation die Tatsache ist, daß die Sinusfunktionen gerade die Eigenfunktionen des Differentialoperators L mit den korrekten Randbedingungen sind.

Zur Illustration betrachten wir die spezielle Anfangsvorgabe

$$u^\circ(x) = x(\pi - x)\,, \qquad 0 \le x \le \pi\,.$$

In diesem Fall lautet die (gleichmäßig konvergente) Sinusreihe der Anfangsvorgabe

$$u^\circ(x) \;=\; \sum_{n=1}^{\infty} \frac{8}{\pi(2n-1)^3}\,\sin(2n-1)x\,, \qquad 0 \le x \le \pi\,,$$

und für die Koeffizientenfunktionen η_n ergeben sich die Anfangswerte

$$b_{2n-1} = \frac{8}{\pi(2n-1)^3}\,, \quad b_{2n} = 0\,, \qquad n \in \mathbb{N}\,.$$

Aus (69.9) erhält man folglich die analytische Darstellung

$$u(x,t) \;=\; \sum_{n=1}^{\infty} \frac{8e^{-(2n-1)^2 t}}{\pi(2n-1)^3}\,\sin(2n-1)x \tag{69.10}$$

der Lösung von (69.7). Abbildung 69.3 zeigt die Temperaturkurven zu einigen ausgewählten Zeitpunkten. Wie man anhand von (69.10) erkennen kann, strebt die Temperatur überall im Stab für $t \to \infty$ gegen die konstante Temperatur $u = 0$. ◊

70 Diffusion im Kraftfeld

Wir wollen nun das Teilchenmodell des vorangegangenen Abschnitts erweitern und davon ausgehen, daß in dem betrachteten Gebiet zusätzlich äußere Kräfte $F = F(x,t)$ wirken. Hierbei folgen wir Feynman [27, Kapitel 43] und betrachten zunächst wieder ein Teilchen, das aufgrund von Kollisionen häufig (und jeweils zufällig) die Richtung wechselt. Nach jeder Kollision startet das Teilchen mit einer Startgeschwindigkeit v_0, die durch eine Zufallsvariable mit Erwartungswert Null gegeben ist. Bezeichnen wir den Zeitpunkt der letzten Kollision mit t_0 und ihren Ort mit x_0, so beschreibt das Teilchen bis zur nächsten Kollision eine Bahn $x(t)$, die der Differentialgleichung

$$x''(t) \,=\, \frac{1}{m}\,F(x,t)\,, \qquad x(t_0) = x_0\,, \quad x'(t_0) = v_0\,,$$

genügt, wobei m die Masse des Teilchens ist (vgl. Abschnitt 63).

Für eine makroskopische Beschreibung dieses Modells werden die physikalischen Größen aller Partikel eines Kontrollvolumens gemittelt (vgl. die Fußnote auf Seite 495). Ist ρ die Dichte einer solchen Größe, so ist nach Abschnitt 65 der zugehörige Fluß durch $J = \rho v$ gegeben, wobei v die mittlere Geschwindigkeit der einzelnen Teilchen in diesem Kontrollvolumen bezeichnet.

Wir setzen nun voraus, daß die Kollisionen der einzelnen Teilchen sehr häufig stattfinden, d. h. auf einer wesentlich feineren Zeitskala als der des makroskopischen Modells, in der die äußeren Kräfte variieren. Unter dieser Voraussetzung kann die auf ein Teilchen zwischen zwei Kollisionen wirkende Kraft als konstant angenommen werden und für das obige Teilchen ergibt sich dann die Geschwindigkeit

$$x'(t) \,\approx\, v_0 \,+\, \frac{t-t_0}{m}\,F(x_0,t_0)$$

für die Zeit zwischen dem letzten und dem nächsten Kollisionszeitpunkt. Werden diese Geschwindigkeiten über alle Teilchen eines Kontrollvolumens gemittelt, so führt dies auf die mittlere Geschwindigkeit

$$v \,=\, \frac{\tau}{2m}F$$

im makroskopischen Modell, wobei τ die mittlere Länge des Zeitintervalls zwischen zwei aufeinanderfolgenden Kollisionen eines Teilchens ist.

Aus diesen Überlegungen heraus ergibt sich also der Fluß

$$J \,=\, \rho v \,=\, \rho\frac{\tau}{2m}F \tag{70.1}$$

und die Erhaltungsgleichung

$$\rho_t + \operatorname{div}(\rho \frac{\tau}{2m} F) = 0\,. \tag{70.2}$$

Beispiel 70.1 (Elektrostatik). Die obigen Überlegungen sind die Grundlage des sogenannten *Drude-Modells* (vgl. [7, Kapitel 1]) eines (zeitlich) konstanten *elektrischen Stroms* ι in einem metallischen Körper $\Omega \subset \mathbb{R}^d$. In diesem Fall sind die Teilchen die freien Elektronen des Metalls und ρ ist die Ladungsdichte. Bei einer Potentialverteilung u im Innern des Körpers und dem zugehörigen elektrischen Feld

$$E = -\operatorname{grad} u \tag{70.3}$$

wirkt auf ein Teilchen die Kraft

$$F = eE\,,$$

wobei $e = 1.602 \cdot 10^{-19}\,\mathrm{C}$ (Coulomb = Ampère · Sekunde) die Ladung eines Elektrons (Elementarladung) ist.

Der elektrische Strom ι (Ladungsfluß) in dem Körper ergibt sich aus (70.1):

$$\iota = \frac{\rho e \tau}{2m} E = -\sigma \operatorname{grad} u \qquad \text{mit} \qquad \sigma = \frac{\rho e \tau}{2m}\,. \tag{70.4}$$

Dies ist nichts anderes als das *Ohmsche Gesetz*: σ ist die elektrische Leitfähigkeit, der Kehrwert des Ohmschen Widerstands, vgl. Abschnitt 64. Während Ladung e und Masse m eines Elektrons konstant sind, hängt das Produkt aus der Ladungsdichte ρ und der mittleren Zeit τ zwischen zwei aufeinanderfolgenden Kollisionen eines Elektrons von dem Material des Körpers ab. Für inhomogene Körper ist also die elektrische Leitfähigkeit σ eine Funktion des Orts.

Aus (70.2) und (70.4) erhalten wir für das Potential die (elliptische) Differentialgleichung

$$\operatorname{div}(\sigma \operatorname{grad} u) = \rho_t = 0 \qquad \text{in } \Omega\,, \tag{70.5}$$

da ρ stationär sein soll. Bei einem homogenen Körper entspricht dies der Laplace-Gleichung $\Delta u = 0$. Um aus der Differentialgleichung (70.5) die Potentialverteilung zu berechnen, benötigen wir (wie in Beispiel 69.2) Randbedingungen auf $\Gamma = \partial\Omega$. Man spricht dann von einem *Randwertproblem*.

Stellen wir uns zum Beispiel einen metallischen Gegenstand vor, auf dessen Oberfläche Γ zwei Elektroden befestigt sind, durch die Strom fließt. Der Rand des Körpers sei ansonsten isoliert, das heißt außer an den Elektroden liegt kein

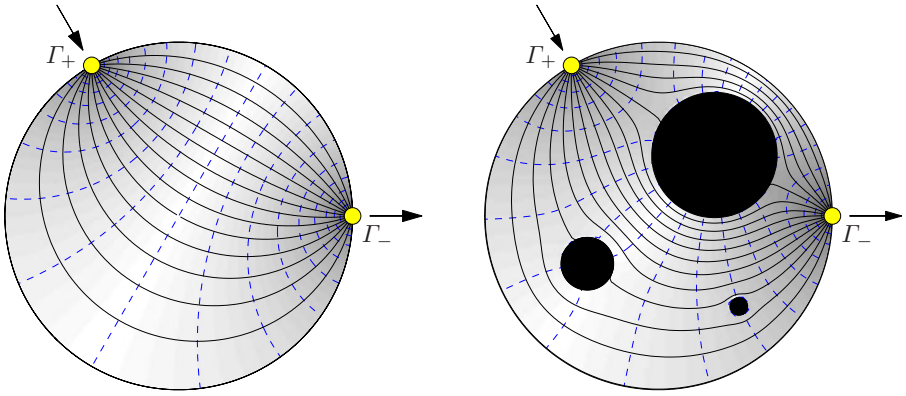


Abb. 70.1: Elektrische Feldlinien und Äquipotentiallinien

Stromfluß durch den Körperrand vor. Wir bezeichnen die Fläche der Elektroden mit $\Gamma_+, \Gamma_- \subset \Gamma$, ihren Flächeninhalt jeweils mit $|\Gamma_\pm|$ und modellieren die Randvorgaben durch

$$\nu \cdot \iota \;=\; \begin{cases} -1/|\Gamma_+|\,, & x \in \Gamma_+\,, \\ \;\;1/|\Gamma_-|\,, & x \in \Gamma_-\,, \\ \;\;\;0\,, & x \in \Gamma \setminus (\Gamma_- \cup \Gamma_+)\,, \end{cases} \tag{70.6}$$

wobei ν wieder die äußere Normale auf dem Rand von Ω bezeichnet. Hierbei handelt es sich um eine sogenannte *Neumann-Randvorgabe*, also eine Randbedingung für den Fluß der gesuchten Lösung von (70.5). Ist u eine Lösung von (70.5), (70.6), so gilt dies offenbar auch für jede Funktion $u + c$, sofern c eine Konstante ist. Dies entspricht dem bekannten Sachverhalt, daß das elektrische Potential erst durch ein Referenzpotential festgelegt ist (etwa die Erdung): In der Praxis wird daher immer die Spannung zwischen zwei Punkten (also die Potentialdifferenz) gemessen.

Das linke Bild in Abbildung 70.1 zeigt die elektrischen Feldlinien (durchgezogene Linien) und die Äquipotentiallinien (gebrochene Linien) dieses Potentials bei einem homogenen kreisförmigen Körper Ω im $\mathbb{R}^2$. Der Einfachheit halber sind hierbei die Elektroden ininitesimal klein, da in diesem Extremfall das Grenzpotential bezüglich einer geeigneten Erdung explizit angegeben werden kann (für $\sigma = 1$ im Einheitskreis Ω):

$$u(x) \;=\; \frac{1}{2\pi}\,\log\frac{1 - 2r\cos\theta + r^2}{1 - 2r\cos(\theta - \varphi) + r^2}\,, \qquad \begin{array}{c} x = (r\cos\theta, r\sin\theta)\,, \\ 0 \le r \le 1,\ 0 \le \theta \le 2\pi\,. \end{array}$$

Die positive Elektrode Γ_+ sitzt in diesem Fall im Punkt $(\cos\varphi, \sin\varphi)$, die negative Elektrode im Punkt $(1, 0)$. Das Potential weist an beiden Elektroden Singularitäten auf.

Die Werte des Potentials können an der Oberfläche des Körpers durch Spannungsmessungen abgegriffen werden. Auf diese Weise läßt sich überprüfen, ob im Innern des Gegenstands eine Anomalie vorliegt, die die homogene Leitfähigkeit des Körpers zerstört; denkbar sind etwa Risse innerhalb des Metallgegenstands. Derartige Ansätze sind für nichtdestruktive Prüfmethoden in der Qualitätssicherung von Bedeutung.

Zur Veranschaulichung betrachte man das rechte Bild in Abbildung 70.1, das den Verlauf der Feldlinien zeigt, wenn der (ansonsten homogene) Körper drei isolierende Hohlkörper K_i, $i = 1, 2, 3$, enthält (letztere entsprechen einer Leitfähigkeit $\sigma = 0$). Aufgrund der Isolierung umfließt der elektrische Strom die Hohlkörper, das heißt der Flußvektor steht senkrecht auf dem Normalenvektor ν,

$$\iota \cdot \nu = \sigma \frac{\partial u}{\partial \nu} = 0$$

auf den Rändern ∂K_i, $i = 1, 2, 3$. Man erkennt in der Tat in dem Bild, daß die Feldlinien die Isolatoren meiden und die Äquipotentiallinien senkrecht in sie einmünden. ◊

Beispiel 70.2. Ein anderes Beispiel eines Diffusionsprozesses tritt in der Bodenphysik auf und ist bei der Ausbreitung von Giftstoffen im Grundwasser oder der Erschließung von Erdölvorräten von Bedeutung, vgl. Knabner und Angermann [61]. Die Erdschichten des Bodens bilden eine mehr oder weniger lockere, gleichmäßige Verteilung feinkörniger Feststoffe, das sogenannte *Bodenskelett*. Dieses Skelett ist zusammenhängend und starr, die freien Zwischenräume werden *Porenraum* genannt und sind mit Luft und/oder Wasser gefüllt. Wir wollen im weiteren annehmen, daß sich die Luft im Porenraum frei bewegen kann, während die Wasserbewegung aufgrund des allgegenwärtigen Bodenskeletts einen diffusiven Charakter besitzt. Diese Wasserbewegung wollen wir im folgenden mathematisch beschreiben.

Zur Vereinfachung des Modells vernachlässigen wir im weiteren die ortsabhängige Mikrostruktur des Bodenskeletts und betrachten den Boden als (makroskopisch) homogen. Mit ρ bezeichnen wir die *volumetrische Wasserdichte* im Boden. Da Wasser inkompressibel ist und der Boden homogen sein soll, ist diese Wasserdichte in jedem Punkt des Bodens durch eine Konstante ρ_1 nach oben beschränkt. Wo dieser Wert erreicht wird, spricht man von *gesättigtem Boden*; der Quotient

$$w = \rho/\rho_1 \in [0, 1] \tag{70.7}$$

beschreibt den *Sättigungsgrad* des Porenraums.

Vereinfacht dargestellt, wird das Wasser von ungesättigten Poren angesaugt, so wie ein Schwamm Wasser aufnimmt. Physikalisch läßt sich dies durch eine *hydraulische* (*volumetrische*) *Energiedichte* E_H des Wassers beschreiben, die für gesättigte Böden durch Null festgelegt wird und für ungesättigte Böden negativ ist. In der makroskopischen Sichtweise hängt diese Energiedichte allein vom Sättigungsgrad ab und die Funktion $w \mapsto -E_H(w)$ ist eine charakteristische Materialeigenschaft des Bodens. In der Bodenphysik wird die hydraulische Energiedichte üblicherweise durch die Formel

$$h(w) \;=\; \frac{E_H(w)}{\rho g} \;=\; \frac{E_H(w)}{\rho_1 w g}$$

in eine äquivalente Größe (Einheit: Meter), die sogenannte *Standrohrspiegelhöhe* (engl. *piezometric head*) umgerechnet; g ist wieder die Erdbeschleunigung.

$E_H(w)/\rho$ ist die spezifische hydraulische Energie des Wassers im teilgesättigten Boden und hieraus errechnet sich die hydraulische Energie $mE_H(w)/\rho = mgh(w)$ eines entsprechenden Wasserpartikels der Masse m. Die Standrohrspiegelhöhe ist also die fiktive Höhe, in der die Lageenergie des Wasserpartikels mit seiner hydraulischen Energie übereinstimmt.

Die (negative) Ableitung dieser Energie nach dem Ort ergibt die hydraulische Kraft, die auf die Wasserpartikel wirkt. Zusammen mit der Schwerkraft errechnen wir hieraus die Gesamtkraft

$$F \;=\; -mg(\operatorname{grad} h(w) + e_3)\,,$$

wobei e_3 den nach oben gerichteten vertikalen Koordinatenvektor und im weiteren ζ die zugehörige Ortskomponente bezeichnen soll. Mit dem Diffusionsmodell (70.1), (70.2), ergibt dies den Fluß

$$J \;=\; -\,\frac{g\rho\tau}{2}\,(\operatorname{grad} h(w) + e_3)\,. \tag{70.8}$$

Die Sättigung selbst genügt dann wegen $w = \rho/\rho_1$ der *Richards-Gleichung*

$$w_t \;=\; \operatorname{div}\bigl(\sigma(w)h'(w)\operatorname{grad} w\bigr) \;+\; \frac{\partial}{\partial\zeta}\,\sigma(w)\,, \tag{70.9}$$

in der

$$\sigma(w) \;=\; g w \tau/2$$

die *hydraulische Leitfähigkeit* bezeichnet. Es sei angeführt, daß der Fluß (70.8) mit experimentellen Messungen gute Übereinstimmung aufweist (*Gesetz von Darcy*).

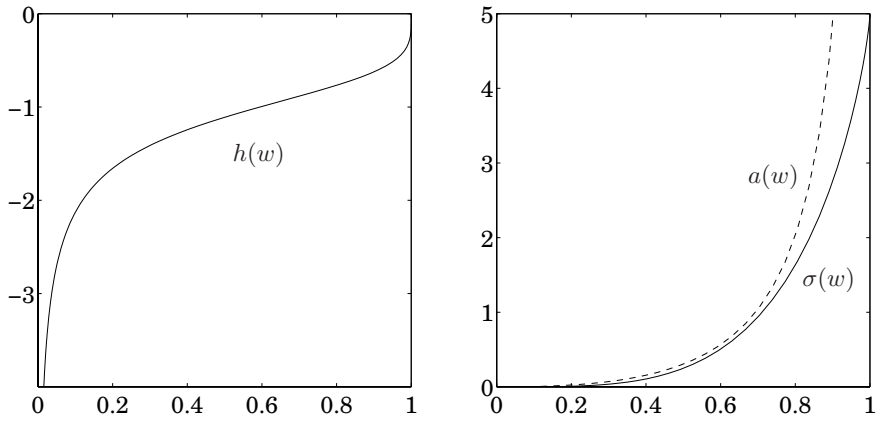


Abb. 70.2: Standrohrspiegelhöhe h (links) sowie hydraulische Leitfähigkeit σ und Diffusionskoeffizient a (rechts) als Funktion des Sättigungsgrads w

Für numerische Simulationen der Wasserbewegung müssen konkrete Funktionen h und σ eingesetzt werden, allerdings sind deren Abhängigkeiten vom Sättigungsgrad experimentell nur schwer zu bestimmen. Üblicherweise wird die Standrohrspiegelhöhe h durch eine monoton wachsende Funktion beschrieben, die wegen der obigen Normierung der hydraulischen Energie die Randbedingung $h(1) = 0$ erfüllen muß. Ein gängiges (dimensionsloses) Modell dieser Art stammt von van Genuchten,

$$h(w) \;=\; -\,\frac{(1-w^{\nu/(\nu-1)})^{1/\nu}}{w^{1/(\nu-1)}}\,, \qquad 0<w<1\,, \tag{70.10}$$

wobei $\nu > 1$ ein freier Parameter ist. Der Graph von h ist für $\nu = 4$ in Abbildung 70.2 (links) dargestellt. Während bei $w = 1$ ein wurzelartiges Verhalten vorliegt, strebt die Standrohrspiegelhöhe für $w \to 0$ gegen $-\infty$. Die Ableitung von h hat demnach sowohl bei $w = 0$ als auch bei $w = 1$ eine Singularität, so daß in der Nähe dieser Extremwerte starke hydraulische Kräfte vorliegen.

Die hydraulische Leitfähigkeit σ nimmt im allgemeinen mit wachsender Sättigung zu und erfüllt $\sigma(0) = 0$. Bei einer trockenen Pore liegt also keine Leitfähigkeit vor; je mehr Teilchen vorhanden sind, also je größer die volumetrische Dichte (und damit der Sättigungsgrad) ist, um so größer ist die Leitfähigkeit. Die durchgezogene Kurve in Abbildung 70.2 (rechts) gehört zu einem entsprechenden Modell von Mualem, welches sich seinerseits an den van Genuchten-Ansatz (70.10) für die Standrohrspiegelhöhe und experimentell gewonnene Daten anlehnt:

$$\sigma(w) \;=\; \sigma_1 w^{\mu}\big(1-(1-w^{\frac{\nu}{\nu-1}})^{\frac{\nu-1}{\nu}}\big)^2\,, \qquad 0\le w\le 1\,. \tag{70.11}$$

Hierbei ist $\sigma_1 = \sigma(1)$ die hydraulische Leitfähigkeit im gesättigten Zustand, ν der Parameter aus dem van Genuchten-Modell und $\mu \in \mathbb{N}$ ein zusätzlicher freier Parameter. Dargestellt ist der Graph von σ für $\mu = 1$ und $\nu = 4$.

Es sei an dieser Stelle betont, daß der Diffusionskoeffizient

$$a(w) = \sigma(w)h'(w)$$

in der Richards-Gleichung (70.9) selbst von der Lösung w abhängt; die Richards-Gleichung ist *quasilinear*. Dies führt dazu, daß nur in wenigen Ausnahmefällen explizite Lösungen dieser Gleichung bekannt sind. Eine weitere Schwierigkeit ist die Tatsache, daß der Diffusionskoeffizient an der Stelle $w = 0$ verschwindet; man sagt, die Gleichung *degeneriert*. Für die Modellfunktionen h und σ aus (70.10), (70.11) gilt

$$a(w) \sim \frac{\sigma_1(\nu-1)}{\nu^2}\, w^{\mu+\frac{\nu}{\nu-1}}, \qquad w \to 0\,. \tag{70.12}$$

Der Graph von a ist als gebrochene Kurve in Abbildung 70.2 (rechts) eingezeichnet.

Bisher waren wir bei Diffusionsgleichungen immer davon ausgegangen, daß der Diffusionskoeffizient strikt positiv ist. Dieser Unterschied hat enorme Auswirkungen auf das qualitative Verhalten der Lösung. Zur Illustration betrachten wir die sogenannte *Poröse-Medien-Gleichung* eine eindimensionale Variante der Richards-Gleichung (70.9) ohne Einfluß der Gravitation,

$$w_t = (w^{n+1})_{xx} = \big(a(w)w_x\big)_x\,, \qquad n \in \mathbb{R}^+\,, \tag{70.13}$$

mit $a(w) = (n+1)w^n$. Dies mag als Modell für die Diffusion von Wasser in einer dünnen horizontalen porösen Bodenschicht gelten, die nach oben und unten durch wasserundurchlässige Gesteinsschichten abgegrenzt ist.

Für einen unendlich lang ausgedehnten ausgetrockneten Boden, in den lediglich an der Stelle $x = 0$ zur Zeit $t = 0$ eine (unendlich) große Menge Wasser eingespeist wird, ist die Lösung dieser Gleichung explizit bekannt, vgl. Grindrod [38, S. 220ff],

$$w(x,t) = \begin{cases} t^{-1/(n+2)} g(xt^{-1/(n+2)})\,, & |x| \le \alpha t^{1/(n+2)}\,, \\ 0\,, & |x| > \alpha t^{1/(n+2)}\,, \end{cases} \tag{70.14}$$

mit

$$g(z) = \Big(\frac{n}{(2n+2)(n+2)}\Big)^{1/n} (\alpha^2 - z^2)^{1/n}\,,$$

$$\alpha = \Big(\frac{(2n+2)(n+2)}{n}\Big)^{1/(n+2)} \Big(\int_{-1}^{1} (1-\zeta^2)^{1/n}\, d\zeta\Big)^{-n/(n+2)}\,.$$

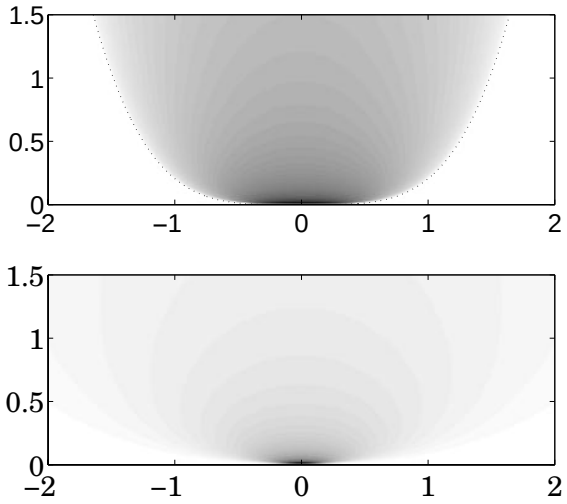


Abb. 70.3: Barenblatt-Lösung w (oben) und Lösung u der Wärmeleitungsgleichung (unten)

Diese Lösung ist für $n = 2$ in Abbildung 70.3 dargestellt. Zum Vergleich zeigt die Abbildung auch die entsprechende Lösung

$$u(x,t) \;=\; \frac{1}{\sqrt{4\pi t}}\,\exp\bigl(-\frac{x^2}{4t}\bigr) \tag{70.15}$$

für den Grenzfall $n = 0$, der der Wärmeleitungsgleichung entspricht (vgl. Abschnitt 69).

Ein entscheidender Unterschied zwischen den beiden Lösungen ist die Ausbreitungsgeschwindigkeit des Wassers. Im Fall der Wärmeleitungsgleichung ist die Ausbreitungsgeschwindigkeit theoretisch beliebig groß: Wie man an der Lösung (70.15) erkennt, ist $u(x,t)$ für beliebiges $x \in \mathbb{R}$ und jedes noch so kleine $t > 0$ positiv. In einem porösen Medium ist die Ausbreitungsgeschwindigkeit des Wassers hingegen endlich: bei der Lösung w aus (70.14) beispielsweise ist das Wasser zum Zeitpunkt t maximal bis $|x| = \alpha t^{1/(n+2)}$ vorgedrungen; es entsteht ein sogenannter *freier Rand* (in der Abbildung gepunktet eingezeichnet), an dem die Lösung Null wird und nicht differenzierbar ist. ◊

71 Kontinuumsmechanik

In Abschnitt 3 haben wir die Verformung eines Tragwerks unter der Einwirkung äußerer Kräfte untersucht. Das dort verwendete Masse-Feder-Modell beruht auf der Annahme, daß die Masse des Tragwerks auf endlich viele Gelenke

konzentriert ist. Im Gegensatz dazu beschäftigt sich die Kontinuumsmechanik mit der Verformung von Körpern, deren Masse kontinuierlich verteilt ist, vgl. etwa die Bücher von Braess [10] und von Temam und Miranville [101]. Wir betrachten im folgenden einen solchen Körper, der im unbelasteten Ruhezustand ein Gebiet $\Omega \subset \mathbb{R}^3$ ausfüllt und dessen Dichte konstant ist. Auf den Körper wirken äußere Kräfte, die wir in Volumenkräfte $f : \Omega \to \mathbb{R}^3$ und Oberflächenkräfte $g : \Gamma \to \mathbb{R}^3$ unterscheiden, wobei letztere nur an der Oberfläche $\Gamma = \partial\Omega$ des Körpers angreifen. Ein Beispiel für eine Volumenkraft ist die Schwerkraft, während eine auf dem Körper ruhende Last eine Oberflächenkraft verursacht. Wir behandeln ausschließlich statische Probleme, bei denen die Kräfte unabhängig von der Zeit sind.

Durch die Krafteinwirkung verformt sich der Körper und es stellt sich ein Gleichgewichtszustand ein, in dem der Körper ein Gebiet $\tilde{\Omega}$ einnimmt. Um diese Verformung zu berechnen, verwenden wir wie in Abschnitt 3 das Schnittprinzip (vgl. auch Abschnitt 67.2), indem wir ein beliebiges Teilgebiet $\tilde{G} \subset \tilde{\Omega}$ freischneiden und die auf $\tilde{G}$ wirkenden Volumen- und Oberflächenkräfte bestimmen; aufgrund des statischen Gleichgewichts müssen sich diese Kräfte zu Null summieren. Da das Gebiet $\tilde{\Omega}$ unbekannt ist, erweist es sich als vorteilhaft, alle Rechnungen auf das ursprüngliche Gebiet zu transformieren. Dazu beschreiben wir die Verformung durch die bijektive Abbildung

$$\Phi : \Omega \to \tilde{\Omega}$$

und bezeichnen mit G das Urbild von $\tilde{G}$ unter Φ; ν und $\tilde{\nu}$ seien die entsprechenden Normalenvektoren in den Punkten x und $\tilde{x} = \Phi(x)$ auf den Oberflächen ∂G beziehungsweise $\partial\tilde{G}$. Aufgrund der Bijektivität ist $\det \Phi'(x)$ von Null verschieden und aus physikalischen Gründen positiv.

Im weiteren wird vorausgesetzt, daß es sich bei den äußeren Kräften um sogenannte *Totlasten* handelt, das sind Kräfte, die sich bei der Deformation nicht ändern. Dies beinhaltet, daß die auf das deformierte Gebiet $\tilde{G}$ wirkende Volumenkraft mit der auf G wirkenden Kraft

$$F = \int_G f(x)\, dx \tag{71.1}$$

identisch ist. Diese Voraussetzung ist bei den meisten Anwendungen erfüllt. Nun wenden wir uns den auf $\partial\tilde{G}$ wirkenden Oberflächenkräften zu. Dazu sei an Abschnitt 67.2 erinnert, wo wir das Schnittprinzip in einem Fluid angewendet haben. Dort wurden reibungsfreie Fluide dadurch charakterisiert, daß die Spannungsvektoren immer den Normalenvektoren entgegengesetzt sind. Bei der Deformation eines festen Körpers können jedoch *Scherspannungen* auftreten, die durch allgemeinere Spannungsvektoren beschrieben werden. Entscheidend ist dabei die folgende Beobachtung von Cauchy: Der Spannungsvektor

$\varphi(\tilde{x})$ auf ein Flächenelement im Punkt $\tilde{x}$ hängt linear von der dortigen Normalen $\tilde{\nu}$ ab. Diese lineare Abhängigkeit wird durch den *Cauchyschen Spannungstensor* $T(\tilde{x}) \in \mathbb{R}^{3\times 3}$ beschrieben: $\varphi(\tilde{x}) = T(\tilde{x})\tilde{\nu}$. Die Oberflächenkraft, die auf das Gebiet $\tilde{G}$ wirkt, hat demnach die Form

$$\int_{\partial\tilde{G}} T(\tilde{x})\tilde{\nu}\, d\tilde{s}\,, \tag{71.2}$$

wobei $d\tilde{s}$ das Oberflächenelement von $\partial\tilde{G}$ bezeichnet.

Das statische Gleichgewicht ist dadurch charakterisiert, daß die beiden Kräfte (71.1) und (71.2) im Gleichgewicht stehen, d. h. es ist

$$\int_G f(x)\, dx \;+\; \int_{\partial\tilde{G}} T(\tilde{x})\tilde{\nu}\, d\tilde{s} \;=\; 0\,.$$

Mit etwas Aufwand kann das zweite Integral in ein Oberflächenintegral über ∂G transformiert werden (vgl. etwa Ciarlet [18, Theorem 1.7-1]), und dann haben wir die Gleichung

$$\int_G f(x)\, dx \;+\; \int_{\partial G} P(x)\nu\, ds \;=\; 0\,, \tag{71.3}$$

in der

$$P(x) \;=\; T(\Phi(x))\Phi'(x)^{-*} \det \Phi'(x) \;\in \mathbb{R}^{3\times 3}$$

den sogenannten *ersten Piola-Kirchhoffschen Spannungstensor* bezeichnet. Das Kräftegleichgewicht (71.3) stellt eine integrale Erhaltungsgleichung dar, die unter den üblichen Glattheitsannahmen zu dem Randwertproblem

$$-\,\mathbf{div}\, P \;=\; f \quad \text{in } \Omega\,, \qquad P(x)\nu \;=\; g \quad \text{auf } \Gamma\,, \tag{71.4}$$

äquivalent ist; hierbei ist die Vektordivergenz **div** wieder zeilenweise auf die Matrix P anzuwenden, vgl. (67.4). Die Randvorgabe $P(x)\nu = g$ beschreibt die wirkenden Oberflächenkräfte.

Wie in Abschnitt 3 hängen die Spannungsvektoren von den *Verzerrungen* des Körpers ab und die Verzerrungen ergeben sich ihrerseits aus der Deformation Φ. Ein Volumenelement dx im Punkt x wird durch Φ in $\Phi'(x)\, dx$ deformiert; die Singulärwertzerlegung von Φ' gibt an, welche orthogonalen Achsen wie stark gestaucht bzw. gestreckt werden, vgl. Abbildung 12.1. Dieselbe Information enthalten die Eigenvektoren und Eigenwerte des sogenannten *Cauchy-Greenschen Verzerrungstensors*

$$C \;=\; \Phi'(x)^{*}\Phi'(x)\,.$$

Für $C = I$ ist die Deformation lokal eine Isometrie[1], die Matrix

$$E \;=\; (C - I)/2$$

beschreibt daher die *Verzerrungen* im Punkt x.

In der *linearen Elastizitätstheorie* wird davon ausgegangen, daß die *Verschiebungen*

$$u(x) \;=\; \Phi(x) - x$$

der einzelnen Punkte des Körpers klein sind; höhere Potenzen von u oder u' werden daher konsequent vernachlässigt. Unter dieser Annahme ergibt sich für die Verzerrungen die Approximation

$$E \;\approx\; \varepsilon \;=\; (u' + u'^*)/2 \;\in \mathbb{R}^{3\times 3}\,. \tag{71.5}$$

Außerdem können die beiden Spannungstensoren T und P der Einfachheit halber miteinander identifiziert werden, da $\Phi'(x)$ in erster Näherung mit der Einheitsmatrix übereinstimmt.

Es verbleibt lediglich noch eine Verbindung zwischen den Spannungen und den Verzerrungen herzustellen. Dabei handelt es sich um ein Materialgesetz. Für homogene elastische Materialien, die isotrop sind (d. h. die Materialeigenschaften sind in alle Richtungen gleich ausgeprägt), ist dies das *Hookesche Gesetz*: Es besagt, daß der Spannungstensor durch

$$P(x) \;=\; \lambda\,\mathrm{Spur}(\varepsilon) I + 2\mu\varepsilon \tag{71.6}$$

gegeben ist; λ und μ sind die sogenannten *Lamé-Konstanten*. Zum Verständnis des Hookeschen Gesetzes sei darauf hingewiesen, daß der Spannungstensor und die Verzerrungen nach (71.6) die gleichen Eigenvektoren haben und die Eigenwerte des Spannungstensors aus zwei Komponenten bestehen. Die zweite Komponente – aus dem hinteren Term in (71.6) – ist proportional zu der Stärke der jeweiligen Verzerrung. Die erste Komponente ist hingegen in alle Richtungen gleich (wie der Druck bei idealen Gasen). Um diesen Term interpretieren zu können, beachten wir, daß

$$\det \Phi' \;=\; \det(I + u') \;\approx\; 1 + \mathrm{Spur}(u') \;=\; 1 + \mathrm{Spur}(\varepsilon)\,,$$

da die Einträge von u' deutlich kleiner als Eins sind. Ein Kontrollvolumen dx im Punkt x wird somit auf ein Volumen

$$\det \Phi'(x)\, dx \;\approx\; (1 + \mathrm{Spur}(\varepsilon))\, dx$$

[1] Wegen $\det \Phi' > 0$ ist Φ lokal eine *Starrkörperbewegung*, d. h. Φ ergibt sich lokal aus einer Translation und einer Rotation.

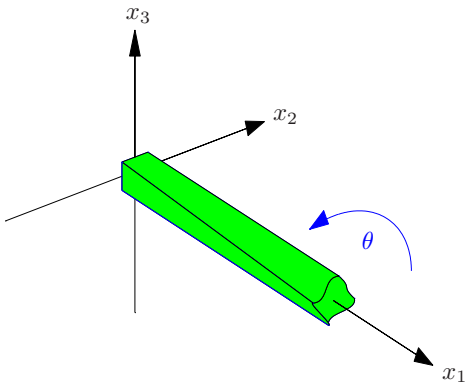


Abb. 71.1:
Torsion eines Balkens

abgebildet und der erste Term in (71.6) ist proportional zur Volumenänderung.

Wird das Hookesche Gesetz (71.6) in die Erhaltungsgleichung (71.4) eingesetzt, so ergibt sich unter Berücksichtigung von $\mathrm{Spur}(\varepsilon) = \operatorname{div} u$ die sogenannte *Lamésche Differentialgleichung* für u:

$$\begin{aligned} -\lambda \operatorname{grad}\operatorname{div} u \;-\; 2\mu\, \mathbf{div}\,\varepsilon &= f \quad \text{in } \Omega\,, \\ \big((\lambda \operatorname{div} u)I + 2\mu\varepsilon\big)\nu &= g \quad \text{auf } \Gamma\,, \end{aligned} \qquad \varepsilon = (u' + u'^{*})/2\,. \quad (71.7)$$

Beispiel 71.1. Zur Illustration untersuchen wir die Torsion eines Balkens der Länge 1 mit quadratischem Querschnitt $Q = [-a, a]^2$. Wie in Abbildung 71.1 liegt der Stab längs der x_1-Achse und ist an einem Ende ($x_1 = 0$) fest eingespannt, während er am anderen Ende ($x_1 = 1$) im mathematisch positiven Sinn um seine Achse gedreht wird. Genauer wollen wir annehmen, daß jeder Punkt x auf den vier rechten äußeren Kanten des Stabs um einen kleinen Winkel θ in der affinen Ebene $x_1 = 1$ gedreht wird: Ein solcher Punkt x hat Koordinaten $[1, x_2, x_3]^T$ mit $|x_2| = a$ oder $|x_3| = a$ und wird durch die Drehung auf

$$\tilde{x} \;=\; \begin{bmatrix} 1 + u_1(x) \\ x_2 \cos\theta - x_3 \sin\theta \\ x_2 \sin\theta + x_3 \cos\theta \end{bmatrix} \;\approx\; x + \begin{bmatrix} u_1(x) \\ -s x_3 \\ s x_2 \end{bmatrix}, \qquad s = \sin\theta\,,$$

abgebildet; die x_1-Koordinate darf sich dabei um $u_1(x)$ verändern, vgl. Abbildung 71.1. Weiterhin nehmen wir an, daß die Torsion des Stabs längs der x_1-Achse proportional zu x_1 zunimmt, d. h. wir geben die Verschiebungen

$$\begin{aligned} u_2(x) &= -s x_3 x_1\,, \\ u_3(x) &= s x_2 x_1\,, \end{aligned} \qquad x \in \Gamma\,, \quad (71.8)$$

vor. Im Gegensatz zu der Neumann-Randbedingung in (71.7) handelt es sich bei (71.8) um Dirichlet-Randvorgaben für die Verschiebungen u_2 und u_3. Lediglich für die Spannungen in x_1-Richtung liegt eine Neumann-artige Randbedingung vor,

$$e_1 \cdot \big(P(x)\nu\big) = 0\,, \qquad x \in \Gamma\,, \tag{71.9}$$

das heißt in x_1-Richtung kann sich der Stab frei bewegen. Volumenkräfte seien vernachlässigbar.

Der Stab kann auf die Torsion nur durch eine Verwölbung der Querschnitte reagieren und wir setzen daher die Lösung wie folgt an: Die Verschiebungen u_2 und u_3 seien für jeden Punkt des Stabs durch (71.8) gegeben und die Verschiebung u_1 in x_1-Richtung sei für alle x_1 gleich:

$$\begin{aligned} u_1(x) &= v(x_2, x_3)\,, \\ u_2(x) &= -s x_3 x_1\,, \qquad x \in \Omega\,. \\ u_3(x) &= s x_2 x_1\,, \end{aligned} \tag{71.10}$$

Um die Verwölbungsfunktion v zu bestimmen, setzen wir diesen Ansatz in die Lamé-Gleichung ein. Wegen $\operatorname{div} u = 0$ und

$$\varepsilon = \frac{1}{2}\begin{bmatrix} 0 & v_{x_2} - s x_3 & v_{x_3} + s x_2 \\ v_{x_2} - s x_3 & 0 & 0 \\ v_{x_3} + s x_2 & 0 & 0 \end{bmatrix}$$

ergibt dies

$$0 = \lambda \operatorname{grad}\operatorname{div} u + 2\mu\, \mathbf{div}\, \varepsilon = \mu\, [\Delta v, 0, 0]^T\,.$$

Somit haben wir gezeigt: Ist v eine Lösung der zweidimensionalen Laplace-Gleichung $\Delta v = 0$ in dem Querschnitt $Q = (-a, a)^2$, so ist die Funktion u aus (71.10) eine Lösung der Lamé-Gleichung. Die Dirichlet-Randvorgaben (71.8) sind aufgrund des Lösungsansatzes automatisch erfüllt. Die Neumann-Randbedingung (71.9) ist äquivalent zu

$$\mu\,\big((v_{x_2} - s x_3)\nu_2 + (v_{x_3} + s x_2)\nu_3\big) = 0$$

und damit äquivalent zu der Neumann-Randvorgabe

$$\frac{\partial v}{\partial \nu_Q} = \begin{cases} \pm s x_3\,, & x_2 = \pm a\,,\ -a < x_3 < a\,, \\ \mp s x_2\,, & x_3 = \pm a\,,\ -a < x_2 < a\,, \end{cases} \tag{71.11}$$

an v (ν_Q bezeichne die äußere Normale auf ∂Q). Abbildung 71.2 zeigt die zugehörige Verwölbung v. ◊

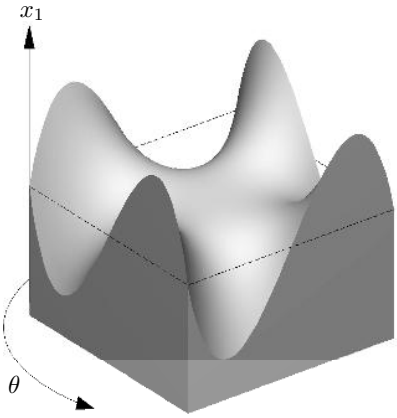


Abb. 71.2:
Form der Verwölbung

72 Finanzmathematik

Brownsche Bewegungen, die uns zur Illustration von Diffusionsprozessen in Abschnitt 69 begegnet waren, spielen auch bei der Modellierung des Aktienmarkts eine grundlegende Rolle. Abbildung 72.1 zeigt einen simulierten Aktienkurs $x = x(t)$: Dargestellt sind die Aktienwerte $x_n = x(t_n)$ über einem äquidistanten Zeitgitter $\{t_n = n\tau \,:\, n = 0, 1, \dots\}$ mit Gitterweite τ. Für die Simulation wurde ein gängiges mathematisches Aktienmodell übernommen, nach dem der Zuwachs des Aktienkurses in jedem Zeitintervall proportional zu dem aktuellen Wert der Aktie ist,

$$x_{n+1} \,=\, x_n + \triangle x_n\,, \qquad \triangle x_n \,=\, \mu_n x_n\,. \tag{72.1}$$

Dabei sind die Zuwachsraten μ_n, $n \in \mathbb{N}_0$, stochastisch unabhängige Zufallsvariablen mit Erwartungswert $\mu\tau$. Wir nehmen im folgenden, wie zu Beginn von Abschnitt 69 der Einfachheit halber an, daß die Differenz $\triangle\mu_n = \mu_n - \mu\tau$ eine stochastische Irrfahrt ist mit

$$\mathcal{P}(\triangle\mu_n = \sqrt{2\sigma\tau}) \,=\, \mathcal{P}(\triangle\mu_n = -\sqrt{2\sigma\tau}) \,=\, 1/2\,. \tag{72.2}$$

In der Finanzmathematik bezeichnet man μ als *Drift* und $\sqrt{2\sigma}$ als *Volatilität*. Die gebrochene Kurve in Abbildung 72.1 zeigt zum Vergleich den Graph der Funktion $t \mapsto e^{\mu t}$, der sich ohne stochastische Einflüsse im Grenzfall $\tau \to 0$ als Aktienkurs ergeben würde.

An der Börse werden neben den klassischen Aktien auch sogenannte *Optionen* gehandelt. Ein Beispiel für eine solche Option ist der *europäische Put*: Mit dem Kauf eines europäischen Puts erwirbt man das Recht, nach einer festgelegten Laufzeit $T > 0$ eine Aktie zum *Basispreis* $X > 0$ zu verkaufen – unabhängig

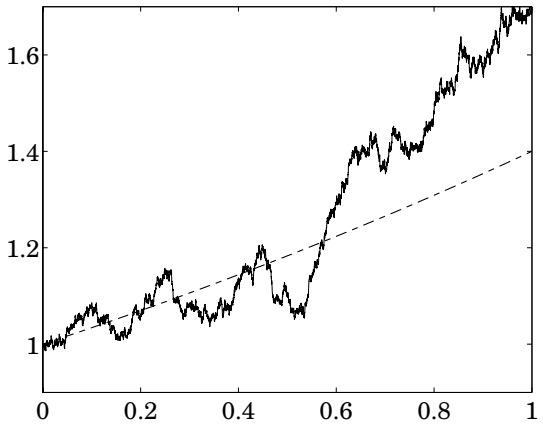


Abb. 72.1: Simulation eines Aktienkurses

vom tatsächlichen Wert $x(T)$ der Aktie zu diesem Zeitpunkt. Ist der Wert X größer als $x(T)$, so bringt der Besitz dieser Option zum Zeitpunkt $t = T$ den Gewinn $X - x(T)$, indem man eine Aktie zum Tageskurs $x(T)$ kauft und anschließend die Option ausübt und die Aktie zum fixierten Basispreis X wieder verkauft. Ist hingegen der aktuelle Wert $x(T)$ größer als X, so erweist sich die Option zum Fälligkeitszeitpunkt als wertlos. Der Wert $u = u(x, T)$ der Option zum Zeitpunkt $t = T$ als Funktion vom Tageskurs x der Aktie errechnet sich also aus der Formel

$$u(x, T) \;=\; \max\{X - x, 0\}\,. \tag{72.3}$$

Sowohl für die Banken als auch für ihre Kunden stellt sich die Frage, welchen (möglichst objektiven) Wert eine Option zum Zeitpunkt $t < T$ besitzt, also wie der Kaufpreis der Option festzulegen ist. Man wird erwarten, daß für feste Optionsparameter T und X dieser Preis $u(x, t)$ sowohl vom Zeitpunkt t als auch vom aktuellen Preis $x = x(t)$ der Aktie abhängt. Es ist weiterhin plausibel, daß zu einem festen Zeitpunkt t der Optionspreis $u(\,\cdot\,, t)$ – wie in (72.3) für $t = T$ – eine monoton fallende Funktion des Aktienpreises ist, denn mit steigendem Aktienwert x sinkt der zu erwartende Erlös (72.3) zum Fälligkeitstermin.

In den siebziger Jahren entwickelten Black und Scholes ein Modell zur Bestimmung eines fairen Optionspreises, das sich als äußerst erfolgreich erwiesen hat und schließlich 1997 mit der Verleihung des Wirtschaftsnobelpreises an Merton und Scholes honoriert wurde. Die Grundlage der *Black-Scholes-Formel* (vgl. Aufgabe 11) zur Berechnung von $u(x, t)$ ist ein Diffusionsprozeß.

Dieses Modell basiert auf verschiedenen Annahmen, man vergleiche etwa das Buch von Seydel [96]. Einerseits wird vorausgesetzt, daß die Transaktions-

kosten für den Handel mit Aktien und Optionen vernachlässigbar sind und daß es sich bei den Käufern der Optionen um „Kleininvestoren“ handelt, deren Transaktionen keine Auswirkungen auf den Aktienkurs besitzen. Natürlich sollen darüber hinaus weder der Käufer noch die Bank über Kenntnisse über die zukünftige Aktienentwicklung verfügen; Drift und Volatilität des Aktienkurses seien hingegen allgemein bekannt.

In das Modell geht außerdem ein, daß jederzeit Geld von der Bank zu einem festen *Bondzinssatz* q geliehen werden kann. Ein Kleininvestor könnte sich beispielsweise zum Zeitpunkt $t = t_n$, $n \in \mathbb{N}_0$, von der Bank den Geldbetrag $u_n + \alpha_n x_n$, $u_n = u(x_n, t_n)$, für den Kauf einer Option und α_n Aktien leihen und würde dann zum Zeitpunkt $t = t_{n+1}$ über einen Nettoerlös

$$g \;=\; u_{n+1} + \alpha_n x_{n+1} - e^{q\tau}(u_n + \alpha_n x_n)\,, \qquad \begin{array}{l} x_{n+1} = x(t_{n+1})\,, \\ u_{n+1} = u(x_{n+1}, t_{n+1})\,, \end{array} \tag{72.4}$$

verfügen; der Term $e^{q\tau}(u_n + \alpha_n x_n)$ in (72.4) entspricht dem an die Bank zu zahlenden verzinsten Kredit.

Um abwägen zu können, ob es sich bei g aus (72.4) um einen Gewinn handelt, also ob g positiv ist, müssen wir den Wert u_{n+1} der Option zum Zeitpunkt t_{n+1} abschätzen. Falls der Optionspreis $u = u(x,t)$ eine hinreichend glatte Funktion von x und t ist, ergibt eine Taylorentwicklung von u um (x_n, t_n)

$$u_{n+1} \;=\; u_n + u_x \triangle x_n + u_t \tau + \tfrac{1}{2} u_{xx} (\triangle x_n)^2 + u_{xt} \tau \triangle x_n + \tfrac{1}{2} u_{tt} \tau^2 + \ldots$$

Das Argument (x_n, t_n) bei den partiellen Ableitungen von u wird hier und im weiteren der Übersichtlichkeit halber weggelassen. Aufgrund des Zuwachsmodells (72.1), (72.2), ist dabei

$$\triangle x_n \;=\; \mu_n x_n \;=\; (\mu\tau + \triangle \mu_n) x_n \;=\; O(\sqrt{\tau})\,, \qquad \tau \to 0\,,$$

und daher

$$u_{n+1} \;=\; u_n \;+\; \mu_n x_n u_x \;+\; \tau u_t \;+\; \tfrac{1}{2} (\triangle \mu_n)^2 x_n^2 u_{xx} \;+\; O(\tau^{3/2})\,.$$

Wegen $(\triangle \mu_n)^2 = 2\sigma\tau$ ergibt sich somit für den Erlös aus (72.4) der Wert

$$\begin{aligned} g \;&=\; u_{n+1} + \alpha_n x_{n+1} - e^{q\tau}(u_n + \alpha_n x_n) \\ &=\; u_{n+1} + \alpha_n x_n + \alpha_n \mu_n x_n - (1 + q\tau + O(\tau^2))(u_n + \alpha_n x_n) \\ &=\; u_{n+1} + \alpha_n \mu_n x_n - u_n - q\tau u_n - \alpha_n q \tau x_n + O(\tau^2) \\ &=\; \mu_n x_n (u_x + \alpha_n) + \tau u_t + \sigma x_n^2 \tau u_{xx} - q\tau u_n - \alpha_n q \tau x_n + O(\tau^{3/2})\,. \end{aligned}$$

In dieser Formel ist zu beachten, daß für kleine Zeitschritte $\tau > 0$ allenfalls im ersten Term der letzten Zeile,

$$\mu_n x_n (u_x + \alpha_n) \;=\; \mu\tau x_n (u_x + \alpha_n) \;\pm\; \sqrt{2\sigma\tau}\, x_n (u_x + \alpha_n)\,,$$

ein Summand der Größenordnung $\sqrt{\tau}$ auftritt und daß ausgerechnet bei diesem dominanten Term das Vorzeichen von der stochastischen Komponente des Aktienkurses abhängt. Ein Kleininvestor ist daher gut beraten, sein Risiko zu minimieren und die Kaufstrategie

$$\alpha_n \;=\; -u_x \;=\; -u_x(x_n, t_n) \tag{72.5}$$

zu verfolgen. Wegen der Monotonie von u ist $\alpha_n \geq 0$, d. h. der Kleininvestor sollte sich neben einer Option immer auch einen entsprechenden Anteil Aktien zulegen. Diese Strategie ergibt einen Nettoerlös

$$g = \tau\big(\, u_t \,+\, \sigma x^2\, u_{xx} \,-\, qu \,+\, qxu_x \,\big) \,+\, O(\tau^{3/2})\,, \tag{72.6}$$

wobei x für x_n und u für u_n steht. Da der objektive Preis der Option gesucht wird, muß der Klammerausdruck Null sein, da ansonsten entweder die Bank oder der Kleininvestor in jedem (hinreichend kleinen) Zeitschritt einen sicheren Verlust hinnehmen müssen.

Mit anderen Worten, der faire Preis $u(x,t)$ eines europäischen Puts erfüllt die Differentialgleichung

$$u_t \;=\; -\sigma x^2 u_{xx} - qxu_x + qu\,, \qquad 0 < t < T\,,\; x > 0\,. \tag{72.7}$$

Dabei ist interessant, daß diese Gleichung nicht von der Drift μ des Aktienkurses sondern lediglich von seiner Volatilität σ abhängt.

Um die Lösung der Differentialgleichung (72.7) eindeutig festzulegen, werden noch Anfangs- und Randbedingungen benötigt. Hier tritt ein neuartiges Phänomen auf: Wir haben lediglich eine *Endbedingung* (72.3) für $u(\,\cdot\,, T)$ aber keine Anfangsbedingung. Um dieses Problem zu beheben, substituiert man einfach t durch $T - t$, vgl. (72.10). Dadurch ändern sich alle Vorzeichen auf der rechten Seite der Differentialgleichung (72.7); dies hat den angenehmen Nebeneffekt, daß der Diffusionskoeffizient vor u_{xx} in (72.7) wie bei den vorangegangenen Beispielen positiv wird.

Wir leiten nun noch Randbedingungen für $x = 0$ und $x = +\infty$ her. An der Stelle $x = 0$ ergibt sich aus der Differentialgleichung (72.7)

$$u_t(0,t) \;=\; qu(0,t)\,, \qquad u(0,T) \;=\; X\,.$$

Folglich ist

$$u(0,t) \;=\; Xe^{-q(T-t)}\,, \qquad t < T\,, \tag{72.8}$$

die korrekte Randbedingung für $x = 0$. Für festes t und $x \to \infty$ sinkt hingegen die Gewinnerwartung (72.3) gegen Null, so daß die Putoption zunehmend wertloser wird. Es ist daher sinnvoll, für u die zweite Randbedingung

$$\lim_{x\to\infty} u(x,t) \;=\; 0\,, \qquad 0 < t < T\,, \tag{72.9}$$

anzusetzen. Die gesuchte Lösung $u(x,t)$ mit diesen Randwerten kann explizit angegeben werden, vgl. Aufgabe 11.

Durch die Substitution

$$v(\xi,t) \;=\; e^{qt}u(e^{\xi},T-t)\,, \qquad t>0\,,\ \xi\in\mathbb{R}\,, \tag{72.10}$$

wird die Differentialgleichung (72.7) in die aus Beispiel 69.1 vertraute Diffusionsgleichung

$$v_t \;=\; \sigma v_{\xi\xi} + (q-\sigma)v_{\xi}\,, \qquad \xi\in\mathbb{R}\,,\ t>0\,, \tag{72.11}$$

transformiert. Dort hatten wir diese Differentialgleichung aus einem Erhaltungsgesetz für v mit zugehörigem Fluß

$$J \;=\; (\sigma-q)v - \sigma v_{\xi} \;=\; \sigma(v-v_{\xi}) - qv \tag{72.12}$$

hergeleitet.

Der Wert von $v(\xi,t)$ entspricht den verzinsten Kosten, die ein Händler zur Zeit T zurückzahlen muß, falls er sich zur Zeit $T-t$ bei der Bank Geld leiht, um eine Option zu erwerben. In der Finanzwirtschaft versteht man daher unter v den *abgezinsten* oder *diskontierten* Wert der Option. Für den fairen Optionspreis ist dieser diskontierte Wert also eine Erhaltungsgröße. Ohne stochastischen Einfluß ($\sigma = 0$) in (72.12) würde die Transportgleichung mit Fluß $-qv$ den Wertzuwachs vollständig beschreiben. Ber zweite Term $\sigma(v-v_{\xi})$ modelliert den Wertverlust der Putoption aufgrund der Schwankungen der Aktie.

Im Gegensatz zu dem europäischen Put, dem eine feste Laufzeit $T>0$ zugrunde liegt, kann eine *amerikanische Putoption* zu jedem Zeitpunkt $0\le t\le T$ eingelöst werden. Der Wert $u(x,t)$ einer amerikanischen Option beträgt also mindestens

$$u(x,t) \;\ge\; \max\{X-x,0\}\,, \qquad 0\le t\le T\,. \tag{72.13}$$

Gilt zu einem Zeitpunkt $t<T$ das echte Ungleichheitszeichen, so wird man die Option nicht einlösen. Statt dessen empfiehlt sich die gleiche Kaufstrategie (72.5) wie bei einer europäischen Option (mit derselben Argumentation wie zuvor).

Andererseits existiert eine Kurve $x=\psi(t)$, $0\le t<T$, auf der der faire Preis $u(x,t)$ der Option die unterer Schranke (72.13) annimmt. Es läßt sich zeigen, daß dann

$$\begin{aligned} u(x,t) \;&>\; \max\{X-x,0\}\,, \qquad x>\psi(t)\,,\\ u(x,t) \;&=\; \max\{X-x,0\}\,, \qquad x\le\psi(t)\,. \end{aligned}$$

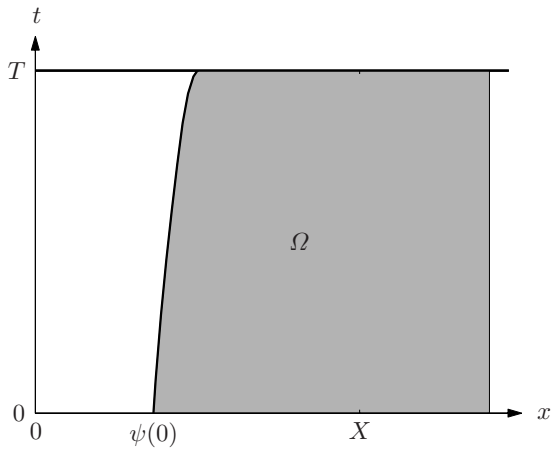


Abb. 72.2: Freies Randwertproblem für den Preis einer amerikanischen Option

Um die Funktion ψ und den genauen Wert der Option für $x > \psi(t)$ zu bestimmen, muß die Differentialgleichung (72.7) in dem Gebiet

$$\Omega \,=\, \{\,(x,t)\,:\, x \geq \psi(t)\,,\; 0 \leq t < T\,\}$$

gelöst werden, vgl. Abbildung 72.2. Für $t = T$ gilt die gleiche Endbedingung (72.3) wie für den europäischen Put. Auch die Randbedingung (72.9) im Unendlichen gilt wie zuvor. Für den linken Rand, $x = \psi(t)$, haben wir offensichtlich die Randbedingung

$$u(\psi(t),t) \,=\, \max\{X - x, 0\}\,, \qquad 0 \leq t < T\,.$$

Allerdings handelt es sich hier um ein freies Randwertproblem, da die Randkurve a priori nicht bekannt ist. Um diesen Rand festzulegen, wird noch eine zweite Bedingung benötigt. Diese Bedingung,

$$u_x(\psi(t),t) \,=\, -1\,, \qquad 0 \leq t < T\,, \tag{72.14}$$

ist nicht so offensichtlich. Sie ergibt sich aus einer Untersuchung der Gewinnerwartung g aus (72.6) für den Fall, daß der Aktienkurs $x_n = x(t_n) = \psi(t_n)$ auf der Randkurve liegt, vgl. Aufgabe 12. Bei einem fairen Optionspreis darf der Nettoerlös g weder bei einem Kursgewinn noch einem Kursverlust der Aktie zu einem kurzfristigen Gewinn führen.

Abbildung 72.3 zeigt den Wert einer amerikanischen Option zu einem Zeitpunkt $t < T$ (die dickere Kurve) zusammen mit der unteren Schranke aus (72.13) (die dünne Gerade). Die gebrochene Linie kennzeichnet den Aktienkurs $x = \psi(t)$, ab dem die beiden Kurven voneinander abweichen. Zum Vergleich enthält die Abbildung noch den Preis der entsprechenden europäischen

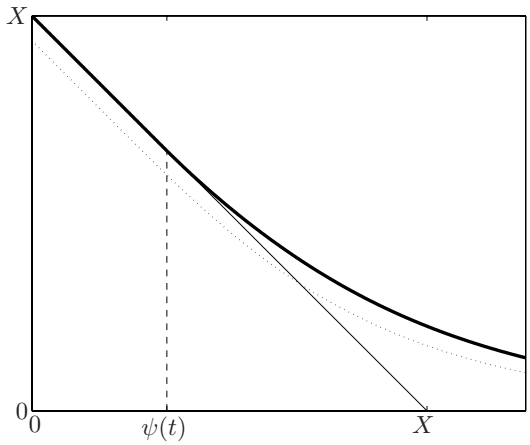


Abb. 72.3: Preis einer amerikanischen Option zu einem festen Zeitpunkt $t < T$ in Abhängigkeit vom Aktienkurs x

Option (gepunktete Kurve). Wie man sieht, unterschreitet der Preis der europäischen Option die Unterschranke (72.13) in einem Intervall $[0, \varphi(t))$, das deutlich größer ist als das Intervall $[0, \psi(t)]$.

Aufgaben

1. Sei u° eine beschränkte Teilchendichte in dem kompakten Intervall $[a,b]$ der reellen Achse, d.h. $|u^\circ(x)| \le C$ für $x \in [a,b]$ und $u^\circ(x) = 0$ für $x \notin [a,b]$. Diffundieren diese Teilchen ab dem Zeitpunkt $t = 0$ auseinander, so folgt aus der Darstellung (69.1), daß zu *jedem* Zeitpunkt $t > 0$ die Teilchendichte $u(x,t)$ an *jedem* Punkt $x \in \mathbb{R}$ positiv ist – die Ausbreitungsgeschwindigkeit der Teilchen ist unendlich groß. Insbesondere breitet sich also Wärme unendlich schnell aus, was im Widerspruch zu unserer physikalischen Vorstellung zu sein scheint.
Tatsächlich ist die Situation etwas komplizierter: Zeigen Sie, daß zu jedem $\varepsilon > 0$ ein $R_\varepsilon > 0$ existiert, so daß u außerhalb des parabelförmigen Gebiets $a - R_\varepsilon\sqrt{t} \le x \le b + R_\varepsilon\sqrt{t}$ durch ε nach oben beschränkt ist.

2. Die Funktion u löse die Differentialgleichung

$$u_t \;=\; -au_x + \sigma u_{xx}\,.$$

(a) Rechnen Sie nach, daß dann die Funktion

$$v(x,t) \;=\; \exp\Big(\frac{a^2}{4\sigma}t - \frac{a}{2\sqrt{\sigma}}x\Big)\, u(\sqrt{\sigma}x, t)$$

eine Lösung der Wärmeleitungsgleichung $v_t = v_{xx}$ ist.
(b) Schließen Sie hieraus, daß die Lösung u aus Beispiel 69.1 die Darstellung

$$u(x,t) \;=\; \frac{1}{2}\,\Big(\,\mathrm{erf}\Big(\frac{0.3 + at - x}{2\sqrt{\varepsilon t}}\Big) \;-\; \mathrm{erf}\Big(\frac{at - x}{2\sqrt{\varepsilon t}}\Big)\,\Big)$$

besitzt. Dabei bezeichnet erf die *Fehlerfunktion* (engl.: *error function*)

$$\mathrm{erf}(x) \;=\; \frac{2}{\sqrt{\pi}} \int_0^x e^{-\xi^2}\, d\xi\,, \qquad x \in \mathbb{R}\,.$$

3. Schreiben Sie zur Lösung des Anfangsrandwertproblems

$$\begin{aligned} u_t - u_{xx} &= f(x,t)\,, \qquad 0 < x < \pi\,,\ t > 0\,,\\ u(x,0) &= u^\circ(x)\,, \quad u(0,t) = u(\pi,t) = 0\,, \end{aligned}$$

ein Programm auf der Grundlage der schnellen Sinustransformation. Interpolieren Sie hierzu f und u_0 durch Sinuspolynome wie in Abschnitt 55 und bestimmen Sie die entsprechende Näherungslösung wie in Beispiel 69.2. Visualisieren Sie das Ergebnis ebenfalls mit Hilfe der schnellen Sinustransformation.

4. Die sogenannte *Boltzmann-Transformation* $u(x,t) = y(x/\sqrt{t})$ kann unter Umständen verwendet werden, um die Wärmeleitungsgleichung auf einem halbunendlichen Intervall auf eine gewöhnliche Differentialgleichung für y zu reduzieren. Leiten Sie auf diese Weise her, daß das Anfangsrandwertproblem

$$\begin{aligned} u_t &= u_{xx}\,, \qquad x > 0\,,\ t > 0\,,\\ u(x,0) &= 0\,, \quad u(0,t) = \omega\,, \end{aligned}$$

mit der Randbedingung im Unendlichen $u(x,t) \to 0$ für $x \to \infty$ die Lösung

$$u(x,t) \;=\; \omega\big(1 - \operatorname{erf}(x/\sqrt{4t})\big)$$

hat. (erf bezeichnet die Fehlerfunktion aus Aufgabe 2.)

5. Um Eis mit einer Temperatur von 0 °C zu schmelzen, muß zusätzliche Energie (*latente Energie*) zugeführt werden, um die zwischenmolekularen Bindungen aufzubrechen. Die latente Energie ist proportional zu der Masse, die geschmolzen werden soll. Umgekehrt wird dieselbe Energiemenge frei, wenn das Wasser gefriert.
Betrachten Sie ein eindimensionales, in eine Richtung unendlich ausgedehntes Wasserreservoir (längs der positiven x-Achse), in dem die Wassertemperatur konstant bei 0 °C liegt und das am linken Ende (bei $x = 0$) ab dem Zeitpunkt $t = 0$ auf die konstante Temperatur -1 °C abgekühlt wird. Das Wasser beginnt nun von links nach rechts zu gefrieren: Zum Zeitpunkt $t > 0$ erstreckt sich der zugefrorene Bereich etwa über das Intervall $[0, \psi(t)]$, am rechten Ende ($x = \psi(t)$) liegt die Temperatur genau bei 0 °C. Der freie Rand, d.h. das Intervallende $\psi(t)$ als Funktion der Zeit, ist gesucht.
(a) Stellen Sie die Energiebilanz für den Fall auf, daß das rechte Intervallende ψ im Zeitintervall dt um $d\psi$ größer wird. Zeigen Sie, daß sich im Grenzübergang $dt \to 0$ die *Stefan-Randbedingung*

$$\sigma \frac{\partial}{\partial x} u(\psi(t), t) \;=\; \lambda \rho \psi'(t)$$

ergibt; λ (Einheit: Joule/kg) bezeichnet die spezifische latente Energie und ρ die Dichte des Wassers.
(b) Für die Wärmeleitungsgleichung zur Berechnung der Temperatur im Eis liegen nun drei Randbedingungen vor (welche?). Auf diese Weise ist es möglich, neben der Temperaturverteilung den freien Rand $x = \psi(t)$ zu berechnen. Berechnen sie die Funktion ψ für die Parameter $\sigma = \lambda = \rho = 1$ mit Hilfe der Boltzmann-Transformation (vgl. Aufgabe 4).

6. Begründen Sie mit dem Drude-Modell das Fouriersche Wärmeleitungsgesetz (69.5) in einem metallischen Körper. Gehen Sie davon aus, daß in Metallen die Energie hauptsächlich durch Elektronen transportiert wird und die Energie eines Elektrons proportional zu der Temperatur am Ort der jeweils letzten Kollision ist. ($\mathcal{E}$ sei die zugehörige Proportionalitätskonstante.) Zeigen Sie, daß im Rahmen dieses Modells und in Abwesenheit eines elektrischen Felds der Wärmeleitfähigkeitskoeffizient durch

$$\sigma \;=\; v \rho \overline{x}\, \mathcal{E}$$

gegeben ist, wobei v der mittlere Geschwindigkeitsbetrag aller Elektronen, ρ die Elektronendichte und $\overline{x}$ die mittlere Weglänge ist, die ein Elektron zwischen zwei Kollisionen zurücklegt.

7. Stellen Sie einen Bezug zwischen elektrischem Strom in einem homogenen Leiter und einer Potentialströmung (im Sinn der Strömungsmechanik, siehe Abschnitt 67) her. Visualisieren Sie die Stromlinien für das Beispiel 70.1 (vgl. Aufgabe XII.7).

8. Betrachten Sie das folgende Anfangsrandwertproblem für die eindimensionale Richards-Gleichung:

$$\begin{aligned} w_t \;&=\; \big(a(w) w_x\big)_x + \gamma w_x \,, \\ w(0,t) = 1 \,, \quad & w(x,0) = 0 \,, \end{aligned} \qquad x > 0 \,,\; t > 0 \,.$$

Hier sei $\gamma \in \mathbb{R}$ und der Diffusionskoeffizient $a = a(w)$ sei positiv für positive w und habe eine Nullstelle in $w = 0$.
Die Lösung w dieser Gleichung entwickelt wegen des degenerierenden Diffusionskoeffizienten typischerweise einen freien Rand $x = \psi(t)$ (wie die Barenblatt-Lösung (70.14)) mit

$$\begin{aligned} w(x,t) &> 0\,, \qquad 0 \le x < \psi(t)\,, \\ w(x,t) &= 0\,, \qquad x \ge \psi(t)\,. \end{aligned}$$

(a) Leiten Sie formal für den freien Rand die Differentialgleichung

$$\psi' \;=\; -a'(0) w_x - \gamma\,, \qquad \psi(0) = 0\,,$$

her, wobei w_x für dessen linksseitigen Grenzwert an der Stelle $(\psi(t), t)$ steht.
(b) Betrachten Sie den Spezialfall ohne Einfluß der Gravitation ($\gamma = 0$). Überführen Sie die Diffusionsgleichung mit Hilfe der Boltzmann-Transformation (vgl. Aufgabe 4) in eine gewöhnliche Differentialgleichung. Wie lautet diese Differentialgleichung und welche Gestalt ergibt sich für den freien Rand?
(c) Läßt sich dies mit der Form des freien Rands bei der Barenblatt-Lösung (70.14) vereinbaren?

9. Die Funktion $f(x) = x$ kann im Intervall $(-\pi, \pi)$ in die Reihe

$$f(x) \;=\; \sum_{k=1}^{\infty} (-1)^{k+1} \frac{2}{\pi (k - 1/2)^2}\, \sin(k - 1/2) x$$

entwickelt werden. Leiten Sie daraus eine Reihendarstellung für die Lösung $v = v(\xi, \eta)$ der Differentialgleichung $\Delta v = 0$ mit den Neumann-Randbedingung (71.11) im Intervall $[-\pi, \pi]^2$ her.
Hinweis: Verwenden Sie einen Separationsansatz für v.

10. In Abschnitt 67.2 wurde die Impulserhaltungsgleichung nur für reibungsfreie Strömungen hergeleitet. Bei *viskosen* Strömungen machen sich Reibungseinflüsse, die von lokalen Geschwindigkeitsunterschieden hervorgerufen werden, durch zusätzliche Oberflächenspannungen bemerkbar. Im einfachsten Fall sogenannter *Newtonscher Fluide* wird dieser zusätzliche Term ähnlich wie in (71.6) durch den Spannungstensor

$$\lambda\, \mathrm{Spur}(\varepsilon) I + 2\mu\varepsilon$$

beschrieben, dabei sind λ und μ Viskositätskoeffizienten und $\varepsilon = (v' + v'^*)/2$ der Deformationsgeschwindigkeitstensor.
Zeigen Sie, daß die Impulserhaltung für Newtonsche Fluide dann durch die *Navier-Stokes-Gleichung*

$$\rho\, v_t \;+\; \rho \frac{\partial v}{\partial x}\, v + \mathrm{grad}\, p - \mu \Delta v - (\lambda + \mu)\, \mathrm{grad}\, \mathrm{div}\, v \;=\; 0\,.$$

ausgedrückt wird.

11. Verifizieren Sie die Black-Scholes-Formel

$$u(x, T - t) \;=\; X e^{-qt} \Phi\Big(\frac{\log(X/x)}{2\sqrt{\sigma t}} - \frac{(q - \sigma)\sqrt{t}}{2\sqrt{\sigma}}\Big) \;-\; x \Phi\Big(\frac{\log(X/x)}{2\sqrt{\sigma t}} - \frac{(q + \sigma)\sqrt{t}}{2\sqrt{\sigma}}\Big)$$

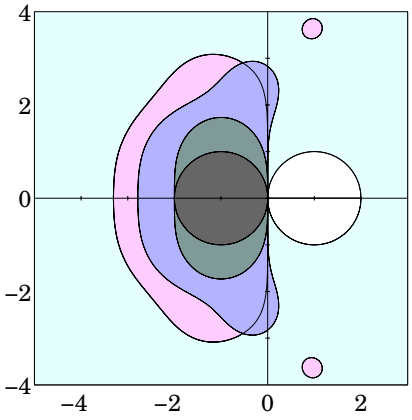
für den fairen Preis einer europäischen Putoption bei Aktienpreis x zum Zeitpunkt $T - t$, $0 \le t \le T$. Hierbei bezeichnet Φ die Verteilungsfunktion der Normalverteilung.
(a) Wenden Sie hierzu die Transformation aus Aufgabe 2 auf die Differentialgleichung (72.11) an.
(b) Bestimmen Sie aus (72.3) die Anfangsvorgabe für die transformierte Differentialgleichung und setzen Sie die Lösungsdarstellung (69.1) an.
(c) Überprüfen Sie, daß das Ergebnis u den korrekten Preis (72.3) zum Zeitpunkt T sowie die beiden Randbedingungen (72.8) und (72.9) erfüllt.

12. Eine Bank verwendet für den Verkauf amerikanischer Putoptionen den in Abschnitt 72 beschriebenen Preis $u = u(x,t)$. An einem Zeitpunkt t, an dem der Aktienkurs x den Wert $x = \psi(t)$ erreicht, kauft ein Anleger bei der Bank eine Option und eine Aktie. Zeigen Sie im Rahmen des Modells aus Abschnitt 72, daß der Anleger dann in einem hinreichend kleinen Zeitintervall τ einen erwarteten Gewinn

$$\mathcal{E}(g) \;=\; \sqrt{\sigma/2}(u_x + 1)x\sqrt{\tau} \;+\; O(\tau)$$

einstreichen kann, sofern $u_x \neq -1$ gilt. Hierbei steht u_x für dessen rechtsseitigen Grenzwert an der Stelle $(\psi(t), t)$. Da es sich dabei in der Tat um einen Gewinn handeln würde (warum?), folgt hieraus die zweite Randbedingung (72.14) für den fairen Optionspreis an der Stelle $x = \psi(t)$.

Gewöhnliche Differentialgleichungen



XIV Anfangswertprobleme

In diesem Kapitel werden numerische Verfahren für Anfangswertprobleme bei gewöhnlichen Differentialgleichungen behandelt. In der Literatur wird zwischen *Einschritt-* und *Mehrschrittverfahren* unterschieden, mit den Runge-Kutta- und den Adams-Verfahren als jeweils bekanntesten Repräsentanten.

Im vorliegenden Text beschränken wir uns auf Einschrittverfahren, gehen dafür aber auf praxisrelevante Themen wie *adaptive Schrittweitensteuerung* und *differential-algebraische Gleichungen* ein und behandeln ausführlich spezielle Fragestellungen für *steife Differentialgleichungen*. Wir hoffen, die Materie in dieser Weise hinreichend gut zu erläutern, um bei Bedarf das Verständnis der Mehrschrittverfahren im Selbststudium zu ermöglichen. Als Literatur empfehlen wir diesbezüglich die beiden Bücher von Hairer, Nørsett und Wanner [44] beziehungsweise Hairer und Wanner [45].

73 Lösungstheorie

Wir betrachten im folgenden Anfangswertprobleme der Form

$$y' = f(t,y)\,, \quad y(0) = y_0\,, \qquad 0 \le t < T \le \infty\,. \tag{73.1}$$

Unter einer Lösung $y = y(t)$ wird eine differenzierbare Funktion über dem Intervall $[0,T)$ verstanden[1], deren Ableitung die Gleichung (73.1) erfüllt.

Dabei braucht y keine skalare Funktion zu sein. Die Beispiele aus Kapitel XI führen beispielsweise fast durchweg auf *Differentialgleichungssysteme*, in denen eine vektorwertige Lösung $y : [0,T) \to \mathbb{R}^d$ gesucht ist und f eine Funktion von $d+1$ Variablen mit Werten in $\mathbb{R}^d$ ist.

Beispiel. Ein Standardbeispiel, das wir bereits aus Abschnitt 60 kennen und das uns noch häufiger als Testgleichung begegnen wird, ist die Differentialgleichung

$$y' = \lambda y\,, \quad y(0) = y_0\,, \qquad y_0, \lambda \in \mathbb{R}\,,$$

[1] Im Randpunkt $t = 0$ wird die Existenz der einseitigen Differentialquotienten gefordert.

mit Lösung

$$y(t) = e^{\lambda t} y_0 \,.$$

Hier ist die Funktion $f(t,y) = \lambda y$ von t unabhängig – man spricht in diesem Fall von einer *autonomen Differentialgleichung*.

Die entsprechende vektorwertige Differentialgleichung lautet

$$y' = Ay \,, \quad y(0) = y_0 \,, \qquad A \in \mathbb{R}^{d\times d},\ y_0 \in \mathbb{R}^d \,. \tag{73.2}$$

Die zugehörige Lösung sieht genauso aus wie zuvor, nämlich

$$y(t) = e^{At} y_0 \,, \tag{73.3}$$

allerdings handelt es sich hierbei zunächst nur um eine formale Schreibweise. Der Ausdruck e^{At} ist nicht komponentenweise zu verstehen, sondern über die Potenzreihenentwicklung der Exponentialfunktion definiert:

$$y(t) = \Big(\sum_{k=0}^{\infty} \frac{1}{k!} A^k t^k \Big) y_0 \,.$$

Man beachte, daß $t \in \mathbb{R}$ und $A \in \mathbb{R}^{d\times d}$, der Ausdruck in der runden Klammer also seinerseits eine reelle $d \times d$ Matrix ist. Die Konvergenz der unendlichen Reihe für jeden möglichen Wert von $t \in \mathbb{R}$ (etwa bezüglich der Spektralnorm $\|\cdot\|_2$) macht man sich wie im Eindimensionalen klar. Ebenso ergibt sich durch gliedweise Differentiation, daß diese unendliche Reihe eine Lösung des Anfangswertproblems darstellt. ◊

Wir wollen im weiteren vektorwertige Funktionen $y : [0,T) \to \mathbb{R}^d$ mit $d \geq 1$ zulassen und fordern, daß die Funktion f in einem Rechteck $\Omega = \mathcal{I} \times \mathcal{J}$ definiert ist mit $\mathcal{I} = [0,T)$ und $\mathcal{J} \subset \mathbb{R}^d$. Dabei kann $\mathcal{J}$ insbesondere ein unbeschränktes Intervall sein. Gemäß (73.1) liefert die rechte Seite $f(t,y)$ für jeden Punkt (t,y) die Steigung der Lösung $y = y(t)$ an dieser Stelle, sofern die Lösungskurve durch diesen Punkt führt. Natürlich ist zunächst nicht bekannt (und hängt vom Anfangswert $y_0 = y(0)$ ab), welche Punkte durchlaufen werden. Dennoch bietet es sich zur Veranschaulichung an, diese Steigungen in Form eines Vektorfelds, des sogenannten *Richtungsfelds*, darzustellen. Abbildung 73.1 zeigt beispielsweise das Richtungsfeld für die *Riccati-Differentialgleichung*

$$y' = y^2 + 1 - t^2 \tag{73.4}$$

in dem Rechteck $\Omega = [0,4) \times [-3,5]$. Anhand der Abbildung kann man bereits erahnen, daß $y(t) = t$ eine Lösung dieser Differentialgleichung mit Anfangswert

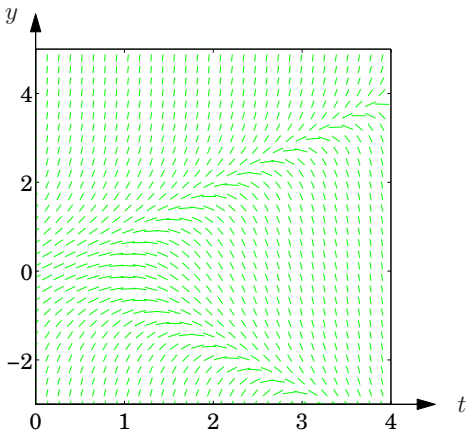


Abb. 73.1:
Richtungsfeld

$y(0) = 0$ ist. Zwei weitere Lösungen mit den Anfangswerten $y_0 = -1$ und $y_0 = 0.1$ sind in Abbildung 73.2 eingezeichnet. Die gebrochene Linie zeigt dabei die vertikale Asymptote $t = t_0$ für die Lösung mit Anfangswert $y_0 = 0.1$ an: diese Lösung divergiert gegen $+\infty$ für $t \to t_0-$.

Grundlegend für die folgenden Überlegungen ist der *Existenzsatz von Picard-Lindelöf*:

Satz 73.1 (Picard-Lindelöf). *f sei stetig in Ω und für alle kompakten Teilmengen $\mathcal{K} \subset \Omega$ gelte eine (lokale) Lipschitz-Bedingung der Form*

$$\|f(t,y) - f(t,z)\|_2 \le L_{\mathcal{K}} \|y - z\|_2 , \qquad (t,y),(t,z) \in \mathcal{K} , \tag{73.5}$$

mit einer positiven Lipschitz-Konstanten $L_{\mathcal{K}}$. Dann existiert für jedes $y_0 \in \mathcal{J}$ ein nichtleeres Teilintervall $\mathcal{I}_0 \subset \mathcal{I}$ mit $0 \in \mathcal{I}_0$ und eine eindeutig bestimmte stetig differenzierbare Lösung $y : \mathcal{I}_0 \to \mathcal{J}$ des Anfangswertproblems (73.1). *Die Lösungskurve $(t, y(t))$ hat zudem eine eindeutig bestimmte Fortsetzung bis an den Rand des Rechtecks Ω.*

Hinreichend für die Gültigkeit der lokalen Lipschitz-Bedingung (73.5) ist etwa, daß f in Ω stetig differenzierbar ist. Dies folgt unmittelbar aus der mehrdimensionalen Verallgemeinerung des Mittelwertsatzes im $\mathbb{R}^n$, wie er bereits im Zusammenhang mit dem Banachschen Fixpunktsatz verwendet wurde, vgl. (17.4). Der Banachsche Fixpunktsatz wird auch zum Beweis des Satzes von Picard-Lindelöf eingesetzt, man vergleiche etwa das Buch von Walter [106, Satz 10.VI].

Ist $\mathcal{I} = [0, T)$ das maximale Intervall, für das f die Voraussetzungen des Satzes 73.1 erfüllt, dann folgt, daß entweder eine eindeutig bestimmte Lösung der Differentialgleichung im gesamten Intervall $[0, T)$ existiert oder daß die Lösung für $t \to t_0- \in (0, T)$ gegen $\partial\mathcal{J}$ konvergiert. Ist $\mathcal{J} = \mathbb{R}^d$ und die Funktion f

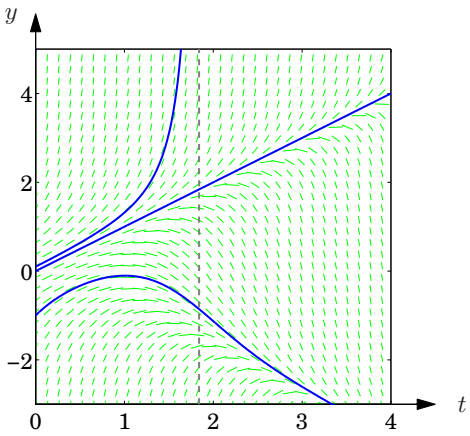


Abb. 73.2:
Richtungsfeld mit Lösungen

gleichmäßig beschränkt in $\mathcal{I} \times \mathcal{J}$, dann macht man sich leicht klar, daß die Lösung y in dem gesamten Intervall wohldefiniert ist.

Beispiele. Beide Fälle treten bei der Riccati-Differentialgleichung (73.4) auf, vgl. Abbildung 73.2. Die Lösung $y(t) = t$ existiert auf dem gesamten Intervall $[0, T)$, während die Lösung mit Anfangswert $y_0 = 0.1$ vorzeitig den oberen Rand von $\Omega = [0, 4) \times [-3, 5]$ erreicht und schließlich für ein endliches t_0 gegen $+\infty$ divergiert.

Ähnlich ist die Situation in dem Beispiel der auf die Erde stürzenden Rakete aus Abschnitt 63.1: Durch die übliche Transformation einer Differentialgleichung höherer Ordnung in ein Differentialgleichungssystem erster Ordnung erhält man für (63.2) das System

$$y_1' = y_2\,, \qquad y_2' = -\frac{1}{2y_1^2}\,,$$

dessen rechte Seite die lokale Lipschitz-Bedingung des Satzes im Rechteck $\Omega = \mathbb{R}^+ \times (\mathbb{R}^+ \times \mathbb{R})$ erfüllt. Die Lösung der Differentialgleichung existiert somit in eindeutiger Weise, solange y_1 nicht Null wird, also solange in diesem Beispiel die Rakete nicht auf die Erde stürzt. Bei der Anfangsvorgabe $y_1(0) = 1$, $y_2(0) = -1$, ist das für $t = 2/3$ der Fall. In diesem Moment erreicht die Lösungskurve den Rand des Rechtecks Ω und dies ist das größtmögliche Existenzintervall der Lösung. Für $y_1(0) = 1$, $y_2(0) = 1$, existiert hingegen eine eindeutige Lösung im gesamten Zeitintervall $[0, \infty)$. ◊

Zum Abschluß dieses Abschnitts beschäftigen wir uns noch mit der stetigen Abhängigkeit der Lösung von der Funktion f und dem Anfangswert y_0.

Satz 73.2. *Die Funktion f erfülle die Voraussetzungen des Satzes* 73.1 *und y sei die Lösung von* (73.1) *im kompakten Intervall $\mathcal{I}_0 \subset [0, T)$ mit Werten*

im Innern des kompakten Intervalls $\mathcal{J}_0 \subset \mathbb{R}^d$. Ferner konvergiere die Funktionenfolge $f_n : \mathcal{I}_0 \times \mathcal{J}_0 \to \mathbb{R}^d$ gleichmäßig gegen f in $\mathcal{I}_0 \times \mathcal{J}_0$ und die reelle Zahlenfolge $\{y_{0n}\} \subset \mathcal{J}_0$ konvergiere für $n \to \infty$ gegen y_0 im Innern von $\mathcal{J}_0$. Schließlich bezeichne y_n, $n \in \mathbb{N}$, eine Lösung des Anfangswertproblems

$$y_n' = f_n(t, y_n)\,, \quad y_n(0) = y_{0n}\,, \qquad 0 \le t < T_n\,.$$

Dann gibt es ein $n_0 \in \mathbb{N}$, so daß y_n für $n \ge n_0$ auf $\mathcal{I}_0$ existiert, und $\{y_n\}_{n \ge n_0}$ konvergiert gleichmäßig gegen y auf $\mathcal{I}_0$.

Unter den Voraussetzungen des Satzes von Picard-Lindelöf garantiert dieser Satz also in relativ natürlicher Weise die stetige Abhängigkeit der Lösung y von der rechten Seite f und dem Anfangswert y_0. Für einen Beweis verweisen wir auf [106, Abschnitt 12]. Für manche Anwendungen ist aber bereits das folgende Ergebnis hinreichend, das unter etwas anderen Voraussetzungen lediglich die stetige Abhängigkeit vom Anfangswert y_0 bei fester rechter Seite f garantiert.

Satz 73.3. *f sei stetig und erfülle die sogenannte* einseitige Lipschitz-Bedingung

$$\big(f(t,y) - f(t,z)\big)^*(y-z) \;\le\; l\,\|y-z\|_2^2 \tag{73.6}$$

für alle $(t,y),(t,z) \in \Omega$ und ein $l \in \mathbb{R}$. Ferner seien $y, z : \mathcal{I} \to \mathcal{J}$ Lösungen der Differentialgleichungen $y' = f(t,y)$ und $z' = f(t,z)$ mit Anfangswerten $y_0, z_0 \in \mathcal{J}$. Dann gilt

$$\|y(t) - z(t)\|_2 \le e^{lt}\,\|y_0 - z_0\|_2 \qquad \textit{für alle } t \in \mathcal{I}\,.$$

Beweis. Da der Beweis sehr einfach ist, soll er hier vorgeführt werden. Wir definieren die Funktion $x(t) = \|y(t) - z(t)\|_2^2$ und betrachten ein beliebiges Intervall $(a,b] \subset \mathcal{I}$, in dem $y - z$ keine Nullstelle besitzt, also x positiv ist. In $(a,b]$ ist somit die Funktion $\log x(t)$ wohldefiniert und differenzierbar. Mit

$$\begin{aligned} x'(t) \;&=\; \frac{d}{dt}\,\|y(t) - z(t)\|_2^2 \;=\; 2\,\big(y'(t) - z'(t)\big)^*\big(y(t) - z(t)\big) \\ &=\; 2\,\big(f(t,y(t)) - f(t,z(t))\big)^*\big(y(t) - z(t)\big) \;\le\; 2l\,\|y(t) - z(t)\|_2^2 \\ &=\; 2l\,x(t) \end{aligned}$$

folgt

$$\frac{d}{dt}\,\log x(t) \;=\; \frac{x'(t)}{x(t)} \;\le 2l\;,$$

und aufintegrieren von $a + \varepsilon$ bis b mit hinreichend kleinem $\varepsilon > 0$ ergibt

$$\log x(b) - \log x(a+\varepsilon) \;\le\; 2l\,(b - a - \epsilon)\,,$$

beziehungsweise

$$x(b) \le x(a+\varepsilon)\, e^{2l\,(b-a-\epsilon)}\,. \tag{73.7}$$

Wäre a eine Nullstelle von x, so ergäbe sich nun aus (73.7) durch Grenzübergang $\varepsilon \to 0$ ein Widerspruch zu der Annahme $x(b) \neq 0$. Somit hat x entweder keine Nullstelle in $[0,T)$ oder es gibt einen Punkt $t_0 \in [0,T)$ mit

$$x(t) > 0\,, \quad 0 < t < t_0 \qquad \text{und} \qquad x(t) = 0\,, \quad t_0 \le t < T\,.$$

Für $t \ge t_0$ ist die Behauptung des Satzes offensichtlich richtig und es müssen nur noch die Zeitpunkte $t \in (0,t_0)$ untersucht werden. Für solche t folgt die Aussage jedoch aus (73.7) mit $a = 0$ und $b = t$ nach Grenzübergang $\varepsilon \to 0$. □

Aus Satz 73.3 folgt insbesondere, daß unter der Bedingung (73.6) Lösungen des Anfangswertproblems $y' = f(t,y)$ mit $y(0) = y_0 \in \mathcal{J}$ *eindeutig* bestimmt sind. Über ihre Existenz wird hingegen nichts ausgesagt.

Die Voraussetzungen von Satz 73.3 sind in gewisser Weise sowohl schwächer als auch stärker als die des Satzes von Picard-Lindelöf. Sie sind einerseits schwächer, da aus einer Lipschitz-Bedingung der Form

$$\|f(t,y) - f(t,z)\|_2 \le L\,\|y-z\|_2 \qquad \text{für alle } (t,y),(t,z) \in \Omega$$

sofort die Bedingung (73.6) des Satzes 73.3 mit $l = L$ folgt. Andererseits ist die Voraussetzung von Satz 73.3 stärker als die Voraussetzung von Satz 73.1, da (73.6) *gleichmäßig* für alle Punkte in Ω gefordert wird.

Die Abschwächung gegenüber dem Satz von Picard-Lindelöf hat jedoch den entscheidenden Vorteil, daß negative l in der Abschätzung von Satz 73.3 möglich sind, während L zwangsläufig positiv sein muß. Differentialgleichungen, die einer einseitigen Lipschitz-Bedingung (73.6) mit einem negativen l genügen, nennt man *strikt dissipativ*.

Beispiel. Für die Differentialgleichung $y' = \lambda y$, $\lambda \in \mathbb{R}$, gilt

$$\big(f(t,y) - f(t,z)\big)^*(y-z) \;=\; \lambda\,\|y-z\|_2^2$$

und die Voraussetzung von Satz 73.3 ist mit $l = \lambda$ in ganz $\mathbb{R}^+ \times \mathbb{R}^d$ erfüllt. Für negative Werte von λ werden Unterschiede (oder Fehler) in den Startwerten also mit dem Faktor $e^{\lambda t}$ gedämpft. Die Lösung wird *asymptotisch stabil* genannt, da solche Fehler für $t \to \infty$ gegen Null gehen. Für $\lambda > 0$ werden Fehler in den Startwerten verstärkt, die Lösungen sind instabil. ◊

Satz 73.3 besagt, daß die Zuordnung $y_0 \mapsto y(t)$ mit $t \in \mathcal{I}$ unter den genannten Voraussetzungen stetig ist, genauer Lipschitz-stetig mit Lipschitz-Konstante $\kappa = e^{lt}$. Die Größe κ kann daher als ein Maß für die lokale Fehlerverstärkung des absoluten Datenfehlers angesehen werden. Sie übernimmt die Rolle einer absoluten *Konditionszahl* der Abbildung $y_0 \mapsto y(t)$.

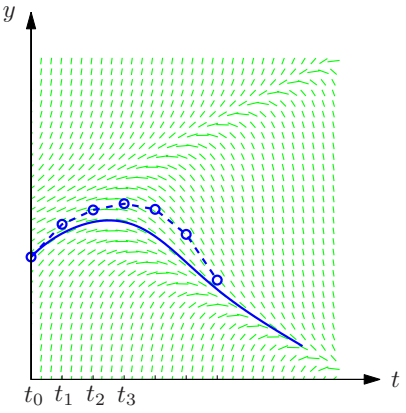


Abb. 74.1:
Euler-Polygonzugverfahren

74 Das Euler-Verfahren

Als erstes numerisches Verfahren zur Lösung von (73.1) betrachten wir das klassische *Euler-Verfahren*, auch *Polygonzugverfahren* genannt. Bei diesem Verfahren wird über einem vorgegebenen Gitter $\triangle = \{0 = t_0 < t_1 < \ldots < t_n\} \subset \mathcal{I}$ ein (vektorwertiger) linearer Spline $y_\triangle \in S^d_{1,\triangle}$ als Approximation an y gewählt, dessen *rechtsseitige Ableitung* in jedem Gitterknoten t_i mit der aus dem Richtungsfeld vorgegebenen Steigung $f(t_i, y_\triangle(t_i))$ übereinstimmt.

Da durch y_0 und $f(0, y_0)$ am linken Rand der Funktionswert und die Anfangssteigung festgelegt sind, lassen sich die Koeffizienten $y_i \in \mathbb{R}^d$ des Splines $y_\triangle(t) = \sum_{i=0}^n y_i \Lambda_i(t)$ in *expliziter* Weise rekursiv von links nach rechts bestimmen:

$$y_{i+1} = y_i + (t_{i+1} - t_i) f(t_i, y_i)\,, \qquad i = 0, 1, \ldots, n-1\,.$$

Beispiel. Für $y' = y$, $y(0) = 1$, mit exakter Lösung $y(t) = e^t$ ergibt das Euler-Verfahren bei einem äquidistanten Gitter ($t_i = ih$)

$$\begin{aligned} y_0 &= 1\,, \\ y_1 &= 1 + h \cdot 1 = 1 + h\,, \\ y_2 &= 1 + h + h(1+h) \;=\; (1+h)^2\,, \\ y_3 &= (1+h)^2 + h(1+h)^2 = (1+h)^3\,. \end{aligned}$$

Durch vollständige Induktion ergibt sich $y_i = (1+h)^i$ für alle $i = 0, \ldots, n$. Wird das Intervall $[0, T]$ in n äquidistante Gitterintervalle unterteilt, folgt hieraus

$$y(T) \;\approx\; y_n \;=\; (1 + T/n)^n\,.$$

Offensichtlich konvergiert y_n tatsächlich für $n \to \infty$ gegen den richtigen Wert $e^T = y(T)$. ◊

Für allgemeine Differentialgleichungen beweisen wir nun eine Fehlerabschätzung für äquidistante Gitter $\triangle$ mit konstanter Gitterweite $h = t_i - t_{i-1}$, $i = 1, \dots, n$.

Satz 74.1. *Sei $\mathcal{I} = [0,T]$ und $f : \mathcal{I} \times \mathbb{R}^d \to \mathbb{R}^d$ stetig differenzierbar und bezüglich y global Lipschitz-stetig,*

$$\|f(t,y) - f(t,z)\|_2 \le L \|y - z\|_2 \qquad \text{für alle } t \in \mathcal{I} \text{ und } y, z \in \mathbb{R}^d .$$

Ist y die eindeutig bestimmte Lösung des Anfangswertproblems (73.1) *und sind y_i, $i = 1, \dots, n$, die Näherungen des Euler-Polygonzugverfahrens an den Gitterpunkten $t_i \in \mathcal{I}$, dann gilt*

$$\|y(t_i) - y_i\|_2 \le \frac{(1+Lh)^i - 1}{2L} \|y''\|_{[0,T]} h \le \frac{e^{LT} - 1}{2L} \|y''\|_{[0,T]} h ,$$

$i = 0, \dots, n$. *Hierbei ist* $\|y''\|_{[0,T]} = \max_{0 \le t \le T} \|y''(t)\|_2$.

Man beachte, daß die Voraussetzungen an f garantieren, daß y zweimal stetig differenzierbar ist: es gilt nämlich

$$y'' = \frac{\partial f}{\partial t} + \frac{\partial f}{\partial y} y' = \frac{\partial f}{\partial t} + \frac{\partial f}{\partial y} f .$$

Beweis von Satz 74.1. Der Beweis ist in drei Schritte gegliedert.

1. Lokaler Fehler: Nehmen wir zunächst an, das Polygonzugverfahren würde zur Zeit t_i mit dem Punkt $\big(t_i, y(t_i)\big)$ auf der exakten Lösungskurve starten und ausgehend von $y(t_i)$ eine Approximation

$$z_{i+1} = y(t_i) + h f(t_i, y(t_i))$$

für $y(t_{i+1})$ berechnen. Dann ergibt sich aus der Differentialgleichung und dem Hauptsatz der Differential- und Integralrechnung der absolute Fehler

$$\begin{aligned} \|y(t_{i+1}) - z_{i+1}\|_2 &= \big\| y(t_{i+1}) - \big(y(t_i) + h f(t_i, y(t_i))\big) \big\|_2 \\ &= \big\| y(t_{i+1}) - y(t_i) - h y'(t_i) \big\|_2 \\ &= \Big\| \int_{t_i}^{t_{i+1}} \big(y'(\tau) - y'(t_i)\big)\, d\tau \Big\|_2 \le \|y''\|_{[0,T]} \int_{t_i}^{t_{i+1}} (\tau - t_i)\, d\tau \\ &= \frac{1}{2} \|y''\|_{[0,T]} h^2 . \end{aligned}$$

2. Lokale Fehlerfortpflanzung: Tatsächlich liegt die Näherung y_i nach i Schritten nicht auf der exakten Lösungskurve. Daher muß noch untersucht werden,

wie der Fehler $y_i - y(t_i)$ im $(i+1)$-ten Schritt fortgepflanzt wird. Aus der Rechenvorschrift des Euler-Verfahrens ergibt sich

$$\begin{aligned} \|y_{i+1} - z_{i+1}\|_2 \;&\le\; \|y_i - y(t_i)\|_2 + h\,\|f(t_i, y_i) - f(t_i, y(t_i))\|_2 \\ &\le\; (1 + hL)\,\|y_i - y(t_i)\|_2\,. \end{aligned} \tag{74.1}$$

3. Kumulierter Fehler: Nun beweisen wir induktiv die erste Teilbehauptung des Satzes, nämlich

$$\|y_i - y(t_i)\|_2 \;\le\; \frac{(1+Lh)^i - 1}{2L}\,\|y''\|_{[0,T]}\,h\,, \qquad i = 0, \ldots, n\,. \tag{74.2}$$

Für $i = 0$ ist diese Behauptung wegen des exakt vorgegebenen Anfangswerts natürlich erfüllt. Aus den ersten beiden Beweisschritten (mit den gleichen Bezeichnungen wie oben) und der Dreiecksungleichung für die $(i+1)$-te Fehlergröße ergibt sich induktiv die Ungleichung

$$\begin{aligned} \|y_{i+1} - y(t_{i+1})\|_2 \;&\le\; \|y_{i+1} - z_{i+1}\|_2 \;+\; \|z_{i+1} - y(t_{i+1})\|_2 \\ &\le\quad (1 + hL)\,\|y_i - y(t_i)\|_2 \;+\; \frac{1}{2}\,\|y''\|_{[0,T]}\,h^2 \qquad (74.3) \\ &\overset{(74.2)}{\le}\quad \frac{1}{2L}\,\big((1+hL)^{i+1} - 1 - hL + hL\big)\,\|y''\|_{[0,T]}\,h \\ &=\quad \frac{(1+hL)^{i+1} - 1}{2L}\,\|y''\|_{[0,T]}\,h\,, \end{aligned}$$

was zu zeigen war. Wegen $1 + hL \le e^{hL}$ und $t_i = ih \in [0, T]$ folgt daraus auch unmittelbar die zweite Behauptung. □

Anhand dieser Fehlerabschätzung erkennt man, daß der Fehler des Euler-Verfahrens linear in h gegen Null geht, falls das Gitter sukzessive verfeinert wird. Eine solche Verfeinerung wird aber in dem Moment nutzlos, in dem zusätzliche Rundungs- oder Rechenfehler die Größenordnung des lokalen Fehlers erreichen. Eine heuristische Überlegung mag das belegen: Angenommen, im $(i+1)$-ten Schritt kommt zu den bereits untersuchten Fehlern (lokaler Fehler und fortgepflanzter Fehler) noch ein additiver Rundungsfehler der Größenordnung ε, etwa der Maschinengenauigkeit, hinzu. Dann erhalten wir anstelle von (74.3) die Ungleichung

$$\|y_{i+1} - y(t_{i+1})\|_2 \;\le\; (1 + hL)\,\|y_i - y(t_i)\|_2 \;+\; \frac{1}{2}\,\|y''\|_{[0,T]}\,h^2 \;+ \varepsilon\,,$$

und induktiv ergibt sich entsprechend

$$\|y_i - y(t_i)\|_2 \le \frac{e^{LT} - 1}{2L}\Big(\,\|y''\|_{[0,T]}\,h + 2\,\frac{\varepsilon}{h}\Big)\,, \quad i = 0, \ldots, n\,. \tag{74.4}$$

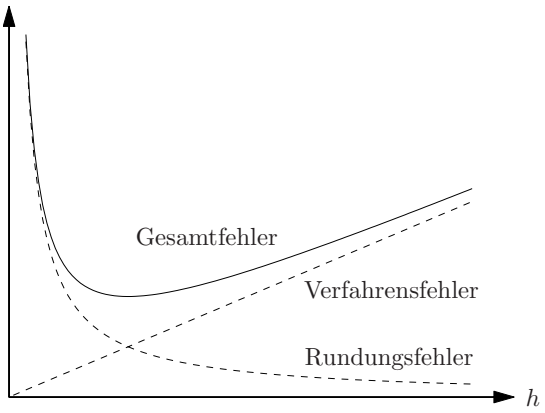


Abb. 74.2: Verfahrensfehler und Rundungsfehler

Mit anderen Worten: Der Gesamtfehler des Euler-Verfahrens setzt sich aus einem (für $h \to 0$ konvergenten) *Verfahrensfehler* und einem (für $h \to 0$ divergenten) *fortgepflanzten Rundungsfehler* zusammen, vgl. Abbildung 74.2. Man sieht leicht, daß die Schranke auf der rechten Seite von (74.4) für $h \sim \sqrt{\varepsilon}$ ihren minimalen Wert von der Größenordnung $O(\sqrt{\varepsilon})$ annimmt.

Wir fassen zusammen:

Bemerkung 74.2. Ist ε der absolute Fehler in einem Schritt des Euler-Verfahrens, so sollte die Schrittweite h nicht kleiner als $\sqrt{\varepsilon}$ gewählt werden. Bei einem Rundungsfehler in der Größenordnung der Maschinengenauigkeit wäre beispielsweise $\varepsilon = \mathsf{eps}$ und $\sqrt{\varepsilon}$ die halbe zur Verfügung stehende Mantissenlänge. ◊

75 Das implizite Euler-Verfahren

Bevor wir ein allgemeines Konstruktionsschema für numerische Algorithmen zur Lösung von Anfangswertaufgaben herleiten, wollen wir noch ein zweites Verfahren angeben, das *implizite Euler-Verfahren*. Bei diesem Verfahren wird die exakte Lösung ebenfalls durch einen linearen Spline approximiert, doch im Unterschied zum expliziten Euler-Verfahren fordert man nun, daß die *linksseitige Ableitung* des Splines in jedem Gitterknoten mit dem Wert von $f(t_i, y_i)$ übereinstimmt. Wie der Name des Verfahrens allerdings bereits andeutet, kann dieser Spline in der Regel nicht mehr explizit berechnet werden. Statt dessen

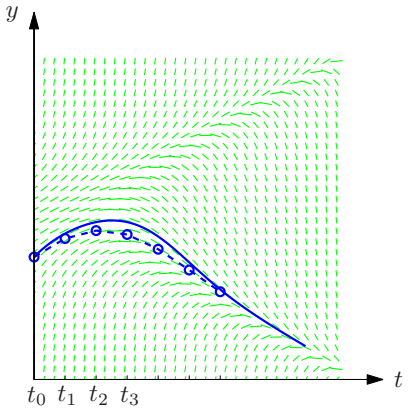


Abb. 75.1:
Implizites Euler-Verfahren

wird y_{i+1} implizit durch die Formel

$$y_{i+1} = y_i + hf(t_{i+1}, y_{i+1}) \tag{75.1}$$

definiert. Für jedes $i = 0, 1, 2, \ldots$ ist daher ein (i.a. nichtlineares) Gleichungssystem zur Berechnung von y_{i+1} zu lösen.

Beispiel. Wir betrachten erneut das Beispiel $y' = \lambda y$, $y(0) = 1$, mit $\lambda < 0$. Die exakte Lösung lautet $y(t) = e^{\lambda t}$. (75.1) ergibt in diesem Fall die Gleichung $y_{i+1} = y_i + h\lambda y_{i+1}$ und daher ist $y_{i+1} = y_i/(1 - h\lambda)$. Wegen $\lambda < 0$ ist y_{i+1} immer wohldefiniert. Induktiv folgt

$$y_i = \Big(\frac{1}{1 - h\lambda}\Big)^i \tag{75.2}$$

und speziell für $T = nh$ ergibt sich

$$y(T) \approx y_n = \big(1 - \frac{T}{n}\lambda\big)^{-n}.$$

Wieder konvergiert diese Näherung für $n \to \infty$ gegen $y(T) = e^{\lambda T}$. ◊

Bevor wir die Konvergenz des impliziten Euler-Verfahrens im allgemeinen Fall untersuchen, diskutieren wir zunächst die Lösbarkeit der nichtlinearen Gleichung (75.1). Wir greifen hierzu auf die einseitige Lipschitz-Bedingung (73.6) zurück.

Satz 75.1. *Sei $\mathcal{I} = [0, T)$ und $f : \mathcal{I} \times \mathbb{R}^d \to \mathbb{R}^d$ stetig differenzierbar und erfülle die einseitige Lipschitz-Bedingung* (73.6) *für ein $l \in \mathbb{R}$. Dann existiert für jedes $y \in \mathbb{R}^d$ und jedes $t \in (0, T)$ eine eindeutig bestimmte Lösung Y der nichtlinearen Gleichung*

$$Y = y + hf(t, Y), \tag{75.3}$$

vorausgesetzt, daß $hl < 1$ ist.

Beweis. Lösungen der Gleichung (75.3) sind offensichtlich Nullstellen der Funktion

$$F(Y) = y + hf(t,Y) - Y\,.$$

Dabei erfüllt die Funktion F ebenfalls eine einseitige Lipschitz-Bedingung:

$$\begin{aligned}\big(F(Y) - F(Z)\big)^*(Y-Z) \;&=\; h\,\big(f(t,Y) - f(t,Z)\big)^*(Y-Z) \;-\; \|Y-Z\|_2^2 \\ &\leq\; -(1-hl)\,\|Y-Z\|_2^2\,.\end{aligned}$$

Da $1 - hl > 0$ ist, folgt hieraus unmittelbar, daß F höchstens eine Nullstelle Y haben kann, also die Eindeutigkeit der Lösung Y von (75.3).

Nullstellen Y der Funktion F ergeben automatisch stationäre (d. h. zeitunabhängige) Lösungen $u \equiv Y$ der Differentialgleichung

$$u' = F(u) \tag{75.4}$$

und umgekehrt. Mit f ist auch F stetig differenzierbar und daher insbesondere lokal Lipschitz-stetig. Also existieren zu jedem $u_0, v_0 \in \mathbb{R}^d$ Lösungen u und v der Differentialgleichung (75.4) mit Anfangswerten $u(0) = u_0$, bzw. $v(0) = v_0$. Ferner gilt nach Satz 73.3 für beliebiges $t_0 > 0$ die Ungleichung

$$\|u(t_0) - v(t_0)\|_2 \;\leq\; e^{-(1-hl)t_0}\,\|u_0 - v_0\|_2\,. \tag{75.5}$$

Aufgrund der Voraussetzung $1 - hl > 0$ ist $q = e^{-(1-hl)t_0}$ kleiner als Eins und die Abbildung $u_0 \mapsto u(t_0)$ folglich eine kontrahierende Selbstabbildung des $\mathbb{R}^d$. Nach dem Banachschen Fixpunktsatz existiert somit ein eindeutiger Fixpunkt dieser Abbildung, den wir mit Y bezeichnen wollen. Die Lösung u mit Startwert $u(0) = Y$ erfüllt also auch $u(t_0) = Y$. Wenn wir zeigen können, daß $u(t) = Y$ für alle $t > 0$ gilt, dann ist offensichtlich $F(Y) = 0$ in (75.4) und Y eine Lösung von (75.3).

Für diesen Nachweis beachten wir zunächst, daß die Differentialgleichung (75.4) autonom ist. Hieraus folgt unmittelbar, daß u t_0-periodisch ist: Mit $v(t) = u(t + t_0)$ gilt nämlich

$$v'(t) = u'(t + t_0) = F\big(u(t + t_0)\big) = F\big(v(t)\big)\,,$$

also sind u und v beides Lösungen von (75.4) mit gleichem Anfangswert $u(0) = v(0) = u(t_0) = Y$. Demnach stimmen u und v überein, d. h. $u(t) = u(t + t_0)$ für alle $t \geq 0$.

Genauso sieht man, daß für festes $t_1 > 0$ die Funktion $v_1(t) = u(t + t_1)$ eine Lösung der Differentialgleichung (75.4) mit Anfangswert $v_1(0) = u(t_1)$ ist. Mit

u ist dann natürlich auch v_1 t_0-periodisch. Also folgt mit (75.5)

$$\begin{aligned}\|u(t_1) - Y\|_2 \;&=\; \|v_1(0) - u(0)\|_2 \;=\; \|v_1(t_0) - u(t_0)\|_2 \\ &\le\; e^{-(1-hl)t_0}\,\|v_1(0) - u(0)\|_2 \;=\; q\,\|u(t_1) - Y\|_2\,.\end{aligned}$$

Da $q < 1$ ist, muß also $u(t_1) = Y$ sein. Da aber $t_1 > 0$ beliebig war, stimmen u und Y überall überein und damit ist Y die gesuchte Lösung von (75.3). □

Die Bedingung $lh < 1$ ist insbesondere dann erfüllt, wenn l negativ ist. Lediglich für positive Werte von l ergeben sich aus Satz 75.1 Einschränkungen an die Schrittweite h. Nun zu dem angekündigten Konvergenzsatz.

Satz 75.2. *Für ein $l \in \mathbb{R}$ gelten die Voraussetzungen von Satz* 75.1 *und die Schrittweite $h > 0$ erfülle die Bedingung $hl < 1$. Dann gilt beim impliziten Euler-Verfahren für alle $t_i = ih \in \mathcal{I}$ die Fehlerabschätzung*

$$\|y(t_i) - y_i\|_2 \;\le\; \frac{1}{2l}\,\Big(\big(\frac{1}{1-lh}\big)^i - 1\Big)\,\|y''\|_{[0,T]}\,h\,. \tag{75.6}$$

Beweis. Der Aufbau des Beweises ist in weiten Strecken derselbe wie im Beweis von Satz 74.1.

1. Lokaler Fehler: Wir betrachten zunächst einen Schritt des impliziten Euler-Verfahrens, ausgehend von einem Punkt $\big(t_i, y(t_i)\big)$ auf der exakten Lösungskurve. Mit dem Satz von Taylor,

$$y(t_i) = y(t_{i+1}) - hy'(t_{i+1}) + r_i\,, \qquad \|r_i\|_2 \le \frac{1}{2}\,\|y''\|_{[0,T]}h^2\,,$$

ergibt sich für die resultierende Approximation z_{i+1} der Fehler

$$\begin{aligned}z_{i+1} - y(t_{i+1}) \;&=\; y(t_i) + hf(t_{i+1}, z_{i+1}) - y(t_{i+1}) \\ &=\; y(t_{i+1}) - hy'(t_{i+1}) + r_i + hf(t_{i+1}, z_{i+1}) - y(t_{i+1}) \\ &=\; h\big(f(t_{i+1}, z_{i+1}) - f(t_{i+1}, y(t_{i+1}))\big) + r_i\,.\end{aligned}$$

Bilden wir auf beiden Seiten der Gleichung das Innenprodukt mit $z_{i+1} - y(t_{i+1})$, so erhalten wir aus der einseitigen Lipschitz-Bedingung die Ungleichung

$$\begin{aligned}\|z_{i+1} - y(t_{i+1})\|_2^2 \;&=\; h\big(z_{i+1} - y(t_{i+1})\big)^*\big(f(t_{i+1}, z_{i+1}) - f(t_{i+1}, y(t_{i+1}))\big) \\ &\quad+\; \big(z_{i+1} - y(t_{i+1})\big)^* r_i \\ &\le\; lh\,\|z_{i+1} - y(t_{i+1})\|_2^2 \;+\; \|r_i\|_2\,\|z_{i+1} - y(t_{i+1})\|_2\end{aligned}$$

und daraus folgt schließlich

$$\|z_{i+1} - y(t_{i+1})\|_2 \;\le\; \frac{1}{1-lh}\,\|r_i\|_2 \;\le\; \frac{1}{2(1-lh)}\,\|y''\|_{[0,T]}\,h^2\,. \tag{75.7}$$

2. Lokale Fehlerfortpflanzung: Ausgehend von y_i bzw. $y(t_i)$ ergeben sich im $(i+1)$-ten Schritt die beiden Näherungen

$$y_{i+1} = y_i + hf(t_{i+1}, y_{i+1}) \qquad \text{und} \qquad z_{i+1} = y(t_i) + hf(t_{i+1}, z_{i+1})$$

für $y(t_{i+1})$ mit

$$y_{i+1} - z_{i+1} = h\big(f(t_{i+1}, y_{i+1}) - f(t_{i+1}, z_{i+1})\big) \,+\, \big(y_i - y(t_i)\big).$$

Bilden wir hier wieder auf beiden Seiten das Innenprodukt mit $y_{i+1} - z_{i+1}$, so folgt wie im ersten Beweisschritt

$$\|y_{i+1} - z_{i+1}\|_2 \,\le\, \frac{1}{1-lh}\,\|y_i - y(t_i)\|_2\,. \tag{75.8}$$

3. Kumulierter Fehler: Für den Gesamtfehler nach $i+1$ Zeitschritten ergibt sich daher beim impliziten Euler-Verfahren die Rekursion

$$\|y_{i+1} - y(t_{i+1})\|_2 \,\le\, \frac{1}{1-lh}\,\|y_i - y(t_i)\|_2 \,+\, \frac{1}{2(1-lh)}\,\|y''\|_{[0,T]}\,h^2\,.$$

Die Behauptung (75.6) folgt nun wieder durch einen einfachen Induktionsschluß. □

Daraus folgt unmittelbar das

Korollar 75.3. *Es gelten die Voraussetzungen von Satz* 75.2 *mit einem* $l < 0$. *Dann gilt für alle* $t_i \in [0,T]$ *die Abschätzung*

$$\|y(t_i) - y_i\|_2 \le \frac{1}{2|l|}\,\|y''\|_{[0,T]}\,h\,.$$

Bei der Implementierung des impliziten Euler-Verfahrens bilden die nichtlinearen Gleichungssysteme das Hauptproblem. Da die naheliegende Fixpunktiteration zur Lösung von (75.1) in der Regel nur für sehr kleine Schrittweiten $h > 0$ konvergiert, vgl. Aufgabe 3, werden in der Praxis meist Newton-artige Verfahren zur Lösung dieser Gleichungssysteme verwendet. Dazu setzt man etwa $y_{i+1}^{(0)} = y_i$ und iteriert für $k = 0, 1, 2, \ldots$

$$\begin{aligned} y_{i+1}^{(k+1)} \;&=\; y_{i+1}^{(k)} - (I - hJ_k)^{-1}\big(y_{i+1}^{(k)} - y_i - hf(t_{i+1}, y_{i+1}^{(k)})\big)\,, \\ J_k \;&=\; f_y(t_{i+1}, y_{i+1}^{(k)})\,. \end{aligned}$$

Wegen der aufwendigen Berechnung der *Jacobi-Matrizen* J_k und der zugehörigen Inversen von $I - hJ_k$ ersetzt man das Newton-Verfahren im allgemeinen

Initialisierung: y_0 und t_0 sowie Schrittweite h seien gegeben
  `for` $i = 0, 1, 2, \ldots$ `do`
    $t_{i+1} = t_i + h$
    $J = f_y(t_{i+1}, y_i)$
    $y_{i+1}^{(0)} = y_i$
    `for` $k = 0, 1, 2, \ldots$ `do` % vereinfachte Newton-Iteration
      löse $(I - hJ)z^{(k)} = y_i + hf(t_{i+1}, y_{i+1}^{(k)}) - y_{i+1}^{(k)}$
      $y_{i+1}^{(k+1)} = y_{i+1}^{(k)} + z^{(k)}$
    `until` stop % end for (k-Schleife)
  `until` $t_{i+1} \geq T$ % end for (i-Schleife)
Ergebnis: $y_i \approx y(t_i), \quad i = 0, 1, 2 \ldots$

Algorithmus 75.1: Implizites Euler-Verfahren

durch das vereinfachte Newton-Verfahren, bei dem in der Jacobi-Matrix immer dieselbe Näherung $y = y_i$ aus dem vorangegangenen Zeitschritt eingesetzt wird. In Programmbibliotheken werden einmal berechnete Jacobi-Matrizen sogar oft über mehrere Zeitschritte hinweg verwendet und adaptiv entschieden, wann sie neu berechnet werden müssen.

