Claus-Dieter Munz · Thomas Westermann

Numerische Behandlung gewöhnlicher und partieller Differenzialgleichungen

Claus-Dieter Munz · Thomas Westermann

Numerische Behandlung gewöhnlicher und partieller Differenzialgleichungen

Ein interaktives Lehrbuch für Ingenieure

Mit 137 Abbildungen

Springer

Professor Dr. Claus-Dieter Munz
Universität Stuttgart
Institut für Aerodynamik und Gasdynamik
Pfaffenwaldring 21
70550 Stuttgart
munz@iag.uni-stuttgart.de
http://www.iag.uni-stuttgart.de/people/claus-dieter.munz/german/Munz.html

Professor Dr. Thomas Westermann
Hochschule Karlsruhe
Technik und Wirtschaft
Postfach 2440
76012 Karlsruhe
thomas.westermann@hs-karlsruhe.de
http://www.home.hs-karlsruhe.de/~wethoo02

Homepage zum Buch
http://www.home.hs-karlsruhe.de/~wethoo02/buecher/revokos/start.htm
oder *http://www.springer.com/de/3-540-29867-3*

Bibliografische Information der Deutschen Bibliothek

Die Deutsche Bibliothek verzeichnet diese Publikation in der Deutschen Nationalbibliografie; detaillierte bibliografische Daten sind im Internet über http://dnb.ddb.de abrufbar.

ISBN-10 3-540-29867-3 Springer Berlin Heidelberg New York
ISBN-13 978-3-540-29867-0 Springer Berlin Heidelberg New York

Springer ist ein Unternehmen von Springer Science+Business Media
springer.de

Printed in Germany

Satz: Reproduktionsfertige Vorlage der Autoren
Herstellung: LE-TEX Jelonek, Schmidt & Vöckler GbR, Leipzig
Umschlaggestaltung: KünkelLopka, Heidelberg
Gedruckt auf säurefreiem Papier 7/3100/YL - 5 4 3 2 1 0

Vorwort

Die Berechnung der meisten natur- und ingenieurwissenschaftlichen Anwendungen führt auf eine oder mehrere Differenzialgleichungen. Deshalb stellen wir in diesem Buch deren numerische Lösung in den Mittelpunkt einer Numerik für Ingenieure. Die notwendigen Kenntnisse aus anderen Bereichen der numerischen Mathematik, wie z.B. der Interpolation oder der numerischen Lösung von linearen Gleichungssystemen, werden an der Stelle erläutert, wo sie benötigt werden. Diese „Hilfsmittel“ beschreiben wir im Anhang, so kurz wie möglich und vervollständigt mit Angabe von weiterführender Literatur.

Dieses Buch wurde zunächst als ein interaktives Skriptum für die Vorlesung „Numerische Methoden für Ingenieure“ entwickelt. Das Ziel war, die beiden Aspekte Theorie und Erfahrung gemeinsam zu vermitteln. Es wurden Rechenprogramme entwickelt, mit denen sich numerische Rechnungen in der Vorlesung ausführen lassen und die Ergebnisse mit den Studierenden zusammen diskutiert werden können. Realisiert als MAPLE-Worksheets ließen sich diese Programme auch aus der pdf-Version des Skripts direkt ausführen und modifizieren. Mit diesem elektronischen Skript lässt sich ein Konzept realisieren, das mit traditionellen Mitteln nicht möglich ist. Das Lernen geschieht in drei Stufen:

Hören und passives Erleben von numerischer Mathematik: Dieses geschieht in der Vorlesung. Gerade bei einem großen Publikum lässt sich hier Interaktion nur eingeschränkt verwirklichen. Das Erleben aber ist bereits möglich, indem die Beispiele aus dem Skriptum im Vortrag verwendet werden. Numerische Methoden können direkt in der Vorlesung ausgeführt werden - mit kleinen Lernspielen für das Publikum.

Lesen und eigenes Experimentieren: Bei der Nachbereitung haben die Teilnehmer ein multimediales Produkt auf dem eigenen PC, das sie einlädt, das Gelesene auch gleich in der Praxis zu erproben. Die vorgearbeiteten Beispiele sind direkt greifbar und die Schnittstellen für eigene Modifikationen einfach. Das Erlernen der Programmiersprache in MAPLE geschieht automatisch.

Schreiben eigener Programme: Solchermaßen vorbereitet ist es für die Studierenden oder den Leser ein relativ kleiner Schritt, für konkrete Aufgabenstellungen eigene Programme zu entwickeln.

Das Buch ist in der Form eines pdf-Files als ein „elektronisches Skript“ auf der beiliegenden CD-ROM. Für fast alle Beispiele gibt es Rechenprogramme, die direkt aus dem Text durch Anklicken aufgerufen und ausgeführt werden können. Auf dem Rechner muss dazu allerdings das Computer-Algebra-System MAPLE zur Verfügung stehen. Damit werden dem „Leser“ neue Möglichkeiten eröffnet. Zur weiteren Bearbeitung können die gesamten Programmier- und Visualisierungsmöglichkeiten von MAPLE eingesetzt werden. Neben der Nachrechnung des Beispiels können in dem aufgerufenen MAPLE-Worksheet die Diskretisierungsparameter geändert werden. Man kann mit den numerischen Verfahren experimentieren und sie auf neue kompliziertere Probleme angewenden. Steht kein MAPLE zur Verfügung, fällt nur die Möglichkeit des interaktiven Arbeitens weg. Zur Veranschaulichung sind die wesentlichen Ergebnisse der Rechnungen als Abbildungen im Buch. Zusätzlich können die Worksheets in einer html-Version betrachtet werden.

Die Möglichkeit, Rechenprogramme in ein Skriptum einzuarbeiten und als Buch zu erweitern, konnte nur durch Unterstützung des Ministeriums für Wissenschaft, Forschung und Kunst Baden-Württemberg im Rahmen eines Projektes „Innovative Lehre“ durchgeführt werden. Wir möchten uns dafür an dieser Stelle herzlichst bedanken. Im Bereich der Lehre gibt es leider sehr wenige Geldgeber für solche Projekte. An mehreren Stellen konnten wir auf das „alte“ Skriptum von Prof. Dr. Karl Förster, überarbeitet und ergänzt von Dr. Alfred Geiger [20] aufbauen. Wir haben hier sehr profitiert und möchten uns dafür sehr herzlich bedanken. Unser herzlicher Dank gebührt dann vor allem unseren Mitarbeitern Mark Ferch, Friedemann Kemm und Mirko Pauli, die uns bei der Erstellung des Buches tatkräftig unterstützt haben. Bedanken möchten wir uns auch bei vielen Studierenden, allen voran Olaf Schönrock, Oliver Hohn und Dominik Saile, bei PD Dr. Horst Parisch für seine zahlreichen Kommentare und bei Dr. Werner Grimm und Dr. Sven Olaf Neumann für die Unterstützung bei den Beispielen. Unser Dank gilt auch Frau Hestermann-Beyerle und Frau Lempe vom Springer-Verlag für die gute und angenehme Zusammenarbeit. Um zukünftig mit neuen MAPLE-Versionen Schritt halten zu können, werden Updates der Worksheets unter der Homepage zum Buch

http://www.home.hs-karlsruhe.de/~weth0002/buecher/revokos/start.htm

zur Verfügung gestellt.

Stuttgart, im Dezember 2005, Claus-Dieter Munz, Thomas Westermann

Inhaltsverzeichnis

0 Einleitung

Die mathematische Modellierung von Vorgängen aus der Physik und dem Ingenieurbereich führt sehr oft auf Differenzialgleichungen. Entweder sind diese gewöhnliche Differenzialgleichungen, bei denen die gesuchte Lösung nur von einer Variablen - Raum oder Zeit - abhängt, oder partielle Differenzialgleichungen, bei denen mehrere Raumdimensionen und die Zeit eine Rolle spielen. Da nur sehr wenige dieser Differenzialgleichungen exakt lösbar sind, ist man in der Praxis auf die näherungsweise Lösung dieser Gleichungen angewiesen. Die numerische Lösung von Differenzialgleichungen ist somit ein wichtiges Teilgebiet der numerischen Mathematik, eng verbunden mit den Ingenieur- und Naturwissenschaften.

Zur numerischen Simulation eines praktischen Problems gehören mehrere Arbeitsschritte: Man benötigt ein geeignetes numerisches Verfahren, formuliert in einem Algorithmus, kann man dieses in ein Rechenprogramm umsetzen und letztlich auf das entsprechende Problem anwenden. Um eine numerische Simulation erfolgreich ausführen zu können, benötigt man gute Kenntnisse und einige Erfahrung in den einzelnen Schritten. In diesem „Buch“ wird der Versuch unternommen, Ideen, Theorie der numerischen Verfahren, Algorithmus und Programmstruktur bis hin zum praktischen Test und der Anwendung mit einzubeziehen. In dem pdf-File der CD-ROM gibt es Links zu elektronischen Arbeitsblättern, in denen die Algorithmen mit grafischer Ausgabe schon fertig umgesetzt sind, und mit deren Hilfe viele Beispiele interaktiv bearbeitet werden können. Im Hintergrund steht das Computer-Algebra-System Maple, welches eine Programmiersprache und geeignete Visualisierungsmöglichkeiten enthält. Damit kann der „Leser“ die verschiedenen Verfahren ausführen und sich die Ergebnisse anschauen, welche als Abbildungen auch im Buch zur Verfügung stehen. Er kann aber darüberhinaus selbständig Diskretisierungsparameter, aber auch Programmteile ändern und neue Probleme lösen. Mit wenigen Voraussetzungen sollte es somit möglich sein, eigene Erfahrungen in der Anwendung numerischer Methoden zu sammeln.

Um diese interaktive Komponente nutzen zu können, muss Maple auf dem Rechner installiert sein. Steht Maple nicht zur Verfügung oder möchte man nur die algorithmische Umsetzung des Verfahrens anschauen bzw. nur die

grafischen Ergebnisse betrachten, kann man mit der html-Version arbeiten, bei der die elektronischen Worksheets als html-Files vorliegen. Dann ist allerdings kein interaktives Arbeiten mehr möglich.

Der Schwerpunkt dieses Buches liegt, neben der Beschreibung der numerischen Verfahren, auf der Vermittlung der praktischen Aspekte in der numerischen Lösung von Differenzialgleichungen bis hin zur Simulation praktischer Probleme. Die Ableitung eines Algorithmus, d.h. der Angabe von aufeinander folgenden Rechenschritten in der Reihenfolge, wie sie dann ein Computer berechnen kann, wird durch die Angabe von Struktogrammen aufgezeigt. Zur Vorbereitung werden in der Einleitung die grundlegenden Elemente eines Struktogramms beschrieben, eine kurze Einführung in die Bedienung der CD-ROM schließt sich daran an. Zunächst wird aber mit Bemerkungen über die möglichen Fehler bei einer numerischen Simulation gestartet.

0.1 Modellierungsfehler, Approximationsfehler und Rundungsfehler

Wird Information über einen technischen oder physikalischen Vorgang gesucht, dann sind in der Regel Daten oder Zahlen gesucht, z.B. wie viel Last hält ein Balken aus? Um solche Aussagen zu bekommen, muss zuerst ein physikalisches Modell vorhanden sein. Mit dem zugehörigen mathematischen Modell, den entsprechenden Gleichungen, welche diesen Vorgang beschreiben, hat man dann den Gegenstand, mit dem gerechnet werden kann. Für viele Probleme ist das mathematische Modell eine oder mehrere Differenzialgleichungen. Überall dort, wo die Änderung einer Größe von deren Wert abhängt, ergibt sich eine solche.

Nehmen wir als Beispiel ein Strömungsproblem. Hier besteht das physikalische Modell aus den Erhaltungssätzen für Masse, Impuls und Energie. Das mathematische Modell wird daraus abgeleitet und besteht aus Differenzialgleichungen für die Massendichte, die Geschwindigkeit und den Druck. In der physikalischen und mathematischen Modellierung eines Vorgangs werden Einflüsse vernachlässigt. Man nimmt an, dass sie die Lösung sehr wenig beeinflussen und im Rahmen der geforderten Genauigkeit keine Rolle spielen. Die Vereinfachung des Problems reduziert den Rechenaufwand. Bei einer Strömungsberechnung könnte dies eine Vernachlässigung der Reibungsterme sein, wenn diese sehr klein sind. Fehler, welche man dadurch macht, nennt man **Modellierungsfehler**. Auf diese Modellierungsfehler werden wir im Folgenden nicht weiter eingehen, da sie Thema der mathematischen Modellierung und nicht der numerischen Approximation sind. Man sollte in praktischen Anwendungen diese Fehler jedoch im Kopf behalten. Falls später die

erzielten Ergebnisse nicht mit der Erwartung und der Wirklichkeit übereinstimmen, könnte dies natürlich auch der Einfluss von unterschätzten Modellierungsfehlern sein. Bei der oben erwähnten Strömung kann die Reibung bei der Umströmung von Körpern eine wichtige Rolle spielen. Die Notwendigkeit der Modellierung der Reibungsterme wird sich dann bei der Diskussion der Ergebnisse zeigen.

Beispiel 0.1 (Flug in konstanter Höhe). Ein einfaches Modell für den Flug eines Flugzeuges in konstanter Höhe ohne Richtungsänderung ist die Differenzialgleichung

$$\dot{v} = \frac{d}{dt}v = \frac{F - W}{m} \tag{0.1}$$

für die Geschwindigkeit v. Dabei bezeichnet F den Schub der Turbinentriebwerke, W den Widerstand und m die Masse des Flugzeuges. Ist die Zeit die unabhängige Variable, schreibt man für die Ableitung statt v' auch gerne $\dot{v}$ mit der Bedeutung der zeitlichen Änderung der Geschwindigkeit.

Das zu Grunde liegende physikalische Modell ist hier der Erhaltungssatz des Impulses. Mathematisch formuliert entspricht er der Gleichung des Newtonschen Gesetzes: Kraft = Masse · Beschleunigung. Der Widerstand hängt von der Geschwindigkeit ab und wird hier mit einer sogenannten quadratischen Polare modelliert,

$$W = q\,S\,(c_{W_0} + k\,c_A^2)\ ,$$

wobei $q = \rho\frac{v^2}{2}$ den Staudruck, S die Bezugsflügelfläche, k den Widerstandsfaktor, c_A den Auftriebsbeiwert und c_{W_0} den Nullwiderstandsbeiwert bezeichnet. Das physikalische Modell ist der Erhaltungssatz des Impulses, der sich als mathematisches Modell in der Gleichung (0.1) ausdrückt.

Es treten hier eine ganze Reihe von Größen auf, die das Problem beeinflussen. Ziel in der physikalischen und mathematischen Modellierung ist es, den Einfluss dieser Größen für das betrachtete Problem zu untersuchen. Dabei können noch verschiedene Annahmen eingehen, welche die Gleichung vereinfachen. Horizontalflug ohne Richtungsänderung bedeutet, dass der Auftrieb A gerade das Gewicht G kompensiert. Es gilt somit mit der Gravitationskonstanten g

$$G = m\,g = q\,S\,c_A = A\ .$$

Aus dieser Gleichung ergibt sich durch Auflösen

$$c_A = \frac{m\,g}{q\,S} = \frac{G}{q\,S}\ .$$

Wird dieser Auftriebsbeiwert oben eingesetzt, erhält man den Widerstand W in der Form

$$W = q\,S\,c_{W_0} + \frac{k\,G^2}{q\,S}\ .$$

Wir machen die weiteren Vereinfachungen, dass der Nullwiderstandsbeiwert c_{W_0} und der Widerstandsfaktor k konstant sind. Dies ist für einen Flug im Unterschallbereich gerechtfertigt. Auch das Gewicht wird als konstant betrachtet, was allerdings nur für kurze Flugzeiten gelten sollte, da Treibstoff verbraucht und das Flugzeug somit leichter wird. Weiterhin nehmen wir an, dass der Schub konstant sei und nicht von der Geschwindigkeit v abhängt. Die rechte Seite der Differenzialgleichung ist damit eine Funktion von v und besitzt die Form

$$\dot{v} = f(v) = a_0 - a_1 v^2 - \frac{a_2}{v^2} \,, \tag{0.2}$$

wobei die Konstanten a_0, a_1, a_2 sich nach

$$a_0 = \frac{F}{m} \,, \qquad a_1 = \frac{\rho\, S\, c_{W_0}}{2m} \,, \qquad a_2 = \frac{2k\, G^2}{m\, \rho\, S}$$

berechnen. Wenn die Geschwindigkeit zum Zeitpunkt $t_0 = 0$ vorgegeben ist, ergibt die Lösung des Anfangswertproblems die Geschwindigkeit als Funktion der Zeit.

Aus dem physikalischen oder ingenieurwissenschaftlichen Problem wurde ein mathematisches Modell abgeleitet. Dabei treten in diesem Anwendungsproblem eine ganze Reihe von Größen und Funktionen auf. Das Problem wurde mit Vereinfachungen übersichtlich geschrieben, so dass die mathematische Struktur des Modells klar wird. Diese Struktur ist wesentlich für die Wahl eines geeigneten Näherungsverfahrens. Wie schon erwähnt, sollte man die verschiedenen Annahmen und Vereinfachungen in der mathematischen Modellierung aber nicht vergessen, wenn es an die Interpretation der numerischen Ergebnisse geht. Wir werden dieses Beispiel in dem Kapitel über Anfangswertprobleme für gewöhnliche Differenzialgleichungen mit der Vorgabe von realistischen Daten wieder aufgreifen und Näherungsverfahren darauf anwenden. □

In den meisten Fällen kann das mathematische Modell nicht exakt gelöst werden - es muss näherungsweise berechnet werden. Es ist im Allgemeinen ein kontinuierliches Problem und passt so nicht in den Computer. Rechenprogramme auf einem Rechner arbeiten mit endlich vielen diskreten Größen. Man sucht etwa eine Approximation der Lösung an einer endlichen Anzahl von Punkten und reduziert das kontinuierliche Problem zunächst auf ein endliches diskretes Problem, das dann in den Rechner mit endlich viel Speicherplatz passt. So kann man z.B. bei einer Fourier-Reihe nicht unendlich vielen Glieder aufsummieren, sondern muss nach endlich vielen abbrechen. Den Fehler, den man dabei macht, nennt man **Approximations-, Verfahrens- oder auch Konsistenzfehler**. Dieser Fehler wird reduziert, wenn die Anzahl der Punkte erhöht wird und das Problem somit besser aufgelöst wird. Bei der Auswertung der Fourier-Reihe wird die Anzahl der aufsummierten Glieder erhöht, wobei dann der Rechenaufwand anwächst.

Ein anderer Fehler, welcher bei numerischen Rechnungen auftritt, ist der **Rundungsfehler**. Dieser Fehler ergibt sich wieder aus der Tatsache, dass der Computer nicht unendlich viele Zahlen speichern kann. So besitzt jede Zahl nur eine endliche Genauigkeit, d.h. eine gewisse Anzahl von Stellen. Die Zahl π wird niemals exakt auf dem Rechner gespeichert sein. Während die Approximationsfehler gewissermaßen die Folge davon sind, dass die numerischen Verfahren eine Approximation des Problems darstellen, sind die Rundungsfehler die Folge davon, dass ein Computer zwar viele aber nicht unendlich viele Daten speichern kann. Wir schauen uns zwei Beispiele an, welche den Einfluss von Rundungsfehlern aufzeigen.

Beispiel 0.2. Man berechne die Lösungen der quadratischen Gleichung

$$x^2 - 2x + 0.999999999999 = 0\,. \tag{0.3}$$

Man hat hier auf der linken Seite fast das Quadrat von $x - 1$ vorliegen und würde somit zwei Lösungen nahe 1 erwarten. Nach der Formel für eine quadratische Gleichung ergeben sich die zwei Lösungen

$$x_1 = 1 + \sqrt{10^{-13}} \text{ und } x_1 = 1 - \sqrt{10^{-13}}\,. \tag{0.4}$$

Ein Taschenrechner liefert bei der Berechnung der Wurzel die Zahl 0.0, so dass sich hier die Lösungen $x_1 = x_2 = 1$ ergeben. Diese Lösung ist nur bis auf die Rundungsfehler genau. Wird die Lösung $x_1 = 1$ oben in die Gleichung eingesetzt, so ergibt sich der Fehler zu 0.0000000000001. Wird mit dieser Lösung weiter gerechnet, so wird dies nur dann Schwierigkeiten machen, wenn die weitere Rechnung bezüglich diesem Fehler empfindlich ist. Durch eine Umformulierung des arithmetischen Ausdrucks kann das Problem oftmals entschärft werden. □

Beispiel 0.3. Wir betrachten die Auswertung der Funktion

$$f(x) = x^3(\frac{x}{x^2 - 1} - \frac{1}{x}) \tag{0.5}$$

für große x. Rechnen wir mit einer endlichen Anzahl von Stellen, dann wird der Ausdruck in der Klammer für größer werdende x gegen Null streben. Wir nehmen an, dass x so groß ist, dass die Addition der Zahl 1 die Darstellung dieser Zahl im Rechner nicht mehr ändert. Werden auf dem Rechner für jede Zahl 10 Stellen und der Exponent abgespeichert, dann wird dies der Fall sein, wenn die Addition die 11. oder eine weiter hinten liegende Stelle ändert. Damit stimmt in dem Raum der Maschinenzahlen $x^2 - 1$ mit x^2 überein. Das Ergebnis der Klammer wird bei dieser Rechnung Null sein. Dies kann man leicht mit dem Taschenrechner ausprobieren. Es tritt ein sogenannter Auslöschungseffekt auf. Diese Klammer wird dann allerdings mit einer sehr

großen Zahl, nämlich x^3 multipliziert, so dass der Fehler bei der Auswertung der Funktion für solch große x-Werte beträchtlich ist.

Nun lässt sich die Funktion etwas umformulieren. Indem man die beiden Brüche zusammenfasst, ergibt sich nach kurzer Rechnung eine andere Darstellung der Funktion f in der Form

$$f(x) = \frac{x^2}{x^2 - 1} = \frac{1}{1 - x^{-2}} \; . \tag{0.6}$$

Der rechte Ausdruck ist für große x sehr gutmütig auszuwerten. Insbesondere sieht man sofort, dass der Wert der Funktion f immer größer als Eins ist. Bei einer Auswertung der Funktion nach der Formel (0.5) lässt sich diese Eigenschaft nicht reproduzieren. Bei der Programmierung arithmetischer Ausdrücke sollte man somit darauf achten, solche Auslöschungsfehler zu umgehen. □

In technischen und physikalischen Anwendungen treten solche Situationen, bei denen die Differenz sehr kleiner Zahlen mit einer großen Zahl multipliziert wird, selten auf. Die Größen dort haben eine physikalische Bedeutung. Es gibt selten einen Grund, wieso sehr große und sehr kleine Werte in den Gleichungen auftreten sollten, deren Multiplikation miteinander einen für den Vorgang bestimmenden Wert liefern. Meist ist es aber sinnvoll, die Gleichungen vor der Rechnung so dimensionslos zu formulieren, dass Zahlen ähnlicher Größenordnungen auftreten.

In Anwendungen gibt es jedoch auch Probleme, bei denen Fehler oder Störungen schnell anwachsen können, wie bei der Berechnung von Instabilitäten. Hier ist vor einer numerischen Simulation zu klären, was die Rechnung an Information liefern kann. Ist die Vorhersagbarkeit des Problems überhaupt gewährleistet oder muss zu einer stochastischen Beschreibung des Problems übergegangen werden. Dies ist aber dann wieder eine Frage der mathematischen Modellierung.

Mit den Rundungsfehlern werden wir uns im Folgenden wenig befassen. Man muss allerdings wissen, dass eine Erhöhung der Genauigkeit in einer numerischen Simulation, z.B. durch die Vergrößerung der Anzahl der Gitterpunkte, nicht beliebig durchgeführt werden kann, da sonst die Rundungsfehler größer als die Verfahrensfehler werden. Dies ist in der Abbildung 0.1 schematisch dargestellt. Benötigt man sehr genaue Ergebnisse, so müssen in diesem Fall bessere numerische Verfahren eingesetzt werden, welche mit der gleichen Anzahl von Punkten die genaueren Ergebnisse liefern. Praktische Beispiele mit Diskussion der Rundungsfehler zeigen wir schon im nächsten Kapitel bei der numerischen Integration und Differenziation.

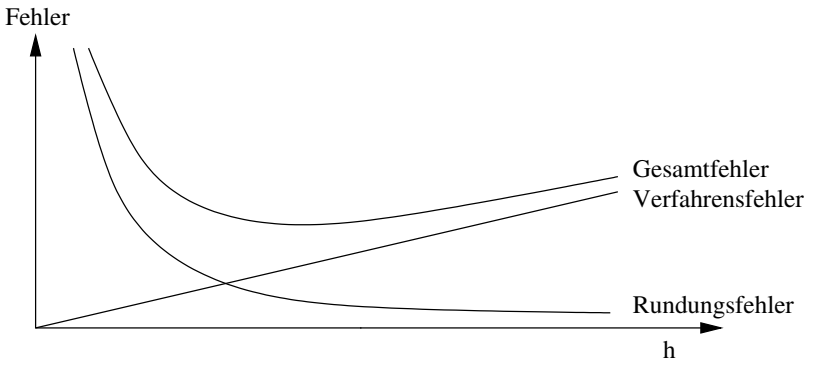


Abb. 0.1. Fehlerverhalten der numerischen Lösung

Weitere Fehler treten dann beim eigenen Programmieren auf: Die **Programmierfehler**, auch **Bugs** genannt, was übersetzt sowohl „technischer Defekt“ als auch im amerikanischen Slang „verrückt machen“ heißen kann. In manchen Fällen passen beide Übersetzungen sehr gut. Diese werden wir hier natürlich auch nicht behandeln. Es kommt oft auf die Erfahrung des Anwenders an, die verschiedenen Fehler richtig zuzuordnen, zu finden und zu korrigieren. Neben der Wahl einer guten numerischen Methode für ein Problem ist die Bewertung der numerischen Daten und die Abschätzung der Fehler eine wichtige Aufgabe der numerischen Simulation.

0.2 Struktogramme

In den folgenden Kapiteln beschreiben wir die numerischen Verfahren. Für die Berechnung eines Problems müssen diese in ein Rechenprogramm umgesetzt werden. Dabei sollte zuvor überlegt werden, wie die Abfolge der einzelnen Schritte aussehen muss. Das numerische Verfahren wird als **Algorithmus** angegeben, durch den die notwendige Reihenfolge der Rechnungen aufgezeigt und die Grundstruktur eines Programmes festgelegt ist.

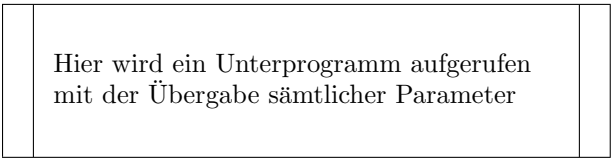
Es gibt verschiedene Möglichkeiten, numerische Algorithmen darzustellen. Man kann sie durch Text beschreiben, den so genannten Pseudocode verwenden, oder eine grafische Darstellung wählen. Unter Pseudocode versteht man ein in einer vereinfachten Programmiersprache geschriebenes Programm. Für die grafische Darstellung kennt die DIN-Norm zwei formale Systeme: Programmablaufpläne (DIN 66001) und Struktogramme (DIN 66261). Letztere sind auch unter dem Namen Nassi-Shneiderman-Diagramme bekannt. Sie wurden eingeführt, nachdem sich die alten Programmablaufpläne für die Anforderungen moderner Informatik und strukturierten Programmierens als

nicht ausreichend erwiesen hatten. Im vorliegenden Buch sollen neben der verbalen Beschreibung ausschließlich diese zum Einsatz kommen. Um ein solches Diagramm zu verstehen, muss man die Bedeutung der verwendeten Symbole kennen. Diese werden im Folgenden beschrieben.

Einfache Anweisungen

Anweisungen stehen in gewöhnlichen Rechtecken:

Wird in der Anweisung ein Unterprogramm oder ein Funktionsunterprogramm aufgerufen, so wird dies durch zusätzliche senkrechte Striche an der Seite gekennzeichnet:

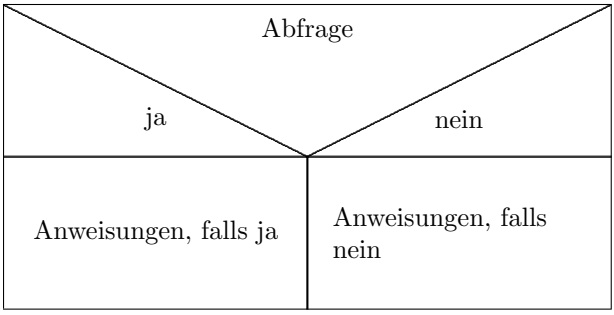


Unterprogramme dienen vor allem der Übersichtlichkeit eines Programms. Je mehr Anweisungen, oft Code genannt, an einem Stück geschrieben werden, desto schwieriger wird es, das Programm zu überblicken. Unterprogramme schaffen hier Ordnung, da sie die Möglichkeit bieten, Teile des Codes aus dem Hauptprogramm auszulagern. Diese Unterprogramme werden dann an den entsprechenden Stellen im Hauptprogramm aufgerufen. Gerade wenn eine Sequenz von Anweisungen an mehreren Stellen im Programm ausgeführt wird, muss sie als Unterprogramm geschrieben nicht jedes mal neu programmiert werden, sondern wird dann mit aktuellen Übergabe-Parametern an den verschiedenen Stellen aufgerufen.

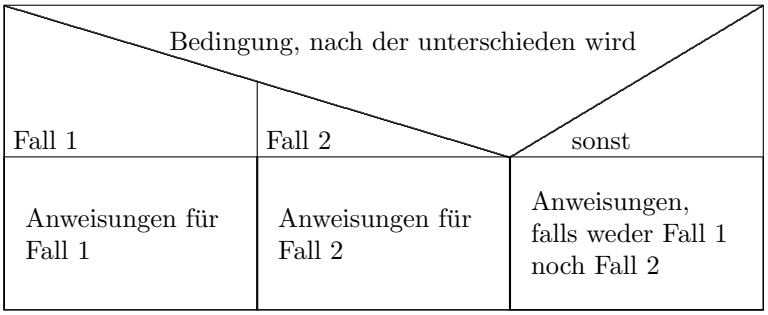
Bei dem Aufruf von Unterprogrammen können mehrere Werte übergeben und auch mehrere Werte zurück gegeben werden. Die Funktion eines Unterprogramms könnte z.B. die Lösung von Gleichungssystemen sein, hier werden die Koeffizientenmatrix und die rechte Seite des Gleichungssystems übergeben und die gesuchte Lösung wird zurück gegeben. Im Spezialfall der Rückgabe eines Wertes spricht man auch von einer Funktion.

Verzweigungen

Bei vielen Algorithmen müssen Fallunterscheidungen getroffen werden, d.h. in verschiedenen Fällen muss ein unterschiedliches Vorgehen gewählt werden. Man braucht Verzweigungen. Dabei unterscheidet man zwischen einfachen:



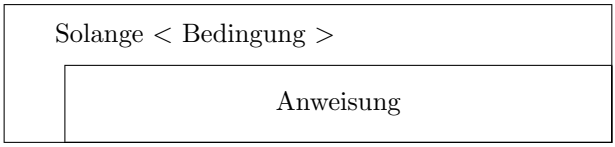
und mehrfachen Verzweigungen:



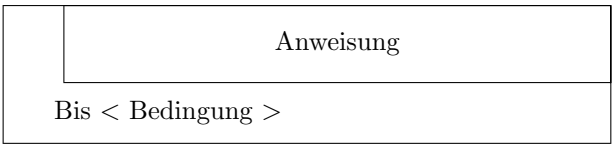
Wiederholungsschleifen

Da oft gewisse Teile eines Algorithmus mehrfach durchlaufen werden müssen, benötigt man Symbole für diese Wiederholungen. Dies ist insbesondere bei numerischen Algorithmen oft der Fall. Es gibt bei den Struktogrammen zwei Möglichkeiten, dieses zu beschreiben:

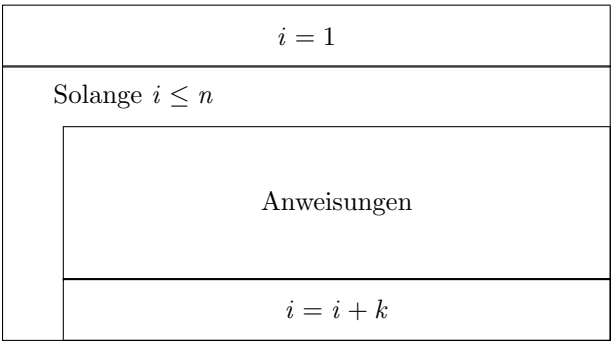
1. Man prüft am Anfang, ob die Schleife überhaupt durchlaufen werden soll:



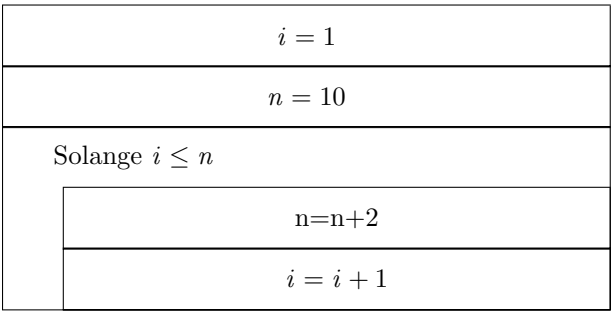
2. Man prüft am Ende, ob ein Abbruchkriterium erfüllt ist:



Ein Symbol für sogenannte Zählschleifen sieht die Norm nicht vor, jedoch gibt es in vielen Programmiersprachen Zählschleifen, um Algorithmen effizient programmieren zu können. Diese können aus den bekannten Formen für die Wiederholungsschleifen erstellt werden:



Die Schrittweite k sowie der maximale Index n sollten nach Möglichkeit in der Schleife nicht verändert werden. Eine sogenannte Endlosschleife wurde im folgenden Beispiel programmiert:



In diesem Beispiel wird der maximale Index n immer größer sein als der aktuelle Index i, was dazu führt, dass das Abbruchkriterium nicht erfüllt werden kann. Der Computer würde hier so lange hochzählen bis der maximal zulässige Wert für die Variablen n bzw. i erreicht ist und dann die Programmausführung mit einer Fehlermeldung beenden. In den seltenen Fällen, in denen eine Änderung des Laufindex sinnvoll ist, sollte der Programmierer aber un-

bedingt sicher stellen, dass das Abbruchkriterium der Schleife erfüllt werden kann.

Struktogramme, die in ein Computerprogramm umgesetzt werden, sollten den folgenden Aufbau besitzen:

1. **Deklarationen**
   Hier werden alle im Programm vorkommenden Variablen mit ihrem Typ deklariert. Auch die Namen der verwendeten Funktionsunterprogramme müssen hier mit ihrem Variablentyp bekannt gemacht werden.

2. **Eingangswerte**
   An dieser Stelle sollten alle Werte eingelesen werden, die von außen, d.h. einer Eingabe oder einem übergeordneten Programm, übernommen werden. In Programmiersprachen geschieht dies oft durch die Übergabeliste eines Unterprogramms. Im Struktogramm sollte es an dieser Stelle explizit geschehen.

3. **Initialisierung von Variablen**
   In einem Programm gibt es Befehlszeilen wie etwa $a = a + b$. Damit so etwas funktioniert, muss a vorher schon einen bestimmten Wert haben. Alle Initialisierungen dieser Art sollten hier zusammengefasst werden.

4. **Der Algorithmus**
   Erst hier steigt man in den eigentlichen Algorithmus ein.

5. **Aus-, Weiter-, Rückgabe von Werten**
   Zum Schluss müssen noch Werte ausgegeben, weitergegeben oder an ein übergeordnetes Programm zurückgegeben werden.

Bemerkungen

1. Beim Erstellen eines Struktogramms sollte man bei den Punkten 1 und 3 zunächst noch Platz übrig lassen. Überprüfen sie am Ende alles auf Vollständigkeit. Die Erfahrung sagt, dass an diesen Stellen meist noch etwas nachkommt.

2. In der Literatur angegebene Struktogramme sind i.A. nicht vollständig. Es werden meist nur die Teile des Algorithmus, nämlich 3 und 4 abgedruckt.

3. Beim Umsetzen von Unterprogrammen in eine Programmiersprache stehen oft Teile von 2 vor 1, weil die Werte durch die Übergabeliste eingelesen werden, und diese üblicherweise direkt beim Namen des Unterprogramms

steht. Der Compiler bringt dies in die oben angegebene Reihenfolge. Daher sollte man diese auch im Struktogramm einhalten.

Beispiel 0.4. Ein Struktogramm für ein Programm zur Berechnung von $a+b$ sieht folgendermaßen aus:

Real: a, b, `AplusB`
Lies a und b ein
`AplusB` $= a + b$
gebe `AplusB` zurück

Die Eingabe der Zahlen erfolgt hier direkt am Bildschirm durch eine Aufforderung, der Betrag der Zahl wird berechnet und für den Benutzer am Bildschirm zurückgegeben. □

Beispiel 0.5. Das zweite Beispiel ist ein Struktogramm für ein Unterprogramm, das von einem übergeordneten Programm aufgerufen wird:

Integer: $i, N,$ `Summe`	
Übernehme N	
$i = 1$	
`Summe` $= 0$	
$i \leq N$	
	`Summe` $=$ `Summe` $+ i$
	$i = i + 1$
Gib `Summe` zurück	

In dem Unterprogramm werden die Zahlen von 1 bis N aufsummiert:

$$\sum_{i=1}^{N} i \ .$$

Der Wert N wird an das Unterprogramm übergeben. Die Initialisierung der Variablen *Summe* zu 0 ist wichtig, da eine nicht initialisierte Variable nicht notwendigerweise den Wert Null enthalten muss. Manche Compiler tun dies allerdings automatisch. □

In den einzelnen Kapiteln geben wir exemplarisch Struktogramme an, so dass man die Schritte vom numerischen Verfahren zum Algorithmus und dem Rechenprogramm nachvollziehen kann.

0.3 Arbeiten mit der CD-ROM

Mit der CD-ROM wurde ein Medium geschaffen, um dem Leser erste Erfahrungen mit dem Umgang numerischer Methoden zu ermöglichen. Es sollte aber auch für Dozenten interessant sein, die ein flexibles Mittel suchen, numerische Methoden sichtbar und anschaulich begreifbar zu machen. Die elektronischen Arbeitsblätter oder Worksheets sind daher so aufgebaut, dass jeder sie nutzen kann, ohne auf dem Gebiet der Computer-Algebra-Systeme ein Experte sein zu müssen. Da die elektronischen Arbeitsblätter lauffähig aufbereitet sind, können sie sofort ausgeführt werden. Wer die CD nutzt, darf einen reibungslosen Einstieg in die Beispiele erwarten.

Systemvoraussetzungen: Um die elektronischen Arbeitsblätter der CD-ROM benutzen zu können, müssen die folgenden Voraussetzungen erfüllt sein:

– Die Worksheets wurden für Maple 9 bzw. Maple 9.5 entwickelt. Eine entsprechende Installation des Computer-Algebra-Systems muss vorhanden sein. Das Ansteuern der Maple-Worksheets erfolgt entweder über die pdf-Version des Buches oder durch Doppelklick auf die Datei index.mws.

– Die html-Versionen der Worksheets können mit jedem üblichen Web-Browser betrachtet werden. Eine interaktive Veränderung der Arbeitsblätter ist dann allerdings nicht möglich. Zum Starten der html-Dateien öffnen sie bitte die Datei index.html durch Doppelklick.

Bedienungserklärung:

– Die Berechnungen auf einem Arbeitsblatt können unter Maple durch Drücken der Enter-Taste zeilenweise oder aber komplett über den Menüpunkt `Edit` - `Execute` - `Worksheet` ausgelöst werden.

– Über die Bildlaufleisten kann ein beliebiger Ausschnitt ausgewählt werden.

– Einige Plots sind als Animationen vorbereitet. Eine Animation ist eine gespeicherte Folge von Einzelbildern. Diese können fortlaufend wie ein Film oder einzeln betrachtet werden. Man erkennt eine Animation daran, dass nach einem Klick auf das Bild eine neue Symbolleiste sichtbar wird. Über diese Symbolleiste oder durch die Menüführung der rechten Maustaste kann die Animation gestartet werden.

– Einige Plots sind dreidimensionale Darstellungen, bei denen der Blickwinkel verändert werden kann. Klicken sie dazu auf das Bild und halten Sie die linke Maustaste gedrückt. Wenn sie jetzt die Maus bewegen, dreht sich das Koordinatenkreuz.

Arbeiten mit der mws-Datei: Durch Doppelklicken der Datei index.mws öffnet man das MAPLE-Inhaltsverzeichnis. Durch anschließendes Anklicken des gewünschten Abschnitts wird das zugehörige MAPLE-Worksheet gestartet und ist dann interaktiv bedienbar. Mit der →-Taste der oberen Taskleiste kommt man vom Worksheet wieder zum Inhaltsverzeichnis zurück. Die einzelnen Worksheets sind auch separat anwählbar.

Arbeiten mit der pdf-Version: Durch Doppelklicken der Datei buch.pdf öffnet man den Inhalt des Buches.

– Die linke Spalte (Lesezeichen) bildet das Inhaltsverzeichnis des Buchs ab. Man wählt das gewünschte Themengebiet aus. Dann springt der Cursor auf die entsprechende Stelle im Buch (rechter Bildschirminhalt).

– Vom rechten Bildschirminhalt aus können die zugehörigen MAPLE-Worksheets durch Anklicken des blau gekennzeichneten Links Worksheet gestartet werden. Durch Schließen der Worksheets kommt man wieder zur pdf-Version des Buches zurück.

– Um innerhalb des Textes zu navigieren, kann auch der Index verwendet werden, da die angegebenen Seitenzahlen mit den Textstellen verlinkt sind, d.h. durch Anklicken der Seitenzahl im Index springt der Cursor an die zugehörige Textstelle.

Hinweise zum Starten der Worksheets aus dem Acrobat Reader unter Linux: Um die Worksheets unter Linux direkt aus dem Acrobat Reader aufrufen zu können, ist es unbedingt erforderlich, daß die Datei von der Konsole aus aufgerufen wird, um zu gewährleisten, dass die Pfade richtig gesetzt sind. Öffnen Sie hierzu eine Konsole und wechseln Sie in das Verzeichnis, in dem die CD gemountet ist, z. B. mit `cd /media/cdrom`. Öffnen Sie dann das Dokument mit der Eingabe `acroread buchlinux.pdf`.

Beim Aufruf eines Worksheets aus dem Acrobat Reader unter Linux werden zunächst zwei Warnmeldungen ausgegeben. Zunächst warnt das Programm davor, dass die Datei nicht unterstützt wird bzw. beschädigt ist. Diese Meldung kann durch Klicken auf „OK“ übergangen werden. Danach warnt der Reader, dass die auszuführende Datei ein potentielles Sicherheitsrisiko darstellt. Klicken Sie bei diesem Dialog auf „Öffnen“, um die Datei auszuführen - das MAPLE-Worksheet sollte nun ausgeführt werden.

Die hier verwendeten Startskripte setzen voraus, dass MAPLE unter Linux mit dem Befehl `xmaple` von der Konsole aus aufgerufen werden kann. Dieser Befehl startet die neue, Java-basierte Version, die unter aktuellen Linux-Versionen mit den Standardeinstellungen nach der Installation nicht korrekt funktioniert. Hier gibt es zwei Möglichkeiten, Abhilfe zu schaffen:

1. Anstatt der Java-basierten Version kann die „Classic-Worksheet-Variante“ von MAPLE verwendet werden, die mit `xmaple -cw` gestartet wird. Hierzu kann z.B. ein Alias gesetzt werden, der den Befehl `xmaple` auf `xmaple -cw` umleitet:

   ```
   alias xmaple="xmaple -cw"
   ```

2. Die Fehlerhafte Ausführung der Java-basierten Version von MAPLE unter Windows liegt an einem Problem der mitgelieferten Java-Runtime-Umgebung mit IPv6. Dieses Problem läßt sich allerdings relativ leicht beheben. Angenommen, MAPLE ist in `/opt/maple9.5` installiert, geht man folgendermaßen vor:

   – Umbenennen der Datei `java` im Verzeichnis `/opt/maple9.5/jre.IBM_INTEL_LINUX/bin` in `java.orig`

   – Erstellen einer neuen Datei mit Namen `java` mit folgendem Inhalt:
   `/opt/maple9.5/jre.IBM_INTEL_LINUX/bin/java.orig -Djava.net.preferIPv4Stack="true" $*` (in eine Zeile schreiben)

   – Die neue erstellte Datei ausführbar machen:
   `chmod 755 java`

   Ersetzen Sie `/opt/maple9.5` ggf. durch Ihren Installationspfad.

Eine kurze Einführung in das Computer-Algebra-System MAPLE findet man z.B. in [73] oder unter der Adresse

http://www.home.hs-karlsruhe.de/~weth0002/veranstaltungen/details/docs/intro.zip.

1 Numerische Integration und Differenziation

Die Approximation bestimmter Integrale tritt bei der numerischen Simulation von praktischen Ingenieurproblemen häufig auf. So führt die Lösung der einfachsten Differenzialgleichungen, bei denen die rechte Seite nicht von der gesuchten Funktion abhängt, auf die Berechnung eines bestimmten Integrals. Dabei ist eine exakte Integration durch die Angabe einer Stammfunktion für viele Anwendungen nicht möglich oder auch für die Berechnung zu aufwändig. Die **numerische Integration**, welche auch **numerische Quadratur** genannt wird, ist für ein Buch über die numerische Lösung von Differenzialgleichungen somit ein guter Startpunkt. Die Approximation von Ableitungen, die numerische Differenziation, liefert ebenso Hilfsmittel für die numerische Behandlung von Differenzialgleichungen.

Die einfachste Differenzialgleichung hat die Form

$$y'(x) = f(x) \; . \tag{1.1}$$

Mit der Vorgabe des Anfangswertes y_a an einer Stelle $x = a$ besitzt diese Gleichung für eine stetige Funktion $f = f(x)$ eine eindeutige Lösung. Man nennt

$$y'(x) = f(x) \; , \qquad y(a) = y_a \tag{1.2}$$

das **Anfangswertproblem (AWP)** der Differenzialgleichung (1.1). Da die rechte Seite f nur von x und nicht von der gesuchten Funktion y abhängt, erhält man die Lösung an einem beliebigen Punkt $x = b > a$ direkt durch die Integration der Differenzialgleichung (1.1) über das Intervall $[a, b]$. Zunächst ergibt sich die Integralgleichung

$$\int_a^b y'(x)\,dx = \int_a^b f(x)\,dx \; .$$

Das Integral auf der linken Seite lässt sich sofort berechnen. Da $y = y(x)$ eine Stammfunktion des Integranden ist, erhält man

$$y(b) = y(a) + \int_a^b f(x)\,dx \; . \tag{1.3}$$

An dem beliebigen Punkt $x = b$ kann man den Wert $y(b)$ somit bestimmen, indem das **bestimmte Integral**

$$Q = \int_a^b f(x)\,dx \tag{1.4}$$

berechnet wird. Wir setzen im Allgemeinen hier voraus, dass die Funktion $f = f(x)$ in $[a, b]$ stetig und die gesuchte Lösung des Anfangswertproblems somit stetig differenzierbar ist.

Grafisch lässt sich das Lösen des Anfangswertproblems (1.2) im Richtungsfeld veranschaulichen. Man zeichnet in den Punkten der (x, y)-Ebene die Steigungen $f(x)$ an. Die Lösung erhält man durch Start bei (a, y_a). Die Steigungen $f(x)$ entsprechen nach (1.1) den Tangentensteigungen an die Kurve $y = y(x)$ (siehe Abbildung 1.1). Da die Steigungen an jeder Stelle unabhängig von y sind, unterscheiden sich zwei Integralkurven nur durch eine additive Konstante. Lösungen der Differenzialgleichung (1.1) zu verschiedenen Anfangswerten gehen durch Parallelverschiebung auseinander hervor.

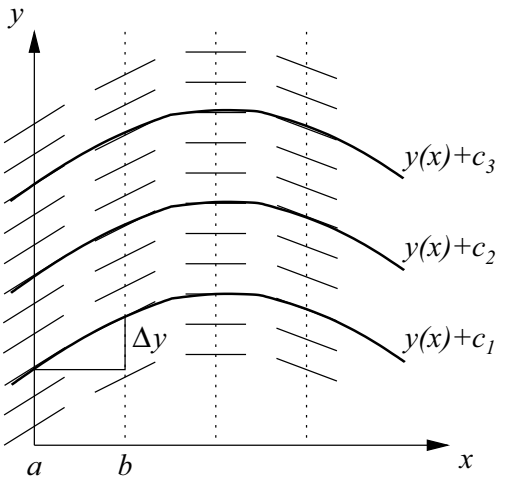


Abb. 1.1. Richtungsfeld und Lösung des Anfangswertproblems (1.2)

Das bestimmte Integral lässt sich bekanntlich als Fläche unter der Kurve $f = f(x)$ zwischen $x = a$ und $x = b$ deuten. Kennt man eine Stammfunktion von $f = f(x)$, dann lässt sich Q nach dem Fundamentalsatz der Differenzial- und Integralrechnung direkt berechnen. Neben den mathematischen Tabellen, in denen für ganze Klassen von Integranden die Stammfunktionen angegeben sind, bieten sich zur Suche einer Stammfunktion heute Computer-Algebra-Systeme an.

Beispiel 1.1 (Exakte Integration, mit MAPLE-Worksheet). Wir berechnen das bestimmte Integral

$$Q = \int_0^3 (x^3 - 2x)\,dx \; . \tag{1.5}$$

In dem Worksheet ist der entsprechende Aufruf zur Berechnung des Integrals dargestellt. Eine Stammfunktion bei diesem Polynom ist $\frac{1}{4}x^4 - x^2$. Für den Wert des Integrals ergibt sich somit

$$Q = \left[\frac{1}{4}x^4 - x^2\right]_0^3 = 11.25 \; , \tag{1.6}$$

welcher mit dem Wert aus dem Worksheet natürlich übereinstimmt. □

Dies ist ein sehr einfaches Beispiel. Sehr häufig existiert jedoch keine geschlossene Lösung eines bestimmten Integrals. In diesem Fall muss das Integral näherungsweise berechnet werden. Es gibt weitere Gründe, wodurch eine numerische Auswertung günstiger oder notwendig sein kann, so z.B. in den Fällen:

– die Auswertung der Stammfunktion ist sehr rechenaufwändig,

– der Integrand liegt nur als Tabelle vor (z.B. Messwerte),

– die gewünschte Genauigkeit ist gering.

Im Folgenden werden wir uns nur noch mit der näherungsweisen Durchführung der Integration beschäftigen.

Als Beispiel für eine Anwendung aus dem Ingenieurbereich, bei dem bestimmte Integrale zu lösen sind, betrachten wir die Biegung eines Balkens.

Beispiel 1.2 (Biegung eines Balkens). Der Lastfall der Balkenbiegung ist einer der am häufigsten in der Mechanik vorkommenden Fälle (siehe z.B. [2], [52]). Wir betrachten im Folgenden ein mathematisches Modell, dessen Vereinfachungen auf folgenden Annahmen beruhen:

– Der Balkenwerkstoff ist homogen und isotrop,
– alle Querschnitte des Balkens bleiben eben und senkrecht zur neutralen Faser,
– Kraft- und Momenteneinleitungspunkte sind weit entfernt von dem zu untersuchenden Bereich,
– das Kräfte-Gleichgewicht wird am unverformten Balken aufgestellt,
– die Längs- sind wesentlich größer als die Querabmessungen.

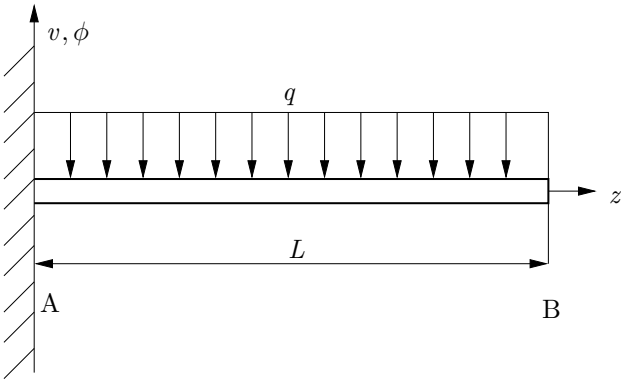


Abb. 1.2. Eingespannter Balken

Diese Annahmen führen auf das folgende mathematische Modell

$$\frac{dQ(z)}{dz} = q(z) \,, \tag{1.7}$$

$$\frac{dM(z)}{dz} = Q(z) \,, \tag{1.8}$$

$$\frac{d\phi(z)}{dz} = -\frac{M(z)}{E(z)I(z)} \left(1 + \phi^2(z)\right)^{\frac{3}{2}} \,, \tag{1.9}$$

$$\frac{dv(z)}{dz} = \phi(z) \,. \tag{1.10}$$

Dabei ist z die unabhängige Variable in Längsrichtung des Balkens. Die gesuchten Funktionen sind die Querkraft $Q = Q(z)$, das Biegemoment $M = M(z)$, der Biegewinkel $\phi = \phi(z)$ und die Durchbiegung $v = v(z)$ in z-Richtung. Bekannt sind die Kräfteverteilung $q = q(z)$, das Elastizitätsmodul $E = E(z)$ und das Trägheitsmoment $I = I(z)$.

In den ersten beiden Gleichungen hängt die rechte Seite nur von z ab. Die Querkraft Q erhält man durch Integration der Kräfteverteilung q. Das Bie-

gemoment M ergibt sich dann durch Integration der Querkraft. Bei kleinen elastischen Verformungen sind die Biegewinkel klein und ϕ^2 kann in (1.9) vernachlässigt werden. Es ergeben sich in diesem linearen Fall vier Differenzialgleichungen, deren rechte Seite nur von z abhängt. Sie können somit sukzessive durch Integration gelöst werden. □

1.1 Die zwei Ideen

Zwei Ideen zur Approximation eines bestimmten Integrals

$$\int_a^b f(x)\,dx \tag{1.11}$$

bieten sich an. Die erste Idee wird motiviert durch die Definition des bestimmten Integrals als Grenzwert der Ober- und Untersumme.

Idee 1.1 (Integral = Grenzwert einer Summe). *Das Integrationsintervall $I = [a, b]$ wird in viele kleine Teilintervalle unterteilt. In jedem Teilintervall wird das Integral durch den Flächeninhalt eines einbeschriebenen Rechtecks angenähert. Das Integral wird durch die Summe der Flächeninhalte der einbeschriebenen Rechtecke angenähert.*

Der Einfachheit halber wird eine äquidistante Zerlegung eingeführt:

$$x_i = a + i\,h,\ \ i = 0, 1, \ldots, n, \qquad \text{mit} \qquad h = \frac{b-a}{n}\,. \tag{1.12}$$

Die Anzahl der sogenannten **Stützstellen** x_i mit $x_0 = a$ und $x_n = b$ ist $n+1$, die Anzahl der **Teilintervalle** ist n. Eine Approximation des Integrals (1.11) im obigen Sinne ist somit zum Beispiel durch

$$Q_h = h \sum_{i=0}^{n-1} f(x_i) \tag{1.13}$$

gegeben.

Die geometrische Interpretation dieser Formel ist einfach: Die Fläche unter der Kurve $y = y(x)$ wird angenähert durch lauter Rechtecke. Man nennt

diese einfache Approximation die **Rechteckregel**. In Abbildung 1.3 sieht man dies skizziert. Der Approximationsfehler, den man macht, entspricht der Fläche zwischen der Kurve und den Rechtecken. Es ist augenscheinlich, dass dieser Fehler kleiner werden sollte, wenn wir nur mehr Rechtecke, d.h. mehr Stützstellen, einführen.

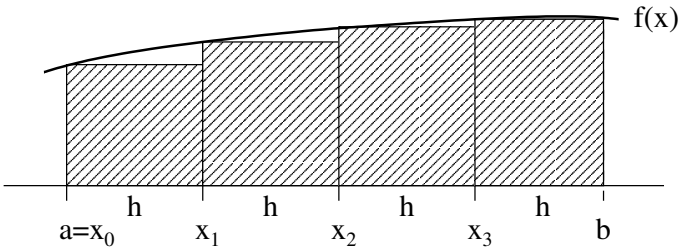


Abb. 1.3. Rechteckregel

In Abbildung 1.4 ist das Struktogramm für diese Rechteckregel als Unterprogramm realisiert. Übergeben werden die Intervallgrenzen und die Zahl der Teilintervalle. Die Funktion $f = f(x)$ ist dabei als Unterprogramm vorgegeben. Nach der Definition der Größen und der Zuweisung von Werten werden in einer Schleife die Funktionswerte aufsummiert. Die Multiplikation mit der Schrittweite erfolgt danach außerhalb der Schleife. Da die Schrittweite hier konstant ist, vermeidet man dadurch Rechenoperationen und spart Rechenzeit. Zwar wird eine solche einfache Optimierung meist von den heutigen Compilern automatisch durchgeführt, es ist aber sicherlich sinnvoll, bei der Programmierung den Aspekt der Einsparung von Rechenoperationen im Blick zu behalten.

Beispiel 1.3 (Rechteckregel, mit MAPLE**-Worksheet).** In dem Worksheet wird mit 20 Stützstellen ein Näherungswert für das Integral Q in (1.5) berechnet. Der Vergleich mit dem exakten Wert liefert den Fehler

$$|Q - Q_h| = |11.25 - 9.725625| = 1.524375$$

und den relativen Fehler

$$\left| \frac{Q - Q_h}{Q} \right| = 13.55\% \, .$$

Erhöht man die Anzahl der Stützstellen im **Worksheet** von 20 auf 50, 100, 200 usw. und vergleicht die Ergebnisse, so zeigt sich, dass die Genauigkeit der Näherung kontinuierlich anwächst. Es sind aber viele Stützstellen erforderlich, um tatsächlich eine gute Genauigkeit zu erreichen. Für das Integral (1.5) mit dem einfachen Integranden $f(x) = x^3 - 2x$ braucht man mehr als

Rechteckregel

Real: a, b, h, Q, x
Real Function: f
Integer: i, N
Übernehme (a, b, N)
$i = 0$
$Q = 0$
$x = a$
$h = \dfrac{b-a}{N}$
Solange $i < N$
$Q = Q + f(x)$
$x = x + h$
$i = i + 1$
$Q = h \cdot Q$
Gib Q zurück

Abb. 1.4. Struktogramm für Rechteckregel als Unterprogramm

280 Stützstellen, um den Fehler kleiner als 1% zu bekommen. Da man an jeder Stützstelle die Funktion f auswerten muss, benötigt man dazu 280 mal die Berechnung von f. □

Bei der Rechteckregel gibt es mehrere Möglichkeiten, die Rechtecke in den Teilintervallen zu definieren. In der Formel (1.13) wird der Funktionswert am linken Randpunkt des Teilintervalls genommen. Man könnte dies auch am rechten Punkt nehmen. Eine einfache Modifikation unserer Formel erhöht im Allgemeinen die Genauigkeit der Approximation. Wird die Funktion f nicht an der linken oder rechten Grenze der Teilintervalle, sondern in der Mitte ausgewertet, erhält man die **Mittelpunktsregel**

$$Q_h = h \sum_{i=0}^{n-1} f_{i+\frac{1}{2}} \quad \text{mit} \quad f_{i+\frac{1}{2}} := f\left(\frac{x_i + x_{i+1}}{2}\right) . \tag{1.14}$$

Der Name ist daraus abgeleitet, dass die Funktion in der Mitte des Teilintervalls ausgewertet wird.

Beispiel 1.4 (Mittelpunktsregel, mit MAPLE-Worksheet). In dem Worksheet wird erneut das Integral

$$Q = \int_0^3 (x^3 - 2x)\, dx$$

numerisch berechnet, jetzt aber mit der Mittelpunktsregel. Es zeigt sich, dass die Mittelpunktsregel für dieses Integral einen Fehler unter 1% schon mit 10 Stützstellen liefert. Die Mittelpunktsregel liefert somit bei der gleichen Anzahl von Stützstellen ein besseres Ergebnis als die Rechteckregel. □

Es ergibt sich hier ein deutlicher Unterschied in der Genauigkeit und Güte der Verfahren. Darauf werden wir im nächsten Abschnitt des Kapitels noch genauer eingehen. Zunächst betrachten wir die zweite Idee der Approximation eines Integrals.

Idee 1.2 (Approximation des Integranden). *Der Integrand $f = f(x)$ wird durch eine einfachere Funktion $u = u(x)$ approximiert, welche sich in möglichst einfacher Weise exakt integrieren lässt, z.B. einem Polynom.*

Die Idee hier besteht also darin, den komplizierten Integranden f durch eine Funktion u zu ersetzen, die

– f im Intervall $I = [a, b]$ möglichst gut approximiert und

– leicht zu integrieren ist.

Von der Approximation u wird verlangt, dass sie mit f an gewissen Stützstellen übereinstimmt:

$$u(x_i) = f_i := f(x_i) \quad \text{mit} \quad x_i = x_0, x_1, \ldots, x_n \ \text{ in } \ [a, b]\,. \tag{1.15}$$

Die einfachsten Funktionen zur Approximation sind Polynome, die sich zudem sehr einfach integrieren lassen. Insofern sind sie auch die idealen Kandidaten zur Definition einer näherungsweisen Integration. Ein Polynom vom Grad n schreibt man meist in der Form

$$u_n(x) = \sum_{i=0}^{n} c_i x^i. \tag{1.16}$$

Die Konstanten c_i bestimmt man aus den Bedingungen (1.15). Eine Möglichkeit zur praktischen Bestimmung ist, alle Punkte $(x_i, f(x_i))$ in (1.16) einzusetzen. Dies führt auf ein lineares Gleichungssystem für die c_i. Im Anhang A sind verschiedene andere Methoden zur Berechnung der Approximation einer

Funktion mit Hilfe eines Polynoms ausgeführt. Die Berechnung einer Funktion, welche vorgegebene Funktionswerte (1.15) annimmt, nennt man auch **Interpolation**. Zur Herleitung von Integrationsformeln wird im Folgenden die Lagrangesche Darstellung des Interpolationspolynoms genommen.

Bei der **Lagrangeschen Interpolationsformel** wird das Interpolationspolynom in der Form

$$u_n(x) = L_0(x)f_0 + L_1(x)f_1 + \cdots + L_n(x)f_n \tag{1.17}$$

angesetzt, wobei

$$L_i(x) = \prod_{\substack{j=0 \\ j \neq i}}^{n} \frac{x - x_j}{x_i - x_j}$$

die Lagrangeschen Stützpolynome sind. Sie genügen den Bedingungen

$$L_i(x_k) = \delta_{ik} = \begin{cases} 1 & \text{für } i = k\ , \\ 0 & \text{sonst}\ . \end{cases} \tag{1.18}$$

Wie man leicht nachrechnet, sind dann die Bedingungen (1.15) erfüllt. Integrieren wir u_n über $[a, b]$, erhalten wir

$$Q_h \;=\; \int_a^b u_n(x)\,dx \;=\; f_0 \int_a^b L_0(x)\,dx \;+\; \cdots \;+\; f_n \int_a^b L_n(x)\,dx \quad . \tag{1.19}$$

Die Polynome L_i lassen sich einfach integrieren, so dass sich die Quadraturformel zu

$$Q_h = a_0 f_0 + a_1 f_1 + \cdots + a_n f_n = \sum_{i=0}^{n} a_i f_i \tag{1.20}$$

mit den Koeffizienten

$$a_i = \int_a^b L_i(x)\,dx \tag{1.21}$$

ergibt. Man bezeichnet (1.20) als **gewichtete Mittelwertformel** mit den Stützkoordinaten f_i und den Gewichten a_i.

Der Ansatz (1.17), (1.18) wird wie schon erwähnt als die Lagrangesche Darstellung eines Interpolationspolynoms bezeichnet. Im Anhang A ist eine kurze

Übersicht über die Approximation und die Interpolation enthalten. Bei der Interpolation bezeichnet man mit n gerne den Grad des Polynoms, wie es auch im Anhang ausgeführt ist. Dies entspricht bei der numerischen Integration der Anzahl der Intervalle, $n+1$ ist dann die Anzahl der Stützstellen. Die einfachsten Beispiele werden wir uns im Folgenden anschauen und die zugehörige Quadraturformel ausführlich herleiten.

Beispiel 1.5 (Mittelwertsformeln). Im **Falle $\mathbf{n=0}$** ist das Interpolationspolynom eine Konstante. Zur Festlegung dieser Konstanten wird im Intervall $[a,b]$ eine Stützstelle und ein dazugehöriger Stützwert vorgegeben. Diese triviale Interpolation führt somit gerade auf die Approximation des Integrals durch den Flächeninhalt eines Rechtecks, also die Rechteckregel.

Im **Falle $\mathbf{n=1}$** sind die Stützstellen

$$x_i = a + ih\ , \qquad i = 0,1$$

gegeben durch

$$x_0 = a,\ x_1 = b\ .$$

Das Interpolationspolynom ist somit eine Gerade, welche in der Lagrangeschen Darstellung

$$u(x) = -\frac{1}{h}(x-b)\,f_0 + \frac{1}{h}(x-a)\,f_1$$

mit

$$L_0(x) = -\frac{1}{h}(x-b)\ , \qquad L_1(x) = \frac{1}{h}(x-a)$$

lautet. Nach (1.19) benötigt man zur Aufstellung der Mittelwertsformel die Integrale der Stützpolynome:

$$\int_a^b L_0(x)\,dx = -\frac{1}{h}\int_a^b (x-b)\,dx = -\frac{1}{h}\Big[\frac{1}{2}x^2 - bx\Big]_a^b = \frac{b-a}{2}\ ,$$

$$\int_a^b L_1(x)\,dx = \frac{1}{h}\int_a^b (x-a)\,dx = \frac{1}{h}\Big[\frac{1}{2}x^2 - ax\Big]_a^b = \frac{b-a}{2}$$

und erhält damit die gewichtete Mittelwertformel für $n=1$ zu

$$Q_h = \frac{b-a}{2}(f_0+f_1)\ . \tag{1.22}$$

Sie wird auch **Sehnentrapezregel** genannt, weil das Integral durch die Fläche des Trapezes approximiert wird, welches unter der Sehne durch die Punkte $(a,f(a))$ und $(b,f(b))$ liegt (siehe Abbildung 1.5). □

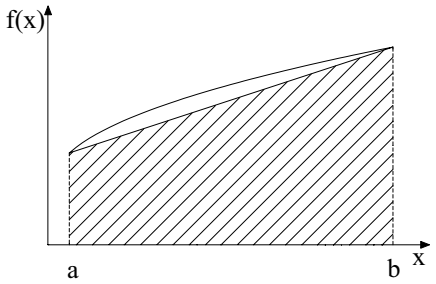


Abb. 1.5. Sehnentrapezregel

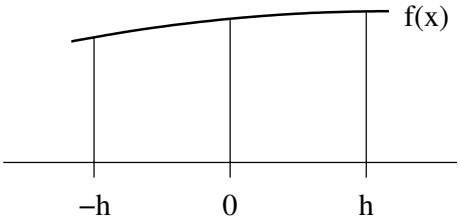
Im **Falle n = 2** werden drei Stützstellen in das Intervall $[a, b]$ eingeführt und die Polynome L_i haben den Grad 2. Man würde dann erwarten, dass die Approximation besser in dem Sinne wird, dass der Fehler

$$R := Q - Q_h$$

betragsmäßig kleiner ist. Wir werden uns im nächsten Kapitel einer Konstruktionsmethode zuwenden, die auch Information über diesen Fehler und somit auch ein Maß der Güte der Approximation liefert.

1.2 Der Taylor-Abgleich

Mit Hilfe der Lagrange-Interpolation erzeugt man Polynome beliebiger Ordnung unter Vorgabe von Werten an den entsprechenden Stützstellen. Wir erwarten, dass sich die Genauigkeit der Integrationsformeln für Polynome höherer Ordnung verbessert. Es müssen dann aber auch mehr Stützstellen vorgegeben und der Integrand an diesen Stützstellen ausgewertet werden. Die Genauigkeit der Mittelwertsformeln werden wir in diesem Abschnitt untersuchen. Zur Illustration greifen wir den Fall $n = 2$ auf. Zur kürzeren Schreib-



weise setzen wir $x_0 = -h,\ x_1 = 0,\ x_2 = h$. Der sogenannte **Taylor-Abgleich** läuft in den folgenden fünf Schritten ab:

Schritt 1: Die Funktion $f = f(x)$ wird in eine Taylor-Reihe um $x = 0$ entwickelt:

$$f(x) = f(0) + x\,f'(0) + \frac{x^2}{2!}f''(0) + \cdots + \frac{x^n}{n!}f^{(n)}(0) + \ldots\,, \tag{1.23}$$

wobei ausreichende Differenzierbarkeitseigenschaften der Funktion f vorausgesetzt werden.

Schritt 2: Mit Hilfe der Taylor-Reihe (1.23) lässt sich der exakte Wert des Integrals in der Form einer Reihe berechnen:

$$Q = \int_{-h}^{h} f(x)dx = \quad 2h\,f_1 + \underbrace{(\frac{h^2}{2} - \frac{h^2}{2})}_{=0} f_1' + \frac{2h^3}{6}f_1'' + \frac{2h^5}{5\cdot 4!}f_1^{(4)} + \ldots \tag{1.24}$$

mit $f_1^{(i)} := f^{(i)}(x_1) = f^{(i)}(0)$ für alle i. Die Integration der Reihe wurde hier elementweise ausgeführt. Alle Terme mit ungeraden Ableitungen fallen dabei wie der oben gekennzeichnete zweite Term auf der rechten Seite heraus.

Schritt 3: Die Funktionswerte an den Stützstellen werden nach der Taylor-Formel (1.23) um den Punkt $x = 0$ entwickelt:

$$\begin{aligned} f_0 &= f(-h) = f_1 - hf_1' + \tfrac{h^2}{2}f_1'' - \tfrac{h^3}{6}f_1''' + \tfrac{h^4}{24}f^{(4)} \pm \ldots\,, \\ f_1 &= \;\; f(0) \;\; = f_1\,, \\ f_2 &= \;\; f(h) \;\; = f_1 + hf_1' + \tfrac{h^2}{2}f_1'' + \tfrac{h^3}{6}f_1''' + \tfrac{h^4}{24}f^{(4)} + \ldots\,. \end{aligned}$$

Schritt 4: Die Mittelwertformel für $n = 2$ lautet allgemein

$$Q_h = a_0 f_0 + a_1 f_1 + a_2 f_2\,. \tag{1.25}$$

Durch Einsetzen der f_i aus dem dritten Schritt ergibt sich

$$\begin{aligned} Q_h &= a_0 f_1 - a_0 h f_1' + a_0\frac{h^2}{2}f_1'' \pm \cdots + a_1 f_1 + a_2 f_1 + h a_2 f_1' + \frac{h^2}{2}a_2 f_1'' + \ldots \\ &= (a_0 + a_1 + a_2)f_1 + (a_2 - a_0)hf_1' + (a_2 + a_0)\frac{h^2}{2}f_1'' + (a_2 - a_0)\frac{h^3}{6}f_1''' + \ldots \end{aligned}$$

Schritt 5: Es wird verlangt, dass die Taylor-Entwicklung der Näherung Q_h mit der des exakten Wertes Q an möglichst vielen der führenden Glieder übereinstimmt. Da wir drei Koeffizienten haben, können wir nur die Erfüllung von drei Gleichungen fordern.