Meist sind wenige (ein bis drei) Iterationsschritte ausreichend, um eine Genauigkeit $\|y_{i+1}^{(k)} - y_{i+1}\|_2 \approx h^2$ zu erreichen. Letzteres ist gerade die Größenordnung des lokalen Fehlers, vgl. (75.7), und wie wir im vergangenen Abschnitt in Bemerkung 74.2 festgehalten haben, ist eine höhere Genauigkeit nicht erforderlich. Dies führt auf den Algorithmus 75.1 zur Lösung von (75.1).

Die oben genannte Abbruchbedingung $\|y_{i+1}^{(k)} - y_{i+1}\|_2 \approx h^2$ ist natürlich in dieser Form nicht verwendbar, da y_{i+1} gerade die gesuchte unbekannte Größe ist. In der Praxis muß dieser Fehler geschätzt werden. Hierzu kann man wie in Bemerkung 19.2 vorgehen und erhält dann die Abbruchbedingung

$$\frac{\|z^{(k)}\|_2^2}{\|z^{(k-1)}\|_2 - \|z^{(k)}\|_2} \lesssim h^2 \tag{75.9}$$

mit $z^{(-1)} = y_{i+1}^{(0)}$ für die vereinfachte Newton-Iteration in Algorithmus 75.1.

76 Runge-Kutta-Verfahren

Der entscheidende Nachteil der beiden Euler-Verfahren ist ihre unbefriedigende Genauigkeit. Anhand der Beweise der Sätze 74.1 und 75.2 wird klar, daß

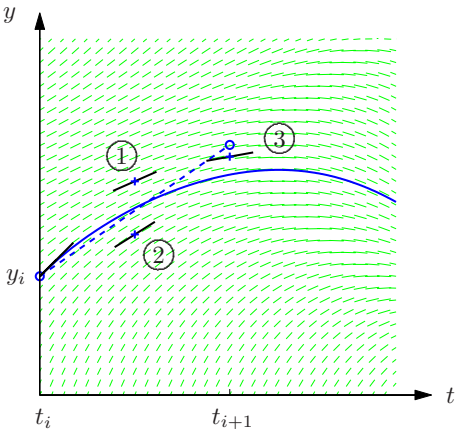


Abb. 76.1:
Runge-Kutta-Ansatz

hierfür allein der lokale Fehler verantwortlich ist, und dieser wiederum ist relativ groß, da die Tangentensteigung an den Randpunkten des Intervalls $[t_i, t_{i+1}]$ die *Sekante* durch die Punkte $(t_i, y(t_i))$ und $(t_{i+1}, y(t_{i+1}))$ auf der Lösungskurve schlecht approximiert. Zur Konvergenzverbesserung kann man einen Ansatz

$$y_{i+1} = y_i + h \sum_{j=1}^{s} b_j f(t_i + c_j h, \eta_j)\,, \qquad \sum_{j=1}^{s} b_j = 1\,, \tag{76.1}$$

wählen (vgl. Abbildung 76.1), mit geeigneten Näherungen η_j (den sogenannten *Stufen*) für y_i beziehungsweise y_{i+1}. Die Parameter c_j sind die *Knoten* des Verfahrens, b_j die *Gewichte*; s ist die *Stufenzahl*. Auch die beiden Euler-Verfahren lassen sich in dieses Schema einordnen: In diesen Fällen ist $s = 1$ und $c_1 = 0$, $\eta_1 = y_i$ (explizites Euler-Verfahren) bzw. $c_1 = 1$, $\eta_1 = y_{i+1}$ (implizites Euler-Verfahren).

Da bei der Rechenvorschrift (76.1) von den alten Näherungswerten y_j, $j \le i$, lediglich y_i eingeht, spricht man bei Verfahren dieser Art von *Einschrittverfahren*. Im Gegensatz dazu verwenden *Mehrschrittverfahren* auch ältere Näherungen y_{i-1}, y_{i-2}, etc. zur Berechnung von y_{i+1}.

Wir nehmen nun an, daß $y_i = y(t_i)$ auf der exakten Lösungskurve liegt und betrachten wie im ersten Schritt des Beweises von Satz 74.1 den lokalen Fehler des Verfahrens (76.1): Aus dem Hauptsatz der Differentialrechnung folgt

$$\begin{aligned} y(t_{i+1}) - y_{i+1} &= y(t_{i+1}) - y(t_i) - h \sum_{j=1}^{s} b_j f(t_i + c_j h, \eta_j) \\ &= \int_{t_i}^{t_{i+1}} y'(t)\,dt \;-\; h \sum_{j=1}^{s} b_j f(t_i + c_j h, \eta_j) \end{aligned}$$

und mit der Differentialgleichung (73.1) ergibt dies

$$y(t_{i+1}) - y_{i+1} \;=\; \int_{t_i}^{t_{i+1}} f(t, y(t))\, dt \;-\; h \sum_{j=1}^{s} b_j f(t_i + c_j h, \eta_j)\,.$$

Der lokale Fehler ist also klein, wenn die Summe $h \sum b_j f(t_i + c_j h, \eta_j)$ eine gute Approximation des entsprechenden Integrals $\int f(t, y(t))\, dt$ ist. Dies motiviert die Verwendung von *Quadraturformeln* zur Wahl geeigneter Koeffizienten b_j, c_j und η_j, $j = 1, \ldots, s$.

Beispiel 76.1. Mit der Mittelpunktformel (36.1) ergibt sich beispielsweise der Ansatz

$$y_{i+1} = y_i + h f(t_i + \tfrac{h}{2}, \eta_1)\,, \tag{76.2}$$

wobei idealerweise $\eta_1 = y(t_i + h/2)$ sein sollte. Allerdings ist dieser Wert nicht bekannt. Mit dem expliziten Euler-Verfahren mit Schrittweite $h/2$ erhält man jedoch eine vernünftige Näherung:

$$\eta_1 \;=\; y(t_i) + \frac{h}{2}\, y'(t_i) \overset{(73.1)}{=} y_i + \frac{h}{2}\, f(t_i, y_i)\,.$$

Dies ist das *Verfahren von Runge* aus dem Jahr 1895. ◊

Beispiel 76.2. Die Trapezformel führt auf

$$y_{i+1} = y_i + \frac{h}{2}\, f(t_i, y_i) + \frac{h}{2}\, f(t_{i+1}, \eta_2)\,,$$

wobei nun $\eta_2 \approx y(t_i + h)$ sein sollte. Geht man wie bei dem Verfahren von Runge vor und ersetzt

$$\eta_2 \;=\; y_i + h y'(t_i) \;=\; y_i + h f(t_i, y_i)\,,$$

dann ergibt sich das *Verfahren von Heun*. ◊

Bei dem Verfahren von Runge (Beispiel 76.1) ergibt eine Taylorentwicklung für hinreichend glattes f

$$y_{i+1} \;=\; y_i \;+\; h\, f(t_i, y_i) \;+\; \frac{h^2}{2}\, f_t(t_i, y_i) \;+\; \frac{h^2}{2} f_y(t_i, y_i) f(t_i, y_i) \;+\; O(h^3)\,.$$

Andererseits folgt aus der Differentialgleichung

$$\begin{aligned} y''(t) \;&=\; \frac{d}{dt} f(t, y(t)) \;=\; f_t(t, y(t)) + f_y(t, y(t)) y'(t) \\ &=\; f_t(t, y(t)) + f_y(t, y(t)) f(t, y(t))\,. \end{aligned}$$

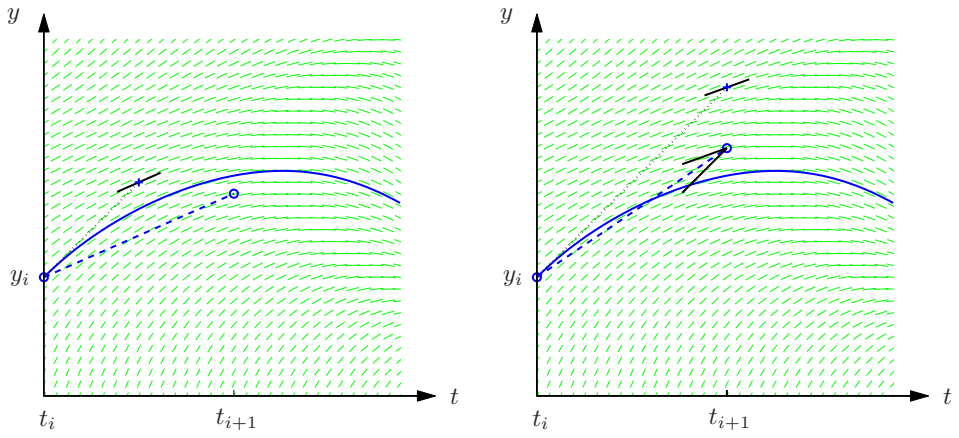


Abb. 76.2: Verfahren von Runge (links) und Heun (rechts)

Unter der Voraussetzung $y_i = y(t_i)$ ist somit

$$\begin{aligned} y(t_{i+1}) &= y(t_i) + h\,y'(t_i) + \frac{h^2}{2}\,y''(t_i) + O(h^3) \\ &= y_i + h\,f(t_i, y_i) + \frac{h^2}{2}\Big(f_t(t_i, y_i) + f_y(t_i, y_i) f(t_i, y_i)\Big) + O(h^3) \end{aligned}$$

und daher gilt $\|y_{i+1} - y(t_{i+1})\|_2 = O(h^3)$. Das Verfahren von Runge hat also einen kleineren lokalen Fehler als die beiden Euler-Verfahren.

Definition 76.3. Ein Einschrittverfahren hat (*Konsistenz-*)*Ordnung* $q \in \mathbb{N}$, falls für jede Lösung $y : \mathcal{I} \to \mathcal{J}$ eines Anfangswertproblems $y' = f(t, y)$ mit $y(0) = y_0 \in \mathcal{J}$ und $f \in C^q(\mathcal{I} \times \mathcal{J})$ eine Konstante c und ein $h_0 > 0$ existiert mit

$$t_i, t_{i+1} \in \mathcal{I},\ y_i = y(t_i) \quad \Longrightarrow \quad \|y_{i+1} - y(t_i + h)\|_2 \le c h^{q+1},\ 0 \le h \le h_0\,.$$

Bemerkung. Die Ordnung wird mit q und nicht mit $q+1$ angegeben, obwohl die entsprechende h-Potenz $q+1$ ist. Wie wir in Satz 76.10 sehen werden, verhält sich nämlich der kumulierte Fehler an einem festen Punkt $t_0 \in (0, T]$ bei einem Verfahren der Ordnung q aufgrund der Fehlerfortpflanzung lediglich wie $O(h^q)$ für $h \to 0$. ◊

Beispiel. Die beiden Euler-Verfahren haben die Ordnung $q = 1$ und das Verfahren von Runge hat die Ordnung $q = 2$. ◊

Der Zusammenhang zwischen einer Quadraturformel und dem zugehörigen Ansatz (76.1) wird durch das folgende Resultat untermauert:

Satz 76.4. *Hat ein Einschrittverfahren der Form* (76.1) *die Ordnung q, dann hat die Quadraturformel*

$$Q[g] \;=\; \sum_{j=1}^{s} b_j g(c_j) \;\approx\; \int_0^1 g(x)\,dx$$

den Exaktheitsgrad $q-1$.

Beweis. Für $0 \le n < q$ betrachten wir das spezielle „Anfangswertproblem“

$$y' = t^n\,, \qquad y(0) = 0\,,$$

mit der eindeutig bestimmten Lösung $y(t) = t^{n+1}/(n+1)$. Mit $y_0 = 0$ ergibt sich aus Definition 76.3 die Abschätzung

$$|y(h) - y_1| \;=\; \Big|\, \frac{1}{n+1}\, h^{n+1} - h \sum_{j=1}^{s} b_j (c_j h)^n \,\Big| \;=\; O(h^{q+1})\,, \qquad h \to 0\,,$$

für ein Einschrittverfahren der Ordnung q. Nach Division durch h^{n+1} erhält man hieraus

$$\Big|\, \frac{1}{n+1} - \sum_{j=1}^{s} b_j c_j^n \,\Big| \;=\; O(h^{q-n}) \;=\; o(1)\,, \qquad h \to 0\,,$$

und durch Grenzübergang $h \to 0$ ergibt sich zwangsläufig für $p_n(t) = t^n$, daß

$$Q[p_n] \;=\; \sum_{j=1}^{s} b_j c_j^n \;=\; \frac{1}{n+1} \;=\; \int_0^1 p_n(t)\,dt\,.$$

Also ist die Quadraturformel $Q[\,\cdot\,]$ für alle Monome t^n, $n = 0, \ldots, q-1$, und damit für den ganzen Unterraum Π_{q-1} exakt. □

Als unmittelbare Folgerung ergibt sich, daß ein s-stufiges Einschrittverfahren maximal die Konsistenzordnung $q = 2s$ haben kann, vgl. Proposition 40.2.

Dieser Zusammenhang zwischen der Ordnung eines Einschrittverfahrens und dem Exaktheitsgrad einer Quadraturformel läßt sich gezielt weiterverfolgen, um Verfahren höherer Ordnung zu konstruieren. Dies ist die Idee der *Runge-Kutta-Verfahren*. Entscheidend ist allerdings auch eine passende Wahl der $\{\eta_j\}$. Wegen

$$\eta_j \;\approx\; y(t_i + c_j h) \;=\; y(t_i) + \int_{t_i}^{t_i + c_j h} y'(t)\,dt \overset{\text{Dgl.}}{=} y(t_i) + \int_{t_i}^{t_i + c_j h} f(t, y(t))\,dt$$

bietet sich hier wieder der Einsatz einer Quadraturformel an. Um zusätzliche Funktionsauswertungen $f(t,y)$ zu vermeiden, beschränkt man sich dabei nach Möglichkeit auf *dieselben* Werte $f(t_i + c_j h, \eta_j)$, $j = 1, \ldots, s$, wie für die Berechnung von y_{i+1}. Dies führt auf den Ansatz

$$\begin{aligned} \eta_j &= y_i + h \sum_{\nu=1}^{s} a_{j\nu} f(t_i + c_\nu h, \eta_\nu)\,, \\ \sum_{\nu=1}^{s} a_{j\nu} &= c_j\,, \end{aligned} \qquad j = 1, \ldots, s\,. \tag{76.3}$$

Wenn $a_{j\nu} = 0$ ist für alle $\nu \geq j$, dann ist diese Rechenvorschrift *explizit* und führt auf ein *explizites Runge-Kutta-Verfahren*, ansonsten ergibt sich ein *implizites Runge-Kutta-Verfahren*.

Üblicherweise werden die Koeffizienten $\{a_{j\nu}, b_j, c_j\}$ in (76.1) und (76.3) in einem quadratischen Tableau zusammengefaßt (dem sogenannten *Runge-Kutta-abc*),

$$\begin{array}{c|c} c & A \\ \hline & b^T \end{array} \quad = \quad \begin{array}{c|ccc} c_1 & a_{11} & a_{12} & \cdots & a_{1s} \\ c_2 & a_{21} & a_{22} & & a_{2s} \\ \vdots & \vdots & & & \vdots \\ c_s & a_{s1} & a_{s2} & \cdots & a_{ss} \\ \hline & b_1 & b_2 & \cdots & b_s \end{array}$$

wobei wir kurzerhand $A = [a_{j\nu}] \in \mathbb{R}^{s\times s}$, $b = [b_1, \ldots, b_s]^T \in \mathbb{R}^s$ und $c = [c_1, \ldots, c_s]^T \in \mathbb{R}^s$ gesetzt haben. Wir sprechen im weiteren kurz von dem Runge-Kutta-Verfahren (A, b, c).

Beispiel 76.5. Für das explizite Euler-Verfahren und für das implizite Euler-Verfahren ergeben sich die Tableaus

$$\begin{array}{c|c} 0 & 0 \\ \hline & 1 \end{array} \qquad \begin{array}{c|c} 1 & 1 \\ \hline & 1 \end{array}$$

Das Verfahren von Runge scheint auf den ersten Blick nicht in das allgemeine Runge-Kutta-Schema hineinzupassen, da zur Berechnung von η_1 auf den Funktionswert $f(t_i, y_i)$ zugegriffen wird, der nicht in der Rechenvorschrift (76.2) vorkommt. Daher behilft man sich mit einem Kunstgriff und führt künstlich $c_0 = 0$ mit $\eta_0 = y_i$ als weitere Stufe ein. Damit wird das Verfahren von Runge zu einem expliziten zweistufigen Runge-Kutta-Verfahren mit Tableau

$$\begin{array}{c|cc} 0 & 0 & 0 \\ 1/2 & 1/2 & 0 \\ \hline & 0 & 1 \end{array}$$

$\Diamond$

Wir wollen nun versuchen, ein Verfahren dritter Ordnung zu konstruieren und leiten uns dafür zunächst Bedingungen an die Parameter her.

Satz 76.6. *Runge-Kutta-Verfahren* (76.1), (76.3), *haben mindestens Konsistenzordnung* $q = 1$. *Ein Runge-Kutta-Verfahren* (A, b, c) *hat genau dann mindestens Konsistenzordnung* $q = 2$, *wenn*

$$\sum_{j=1}^{s} b_j c_j \;=\; \frac{1}{2}\;. \tag{76.4}$$

Es hat genau dann eine Konsistenzordnung von mindestens $q = 3$, *wenn darüber hinaus*

$$\sum_{j=1}^{s} b_j c_j^2 \;=\; \frac{1}{3} \qquad \textit{und} \qquad \sum_{j=1}^{s} b_j \sum_{\nu=1}^{s} a_{j\nu} c_\nu \;=\; \frac{1}{6}\;. \tag{76.5}$$

Bemerkung. Die ersten zwei Gleichungen von (76.4), (76.5) sind dazu äquivalent, daß die Quadraturformel

$$Q[p] \;=\; \sum_{j=1}^{s} b_j p(c_j) \;\approx\; \int_0^1 p(x)\,dx$$

für alle Polynome $p \in \Pi_2$ exakt ist (vgl. Satz 76.4). ◊

Beweis von Satz 76.6. Nach Definition 76.3 können wir annehmen, daß die rechte Seite f der Differentialgleichung hinreichend glatt ist. Nach Aufgabe 7 können wir uns zudem auf autonome Differentialgleichungen der Form $y' = f(y)$ beschränken. Dann folgt durch Taylorentwicklung und unter Berücksichtigung der Differentialgleichung[2]

$$\begin{aligned} y(t_i + h) \;&=\; y(t_i) \;+\; hy'(t_i) \;+\; \tfrac{1}{2}\,h^2 y''(t_i) \;+\; \tfrac{1}{6}\,h^3 y'''(t_i) \;+\; O(h^4) \\ &=\; y \;+\; hf \;+\; \tfrac{1}{2}\,h^2 f_y f \;+\; \tfrac{1}{6}\,h^3 (f^* f_{yy} f + f_y^2 f) \;+\; O(h^4)\,, \end{aligned} \tag{76.6}$$

wobei bei der Funktion f und ihren partiellen Ableitungen sowie bei y immer das Argument $y(t_i)$ bzw. t_i weggelassen wurde.

[2]Zur Vermeidung weiterer Klammern verwenden wir durchweg die formale Schreibweise $u^* f_{yy} v$ für die Auswertung $f_{yy}(u, v) \in \mathbb{R}^d$ des Tensors f_{yy} an gegebenen $u, v \in \mathbb{R}^d$:

$$u^* f_{yy} v \;=\; \Big[\sum_{j,k=1}^{d} \frac{\partial^2 f_i}{\partial y_j \partial y_k}\, u_j v_k \Big]_{i=1}^{d}\,.$$

Zum Vergleich entwickeln wir nun y_{i+1} in Potenzen von h. Dazu ist zunächst zu beachten, daß wegen (76.3)

$$\eta_j - y_i \;=\; h\sum_{\nu=1}^{s} a_{j\nu} f(\eta_\nu) \;=\; h\sum_{\nu=1}^{s} a_{j\nu} f(y_i) + O(h^2) \;=\; hc_j f(y_i) + O(h^2)$$

gilt. Wieder in die rechte Seite von (76.3) eingesetzt, folgt weiterhin durch Taylorentwicklung (das Argument y_i bei f und den partiellen Ableitungen von f wird der Einfachheit halber wieder weggelassen)

$$\begin{aligned}\eta_j \;&=\; y_i + h\sum_{\nu=1}^{s} a_{j\nu} f(\eta_\nu) \;=\; y_i + h\sum_{\nu=1}^{s} a_{j\nu}\big(f + f_y\cdot(\eta_\nu - y_i) + O(h^2)\big)\\ &=\; y_i \;+\; h\, f\sum_{\nu=1}^{s} a_{j\nu} \;+\; h^2 f_y \sum_{\nu=1}^{s} a_{j\nu}c_\nu f \;+\; O(h^3)\,,\end{aligned}$$

also

$$\eta_j \;=\; y_i \;+\; h\, f c_j \;+\; h^2 f_y f \sum_{\nu=1}^{s} a_{j\nu}c_\nu \;+\; O(h^3)\,.$$

Damit ergibt sich schließlich aus (76.1)

$$\begin{aligned} y_{i+1} \;&=\; y_i \;+\; h\sum_{j=1}^{s} b_j f(\eta_j)\\ &=\; y_i \;+\; h\sum_{j=1}^{s} b_j\Big(f + f_y\cdot(\eta_j - y_i) + \tfrac{1}{2}\,(\eta_j - y_i)^* f_{yy}(\eta_j - y_i) + O(h^3)\Big)\\ &=\; y_i \;+\; h\, f\sum_{j=1}^{s} b_j \;+\; h^2 f_y f \sum_{j=1}^{s} b_j c_j \;+\; h^3 f_y^2 f \sum_{j=1}^{s} b_j \sum_{\nu=1}^{s} a_{j\nu}c_\nu\\ &\qquad +\; h^3 f^* f_{yy} f\,\frac{1}{2}\sum_{j=1}^{s} b_j c_j^2 \;+\; O(h^4)\\ &=\; y_i \;+\; h\, f \;+\; h^2 f_y f \sum_{j=1}^{s} b_j c_j \;+\; h^3 f^* f_{yy} f\,\frac{1}{2}\sum_{j=1}^{s} b_j c_j^2\\ &\qquad +\; h^3 f_y^2 f \sum_{j=1}^{s} b_j \sum_{\nu=1}^{s} a_{j\nu}c_\nu \;+\; O(h^4)\,,\end{aligned}$$

wobei wir zuletzt verwendet haben, daß $\sum_{j=1}^{s} b_j = 1$ ist, vgl. (76.1).

Laut Definition 76.3 muß dieses Ergebnis unter der Voraussetzung $y_i = y(t_i)$ mit (76.6) verglichen werden. Demnach ist jedes Runge-Kutta-Verfahren ein Verfahren erster Ordnung; es hat Konsistenzordnung $q = 2$, wenn (76.4) gilt, und es hat Konsistenzordnung $q = 3$, wenn darüber hinaus die beiden Bedingungen (76.5) erfüllt sind. □

Das Verfahren von Runge haben wir bereits als Runge-Kutta-Verfahren zweiter Ordnung identifiziert, d. h. Gleichung (76.4) ist erfüllt:

$$b_1 c_1 + b_2 c_2 \;=\; 0 \cdot 0 \,+\, 1/2 \cdot 1 \;=\; 1/2\,.$$

Die Gleichungen (76.5) sind hingegen beide *nicht* erfüllt.

Beispiel 76.7. Wann immer in der Literatur oder in den Anwendungen von *dem* Runge-Kutta-Verfahren oder dem klassischen Runge-Kutta-Verfahren gesprochen wird, dann ist das folgende explizite Verfahren von *Kutta* (1901) auf der Basis der Simpson-Formel gemeint. Das besondere an diesem Verfahren ist die Verdoppelung des mittleren Knotens $c = 1/2$ bei gleichzeitiger Halbierung des zugehörigen Gewichts $b = 2/3$. Dadurch ergeben sich die vier Stufen

$$c_1 = 0, \quad c_2 = 1/2, \quad c_3 = 1/2, \quad c_4 = 1\,,$$

mit den Gewichten

$$b_1 = 1/6, \quad b_2 = 1/3, \quad b_3 = 1/3, \quad b_4 = 1/6\,.$$

Aufgrund des Exaktheitsgrads $q = 3$ der Simpson-Formel sind die Bedingung (76.4) und die erste, von den Koeffizienten $a_{j\nu}$ unabhängige Bedingung in (76.5) zwangsläufig erfüllt. Berücksichtigt man, daß das Verfahren explizit sein soll, vereinfacht sich die verbleibende Ordnungsbedingung in (76.5) zu

$$a_{32}/6 + a_{42}/12 + a_{43}/12 \;=\; 1/6\,,$$

so daß die Koeffizienten der Matrix A durch diese Gleichungen unterbestimmt sind. Erweitert man Satz 76.6 noch um die Bedingungen für ein Verfahren vierter Ordnung (vgl. Aufgabe 8), dann ist die Matrix A hingegen eindeutig festgelegt und es ergibt sich das Tableau

0				
1/2	1/2			
1/2	0	1/2		
1	0	0	1	
	1/6	1/3	1/3	1/6

Abbildung 76.1 illustriert das Verfahren von Kutta im Richtungsfeld: Zunächst wird mit einem halben (expliziten) Euler-Schritt die mit ① markierte Steigung

$k_2 = f(t_i + h/2, \eta_2)$ bestimmt, ein weiterer halber Euler-Schritt mit dieser Steigung k_2 führt auf die mit ② markierte Steigung $k_3 = f(t_i + h/2, \eta_3)$ und ein ganzer Euler-Schritt mit dieser Steigung definiert $k_4 = f(t_i + h, \eta_4)$. Eine abschließende Mittelung von $k_1 = f(t_i, y_i)$, k_2, k_3 und k_4 ergibt die Steigung der gebrochenen Linie zu y_{i+1}. ◊

Es stellt sich nun zwangsläufig die Frage, welche Ordnung mit einem s-stufigen Verfahren überhaupt erreichbar ist. Wie bereits oben bemerkt wurde, kann diese Ordnung höchstens $2s$ sein. Das Verfahren von Kutta ist trotzdem in der Klasse der vierstufigen Verfahren in gewisser Weise optimal, wie das folgende Resultat zeigt, das wir allerdings erst im nächsten Abschnitt (Seite 583) beweisen können.

Bemerkung 76.8. Ein s-stufiges *explizites* Runge-Kutta-Verfahren hat höchstens die Konsistenzordnung $q = s$. ◊

Beispiel 76.9. Eines der meist verwendeten expliziten Runge-Kutta-Verfahren ist das Verfahren von Dormand und Prince, vgl. [44]. Es ist in vielen Programmbibliotheken unter dem Namen `dopri5` implementiert, MATLAB stellt dieses Verfahren in der Routine `ode45` zur Verfügung. Das Runge-Kutta-Tableau dieses sechsstufigen Verfahrens lautet

0						
$\frac{1}{5}$	$\frac{1}{5}$					
$\frac{3}{10}$	$\frac{3}{40}$	$\frac{9}{40}$				
$\frac{4}{5}$	$\frac{44}{45}$	$-\frac{56}{15}$	$\frac{32}{9}$			
$\frac{8}{9}$	$\frac{19372}{6561}$	$-\frac{25360}{2187}$	$\frac{64448}{6561}$	$-\frac{212}{729}$		
1	$\frac{9017}{3168}$	$-\frac{355}{33}$	$\frac{46732}{5247}$	$\frac{49}{176}$	$-\frac{5103}{18656}$	
	$\frac{35}{384}$	0	$\frac{500}{1113}$	$\frac{125}{192}$	$-\frac{2187}{6784}$	$\frac{11}{84}$

◊

Wir kommen nun zu dem bereits angekündigten Satz über die Konvergenz allgemeiner Runge-Kutta-Verfahren.

Satz 76.10. *Sei $\mathcal{I} = [0, T]$, $f \in C^q(\mathcal{I} \times \mathbb{R}^d)$ mit in $\mathcal{I} \times \mathbb{R}^d$ beschränkter partieller Ableitung f_y und das Runge-Kutta-Verfahren habe Konsistenzordnung q. Dann existiert ein $h_0 > 0$, so daß bei Schrittweite $h \in (0, h_0)$ alle Näherungen y_i des Runge-Kutta-Verfahrens für $t_i = ih \in \mathcal{I}$ eindeutig definiert sind. Ferner gilt*

$$\|y(t_i) - y_i\|_\infty \le Ch^q\,, \qquad h < h_0\,,$$

wobei die Konstante C von i und h unabhängig ist, solange t_i in $[0, T]$ liegt.

Beweis. Das Grundgerüst des Beweises ist das gleiche wie bei den Konvergenzbeweisen für die beiden Euler-Verfahren. Zusätzliche Schwierigkeiten entstehen allerdings dadurch, daß die Existenz der Zwischenstellen η_j in jedem Zeitschritt sicherzustellen ist.

1. Lokaler Fehler: Die notwendige Abschätzung für den lokalen Fehler entspricht gerade der Ordnungsdefinition des Runge-Kutta-Verfahrens: Beginnt ein Runge-Kutta-Schritt mit einem Punkt $\big(t_i, y(t_i)\big)$ auf der Lösungskurve, dann ergibt sich eine Näherung z_{i+1} für $y(t_{i+1})$ mit

$$\|y(t_{i+1}) - z_{i+1}\|_\infty \;\le\; C_1 h^{q+1}\,, \tag{76.7}$$

wobei C_1 eine geeignete Konstante ist, vgl. Definition 76.3. (Man beachte dabei, daß Euklid- und Maximumnorm im $\mathbb{R}^d$ äquivalente Normen sind.)

2a. Existenz der Zwischenstellen: Wir betrachten nun einen festen Zeitschritt des Verfahrens und fixieren somit t_i und die Schrittweite h. Um die lokale Existenz der Lösung des (ggf. impliziten) Runge-Kutta-Schritts zu klären, führen wir zu beliebigem $u \in \mathbb{R}^d$ die Vektoren

$$\Phi(u,\eta) = \begin{bmatrix} \phi_1(u,\eta) \\ \vdots \\ \phi_s(u,\eta) \end{bmatrix} \in \mathbb{R}^{sd} \qquad \text{mit} \qquad \eta = \begin{bmatrix} \eta_1 \\ \vdots \\ \eta_s \end{bmatrix} \in \mathbb{R}^{sd}$$

ein, wobei die einzelnen Komponenten von Φ durch

$$\phi_j(u,\eta) \;=\; u + h\sum_{\nu=1}^{s} a_{j\nu} f(t_i + c_\nu h, \eta_\nu) \;=:\; u + h\psi_j(\eta)\,,$$

$j = 1, \ldots, s$, definiert sind. Der Runge-Kutta-Schritt ist durchführbar, sofern das Gleichungssystem (76.3) eine Lösung besitzt, also sofern die Fixpunktgleichung

$$\eta = \Phi(u,\eta) \tag{76.8}$$

für $u = y_i$ lösbar ist. Um dies zu klären, zeigen wir zunächst, daß die Abbildung $\Phi(u, \cdot\,)$ für jedes $u \in \mathbb{R}^d$ eine Kontraktion im Raum $\mathbb{R}^{sd}$ ist. Die Ableitung Φ_η von Φ ergibt sich aus den partiellen Ableitungen

$$\frac{\partial \phi_j}{\partial \eta_\nu} \;=\; h\,\frac{\partial \psi_j}{\partial \eta_\nu} \;=\; h a_{j\nu} f_y(t_i + c_\nu h, \eta_\nu)$$

zu

$$\Phi_\eta(u,\eta) \;=\; h \begin{bmatrix} a_{11} f_y(t_i + c_1 h, \eta_1) & \cdots & a_{1s} f_y(t_i + c_s h, \eta_s) \\ \vdots & & \vdots \\ a_{s1} f_y(t_i + c_1 h, \eta_1) & \cdots & a_{ss} f_y(t_i + c_s h, \eta_s) \end{bmatrix}.$$

Ist L eine obere Schranke für $\|f_y\|_\infty$ in $\mathcal{I} \times \mathbb{R}^d$ und $A = [a_{j\nu}]$ die Koeffizientenmatrix des Runge-Kutta-Verfahrens, dann ergibt sich hieraus

$$\|\Phi_\eta(u,\eta)\|_\infty \le hL\|A\|_\infty$$

und somit ist die Funktion $\Phi(u,\cdot)$ für $h < h_0 = 1/(2L\|A\|_\infty)$ eine Kontraktion mit Kontraktionsfaktor $1/2$, vgl. (17.4). Nach dem Banachschen Fixpunktsatz existiert somit für jedes $u \in \mathbb{R}^d$ eine (eindeutig bestimmte) Lösung $\eta(u) = [\eta_j(u)]_j \in \mathbb{R}^{sd}$ der Fixpunktgleichung (76.8). Zudem gilt für beliebige $u, v \in \mathbb{R}^d$ nach dem Mittelwertsatz

$$\begin{aligned}\|\eta_j(u) - \eta_j(v)\|_\infty &= \|\phi_j(u,\eta(u)) - \phi_j(v,\eta(v))\|_\infty \\ &= \|u - v + h\big(\psi_j(\eta(u)) - \psi_j(\eta(v))\big)\|_\infty \\ &\le \|u-v\|_\infty + h\|A\|_\infty L\|\eta(u) - \eta(v)\|_\infty .\end{aligned}$$

Wegen $h < h_0$ folgt hieraus

$$\|\eta(u) - \eta(v)\|_\infty \le \|u-v\|_\infty + \frac{1}{2}\|\eta(u) - \eta(v)\|_\infty$$

beziehungsweise

$$\|\eta(u) - \eta(v)\|_\infty \le 2\|u-v\|_\infty . \tag{76.9}$$

2b. Lokale Fehlerfortpflanzung: Bezeichnen wir nun wieder wie in den Beweisen der Sätze 74.1 und 75.2 mit z_{i+1} und y_{i+1} die Näherungen des $(i+1)$-ten Runge-Kutta-Schritts, falls wir mit $u = y(t_i)$ bzw. $v = y_i$ beginnen. Dann folgt wiederum aus dem Mittelwertsatz

$$\begin{aligned}&\|z_{i+1} - y_{i+1}\|_\infty \\ &\quad= \Big\| y(t_i) - y_i + h\sum_{j=1}^{s} b_j\Big(f\big(t_i + c_j h, \eta_j(y(t_i))\big) - f\big(t_i + c_j h, \eta_j(y_i)\big)\Big)\Big\|_\infty \\ &\quad\le \|y(t_i) - y_i\|_\infty + h\|b\|_1 L\|\eta(y(t_i)) - \eta(y_i)\|_\infty \\ &\quad\le (1 + hC_2)\,\|y(t_i) - y_i\|_\infty\end{aligned}$$

mit $C_2 = 2\|b\|_1 L$, vgl. (76.9).

3. Kumulierter Fehler: Für den Gesamtfehler nach i Zeitschritten ergibt sich daher bei einem Runge-Kutta-Verfahren mit Konsistenzordnung q die Rekursion

$$\|y(t_{i+1}) - y_{i+1}\|_\infty \le (1 + hC_2)\,\|y(t_i) - y_i\|_\infty + C_1 h^{q+1} \tag{76.10}$$

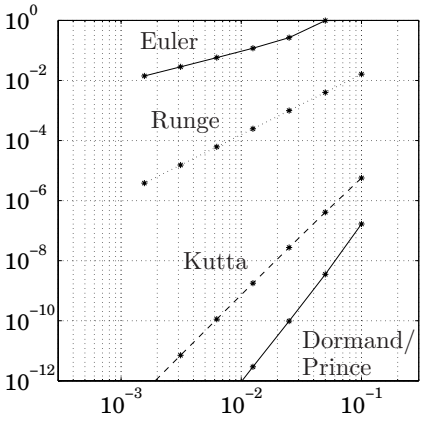


Abb. 76.3:
Relative Fehler der einzelnen Verfahren

und durch Induktion erhält man schließlich wie in dem Beweis von Satz 74.1 für $h = T/n$ und alle $i = 0, \ldots, n$ die Ungleichung

$$\begin{aligned} \|y(t_i) - y_i\|_\infty &\leq \frac{C_1}{C_2}\,(1 + 2C_2 h)^i\, h^q \leq \frac{C_1}{C_2}\,\big(1 + 2C_2 T/n\big)^n\, h^q \\ &\leq \frac{C_1}{C_2}\, e^{2C_2 T}\, h^q\,. \end{aligned}$$

□

Bemerkung. Die Voraussetzungen an f_y lassen sich mit entsprechend höherem technischen Aufwand abschwächen. ◊

Beispiel 76.11. Wir wenden die expliziten Verfahren von Euler, Runge, Kutta sowie das Verfahren von Dormand und Prince auf die Räuber-Beute-Gleichung (60.7)

$$\begin{aligned} x_1' &= x_1\,(\,d_1 - a_1 x_1 - r\,x_2)\,, \qquad x_1(0) = \xi_1\,, \\ x_2' &= x_2(-d_2 + b\,x_1 - a_2 x_2)\,, \qquad x_2(0) = \xi_2\,, \end{aligned}$$

mit den Parametern $d_1 = d_2 = r = 1$, $b = 1.2$ und $a_1 = a_2 = 0$ sowie den Startwerten $\xi_1 = \xi_2 = 2$ an. In Abbildung 76.3 sind die Euklidnormen der zugehörigen relativen Fehler zur Zeit $t = 15$ über der Zeitschrittweite h in einer doppelt logarithmischen Darstellung geplottet. Anhand der gepunkteten Linien kann die Konsistenzordnung abgelesen werden: Bei dem Verfahren von Dormand und Prince beispielsweise bewirkt der Übergang von $h = 10^{-1}$ auf $h = 10^{-2}$ eine Reduktion des Fehlers von ungefähr 10^{-7} auf 10^{-12}, also um fünf Zehnerpotenzen, wie man das bei einem Verfahren der Ordnung $q = 5$ erwartet. ◊

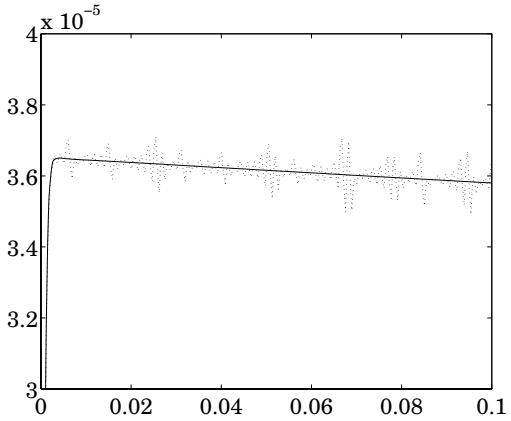


Abb. 77.1: Exakte Lösung und Simulation mit dem Verfahren von Dormand und Prince

77 Stabilitätstheorie

Zur Einleitung dieses Abschnitts berechnen wir mit zwei verschiedenen numerischen Verfahren die Konzentration der Substanz B aus dem chemischen Reaktionsmodell aus Beispiel 62.1. Dabei stellen wir dem expliziten Verfahren von Dormand und Prince mit Konsistenzordnung $q = 5$ ein implizites Verfahren zweiter Ordnung (`ode23s`) gegenüber, das wir in Abschnitt 80 genauer untersuchen werden. Für beide Verfahren verwenden wir die MATLAB-Routinen, die mit einer automatischen Schrittweitensteuerung implementiert sind, ein Thema, auf das wir in Abschnitt 81 noch einmal zurückkommen werden. Für die zugehörigen Steuerungsparameter wählen wir bei beiden Verfahren die gleichen Standardvoreinstellungen.

Abbildung 77.1 zeigt die numerischen Ergebnisse. Die durchgezogene Kurve gibt den exakten Lösungsverlauf an, der mit dieser Achsenskalierung optisch nicht von dem Ergebnis von `ode23s` unterschieden werden kann. Die Punkte in der Abbildung sind die Näherungen, die sich bei dem Verfahren von Dormand und Prince ergeben. Offensichtlich ist das Resultat des Verfahrens von Dormand und Prince trotz der erheblich höheren Konsistenzordnung wesentlich schlechter. Dies bedeutet, daß neben der Konsistenzordnung noch andere Kriterien für eine hohe Genauigkeit wesentlich sind. Ein solches Kriterium ist die Stabilität, die wir in diesem Abschnitt genauer untersuchen.

Die Stabilität von Runge-Kutta-Verfahren haben wir bereits am Ende von Abschnitt 74 angesprochen. Demnach muß neben dem Verfahrensfehler in der Regel noch der fortgepflanzte Rundungsfehler in jedem Runge-Kutta-Schritt berücksichtigt werden. Dies entspricht der Frage, in welcher Weise Approxi-

mationsfehler in y_i im $(i+1)$-ten Schritt propagiert werden:

$$y(t_i) - y_i \qquad \overset{?}{\rightsquigarrow} \qquad y(t_{i+1}) - y_{i+1} \ .$$

Abbildung 74.2 illustriert die pessimistische „worst-case“ Abschätzung für die lokale Fehlerfortpflanzung aus dem zweiten Beweisschritt von Satz 74.1. Ziel dieses Abschnitts ist eine realistischere Untersuchung der lokalen Fehlerfortpflanzung für allgemeine Runge-Kutta-Verfahren. Dazu ist es sinnvoll, das allgemeine Anfangswertproblem (73.1) zunächst auf ein einfaches Modellproblem zu reduzieren. Diese Reduktion kann für Differentialgleichung und Runge-Kutta-Verfahren parallel erfolgen, vgl. Aufgabe 9:

1. *Linearisierung*
Wir interessieren uns für den Einfluß einer (kleinen) Störung u_i des exakten Werts $y(t_i)$ unserer Differentialgleichung: Bezeichnen wir mit $y+u$ die Lösung für den Anfangswert $(y+u)(t_i) = y(t_i) + u_i$, dann gilt – zumindest in einem Zeitintervall, in dem die resultierende Störung u „klein“ bleibt –

$$(y+u)' \ = \ f(t, y+u) \ \approx \ f(t,y) + f_y(t,y)u \,, \qquad f_y(t,y) \in \mathbb{R}^{d\times d} \,.$$

2. *Einfrieren der Zeit*
In einem zweiten Schritt betrachten wir einen kurzen Zeitschritt und vernachlässigen dabei den Einfluß der Zeit in der Jacobi-Matrix $f_y(t,y)$, gehen also davon aus, daß sich lokal der fortgepflanzte Fehler u wie die Lösung der autonomen Differentialgleichung

$$u' = Ju \,, \qquad J = f_y(t_i, y_i) \in \mathbb{R}^{d\times d} \,,$$

mit gleichem Anfangswert $u(t_i) = u_i$ verhält.

3. *Diagonalisierung*
Dieser letzte Schritt beruht auf der Annahme, daß die Matrix J diagonalisierbar ist: Es existiere also eine Basis $\{x_1, \ldots, x_d\} \subset \mathbb{R}^d$ mit $Jx_n = \lambda_n x_n$, $n = 1, \ldots, d$. Entwickelt man für jedes t den Vektor $u(t) \in \mathbb{R}^d$ in dieser Basis, $u(t) = \sum_{n=1}^d \eta_n(t)\, x_n$, dann ergeben sich die Differentialgleichungen

$$\eta_n' = \lambda_n \eta_n \,, \qquad n = 1, \ldots, d \,,$$

für die skalaren Koeffizientenfunktionen η_n. Diese Differentialgleichung ist in den vorigen Abschnitten bereits mehrfach als Beispiel aufgetaucht.

Unter den getroffenen Annahmen ist es ausreichend, die Fortpflanzung kleiner Ausgangsstörungen in diesen entkoppelten Differentialgleichungen zu betrachten. Man mag erwarten, daß für hinreichend kleine $t > 0$ das Resultat

$$u(t) \ = \ \sum_{n=1}^{d} \eta_n(t)\, x_n$$

eine gute Approximation an den fortgepflanzten Fehler des nichtlinearen zeitabhängigen Problems darstellt.

Im Rest dieses Abschnitts betrachten wir daher nur noch die eindimensionale *Testgleichung*

$$y' = \lambda y, \quad y(0) = 1\,, \qquad \lambda \in \mathbb{C}\,. \tag{77.1}$$

Die Lösung $y(t) = e^{\lambda t}$ verhält sich dabei in Abhängigkeit von λ wie folgt:

$$\begin{aligned} &\operatorname{Re}\lambda > 0 : && |y(t)| \to \infty \quad \text{für } t \to \infty\,; \\ &\operatorname{Re}\lambda < 0 : && |y(t)| \to 0 \quad \text{für } t \to \infty\,; \\ &\operatorname{Re}\lambda = 0 : && |y(t)| = 1 \quad \text{für alle } t \in \mathbb{R}_0^+\,. \end{aligned}$$

Eine Grundregel der Numerik lautet, daß eine numerische Lösung möglichst viele Eigenschaften der kontinuierlichen Lösung besitzen sollte. Demnach sind wir vor allem an solchen numerischen Algorithmen für Anfangswertaufgaben interessiert, die die obigen drei Eigenschaften auch für die Näherungslösungen dieser einfachen Testgleichung realisieren.

Definition 77.1. Seien $\{y_i\}$ die Näherungen eines numerischen Verfahrens zur Lösung der Testgleichung (77.1). Dann bezeichnet man das Verfahren als *A-stabil*[3], falls bei beliebigem $\lambda \in \mathbb{C}^- = \{\lambda : \operatorname{Re}\lambda \le 0\}$ die Näherungen mit beliebiger, aber fester Schrittweite $h > 0$ kontraktiv sind, also wenn

$$|y_{i+1}| \le |y_i| \qquad \text{für alle } i \text{ und beliebiges } h > 0\,.$$

Das Verfahren heißt *Isometrie erhaltend*, wenn für beliebiges λ mit $\operatorname{Re}\lambda = 0$ gilt, daß

$$|y_{i+1}| = |y_i| = 1 \qquad \text{für alle } i \text{ und beliebiges } h > 0\,.$$

Definition und Satz 77.2. *Es sei* $\mathbb{1} = [1, \ldots, 1]^T \in \mathbb{R}^s$ *und* $\widehat{\mathbb{C}} = \mathbb{C} \cup \{\infty\}$. *Die zu einem Runge-Kutta-Verfahren* (A, b, c) *gehörige Funktion*

$$R : \widehat{\mathbb{C}} \longrightarrow \widehat{\mathbb{C}}\,, \qquad R(\zeta) = 1 + \zeta b^*(I - \zeta A)^{-1}\mathbb{1}\,,$$

heißt Stabilitätsfunktion. *R ist eine (komplexwertige) rationale Funktion über $\widehat{\mathbb{C}}$, deren Zähler- und Nennergrad höchstens s ist und die allenfalls in den Kehrwerten der Eigenwerte von A Polstellen besitzt; bei einem expliziten Runge-Kutta-Verfahren ist R ein Polynom.*

[3]In der Literatur wird der Begriff der A-Stabilität nicht ganz einheitlich verwendet: Man findet manchmal auch die Forderung, daß $|y_i|$ streng monoton fällt, falls Re $\lambda < 0$ ist. Bei Runge-Kutta-Verfahren sind diese beiden Definitionen äquivalent.

Beweis. Wenn das Runge-Kutta-Verfahren explizit ist, dann ist $A \in \mathbb{R}^{s\times s}$ eine strikte untere Dreiecksmatrix und folglich $A^s = 0$. Man rechnet daher unmittelbar nach, daß $I - \zeta A$ für alle $\zeta \in \mathbb{C}$ invertierbar ist mit

$$(I - \zeta A)^{-1} \;=\; I \,+\, \zeta A \,+\, \ldots \,+\, \zeta^{s-1} A^{s-1}\,. \tag{77.2}$$

Folglich ist R ein Polynom vom Grad s. Um zu sehen, daß R im allgemeinen Fall eine rationale Funktion ist, greifen wir auf die Cramersche Regel zurück, wonach sich die einzelnen Komponenten z_ν von $z = (I - \zeta A)^{-1}\mathbb{1}$ in der Form

$$z_\nu = p_\nu(\zeta)/\det(I - \zeta A) \quad \text{mit} \;\; p_\nu \in \Pi_{s-1}\,, \qquad \nu = 1, \ldots, s\,,$$

schreiben lassen. Die Aussage folgt dann aus der Definition von R. Man beachte, daß R nur dann einen Pol haben kann, wenn einer der Koeffizienten z_ν von z einen Pol besitzt, also nur an den Nullstellen von $\det(I - \zeta A)$. Diese Nullstellen sind aber gerade die Kehrwerte der Eigenwerte von A. □

Die Bedeutung der Stabilitätsfunktion liegt in dem folgenden Resultat:

Proposition 77.3. *Sofern $1/(h\lambda)$ nicht im Spektrum von A liegt, sind alle Näherungen $\{y_i\}$ des Runge-Kutta-Verfahrens (A, b, c) mit Schrittweite h für die Testgleichung* (77.1) *wohldefiniert. In diesem Fall gilt*

$$y_i = \big(R(h\lambda)\big)^i\,, \qquad i = 0, 1, \ldots\,. \tag{77.3}$$

Beweis. Wir definieren die beiden s-dimensionalen Vektoren

$$\eta = \begin{bmatrix}\eta_1\\ \vdots\\ \eta_s\end{bmatrix} \qquad \text{und} \qquad k = \begin{bmatrix}f(t_i + c_1 h, \eta_1)\\ \vdots\\ f(t_i + c_s h, \eta_s)\end{bmatrix} \overset{(77.1)}{=} \lambda\eta\,.$$

Gemäß (76.1) und (76.3) ergibt sich die Näherung y_{i+1} aus y_i durch Lösen des Gleichungssystems

$$\begin{aligned} \eta \;&=\; y_i\mathbb{1} + hAk \;&=\; y_i\mathbb{1} + h\lambda A\eta\,,\\ y_{i+1} \;&=\; y_i + hb^*k \;&=\; y_i + h\lambda b^*\eta\,. \end{aligned}$$

Demnach ist $(I - h\lambda A)\eta = y_i\mathbb{1}$ und

$$y_{i+1} \;=\; y_i + h\lambda b^*(I - h\lambda A)^{-1}(y_i\mathbb{1}) \;=\; R(h\lambda)\, y_i\,, \tag{77.4}$$

sofern $I - h\lambda A$ invertierbar ist, also wenn $\zeta = h\lambda$ nicht der Kehrwert eines Eigenwerts von A ist. Aus (77.4) folgt die behauptete Darstellung (77.3) von y_i durch Induktion. □

Man beachte, daß die Stabilitätsfunktion *nicht* von der Wahl der Knoten $\{c_j\}$ abhängt. Dies ist auch durchaus plausibel, da die verwendete Testgleichung autonom ist, also nicht explizit von der Zeit abhängt.

Beispiel 77.4. In (75.2) haben wir bereits gesehen, daß das implizite Euler-Verfahren die Stabilitätsfunktion $R(\zeta) = 1/(1-\zeta)$ besitzt. Zum expliziten Euler-Verfahren gehört hingegen die Stabilitätsfunktion

$$R(\zeta) = 1 + \zeta(1-0)^{-1}1 = 1+\zeta\,.$$

Für das Runge-Kutta-Verfahren aus Beispiel 76.7 ergibt sich mit Hilfe von (77.2) die Stabilitätsfunktion

$$R(\zeta) \;=\; 1 \,+\, \zeta\;[1/6,1/3,1/3,1/6] \begin{bmatrix} 1 & 0 & 0 & 0 \\ \zeta/2 & 1 & 0 & 0 \\ \zeta^2/4 & \zeta/2 & 1 & 0 \\ \zeta^3/4 & \zeta^2/2 & \zeta & 1 \end{bmatrix} \begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \end{bmatrix}$$

$$= \;1 \,+\, \zeta\;[1/6,1/3,1/3,1/6] \begin{bmatrix} 1 \\ 1+\zeta/2 \\ 1+\zeta/2+\zeta^2/4 \\ 1+\zeta+\zeta^2/2+\zeta^3/4 \end{bmatrix}$$

$$= \;1+\zeta\,(1+\zeta/2+\zeta^2/6+\zeta^3/24) \;=\; 1+\zeta+\tfrac{1}{2}\zeta^2+\tfrac{1}{6}\zeta^3+\tfrac{1}{24}\zeta^4\,. \qquad \lozenge$$

Man erkennt, daß R im letzten Beispiel aus den ersten fünf Summanden der Taylorreihe von e^ζ besteht. Dies hat einen tieferen Grund:

Satz 77.5. *Ist R die Stabilitätsfunktion eines Einschrittverfahrens der Ordnung q, dann gilt*

$$|R(\zeta) - e^\zeta| \;=\; O(|\zeta|^{q+1})\,, \qquad \zeta \to 0\,.$$

Somit stimmen die Taylorpolynome der Exponentialfunktion und der Stabilitätsfunktion um $\zeta = 0$ bis zum Grad q überein.

Beweis. $R(\zeta)$ und e^ζ können beide in eine lokal konvergente Taylorreihe um $\zeta = 0$ entwickelt werden. Ferner gilt nach Definition 76.3 bei Anwendung des Runge-Kutta-Verfahrens auf die Testgleichung (77.1) mit $\lambda = 1$ und Schrittweite $h > 0$:

$$y_1 - y(h) \;=\; y_1 - e^h \;=\; R(h) - e^h \;=\; O(h^{q+1})\,, \qquad h \to 0\,.$$

Daher müssen die Taylorpolynome vom Grad q von R und der Exponentialfunktion übereinstimmen. $\square$

Die Umkehrung dieses Satzes ist offensichtlich falsch, da – wie bereits oben erwähnt – die Stabilitätsfunktion nicht von dem Vektor c abhängt, während c bereits für ein Verfahren zweiter Ordnung spezielle Bedingungen erfüllen muß, vgl. Satz 76.6.

Als einfache Anwendung von Satz 77.5 beweisen wir nun die im vorigen Abschnitt formulierte Bemerkung 76.8.

Beweis von Bemerkung 76.8. Für ein s-stufiges explizites Runge-Kutta-Verfahren ist R ein Polynom vom Grad s, und daher ergibt sich in Satz 77.5 notwendigerweise $q \leq s$. Für $q = s$ muß R das Taylorpolynom der Exponentialfunktion um $\zeta = 0$ vom Grad s sein. □

Mit dem folgenden Satz kommen wir auf die in Definition 77.1 eingeführten Stabilitätseigenschaften eines Runge-Kutta-Verfahrens zurück. Diese Eigenschaften können nämlich unmittelbar an der Stabilitätsfunktion abgelesen werden. Der Beweis folgt unmittelbar aus (77.3).

Proposition 77.6. *Gegeben sei ein Runge-Kutta-Verfahren mit Stabilitätsfunktion R. Dann gilt:*

(a) Das Verfahren ist genau dann A-stabil wenn $|R(\zeta)| \leq 1$ für alle $\zeta \in \mathbb{C}^-$;
(b) Das Verfahren ist genau dann Isometrie erhaltend, wenn $|R(\zeta)| = 1$ für alle ζ mit $\operatorname{Re}\zeta = 0$.

Bemerkungen. Aus diesem Resultat kann man sofort schließen, daß alle expliziten Runge-Kutta-Verfahren weder A-stabil noch Isometrie erhaltend sind, da für Polynome grundsätzlich mit $|\zeta| \to \infty$ auch $|R(\zeta)|$ gegen unendlich strebt.

Das implizite Euler-Verfahren mit Stabilitätsfunktion $R(\zeta) = (1-\zeta)^{-1}$ hingegen ist A-stabil, denn

$$|1-\zeta|^2 \;=\; (1-\operatorname{Re}\zeta)^2 + (\operatorname{Im}\zeta)^2 \;=\; 1 - 2\operatorname{Re}\zeta + |\zeta|^2 \;\geq\; 1$$

für $\operatorname{Re}\zeta \leq 0$. ◊

Da jedoch die Ordnung des impliziten Euler-Verfahrens für die meisten praktischen Anwendungen zu schlecht ist, wenden wir uns in den Abschnitten 78 und 79 der Konstruktion impliziter Runge-Kutta-Verfahren mit höherer Ordnung zu. Zuvor aber noch einige Ergebnisse zu expliziten Verfahren.

Definition 77.7. Mit $\mathcal{S} = \{\zeta \in \mathbb{C} : |R(\zeta)| \leq 1\}$ wird das *Stabilitätsgebiet* eines Runge-Kutta-Verfahrens bezeichnet.

Abbildung 77.2 zeigt die Stabilitätsgebiete der bislang eingeführten Runge-Kutta-Verfahren. Für A-stabile Runge-Kutta-Verfahren wie das implizite Euler-Verfahren ist die gesamte abgeschlossene linke Halbebene $\mathbb{C}^-$ von $\mathbb{C}$ in $\mathcal{S}$

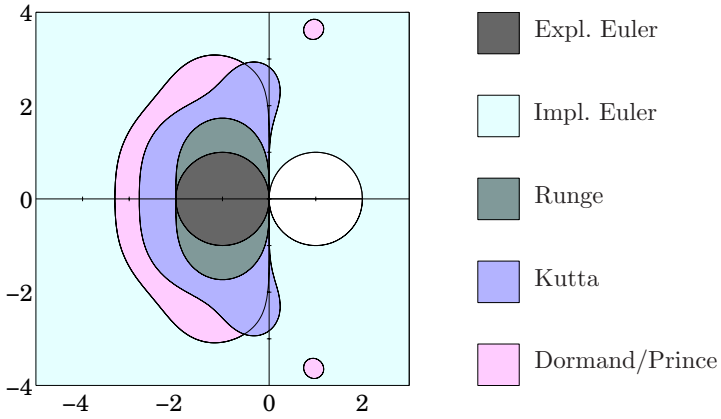


Abb. 77.2: Stabilitätsgebiete

enthalten. Im folgenden sammeln wir einige allgemeine Aussagen über Stabilitätsgebiete.

Lemma 77.8. *Für jedes Runge-Kutta-Verfahren ist* $0 \in \partial\mathcal{S}$.

Beweis. Da jedes Runge-Kutta-Verfahren (mindestens) die Ordnung Eins besitzt, gilt nach Satz 77.5, daß

$$R(\zeta) = 1 + \zeta + O(\zeta^2)\,, \qquad \zeta \to 0\,.$$

Somit ist $R(0) = 1$, also $0 \in \mathcal{S}$, aber es gibt ein ganzes Teilintervall $(0, \varepsilon) \notin \mathcal{S}$, denn

$$|R(\zeta)| > 1 + \zeta/2 > 1 \qquad \text{für } \zeta \in (0, \varepsilon)$$

mit $\varepsilon > 0$ hinreichend klein. □

Proposition 77.9. *Das Stabilitätsgebiet eines expliziten Runge-Kutta-Verfahrens ist immer beschränkt.*

Beweis. Dies folgt sofort aus der Tatsache, daß die Stabilitätsfunktion eines expliziten Verfahrens ein Polynom ist, welches notwendigerweise für $|\zeta| \to \infty$ gegen unendlich strebt. □

Proposition 77.10. *Gegeben sei ein Runge-Kutta-Verfahren, dessen Stabilitätsgebiet* $\mathcal{S}$ *den Halbkreis*

$$\mathcal{B}_\tau^- \;=\; \{\, \zeta \in \mathbb{C}^- : |\zeta| \le \tau \,\}$$

enthält. Dann sind bei jedem $\lambda \in \mathbb{C}^-$ *die Runge-Kutta-Näherungen für die Testgleichung* (77.1) *monoton,* $|y_{i+1}| \le |y_i|$, *sofern eine Schrittweite* $h \le \tau/|\lambda|$ *gewählt wird.*

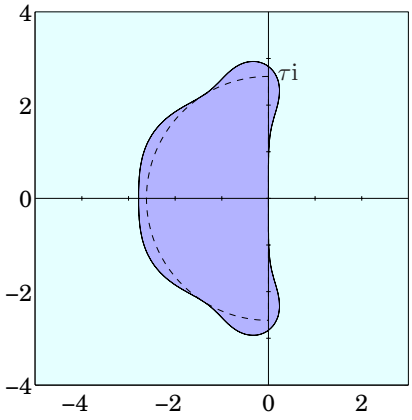


Abb. 77.3:
Illustration von Proposition 77.10

Beweis. Für Re $\lambda \le 0$ und $h \le \tau/|\lambda|$ liegt $\zeta = h\lambda$ in $\mathcal{B}_\tau^-$ und daher folgt die Aussage aus der Rekursion $y_{i+1} = R(h\lambda)y_i$. □

Unter den Voraussetzungen von Proposition 77.10 kann die Stabilität eines (evtl. expliziten) Runge-Kutta-Verfahrens auf Kosten einer gegebenenfalls sehr kleinen Schrittweite $h > 0$ gerettet werden. Die Größe von h ist abhängig von der Größe von $|\lambda|$ aber bemerkenswerterweise *unabhängig* von der Ordnung des Verfahrens.

Zur Illustration sei auf Abbildung 77.3 verwiesen. Sie zeigt das Stabilitätsgebiet des klassischen Runge-Kutta-Verfahrens mit dem größtmöglichen eingeschlossenen Halbkreis B_τ^- (der entsprechende Wert von τ ist $\tau = 2.615\ldots$). Außer dem klassischen Runge-Kutta-Verfahren erfüllt keines der behandelten expliziten Verfahren die Voraussetzung von Proposition 77.10.

Beispiel 77.11. Wir greifen das Beispiel 69.2 der Wärmeleitungsgleichung aus Abschnitt 69 auf,

$$u_t(x,t) = f(u) = u_{xx}(x,t)\,, \qquad u(x,0) = x(\pi - x)\,, \tag{77.5}$$

mit $0 \le x \le \pi$ und $t \ge 0$, dessen Lösung u für alle $t \ge 0$ homogene Randbedingungen $u(0,t) = u(\pi,t) = 0$ haben soll.

Da die rechte Seite $f(u) = u_{xx}$ von (77.5) linear in u und von der Zeit unabhängig ist, entfallen die ersten beiden Reduktionsschritte des vorigen Abschnitts. Für die Diagonalisierung des linearen Operators $Av = v_{xx}$, wobei v ebenfalls die Randbedingungen $v(0) = v(\pi) = 0$ erfüllen soll, greifen wir auf Beispiel 69.2 zurück: Dort haben wir gesehen, daß Eigenwerte und Eigenfunktionen von A durch $\lambda_n = -n^2$ und $v_n(x) = \sin nx$, $n \in \mathbb{N}$, gegeben sind.

Stellen wir uns nun vor, wir wenden ein explizites Runge-Kutta-Verfahren mit Stabilitätsfunktion R zur Lösung der Differentialgleichung (77.5) an und igno-

rieren dabei den zusätzlichen Fehler, der bei der Diskretisierung der Ortsvariablen entsteht. Da das Runge-Kutta-Verfahren nicht die exakte Lösung, sondern nur eine Approximation davon berechnen kann, haben wir bereits nach einem einzigen Zeitschritt in jeden Term der Reihe (69.10) einen Fehler der Größe ε_n eingeschleppt, und nach i weiteren Zeitschritten hat sich dieser Fehler zu

$$R(h\lambda_n)^i\,\varepsilon_n \;=\; R(-hn^2)^i\,\varepsilon_n$$

fortgepflanzt. Da das Stabilitätsgebiet $\mathcal{S}$ eines expliziten Verfahrens immer beschränkt ist, findet sich immer ein $n_0 \in \mathbb{N}$, so daß

$$-hn^2 \notin \mathcal{S} \qquad \text{für } n \geq n_0\,.$$

Für diese n ist $|R(-hn^2)| > 1$ und der fortgepflanzte Fehler explodiert. Wegen der zugehörigen Sinus-Eigenfunktion v_n führt dies zu starken hochoszillierenden Artefakten in der Näherungslösung.

Diese Situation ist besonders unangenehm, da in der exakten Lösung die hochfrequenten Anteile bereits nach kurzer Zeit überhaupt keine Rolle mehr spielen, vgl. (69.10), weil die Vorfaktoren e^{-n^2t} sehr schnell sehr klein werden. Trotzdem bestimmen gerade diese Anteile das Stabilitätsverhalten des Runge-Kutta-Verfahrens.

Für die numerische Lösung der Wärmeleitungsgleichung kommen somit nur A-stabile Runge-Kutta-Verfahren in Frage. ◊

Beispiel 77.12. Abschließend betrachten wir noch einmal das chemische Reaktionsmodell aus Beispiel 62.1. Zu Beginn dieses Abschnitts haben wir festgestellt, daß das explizite Verfahren von Dormand und Prince bei der Lösung der Differentialgleichung (62.2)

$$\begin{aligned} y_1' &= -0.04\,y_1 + 10^4\,y_2y_3\,, \\ y_2' &= 0.04\,y_1 - 10^4y_2y_3 - 3\cdot 10^7 y_2^2\,, \\ y_3' &= 3\cdot 10^7 y_2^2 \end{aligned}$$

größere Schwierigkeiten mit der Berechnung von y_2 hat. Diese Schwierigkeiten treten bei den Funktionswerten

$$y_1 \approx 1\,, \qquad y_2 \approx 3.5\cdot 10^{-5} \qquad \text{und} \qquad y_3 \lesssim 10^{-2}$$

auf.

Bei der Linearisierung um diese Funktionswerte – dem ersten Reduktionsschritt – ergibt sich die Jacobi-Matrix

$$f_y = \begin{bmatrix} -0.04 & 10^4 y_3 & 10^4 y_2 \\ 0.04 & -6\cdot 10^7 y_2 - 10^4 y_3 & -10^4 y_2 \\ 0 & 6\cdot 10^7 y_2 & 0 \end{bmatrix} \approx \begin{bmatrix} -4\cdot 10^{-2} & 10^2 & 3.5\cdot 10^{-1} \\ 4\cdot 10^{-2} & -2.1\cdot 10^3 & -3.5\cdot 10^{-1} \\ 0 & 2.1\cdot 10^3 & 0 \end{bmatrix}.$$

Der zweite Reduktionsschritt (Einfrieren der Zeit) entfällt, da die Zeit in der Differentialgleichung nicht auftritt. Wegen der Massenerhaltung ($y_1+y_2+y_3$ ist konstant) hat die Jacobi-Matrix den Eigenwert $\lambda_1 = 0$ mit linkem Eigenvektor $[1,1,1]^T$. Die (rechten) Eigenvektoren zu den verbleibenden beiden Eigenwerten sind also senkrecht zu $[1,1,1]^T$, d. h. sie sind von der Form $[-\alpha-1, \alpha, 1]^T$. Eine recht einfache Rechnung ergibt dann einen Eigenwert im Intervall $(-1,0)$ und einen Eigenwert in der Größenordnung von -2000.

Aufgrund dieses stark negativen Eigenwerts ist das Verfahren von Dormand und Prince nur für sehr kleine Schrittweiten stabil. Da andererseits der Wert von y_2 sehr viel kleiner ist als die Werte von y_1 und y_3, macht sich eine zu große Schrittweite nicht unmittelbar bemerkbar. Dies bewirkt das Scheitern der automatischen Schrittweitensteuerung bei einer moderat vorgeschriebenen Fehlertoleranz.

Das Problem kann umgangen werden, wenn eine sehr geringe absolute Fehlertoleranz vorgeschrieben wird. Dann wird die exakte Lösung im Intervall $[0, 0.1]$ mit fast der gleichen Schrittzahl von knapp 300 Schritten gut approximiert. Dies belegt, daß nicht der lokale Fehler sondern der fortgeplanzte Datenfehler für das Scheitern des Verfahrens verantwortlich ist. Interessanterweise benötigt das Verfahren `ode23s` für das gleiche Zeitintervall nur 12 (!) Zeitschritte, obwohl `ode23s` nur ein Verfahren zweiter Ordnung ist. Dies liegt daran, daß `ode23s` A-stabil ist und somit ohne Stabilitätsverluste wesentlich größere Schrittweiten wählen kann: In dem nahezu stationären Bereich hinter dem Maximum von y_2 liegt die Schrittweite von `ode23s` relativ konstant im Bereich $h \approx 0.03$. ◊

Differentialgleichungen wie in den beiden obigen Beispielen, bei denen explizite Einschrittverfahren versagen, werden *steife Differentialgleichungen* genannt. Für steife Differentialgleichungen ist oft das Auftreten betragsgroßer negativer Eigenwerte in der Jacobi-Matrix f_y kennzeichnend.

78 Gauß-Verfahren

Aufgrund der Resultate des vorigen Abschnitts konstruieren wir nun implizite A-stabile Runge-Kutta-Verfahren, die eine höhere Konsistenzordnung als das implizite Euler-Verfahren besitzen.

Dabei werden wir sehen, daß implizite Runge-Kutta-Verfahren mit der maximal möglichen Konsistenzordnung $q = 2s$ existieren (zur Erinnerung: explizite Runge-Kutta-Verfahren haben nach Bemerkung 76.8 bestenfalls Konsistenzordnung $q = s$). Nach Satz 76.4 muß bei einem solchen Verfahren die Quadra-

turformel

$$Q[g] \;=\; \sum_{j=1}^{s} b_j g(c_j) \;\approx\; \int_0^1 g(t)\,dt \tag{78.1}$$

den maximal möglichen Exaktheitsgrad $2s-1$ haben. Demnach kann $Q[\,\cdot\,]$ nur die s-stufige Gauß-Legendre-Quadraturformel $G_s[\,\cdot\,]$ sein, vgl. die Abschnitte 40 und 41.

Folglich wählen wir im weiteren für c_j, $j=1,\dots,s$, die Nullstellen des s-ten Legendre-Polynoms (umskaliert auf das Intervall $[0,1]$) und für b_j, $j=1,\dots,s$, die zugehörigen Gewichte der Gauß-Quadraturformel. Die verbleibenden Koeffizienten $a_{j\nu}$, $j,\nu=1,\dots,s$, werden wie in (76.3) so gewählt, daß

$$\begin{aligned} h\sum_{\nu=1}^{s} a_{j\nu} f(t_i+c_\nu h,\eta_\nu) \;&\approx\; \int_{t_i}^{t_i+c_j h} f(t,y(t))\,dt \\ &=\; h\int_0^{c_j} f\big(t_i+\tau h, y(t_i+\tau h)\big)\,d\tau\,, \qquad j=1,\dots,s\,. \end{aligned}$$

Da die η_ν nicht bekannt sind, bietet es sich wieder an, maximale Exaktheit der Quadraturformel

$$\sum_{\nu=1}^{s} a_{j\nu} g(c_\nu) \;\approx\; \int_0^{c_j} g(t)\,dt\,, \qquad j=1,\dots,s\,, \tag{78.2}$$

als Konstruktionsprinzip zu verwenden. Man beachte, daß die zugehörigen Knoten bereits bestimmt sind, lediglich die Wahl der Gewichte ist noch frei. Aus Proposition 38.1 ergibt sich die optimale Wahl dieser Gewichte durch Integration der Lagrange-Grundpolynome:

$$a_{j\nu} = \int_0^{c_j} l_\nu(t)\,dt\,, \qquad l_\nu(t) = \prod_{\substack{i=1\\ i\neq\nu}}^{s} \frac{t-c_i}{c_\nu-c_i}\,. \tag{78.3}$$

Die resultierende Quadraturformel (78.2) hat Exaktheitsgrad $s-1$. Das durch (78.1) und (78.3) bestimmte implizite Runge-Kutta-Verfahren heißt s-stufiges (*Runge-Kutta-*)*Gauß-Verfahren.*

Bemerkung. Im Gegensatz zu Abschnitt 38 befinden sich einzelne Knoten der Quadraturformel (78.2) *außerhalb* des Integrationsintervalls $[0,c_j]$. Alle Eigenschaften der entsprechenden Newton-Cotes-Formeln aus Abschnitt 38 bleiben jedoch auch in diesem Fall gültig. ◊

Beispiel 78.1. Das einfachste Gauß-Verfahren ($s = 1$) führt auf das sogenannte *implizite Mittelpunktverfahren*. Nach Beispiel 41.1 ist die erste Gauß-Legendre-Formel die Mittelpunktformel, d. h. die entsprechenden Runge-Kutta-Parameter lauten $c_1 = 1/2$ und $b_1 = 1$. a_{11} ergibt sich aus (78.3), nämlich $a_{11} = \int_0^{1/2} 1\,dt = 1/2$. Das zugehörige Tableau lautet somit

$$\begin{array}{c|c} 1/2 & 1/2 \\ \hline & 1 \end{array}$$

und das Einschrittverfahren hat die Form

$$y_{i+1} = y_i + hf(t_i + h/2, \eta_1)\,, \qquad \eta_1 = y_i + \frac{h}{2}\,f(t_i + h/2, \eta_1)\,,$$

beziehungsweise nach Kombination dieser beiden Gleichungen,

$$y_{i+1} = y_i + hf\big(t_i + h/2, (y_i + y_{i+1})/2\big)\,. \tag{78.4}$$

Man kann das implizite Mittelpunktverfahren auch als eine Kombination des impliziten und des expliziten Euler-Verfahrens interpretieren. Demnach wird zunächst in einem ersten Halbschritt mit Zeitschrittweite $h/2$ eine Näherung η_1 für $y(t_i + h/2)$ mit dem impliziten Euler-Verfahren berechnet, die dann einem zweiten Halbschritt mit dem expliziten Euler-Verfahren als Ausgangspunkt dient:

$$\eta_1 = y_i + \frac{h}{2}\,f(t_i + h/2, \eta_1)\,, \qquad y_{i+1} = \eta_1 + \frac{h}{2}\,f(t_i + h/2, \eta_1)\,.$$

Das implizite Mittelpunktverfahren wird oft bei zeitabhängigen Diffusionsgleichungen wie der Wärmeleitungsgleichung eingesetzt und läuft in diesem Kontext unter dem Namen *Crank-Nicolson-Verfahren*, vgl. Abschnitt 99.

Zur Ergänzung seien noch die Runge-Kutta-Tableaus der Gauß-Verfahren mit $s = 2$ beziehungsweise $s = 3$ Stufen angeführt:

$$\begin{array}{c|cc} \frac{1}{2} - \frac{\sqrt{3}}{6} & \frac{1}{4} & \frac{1}{4} - \frac{\sqrt{3}}{6} \\ \frac{1}{2} + \frac{\sqrt{3}}{6} & \frac{1}{4} + \frac{\sqrt{3}}{6} & \frac{1}{4} \\ \hline & \frac{1}{2} & \frac{1}{2} \end{array} \qquad \begin{array}{c|ccc} \frac{1}{2} - \frac{\sqrt{15}}{10} & \frac{5}{36} & \frac{2}{9} - \frac{\sqrt{15}}{15} & \frac{5}{36} - \frac{\sqrt{15}}{30} \\ \frac{1}{2} & \frac{5}{36} + \frac{\sqrt{15}}{24} & \frac{2}{9} & \frac{5}{36} - \frac{\sqrt{15}}{24} \\ \frac{1}{2} + \frac{\sqrt{15}}{10} & \frac{5}{36} + \frac{\sqrt{15}}{30} & \frac{2}{9} + \frac{\sqrt{15}}{15} & \frac{5}{36} \\ \hline & \frac{5}{18} & \frac{4}{9} & \frac{5}{18} \end{array}$$

◊

Für die späteren Resultate erweist es sich als nützlich, die Fehlerabschätzung aus Satz 41.4 für den Quadraturfehler auf die hier vorliegende Situation anzupassen.

Lemma 78.2. *Sei $g \in C^{2s}[0,T]$ eine reellwertige Funktion und $[\tau, \tau+h] \subset [0,T]$. Dann gilt für die Gauß-Legendre-Quadraturformel* (78.1) *die Fehlerabschätzung*

$$\Big| h \sum_{j=1}^{s} b_j g(\tau + c_j h) \; - \; \int_{\tau}^{\tau+h} g(t)\, dt \Big| \;\; \le \;\; \kappa_s \, \|g^{(2s)}\|_{\,[0,T]} \, h^{2s+1}$$

mit $\kappa_s = \frac{1}{2s+1} \, (s!)^4 / \big((2s)!\big)^3$.

Beweis. Sei $G(x) = g(\tau + (x+1)h/2)$, $-1 \le x \le 1$. Dann ist nach Satz 41.4

$$\begin{aligned} &\Big| h \sum_{j=1}^{s} b_j g(\tau + c_j h) \; - \; \int_{\tau}^{\tau+h} g(t)\, dt \Big| \\ &\qquad = \; \frac{h}{2} \Big| \sum_{j=1}^{s} 2 b_j G(2c_j - 1) \; - \; \int_{-1}^{1} G(x)\, dx \Big| \\ &\qquad \le \; \varepsilon_s \|G^{(2s)}\|_{\,[-1,1]} \, \frac{h}{2} \, . \end{aligned}$$

Dabei sind $2c_j - 1$ und $2b_j$ die Knoten und Gewichte der klassischen Gauß-Legendre-Formel über dem Intervall $[-1,1]$ und $\varepsilon_s = 2/(2s+1)\, 4^s (s!)^4 / ((2s)!)^3$. Daraus folgt nun unmittelbar die Behauptung, denn nach der Kettenregel ist $G^{(k)}(x) = (h/2)^k g^{(k)}(\tau + (x+1)h/2)$. □

Wir benötigen außerdem noch den folgenden Hilfssatz, in dem Π_s^d den Raum aller Funktionen mit Werten im $\mathbb{R}^d$ bezeichnet, deren einzelne Komponenten Polynome vom Grad s sind.

Lemma 78.3. *Sei $f \in C^{2s}(\mathcal{I} \times \mathcal{J})$ und h so klein, daß die Näherungen y_i des s-stufigen Gauß-Verfahrens wohldefiniert sind* (*vgl. Satz* 76.10). *Dann existiert zu jedem t_i mit $[t_i, t_i + h] \subset \mathcal{I}$ ein Polynom $p \in \Pi_s^d$ mit*

$$\begin{aligned} p(t_i) = y_i \, , \qquad & p(t_i + h) = y_{i+1} \, , \\ p'(t_i + c_j h) = f\big(t_i + c_j h, p(t_i + c_j h)\big) \, , \qquad & j = 1, \dots, s \, . \end{aligned}$$

Das Polynom p heißt *Kollokationspolynom*, denn es ist die Lösung eines sogenannten *Kollokationsverfahrens*, bei dem die Gleichung $p'(t) = f(t, p(t))$ nur an s *isolierten* Punkten $t_i + c_j h$, $j = 1, \dots, s$, erfüllt wird.

Beweis von Lemma 78.3. Wir wählen für p' das (komponentenweise definierte) Interpolationspolynom in Π_{s-1}^d mit $p'(t_i + c_j h) = f(t_i + c_j h, \eta_j)$, $j = 1, \dots, s$, und für $p \in \Pi_s^d$ diejenige Stammfunktion von p' mit $p(t_i) = y_i$. Wegen des

Exaktheitsgrads $q = s - 1$ der Quadraturformeln (78.2) ergibt sich

$$\begin{aligned} \eta_j &= y_i + h \sum_{\nu=1}^{s} a_{j\nu} f(t_i + c_\nu h, \eta_\nu) \;=\; p(t_i) + h \sum_{\nu=1}^{s} a_{j\nu} p'(t_i + c_\nu h) \\ &= p(t_i) + \int_{t_i}^{t_i + c_j h} p'(t)\, dt \;=\; p(t_i + c_j h)\,, \qquad j = 1, \ldots, s\,. \end{aligned}$$

Die s Interpolationsbedingungen für p' können somit umformuliert werden in

$$p'(t_i + c_j h) \;=\; f(t_i + c_j h, \eta_j) \;=\; f\big(t_i + c_j h, p(t_i + c_j h)\big)\,.$$

Damit ist der zweite Teil der Behauptung nachgewiesen. Der verbleibende Teil $p(t_i + h) = y_{i+1}$ folgt entsprechend, da der Exaktheitsgrad $2s - 1$ der Quadraturformel (78.1) größer als $s - 1$ ist:

$$\begin{aligned} y_{i+1} &= y_i + h \sum_{j=1}^{s} b_j f(t_i + c_j h, \eta_j) \;=\; p(t_i) + h \sum_{j=1}^{s} b_j p'(t_i + c_j h) \\ &= p(t_i) + \int_{t_i}^{t_{i+1}} p'(t)\, dt \;=\; p(t_{i+1})\,. \end{aligned}$$

□

Bemerkung. In der Formulierung des Lemmas wird Satz 76.10 angesprochen, nach dem für jedes implizite Runge-Kutta-Verfahren bei hinreichend glatter rechter Seite f alle Näherungen y_i für hinreichend kleines $h > 0$ wohldefiniert sind. Für strikt dissipative Differentialgleichungen, also Differentialgleichungen, bei denen f einer Lipschitz-Bedingung (73.6) mit einem negativen l genügt, läßt sich zeigen, daß die Näherungen der Gauß-Verfahren für *alle* $h > 0$ wohldefiniert sind, vgl. [45, Section IV.14]. Dies entspricht Satz 75.1 für das implizite Euler-Verfahren (welches allerdings kein Gauß-Verfahren ist). ◊

Mit Hilfe von Lemma 78.3 läßt sich nun weiter zeigen, daß das s-stufige Gauß-Verfahren die maximal mögliche Konsistenzordnung $2s$ besitzt. Dazu sei kurz an die Methode der *Variation der Konstanten* erinnert:

Bei einer linearen Differentialgleichung $y' = A(t)y$, $t \geq t_0 \geq 0$, hängt die Lösung y linear vom Anfangswert $y_0 = y(t_0)$ zur Zeit $t = t_0$ ab. Demnach existieren Matrizen $Y(t, t_0) \in \mathbb{R}^{d \times d}$ mit

$$y(t) = Y(t, t_0)\, y(t_0)\,, \qquad t \geq t_0\,,$$

und $Y(t, t) = I$ für alle $t \geq t_0$; für ein konstantes A ergibt sich beispielsweise $Y(t, t_0) = e^{A(t - t_0)}$. Durch Einsetzen in das Anfangswertproblem folgt

$$A(t) Y(t, t_0) y_0 \;=\; A(t)\, y(t) \;=\; y'(t) \;=\; \frac{\partial}{\partial t}\, Y(t, t_0)\, y_0\,, \qquad y_0 = y(t_0)\,.$$

Da dies für alle $y_0 \in \mathbb{R}^d$ richtig ist, ergibt sich

$$A(t)Y(t,t_0) \;=\; \frac{\partial}{\partial t}\,Y(t,t_0)\,. \tag{78.5}$$

Für die *inhomogene* Differentialgleichung

$$y' = A(t)\,y + \varphi(t)\,, \qquad y(0) = y_0\,,$$

mit einer Funktion $\varphi : [0,T] \to \mathbb{R}^d$ ergibt sich damit die Lösungsformel

$$y(t) = Y(t,0)\,y_0 + \int_0^t Y(t,\tau)\varphi(\tau)\,d\tau\,. \tag{78.6}$$

Davon überzeugt man sich durch explizite Differentiation von (78.6):

$$\begin{aligned} y'(t) \;&=\; \frac{\partial}{\partial t}\,Y(t,0)\,y_0 + Y(t,t)\,\varphi(t) + \int_0^t \frac{\partial}{\partial t}\,Y(t,\tau)\varphi(\tau)\,d\tau \\ &\overset{(78.5)}{=} A(t)Y(t,0)\,y_0 + \varphi(t) + \int_0^t A(t)Y(t,\tau)\varphi(\tau)\,d\tau \\ &=\; \varphi(t) + A(t)\Big(Y(t,0)\,y_0 + \int_0^t Y(t,\tau)\varphi(\tau)\,d\tau\Big) \\ &\overset{(78.6)}{=} \varphi(t) + A(t)\,y(t)\,. \end{aligned}$$

Man beachte, daß das Integral in (78.6) vollständig den Einfluß der Inhomogenität auf die Lösung y beschreibt.

Satz 78.4. *Die Ordnung des s-stufigen Gauß-Verfahrens ist $q = 2s$.*

Beweisskizze. Sei $h > 0$ so klein gewählt, daß alle Zeitschritte des Gauß-Verfahrens wohldefiniert sind. Wir bezeichnen mit $p \in \Pi_s^d$ das Kollokationspolynom aus Lemma 78.3 und setzen

$$\varepsilon(t) = p'(t) - f\big(t,p(t)\big)\,, \qquad t_i \le t \le t_i + h\,. \tag{78.7}$$

Aufgrund der Kollokationseigenschaft ist $\varepsilon(t_i + c_j h) = 0$, $j = 1,\dots,s$. Um die Ordnung des Gauß-Verfahrens zu bestimmen, gehen wir davon aus, daß $p(t_i) = y_i = y(t_i)$ auf der Lösungskurve liegt und fragen nach dem Fehler $\|p(t_{i+1}) - y(t_{i+1})\|_2 = \|y_{i+1} - y(t_{i+1})\|_2$.

Dazu betrachten wir für festes $\theta \in [0,1]$ die Differentialgleichung

$$u' = f(t,u) + \theta\varepsilon(t)\,, \qquad u(t_i) = y_i\,, \tag{78.8}$$

deren Lösung mit $u(t,\theta)$ bezeichnet wird. Offensichtlich ist $u(t,0) = y(t)$ und $u(t,1) = p(t)$ für alle $t \in [t_i, t_i + h]$. Somit folgt

$$p(t) - y(t) \;=\; u(t,1) - u(t,0) \;=\; \int_0^1 u_\theta(t,\theta)\,d\theta \tag{78.9}$$

und Differentiation von (78.8) nach θ ergibt

$$u'_\theta = f_y(t,u)u_\theta + \varepsilon(t)\,, \qquad u_\theta(t_i) = 0\,.$$

Da $A(t,\theta) = f_y(t,u) \in \mathbb{R}^{d\times d}$ nicht von u_θ abhängt, erhalten wir auf diese Weise eine inhomogene lineare Differentialgleichung für u_θ, auf die wir die Darstellungsformel (78.6) der Variation der Konstanten anwenden können. Wegen des Anfangswerts $u_\theta(t_i) = 0$ ist die zugehörige homogene Lösung identisch Null, und somit ergibt sich

$$u_\theta(t,\theta) = \int_{t_i}^t Y(t,\tau;\theta)\varepsilon(\tau)\,d\tau\,, \qquad t_i \le t \le t_{i+1}\,,\ 0 \le \theta \le 1\,,$$

mit einem geeigneten Lösungsoperator $Y(\,\cdot\,,\,\cdot\,;\theta)$. In die Fehlerformel (78.9) eingesetzt, erhält man

$$\begin{aligned} y_{i+1} - y(t_i+h) \;&=\; p(t_{i+1}) - y(t_{i+1}) \;=\; \int_0^1 \int_{t_i}^{t_i+h} Y(t_i+h,\tau;\theta)\varepsilon(\tau)\,d\tau\,d\theta \\ &= \int_{t_i}^{t_i+h} \Big(\int_0^1 (Y(t_i+h,\tau;\theta)\,d\theta \Big) \varepsilon(\tau)\,d\tau \;=\; \int_{t_i}^{t_i+h} g(\tau)\varepsilon(\tau)\,d\tau \end{aligned}$$

mit

$$g(\tau) \;=\; \int_0^1 Y(t_i+h,\tau;\theta)\,d\theta\,, \qquad t_i \le \tau \le t_i + h\,,$$

und diese Fehlerdarstellung kann gemäß Lemma 78.2 durch die Gauß-Quadraturformel abgeschätzt werden,

$$\begin{aligned} y_{i+1} - y(t_i+h) \;&=\; h\sum_{j=1}^s b_j g(t_i+c_jh)\varepsilon(t_i+c_jh) \;+\; O(h^{2s+1}) \\ &=\; O(h^{2s+1})\,, \end{aligned}$$

wobei im letzten Schritt $\varepsilon(t_i + c_jh) = 0$, $j = 1,\dots,s$, verwendet wurde.

Leider ist der Beweis so nicht vollständig, da die Konstante in dem O-Term nach Lemma 78.2 von der $2s$-ten Ableitung von $g\varepsilon$ und damit implizit von dem unbekannten Polynom p selbst abhängt. Allerdings läßt sich mit erheblich mehr Aufwand zeigen, daß die „O-Konstante“ unabhängig von p beschränkt bleibt (vgl. [44, Theorem 7.10]). □

Beispiel. Das implizite Mittelpunktverfahren oder Crank-Nicolson-Verfahren aus Beispiel 78.1 hat demnach die Ordung $q = 2$ und ist somit dem impliziten Euler-Verfahren (zumindest in dieser Hinsicht) überlegen. Aus Definition 77.2 erhalten wir die Stabilitätsfunktion des impliziten Mittelpunktverfahrens:

$$R(\zeta) \;=\; 1 + \zeta(1-\zeta/2)^{-1} \;=\; \frac{1+\zeta/2}{1-\zeta/2} \;=\; 1 + \zeta + \tfrac{1}{2}\,\zeta^2 + \tfrac{1}{4}\,\zeta^3 + \dots$$

Zerlegen wir ζ in Real- und Imaginärteil, so folgt

$$0 \;\le\; |R(\zeta)|^2 \;=\; \frac{1 + \operatorname{Re}\zeta + |\zeta|^2/4}{1 - \operatorname{Re}\zeta + |\zeta|^2/4} \;=\; 1 \;+\; \frac{2\operatorname{Re}\zeta}{|1-\zeta/2|^2}\,.$$

Hieran erkennt man, daß die Stabilitätsfunktion die imaginäre Achse auf den Einheitskreisrand und die offene linke Halbebene von $\mathbb{C}$ auf das Innere des Einheitskreises abbildet. Also ist das implizite Mittelpunktverfahren nach Proposition 77.6 A-stabil und zudem Isometrie erhaltend. ◊

Diese beiden Eigenschaften treffen auf *alle* Gauß-Verfahren zu:

Satz 78.5. *Alle Gauß-Verfahren sind A-stabil und Isometrie erhaltend.*

Beweis. Wir wenden das s-stufige Gauß-Verfahren (A, b, c) auf die Testgleichung $y' = \lambda y$, $y_0 = 1$, mit $\lambda \in \mathbb{C}^-$ und Schrittweite $h = 1$ an. Nach Aufgabe 10 ist $I - \lambda A$ invertierbar und somit sind neben

$$y_1 \;=\; R(\lambda) \;=\; 1 + \lambda b^*(I - \lambda A)^{-1}\mathbb{1}$$

auch die Zwischenstellen η_j, $j = 1, \dots, s$, des Gauß-Verfahrens wohldefiniert, vgl. den Beweis von Proposition 77.3. Wir greifen nun auf das in Lemma 78.3 konstruierte Kollokationspolynom p zu diesem Runge-Kutta-Schritt zurück und setzen $g = |p|^2 = \overline{p}p$. Demnach ist

$$\begin{aligned} |R(\lambda)|^2 \;&=\; |y_1|^2 \;=\; |p(1)|^2 \;=\; g(1) \;=\; g(0) + \int_0^1 g'(\tau)\,d\tau \\ &=\; 1 + 2\operatorname{Re}\int_0^1 \overline{p(\tau)}\,p'(\tau)\,d\tau\,. \end{aligned}$$

Da $\overline{p}p'$ zu Π_{2s-1} gehört, kann das Integral mit der Gauß-Legendre-Formel exakt ausgewertet werden und es ergibt sich

$$|R(\lambda)|^2 \;=\; 1 + 2\operatorname{Re}\sum_{j=1}^{s} b_j \overline{p(c_j)}\,p'(c_j)\,.$$

Die Gewichte b_j der Gaußformeln sind immer positiv (vgl. Satz 40.3). Aus der Kollokationseigenschaft von p folgt ferner, daß $p'(c_j) = f(c_j, p(c_j)) = \lambda p(c_j)$.

Folglich ist

$$|R(\lambda)|^2 \;=\; 1 + 2\,\mathrm{Re}\;\lambda \sum_{j=1}^{s} b_j |p(c_j)|^2\,, \tag{78.10}$$

und dies ist genau dann kleiner, größer oder gleich Eins, wenn Re λ kleiner, größer oder gleich Null ist. □

Zum Abschluß dieses Paragraphen skizzieren wir die effiziente Implementierung der Gauß-Verfahren (bzw. allgemeiner impliziter Runge-Kutta-Verfahren). Der Hauptaufwand steckt dabei in der Lösung des nichtlinearen Gleichungssystems

$$\eta_j = y_i + h \sum_{\nu=1}^{s} a_{j\nu} f(t_i + c_\nu h, \eta_\nu)\,, \qquad j = 1, \ldots, s\,.$$

Dieses Gleichungssystem wird in der Regel iterativ mit einem Newton-artigen Verfahren gelöst. Auf diese Weise ergeben sich Näherungen für η_ν, aus denen dann die Steigungen $f(t_i + c_j h, \eta_j)$ für den Schlußschritt

$$y_{i+1} = y_i + h \sum_{j=1}^{s} b_j f(t_i + c_j h, \eta_j) \tag{78.11}$$

berechnet werden. Für den folgenden Algorithmus führen wir für diese Steigungen Hilfsvariablen k_j ein,

$$k_j = f(t_i + c_j h, \eta_j)\,,$$

mit denen das Runge-Kutta-Verfahren folgendermaßen umformuliert werden kann:

$$k_j \;=\; f\Big(t_i + c_j h, y_i + h \sum_{\nu=1}^{s} a_{j\nu} k_\nu\Big)\,, \qquad j = 1, \ldots, s\,, \tag{78.12a}$$

$$y_{i+1} \;=\; y_i + h \sum_{\nu=1}^{s} b_\nu k_\nu\,. \tag{78.12b}$$

Zur Lösung des $sd \times sd$-dimensionalen Gleichungssystems (78.12a) für die Steigungen $\{k_j\}$ wird zumeist ein vereinfachtes Newton-Verfahren verwendet, bei dem die exakten Ableitungen $f_y(t_i + c_j h, y_i + h \sum a_{j\nu} k_\nu)$ durch die „eingefrorene“ Näherung $J = f_y(t_i, y_i)$ approximiert werden. Mit der abkürzenden Schreibweise

$$k = \Phi(k)\,, \qquad k = \begin{bmatrix} k_1 \\ \vdots \\ k_s \end{bmatrix} \in \mathbb{R}^{sd}\,,$$

Initialisierung: (A, b, c) sei ein Runge-Kutta-Verfahren der Ordnung q,
y_0 und t_0 sowie Schrittweite h seien gegeben

`for` $i = 0, 1, 2, \ldots$ `do`
  $J = f_y(t_i, y_i)$
  werte D_Φ aus (78.13) aus
  % berechne Vektorkomponenten $k_j^{(0)}$ von $k^{(0)}$:
  `for` $j = 1, \ldots, s$ `do`
    $k_j^{(0)} = f(t_i, y_i)$
  `end for`
  `for` $n = 0, 1, 2, \ldots$ `do` % vereinfachte Newton-Iteration
    löse $(I - D_\Phi) z^{(n)} = \Phi(k^{(n)}) - k^{(n)}$
    $k^{(n+1)} = k^{(n)} + z^{(n)}$
  `until` stop % end for (Newton-Iteration), vgl. (75.9) mit h^{q+1} statt h^2
  $t_{i+1} = t_i + h$
  $y_{i+1} = y_i + h \sum_{j=1}^{s} b_j k_j^{(n)}$
`until` $t_{i+1} > T$ % end for (i-Schleife)

Ergebnis: $y_i \approx y(t_i), \quad i = 0, 1, 2, \ldots$

Algorithmus 78.1: Implizites Runge-Kutta-Verfahren

für das nichtlineare Gleichungssystem (78.12a) ergibt dies die Näherung

$$D_\Phi = h \begin{bmatrix} a_{11}J & a_{12}J & \cdots & a_{1s}J \\ a_{21}J & a_{22}J & & a_{2s}J \\ \vdots & & & \vdots \\ a_{s1}J & a_{s2}J & \cdots & a_{ss}J \end{bmatrix} \tag{78.13}$$

für die Jacobi-Matrix Φ'. Als Startnäherung für die Iteration bieten sich die Steigungen $k_j = f(t_i, y_i)$, $j = 1, \ldots, s$, an. Ein entsprechender Implementierungsvorschlag ist in Algorithmus 78.1 enthalten.

79 Radau-IIA-Verfahren

Die Runge-Kutta-Gauß-Verfahren des vorigen Abschnitts besitzen (fast) alle erstrebenswerten Eigenschaften, die wir bisher kennengelernt haben: sie haben maximale Ordnung, sind A-stabil und zudem noch Isometrie erhaltend. Es gibt allerdings wichtige Anwendungen, etwa die Wärmeleitungsgleichung, bei

denen sich gerade die letzte Eigenschaft als Nachteil herausstellt. Wie wir in Beispiel 77.11 gesehen haben, konvergieren bei der Wärmeleitungsgleichung die für die Stabilitätsanalyse relevanten Eigenwerte λ_n gegen $-\infty$ und die zugehörigen Lösungskomponenten klingen schnell ab. Für ein Runge-Kutta-Verfahren wäre es daher wünschenswert, daß auch $R(\lambda_n)$ für $n \to \infty$ gegen Null konvergiert. Für ein Isometrie erhaltendes Runge-Kutta-Verfahren gilt hingegen $|R(\infty)| = 1$.

Definition 79.1. Ein Runge-Kutta-Verfahren (A, b, c) heißt *steifgenau* (engl. *stiffly accurate*), falls $A = [a_{ij}]$ invertierbar ist und $a_{sj} = b_j$, $j = 1, \ldots, s$, gilt. Ein A-stabiles Runge-Kutta-Verfahren mit Stabilitätsfunktion R heißt *L-stabil*, falls $R(\infty) = 0$ ist.

Gauß-Verfahren sind somit nicht L-stabil. Aufgrund der Stetigkeit der Stabilitätsfunktion im Punkt ∞ sind Isometrie erhaltende Verfahren generell niemals L-stabil und umgekehrt L-stabile Verfahren niemals Isometrie erhaltend.

Proposition 79.2. *Ein A-stabiles steifgenaues Runge-Kutta-Verfahren ist L-stabil.*

Beweis. Für ein allgemeines Runge-Kutta-Verfahren (A, b, c) ist

$$R(\zeta) \;=\; 1 + \zeta b^*(I - \zeta A)^{-1}\mathbb{1} \;=\; 1 - b^*(A - I/\zeta)^{-1}\mathbb{1}\,.$$

Bei einem steifgenauen Runge-Kutta-Verfahren konvergiert $(A - I/\zeta)^{-1}$ für $\zeta \to \infty$ gegen A^{-1} und somit ist

$$R(\infty) \;=\; 1 - (A^{-*}b)^*\mathbb{1}\,.$$

Wegen $a_{sj} = b_j$ ist $A^* e_s = b$ und somit folgt $R(\infty) = 1 - e_s^*\mathbb{1} = 1 - 1 = 0$. Also sind A-stabile steifgenaue Runge-Kutta-Verfahren L-stabil. □

Für steifgenaue Verfahren gilt

$$c_s = \sum_{j=1}^{s} a_{sj} = \sum_{j=1}^{s} b_j = 1\,,$$

folglich stimmt in jedem Zeitschritt die letzte Stützstelle eines solchen Verfahrens mit dem neuen Zeitpunkt t_{i+1} überein. Ferner sind die definierenden Gleichungen für η_s und y_{i+1} identisch, das heißt in jedem Zeitschritt ist

$$t_i + c_s h = t_{i+1}\,, \qquad \eta_s = y_{i+1}\,.$$

Aus diesem Grund bieten sich die Radau-Legendre-Formeln aus Abschnitt 41 als Grundlage für steifgenaue (und L-stabile) Runge-Kutta-Verfahren an. Sie

haben den optimalen Exaktheitsgrad $2s-2$ unter der Nebenbedingung $c_s = 1$. Die Konstruktion ist dabei völlig analog zur Vorgehensweise bei den Runge-Kutta-Gauß-Verfahren.

Wir übernehmen die (aufsteigend angeordneten) Knoten c_j und die Gewichte b_j der s-stufigen Radau-Legendre-Formel

$$R_s[f] \;=\; \sum_{j=1}^{s} b_j g(c_j) \;\approx\; \int_0^1 g(t)\,dt\,,$$

und bestimmen die Koeffizienten $a_{j\nu}$ durch Integration der Lagrange-Grundpolynome l_ν, vgl. (78.3). Da die Radau-Legendre-Formel R_s alle Polynome vom Grad $2s-2$ über $[0,1]$ exakt integriert, also insbesondere auch die Lagrange-Grundpolynome, stimmen die Koeffizienten $a_{s\nu}$ wegen $c_s = 1$ mit den Radau-Legendre-Gewichten b_ν überein. Das resultierende Runge-Kutta-Verfahren heißt *Radau-IIA-Verfahren* und hat Konsistenzordnung $q = 2s-1$, vgl. [45].

Um nachzuweisen, daß Radau-IIA-Verfahren steifgenau sind, müssen wir lediglich noch die Nichtsingularität der Matrix A überprüfen. Dazu nehmen wir an, daß ein Vektor $x = [x_1, \ldots, x_s]^T$ existiert mit $Ax = 0$. Das zugehörige Polynom

$$p \;=\; \sum_{\nu=1}^{s} x_\nu l_\nu \;\in\; \Pi_{s-1} \qquad \text{mit} \qquad p(c_\nu) = x_\nu\,,\ \nu = 1,\ldots,s\,,$$

erfüllt wegen (78.3) die Gleichungen

$$0 \;=\; \sum_{\nu=1}^{s} a_{j\nu} x_\nu \;=\; \sum_{\nu=1}^{s} x_\nu \int_0^{c_j} l_\nu(t)\,dt \;=\; \int_0^{c_j} p(t)\,dt\,, \qquad j = 1,\ldots,s\,,$$

und durch Subtraktion der einzelnen Gleichungen ergibt sich

$$\int_{c_{j-1}}^{c_j} p(t)\,dt \;=\; 0\,, \qquad j = 1,\ldots,s\,,$$

wobei wir $c_0 = 0$ gesetzt haben. Die Intervalle (c_{j-1}, c_j), $j = 1,\ldots,s$, sind dabei nichtleere offene Mengen, da alle Knoten der Radau-Formel paarweise verschieden und positiv sind. Folglich muß p in jedem solchen Intervall mindestens eine Nullstelle haben. Dies ergibt mindestens s Nullstellen und somit ist p das Nullpolynom, also x der Nullvektor in $\mathbb{R}^s$. Damit haben wir die erste Aussage des folgenden Satzes bewiesen:

Satz 79.3. *Die Radau-IIA-Verfahren sind steifgenau und L-stabil.*

Beweis. Zum Nachweis der L-Stabilität müssen wir lediglich noch die A-Stabilität nachweisen. Dazu gehen wir wie in Abschnitt 78 vor und berücksichtigen, daß der Beweis von Lemma 78.3 und daher auch die Aussage des Lemmas wortwörtlich auf Radau-IIA-Verfahren anwendbar sind. Wie im Beweis von Satz 78.5 erhalten wir daher für die Stabilitätsfunktion R des s-stufigen Radau-IIA-Verfahrens die Identität

$$|R(\lambda)|^2 \;=\; 1 + 2\,\mathrm{Re} \int_0^1 \overline{p(\tau)}\, p'(\tau)\, d\tau\,, \tag{79.1}$$

wobei $p \in \Pi_s$ mit $p(0) = 1$ das Kollokationspolynom mit $p(1) = R(\lambda)$ und

$$p'(c_j) \;=\; \lambda p(c_j)\,, \qquad j = 1, \ldots, s\,, \tag{79.2}$$

bezeichnet. Im Gegensatz zu dem Beweis von Satz 78.5 kann jedoch das Integral in (79.1) im allgemeinen nicht mit der Radau-Legendre-Formel $R_s[\,\cdot\,]$ exakt ausgewertet werden, da $\overline{p}p'$ zu Π_{2s-1} gehört. Statt dessen ergibt sich bei der Auswertung ein Quadraturfehler

$$\varepsilon \;=\; \int_0^1 \overline{p(\tau)}\, p'(\tau)\, d\tau \;-\; R_s[\overline{p}p']\,,$$

den es genauer zu untersuchen gilt. Dazu interpolieren wir $\overline{p}p'$ ähnlich wie im Beweis von Satz 40.5 durch ein Polynom $q_{2s-2} \in \Pi_{2s-2}$ derart, daß q_{2s-2} und $\overline{p}p'$ und die zugehörigen Ableitungen in den Knoten $c_1, \ldots, c_{s-1} \in (0,1)$ übereinstimmen sowie zusätzlich noch $q_{2s-2}(1) = \overline{p(1)}p'(1)$ gilt. In Analogie zu Satz 37.4 läßt sich dann der resultierende Interpolationsfehler in der Form

$$\overline{p(t)}p'(t) - q_{2s-2}(t) \;=\; \frac{1}{(2s-1)!}\, \frac{d^{2s-1}\big(\overline{p}p'\big)}{dt^{2s-1}}(\tau)\,(t-1) \prod_{j=1}^{s-1} (t - c_j)^2$$

mit einer Zwischenstelle $\tau \in (0,1)$ schreiben. Da $\overline{p}p'$ zu Π_{2s-1} gehört und für jedes nicht konstante Polynom p immer einen positiven Höchstkoeffizienten besitzt, ist die $(2s-1)$-te Ableitung jedoch konstant und größer gleich Null. Demnach ist

$$\overline{p(t)}p'(t) - q_{2s-2}(t) \;\le\; 0 \qquad \text{für alle } t \in [0,1]\,,$$

und für den Quadraturfehler ergibt sich hieraus

$$\varepsilon \;=\; \int_0^1 \overline{p(\tau)}\, p'(\tau)\, d\tau \;-\; R_s[q_{2s-2}] \;=\; \int_0^1 \big(\overline{p(\tau)}\, p'(\tau) - q_{2s-2}(\tau)\big)\, d\tau \;\le\; 0\,.$$

Aus (79.1) und (79.2) folgt somit

$$|R(\lambda)|^2 = 1 + 2\,\mathrm{Re}\; R_s[\overline{p}p'] + \varepsilon \;\le\; 1 + 2\,\mathrm{Re} \sum_{j=1}^{s} b_j \overline{p(c_j)}\, p'(c_j)$$

$$= 1 + 2\,\mathrm{Re}\; \lambda \sum_{j=1}^{s} b_j |p(c_j)|^2 \,. \tag{79.3}$$

Da die Gewichte b_j der Radau-Legendre-Formeln positiv sind, vgl. Proposition 41.5, ist die Summe in (79.3) nichtnegativ und somit $|R(\lambda)| \le 1$ für $\mathrm{Re}\;\lambda \le 0$. Radau-IIA-Verfahren sind somit A-stabil und wegen der Steifgenauigkeit auch L-stabil. □

Beispiel 79.4. Das erste Radau-IIA-Verfahren ($s = 1$) ist das implizite Euler-Verfahren mit Konsistenzordnung $q = 2s - 1 = 1$. Die zweite Radau-Legendre-Formel wurde in Aufgabe VII.13 bestimmt, das zugehörige Radau-IIA-Verfahren hat die Parameter

$$b_1 = 3/4\,,\; b_2 = 1/4\,, \qquad c_1 = 1/3\,,\; c_2 = 1\,.$$

Wegen der Steifgenauigkeit sind damit auch $a_{21} = 3/4$ und $a_{22} = 1/4$ festgelegt und es verbleibt lediglich die Berechnung von a_{11} und a_{12} aus (78.3). Dies ergibt $a_{11} = 5/12$ und $a_{12} = -1/12$. Daher erhalten wir für das Radau-IIA-Verfahren der Ordnung $q = 3$ das Runge-Kutta-Tableau

$1/3$	$5/12$	$-1/12$
1	$3/4$	$1/4$
	$3/4$	$1/4$

Das Radau-IIA-Verfahren der Ordnung $q = 5$ lautet

$\frac{2}{5} - \frac{\sqrt{6}}{10}$	$\frac{11}{45} - \frac{7\sqrt{6}}{360}$	$\frac{37}{225} - \frac{169\sqrt{6}}{1800}$	$-\frac{2}{225} + \frac{\sqrt{6}}{75}$
$\frac{2}{5} + \frac{\sqrt{6}}{10}$	$\frac{37}{225} + \frac{169\sqrt{6}}{1800}$	$\frac{11}{45} + \frac{7\sqrt{6}}{360}$	$-\frac{2}{225} - \frac{\sqrt{6}}{75}$
1	$\frac{4}{9} - \frac{\sqrt{6}}{36}$	$\frac{4}{9} + \frac{\sqrt{6}}{36}$	$\frac{1}{9}$
	$\frac{4}{9} - \frac{\sqrt{6}}{36}$	$\frac{4}{9} + \frac{\sqrt{6}}{36}$	$\frac{1}{9}$

Eine Implementierung dieses dreistufigen Verfahrens mit automatischer Schrittweitensteuerung und weiteren Optionen wird in [45, Appendix] beschrieben.[4]

◊

[4] Das zugehörige Programmpaket `radau5` der Autoren Hairer und Wanner ist auf der Internetseite `http://www.unige.ch/math/folks/hairer/software.html` frei zugänglich.

80 Rosenbrock-Typ-Verfahren

In diesem Abschnitt führen wir eine andere Klasse von Einschrittverfahren ein, die aus den Runge-Kutta-Verfahren abgeleitet werden können. Wir beschränken uns dabei ausschließlich auf autonome Differentialgleichungen

$$y' = f(y)\,, \qquad y(0) = y_0\,.$$

Es sei daran erinnert, daß durch Hinzunahme einer weiteren Lösungskomponente jede Differentialgleichung in eine autonome Differentialgleichung transformiert werden kann, vgl. Aufgabe 7.

Implizite Runge-Kutta-Verfahren haben eine Koeffizientenmatrix A, die keine strikte untere Dreiecksgestalt hat. Die nichtlinearen Gleichungssysteme, die in jedem Schritt gelöst werden müssen – etwa (78.12a) für die Steigungen k_j – haben im allgemeinen die Dimension $sd \times sd$. Einfacher ist die Situation, wenn die Matrix A eine untere Dreiecksmatrix mit nicht-verschwindender Diagonalen ist. In diesem Fall sind die s Gleichungen für die Steigungen *entkoppelt*,

$$k_j = f\Big(y_i + h\sum_{\nu=1}^{j-1} a_{j\nu}k_\nu + a_{jj}hk_j\Big)\,, \qquad j = 1,\dots,s\,, \tag{80.1}$$

und können daher sequentiell gelöst werden. Damit reduziert sich der Aufwand auf die Lösung von s lediglich d-dimensionalen nichtlinearen Gleichungssystemen. Das (vereinfachte) Newton-Verfahren zur Lösung der j-ten Gleichung (80.1) führt auf die Iteration

$$(I - a_{jj}hJ)\big(k_j^{(n+1)} - k_j^{(n)}\big) \;=\; f\Big(y_i + h\sum_{\nu=1}^{j-1} a_{j\nu}k_\nu + a_{jj}hk_j^{(n)}\Big) - k_j^{(n)}$$

beziehungsweise

$$(I - a_{jj}hJ)k_j^{(n+1)} \;=\; f\Big(y_i + h\sum_{\nu=1}^{j-1} a_{j\nu}k_\nu + a_{jj}hk_j^{(n)}\Big) - a_{jj}hJk_j^{(n)}\,,$$

wobei in der Regel wieder die Näherungsableitung $J = f_y(y_i)$ anstelle der Ableitung an der aktuellen y-Koordinate verwendet wird. Für hinreichend kleine h sind diese linearen Gleichungssysteme eindeutig lösbar.

Häufig ist es für die Genauigkeit völlig ausreichend, lediglich einen einzigen Iterationsschritt durchzuführen – zumindest bei vernünftiger Schrittweite h und hinreichend guter Startnäherung $k_j^{(0)}$. Bei einem Ansatz $k_j^{(0)} = \sum_{\nu=1}^{j-1} d_{j\nu}/a_{jj}\, k_\nu$

für die jeweilige Startnäherung mit noch zu spezifizierenden $d_{j\nu}$ führt diese Vereinfachung auf ein sogenanntes *linear-implizites* Runge-Kutta-Verfahren

$$\begin{aligned}(I - a_{jj}hJ)k_j &= f\Big(y_i + h\sum_{\nu=1}^{j-1}(a_{j\nu}+d_{j\nu})k_\nu\Big) - hJ\sum_{\nu=1}^{j-1} d_{j\nu}k_\nu\,, \\ & \qquad\qquad j = 1,\dots,s\,, \\ y_{i+1} &= y_i + h\textstyle\sum_{j=1}^{s} b_j k_j\,.\end{aligned} \tag{80.2}$$

Es handelt sich dabei um ein implizites Einschrittverfahren, da zur Berechnung jeder einzelnen Stufe k_j ein Gleichungssystem gelöst werden muß. Die Gleichungssysteme sind allerdings linear, so daß der Arbeitsaufwand um ein Vielfaches niedriger ist als bei (echt) impliziten Runge-Kutta-Verfahren. In gewisser Weise stellen linear-implizite Verfahren also einen Kompromiß aus expliziten und impliziten Runge-Kutta-Verfahren dar. Sie erben von den expliziten Verfahren den moderaten Arbeitsaufwand und behalten zudem den Hauptvorteil der impliziten Verfahren, nämlich die Stabilität.