Der Koeffizientenvergleich mit (1.24) liefert:

$$\begin{array}{lll} a_2 & +a_1 & +a_0 = 2h\ , \\ a_2 & & -a_0 = 0\ , \\ a_2 & & +a_0 = \frac{2h}{3}\ . \end{array}$$

Die Lösung dieses Gleichungssystems ist

$$a_0 = a_2 = \frac{h}{3}\ , \qquad a_1 = \frac{4h}{3}\ ,$$

und die gesuchte Integrationsformel für $\mathbf{n = 2}$ lautet somit

$$Q_h = \frac{h}{3}(f_0 + 4f_1 + f_2)\ . \tag{1.26}$$

Sie ist unter dem Namen **Simpson-Formel** bekannt.

Es zeigt sich, dass der nächste Koeffizientenvergleich auf die Gleichung

$$a_2 - a_0 = h/3$$

führt, welche automatisch erfüllt ist. Für den Fehler $Q - Q_h$ ergibt sich

$$Q - Q_h = -\frac{h^5}{90} f_i^{(4)} - \frac{h^7}{1890} f_i^{(6)} - \dots$$

oder nach dem Zwischenwertsatz der Differenzialrechnung

$$R := Q - Q_h = -\frac{h^5}{90} f^{(4)}(\xi) \tag{1.27}$$

mit einer Zwischenstelle $\xi \in (-h, h)$. R wird üblicherweise **Restglied** genannt. Bei dieser Ableitung muss natürlich vorausgesetzt werden, dass der Integrand f mindestens viermal stetig differenzierbar ist.

Beispiel 1.6 (Simpson-Regel, mit MAPLE-Worksheet). Greifen wir als Beispiel das Integral (1.5) vom letzten Abschnitt auf:

$$Q = \int_0^3 (x^3 - 2x)\ dx\ .$$

Mit den Stützstellen $x_0 = 0,\ x_1 = 1.5,\ x_2 = 3$ ergibt sich nach der Simpson-Regel

$$Q_h = \frac{1.5}{3}(0 + 4 \cdot (-0.75) + 21) = 11.25$$

und stimmt genau mit dem exakten Wert überein. Ist dies Zufall? Natürlich nicht! Ein Blick auf das Restglied in Gleichung (1.27) zeigt, dass dort die vierte Ableitung des Integranden steht. Diese ist bei dem Polynom dritten Grades Null und somit liefert die Näherungsformel den exakten Wert. □

Wie in Beispiel 1.6 diskutiert, liefert die Simpson-Regel für jedes Polynom vom Grade kleiner oder gleich drei den exakten Wert. Dies kann man als ein Maß der Güte der Quadraturformel festhalten:

Die **Simpson-Regel** integriert Polynome bis Grad 3 exakt. Man sagt: Sie hat den **Exaktheitsgrad 3**.

Wie hier im Fall $n = 2$ kann man den Taylor-Abgleich für andere Werte von n ausführen. Für $n = 1$ ergibt sich entsprechend die Sehnentrapezregel. Diese Quadraturformel integriert Polynome bis zum Grad 1 exakt. Der **Exaktheitsgrad der Sehnentrapezregel ist somit 1**.

Die Tabelle 1.1 enthält unter der Rubrik **abgeschlossene Newton-Cotes-Formeln** solche Quadraturformeln für $k = 2, 3, 4$ und 5 Stützstellen bzw. $n = 1, 2, 3, 4$ Streifen (Intervalle) der Breite h. Weiter sind der Exaktheitsgrad (E), der Namen, unter dem sie in der Literatur bekannt sind, und der Betrag des Restglieds aufgeführt. **Abgeschlossen** heißen sie, weil die äußeren Stützstellen mit den Intervallgrenzen zusammenfallen: $x_0 = a$, $x_n = b$; ihre Streifenzahl ist demnach n und die Streifenbreite $h = (b - a)/n$, siehe Abbildung 1.6. Das zu Grunde liegende Interpolationspolynom hat den Grad n.

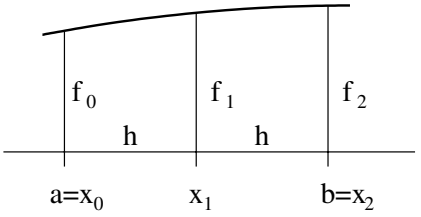


Abb. 1.6. Abgeschlossene Newton-Cotes-Formeln

Offene Newton-Cotes-Formeln entstehen, wenn die Intervallgrenzen nicht als Stützstellen benutzt werden: Bleiben wir bei der alten Bezeichnung, dann ist die erste Stützstelle $x_1 = a + h$ und die letzte $x_{n-1} = b - h$. Für diese

Integrationsformel ist die Anzahl der Streifen n mit der Breite $h = (b-a)/n$, aber die Anzahl der Stützstellen beträgt $k = n-1$. Diese Situation ist in Abbildung 1.7 skizziert.

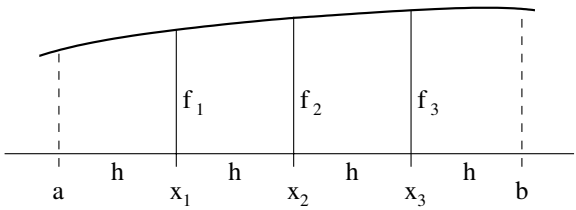


Abb. 1.7. Offene Newton-Cotes-Formeln

Bei der dritten Formelgruppe, den **MacLaurin-Formeln**, liegen die Stützstellen in der Mitte der Streifen, so dass die Anzahl der Streifen n mit der Breite $h = (b-a)/n$ gleich der Anzahl der verwendeten Stützstellen $x_{1/2} = a+h/2$, ..., $x_{n-1/2} = b-h/2$ ist. Dies ist in Abbildung 1.8 dargestellt. Dabei haben wir die alte Bezeichnung wieder beibehalten. Die Teilintervalle

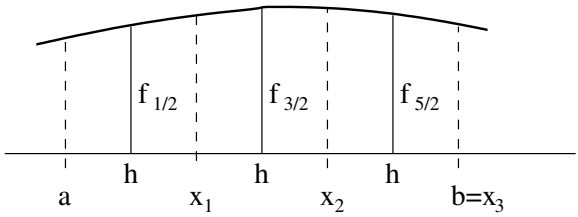


Abb. 1.8. MacLaurin-Formeln

sind durch $[x_{i-1},\, x_i]$ für $i = 1, \ldots,\, n$ gegeben.

Zum Exaktheitsgrad all dieser Formeln lässt sich feststellen:

Eine Newton-Cotes-Formel mit k äquidistanten Stützstellen hat
- den **Exaktheitsgrad** k, falls k **ungerade**
- den **Exaktheitsgrad** $k-1$, falls k **gerade** ist,

das heißt, Polynome bis zum k-ten bzw. $(k-1)$-ten Grad werden exakt integriert; für Polynome höheren Grades und anderen Funktionen ist das Restglied ein Maß für den Fehler.

Bemerkung: In praktischen Rechnungen zeigt es sich, dass Newton-Cotes-Formeln mit Polynomen hohen Grades zur Instabilität neigen. Ab Polynom-

Familie	E	k	n	$I = \int_a^b f(x)\,dx$	*Name*	$\lvert R\rvert$
Abgeschlossene	1	2	1	$I = \frac{h}{2}(f_0 + f_1)$	*'Sehnentrapez'*	$\frac{1}{12}h^3 f''(\xi)$
Newton-Cotes-	3	3	2	$I = \frac{h}{3}(f_0 + 4f_1 + f_2)$	*'Simpson'*	$\frac{1}{90}h^5 f^{(4)}(\xi)$
Formeln	3	4	3	$I = \frac{3h}{8}(f_0 + 3f_1 + 3f_2 + f_3)$	*'3/8'*	$\frac{3}{80}h^5 f^{(4)}(\xi)$
	5	5	4	$I = \frac{2h}{45}(7f_0 + 32f_1 + 12f_2 + 32f_3 + 7f_4)$	*'Milne'*	$\frac{8}{945}h^7 f^{(6)}(\xi)$
Offene	1	1	2	$I = 2hf_1$	*'Rechteck'*	$\frac{1}{3}h^3 f''(\xi)$
Newton-Cotes-	1	2	3	$I = \frac{3h}{2}(f_1 + f_2)$		$\frac{3}{4}h^3 f''(\xi)$
Formeln	3	3	4	$I = \frac{4h}{3}(2f_1 - f_2 + 2f_3)$		$\frac{14}{45}h^5 f^{(4)}(\xi)$
	3	4	5	$I = \frac{5h}{24}(11f_1 + f_2 + f_3 + 11f_4)$		$\frac{95}{144}h^5 f^{(4)}(\xi)$
MacLaurin-	1	1	1	$I = hf_{1/2}$		$\frac{1}{24}h^3 f''(\xi)$
Formeln	1	2	2	$I = h(f_{1/2} + f_{3/2})$		$\frac{1}{12}h^3 f''(\xi)$
	3	3	3	$I = \frac{3h}{8}(3f_{1/2} + 2f_{3/2} + 3f_{5/2})$		$\frac{21}{640}h^5 f^{(4)}(\xi)$
	3	4	4	$I = \frac{h}{12}(13f_{1/2} + 11f_{3/2} + 11f_{5/2} + 13f_{7/2})$		$\frac{103}{1440}h^5 f^{(4)}(\xi)$

Tabelle 1.1. Quadraturformeln für k äquidistante Stützstellen mit n-Streifen und Exaktheitsgrad E

grad ≥ 8 treten negative Gewichte auf. Die zugehörigen Integrationsformeln sind zur numerischen Integration nicht mehr geeignet.

1.3 Summierte Mittelwertformeln

Wir haben gesehen, dass die erste Idee - das Integral durch eine einfache Summe, z.B. der Rechteckregel, zu ersetzen - im Allgemeinen keine gute Näherung liefert. Die Erhöhung der Zahl der Stützstellen, um die Näherung zu verbessern, kostet viele Funktionsauswertungen und damit viel Rechenzeit. Die zweite Idee ist ebenso nicht ideal. Wenn die Genauigkeit erhöht werden soll, dann müssen Polynome immer höheren Grades zur Approximation des Integranden benutzt werden. Da die Polynome hohen Grades auf äquidistanten Gittern zu Oszillationen neigen, sind die zugehörigen Newton-Cotes-Formeln für praktische Berechnungen nicht mehr geeignet. Beide Wege sind somit ungünstig. Die **Kombination beider Ideen** liefert aber brauchbare Verfahren.

Wir zerlegen dazu das gesamte Integral in eine Summe von Teilintegralen

$$\int_a^b = \int_a^{\alpha_1} + \int_{\alpha_1}^{\alpha_2} + \cdots + \int_{\alpha_k}^b$$

und integrieren die Teilintegrale mit einer Newton-Cotes-Formel. Man nennt diese Formeln dann **summierte** oder **zusammengesetzte Mittelwertformeln**.

Beispiel 1.7 (Summierte Simpson-Formel). Als Beispiel nehmen wir eine Zerlegung des Integrationsintervalls in drei Teilgebiete $[a, \alpha_1]$, $[\alpha_1, \alpha_2]$ und $[\alpha_2, b]$ und die Simpson-Formel zur Approximation innerhalb dieser Teilgebiete. Für die einfache Simpson-Formel benötigt man zwei Streifen und drei Stützstellen. Die Gesamtanzahl der Streifen bei der Aufspaltung in drei Teilintegrale ist somit $3 \cdot 2 = 6$ und die Anzahl der Stützstellen beträgt 7. Ein Diagramm für die aufsummierte Simpson-Regel ist in Abbildung 1.9 gezeichnet. Die Simpson-Formeln für die verschiedenen Teilgebiete schreiben wir im Folgenden untereinander und summieren sie auf:

$$\begin{array}{rlllllllll}
 & Q_{hI} & = \frac{h}{3}\,(f_0 + & 4f_1 + & f_2) & & & & & \\
+ & Q_{hII} & = \frac{h}{3}\,(& & f_2 + & 4f_3 + & f_4) & & & \\
+ & Q_{hIII} & = \frac{h}{3}\,(& & & & f_4 + & 4f_5 + & f_6) \\
\hline
 & Q_h & = \frac{h}{3}\,(f_0 + & 4f_1 + & 2f_2 + & 4f_3 + & 2f_4 + & 4f_5 + & f_6)\,.
\end{array}$$

Die Gewichte an den Stützstellen, welche mit dem Rand der Teilgebiete zusammenfallen, addieren sich. □

Allgemein ergibt sich die **zusammengesetzte Simpson-Formel** zu

$$Q_h = \frac{h}{3}(f_0 + 4f_1 + 2f_2 + 4f_3 + \cdots + f_n)\,. \tag{1.28}$$

In Abbildung 1.9 ist diese aufsummierte Formel dargestellt. Jedes Teilintervall besteht aus zwei Streifen und drei Stützstellen. Dabei gehören die Stützstellen am Rand der Teilintervalle auch zu dem Nachbarintervall. Die Gesamtanzahl der Stützstellen der aufsummierten Simpson-Regel ist somit Streifenzahl mal Anzahl der Teilintervalle plus Eins.

Durch Aufsummieren der Sehnentrapezregel erhalten wir die **zusammengesetzte Sehnentrapezregel:**

$$Q_h = h\left(\frac{1}{2}f_0 + f_1 + f_2 + \cdots + f_{n-1} + \frac{1}{2}f_n\right)\,. \tag{1.29}$$

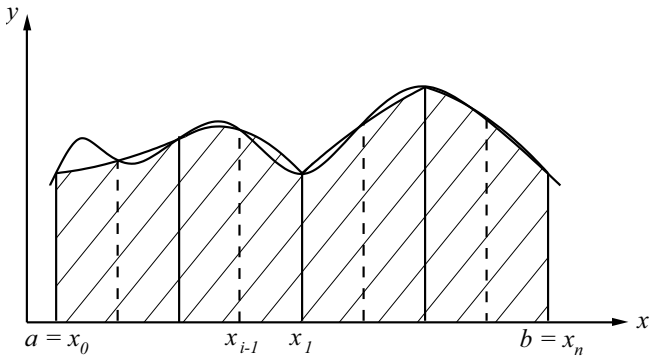


Abb. 1.9. Aufsummierte Simpson-Regel

In Abbildung 1.10 ist diese aufsummierte Formel dargestellt. Die Gesamtanzahl der Stützstellen der aufsummierten Sehnentrapezformel ist somit Anzahl der Teilintervalle plus Eins.

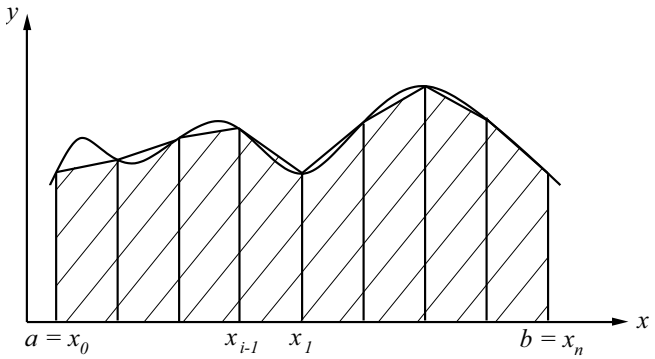


Abb. 1.10. Aufsummierte Sehnentrapezregel

Der Exaktheitsgrad bleibt bei der Summierung erhalten. Für die über s Teilgebiete aufsummierte Simpson-Regel erhalten wir zum Beispiel

$$R = Q - Q_h = \sum_{j=1}^{s} \frac{1}{90} h^5 f^{(4)}(\xi_j) \,, \tag{1.30}$$

wobei die ξ_j die Zwischenstellen in den einzelnen Teilgebieten bezeichnen. Wir können den Betrag dieses Ausdrucks nach oben abschätzen, wenn wir

zum Maximalwert der vierten Ableitung in $[a, b]$ übergehen und erhalten

$$\begin{aligned} |Q - Q_h| &\leq \frac{1}{90} h^5 \, s \;\; \max_{\xi \in [a,b]} \left| f^{(4)}(\xi) \right| \\ &= \frac{1}{90} h^5 \, \frac{n}{2} \;\; \max_{\xi \in [a,b]} \left| f^{(4)}(\xi) \right| \\ &= \frac{b-a}{180} h^4 \, \max_{\xi \in [a,b]} \left| f^{(4)}(\xi) \right| \; . \end{aligned} \tag{1.31}$$

Hier nutzten wir aus, dass die Simpson-Formel zwei Streifen benutzt. Die Gesamtanzahl n von Streifen ist somit gerade $2s$. Der Wert n wird dann durch $(b-a)/h$ ersetzt, wobei sich die Schrittweite herauskürzt und sich die Potenz von h um 1 erniedrigt.

Diese Fehlerabschätzung (1.31) enthält nur noch die Schrittweite h als Diskretisierungsparameter und kann nun folgendermaßen interpretiert werden: Wird die Anzahl der Teilgebiete erhöht, das heißt die Anzahl der Stützstellen erhöht oder die Schrittweite verkleinert, dann geht der Fehler mit der vierten Potenz der Schrittweite gegen Null. Dies kann man dann als ein Qualitätsmaß für die aufsummierte Simpson-Formel interpretieren.

Man definiert ganz allgemein:

Ein numerisches Integrationsverfahren hat die **Konsistenz- oder Genauigkeitsordnung k**, falls

$$|Q - Q_h| = \mathcal{O}(h^k) \;\; \text{für} \;\; n \to \infty \tag{1.32}$$

bei hinreichend glatten Integranden gilt.

Bemerkungen:

1. Die aufsummierte Simpson-Formel hat somit die Genauigkeitsordnung 4, ebenso die 3/8-Regel. Aus der Tabelle 1.1 erhält man weiter die Genauigkeitsordnung der aufsummierten Trapezregel zu 2, die der Milne-Regel zu 6.

2. Die Rechteckregel mit der Auswertung des Integranden links oder rechts des Teilintervalls kann als aufsummierte Formel interpretiert werden. Sie besitzt die Genauigkeit 1.

3. Die Mittelpunktsregel ist die einfachste aufsummierte MacLaurin-Formel und besitzt die Genauigkeitsordnung 2.

Klar ist, dass je größer die Genauigkeitsordnung k eines Verfahrens ist, desto schneller wird der Fehler kleiner, wenn h gegen Null strebt. Um den Fehler für ein vorgegebenes h vor der Rechnung zu bestimmen, müssten wir das Restglied nach der Formel (1.31) bestimmen. Das Abschätzen der Ableitungen hoher Ordnung wird im Allgemeinen sehr aufwändig und damit für die Praxis nicht geeignet sein. Die Genauigkeitsordnung sagt für ein bestimmtes h nicht, wie groß der Fehler ist, sondern dass bei Verkleinerung der Schrittweite die Ergebnisse proportional zu h^k besser werden. Man kann also annehmen, dass ein Verfahren mit hoher Ordnung bei Erhöhung der Anzahl der Stützstellen sehr schnell gute Ergebnisse liefert.

Beispiel 1.8 (Aufsummierte Sehnentrapezformel, mit MAPLE-Worksheet). Zur Berechnung des Integrals (1.5) wird im Folgenden die aufsummierte Sehnentrapezregel ausgeführt. Es ergibt sich für den relativen Fehler bei Rechnung mit verschiedenen Schrittweiten das folgende Bild:

Anzahl der Stützstellen	Wert	relativer Fehler
10	11.50	2.22
20	11.31	0.50
40	11.26	0.12

Die Genauigkeitsordnung 2 sagt aus, dass der Fehler für kleine Schrittweiten wie h^2 klein werden sollte. Wenn wir auf die obige Tabelle schauen, dann zeigt es sich, dass bei einer Verdopplung der Anzahl der Stützstellen, d.h. einer Halbierung der Schrittweite, der Fehler etwa um den Faktor 4 kleiner wird. Wenn wir diese Ergebnisse mit denen für die Rechteckregel in Beispiel 1.3 vergleichen, dann sehen wir, dass die Genauigkeit der Näherung bei der Trapezregel für die gleiche Anzahl von Stützstellen deutlich besser ist. Allerdings liefert für dieses Beispiel die Mittelpunktsregel bei 10 Stützstellen einen Fehler von 0.90. Aufsummierte Sehnentrapez- und Mittelpunktsregel sind beides Verfahren mit der Konsistenzordnung 2. Der Vorfaktor vor dem h^2 ist problemabhängig, so dass für verschiedene Probleme und Schrittweiten sich bei Verfahren mit gleicher Ordnung durchaus unterschiedliche Fehler ergeben. □

Beispiel 1.9 (Vergleich von Verfahren unterschiedlicher Ordnung, mit MAPLE-Worksheet). In dem Worksheet wird die Rechteckregel mit der Trapezregel und der Simpson-Regel verglichen. Wir gehen zu einem anderen bestimmten Integral über, da das Polynom 3. Grades von der Simpson-Formel und damit auch von der aufsummierten Simpson-Formel exakt gelöst wird. Das Integral, welches berechnet wird, lautet

$$Q = \int_1^2 \frac{1}{x} dx = \left[\ln x \right]_1^2 = 0.693147 \, .$$

Dabei wird in den Rechnungen mit allen drei Verfahren die Anzahl der Stützstellen sukzessive um den Faktor 10 erhöht. Es ergeben sich die folgenden Ergebnisse:

	Fehler		
Stützstellen	Rechteckregel	Sehnentrapezregel	Simpson-Regel
10	$0.256 \cdot 10^{-1}$	$0.624 \cdot 10^{-3}$	$0.305 \cdot 10^{-5}$
100	$0.251 \cdot 10^{-2}$	$0.625 \cdot 10^{-5}$	$0.311 \cdot 10^{-9}$
1000	$0.250 \cdot 10^{-3}$	$0.625 \cdot 10^{-7}$	$0.100 \cdot 10^{-11}$
10000	$0.250 \cdot 10^{-4}$	$0.625 \cdot 10^{-9}$	$0.100 \cdot 10^{-11}$
100000	$0.250 \cdot 10^{-5}$	$0.600 \cdot 10^{-11}$	$0.100 \cdot 10^{-11}$

Man erkennt in dieser Tabelle sehr gut, dass der Fehler der Rechteckregel proportional h, der Fehler der Sehnentrapezregel proportional h^2 und der Fehler der Simpson-Regel proportional h^4 ist, wie es die Genauigkeitsordnung der Verfahren aussagt. Bei der hinteren Spalte ist dies allerdings nur bis zu einer gewissen Anzahl von Stützstellen, dann bleibt der Fehler auf dem Wert $0.1 \cdot 10^{-12}$ stehen. Dies liegt an der eingestellten Rechengenauigkeit. Im Worksheet steht in der zweiten Zeile die Angabe `Digits:=12`. Dies bedeutet, dass die Rechnung mit 12-stelligen Zahlen ausgeführt wird. Die Genauigkeit, welche wir durch Verfeinerung der Schrittweiten hier erhalten, ist durch diese Rechengenauigkeit beschränkt. In diesem Worksheet kann man zum Test die Genauigkeit erhöhen. Rechnet man mit zusätzlichen Stellen, dann verbessert sich die Genauigkeit der Simpson-Regel wieder. Man sollte etwas vorsichtig sein, da die Rechenzeit kräftig in die Höhe geht.

Für praktische Rechnungen mit einem Programm, welches in einer Programmiersprache erstellt wurde und nicht im Rahmen eines Computer-Algebra-Systems abläuft, hat man diesen Effekt auf Grund von Rundungsfehler schon früher. Es wird hier im Allgemeinen sogar so sein, dass bei Verkleinerung der Schrittweite der Fehler anwächst. Im Rahmen von MAPLE wird exakt gerechnet und dann erst zum Schluss gerundet, so dass die Rundungsfehler hier keine Rolle spielen. □

1.4 Die Gaußschen Integrationsformeln

Bei den Newton-Cotes-Formeln wurden die Stützstellen äquidistant verteilt und dann die Gewichte a_i so bestimmt, dass eine möglichst hohe Genauig-

keitsordnung erzielt wird. Dies führt auf eine Mittelwertsformel in der folgenden Form:

$$Q_h = \sum_{i=0}^{n} a_i f(x_i), \qquad x_i = a + ih, \qquad i = 0, 1, \ldots, n \; .$$

Es stellt sich die Frage, ob die Stützstellen nicht besser gewählt werden können. Die Antwort lautet: Ja, die zugehörigen Quadraturformeln sind die sogenannten **Gauß-Formeln**.

Der Ansatz bei der Gauß-Quadratur ist die allgemeine Mittelwertsformel

$$Q_h = \sum_{i=0}^{n} a_i f(x_i) \; . \tag{1.33}$$

Die Aufgabe ist: Bestimme sowohl a_i als auch x_i so, dass die Integrationsformel für ein gegebenes n optimal ist. Der Taylor-Abgleich als Weg, eine möglichst hohe Fehlerordnung zu erhalten, wird für nicht-äquidistante Stützstellen sehr aufwändig. Wir werden deshalb im Folgenden den Exaktheitsgrad als Kriterium nehmen.

Es wird gefordert, dass die Integrationsformel (1.33) Polynome

$$p_k(x) = \sum_{j=0}^{k} c_j x^j$$

bis zu einem möglichst großen Grad k exakt integriert:

$$Q_h = \sum_{i=0}^{n} a_i f_i = \int_a^b \sum_{j=0}^{k} c_j x^j \, dx = Q \; . \tag{1.34}$$

Der Näherungswert, wie er in (1.33) berechnet ist, muss mit dem Integral des Polynoms übereinstimmen. Die Lage der Stützstellen hängt natürlich vom Intervall $[a, b]$ ab. Wir setzen hier $a = -1$ und $b = 1$. Beliebige Intervalle können dann durch eine Transformation auf $[-1, 1]$ abgebildet werden. Darauf werden wir anschließend zurückkommen.

Nützt man in Gleichung (1.34) aus, dass die Funktion f ein Polynom ist und somit $f_i = p_k(x_i)$ gilt, dann ergibt sich

$$\sum_{i=0}^{n} a_i \sum_{j=0}^{k} c_j x_i^j = \sum_{j=0}^{k} c_j \int_{-1}^{1} x^j \, dx \; .$$

Integration des Polynoms und Umordnen liefern weiter

$$\sum_{j=0}^{k} c_j \sum_{i=0}^{n} a_i x_i^j = \sum_{j=0}^{k} c_j \left[\frac{x^{j+1}}{j+1}\right]_{-1}^{1} . \tag{1.35}$$

Wir fordern, dass diese Integrationsformel für alle Polynome bis zum Grad k exakt ist.

Es wird nun ein Koeffizientenvergleich durchgeführt, indem die Koeffizienten vor den Monomen x^k gleichgesetzt werden. Dann zerlegt sich (1.35) in die Gleichungen

$$\sum_{i=0}^{n} a_i x_i^j = \left[\frac{x^{j+1}}{j+1}\right]_{-1}^{1} \quad \text{für } j = 0, 1, \ldots, k \tag{1.36}$$

oder als Gleichungssystem ausgeschrieben

$$\begin{aligned}
j = 0 \;:\; & \sum_{i=0}^{n} a_i = \Big[x\Big]_{-1}^{1} , \\
j = 1 \;:\; & \sum_{i=0}^{n} a_i x_i = \left[\frac{x^2}{2}\right]_{-1}^{1} , \\
j = 2 \;:\; & \sum_{i=0}^{n} a_i x_i^2 = \left[\frac{x^3}{3}\right]_{-1}^{1} , \\
& \vdots \\
j = k \;:\; & \sum_{i=0}^{n} a_i x_i^k = \left[\frac{x^{k+1}}{k+1}\right]_{-1}^{1} .
\end{aligned}$$

Die obigen Gleichungen bilden ein nichtlineares Gleichungssystem, welches nur für den Fall $n = 1$ und $n = 2$ geschlossen lösbar ist, wenn gewisse Symmetrieeigenschaften angenommen werden.

Zunächst der **Fall n = 1:** Wir gehen von einem symmetrischen Aufbau der Formel aus: $a_0 = a_1 =: a$, $x_1 = -x_0$. Dann ergibt sich aus den ersten drei Gleichungen

$$\begin{aligned}
j = 0 \;:\; & a + a = 2 && \Rightarrow a = 1 , \\
j = 1 \;:\; & x_0 + x_1 = 0 && \Rightarrow \text{automatisch erfüllt} , \\
j = 2 \;:\; & x_0^2 + x_1^2 = \frac{2}{3} && \Rightarrow 2x_1^2 = \frac{2}{3} .
\end{aligned}$$

Damit ergeben sich die Stützstellen im Intervall $[-1, 1]$ zu

$$x_0 = -\frac{1}{\sqrt{3}}, \quad x_1 = \frac{1}{\sqrt{3}}$$

und die **Gaußsche Integrationsformel für n = 1** zu

$$Q_1 = f\left(-\frac{1}{\sqrt{3}}\right) + f\left(\frac{1}{\sqrt{3}}\right) . \tag{1.37}$$

Im Gegensatz zur Trapezregel, bei der ebenso zwei Stützstellen benutzt werden, integriert die Gauß-Formel nicht nur lineare Polynome, sondern auch kubische Polynome exakt.

Fall n = 2: Wählen wir wiederum einen symmetrischen Aufbau mit $x_2 = -x_0,\ x_1 = 0,\ a_2 = a_0$, ergeben sich die Gleichungen

$$\begin{aligned}
j = 0 &: a_0 + a_1 + a_0 = 2 && \Rightarrow 2a_0 + a_1 = 2 ,\\
j = 1 &: a_0x_0 + a_1x_1 - a_0x_0 = 0 && \curvearrowright \text{automatisch erfüllt} ,\\
j = 2 &: a_0x_0^2 + a_1x_1^2 + a_0x_0^2 = \frac{2}{3} && \Rightarrow 2a_0x_0^2 = \frac{2}{3} ,\\
j = 3 &: a_0x_0^3 + a_1x_1^3 - a_0x_0^3 = 0 && \curvearrowright \text{automatisch erfüllt} ,\\
j = 4 &: a_0x_0^4 + a_1x_1^4 + a_0x_0^4 = \frac{2}{5} && \Rightarrow 2a_0x_0^4 = \frac{2}{5} ,\\
j = 5 &: a_0x_0^5 + a_1x_1^5 - a_0x_0^5 = 0 && \curvearrowright \text{automatisch erfüllt} .
\end{aligned}$$

Dividiert man die dritte Gleichung ($j = 4$) durch die zweite Gleichung ($j = 2$), ergibt sich in der rechten Spalte im obigen Schema die Gleichung

$$\frac{2a_0x_0^4}{2a_0x_0^2} = \frac{2}{5}\cdot\frac{3}{2} \qquad \text{und daraus} \qquad x_0^2 = \frac{3}{5} .$$

Mit Hilfe der anderen Gleichungen ergibt sich weiter $a_0 = \frac{5}{9}$ und $a_1 = \frac{8}{9}$.

Die **Gaußsche Integrationsformel für n = 2** lautet somit

$$Q_2 = \frac{1}{9}\Big(5f(-\sqrt{\frac{3}{5}}) + 8f(0) + 5f(\sqrt{\frac{3}{5}})\Big) . \tag{1.38}$$

Nicht mehr erfüllt ist die Gleichung für $j = 6$. Die Formel besitzt somit den Exaktheitsgrad 5, da für das Restglied $R \sim f^{(6)}(\xi)$ gilt. Dagegen ist die Simpson-Regel, für welche man ebenso 3 Stützstellen verwendet, von 3. Ordnung genau. Durch die geschickte Wahl der Stützstellen wurde Genauigkeit gewonnen.

n	x_i	a_i	$\lvert R\rvert$
0	0.	2.	$\frac{1}{3}f''(\xi)$
1	$\pm 1/\sqrt{3}$	1.	$\frac{1}{135}f^{(4)}(\xi)$
2	$0., \pm\sqrt{3/5}$	$\frac{8}{9},\ \frac{5}{9}$	$\frac{1}{15750}f^{(6)}(\xi)$
3	$\pm 0.86113631, \pm 0.33998104$	$0.34785485,\ 0.65214515$	$\frac{1}{3472875}f^{(8)}(\xi)$
4	$0., \pm 0.90617985, \pm 0.53846931$	$\frac{128}{225},\ 0.23692689,\ 0.47862867$	$\frac{1}{1237732650}f^{(10)}(\xi)$

Tabelle 1.2. Quadraturformeln nach Gauß

Für mehr Stützstellen, d.h. $n \geq 3$, kann das nichtlineare Gleichungssystem (1.36) nicht mehr exakt gelöst werden. Es zeigt sich, dass die n Stützstellen der Gauß-Formel gerade die Nullstellen des **n-ten Legendre-Polynoms** sind. Die Legendre-Polynome sind Kugelfunktionen und bilden ein Orthogonalsystem. Ihre Nullstellen sind reell und liegen im Intervall $[-1, 1]$. Sie sind für $n > 2$ nur numerisch berechenbar. Die Gauß-Punkte sind bis $n = 4$ im Worksheet **Gauß** berechnet und in Tabelle 1.2 zusammen mit den zugehörigen Gewichten und dem Restglied aufgelistet. Die Gewichte a_i müssen symmetrisch vervollständigt werden.

Mit den Gauß-Formeln können wir bislang nur Integrale mit den Grenzen -1 und 1 lösen. Da sich jedes Intervall auf $[-1, 1]$ transformieren lässt, kann dies verallgemeinert werden. Die gesuchte Transformationsvorschrift

$$\begin{aligned} T:\ [a,b] &\to [-1,1] \\ x &\to z = T(x) \end{aligned}$$

lautet

$$z = \frac{2}{b-a}x - \frac{b+a}{b-a}\ ,$$

woraus sich die Umkehrtransformation zu

$$x = \frac{b-a}{2}z + \frac{b+a}{2}$$

berechnet. Mit

$$dx = \frac{b-a}{2}dz \quad \text{und} \quad g(z) = f(\frac{b-a}{2}z + \frac{b+a}{2})$$

ergibt sich für das Integral die Formel

$$\int_a^b f(x)dx = \int_{-1}^{1} f(\frac{b-a}{2}z + \frac{b+a}{2})\frac{b-a}{2}dz = \frac{b-a}{2}\int_{-1}^{1} g(z)dz\ . \tag{1.39}$$

Wie bei den Newton-Cotes-Formeln wird man die Gaußsche Quadratur im Allgemeinen nicht auf das gesamte Intervall anwenden, sondern das Integral zunächst in Teilintervalle zerlegen. Nach der Formel (1.39) muss die Gauß-Quadratur in jedem Teilbereich ausgeführt werden. Die Näherungswerte der Integrale in den Teilbereichen werden dann entsprechend aufsummiert.

1.5 Adaptivität und Fehlerextrapolation

Wie schon im letzten Abschnitt diskutiert, kann man mit der Abschätzung des Restglieds Fehlerschranken für eine benutzte Formel erhalten. Dazu müssen aber Ableitungen hoher Ordnung berechnet und abgeschätzt werden. Dies ist oft sehr aufwändig. Ist die Abschätzung schlecht, dann nützt die ganze Rechnung wenig. Die Ordnung des Verfahrens sagt etwas über die Güte des Verfahrens aus, wenn h klein wird, aber nichts über die Größe des Fehlers für ein endliches h. Wenn wir Aussagen über den aktuellen Fehler wollen, sind diese Fehlerabschätzungen somit eher von dem Typ „für die Praxis nutzlos“.

Das Dilemma mit den Fehlerabschätzungen ist sogar noch größer, da in der Praxis oft Situationen auftreten, in denen eine gewisse Genauigkeit der Ergebnisse gefordert wird, z.B. die Anforderung, dass der Fehler $< 0.1\%$ ist. Der Fehler sollte somit eigentlich der Steuerparameter des Verfahrens sein. Es stellt sich die wichtige Frage: Wie bekommt man eine Aussage über den aktuellen Fehler und wie beurteilt man das Ergebnis einer Näherungsformel? Neben der Erzeugung von Näherungswerten ist die Beurteilung der Ergebnisse eine sehr wichtige Aufgabe in der numerischen Simulation.

Beispiel 1.10 (Aufsummierte Simpson-Formel, mit MAPLE-Worksheet). Wir berechnen das Integral

$$\int_0^{\frac{\pi}{4}} \cos(2x)\sin^3(2x)\,dx \tag{1.40}$$

mit der aufsummierten Simpson-Formel für 21 und 41 Stützstellen und vergleichen die Ergebnisse miteinander. Es ergibt sich

$$Q_{21} = 0.12500042\,,$$
$$Q_{41} = 0.12500003\,.$$

Wie groß könnte der Fehler sein? Es sieht beim Vergleich der beiden Ergebnisse so aus, als ob die ersten sechs Stellen gesichert sind. Bei einer weiteren Erhöhung der Anzahl der Stützstellen lässt sich dies bestätigen. □

Dieses Beispiel zeigt: Wenn man die Ergebnisse für verschiedene Schrittweiten vergleicht, bekommt man vielleicht einen Hinweis auf den Fehler. Stimmen die Ergebnisse bei Schrittweitenverfeinerung auf sechs Stellen überein, so weist dies darauf hin, dass das Ergebnis in diesen sechs Stellen genau sein sollte. Schauen wir uns dieses Vorgehen noch genauer an und gehen dazu folgendermaßen vor:

Schritt 1: Berechnung mit der Schrittweite h_1. Der Fehler für eine Newton-Cotes-Formel mit der Genauigkeitsordnung k hat die Form

$$R_1 := Q - Q_1 = c\, h_1^k f^{(k)}(\xi_1) =: \alpha(\xi_1) h_1^k$$

mit einer Konstanten c.

Schritt 2: Berechnung mit der Schrittweite h_2. Jetzt lautet der Fehler

$$R_2 := Q - Q_2 = \alpha(\xi_2) h_2^k \ .$$

Nehmen wir an, dass $\alpha(\xi_1) \approx \alpha(\xi_2)$ gilt, und nennen diesen Term α, dann erhalten wir für die Differenz

$$R_1 - R_2 = Q_2 - Q_1 = \alpha \left(h_1^k - h_2^k \right) \ .$$

Diese Gleichung können wir nach α auflösen:

$$\alpha = \frac{Q_2 - Q_1}{h_1^k - h_2^k} \ . \tag{1.41}$$

Damit haben wir den unbekannten exakten Wert Q des Integrals eliminiert. Oben eingesetzt erhalten wir die **Fehlerabschätzungen**

$$R_1 \approx \alpha\, h_1^k, \quad R_2 \approx \alpha\, h_2^k \ . \tag{1.42}$$

Zusätzlich kann man auch eine **Fehlerkorrektur** ausführen

$$Q_2^* = Q_2 - R_2 = \frac{(\frac{h_1}{h_2})^k Q_2 - Q_1}{(\frac{h_1}{h_2})^k - 1} \tag{1.43}$$

und erhält einen verbesserten Näherungswert Q_2^*.

Man nennt dieses Vorgehen **Fehler- oder Richardson-Extrapolation**. Wir können uns die Bezeichnung „Extrapolation“ folgendermaßen deutlich machen. Wir berechnen Näherungswerte $Q(h_1)$ und $Q(h_2)$ und berechnen danach ein Interpolationspolynom durch die Stützpaare $h_1, Q(h_1)$ und $h_2, Q(h_2)$. Dieses Interpolationspolynom wird dann ausgewertet an der Stelle $h = 0$. Geometrisch sieht dieser Sachverhalt in einem (h, Q)-Schaubild folgendermaßen aus: Den exakten Wert erhält man für $h \to 0$, also verbinden

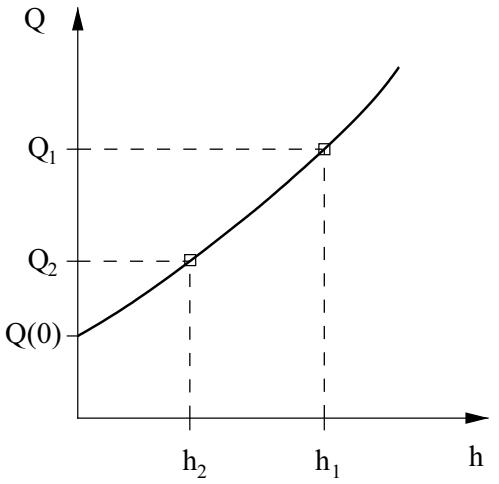


Abb. 1.11. Fehlerextrapolation

wir die Punkte (h_1, Q_1) und (h_2, Q_2) mit einem Interpolationspolynom der Form $a + bh^p$, wobei p die Ordnung des Verfahrens ist. Berechnet man den Wert dieses Polynoms bei $h = 0$, so erhält man gerade (1.43) mit $p = n$. Extrapolation nennt man dies, da der Punkt $h = 0$, an dem man das Interpolationspolynom auswertet, außerhalb des Intervalls der Stützstellen $[h_2, h_1]$ liegt.

Diese Fehlerextrapolation wird sehr effizient, wenn das zugrunde liegende Verfahren eine quadratische **asymptotische Fehlerentwicklung** der Form

$$Q(h) = Q + c_1 h^2 + c_2 h^4 + \cdots + c_m h^{2m} + \mathcal{O}(h^{2m+2})$$

besitzt. Durch eine Fehlerkorrektur kann die Ordnung des Verfahrens gleich um zwei erhöht werden. Die ganze Prozedur wird mehrmals hintereinander angewandt und so Schritt für Schritt die Genauigkeit erhöht.

Im 1. Schritt wird mit den Näherungswerten für $p = 2$ eine Näherung der Ordnung 4 erzeugt. Die Formel für $p = 4$ führt zu einem Wert der Ordnung

6. Diese sukzessive Anwendung des Extrapolationsprinzips auf die aufsummierte Trapezregel heißt **Romberg-Verfahren**. Das Romberg-Schema ist in Tabelle 1.3 dargestellt und verläuft in drei Schritten.

Schrittw.	$(2.O)$	$(4.O)$	$(6.O)$	$(8.O)$	$(10.O)$
h_1	T_{11}				
$h_2 = \frac{1}{2}h_1$	T_{21}	T_{22}			
$h_3 = \frac{1}{2}h_2$	T_{31}	T_{32}	T_{33}		
$h_4 = \frac{1}{2}h_3$	T_{41}	T_{42}	T_{43}	T_{44}	
$h_5 = \frac{1}{2}h_4$	T_{51}	T_{52}	T_{53}	T_{54}	T_{55}

Tabelle 1.3. Romberg-Verfahren

Schritt 1: Man berechnet das gegebene Problem mehrmals mit den sukzessiv halbierten Schrittweiten nach der aufsummierten Sehnentrapezregel. Die Ergebnisse stellen die erste Spalte des Romberg-Schemas dar. $T_{i,1}$ ist der Wert der aufsummierten Trapezregel zur Schrittweite h_i.

Schritt 2: Aus je zwei Näherungswerten $T_{i-1,1}$ und $T_{i,1}$ der ersten Spalte berechnet man nun, wie dies durch die Pfeile angedeutet ist, einen verbesserten Wert entsprechend der Formel

$$T_{i,2} = T_{i,1} + \frac{T_{i,1} - T_{i-1,1}}{\left(\frac{h_{i-1}}{h_i}\right)^2 - 1} = T_{i,1} + \frac{1}{3} \cdot (T_{i,1} - T_{i-1,1}) \,.$$

Da der Term $\alpha_2 h^2$ eliminiert wurde, sind sie jeweils von 4. Ordnung genau.

Schritt 3: Aufbauend auf $T_{i,2}$ werden in gleicher Weise Näherungen

$$T_{i,3} = T_{i,2} + \frac{T_{i,2} - T_{i-1,2}}{(\frac{h_{i-1}}{h_i})^4 - 1} = T_{i,2} + \frac{1}{15}(T_{i,2} - T_{i-1,2})$$

berechnet, die von 6. Ordnung sind.

Allgemein gilt

$$T_{i,j} = T_{i,j-1} + \frac{T_{i,j-1} - T_{i-1,j-1}}{\left(\frac{h_{i-1}}{h_i}\right)^{2(j-1)} - 1}, \tag{1.44}$$

wobei $2j$ die Konsistenzordnung der entsprechenden Spalte bezeichnet.

Beispiel 1.11 (Romberg-Verfahren, mit MAPLE-Worksheet). Dieses Worksheet berechnet das Integral (1.40) mit dem Romberg-Verfahren. Dabei wird die Anzahl der Spalten bzw. die Ordnung des Verfahrens vorgegeben. Bei einer vorgegebenen Ordnung von 14 wird das exakte Ergebnis auf 9 Stellen genau berechnet. Das Schema aus Tabelle 1.3 wird in dem Worksheet exemplarisch berechnet und ausgegeben. In Abbildung 1.12 ist der Algorithmus als Struktogramm dargestellt. □

Bemerkungen:

1. Möchte man im Programm nicht die Ordnung vorgeben, sondern die gewünschte Genauigkeit, dann wird man folgendermaßen vorgehen: Man beginnt mit $T_{1,1}$, $T_{2,1}$. Hier kann man schon prüfen, ob soviel Stellen übereinstimmen, wie die vorgegebene Genauigkeit dies erfordert. Falls ja, ist man fertig. Da es kaum Rechenaufwand kostet, wird man $T_{2,2}$ berechnen und als Ergebnis nehmen. Falls nein, wird im Romberg-Schema die Zeile mit h_3 hinzugenommen. Dann vergleicht man $T_{2,2}$ und $T_{3,2}$. Ist die gewünschte Genauigkeit nicht erreicht, kommt eine neue Zeile dazu bis

   $$|T_{j,J} - T_{j+1,J}| < \varepsilon,$$

   so dass $T_{j+1,j+1}$ als Näherungswert akzeptiert wird. Das Verfahren sucht sich so bei vorgegebener Genauigkeit die Ordnung selbst.

2. Wenn man zu feineren Unterteilungen übergeht, steigt der Rechenaufwand des Romberg-Verfahrens relativ stark an, wie die Zusammenstellung der Stützordinaten zeigt:

Schrittw.	a	b	Funktionsberechnungen	
h_1	*	*	2	(2.O)
h_2	* *	*	+1 = 3	(4.O)
h_3	* * * *	*	+2 = 5	(6.O)
h_4	* * * * * * * *	*	+4 = 9	(8.O)
h_5	* * * * * * * * * * * * * * * *	*	+8 = 17	(10. O)

Romberg

Real: a, b, q, h

Real Function: f

Real Feld $(0..NN, 0..NN)$: T

Integer: i, n, m, N, p

Übernehme (a, b, N)

$T_{1,1} = (b-a)\dfrac{f(a)+f(b)}{2}$

$n = 2$

Solange $n \leq N$

$h = \dfrac{b-a}{2^{n-1}}$

$q = 0$

$i = 1$

Solange $i \leq 2^{n-1} - 1$

$q = q + f(a + ih)$

$i = i + 2$

$T_{n,1} = \dfrac{T_{n-1,1}}{2} + hq$

$p = 4$

$m = 1$

Solange $m \leq n$

$T_{n,m} = T_{n,m-1} + \dfrac{T_{n,m-1} - T_{n-1,m-1}}{p-1}$

$p = 4p$

$m = m + 1$

$n = n + 1$

Gib $T_{N,N}$ zurück.

Abb. 1.12. Das Romberg-Verfahren mit vorgegebener Ordnung N. Dabei muss im aufrufenden Programm sichergestellt werden, dass N nicht größer ist als NN.

Wählt man für die h_i statt der **Romberg-Folge**

$$1, \frac{1}{2}, \frac{1}{4}, \frac{1}{8}, \frac{1}{16}, \frac{1}{32} \ \ldots \text{ entspricht } h_i = \frac{h_0}{2^i},$$

die von **Bulirsch** angegebene:

$$1, \frac{1}{2}, \frac{1}{3}, \frac{1}{4}, \frac{1}{6}, \frac{1}{8} \ \ldots \text{ entspricht } h_i = \frac{h_0}{\nu_i},$$

$$\text{wobei } i \in \{2k-1, 2k\}\,,\ k \in 1, 2, 3 \ldots \text{ und } \nu_i = 2 \cdot 2^{k-1} \text{ falls } i = 2k-1,$$

$$\nu_i = 3 \cdot 2^{k-1} \text{ falls } i = 2k,$$

so vermindert sich der Aufwand für die Funktionsberechnungen, wenn man zu größeren i kommt:

Schrittw.	a	b Funktionsberechnungen	
h_1	*	* 2	(2.O)
h_2	* *	* +1 = 3	(4.O)
h_3	* * *	* +2 = 5	(6.O)
h_4	* * * *	* +2 = 7	(8.O)
h_5	* * * * * * *	+2 = 9	(10.O)

Zu beachten ist, dass die Teiler $(h_{i-1}/h_i)^{n+1} - 1$ für die Bulirsch-Folge von denen der Romberg-Folge abweichen. Die Bulirsch-Folge lässt sich als Romberg-Folge mit zusätzlich zwischengeschalteten Gliedern darstellen:

$$\begin{array}{ccccccc}
\textcircled{1} & \xrightarrow{\cdot\frac{1}{2}} & \textcircled{\frac{1}{2}} & \xrightarrow{\cdot\frac{1}{2}} & \textcircled{\frac{1}{4}} & \xrightarrow{\cdot\frac{1}{2}} & \textcircled{\frac{1}{8}} \cdots \\
 & & \cdot\frac{2}{3}\downarrow & & \cdot\frac{2}{3}\downarrow & & \cdot\frac{2}{3}\downarrow \\
 & & \textcircled{\frac{1}{3}} & \xrightarrow{\cdot\frac{1}{2}} & \textcircled{\frac{1}{6}} & \xrightarrow{\cdot\frac{1}{2}} & \textcircled{\frac{1}{12}} \quad \cdots
\end{array}$$

Vorsicht: Wir setzen hier wieder voraus, dass der Integrand genügend oft stetig differenzierbar ist. Ist dies nicht der Fall, dann lässt sich die entsprechende Ordnung nicht erreichen.

1.6 Numerische Differenziation

Im Gegensatz zur Integration ist die Differenziation ein konstruktiver Vorgang. Ist die Funktion explizit gegeben, dann wird man die Differenziation

exakt entweder per Hand oder mit Hilfe eines Computer-Algebra-Systems ausführen. Ist die Funktion nur an Hand einer Wertetabelle gegeben, dann ist eine numerische Differenziation notwendig: Die Werte werden interpoliert und das Interpolationspolynom wird entsprechend differenziert. Aber auch bei der numerischen Lösung von gewöhnlichen als auch partiellen Differenzialgleichungen mit Hilfe von Differenzenverfahren benötigt man entsprechende Approximationen. Die Idee bei den Differenzenverfahren besteht darin, dass man die kontinuierlichen Ableitungen einer Differenzialgleichung durch Näherungsausdrücke ersetzt. In diesem Abschnitt über die numerische Differenziation werden wir solche Näherungsformeln diskutieren.

Um die Ableitung einer Funktion f an der Stelle x_0 auf einem Rechner numerisch zu berechnen, geht man auf die Definition der Ableitung über den Differenzialquotienten zurück:

$$f'(x_0) = \lim_{h \to 0} \frac{f(x_0 + h) - f(x_0)}{h} \; .$$

Die Ableitung bedeutet geometrisch die Steigung der Tangente im Punkte $f(x_0)$. Die Tangentensteigung erhält man, indem man die Sekante durch die Funktionswerte an den Stellen x_0 und $x_0 + h$ aufstellt, die Sekantensteigung

$$\frac{f(x_0 + h) - f(x_0)}{(x_0 + h) - x_0}$$

bestimmt und den Grenzübergang $h \to 0$ berechnet.

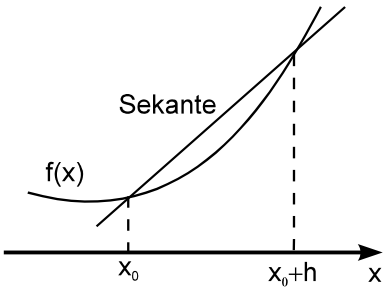


Abb. 1.13. Sekantensteigung

Der Grenzübergang $h \to 0$ kann numerisch nicht durchgeführt werden, da hier Differenzen aus sehr kleinen Zahlen auftreten, die miteinander dividiert

werden und somit Rundungsfehler stark anwachsen. Daher nähert man numerisch die Ableitung einer stetig differenzierbaren Funktion $f = f(x)$ im Punkte x_0 durch die Sekantensteigung

$$D^+f(x_0) = \frac{f(x_0+h) - f(x_0)}{h}$$

mit $h > 0$ an. Dies ist eine **einseitige** oder die **rechtsseitige Differenzenformel**. Man beachte, dass im Gegensatz zu einer analytischen Rechnung numerisch nicht die Ableitung einer Funktion, sondern nur der Wert der Ableitung in einem speziell vorgegebenen Punkt x_0 berechnet wird!

Diese einseitige Differenzenformel hat die folgenden Eigenschaften:

1. Für $h \to 0$ geht der numerische Wert gegen den Wert der exakten Ableitung im Punkt $x = x_0$, wenn Rundungsfehler vernachlässigt werden.

2. Polynome vom Grade $n = 1$ (d.h. Geraden) werden exakt differenziert.

Ganz analog kann man auch die **linksseitige Differenzenformel**

$$D^-f(x_0) = \frac{f(x_0) - f(x_0-h)}{h}$$

mit $h > 0$ definieren.

Eine genauere Differenzenformel erhält man, wenn man den Mittelwert der rechtsseitigen und linksseitigen Differenzenformel nimmt:

$$Df(x_0) = \frac{1}{2}(D^+f(x_0) + D^-f(x_0))$$

$$\Rightarrow \; Df(x) = \frac{1}{2}\frac{f(x_0+h) - f(x_0-h)}{h} \; . \qquad (1.45)$$

Man nennt diese Formel die **zentrale Differenzenformel**.

Mit dieser zentralen Differenzenformel werden Polynome bis zum Grad 2 exakt differenziert, wie wir im Folgenden nachrechnen werden. Dazu gehen wir von einem beliebigen Polynom 2. Grades, $f(x) = a + b\,x + c\,x^2$, aus. Dann gilt

$$\begin{aligned} Df(x) &= \frac{1}{2h}[a + b(x+h) + c(x+h)^2 - a - b(x-h) - c(x-h)^2] \\ &= \frac{1}{2h}[2\,b\,h + 4\,c\,x\,h] = b + 2\,c\,x = f'(x) \; . \end{aligned}$$

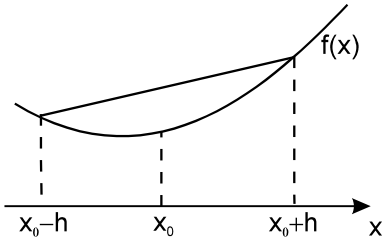


Abb. 1.14. Zentrale Differenzenformel

Beispiel 1.12 (Mit MAPLE-Worksheet). Gesucht ist die Ableitung der Funktion

$$f(x) = \sin x \cdot \ln x \quad \text{an der Stelle} \quad x_0 = \frac{1}{2} \, .$$

Die exakte Ableitung dieser Funktion lautet

$$f'(x) = \cos x \cdot \ln x + \frac{\sin x}{x} \quad \text{und damit} \quad f'(x_0) = 0.3505571 \, .$$

In Tabelle 1.4 sind für unterschiedliche Schrittweiten h die Fehler der numerischen Differenziation betragsmäßig aufgelistet. In der zweiten Spalte steht

	Fehler für einseitige Formel	Fehler für zentrale Differenzen
$h = 10^{-1}$	$8.8 \cdot 10^{-2}$	$8.6 \cdot 10^{-3}$
$h = 10^{-2}$	$9.5 \cdot 10^{-3}$	$8.5 \cdot 10^{-5}$
$h = 10^{-3}$	$9.6 \cdot 10^{-4}$	$8.5 \cdot 10^{-7}$
$h = 10^{-4}$	$9.6 \cdot 10^{-5}$	$8.5 \cdot 10^{-9}$
	$\sim h$	$\sim h^2$

Tabelle 1.4. Verfahrensfehler

die Abweichung der exakten Ableitung zum Wert der einseitigen und in der dritten Spalte zum Wert der zentralen Differenzenformel. □

Man entnimmt Tabelle 1.4 das Fehlerverhalten der beiden Verfahren: Der Fehler bei der einseitigen Differenzenformel ist proportional zu h, während der Fehler bei der zentralen Differenzenformel proportional zu h^2 ist. Dieses Verhalten spiegelt die Ordnung des Verfahrens wider. Die einseitige Differenzenformel ist von 1. Ordnung und die zentrale Differenzenformel von 2. Ordnung genau. Wir werden darauf gleich zurückkommen.

Bemerkung: Die Differenzenformel für die erste Ableitung einer Funktion kann man systematischer gewinnen, indem man durch die Punkte (x_0, f_0), (x_1, f_1), (x_2, f_2) das Interpolationspolynom vom Grade 2, $p_2(x)$, bestimmt, dieses Polynom anschließend ableitet und an der gesuchten Zwischenstelle auswertet. Wir führen diese Vorgehensweise für die Ableitung an der Stelle $x = x_1$ durch. Es ist hier günstig, die Darstellung des **Interpolationspolynoms nach Newton** zu benutzen, welche im Anhang A ausführlich beschrieben ist. Der Ansatz in dieser Darstellung lautet

$$p_2(x) = a_0 + a_1\,(x - x_0) + a_2\,(x - x_0)\,(x - x_1)$$

$$p_2'(x) = a_1 + a_2\,(x - x_1) + a_2\,(x - x_0)$$

$$\Rightarrow p_2'(x_1) = a_1 + a_2\,(x_1 - x_0)\,.$$

Die Koeffizienten a_0, a_1 und a_2 des Interpolationspolynoms werden durch die Stützpaare bestimmt. Nach dem Schema der dividierten Differenzen ergeben sie sich aus der folgenden Tabelle:

x_0	f_0			
		$\searrow$		
x_1	f_1	$\rightarrow \frac{f_1-f_0}{x_1-x_0}$		
		$\searrow$	$\searrow$	
x_2	f_2	$\rightarrow \frac{f_2-f_1}{x_2-x_1}$	$\rightarrow \left(\frac{f_2-f_1}{x_2-x_1} - \frac{f_1-f_0}{x_1-x_0}\right) \Big/ (x_2 - x_0)$	

und lauten

$$a_0 = f_0, \qquad a_1 = \frac{f_1 - f_0}{x_1 - x_0}\,,$$
$$a_2 = \frac{(x_1 - x_0)\,(f_2 - f_1) - (x_2 - x_1)\,(f_1 - f_0)}{(x_2 - x_0)\,(x_2 - x_1)\,(x_1 - x_0)}\,.$$

Setzen wir diese Koeffizienten in $p_2'(x_1)$ ein, ist

$$p_2'(x_1) = \frac{f_1 - f_0}{x_1 - x_0} + \frac{(x_1 - x_0)\,(f_2 - f_1) - (x_2 - x_1)\,(f_1 - f_0)}{(x_2 - x_0)\,(x_2 - x_1)}\,.$$

Speziell für eine äquidistante Unterteilung $h = (x_1 - x_0) = (x_2 - x_1)$ ergibt sich

$$p_2'(x_1) = \frac{f_1 - f_0}{h} + \frac{h\,(f_2 - f_1) - h\,(f_1 - f_0)}{2\,h\,h} = \frac{f_2 - f_0}{2\,h}\,.$$

Dies ist wieder die zentrale Differenzenformel.

Genauere Formeln erhält man, indem das Interpolationspolynom durch mehr als drei Punkte gelegt, abgeleitet und an der gesuchten Stelle ausgewertet

wird. Die Genauigkeit der so bestimmten Differenzenformeln berechnet man mit dem **Taylor-Abgleich**, ganz ähnlich wie bei der Herleitung der Newton-Cotes-Formeln im Rahmen der numerischen Integration. Wir führen diese Methode für den zentralen Differenzenquotienten bei einer äquidistanten Unterteilung exemplarisch vor.

Berechnung der Ordnung der Differenzenformeln: Sei $f = f(x)$ eine 4-mal stetig differenzierbare Funktion, dann gilt nach dem Taylorschen Satz

$$f(x) = f(x_0) + f'(x_0)\,(x - x_0) + \frac{1}{2!} f''(x_0)\,(x - x_0)^2 + \frac{1}{3!} f'''(x_0)\,(x - x_0)^3 + R_3(x)\;.$$

Wir setzen diesen Ausdruck in die zentrale Differenzenformel ein. Dazu bestimmen wir $f(x_0 + h)$ und $f(x_0 - h)$:

$$f(x_0 + h) = f(x_0) + f'(x_0)\,h + \frac{1}{2!} f''(x_0)\,h^2 + \frac{1}{3!} f'''(x_0)\,h^3 + R_3(h)\;,$$
$$f(x_0 - h) = f(x_0) - f'(x_0)\,h + \frac{1}{2!} f''(x_0)\,h^2 - \frac{1}{3!} f'''(x_0)\,h^3 + R_3(-h)\;.$$

Als Differenz erhalten wir

$$f(x_0 + h) - f(x_0 - h) = 2\,h\,f'(x_0) + \frac{1}{3} f'''(x_0)\,h^3 + R_3(h) - R_3(-h)$$

und somit den zentralen Differenzenquotient als

$$\frac{1}{2\,h}\,(f(x_0 + h) - f(x_0 - h)) = f'(x_0) + \mathcal{O}(h^2)\;. \qquad (1.46)$$

Auf der linken Seite steht der zentrale Differenzenquotient und auf der rechten Seite die Ableitung der Funktion plus einem Term $\mathcal{O}(h^2)$ der proportional zu h^2 ist. Bis auf diesen Term $\mathcal{O}(h^2)$ stimmen Ableitung und zentraler Differenzenquotient überein. Man nennt den Exponenten die **Ordnung der Differenzenformel**. Dies spiegelt genau unsere experimentelle Beobachtung aus Tabelle 1.4 wider, dass der zentrale Differenzenquotient von der Ordnung 2 ist.

Obige Aussagen gelten allerdings nur, wenn man die Rundungsfehler vernachlässigt. Denn setzen wir Tabelle 1.4 für kleinere h-Werte fort, so erhält man für eine Rechengenauigkeit von 10 Stellen das in Tabelle 1.5 dargestellte Verhalten. Man erkennt, dass obwohl h sich verkleinert, der Fehler ab einem gewissen h wieder ansteigt. Obwohl der Verfahrensfehler (= *Diskretisierungsfehler*) gegen Null geht, steigt der Gesamtfehler an. Es gilt

h	Fehler für einseitige Formel	Fehler für zentrale Differenzen
10^{-1}	$8.8 \cdot 10^{-2}$	$8.6 \cdot 10^{-3}$
10^{-2}	$9.5 \cdot 10^{-3}$	$8.5 \cdot 10^{-5}$
10^{-3}	$9.6 \cdot 10^{-4}$	$8.5 \cdot 10^{-7}$
10^{-4}	$9.6 \cdot 10^{-5}$	$8.5 \cdot 10^{-9}$
10^{-5}	$9.6 \cdot 10^{-6}$	$1.4 \cdot 10^{-8}$
10^{-6}	$1.1 \cdot 10^{-6}$	$3.0 \cdot 10^{-8}$
10^{-7}	$2.9 \cdot 10^{-6}$	$7.1 \cdot 10^{-7}$
10^{-8}	$7.5 \cdot 10^{-6}$	$1.5 \cdot 10^{-5}$
10^{-9}	$5.0 \cdot 10^{-4}$	$5.3 \cdot 10^{-5}$
10^{-10}	$4.1 \cdot 10^{-3}$	$1.8 \cdot 10^{-3}$
10^{-11}	$1.3 \cdot 10^{-2}$	$9.4 \cdot 10^{-3}$
10^{-12}	$1.0 \cdot 10^{-1}$	$1.2 \cdot 10^{-1}$

Tabelle 1.5. Gesamtfehler

Gesamtfehler = Verfahrensfehler + Rundungsfehler.

Der Einfluss der Rundungsfehler wurde in Abschnitt 0.1 und bei der numerischen Integration schon diskutiert. Der Verfahrens- oder Diskretisierungsfehler ist der Fehler, den man durch die numerische Approximation erhält: Die Ableitung oder Tangentensteigung wird durch die Sekantensteigung mit $h > 0$ ersetzt. Der Rundungsfehler beruht auf der Tatsache, dass bei einer numerischen Rechnung die Zahlen nur näherungsweise dargestellt werden und mit endlicher Genauigkeit gerechnet wird. Der Diskretisierungsfehler geht für $h \to 0$ gegen Null, der *Rundungsfehler* wächst für kleine h wie $\frac{1}{h}$, so dass der Gesamtfehler für sehr kleine h durch den Rundungsfehler bestimmt ist. Diesen Sachverhalt zeigte schon die Abbildung 0.1 im Kapitel 0.1 bei der Diskussion der unterschiedlichen Fehler.

Differenzenformeln für Ableitungen höherer Ordnung erhält man analog zu den Betrachtungen für die erste Ableitung. Das Interpolationspolynom muss entsprechend oft abgeleitet werden. Die Ordnung des Polynoms muss natürlich mindestens der Ordnung der Ableitung entsprechen. Das MAPLE-Worksheet **DiffFormeln** bestimmt zu vorgegebenen Punkten $(t_1, s_1), (t_2, s_2), \ldots, (t_k, s_k)$ Diskretisierungsformeln für die n-te Ableitung. Zur sinnvollen Anwendung der Prozedur sollte $k > n$ gewählt werden! Die Prozedur legt zunächst durch die Punkte das Interpolationspolynom und leitet dieses n-mal ab. Anschließend wird dieses Polynom an der spezifizierten Stelle t_i, $(1 \leq i \leq k)$ ausgewertet.

Bemerkungen:

1. Allgemeine Diskretisierungsformeln für die zweite Ableitung mit höherer Ordnung sowie bei nicht-äquidistanter Unterteilung erhält man, indem durch vorgegebene Punkte $s(t_1),\ s(t_2), \ldots,\ s(t_n)$ das Interpolationspolynom gelegt, dieses zweimal differenziert und anschließend die auszuwertende Stelle eingesetzt wird ($\rightarrow$ analoges Vorgehen wie bei den Differenzenformeln für die erste Ableitung).

2. Der Verfahrensfehler wird wie im Falle der ersten Ableitung durch Taylor-Abgleich berechnet. Dies ist in der Prozedur **DiffFormelnOrdnung** realisiert (siehe auch das zugehörige **Worksheet**).

In der Tabelle 1.6 sind Differenzenformeln für die erste bis vierte Ableitung einer Funktion an der Stelle x_i sowie das Restglied bei einer äquidistanten, **symmetrischen** Anordnung der Stützstellen angegeben. Die oberste Zeile spezifiziert die Indizes k. Dabei kann k kann von $i-3$ bis $i+3$ laufen. Darunter stehen im ersten Abschnitt der Tabelle Differenzenquotienten für die 1. Ableitung, vor der Klammer steht der Vorfaktor, dann kommen die Koeffizienten der Werte an den Stützstellen x_k, hinter der Klammer symbolisch der Wert y_k an der entsprechenden Stützstelle. In der Spalte ganz rechts ist der Fehlerterm angegeben. Der erste so dargestellte Differenzenquotient ist der zentrale Differenzenquotient 2. Ordnung, darunter die zentralen Differenzenquotienten 4. und 6. Ordnung. Es ist klar, dass für die höhere Ordnung immer mehr Stützstellen mit einbezogen werden müssen. In dem nächsten Abschnitt sind dann die Differenzenquotienten für die 2. Ableitung mit der Genauigkeitsordnung 2, 4 und 6 und danach die 3. und 4. Ableitung mit den Genauigkeitsordnungen 2 und 4 angegeben.

Bei einer **unsymmetrischen** Auswahl der Stützstellen ergeben sich die Differenzenformeln in Tabelle 1.7. Die Tabelle ist wie bei den symmetrischen Formeln aufgebaut. Es sind hier die rechtsseitigen Differenzenquotienten ausgewählt. Die linksseitigen erhält man durch Spiegelung bei $k = i$ und Vertauschen der Vorzeichen.

1.7 Bemerkungen und Entscheidungshilfen

Die Wahl einer Näherungsmethode zur Berechnung eines bestimmten Integrals hängt von verschiedenen Faktoren ab. Die Glattheit des Integranden ist zunächst ein Kriterium. Ist die zu integrierende Funktion nur ein- oder zweimal stetig differenzierbar, dann macht ein Verfahren hoher Ordnung keinen Sinn. Höhere Ordnung bedeutet in diesem Fall nicht gleichzeitig höhe-

$k=$	i	-3	-2	-1	0	$+1$	$+2$	$+3$		R
$y_i' =$	$\frac{1}{2h}$ (			-1		$+1$		)	$*y_k -$	$\frac{h^2}{6}\, y'''$
$=$	$\frac{1}{12h}$ (		$+1$	-8		$+8$	-1	)	$+$	$\frac{h^4}{30}\, y^{(5)}$
$=$	$\frac{1}{60h}$ (	-1	$+9$	-45		$+45$	-9	$+1$)	$-$	$\frac{h^6}{140}\, y^{(7)}$
$y_i'' =$	$\frac{1}{h^2}$ (			$+1$	-2	$+1$		)	$*y_k -$	$\frac{h^2}{12}\, y^{(4)}$
$=$	$\frac{1}{12h^2}$ (		-1	$+16$	-30	$+16$	-1	)	$+$	$\frac{h^4}{90}\, y^{(6)}$
$=$	$\frac{1}{180h^2}$ (	$+2$	-27	$+270$	-490	$+270$	-27	$+2$)	$-$	$\frac{h^6}{560}\, y^{(8)}$
$y_i''' =$	$\frac{1}{2h^3}$ (		-1	$+2$		-2	$+1$	)	$*y_k -$	$\frac{h^2}{4}\, y^{(5)}$
$=$	$\frac{1}{8h^3}$ (	$+1$	-8	$+13$		-13	$+8$	-1)	$+$	$\frac{7h^4}{120}\, y^{(7)}$
$y_i^{(4)} =$	$\frac{1}{h^4}$ (		$+1$	-4	$+6$	-4	$+1$	)	$*y_k -$	$\frac{h^2}{6}\, y^{(6)}$
$=$	$\frac{1}{6h^4}$ (	-1	$+12$	-39	$+56$	-39	$+12$	-1)	$+$	$\frac{7h^4}{240}\, y^{(7)}$

Tabelle 1.6. Symmetrische Differenzenformeln

$k=$	i	-2	-1	0	$+1$	$+2$	$+3$	$+4$		R
$y_i' =$	$\frac{1}{h}$ (			-1	$+1$			)	$*y_k -$	$\frac{h}{2}\, y''$
$=$	$\frac{1}{2h}$ (			-3	$+4$	-1		)	$+$	$\frac{h^2}{3}\, y'''$
$=$	$\frac{1}{12h}$ (		-3	-10	$+18$	-6	$+1$	)	$-$	$\frac{h^4}{20}\, y^{(5)}$
$=$	$\frac{1}{60h}$ (	$+2$	-24	-35	$+80$	-30	$+8$	-1)	$+$	$\frac{h^6}{105}\, y^{(7)}$
$y_i'' =$	$\frac{1}{h^2}$ (			$+2$	-5	$+4$	-1	)	$*y_k +$	$\frac{11h^2}{12}\, y^{(4)}$
$=$	$\frac{1}{12h^2}$ (		$+11$	-20	$+6$	$+4$	-1	)	$+$	$\frac{h^3}{12}\, y^{(5)}$
$=$	$\frac{1}{180h^2}$ (	-13	$+228$	-420	$+200$	$+15$	-12	$+2$)	$-$	$\frac{h^5}{90}\, y^{(7)}$
$y_i''' =$	$\frac{1}{2h^3}$ (		-3	$+10$	-12	$+6$	-1	)	$*y_k +$	$\frac{h^2}{4}\, y^{(5)}$
$=$	$\frac{1}{8h^3}$ (	-1	-8	$+35$	-48	$+29$	-8	$+1$)	$-$	$\frac{h^4}{15}\, y^{(7)}$

Tabelle 1.7. Unsymmetrische Differenzenformeln

re Genauigkeit. Der nächste Parameter ist die Anzahl der Stützstellen. Bei Integranden, welche sich stark ändern, z.B. oszillierende Funktionen, ist es wichtig, genügend Stützstellen in das Integrationsintervall einzuführen. An den folgenden Richtlinien kann man sich orientieren.