Proposition 80.1. *Sei (A,b,c) ein Runge-Kutta-Verfahren mit nichtsingulärer linker unterer Dreiecksmatrix $A \in \mathbb{R}^{s\times s}$ und Stabilitätsfunktion R. Dann hat das daraus abgeleitete linear-implizite Verfahren* (80.2) *mit $J = f_y(y_i)$ unabhängig von der Wahl der $d_{j\nu}$ die gleiche Stabilitätsfunktion wie das Ausgangsverfahren, d. h. bei Anwendung auf die Testgleichung $y' = \lambda y$, $y(0) = 1$, ergeben sich die Näherungen*

$$y_i = R(h\lambda)^i\,, \qquad i = 1,2,\dots\,,$$

sofern $1/(h\lambda) \,/\in \sigma(A) = \{a_{11},\dots,a_{ss}\}$.

Beweis. Wir müssen das linear-implizite Verfahren auf die eindimensionale Testgleichung $y' = \lambda y$, $y(0) = 1$, anwenden. Wegen $f(y) = \lambda y$ ergibt sich in diesem Fall $J = \lambda$ und die j-te Gleichung der Rekursion (80.2) hat die Form

$$(1 - a_{jj}h\lambda)k_j = \lambda + h\lambda\sum_{\nu=1}^{j-1}(a_{j\nu}+d_{j\nu})k_\nu - h\lambda\sum_{\nu=1}^{j-1} d_{j\nu}k_\nu = \lambda + h\lambda\sum_{\nu=1}^{j-1} a_{j\nu}k_\nu\,.$$

Multiplikation mit h und Substitution von $\zeta = h\lambda$ ergibt folglich

$$k_j h - \zeta\sum_{\nu=1}^{j} a_{j\nu}k_\nu h = \zeta\,, \qquad j = 1,\dots,s\,,$$

beziehungsweise

$$(I - \zeta A)kh = \zeta\mathbb{1}\,, \qquad k = [k_1,\dots,k_s]^T \in \mathbb{R}^s\,.$$

Eingesetzt in die Definition von y_1 folgt daraus

$$y_1 \;=\; 1 + b^* k h \;=\; 1 + \zeta b^* (I - \zeta A)^{-1} \mathbb{1}\,,$$

da $I - \zeta A$ nach Voraussetzung invertierbar ist. Damit ist y_1 aber gerade der Wert der Stabilitätsfunktion des Ausgangsverfahrens an der Stelle $h\lambda$, vgl. Definition 77.2. □

Linear-impliziten Einschrittverfahren läßt sich also genau wie Runge-Kutta-Verfahren eine Stabilitätsfunktion im Sinne von Proposition 77.3 zuordnen. Die Resultate von Abschnitt 77 gelten entsprechend. Da die wichtigsten Stabilitätseigenschaften gemäß Proposition 77.6 allein von der Stabilitätsfunktion abhängen, übertragen sich dank des eben bewiesenen Resultats entsprechende Eigenschaften eines impliziten Runge-Kutta-Verfahrens unmittelbar auf das dazugehörige linear-implizite Verfahren.

Es sei aber an dieser Stelle darauf hingewiesen, daß die Konsistenzordnung des Verfahrens (80.2) bei ungünstiger Wahl der d_{jk} geringer sein kann als die Ordnung des impliziten Ausgangsverfahrens.

Beispiel. Zur Herleitung des *linear-impliziten Euler-Verfahrens* verwenden wir die Koeffizienten $A = 1$ und $b = 1$, vgl. Beispiel 76.5. Demnach ergibt sich aus (80.2) das Verfahren

$$y_{i+1} = y_i + hk\,, \qquad (I - h f_y(y_i))\, k = f(y_i)\,. \tag{80.3}$$

Wie das implizite Euler-Verfahren ist auch das linear-implizite Euler-Verfahren A-stabil. Im Gegensatz zum impliziten Euler-Verfahren muß aber bei der linear-impliziten Variante in jedem Zeitschritt lediglich ein lineares Gleichungssystem gelöst werden.

Das A-stabile und Isometrie erhaltende *linear-implizite Mittelpunktverfahren* lautet entsprechend

$$y_{i+1} = y_i + hk\,, \qquad (I - \tfrac{h}{2} f_y(y_i))\, k = f(y_i)\,. \tag{80.4}$$

◊

Für die Praxis ergibt sich noch eine erhebliche Vereinfachung, wenn bei allen Stufen der gleiche Koeffizient $a_{jj} = a$ verwendet wird. In diesem Fall haben alle linearen Gleichungssysteme (80.2) die gleiche Koeffizientenmatrix $I - ahJ$ und daher wird lediglich die LR-Zerlegung einer einzigen Matrix benötigt. Meist ist es sogar möglich, *dieselbe* Jacobi-Matrix J über mehrere Zeitschritte hinweg zu verwenden. Auf diese Weise erspart man sich zusätzlichen Aufwand.

Initialisierung: y_0 und t_0 sowie Schrittweite h seien gegeben

`for` $i = 0, 1, 2, \ldots$ `do`

$t_{i+1} = t_i + h$

$J = f_y(y_i)$ % nicht notwendig in jedem Zeitschritt

faktorisiere $I - ahJ = LR$ % LR-Zerlegung

`for` $j = 1, \ldots, s$ `do`

$$LR\,k_j = f\big(y_i + h\sum_{\nu=1}^{j-1} a_{j\nu}k_\nu\big) \;-\; hJ\sum_{\nu=1}^{j-1} d_{j\nu}k_\nu$$

`end for`

$$y_{i+1} = y_i + h\sum_{j=1}^{s} b_j k_j$$

`until` $t_{i+1} \geq T$ % end for (i-Schleife)

Ergebnis: $y_i \approx y(t_i), \quad i = 0, 1, 2 \ldots$

Algorithmus 80.1: Rosenbrock-Typ-Verfahren

Lösen wir uns von der bisherigen Bedeutung der Koeffizienten $a_{j\nu}$, dann kann die Definition (80.2) umformuliert werden in

$$\begin{aligned} (I - ahJ)k_j &= f(y_i + h\sum_{\nu=1}^{j-1} a_{j\nu}k_\nu) - hJ\sum_{\nu=1}^{j-1} d_{j\nu}k_\nu\,, \qquad j = 1, \ldots, s\,, \\ y_{i+1} &= y_i + h\sum_{j=1}^{s} b_j k_j \end{aligned} \tag{80.5}$$

einem sogenannten *Rosenbrock-Typ-Verfahren*. Dabei sind a, $a_{j\nu}$, $d_{j\nu}$ und b_j geeignet zu wählende Parameter, die allein im Hinblick auf Ordnung und Stabilität des Verfahrens optimiert werden können.

Beispiel 80.2. Das Programmpaket MATLAB bietet das Rosenbrock-Typ-Verfahren `ode23s` (das wir bereits in Beispiel 77.12 verwendet haben) zur Lösung steifer Differentialgleichungen an. In seiner Grundform ist das Verfahren folgendermaßen definiert (vgl. Shampine und Reichelt [97]):

$$\begin{aligned} (I - ahJ)k_1 &= f(y_i)\,, \\ (I - ahJ)k_2 &= f(y_i + h\,\tfrac{1}{2}k_1) - ahJk_1\,, \\ y_{i+1} &= y_i + hk_2\,, \end{aligned}$$

mit $a = 1/(2 + \sqrt{2})$. Offensichtlich ist dieses Verfahren ein Spezialfall von Algorithmus 80.1 mit den Parametern $a_{21} = 1/2$ und $d_{21} = a$ sowie $b_1 = 0$ und $b_2 = 1$. ◊

Weitere Rosenbrock-Typ-Verfahren finden sich in [45, Abschnitt IV.7]. Wir wollen hier jedoch auf keine anderen Beispiele mehr eingehen, sondern statt dessen exemplarisch die Ordnung und die Stabilitätseigenschaften des Verfahrens `ode23s` untersuchen.

Satz 80.3. *Das Rosenbrock-Typ-Verfahren* `ode23s` *ist ein Verfahren zweiter Ordnung.*

Beweis. Wir beginnen mit folgender Beobachtung: Für $0 < h < 1/(2a\|J\|_2)$ ist die $s \times s$-Matrix $I - ahJ$ invertierbar, und für die Lösung k eines linearen Gleichungssystems $(I - ahJ)k = f$ ergibt sich wegen

$$\|(I - ahJ)k\|_2 \;\geq\; \|k\|_2 - ah\|J\|_2\|k\|_2 \;\geq\; \frac{1}{2}\|k\|_2$$

unmittelbar die Abschätzung

$$\|k\|_2 \leq 2\|f\|_2 .$$

Aus der Definition von k_1 folgt daher durch rekursives Einsetzen

$$\begin{aligned} k_1 \;&=\; f + ahJk_1 \;=\; f + ahJ\big(f + ahJk_1\big) \\ &=\; f + ahJf + O(h^2) , \end{aligned} \tag{80.6}$$

wobei wir wie schon früher das Argument y_i von f wieder weggelassen haben. Dies werden wir im weiteren auch bei Ableitungen von f so halten. Mittels Taylorentwicklung ergibt sich nun in entsprechender Weise unter Berücksichtigung von (80.6)

$$\begin{aligned} k_2 \;&=\; f(y_i + h\tfrac{1}{2}k_1) - ahJk_1 + ahJk_2 \\ &=\; f + f_y h\tfrac{1}{2}k_1 - ahJk_1 + ahJk_2 + O(h^2) \\ &=\; f + f_y h\tfrac{1}{2}f - ahJf + ahJk_2 + O(h^2) \\ &=\; f + f_y h\tfrac{1}{2}f - ahJf + ahJf + O(h^2) \\ &=\; f + f_y h\tfrac{1}{2}f + O(h^2) . \end{aligned}$$

Somit ist

$$y_{i+1} \;=\; y_i + hk_2 \;=\; y_i \;+\; hf \;+\; h^2\tfrac{1}{2}f_y f \;+\; O(h^3) .$$

Ein Vergleich der obigen Entwicklung mit der Taylorentwicklung (76.6) der exakten Lösung ergibt eine Übereinstimmung bis auf den Term $O(h^3)$, d. h. das Rosenbrock-Typ-Verfahren `ode23s` ist ein Einschrittverfahren zweiter Ordnung. □

Man beachte, daß in diesem Beweis an keiner Stelle verwendet wurde, daß J die *exakte Ableitung* $f_y(y_i)$ ist. Mit anderen Worten: Das Verfahren `ode23s` ist selbst dann noch ein Verfahren zweiter Ordnung, wenn J nur eine Näherung an $f_y(y_i)$ darstellt (genau genommen ist noch nicht einmal das notwendig). Aus dem folgenden Lemma folgt, daß `ode23s` selbst mit exakter Jacobi-Matrix kein Verfahren dritter Ordnung ist, da nur die ersten drei Terme der Taylorentwicklungen der Stabilitätsfunktion und der Exponentialfunktion übereinstimmen. Es besteht also (zumindest aus diesem Blickwinkel heraus) kein unmittelbarer Zwang, die Jacobi-Matrix *in jedem Zeitschritt* exakt auszuwerten. Es reicht, J ab und an neu zu berechnen.

Lemma 80.4. *Für $J = f_y(y_i)$ hat die Stabilitätsfunktion $R(\zeta)$ des Rosenbrock-Typ-Verfahrens* `ode23s` *die Form*

$$R(\zeta) \;=\; \frac{1+(1-2a)\zeta}{(1-a\zeta)^2} \;=\; 1 \,+\, \zeta \,+\, \tfrac{1}{2}\zeta^2 \,+\, (\tfrac{1}{2}-a)\zeta^3 \,+\, O(\zeta^4)$$

für $\zeta \to 0$. Dabei ist weiterhin $a = 1/(2+\sqrt{2})$.

Beweis. Für die Testgleichung $y' = \lambda y$ mit $y(0) = y_0 = 1$ ist $f(y) = \lambda y$ und $J = f_y(y_0) = \lambda$. Für hinreichend kleines $h > 0$ ergibt sich somit

$$k_1 = \lambda/(1-ah\lambda) \quad \text{und} \quad k_2 = \big(\lambda(1+h\tfrac{1}{2}k_1) - ah\lambda k_1\big)/(1-ah\lambda)\,.$$

Durch Einsetzen in $R(\zeta) = y_1 = y_0 + hk_2 = 1 + hk_2$ ergibt das mit $\zeta = h\lambda$

$$\begin{aligned} R(\zeta) \;&=\; 1 + \frac{h\lambda + h^2\lambda k_1(\frac{1}{2}-a)}{1-ah\lambda} \;=\; 1 + \frac{\zeta(1-a\zeta)+\zeta^2(\frac{1}{2}-a)}{(1-a\zeta)^2} \\ &=\; \frac{1+(1-2a)\zeta+(a^2-2a+\frac{1}{2})\zeta^2}{(1-a\zeta)^2}\,. \end{aligned}$$

Für $a = 1/(2+\sqrt{2}) = 1-1/\sqrt{2}$ ist $a^2-2a+1/2 = 0$ und somit folgt unmittelbar die erste Behauptung.

Mit Hilfe der geometrischen Reihe ergibt sich schließlich die Reihenentwicklung von $R(\zeta)$ um $\zeta = 0$:

$$\begin{aligned} R(\zeta) \;&=\; \big(1+a\zeta+a^2\zeta^2+a^3\zeta^3+O(\zeta^4)\big)^2\big(1+(1-2a)\zeta\big) \\ &=\; \big(1+2a\zeta+3a^2\zeta^2+4a^3\zeta^3+O(\zeta^4)\big)\big(1+(1-2a)\zeta\big) \\ &=\; 1 \,+\, (2a+1-2a)\zeta \,+\, (3a^2+2a-4a^2)\zeta^2 \\ &\qquad +\, (4a^3+3a^2-6a^3)\zeta^3 \,+\, O(\zeta^4) \\ &=\; 1 \,+\, \zeta \,+\, (2a-a^2)\zeta^2 \,+\, (3a^2-2a^3)\zeta^3 \,+\, O(\zeta^4)\,. \end{aligned}$$

Wegen $a^2 - 2a + 1/2 = 0$ folgt $2a - a^2 = 1/2$ und $3a^2 - 2a^3 = -2a(a^2 - 2a + 1/2) - a^2 + a = 1/2 - a$. Folglich ist auch die zweite Darstellung nachgewiesen. □

Nun wenden wir uns der Stabilität dieses Rosenbrock-Typ-Verfahrens zu.

Satz 80.5. *Das Verfahren* `ode23s` *ist L-stabil, falls in jedem Zeitschritt die exakte Jacobi-Matrix $J = f_y(y_i)$ verwendet wird.*

Beweis. Wir beweisen zunächst die A-Stabilität, also daß $|R(\zeta)| \leq 1$ ist für alle $\zeta \in \mathbb{C}$ mit $\operatorname{Re} \zeta \leq 0$. Dazu betrachten wir die Stabilitätsfunktion auf der imaginären Achse:

$$|R(\mathrm{i}t)|^2 = \frac{|\,1 + \mathrm{i}(1-2a)t\,|^2}{|\,1 - \mathrm{i}at\,|^4} = \frac{1 + (1-2a)^2t^2}{(1+a^2t^2)^2} = \frac{1 + (1-4a+4a^2)t^2}{1 + 2a^2t^2 + a^4t^4}\,.$$

Wegen $a = 1/(2+\sqrt{2})$ ergibt sich $1 - 4a + 4a^2 = 2a^2$ und folglich ist

$$|R(\mathrm{i}t)|^2 = \frac{1 + 2a^2t^2}{1 + 2a^2t^2 + a^4t^4} \leq 1\,, \qquad t \in \mathbb{R}\,.$$

Da die rationale Funktion R lediglich eine Polstelle für $\zeta = 1/a = 2+\sqrt{2}$ besitzt und diese in der rechten Halbebene von $\mathbb{C}$ liegt, ist R in der linken Halbebene holomorph. Da R auf dem Rand dieser Halbebene (der imaginären Achse) durch Eins beschränkt ist, ist R nach dem Maximumprinzip für holomorphe Funktionen[5] in der ganzen linken Halbebene durch Eins beschränkt. Also ist `ode23s` nach Proposition 77.6 A-stabil. Die L-Stabilität folgt aus der Tatsache, daß der Nennergrad von R größer ist als der Zählergrad, also R für $|\zeta| \to \infty$ gegen 0 konvergiert. □

Man beachte, daß dieses letzte Resultat unter der Voraussetzung bewiesen wurde, daß die Jacobi-Matrix J in jedem Schritt exakt ausgerechnet wird. Eine inexakte Jacobi-Matrix J beeinflußt also nicht die Konsistenzordnung aber möglicherweise das Stabilitätsverhalten von `ode23s`.

81 Schrittweitensteuerung

Bei der Implementierung eines Einschrittverfahrens ist es unerläßlich, sich über eine optimale Wahl der Schrittweite h Gedanken zu machen. Um den Aufwand zu minimieren, sollte h natürlich so groß wie möglich sein: Je größer h ist, desto weniger Zeitschritte sind nötig, um ein Zeitintervall $[0, T]$ vollständig zu

[5] vgl. Ahlfors [2]

durchlaufen. Andererseits ist der Fehler $y_i - y(t_i)$ im Intervall $[0, T]$ bei einem Verfahren der Ordnung q nach Satz 76.10 von der Größe $O(h^q)$, also an die Größe von h gekoppelt. Schließlich haben wir für das implizite Euler-Verfahren in Korollar 75.3 gesehen, daß die Konstante in dieser O-Abschätzung sehr klein sein kann und zwar in Abhängigkeit von dem lokalen Verhalten der rechten Seite f. In solchen Situationen kann dann h natürlich größer gewählt werden, ohne eine vorgegebene Genauigkeit der Näherungslösung zu verletzen. Entsprechende Beobachtungen haben wir auch bei dem numerischen Beispiel 77.12 mit dem Verfahren `ode23s` gemacht.

Hieraus ergibt sich für die Praxis die Notwendigkeit, den Fehler der aktuellen Approximation zu schätzen, um gegebenenfalls Korrekturen an der Schrittweite vornehmen zu können oder um eine Näherung wegen mangelnder Genauigkeit zu verwerfen.

Solche Fehlerschätzer beruhen zumeist auf dem Ergebnis eines zweiten Verfahrens, dem sogenannten *Kontrollverfahren*. Das Kontrollverfahren ist in der Regel ein Verfahren *höherer Ordnung* mit vergleichbaren (oder besseren) Stabilitätseigenschaften. Wir bezeichnen mit $\widehat{y}_i$ die Näherungen des Kontrollverfahrens, während y_i weiterhin die Näherungen des Ausgangsverfahrens sind. Aufgrund der Konstruktion erwarten wir $\|\widehat{y}_i - y(t_i)\|_2 \ll \|y_i - y(t_i)\|_2$, so daß

$$\begin{aligned} \delta_i \;=\; \|y_i - \widehat{y}_i\|_2 &\lesseqgtr \|y_i - y(t_i)\|_2 \pm \|\widehat{y}_i - y(t_i)\|_2 \\ &\approx \|y_i - y(t_i)\|_2 \end{aligned} \tag{81.1}$$

ein plausibler *Fehlerschätzer* ist.

Natürlich sollte das Kontrollverfahren nicht die Kosten des Algorithmus in die Höhe treiben. Daher verwendet man meist Verfahren, die *dieselben* Steigungen $f(t_i + c_j h, \eta_j)$ (bzw. k_j im Rosenbrock-Fall) verwenden. Auf diese Weise erspart man sich zusätzliche Funktionsauswertungen. Für Runge-Kutta-Verfahren führen diese Überlegungen auf Schemata der Form

$$\begin{aligned} y_{i+1} &= y_i + h \sum_{j=1}^{s} b_j f(t_i + c_j h, \eta_j)\,, \\ \widehat{y}_{i+1} &= y_i + h \sum_{j=1}^{s} \widehat{b}_j f(t_i + c_j h, \eta_j)\,, \end{aligned} \tag{81.2}$$

mit neuen Koeffizienten $\widehat{b}_j$ für das Kontrollverfahren, wobei die Steigungen $f(t_i + c_j h, \eta_j)$ bei beiden Verfahren gleich sind und wie in (76.3) berechnet werden, also

$$\eta_j = y_i + h \sum_{\nu=1}^{s} a_{j\nu} f(t_i + c_\nu h, \eta_\nu)\,, \qquad j = 1, \ldots, s\,.$$

Initialisierung: q und $q+1$ seien die Konsistenzordnungen der beiden Verfahren, ϵ die Fehlertoleranz und $\tau \in [0.8, 0.9]$ sei beliebig gewählt

$\delta = \epsilon$
wähle Startschrittweite $h > 0$
`for` $i = 0, 1, 2, \ldots$ `do`
  `repeat` % bis der Zeitschritt erfolgreich war
    $h = \tau\big(\epsilon/\delta\big)^{1/(q+1)} h$
    $y_{i+1} = y_i + h \sum_{j=1}^{s} b_j f(t_i + c_j h, \eta_j)$
    $\widehat{y}_{i+1} = y_i + h \sum_{j=1}^{s} \widehat{b}_j f(t_i + c_j h, \eta_j)$
    $\delta = \|y_{i+1} - \widehat{y}_{i+1}\|$
  `until` $\delta \le \epsilon$
  $t_{i+1} = t_i + h$
`until` $t_{i+1} > T$ % end for

Algorithmus 81.1: Schrittweitensteuerung

Kontrollverfahren der Form (81.2) heißen *eingebettete Runge-Kutta-Verfahren*.

Für die Schrittweitensteuerung muß eine Fehlertoleranz $\epsilon > 0$ vorgeben werden mit dem Ziel, in jedem Zeitschritt die Ungleichung $\delta_{i+1} \le \epsilon$ zu erfüllen. Dabei ist $\delta_{i+1} = \|y_{i+1} - \widehat{y}_{i+1}\|_2$ der jeweilige Fehlerschätzwert (81.1). Ist die Ungleichung erfüllt, so sprechen wir von einem erfolgreichen Zeitschritt und die aktuelle Näherung y_{i+1} wird als solche akzeptiert. Andernfalls muß der Zeitschritt mit einer kleineren Schrittweite wiederholt werden. Nach einem erfolgreichen Zeitschritt muß schließlich noch eine geeignete Schrittweite für den nächsten Zeitschritt festgelegt werden.

Die Schrittweite wird in beiden Fällen nach dem gleichen Muster modifiziert. Angenommen, das Ausgangsverfahren hat die Ordnung q und das Kontrollverfahren die Ordnung $q+1$. Dann hat der lokale Fehler der beiden Verfahren im i-ten Zeitschritt die Form

$$y_{i+1} - y(t_{i+1}) = h^{q+1} w_i + O(h^{q+2}), \qquad \widehat{y}_{i+1} - y(t_{i+1}) = O(h^{q+2}),$$

mit einem allgemein unbekannten $w_i \in \mathbb{R}^d$. Der Fehlerschätzer ist also ungefähr durch

$$\delta_{i+1} = \delta_{i+1}(h) = \|y_{i+1} - \widehat{y}_{i+1}\|_2 = \|h^{q+1} w_i + O(h^{q+2})\|_2 \approx h^{q+1} \|w_i\|_2$$

gegeben. Mit Schrittweite $\widetilde{h}$ anstelle von h würde sich entsprechend

$$\delta_{i+1}(\widetilde{h}) \approx \widetilde{h}^{q+1} \|w_i\|_2 = \big(\widetilde{h}/h\big)^{q+1} h^{q+1} \|w_i\|_2 \approx \big(\widetilde{h}/h\big)^{q+1} \delta_{i+1}(h)$$

ergeben. Folglich ist

$$\widetilde{h} = \tau \left(\frac{\epsilon}{\delta_{i+1}(h)}\right)^{1/(q+1)} h \tag{81.3}$$

mit $\tau = 1$ die größtmögliche Schrittweite, für die noch $\delta_{i+1}(\widetilde{h}) \lesssim \epsilon$ gilt. In diesem Sinn liefert $\widetilde{h}$ einen optimalen Kompromiß aus Genauigkeit und (zukünftigem) Rechenaufwand. Daher ersetzt man für die nachfolgenden Zeitschritte die alte Schrittweite h durch $\widetilde{h}$ aus (81.3), wobei der Toleranzparameter τ in (81.3) in der Praxis meist ein wenig kleiner als 1 gewählt wird, etwa $\tau = 0.8$ oder 0.9.

Bemerkung. Da die skizzierte Vorgehensweise auf einer Schätzung des *lokalen Fehlers* beruht, wird die Fehlerfortpflanzung bei dieser Vorgehensweise nicht berücksichtigt. Dies bedeutet, daß das Einschrittverfahren durchaus am Intervallende T einen Fehler oberhalb der vorgegebenen Fehlertoleranz haben kann. Eine gute Faustregel ist etwa der Schätzwert $T\epsilon$ für den maximalen Fehler in dem gesamten Zeitintervall. ◊

Die meisten Runge-Kutta-Verfahren sind so konstruiert, daß ihre Ordnung angesichts der Stufenzahl s größtmöglich ist. Dies bedeutet, daß das Kontrollverfahren – welches ja die gleichen Steigungen verwenden soll, vgl. (81.2) – in der Regel nur dann eine höhere Ordnung haben kann, wenn es zusätzliche Steigungen verwendet. Zusätzliche Steigungen bedeuten aber zusätzliche Funktionsauswertungen, also zusätzlichen Aufwand.

Einen Ausweg aus diesem Problem bietet der sogenannte *Fehlberg-Trick*: Fehlberg hat vorgeschlagen, als zusätzliche Stufe die erste Stufe des folgenden Zeitschritts zu verwenden (in der Regel ist das die Stufe $t_{i+1} = t_i + h$, wobei h noch die alte Schrittweite ist). Man beachte, daß die zugehörige Steigung bei einem erfolgreichen Zeitschritt ohnehin berechnet wird. Lediglich wenn die Berechnung von y_{i+1} aufgrund einer überschrittenen Fehlertoleranz mit kleinerer Schrittweite wiederholt werden muß, führt der Fehlberg-Trick zu einem zusätzlichen Aufwand. Üblicherweise ist allerdings die Mehrheit der Zeitschritte erfolgreich, so daß mit dieser Technik das Kontrollverfahren lediglich einen geringfügigen Mehraufwand mit sich bringt.

Beispiel 81.1 (klassisches Runge-Kutta-Verfahren). Das klassische Runge-Kutta-Verfahren mit $s = q = 4$ aus Beispiel 76.7 ist durch das Tableau

0				
1/2	1/2			
1/2	0	1/2		
1	0	0	1	
	1/6	1/3	1/3	1/6

gegeben. Wir wollen im folgenden dieses Verfahren als Kontrollverfahren wählen und ein eingebettetes Verfahren *dritter Ordnung* suchen, das mit denselben Steigungen auskommt.

Die Gewichte des Kontrollverfahrens lauten hier also

$$\widehat{b}_1 = 1/6\,,\ \widehat{b}_2 = 1/3\,,\ \widehat{b}_3 = 1/3\,,\ \widehat{b}_4 = 1/6\,.$$

Nach Satz 76.6 müssen die Gewichte b_j, $j = 1, \ldots, 4$, des eingebetteten Verfahrens dritter Ordnung neben der grundlegenden Gleichung $\sum b_j = 1$ aus (76.1) noch die drei Gleichungen

$$\sum_{j=1}^{4} b_j c_j \;=\; \frac{1}{2}\,, \qquad \sum_{j=1}^{4} b_j c_j^2 \;=\; \frac{1}{3}\,, \qquad \sum_{j=1}^{4} b_j \sum_{\nu=1}^{4} a_{j\nu} c_\nu \;=\; \frac{1}{6}$$

erfüllen, also das folgende Gleichungssystem lösen:

$$\begin{bmatrix} 1 & 1 & 1 & 1 \\ 0 & 1/2 & 1/2 & 1 \\ 0 & 1/4 & 1/4 & 1 \\ 0 & 0 & 1/4 & 1/2 \end{bmatrix} \begin{bmatrix} b_1 \\ b_2 \\ b_3 \\ b_4 \end{bmatrix} \;=\; \begin{bmatrix} 1 \\ 1/2 \\ 1/3 \\ 1/6 \end{bmatrix}.$$

Man sieht sofort, daß die Matrix nicht singulär ist und das Gleichungssystem daher nur eine Lösung hat, nämlich gerade die Gewichte des klassischen Runge-Kutta-Verfahrens. Mit anderen Worten: Es gibt *kein* eingebettetes Einschrittverfahren dritter Ordnung – außer dem Kontrollverfahren selbst.

Auch hier liefert der Fehlberg-Trick einen Ausweg: Wir fügen künstlich die fünfte Stufe $1 \;\big|\; \frac{1}{6}\ \frac{1}{3}\ \frac{1}{3}\ \frac{1}{6}\ 0$ in das Runge-Kutta-Tableau ein, setzen also

$$\widehat{b}_5 = 0 \qquad \text{und} \qquad \eta_5 \;=\; \widehat{y}_{i+1} \;=\; y_i + h \sum_{j=1}^{4} \widehat{b}_j f(t_i + c_j h, \eta_j)\,.$$

Die Ordnungsbedingungen führen dann in entsprechender Weise auf das um eine Spalte erweiterte Gleichungssystem für die Gewichte b_j, $j = 1, \ldots, 5$:

$$\begin{bmatrix} 1 & 1 & 1 & 1 & 1 \\ 0 & 1/2 & 1/2 & 1 & 1 \\ 0 & 1/4 & 1/4 & 1 & 1 \\ 0 & 0 & 1/4 & 1/2 & 1/2 \end{bmatrix} \begin{bmatrix} b_1 \\ b_2 \\ b_3 \\ b_4 \\ b_5 \end{bmatrix} \;=\; \begin{bmatrix} 1 \\ 1/2 \\ 1/3 \\ 1/6 \end{bmatrix}.$$

Die zusätzliche Spalte verändert nicht den Rang der Matrix und der Vektor $\widehat{b} = [\widehat{b}_1, \ldots, \widehat{b}_5]^T$ ist eine Lösung des Gleichungssystems. Alle anderen Lösungen ergeben sich durch Addition eines nichttrivialen Vertreters aus dem (eindimensionalen) Nullraum dieser Matrix, der durch den Vektor $[0, 0, 0, 1, -1]^T$ aufgespannt wird.

Ein möglicher Parametersatz für (81.2) ist daher

$$b = [1/6, 1/3, 1/3, 0, 1/6]^T\,, \qquad \widehat{b} = [1/6, 1/3, 1/3, 1/6, 0]^T\,.$$

Der Fehlerschätzer (81.1) macht natürlich nur dann Sinn, wenn das eingebettete Verfahren selber *kein* Verfahren vierter Ordnung ist. Dies ergibt sich mit einer etwas aufwendigeren Rechnung analog zu Beispiel 77.4, indem man zeigt, daß die Stabilitätsfunktion des eingebetteten Verfahrens durch

$$R(\zeta) \;=\; 1 + \zeta + \tfrac{1}{2}\zeta^2 + \tfrac{1}{6}\zeta^3 + \tfrac{1}{36}\zeta^4 + \tfrac{1}{144}\zeta^5$$

gegeben ist. Da der Koeffizient vor ζ^4 nicht der Koeffizient $1/24$ der Exponentialreihe ist, hat das eingebettete Runge-Kutta-Verfahren nach Satz 77.5 bestenfalls Konsistenzordnung $q = 3$. Aus (81.1) ergibt sich schließlich der Fehlerschätzer

$$\delta_{i+1} \;=\; \|y_{i+1} - \widehat{y}_{i+1}\|_2 \;=\; \frac{h}{6}\,\|f(t_i + h, \eta_5) - f(t_i + h, \eta_4)\|_2\,. \qquad \Diamond$$

Beispiel 81.2 (`ode23s`). Der MATLAB-Code `ode23s` verwendet als Kontrollverfahren ein Verfahren dritter Ordnung, das ebenfalls den Fehlberg-Trick verwendet. Bei einem s-stufigen Rosenbrock-Typ-Verfahren entspricht dies der Einführung einer zusätzlichen Steigung k_{s+1}, die durch die Gleichung

$$(I - ahJ)k_{s+1} \;=\; f(y_{i+1}) - hJ \sum_{\nu=1}^{s} d_{s+1,\nu} k_\nu$$

mit geeigneten Parametern $d_{s+1,\nu}$, $\nu = 1, \ldots, s$, gegeben ist. Bei `ode23s` sind es somit drei Gleichungen

$$\begin{aligned}
(I - ahJ)k_1 &= f(y_i)\,,\\
(I - ahJ)k_2 &= f(y_i + h\,\tfrac{1}{2}k_1) - ahJk_1\,,\\
(I - ahJ)k_3 &= f(y_i + hk_2) - d_{31}hJk_1 - d_{32}hJk_2
\end{aligned}$$

mit $a = 1/(2+\sqrt{2})$ und den neuen Parametern $d_{31} = -(4+\sqrt{2})/(2+\sqrt{2})$ und $d_{32} = (6+\sqrt{2})/(2+\sqrt{2})$. Nach wie vor ist $y_{i+1} = y_i + hk_2$ die berechnete Näherung für $y(t_{i+1})$, während das Kontrollverfahren durch

$$\widehat{y}_{i+1} = y_i \;+\; \frac{h}{6}\,(k_1 + 4k_2 + k_3)$$

gegeben ist. Für die folgenden Rechnungen nehmen wir der Einfachheit halber an, daß durchweg die exakte Jacobi-Matrix $J = f_y(y_i)$ verwendet wird. Selbst für eine davon abweichende Matrix J bleibt das Kontrollverfahren jedoch ein Verfahren dritter Ordnung.

Wie im Beweis von Satz 80.3 erhalten wir aus (80.6)

$$k_1 = f + ahf_y k_1 = f + ahf_y f + a^2h^2 f_y^2 f + O(h^3)$$

und entsprechend durch rekursives Einsetzen von k_2 (unter Berücksichtigung, daß f von t unabhängig ist)

$$\begin{aligned}
k_2 &= f + h\tfrac{1}{2}f_y k_1 + h^2\tfrac{1}{8}k_1^* f_{yy} k_1 - ahf_y k_1 + ahf_y k_2 + O(h^3) \\
&= f + h\tfrac{1}{2}f_y(f + ahf_y f) + h^2\tfrac{1}{8}f^* f_{yy} f - ahf_y(f + ahf_y f) \\
&\quad + ahf_y k_2 + O(h^3) \\
&= f + h\tfrac{1}{2}f_y(f + ahf_y f) + h^2\tfrac{1}{8}f^* f_{yy} f - ahf_y(f + ahf_y f) \\
&\quad + ahf_y\big(f + h\tfrac{1}{2}f_y f \underbrace{-ahf_y f + ahf_y k_2}_{=\,O(h^2)}\big) + O(h^3) \\
&= f + h\tfrac{1}{2}f_y f + h^2\big((a - a^2)f_y^2 f + \tfrac{1}{8}f^* f_{yy} f\big) + O(h^3)\,.
\end{aligned}$$

Schließlich ist

$$\begin{aligned}
k_3 &= f + hf_y k_2 + h^2\tfrac{1}{2}k_2^* f_{yy} k_2 - d_{31}hf_y k_1 - d_{32}hf_y k_2 \\
&\quad + ahf_y k_3 + O(h^3) \\
&= f + hf_y(f + h\tfrac{1}{2}f_y f) + h^2\tfrac{1}{2}f^* f_{yy} f - d_{31}hf_y(f + ahf_y f) \\
&\quad - d_{32}hf_y(f + h\tfrac{1}{2}f_y f) + ahf_y k_3 + O(h^3) \\
&= f + h(1 - d_{31} - d_{32})f_y f \\
&\quad + h^2\big((\tfrac{1}{2} - d_{31}a - \tfrac{1}{2}d_{32})f_y^2 f + \tfrac{1}{2}f^* f_{yy} f\big) + ahf_y k_3 + O(h^3)\,.
\end{aligned}$$

Damit ergibt sich durch rekursives Einsetzen

$$\begin{aligned}
k_3 &= f + h(1 - d_{31} - d_{32})f_y f \\
&\quad + h^2\big((\tfrac{1}{2} - d_{31}a - \tfrac{1}{2}d_{32})f_y^2 f + \tfrac{1}{2}f^* f_{yy} f\big) \\
&\quad + ahf_y\big(f + h(1 - d_{31} - d_{32})f_y f + ahf_y k_3\big) + O(h^3) \\
&= f + h(1 - d_{31} - d_{32} + a)f_y f \\
&\quad + h^2\big((\tfrac{1}{2} - 2d_{31}a - d_{32}a - \tfrac{1}{2}d_{32} + a)f_y^2 f + \tfrac{1}{2}f^* f_{yy} f\big) \\
&\quad + h^2a^2 f_y^2 k_3 + O(h^3) \\
&= f + h(1 - d_{31} - d_{32} + a)f_y f \\
&\quad + h^2\big((\tfrac{1}{2} - 2d_{31}a - d_{32}a - \tfrac{1}{2}d_{32} + a + a^2)f_y^2 f + \tfrac{1}{2}f^* f_{yy} f\big) \\
&\quad + O(h^3)\,.
\end{aligned}$$

Wegen $d_{31} + d_{32} = 2a$ und $\frac{1}{2} - d_{31}a - \frac{1}{2}d_{32} + a = 2a^2$ vereinfacht sich dies zu

$$\begin{aligned}
k_3 &= f + h(1 - a)f_y f \\
&\quad + h^2\big((\tfrac{1}{2} - d_{31}a - \tfrac{1}{2}d_{32} + a - a^2)f_y^2 f + \tfrac{1}{2}f^* f_{yy} f\big) + O(h^3) \\
&= f + h(1 - a)f_y f + h^2\big(a^2 f_y^2 f + \tfrac{1}{2}f^* f_{yy} f\big) + O(h^3)\,.
\end{aligned}$$

Somit ist

$$\begin{aligned}\widehat{y}_{i+1} &= y_i + \frac{h}{6}(k_1 + 4k_2 + k_3)\\ &= y_i + \frac{h}{6}\Big(6f + h(a+2+1-a)f_y f\\ &\qquad + h^2(a^2+4a-4a^2+a^2)f_y^2 f + h^2 f^* f_{yy} f\Big)\\ &\quad + O(h^4)\end{aligned}$$

und aus $4a - 2a^2 = 1$ folgt schließlich

$$\widehat{y}_{i+1} = y_i + hf + \tfrac{1}{2}h^2 f_y f + \tfrac{1}{6}h^3(f_y^2 f + f^* f_{yy} f) + O(h^4)\,.$$

Dies ist in der Tat die notwendige Potenzreihenentwicklung (76.6) für ein Verfahren dritter Ordnung (bei autonomer rechter Seite f).

An dieser Stelle sei jedoch darauf hingewiesen, daß das eingebettete Verfahren für den Fehlerschätzer von `ode23s` *nicht* A-stabil ist, vgl. Aufgabe 15. Dies hat zur Folge, daß der Fehlerschätzer bei steifen Differentialgleichungen unter Umständen unnötig kleine Zeitschritte vorschreibt, man vergleiche etwa die numerischen Ergebnisse in Beispiel 103.1. ◊

Bei Rosenbrock-Typ-Verfahren bedeutet eine Schrittweitenänderung auch immer, daß sich die Matrix $I - ahJ$ ändert, selbst wenn die alte Näherung J von $f_y(y_i)$ weiter benutzt werden soll. In dem Fall muß im zweiten Schritt von Algorithmus 80.1 eine neue LR-Zerlegung berechnet werden. Um diese Berechnung zu sparen, wird man bei Rosenbrock-Typ-Verfahren die Schrittweite nur dann modifizieren, wenn J ohnehin neu berechnet werden muß oder wenn die Fehlertoleranz ϵ deutlich unter- oder überschritten wurde.

Nehmen wir an, das Rosenbrock-Typ-Verfahren hat die Ordnung q und das Kontrollverfahren für den Fehlerschätzer die Ordnung $q+1$, unabhängig von der Wahl von J – wie bei `ode23s`. Die Kontrolle über die Jacobi-Matrix könnte nun anhand eines zweiten eingebetteten Verfahrens mit Näherungen $\check{y}_i$ erfolgen, welches nur mit exakter Jacobi-Matrix die Ordnung $q+1$ hat. Häufig ist es dann so, daß ein solches Verfahren zumindest noch die Ordnung q hat, solange $J = f_y(t_i, y_i) + O(h)$ gilt, also etwa, wenn J die exakte Jacobi-Matrix des vorigen Zeitschritts ist. Ist auch diese Eigenschaft verletzt, ist die Ordnung in der Regel kleiner als q.

Bei exakter Jacobi-Matrix erwartet man daher $\|\check{y}_i - \widehat{y}_i\| \ll \|y_i - \widehat{y}_i\|$. Bei darauffolgenden Zeitschritten gilt wenigstens noch $J = f_y + O(h)$ und es ergibt sich $\|\check{y}_i - \widehat{y}_i\| \approx \|y_i - \widehat{y}_i\|$, bis irgendwann $\|\check{y}_i - \widehat{y}_i\| \gg \|y_i - \widehat{y}_i\|$ ist. Daher liegt es nahe, die Jacobi-Matrix in dem Moment neu zu berechnen, in dem $\|\check{y}_i - \widehat{y}_i\| > \|y_i - \widehat{y}_i\|$ wird.

82 Differential-algebraische Gleichungen

Zum Abschluß wenden wir uns differential-algebraischen Gleichungen zu, also Differentialgleichungen mit algebraischen Nebenbedingungen:

$$\begin{aligned} y' &= f(y,z)\,, \qquad y(0) = y_0\,, \\ 0 &= g(y,z)\,, \qquad z(0) = z_0\,. \end{aligned} \tag{82.1}$$

Dabei sind $y \in \mathbb{R}^d$ und $z \in \mathbb{R}^p$ gesuchte Funktionen der Zeit t, während $f : \mathbb{R}^{d+p} \to \mathbb{R}^d$ und $g : \mathbb{R}^{d+p} \to \mathbb{R}^p$ gegeben sind. Damit eine Lösung des Systems (82.1) existieren kann, müssen die Anfangswerte *konsistent* sein, das heißt es gilt

$$g(y_0, z_0) = 0\,. \tag{82.2}$$

Ferner soll im weiteren angenommen werden, daß f und g stetig differenzierbar sind. Abhängig davon, ob die partielle Ableitungsmatrix $g_z(y_0, z_0) \in \mathbb{R}^{p\times p}$ invertierbar ist oder nicht, nennt man (82.1) ein differential-algebraisches System mit *Index Eins* oder *Index größer als Eins*.[6]

Beispiele. Das Gleichungssystem (64.3) für den Verstärkerschaltkreis aus Abschnitt 64 kann mit Hilfe der Variablensubstitution $y_1 = u_B$, $y_2 = u_C - u_L$, $y_3 = u_E$, $z = u_C - u_B$, in die Form (82.1) gebracht werden:

$$\begin{aligned} y_1' &= -\Big(\frac{1}{c_1 r_2} + \frac{1}{c_1 r_1}\Big) y_1 - \iota_B/c_1 + \frac{u_0}{c_1 r_1} + u_S'\,, \\ y_2' &= -\frac{1}{c_2 r_C}(y_1 + z) - \iota_C/c_2 + \frac{u_0}{c_2 r_C}\,, \\ y_3' &= -\frac{1}{c_E r_E} y_3 + (\iota_B + \iota_C)/c_E\,, \\ 0 &= \Big(\frac{1}{r_C} + \frac{1}{r_L}\Big)(y_1 + z) - \frac{1}{r_L} y_2 + \iota_C - u_0/r_C\,. \end{aligned}$$

Dabei hängen ι_B und ι_C ebenfalls von den gesuchten Spannungen ab: aus (64.6) erhalten wir beispielsweise mit $u_{BE} = u_B - u_E = y_1 - y_3$ und $u_{BE} - u_{CE} = u_B - u_C = -z$ für ι_C die Darstellung

$$\iota_C = \gamma(e^{(y_1 - y_3)/u_*} - 1) - \delta(e^{-z/u_*} - 1)$$

mit positiven Konstanten γ, δ und u_*. Für die Untersuchung des Index beachten wir, daß

$$g_z = \frac{1}{r_C} + \frac{1}{r_L} + \frac{\partial \iota_C}{\partial z} = \frac{1}{r_C} + \frac{1}{r_L} + \frac{\delta}{u_*} e^{-z/u_*}$$

[6] Für eine formale Definition des (*Differentiations-*)*Index* eines differential-algebraischen Systems sei auf die Literatur verwiesen, etwa auf [45] oder das Buch von Ascher und Petzold [6].

positiv ist und das Beispiel daher Index Eins besitzt.

Als zweites betrachten wir Beispiel 63.1. Durch Substitution der Geschwindigkeit v ergibt sich unter Vernachlässigung der Reibung ($\mu = 0$) aus dem Modell (63.12) eine Gleichung erster Ordnung der Form (82.1),

$$\begin{aligned} mv' &= F + \lambda \operatorname{grad} g(x)\,, \\ x' &= v\,, \\ 0 &= g(x)\,, \end{aligned} \tag{82.3}$$

bei der der Vektor y aus der Geschwindigkeit v und dem Ort x besteht und der Lagrange-Parameter λ die algebraische Variable z ist. Da g nicht von λ abhängt, ist $g_\lambda = 0$ und somit hat die differential-algebraische Gleichung einen Index größer als Eins. ◊

Wir wollen zunächst annehmen, daß das System (82.1) Index Eins besitzt. Unter dieser Voraussetzung existiert nach dem Satz über implizite Funktionen eine eindeutig bestimmte Funktion $\psi : \mathcal{U} \to \mathbb{R}^p$ in einer Umgebung $\mathcal{U} \subset \mathbb{R}^d$ von y_0 mit

$$\psi(y_0) = z_0 \qquad \text{und} \qquad g(y, \psi(y)) = 0\,.$$

Mit Hilfe dieser Funktion kann (82.1) in das äquivalente System

$$y' = f(y, \psi(y))\,, \quad y(0) = y_0\,, \qquad z = \psi(y)\,, \tag{82.4}$$

überführt werden, dessen Lösbarkeit durch den Satz von Picard-Lindelöf garantiert wird.

Zur numerischen Lösung differential-algebraischer Systeme ist die Transformation (82.4) glücklicherweise nicht erforderlich, da Runge-Kutta-Verfahren unmittelbar auf das System (82.1) angewandt werden können. Um dies zu erkennen, führen wir zunächst einen Parameter $\varepsilon \neq 0$ ein und betrachten das modifizierte Problem

$$\begin{aligned} y' &= f(y,z)\,, \qquad y(0) = y_0\,, \\ \varepsilon z' &= g(y,z)\,, \qquad z(0) = z_0\,. \end{aligned} \tag{82.5}$$

Ausgehend von Näherungswerten $y_i \approx y(t_i)$ und $z_i \approx z(t_i)$ zum Zeitpunkt $t = t_i$ ergibt das Runge-Kutta-Verfahren angewandt auf (82.5) die Zwischenstellen

$$\begin{aligned} \eta_j &= y_i + h \sum_{\nu=1}^{s} a_{j\nu} f(\eta_\nu, \zeta_\nu)\,, \\ \varepsilon \zeta_j &= \varepsilon z_i + h \sum_{\nu=1}^{s} a_{j\nu} g(\eta_\nu, \zeta_\nu)\,, \end{aligned} \qquad j = 1, \ldots, s\,, \tag{82.6}$$

und anschließend die Näherungen

$$\begin{aligned} y_{i+1} &= y_i + h\sum_{j=1}^{s} b_j f(\eta_j,\zeta_j)\,, \\ \varepsilon z_{i+1} &= \varepsilon z_i + h\sum_{j=1}^{s} b_j g(\eta_j,\zeta_j)\,, \end{aligned} \tag{82.7}$$

für den nächsten Zeitpunkt.

Die Gleichungen für $\zeta_j \in \mathbb{R}^p$ in (82.6) können mit Hilfe der Matrizen

$$G = [g(\eta_1,\zeta_1),\ldots,g(\eta_s,\zeta_s)] \qquad \text{und} \qquad Z = [\zeta_1 - z_i,\ldots,\zeta_s - z_i]$$

aus $\mathbb{R}^{p\times s}$ in der Form

$$\varepsilon Z = hGA^* \tag{82.8}$$

geschrieben werden, wobei A die Koeffizientenmatrix des Runge-Kutta-Verfahrens ist. Sofern A invertierbar ist, kann dieses Gleichungssystem formal nach $G = \varepsilon Z A^{-*}/h$ aufgelöst werden und in (82.7) eingesetzt ergibt sich

$$z_{i+1} \;=\; z_i \;+\; \frac{h}{\varepsilon}\,Gb \;=\; z_i + ZA^{-*}b\,. \tag{82.9}$$

Lassen wir nun ε in (82.8) gegen Null gehen, so strebt G wegen der Nichtsingularität von A gegen die Nullmatrix und die Zwischenstellen streben gegen die Lösungen des nichtlinearen Gleichungssystems

$$\eta_j \;=\; y_i + h\sum_{\nu=1}^{s} a_{j\nu} f(\eta_\nu,\zeta_\nu)\,, \qquad 0 \;=\; g(\eta_j,\zeta_j)\,, \tag{82.10}$$

$j = 1,\ldots,s$. Hieraus berechnet sich dann

$$y_{i+1} \;=\; y_i + h\sum_{j=1}^{s} b_j f(\eta_j,\zeta_j)\,, \tag{82.11}$$

während z_{i+1} auch für $\varepsilon = 0$ durch die rechte Seite von (82.9) gegeben ist.

Satz 82.1. *Die Funktionen f und g seien hinreichend glatt, g_z sei überall invertierbar und g_z^{-1} gleichmäßig beschränkt sowie $\big(y(t), z(t)\big)$ die exakte Lösung von* (82.1) *über $[0,T]$. Ferner sei das Runge-Kutta-Verfahren (A,b,c) steifgenau und habe Ordnung q. Dann ist das Verfahren* (82.10), (82.11), (82.9) *für hinreichend kleine Schrittweiten $h > 0$ wohldefiniert und es gilt*

$$\|y(t_i) - y_i\|_2 \;\le\; O(h^q)\,, \qquad \|z(t_i) - z_i\|_2 \;\le\; O(h^q)\,,$$

für alle $t_i = ih \in [0,T]$.

Beweis. Wir zeigen, daß sich unter den genannten Bedingungen die gleichen Näherungen ergeben, wenn das Runge-Kutta-Verfahren (A, b, c) auf das transformierte System (82.4) angewandt wird und $z_i = \psi(y_i)$ für $i = 1, 2, \ldots$, gesetzt wird. Zusammen mit Satz 76.10 folgt dann unmittelbar die Behauptung.

Bei Anwendung des Runge-Kutta-Verfahrens auf das System (82.4) erhält man für die Zwischenstellen η_j die Gleichungen

$$\eta_j \;=\; y_i + h \sum_{\nu=1}^{s} a_{j\nu} f(\eta_\nu, \psi(\eta_\nu))\,, \qquad j = 1, \ldots, s\,,$$

und mit $\zeta_\nu = \psi(\eta_\nu)$ entspricht dies gerade dem Gleichungssystem (82.10). Somit sind alle Zwischenstellen und die Näherungen y_{i+1} für $y(t_{i+1})$ der beiden Verfahren gleich. Es verbleibt noch zu zeigen, daß die z-Approximationen übereinstimmen. Dazu beachten wir, daß für ein steifgenaues Runge-Kutta-Verfahren $b^* A^{-1} = e_s^*$ gilt (vgl. den Beweis von Proposition 79.2), und somit aus (82.9)

$$z_{i+1} \;=\; z_i + (\zeta_s - z_i) \;=\; \zeta_s \;=\; \psi(\eta_s)$$

folgt. Da bei steifgenauen Runge-Kutta-Verfahren zudem η_s und y_{i+1} übereinstimmen, ist $z_{i+1} = \psi(y_{i+1})$. □

Bemerkungen. Ergänzt man das Schema (82.10), (82.11) anstelle von (82.9) durch die Gleichung

$$g(y_{i+1}, z_{i+1}) = 0 \tag{82.12}$$

zur Spezifikation von z_{i+1}, so bleibt die Konsistenzordnung des Runge-Kutta-Verfahrens auch bei nicht steifgenauen Verfahren erhalten. Dies folgt aus dem Beweis des letzten Satzes.

Allerdings kann die Konsistenz (82.12) der neuen Näherungen (y_{i+1}, z_{i+1}) nur dann erzwungen werden, wenn die algebraischen Nebenbedingungen wie in (82.1) explizit gegeben sind. Für differential-algebraische Gleichungen der Form

$$Mx' = \phi(x)\,, \qquad x(0) = x_0\,, \tag{82.13}$$

mit einer singulären quadratischen Matrix M sind die algebraischen Nebenbedingungen implizit durch die Konsistenzbedingung

$$\phi(x) \in \mathcal{R}(M) \tag{82.14}$$

gegeben und lassen sich nicht durch einen so einfachen Trick wie in (82.12) gewährleisten. Steifgenaue Runge-Kutta-Verfahren können jedoch ähnlich wie

zuvor auf Systeme der Form (82.13) verallgemeinert werden, vgl. Aufgabe 16. Bei diesen Verfahren muß die Matrix M nicht faktorisiert werden, was von entscheidendem Vorteil ist, wenn M sehr groß und dünn besetzt ist, etwa bei den Beispielen aus Abschnitt 64. ◊

Die Entwicklung geeigneter numerischer Verfahren zur Lösung differential-algebraischer Gleichungen mit Index größer als Eins stellt gegenwärtig ein sehr aktives Forschungsgebiet dar, da in diesem Fall die oben beschriebenen Verfahren im allgemeinen nicht mehr mit der gleichen Konsistenzordnung konvergieren: Man beobachtet eine *Ordnungsreduktion* (siehe [45]).

Die Schwierigkeiten sollen im folgenden anhand der Differentialgleichung (82.3) illustriert werden (vgl. Beispiel 63.1), bei der die Kraft F der Einfachheit halber von Ort und Zeit unabhängig sein soll. Das einfachste steifgenaue Verfahren, das implizite Euler-Verfahren, führt bei diesem Beispiel im $(i+1)$-ten Zeitschritt auf das nichtlineare Gleichungssystem

$$\begin{aligned} mv_{i+1} &= mv_i + h\big(F + \lambda_{i+1} \operatorname{grad} g(x_{i+1})\big)\,, \\ x_{i+1} &= x_i + hv_{i+1}\,, \\ g(x_{i+1}) &= 0\,, \end{aligned} \tag{82.15}$$

dessen Jacobi-Matrix

$$\begin{bmatrix} mI & -h\lambda g''(x) & -h \operatorname{grad} g(x) \\ -hI & I & 0 \\ 0 & \operatorname{grad} g(x)^* & 0 \end{bmatrix}$$

für kleine h fast singulär ist. Entsprechend schwierig ist die numerische Lösung des nichtlinearen Gleichungssystems. Man beachte, daß diese Schwierigkeiten widerspiegeln, daß die differential-algebraische Gleichung einen Index größer als Eins besitzt, da ansonsten der $(3,3)$-Block der Jacobi-Matrix ungleich Null wäre.

Davon abgesehen ist jedoch auch die Konsistenzordnung des impliziten Euler-Verfahrens nicht mehr $q = 1$ wie bei gewöhnlichen Anfangswertaufgaben. Um dies zu sehen, rufen wir uns zunächst aus Abschnitt 63.3 in Erinnerung, daß die exakte Lösung von (82.3) außer der Nebenbedingung $g(x) = 0$ noch die *versteckte Nebenbedingung*

$$\frac{d}{dt}\, g(x) = v^* \operatorname{grad} g(x) = 0 \tag{82.16}$$

erfüllt, vgl. (63.6). Außerdem ist der exakte Lagrange-Parameter nach (63.10) mit $M = m$ durch

$$\lambda \;=\; -\,\frac{m\,v^*g''(x)v \;+\; F^*\operatorname{grad} g(x)}{\|\operatorname{grad} g(x)\|_2^2} \tag{82.17}$$

gegeben. Eingesetzt in die Differentialgleichung (82.3) folgt hieraus

$$v' \;=\; \frac{F}{m} \;-\; \frac{m\,v^*g''(x)v \;+\; F^*\operatorname{grad} g(x)}{m\,\|\operatorname{grad} g(x)\|_2^2}\,\operatorname{grad} g(x)\,. \tag{82.18}$$

Nun nehmen wir an, daß $v_i = v(t_i)$ und $x_i = x(t_i)$ die exakte Geschwindigkeit und Position des Massekörpers zum Zeitpunkt $t = t_i$ sind und verwenden für den nächsten Zeitschritt des impliziten Euler-Verfahrens den Ansatz

$$v_{i+1} \;=\; v_i + hv_i' + O(h^2)\,, \qquad x_{i+1} \;=\; x_i + hx_i' + \tfrac{1}{2}h^2x_i'' + O(h^3)\,.$$

Ferner schreiben wir im weiteren ∇_i für $\operatorname{grad} g(x_i)$ und $G_i = g''(x_i)$ für die Hesse-Matrix von g an der Stelle x_i. Eingesetzt in (82.15) ergibt dies

$$\begin{aligned}
mv_i \;+\; hmv_i' \;&=\; mv_i \;+\; hF \;+\; h\lambda_{i+1}\operatorname{grad} g\big(x_i + O(h)\big) \;+\; O(h^2)\\
&=\; mv_i \;+\; hF \;+\; h\lambda_{i+1}\nabla_i + O(h^2)\,,\\
x_i \;+\; hx_i' \;+\; \tfrac{1}{2}h^2x_i'' \;&=\; x_i \;+\; hv_i \;+\; h^2v_i' \;+\; O(h^3)\,,\\
0 \;&=\; g\big(x_i + hx_i' + \tfrac{1}{2}h^2x_i'' + O(h^3)\big)\\
&=\; g(x_i) \;+\; \nabla_i^*(hx_i' + \tfrac{1}{2}h^2x_i'') \;+\; \tfrac{1}{2}h^2x_i'^*G_ix_i' \;+\; O(h^3)\,,
\end{aligned}$$

und ein Vergleich der entsprechenden h-Potenzen liefert die Gleichungen

$$\lambda_{i+1}\nabla_i \;=\; mv_i' - F + O(h)\,, \tag{82.19a}$$

$$x_i' \;=\; v_i\,, \qquad\qquad x_i'' \;=\; 2v_i'\,, \tag{82.19b}$$

$$\nabla_i^*x_i' \;=\; 0\,, \qquad\qquad \nabla_i^*x_i'' \;=\; -x_i'^*G_ix_i'\,, \tag{82.19c}$$

für die Koeffizienten der asymptotischen Entwicklung der numerischen Lösung. Die erste Gleichung in (82.19c) ist dabei redundant, denn sie entspricht der versteckten Nebenbedingung (82.16). Ersetzt man in der zweiten Gleichung von (82.19c) x_i' und x_i'' gemäß (82.19b) und multipliziert (82.19a) von links mit ∇_i^*, so ergibt sich

$$\nabla_i^*v_i' \;=\; -\,\frac{1}{2}\,v_i^*G_iv_i\,, \qquad \lambda_{i+1} \;=\; \frac{\nabla_i^*(mv_i' - F)}{\|\nabla_i\|_2^2} \;+\; O(h)\,,$$

und daraus folgt schließlich das Ergebnis

$$\lambda_{i+1} \;=\; -\,\frac{(m/2)\,v_i^* G_i v_i \;+\; F^*\nabla_i}{\|\nabla_i\|_2^2} \;+\; O(h)\,,$$
$$v_i' \;=\; \frac{F}{m} \;-\; \frac{(m/2)\,v_i^* G_i v_i \;+\; F^*\nabla_i}{m\,\|\nabla_i\|_2^2}\,\nabla_i \;+\; O(h)\,.$$

Ein Vergleich mit (82.17) und (82.18) zeigt, daß beim impliziten Euler-Verfahren lediglich die Ortskoordinaten die korrekte Konsistenzordnung aufweisen, während die numerisch berechnete Geschwindigkeit allmählich abdriftet und der Lagrange-Parameter unabhängig von der Zeitschrittweite bereits nach dem ersten Schritt völlig daneben liegen kann.

Einen Ausweg bietet ein sogenanntes *halbexplizites* Euler-Verfahren, bei dem die neuen Ortskoordinaten x_{i+1} durch einen expliziten Euler-Schritt

$$\tilde{x}_{i+1} \;=\; x_i \;+\; h v_i \tag{82.20a}$$

und eine anschließende Projektion auf die Nebenbedingungsmenge

$$x_{i+1} \;=\; \tilde{x}_{i+1} - \alpha_i \operatorname{grad} g(\tilde{x}_{i+1})\,, \qquad g(x_{i+1}) = 0\,, \tag{82.20b}$$

bestimmt werden. Die neue Geschwindigkeit und der neue Lagrange-Parameter können anschließend über einen impliziten Schritt berechnet werden, für den nun allerdings nur noch die erste Gleichung aus (82.15) neue Information liefert. Als zweite Gleichung kann jedoch die versteckte Nebenbedingung (82.16) herangezogen werden. Mit diesen beiden Gleichungen führt das implizite Euler-Verfahren auf das symmetrische lineare Gleichungssystem

$$\begin{bmatrix} mI & \operatorname{grad} g(x_{i+1}) \\ \operatorname{grad} g(x_{i+1})^* & 0 \end{bmatrix} \begin{bmatrix} v_{i+1} \\ -h\lambda_{i+1} \end{bmatrix} = \begin{bmatrix} mv_i + hF \\ 0 \end{bmatrix}. \tag{82.20c}$$

Man beachte, daß dieses halbexplizite Verfahren gegenüber dem impliziten Euler-Verfahren den Vorteil aufweist, daß die berechneten Approximationen (x_i, v_i, λ_i) nicht nur die Nebenbedingung $g(x_i) = 0$ sondern auch die versteckte Nebenbedingung (82.16) erfüllen. Darüber hinaus ist in jedem Zeitschritt lediglich eine skalare (und gut konditionierte) nichtlineare Gleichung für α_i zu lösen. Das lineare Gleichungssystem (82.20c) ist im allgemeinen auch gut konditioniert, die schlechte Kondition steckt wegen des Produkts $h\lambda_{i+1}$ in (82.20c) allein in der Berechnung des Lagrange-Parameters. Ein dritter Vorteil ist die Konsistenzordnung des Verfahrens (82.20):

Proposition 82.2. *Sei $g : \mathbb{R}^3 \to \mathbb{R}$ hinreichend glatt und $1/\|\operatorname{grad} g\|_2$ gleichmäßig beschränkt. Ist $\big(x(t), v(t), \lambda(t)\big)$ die exakte Lösung der Differentialgleichung* (82.3) *über $[0,T]$, dann erfüllen die Näherungen des halbexpliziten Verfahrens* (82.20) *die Abschätzungen*

$$\|x_i - x(t_i)\|_\infty \le O(h)\,, \qquad \|v_i - v(t_i)\|_\infty \le O(h)\,, \qquad \|\lambda_i - \lambda(t_i)\|_\infty \le O(h)\,,$$

für alle $t_i \in [0,T]$.

Beweisskizze. Der Beweis ist ähnlich strukturiert wie die vorangegangenen Konvergenzbeweise für Runge-Kutta-Verfahren zur Lösung gewöhnlicher Differentialgleichungen. Ein erheblicher Unterschied besteht jedoch darin, daß der lokale Fehler lediglich für die Orts- und Geschwindigkeitsvektoren die Größenordnung $O(h^2)$ aufweist. Um die Beweisidee herauszustellen, beschränken wir uns im folgenden auf formale asymptotische Entwicklungen der einzelnen Größen und verzichten auf eine detaillierte Abschätzung der Terme höherer Ordnung.

1. Lokaler Fehler: Sei $x_i = x(t_i)$ ein Punkt auf der Lösungskurve und $v_i = v(t_i)$ die zugehörige Geschwindigkeit des Massepunkts. Wir verwenden im weiteren wieder die Abkürzungen $\nabla_i = \operatorname{grad} g(x_i)$ und $G_i = g''(x_i)$. Aus (82.20a) und (82.20b) folgt

$$0 \;=\; g\big(x_i + hv_i - \alpha_i \operatorname{grad} g(x_i + hv_i)\big) \;=:\; \psi(h, \alpha_i)\,,$$

und diese Gleichung definiert implizit eine Funktion $\alpha_i = \alpha_i(h)$ mit $\alpha_i(0) = 0$. Nach dem Satz über implizite Funktionen ist

$$\frac{d\alpha_i}{dh}(0) \;=\; -\,\frac{\psi_h(0,0)}{\psi_\alpha(0,0)} \;=\; -\,\frac{\nabla_i^* v_i}{\|\nabla_i\|_2^2}$$

und letzteres ist gleich Null wegen der versteckten Nebenbedingung (82.16). Hinreichende Glattheit von g vorausgesetzt, beginnt demnach die Taylorentwicklung von $\alpha_i(h)$ um $h = 0$ erst mit dem Term zweiter Ordnung, d. h.

$$\alpha_i = O(h^2)\,, \qquad h \to 0\,.$$

Eingesetzt in (82.20b) folgt somit

$$x_{i+1} \;=\; \tilde{x}_{i+1} + O(h^2) \;=\; x_i + hv_i + O(h^2)\,, \tag{82.21}$$

d. h. die neue Ortskoordinate hat Konsistenzordnung $q = 1$.

Aus (82.20c) erhalten wir nun durch Multiplikation der ersten Gleichung mit $\nabla_{i+1} = \operatorname{grad} g(x_{i+1}) = \nabla_i + hG_iv_i + O(h^2)$ die Beziehung

$$\begin{aligned} m\nabla_{i+1}^* v_{i+1} - h\lambda_{i+1}\|\nabla_{i+1}\|_2^2 \;&=\; m\nabla_{i+1}^* v_i + hF^*\nabla_{i+1} \\ &=\; m\nabla_i^* v_i + hm\, v_i^* G_i v_i + hF^*\nabla_i + O(h^2) \end{aligned}$$

und wegen $\nabla_i^* v_i = \nabla_{i+1}^* v_{i+1} = 0$ folgt

$$\lambda_{i+1} \;=\; -\,\frac{m\,v_i^* G_i v_i + F^* \nabla_i}{\|\nabla_i\|_2^2} \;+\; O(h)\,. \tag{82.22}$$

Diese Entwicklung stimmt mit (82.17) überein, allerdings nicht mit Konsistenzordnung Eins, wie eine detailliertere Analyse zeigt. Dafür ergibt sich nun aus (82.20c) die Konsistenzordnung $q = 1$ für die berechnete Geschwindigkeit:

$$\begin{aligned} v_{i+1} \;&=\; v_i \;+\; \frac{h}{m} F \;+\; h\,\frac{\lambda_{i+1}}{m} \nabla_i \;+\; O(h^2) \\ &=\; v_i \;+\; \frac{h}{m} F \;-\; h\,\frac{m\,v_i^* G_i v_i + F^* \nabla_i}{m\,\|\nabla_i\|_2^2} \nabla_i \;+\; O(h^2)\,, \end{aligned} \tag{82.23}$$

vgl. (82.18).

2. Lokale Fehlerfortpflanzung: Für die Untersuchung des lokalen Fehlers wurde nur verwendet, daß für die Startnäherung (x_i, v_i) die Nebenbedingungen $g(x_i) = 0$ und $\operatorname{grad} g(x_i)^* v_i = 0$ gültig sind. Umgekehrt erfüllen alle berechneten Näherungen (x_i, v_i) des halbexpliziten Euler-Verfahrens (82.20) diese beiden Nebenbedingungen. Nehmen wir also an, (x_i, v_i) und $(\bar{x}_i, \bar{v}_i)$ seien zwei verschiedene Startnäherungen für die Rekursion (82.20), die in dem vorgenannten Sinn zulässig sind. Dann folgt aus (82.21) für die neuen Positionen x_{i+1} und $\bar{x}_{i+1}$, daß

$$\bar{x}_{i+1} - x_{i+1} \;=\; \bar{x}_i - x_i + h(\bar{v}_i - v_i) + O(h^2)\,.$$

Entsprechend folgt mit etwas mehr Aufwand aus (82.23), daß

$$\|\bar{v}_{i+1} - v_{i+1}\|_\infty \;\le\; \|\bar{v}_i - v_i\|_\infty \;+\; Ch\big(\|\bar{v}_i - v_i\|_\infty + \|\bar{x}_i - x_i\|_\infty\big) \;+\; O(h^2)$$

für ein gewisses $C > 0$.

3. Kumulierter Fehler: Da die Näherung λ_i des Lagrange-Parameters $\lambda(t_i)$ *nicht* in die Rekursionsvorschrift des Verfahrens (82.20) eingeht, können wir nun wie in früheren Beweisen (vergleiche etwa Satz 76.10) den globalen Fehler der Approximationen x_i und v_i, $i = 1, 2, \ldots$, durch einen Induktionsbeweis abschätzen. Dies führt auf die Behauptung des Satzes für die Ortskoordinaten und die Geschwindigkeitsvektoren. Die Genauigkeit von λ_i ergibt sich dann schließlich aus (82.22), denn aufgrund der bereits bewiesenen Behauptung folgt

$$\begin{aligned} \lambda_{i+1} \;&=\; -\,\frac{m\,v(t_i)^* g''\big(x(t_i)\big) v(t_i) \;+\; F^* \operatorname{grad} g\big(x(t_i)\big)}{\|\operatorname{grad} g\big(x(t_i)\big)\|_2^2} \;+\; O(h) \\ &=\; \lambda(t_i) + O(h) \;=\; \lambda(t_{i+1}) + O(h)\,, \end{aligned}$$

was zu zeigen war. □

Das Verfahren (82.20) läßt sich in naheliegender Weise auf die allgemeineren restringierten Mehrkörperprobleme aus Abschnitt 63.3 übertragen. Darüberhinaus kann mit Extrapolationsmethoden[7] die Konsistenzordnung erhöht werden. Diese Erweiterungen sind in dem Programmpaket MEXAX von Lubich et al. [72] implementiert worden.

[7]Das Prinzip der Extrapolation wird dabei ähnlich wie bei der Romberg-Quadratur verwendet.

Aufgaben

1. Gegeben sei das Anfangswertproblem

$$y' \,=\, f(t,y;a)\,, \qquad y(0) \,=\, y_0\,,$$

mit einem Parameter $a \in \mathbb{R}$. Die Funktion f sei stetig auf $\Omega = [0,T) \times \mathcal{J} \times \mathbb{R}$ und für alle kompakten Teilmengen $\mathcal{K} \subset \Omega$ gelte eine lokale Lipschitz-Bedingung der Form

$$\|f(t,y;a) - f(t,z;\tilde{a})\|_2 \,\le\, L_{\mathcal{K}} \|(y,a) - (z,\tilde{a})\|_2\,, \qquad (t,y,\hat{a}), (t,z,\tilde{a}) \in \mathcal{K}\,.$$

Zeigen Sie, daß unter diesen Voraussetzungen die Lösung des Anfangswertproblems stetig vom Parameter a abhängt.

Hinweis: Verwenden Sie eine geeignete Differentialgleichung für die Funktion $Y = \begin{bmatrix} y \\ a \end{bmatrix}$.

2. Sei $A \in \mathbb{R}^{n\times n}$, $\varphi \in \big(\mathcal{L}^2(0,T)\big)^n$ und $y_0 \in \mathbb{R}^n$. Überprüfen Sie, daß die Formel (78.6) aus der Variation der Konstanten unter diesen Voraussetzungen eine Funktion $y \in \big(H^1(0,T)\big)^n$ mit $y(0) = 0$ definiert, deren (komponentenweise gebildete) schwache Ableitung die Differentialgleichung

$$y' \,=\, Ay + \varphi(t)$$

im $\mathcal{L}^2$-Sinn erfüllt. Zeigen Sie darüber hinaus: Sind $y_1, y_2 \in \big(H^1(0,T)\big)^n$ zwei Lösungen dieser Differentialgleichung mit demselben Anfangswert $y_1(0) = y_2(0)$, so ist die Differenz $y_1 - y_2$ stetig differenzierbar und es gilt $y_1 = y_2$.

3. Untersuchen Sie die Fixpunktiteration zur Lösung der Rekursionsgleichung

$$y_{i+1} = y_i + hf(t_{i+1}, y_{i+1})$$

des impliziten Euler-Verfahrens. Zeigen Sie, daß die Fixpunktiteration konvergiert, falls f der Lipschitz-Bedingung

$$\|f(t,y) - f(t,z)\|_2 \,\le\, L\|y - z\|_2\,, \qquad y,z \in \mathbb{R}^d\,,\ t \in [0,T)\,,$$

für ein $L > 0$ genügt und die Schrittweite h kleiner als $1/L$ gewählt wird. Geben Sie eine Funktion f an, für die die Fixpunktiteration mit $h = 1/L$ divergiert.

4. Sei $G \in \mathbb{R}^{n\times n}$ symmetrisch und positiv definit sowie $\|y\|_G = (y^*Gy)^{1/2}$ die entsprechende Energienorm im $\mathbb{R}^n$. Ferner erfülle die stetig differenzierbare Funktion $f : [0,T) \times \mathbb{R}^n \to \mathbb{R}^n$ mit einem $l < 0$ die einseitige Lipschitz-Bedingung

$$\big(f(t,y) - f(t,z)\big)^* G(y-z) \,\le\, l\,\|y - z\|_G^2 \qquad \text{für alle } y,z \in \mathbb{R}^n\,.$$

Zeigen Sie, daß unter diesen Voraussetzungen für die Näherungen y_i des impliziten Euler-Verfahrens an die Lösung des Anfangswertproblems $y' = f(t,y)$, $y(0) = y_0$, die folgende Fehlerabschätzung gilt:

$$\|y_i - y(t_i)\|_G \,\le\, \max_{0\le t\le T} \|y''(t)\|_G \,\frac{h}{2|l|}\,.$$

5. Approximieren Sie die Lösung des Anfangswertproblems

$$l\,\theta'' = -g\sin\theta\,, \qquad \theta(0) = \pi/4\,,\ \theta'(0) = 0\,,$$

für die Auslenkung eines mathematischen Pendels (vgl. Aufgabe XI.6) mit dem expliziten und dem impliziten Euler-Verfahren und visualisieren Sie die Ergebnisse. Was fällt Ihnen auf? Implementieren Sie zum Vergleich das implizite Mittelpunktverfahren.

6. Geben Sie das Runge-Kutta-Tableau des Verfahrens von Heun an. Welche Konsistenzordnung hat dieses Verfahren?

7. Jede nicht-autonome Differentialgleichung $y' = f(t,y)$ läßt sich als autonomes System umformulieren, indem man $Y = [t, y^T]^T \in \mathbb{R}^{d+1}$ als neue Funktion einführt:

$$Y' = F(Y) \qquad \text{mit} \qquad F(Y) = \begin{bmatrix} 1 \\ f(Y) \end{bmatrix}.$$

Zeigen Sie: Ein Runge-Kutta-Verfahren liefert für beide Anfangswertprobleme $y' = f(t,y)$, $y(t_0) = y_0$, und $Y' = F(Y)$, $Y(t_0) = [t_0, y_0^T]^T$ die gleichen Näherungen $y_i \approx y(t_i)$.

8. Bestimmen Sie alle Bedingungen an die Koeffizienten eines Runge-Kutta-Verfahrens, damit sich für autonome Differentialgleichungen ein Verfahren vierter Ordnung ergibt.

9. Gegeben sei das Anfangswertproblem

$$y' = Ay + \varphi(t)\,, \quad y(0) = y_0\,,$$

mit einer diagonalisierbaren Matrix $A \in \mathbb{R}^{d\times d}$ und vektorwertigen Funktionen $\varphi, y \in \mathbb{R}^d$. Es seien v_k die Eigenvektoren und λ_k, $k = 1,\dots,d$, die Eigenwerte von A. Ferner bezeichnen

$$y_i = \sum_{k=1}^{d} \eta_{ik} v_k\,, \qquad i \in \mathbb{N}_0\,,$$

die Approximationen eines Runge-Kutta-Verfahrens mit Schrittweite $h > 0$. Schließlich sei noch für jedes $t > 0$

$$\varphi(t) = \sum_{k=1}^{d} \varphi_k(t) v_k$$

gesetzt. Zeigen Sie, daß sich die Komponenten $\{\eta_{ik} : i \in \mathbb{N}_0\}$ auch ergeben, wenn das gleiche Runge-Kutta-Verfahren auf das Anfangswertproblem

$$\eta' = \varphi_k(t) + \lambda_k \eta\,, \quad \eta(0) = \eta_{0k}\,,$$

angewandt wird.

10. Sei (A,b,c) das s-stufige Gauß-Verfahren oder das s-stufige Radau-IIA-Verfahren und $D \in \mathbb{R}^{s\times s}$ die Diagonalmatrix mit den Einträgen b_i/c_i, $i = 1,\dots,s$, auf der Diagonalen.
(a) Zeigen Sie, daß $DA + A^*D$ positiv definit ist.

(b) Folgern Sie daraus, daß alle Eigenwerte von A in der Halbebene $\mathbb{C}^+ = \{\lambda : \operatorname{Re}\lambda > 0\}$ liegen.
Hinweis zu (a): Drücken Sie für $x = [x_1, \ldots, x_s]^T$ den Ausdruck $x^*(DA + A^*D)x$ mit Hilfe des zugehörigen Interpolationspolynoms $p' \in \Pi_{s-1}$ mit $p'(c_j) = x_j$, $j = 1, \ldots, s$, aus.

11. Seien (A, b, c) die Koeffizienten des s-stufigen Gauß-Verfahrens. Berechnen Sie $b^*A^{-1}\mathbb{1}$ und $b^*A^{-2}\mathbb{1}$ und weisen Sie damit nach, daß

$$R(\zeta) \;=\; (-1)^s \;+\; 2(-1)^s s(s+1)\zeta^{-1} \;+\; O(|\zeta|^{-2})\,, \qquad |\zeta| \to \infty\,.$$

Hinweis: Aufgabe VI.9.

12. Implementieren Sie das zweistufige Gauß-Verfahren für das Räuber-Beute-Modell aus Beispiel 76.11. Variieren Sie die Schrittweite und vergleichen Sie Ihre Ergebnisse mit denen aus Abbildung 76.3.

13. Bestimmen Sie (für autonome Differentialgleichungen) die Konsistenzordnungen
(a) des linear-impliziten Mittelpunktverfahrens (80.4);
(b) des linear-impliziten Euler-Verfahrens (80.3).
Welche Ordnung liegt jeweils vor, wenn anstelle von f_y eine Approximation $J = f_y + O(h)$ verwendet wird; welche Ordnung liegt für ein beliebiges J anstelle von f_y vor?

14. Leiten Sie eine Variante des linear-impliziten Mittelpunktverfahrens für nicht-autonome Differentialgleichungen her. Transformiere Sie dazu zunächst die Differentialgleichung wie in Aufgabe 7 in ein autonomes System und wenden Sie darauf das linear-implizite Mittelpunktverfahren (80.4) an. Eliminieren Sie abschließend die Hilfsvariable für die Zeit.

15. (a) Beweisen Sie, daß die Stabilitätsfunktion eines s-stufigen Rosenbrock-Typ-Verfahrens (80.5) der Ordnung $q \geq s$ lediglich von dem Parameter a und der Stufenzahl abhängt. Zeigen Sie dazu, daß das Verfahren die gleiche Stabilitätsfunktion wie ein geeignetes linear-implizites Runge-Kutta-Verfahren hat und verwenden Sie Satz 77.5.
(b) Berechnen Sie die Stabilitätsfunktion des Fehlerschätzers von `ode23s` aus Beispiel 81.2 und überprüfen Sie die A-Stabilität dieses eingebetteten Verfahrens.

16. Gegeben sei eine differential-algebraische Gleichung der Gestalt (82.13),

$$Mx' = \phi(x)\,, \qquad x(0) \;=\; x_0\,,$$

mit einer singulären Matrix $M \in \mathbb{R}^{(d+p)\times(d+p)}$ mit Rang d sowie ein steifgenaues Runge-Kutta-Verfahren (A, b, c).
(a) Ersetzen Sie die Matrix M durch eine leicht gestörte nichtsinguläre Matrix M_ε und leiten Sie auf diese Weise eine Runge-Kutta-Variante für differential-algebraische Gleichungen der obigen Form her. Zeigen Sie, daß die Näherungen x_i dieses Verfahrens konsistent sind, also daß $\phi(x_i) \in \mathcal{R}(M)$.
(b) Wann hat ein System der obigen Gestalt Index Eins?
(c) Beweisen Sie die Aussage von Satz 82.1 für den Fall, daß die differential-algebraische Gleichung (82.13) Index Eins hat.
(d) Überlegen Sie sich entsprechend, wie das linear-implizite Euler-Verfahren auf differential-algebraische Gleichungen der obigen Form übertragen werden kann. Bestimmen Sie die Ordnung dieses Verfahrens.

Hinweis zu (b): Verwenden Sie die Singulärwertzerlegung von M und transformieren Sie so die Gleichung auf die Form (82.1).

17. Die Funktionen $f : \mathbb{R}^{d+p} \to \mathbb{R}^d$ und $g : \mathbb{R}^d \to \mathbb{R}^p$ seien hinreichend glatt und $g'(y)f_z(y,z)$ sei invertierbar. Dann hat das differential-algebraische System

$$y' = f(y,z)\,, \qquad 0 = g(y)\,,$$

mit konsistenten Anfangswerten $y(0) = y_0$ und $z(0) = z_0$ einen Index größer als Eins.
(a) Zeigen Sie, daß z unter diesen Voraussetzungen stetig differenzierbar ist mit

$$z' \;=\; -\big(g' f_z\big)^{-1}\big(g' f_y f + g''(f,f)\big)\,.$$

(b) Betrachten Sie die Approximationen

$$y_{i+1} \;=\; y_i + h f(y_{i+1}, z_{i+1})\,, \qquad g(y_{i+1}) = 0\,,$$

des impliziten Euler-Verfahrens. Zeigen Sie, daß unter der Annahme $y_i = y(t_i)$, $z_i = z(t_i)$ für die neue Näherung die asymptotische Entwicklung

$$z_{i+1} \;=\; z_i - h\big(g' f_z\big)^{-1}\big(g' f_y f + \tfrac{1}{2} g''(f,f)\big) + O(h^2)$$

gilt, wobei die Argumente auf der rechten Seite immer y_i und z_i sind.
Interpretieren Sie dieses Ergebnis.
Hinweis zu (b): Gehen Sie wie bei der Differentialgleichung (82.15) vor.

XV Randwertprobleme

Bei gewöhnlichen Differentialgleichungen zweiter Ordnung werden neben Anfangswertaufgaben oft auch Randwertprobleme betrachtet. Gängige Verfahren zur Lösung solcher Randwertaufgaben sind Differenzenverfahren und die Methode der finiten Elemente. Wir betrachten in diesem Kapitel nur Differenzenverfahren und behandeln die Methode der finiten Elemente im nachfolgenden Kapitel in größerer Allgemeinheit für elliptische Randwertaufgaben. Differenzenverfahren werden sehr ausführlich in dem Buch von Großmann und Roos [40] beschrieben. Schöne Darstellungen der Finite-Elemente-Methode für eindimensionale Probleme finden sich in den Büchern von Kreß [63] und von Quarteroni, Sacco und Saleri [86].

In den letzten beiden Abschnitten gehen wir noch auf Schießverfahren zur Lösung *nichtlinearer* Randwertprobleme ein.

83 Differenzenverfahren

Im weiteren seien (reellwertige) stetige Funktionen b, c und f der Variablen $x \in [0,1]$ gegeben; gesucht ist eine hinreichend glatte Funktion u mit

$$\begin{aligned} L[u] = -u'' + bu' + cu = f \qquad & \text{in } (0,1)\,, \\ u(0) = u(1) = 0\,. & \end{aligned} \tag{83.1}$$

An beiden Randpunkten liegt also eine Vorgabe an die Funktionswerte der Lösung vor, eine Dirichlet-Randbedingung. Man spricht daher von einem Randwertproblem. Es läßt sich zeigen (vgl. etwa Heuser [52, Kapitel VI]), daß dieses Randwertproblem eine eindeutige Lösung $u \in C^2[0,1]$ besitzt, falls die Funktion c nichtnegativ ist. Wir setzen also im weiteren

$$c(x) \geq 0 \qquad \text{für } x \in [0,1]$$

voraus.

Bei den numerischen Verfahren dieses Kapitels suchen wir Näherungen an die Funktionswerte von u auf äquidistanten Gittern

$$\triangle_h \;=\; \{\, x_i = ih \,:\, i = 1, \ldots, n-1,\; h = 1/n \,\} \;\subset\; (0,1)\,. \tag{83.2}$$

Mit $\boldsymbol{u} = \big[u(x_1), \ldots, u(x_{n-1})\big]^T \in \mathbb{R}^{n-1}$ bezeichnen wir den Vektor der Funktionswerte der exakten Lösung u von (83.1) auf $\triangle_h$. Die Randpunkte $x_0 = 0$ und $x_n = 1$ haben wir dabei sowohl bei dem Gitter als auch bei den Funktionswerten weggelassen, da die Lösung aufgrund der Randvorgaben (83.1) dort ohnehin durch Null fixiert ist.

Gesucht ist ein Näherungsvektor $\boldsymbol{u}_h = [u_1, \ldots, u_{n-1}]^T$ für $\boldsymbol{u}$. Zu diesem Zweck wird der Differentialoperator L aus (83.1) diskretisiert, indem die Ableitungen von u an den Stellen $x = x_i$ durch Differenzenquotienten ersetzt werden. Dabei stehen für die erste Ableitung im wesentlichen drei Alternativen zur Verfügung:

$$(i) \qquad D_h^+[u](x) \;=\; \frac{u(x+h) - u(x)}{h}\,,$$

$$(ii) \qquad D_h^-[u](x) \;=\; \frac{u(x) - u(x-h)}{h}\,,$$

$$(iii) \qquad D_h[u](x) \;=\; \frac{u(x+h) - u(x-h)}{2h}\,.$$

Die ersten beiden Näherungen heißen *einseitige Differenzenquotienten*, die Nähung (iii) ist ein *zentraler Differenzenquotient*. Die zweite Ableitung von u wird durch den zentralen Differenzenquotienten

$$D_h^2[u](x) \;=\; \frac{u(x+h) - 2u(x) + u(x-h)}{h^2} \;\approx\; u''(x) \tag{83.3}$$

approximiert.

Beispiel 83.1. Wir betrachten den einfachsten Fall in (83.1), und zwar $b(x) = c(x) = 0$, und ersetzen die zweite Ableitung u'' an allen Gitterpunkten von $\triangle_h$ durch den Differenzenquotienten (83.3). Die Güte dieser Näherung wird in Lemma 83.3 genauer untersucht. Unter Berücksichtigung der Randwerte $u(x_0) = u(x_n) = 0$ ergibt sich somit

$$\begin{bmatrix} f(x_1) \\ f(x_2) \\ \vdots \\ f(x_{n-1}) \end{bmatrix} = \begin{bmatrix} -u''(x_1) \\ -u''(x_2) \\ \vdots \\ -u''(x_{n-1}) \end{bmatrix} \approx h^{-2} \begin{bmatrix} 2 & -1 & & 0 \\ -1 & 2 & \ddots & \\ & \ddots & \ddots & -1 \\ 0 & & -1 & 2 \end{bmatrix} \begin{bmatrix} u(x_1) \\ u(x_2) \\ \vdots \\ u(x_{n-1}) \end{bmatrix}.$$

Wir schreiben kurz $\boldsymbol{f}$ für den Vektor aus $\mathbb{R}^{n-1}$ mit den Funktionswerten von f über $\triangle_h$, sowie L_h für die Matrix auf der rechten Seite (inklusive dem Vorfaktor

h^{-2}) und erhalten somit $L_h \boldsymbol{u} \approx \boldsymbol{f}$. Es ist nun naheliegend, die Lösung $\boldsymbol{u}_h$ des Gleichungssystems

$$L_h \boldsymbol{u}_h = \boldsymbol{f}$$

als Approximation von $\boldsymbol{u}$ anzusehen.

Die Eigenwerte der symmetrischen Matrix L_h sind nach Aufgabe VI.13 durch $4h^{-2}\sin^2(kh\pi/2)$ gegeben, wobei k von 1 bis $n-1$ läuft. Insbesondere ist L_h invertierbar. Wegen

$$\sin(kh\pi/2) \;\geq\; \frac{2}{\pi}\, kh\pi/2 \;=\; kh$$

ergibt sich die Abschätzung

$$\|L_h^{-1}\|_2 \;=\; \max_{1\leq k\leq n-1} \frac{h^2}{4\sin^2(kh\pi/2)} \;\leq\; \max_{1\leq k\leq n-1} \frac{h^2}{4k^2h^2} \;=\; \frac{1}{4}\,. \tag{83.4}$$

Folglich ist

$$\begin{aligned} \|\boldsymbol{u} - \boldsymbol{u}_h\|_2 \;&=\; \|L_h^{-1}(L_h\boldsymbol{u} - \boldsymbol{f})\|_2 \;\leq\; \|L_h^{-1}\|_2 \|L_h\boldsymbol{u} - \boldsymbol{f}\|_2 \\ &\leq\; \frac{1}{4}\,\|L_h\boldsymbol{u} - \boldsymbol{f}\|_2\,. \end{aligned} \tag{83.5}$$

◊

Dieses konkrete Beispiel demonstriert die allgemeine Vorgehensweise bei der Konvergenzanalyse eines Differenzenverfahrens, die aus zwei wesentlichen Komponenten besteht: einer Schranke für die Norm von L_h^{-1} und einer Abschätzung des Residuums $L_h\boldsymbol{u} - \boldsymbol{f}$; ersteres führt auf den Begriff der *Stabilität*, letzteres ist mit der *Konsistenz* des Differenzenverfahrens verknüpft.

Im Unterschied zu Beispiel 83.1 werden wir uns allerdings im weiteren auf stärkere Fehlerabschätzungen bezüglich der Maximumnorm konzentrieren, d. h. in der Ungleichungskette (83.5) wird die Euklidnorm des Residuums durch die Maximumnorm und die Spektralnorm von L_h^{-1} durch die Zeilensummennorm ersetzt.

Im weiteren werden die folgenden Fehlerabschätzungen verwendet.

Lemma 83.2. *Sei $u \in C^2[0,1]$ und $x \in [h, 1-h]$. Dann gelten für die einseitigen Differenzenquotienten die Abschätzungen*

$$\big|\, D_h^{\pm}[u](x) - u'(x) \,\big| \;\leq\; \frac{1}{2}\|u''\|_{[0,1]} h\,,$$

während der Approximationsfehler des zentralen Differenzenquotienten für u aus $C^3[0,1]$ und $x \in [h, 1-h]$ durch

$$\big|\, D_h[u](x) - u'(x) \,\big| \;\leq\; \frac{1}{6}\|u'''\|_{[0,1]} h^2$$

beschränkt ist.

Beweis. Wir beweisen lediglich die dritte Behauptung (für $u \in C^3[0,1]$): Durch Taylorentwicklung ergibt sich

$$\begin{aligned} u(x+h) &= u(x) + hu'(x) + \frac{1}{2}h^2u''(x) + \frac{1}{6}h^3u'''(\xi_+)\,, \\ u(x-h) &= u(x) - hu'(x) + \frac{1}{2}h^2u''(x) - \frac{1}{6}h^3u'''(\xi_-)\,, \end{aligned}$$

mit $x-h < \xi_- < x < \xi_+ < x+h$. Daraus folgt

$$u(x+h) - u(x-h) = 2hu'(x) + \frac{1}{6}h^3\left(u'''(\xi_+) + u'''(\xi_-)\right)$$

und Division durch $2h$ ergibt unmittelbar die Behauptung. Es ist nun offensichtlich, wie der Beweis abgeändert werden muß, um die anderen beiden Behauptungen nachzuweisen. □

Für den zentralen Differenzenquotienten für die zweite Ableitung gilt die folgende analoge Abschätzung.

Lemma 83.3. *Sei $u \in C^4[0,1]$. Dann ist für alle $x \in [h, 1-h]$*

$$\left| D_h^2[u](x) - u''(x) \right| \le \frac{1}{12}\,\|u''''\|_{[0,1]}h^2\,.$$

Beweis. Taylorentwicklung von u um x ergibt

$$\begin{aligned} u(x+h) &= u(x) + hu'(x) + \frac{1}{2}h^2u''(x) + \frac{1}{6}h^3u'''(x) + \frac{1}{24}h^4u''''(\xi_+)\,, \\ u(x-h) &= u(x) - hu'(x) + \frac{1}{2}h^2u''(x) - \frac{1}{6}h^3u'''(x) + \frac{1}{24}h^4u''''(\xi_-)\,, \end{aligned}$$

mit $x-h < \xi_- < x < \xi_+ < x+h$. Daraus folgt

$$u(x+h) + u(x-h) - 2u(x) = h^2u''(x) + \frac{1}{24}h^4\left(u''''(\xi_+) + u''''(\xi_-)\right)$$

und Division durch h^2 ergibt die Behauptung. □

Beispiel. Für das Residuum $L_h\boldsymbol{u} - \boldsymbol{f}$ aus Beispiel 83.1 folgt aus Lemma 83.3 unmittelbar die Abschätzung

$$\|L_h\boldsymbol{u} - \boldsymbol{f}\|_\infty \le \frac{1}{12}\,\|u''''\|_{[0,1]}\,h^2 = \frac{1}{12}\,\|f''\|_{[0,1]}\,h^2 \tag{83.6}$$

bezüglich der Maximumnorm, falls u in $[0,1]$ viermal stetig differenzierbar bzw. f zweimal stetig differenzierbar ist. ◊

Wir diskretisieren nun die allgemeine Randwertaufgabe (83.1). Dazu gehen wir zunächst wie in Beispiel 83.1 vor und ersetzen die zweite Ableitung u'' durch den Differenzenquotienten $D_h^2[u]$ aus (83.3). Entsprechend ersetzen wir die erste Ableitung $u'(x)$ durch einen der drei Differenzenquotienten $D_h^+[u]$, $D_h^-[u]$ oder $D_h[u]$. In jedem Fall wird die linke Seite von (83.1) durch ein Matrix-Vektor-Produkt $L_h\boldsymbol{u}$ diskretisiert mit einer Tridiagonalmatrix

$$L_h \;=\; h^{-2}\begin{bmatrix} d_1 & s_1 & & 0 \\ r_2 & d_2 & \ddots & \\ & \ddots & \ddots & s_{n-2} \\ 0 & & r_{n-1} & d_{n-1} \end{bmatrix} \in \mathbb{R}^{(n-1)\times(n-1)}\,. \tag{83.7}$$

In den verschiedenen Fällen ergeben sich die Einträge in den einzelnen Zeilen von L_h wie folgt:

$$\begin{aligned} D_h^+ :\qquad d_i &= \;\; 2 - hb(x_i) + h^2c(x_i)\,, \\ r_i &= -1\,, \\ s_i &= -1 + hb(x_i)\,, \end{aligned} \tag{83.8a}$$

$$\begin{aligned} D_h^- :\qquad d_i &= \;\; 2 + hb(x_i) + h^2c(x_i)\,, \\ r_i &= -1 - hb(x_i)\,, \\ s_i &= -1\,, \end{aligned} \tag{83.8b}$$

$$\begin{aligned} D_h :\qquad d_i &= \;\; 2 + h^2c(x_i)\,, \\ r_i &= -1 - hb(x_i)/2\,, \\ s_i &= -1 + hb(x_i)/2\,. \end{aligned} \tag{83.8c}$$

Wie in Beispiel 83.1 erhält man nun eine Näherungslösung $\boldsymbol{u}_h$ auf $\triangle_h$ durch Lösen des linearen Gleichungssystems

$$L_h\boldsymbol{u}_h = \boldsymbol{f}\,, \tag{83.9}$$

sofern selbiges lösbar ist (dieser Frage gehen wir im folgenden Abschnitt nach). Da L_h eine Tridiagonalmatrix ist, kann (83.9) leicht mit etwa $5n$ Multiplikationen durch Gauß-Elimination gelöst werden.

Wir wenden uns nun der Konsistenz dieser Differenzenapproximation zu.

Definition 83.4. Ein Differenzenverfahren hat die *Konsistenzordnung* q (bezüglich der Maximumnorm), wenn unter den genannten Voraussetzungen an b, c und f für jede hinreichend glatte Lösung u ein $C > 0$ existiert, so daß für alle $h > 0$

$$\|L_h\boldsymbol{u} - \boldsymbol{f}\|_\infty \le Ch^q\,.$$

Satz 83.5. *Die Lösung u des Randwertproblems* (83.1) *sei viermal stetig differenzierbar in* $[0,1]$*; dann hat das Differenzenverfahren* (83.7), (83.8) *die Konsistenzordnung $q = 2$, wenn die erste Ableitung durch den zentralen Differenzenquotienten D_h approximiert wird, beziehungsweise die Konsistenzordnung $q = 1$ bei Verwendung der einseitigen Differenzenquotienten D_h^+ oder D_h^-.*

Beweis. Wir zerlegen L_h in die drei Anteile $A_h + B_h + C_h$ für die drei Summanden aus (83.1). Insbesondere ist A_h die Matrix aus Beispiel 83.1 und C_h die Diagonalmatrix mit den Funktionswerten von c auf $\triangle_h$; lediglich B_h hängt von der Wahl des Differenzenquotienten zur Approximation der ersten Ableitung ab. Bezeichnen wir wieder mit $(Mz)_i$ die i-te Komponente des Matrix-Vektor-Produkts Mz, dann folgt aus Lemma 83.3

$$\big|\,(A_h \boldsymbol{u})_i + u''(x_i)\,\big| \;\le\; \frac{1}{12}\,\|u''''\|_{\,[0,1]}\,h^2\,.$$

Entsprechend ergibt sich aus Lemma 83.2 bei der Verwendung zentraler Differenzenquotienten die Abschätzung

$$\big|\,(B_h \boldsymbol{u})_i - b(x_i)u'(x_i)\,\big| \;\le\; |b(x_i)|\,\frac{1}{6}\,\|u'''\|_{\,[0,1]}\,h^2\,, \tag{83.10}$$

und daher ist

$$\begin{aligned}\|L_h\boldsymbol{u} - \boldsymbol{f}\|_\infty &= \|A_h\boldsymbol{u} + B_h\boldsymbol{u} + C_h\boldsymbol{u} - \boldsymbol{f}\|_\infty \\ &\le \Big(\frac{1}{12}\,\|u''''\|_{\,[0,1]} + \frac{1}{6}\,\|u'''\|_{\,[0,1]}\|b\|_{\,[0,1]}\Big)h^2\,.\end{aligned}$$

Damit ist für diesen Fall die Konsistenzordnung $q = 2$ nachgewiesen. Verwendet man einseitige Differenzen zur Approximation von u', dann verschlechtert sich die Abschätzung (83.10) gemäß Lemma 83.2 auf $O(h)$ anstelle von $O(h^2)$ und das Differenzenschema hat lediglich die Konsistenzordnung $q = 1$. □

Bemerkung 83.6. Inhomogene Dirichlet-Randbedingungen

$$u(0) = \alpha\,, \qquad u(1) = \beta\,,$$

können entsprechend behandelt werden. Anhand von Beispiel 83.1 macht man sich klar, daß in diesem Fall lediglich die erste und die letzte Gleichung des Gleichungssystems und auch hier nur die rechten Seiten abgeändert werden müssen. Speziell für Beispiel 83.1 lautet das zugehörige diskrete System dann etwa

$$L_h\boldsymbol{u}_h = \widetilde{\boldsymbol{f}} = \boldsymbol{f} + h^{-2}\big[\,\alpha, 0, \ldots, 0, \beta\,\big]^T\,.$$

Oft treten auch sogenannte *Neumann-Randbedingungen*

$$u'(0) = \alpha\,, \qquad u'(1) = \beta\,, \tag{83.11}$$

auf. Da nun $u(0)$ und $u(1)$ nicht mehr bekannt sind, werden die Randpunkte $x_0 = 0$ und $x_n = 1$ in das Gitter $\triangle_h$ mit aufgenommen. Bei den entsprechenden Gleichungen des Differenzenverfahrens ergeben sich dann allerdings auch Verweise auf Werte von u an Stellen außerhalb des Intervalls $[0, 1]$, nämlich auf $u(-h)$ und $u(1 + h)$. Diese Verweise können mit Hilfe eines der drei Differenzenquotienten $D_h^+[u]$, $D_h^-[u]$ oder $D_h[u]$ an den Stellen $x = 0$ bzw. $x = 1$ unter Verwendung der Randbedingungen (83.11) wieder eliminiert werden. Anhand des Modellproblems aus Beispiel 83.1 lassen sich die zwei meist verwendeten Varianten leicht erläutern. Im ersten Fall ersetzt man beispielsweise in (83.3) Verweise auf diese äußeren Punkte durch einseitige Differenzenquotienten, etwa

$$u(-h) \approx u(0) - hu'(0) = u(0) - \alpha h\,,$$

und erhält dann

$$\begin{aligned} -u''(0) &\approx \frac{-u(h) + 2u(0) - u(-h)}{h^2} \approx \frac{-u(h) + u(0) + \alpha h}{h^2} \\ &= \frac{u(x_0) - u(x_1)}{h^2} + \frac{\alpha}{h}\,; \end{aligned}$$

entsprechend ergibt sich

$$-u''(1) \approx \frac{u(x_n) - u(x_{n-1})}{h^2} - \frac{\beta}{h}\,.$$

Dies führt auf das lineare Gleichungssystem

$$h^{-2} \begin{bmatrix} 1 & -1 & & & 0 \\ -1 & 2 & -1 & & \\ & -1 & \ddots & \ddots & \\ & & \ddots & 2 & -1 \\ 0 & & & -1 & 1 \end{bmatrix} \begin{bmatrix} u_0 \\ u_1 \\ \vdots \\ u_{n-1} \\ u_n \end{bmatrix} = \boldsymbol{f} + h^{-1} \begin{bmatrix} -\alpha \\ 0 \\ \vdots \\ 0 \\ \beta \end{bmatrix}.$$

Die andere Möglichkeit ist die Verwendung des zentralen Differenzenquotienten für die Randableitungen:

$$\begin{aligned} u(1 + h) &\approx u(1 - h) + 2hu'(1) \approx u_{n-1} + 2\beta h\,, \\ u(-h) &\approx u(h) - 2hu'(0) \approx u_1 - 2\alpha h\,. \end{aligned}$$

Dies ergibt das lineare Gleichungssystem

$$h^{-2} \begin{bmatrix} 2 & -2 & & & 0 \\ -1 & 2 & -1 & & \\ & -1 & \ddots & \ddots & \\ & & \ddots & 2 & -1 \\ 0 & & & -2 & 2 \end{bmatrix} \begin{bmatrix} u_0 \\ u_1 \\ \vdots \\ u_{n-1} \\ u_n \end{bmatrix} = \boldsymbol{f} + h^{-1} \begin{bmatrix} -2\alpha \\ 0 \\ \vdots \\ 0 \\ 2\beta \end{bmatrix}.$$

Der Vorteil dieser Variante ist die höhere Konsistenzordnung; dafür ist bei der ersten Variante die Koeffizientenmatrix symmetrisch.

Man beachte, daß beide Möglichkeiten auf singuläre Gleichungssysteme führen; der Vektor $\mathbb{1} = [1, 1, \ldots, 1]^T$ liegt im Kern beider Matrizen. Dies hängt damit zusammen, daß die Lösung von $-u'' = f$, $u'(0) = \alpha$, $u'(1) = \beta$, nur bis auf additive Konstanten eindeutig bestimmt ist. Die Gleichungssysteme können daher (wie auch die Differentialgleichung) unlösbar sein. Wegen

$$u'(1) \;=\; u'(0) \;+\; \int_0^1 u''(x)\,dx$$

ist die Differentialgleichung nur dann lösbar, wenn die Bedingung

$$\alpha - \beta \;=\; \int_0^1 f(x)\,dx$$

erfüllt ist (vgl. Satz 89.7). In diesem Fall kann eine sinnvolle Näherungslösung berechnet werden, indem beispielsweise die erste Komponente des Lösungsvektors $\boldsymbol{u}_h$ durch Eins fixiert und aus dem Gleichungssystem eliminiert wird. Streicht man dann noch die erste Zeile des Systems, dann ergibt sich ein nichtsinguläres Gleichungssystem für die restlichen Komponenten von $\boldsymbol{u}_h$.

Es sei angemerkt, daß die Singularität des Gleichungssystems auch dadurch begründet ist, daß in dem betrachteten Beispiel die Koeffizientenfunktion c identisch verschwindet, vgl. Abschnitt 89.2. ◊

84 Stabilitätsabschätzungen

Ziel dieses Abschnitts sind Abschätzungen für $\|L_h^{-1}\|_\infty$, wobei L_h eine der Matrizen (83.7), (83.8) bezeichnen soll. Mit Hilfe des Konsistenzresultats aus Satz 83.5 ergeben sich dann unmittelbar Fehlerabschätzungen für $\boldsymbol{u}_h - \boldsymbol{u}$.

Definition 84.1. Ein Differenzenverfahren für das Randwertproblem (83.1) heißt *stabil* (bezüglich der diskreten Maximumnorm), wenn positive Konstanten C und h_0^* existieren, so daß für alle $0 < h < h_0^*$ die Matrix L_h invertierbar ist mit $\|L_h^{-1}\|_\infty \le C$.

Der Nachweis der Stabilität eines Differenzenschemas ist eng mit gewissen Monotonieeigenschaften der Matrix L_h verknüpft.

Definition 84.2. Ist $A = [a_{ij}] \in \mathbb{R}^{n\times n}$ invertierbar, $a_{ij} \le 0$ für $i \neq j$ und A^{-1} komponentenweise nichtnegativ (für letzteres wird im folgenden $A^{-1} \ge 0$ geschrieben), dann nennt man A eine *M-Matrix*.

Eine M-Matrix hat die angesprochene *Monotonieeigenschaft*: Sind $x, y \in \mathbb{R}^n$ und ist A eine M-Matrix, dann gilt (komponentenweise)

$$x \le y \qquad \Longrightarrow \qquad A^{-1}x \le A^{-1}y\,. \tag{84.1}$$

In der Regel ist es jedoch schwierig, einer Matrix A anzusehen, ob sie eine M-Matrix ist. Eine Ausnahme bilden die irreduziblen Tridiagonalmatrizen:

Satz 84.3. *Jede irreduzibel diagonaldominante Tridiagonalmatrix mit positiven Diagonalelementen und negativen Nebendiagonalelementen ist eine M-Matrix.*

Beweis. Sei $T = [t_{ij}]$ eine Tridiagonalmatrix, die die genannten Voraussetzungen erfüllt. Dann ist T nach Satz 23.3 invertierbar und es muß gezeigt werden, daß T^{-1} nichtnegativ ist. Dazu zerlegen wir $T = D - N$ in den Diagonal- und Nebendiagonalanteil; nach Voraussetzung ist $D \ge 0$ und $N \ge 0$. Für beliebiges $\varepsilon > 0$ ist $T + \varepsilon I$ strikt diagonaldominant (vgl. Definition 4.5) und es gilt

$$T + \varepsilon I = D + \varepsilon I - N = (D + \varepsilon I)(I - R)$$

mit $R = (D + \varepsilon I)^{-1}N$. Da die Diagonaleinträge der Diagonalmatrix $D + \varepsilon I$ allesamt positiv sind, ist $(D + \varepsilon I)^{-1} \ge 0$ und somit auch R eine nichtnegative Matrix. Zudem ist

$$\|R\|_\infty = \|(D + \varepsilon I)^{-1}N\|_\infty < 1\,,$$

da $T + \varepsilon I$ strikt diagonaldominant ist. Damit konvergiert die Neumannsche Reihe[1] in der folgenden Umformung

$$\begin{aligned}(T + \varepsilon I)^{-1} &= (I - R)^{-1}(D + \varepsilon I)^{-1} = \Big(\sum_{k=0}^{\infty} R^k\Big)(D + \varepsilon I)^{-1} \\ &= \sum_{k=0}^{\infty} R^k (D + \varepsilon I)^{-1}\,,\end{aligned}$$

und es folgt $(T+\varepsilon I)^{-1} \ge 0$, da alle Terme dieser Reihe nichtnegativ sind. Beim Grenzübergang $\varepsilon \to 0$ strebt schließlich $(T + \varepsilon I)^{-1} \to T^{-1}$, so daß auch die Grenzmatrix T^{-1} nichtnegativ ist. Damit ist der Satz vollständig bewiesen. □

Korollar 84.4. *Bei Verwendung des zentralen Differenzenquotienten $D_h[u](x_i)$ für $u'(x_i)$, $i = 1, \ldots, n-1$, ist die Koeffizientenmatrix L_h aus* (83.7) *für*

$$0 < h < h_0 = 2/\|b\|_{[0,1]} \tag{84.2}$$

eine M-Matrix.

[1] vgl. Heuser [51]

Beweis. Für $0 < h < h_0$ ist $|hb(x_i)/2| < 1$, und aus (83.8c) folgt, daß für diese Werte von h die Nebendiagonalelemente von L_h negativ sind und die Summe ihrer Beträge $2h^{-2}$ ergibt. Demnach ist L_h irreduzibel diagonaldominant mit positiven Diagonaleinträgen, da c nichtnegativ angenommen wurde (falls c strikt positiv ist, ist L_h sogar strikt diagonaldominant); die echte Ungleichung in (23.3) gilt zumindest für $i = 1$ und $i = n - 1$. Also ist L_h nach Satz 84.3 eine M-Matrix. □

Nun zu dem angekündigten Stabilitätsresultat für das Differenzenschema mit zentralen Differenzen.

Satz 84.5. *Sei $b \in C^2[0,1]$ und $c \in C[0,1]$ nichtnegativ. Dann ist bei Verwendung zentraler Differenzen für $u'(x_i)$ das Differenzenschema* (83.9) *mit L_h aus* (83.7)*,* (83.8c) *stabil.*

Beweis. $w \in C^4[0,1]$ bezeichne die Lösung des Randwertproblems

$$-w'' + bw' = 1 \quad \text{in } (0,1)\,, \qquad w(0) = w(1) = 0\,, \tag{84.3}$$

vgl. Aufgabe 8. Dann kann w kein lokales Minimum $x_0 \in (0,1)$ haben, denn sonst wäre $w'(x_0) = 0$ und $w''(x_0) \geq 0$ im Widerspruch zu (84.3). Folglich ist $w(x) \geq 0$ für alle $x \in [0,1]$. Wie zuvor definieren wir $\boldsymbol{w} \in \mathbb{R}^{n-1}$ als den Vektor der Funktionswerte von w über $\triangle_h$. Da c nichtnegativ ist, folgt aus (84.3) punktweise

$$L[w] \;=\; -w'' + bw' + cw \;\geq\; 1 \qquad \text{in } (0,1)$$

und aus der Konsistenz von L_h, vgl. Satz 83.5, ergibt sich daher

$$L_h\boldsymbol{w} \;\geq\; \mathbb{1} - Ch^2\mathbb{1} \;\geq\; \tfrac{1}{2}\,\mathbb{1} \tag{84.4}$$

für h hinreichend klein, etwa für $0 < h < h_0^* < h_0$ mit h_0 aus (84.2). Nun kann $\|L_h^{-1}\|_\infty$ abgeschätzt werden: Da L_h eine M-Matrix ist, folgt nämlich aus (84.1) und (84.4)

$$L_h^{-1}\mathbb{1} \;\leq\; 2L_h^{-1}L_h\boldsymbol{w} \;=\; 2\boldsymbol{w}\,,$$

und da L_h^{-1} nichtnegativ ist, gilt

$$\|L_h^{-1}\|_\infty \;=\; \|L_h^{-1}\mathbb{1}\|_\infty \;\leq\; 2\|\boldsymbol{w}\|_\infty \;\leq\; 2\|w\|_{C[0,1]}$$

für alle $0 < h < h_0^*$. □

Bemerkung 84.6. Dieser Satz bleibt auch noch richtig, wenn b lediglich stetig ist: In diesem Fall gehört w aus (84.3) zu $C^2[0,1]$ und w'' ist gleichmäßig stetig über $[0,1]$; letzteres reicht aus, um nachzuweisen, daß $L_h\boldsymbol{w} \geq \frac{1}{2}\mathbb{1}$ für hinreichend kleine $h > 0$, vgl. Aufgabe 8. ◊

Beispiel. Wir greifen noch einmal die Differentialgleichung $-u'' = f$ aus Beispiel 83.1 auf. Die Methode aus dem Beweis von Satz 84.5 legt nahe, zur Abschätzung von $\|L_h^{-1}\|_\infty$ den Vektor $\boldsymbol{w}$ heranzuziehen, der zur Lösung w des Randwertproblems

$$-w'' = 1 \quad \text{in } (0,1)\,, \qquad w(0) = w(1) = 0\,,$$

gehört, vgl. (84.3). Erwartungsgemäß ist die Lösung $w(x) = (x - x^2)/2$ positiv für $x \in (0,1)$. Aus Lemma 83.3 folgt nun

$$(L_h \boldsymbol{w})_i \;=\; -D_h^2[w](x_i) \;=\; -w''(x_i) \;=\; 1\,, \qquad i = 1, \ldots, n-1\,,$$

da $w'''' = 0$ ist. Somit ist $L_h \boldsymbol{w} = \mathbb{1}$ für jede Gitterweite $h > 0$. Zudem ist L_h nach Korollar 84.4 für jedes $h > 0$ eine M-Matrix, also L_h^{-1} nichtnegativ. Wie im Beweis von Satz 84.5 folgt hieraus

$$\|L_h^{-1}\|_\infty \;=\; \|L_h^{-1}\mathbb{1}\|_\infty \;=\; \|\boldsymbol{w}\|_\infty \;=\; \begin{cases} w(1/2)\,, & n \text{ gerade}\,, \\ w(1/2 \pm h/2)\,, & n \text{ ungerade}\,, \end{cases}$$

also

$$\|L_h^{-1}\|_\infty \;=\; \begin{cases} 1/8\,, & n \text{ gerade}\,, \\ 1/8 - h^2/8\,, & n \text{ ungerade}\,. \end{cases} \tag{84.5}$$

◊

Nun haben wir alle Zutaten für die gesuchte Fehlerabschätzung parat.

Satz 84.7. *Das Randwertproblem* (83.1) *habe eine Lösung* $u \in C^4[0,1]$. *Dann genügt die Lösung* $\boldsymbol{u}_h$ *von* (83.9) *bei Verwendung des zentralen Differenzenquotienten zur Approximation von* $u'(x_i)$ *einer Fehlerabschätzung*

$$\|\boldsymbol{u} - \boldsymbol{u}_h\|_\infty \le Ch^2$$

mit einer Konstanten $C > 0$ *für alle hinreichend kleinen* $h > 0$.

Beweis. Nach Satz 84.5 existieren $C_1 > 0$ und $h_0^* > 0$ mit $\|L_h^{-1}\|_\infty \le C_1$ für $0 < h < h_0^*$. Damit ergibt sich

$$\begin{aligned} \|\boldsymbol{u} - \boldsymbol{u}_h\|_\infty \;&=\; \|L_h^{-1} L_h(\boldsymbol{u} - \boldsymbol{u}_h)\|_\infty \;=\; \|L_h^{-1}(L_h\boldsymbol{u} - \boldsymbol{f})\|_\infty \\ &\le\; C_1 \|L_h\boldsymbol{u} - \boldsymbol{f}\|_\infty \end{aligned}$$

für $0 < h < h_0^*$, und nach Satz 83.5 existiert weiterhin eine Konstante $C_2 > 0$ mit $\|L_h\boldsymbol{u} - \boldsymbol{f}\|_\infty \le C_2 h^2$. Daher folgt die Behauptung mit $C = C_1 C_2$. □

Beispiel. Für Beispiel 83.1 ergibt sich somit aus (84.5) und (83.6)

$$\|\boldsymbol{u} - \boldsymbol{u}_h\|_\infty \;\le\; \|L_h^{-1}\|_\infty \|L_h\boldsymbol{u} - \boldsymbol{f}\|_\infty \;\le\; \frac{1}{8}\,\|L_h\boldsymbol{u} - \boldsymbol{f}\|_\infty \;\le\; \frac{1}{96}\,\|f''\|_{[0,1]} h^2\,,$$

sofern f in $[0,1]$ zweimal stetig differenzierbar ist. ◊

85 Singulär gestörte Probleme

Die Schranke $h_0 = 2/\|b\|_{[0,1]}$ aus (84.2) gibt einen Hinweis, daß das Differenzenverfahren (83.9) für große Gitterweiten instabil ist.

Beispiel 85.1. Zur Illustration untersuchen wir das Randwertproblem

$$-\varepsilon u'' + u' = 1 \quad \text{in } (0,1)\,, \qquad u(0) = u(1) = 0\,, \tag{85.1}$$

mit einem kleinen Parameter $\varepsilon > 0$. Die exakte Lösung dieses Problems lautet

$$u_\varepsilon(x) \;=\; x - v_\varepsilon(x)\,, \qquad v_\varepsilon(x) \;=\; \frac{e^{(x-1)/\varepsilon} - e^{-1/\varepsilon}}{1 - e^{-1/\varepsilon}}\,. \tag{85.2}$$

Für sehr kleine Werte von ε ist der Nenner von v_ε ungefähr 1 und der Zähler ist für $1 - x \gg \varepsilon$ nahe bei Null. Das bedeutet, daß sich der Graph der Lösung außerhalb eines sogenannten *Grenzschichtintervalls* $(1 - \gamma\varepsilon, 1)$ wie der von $u(x) = x$ verhält, bevor er schließlich „abknickt“, um die rechte Randbedingung $u_\varepsilon(1) = 0$ zu erfüllen, vgl. Abbildung 85.1.

Das Differenzenverfahren aus Abschnitt 83 mit zentralen Differenzen für u' ergibt das lineare Gleichungssystem

$$L_h \boldsymbol{u}_h = \boldsymbol{f}\,,$$

wobei $h^2 L_h/\varepsilon$ eine Tridiagonal-Toeplitz-Matrix ist mit den Einträgen

$$d_i = 2\,, \qquad r_i = -1 - h/(2\varepsilon)\,, \qquad s_i = -1 + h/(2\varepsilon)\,,$$

auf der Diagonalen bzw. der unteren und oberen Nebendiagonalen. Nach Definition 84.2 und Korollar 84.4 ist L_h lediglich für $h < 2\varepsilon$ eine M-Matrix. Tatsächlich geht für größere h die Stabilität des Verfahrens verloren: An der Grenzschicht treten starke Oszillationen auf, vgl. Abbildung 85.1; die Parameter in diesem Beispiel sind $h = 1/14$ und $\varepsilon = 1/50$, d. h. $h > 2\varepsilon$. ◊

Falls der Koeffizient b wesentlich größer als der Koeffizient vor u'' ist, muß die Gitterweite somit extrem klein gewählt werden, um Stabilität zu gewährleisten. Üblicherweise formuliert man solche Probleme wie in (85.1) mit einem kleinen Vorfaktor ε vor u'' anstelle eines großen Faktors vor u', also

$$\begin{aligned} L[u] = -\varepsilon u'' + bu' + cu = f \quad &\text{in } (0,1)\,, \\ u(0) = u(1) &= 0\,. \end{aligned} \tag{85.3}$$

Für $\varepsilon = 0$ wird aus diesem Randwertproblem zweiter Ordnung eine Differentialgleichung erster Ordnung mit zwei Randbedingungen; dieses Problem ist

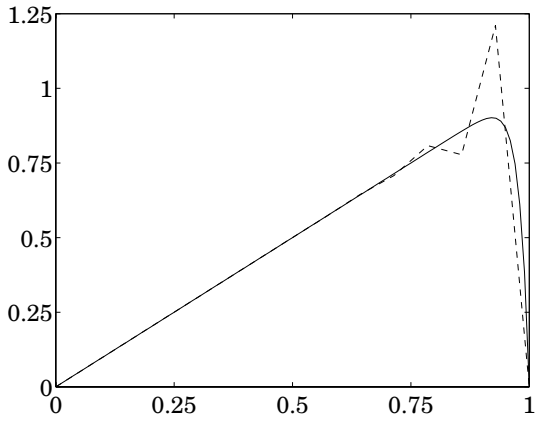


Abb. 85.1: Lösung u_ε und Approximation durch zentrale Differenzen

formal überbestimmt und das erklärt die Gestalt der Lösungskurve. Die Lösung u_ε aus (85.2) stimmt in einem großen Bereich des Intervalls nahezu mit der Lösung $u(x) = x$ des Anfangswertproblems $u' = 1$, $u(0) = 0$, überein. Da u jedoch nicht die rechte Randbedingung erfüllt, entsteht die scharfe Grenzschicht in der Nähe von $x = 1$.

Differentialgleichungen der Form (85.3) mit kleinem $\varepsilon > 0$ heißen *singulär gestört*. Die Lösungen weisen oft solche Grenzschichten auf, und entsprechend schwierig ist die Konstruktion verläßlicher numerischer Algorithmen.

Es liegt natürlich nahe, die Gitterweite h so klein zu wählen, daß die Grenzschicht auf dem jeweiligen Gitter gut aufgelöst werden kann, also $h \ll \varepsilon$. Dann ist die entsprechende Einschränkung $h \lesssim h_0 = 2\varepsilon / \|b\|_{[0,1]}$ für Probleme der Form (85.3) erfüllt und die Voraussetzungen für die Stabilitätsanalyse aus Abschnitt 84 sind gegeben. Die Schranke h_0 aus Korollar 84.4 bezieht sich allerdings nur auf die Verwendung zentraler Differenzenquotienten zur Diskretisierung der ersten Ableitung. Bei einseitigen Differenzenquotienten sieht die Situation anders aus. Betrachtet man etwa die Matrixeinträge (83.8) von L_h, so erkennt man, daß L_h (nach Satz 84.3) automatisch eine M-Matrix ist, wenn die Alternative (83.8a) im Fall $b(x_i) < 0$ und (83.8b) im Fall $b(x_i) > 0$ gewählt wird. Dies ist das sogenannte *Upwind-Schema* mit den Matrixeinträgen

$$\begin{aligned} d_i &= 2 + h|b(x_i)| + h^2 c(x_i)\,, \\ r_i &= -1 - hb^+(x_i)\,, \qquad & b^+(x) &= \max\{b(x), 0\}\,, \\ s_i &= -1 + hb^-(x_i)\,, \qquad & b^-(x) &= \min\{b(x), 0\}\,, \end{aligned} \tag{85.4}$$

in (83.7), bezogen auf den Differentialoperator L_h aus (83.1).

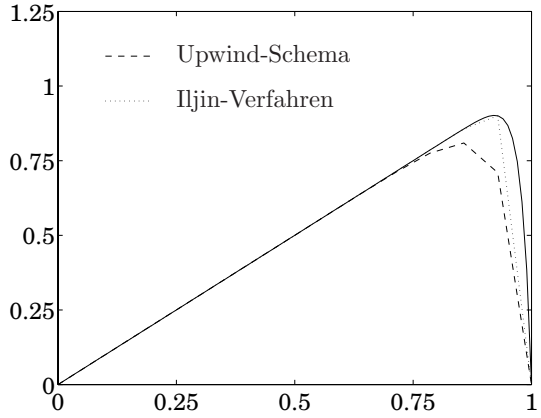


Abb. 85.2: Lösung u_ε und stabile Näherungen

Die Stabilität des Upwind-Schemas steht in Analogie zu der Stabilität des impliziten bzw. expliziten Euler-Verfahrens: Für das Grenzproblem $bu' = -cu$, das für $f = 0$ beim Übergang $\varepsilon \to 0$ aus (85.3) entsteht, würde man aufgrund der Ergebnisse aus Abschnitt 77 für $b > 0$ das implizite Euler-Verfahren und für $b < 0$ das explizite Euler-Verfahren verwenden; ersteres entspricht der Approximation von u' durch den linksseitigen Differenzenquotienten, letzteres einer Approximation durch den rechtsseitigen Differenzenquotienten.

Das Upwind-Verfahren läßt sich schließlich noch physikalisch interpretieren: Der Term $b(x)$ in (83.1) beschreibt für $b > 0$ einen stationären Fluß nach rechts und für $b < 0$ einen Fluß nach links, vgl. Beispiel 69.1. Die obigen Resultate zeigen, daß die Diskretisierung stabil ist, wenn der Differenzenquotient der Flußrichtung entgegengesetzt gewählt wird (gegen den Wind, engl. *upwind*), zur Approximation von u' also nur Information „aus der Vergangenheit“ herangezogen wird.

Beispiel. Abbildung 85.2 zeigt das Ergebnis des Upwind-Schemas für das Modellproblem aus Beispiel 85.1. Wie man sieht, weist die Näherungslösung in diesem Fall keine Oszillationen auf. Andererseits wird mit dem Upwind-Schema auch die Grenzschicht nicht scharf nachgezeichnet; sie wird stark verschmiert. Dies erklärt sich zum Teil durch die schlechtere Konsistenzordnung $q = 1$ des Upwind-Schemas wegen der Verwendung einseitiger Differenzenquotienten. ◊

Das Upwind-Schema angewandt auf (85.3) ergibt die Matrixeinträge

$$\begin{aligned} d_i &= 2\varepsilon + h|b(x_i)| + h^2 c(x_i)\,, \\ r_i &= -\varepsilon - hb^+(x_i)\,, \\ s_i &= -\varepsilon + hb^-(x_i)\,, \end{aligned}$$

für die umskalierte Matrix $h^2 L_h$. Wegen $b^+ = (b + |b|)/2$ und $b^- = (b - |b|)/2$ kann dies wie folgt umgeformt werden,

$$\begin{aligned} d_i &= 2\varepsilon + h\,|b(x_i)| + h^2 c(x_i) \;=\; 2\varepsilon(1+\sigma_i) + h^2 c(x_i)\,, \\ r_i &= -\varepsilon - hb^+(x_i) \;=\; -\varepsilon(1+\sigma_i) - hb(x_i)/2\,, \\ s_i &= -\varepsilon + hb^-(x_i) \;=\; -\varepsilon(1+\sigma_i) + hb(x_i)/2\,, \end{aligned} \tag{85.5}$$

wobei

$$\sigma_i = \sigma(x_i) \qquad \text{mit} \qquad \sigma(x) = \frac{h}{2\varepsilon}\,|b(x)|\,. \tag{85.6}$$

Damit entspricht das Upwind-Schema einer Approximation durch zentrale Differenzen (83.8c) des modifizierten Differentialoperators

$$\widetilde{L}[u] = -\varepsilon a(x) u'' + b(x) u' + c(x) u\,, \qquad a(x) = 1 + \sigma(x)\,, \tag{85.7}$$

anstelle von $L[\,\cdot\,]$ aus (85.3). Der Term $\sigma(x)$ wirkt aufgrund der obigen Resultate offensichtlich stabilisierend: man spricht von *künstlicher Diffusion*.

Neben der geringeren Konsistenzordnung $q = 1$ hat das Upwind-Schema den Nachteil, daß die Konstante C in der Definition 83.4 der Konsistenzordnung von der Größe von b abhängt bzw. im hiesigen Fall von ε. Für $\varepsilon \to 0$ strebt $C \to \infty$, wie man leicht an dem Beweis von Satz 83.5 sieht. Die Suche nach Verfahren erster Ordnung mit einer von ε unabhängigen Konstanten ist Gegenstand intensiver Forschung.

Eine Möglichkeit hierfür bietet das *Iljin-Verfahren*. Auch bei diesem Verfahren ersetzt man den Faktor vor u'' wie in (85.7) durch einen künstlichen Diffusionsterm, diesmal von der Form

$$a_I(x) = \sigma(x) \coth \sigma(x) \tag{85.8}$$

mit σ aus (85.6). Bei der Diskretisierung werden zentrale Differenzen zur Approximation von u' verwendet. Dies führt dann auf die Formeln (85.5) mit $a_I(x_i)$ anstelle von $1 + \sigma_i = a(x_i)$.

Abbildung 85.3 vergleicht die beiden künstlichen Diffusionsterme $a = 1+\sigma$ und $a_I = \sigma \coth \sigma$ über dem Intervall $[0, 20]$. Offensichtlich ist $a \approx a_I \approx 1$ für kleine $\sigma = h(2\varepsilon)^{-1}|b|$, d. h. bei einem kleinen Transportterm approximieren beide Verfahren das Differenzenschema mit zentralen Differenzen. Für dominante Flüsse mit sehr großem σ liegt der Quotient a/a_I nahe bei Eins; das Iljin-Verfahren und das Upwind-Schema sind in diesem Fall also sehr ähnlich.

Man kann das Iljin-Verfahren schließlich noch wie folgt motivieren:

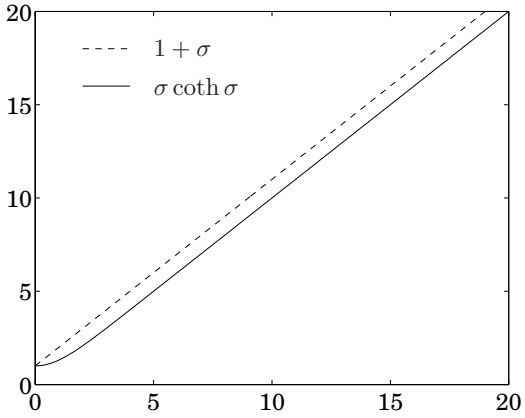


Abb. 85.3: Vergleich der beiden Diffusionsterme

Proposition 85.2. *Für das Modellproblem aus Beispiel* 85.1 *liefert das Iljin-Verfahren die exakte Lösung* $u_i = u_\varepsilon(x_i)$, $i = 1, \dots, n-1$.

Beweis. Sei L_h die Matrix des Iljin-Differenzenschemas und $\boldsymbol{u} \in \mathbb{R}^{n-1}$ der Vektor mit den Werten der exakten Lösung u_ε aus (85.2) über $\triangle_h$. Dann hat die i-te Komponente $(L_h\boldsymbol{u})_i$ von $L_h\boldsymbol{u}$ die Form

$$h^2(L_h\boldsymbol{u})_i \;=\; \varepsilon a_I\big(-u_\varepsilon(x_{i-1}) + 2u_\varepsilon(x_i) - u_\varepsilon(x_{i+1})\big) \\ +\; \varepsilon\sigma\big(-u_\varepsilon(x_{i-1}) + u_\varepsilon(x_{i+1})\big)$$

mit $\sigma = h/(2\varepsilon)$ und $a_I = \sigma\coth\sigma$, vgl. (85.6) und (85.8). Nun verwenden wir die Darstellung (85.2) für die exakte Lösung

$$u_\varepsilon(x) \;=\; x - v_\varepsilon(x) \;=\; x - c_\varepsilon - c_\varepsilon e^{x/\varepsilon}\,,$$

mit der Konstanten $c_\varepsilon = (e^{1/\varepsilon} - 1)^{-1}$ und erhalten

$$h^2(L_h\boldsymbol{u})_i \;=\; 2\varepsilon\sigma h + \varepsilon a_I c_\varepsilon(e^{x_{i-1}/\varepsilon} - 2e^{x_i/\varepsilon} + e^{x_{i+1}/\varepsilon}) + \varepsilon\sigma c_\varepsilon(e^{x_{i-1}/\varepsilon} - e^{x_{i+1}/\varepsilon})\,.$$

Die zentralen Differenzen innerhalb der beiden Klammern vereinfachen sich zu

$$e^{x_{i-1}/\varepsilon} - 2e^{x_i/\varepsilon} + e^{x_{i+1}/\varepsilon} \;=\; e^{x_i/\varepsilon}(e^{-h/\varepsilon} - 2 + e^{h/\varepsilon}) \;=\; e^{x_i/\varepsilon}(e^{\sigma} - e^{-\sigma})^2$$

beziehungsweise

$$e^{x_{i-1}/\varepsilon} - e^{x_{i+1}/\varepsilon} \;=\; e^{x_i/\varepsilon}(e^{-h/\varepsilon} - e^{h/\varepsilon}) \;=\; e^{x_i/\varepsilon}(e^{-2\sigma} - e^{2\sigma})\,,$$

und daher ergibt sich

$$\begin{aligned} h^2(L_h\boldsymbol{u})_i &= 2\varepsilon\sigma h + \varepsilon c_\varepsilon e^{x_i/\varepsilon}\big(a_I(e^\sigma - e^{-\sigma})^2 + \sigma(e^{-2\sigma} - e^{2\sigma})\big) \\ &= h^2 + \sigma\varepsilon c_\varepsilon e^{x_i/\varepsilon}\big(\coth\sigma(e^\sigma - e^{-\sigma})^2 + e^{-2\sigma} - e^{2\sigma}\big) \\ &= h^2 + \frac{hc_\varepsilon e^{x_i/\varepsilon}}{2}(e^{2\sigma} - e^{-2\sigma} + e^{-2\sigma} - e^{2\sigma}) \\ &= h^2 \,. \end{aligned}$$

Folglich ist $L_h\boldsymbol{u} = \mathbb{1}$, was zu beweisen war, vgl. (85.1). □

Tatsächlich ist für das Iljin-Verfahren die erhoffte Fehlerabschätzung gültig:

Satz 85.3. *Gegeben sei das Randwertproblem* (85.3) *mit* $b(x) \geq b_0 > 0$ *in* $[0,1]$. $\boldsymbol{u} \in \mathbb{R}^{n-1}$ *enthalte die Werte der exakten Lösung über dem Gitter* $\triangle_h$ *und* $\boldsymbol{u}_h$ *sei die Näherungslösung des Iljin-Differenzenverfahrens* $L_h\boldsymbol{u}_h = \boldsymbol{f}$. *Dann gilt*

$$\|\boldsymbol{u} - \boldsymbol{u}_h\|_\infty \leq Ch \,,$$

wobei die Konstante C *von* ε *und* h *unabhängig ist.*

Für einen Beweis und weiterführende Resultate sei auf die Monographie von Roos, Stynes und Tobiska [91] verwiesen.

86 Adaptive Gitterverfeinerung

Wie bei Anfangswertaufgaben ist auch bei Randwertaufgaben ein Konzept zur Fehlerkontrolle in einem effizienten Code unerläßlich. Idealerweise schreibt der Anwender eine gewisse Genauigkeit vor, und das Programm wählt automatisch ein geeignetes Gitter, um die gewünschten Anforderungen zu verwirklichen. Dabei sollte das Gitter nach Möglichkeit adaptiv, etwa im Hinblick auf die berechneten Fehlergrößen, verfeinert werden.

Im folgenden wird eine Möglichkeit beschrieben, mit der ein solches Konzept zumindest ansatzweise realisiert werden kann. Dazu nehmen wir an, wir haben bereits eine Näherungslösung $\boldsymbol{u}_h \in \mathbb{R}^{n-1}$ des Randwertproblems

$$\begin{aligned} L[u] &= -u'' + bu' + cu = f \qquad \text{in } (0,1)\,, \\ & u(0) = u(1) = 0\,, \end{aligned} \tag{86.1}$$

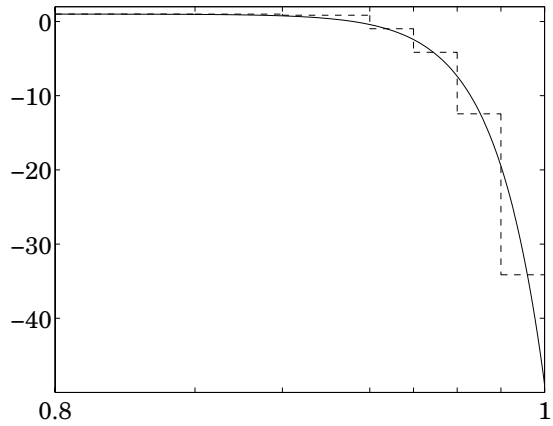


Abb. 86.1: u' und u_h' (gebrochene Linie) über einem nicht äquidistanten Gitter

über einem Gitter $\triangle = \{\, 0 < x_1 < \ldots < x_{n-1} < 1 \,\}$ mit einem Differenzenverfahren berechnet; das Gitter $\triangle$ braucht dazu nicht äquidistant zu sein. Bezeichnen $\Lambda_i(x)$, $i = 0, \ldots, n$, die nodalen Basisfunktionen der linearen Splines über dem erweiterten Gitter $\{x_0 = 0\} \cup \triangle \cup \{x_n = 1\}$, dann korrespondiert $\boldsymbol{u}_h = [u_1, \ldots, u_{n-1}]^T$ mit dem interpolierenden linearen Spline

$$u_h(x) \;=\; \sum_{i=1}^{n-1} u_i \Lambda_i(x)\,, \qquad 0 \le x \le 1\,. \tag{86.2}$$

Allerdings gehört u_h nicht zu $C^2[0,1]$ (sofern u_h nicht die Nullfunktion ist), noch nicht einmal zu $H^2(0,1)$, so daß $L[u_h]$ nicht wohldefiniert ist. Allerdings ist $L[u_h]$ auf jedem Gitterintervall (x_{i-1}, x_i) definiert und für den Fehler $e_h = u - u_h$ gilt

$$\begin{aligned} L[e_h] \;&=\; L[u] - L[u_h] \;=\; f + u_h'' - b u_h' - c u_h \\ &=\; f - b u_h' - c u_h \qquad \text{in } (x_{i-1}, x_i)\,. \end{aligned} \tag{86.3}$$

Auf diese Weise ergibt sich also auf jedem Teilintervall des Gitters eine Differentialgleichung für e_h. Um diese Differentialgleichung lösen zu können, bedarf es allerdings noch zweier Randbedingungen auf jedem Gitterintervall. Da keine exakten Randbedingungen bekannt sind, muß man sich mit heuristischen Überlegungen weiterhelfen.

Die Ableitung u_h' springt beim Übergang von einem Teilintervall ins nächste; wir schreiben

$$\big[\, u_h' \,\big]_{x_i} \;=\; u_h'(x_i+) \;-\; u_h'(x_i-)\,, \qquad i = 1, \ldots, n-1\,, \tag{86.4}$$

für die Sprunghöhe beim Übergang an der Stelle $x = x_i$. Da die exakte Lösung von (86.1) zu $C^2[0,1]$ gehören soll, ist $[\,u'\,]_{x_i} = 0$, $i = 1, \ldots, n-1$. Um eine Randbedingung für e_h' an den Punkten x_{i-1} und x_i zu konstruieren, nehmen wir an, daß

$$u'(x_i) \;\approx\; \frac{1}{2}\Big(u_h'(x_i-) \,+\, u_h'(x_i+)\Big)\,, \qquad i = 1, \ldots, n-1\,, \tag{86.5}$$

vgl. Abbildung 86.1 zur Motivation. Mit dieser Annahme ergibt sich dann

$$\lim_{x \to x_i-} e_h'(x) \;=\; u'(x_i) - u_h'(x_i-) \;\approx\; \frac{1}{2}\,\big[\,u_h'\,\big]_{x_i} \tag{86.6a}$$

und

$$\lim_{x \to x_i+} e_h'(x) \;=\; u'(x_i) - u_h'(x_i+) \;\approx\; -\,\frac{1}{2}\,\big[\,u_h'\,\big]_{x_i} \tag{86.6b}$$

für $i = 1, \ldots, n-1$. An den beiden äußeren Randpunkten $x = 0$ und $x = 1$ liegen hingegen exakte Randbedingungen für e_h vor, nämlich

$$e_h(0) = 0\,, \qquad e_h(1) = 0\,. \tag{86.7}$$

Mit den Randbedingungen (86.6) und (86.7) können nun Näherungen $\tilde{e}_h$ für e_h in allen Teilintervallen (x_{i-1}, x_i), $i = 1, \ldots, n$, des Gitters berechnet werden. Dazu fügt man in dem jeweiligen Gitterintervall weitere Gitterpunkte ein und verwendet wiederum ein Differenzenverfahren zur Lösung der Randwertaufgabe für $\tilde{e}_h$. Falls $c = 0$ ist, können dabei Lösbarkeits- und Eindeutigkeitsprobleme auftreten, vgl. Bemerkung 83.6; gegebenenfalls ist dann die Annahme (86.5) geeignet zu modifizieren, etwa wie in dem nachfolgenden Beispiel.

Als Maß für die Genauigkeit der Näherungslösung $\boldsymbol{u}_h$ bietet es sich nun etwa an, die Größe von $\tilde{e}_h$ in $H^1(x_{i-1}, x_i)$ oder $C^1[x_{i-1}, x_i]$ heranzuziehen. Diagnostiziert man auf diese Weise in einzelnen Teilintervallen unverhältnismäßig große Fehler, so wird das Gitter $\triangle$ im entsprechenden Bereich mit zusätzlichen Punkten verfeinert und anschließend eine neue Näherung über dem verfeinerten Gitter berechnet.

Aufwand. Der Aufwand dieses Fehlerschätzers hängt von der Anzahl Gitterpunkte ab, die bei der Lösung von (86.3) in jedem Teilintervall verwendet werden. Bei q zusätzlichen Gitterpunkten je Teilintervall kostet die Berechnung des Fehlerschätzers etwa $5q$ Multiplikationen *pro Teilintervall*, also etwa $5qn$ Multiplikationen insgesamt. Damit übersteigt die Berechnung des Fehlerschätzers den Aufwand des gesamten Verfahrens um den Faktor q. Typischerweise wird man daher nur an kritischen Stellen den Fehlerschätzer einsetzen, etwa dort, wo zuvor bereits einmal verfeinert werden mußte. $\Diamond$

Bemerkung. Die relativ hohen Kosten des Fehlerschätzers (verglichen mit den Kosten für die Lösung des diskreten Problems an und für sich) beruhen darauf, daß die Berechnung von u_h mit nur $O(n)$ Operationen *billig* ist. Ist die Lösung des diskreten Problems teurer, etwa proportional zu n^ν mit einem $\nu > 1$, dann ergibt sich entsprechend der Kostenaufwand $O(q^\nu)$ für den Fehlerschätzer je Teilintervall, also $O(q^\nu n)$ für den Gesamtaufwand des Fehlerschätzers. Diese Kosten wären dann gegenüber dem Aufwand $O(n^\nu)$ zur Berechnung der Näherungslösung vernachlässigbar. Ein entsprechendes Szenarium ergibt sich bei höherdimensionalen Problemen, also bei Randwertaufgaben für elliptische Differentialgleichungen in $\mathbb{R}^d$ mit $d \geq 2$, auf die der besprochene Fehlerschätzer verallgemeinert werden kann. Für solche Randwertaufgaben spielen die Kosten des Fehlerschätzers nur eine sehr untergeordnete Rolle. $\Diamond$

Beispiel 86.1. Gegeben sei die Differentialgleichung $-\varepsilon u'' + u' = 1$ mit $\varepsilon = 1/50$ aus Beispiel 85.1. Dort haben wir gesehen, daß die Näherungslösung bei Verwendung zentraler Differenzen stark oszilliert, sofern die Gitterweite h nicht hinreichend klein ist. Die resultierende Einschränkung an h führt jedoch zu einem feinen Gitter (und damit zu einem hohen Aufwand), obwohl im Bereich $x \in [0, 0.75]$ die Lösung bereits mit einem groben Gitter gut rekonstruiert werden kann. Hier bietet sich also eine adaptive Gitterverfeinerung an, die über einen kontrollierenden Fehlerschätzer gesteuert wird.

Wir beginnen mit derselben Diskretisierung wie in Beispiel 85.1, also einem relativ groben äquidistanten Gitter $\triangle_h$ mit $h = 1/14$ und zentralen Differenzen zur Approximation von u'. Für den oben vorgestellten Fehlerschätzer unterteilen wir jedes Gitterintervall in sechs äquidistante Teilintervalle und lösen dort die jeweiligen Differentialgleichungen (86.3), die sich in unserem Fall zu

$$-\varepsilon e_h'' + e_h' = 1 - u_h' \qquad \text{in } (x_{i-1}, x_i) \tag{86.8}$$

vereinfachen. Man beachte, daß $u_h' = (u_i - u_{i-1})/(x_i - x_{i-1})$ im gesamten Gitterintervall konstant ist (wobei wir $u_0 = u_n = 0$ setzen); für jedes $i = 1, \ldots, n$ ist also die rechte Seite von (86.8) eine Konstante.

In diesem Beispiel ist die Randvorgabe (86.6) problematisch, da die Differentialgleichung (86.8) nicht für jede Neumann-Randvorgabe lösbar ist. Statt dessen verwenden wir hier für die Berechnung von $\tilde{e}_h$ im Intervall (x_{i-1}, x_i), $i = 1, \ldots, n-1$, die Randvorgaben

$$\tilde{e}_h(x_{i-1}) = 0\,, \qquad \tilde{e}_h'(x_i) = \frac{1}{2}\left[\, u_h' \,\right]_{x_i}.$$

Als Verfeinerungsindikatoren ϵ_i betrachten wir die H^1-Halbnormen

$$\epsilon_i = \Big(\int_{x_{i-1}}^{x_i} \big|\, \tilde{e}_h'(x) \,\big|^2 \, dx \Big)^{1/2} \tag{86.9}$$

und verfeinern das oder die Teilintervalle, deren Fehlerindikatoren ϵ_i am größten sind. Dabei kann ein großer Fehlerindikator ϵ_i einerseits auf eine Verletzung der Randbedingung $\tilde{e}_h(x_{i-1}) = 0$ oder auf einen großen Fehler e_h im Intervallinneren hindeuten. Bei einer Verfeinerung aufgrund eines großen Fehlerindikators ϵ_i werden daher *beide* an x_{i-1} angrenzenden Gitterintervalle durch je einen zusätzlichen Gitterpunkt in der Intervallmitte halbiert.

Bei der Diskretisierung der Differentialgleichung $-\varepsilon u'' + u' = 1$ muß man beachten, daß die Gitter nach dem ersten Schritt nicht mehr äquidistant sind. Die zentralen Differenzenquotienten für u' und u'' werden dann komplizierter; sie lauten

$$D_h[u](x_i) = \frac{1}{h_i + h_{i+1}} \Big(\frac{h_i}{h_{i+1}} u_{i+1} + \big(\frac{h_{i+1}}{h_i} - \frac{h_i}{h_{i+1}} \big) u_i - \frac{h_{i+1}}{h_i} u_{i-1} \Big),$$

$$D_h^2[u](x_i) = \frac{2}{h_i + h_{i+1}} \Big(\frac{1}{h_{i+1}} u_{i+1} - \big(\frac{1}{h_{i+1}} + \frac{1}{h_i} \big) u_i + \frac{1}{h_i} u_{i-1} \Big);$$

hierbei ist $h_i = x_i - x_{i-1}$, $i = 1, \ldots, n$.

Die Abbildung 86.2 zeigt die ersten sechs Schritte einer solchen Verfeinerungssequenz. Für jeden Schritt enthält das obere Bild die exakte Lösung u (durchgezogene Kurve) und die Rekonstruktion u_h aus (86.2) als linearen Spline (gebrochene Kurve); das jeweils untere Bild stellt mit gebrochener Linie die Fehlerindikatoren dar (als Treppenfunktion mit Wert ϵ_i in (x_{i-1}, x_i)) und entsprechend den exakten H^1-Fehler (durchgezogene Kurve), der sich ergibt, wenn in (86.9) $\tilde{e}_h$ durch $e_h = u - u_h$ ersetzt wird. Anhand der Treppenstufen bei den Fehlerindikatoren kann die aktuelle Gitterunterteilung erkannt werden; die Sterne auf der x-Achse zeigen diejenigen Gitterpunkte an, an denen verfeinert wird.

Wie man erkennen kann, folgen die Fehlerindikatoren in sehr überzeugender Weise dem Fehlerverlauf, und man kann anhand der Fehlerindikatoren verläßlich entscheiden, wo der Fehler am größten ist. Nach den sechs Verfeinerungsstufen besteht das Gitter aus 36 Teilintervallen und die Näherungslösung u_h liegt so nah an der exakten Lösung, daß mit dem Auge fast kein Unterschied mehr zu erkennen ist, vgl. Abbildung 86.3. Bis zuletzt wird keines der Intervalle im Bereich $[0, 0.75]$ verfeinert. Abbildung 86.3 enthält zum Vergleich auch die Näherung $u_{1/36}$ (gepunktet), die sich bei einem äquidistanten Gitter mit derselben Anzahl Gitterintervalle ergeben würde. Das rechte Teilbild zeigt die absoluten Fehler der beiden Näherungslösungen u_h und $u_{1/36}$. Wie man sieht, ist die Approximation u_h im Bereich der Grenzschicht um fast eine Dezimalstelle genauer. ◊

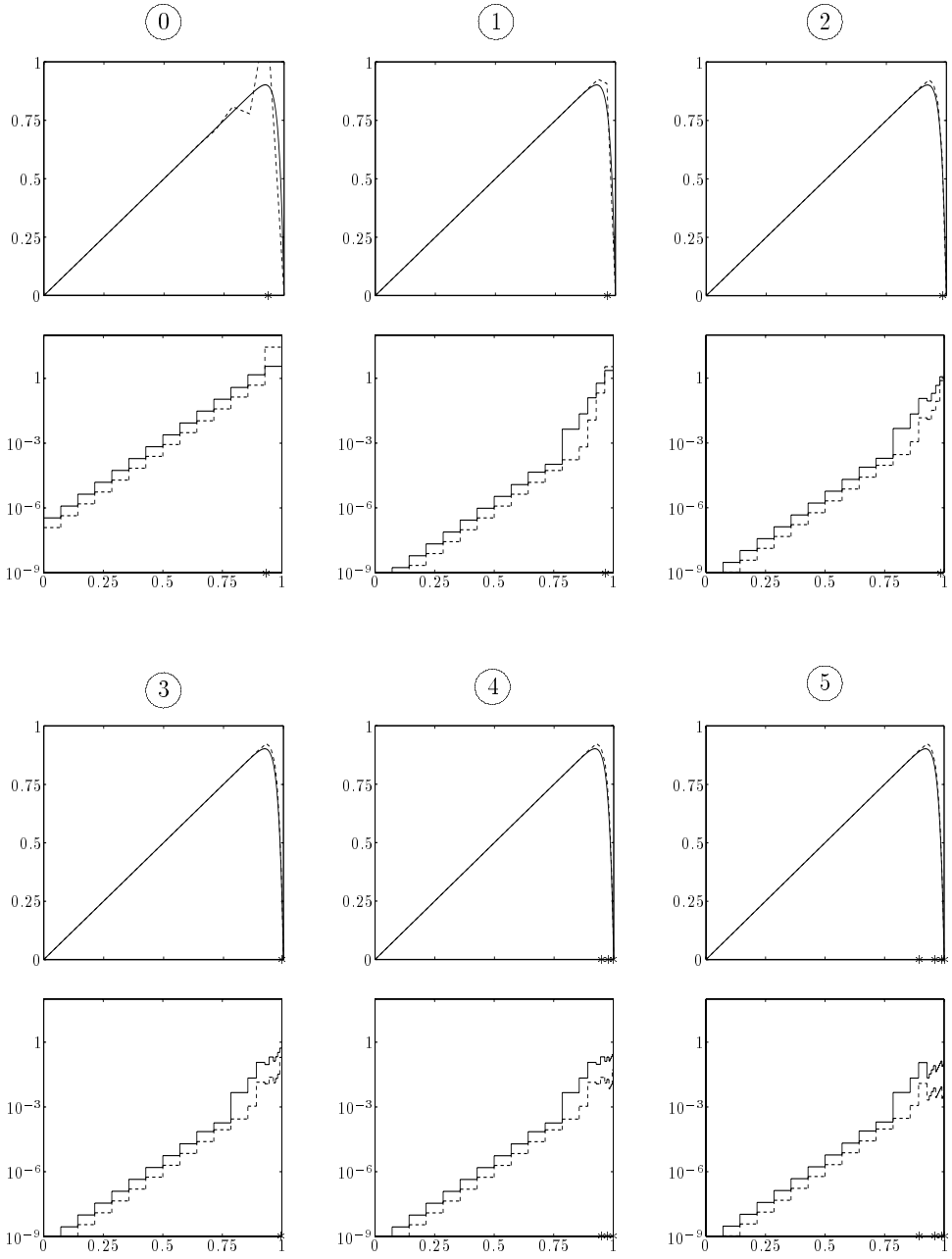


Abb. 86.2: Näherungen und Fehlerschätzer

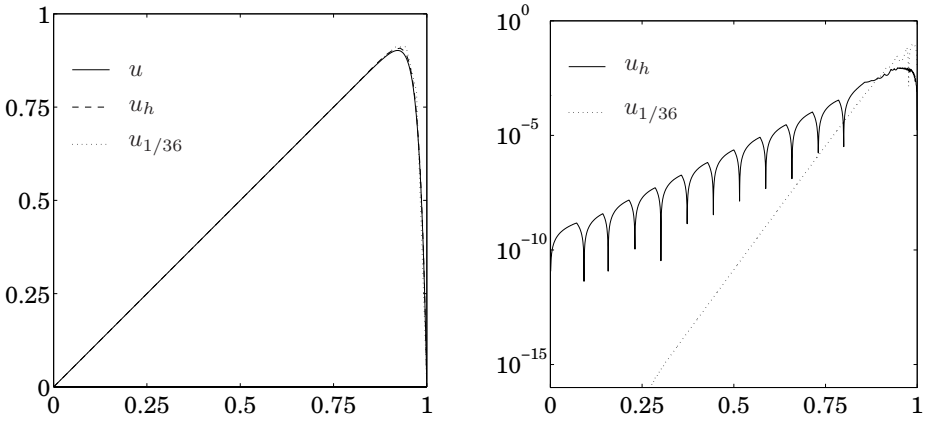


Abb. 86.3: Rekonstruktionen u_h und $u_{1/36}$ (links) sowie der absolute Fehler (rechts)

87 Das Schießverfahren

Für den Rest dieses Kapitels wenden wir uns (skalaren) nichtlinearen Randwertproblemen der Form

$$u'' = f(x, u, u') \quad \text{in } (0,1)\,, \qquad u(0) = u(1) = 0\,, \tag{87.1}$$

zu. Wir setzen dabei voraus, daß eine Lösung $u \in C^2[0,1]$ von (87.1) existiert. Die Verfahren, die im weiteren vorgestellt werden, lassen sich ohne große Modifikationen auf Systeme von Randwertaufgaben mit komplizierteren (auch nichtlinearen) Randbedingungen übertragen.

Nehmen wir an, die Ableitung $\alpha = u'(0)$ am linken Rand wäre bekannt. Dann löst u neben (87.1) auch das *Anfangswertproblem*

$$v'' = f(x, v, v')\,, \qquad v(0) = 0\,, \quad v'(0) = \alpha\,. \tag{87.2}$$

Dieses Anfangswertproblems ist unter milden Glattheitsbedingungen an die Funktion f eindeutig lösbar (vgl. etwa Satz 73.1) und in Kapitel XIV haben wir effiziente Algorithmen kennengelernt, die zur Lösung von (87.2) eingesetzt werden können. Das Problem ist nur: Der genaue Wert von $\alpha = u'(0)$ ist in der Regel *nicht* bekannt.

Das Schießverfahren liefert eine sehr intuitive algorithmische Lösung dieses Problems. Man wählt einen Schätzwert für α, berechnet die zugehörige Lösung v_α des Anfangswertproblems (87.2) mit einem der Verfahren aus Kapitel XIV und optimiert α solange, bis $v_\alpha(1) \approx u(1) = 0$. Diese Vorgehensweise ist in Abbildung 87.1 illustriert.

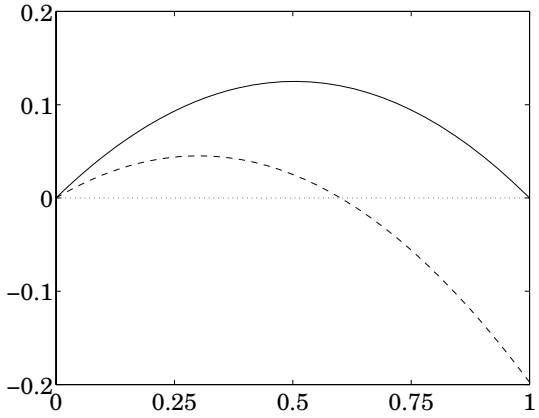


Abb. 87.1: Lösung u und Näherung v_α (gebrochene Linie)

Für diejenigen Parameter α, für die die Lösung v_α von (87.2) im gesamten Intervall $[0,1]$ existiert, kann die Abbildung

$$F : \mathcal{D}(F) \to \mathbb{R}\,, \qquad F : \alpha \mapsto v_\alpha(1)\,, \tag{87.3}$$

definiert werden, und zur Lösung von (87.1) wird eine Nullstelle $\widehat{\alpha}$ dieser Abbildung gesucht: Ist $F(\widehat{\alpha}) = 0$, dann löst $v_{\widehat{\alpha}}$ das Randwertproblem (87.1); ist umgekehrt u eine Lösung von (87.1), dann hat F eine Nullstelle für $\widehat{\alpha} = u'(0)$.

Die Funktion F ist im allgemeinen nichtlinear und die Nullstelle $\widehat{\alpha}$ von F kann prinzipiell mit jedem Verfahren zur Lösung nichtlinearer Gleichungen gelöst werden, etwa mit dem Sekanten- oder dem Newton-Verfahren. Für letzteres muß die Ableitung $F'(\alpha)$ von F ausgerechnet oder zumindest approximiert werden.

Proposition 87.1. *Die Funktion f sei stetig differenzierbar bezüglich u und u', und v_α und F seien wie in* (87.2) *bzw.* (87.3) *definiert. Dann ist F differenzierbar mit $F'(\alpha) = w_\alpha(1)$, wobei w_α die Lösung des Anfangswertproblems*

$$\begin{aligned} &w'' = f_u(x, v_\alpha, v'_\alpha)w + f_{u'}(x, v_\alpha, v'_\alpha)w'\,, \\ &w(0) = 0\,, \quad w'(0) = 1\,, \end{aligned} \tag{87.4}$$

ist.

Beweis. Sei $\alpha \in \mathbb{R}$ fest gewählt und $\beta \neq \alpha$. Dann bezeichnen v_α und v_β die beiden Lösungen des Anfangswertproblems (87.2) mit $v'(0) = \alpha$ bzw. β. Ferner definieren wir

$$w_\beta(x) \;=\; \frac{v_\beta(x) - v_\alpha(x)}{\beta - \alpha}\,.$$

Dann gilt offensichtlich

$$\frac{F(\beta)-F(\alpha)}{\beta-\alpha} \;=\; \frac{v_\beta(1)-v_\alpha(1)}{\beta-\alpha} \;=\; w_\beta(1)\,,$$

und durch Grenzübergang $\beta \to \alpha$ erhalten wir die Darstellung

$$F'(\alpha) = \lim_{\beta\to\alpha} w_\beta(1)\,, \tag{87.5}$$

sofern der Grenzwert auf der rechten Seite existiert. Wegen (87.2) erfüllt w_β die beiden Anfangsbedingungen

$$w_\beta(0) = 0\,, \qquad w'_\beta(0) = 1\,,$$

und darüberhinaus gilt für festes $x \in (0,1)$

$$(\beta-\alpha)w''_\beta \;=\; v''_\beta - v''_\alpha \;=\; f(x,v_\beta,v'_\beta) - f(x,v_\alpha,v'_\alpha) \;=\; \phi(1)-\phi(0)\,,$$

wobei

$$\phi(\xi) \;=\; f\Big(x, v_\alpha+\xi(v_\beta-v_\alpha), v'_\alpha+\xi(v'_\beta-v'_\alpha)\Big)\,.$$

Mit der Abkürzung $v = v_\alpha + \xi(v_\beta - v_\alpha)$ ergibt sich

$$\begin{aligned}\phi'(\xi) &= f_u(x,v,v')(v_\beta-v_\alpha) + f_{u'}(x,v,v')(v'_\beta-v'_\alpha)\\ &= (\beta-\alpha)\big(f_u(x,v,v')w_\beta + f_{u'}(x,v,v')w'_\beta\big)\,,\end{aligned}$$

und oben eingesetzt folgt

$$\begin{aligned}w''_\beta &= \frac{\phi(1)-\phi(0)}{\beta-\alpha} \;=\; \frac{1}{\beta-\alpha}\int_0^1 \phi'(\xi)\,d\xi\\ &= \int_0^1 \big(f_u(x,v,v')w_\beta + f_{u'}(x,v,v')w'_\beta\big)\,d\xi\\ &= \Big(\int_0^1 f_u(x,v,v')\,d\xi\Big)\,w_\beta \;+\; \Big(\int_0^1 f_{u'}(x,v,v')\,d\xi\Big)\,w'_\beta\,.\end{aligned}$$

Nach Satz 73.2 streben für $\beta \to \alpha$ die Argumente v und v' in den beiden Integralen jeweils gleichmäßig gegen v_α und v'_α. Aus dem gleichen Satz folgt somit, daß die Lösung w_β der linearen Differentialgleichung gleichmäßig gegen die eindeutig bestimmte Lösung w der Grenz-Differentialgleichung (87.4) konvergiert. Daher ergibt sich die Behauptung aus (87.5) und der Beweis ist fertig.

□

Beispiel 87.2. Wir betrachten das Beispiel

$$u'' + uu' = -1\,, \qquad u(0) = u(1) = 0\,.$$

Abbildung 87.1 zeigt die exakte Lösung sowie die Näherung für $\alpha_0 = 0.3$. Zur Berechnung einer Nullstelle $\widehat{\alpha}$ der zugehörigen Abbildung F aus (87.3) soll das Newton-Verfahren

$$\alpha_{k+1} = \alpha_k - F(\alpha_k)/F'(\alpha_k)\,, \qquad k = 0, 1, \dots\,,$$

verwendet werden. Dazu ist in jedem Iterationsschritt $F(\alpha_k)$ und $F'(\alpha_k)$ auszuwerten. Dies erfordert die Lösung der beiden Anfangswertprobleme (87.2) und (87.4):

$$\begin{aligned} v'' + vv' &= -1\,, \qquad v(0) = 0\,, \quad v'(0) = \alpha_k\,, \\ w'' + vw' + v'w &= 0\,, \qquad w(0) = 0\,, \quad w'(0) = 1\,; \end{aligned}$$

dann folgt

$$F(\alpha_k) = v(1) \qquad \text{und} \qquad F'(\alpha_k) = w(1)\,.$$

Die beiden Differentialgleichungen sind gekoppelt (w hängt von v ab) und sollten daher als ein gemeinsames Anfangswertproblem für die vier Funktionen $y_1 = v$, $y_2 = v'$, $y_3 = w$ und $y_4 = w'$ gelöst werden. Dieses System lautet

$$\begin{aligned} y_1' &= y_2\,, & y_1(0) &= 0\,, \\ y_2' &= -1 - y_1 y_2\,, & y_2(0) &= \alpha_k\,, \\ y_3' &= y_4\,, & y_3(0) &= 0\,, \\ y_4' &= -y_1 y_4 - y_2 y_3\,, & y_4(0) &= 1\,. \end{aligned}$$

In dieser Weise findet das Newton-Verfahren nach drei Iterationsschritten den Näherungswert $\alpha = 0.4958$ für $u'(0)$. ◊

In der bislang beschriebenen Form ist das Schießverfahren jedoch nicht sehr stabil. Beispielsweise wird das maximale Existenzintervall einer Lösung v_α von (87.2) in der Regel von α abhängen; unter Umständen gehört bereits der „Zielpunkt“ $x = 1$ nicht mehr zum Existenzintervall, d. h. $\alpha \notin \mathcal{D}(F)$. Selbst wenn $\mathcal{D}(F) = \mathbb{R}$, also wenn für jedes α die Lösung v_α im gesamten Intervall $[0, 1]$ existiert, kann es aus anderen Gründen zu Stabilitätsproblemen kommen:

Beispiel. Der Einfachheit halber betrachten wir das *lineare* Randwertproblem

$$-u'' + c^2 u = c^2 x\,, \qquad u(0) = 0\,, \quad u(1) = 0\,, \tag{87.6}$$

über $(0,1)$ mit konstantem Parameter $c > 0$. Durch $u = w + x$ wird dieses Randwertproblem in das homogene Problem

$$-w'' + c^2 w = 0\,, \qquad w(0) = 0\,, \quad w(1) = -1\,,$$

überführt. Die Lösung von (87.6) hat daher die Form

$$u(x) = x + \xi_1 e^{cx} + \xi_2 e^{-cx}\,,$$

wobei die Koeffizienten ξ_1 und ξ_2 so einzustellen sind, daß die beiden Randwerte $u(0) = u(1) = 0$ angenommen werden. Die zugehörige Differentialgleichung (87.2) für v_α hat die entsprechende Lösung

$$v_\alpha(x) = x + \eta_1 e^{cx} + \eta_2 e^{-cx}\,,$$

wobei nun η_1 und η_2 aus dem linearen Gleichungssystem

$$\begin{bmatrix} 1 & 1 \\ c & -c \end{bmatrix} \begin{bmatrix} \eta_1 \\ \eta_2 \end{bmatrix} = \begin{bmatrix} 0 \\ \alpha - 1 \end{bmatrix} \tag{87.7}$$

bestimmt werden müssen. Hieraus folgt $\eta_2 = -\eta_1$ und schließlich die Lösung

$$v_\alpha(x) = x + \frac{\alpha - 1}{2c} \left(e^{cx} - e^{-cx} \right).$$

Für $c \gg 0$ gilt daher näherungsweise

$$F(\alpha) = v_\alpha(1) \approx \frac{\alpha - 1}{2c}\, e^c\,,$$

und bei einer Implementierung des Schließverfahrens zur numerischen Lösung von (87.6) droht die Gefahr eines Overflows oder zumindest einer starken Beeinträchtigung des Ergebnisses durch große absolute Rundungsfehler. Auf der anderen Seite kann (87.6) mit einem der Differenzenverfahren aus Abschnitt 83 stabil gelöst werden. Mit anderen Worten: Die Randwertaufgabe ist gut konditioniert, lediglich das Schießverfahren als Lösungsalgorithmus ist numerisch instabil, vgl. Abschnitt 1. ◊

Der Grund für dieses instabile Verhalten ist das Auftreten stark anwachsender Lösungskurven bei allgemeiner Wahl der Anfangswerte. Um dies zu verhindern und darüberhinaus die Existenz der Näherungslösung v_α zu garantieren, verwendet man das Schießverfahren meist in der folgenden Variante, der *Mehrzielmethode*.

Dazu gibt man sich ein Gitter $\triangle = \{0 = x_0 < x_1 < \ldots < x_n = 1\} \subset [0,1]$ vor und parametrisiert die Lösung nicht nur durch $\alpha_0 = u'(0)$ sondern zusätzlich noch durch

$$\eta_i = u(x_i)\,, \quad \alpha_i = u'(x_i)\,, \qquad i = 1, \ldots, n-1\,.$$

Der Vollständigkeit halber setzen wir noch $\eta_0 = u(0) = 0$. Sofern die Parameter η_i und α_i, $i = 1, \dots, n-1$, sowie α_0 allesamt bekannt sind, kann die Lösung u von (87.1) aus den Lösungen $v_i : [x_{i-1}, x_i] \to \mathbb{R}$ der n Anfangswertprobleme

$$v_i'' = f(x, v_i, v_i')\,, \qquad v_i(x_{i-1}) = \eta_{i-1}\,, \quad v_i'(x_{i-1}) = \alpha_{i-1}\,, \tag{87.8}$$

für $i = 1, \dots, n$, zusammengestückelt werden. Damit gilt es bei der Mehrzielmethode $2n-1$ Unbekannte (anstelle der einen Unbekannten α beim bisherigen Schießverfahren) zu bestimmen; dafür stehen allerdings auch $2n-2$ Bedingungen mehr als zuvor zur Verfügung, nämlich

$$\eta_i = v_i(x_i)\,, \quad \alpha_i = v_i'(x_i)\,, \qquad i = 1, \dots, n-1\,.$$

Gesucht ist daher eine Nullstelle der (nichtlinearen) Abbildung

$$F(z) \;=\; \begin{bmatrix} v_1(x_1) - \eta_1 \\ v_1'(x_1) - \alpha_1 \\ \vdots \\ v_{n-1}(x_{n-1}) - \eta_{n-1} \\ v_{n-1}'(x_{n-1}) - \alpha_{n-1} \\ v_n(1) \end{bmatrix} \qquad \text{mit} \qquad z \;=\; \begin{bmatrix} \alpha_0 \\ \eta_1 \\ \alpha_1 \\ \vdots \\ \eta_{n-1} \\ \alpha_{n-1} \end{bmatrix}.$$

Bei der Anwendung des Newton-Verfahrens zur Bestimmung einer solchen Nullstelle kann man aus der Struktur der Jacobi-Matrix F' Nutzen ziehen. Es ist nämlich

$$F'(z) \;=\; \begin{bmatrix} \omega_{1,\alpha} & -1 & & & & & & & \\ \mu_{1,\alpha} & 0 & -1 & & & & & & \\ & \omega_{2,\eta} & \omega_{2,\alpha} & -1 & & & & & \\ & \mu_{2,\eta} & \mu_{2,\alpha} & 0 & -1 & & & & \\ & & & \omega_{3,\eta} & \omega_{3,\alpha} & -1 & & & \\ & & & \mu_{3,\eta} & \mu_{3,\alpha} & 0 & -1 & & \\ & & & & & \ddots & & \ddots & \\ & & & & & & \omega_{n-1,\eta} & \omega_{n-1,\alpha} & -1 \\ & & & & & & \mu_{n-1,\eta} & \mu_{n-1,\alpha} & 0 & -1 \\ & & & & & & & & \omega_{n,\eta} & \omega_{n,\alpha} \end{bmatrix},$$

wobei jeweils

$$\begin{aligned} \omega_{i,\eta} \;&=\; \frac{\partial}{\partial \eta_{i-1}}\, v_i(x_i)\,, \qquad \omega_{i,\alpha} \;=\; \frac{\partial}{\partial \alpha_{i-1}}\, v_i(x_i)\,, \\ \mu_{i,\eta} \;&=\; \frac{\partial}{\partial \eta_{i-1}}\, v_i'(x_i)\,, \qquad \mu_{i,\alpha} \;=\; \frac{\partial}{\partial \alpha_{i-1}}\, v_i'(x_i)\,. \end{aligned} \tag{87.9}$$

Mit einem entsprechenden Beweis wie in Proposition 87.1 können diese Ableitungen wieder über die Lösung gewisser Anfangswertaufgaben bestimmt werden: Es gilt

$$\begin{aligned} \omega_{i,\eta} &= W_i(x_i)\,, & \omega_{i,\alpha} &= w_i(x_i)\,, \\ \mu_{i,\eta} &= W_i'(x_i)\,, & \mu_{i,\alpha} &= w_i'(x_i)\,, \end{aligned} \tag{87.10}$$

wobei w_i und W_i die Lösungen der Differentialgleichungen

$$\begin{aligned} &w_i'' = f_u(x, v_i, v_i')w_i + f_{u'}(x, v_i, v_i')w_i'\,, \\ &w(x_{i-1}) = 0\,, \quad w'(x_{i-1}) = 1\,, \\ &W_i'' = f_u(x, v_i, v_i')W_i + f_{u'}(x, v_i, v_i')W_i'\,, \\ &W(x_{i-1}) = 1\,, \quad W'(x_{i-1}) = 0\,, \end{aligned} \tag{87.11}$$

bezeichnen.

Der Hauptaufwand eines einzelnen Newton-Schritts besteht also in der Lösung der gekoppelten Anfangswertaufgaben (87.8) und (87.11) für (v_i, w_i, W_i), $i = 1, \dots, n$. Diese n Anfangswertaufgaben in den einzelnen Teilintervallen können dabei prinzipiell unabhängig voneinander (auch parallel) gelöst werden. Die Lösung des abschließenden linearen Gleichungssystems mit der Koeffizientenmatrix $F'(z)$ kann wieder sehr effizient durch Gauß-Elimination bestimmt werden, da $F'(z)$ eine Bandmatrix ist.

88 Optimierungsrandwertaufgaben

Zur Abrundung dieses Kapitels betrachten wir Anfangswertprobleme für Differentialgleichungen, die von gewissen Parametern abhängen:

$$u' = f(t, u; a)\,, \quad t_0 < t < T\,, \qquad u(t_0) = \eta_0\,. \tag{88.1}$$

Dabei ist $a = [a_1, \dots, a_p]^T \in \mathbb{R}^p$ ein Parametervektor und der Einfachheit halber beschränken wir uns wieder auf skalare Funktionen $u = u(t; a, \eta_0)$.

In Kapitel XIV haben wir Algorithmen behandelt, mit denen die Lösung eines solchen Anfangswertproblems für gegebene Parameterwerte numerisch berechnet werden kann. Im Gegensatz dazu sei nun angenommen, daß Meßwerte $y_j \approx u(t_j; \widehat{a}, \widehat{\eta}_0)$ der Lösung von (88.1) für bestimmte Zeitpunkte t_j, $j = 0, \dots, m$, gegeben sind und der zugehörige Parametervektor $\widehat{a}$ und gegebenenfalls der Anfangswert $\widehat{\eta}_0$ gesucht sind. Dabei sei m größer als p und

die Parameter a und η_0 sind im Sinne eines nichtlinearen $(m+1)\times(p+1)$-dimensionalen Ausgleichsproblems

$$\text{minimiere} \quad \frac{1}{2}\sum_{j=0}^{m} |u(t_j;a,\eta_0)-y_j|^2 \tag{88.2}$$

zu optimieren.

Derartige *Parameteridentifikationsprobleme* treten in vielen Anwendungen auf und können mit ähnlichen Methoden wie im vorangegangenen Abschnitt angegangen werden. Dazu wird das Zeitintervall $[t_0,t_m]$ in ein Gitter

$$\triangle \;=\; \{\, t_0 = x_0 < x_1 < \ldots < x_n = t_m \,\}\,, \qquad n<m\,,$$

unterteilt, bei dem nicht nur Meßpunkte als Gitterknoten in Frage kommen. Zu jedem Parametervektor a und jedem Satz von Anfangswerten $\eta=[\eta_i]_{i=0}^{n-1}$ lassen sich auf den einzelnen Teilintervallen $[x_{i-1},x_i]$, $i=1,\ldots,n$, die n Anfangswertprobleme

$$u_i' = f(t,u_i;a)\,, \quad x_{i-1}\le t\le x_i\,, \qquad u_i(x_{i-1})=\eta_{i-1}\,, \tag{88.3}$$

lösen, und die zugehörigen Lösungen $u_i = u_i(t;a,\eta_{i-1})$ können anschließend wie in Abschnitt 87 zu einer Funktion

$$u(t;a,\eta) \;=\; u_i(t;a,\eta_{i-1}) \qquad \text{für} \quad t\in[x_{i-1},x_i) \tag{88.4}$$

über $[t_0,T)$ zusammengesetzt werden.

Der Parametervektor a und die Anfangswerte η_i sind einerseits so einzustellen, daß die zusammengesetzte Funktion in jedem Gitterpunkt stetig ist, also daß die Nebenbedingungen

$$u_i(x_i;a,\eta_{i-1})=\eta_i\,, \qquad i=1,\ldots,n-1\,, \tag{88.5}$$

erfüllt sind. Andererseits sind die Parameter so einzustellen, daß die Meßwerte y_j möglichst gut approximiert werden. Führen wir die Funktionen $F(a,\eta)=[F_j]\in\mathbb{R}^{m+1}$ und $G(a,\eta)=[G_j]\in\mathbb{R}^{n-1}$ über ihre Komponenten

$$\begin{aligned} F_j(a,\eta) \;&=\; u(t_j;a,\eta)-y_j\,, \qquad j=0,\ldots,m\,,\\ G_i(a,\eta) \;&=\; u_i(x_i;a,\eta_{i-1})-\eta_i\,, \quad i=1,\ldots,n-1\,, \end{aligned}$$

ein, so kann das Parameteridentifikationsproblem auch als *restringiertes Ausgleichsproblem* geschrieben werden:

$$\text{minimiere} \quad \frac{1}{2}\|F(a,\eta)\|_2^2 \quad \text{unter der Nebenbedingung} \quad G(a,\eta)=0\,. \tag{88.6}$$

Dies ist ein nichtlineares Ausgleichsproblem mit $m+1$ Gleichungen für $p+n$ Variablen und $n-1$ Nebenbedingungen, das mit Varianten der Methoden aus den Abschnitten 20 und 21 iterativ gelöst werden kann (vgl. etwa das Buch von Nash und Sofer [74]).

Wir skizzieren im folgenden eine Variante des Gauß-Newton-Verfahrens, bei der F und G jeweils um eine aktuelle Iterierte $(a^{(k)}, \eta^{(k)})$ linearisiert werden. Dadurch wird das Problem (88.6) in ein lineares Ausgleichsproblem mit linearen Nebenbedingungen überführt:

$$\text{minimiere} \quad \frac{1}{2}\,\|F + F_a h + F_\eta \zeta\|_2^2 \tag{88.7a}$$

bezüglich $h \in \mathbb{R}^p$ und $\zeta \in \mathbb{R}^n$ unter der Nebenbedingung

$$G + G_a h + G_\eta \zeta = 0\,. \tag{88.7b}$$

Dabei sind F, G und deren partielle Ableitungen jeweils an den aktuellen Näherungen $a^{(k)}$ und $\eta^{(k)}$ auszuwerten.

Das linearisierte Problem (88.7) führt auf das erweiterte lineare Gleichungssystem

$$\begin{bmatrix} I & 0 & -F_a & -F_\eta \\ 0 & 0 & -G_a & -G_\eta \\ -F_a^* & -G_a^* & 0 & 0 \\ -F_\eta^* & -G_\eta^* & 0 & 0 \end{bmatrix} \begin{bmatrix} r \\ \lambda \\ h \\ \zeta \end{bmatrix} = \begin{bmatrix} F \\ G \\ 0 \\ 0 \end{bmatrix}, \tag{88.8}$$

vgl. Aufgabe III.3. Hierbei ist $r = F + F_a h + F_\eta \zeta$ das zu minimierende Residuum und der Vektor λ hat die Rolle eines Lagrange-Parameters. Die Koeffizientenmatrix des Gleichungssystems (88.8) ist hermitesch, aber indefinit, vgl. Aufgabe V.9. Bei großen Problemen empfiehlt sich daher der Einsatz des GMRES-Verfahrens zur Lösung von (88.8).

Die Lösungskomponenten h und ζ aus (88.8) sind wie in Abschnitt 20 als Suchrichtungen zu verstehen; der alte Parametervektor $a^{(k)}$ und die alten Zwischenwerte $\eta^{(k)}$ sind also durch $a^{(k+1)} = a^{(k)} + \alpha h$ beziehungsweise $\eta^{(k+1)} = \eta^{(k)} + \alpha\zeta$ zu ersetzen, wobei $\alpha > 0$ eine sinnvolle Schrittweite bezeichnet. Ausgehend von $\alpha = 1$ ist diese Schrittweite solange zu halbieren, bis der Wert einer geeigneten Zielfunktion im Sinne des Armijo-Goldstein-Kriteriums (20.8) verbessert worden ist. Die Zielfunktion muß dabei sowohl das Kleinste-Quadrate-Funktional als auch die Nebenbedingung (88.5) berücksichtigen. Denkbar ist etwa eine Zielfunktion der Form

$$\Phi(a,\eta) \;=\; \frac{1}{2}\,\|F(a,\eta)\|_2^2 \;+\; \rho\|G(a,\eta)\|_2^2\,,$$

bei der die Nebenbedingung durch einen Strafterm mit großem Strafparameter $\rho \gg 0$ eingeht.

Der Vollständigkeit halber bestimmen wir noch die Einträge der Ableitungsmatrizen F_a, F_η und G_a, G_η. Die Einträge von F_η und G_η ergeben sich analog zu Proposition 87.1: Hinreichende Glattheit von f vorausgesetzt, erhält man im Intervall $x_{i-1} \le t \le x_i$ aus (88.3)

$$\frac{\partial}{\partial \eta_{i-1}} u_i(t; a, \eta_{i-1}) = w_i(t; a) \qquad \text{und} \qquad \frac{\partial}{\partial \eta_j} u_i(t; a, \eta_{i-1}) = 0\,, \quad j \neq i-1\,,$$

wobei w_i das lineare Anfangswertproblem

$$w_i' = f_u(t, u_i; a) w_i\,, \quad x_{i-1} \le t < x_i\,, \qquad w_i(x_{i-1}) = 1\,,$$

löst. Demnach ist G_η die Bidiagonalmatrix

$$G_\eta = \begin{bmatrix} w_1(x_1; a) & -1 & & & \\ & w_2(x_2; a) & -1 & & \\ & & \ddots & \ddots & \\ & & & w_{n-1}(x_{n-1}; a) & -1 \end{bmatrix} \in \mathbb{R}^{(n-1)\times n}\,.$$

Die Einträge von F_η ergeben sich entsprechend; es ist lediglich zu berücksichtigen, in welchem Gitterintervall von $\triangle$ sich die jeweiligen Meßpunkte befinden.

Darüber hinaus müssen noch die Ableitungen F_a und G_a ausgewertet werden. Da alle Teillösungen u_i, $i = 1, \ldots, m$, gleichermaßen von a abhängen, sind die Blockmatrizen F_a und G_a in der Regel voll besetzt. Eine analoge Überlegung wie in Proposition 87.1 führt auf das folgende Ergebnis:

$$\frac{\partial}{\partial a_k} u_i(t; a, \eta_{i-1}) = v_{ik}(t; \eta_{i-1})\,, \qquad t \in [x_{i-1}, x_i]\,, \tag{88.9}$$

wobei die Funktionen v_{ik}, $i = 1, \ldots, n$, $k = 1, \ldots, p$, die Anfangswertprobleme

$$v_{ik}' = f_u(t, u_i; a) v_{ik} + \frac{\partial f}{\partial a_k}(t, u_i; a)\,, \quad x_{i-1} \le t \le x_i\,, \qquad v_{ik}(x_{i-1}) = 0\,,$$

lösen, vgl. Aufgabe 10 (b). Aus den Funktionswerten von v_{ik} setzen sich die partiellen Ableitungen F_a und G_a zusammen.

Beispiel 88.1. In Abschnitt 60 wurde für die Weltbevölkerung $y(t)$ nach dem Verhulst-Modell ein Wachstum gemäß der logistischen Differentialgleichung

$$y' = y(d - ay) \tag{88.10}$$

mit den Parametern

$$d = 0.029 \qquad \text{und} \qquad a = 2.941 \cdot 10^{-3} \tag{88.11}$$

aus Braun [11, S. 39] angesetzt. Alternativ versuchen wir nun, die Parameter a und d dieses Modells in optimaler Weise an die auf Seite 467 angeführten sechs Bevölkerungszahlen $y_j = y(t_j)$, $j = 0, \ldots, 5$, zwischen den Jahren 1950 und 2000 anzupassen. Dazu verwenden wir die beschriebene Mehrzielmethode mit zwei Zeitintervallen und $x_0 = t_0 = 1950$, $x_1 = (t_2 + t_3)/2 = 1975$, $x_2 = 2000$. Neben a und d sind also auch die unbekannten Anfangswerte $\eta_0 = u_1(x_0)$ und $\eta_1 = u_2(x_1)$ für die Teillösungen u_i, $i = 1, 2$, in den beiden Gitterintervallen gesucht.

In diesem Beispiel ist $m = 5$, $n = 2$ sowie

$$f = u(d - au)\,, \qquad f_u = d - 2au\,, \qquad f_a = -u^2\,, \qquad f_d = u\,.$$

In jedem Iterationsschritt der beschriebenen Gauß-Newton-Variante ergibt sich daher das linearisierte Teilproblem (88.7) wie folgt: minimiere

$$\begin{bmatrix} v_{11}(t_0) & v_{12}(t_0) & w_1(t_0) & 0 \\ v_{11}(t_1) & v_{12}(t_1) & w_1(t_1) & 0 \\ v_{11}(t_2) & v_{12}(t_2) & w_1(t_2) & 0 \\ v_{21}(t_3) & v_{22}(t_3) & 0 & w_2(t_3) \\ v_{21}(t_4) & v_{22}(t_4) & 0 & w_2(t_4) \\ v_{21}(t_5) & v_{22}(t_5) & 0 & w_2(t_5) \end{bmatrix} \begin{bmatrix} \alpha \\ \delta \\ \zeta_0 \\ \zeta_1 \end{bmatrix} - \begin{bmatrix} y_0 - \eta_0 \\ y_1 - u_1(t_1) \\ y_2 - u_1(t_2) \\ y_3 - u_2(t_3) \\ y_4 - u_2(t_4) \\ y_5 - u_2(t_5) \end{bmatrix}$$

bezüglich der Euklidnorm unter der Nebenbedingung

$$\alpha v_{11}(x_1) + \delta v_{12}(x_1) + \zeta_0 w_1(x_1) - \zeta_1 \;=\; \eta_1 - u_1(x_1)\,.$$

Die acht Funktionen u_i, w_i, v_{i1} und v_{i2}, $i = 1, 2$, sind dabei aus den gekoppelten Differentialgleichungssystemen

$$\begin{aligned} u_i' &= u_i(d - au_i)\,, & u_i(x_{i-1}) &= \eta_{i-1}\,, \\ w_i' &= (d - 2au_i)w_i\,, & w_i(x_{i-1}) &= 1\,, \\ v_{i1}' &= (d - 2au_i)v_{i1} - u_i^2\,, & v_{i1}(x_{i-1}) &= 0\,, \\ v_{i2}' &= (d - 2au_i)v_{i2} + u_i\,, & v_{i2}(x_{i-1}) &= 0\,, \end{aligned} \qquad i = 1, 2\,,$$

zu bestimmen.

Für die Iteration wählen wir die Startparameter $a^{(0)} = 10^{-3}$ und $d^{(0)} = 0.01$ und interpolieren die Daten linear, um für den ersten Iterationsschritt geeignete

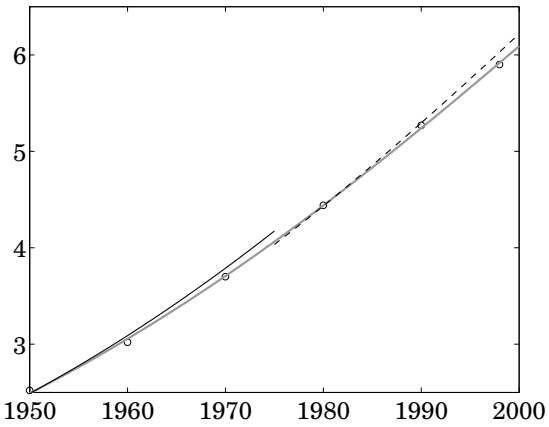


Abb. 88.1: Die optimale Lösung und die Funktion $u(t; a, \eta)$ nach einer Iteration

Anfangswerte $\eta_0^{(0)}$ und $\eta_1^{(0)}$ für die Differentialgleichungen zu konstruieren. Als Strafparameter für die Zielfunktion verwenden wir $\rho = 100$.

Das Verfahren benötigt drei Iterationen, um die optimalen Parameter

$$d = 2.559\ldots \cdot 10^{-2} \qquad \text{und} \qquad a = 1.862\ldots \cdot 10^{-3}$$

sowie den Startwert $\eta_0 = 2.490\ldots$ auf vier signifikante Stellen genau zu berechnen; man vergleiche diese Koeffizienten mit den Parametern aus (88.11). Abbildung 88.1 zeigt die optimale Lösung (die hellere Kurve). Darüber ist die Näherung u nach der ersten Iteration eingezeichnet; der durchgezogene Teil ist der Graph von u_1, die gebrochene Kurve gehört zu u_2. Wie man sieht, ist die Nebenbedingung zu diesem Zeitpunkt noch nicht erfüllt, d. h. $u_1(1975) \not\approx u_2(1975)$. ◊

Aufgaben

1. (a) Vervollständigen Sie den Beweis von Lemma 83.2.
(b) Geben Sie analoge Abschätzungen bezüglich der $\mathcal{L}^2$-Norm an, etwa

$$\|D_h[u] - u'\|_{\mathcal{L}^2(h,1-h)} \le C\|u'''\|_{\mathcal{L}^2(0,1)}h^2 \qquad \text{für ein } C > 0\,.$$

2. Seien $h_1, h_2 > 0$. Zeigen Sie:
(a) Der gewichtete Differenzenquotient

$$D[u](x) = \frac{h_2}{h_1+h_2}D^-_{h_1}[u](x) + \frac{h_1}{h_1+h_2}D^+_{h_2}[u](x)$$

stellt eine Approximation zweiter Ordnung an $u'(x)$ dar, d. h. es gilt $|D[u](x) - u'(x)| \le ch^2$ mit $h = \max\{h_1, h_2\}$ und einer von h unabhängigen Konstante $c > 0$.
(b) Im allgemeinen ist

$$D^2[u](x) = \frac{2}{h_1+h_2}\big(D^+_{h_2}[u](x) - D^-_{h_1}[u](x)\big)$$

lediglich eine Approximation erster Ordnung für $u''(x)$. Unter welcher Voraussetzung ist die Konsistenzordnung gleich Zwei?

3. Zeigen Sie, daß der Differenzenquotient

$$\frac{-11f(0) + 18f(h) - 9f(2h) + 2f(3h)}{6h}$$

eine Approximation dritter Ordnung von $f'(0)$ liefert, sofern f in einer Umgebung des Nullpunkts viermal stetig differenzierbar ist.

4. Gegeben sei das Randwertproblem (83.1) mit konstanten Koeffizienten $b \neq 0$ und $c \ge 0$. Betrachten Sie die Tridiagonalmatrix L_h aus (83.7), (83.8c) bei Verwendung zentraler Differenzen.
(a) Zeigen Sie, daß genau für $h < 2/|b|$ eine Diagonalmatrix $D \in \mathbb{R}^{(n-1)\times(n-1)}$ existiert, so daß DL_hD^{-1} symmetrisch ist, und geben Sie eine solche Matrix D an.
(b) Berechnen Sie für $h < 2/|b|$ die Eigenwerte von L_h.

5. Gegeben sei das Randwertproblem

$$-\big(a(x)u'(x)\big)' = f(x) \quad \text{in } (0,1)\,, \qquad u(0) = u(1) = 0\,,$$

mit einer in $[0,1]$ differenzierbaren Funktion $a > 0$. Diskretisieren Sie dieses Problem unter zweimaliger Verwendung des zentralen Differenzenquotienten $D_{h/2}[v]$ (also mit Gitterweite $h/2$ statt h) für die beiden Ableitungen. Geben Sie Bedingungen an a und u an, unter denen die Diskretisierung die Konsistenzordnung $q = 2$ hat.

6. Sei $A = [a_{ij}] \in \mathbb{R}^{n\times n}$ mit $a_{ij} \le 0$ für $i \neq j$. Zeigen Sie die Äquivalenz folgender Aussagen:
(a) A ist M-Matrix.
(b) Es gibt einen Vektor $0 < x \in \mathbb{R}^n$ mit $Ax > 0$ und es gilt $\|A^{-1}\|_\infty \le \|x\|_\infty / \min\limits_{1\le k\le n}(Ax)_k$.

(c) A hat positive Diagonaleinträge und für die Zerlegung von $A = D - N$ in Diagonal- und Nebendiagonalanteil gilt $\varrho(D^{-1}N) < 1$.
Hinweis zum Beweis von „(b)⇒(c)“: Zeigen Sie, daß $\|X^{-1}D^{-1}NX\|_\infty < 1$ für eine geeignete Diagonalmatrix X.

7. Die *Besselsche Differentialgleichung* der Ordnung $\nu \geq 0$,

$$x^2u'' + xu' + (x^2 - \nu^2)u = 0\,, \qquad x \in (0,2)\,,$$

besitzt genau eine Lösung $u \in C^2[0,2]$ mit Randwert $u(2) = 1$. (Darüber hinaus existieren noch weitere Lösungen, die im Nullpunkt jedoch unbeschränkt sind.)
(a) Mit dem Ansatz $u(x) = x^r \sum_{j=0}^{\infty} a_j x^j$, $a_0 \neq 0$, läßt sich diese beschränkte Lösung darstellen. Berechnen Sie die Koeffizienten a_j.
(b) Leiten Sie aus der Differentialgleichung eine Randbedingung in $x = 0$ her. Unterscheiden Sie die Fälle $\nu = 0$ und $\nu \neq 0$.
(c) Diskretisieren Sie das Randwertproblem mit zentralen Differenzenquotienten auf einem äquidistanten Gitter und stellen Sie das zugehörige Gleichungssystem auf. Verwenden Sie die Randbedingung aus (b).
(d) Implementieren Sie das Verfahren und vergleichen Sie für $\nu = 0$ und $\nu = 1$ und Gitterweiten $h = 2/N$, $N = 2^s$, $s = 4, 5, \ldots, 10$, den Fehler zwischen der numerischen und der exakten Lösung $J_\nu(x)/J_\nu(2)$. (In MATLAB werden die Besselfunktionen J_ν mit dem Befehl `besselj` zur Verfügung gestellt.) Erläutern Sie das unterschiedliche Fehlerverhalten für die beiden Parameter.

8. (a) Zeigen Sie, daß das Randwertproblem (84.3) mit $b \in C^k[0,1]$, $k \in \mathbb{N}_0$, eine eindeutig bestimmte Lösung $w \in C^{k+2}[0,1]$ besitzt. Substituieren Sie zunächst $v = w'$ und machen Sie einen Ansatz $v(x) = a(x)\exp(\int_0^x b(t)dt)$.
(b) Beweisen Sie die Darstellung

$$w(x+h) - 2w(x) + w(x-h) - h^2w''(x) \;=\; \int_0^h \int_{-\xi}^{\xi} \big(w''(x+t) - w''(x)\big)\,dt\,d\xi$$

für die zweite zentrale Differenz von w und folgern Sie daraus die Aussage von Bemerkung 84.6.

9. Implementieren Sie das Schießverfahren für Beispiel 87.2. Verwenden Sie etwa das klassische Runge-Kutta-Verfahren zur Lösung der Anfangswertprobleme.

10. (a) Sei $f(t,u;a) : [t_0,t_1] \times \mathbb{R}^n \times \mathbb{R}^p \to \mathbb{R}^n$ bezüglich u und dem Parametervektor $a = (a_1,\ldots,a_p)$ stetig differenzierbar und sei $u = u(\,\cdot\,;a,\eta)$ die Lösung des Anfangswertproblems

$$u' = f(t,u;a)\,, \quad t_0 \leq t \leq t_1\,, \qquad u(t_0) = \eta \quad \text{mit } \eta = (\eta_1,\ldots,\eta_n) \in \mathbb{R}^n\,.$$

Zeigen Sie, daß $u(t_1)$ partiell nach a und nach η differenzierbar ist mit

$$\frac{\partial}{\partial a_j}u(t_1;a,\eta) = v_j(t_1) \qquad \text{und} \qquad \frac{\partial}{\partial \eta_k}u(t_1;a,\eta) = w_k(t_1)\,,$$

wobei v_j, $j = 1,\ldots,p$, und w_k, $k = 1,\ldots,n$, durch die Anfangswertprobleme

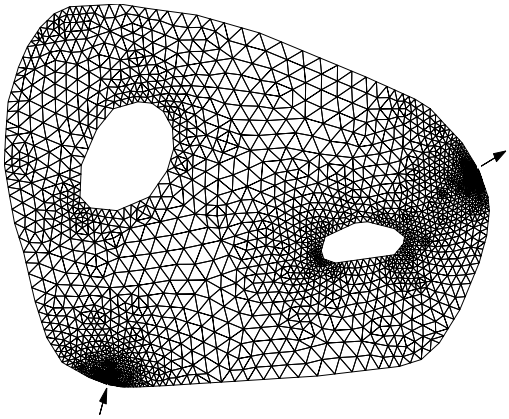
$$v_j' = f_u(t,u;a)v_j + \frac{\partial f}{\partial a_j}(t,u;a)\,, \quad t_0 \leq t \leq t_1\,, \qquad v(t_0) = 0\,,$$

beziehungsweise

$$w_k' = f_u(t,u)w_k\,, \quad t_0 \le t \le t_1\,, \qquad w(t_0) = e_k\,,$$

definiert sind. Hierbei ist f_u die partielle Ableitung von f nach u und e_k der k-te Einheitsvektor im $\mathbb{R}^n$.
(b) Verifizieren Sie die Darstellungen (87.10) und (88.9).
Hinweis zu (a): Gehen Sie vor wie im Beweis von Proposition 87.1 und verwenden Sie Aufgabe XIV.1.

Partielle Differentialgleichungen



XVI Elliptische Differentialgleichungen

Nach den gewöhnlichen Differentialgleichungen wenden wir uns nun partiellen Differentialgleichungen zu. Wir beginnen mit stationären Diffusionsprozessen, also Randwertaufgaben für elliptische Differentialgleichungen, deren numerische Behandlung grundsätzlich ähnlich ist zu der von Randwertproblemen für gewöhnliche Differentialgleichungen zweiter Ordnung. Da wir in Kapitel XV ausführlich Differenzenverfahren behandelt haben, konzentrieren wir uns im weiteren auf die *Methode der finiten Elemente* (FEM).

Das Buch von Hackbusch [43] ist als umfassende Einführung in numerische Methoden zur Lösung elliptischer Differentialgleichungen zu empfehlen, in der auch die funktionalanalytischen Grundlagen erarbeitet werden. Das Buch behandelt sowohl die Methode der finiten Elemente als auch Differenzenverfahren. Daneben seien die Bücher [10, 40, 61] genannt.

89 Schwache Lösungen

Sowohl die Theorie wie die Numerik elliptischer Randwertprobleme kann weitgehend unabhängig von der Raumdimension behandelt werden. Um die Notation möglichst einfach zu halten, beschränken wir uns daher durchweg auf Gleichungen in zwei (reellen) Variablen.

Punkte dieses zweidimensionalen Ortsraums werden im weiteren mit $x = [\xi, \eta]^T$ bezeichnet; wie in dem Modellierungsteil verwenden wir die Notationen $x_1 \cdot x_2 = \xi_1\xi_2 + \eta_1\eta_2$ und $|x| = (\xi^2 + \eta^2)^{1/2}$ für das euklidische Innenprodukt und die Euklidnorm im Ortsraum. Für eine reellwertige differenzierbare Funktion $u = u(\xi, \eta)$ interpretieren wir $\operatorname{grad} u = [u_\xi, u_\eta]^T$ wie zuvor als ein Element des Ortsraums.

Wir betrachten im folgenden beschränkte *polygonale Gebiete* $\Omega \subset \mathbb{R}^2$. Dies sind Gebiete, deren Rand $\Gamma = \partial\Omega$ aus einer endlichen Vereinigung geradliniger Kanten besteht. In wenigen Ausnahmefällen veranschaulichen wir jedoch Resultate der Einfachheit halber am Beispiel des Einheitskreises. In den zuge-

lassenen Gebieten gilt die *Greensche Formel*

$$\int_\Omega v\Delta u\,dx \;=\; \int_\Gamma v\,\frac{\partial u}{\partial \nu}\,ds \;-\; \int_\Omega \operatorname{grad} v\cdot\operatorname{grad} u\,dx \tag{89.1}$$

für $u\in C^2(\overline{\Omega})$ und $v\in C^1(\overline{\Omega})$; hierbei ist $\partial u/\partial\nu=\nu\cdot\operatorname{grad} u$ die Ableitung von u in Richtung der *äußeren* Normalen ν an den Rand des Gebiets (die Normalenableitung) und $\int_\Gamma\ldots ds$ das entsprechende Randintegral über Γ. Es sei bereits an dieser Stelle darauf hingewiesen, daß die Greensche Formel auch unter schwächeren Glattheitsvoraussetzungen an u und v gültig bleibt.

Eine wichtige Verallgemeinerung der Greenschen Formel ist die folgende Integralbeziehung (*partielle Integration*), vgl. Aufgabe 1: Sind $\sigma,u,v:\Omega\to\mathbb{R}$ hinreichend oft differenzierbar, dann gilt

$$\int_\Omega v\operatorname{div}(\sigma\operatorname{grad} u)\,dx \;=\; \int_\Gamma v\sigma\frac{\partial u}{\partial\nu}\,ds-\int_\Omega \sigma\operatorname{grad} v\cdot\operatorname{grad} u\,dx\,. \tag{89.2}$$

Bevor wir uns elliptischen Differentialgleichungen und ihrer Lösungstheorie zuwenden können, benötigen wir noch einige vorbereitende Definitionen. Als erstes führen wir in Analogie zum Eindimensionalen (vgl. Abschnitt 31) den *Sobolevraum* $H^1(\Omega)$ ein.

Definition 89.1. Der Raum $H^1(\Omega)\subset\mathcal{L}^2(\Omega)$ enthält alle reellwertigen Funktionen, deren partielle Ableitungen ebenfalls quadratisch integrierbar sind.[1] Das Innenprodukt in $H^1(\Omega)$ lautet

$$\langle\, u,v\,\rangle_{H^1(\Omega)} \;=\; \int_\Omega \operatorname{grad} u\cdot\operatorname{grad} v\,dx \;+\; \int_\Omega uv\,dx\,.$$

Neben der zugehörigen Norm $\|\cdot\|_{H^1(\Omega)}$ definieren wir noch eine Halbnorm in $H^1(\Omega)$ durch

$$|u|^2_{H^1(\Omega)} \;=\; \int_\Omega |\operatorname{grad} u|^2\,dx\,.$$

Der Raum $H^1(\Omega)$ ist für $\Omega\subset\mathbb{R}^2$ mit noch mehr Vorsicht anzugehen als für $\Omega\subset\mathbb{R}$; dies wird an dem folgenden Beispiel deutlich.

[1] Streng genommen wird hier ein verallgemeinerter (schwacher) Ableitungsbegriff verwendet. Für das Verständnis dieses Kapitels ist es jedoch ausreichend, sich unter Elementen von $H^1(\Omega)$ Funktionen vorzustellen, deren Gradient fast überall existiert und quadratisch integrierbar ist.

Beispiel 89.2. Ω sei der Einheitskreis im $\mathbb{R}^2$ und $u : \Omega \to \mathbb{R}$ sei definiert durch $u(x) = \log|\log(r/2)|$ mit $r = |x|$. Dann ist u stetig differenzierbar in $\Omega \setminus \{0\}$ mit Gradient

$$\operatorname{grad} u = \frac{1}{-\log(r/2)} \frac{-2}{r} \frac{1}{2} \operatorname{grad} r\,.$$

Wegen $r_\xi = \xi/r$ und $r_\eta = \eta/r$ ergibt sich

$$|\operatorname{grad} u|^2 = \frac{1}{r^2} \frac{1}{\log^2(r/2)} |\operatorname{grad} r|^2 = \frac{1}{r^2} \frac{1}{\log^2(r/2)}\,,$$

und folglich existiert das uneigentliche Integral

$$\int_\Omega |\operatorname{grad} u|^2\, dx = \int_0^{2\pi} \int_0^1 \frac{1}{r^2} \frac{1}{\log^2(r/2)}\, r\, dr\, d\theta = -2\pi \frac{1}{\log(r/2)}\Big|_0^1 = \frac{2\pi}{\log 2}\,.$$

Mit anderen Worten: Die Funktion u gehört zu $H^1(\Omega)$, obwohl u noch nicht einmal stetig ist ($u(x) \to \infty$ für $x \to 0$). ◊

Obwohl Funktionen $u \in H^1(\Omega)$ in einzelnen Punkten nicht stetig zu sein brauchen, kann man ihnen in sinnvoller Weise Randwerte $u|_\Gamma$ auf der Randkurve $\Gamma = \partial\Omega$ zuordnen. Dieser Sachverhalt ist der Gegenstand sogenannter *Spursätze*. ($u|_\Gamma$ wird auch als *Spur* von u auf Γ bezeichnet.)

Satz 89.3. *Jede Funktion $u \in H^1(\Omega)$ besitzt eine (im $\mathcal{L}^2$-Sinn) eindeutig bestimmte Spur $u|_\Gamma \in \mathcal{L}^2(\Gamma)$; die Zuordnung $u \mapsto u|_\Gamma$ ist eine lineare Abbildung mit*

$$\|u|_\Gamma\|^2_{\mathcal{L}^2(\Gamma)} \le c_\Omega \|u\|_{\mathcal{L}^2(\Omega)} \|u\|_{H^1(\Omega)}\,, \tag{89.3}$$

wobei die Konstante c_Ω nur von dem Gebiet Ω abhängt. Ist Ω ein Dreieck, so hängt c_Ω lediglich vom kleinsten Innenwinkel von Ω ab.

Beweis. Wir geben einen vergleichsweise einfachen Beweis der Ungleichung (89.3) für stetig differenzierbare Funktionen im Einheitskreis Ω, vgl. auch Aufgabe 3. Dazu führen wir wieder Polarkoordinaten ein, $x = [r\cos\theta, r\sin\theta]^T$, und verwenden die Darstellung

$$\begin{aligned} u^2(\cos\theta, \sin\theta) &= \int_0^1 \frac{d}{dr}\big(r^2 u^2(r\cos\theta, r\sin\theta)\big)\, dr \\ &= \int_0^1 2\big(ru^2 + r^2 u(u_\xi \cos\theta + u_\eta \sin\theta)\big)\, dr \\ &= \int_0^1 2(ru^2 + ru \operatorname{grad} u \cdot x)\, dr\,. \end{aligned}$$

Hieraus folgt mit der Cauchy-Schwarz-Ungleichung in $\mathbb{R}^2$, daß

$$u^2(\cos\theta,\sin\theta) \;\le\; \int_0^1 2(u^2+|u|\,|\operatorname{grad} u|)\,r\,dr\,,$$

und Integration über θ sowie die Substitutionsregel ergeben

$$\begin{aligned}\|u\|^2_{\mathcal{L}^2(\Gamma)} &= \int_0^{2\pi} u^2(\cos\theta,\sin\theta)\,d\theta \\ &\le\; 2\int_0^{2\pi}\int_0^1 (u^2+|u|\,|\operatorname{grad} u|)\,r\,dr\,d\theta \\ &=\; 2\int_\Omega (u^2+|u|\,|\operatorname{grad} u|)\,dx\,.\end{aligned}$$

Aus der Cauchy-Schwarz-Ungleichung folgt schließlich die Behauptung für den betrachteten Spezialfall:

$$\|u\|^2_{\mathcal{L}^2(\Gamma)} \;\le\; 2\left(\|u\|^2_{\mathcal{L}^2(\Omega)} + \|u\|_{\mathcal{L}^2(\Omega)}|u|_{H^1(\Omega)}\right) \;\le\; 4\|u\|_{\mathcal{L}^2(\Omega)}\,\|u\|_{H^1(\Omega)}\,.$$

Für den Beweis dieses Satzes bei polygonalen Gebieten sei auf die Bücher von Hackbusch [43] oder Braess [10] verwiesen. □

Als Konsequenz aus Satz 89.3 ist

$$H^1_0(\Omega) = \{\, u\in H^1(\Omega) \,:\, u|_\Gamma = 0\,\} \tag{89.4}$$

ein abgeschlossener linearer Unterraum von $H^1(\Omega)$. Für Funktionen $u\in H^1_0(\Omega)$ gilt darüber hinaus die wichtige *Poincaré-Friedrichs-Ungleichung*, vgl. [43, Lemma 6.2.11].

Lemma 89.4. *Es gibt eine Konstante $\gamma_\Omega>0$, die nur vom Gebiet Ω abhängt, so daß*

$$\gamma_\Omega\,\|u\|_{H^1(\Omega)} \;\le\; |u|_{H^1(\Omega)} \qquad \textit{für alle } u\in H^1_0(\Omega)\,. \tag{89.5}$$

Eine Abschätzung in die andere Richtung ist nach Definition 89.1 trivial (mit $\gamma_\Omega=1$). Für die Gültigkeit der hier formulierten Ungleichung ist entscheidend, daß die Randwerte $u|_\Gamma$ aufgrund der Voraussetzung $u\in H^1_0(\Omega)$ durch Null fixiert sind. Andernfalls könnte man zu u eine beliebige Konstante addieren, ohne daß sich die rechte Seite der Ungleichung ändert – wohl aber die linke Seite.

Bemerkung. Ein entsprechendes Resultat ist uns aus dem Eindimensionalen bekannt: Nach Aufgabe IX.8 (a) gilt nämlich

$$\begin{aligned} \tfrac{\pi^2}{\pi^2+1}\,\|u\|^2_{H^1(0,1)} &= \tfrac{\pi^2}{\pi^2+1}\left(\|u\|^2_{\mathcal{L}^2(0,1)} + \|u'\|^2_{\mathcal{L}^2(0,1)}\right) \\ &\le \left(\tfrac{1}{\pi^2+1} + \tfrac{\pi^2}{\pi^2+1}\right)\|u'\|^2_{\mathcal{L}^2(0,1)} = \|u'\|^2_{\mathcal{L}^2(0,1)} \end{aligned}$$

für alle $u \in H^1_0(0,1)$. Ohne die Randbedingung $u(0) = u(1) = 0$ ist das Ergebnis hingegen falsch, wie das Beispiel der konstanten Funktion $u = 1$ aus $H^1(0,1)$ belegt. ◊

Nach diesen Vorbereitungen betrachten wir nun die Differentialgleichung

$$L[u] = -\operatorname{div}(\sigma \operatorname{grad} u) + cu = f \qquad \text{in } \Omega \tag{89.6a}$$

mit der Dirichlet-Randbedingung

$$u = 0 \qquad \text{auf } \Gamma\,. \tag{89.6b}$$

Hierbei seien σ, c und f stetige Funktionen in $\overline{\Omega}$ mit $c(x) \ge 0$. Der Differentialoperator L aus (89.6a) heißt *elliptisch*, falls σ von unten durch eine positive Konstante beschränkt ist, $\sigma(x) \ge \sigma_0 > 0$. Für den Moment fordern wir zusätzlich, daß σ in Ω stetig differenzierbar ist. Für ein eindimensionales Intervall $\Omega = (0,1)$ entsprechen die obigen Bedingungen unseren Voraussetzungen aus Kapitel XV an das Randwertproblem (83.1).

Unter einer *klassischen Lösung* u von (89.6) verstehen wir eine Funktion $u \in C^2(\overline{\Omega})$ mit $u|_\Gamma = 0$, die der Differentialgleichung genügt. Ist u eine solche Lösung und $v \in C^1(\overline{\Omega})$, dann ergibt sich aus (89.6a) und (89.2)

$$\begin{aligned} \int_\Omega f v\,dx &= -\int_\Omega v\,\operatorname{div}(\sigma \operatorname{grad} u)\,dx + \int_\Omega cuv\,dx \\ &= \int_\Omega \sigma \operatorname{grad} u \cdot \operatorname{grad} v\,dx + \int_\Omega cuv\,dx - \int_\Gamma v\,\sigma\,\frac{\partial u}{\partial \nu}\,ds\,. \end{aligned}$$

Falls darüberhinaus $v|_\Gamma = 0$ ist, erhalten wir also

$$\int_\Omega \sigma \operatorname{grad} u \cdot \operatorname{grad} v\,dx + \int_\Omega cuv\,dx = \int_\Omega f v\,dx\,. \tag{89.7}$$

Dies ist die sogenannte *schwache Form* des Randwertproblems.

Bisher hatten wir gefordert, daß die Lösung der Randwertaufgabe (89.6) zweimal stetig differenzierbar ist und daß v in (89.7) zumindest einmal stetig differenzierbar ist. Tatsächlich kann die Gleichung (89.7) bereits unter schwächeren Anforderungen an u und v sinnvoll formuliert werden: Wegen der Gültigkeit der Cauchy-Schwarz-Ungleichung existieren die Integrale aus (89.7) bereits dann, wenn u und v zu $H^1(\Omega)$ gehören. Daher kann der Lösungsbegriff für (89.6) folgendermaßen abgeschwächt werden:

Definition und Satz 89.5. *Es seien σ und c beschränkte, meßbare Funktionen mit $0 \le c(x) \le c_\infty$ und $0 < \sigma_0 \le \sigma(x) \le \sigma_\infty$. Dann hat das Randwertproblem* (89.6) *für jedes $f \in \mathcal{L}^2(\Omega)$ eine eindeutig bestimmte* schwache Lösung *$u \in H_0^1(\Omega)$ mit*

$$\int_\Omega \sigma \operatorname{grad} u \cdot \operatorname{grad} v \, dx + \int_\Omega cuv \, dx = \int_\Omega fv \, dx$$

für alle $v \in H_0^1(\Omega)$. Ist zudem σ differenzierbar und liegt u in $C^2(\overline{\Omega})$, dann ist u auch eine klassische Lösung von (89.6).

Für einen Beweis dieses Resultats verweisen wir wieder auf das Buch von Hackbusch [43].

Zur Illustration dieses schwächeren Lösungsbegriffs untersuchen wir das folgende eindimensionale Beispiel, für das Satz 89.5 entsprechend gilt.

Beispiel 89.6. Sei $\xi \in (0,1)$. Das Randwertproblem

$$-(\sigma u')' = x \quad \text{in } (0,1)\,, \qquad u(0) = u(1) = 0\,, \tag{89.8}$$

mit

$$\sigma(x) = \begin{cases} 1\,, & x < \xi\,, \\ 2\,, & x \ge \xi\,, \end{cases}$$

kann keine klassische Lösung besitzen. Wäre nämlich u eine solche Lösung, so gilt offenbar $-u''(x) = x$ für $x < \xi$ und $-2u''(x) = x$ für $x > \xi$ und dies führt unmittelbar auf einen Widerspruch:

$$-\xi = \lim_{x \to \xi -} u''(x) \neq \lim_{x \to \xi +} u''(x) = -\xi/2\,.$$

Nach Satz 89.5 existiert jedoch eine schwache Lösung $u \in H^1(0,1)$ dieses Problems; insbesondere ist u stetig (im Eindimensionalen ist das nach Abschnitt 31 der Fall) und der Funktionswert $u_0 = u(\xi)$ ist wohldefiniert. Wir setzen nun u in den zwei Teilintervallen durch die Lösungen der beiden Randwertaufgaben

$$\begin{aligned} -u'' &= x \quad \text{in } (0,\xi)\,, \qquad u(0) = 0\,, \;\; u(\xi) = u_0\,, \\ -2u'' &= x \quad \text{in } (\xi,1)\,, \qquad u(\xi) = u_0\,,\, u(1) = 0\,, \end{aligned}$$

an. Diese Probleme haben die klassischen Lösungen

$$\begin{aligned} u(x) &= -x^3/6 + a_0 x + b_0\,, \qquad 0 \le x \le \xi\,, \\ u(x) &= -x^3/12 + a_1 x + b_1\,, \qquad \xi \le x \le 1\,, \end{aligned} \tag{89.9}$$

mit geeigneten Koeffizienten a_0, b_0, a_1, b_1. Zusammen mit u_0 sind das fünf unbekannte Parameter, für die bislang vier Gleichungen vorliegen (die Randbedingungen). Eine fünfte Gleichung wird durch die Differentialgleichung nahegelegt: Interpretieren wir (89.8) so, daß die Funktion $\sigma u'$ in $H^1(0,1)$ liegt und die lineare Funktion x auf der rechten Seite von (89.8) ihre schwache Ableitung ist, so folgt daraus, daß $\sigma u'$ insbesondere im Punkt $x = \xi$ stetig sein muß. Wir setzen dies als fünfte Bedingung an. Dann erhalten wir das folgende Gleichungssystem für die fünf Parameter:

$$\begin{aligned} b_0 &= 0\,, & a_1 + b_1 &= 1/12\,, \\ \xi a_0 + b_0 &= u_0 + \xi^3/6\,, & \xi a_1 + b_1 &= u_0 + \xi^3/12\,, \\ a_0 - 2a_1 &= \xi^2/2 - \xi^2/2 = 0\,. \end{aligned}$$

Es ist nicht schwer zu zeigen, daß dieses Gleichungssystem eindeutig lösbar ist. Da die zugehörige Funktion u aus (89.9) in den jeweiligen Teilintervallen die Differentialgleichung löst, ergibt sich durch partielle Integration für jedes $v \in H_0^1(0,1)$

$$\begin{aligned} \int_0^1 \sigma u'v'\,dx \;&=\; \int_0^\xi u'v'\,dx \;+\; 2\int_\xi^1 u'v'\,dx \\ &=\; u'v\big|_0^\xi \;-\; \int_0^\xi vu''\,dx \;+\; 2u'v\big|_\xi^1 \;-\; 2\int_\xi^1 vu''\,dx \\ &=\; u'(\xi-)v(\xi) \;+\; \int_0^\xi xv\,dx \;-\; 2u'(\xi+)v(\xi) \;+\; \int_\xi^1 xv\,dx \\ &=\; v(\xi)\big(u'(\xi-) - 2u'(\xi+)\big) \;+\; \int_0^1 xv\,dx\,. \end{aligned}$$

Da nach Konstruktion $\sigma u'$ im Punkt $x = \xi$ stetig ist, hat die Differenz in der runden Klammer den Wert 0 und demnach gilt tatsächlich

$$\int_0^1 \sigma u'v'\,dx = \int_0^1 xv\,dx \qquad \text{für alle } v \in H_0^1(0,1)\,,$$

d. h. u ist die gesuchte schwache Lösung von (89.8); die gepunkteten Kurven in Abbildung 89.1 zeigen die beiden Funktionen aus der Definition (89.9) über dem gesamten Intervall $[0,1]$, die durchgezogene Kurve ist der Graph der schwachen Lösung u.

Diese Lösung ist darüber hinaus physikalisch sinnvoll: Wie wir aus Beispiel 70.1 wissen, beschreibt die Differentialgleichung (89.8) beispielsweise das elektrische Potential u in einem inhomogenen Stab mit Leitfähigkeit $\sigma = 1$ im linken

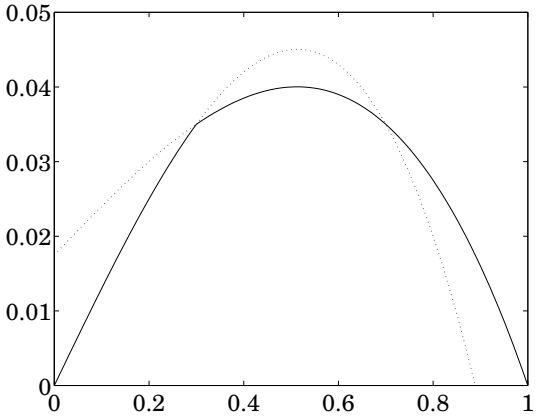


Abb. 89.1: Lösung von (89.8) für $\xi = 0.3$

Teil ($0 < x < \xi$) und Leitfähigkeit $\sigma = 2$ im rechten Teil ($\xi < x < 1$). In diesem Fall gibt $J = -\sigma u'$ den elektrischen Strom an, der aufgrund des Erhaltungsgesetzes im Stab stetig sein muß, sofern die rechte Seite (die Quelle $f = x$) keine Sprünge aufweist. Dies ist gerade die Bedingung, die zuvor die fünfte Gleichung ergeben hat und die dafür verantwortlich ist, daß die Lösung u im Punkt $x = \xi$ in der Regel nicht stetig differenzierbar ist. ◊

89.1 Inhomogene Dirichlet-Randbedingungen

Bislang hatten wir eine Lösung u der Differentialgleichung gesucht, die auf dem Rand Γ verschwindet. Wir wollen nun noch zwei alternative Randbedingungen untersuchen. Als erstes betrachten wir den Fall, daß eine von Null verschiedene Dirichlet-Randbedingung

$$u|_\Gamma = g \tag{89.10}$$

für die Lösung u von (89.6a) vorgegeben ist.

Ist u eine klassische Lösung dieses *inhomogenen Dirichlet-Problems*, so ergibt sich auf die gleiche Weise wie in (89.7) die schwache Form des Randwertproblems:

$$\int_\Omega \sigma \operatorname{grad} u \cdot \operatorname{grad} v \, dx \; + \; \int_\Omega cuv \, dx \; = \; \int_\Omega fv \, dx \tag{89.11}$$

für alle $v \in C^1(\overline{\Omega})$ mit $v|_\Gamma = 0$. Entsprechend nennen wir eine Lösung $u \in H^1(\Omega)$ eine schwache Lösung des inhomogenen Dirichletproblems (89.6a),

(89.10), falls u die Randbedingung (89.10) erfüllt und (89.11) für alle $v \in H_0^1(\Omega)$ gilt.

Bei linearen Problemen läßt sich das inhomogene Dirichlet-Problem auf ein homogenes Problem mit Randvorgabe Null zurückführen, wenn man eine spezielle Funktion u_0 mit korrekten Randwerten $u_0|_\Gamma = g$ konstruieren kann: In diesem Fall löst nämlich $w = u - u_0$ das homogene Dirichletproblem

$$L[w] \;=\; f - L[u_0]\,, \qquad w|_\Gamma = 0\,.$$

Allerdings ist die Existenz einer entsprechenden Funktion u_0 nicht für jede beliebige Randvorgabe $g \in \mathcal{L}^2(\Gamma)$ gesichert. Dies führt auf die Fragestellung sogenannter *Fortsetzungssätze*: Für welche Funktionen $g \in \mathcal{L}^2(\Gamma)$ existiert eine Funktion $u_0 \in H^1(\Omega)$ mit $u_0|_\Gamma = g$? Zur Beantwortung dieser Frage verweisen wir auf die Literatur, etwa auf [43, Satz 6.2.40]. Wir stellen lediglich fest, daß das inhomogene Dirichlet-Problem (89.6a), (89.10) nicht für jede Randvorgabe $g \in \mathcal{L}^2(\Gamma)$ eine (schwache) Lösung $u \in H^1(\Omega)$ besitzt.

89.2 Neumann-Randbedingungen

Zum Abschluß dieses Abschnitts wenden wir uns nun noch *Neumann-Randbedingungen* zu, also einer Vorgabe von $\partial u/\partial\nu$ auf dem Rand Γ; üblich ist dabei meist eine Formulierung über den *Fluß* $\sigma\partial u/\partial\nu$. Wie zuvor überführen wir zunächst das Randwertproblem

$$-\operatorname{div}(\sigma \operatorname{grad} u) + cu = f \quad \text{in } \Omega\,, \qquad \sigma\,\frac{\partial u}{\partial \nu}\Big|_\Gamma = g\,, \tag{89.12}$$

mit vorgegebenem $g \in \mathcal{L}^2(\Gamma)$ in schwache Form. Dazu multiplizieren wir die Differentialgleichung diesmal mit einer beliebigen Funktion $v \in H^1(\Omega)$ und integrieren:

$$\begin{aligned}
\int_\Omega f v\,dx \;&=\; -\int_\Omega v\,\operatorname{div}(\sigma \operatorname{grad} u)\,dx \;+\; \int_\Omega cuv\,dx \\
&= \int_\Omega \sigma \operatorname{grad} u \cdot \operatorname{grad} v\,dx \;+\; \int_\Omega cuv\,dx \;-\; \int_\Gamma v\,\sigma\,\frac{\partial u}{\partial \nu}\,ds \\
&= \int_\Omega \sigma \operatorname{grad} u \cdot \operatorname{grad} v\,dx \;+\; \int_\Omega cuv\,dx \;-\; \int_\Gamma vg\,ds\,.
\end{aligned}$$

Damit ist

$$\int_\Omega \sigma \operatorname{grad} u \cdot \operatorname{grad} v\,dx \;+\; \int_\Omega cuv\,dx \;=\; \int_\Omega f v\,dx \;+\; \int_\Gamma vg\,ds \tag{89.13}$$

für alle $v \in H^1(\Omega)$ die schwache Form der Differentialgleichung (89.12) mit Neumann-Randbedingungen. Eine Lösung $u \in H^1(\Omega)$ von (89.13) heißt schwache Lösung des Randwertproblems (89.12).

Im Gegensatz zum Dirichlet-Problem wird bei Neumann-Problemen die Randbedingung nicht explizit im *Ansatzraum* für u vorgegeben, sondern steckt implizit in der Präsenz des Randintegrals in (89.13), das aufgrund der Variation von v in dem größeren *Testraum* $H^1(\Omega)$ hier nicht wegfällt.

Satz 89.7. *Es seien σ und c beschränkte Funktionen mit $\sigma(x) \geq \sigma_0 > 0$ und $c(x) \geq c_0 > 0$. Dann hat die Differentialgleichung* (89.12) *für jedes $f \in \mathcal{L}^2(\Omega)$ und jedes $g \in \mathcal{L}^2(\Gamma)$ eine eindeutig bestimmte schwache Lösung $u \in H^1(\Omega)$. Für $c = 0$ existieren schwache Lösungen genau dann, wenn zusätzlich*

$$\int_\Omega f(x)\,dx = -\int_\Gamma g(s)\,ds\,; \tag{89.14}$$

in diesem Fall wird durch $\int_\Omega u(x)\,dx = 0$ eine Lösung u eindeutig festgelegt, alle anderen Lösungen haben die Form $u+u_0$ mit beliebiger Konstanten $u_0 \in \mathbb{R}$.

Für einen Beweis von Satz 89.7 vergleiche man die Diskussion in [43, Abschnitt 7.4]. Hier sei lediglich angemerkt, daß sich im Fall $c = 0$ durch Einsetzen der konstanten Funktion $v = 1$ in (89.13) die notwendige Lösbarkeitsbedingung (89.14) ergibt. Daß umgekehrt die Differentialgleichung (89.12) für $c = 0$ nicht eindeutig lösbar sein kann, liegt auf der Hand; in der Elektrostatik können beispielsweise aufgrund der unbekannten Konstanten u_0 nur Spannungen, d. h. Potentialdifferenzen, bestimmt werden, vgl. Beispiel 70.1.

90 Das Galerkin-Verfahren

Wir beschränken uns im weiteren größtenteils auf das homogene Dirichlet-Problem und verweisen für die anderen Randbedingungen auf das Ende dieses Abschnitts. Zunächst vereinfachen wir die Schreibweise und definieren

$$a(u,v) = \int_\Omega \sigma \operatorname{grad} u \cdot \operatorname{grad} v\,dx \;+\; \int_\Omega cuv\,dx \tag{90.1}$$

und

$$\ell(v) \;=\; \int_\Omega fv\,dx\,. \tag{90.2}$$

Offensichtlich ist ℓ eine lineare Abbildung von $V = H^1(\Omega)$ nach $\mathbb{R}$. Die reellwertige Funktion a ist in beiden Argumenten linear, also eine *Bilinearform* über V.

Definition 90.1. Sei V ein reeller Vektorraum mit Norm $\|\cdot\|_V$ und W ein abgeschlossener linearer Unterraum von V. Eine Bilinearform $a : V \times V \to \mathbb{R}$ heißt

- *symmetrisch*, falls $a(u,v) = a(v,u)$ für alle $u, v \in V$,
- *stetig*, falls eine Konstante a_∞ existiert mit

$$|\, a(u,v)\,| \;\le\; a_\infty \|u\|_V \|v\|_V \qquad \text{für alle } u, v \in V,$$

- *W-elliptisch*, falls eine Konstante $a_0 > 0$ existiert mit

$$a(w,w) \;\ge\; a_0 \|w\|_V^2 \qquad \text{für alle } w \in W.$$

Beispiel. Induziert die Norm $\|\cdot\|_V$ ein (reelles) Innenprodukt, dann ist dieses Innenprodukt ein Beispiel für eine symmetrische, stetige und V-elliptische Bilinearform. Dabei ist $a_0 = a_\infty = 1$ und die Stetigkeitsforderung entspricht gerade der Cauchy-Schwarz-Ungleichung. ◊

Ein weiteres Beispiel ist die Bilinearform (90.1):

Proposition 90.2. *Unter den Voraussetzungen von Satz* 89.5 *ist die Bilinearform a aus* (90.1) *symmetrisch, stetig in $H^1(\Omega)$ mit $a_\infty = \max\{\sigma_\infty, c_\infty\}$ und $H_0^1(\Omega)$-elliptisch mit $a_0 = \gamma_\Omega^2 \sigma_0$. Hierbei ist γ_Ω die Konstante aus der Poincaré-Friedrichs-Ungleichung* (89.5). *Darüber hinaus ist $a(v,v) \ge 0$ für alle $v \in H^1(\Omega)$.*

Beweis. Die Symmetrie von a ist offensichtlich. Aus der Cauchy-Schwarz-Ungleichung (zunächst im Ortsraum und dann in $\mathcal{L}^2(\Omega)$) folgt für beliebige $u, v \in H^1(\Omega)$

$$\begin{aligned} |\, a(u,v)\,| \;&\le\; \int_\Omega \sigma\, |\operatorname{grad} u|\, |\operatorname{grad} v|\, dx \;+\; \int_\Omega c\, |uv|\, dx \\ &\le\; \sigma_\infty |u|_{H^1(\Omega)} |v|_{H^1(\Omega)} \;+\; c_\infty \|u\|_{\mathcal{L}^2(\Omega)} \|v\|_{\mathcal{L}^2(\Omega)} \\ &\le\; \max\{\sigma_\infty, c_\infty\}\, \|u\|_{H^1(\Omega)} \|v\|_{H^1(\Omega)}\,. \end{aligned}$$

Dies ist gerade die behauptete Stetigkeit von a. Ferner ist

$$a(v,v) \;=\; \int_\Omega \sigma\, |\operatorname{grad} v|^2\, dx + \int_\Omega c v^2\, dx \;\ge\; \sigma_0 |v|_{H^1(\Omega)}^2$$

und die rechte Seite ist für jedes $v \in H^1(\Omega)$ nichtnegativ. Speziell für $v \in H_0^1(\Omega)$ kann diese rechte Seite mit der Poincaré-Friedrichs-Ungleichung (89.5) weiter abgeschätzt werden:

$$a(v,v) \;\ge\; \sigma_0 |v|_{H^1(\Omega)}^2 \;\ge\; \gamma_\Omega^2 \sigma_0 \|v\|_{H^1(\Omega)}^2\,.$$

Daher ist a $H_0^1(\Omega)$-elliptisch. □

Mit diesen Definitionen läßt sich die schwache Lösung $u \in H_0^1(\Omega)$ des Randwertproblems (89.6) als Lösung des *Variationsproblems*

$$a(u,v) = \ell(v) \qquad \text{für alle } v \in H_0^1(\Omega) \tag{90.3}$$

charakterisieren, der sogenannten *Variationsformulierung* des Randwertproblems.