Die aufsummierten Newton-Cotes-Formeln werden auf einem äquidistanten Gitter angewandt. Sie sind sehr einfach zu programmieren, allerdings erfüllen sie auch nur geringe Genauigkeitsansprüche. Simpson- oder 3/8-Regel sind für die Praxis sehr gut geeignet. Bei den Newton-Cotes-Formeln höherer Ordnung können verstärkt Rundungsfehler anwachsen, es treten Instabilitäten auf, so dass sie für praktische Rechnungen nicht mehr zu gebrauchen sind.

Bei der Gauß-Quadratur bekommt man etwa die doppelte Ordnung bei der gleichen Zahl von Funktionsauswertungen. Allerdings erfordert die Berechnung der Stützstellen und der Gewichte einen nicht zu vernachlässigenden Aufwand. Ist die Funktionsauswertung sehr Rechenzeit aufwändig, dann wird die Gauß-Quadratur Vorteile haben. Ein Spezialfall der Gauß-Integration sind die sogenannten Tschebyscheffschen Formeln, die wir in unseren Ausführungen nicht vorgestellt haben. Der Ansatz ist hier, dass die Gewichte konstant gehalten, aber die Stützstellen optimal gewählt werden. Durch die gleichen Gewichte sind die Tschebyscheffschen Formeln bezüglich Rundungsfehler sehr unempfindlich.

Für glatte, nicht zu stark oszillierende Funktionen ist die Romberg-Integration sehr gut geeignet. Das Romberg-Verfahren ist ein effektives und stabiles adaptives Verfahren. Eine Überprüfung, ob der Integrand auf den Teilintervallen genügend glatt ist, sollte mit dem Romberg-Verfahren gekoppelt werden. Der Ansatz der Adaptivität wird auch mit anderen Integrationsverfahren gekoppelt. Eine optimale Steuerung des Verfahrens wird so aussehen, dass in den Bereichen, in denen sich der Integrand stark ändert, automatisch sehr kleine Teilintervalle gewählt werden. Zur Minimierung der Anzahl der Stützstellen kann man in den Teilintervallen genaue Gauß-Formeln benutzen.

Die numerische Integration ist in vielen Büchern über numerische Methoden ausführlicher beschrieben. Wir verweisen hier auf die Bücher von Stoer [58], Deuflhard [15] und Schwarz [55]. Ein Standardwerk zur numerischen Integration ist das Buch von P. Davis und P. Rabinowitz [12]. Wir haben uns hier auf die Integration und Differenziation von Funktionen einer reellen Variablen beschränkt. Die Erweiterung auf Mehrfachintegrale wird in der oben angegebenen Literatur diskutiert. Es gibt eine ganze Reihe von Rechenprogrammen in den großen Software-Bibliotheken IMSL [68] und NAG [45]. Matlab und Computer-Algebra-Systeme wie MAPLE stellen hier ebenso Rechenprogramme zur Verfügung. Das Hauptanwendungsgebiet der verschiedenen abgeleiteten Differenzenformeln im Abschnitt über die numerische Differenziation ist

die numerische Lösung von Differenzialgleichungen mit Hilfe der Differenzenverfahren. Bei der Lösung von gewöhnlichen und partiellen Differenzialgleichungen werden wir darauf ausführlich eingehen.

1.8 Beispiele und Aufgaben

Aufgabe 1.1. Lösen sie mit der *aufsummierten* 3/8 *Regel* das Integral

$$\int_0^3 \left(e^{2x} + \sin(2x)^3\right) dx \ .$$

Schreiben sie für die Programmentwicklung zuerst ein Struktogramm. Wählen sie die Anzahl der Stützstellen so aus, dass der Fehler weniger als 0.001, $1.0 \cdot 10^{-5}$, $7.0 \cdot 10^{-6}$ beträgt.

Lösung: MAPLE**-Worksheet** □

Aufgabe 1.2. Lösen sie das Integral von Aufgabe 1.1 mit der *aufsummierten Simpson-Regel* und gehen sie wie dort vor. Wählen sie die Stützstellen jetzt so aus, dass der Fehler weniger als 0.01, $1.0 \cdot 10^{-3}$, $1.0 \cdot 10^{-5}$ beträgt.

Lösung: MAPLE**-Worksheet** □

Aufgabe 1.3. Lösen sie das Integral von Aufgabe 1.1 mit der *aufsummierten MacLaurin-Regel* vom Grad $n = 3$ und gehen sie wie dort vor. Wählen sie die Stützstellen jetzt so aus, dass der Fehler weniger als 0.01, $1.0 \cdot 10^{-3}$, $1.0 \cdot 10^{-5}$ beträgt.

Lösung: MAPLE**-Worksheet** □

Aufgabe 1.4. Erstellen sie mit Hilfe der Prozedur ***DiffFormeln*** eine Diskretisierungsformel für die dritte Ableitung ($n = 3$) bei äquidistanter Unterteilung des Intervalls t_1, t_2, t_3, t_4, t_5 an der Stelle t_2 ($i = 2$).

Lösung: MAPLE**-Worksheet** □

Aufgabe 1.5. Bestimmen sie die Ordnung obiger Differenzenformel mit der Prozedur ***DiffFormelnOrdnung***. □

2 Anfangswertprobleme gewöhnlicher Differenzialgleichungen

Bei der mathematischen Modellierung von ingenieur- oder naturwissenschaftlichen Problemen treten oft Differenzialgleichungen auf. Überall dort, wo die gesuchte Größe und deren Änderung in das mathematische Modell eingehen, wird sich eine solche ergeben. Ist der Anfangszustand bekannt und die Differenzialgleichung beschreibt die zeitliche oder räumliche Änderung des Vorgangs, so liegt ein sogenanntes Anfangswertproblem vor. In diesem Kapitel betrachten wir Probleme, welche nur von einer Variablen, Zeit- oder Raumvariable, abhängen, also Anfangswertprobleme für gewöhnliche Differenzialgleichungen.

Ein **Anfangswertproblem für eine gewöhnliche Differenzialgleichung (AWP)** schreibt man in der Form

$$y'(x) = f\big(x, y(x)\big) \ , \qquad y(x_0) = y_0 \ . \tag{2.1}$$

Da sehr wenige Anfangswertprobleme exakt lösbar sind, müssen für praktische Probleme meist numerische Verfahren eingesetzt werden. Der einfache Fall, wenn die rechte Seite f nur von x abhängt, führt durch Integration auf die Berechnung eines bestimmten Integrals, welches sich mit den Methoden aus Kapitel 1 numerisch lösen lässt. Im Folgenden werden wir von dem allgemeinen Fall ausgehen, bei dem $f = f(x, y)$, also auch von der Lösung $y = y(x)$ abhängt.

Oft treten in den Anwendungen **Systeme von Differenzialgleichungen** auf. Der Vorgang wird durch mehrere Zustandsgrößen bestimmt. Besteht das System aus m Gleichungen, so werden m unbekannte Funktionen $y_1,\ y_2,\ \ldots\ ,\ y_m$ gesucht, welche die Differenzialgleichungen

$$y_i'(x) = f_i(x, y_1(x), y_2(x), \ldots, y_m(x)) \ , \qquad y_i(x_0) = y_{i0} \ , \tag{2.2}$$

für $i = 1, 2, \ldots, m$ erfüllen. Führt man die vektorwertigen Funktionen

$$\mathbf{y} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_m \end{pmatrix}, \qquad \mathbf{f}(x, \mathbf{y}) = \begin{pmatrix} f_1(x, y_1, \ldots, y_m) \\ f_2(x, y_1, \ldots, y_m) \\ \vdots \\ f_m(x, y_1, \ldots, y_m) \end{pmatrix}$$

ein, dann kann man das Anfangswertproblem für das System (2.2) formal in der Form (2.1) schreiben. Die skalaren Funktionen y und $f(x, y)$ werden durch die Vektorfunktionen $\mathbf{y}$ und $\mathbf{f}(x, \mathbf{y})$ in (2.1) ersetzt. Alle folgenden Herleitungen von numerischen Verfahren werden wir für eine einzelne Gleichung (2.1) ausführen. Eine Erweiterung auf Systeme ist direkt über die oben angegebene Vektorformulierung möglich. An Hand von einzelnen Beispielen wird dies im Abschnitt 2.7 aufgegriffen. Differenzialgleichungen höherer Ordnung lassen sich auf Systeme erster Ordnung zurückführen. Auch dies werden wir in diesem Abschnitt kurz behandeln.

Im Folgenden setzen wir immer voraus, dass eine eindeutige, stetig differenzierbare Lösung des Anfangswertproblems existiert. Ein zentraler Satz zur globalen Existenz und Eindeutigkeit ist der Satz von *Picard-Lindelöf*. Die wesentliche Voraussetzung in diesem Satz ist neben der Stetigkeit von f die *Lipschitz-Stetigkeit* von f bezüglich dem zweiten Argument y. Ist f stetig differenzierbar, dann ist diese Bedingung automatisch erfüllt. Einen Überblick über die Theorie von Anfangswertproblemen und die elementar lösbaren Typen finden sich z. B. bei Walter [69] und Heuser [32].

Bevor wir mit der Beschreibung der numerischen Methoden starten, betrachten wir einige Beispiele von einfachen Anwendungen.

Beispiel 2.1 (Mathematisches Modell zur Populationsdynamik). Für den ersten Test und den Vergleich der numerischen Methoden wird man ein einfaches Beispiel wählen, bei dem die exakte Lösung bekannt und einfach zu berechnen ist. Eine für diesen Zweck oft benutzte Differenzialgleichung ist

$$y' = s\, y \tag{2.3}$$

mit dem Anfangswert

$$y(0) = y_0\,. \tag{2.4}$$

Dieses Problem lässt sich auch als ein einfaches mathematisches Modell im Bereich der **Populationsdynamik** interpretieren. Ist die unabhängige Variable die Zeit t und die Lösung eine Funktion $y = y(t)$, dann sagt diese Differenzialgleichung aus, dass die **Änderung der Population** y proportional zu y ist. Den Wert s nennt man den Proportionalitätsfaktor. Mit einem **Proportionalitätsfaktor** $s > 0$ kann man es als sehr einfaches Modell für das Anwachsen einer Bakterienpopulation vorstellen. Die zeitliche Änderung der Anzahl der Bakterien ist mit der Menge der Bakterien direkt gekoppelt.

Das Anfangswertproblem (2.3), (2.4) lässt sich mit Hilfe der Trennung der Veränderlichen exakt lösen und besitzt die Lösung

$$y(t) = y_0 \, e^{st} \, . \tag{2.5}$$

Die Anzahl der Bakterien wird in diesem mathematischen Modell somit exponentiell anwachsen. Für $s = 0$ erhält man die konstante Lösung, für $s < 0$ eine mit einem negativen Exponenten fallende Lösung.

Das Modell mit einem exponentiellen Wachstum für alle Zeiten ist für eine Population nicht sehr realistisch. Mit dem Modell (2.3), (2.4) wurde eine kontinuierliche Approximation einer diskreten Population eingeführt unter der Vernachlässigung von verschiedenen Faktoren, wie z. B. Nahrungsmangel. Ein Modell, welches zumindest bei Überbevölkerung mit einer Verringerung der Populationsgröße reagiert, würde so aussehen:

$$y'(t) = s \, y(t) \, (y(t) - K) \, . \tag{2.6}$$

Für $s > 0$ wird die rechte Seite beim Überschreiten der Schranke $y = K$ negativ und die Größe der Population fällt dementsprechend. Als Grenzwert für $t \to \infty$ wird sich $y = K$ einstellen. Eine sehr lesenswerte Beschreibung der Modellierung von Populationsmodellen findet sich in dem Buch von Sonar [56] und weiterführend bei Bossel [4]. □

Beispiel 2.2 (Flug in konstanter Höhe). Ein einfaches Modell für den Flug eines Flugzeuges in konstanter Höhe wurde in der Einleitung bei der Diskussion der einzelnen Schritte der Simulation eines Anwendungsproblems eingeführt. Das mathematische Modell hierfür ist die Differenzialgleichung

$$\dot{v} = f(v) = a_0 - a_1 v^2 - \frac{a_2}{v^2} \, , \tag{2.7}$$

wobei die Konstanten a_0, a_1 und a_2 sich aus

$$a_0 = \frac{F}{m} \, , \qquad a_1 = \frac{\rho \, S \, c_{W_0}}{2m} \, , \qquad a_2 = \frac{2k \, G^2}{m \, \rho \, S}$$

berechnen. Wenn die Geschwindigkeit zum Zeitpunkt $t_0 = 0$ vorgegeben ist, ergibt die Lösung des Anfangswertproblems die Geschwindigkeit als Funktion der Zeit. □

Bemerkung: Eine Differenzialgleichung wie (2.7) oder ein System von Differenzialgleichungen, deren rechte Seite nur von der gesuchten Funktion abhängt, nennt man auch ein **autonomes dynamisches System**.

Beispiel 2.3 (Biegung eines Balkens). Bei der numerischen Integration in Kapitel 1 wurde das mathematische Modell der Biegung eines Balken schon aufgegriffen und der linearisierte Fall behandelt.

Allgemein lautet es:

$$\frac{dQ(z)}{dz} = q(z)\ , \tag{2.8}$$

$$\frac{dM(z)}{dz} = Q(z)\ , \tag{2.9}$$

$$\frac{d\phi(z)}{dz} = -\frac{M(z)}{E(z)I(z)}\left(1+\phi^2(z)\right)^{\frac{3}{2}}\ , \tag{2.10}$$

$$\frac{dv(z)}{dz} = \phi(z)\ . \tag{2.11}$$

Dabei beschreibt $q = q(z)$ die Einzelkräfte, $Q = Q(z)$ die Querkräfte, $M = M(z)$ das Biegemoment. Die Variable z ist die unabhängige Variable in Längsrichtung des Balkens, $\phi = \phi(z)$ der Biegewinkel, $v = v(z)$ die Durchbiegung in z-Richtung, $E = E(z)$ das Elastizitätsmodul und $I = I(z)$ das Trägheitsmoment.

Es liegt somit ein System von vier gewöhnlichen Differenzialgleichungen vor. In die Form (2.1) kann es durch die Definitionen

$$\mathbf{y} = \begin{pmatrix} y_1 \\ y_2 \\ y_3 \\ y_4 \end{pmatrix} = \begin{pmatrix} Q \\ M \\ \phi \\ v \end{pmatrix}, \qquad \mathbf{f} = \mathbf{f}(z,\mathbf{y}) = \begin{pmatrix} q \\ y_1 \\ -\frac{y_2}{EI}\left(1+y_3^2\right)^{\frac{3}{2}} \\ y_3 \end{pmatrix}$$

umgeschrieben werden.

Ist der Balken auf der linken Seite bei $z = A$ fest eingespannt, wie es in Abbildung 1.2 in Kapitel 1 dargestellt ist, dann sind die Werte der Lösung an diesem Punkt vorgegeben: Sowohl die Durchbiegung v als auch der Biegewinkel ϕ haben dort den Wert Null, Biegemoment und Querkraft sind ebenso Null. Es liegen somit die Anfangsbedingungen

$$y_i(A) = 0 \qquad \text{mit} \quad i = 1,2,3,4$$

vor.

Die rechte Seite der ersten beiden Gleichungen im Differenzialgleichungssystem für die Balkenbiegung hängen nur von der unabhängigen Variablen z ab. Die Querkraft Q berechnet sich aus der Verteilung der Einzelkräfte und daraus das resultierende Biegemoment M jeweils durch eine einfache Integration. Nehmen wir an, diese sei ausgeführt, dann ist die Gleichung für das Biegemoment in der Form

$$\phi'(z) = -r(z)\,(1+\phi^2(z))^{\frac{3}{2}} \tag{2.12}$$

mit

$$r(z) = \frac{M(z)}{E(z)I(z)}$$

gegeben. Diese muss numerisch gelöst werden. Die Durchbiegung v ergibt sich nach (2.11) wieder durch eine einfache Integration.

Bei kleinen Biegewinkeln, welche bei kleinen elastischen Verschiebungen meist vorliegen, ist ϕ^2 sehr klein und kann vernachlässigt werden. In diesem Fall kann die Gleichung (2.10) durch

$$\phi'(z) = -r(z)$$

angenähert werden und man erhält mit (2.8) - (2.11) ein lineares Differenzialgleichungssystem, deren rechte Seite von y nicht mehr abhängt. Es kann somit wie in Kapitel 1 ausgeführt, sukzessive direkt integriert werden. Dieser lineare Fall kann auch für die Verifikation numerischer Methoden für das vollständige System eingesetzt werden. Für ein Problem mit einem kleinen Biegewinkel müssen die numerischen Ergebnisse mit beiden mathematischen Modellen gut übereinstimmen. □

Beispiel 2.4 (Elektrische Schaltungen). Ein weiteres Beispiel, welches zu einem Differenzialgleichungssystem 1. Ordnung führt, ist die Modellierung elektrischer Schaltungen, wenn sich die Schaltungen aus Ohmschen Widerständen R, Kapazitäten C und Induktivitäten L zusammensetzen. Das Ziel bei der numerischen Behandlung ist, für beliebige Eingangssignale die zugehörigen Ausgangssignale zu berechnen. Um die Schaltungen zu modellieren, benötigt man die folgenden physikalischen Gesetze:

Ohmscher Widerstand: 

$$\boxed{U(t) = R \cdot I(t)}$$

Der Strom durch den Widerstand R ist $I(t) = \frac{1}{R}U(t)$, wenn $U(t)$ der Spannungsabfall über R ist.

Spule mit Induktivität L: 

$$\boxed{U(t) = L \cdot \frac{dI(t)}{dt}}$$

Fließt durch eine Spule L ein Strom I, so ist der Spannungsabfall an der Spule proportional $\frac{dI}{dt}$. Die Proportionalitätskonstante bezeichnet man mit Induktivität L: $U(t) = L\frac{dI(t)}{dt}$.

Kondensator mit Kapazität C: 

$$\boxed{I(t) = C \cdot \frac{dU(t)}{dt}}$$

Liegt am Kondensator die Spannung $U(t)$, so ist die auf dem Kondensator

induzierte Ladung $Q(t) = C \cdot U(t)$, wenn C die Kapazität des Kondensators ist. Wegen $I = \frac{dQ}{dt}$ folgt für den Strom $I(t) = C \cdot \frac{dU(t)}{dt}$.

Um die Differenzialgleichungen für die elektrischen Schaltungen aufzustellen, wendet man den folgenden systematischen Ansatz an:

1. Jedem Energiespeicher wird eine Zustandsvariable zugeordnet:
   Jedem Kondensator C_i wird seine Spannung U_i, jeder Spule L_i wird ihr Strom I_i als Zustandsvariable zugeordnet. Den Widerständen wird keine Zustandsvariable zugeordnet, da Widerstände keine Energiespeicher, sondern nur Energieverbraucher darstellen.

2. Der Maschensatz und der Knotensatz werden auf die Schaltung angewendet. Das Ziel ist für jede Zustandsvariable eine Differenzialgleichung 1. Ordnung zu erhalten. Sind Spule und Kondensator in Reihe geschaltet, wird formal ein weiterer Knoten eingeführt.

3. In der Regel führt die Vorgehensweise zu je einer Differenzialgleichung für jede Zustandsvariable. In manchen Fällen sind aber mehrere Ableitungen 1. Ordnung in einer Differenzialgleichung vorhanden. Es muss dann zuerst ein lineares Gleichungssystem gelöst werden, um die übliche Struktur des Systems von Differenzialgleichungen zu erhalten.

Exemplarisch wird für einen Tiefpass 5. Ordnung das System der Modellgleichungen aufgestellt. Der Tiefpass besteht aus 5 Energiespeicher C_1, C_2, C_3, L_1, L_2; diesen Energiespeichern werden 5 Zustandsvariablen U_1, U_2, U_3, I_1, I_2 zugeordnet. Das Anwenden der Knoten- und Maschenregel liefert:

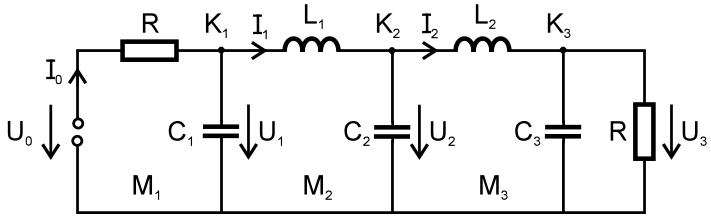


Abb. 2.1. Tiefpass

$$
\begin{array}{llll}
M_1: & U_0 = RI_0 + U_1 & K_1: & I_0 = C_1\dot{U}_1 + I_1 \\
M_2: & U_1 = L_1\dot{I}_1 + U_2 & K_2: & I_1 = C_2\dot{U}_2 + I_2 \\
M_3: & U_2 = L_2\dot{I}_2 + U_3 & K_3: & I_2 = C_3\dot{U}_3 + U_3/R
\end{array}
$$

I_0 muss ersetzt werden, da es **keinem** Energiespeicher zugeordnet ist!

Man beachte, dass in den Gleichungen die Größe I_0 auftritt, die keinem Energiespeicher zugeordnet ist. Sie muss aus den Gleichungen eliminiert werden, indem wir sie durch die Gleichung des Knotensatzes K_1 ersetzen. Wir erhalten fünf Differenzialgleichungen 1. Ordnung:

$$\frac{dU_1}{dt} = ((U_0 - U_1)/R - I_1)/C_1 \ , \tag{2.13}$$

$$\frac{dI_1}{dt} = (U_1 - U_2)/L_1 \ , \tag{2.14}$$

$$\frac{dU_2}{dt} = (I_1 - I_2)/C_2 \ , \tag{2.15}$$

$$\frac{dI_2}{dt} = (U_2 - U_3)/L_2 \ , \tag{2.16}$$

$$\frac{dU_3}{dt} = (I_2 - U_3/R)/C_3 \ . \tag{2.17}$$

Wir werden nach der Vorstellung der numerischer Methoden für Anfangswertprobleme auch auf dieses Beispiel zurückkommen. □

2.1 Das Euler-Cauchy-Verfahren

Wir betrachten in diesem Abschnitt zunächst die einfachsten numerischen Approximationen des Anfangswertproblems, welche schon lange vor der Verfügbarkeit von Computern zur näherungsweisen Lösung benutzt wurde. Damals mussten die Näherungswerte per Hand berechnet werden.

Zur Konstruktion eines Näherungsverfahrens geht man zunächst wie bei der numerischen Quadratur vor. Es werden in das betrachtete Intervall $I = [a, b]$ Stützstellen x_i eingeführt. Wir nehmen an, dass wir die Lösung am Punkt x_i kennen und suchen die Lösung der Differenzialgleichung

$$y' = f(x, y) \tag{2.18}$$

am nächsten Punkt x_{i+1}. Wird die Differenzialgleichung (2.18) bezüglich x über das Intervall $[x_i, x_{i+1}]$ integriert, erhält man die Integralgleichung

$$y(x_{i+1}) = y(x_i) + \int_{x_i}^{x_{i+1}} f(x, y(x)) \, dx \ . \tag{2.19}$$

Die erste Idee ist, das Integral durch eine der in Kapitel 1 abgeleiteten Integrationsformeln zu approximieren. Leider hängt aber die rechte Seite von (2.18)

und damit der Integrand von der Lösung $y = y(x)$ selbst ab. Die Sache wird also komplizierter.

Beginnen wir mit der einfachsten Integrationsformel und betrachten die Rechteckregel mit Auswertung des Integranden an dem linken Randpunkt x_i des Intervalls. Man erhält

$$\int_{x_i}^{x_{i+1}} f(x, y(x))\, dx \approx h\, f(x_i, y_i) \qquad \text{mit} \quad h := x_{i+1} - x_i$$

und aus (2.19) dann die Näherungsformel

$$y_{i+1} = y_i + h\, f(x_i, y_i)\ . \tag{2.20}$$

Diese Formel nennt man die **Euler-Cauchy-** oder **Streckenzug-Formel**.

Geometrisch kann man sich dies folgendermaßen vorstellen. Die Funktion y wird im Intervall $[x_i, x_{i+1}]$ durch eine Gerade approximiert, die den Punkt (x_i, y_i) durchläuft und die Steigung $f(x_i, y_i)$ besitzt. In Abbildung 2.2 ist diese Situation in dem rechten Schaubild dargestellt. Links sieht man f über x aufgetragen, eingezeichnet ist die Approximation des Integrals durch die Rechteckregel.

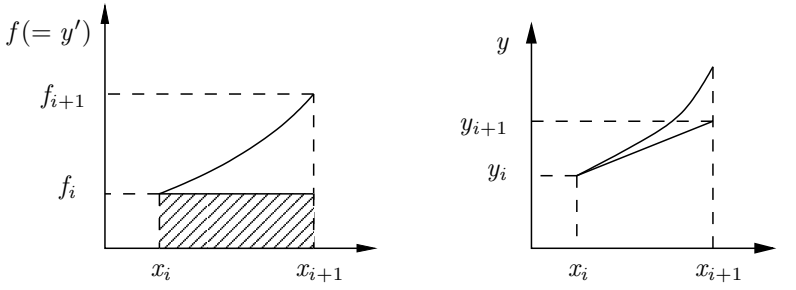


Abb. 2.2. Euler-Cauchy-Einschrittverfahren

Wendet man die Euler-Cauchy-Formel auf das Anfangswertproblem

$$y' = f(x, y)\ , \quad y(a) = y_a \tag{2.21}$$

im Intervall $[a, b]$ an, so zerlegt man das Intervall in n Teilintervalle der Länge

$$h = \frac{b - a}{n}\ . \tag{2.22}$$

Wie schon bei der numerischen Integration nennen wir h die **Schrittweite**. Die Punkte

$$x_i = a + ih \ , \qquad i = 0, 1, \ldots, n$$

sind die **Stützstellen**. Wenden wir sukzessive mit Start bei $x_0 = a$ die Näherungsformel (2.20) an, dann erhalten wir der Reihe nach die Näherungen $y_1, y_2, \ldots$ für die Funktionswerte $y(x_1), y(x_2), \ldots$. Man berechnet die Werte also sukzessive vom Punkt $P_a = (a, y_a)$ aus. Die Lösungskurve setzt sich aus geradlinigen Strecken zusammen. Deshalb nennt man das Euler-Cauchy-Verfahren auch **Polygonzug-** oder **Streckenzugverfahren**.

Beispiel 2.5 (Euler-Cauchy-Verfahren, mit MAPLE-Worksheet). Wir greifen ein Anfangswertproblem für das einfache Beispiel 2.1 mit $s = 1$ auf. Gesucht ist die Lösung von

$$y' = -y \ , \ y(0) = 1 \tag{2.23}$$

im Intervall $[0, 4]$. Die exakte Lösung zum Vergleich mit den Näherungswerten ist

$$y(x) = e^{-x}.$$

Im MAPLE-Worksheet wird dieses Anfangswertproblem mit dem Euler-Cauchy-Verfahren gelöst. Dabei werden $n = 10$ Schritte ausgeführt und somit 11 Stützstellen mit Anfangs- und Endpunkt des Intervalls eingeführt. Das Struktogramm für ein solches Computerprogramm ist in Abbildung 2.3 zu sehen. Die Anfangswerte, der Endpunkt und die Anzahl der Schritte werden dabei eingegeben. Es wird eine Plausibilitätsprüfung der eingegebenen Daten ausgeführt. Einen Vergleich von exakter Lösung und Näherungslösung mit dem Euler-Cauchy-Verfahren für 11 Stützstellen sieht man in Abbildung 2.4, wie sie von dem Worksheet erzeugt wird.

Mit 11 Stützstellen erhält man bei $x = 4$ einen relativen Fehler von ca. 67 Prozent. Bei Verfeinerung des Gitters verringert sich der Fehler, wie zu erwarten ist. Für die Anzahl 20, 40, 100, und 200 Schritte erhält man die Werte in der folgenden Tabelle:

Schrittweite h	Relativer Fehler in Prozent
0.4	67
0.2	37
0.04	8
0.02	4

Der Fehler wird proportional zur Schrittweite kleiner. □

Euler-Cauchy

Real: $x, y, x_0, y_0, x_{end}, h$
Integer: *Schritte*
Lies x_0 und y_0 ein
Lies x_{end} und *Schritte* ein
$h = \dfrac{x_{end} - x_0}{Schritte}$
Schreibe x_0, y_0 in Datei
$x = x_0$
$y = y_0$
$x = x + h$
$y = y - h\,y$
Schreibe x, y in Datei
bis $x = x_{end}$

Abb. 2.3. Struktogramm zum Euler-Cauchy-Verfahren

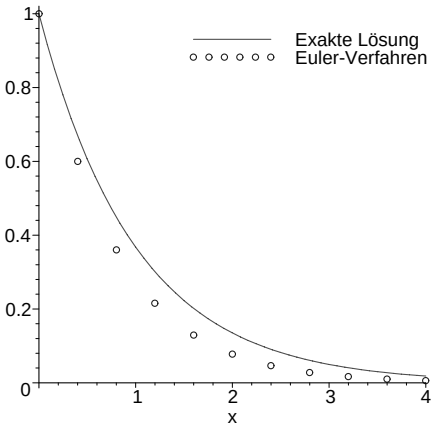


Abb. 2.4. Näherungslösung mit Euler-Cauchy-Verfahren, berechnet mit 10 Teilintervallen

Bei der numerischen Integration in Kapitel 1 haben wir gesehen, dass die Rechteckregel bezüglich der Genauigkeit schlecht ist. Es kommt hier hinzu,

dass die numerische Lösung eines Anfangswertproblems im Hinblick auf Approximationsfehler empfindlicher als die Integration ist. Hängt die rechte Seite zusätzlich von y ab, muss man schrittweise von einem Stützpunkt zum nächsten fortschreiten. Approximationsfehler im n-ten Schritt ergeben im nächsten Schritt Steigungsfehler, also Fehler bei der Auswertung von f. Dies bedeutet, dass sich Approximationsfehler fortpflanzen. Ein genaueres Verfahren als das Euler-Cauchy-Verfahren ist für praktische Berechnungen unumgänglich.

Wird das Integral statt mit der Rechteckregel mit der **Mittelpunktsformel** approximiert, so ergibt sich die folgende Situation. Die Steigung wird am Punkt $x_{i+1/2} = x_i + \frac{h}{2}$ genommen, und wir erhalten die Formel

$$y_{i+1} = y_i + hf(x_{i+\frac{1}{2}}, y_{i+\frac{1}{2}}) \; . \tag{2.24}$$

Der Wert $y_{i+\frac{1}{2}}$ ist allerdings unbekannt. Man möchte zudem an einem Mittelpunkt zwischen zwei Stützstellen keinen zusätzlichen Näherungswert bestimmen. Es hilft hier die Taylor-Entwicklung

$$\begin{aligned} y(x_{i+\frac{1}{2}}) &= y(x_i) + (x_{i+\frac{1}{2}} - x_i)\; y'(x_i) + \mathcal{O}(h^2) \\ &= y(x_i) + \frac{h}{2} f(x_i, y_i) + \mathcal{O}(h^2) \; . \end{aligned}$$

Eingesetzt in Formel (2.24) und bei Vernachlässigung der $\mathcal{O}(h^2)$-Terme erhält man die Formel

$$y_{i+1} = y_i + h\, f\big(x_{i+\frac{1}{2}}, y_i + \frac{h}{2} f(x_i, y_i)\big) \; , \tag{2.25}$$

welche nun direkt ausgewertet werden kann. Man nennt dieses Verfahren das **verbesserte Euler-Cauchy-Verfahren** oder die **verbesserte Polygonzugmethode**. Auf dem linken Bild in der Abbildung 2.5 ist in das Schaubild von f die Approximation des Integrals durch die Mittelpunktsregel eingezeichnet. Der Näherungswert an der Stelle x_{i+1} ergibt sich dann nach Formel (2.25) wie im rechten Bild.

Ein drittes Verfahren ergibt sich aus der Anwendung der **Sehnentrapezformel** auf das Integral in der Integralgleichung (2.19):

$$y_{i+1} = y_i + \frac{h}{2}\big(f(x_i, y_i) + f(x_{i+1}, y_{i+1})\big) \; . \tag{2.26}$$

Hier tritt allerdings auf der rechten Seite das Argument y_{i+1} auf. Da der

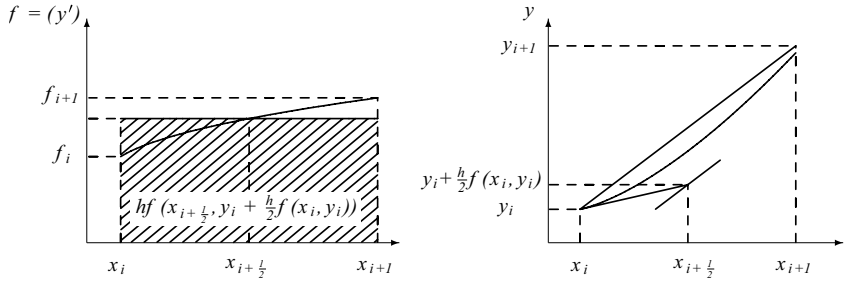


Abb. 2.5. Verbessertes Euler-Cauchy-Verfahren

gesuchte neue Näherungswert somit nur **implizit** gegeben ist, spricht man von einer **impliziten Methode**. Nur in Spezialfällen lässt sich die Gleichung (2.26) nach y_{i+1} direkt auflösen. Das (f, x)- und das (y, x)-Diagramm ist hier in Abbildung 2.6 aufgezeichnet.

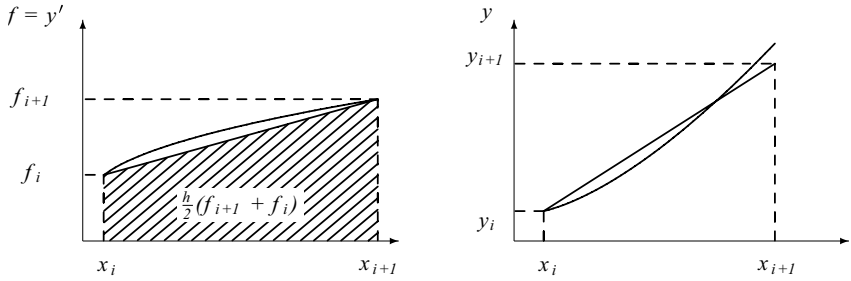


Abb. 2.6. Verfahren mit Anwendung der Sehnentrapezformel

Beispiel 2.6. Für die Differenzialgleichung

$$y'(x) = x\, y(x)$$

ergibt sich das implizite Verfahren (2.26) zu

$$y_{i+1} = y_i + \frac{h}{2}(x_i y_i + x_{i+1} y_{i+1}) \,.$$

Nach y_{i+1} aufgelöst erhält man

$$y_{i+1} = \frac{y_i + \frac{h}{2} x_i y_i}{1 - \frac{h}{2} x_{i+1}} \,.$$

Für jede lineare Gleichung kann das oben beschriebene Verfahren in dieser Art ausgeführt werden. □

Falls f nichtlinear in y ist, geht eine solche Auflösung im Allgemeinen nicht mehr. Die nichtlineare Gleichung muss mit Hilfe eines Iterationsverfahrens näherungsweise gelöst werden.

Wir betrachten hier die einfache Iterationsformel

$$y_{i+1}^{(\nu+1)} = y_i + \frac{h}{2}\big(f(x_i, y_i) + f(x_{i+1}, y_{i+1}^{(\nu)})\big) \ , \quad \nu = 0, 1, 2, \ldots \ . \tag{2.27}$$

Der obere Index bezeichnet dabei den Iterationsschritt. Starten kann man bei $\nu = 0$ etwa mit dem Wert $y_{i+1}^0 = y_i$ oder besser mit dem Näherungswert am Punkt x_{i+1} aus Euler-Cauchy-Verfahren.

Wir können (2.27) abgekürzt in der Form einer allgemeinen Iterationsvorschrift

$$y^{(\nu+1)} = \varphi(y^{(\nu)}) \ , \quad \nu = 0, 1, 2, \ldots$$

schreiben. Man nennt $y = \varphi(y)$ eine **Fixpunktgleichung**. Ein Iterationverfahren zur Ermittlung der Lösung einer Fixpunktgleichung, des Fixpunktes, konvergiert, wenn für die Funktion φ die Ungleichung

$$\left| \frac{d\varphi(y)}{dy} \right| < 1$$

erfüllt ist. In unserem Fall (2.27) muss somit

$$\left| \frac{\partial}{\partial y_{i+1}} \big(\frac{h}{2} f(x_{i+1}, y_{i+1}) \big) \right| = \frac{h}{2} |f_y(x_{i+1}, y_{i+1})| < 1$$

erfüllt sein. Dabei bedeutet f_y die partielle Ableitung von $f = f(x, y)$ nach dem zweiten Argument. Es ergibt sich somit eine Bedingung an die Schrittweite h in der Form

$$\kappa := h \, |f_y| < 2 \ , \tag{2.28}$$

wobei $\boldsymbol{\kappa}$ mit **Schrittkennzahl** bezeichnet wird.

Um die Konvergenz zu garantieren, muss die Bedingung nach (2.28) $\kappa < 2$ erfüllt sein. Die Iterationsvorschrift (2.27) wird so lange wiederholt bis eine ausreichende Genauigkeit erreicht ist. Was bedeutet hier ausreichend? Es macht keinen Sinn, den Abbruchfehler der Iteration mit vielen Iterationen

sehr viel kleiner als den Approximations- oder Diskretisierungsfehler zu machen. Je kleiner die Schrittkennzahl κ ist, desto größer ist die Konvergenzgeschwindigkeit und man ist mit einigen wenigen Iterationen am Ziel.

Wird nur eine Iteration ausgeführt, dann spricht man auch von einem Prädiktor-Korrektor-Verfahren, dem **Verfahren von Heun**:

Prädiktor (Euler-Cauchy-Verfahren)

$$\tilde{y}_{i+1} = y_i + hf(x_i, y_i)$$

Korrektor (Sehnentrapezregel)

$$y_{i+1} = y_i + \frac{h}{2}\big(f(x_i, y_i) + f(x_{i+1}, \tilde{y}_{i+1})\big) \ .$$

Beim Verfahren von Heun mit nur einer Iteration wird man in der praktischen Berechnung nicht mit $\kappa = 2$ rechnen, sondern mit kleineren Werten. Als günstig erwiesen hat sich der Bereich:

$$0.05 < \kappa < 0.2 \ .$$

Da sich mit der Lösung von einer zur anderen Stützstelle auch der Wert von f_y ändern kann, ist es ungünstig mit einer konstanten Schrittweite zu rechnen, wie wir dies bisher der Einfachheit halber vorausgesetzt haben. In den angegebenen Formeln muss dann h durch h_i ersetzt werden. Wir werden auf eine solche Schrittweitensteuerung im Abschnitt 2.6 zurückkommen.

Bemerkung: Die Gleichung (2.26) ist im allgemeinen Fall eine nichtlineare Gleichung und kann mit einem Näherungsverfahren für gelöst werden. Eine kurze Einführung in die numerische Lösung nichtlinearer Gleichungen ist in Anhang B aufgeführt. Das hier benutzte einfache Iterationsverfahren wird **Fixpunktiteration** genannt. Das benutzte Konvergenzkriterium ist der Banachsche Fixpunktsatz. Eine ausführliche Beschreibung von Iterationsverfahren findet sich z.B. bei Westermann [72, 71] und in Numerik-Lehrbüchern, wie [58] oder [55].

Beispiel 2.7 (Verfahren von Heun, mit MAPLE-Worksheet). In dem Worksheet wird das Anfangswertproblem (2.23) mit dem Verfahren von Heun berechnet. Halbiert man beim Euler-Cauchy- und Heun-Verfahren die Schrittweiten und führt die Rechnungen nochmals durch, bekommt man mit beiden Methoden bessere Ergebnisse. Dies hatten wir schon mit Euler-Cauchy-Verfahren in Beispiel 2.5 ausprobiert. Berechnet man die Fehler zur exakten Lösung, erhält man die beiden Fehlerkurven in Abbildung 2.7.

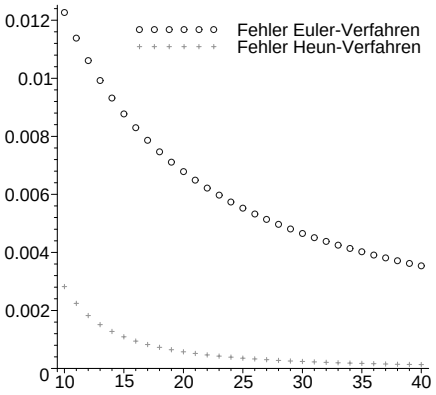


Abb. 2.7. Fehler bei Euler-Cauchy- und Heun-Verfahren

Es zeigt sich, dass das Verfahren von Heun bei 11 Stützstellen um zwei Größenordnungen genauer ist als das Euler-Cauchy-Verfahren. Bei der Verkleinerung der Schrittweiten wird der Fehler beim Heun-Verfahren deutlich schneller kleiner. Dies war eigentlich zu erwarten, da die zugrunde liegende Integrationsformel genauer ist. □

Wie man in dem Beispiel sieht, gibt es hier in Analogie zur numerischen Integration gewisse Qualitätskriterien, wie die **Genauigkeit des Verfahrens**. Diese wird im Folgenden näher betrachtet. Wie bei der numerischen Integration erhält man Aussagen zur Genauigkeit mit Hilfe von Taylor-Entwicklungen. Es wird hier von der Taylor-Entwicklung

$$y(x_{i+1}) = y(x_i) + hy'(x_i) + \frac{h^2}{2}y''(x_i) + \dots \ ,$$

ausgegangen, wobei $y(x_i)$ wieder den Wert der exakten Lösung an der Stützstelle x_i bezeichnet. Die Ableitungen von $y = y(x)$ können mit Hilfe der Differenzialgleichung wie folgt ersetzt werden:

$$\begin{aligned} y'(x) &= f\big(x, y(x)\big) \ , \\ y''(x) &= \frac{d}{dx}y'(x) = \frac{d}{dx}f\big(x, y(x)\big) = f_x + f\, f_y \ , \end{aligned}$$

$$\begin{aligned} y'''(x) &= f_{xx} + 2f\, f_{xy} + f^2 f_{yy} + f_x f_y + f\, f_y^2 \ , \\ &\ \vdots \quad \vdots \qquad \vdots \qquad \vdots \qquad \vdots \qquad . \end{aligned}$$

Dabei bezeichnen f_x und f_y die partiellen Ableitungen von $f = f(x, y)$ nach dem ersten bzw. zweiten Argument im Punkt $(x, y(x))$. Entsprechend sind die Ableitungen höherer Ordnung bezeichnet.

Wir vergleichen die Euler-Cauchy-Formel mit dieser Taylor-Entwicklung und bilden die Differenz des Näherungswertes y_{i+1} und des exakten Wertes $y(x_{i+1})$

$$y_{i+1} - y(x_{i+1}) = y_i + hf(x_i, y_i) - y(x_i) - hy'(x_i) - \dots$$

Nimmt man an, dass $y_i = y(x_i)$ gilt, d.h. an der Stelle x_i stimmt der Näherungswert mit dem Wert der exakten Lösung überein, dann erhält man

$$y_{i+1} - y(x_{i+1}) = -\frac{h^2}{2}(f_x + f\, f_y)(x_i, y_i) + \mathcal{O}(h^3) \; .$$

Der Fehler, den man auch das lokale Restglied nennt, lautet also

$$R = -\frac{h^2}{2}(f_x + f\, f_y) \; ,$$

ausgewertet an einer Zwischenstelle $\xi \in (x_i, x_{i+1})$. Dieser Fehler, der in jedem Schritt gemacht wird, ist somit proportional zu h^2. Der gesamte Fehler im Intervall $[a, b]$, in dem n Schritte mit dem Euler-Verfahren durchgeführt werden, ergibt sich durch Akkumulation dieses lokalen Fehlers. Wegen $n = (b - a)/h$ hebt sich bei der Berechnung des globalen Fehlers ein h heraus. Der globale Fehler geht beim Euler-Cauchy-Verfahren somit linear gegen Null, wenn die Schrittweite h gegen Null strebt. Dies hat sich auch in Beispiel 2.5 so ergeben.

Für das Verfahren von Heun wird die Berechnung des Diskretisierungsfehlers aufwändiger. Wegen $f(x_{i+1}, \tilde{y}_{i+1})$ mit dem Wert $\tilde{y}_{i+1}$ aus dem Prädiktor

$$\tilde{y}_{i+1} = y_i + hf(x_i, y_i) \; ,$$

benötigt man eine Taylor-Entwicklung von f in zwei Variablen:

$$\begin{aligned} f(x_{i+1}, \tilde{y}_{i+1}) = f(x_i, y_i) &+ hf_x + hf\, f_y \\ &+ \frac{h^2}{2}\left(f_{xx} + 2f\, f_{xy} + f^2\, f_{yy}\right) \\ &+ \frac{h^3}{6}\left(f_{xxx} + 3f\, f_{xxy} + 3f^2\, f_{xyy} + f^3\, f_{yyy}\right) \\ &+ \dots \; , \end{aligned}$$

wobei auf der rechten Seite die Argumente (x_i, y_i) weggelassen wurden. In die Heunsche Formel

$$y_{i+1} = y_i + \frac{h}{2}(f(x_i, y_i) + f(x_{i+1}, \tilde{y}_{i+1}))$$

eingesetzt, erhält man

$$y_{i+1} - y(x_{i+1}) = -\frac{h^3}{12}(f_{xx} + 2f\, f_{xy} + f^2\, f_{yy} - 2f_y\, (f_x + f\, f_y)) + \mathcal{O}(h^4)\ .$$

Das lokale Restglied ist proportional zu h^3. Es ergibt sich somit ein globaler Fehler proportional zu h^2. Bei dieser Berechnung des Fehlerterms muss man voraussetzen, dass die Taylor-Entwicklungen bis zur entsprechenden Ordnung existieren, d.h. dass die entsprechende Differenzierbarkeit von f vorliegt. Sind diese Stetigkeitsanforderungen erfüllt, dann wird der Fehler beim Heun-Verfahren schneller gegen Null streben, wenn h gegen Null strebt. Dieser Sachverhalt lässt sich in Abbildung 2.7 ablesen.

Nach den Erfahrungen mit der numerischen Integration wissen wir, dass eine höhere Genauigkeit und damit oft auch eine bessere Effizienz des Verfahrens erzielt werden kann, wenn Integrationsformeln mit mehreren Stützstellen benutzt werden. Statt der Sehnentrapezregel können wir es z. B. mit der Simpson-Regel versuchen. Die Differenzialgleichung wird dazu über das Intervall $[x_{i-1}, x_{i+1}]$ integriert:

$$y(x_{i+1}) = y(x_{i-1}) + \int_{x_{i-1}}^{x_{i+1}} f(x, y(x))\ dx\ .$$

Wird das Integral mit der **Simpson-Regel** approximiert, ergibt sich die Näherungsformel

$$y_{i+1} = y_{i-1} + \frac{h}{3}(f(x_{i-1}, y_{i-1}) + 4f(x_i, y_i) + f(x_{i+1}, y_{i+1}))\ , \qquad (2.29)$$

in die drei Stützwerte eingehen, siehe Abbildung 2.8.

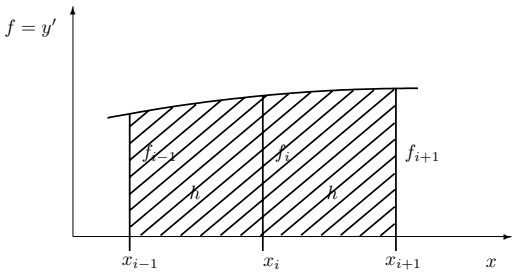


Abb. 2.8. Mehrschrittverfahren (2.29)

Auch die Simpson-Regel führt auf ein implizites Verfahren, d. h. erfordert im Allgemeinen eine iterative Lösung. Weiter erstreckt sich die Formel über meh-

rere Intervalle oder Stützstellen, so dass eine gewisse Anlaufrechnung nötig ist. Neben dem Anfangswert $y_0 = y(a) =: y_a$ benötigt man den Näherungswert y_1 an der Stelle x_1. Man nennt ein solches Verfahren ein **Mehrschrittverfahren**.

Beispiel 2.8 (Instabiles Verfahren, mit MAPLE-Worksheet). In diesem Worksheet wird das Anfangswertproblem

$$y' = -y\ , \qquad y(0) = 1 \tag{2.30}$$

im Intervall $[0, 4]$ mit 10, 20 und 40 Stützstellen berechnet. Als zusätzlicher Startwert wird einfach der Wert der exakten Lösung $y_1 = y(h) = e^{-h}$ vorgegeben. Der relative Fehler

$$R = \frac{y_i - y(x_i)}{y(x_i)}$$

zeigt Oszillationen, deren Amplitude mit x anwächst. Bei einer Verfeinerung der Diskretisierung bleiben die Oszillationen erhalten. Die Amplitude nimmt zwar etwas ab, das Wachstum mit x aber bleibt, und die Frequenz der Oszillationen erhöht sich.

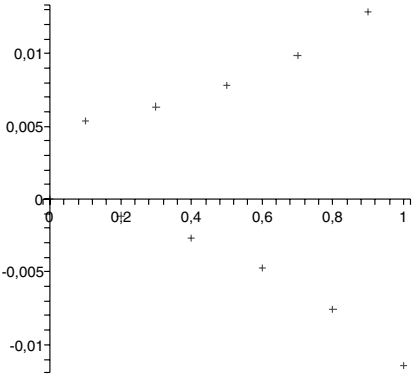


Abb. 2.9. Relativer Fehler beim Verfahren (2.29)

Ein physikalisches Phänomen kann dies nicht sein. Wie wir wissen, ist die eindeutige exakte Lösung monoton fallend. Für praktische Berechnungen ist diese Methode somit nicht brauchbar. □

Bevor es mit der Konstruktion numerischer Methoden weiter gehen kann, muss das Phänomen dieser „Instabilität“ genauer betrachtet werden. Es wird sich im folgenden Kapitel zeigen, dass die in dieser Art konstruierten Methoden **instabil** und damit für die numerische Simulation unbrauchbar sind.

2.2 Stabilität, Konsistenz und Konvergenz

In diesem Abschnitt werden theoretische Aussagen über Näherungsverfahren beschrieben. Diese geben Auskunft über die Qualität der Lösungen eines numerischen Verfahrens. Die Genauigkeit oder die Konsistenz des Verfahrens bezeichnet die Güte der Approximation des Ausgangsproblems. Für die Ausführung des Verfahrens und die praktische Berechnung spielt die Stabilität des Verfahrens eine zentrale Rolle. Die Konvergenz eines Verfahrens bedeutet, dass die Näherungslösung gegen die exakte Lösung konvergiert, wenn die Schrittweite gegen Null strebt.

2.2.1 Stabilität

Um die Stabilität des Näherungsverfahrens beurteilen zu können, muss man die Eigenschaften der gegebenen Differenzialgleichung und ihrer Lösung kennen und berücksichtigen. Erweist sich die numerische Lösung eines Anfangswertproblems als instabil, wie dies im vorigen Abschnitt auftrat, so kann dies an zwei Dingen liegen:

1. Das Anfangswertproblem selbst ist instabil,

2. das numerische Verfahren ist instabil.

Wir benötigen somit Information über das betrachtete Problem. Grundsätzlich setzen wir voraus, dass es sachgemäß gestellt ist:

Eine Differenzialgleichung heißt **sachgemäß gestellt** genau dann, wenn sie eine eindeutige Lösung besitzt, welche stetig von den Daten abhängt. Andernfalls heißt sie **schlecht gestellt**.

Dies ist gewissermaßen eine Grundvoraussetzung für einen physikalischen Vorgang. Die stetige Abhängigkeit von den Daten bezeichnet man schon als Stabilität des mathematischen Modells. Trotzdem gibt es ein sehr unterschiedliches Verhalten von Lösungen.

Beispiel 2.9 (Mit MAPLE-Worksheet). Das Anfangswertproblem

$$y'(x) = y(x) - 3e^{-2x} \,, \tag{2.31}$$

$$y(0) = y_0 \,. \tag{2.32}$$

besitzt die eindeutige Lösung

$$y(x) = e^{-2x} + (y_0 - 1)e^x \ .$$

Wie man sieht, hängt die Lösung stetig von den Anfangsdaten ab, das AWP ist daher sachgemäß gestellt. Für große Werte von x ergibt sich das folgende Verhalten der Lösung

$$\lim_{x\to\infty} y(x) = \begin{cases} -\infty & \text{für } y_0 < 1 \ , \\ 0 & \text{für } y_0 = 1 \ , \\ \infty & \text{für } y_0 > 1 \ . \end{cases} \tag{2.33}$$

Bei der Approximation des Anfangswertproblems (2.31), (2.32) mit $y_0 = 1$ wird schon der kleinste Rundungs- oder Diskretisierungsfehler für große x zu einem qualitativ falschen Verhalten der numerischen Lösung führen. Der Fehler wächst oder fällt exponentiell. □

Probleme, bei denen kleine Störungen sehr schnell anwachsen können, kommen in den Natur- und Ingenieurwissenschaften öfters vor. So reagiert eine brennende Kerze auf einen leichten Windzug mit einem Flackern der Flamme. Die Strömung in der Kaffeetasse beim Eingießen der Milch zeigt Strukturen, welche sich chaotisch verhalten: Kleine Störungen ergeben große Änderungen. Bei der Wettervorhersage kennt man Wetterlagen, bei denen eine Voraussage sehr schwierig ist. Ungenauigkeiten in den Messdaten können schnell anwachsen und führen dazu, dass das Wetter für große Zeiten nur sehr ungenau vorhersagbar ist. Ein Phänomen, welches hinlänglich bekannt ist.

Wenn man die Stabilität und die Eigenschaften eines numerischen Verfahrens untersucht, dann wird man ein Problem als Testbeispiel aussuchen, bei dem Störungen in der exakten Lösungsfunktion nicht schnell anwachsen, am besten sogar kleiner werden. Wendet man auf ein solches Problem das numerische Verfahren an, dann muss diese Eigenschaft der Lösung wiedergegeben werden. Wachsen bei der numerischen Simulation die Störungen an, dann handelt es sich um die **Instabilität des numerischen Verfahrens**. Für praktische Rechnungen wird dieses Verfahren nicht sinnvoll anwendbar sein.

Beispiel 2.10 (Mit MAPLE-Worksheet). Ein **Standardbeispiel** für den Stabilitätstest ist das Anfangswertproblem

$$y'(x) = \lambda y(x) \ , \qquad y(0) = y_0 \tag{2.34}$$

mit y_0, $\lambda \in \mathbb{R}$. Als Beispiel 2.1 wurde dieses als Modell zur Populationsdynamik schon betrachtet. Mit $\lambda = -1$ wurde es als Rechenbeispiel im letzten Abschnitt benutzt.

Die exakte Lösung lautet

$$y(x) = y_0 e^{\lambda x} \ . \tag{2.35}$$

Stören wir die Anfangswerte um eine kleine Größe δ, d.h. betrachten wir das Anfangswertproblem

$$y' = \lambda y \ , \qquad y(0) = y_0 + \delta \ , \tag{2.36}$$

dann ergibt sich analog zu (2.35) die Lösung

$$\widetilde{y}(x) = (y_0 + \delta) e^{\lambda x} \ .$$

Für die Differenz gilt somit

$$y - \widetilde{y} = \delta e^{\lambda x} \ . \tag{2.37}$$

Der Betrag des Fehlers zeigt ein exponentielles Wachstum, falls $\lambda > 0$ und ein Abklingen falls $\lambda < 0$ ist. Der Fall $\lambda = 0$ ist der Fall mit der konstanten Lösung y_0. □

Das Anfangswertproblem (2.34) ist für $\lambda < 0$ also ein sehr gutes Testproblem für ein numerisches Verfahren. Die exakte Lösung besitzt die Eigenschaft, dass ein Fehler in den Daten mit x abklingt. Wird ein numerisches Verfahren auf ein solches Problem angewandt, dann fordert man, dass diese Eigenschaft erhalten bleibt. Das im letzten Abschnitt vorgestellte Verfahren hat für dieses Beispiel mit $\lambda = -1$ einen oszillierenden Fehler mit wachsender Amplitude geliefert. Es ist somit instabil und für praktische Berechnungen unbrauchbar.

Das Anfangswertproblem (2.34) hat nicht nur als Beispiel für den Test numerischer Methoden, sondern auch als sogenannte **Störungsgleichung** eine Bedeutung. Betrachten wir die allgemeine Gleichung

$$y' = f(x, y) \tag{2.38}$$

und untersuchen das Anwachsen einer Störung η in der Lösung y. Das Verhalten der Störung wird bestimmt durch die Differenzialgleichung

$$\widetilde{y}' = f(x, \widetilde{y}) \quad \text{mit} \quad \widetilde{y} := y + \eta \ . \tag{2.39}$$

Ziehen wir (2.38) von dieser Gleichung ab, ergibt sich

$$\eta' = f(x, y + \eta) - f(x, y) \ .$$

Setzt man für den ersten Term auf der rechten Seite eine Taylor-Entwicklung bezüglich des zweiten Arguments um (x, y) ein, erhält man

$$\eta' = \eta f_y(x, y) + \frac{1}{2}\eta^2 f_{yy}(x, y) + \dots \ .$$

Wir betrachten kleine Störungen η und vernachlässigen alle Terme mit Potenzen von η gleich und größer zwei. Damit ergibt sich die Differenzialgleichung für kleine Störungen zu

$$\eta' = \eta f_y \ . \tag{2.40}$$

Betrachten wir weiter f_y für kleine η als konstant, dann ergibt sich genau die Differenzialgleichung (2.34) mit der Lösung

$$\eta(x) = \eta_0 e^{f_y x} \ . \tag{2.41}$$

Ist $f_y < 0$, dann werden kleine Störungen in der exakten Lösung abklingen.

Bemerkung: Bei dieser Störungsrechnung und Linearisierung wird immer angenommen, dass die Störungen klein sind und die entsprechenden Terme höherer Ordnung vernachlässigt werden können. Für Probleme, bei denen nichtlineare Terme wesentlich die Lösung bestimmen, gilt dies nur in einer kleinen Umgebung um die Störung. Es müssen ansonsten nichtlineare Betrachtungen zur Stabilität durchgeführt werden.

Zur Untersuchung der **Stabilität des numerischen Verfahrens** geht man ähnlich zum kontinuierlichen Fall vor. Das Euler-Cauchy-Verfahren wird auf das Anfangswertproblem ohne und mit gestörten Anfangswerten angewandt. Die Differenz der beiden Approximationen lautet

$$\eta_{i+1} = \eta_i + h(f(x_i, y_i) - f(x_i, y_i + \eta_i)) \ , \qquad i = 0, 1, 2, \dots \ .$$

Wird die Differenz der verschiedenen Funktionswerte von f wieder mit Hilfe des Zwischenwertsatzes ersetzt, ergibt sich

$$\eta_{i+1} = \eta_i + h f_y \, \eta_i = (1 + h \, f_y)\eta_i \ ,$$

wobei f_y an einer Zwischenstelle auszuwerten ist. Wir nehmen an, dass hf_y konstant ist, etwa indem durch eine Schrittweitensteuerung h in jedem Schritt geeignet gewählt wird. Mit dem sogenannten **Verstärkungsfaktor** $s := 1 + h \, f_y$ ergibt sich dann

$$\eta_{i+1} = s\eta_i = s^2\eta_{i-1} = \cdots = s^{i+1}\eta_0 \ .$$

Der Fehler klingt somit ab, wenn $|s| < 1$ gilt, bzw.

$$|1 + h \, f_y| < 1 \ .$$

Wegen $h > 0$ führt dies auf die Bedingung

$$f_y < 0 \qquad \text{und} \qquad -2 < h \, f_y \ .$$

Diese Untersuchung nennt man eine **Stabilitätsanalyse des numerischen Verfahrens**. Die Eigenschaft der exakten Lösung einer Differenzialgleichung mit $f_y < 0$, dass nach der linearisierten Stabilitätsanalyse Störungen abklingen, bleibt unter der Schrittweitenbedingung $-2 < h \, f_y < 0$ erhalten. Man sagt:

Das Euler-Cauchy-Verfahren ist **stabil**, falls die Schrittweitenbedingung

$$\kappa = h\,|f_y| < 1 \tag{2.42}$$

erfüllt ist.

Bemerkung: Die so definierte Stabilität nennt man auch **A-Stabilität**. Es gibt weitere Stabilitätsbegriffe, die insbesondere für die Anwendung auf Probleme, deren Lösungen sich stark ändern, wichtig sind. Dies sind z.B. Probleme, bei denen die Nichtlinearität von f eine wesentliche Rolle für das Lösungsverhalten spielen. Die Forderung an die Stabilität des Verfahrens hängt dabei durchaus von dem betrachteten Problem ab. Wir werden dies hier nicht vertiefen und auf die Spezialliteratur verweisen. Eine kurze Diskussion von sogenannten steifen Problemen ist im Abschnitt Bemerkungen und Entscheidungshilfen ausgeführt.

Wir wenden nun das Mehrschrittverfahren (2.29), welches in Kapitel 2.1 konstruiert wurde, direkt auf die Störungsgleichung an. Die diskretisierte Störungsgleichung lautet

$$\eta_{i+1} = \eta_{i-1} + \frac{h}{3}(\eta_{i-1}f_y + 4\eta_i f_y + \eta_{i+1}f_y)\ .$$

Wir machen den Ansatz $\eta_i = s^i\eta_0$ mit dem Verstärkungsfaktor s und setzen dies oben ein. Nach der Division mit $s^{i-1}\eta_0$ ergibt sich die quadratische Gleichung

$$s^2 = 1 + \frac{1}{3}hf_y(1 + 4s + s^2).$$

Diese quadratische Gleichung besitzt die Lösungen

$$s_1 = \frac{1}{3 - hf_y}(2hf_y + \sqrt{3(3 + (hf_y)^2)})\ ,$$
$$s_2 = \frac{1}{3 - hf_y}(2hf_y - \sqrt{3(3 + (hf_y)^2)})\ ,$$

welche in Abbildung 2.2.1 grafisch als Funktion von hf_y-Ebene dargestellt ist. Diese zwei Lösungen überlagern sich zur gesamten Störlösung bzw. dem Verstärkungsfaktor $s = c_1 s_1 + c_2 s_2$. Im Fall $f_y < 0$, d.h. einem Ausgangsproblem, bei dem die Störungen abklingen, ergibt sich

$$|s_2| > 1\ .$$

Es kann somit einer der beiden Anteile in den Störlösungen anwachsen und zu dem oszillierenden Verhalten führen. Das numerische Verfahren (2.29) ist für praktische Probleme somit instabil und unbrauchbar.

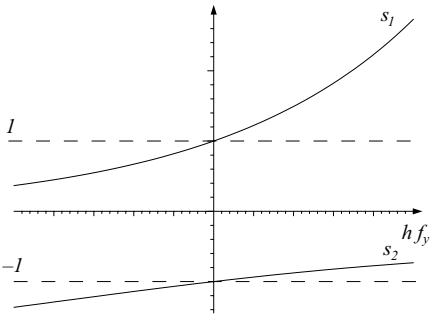


Abb. 2.10. Stabilitätsbereich des Verfahrens (2.29)

2.2.2 Konsistenz

Neben der Stabilität von Näherungsverfahren spielt die **Genauigkeit** des numerischen Verfahrens für die Praxis eine wichtige Rolle. Über die Genauigkeit der Verfahren von *Euler-Cauchy* und *Heun* haben wir uns bei der Definition der Verfahren schon Gedanken gemacht. Wir haben bei den numerischen Ergebnissen gesehen, dass der Fehler mit h beziehungsweise mit h^2 bei einer Verfeinerung der Gitters abnimmt und haben dies auch für einen Rechenschritt unter Benutzung einer Taylor-Entwicklung bestätigt. Im Folgenden werden wir dies verallgemeinern und erweitern.

Wir beginnen mit dem schon in dem vorausgehenden Abschnitt betrachteten lokalen Fehler, welcher in einem einzelnen Schritt des numerischen Verfahrens entsteht, und definieren:

Setzt man $y_i = y(x_i)$ voraus, dann heißt die Differenz

$$\varepsilon_{i+1} = y(x_{i+1}) - y_{i+1} \tag{2.43}$$

lokaler Diskretisierungs- oder **Verfahrensfehler** im Gitterpunkt x_{i+1}.

Wie schon im vorigen Abschnitt bezeichnet $y(x_{i+1})$ die exakte Lösung am Punkt x_{i+1} und y_{i+1} den Wert des numerischen Verfahrens an diesem Punkt. Betrachtet man das *Euler-Cauchy-Verfahren*, so erhält man

$$\begin{aligned}\varepsilon_{i+1} &= y(x_{i+1}) - y_{i+1} \\ &= y(x_{i+1}) - y(x_i) + hf(x_i, y(x_i)).\end{aligned}$$

Der lokale Diskretisierungsfehler ergibt sich somit als das **Residuum** oder den **Defekt**, wenn die exakte Lösung in das numerische Verfahren eingesetzt wird.

Wird eine Taylor-Entwicklung um den Punkt $(x_i, y(x_i))$ in die Gleichung eingesetzt und nach Potenzen von h geordnet, so erhält man die Information, wie der lokale Diskretisierungsfehler mit der Schrittweite h gegen Null strebt. Dies führt auf die Definition:

Ein numerisches Verfahren besitzt die **Konsistenzordnung p**, wenn für den lokalen Diskretisierungsfehler

$$|\varepsilon_{i+1}| \leq C\ h^{p+1}\ , \qquad i = 0, 1, 2, \ldots \tag{2.44}$$

mit einer von h und x unabhängigen Konstanten C gilt.

Nach unserer Rechnung im Kapitel 2.1 ergibt sich somit die folgende Aussage:

Das **Euler-Cauchy-Verfahren** besitzt die **Konsistenzordnung 1**, das **Verfahren von Heun** besitzt die **Konsistenzordnung 2**.

2.2.3 Konvergenz

Der lokale Diskretisierungsfehler sagt etwas über die Genauigkeit des Verfahrens aus, also über die Fehler, welche in einem Schritt des Verfahrens auftreten. Wenn diese Fehler im Laufe der Rechnung stark anwachsen - auch bei Problemen, deren exakte Lösung diese Eigenschaft nicht besitzt - oder bei der Verfeinerung des Gitters nicht kleiner werden, dann ist das Verfahren für praktische Probleme unbrauchbar: Der akkumulierte oder globale Fehler wird groß.

Die Differenz

$$e_{i+1} = y(x_{i+1}) - y_{i+1} \tag{2.45}$$

heißt der **globale Diskretisierungs-** oder **Verfahrensfehler im Punkt $\mathbf{x_{i+1}}$**.

Dies ist der akkumulierte Fehler, welcher bei Berücksichtigung der vorausgegangenen Fehler tatsächlich im Punkt x_{i+1} auftritt. Beim Euler-Cauchy-Verfahren ist dieser Fehler proportional h, während der lokale Fehler proportional h^2 ist. Dies wurde schon bei Definition der Konsistenzordnung mit p

und nicht mit $p+1$ berücksichtigt. Dieser Zusammenhang ist auch für andere Näherungsverfahren typisch. Die Größe des globalen Verfahrensfehlers sagt etwas aus über die Konvergenz der Näherung gegen die exakte Lösung, wenn die Schrittweite h gegen Null strebt.

Ein numerisches Verfahren heißt **konvergent**, wenn für den globalen Diskretisierungsfehler

$$\max_{i=1,2,\ldots,n} |e_i| \to 0 \quad \text{für } h \to 0$$

gilt. Es hat die **Konvergenzordnung p**, wenn

$$\max_{i=1,2,\ldots,n} |e_i| \leq K\ h^p \tag{2.46}$$

mit einer von h unabhängigen Konstanten K gilt.

Der lokale Diskretisierungsfehler gibt Auskunft darüber, welche Fehler in einem Schritt des Näherungsverfahrens auftreten. Die Stabilität ist ein Maß dafür, wie schnell Störungen anwachsen können. Es verwundert somit nicht, dass mit einer geeigneten Definition der Stabilität der folgende Sachverhalt gilt:

Stabilität	+	**Konsistenz**	=	**Konvergenz**

Weitere Aussagen zu diesen Begriffen der numerischen Analysis, verschiedene Definitionen der Stabilität findet man z.B. in den Büchern von Strehmel und Weiner [62], und Hairer und Wanner [29].

Bemerkung: Für ein Verfahren mit der Konvergenzordnung p strebt der Fehler gegen Null mit h^p, wenn die Schrittweite gegen Null strebt, d.h. das Gitter verfeinert wird. Wie bei der numerischen Integration zeigt sich dies bei praktischen Berechnungen. Wird die Schrittweite sehr klein, dann werden die Rundungsfehler durch die Zahlendarstellung auf dem Rechner das Ergebnis beeinflussen. Bei der numerischen Lösung von Anfangswertproblemen ist dieser Sachverhalt noch deutlicher, da Rundungsfehler an einem Gitterpunkt x_i wieder die Steigungsberechnung an allen weiteren Gitterpunkten beeinflussen. Es kommt somit ein lokaler und auch akkumulierter Rundungsfehler zu dem Gesamtfehler hinzu. Während für $h \to 0$ der Verfahrensfehler gegen Null konvergiert, wächst der Rundungsfehler an und divergiert gegen unendlich. Diese Situation ist in Abbildung 0.1 im Kapitel 1 schematisch dargestellt. Eine Merkregel ist, dass die Schrittweite nicht kleiner als die Wurzel der Maschinengenauigkeit gewählt werden sollte.