Um eine Näherungslösung u_h für u zu bestimmen, bietet sich das *Galerkin-Verfahren* an, das uns bereits in Abschnitt 59 begegnet ist: Hierzu wählt man einen (endlichdimensionalen) Teilraum $V_h \subset H_0^1(\Omega)$ als Ansatz- und Testraum, sucht also eine Funktion $u_h \in V_h$ mit

$$a(u_h,v) = \ell(v) \qquad \text{für alle } v \in V_h\,. \tag{90.4}$$

Ist $\{\phi_1,\ldots,\phi_n\}$ eine Basis von V_h, dann führt der Lösungsansatz

$$u_h = \sum_{i=1}^{n} u_i \phi_i$$

unmittelbar auf das lineare Gleichungssystem $A\boldsymbol{u}_h = b$ für $\boldsymbol{u}_h = [u_1,\ldots,u_n]^T$ mit

$$A = \big[\, a(\phi_i,\phi_j) \,\big]_{ij} \in \mathbb{R}^{n\times n}\,, \qquad b = \big[\, \ell(\phi_j) \,\big]_j \in \mathbb{R}^n\,. \tag{90.5}$$

Die Matrix A wird *Steifigkeitsmatrix* genannt.

Beispiel 90.3. Für das eindimensionale Beispiel $-u'' = f$ in $(0,1)$ mit Randvorgabe $u(0) = u(1) = 0$ wählen wir für V_h den Raum der linearen Splines über einem äquidistanten Gitter $\triangle_h = \{x_i = ih \,:\, 0 \le i \le n,\, h = 1/n\}$ mit homogenen Randwerten. Für die nodale Basis $\phi_i = \Lambda_i$, $i = 1,\ldots,n-1$, von V_h ergeben sich dann die folgenden Matrixeinträge der Steifigkeitsmatrix:

$$a(\Lambda_i,\Lambda_j) \;=\; \int_0^1 \Lambda_i'(x)\Lambda_j'(x)\,dx \;=\; \begin{cases} 2/h\,, & i = j\,, \\ -1/h\,, & |i-j| = 1\,, \\ 0\,, & \text{sonst}\,. \end{cases}$$

Somit hat das Gleichungssystem $A\boldsymbol{u}_h = b$ die Gestalt

$$\frac{1}{h}\begin{bmatrix} 2 & -1 & & \\ -1 & 2 & \ddots & \\ & \ddots & \ddots & -1 \\ & & -1 & 2 \end{bmatrix} \begin{bmatrix} u_1 \\ u_2 \\ \vdots \\ u_{n-1} \end{bmatrix} = \begin{bmatrix} b_1 \\ b_2 \\ \vdots \\ b_{n-1} \end{bmatrix}$$

mit $b_i = \int_0^1 f(x)\Lambda_i(x)\,dx$, $i = 1, \ldots, n-1$. Man vergleiche dies mit Beispiel 83.1, in dem dieselbe Randwertaufgabe mit einem Differenzenverfahren diskretisiert wird: Die beiden resultierenden Gleichungssysteme stimmen überein, wenn b_i durch die Trapezsumme über $\triangle_h$ approximiert wird. ◊

Proposition 90.4. *Es gelten die Voraussetzungen von Satz* 89.5. *Dann ist die Matrix A aus* (90.5) *symmetrisch und positiv definit.*

Beweis. Zunächst ist $A = [a_{ij}]$ offensichtlich symmetrisch, da a eine symmetrische Bilinearform ist:

$$a_{ij} = a(\phi_i, \phi_j) = a(\phi_j, \phi_i) = a_{ji}\,, \qquad i, j = 1, \ldots, n\,.$$

Ferner gilt für einen beliebigen Vektor $\boldsymbol{v} = [v_1, \ldots, v_n]^T \in \mathbb{R}^n$ und die zugehörige Funktion $v = \sum_{i=1}^n v_i\phi_i \in V_h$

$$\boldsymbol{v}^* A\boldsymbol{v} \;=\; \sum_{i,j=1}^{n} v_i v_j\, a(\phi_i, \phi_j) \;=\; a\Big(\sum_{i=1}^{n} v_i\phi_i\,,\, \sum_{j=1}^{n} v_j\phi_j\Big) \;=\; a(v, v)\,.$$

Nach Proposition 90.2 ist die rechte Seite nichtnegativ und daher A positiv semidefinit. Darüber hinaus ist a auch $H_0^1(\Omega)$-elliptisch und $V_h \subset H_0^1(\Omega)$; daher ist $\boldsymbol{v}^* A\boldsymbol{v}$ genau dann gleich Null, wenn die Funktion v identisch Null ist, also nur für $\boldsymbol{v} = 0$. □

Unter den genannten Voraussetzungen hat somit das lineare Gleichungssystem $A\boldsymbol{u}_h = b$ für das Dirichlet-Problem eine eindeutige Lösung $\boldsymbol{u_h}$, d. h. das Galerkin-Verfahren liefert eine eindeutig bestimmte Lösung $u_h \in V_h$.

Lemma 90.5 (Lemma von Céa). *Es seien σ und c beschränkte Funktionen mit $0 < \sigma_0 \le \sigma(x) \le \sigma_\infty$ und $0 \le c(x) \le c_\infty$. Ferner sei $f \in \mathcal{L}^2(\Omega)$ und u und u_h bezeichnen die zugehörige schwache Lösung des Dirichlet-Problems* (89.6) *beziehungsweise die Approximation* (90.4) *des Galerkin-Verfahrens. Dann ist*

$$\|u - u_h\|_{H^1(\Omega)} \;\le\; \frac{\max\{\sigma_\infty, c_\infty\}}{\gamma_\Omega^2 \sigma_0} \inf_{v \in V_h} \|u - v\|_{H^1(\Omega)}\,,$$

wobei γ_Ω die Konstante aus der Poincaré-Friedrichs-Ungleichung (89.5) *bezeichnet.*

Beweis. Aus (90.3) und (90.4) folgt für jedes beliebige $v \in V_h$

$$\begin{aligned} a(u - u_h, u - u_h) &= a(u - u_h, u - v) + a(u, v - u_h) - a(u_h, v - u_h) \\ &= a(u - u_h, u - v) + \ell(v - u_h) - \ell(v - u_h) \\ &= a(u - u_h, u - v)\,. \end{aligned}$$

Mit der Stetigkeit und der $H^1_0(\Omega)$-Elliptizität von a folgt hieraus unmittelbar

$$\begin{aligned} a_0 \|u-u_h\|^2_{H^1(\Omega)} &\le a(u-u_h, u-u_h) = a(u-u_h, u-v) \\ &\le a_\infty \|u-u_h\|_{H^1(\Omega)} \|u-v\|_{H^1(\Omega)} \end{aligned}$$

für jedes $v \in V_h$. Nach Proposition 90.2 kann $a_0 = \gamma_\Omega^2 \sigma_0$ und $a_\infty = \max\{\sigma_\infty, c_\infty\}$ gewählt werden. Hieraus folgt die Behauptung. □

Das Lemma von Céa zeigt also, daß der Fehler der Galerkin-Approximation u_h bezüglich der H^1-Norm um höchstens einen konstanten Faktor schlechter ist als der Abstand der Lösung zu ihrer Bestapproximation aus V_h. Damit läßt sich der Fehler im wesentlichen über die Wahl geeigneter Ansatzräume V_h steuern.

Bemerkung 90.6 (Inhomogenes Dirichlet-Problem). Bei dem inhomogenen Dirichlet-Problem wählen wir wie in Abschnitt 89.1 zunächst eine Funktion $u_0 \in H^1(\Omega)$ mit $u_0|_\Gamma = g$ und betrachten die Variationsgleichung

$$a(w,v) = \ell(v) - a(u_0, v) \qquad \text{für alle } v \in H^1_0(\Omega) \tag{90.6}$$

für die Differenz $w = u - u_0 \in H^1_0(\Omega)$. Das Galerkin-Verfahren liefert wie zuvor eine Approximation $w_h \in V_h$ von w und für $u_h = w_h + u_0$ folgt aus dem Céa-Lemma 90.5 die Abschätzung

$$\|u-u_h\|_{H^1(\Omega)} = \|w-w_h\|_{H^1(\Omega)} \le \frac{\max\{\sigma_\infty, c_\infty\}}{\gamma_\Omega^2 \sigma_0} \inf_{v \in V_h} \|w-v\|_{H^1(\Omega)} .$$

◊

Bemerkung 90.7 (Neumann-Problem). Für das Neumann-Problem muß die rechte Seite $\ell(v)$ des Variationsproblems (90.3) zu

$$\ell(v) = \int_\Omega f v \, dx + \int_\Gamma g v \, ds$$

abgeändert werden, vgl. (89.13). Zudem durchläuft v in (90.3) alle Elemente aus $H^1(\Omega)$ und die Lösung u wird seinerseits in $H^1(\Omega)$ gesucht; für den Ansatz- und Testraum V_h darf man sich daher nicht nur auf Funktionen mit homogenen Randwerten auf Γ beschränken. Als Konsequenz aus diesen Modifikationen ist die Steifigkeitsmatrix A in der Regel nur positiv semidefinit; wenn etwa $c = 0$ ist und die konstanten Funktionen zum Ansatzraum V_h gehören, liegen die zugehörigen Koeffizientenvektoren im Kern von A. Dies spiegelt den kontinuierlichen Lösungsraum wider, da sich in dem betrachteten Fall verschiedene Lösungen um eine additive Konstante unterscheiden können. Ist hingegen $c \ge c_0 > 0$, dann ergibt sich auch in diesem Fall eine positiv definite Steifigkeitsmatrix und das Céa-Lemma 90.5 gilt entsprechend mit der Konstanten $\max\{\sigma_\infty, c_\infty\} / \min\{\sigma_0, c_0\}$. ◊

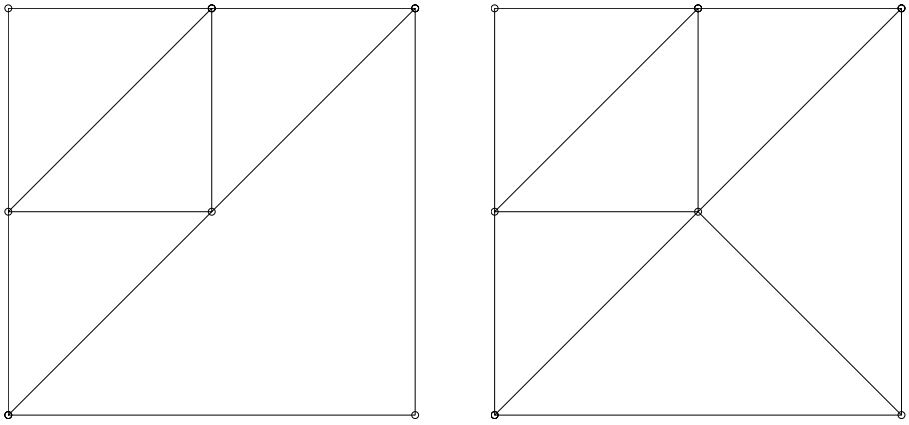


Abb. 91.1: Nicht reguläre (links) und reguläre Triangulierung (rechts)

91 Finite Elemente

Zur praktischen Realisierung des Galerkin-Verfahrens bedarf es der Auswahl geeigneter endlichdimensionaler Ansatzräume $V_h \subset H_0^1(\Omega)$ beziehungsweise $V_h \subset H^1(\Omega)$. Im $\mathbb{R}^1$ können etwa lineare Splines verwendet werden, vgl. Beispiel 90.3. Wir stellen nun entsprechende Ansatzräume auf zweidimensionalen polygonalen Gebieten vor. Dazu wird das Gebiet in Dreiecke unterteilt, deren Ecken wieder als *Knoten* bezeichnet werden. Eine solche *Triangulierung* des Gebiets darf allerdings nicht völlig willkürlich vorgenommen werden.

Definition 91.1. Ein Mengensystem $\mathcal{T} = \{T_1, \ldots, T_m\}$ heißt *reguläre Triangulierung* von Ω, falls alle Elemente $T_i \in \mathcal{T}$ offene Dreiecke sind mit $T_i \cap T_j = \emptyset$ für $i \neq j$ und $\bigcup_{i=1}^m \overline{T_i} = \overline{\Omega}$ und falls darüber hinaus $\overline{T_i} \cap \overline{T_j}$ für $i \neq j$ entweder

(i) leer ist oder
(ii) eine gemeinsame Ecke von T_i und T_j ist oder
(iii) eine gemeinsame Kante von T_i und T_j ist.

Diese Definition wird durch Abbildung 91.1 illustriert. Das linke Bild zeigt eine Triangulierung, die nicht regulär ist. Durch Unterteilung des großen Dreiecks ergibt sich die reguläre Triangulierung aus dem rechten Bild.

In Analogie zum $\mathbb{R}^1$ führen wir nun über einer regulären Triangulierung $\mathcal{T}$ einen Raum stückweise linearer Funktionen ein.

Definition und Satz 91.2. *Sei $\mathcal{T}$ eine reguläre Triangulierung des polygonalen Gebiets Ω mit den Knoten x_i, $i = 1, \ldots, n$. Dann existieren stetige Funktionen $\Lambda_i : \Omega \to \mathbb{R}$, $i = 1, \ldots, n$, mit den folgenden Eigenschaften: In*

jedem Dreieck $T_k \in \mathcal{T}$ stimmt Λ_i mit einem Polynom 1. Grades überein, d. h.

$$\Lambda_i(x) = \beta_{ik} + \alpha_{ik} \cdot x \qquad \text{für gewisse } \beta_{ik} \in \mathbb{R},\ \alpha_{ik} \in \mathbb{R}^2 \text{ und } x \in T_k\,,$$

und es gilt

$$\Lambda_i(x_j) = \delta_{ij}\,, \qquad i,j = 1,\dots,n\,. \tag{91.1}$$

Die lineare Hülle $V^{\mathcal{T}} = \operatorname{span}\{\Lambda_1,\dots,\Lambda_n\}$ dieser Funktionen ist der Raum der stetigen, stückweise linearen Funktionen *bezüglich der Triangulierung $\mathcal{T}$. Wie bei linearen Splines wird $\{\Lambda_1,\dots,\Lambda_n\}$ die* nodale Basis *der* Hutfunktionen *genannt. Das Tupel $\big(\mathcal{T}, V^{\mathcal{T}}\big)$ nennt man* finite Elemente*. Der Raum $V^{\mathcal{T}}$ ist ein Unterraum von $H^1(\Omega)$: Die* schwache Ableitung bzw. Gradient *einer Funktion $\phi \in V^{\mathcal{T}}$ ist stückweise konstant und stimmt im Innern eines Dreiecks $T \in \mathcal{T}$ mit dem klassischen Gradienten von ϕ überein. Der Teilraum $V^{\mathcal{T}} \cap H_0^1(\Omega)$ wird im folgenden mit $V_0^{\mathcal{T}}$ bezeichnet.*

Man beachte bei dieser Definition, daß die Anzahl n der Knoten nicht unmittelbar an die Anzahl m der Dreiecke gekoppelt ist, sondern auch von der Anordnung der Dreiecke abhängt.

Wir verzichten auf den Beweis von Satz 91.2, betonen aber, daß für die Existenz der Basisfunktionen Λ_i die Voraussetzung einer *regulären* Triangulierung wesentlich ist. Man macht sich beispielsweise unmittelbar klar, daß für den Knoten in der Mitte der nicht regulären Triangulierung aus Abbildung 91.1 (links) *keine* entsprechende Hutfunktion existiert.

Aus der Lagrange-Eigenschaft (91.1) folgt wie bei eindimensionalen linearen Splines der folgende Interpolationssatz.

Satz 91.3. *Sei $\mathcal{T}$ eine reguläre Triangulierung von $\Omega \subset \mathbb{R}^2$ mit Knoten x_i, $i = 1,\dots,n$. Dann gibt es genau eine Funktion $\psi \in V^{\mathcal{T}}$, die zu vorgegebenen Werten y_i, $i = 1,\dots,n$, die Interpolationsaufgabe*

$$\psi(x_i) = y_i\,, \qquad i = 1,\dots,n\,,$$

löst, und zwar $\psi = \sum_{i=1}^n y_i \Lambda_i$.

Die Approximationseigenschaften finiter Elemente sind ebenfalls analog zu denen linearer Splines, vgl. Satz 45.4. Um dies zu beweisen, führen wir zunächst noch den Sobolevraum $H^2(\Omega)$ für ein Gebiet $\Omega \subset \mathbb{R}^d$ ein.

Definition 91.4. Sei $\Omega \subset \mathbb{R}^d$ mit $d \geq 1$. Unter dem Unterraum $H^2(\Omega) \subset H^1(\Omega)$ versteht man die Menge aller Funktionen $f \in H^1(\Omega)$, deren zweite partielle Ableitungen allesamt quadratisch integrierbar sind. In diesem Unterraum wird durch

$$\|f\|^2_{H^2(\Omega)} \;=\; \|f\|^2_{H^1(\Omega)} + \int_\Omega \|f''(x)\|^2_F \, dx$$

eine Norm definiert; dabei bezeichnet $f'' = \left[\frac{\partial^2 f}{\partial x_i \partial x_j} \right]_{ij} \in \mathbb{R}^{d\times d}$ die Hesse-Matrix von f.

Ist $\Omega = (a,b) \subset \mathbb{R}^1$ ein eindimensionales Intervall, dann sind die Definitionen 91.4 und 45.3 äquivalent. In diesem Fall ist

$$\|f\|^2_{H^2(\Omega)} \;=\; \|f\|^2_{\mathcal{L}^2(a,b)} + \|f'\|^2_{\mathcal{L}^2(a,b)} + \|f''\|^2_{\mathcal{L}^2(a,b)} \,.$$

Für $\Omega \subset \mathbb{R}^2$ ergibt sich

$$\begin{aligned} \|f\|^2_{H^2(\Omega)} \;&=\; \|f\|^2_{\mathcal{L}^2(\Omega)} + \|f_\xi\|^2_{\mathcal{L}^2(\Omega)} + \|f_\eta\|^2_{\mathcal{L}^2(\Omega)} + \\ &\qquad \|f_{\xi\xi}\|^2_{\mathcal{L}^2(\Omega)} + \|f_{\xi\eta}\|^2_{\mathcal{L}^2(\Omega)} + \|f_{\eta\xi}\|^2_{\mathcal{L}^2(\Omega)} + \|f_{\eta\eta}\|^2_{\mathcal{L}^2(\Omega)} \,. \end{aligned}$$

Für eine Funktion $f \in H^2(\Omega)$ gelten stärkere Spursätze als für Funktionen aus $H^1(\Omega)$. Wir benötigen das folgende Resultat, dessen Beweis etwa in dem Buch von Braess [10, Abschnitt II.3] nachgelesen werden kann.

Satz 91.5. *In einem beschränkten polygonalen Gebiet $\Omega \subset \mathbb{R}^2$ ist jede Funktion $f \in H^2(\Omega)$ stetig in $\overline{\Omega}$. Insbesondere ist die Spur $f|_\Gamma$ eine stetige Funktion.*

Aufgrund dieser Stetigkeit können wir jede Funktion $f \in H^2(\Omega)$ durch eine Funktion $\psi \in V^{\mathcal{T}}$ gemäß Satz 91.3 in den Knoten der Triangulierung $\mathcal{T}$ von Ω linear interpolieren, sofern diese Triangulierung regulär ist. Der Interpolationsfehler kann dann in der folgenden Weise abgeschätzt werden.

Satz 91.6. *Sei $\mathcal{T}$ eine reguläre Triangulierung eines polygonalen Gebiets Ω, h die maximale Kantenlänge und $\alpha_0 > 0$ der minimale Innenwinkel aller Dreiecke $T_i \in \mathcal{T}$. Ferner sei $f \in H^2(\Omega)$ und $\psi \in V^{\mathcal{T}}$ der zugehörige stückweise lineare Interpolant. Dann ist*

$$\|f - \psi\|_{\mathcal{L}^2(\Omega)} \;\le\; \sqrt{3/8}\, h^2 \, \|f\|_{H^2(\Omega)} \tag{91.2}$$

und

$$|f - \psi|_{H^1(\Omega)} \;\le\; \frac{3}{\sqrt{8}\sin^2\alpha_0}\, h \, \|f\|_{H^2(\Omega)} \,. \tag{91.3}$$

Beweis. Wir beweisen die beiden Ungleichungen nur für Funktionen $f \in C^2(\overline{\Omega})$. Der allgemeine Fall folgt dann aus der Tatsache, daß $C^2(\overline{\Omega})$ dicht in $H^2(\Omega)$ liegt, vgl. Hackbusch [43].

1. Wir beginnen mit der Ungleichung (91.2) und betrachten hierzu zunächst ein festes Dreieck $T \in \mathcal{T}$, dessen Ecken im folgenden mit x_1, x_2 und x_3 bezeichnet werden. Ist x ein beliebiger Punkt des Dreiecks, so erhalten wir aus dem Satz von Taylor

$$f(x_k) \;=\; f(x) \;+\; \operatorname{grad} f(x) \cdot d_k \;+\; \int_0^1 d_k^T f''(x + t d_k) d_k (1-t)\, dt \tag{91.4}$$

mit $d_k = x_k - x$, $k = 1, 2, 3$. Mit der Darstellung des linearen Interpolanten ψ aus Satz 91.3 folgt hieraus

$$\psi(x) \;=\; \sum_{k=1}^{3} f(x_k)\Lambda_k(x) \;=\; f(x)\sum_{k=1}^{3}\Lambda_k(x) \;+\; \operatorname{grad} f(x)\cdot\sum_{k=1}^{3} d_k\Lambda_k(x) \\ +\;\sum_{k=1}^{3}\Big(\int_0^1 d_k^T f''(x+td_k)d_k(1-t)\,dt\Big)\Lambda_k(x)\,.$$

Offensichtlich gelten für $x \in T$ die Summenformeln

$$\sum_{k=1}^{3}\Lambda_k(x) = 1 \qquad \text{und} \qquad \sum_{k=1}^{3} x_k\Lambda_k(x) = x\,, \tag{91.5}$$

so daß

$$\sum_{k=1}^{3} d_k\Lambda_k(x) \;=\; \sum_{k=1}^{3} x_k\Lambda_k(x) \;-\; \sum_{k=1}^{3} x\Lambda_k(x) \;=\; x - x \;=\; 0\,.$$

In die Gleichung für ψ eingesetzt erhalten wir die Fehlerdarstellung

$$\psi(x) - f(x) \;=\; \sum_{k=1}^{3}\Big(\int_0^1 d_k^T f''(x+td_k)d_k(1-t)\,dt\Big)\Lambda_k(x)\,,$$

und die Cauchy-Schwarz-Ungleichung in $\mathbb{R}^3$ ergibt

$$|\psi(x) - f(x)|^2 \;\le\; \sum_{k=1}^{3}\Big(\int_0^1 d_k^T f''(x+td_k)d_k(1-t)\,dt\Big)^2 \sum_{k=1}^{3}\Lambda_k^2(x) \\ \le\; \sum_{k=1}^{3}\Big(\int_0^1 d_k^T f''(x+td_k)d_k(1-t)\,dt\Big)^2\,,$$

wobei die letzte Abschätzung aus der ersten Summe in (91.5) zusammen mit den Ungleichungen $\Lambda_k^2(x) \le \Lambda_k(x)$, $k = 1, 2, 3$, folgt. Durch Integration über T erhalten wir schließlich eine Darstellung des $\mathcal{L}^2$-Fehlers:

$$\|\psi - f\|^2_{\mathcal{L}^2(T)} \;\le\; \sum_{k=1}^{3} I_k \tag{91.6}$$

mit

$$I_k \;=\; \int_T\Big(\int_0^1 d_k^T f''(x+td_k)d_k(1-t)\,dt\Big)^2 dx\,, \qquad k = 1, 2, 3\,.$$

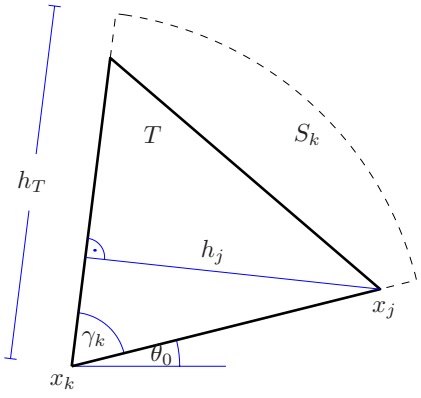


Abb. 91.2:
Skizze zum Beweis von Satz 91.6

2. Wir untersuchen als nächstes das Integral I_k für ein beliebiges, aber festes $k \in \{1,2,3\}$. Hierzu setzen wir den Integranden (bzw. f'') durch Null auf den Kreissektor S_k mit Zentrum in x_k und Öffnungswinkel γ_k des Dreiecks wie in Abbildung 91.2 fort; als Radius h_T des Sektors wählen wir die maximale Seitenlänge des Dreiecks. In S_k führen wir anschließend Polarkoordinaten ein,

$$x = x_k + r x_\theta \qquad \text{mit} \qquad x_\theta = [\cos\theta, \sin\theta]^T\,, \qquad \begin{array}{l} \theta_0 \le \theta \le \theta_0 + \gamma_k\,, \\ 0 \le r \le h_T\,. \end{array}$$

In diesen Koordinaten ist $d_k = x_k - x = -r x_\theta$ und $x + t d_k = x_k + (1-t) r x_\theta$, und wir erhalten

$$I_k \;=\; \int_{\theta_0}^{\theta_0+\gamma_k} \int_0^{h_T} \Big(\int_0^1 r^2 x_\theta^T f''\big(x_k + (1-t) r x_\theta\big) x_\theta (1-t)\, dt \Big)^2 r\, dr d\theta\,.$$

Eine neuerliche Substitution $s = (1-t)r$ ergibt dann zusammen mit der Cauchy-Schwarz-Ungleichung die Abschätzung

$$\begin{aligned} I_k \;&=\; \int_{\theta_0}^{\theta_0+\gamma_k} \int_0^{h_T} \Big(\int_0^r x_\theta^T f''(x_k + s x_\theta) x_\theta s\, ds \Big)^2 r\, dr\, d\theta \\ &\le \int_{\theta_0}^{\theta_0+\gamma_k} \int_0^{h_T} \Big(\int_0^r \big|\, x_\theta^T f''(x_k + s x_\theta) x_\theta \,\big|^2 s\, ds \Big) \Big(\int_0^r s\, ds \Big) r\, dr\, d\theta \\ &= \frac{1}{2} \int_{\theta_0}^{\theta_0+\gamma_k} \int_0^{h_T} \Big(\int_0^r \big|\, x_\theta^T f''(x_k + s x_\theta) x_\theta \,\big|^2 s\, ds \Big) r^3\, dr\, d\theta\,. \end{aligned}$$

Nach Abschnitt 2 ist

$$\big|\, x_\theta^T f''(x_k + s x_\theta) x_\theta \,\big| \;\le\; |x_\theta| \,\big|\, f''(x_k + s x_\theta) x_\theta \,\big| \;\le\; \|f''(x_k + s x_\theta)\|_F$$

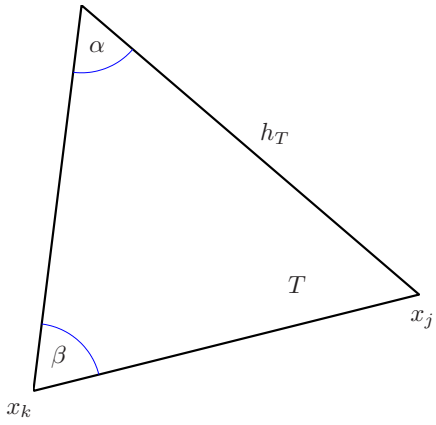


Abb. 91.3:
Skizze zum Sinussatz

und somit haben wir die Ungleichung

$$I_k \;\le\; \frac{1}{2}\int_{\theta_0}^{\theta_0+\gamma_k}\int_0^{h_T}\Big(\int_0^r \|f''(x_k+sx_\theta)\|_F^2\, s\,ds\;\Big)\, r^3\,dr\,d\theta\,.$$

Hieraus folgt schließlich

$$\begin{aligned} I_k &\le \frac{1}{2}\int_{\theta_0}^{\theta_0+\gamma_k}\Big(\int_0^{h_T}\|f''(x_k+sx_\theta)\|_F^2\, s\,ds\Big)\Big(\int_0^{h_T} r^3 dr\Big)\,d\theta \\ &= \frac{1}{8}\,h_T^4\int_{\theta_0}^{\theta_0+\gamma_k}\int_0^{h_T}\|f''(x_k+sx_\theta)\|_F^2\, s\,ds\,d\theta \\ &\le \frac{1}{8}\,h_T^4\,\|f\|_{H^2(T)}^2\,. \end{aligned} \tag{91.7}$$

Aus (91.6) und (91.7) ergibt sich wegen $h_T \le h$ dann die erste Behauptung (91.2) durch Summation über alle Dreiecke $T \in \mathcal{T}$.

3. Die H^1-Abschätzung kann ebenfalls auf die Ungleichung (91.7) zurückgeführt werden. Zuvor benötigen wir jedoch noch eine Abschätzung der Gradienten der Hutfunktionen Λ_j, $j = 1, 2, 3$. Die Hutfunktion Λ_j verschwindet auf der dem Knoten x_j gegenüberliegenden Seite von T; der (in T konstante) Gradient von Λ_j steht somit senkrecht auf dieser Seite und $|\operatorname{grad}\Lambda_j|$ ist der Kehrwert der in Abbildung 91.2 eingezeichneten Höhe h_j. Mit den dortigen Bezeichnungen errechnet sich diese Höhe aus

$$h_j \;=\; |x_j - x_k|\sin\gamma_k\,. \tag{91.8}$$

Zur Abschätzung von $|x_j - x_k|$ verwenden wir nun den Sinussatz, den wir auf die Kante zwischen x_j und x_k sowie die längste Seite des Dreiecks (mit der Länge h_T) anwenden, vgl. Abbildung 91.3. Mit den dortigen Bezeichnungen

ist

$$|x_j - x_k| \;=\; h_T \,\frac{\sin\alpha}{\sin\beta} \;\geq\; h_T \sin\alpha \;\geq\; h_T \sin\alpha_0\,, \tag{91.9}$$

wobei wir die Voraussetzung ausgenutzt haben, daß alle Innenwinkel von $\mathcal{T}$ durch ein $\alpha_0 > 0$ nach unten beschränkt sind. Zusammen mit (91.8) ergibt dies $h_j \geq h_T \sin^2\alpha_0$ beziehungsweise

$$|\operatorname{grad}\Lambda_j| \;=\; \frac{1}{h_j} \;\leq\; \frac{1}{h_T \sin^2\alpha_0}\,, \qquad j = 1,2,3\,. \tag{91.10}$$

4. Nun kommen wir zum Beweis von (91.3). Wir multiplizieren (91.4) mit $\operatorname{grad}\Lambda_k(x)$ und summieren von $k = 1$ bis 3. Dabei ist zu berücksichtigen, daß für Vektoren $x, y_k, z_k \in \mathbb{R}^2$, $k = 1,\ldots,p$, das Distributivgesetz

$$\sum_{k=1}^{p}(x\cdot y_k)z_k \;=\; \sum_{k=1}^{p} z_k(y_k^T x) \;=\; \sum_{k=1}^{p}(z_k y_k^T)x \;=\; \Big(\sum_{k=1}^{p} z_k y_k^T\Big)x$$

gilt, wobei $\sum_k z_k y_k^T \in \mathbb{R}^{2\times 2}$. Daher erhalten wir bei der genannten Summation

$$\begin{aligned}\operatorname{grad}\psi(x) \;&=\; \sum_{k=1}^{3} f(x_k)\operatorname{grad}\Lambda_k(x)\\ &=\; f(x)\sum_{k=1}^{3}\operatorname{grad}\Lambda_k(x) \;+\; \Big(\sum_{k=1}^{3}\operatorname{grad}\Lambda_k(x)\,d_k^T\Big)\operatorname{grad} f(x)\\ &\qquad +\; \sum_{k=1}^{3}\Big(\int_0^1 d_k^T f''(x+td_k)d_k(1-t)\,dt\Big)\operatorname{grad}\Lambda_k(x)\,.\end{aligned}$$

Aufgrund von (91.5) ist $\sum_k \operatorname{grad}\Lambda_k(x) = 0$ und $\sum_k x_k(\operatorname{grad}\Lambda_k(x))^T$ die zweidimensionale Einheitsmatrix, so daß sich die obige Darstellung zu

$$\operatorname{grad}\psi(x) \;=\; \operatorname{grad} f(x) \;+\; \sum_{k=1}^{3}\Big(\int_0^1 d_k^T f''(x+td_k)d_k(1-t)\,dt\Big)\operatorname{grad}\Lambda_k(x)$$

vereinfacht. Mit der Cauchy-Schwarz-Ungleichung im $\mathbb{R}^3$ und mit (91.10) ergibt sich somit

$$|\operatorname{grad}(\psi - f)(x)|^2 \;\leq\; \frac{3}{\sin^4\alpha_0}\,\frac{1}{h_T^2}\sum_{k=1}^{3}\Big(\int_0^1 d_k^T f''(x+td_k)d_k(1-t)\,dt\Big)^2$$

und Integration über T liefert

$$|\psi - f|^2_{H^1(T)} \le \frac{3}{\sin^4 \alpha_0} \frac{1}{h_T^2} \sum_{k=1}^{3} I_k$$

mit demselben Integral I_k wie in (91.6). Aus (91.7) folgt somit

$$|\psi - f|^2_{H^1(T)} \le \frac{9}{8 \sin^4 \alpha_0} h_T^2 \, \|f\|^2_{H^2(T)} \, .$$

Eine letzte Summation über alle Dreiecke T von $\mathcal{T}$ ergibt wegen $h_T \le h$ somit die zweite Behauptung (91.3) des Satzes für eine Funktion $f \in C^2(\overline{\Omega})$. □

92 Fehlerschranken für die Finite-Elemente-Methode

Werden finite Elemente als Ansatzfunktionen für das Galerkin-Verfahren verwendet, so spricht man von der *Methode der finiten Elemente*. Wir wollen nun den Fehler dieses Verfahrens genauer quantifizieren und betrachten dazu wieder ein elliptisches Dirichlet-Problem

$$-\operatorname{div}(\sigma \operatorname{grad} u) + cu = f \quad \text{in } \Omega \, , \qquad u|_\Gamma = 0 \, , \tag{92.1}$$

mit $0 < \sigma_0 \le \sigma(x) \le \sigma_\infty$ und $0 \le c(x) \le c_\infty$. Wir bezeichnen mit u die (schwache) Lösung von (92.1) und mit u_h die Galerkin-Näherung aus dem Raum $V_h = V_0^{\mathcal{T}}$. Es wird im weiteren durchweg vorausgesetzt, daß Ω ein polygonales Gebiet und $\mathcal{T}$ eine reguläre Triangulierung mit maximaler Kantenlänge h und minimalem Innenwinkel $\alpha_0 > 0$ ist. Als erstes beweisen wir die folgende Fehlerabschätzung für $u - u_h$ bezüglich der H^1-Norm.

Satz 92.1. *Die schwache Lösung u von (92.1) gehöre zu $H^2(\Omega) \cap H_0^1(\Omega)$. Dann gilt*

$$\|u - u_h\|_{H^1(\Omega)} \le c_1 h \|u\|_{H^2(\Omega)}$$

mit einer geeigneten Konstanten $c_1 > 0$.

Beweis. Nach dem Lemma 90.5 von Céa ist

$$\begin{aligned} \|u - u_h\|_{H^1(\Omega)} &\le \frac{\max\{\sigma_\infty, c_\infty\}}{\gamma_\Omega^2 \sigma_0} \inf_{v \in V_0^{\mathcal{T}}} \|u - v\|_{H^1(\Omega)} \\ &\le \frac{\max\{\sigma_\infty, c_\infty\}}{\gamma_\Omega^2 \sigma_0} \|u - \psi\|_{H^1(\Omega)} \, , \end{aligned}$$

wobei $\psi \in V_0^{\mathcal{T}}$ die stückweise lineare Funktion bezeichnet, die u in den Knoten von $\mathcal{T}$ interpoliert. Nach Satz 91.6 existiert ein $\kappa_1 > 0$ mit

$$\|u - \psi\|_{H^1(\Omega)} \leq \kappa_1 h \|u\|_{H^2(\Omega)} ,$$

und somit folgt unmittelbar die Behauptung. □

Für Fehlerabschätzungen bezüglich der $\mathcal{L}^2$-Norm spielen sogenannte *Regularitätssätze* eine entscheidende Rolle, die garantieren, daß die schwache Lösung des Randwertproblems (92.1) für jedes $f \in \mathcal{L}^2(\Omega)$ in $H^2(\Omega)$ liegt. Stellvertretend sei hier der folgende Satz angeführt, dessen Beweis sich in dem Buch von Grisvard [39, Theorem 3.2.1.2] findet (vgl. Aufgabe 6 für einen Beweis im $\mathbb{R}^1$).

Satz 92.2. *Sei Ω ein konvexes polygonales Gebiet und zusätzlich zu den Voraussetzungen von Satz* 89.5 *gehöre σ zu $C^1(\overline{\Omega})$. Dann hat die Differentialgleichung* (92.1) *eine schwache Lösung u in $H^2(\Omega) \cap H_0^1(\Omega)$ und es gilt*

$$\|u\|_{H^2(\Omega)} \leq c_2 \|f\|_{\mathcal{L}^2(\Omega)} , \tag{92.2}$$

wobei c_2 nur vom Gebiet Ω abhängt.

Unter dieser zusätzlichen Glattheitsannahme an die Koeffizienten des Differentialoperators kann nun die folgende Fehlerabschätzung bewiesen werden.

Satz 92.3. *Unter den Voraussetzungen des Regularitätssatzes* 92.2 *gilt*

$$\|u - u_h\|_{\mathcal{L}^2(\Omega)} \leq c_0 \, h^2 \, \|u\|_{H^2(\Omega)}$$

mit $c_0 = \max\{\sigma_\infty, c_\infty\} c_1^2 c_2$ und c_1, c_2 wie in den Sätzen 92.1 *und* 92.2.

Beweis. Sei w die (schwache) Lösung des Dirichlet-Problems

$$-\operatorname{div}(\sigma \operatorname{grad} w) + cw = u - u_h \quad \text{in } \Omega , \qquad w|_\Gamma = 0 .$$

Nach Satz 92.2 gehört w zu $H^2(\Omega) \cap H_0^1(\Omega)$ und löst das Variationsproblem

$$a(w, v) = \int_\Omega (u - u_h) v \, dx \qquad \text{für alle } v \in H_0^1(\Omega) . \tag{92.3}$$

Aufgrund der Variationsformulierung (90.3) des Randwertproblems für u und der Galerkin-Bedingung (90.4) für u_h gilt entsprechend

$$a(u, v) = \int_\Omega f v \, dx , \quad a(u_h, v) = \int_\Omega f v \, dx \qquad \text{für alle } v \in V_0^{\mathcal{T}} .$$

Insbesondere ist dies für die Galerkin-Approximation $v = w_h$ an w aus $V_0^{\mathcal{T}}$ richtig, d. h.

$$0 = a(u - u_h, w_h) = a(w_h, u - u_h) .$$

Zusammen mit (92.3) folgt hieraus

$$\begin{aligned}\|u-u_h\|^2_{\mathcal{L}^2(\Omega)} &= \int_\Omega (u-u_h)^2\,dx = a(w,u-u_h) = a(w-w_h,u-u_h)\\ &\le a_\infty \|w-w_h\|_{H^1(\Omega)}\|u-u_h\|_{H^1(\Omega)}\end{aligned}$$

mit $a_\infty = \max\{\sigma_\infty, c_\infty\}$, vgl. Proposition 90.2. Aus Satz 92.1 (angewandt auf w und auf u) und der Ungleichung (92.2) aus Satz 92.2 ergibt sich dann

$$\begin{aligned}\|u-u_h\|^2_{\mathcal{L}^2(\Omega)} &\le a_\infty c_1 h \|w\|_{H^2(\Omega)} c_1 h \|u\|_{H^2(\Omega)}\\ &\le a_\infty c_1^2 c_2 h^2 \|u-u_h\|_{\mathcal{L}^2(\Omega)}\|u\|_{H^2(\Omega)},\end{aligned}$$

und nach Division durch $\|u-u_h\|_{\mathcal{L}^2(\Omega)}$ erhalten wir die Behauptung. □

Man beachte, daß dieses Resultat mit der Fehlerabschätzung aus Satz 84.7 für Randwertaufgaben im $\mathbb{R}^1$ vergleichbar ist. Dort war aber die wesentlich stärkere Voraussetzung $u \in C^4[0,1]$ an die Lösung gestellt worden.

Abschließend sei erwähnt, daß bessere Fehlerschranken möglich sind, falls die Lösung des Randwertproblems sehr glatt ist und das Galerkin-Verfahren mit glatteren Ansatzfunktionen implementiert wird. Entsprechende Ansatzräume finden sich in den eingangs genannten Büchern.

93 Die Steifigkeitsmatrix

Bei der Implementierung der Finite-Elemente-Methode steht die Steifigkeitsmatrix A aus (90.5) und die anschließende Lösung des Gleichungssystems $A\boldsymbol{u}_h = b$ im Mittelpunkt. In folgenden wenden wir uns zunächst der Berechnung der Steifigkeitsmatrix zu, Abschnitt 93.5 schließt mit einigen Bemerkungen zur Cholesky-Faktorisierung von A. Wir beschränken uns durchweg auf den Fall, daß die Funktionen ϕ_i die nodale Basis eines Finite-Elemente-Ansatzraums bezeichnen.

93.1 Die Assemblierung

Für die Einträge von A müssen gemäß (90.5) die Integrale

$$a(\phi_i,\phi_j) = \int_\Omega \sigma \operatorname{grad}\phi_i \cdot \operatorname{grad}\phi_j\,dx + \int_\Omega c\phi_i\phi_j\,dx$$

ausgewertet werden. Die meisten Einträge sind allerdings Null, da die Träger der Hutfunktionen nur aus wenigen benachbarten Dreiecken bestehen. Die Matrix A ist also dünn besetzt.

Auf den ersten Blick erscheint es naheliegend, die Integrale nacheinander für alle $i, j = 1, \ldots, n$ auszuwerten. Dies erweist sich jedoch nicht als allzu geschickt, da bei dieser Vorgehensweise oft nicht unmittelbar klar ist, für welche Paare (i, j) der Integralwert ungleich Null ist. Besser ist es, die Integrale als Summe der Teilintegrale über die einzelnen Dreiecke von $\mathcal{T}$ zu berechnen. Dies ist einfach zu implementieren, wenn man in einer Liste `x` die Ortskoordinaten (ξ, η) der einzelnen Knoten und in einer zweiten Liste `T` für jedes Dreieck die Indizes i_1, i_2, i_3 seiner Eckknoten abspeichert. Für die Rechnung erweist es sich als vorteilhaft, in `T` überdies den jeweiligen Wert von $d = \det \Phi'$ aus (93.4) mit abzuspeichern.

Für jedes Dreieck $T_k \in \mathcal{T}$ ist

$$S_k \;=\; \Big[\int_{T_k} \sigma \operatorname{grad} \phi_i \cdot \operatorname{grad} \phi_j \, dx \;+\; \int_{T_k} c \phi_i \phi_j \, dx \Big]_{ij} \;\in\; \mathbb{R}^{n \times n}$$

eine Matrix, die aus allen Teilintegralen über dieses eine Dreieck besteht. Wegen

$$\begin{aligned} a(\phi_i, \phi_j) \;&=\; \int_\Omega \sigma \operatorname{grad} \phi_i \cdot \operatorname{grad} \phi_j \, dx \;+\; \int_\Omega c \phi_i \phi_j \, dx \\ &=\; \sum_{k=1}^{m} \Big(\int_{T_k} \sigma \operatorname{grad} \phi_i \cdot \operatorname{grad} \phi_j \, dx \;+\; \int_{T_k} c \phi_i \phi_j \, dx \Big) \end{aligned}$$

ergibt dies

$$A = \sum_{k=1}^{m} S_k \,. \tag{93.1}$$

Die Matrizen S_k heißen *Elementsteifigkeitsmatrizen*. Bei der nodalen Basis besitzt jede Elementsteifigkeitsmatrix S_k höchstens 3×3 von Null verschiedene Einträge, da nur die Integrale mit Hutfunktionen Λ_i und Λ_j zu Eckknoten x_i und x_j von T_k von Null verschieden sein können. In der Praxis werden lediglich diese von Null verschiedenen Einträge von S_k abgespeichert. Bei der sogenannten *Assemblierung* (93.1) muß dann lediglich die Numerierung der Knoten beachtet werden.

93.2 Übergang zum Referenzdreieck

Zur Berechnung einer Elementsteifigkeitsmatrix empfiehlt sich eine Transformation des jeweiligen Dreiecks T mit den Ecken $x_i = [\xi_i, \eta_i]^T$, $i = 1, 2, 3$, auf das sogenannte *Referenzdreieck*

$$D \;=\; \{ z = [s, t]^T \,:\, s > 0,\, t > 0,\, s + t < 1 \} \,. \tag{93.2}$$

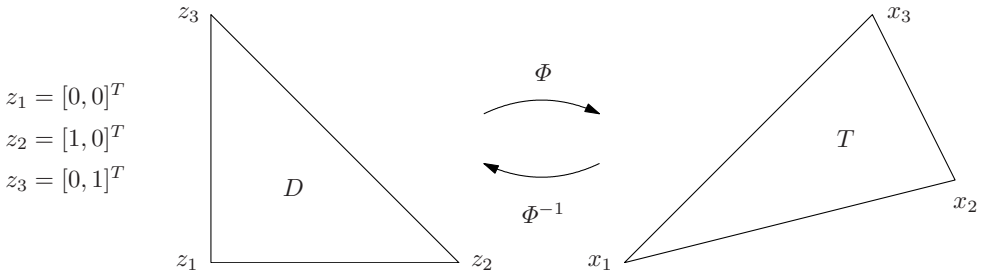


Abb. 93.1: Übergang zum Referenzdreieck

Dieser Übergang erfolgt mittels der affinen Abbildung

$$\Phi(s,t) \;=\; x_1 + s(x_2 - x_1) + t(x_3 - x_1)\,, \tag{93.3}$$

die das Referenzdreieck D bijektiv auf T abbildet, vgl. Abbildung 93.1. Φ' ist dabei von z unabhängig,

$$\Phi' \;=\; \begin{bmatrix} x_2 - x_1 & x_3 - x_1 \end{bmatrix} \;=\; \begin{bmatrix} \xi_2 - \xi_1 & \xi_3 - \xi_1 \\ \eta_2 - \eta_1 & \eta_3 - \eta_1 \end{bmatrix},$$

und der Betrag der Determinante

$$d \;=\; \det \Phi' \;=\; (\xi_2 - \xi_1)(\eta_3 - \eta_1) \;-\; (\xi_3 - \xi_1)(\eta_2 - \eta_1) \tag{93.4}$$

ist gerade das Doppelte des Flächeninhalts von T; folglich ist $d \neq 0$ und

$$\Phi'^{-1} \;=\; \frac{1}{d} \begin{bmatrix} \eta_3 - \eta_1 & \xi_1 - \xi_3 \\ \eta_1 - \eta_2 & \xi_2 - \xi_1 \end{bmatrix}.$$

Beispiel 93.1. Wir berechnen die Elementsteifigkeitsmatrix $S = [s_{ij}]$ für den Fall des Laplace-Operators $L[u] = -\Delta u$ und ein Dreieck T mit den drei zugehörigen Hutfunktionen $\Lambda_1, \Lambda_2, \Lambda_3$ über den Ecken x_1, x_2 und x_3 von T. In diesem Beispiel ergibt sich für $i, j = 1, 2, 3$ mit der Substitutions- und Kettenregel

$$\begin{aligned} s_{ij} \;&=\; \int_T \operatorname{grad}_x \Lambda_i(x) \cdot \operatorname{grad}_x \Lambda_j(x)\, dx \\ &=\; \int_D \operatorname{grad}_x \big(\Lambda_i(\Phi(z))\big) \cdot \operatorname{grad}_x \big(\Lambda_j(\Phi(z))\big)\, |\det \Phi'|\, dz \\ &=\; |d| \int_D \Big(\Phi'^{-T} \operatorname{grad}_z \big(\Lambda_i(\Phi(z))\big)\Big) \cdot \Big(\Phi'^{-T} \operatorname{grad}_z \big(\Lambda_j(\Phi(z))\big)\Big)\, dz\,. \end{aligned}$$

Hierbei ist $\Lambda_i \circ \Phi$ die Hutfunktion über D zur Ecke z_i und daher ist

$$G := \begin{bmatrix} \operatorname{grad}_z(\Lambda_1 \circ \Phi)^T \\ \operatorname{grad}_z(\Lambda_2 \circ \Phi)^T \\ \operatorname{grad}_z(\Lambda_3 \circ \Phi)^T \end{bmatrix} = \begin{bmatrix} -1 & -1 \\ 1 & 0 \\ 0 & 1 \end{bmatrix}.$$

Die Integranden von s_{ij} sind also über D konstant. Da der Flächeninhalt von D gerade $1/2$ ist, erhält man schließlich den linken oberen 3×3-Block

$$\frac{|d|}{2} G \Phi'^{-1} \Phi'^{-T} G^T = \frac{1}{2|d|} \begin{bmatrix} \eta_2 - \eta_3 & \xi_3 - \xi_2 \\ \eta_3 - \eta_1 & \xi_1 - \xi_3 \\ \eta_1 - \eta_2 & \xi_2 - \xi_1 \end{bmatrix} \begin{bmatrix} \eta_2 - \eta_3 & \eta_3 - \eta_1 & \eta_1 - \eta_2 \\ \xi_3 - \xi_2 & \xi_1 - \xi_3 & \xi_2 - \xi_1 \end{bmatrix} \quad (93.5)$$

von S. ◊

93.3 Quadraturformeln

Im vorangegangenen Beispiel 93.1 konnten die Integrale noch explizit ausgerechnet werden. Für $c \neq 0$ ist dies bereits schwieriger, da in diesem Fall zur Berechnung der Elementsteifigkeitsmatrizen auch Integrale der Form

$$\int_T c \Lambda_i \Lambda_j \, dx$$

auszuwerten sind. Falls c konstant ist, ist der Integrand ein Polynom zweiten Grades, also eine Funktion der Form

$$p(\xi, \eta) = \alpha_{00} + \alpha_{10}\xi + \alpha_{01}\eta + \alpha_{20}\xi^2 + \alpha_{11}\xi\eta + \alpha_{02}\eta^2 . \quad (93.6)$$

Noch komplizierter wird es, wenn σ und c nicht konstant oder stückweise konstant sind. Dann müssen zur Berechnung der Elementsteifigkeitsmatrizen in der Regel Quadraturformeln verwendet werden.

Wie in Kapitel VII kann man auch in höheren Dimensionen Quadraturformeln nach der Vorgabe konstruieren, daß Polynome niederen Grades exakt integriert werden. Entsprechend wird wie im Eindimensionalen der *Exaktheitsgrad* einer Quadraturformel definiert. Um Formeln mit vorgegebenem Exaktheitsgrad bestimmen zu können, ist das folgende Resultat hilfreich.

Lemma 93.2. *Ist D das Referenzdreieck aus* (93.2)*, dann gilt*

$$\int_D s^j t^k \, dz = \frac{j!\, k!}{(j+k+2)!} . \quad (93.7)$$

Dabei ist $0! = 1$.

Beweis. Nach dem Satz von Fubini ist

$$\int_D s^j t^k \, dz \;=\; \int_0^1 s^j \int_0^{1-s} t^k \, dt \, ds \;=\; \frac{1}{k+1} \int_0^1 s^j (1-s)^{k+1} \, ds \,.$$

Für $j = 0$ hat das Integral auf der rechten Seite den Wert $1/(k+2)$. Für $j \geq 1$ ergibt sich durch partielle Integration

$$\begin{aligned} \int_D s^j t^k \, dz \;&=\; \frac{1}{(k+1)(k+2)} \Big(-s^j (1-s)^{k+2} \Big|_0^1 \;+\; j \int_0^1 s^{j-1} (1-s)^{k+2} \, ds \Big) \\ &=\; \frac{j}{(k+1)(k+2)} \int_0^1 s^{j-1} (1-s)^{k+2} \, ds \,, \end{aligned}$$

und durch Induktion folgt schließlich

$$\begin{aligned} \int_D s^j t^k \, dz \;&=\; \frac{j!}{(k+1)\cdots(k+j+1)} \int_0^1 (1-s)^{k+j+1} \, ds \\ &=\; -\,\frac{j!\,k!}{(k+j+2)!} (1-s)^{k+j+2} \Big|_0^1 \;=\; \frac{j!\,k!}{(k+j+2)!} \,. \end{aligned}$$

□

Die folgende einfache Quadraturformel hat den Exaktheitsgrad $q = 2$.

Proposition 93.3. *Sei T ein Dreieck mit Flächeninhalt $|d|/2$ und Seitenmittelpunkten y_1, y_2 und y_3. Dann ist*

$$\int_T p(x) \, dx \;=\; \frac{|d|}{6} \big(p(y_1) + p(y_2) + p(y_3) \big) \tag{93.8}$$

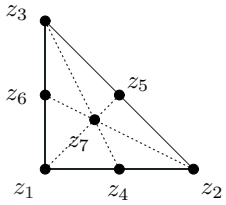
für jedes quadratische Polynom p aus (93.6).

Beweis. Sei zunächst T das Referenzdreieck D. Dann überprüfen wir die Exaktheit der Quadraturformel anhand von Lemma 93.2 über die folgende Tabelle:

Basismonom	1	s	t	s^2	st	t^2
(93.7)	1/2	1/6	1/6	1/12	1/24	1/12
(93.8)	1/2	1/6	1/6	1/12	1/24	1/12

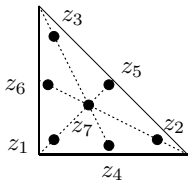
Auf D ist diese Quadraturformel also für alle Monome und damit alle Polynome vom Grad $q \leq 2$ exakt. Ist T ein beliebiges Dreieck und p ein Polynom vom Grad kleiner oder gleich Zwei, dann gilt

$$\int_T p(x) \, dx \;=\; |d| \int_D p(\Phi(z)) \, dz \,,$$



$z_7 = [1/3, 1/3]^T$

$\omega_1 = 1/40$
$\omega_2 = 1/40$
$\omega_3 = 1/40$
$\omega_4 = 1/15$
$\omega_5 = 1/15$
$\omega_6 = 1/15$
$\omega_7 = 27/120$



i	s	t	ω
1	$(6-\sqrt{15})/21$	$(6-\sqrt{15})/21$	$(155-\sqrt{15})/2400$
2	$(9+2\sqrt{15})/21$	$6-\sqrt{15})/21$	$(155-\sqrt{15})/2400$
3	$(6-\sqrt{15})/21$	$(9+2\sqrt{15})/21$	$(155-\sqrt{15})/2400$
4	$(6+\sqrt{15})/21$	$(9-2\sqrt{15})/21$	$(155+\sqrt{15})/2400$
5	$(6+\sqrt{15})/21$	$(6+\sqrt{15})/21$	$(155+\sqrt{15})/2400$
6	$(9-2\sqrt{15})/21$	$(6+\sqrt{15})/21$	$(155+\sqrt{15})/2400$
7	1/3	1/3	9/80

Abb. 93.2: Quadraturformeln mit Exaktheitsgrad $q = 3$ und $q = 5$

wobei $p \circ \Phi$ wiederum ein Polynom vom Grad kleiner oder gleich Zwei ist (Φ ist weiterhin durch (93.3) gegeben und demnach ist $|d| = |\det \Phi'|$). Da die Exaktheit der Quadraturformel (93.8) über D für Polynome vom Grad $q \leq 2$ bereits nachgewiesen ist, folgt

$$\begin{aligned}\int_T p(x)\,dx \;&=\; |d|\,\frac{1}{6}\left(p\circ\Phi(0.5,0) \;+\; p\circ\Phi(0,0.5) \;+\; p\circ\Phi(0.5,0.5)\right)\\ &=\; \frac{|d|}{6}\big(p(y_1)+p(y_2)+p(y_3)\big)\,.\end{aligned}$$

□

Neben dieser einfachen Quadraturformel verwendet man in der Praxis für Integrale über das Referenzdreieck D häufig noch die beiden Formeln

$$\int_D f(z)\,dz \;\approx\; \sum_{i=1}^{7} \omega_i f(z_i)$$

mit den Knoten $z_i = [s_i, t_i]^T$ und Gewichten ω_i aus Abbildung 93.2. Die obere Quadraturformel hat Exaktheitsgrad $q = 3$, die untere sogar Exaktheitsgrad $q = 5$.

Aus dem Exaktheitsgrad einer Quadraturformel ergeben sich bei hinreichend glatten Funktionen ähnliche Fehlerabschätzungen für die Integralnäherungen wie im Eindimensionalen, vgl. Abschnitt 38:

Lemma 93.4. *Sei T ein beliebiges Dreieck mit maximaler Kantenlänge h, D das Referenzdreieck und Φ bezeichne wie zuvor die Transformation* (93.3)*. Ferner sei $Q[f]$ eine Quadraturformel für $\int_D f(z)\,dz$ mit Exaktheitsgrad q. Dann gilt für jede Funktion $g \in C^{q+1}(\overline{T})$ die Quadraturfehlerabschätzung*

$$\left| \, |d|\, Q[g \circ \Phi] \, - \, \int_T g(x)\,dx \, \right| \, = \, O(|d| h^{q+1}) \,,$$

wobei die Fehlerkonstante von der Funktion g abhängt.

Beweis. Mit dem Satz von Taylor kann man g schreiben als

$$g(x) = p(x) + r(x) \,,$$

wobei p das Taylorpolynom vom Grad q um einen Punkt $x_0 \in T$ und r das Restglied ist. Für das Restglied gilt die Abschätzung

$$|\, r(x) \,| \, \le \, \frac{1}{q!} \, \max_{\overline{T}} \left| \, \frac{\partial^{q+1}}{\partial v^{q+1}} \, g \, \right| h^{q+1} \, = \, O(h^{q+1}) \,, \tag{93.9}$$

wobei $v = (x - x_0)/\|x - x_0\|$ die Richtung von x_0 nach x ist und die Fehlerkonstante in dem O-Term nur von der Funktion g abhängt. Da die Quadraturformel linear und für $p \circ \Phi$ exakt ist, folgt weiter

$$|d|\, Q[g \circ \Phi] \, = \, |d|\, Q[p \circ \Phi] + |d|\, Q[r \circ \Phi] \, = \int_T p(x)\,dx + |d|\, Q[r \circ \Phi] \,.$$

Somit ist der Quadraturfehler gleich $|d|\, Q[r \circ \Phi] - \int_T r(x)\,dx$ und wegen (93.9) ist jeder der beiden Terme von der Größenordnung $|d| h^{q+1}$. Damit folgt die Behauptung aus der Dreiecksungleichung. □

Aus Lemma 93.4 folgt, daß die Summe der so ausgerechneten Elementsteifigkeitsmatrizen die Steifigkeitsmatrix A bei hinreichend glatten Koeffizientenfunktionen σ und c bis auf einen Fehler $O(h^{q+1})$ bezüglich der Zeilensummennorm approximiert. Da die Steifigkeitsmatrix symmetrisch ist, ergibt sich aus Satz 2.8 die gleiche Approximationsordnung auch für die Spektralnorm.

93.4 Ein Modellproblem

Im folgenden stellen wir die Steifigkeitsmatrix A für den negativen Laplace-Operator $-\Delta$ mit homogenen Dirichlet-Randbedingungen und einer Standard-

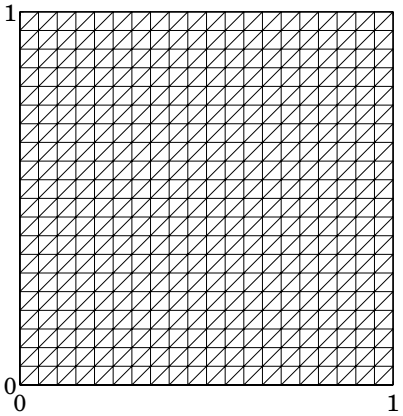
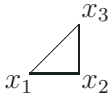
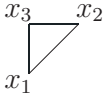


Abb. 93.3:
Triangulierung des Einheitsquadrats

triangulierung $\mathcal{T}$ des Einheitsquadrats $\Omega = (0,1) \times (0,1)$ auf, vgl. Abbildung 93.3. Abweichend von der bisherigen Konvention bezeichnen wir hier mit $h = 1/(\nu+1)$, $\nu \in \mathbb{N}$, den konstanten Abstand zwischen je zwei horizontalen bzw. vertikalen Gitterlinien. Die maximale Kantenlänge eines Dreiecks von $\mathcal{T}$ ist also $h\sqrt{2}$.

Diese Triangulierung enthält zwei Arten von Dreiecken:

(i) 

(ii) 

In beiden Fällen ist $d = \det \Phi' = h^2$. Betrachten wir zunächst ein Dreieck der Form (i): Die relevanten Einträge der zugehörigen Elementsteifigkeitsmatrix lauten nach (93.5)

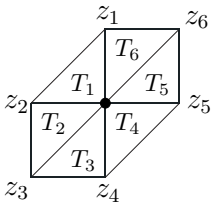
$$\frac{1}{2h^2} \begin{bmatrix} -h & 0 \\ h & -h \\ 0 & h \end{bmatrix} \begin{bmatrix} -h & h & 0 \\ 0 & -h & h \end{bmatrix} = \begin{bmatrix} 1/2 & -1/2 & 0 \\ -1/2 & 1 & -1/2 \\ 0 & -1/2 & 1/2 \end{bmatrix}.$$

Für Dreiecke der Form (ii) ergibt sich entsprechend

$$\frac{1}{2h^2} \begin{bmatrix} 0 & -h \\ h & 0 \\ -h & h \end{bmatrix} \begin{bmatrix} 0 & h & -h \\ -h & 0 & h \end{bmatrix} = \begin{bmatrix} 1/2 & 0 & -1/2 \\ 0 & 1/2 & -1/2 \\ -1/2 & -1/2 & 1 \end{bmatrix}.$$

Jeder innere Knoten z_0 der Triangulierung $\mathcal{T}$ grenzt an sechs Dreiecke $T_1, \ldots, T_6$

und hat sechs unmittelbare Nachbarknoten $z_1, \ldots, z_6$:



Bezeichnen wir mit $S_1, \ldots, S_6$ die Elementsteifigkeitsmatrizen zu $T_1, \ldots, T_6$, dann ergibt die Assemblierung (93.1) die folgenden Einträge in der zu z_0 gehörenden Zeile von A:

	z_0	z_1	z_2	z_3	z_4	z_5	z_6
S_1	1	$-1/2$	$-1/2$				
S_2	$1/2$		$-1/2$	0			
S_3	$1/2$			0	$-1/2$		
S_4	1				$-1/2$	$-1/2$	
S_5	$1/2$					$-1/2$	0
S_6	$1/2$	$-1/2$					0
A	4	-1	-1	0	-1	-1	0

Schließlich numerieren wir die inneren Knoten der Triangulierung $\mathcal{T}$ aus Abbildung 93.3 (also diejenigen Knoten der Triangulierung, die nicht zum Rand Γ gehören) in lexikographischer Reihenfolge, d. h. von links nach rechts und von oben nach unten,

$$\begin{array}{cccc} x_1 & x_2 & \cdots & x_\nu \\ x_{\nu+1} & x_{\nu+2} & \cdots & x_{2\nu} \\ \vdots & \vdots & & \vdots \\ & & \cdots & x_{\nu^2} \end{array} \tag{93.10}$$

wobei $n = \nu^2 = \dim V_0^{\mathcal{T}}$. Man beachte, daß Randknoten von $\mathcal{T}$ wegen der homogenen Dirichlet-Randbedingung nicht im Gleichungssystem auftreten. Mit dieser Numerierung erhalten wir die Steifigkeitsmatrix in der Blockform

$$A = \begin{bmatrix} C & -I & & & \\ -I & C & -I & & \\ & -I & \ddots & \ddots & \\ & & \ddots & \ddots & -I \\ & & & -I & C \end{bmatrix} \quad \text{mit} \quad C = \begin{bmatrix} 4 & -1 & & & \\ -1 & 4 & -1 & & \\ & -1 & \ddots & \ddots & \\ & & \ddots & \ddots & -1 \\ & & & -1 & 4 \end{bmatrix}. \tag{93.11}$$

Dabei ist $A \in \mathbb{R}^{n \times n}$ und $C \in \mathbb{R}^{\nu \times \nu}$; I bezeichnet die Einheitsmatrix aus $\mathbb{R}^{\nu \times \nu}$.

Bemerkung 93.5. Die gleiche Matrix A aus (93.11) ergibt sich übrigens auch bis auf einen Vorfaktor h^{-2}, wenn man anstelle finiter Elemente ein Differenzenverfahren zur Diskretisierung des Laplace-Operators verwendet. Für gegebene Näherungen $u_k \approx u(x_k)$ an den inneren Knoten x_k aus (93.10) approximiert man dabei die zweiten partiellen Ableitungen $u_{\xi\xi}$ und $u_{\eta\eta}$ durch zentrale Differenzenquotienten wie in Beispiel 83.1 und erhält

$$-\Delta u(x_k) = -u_{\xi\xi}(x_k) - u_{\eta\eta}(x_k) \approx \frac{-u_{k-1} + 2u_k - u_{k+1} - u_{k-\nu} + 2u_k - u_{k+\nu}}{h^2}$$
$$= \frac{4u_k - u_{k-1} - u_{k+1} - u_{k-\nu} - u_{k+\nu}}{h^2} .$$

Dabei sind die Funktionswerte von u, die zu Randpunkten von Ω gehören, wegen der Randbedingung durch Null zu ersetzen. Diese Diskretisierung von $L[u] = -\Delta$ durch ein Differenzenverfahren führt also auf den diskreten Operator $L_h = A/h^2$ mit der Matrix A aus (93.11). Man vergleiche dieses Ergebnis mit Beispiel 90.3. ◊

93.5 Eigenschaften der Steifigkeitsmatrix

Die Steifigkeitsmatrizen zur nodalen Basis haben in der Regel folgende charakteristische Eigenschaften, vgl. (93.11) zur Illustration:

- die Steifigkeitsmatrix ist eine Bandmatrix mit einer Bandbreite $O(\sqrt{n})$, wobei n die Dimension von $V_0^{\mathcal{T}}$ ist (die Gesamtzahl aller inneren Knoten);
- die Matrix ist innerhalb der Bandbreite dünn besetzt; die Zahl der von Null verschiedenen Nebendiagonaleinträge in der i-ten Zeile ist beschränkt durch die Anzahl der unmittelbaren Nachbarknoten des Knotens x_i.

Die Lösung des resultierenden linearen Gleichungssystems $A\boldsymbol{u}_h = b$ ist wesentlich aufwendiger als bei eindimensionalen Randwertaufgaben. So werden bei einer Cholesky-Zerlegung von A im allgemeinen die vielen leeren Nebendiagonalen innerhalb der Bandbreite aufgefüllt. Der Cholesky-Faktor hat somit $O(n\sqrt{n})$ von Null verschiedene Einträge und die Berechnung der Cholesky-Faktorisierung kostet $O(n^2)$ Operationen. Daher ist eine Cholesky-Zerlegung nur für relativ grobe Triangulierungen praktikabel. Statt dessen werden häufig Iterationsverfahren wie das (präkonditionierte) CG-Verfahren zur Lösung dieser Gleichungssysteme verwendet (vgl. auch die Beispiele 9.7 und 10.3) Bei Iterationsverfahren ist zu beachten, daß die Lösung wegen des Diskretisierungsfehlers nur bis auf eine $\mathcal{L}^2$-Genauigkeit $O(h^2)$ berechnet werden muß.

Beispiel 93.6. Abbildung 93.4 zeigt links eine Triangulierung zur Berechnung eines elektrischen Potentials in dem dargestellten Körper Ω mit zwei isolierenden Einschlüssen (vgl. Beispiel 70.1). Die Pfeile markieren zwei punktförmige

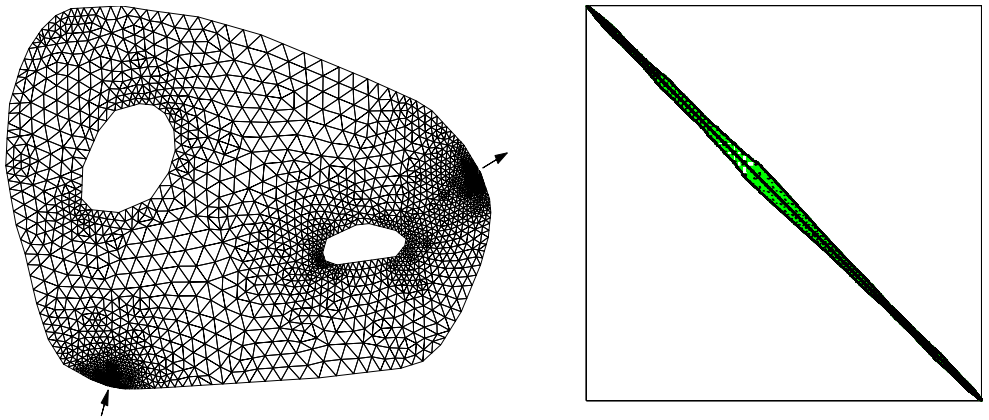


Abb. 93.4: Beispielgitter und die zugehörige Steifigkeitsmatrix

Elektroden, durch die Strom in den Körper eingespeist wird; dies entspricht der Neumann-Randbedingung aus (70.6). Man beachte die notwendige Verfeinerung der Triangulierung in der Nähe dieser Elektroden, um das Potential dort hinreichend genau approximieren zu können.

Die Triangulierung besteht aus $n = 2466$ Knoten und $m = 4636$ Dreiecken. Die Steifigkeitsmatrix $A \in \mathbb{R}^{n\times n}$ enthält lediglich $16672 \approx 6.8\,n$ von Null verschiedene Einträge. Bei optimierter Numerierung der Knoten sind diese von Null verschiedenen Einträge wie im rechten Teil der Abbildung angeordnet (die dunklen Punkte). Hellgrau hinterlegt ist der Bereich der Matrix, in dem Nulleinträge bei einer Cholesky-Zerlegung der Steifigkeitsmatrix zerstört werden. Wie man sehen kann, wird fast die gesamte Bandbreite von A aufgefüllt: Der Cholesky-Faktor hat 93344 von Null verschiedene Einträge. ◊

94 Schnelle direkte Löser

In einigen wenigen Spezialfällen weist das Galerkin-Gleichungssystem zusätzliche Strukturen auf, die besonders effiziente Lösungsverfahren ermöglichen.

Definition 94.1. Unter dem *Kronecker-Produkt* zweier Matrizen $B = [b_{ij}] \in \mathbb{K}^{m\times n}$ und $C \in \mathbb{K}^{\mu\times\nu}$ versteht man die Blockmatrix

$$B \otimes C = \begin{bmatrix} b_{11}C & b_{12}C & \cdots & b_{1n}C \\ b_{21}C & b_{22}C & \cdots & b_{2n}C \\ \vdots & \vdots & & \vdots \\ b_{m1}C & b_{m2}C & \cdots & b_{mn}C \end{bmatrix} \in \mathbb{K}^{m\mu\times n\nu}.$$

Beispiel 94.2. In dem Modellproblem aus Abschnitt 93.4 kann die Steifigkeitsmatrix A aus (93.11) mit Hilfe von Kronecker-Produkten in der Form

$$A = I \otimes T + T \otimes I \qquad \text{mit} \qquad T = \begin{bmatrix} 2 & -1 & & \\ -1 & 2 & \ddots & \\ & \ddots & \ddots & -1 \\ & & -1 & 2 \end{bmatrix}$$

geschrieben werden. Die Tridiagonalmatrix T stimmt bis auf einen Faktor h mit der Matrix aus dem eindimensionalen Beispiel 90.3 überein. Die obige Summendarstellung von A entspricht einer Zerlegung des Laplace-Operators in die beiden Summanden $u_{\xi\xi}$ und $u_{\eta\eta}$. ◊

Das folgende Resultat zeigt, daß die Eigenwerte und Eigenvektoren quadratischer Kronecker-Produkte einfach angegeben werden können.

Lemma 94.3. *Seien B und C diagonalisierbare $\nu \times \nu$-Matrizen, d. h. es gibt Eigenvektorbasen $\{v_i\}$ und $\{w_j\}$ des $\mathbb{C}^\nu$ mit zugehörigen (komplexen) Eigenwerten, $Bv_i = \mu_i v_i$ und $Cw_j = \lambda_j w_j$, $i, j = 1, \dots, \nu$. Dann hat die Matrix $B \otimes C$ die Eigenvektoren $z_{ij} = v_i \otimes w_j \in \mathbb{C}^{\nu^2}$ und es gilt $(B \otimes C) z_{ij} = \mu_i \lambda_j z_{ij}$, $i, j = 1, \dots, \nu$.*

Beweis. Dies folgt sofort aus Teil (a) von Aufgabe 9:

$$(B \otimes C)(v_i \otimes w_j) \;=\; (Bv_i) \otimes (Cw_j) \;=\; (\mu_i v_i) \otimes (\lambda_j w_j) \;=\; \mu_i \lambda_j z_{ij}\,.$$

Nach Teil (c) derselben Aufgabe bilden die Vektoren z_{ij} eine Basis von $\mathbb{C}^{\nu^2}$. □

Korollar 94.4. *Die Steifigkeitsmatrix A des Modellproblems aus Abschnitt* 93.4 *mit Dimension $n = \nu^2$ hat die Eigenwerte*

$$\lambda_{ij} = 4\big(\sin^2 \frac{\theta_i}{2} \,+\, \sin^2 \frac{\theta_j}{2}\big)\,, \qquad i, j = 1, \dots, \nu\,,$$

mit zugehörigen Eigenvektoren $z_{ij} = s_i \otimes s_j$, wobei $\theta_l = l\pi/(\nu+1)$ und

$$s_l = \big[\, \sin k\theta_l \,\big]_{k=1}^{\nu}\,, \qquad l = 1, \dots, \nu\,.$$

Beweis. Die Sinusvektoren s_j, $j = 1, \dots, \nu$, diagonalisieren nach Aufgabe VI.13 die Tridiagonalmatrix T aus Beispiel 94.2,

$$T s_j = (2 - 2\cos\theta_j) s_j \;=\; 4 \sin^2 \frac{\theta_j}{2}\, s_j\,, \qquad j = 1, \dots, \nu\,. \tag{94.1}$$

Da die Sinusvektoren natürlich auch Eigenvektoren der Einheitsmatrix sind, hat A nach Lemma 94.3 die Eigenvektoren $s_i \otimes s_j$ mit Eigenwerten

$$\lambda_{ij} \;=\; 4 \sin^2 \frac{\theta_j}{2} \,+\, 4 \sin^2 \frac{\theta_i}{2}\,.$$

□

Aus diesem Korollar folgt unmittelbar, daß die Inverse dieser Steifigkeitsmatrix die Faktorisierung

$$A^{-1} \;=\; (S \otimes S)(I \otimes D + D \otimes I)^{-1}(S \otimes S)^{-1} \tag{94.2}$$

besitzt. Hierbei ist $S \in \mathbb{R}^{\nu\times\nu}$ die Sinusmatrix aus (55.3) mit den Sinusvektoren s_l, $l = 1, \ldots, \nu$, als Spalten, und D ist die Diagonalmatrix mit den Diagonaleinträgen $d_i = 4\sin^2(\theta_i/2)$, $i = 1, \ldots, \nu$. Da die Inverse der Sinusmatrix nach Proposition 55.3 durch $2S/(\nu+1)$ gegeben ist, folgt

$$(S \otimes S)^{-1} \;=\; S^{-1} \otimes S^{-1} \;=\; \frac{4}{(\nu+1)^2}(S \otimes S)\,.$$

Für das erste Gleichheitszeichen vergleiche man Aufgabe 9 (a).

Der Koeffizientenvektor $\boldsymbol{u}_h$ der Galerkin-Näherung ergibt sich aus dem Gleichungssystem $A\boldsymbol{u}_h = b$ mit b aus (90.5). Nach (94.2) ist daher

$$\boldsymbol{u}_h \;=\; \frac{4}{\nu^2}(S \otimes S)(I \otimes D + D \otimes I)^{-1}(S \otimes S)\,b\,,$$

und im folgenden beschreiben wir, wie diese Darstellung in $O(n \log n)$ Operationen ausgewertet werden kann. Dabei ist $n = \nu^2$ die Anzahl der inneren Knoten der Triangulierung des Modellproblems.

Die Matrix $I \otimes D + D \otimes I$ ist eine Diagonalmatrix, deren Berechnung und anschließende Multiplikation mit einem Vektor lediglich $O(n)$ Operationen kostet. Interessanter sind die beiden Matrix-Vektor-Multiplikationen mit der Matrix $S \otimes S$, die mit der schnellen Sinustransformation (DST) realisiert werden können. Wir betrachten dazu konkret die Multiplikation

$$\widehat{b} = (S \otimes S)b\,. \tag{94.3}$$

Mit $b = [b_j]$ assoziieren wir im weiteren die Matrix

$$B \;=\; \begin{bmatrix} b_1 & b_2 & \cdots & b_\nu \\ b_{\nu+1} & b_{\nu+2} & \cdots & b_{2\nu} \\ \vdots & \vdots & & \vdots \\ & & \cdots & b_{\nu^2} \end{bmatrix} \in \mathbb{R}^{\nu\times\nu} \tag{94.4}$$

und entsprechend identifizieren wir $\widehat{b}$ mit der Matrix $\widehat{B} \in \mathbb{R}^{\nu\times\nu}$. Aus (94.3) erhalten wir dann die Gleichung

$$\widehat{B} \;=\; SBS^T \;=\; S(SB^T)^T\,, \tag{94.5}$$

Initialisierung: Gegeben sei die rechte Seite b aus (90.5)

% Für eine Matrix $M = [m_1 \cdots m_\nu]$ mit Spalten $m_j \in \mathbb{R}^\nu$, $j = 1, \ldots, \nu$, bezeichne
% $\mathrm{DST}(M)$ die Matrix mit den Spalten $\mathrm{DST}(m_j, \nu)$ (vgl. Algorithmus 55.1)

$h = 1/(\nu + 1)$

definiere $B \in \mathbb{R}^{\nu \times \nu}$ wie in (94.4)

% zweidimensionale schnelle Sinustransformation von B:

$Z = \mathrm{DST}(B^T)$

$\widehat{B} = \mathrm{DST}(Z^T)$

$E = \frac{1}{4}\big[\,(\sin^2(\theta_i/2) + \sin^2(\theta_j/2))^{-1}\,\big]_{ij}$ % $\theta_k = k\pi/(\nu+1)$

$\widehat{U} = \widehat{B} \bullet E$ % komponentenweise Multiplikation

% schnelle Rücktransformation:

$W = \mathrm{DST}(\hat{U}^T)$

$U = \frac{4}{(\nu+1)^2}\mathrm{DST}(W^T)$

Ergebnis: U enthält die Werte von $u_k = u_h(x_k)$ in der Anordnung (94.6)

Algorithmus 94.1: Schneller direkter Löser im Einheitsquadrat

vgl. Aufgabe 9 (b), an der man ablesen kann, daß die Matrix-Vektor-Multiplikation (94.3) mit zwei Matrix-Matrix-Multiplikationen mit der Sinusmatrix S, also mit 2ν schnellen Sinustransformationen (Algorithmus 55.1) implementiert werden kann.

Die Transformation (94.5) heißt *schnelle zweidimensionale Sinustransformation*, in Analogie zu der schnellen zweidimensionalen Kosinustransformation aus Beispiel 55.5. Algorithmus 94.1 faßt alle Schritte zur Lösung des Gleichungssystems $A\boldsymbol{u}_h = b$ zusammen. Das Ergebnis

$$U = \begin{bmatrix} u_1 & u_2 & \cdots & u_\nu \\ u_{\nu+1} & u_{\nu+2} & \cdots & u_{2\nu} \\ \vdots & \vdots & & \vdots \\ & & \cdots & u_{\nu^2} \end{bmatrix} \in \mathbb{R}^{\nu \times \nu} \tag{94.6}$$

dieses Algorithmus enthält die Werte $u_k = u_h(x_k)$ der Galerkin-Näherung u_h an den Knoten x_k in der durch die Numerierung (93.10) vorgegebenen Anordnung.

Aufwand. Der Algorithmus erfordert 4ν schnelle Sinustransformationen, also $O(\nu^2 \log \nu) = O(n \log n)$ Operationen. Demgegenüber sind die n Divisionen durch die Eigenwerte von A vernachlässigbar. ◊

Ergänzend sei noch auf das Buch von Iserles [56] verwiesen, in dem ein schneller direkter Löser für die Laplace-Gleichung im Einheitskreis vorgestellt wird.

95 Mehrgitterverfahren

Während die FFT-Techniken des vorangegangenen Abschnitts fast ausschließlich auf die Laplace-Gleichung und einige wenige Gebiete Ω anwendbar sind, bilden die sogenannten Mehrgitterverfahren im allgemeinen die wohl effizientesten Methoden zur Lösung der Galerkin-Gleichungssysteme $A\boldsymbol{u}_h = b$. Diesen Verfahren liegt die Beobachtung zugrunde, daß die konventionellen Iterationsverfahren wie das Gesamtschritt- oder das Einzelschrittverfahren zwar nur sehr langsam konvergieren, aber dennoch bereits nach wenigen Iterationsschritten sehr gute Approximationen an die hochfrequenten Anteile der Lösung liefern. Die niederfrequenten Lösungskomponenten können anschließend anderweitig berechnet werden.

Exemplarisch betrachten wir das Iterationsverfahren

$$\boldsymbol{u}^{(k+1)} = \boldsymbol{u}^{(k)} + \omega(b - A\boldsymbol{u}^{(k)})\,, \qquad k = 0, 1, 2, \dots\,, \tag{95.1}$$

bei dem $\omega > 0$ ein geeignet zu wählender Parameter ist. Das Verfahren (95.1) wird bisweilen *Richardson-Iteration* genannt; für das Modellproblem aus Abschnitt 93.4 und $\omega = 1/4$ stimmt (95.1) mit dem Gesamtschrittverfahren aus Abschnitt 8 überein. In diesem Beispiel hat die Iterationsmatrix $T_\omega = I - \omega A$ nach Korollar 94.4 die Eigenwerte

$$\lambda_{ij} \;=\; 1 - 4\omega\,\big(\sin^2\frac{i\pi}{2(\nu+1)} \;+\; \sin^2\frac{j\pi}{2(\nu+1)}\big)\,, \tag{95.2}$$

$i, j = 1, \dots, \nu$. Die zugehörigen Eigenvektoren z_{ij} sind wie bei der Steifigkeitsmatrix A die Kronecker-Produkte der Sinusvektoren

$$z_{ij} = s_i \otimes s_j \qquad \text{mit} \qquad s_l = \big[\,\sin\frac{kl\pi}{\nu+1}\,\big]_{k=1}^{\nu}\,. \tag{95.3}$$

Für $\omega \le 1/4$ sind alle Eigenwerte betragsmäßig kleiner als Eins, aber wegen

$$\varrho(T_\omega) \;=\; \lambda_{1,1} \;=\; 1 - O(h^2)\,, \qquad h = 1/(\nu+1)\,, \tag{95.4}$$

ist die Konvergenz des Verfahrens für kleine h viel zu langsam.

Andererseits liegen alle Eigenwerte von T_ω für ein festes $\omega \in (0, 1/8]$ zwischen Null und Eins und werden mit wachsendem i und j immer kleiner. Die niederfrequenten Eigenvektoren ($i, j \ll \nu/2$) von T_ω gehören zu Eigenwerten nahe bei Eins, die hochfrequenten Eigenvektoren ($i, j \gg \nu/2$) gehören zu Eigenwerten in der Nähe von $1 - 8\omega$. Speziell für $\omega = 1/8$ liegen diese Eigenwerte also sehr nahe bei Null.

Betrachten wir nun den Fehler $\boldsymbol{e}^{(k)} = \boldsymbol{u}_h - \boldsymbol{u}^{(k)}$, $k \in \mathbb{N}$, etwas genauer: Aus der Iterationsvorschrift folgt

$$\boldsymbol{e}^{(k)} \;=\; \boldsymbol{e}^{(k-1)} - \omega A(\boldsymbol{u}_h - \boldsymbol{u}^{(k-1)}) \;=\; T_\omega \boldsymbol{e}^{(k-1)} \;=\; T_\omega^k \boldsymbol{e}^{(0)}\,,$$

und mit einer ähnlichen Argumentation wie bei der Untersuchung der Potenzmethode in Abschnitt 25 ergibt sich aus der Verteilung der Eigenwerte, daß die hochfrequenten Anteile des Ausgangsfehlers $\boldsymbol{e}^{(0)}$ für $\omega \leq 1/8$ wesentlich schneller gedämpft werden als die niederfrequenten Anteile. Der Fehler wird somit immer „glatter“ im Verlauf der Iteration; die Iterationsmatrix T_ω wird daher *Glätter* genannt. Hieraus folgt, daß $\boldsymbol{u}^{(k)} = \boldsymbol{u}_h - \boldsymbol{e}^{(k)}$ eine gute Approximation der hochfrequenten Lösungskomponenten ist.

Die Iteration (95.1) hat bei allgemeineren elliptischen Differentialoperatoren vergleichbare Glättungseigenschaften, sofern die Produkte $\omega\lambda$, $\lambda \in \sigma(A)$, nicht viel größer als Eins werden; diese Voraussetzung ist zum Beispiel für

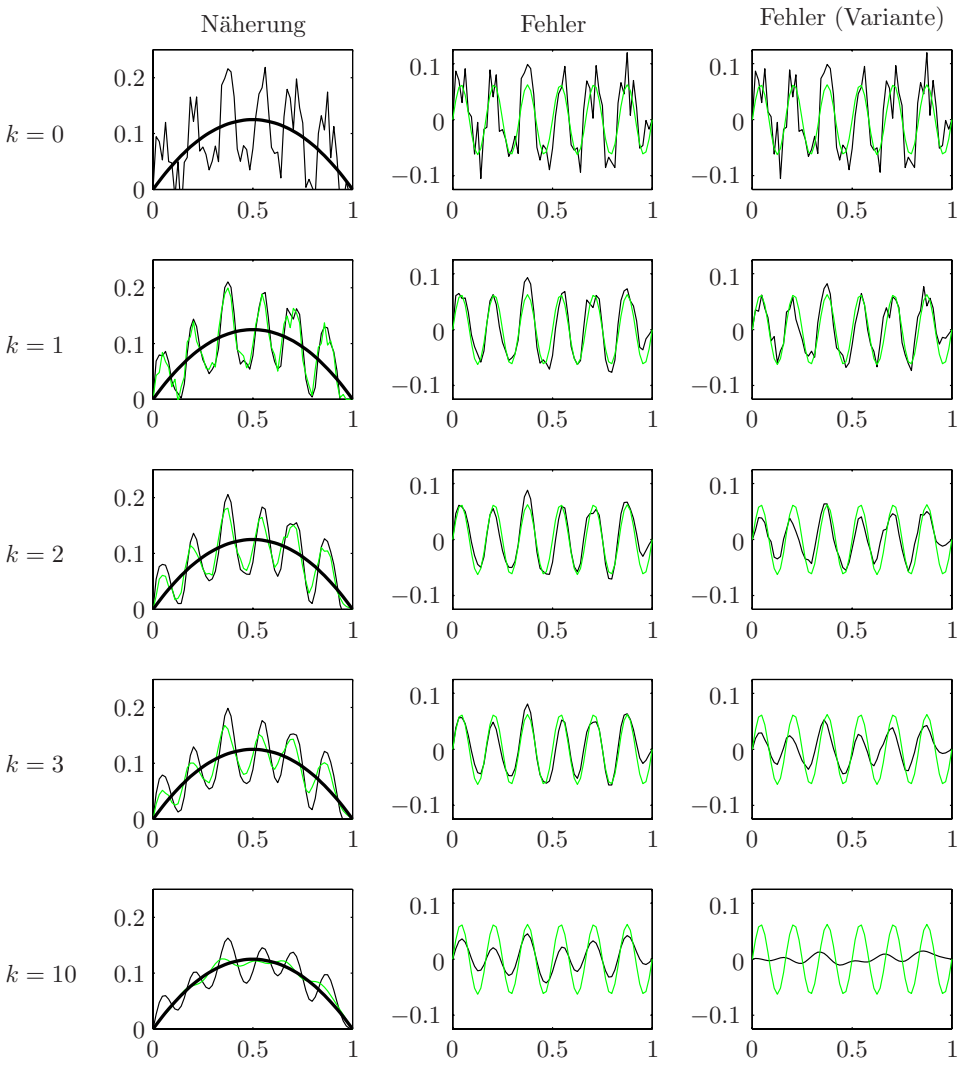
$$\omega = 1/\|A\|_\infty \tag{95.5}$$

erfüllt, da aufgrund der Symmetrie von A die rechte Seite von (95.5) nach Satz 2.8 immer eine untere Schranke von $1/\varrho(A)$ bildet. Für das Modellproblem führt (95.5) wieder auf $\omega = 1/8$.

Beispiel 95.1. Zur Illustration betrachten wir die Approximationen $\boldsymbol{u}^{(k)}$ und die Fehler $\boldsymbol{e}^{(k)}$ der Iteration (95.1) im eindimensionalen Spezialfall des Modellproblems, und zwar für das Randwertproblem

$$-u'' = f \quad \text{in } (0,1)\,, \qquad u(0) = u(1) = 0\,,$$

mit $f = 1$ und der Lösung $u(x) = x(1-x)/2$, vgl. Beispiel 90.3. Gemäß (95.5) wird der Parameter $\omega = h/4$ verwendet. Abbildung 95.1 zeigt in der mittleren Spalte den Fehler (den zu $\boldsymbol{e}^{(k)}$ gehörenden linearen Spline über $[0,1]$) nach $k = 0, 1, 2, 3$ und 10 Iterationen. Ausgangspunkt ist ein Fehler im mittleren Bild der obersten Zeile, der aus einer Überlagerung einer niederfrequenten Sinusschwingung (sechs Perioden, heller dargestellt) und einem zweiten hochfrequenten Anteil besteht. Nach drei Iterationen ist der Fehler bereits recht glatt und ähnelt stark der Sinusschwingung aus dem Anfangsfehler. Nach zehn Iterationen erkennt man auch eine zunehmende Dämpfung des niederfrequenten Fehleranteils (zum Vergleich im helleren Farbton immer die Ausgangssinusschwingung). Anstelle der Iteration (95.1) kann auch das Einzelschrittverfahren (vgl. Abschnitt 8) als Glätter verwendet werden; für dieses Verfahren wird die Fehlerentwicklung in der rechten Spalte der Abbildung dargestellt. Die erste Spalte zeigt schließlich die exakte Lösung (fett dargestellt) sowie die Näherungen der Iteration (95.1) (dunklere Kurve) und des Einzelschrittverfahrens (hellere Kurve).



Ein Iterationsschritt von Algorithmus 95.1 mit $\kappa = 3$:

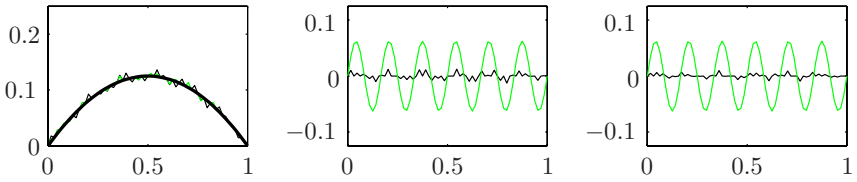


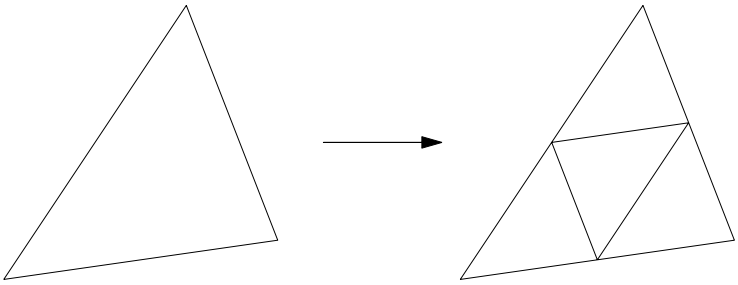
Abb. 95.1: Fehlerentwicklung für die Iteration (95.1)

Die letzte Zeile der Abbildung illustriert das unten folgende Zweigitterverfahren (Algorithmus 95.1), bei dem die Iteration (95.1) nach drei Schritten abgebrochen und der verbliebene Fehler auf einem gröberen Gitter weiter reduziert wird. Wie man sieht, ist diese Approximation der Lösung wesentlich besser. ◊

Zur Definition des angesprochenen Zweigitterverfahrens übernehmen wir aus Kapitel X das Konzept der Multiskalenbasen und konstruieren (zunächst) zwei Teilräume V_H und V_h mit

$$V_H \subset V_h \subset H_0^1(\Omega)\,. \tag{95.6}$$

Hierfür gehen wir von einer regulären Triangulierung $\mathcal{T}_H$ von Ω mit maximaler Kantenlänge H und minimalem Innenwinkel $\alpha_0 > 0$ aus und definieren die feine Triangulierung $\mathcal{T}_h$, indem alle Dreiecke aus $\mathcal{T}_H$ wie folgt in vier kongruente Teildreiecke zerlegt werden:



Die zugehörigen Unterräume $V_H = V_0^{\mathcal{T}_H}$ und $V_h = V_0^{\mathcal{T}_h}$ der stückweise linearen Funktionen mit homogenen Randwerten auf $\Gamma = \partial\Omega$ erfüllen dann die Zweiskalenbedingung (95.6). Dabei können in V_H nur „relativ niederfrequente“ Funktionen aus $H_0^1(\Omega)$ gut approximiert werden, während zur Darstellung höherer Frequenzen der feinere Raum V_h notwendig ist.

In einem Iterationsschritt des Zweigitterverfahrens mit Startnäherung $u^{(l)} \in V_h$ werden zunächst κ Iterationen (95.1) mit ω aus (95.5) für das Galerkin-Gleichungssystem $A_h \boldsymbol{u}_h = b_h$ durchgeführt, das zu der feinen Triangulierung $\mathcal{T}_h$ gehört. Die zuletzt berechnete Iterierte bezeichnen wir mit $\tilde{\boldsymbol{u}}_h$; sie gehört zu einer Funktion $\tilde{u}_h \in V_h$. Der Fehler $e_h = u_h - \tilde{u}_h$ erfüllt das Variationsproblem

$$a(e_h, v) \;=\; a(u_h, v) - a(\tilde{u}_h, v) \;=\; \ell(v) - a(\tilde{u}_h, v) \qquad \text{für alle } v \in V_h\,,$$

und aufgrund unserer Vorüberlegungen können wir erwarten, daß dieser Fehler bereits für relativ kleine κ im Raum V_H der niederfrequenten Funktionen gut approximiert werden kann. Es bietet sich daher an, eine Approximation $e_H \approx e_h$ aus V_H zu bestimmen, indem wir dieses Variationsproblem in V_H lösen:

$$a(e_H, v) \;=\; \ell(v) - a(\tilde{u}_h, v) \qquad \text{für alle } v \in V_H\,. \tag{95.7}$$

Mit Hilfe dieser sogenannten *Grobgitterkorrektur* erhalten wir eine neue Approximation

$$u^{(l+1)} \;=\; \tilde{u}_h + e_H \;\approx\; \tilde{u}_h + e_h \;=\; u_h$$

der Lösung u_h und der Iterationsschritt des Zweigitterverfahrens ist damit abgeschlossen.

Die Realisierung von (95.7) führt wieder auf ein lineares Gleichungssystem, diesmal mit der Steifigkeitsmatrix A_H für den Unterraum V_H. Um die Matrix A_H zu berechnen, entwickeln wir die Basisfunktionen $\Lambda_{H,i}$, $i = 1, \ldots, n_H$, von V_H in die Basis $\{\Lambda_{h,j}\}_{j=1}^{n_h}$ von V_h,

$$\Lambda_{H,i} \;=\; \sum_{k=1}^{n_h} \xi_{ik} \Lambda_{h,k}\,, \qquad i = 1, \ldots, n_H\,,$$

und sammeln die Koeffizienten in der Matrix $R = [\xi_{ik}]$. Dann ergibt sich

$$\begin{aligned}(A_H)_{ij} \;&=\; a(\Lambda_{H,i}, \Lambda_{H,j}) \;=\; \sum_{k,l=1}^{n_h} \xi_{ik}\, a(\Lambda_{h,k}, \Lambda_{h,l})\, \xi_{jl} \;=\; \sum_{k,l=1}^{n_h} \xi_{ik} (A_h)_{kl} \xi_{jl} \\ &=\; (R A_h R^*)_{ij}\,,\end{aligned}$$

das heißt es ist

$$A_H = R A_h R^*\,, \qquad R = [\xi_{ik}] \in \mathbb{R}^{n_H \times n_h}\,. \tag{95.8}$$

Bezeichnen wir schließlich mit $\boldsymbol{e}_H$ den Koeffizientenvektor von e_H bezüglich der Grobgitterbasis, so folgt aus (95.7) entsprechend, daß

$$\begin{aligned}(A_H \boldsymbol{e}_H)_i \;&=\; a(e_H, \Lambda_{H,i}) \;=\; \ell(\Lambda_{H,i}) - a(\tilde{u}_h, \Lambda_{H,i}) \\ &=\; \sum_{k=1}^{n_h} \xi_{ik} \big(\ell(\Lambda_{h,k}) - a(\tilde{u}_h, \Lambda_{h,k})\big) \;=\; \sum_{k=1}^{n_h} \xi_{ik} (b_h - A_h \tilde{\boldsymbol{u}}_h)_k \\ &=\; \big(R(b_h - A_h \tilde{\boldsymbol{u}}_h)\big)_i\,.\end{aligned}$$

Das Variationsproblem (95.7) entspricht damit dem Gleichungssystem

$$A_H \boldsymbol{e}_H \;=\; R(b_h - A_h \tilde{\boldsymbol{u}}_h)\,.$$

Schließlich ist $R^* \boldsymbol{e}_H$ der Koeffizientenvektor von e_H bezüglich der Basis von V_h (vgl. auch Aufgabe VI.1); somit hat die neue Näherung $u^{(l+1)}$ des Zweigitteriterationsschritts den Koeffizientenvektor

$$\tilde{\boldsymbol{u}}_h + R^* \boldsymbol{e}_H\,, \qquad A_H \boldsymbol{e}_H \;=\; R(b_h - A_h \tilde{\boldsymbol{u}}_h)\,, \tag{95.9}$$

Initialisierung: A_h, b_h und R seien gegeben; $\boldsymbol{u}$ sei der Koeffizientenvektor einer Startnäherung $u^{(0)} \in V_h$

% Während der Iteration: $\boldsymbol{u}$ ist der Koeffizientenvektor von $u^{(l)} \in V_h$
% $\boldsymbol{u}^{(k)}$ ist die aktuelle Iterierte von (95.1)

$A_H = RA_hR^*$
$\omega = 1/\|A_h\|_\infty$
`for` $l = 0, 1, 2, \ldots$ `do` % Iterationsschritt des Zweigitterverfahrens
  $\boldsymbol{u}^{(0)} = \boldsymbol{u}$
  `for` $k = 1, 2, \ldots, \kappa$ `do` % Glättungsschritte
    $\boldsymbol{u}^{(k)} = \boldsymbol{u}^{(k-1)} + \omega(b_h - A_h\boldsymbol{u}^{(k-1)})$
  `end for` % k-Schleife
  $\tilde{\boldsymbol{u}}_h = \boldsymbol{u}^{(\kappa)}$
  löse $A_H\boldsymbol{e}_H = R(b_h - A_h\tilde{\boldsymbol{u}}_h)$ % Grobgitterkorrektur
  $\boldsymbol{u} = \tilde{\boldsymbol{u}}_h + R^*\boldsymbol{e}_H$ % Koeffizientenvektor von $u^{(l+1)} \in V_h$
`until` stop % end for l

Ergebnis: $\boldsymbol{u} = [u_j]$ enthält die Entwicklungskoeffizienten von $u^{(l)} = \sum_{j=1}^{n_h} u_j\Lambda_{h,j} \approx u_h$

Algorithmus 95.1: Zweigitterverfahren

vgl. Algorithmus 95.1.

Die Multiplikationen mit R entsprechen sogenannten *Restriktionen* von V_h nach V_H, Multiplikationen mit R^* sind *Prolongationen* von V_H nach V_h. Die Anzahl der von Null verschiedenen Elemente in R ist ähnlich wie bei der Steifigkeitsmatrix A_h.

Beispiel. Wir betrachten die eindimensionale Situation, bei der $\triangle_h$ und $\triangle_H$ äquidistante Gitter mit Gitterweiten h und $H = 2h$ sind. Die Hutfunktionen $\Lambda_{H,i}$ der inneren Gitterpunkte setzen sich aus jeweils drei Hutfunktionen über dem feinen Gitter zusammen,

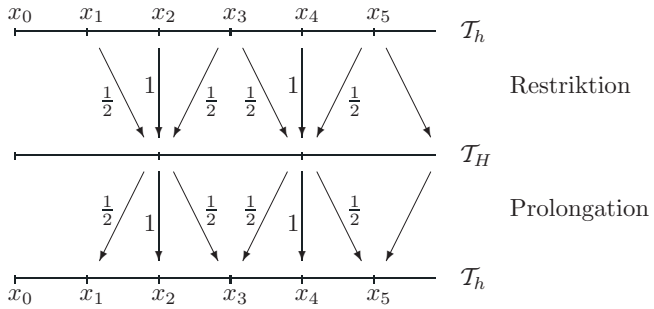
$$\Lambda_{H,i} \;=\; \frac{1}{2}\Lambda_{h,2i-1} + \Lambda_{h,2i} + \frac{1}{2}\Lambda_{h,2i+1}\,, \qquad i = 1, \ldots, n-1\,,$$

und somit ist

$$R = \begin{bmatrix} 1/2 & 1 & 1/2 & & & & & \\ & & 1/2 & 1 & 1/2 & & & \\ & & & & & \ddots & & \\ & & & & & & 1/2 & 1 & 1/2 \end{bmatrix} \in \mathbb{R}^{(n-1)\times(2n-1)}$$

mit $n = 1/H$. Die folgende Skizze veranschaulicht die Linearkombinationen der Näherungswerte an den Gitterknoten, die bei einer Restriktion beziehungsweise

einer Prolongation berechnet werden müssen:



$\Diamond$

Wir verweisen schließlich noch einmal auf die numerischen Ergebnisse in Abbildung 95.1: Die unterste Zeile dieser Abbildung zeigt das Resultat *eines* Iterationsschritts dieses Zweigitterverfahrens mit $\kappa = 3$ Glättungsschritten und anschließender Grobgitterkorrektur; das Zwischenergebnis vor der Grobgitterkorrektur ist in der vierten Zeile dieser Abbildung dargestellt ($k = 3$). Die Glättungsschritte reduzieren die hochfrequenten Fehlerkomponenten, die Grobgitterkorrektur entfernt hingegen die niederfrequenten Anteile. Mit dem Einzelschrittverfahren als Glätter ist die Fehlerreduktion insgesamt noch etwas besser, wie man anhand der rechten Spalte von Abbildung 95.1 erkennen kann.

Eine sorgfältige Konvergenzanalyse des Algorithmus zeigt, daß die Zweigitteriteration eine Kontraktion ist und der Kontraktionsfaktor unter milden Voraussetzungen durch eine Konstante abgeschätzt werden kann, die nur über den kleinsten Innenwinkel α_0 von der Triangulierung $\mathcal{T}_h$ abhängt (vgl. Braess [10]). Insbesondere ist diese Konstante unabhängig von h. Die angesprochenen Voraussetzungen sind im wesentlichen mit denen des Regularitätssatzes 92.2 identisch. Daß der Kontraktionsfaktor unabhängig von h ist, also unabhängig von der Feinheit der Triangulierung, ist von nicht zu unterschätzender Bedeutung, da hohe Genauigkeitsanforderungen feine Triangulierungen erfordern. Zum Vergleich sei in Erinnerung gerufen, vgl. (95.4), daß die Iteration (95.1) ohne Grobgitterkorrektur für $h \to 0$ immer langsamer konvergiert. Auf der anderen Seite wird natürlich die Grobgitterkorrektur, also das Lösen eines linearen Gleichungssystems mit der Matrix A_H für $h \to 0$ auch immer aufwendiger.

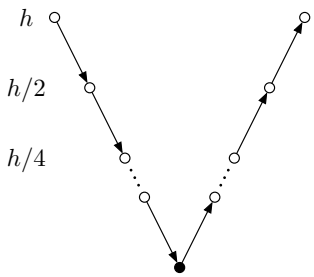
Beispiel. Für das zweidimensionale Modellproblem aus Abschnitt 93.4 mit $\nu = 100$ Gitterpunkten in jeder Dimension der feinen Triangulierung ergeben sich in Abhängigkeit von der Anzahl κ der Glättungsschritte die folgenden Konvergenzfaktoren (auf zwei Dezimalstellen genau):

κ	1	2	3	4	5	6	7	8	9	10
$T_{1/8}$	0.75	0.56	0.42	0.32	0.24	0.18	0.13	0.12	0.11	0.10
$\mathcal{L}$	0.35	0.14	0.10	0.07	0.05	0.04	0.04	0.03	0.03	0.03

Die mittlere Zeile bezieht sich auf das Zweigitterverfahren in der in Algorithmus 95.1 angegebenen Form, die untere Zeile gehört zu der Variante mit dem Einzelschrittoperator als Glätter. Ohne Grobgitterkorrektur ergeben sich die Konvergenzfaktoren $\varrho(T_{1/8}) \approx 0.9998$ bzw. $\varrho(\mathcal{L}) \approx 0.9990$.

Neben der generellen Überlegenheit des Zweigitterverfahrens bestätigt die Tabelle auch die numerischen Ergebnisse aus Abbildung 95.1, wonach das Einzelschrittverfahren der Iteration (95.1) als Glätter überlegen ist. Schließlich sei noch erwähnt, daß sich die Einträge in der mittleren Zeile der Tabelle bei einem Übergang zu doppelt so vielen Gitterpunkten in jeder Dimension nicht mehr verändern; die Konvergenzfaktoren beim Einzelschrittverfahren werden zum Teil geringfügig schlechter. ◊

Als Faustregel kann damit gerechnet werden, daß die feine Triangulierung $\mathcal{T}_h$ etwa viermal so viele Knoten enthält wie die grobe Triangulierung $\mathcal{T}_H$ und daß die Bandbreite von A_h etwa doppelt so groß ist wie die Bandbreite von A_H. Daher wird man erwarten, daß der Aufwand für eine Cholesky-Zerlegung von A_H im Vergleich zu A_h um etwa den Faktor 16 günstiger ist. Dies wird in vielen Fällen jedoch immer noch so teuer sein, daß man das Problem auch in dem groben Finite-Elemente-Raum nicht exakt lösen möchte. Statt dessen bietet es sich an, Algorithmus 95.1 rekursiv zu implementieren, also das Problem auf jeder Triangulierung nach einigen Glättungsschritten auf die nächstgröbere Triangulierung zu restringieren, dort weiter zu iterieren und so fort. Auf diese Weise erhält man eine von mehreren möglichen Varianten des *Mehrgitterverfahrens*, den sogenannten *V-Zyklus*:



Diese Skizze soll die Vorgehensweise im V-Zyklus illustrieren, also *einen* vollständigen Iterationsschritt des Mehrgitterverfahrens. Der V-Zyklus startet auf der feinsten Triangulierung (dargestellt durch den Knoten links oben) und steigt dann entlang der Knoten schrittweise zur gröbsten Triangulierung ab. Ein nicht ausgefüllter Knoten (○) zeigt an, daß auf der jeweiligen Triangulierung κ Glättungsschritte durchgeführt werden. Der Knoten für die gröbste Triangulierung ist ausgefüllt (●); hier wird nicht geglättet, statt dessen wird das (kleine) Gleichungssystem exakt gelöst. Nach der Prolongation auf die nächstgröbere Triangulierung werden wieder κ Glättungsschritte durchge-

führt ($\circ$) bevor weiter prolongiert wird. Auch für den V-Zyklus läßt sich lineare Konvergenz mit einem von h unabhängigen Konvergenzfaktor beweisen (vgl. Braess [10]).

Aufwand. Wird der Aufwand für die einfachen Glättungsoperatoren auf je cn_h Operationen pro Glättungsschritt bei einer Triangulierung mit n_h Knoten geschätzt, dann führt die Faustregel $n_H = n_h/4$ für $H = 2h$ auf etwa

$$2(c\kappa n + c\kappa n/4 + c\kappa n/16 + \dots) \approx 8/3 c\kappa n$$

Operationen für alle Glättungsschritte eines V-Zyklus, wobei n die Knotenzahl der feinsten Triangulierung angibt. Der Aufwand aller Restriktionen und Prolongationen liegt ebenfalls in der Größenordnung $O(n)$. Rechnet man schließlich noch einen konstanten Aufwand für die Lösung der Galerkin-Gleichung auf dem gröbsten Unterraum hinzu, so ergeben sich insgesamt $O(n)$ Operationen für einen Iterationsschritt des V-Zyklus. Da für eine vorgeschriebene Genauigkeit wegen des von h unabhängigen Konvergenzfaktors eine feste (und meist kleine) Anzahl von Mehrgitter-Iterationen ausreichend ist, liegt auch der Gesamtaufwand zur Berechnung von u_h mit dieser Genauigkeit in der Größenordnung von $O(n)$ Operationen. Dieser Aufwand ist optimal, da insgesamt n Näherungswerte zu berechnen sind. $\Diamond$

96 Ein Fehlerschätzer

Zum Abschluß dieses Kapitels widmen wir uns noch der Frage, wie die Genauigkeit der berechneten Galerkin-Lösung überprüft werden kann. Wir verwenden weiterhin die Notation der vorangegangenen Abschnitte und betrachten das Randwertproblem

$$L[u] = -\operatorname{div}(\sigma \operatorname{grad} u) + cu = f \quad \text{in } \Omega\,, \qquad u|_\Gamma = 0\,, \tag{96.1}$$

unter den üblichen Voraussetzungen, daß $f \in \mathcal{L}^2(\Omega)$ und σ und c beschränkte Funktionen sind mit $0 < \sigma_0 \le \sigma \le \sigma_\infty$ und $c \ge 0$; wir nehmen lediglich zusätzlich an, daß für jedes Dreieck $T \in \mathcal{T}$ die Einschränkung $\sigma|_T$ zu $H^1(T)$ gehört. Ziel der nachfolgenden Umformungen ist es, eine berechenbare A-posteriori-Abschätzung für den Fehler der Galerkin-Näherung herzuleiten. Eine ganze Reihe weiterer Fehlerschätzer findet sich in der Monographie von Ainsworth und Oden [3].

Die exakte (schwache) Lösung u und die Galerkin-Näherung u_h lösen das Variationsproblem

$$a(u,v) = \int_\Omega f v\,dx \qquad \text{bzw.} \qquad a(u_h,v) = \int_\Omega f v\,dx$$

für alle $v \in H_0^1(\Omega)$ bzw. alle $v \in V_h = V_0^{\mathcal{T}}$. Für den Fehler $e_h = u - u_h$ gilt somit einerseits

$$a(e_h, v) = 0 \qquad \text{für alle } v \in V_0^{\mathcal{T}}\,, \tag{96.2}$$

während andererseits für beliebiges $v \in H_0^1(\Omega)$

$$\begin{aligned} a(e_h, v) \;&=\; \int_\Omega f v\,dx - \int_\Omega \sigma \operatorname{grad} u_h \cdot \operatorname{grad} v\,dx - \int_\Omega c u_h v\,dx \\ &=\; \int_\Omega (f - c u_h) v\,dx - \int_\Omega \sigma \operatorname{grad} u_h \cdot \operatorname{grad} v\,dx\,. \end{aligned} \tag{96.3}$$

Aus der $H_0^1(\Omega)$-Elliptizität der Bilinearform a, vgl. Proposition 90.2, und aus (96.2) ergibt sich weiterhin

$$a_0 \|u - u_h\|^2_{H^1(\Omega)} \;\le\; a(e_h, e_h) \;=\; a(e_h, e_h - \psi) \tag{96.4}$$

für jedes beliebige $\psi \in V_0^{\mathcal{T}}$. Für eine scharfe A-posteriori-Abschätzung von $\|u - u_h\|_{H^1(\Omega)}$ werden wir weiter unten in (96.7) für ψ eine möglichst gute Approximation an e_h auswählen. Zunächst erhalten wir jedoch für beliebiges $\psi \in V_0^{\mathcal{T}}$ durch Einsetzen von (96.3) in (96.4) die Abschätzung

$$a_0 \|u - u_h\|^2_{H^1(\Omega)} \;\le\; \int_\Omega (f - c u_h)(e_h - \psi)\,dx - \int_\Omega \sigma \operatorname{grad} u_h \cdot \operatorname{grad}(e_h - \psi)\,dx\,,$$

und der zweite Term kann durch partielle Integration auf jedem Teildreieck weiter umgeformt werden, vgl. (89.2):

$$\begin{aligned} \int_\Omega \sigma \operatorname{grad} u_h \cdot \operatorname{grad}(e_h - \psi)\,dx \;&=\; \sum_{T_i \in \mathcal{T}} \int_{T_i} \sigma \operatorname{grad} u_h \cdot \operatorname{grad}(e_h - \psi)\,dx \\ &=\; \sum_{T_i \in \mathcal{T}} \Big(\int_{\partial T_i} (e_h - \psi)\, \sigma \frac{\partial}{\partial \nu} u_h\,ds \;-\; \int_{T_i} (e_h - \psi) \operatorname{div}(\sigma \operatorname{grad} u_h)\,dx \Big)\,. \end{aligned}$$

Da u_h auf jedem Dreieck T_i eine lineare Funktion ist, liefert die Produktregel

$$\operatorname{div}(\sigma \operatorname{grad} u_h) \;=\; \operatorname{grad}\sigma \cdot \operatorname{grad} u_h \;+\; \sigma \Delta u_h \;=\; \operatorname{grad}\sigma \cdot \operatorname{grad} u_h\,,$$

und somit folgt insgesamt

$$a_0 \|u - u_h\|^2_{H^1(\Omega)} \;\le\; \int_\Omega r_h (e_h - \psi)\,dx \;-\; \sum_{T_i \in \mathcal{T}} \int_{\partial T_i} (e_h - \psi)\, \sigma \frac{\partial}{\partial \nu} u_h\,ds \tag{96.5}$$

mit

$$r_h = f - c u_h + \operatorname{grad}\sigma \cdot \operatorname{grad} u_h\,. \tag{96.6}$$

Man beachte, daß r_h nichts anderes ist als das (in jedem einzelnen Teildreieck wohldefinierte und berechenbare) Residuum $f - L[u_h]$ der Differentialgleichung (96.1) in ihrer klassischen Form.