2.3 Mehrschrittverfahren

In Abschnitt 2.1 haben wir gesehen, dass die einfache Übertragung der Newton-Cotes-Formeln für die numerische Integration auf die Integralgleichung eines Anfangswertproblems zu einem instabilen Verfahren führt. Es zeigt sich allgemein, dass die Integration wie bei der Herleitung des Euler-Cauchy-Verfahrens nur über ein einzelnes Gitterintervall ausgeführt werden darf:

$$y(x_{i+1}) = y(x_i) + \int_{x_i}^{x_{i+1}} f(x, y(x))\ dx\ . \tag{2.19}$$

Um ein genaueres Verfahren zu bekommen, kann man allerdings den Integranden mit einem Interpolationspolynom höherer Ordnung approximieren, welches durch mehr als zwei Stützstellen verläuft. Man betrachtet somit ein Näherungsverfahren der Form

$$y_{i+1} = y_i + \int_{x_i}^{x_{i+1}} p_n(x)\ dx\,, \tag{2.47}$$

wobei p_n ein Polynom n-ten Grades ist.

Werden nur bekannte Werte für das Approximationspolynom $p_n(x)$ benutzt, dann ergibt sich eine explizite Formel. Für ein Polynom n-ten Grades wird gefordert, dass

$$p_{n,i}(x_{i-j}) = f_{i-j} \quad \text{für} \quad j = 0, 1, \ldots, n$$

gilt. Mit der Lagrangesche Interpolationsformel erhält man für das Polynom die Darstellung

$$p_{n,i}(x) = \sum_{j=0}^{n} f_{i-j} L_j(x)\ , \tag{2.48}$$

wobei L_j das entsprechende Lagrangesche Stützpolynom bezeichnet (siehe Anhang A.1).

Beispiel 2.11. Der einfachste Fall ist ein Zweischrittverfahren. Mit den Lagrangeschen Stützpolynomen

$$L_0(x) = \frac{x - x_i}{x_{i-1} - x_i} \quad \text{und} \quad L_1(x) = \frac{x - x_{i-1}}{x_i - x_{i-1}} \tag{2.49}$$

ergibt sich das Interpolationspolynom vom Grad $n = 1$ zu

$$p_{1,i}(x) = f_{i-1}\ \frac{x - x_i}{x_{i-1} - x_i} + f_i\ \frac{x - x_{i-1}}{x_i - x_{i-1}} = f_{i-1}\ \frac{1}{h}(x - x_i) + f_i\ \frac{1}{h}(x - x_{i-1})\ .$$

Die Integration liefert nach Gleichung (2.47)

$$\int_{x_i}^{x_{i+1}} p_{1,i}(x)\ dx = \left[f_{i-1}\ \frac{1}{h}\ \frac{1}{2}(x - x_i)^2 + f_i\ \frac{1}{h}\ \frac{1}{2}(x - x_{i-1})^2\right]_{x_i}^{x_{i+1}}.$$

Da $x_{i+1} - x_i = h$ und $x_{i+1} - x_{i-1} = 2h$ lautet das Verfahren somit

$$y_{i+1} = y_i + \frac{h}{2}(-f_{i-1} + 3f_i) \quad i = 1, 2, .. \tag{2.50}$$

In das Zweischrittverfahren gehen als Daten die Werte aus den zwei zurückliegenden Stützstellen ein. □

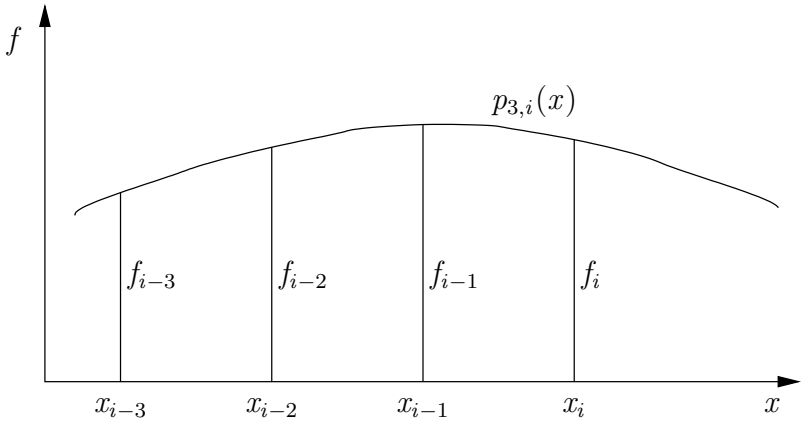


Abb. 2.11. Adams-Bashforth-Verfahren

Man nennt ein solchermaßen konstruiertes Verfahren ein **Adams-Bashforth-Verfahren**. In Tabelle 2.1 sind die Koeffizienten dieser Verfahren für die Polynomgrade $n = 1, 2, \ldots, 6$ aufgeführt. Die Koeffizienten ergeben sich aus der Bedingung

$$\int_{x_i}^{x_{i+1}} p_{n,i}(x)\, dx = a_{i-n} f_{i-n} + a_{i-n+1} f_{i-n+1} + \ldots + a_{i-1} f_{i-1} + a_i f_i. \tag{2.51}$$

Die globale Konsistenzordnung dieser Verfahren ist $n+1$, wenn n den Grad des Interpolationspolynoms bezeichnet. In Abbildung 2.11 ist die Situation für den Polynomgrad $n = 3$ aufgezeichnet. Zur Berechnung von y_{i+1} werden hier die Funktionswerte von f an den Stützstellen x_{i-3}, x_{i-2}, x_{i-1} und x_i benutzt.

	a_{i-6}	a_{i-5}	a_{i-4}	a_{i-3}	a_{i-2}	a_{i-1}	a_i
$n=1$:						$-\frac{1}{2}$	$\frac{3}{2}$
$n=2$:					$\frac{5}{12}$	$-\frac{16}{12}$	$\frac{23}{12}$
$n=3$:				$-\frac{9}{24}$	$\frac{37}{24}$	$-\frac{59}{24}$	$\frac{55}{24}$
$n=4$:			$\frac{251}{720}$	$-\frac{1274}{720}$	$\frac{2616}{720}$	$-\frac{2774}{720}$	$\frac{1901}{720}$
$n=5$:		$-\frac{475}{1440}$	$\frac{2877}{1440}$	$-\frac{7298}{1440}$	$\frac{9982}{1440}$	$-\frac{7923}{1440}$	$\frac{4277}{1440}$
$n=6$:	$\frac{19087}{60480}$	$-\frac{134472}{60480}$	$\frac{407139}{60480}$	$-\frac{688256}{60480}$	$\frac{705549}{60480}$	$-\frac{447288}{60480}$	$\frac{198721}{60480}$

Tabelle 2.1. Explizite Mehrschrittverfahren (Adams-Bashforth-Verfahren)

Im MAPLE-**Worksheet** zu den Adams-Bashforth-Verfahren werden die Formeln für beliebiges n erzeugt. Dabei sind auch unterschiedliche Schrittweiten zwischen den Stützstellen erlaubt.

Wird das Stützpaar (x_{i+1}, f_{i+1}) mit in die Interpolation einbezogen, dann ergibt sich ein implizites Verfahren. Man nennt diese Klasse von Verfahren die **Adams-Moulton-Verfahren**. Greifen wir das Beispiel mit dem Zweischrittverfahren nochmals auf und suchen jetzt ein Polynom durch die Punkte

$$(x_{i-1}, f_{i-1}), \quad (x_i, f_i) \quad \text{und} \quad (x_{i+1}, f_{i+1})\,.$$

Dies lässt sich durch eine Parabel bewerkstelligen. Mit der analogen Rechnung unter Zuhilfenahme der Lagrangeschen Interpolationsformel erhält man ein numerisches Verfahren in der Form

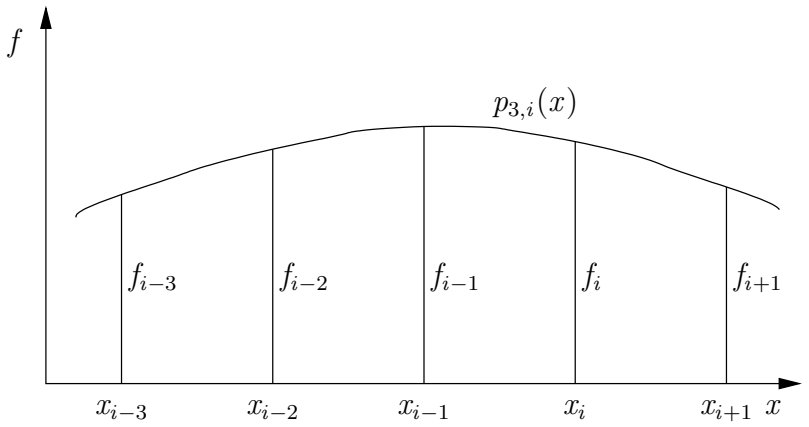


Abb. 2.12. Adams-Moulton-Verfahren

$$y_{i+1} = y_i + \frac{h}{12}(-f_{i-1} + 8f_i + 5f_{i+1}) \quad i = 1, 2, .. \tag{2.52}$$

Nur für eine lineare oder einfache nichtlineare Funktion f ist hier eine explizite Auflösung nach y_{i+1} möglich. Im Allgemeinen muss die Gleichung (2.52) iterativ gelöst werden. Die Koeffizienten der Adams-Moulton-Verfahren sind gemäß dem Ansatz

$$\int_{x_i}^{x_{i+1}} p_{n,i}(x)\, dx = a_{i-n+1}f_{i-n+1} + \ldots + a_{i-1}f_{i-1} + a_i f_i + a_{i+1}f_{i+1} \tag{2.53}$$

in Tabelle 2.2 für die Polynomgrade $n = 2, 3, \ldots, 7$ aufgeführt. Die Konsistenzordnung der Adams-Moulton-Verfahren ist ebenso $n + 1$.

Im **Worksheet** zu den Adams-Moulton-Verfahren werden die Formeln für beliebiges n erzeugt. Dabei können auch unterschiedliche Schrittweiten zwischen den Stützstellen gewählt werden.

Bemerkung: Ein Nachteil dieser Mehrschrittverfahren ist, dass im ersten Schritt nicht nur die Anfangswerte eingehen. Es ist ein Anlaufstück erforderlich, für das Näherungswerte mit Hilfe eines anderen Verfahrens bestimmt werden müssen. In dem betrachteten Beispiel des Zweischrittverfahrens muss neben dem Anfangswert der Wert am Punkt x_1 bereitgestellt werden. Da-

a_{i-6}	a_{i-5}	a_{i-4}	a_{i-3}	a_{i-2}	a_{i-1}	a_i	a_{i+1}
					$-\frac{1}{12}$	$\frac{8}{12}$	$\frac{5}{12}$
				$\frac{1}{24}$	$-\frac{5}{24}$	$\frac{19}{24}$	$\frac{9}{24}$
			$-\frac{19}{720}$	$\frac{106}{720}$	$-\frac{264}{720}$	$\frac{646}{720}$	$\frac{251}{720}$
		$\frac{27}{1440}$	$-\frac{173}{1440}$	$\frac{482}{1440}$	$-\frac{798}{1440}$	$\frac{1427}{1440}$	$\frac{475}{1440}$
	$-\frac{863}{60480}$	$\frac{6312}{60480}$	$-\frac{20211}{60480}$	$\frac{37504}{60480}$	$-\frac{46461}{60480}$	$\frac{65112}{60480}$	$\frac{19087}{60480}$
$\frac{1375}{120960}$	$-\frac{11351}{120960}$	$\frac{41499}{120960}$	$-\frac{88547}{120960}$	$\frac{123133}{120960}$	$-\frac{121797}{120960}$	$\frac{139849}{60480}$	$\frac{36799}{120960}$

Tabelle 2.2. Implizite Mehrschrittverfahren (Adams-Moulton-Verfahren)

bei sollte das Verfahren für das Anlaufstück die gleiche lokale Fehlerordnung besitzen.

Wie schon bei dem Verfahren von Heun diskutiert wurde, vermeidet man die Iteration der impliziten Verfahren, indem man einen Prädiktor-Korrektor-Ansatz wählt.

Prädiktor-Korrektor-Verfahren: Eine Adams-Bashforth-Formel wird mit einer impliziten Adams-Moulton-Formel als Korrektor kombiniert.

Dabei empfiehlt es sich, eine Korrektor-Formel mit einer um eins höheren Fehlerordnung zu nehmen. Man führt somit nur einen Iterationsschritt im Iterationsverfahren für das implizite Adams-Moulton-Verfahren aus. Die Schrittweite h darf dann nicht zu groß gewählt werden. Gewissermaßen ist das Verfahren von Heun der einfachste Vertreter dieser Klasse von Prädiktor-Korrektor-Verfahren.

Das Beispiel $y' = -y$ ist sehr gut geeignet, die Stabilität von Näherungsverfahren zu untersuchen. Um am praktischen Beispiel die Genauigkeit von Verfahren aufzuzeigen, gehen wir zu einem etwas schwierigeren Problem über.

Beispiel 2.12 (Mit MAPLE-Worksheet). Gesucht ist die Lösung des Anfangswertproblems

$$y'(x) = y^2(x)\sin(x)\,, \qquad y(0) = -4 \tag{2.54}$$

im Intervall $[0, 0.5]$. Die rechte Seite ist nichtlinear und hängt von x und y ab. Die exakte Lösung lässt sich durch die Trennung der Veränderlichen berechnen. Es folgt aus (2.54) für $y \neq 0$

$$\frac{1}{y^2}\frac{dy}{dx} = \sin(x)$$

und damit weiter

$$\frac{1}{y^2}\,dy = \sin(x)\,dx\,.$$

Die Integration liefert

$$-\frac{1}{y} = -\cos(x) + C.$$

Wird die Konstante C über die Anfangsbedingungen bestimmt, ergibt sich die Lösung zu

$$y = \frac{1}{\cos(x) - \frac{5}{4}}\,.$$

Dieses Problem wird mit einem Adams-Bashforth-Verfahren 3. Ordnung und der Schrittweite $h = 0.1$ im Worksheet gelöst. Für das Anlaufstück haben wir hier einfach die exakte Lösung vorgeschrieben. Dabei wurden die künstlichen Stützstellen

$$x_{-1} = -h\,, \qquad x_{-2} = -2h$$

eingeführt, so dass das Verfahren beim Anfangswert $x = 0$ startet. In Abbildung 2.13 ist die exakte Lösung und die Näherungswerte aufgezeichnet. Es ergibt sich für die Lösung im Punkte x_5

$$y_5 = -2.6745658\,, \quad \text{rel. Fehler} = 0.010593088\,.\square$$

Beispiel 2.13 (Mit MAPLE-Worksheet). In diesem Worksheet wird das Adams-Moulton-Verfahren 3. Ordnung mit dem Adams-Bashforth-Verfahren 2. Ordnung als Prädiktor kombiniert. Es ergeben sich mit diesem Prädiktor-Korrektor-Verfahren etwas bessere Werte:

$$y_5 = -2.69020826\,, \quad \text{rel. Fehler} = 0.0049492\,.$$

Man sieht, dass dieses Verfahren die 3. Ordnung besitzt, wenn man die Schrittweite mit dem Faktor 10 verkleinert. Das Ergebnis wird dann mit Hilfe von 50 Stützstellen berechnet und hat den Wert

$$y_5 = -2.6851624\,, \quad \text{rel. Fehler} = 0.0000035\,.$$

Der Fehler wurde um den Faktor 10^{-3} kleiner. $\square$

Im Allgemeinen hat man natürlich keine exakte Lösung für das Anlaufstück zur Verfügung. Im nächsten Abschnitt werden wir Einschrittverfahren höherer Ordnung kennen lernen, mit denen man die Werte des Anlaufstückes berechnen kann.

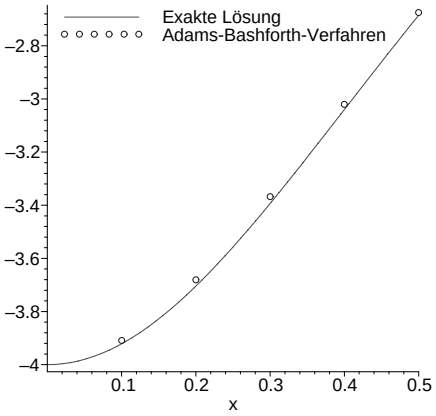


Abb. 2.13. Exakte Lösung und Näherungslösung mit dem Adams-Bashforth-Verfahren für Beispiel 2.12

2.4 Runge-Kutta-Verfahren

Neben den Mehrschrittverfahren lassen sich jedoch auch Einschrittverfahren mit einer Fehlerordnung größer als zwei konstruieren. Die für die höhere Ordnung erforderlichen Werte an den Stützstellen werden als Hilfswerte im Innern des Intervalls erzeugt. Sie sind nicht Teil der Lösung und werden nach Ausführung des Schrittes wieder vergessen.

Der Ansatz eines solchen so genannten **Runge-Kutta-Verfahrens** ist

$$y_{i+1} = y_i + h \sum_{j=1}^{n} a_j k_j \ , \tag{2.55}$$

wobei die k_j die Funktionswerte von f im Innern des Gitterintervalls $[x_i, x_{i+1}]$ approximieren und a_j geeignete Koeffizienten sind.

Den einfachsten Ansatz erhält man durch die Berücksichtigung von zwei Hilfswerten:

$$y_{i+1} = y_i + h\,(a_1 k_1 + a_2 k_2) \ , \tag{2.56}$$

wobei wir annehmen

$$k_1 := f_i, \ \ k_2 = f(x_i + \alpha h, \ y_i + \beta\, h k_1) \ . \tag{2.57}$$

Die Wahl von k_1 liegt auf der Hand, da $f_i = f(x_i, y_i)$ bekannt ist, k_2 ist als Wert an einer beliebigen Zwischenstelle gewählt. Die freien Parameter sind a_1, a_2, α und β. Es geht als nächstes darum, geeignete Werte für diese Parameter zu finden.

Schritt 1: Eine Taylor-Entwicklung liefert zunächst

$$y(x_{i+1}) = y(x_i) + hy'(x_i) + \frac{h^2}{2!}y''(x_i) + \dots \ . \tag{2.58}$$

Die Ableitungen von y lassen sich durch die rechte Seite f ausdrücken. So gilt

$$\begin{aligned} y'(x) &= f(x, y(x)) \\ y''(x) &= \frac{d}{dx}f'(x, y(x)) = \frac{\partial f}{\partial x}(x, y(x)) + \frac{\partial f}{\partial y}(x, y(x))y'(x) \\ &= (f_x + f\, f_y)(x, y(x)) \ , \end{aligned}$$

wobei hier die Kettenregel angewandt werden muss und die abkürzende Schreibweise f_x bzw. f_y für die partiellen Ableitungen von f nach dem ersten bzw. zweiten Argument wiederum benutzt wurde. Wird dies in die obige Taylor-Entwicklung (2.58) eingesetzt, erhält man

$$y(x_{i+1}) = y(x_i) + hf(x_i, y(x_i)) + \frac{h^2}{2!}(f_x + f\, f_y)(x_i, y(x_i)) + \frac{h^3}{3!}y'''(\xi)$$

mit einer Zwischenstelle ξ aus dem Intervall $(x_i,\ x_{i+1})$.

Schritt 2: Eine Taylor-Entwicklung von k_2 um $f(x_i, y_i)$ ergibt

$$k_2 = f(x_i, y_i) + \alpha h f_x(x_i, y_i) + \beta h f_y(x_i, y_i) + \mathcal{O}(h^2) \ .$$

Wird dies in die Formel für das Runge-Kutta-Verfahren eingesetzt, erhält man

$$y_{i+1} = y_i + h(a_1 + a_2)f(x_i, y_i) + h^2(a_2\alpha f_x(x_i, y_i) + a_2\beta f\, f_y(x_i, y_i)) + \mathcal{O}(h^3) \ .$$

Schritt 3: Der dritte Schritt ist der Koeffizientenvergleich der beiden Entwicklungen aus Schritt 1 und Schritt 2. Man erhält ein Verfahren zweiter Ordnung, wenn die folgenden Gleichungen

$$a_1 + a_2 = 1, \quad \alpha a_2 = \frac{1}{2}, \quad \beta a_2 = \frac{1}{2}$$

erfüllt sind. Aus den beiden letzteren Gleichungen folgt sofort $\alpha = \beta$. Danach bleiben zwei Gleichungen für drei Unbekannte übrig, so dass die Lösung nicht eindeutig bestimmt ist.

Die Lösung $a_1 = 0$, $a_2 = 1$, $\alpha = \beta = 0.5$ führt auf das Verfahren

$$y_{i+1} = y_i + h\, f\left(x_i + \frac{h}{2}, y_i + \frac{h}{2} f(x_i, y_i)\right) , \tag{2.59}$$

welches schon in Abschnitt 2.1 als **verbessertes Euler-Cauchy-Verfahren** auftauchte. Die Berechnung in einem Rechenprogramm könnte in der Form des Runge-Kutta-Verfahrens entsprechend der Abfolge

$$\begin{aligned} k_1 &= f(x_i, y_i) , \\ k_2 &= f\left(x_i + \frac{h}{2}, y_i + \frac{h}{2} k_1\right) , \\ y_{i+1} &= y_i + h k_2 \end{aligned}$$

eingesetzt werden.

Die Lösung $a_1 = a_2 = 0.5$, $\alpha = \beta = 1$ führt auf das Verfahren

$$y_{i+1} = y_i + \frac{h}{2} \left(f(x_i, y_i) + f\left(x_{i+1}, y_i + h f(x_i, y_i)\right)\right) , \tag{2.60}$$

welches gerade das **Verfahren von Heun** darstellt. Die Berechnung im Rechenprogramm würde hier in der Form

$$\begin{aligned} k_1 &= f(x_i, y_i) , \\ k_2 &= f(x_i + h, y_i + h k_1) , \\ y_{i+1} &= y_i + \frac{h}{2}(k_1 + k_2) \end{aligned}$$

durchgeführt.

Runge-Kutta-Verfahren höherer Ordnung können analog über den Taylor-Abgleich abgeleitet werden. Ein **Verfahren dritter Ordnung** erhält man mit dem Ansatz

$$y_{i+1} = y_i + h(a_1 k_1 + a_2 k_2 + a_3 k_3) \tag{2.61}$$

mit

$$\begin{aligned} k_1 &= f(x_i, y_i)\,, \\ k_2 &= f(x_i + \alpha h, y_i + \alpha h k_1)\,, \\ k_3 &= f(x_i + \beta h, y_i + \gamma_1 h k_1 + \gamma_2 h k_2)\,. \end{aligned}$$

Für k_2 wurde schon die Konsistenzbedingung, wie sie bei den Verfahren zweiter Ordnung als $\alpha = \beta$ herauskam, eingesetzt. Für k_3 muss aus Konsistenzgründen $\beta = \gamma_1 + \gamma_2$ gelten. Der Taylor-Abgleich wird in diesem Falle deutlich umfangreicher, da die dritten Ableitungen von y bzw. die zweiten Ableitungen von $f(x, y(x))$ eingehen, die mit Hilfe von Produkt- und Kettenregel berechnet werden müssen. Wir wollen hier auf diese Rechnung verzichten. Für die sieben freien Parameter ergeben sich fünf Gleichungen, so dass auch in diese Fall mehrere Varianten existieren. In Tabelle 2.4 sind die gebräuchlichsten Runge-Kutta-Verfahren aufgelistet und im MAPLE-**Worksheet** zum Runge-Kutta-Verfahren werden die Koeffizienten gemäß dem oben diskutierten Vorgehen berechnet. Es zeigt sich, dass beim Verfahren 5. Ordnung sechs Zwischenwerte k_i berechnet werden müssen. Diese Tendenz setzt sich bei noch höherer Ordnung fort: Es werden mehr Zwischenwerte benötigt als der Wert der Ordnung und die Verfahren werden ineffizient.

Beispiel 2.14 (Runge-Kutta-Verfahren, mit MAPLE-Worksheet). Wir greifen das Beispiel (2.54) wieder auf. Eine Lösung des Anfangswertproblems

$$y'(x) = y^2(x)\sin(x)\,, \qquad y(0) = -4 \tag{2.62}$$

wird in dem MAPLE-Worksheet mit dem Runge-Kutta-Verfahren (3. Ordnung, Typ a) gelöst.

Die Schrittweite ist wieder $h = 0.1$. Nach fünf Schritten ergibt sich bei $x = 0.5$ die Näherungslösung und der relative Fehler zu

$$y_5 = 2.6858695\,, \quad \text{rel. Fehler} = 0.0007106\,.$$

In Abbildung 2.14 sieht man bei einer grafischen Darstellung fast keinen Unterschied zur exakten Lösung.

Ein Struktogramm für das klassische Verfahren vierter Ordnung ist in Abbildung 2.15 angegeben. □

Bemerkung: Die Runge-Kutta-Verfahren lassen sich als Einschrittverfahren auch zur Berechnung des Anlaufstückes bei Mehrschrittverfahren einsetzen. Die ersten Schritte werden mit dem Runge-Kutta-Verfahren ausgeführt. Danach startet das Prädiktor-Korrektor-Verfahren.

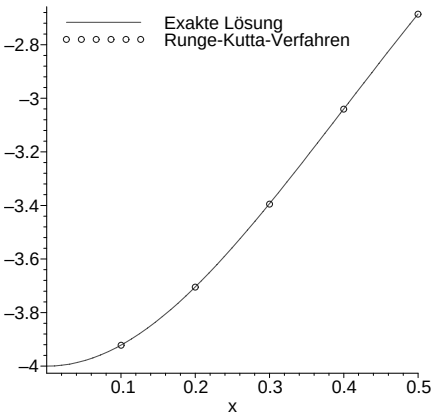


Abb. 2.14. Exakte und numerische Lösung mit dem Runge-Kutta-Verfahren

2.5 Extrapolationsverfahren

Aus den Taylor-Entwicklungen der Runge-Kutta-Verfahren können auch Fehlerterme abgeleitet werden. Um eine Abschätzung des Fehlers zu bekommen, muss man allerdings höhere Ableitungen berechnen. Zudem kennt man die Zwischenstelle ξ natürlich nicht, d.h. diese Ableitungen müssen nach oben und unten abgeschätzt werden. Für praktische Rechnungen ist dies nicht geeignet.

Auf der anderen Seite ist man oft in der Situation, dass man eine gewisse Genauigkeit der Lösung verlangt, z.B. auf 4 oder 5 Stellen genau. Die Ordnung eines Verfahrens ist dabei nur ein Hinweis auf die Güte des Verfahrens und liefert nur die Information, wie sich der Fehler verhält, wenn die Schrittweite gegen Null strebt. Man hat jedoch keine Auskunft darüber, wie groß der tatsächliche Fehler bei der gerade benutzten Schrittweite ist. Bei der numerischen Integration hatten wir mit dem Romberg-Verfahren gesehen, dass mit Hilfe einer Fehlerextrapolation die Möglichkeit besteht, die Genauigkeit über eine Fehlerberechnung zu steuern. Die Methode lässt sich auf Anfangswertprobleme übertragen.

Um das Extrapolationsverfahren effizient durchführen zu können, sollte wieder eine quadratische Entwicklung des Fehlers existieren. Wir nehmen an, dass

$$y_{i+1} - y(x_{i+1}) = A_1 h^2 + A_2 h^4 + A_3 h^6 + \dots \,, \tag{2.63}$$

gilt, mit Koeffizienten A_i, welche unabhängig von h sind. Im Falle des Romberg-Verfahrens für die numerische Integration wurde dies von der Seh-

Real $k_1, k_2, k_3, k_4, x, y, b, h$
Integer n, i
Schreibe auf Bildschirm: „Wie weit soll gerechnet werden, wie viele Schritte sollen genommen werden?“
Lies b ein
Lies n ein
$h = \frac{b}{n}$
$x = 0$
$y = 1$
Schreibe x und y formatiert in Datei `datei_1.txt`
$i = 1$
Solange $i \leq n$
$k_1 = y^2 \sin(x)$
$k_2 = (y + \frac{h}{2} k_1)^2 \sin(x + \frac{h}{2})$
$k_3 = (y + \frac{h}{2} k_2)^2 \sin(x + \frac{h}{2})$
$k_4 = (y + hk_3)^2 \sin(x + h)$
$y = y + \frac{h}{6}(k_1 + 2(k_2 + k_3) + k_4)$
$x = x + h$
Schreibe x und y formatiert in Datei `datei_1.txt`
$i = i + 1$

Abb. 2.15. Struktogramm für das Runge-Kutta-Verfahren 4. Ordnung angewandt auf die Differenzialgleichung $y'(x) = -y^2(x)\sin(x)$ (2.62)

nentrapezformel erfüllt. Das aus der Sehnentrapezformel abgeleitete Heunsche Verfahren erfüllt dies allerdings nicht. Von Gragg, Bulirsch und Stoer stammt eine **Modifikation der Tangententrapezformel**, die diese asymptotische Fehlerentwicklung besitzt (siehe z.B. [59]).

Wir wollen dieses Verfahren auf ein beliebiges Teilintervall $[x_i, x_{i+1}]$ anwenden. Der Näherungswert y_i wurde im vorherigen Schritt berechnet. Gesucht ist der neue Näherungswert y_{i+1} am Punkt x_{i+1}. Es wird so vorgegangen, dass

Gebräuchliche Varianten des Runge-Kutta-Verfahrens

2. Ordnung:

a)

$$
\begin{aligned}
k_1 &= f(x_i\ ,\ y_i)\\
k_2 &= f(x_i + \tfrac{h}{2}\ ,\ y_i + \tfrac{h}{2}k_1)\\
\hline
y_{i+1} &= y_i + hk_2
\end{aligned}
$$

b)

$$
\begin{aligned}
k_1 &= f(x_i\ ,\ y_i)\\
k_2 &= f(x_i + h\ ,\ y_i + hk_1)\\
\hline
y_{i+1} &= y_i + \tfrac{h}{2}(k_1 + k_2)
\end{aligned}
$$

3. Ordnung:

a)

$$
\begin{aligned}
k_1 &= f(x_i\ ,\ y_i)\\
k_2 &= f(x_i + \tfrac{h}{2}\ ,\ y_i + \tfrac{h}{2}k_1)\\
k_3 &= f(x_i + h\ ,\ y_i - hk_1 + 2hk_2)\\
\hline
y_{i+1} &= y_i + \tfrac{h}{6}(k_1 + 4k_2 + k_3)
\end{aligned}
$$

b)

$$
\begin{aligned}
k_1 &= f(x_i\ ,\ y_i)\\
k_2 &= f(x_i + \tfrac{h}{3}\ ,\ y_i + \tfrac{h}{3}k_1)\\
k_3 &= f(x_i + \tfrac{2h}{3}\ ,\ y_i + \tfrac{2h}{3}k_2)\\
\hline
y_{i+1} &= y_i + \tfrac{h}{4}(k_1 + 3k_3)
\end{aligned}
$$

4. Ordnung:

a)

$$
\begin{aligned}
k_1 &= f(x_i\ ,\ y_i)\\
k_2 &= f(x_i + \tfrac{h}{3}\ ,\ y_i + \tfrac{h}{3}k_1)\\
k_3 &= f(x_i + \tfrac{2h}{3}\ ,\ y_i - \tfrac{h}{3}k_1 + hk_2)\\
k_4 &= f(x_i + h\ ,\ y_i + h(k_1 - k_2 + k_3))\\
\hline
y_{i+1} &= y_i + \tfrac{h}{8}(k_1 + 3(k_2 + k_3) + k_4)
\end{aligned}
$$

b) (*klassische Variante*)

$$
\begin{aligned}
k_1 &= f(x_i\ ,\ y_i)\\
k_2 &= f(x_i + \tfrac{h}{2}\ ,\ y_i + \tfrac{h}{2}k_1)\\
k_3 &= f(x_i + \tfrac{h}{2}\ ,\ y_i + \tfrac{h}{2}k_2)\\
k_4 &= f(x_i + h\ ,\ y_i + hk_3)\\
\hline
y_{i+1} &= y_i + \tfrac{h}{6}(k_1 + 2(k_2 + k_3) + k_4)
\end{aligned}
$$

c) (*Gill*)

$$
\begin{aligned}
k_1 &= f(x_i\ ,\ y_i)\\
k_2 &= f(x_i + \tfrac{h}{2}\ ,\ y_i + \tfrac{h}{2}k_1)\\
k_3 &= f(x_i + \tfrac{h}{2}\ ,\ y_i = \tfrac{h}{2}(\sqrt{2} - 1)k_1 + \tfrac{h}{2}(2 - \sqrt{2})k_2)\\
k_4 &= f(x_i + \tfrac{h}{2}\ ,\ y_i - \tfrac{h}{2}\sqrt{2}k_2 + \tfrac{h}{2}(2 + \sqrt{2})k_3)\\
\hline
y_{i+1} &= y_i + \tfrac{h}{6}(k_1 + (2 - \sqrt{2})k_2 + (2 + \sqrt{2})k_3 + k_4)
\end{aligned}
$$

5. Ordnung:

(*Butcher*)

$$
\begin{aligned}
k_1 &= f(x_i\ ,\ y_i)\\
k_2 &= f(x_i + \tfrac{h}{4}\ ,\ y_i + \tfrac{h}{4}k_1)\\
k_3 &= f(x_i + \tfrac{h}{4}\ ,\ y_i + \tfrac{h}{8}(k_1 + k_2))\\
k_4 &= f(x_i + \tfrac{h}{2}\ ,\ y_i - \tfrac{h}{2}k_2 + hk_3)\\
k_5 &= f(x_i + \tfrac{3h}{4}\ ,\ y_i + \tfrac{3h}{16}(k_1 + 3k_4))\\
k_6 &= f(x_i + h\ ,\ y_i - \tfrac{h}{7}(3k_1 - 2k_2 - 12(k_3 - k_4) - 8k_5))\\
\hline
y_{i+1} &= y_i + \tfrac{h}{90}(7(k_1 + k_6) + 32(k_3 + k_5) + 12k_4)
\end{aligned}
$$

dieses Teilintervall sukzessive entsprechend einer Folge von Schrittweiten verfeinert wird bis die gewünschte Ordnung oder Genauigkeit erzielt ist. Um die Schreibweise zu vereinfachen und mehrere Indizes zu vermeiden, nennen wir das Teilintervall im Folgenden $[a, b]$. Der Anfangswert bei $x_0 = a$ sei mit

y_0 bezeichnet. Im ersten Schritt lautet die modifizierte Tangententrapezregel dann folgendermaßen.

Mit der Schrittweite $h_0/2$ mit $h_0 = b - a$ wird zuerst das Euler-Verfahren angewandt:

$$y_1 = y_0 + \frac{h_0}{2} f(x_0, y_0) \ ,$$

gefolgt von der Tangententrapezformel

$$y_2 = y_0 + h_0 \, f\left(\frac{a+b}{2}, y_1\right) \ .$$

Das Verfahren wird dann mit der Endformel

$$y_2 = \frac{1}{2} \left(y_1 + y_2 + h_0 f(b, y_2)\right)$$

abgeschlossen. Der Endwert y_2 entspricht der mit der Schrittweite $h_0/2$ berechneten Näherung von $y(b)$. Die ganze Prozedur wird in den nächsten Schritten mit kleineren Schrittweiten sukzessive wiederholt und verbesserte Näherungswerte mit der Fehlerextrapolationsformel berechnet. Allgemein im k-ten Schritt sieht das Extrapolationsverfahren dann wie folgt aus:

Schritt 1: Berechne die Schrittweite

$$h_k = \frac{b-a}{n_k} \ .$$

Schritt 2: Anlaufrechnung mit dem Euler-Verfahren

$$y_1 = y_0 + h_k f(x_0, y_0) \ . \tag{2.64}$$

Schritt 3: Tangententrapezformel

$$y_j = y_{j-2} + 2h_k f(x_{j-1}, y_{j-1}), \ j = 2, 3, \ldots, n_k \ . \tag{2.65}$$

Schritt 4: Endformel

$$T_{k,1} = \frac{1}{2} \left(y_{n_k - 1} + y_{n_k} + h_k f(x_{n_k}, y_{n_k})\right) \ . \tag{2.66}$$

Diese Schritte werden für verschiedene n_k durchgeführt.

Bei der **Romberg-Folge** ergibt sich

k	1	2	3	4	5	...
n_k	2	4	8	16	32	...

Dies ist einfacher bei der Programmierung, bringt aber für größere n_k wie schon beim Romberg-Verfahren einen höheren Rechenaufwand mit sich als die **Bulirsch-Folge**:

k	1	2	3	4	5	6
n_k	2	4	6	8	12	16

.

Wie beim Romberg-Verfahren werden aus den $T_{k,1}$ dann die $T_{k,2}$ nach der Formel

$$T_{k,2} = T_{k,1} + \frac{T_{k,1} - T_{k-1,1}}{\left(\frac{h_{k-1}}{h_k}\right)^2 - 1} \tag{2.67}$$

und allgemein

$$T_{k,j} = T_{k,j-1} + \frac{T_{k,j-1} - T_{k-1,j-1}}{\left(\frac{h_{k-1}}{h_k}\right)^{2(j-1)} - 1} \tag{2.68}$$

berechnet.

Bemerkung: Dieses Extrapolationsverfahren kann auf verschiedene Art und Weise programmiert werden. Es ist möglich, die Ordnung des Verfahrens vorzugeben und dann in jedem Teilintervall entsprechend dieser Vorgabe abzubrechen. Der große Vorteil der Extrapolation ist jedoch die mitgelieferte Fehlerabschätzung. So ist die sinnvollere Implementierung eine, welche durch die gewünschte Genauigkeit gesteuert wird. Zu der vorgegebenen Genauigkeit wird durch Vergleich der Werte $T_{k,j}$ entschieden, ob zu einer weiteren Rechnung mit n_{k+1} Stützstellen übergegangen werden muss.

2.6 Schrittweitenkontrolle und Fehlerschätzer

Bis auf das Extrapolationsverfahren wurden bislang die Verfahren mit einer vorgegebenen Schrittweite ausgeführt. Um die Stabilität des Verfahrens zu sichern, aber auch um die gewünschte Genauigkeit des Verfahrens zu erhalten, sollte eine Schrittweitenkontrolle mit der numerischen Simulation einhergehen. In Bereichen, in denen sich die Lösung sehr stark ändert, muss eine relativ kleine Schrittweite gewählt werden, um den Fehler klein zu halten.

In den Bereichen, in denen die Lösung sich wenig ändert, kann dann eine große Schrittweite gewählt werden, um den Rechenaufwand zu verringern. Die Schrittweite muss gewissermaßen dem Lösungsverlauf angepasst werden. Die Effizienz des Verfahrens hängt sehr stark von einer guten Schrittweitensteuerung ab.

Bei der Untersuchung des Konvergenzverhaltens der Sehnentrapezregel hatten wir festgestellt, dass die Iteration nur konvergiert, falls die Schrittkennzahl kleiner zwei ist. Bei dem Verfahren von Heun wird die Iteration auf nur einen Iterationsschritt reduziert. Dies reicht nur dann aus - ohne dass die Genauigkeit des Verfahrens sehr schlecht wird - falls eine deutlich kleinere Schrittkennzahl benutzt wird, z.B. kleiner um den Faktor 10. Eine **automatische Schrittweitensteuerung** nach diesem Kriterium wurde in Kapitel 2.1 sieht wie folgt aus:

1. Berechne y_{i+1}
2. $\kappa_{i+1} = h\,|f_y(x_{i+1}, y_{i+1})|$
3. Mittle: $\kappa := \frac{1}{2}(\kappa_i + \kappa_{i+1})$
4. Ist $\kappa \geq 0.2 \;\Rightarrow\; h := \frac{1}{2}h$ und letzten Schritt wiederholen.
   Ist $\kappa \leq 0.05 \;\Rightarrow\; h := 2h$ und weiter rechnen.
   Ist $0.05 < \kappa < 0.2 \;\Rightarrow$ Schrittweite beibehalten und weiter rechnen.

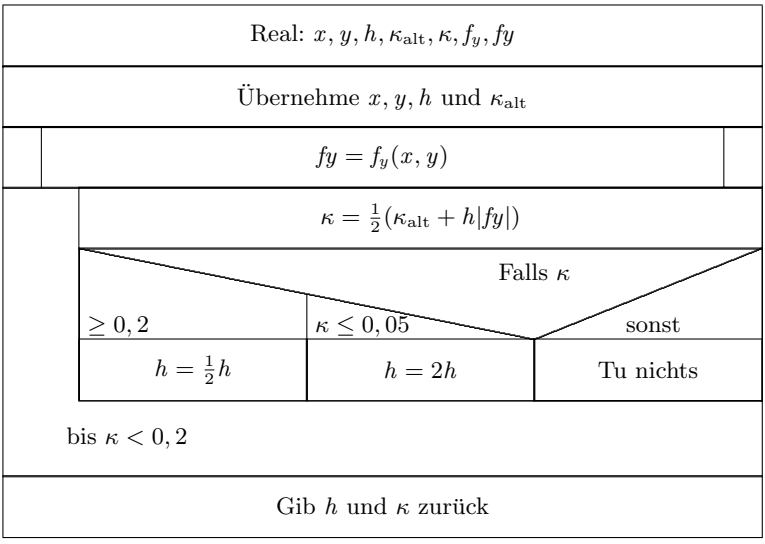


Abb. 2.16. Schrittweitensteuerung als Unterprogramm

Ein Struktogramm, in dem die Schrittweitenkontrolle in einem separaten Unterprogramm ausgeführt wird, ist in Abbildung 2.16 dargestellt. Hier werden die aktuellen Werte von x und y sowie der jeweils letzte Wert von κ und der Schrittweite h übergeben.

Bei der klassischen Variante des Runge-Kutta-Verfahrens 4. Ordnung erhält man automatisch eine Abschätzung der Ableitung f_y mitgeliefert und kann damit die Schrittkennzahl berechnen. Nach der Tabelle 2.4 ergibt sich aus

$$k_2 = f\left(x_i + \frac{h}{2}, y_i + \frac{h}{2}k_1\right) ,$$
$$k_3 = f\left(x_i + \frac{h}{2}, y_i + \frac{h}{2}k_2\right)$$

eine Näherung für f_y mit Hilfe des Differenzenquotienten zu

$$f_y \approx \frac{k_3 - k_2}{y_i + \frac{h}{2}k_2 - y_i - \frac{h}{2}k_1} = \frac{k_3 - k_2}{\frac{h}{2}(k_2 - k_1)} . \tag{2.69}$$

Dies wird dadurch ermöglicht, dass k_2 und k_3 das gleiche erste Argument $x_i + h/2$ besitzen. Die abgeschätzte Schrittkennzahl ergibt sich dann in der einfachen Form zu

$$\kappa = h|f_y| = 2\left|\frac{k_3 - k_2}{k_2 - k_1}\right| . \tag{2.70}$$

Eine Schrittweitensteuerung kann analog wie oben durchgeführt werden.

Eine andere Möglichkeit ist die direkte **Abschätzung des Fehlers** wie bei einem Extrapolationsverfahren. Es werden zwei Rechnungen mit verschiedenen Schrittweiten durchführen. Nehmen wir z.B. h und $h/2$ und bezeichnen die Näherungslösungen mit y_h und $y_{h/2}$. Möchte man eine Abschätzung des Fehlers bekommen, dann kann der Vergleich dieser Näherungswerte Aufschluss geben, auf wie viele Stellen die Näherungswerte übereinstimmen und man kann dies als Maß der Genauigkeit nehmen. Dies ist allerdings keine garantierte Abschätzung des Fehlers und man sollte hier vorsichtig sein. Eine weitere Gitterverfeinerung sollte zur Bestätigung durchgeführt werden.

Wenn zwei Rechnungen mit verschiedenen Schrittweiten vorliegen, dann kann man noch einen Schritt weitergehen und eine Fehlerextrapolation durchführen. Weiss man die Konvergenzordnung des Verfahrens und gilt für den globalen Diskretisierungsfehler

$$\begin{aligned} y_h(x_{i+1}) - y(x_{i+1}) &= A_0 h^p + \mathcal{O}(h^{p+1}) , \\ y_{h/2}(x_{i+1}) - y(x_{i+1}) &= 2A_0\left(\frac{h}{2}\right)^p + \mathcal{O}(h^{p+1}) , \end{aligned} \tag{2.71}$$

wobei die Funktion A_0 nur von x_i und nicht von h abhängt, dann können wir den Wert der exakten Lösung $y(x_{i+1})$ dort eliminieren. Der Faktor 2 in der zweiten Entwicklung kommt dadurch zustande, dass für $h/2$ zwei Schritte durchgeführt werden müssen. Für Runge-Kutta-Verfahren ist dies ausführlich in [29] abgeleitet. Werden die beiden Gleichungen in (2.71) subtrahiert, dann wird der exakte Wert eliminiert und man erhält

$$y_{h/2} - y_h = A_0 \left(2 \left(\frac{h}{2} \right)^p - h^p \right) . \tag{2.72}$$

Daraus berechnet sich die führende Fehlerkonstante A_0 zu

$$A_0 = \frac{y_{h/2} - y_h}{2(1 - 2^{p-1}) \left(\frac{h}{2} \right)^p} .$$

Dies kann zum Einen benutzt werden, um den Hauptfehlerterm auf der rechten Seite von (2.71) abzuschätzen, zum Anderen können die Näherungswerte noch verbessert werden, indem man diesen Fehlerterm abzieht.

2.7 Systeme von Differenzialgleichungen und Differenzialgleichungen höherer Ordnung

Alle hier vorgestellten Verfahren lassen sich auf Systeme erster Ordnung übertragen. Ein **System erster Ordnung** hat zunächst die Form

$$\begin{aligned} y_1'(x) &= f_1(x, y_1, y_2, \ldots, y_n) , \\ y_2'(x) &= f_2(x, y_1, y_2, \ldots, y_n) , \\ \vdots \qquad &\qquad \vdots \\ y_n'(x) &= f_n(x, y_1, y_2, \ldots, y_n) . \end{aligned} \tag{2.73}$$

Die n Funktionen $y_1, \ldots, y_n$ sind eine Lösung des Anfangswertproblems dieses Differenzialgleichungssystems im Intervall $I = [a, b]$, wenn sie die Gleichungen für alle $x \in I$ mit den Anfangsdaten

$$y_1(a) = y_{1,a},\ y_2(a) = y_{2,a},\ \ldots,\ y_n(a) = y_{n,a}$$

erfüllen. In der Vektorschreibweise mit den Vektoren

$$\mathbf{y} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix} ,\ \mathbf{f}(x, \mathbf{y}) = \begin{pmatrix} f_1(x, \mathbf{y}) \\ f_2(x, \mathbf{y}) \\ \vdots \\ f_n(x, \mathbf{y}) \end{pmatrix}$$

können wir das System kurz in der Form

$$\mathbf{y}' = \mathbf{f}(x, \mathbf{y}) \tag{2.74}$$

schreiben.

Das zugehörige Anfangswertproblem lautet

$$\mathbf{y}' = \mathbf{f}(x, \mathbf{y}), \quad \mathbf{y}(a) = \mathbf{y}_a \, . \tag{2.75}$$

Damit hat das System von Differenzialgleichungen erster Ordnung formal genau dieselbe Schreibweise wie eine einzelne Differenzialgleichung. Die Funktionen sind hier vektorwertige Funktionen von x und dem Vektor $\mathbf{y}$. Dementsprechend können die numerischen Methoden formal in der gleichen Art direkt auf Systeme übertragen werden.

Beispiel 2.15. Wir betrachten ein einfaches System bestehend aus den zwei Differenzialgleichungen

$$u' = 2u + 2v \, , \tag{2.76}$$
$$v' = 3u + v \, . \tag{2.77}$$

Dabei seien für das Anfangswertproblem geeignete Anfangswerte $u(0) = u_0$ und $v(0) = v_0$ vorgeschrieben.

Wenden wir ein numerisches Verfahren zur Lösung des Anfangswertproblems für dieses System an, dann wird das numerische Verfahren auf jede dieser Gleichungen angewandt. Der Einfachheit halber führen wir dies im Folgenden für das Euler-Cauchy-Verfahren aus. Die diskretisierte Form der beiden Gleichungen beim i-ten Schritt lautet

$$u_{i+1} = u_i + h(2u_i + 2v_i) \, , \tag{2.78}$$
$$v_{i+1} = v_i + h(3u_i + v_i) \tag{2.79}$$

oder in der Vektorschreibweise

$$\mathbf{y}_{i+1} = \mathbf{y}_i + h \, \mathbf{f}(x_i, \mathbf{y_i}) \tag{2.80}$$

mit

$$\mathbf{y} = \begin{pmatrix} u \\ v \end{pmatrix} \, , \quad \mathbf{f}(x, \mathbf{y}) = \mathbf{f}(x, u, v) = \begin{pmatrix} 2u + 2v \\ 3u + v \end{pmatrix} \, .$$

Ganz analog übertragen sich andere Verfahren, indem man diese auf die einzelnen Gleichungen anwendet. □

Sogenannte **steife** Differenzialgleichungssysteme ergeben sich bei Vorgängen, bei denen die Lösung verschiedenartige Anteile enthalten, z.B. sehr unterschiedliches Abklingverhalten. Man benötigt hier spezielle numerische Methoden, welche im nächsten Abschnitt kurz angesprochen werden.

In den Anwendungen treten auch oft **Anfangswertprobleme für Differenzialgleichungen höherer Ordnung** auf. Ein Anfangswertproblem für eine Differenzialgleichung m-ter Ordnung hat die Form

$$y^{(m)} = g\left(x, y, y', y'', \ldots, y^{(m-1)}\right)$$

mit vorgegebenen Anfangswerten

$$y^{(i)}(a) = y_{a,i} \quad \text{für } i = 0, 1, \ldots, m-1 \,.$$

Die Differenzialgleichungen m-ter Ordnung können in ein System von m Differenzialgleichungen erster Ordnung überführt werden. Definiert man

$$u_i := y^{(i-1)} \quad \text{für } i = 1, 2, \ldots, m \,,$$

so erhält man das System

$$\begin{aligned} u_i' &= u_{i-1} \quad \text{für } i = 1, 2, \ldots, m-1 \\ u_m' &= g(x, u_1, u_2, \ldots, u_m) \end{aligned}$$

mit den entsprechenden Anfangswerten

$$u_i(a) = y_{a,i-1} \quad \text{für } i = 1, 2, \ldots, m \,.$$

Beispiel 2.16. Wir betrachten in diesem Beispiel die Schwingung einer Masse m, welche an zwei gleichen Federn zwischen zwei Wänden aufgehängt ist. Die Schwerkraft wird dabei vernachlässigt. Die Gleichung für die Auslenkung aus der Ruhelage $y = y(t)$ lautet

$$m\,\ddot{y} + r\dot{y} + ky = g, \tag{2.81}$$

wobei r der Reibungskoeffizient, m die Masse, k die Federkonstante und g eine äußere Kraft ist. Möchte man diese Differenzialgleichung 2. Ordnung in ein System erster Ordnung überführen, so definiert man nach Gleichung (2.81)

$$u_1 := y, \quad u_2 := y' \,. \tag{2.82}$$

Man erhält dann die beiden Gleichungen

$$\dot{u_1} = u_2, \quad \dot{u_2} = -\frac{r}{m}u_2 - \frac{k}{m}u_1 + g \,. \tag{2.83}$$

Diese bilden ein System erster Ordnung, welches dann mit Anfangswerten für u_1 und u_2 numerisch gelöst werden können, wie es am Anfang dieses Abschnittes gezeigt wurde. Als System können wir es in der Form

$$\mathbf{u} = \begin{pmatrix} u_1 \\ u_2 \end{pmatrix} , \qquad \mathbf{f}(x, \mathbf{u}) = \mathbf{f}(x, u, v) = \begin{pmatrix} u_2 \\ -\frac{k}{m} u_1 - \frac{r}{m} u_2 + g \end{pmatrix} .$$

schreiben. □

Beispiel 2.17. In Beispiel 2.3 wurde das mathematische Modell der Biegung eines Balkens behandelt. Wird durch Integration der Einzelkräfte nach (2.8) und (2.9) das Biegemoment $M = M(z)$ berechnet, dann erhält man das System erster Ordnung

$$\frac{d\phi(z)}{dz} = -\frac{M(z)}{E(z)I(z)} \left(1 + \phi^2\right)^{\frac{3}{2}} , \tag{2.84}$$

$$\frac{dv(z)}{dz} = \phi(z) , \tag{2.85}$$

bestehend aus zwei Gleichungen für den Biegewinkel ϕ und die Durchbiegung v. Dieses System kann man noch zusammenfassen, indem man (2.85) in die Gleichung (2.85) einsetzt:

$$v''(z) = -\frac{M(z)}{E(z)I(z)} (1 + (v'(z))^2)^{\frac{3}{2}} . \tag{2.86}$$

Das System 1. Ordnung lässt sich somit umgekehrt als eine Differenzialgleichung zweiter Ordnung umschreiben. □

2.8 Bemerkungen und Entscheidungshilfen

Welches Verfahren ist denn nun das Beste? Wie so oft im Bereich der numerischen Simulation lässt sich diese Frage leider so nicht eindeutig beantworten. Wir hätten hier dann natürlich auch nur dieses Verfahren vorgestellt. Das Standardverfahren zur numerischen Behandlung von Anfangswertproblemen gibt es nicht. Um eine Antwort zu bekommen, muss die Frage etwas anders gestellt werden: Welches Verfahren ist das richtige für ein vorgegebenes Problem? Betrachtet man die Vor- und Nachteile und die Eigenschaften der Verfahren, dann bekommt man Hinweise darauf, welches Verfahren das am besten geeignete für ein gegebenes Problem sein sollte.

Das **Euler-Cauchy-Verfahren** oder auch das **Verfahren von Heun** sind im Allgemeinen zu ungenau und werden für praktische Rechnungen nur in

Ausnahmefällen eingesetzt. Allerdings sind diese Verfahren sehr schnell, so dass sie für Anwendungen, die zeitkritisch sind, aber keine hohe Genauigkeit erfordern, sinnvoll sein können.

Bei den **Runge-Kutta-Verfahren** ist kein Anlaufstück nötig, sie sind gewissermaßen selbststartend. Sie haben eine feste lokale Fehlerordnung. Es handelt sich um eine ganze Klasse von Verfahren, so dass, abhängig von der gewünschten Genauigkeit, ein entsprechendes Verfahren ausgewählt werden kann. Runge-Kutta-Verfahren sind einfach zu handhaben. Eine automatische Schrittweitensteuerung, wie sie im vorherigen Abschnitt angesprochen wurde, ist verhältnismäßig einfach zu implementieren. All diesen Vorteilen steht der Nachteil gegenüber, dass zur Berechnung eines neuen Näherungswertes mehrere Funktionsauswertungen zu machen sind, etwa beim klassischen Runge-Kutta-Verfahren vier. Runge-Kutta-Verfahren mit einer Genauigkeitsordnung 5 benötigen 6 Stufen. Das Verhalten, dass die Anzahl der Stufen größer als die Ordnung sein muss, setzt sich so fort. Der Rechenaufwand wächst bei höherer Ordnung somit kräftig an. Die Runge-Kutta-Verfahren sind somit günstig für Probleme, für die keine sehr hohe Genauigkeit gefordert wird und der Rechenaufwand für die Berechnung der Funktionswerte klein ist.

Der Vorteil der **Mehrschrittverfahren** ist, dass pro Schritt nur ein oder auch bei dem Prädiktor-Korrektor-Verfahren nur zwei Funktionsauswertungen nötig sind. Es können Formeln mit beliebig hoher Genauigkeit abgeleitet werden. Die Nachteile sind demgegenüber, dass die Verfahren ein Anlaufstück benötigen, etwa durch ein Runge-Kutta-Verfahren. Eine Schrittweitensteuerung ist möglich, aber dadurch, dass bei Änderung der Schrittweite ein Anlaufstück jeweils neu berechnet werden muss, ist dies deutlich aufwändiger. Ein Mehrschrittverfahren wird man besonders dann anwenden, wenn die Auswertung von f viel Rechenaufwand benötigt. Statt oder zusätzlich zu einer Steuerung der Schrittweite können auch Verfahren mit variabler Ordnung angewandt werden.

Ein **Extrapolationsverfahren** ist selbststartend und hat keine starre Ordnung. Das Verfahren wird gesteuert, indem die Genauigkeit vorgegeben wird, welche die Näherungslösung haben soll. Neben einer Schrittweitensteuerung kann die Ordnung des Verfahrens erhöht werden. Dies führt auf die Konstruktion sehr effizienter Rechenprogramme. Der Nachteil der Extrapolationsverfahren ist, dass der Rechenaufwand pro Schritt sehr groß werden kann. Automatische Schrittweitensteuerung und variable Ordnung machen die Implementierung aufwändig. Das Extrapolationsverfahren ist aber oft das Verfahren der Wahl, falls der Aufwand zur Berechnung der Funktion f nicht zu groß und f sehr glatt ist.

Die Verfahren, die wir in diesem Kapitel vorgestellt haben, sind fast alle explizite Verfahren. Implizite Verfahren führen auf die Lösung von im All-

gemeinen nichtlinearen Gleichungssystemen und damit auf einen großen Rechenaufwand. Allerdings haben diese Verfahren ein besseres Stabilitätsverhalten. Dies ist wichtig bei numerischen Methoden für **steife Differenzialgleichungssysteme**. Diese Differenzialgleichungssysteme kommen in verschiedenen Anwendungen der Naturwissenschaften und der Technik vor. Der Begriff der Steifheit lässt sich nicht genau definieren, sondern an Hand der Eigenschaft der Lösungen erklären. Steife Differenzialgleichungssysteme beschreiben Vorgänge, deren Lösungen sich aus zwei stark unterschiedlichen Anteilen zusammensetzen, z. B. aus einem sehr schnell und einem langsam exponentiell abklingenden Anteil. Typische Beispiele sind mathematische Modelle zur Beschreibung chemischer Reaktionen. Hier gibt es Reaktionen mit sehr unterschiedlichen Geschwindigkeiten. Im Spezialfall eines linearen Problems

$$\mathbf{y}'(x) = \mathbf{A} \cdot \mathbf{y}(x) \,, \qquad \mathbf{y}(x_0) = \mathbf{y_0} \,. \tag{2.87}$$

nennt man das System steif, wenn die negativen Realteile der Eigenwerte der Matrix $\mathbf{A}$ sehr unterschiedlich sind. Der Quotient des größten und des kleinsten Eigenwertes ist somit ein Maß der Steifheit. Die Realteile der Eigenwerte bestimmen das exponentielle Abklingverhalten in der Lösung. Die Schrittweite bei einem expliziten Verfahren muss so klein sein, dass das Verfahren bezüglich des schnellsten Abklingens stabil ist.

Betrachtet man chemische Reaktionen und möchte wissen, aus was sich der Stoff nach einer längeren Zeit zusammensetzt, so erfordert das explizite Verfahren Zeitschritte, welche eine Auflösung der zeitlichen Entwicklung jeder Reaktion erlaubt. Ist die zeitliche Auflösung einer schnellen Reaktion zur Beschreibung des Gesamtvorgang nicht nötig, könnte man die entsprechenden Reaktionen direkt durch das Gleichgewicht ersetzen, dann erfordert das numerische Verfahren viel zu kleine Zeitschritte. Das Gesamtreaktion ist dadurch vielleicht in sinnvoller Zeit nicht mehr zu lösen oder Rundungsfehler wachsen an. In diesem Fall wendet man ein implizites Verfahren mit einem besseren Stabilitätsverhalten an. Wir werden steife Problem hier nicht weiter betrachten und verweisen auf die Diskussion und Ausführungen in der Spezialliteratur.

Weiterführende Literatur zur numerischen Lösung von steifen und nichtsteifen Anfangswertproblemen für gewöhnliche Differenzialgleichungen sind die Bücher über numerische Mathematik von Deuflhard und Bornemann [14], Stoer und Bulirsch [59]. Eine projektorinetierte Einführung insbesondere für Ingenieure findet sich in [3]. Spezielle Monographien über AWP sind die Bücher von Hairer, Nørsett und Wanner [28], Hairer und Wanner [29], Strehmel und Weiner [62] sowie Grigorieff [22],[23]. Programme zur numerischen Lösungen von gewöhnlichen Differenzialgleichungen gibt es sowohl für steife als auch nicht-steife Probleme in [47] und den großen Bibliotheken von NAG [45] und IMSL [68]. Auch in MATLAB werden mehrere Programme angeboten.

Im Internet werden verschiedene Programme von den Autoren der Bücher Deuflhard und Bornemann [14] und Hairer, Nørsett und Wanner [28], [29] angeboten, siehe auch http://www.netlib.org/.

2.9 Beispiele und Aufgaben

Nach der Herleitung von Näherungsverfahren werden in diesem Abschnitt Beispiele aufgegriffen, welche wir am Anfang des Kapitels zur Motivation vorgestellt haben, und numerisch gelöst. Das einfachste Modell zur Populationsdynamik in Beispiel 2.1 haben wir schon als Testproblem zur Untersuchung der Stabilität von Näherungsverfahren herangezogen. In Beispiel 2.2 wurde ein einfache Bewegungsgleichung für den Flug in konstanter Höhe ohne Richtungsänderung angegeben. Wir wollen dieses Problem als erstes hier wieder aufgreifen.

Beispiel 2.18 (Flug in konstanter Höhe, mit MAPLE-Worksheet). Wir lösen die am Anfang des Kapitels vorgestellte Bewegungsdifferenzialgleichung für ein Flugzeug in konstanter Höhe

$$\dot{v} = a_0 - a_1 v^2 - \frac{a_2}{v^2} := f(v) \tag{2.88}$$

mit den konstanten Koeffizienten

$$a_0 = \frac{F}{m}\,, \quad a_1 = \frac{\rho\, S\, c_{W_0}}{2m}\,, \quad a_2 = \frac{2k\, G^2}{m\,\rho\, S}\,.$$

Für die Zahlenwerte

$$c_{W_0} = 0.02\,, \quad G = 200000\,\mathrm{N}\,, \quad g = 9.81\,\mathrm{m/s^2}\,, \quad k = 0.2\,, \quad S = 50\,\mathrm{m^2}$$

ergeben sich die Koeffizienten zu

$$a_0 = 7.0898\,\mathrm{m/s^2}\,, \quad a_1 = 2.3633 \cdot 10^{-5}\,\mathrm{1/m}\,, \quad a_2 = 16289\,\mathrm{kg/s^4}\,.$$

Dabei hat man noch ausgenutzt, dass sich die Masse aus dem Gewicht entsprechend $m = G/g$ ergibt. Die Luftdichte wurde als exponentielle Funktion der Höhe modelliert:

$$\rho = \rho_0\, e^{-\beta h}$$

mit

$$\beta = 1.2 \cdot 10^{-4}\,\mathrm{1/m} \qquad \text{und} \qquad \rho_0 = 1.225\,\mathrm{kg/m^3}\,.$$

Der Schub wird als proportional zur Luftdichte angenommen:

$$F = F_0\rho \quad \text{mit} \quad F_0 = 150000\,\mathrm{m^4/s^2}\,.$$

In dem vorliegenden Modell gibt es zwei Nullstellen v_1 und v_2 der Funktion

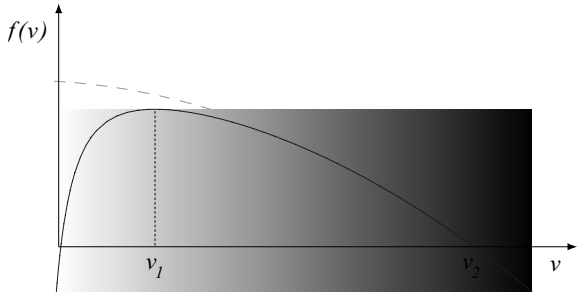


Abb. 2.17. Längsbeschleunigung als Funktion der Geschwindigkeit

$f(v)$, beide mit positiver Geschwindigkeit. Es gilt

$$f(v) > 0 \quad \text{für} \quad v_1 < v < v_2\,.$$

Die Werte für v_1 und v_2 ergeben sich aus der Gleichung

$$a_0 - a_1 v^2 - \frac{a_2}{v^2} = 0\,.$$

Durch Substitution lässt sie sich in eine quadratische Gleichung überführen und lösen. In diesem Modell müsste somit für jeden Anfangswert größer als v_1 der stationäre Zustand v_2 nach einer gewissen Zeit erreicht werden. Die Werte v_1 und v_2 sind die Grenzen der Flugenvelope. Der Wert v_2 stellt ein stabiles Gleichgewicht dar. Ist die Geschwindigkeit etwas größer als v_2, dann ist die Funktion $f(v)$ negativ und die Geschwindigkeit wird bis v_2 wieder reduziert. Der Wert v_1 ist ein instabiles Gleichgewicht. Ist v ein klein wenig größer als v_1, dann wird die Lösung gegen den stabilen Gleichgewichtswert v_2 streben. Ist die Geschwindigkeit kleiner als v_1, dann nimmt die Geschwindigkeit weiter kontinuierlich ab. Das sollte für das Flugzeug nicht so günstig sein. Das Modell ist somit nur in dem Bereich $> v_1$ sinnvoll. In dem Worksheet wird das Runge-Kutta-Verfahren auf dieses Problem angewandt. □

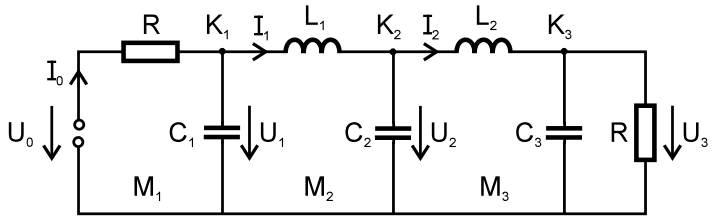


Abb. 2.18. Tiefpass

Beispiel 2.19 (Elektrische Schaltungen, mit MAPLE-Worksheet). In diesem Beispiel kommen wir auf die in Beispiel 2.4 modellierte Schaltung eines Tiefpasses 5. Ordnung zurück (siehe Abbildung 2.9.) Auf jede der Differenzialgleichungen

$$\frac{dU_1}{dt} = ((U_0 - U_1)/R - I_1)/C_1 \tag{2.89}$$

$$\frac{dI_1}{dt} = (U_1 - U_2)/L_1 \tag{2.90}$$

$$\frac{dU_2}{dt} = (I_1 - I_2)/C_2 \tag{2.91}$$

$$\frac{dI_2}{dt} = (U_2 - U_3)/L_2 \tag{2.92}$$

$$\frac{dU_3}{dt} = (I_2 - U_3/R)/C_3 \tag{2.93}$$

wenden wir das Euler-Cauchy-Verfahren an. Für die Rechnung wählen wir $L_1 = L_2 = 1$; $C_1 = C_3 = 1$; $C_2 = 2$; $R = 0.8$; setzen $t_{max} = 90$, $dt = t_{max}/500$ und simulieren die Schaltung für unterschiedliche Eingangssignale.

(1.) Die Eingangsspannung sei eine Wechselspannung

$$U_0(t) = sin(\omega t).$$

Variiert man ω zwischen 0.5 und 2.0, so erhält man für die Amplitude der Ausgangsspannung

ω	0.5	0.75	1.0	1.25	1.5	1.75	2.0
U_A	0.5	0.5	0.5	0.42	0.17	0.065	0.029

Man erkennt, dass die Ausgangsamplitude bei Frequenzen zwischen 1.25 und 1.5 drastisch abfällt; dies spiegelt den Tiefpasscharakter der Schaltung wider: Tiefe Frequenzen können die Schaltung passieren ($U_A \approx 0.5$), hohe Frequenzen werden gesperrt ($U_A \ll 0.5$).

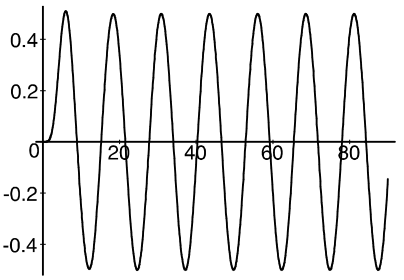
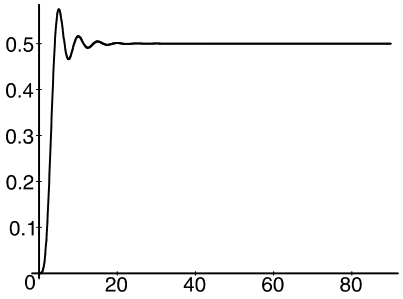
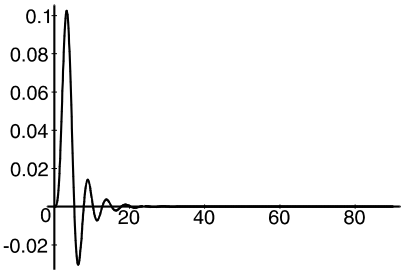


Abb. 2.19. Ausgangssignal für das Eingangssignal $U_0 = sin(\omega t)$ mit $\omega = 1$

(2.) Um einen Einschaltvorgang zu simulieren, setzen wir als Eingangsspannung die Sprung- oder Heaviside-Funktion. Für negative Zeiten hat diese den Wert 0 und springt bei $t = 0$ auf den Wert 1. In MAPLE wird diese Funktion mit **Heaviside(t)** bezeichnet: $U_0(t) :=$ Heaviside(t). Der Wert an der Stelle $t = 0$ muss explizit definiert werden: Heaviside(0.):=0.



(a) Sprungantwort



(b) Impulsantwort für $T = 0.5$

Abb. 2.20. Simulationsergebnisse

Das zur Sprungfunktion gehörende Antwortsignal ist in Abbildung 2.19a gezeichnet. Da dieses Antwortsignal die Reaktion des Systems auf die Sprungfunktion ist, nennt man sie die *Sprungantwort*.

(3.) Wir wählen als weiteres Eingangssignal einen Impulsstoß mit Breite T und Höhe $1/T$. Damit regen wir das System impulsartig an. Der Impulsstoß lässt sich in Maple mit

$$U_0(t) := 1/T * (Heaviside(t) - Heaviside(t - T)), \quad Heaviside(0.) := 0$$

realisieren. Die Ausgangsspannung nennt man zugehörig die *Impulsantwort*, siehe Abbildung 2.19b. □

Aufgabe 2.1 (Vergleich verschiedener AWP-Löser). Lösen sie das Anfangswertproblem

$$y'(x) = \sin(x)\, y, \quad y(0) = 1$$

im Intervall $[0, 4]$ mit den in diesem Kapitel vorgestellten Einschrittverfahren (Euler-Cauchy, verb. Euler-Cauchy, Heun sowie Runge-Kutta 3. und 4. Ordnung). Im Worksheet wird der L_2-Fehler

$$L_2 = \sqrt{\int \left(u_{\text{exakt}} - u_{\text{numerisch}}\right)^2 dx}$$

dieser Verfahren berechnet. Verkleinern sie die Schrittweiten aller Verfahren, um die Genauigkeit des Runge-Kutta-Verfahrens 4. Ordnung mit 10 Schritten zu erhalten. An Hand der Verfahrensordnung kann man dabei abschätzen, wie groß die Schrittweite sein müsste.