Wie bereits angekündigt, wird nun ein $\psi \in V_0^{\mathcal{T}}$ gesucht, für das die rechte Seite von (96.5) möglichst klein wird. Während sich im Eindimensionalen hierfür der interpolierende lineare Spline anbieten würde (vgl. Aufgabe 11), verbietet sich ein entsprechender Zugang in $\mathbb{R}^2$, da dort die Funktion $e_h \in H_0^1(\Omega)$ nicht stetig zu sein braucht (vgl. Beispiel 89.2). Statt dessen wird auf eine Konstruktion von Clément [19] zurückgegriffen (im wesentlichen eine lokale Mittelung von e_h), die ähnliche Approximationseigenschaften wie der interpolierende lineare Spline im $\mathbb{R}^1$ aufweist (vgl. Satz 45.2):

$$\|e_h - \psi\|_{\mathcal{L}^2(T_i)} \;\le\; c_1\, h_i^{\triangle}\, \|e_h\|_{H^1(T_i^{\triangle})}\,, \tag{96.7a}$$

$$\|e_h - \psi\|_{\mathcal{L}^2(\Gamma_j)} \;\le\; c_0 \sqrt{h_j'}\, \|e_h\|_{H^1(T_j')}\,. \tag{96.7b}$$

Dabei bezeichnen $T_i^{\triangle}$ und T_j' jeweils die Vereinigung aller angrenzenden Dreiecke von T_i bzw. Γ_j, $h_i^{\triangle}$ ist der Durchmesser von T_i und h_j' die Länge der Kante Γ_j. Die beiden Konstanten c_1 und c_0 hängen lediglich vom kleinsten Innenwinkel α_0 von $\mathcal{T}$ ab.

Die beiden Terme auf der rechten Seite von (96.5) behandeln wir nun separat weiter. Aus der Cauchy-Schwarz-Ungleichung und aus (96.7a) folgt

$$\begin{aligned}
\int_\Omega r_h(e_h - \psi)\,dx \;&=\; \sum_{i=1}^m \int_{T_i} r_h(e_h - \psi)\,dx \;\le\; \sum_{i=1}^m \|r_h\|_{\mathcal{L}^2(T_i)}\, \|e_h - \psi\|_{\mathcal{L}^2(T_i)} \\
&\le\; \sum_{i=1}^m c_1 h_i^{\triangle}\, \|r_h\|_{\mathcal{L}^2(T_i)}\, \|e_h\|_{H^1(T_i^{\triangle})} \\
&\le\; \Big(\sum_{i=1}^m c_1^2 (h_i^{\triangle})^2\, \|r_h\|^2_{\mathcal{L}^2(T_i)}\Big)^{1/2} \Big(\sum_{i=1}^m \|e_h\|^2_{H^1(T_i^{\triangle})}\Big)^{1/2},
\end{aligned}$$

und wegen der beschränkten Anzahl von Dreiecken, an die ein einzelnes Dreieck grenzen kann, ergibt sich

$$\int_\Omega r_h(e_h - \psi)\,dx \;\le\; C_1 \Big(\sum_{i=1}^m (h_i^{\triangle})^2\, \|r_h\|^2_{\mathcal{L}^2(T_i)}\Big)^{1/2} \|e_h\|_{H^1(\Omega)} \tag{96.8}$$

für eine gewisse Konstante $C_1 > 0$.

Bei dem zweiten Term auf der rechten Seite von (96.5) beachten wir zunächst, daß über jede innere Kante Γ_j der Triangulierung $\mathcal{T}$ genau zweimal integriert wird, und zwar einmal bei jedem der beiden angrenzenden Dreiecke. Dabei

hat die Normale ν in den beiden zusammengehörenden Termen jeweils unterschiedliches Vorzeichen, das heißt die entsprechenden Randintegrale werden jeweils voneinander abgezogen. Trotzdem ist diese Differenz nicht Null, da die Richtungsableitung von u_h senkrecht zur Kante beim Übergang von einem Dreieck ins Nachbardreieck springt; auch σ ist in der Regel nicht stetig. Daher definieren wir auf Γ_j den Sprung

$$\Big[\sigma \frac{\partial}{\partial \nu} u_h\Big]_{\Gamma_j}$$

von $\sigma \partial u_h/\partial \nu$ bei Überquerung von Γ_j in Richtung der Normalen; das Vorzeichen ist im weiteren ohne Belang und wir können uns daher eine genauere Spezifikation ersparen. Entsprechend verfahren wir für Randkanten von Ω, indem wir u_h außerhalb von Ω formal durch Null fortsetzen.

Auf der anderen Seite gehört $e_h - \psi$ zu $H^1(\Omega)$, hat also nach dem Spursatz 89.3 die gleichen Randwerte auf der Kante Γ_j, unabhängig davon, von welchem Dreieck aus die Spur gebildet wird. Somit erhalten wir für den zweiten Term auf der rechten Seite von (96.5) die obere Schranke

$$\sum_j \Big| \int_{\Gamma_j} (e_h - \psi) \Big[\sigma \frac{\partial}{\partial \nu} u_h\Big]_{\Gamma_j} ds \Big| .$$

Zur Abschätzung der einzelnen Integrale verwenden wir die Cauchy-Schwarz-Ungleichung und die Fehlerabschätzung (96.7b) für die Approximationsgüte von ψ auf den Kanten Γ_j:

$$\begin{aligned} \Big| \int_{\Gamma_j} (e_h - \psi) \Big[\sigma \frac{\partial}{\partial \nu} u_h\Big]_{\Gamma_j} ds \Big| \;&\le\; \|e_h - \psi\|_{\mathcal{L}^2(\Gamma_j)} \Big\| \Big[\sigma \frac{\partial}{\partial \nu} u_h\Big]_{\Gamma_j} \Big\|_{\mathcal{L}^2(\Gamma_j)} \\ &\le\; c_0 \sqrt{h'_j}\, \|e_h\|_{H^1(T'_j)} \Big\| \Big[\sigma \frac{\partial}{\partial \nu} u_h\Big]_{\Gamma_j} \Big\|_{\mathcal{L}^2(\Gamma_j)} ; \end{aligned}$$

hierbei bezeichnet T'_j wieder die Vereinigung aller an Γ_j angrenzenden Dreiecke aus $\mathcal{T}$. Durch Aufsummieren über alle Kanten Γ_j und eine anschließende Anwendung der Cauchy-Schwarz-Ungleichung für das euklidische Innenprodukt ergibt sich schließlich die obere Schranke

$$\begin{aligned} & c_0 \Big(C_2 \sum_{T \in \mathcal{T}} \|e_h\|^2_{H^1(T)} \Big)^{1/2} \Big(\sum_j h'_j \Big\| \Big[\sigma \frac{\partial}{\partial \nu} u_h\Big]_{\Gamma_j} \Big\|^2_{\mathcal{L}^2(\Gamma_j)} \Big)^{1/2} \\ &\qquad = \; c_0\, \|e_h\|_{H^1(\Omega)} \Big(C_2 \sum_j h'_j \Big\| \Big[\sigma \frac{\partial}{\partial \nu} u_h\Big]_{\Gamma_j} \Big\|^2_{\mathcal{L}^2(\Gamma_j)} \Big)^{1/2} \end{aligned}$$

für den zweiten Term in (96.5), wobei C_2 eine geeignete Schranke für die mögliche Anzahl an Kanten ist, die ein Dreieck berühren kann.

Zusammen mit (96.8) in (96.5) eingesetzt erhalten wir (mit einer neuen Konstante $C > 0$)

$$\begin{aligned}\|e_h\|^2_{H^1(\Omega)} \;\le\; & C\Big(\sum_{i=1}^{m}(h_i^{\triangle})^2\,\|r_h\|^2_{\mathcal{L}^2(T_i)}\Big)^{1/2}\|e_h\|_{H^1(\Omega)} \\ & + C\,\|e_h\|_{H^1(\Omega)}\Big(\sum_j h'_j\,\Big\|\Big[\sigma\,\frac{\partial}{\partial\nu}\,u_h\Big]_{\Gamma_j}\Big\|^2_{\mathcal{L}^2(\Gamma_j)}\Big)^{1/2},\end{aligned}$$

und nach Division durch $\|e_h\|_{H^1(\Omega)}$ ergibt sich schließlich die Fehlerschranke

$$\begin{aligned}\|e_h\|_{H^1(\Omega)} \;\le\; & C\Big(\sum_{i=1}^{m}(h_i^{\triangle})^2\,\|r_h\|^2_{\mathcal{L}^2(T_i)}\Big)^{1/2} \\ & + C\Big(\sum_j h'_j\,\Big\|\Big[\sigma\,\frac{\partial}{\partial\nu}\,u_h\Big]_{\Gamma_j}\Big\|^2_{\mathcal{L}^2(\Gamma_j)}\Big)^{1/2}\end{aligned} \tag{96.9}$$

mit einer universellen Konstanten $C > 0$, die nur von den Koeffizienten der Differentialgleichung, vom Gebiet Ω und vom kleinsten Innenwinkel der Triangulierung abhängt; r_h ist dabei das Residuum (96.6) der klassischen Form der Differentialgleichung, $h_i^{\triangle}$ der Durchmesser von Dreieck T_i und h'_j die Länge der Kante Γ_j.

Man beachte, daß die rechte Seite von (96.9) bis auf die unbekannte Konstante C eine berechenbare Größe darstellt, die zur Fehlerkontrolle immer mit berechnet werden kann. Die beiden Anteile dieses Fehlers lassen sich wie folgt interpretieren: Der erste Term mißt, wie gut u_h die klassische Form der Differentialgleichung auf jedem Teildreieck löst (also dort, wo u_h ohnehin glatt ist); der zweite Term überprüft, wie gut die Sprungbedingungen an den Kanten der Triangulierung approximiert werden (zur Motivation hierfür sei noch einmal auf Beispiel 89.6 verwiesen). Ist beispielsweise σ eine glatte Funktion, dann erwarten wir auch eine glatte Lösung u_h und in diesem Fall ist der zweite Term ein Maß für die Glattheit von u_h.

Anhand der Abschätzung (96.9) kann man auch entscheiden, in welchem Bereich von Ω die Triangulierung verfeinert werden muß, um eine gewünschte Genauigkeit zu erreichen. Aufgrund der einzelnen Beiträge in (96.9) bietet es sich etwa an, diejenigen Dreiecke zu verfeinern, in denen das gewichtete Residuum $h_i^{\triangle} r_h$ am größten ist oder auf deren Kanten der gewichtete Sprung $\sqrt{h'_j}\,(\sigma\partial u_h/\partial\nu)$ am größten ist.

Aufgaben

1. Seien $\sigma, u, v \in C^2(\overline{\Omega})$. Beweisen Sie die verallgemeinerte Greensche Formel

$$\int_\Omega v \operatorname{div}(\sigma \operatorname{grad} u)\, dx \;=\; \int_\Gamma v\sigma \frac{\partial u}{\partial \nu}\, ds \;-\; \int_\Omega \sigma \operatorname{grad} v \cdot \operatorname{grad} u\, dx\,.$$

Hinweis: Wenden Sie die Greensche Formel (89.1) auf v und σu bzw. σ und uv an.

2. Sei $\Omega \subset \mathbb{C}$ ein Gebiet, $f : \Omega \to \mathbb{C}$ eine holomorphe Funktion, und die reellwertige Funktion u sei durch

$$u : (\xi, \eta) \mapsto u(\xi, \eta) = \operatorname{Re} f(\xi + \mathrm{i}\eta)\,, \qquad \xi + \mathrm{i}\eta \in \Omega\,,$$

definiert. Zeigen Sie, daß dann $|\operatorname{grad} u(\xi, \eta)|^2 = |f'(\xi + \mathrm{i}\eta)|^2$ gilt und u die Laplace-Gleichung $\Delta u = 0$ löst. Folgern Sie, daß $u_k(x) = r^k \cos k\theta$ und $v_k(x) = r^k \sin k\theta$ in der Polarkoordinatendarstellung $x = (r\cos\theta, r\sin\theta)$ mit $r \geq 0$ und $0 \leq \theta \leq 2\pi$ ebenfalls Lösungen der Laplace-Gleichung sind.

3. Gegeben sei die reellwertige Funktion $g(\theta) \sim \sum_{-\infty}^{\infty} \alpha_k e^{\mathrm{i}k\theta} \in H^s_\pi(0, 2\pi)$. Zeigen Sie, daß die Funktion

$$u(x) \;=\; \alpha_0 \;+\; \sum_{k=1}^{\infty} 2r^k (\operatorname{Re}\alpha_k \cos k\theta \;-\; \operatorname{Im}\alpha_k \sin k\theta)\,, \qquad \begin{array}{c} x = (r\cos\theta, r\sin\theta)\,, \\ r \geq 0,\ 0 \leq \theta < 2\pi\,, \end{array}$$

im Innern des Einheitskreises Ω eine klassische Lösung der Laplace-Gleichung ist, die auf dem Einheitskreisrand mit g übereinstimmt, d. h. $u(\cos\theta, \sin\theta) = g(\theta)$, $0 \leq \theta < 2\pi$. Beweisen Sie, daß

$$u \in H^1(\Omega) \qquad \text{genau dann, wenn} \qquad g \in H^{1/2}_\pi(0, 2\pi)\,.$$

Hinweis: Aufgabe 2

4. Betrachten Sie die Bilinearform

$$a(u, v) \;=\; \int_\Omega \sigma \operatorname{grad} u \cdot \operatorname{grad} v\, dx \;+\; \int_\Omega cuv\, dx\,,$$

wobei σ und c den Voraussetzungen von Satz 89.5 genügen sollen. (Im Fall von Neumann-Randbedingungen sei zusätzlich $c(x) \geq c_0 > 0$ vorausgesetzt.) Ferner bezeichne $\ell(v)$ die rechte Seite von (89.7) (Dirichlet-Problem) bzw. (89.13) (Neumann-Problem). Schließlich sei u die schwache Lösung des zugehörigen Dirichlet- bzw. Neumann-Randwertproblems. Weisen Sie für $\Phi[u] = \frac{1}{2}a(u, u) - \ell(u)$ die Identität

$$\Phi[v] - \Phi[u] \;=\; \frac{1}{2}\, a(v - u, v - u)$$

für alle $v \in H^1_0(\Omega)$ (Dirichlet) bzw. $v \in H^1(\Omega)$ (Neumann) nach. Folgern Sie, daß u die eindeutige Minimalstelle von Φ ist.

5. $\mathcal{T}_h$ bezeichne eine reguläre Triangulierung eines Gebiets Ω mit maximaler Kantenlänge $h > 0$. Ferner gelte $|T_j| \geq \alpha h^2$ für alle $T_j \in \mathcal{T}_h$ mit einer Konstanten $\alpha > 0$; dabei bezeichne

$|T_j|$ den Flächeninhalt des Dreiecks T_j. Schließlich sei v eine stückweise lineare Funktion über $\mathcal{T}_h$ und der Vektor $\boldsymbol{v} = [v(x_i)]$ enthalte die Werte von v an den Knoten x_i von $\mathcal{T}_h$. Zeigen Sie, daß dann Konstanten $c, C > 0$ existieren mit

$$c\|v\|_{\mathcal{L}^2(\Omega)} \le h\|\boldsymbol{v}\|_2 \le C\|v\|_{\mathcal{L}^2(\Omega)}\,.$$

6. Es sei $c : [0,1] \to \mathbb{R}$ eine beschränkte nichtnegative Funktion, f sei aus $\mathcal{L}^2(0,1)$ und $\sigma \in C^1[0,1]$ sei positiv über $[0,1]$. Schließlich sei u die schwache Lösung des Randwertproblems

$$-(\sigma u')' + cu = f \quad \text{in } (0,1)\,, \qquad u(0) = u(1) = 0\,.$$

(a) Zeigen Sie, daß $w = (cu - f - \sigma' u')/\sigma$ zu $\mathcal{L}^2(0,1)$ gehört.
(b) Betrachten Sie $W(x) = \int_0^x w(t)dt \in H^1(0,1)$ und setzen Sie $\omega = \int_0^1 W(t)dt$. Beweisen Sie, daß

$$\langle\, W, v\,\rangle_{\mathcal{L}^2(0,1)} = \langle\, u' + \omega, v\,\rangle_{\mathcal{L}^2(0,1)} \qquad \text{für alle } v \in \mathcal{L}^2(0,1)\,.$$

Nutzen Sie dabei aus, daß die Funktion $V(x) = \int_0^x v(t)dt - x\int_0^1 v(t)dt$ zu $H_0^1(0,1)$ gehört.
(c) Setzen Sie $v = W - u' - \omega$ in (b) und folgern Sie, daß $u \in H^2(0,1) \cap H_0^1(0,1)$.

7. (a) Sei V_h ein endlichdimensionaler Teilraum des reellen Vektorraums V und $a : V \times V \to \mathbb{R}$ eine symmetrische stetige und V-elliptische Bilinearform.
Zeigen Sie, daß zu jedem $v \in V$ eine eindeutig bestimmte *Ritz-Projektion* $R_h v = v_h \in V_h$ existiert, welche das Variationsproblem

$$a(v_h, w) = a(v, w) \qquad \text{für alle } w \in V_h \tag{96.10}$$

löst. Weisen Sie nach, daß $R_h : V \to V_h$ tatsächlich eine Projektion ist, die bezüglich des Innenprodukts

$$\langle\, u, v\,\rangle = a(u, v)$$

orthogonal ist.
(b) Sei nun speziell $a(\,\cdot\,,\cdot\,)$ die Bilinearform aus (90.1) und darüber hinaus seien die Voraussetzungen von Satz 92.2 erfüllt. Betrachten Sie $V = H_0^1(\Omega)$ und den Teilraum $V_h = V_0^{\mathcal{T}}$ der stückweise linearen finiten Elemente über einer festen Triangulierung $\mathcal{T}$ von Ω mit homogenen Randwerten. Zeigen Sie, daß für $v \in H^2(\Omega)$ die Ritz-Projektion v_h aus $V_0^{\mathcal{T}}$ der Fehlerabschätzung

$$\|v_h - v\|_{\mathcal{L}^2(\Omega)} \le c_0 h^2 \|v\|_{H^2(\Omega)}$$

mit der Konstanten c_0 aus Satz 92.3 genügt.

8. Lösen Sie das Neumann-Randwertproblem (71.11) für die Laplace-Gleichung mit der Methode der finiten Elemente. Triangulieren Sie dazu das Quadrat $(-a,a) \times (-a,a)$ wie in Abbildung 93.3 und verwenden Sie den zugehörigen Ansatzraum $V^{\mathcal{T}}$, vgl. Bemerkung 90.7.
(a) Berechnen Sie die Steifigkeitsmatrix A und geben Sie Matrizen D und T an, so daß wie in Beispiel 94.2 eine Darstellung

$$A = D \otimes T + T \otimes D$$

gilt.

(b) Bestimmen Sie die rechte Seite b des linearen Gleichungssystems und zeigen Sie, daß b im Bild von A liegt. Ergänzen Sie das System um eine zusätzliche Gleichung, so daß die Koeffizientenmatrix invertierbar wird.

9. (a) Rechnen Sie für $A_1 \in \mathbb{K}^{m\times n}$, $A_2 \in \mathbb{K}^{n\times p}$, $B_1 \in \mathbb{K}^{\mu\times\nu}$ und $B_2 \in \mathbb{K}^{\nu\times\rho}$ nach, daß $(A_1 \otimes B_1)(A_2 \otimes B_2) = (A_1A_2 \otimes B_1B_2)$.
(b) Seien $A \in \mathbb{K}^{m\times n}$, $B \in \mathbb{K}^{p\times q}$, $x \in \mathbb{K}^{nq}$ und $y = (A \otimes B)x \in \mathbb{K}^{mp}$. Zeigen Sie, daß $Y = AXB^T$, wenn $X \in \mathbb{K}^{n\times q}$ bzw. $Y \in \mathbb{K}^{m\times p}$ durch zeilenweise Anordnung von x bzw. y gebildet werden, also

$$X = \begin{bmatrix} x_1 & x_2 & \dots & x_q \\ x_{q+1} & x_{q+2} & \dots & x_{2q} \\ \vdots & \vdots & & \vdots \\ & & \dots & x_{nq} \end{bmatrix} \qquad \text{bzw.} \qquad Y = \begin{bmatrix} y_1 & y_2 & \dots & y_p \\ y_{p+1} & y_{p+2} & \dots & y_{2p} \\ \vdots & \vdots & & \vdots \\ & & \dots & y_{mp} \end{bmatrix}.$$

(c) Seien $\{x_i\}_{i=1}^{\nu}$ und $\{z_j\}_{j=1}^{\nu}$ zwei Basen des $\mathbb{K}^\nu$. Weisen Sie nach, daß $\{x_i \otimes z_j\}_{i,j=1}^{\nu}$ eine Basis des $\mathbb{K}^{\nu^2}$ bildet.

10. Es seien $u^{(l)} \in V_h$ die Iterierten aus Algorithmus 95.1 und u_h die exakte Galerkin-Näherung. Ferner seien mit $\boldsymbol{e}^{(l)}$ die zu den Fehlern $u^{(l)} - u_h$ gehörigen Koeffizientenvektoren bezüglich der nodalen Basis bezeichnet.
(a) Zeigen Sie, daß diese Vektoren der Rekursion

$$\boldsymbol{e}^{(l+1)} \;=\; (I - R^* A_H^{-1} R A_h)(I - \omega A_h)^\kappa \boldsymbol{e}^{(l)}\,, \qquad l = 0, 1, 2, \dots\,,$$

genügen.
(b) Rechnen Sie nach, daß die Iterationsmatrix aus Teil (a) für das eindimensionale Modellproblem aus Beispiel 90.3 mit Gitterweite $h = 1/(2n)$ neben dem Eigenwert Null lediglich noch die Eigenwerte

$$\lambda_j \;=\; \cos^2\theta_j \sin^2\theta_j \big(\cos^{2\kappa-2}\theta_j + \sin^{2\kappa-2}\theta_j\big)\,, \qquad j = 1, 2, \dots, n\,,$$

mit $\theta_j = j\pi h/2$ besitzt. Folgern Sie, daß der asymptotische Konvergenzfaktor des Zweigitterverfahrens in diesem Beispiel durch $1/(\kappa+1)$ abgeschätzt werden kann.

11. Übertragen Sie die Konstruktion des Fehlerschätzers aus Abschnitt 96 auf das eindimensionale Randwertproblem $-(\sigma u')' + cu = f$ über $(0,1)$ mit $u(0) = u(1) = 0$. Die Approximation u_h bezeichne die Lösung des Finite-Elemente-Verfahrens aus dem Ansatzraum der linearen Splines über einem Gitter $\triangle = \{0 = x_0 < x_1 < \ldots < x_n = 1\}$ mit Teilintervallen $\mathcal{I}_k = (x_{k-1}, x_k)$ der Länge h_k. Leiten Sie für u_h die Fehlerabschätzung

$$\|u - u_h\|^2_{H^1(\Omega)} \;\le\; \frac{1}{2a_0^2} \sum_{k=1}^{n} h_k^2 \,\|f - cu_h + \sigma' u_h'\|^2_{L^2(\mathcal{I}_k)}$$

her, wobei a_0 die Konstante aus der Elliptizitätsbedingung ist, vgl. Proposition 90.2.

XVII Parabolische Differentialgleichungen

Als nächstes kommen wir zu zeitabhängigen Diffusionsprozessen. Wir untersuchen numerische Verfahren, bei denen die Differentialgleichung mittels einer Ortsdiskretisierung in ein System gewöhnlicher Differentialgleichungen bezüglich der Zeit transformiert wird (in Form eines Anfangswertproblems). Die Ortsdiskretisierung kann wahlweise mit einem Differenzenverfahren oder dem Galerkin-Verfahren erfolgen; wie bei den elliptischen Differentialgleichungen beschränken wir uns auf letzteres. Das resultierende Anfangswertproblem muß schließlich mit einem A-stabilen Runge-Kutta-Verfahren gelöst werden.

Zur Vertiefung der hier behandelten Theorie sei auf das Buch von Thomée [102] verwiesen. Daneben seien noch die Bücher von Großmann und Roos [40] und Quarteroni und Valli [87] empfohlen, die beide auch auf Differenzenverfahren eingehen.

97 Schwache Lösungen und Regularität

Sei $\Omega \subset \mathbb{R}^2$ ein polygonales Gebiet mit Rand $\Gamma = \partial\Omega$ und

$$L[u] = -\operatorname{div}(\sigma \operatorname{grad} u) + cu \tag{97.1}$$

wie in Kapitel XVI ein *elliptischer Differentialoperator* in Ω, das heißt σ und c sind nichtnegative beschränkte Funktionen in Ω mit

$$\sigma(x) \geq \sigma_0 > 0\,, \qquad x \in \Omega\,.$$

Unter diesen Voraussetzungen nennt man

$$\begin{aligned} u_t + L[u] &= f && \text{für } x \in \Omega \text{ und } t \geq 0\,, \\ u(x,t) &= 0 && \text{für } x \in \Gamma \text{ und } t \geq 0\,, \\ u(x,0) &= u^\circ(x) && \text{für } x \in \Omega\,, \end{aligned} \tag{97.2}$$

ein *parabolisches Anfangsrandwertproblem*, u° ist die Anfangsvorgabe zum Zeitpunkt $t = 0$ und $u = 0$ die (homogene) Dirichlet-Randbedingung auf Γ für alle

Zeiten $t \geq 0$. Letztere kann auch durch eine inhomogene (auch zeitabhängige) Dirichlet- oder Neumann-Randbedingung ersetzt werden. Unter einer *klassischen Lösung* dieses Anfangsrandwertproblems wird eine Funktion $u = u(x,t)$ verstanden, die in $\Omega \times \mathbb{R}_0^+$ zweimal bezüglich der Ortsvariablen und einmal bezüglich der Zeit differenzierbar ist und die drei Gleichungen aus (97.2) erfüllt.

Der Quellterm f ist in der Regel eine Funktion von Ort *und* Zeit, $f = f(x,t)$, während wir hier und im weiteren voraussetzen wollen, daß die Koeffizientenfunktionen σ und c nur von der Ortsvariablen abhängen. Ist $f = f(x)$ ebenfalls von der Zeit unabhängig und $v = v(x)$ eine Lösung der elliptischen Differentialgleichung

$$L[v] = f \quad \text{in } \Omega\,, \qquad v|_\Gamma = 0\,,$$

so ist offensichtlich $u(x,t) = v(x)$ eine zeitunabhängige (*stationäre*) Lösung des parabolischen Anfangsrandwertproblems (97.2) mit $u^\circ(x) = v(x)$. Da wir aus Abschnitt 89 wissen, daß die Lösungen elliptischer Gleichungen im allgemeinen nicht die oben geforderten Differenzierbarkeitseigenschaften aufweisen, empfiehlt es sich, auch für parabolische Gleichungen einen abgeschwächten Lösungsbegriff einzuführen. Dabei erweist es sich als vorteilhaft, der Zeitvariablen eine bevorzugte Stellung einzuräumen: Ist $v : \Omega \times \mathbb{R}_0^+$ eine Funktion des Ortes $x \in \Omega$ und der Zeit $t > 0$, so werden wir oftmals das x-Argument weglassen und einfach $v(t)$ für $v(\,\cdot\,,t)$ schreiben, wenn die Lösung zu einem festen Zeitpunkt „als Ganzes“ gemeint ist.[1] Zudem verwenden wir gelegentlich die Notation v' für die Zeitableitung v_t.

Sei nun ein $t > 0$ fest gewählt und u eine klassische Lösung der Differentialgleichung (97.2), dann ergibt sich durch Multiplikation mit $w \in H_0^1(\Omega)$ und Integration über Ω wie in Abschnitt 89 die Integralidentität

$$\int_\Omega u'(t) w \, dx + \int_\Omega \sigma \operatorname{grad} u(t) \cdot \operatorname{grad} w \, dx + \int_\Omega c u(t) w \, dx = \int_\Omega f(t) w \, dx\,.$$

Mit der Bilinearform $a(\,\cdot\,,\,\cdot\,)$ aus (90.1) kann dies auch in der Form

$$\langle\, u'(t), w \,\rangle_{\mathcal{L}^2(\Omega)} + a(u(t), w) = \langle\, f(t), w \,\rangle_{\mathcal{L}^2(\Omega)} \tag{97.3}$$

geschrieben werden. An dieser Stelle sei daran erinnert, vgl. Proposition 90.2, daß aufgrund der Voraussetzungen an den Differentialoperator L die Bilinearform a $H_0^1(\Omega)$-elliptisch ist, d. h. es gibt eine Konstante $a_0 > 0$, so daß

$$a(w,w) \geq a_0 \|w\|^2_{H^1(\Omega)} \qquad \text{für alle } w \in H_0^1(\Omega)\,. \tag{97.4}$$

Nun können wir schwache Lösungen von (97.2) definieren.

[1] Streng genommen interpretieren wir v als Banachraum-wertige Abbildung $v : \mathbb{R}_0^+ \to \mathcal{L}^2(\Omega)$.

Definition 97.1. Eine Funktion $u : \Omega \times \mathbb{R}_0^+ \to \mathbb{R}$ mit $u(x,t) = 0$ für $x \in \Gamma$ und $t \geq 0$ sowie $\|u(t) - u^\circ\|_{\mathcal{L}^2(\Omega)} \to 0$ für $t \to 0$ heißt *schwache Lösung* des Anfangsrandwertproblems (97.2), falls $u(t)$ für fast alle $t > 0$ zu $H^1(\Omega)$ gehört und für solche t die schwache Form (97.3) der Differentialgleichung für alle $w \in H_0^1(\Omega)$ erfüllt.

Man kann zeigen, daß unter den getroffenen Voraussetzungen an σ und c für jedes $f \in \mathcal{L}^2(\Omega \times \mathbb{R}_0^+)$ eine schwache Lösung in einem geeigneten Funktionenraum existiert, vgl. etwa Quarteroni und Valli [87]. Unter den Voraussetzungen des Regularitätssatzes 92.2 für den elliptischen Differentialoperator hat diese Lösung die folgenden Glattheitseigenschaften, vgl. [87, S. 370 ff]:

Satz 97.2. *Ist Ω ein konvexes polygonales Gebiet, c eine nichtnegative beschränkte Funktion und σ stetig differenzierbar in $\overline{\Omega}$ mit $\sigma(x) \geq \sigma_0$, dann hat das Anfangsrandwertproblem* (97.2) *für jedes $f \in \mathcal{L}^2(\Omega \times \mathbb{R}_0^+)$ und $u^\circ \in H_0^1(\Omega)$ eine schwache Lösung u mit $u(t) \in H_0^1(\Omega)$ für alle $t \geq 0$. Zudem gehört $u(t)$ für fast alle $t > 0$ zu $H^2(\Omega)$ und $u'(t)$ zu $\mathcal{L}^2(\Omega)$.*

Damit eine klassische Lösung existiert, müssen die Voraussetzungen noch weiter verschärft werden. Insbesondere muß f stetig sein. Darüber hinaus muß die Anfangsvorgabe u° zweimal stetig differenzierbar sein und die naheliegenden *Kompatibilitätsbedingungen*

$$u^\circ|_\Gamma = 0 \qquad \text{und} \qquad L[u^\circ](x) = f(x,0) \;\text{ für } x \in \Gamma$$

erfüllen, die sich aus den aufeinandertreffenden Anfangs- und Randvorgaben bei gleichzeitiger Gültigkeit der Differentialgleichung (97.2) ergeben.

Es sei jedoch angemerkt, daß unter den eingangs genannten Voraussetzungen an die Koeffizienten σ und c eine schwache Lösung der *homogenen* parabolischen Anfangsrandwertaufgabe (also für $f = 0$) zu jedem positiven Zeitpunkt $t > 0$ unendlich oft nach t differenzierbar ist. Wir erläutern diese wichtige Eigenschaft anhand der eindimensionalen Wärmeleitungsgleichung aus Beispiel 69.2,

$$\begin{aligned} u_t = u_{xx}\,, \qquad & u(0) = u^\circ\,, \qquad u(0,t) = u(\pi,t) = 0\,, \\ & 0 < x < \pi\,, \qquad t \geq 0\,. \end{aligned} \tag{97.5}$$

Hier sei $u^\circ \in \mathcal{L}^2(0,\pi)$ eine beliebige Anfangsvorgabe, die im Intervall $(0,\pi)$ wie in Abschnitt 55.1 in eine Sinusreihe

$$u^\circ(x) \;=\; \sum_{j=1}^{\infty} b_j \sin jx\,, \qquad 0 \leq x \leq \pi\,,$$

entwickelt werden kann. Nach Aufgabe IX.18 sind die Koeffizienten b_j quadratisch summierbar mit

$$\|u^\circ\|^2_{\mathcal{L}^2(0,\pi)} \;=\; \frac{\pi}{2}\sum_{j=1}^{\infty} b_j^2\,.$$

In Beispiel 69.2 haben wir bereits die formale Lösung

$$u(x,t) \;=\; \sum_{j=1}^{\infty} b_j e^{-j^2 t}\sin jx \tag{97.6}$$

dieses Anfangsrandwertproblems bestimmt. Genauso ergibt sich, zunächst formal,

$$\frac{d^\nu}{dt^\nu}u(x,t) \;=\; \sum_{j=1}^{\infty}(-1)^\nu b_j\, j^{2\nu} e^{-j^2 t}\sin jx\,, \qquad \nu\in\mathbb{N}\,,$$

wobei die Reihen aufgrund des Exponentialterms für $t>0$ gleichmäßig konvergieren und somit Summation und Differentiation jeweils vertauschbar sind. Wegen

$$\max_{j\ge 0} j^{2\nu}e^{-j^2 t} \;=\; (\nu/e)^\nu t^{-\nu} \qquad \text{für } t>0$$

erhalten wir für die $\mathcal{L}^2$-Norm dieser Zeitableitungen zudem die Schranke

$$\Big\|\frac{d^\nu}{dt^\nu}u(t)\Big\|_{\mathcal{L}^2(0,\pi)} \;\le\; (\nu/e)^\nu t^{-\nu}\,\|u^\circ\|_{\mathcal{L}^2(0,\pi)}\,, \qquad t>0\,,\ \nu\in\mathbb{N}\,,$$

d. h. alle Zeitableitungen haben ein algebraisches Abklingverhalten für $t\to\infty$.

Während entsprechende Überlegungen auch auf allgemeine elliptische Differentialoperatoren L übertragen werden können, sind Lösungen der homogenen Wärmeleitungsgleichung zusätzlich noch unendlich oft bezüglich der Ortsvariablen differenzierbar, wie man leicht in der gleichen Weise überprüft. Diese Eigenschaft rührt allerdings wesentlich daher, daß die zugehörigen Koeffizientenfunktionen σ und c bei der Wärmeleitungsgleichung konstant und somit selbst unendlich oft differenzierbar sind. Die Glattheit der Lösung u von (97.3) bezüglich der Ortsvariablen wird wesentlich durch die Glattheit der Koeffizientenfunktionen gesteuert.

Man beachte, daß diese Ergebnisse im Gegensatz zu Satz 97.2 lediglich die schwache Voraussetzung $u^\circ\in\mathcal{L}^2(\Omega)$ benötigen.

98 Die Linienmethode

Aufgrund der Verwandtschaft mit dem Lösungsbegriff bei elliptischen Differentialgleichungen bietet sich zur Ortsdiskretisierung eines parabolischen Anfangsrandwertproblems ein Ansatz auf der Basis des Galerkin-Verfahrens an. Dazu wird wie in Abschnitt 90 ein endlichdimensionaler Ansatzraum $V_h \subset H_0^1(\Omega)$ gewählt und für jedes $t > 0$ eine Approximation $u_h(t) \in V_h$ bestimmt, welche das *Variationsproblem*

$$\langle\, u_h'(t), w \,\rangle_{\mathcal{L}^2(\Omega)} + a(u_h(t), w) = \langle\, f(t), w \,\rangle_{\mathcal{L}^2(\Omega)} \qquad \text{für alle } w \in V_h \tag{98.1}$$

löst. Wie im vorherigen Kapitel beschränken wir uns im weiteren durchweg auf die Teilräume $V_h = V_0^{\mathcal{T}} = V^{\mathcal{T}} \cap H_0^1(\Omega)$ der stückweise linearen finiten Elemente zu einer Triangulierung $\mathcal{T}$ von Ω mit homogenen Randwerten. In diesem Fall gibt der Index h die maximale Kantenlänge der Triangulierung an.

Bei der obigen Vorgehensweise handelt es sich um eine *Semidiskretisierung* bezüglich des Orts, denn die Zeitvariable wird nach wie vor als kontinuierliche Größe betrachtet. Wie noch erläutert wird, ist die Lösung u_h von (98.1) wohldefiniert; sie genügt darüber hinaus einer Fehlerabschätzung, wie wir sie von der Galerkin-Approximation elliptischer Differentialgleichungen her kennen. In dem Beweis dieser Fehlerabschätzung greifen wir auf die in Aufgabe XVI.7 eingeführte *Ritz-Projektion* von $H_0^1(\Omega)$ nach V_h zurück: Ist $v \in H_0^1(\Omega)$, so wird die Ritz-Projektion $v_h \in V_h$ von v durch das Variationsproblem

$$a(v_h, w) = a(v, w) \qquad \text{für alle } w \in V_h$$

definiert.

Satz 98.1. *Der Differentialoperator L aus* (97.1) *erfülle die Voraussetzungen aus Satz* 97.2. *Ferner sei $u^\circ \in H_0^1(\Omega) \cap H^2(\Omega)$ und die Lösung u des Anfangsrandwertproblems* (97.2) *habe eine (Zeit-)Ableitung u' mit $u'(t) \in H^2(\Omega)$ für fast alle $t \geq 0$. Ist u_h die Lösung von* (98.1)*, so gilt für jedes $t \geq 0$ die Fehlerabschätzung*

$$\begin{aligned}\|u_h(t) - u(t)\|_{\mathcal{L}^2(\Omega)} \leq{}& \|u_h(0) - u^\circ\|_{\mathcal{L}^2(\Omega)} \\ &+ 2c_0 h^2 \big(\|u^\circ\|_{H^2(\Omega)} + \int_0^t \|u'(\tau)\|_{H^2(\Omega)}\, d\tau \big)\end{aligned}$$

mit der Konstanten c_0 aus Satz 92.3.

Beweis. Für beliebiges $t \geq 0$ ist die Ritz-Projektion $v_h(t) \in V_0^{\mathcal{T}}$ von $u(t)$ durch das Variationsproblem

$$a(v_h(t), w) = a(u(t), w) = \int_\Omega \big(f(t) - u'(t)\big) w \, dx \quad \text{für alle } w \in V_0^{\mathcal{T}} \tag{98.2}$$

definiert. Durch Differentiation von (98.2) nach der Zeit erkennt man, daß v_h nach t differenzierbar ist und v_h' die Gleichung

$$a(v_h', w) \;=\; a(u', w) \qquad \text{für alle } w \in V_0^{\mathcal{T}} \tag{98.3}$$

erfüllt; v_h' ist also die Ritz-Projektion von u' aus $V_0^{\mathcal{T}}$.

Wir betrachten nun den Abstand $d_h = v_h - u_h$. Nach (98.2) und (98.1) erfüllt d_h das Variationsproblem

$$\begin{aligned} \langle\, d_h', w \,\rangle_{\mathcal{L}^2(\Omega)} + a(d_h, w) \;&=\; \langle\, v_h' - u_h', w \,\rangle_{\mathcal{L}^2(\Omega)} + a(v_h, w) - a(u_h, w) \\ &=\; \langle\, v_h' - u_h' + f - u' - (f - u_h'), w \,\rangle_{\mathcal{L}^2(\Omega)} \;=\; \langle\, v_h' - u', w \,\rangle_{\mathcal{L}^2(\Omega)} \end{aligned}$$

für alle $w \in V_0^{\mathcal{T}}$. Da d_h seinerseits zu $V_0^{\mathcal{T}}$ gehört, kann hierbei $w = d_h$ gewählt werden und wir erhalten

$$\frac{1}{2}\frac{d}{dt}\,\|d_h\|_{\mathcal{L}^2(\Omega)}^2 + a(d_h, d_h) \;=\; \langle\, v_h' - u', d_h \,\rangle_{\mathcal{L}^2(\Omega)}\,.$$

Aufgrund der Elliptizität (97.4) der Bilinearform a führt dies mit der Cauchy-Schwarz-Ungleichung zu der Abschätzung

$$\frac{1}{2}\frac{d}{dt}\,\|d_h\|_{\mathcal{L}^2(\Omega)}^2 \;\le\; \|v_h' - u'\|_{\mathcal{L}^2(\Omega)}\,\|d_h\|_{\mathcal{L}^2(\Omega)}\,. \tag{98.4}$$

Bis auf zeitliche Nullmengen, in denen d_h verschwindet, kann der erste Term durch

$$\frac{1}{2}\frac{d}{dt}\,\|d_h\|_{\mathcal{L}^2(\Omega)}^2 = \|d_h\|_{\mathcal{L}^2(\Omega)}\frac{d}{dt}\,\|d_h\|_{\mathcal{L}^2(\Omega)}$$

ersetzt werden, so daß die Ungleichung (98.4) durch $\|d_h\|_{\mathcal{L}^2(\Omega)}$ gekürzt werden kann. Dies führt auf

$$\frac{d}{dt}\,\|d_h\|_{\mathcal{L}^2(\Omega)} \;\le\; \|v_h' - u'\|_{\mathcal{L}^2(\Omega)}$$

und Integration von 0 bis t über die Zeit ergibt

$$\|d_h(t)\|_{\mathcal{L}^2(\Omega)} \;\le\; \|d_h(0)\|_{\mathcal{L}^2(\Omega)} + \int_0^t \|v_h'(\tau) - u'(\tau)\|_{\mathcal{L}^2(\Omega)}\,d\tau\,. \tag{98.5}$$

Aus der Darstellung $u - u_h = u - v_h + d_h$ erhalten wir hieraus

$$\begin{aligned} \|u - u_h\|_{\mathcal{L}^2(\Omega)} \;&\le\; \|u - v_h\|_{\mathcal{L}^2(\Omega)} + \|d_h\|_{\mathcal{L}^2(\Omega)} \\ &\le\; \|u - v_h\|_{\mathcal{L}^2(\Omega)} + \|d_h(0)\|_{\mathcal{L}^2(\Omega)} + \int_0^t \|v_h'(\tau) - u'(\tau)\|_{\mathcal{L}^2(\Omega)}\,d\tau\,. \end{aligned} \tag{98.6}$$

Der Fehler der Ritz-Projektion ist aufgrund der Regularitätsannahmen für den Differentialoperator L nach Aufgabe XVI.7 (b) für fast alle $t > 0$ durch

$$\|u - v_h\|_{\mathcal{L}^2(\Omega)} \le c_0 h^2 \|u\|_{H^2(\Omega)} \tag{98.7}$$

beschränkt. Entsprechend folgt aus (98.3) für den Fehler $u' - v_h'$ die Abschätzung

$$\|u' - v_h'\|_{\mathcal{L}^2(\Omega)} \le c_0 h^2 \|u'\|_{H^2(\Omega)} .$$

In (98.6) eingesetzt, ergibt dies

$$\|u - u_h\|_{\mathcal{L}^2(\Omega)} \le \|d_h(0)\|_{\mathcal{L}^2(\Omega)} + c_0 h^2 \Big(\|u\|_{H^2(\Omega)} + \int_0^t \|u'(\tau)\|_{H^2(\Omega)} \, d\tau \Big) .$$

Hierbei ist wegen (98.7)

$$\begin{aligned} \|d_h(0)\|_{\mathcal{L}^2(\Omega)} &\le \|u_h(0) - u^\circ\|_{\mathcal{L}^2(\Omega)} + \|u^\circ - v_h(0)\|_{\mathcal{L}^2(\Omega)} \\ &\le \|u_h(0) - u^\circ\|_{\mathcal{L}^2(\Omega)} + c_0 h^2 \|u^\circ\|_{H^2(\Omega)} , \end{aligned}$$

so daß wir insgesamt die folgende Fehlerabschätzung erhalten:

$$\begin{aligned} \|u_h(t) - u(t)\|_{\mathcal{L}^2(\Omega)} \le \; & \|u_h(0) - u^\circ\|_{\mathcal{L}^2(\Omega)} + \\ & c_0 h^2 \big(\|u^\circ\|_{H^2(\Omega)} + \|u(t)\|_{H^2(\Omega)} + \int_0^t \|u'(\tau)\|_{H^2(\Omega)} \, d\tau \big) . \end{aligned}$$

Um den Beweis abzuschließen verwenden wir jetzt nur noch die Darstellung

$$u(t) = u^\circ + \int_0^t u'(\tau) \, d\tau ,$$

aus der dann mit der Dreiecksungleichung die benötigte Abschätzung für die H^2-Norm von $u(t)$ folgt. □

Um die Galerkin-Approximation $u_h \in V_0^{\mathcal{T}}$ aus (98.1) berechnen zu können, stellen wir u_h in der nodalen Basis $\{\Lambda_1, \ldots, \Lambda_n\}$ dar, also als

$$u_h(t) = \sum_{j=1}^{n} \eta_j(t) \Lambda_j \tag{98.8}$$

mit zeitabhängigen Koeffizienten $\eta_j(t) \in \mathbb{R}$, $j = 1, \ldots, n$. Das Variationsproblem (98.1) führt dann mit $w = \Lambda_k$, $k = 1, \ldots, n$, auf das Differentialgleichungssystem

$$\sum_{j=1}^{n} \eta_j'(t) \langle \Lambda_j, \Lambda_k \rangle_{\mathcal{L}^2(\Omega)} + \sum_{j=1}^{n} \eta_j(t) a(\Lambda_j, \Lambda_k) = \langle f(t), \Lambda_k \rangle_{\mathcal{L}^2(\Omega)} , \quad k = 1, \ldots, n .$$

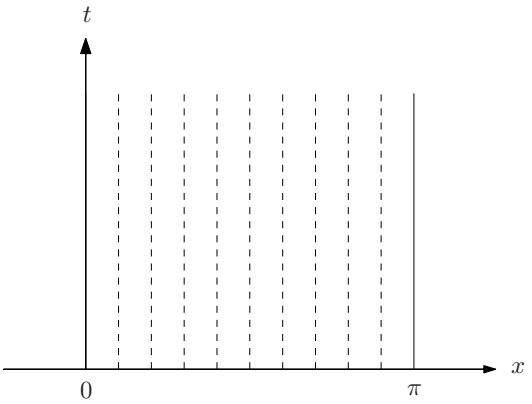


Abb. 98.1: Linienmethode

Mit den Vektoren

$$y(t) = [\,\eta_j(t)\,]_j \in \mathbb{R}^n \qquad \text{und} \qquad b(t) = [\,\langle\, f(t), \Lambda_k \,\rangle_{\mathcal{L}^2(\Omega)}\,]_k \in \mathbb{R}^n$$

läßt sich dieses Differentialgleichungssystem als

$$G y' = b - A y\,, \qquad y(0) = [\,\eta_j(0)\,]_j \in \mathbb{R}^n\,, \tag{98.9}$$

schreiben, wobei $G \in \mathbb{R}^{n\times n}$ die Gramsche Matrix der Hutfunktionen in $\mathcal{L}^2(\Omega)$ und $A = [\,a(\Lambda_j, \Lambda_k)\,]_{jk} \in \mathbb{R}^{n\times n}$ die bereits aus Abschnitt 93 vertraute Steifigkeitsmatrix ist. Der Startvektor $y(0)$ enthält die Entwicklungskoeffizienten der Funktion $u_h(0) \in V_0^{\mathcal{T}}$, die eine möglichst gute Approximation der exakten Anfangsvorgabe u° sein sollte, z. B. die interpolierende stückweise lineare Funktion vgl. Abschnitt 91.

Die Differentialgleichung (98.9) hat nach Aufgabe XIV.2 für jedes Zeitintervall $[0,T]$ eine eindeutig bestimmte Lösung $y \in \big(H^1(0,T)\big)^n$, aus der sich dann über (98.8) die Lösung des Variationsproblems (98.1) der Semidiskretisierung ergibt. Dies klärt übrigens nachträglich auch noch die Frage nach Existenz und Eindeutigkeit der Lösung von (98.1). Aus anschaulichen Gründen wird diese Semidiskretisierung auch (*vertikale*) *Linienmethode* genannt: Für alle $t \geq 0$ enthält nämlich die vektorwertige Funktion $y(t)$ die Funktionswerte $\eta_j(t)$, $j = 1, \ldots, n$, der Approximation u_h an den inneren Knoten der Triangulierung, vgl. Abbildung 98.1 für den räumlich eindimensionalen Fall mit $\Omega = (0,\pi)$.

Beispiel 98.2. Für die eindimensionale Wärmeleitungsgleichung im Intervall $(0,\pi)$, vgl. Beispiel 69.2, wählen wir für V_h den Raum der linearen Splines über dem äquidistanten Gitter

$$\triangle_h \;=\; \{\, x_j = jh \;:\; 0 \leq j \leq n+1 \,\}$$

mit Gitterweite $h = \pi/(n+1)$ und mit homogenen Randwerten. In diesem Fall sind

$$G = \frac{h}{6}\begin{bmatrix} 4 & 1 & & \\ 1 & 4 & \ddots & \\ & \ddots & \ddots & 1 \\ & & 1 & 4 \end{bmatrix} \quad \text{und} \quad A = \frac{1}{h}\begin{bmatrix} 2 & -1 & & \\ -1 & 2 & \ddots & \\ & \ddots & \ddots & -1 \\ & & -1 & 2 \end{bmatrix} \tag{98.10}$$

die entsprechenden Matrizen aus (98.9), vgl. die Beispiele 45.5 und 90.3. ◊

Die Semidiskretisierung ist allerdings nur der erste Schritt zur Berechnung konkreter numerischer Werte. Als nächstes gilt es, das Anfangswertproblem (98.9) zu lösen. Hierzu können etwa Runge-Kutta-Verfahren eingesetzt werden, sofern die zeitabhängige Funktion b zumindest stetig ist. Wie wir bereits in Beispiel 77.11 gesehen haben, sind die resultierenden Differentialgleichungen in der Regel steif, so daß lediglich implizite Verfahren in Frage kommen.

Beispiel 98.3. Wir unterteilen den Zeitbereich in äquidistante Intervalle der Länge $\tau > 0$ und bezeichnen mit

$$y_i \approx y(i\tau) \in \mathbb{R}^n\,, \qquad i = 0, 1, 2, \dots\,,$$

die Näherungen des *impliziten Euler-Verfahrens* an die exakte Lösung der Differentialgleichung (98.9). Dabei ist y_0 durch die Entwicklungskoeffizienten von $u_h(0) \in V_0^{\mathcal{T}}$ vorgegeben.

Für das implizite Euler-Verfahren ergeben sich die Näherungen aus der Rekursion

$$y_{i+1} = y_i + \tau G^{-1}(b_{i+1} - Ay_{i+1})$$

beziehungsweise

$$(G + \tau A)y_{i+1} = Gy_i + \tau b_{i+1}\,, \qquad i = 0, 1, 2, \dots\,, \tag{98.11}$$

wobei wir $b_i = b(i\tau)$, $i \in \mathbb{N}_0$, gesetzt haben. In jedem impliziten Euler-Schritt ist also ein lineares Gleichungssystem mit der positiv definiten Matrix $G + \tau A$ zu lösen.

Die Konvergenzanalyse dieses Verfahrens können wir auf die Ergebnisse aus Abschnitt 75 zurückführen. Dazu leiten wir zunächst eine einseitige Lipschitz-Bedingung für die rechte Seite des Differentialgleichungssystems (98.9) her. Hierfür seien $y = [\eta_j]$ und $\tilde{y} = [\tilde{\eta}_j]_j$ beliebige Vektoren im $\mathbb{R}^n$ und

$$u_h = \sum_{j=1}^{n} \eta_j \Lambda_j \qquad \text{und} \qquad \tilde{u}_h = \sum_{j=1}^{n} \tilde{\eta}_j \Lambda_j$$

die entsprechenden Ansatzfunktionen aus $V_0^{\mathcal{T}}$. Da G und A symmetrisch sind, gilt

$$\begin{aligned}\big(G^{-1}(b-Ay)-G^{-1}(b-A\tilde{y})\big)^* G(y-\tilde{y}) \;&=\; -(y-\tilde{y})^* A(y-\tilde{y})\\ &=\; -a(u_h-\tilde{u}_h,u_h-\tilde{u}_h)\,,\end{aligned}$$

vgl. den Beweis von Proposition 90.4. Aus der Elliptizität (97.4) der Bilinearform und der Definition der Gramschen Matrix G folgt somit

$$\begin{aligned}\big(G^{-1}(b-Ay)-G^{-1}(b-A\tilde{y})\big)^* G(y-\tilde{y}) \;&\le\; -a_0\,\|u_h-\tilde{u}_h\|^2_{H^1(\Omega)}\\ &\le\; -a_0\,\|u_h-\tilde{u}_h\|^2_{\mathcal{L}^2(\Omega)} \;=\; -a_0(y-\tilde{y})^* G(y-\tilde{y})\,.\end{aligned}$$

Demnach ist das Differentialgleichungssystem (98.9) bezüglich des zu G gehörigen (Energie-)Innenprodukts im $\mathbb{R}^n$ strikt dissipativ.

Bezeichnen wir daher mit

$$u_i=\sum_{j=1}^{n}\eta_{ij}\Lambda_j\in V_0^{\mathcal{T}}$$

die zu $y_i=[\eta_{ij}]_{j=1}^n$ gehörende stückweise lineare Ansatzfunktion, so folgt mit Aufgabe XIV.4 für das implizite Euler-Verfahren die Fehlerabschätzung

$$\|u_i-u_h(i\tau)\|_{\mathcal{L}^2(\Omega)} \;\le\; \frac{1}{2a_0}\,\tau\max_{0\le t\le i\tau}\|u_h''(t)\|_{\mathcal{L}^2(\Omega)}\,.$$

$\Diamond$

Um den zusätzlichen Fehler durch die Zeitdiskretisierung in den Bereich der Schranke aus der örtlichen Fehlerabschätzung aus Satz 98.1 zu drücken, ist also beim impliziten Euler-Verfahren eine Zeitschrittweite $\tau=O(h^2)$ erforderlich. Diese Diskrepanz zwischen den beiden Diskretisierungsparametern τ und h liegt daran, daß das Galerkin-Verfahren mit stückweise linearen Ansatzfunktionen Konsistenzordnung Zwei besitzt, während das implizite Euler-Verfahren nur ein Runge-Kutta-Verfahren erster Ordnung ist.

Bemerkung 98.4. Das lineare Gleichungssystem (98.11), das in einem Zeitschritt des impliziten Euler-Verfahrens auftritt, entspricht der Diskretisierung einer elliptischen Differentialgleichung: Die Funktion $u_{i+1}\in V_0^{\mathcal{T}}$, deren Entwicklungskoeffizienten in dem Vektor y_{i+1} stehen, ist die Galerkin-Approximation der Lösung w der Differentialgleichung

$$-\operatorname{div}(\tau\sigma\,\operatorname{grad} w)+(1+\tau c)w \;=\; u_i+\tau f(t_{i+1})\,.$$

Entsprechend aufwendig ist die Lösung dieser Gleichungssysteme bei höherdimensionalen Problemen. Allerdings ist die Kondition der zugehörigen Koeffizientenmatrix $G+\tau A$ aufgrund des Faktors τ vor der Steifigkeitsmatrix bei

kleinen Zeitschrittweiten nicht so groß wie bei den Gleichungsystemen für die elliptische Gleichung. Daher kann das Gleichungssystem (98.11) mit dem CG-Verfahren effizient gelöst werden. Bei größeren Zeitschrittweiten bieten sich Mehrgitterverfahren an. ◊

99 Das Crank-Nicolson-Verfahren

Um größere Zeitschritte wählen zu können, müssen Runge-Kutta-Verfahren höherer Ordnung eingesetzt werden, die wie das implizite Euler-Verfahren die notwendigen Stabilitätsanforderungen erfüllen. Wir greifen für die entsprechenden Verfahren wie in Beispiel 98.3 auf ein äquidistantes Zeitgitter

$$\triangle_\tau = \{\, t_i = i\tau \,:\, i \in \mathbb{N}_0 \,\}$$

mit Gitterweite $\tau > 0$ zurück und bezeichnen mit $y_i = [\eta_{ij}]$ die Approximationen der entsprechenden Runge-Kutta-Verfahren für $y(t_i)$. Ferner assoziieren wir mit y_i die stückweise lineare Funktion

$$u_i = \sum_{j=1}^{n} \eta_{ij} \Lambda_j \approx u_h(t_i)\,.$$

Die Werte η_{ij} approximieren also die Funktionswerte von u an den inneren Knoten $\{x_j \,:\, j = 1, \ldots, n\}$ der Triangulierung $\mathcal{T}$:

$$\eta_{ij} \approx u_h(x_j, t_i) \approx u(x_j, t_i)\,.$$

Wir betrachten in diesem Abschnitt das einfachste A-stabile Runge-Kutta-Verfahren zweiter Ordnung, das einstufige Gauß-Verfahren (implizites Mittelpunktverfahren). Wie in Beispiel 78.1 erläutert wurde, vgl. (78.4), führt dieses Verfahren auf die Rekursion

$$(G + \frac{\tau}{2} A) y_{i+1} = (G - \frac{\tau}{2} A) y_i + \tau b_{i+1/2}\,, \qquad i = 0, 1, 2, \ldots\,, \tag{99.1}$$

wobei wir

$$b_{i+1/2} = \big[\, \langle\, f_{i+1/2}, \Lambda_j \,\rangle_{\mathcal{L}^2(\Omega)} \,\big]_j \qquad \text{mit} \qquad f_{i+1/2} = f(t_i + \tau/2)$$

gesetzt haben. Man beachte, daß sowohl G als auch A positiv definit sind; daher ist die Koeffizientenmatrix $G + \tau/2\, A$ für jedes $\tau > 0$ positiv definit und

(99.1) eindeutig lösbar. Die Rekursion (99.1) kann auch als endlichdimensionales Variationsproblem für $u_{i+1} \in V_0^{\mathcal{T}}$ formuliert werden:

$$\begin{aligned} &\langle u_{i+1}, w \rangle_{\mathcal{L}^2(\Omega)} + \frac{\tau}{2}\, a(u_{i+1}, w) \\ &\quad = \langle u_i, w \rangle_{\mathcal{L}^2(\Omega)} - \frac{\tau}{2}\, a(u_i, w) + \tau \langle f_{i+1/2}, w \rangle_{\mathcal{L}^2(\Omega)} \quad \text{für alle } w \in V_0^{\mathcal{T}}\,. \end{aligned} \tag{99.2}$$

Das Verfahren (99.1), (99.2) wird *Crank-Nicolson-Verfahren* genannt und ist bezüglich des Aufwands mit dem impliziten Euler-Verfahren aus Beispiel 98.3 vergleichbar: Auch hier ist in jedem Zeitschritt ein lineares Gleichungssystem zu lösen, Bemerkung 98.4 gilt dabei entsprechend.

Für das implizite Mittelpunktverfahren haben wir in Kapitel XIV lediglich die allgemeine Fehlerabschätzung aus Satz 76.10 kennengelernt, für die vorausgesetzt wurde, daß die rechte Seite der Differentialgleichung eine beschränkte partielle Ableitung nach y hat. Diese Abschätzung ist für die hiesigen Zwecke unbrauchbar, da hier die Ableitung nach y durch die Matrix $-G^{-1}A$ gegeben ist, deren kleinster Eigenwert in der Regel mit wachsender Knotenzahl gegen $-\infty$ strebt, vgl. Aufgabe 3.

Der folgende Satz präsentiert daher eine Fehleranalyse des Crank-Nicolson-Verfahrens, die speziell auf parabolische Anfangsrandwertprobleme zugeschnitten ist.

Satz 99.1. *Es seien* $\|u_h(0) - u^\circ\|_{\mathcal{L}^2(\Omega)} \le C_0 h^2$ *und* $u_i \in V_0^{\mathcal{T}}$, $i \in \mathbb{N}$, *durch* (99.2) *definiert. Erfüllt* $L[\,\cdot\,]$ *die Regularitätsanforderungen aus Satz* 97.2 *und ist die exakte Lösung* u *von* (97.3) *hinreichend glatt, so gilt für jedes feste* $T > 0$ *die Abschätzung*

$$\|u_i - u(t_i)\|_{\mathcal{L}^2(\Omega)} \;\le\; O(h^2 + \tau^2)\,, \qquad 0 \le t_i \le T\,,$$

wobei die Konstante in der O-Notation nur von u *und* T *abhängt.*

Beweis. Wie im Beweis von Satz 98.1 greifen wir auf die Ritz-Projektion $v_h(t_i) \in V_0^{\mathcal{T}}$ von $u(t_i)$ aus (98.2) zurück. Zunächst ist jedoch wegen (99.2)

$$\langle u_{i+1} - u_i, w \rangle_{\mathcal{L}^2(\Omega)} + \frac{\tau}{2}\, a(u_{i+1} + u_i, w) \;=\; \tau \langle f_{i+1/2}, w \rangle_{\mathcal{L}^2(\Omega)}$$

für alle $w \in V_0^{\mathcal{T}}$, und für die Differenzen $d_i = v_h(t_i) - u_i$ und $d_{i+1} = v_h(t_{i+1}) - u_{i+1}$ ergibt sich hieraus

$$\begin{aligned} &\langle d_{i+1} - d_i, w \rangle_{\mathcal{L}^2(\Omega)} + \frac{\tau}{2}\, a(d_{i+1} + d_i, w) \\ &\quad = \quad \langle v_h(t_{i+1}) - v_h(t_i), w \rangle_{\mathcal{L}^2(\Omega)} + \frac{\tau}{2}\, a(v_h(t_{i+1}) + v_h(t_i), w) \\ &\qquad\qquad - \tau \langle f_{i+1/2}, w \rangle_{\mathcal{L}^2(\Omega)} \\ &\quad \overset{(98.2)}{=} \quad \langle g_i, w \rangle_{\mathcal{L}^2(\Omega)} \end{aligned} \tag{99.3}$$

mit

$$g_i = v_h(t_{i+1}) - v_h(t_i) + \frac{\tau}{2}\big(f_{i+1} - 2f_{i+1/2} + f_i\big) - \frac{\tau}{2}\big(u'(t_{i+1}) + u'(t_i)\big), \quad (99.4)$$

wobei wir f_i für $f(t_i)$ und f_{i+1} für $f(t_{i+1})$ gesetzt haben.

Speziell für $w = d_{i+1} + d_i$ ist der zweite Term in (99.3) nichtnegativ und daher folgt mit Hilfe der Cauchy-Schwarz-Ungleichung

$$\|d_{i+1}\|^2_{\mathcal{L}^2(\Omega)} - \|d_i\|^2_{\mathcal{L}^2(\Omega)} \;\le\; \|g_i\|_{\mathcal{L}^2(\Omega)}\big(\|d_{i+1}\|_{\mathcal{L}^2(\Omega)} + \|d_i\|_{\mathcal{L}^2(\Omega)}\big)$$

beziehungsweise

$$\|d_{i+1}\|_{\mathcal{L}^2(\Omega)} - \|d_i\|_{\mathcal{L}^2(\Omega)} \;\le\; \|g_i\|_{\mathcal{L}^2(\Omega)}\,.$$

Somit haben wir

$$\|d_{i+1}\|_{\mathcal{L}^2(\Omega)} \;\le\; \|d_i\|_{\mathcal{L}^2(\Omega)} + \|g_i\|_{\mathcal{L}^2(\Omega)} \;\le\; \ldots \;\le\; \|d_0\|_{\mathcal{L}^2(\Omega)} + \sum_{l=0}^{i} \|g_l\|_{\mathcal{L}^2(\Omega)}\,.$$

Aufgrund der Voraussetzungen und Aufgabe XVI.7 (b) ist der erste Term auf der rechten Seite von der Größenordnung $O(h^2)$, so daß sich wiederum mit Hilfe derselben Aufgabe die Ungleichung

$$\begin{aligned}\|u_i - u(t_i)\|_{\mathcal{L}^2(\Omega)} \;&\le\; \|u_i - v_h(t_i)\|_{\mathcal{L}^2(\Omega)} + \|v_h(t_i) - u(t_i)\|_{\mathcal{L}^2(\Omega)} \\ &\le\; O(h^2) + \sum_{l=0}^{i-1} \|g_l\|_{\mathcal{L}^2(\Omega)}\end{aligned} \quad (99.5)$$

ergibt.

Es verbleibt nun noch die Abschätzung von g_i aus (99.4). Dazu schreiben wir

$$\begin{aligned} g_i \;&=\; v_h(t_{i+1}) - v_h(t_i) - \big(u(t_{i+1}) - u(t_i)\big) && (99.6a)\\ &+\; \frac{\tau}{2}\,(f_{i+1} - 2f_{i+1/2} + f_i) && (99.6b)\\ &+\; u(t_{i+1}) - u(t_i) - \frac{\tau}{2}\,u'(t_{i+1}) - \frac{\tau}{2}\,u'(t_i) && (99.6c)\end{aligned}$$

und schätzen die drei Terme in den einzelnen Zeilen separat ab.

Wir beginnen mit (99.6a): Aufgrund der Definition der Ritz-Projektion gilt

$$a\big(v_h(t_{i+1}) - v_h(t_i), w\big) \;=\; a\big(u(t_{i+1}) - u(t_i), w\big) \qquad \text{für alle } w \in V_0^{\mathcal{T}}\,,$$

d. h. $v_h(t_{i+1}) - v_h(t_i)$ ist die Ritz-Projektion von $u(t_{i+1}) - u(t_i)$. Aus Aufgabe XVI.7 (b) folgt daher

$$\begin{aligned}\|v_h(t_{i+1}) - v_h(t_i) - \big(u(t_{i+1}) - u(t_i)\big)\|_{\mathcal{L}^2(\Omega)} \;&\le\; c_0 h^2 \|u(t_{i+1}) - u(t_i)\|_{H^2(\Omega)} \\ &\le\; c_0 h^2 \int_{t_i}^{t_{i+1}} \|u'(t)\|_{H^2(\Omega)}\,dt\,,\end{aligned}$$

und dieses Integral kann durch $\tau \|u'(\tilde{t}_i)\|_{H^2(\Omega)}$ mit einer geeigneten Zwischenstelle $\tilde{t}_i \in [t_i, t_{i+1}]$ nach oben abgeschätzt werden. Somit haben wir für (99.6a) die Ungleichung

$$\|v_h(t_{i+1}) - v_h(t_i) - \big(u(t_{i+1}) - u(t_i)\big)\|_{\mathcal{L}^2(\Omega)} \le c_0 h^2 \tau \|u'(\tilde{t}_i)\|_{H^2(\Omega)} . \tag{99.7a}$$

Um (99.6b) abzuschätzen, gehen wir wie in Lemma 83.3 vor:[2] Demnach ist

$$f_{i+1} - 2f_{i+1/2} + f_i = \frac{\tau^2}{4} f''(t_i + \tau/2) + O(\tau^4)$$

und hieraus folgt

$$\|\frac{\tau}{2}\,(f_{i+1} - 2f_{i+1/2} + f_i)\|_{\mathcal{L}^2(\Omega)} = O(\tau^3) . \tag{99.7b}$$

Bezüglich des letzten Terms (99.6c) erhalten wir aus Lemma 83.2 die Darstellung

$$\begin{aligned} &2\big(u(t_{i+1}) - u(t_i)\big) - \tau u'(t_{i+1}) - \tau u'(t_i) \\ &\qquad = 2\tau u'(t_i + \tau/2) - \tau u'(t_{i+1}) - \tau u'(t_i) + O(\tau^3) , \end{aligned}$$

und nun folgt wie zuvor mit Lemma 83.3 die Abschätzung

$$\|u(t_{i+1}) - u(t_i) - \frac{\tau}{2}\,u'(t_{i+1}) - \frac{\tau}{2}\,u'(t_i)\|_{\mathcal{L}^2(\Omega)} = O(\tau^3) . \tag{99.7c}$$

Verwenden wir die Abschätzungen (99.7) in (99.6), so ergibt dies die Schranke

$$\|g_l\|_{\mathcal{L}^2(\Omega)} \le C\tau(h^2 + \tau^2) , \qquad l = 0, 1, \dots, i-1 ,$$

mit einer nur von dem Zeitintervall $[0, t_i]$ abhängigen Konstanten $C > 0$, und somit ist

$$\sum_{l=0}^{i-1} \|g_l\|_{\mathcal{L}^2(\Omega)} \le iC\tau(h^2 + \tau^2) \le CT(h^2 + \tau^2) ,$$

solange $t_i = i\tau$ im Intervall $[0, T]$ liegt. Zusammen mit (99.5) folgt hieraus schließlich die Behauptung. □

Beispiel 99.2. Zur Illustration dieses theoretischen Ergebnisses vergleichen wir abschließend die Genauigkeit des Crank-Nicolson-Verfahrens mit der des

[2] Da hier Funktionen mehrerer Veränderlicher abzuschätzen sind, muß im Beweis von Lemma 83.3 der Taylorrest in integraler Form dargestellt werden. Die folgenden Abschätzungen sind daher mit großer Sorgfalt nachzuvollziehen, vgl. Aufgabe 5.

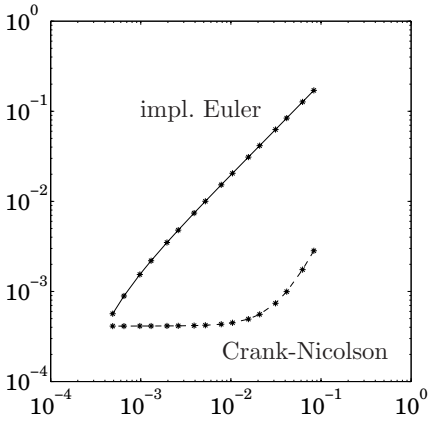


Abb. 99.1:
Relative $\mathcal{L}^2$-Fehler für das implizite Euler- und das Crank-Nicolson-Verfahren

impliziten Euler-Verfahrens. Dazu wenden wir beide Verfahren mit verschiedenen Zeitschrittweiten auf die homogene Wärmeleitungsgleichung aus Beispiel 69.2 mit der dort angegebenen Ausgangstemperatur $u^\circ(x) = x(\pi - x)$ an, und bestimmen die relativen Fehler der beiden Näherungen von $u(t)$ für $t = 4$ in der $\mathcal{L}^2$-Norm. Als Ansatzraum $V_0^{\mathcal{T}}$ für die Ortsdiskretisierung wählen wir lineare Splines über einem äquidistanten Gitter mit Gitterweite $h = \pi/100$.

In Abbildung 99.1 sind die relativen Fehler in Abhängigkeit von der Zeitschrittweite τ aufgetragen. Das Crank-Nicolson-Verfahrens erreicht bereits für relativ große Zeitschrittweiten $\tau \approx 10^{-2}$ (vergleichbar mit der Gitterweite h bezüglich des Orts) die optimale Genauigkeit, kleinere Zeitschrittweiten können den örtlichen Diskretisierungsfehler nicht aufwiegen. Im Unterschied dazu muß beim impliziten Euler-Verfahren eine Zeitschrittweite $\tau \sim h^2$ gewählt werden, um dieselbe Genauigkeit zu erreichen. Dank der doppeltlogarithmischen Darstellung läßt sich aus der Steigung der beiden Fehlerkurven oberhalb des Diskretisierungsfehlers die jeweilige Ordnung der Verfahren ablesen: Für das implizite Euler-Verfahren ist die Steigung gleich Eins, für das Crank-Nicolson-Verfahren ist die Steigung gleich Zwei. ◊

100 Maximumprinzipien

Bei der Frage nach der Eindeutigkeit der Lösung parabolischer Differentialgleichungen, aber auch bei der Untersuchung ihrer qualitativen Eigenschaften spielen sogenannte Maximumprinzipien eine entscheidende Rolle. Das folgende Resultat ist ein Prototyp eines *schwachen Maximumprinzips*.

Satz 100.1. *Der Differentialoperator L sei durch* (97.1) *mit $c = 0$ und stetig*

differenzierbarem Koeffizienten $\sigma \geq \sigma_0 > 0$ gegeben. Ferner sei u eine klassische Lösung der Differentialgleichung $u_t + L[u] = f$ mit $f(x,t) \leq 0$ für alle $(x,t) \in \Omega \times [0,T]$ mit $T < \infty$. Dann nimmt die Funktion u ihr Maximum am Rand Γ von Ω oder für $t = 0$ an.

Beweis. Wir betrachten zunächst den Fall, daß f in $\Omega \times (0,T]$ strikt negativ ist. Dann kann das Maximum von u in keinem inneren Punkt $(x_0,t_0) \in \Omega \times (0,T)$ angenommen werden, da ansonsten nach den bekannten Sätzen der Analysis an diesem Punkt

$$u_t(x_0,t_0) = 0\,, \quad \operatorname{grad} u(x_0,t_0) = 0, \quad u_{\xi\xi}(x_0,t_0),\, u_{\eta\eta}(x_0,t_0) \leq 0 \tag{100.1}$$

gelten müßte. Hieraus würde dann aber aufgrund der Elliptizitätsbedingung $\sigma(x) \geq \sigma_0 > 0$ die Ungleichung

$$\begin{aligned} f(x_0,t_0) &= u_t(x_0,t_0) + L[u](x_0,t_0) \\ &= u_t(x_0,t_0) - \sigma(x_0)\Delta u(x_0,t_0) - \operatorname{grad}\sigma(x_0) \cdot \operatorname{grad} u(x_0,t_0) \\ &\geq 0 \end{aligned}$$

folgen, doch diese steht im Widerspruch zu der getroffenen Annahme $f < 0$. Entsprechend sieht man auch, daß das Maximum nicht für $t_0 = T$ angenommen werden kann, denn ansonsten gilt (100.1) mit der Modifikation $u_t(x_0,T) \geq 0$, und der anschließende Widerspruch zu dem Vorzeichen von f an der Maximalstelle (x_0,T) bleibt weiterhin gültig.

Falls f in $\Omega \times (0,T]$ nicht strikt negativ ist, betrachtet man die Funktion $u_\varepsilon = u + \varepsilon e^{-t}$ mit $\varepsilon > 0$, die die Differentialgleichung

$$\frac{d}{dt}u_\varepsilon + L[u_\varepsilon] = \frac{d}{dt}u + \varepsilon\frac{d}{dt}e^{-t} + L[u] = f - \varepsilon e^{-t} =: f_\varepsilon$$

erfüllt. Da f_ε in $\Omega \times (0,T]$ strikt negativ ist, kann aus dem bereits bewiesenen Teilresultat geschlossen werden, daß u_ε das Maximum auf dem Rand Γ oder für $t = 0$ annimmt. Durch Grenzübergang $\varepsilon \to 0$ überträgt sich diese Aussage auch auf die Grenzfunktion u. □

Bemerkung. Für den Fall $f < 0$ liefert der Beweis das schärfere Ergebnis, daß das Maximum von u *nur* auf dem Rand Γ oder durch die Anfangsvorgabe für $t = 0$ angenommen werden kann. Durch den Grenzübergang im Fall $f \leq 0$ geht diese Eigenschaft verloren, das heißt prinzipiell können für $t > 0$ weitere Maxima im Innern von Ω *hinzukommen*. Tatsächlich kann man zeigen (vgl. etwa Protter und Weinberger [85]), daß unter unseren Voraussetzungen zu einer Zeit $t_0 > 0$ nur dann ein Maximum in einem inneren Punkt x_0 von Ω vorliegen kann, wenn die Funktion u in $\Omega \times [0,t_0]$ konstant ist. Dies ist das sogenannte *starke Maximumprinzip*. ◊

Gibt u die Temperaturverteilung in einem Gebiet Ω an, so besagt Satz 100.1, daß bei Abwesenheit von Wärmequellen im Innern des Gebiets ($f \le 0$, vgl. Abschnitt 69) die Höchsttemperatur am Rand des Gebiets oder zum Zeitpunkt $t = 0$ angenommen wird.

Wir wollen nun exemplarisch anhand der eindimensionalen Wärmeleitungsgleichung

$$\begin{aligned} u_t - u_{xx} \;=\; f\,, \quad u(t,0) = u(t,\pi) = 0\,, \quad u(0) = u^\circ \in V_0^{\mathcal{T}}\,, \\ 0 \le x \le \pi\,, \qquad t \ge 0\,, \end{aligned} \tag{100.2}$$

untersuchen, ob die berechneten Näherungen $u_i \approx u(t_i)$ des Crank-Nicolson-Verfahrens ein entsprechendes Maximumprinzip erfüllen. Für die Ortsdiskretisierung sei $V_0^{\mathcal{T}}$ der Raum der linearen Splines über dem äquidistanten Gitter

$$\triangle_h \;=\; \{\, x_j = jh \,:\, 0 \le j \le n+1 \,\}$$

mit Gitterweite $h = \pi/(n+1)$. Die entsprechenden Matrizen G und A wurden in Beispiel 98.2 angegeben. Für diesen Spezialfall gilt das folgende Resultat:

Proposition 100.2. *Die Approximationen $y_i = [\eta_{ij}]$ des Crank-Nicolson-Verfahrens* (99.1) *für die Wärmeleitungsgleichung* (100.2) *genügen für*

$$h^2/3 \le \tau \le 2h^2/3 \tag{100.3}$$

dem (schwachen) Maximumprinzip, d. h. unter der Voraussetzung $f \le 0$ gilt

$$\max_{1\le j\le n} \eta_{ij} \;\le\; \max_{0\le j\le n+1} \eta_{0j}\,, \qquad i \in \mathbb{N}\,,$$

wobei $\eta_{00} = \eta_{0,n+1} = 0$ gesetzt seien.

Beweis. Durch Vergleich der Nebendiagonalelemente in (98.10) sieht man mit Hilfe von Satz 84.3, daß $G + \tau/2\,A$ für $\tau/(2h) > h/6$ eine M-Matrix ist, also $(G+\tau/2\,A)^{-1}$ für $\tau > h^2/3$ nichtnegativ ist; letzteres bleibt auch für $\tau = h^2/3$ richtig, da in diesem Fall $G+\tau/2\,A$ eine nichtnegative Diagonalmatrix ist. Entsprechend ergibt sich durch Vergleich der Hauptdiagonalelemente in (98.10), daß die Matrix $G-\tau/2\,A$ für $\tau/h \le 4h/6$ nichtnegativ ist. Die Rekursion (99.1) des Crank-Nicolson-Verfahrens lautet also

$$y_{i+1} \;=\; Ty_i + \tau M^{-1} b_{i+1/2}\,, \qquad i = 0, 1, 2, \dots\,, \tag{100.4}$$

wobei die beiden Matrizen

$$T = (G + \frac{\tau}{2}A)^{-1}(G - \frac{\tau}{2}A) \qquad \text{und} \qquad M^{-1} = (G + \frac{\tau}{2}A)^{-1}$$

nichtnegativ sind, sofern die Zeitschrittweite τ in dem angegebenen Bereich $h^2/3 \le \tau \le 2h^2/3$ liegt. Die Einträge $\langle\, f(t_i + \tau/2), \Lambda_j \,\rangle_{\mathcal{L}^2(\Omega)}$ von $b_{i+1/2}$ sind unter der Voraussetzung an das Vorzeichen von f nichtpositiv und somit ist

$$y_{i+1} \le T y_i\,, \qquad i = 0, 1, 2, \ldots \tag{100.5}$$

Setzen wir nun $\bar{u}_i = \max_j \eta_{ij}$, so haben wir (komponentenweise) $y_i \le \bar{u}_i \mathbb{1}$ und aus (100.5) und der Nichtnegativität von T folgt

$$y_{i+1} \le T\bar{u}_i\mathbb{1} = \bar{u}_i T\mathbb{1}\,.$$

Aus der Ungleichungskette

$$(G - \frac{\tau}{2}A)\mathbb{1} = G\mathbb{1} - \frac{\tau}{2h}\begin{bmatrix}1\\0\\\vdots\\0\\1\end{bmatrix} \le G\mathbb{1} + \frac{\tau}{2h}\begin{bmatrix}1\\0\\\vdots\\0\\1\end{bmatrix} = M\mathbb{1}$$

erhalten wir zudem $T\mathbb{1} \le \mathbb{1}$ wegen der Nichtnegativität von M^{-1}. Somit ist in diesem Fall $y_{i+1} \le \bar{u}_i T\mathbb{1} \le \bar{u}_i\mathbb{1}$ und $\max_j \eta_{i+1,j} \le \bar{u}_i = \max_j \eta_{ij}$. Durch vollständige Induktion folgt schließlich die Behauptung. □

Die obere Schranke für die Zeitschrittweite aus Proposition 100.2 weist auf einen Nachteil des Crank-Nicolson-Verfahrens hin. Obwohl aufgrund der Fehlerabschätzung aus Satz 99.1 für dieses Verfahren recht große Zeitschrittweiten $\tau \sim h$ für eine vorgegebene Genauigkeit $\varepsilon \approx h^2$ in der $\mathcal{L}^2$-Norm möglich sind, offenbart Proposition 100.2, daß für derart große Schrittweiten Stabilitätsprobleme bezüglich der Maximumnorm auftreten können.

Beispiel 100.3. Die Verletzung des Maximumprinzips für große Zeitschrittweiten macht sich vor allem dann bemerkbar, wenn die exakte Lösung wenig glatt ist. Als Beispiel wählen wir für die Anfangsvorgabe u° die Hutfunktion aus Abbildung 100.1, die sich über die mittleren acht der insgesamt hundert Gitterintervalle erstreckt ($h = \pi/100$). Diese Funktion gehört zwar zu $H_0^1(\Omega)$, ihre H^1-Norm ist wegen der starken Steigungen jedoch relativ groß. Mit dieser Anfangstemperatur führen große Zeitschritte zu starken Oszillationen in den berechneten Näherungslösungen. Beispielhaft zeigt Abbildung 100.1 rechts die exakte Lösung und die numerische Rekonstruktion zum Zeitpunkt $t = 0.25$ mit Zeitschrittweite $\tau = h/4$.

Trotz der unteren Schranke aus Proposition 100.2 lassen sich selbst bei extrem kleinen Zeitschrittweiten keine derartig starken Oszillationen beobachten. Allerdings treten bei kleinem τ für kleine Zeiten t_i geringfügig negative Temperaturen auf, die belegen, daß das Maximumprinzip verletzt ist. Bei glatten

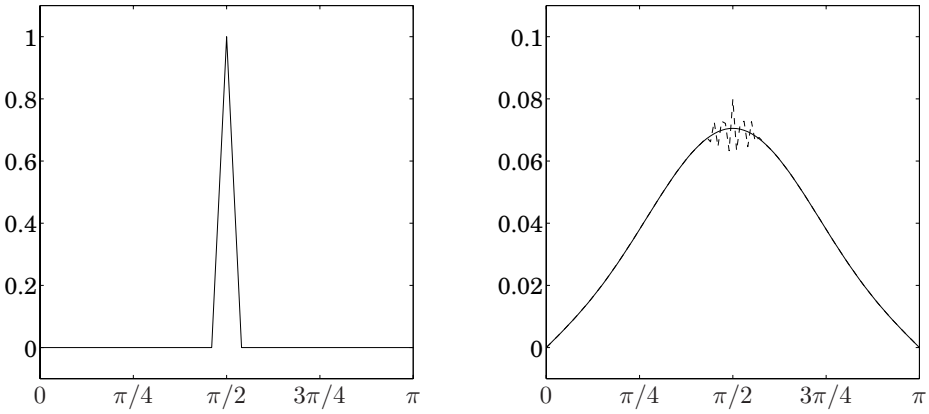


Abb. 100.1: Vorgabe u° (links) und berechnete Näherung von $u(0.25)$ für $\tau = h/4$ (rechts)

Anfangsvorgaben, etwa der Funktion $u^\circ(x) = x(\pi - x)$ aus Beispiel 69.2, und glatten Quelltermen f ergeben sich weder bei sehr großen noch bei extrem kleinen Zeitschrittweiten störende Oszillationen. ◊

Für allgemeinere Differentialoperatoren als $L[u] = -u_{xx}$ können entsprechende Abschätzungen wie in Proposition 100.2 bewiesen werden. Dabei kann es allerdings passieren, daß die untere Schranke in (100.3) den Wert der oberen Schranke überschreitet. Aus dem Beweis der Proposition erkennt man, daß die untere Schranke für die Nichtnegativität von M^{-1} wesentlich ist; sofern M^{-1} negative Einträge besitzt, erhält man etwa für $y_i = 0$ und geeignetes $f \leq 0$ eine Approximation y_{i+1} mit positiven Komponenten, die das Maximumprinzip verletzt.

Tatsächlich kann die untere Schranke vollständig umgangen und gleichzeitig die obere Schranke ein wenig vergrößert werden, wenn man die Koeffizientenmatrix $G + \tau/2A$ des Crank-Nicolson-Verfahrens geeignet modifiziert. Dies ist die sogenannte *Lumping-Variante*, die abschließend noch kurz vorgestellt werden soll. Hierbei wird die Gramsche Matrix $G = \left[\langle\, \Lambda_i, \Lambda_j \,\rangle_{\mathcal{L}^2(\Omega)}\right]$ in (99.1) durch eine Diagonalmatrix $D = [d_{ij}]$ ersetzt, deren Diagonaleinträge gerade die Zeilensummen der Gramschen Matrix sind:

$$d_{ii} = \sum_{j=1}^{n} \langle\, \Lambda_i, \Lambda_j \,\rangle_{\mathcal{L}^2(\Omega)}\,, \qquad i = 1, \ldots, n\,.$$

Der Name dieser Variante stammt aus dem Englischen und umschreibt, daß die Masseanteile der Massematrix G zu einem „Masse-Klumpen“ (engl.: *lump*) auf der Hauptdiagonalen von D zusammengefaßt werden.

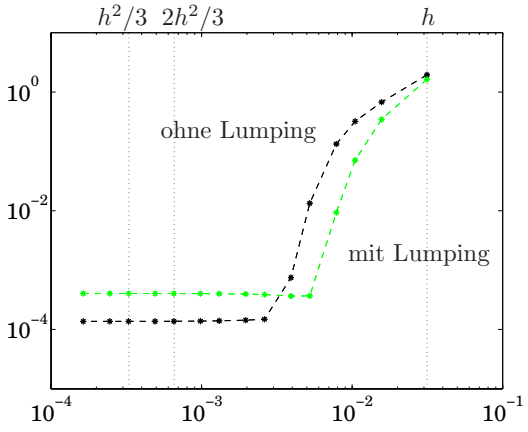


Abb. 100.2: Relativer Fehler bzgl. der Maximumnorm für verschiedene Zeitschrittweiten τ

Mit der Lumping-Variante lautet das Crank-Nicolson-Verfahren

$$(D + \frac{\tau}{2}A)y_{i+1} = (D - \frac{\tau}{2}A)y_i + \tau b_{i+1/2}\,, \qquad i = 0, 1, 2, \ldots \tag{100.6}$$

Da D eine nichtnegative Diagonalmatrix ist, ist $D + \tau/2A$ immer eine M-Matrix. In diesem Fall muß für ein diskretes Maximumprinzip also lediglich darauf geachtet werden, daß $D - \tau/2A$ nichtnegativ ist.

Während für die Lumping-Variante ähnliche Fehlerabschätzungen bezüglich der $\mathcal{L}^2$-Norm bewiesen werden können (vgl. Quarteroni und Valli [87, S. 402ff]), werden die Fehler bezüglich der Maximumnorm aufgrund der erhöhten Stabilität im kritischen Schrittweitenbereich $\tau > h^2$ zum Teil deutlich verringert.

Beispiel. Für die Wärmeleitungsgleichung (100.2) und die bereits oben verwendete Galerkin-Diskretisierung ergibt sich die Lumping-Matrix

$$D = \frac{h}{6}\begin{bmatrix} 5 & & & & \\ & 6 & & & \\ & & \ddots & & \\ & & & 6 & \\ & & & & 5 \end{bmatrix}.$$

Somit ist $D - \tau/2A$ genau dann nichtnegativ, wenn $\frac{5h}{6} - \frac{\tau}{2}\frac{2}{h}$ nichtnegativ ist, also für $\tau \leq 5h^2/6$. Diese obere Schranke ist geringfügig besser als die aus Proposition 100.2.

In Abbildung 100.2 wird die Lumping-Methode mit dem klassischen Galerkin-Verfahren für die Anfangstemperatur u° aus Beispiel 100.3 verglichen. Die

beiden Kurven zeigen die relativen Fehler des Crank-Nicolson-Verfahrens bezüglich der Maximumnorm zum Zeitpunkt $t = 0.25$ für verschiedene Zeitschrittweiten $\tau > 0$, die vertikalen Linien markieren die Schranken $\tau = h^2/3$ und $2h^2/3$ aus Proposition 100.2 sowie die Zeitschrittweite $\tau = h$, die nach Satz 99.1 für einen optimalen $\mathcal{L}^2$-Fehler ausreichen sollte. Die numerischen Resultate mit der Lumping-Technik weisen selbst bei größeren Zeitschrittweiten keine Oszillationen wie in Abbildung 100.1 auf, und die relativen Fehler sind in diesem Bereich ebenfalls etwas besser als bei dem Galerkin-Ansatz. Das Galerkin-Verfahren erreicht dafür bei kleineren Schrittweiten eine höhere Genauigkeit. ◊

101 Verfahren höherer Ordnung

Wenn die Lösung u bezüglich der Zeit sehr glatt ist, wird man dem impliziten Mittelpunktverfahren ein Runge-Kutta-Verfahren höherer Ordnung vorziehen. Ein solches Verfahren sollte jedoch A-stabil, am besten sogar L-stabil sein.

In Kapitel XIV haben wir bei der Berechnung der Konsistenzordnung eines Runge-Kutta-Verfahrens immer (stillschweigend) vorausgesetzt, daß alle Ableitungen der rechten Seite der gewöhnlichen Differentialgleichung beschränkt sind. Bei parabolischen Differentialgleichungen ist diese Voraussetzung nicht erfüllt und daher beobachtet man in der Regel eine sogenannte *Ordnungsreduktion*, d. h. die numerischen Verfahren haben einen globalen Fehler mit einer geringeren Konsistenzordnung.

Wir wollen das im folgenden anhand der (L-stabilen) Radau-IIA-Verfahren für die eindimensionale inhomogene Wärmeleitungsgleichung

$$u_t = u_{xx} + f\,, \quad u(0) = u^\circ\,, \quad u(0,t) = u(\pi,t) = 0\,, \tag{101.1}$$

im Intervall $0 < x < \pi$ demonstrieren. Dabei sei $u^\circ \in \mathcal{L}^2(0,\pi)$ und $f = f(x,t)$ eine bezüglich der Zeit hinreichend glatte Funktion mit $f(t) \in \mathcal{L}^2(0,\pi)$ für alle $t \geq 0$.

Analog zu Beispiel 69.2 entwickeln wir alle Funktionen in Sinusreihen,

$$f(x,t) \;=\; \sum_{j=1}^{\infty} \varphi_j(t) \sin jx\,, \qquad u^\circ(x) \;=\; \sum_{j=1}^{\infty} \eta_{0j} \sin jx\,,$$

und erhalten die Lösung

$$u(x,t) \;=\; \sum_{j=1}^{\infty} \eta_j(t) \sin jx\,, \tag{101.2}$$

deren Entwicklungskoeffizienten das Differentialgleichungssystem

$$\eta_j' = -j^2\eta_j + \varphi_j(t)\,, \quad \eta_j(0) = \eta_{0j}\,, \qquad j \in \mathbb{N}\,, \tag{101.3}$$

lösen.

Um das Phänomen der Ordnungsreduktion zu verstehen, ist es daher hilfreich, das Radau-IIA-Verfahren zunächst für das inhomogene Anfangswertproblem

$$y' = \lambda y + \varphi(t)\,, \quad y(0) = y_0\,, \qquad t \geq 0\,, \tag{101.4}$$

mit $y_0 \in \mathbb{R}$ und $\lambda < 0$ als Testgleichung für konstante Zeitschrittweiten $\tau > 0$ zu untersuchen. Dazu übernehmen wir die Notationen aus Kapitel XIV und bezeichnen mit (A, b, c) die Koeffizienten des s-stufigen Radau-IIA-Verfahrens.[3] Die Näherungen y_i des Radau-IIA-Verfahrens an die Werte $y(t_i)$ der exakten Lösung von (101.4) ergeben sich mit diesen Koeffizienten aus der Rekursion

$$\begin{aligned} \eta_j &= y_i + \tau \sum_{\nu=1}^{s} a_{j\nu}\big(\lambda\eta_\nu + \varphi(t_i + c_\nu\tau)\big)\,, \qquad j = 1, \ldots, s\,, \\ y_{i+1} &= y_i + \tau \sum_{j=1}^{s} b_j\big(\lambda\eta_j + \varphi(t_i + c_j\tau)\big)\,. \end{aligned} \tag{101.5}$$

Im folgenden Hilfsresultat schätzen wir zunächst die sogenannten *Defekte* ab, die sich beim Einsetzen der exakten Lösung in diese Rekursion ergeben.

Lemma 101.1. *Sei $s \geq 3$, $\varphi^{(s+2)} \in \mathcal{L}^2(0,T)$ und ω das Knotenpolynom der s-stufigen Radau-Legendre-Formel über dem Intervall $[0,1]$. Dann sind*

$$\begin{aligned} y(t_{i+1}) - y(t_i) - \tau \sum_{j=1}^{s} b_j y'(t_i + c_j\tau) &= \tau^{s+3}\varepsilon_i\,, \\ y(t_i + c_j\tau) - y(t_i) - \tau \sum_{\nu=1}^{s} a_{j\nu} y'(t_i + c_\nu\tau) &= \tau^{s+1} y^{(s+1)}(t_i) w_j + \frac{\tau^{s+2}}{\lambda}\,\varepsilon_{ij}\,, \end{aligned}$$

$j = 1, \ldots, s$, die Defekte der exakten Lösung y von (101.4) bezüglich des Radau-IIA-Verfahrens. Hierbei ist

$$w_j = \frac{1}{s!}\int_0^{c_j} \omega(t)\,dt\,, \qquad j = 1, \ldots, s\,, \tag{101.6}$$

[3] In diesem Abschnitt bezeichnet A also die Koeffizientenmatrix des Radau-IIA-Verfahrens und nicht die Steifigkeitsmatrix der Finite-Elemente-Diskretisierung, b enthält die Gewichte der Radau-Legendre-Formel und nicht die rechte Seite des gewöhnlichen Differentialgleichungssystems (98.9). Da die Differentialgleichung (98.9) in diesem Abschnitt nicht auftritt, dürften Verwechslungen ausgeschlossen sein.

und $|\varepsilon_i|$ und $|\varepsilon_{ij}|$ sind jeweils beschränkt durch

$$C'_s\Big(\frac{1}{\tau}\int_{t_i}^{t_{i+1}} |y^{(s+3)}(t)|\,dt \;+\; \frac{1}{\tau}\int_{t_i}^{t_{i+1}} |\varphi^{(s+2)}(t)|\,dt\Big) \tag{101.7}$$

mit einer Konstanten $C'_s > 0$, die von den Koeffizienten des Radau-IIA-Verfahrens, also letztlich nur von der Stufenzahl des Verfahrens abhängt.

Beweis. Durch Taylorentwicklung um $t = t_i$ erhalten wir

$$\begin{aligned}
&y(t_i + c_j\tau) - y(t_i) - \tau\sum_{\nu=1}^{s} a_{j\nu} y'(t_i + c_\nu\tau) \\
&\qquad = \sum_{k=1}^{s+1} \frac{y^{(k)}(t_i)}{k!}\, c_j^k \tau^k \;-\; \sum_{\nu=1}^{s} a_{j\nu} \sum_{k=1}^{s+1} \frac{y^{(k)}(t_i)}{(k-1)!}\, c_\nu^{k-1}\tau^k \\
&\qquad\quad + \int_{t_i}^{t_i+c_j\tau} \frac{(t_i + c_j\tau - t)^{s+1}}{(s+1)!}\, y^{(s+2)}(t)\,dt \\
&\qquad\quad - \tau\sum_{\nu=1}^{s} a_{j\nu} \int_{t_i}^{t_i+c_\nu\tau} \frac{(t_i + c_\nu\tau - t)^{s}}{s!}\, y^{(s+2)}(t)\,dt \\
&\qquad = \sum_{k=1}^{s+1} \frac{y^{(k)}(t_i)}{(k-1)!}\Big(\frac{1}{k}\, c_j^k - \sum_{\nu=1}^{s} a_{j\nu} c_\nu^{k-1}\Big)\tau^k \;+\; \frac{\tau^{s+2}}{\lambda}\,\varepsilon_{ij}
\end{aligned} \tag{101.8}$$

mit

$$\begin{aligned}
\varepsilon_{ij} \;=\; &\frac{\lambda}{\tau}\int_{t_i}^{t_i+c_j\tau} \Big(\frac{t_i + c_j\tau - t}{\tau}\Big)^{s+1} \frac{y^{(s+2)}(t)}{(s+1)!}\,dt \\
&- \sum_{\nu=1}^{s} a_{j\nu}\,\frac{\lambda}{\tau}\int_{t_i}^{t_i+c_\nu\tau} \Big(\frac{t_i + c_\nu\tau - t}{\tau}\Big)^{s} \frac{y^{(s+2)}(t)}{s!}\,dt\,.
\end{aligned} \tag{101.9}$$

Die Koeffizienten $a_{j\nu}$ des Radau-IIA-Verfahrens sind so gewählt, daß die Quadraturformeln

$$Q_j[f] \;=\; \sum_{\nu=1}^{s} a_{j\nu} f(c_\nu) \;\approx\; \int_0^{c_j} f(t)\,dt\,, \qquad j = 1, \ldots, s\,,$$

für alle $f \in \Pi_{s-1}$ exakt sind. Somit ist

$$\sum_{\nu=1}^{s} a_{j\nu} c_\nu^{k-1} \;=\; \int_0^{c_j} t^{k-1}\,dt \;=\; \frac{1}{k}\, c_j^k\,, \qquad j, k = 1, \ldots, s\,,$$

und für $p(t) = t^s$ ergibt sich wegen $p - \omega \in \Pi_{s-1}$ für $j = 1, \dots, s$

$$\sum_{\nu=1}^{s} a_{j\nu} c_\nu^s \;=\; Q_j[p] \;=\; Q_j[p-\omega] + Q_j[\omega] \;=\; \int_0^{c_j} (p-\omega)(t)\,dt + 0$$
$$= \frac{1}{s+1}\, c_j^{s+1} - \int_0^{c_j} \omega(t)\,dt\,.$$

Setzen wir dies in die Defektdarstellung (101.8) ein, so folgt

$$y(t_i + c_j\tau) - y(t_i) - \tau \sum_{\nu=1}^{s} a_{j\nu} y'(t_i + c_\nu \tau)$$
$$= \Big(\frac{1}{s!}\int_0^{c_j} \omega(t)\,dt\Big) y^{(s+1)}(t_i)\tau^{s+1} \;+\; \frac{\tau^{s+2}}{\lambda}\,\varepsilon_{ij}\,,$$

und es verbleibt nur noch die Abschätzung des Restterms ε_{ij} aus (101.9):

$$|\varepsilon_{ij}| \;\le\; \frac{|\lambda|}{\tau} \int_{t_i}^{t_i + c_j\tau} \frac{|y^{(s+2)}(t)|}{(s+1)!}\,dt \;+\; \sum_{\nu=1}^{s} |a_{j\nu}|\, \frac{|\lambda|}{\tau} \int_{t_i}^{t_i + c_\nu \tau} \frac{|y^{(s+2)}(t)|}{s!}\,dt$$
$$\le\; \frac{1}{s!}(\,\|A\|_\infty + 1)\,\frac{1}{\tau}\int_{t_i}^{t_{i+1}} |\lambda y^{(s+2)}(t)|\,dt\,.$$

Durch $(s+2)$-maliges Ableiten der Differentialgleichung (101.4) können wir hierin $\lambda y^{(s+2)}$ durch $y^{(s+3)} - \varphi^{(s+2)}$ ersetzen und erhalten so die gewünschte Schranke (101.7) für ε_{ij}.

Wir verzichten auf den Beweis der anderen Behauptung, da die Berechnung des Defekts $y(t_{i+1}) - y(t_i) - \tau \sum_j b_j y'(t_i + c_j\tau)$ ganz entsprechend erfolgt. Im Unterschied zu vorher muß dieser Defekt allerdings in ein Taylorpolynom vom Grad $s+2$ entwickelt werden, und man muß beachten, daß der Exaktheitsgrad $2s-2$ der s-stufigen Radau-Legendre-Formel aufgrund der Voraussetzung $s \ge 3$ mindestens $s+1$ ist. □

Für die weiteren Umformungen führen wir den Vektor $w = [w_j] \in \mathbb{R}^s$ mit den Koeffizienten w_j aus (101.6) und die Funktionen

$$Q(\zeta) \;=\; (I - \zeta A)^{-*} b \qquad \text{und} \qquad q(\zeta) \;=\; \frac{b^*(I-\zeta A)^{-1} w}{R(\zeta) - 1} \tag{101.10}$$

ein, die mit den Koeffizienten und der Stabilitätsfunktion R des s-stufigen Radau-IIA-Verfahrens gebildet werden. Man beachte, daß $Q : \mathbb{R} \to \mathbb{R}^s$ vektorwertig und $q : \mathbb{R} \to \mathbb{R}$ eine skalare Funktion ist.

Lemma 101.2. *Die Funktionen Q und q aus* (101.10) *sind rationale Funktionen, die für $s \geq 3$ über $\mathbb{R}_0^-$ beschränkt sind, d. h. für jedes $s \geq 3$ gibt es eine Konstante $c_s > 0$ mit*

$$|q(\zeta)| \leq c_s\,, \quad \|Q(\zeta)\|_1 \leq c_s\,, \qquad \zeta \in \mathbb{R}_0^-\,.$$

Beweis. Die Einträge von Q sind durch $b^*(I - \zeta A)^{-1}e_j$, $j = 1, \ldots, s$, gegeben, wobei die e_j wieder die kartesischen Basisvektoren des $\mathbb{R}^s$ bezeichnen. Wie im Beweis von Satz 77.2 sieht man daher, daß diese Einträge rationale Funktionen sind, die (höchstens) an den Eigenwerten von A Polstellen besitzen. Für die Radau-IIA-Verfahren liegen diese Eigenwerte nach Aufgabe XIV.10 in der rechten Halbebene. Insbesondere ist A invertierbar und

$$b^*(I - \zeta A)^{-1}e_j \;=\; \frac{1}{\zeta}\, b^*(I/\zeta - A)^{-1}e_j \;\sim\; -\frac{1}{\zeta}\, b^*A^{-1}e_j\,, \qquad |\zeta| \to \infty\,,$$

das heißt

$$Q(\zeta) \to 0 \qquad \text{für} \;\; |\zeta| \to \infty\,. \tag{101.11}$$

Daher ist die Betragssummennorm von Q über $\mathbb{R}_0^-$ beschränkt.

Nun wenden wir uns der Funktion $q = Q^*w/(R-1)$ zu, die offensichtlich auch eine rationale Funktion ist. Da die Radau-IIA-Verfahren nach Satz 79.3 L-stabil sind, also $R(\infty) = 0$ ist, folgt aus (101.11) unmittelbar, daß auch $q(\zeta)$ für $|\zeta| \to \infty$ gegen Null konvergiert. Wir müssen somit nur noch die Lage der Polstellen von q untersuchen. Auf $\mathbb{R}_0^-$ kommen nur Nullstellen von $R - 1$ als Polstellen von q in Betracht, da Q dort beschränkt bleibt. Die Nullstellen von $R - 1$ liegen jedoch bis auf den Ursprung ebenfalls in der rechten Halbebene, vgl. (79.3) im Beweis von Satz 79.3, und der Ursprung selbst ist eine einfache Nullstelle von $R - 1$, da nach Satz 77.5 $R'(0) = 1$ gilt. Betrachten wir den Zähler b^*w von q für $\zeta = 0$, so sehen wir unter Berücksichtigung von (101.6), daß

$$b^*w \;=\; \sum_{j=1}^{s} b_j W(c_j) \qquad \text{mit} \qquad W(t) \;=\; \frac{1}{s!}\int_0^t \omega(\xi)\,d\xi\,, \tag{101.12}$$

wobei ω das Knotenpolynom der s-stufigen Radau-Legendre-Formel über dem Intervall $[0, 1]$ bezeichnet. Das Innenprodukt b^*w in (101.12) entspricht gerade dieser Quadraturformel für das Integral über W, und wegen $W \in \Pi_{s+1} \subset \Pi_{2s-2}$ ist diese Quadraturformel für $s \geq 3$ exakt. Also gilt

$$\begin{aligned} b^*w \;&=\; \int_0^1 W(t)\,dt \;=\; \int_0^1 \int_0^t \omega(\xi)\,d\xi\,dt \;=\; \int_0^1 \omega(\xi) \int_\xi^1 dt\,d\xi \\ &=\; \int_0^1 \omega(\xi)(1-\xi)\,d\xi\,, \end{aligned}$$

und das letzte Integral verschwindet aufgrund der Orthogonalitätseigenschaften des Knotenpolynoms, vgl. Aufgabe VII.15. Daher kann q stetig in den Nullpunkt fortgesetzt werden und bleibt somit über $\mathbb{R}_0^-$ beschränkt. □

Bemerkung. Die Voraussetzung $s \geq 3$ in Lemma 101.2 ist wesentlich. Betrachten wir nämlich das zweistufige Radau-IIA-Verfahren aus Beispiel 79.4 mit

$$(I - \zeta A)^{-1} = \frac{6}{\zeta^2 - 4\zeta + 6} \begin{bmatrix} 1 - \zeta/4 & -\zeta/12 \\ 3\zeta/4 & 1 - 5\zeta/12 \end{bmatrix},$$

so ergibt sich

$$R(\zeta) = \frac{2\zeta + 6}{\zeta^2 - 4\zeta + 6} \quad \text{und} \quad Q(\zeta) = \Big[\frac{9/2}{\zeta^2 - 4\zeta + 6}, \frac{-\zeta + 3/2}{\zeta^2 - 4\zeta + 6} \Big]^T.$$

Auf der negativen reellen Achse ist $\|Q\|_1$ zwar durch 1 beschränkt, aber die Funktion

$$q(\zeta) = \frac{1/9}{-\zeta^2 + 6\zeta} \tag{101.13}$$

hat im Nullpunkt einen Pol und ist somit nicht beschränkt über $\mathbb{R}_0^-$. ◊

Mit den beiden Funktionen Q und q aus (101.10) kann nun eine Rekursion für den Fehler des Radau-IIA-Verfahrens für die inhomogene Testgleichung (101.4) hergeleitet werden.

Lemma 101.3. *Sei $s \geq 3$, $\varphi^{(s+2)} \in \mathcal{L}^2(0,T)$ und $\tau > 0$ die Zeitschrittweite des s-stufigen Radau-IIA-Verfahrens. Dann genügen die Fehler der numerischen Approximationen der Rekursion*

$$\begin{aligned} y_{i+1} - y(t_{i+1}) &= R(\zeta)(y_i - y(t_i)) \\ &\quad + \tau^{s+2} \lambda y^{(s+1)}(t_i) q(\zeta)(R(\zeta) - 1) + \tau^{s+3} r_i \,, \end{aligned} \tag{101.14}$$

wobei $\zeta = \tau\lambda$ und das Restglied durch

$$|r_i| \leq C_s \Big(\frac{1}{\tau} \int_{t_i}^{t_{i+1}} |y^{(s+3)}(t)|\, dt + \frac{1}{\tau} \int_{t_i}^{t_{i+1}} |\varphi^{(s+2)}(t)|\, dt \Big) \tag{101.15}$$

beschränkt ist. Die Konstante C_s hängt nur von der Stufenzahl s ab.

Beweis. Die Zwischenwerte des $(i+1)$-ten Runge-Kutta-Schritts errechnen sich aus den Gleichungen

$$\eta_j = y_i + \tau \sum_{\nu=1}^{s} a_{j\nu} \big(\lambda \eta_\nu + \varphi(t_i + c_\nu \tau) \big), \qquad j = 1, \ldots, s \,.$$

Der Ansatz

$$\eta_j = y(t_i + c_j\tau) + d_j\,, \qquad j = 1, \ldots, s\,, \tag{101.16}$$

führt dann unter Verwendung der Differentialgleichung (101.4) auf

$$\begin{aligned} d_j &= y_i - y(t_i + c_j\tau) + \tau \sum_{\nu=1}^{s} a_{j\nu}\big(\lambda y(t_i + c_\nu\tau) + \varphi(t_i + c_\nu\tau)\big) + \tau \sum_{\nu=1}^{s} a_{j\nu}\lambda d_\nu \\ &= y_i - y(t_i) + y(t_i) + \tau \sum_{\nu=1}^{s} a_{j\nu} y'(t_i + c_\nu\tau) - y(t_i + c_j\tau) + \zeta \sum_{\nu=1}^{s} a_{j\nu} d_\nu\,, \end{aligned}$$

und aus Lemma 101.1 folgt

$$d_j = y_i - y(t_i) + \tau^{s+1} y^{(s+1)}(t_i) w_j + \frac{\tau^{s+2}}{\lambda}\varepsilon_{ij} + \zeta \sum_{\nu=1}^{s} a_{j\nu} d_\nu\,, \qquad j = 1, \ldots, s\,.$$

In Vektorschreibweise entspricht dies dem Gleichungssystem

$$(I - \zeta A)d = (y_i - y(t_i))\mathbb{1} + \tau^{s+1} y^{(s+1)}(t_i)\, w + \frac{\tau^{s+2}}{\lambda}\boldsymbol{\varepsilon}_i \tag{101.17}$$

mit den s-dimensionalen Vektoren $d = [d_j]$, $w = [w_j]$ und $\boldsymbol{\varepsilon}_i = [\varepsilon_{ij}]$. Aus den Zwischenwerten ergibt sich die nächste Näherung

$$y_{i+1} = y_i + \tau \sum_{j=1}^{s} b_j\big(\lambda\eta_j + \varphi(t_i + c_j\tau)\big)$$

des Runge-Kutta-Verfahrens. Mit Hilfe von (101.16) und (101.4) führt dies auf

$$\begin{aligned} y_{i+1} &= y_i + \tau \sum_{j=1}^{s} b_j\big(\lambda y(t_i + c_j\tau) + \varphi(t_i + c_j\tau)\big) + \tau \sum_{j=1}^{s} b_j\lambda d_j \\ &= y_i + \tau \sum_{j=1}^{s} b_j y'(t_i + c_j\tau) + \zeta b^* d\,, \end{aligned}$$

und mit Lemma 101.1 folgt hieraus

$$\begin{aligned} y_{i+1} - y(t_{i+1}) &= y_i - y(t_i) + y(t_i) + \tau \sum_{j=1}^{s} b_j y'(t_i + c_j\tau) - y(t_{i+1}) + \zeta b^* d \\ &= y_i - y(t_i) + \tau^{s+3}\varepsilon_i + \zeta b^* d\,. \end{aligned}$$

Den Vektor d ersetzen wir nun mit Hilfe von (101.17): Mit den Funktionen Q und q aus (101.10) erhalten wir

$$\begin{aligned} y_{i+1} - y(t_{i+1}) \\ &= y_i - y(t_i) + \tau^{s+3}\varepsilon_i + \zeta b^*(I-\zeta A)^{-1}\mathbb{1}(y_i - y(t_i)) \\ &\qquad + \tau^{s+1}y^{(s+1)}(t_i)\zeta b^*(I-\zeta A)^{-1}w + \frac{\tau^{s+2}}{\lambda}\zeta b^*(I-\zeta A)^{-1}\boldsymbol{\varepsilon}_i \\ &= R(\zeta)(y_i - y(t_i)) + \tau^{s+2}\lambda y^{(s+1)}(t_i)q(\zeta)(R(\zeta)-1) + \tau^{s+3}r_i \end{aligned}$$

mit dem Restglied $r_i = \varepsilon_i + Q(\zeta)^*\boldsymbol{\varepsilon}_i$. Die Abschätzung (101.15) folgt dann unmittelbar aus den Lemma 101.1 und Lemma 101.2. □

Nach diesen Vorüberlegungen können wir nun den globalen Fehler des s-stufigen Radau-IIA-Verfahrens für das Anfangswertproblem (101.4) abschätzen.

Satz 101.4. *Sei $s \geq 3$, $T < \infty$ und $\varphi^{(s+2)} \in \mathcal{L}^2(0,T)$. Dann ist $y^{(s+3)} \in \mathcal{L}^2(0,T)$ und die Näherungen des s-stufigen Radau-IIA-Verfahrens für das Anfangswertproblem* (101.4) *genügen einer Abschätzung*

$$|y_i - y(t_i)| \leq C\tau^{s+2}, \qquad t_i \in [0,T],$$

wobei die Konstante C nur von der Stufenzahl s, den ersten $s+2$ Ableitungen von φ und den ersten $s+3$ Ableitungen der Lösung y abhängt.