Lösung: MAPLE**-Worksheet** □

Aufgabe 2.2 (Schrittweitensteuerung). Lösen sie das Anfangswertproblem

$$y''(x) = \sin(x)\, y^2, \quad y(0) = -4$$

im Intervall $[0, 5]$ mit dem Runge-Kutta-Verfahren 4. Ordnung mit Schrittweitensteuerung. Die Abbildung 2.21 zeigt eine sehr gute Übereinstimmung der Näherungswerte mit der exakten Lösung. Man sieht, wie die Stützstellen abhängig vom Lösungsverhalten ungleichmäßig verteilt sind.

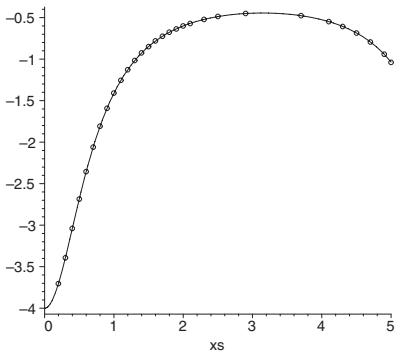


Abb. 2.21. Lösung des Anfangswertproblems mit Schrittweitensteuerung

Lösung: MAPLE**-Worksheet** □

3 Rand- und Eigenwertprobleme gewöhnlicher Differenzialgleichungen

Am Anfang des Kapitels 2 über Anfangswertprobleme hatten wir als Beispiel die Durchbiegung eines eingespannten Balkens, mit konstanter Streckenlast betrachtet. Dieses Problem könnte sich in der Anwendung auch in der Form stellen, dass der Balken auf beiden Seiten aufliegt. Diese beidseitige Lagerung führt auf zwei Randwerte für das Problem (siehe Abbildung 3.1). Die Differenzialgleichung, welche das Verbiegen beschreibt und in (2.86) als Differenzialgleichung 2. Ordnung geschrieben wird, bleibt dieselbe

$$y''(x) = -\frac{M(x)}{E(x)I(x)} \left(1 + (y'(x))^2\right)^{\frac{3}{2}} \quad \text{in } [0, l] \,, \tag{3.1}$$

aber es sind jetzt die zwei Randwerte

$$y(0) = 0, \qquad y(l) = 0 \tag{3.2}$$

vorgegeben. Es ist hier eine Lösung der Differenzialgleichung gesucht, welche an zwei verschiedenen Punkten vorgegebene Werte annimmt. Man nennt ein solches Problem ein **Randwertproblem**.

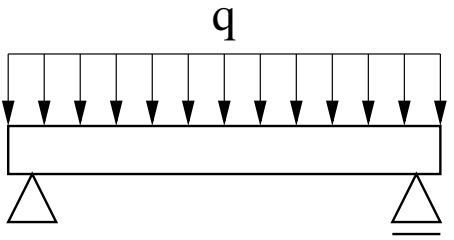


Abb. 3.1. Aufgelegter Balken

Im Gegensatz zum Anfangswertproblem können wir nun nicht mehr von einem Punkt aus losrechnen. Auch die Frage von Existenz- und Eindeutigkeitsaussagen für die Lösung von Randwertproblemen ist schwieriger zu beantworten. Wir können hier auf die Theorie nicht im Detail eingehen, sondern werden die verschiedenen Möglichkeiten an Hand des linearen Falls erörtern.

3.1 Vorbemerkungen und Begriffsbestimmungen

Wir befassen uns zunächst mit linearen Differenzialgleichungen höherer Ordnung, wie z.B. der Differenzialgleichung 2. Ordnung

$$y''(x) = r(x) - q(x)y(x) - p(x)y'(x)$$

mit zumindest stetigen Funktionen $r = r(x)$, $q = q(x)$ und $p = p(x)$. Linear heißt, dass die gesuchte Funktion $y = y(x)$ und alle auftretenden Ableitungen von y linear vorkommen; eine Nichtlinearität in x ist zugelassen.

Die Mathematik besagt, dass die allgemeine Lösung von linearen inhomogenen Differenzialgleichungen n-ter Ordnung, $\mathbb{L}_i$, sich zusammensetzt aus der allgemeinen homogenen Lösung, $\mathbb{L}_h$, plus einer speziellen Lösung $y_p(x)$ der inhomogenen Differenzialgleichung:

$$\mathbb{L}_i = \mathbb{L}_h + y_p(x) \; .$$

Dabei stellt die Lösungsmenge des homogenen Problems, $\mathbb{L}_h$, einen n-dimensionalen Vektorraum dar. In unserem Beispiel für die Differenzialgleichung 2. Ordnung ist

$$y(x) = c_1 y_1(x) + c_2 y_2(x) + y_p(x) \tag{3.3}$$

die allgemeine Lösung, wenn $y_1 = y_1(x)$, $y_2 = y_2(x)$ ein Fundamental-System bildet, d.h. y_1 und y_2 sind linear unabhängige Lösungen der homogenen Differenzialgleichung, und $y_p = y_p(x)$ ist eine spezielle (partikuläre) Lösung der inhomogenen Differenzialgleichung. Die Situation bei den Anfangswertproblemen ist die folgende: Durch die Angabe zweier Anfangsbedingungen $y(a) = y_0$ und $y'(a) = y'_0$ werden die freien Parameter c_1 und c_2 bestimmt. Damit ist das Anfangswertproblem eindeutig festgelegt und mit gewissen Stetigkeitsvoraussetzungen an die rechte Seite f immer lösbar.

Um die beiden Konstanten c_1 und c_2 festzulegen, kann man statt der Vorgabe der Anfangswerte $y(a)$, $y'(a)$ zwei Randwerte $y(a)$ und $y(b)$ spezifizieren.

Randwertproblem (RWP): Sind für die Differenzialgleichung der Form

$$y''(x) = r(x) - q(x)y(x) - p(x)y'(x) \qquad \text{in } [a, b] \tag{3.4}$$

die Randbedingungen

$$y(a) = \alpha, \qquad y(b) = \beta \tag{3.5}$$

gegeben, nennt man dies das **Randwertproblem** für (3.4).

Die Differenzialgleichung 2. Ordnung ist gewissermaßen ein Standard-Typ für ein Randwertproblem. Sie treten aber auch in anderer Form auf.

Da man jede Differenzialgleichung n-ter Ordnung der Form

$$y^{(n)}(x) = f(x, y(x), \ldots, y^{(n-1)}(x))$$

in ein Differenzialgleichungssystem 1. Ordnung transformieren kann, ist ein weiterer Standard-Typ allgemein das System erster Ordnung

$$\mathbf{y}'(x) = \mathbf{f}(x, \mathbf{y}(x)) \qquad \text{in } [a, b] \,. \tag{3.6}$$

Die Randbedingungen können dabei in verschiedenen Formen auftreten. Für ein System aus n Gleichungen wird man erwarten, dass n Randbedingungen gefordert werden - eine gewisse Anzahl bei $x = a$ und die restlichen bei $x = b$. Ganz allgemein kann man Randbedingungen in der Form

$$\mathbf{A}\,\mathbf{y}(a) + \mathbf{B}\,\mathbf{y}(b) = \mathbf{r} \tag{3.7}$$

mit den $n \times n$-Matrizen $\mathbf{A}$ und $\mathbf{B}$ fordern.

Die Frage der Existenz einer Lösung des Randwertproblems ist weitaus schwieriger als bei einem Anfangswertproblem. Ein Randwertproblem besitzt nicht immer eine Lösung. Wesentlich ist, ob das Randwertproblem homogen ist oder nicht. Für das eingangs gestellte lineare Problem in Gleichung (3.4) wird dies im Folgenden kurz diskutiert.

Definition: Gegeben sei das Randwertproblem (RWP)

$$y''(x) + p(x)y'(x) + q(x)y(x) = r(x) \qquad x \in [a, b] \,,$$

$$y(a) = \alpha, \qquad y(b) = \beta \,.$$

1. Für $r(x) \equiv 0$, $\alpha = \beta = 0$ nennt man das RWP **(voll) homogen** bzw. ein **Eigenwertproblem** (EWP).
2. Für $r(x) = 0$ nennt man das RWP **halb homogen**.
3. Für $r(x) \neq 0$ nennt man das RWP **inhomogen**.

3.1.1 Homogenes Randwertproblem (Eigenwertproblem)

Das homogene Problem

$$y''(x) + p\,y'(x) + q\,y(x) = 0 \qquad \text{mit} \quad y(a) = y(b) = 0$$

hat als allgemeine Lösung

$$y(x) = c_1\,y_1(x) + c_2\,y_2(x)$$

mit zwei linear unabhängigen Lösungen y_1 und y_2. Die Koeffizienten c_1 und c_2 bestimmt man aus den Randbedingungen

$$\left.\begin{aligned} y(a) &= c_1 y_1(a) + c_2 y_2(a) = 0 \\ y(b) &= c_1 y_1(b) + c_2 y_2(b) = 0 \end{aligned}\right\} \Rightarrow \begin{pmatrix} y_1(a) & y_2(a) \\ y_1(b) & y_2(b) \end{pmatrix} \begin{pmatrix} c_1 \\ c_2 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}.$$

Damit nichttriviale Lösungen (d.h. Koeffizienten $c_i \neq 0$) des Randwertproblems existieren, muss die Determinante der Matrix verschwinden:

$$D = \begin{vmatrix} y_1(a) & y_2(a) \\ y_1(b) & y_2(b) \end{vmatrix} = y_1(a)y_2(b) - y_1(b)y_2(a) = 0.$$

Nur wenn obige Randwert-Determinante D für die Fundamentallösungen Null ergibt, ist das Problem nichttrivial lösbar. Unter der Bedingung $D = 0$ gilt dann

$$c_2 = -c_1 \frac{y_1(a)}{y_2(a)},$$

so dass

$$y(x) = c\left(y_1(x) - \frac{y_1(a)}{y_2(a)} y_2(x) \right). \tag{3.8}$$

Das Randwertproblem ist also nur bis auf eine Konstante bestimmt. Wenn ein homogenes Randwertproblem lösbar ist, dann hat es "ähnliche" Lösungen, welche sich durch eine Konstante unterscheiden. Man spricht hier von einem Eigenwertproblem.

3.1.2 Inhomogenes Randwertproblem

Das inhomogene Randwertproblem

$$y''(x) + p\, y'(x) + q\, y(x) = r(x) \qquad \text{mit } y(a) = \alpha, \quad y(b) = \beta$$

hat die allgemeine Lösung

$$y(x) = c_1 y_1(x) + c_2 y_2(x) + y_p(x),$$

wenn $y_p(x)$ eine partikuläre Lösung darstellt. Die Konstanten c_i erhält man wieder aus den Randbedingungen

$$\left.\begin{aligned} y(a) &= c_1 y_1(a) + c_2 y_2(a) + y_p(a) = \alpha \\ y(b) &= c_1 y_1(b) + c_2 y_2(b) + y_p(b) = \beta \end{aligned}\right\}$$
$$\Rightarrow \begin{pmatrix} y_1(a) & y_2(a) \\ y_1(b) & y_2(b) \end{pmatrix} \begin{pmatrix} c_1 \\ c_2 \end{pmatrix} = \begin{pmatrix} \alpha - y_p(a) \\ \beta - y_p(b) \end{pmatrix}.$$

Damit dieses lineare Gleichungssystem für c_1, c_2 lösbar ist, muss die Randwert-Determinante

$$D \neq 0$$

sein. Dann gilt z.B. nach der Cramerschen Regel

$$c_1 = \frac{1}{D}\begin{vmatrix} \alpha - y_p(a) & y_2(a) \\ \beta - y_p(b) & y_2(b) \end{vmatrix}, \qquad c_2 = \frac{1}{D}\begin{vmatrix} y_1(a) & \alpha - y_p(a) \\ y_1(b) & \beta - y_p(b) \end{vmatrix}.$$

Also ist das inhomogene Randwertproblem nur lösbar, wenn das homogene nur die triviale Null-Lösung besitzt. Für alle Randwertprobleme, also auch für solche höherer als 2. Ordnung, gilt der **Fredholmsche Alternativsatz**

	RWP homogen	RWP inhomogen
$D = 0$	nichttriviale Lösungen existieren, sind aber nur bis auf eine Konstante eindeutig.	i.A. keine Lösung bei beliebigen Randwerten.
$D \neq 0$	nur triviale Lösung $y(x) = 0$ existiert.	eindeutig lösbar.

Die eigentlichen Randwertprobleme treten bei inhomogenen Gleichungen auf, für die $D \neq 0$ zu fordern ist. Diese werden in diesem Kapitel im Zentrum stehen.

3.2 Schießverfahren

Wir betrachten im Folgenden das Randwertproblem für eine Differenzialgleichung 2. Ordnung in der Form

$$y'' = f(x, y, y') \,, \tag{3.9}$$

$$y(a) = \alpha, \qquad y(b) = \beta \,. \tag{3.10}$$

Diese Differenzialgleichung 2. Ordnung schreiben wir als System 1. Ordnung um, wie wir dies schon im Kapitel 2 durchgeführt haben:

$$\mathbf{y}' = \mathbf{f}(x, \mathbf{y}) \tag{3.11}$$

$$\text{mit } \mathbf{y} = \begin{pmatrix} y_1 \\ y_2 \end{pmatrix} = \begin{pmatrix} y \\ y' \end{pmatrix}, \ \mathbf{f}(x, \mathbf{y}) = \begin{pmatrix} y_2 \\ f(x, y_1, y_2) \end{pmatrix}. \tag{3.12}$$

Die zugehörigen Randwerte ergeben sich dann in der Form

$$y_1(a) = \alpha, \qquad y_1(b) = \beta \,. \tag{3.13}$$

In Gleichung (3.7) hat in diesem Fall die Matrix $\mathbf{A}$ eine 1 in der ersten Zeile und ersten Spalte, die Matrix $\mathbf{B}$ eine 1 in der ersten Spalte und zweiten Zeile und sonst überall Nullen, der Vektor $\mathbf{r}$ besitzt die Komponenten α und β.

Bei einem Randwertproblem können wir nicht mehr einfach schrittweise vom Anfangspunkt bis zum Endpunkt durchrechnen wie bei einem Anfangswertproblem. Betrachten wir das AWP für die Differenzialgleichung (3.9), dann müssen zwei Anfangswerte vorgegeben sein, zum Beispiel

$$y(a) = \alpha, \qquad y'(a) = \gamma\,. \tag{3.14}$$

Das entspräche bei der Umformulierung in ein System 1. Ordnung (3.11) gerade der Vorgabe der Werte der beiden Komponenten $y_1(a) = \alpha$ und $y_2(a) = \gamma$ an der Stelle $x = a$. Dieses AWP (3.9), (3.14) können wir so deuten, dass neben dem Funktionswert die Steigung von y an der Stelle $x = a$ vorgegeben wird.

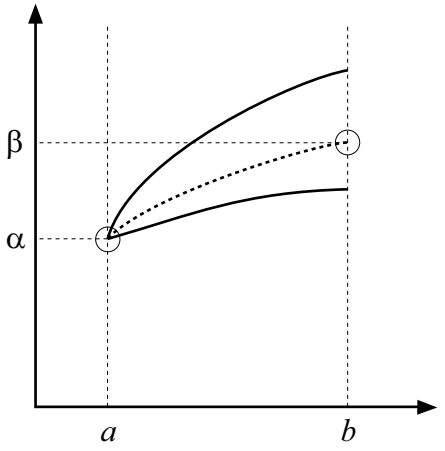


Abb. 3.2. Schießverfahren

Die Idee beim Schießverfahren (Shooting-Methode) ist somit, das AWP in der Art und Weise zu lösen, dass im Laufe des Verfahrens eine Steigung α_2 so gefunden wird, dass die Lösungskurve $y = y(x)$ durch (b, β) verläuft. In Abbildung 3.2 ist ein „Schuss“ über das Ziel, ein Schuss unter das Ziel und ein Treffer eingezeichnet. Alle Methoden zur Lösung der AWP als System 1. Ordnung in der Form (3.11), welche wir in Kapitel 2 kennen gelernt haben, können hier angewandt werden. Das RWP ist damit auf die Lösung von AWPen zurückgeführt. Es bleibt eine gute Strategie zu entwickeln, wie die „richtige“ Steigung von y im Punkt $x = a$ schnell gefunden wird.

Zusammenfassung: Die Idee beim Schießverfahren ist, statt dem Randwertproblem für die Differenzialgleichung 2. Ordnung (3.9), (3.10) das Anfangswertproblem für das System 1. Ordnung

$$y_1' = y_2 \,, \tag{3.15}$$
$$y_2' = f(x, y_1, y_2) \tag{3.16}$$

mit

$$y_1(a) = \alpha, \qquad y_2(a) = \gamma \,, \tag{3.17}$$

zu lösen und γ so zu bestimmen, dass $y_1(b) = \beta$ wird.

Wir betrachten zunächst lineare Differenzialgleichungen. Für ein lineares Problem ist die Strategie einfacher, da die Lösung linear aus zwei Lösungen des Anfangswertproblems zusammengesetzt werden kann.

3.2.1 Lineare Probleme

Die numerische Lösung eines linearen Randwertproblems einer Differenzialgleichung 2. Ordnung in der Form (3.4) mit den Randwerten (3.17) wird im Folgenden beschrieben. Diese Lösung bezeichnen wir mit y_h. Wie oben kurz beschrieben erhalten wir die numerische Lösung dieses Randwertproblems mit Hilfe der numerischen Lösung des Anfangswertproblems für das System 1. Ordnung (3.15). Diese Lösung ist ein Vektor $\mathbf{y}_h$ mit den Komponenten $y_{h,1}$ und $y_{h,2}$. Dabei stimmt die erste Komponente des Vektors $\mathbf{y}$ bzw. des Vektors mit den Näherungswerten $\mathbf{y_h}$ mit der Lösung der Differenzialgleichung 2. Ordnung (3.9) y_h überein: $y_h = y_{h,1}$.

1. Schritt: Löse zweimal das Anfangswertproblem (3.15) mit den Anfangswerten $y_1(a) = \alpha$ und $y_2(a) = \gamma^{(1)}$ bzw. $y_2(a) = \gamma^{(2)}$. Dabei sind die Anfangswerte $\gamma^{(1)}$ und $\gamma^{(2)}$ beliebig gewählt. Sie stellen Anfangswerte für die Steigung der Lösung $y = y(x)$ dar, da $y_2(a) = y'(a)$. Dies liefert die numerischen Lösungen $\mathbf{y}_h^{(1)}$ und $\mathbf{y}_h^{(2)}$:

$$\gamma^{(1)} \Longrightarrow \mathbf{y}_h^{(1)} \,,$$
$$\gamma^{(2)} \Longrightarrow \mathbf{y}_h^{(2)} \,.$$

2. Schritt: Superposition:

$$\mathbf{y}(x) = K_1 \mathbf{y}_h^{(1)} + K_2 \mathbf{y}_h^{(2)} \,. \tag{3.18}$$

K_1 und K_2 werden so bestimmt, dass die richtigen Randwerte angenommen werden:

$$x = a: \qquad K_1 y_{h,1}^{(1)}(a) + K_2 y_{h,1}^{(2)}(a) = \alpha \ ,$$
$$x = b: \qquad K_1 y_{h,1}^{(1)}(b) + K_2 y_{h,1}^{(2)}(b) = \beta \ .$$

Dies ist ein lineares Gleichungssystem, welches sich nach K_1 und K_2 auflösen lässt. Man erhält:

$$K_1 = \frac{y_{h,1}^{(2)}(b) - \beta}{y_{h,1}^{(2)}(b) - y_{h,1}^{(1)}(b)}, \qquad K_2 = 1 - K_1 \tag{3.19}$$

und damit die gesuchte Näherungslösung.

Lineares Schießverfahren

Real: $\alpha, \beta, a, b, h, K_1, K_2, steig1, steig2$
Real Feld $(0..NN)$: $y, y^{(1)}, y^{(2)}, x$
Integer: N, i
Übernehme $(a, b, \alpha, \beta, N, steig1, steig2)$
$h = \dfrac{b-a}{N}$
$i = 0$
Solange $i \leq N$
$x_i = a + ih$
$i = i + 1$
AWP-Löser$(x, y^{(1)}, \alpha, steig1, N)$
AWP-Löser$(x, y^{(2)}, \alpha, steig2, N)$
$K_1 = \dfrac{y_N^{(2)} - \beta}{y_N^{(2)} - y_N^{(1)}}$
$K_2 = 1 - K_1$
$i = 0$
Solange $i \leq N$
$y_i = K_1 y_i^{(1)} + K_2 y_i^{(2)}$
$i = i + 1$
Gib (x_i, y_i) für $i = 1, 2, \ldots, N$ zurück

Abb. 3.3. Das Schießverfahren für ein lineares Problem 2. Ordnung

Ein Unterprogramm in Form eines Struktogramms ist in Abbildung 3.3 angegeben.

3.2.2 Nichtlineare Probleme

Für das nichtlineare Randwertproblem (3.9), (3.10) wird das Verfahren etwas aufwändiger. Die Vorgehensweise erfolgt dann iterativ. Es wird eine Folge $\gamma^{(\nu)}$ erzeugt, die gegen die Steigung konvergiert, für welche die zugehörige Lösung den Randwert $y(b) = \beta$ erfüllt. Wir bezeichnen wiederum mit $y = y(x)$ die Lösung der Differenzialgleichung 2. Ordnung (3.9) und mit dem Vektor $\mathbf{y} = \mathbf{y}(x)$ die Lösung des zugehörigen Systems 1. Ordnung (3.11). Es gilt dabei insbesondere $y = y_1$, wobei y_1 die erste Komponente des Vektors $\mathbf{y}$ ist. Bezeichnen wir die numerische Lösung des AWP (3.15) mit dem freien Parameter $\gamma^{(\nu)}$ als $y_h(x;\gamma^{(\nu)})$, dann suchen wir das $\gamma^{(\nu)}$, für welches $y_h(b;\gamma^{(\nu)}) \approx \beta$ im Rahmen der gewünschten Genauigkeit gilt. Der Wert $y_h(x;\gamma^{(\nu)})$ ist dabei der Wert der ersten Komponente des Vektors $\mathbf{y}_h(x;\gamma^{(\nu)})$. Anders ausgedrückt: Wir suchen die Nullstelle der Funktion

$$\varphi = \varphi(\gamma^{(\nu)}) = y_h(b;\gamma^{(\nu)}) - \beta \,. \tag{3.20}$$

Hier wurde der Einfachheit halber für die numerische Lösung $y_h = y_h(x)$ geschrieben, obwohl diese nur an den Gitterpunkten definiert ist. Das nichtlineare Schießverfahren läuft jetzt in den folgenden drei Schritten ab.

1. Schritt: Es werden zwei Testrechnungen mit vorgegebenen beliebigen Werten für $\gamma^{(1)}$ und $\gamma^{(2)}$ berechnet. Die numerische Lösung des AWP liefert dann

$$\gamma^{(1)} \Longrightarrow y_h(b;\gamma^{(1)}) := y_b^{(1)} \,,$$
$$\gamma^{(2)} \Longrightarrow y_h(b;\gamma^{(2)}) := y_b^{(2)} \,,$$

wobei $y_b^{(1)}$ und $y_b^{(2)}$ die Näherungswerte an der Stützstelle $x = b$ bezeichnen.

2. Schritt: Berechne

$$y_b^{(1)} - \beta, \qquad y_b^{(2)} - \beta \,.$$

3. Schritt: Auffinden der Nullstelle. Im Allgemeinen wird keiner der Werte im Rahmen der gewünschten Genauigkeit ungefähr Null sein. Als einfachste Methode, eine neue Steigung anzugeben, wenden wir das Bisektionsverfahren an. Gilt

$$(y_b^{(1)} - \beta)(y_b^{(2)} - \beta) < 0 \,, \tag{3.21}$$

dann sei

$$\gamma^{(neu)} = \frac{1}{2}(\gamma^{(1)} + \gamma^{(2)}) \,. \tag{3.22}$$

Nun setzt man

$$\gamma^{(2)} := \gamma^{(neu)}, \text{ falls } (y_b^{(neu)} - \beta)(y_b^{(1)} - \beta) < 0$$

oder

$$\gamma^{(1)} := \gamma^{(neu)}, \text{ falls } (y_b^{(neu)} - \beta)(y_b^{(2)} - \beta) < 0$$

und lässt den anderen Wert ungeändert. Danach geht es mit Schritt 1 wieder weiter: Es wird zu der neuen Anfangssteigung wiederum das Anfangswertproblem gelöst und die Schritte 2 und 3 ausgeführt bis die gesuchte Genauigkeit erreicht ist.

Ist (3.21) nicht erfüllt, dann liegt die gesuchte Nullstelle nicht im Intervall, welches durch die Grenzen $\gamma^{(1)}$ und $\gamma^{(2)}$ gebildet wird. Man muss dann das betrachtete Intervall vergrößern. Es wird dabei vorausgesetzt, dass es genau eine Nullstelle gibt. Nimmt man den Schnittpunkt der Geraden durch die Punkte $(\gamma^{(1)}, y_b^{(1)})$ und $(\gamma^{(2)}, y_b^{(2)})$ mit der γ-Achse, so erhält man

$$\gamma = \gamma^{(1)} - \frac{(y_b^{(1)} - \beta)(\gamma^{(2)} - \gamma^{(1)})}{y_b^{(2)} - y_b^{(1)}} \, .$$

Man setzt dann $\gamma^{(1)} = \gamma$ oder $\gamma^{(2)} = \gamma$ und versucht damit, die Nullstellen einzufangen.

Bemerkung: Das Bisektionsverfahren konvergiert recht langsam. Man wird hier im Allgemeinen bessere Iterationsverfahren für nichtlineare Gleichungssysteme einsetzen. Eine kurze Übersicht und weiterführende Literatur sind in Anhang B aufgeführt.

Beispiel 3.1 (Balkenbiegung, mit MAPLE-Worksheet). Das Schießverfahren wird auf das *nichtlineare* Randwertproblem

$$y''(x) = -\frac{q}{2} \frac{x(x-1)}{EI} (1 + y'(x)^2)^{(\frac{3}{2})}, \quad y(0) = y(1) = 0 \tag{3.23}$$

angewendet. Dies entspricht der Differenzialgleichung (3.1), welche die Biegung eines Balkens der Länge 1 beschreibt und in Kapitel 2 und am Anfang dieses Kapitels diskutiert wird. Vorgegeben ist hier das Biegemoment $M(x) = \frac{q}{2}x(x-1)$, der Elastizitätsmodul E und das Trägheitsmoment I sind als konstant angenommen. In dem MAPLE-Worksheet ist das iterative Verfahren in Form einer Animation visualisiert. □

Beispiel 3.2 (Balkenbiegung, mit MAPLE-Worksheet). Ist das Biegemoment klein, dann kann das Quadrat der Ableitung der Durchbiegung y in Beispiel 3.1 vernachlässigt werden.

Die nichtlineare Differenzialgleichung wird dann linear:

$$y''(x) = -\frac{q}{2}\,\frac{x(x-1)}{EI}, \quad y(0) = y(1) = 0\,.$$

Dabei ist das Biegemoment wieder mit $M(x) = \frac{q}{2}x(x-1)$ angegeben. Zur Validierung des Verfahrens für die nichtlineare Gleichung in Beispiel 3.1 ist dies ein guter Test. Mit den vorgegebenen Daten ist man in dem Bereich, in der die Linearisierung noch gut anwendbar ist. Die Ergebnisse mit dem nichtlinearen und dem linearen Verfahren stimmen sehr gut überein. Da die rechte Seite im linearisierten Fall nicht von y abhängt, kann man diese Lösung durch Integration erhalten. □

3.3 Differenzenverfahren

Bei der Shooting-Methode zur Lösung von Randwertproblemen variiert man die Anfangsbedingung $y'(a) = \gamma^{(\nu)}$ so lange, bis sich die zweite Randbedingung $y(b) = \beta$ einstellt. Jetzt werden wir das Randwertproblem direkt angehen, indem wir die Ableitungen der Differenzialgleichung durch Näherungsausdrücke ersetzen und beide Randbedingungen $y(a) = \alpha$ und $y(b) = \beta$ berücksichtigen.

Im Kapitel über die numerische Differenziation 1.6 werden Formeln angegeben, wie man aus einer direkten Vorgabe von Funktionswerten y_{i-1}, y_i, y_{i+1} Ableitungen näherungsweise bestimmen kann. Für die erste und zweite Ableitung einer Funktion $y = y(x)$ an der Stelle x_i gelten die Näherungen

$$y_i' \approx \frac{1}{2\Delta x}(y_{i+1} - y_{i-1})\,, \tag{3.24}$$

$$y_i'' \approx \frac{1}{\Delta x^2}(y_{i-1} - 2y_i + y_{i+1})\,. \tag{3.25}$$

Obige Näherungsformeln gelten für eine äquidistante Unterteilung; in gleicher Weise lassen sich aber auch nicht-äquidistante Stützstellen berücksichtigen.

Die grundlegende Idee bei einem Differenzenverfahren ist, die Ableitungen in der Differenzialgleichung durch Differenzenquotienten zu ersetzen. Aus der Differenzialgleichung ergibt sich somit ein System von Differenzengleichungen an den Stützstellen. Im Folgenden werden wir das Vorgehen der Differenzenmethode am Beispiel der Knickbiegung demonstrieren.

Beispiel 3.3 (Knickbiegung). Den Balken mit veränderlicher Steifigkeit $EI(\xi)$ unter veränderlicher Querbelastung $p(\xi)$ und Axiallast P beschreibt das Randwertproblem 2. Ordnung (siehe [2])

$$\frac{d^2M(\xi)}{d\xi^2} + \frac{P}{EI(\xi)}M(\xi) = -p(\xi) \qquad \text{mit} \qquad M(-l) = M(l) = 0\,. \tag{3.26}$$

Ist z.B.

$$EI(\xi) = \frac{EI_0}{1 + (\frac{\xi}{l})^2}, \quad p(\xi) = p_0 = konstant$$

und führen wir die dimensionslosen Größen

$$x := \frac{\xi}{l}, \qquad y := \frac{M}{l^2 p_0}, \qquad c := \frac{p_0 l^2 P}{EI_0} = konstant = 1$$

ein, folgt das Randwertproblem

$$y''(x) + c\,(1 + x^2)\,y(x) = -1 \qquad \text{für } x \in [-1, 1]\,, \tag{3.27}$$

$$y(-1) = y(1) = 0\,. \tag{3.28}$$

Es ist ein inhomogenes lineares Randwertproblem mit homogenen Randbedingungen. □

An diesem Randwertproblem werden wir die verschiedenen Schritte in einem Differenzenverfahren im Folgenden erklären.

1. Schritt: Unterteilung des Intervalls $[-1, 1]$ **in** n **Teilintervalle**. Für unser Beispiel wählen wir $n = 6$, d.h. $h = \Delta x = \frac{b-a}{6} = \frac{1}{3}$.

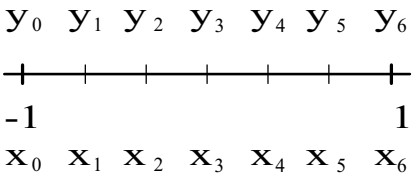


Abb. 3.4. Intervallunterteilung

2. Schritt: Ersetzen der Ableitungen durch Differenzenquotienten. Für jeden inneren Punkt x_i des Intervalls gilt die Differenzialgleichung:

$$y''(x_i) + c\,(1 + x_i^2)\,y(x_i) = -1\ .$$

Für die Randpunkte gelten die Randbedingungen $y(x_0) = y(x_6) = 0$. Bezeichnen wir die Näherungswerte an der Stelle x_i mit y_i und ersetzen die kontinuierliche Ableitung durch Differenzenformeln, gilt für $i = 1, \ldots, 5$:

$$\frac{1}{h^2}(y_{i-1} - 2y_i + y_{i+1}) + c\,(1 + x_i^2)\,y_i = -1$$

bzw.

$$\frac{1}{h^2}\,y_{i-1} + (c\,(1 + x_i^2) - \frac{2}{h^2})y_i + \frac{1}{h^2}\,y_{i+1} = -1\ .$$

Damit erhalten wir 5 Gleichungen für die 5 unbekannten Funktionswerte $y_1, \ldots, y_5$:

$$\begin{aligned}
x_1: &\quad \tfrac{1}{h^2}\,y_0 + (c(1 + x_1^2) - \tfrac{2}{h^2})y_1 + \tfrac{1}{h^2}\,y_2 = -1\ ,\\
x_2: &\quad \tfrac{1}{h^2}\,y_1 + (c(1 + x_2^2) - \tfrac{2}{h^2})y_2 + \tfrac{1}{h^2}\,y_3 = -1\ ,\\
x_3: &\quad \tfrac{1}{h^2}\,y_2 + (c(1 + x_3^2) - \tfrac{2}{h^2})y_3 + \tfrac{1}{h^2}\,y_4 = -1\ ,\\
x_4: &\quad \tfrac{1}{h^2}\,y_3 + (c(1 + x_4^2) - \tfrac{2}{h^2})y_4 + \tfrac{1}{h^2}\,y_5 = -1\ ,\\
x_5: &\quad \tfrac{1}{h^2}\,y_4 + (c(1 + x_5^2) - \tfrac{2}{h^2})y_5 + \tfrac{1}{h^2}\,y_6 = -1\ .
\end{aligned}$$

Für die numerische Rechnung setzen wir $c = 1$. Berücksichtigen wir nun, dass $y_0 = y(-1) = 0$, $y_6 = y(1) = 0$ bzw. $h^2 = \frac{1}{9}$ und $x_i = -1 + ih = -1 + i\frac{1}{3}$ gilt, so erhalten wir

$$\begin{aligned}
-\frac{149}{9}\,y_1 + 9y_2 &= -1\ ,\\
+9y_1 - \frac{152}{9}\,y_2 + 9y_3 &= -1\ ,\\
+9y_2 - 17y_3 + 9y_4 &= -1\ ,\\
+9y_3 - \frac{152}{9}\,y_4 + 9y_5 &= -1\ ,\\
+9y_4 - \frac{149}{9}\,y_5 &= -1\ .
\end{aligned}$$

Dies ist ein lineares inhomogenes Gleichungssystem für die 5 Unbekannten $y_1, \ldots, y_5$:

$$\left(\begin{array}{ccccc|c}
-\frac{149}{9} & 9 & 0 & 0 & 0 & -1\\
9 & -\frac{152}{9} & 9 & 0 & 0 & -1\\
0 & 9 & -17 & 9 & 0 & -1\\
0 & 0 & 9 & -\frac{152}{9} & 9 & -1\\
0 & 0 & 0 & 9 & -\frac{149}{9} & -1
\end{array}\right)\ .$$

3. Schritt: Lösen des LGS mit geeigneten Methoden. Die Matrix auf der linken Seite ist eine **Tridiagonalmatrix**. Man nennt daher das System auch ein **tridiagonales Gleichungssystem**. Die Auflösung des LGS erfolgt mit dem Gauß-Algorithmus, der hier zu einem besonders einfachen Algorithmus wird, dem **Tridiagonal-** bzw. **Thomas-Algorithmus**.

Gehen wir zur Beschreibung des Thomas-Algorithmus von einem allgemeinen tridiagonalen LGS aus

$$\left(\begin{array}{cccccc|c} b_1 & c_1 & 0 & \dots & 0 & 0 & r_1 \\ a_2 & b_2 & c_2 & \dots & 0 & 0 & r_2 \\ \vdots & \ddots & \ddots & \ddots & \ddots & \vdots & \vdots \\ \vdots & \ddots & \ddots & \ddots & \ddots & \vdots & \vdots \\ 0 & 0 & \dots & a_{m-1} & b_{m-1} & c_{m-1} & r_{m-1} \\ 0 & 0 & \dots & 0 & a_m & b_m & r_m \end{array}\right) .$$

Dann lautet die Gauß-Elimination

$$\left.\begin{array}{l} q := \frac{a_i}{b_{i-1}} \\ b_i := b_i - q c_{i-1} \\ r_i := r_i - q r_{i-1} \end{array}\right\} i = 2, \dots, m ,$$

und das Rückwärtsauflösen erfolgt durch

$$\begin{aligned} x_m &:= \frac{r_m}{b_m} \\ x_i &:= \frac{r_i - c_i x_{i+1}}{b_i} \qquad i = m-1, \dots, 1 . \end{aligned}$$

In der Prozedur **Thomas** wird dieser Algorithmus in MAPLE realisiert. Im **Worksheet zum Algorithmus** wird er verwendet, um das obige LGS zu lösen und anschließend die numerische Lösung grafisch darzustellen. Die gesuchten Näherungswerte an den Stellen x_i $(i = 1, \dots, 5)$ sind:

$$y_1 = 0.51724, y_2 = 0.84035, y_3 = 0.94861, y_4 = 0.84035, y_5 = 0.51724$$

vervollständigt durch die Randwerte $y_0 = 0, y_6 = 0$.

4. Schritt: Darstellung der Ergebnisse. Den angenäherten Verlauf der Lösung zwischen den Stützstellen erhält man z.B. durch lineare Interpolation, wie er in Abbildung 3.5 angegeben ist.

Bemerkungen:

1. Für $n \to \infty$, also für $h \to 0$, geht der systematische Fehler beim Ersetzen der kontinuierlichen Ableitung durch Differenzenformeln gegen Null, so

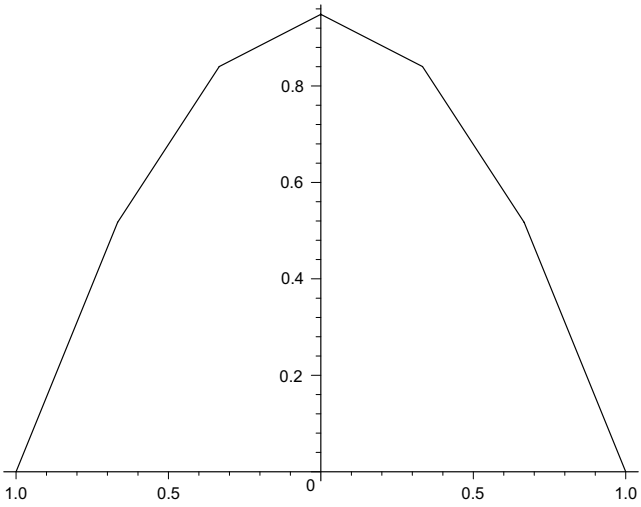


Abb. 3.5. Numerische Lösung für die Knickbiegung

dass die numerische Lösung die exakte immer besser approximiert. Hier gelten die analogen Aussagen wie für die Anfangswertprobleme. Auch hier spielt die Stabilität der Approximation eine wichtige Rolle. Ist die Stabilität der numerischen Approximation gesichert, dann wird die Näherungslösung gegen die exakte Lösung konvergieren. Wie schon in der Theorie exakter Lösungen von Randwertproblemen sind theoretische Aussagen auch für Näherungsverfahren bei Randwertproblem deutlich schwieriger als bei Anfangswertproblem. Wir werden dies hier nicht vertiefen und auf die Spezialliteratur verweisen.

2. Die Struktur des LGS ändert sich nicht, wenn man für n größere Werte wählt.

3. Die Differenzenverfahren sind einfach anzuwenden und laufen in den folgenden Schritten ab:

   a) Man teilt das Intervall in die diskreten Punkte x_i, den Stützstellen, ein.

   b) Man ersetzt die kontinuierlichen Ableitungen der Differenzialgleichung durch Differenzenquotienten an den verschiedenen Stützstellen.

c) Somit erhält man ein Gleichungssystem für die unbekannten Näherungswerte y_i an den Stützstellen.

d) Falls erforderlich werden die diskreten Näherungswerte interpoliert, um eine kontinuierliche Näherungsfunktion zu erhalten.

4. Das Auflösen des Gleichungssystems ist sehr einfach, falls die Differenzialgleichung linear und nicht von höherer als 2. Ordnung ist. Für nichtlineare Differenzialgleichungen erhält man nichtlineare Gleichungssysteme die wesentlich schwieriger aufzulösen sind. Man muss hier die entsprechenden Verfahren einsetzen, wie sie im Anhang B aufgeführt werden. Für Differenzialgleichungen höherer Ordnung müssen am Rand teilweise unsymmetrische Formeln gewählt bzw. auch höhere Ableitungen berücksichtigt werden. Hierzu kann man dann auch Differenzenformeln mit Ableitungen einsetzen, die im nächsten Abschnitt eingeführt werden.

5. Die MAPLE-Prozedur **DiffFormeln** erzeugt Differenzenformeln für die n-te Ableitung unter Berücksichtigung von Funktionswerten $(y_{i-k}, \ldots, y_i, \ldots, y_{i+l})$. Die Prozedur **DiffFormelnOrdnung** bestimmt die Ordnung dieser Formeln durch Taylor-Abgleich. Exemplarisch wird die Verwendung dieser Prozeduren im zugehörigen **Worksheet** aufgezeigt.

3.4 Differenzenformeln mit Ableitungen

Gegeben ist das Problem eines Balkens auf elastischer Bettung (z.B. [9])

$$cy^{(4)}(x) + y(x) = 1 \qquad \text{in } [0,1]\,, \tag{3.29}$$

$$y(0) = y(1) = 0, \quad y''(0) = y''(1) = 0\,. \tag{3.30}$$

Wenden wir auf dieses Problem das Differenzenverfahren an, dann müssen wir im ersten Schritt das Intervall $[0,1]$ in Teilintervalle $x_0 = 0 < x_1 < \cdots < x_n = 1$ unterteilen und die 4. Ableitung durch eine Differenzenformel ersetzen. Um auf eine geeignete Näherung zu kommen, wählt man den Ansatz

$$y^{(4)}(x_1) \approx \alpha_0 y(x_0) + \alpha_1 y(x_1) + \alpha_2 y(x_2) + \alpha_3 y(x_3) + \alpha_4 y(x_4)\,. \tag{3.31}$$

Man beachte, dass wir im Ansatz mindestens 5 freie Parameter benötigen, da es sich um die 4. Ableitung handelt. Allerdings geht im obigen Ansatz nur $y(x_0)$ ein aber nicht die weitere Randbedingung $y''(0)$! Daher modifizieren wir obigen Ansatz zu

$$y^{(4)}(x_1) \approx \alpha_0 y(x_0) + \alpha_1 y(x_1) + \alpha_2 y(x_2) + \alpha_3 y(x_3) + \underline{\alpha_4 y''(x_0)}\,. \tag{3.32}$$

Denn dann werden bei der Diskretisierung der DGL beide Randbedingungen berücksichtigt. Um Bedingungen für die Konstanten α_i zu erhalten, wählen wir einen Taylor-Abgleich. Wir entwickeln $y(x)$ in eine Taylor-Reihe mit Entwicklungspunkt x_1. Anschließend sortieren wir nach den Ableitungen von y an der Stelle x_1 und bestimmen die α_i so, dass bis auf die 4. Ableitung alle anderen verschwinden.

Zur Vereinfachung der Formeln setzen wir

$$x_0 = x_1 - h, \qquad x_2 = x_1 + h, \qquad x_3 = x_1 + 2h \; .$$

Die Taylor-Entwicklung von $y(x)$ lautet an der Stelle x_1

$$y(x) = \sum_{n=0}^{N} \frac{1}{n!} y^{(n)}(x_1)(x - x_1)^n + \mathcal{O}(h^{N+1}) \; . \tag{3.33}$$

Speziell für die Punkte x_0, x_1, x_2 gilt mit $y^{(k)}(x_1) = y_1^{(k)}$

$$\begin{aligned}
x = x_0 \; &= x_1 - h \; : y(x_0) = y_1 - y_1' h + \frac{1}{2!} y_1'' h^2 - \frac{1}{3!} y_1''' h^3 + \frac{1}{4!} y_1^{(4)} h^4 \pm \dots \; , \\
x = x_1 \; &\qquad\quad : y(x_1) = y_1 \; , \\
x = x_2 \; &= x_1 + h \; : y(x_2) = y_1 + y_1' h + \frac{1}{2!} y_1'' h^2 + \frac{1}{3!} y_1''' h^3 + \frac{1}{4!} y_1^{(4)} h^4 + \dots \; , \\
x = x_3 = x_1 + 2h \; &: y(x_3) = y_1 + y_1' 2h + \frac{1}{2!} y_1'' 4h^2 + \frac{1}{3!} y_1''' 8h^3 + \frac{1}{4!} y_1^{(4)} 16h^4 + \dots \; .
\end{aligned}$$

Außerdem gilt nach Gleichung (3.33)

$$y''(x) = \sum_{n=0}^{N-2} \frac{1}{n!} y^{(n+2)}(x_1)(x - x_1)^n + \mathcal{O}(h^{N-1}) \; .$$

Speziell für $y_0'' = y''(x_0)$ gilt somit

$$x = x_0 = x_1 - h : \qquad y''(x_0) = y_1'' - y_1''' h + \frac{1}{2!} y_1^{(4)} h^2 - \frac{1}{3!} y_1^{(5)} h^3 \pm \dots \; .$$

Setzt man die Taylor-Entwicklungen in den Ansatz (3.32) ein und sortiert nach y_1, y_1', y_1'', y_1''', erhält man

$$y^{(4)}(x_1) \approx \underbrace{(\alpha_0+\alpha_1+\alpha_2+\alpha_3)}_{0} y_1 + \underbrace{(-h\alpha_0+h\alpha_2+2h\alpha_3)}_{0} y_1' + \\ + \underbrace{(\frac{1}{2!}h^2\alpha_0+\frac{1}{2!}h^2\alpha_2+\frac{1}{2!}4h^2\alpha_3+\alpha_4)}_{0} y_1'' + \\ + \underbrace{(-\frac{1}{3!}h^3\alpha_0+\frac{1}{3!}h^3\alpha_2+\frac{1}{3!}8h^3\alpha_3-h\alpha_4)}_{0} y_1''' + \\ + \underbrace{(\frac{1}{4!}h^4\alpha_0+\frac{1}{4!}h^4\alpha_2+\frac{1}{4!}16h^4\alpha_3+\frac{1}{2!}h^2\alpha_4)}_{1} y_1'''' + \dots \ .$$

Damit hat man 5 lineare inhomogene Gleichungen für die 5 unbekannten Gewichte $\alpha_0, \alpha_1, \alpha_1, \alpha_3, \alpha_4$. Die Lösung des LGS lautet:

$$\alpha_0 = -\frac{24}{11}\frac{1}{h^4}, \quad \alpha_1 = \frac{60}{11}\frac{1}{h^4}, \quad \alpha_2 = -\frac{48}{11}\frac{1}{h^4}, \quad \alpha_3 = \frac{12}{11}\frac{1}{h^4}, \quad \alpha_4 = \frac{12}{11}\frac{1}{h^2} \ . \tag{3.34}$$

Die Rechnung ist im MAPLE-**Worksheet** ausgeführt. Die Ordnung der Differenzenformel ist 1, d.h.

$$y^{(4)}(x_1) = \alpha_0 y_0 + \alpha_1 y_1 + \alpha_2 y_2 + \alpha_3 y_3 + \alpha_4 y_0'' + \mathcal{O}(h) \ . \tag{3.35}$$

Wird eine höhere Genauigkeit bzw. Ordnung des Verfahrens gewünscht, müssen noch weitere Stützpunkte in die Formel hinzugenommen werden, wie dies in zugehörigen **Worksheet** realisiert ist.

Bemerkungen:

1. Das oben beschriebene Verfahren lässt sich auf Differenzenformeln für die n-te Ableitung verallgemeinern, wobei dann auch für die Randwerte Ableitungen bis zur Ordnung $n-1$ zugelassen sind. Die Prozedur **FDFormeln** erstellt für eine vorgegebene Anzahl von Stützstellen und Ableitungen die Koeffizienten für die Differenzenformel der n-ten Ableitung. Die Prozedur **FDOrdnung** bestimmt hierzu die Ordnung der Formel. Die Prozedur FDFormeln ist so allgemein gehalten, dass Ableitungen niedriger Ordnung an beliebigen Stützstellen berücksichtigt werden können.

2. Speziell durch den Ansatz

   $$\alpha_{-1} y_{i-1} + \alpha_0 y_i + \alpha_1 y_{i+1} + A_{-1} y_{i-1}' + A_0 y_i' + A_1 y_{i+1}' = 0 + \mathcal{O}(h^4)$$

   erhält man die **Mehrstellenansatz** in engerem Sinne. Der Unterschied zu den diskutierten Differenzenformeln besteht darin, dass die Bestimmungsgleichungen für die Koeffizienten $\alpha_{-1}, \alpha_0, \alpha_1, A_{-1}, A_0, A_1$ zu einem

homogenen LGS führen, so dass eine Konstante (z.B. $\alpha_1 = 1$) frei wählbar ist.

Beispiel 3.4 (Knickbiegung, mit MAPLE-Worksheet). Wir wenden jetzt die Differenzenmethode auf das Balkenproblem an:

$$y^{(4)}(x) + y(x) = 1; \qquad y(0) = y(1) = 0 \\ y''(0) = y''(1) = 0.$$

1. Schritt: Unterteilung des Intervalls $[0, 1]$ in $n = 6$ Teilintervalle:

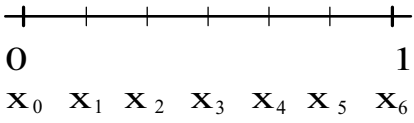


Abb. 3.6. Intervallunterteilung

2. Schritt: Für jeden inneren Punkt x_i $(i = 1 \dots 5)$ des Intervalls ersetzen wir die Differenzialgleichung durch Differenzenformeln. Dabei unterscheiden wir zwischen randnahen und randfernen Punkten. An den randnahen Punkten, x_1 und x_5, diskretisieren wir die DGL unter Berücksichtigung der Randbedingungen $y_0 = 0, y_0'' = 0$ bzw. $y_6 = 0, y_6'' = 0$. An den Rand fernen Punkten, x_2, x_3, x_4, ersetzen wir dabei die 4. Ableitung durch eine symmetrische Formel:

$$\begin{aligned}
x_1 &: -\tfrac{12}{11}\tfrac{1}{h^4}(-2y_0 + 5y_1 - 4y_2 + y_3 - y_0'' h^2) + y_1 = 1\,, \\
x_2 &: \quad \tfrac{1}{h^4}(y_0 - 4y_1 + 6y_2 - 4y_3 + y_4) + y_2 \quad = 1\,, \\
x_3 &: \quad \tfrac{1}{h^4}(y_1 - 4y_2 + 6y_3 - 4y_4 + y_5) + y_3 \quad = 1\,, \\
x_4 &: \quad \tfrac{1}{h^4}(y_2 - 4y_3 + 6y_4 - 4y_5 + y_6) + y_4 \quad = 1\,, \\
x_5 &: -\tfrac{12}{11}\tfrac{1}{h^4}(-y_3 + 4y_4 - 5y_5 + 2y_6 - y_6'' h^2) + y_5 = 1\,.
\end{aligned}$$

3. Schritt: Im zugehörigen MAPLE-**Worksheet** werden die Gleichungen mit Hilfe der Prozedur **FDFormeln** aufgestellt und das LGS für $y_1, \dots, y_5$ mit $y_0'' = y_6'' = 0$, $y_0 = y_6 = 0$, $h = \frac{b-n}{n} = \frac{1}{6}$ gelöst. Es ergibt sich

$$y_1 = 6.522 \cdot 10^{-3}, \quad y_2 = 11.198 \cdot 10^{-3}, \quad y_3 = 12.884 \cdot 10^{-3}\,, \\ y_4 = 11.198 \cdot 10^{-3}, \quad y_5 = 6.522 \cdot 10^{-3}\,.$$

4. Schritt: In Abbildung 3.7 wird die numerisch berechnete Lösung mit der von MAPLE bestimmten exakten Lösung des Randwertproblems grafisch verglichen. Es ergibt sich in den Rechenpunkten eine erstaunlich gute Übereinstimmung. Die Übereinstimmung ist selbst für $n = 4$ noch zu erkennen.

3.5 Eigenwertproblem

Homogene Randwertprobleme führen auf sogenannte Eigenwertprobleme. Diese treten zum Beispiel in der Elastizitätstheorie oder bei Schwingungsproblemen auf. Wir zeigen im Folgenden am Beispiel des einseitig geführten, anderseitig eingespannten **Knickstabs** [2], dass Eigenwertprobleme mit der Differenzenmethode ebenfalls behandelt werden können. Die Approximation des Randwertproblems für die gewöhnliche Differenzialgleichung führt auf das Problem der Bestimmung der Eigenwerte von Matrizen. Wir werden dies exemplarisch an dem Beispiel ausführen; für das Lösen des Eigenwertproblems für Matrizen dann aber auf die Literatur verweisen.

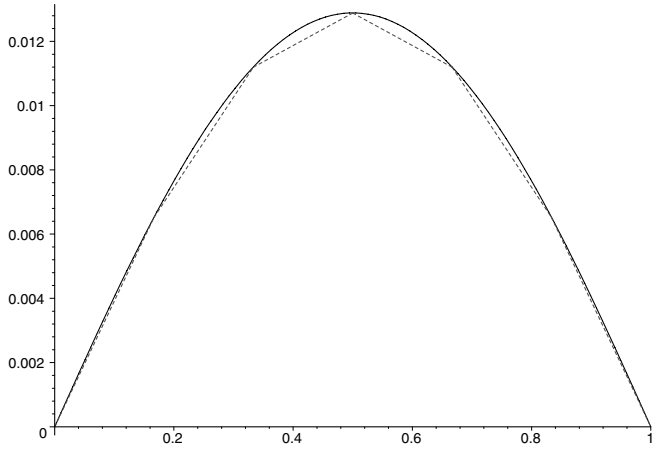


Abb. 3.7. Vergleich der numerischen Lösung (gestrichelt) mit der exakten

Das mathematische Modell des Knickstabes lautet

$$y^{(4)}(x) + \lambda^2 y''(x) = 0 \qquad \text{in } [0,1]\,, \tag{3.36}$$

$$y(0) = y(1) = 0,\ y''(0) = 0,\ y'(1) = 0\,. \tag{3.37}$$

Die physikalische Situation ist in Abbildung 3.8 dargestellt. In dem mathematischen Modell (3.36) ist λ ein zunächst unbestimmter Parameter. Eine Lösung des homogenen mathematischen Modells ist sofort klar: $y \equiv 0$, d.h. die triviale Lösung. Es geht nun darum, alle Werte von λ zu finden, für die mehr als diese Null-Lösung existiert, und die zugehörigen Lösungen anzugeben. Diese speziellen Werte λ heißen dann Eigenwerte und die zugehörigen Lösungen Eigenlösungen.

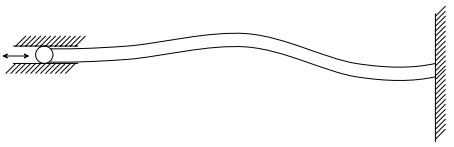


Abb. 3.8. Knickstab

3.5.1 Exakte Lösung der Differenzialgleichung

Wie man durch Einsetzen in die Differenzialgleichung nachprüfen kann, sind

$$sin(\lambda x),\ cos(\lambda x),\ x,\ 1$$

Lösungen, welche auch linear unabhängig sind. Die allgemeine Lösung der Differenzialgleichung lautet damit

$$y(x) = c_1 sin(\lambda x) + c_2 cos(\lambda x) + c_3 x + c_4\ . \tag{3.38}$$

Setzen wir die Randbedingungen ein, so erhalten wir für die Konstanten c_i ein lineares Gleichungssystem

$$\begin{array}{l} y(0) \to \\ y(1) \to \\ y'(1) \to \\ y''(0) \to \end{array} \begin{pmatrix} 0 & 1 & 0 & 1 \\ sin(\lambda) & cos(\lambda) & 1 & 1 \\ \lambda cos(\lambda) & -\lambda sin(\lambda) & 1 & 0 \\ 0 & \lambda^2 & 0 & 0 \end{pmatrix} \begin{pmatrix} c_1 \\ c_2 \\ c_3 \\ c_4 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ 0 \\ 0 \end{pmatrix}.$$

Damit das Eigenwertproblem nichttrivial lösbar ist, muss die Randwert-Determinante Null ergeben. Die Entwicklung nach der letzten Zeile liefert

$$\det D = \lambda^2 \begin{vmatrix} 0 & 0 & 1 \\ sin(\lambda) & 1 & 1 \\ \lambda cos(\lambda) & 1 & 0 \end{vmatrix} = \lambda^2 (sin(\lambda) - \lambda cos(\lambda)) \overset{!}{=} 0\ .$$

Da wir den Trivialfall $\lambda = 0$ ausschließen wollen, lautet die Lösbarkeitsbedingung

$$\begin{aligned} sin(\lambda) - \lambda cos(\lambda) = 0 &\Rightarrow tan(\lambda) = \lambda \\ &\Rightarrow \lambda = arctan(\lambda)\ . \end{aligned} \tag{3.39}$$

Eine Iteration der Form

$$\lambda_n^{(\nu+1)} = \arctan(\lambda_n^{(\nu)}) + n\pi$$

konvergiert sehr rasch und liefert die Zahlenwerte

$$\lambda_1 = 4.493409, \quad \lambda_2 = 7.725252, \quad \lambda_3 = 10.904122, \quad \text{usw.}\ . \tag{3.40}$$

Für jedes λ_n, welches die Eigenwertgleichung (3.39) erfüllt, ist das Eigenwertproblem des Knickstabs nichttrivial lösbar. Für die allgemeine Lösung (3.38) gilt dann

$$c_2 = c_4 = 0, \quad \frac{c_3}{c_1} = -\sin(\lambda_n) \; .$$

Für jedes n gibt es damit eine unnormierte (bis auf eine Konstante bestimmte) Lösung

$$y_n(x) = c(sin(\lambda_n x) - x\, sin(\lambda_n)) \; . \tag{3.41}$$

Man nennt die $y_n(x)$ auch die **Eigenfunktionen**.

3.5.2 Numerische Lösung mit dem Differenzenverfahren

1. Schritt: Wir unterteilen das Intervall $[0, 1]$ in $n = 6$ gleichgroße Teilintervalle mit der Bezeichnungsweise wie in Abbildung 3.6.

2. Schritt: Für jeden inneren Punkt x_i, $i = 1, \ldots, 5$ des Intervalls ersetzen wir die kontinuierlichen Ableitungen in der Differenzialgleichung durch Differenzenformeln. Dabei unterscheiden wir wie in §3.4 zwischen randnahen und randfernen Punkten. In den randnahen Punkten, x_1 und x_5, diskretisieren wir die DGL unter Berücksichtigung der Randwerte $y(0) = 0$, $y''(0) = 0$ bzw. $y(1) = 0$, $y'(1) = 0$. In den randfernen Punkten x_2, x_3, x_4 ersetzen wir die 4. Ableitung durch symmetrische Formeln. Für die 2. Ableitung können an allen inneren Punkten symmetrische Formeln gewählt werden:

$$\begin{aligned}
x_1 &: -\tfrac{12}{11}\tfrac{1}{h^4}(-2y_0 + 5y_1 - 4y_2 + y_3 - y_0'' h^2) + \lambda^2 \tfrac{1}{h^2}(y_0 - 2y_1 + y_2) = 0 \; , \\
x_2 &: \tfrac{1}{h^4}(y_0 - 4y_1 + 6y_2 - 4y_3 + y_4) + \lambda^2 \tfrac{1}{h^2}(y_1 - 2y_2 + y_3) = 0 \; , \\
x_3 &: \tfrac{1}{h^4}(y_1 - 4y_2 + 6y_3 - 4y_4 + y_5) + \lambda^2 \tfrac{1}{h^2}(y_2 - 2y_3 + y_4) = 0 \; , \\
x_4 &: \tfrac{1}{h^4}(y_2 - 4y_3 + 6y_4 - 4y_5 + y_6) + \lambda^2 \tfrac{1}{h^2}(y_3 - 2y_4 + y_5) = 0 \; , \\
x_5 &: -\tfrac{2}{3}\tfrac{1}{h^4}(2y_3 - 9y_4 + 18y_5 - 11y_6 + 6y_6' h) + \lambda^2 \tfrac{1}{h^2}(y_4 - 2y_5 + y_6) = 0 \; .
\end{aligned}$$

3. Schritt: Im zugehörigen MAPLE-**Worksheet** werden die Gleichungen mit Hilfe der Prozedur **FDFormeln** erstellt. Für $h = \frac{1}{6}$, $y_0 = y_6 = 0$, $y_0'' = 0$ und $y_6' = 0$ ergibt sich das LGS $A\,\mathbf{y} = 0$ mit der Matrix

$$\mathbf{A} = \begin{pmatrix}
\frac{77760}{11} - 72\lambda^2 & -\frac{62208}{11} + 36\lambda^2 & \frac{15552}{11} & 0 & 0 \\
-5184 + 36\lambda^2 & 7776 - 72\lambda^2 & -5184 + 36\lambda^2 & 1296 & 0 \\
1296 & -5734 + 36\lambda^2 & 7776 - 72\lambda^2 & -5784 + 36\lambda^2 & 1296 \\
0 & 1296 & -5734 + 36\lambda^2 & 7776 - 72\lambda^2 & -5784 + 36\lambda^2 \\
0 & 0 & 1728 & -7776 + 36\lambda^2 & 15552 - 72\lambda^2
\end{pmatrix}$$

Damit das Eigenwertproblem nicht nur die triviale Null-Lösung besitzt, muss die Determinante der Matrix **A** verschwinden. Wir bestimmen die Nullstellen der Determinante (Eigenwerte) mit MAPLE durch **fsolve (det(A) = 0,**

lambda); und erhalten für die Werte von λ in Abhängigkeit der Gitterpunkte:

	$n = 6$	$n = 10$	$n = 20$	EW
λ_1	4.44	4.47	4.49	4.49
λ_2	7.37	7.57	7.68	7.76
λ_3	9.86	10.47	10.78	10.90
λ_4	11.46	13.14	13.81	14.07
λ_5	15.23	15.51	16.75	17.22

Die Nullstellen, die für unterschiedliche Unterteilungen n des Intervalls berechnet wurden, werden mit den direkt bestimmten Eigenwerten des Problems verglichen. Je größer die Anzahl der Stützstellen ist, umso besser stimmen die numerischen Eigenwerte. Der erste Eigenwert ist schon bei der Unterteilung $n = 6$ bis auf 1% richtig. Je höher die Zahlenwerte für die Eigenwerte sind, desto mehr Stützstellen werden zur genauen Bestimmung benötigt.

4. Schritt: Grafische Darstellung. Für die exakten Eigenfunktionen des Eigenwertproblems gilt

$$y_n(x) = c(sin(\lambda_n x) - xsin(\lambda_n)).$$

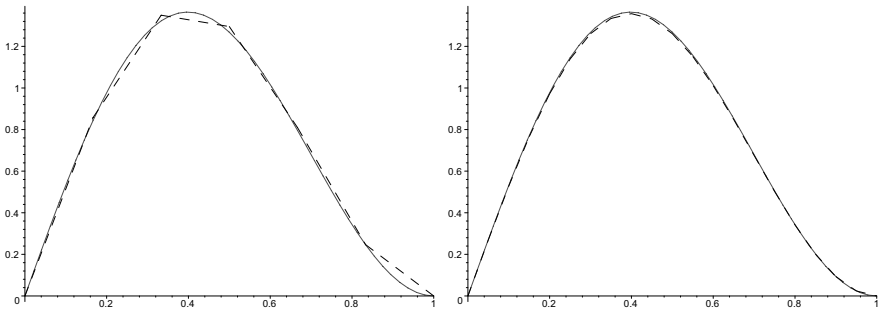
Diese Eigenfunktionen vergleichen wir mit der Lösung des LGS. Aufgrund der numerischen Ungenauigkeiten ist die Lösung des LGS $A\,\mathbf{y} = 0$ immer die Null-Lösung, da die Determinante einen zwar sehr kleinen aber doch von Null verschiedenen Wert besitzt. Daher setzen wir im zugehörigen MAPLE-**Worksheet** die letzte Zeile der Matrix auf Null und erhalten dann als Lösung

$$y_1 = 0.6352y_2,\ y_3 = 0.9594y_2,\ y_4 = 0.6000y_2,\ y_5 = 0.1827y_2\ .$$

Wir wählen in der exakten Lösung den freien Parameter $c = 1$ und passen in der numerischen Lösung y_2 so an, dass die Grafen vergleichbar werden. Es ergibt sich eine qualitativ gute Übereinstimmung zwischen exakter Eigenfunktion $y_1(x)$ und der numerisch berechneten: Um die Nichtlinearität im rechten Intervallbereich besser aufzulösen, würde man von der äquidistanten zu einer nicht äquidistanten Unterteilung des Intervalls übergehen und den rechten Bereich verfeinern. Dies kann man z.B. durch eine geeignete Wahl der Stützstellen ($x_i = L - \frac{L}{N^2}\,(N - i)^2, \quad i = 0 \ldots N$) erreichen. Die Rechnung und der Vergleich ist im MAPLE-**Worksheet** durchgeführt.

3.6 Methode der gewichteten Residuen

Wir betrachten nun eine Methode mit einer anderen Vorgehenswiese, die Methode der gewichteten Residuen. Es wird bei dieser Methode nicht das Intervall diskretisiert und Näherungswerte an Stützstellen oder Gitterpunkten



(a) Vergleich für n=6 (b) Vergleich für n=20

Abb. 3.9. Vergleich der exakten Eigenfunktion $y_1(x)$ mit den numerischen (gestrichelt)

gesucht, sondern es wird hier der Funktionenraum diskretisiert. Die Näherung nennt man **Ansatzfunktion**. Sie ist eine kontinuierliche Funktion, welche aus endlich vielen elementaren Funktionen als Linearkombination zusammengesetzt ist. Die Koeffizienten der Linearkombination werden dann so berechnet, dass die Ansatzfunktion eine möglichst gute Approximation der Lösung ist. Das Kriterium für diese Optimierung ist das gewichtete Residuum. Die Vorgehensweise werden wir am Beispiel eines Balkens auf elastischer Bettung

$$y^{(4)}(x) + y(x) = 1 \quad \text{in } [0,1]\,, \tag{3.42}$$
$$y(0) = y(1) = 0, \quad y''(0) = y''(1) = 0 \tag{3.43}$$

erläutern (mit MAPLE-**Worksheet**).

1. Schritt: Wahl der Ansatzfunktion. Als Ansatzfunktion sind alle Funktionen erlaubt, welche die Randbedingungen des Problems erfüllen, z.B:

$$sin(\pi x),\ sin(3\pi x),\ x - 2x^3 + x^4,$$
$$2.996e^x - 2.638e^{-x} - 1.571e^{2x} + e^{-2x} + 0.214e^{3x} \quad \text{usw.}\,.$$

Es ist im Allgemeinen nicht immer einfach, geeignete Funktionen zu finden, welche die Randbedingungen erfüllen. In einem separaten **Worksheet** werden solche Funktionen durch einen allgemeinen Ansatz der Form

$$f(x) = a\,f_1(x) + b\,f_2(x) + c\,f_3(x) + d\,f_4(x) + e\,f_5(x)$$

erzeugt, indem man beliebige Funktionen $f_i = f_i(x)$ vorgibt und die Parameter a, b, c, d, e an die Randwerte anpasst. Für obiges Problem der Balkenbiegung wählen wir als Ansatzfunktion

$$\tilde{y}(x) = c_1 sin(\pi x) + c_2 sin(3\pi x) \; . \tag{3.44}$$

Um diese Näherungslösung nicht mit der exakten Lösung $y = y(x)$ des Problems zu verwechseln, schreiben wir für sie $\tilde{y} = \tilde{y}(x)$.

2. Schritt: Berechnung des Residuums. Durch den vorgegebenen Lösungsansatz bekommt man nach Einsetzen in die Differenzialgleichung das Residuum

$$R = R(x) = \tilde{y}^{(4)}(x) + \tilde{y}(x) - 1 = \tag{3.45}$$

$$= c_1(\pi^4 + 1) sin(\pi x) + c_2(81\pi^4 + 1) sin(3\pi x) - 1 \; . \tag{3.46}$$

Nur die exakte Lösung der DGL würde $R(x) = 0$ für alle x ergeben. Wir können aber versuchen, die freien Parameter in der Ansatzfunktion so zu bestimmen, dass das Residuum möglichst klein wird. Man benötigt dabei zwei Bedingungen, um die beiden Koeffizienten c_1 und c_2 zu bestimmen.

3. Schritt: Minimum des Residuums. Es gibt verschiedene Kriterien, nach denen die Ansatzfunktion optimiert werden kann und das Residuum damit möglichst klein wird. Dies sind (a) die Methode der Kollokation, (b) die Integration über Teilintervalle, (c) die Methode der kleinsten Quadrate und (d) die sogenannte Galerkin-Methode. Die verschiedenen Methoden sind nicht immer einfach durchzuführen. Für unser einfaches Beispiel und die vorgeschlagene Ansatzfunktion lassen sich jedoch alle vier Gewichtungsmethoden explizit angeben.

1. **Kollokation:** Bei dieser Methode werden die Koeffizienten so bestimmt, dass das Residuum $R = R(x)$ an vorgegebenen Punkten verschwindet. Da wir zwei freie Koeffizienten in unserer Ansatzfunktion haben, benötigen wir zwei solche Stützstellen. Wegen der Symmetrie des Problems wählen wir $x_1 = \frac{1}{4}$ und $x_2 = \frac{1}{2}$ als Stützstellen zum Auswerten des Residuums $R(x)$:

$$x_1 = \frac{1}{4} \rightsquigarrow R(\frac{1}{4}) = \frac{\pi^4 + 1}{\sqrt{2}} c_1 + \frac{81\pi^4 + 1}{\sqrt{2}} c_2 - 1 = 0 \; ,$$

$$x_2 = \frac{1}{2} \rightsquigarrow R(\frac{1}{2}) = (\pi^4 + 1) c_1 + (81\pi^4 + 1) c_2 - 1 = 0 \; ,$$

$$\Rightarrow c_1 = \frac{1 + \sqrt{2}}{2(\pi^4 + 1)} \approx 0.012266, \; c_2 = \frac{(\pi^4 + 1) c_1 - 1}{81\pi^4 + 1} \approx 0.000026 \; .$$

Die Methode der Kollokation lässt sich immer durchführen. Man kann damit die Näherung so bestimmen, dass sie an möglichen „kritischen“ Punkten den Wert der exakten Lösung besitzt.

2. Teilintervall: Hier verlangt man, dass das Integral über das Residuum $\int R(x)dx$ verschwindet. Da wir bei unserem Problem zwei Koeffizienten haben, fordern wir, dass das Intervall auf zwei Teilbereichen verschwindet. Wegen der Symmetrie des Problems wählen wir $[0, \frac{1}{4}]$ und $[\frac{1}{4}, \frac{1}{2}]$ als Teilintervalle, über welche das Residuum integriert wird:

$$\int_0^{\frac{1}{4}} R(x)dx = \left[c_1(\pi^4+1)\frac{-cos(\pi x)}{\pi} + c_2(81\pi^4+1)\frac{-cos(3\pi x)}{3\pi}\right]_0^{\frac{1}{4}} - \frac{1}{4} \stackrel{!}{=} 0$$

$$\int_{\frac{1}{2}}^{\frac{1}{4}} R(x)dx = \left[c_1(\pi^4+1)\frac{-cos(\pi x)}{\pi} + c_2(81\pi^4+1)\frac{-cos(3\pi x)}{3\pi}\right]_{\frac{1}{4}}^{\frac{1}{2}} - \frac{1}{4} \stackrel{!}{=} 0$$

$$\Rightarrow c_1 \approx 0.013624\,, \quad c_2 \approx 0.000087\,.$$

Diese Methode ist bei den hier vorgegebenen leicht integrierbaren elementaren Funktionen sehr gut durchführbar.

3. **Kleinste Quadrate:** Die Methode der kleinsten Quadrate fordert, dass der Betrag des Residuums bzw. äquivalent das Quadrat des Residuums bezüglich der Koeffizienten c_i minimal wird: $\int R^2(x)dx = min$. Bei einer Funktion einer Variablen bedeutet Minimum oder Maximum, dass die Ableitung an diesen Punkten verschwindet. Hier bedeutet dies, dass die partiellen Ableitungen verschwinden: $\frac{\partial}{\partial c_i} \int R^2(x)dx = 0$ bzw. $\int R\frac{\partial R}{\partial c_i}dx = 0$ für $i = 1, 2$. Damit erhält man

$$\int_0^1 \frac{\partial R}{\partial c_1} R dx = \int_0^1 (\pi^4+1)sin(\pi x)[c_1(\pi^4+1)sin(\pi x) + c_2(81\pi^4+1)sin(3\pi x) - 1]dx \stackrel{!}{=} 0\,,$$

$$\int_0^1 \frac{\partial R}{\partial c_2} R dx = \int_0^1 (81\pi^4+1)sin(3\pi x)[c_1(\pi^4+1)sin(\pi x) + c_2(81\pi^4+1)sin(3\pi x) - 1]dx \stackrel{!}{=} 0\,,$$

$$\Rightarrow c_1 \approx 0.012938, \qquad c_2 \approx 0.000053\,.$$

4. Galerkin-Methode: Bei der Galerkin-Methode wird gefordert, dass das Residuum orthogonal zu der Ansatzfunktion steht: Das Skalarprodukt der

Ansatzfunktion mit dem Residuum muss verschwinden $\int f_i R dx = 0$. In diesem Fall stimmt die Galerkin-Methode bis auf die Faktoren $(\pi^4 + 1)$ bzw. $(81\pi^4 + 1)$ mit der Methode der kleinsten Quadrate überein und man erhält die gleichen Zahlenergebnisse für die Koeffizienten:

$$c_1 \approx 0.012938, \qquad c_2 \approx 0.000053 \; .$$

Die Orthogonalität des Residuums mit den Ansatzfunktionen bedeutet, dass keine Linearkombination der gewählten elementaren Funktionen gefunden werden kann, welche die Lösung besser approximiert, d.h. mit einem kleineren Residuum.

Diskussion: Bei allen vier Gewichtungsmethoden zeigt sich, dass die zweite Funktion, $sin(3\pi x)$, erheblich weniger ins Gewicht fällt und der Verlauf der "Lösung" im Wesentlichen durch $sin(\pi x)$ dominiert wird. Im zugehörigen **Worksheet** zeigt sich grafisch kein Unterschied zwischen den so ermittelten Funktionen und der exakten Lösung

$$y(x) = 1 + k_1 e^{-\frac{1}{2}\sqrt{2}x} sin(\frac{1}{2}\sqrt{2}x) + k_2 e^{\frac{1}{2}\sqrt{2}x} sin(\frac{1}{2}\sqrt{2}x) + k_2 e^{-\frac{1}{2}\sqrt{2}x} cos(\frac{1}{2}\sqrt{2}x) + k_3 e^{\frac{1}{2}\sqrt{2}x} cos(\frac{1}{2}\sqrt{2}x) \; ,$$

wobei die Koeffizienten k_1, k_2, k_3, k_4 an die Randbedingungen angepasst werden. Der Ausdruck für diese exakte Lösung unter Berücksichtigung der Randwerte erstreckt sich über 4 Formelzeilen und wird daher hier unterdrückt. In Abbildung 3.10 ist die gewichtete Ansatzfunktion gepunktet, die exakte Lösung durchgezogen gezeichnet.

Bemerkungen/Verallgemeinerungen:

1. Man kann die Bestimmung des Residuenminimums durch das Kollokationsverfahren dadurch verbessern, dass man für die beiden Koeffizienten c_1 und c_2 nicht zwei, sondern mehrere Stützstellen auswählt und die Methode der kleinsten Quadrate nimmt, um die c_1 und c_2 an die Stützstellen anzupassen. In einem eigenen **Worksheet** wird die Rechnung mit 8 Stützstellen ($\frac{1}{16}$, $\frac{2}{16}$, ..., $\frac{8}{16}$) durchgeführt. Die Koeffizienten ergeben sich dann zu

   $$c_1 = 0.01304187075, \quad c_2 = 0.00005040812997 \; .$$

   In den Abbildungen 3.11(a) und 3.11(b) sind die Ergebnisse für 2 und 8 Stützstellen gegenüber gestellt.

2. Wählt man statt $\tilde{y}(x) = c_1 sin(\pi x) + c_2 sin(3\pi x)$ einen Ansatz mit Exponentialfunktionen oder Polynomfunktionen bzw. Kombinationen davon, zeigt sich ebenfalls eine überraschend gute Übereinstimmung mit der exakten Lösung, wie in dem zusätzlichen **Worksheet** zu erkennen ist.

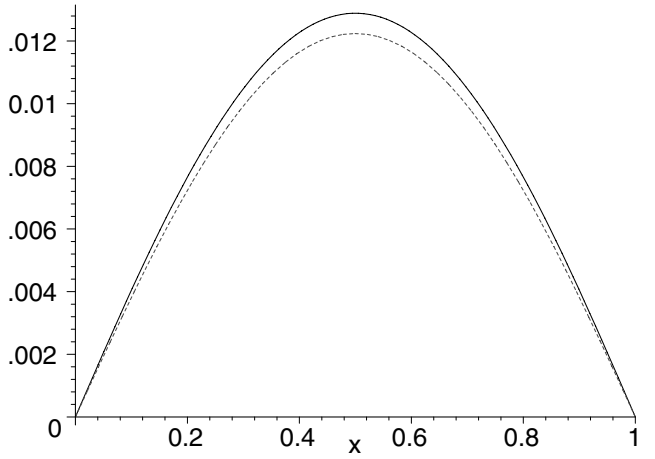
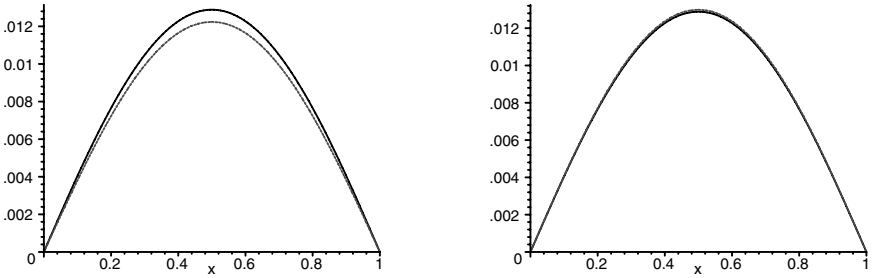


Abb. 3.10. Vergleich der gewichteten Ansatzfunktion (gestrichelt) mit der analytischen Lösung

3. Bei **inhomogenen** Randwerten wählt man den Ansatz

$$\tilde{y}(x) = \tilde{y}_0(x) + \sum_{i=1}^{n} c_i \, \tilde{y}_i(x),$$

wobei $\tilde{y}_0(x)$ die *inhomogenen* und $\tilde{y}_i(x)$ $(i = 1, ..., n)$ die *homogenen* Randbedingungen erfüllen. Dann besitzt die Ansatzfunktion $\tilde{y}(x)$ die geforderten Randwerte.



(a) Kollokation bei 2 Stützstellen (b) Kollokation bei 8 Stützstellen

Abb. 3.11. Vergleich der exakten mit den numerischen Lösungen (gestrichelt)

Wir wollen die Methode der gewichteten Residuen etwas allgemeiner zusammenfassen. Man wählt zunächst als Ansatzfunktion eine Linearkombination

$$\tilde{y}(x) = \sum_{i=1}^{n} c_i \varphi_i(x) \tag{3.47}$$

mit elementaren Funktionen $\varphi_i = \varphi_i(x)$. Dabei müssen die Funktionen φ_i die homogenen Randbedingungen erfüllen. Die Bestimmung der Freiheitsgrade c_i kann dann im zweiten Schritt nach mehreren Methoden erfolgen: Kollokation, Integration über Teilintervalle, Methode der kleinsten Quadrate oder durch die Galerkin-Bedingung. Das Einsetzen der Ansatzfunktion $\tilde{y}(x)$ in das entsprechende Kriterium liefert ein Gleichungssystem für die Freiheitsgrade c_i.

Bemerkung: Die Funktionen $\varphi_i(x)$ werden als Basisfunktionen bezeichnet, da die Menge aller Überlagerungen (Linearkombinationen)

$$\{\varphi(x) = \sum_{i=1}^{n} c_i \varphi_i(x), \quad c_i \in \mathbb{R}\}$$

einen Vektorraum aufspannt. Sind die Funktionen $\varphi_i(x)$ linear unabhängig gewählt, so stellen sie eine Basis dieses n-dimensionalen Vektorraums dar. Entscheidend für die lineare Unabhängigkeit ist, dass die Wronski-Determinante

$$W(x) = det \begin{pmatrix} \varphi_1(x) & \varphi_2(x) & \dots & \varphi_n(x) \\ \varphi_1'(x) & \varphi_2'(x) & \dots & \varphi_n'(x) \\ \vdots & \vdots & & \vdots \\ \varphi_1^{(n-1)}(x) & \varphi_2^{(n-1)}(x) & \dots & \varphi_n^{(n-1)}(x) \end{pmatrix}.$$

ungleich Null ist.

Es wird nach den verschiedenen Kriterien innerhalb des Vektorraums der ausgewählten Basisfunktionen eine optimale Näherungslösung berechnet. Das Wesentliche und auch Schwierige an der Methode der gewichteten Residuen ist das Auffinden von geeigneten Ansatzfunktionen. Es hat sich hier in den letzten Jahren ein mächtiges Konzept durchgesetzt, welches man die **Finite-Elemente-Methode (FE-Methode)** nennt. Hier werden als Basisfunktionen „lokal definierte“ Funktionen eingesetzt, welche nur auf einem kleinen Teilintervall von Null verschieden sind. Die Näherungslösung wird dann als Linearkombination dieser lokalen Funktionen gewählt. Das lokale Verhalten einer Lösung lässt sich hier sehr gut wiedergeben, ohne dass andere Bereiche sehr stark davon beeinflusst werden. Dies macht bei Basisfunktionen, welche im ganzen Rechengebiet ungleich Null sind, oft Schwierigkeiten. Bevor wir zu den finiten Elementen kommen, wird in den nächsten Abschnitten noch eine weitere wichtige Methode beschrieben, um die Koeffizienten der Ansatzfunktion zu bestimmen, das sogenannte Ritzsche Verfahren. Dazu müssen wir uns zunächst mit dem Variationsproblem beschäftigen.

3.7 Variationsproblem

Viele Probleme der Ingenieur- und Naturwissenschaften ergeben sich nicht direkt als Differenzialgleichung, sondern als Variationsproblem. Bei einem Variationsproblem sucht man eine Lösung, welche sich als Maximum, Minimum oder Sattelpunkt ergibt. So wird in der Physik oft der Zustand gesucht, bei dem die potentielle Energie am kleinsten ist. Auch in der Steuerungs- und Regelungstheorie findet die Variationsrechnung Anwendung, wenn es um die Bestimmung des optimalen Reglers geht. Auch das Problem der minimalen Oberflächen, die z.B. bei Seifenblasen auftreten, ist ein Variationsproblem. Die Variationsrechnung wurde um 1800 von Lagrange entwickelt.

Als Beispiel soll die Modellierung der Balkenbiegung mit konstanter Streckl ast auf elastischer Bettung dienen:

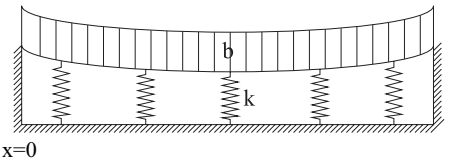


Abb. 3.12. Balken mit konstanter Streckl ast b auf elastischer Bettung (Federkonstante k), Bettungskraft $-k \cdot dx$ (nach oben gerichtet)

Die direkte Beschreibung des Problems erhält man über die Momentengleichung

$$dM(x) = (k\, y(x) - b)dx \cdot x \rightsquigarrow -\frac{d^2M(x)}{dx^2} = b - k\ y(x)\ .$$

Wegen $M(x) = -EI\ y''(x)$ folgt

$$-\frac{d^2M(x)}{dx^2} = EI\ y^{(4)}(x)\ .$$

Somit erhalten wir das Randwertproblem

$$y^{(4)}(x) = \frac{b}{EI} - \frac{k}{EI}y(x)\ , \tag{3.48}$$

$$\text{mit} \quad y(0) = y(1) = 0, \quad y''(0) = y''(1) = 0\ . \tag{3.49}$$

Generell lassen sich Gleichgewichtsprobleme direkt über die Momentengleichung bzw. Kräftebilanz bei Massenpunkten angeben. Alternativ dazu wird

das Problem auch über die potentielle Energie beschrieben: Das Extremalprinzip besagt, dass die potentielle Energie im Gleichgewichtszustand zum Minimum wird. Diesen indirekten Weg zeigen wir am Beispiel obiger Balkenbiegung auf:

Die potentielle Energie des Balkens setzt sich zusammen aus

$$E_{pot} = E_{Biegung} + E_{Bettung} + E_{Last} \ ,$$

wobei der erste Term der Summe die gespeicherte elastische Biegeenergie, der zweite Term die Energie der Bettung durch die Druckfedern und der dritte die potentielle Energie der Lasten darstellt. Unser Problem lautet nun

$$E(y) = \int_0^1 (\frac{EI}{2} y''^2(x) - \frac{k}{2} y^2(x) - by(x)) dx \stackrel{!}{=} \text{extremal} \tag{3.50}$$

$$\text{mit} \quad y(0) = y(1) = 0. \tag{3.51}$$

Dieses Problem wird nicht mehr Randwertproblem (RWP) genannt, sondern Variationsaufgabe (VA). Variationsproblem deshalb, weil unter allen möglichen Verbiegungen $y = y(x)$ des Balkens nur die physikalisch realisiert wird, deren Energie $E(y)$ verglichen mit allen anderen Biegungen minimal ist.

Der Vorteil bei der Formulierung über die Variationsaufgabe liegt zunächst darin, dass in der Variationsaufgabe nur Ableitungen von 2. Ordnung auftreten. Man benötigt nur zwei Randbedingungen; es genügen dabei die beiden, welche direkt aus der Geometrie folgen. Die Variationsaufgabe ist gerade bei mechanischen Anwendungen manchmal bedeutend einfacher zu formulieren als das direkte Problem.

Bemerkungen:

1. Bei dynamischen Problemen verwendet man oft das Hamiltonsche Prinzip:
$$\int_{t_1}^{t_2} (E_{kin} - E_{pot}) dt = \text{minimal}.$$
Dies führt direkt auf ein Variationsproblem.

2. Der Integralausdruck ist in beiden Fällen ein *Funktional*, d.h. eine Abbildung in die reellen Zahlen, deren Argumente nicht variable Zahlenwerte, sondern variable Funktionen sind.

3. Das Randwertproblem des Beispiels benötigt entsprechend der Ordnung der Differenzialgleichung 4 Randbedingungen, zwei davon sind nicht „wesentlich“, sondern zusätzlich. Die Variationsaufgabe dagegen braucht nur 2 Randbedingungen, beide sind „wesentlich“ durch die Konstruktion bestimmt.

4. Allgemein und exakt gilt folgendes: Die Randbedingungen einer Variationsaufgabe sind in aller Regel einfacher als die des äquivalenten Randwertproblems, und zwar gilt:
Sei die Ordnung des RWP $= 2m$, dann sind die Randbedingungen, die sich in $y, y', \ldots, y^{(m-1)}$ ausdrücken lassen, „wesentlich“, und nur diese treten in der Variationsaufgabe auf. Alle anderen Randbedingungen, $y^{(m)}, \ldots, y^{(2m-1)}$, sind „zusätzlich“ und werden bei der Variationsaufgabe nicht berücksichtigt. Physikalisch ist es meist so, dass wesentliche Randbedingungen sich auf die Geometrie beziehen und zusätzliche durch Berücksichtigung von Kräften oder Momenten entstehen.

3.7.1 Variationsproblem

Nach diesen physikalisch motivierten Betrachtungen nun zur mathematischen Sicht. Das Teilgebiet der Mathematik, mit dem wir uns beschäftigen, ist die Variationsrechnung - wir beschränken uns hier im Weiteren auf die Angabe von Ergebnissen. So gibt es den mathematischen **Satz:** Zu einem Variationsproblem existiert immer ein zugehöriges Rand- oder Eigenwertproblem. Die Herleitung des RWP/EWP aus dem Variationsproblem ist relativ leicht durchzuführen, was wir an einem einfachen Modellproblem demonstrieren werden. Gegeben sei die Variationsaufgabe

$$I(y) = \int_{x_1}^{x_2} f(x, y, y')\, dx \overset{!}{=} \text{minimal} .$$

Wir suchen also aus einer Klasse von Konkurrenzfunktionen $y = y(x)$ diejenige (y_0), die diese Extremalbedingung erfüllt. Zur Konkurrenz zugelassen sind alle zweimal stetig differenzierbaren beschränkten Funktionen, welche die geometrisch wesentlichen Randbedingungen erfüllen: $y(x_1) = y_1$ und $y(x_2) = y_2$ (siehe Abbildung 3.13). Alle Konkurrenzfunktionen y können wir formal unter Verwendung der Lösung y_0 als

$$y = y_0 + \varepsilon \cdot \eta$$

schreiben mit der beschränkten und die Randbedingungen $\eta(x_1) = \eta(x_2) = 0$ erfüllenden Funktion η. Da y_0 das Minimum ist, gilt

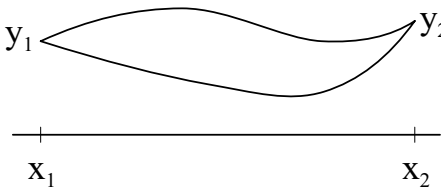


Abb. 3.13. Konkurrenzfunktionen

$$I(y_0, 0) \leq I(y_0, \varepsilon) = \int_{x_1}^{x_2} f(x, y_0 + \varepsilon \cdot \eta, y_0' + \varepsilon \cdot \eta') dx \ .$$

Wenn die Funktion y_0 das Integral $I(y)$ zum Minimum macht, dann hat das Integral $I(y_0, \varepsilon)$ als Funktion von ε betrachtet bei $\varepsilon = 0$ ein relatives Minimum, also $I(y_0, 0) \leq I(y_0, \varepsilon)$. Eine notwendige Bedingung für $I(y_0, 0) \leq I(y_0, \varepsilon)$ ist, dass

$$\delta I := \frac{dI}{d\varepsilon}(\varepsilon = 0) = \int_{x_1}^{x_2} [\frac{\partial f}{\partial y}(x, y, y')\eta + \frac{\partial f}{\partial y'}(x, y, y')\eta'] dx = 0$$

(“erste Variation des Grundintegrals“).

Eine partielle Integration liefert

$$\delta I = \int_{x_1}^{x_2} (f_y - \frac{d}{dx} f_{y'})\eta dx = 0 \ .$$

Da $\eta \neq 0$ ist die Bedingung sicher erfüllt, wenn der Integrand verschwindet, also

$$\frac{d}{dx}\frac{\partial f}{\partial y'} - \frac{\partial f}{\partial y} = 0.$$

Man bezeichnet diese Gleichung auch als die zur Variationsaufgabe gehörende Eulerschen Differenzialgleichung.

3.7.2 Eulersche Differenzialgleichungen

In Verallgemeinerung der oben gezeigten Vorgehensweise gilt im Falle einer allgemeinen Extremalbedingung der folgende Satz:

Satz: Gegeben sei die Extremalbedingung

$$E(y) = \int f(x, y, y', ..., y^{(n)}) dx \overset{!}{=} \text{extremal} . \tag{3.52}$$

Dann hat die zugehörige **Eulersche Differenzialgleichung** die Form

$$\frac{\partial f}{\partial y} - \frac{d}{dx}\frac{\partial f}{\partial y'} + \frac{d^2}{dx^2}\frac{\partial f}{\partial y''} - ... + (-1)^n \frac{d^n}{dx^n}\frac{\partial f}{\partial y^{(n)}} = 0. \tag{3.53}$$

Ist zu einem Problem die Variationsaufgabe bekannt, dann lässt sich über die Eulersche Differenzialgleichung auf einfache Weise die zugehörige Differenzialgleichung angeben.

Für das inverse Problem der Variationsrechnung, also dem Übergang von einem Randwertproblem zu einer Variationsaufgabe, existiert nicht immer eine Lösung und falls sie existiert, ist es im Allgemeinen schwierig, sie herzuleiten. Eine Ausnahme bilden hier lineare Differenzialgleichungen, bei denen ein Variationsproblem genau dann existiert, wenn der Differenzialoperator selbstadjungiert und positiv definit ist. Das Variationsproblem lässt sich dann direkt angeben. Gesichert ist auch die Existenz des zugehörigen Variationsproblems für Differenzialgleichungen zweiter Ordnung (auch nichtlineare).

3.7.3 Das Ritzsche Verfahren

Die Idee beim Ritzschen Verfahren besteht darin, die Variationsaufgabe mit der Methode der Ansatzfunktionen zu lösen. Man wählt als Ansatzfunktion wieder eine Linearkombination

$$\tilde{y}(x) = \sum_{i=1}^{n} c_i \varphi_i(x) ,$$

wobei die Basisfunktionen $\varphi_i = \varphi_i(x)$ nur die wesentlichen Randbedingungen erfüllen müssen. Die Ansatzfunktion $y(x)$ wird in das Funktional eingesetzt. Die Minimierung erfolgt durch Nullsetzen der partiellen Ableitungen nach den c_i.

Die Näherung bei dem Ritzschen Verfahren besteht darin, dass das Minimum des Funktionals nicht unter den unendlich vielen erlaubten Konkurrenzfunktionen gesucht wird, sondern nur innerhalb des endlich dimensionalen Unterraums der ausgewählten Basisfunktionen.

Beispiel 3.5 (Knickstab, mit MAPLE- Worksheet). Wir wenden das Ritzsche Verfahren auf das Beispiel der Knickstabs an, das wir schon im Abschnitt 3.3 mit Hilfe der Differenzenmethode näherungsweise gelöst haben (siehe Abbildung 3.8).

Für den einseitig geführten, anderseitig eingespannten Knickstab gilt das normierte Randwertproblem

$$y^{(4)}(x) + \lambda^2 y''(x) = 0 \quad \text{in } [0,1]\,, \tag{3.54}$$

$$y(0) = y(1) = 0,\ y''(0) = 0,\ y'(1) = 0\,. \tag{3.55}$$

In der technischen Mechanik wird gezeigt, dass das Energiefunktional zu diesem Problem

$$E(y) = \int_0^1 (y''^2(x) - \lambda^2 y'^2(x))\,dx \tag{3.56}$$

lautet. Man überprüfe diese Aussage, indem man zur Variationsaufgabe die Eulersche Differenzialgleichung bestimme! Die wesentlichen Randbedingungen der Variationsaufgabe lauten

$$y(0) = y(1) = 0,\ y'(1) = 0\,.$$

Das Verfahren von Ritz verlangt, dass die benutzte Ansatzfunktion die kinematischen Randbedingungen erfüllt, im betrachteten Beispiel ist das die Lagerung des Knickstabs. Wir benötigen bei der Formulierung über die Variationsaufgabe nur diese 3 Randbedingungen, da in der Variationsaufgabe maximal die zweite Ableitung vorkommt und daher nur die Randwerte für y und y' wesentlich sind.

Als Basisfunktionen wählen wir

$$\varphi_1(x) = x(1-x)^2, \quad \varphi_2(x) = x(1-x)^3\,,$$

welche beide die Randbedingungen erfüllen, und definieren als Ansatzfunktion

$$\begin{aligned}\tilde{y}(x) &= c_1\varphi_1(x) + c_2\varphi_2(x)\\ &= c_1 x(1-x)^2 + c_2 x(1-x)^3\,.\end{aligned}$$

Setzen wir die Ableitungen

$$\begin{aligned}\tilde{y}'(x) &= c_1(1-x)(1-3x) + c_2(1-x)^2(1-4x)\,,\\ \tilde{y}''(x) &= 2[c_1(3x-2) + 3c_2(1-x)(2x-1)]\end{aligned}$$

in das Energiefunktional $E = E(y)$ ein, so müssen die partiellen Ableitungen nach c_1 und c_2 als notwendige Bedingung für das Minimum verschwinden. Damit interpretieren wir $E(y)$ als Funktion $F(c_1, c_2)$ der beiden Variablen c_1 und c_2. Es gilt

$$\frac{\partial F}{\partial c_1} = 2\int_0^1 (\frac{\partial \tilde{y}''}{\partial c_1}\tilde{y}'' - \lambda^2 \frac{\partial \tilde{y}'}{\partial c_1}\tilde{y}'')dx = (4 - \frac{2}{15}\lambda^2)c_1 + (4 - \frac{1}{10}\lambda^2)c_2 \stackrel{!}{=} 0 \,,$$

$$\frac{\partial F}{\partial c_2} = 2\int_0^1 (\frac{\partial \tilde{y}''}{\partial c_2}\tilde{y}'' - \lambda^2 \frac{\partial \tilde{y}'}{\partial c_2}\tilde{y}'')dx = (4 - \frac{1}{10}\lambda^2)c_1 + (\frac{24}{5} - \frac{3}{35}\lambda^2)c_2 \stackrel{!}{=} 0 \,.$$

Dies ist ein lineares, homogenes Gleichungssystem für die Koeffizienten c_1 und c_2. Damit wir nicht nur die triviale Null-Lösung $c_1 = c_2 = 0$ und damit $y \equiv 0$ erhalten, muss die Koeffizientendeterminante Null ergeben:

$$det \begin{pmatrix} 4 - \frac{2}{15}\lambda^2 & 4 - \frac{1}{10}\lambda^2 \\ 4 - \frac{1}{10}\lambda^2 & \frac{24}{5} - \frac{3}{35}\lambda^2 \end{pmatrix} = 0 \,.$$

Dies liefert die beiden Eigenwerte

$$\lambda_1 = 4.5737, \quad \lambda_2 = 10.3480$$

mit einem Fehler von $\varepsilon_1 = +1.8\%$ bzw. $\varepsilon_2 = +34\%$ im Vergleich mit den beiden richtigen Eigenwerten.

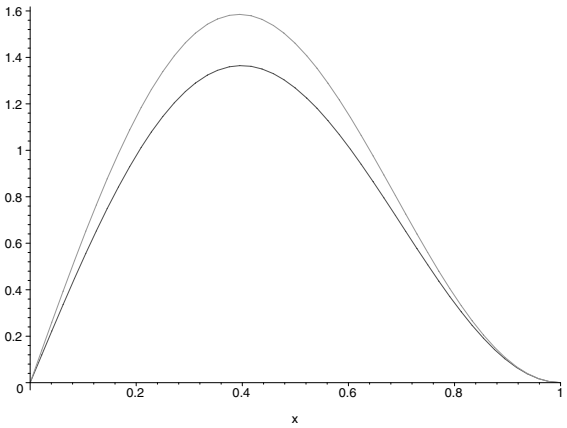


Abb. 3.14. Vergleich mit der numerischen (gestrichelt) mit der exakten (durchgezogen) Lösung

Die genaue Rechnung und Details sowie der grafische Vergleich mit der exakten Lösung sind im zugehörigen **Worksheet** angegeben. Dort wird die Rechnung auch alternativ mit den Basisfunktionen $\varphi_1(x) = -3sin(\pi x) + sin(3\pi x)$ und $\varphi_2(x) = sin(\frac{\pi}{2}x) + sin(\frac{3}{2}\pi x)$ durchgeführt. Die Ergebnisse sind vergleichbar. Als Eigenwerte erhält man dann

1. Eigenwert $\lambda_1 = 4.4976$ mit einem Fehler $\varepsilon_1 = +0.1\%$,
2. Eigenwert $\lambda_2 = 8.7945$ mit einem Fehler $\varepsilon_2 = +13.8\%$. □

Bemerkungen:

1. Das Ritzsche Verfahren liefert grundsätzlich zu große Eigenwerte, so dass von mehreren unterschiedlichen Rechenergebnissen stets das Kleinste das Beste ist.

2. Es lässt sich zeigen, dass die Methode von Galerkin in den Fällen, in denen eine zum gegebenen Randwertproblem äquivalente Variationsaufgabe existiert, dieselben Ergebnisse liefert wie das Ritzsche Verfahren.

3. Da es sich beim obigen Beispiel um ein Eigenwertproblem handelt, gibt es zu jedem Eigenwert λ ähnliche Lösungen; die Lösungen sind nur bis auf eine Konstante bestimmt. Im Worksheet wird der freie Parameter der numerischen Lösung an die exakte Lösung $sin(\lambda x) - x\, sin(\lambda)$ angepasst.

4. Das Problem beim Ritzschen Verfahren ist wie bei der Methode der gewichteten Residuen geeignete, globale Ansatzfunktionen zu finden, welche die Randbedingungen erfüllen. Auch ist die Auswertung der Integrale meist sehr aufwändig. Einen Ausweg bildet hier die Finite-Elemente-Methode, welche allgemein anwendbare Basisfunktionen liefert.

3.8 Die Finite-Elemente-Methode

Die Methode der finiten Elemente wurde in den letzten 50 Jahren entwickelt und ist zu einem der wichtigsten Näherungsverfahren vor allem für Randwertprobleme für partielle Differenzialgleichungen geworden. Pionierarbeiten leisteten auf diesem Gebiet neben R. Courant, J. H. Argyris und O. C. Zienkiewiecz. Eine kurze Geschichte der Finite-Element-Methode wird in dem Buch von Jung und Langer [35] gegeben.

Wir haben bei der Diskussion der Methode der gewichteten Residuen und bei dem Ritzschen Verfahren bislang globale Ansatzfunktionen benutzt. Bei

der **Finite-Elemente-Methode (FE-Methode)** definiert man die Basisfunktionen nicht mehr global auf dem gesamten Intervall $[a, b]$, sondern nur lokal auf Teilintervallen, genauer ausgedrückt: Sie haben nur Werte ungleich Null auf einzelnen Teilintervallen. Damit umgeht man das Problem, geeignete Funktionen zu finden, welche die Randbedingungen erfüllen, im mehrdimensionalen Fall ist dies oft unmöglich! Die Gleichungssysteme sind bei den lokalen Basisfunktionen schwach besetzt und können auch für viele Gitterintervalle noch sehr effizient gelöst werden. Theoretische Aussagen über die Genauigkeit werden möglich.

Der Grundgedanke der Finiten-Elemente-Methode besteht darin, das Intervall $[a, b]$ in n Teilintervalle zu zerlegen („finite Elemente“) und als Basisfunktionen einfache, lokal definierte Polynome zu wählen. Wir werden die Vorgehensweise für die folgenden, allgemeinen Differenzialgleichung 2. Ordnung vorführen. Dazu betrachten wir das Randwertproblem

$$-(p\,y'(x))' + q\,y(x) = f(x) \quad \text{in } [a, b]\,, \tag{3.57}$$

$$\text{mit} \quad y(a) = y(b) = 0 \tag{3.58}$$

$$\text{und} \quad p > 0, q \geq 0\,. \tag{3.59}$$

Bemerkungen:

1. Wir können ohne Beschränkung der Allgemeinheit annehmen, dass die Randbedingungen $y(a) = y(b) = 0$ sind. Denn ist dies nicht der Fall, transformiert man das Randwertproblem mit von Null verschiedenen Randwerten

   $$-(p\,u')' + q\,u = g; \quad u(a) = u_a, \quad u(b) = u_b$$

   durch

   $$y(x) := u(x) - \frac{b-x}{b-a}u_a - \frac{a-x}{a-b}u_b$$

   auf ein Problem mit verschwindenden Randwerten

   $$-(p\,y')' + q\,y = f; \quad y(a) = 0, \quad y(b) = 0$$

   mit der modifizierten Inhomogenität

   $$f = g - \frac{u_a}{b-a}p' - \frac{u_b}{a-b}p' - q\frac{b-x}{b-a}u_a - q\frac{a-x}{a-b}u_b\,.$$

2. In der folgenden Diskussion nehmen wir der Einfachheit halber an, dass p und q konstante Funktionen sind. Die Vorgehensweise gilt aber auch allgemein, wenn p und q Funktionen in x sind.

Wie man über die Eulerschen Differenzialgleichung nachprüfen kann, ist

$$E(u) = \int_a^b (p\, u'^2(x) + q\, u^2(x))\, dx - 2 \int_a^b u(x)\, f(x)\, dx \qquad (3.60)$$

das zum Randwertproblem gehörende Energiefunktional. Es ist damit mathematisch äquivalent das Randwertproblem zu lösen oder unter allen zulässigen Funktionen $u(x)$ (zweimal stetig differenzierbar mit $u(a) = u(b) = 0$) die Funktion zu finden, welche das Energiefunktional minimiert. Auch das Variationsproblem ist in der Regel **nicht** exakt lösbar. Daher muss man sich wie beim Ritzschen Verfahren auf eine kleinere Funktionenklasse beschränken und sucht darin das Minimum.

Das Finite-Elemente-Verfahren läuft dann in den folgenden Schritten ab.

1. Schritt: Unterteilung des Intervalls. Wir unterteilen das Intervall in n Teilintervalle (für die grafische Darstellung wählen wir $n = 7$).

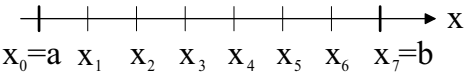


Abb. 3.15. Intervallunterteilung

2. Schritt: Wahl der Basisfunktionen. Für jeden inneren Punkt des Intervalls $x_i,\ i = 1, ..., n-1$, definieren wir eine Basisfunktion $u_i(x)$ mit der Eigenschaft

$$u_i(x_j) = \delta_{ij} = \begin{cases} 0 & \text{für } i \neq j \\ 1 & \text{für } i = j \end{cases}.$$

Zwischen den Punkten soll die Funktion linear verlaufen. Man beachte, dass die Beschränkung auf stückweise lineare Funktionen hier der Einfachheit halber vorgenommen wird. Es können auch stückweise Polynome höherer Ordnung gewählt werden. Aufgrund des Schaubilds nennen wir obige Funktionen auch Dreiecksfunktionen oder Hutfunktionen.

3. Schritt: Berechnung des Energiefunktionals. Die Ansatzfunktion $\tilde{u} = \tilde{u}(x)$ wird als Überlagerung (Linearkombination) dieser Dreiecksfunktionen definiert. Man hat hier dann doch wieder wie im Falle eines Differenzenverfahrens eine Art Diskretisierung des Intervalls eingeführt. Bei der

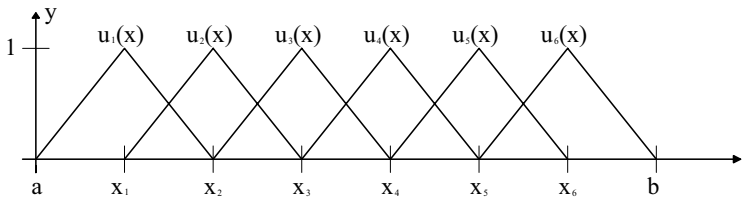


Abb. 3.16. Dreiecksfunktionen

FE-Methode ist dies aber nur der Ausgangspunkt für die Definition der Basisfunktionen. Die Ansatzfunktion und nach Bestimmung der Freiheitsgrade die Näherungsfunktion ist eine kontinuierliche Funktion. Dieser wesentliche Unterschied bleibt. Man bezeichnet in Anlehnung an die Differenzenverfahren die Ansatzfunktion trotzdem gerne mit $u_h = u_h(x)$ statt mit $\tilde{u}$, wobei die „Schrittweite“ h gewissermaßen das Symbol für die Diskretisierung oder allgemein die Approximation ist. Wir werden diese Bezeichnung im Folgenden so übernehmen und erhalten die Ansatzfunktion in der Form

$$u_h(x) = \sum_{i=1}^{n-1} c_i\, u_i(x)\ . \tag{3.61}$$

Man bezeichnet die Dreiecksfunktionen $u_i(x)$ als Basisfunktionen, da sie eine Basis des $(n-1)$-dimensionalen Vektorraums

$$S = \{u_h(x) = \sum_{i=1}^{n-1} c_i\, u_i(x) \quad \text{mit } c_i \in \mathbb{R}\}$$

bilden. Gelegentlich werden sie auch als Elementfunktionen bzw. Shapefunktionen bezeichnet, da sie auf den finiten Elementen definiert sind. Für eine gegebene Unterteilung des Intervalls ist S die Menge der stetigen, stückweise linearen Funktionen, die in den Punkten $x_i (i = 1, ..., n-1)$ vorgegeben sind. Die Funktionswerte an den Punkten x_i sind in diesem Falle die Koeffizienten c_i, d.h.

$$u_h(x_i) = c_i \quad \text{für alle} \quad i = 1, ..., n-1.$$

Man nennt die Koeffizienten c_i auch die Freiheitsgrade der Ansatzfunktion. Nach dieser Diskretisierung des Funktionenraumes suchen wir das Minimum des Energiefunktionals $E = E(u)$ nicht mehr unter allen zulässigen Funktionen, sondern nur unter den Funktionen aus der Menge S, also unter den stückweisen linearen Funktionen. Es sind somit Koeffizienten c_i gesucht, so dass für die Funktion $u_h(x) = \sum_{i=1}^{n-1} c_i\, u_i(x)$ das Energiefunktional $E(u)$ minimal wird.

Um eine notwendige Bedingung für das Minimum zu erhalten, nehmen wir

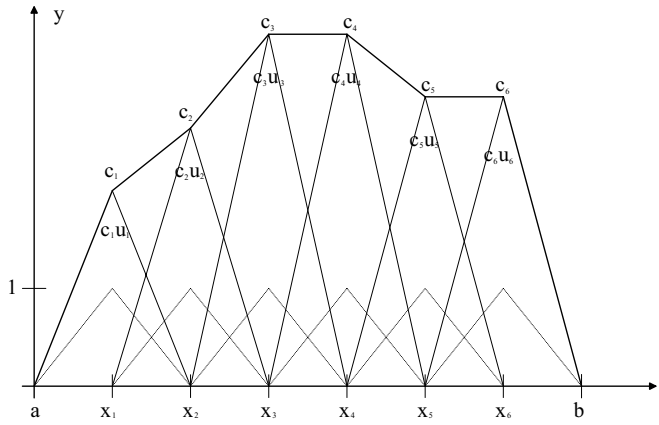


Abb. 3.17. Linearkombination der Dreiecksfunktionen

eine beliebige Funktion aus S und bestimmen dazu die Energie $E(u)$. Zur übersichtlicheren Berechnung führen wir die folgenden Abkürzungen ein:

$$[u, v] = \int_a^b (p\, u'(x)\, v'(x) + q\, u(x)\, v(x))dx \;, \tag{3.62}$$

$$< u, v > = \int_a^b u(x)\, v(x)dx \;. \tag{3.63}$$

Sowohl $<,>$ als auch $[\,,]$ sind Skalarprodukte, d.h. es gelten die folgenden Eigenschaften:

(S1) Linearität im ersten und zweiten Argument:

$$< \alpha_1 u_1 + \alpha_2 u_2, v > = \alpha_1 < u_1, v > + \alpha_2 < u_2, v > \;,$$
$$< u, \alpha_1 v_1 + \alpha_2 v_2 > = \alpha_1 < u, v_1 > + \alpha_2 < u, v_2 > \;,$$

(S2) Symmetrie:

$$< u, v > = < v, u > \;,$$

(S3) Positiv definit:

$$< u, u > \geq 0 \quad \text{für alle } u \;,$$
$$< u, u > = 0 \;\Leftrightarrow\; u = 0 \;.$$

Beweis der Eigenschaften $\Rightarrow$ Übungsaufgabe.

Mit den eingeführten Abkürzungen und den Eigenschaften von $[\,,\,]$ bzw. $<,>$ setzt man die Ansatzfunktion in das Energiefunktional ein:

$$E(u_h) = [u_h, u_h] - 2 < u_h, f > =: F(c_1, c_2, ..., c_{n-1}) \; . \tag{3.64}$$

Es ergibt sich für die Funktion $F = F(u_h)$ in den Variablen $c_1, ..., c_{n-1}$

$$\begin{aligned} F(u_h) &= [\sum_{i=1}^{n-1} c_i u_i, \sum_{j=1}^{n-1} c_j u_j] - 2 < \sum_{k=1}^{n-1} c_k u_k, f > \\ &\overset{S1}{=} \sum_{i=1}^{n-1} c_i [u_i, \sum_{j=1}^{n-1} c_j u_j] - 2 \sum_{k=1}^{n-1} c_k < u_k, f > \\ &\overset{S2}{=} \sum_{i=1}^{n-1} \sum_{j=1}^{n-1} c_i c_j [u_i, u_j] - 2 \sum_{k=1}^{n-1} c_k < u_k, f > \; . \end{aligned}$$

Man beachte, dass $[u_i, u_j]$ bzw. $< u_k, f >$ bestimmte Integrale über bekannte Funktionen sind und damit reine Zahlenwerte darstellen.

4. Schritt: Minimierung des Energiefunktionals. Von der Funktion F muss nun das Minimum gefunden werden. Eine notwendige Bedingung für das Minimum ist, dass alle partiellen Ableitungen nach den Variablen c_l verschwinden:

$$\frac{\partial F}{\partial c_l} = 0 \quad \text{für alle } l = 1, ..., n-1 \; . \tag{3.65}$$

Ausführlich geschrieben lautet die Funktion F:

$$\begin{aligned} F(c_1, ..., c_{n-1}) = {} & c_1^2 [u_1, u_1] + c_1 c_2 [u_1, u_2] + c_1 c_3 [u_1, u_3] + ... + c_1 c_{n-1} [u_1, u_{n-1}] + \\ & c_2 c_1 [u_2, u_1] + c_2^2 [u_2, u_2] + c_2 c_3 [u_2, u_3] + ... + c_2 c_{n-1} [u_2, u_{n-1}] + \\ & \vdots \\ & c_{n-1} c_1 [u_{n-1}, u_1] + c_{n-1} c_2 [u_{n-1}, u_2] + c_{n-1} c_3 [u_{n-1}, u_3] + ... \\ & \qquad + c_{n-1}^2 [u_{n-1}, u_{n-1}] \\ & -2c_1 < u_1, f > -2c_2 < u_2, f > - ... - 2c_{n-1} < u_{n-1}, f > . \end{aligned}$$

Wir leiten F partiell nach den Koeffizienten c_i ab. Mit der Symmetrieeigenschaft $[u_i, u_j] = [u_j, u_i]$ gilt dann

$$\begin{aligned}
\frac{\partial F}{\partial c_1} = 0 : 2c_1[u_1,u_1] + 2c_2[u_1,u_2] + 2c_3[u_1,u_3] + ... \\
+2c_{n-1}[u_1,u_{n-1}] - 2 < u_1, f >= 0 \,, \\
\frac{\partial F}{\partial c_2} = 0 : 2c_1[u_2,u_1] + 2c_2[u_2,u_2] + 2c_3[u_2,u_3] + ... \\
+2c_{n-1}[u_2,u_{n-1}] - 2 < u_2, f >= 0 \,, \\
\vdots \\
\frac{\partial F}{\partial c_{n-1}} = 0 : 2c_1[u_{n-1},u_1] + 2c_2[u_{n-1},u_2] + 2c_3[u_{n-1},u_3] + ... \\
+2c_{n-1}[u_{n-1},u_{n-1}] - 2 < u_{n-1}, f >= 0 \,.
\end{aligned}$$

Dies ist ein LGS für die gesuchten Größen $c_1, ..., c_{n-1}$. Führen wir die Matrix $\mathbf{A}$ und die rechte Seite $\mathbf{b}$ ein, so erhalten wir das LGS in der Form

$$\underbrace{\begin{pmatrix} [u_1,u_1] & [u_1,u_2] & [u_1,u_3] & \dots & [u_1,u_{n-1}] \\ [u_2,u_1] & [u_2,u_2] & [u_2,u_3] & \dots & [u_2,u_{n-1}] \\ & & \vdots & & \\ [u_{n-1},u_1] & [u_{n-1},u_2] & [u_{n-1},u_3] & \dots & [u_{n-1},u_{n-1}] \end{pmatrix}}_{\mathbf{A}} \underbrace{\begin{pmatrix} c_1 \\ c_2 \\ \vdots \\ c_{n-1} \end{pmatrix}}_{\mathbf{c}} = \underbrace{\begin{pmatrix} < u_1, f > \\ < u_2, f > \\ \vdots \\ < u_{n-1}, f > \end{pmatrix}}_{\mathbf{b}}$$

oder in Kurzschreibweise

$$\mathbf{Ac} = \mathbf{b} \,.$$

Ist $\bar{\mathbf{c}} = (\bar{c}_1, \bar{c}_2, \dots, \bar{c}_{n-1})$ die Lösung des LGS, dann ist

$$\bar{u}_h(x) = \sum_{i=1}^{n-1} \bar{c}_i u_i(x) \tag{3.66}$$

die gesuchte Funktion aus S, für die das Energieminimum angenommen wird. Zunächst ist die Bedingung, dass die partiellen Ableitungen Null sind, nur eine notwendige Bedingung für ein lokales Extremum. Man kann jedoch weiter zeigen, dass die Matrix $\mathbf{A}$ positiv definit ist, d.h. alle Eigenwerte positiv sind. Daher ist die Lösung des LGS gleichzeitig das lokale Minimum der Funktion F.

Bemerkung: In der oben durchgeführten Diskussion wurde noch nicht die spezielle Wahl der Basisfunktionen eingegangen. D.h. das angegebene LGS mit der Matrix $\mathbf{A}$ ist noch unabhängig von der Wahl dieser Basisfunktionen. Wir werden erst im folgenden Beispiel die spezielle Wahl der Basisfunktionen berücksichtigen.

Beispiel 3.6 (Mit MAPLE-Worksheet). Gegeben sei das Randwertproblem

$$-y''(x) = 10 \quad \text{in } [0,1], \qquad y(0) = y(1) = 0\ .$$

Wählen wir eine Unterteilung des Intervalls $[0,1]$ in n Teilintervalle der Länge $h = \frac{1}{n}$ und als Basisfunktionen die in Schritt 2 eingeführten Dreiecksfunktionen $u_i(x)$, müssen wir die Ansatzfunktion

$$u_h(x) = \sum_{i=1}^{n-1} c_i u_i(x)$$

in das Energiefunktional

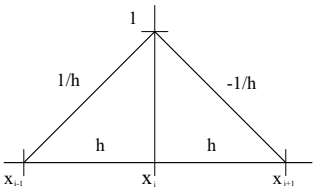
$$E(u) = [u,u] - 2 < u, f >$$

einsetzen. In unserem Beispiel ist $p = 1, q = 0$ und $f = 10$, so dass

$$[u,v] = \int_0^1 u'(x)\, v'(x)\, dx \quad \text{und} \quad < u, f >= \int_0^1 u(x)\, 10\, dx\ .$$

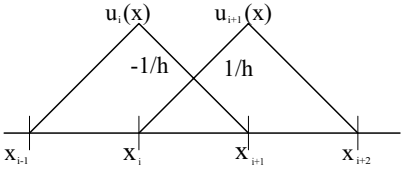
Beim Aufstellen der Matrix $\mathbf{A} = ([u_i, u_j])_{i,j}$ stellt man fest, dass $[u_i, u_j] = 0$ für $|i-j| > 1$, da sich die Träger der Dreiecksfunktionen nicht überschneiden und somit das Integral Null ergibt. Daher müssen wir nur $[u_i, u_i]$ und $[u_i, u_{i+1}]$ bzw. $< u_i, f >$ für die rechte Seite $\mathbf{b}$ berechnen:

(1) $[u_i, u_i]$:



$$\begin{aligned}[u_i, u_i] &= \int_0^1 (u_i'(x))^2\, dx \\ &= \int_{x_{i-1}}^{x_i} (\tfrac{1}{h})^2\, dx + \int_{x_i}^{x_{i+1}} (-\tfrac{1}{h})^2\, dx \\ &= \tfrac{1}{h^2} h + \tfrac{1}{h^2} h = \tfrac{2}{h}.\end{aligned}$$

(2) $[u_i, u_{i+1}]$:



$$\begin{aligned}[u_i, u_{i+1}] &= \int_{x_i}^{x_{i+1}} u_i'(x)\ u_{i+1}'(x)\ dx \\ &= \int_{x_i}^{x_{i+1}} (-\tfrac{1}{h})\ \tfrac{1}{h}\ dx = -\tfrac{1}{h}\ .\end{aligned}$$

(3) Für die Berechnung der rechten Seite des LGS gilt

$$b_i =< u_i, f >= \int_{x_{i-1}}^{x_{i+1}} u_i(x) \cdot 10 \, dx = \frac{1}{2} 2h \cdot 1 \cdot 10 = 10\, h \; .$$

Der Wert $[u_{i-1}, u_i]$ ergibt sich aus der Symmetriebedingung $[u_{i-1}, u_i] = [u_i, u_{i+1}]$.

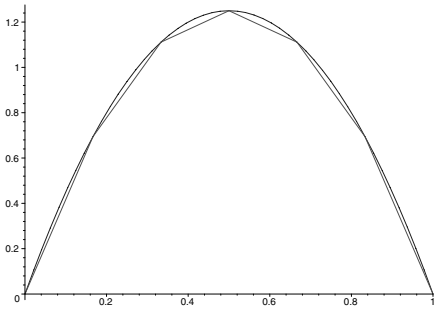


Abb. 3.18. Vergleich der numerischen mit der exakten Lösung

Insgesamt erhalten wir das lineare Gleichungssystem

$$\mathbf{Ac} = \mathbf{b}$$

mit der Tridiagonalmatrix $\mathbf{A}$ und der rechten Seite $\mathbf{b}$

$$\mathbf{A} = \frac{1}{h} \begin{pmatrix} 2 & -1 & \dots & 0 \\ -1 & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & -1 \\ 0 & \dots & -1 & 2 \end{pmatrix}, \qquad \mathbf{b} = h \begin{pmatrix} 10 \\ 10 \\ \vdots \\ 10 \end{pmatrix} \; .$$

Dieses LGS ist für eine Unterteilung von $n = 6$ im **Worksheet** gelöst. Man erhält

$$\bar{c}_1 = \frac{25}{36},\ \bar{c}_2 = \frac{10}{9},\ \bar{c}_3 = \frac{5}{4},\ \bar{c}_4 = \frac{10}{9},\ \bar{c}_5 = \frac{25}{36} \; .$$

Damit ist

$$\bar{u}_h(x) = \sum_{i=1}^{5} \bar{c}_i \, u_i(x)$$

eine Näherung für die Lösung der Differenzialgleichung. Zum grafischen Vergleich wird die exakte Lösung des Randwertproblem $y(x) = -5x^2 + 5x$ mit in das Schaubild aufgenommen. □

Beispiel 3.7 (Mit MAPLE-Worksheet). Gesucht ist eine Näherung für die Lösung des Randwertproblems

$$-y''(x) = 10 \quad \text{in } [0,1], y(0) = y_a \text{ und } y(1) = y_b \,. \tag{3.67}$$

Um dieses RWP mit nicht verschwindenden Randbedingungen zu lösen, kann man zwei alternative Wege wählen:

1. Man transformiert nach Bemerkung (1) das Randwertproblem auf

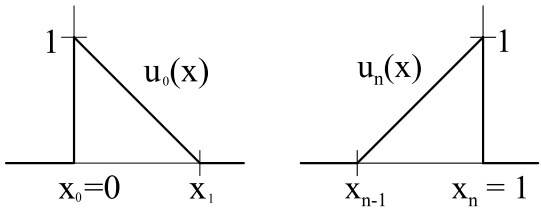
$$-u''(x) = 10, \quad u(0) = u(1) = 0 \,,$$

löst dieses Problem analog zu Beispiel 1 mit der gleichen Matrix A und rechten Seite b, da $p' = 0$ und $q = 0$. Anschließend setzt man

$$y(x) = u(x) + (1-x)y_a + x\, y_b,$$

um die gesuchte Näherung zu erhalten, siehe **Worksheet**.

2. Alternativ erweitert man das Konzept der Basisfunktionen, indem man zu $u_1(x), \ldots, u_{n-1}(x)$ noch die beiden "halben" Dreiecksfunktionen $u_0(x)$ und $u_n(x)$ für den linken bzw. rechten Rand hinzunimmt.



(a) Linker Rand (b) Rechter Rand

Abb. 3.19. Basisfunktionen am linken und rechten Rand

Als Ansatzfunktion wählt man

$$u_h(x) = y_a\, u_0(x) + \sum_{i=1}^{n-1} c_i\, u_i(x) + y_b\, u_n(x)$$

mit den fest vorgegebenen Randwerten y_a und y_b. Bei der Minimierung des Energiefunktionals bleiben alle Gleichungen bis auf die erste und letzte erhalten. Diese werden modifiziert zu

$$\underline{2y_a[u_0,u_1]} + 2c_1[u_1,u_1] + \cdots + 2c_{n-1}[u_1,u_{n-1}] - 2 < u_1, f > = 0$$

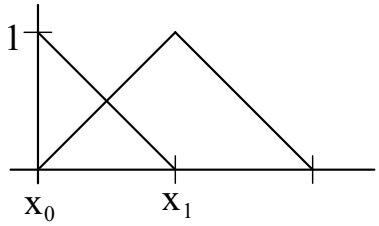
bzw.

$$2c_1[u_{n-1},u_1] + \cdots + 2c_n[u_{n-1},u_{n-1}] + \underline{2y_b[y_{n-1},y_n]} - 2 < u_{n-1}, f > = 0.$$

Daher bleibt die Matrix **A** wieder erhalten, lediglich die rechte Seite **b** des LGS wird an der ersten und letzten Stelle modifiziert:

$$\mathbf{b} = \begin{pmatrix} < u_1, f > & - & y_a[u_0,u_1] \\ < u_2, f > & & \\ \vdots & & \\ < u_{n-2}, f > & & \\ < u_{n-1}, f > & - & y_b[u_{n-1},u_n] \end{pmatrix}.$$

Speziell für obiges Randwertproblem gilt



$$\begin{aligned}[u_0,u_1] &= \int_{x_0}^{x_1} u_0'(x)\ u_1'(x)dx \\ &= \int_{x_0}^{x_1} (-\tfrac{1}{h})\tfrac{1}{h}dx \\ &= -\tfrac{1}{h^2}h = -\tfrac{1}{h}\end{aligned}$$

bzw. analog $[u_{n-1},u_n] = -\frac{1}{h}$.

Damit ist

$$\mathbf{b} = \begin{pmatrix} < u_1, f > & + & \frac{y_a}{h} \\ < u_2, f > & & \\ \vdots & & \\ < u_{n-2}, f > & & \\ < u_{n-1}, f > & + & \frac{y_b}{h} \end{pmatrix}.$$

Die Rechnung ist in einem separaten **Worksheet** ausgeführt und mit der exakten Lösung des Problems verglichen.

Zusammenfassung: Bei der Beschreibung der Finiten-Elemente-Methode wurde das Randwertproblem nicht direkt über die Differenzialgleichung gelöst, sondern über die zugehörige Variationsaufgabe. Es wurde ein Minimum des zur Differenzialgleichung gehörenden Energiefunktionals gesucht. Da auch dieses Problem in der Regel nicht exakt lösbar ist, muss eine Approximation ausgeführt werden. Dazu beschränkt man sich bei der Suche nach dem Minimum auf eine kleinere Klasse von Funktionen: Man unterteilt das Intervall

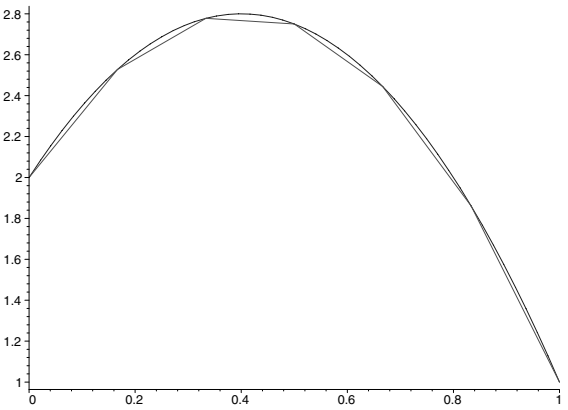


Abb. 3.20. Vergleich der numerischen (gestrichelt) mit der exakten (durchgezogen) Lösung

$[a, b]$ in n Teilintervalle (finite Elemente) und betrachtet in jedem Punkt x_i nur Dreiecksfunktionen (Basisfunktionen). Das Minimum wird dann in der Menge der Überlagerungen dieser Basisfunktionen

$$u_h(x) = \sum_{i=1}^{n-1} c_i \; u_i(x)$$

gesucht. Durch die Einschränkung auf lokale Basisfunktionen (hier stetige, stückweise lineare Funktionen) reduziert sich das Problem auf das Lösen eines LGS für die Koeffizienten c_i, welche genau die gesuchten Funktionswerte an den Stellen x_i sind. Durch die lokalen Basisfunktionen ist die Matrix des LGS tridiagonal und somit schnell lösbar.

Bemerkung: Wir haben hier die Idee der finiten Elemente mit den lokalen Basisfunktionen im Rahmen des Ritzschen Verfahrens benutzt. Hier wurden die Freiheitsgrade aus der Variationsaufgabe bestimmt. Die Finite-Elemente-Methode kann auch mit der Methode der gewichteten Residuen kombiniert werden. Es ist sogar so, dass bei den Problemen, für die eine Variationsformulierung existiert, das Galerkin-Verfahren die gleiche Näherungslösung wie das Ritzsche Verfahren liefert.

3.9 Bemerkungen und Entscheidungshilfen

Die Theorie der Randwertprobleme ist weitaus komplizierter als die der Anfangswertprobleme. Sie wird zum Beispiel bei Heuser [32] einführend be-

handelt. Das Schießverfahren, wie wir es hier vorgestellt haben, wird auch Einfachschießverfahren genannt. Neben diesem gibt es noch den Ansatz des Mehrfachschießverfahrens oder auch Mehrzielmethode genannt, bei dem das Gesamtintervall in einzelne Intervalle aufgeteilt wird und dann in jedem Teilintervall das Schießverfahren durchgeführt wird. Dieser Ansatz ist im Allgemeinen stabiler und wird für praktische Probleme dem einfachen Verfahren vorgezogen. Eine Einführung in das Mehrfachschießverfahren findet sich zum Beispiel bei Stoer und Bulirsch [59], Rechenprogramme in [47].

Differenzenverfahren sind von der Idee her einfach zu verstehen. Die Differenzengleichungen an den Stützstellen liefern ein Gleichungssystem. Dies ist im Falle einer linearen Differenzialgleichungen ein lineares Gleichungssystem. Insbesondere für lineare Differenzialgleichungen 2. Ordnung ist es als tridiagonales Gleichungssystem einfach und schnell mit dem Thomas-Algorithmus zu lösen. Bei nichtlinearen Problemen ergibt sich ein nichtlineares Gleichungssystem. Hier muss ein geeignetes Iterationsverfahren gefunden werden. Das Newton-Verfahren liefert sehr schnell eine genaue Lösung, allerdings muss der Startwert im Allgemeinen gut sein, d.h. in der Nähe der Lösung liegen. Es wird somit gern kombiniert mit einem anderen Iterationsverfahren, wie sie im Anhang B aufgeführt sind, welche dann ein paar Iterationen einen guten Sartwert für das Newton-Verfahren liefert. Differenzenverfahren werden meist dann eingesetzt, wenn die erforderliche Genauigkeit nicht sehr hoch sein muss. Differenzenquotienten höherer Ordnung erhöhen den Rechenaufwand deutlich. Extrapolationsmethoden lassen sich ebenso zur Erhöhung der Ordnung einsetzen. Ein Standardwerk zur numerischen Lösung von Randwertproblemen mit Differenzenverfahren ist das Buch von Keller [36].

Die Methode der gewichteten Residuen wird nur für spezielle Problemklassen mit globalen Ansatzfunktionen eingesetzt. Für diese Probleme kennt man die Eigenschaften der Lösung und geeignete zugehörige Basisfunktionen. Ansonsten benutzt man die Basis der lokal definierten Ansatzfunktionen, d.h. die Finite-Elemente-Methode. Die Berechnung der optimalen Koeffizienten kann über die verschiedenen Methoden der gewichteten Residuen erfolgen oder falls für die Differenzialgleichung ein zugehöriges Variationsproblem existiert, über dessen Minimierung. Dies haben wir hier exemplarisch ausgeführt. Die Hauptanwendungsgebiete der Finiten-Elemente sind allerdings die partiellen Differenzialgleichungen. Darauf werden wir zurückkommen. Bücher über die Methode der finiten Elemente enthalten als Einführung meist die Behandlung von Randwertproblemen von gewöhnlichen Differenzialgleichungen, so z.B. Jung und Langer [35], Braess [5], Hanke-Bourgeois [30] und Großmann und Roos [24].

Kapitel mit der numerischen Behandlung von Randwertproblemen für gewöhnliche Differenzialgleichungen allgemein gibt es z.B. in den Numerik-Büchern von Kress [38], Schwarz und Köckler [55], Stoer und Bulirsch [59],

Faires und Burden [19] und Plato [46]. Dort werden im Rahmen der numerischen linearen Algebra auch die numerische Lösung von Eignewertproblemen behandelt. Ein klassisches Werk zu Eigenwertproblemen, bei dem die Anwendungen im Vordergrund stehen, ist das Buch von Collatz [9]. Software zu Randwertproblemen für gewöhnliche Differenzialgleichungen enthalten die großen Software-Bibliotheken für gewöhnliche Differenzialgleichungen ebenso die IMSL- [68] oder NAG-Bibliothek [45]. In MATLAB gibt es entsprechende Programme. Für die numerische Lösung mit Finiten-Elemente-Verfahren gibt es eine ganze Reihe von kommerziellen Programmpaketen wie ABAQUS, ANSYS und NASTRAN und PERMAS.

3.10 Beispiele und Aufgaben

Aufgabe 3.1 (Nichtlineares Schießverfahren). Lösen sie das RWP für die *nichtlineare* Differenzialgleichung

$$\begin{aligned} y''(x) &= (1 + y'^2(x))^{(3/2)}\, x(x-1)\,, \\ y(0) &= y(1) = 0 \end{aligned}$$

mit dem allgemeinen Schießverfahren.

Lösung: MAPLE**-Worksheet** □

Aufgabe 3.2 (Lineares Schießverfahren). Lösen sie die *lineare* Differenzialgleichung

$$\begin{aligned} y''(x) &= x(x-1)\, y'(x) + 1\,, \\ y(0) &= y(1) = 0 \end{aligned}$$

mit dem vereinfachten Schießverfahren.

Lösung: MAPLE**-Worksheet** □

Aufgabe 3.3 (Schießverfahren). Gegeben ist das Randwertproblem

$$\begin{aligned} y''(x) + (1 + x^2)\, y(x) &= -1, \\ y(-1) = y(1) &= 0, \end{aligned}$$

welches die Knickbiegung eines elastischen Stabs unter Einwirkung einer axial angreifenden Druckkraft und konstanter Querbelastung beschreibt.

Lösen sie dieses Problem mit dem Schießverfahren. Zeigen sie, indem sie $h = 0.1, 0.05, 0.01$ setzen, dass die mit dem Schießverfahren berechneten

Näherungen dieselbe Fehlerordnung besitzen wie die Differenzennäherungen, aus denen sie hervorgegangen sind. □

Aufgabe 3.4 (Differenzenverfahren). Lösen sie das Randwertproblem

$$-y''(x) + y(x) = 1 \qquad \text{mit} \qquad y(0) = y(1) = 0$$

mit der Differenzenmethode und vergleichen sie die numerische Lösung mit der exakten Lösung

$$y(x) = \frac{e-1}{1-e^2}\,(e^x + e^{-x+1}) + 1\ .$$

□

Aufgabe 3.5 (Differenzenverfahren). Lösen sie das Randwertproblem

$$y^{(4)}(x) + y''(x) = -0.5\ ,$$
$$y(0) = y(1) = 0,\ y'(0) = 1,\ y'(1) = 0$$

mit der Differenzenmethode. Berücksichtigen sie dabei die Ableitungen an den Rändern, indem sie für die Formeln am Rand die Prozedur **FDFormeln** verwenden. Vergleichen sie die numerische Lösung mit der analytischen.

Lösung: MAPLE**-Worksheet** □

Aufgabe 3.6 (Eigenwertproblem). Lösen sie das Eigenwertproblem

$$y''(x) + \lambda\, y(x) = 0, \qquad y(0) = y(1) = 0$$

mit dem Differenzenverfahren. Vergleichen sie die numerisch gefundenen Eigenwerte mit den exakten $\lambda_n = n^2\,\pi^2$. Wählen sie dazu als Anzahl der Intervallunterteilung $7, 13, 21$. Vergleichen sie grafisch die numerischen Eigenfunktionen mit den Eigenfunktionen der exakten Lösung

$$y_n(x) = c_n\ sin(n\,\pi\,x)\ .$$

Lösung: MAPLE**-Worksheet** □

Aufgabe 3.7 (Methode der gewichteten Residuen). Lösen sie das Randwertproblem

$$-y''(x) + y(x) = 1; \qquad y(0) = y(1) = 0$$

durch die Methode der gewichteten Residuen, indem sie geeignete, globale **Ansatzfunktionen** erzeugen. Vergleichen sie die so gefundene Lösung mit der exakten

$$y(x) = \frac{e-1}{1-e^2}\,(e^x + e^{-x+1}) + 1\,.$$

Lösung: MAPLE**-Worksheet** □

Aufgabe 3.8 (Variationsaufgabe). Prüfen sie durch die Anwendung der Eulerschen Differenzialgleichung, dass das Energiefunktional

$$E(y) = \int_0^1 (y''^2(x) - \lambda^2 y'^2(x))\,dx$$

auf die Differenzialgleichung

$$y^{(4)}(x) + \lambda^2 y''(x) = 0 \quad \text{in} \quad [0,1]$$

führt. □

Aufgabe 3.9 (Ritzsches Verfahren).

a) Zeigen sie, dass das zur Variationsaufgabe (unter den Nebenbedingungen $y(0) = 4, y(1) = 1$)

$$E(\varphi) = \frac{1}{2}\int_0^1 (\varphi'^2(x) + \varphi^3(x))\,dx$$

gehörende Randwertproblem

$$y''(x) = \frac{3}{2}\,y^2(x) \;\text{ mit }\; y(0) = 4, y(1) = 1$$

lautet.

b) Zeigen sie, dass

$$y(x) = \frac{4}{(1+x)^2}$$

die Lösung des Randwertproblems ist.

c) Wenden sie auf das Problem das Ritz-Verfahren mit geeigneten Ansatzfunktionen an und vergleichen sie die numerische Lösung mit der exakten.

d) Prüfen sie, ob

$$u_0(x) = 4 - 3x,\ u_n(x) = x - x^{(n+1)} \quad (n \geq 1)$$

erlaubte Ansatzfunktionen sind, d.h. u_0 die inhomogenen und u_n die homogenen Randbedingungen erfüllen.

e) Wählen sie als Ansatz ($n = 1$)

$$u(x) = u_0(x) + c_1\, u_1(x)$$

und prüfen sie, dass $c_1 = -3.041320$ bzw. $c_1 = -51.403125$ gilt. Zeichnen Sie beide Varianten und vergleichen sie diese mit der exakten Lösung.

f) Wählen sie als Ansatz ($n = 2$)

$$u(x) = u_0(x) + c_1\, u_1(x) + c_2\, u_2(x)$$

und prüfen sie durch Lösen des nichtlinearen Gleichungssystems,

$$\begin{aligned} 18c_1^2 + 54c_1c_2 + 41c_2^2 + 980c_1 + 1452c_2 + 2814 &= 0 \\ 27c_1^2 + 82c_1c_2 + 63c_2^2 + 1452c_1 + 2250c_2 + 3906 &= 0, \end{aligned}$$

ob ($c_1 = -7.07003620$ und $c_2 = 2.72044119$) bzw. ($c_1 = -32.15840455$ und $c_2 = -12.59374898$) gilt. Zeichnen sie beide Varianten und vergleichen sie diese mit der exakten Lösung.

Lösung: MAPLE**-Worksheet** □

Aufgabe 3.10 (Ritzsches Verfahren). Wenden sie auf das Eigenwertproblem

$$y''(x) + \lambda\, y(x) = 0, \qquad y(0) = y(1) = 0$$

das Ritz-Verfahren an. Ein zugehöriges Energiefunktional lautet

$$E(u) = \frac{1}{2} \int_0^1 (u'^2(x) - \lambda u^2(x))\, dx\ .$$

Vergleichen sie die numerisch gefundenen Eigenwerte mit den exakten $\lambda_n = n^2\, \pi^2$. Vergleichen sie grafisch die numerischen Eigenfunktionen mit den richtigen Eigenfunktionen

$$y_n(x) = c_n\, sin(n\, \pi\, x)\ .$$

□

Aufgabe 3.11 (Variationsaufgabe). Prüfen sie durch die Anwendung der Eulerschen Differenzialgleichung, dass

$$E(u) = \frac{1}{2} \int_a^b (p\, u'^2(x) + q\, u^2(x) - 2u(x)\, f(x))\, dx$$

unter der Nebenbedingung $u(a) = u_a$, $u(b) = u_b$ ein zum Randwertproblem

$$-(p\, u')' + q\, u = f, \quad u(a) = u_a, \quad u(b) = u_b$$

gehörendes Energiefunktional ist. □

Aufgabe 3.12 (Finite-Elemente-Methode). Lösen sie das Randwertproblem

$$-y''(x) = \pi^2 sin(\pi\, x), \qquad y(0) = 0,\ y(1) = 1$$

mit der Finiten-Elemente-Methode. Verwenden sie hierzu das Energiefunktional aus obiger Aufgabe. Vergleichen sie die numerische Lösung mit der exakten Lösung

$$y(x) = sin(\pi\, x) + x\ .$$

□

Aufgabe 3.13 (Finite-Elemente-Methode). Lösen sie das Randwertproblem

$$\begin{aligned} -y''(x) + y(x) &= 0 \\ y(0) = 5, \quad y(2) &= 4 \end{aligned}$$

mit der Finiten-Elemente-Methode. Beachten sie dabei, dass das Energiefunktional $E(u) = [u, u]$ durch

$$[u, v] = \int_1^2 (u'(x)\ v'(x) + u(x)\ v(x))\, dx$$

gegeben ist.
(a) Transformieren sie dazu das Problem auf eines mit verschwindenden Randbedingungen (mit MAPLE-**Worksheet**).
(b) Modifizieren sie alternativ zu (a) die Basisfunktionen an den Intervallgrenzen (mit MAPLE-**Worksheet**). □

4 Grundlagen der partiellen Differenzialgleichungen

Viele technische und physikalische Vorgänge lassen sich nicht oder nur in Spezialfällen durch eine Koordinate und mit gewöhnlichen Differenzialgleichungen beschreiben. Bei mehr als einer unabhängigen Variablen ist das mathematische Modell dann eine **partielle Differenzialgleichung**. Man denke nur an eine Strömung, welche sich im ganzen Raum ausbreitet. In diesem Fall hat man als unabhängige Variablen die drei Raumvariablen x, y, z, und für instationäre Probleme zusätzlich die Zeit t. Bei einer instationären Strömung sind zum Beispiel Geschwindigkeit v und der Druck p als Funktionen

$$v = v(x, y, z, t)\,, \quad p = p(x, y, z, t)$$

gesucht. Die Lösung ist eine Funktion von mehreren Variablen und das mathematische Modell eine oder ein System von partiellen Differenzialgleichungen.