Bemerkung. Lediglich für $s = 3$ entspricht diese Abschätzung der in Kapitel XIV eingeführten (klassischen) Ordnung $2s-1$ des s-stufigen Radau-IIA-Verfahrens. Dennoch ist diese Abschätzung bestmöglich, wenn die Fehlerkonstante von λ unabhängig sein soll. Für $s > 3$ liegt also eine *Ordnungsreduktion* vor. ◊

Beweis von Satz 101.4. Die Glattheit von y folgt aus der Lösungsdarstellung

$$y(t) = y_0 e^{\lambda t} + \int_0^t e^{\lambda(t-\xi)}\varphi(\xi)\,d\xi,$$

die man durch Variation der Konstanten erhält, vgl. (78.6). Für den Fehler des Radau-IIA-Verfahrens ergibt sich durch Induktion aus (101.14)

$$\begin{aligned} y_{i+1} - y(t_{i+1}) &= R(\zeta)(y_i - y(t_i)) + \tau^{s+2}\lambda y^{(s+1)}(t_i)q(\zeta)(R(\zeta)-1) + \tau^{s+3}r_i \\ &= \ldots = \tau^{s+2}q(\zeta)\sum_{k=0}^{i} R(\zeta)^k(R(\zeta)-1)\lambda y^{(s+1)}(t_{i-k}) + \tau^{s+3}\sum_{k=0}^{i} R(\zeta)^k r_{i-k}, \end{aligned}$$

was im folgenden in der Form

$$y_{i+1} - y(t_{i+1}) = \tau^{s+2}q(\zeta)\psi_i + \tau^{s+3}\sum_{k=0}^{i} R(\zeta)^k r_{i-k} \tag{101.18}$$

mit

$$\psi_i \;=\; \sum_{k=0}^{i} R(\zeta)^k (R(\zeta)-1)\lambda y^{(s+1)}(t_{i-k}) \tag{101.19}$$

geschrieben wird. Durch partielle Summation kann ψ_i umgeformt werden,

$$\begin{aligned}\psi_i \;&=\; R(\zeta)^{i+1}\lambda y^{(s+1)}(t_0) \;-\; \lambda y^{(s+1)}(t_i) \\ &\quad + \sum_{k=1}^{i} R(\zeta)^k \lambda\big(y^{(s+1)}(t_{i-k+1}) - y^{(s+1)}(t_{i-k})\big) \\ &= R(\zeta)^{i+1}\lambda y^{(s+1)}(t_0) \;-\; \lambda y^{(s+1)}(t_i) \;+\; \sum_{k=1}^{i} R(\zeta)^k \int_{t_{i-k}}^{t_{i-k+1}} \lambda y^{(s+2)}(t)\,dt\,,\end{aligned}$$

und da der Betrag der Stabilitätsfunktion R wegen der A-Stabilität des Radau-IIA-Verfahrens in der linken Halbebene durch Eins beschränkt ist, ergibt dies für $t_{i+1} \le T$ die obere Schranke

$$|\psi_i| \;\le\; |\lambda y^{(s+1)}(t_0)| \;+\; |\lambda y^{(s+1)}(t_i)| \;+\; \int_0^T |\lambda y^{(s+2)}(t)|\,dt\,.$$

Die Terme der Form $\lambda y^{(\nu)}$ können wegen (101.4) durch $y^{(\nu+1)} - \varphi^{(\nu)}$ ersetzt werden, so daß

$$\begin{aligned}|\psi_i| \;\le\; &|y^{(s+2)}(t_0)| \;+\; |y^{(s+2)}(t_i)| \;+\; \int_0^T |y^{(s+3)}(t)|\,dt \\ &+\; |\varphi^{(s+1)}(t_0)| \;+\; |\varphi^{(s+1)}(t_i)| \;+\; \int_0^T |\varphi^{(s+2)}(t)|\,dt\,.\end{aligned} \tag{101.20}$$

Demnach ist ψ_i durch eine Konstante c beschränkt, die nur von den ersten $s+3$ Ableitungen von y und den ersten $s+2$ Ableitungen von φ abhängt. Mit demselben c ergibt sich aus (101.15) für die Restglieder r_i die Abschätzung

$$\sum_{k=0}^{i} |r_{i-k}| \;\le\; \frac{C_s}{\tau} \int_0^{t_{i+1}} \big(|y^{(s+3)}(t)| + |\varphi^{(s+2)}(t)|\big)\,dt \;\le\; \frac{cC_s}{\tau}\,. \tag{101.21}$$

Da die Stabilitätsfunktion R durch Eins und q nach Lemma 101.2 durch c_s beschränkt ist, folgt somit aus (101.18) die gewünschte Abschätzung

$$|y_{i+1} - y(t_{i+1})| \;\le\; \tau^{s+2} c_s |\psi_i| \;+\; \tau^{s+3} \sum_{k=0}^{i} |r_{i-k}| \;\le\; (c_s + C_s) c \tau^{s+2}\,.$$

□

Nun wenden wir uns schließlich der Wärmeleitungsgleichung (101.1) zu, vereinfachen aber die folgenden Betrachtungen, indem wir wie in Beispiel 77.11 die zusätzlichen Fehler aufgrund der Ortsdiskretisierung ignorieren. Formal wenden wir also das Radau-IIA-Verfahren auf das unendlichdimensionale Anfangswertproblem (101.3) an und erhalten auf diese Weise nach i Zeitschritten eine Näherung $u_i \approx u(t_i)$, die sich ebenfalls in eine Sinusreihe entwickeln läßt:

$$u_i(x) \;=\; \sum_{j=1}^{\infty} \eta_{ij} \sin jx \,. \tag{101.22}$$

Nach Aufgabe XIV.9 ist dabei η_{ij} das Ergebnis des i-ten Zeitschrittes des Radau-IIA-Verfahrens, angewendet auf das skalare Anfangswertproblem

$$\eta_j' = -j^2 \eta_j + \varphi_j(t) \,, \qquad \eta_j(0) = \eta_{0j} \,. \tag{101.23}$$

An dieser Stelle können wir also das Ergebnis von Satz 101.4 ausnutzen.

Satz 101.5. *Es sei $s \geq 3$, $T < \infty$ und $f^{(s+2)} \in \mathcal{L}^2((0,\pi) \times [0,T])$. Dann liegt $u^{(s+3)} \in \mathcal{L}^2((0,\pi) \times [0,T])$ und die Näherungen $\{u_i\}$ aus* (101.22) *erfüllen*

$$\|u_i - u(t_i)\|_{\mathcal{L}^2(0,\pi)} = O(\tau^{s+2}) \,, \qquad t_i \in [0,T] \,.$$

Beweis. Nach unseren Vorarbeiten ist der Beweis nicht mehr schwierig. Nach Aufgabe IX.18 ist

$$\|u_i - u(t_i)\|^2_{\mathcal{L}^2(0,\pi)} \;=\; \frac{\pi}{2} \sum_{j=1}^{\infty} |\eta_{ij} - \eta_j(t_i)|^2 \,,$$

wobei η_{ij} und $\eta_j(t_i)$ die Koeffizienten der Sinusreihen (101.22) und (101.2) sind. Die Abschätzung der einzelnen Summanden dieser Reihe erfolgte in Satz 101.4. Wir benötigen hier die Darstellung (101.18) aus dessen Beweis und erhalten mit Lemma 101.2

$$\|u_{i+1} - u(t_{i+1})\|^2_{\mathcal{L}^2(0,\pi)} \;\leq\; \pi \Big(c_s^2 \sum_{j=1}^{\infty} |\psi_{ij}|^2 + \tau^2 \sum_{j=1}^{\infty} \big(\sum_{k=0}^{i} |r_{ij}| \big)^2 \Big) \tau^{2s+4} \,,$$

wobei ψ_{ij} und r_{ij} den Größen ψ_i bzw. r_i aus (101.19) und (101.14) für das j-te Anfangswertproblem (101.23) entsprechen. Diese Terme können somit durch (101.20) bzw. (101.21) abgeschätzt werden. Wir demonstrieren dies für den zweiten Summanden: Mittels (101.21) und der Cauchy-Schwarz-Ungleichung

in $\mathcal{L}^2(0,T)$ folgt für $t_{i+1} \le T$

$$\begin{aligned}
\sum_{j=1}^{\infty}\Big(\sum_{k=0}^{i}|r_{ij}|\Big)^2 &\le \frac{C_s^2}{\tau^2}\sum_{j=1}^{\infty}\Big(\int_0^T\big(|\eta_j^{(s+3)}(t)|+|\varphi_j^{(s+2)}(t)|\big)\,dt\Big)^2\\
&\le \frac{2C_s^2T}{\tau^2}\sum_{j=1}^{\infty}\Big(\int_0^T|\eta_j^{(s+3)}(t)|^2\,dt+\int_0^T|\varphi_j^{(s+2)}(t)|^2\,dt\Big)\\
&= \frac{2C_s^2T}{\tau^2}\Big(\int_0^T\sum_{j=1}^{\infty}|\eta_j^{(s+3)}(t)|^2\,dt+\int_0^T\sum_{j=1}^{\infty}|\varphi_j^{(s+2)}(t)|^2\,dt\Big)\\
&= \frac{8C_s^2T}{\pi\tau^2}\Big(\int_0^T\|u^{(s+3)}(t)\|^2_{\mathcal{L}^2(0,\pi)}\,dt+\int_0^T\|f^{(s+2)}(t)\|^2_{\mathcal{L}^2(0,\pi)}\,dt\Big).
\end{aligned}$$

Der Klammerausdruck in der letzten Zeile ist aufgrund der Voraussetzungen an f durch eine Konstante γ_2 beschränkt. Entsprechend ergibt sich $\sum|\psi_{ij}|^2 \le \gamma_1$, und somit folgt

$$\|u_{i+1}-u(t_{i+1})\|^2_{\mathcal{L}^2(0,\pi)} \le \big(\pi\gamma_1c_s^2+8\gamma_2C_s^2T\big)\tau^{2s+4}.$$

□

Bemerkungen. Das zweistufige Radau-IIA-Verfahren hat die klassische Ordnung $q=3$; hier tritt keine Ordnungsreduktion auf. Für den Beweis genügt eine um einen Term verkürzte Taylorentwicklung der Defekte in Lemma 101.1.

Bei den Gauß-Verfahren aus Abschnitt 78 reduziert sich für $s \ge 2$ die Ordnung entsprechend von $q=2s$ auf $s+2$. Der Beweis verläuft völlig analog, lediglich in Lemma 101.2 ist zu beachten, daß der Nenner von q für $|\zeta| \to \infty$ eine Nullstelle besitzen kann. Dennoch ist q über $\mathbb{R}_0^-$ beschränkt, vgl. Aufgabe 8.

Für $s \ge 3$ stimmen also die reduzierten Ordnungen der Gauß- und Radau-IIA-Verfahren überein. Wegen der L-Stabilität sind daher die Radau-IIA-Verfahren bei parabolischen Differentialgleichungen vorzuziehen. Für einen Vergleich der zweistufigen Gauß- und Radau-IIA-Verfahren sei auf das nachfolgende Beispiel verwiesen. ◊

In Satz 101.5 wird die Ortsdiskretisierung außer Betracht gelassen. Man kann jedoch zeigen, daß mit der Linienmethode die Fehlerabschätzung um einen zusätzlichen Term $O(h^2)$ für die Approximation des Galerkin-Verfahrens ergänzt werden muß. Die Ergebnisse lassen sich zudem auf die allgemeineren parabolischen Anfangsrandwertaufgaben dieses Kapitels übertragen. Die entsprechenden Beweise finden sich in dem Buch von Thomée [102].

Beispiel. Abbildung 101.1 illustriert die höhere Konsistenzordnung der zweistufigen Gauß- und Radau-IIA-Verfahren gegenüber dem Crank-Nicolson- und dem Euler-Verfahren am Beispiel der eindimensionalen homogenen Wärmeleitungsgleichung. Als Anfangsvorgabe u° dient der kubischen B-Spline aus

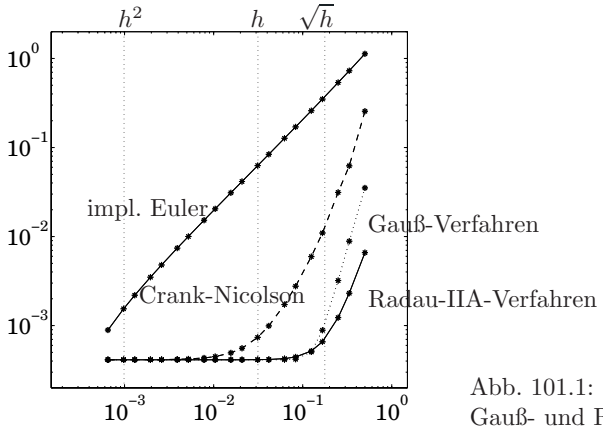


Abb. 101.1:
Gauß- und Radau-IIA-Verfahren

Aufgabe IX.19. Die Kurven zeigen die relativen Fehler der vier Verfahren in Abhängigkeit von der Zeitschrittweite. Als Maßstab dient die $\mathcal{L}^2$-Norm des Fehlers zum Zeitpunkt $t = 4$.

Die höhere Konsistenzordnung macht sich dadurch bemerkbar, daß das Gauß- und das Radau-IIA-Verfahren die durch das Galerkin-Verfahren limitierte Genauigkeit $O(h^2)$ bereits für $\tau \approx \sqrt{h}$ erreichen. Interessant ist ein Vergleich zwischen Gauß- und Radau-IIA-Verfahren: An der Steigung der Fehlerkurven für $\tau > \sqrt{h}$ erkennt man die höhere Konsistenzordnung des Gauß-Verfahrens. Dennoch ist das Radau-IIA-Verfahren für große Zeitschrittweiten überlegen, und zwar wegen seiner L-Stabilität. ◊

102 Eine quasilineare Diffusionsgleichung

Die bisher betrachteten Diffusionsgleichungen waren allesamt linear. Wir betrachten nun noch ausgewählte numerische Verfahren zur Lösung *quasilinearer* Anfangsrandwertaufgaben der Gestalt

$$\begin{aligned} u_t - \operatorname{div}\bigl(\sigma(u)\operatorname{grad} u\bigr) \;&=\; f(u)\,, \qquad u(0) = u^\circ\,, \\ u(x,t) = 0 \quad &\text{für } x \in \Gamma \text{ und } t \geq 0\,, \end{aligned} \tag{102.1}$$

bei der der Diffusionskoeffizient σ und der Quellterm f von der Lösung u abhängen. Solche Gleichungen treten beispielsweise bei Grundwasserströmungen auf, vgl. Beispiel 70.2. Wie im linearen Fall sei

$$0 < \sigma_0 \leq \sigma(\omega) \leq \sigma_\infty \qquad \text{für alle } \omega \in \mathbb{R}\,.$$

Darüber hinaus sei vorausgesetzt, daß σ und f stetig differenzierbar sind.

Die schwache Form der Differentialgleichung (102.1) ergibt sich völlig analog zum linearen Fall, nämlich

$$\langle\, u'(t), w \,\rangle_{\mathcal{L}^2(\Omega)} + a(u(t), w) = \langle\, f(u(t)), w \,\rangle_{\mathcal{L}^2(\Omega)} \tag{102.2}$$

für alle $w \in H_0^1(\Omega)$. Dabei ist zu beachten, daß sowohl die rechte Seite von (102.2) als auch die Bilinearform

$$a(v, w) = \int_\Omega \sigma(u(t)) \operatorname{grad} v \cdot \operatorname{grad} w \, dx$$

von der exakten Lösung $u(t)$ und damit indirekt auch von der Zeit abhängen.

Die Linienmethode läßt sich ohne größere Schwierigkeiten auf den hier vorliegenden Fall übertragen: Gesucht ist eine Funktion $u_h : V_0^{\mathcal{T}} \times [0, T] \to \mathbb{R}$, die für alle $w \in V_0^{\mathcal{T}}$ das Variationsproblem

$$\langle\, u_h'(t), w \,\rangle_{\mathcal{L}^2(\Omega)} + a_h(u_h(t), w) = \langle\, f(u_h(t)), w \,\rangle_{\mathcal{L}^2(\Omega)} \tag{102.3}$$

löst, wobei die Bilinearform a_h nun über die Approximation u_h definiert ist:

$$a_h(v, w) = \int_\Omega \sigma(u_h(t)) \operatorname{grad} v \cdot \operatorname{grad} w \, dx\,.$$

Unter Vorgabe einer Approximation $u_h(0) \approx u^\circ$ kann die Existenz und Eindeutigkeit einer solchen Näherungslösung wie in Abschnitt 98 nachgewiesen werden, vgl. Aufgabe 9.

Beispiel 102.1. Wir betrachten eine Variante der *Poröse-Medien-Gleichung* aus Beispiel 70.2,

$$u_t = \frac{1}{3}(u^3)_{xx} = (u^2 u_x)_x\,, \qquad |x| < 1\,, \quad t \geq 0\,. \tag{102.4}$$

In diesem Beispiel ist $\sigma(u) = u^2$, d. h. die Elliptizität des Operators ist nur solange gewährleistet, wie die Lösung positiv bleibt. Dies ist für die Anfangs- und Randvorgaben

$$\begin{aligned} u(x, 0) &= u^\circ(x) = \big(1 - (x/2)^2\big)^{1/2}\,, && |x| \leq 1\,, \\ u(\pm 1, t) &= g(t) = \chi(t)\big(1 - \chi^2(t)/4\big)^{1/2}\,, && t \geq 0\,, \end{aligned}$$

mit $\chi(t) = (1 + t)^{-1/4}$ in dem gesamten Halbstreifen $|x| \leq 1$, $t \geq 0$, der Fall. Die zugehörige Lösung lautet

$$u(x, t) = \chi(t)\big(1 - x^2\chi^2(t)/4\big)^{1/2}\,. \tag{102.5}$$

Man beachte, daß die Randvorgabe g von Null verschieden ist; es handelt sich also hierbei um eine inhomogene Dirichlet-Randbedingung. Für $u_h \in V^{\mathcal{T}}$ wählen wir daher analog zu Abschnitt 89.1 einen Ansatz

$$u_h(x,t) \;=\; \sum_{j=0}^{n+1} \eta_j(t) \Lambda_j(x)\,,$$

wobei die Hutfunktionen Λ_0 und Λ_{n+1} zu den Randpunkten $x_0 = -1$ und $x_{n+1} = 1$ gehören und die zugehörigen Koeffizienten somit durch

$$\eta_0(t) = \eta_{n+1}(t) = g(t)\,, \qquad t \geq 0\,,$$

festgelegt sind. Die verbleibenden Koeffizienten werden wie vorher in dem Vektor $y(t) = \big[\eta_j(t)\big]_{j=1}^{n} \in \mathbb{R}^n$ gesammelt. Für y ergibt sich auf diese Weise analog zu (102.3) das Anfangswertproblem

$$Gy' + A(t,y)y = b(t,y)\,, \qquad y(0) = \big[\,\eta_j(0)\,\big]_j\,, \tag{102.6}$$

mit $A(t,y) = [a_{ij}(t,y)]_{ij} \in \mathbb{R}^{n\times n}$ und $b(t,y) = [b_j(t,y)]_j \in \mathbb{R}^n$,

$$\begin{aligned} a_{ij}(t,y) \;&=\; \int_\Omega \sigma(u_h(t)) \operatorname{grad}\Lambda_i \cdot \operatorname{grad}\Lambda_j \, dx\,, \\ b_j(t,y) \;&=\; -g(t) \int_\Omega \sigma(u_h(t)) \operatorname{grad}\Lambda_j \cdot \operatorname{grad}(\Lambda_0 + \Lambda_{n+1}) \, dx \\ &\qquad -\; g'(t) \int_\Omega \Lambda_j(\Lambda_0 + \Lambda_{n+1}) \, dx\,. \end{aligned}$$

Wie zuvor ist $G \in \mathbb{R}^{n\times n}$ die Gramsche Matrix der nodalen Basis $\{\Lambda_1, \ldots, \Lambda_n\}$ von $V_0^{\mathcal{T}}$.

Obwohl in diesem Beispiel kein Quellterm vorliegt ($f = 0$), ergibt sich aufgrund der inhomogenen Dirichlet-Randvorgabe auf der rechten Seite von (102.6) ein Term, der die Rolle des Quellterms b aus (98.9) übernimmt. Weiter ist zu beachten, daß die Differentialgleichung (102.6) aufgrund der inhomogenen Dirichlet-Randvorgabe nicht mehr in autonomer Form vorliegt; bei der Matrix A entsteht beispielsweise eine *explizite* Abhängigkeit von t durch die einfließenden Randwerte $\eta_0 = \eta_n = g(t)$ im Argument von σ. ◊

Im Rest dieses Abschnitts diskutieren wir verschiedene numerische Verfahren zweiter Ordnung zur Approximation der Lösung von (102.6), wobei die spezielle Form von b ohne Bedeutung ist. Die Verfahren lassen sich also in entsprechender Weise auf quasilineare Differentialgleichungen mit inhomogenen Quelltermen und homogener Dirichlet-Randbedingung übertragen.

Die Verfahren sind bis auf eine Ausnahme aus den vorangegangenen Abschnitten oder aus Kapitel XIV bekannt. Wir konzentrieren uns daher besonders auf Implementierungsdetails. Um den Aufwand der Verfahren zu vergleichen, wird die Anzahl der zu lösenden linearen Gleichungssysteme gezählt. Für das eindimensionale Beispiel 102.1 sind dies zwar allesamt Tridiagonalsysteme, deren numerische Lösung sehr billig ist, bei höherdimensionalen Problemen dominieren jedoch die Kosten dieser Teilaufgaben die Gesamtkosten eines Zeitschritts.

102.1 Das Crank-Nicolson-Verfahren

Wir beginnen mit dem *Crank-Nicolson-Verfahren*, also dem impliziten Mittelpunktverfahren für das Differentialgleichungssystem (102.6):

$$\begin{aligned} G(y_{i+1} - y_i) &= \tau b(t_{i+1/2}, y_{i+1/2}) - \tau A(t_{i+1/2}, y_{i+1/2}) y_{i+1/2}\,, \\ y_{i+1/2} &= (y_i + y_{i+1})/2\,, \qquad t_{i+1/2} = t_i + \tau/2\,, \end{aligned} \tag{102.7}$$

$i = 0, 1, 2, \ldots$, wobei wir wieder die Notation der vorangegangenen Abschnitte verwenden.

Im Gegensatz zu Abschnitt 99 führt die Rekursion (102.7) des Crank-Nicolson-Verfahrens in jedem Zeitschritt auf die Lösung eines *nichtlinearen* Gleichungssystems. In Kapitel XIV wurde zur Lösung dieser Gleichungssysteme ein vereinfachtes Newton-Verfahren empfohlen, vgl. Algorithmus 78.1, für das in jedem Zeitschritt die Jacobi-Matrix

$$J(t, y) = \frac{\partial}{\partial y} \Big(b(t, y) - A(t, y) y \Big) \in \mathbb{R}^{n \times n} \tag{102.8}$$

an der Stelle $(t, y) = (t_i, y_i)$ zu bestimmen ist. Algorithmus 102.1 ist eine Adaption von Algorithmus 78.1 auf das hier vorliegende Problem. Man beachte allerdings die leichte Modifikation der Notation: Die Matrix J aus Algorithmus 78.1 ist durch das Produkt $G^{-1}J(t, y)$ mit $J(t, y)$ aus (102.8) zu ersetzen. In der vorgeschlagenen Realisierung aus Algorithmus 102.1 ist jedoch keine explizite Berechnung von G^{-1} erforderlich.

Aufwand. Der Aufwand von Algorithmus 102.1 hängt wesentlich von der Anzahl der Newton-Iterationen je Zeitschritt ab. In Beispiel 102.1 reicht bei einer vernünftigen Zeitschrittweite in der Regel ein Newton-Schritt aus, so daß insgesamt zwei lineare Gleichungssysteme zur Berechnung von $k^{(0)}$ und $z^{(0)}$ in jedem Zeitschritt zu lösen sind. Der Aufwand kann dabei mit der Lumping-Technik aus Abschnitt 100 um die Hälfte reduziert werden, da bei dieser Variante G durch eine Diagonalmatrix ersetzt wird und die Kosten zur Berechnung von $k^{(0)}$ vernachlässigt werden können. ◊

Initialisierung: y_0, $t_0 = 0$ und Schrittweite τ seien gegeben
  `for` $i = 0, 1, 2, \ldots$ `do`
    $J = J(t_i, y_i)$ % vgl. (102.8)
    $b = b(t_i, y_i)$
    $A = A(t_i, y_i)$
    $Gk^{(0)} = b - Ay_i$
    $\varepsilon_0 = \left(k^{(0)*} Gk^{(0)}\right)^{1/2}$ % für Abbruchbedingung (s.u.)
    `for` $m = 0, 1, 2, \ldots$ `do` % vereinfachte Newton-Iteration
      $y_{i+1/2} = y_i + \tau/2\, k^{(m)}$
      $b = b(t_{i+1/2}, y_{i+1/2})$
      $A = A(t_{i+1/2}, y_{i+1/2})$
      $(G - \tau/2\, J) z^{(m)} = b - Ay_{i+1/2} - Gk^{(m)}$
      $k^{(m+1)} = k^{(m)} + z^{(m)}$
      $\varepsilon_{m+1} = \left(z^{(m)*} Gz^{(m)}\right)^{1/2}$ % entspricht $\mathcal{L}^2$-Norm der zugehörigen Funktion
    `until` $\varepsilon_{m+1}^2/(\varepsilon_m - \varepsilon_{m+1}) \le \tau^2 \varepsilon_0$ % end for (m-Schleife)
    $y_{i+1} = y_i + \tau k^{(m+1)}$
    $t_{i+1} = t_i + \tau$
  `until` $t_{i+1} \ge T$ % end for (i-Schleife)
Ergebnis: $y_i \approx y(t_i), \quad i = 0, 1, 2 \ldots$

Algorithmus 102.1: Crank-Nicolson-Verfahren

102.2 Ein linearisiertes Crank-Nicolson-Verfahren

Die Nichtlinearität des Crank-Nicolson-Verfahrens (102.7) liegt an der unbekannten Näherung $y_{i+1/2}$ im Argument von b und A. Es liegt daher nahe, an diesen beiden Stellen eine Approximation $\tilde{y}_{i+1/2}$ von $y_{i+1/2}$ zu verwenden:

$$G(y_{i+1} - y_i) \;=\; \tau b(t_{i+1/2}, \tilde{y}_{i+1/2}) - \frac{\tau}{2} A(t_{i+1/2}, \tilde{y}_{i+1/2})(y_{i+1} + y_i)\,.$$

Um die Konsistenzordnung des Verfahrens zu erhalten, darf dabei maximal ein zusätzlicher Fehler $O(\tau^2)$ entstehen.

In dem Buch von Thomée [102, S. 218 ff] findet sich diesbezüglich der Vorschlag, für $i \ge 1$ zunächst y_{i+1} durch Extrapolation aus y_i und y_{i-1} zu approximieren,

$$\tilde{y}_{i+1} \;=\; 2y_i - y_{i-1} \;=\; y_{i+1} + O(\tau^2)\,,$$

und dann

$$\tilde{y}_{i+1/2} = (y_i + \tilde{y}_{i+1})/2 = (3y_i - y_{i-1})/2\,, \qquad i \ge 1\,, \tag{102.9a}$$

zu setzen. Die Rekursion lautet dann

$$\begin{aligned} (G + \frac{\tau}{2}A_{i+1/2})y_{i+1} \;&=\; (G - \frac{\tau}{2}A_{i+1/2})y_i + \tau b_{i+1/2}\,, \\ A_{i+1/2} = A(t_{i+1/2}, \tilde{y}_{i+1/2})\,, &\qquad b_{i+1/2} = b(t_{i+1/2}, \tilde{y}_{i+1/2})\,, \end{aligned} \tag{102.9b}$$

$i = 0, 1, 2, \ldots$ Da bei der Berechnung von y_{i+1} neben der aktuellen Näherung y_i auch noch die vorangegangene Näherung y_{i-1} verwendet wird, ist das Schema (102.9) kein echtes Einschrittverfahren mehr. Zudem muß die Rechenvorschrift (102.9a) im ersten Zeitschritt ($i = 0$) modifiziert werden, da kein y_{-1} zur Verfügung steht. Denkbar ist etwa, zunächst in einem „Vorlaufschritt“ (*Prädiktorschritt*) mit einem anderen Verfahren eine geeignete Näherung $\tilde{y}_1 \approx y(\tau)$ zu berechnen, um dann im eigentlichen ersten Schritt (102.9b) mit $i = 0$ das Argument

$$\tilde{y}_{1/2} = (y_0 + \tilde{y}_1)/2$$

zu verwenden. Für das numerische Beispiel in Abschnitt 102.4 folgen wir dem Vorschlag aus [102] und verwenden hierfür das Crank-Nicolson-Verfahren mit den „eingefrorenen“ Approximationen $A \approx A(0, y_0)$ und $b \approx b(0, y_0)$:

$$\big(G + \frac{\tau}{2}A(0, y_0)\big)\tilde{y}_1 \;=\; \big(G - \frac{\tau}{2}A(0, y_0)\big)y_0 \;+\; \tau b(0, y_0)\,.$$

Aufwand. Für das linearisierte Crank-Nicolson-Verfahren (102.9) ist (abgesehen von dem Prädiktorschritt) jeweils ein lineares Gleichungssystem pro Zeitschritt zu lösen. ◊

102.3 Das linear-implizite Mittelpunktverfahren

Die nichtlinearen Gleichungssysteme des Crank-Nicolson-Verfahrens können auch mit dem *linear-impliziten Mittelpunktverfahren* umgangen werden. Dabei ist allerdings zu beachten, daß das Verfahren in der in Kapitel XIV eingeführten Form (80.4) nur für autonome Differentialgleichungen die Konsistenzordnung Zwei besitzt. Da die Gleichung (102.6) nicht autonom ist, muß das linear-implizite Mittelpunktverfahren in der Variante aus Aufgabe XIV.14 verwendet werden.

Hierzu ist die Zeitableitung

$$\psi(t, y) \;=\; \frac{\partial}{\partial t}\Big(b(t, y) - A(t, y)y\Big) \;\in\; \mathbb{R}^n \tag{102.10}$$

auszuwerten, mit der dann der $(i+1)$-te Zeitschritt dieses Verfahrens durch

$$\begin{aligned} (G - \frac{\tau}{2}J)k \;&=\; b(t_i, y_i) - A(t_i, y_i)y_i + \frac{\tau}{2}\,\psi(t_i, y_i)\,, \\ y_{i+1} \;&=\; y_i + \tau k\,, \end{aligned} \tag{102.11}$$

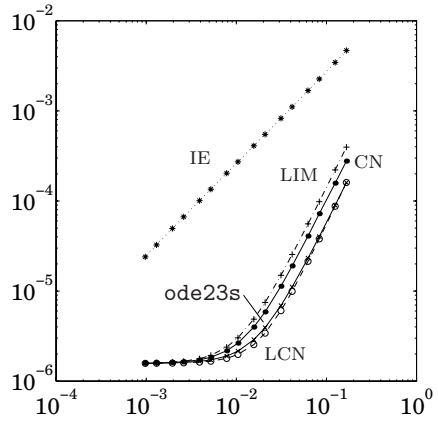


Abb. 102.1:
Ergebnisse der verschiedenen Verfahren für Beispiel 102.1

gegeben ist, wobei $J = J(t_i, y_i)$ wieder die Jacobi-Matrix aus (102.8) bezeichnet.

Aufwand. Der Aufwand des linear-impliziten Mittelpunktverfahrens ist im wesentlichen derselbe wie der des linearisierten Crank-Nicolson-Verfahrens: In jedem Zeitschritt ist ein lineares Gleichungssystem zu lösen. Das linear-implizite Verfahren ist jedoch etwas aufwendiger zu programmieren, da neben der Jacobi-Matrix J noch die Zeitableitung ψ in jedem Zeitschritt ausgewertet werden muß. Insbesondere werden hierfür die zweiten Ableitungen der Randvorgabe g benötigt. ◊

102.4 Numerische Resultate

Abbildung 102.1 zeigt die numerischen Resultate der drei Verfahren für Beispiel 102.1. Aufgetragen ist der relative $\mathcal{L}^2$-Fehler zum Zeitpunkt $t = 1$ über der Zeitschrittweite τ (die Ortsdiskretisierung mit $h = 2/100$ ist immer dieselbe). Zum Vergleich enthält die Abbildung noch die Ergebnisse des impliziten Euler-Verfahrens (IE) und des Rosenbrock-Typ-Verfahrens `ode23s` (durchgezogene Kurve mit Kreuzen) aus Beispiel 80.2. Letzteres wurde für dieses Beispiel wie die anderen Verfahren mit konstanter Zeitschrittweite, d. h. ohne Schrittweitensteuerung implementiert. Dabei ist zu berücksichtigen, daß der Aufwand von `ode23s` größer ist, da in jedem Zeitschritt *zwei* lineare Gleichungssysteme (mit derselben Matrix) zu lösen sind, während bei den anderen Verfahren im wesentlichen ein lineares Gleichungssystem gelöst werden muß. Für eine Anwendung von `ode23s` muß die Differentialgleichung wie beim linear-impliziten Mittelpunktverfahren zunächst in autonome Form transformiert werden.

Das Crank-Nicolson-Verfahren (CN), das linearisierte Crank-Nicolson-Verfah-

ren (LCN), das linear-implizite Mittelpunktverfahren (LIM, gebrochene Kurve mit Kreuzen) und `ode23s` liefern ähnliche Ergebnisse; die entsprechenden Kurven weisen allesamt eine Konvergenzgeschwindigkeit $O(h^2 + \tau^2)$ auf; für entsprechende theoretische Resultate sei auf das Buch von Thomée [102] verwiesen. Erwartungsgemäß müssen hingegen beim Euler-Verfahren für vergleichbare Genauigkeiten deutlich kleinere Zeitschritte gewählt werden.

103 Schrittweitensteuerung und adaptive Gitter

Wir beschreiben nun noch eine Möglichkeit, wie Algorithmen zur Schrittweitensteuerung für Runge-Kutta-Verfahren und Fehlerschätzer für elliptische Differentialgleichungen in numerische Verfahren zur Lösung parabolischer Anfangsrandwertprobleme integriert werden können. Ziel ist ein vollständig adaptives Programm, bei dem sowohl die Zeitschrittweite als auch die Triangulierung des Gebiets während der Rechnung optimiert werden, um eine vorgeschriebene Genauigkeit

$$\|u_i - u(t_i)\|_{\mathcal{L}^2(\Omega)} \leq \epsilon$$

für alle Zeitschritte $0 \leq t_i \leq T$ mit geringstmöglichem Rechenaufwand zu gewährleisten.

Zur Motivation des nachfolgenden Ansatzes schätzen wir den Fehler mit der Dreiecksungleichung ab,

$$\|u_i - u(t_i)\|_{\mathcal{L}^2(\Omega)} \leq \|u_i - u_h(t_i)\|_{\mathcal{L}^2(\Omega)} + \|u_h(t_i) - u(t_i)\|_{\mathcal{L}^2(\Omega)}, \tag{103.1}$$

wobei $u_h(t)$ die Näherung aus der Linienmethode bezeichnet. Der zweite Fehlerterm auf der rechten Seite von (103.1) hängt ausschließlich von der Güte der Triangulierung ab, vgl. Satz 98.1. Der erste Term ist der Fehler der Zeitintegration: Wird hierfür ein A-stabiles Runge-Kutta-Verfahren mit einem verläßlichen eingebetteten Fehlerschätzer und korrekt eingestellten Fehlertoleranzen verwendet, so können wir davon ausgehen, daß dieser Term die gewünschte Genauigkeit aufweist. Im Umkehrschluß kann daher gefolgert werden, daß die Ursache für eine ungenaue Näherung u_i in einer mangelhaften Triangulierung des Gebiets zu suchen ist. Mit dieser Argumentation wird die Zeitadaptivität vollständig in die Verantwortung des Runge-Kutta-Verfahrens verschoben. Die Initialisierung der entsprechenden Fehlertoleranzen muß lediglich eine Abschätzung

$$\|u_i - u_h(t_i)\|_{\mathcal{L}^2(\Omega)} \ll \epsilon \tag{103.2}$$

sicherstellen.

Es bleibt also zu klären, wie die Triangulierung an das Problem anzupassen ist, so daß auch für den letzten Term in (103.1) eine Fehlertoleranz ϵ gewährleistet wird. Dies beschreiben wir im folgenden anhand der eindimensionalen Wärmeleitungsgleichung

$$u_t - u_{xx} = f \quad \text{in } (0,\pi) \times [0,T]\,, \qquad u(0) = u(\pi) = 0\,,$$

mit einer zeitabhängigen Wärmequelle $f = f(x,t)$. Die Übertragung auf die allgemeineren parabolischen Anfangsrandwertaufgaben dieses Kapitels bereitet keine nennenswerten prinzipiellen Schwierigkeiten.

Wir fixieren einen festen Zeitpunkt t_i, an dem die Güte des aktuellen Gitters $\triangle = \{x_j : j = 0,\dots,m = n+1\}$ anhand der Genauigkeit der Näherungslösung $u_i \approx u(t_i)$ überprüft werden soll. In Anbetracht der Vorüberlegungen gehen wir davon aus, daß der erste Term auf der rechten Seite von (103.1) vernachlässigbar ist, also daß

$$\|u_i - u(t_i)\|_{\mathcal{L}^2(0,\pi)} \approx \|u_h(t_i) - u(t_i)\|_{\mathcal{L}^2(0,\pi)} \tag{103.3}$$

gilt. Um hier die rechte Seite abzuschätzen, nutzen wir aus, daß $v = u(t_i)$ die elliptische Differentialgleichung

$$-v_{xx} = F \qquad \text{mit } F = f(t_i) - u_t(t_i) \tag{103.4}$$

löst, während $u_h(t_i)$ gemäß der Linienmethode das Variationsproblem

$$a(u_h(t_i), w) = \langle\, f(t_i) - u_h'(t_i), w \,\rangle_{\mathcal{L}^2(0,\pi)} \qquad \text{für alle } w \in V_0^{\mathcal{T}} \tag{103.5}$$

erfüllt.

Für die eindimensionale Wärmeleitungsgleichung liefert daher die Methode aus Abschnitt 96 die Abschätzung

$$\|u_h(t_i) - u(t_i)\|^2_{H^1(0,\pi)} \le C \sum_{j=1}^{m} h_j^2 \|f(t_i) - u_h'(t_i)\|^2_{\mathcal{L}^2(\mathcal{I}_j)} \tag{103.6}$$

für den H^1-Fehler (vgl. Aufgabe XVI.11), wobei h_j die Länge des Gitterintervalls $\mathcal{I}_j = (x_{j-1}, x_j)$ angibt. Da die Wärmeleitungsgleichung die Voraussetzungen des Regularitätssatzes 92.2 erfüllt, lassen wir uns von Abschnitt 92 leiten und setzen für den $\mathcal{L}^2$-Fehler die Größenordnung

$$\|u_h(t_i) - u(t_i)\|_{\mathcal{L}^2(0,\pi)} \lesssim c\|u_h(t_i) - u(t_i)\|^2_{H^1(0,\pi)}$$

an. Auf dieser Grundlage erhalten wir also aus (103.3) und (103.6) den Fehlerschätzer

$$\|u_i - u(t_i)\|_{\mathcal{L}^2(0,\pi)} \lesssim C \sum_{j=1}^{m} h_j^2 \|f(t_i) - u_h'(t_i)\|^2_{\mathcal{L}^2(\mathcal{I}_j)}\,, \tag{103.7}$$

dessen rechte Seite allerdings nicht ausgewertet werden kann, da – abgesehen von der neuen Konstante $C > 0$ – die Zeitableitung $u_h'(t_i)$ nicht bekannt ist. Wir ersetzen daher $u_h'(t_i)$ durch die „Steigung“ $u_i' \in V_0^{\mathcal{T}}$ aus dem Richtungsfeld der gewöhnlichen Differentialgleichung im Punkt $(t_i, u(t_i))$: Diese Steigung erfüllt das Variationsproblem

$$\langle u_i', w \rangle_{\mathcal{L}^2(0,\pi)} = \langle f(t_i), w \rangle_{\mathcal{L}^2(0,\pi)} - a(u_i, w) \quad \text{für alle } w \in V_0^{\mathcal{T}} \tag{103.8}$$

in Analogie zu (103.5). Eingesetzt in (103.7) erhalten wir schließlich den berechenbaren Fehlerschätzer

$$\|u_i - u(t_i)\|_{\mathcal{L}^2(0,\pi)} \lesssim C \sum_{j=1}^{m} h_j^2 \|f(t_i) - u_i'\|^2_{\mathcal{L}^2(\mathcal{I}_j)}, \tag{103.9}$$

dessen einzelne Summanden

$$\epsilon_j = h_j^2 \|f(t_i) - u_i'\|^2_{\mathcal{L}^2(\mathcal{I}_j)}, \qquad j = 1, \ldots, m,$$

als Beiträge aus den einzelnen Gitterintervallen zum Gesamtfehler aufgefaßt werden können.

Ist die Summe auf der rechten Seite von (103.9) größer als die vorgegebene Fehlertoleranz ϵ, so kann dies gemäß der Vorüberlegung nur an dem aktuellen Gitter $\triangle$ liegen. Wir fassen daher die Werte ϵ_j als Fehlerindikatoren für die einzelnen Gitterintervalle $\mathcal{I}_j$ auf und streben einen Zustand an, bei dem alle Fehlerindikatoren gleich groß sind. Um diesen Zustand zu erreichen, wählen wir zwei Toleranzparameter $c_1 < 1 < c_2$ (etwa $c_1 = 1/2$ und $c_2 = 2$) und modifizieren das Gitter nach den folgenden Regeln:

- *Verfeinerung*:
Gitterintervalle $\mathcal{I}_j$ mit $\epsilon_j > c_2\epsilon/m$ werden durch n_j zusätzliche (äquidistant verteilte) Gitterpunkte unterteilt. Die Heuristik, daß der Fehler proportional zu h^2 fällt, führt auf den Schätzwert $\epsilon_j/(n_j+1)^2$ für den resultierenden Fehlerindikator nach der Verfeinerung. Daher sollte etwa $n_j \approx (m\epsilon_j/\epsilon)^{1/2}$ gelten.
- *Vergröberung*:
Gitterpunkte x_j, für die die beiden Fehlerindikatoren ϵ_j und ϵ_{j+1} der benachbarten Gitterintervalle unterhalb der Schranke $c_1\epsilon/m$ liegen, werden aus dem Gitter entfernt.

Wenn die Fehlertoleranz ϵ verletzt war und das Gitter entsprechend angepaßt worden ist, wird der letzte Zeitschritt mit dem neuen Gitter wiederholt, d. h. das Runge-Kutta-Verfahren wird zum Zeitpunkt t_{i-1} neu gestartet. Generell ist zu beachten, daß im Fall einer Vergröberung des Gitters die aktuelle Runge-Kutta-Näherung nicht mehr in dem neuen Finite-Elemente-Raum V_h liegt. Statt dessen verwenden wir ihre Bestapproximation (bezüglich der $\mathcal{L}^2$-Norm) aus V_h.

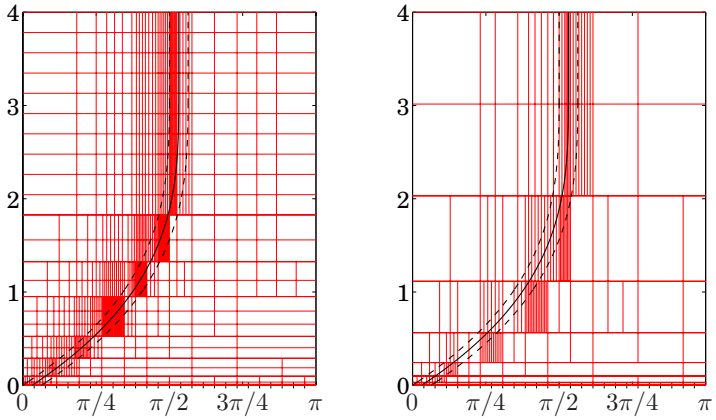


Abb. 103.1: Das örtliche Gitter $\triangle$ aus Beispiel 103.1 in Abhängigkeit von der Zeit: links für `ode23s`, rechts für `radau5`

Beispiel 103.1. Wir betrachten die Wärmeleitungsgleichung (103.4) mit $u^\circ = 0$ und der Wärmequelle

$$f(x,t) \;=\; \frac{1}{\delta}\,\Lambda\Big(\,\frac{x-\psi(t)}{\delta}\,\Big)\,,$$

wobei Λ die symmetrische Hutfunktion mit Höhe Eins und Träger $[0,2]$ bezeichnet und

$$\psi(t) \;=\; \begin{cases} \dfrac{\pi}{2} \;-\; \dfrac{(\pi-t)^3}{2\pi^2}\,, & 0 \le t \le \pi\,, \\[2ex] \dfrac{\pi}{2}\,, & \pi < t\,, \end{cases} \qquad .$$

den linken Randpunkt von $\operatorname{supp}(f(t))$ angibt. Der Parameter δ reguliert die örtliche Ausbreitung der Wärmequelle; für dieses Beispiel wählen wir $\delta = 1/10$.

Man kann sich unter der Lösung u dieser Differentialgleichung etwa die Wärmeverteilung in einem Stab vorstellen, der an beiden Enden auf einer konstanten Temperatur gehalten wird, während er gleichzeitig durch ein Feuerzeug erwärmt wird, das vom linken Rand zur Stabmitte wandert.

Bei der Lösung der Differentialgleichung mit der oben beschriebenen Methode vergleichen wir das Rosenbrock-Typ-Verfahren `ode23s` mit dem Runge-Kutta-Verfahren `radau5` aus Beispiel 79.4 und den jeweils zugehörigen Algorithmen zur adaptiven Steuerung der Zeitschrittweite. Das Ortsgitter wird in beiden Fällen so angepaßt, daß die Fehlertoleranz $\epsilon = 5 \cdot 10^{-3}$ eingehalten wird. Abbildung 103.1 zeigt die resultierenden Gitter in der (x,t)-Ebene: Die horizontalen Linien veranschaulichen die gewählten Zeitschritte, während die vertikalen

Linien die entsprechenden Ortsgitter in Abhängigkeit von der Zeit darstellen. Zum besseren Verständnis beachte man die fett eingezeichneten Kurven, die die Bewegung der Wärmequelle nachvollziehen; die Spitze der Hutfunktion f bewegt sich entlang der durchgezogenen Kurve. Man kann gut erkennen, daß das Gitter in der Nähe der Wärmequelle zunächst verfeinert und dann auch wieder vergröbert wird, wenn die Wärmequelle weitergewandert ist.

An diesem Beispiel läßt sich auch erkennen, daß der Fehlerschätzer von `ode23s` nicht A-stabil ist. Obwohl sich die Wärmequelle ab $t = \pi$ nicht mehr bewegt und die Lösung u ab diesem Zeitpunkt sehr schnell gegen einen stationären Zustand konvergiert (vgl. Aufgabe 1), werden die Zeitschritte von `ode23s` für $t > \pi$ nicht mehr größer. Das L-stabile Radau-IIA-Verfahren nutzt diesen Umstand hingegen aus und wählt rasch anwachsende Zeitschritte, ohne dabei die Genauigkeitsanforderungen zu verletzen. ◊

Abschließend sei darauf hingewiesen, daß die Programmierung adaptiver Gitter in höheren Dimensionen wesentlich aufwendiger wird, da sowohl bei den Verfeinerungen als auch den Vergröberungen darauf geachtet werden muß, daß alle Triangulierungen regulär sind. Entsprechende Algorithmen werden zum Beispiel in der Monographie von Lang [67] diskutiert.

Aufgaben

1. Gegeben sei das parabolische Anfangsrandwertproblem $u' + L[u] = f$ für $x \in \Omega$ und $t \geq 0$ mit $u(0) = u^\circ$ und homogenen Dirchlet-Randwerten auf $\Gamma = \partial\Omega$. Die rechte Seite f hänge dabei nur von x und nicht von der Zeit t ab. $v \in H_0^1(\Omega)$ bezeichne die schwache Lösung des elliptischen Randwertproblems $L[v] = f$. Beweisen Sie, daß

$$\|u(t) - v\|_{\mathcal{L}^2(\Omega)} \longrightarrow 0 \qquad \text{für } t \to \infty\,.$$

Hinweis: Betrachten Sie das Funktional

$$\frac{1}{2}\frac{d}{dt}\,\|u(t) - v\|^2_{\mathcal{L}^2(\Omega)} + a(u(t) - v, u(t) - v)\,.$$

2. Sei u die schwache Lösung des Anfangsrandwertproblems (97.5) mit Anfangswerten $u^\circ \in \mathcal{L}^2(0,\pi)$. Zeigen Sie, daß

$$\Big\| \frac{\partial^\nu}{\partial x^\nu} u(\,\cdot\,,t) \Big\|_{\mathcal{L}^2(0,\pi)} \leq \Big(\frac{\nu}{2e}\Big)^{\nu/2} t^{-\nu/2} \|u^\circ\|_{\mathcal{L}^2(0,\pi)}\,, \qquad t > 0\,,\ \nu \in \mathbb{N}\,.$$

3. Sei $\mathcal{T}_h$ eine reguläre Triangulierung eines polygonalen Gebiets Ω und $a(\,\cdot\,,\cdot\,)$ die Bilinearform aus (90.1). Weiter bezeichne G die Gramsche Matrix und A die Steifigkeitsmatrix der zugehörigen Hutfunktionen in $V_0^{\mathcal{T}}$. Zeigen Sie, daß $\varrho(G^{-1}A) \geq c/h^2$ für ein $c > 0$.
Hinweis: Zeigen Sie zunächst, daß $\varrho(G^{-1}A) = \sup\{a(u,u)/\|u\|^2_{\mathcal{L}^2(\Omega)} : 0 \neq u \in V_0^{\mathcal{T}}\}$.

4. Bei der horizontalen Linienmethode (oder *Rothe-Methode*) wird die parabolische Differentialgleichung bezüglich der Zeit diskretisiert und in jedem Zeitschritt ein elliptisches Randwertproblem gelöst. Für die eindimensionale Wärmeleitungsgleichung (97.5) ergeben sich bei Verwendung des impliziten Euler-Verfahrens auf diese Weise die Näherungen $u_i(x)$ für $u(x,t_i)$, $t_i = i\tau$, aus dem sukzessiven Lösen der Randwertprobleme

$$u_i = u_{i-1} + \tau \frac{d^2}{dx^2} u_i\,, \quad 0 < x < \pi\,, \qquad u_i(0) = u_i(\pi) = 0\,,$$

$i = 1, 2, \ldots,$ und $u_0 = u^\circ$.
(a) Zeigen Sie, daß sich für die Anfangwerte $u^\circ(x) = \sum_{j=1}^\infty \beta_j \sin jx$ hierbei die Näherungen $u_i(x) = \sum_{j=1}^\infty (1 + \tau j^2)^{-i} \beta_j \sin jx$ ergeben. Vergleichen Sie die dies mit der exakten Lösung $u(x,t)$ aus (97.6).
(b) Verallgemeinern Sie die Methode auf beliebige Runge-Kutta-Verfahren und zeigen Sie, daß die Näherungen dann durch $u_i(x) = \sum_{j=1}^\infty R(-\tau j^2)^i \beta_j \sin jx$ gegeben sind, wobei R die Stabilitätsfunktion des Runge-Kutta-Verfahrens bezeichnet.

5. Die Funktion $f : \Omega \times [0,T] \to \mathbb{R}$ sei viermal bezüglich der Zeit stetig differenzierbar mit $f^{(k)}(t) \in \mathcal{L}^2(\Omega)$ für $k = 0, 1, \ldots, 4$ und alle $0 \leq t \leq T$. Zeigen Sie, daß unter diesen Voraussetzungen

$$\|f(t+\tau) - 2f(t) + f(t-\tau) - \tau^2 f''(t)\|_{\mathcal{L}^2(\Omega)} \leq \frac{\tau^4}{\sqrt{63}} \sup_{t-\tau \leq s \leq t+\tau} \|f^{(4)}(s)\|_{\mathcal{L}^2(\Omega)}\,,$$

sofern $t - \tau \geq 0$ und $t + \tau \leq T$.

6. (a) Es seien a und c stetige und beschränkte Funktionen über $\Omega \times [0,T)$, $b \in \mathbb{R}^2$ sei ein zweidimensionaler Vektor ebensolcher Funktionen und es gelte $a \geq a_0$ für eine Konstante $a_0 > 0$. Ferner sei $u : \Omega \times (0,T)$ zweimal bezüglich $x \in \Omega$ und einmal bezüglich der Zeit $t \in (0,T)$ stetig differenzierbar sowie stetig auf den Rand von $\Omega \times (0,T)$ fortsetzbar. Zeigen Sie: Gilt

$$u_t - a\Delta u + b \cdot \operatorname{grad} u + cu \;\leq\; 0 \qquad \text{in } \Omega \times (0,T)$$

und ist $u(0) \leq 0$ in Ω und $u(t) \leq 0$ auf dem Rand Γ von Ω für $0 < t < T$, so ist u in $\Omega \times [0,T)$ nichtpositiv.
(b) Nutzen Sie das Resultat aus (a) für einen Beweis, daß das quasilineare Anfangsrandwertproblem (102.1) mit $\sigma \in C^2(\mathbb{R})$ und $f \in C^1(\mathbb{R})$ höchstens eine klassische Lösung besitzt, falls $0 < \sigma_0 \leq \sigma(\omega) \leq \sigma_\infty$ für alle $\omega \in \mathbb{R}$.
Hinweis zu (a): Betrachten Sie die Funktion $v(x,t) = u(x,t)e^{-\gamma t}$ mit geeignetem $\gamma \in \mathbb{R}$.

7. Zeigen Sie, daß die Diagonalmatrix D bei der Lumping-Variante im $\mathbb{R}^2$ die Gramsche Matrix der Hutfunktionen in $V^{\mathcal{T}}$ bezüglich des „diskreten $\mathcal{L}^2$-Innenprodukts“

$$\langle\, \varphi, \psi \,\rangle_h \;=\; \sum_{T_k \in \mathcal{T}} \frac{1}{3}\, |T_k| \sum_{\nu=1}^{3} \varphi(x_{k\nu})\psi(x_{k\nu})$$

ist, wobei $|T_k|$ den Flächeninhalt des Dreiecks T_k und $x_{k\nu}$, $\nu = 1,2,3$, die drei Ecken von T_k bezeichnen.

8. Seien (A,b,c) die Koeffizienten des s-stufigen Gauß-Verfahrens für $s \geq 2$. Beweisen Sie, daß die zugehörige rationale Funktion q gemäß (101.10) auf der negativen reellen Achse beschränkt ist. Rechnen Sie nach, daß für $s = 2$ die Funktion q eine Konstante ist.
Hinweis: Aufgaben XIV.10 und XIV.11.

9. Beweisen Sie, daß das quasilineare Variationsproblem (102.3) eine eindeutige Lösung besitzt.

10. Bestimmen Sie für Beispiel 102.1 die Einträge der Jacobi-Matrix $J(t,y)$ aus (102.8) sowie der Zeitableitung $\psi(t,y)$ aus (102.10), wenn für die Ortsdiskretisierung ein äquidistantes Gitter $\triangle_h \;=\; \{\, x_j = -1 + jh \;:\; 0 \leq j \leq n+1 \,\}$ mit Gitterweite $h = 2/(n+1)$ verwendet wird.

11. Programmieren Sie das adaptive Programm aus Abschnitt 103 unter Verwendung eines A-stabilen Einschrittverfahrens mit Schrittweitensteuerung aus einer Programmbibliothek Ihrer Wahl.
Hinweis: Für das Verfahren `ode23s` in MATLAB kann mit der Option `OutputFcn` ein Unterprogramm benannt werden, das nach jedem Zeitschritt die Güte des Ortsgitters überprüft.

XVIII Hyperbolische Erhaltungsgleichungen

Als letztes wenden wir uns Erhaltungsgleichungen zu, die auf hyperbolische Differentialgleichungen erster Ordnung führen. Wie wir bereits in Kapitel XII gesehen haben, können die Lösungen solcher Differentialgleichungen Sprungstellen aufweisen, deren numerische Approximation äußerst diffizil ist. Daher kann im Rahmen dieses Buchs allenfalls eine Einführung in die relevanten Problemstellungen gegeben werden, wobei wir uns zudem fast ausschließlich auf den einfachsten Fall von einer Gleichung mit lediglich einer skalaren Ortsvariablen (neben der Zeit) beschränken.

Als Lektüre zu diesem Kapitel sei das Buch von LeVeque [68] empfohlen. Der weitaus schwierigere Fall eines Systems von Erhaltungsgleichungen wird ausführlich in dem Buch von Kröner [66] behandelt.

104 Die Transportgleichung

Wir beginnen mit der einfachsten Erhaltungsgleichung, der linearen Transportgleichung

$$u_t = -a u_x \,, \qquad x \in \mathbb{R} \,,\ t \geq 0 \,, \tag{104.1a}$$

bei der a eine reelle Konstante ist. Diese Gleichung bietet den Vorteil, daß für ein Anfangswertproblem mit Anfangsvorgabe

$$u(x,0) = u^\circ(x) \,, \qquad x \in \mathbb{R} \,, \tag{104.1b}$$

zur Zeit $t = 0$ die exakte Lösung explizit angegeben werden kann, nämlich $u(x,t) = u^\circ(x - at)$, vgl. Abschnitt 65. In diesem Abschnitt haben wir auch kurz angesprochen (auf Seite 497), daß bei einer Anfangsbedingung der Gestalt (104.1b) über einem endlichen Intervall eine zusätzliche Randbedingung vorgeschrieben werden muß. Um die Darstellung zu vereinfachen, beschränken wir uns hier jedoch weitgehend auf die Anfangsbedingung (104.1b).

Differenzenverfahren sind die gängigen Verfahren zur numerischen Lösung hyperbolischer Erhaltungsgleichungen. Wir definieren zunächst ein (äquidistantes) Ortsgitter

$$\triangle \;=\; \{\, x_j = jh \;:\; j \in \mathbb{Z} \,\}$$

mit einer Gitterweite $h > 0$ und bestimmen eine Zeitschrittweite

$$\tau = \gamma h \tag{104.2}$$

über den Parameter $\gamma > 0$, dem in diesem Kapitel eine tragende Rolle zukommt. Auf diese Weise ergibt sich ein zweidimensionales kartesisches Gitter

$$\{\, x_j = jh \;:\; j \in \mathbb{Z} \,\} \;\times\; \{\, t_i = i\tau \;:\; i \in \mathbb{N}_0 \,\} \;\subset\; \mathbb{R} \times \mathbb{R}_0^+ \,,$$

und jedem Knoten dieses Gitters wird analog zu Kapitel XV ein Näherungswert $u_{ij} \approx u(x_j, t_i)$ zugewiesen. Dabei ist es wie im vorangegangenen Kapitel gelegentlich von Vorteil, Näherungswerte u_{ij} mit festem i zu einem (hier unendlichdimensionalen) Vektor $u_i = [u_{ij}]_{j\in\mathbb{Z}}$ zusammenzufassen.

Bei einem Differenzenverfahren werden die Ableitungen von u durch geeignete Differenzenquotienten ersetzt. Wird die Zeitableitung etwa durch

$$u_t(x_j, t_i) \;\approx\; \frac{u_{i+1,j} - u_{ij}}{\tau}$$

approximiert, so können die Vektoren u_i wie beim expliziten Euler-Verfahren rekursiv berechnet werden. Dazu muß lediglich noch die Ortsableitung durch einen der drei Differenzenquotienten D_h, D_h^+ oder D_h^- aus Abschnitt 83 diskretisiert werden. Wie wir noch sehen werden, liefert eine dieser drei Möglichkeiten ein stabiles Verfahren, so daß der Ansatz über das explizite Euler-Verfahren für die Zeitintegration im nachhinein gerechtfertigt ist.

Das Euler-Verfahren führt auf die Rekursionsvorschrift

$$u_{i+1} \;=\; u_i - \tau a L_h u_i \;=\; A_h u_i \,, \qquad A_h = I - \tau a L_h \,, \tag{104.3}$$

wobei L_h die Matrixdarstellung des entsprechenden Differenzenquotienten und A_h die resultierende (unendlichdimensionale) Tridiagonalmatrix mit Toeplitz-Struktur ist,

$$A_h \;=\; \left[\begin{array}{ccc:ccc}
\ddots & \ddots & & & & \\
\ddots & \alpha_0 & \alpha_1 & & & \\
 & \alpha_{-1} & \alpha_0 & \alpha_1 & & \\
\hdashline
 & & \alpha_{-1} & \alpha_0 & \alpha_1 & \\
 & & & \alpha_{-1} & \alpha_0 & \ddots \\
 & & & & \ddots & \ddots
\end{array}\right] , \tag{104.4}$$

deren Einträge je nach Wahl von L_h folgendermaßen lauten:

$$D_h^- : \qquad \alpha_{-1} = \gamma a, \quad \alpha_0 = 1 - \gamma a, \quad \alpha_1 = 0, \tag{104.5a}$$

$$D_h^+ : \qquad \alpha_{-1} = 0, \quad \alpha_0 = 1 + \gamma a, \quad \alpha_1 = -\gamma a, \tag{104.5b}$$

$$D_h : \qquad \alpha_{-1} = \gamma a/2, \quad \alpha_0 = 1, \quad \alpha_1 = -\gamma a/2. \tag{104.5c}$$

Der Vektor u_0 für den ersten Rekursionsschritt ergibt sich aus der Anfangsvorgabe:

$$u_0 = [u_{0j}]_{j\in\mathbb{Z}} \qquad \text{mit} \qquad u_{0j} = u^\circ(x_j)\,.$$

Wie schon bei der Lösung von Randwertaufgaben in Kapitel XV spielen Konsistenz und Stabilität die entscheidende Rolle bei der Konvergenzanalyse des Differenzenverfahrens (104.3). Für die Definition der Konsistenz müssen wir noch die Vektoren

$$\widehat{u}_i \;=\; \big[\, u(x_j, t_i) \,\big]_{j\in\mathbb{Z}}\,, \qquad i \in \mathbb{N}_0\,,$$

mit den Funktionswerten der exakten Lösung einführen.

Definition 104.1. Wir nennen das Differenzenverfahren $u_{i+1} = A_h u_i$ *konsistent* bezüglich der Maximumnorm $\|\cdot\|_\infty$ (im Raum $\ell^\infty(\mathbb{Z})$ aller beschränkten doppelt unendlichen Folgen), wenn bei hinreichend glatter Lösung u für jedes t_i, $i \in \mathbb{N}_0$,

$$\|\, \widehat{u}_{i+1} \;-\; A_h \widehat{u}_i \,\|_\infty = o(h)\,, \qquad h \to 0\,,\ \tau = \gamma h\,,$$

gilt. Haben wir zudem eine Abschätzung der Gestalt

$$\|\, \widehat{u}_{i+1} \;-\; A_h \widehat{u}_i \,\|_\infty \;\le\; C\, h^q \tau \tag{104.6}$$

für ein $C > 0$ und $q \in \mathbb{N}$, dann hat das Verfahren die *Konsistenzordnung* q.

Man vergleiche diese Definition mit den Definitionen 76.3 und 83.4 für gewöhnliche Differentialgleichungen; der Faktor τ in (104.6) entspricht der um Eins höheren Potenz in Definition 76.3.

Da das Euler-Verfahren nur Konsistenzordnung $q = 1$ besitzt, können die betrachteten Differenzenverfahren (104.3)–(104.5) keine höhere Konsistenzordnung erreichen. Für ein Verfahren höherer Konsistenzordnung muß deshalb eine andere Zeitintegration gewählt werden, vgl. Abschnitt 110.

Satz 104.2. *Das Differenzenverfahren* (104.3) *hat Konsistenzordnung* $q = 1$, *wenn* L_h *mittels eines der Differenzenquotienten* D_h, D_h^+ *oder* D_h^- *definiert wird.*

Beweis. Dies folgt nach den obigen Überlegungen sofort aus Lemma 83.2. □

Definition 104.3. Das Differenzenverfahren $u_{i+1} = A_h u_i$ heißt *stabil* bezüglich der Maximumnorm, falls für jede Zeit $T > 0$ eine Konstante $C_T > 0$ existiert mit

$$\|A_h^i\|_\infty \le C_T\,, \qquad \text{sofern} \quad i\tau = i\gamma h \le T\,.$$

Hierbei bezeichnet $\|\cdot\|_\infty$ die von der Maximumnorm in $\ell^\infty(\mathbb{Z})$ induzierte Norm, also das unendlichdimensionale Analogon der Zeilensummennorm.

Bevor wir uns der Stabilität der Differenzenverfahren (104.3)–(104.5) zuwenden, halten wir fest, daß nach dem *Äquivalenzsatz von Lax* – grob gesprochen – bei der Transportgleichung Konsistenz und Stabilität eines Differenzenverfahrens zusammen äquivalent sind zu der Konvergenz des Verfahrens. Für unsere Zwecke ist nur eine Richtung dieses Satzes von Relevanz (aus Konsistenz und Stabilität folgt Konvergenz) und auch diese Aussage benötigen wir nicht in der allgemeinsten Form. Wir beweisen daher lediglich den folgenden Satz:

Satz 104.4. *Das Differenzenverfahren* (104.3) *besitze die Konsistenzordnung* $q \in \mathbb{N}$ *und es gelte*

$$\|A_h\|_\infty \le 1 + \sigma h \tag{104.7}$$

für ein festes $\sigma > 0$*. Dann ist das Differenzenverfahren stabil und es gilt*

$$\|\widehat{u}_i - u_i\|_\infty \;\le\; \frac{\gamma C}{\sigma}\, e^{\sigma T/\gamma}\, h^q \tag{104.8}$$

für alle $t_i = i\tau$ *mit* $0 \le t_i \le T$*, sofern die exakte Lösung* u *der Transportgleichung hinreichend glatt ist. Dabei ist* C *die Konstante aus Definition* 104.1*.*

Beweis. Für den Nachweis der Stabilität seien $h > 0$ und $i \in \mathbb{N}_0$ so, daß $i\tau \le T$ gilt. Dann folgt aus (104.7)

$$\|A_h^i\|_\infty \;\le\; \|A_h\|_\infty^i \;\le\; (1+\sigma h)^i \;\le\; e^{i\sigma h} \;\le\; e^{\sigma T/\gamma}\,. \tag{104.9}$$

Als nächstes beweisen wir induktiv die Ungleichung

$$\|\widehat{u}_i - u_i\|_\infty \;\le\; \gamma C\, \frac{(1+\sigma h)^i - 1}{\sigma}\, h^q\,, \tag{104.10}$$

die für $i = 0$ trivialerweise erfüllt ist. Für den Induktionsschluß $i \mapsto i+1 \in \mathbb{N}$ verwenden wir

$$\begin{aligned} \|\widehat{u}_{i+1} - u_{i+1}\|_\infty \;&=\; \|\widehat{u}_{i+1} - A_h u_i\|_\infty \\ &\le\; \|\widehat{u}_{i+1} - A_h \widehat{u}_i\|_\infty \;+\; \|A_h(\widehat{u}_i - u_i)\|_\infty \end{aligned}$$

und folgern hieraus mit Hilfe der Definition 104.1 und (104.7), daß

$$\begin{aligned}\|\widehat{u}_{i+1} - u_{i+1}\|_\infty &\le \tau C h^q + \|A_h\|_\infty \|\widehat{u}_i - u_i\|_\infty \\ &\le \gamma C h^{q+1} + (1+\sigma h)\,\|\widehat{u}_i - u_i\|_\infty .\end{aligned}$$

Mit der Induktionsannahme (104.10) für den Fehler nach i Zeitschritten ergibt dies

$$\begin{aligned}\|\widehat{u}_{i+1} - u_{i+1}\|_\infty &\le h\gamma C h^q + (1+\sigma h)\,\gamma C\,\frac{(1+\sigma h)^i - 1}{\sigma}\,h^q \\ &\le \gamma C\,\frac{\sigma h + (1+\sigma h)^{i+1} - 1 - \sigma h}{\sigma}\,h^q .\end{aligned}$$

Damit ist die Ungleichung (104.10) bewiesen. Die Behauptung folgt dann ähnlich wie in der Stabilitätsabschätzung (104.9). □

Wir wenden uns nun der Stabilität des Differenzenverfahrens (104.3), (104.4) mit den drei Alternativen (104.5) zu und beschränken uns dabei auf den Fall, daß $a > 0$ ist. Der Fall $a < 0$ kann durch die Transformation $\tilde{u}(t,x) = u(t,-x)$, $\tilde{u}^\circ(x) = u^\circ(-x)$, auf diesen Fall zurückgeführt werden.

Für die Stabilität spielt neben der Wahl des richtigen Differenzenquotienten die *Courant-Friedrichs-Levi-Bedingung* (CFL-Bedingung)

$$a\gamma = a\tau/h \le 1 \tag{104.11}$$

eine entscheidende Rolle.

Satz 104.5. *Sei $a > 0$. Von den drei Möglichkeiten aus Satz* 104.2 *ist lediglich die Verwendung des linksseitigen Differenzenquotienten D_h^- für geeignete h und τ bezüglich der Maximumnorm stabil. Genauer gilt bei Rückwärtsdifferenzen: Ist die CFL-Bedingung* (104.11) *erfüllt, dann ist $\|A_h\|_\infty = 1$ und es gilt das* Maximumprinzip

$$\inf_{x\in\mathbb{R}} u^\circ(x) \le u_{ij} \le \sup_{x\in\mathbb{R}} u^\circ(x) \tag{104.12}$$

für alle $i \in \mathbb{N}_0$ und $j \in \mathbb{Z}$.

Beweis. Unendliche Toeplitz-Matrizen werden ähnlich wie zirkulante Matrizen diagonalisiert, vgl. Abschnitt 54: Die Eigenvektoren sind gegeben durch $x_\theta = [e^{\mathrm{i}j\theta}]_{j\in\mathbb{Z}} \in \ell^\infty(\mathbb{Z})$ für beliebige $\theta \in \mathbb{R}$. Für A_h aus (104.4) lautet die j-te Zeile des Matrix-Vektor-Produkts $A_h x_\theta$

$$\begin{aligned}(A_h x_\theta)_j &= \alpha_{-1} e^{\mathrm{i}\theta(j-1)} + \alpha_0 e^{\mathrm{i}\theta j} + \alpha_1 e^{\mathrm{i}\theta(j+1)} \\ &= (x_\theta)_j \big(\alpha_{-1} e^{-\mathrm{i}\theta} + \alpha_0 + \alpha_1 e^{\mathrm{i}\theta}\big), \qquad j \in \mathbb{Z},\end{aligned}$$

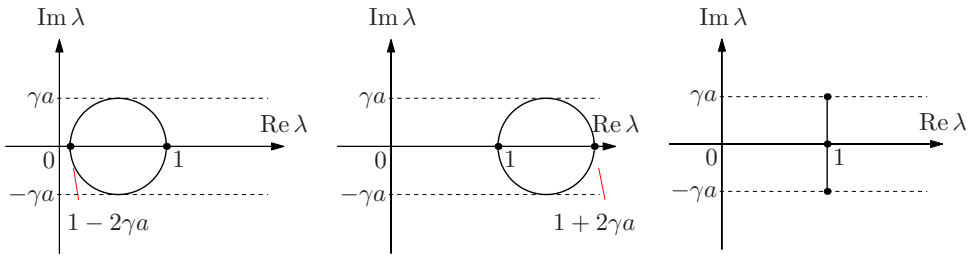


Abb. 104.1: Eigenwerte von A_h für D_h^-, D_h^+ und D_h unter der CFL-Bedingung

und demnach hat A_h den zugehörigen Eigenwert

$$\lambda_\theta \;=\; f(\theta) \;=\; \alpha_{-1}e^{-\mathrm{i}\theta} + \alpha_0 + \alpha_1 e^{\mathrm{i}\theta}\,, \qquad \theta \in \mathbb{R}\,.$$

Die 2π-periodische Funktion f nennt man das *Symbol* der Toeplitz-Matrix. In den drei Fällen (104.5) ergibt sich

$$\begin{aligned} D_h^- &: \quad f(\theta) = 1 - \gamma a + \gamma a\, e^{-\mathrm{i}\theta}\,, \\ D_h^+ &: \quad f(\theta) = 1 + \gamma a - \gamma a\, e^{\mathrm{i}\theta}\,, \\ D_h &: \quad f(\theta) = 1 - \mathrm{i}\,\gamma a \sin\theta\,. \end{aligned}$$

Die Eigenwerte liegen auf den in Abbildung 104.1 dargestellten Kurven. Wie im endlichdimensionalen Fall ist die Norm $\|A_h\|_\infty$ mindestens so groß wie der betragsgrößte Eigenwert von A_h, vgl. Aufgabe I.13. Daher ist

$$\|A_h\|_\infty \;\geq\; \max\{\,|f(\theta)| \,:\, \theta \in \mathbb{R}\}$$

und dies ergibt die unteren Schranken $\|A_h\|_\infty \geq 1+2\gamma a$ für D_h^+ und $\|A_h\|_\infty \geq (1+\gamma^2 a^2)^{1/2}$ für D_h. Wegen $\gamma a > 0$ können diese beiden Differenzverfahren also nicht stabil sein.

Um nachzuweisen, daß D_h^- bezüglich der Maximumnorm stabil ist, müssen wir allerdings anders vorgehen und zeigen im folgenden direkt, daß $\|A_h\|_\infty = 1$ ist. Zunächst ist wegen der gefundenen Eigenwerte sicher $\|A_h\|_\infty \geq 1$. Nun wählen wir einen beliebigen Vektor $v = [v_j]_{j\in\mathbb{Z}} \in \ell^\infty(\mathbb{Z})$ und betrachten die Einträge von $A_h v$. Die j-te Komponente errechnet sich aus

$$(A_h v)_j \;=\; \gamma a v_{j-1} + (1-\gamma a) v_j\,,$$

wobei die rechte Seite wegen der CFL-Bedingung eine Konvexkombination von v_j und v_{j-1} ist; also ist

$$\inf_{j\in\mathbb{Z}} v_j \;\leq\; (A_h v)_{j'} \;\leq\; \sup_{j\in\mathbb{Z}} v_j \qquad \text{für alle } j' \in \mathbb{Z}\,. \tag{104.13}$$

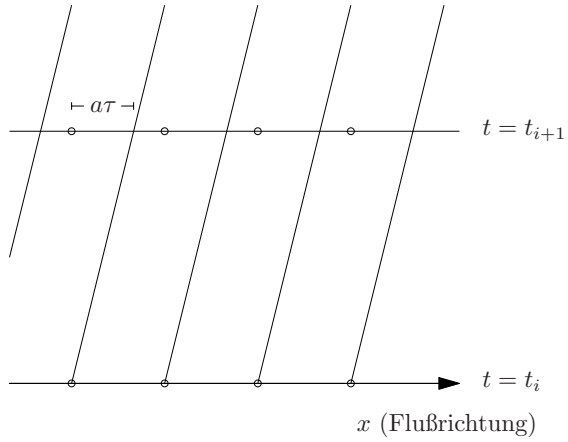


Abb. 104.2: Illustration des Upwind-Schemas

Hieraus folgt $\|A_h v\|_\infty \leq \|v\|_\infty$ und $\|A_h\|_\infty \leq 1$, d. h. das Differenzenverfahren ist stabil.

Ferner erhält man aus (104.3) und (104.13) unmittelbar

$$\inf_{j\in\mathbb{Z}} u_{i-1,j} \;\leq\; u_{ij'} \;\leq\; \sup_{j\in\mathbb{Z}} u_{i-1,j} \qquad \text{für alle } j' \in \mathbb{Z}$$

und durch Induktion ergibt sich daraus die verbliebene Behauptung (104.12).

□

Bemerkung 104.6. Man beachte die Analogie zu singulär gestörten Randwertproblemen, vgl. Abschnitt 85: Die Transportgleichung kann als Grenzfall der parabolischen Differentialgleichung

$$u_t = \varepsilon u_{xx} - a u_x$$

für $\varepsilon \to 0$ interpretiert werden. In Abschnitt 85 haben wir gesehen, daß eine stabile Diskretisierung des elliptischen Operators $L[u] = \varepsilon u_{xx} - a u_x$ nur mit Hilfe von einseitigen Differenzenquotienten *entgegen* der Flußrichtung möglich ist. Dieselbe Beobachtung trifft auch hier zu.

Wir nennen das Differenzenschema mit D_h^- (für $a > 0$) daher auch bei hyperbolischen Erhaltungsgleichungen ein *Upwind-Schema*; für $a < 0$ muß entsprechend der Differenzenoperator D_h^+ verwendet werden. ◊

Die Auswahl des Differenzenquotienten D_h^- für u_x kann auch anders motiviert werden. Nach unseren Überlegungen aus Beispiel 65.1 ist die Lösung u des Anfangswertproblems entlang der Geraden $x(t) = x_j + a(t - t_i)$ konstant. Wie in Abbildung 104.2 verdeutlicht, gehen diese Geraden durch die Gitterpunkte

(x_j, t_i) und schneiden die Zeitlinie $t = t_{i+1}$ in den Punkten $x = x_j + a\tau$. Bei Vorgabe von $u_{ij} = u(x_j, t_i)$ und $u_{i,j+1}$ ist somit

$$u(x_j + a\tau, t_{i+1}) = u_{ij}\,, \qquad u(x_j + a\tau + h, t_{i+1}) = u_{i,j+1}\,.$$

Allerdings sind $x_j + a\tau$ und $x_j + a\tau + h$ in der Regel keine Gitterpunkte von $\triangle$; statt dessen liegt genau ein Gitterpunkt $x_{j'}$ im Intervall $[x_j + a\tau, x_j + a\tau + h)$ und für diesen Gitterpunkt erhält man durch lineare Interpolation die Näherung

$$u(x_{j'}, t_{i+1}) \;\approx\; \frac{x_{j'} - x_j - a\tau}{h}\, u_{i,j+1} + \frac{x_j + a\tau + h - x_{j'}}{h}\, u_{ij}\,.$$

Da die Flußrichtung $a > 0$ angenommen war, ist sicher $j' > j$; die CFL-Bedingung garantiert ferner, daß $j' = j + 1$ ist. In diesem Fall ergibt die obige Interpolation wieder das Upwind-Schema:

$$u(x_{j+1}, t_{i+1}) \;\approx\; (1 - a\gamma)u_{i,j+1} + a\gamma u_{ij} \;=\; u_{i+1,j+1}\,. \tag{104.14}$$

Bemerkung. Wie bereits erwähnt, sind bei Anfangsvorgaben über beschränkten Ortsintervallen für die Transportgleichung mit $a > 0$ Randvorgaben auf dem linken Rand (Einflußrand) nötig, vgl. Seite 497. Das obige Differenzenverfahren läßt sich problemlos auf diesen Fall übertragen, sofern die entsprechenden Randvorgaben zur Verfügung stehen. ◊

105 Die Methode der Charakteristiken

Wir kommen nun zu Anfangswertproblemen für nichtlineare skalare Erhaltungsgleichungen,

$$\begin{aligned} u_t + \frac{d}{dx}\, F(u) \;&=\; u_t + a(u)u_x \;=\; f(x,t)\,, \\ u(x,0) = u^\circ(x)\,, &\qquad a(u) = F'(u)\,. \end{aligned} \tag{105.1}$$

Dabei beschreibt u eine Erhaltungsdichte und F ihren Fluß, von dem wir im weiteren annehmen wollen, daß er nur von der Dichte selbst abhängt. Bei der Untersuchung dieser Gleichung spielen sogenannte Charakteristiken eine wichtige Rolle.

Definition 105.1. Die Funktion u löse die Differentialgleichung (105.1) für $0 \le t \le t_0$. Dann heißen die Lösungen $\chi = \chi(\tau; x_0)$ der Anfangswertprobleme

$$\chi'(\tau) = a\big(u(\chi(\tau), \tau)\big)\,, \quad 0 \le \tau \le t_0\,; \qquad \chi(0) = x_0\,,$$

Charakteristiken der Erhaltungsgleichung (105.1).

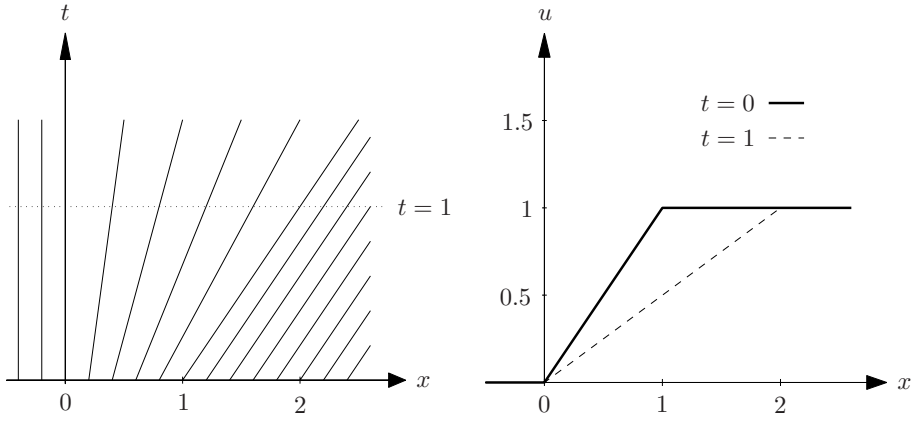


Abb. 105.1: Charakteristiken in der (x,t)-Ebene (links) und die Lösung u zu zwei festen Zeitpunkten als Funktion des Orts (rechts)

Proposition 105.2. *Sei $U(\tau) = u(\chi(\tau),\tau)$ die Einschränkung der Lösung u von* (105.1) *auf den Graph der Charakteristik $\chi = \chi(\,\cdot\,;x_0)$. Dann löst U das Anfangswertproblem*

$$\frac{d}{d\tau}\,U(\tau) \;=\; f\big(\chi(\tau),\tau\big)\,, \qquad U(0) = u^\circ(x_0)\,. \tag{105.2}$$

Beweis. Die Behauptung folgt unmittelbar aus der Kettenregel und (105.1):

$$\frac{d}{d\tau}\,U(\tau) \;=\; u_x\,\chi'(\tau) + u_t \;=\; u_x a(u) + u_t \;=\; f\,.$$

□

Korollar 105.3. *Ist die Differentialgleichung* (105.1) *homogen, also $f = 0$, dann sind die Graphen der Charakteristiken die* (*in diesem Fall geradlinigen*) *Höhenlinien von u.*

Beweis. Für $f = 0$ ist die Lösung U von (105.2) eine konstante Funktion der Zeit, der Wert von $u(\chi(\tau),\tau)$ also konstant. Somit ist der Graph von χ eine Höhenlinie von u. Nach Definition 105.1 löst χ das Anfangswertproblem

$$\chi' = a(U), \qquad \chi(0) = x_0\,,$$

also gilt $\chi(t) = x_0 + a(U)t$, d. h. der Graph von χ ist eine Gerade. □

Abbildung 105.1 illustriert die Aussage dieses Korollars. Zu jedem Punkt auf der Anfangskurve existiert eine lokal eindeutig bestimmte charakteristische Gerade, auf der die Lösung den zugehörigen Anfangswert, in diesem Fall die

Initialisierung: Gegeben sei das Ortsgitter $\triangle = \{x_j = jh : j \in \mathbb{Z}\} \subset \mathbb{R}$

`for` $j \in \mathbb{Z}$ `do`

löse für $\tau \in [0,T]$ das gekoppelte Anfangswertproblem

$$\chi_j'(\tau) = a(U_j)\,, \qquad \chi_j(0) = x_j\,,$$
$$U_j'(\tau) = f(\chi_j,\tau)\,, \qquad U_j(0) = u^\circ(x_j)$$

% verwende hierzu ein geeignetes Verfahren aus Kapitel XIV

`end for`

Ergebnis: Falls die exakte Lösung u von (105.1) im Zeitintervall $0 \le t \le T$ glatt ist, gilt $u(\chi_j(t),t) = U_j(t)$ für $t \in [0,T]$

Algorithmus 105.1: Charakteristikenmethode

Vorgabe

$$u^\circ(x) = \begin{cases} 0\,, & x < 0\,, \\ x\,, & 0 \le x < 1\,, \\ 1\,, & x \ge 1\,, \end{cases}$$

beibehält.

Während Proposition 105.2 auch noch richtig bleibt, wenn a zusätzlich von t oder x abhängt, sind die Charakteristiken in diesem Fall im allgemeinen keine Geraden mehr, da ihre Steigungen dann neben u auch von Ort und Zeit abhängen.

Beispiel. In Abschnitt 65 haben wir bereits die Charakteristiken der Transportgleichung ausgerechnet, denn dort haben wir gesehen, daß die Lösung entlang der Geraden $x = x_0 + at$ jeweils konstant ist. Tatsächlich überprüft man leicht anhand der obigen Definition, daß diese Geraden die Charakteristikengleichung

$$\chi'(\tau) = a\,, \qquad \chi(0) = x_0\,,$$

für die Transportgleichung erfüllen. ◊

Bei der *Charakteristikenmethode* aus Algorithmus 105.1 wird Proposition 105.2 für ein numerisches Verfahren verwendet, indem die Lösung u der Differentialgleichung (105.1) entlang der Charakteristiken rekonstruiert wird.

Bei diesem Verfahren muß man sich bewußt sein, daß die definierenden Differentialgleichungen für die Charakteristiken unter Umständen nur in einem schmalen Bereich $0 \le t < t_0$ eindeutig lösbar sind, also solange die Lösung glatt ist. Als warnendes Beispiel betrachten wir die folgende Situation:

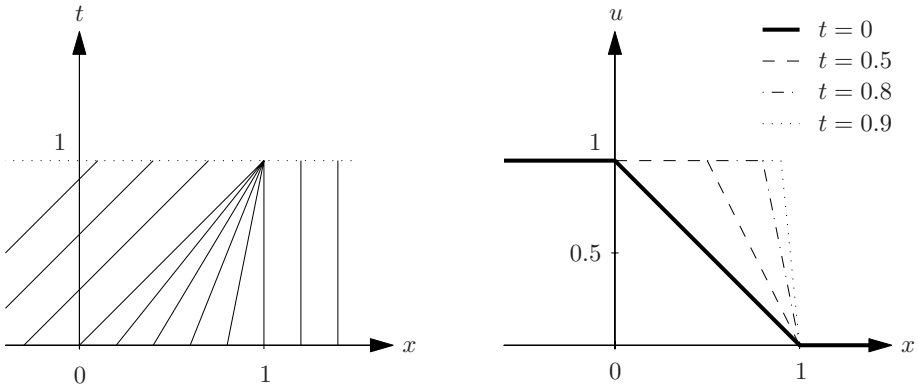


Abb. 105.2: Charakteristiken (links) und der Lösungsgraph von $u(t)$ zu verschiedenen Zeitpunkten (rechts)

Beispiel 105.4 (Burgers-Gleichung). Für den Fluß $F(u) = u^2/2$ lautet die Differentialgleichung

$$u_t + uu_x = 0\,, \qquad u(x,0) = u^\circ(x)\,. \tag{105.3}$$

Dies ist die sogenannte *Burgers-Gleichung*. Die Charakteristiken der Lösung ergeben sich aus den Differentialgleichungen

$$\chi'(\tau) = u\big(\chi(\tau),\tau\big)\,, \qquad \chi(0) = x_0\,,$$

und der Graph von $\chi(\tau;x_0)$ ist eine Gerade in der (x,t)-Ebene mit Steigung $1/u^\circ(x_0)$. Bei der Anfangsvorgabe

$$u^\circ(x) \;=\; \begin{cases} 0\,, & x < 0\,, \\ x\,, & 0 \le x < 1\,, \\ 1\,, & x \ge 1\,, \end{cases}$$

verlaufen die Charakteristiken beispielsweise wie in Abbildung 105.1; u° ist in der rechten Hälfte der Abbildung fett eingezeichnet.

Für die Anfangsvorgabe

$$u^\circ(x) \;=\; \begin{cases} 1\,, & x < 0\,, \\ 1-x\,, & 0 \le x < 1\,, \\ 0\,, & x \ge 1\,, \end{cases} \tag{105.4}$$

ergibt sich hingegen ein ganz anderes Bild, nämlich die linke Skizze in Abbildung 105.2. In diesem Fall schneidet sich ein ganzes Bündel von Charakteristiken im Punkt $(1,1)$; die Lösung kann also in diesen Punkt nicht mehr stetig

fortsetzbar sein. Ergänzend zeigt das rechte Bild von Abbildung 105.2 wieder die Lösungskurven[1] $u(t)$ für verschiedene Werte von $t \in [0,1)$. Wie man sieht, „staut" sich die Lösungskurve für $t \to 1$ im Punkt $x = 1$ auf; man spricht daher von einem *Schock*. ◊

Nur solange keine derartigen Schocks auftreten, ist Algorithmus 105.1 ein verläßliches numerisches Verfahren zur Lösung solcher Erhaltungsgleichungen.

106 Schwache Lösungen und der Begriff der Entropie

Wie in den vorangegengenen Kapiteln führen wir nun einen schwächeren Lösungsbegriff für hyperbolische Erhaltungsgleichungen der Gestalt

$$u_t + \frac{d}{dx} F(u) = f\,, \quad x \in \mathbb{R}\,,\ t > 0\,, \qquad u(x,0) = u^\circ(x)\,, \tag{106.1}$$

ein. Dazu integrieren wir (106.1) über ein beliebiges Gebiet $\Omega' \subset \mathbb{R} \times \mathbb{R}_0^+$ (in der (x,t)-Ebene) und erhalten mit dem Gaußschen Integralsatz

$$\begin{aligned} \int_{\Omega'} f\, d(x,t) \;&=\; \int_{\Omega'} \big(\frac{d}{dx} F(u) + u_t\big)\, d(x,t) \\ &=\; \int_{\Omega'} \operatorname{div}\big[F(u),u\big]^T\, d(x,t) \;=\; \int_{\partial\Omega'} \nu' \cdot \big[F(u),u\big]^T\, ds\,, \end{aligned}$$

wobei die Divergenz hier ausnahmsweise bezüglich beider Variablen x *und* t zu bilden ist und ν' die äußere Normale des Gebiets Ω' bezeichnet.

Definition 106.1. Sei $T \le \infty$. Eine beschränkte Funktion $u : \mathbb{R} \times [0,T] \to \mathbb{R}$ mit $u(x,0) = u^\circ(x)$ heißt *schwache Lösung* des Anfangswertproblems (106.1) im Zeitintervall $0 \le t \le T$, falls

$$\int_{\partial\Omega'} \nu' \cdot \big[F(u),u\big]^T\, ds \;=\; \int_{\Omega'} f\; d(x,t) \tag{106.2}$$

für alle beschränkten Gebiete $\Omega' \subset \mathbb{R} \times [0,T]$.

Bemerkung 106.2. Für den praktischen Nachweis einer schwachen Lösung reicht es aus, die Bedingung (106.2) für alle Rechtecke $\Omega' = (x_1,x_2) \times (t_1,t_2)$

[1] Wir schreiben im folgenden wieder kurz $u(t)$ für die Funktion $u(\,\cdot\,,t)$ des Orts, vgl. die Fußnote auf Seite 724.

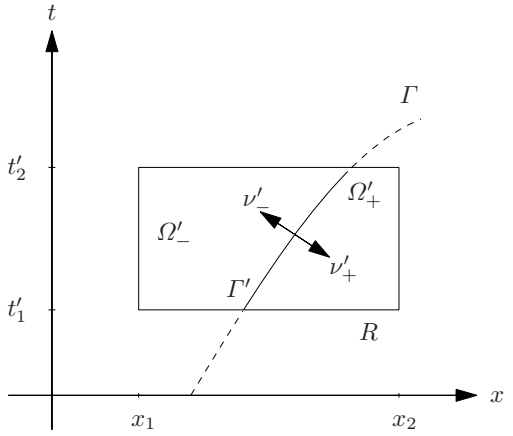


Abb. 106.1: Skizze zur Herleitung der Rankine-Hugoniot-Bedingung

mit $x_1 < x_2$ und $0 \le t_1 < t_2 \le T$ zu überprüfen. In diesem Fall hat (106.2) die Form

$$\begin{aligned}\int_{t_1}^{t_2}\int_{x_1}^{x_2} f(x,t)\,dx\,dt \;=\; &\int_{x_1}^{x_2} u(x,t_2)\,dx \;-\; \int_{x_1}^{x_2} u(x,t_1)\,dx \\ &+\int_{t_1}^{t_2} F(u(x_2,t))\,dt \;-\; \int_{t_1}^{t_2} F(u(x_1,t))\,dt\,.\end{aligned} \tag{106.3}$$

Diese Integralidentität kann übrigens auch direkt aus der integralen Erhaltungsgleichung (65.1) durch Integration über die Zeit hergeleitet werden. ◊

Definition 106.1 läßt auch unstetige Funktionen u als Kandidaten für eine schwache Lösung zu. Wir wollen diese Möglichkeit im folgenden genauer untersuchen. Dazu nehmen wir an, daß f integrierbar und u eine schwache Lösung von (106.1) ist, die entlang einer glatten Kurve

$$\Gamma \;=\; \{\,(x,t)\,:\,x=\psi(t),\,t_1<t<t_2\,\}$$

unstetig oder nicht differenzierbar ist, aber von links und rechts jeweils stetig auf Γ fortgesetzt werden kann; die entsprechenden Grenzwerte werden mit

$$u_-(t) = \lim_{x\to\psi(t)-} u(x,t) \qquad \text{und} \qquad u_+(t) = \lim_{x\to\psi(t)+} u(x,t) \tag{106.4}$$

bezeichnet. Wir schließen nun einen Teil Γ' von Γ in ein Rechteck $R = (x_1,x_2)\times(t_1',t_2')$ mit $t_1 \le t_1' < t_2' \le t_2$ vollständig ein und zerlegen dieses Rechteck in $R = \Omega_-' \cup \Gamma' \cup \Omega_+'$ wie in Abbildung 106.1; Ω_-' und Ω_+' bezeichnen die Teile des Rechtecks links bzw. rechts von Γ'. Sind schließlich ν_-' und

$\nu'_+ = -\nu'_-$ die Normalenvektoren der Kurve Γ nach links bzw. rechts, dann ergibt Definition 106.1 (angewendet auf R, Ω'_- und Ω'_+)

$$\begin{aligned}
\int\limits_R f\,d(x,t) \;&=\; \int\limits_{\partial R} \nu' \cdot \big[F(u),u\big]^T\,ds \\
&= \int\limits_{\partial\Omega'_-} \nu' \cdot \big[F(u),u\big]^T\,ds \;-\; \int\limits_{\Gamma'} \nu'_+ \cdot \big[F(u_-),u_-\big]^T\,ds \\
&\qquad + \int\limits_{\partial\Omega'_+} \nu' \cdot \big[F(u),u\big]^T\,ds \;-\; \int\limits_{\Gamma'} \nu'_- \cdot \big[F(u_+),u_+\big]^T\,ds \\
&= \int\limits_{\Omega'_-} f\,d(x,t) \;+\; \int\limits_{\Omega'_+} f\,d(x,t) \;-\; \int\limits_{\Gamma'} \nu'_- \cdot \big[F(u_+)-F(u_-),u_+-u_-\big]^T\,ds \\
&= \int\limits_R f\;d(x,t) \;-\; \int\limits_{\Gamma'} \nu'_- \cdot \big[F(u_+)-F(u_-),u_+-u_-\big]^T\,ds\,.
\end{aligned}$$

Da t'_1 und t'_2 beliebig nahe beieinander gewählt werden können, ergibt sich hieraus, daß der Integrand des Kurvenintegrals in der letzten Zeile auf ganz Γ verschwinden muß, also daß $[F(u_+)-F(u_-),u_+-u_-]^T$ parallel zum Tangentialvektor $[\psi'(t),1]^T$ an Γ ist. Für $u_+ \neq u_-$ ist dies die sogenannte *Rankine-Hugoniot-Bedingung*

$$\frac{F(u_+)-F(u_-)}{u_+-u_-} \;=\; \psi'(t)\,. \tag{106.5a}$$

Im Grenzfall, in dem u_+ mit u_- übereinstimmt, ergibt sich aus (106.5a) durch einen formalen Grenzübergang $u_+ \to u_- = u$ die *verallgemeinerte Rankine-Hugoniot-Bedingung*

$$F'(u) = \psi'(t)\,, \qquad \text{falls } u_+ = u_- = u\,, \tag{106.5b}$$

also die Charakteristik Γ aus Definition 105.1.

Wir formulieren nun ein hinreichendes Kriterium für eine schwache Lösung von (106.1).

Definition 106.3. Eine Funktion $u : \mathbb{R}\times\mathbb{R}_0^+ \to \mathbb{R}$ gehört zu der Funktionenklasse $\mathcal{U}$, wenn in jedem Rechteckgebiet $\Omega' \subset \mathbb{R}\times\mathbb{R}_0^+$ höchstens endlich viele Kurven $\Gamma_k = \{\,(\psi_k(t),t) \,:\, 0 \le t_k \le t \le T_k\,\}$, $k = 1,\dots,K$, existieren, so daß u in $\Omega' \setminus \bigcup_k \Gamma_k$ stetig differenzierbar ist und die links- und rechtsseitigen Grenzwerte von u auf Γ_k im Sinne von (106.4) existieren.

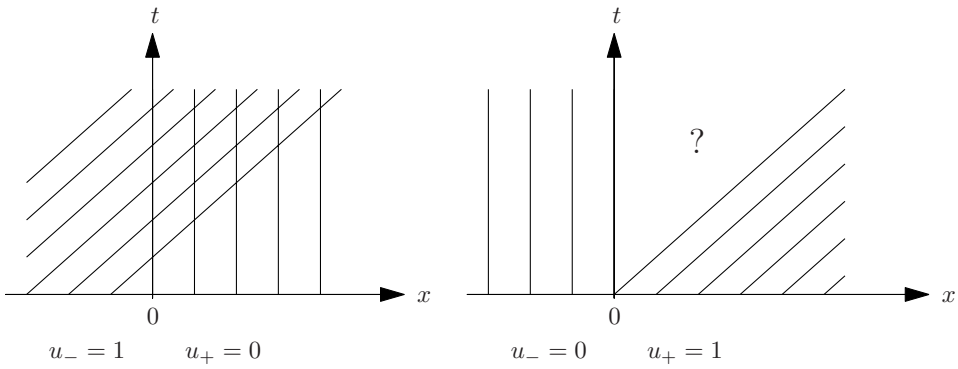


Abb. 106.2: Die Charakteristiken der Burgers-Gleichung für das Riemann-Problem

Satz 106.4. *Eine Funktion $u \in \mathcal{U}$ ist eine schwache Lösung der Erhaltungsgleichung* (106.1)*, falls u in jedem Rechteckgebiet $\Omega' \subset \mathbb{R} \times \mathbb{R}_0^+$ entlang der Unstetigkeitskurven Γ_k aus Definition* 106.3 *die Rankine-Hugoniot-Bedingung* (106.5) *erfüllt und in $\Omega' \setminus \bigcup_k \Gamma_k$ die Diffentialgleichung* (106.1) *im klassischen Sinn löst.*

Beweis. Der Beweis folgt der Argumentation aus Bemerkung 106.2: Für Rechtecke $\Omega' \subset \mathbb{R} \times [0, T]$, die keine Kurve Γ_k enthalten, ergibt sich die Bedingung (106.3) unmittelbar aus dem Gaußschen Satz. Andernfalls zerlegt man das Rechteck analog zu Abbildung 106.1 und erhält dann (106.3) mit Hilfe der Rankine-Hugoniot-Bedingung. □

Beispiel 106.5. Ein wichtiges Testbeispiel für nichtlineare Erhaltungsgleichungen ist das sogenannte *Riemann-Problem*[2]: Hier gibt man als Anfangswert die stückweise konstante Funktion

$$u^\circ(x) = \begin{cases} u_-\,, & x < 0\,, \\ u_+\,, & x \geq 0\,, \end{cases} \tag{106.6}$$

vor. Für die Burgers-Gleichung (Beispiel 105.4) ergeben sich bei der Anfangsbedingung (106.6) die Charakteristiken aus Abbildung 106.2. Dabei zeigt die linke Abbildung den Fall $u_- = 1$, $u_+ = 0$ und die rechte Abbildung den umgekehrten Fall $u_- = 0$, $u_+ = 1$. Eine Situation wie im ersten Fall haben wir

[2]Der Name geht auf eine Arbeit von Riemann zurück, der 1892 ein Problem aus der Strömungsmechanik betrachtet hat, vgl. Abschnitt 67: Eine mit Gas gefüllte Kammer wird durch eine Membran derart in zwei Kammern unterteilt, daß die Dichte und damit der Druck des Gases in den beiden Kammerhälften unterschiedlich ist, u_- bzw. u_+. Die Anfangsbedingung (106.6) ist die entsprechende Vorgabe für den Druck u, wenn zu einem Zeitpunkt $t_0 = 0$ die Membran entfernt wird; es entsteht eine Druckwelle.

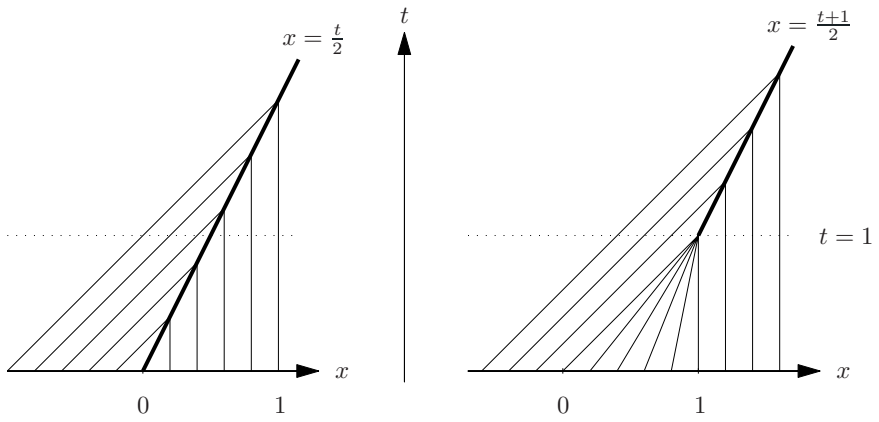


Abb. 106.3: Zwei Schockwellen

bereits in Beispiel 105.4 gesehen: Die Charakteristiken schneiden sich auf der Achse $x = 0$. Man wird erwarten, daß die Lösung dieser Anfangswertaufgabe nicht stetig ist, sondern entlang einer Unstetigkeitskurve $x = \psi(t)$ von $u_- = 1$ nach $u_+ = 0$ springt. In diesem Fall folgt aus der Rankine-Hugoniot-Bedingung (wegen $F(u) = u^2/2$) die Gleichung $\psi'(t) = 1/2$, d. h. die Unstetigkeitskurve Γ ist die Gerade $x = t/2$. Tatsächlich folgt aus Satz 106.4, daß

$$u(x,t) = \begin{cases} 1\,, & x < t/2\,, \\ 0\,, & x \geq t/2\,, \end{cases} \tag{106.7}$$

eine schwache Lösung dieses Anfangswertproblems ist. Die Charakteristiken dieser Schockwelle sind im linken Teil von Abbildung 106.3 dargestellt. Die entsprechende Fortsetzung der Schockwelle aus Beispiel 105.4 bzw. Abbildung 105.2 befindet sich im rechten Teil dieser Abbildung.

Das zweite Riemann-Problem mit $u_- = 0$ und $u_+ = 1$ zeigt ein neues Phänomen, ein „Aufreißen“ der Charakteristiken, vgl. Abbildung 106.2 rechts. Das Fragezeichen in dieser Abbildung deutet an, daß Punkte in dem Bereich $0 < x < t$ nicht durch die eingezeichneten Charakteristiken mit der Anfangskurve verbunden sind; der Funktionswert der Lösung u ist in diesem Bereich also nicht unmittelbar klar.

Ist $\alpha \in (0, 1]$ ein beliebiger Parameter, dann kann man wieder mit Satz 106.4 begründen, daß die Funktion

$$u(x,t) = \begin{cases} 0\,, & x < \alpha t/2\,, \\ \alpha\,, & \alpha t/2 \leq x < (1+\alpha)t/2\,, \\ 1\,, & x \geq (1+\alpha)t/2\,, \end{cases} \tag{106.8}$$

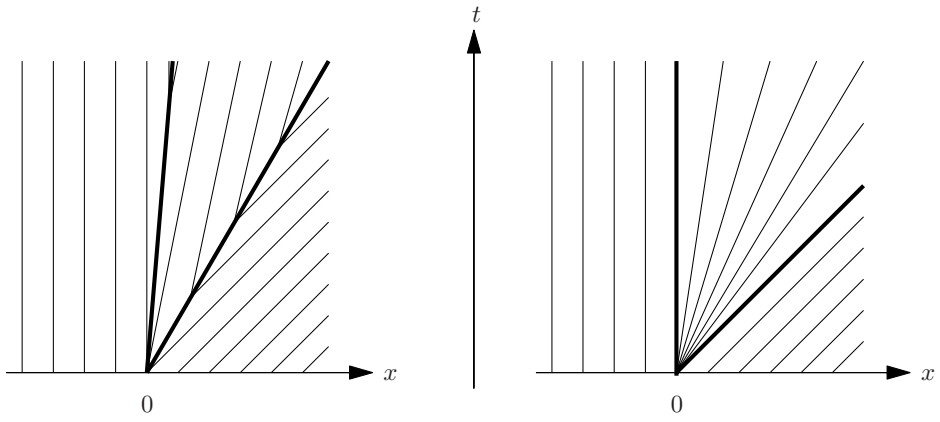


Abb. 106.4: Eine schwache Lösung (links) und die physikalisch korrekte Lösung (rechts)

für jedes solche α eine schwache Lösung dieses Anfangswertproblems ist, denn an den beiden Unstetigkeitskurven sind die Rankine-Hugoniot-Bedingungen erfüllt. Für $\alpha = 1/6$ sind die Charakteristiken dieser Lösung im linken Teil von Abbildung 106.4 dargestellt. Man beachte, daß in dieser Abbildung die zu $u = \alpha$ gehörenden Charakteristiken (bis auf eine Ausnahme) *nicht* auf die Anfangskurve zurückverfolgt werden können, sondern daß sie aus einer der beiden Unstetigkeitskurven entspringen (siehe dazu auch Aufgabe 5).

Wie wir in Beispiel 105.4 und Abbildung 105.1 gesehen haben, ist der Lösungsverlauf unstrittig, wenn die Anfangsvorgabe stetig und monoton wächst, etwa

$$u_\varepsilon^\circ(x) = \begin{cases} 0\,, & x < 0\,, \\ x/\varepsilon\,, & 0 \le x < \varepsilon\,, \\ 1\,, & x \ge \varepsilon\,. \end{cases}$$

Für festes $t > 0$ macht man sich anhand von Abbildung 105.1 oder Satz 106.4 klar, daß die Lösung u_ε durch

$$u_\varepsilon(x,t) = \begin{cases} 0\,, & x < 0\,, \\ x/(\varepsilon + t)\,, & 0 \le x < \varepsilon + t\,, \\ 1\,, & x \ge \varepsilon + t\,, \end{cases}$$

gegeben ist. Der (punktweise) Grenzübergang $\varepsilon \to 0$ führt auf die Funktion

$$u(x,t) = \begin{cases} 0\,, & x < 0\,, \\ x/t\,, & 0 \le x < t\,, \\ 1\,, & x \ge t\,. \end{cases} \tag{106.9}$$

Unterstellt man, daß eine physikalisch sinnvolle Lösung stetig von den Anfangswerten abhängen soll, so drängt sich daher diese Funktion u als physikalisch korrekte Lösung des Riemann-Problems (106.6) mit $u_- = 0$ und $u_+ = 1$ auf. Tatsächlich ist (106.9) eine weitere schwache Lösung der Anfangswertaufgabe: Die beiden Kurven $\Gamma_1 = \{x = 0\}$ und $\Gamma_2 = \{x = t\}$ zerteilen das Gebiet in drei Teilgebiete, in denen die Burgers-Gleichung jeweils erfüllt ist; darüber hinaus sind entlang der beiden Kurven Γ_1 und Γ_2 die (verallgemeinerten) Rankine-Hugoniot-Bedingungen erfüllt; damit folgt die Behauptung aus Satz 106.4. Die zugehörigen sich auffächernden Charakteristiken sind in der rechten Hälfte von Abbildung 106.4 dargestellt. Diese Lösung wird *Verdünnungswelle* genannt. ◊

Aus diesem zweiten Beispiel mag man den Schluß ziehen, daß durch das Zulassen unstetiger (schwacher) Lösungen der Lösungsbegriff zu weit gefaßt wurde; nicht jede unstetige Lösung ist physikalisch relevant. Um die Spreu vom Weizen zu trennen, benötigen wir den Begriff der Entropie.

Definition 106.6. Eine schwache Lösung $u \in \mathcal{U}$ von (106.1) erfüllt die *Entropiebedingung von Oleinik*, wenn längs jeder Unstetigkeitskurve $x = \psi(t)$, $t_1 < t < t_2$, mit zugehörigen links- und rechtsseitigen Grenzwerten $u_-(t)$ und $u_+(t)$ von u die Ungleichung

$$\frac{F(u_-) - F(v)}{u_- - v} \;\geq\; \psi'(t) \;\geq\; \frac{F(u_+) - F(v)}{u_+ - v} \tag{106.10}$$

für jedes $t \in (t_1, t_2)$ und alle v zwischen $u_-(t)$ und $u_+(t)$ erfüllt ist.

Beispiel. Bei dem Riemann-Problem für die Burgers-Gleichung (Beispiel 106.5) mit $u_- = 0$ und $u_+ = 1$ verletzen die schwachen Lösungen (106.8) die Entropiebedingung von Oleinik: An der ersten Unstetigkeitskurve $\psi(t) = \alpha t/2$ haben wir beispielsweise

$$\frac{F(u_-) - F(v)}{u_- - v} \;=\; \frac{F(v) - F(0)}{v} \;=\; \frac{v}{2} \;<\; \psi'(t) \;=\; \frac{\alpha}{2}$$

für alle v zwischen 0 und α im Widerspruch zur Forderung (106.10). Die physikalisch korrekte Lösung (106.9) steht hingegen im Einklang mit der Entropiebedingung, da diese Lösung stetig ist. ◊

Auf der Basis dieser Entropiebedingung kann die Eindeutigkeit der Lösung des Anfangswertproblems (106.1) nachgewiesen werden. Wir zitieren lediglich dieses Resultat und verweisen für einen Beweis auf das Buch von Warnecke [107, Abschnitt 10.3] (für einen Zusammenhang der dort verwendeten Entropiebedingung von Kružkov mit der Bedingung von Oleinik vergleiche man Aufgabe 7).

Satz 106.7. *Ist $F \in C^\infty(\mathbb{R})$ und u° beschränkt mit*

$$\int_{\mathbb{R}} |u^\circ(x)|\, dx \,<\, \infty\,,$$

so hat das Anfangswertproblem (106.1) *höchstens eine schwache Lösung $u \in \mathcal{U}$, die die Entropiebedingung von Oleinik erfüllt.*

Wir können und wollen in diesem Buch nicht herleiten, inwieweit Definition 106.6 mit dem Entropiebegriff aus der Thermodynamik zusammenhängt.[3] Es sei hier lediglich angemerkt, daß die Entropie in der Thermodynamik als Maß für die Unordnung interpretiert wird. Bei natürlichen Vorgängen nimmt die Entropie (also die Unordnung) mit der Zeit zu. Wie wir in dem obigen Beispiel gesehen haben, zeichnet die Entropiebedingung von Oleinik schwache Lösungen mit Verdünnungswellen aus, also Lösungen, bei denen alle Werte gewisser Intervalle durchlaufen werden. Aufgabe 6 bestätigt diese Beobachtung anhand eines anderen Beispiels mit einer unstetigen Lösung. Die Natur versucht also offensichtlich, die Ordnung getrennter Zustände durch einen „verschmierten“ Übergang zu ersetzen. Wenn wir im folgenden von “physikalisch korrekten” Lösungen eines Anfangswertproblems (106.1) sprechen, meinen wir also immer Lösungen im Sinne des Satzes 106.7.

Beispiel. Schock- und Verdünnungswellen sind uns auch in dem Chromatographie-Beispiel aus Abschnitt 66 begegnet. Abbildung 66.2 zeigt links die Konzentration u der Substanz in der Säule in Abhängigkeit von Zeit und Ort. Von links nach rechts erkennt man zunächst einen abrupten Sprung von $u = 0$ auf die maximale Konzentration u_*, also eine Schockwelle. Danach nimmt die Konzentration in Form einer Verdünnungswelle allmählich wieder ab. ◊

107 Das Godunov-Verfahren

Im vergangenen Abschnitt haben wir für die Burgers-Gleichung die physikalisch relevanten schwachen Lösungen des Riemann-Problems bestimmt. Darauf aufbauend kann nun für eine beliebige Erhaltungsgleichung

$$u_t + \frac{d}{dx} F(u) \,=\, 0\,, \quad x \in \mathbb{R}\,,\ t \geq 0\,, \qquad u(x,0) = u^\circ(x)\,, \qquad (107.1)$$

mit Anfangsvorgabe

$$u^\circ(x) \,=\, \begin{cases} u_-\,, & x < 0\,, \\ u_+\,, & x \geq 0\,, \end{cases} \qquad u_- \neq u_+\,,$$

[3] Interessierte Leser/innen seien auf die Arbeit von Ansorge und Sonar [4] verwiesen.

die Lösung angegeben werden.

Satz 107.1. *Sei $a = F'$ stetig differenzierbar mit $a'(v) > 0$ für alle $v \in \mathbb{R}$. Dann lautet die physikalisch relevante Lösung $u \in \mathcal{U}$ von* (107.1) *wie folgt:*

(*a*) $u_- > u_+$: *Mit $s = (F(u_-) - F(u_+))/(u_- - u_+)$ gilt*

$$u(x,t) = \begin{cases} u_-\,, & x < st\,, \\ u_+\,, & x \geq st\,; \end{cases}$$

(*b*) $u_+ > u_-$: *Bezeichnet $a^{-1}(s)$ die Umkehrfunktion von a, dann gilt*

$$u(x,t) = \begin{cases} u_-\,, & x < a(u_-)t\,, \\ a^{-1}(x/t)\,, & a(u_-)t \leq x < a(u_+)t\,, \\ u_+\,, & x \geq a(u_+)t\,. \end{cases}$$

Beweis. Die angeführte Funktion u ist bis auf die Trenngerade $x = st$ im ersten Fall und die beiden Geraden $x = a(u_-)t$ bzw. $x = a(u_+)t$ im zweiten Fall stetig differenzierbar und genügt in den jeweiligen Teilgebieten der Differentialgleichung (107.1). Die einzige Schwierigkeit besteht hierbei im zweiten Fall zwischen den beiden Trenngeraden. Dort ist $u(x,t) = a^{-1}(x/t)$ beziehungsweise $a(u(x,t)) = x/t$, und es folgt

$$u_x = \frac{1}{a'(u)}\,\frac{1}{t}\,, \qquad u_t = \frac{1}{a'(u)}\,\frac{-x}{t^2}\,.$$

Daraus ergibt sich dann aber $u_t + a(u)u_x = u_t + (x/t)u_x = 0$.

Im Fall (*b*) ist die (verallgemeinerte) Rankine-Hugoniot-Bedingung an den Nahtstellen $x = a(u_-)t$ und $x = a(u_+)t$ definitionsgemäß erfüllt; also ist u nach Satz 106.4 eine schwache Lösung von (107.1). Die Entropiebedingung (106.10) muß nirgends überprüft werden, da u stetig ist. Folglich ist u die physikalisch korrekte Lösung.

Im Fall (*a*) ist noch die Unstetigkeitsstelle $x = st$ zu überprüfen, an der die Rankine-Hugoniot-Bedingung definitionsgemäß erfüllt ist. Die Entropiebedingung (106.10) ergibt in diesem Fall die Forderung

$$\frac{F(u_-) - F(u)}{u_- - u} \geq \frac{F(u_-) - F(u_+)}{u_- - u_+} \geq \frac{F(u_+) - F(u)}{u_+ - u}\,, \tag{107.2}$$

die für jedes u zwischen u_+ und u_- erfüllt sein muß. Da $a' = F''$ positiv ist, ist F konvex und wegen $u_+ < u < u_-$ folgt hieraus die gewünschte Ungleichungskette, vgl. etwa Rudin [93]. Damit ist der Beweis vollständig. □

Für monoton fallende Funktionen a gilt ein entsprechendes Resultat, bei dem lediglich die Rollen von Schock- und Verdünnungswelle vertauscht sind.

Man beachte, daß sich gemäß Satz 107.1 die Singularität der physikalisch korrekten Lösung u eines Riemann-Problems für eine homogene Erhaltungsgleichung immer entlang von Geraden in der (x,t)-Ebene ausbreitet, solange der Fluß F lediglich von u und nicht von x oder t abhängt, vgl. auch Korollar 105.3. Lediglich die „Form“ der Verdünnungswelle und die Ausbreitungsgeschwindigkeit der Schockwelle hängen von F ab.