Wir werden uns der Einfachheit halber oft auf zwei unabhängige Variablen x, y oder auch x, t beschränken, falls ein zeitabhängiges Problem vorliegt. Ganz allgemein hat eine solche zweidimensionale **partielle Differenzialgleichung erster Ordnung** die Form

$$F(x, y, u(x, y), u_x(x, y), u_y(x, y)) = F(x, y, u, u_x, u_y) = 0\,.$$

Die partiellen Ableitungen werden der kürzeren Schreibweise halber meist durch Indizes bezeichnet, d.h. $u_x = \frac{\partial u}{\partial x}$ zum Beispiel. Die Ordnung der partiellen Differenzialgleichung ist der Grad der höchsten Ableitung, welche in der Gleichung auftritt. Die allgemeinste Form einer **partiellen Differenzialgleichung zweiter Ordnung** lautet somit

$$F(x, y, u, u_x, u_y, u_{xx}, u_{xy}, u_{yy}) = 0\,.$$

Der Schritt von einem zwei- zu einem dreidimensionalen Problem ist bezüglich der Konstruktion numerischer Methoden oft nur eine einfache Übertragung der Algorithmen und der Ideen. Der Übergang von der eindimensionalen gewöhnlichen zur zweidimensionalen partiellen Gleichung ist meist der wesentliche Schritt, der neue Ansätze und Konstruktionen erfordert. Bei der Übertragung der Algorithmen in Rechenprogramme und der numerischen Simulation praktischer Probleme muss man bei dieser Aussage allerdings etwas

skeptisch sein. Bezüglich Programmstruktur, Datenmenge und Rechenzeit sind die dreidimensionalen numerischen Simulationen weitaus aufwändiger. Zudem sind die simulierten dreidimensionalen physikalischen Phänomene oft auch komplexer. Kurz ausgedrückt, das dreidimensionale praktische Problem kann eine ganz neue Dimension bekommen.

Schon bei den gewöhnlichen Differenzialgleichungen sah man, dass die Lösungen problemabhängig ein sehr unterschiedliches Verhalten zeigen können. Bei partiellen Differenzialgleichungen ist dies noch stärker ausgeprägt. Die erste Orientierung bezüglich der Eigenschaften einer partiellen Differenzialgleichung gibt die Einteilung in drei Klassen. Die Vertreter dieser Klassen beschreiben grundlegende verschiedene physikalische Vorgänge: **Stationäre Vorgänge**, **dissipative Phänomene** und **Wellenausbreitungsvorgänge**. Zwischen den einzelnen Vertretern einer Klasse gibt es weitere Unterschiede. Das Verhalten der Lösungen nichtlinearer Differenzialgleichungen sind oft sehr verschieden von denen linearer Probleme. Für die numerische Simulation ist es somit wichtig, über die Eigenschaften des Problems und seiner Lösungen Information zu haben. Diese spielen eine wichtige Rolle bei der Konstruktion der numerischen Verfahren.

Dieses Kapitel enthält deshalb etwas Theorie zu den partiellen Differenzialgleichungen. Es wird zunächst auf die Klassifizierung eingegangen. An einigen typischen Vertretern aus technischen und physikalischen Anwendungen werden die grundlegenden Eigenschaften der Lösungen skizziert, die Vorgabe von Nebenbedingungen wie Anfangs- oder Randwerte diskutiert. Neben den drei Klassen von partiellen Differenzialgleichungen gibt es auch drei verschiedene Arten von Nebenbedingungen: **die Anfangswertprobleme (AWP), die Randwertprobleme (RWP) und die Anfangs-Randwertprobleme (ARWP)**. Wir können in diesem Kapitel natürlich nur einen kurzen Überblick geben, der aber für das Verständnis der Konstruktion der numerischen Methoden für die hier behandelten partiellen Differenzialgleichungen zwingend nötig ist.

4.1 Klassifizierung der partiellen Differenzialgleichungen 2. Ordnung

Die Klassifizierung möchten wir zunächst an einer partiellen Differenzialgleichung zweiter Ordnung der Form

$$a\, u_{xx} + b\, u_{xy} + c\, u_{yy} = h \tag{4.1}$$

ausführen. Wir beschränken uns dabei auf zwei Raumdimensionen, die gesuchte Lösung u hängt von den zwei unabhängigen Variablen x und y ab:

$u = u(x, y)$. Vorausgesetzt wird hier und im Folgenden, dass die Lösung mindestens zweimal stetig differenzierbar ist.

Die Gleichung (4.1) heißt **linear mit konstanten Koeffizienten**, **linear** oder **quasilinear**, wenn

$$a \in \mathbb{R},\ b \in \mathbb{R},\ c \in \mathbb{R} \tag{4.2}$$

oder

$$a = a(x, y),\quad b = b(x, y),\quad c = c(x, y) \tag{4.3}$$

beziehungsweise

$$a = a(x, y, u),\quad b = b(x, y, u),\quad c = c(x, y, u) \tag{4.4}$$

gilt.

Für die rechte Seite h wird nur vorausgesetzt, dass sie keine zweiten Ableitungen enthält:

$$h = h(x, y, u, u_x, u_y)\ . \tag{4.5}$$

Solange keine Verwechslungen auftreten können, werden die partiellen Ableitungen mit den entsprechenden Indizes abgekürzt:

$$u_{xx} = \frac{\partial^2 u}{\partial x^2},\quad u_{xy} = \frac{\partial^2 u}{\partial x \partial y},\quad u_{yy} = \frac{\partial^2 u}{\partial y^2}\ ,$$
$$u_x = \frac{\partial u}{\partial x},\quad u_y = \frac{\partial u}{\partial y}\ ,$$

wie schon in Gleichung(4.1) ausgeführt. Im Folgenden wird zunächst angenommen, dass die Gleichung linear mit konstanten Koeffizienten ist.

Die lineare Differenzialgleichung zweiter Ordnung (4.1) mit konstanten Koeffizienten heißt **elliptisch**, wenn für die Koeffizienten

$$b^2 - 4ac < 0\ , \tag{4.6}$$

sie heißt **parabolisch**, wenn

$$b^2 - 4ac = 0 \tag{4.7}$$

und **hyperbolisch**, wenn

$$b^2 - 4ac > 0 \tag{4.8}$$

erfüllt ist.

Bemerkung: Was haben die partiellen Differenzialgleichungen mit den Kegelschnitten, wie der Ellipse, Parabel oder Hyperbel zu tun? Nun, die Gleichung der Kegelschnitte lautet

$$a\,x^2 + b\,xy + c\,y^2 = h \tag{4.9}$$

mit Konstanten a, b, c und der linearen Funktion $h = h(x, y)$. Gilt (4.6), (4.7) oder (4.8), dann ist die zugehörige Gleichung eine Ellipse, eine Parabel oder eine Hyperbel. Die Namen für die Klassifizierung bei den partiellen Differenzialgleichungen wird aus dieser Analogie abgeleitet.

Eine solche Klasseneinteilung macht zum Einen nur dann Sinn, wenn die Vertreter einer Klasse ähnliche Eigenschaften besitzen. Zum Anderen: Wird das mathematische Modell umformuliert oder ein anderes Koordinatensystem gewählt, darf sich der Typ der Differenzialgleichung nicht ändern: Es wird noch immer das gleiche Problem beschrieben. Es lässt sich durch einfaches Ausrechnen zeigen, dass jede umkehrbare Transformation den Typ der Differenzialgleichung erhält. Man zeigt, dass die Koeffizienten der transformierten Differenzialgleichung wiederum den entsprechenden Bedingungen (4.6) - (4.8) genügen.

Bei den Kegelschnitten kann man mit Hilfe der Hauptachsentransformation die Gleichung (4.9) des Kegelschnittes auf eine einfache Form bringen. So ist die Normalform der Ellipse durch die Gleichung

$$a\,x^2 + c\,y^2 = 1 \quad \text{mit} \quad a > 0,\ c > 0\ ,$$

die Hyperbel durch

$$a\,x^2 - c\,y^2 = 1 \quad \text{mit} \quad a > 0,\ c > 0$$

und die Parabel durch

$$y = a\,x^2 + h \quad \text{mit} \quad a \in \mathbb{R}$$

bestimmt. Bei den partiellen Differenzialgleichungen zweiter Ordnung lässt sich dies in ähnlicher Form durchführen. Es gibt somit einfache kanonische Vertreter der entsprechenden Klassen, an denen in den folgenden Abschnitten die wichtigsten Eigenschaften erklärt werden.

Bemerkungen:

1. Für die Bestimmung des Typs einer partiellen Differenzialgleichung sind nur die Koeffizienten der höchsten Ableitungen maßgebend. Die rechte Seite darf von x und y, von der Funktion u selbst oder auch von den ersten Ableitungen von u abhängen.

2. Die Klassifizierung der Differenzialgleichung zweiter Ordnung (4.1) wurde für konstante Koeffizienten durchgeführt. Sind die Koeffizienten nicht konstant, dann kann sich der Typ in Abhängigkeit von x und y auch ändern, so dass die Klassifizierung nur lokal durchgeführt werden kann. So heißt die lineare Differenzialgleichung (4.1) **elliptisch, parabolisch oder hyperbolisch im Punkte** $(\mathbf{x_0}, \mathbf{y_0})$, wenn die Bedingungen (4.6), (4.7) bzw. (4.8) im Punkte (x_0, y_0) erfüllt sind.

3. Beim quasilinearen Fall kann sich der Typ der Differenzialgleichung auch noch in Abhängigkeit der Lösung u ändern. Man muss dann bei der lokalen Typisierung die Lösung mit berücksichtigen.

4.2 Elliptische Differenzialgleichungen 2. Ordnung

Elliptische Differenzialgleichungen beschreiben stationäre Gleichgewichtszustände in verschiedenen physikalischen Bereichen wie Festigkeitslehre, Magnetostatik, Strömungsmechanik oder Thermodynamik. So ist die stationäre Wärmeverteilung oder das elektrische Potenzial Lösung einer elliptischen Gleichung.

Beispiel 4.1 (Gleichungen der Elektrostatik). Für das elektrostatische Potenzial $\phi = \phi(x, y)$ und das elektrische Feld $\mathbf{E} = \mathbf{E}(x, y)$ eines ebenen stationären Problems gelten die Grundgleichungen

$$\mathbf{E} = -\nabla\phi \,, \tag{4.10}$$

$$\nabla \cdot \mathbf{E} = \frac{1}{\varepsilon}\rho \,. \tag{4.11}$$

Dabei bezeichnet ∇ den Gradienten einer skalaren Funktion, $\nabla\cdot$ die Divergenz einer vektorwertigen Funktion:

$$\nabla\phi = \begin{pmatrix} \phi_x \\ \phi_y \end{pmatrix}, \quad \nabla \cdot \mathbf{E} = \nabla \cdot \begin{pmatrix} E_1 \\ E_2 \end{pmatrix} = (E_1)_x + (E_2)_y \,.$$

Die Variable $\rho = \rho(x, y)$ ist die Ladungsdichte und ε die Dielektrizitätskonstante.

Setzt man die Gleichung (4.10) in die Gleichung (4.11) ein, so erhält man

$$\nabla \cdot \mathbf{E} = (E_1)_x + (E_2)_y = (-\phi_x)_x + (-\phi_y)_y = \frac{1}{\varepsilon}\rho$$

und weiter die sogenannte Poisson-Gleichung für das Potenzial ϕ:

$$\phi_{xx} + \phi_{yy} = -\frac{1}{\varepsilon}\rho \; .$$

Für den ladungsfreien Raum ist $\rho \equiv 0$ und man erhält die sogenannte Laplace-Gleichung

$$\phi_{xx} + \phi_{yy} = 0 \; .$$

Zur Abkürzung der linken Seite schreibt man dabei oft

$$\Delta\phi := \phi_{xx} + \phi_{yy}$$

und nennt Δ den **Laplace-Operator**. □

Zusammenfassung: Allgemein nennt man

$$\Delta u = h \tag{4.12}$$

mit der rechten Seite $h = h(x, y, u, u_x, u_y)$ die **Poisson-Gleichung** und

$$\Delta u = 0 \tag{4.13}$$

die **Laplace-Gleichung** mit $\Delta u = u_{xx} + u_{yy}$.

Man überprüft leicht, dass die Poisson- und die Laplace-Gleichung elliptische Differenzialgleichungen im Sinne unser Typ-Einteilung sind. So ist mit $a = 1$, $c = 1$ und $b = 0$ die Bedingung (4.6) der Elliptizität erfüllt.

Wir haben bisher nur die Differenzialgleichung betrachtet. Das praktische Problem besteht neben den Gleichungen noch aus dem Rechengebiet, in dem dieser Vorgang stattfindet. Wie wir bei den gewöhnlichen Differenzialgleichungen zweiter Ordnung schon gesehen haben, werden zum Erhalt einer eindeutigen Lösung Nebenbedingungen wie Randwerte oder Anfangswerte benötigt. Da elliptische Differenzialgleichungen stationäre Probleme beschreiben, sind die richtigen Nebenbedingungen die Vorgabe von Randwerten. Das Randwertproblem für gewöhnliche Differenzialgleichungen, wie es in Kapitel 3 behandelt wurde, ist der eindimensionale Fall eines elliptischen Problems.

Wir betrachten im Folgenden die Poisson-Gleichung (4.12) in einem Gebiet G mit einem stetigen Rand $\Gamma = \partial G$. Auf dem Rand Γ des Gebietes G werden zusätzliche Bedingungen vorgegeben (siehe Abbildung 4.1).

Das physikalische Problem kann sich so stellen, dass die Werte der Lösung auf dem Rand vorgegeben sind. Eine typische Situation ist hier eine eingespannte Membran. Die Forderung an die Lösung u ist

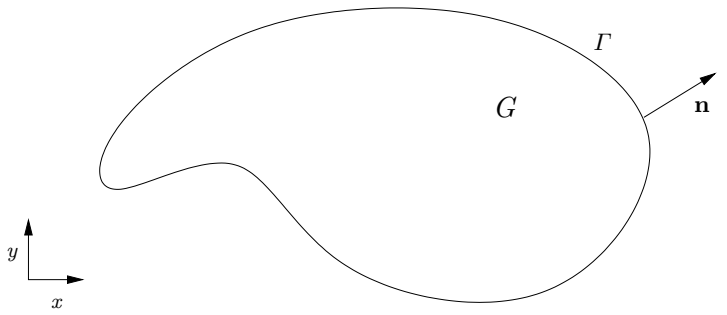


Abb. 4.1. Rechengebiet

$$u = f \quad \text{auf} \quad \Gamma\,, \tag{4.14}$$

wobei $f = f(x, y)$ eine auf dem Rand Γ definierte vorgegebene Funktion ist. Diese Vorgabe von Randwerten nennt man die **Dirichletschen Randbedingungen** oder auch Randbedingungen der 1. Art. Das zugehörige Randwertproblem für die Poisson-Gleichung (4.12), (4.14) nennt man das **Dirichlet-Problem für die Poisson-Gleichung**. Für die speziellen Randwerte (4.14)

$$f = 0 \quad \text{auf} \quad \Gamma \tag{4.15}$$

nennt man die Randbedingungen **homogen** und spricht von **homogenen Dirichlet-Randwerten** oder **homogenem Dirichlet-Problem**.

Bei einem elektrostatischen Problem und der Berechnung des zugehörigen Potenzials liegen oft andere Randbedingungen vor. So verschwindet an einem idealen Leiter das elektrische Feld senkrecht zum Rand und damit die Normalableitung des Potenzials u, die Äquipotenziallinien laufen senkrecht zum Rand. Dies führt auf die **Neumann-Randbedingungen** oder die Randbedingungen der 2. Art. Hier ist die Normalableitung von u auf dem Rand vorgegeben:

$$\frac{\partial u}{\partial \mathbf{n}} = \nabla u \cdot \mathbf{n} = g\,. \tag{4.16}$$

Dabei bezeichnet $\mathbf{n}$ die nach außen gerichtete Normale auf dem Rand Γ von G und $g = g(x, y)$ ist eine auf Γ definierte vorgegebene Funktion. Das zugehörige Randwertproblem (4.12), (4.16) nennt man das **Neumann-Problem für die Poisson-Gleichung**. Ganz analog zum Dirichlet-Problem spricht man von **homogenen Neumann-Randbedingungen**, wenn $g = 0$ auf dem ganzen Rand gilt.

Die Kombination dieser beiden Randbedingungen ist die **Robinsche Randbedingung** oder Randbedingung der 3. Art:

$$\frac{\partial u}{\partial \mathbf{n}} + \sigma u = h \quad \text{auf} \quad \Gamma$$

mit $\sigma = \sigma(x, y)$, $h = h(x, y)$ definiert für alle $(x, y) \in \Gamma$.

Zusammenfassung: Das physikalische Problem für die Poisson-Gleichung stellt sich als **Randwertproblem (RWP)**. Als mögliche Randbedingungen treten die **Dirichletsche Randbedingung**

$$u = f \quad \text{auf} \quad \Gamma\,, \tag{4.17}$$

die **Neumannsche Randbedingung**

$$\frac{\partial u}{\partial \mathbf{n}} = \nabla u \cdot \mathbf{n} = g \quad \text{auf} \quad \Gamma \tag{4.18}$$

oder die **Robinsche Randbedingung**

$$\frac{\partial u}{\partial \mathbf{n}} + \sigma u = h \quad \text{auf} \quad \Gamma \tag{4.19}$$

auf. Der gesamte Rand kann auch in einzelne Abschnitte mit unterschiedlichen Randbedingungen aufgeteilt sein.

Bemerkung: Bei der Vorgabe von Neumannschen Randbedingungen auf dem gesamten Rand des betrachteten Rechengebietes ist die Lösung nicht eindeutig, sondern nur bis auf eine additive Konstante bestimmt.

Schon bei den gewöhnlichen Differenzialgleichungen ist es so, dass nur sehr wenige exakt lösbar sind. Dies gilt noch verstärkt für die partiellen Differenzialgleichungen. Zur Demonstration der Eigenschaften der Lösungen der verschiedenen Typen von Differenzialgleichungen aber auch zur Bereitstellung exakter Lösungen für die Validierung von numerischen Methoden werden im Folgenden einige wenige exakte Lösungen für einfache Probleme angegeben. Weitere findet man, neben den Aussagen von Existenz und Eindeutigkeit von Lösungen, in Büchern über partielle Differenzialgleichungen oder auch über die speziellen Anwendungen.

Beispiel 4.2 (RWP für die Laplace-Gleichung, mit MAPLE-Worksheet). Als Modellfall für die Lösung der Laplace-Gleichung betrachten wir eine Membran, die durch einen rechteckigen Drahtrahmen aufgespannt wird. Eine der Rahmenseiten ist aus der Ebene heraus gebogen. Gesucht ist der Verlauf dieser Membran. Das Ganze ist ein zweidimensionales Problem - eine Fläche im Raum. Wir wählen als unabhängige Koordinaten x und y. Bezeichnet $u = u(x, y)$ die Auslenkung der Membran im Punkt (x, y) in z-Richtung, dann ist u bestimmt durch die Laplace-Gleichung

$$\Delta u = 0 \quad \text{in } [0, a] \times [0, b] \tag{4.20}$$

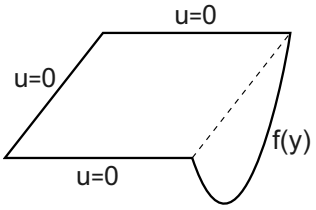


Abb. 4.2. Randbedingungen einer eingespannten Membran

mit den Dirichlet-Randdaten

$$\begin{array}{ll} u(x,0)=0 & \text{für alle } x\in[0,a]\ , \\ u(x,b)=0 & \text{für alle } x\in[0,a]\ , \\ u(0,y)=0 & \text{für alle } x\in[0,b]\ , \\ u(a,y)=f(y) & \text{für alle } x\in[0,b]\ . \end{array} \tag{4.21}$$

Dabei ist die rechte Rechteckseite $x = a$ in z-Richtung gemäß der vorgegebenen Randfunktion $f = f(y)$ gebogen, wie es in Abbildung 4.2 dargestellt ist.

Dieses Problem lässt sich mit einem **Separationsansatz**

$$u(x,y) = X(x)\cdot Y(y)$$

lösen. Durch Einsetzen in die Laplace-Gleichung erhält man zwei gewöhnliche Differenzialgleichungen, eine in Abhängigkeit von x, die andere in y, welche exakt gelöst werden können. Unter Berücksichtigung der Randwerte (4.21) lautet die Lösung in der Form einer Fourierreihe

$$u(x,y) = \sum_{n=1}^{\infty} b_n \sinh\left(n\frac{\pi}{b}x\right)\sin\left(n\frac{\pi}{b}y\right) \tag{4.22}$$

mit den Koeffizienten

$$b_n = \frac{2}{b\sinh\left(n\frac{\pi}{b}a\right)}\int\limits_0^b f(y)\sin\left(n\frac{\pi}{b}y\right)\,dy\ ,$$

welche sich aus der Fourierreihe für die Funktion $f = f(y)$ ableiten. Ausführlich wird dieses Beispiel in [71] behandelt.

Im MAPLE-**Worksheet** werden diese Formeln auf die Randverbiegung

$$f(y) = y(y - b) \quad \text{für } y \in [0, b]$$

angewandt. Gesucht ist die Auslenkung $u = u(x, y)$ der Membran in z-Richtung im gesamten Innern des Rechtecks. Für die Rechnung setzen wir a=2 und b=1.

1. Schritt: Fourier-Zerlegung der Funktion $f(y)$ als punktsymmetrische $2b-$periodische Funktion.

In der praktischen Auswertung der Fourier-Reihe (4.22) können wir nicht unendlich viele Summanden, sondern nur endlich viele, d.h. eine Partialsumme, berechnen. Mit den Randwerten $f = f(y)$ haben wir eine einfache Parabel vorgegeben, so dass einige wenige Glieder der Fourier-Reihe ausreichen sollten. In dem MAPLE-Worksheet vergleichen wir die Partialsumme der Fourier-Reihe für die Randfunktion mit der exakten Funktion $f = f(y)$.

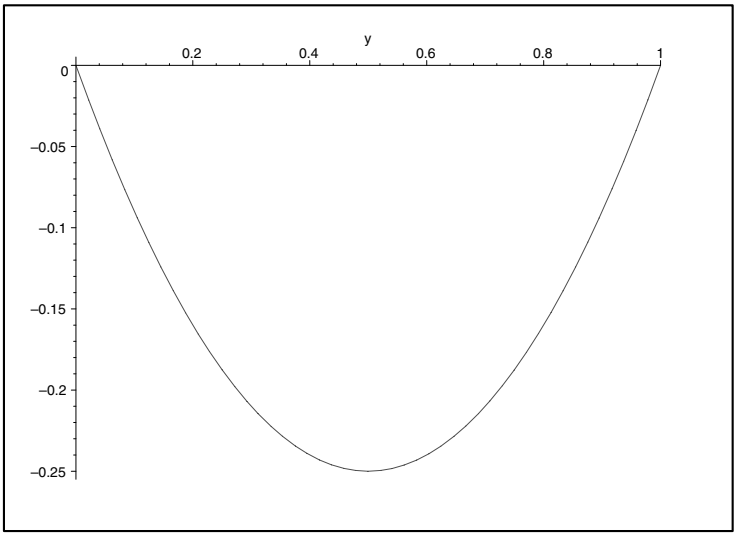


Abb. 4.3. Funktion f und Partialsumme der Fourier-Reihe

Der Vergleich in Abbildung 4.3 zeigt, dass sich zwischen der Partialsumme der Fourier-Reihe mit nur drei Summanden und der exakten Auslenkung $f = f(y)$ grafisch schon kein Unterschied mehr feststellen lässt. Nachdem die Randfunktion somit mit einigen wenigen Gliedern der Reihe sehr gut erfasst wird, sollte man sich bei der Berechnung der Lösung auch auf wenige Glieder beschränken können. Eine Überprüfung der Genauigkeit ist hier leicht möglich, indem eine Rechnung mit mehr Summanden zusätzlich durchgeführt wird.

2. Schritt: Nachdem die Fourier-Koeffizienten b_n von $f = f(y)$ bestimmt wurden, kann man die Formel für die Lösungsdarstellung verwenden. Bei dieser einfachen Verbiegung genügt es, nur wenige Terme der Reihe mitzunehmen. Die Lösung ist in Abbildung 4.4 dargestellt, so wie sie von dem MAPLE-Worksheet geliefert wird. □

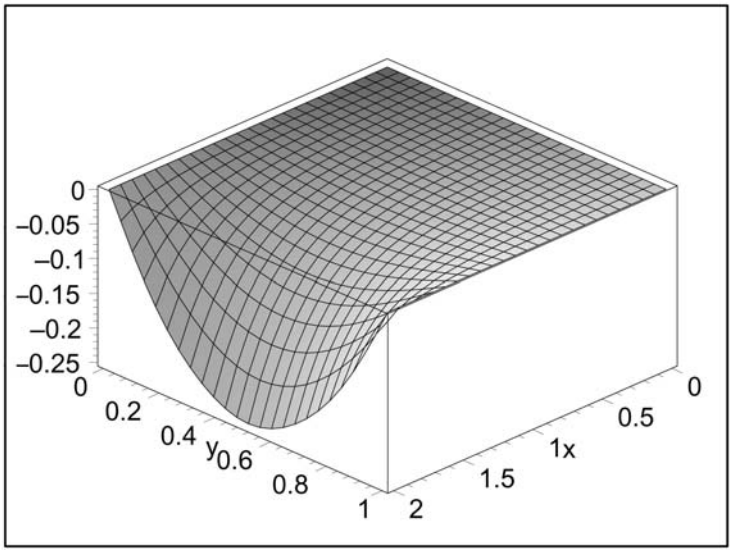


Abb. 4.4. Verbiegung einer Membran mit den Randwerten (4.21)

4.3 Parabolische Differenzialgleichungen 2. Ordnung

Der Modellfall für eine parabolische Gleichung ist die **Wärmeleitungsgleichung**.

Beispiel 4.3. Wir betrachten einen dünnen Metallstab der Länge L mit rechteckigem Querschnitt der Breite b und der Höhe h. $T = T(x,t)$ bezeichnet die Temperatur im Stab an der Stelle x zur Zeit t. Modelliert wird der Wärmetransport in x-Richtung als eindimensionales Problem.

Aus der Thermodynamik sind die folgenden Gesetzmäßigkeiten bekannt:

1. Die Wärmemenge δQ, die in x-Richtung durch eine Fläche $A = b \cdot h$ in der Zeit $\triangle t$ fließt, ist gegeben durch

$$\delta Q = -\lambda \, A \, \Delta t \, \frac{\partial T}{\partial x} \, .$$

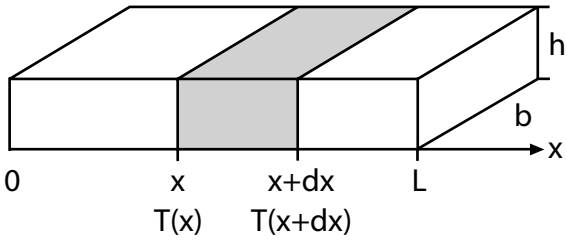


Abb. 4.5. Wärmetransport in x-Richtung

Die Wärmemenge ist proportional zum Temperaturgefälle $\frac{\partial T}{\partial x}$, λ ist die materialspezifische Wärmeleitfähigkeit.

2. Die Wärmemenge δQ, die in einem Volumen mit Masse m und spezifischer Wärme c gespeichert wird, ist gegeben durch

$$\delta Q = c\, m\, (T_2 - T_1) = c\, m\, \Delta T\ ,$$

wenn $\Delta T = (T_2 - T_1)$ die Temperaturdifferenz an den Enden des Volumens ist.

Die Energiebilanz auf das in Abbildung 4.5 gezeigte Volumenelement $dV = b \cdot h \cdot dx$ ergibt:

Änderung der Energie pro Zeiteinheit $\triangle t$ im Massenelement dm ist

= Zufluss der Wärmemenge in das Massenelement durch die Fläche A an der Stelle x

+ Abfluss der Wärmemenge aus dem Massenelement durch die Fläche A an der Stelle $x + dx$

$$\frac{\delta Q}{\partial t} = c\, dm\, \frac{\partial T}{\partial t} = -\lambda\, A \left(\frac{\partial T}{\partial x}\right)_x + \lambda\, A \left(\frac{\partial T}{\partial x}\right)_{x+dx}\ .$$

Linearisiert man den Term $\left(\frac{\partial T}{\partial x}\right)_{x+dx} \approx \left(\frac{\partial T}{\partial x}\right)_x + dx \left(\frac{\partial^2 T}{\partial x^2}\right)_x$ und setzt diese Linearisierung in die obige Gleichung ein, folgt mit $dm = \rho \cdot dV = \rho \cdot b \cdot h \cdot dx$

$$c\, \rho\, b\, h\, dx\, \frac{\partial T}{\partial t} = -\lambda\, A\, \frac{\partial T}{\partial x} + \lambda\, A \left[\frac{\partial T}{\partial x} + dx\, \frac{\partial^2 T}{\partial x^2}\right]. \qquad \square$$

Die allgemeine Form der **Wärmeleitungsgleichung** für die Temperaturverteilung $u = u(x,t)$ ergibt sich dann zu

$$u_t = \kappa u_{xx} \quad .$$

Dabei bezeichnet $\kappa := \frac{\lambda}{c\rho}$ die Temperaturleitzahl.

Um die Temperatur zu einer beliebigen Zeit t bestimmen zu können, benötigen wir sowohl die anfängliche Temperaturverteilung $T_0 = T_0(x)$ als auch die Randbedingungen am Stabende:

$$u(x,0) = T_0(x) \quad \text{für } x \in [0,L] \,,$$

$$u(0,t) = T_e, \qquad u(L,t) = T_r \,,$$

T_e und T_r sind die Temperaturen am Stabende. Weiterhin setzen wir die Verträglichkeitsbedingungen der Rand- und Anfangswerte voraus:

$$T_0(0) = T_e, \qquad T_0(L) = T_r \,.$$

Man hat hier somit ein Anfangs-Randwertproblem vor sich.

Zusammenfassung: Die **Wärmeleitungsgleichung** in einer Raumdimension hat die Form

$$u_t = \kappa \; u_{xx} \tag{4.23}$$

mit der Temperaturleitzahl $\kappa > 0$. In mehreren Raumdimensionen besitzt sie die Form

$$u_t = \kappa \; \Delta u \; . \tag{4.24}$$

Für eine eindeutige Lösung sind Anfangs- und Randwerte notwendig. Das Problem der instationären Wärmeleitung ist somit ein **Anfangs-Randwertproblem (ARWP).**

Die Wärmeleitungsgleichung ist wegen $a = \kappa > 0$ und $c = 0$ nach der Bedingung (4.7) eine parabolische Differenzialgleichung.

Bemerkung: Im physikalisch unsinnigen Fall eines negativen κ existieren für die Probleme (4.23) und (4.24) keine stabilen Lösungen.

Beispiel 4.4 (ARWP für die Wärmeleitungsgleichung, mit MAPLE-Worksheet). Wir betrachten das spezielle Beispiel für die Wärmeleitungsgleichung, wenn die Randwerte identisch Null sind, also das homogene Randwertproblem.

In diesem Fall erhält man die exakte Lösung wiederum über einen Separationsansatz in der Form

$$u(x,t) = X(x) \cdot T(t) \, .$$

Die ausführliche Berechnung der Lösung kann in [71] nachgelesen werden. Die Lösung lautet

$$u(x,t) = \sum_{n=1}^{\infty} b_n e^{-\kappa \left(n\frac{\pi}{L}\right)^2 t} \sin\left(n\frac{\pi}{L}x\right) \, , \tag{4.25}$$

wobei die Koeffizienten sich aus dem anfänglichen Temperaturprofil $T_0(x)$ nach der Formel

$$b_n = \frac{1}{L}\int_0^L T_0(x) \sin\left(n\frac{\pi}{L}x\right) dx$$

bestimmen. Für die spezielle Anfangsverteilung

$$T_0(x) = 2\sin(3\pi x) + 5\sin(8\pi x)$$

ergibt sich die Lösung für die Temperaturleitzahl $\kappa = 1.4$ zu

$$u(x,t) = 2e^{-1.14\cdot 9\pi^2 t}\sin(3\pi x) + 5e^{-1.14\cdot 64\pi^2 t}\sin(8\pi x) \, .$$

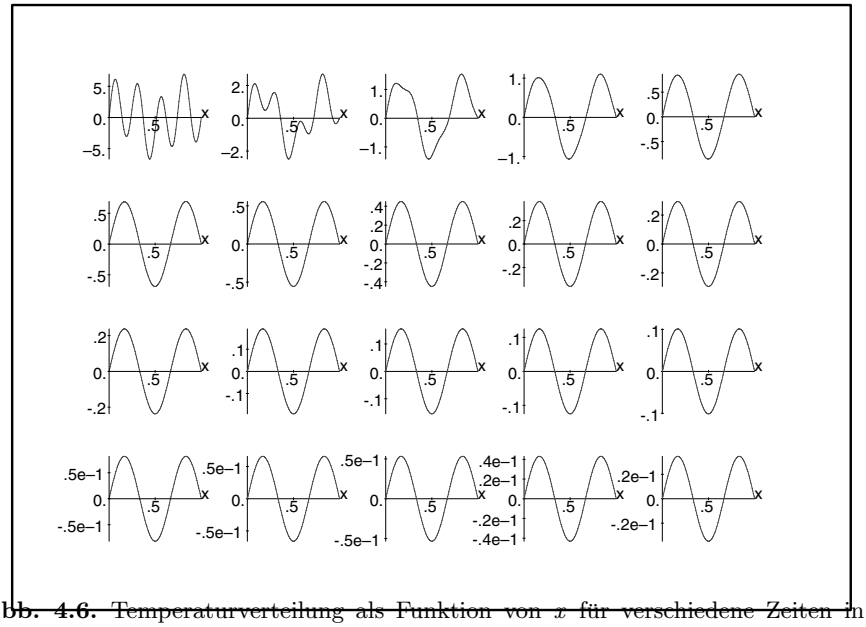


Abb. 4.6. Temperaturverteilung als Funktion von x für verschiedene Zeiten in Beispiel 4.4

Die Lösung strebt für $t \to \infty$ gegen die konstante Lösung $u \equiv 0$. Die zeitliche Entwicklung der Lösung ist in dem Worksheet animiert und in Abbildung 4.6

in einer Folge von Bildern zu verschiedenen Zeiten dargestellt. Man erkennt, dass das Temperaturprofil gegen Null strebt. Die hohen Frequenzen werden dabei schneller gedämpft, so dass gegen Ende der Animation nur noch der Anteil von $\sin(3\pi x)$ sichtbar ist. Die Amplitude wird mit der Zeit immer kleiner. Man beachte, dass sich in Abbildung 4.6 die Skalierung der y-Achse von Bild zu Bild ändert. □

Beispiel 4.5 (ARWP für die Wärmeleitungsgleichung, mit MAPLE-Worksheet). Für die Bewertung numerischer Verfahren schauen wir uns noch eine weitere exakte Lösung an. Für die Anfangsverteilung

$$u(x, t = 0) = -1.5x + 2 + \sin(\pi x)$$

und der Temperatur an den Stabenden

$$u(x = 0, t) = T_l = 2\,, \qquad u(x = 1, t) = T_r = 0.5$$

lautet die exakte Lösung

$$u(x, t) = -1.5x + 2 + e^{-1.14\pi^2 t} \sin(\pi t)\,.$$

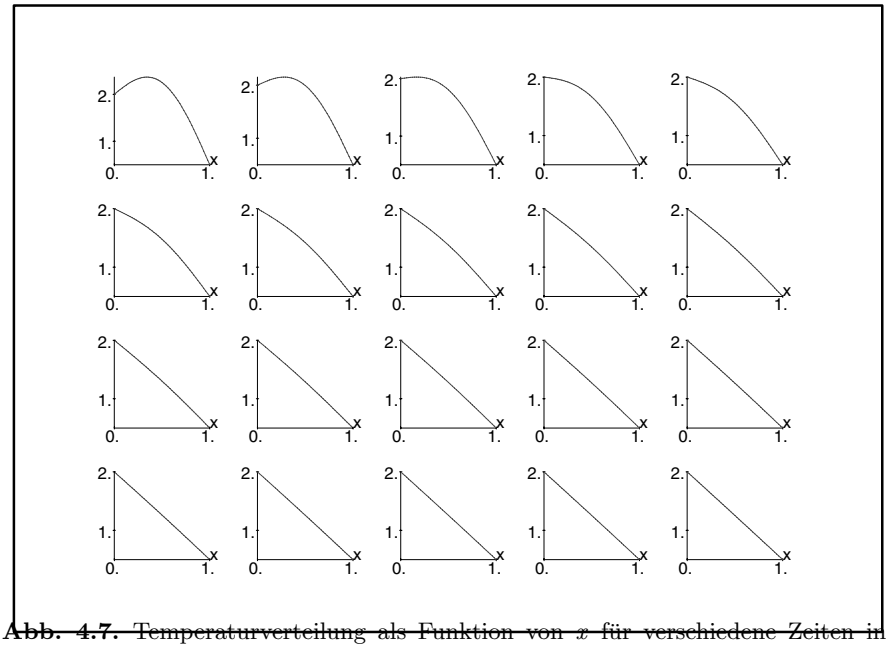


Abb. 4.7. Temperaturverteilung als Funktion von x für verschiedene Zeiten in Beispiel 4.5

Der zeitliche Verlauf dieser Lösung ist in dem Worksheet berechnet und animiert. In einer Folge von Bildern in Abbildung 4.7 erkennt man, dass sich

nach einer gewissen Zeit ein lineares Temperaturprofil zwischen den Randwerten einstellt. In dem Diagramm ist zu beachten, dass die Skalierung der y-Achse unterschiedlich ist. □

4.4 Hyperbolische Differenzialgleichungen 2. Ordnung

Hyperbolische Differenzialgleichungen beschreiben Wellen- und Ausbreitungsphänomene. Dies ist ein instationärer Vorgang, so dass die zweite Variable statt mit y wieder mit t bezeichnet wird.

Beispiel 4.6 (Wellengleichung). Wir betrachten als Modellfall eine an den Enden fest eingespannte, elastische Saite der Länge L, welche in der vertikalen Richtung Schwingungen ausführt. Gesucht ist die, vom Ort x und der Zeit t abhängige, vertikale Auslenkung $u = u(x, t)$.

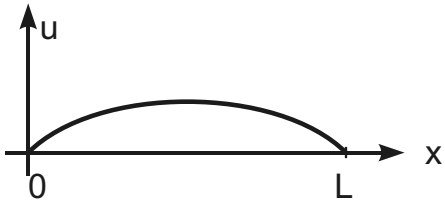


Abb. 4.8. Eingespannte Saite

Abgeleitet wird die Wellengleichung (4.26) aus dem Newtonschen Gesetz $F = m\,a$ (die Beschleunigungskraft $m\,a$ ist gleich der Summe aller angreifenden Kräfte) unter den folgenden Voraussetzungen:

1. Die Spannkraft $F_o = |\mathbf{F}(x)|$ der Saite ist konstant.

2. Es werden nur kleine Auslenkungen für u betrachtet.

Für die Querkraft F_u in Richtung u gilt im Punkte x für kleine Auslenkungen

$$F_u(x) = -F_o \sin\alpha \approx -F_o\alpha \approx -F_o \tan\alpha = -F_o \left(\frac{\partial u}{\partial x}\right)_x .$$

Analog gilt für die Querkraft an der Stelle $x + \Delta x$

$$F_u(x+\Delta x) \approx F_o \left(\frac{\partial u}{\partial x}\right)_{x+\Delta x} \approx F_o \left[\left(\frac{\partial u}{\partial x}\right)_x + \Delta x \left(\frac{\partial^2 u}{\partial x^2}\right)_x\right] ,$$

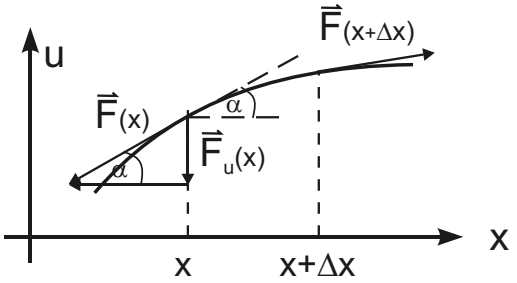


Abb. 4.9. Herleitung der Wellengleichung

wobei bei der letzten Umformulierung nur die ersten zwei Glieder in der Taylor-Entwicklung der Ableitung um den Punkt x berücksichtigt wurden. Die resultierende Querkraft auf das zwischen x und $x + \Delta x$ gelegene Saitenelement lautet dann

$$\Delta F_u = F_u(x + \Delta x) - F_u(x) \approx F_o \Delta x \left(\frac{\partial^2 u}{\partial x^2} \right) .$$

Diese Querkraft beschleunigt das Massenelement $\Delta m = \rho \cdot \Delta x \cdot A$, wenn ρ die Dichte und A die Querschnittsfläche der Saite ist. Nach dem Newtonschen Gesetz gilt somit

$$\Delta m \, u_{tt} = \Delta F_u = F_o \Delta x \, u_{xx} , \tag{4.26}$$

woraus sich mit $c^2 = F_o/(\rho A)$ die Wellengleichung ergibt. □

Die **Wellengleichung in einer Raumdimension** hat somit die Form

$$u_{tt} - c^2 \, u_{xx} = 0 . \tag{4.27}$$

Über eine Dimensionsbetrachtung ergibt sich für c die Dimension einer Geschwindigkeit. Es ist die Ausbreitungsgeschwindigkeit der Welle. Wie beschrieben ist sie das mathematische Modell einer schwingenden Saite, wobei u hier der vertikale Auslenkung bezeichnet. Die Wellengleichung ist wegen $a = 1$ und $b = -c^2$ nach (4.8) eine hyperbolische Gleichung.

Als nächstes betrachten wir Lösungen der Wellengleichung. Ist f eine beliebige 2-mal stetig differenzierbare Funktion einer Variablen $f = f(x)$, dann erhält man durch den Ansatz

$$f = f(x + ct) \quad \text{oder} \quad f = f(x - ct)$$

mit $c > 0$ Lösungen der Wellengleichung, wie man durch direktes Einsetzen bestätigen kann.

Die allgemeine Lösung lautet somit

$$u(x,t) = f_1(x+ct) + f_2(x-ct)$$

mit zwei beliebigen 2-mal stetig differenzierbaren Funktionen f_1 und f_2. Dabei beschreibt $f_1(x+ct)$ eine mit der Geschwindigkeit c in negative x-Richtung laufende Welle und $f_2(x-ct)$ eine mit der Geschwindigkeit c in positive x-Richtung laufende Welle.

Die Lösung wird eindeutig, wenn die Anfangswerte

$$u(x,0) = u_0(x)\,, \quad u_t(x,0) = v_0(x) \qquad \text{für alle } x \in \mathbb{R}$$

mit 2-mal stetig differenzierbaren Funktionen u_0 und v_0 vorgegeben werden. Dabei ist u_0 die Anfangsauslenkung und v_0 die Anfangsgeschwindigkeit. Sind diese Funktionen vorgegeben, lässt sich die exakte Lösung des Anfangswertproblems mit Hilfe der **d'Alembertschen Formel** zu

$$u(x,t) = \frac{1}{2}\big(u_0(x+ct) + u_0(x-ct)\big) + \frac{1}{2c}\int\limits_{x-ct}^{x+ct} v_0(\tau)\,d\tau \tag{4.28}$$

angeben, siehe z.B. [71]. Für den Spezialfall, dass die Saite keine Anfangsgeschwindigkeit besitzt, vereinfacht sich die Formel zu

$$u(x,t) = \frac{1}{2}\big(u_0(x+ct) + u_0(x-ct)\big)\,. \tag{4.29}$$

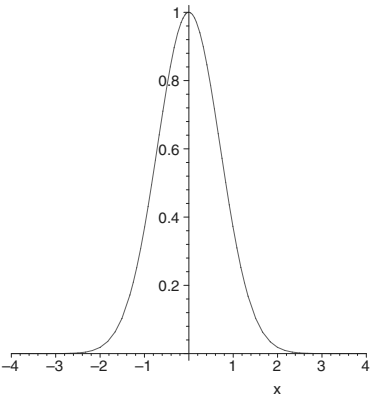
Beispiel 4.7 (AWP für die Wellengleichung, mit MAPLE-Worksheet). Gesucht ist die Lösung der Wellengleichung mit den Anfangswerten

$$u_0(x) = e^{-x^2}\,, \quad v_0(x) = 0 \qquad \text{für alle } x \in \mathbb{R}. \tag{4.30}$$

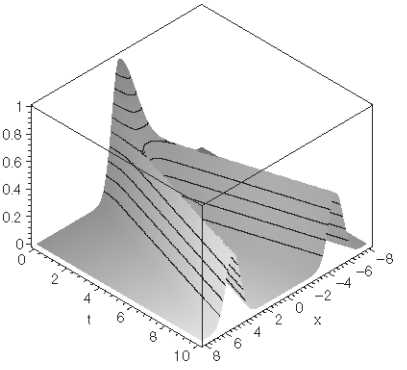
Die Anfangsdaten beschreiben eine Auslenkung in Form einer Gauß- oder Glockenkurve (siehe Abbildung 4.7a). Nach der d'Alembertschen Formel (4.29) ergibt sich als Lösung

$$u(x,t) = \frac{1}{2}e^{-(ct+x)^2} + \frac{1}{2}e^{-(ct-x)^2}$$

für alle x und t. Für die Ausbreitungsgeschwindigkeit $c = 0.5$ ist die Lösung als Funktion von x und t in Abbildung 4.7b gezeichnet. Sie wird in dem MAPLE-Worksheet berechnet und gezeichnet. Die anfängliche Auslenkung erzeugt zwei Wellen, welche mit der halben Amplitude nach links und rechts laufen. □



(a) Anfangsauslenkung aus Beispiel 4.7



(b) Lösung des Anfangswertproblems (4.30) für die Wellengleichung

Die Wellengleichung lässt sich auch in einer anderen Form schreiben. Führen wir die neuen Variablen

$$q = c\,u_x, \qquad p = u_t$$

ein, so können wir die Wellengleichung wegen

$$p_t = u_{tt} = c^2 u_{xx} = c\,q_x$$

$$q_t = c\;u_{xt} = c\;u_{tx} = c\,p_x$$

auch als ein **System von Differenzialgleichungen erster Ordnung** in der Form

$$p_t - c\,q_x = 0, \qquad q_t - c\,p_x = 0 \tag{4.31}$$

schreiben. Dies ist ein System von sogenannten **Evolutionsgleichungen**, welches sich auch in der Form einer Vektorgleichung

$$\mathbf{w}_t + \mathbf{A}\mathbf{w}_x = 0 \tag{4.32}$$

mit dem Vektor $\mathbf{w}$ und der Matrix $\mathbf{A}$,

$$\mathbf{w} = \begin{pmatrix} p \\ q \end{pmatrix}, \qquad \mathbf{A} = \begin{pmatrix} 0 & -c \\ -c & 0 \end{pmatrix}$$

umschreiben lässt. Die Ableitungen $\mathbf{w}_t$, $\mathbf{w}_x$ sind dabei komponentenweise zu verstehen:

$$\mathbf{w}_x = \begin{pmatrix} p_x \\ q_x \end{pmatrix}, \qquad \mathbf{w}_t = \begin{pmatrix} p_t \\ q_t \end{pmatrix}.$$

Die 2×2-Matrix $\mathbf{A}$ besitzt die reellen Eigenwerte

$$\lambda_1 = -c\,, \qquad \lambda_2 = c\,,$$

welche den verschiedenen Wellengeschwindigkeiten, einer Welle nach links mit der Ausbreitungsgeschwindigkeit $-c$ und einer nach rechts mit der Geschwindigkeit c entsprechen. Evolutionsgleichungen dieser Art werden wir uns im folgenden Abschnitt noch etwas genauer anschauen.

In **mehreren Raumdimensionen** lautet die Wellengleichung

$$u_{tt} - c^2 \Delta u = 0 \, . \tag{4.33}$$

Die Raumableitung 2. Ordnung ist durch den Laplace-Operator ersetzt. Wie schon im räumlich eindimensionalen Fall lässt sich die mehrdimensionale Wellengleichung auch als Evolutionsgleichung schreiben. Die Transformation hat hier die Form

$$p = u_t \, , \qquad \mathbf{v} = c \, \nabla u \tag{4.34}$$

und das System von Evolutionsgleichungen erster Ordnung ergibt sich zu

$$p_t - c \nabla \cdot \mathbf{v} = 0 \tag{4.35}$$

$$\mathbf{v}_t - c \nabla p = 0 \, . \tag{4.36}$$

Dabei bezeichnet $\nabla \cdot$ die Divergenz eines Vektors und ∇ den Gradienten einer skalaren Funktion. Das Differenzialgleichungssystems 1. Ordnung besteht somit im drei-dimensionalen Fall aus vier Gleichungen, eine für die skalare Funktion p und drei für den Vektor $\mathbf{v}$.

Zusammenfassung: Die **Wellengleichung in einer Raumdimension** hat die Form

$$u_{tt} - c^2 \, u_{xx} = 0 \, , \tag{4.37}$$

wobei $c > 0$ die Wellengeschwindigkeit ist. Für die Existenz einer eindeutigen Lösung müssen Anfangswerte auf der ganzen reellen Achse vorgegeben werden. Das wohldefinierte Problem für die Wellengleichung ist somit ein **Anfangswertproblem (AWP)**. Stellt sich das physikalische Problem als Anfangs-Randwertproblem (ARWP), so treten am Rand **Verträglichkeitsbedingungen** auf.
Die **Wellengleichung in mehreren Raumdimensionen** lautet

$$u_{tt} - c^2 \Delta u = 0 \, . \tag{4.38}$$

4.5 Evolutionsgleichungen

Ein System von m Differenzialgleichungen erster Ordnung in der Form

$$\mathbf{u}_t + \mathbf{A}\mathbf{u}_x = 0 \tag{4.39}$$

bezeichnet man als ein **System von Evolutionsgleichungen**. Dabei ist $\mathbf{u}$ ein Vektor mit m Komponenten und $\mathbf{A}$ eine reelle $m \times m$-Matrix. Ist $\mathbf{A}$ eine konstante Matrix, d.h. alle Koeffizienten sind reelle Zahlen, dann ist (4.39) ein **lineares System von Evolutionsgleichungen mit konstanten Koeffizienten**. Ist $\mathbf{A}$ abhängig von x und t, so ist (4.39) ein **lineares System**. Hängt $\mathbf{A}$ von $\mathbf{u}$ ab, dann nennt man es ein **quasilineares System**.

Es gibt die folgende Aussage bezüglich des Typs der Evolutionsgleichungen:

Das System (4.39) von Evolutionsgleichungen heißt **hyperbolisch**, wenn alle Eigenwerte von $\mathbf{A}$ reell sind. Die Eigenwerte von $\mathbf{A}$ entsprechen den verschiedenen Wellengeschwindigkeiten.

Die einfachste hyperbolische Differenzialgleichung ist demnach nicht die Wellengleichung (4.32), sondern für $m = 1$ die lineare Evolutionsgleichung

$$u_t + a u_x = 0. \tag{4.40}$$

Diese nennt man **lineare Transportgleichung** oder auch **lineare Konvektionsgleichung** oder **Advektionsgleichung**.

Gilt $a \in \mathbb{R}$, dann ist (4.40) eine Transportgleichung mit einem konstanten Koeffizienten. In diesem Falle lässt sich diese Gleichung in einfacher Weise exakt lösen. Dazu betrachten wir die Geradenschar

$$C: \quad x = x(t) \quad \text{mit} \quad \frac{dx(t)}{dt} = a\ . \tag{4.41}$$

Jede dieser Geraden besitzt die Steigung $1/a$ in der (x,t)-Ebene. Die Zeitvariable t wird hierbei als Kurvenparameter benutzt. Für die Änderung einer Funktion u auf einer Geraden der Schar C gilt

$$\frac{d}{dt}u(x(t),t) = u_x \frac{dx}{dt} + u_t \cdot 1 = u_t + a u_x\ . \tag{4.42}$$

Jede Lösung u der linearen Transportgleichung (4.40) ändert sich somit entlang einer Geraden C nicht. Die Geraden mit der Eigenschaft (4.41) nennt man die **Charakteristiken** der Transportgleichung (4.40). Gewissermaßen stellt $u_t + au_x$ eine Richtungsableitung in Richtung der Charakteristiken dar. Die Transportgleichung (4.40) besagt, dass sich u entlang dieser Bahn nicht ändert.

Ist eine Lösung der Transportgleichung (4.40) mit vorgegebenen Anfangswerten

$$u(x,0) = q(x) \quad \text{für alle} \quad x \in \mathbb{R}$$

gesucht, so wird man die Lösung folgendermaßen erhalten: An einem beliebigen Punkt (x_0, t_0) trägt man die Gerade mit der Steigung $1/a$ an, d.h. die Charakteristik durch diesen Punkt. Verfolgt man diese Gerade bis zum Schnittpunkt Q mit der x-Achse, so ist der Wert der Lösung am Punkt P gleich dem Wert am Punkt Q, da sich die Lösung entlang der Charakteristik nicht ändert. Ganz allgemein gilt für die Lösung des Anfangswertproblems

$$u(x,t) = q(x - at) \quad \text{für alle} \quad (x,t) \in \mathbb{R} \times \mathbb{R}^+ , \tag{4.43}$$

wie es in Abbildung 4.10 skizziert ist. Die lineare Transportgleichung beschreibt den Transport der Anfangswerte mit der Geschwindigkeit a.

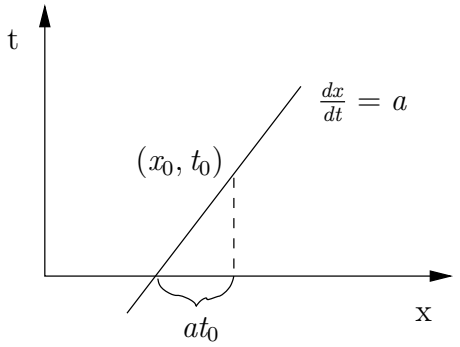


Abb. 4.10. Charakteristik im Punkt (x_0, t_0)

Hängt a von x und oder t ab, dann sind die Charakteristiken im Allgemeinen keine Geraden, sondern Kurven in der (x,t)-Ebene. Die Lösung ist wiederum konstant entlang dieser Charakteristiken.

Beispiel 4.8 (AWP für die lineare Transportgleichung, mit MAPLE-Worksheet). Wir betrachten für die lineare Transportgleichung (4.40) Anfangswerte in der Form einer Gaußschen Glockenkurve

$$u(x,0) = q(x) = e^{-x^2} \quad \text{für alle } x \in \mathbb{R} .$$

Mit der Ausbreitungsgeschwindigkeit $a = 0.5$ ergibt sich die exakte Lösung zu

$$u(x,t) = q(x - 0.5t) = e^{-(x-0.5t)^2} ,$$

welche sich aus dem Transport der Anfangswerte mit der Geschwindigkeit $a = 0.5$ ergibt.

Dieses Beispiel ist ein guter Test für die numerischen Lösungsverfahren. Man wird dazu ein endliches Intervall wählen und periodische Randbedingungen vorschreiben. Die Glockenkurve wird mit der Geschwindigkeit $a = 0.5$ nach rechts transportiert. Nach einer gewissen Zeit kommt sie an dem rechten Rand des Rechengebiets an, läuft dort hinaus und durch die periodischen Randbedingungen am linken Rand entsprechend wieder hinein. Ist sie die ganze Intervalllänge durchgelaufen, dann stimmt die Lösung wieder mit den Anfangsdaten überein. Für einen Vergleich mit Resultaten einer numerischen Rechnung ist dies natürlich sehr günstig. In dem MAPLE-Worksheet wird die exakte Lösung des periodischen Problems berechnet und gezeichnet. In Abbildung 4.11 ist diese Lösung zu verschiedenen Zeiten gezeichnet. Bei dem Test der numerischen Methoden werden wir dieses Problem wieder aufgreifen. □

Das Finden von Lösungen mit Hilfe der Charakteristikentheorie lässt sich auf **Systeme von hyperbolischen Evolutionsgleichungen**

$$\mathbf{u}_t + \mathbf{A}\mathbf{u}_x = 0$$

erweitern, wenn die Matrix A durch eine Ähnlichkeitstransformation diagonalisierbar ist. Die Bedingung dafür ist, dass ein vollständiger Satz von Eigenvektoren existiert. Man nennt das System in diesem Falle auch **streng hyperbolisch**. Eine hinreichende Bedingung ist, dass die reellen Eigenwerte von A alle voneinander verschieden sind. Die Ähnlichkeitstransformation sieht folgendermaßen aus:

$$\mathbf{\Lambda} := \mathbf{R}^{-1}\mathbf{A}\mathbf{R} = \begin{pmatrix} a_1 & & 0 \\ & a_2 & \\ & & \ddots \\ 0 & & a_m \end{pmatrix} , \tag{4.44}$$

wobei die $a_1, a_2, \ldots, a_m$ die Eigenwerte der Matrix $\mathbf{A}$ sind. Die $m \times m$-Matrix $\mathbf{R}$ enthält als Spalten die zugehörigen Eigenvektoren, $\mathbf{R}^{-1}$ ist die inverse Matrix zu $\mathbf{R}$.

Die Diagonalisierung gibt die Möglichkeit, das System in einzelne Gleichungen zu zerlegen. Dazu multiplizieren wir das System von Evolutionsgleichungen mit der Matrix $\mathbf{R}^{-1}$ von links und erhalten

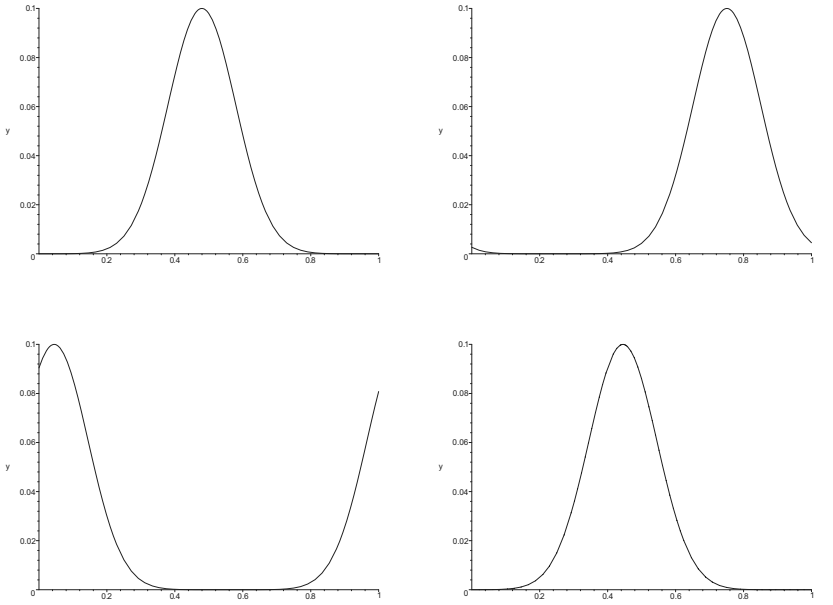


Abb. 4.11. Lösung des Anfangswertproblems für eine lineare Transportgleichung in Beispiel 4.8 mit periodischen Randbedingungen zu verschiedenen Zeitpunkten

$$\mathbf{R}^{-1}\mathbf{u}_t + \mathbf{R}^{-1}\mathbf{A}\mathbf{u}_x = 0 \,.$$

Betrachten wir den linearen Fall mit konstanten Koeffizienten, dann können wir die konstante Matrix $\mathbf{R}^{-1}$ mit unter die Ableitung ziehen:

$$(\mathbf{R}^{-1}\mathbf{u})_t + \mathbf{R}^{-1}\mathbf{A}\mathbf{R}(\mathbf{R}^{-1}\mathbf{u})_x = 0 \,.$$

Im zweiten Term haben wir dabei die Matrix $\mathbf{R}\mathbf{R}^{-1}$ eingeschoben. Diese Matrix ist nach Definition der Inversen gerade die Einheitsmatrix und ändert den Wert des Ausdrucks nicht. Durch Ersetzen von $\mathbf{R}^{-1}\mathbf{A}\mathbf{R}$ mit $\mathbf{\Lambda}$ erhalten wir daraus

$$(\mathbf{R}^{-1}\mathbf{u})_t + \mathbf{\Lambda}(\mathbf{R}^{-1}\mathbf{u})_x = 0 \,.$$

Es wird jetzt die neue abhängige Variable

$$\mathbf{w} = \mathbf{R}^{-1}\mathbf{u} \tag{4.45}$$

eingeführt, deren zeitliche Entwicklung durch das System von Evolutionsgleichungen

$$\mathbf{w}_t + \mathbf{\Lambda}\,\mathbf{w}_x = 0 \tag{4.46}$$

bestimmt ist.

Dies sollten wir uns noch etwas genauer ansehen, um den Sinn der Procedere zu erkennen. Die Matrix dieser Evolutionsgleichungen ist eine Diagonalmatrix. Dies bedeutet, dass die einzelnen Evolutionsgleichungen untereinander nicht mehr gekoppelt sind. Einzeln ausgeschrieben lauten sie

$$(w_i)_t + a_i(w_i)_x = 0 \,, \qquad i = 1, 2, \ldots, m \,,$$

wobei w_i die i-te Komponente des Vektors $\mathbf{w}$ bezeichnet. Das gekoppelte Problem ist also zerlegt in m einzelne skalare Probleme, welche wir einzeln nach dem vorne vorgestellten Verfahren lösen können.

Die Lösungen der einzelnen Probleme fassen wir wieder zu dem Vektor $\mathbf{w} = \mathbf{w}(x,t)$ als Lösung des Systems (4.46) zusammen. Die Lösung von unserem ursprünglichen System von Evolutionsgleichungen ergibt sich dann aus der Rücktransformation

$$\mathbf{u} = \mathbf{R}\mathbf{w} \,.$$

Den Variablenvektor $\mathbf{w}$ nennt man die **charakteristischen Variablen** und die Gleichung (4.46) die **charakteristische Normalform** des Systems (4.44).

Zusammenfassung: Ist A diagonalisierbar, dann ergibt sich die Lösung von

$$\mathbf{u}_t + \mathbf{A}\,\mathbf{u}_x = 0$$

mit Hilfe der Charakteristikentheorie. Durch die Einführung der **charakteristischen Variablen** $\mathbf{w} = \mathbf{R}^{-1}\mathbf{u}$ ergibt sich die **charakteristische Normalform**

$$\mathbf{w}_t + \mathbf{\Lambda}\,\mathbf{w}_x = 0$$

bestehend aus nicht gekoppelten einzelnen Gleichungen.

Bemerkung: Mit dieser Anwendung der Charakteristikentheorie werden wir später numerische Verfahren für eine einzelne Transportgleichung auf Systeme übertragen.

Beispiel 4.9. Wir wollen die Charakteristikentheorie auf die Wellengleichung (4.32) anwenden, um die exakte Lösung zu ermitteln. Die 2×2-Matrix $\mathbf{A}$ besitzt die reellen Eigenwerte

$$\lambda_1 = -c \,, \qquad \lambda_2 = c \,.$$

Sie entsprechen den verschiedenen Wellengeschwindigkeiten, es existiert somit eine Welle nach links mit der Ausbreitungsgeschwindigkeit $-c$ und eine nach

rechts mit der Geschwindigkeit c. Die Eigenvektoren sind schnell berechnet. Als Spalten der Matrix $\mathbf{R}$ geschrieben lauten sie

$$\mathbf{R} = \begin{pmatrix} 1 & 1 \\ 1 & -1 \end{pmatrix} \quad \text{und} \quad \mathbf{R}^{-1} = \frac{1}{2} \begin{pmatrix} 1 & 1 \\ 1 & -1 \end{pmatrix} .$$

Die charakteristischen Variablen lauten somit im Falle der Wellengleichung

$$\mathbf{w} = \mathbf{R}^{-1}\mathbf{u} = \frac{1}{2} \begin{pmatrix} p+q \\ p-q \end{pmatrix}$$

und die charakteristische Form ergibt sich zu

$$\begin{aligned} (p+q)_t - c(p+q)_x &= 0 , \\ (p-q)_t + c(p-q)_x &= 0 . \end{aligned}$$

Das Anfangswertproblem lässt sich somit mit der Charakteristikentheorie folgendermaßen lösen: Die Funktion $p+q$ ist entlang der Charakteristik mit der Steigung $-\frac{1}{c}$ konstant. Wir können in einem beliebigen Punkt (x_1,t_1) der (x,t)-Ebene diese Gerade anzeichnen und auf die Anfangsdaten zurückverfolgen. Den Fußpunkt nennen wir Q_-. In Abbildung 4.12 ist dies entsprechend eingetragen. Bezeichnen wir die Anfangsdaten der Funktion $p+q$ mit einem Index „0“, so gilt

$$(p+q)(x_1,t_1) = (p+q)\,Q_- = (p_0+q_0)(x_1+ct_1) .$$

Ganz analog erhalten wir entlang der Charakteristik mit der Steigung $\frac{1}{c}$ die Beziehung

$$(p-q)(x_1,t_1) = (p-q)\,Q_+ = (p_0-q_0)(x_1-ct_1) .$$

Dies sind zwei Gleichungen für die zwei Unbekannten $p(x_1,t_1)$ und $q(x_1,t_1)$ mit der Lösung

$$\begin{aligned} p(x_1,t_1) &= \frac{1}{2}\big(p_0(x_1+ct_1) + q_0(x_1+ct_1) + p_0(x_1-ct_1) - q_0(x_1-ct_1)\big) \\ q(x_1,t_1) &= \frac{1}{2}\big(p_0(x_1+ct_1) + q_0(x_1+ct_1) - p_0(x_1-ct_1) + q_0(x_1-ct_1)\big) \end{aligned}$$

für beliebige $(x_1,t_1) \in \mathbb{R}\times\mathbb{R}^+$. Die Lösung der Wellengleichung ist die Summe einer Funktion von $x-ct$ und $x+ct$, wie dies im Ansatz (4.28) und (4.29) benutzt wurde. □

Der **quasilineare skalare Fall** mit

$$u_t + a(u)u_x = 0 \tag{4.47}$$

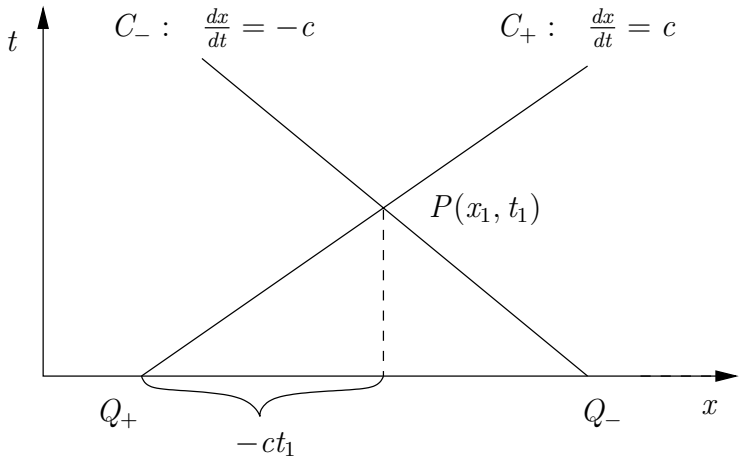


Abb. 4.12. Lösung der Wellengleichung aus Beispiel 4.9

bringt neue Schwierigkeiten. Die Charakteristikentheorie lässt sich zunächst ganz analog zum linearen Fall ausführen. So sind die Charakteristiken wieder als die Kurven

$$C : \quad x = x(t) \quad \text{mit} \quad \frac{dx(t)}{dt} = a(u) \tag{4.48}$$

definiert. Die **Steigungen der Charakteristiken** sind dabei allerdings **lösungsabhängig**. Analog zum linearen Fall ergibt sich die Konstanz der Lösung entlang den Charakteristiken. Wenn die Lösung u aber entlang einer Charakteristik konstant ist, dann ist dies natürlich auch $a(u)$ und die Steigung der Charakteristik ändert sich entlang der Charakteristik nicht. Eine Charakteristik ist somit auch im nichtlinearen skalaren Fall eine Gerade.

Für die exakte Lösung eines Anfangswertproblems können wir damit wie schon im linearen Fall nach Abbildung 4.10 die Formel

$$u(x,t) = q(x - a(u(x,t))t) \tag{4.49}$$

schreiben. Es gibt jetzt aber einen großen Unterschied. Im Argument auf der rechten Seite der Anfangswerte steht $x - a(u)t$. Ist die Gleichung (4.47) tatsächlich nichtlinear und $a'(u) \neq 0$, dann kommt die Lösung u auch in diesem Argument vor und nicht nur auf der linken Seite. Die Gleichung (4.49) ist somit eine implizite Gleichung. Die Anfangsdaten $q = q(x)$ seien stetig differenzierbar, dann können wir (4.49) nach t oder x differenzieren. Differenzieren wir die implizite Gleichung nach x, dann erhalten wir unter Benutzung der Kettenregel

$$u_x = q'(1 - a'(u)u_x t)$$

und weiter nach u_x aufgelöst:

$$\begin{aligned} u_x + q' a'(u) t u_x &= q' \\ \Rightarrow u_x &= \frac{q'}{1 + q' a'(u) t} \, . \end{aligned} \tag{4.50}$$

Der Nenner auf der rechten Seite kann Null werden. Ist etwa $a(u)$ monoton wachsend ($a'(u) > 0$) und sind die Anfangswerte monoton fallend ($q' < 0$), dann gibt es eine Zeit t_s, bei welcher der Nenner Null wird, die Ableitung u_x somit unendlich wird und keine stetige Lösung existiert.