Beispiel. Für die Flußfunktion des Chromatographie-Beispiels (Abschnitt 66) gilt $F(u) = av(u)$ mit v aus (66.6), also

$$F(u) \;=\; \frac{a}{2}\,(u-\rho-1/\kappa) \;+\; \frac{a}{2}\left(\frac{4u}{\kappa} + (u-\rho-1/\kappa)^2\right)^{1/2}$$

mit Umkehrfunktion

$$u \;=\; v + \rho\,\frac{\kappa v}{1+\kappa v}\,, \qquad v \;=\; \frac{F}{a}\,.$$

Die Parameter a, ρ und κ sind positiv. Man rechnet schnell nach, daß F^{-1} für $v > 0$ monoton wachsend und konkav ist, d. h. F ist im relevanten Bereich $u > 0$ monoton wachsend und konvex. Für $u_- = u_*$ und $u_+ = 0$ (zu Beginn der Einspritzung) entsteht also nach Satz 107.1 eine Schockwelle. Für $u_- = 0$ und $u_+ = u_*$ (am Ende der Einspritzung) bildet sich eine Verdünnungswelle, vgl. Abbildung 66.2. ◊

Wir ziehen nun Satz 107.1 heran, um das *Godunov-Verfahren* herzuleiten. Dazu führen wir im Ortsbereich neben dem Gitter $\triangle$ noch ein zweites verschobenes Gitter

$$\triangle' \;=\; \{\, x_{j-1/2} = x_j - h/2 \;:\; j \in \mathbb{Z}\,\}$$

ein und approximieren für jedes $t_i = i\tau$ die exakte Lösung $u(t_i)$ durch eine Treppenfunktion

$$u_h(t_i) \;=\; \sum_{j\in\mathbb{Z}} u_{ij}\chi_j$$

über dem verschobenen Gitter $\triangle'$, wobei χ_j die charakteristische Funktion des Gitterintervalls $[x_{j-1/2}, x_{j+1/2})$ bezeichnet. Die Koeffizienten u_{ij} werden weiterhin in den Vektoren $u_i = [u_{ij}]_{j\in\mathbb{Z}}$ zusammengefaßt, allerdings verstehen wir unter u_{ij} im folgenden eine Approximation der mittleren Masse der Erhaltungsgröße in der Gitterzelle $(x_{j-1/2}, x_{j+1/2})$ von $\triangle'$:

$$u_{ij} \;=\; \frac{1}{h}\int_{x_{j-1/2}}^{x_{j+1/2}} u_h(x,t_i)\,dx \;\approx\; \frac{1}{h}\int_{x_{j-1/2}}^{x_{j+1/2}} u(x,t_i)\,dx\,. \tag{107.3}$$

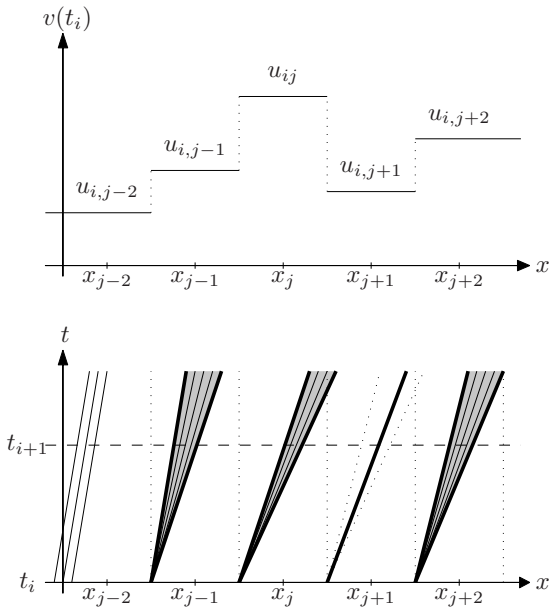


Abb. 107.1: Godunov-Verfahren: Anfangswert (oben) und Höhenlinien der Lösung (unten)

Mit Hilfe von Satz 107.1 können wir die exakte Lösung von

$$v_t + a(v)v_x = 0\,, \quad x \in \mathbb{R}\,,\ t \geq t_i\,, \qquad v(t_i) = u_h(t_i)\,, \tag{107.4}$$

in einem hinreichend kleinen Zeitintervall bestimmen, indem wir die Lösungen der einzelnen Riemann-Probleme zusammensetzen, vgl. Abbildung 107.1. Dazu setzen wir wie in dem Satz voraus, daß $a' = F'' > 0$ gilt.

Tatsächlich ist die Lösung v von (107.4) durch Satz 107.1 festgelegt, solange sich keine Schock- bzw. Verdünnungswellen gegenseitig stören. Dies ist lediglich für benachbarte Verdünnungswellen sichergestellt, da deren einander zugewandte Randkurven die gleiche Steigung aufweisen, nämlich $1/a(u_{ij})$. Anders ist die Situation, wenn an einem Intervallende (etwa links) eine Verdünnungswelle und am anderen Intervallende (also rechts) eine Schockwelle entsteht. In Abbildung 107.1 ist das in dem Gitterintervall um x_j der Fall: Das rechte Ende der am linken Randpunkt entspringenden Verdünnungswelle ist in diesem Fall durch die Gleichung

$$x_l(t) \;=\; x_{j-1/2} + a(u_{ij})(t - t_i)$$

gegeben, während die Schockwelle rechts nach Satz 107.1(a) der Gleichung

$$x_r(t) \;=\; x_{j+1/2} + s(t - t_i) \;=\; x_{j+1/2} + \frac{F(u_{ij}) - F(u_{i,j+1})}{u_{ij} - u_{i,j+1}}\,(t - t_i)$$

genügt. Nach dem Mittelwertsatz kann dies auch in der Form

$$x_r(t) \;=\; x_{j+1/2} + a(\tilde{u})(t - t_i)$$

mit einem $\tilde{u} \in (u_{i,j+1}, u_{ij})$ geschrieben werden. Die beiden Wellen treffen sich, wenn $x_l(t) = x_r(t)$ wird, also wenn

$$\big(a(u_{ij}) - a(\tilde{u})\big)(t - t_i) = h \tag{107.5}$$

gilt. Die gleiche Bedingung ergibt sich, wenn die Schockwelle vom linken Intervallrand und die Verdünnungswelle vom rechten Rand kommt oder wenn sich von beiden Intervallenden her Schockwellen ausbreiten. Da $\tilde{u} < u_{ij}$ und a monoton wachsend ist, hat die Gleichung (107.5) eine Lösung $t^* > t_i$. Über das Verhalten von u für $t > t^*$ gibt Satz 107.1 keine Information.

Um ein Aufeinandertreffen solcher Wellen zu vermeiden, bedarf es einer Einschränkung an die Zeitschrittweite $\tau = t_{i+1} - t_i$: Die zu (104.11) analoge CFL-Bedingung

$$2 \sup_{j\in\mathbb{Z}} |a(u_{ij})|\, \tau/h \le 1 \tag{107.6}$$

ist hinreichend dafür, daß die Lösung t^* von (107.5) nicht kleiner als t_{i+1} ist. Unter der Einschränkung (107.6) an die Zeitschrittweite τ läßt sich $u_h(t_i)$ also zu einer Lösung $v(t)$, $t_i \le t \le t_{i+1}$, von (107.4) fortsetzen. Dieses Verfahren kann induktiv weitergeführt werden, wenn die Lösung $v(t_{i+1})$ von (107.4) wieder durch eine Treppenfunktion $u_h(t_{i+1})$ über dem Gitter $\triangle'$ approximiert wird. Im Hinblick auf (107.3) bietet sich hierfür die Treppenfunktion mit den Koeffizienten

$$u_{i+1,j} \;=\; \frac{1}{h} \int_{x_{j-1/2}}^{x_{j+1/2}} v(x, t_{i+1})\, dx\,, \qquad j \in \mathbb{Z}\,,$$

an. Auf diese Weise wird erreicht, daß die numerische Lösung u_h wie die exakte Lösung eine Erhaltungsgröße darstellt, wie im nachfolgenden Abschnitt genauer dargestellt werden wird. Man beachte zudem, daß die so definierte Treppenfunktion $u_h(t_{i+1})$ nach Satz 43.1 gerade die Bestapproximation an $v(t_{i+1})$ ist.

Diese Approximation kann überraschend einfach implementiert werden, da das Integral über v explizit ausgerechnet werden kann. Hierzu muß die Bedingung (106.3) für die schwache Lösung v verwendet werden: Demnach ist

$$\begin{aligned} u_{i+1,j} - u_{ij} \;&=\; \frac{1}{h} \int_{x_{j-1/2}}^{x_{j+1/2}} v(x, t_{i+1})\, dx \;-\; \frac{1}{h} \int_{x_{j-1/2}}^{x_{j+1/2}} v(x, t_i)\, dx \\ &=\; \frac{1}{h} \int_{t_i}^{t_{i+1}} F(v(x_{j-1/2}, t))\, dt \;-\; \frac{1}{h} \int_{t_i}^{t_{i+1}} F(v(x_{j+1/2}, t))\, dt\,. \end{aligned}$$

Initialisierung: $a = F'$ sei monoton wachsend und es sei $\gamma \le 1/(2\|a\|_{\mathbb{R}})$
  wähle Gitter $\triangle = \{x_j = jh : j \in \mathbb{Z}\} \subset \mathbb{R}$, $\tau = \gamma h$, $t_0 = 0$
  `for` $j \in \mathbb{Z}$ `do` % berechne $u_h(0)$
    $u_{0j} = \dfrac{1}{h} \displaystyle\int_{x_{j-1/2}}^{x_{j+1/2}} u^\circ(x)\, dx$
  `end for`
  `for` $i = 0, 1, 2, \ldots$ `do` % Zeitschritte
    $t_{i+1} = t_i + \tau$
    `for` $j \in \mathbb{Z}$ `do` % berechne $g_{j-1/2} = G(u_{i,j-1}, u_{ij})$
      `if` $u_{i,j-1} \ge u_{ij}$ `then`
        $g_{j-1/2} = \max F(u)$ (bzgl. $u \in [u_{ij}, u_{i,j-1}]$)
      `else`
        $g_{j-1/2} = \min F(u)$ (bzgl. $u \in [u_{i,j-1}, u_{ij}]$)
      `end if`
    `end for`
    `for` $j \in \mathbb{Z}$ `do`
      $u_{i+1,j} = u_{ij} - \gamma(g_{j+1/2} - g_{j-1/2})$
    `end for`
  `until` $t_{i+1} \ge T$
Ergebnis: $u_h(t_i) = \sum_{j\in\mathbb{Z}} u_{ij}\chi_j \approx u(t_i)$

Algorithmus 107.1: Godunov-Verfahren für konvexes F

Da $v(x,t)$ zudem entlang der Geraden $x = x_{j\pm1/2}$ jeweils konstant ist, ergibt dies

$$u_{i+1,j} \;=\; u_{ij} + \gamma\left(G(u_{i,j-1}, u_{ij}) - G(u_{ij}, u_{i,j+1})\right) \tag{107.7}$$

mit $\gamma = \tau/h$, wobei $G(u_-, u_+)$ den Funktionswert $F(u)$ mit demjenigen Wert von u bezeichnet, der sich beim Riemann-Problem (107.1) gemäß Satz 107.1 entlang der Geraden $x = 0$ ergibt. Es ist nicht schwer zu zeigen, vgl. Aufgabe 10, daß wegen der Konvexität von F

$$G(u_-, u_+) \;=\; \begin{cases} \max\limits_{u_+ \le u \le u_-} F(u)\,, & u_- \ge u_+\,, \\ \min\limits_{u_- \le u \le u_+} F(u)\,, & u_- \le u_+\,, \end{cases} \tag{107.8}$$

gilt. Das gesamte Godunov-Verfahren ist in Algorithmus 107.1 zusammengefaßt.

Beispiel. Bei der (linearen) Transportgleichung ist $F(u) = au$. Für $a > 0$ ergibt (107.8)

$$G(u_-, u_+) = F(u_-) = au_-$$

und die Rekursion (107.7) lautet

$$u_{i+1,j} \;=\; u_{ij} + a\gamma(u_{i,j-1} - u_{ij})\,.$$

Dies entspricht der Rekursion (104.14) des Upwind-Schemas, d. h. das Godunov-Verfahren ist eine Verallgemeinerung des Upwind-Schemas auf nichtlineare Erhaltungsgleichungen. ◊

Beispiel 107.2. Zum Abschluß wenden wir das Godunov-Verfahren auf das Chromatographie-Beispiel aus Abschnitt 66 an. Dazu verwenden wir die Anfangsbedingung

$$u(x,0) = 0 \qquad \text{für } 0 \le x \le L$$

und die Randbedingung

$$u(0,t) \;=\; \begin{cases} u_*\,, & 0 \le t \le 1\,, \\ 0\,, & t > 1\,, \end{cases}$$

am Einflußrand. Dabei ist L die Länge der chromatographischen Säule, $x = 0$ der Einspritzpunkt und u_* entspricht dem Zustand maximaler Sättigung. Im folgenden werden wie in Abschnitt 66 die Parameter $u_* = 1$, $L = 10$ und $\rho = \kappa = a = 1$ gewählt.

Für die numerische Simulation wird die Säule in hundert Gitterintervalle unterteilt, d. h. die Gitterweite ist $h = L/100$. Die Zeitschrittweite ergibt sich dann aus der CFL-Bedingung (107.6), nämlich

$$\tau \;=\; \delta h / F'(u_*) \tag{107.9}$$

mit $\delta \le 1/2$. Tatsächlich kann in diesem Beispiel sogar $\delta = 1$ gewählt werden, denn wegen der Monotonie von F läßt sich der Faktor $a(u_{ij}) - a(\tilde{u})$ in (107.5) durch $a(u_{ij})$ nach oben abschätzen.

An Abbildung 107.2 kann man ablesen, daß die Wahl von δ für die Genauigkeit des Verfahrens von entscheidender Bedeutung ist. Die beiden Bilder zeigen die aus der Säule austretende Substanz, gegeben durch $av(L,t)$ in Abhängigkeit von der Zeit. Dabei enthält die linke Abbildung die exakte Lösung (dunkle Kurve) und das berechnete Ergebnis (helle Kurve) für $\delta = 1$, also mit der größten erlaubten Zeitschrittweite; die rechte Abbildung zeigt die numerischen Lösungen für den zulässigen Wert $\delta = 1/2$ (gebrochene helle Kurve) und den die CFL-Bedingung verletzenden Wert $\delta = 1.5$ (durchgezogene helle Kurve). Bei zu vorsichtiger Wahl von δ wird die Lösung offensichtlich stark geglättet, während eine zu große Wahl von δ zu Instabilitäten führt. ◊

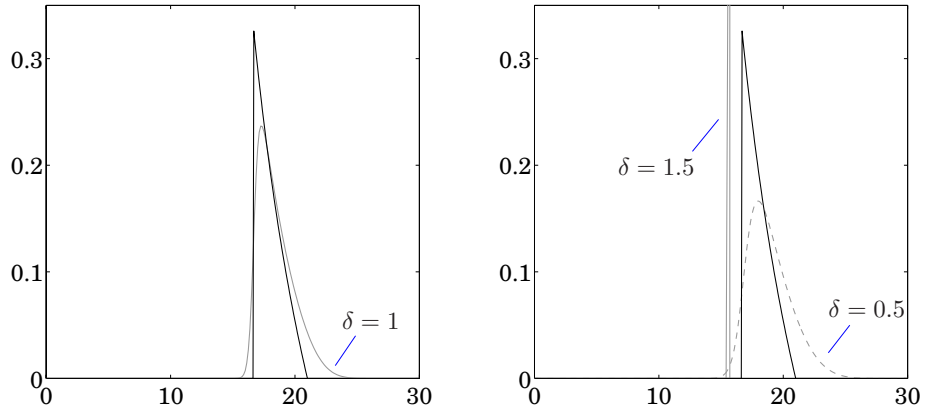


Abb. 107.2: Numerische Ergebnisse des Godunov-Verfahrens für verschiedene γ

108 Differenzenverfahren in Erhaltungsform

Erinnern wir uns an die Ausgangssituation bei der Modellierung eines homogenen Erhaltungsgesetzes (im eindimensionalen Spezialfall): Wenn sich in einem Referenzgebiet $\Omega = [a, b] \subset \mathbb{R}$ die Dichte u aufgrund eines Flusses $F(u)$ verändert, dann lautet die integrale Erhaltungsgleichung

$$\frac{d}{dt} \int_\Omega u \, dx \;=\; -\int_{\partial\Omega} \nu \cdot F(u) \, ds \;=\; F(u(a,t)) - F(u(b,t)) \,.$$

Durch Integration über die Zeit $t_i < t < t_{i+1}$ ergibt sich die schwache Form

$$\begin{aligned} &\int_a^b u(x, t_{i+1}) \, dx - \int_a^b u(x, t_i) \, dx \\ &\qquad = \int_{t_i}^{t_{i+1}} F(u(a,t)) \, dt - \int_{t_i}^{t_{i+1}} F(u(b,t)) \, dt \,. \end{aligned} \tag{108.1}$$

Setzen wir $\tau = t_{i+1} - t_i$ und

$$\mathcal{F}_i(x) = \frac{1}{\tau} \int_{t_i}^{t_{i+1}} F(u(x,t)) \, dt \tag{108.2}$$

für den mittleren Fluß in dem Zeitintervall $[t_i, t_{i+1}]$, dann kann (108.1) weiter umgeformt werden in

$$\int_a^b u(x, t_{i+1}) \, dx - \int_a^b u(x, t_i) \, dx \;=\; \tau \big(\mathcal{F}_i(a) - \mathcal{F}_i(b) \big) \,. \tag{108.3}$$

Besonders interessant ist der Fall $[a,b] = [x_{j-1/2}, x_{j+1/2}]$, denn dann werden die Ortsintegrale über u (bis auf den Faktor h) durch die Dichtemittel u_{ij} aus (107.3) approximiert. Dies führt auf die folgende Definition:

Definition 108.1. Ein Verfahren der Form

$$u_{i+1,j} = u_{ij} - \gamma(g_{i,j+1/2} - g_{i,j-1/2})\,, \qquad \gamma = \tau/h\,, \tag{108.4}$$

heißt *Differenzenverfahren in Erhaltungsform*, falls eine Lipschitz-stetige Funktion $G : \mathbb{R}^2 \to \mathbb{R}$ existiert mit

$$g_{i,j-1/2} = G(u_{i,j-1}, u_{ij}) \qquad \text{für alle } j \in \mathbb{Z}\,; \tag{108.5}$$

die Funktion G heißt *numerischer Fluß*. Ein Differenzenverfahren in Erhaltungsform heißt *konsistent*, falls

$$G(u,u) = F(u)\,. \tag{108.6}$$

Bemerkungen. Für ein Differenzenverfahren in Erhaltungsform, bei dem u_{ij} gemäß (107.3) dem Integralmittel von $u_h(t_i)$ über $[x_{j-1/2}, x_{j+1/2}]$ entspricht, gilt offensichtlich (die absolute Konvergenz aller Reihen sei stillschweigend vorausgesetzt)

$$\begin{aligned}\int_{\mathbb{R}} u_h(t_{i+1})\,dx \;&=\; h\sum_{j\in\mathbb{Z}} u_{i+1,j} \;=\; h\sum_{j\in\mathbb{Z}} u_{ij} - \tau\sum_{j\in\mathbb{Z}} g_{i,j+1/2} + \tau\sum_{j\in\mathbb{Z}} g_{i,j-1/2} \\ &=\; h\sum_{j\in\mathbb{Z}} u_{ij} \;=\; \int_{\mathbb{R}} u_h(t_i)\,dx\end{aligned}$$

und induktiv folgt hieraus

$$\int_{\mathbb{R}} u_h(t_i)\,dx \;=\; \int_{\mathbb{R}} u_h(0)\,dx\,, \qquad i \in \mathbb{N}\,.$$

Wählt man nun noch $u_h(0) \in S_{0,\triangle'}$ als

$$u_h(0) \;=\; \sum_{j\in\mathbb{Z}} u_{0j}\chi_j \quad \text{mit} \quad u_{0j} \;=\; \frac{1}{h}\int_{x_{j-1/2}}^{x_{j+1/2}} u^\circ\,dx\,, \quad j \in \mathbb{Z}\,, \tag{108.7}$$

dann gilt also

$$\int_{\mathbb{R}} u_h(t_i)\,dx \;=\; \int_{\mathbb{R}} u^\circ\,dx \qquad \text{für alle} \quad i \in \mathbb{N}_0\,,$$

d. h. die numerische Approximation ist eine Erhaltungsgröße.

Die Bedingung (108.5) ist durch einen Vergleich von (108.3) und (108.4) motiviert: Demnach empfiehlt es sich, für die Definition von $g_{i,j-1/2}$ die diskreten Werte $u_{i,j-1}$ und u_{ij} an der „Nahtstelle“ $x_{j-1/2}$ heranzuziehen. Ist das Differenzenverfahren konsistent und die exakte Lösung hinreichend glatt, so konvergieren zwei benachbarte Näherungswerte $u_{i,j-1}$ und u_{ij} für $h, \tau \to 0$ gegen den gleichen Funktionswert u und $\mathcal{F}_i$ aus (108.2) gegen $F(u)$. Die Bedingung (108.6) ist daher für ein konsistentes Verfahren notwendig. ◊

Beispiele. Nach (107.7) ist das Godunov-Verfahren ein Differenzenverfahren in Erhaltungsform. Für konvexe Flüsse F ist der numerische Fluß $G(u, v)$ durch (107.8) gegeben; das Godunov-Verfahren erfüllt daher die Konsistenzbedingung (108.6).

Als zweites Beispiel betrachten wir das *Lax-Friedrichs-Verfahren*

$$u_{i+1,j} = u_{ij} - \gamma \, \frac{F(u_{i,j+1}) - F(u_{i,j-1})}{2} + \frac{u_{i,j+1} - 2u_{ij} + u_{i,j-1}}{2} \,. \tag{108.8}$$

Der zugehörige numerische Fluß lautet

$$G(u, v) \;=\; \frac{F(u) + F(v)}{2} \;+\; \frac{u - v}{2\gamma} \tag{108.9}$$

und daher ist auch das Lax-Friedrichs-Verfahren konsistent. ◊

Der mittlere Term auf der rechten Seite von (108.8) entspricht einer zentralen Differenz zur Approximation der Ortsableitung von $F(u)$. Speziell für lineare Probleme $F(u) = au$ scheint das der Empfehlung aus Abschnitt 104 zu widersprechen, einseitige Differenzen entgegen der Flußrichtung zu verwenden. Tatsächlich kann man jedoch (108.8) umschreiben als

$$\frac{u_{i+1,j} - u_{ij}}{\tau} \;+\; \frac{F(u_{i,j+1}) - F(u_{i,j-1})}{2h} \;=\; h \, \frac{1}{2\gamma} \, \frac{u_{i,j+1} - 2u_{ij} + u_{i,j-1}}{h^2} \,,$$

was für festes γ bezüglich der Ortsvariablen eine Approximation zweiter Ordnung an die *regularisierte Gleichung*

$$u_t + \frac{d}{dx} F(u) \;=\; \frac{h}{2\gamma} \, u_{xx}$$

ist. Dies erinnert an die Vorgehensweise in Abschnitt 85, vgl. (85.7), wo zentrale Differenzenquotienten für singulär gestörte Probleme durch Hinzunahme eines künstlichen Diffusionsterms im Differentialoperator stabilisiert wurden. Auf den Zusammenhang zwischen Erhaltungsgleichungen und singulär gestörten Problemen haben wir bereits in Bemerkung 104.6 hingewiesen.

Für die Stabilität des Lax-Friedrichs-Verfahrens spielt wieder eine CFL-Bedingung eine entscheidende Rolle, nämlich

$$\gamma \sup_{v\in\mathbb{R}} |a(v)| \;\le\; 1\,, \qquad a = F'\,. \tag{108.10}$$

Lemma 108.2. *Sei $u_h(0) \in S_{0,\triangle'}$ durch* (108.7) *gegeben. Dann ist das Lax-Friedrichs-Verfahren unter der CFL-Bedingung* (108.10) *bezüglich der Maximumnorm stabil, denn es gilt*

$$\inf_{z\in\mathbb{R}} u^\circ(z) \;\le\; u_h(x,t_i) \;\le\; \sup_{z\in\mathbb{R}} u^\circ(z) \qquad \textit{für alle } i \in \mathbb{N}_0 \textit{ und } x \in \mathbb{R}\,.$$

Beweis. Aus (108.8) folgt mit dem Mittelwertsatz

$$\begin{aligned} u_{i+1,j} \;&=\; \frac{1}{2}\,u_{i,j-1} \;+\; \frac{1}{2}\,u_{i,j+1} \;-\; \gamma\,\frac{a(\tilde u)}{2}\,(u_{i,j+1} - u_{i,j-1}) \\ &=\; \frac{1+\gamma a(\tilde u)}{2}\,u_{i,j-1} \;+\; \frac{1-\gamma a(\tilde u)}{2}\,u_{i,j+1}\,, \end{aligned}$$

wobei wieder $a = F'$ und $\tilde u$ eine Zwischenstelle zwischen $u_{i,j-1}$ und $u_{i,j+1}$ bezeichnet. Aufgrund der CFL-Bedingung sind beide Brüche in der unteren Zeile positiv und summieren sich zu Eins auf. Mit anderen Worten: $u_{i+1,j}$ ist eine Konvexkombination von $u_{i,j-1}$ und $u_{i,j+1}$ und daher folgt induktiv die Behauptung des Lemmas. (Die Induktionsvoraussetzung folgt aus der Definition von u_{0j} als Integralmittel von u°.) □

Während dieses Stabilitätsresultat sehr einfach zu beweisen ist, ist die Konvergenzuntersuchung des Lax-Friedrichs-Verfahrens schwieriger und verwendet tiefliegende Hilfsmittel aus der Analysis. Ein zentraler Pfeiler dieser Konvergenzanalyse ist dabei die Tatsache, daß die Totalvariation der Funktionen $u_h(t_i)$ in der Zeit monoton fällt. Die Totalvariation der stückweise konstanten Funktion $u_h(t_i)$ kann dabei durch

$$|u_h(t_i)|_{BV(\mathbb{R})} \;=\; \sum_{j\in\mathbb{Z}} |u_{i,j+1} - u_{ij}|$$

berechnet werden, vorausgesetzt die Summe konvergiert; $BV(\mathbb{R})$ ist der Raum aller Funktionen beschränkter Variation. Die Totalvariation ist eine Halbnorm in $BV(\mathbb{R})$. Numerische Verfahren, für die die Totalvariation von $u_h(t_i)$ monoton fallend ist, heißen *TVD-Verfahren* (engl.: *total variation diminishing*).

Lemma 108.3. *Das Lax-Friedrichs-Verfahren wird mit der CFL-Bedingung* (108.10) *ein TVD-Verfahren.*

Beweis. Gemäß (108.8) ist

$$\begin{aligned} u_{i+1,j+1} &= \frac{1}{2}\,(u_{i,j+2} + u_{ij}) \;-\; \frac{\gamma}{2}\,\big(F(u_{i,j+2}) - F(u_{ij})\big)\,, \\ u_{i+1,j} &= \frac{1}{2}\,(u_{i,j+1} + u_{i,j-1}) \;-\; \frac{\gamma}{2}\,\big(F(u_{i,j+1}) - F(u_{i,j-1})\big)\,, \end{aligned}$$

und nach Subtraktion folgt mit dem Mittelwertsatz

$$\begin{aligned} u_{i+1,j+1} - u_{i+1,j} &= \frac{u_{i,j+2} - u_{i,j+1}}{2} + \frac{u_{ij} - u_{i,j-1}}{2} \\ &\quad - \frac{\gamma}{2}\,\big(F(u_{i,j+2}) - F(u_{i,j+1})\big) + \frac{\gamma}{2}\,\big(F(u_{ij}) - F(u_{i,j-1})\big) \\ &= (u_{i,j+2} - u_{i,j+1})\big(\frac{1}{2} - \frac{\gamma}{2}\,a_{j+3/2}\big) + (u_{ij} - u_{i,j-1})\big(\frac{1}{2} + \frac{\gamma}{2}\,a_{j-1/2}\big), \end{aligned}$$

wobei $a_{j-1/2} = a(\tilde{u})$ durch eine Zwischenstelle $\tilde{u}$ zwischen $u_{i,j-1}$ und u_{ij} gegeben ist. Die Faktoren $(1 \pm a_{j+1/2}\gamma)/2$, $j \in \mathbb{Z}$, sind wegen der CFL-Bedingung (108.10) nichtnegativ, und daher ergibt die Dreiecksungleichung

$$\begin{aligned} |u_{i+1,j+1} - u_{i+1,j}| \;\le\; |u_{i,j+2} - u_{i,j+1}|\,\big(\frac{1}{2} - \frac{\gamma}{2}\,a_{j+3/2}\big) \\ + |u_{ij} - u_{i,j-1}|\,\big(\frac{1}{2} + \frac{\gamma}{2}\,a_{j-1/2}\big)\,. \end{aligned}$$

Durch Summation über j folgt schließlich

$$\begin{aligned} \sum_{j\in\mathbb{Z}} |u_{i+1,j+1} - u_{i+1,j}| &\le \sum_{j\in\mathbb{Z}} |u_{i,j+2} - u_{i,j+1}|\,\big(\frac{1}{2} - \frac{\gamma}{2}\,a_{j+3/2}\big) \\ &\qquad + \sum_{j\in\mathbb{Z}} |u_{ij} - u_{i,j-1}|\,\big(\frac{1}{2} + \frac{\gamma}{2}\,a_{j-1/2}\big) \\ &= \sum_{j\in\mathbb{Z}} |u_{ij} - u_{i,j-1}|\,\big(\frac{1}{2} - \frac{\gamma}{2}\,a_{j-1/2} + \frac{1}{2} + \frac{\gamma}{2}\,a_{j-1/2}\big) \\ &= \sum_{j\in\mathbb{Z}} |u_{ij} - u_{i,j-1}|\,, \end{aligned}$$

was $|u_h(t_{i+1})|_{BV(\mathbb{R})} \le |u_h(t_i)|_{BV(\mathbb{R})}$ entspricht. □

Die Stabilitätsanalyse aus diesem Abschnitt läßt sich auf einige andere Differenzenverfahren in Erhaltungsform übertragen, insbesondere auch auf das Godunov-Verfahren. Für eine weitergehende Konvergenzanalyse dieser Verfahren sei auf die eingangs genannten Bücher verwiesen.

109 Eine Ortsdiskretisierung höherer Ordnung

Die Treppenfunktionen u_h, die bisher verwendet wurden, sind gut zur Approximation unstetiger Lösungen geeignet. Auf der anderen Seite erlauben sie bei glatten Lösungen lediglich eine Approximation der Ordnung Eins. Im folgenden stellen wir andere (unstetige) Ansatzfunktionen vor, mit denen eine höhere Konsistenzordnung erzielt werden kann.

Definition 109.1. Eine Funktion $s : \mathbb{R} \to \mathbb{R}$ heißt *monotonieerhaltend stückweise linear* über dem verschobenen Gitter $\triangle' = \{x_{j-1/2} : j \in \mathbb{Z}\}$, falls s auf jedem Teilintervall $[x_{j-1/2}, x_{j+1/2})$ von $\triangle'$ eine lineare Funktion ist und falls gilt:

$$\begin{aligned} s(x_j) \le s(x_{j+1}) &\implies s \text{ ist monoton wachsend in } (x_j, x_{j+1}) , \\ s(x_j) \ge s(x_{j+1}) &\implies s \text{ ist monoton fallend in } (x_j, x_{j+1}) . \end{aligned}$$

Wir bezeichnen im weiteren mit $s_j = s(x_j)$, $j \in \mathbb{Z}$, die Funktionswerte einer solchen Funktion über $\triangle$. Aus Definition 109.1 folgt, daß s im Intervall $[x_{j-1/2}, x_{j+1/2})$ konstant ist, falls die Teilfolge $\{s_{j-1}, s_j, s_{j+1}\}$ nicht monoton ist. Aus dieser Beobachtung kann man leicht folgern, daß die Menge der monotonieerhaltenden stückweise linearen Funktionen *kein* linearer Raum ist.

Für monotonieerhaltende stückweise lineare Funktionen kann die Totalvariation wie folgt berechnet werden:

Proposition 109.2. *Sei s monotonieerhaltend stückweise linear über dem Gitter $\triangle'$. Dann ist*

$$|s|_{BV(\mathbb{R})} = \sum_{j\in\mathbb{Z}} |s(x_{j+1}) - s(x_j)| = \sum_{j\in\mathbb{Z}} |s(x_{j+1/2}\pm) - s(x_{j-1/2}\pm)| .$$

Beweis. Da s nach Voraussetzung in (x_j, x_{j+1}) für jedes $j \in \mathbb{Z}$ monoton ist, berechnet sich die Totalvariation von s über einem solchen Teilintervall als $|s(x_{j+1}) - s(x_j)|$. Durch Zusammensetzen der einzelnen Teilinvervalle ergibt sich unmittelbar die erste Gleichung. Die anderen beiden Darstellungen folgen entsprechend, da s nicht nur in (x_j, x_{j+1}) sondern damit automatisch auch in $(x_{j-1/2}, x_{j+3/2})$ monoton ist. □

Es sei an dieser Stelle betont, daß eine monotonieerhaltende stückweise lineare Funktion *nicht* durch ihre Funktionswerte $s_j = s(x_j)$ eindeutig festgelegt ist, da sie an den Punkten $x_{j-1/2}$ nicht stetig zu sein braucht. Allerdings existiert zu jeder Folge $\{s_j\}$ mindestens eine Funktion mit den entsprechenden Werten, nämlich die interpolierende Treppenfunktion über $\triangle'$. Eine weitere interpolierende Funktion, die sogenannte Minmod-Interpolierende, ist in Abbildung 109.1 dargestellt.

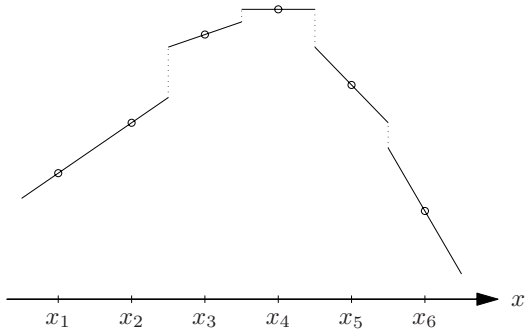


Abb. 109.1: Die Minmod-Interpolierende

Definition und Satz 109.3. *Zu jeder Folge* $\{s_j\}_{j\in\mathbb{Z}} \subset \mathbb{R}$ *existiert eine über* $\triangle'$ *monotonieerhaltende stückweise lineare Funktion* s *mit* $s(x_j) = s_j$ *und mit den Ableitungswerten*

$$s'(x_j) = s'_j = \frac{\operatorname{minmod}(s_{j+1} - s_j, s_j - s_{j-1})}{h}\,, \qquad j \in \mathbb{Z}\,, \tag{109.1}$$

wobei

$$\operatorname{minmod}(\xi,\eta) \;=\; \begin{cases} \xi\,, & \xi\eta > 0\,, \quad |\xi| \le |\eta|\,, \\ \eta\,, & \xi\eta > 0\,, \quad |\xi| > |\eta|\,, \\ 0\,, & \xi\eta \le 0\,. \end{cases}$$

Wir nennen s *die* Minmod-Interpolierende *der Daten* $\{s_j\}$ *über* $\triangle$. *Sind diese Daten die Funktionswerte einer zweimal stetig differenzierbaren Funktion* f, $s_j = f(x_j)$, $j \in \mathbb{Z}$, *dann gilt*

$$|f(x) - s(x)| \le \frac{5}{8}\, h^2\, \|f''\|_{[x_{j-1},x_{j+1}]} \tag{109.2}$$

für $x \in [x_{j-1/2}, x_{j+1/2})$.

Beweis. Die stückweise lineare Funktion s über $\triangle'$ ist durch die Interpolationsbedingung und die Ableitungswerte (109.1) festgelegt. Es bleibt nachzuweisen, daß s monotonieerhaltend ist. Zunächst zeigen wir, daß s in (x_j, x_{j+1}) monoton wachsend ist, falls $\xi = s_{j+1} - s_j \ge 0$ gilt. Für $\eta = s_j - s_{j-1} \le 0$ ergibt sich $s'_j = 0$ aus (109.1), andernfalls ist s'_j einer der beiden nichtnegativen Werte ξ oder η. Damit ist s'_j immer nichtnegativ, also s im Intervall $(x_j, x_{j+1/2})$

monoton wachsend. Im anderen Teilintervall $(x_{j+1/2}, x_{j+1})$ argumentiert man entsprechend mit $\eta = s_{j+2} - s_{j+1}$. Es bleibt noch die Monotonie an der Sprungstelle zu untersuchen. Hierzu beweisen wir die Ungleichungskette

$$s(x_{j+1/2}-) \;\le\; (s_j + s_{j+1})/2 \;\le\; s(x_{j+1/2}+)\,. \tag{109.3}$$

Da wir $s_{j+1} \ge s_j$ angenommen haben, ist die erste Ungleichung sicher richtig, wenn $s'_j = 0$ ist. Andernfalls haben $s_{j+1} - s_j$ und $s_j - s_{j-1}$ das gleiche (positive) Vorzeichen und folglich ist

$$s'_j \;=\; \min\Big\{\, \frac{s_{j+1} - s_j}{h}\,,\, \frac{s_j - s_{j-1}}{h}\,\Big\} \;\le\; \frac{s_{j+1} - s_j}{h}\,.$$

Damit ergibt sich jedoch

$$s(x_{j+1/2}-) \;=\; s_j + s'_j h/2 \;\le\; (s_j + s_{j+1})/2\,.$$

Entsprechend weist man die zweite Ungleichung in (109.3) nach. Damit ist der Monotoniebeweis für den Fall $s_j \le s_{j+1}$ abgeschlossen. Der Beweis für $s_j \ge s_{j+1}$ wird in der gleichen Weise geführt.

Für die Abschätzung (109.2) betrachten wir zunächst den Fall $s'_j = 0$. In diesem Fall folgt aus (109.1), daß die Folge $\{s_{j-1}, s_j, s_{j+1}\}$ ein Extremum in s_j besitzt. Die Funktion f hat dann an einer Stelle $\tilde{x} \in (x_{j-1}, x_{j+1})$ ebenfalls ein lokales Extremum, d. h. $f'(\tilde{x}) = 0 = s'_j$, und daraus folgt

$$\begin{aligned} |f'(x_j) - s'_j| \;&=\; |f'(x_j) - f'(\tilde{x})| \;=\; \Big|\, \int_{\tilde{x}}^{x_j} f''(\xi)\, d\xi \,\Big| \\ &\le\; h\, \|f''\|_{[x_{j-1}, x_{j+1}]}\,. \end{aligned} \tag{109.4}$$

Ist hingegen $s'_j \ne 0$, dann stimmt s'_j mit der Steigung der Sekante an f in einem der beiden Teilintervalle $[x_{j-1}, x_j]$ oder $[x_j, x_{j+1}]$ überein und nach dem Mittelwertsatz gibt es wieder eine Stelle $\tilde{x} \in [x_{j-1}, x_{j+1}]$ mit $f'(\tilde{x}) = s'_j$. Demnach gilt (109.4) auch in diesem Fall. Durch Taylorentwicklung von $f - s$ um x_j folgt nun für jedes $x \in [x_{j-1/2}, x_{j+1/2})$ die Existenz eines Zwischenpunktes ξ im gleichen Intervall mit

$$f(x) - s(x) = (f'(x_j) - s'_j)(x - x_j) + \frac{1}{2} f''(\xi)(x - x_j)^2\,,$$

und wegen $|x - x_j| \le h/2$ folgt daraus zusammen mit (109.4) die Behauptung

$$|f(x) - s(x)| \;\le\; \frac{5}{8}\, h^2\, \|f''\|_{[x_{j-1}, x_{j+1}]}$$

für alle $x \in [x_{j-1/2}, x_{j+1/2})$. □

Wir untersuchen nun noch die Sprungstellen der Minmod-Interpolierenden.

Lemma 109.4. *Sei $f \in C^3(\mathbb{R})$ und s die Minmod-Interpolierende zu f über $\triangle$. Ferner sei $f'(0) \neq 0$. Dann gilt für hinreichend kleine $h > 0$*

$$s(h/2-) \;=\; \begin{cases} f(h/2) - \frac{3}{8} f''(0)h^2 + O(h^3)\,, & f'(0)f''(0) \geq 0\,, \\ f(h/2) + \frac{1}{8} f''(0)h^2 + O(h^3)\,, & f'(0)f''(0) < 0\,, \end{cases}$$

und

$$s(-h/2+) \;=\; \begin{cases} f(-h/2) + \frac{1}{8} f''(0)h^2 + O(h^3)\,, & f'(0)f''(0) \geq 0\,, \\ f(-h/2) - \frac{3}{8} f''(0)h^2 + O(h^3)\,, & f'(0)f''(0) < 0\,. \end{cases}$$

Beweis. Für den Beweis unterscheiden wir verschiedene Fälle.

1. $f'(0) > 0$, $f''(0) > 0$:

In diesem Fall ist f in einer Umgebung des Nullpunkts monoton wachsend und konvex; folglich ist für hinreichend kleine $h > 0$

$$s_0' \;=\; \frac{s_0 - s_{-1}}{h} \;=\; \frac{f(0) - f(-h)}{h} \;=\; f'(0) - \frac{1}{2} f''(0)h + O(h^2)$$

und

$$\begin{aligned} s(h/2-) \;&=\; s_0 + \frac{1}{2} s_0' h \;=\; f(0) + \frac{1}{2} f'(0)h - \frac{1}{4} f''(0)h^2 + O(h^3) \\ &=\; f(h/2) - \frac{3}{8} f''(0)h^2 + O(h^3)\,. \end{aligned}$$

Die gleiche Auswahl von s_0' tritt ein, wenn f monoton fallend und konkav ist, also für $f'(0) < 0$ und $f''(0) < 0$.

2. $f'(0) > 0$, $f''(0) < 0$:

In diesem Fall ergibt sich entsprechend

$$s_0' \;=\; \frac{s_1 - s_0}{h} \;=\; \frac{f(h) - f(0)}{h} \;=\; f'(0) + \frac{1}{2} f''(0)h + O(h^2)$$

und daher ist

$$\begin{aligned} s(h/2-) \;&=\; s_0 + \frac{1}{2} s_0' h \;=\; f(0) + \frac{1}{2} f'(0)h + \frac{1}{4} f''(0)h^2 + O(h^3) \\ &=\; f(h/2) + \frac{1}{8} f''(0)h^2 + O(h^3)\,. \end{aligned}$$

Auch hier gilt dasselbe Resultat für $f'(0) < 0$ und $f''(0) > 0$.

3. $f'(0) \neq 0$, $f''(0) = 0$:

Falls $f''(0)$ verschwindet, ergibt sich in jedem Fall

$$s(h/2-) \;=\; f(h/2) + O(h^3)$$

wie behauptet.

Die Behauptung für $s(-h/2+)$ wird entsprechend bewiesen. □

Um die Lösung u der Erhaltungsgleichung durch stückweise lineare Funktionen zu approximieren, gehen wir analog zu Abschnitt 107 vor und interpolieren für festes $t = t_i$ die Dichtemittel

$$u_{ij} \;\approx\; \frac{1}{h} \int_{x_{j-1/2}}^{x_{j+1/2}} u(x, t_i)\, dx$$

durch die Minmod-Interpolierende $u_h(t_i)$ über $\triangle$:

$$u_h(x_j, t_i) \;=\; u_{ij}\,, \qquad j \in \mathbb{Z}\,.$$

Für ein Differenzenschema (108.4) in Erhaltungsform bietet es sich dann an, den numerischen Fluß an den Grenzwerten von u_h an den Gitterpunkten von $\triangle'$ auszuwerten:

$$g_{i,j-1/2} \;=\; G\big(u_h(x_{j-1/2}-, t_i), u_h(x_{j-1/2}+, t_i)\big)\,. \tag{109.5}$$

Auf diese Weise erreichen wir Konsistenzordnung zwei für die Diskretisierung der Ortsableitung in der Erhaltungsgleichung.

Proposition 109.5. *Es seien $f, F \in C^3(\mathbb{R})$ mit $f'(0) \neq 0$ und s die Minmod-Interpolierende zu f über $\triangle$, d. h. $s(x_j) = f(x_j)$ für alle $j \in \mathbb{Z}$. Ferner sei $G(u, v)$ ein stetig differenzierbarer konsistenter numerischer Fluß mit Lipschitzstetigem Gradienten* $\operatorname{grad} G = [G_u, G_v]^T$. *Ist $g_{\pm 1/2} = G\big(s(\pm h/2-), s(\pm h/2+)\big)$, dann gilt*

$$\frac{g_{1/2} - g_{-1/2}}{h} \;=\; \frac{d}{dx}\, F\big(f(x)\big)\Big|_{x=0} \;+\; O(h^2)\,, \qquad h \to 0\,.$$

Beweis. Nach Satz 109.3 ist

$$|s(h/2-) - f(h/2)| \;=\; O(h^2)\,, \qquad |s(h/2+) - f(h/2)| \;=\; O(h^2)\,,$$

und entsprechende Abschätzungen gelten am linken Gitterrand $-h/2$. Daher ergibt eine Taylorentwicklung von G um $\hat{u} = f(h/2)$ mit $s_0 = s(0) = f(0)$:

$$\begin{aligned} g_{1/2} \;&=\; G(\hat{u}, \hat{u}) + G_u(\hat{u}, \hat{u})\big(s(h/2-) - \hat{u}\big) + G_v(\hat{u}, \hat{u})\big(s(h/2+) - \hat{u}\big) + O(h^4) \\ &=\; F(\hat{u}) \;+\; G_u(s_0, s_0)\big(s(h/2-) - \hat{u}\big) \;+\; G_v(s_0, s_0)\big(s(h/2+) - \hat{u}\big) \\ &\qquad +\; \big(G_u(\hat{u}, \hat{u}) - G_u(s_0, s_0)\big)\big(s(h/2-) - \hat{u}\big) \\ &\qquad +\; \big(G_v(\hat{u}, \hat{u}) - G_v(s_0, s_0)\big)\big(s(h/2+) - \hat{u}\big) \;+\; O(h^4)\,. \end{aligned}$$

Da G_u und G_v Lipschitz-stetig sind und da $\hat{u} - s_0 = f(h/2) - f(0) = O(h)$, vereinfacht sich dies zu

$$\begin{aligned} g_{1/2} \;=\; & F(\hat{u}) + G_u(s_0, s_0)\big(s(h/2-) - \hat{u}\big) \\ & + G_v(s_0, s_0)\big(s(h/2+) - \hat{u}\big) \;+\; O(h^3)\,. \end{aligned}$$

Entsprechend ergibt sich

$$\begin{aligned} g_{-1/2} \;=\; & F(\check{u}) + G_u(s_0, s_0)\big(s(-h/2-) - \check{u}\big) \\ & + G_v(s_0, s_0)\big(s(-h/2+) - \check{u}\big) \;+\; O(h^3) \end{aligned}$$

mit der Abkürzung $\check{u}$ für $f(-h/2)$. Durch Subtraktion dieser beiden Resultate erhalten wir daher

$$\begin{aligned} g_{1/2} - g_{-1/2} \;=\; & F(\hat{u}) - F(\check{u}) \\ & + G_u(s_0, s_0)\big(s(h/2-) - \hat{u} - s(-h/2-) + \check{u}\big) \\ & + G_v(s_0, s_0)\big(s(h/2+) - \hat{u} - s(-h/2+) + \check{u}\big) \;+\; O(h^3)\,. \end{aligned} \tag{109.6}$$

Da $f'(0) \neq 0$ angenommen war, kann für hinreichend kleine $h > 0$ der Abstand $s(h/2-) - \hat{u} = s(h/2-) - f(h/2)$ gemäß Lemma 109.4 durch

$$s(h/2-) - \hat{u} \;=\; \alpha f''(0)h^2 + O(h^3)$$

abgeschätzt werden, wobei α entweder $-3/8$ oder $1/8$ ist, abhängig vom Vorzeichen von $f'(0)f''(0)$. Nach Voraussetzung ist $f'(0)$ von Null verschieden und wir können ohne Einschränkung annehmen, daß auch $f''(0) \neq 0$ ist, denn ansonsten ist der Wert von α ohnehin nicht von Bedeutung, da er mit Null multipliziert wird. Unter diesen Voraussetzungen haben $f'(0)f''(0)$ und $f'(-h)f''(-h)$ für hinreichend kleine h das gleiche Vorzeichen und damit ergibt Lemma 109.4 (auf das benachbarte Gitterintervall angewendet)

$$\begin{aligned} s(-h/2-) - \check{u} \;&=\; s(-h/2-) - f(-h/2) \;=\; \alpha f''(-h)h^2 + O(h^3) \\ &=\; \alpha f''(0)h^2 + O(h^3) \end{aligned}$$

mit dem gleichen Wert von α. Daraus folgt

$$s(h/2-) - \hat{u} - s(-h/2-) + \check{u} \;=\; O(h^3) \tag{109.7}$$

und entsprechend ist

$$s(h/2+) - \hat{u} - s(-h/2+) + \check{u} \;=\; O(h^3)\,.$$

Eingesetzt in (109.6) ergibt sich daher die Behauptung

$$\begin{aligned} g_{1/2} - g_{-1/2} \;&=\; F(f(h/2)) - F(f(-h/2)) + O(h^3) \\ &=\; h\,\frac{d}{dx} F(f(x))\,\Big|_{x=0} + O(h^3)\,, \end{aligned}$$

wobei in der letzten Zeile die Approximationsordnung des zentralen Differenzenquotienten ausgenutzt wurde, vgl. Lemma 83.2. □

Bemerkung. An lokalen Extrema von f geht die zweite Approximationsordnung verloren. Bestimmt man an diesen Stellen die Sprunghöhen der Minmod-Interpolierenden in entsprechender Weise, so stellt sich heraus, daß in diesem Fall die linke Seite von (109.7) nur durch $O(h^2)$ abgeschätzt werden kann. ◊

Ersetzt man in Proposition 109.5 die Funktion f durch $u(t)$, dann ergibt $(g_{-1/2} - g_{1/2})/h$ eine Approximation zweiter Ordnung für die rechte Seite der Differentialgleichung

$$u_t \;=\; -\,\frac{d}{dx}F(u)$$

an der Stelle $x = 0$, sofern $u(0,t)$ kein lokales Extremum von $u(t)$ ist. Die in diesem Abschnitt vorgestellte Ortsdiskretisierung wird *MUSCL-Schema* (engl.: *Monotone Upwind Schemes for Conservation Laws*) genannt.

110 Zeitintegration des MUSCL-Schemas

Das MUSCL-Schema aus dem vorangegangenen Abschnitt liefert eine örtliche Diskretisierung der Erhaltungsgleichung mit Konsistenzordnung zwei. Für die Zeitintegration sollte daher anstelle des Euler-Verfahrens ein Runge-Kutta-Verfahren mit derselben Konsistenzordnung verwendet werden.

Ausgangspunkt unserer Überlegungen ist erneut die integrale Form (108.3) der Erhaltungsgleichung. Verwenden wir für $u_h(t_i)$ die Minmod-Interpolierende der diskreten Werte u_{ij} über $\triangle$, so ist

$$u_{ij} \;=\; \frac{1}{h}\int_{x_{j-1/2}}^{x_{j+1/2}} u_h(x,t_i)\,dx \;\approx\; \frac{1}{h}\int_{x_{j-1/2}}^{x_{j+1/2}} u(x,t_i)\,dx$$

und (108.3) ergibt

$$u_{i+1,j} \;\approx\; u_{ij} \;+\; \tau(\mathcal{F}_i(x_{j-1/2}) - \mathcal{F}_i(x_{j+1/2}))\,, \tag{110.1}$$

wobei $\mathcal{F}_i$ den mittleren Fluß

$$\mathcal{F}_i(x) \;=\; \frac{1}{\tau}\int_{t_i}^{t_{i+1}} F(u(x,t))\,dt$$

bezeichnet.

Der Ansatzpunkt der Runge-Kutta-Verfahren besteht darin, dieses Integral durch eine Quadraturformel zu approximieren. Entsprechende Verfahren zweiter Ordnung beruhen auf der Mittelpunktformel (Verfahren von Runge, Beispiel 76.1) oder der Trapezformel (Verfahren von Heun, Beispiel 76.2). Für die Integration von Erhaltungsgleichungen ist das *Verfahren von Heun* überlegen, das (für $j \in \mathbb{Z}$) auf den Ansatz

$$\mathcal{F}_i(x_{j+1/2}) \approx \frac{1}{2} F(u_h(x_{j+1/2}, t_i)) + \frac{1}{2} F(\eta_i(x_{j+1/2})) \tag{110.2}$$

mit $\eta_i \approx u(t_{i+1})$ führt. Im Hinblick auf mögliche Unstetigkeiten der Lösung wird dieser Ausdruck wieder durch einen numerischen Fluß approximiert, der hier analog zu (109.5) die Gestalt

$$g_{i,j+1/2} = \frac{1}{2} G(u_{i,j+1/2-}, u_{i,j+1/2+}) + \frac{1}{2} G(\eta_{i,j+1/2-}, \eta_{i,j+1/2+})$$

hat; dabei sind

$$u_{i,j+1/2\pm} = u_h(x_{j+1/2}\pm, t_i)\,, \qquad \eta_{i,j+1/2\pm} = \eta_i(x_{j+1/2}\pm)\,,$$

die zugehörigen rechts- und linksseitigen Grenzwerte von $u_h(t_i)$ und η_i an den Gitterpunkten des verschobenen Gitters $\triangle'$.

Die Berechnung von $\eta_{i,j+1/2\pm}$ erfolgt gemäß dem Verfahren von Heun mit einem expliziten Euler-Schritt. Da hier Funktionswerte von η_i gesucht sind (und keine Integralmittel) wird hierfür die Erhaltungsgleichung in ihrer differentiellen Form verwendet: Die Ortsableitung von $F(u_h)$ wird dabei durch einseitige Differenzenquotienten innerhalb des jeweiligen Gitterintervalls von $\triangle'$ approximiert. Dies ergibt das *MUSCL-Verfahren*

$$\eta_{i,j+1/2-} = u_{i,j+1/2-} - \gamma(F(u_{i,j+1/2-}) - F(u_{i,j-1/2+}))\,, \tag{110.3a}$$

$$\eta_{i,j+1/2+} = u_{i,j+1/2+} - \gamma(F(u_{i,j+3/2-}) - F(u_{i,j+1/2+}))\,, \tag{110.3b}$$

$$g_{i,j+1/2} = \frac{1}{2} G(u_{i,j+1/2-}, u_{i,j+1/2+}) + \frac{1}{2} G(\eta_{i,j+1/2-}, \eta_{i,j+1/2+})\,, \tag{110.3c}$$

$$u_{i+1,j} = u_{ij} - \gamma(g_{i,j+1/2} - g_{i,j-1/2})\,, \tag{110.3d}$$

das bei konsistentem (und hinreichend glattem) numerischen Fluß Konsistenzordnung zwei aufweist.

Die Startwerte des Verfahrens ergeben sich aus den Massemitteln

$$u_{0j} = \frac{1}{h} \int_{x_{j-1/2}}^{x_{j+1/2}} u^\circ(x)\, dx\,, \qquad j \in \mathbb{Z}\,, \tag{110.4}$$

und der zugehörigen Minmod-Interpolierenden $u_h(0)$ über $\triangle$.

Lemma 110.1. *Sei $u_h(0)$ die Minmod-Interpolierende der mittleren Massen $\{u_{0j}\}$ aus (110.4) über $\triangle$. Dann gilt*

$$|u_h(0)|_{BV(\mathbb{R})} \;\le\; |u^\circ|_{BV(\mathbb{R})}\,.$$

Beweis. Nach Proposition 109.2 ist

$$\begin{aligned}
|u_h(0)|_{BV(\mathbb{R})} \;&=\; \sum_{j\in\mathbb{Z}} |u_{0,j+1} - u_{0j}| \;=\; \sum_{j\in\mathbb{Z}} \frac{1}{h} \Big| \int_{x_{j-1/2}}^{x_{j+1/2}} \big(u^\circ(x+h) - u^\circ(x)\big)\,dx \Big| \\
&\le\; \frac{1}{h} \sum_{j\in\mathbb{Z}} \int_{x_{j-1/2}}^{x_{j+1/2}} \big|u^\circ(x+h) - u^\circ(x)\big|\,dx \\
&=\; \frac{1}{h} \sum_{j\in\mathbb{Z}} \int_{-h/2}^{h/2} \big|u^\circ(x_j+z+h) - u^\circ(x_j+z)\big|\,dz\,.
\end{aligned}$$

Eine Vertauschung von Summation und Integration ergibt daher

$$|u_h(0)|_{BV(\mathbb{R})} \;\le\; \frac{1}{h} \int_{-h/2}^{h/2} \Big(\sum_{j\in\mathbb{Z}} \big|u^\circ(x_j+z+h) - u^\circ(x_j+z)\big| \Big)\,dz$$

und die Summe in der runden Klammer ist für jedes z eine untere Schranke für die Totalvariation von u°. Folglich ist

$$|u_h(0)|_{BV(\mathbb{R})} \;\le\; \frac{1}{h} \int_{-h/2}^{h/2} |u^\circ|_{BV(\mathbb{R})}\,dz \;=\; |u^\circ|_{BV(\mathbb{R})}\,.$$

□

Im weiteren soll exemplarisch der Fall untersucht werden, daß G der Godunov-Fluß und F monoton wachsend und konvex ist. Nach (107.8) hat der Godunov-Fluß dann die einfache Form

$$G(u_-, u_+) \;=\; F(u_-)\,, \tag{110.5}$$

so daß der Schritt (110.3b) überflüssig wird. Algorithmus 110.1 faßt diesen Spezialfall des MUSCL-Verfahrens (110.3) zusammen.

Satz 110.2. *Sei F monoton wachsend und konvex. Dann erfüllen die Näherungsfunktionen $u_h(t_i)$ aus Algorithmus* 110.1 *unter der CFL-Bedingung*

$$\gamma \sup_{v\in\mathbb{R}} |a(v)| \;\le\; 1 \tag{110.6}$$

die TVD-Eigenschaft

$$|u_h(t_{i+1})|_{BV(\mathbb{R})} \;\le\; |u_h(t_i)|_{BV(\mathbb{R})} \;\le\; |u^\circ|_{BV(\mathbb{R})}\,, \qquad i = 0, 1, \ldots$$

Initialisierung: F sei monoton wachsend und konvex; $\triangle$ habe Gitterweite h; ferner sei $\gamma \le 1$

bestimme u_{0j} durch (110.4)
$\tau = \gamma h$
`for` $i = 0, 1, 2, \ldots$ `do`
  $t_i = i\tau$
  berechne Minmod-Interpolierende $u_h(t_i)$ von $\{u_{ij}\}_{j\in\mathbb{Z}}$ über $\triangle$
  `for` $j \in \mathbb{Z}$ `do`
    % $u_{i,j+1/2\pm} = u_h(x_{j+1/2}\pm, t_i)$
    $\eta_{i,j+1/2-} = u_{i,j+1/2-} - \gamma(F(u_{i,j+1/2-}) - F(u_{i,j-1/2+}))$
    $g_{i,j+1/2} = (F(u_{i,j+1/2-}) + F(\eta_{i,j+1/2-}))/2$
    $u_{i+1,j} = u_{ij} - \gamma(g_{i,j+1/2} - g_{i,j-1/2})$
  `end for`
`end for`

Algorithmus 110.1: MUSCL-Verfahren mit Godunov-Fluß (F monoton und konvex)

Beweis. Wir betrachten im folgenden einen festen Zeitschritt des Verfahrens und vereinfachen die Notation, indem wir den Index i weglassen; statt dessen schreiben wir u_j für u_{ij} und v_j für $u_{i+1,j}$.

Wir beginnen mit dem numerischen Fluß: Mit dem Mittelwertsatz ergibt sich aus (110.3c) und (110.5)

$$\begin{aligned} g_{j+1/2} - g_{j-1/2} &= \frac{1}{2}\big(F(u_{j+1/2-}) - F(u_{j-1/2-}) + F(\eta_{j+1/2-}) - F(\eta_{j-1/2-})\big) \\ &= \frac{1}{2}\big(a_j^-(u_{j+1/2-} - u_{j-1/2-}) + \tilde{a}_j(\eta_{j+1/2-} - \eta_{j-1/2-})\big)\,, \end{aligned} \tag{110.7}$$

wobei $a_j^- = F'(u_j^-)$ und $\tilde{a}_j = F'(\tilde{\eta}_j)$ geeignete Werte von $a = F'$ an gewissen Stellen u_j^- und $\tilde{\eta}_j$ zwischen $u_{j-1/2-}$ und $u_{j+1/2-}$ bzw. zwischen $\eta_{j-1/2-}$ und $\eta_{j+1/2-}$ bezeichnen.

Entsprechend kann die Differenz $\eta_{j+1/2-} - \eta_{j-1/2-}$ ausgedrückt werden: Aus (110.3a) folgt

$$\begin{aligned} \eta_{j+1/2-} - \eta_{j-1/2-} &= u_{j+1/2-} - u_{j-1/2-} \\ &\quad - \gamma\big(F(u_{j+1/2-}) - F(u_{j-1/2-})\big) + \gamma\big(F(u_{j-1/2+}) - F(u_{j-3/2+})\big) \\ &= (1 - \gamma a_j^-)(u_{j+1/2-} - u_{j-1/2-}) + \gamma a_{j-1}^+(u_{j-1/2+} - u_{j-3/2+}) \end{aligned}$$

mit demselben a_j^- wie zuvor und $a_{j-1}^+ = F'(u_{j-1}^+)$ an einer Zwischenstelle u_{j-1}^+

zwischen $u_{j-3/2+}$ und $u_{j-1/2+}$. Eingesetzt in (110.7) ergibt dies

$$\begin{aligned} g_{j+1/2} - g_{j-1/2} \;=\; & \frac{1}{2\gamma}\alpha_j(u_{j+1/2-} - u_{j-1/2-}) \\ & + \frac{1}{2\gamma}\beta_{j-1}(u_{j-1/2+} - u_{j-3/2+}) \end{aligned} \tag{110.8}$$

mit

$$\alpha_j \;=\; \gamma a_j^- + \gamma\tilde{a}_j(1-\gamma a_j^-)\,, \qquad \beta_{j-1} \;=\; \gamma\tilde{a}_j\gamma a_{j-1}^+\,. \tag{110.9}$$

Aufgrund unserer Voraussetzungen an $a = F'$ und der CFL-Bedingung (110.6) sind α_j und β_{j-1} nichtnegativ; ferner ist

$$\alpha_j - 1 \;=\; (\gamma a_j^- - 1)(1-\gamma\tilde{a}_j) \le 0$$

aufgrund der CFL-Bedingung. Damit haben wir die Schranken

$$0 \;\le\; \alpha_j, \beta_{j-1} \;\le\; 1\,. \tag{110.10}$$

Nun kommen wir zu dem neuen Wert v_j von $u_h(t_{i+1})$ an der Stelle x_j. Aus (110.3d) und (110.8) folgt

$$v_j \;=\; u_j \;-\; \frac{\alpha_j}{2}\,(u_{j+1/2-} - u_{j-1/2-}) \;-\; \frac{\beta_{j-1}}{2}\,(u_{j-1/2+} - u_{j-3/2+})\,.$$

Da u_h im Gitterintervall $(x_{j-1/2}, x_{j+1/2})$ linear ist, gilt die Mittelwerteigenschaft $u_j = (u_{j-1/2+} + u_{j+1/2-})/2$ und somit ist

$$\begin{aligned} v_j \;=\; & \frac{1}{2}\,u_{j-1/2+} + \frac{1}{2}\,u_{j+1/2-} - \frac{\alpha_j}{2}\,(u_{j+1/2-} - u_{j-1/2-}) \\ & - \frac{\beta_{j-1}}{2}\,(u_{j-1/2+} - u_{j-3/2+})\,. \end{aligned} \tag{110.11}$$

Für die Differenz $v_{j+1} - v_j$ folgt hieraus

$$\begin{aligned} v_{j+1} - v_j \;=\; & \frac{1-\beta_j}{2}(u_{j+1/2+} - u_{j-1/2+}) + \frac{\beta_{j-1}}{2}(u_{j-1/2+} - u_{j-3/2+}) \\ & + \frac{1-\alpha_{j+1}}{2}(u_{j+3/2-} - u_{j+1/2-}) + \frac{\alpha_j}{2}(u_{j+1/2-} - u_{j-1/2-})\,. \end{aligned}$$

Da wegen (110.10) alle Brüche auf der rechten Seite nichtnegativ sind, ergibt dies

$$\begin{aligned} |v_{j+1} - v_j| \;\le\; & \frac{1-\beta_j}{2}\,|u_{j+1/2+} - u_{j-1/2+}| + \frac{\beta_{j-1}}{2}\,|u_{j-1/2+} - u_{j-3/2+}| \\ & + \frac{1-\alpha_{j+1}}{2}\,|u_{j+3/2-} - u_{j+1/2-}| + \frac{\alpha_j}{2}\,|u_{j+1/2-} - u_{j-1/2-}| \end{aligned}$$

und Summation über j liefert

$$\sum_{j\in\mathbb{Z}} |v_{j+1} - v_j| \;\le\; \frac{1}{2}\sum_{j\in\mathbb{Z}} \bigl(1 - \beta_j + \beta_j\bigr)\, |u_{j+1/2+} - u_{j-1/2+}| \\ + \;\frac{1}{2}\sum_{j\in\mathbb{Z}} \bigl(1 - \alpha_j + \alpha_j\bigr)\, |u_{j+1/2-} - u_{j-1/2-}|\,.$$

Die beiden Summen auf der rechten Seite dieser Ungleichung entsprechen nach Proposition 109.2 der Totalvariation von $u_h(t_i)$; aus demselben Grund ist die linke Seite die Totalvariation von $u_h(t_{i+1})$. Somit ist gezeigt, daß

$$|u_h(t_{i+1})|_{BV(\mathbb{R})} \;\le\; \frac{1}{2}\,|u_h(t_i)|_{BV(\mathbb{R})} \;+\; \frac{1}{2}\,|u_h(t_i)|_{BV(\mathbb{R})} \;=\; |u_h(t_i)|_{BV(\mathbb{R})}\,,$$

und durch Induktion folgt

$$|u_h(t_i)|_{BV(\mathbb{R})} \;\le\; |u_h(0)|_{BV(\mathbb{R})}\,, \qquad i = 1, 2, \ldots$$

Zusammen mit Lemma 110.1 folgt schließlich die Behauptung des Satzes. □

Zum Abschluß zeigen wir noch, daß Algorithmus 110.1 bezüglich der Maximumnorm stabil ist.

Satz 110.3. *Unter den Voraussetzungen von Satz* 110.2 *gilt für die Näherungslösung u_h von Algorithmus* 110.1 *die Ungleichung*

$$\inf_{z\in\mathbb{R}} u^\circ(z) \;\le\; u_h(x, t_i) \;\le\; \sup_{z\in\mathbb{R}} u^\circ(z)$$

für alle $i \in \mathbb{N}_0$ und $x \in \mathbb{R}$.

Beweis. Aus der Gleichung (110.11) im Beweis von Satz 110.2 folgt

$$u_{i+1,j} \;=\; \frac{1-\beta_{j-1}}{2}\, u_{i,j-1/2+} \;+\; \frac{1-\alpha_j}{2}\, u_{i,j+1/2-} \\ + \;\frac{\alpha_j}{2}\, u_{i,j-1/2-} \;+\; \frac{\beta_{j-1}}{2}\, u_{i,j-3/2+}\,.$$

Wegen (110.10) sind alle vier Faktoren vor den Werten $u_{i,j-1/2\pm}$, $u_{i,j+1/2-}$ und $u_{i,j-3/2+}$ nichtnegativ und summieren sich zu Eins. Folglich ist $u_{i+1,j}$ eine Konvexkombination dieser vier Grenzwerte von u_h und es gilt

$$\inf_{z\in\mathbb{R}} u_h(z, t_i) \;\le\; u_{i+1,j} \;\le\; \sup_{z\in\mathbb{R}} u_h(z, t_i), \qquad j \in \mathbb{Z}\,.$$

Aus der Monotonieerhaltung der Minmod-Interpolierenden $u_h(t_{i+1})$ der Werte $\{u_{i+1,j}\}$ über $\triangle$ folgt ferner, daß

$$\inf_{j\in\mathbb{Z}} u_{i+1,j} \;\le\; u_h(x, t_{i+1}) \;\le\; \sup_{j\in\mathbb{Z}} u_{i+1,j}$$

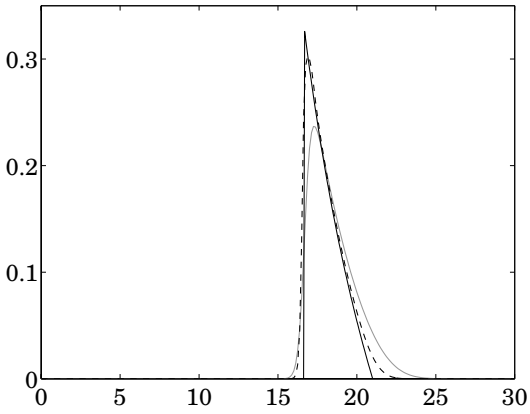


Abb. 110.1: Numerisches Ergebnis des MUSCL-Verfahrens

und somit gilt

$$\inf_{z\in\mathbb{R}} u_h(z,t_i) \;\le\; u_h(x,t_{i+1}) \;\le\; \sup_{z\in\mathbb{R}} u_h(z,t_i)$$

für alle $x \in \mathbb{R}$. Die Behauptung folgt somit durch vollständige Induktion, wobei noch beachtet werden muß, daß $u_h(0)$ aufgrund der Konstruktion (110.4) seinerseits durch das Supremum beziehungsweise Infimum von u° beschränkt ist. □

Beispiel. Abbildung 110.1 vergleicht die Ergebnisse des Godunov-Verfahrens und des MUSCL-Algorithmus 110.1 (jeweils mit $\gamma = 1$) für das Chromatographie-Beispiel 107.2. Während beide Verfahren die Sprungstelle der Ausflußkurve gut erkennen, approximiert das MUSCL-Verfahren (die gebrochene dunkle Linie) den exakten Ausfluß (dunkle Linie) deutlich besser als das Godunov-Verfahren (helle Linie). Für beide Verfahren wurde die Chromatographiesäule (der Ortsbereich) mit 100 Gitterintervallen diskretisiert. Um mit dem MUSCL-Verfahren die Qualität der Godunov-Näherung zu erzielen, hätten jedoch schon 40 Gitterintervalle ausgereicht. ◊

111 Systeme von Erhaltungsgleichungen

Als letztes gehen wir noch kurz auf den Fall mehrerer Erhaltungsgrößen ein, bei dem die Funktionswerte $u(x,t)$ Vektoren mit m reellen Komponenten sind.

In der differentiellen Form lautet die Erhaltungsgleichung dann

$$u_t + \frac{d}{dx}F(u) = u_t + A(u)u_x = 0\,, \qquad u(0) = u^\circ\,, \tag{111.1}$$

wobei $A \in \mathbb{R}^{m\times m}$ die Jacobi-Matrix von F ist. Ein solches System von Erhaltungsgleichungen heißt *hyperbolisch*, falls $A(u)$ für jedes u *reell* diagonalisierbar ist. Dies soll durchweg vorausgesetzt werden.

Wenn A konstant ist, ist die Lösung von (111.1) noch recht übersichtlich. Sind $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_m$ die Eigenwerte von A und $v_k \in \mathbb{R}^m$, $k = 1, \ldots, m$, die zugehörigen Eigenvektoren, dann zerfällt das System (111.1) mit dem Lösungsansatz

$$u(x,t) = \sum_{k=1}^{m} \omega_k(x,t)v_k\,, \qquad \omega_k : \mathbb{R}\times[0,T] \to \mathbb{R}\,,$$

in m skalare Transportgleichungen

$$\frac{d}{dt}\,\omega_k + \lambda_k \frac{d}{dx}\,\omega_k = 0\,, \qquad k = 1, \ldots, m\,. \tag{111.2}$$

Genauso wurde bereits in Abschnitt 68 das System (68.2) für die eindimensionale Wellengleichung in zwei skalare Transportgleichungen entkoppelt.

Betrachten wir für diesen linearen Fall ein Riemann-Problem mit

$$u^\circ(x) = \begin{cases} u_-\,, & x < 0\,, \\ u_+\,, & x \geq 0\,, \end{cases} \qquad u_- \neq u_+\,, \tag{111.3}$$

und entwickeln die Vektoren $u_\pm \in \mathbb{R}^m$ in die Eigenvektoren von A,

$$u_\pm = \sum_{k=1}^{m} \omega_k^\pm v_k\,,$$

so werden gemäß (111.2) die einzelnen Eigenkomponenten von $u_\pm$ mit unterschiedlichen Geschwindigkeiten λ_k, $k = 1, \ldots, m$, weitertransportiert. Im allgemeinen entstehen m Unstetigkeitsstellen längs der linearen Charakteristiken $\chi_i(t) = \lambda_i t$, vgl. Abbildung 111.1. In den Sektoren W_i, $i = 0, \ldots, m$, ist die Lösung u des Riemann-Problems dann jeweils konstant:

$$u(x,t) = u_i = u_+ + \sum_{k=1}^{i} (\omega_k^- - \omega_k^+)v_k\,, \qquad (x,t) \in W_i\,.$$

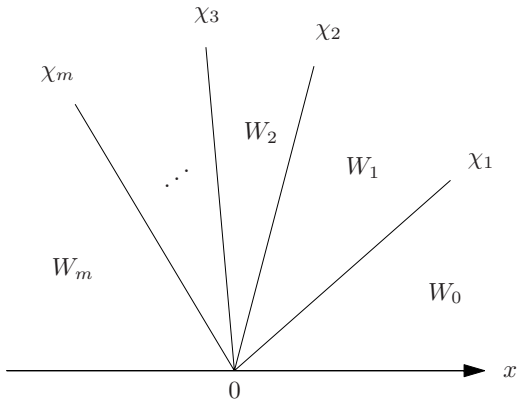


Abb. 111.1: Charakteristiken im linearen Fall

Bei nichtlinearen Erhaltungsgleichungen ist die Situation erwartungsgemäß komplizierter, da die Eigenwerte $\lambda_k^{\pm}$ von $A(u_{\pm})$ in der Regel verschieden sind. In der Regel entstehen jedoch bei der Lösung des Riemann-Problems wie im linearen Fall $m+1$ Sektoren W_i, $i = 0, \ldots, m$, in denen die Lösung $u = u_i$ jeweils konstant ist, vgl. Warnecke [107] oder LeVeque [68]. Berühren sich zwei solche Sektoren wie in Abbildung 111.1, so spricht man von einer Schockwelle; entlang der zugehörigen Geraden $\chi(t) = st$ ist dann eine Rankine-Hugoniot-Bedingung in der Form

$$F(u_i) - F(u_{i+1}) \;=\; s(u_i - u_{i+1})$$

erfüllt. Zwei benachbarte Sektoren müssen sich jedoch nicht berühren, sondern können ihrerseits durch Verdünnungswellen voneinander getrennt sein. Die genaue Anordnung der Sektoren und die Funktionswerte u_i können im allgemeinen jedoch nicht mehr explizit ausgerechnet werden.

Beispiel 111.1. Wir erläutern den nichtlinearen Fall wieder an dem Chromatographie-Beispiel aus Abschnitt 66. Der Übersichtlichkeit halber setzen wir die Geschwindigkeit a der Trägersubstanz und die maximale Teilchenkonzentration ρ jeweils auf Eins und beschränken uns auf $m = 2$ verschiedene Stoffe, die durch die chromatographische Säule fließen. Die Erhaltungsgleichung (66.5) hat in diesem Fall die Form (111.1), wobei die Funktion F nicht explizit sondern nur über ihre Umkehrfunktion

$$F^{-1}(v) \;=\; H(v) \;=\; \begin{bmatrix} v_1 \;+\; \dfrac{\kappa_1 v_1}{1 + \kappa_1 v_1 + \kappa_2 v_2} \\[2ex] v_2 \;+\; \dfrac{\kappa_2 v_2}{1 + \kappa_1 v_1 + \kappa_2 v_2} \end{bmatrix}, \qquad v = \begin{bmatrix} v_1 \\ v_2 \end{bmatrix},$$

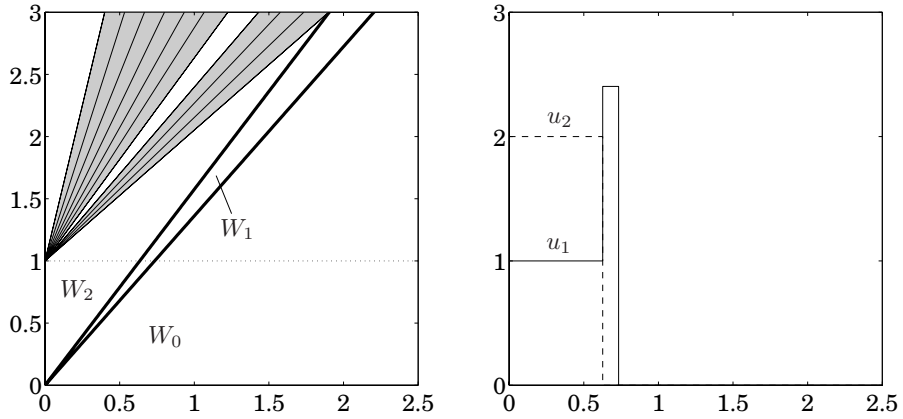


Abb. 111.2: Gemeinsame Höhenlinien der Lösung in der (x,t)-Ebene (links) und Lösung u zum Zeitpunkt $t = 1$ (rechts)

gegeben ist. Mit $v = F(u)$ erhalten wir hieraus

$$A(u) \;=\; H'(v)^{-1} \;=\; \begin{bmatrix} 1 \,+\, \dfrac{\kappa_1 + \kappa_1\kappa_2 v_2}{(1+\kappa_1 v_1 + \kappa_2 v_2)^2} & -\dfrac{\kappa_1\kappa_2 v_1}{(1+\kappa_1 v_1 + \kappa_2 v_2)^2} \\ -\dfrac{\kappa_1\kappa_2 v_2}{(1+\kappa_1 v_1 + \kappa_2 v_2)^2} & 1 \,+\, \dfrac{\kappa_2 + \kappa_1\kappa_2 v_1}{(1+\kappa_1 v_1 + \kappa_2 v_2)^2} \end{bmatrix}^{-1}.$$

Für jedes $v \in \mathbb{R}^2$ sind beide Eigenwerte von $H'(v)$ reell und positiv, also ist das Erhaltungssystem hyperbolisch (vgl. Aufgabe 13).

In dem Chromatographie-Beispiel 66.2 treten sowohl zu Beginn als auch am Ende des Einspritzvorgangs ($t = 0$ bzw. $t = 1$) Riemann-Probleme auf, deren Lösung sehr ausführlich in dem Buch von Rhee, Aris und Amundson [90, Chapter 2] hergeleitet wird; Abbildung 111.2 zeigt links den Verlauf der entstehenden Schock- und Verdünnungswellen.

Zum Zeitpunkt $t = 0$ sind die Daten $u_\pm$ des Riemann-Problems im Punkt $x = 0$ durch die Randbedingung $u_- = [1,2]^T$ (die eingespritzten Konzentrationenen) bzw. $u_+ = [0,0]^T$ (aufgrund der leeren Säule) gegeben. Es entstehen zwei Schockwellen, die sich (wegen der positiven Eigenwerte von $A(u)$) nach rechts ausbreiten. In den äußeren beiden Sektoren (in Abbildung 111.2 mit W_0 und W_2 bezeichnet) ist $u = u_+$ bzw. u_-, im Sektor W_1 stellt sich ein neuer Zustand $u = [u_1, 0]^T$ mit erhöhter Konzentration u_1 ein, vgl. Abbildung 111.2 rechts.

Nach dem Einspritzvorgang ergibt sich im Punkt $x = 0$ ein neues Riemann-Problem mit $u_- = [0,0]^T$ und $u_+ = [1,2]^T$. Bei dieser Konstellation treten zwei Verdünnungswellen auf (in der Abbildung grau hinterlegt), zwischen denen ein neuer Sektor entsteht, in dem $u = [0, u_2]^T$ konstant ist. Eine Vorstellung

von der Lösung in diesem Bereich gewinnt man aus Abbildung 66.3, in der u für $t = 2.24$ dargestellt ist. Der Lösungsverlauf wird komplizierter, wenn die Schock- und Verdünnungswellen schließlich aufeinandertreffen, vgl. [90]. $\Diamond$

Das *HLL-Verfahren* von Harten, Lax und van Leer [49] ist eines der einfachsten numerischen Verfahren zur Lösung solcher Erhaltungssysteme. Es gehört zur Klasse der *Godunov-Typ-Verfahren*, die wie in Abschnitt 107 konstruiert werden, wobei allerdings geeignete Näherungslösungen die exakte Lösung der einzelnen Riemann-Probleme ersetzen.

Im HLL-Verfahren wird die Lösung des Riemann-Problems (111.1), (111.3) durch eine Sprungfunktion

$$\widetilde{v}(x,t) \;=\; \begin{cases} u_-\,, & x < \mu_- t\,, \\ v_0\,, & \mu_- t \le x < \mu_+ t\,, \\ u_+\,, & x \ge \mu_+ t\,, \end{cases} \tag{111.4}$$

approximiert, die in drei Sektoren jeweils konstant ist, vgl. Abbildung 111.3. Dabei sind die Ausbreitungsgeschwindigkeiten μ_- und μ_+ der Schockwellen sowie der mittlere Zustand $v_0 \in \mathbb{R}^m$ freie Parameter.

Sind $\lambda_k(u_\pm)$, $k = 1, \dots, m$, die jeweils absteigend angeordneten Eigenwerte der beiden Matrizen $A(u_\pm)$, so werden für die Ausbreitungsgeschwindigkeiten $\mu_\pm$ die Formeln

$$\begin{aligned} \mu_+ \;&=\; \max\{\,\lambda_1(u_+)\,,\,((\lambda_1(u_+) + \lambda_1(u_-))/2\,\}\,, \\ \mu_- \;&=\; \min\{\,\lambda_m(u_-)\,,\,((\lambda_m(u_+) + \lambda_m(u_-))/2\,\} \end{aligned} \tag{111.5}$$

empfohlen. Der mittlere Zustand v_0 in (111.4) sollte so gewählt werden, daß die Näherungslösung $\widetilde{v}$ das integrale Erhaltungsgesetz (108.1) für jedes Rechteck $[-h/2, h/2] \times [0, \tau]$ erfüllt, das die Schockwellen am oberen Ende verlassen (vgl. das gepunktet eingezeichnete Rechteck in Abbildung 111.3):

$$\begin{aligned} &\int_{-h/2}^{h/2} \widetilde{v}(x,\tau)\,dx - \int_{-h/2}^{h/2} u^\circ(x)\,dx \\ &\quad= \int_0^\tau F(\widetilde{v}(-h/2,t))\,dt - \int_0^\tau F(\widetilde{v}(h/2,t))\,dt \;=\; \tau(F(u_-) - F(u_+))\,. \end{aligned}$$

Die auftretenden Integrale sind einfach auszuwerten und führen auf

$$v_0 \;=\; \frac{\mu_+ u_+ - \mu_- u_-}{\mu_+ - \mu_-} \;-\; \frac{F(u_+) - F(u_-)}{\mu_+ - \mu_-}\,. \tag{111.6}$$

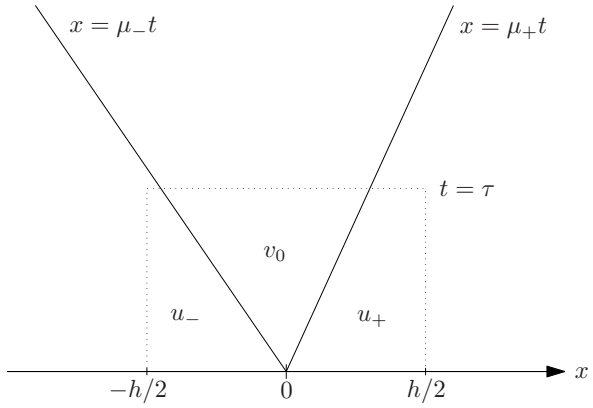


Abb. 111.3: HLL-Verfahren

Die verbleibenden Details des HLL-Verfahrens sind ähnlich wie beim Godunov-Verfahren: Zu den diskreten Zeitpunkten $t_i = i\tau$, $i \in \mathbb{N}_0$, werden stückweise konstante Näherungsfunktionen

$$u_h(t_i) \;=\; \sum_{j\in\mathbb{Z}} u_{ij}\chi_j \;\approx\; u(t_i)$$

über dem verschobenen Gitter $\triangle'$ bestimmt, wobei $u_h(t_{i+1})$ aus $u_h(t_i)$ durch eine Aneinanderreihung von Riemann-Problemen an den Gitterpunkten von $\triangle'$ konstruiert wird. Deren Lösungen lassen sich wie in (111.4)–(111.6) approximieren und können, sofern die entsprechende CFL-Bedingung

$$2 \sup_u \varrho(A(u))\,\tau/h \;\le\; 1 \tag{111.7}$$

erfüllt ist, zu einer Approximation $\widetilde{v}$ über $\mathbb{R}$ zusammengesetzt werden. Die CFL-Bedingung garantiert, daß sich die Mittelsektoren der einzelnen Teillösungen nicht überschneiden, vgl. Abbildung 111.3. Für den Wert $u_{i+1,j}$ von $u_h(t_{i+1})$ wird schließlich das entsprechende Integralmittel von $\widetilde{v}$ eingesetzt.

Unter Berücksichtigung aller möglichen Vorzeichen von $\mu_\pm$ ergibt sich für das HLL-Verfahren eine Rekursion in Erhaltungsform (108.4) mit dem (konsistenten) numerischen Fluß

$$G(u_-,u_+) \;=\; \frac{\mu_+^+F(u_-)-\mu_-^-F(u_+)}{\mu_+^+-\mu_-^-} \;+\; \frac{\mu_+^+\mu_-^-}{\mu_+^+-\mu_-^-}(u_+-u_-)\,, \tag{111.8}$$

vgl. Aufgabe 14. Hierbei ist

$$\mu_+^+ \;=\; \max\{\mu_+,0\} \qquad \text{und} \qquad \mu_-^- \;=\; \min\{\mu_-,0\}\,.$$

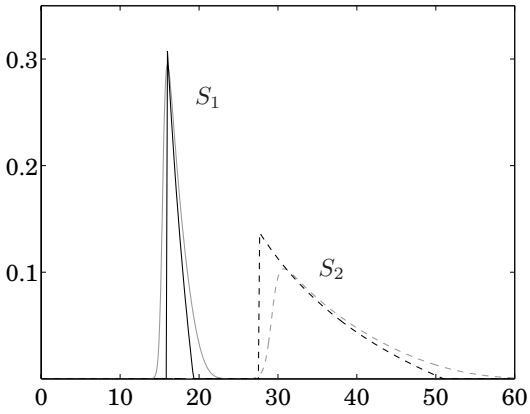


Abb. 111.4: Numerisches Ergebnis des HLL-Verfahrens

Sind die Eigenwerte von $A(u)$ allesamt positiv, so ist offensichtlich $\mu_+^+ = \mu_+$ und $\mu_-^- = 0$ und der numerische Fluß vereinfacht sich zu

$$G(u_-, u_+) \;=\; F(u_-)\,.$$

Das Verfahren stimmt in diesem Fall also mit dem Upwind-Schema überein und beide Steigungen $\mu_\pm$ sind positiv. Eine entsprechende Modifikation der Skizze in Abbildung 111.3 macht deutlich, daß dann selbst für

$$\tau \;\le\; h/\sup_u \varrho(A(u)) \tag{111.9}$$

gewährleistet ist, daß sich die Mittelsektoren benachbarter Teillösungen nicht überschneiden. Die Zeitschrittweite kann also in diesem Fall gegenüber (111.7) verdoppelt werden.

Beispiel. In Beispiel 111.1 sind die Eigenwerte von $A(u)$ jeweils positiv. Das HLL-Verfahren entspricht auch in diesem Fall dem Upwind-Schema und die Zeitschrittweite muß die CFL-Bedingung (111.9) erfüllen. Für die numerischen Resultate in Abbildung 111.4 wurde die chromatographische Säule der Länge $L = 10$ wie in den vorangegangenen Beispielen in 100 Gitterintervalle unterteilt. Die Abbildung zeigt den tatsächlichen Ausfluß der beiden Substanzen aus der Säule (vgl. Abbildung 66.3) mit der Approximation des HLL-Verfahrens (hellgrau eingezeichnet). Bei der numerischen Lösung wird der sprunghafte Ausfluß der zweiten Substanz S_2 nur schlecht approximiert. Dies liegt an der Zeitschrittweite, die im Verlauf der Rechnung durch die CFL-Bedingung (111.9) zu stark eingeschränkt wird. ◊

Aufgaben

1. Zeigen Sie, daß (mit der Notation aus Abschnitt 104) das Verfahren

$$u_{i+1,j} \;=\; 2(1-\rho^2)u_{ij} + \rho^2(u_{i,j+1} + u_{i,j-1}) - u_{i-1,j}\,, \qquad \rho = a\tau/h\,,$$

ein Differenzenschema zweiter Ordnung zur Lösung der eindimensionalen Wellengleichung $u_{tt} = a^2 u_{xx}$ ist. Machen Sie sich klar, wie analog zu Bemerkung 83.6 Anfangsbedingungen für u und u_t an der Stelle $t = 0$ eingebaut werden können, vgl. (68.5). Beweisen Sie, daß das Verfahren höchstens unter der CFL-Bedingung $a\tau/h < 1$ stabil ist.

2. Bestimmen Sie für die singulär gestörte Burgers-Gleichung

$$u_t + uu_x \;=\; \varepsilon u_{xx}$$

eine Lösung der Form

$$u(x,t;\varepsilon) \;=\; U(x-at)\,, \quad a \in \mathbb{R}\,,$$

die für $u_- > u_+$ im Unendlichen die Randbedingungen

$$\lim_{s\to-\infty} U(s) = u_-\,, \qquad \lim_{s\to\infty} U(s) = u_+\,, \qquad \lim_{s\to\pm\infty} U'(s) = 0\,,$$

sowie $U(0) = a$ erfüllt. Untersuchen Sie das Verhalten dieser Lösung für kleine ε.

3. Geben Sie für die Burgers-Gleichung zwei unterschiedliche Anfangsvorgaben u° über $\mathbb{R}$ zur Zeit $t = 0$ an, die für $t \geq 1$ auf die gleiche Lösung

$$u(x,t) \;=\; \begin{cases} 1\,, & x < t/2\,, \\ 0\,, & x \geq t/2\,, \end{cases}$$

führen.

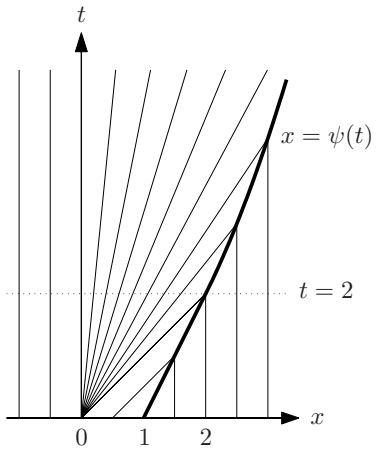
4. Betrachten Sie die Burgers-Gleichung mit der Anfangsvorgabe

$$u^\circ(x) \;=\; \begin{cases} 0\,, & x < 0\,, \\ 1\,, & 0 \leq x < 1\,, \\ 0\,, & x \geq 1\,. \end{cases}$$

In diesem Fall breiten sich in $x = 0$ eine Verdünnungswelle und in $x = 1$ eine Schockwelle aus, die sich für $t = 2$ im Punkt $(2, 2)$ treffen.
Zeigen Sie, daß die Lösung u für $t \geq 2$ durch

$$u(x,t) \;=\; \begin{cases} 0\,, & x < 0\,, \\ x/t\,, & 0 \leq x < \psi(t)\,, \\ 0\,, & x \geq \psi(t)\,, \end{cases}$$

mit einer gewissen Unstetigkeitskurve $(\psi(t), t)$, $t \geq 2$, gegeben ist. Bestimmen Sie ψ. (Die Unstetigkeitskurve ist in der Skizze fett eingezeichnet.)



5. Es sei $u \in \mathcal{U}$ eine schwache Lösung der hyperbolischen Anfangswertaufgabe (106.1), die der Entropiebedingung von Oleinik genügt. Weiter seien $\Gamma = \{(\psi(t), t) : t_1 < t < t_2\}$ eine Unstetigkeitskurve und χ eine Charakteristik von u, die zur Zeit $t_0 \in (t_1, t_2)$ zusammentreffen. Zeigen Sie, daß χ entweder in t_0 endet oder für $t > t_0$ mit Γ übereinstimmt.

6. Gegeben sei die Differentialgleichung

$$u_t + \frac{d}{dx} F(u) = 0 \quad \text{in } \mathbb{R} \times \mathbb{R}^+ \qquad \text{mit } F(u) = u(u-2)(u-3)(u-4)\,.$$

Bestimmen Sie die Lösung des zugehörigen Riemann-Problems mit $u_- = 4$ und $u_+ = 0$, die die Entropiebedingung von Oleinik erfüllt.

7. Die Funktion $u \in \mathcal{U}$ mit Werten in einem beschränkten Intervall $\mathcal{I} = [c, d]$ sei eine schwache Lösung des Anfangswertproblems (106.1), die die Entropiebedingung von Oleinik erfüllt. Ferner sei $\eta \in C^2(\mathcal{I})$ konvex und

$$q(u) = \int_c^u F'(w)\eta'(w)\,dw\,.$$

Beweisen Sie:
(a) Besitzt u entlang der Unstetigkeitskurve $\Gamma = \{\,(\psi(t), t) : 0 \le t_0 \le t \le t_1 \le T\,\}$ links- bzw. rechtsseitige Grenzwerte u_- und u_+, so gilt

$$q(u_+) - q(u_-) \le \psi'\big(\eta(u_+) - \eta(u_-)\big)\,.$$

(b) Ist $\varphi \in C^\infty(\mathbb{R} \times \mathbb{R})$ nichtnegativ mit kompaktem Träger, so gilt

$$\int_{\mathbb{R}\times[0,T]} \Big(\eta(u)\varphi_t + q(u)\varphi_x\Big)\,d(x,t) + \int_{\mathbb{R}} \eta(u^\circ)\varphi(0)\,dx \ge 0\,.$$

Hinweis zu (a): Verwenden Sie die Darstellung

$$\eta(u) = \eta_0 + \eta_1 u + \int_c^d \eta''(w)(u-w)^+\,dw\,, \qquad (u-w)^+ = \max\{u-w, 0\}\,,$$

mit geeigneten $\eta_0, \eta_1 \in \mathbb{R}$.

8. (a) Geben Sie die Lösung der Burgers-Gleichung mit vorgegebenen Anfangswerten

$$u^\circ(x) = \begin{cases} -1/(2\varepsilon) & \text{für } -\varepsilon < x < 0\,, \\ 1/(2\varepsilon) & \text{für } 0 < x < \varepsilon\,, \\ 0 & \text{sonst}\,, \end{cases}$$

an und zeigen Sie, daß diese für $\varepsilon \to 0$ punktweise gegen

$$U(x,t) = \begin{cases} x/t & \text{für } x^2 < t\,, \\ 0 & \text{für } x^2 > t\,, \end{cases}$$

konvergiert.

(b) Zeigen Sie, daß U in $\mathbb{R}\times(0,\infty)$ zwar eine schwache Lösung der Burgers-Gleichung ist und die Entropiebedingung von Oleinik erfüllt, daß es aber keine Anfangsvorgabe $u^\circ \in \mathcal{L}^1(\mathbb{R})$ geben kann, so daß U die schwache Lösung des zugehörigen Anfangswertproblems ist.

9. Gegenstand dieser Aufgabe ist das Verkehrsbeispiel aus Aufgabe XII.3 mit der Geschwindigkeitsfunktion

$$v(u) \,=\, v_\infty(1-u/u_\infty)\,, \qquad 0<u<u_\infty\,.$$

(a) An der Stelle $x=0$ sei eine Ampel, die zur Zeit $t=0$ von Rot auf Grün umspringt. Links von der Ampel ($x<0$) habe sich der Verkehr aufgestaut ($u=u_\infty$), für $x>0$ sei die Straße leer ($u=0$). Bestimmen Sie die Verkehrsdichte $u(x,t)$ für $t>0$ und interpretieren Sie das Ergebnis im Rahmen des Modells.
(b) Wie lautet die Lösung für das Riemann-Problem mit $u_-\in(0,u_\infty)$ und $u_+=u_\infty$, und wie läßt sich diese Lösung im Verkehrskontext interpretieren?
(c) Berechnen Sie die Lösung $u(x,t)$ für das Anfangswertproblem

$$u(x,0) \,=\, \begin{cases} 0\,, & x<a\,,\\ u_\infty(x-a)/(b-a)\,, & a\le x\le b<0\,,\\ u_\infty\,, & b\le x<0\,,\\ 0\,, & x\ge 0\,, \end{cases}$$

mit dem Godunov-Verfahren. Gehen Sie davon aus, daß alle Autos vier Meter lang sind und die Höchstgeschwindigkeit bei 100 km/h liegt. Rechnen Sie auf dem Ortsintervall $[-5,4]$ mit den Parametern $a=-5$ und $b=-0.1$ und nehmen Sie an, daß von links keine Autos mehr nachkommen. Diskretisieren Sie das Ortsintervall durch 900 Gitterintervalle.

10. Es sei F eine monotone konvexe Flußfunktion. Zeigen Sie, daß in diesem Fall der Godunov-Fluß durch

$$G(u_-,u_+) \,=\, \begin{cases} F(u_-)\,, & \text{falls } F \text{ monoton wachsend}\,,\\ F(u_+)\,, & \text{falls } F \text{ monoton fallend}\,, \end{cases}$$

gegeben ist und daß das Godunov-Verfahren in diesem Fall mit dem Upwind-Schema übereinstimmt.

11. Ein Differenzenverfahren in Erhaltungsform wird *monoton* genannt, falls

$$H(\mu_-,\mu_0,\mu_+) \,=\, \mu_0-\gamma(G(\mu_0,\mu_+)-G(\mu_-,\mu_0))$$

in allen drei Argumenten monoton wachsend ist. Zeigen Sie, daß für ein monotones Verfahren

$$\min\{u_{i,j-1},u_{i,j},u_{i,j+1}\} \,\le\, u_{i+1,j} \,\le\, \max\{u_{i,j-1},u_{i,j},u_{i,j+1}\}\,, \qquad j\in\mathbb{Z}\,,\ i\in\mathbb{N}_0\,,$$

gilt und daß jedes monotone Verfahren ein TVD-Verfahren ist.
Hinweis: Beweisen Sie für die zweite Aussage zunächst, daß

$$\sum_{j\in\mathbb{Z}}(u_{i+1,j+1}-u_{i+1,j})^+ \,\le\, \sum_{j\in\mathbb{Z}}(u_{i,j+1}-u_{ij})^+\,,$$

wobei $c^+=\max\{c,0\}$.

12. Welches der zu den folgenden numerischen Flußfunktionen gehörende Differenzenverfahren in Erhaltungsform ist unter der CFL-Bedingung $\gamma \max_u |F'(u)| \le 1$ ein TVD-Verfahren?

(a) *Lax-Friedrichs-Verfahren*: $G_{\mathrm{LF}}(v,w) = \dfrac{F(v)+F(w)}{2} + \dfrac{(v-w)}{2\gamma}$

(b) *Lax-Wendroff-Verfahren*: $G_{\mathrm{LW}}(v,w) = \dfrac{F(v)+F(w)}{2} - \dfrac{\gamma}{2}F'\big(\dfrac{v+w}{2}\big)\big(F(w)-F(v)\big)$

(c) *Enquist-Osher-Verfahren*: $G_{\mathrm{EO}}(v,w) = \dfrac{F(v)+F(w)}{2} - \dfrac{1}{2}\displaystyle\int_v^w |F'(s)|\,ds$

13. Zeigen Sie, daß das System (66.5) aus Erhaltungsgleichungen für jedes $m \ge 2$ hyperbolisch ist und daß alle Eigenwerte von $A(u)$ aus der Darstellung (111.1) positiv und kleiner gleich Eins sind.
Hinweis: Verwenden Sie Proposition 29.1.

14. Sei $\widetilde{v}(x,t)$ die beim HLL-Verfahren verwendete approximative Lösung (111.4) des Riemann-Problems. Zeigen Sie, daß

(a) $\dfrac{1}{h}\displaystyle\int_{-h/2}^{0} \widetilde{v}(x,\tau)\,dx = \dfrac{u_-}{2} - \gamma\Big(\dfrac{\mu_+^+\mu_-^-}{\mu_+^+-\mu_-^-}(u_+-u_-) + \dfrac{\mu_+^+F(u_-)-\mu_-^-F(u_+)}{\mu_+^+-\mu_-^-} - F(u_-)\Big)$,

(b) $\dfrac{1}{h}\displaystyle\int_{0}^{h/2} \widetilde{v}(x,\tau)\,dx = \dfrac{u_+}{2} + \gamma\Big(\dfrac{\mu_+^+\mu_-^-}{\mu_+^+-\mu_-^-}(u_+-u_-) + \dfrac{\mu_+^+F(u_-)-\mu_-^-F(u_+)}{\mu_+^+-\mu_-^-} - F(u_+)\Big)$.

Hierbei sei wieder $\mu_+^+ = \max\{\mu_+, 0\}$ und $\mu_-^- = \min\{\mu_-, 0\}$ mit den Ausbreitungsgeschwindigkeiten μ_+ und μ_- aus (111.5). Es sei vorausgesetzt, daß die CFL-Bedingung (111.7) erfüllt ist.
Benutzen Sie dies, um zu zeigen, daß der numerische Fluß des HLL-Verfahrens die in (111.8) angegebene Form hat.

Literaturverzeichnis

[1] Abramowitz, M.; Stegun, I.: Handbook of Mathematical Functions. New York: Dover 1972

[2] Ahlfors, L. V.: Complex Analysis. New York: McGraw-Hill 1979

[3] Ainsworth, M.; Oden, J. T.: A Posteriori Error Estimation in Finite Element Analysis. New York: Wiley 2000

[4] Ansorge, R.; Sonar, T.: Informationsverlust, abstrakte Entropie und die mathematische Beschreibung des zweiten Hauptsatzes der Thermodynamik. ZAMM **77** (1997) 803–821

[5] Aris, R.: Mathematical Modelling. San Diego: Academic Press 1999

[6] Ascher, U. M.; Petzold, L. R.: Computer Methods for Ordinary and Differential Equations and Differential-Algebraic Equations. Philadelphia: SIAM 1998

[7] Ashcroft, N. W.; Mermin, N. D.: Festkörperphysik. München: Oldenbourg 2001

[8] Björck, Å.: Numerical Methods for Least Squares Problems. Philadelphia: SIAM 1996

[9] Björck, Å.; Dahlquist, G.: Numerical Methods and Scientific Computations. Philadelphia: SIAM to appear

[10] Braess, D.: Finite Elemente. Berlin: Springer 1997

[11] Braun, M.: Differentialgleichungen und ihre Anwendungen. Berlin: Springer 1994

[12] Bulirsch, R.: Bemerkungen zur Romberg-Integration. Numer. Math. **6** (1964) 6–16

[13] Cannon, J. R.: The One-Dimensional Heat Equation. Menlo Park, CA: Addison-Wesley 1984

[14] Chan, R. H.; Ng, M. K.: Conjugate gradient methods for Toeplitz systems. SIAM Rev. **38** (1996) 427–482

[15] Chihara, T. S.: An Introduction to Orthogonal Polynomials. New York: Gordon and Breach 1978

[16] Chorin, A. J.; Marsden, J. E.: A Mathematical Introduction to Fluid Mechanics. New York: Springer 1998

[17] Chui, C. K.: Wavelets: A Mathematical Tool for Signal Analysis. Philadelphia: SIAM 1997
[18] Ciarlet, P. G.: Mathematical Elasticity, Bd. 1: Three Dimensional Elasticity. Amsterdam: North-Holland 1993
[19] Clément, P.: Approximation by finite element functions using local regularization. RAIRO Anal. Numér. **9 (R-2)** (1975) 77–84
[20] Cohen, A.: Wavelet methods in numerical analysis. In: P. G. Ciarlet; J. L. Lions (Hrsg.), *Handbook of Numerical Analysis*, Bd. VII. Amsterdam: Elsevier, 2000 S. 417–711
[21] Davis, P. J.; Rabinowitz, P.: Methods of Numerical Integration. New York: Academic Press 1975
[22] Demmel, J. W.: Applied Numerical Linear Algebra. Philadelphia: SIAM 1997
[23] Deuflhard, P.; Hohmann, A.: Numerische Mathematik I. Eine algorithmisch orientierte Einführung. Berlin, New York: de Gruyter 1991
[24] Eich-Soellner, E.; Führer, C.: Numerical Methods in Multibody Dynamics. Stuttgart: Teubner 1998
[25] Eisenstat, S. C.: Efficient implementation of a class of preconditioned conjugate gradient methods. SIAM J. Sci. Stat. Comput. **2** (1981) 1–4
[26] Engl, H. W.; Hanke, M.; Neubauer, A.: Regularization of Inverse Problems. Dordrecht: Kluwer 1996
[27] Feynman, R. P.; Leighton, R. B.; Sands, M.: Feynman Vorlesungen über Physik, Bd. I. München: Oldenbourg 2001
[28] Forsythe, G. E.; Henrici, P.: The cyclic Jacobi method for computing the principal values of a complex matrix. Trans. Amer. Math. Soc. **94** (1960) 1–23
[29] Fowkes, N. D.; Mahony, J. J.: Einführung in die mathematische Modellierung. Heidelberg: Spektrum Akadademischer Verlag 1996
[30] Gander, W.; Gautschi, W.: Adaptive quadrature—revisited. BIT **40** (2000) 84–101
[31] Gautschi, W.: Numerical Analysis. Basel: Birkhäuser 1997
[32] Geiger, C.; Kanzow, C.: Numerische Verfahren zur Lösung unrestringierter Optimierungsaufgaben. Berlin: Springer 1999
[33] Goldstein, H.: Klassische Mechanik. Wiesbaden: Aula-Verlag 1991
[34] Golub, G. H.; Van Loan, C.: Matrix Computations. 3. Aufl. Baltimore: John Hopkins University Press 1996
[35] Golub, G. H.; Welsch, J. H.: Calculation of Gauss quadrature rules. Math. Comp. **23** (1969) 221–230
[36] Greenbaum, A.: Iterative Methods for Solving Linear Systems. Philadelphia: SIAM 1997
[37] Greenspan, D.: Particle Modeling. Boston: Birkhäuser 1997

[38] Grindrod, P.: The Theory and Applications of Reaction-Diffusion Equations—Patterns and Waves. Oxford: Clarendon Press 1996
[39] Grisvard, P.: Elliptic Problems in Nonsmooth Domains. Boston: Pitman 1985
[40] Großmann, C.; Roos, H.-G.: Numerik partieller Differentialgleichungen. Stuttgart: Teubner 1994
[41] Haar, A.: Zur Theorie der orthogonalen Funktionensysteme. Math. Ann. **69** (1910) 331–371
[42] Haberman, R.: Mathematical Models. Philadelphia: SIAM 1998
[43] Hackbusch, W.: Theorie und Numerik elliptischer Differentialgleichungen. Stuttgart: Teubner 1996
[44] Hairer, E.; Nørsett, S. P.; Wanner, G.: Solving Ordinary Differential Equations I. Berlin: Springer 2000
[45] Hairer, E.; Wanner, G.: Solving Ordinary Differential Equations II. Berlin: Springer 2002
[46] Hämmerlin, G.; Hoffmann, K.-H.: Numerische Mathematik. Berlin: Springer 1994
[47] Hanke, M.; Scherzer, O.: Error analysis of an equation error method for the identification of the diffusion coefficient in a quasilinear parabolic differential equation. SIAM J. Appl. Math. **59** (1999) 1012–1027
[48] Hanke, M.; Scherzer, O.: Inverse problems light: numerical differentiation. Amer. Math. Monthly **108** (2001) 512–521
[49] Harten, A.; Lax, P. D.; van Leer, B.: On upstream differencing and Godunov-type schemes for hyperbolic conservation laws. SIAM Rev. **25** (1983) 35–62
[50] Henrici, P.: Applied and Computational Complex Analysis. New York: Wiley 1993
[51] Heuser, H.: Funktionalanalysis. Stuttgart: Teubner 1992
[52] Heuser, H.: Gewöhnliche Differentialgleichungen. Stuttgart: Teubner 1995
[53] Heuser, H.: Lehrbuch der Analysis. Stuttgart: Teubner 2000
[54] Higham, D. J.; Higham, N. J.: MATLAB Guide. Philadelphia: SIAM 2000
[55] Higham, N. J.: Accuracy and Stability of Numerical Algorithms. Philadelphia: SIAM 1996
[56] Iserles, A.: A First Course in the Numerical Analysis of Differential Equations. Cambridge: Cambridge University Press 1996
[57] Jensen, A.; la Cour-Harbo, A.: Ripples in Mathematics: The Discrete Wavelet Transform. Berlin: Springer 2001
[58] Jordan, C.: Calculus of Finite Differences. New York: Chelsea Publ. Co. 1965
[59] Kirschner, D.: Using mathematics to understand HIV immune dynamics. Notices Amer. Math. Soc. **43** (1996) 191–202

[60] Klamkin, M. S. (Hrsg.): Mathematical Modelling. Philadelphia: SIAM 1987
[61] Knabner, P.; Angermann, L.: Numerik partieller Differentialgleichungen. Berlin: Springer 2000
[62] Köckler, F.: Numerische Algorithmen in Softwaresystemen. Stuttgart: Teubner 1990
[63] Kreß, R.: Numerical Analysis. New York: Springer 1998
[64] Kreß, R.: Linear Integral Equations. 2. Aufl. New York: Springer 1999
[65] Krommer, A. R.; Überhuber, C. W.: Computational Integration. Philadelphia: SIAM 1998
[66] Kröner, D.: Numerical Schemes for Conservation Laws. New York: Wiley 1997
[67] Lang, J.: Adaptive Multilevel Solution of Nonlinear Parabolic PDE Systems. Berlin: Springer 2001
[68] LeVeque, R. J.: Numerical Methods for Conservation Laws. Basel: Birkhäuser 1990
[69] Lin, C.-C.; Segel, L. A.: Mathematics Applied to Deterministic Problems in the Natural Sciences. Philadelphia: SIAM 1994
[70] Louis, A. K.; Maaß, P.; Rieder, A.: Wavelets: Theorie and Anwendungen. 2. Aufl. Stuttgart: Teubner 1998
[71] Löwe, A.: Chemische Reaktionskinetik. Weinheim: Wiley-VCH 2001
[72] Lubich, C.; Engstler, C.; Nowak, U.; Pöhle, U.: Numerical integration of constrained mechanical systems using MEXX. Mech. Structures Mach. **23** (1995) 473–495
[73] McKenna, P. J.: Large torsional oscillations in suspension bridges revisited: fixing an old approximation. Amer. Math. Monthly **106** (1999) 1–18
[74] Nash, S.; Sofer, A.: Linear and Nonlinear Programming. New York: McGraw-Hill 1996
[75] Natterer, F.: The Mathematics of Computerized Tomography. Stuttgart: Wiley 1986
[76] Niethammer, W.: Relaxation bei komplexen Matrizen. Math. Z. **86** (1964) 34–40
[77] Nowak, M. A.; May, R. M.: Virus Dynamics. Oxford: Oxford University Press 2000
[78] Ortega, J. M.; Rheinboldt, W. C.: Iterative Solution of Nonlinear Equations in Several Variables. New York: Academic Press 1970
[79] Overton, M. L.: Numerical Computing With IEEE Floating Point Arithmetic. Philadelphia: SIAM 2001
[80] Paige, C. C.; Saunders, M. A.: Solution of sparse indefinite systems of linear equations. SIAM J. Numer. Anal. **12** (1975) 617–629

[81] Parlett, B. N.: The Symmetric Eigenvalue Problem. Englewood Cliffs, NJ: Prentice-Hall 1980
[82] Parlett, B. N.; Dhillon, I. S.: Fernando's solution to Wilkinson's problem: An application of double factorization. Linear Algebra Appl. **267** (1997) 247–279
[83] Pennebaker, W. B.; Mitchell, J. L.: JPEG: Still Image Data Compression Standard. New York: Van Nostrand Reinhold 1993
[84] Perelson, A. S.; Kirschner, D. E.; De Boer, R.: Dynamics of HIV infection of $CD4^+$ T cells. Math. Biosci. **114** (1993) 81–125
[85] Protter, M. H.; Weinberger, H. F.: Maximum Principles in Differential Equations. Englewood Cliffs, NJ: Prentice-Hall 1967
[86] Quarteroni, A.; Sacco, R.; Saleri, F.: Numerische Mathematik. Heidelberg: Springer 2002
[87] Quarteroni, A.; Valli, A.: Numerical Approximation of Partial Differential Equations. Berlin: Springer 1997
[88] Rabier, P. J.; Rheinboldt, W. C.: Nonholonomic Motion of Rigid Mechanical Systems from a DAE Viewpoint. Philadelphia: SIAM 2000
[89] Rhee, H.-K.; Aris, R.; Amundson, N. R.: First-Order Partial Differential Equations I. Englewood Cliffs, NJ: Prentice-Hall 1986
[90] Rhee, H.-K.; Aris, R.; Amundson, N. R.: First-Order Partial Differential Equations II. Englewood Cliffs, NJ: Prentice-Hall 1989
[91] Roos, H.-G.; Stynes, M.; Tobiska, L.: Numerical Methods for Singularly Perturbed Differential Equations. Berlin: Springer 1996
[92] Rubinstein, I.; Rubinstein, L.: Partial Differential Equations in Classical Mathematical Phyics. Cambridge: Cambridge University Press 1998
[93] Rudin, W.: Reelle und Komplexe Analysis. München: Oldenbourg 1999
[94] Rump, S. M.: Wie zuverlässig sind die Ergebnisse unserer Rechenanlagen? Jahrb. Überbl. Math. (1983) 163–168
[95] Schumaker, L.: Spline Functions: Basic Theory. New York: Wiley 1981
[96] Seydel, R.: Einführung in die numerische Berechnung von Finanz-Derivaten · computational finance. Berlin: Springer 2000
[97] Shampine, L. F.; Reichelt, M. W.: The MATLAB ODE Suite. SIAM J. Sci. Comput. **18** (1997) 1–22
[98] Stewart, G. W.: Introduction to Matrix Computations. New York: Academic Press 1973
[99] Swartz, B. K.; Varga, R. S.: Error bounds for spline and L-spline interpolation. J. Approx. Th. **6** (1972) 6–49
[100] Szegö, G.: Orthogonal Polynomials, Bd. 23 von *Amer. Math. Soc. Colloq. Publ.* Providence, RI: Amer. Math. Soc. 1975
[101] Temam, R.; Miranville, A.: Mathematical Modeling in Continuum Mechanics. Cambridge: Cambridge University Press 2001

[102] Thomée, V.: Galerkin Finite Element Methods for Parabolic Problems. Berlin: Springer 1997
[103] Überhuber, C.; Katzenbeisser, S.: MATLAB 6. Wien: Springer 2000
[104] Van Loan, C. F.: Computational Frameworks for the Fast Fourier Transform. Philadelphia: SIAM 1992
[105] Vlach, J.; Singhal, K.: Computer Methods for Circuit Analysis and Design. New York: Van Nostrand Reinhold 1983
[106] Walter, W.: Gewöhnliche Differentialgleichungen. Berlin: Springer 2000
[107] Warnecke, G.: Analytische Methoden in der Theorie der Erhaltungsgleichungen. Stuttgart, Leipzig: Teubner 1999
[108] Wilkinson, J. H.: The Algebraic Eigenvalue Problem. Oxford: Clarendon Press 1965
[109] Yeargers, E. K.; Shonkwiler, R. W.; Herod, J. V.: An Introduction to the Mathematics of Biology. Boston: Birkhäuser 1996
[110] Zygmund, A.: Trigonometric Series. New York: Cambridge University Press 1959

Sachverzeichnis