Bemerkung: Der Satz über implizite Funktionen garantiert, dass eine Lösung der impliziten Gleichung (4.49) zumindest für kurze Zeit existiert. Diese stetig differenzierbare Lösung kann aber zusammenbrechen, indem sich eine Unstetigkeit entwickelt. Wir sehen in (4.50) auch, dass dies ein echtes nichtlineares Phänomen ist. Im linearen Fall gilt $a'(u) = 0$ und der Nenner bleibt für alle Zeiten t von Null verschieden.

Das Problem bleibt ein reines **Anfangswertproblem**. Die Einschränkung auf ein endliches Gebiet zeigt wiederum den Charakter der hyperbolischen Gleichungen. Überall dort, wo Charakteristiken das Rechengebiet verlassen, tritt eine Verträglichkeitsbedingung auf: Überall dort, wo eine Charakteristik in das Rechengebiet hinein läuft, muss ein Randwert vorgegeben werden.

Die allgemeine Form einer **Evolutionsgleichung in zwei Raumdimensionen** lautet

$$\mathbf{u}_t + \mathbf{A}\mathbf{u}_x + \mathbf{B}\mathbf{u}_y = 0 \, , \tag{4.51}$$

wobei $\mathbf{u}$ den Lösungsvektor mit m Komponenten bezeichnet und $\mathbf{A}$ und $\mathbf{B}$ $m \times m$-Matrizen sind. Die Evolutionsgleichung ist hyperbolisch, wenn $\alpha\mathbf{A} + (1 - \alpha)\mathbf{B}$ für jedes $\alpha \in \mathbb{R}$ nur reelle Eigenwerte besitzt.

4.6 Erhaltungsgleichungen

Der Zusammenbruch der stetig differenzierbaren Lösung einer nichtlinearen hyperbolischen Gleichung und das Auftreten einer Unstetigkeit kann physikalische Relevanz haben. Eine Klasse von Differenzialgleichungen, bei deren Lösung diese Phänomene auftreten können, sind Erhaltungsgleichungen. Die

Erhaltungsgleichungen werden aus integralen Erhaltungssätzen der Physik abgeleitet. Beispiele hierfür sind Massen-, Impuls- oder Energieerhaltung.

Wir betrachten einen solchen Erhaltungssatz genauer, z.B. die Erhaltung der Masse. Dieser Satz bedeutet, dass Masse weder verschwinden noch plötzlich entstehen kann. In einer Strömung bedeutet dies nicht, dass in einem beliebigen Volumen des Strömungsgebiets sich die Masse nicht ändern kann. Es wird Masse in den betrachteten Bereich hinein- und heraus strömen, d.h. die Masse in dem betrachteten Gebiet wird von der Strömung durch den Rand des Gebietes abhängen. Fließt mehr rein als raus, wird die Masse in diesem Gebiet größer. Mathematisch können wir dies folgendermaßen aufschreiben:

$$\frac{d}{dt}\int\limits_{\Omega} \rho dV = -\int\limits_{\partial\Omega} \mathbf{f}_m \cdot \mathbf{n} dS \; . \tag{4.52}$$

Dabei ist Ω ein beliebiges Teilgebiet des Strömungsgebiets mit Rand $\partial\Omega$. Die Funktion ρ ist die **Dichte** oder besser die **Massendichte** des Fluids, $\mathbf{f}_m$ ist der Massenfluss und $\mathbf{n}$ die nach außen gerichteten Einheitsnormale auf $\partial\Omega$.

Der Gleichung (4.52) liegt die Kontinuumsannahme zugrunde: Die Masse in irgendeinem Bereich lässt sich durch Integration einer im ganzen Raum kontinuierlichen Funktion, der Massendichte, berechnen. Die zeitliche Änderung der Masse in Ω ist auf einen Fluss durch den Rand zurückzuführen: dem Integral über den gesamten Rand $\partial\Omega$ des Massenflusses $\mathbf{f}_m$ in Richtung der Einheitsnormalen $\mathbf{n}$ (siehe Abbildung 4.13). Stillschweigend ist dabei für Ω vorausgesetzt, dass auf dem Rand in jedem Punkt die Normale existiert.

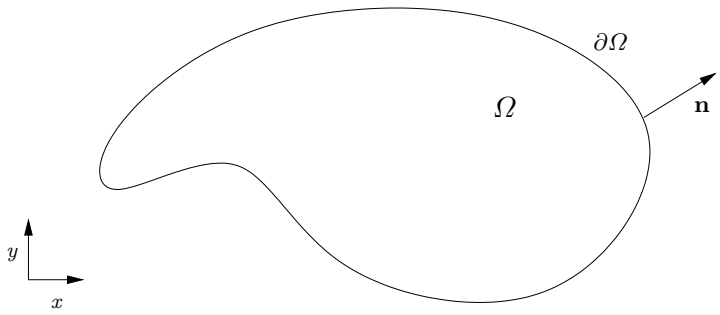


Abb. 4.13. Erhaltungssatz

Nehmen wir an, dass die Dichte ρ und der Massenfluss $\mathbf{f}_m$ stetig differenzierbare Funktionen sind, dann kann man auf der linken Seite in (4.52) Integration und Differenziation vertauschen. Die zeitliche Ableitung kann man als partielle Ableitung unter das Integral ziehen. Auf der linken Seite wird der

Gaußsche Satz angewandt: Ist $\mathbf{f}$ eine stetig differenzierbare vektorwertige Funktion, dann gilt

$$\int_{\Omega} \nabla \cdot \mathbf{f}\, dV = \int_{\partial\Omega} \mathbf{f} \cdot \mathbf{n} dS \; . \tag{4.53}$$

Die rechte Seite von (4.52) können wir somit mit dem Satz von Gauß als Volumenintegral umschreiben und erhalten insgesamt

$$\int_{\Omega} (\rho_t + \nabla \cdot \mathbf{f}_m)\, dV = 0 \; . \tag{4.54}$$

Für die nächste Folgerung ist die wesentliche Bedingung, dass (4.54) für jedes beliebige Teilgebiet Ω gilt. Nur dann kann man nach dem **Lemma der Variationsrechnung**(siehe z.B. [44]) daraus folgern, dass der Integrand in jedem Punkt Null sein muss:

$$\rho_t + \nabla \cdot \mathbf{f}_m = 0 \; .$$

Damit hat man eine lokale Formulierung der Massenerhaltung in Form einer **Differenzialgleichung** gefunden. Da der lokale Massenstrom eines bewegten Fluids gleich dem Produkt aus Dichte und Geschwindigkeit ist, bekommt die Gleichung der Massenerhaltung die Form

$$\rho_t + \nabla \cdot (\rho \mathbf{v}) = 0 \; ,$$

wobei $\mathbf{v}$ den Geschwindigkeitsvektor bezeichnet.

Allgemein ist ein physikalisches Erhaltungsgesetz eine Integralgleichung der Form

$$\frac{d}{dt} \int_{\Omega} u\, dV = - \int_{\partial\Omega} \mathbf{f} \cdot \mathbf{n}\, dS + \int_{\Omega} g\, dV \; . \tag{4.55}$$

Dabei ist u die **Dichte der Erhaltungsgröße**. Das erste Integral auf der rechten Seite enthält die **Oberflächeneffekte**, den Fluss. Auf der rechten Seite steht das **Volumenintegral** einer **Quelle** g. Führen wir das Procedere wie oben durch, ergibt sich die allgemeine **lokale Formulierung der Erhaltungsgleichung als Differenzialgleichung** (4.55):

$$u_t + \nabla \cdot \mathbf{f} = g \; . \tag{4.56}$$

Meist ist ein physikalischer Vorgang erst durch mehrere Erhaltungsgesetze bestimmt.

Wir nehmen im Folgenden zunächst an, dass eine einzelne Erhaltungsgleichung vorliegt und $f = f(u)$ eine skalare Funktion von u ist. Ändert sich die Lösung nur in eine Raumrichtung, z. B. in x-Richtung, und ist in die

anderen Richtungen konstant, reduziert sich das Problem auf eine Raumdimension. In der Divergenz des Flusses bleibt nur die x-Ableitung übrig und die Erhaltungsgleichung lautet

$$u_t + f(u)_x = 0 \; . \tag{4.57}$$

Dabei wird angenommen, dass der Fluss $f(u)$ stetig differenzierbar ist. Die Ableitung von $f(u)$ nach x kann dann unter Benutzung der Kettenregel ausgeführt werden und man erhält die **quasilineare Form** der Erhaltungsgleichung:

$$u_t + a(u)\, u_x = 0 \tag{4.58}$$

mit

$$a(u) = \frac{df(u)}{du} \; .$$

Den Fall einer quasilinearen Evolutionsgleichung haben wir in Abschnitt 4.5 schon erläutert. Die Charakeristikentheorie liefert die Lösungsformel (4.49), welche die Lösung nur implizit angibt, falls $f(u)$ nichtlinear ist. Im linearen Fall kann man an einem beliebigen Punkt der (x,t)-Ebene die Charakteristik einzeichnen und diese auf die Anfangsdaten zurückverfolgen. Im nichtlinearen Fall weiß man hingegen die Steigung der Charakteristik nur, wenn die Lösung bekannt ist. Dies bedeutet für ein Anfangswertproblem jedoch, dass man ausgehend von den Anfangsdaten die Charakteristiken zeichnen und verfolgen kann.

Beispiel 4.10. Wir betrachten die einfachste nichtlineare Erhaltungsgleichung:

$$u_t + (\frac{1}{2}u^2)_x = 0, \tag{4.59}$$

welche meist **Burgers-Gleichung** genannt wird. Als Anfangswerte seien vorgegeben

$$u(x,0) = q(x) = \begin{cases} 1 & \text{für } x < 0 \; , \\ 1 - x & \text{für } 0 < x < 1 \; , \\ 0 & \text{für } x > 1 \; . \end{cases} \tag{4.60}$$

Das Bild der Charakteristiken kann man ausgehend von den Anfangswerten zeichnen. Wegen $a(u) = u$ im Falle der Burgers-Gleichung sind die Charakteristiken für $x < 0$ Parallelen zur Winkelhalbierenden. Für $x > 1$ sind sie Parallelen zur t-Achse, da dort $u = 0$ gilt. Der stetige Übergang in den Anfangswerten (4.60) bewirkt, dass die Steigung der Charakteristiken größer werden und sich aufstellen.

Der Verlauf der Anfangswerte und die Charakteristiken in der (x,t)-Ebene sind in Abbildung 4.14 im oberen Bild gezeichnet. Das untere zeigt den Verlauf der Charakteristiken in der (x,t)-Ebene. Man sieht hier zwischen $x = 0$

und $x = 1$, wie sich die Strecke des Übergangs zwischen $u = 1$ und $u = 0$ mit der Zeit verkürzt und gegen Null strebt. Bei $t = 1$ passiert etwas seltsames: Es kommt zum Schneiden der Charakteristiken. Die Charakteristiken von links bringen die Information $u = 1$, die Charakteristiken auf der rechten Seite für $x > 1$ transportieren die Information $u = 0$. Es kommt somit zu einem Widerspruch und zum Zusammenbruch einer stetigen Lösung. Man kann im Bild der Charakteristiken ablesen, wie sich die Lösung für $t < 1$ verhält. Der Bereich des Übergangs von $u = 1$ nach $u = 0$ wird mit wachsender Zeit immer kleiner. Dies bedeutet, dass sich die Lösung aufstellt bis ein unstetiger Übergang bei $t = 1$ vorliegt. Für das betrachtete Anfangswertproblem existiert somit keine globale stetige Lösung.

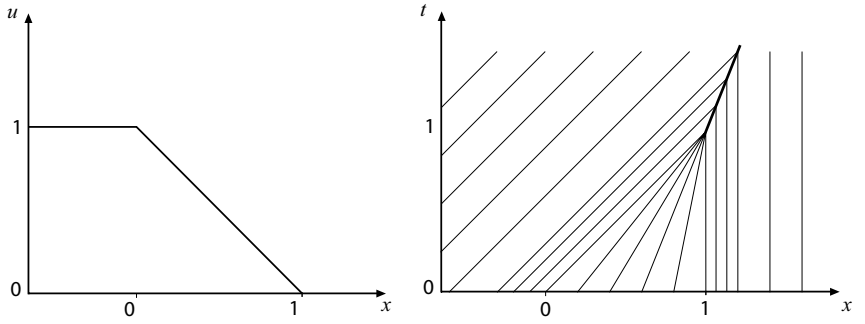


Abb. 4.14. Links: Anfangswerte (4.60), rechts: Bild der Charakteristiken in der (x, t)-Ebene

In dem Bereich zwischen $t < x < 1$ folgt für die Lösung nach Gleichung (4.49)

$$u(x, t) = q(x - ut) = 1 - (x - ut) \,,$$

welche sich nach u auflösen lässt. Man erhält

$$u(x, t) = \frac{1 - x}{1 - t} \,.$$

Die Charakteristiken, welche im Intervall $[0, 1]$ starten, laufen alle im Punkt $P = (1, 1)$ zusammen. Wir haben bei den Anfangsdaten der Einfachheit halber einen linearen Übergang von $u = 1$ nach $u = 0$ gewählt. Das Problem mit dem Schneiden der Charakteristiken ist von der Stetigkeit der Anfangsdaten unabhängig. Auch bei einem stetig differenzierbaren Übergang wird dieses Phänomen auftreten. □

Mit der Nichtexistenz einer stetig differenzierbaren Lösung für alle Zeiten könnte man sich gut abfinden, wenn diese Unstetigkeit keinen physikalischen

Hintergrund hätte. Diese Unstetigkeit ist eine sogenannte **Stoßwelle** oder **Verdichtungsstoß**. Der Namen ist abgeleitet von einem Phänomen in einer Gasströmung, welches in Überschallströmungen auftreten kann. Es ist als Überschallknall beim Überschallflug von Düsenflugzeugen wohlbekannt. Die Stoßwellen sind Änderungen der physikalischen Größen im mikroskopischen Bereich, die sich in den makroskopischen Erhaltungsgleichungen als Unstetigkeiten in den Lösungen zeigen.

Den Schritt der Ableitung der lokalen Formulierung als Differenzialgleichung dürfen wir an diesen Stellen nicht ausführen, weil die Stetigkeitseigenschaften des Gaußschen Satzes nicht erfüllt sind. Möchte man wissen, wie sich eine solche Unstetigkeit zeitlich entwickelt, muss man wieder zu den integralen Erhaltungssätzen zurückgehen und diese lösen. Setzt man eine Unstetigkeit in die integralen Gleichungen ein, so erhält man eine Bedingung, welche den Sprung der Variablen, der Flüsse und der Ausbreitungsgeschwindigkeit koppelt.

Bemerkung: Die Ausbreitungsgeschwindigkeit der Stoßwelle berechnet sich aus der Sprungbedingung

$$s(u_r - u_l) = f(u_r) - f(u_l) \,, \tag{4.61}$$

welche sich aus der integralen Erhaltung ableiten. Man nennt diese Sprungbedingung in Anlehnung an die Gasdynamik **Rankine-Hugoniot-Bedingung**.

Beispiel 4.11. Wir greifen das Problem in Beispiel 4.10 noch einmal auf. Für Zeitpunkte $t > 1$ hat man die Werte $u_r = 1$ rechts und $u_l = 0$ links der Unstetigkeit. Nach der Rankine-Hugoniot-Bedingung (4.61) ergibt sich die Geschwindigkeit der Stoßwelle zu

$$s = \frac{\frac{1}{2}u_r^2 - \frac{1}{2}u_l^2}{u_r - u_l} = \frac{1}{2}(u_r + u_l) = \frac{1}{2} \,. \tag{4.62}$$

In diesem Fall lautet die Lösung für $t > 1$ dann

$$u(x,t) = \begin{cases} 1 & \text{für } x < \frac{t+1}{2} \,, \\ 0 & \text{für } x > \frac{t+1}{2} \,. \end{cases} \tag{4.63}$$

Sie ist stückweise konstant und springt an der Unstetigkeitsstelle. □

Unstetigkeiten dieser Art sind auch Flammen- und Detonationsfronten in den makroskopischen Erhaltungsgleichungen für reaktive Strömungen.

Eine einzelne **mehrdimensionale Erhaltungsgleichung** hat die Gestalt

$$u_t + \nabla \cdot \mathbf{f}(u) = 0 \,. \tag{4.64}$$

Dabei ist der Fluss $\mathbf{f}(u)$ eine vektorwertige Funktion. In der Komponentenschreibweise lautet die zweidimensionale Erhaltungsgleichung

$$u_t + f_1(u)_x + f_2(u)_y = 0 \tag{4.65}$$

mit den Komponenten f_1 und f_2 des Flusses $\mathbf{f}$.

Ein **System von Erhaltungsgleichungen in zwei Raumdimensionen** lautet

$$\mathbf{u}_t + \mathbf{f}_1(\mathbf{u})_x + \mathbf{f}_2(\mathbf{u})_y = \mathbf{0} \,. \tag{4.66}$$

Dabei ist $\mathbf{u}$ der Lösungsvektor mit m Komponenten, $\mathbf{f_1}(\mathbf{u})$ und $\mathbf{f_2}(\mathbf{u})$ die vektorwertigen Flüsse in x- bzw. in y-Richtung. Sind die Flüsse stetig differenzierbar, dann kann man das System von Erhaltungsgleichungen in die quasilineare Form (4.51) mit den Funktionalmatrizen

$$\mathbf{A}(\mathbf{u}) = \frac{d\mathbf{f}_1(\mathbf{u})}{d\mathbf{u}} \,, \qquad \mathbf{B}(\mathbf{u}) = \frac{d\mathbf{f}_2(\mathbf{u})}{d\mathbf{u}}$$

umschreiben.

4.7 Anwendungen

In diesem Abschnitt gehen wir auf die mathematische Modellierung verschiedener physikalischer Probleme ein. Die Typeinteilung der partiellen Differenzialgleichungen wird dabei aufgegriffen und Eigenschaften der Lösungen der verschiedenen Gleichungen an Hand von Beispielen erläutert.

4.7.1 Die kompressiblen Navier-Stokes-Gleichungen

Spricht man von einem Fluid, dann ist dies ein System von vielen Teilchen, welche miteinander wechselwirken. Es ist eine Materie, welche sich durch Oberflächenkräfte verformt, wobei sich die Teilchen frei bewegen können. Die Gleichung (4.55) in dem Kapitel über Erhaltungsgleichungen ist allgemein die Integralgleichung eines physikalischen Erhaltungssatzes. Als Beispiel hatten wir im vorigen Abschnitt schon die Massenerhaltung angeführt. Die zentralen

Erhaltungssätze in der Strömungsmechanik sind neben der Massenerhaltung die Impuls- und die Energieerhaltung.

Die **Impulsgleichung** ist eine Vektorgleichung. Die Impulsdichte ist der Impuls pro Einheitsvolumen $\rho\mathbf{v}$. Es ist ein Vektor mit den verschiedenen Komponenten des Impulses in x-, y- und z-Richtung. Die integrale Impulsgleichung lautet

$$\frac{d}{dt}\int\limits_{\Omega}\rho\mathbf{v}\,dV = -\int\limits_{\partial\Omega}\big((\rho\mathbf{v})\circ\mathbf{v}\big)\cdot\mathbf{n}dS + \int\limits_{\partial\Omega}\boldsymbol{\sigma}\cdot\mathbf{n}dS + \int\limits_{\partial\Omega}\mathbf{g}dV\ .$$

Das linke Integral steht für die zeitliche Änderung des Impulses. Integration und Differenziation eines Vektors werden dabei komponentenweise ausgeführt. Das erste und zweite Integral auf der rechten Seite enthält die Oberflächenkräfte. Das erste Integral steht für die Impulsänderung, verursacht durch den Massentransport und damit ein Konvektions- oder Transportterm. Das Zeichen „$\circ$“ bezeichnet dabei das dyadische Produkt: $(\rho\mathbf{v}\circ\mathbf{v})$, ein Tensor, der zeilenweise mit der Normalen $\mathbf{n}$ multipliziert wird. Das zweite Integral enthält die Spannungskräfte, die von den Normal- und Tangentialspannungen an der Oberfläche herrühren. Dabei kann man den **Spannungstensor** $\boldsymbol{\sigma}$ zerlegen in die Druckspannungen p und die Zähigkeits- oder Reibungsspannungen $\boldsymbol{\tau}$:

$$\boldsymbol{\sigma} = \boldsymbol{\tau} - p\mathbf{I}$$

mit der Einheitsmatrix I. Das dritte Integral auf der rechten Seite enthält die Volumenkräfte, welche durch eine äußere Kraft $\mathbf{g}$, wie z.B. die Gravitation, entstehen.

Leiten wir wieder unter der Annahme, dass alle auftretenden Funktionen stetig differenzierbar sind, eine lokale Formulierung ab, dann erhalten wir die Impulsgleichung als Differenzialgleichung

$$(\rho\mathbf{v})_t + \nabla\cdot\big((\rho\mathbf{v})\circ\mathbf{v}\big) + \nabla p = \nabla\cdot\boldsymbol{\tau} + \mathbf{g} \tag{4.67}$$

als das mathematische Modell in Form einer Differenzialgleichung. Die Divergenz der Tensoren $\rho\mathbf{v})\circ\mathbf{v}$ und $boldsymbol\tau$ ist so zu verstehen, dass die Divergenz auf jede Zeile der Matrix angewandt wird. Jede der Komponenten der Vektorgleichung entspricht der Form einer Erhaltungsgleichung.

Als nächstes betrachten wir die **Energiegleichung**. Die Energiedichte e ist die Gesamtenergie pro Einheitsvolumen. Die integrale Erhaltung lautet

$$\frac{d}{dt}\int\limits_{\Omega}e dV = -\int\limits_{\partial\Omega}\mathbf{v}(e+p)\cdot\mathbf{n}dS + \int\limits_{\partial\Omega}(\boldsymbol{\tau}\mathbf{v})\cdot\mathbf{n}dS - \int\limits_{\partial\Omega}\mathbf{q}\cdot\mathbf{n}dS + \int\limits_{\Omega}\mathbf{g}\cdot\mathbf{v}dV + \int\limits_{\Omega}Q dV\ .$$

In den Oberflächenintegralen geht die Arbeit pro Zeiteinheit ein, welche die Umgebung an dem Fluid leistet. Rechts haben wir als ersten Term wieder den

Transport von Energie durch die Strömung. Die Arbeit, welche die Druckkräfte leisten, ergeben sich aus den Druckkräften multipliziert mit der Geschwindigkeit. Das zweite Integral auf der rechten Seite enthält die Arbeit durch die Reibung, das dritte den Wärmefluss $\mathbf{q}$ durch den Rand. Die geleistete Arbeit der äußeren Kräfte ist durch das erste Volumenintegral und eine von außen zugeführte Wärmemenge Q durch das zweite gegeben.

Ganz analog zu den beiden anderen Gleichungen kann man eine lokale Formulierung ableiten. Das aus Massen-, Impuls- und Energiegleichung abgeleitete Differenzialgleichungssystem nennt man die **kompressiblen Navier-Stokes-Gleichungen**:

$$\rho_t + \nabla \cdot (\rho \mathbf{v}) = 0 \,, \tag{4.68}$$

$$(\rho \mathbf{v})_t + \nabla \cdot \big((\rho \mathbf{v}) \circ \mathbf{v}\big) + \nabla p = \nabla \cdot \boldsymbol{\tau} + \mathbf{g} \,, \tag{4.69}$$

$$e_t + \nabla \cdot \big(\mathbf{v}(e+p)\big) = \nabla \cdot (\boldsymbol{\tau}\mathbf{v}) - \nabla \cdot \mathbf{q} + Q \,. \tag{4.70}$$

Es gibt in diesem System zunächst mehr gesuchte physikalische Größen als Gleichungen. Wir haben drei Gleichungen, zwei skalare und eine Vektorgleichung, Als Unbekannte sind die vier physikalische Grundgrößen: ρ,$\mathbf{v}$,p,e zu bestimmen. Eine zusätzliche Gleichung ist die **Zustandsgleichung**. Der Druck ist eine Funktion der anderen physikalischen Größen. Oft wird die Zustandsgleichung in der Form

$$p = p(\rho, \varepsilon)$$

mit der spezifischen inneren Energie ε geschrieben. Gesamt-, innere und kinetische Energie erfüllen dabei die Gleichung

$$e = \rho\varepsilon + \frac{1}{2}\rho \mathbf{v}^2 \,. \tag{4.71}$$

Für ein ideales Gas lautet die Zustandsgleichung

$$p(\rho, \varepsilon) = (\gamma - 1)\rho\varepsilon \tag{4.72}$$

mit dem Adiabatenexponenten γ. Die Wärmeleitung ist dem Temperaturgradienten proportional

$$\mathbf{q} = -k \, \nabla T \tag{4.73}$$

mit der Wärmeleitfähigkeit k.

Ist die Wärmeleitfähigkeit k konstant, dann ergibt sich auf der rechten Seite der Energiegleichung der Term

$$k\Delta T \ .$$

Der Typ der Energiegleichung ist demnach **parabolisch**. Er wird festgelegt durch die höchsten Ortsableitungen. Auch bei der Impulsgleichung haben wir einen parabolischen Term auf der rechten Seite stehen: den Reibungstensor $\boldsymbol{\tau}$. Für ein Newtonsches Fluid lautet dieser unter der Annahme der Stokesschen Hypothese

$$\boldsymbol{\tau} = 2\eta\mathbf{D} - \frac{2}{3}\eta(\nabla \ \cdot \mathbf{v})\mathbf{I} \quad \text{mit} \quad D = \frac{1}{2}\Big[\nabla \circ \mathbf{v} + (\nabla \circ \mathbf{v})^T\Big] \ .$$

Dabei bezeichnet η den Viskositätskoeffizienten. Bei der Massenerhaltung treten keine zweiten Ableitungen auf. Es ist eine hyperbolische Gleichung. Man nennt deshalb die **kompressiblen Navier-Stokes-Gleichungen** ein **hyperbolisch-parabolisches System** oder auch **unvollständig parabolisch**.

Besonders bei Gasströmungen ist die Reibung und die Wärmeleitung sehr klein und kann deshalb oft vernachlässigt werden. Aus den kompressiblen Navier-Stokes-Gleichungen ergeben sich die **Euler-Gleichungen** oder auch die **Gleichungen der Gasdynamik** genannt:

$$\rho_t + \nabla \cdot (\rho\mathbf{v}) = 0 \ , \tag{4.74}$$

$$(\rho\mathbf{v})_t + \nabla \cdot \big((\rho\mathbf{v}) \circ \mathbf{v}\big) + \nabla p = 0 \ , \tag{4.75}$$

$$e_t + \nabla \cdot \big(\mathbf{v}(e+p)\big) = 0 \ , \tag{4.76}$$

hier ohne Berücksichtigung von äußeren Kräfte.

Betrachten wir eine Rohrströmung, welche sich in x-Richtung ausbreitet, und nehmen wir an, dass sich die physikalischen Größen nur in Strömungsrichtung ändern, dann fallen in dem System der Euler-Gleichungen die Ableitungen nach y und z heraus. Ebenso verschwinden die Geschwindigkeitskomponenten in diesen beiden Richtungen. Was übrig bleibt, ist ein System aus drei skalaren Gleichungen, den **eindimensionalen Euler-Gleichungen**. Dabei bezeichnet v hier die Geschwindigkeitskomponente in x-Richtung. Meist schreibt man die eindimensionalen Euler-Gleichungen als System in Erhaltungsform

$$\mathbf{u}_t + \mathbf{f}(\mathbf{u})_x = 0 \tag{4.77}$$

mit dem Vektor der Erhaltungsgrößen $\mathbf{u}$ und dem Fluss $\mathbf{f}(\mathbf{u})$:

$$\mathbf{u} = \begin{pmatrix} \rho \\ \rho v \\ e \end{pmatrix}, \qquad \mathbf{f}(\mathbf{u}) = \begin{pmatrix} \rho v \\ \rho v^2 + p \\ v(e+p) \end{pmatrix}.$$

Dieses System kann man in die quasilineare Form umschreiben. Es ergibt sich ein System von Evolutionsgleichungen.

$$\mathbf{u}_t + \mathbf{A}(\mathbf{u})\mathbf{u}_x = 0 \tag{4.78}$$

mit der Funktionalmatrix

$$\mathbf{A}(\mathbf{u}) = \frac{d\mathbf{f}(\mathbf{u})}{d\mathbf{u}}.$$

Die Matrix $\mathbf{A}(\mathbf{u})$ besitzt die Eigenwerte

$$a_1 = v - c\,, \qquad a_2 = v\,, \qquad a_3 = v + c\,. \tag{4.79}$$

Dabei bezeichnet c die Schallgeschwindigkeit, welche sich über die Beziehung

$$c^2 = \gamma \frac{p}{\rho}$$

ergibt. Die Eigenwerte entsprechen den verschiedenen Wellengeschwindigkeiten in einem Gas. Die eindimensionalen **Euler-Gleichungen** besitzen somit drei reelle Eigenwerte und sind ein System von **nichtlinearen streng hyperbolischen Gleichungen**.

Bemerkung: Die Druckwellen breiten sich mit der Geschwindigkeit $v - c$ oder $v + c$ aus. Der Eigenwert $a_2 = v$ entspricht dem Transport mit der Strömungsgeschwindigkeit. Ist $|v| > c$, dann ist die vorliegende Strömung eine Überschallströmung. Sind alle drei Eigenwerte positiv, so laufen die Charakteristiken in der (x, t)-Ebene nach rechts. Damit wird Information nur von links nach rechts transportiert. Keine Information kann gegen die Strömungsrichtung laufen.

Die Steigungen der Charakteristiken für die nichtlinearen Euler-Gleichungen sind lösungsabhängig und damit im Allgemeinen keine Geraden. Schneiden sich Charakteristiken zum gleichen Eigenwert, dann kann keine stetig differenzierbare Lösung existieren. Wir haben dies im Abschnitt 4.4 an einer einzelnen Gleichung erläutert. In diesem Fall kommt es zum Auftreten einer **Stoßwelle** oder auch **Verdichtungsstoß** genannt. Der Überschallknall bei einem Düsenjäger ist ein bekanntes Beispiel für dieses physikalische Phänomen.

Es ist interessant, dass das Phänomen des Verdichtungsstoßes als ein Phänomen der Lösung der zugehörigen nichtlinearen Gleichungen entdeckt wurde.

Es war der Mathematiker Bernhard Riemann, welcher dies in einer Arbeit in den Göttinger Nachrichten 1860 beschrieb. Er führte den Begriff des Verdichtungsstoßes ein. Erst einige Jahre später wurde z. B. durch Ernst Mach die Stoßwelle oder der Verdichtungsstoß experimentell bestätigt.

Bemerkung: Existiert für das betrachtete Problem eine stationäre Lösung, dann ist diese auch Lösung der hier vorgestellten Gleichungen. Stationär bedeutet, dass die Zeitableitungen verschwinden. Man kann somit ein vereinfachtes System angeben, in dem man die Zeitableitungen heraus streicht. Dieses System ist hyperbolisch für eine Überschall- und elliptisch für eine Unterschall-Strömung.

Zusammenfassung: Die **kompressiblen Navier-Stokes-Gleichungen** (4.73)-(4.75) sind **unvollständig parabolisch**. Die Massenerhaltung (4.73) ist hyperbolisch, während die Reibungs- und Wärmeleitungsterme dafür sorgen, dass die Impuls- und Energiegleichung parabolisch sind.
Die **Euler-Gleichungen** oder Gleichungen der Gasdynamik, bei denen Reibung und Wärmeleitung vernachlässigt werden, sind ein **System nichtlinearer hyperbolischer Differenzialgleichungen**.

4.7.2 Die inkompressiblen Navier-Stokes-Gleichungen

Die kompressiblen Navier-Stokes-Gleichungen beschreiben Strömungen, bei denen die Kompressibilität, das heißt Veränderungen der Dichte, eine wesentliche Rolle spielen. Man nennt diesen Bereich der Strömungsmechanik auch die Strömungsmechanik der hohen Geschwindigkeiten. Die Strömung von Flüssigkeiten kann meist als inkompressibel betrachtet werden. Der eigentliche physikalische Hintergrund liegt in dem Unterschied der Ausbreitungsgeschwindigkeit des Drucks und der Strömung, das heißt der Schall- und der Strömungsgeschwindigkeit. Das Verhältnis dieser beiden wird als Machzahl bezeichnet:

$$M = \frac{v}{c} \, .$$

Ist die Machzahl in der Größenordnung von Eins, dann haben die Geschwindigkeit der Druckwellen und der Strömungstransport etwa die gleiche Größenordnung. Wird die Machzahl klein, dann sind die Druck- oder Schallwellen sehr viel schneller als die lokale Strömung. Dies bedeutet, dass ein schneller Druckausgleich erfolgt und lokale Änderungen im Geschwindigkeitsfeld keine nennenswerten lokalen Druckunterschiede erzeugen können. Dies bedeutet dann natürlich auch keine nennenswerten Dichteunterschiede: Die Strömung ist inkompressibel.

Die Frage kompressibel oder inkompressibel ist somit keine Stoffeigenschaft, sondern ein Strömungsbereich, welcher durch das Verhältnis Schall- und Strömungsgeschwindigkeit bestimmt ist. Da Flüssigkeiten eine große Schallgeschwindigkeit besitzen, gelten viele Strömungen von Flüssigkeiten als inkompressibel. Explosionen im Wasser müssen dann aber wieder mit dem kompressiblen mathematischen Modell beschrieben werden. Gasströmungen mit kleinen Geschwindigkeiten werden oft mit den inkompressiblen Navier-Stokes-Gleichungen berechnet, wie die Umströmung von Kraftfahrzeugen.

Die inkompressiblen Gleichungen erhält man als Grenzwert der kompressiblen Gleichungen, wenn die Machzahl gegen Null strebt. Liegt eine Strömung mit anfangs räumlich konstanter Dichte vor, dann bleibt diese Dichte auch zeitlich konstant. Damit ergibt sich aus dem Massenerhalt (4.68) direkt die Bedingung der Divergenzfreiheit des Geschwindigkeitsfeldes. Die Impulsgleichung kann durch die konstante Dichte dividiert werden.

In dimensionsloser Form haben die **inkompressiblen Navier-Stokes-Gleichungen** die Gestalt

$$\mathbf{v}_t + (\mathbf{v} \cdot \nabla)\mathbf{v} + \nabla p = \frac{1}{Re} \Delta \mathbf{v} \,, \tag{4.80}$$

$$\nabla \cdot \mathbf{v} = 0 \,. \tag{4.81}$$

Dabei bezeichnet Re die Reynoldszahl - eine dimensionslose Kennzahl, welche ein Maß für die Reibung darstellt. Die Energiegleichung ist bei Gültigkeit von (4.81) automatisch erfüllt. Es sind zwei Gleichungen, eine Vektorgleichung und einer skalaren Gleichung für die zwei Unbekannten Geschwindigkeit v und Druck p.

Die zweite Gleichung, als zusätzliche Bedingung an die Geschwindigkeit, enthält keine Zeitableitungen. Die inkompressiblen Gleichungen sind somit kein typisches System von Evolutionsgleichungen. Da die Gleichung (4.81) keine Zeitableitungen enthält, ist die Vermutung, dass es sich hierbei um eine elliptische Gleichung handelt. Den elliptischen Charakter der Gleichungen sieht man folgendermaßen. Man wendet die Divergenz als Differenzialoperator auf die Geschwindigkeitsgleichung an. Da das Geschwindigkeitsfeld divergenzfrei ist, fällt die Zeitableitung der Divergenz von v heraus. Auf der linken Seite schreiben wir die Divergenz des Gradienten des Drucks als Δp. Alle anderen Terme schreiben wir auf die rechte Seite und erhalten wegen $\nabla \cdot \Delta v = 0$ die Gleichung

$$\Delta p = -\nabla \cdot \big[(v \cdot \nabla) v\big] \,. \tag{4.82}$$

Ist die Geschwindigkeit bekannt, so stehen auf der rechten Seite bekannte Terme. Zu einem bekannten Geschwindigkeitsfeld ergibt sich somit der Druck als Lösung der **elliptischen Poisson-Gleichung** (4.82).

Ein vereinfachtes mathematisches Modell für eine inkompressible Strömung ist die Potenzialströmung. Neben der Annahme, dass die Strömung reibungsfrei wie bei den Euler-Gleichungen ist, wird zusätzlich vorausgesetzt, dass die Strömung rotationsfrei ist:

$$\nabla \times \mathbf{v} = 0 \, .$$

Es folgt daraus, dass ein Geschwindigkeitspotenzial mit $\mathbf{v} = -\nabla \phi$ existiert. Die Bedingung (4.81) ergibt dann die Forderung

$$\nabla \cdot \mathbf{v} = -\nabla \cdot \nabla \phi = 0$$

an das Geschwindigkeitspotenzial. Die Geschwindigkeitsgleichung kann danach integriert werden und ergibt eine algebraische Gleichung, die Bernoulli-Gleichung.

Zusammenfassung: Die inkompressiblen Navier-Stokes-Gleichungen (4.80), (4.81) sind ein **parabolisch-elliptisches System**. Als Grenzwert, wenn die Reynoldszahl gegen unendlich strebt und die Reibungsterme vernachlässigt werden können, erhält man die **hyperbolisch-elliptischen inkompressiblen Euler-Gleichungen**. Wird zusätzlich eine rotationsfreie Strömung vorausgesetzt, ergibt sich die Laplace-Gleichung für das Geschwindigkeitspotenzial

4.7.3 Die Gleichungen der Akustik

Die Schall- oder Lärmausbreitung sind Vorgänge, welche im Prinzip Lösungen der kompressiblen Navier-Stokes-Gleichungen sind. Jedoch sind die Amplituden sehr klein, so dass man das System der Gleichungen der Strömungsmechanik vereinfachen kann. Betrachtet man Druckwellen mit kleinen Amplituden in einer Ruheströmung, d.h. die Geschwindigkeit ist nahezu Null, dann kann man den Störungsansatz

$$\rho = \rho_0 + \rho' \, , \qquad \mathbf{v} = \mathbf{v}' \, , \qquad p = p_0 + p' \tag{4.83}$$

in die Euler-Gleichungen einsetzen. Die Größen ρ_0 und p_0 werden dabei als konstant betrachtet, die Strichgrößen sind die Fluktuationen der konstanten Hintergrundströmung. Eingesetzt in die Gleichungen erhält man z.B. für die Massenerhaltung

$$\rho'_t + \rho_0 \nabla \cdot \mathbf{v}' + \nabla \cdot (\rho' \mathbf{v}') = 0 \, .$$

Dabei wurde schon ausgenutzt, dass die Hintergrundströmung räumlich und zeitlich konstant ist und die Ableitungen von ρ_0 und p_0 verschwinden.

Die Strichgrößen sind sehr klein. Es wird daher angenommen, dass die Produkte dieser Größen und auch deren Ableitungen vernachlässigt werden können. Es bleibt dann die Gleichung

$$\rho'_t + \rho_0 \nabla \cdot \mathbf{v}' = 0 \tag{4.84}$$

übrig. Dies kann man analog mit Impuls- und Energiegleichung ausführen. Setzt man die Zustandsgleichung ein, bleiben die Gleichungen für die Fluktuationen in der Form

$$\mathbf{v}'_t + \frac{1}{\rho_0} \nabla p' = 0 \,, \tag{4.85}$$

$$p'_t + \gamma p_0 \nabla \cdot \mathbf{v}' = 0 \tag{4.86}$$

zurück. Die Kombination der Gleichungen für die Dichte (4.84) und den Druck (4.86) führt auf die Beziehung $p' - c_0^2 \rho' =$ konstant, wobei c_0 die Schallgeschwindigkeit in der Ruheströmung bezeichnet. Mit dieser Beziehung lassen sich die Dichtefluktuationen aus den Druckfluktuationen bestimmen.

Es bleibt das System (4.85), (4.86) übrig, welches eine ganz ähnliche Form wie die Wellengleichung (4.35), (4.36) besitzt. In der Tat, wird auf die erste Gleichung die Divergenz angewandt und mit γp_0 multipliziert, die zweite nach der Zeit differenziert und die beiden Gleichungen subtrahiert, so ergibt sich

$$p_{tt} - c_0^2 \Delta p = 0 \,, \tag{4.87}$$

d. h. die lineare Wellengleichung. Die Wellengeschwindigkeiten in einer Raumdimension sind c_0 und $-c_0$. Man nennt die Gleichungen (4.85), (4.86) oder die Gleichung zweiter Ordnung für die Druckfluktuationen die **Gleichungen der linearen Akustik**. Die Ausbreitung von Schall und Lärm in das Fernfeld in einer Ruheströmung wird mit diesem mathematischen Modell beschrieben.

4.8 Bemerkungen

Wie wir in den einzelnen Kapiteln gesehen haben, hat die Klassifizierung der partiellen Differenzialgleichungen eine physikalische Motivation. Die verschiedenen Typen von Differenzialgleichungen beschreiben verschiedene physikalische Vorgänge. Die Lösungen dieser Probleme haben somit verschiedene Eigenschaften. Die Kenntnis darüber ist ganz wesentlich für die numerische Approximation, da hier die wichtigen Eigenschaften richtig wiedergegeben

werden sollen. Möchte man ein Problem mit einem numerischen Verfahren näherungsweise lösen, so ist der Typ der Differenzialgleichung zunächst die wichtigste Information für die Auswahl eines numerischen Verfahrens. Die Klassifizierung gibt Hinweise auf die vorzugebenden Rand- oder Anfangsbedingungen und über die Stetigkeitseigenschaften der Lösung.

Elliptische Probleme beschreiben vor allem stationäre Vorgänge. Wohldefiniert ist ein elliptisches Problem durch die Vorgabe von geeigneten Randbedingungen. Die Lösungen von elliptischen Problemen, speziell von linearen Problemen haben sehr gute Stetigkeitseigenschaften. Die Konstruktion von Verfahren hoher Ordnung ist für diese Problemklasse somit möglich. Es gilt für elliptische Probleme ein sogenanntes Maximumprinzip. Das Maximum und das Minimum der Lösung wird auf dem Rand angenommen und niemals im Innern des Gebietes.

Parabolische Probleme sind im Allgemeinen instationäre Probleme. Sie beschreiben Wärmeleitung, Dissipation und Reibungseinflüsse. Ein parabolisches Problem ist ein Anfangs-Randwertproblem. Ähnlich wie bei den elliptischen Problemen ist die Lösung im Allgemeinen sehr glatt. Dies hat sich sehr deutlich in dem Beispiel 4.3 gezeigt. Die kurzwelligen Anteile in den Anfangswerten wurden schneller gedämpft als die langwelligen Anteile. Die Glattheit der Lösung erhöht sich mit fortschreitender Zeit. Auch ein parabolisches Problem genügt einem Maximumprinzip. So fällt der Wert des Maximums in den Anfangswerten monoton, während der Wert des Minimums monoton wächst.

Hyperbolische Probleme beschreiben Wellen und Transportvorgänge. Die Charakteristiken spielen hier eine entscheidende Rolle. Die Information wird entlang der Charakteristiken transportiert. Das zugehörige Problem ist ein reines Anfangswertproblem. Läuft das physikalische Problem in einem endlichen Gebiet ab, so können dort die Randwerte nicht beliebig vorgeschrieben werden, sondern es gibt Verträglichkeitsbedingungen: Die Randwerte müssen mit der Wellenausbreitung verträglich sein. Nichtlineare hyperbolische Gleichungen können für beliebig oft stetig differenzierbare Anfangswerte Unstetigkeiten entwickeln. Diese Unstetigkeiten können physikalische Bedeutung haben, wie Stoßwellen in der Gasdynamik. Unstetigkeiten in den Anfangswerten werden bei linearen Problemen entlang der Charakteristiken transportiert. Eine Regularitätserhöhung der Lösung mit fortschreitender Zeit tritt bei hyperbolischen Problemen nicht auf.

Wir haben in unseren Betrachtungen als Rechengebiet immer ein endliches Gebiet vorausgesetzt. Für die numerische Simulation ist dies unumgänglich, da der Rechner nur endlich viel Speicherplatz besitzt. Physikalische Vorgänge können auf einem sehr großen Gebiet ablaufen und die entsprechende mathematische Modellierung führt auf ein unendliches Gebiet, in dem die Lösung gesucht wird. Für die numerische Simulation muss man dieses Gebiet ein-

schränken und geeignete Randbedingungen finden. Man nennt diese numerische oder künstliche Randbedingungen. Um das Rechengebiet zu verkleinern und damit die Effizienz der numerischen Simulation zu erhöhen, wird man dieses Vorgehen immer dann anwenden, wenn die Einflüsse des Randes vernachlässigbar sind. Allerdings ist die Vorgabe von künstlichen Randwerten nicht einfach und für viele Probleme ein Bereich der aktuellen Forschung. Bei der Behandlung der numerischen Methoden werden wir darauf zurückkommen.

Über die Theorie von partiellen Differenzialgleichungen gibt es eine ganze Reihe von Monographien. Eine Übersicht enthalten z. B. die zwei klassischen Bände der Methoden der mathematischen Physik von Courant und Hilbert [11], immer noch sehr zu empfehlen. Weitere Monographien über partielle Differenzialgleichungen sind z.B. die Bücher von Strauss [61], Hellwig [31], John [34], Evans [18] und Quarteroni und Valli [48]. Da die Eigenschaften der verschiedenen Typen von partiellen Differenzialgleichungen sehr unterschiedlich und vielfältig sind und eine Vielzahl unterschiedlicher physikalischer Phänomene beschreiben, gibt es auch Bücher über die einzelnen Klassen der Differenzialgleichungen.

4.9 Beispiele und Aufgaben

Beispiel 4.12 (Mit MAPLE-Worksheet). Das Anfangswertproblem für die Differenzialgleichung

$$u_t + \frac{1}{2}\sqrt{t}\, u_x = 0$$

mit den Anfangswerten $u(x,0) = q(x) = e^{-x^2}$ wird im Folgenden gelöst. Die Transportgeschwindigkeit hängt hier von der Zeit ab. Die Charakteristiken sind demnach gegeben durch

$$C : x = x(t) \quad \text{mit} \quad \frac{dx(t)}{dt} = \frac{1}{2}\sqrt{t}$$

und sind Kurven in der (x,t)-Ebene. Die Lösung u ist konstant entlang den Charakteristiken und ergibt sich zu

$$u(x,t) = q\left(x - \frac{1}{3}t^{\frac{3}{2}}\right) = e^{-\left(x-\frac{1}{3}t^{\frac{3}{2}}\right)^2}.$$

Das MAPLE-Worksheet zeigt die Lösung als Funktion von x und t und ist in Abbildung 4.15 gezeichnet. □

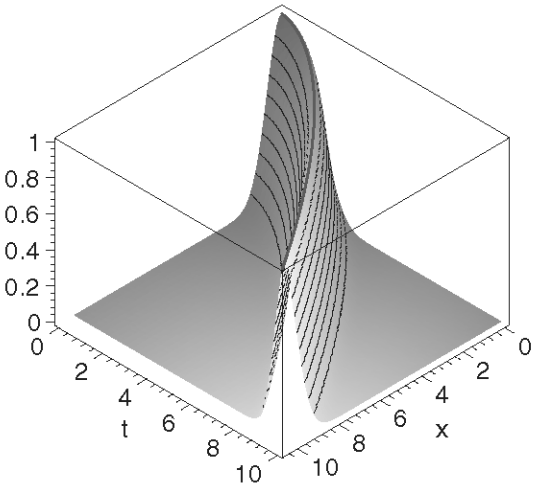


Abb. 4.15. Lösung von Beispiel 4.12

Beispiel 4.13 (Mit MAPLE-Worksheet). Bei dem AWP für die lineare Transportgleichung

$$u_t + \sqrt{x}\, u_x = 0$$

und den Anfangswerten $u(x,0) = q(x) = e^{-x^2}$ hängt die Ausbreitungsgeschwindigkeit von x ab. Trotzdem lassen sich analog zu Beispiel 4.12 die Charakteristiken definieren. Die Lösung ist entlang den Charakteristiken konstant und lautet

$$u(x,t) = q(t - 2\sqrt{x}) = e^{-(t-2\sqrt{x})^2} \ .$$

Das MAPLE-Worksheet liefert Abbildung 4.16. Man erkennt, wenn man entlang der Zeitachse läuft, dass die Anfangsauslenkung entlang der Charakteristiken durch den Punkt ($x = 0$, $t = 0$) läuft. Diese hat jetzt die Gleichung

$$x = \left(\frac{t}{2}\right)^2 \ .$$

□

Beispiel 4.14 (Riemann-Problem für die Burgers-Gleichung). Ein Anfangswertproblem mit stückweisen konstanten Anfangswerten wird **Riemann-Problem** genannt. Wir betrachten hier die Lösung eines Riemann-Problems für die Burgers-Gleichung

$$u_t + (\frac{1}{2}u^2)_x = 0 \tag{4.88}$$

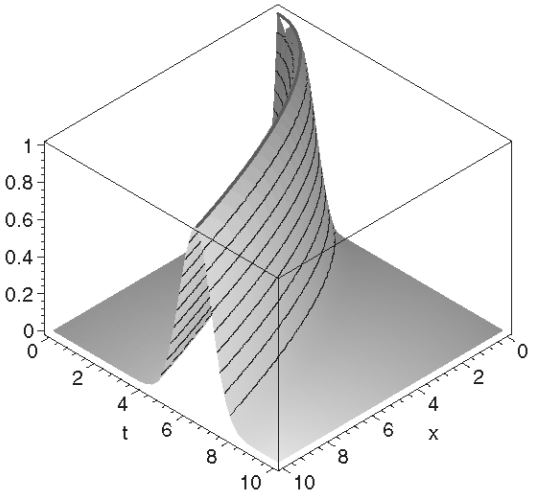


Abb. 4.16. Lösung von Aufgabe 4.13

mit den Anfangswerten

$$u(x,0) = \begin{cases} 1 & \text{für } x < 0\,, \\ 0 & \text{für } x > 0\,. \end{cases} \tag{4.89}$$

Die Steigungen der Charakteristiken in der (x,t)-Ebene haben wegen $a(u) = u$ den Wert $1/u$. In der linken Halbebene sind die Charakteristiken somit Parallelen zur Winkelhalbierenden, auf der rechten Seite Parallelen zur t-Achse. Das entsprechende Diagramm ist in der Abbildung 4.17 auf der linken Seite gezeichnet. Die Charakteristiken schneiden sich im Bereich zwischen der t-Achse und der Winkelhalbierenden. Das Schneiden der Charakteristiken bedeutet, dass die Lösung eine Unstetigkeit - Stoßwelle - enthält. Die Ausbreitungsgeschwindigkeit der Stoßwelle berechnet sich aus der Sprungbedingung (4.61) zu

$$s = \frac{\frac{1}{2}u_r^2 - \frac{1}{2}u_l^2}{u_r - u_l} = \frac{1}{2}(u_r + u_l) = \frac{1}{2}\,. \tag{4.90}$$

Die Ausbreitungskurve der Stoßwelle ist somit eine Gerade mit der Gleichung $t = 2x$. In das rechte Bild von Abbildung 4.17 ist diese Gerade eingezeichnet. Die schwache Lösung für das angegebene Riemann-Problem ist somit

$$u(x,t) = \begin{cases} 1 & \text{für } \frac{x}{t} < \frac{1}{2}\,, \\ 0 & \text{für } \frac{x}{t} > \frac{1}{2}\,. \end{cases} \tag{4.91}$$

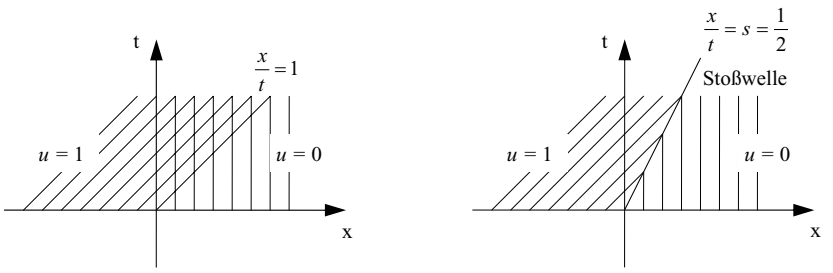


Abb. 4.17. Verlauf der Charakteristiken für Beispiel 4.14

Die Stoßwelle trennt die beiden konstanten Zustände $u_l = 1$ und $u_r = 0$, die Charakteristiken laufen von beiden Seiten in die Stoßwelle hinein. □

Beispiel 4.15 (Riemann-Problem für die Burgers-Gleichung). Wir betrachten die Lösung des Riemann-Problems für die Burgers-Gleichung mit den Anfangswerten

$$u(x,0) = \begin{cases} 0 & \text{für } x < 0\,, \\ 1 & \text{für } x > 0\,. \end{cases} \tag{4.92}$$

Hier hat man den umgekehrten Fall vorliegen. Die Charakteristiken in der linken Halbebene sind Parallelen zur t-Achse und in der rechten Halbebene Parallelen zur Winkelhalbierenden. Abbildung 4.18 zeigt, dass hier die Charakteristiken aufbrechen und es einen Bereich ohne Charakteristiken gibt. In diesem Fall führt die Unstetigkeit in den Anfangswerten auf eine stetige Lösung, bei der die Charakteristiken wie bei einem Fächer das Loch auffüllen und eine Verbindung der t-Achse mit der ersten Winkelhalbierenden schaffen. Die Lösung lautet

$$u(x,t) = \begin{cases} 0 & \text{für } \frac{x}{t} < 0\,, \\ \frac{x}{t} & \text{für } 0 \le \frac{x}{t} \le 1\,, \\ 1 & \text{für } \frac{x}{t} > 1\,. \end{cases} \tag{4.93}$$

Sie ist stetig und stetig differenzierbar bis an dem Knick bei $x = 0$ und $x = t$. □

Beispiel 4.16 (Verkehrsflussmodell). Wir betrachten ein einfaches mathematisches Modell für die Beschreibung des Verkehrsflusses auf einer Autobahn. Ist der Verkehr dicht, was ab und zu auf den deutschen Autobahnen vorkommt, kann man ihn als Kontinuum betrachten. Die Anzahl der Fahrzeuge in einem Abschnitt I der Autobahn berechnet sich somit als

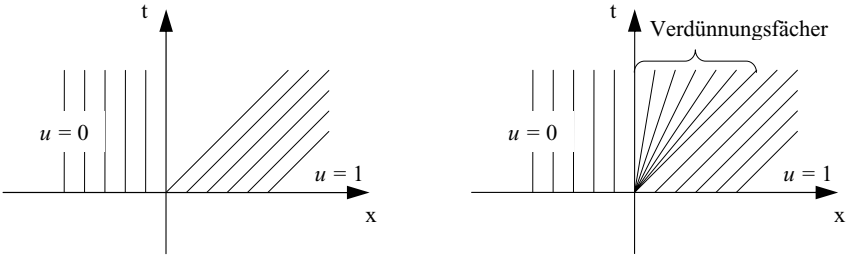


Abb. 4.18. Verlauf der Charakteristiken bei einer Verdünnungswelle

$$M(t;I) = \int\limits_I \rho(x,t)dx \ ,$$

mit der Verkehrsdichte ρ. Die zugehörige Erhaltungsgleichung lautet

$$\rho_t + f(\rho)_x = 0 \ .$$

Dabei ist der Verkehrsfluss durch $f(\rho) = \rho\, v$ gegeben wie bei der Kontinuitätsgleichung in einem Gas. Die Variable v bezeichnet die Geschwindigkeit in der Bedeutung einer lokalen Durchschnittsgeschwindigkeit.

Man benötigt nun eine Beziehung für die Geschwindigkeit als Funktion der Verkehrsdichte $v = v(\rho)$. Eine einfache Modellierung liegt auf der Hand: Die Geschwindigkeit ist maximal bei kleiner Dichte, während bei sehr großer Dichte der Verkehr letztlich zum Erliegen kommt: Verkehrsstau. Die einfachste Annahme ist eine lineare Funktion: Das Maximum v_{max} bei $\rho = 0$ und das Minimum $v = 0$ bei $\rho = \rho_{max}$ wird durch eine Gerade verbunden. Der zugehörige Fluss lautet in diesem Fall

$$f(\rho) = \rho\, v = \rho\,(v_{max} - \frac{v_{max}}{\rho_{max}}\,\rho) = -\frac{v_{max}}{\rho_{max}}\,\rho^2 + v_{max}\,\rho.$$

In Abbildung 4.19 ist die Geschwindigkeit $v(\rho)$ und der Fluss $f(\rho)$ mit diesen Annahmen gezeichnet. Die Charakteristiken der skalaren Erhaltungsgleichung sind die Geraden

$$C: \ x = x(t) \quad \text{mit} \quad \frac{dx(t)}{dt} = -2\frac{v_{max}}{\rho_{max}}\rho + v_{max} =: a(\rho) \ .$$

Entlang dieser Charakteristiken ist die Verkehrsdichte konstant.

Wir schauen uns einen Verkehrsstau mit den Anfangswerten

$$u(x,0) = \begin{cases} \rho_l & \text{für } x < 0 \ , \\ \rho_{max} & \text{für } x > 0 \ . \end{cases} \tag{4.94}$$

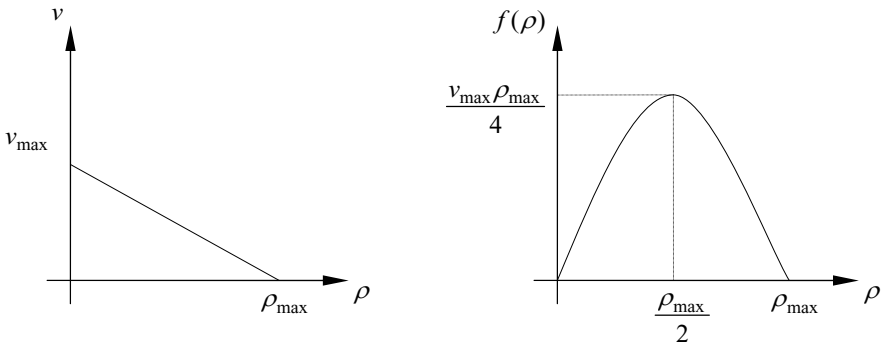


Abb. 4.19. Abhängigkeit der Geschwindigkeit v von der Verkehrsdichte ρ (links) und der zugehörige Verkehrsfluss (rechts)

genauer an. An der Stelle $x = 0$ befindet sich zum Zeitpunkt $t = 0$ das Stauende. Die Werte der Verkehrsdichte sind nach rechts ρ_{max}. Der Wert ρ_l sei kleiner als der Maximalwert. Auf der linken Seite fließt somit der Verkehr.

Die Charakteristiken sind lösungsabhängig und können nur von den bekannten Anfangswerten aus gezeichnet werden. Auf der rechten Seite, also für $x > 0$, haben die Charakteristiken in der (x, t)-Ebene die Steigung $1/a(\rho_{max}) = -1/v_{max}$. Auf der rechten Seite $1/a(\rho_l)$. Diese Steigung ist kleiner als auf der linken Seite. Nehmen wir als Beispiel $\rho_l = \rho_{max}/2$, dann ist $a(\rho_l)$ gerade Null und die Charakteristiken sind Parallelen zur t-Achse. Das Bild der Charakteristiken ist in Abbildung 4.20 in dem linken Bild eingezeichnet. Es kommt zu einem Schneiden der Charakteristiken.

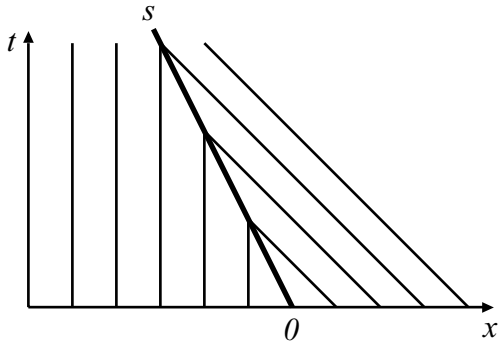


Abb. 4.20. Bild der Charakteristiken im (x, t)-Diagramm bei einem Verkehrsstau-Problem (links) und mit eingezeichneter Verdichtungsstoßwelle (rechts)

Die Lösung enthält eine Unstetigkeit, an der die Verkehrsdichte von dem linken Wert ρ_l auf den Wert ρ_{max} springt und das Auto im Stau steht. Die Ausbreitungsgeschwindigkeit des Verdichtungsstoßes ergibt sich aus der Sprung- oder Rankine-Hugoniot-Bedingung

$$s = \frac{f(\rho_r) - f(\rho_l)}{\rho_r - \rho_l} = v_{max}(1 - \frac{\rho_l + \rho_r}{\rho_{max}}) = -\frac{v_{max}}{2}$$

für unser Beispiel $\rho_l = \rho_{max}/2$. Die Unstetigkeitskurve mit der Gleichung $x/t = s$ trennt die beiden konstanten Zustände ρ_l und ρ_{max} und die Charakteristiken von den beiden Seiten.

Kommt in diesem Modell ein Auto an den Verkehrsstau, dann bremst es am Stau abrupt bis zum Stand. Diese Verdichtungswelle läuft gegen den Verkehrsstrom nach links, wie es in Abbildung 4.20 zu sehen ist. Wir betrachten Anfangswerte der Form

$$\rho(x) = 0.5\,\rho_{max}\,e^{-x^2}\ ,$$

die in dieser räumlichen Verteilung beim Feierabendverkehr so auftreten könnten: Der Verkehr ist zunächst dünn, erhöht sich dann eine gewisse Zeit kräftig und geht danach wieder zurück. Ist die Verkehrsdichte klein, dann ist die Geschwindigkeit hoch und eine Verdichtungswelle kann gegen die Fahrtrichtung laufen und gegebenenfalls zu einem Stau führen.

Das hier betrachtete Verkehrsmodell ist sehr einfach und vernachlässigt einige wichtige Effekte. Mögliche Ein- und Ausfahrten oder Geschwindigkeitsbeschränkungen sind hier nicht berücksichtigt. Berücksichtigt man die Reaktion der Autofahrer, die auf die Änderung der Verkehrsdichte mit einer Verringerung der Geschwindigkeit reagieren, dann lautet der Verkehrsfluss z.B.

$$f(\rho) = \rho v - \delta\rho_x$$

und man erhält nun statt der hyperbolischen Erhaltungsgleichung eine **Konvektions-Diffusions-Gleichung**

$$\rho_t + f(\rho)_x = \delta\rho_{xx}.$$

Die Dissipation wird hier ein Verschmieren der Verdichtungsstöße bewirken. Vom Hubschrauber aus lassen sich solche Kompressions- und Verdünnungswellen sehr gut beobachten. Der Stop- und Go-Verkehr ergibt sich aus der Interaktion dieser Wellen. Bei großer Verkehrsdichte kann eine Kompressionswelle große Strecken zurücklegen und dazu führen, dass man ohne einen offensichtlichen Grund plötzlich in einem Stau steht. Der Schuldige ist hier lediglich nur das Lösungsverhalten von nichtlinearen hyperbolischer Differenzialgleichungen. □

Aufgabe 4.1. Bestimmen sie den Typ der Differenzialgleichungen:

1. $2u_{xx} + 2u_{xy} + 7u_{yy} + 3u_x + 2u_y = x^2$,
2. $\varphi_t + \alpha\varphi_{xx} + \beta\varphi_x = 0 \quad \text{mit} \quad \alpha, \beta \in \mathbb{R}$,
3. $\sin(x)u_{xx} + 2u_{xy} + \sin(x)u_{yy} = 0$,
4. $u^2 u_{xx} + 2u_{xy} + u_{yy} = 0$.

Lösung:

1. Diese DGL 2. Ordnung ist linear mit konstanten Koeffizienten und elliptisch.

2. Es handelt sich auch hier um eine DGL mit konstanten Koeffizienten. Sie ist von 2. Ordnung und parabolisch für $\alpha \neq 0$, **Konvektions-Diffusions-Gleichung** genannt. Im Falle $\alpha = 0$ fällt der Term mit der 2. Ableitung heraus und es ergibt sich eine lineare hyperbolische Transportgleichung. Sind beide Koeffizienten null, dann entartet die Gleichung zu einer gewöhnlichen Differenzialgleichung.

3. Diese Gleichung ist eine lineare DGL 2. Ordnung. Die Koeffizienten hängen allerdings von der unabhängigen Variablen x ab. Bei die Bestimmung des Typs ergibt sich $b^2 - 4ac = 4 - 4\sin^2(x) > 0$. Die Gleichung ist somit hyperbolisch bis auf parabolische Entartungspunkte mit $sin(x) = 1$.

4. Diese Gleichung 2. Ordnung ist eine quasilineare Differenzialgleichung. Der Typ der Differenzialgleichung hängt somit von der Lösung selbst ab. Durch Einsetzen ergibt sich $b^2 - 4ac = 4 - 4u^2 \geq 4(u+1)(u-1)$. Die Gleichung ist somit hyperbolisch für $-1 < u < 1$, für $u^2 = 1$ ist sie parabolisch und in den anderen Fällen elliptisch. □

Aufgabe 4.2. Bestimmen sie den Typ des Systems 1. Ordnung

$$\begin{aligned} u_t + au_x + bv_x &= 0 \,, \\ v_t + bu_x + av_x &= 0 \end{aligned}$$

für beliebige $a, b \in \mathbb{R}$.

Lösung: Zur Bestimmung des Typs dieses Systems von Evolutionsgleichungen benötigen wir die Gleichungen in der Matrixschreibweise

$$\mathbf{w}_t + \mathbf{A}\mathbf{w}_x = 0$$

mit dem Vektor $\mathbf{w}$, welcher die Komponenten u und v besitzt. Die Eigenwerte von $\mathbf{A}$ lauten $a-b$ und $= a+b$, sind reell und das System somit hyperbolisch. □

Aufgabe 4.3. Das System der Flachwasserwellen lautet

$$\begin{aligned} h_t + hv_x + vh_x &= 0 \,, \\ v_t + vv_x + gh_x &= 0 \,. \end{aligned}$$

Dabei bedeutet v die Geschwindigkeit, h die Wassertiefe und g die Gravitationskonstante. Zeigen sie, dass dieses System hyperbolisch ist. Schreiben sie die Gleichungen in Erhaltungsform um.

Lösung: Es ist ein hyperbolisches System, da die Matrix

$$\mathbf{A} = \begin{pmatrix} v & h \\ g & v \end{pmatrix} \,.$$

wegen der nicht negativen Gravitationskonstante und Wassertiefe die reellen Eigenwerte

$$a_1 = v - \sqrt{gh} \,, \qquad a_2 = v + \sqrt{gh}$$

besitzt. Die Erhaltungsform lautet

$$h_t + (vh)_x = 0, \quad v_t + (\frac{1}{2}v^2 + gh)_x = 0 \,.$$

und ergibt sich durch Anwendung der Produktregel in der 1. und über die Stammfunktion in der 2. Gleichung. □

Aufgabe 4.4. Man zeige, dass die eindimensionalen Gleichungen der Gasdynamik in primitiven Variablen

$$\begin{aligned} \rho_t + v\rho_x + \rho v_x &= 0, \\ v_t + vv_x + \frac{1}{\rho}p_x &= 0, \\ p_t + vp_x + \gamma p v_x &= 0, \end{aligned}$$

ein hyperbolisches System sind.

Lösung: Die Eigenwerte bleiben bei der Umformulierung erhalten und sind $v - c, v, v + c$. □

5 Grundlagen der numerischen Verfahren für partielle Differenzialgleichungen

Schon gewöhnliche Differenzialgleichungen lassen sich meist nicht exakt lösen und man ist auf die numerische Lösung angewiesen. Diese Situation liegt bei den partiellen Differenzialgleichungen noch verstärkt vor. Die drei wichtigsten Ansätze für die Konstruktion von Näherungsverfahren werden in den nächsten Kapiteln vorgestellt: **Die Differenzen-, die Finite-Elemente- und die Finite-Volumen-Verfahren**. Auf die Herleitungen und die Erfahrungen mit den Methoden für gewöhnliche Differenzialgleichungen wird dabei aufgebaut. Die Ideen und Konzepte der Differenzenverfahren und der Finite-Elemente-Verfahren wurden dort in einer Raumdimension schon vorgestellt.

Bevor jedoch numerische Methoden für die verschiedenen Typen von Differenzialgleichungen vorgestellt werden, werden in diesem ersten Abschnitt grundlegende Begriffe und Eigenschaften dargestellt. Qualitätskriterien für Näherungsverfahren sind Konsistenz, Stabilität und Konvergenz. Diese Begriffe wurden ebenso schon bei den gewöhnlichen Differenzialgleichungen eingeführt und werden für die numerische Approximationen von partiellen Differenzialgleichungen erweitert oder angepasst. Die Gittererzeugung ist im Falle mehrdimensionaler Probleme weitaus komplizierter und für alle hier betrachteten Verfahren gleich, so dass wir auch diese Thematik in dieses Grundlagenkapitel eingefügt haben.

5.1 Konsistenz, Stabilität und Konvergenz

Die **Konsistenz** der numerischen Lösungsverfahren für gewöhnliche Differenzialgleichungen wurde über den lokalen Diskretisierungsfehler definiert. Unter der Annahme, dass der Wert der Näherungslösung im n-ten Schritt mit dem Wert der exakten Lösung übereinstimmt, ist der lokale Diskretisierungsfehler die Differenz des Näherungswertes und der exakten Lösung im neuen Punkt, also der Fehler, den man in einem Schritt macht. Konsistenz bedeutet, dass der lokale Diskretisierungsfehler gegen Null strebt, wenn die Schrittweite gegen Null strebt. Wir haben in dem Abschnitt 2.2.2 über Konsistenz gesehen,

dass man den lokalen Diskretisierungsfehler auch als Residuum erhält, wenn die exakte Lösung in das numerische Verfahren eingesetzt wird. Diese Vorgehensweise kann man als Definition auf den Fall der Näherungsverfahren für partielle Differenzialgleichungen entsprechend ausdehnen.

Sei unser Problem ganz allgemein durch die Gleichung

$$Lu - f = 0 \tag{5.1}$$

beschrieben. Dabei ist L ein Differenzialoperator. Wir nehmen an, dass die Rand- oder Anfangswertbedingungen mit in L enthalten sind. Das Näherungsverfahren für (5.1) ist durch eine Vorschrift

$$L_h v_h - f_h = 0 \tag{5.2}$$

gegeben. Dabei bezeichnet L_h die Approximation von L und v_h die Näherungslösung zu v. Der Parameter h deutet hier die Diskretisierung an.

Wird die exakte Lösung u von (5.1) in das Näherungsverfahren (5.2) eingesetzt, dann wird sich im Allgemeinen ein Fehler ergeben: Das **Residuum** oder der **Defekt**. Die entsprechende diskretisierte exakte Lösung wird mit u_h bezeichnet. Konsistent bedeutet dann die folgende Aussage:

Ein numerisches Verfahren ist **konsistent** und besitzt die **Approximations- oder Konsistenzordnung** p, wenn für den lokalen Diskretisierungsfehler

$$||L_h u_h - f_h|| \leq Ch^p \tag{5.3}$$

mit einer von h unabhängigen Konstanten C gilt.

Bemerkungen:

1. Mit h wurde ganz allgemein ein Diskretisierungsparameter eingeführt. Für ein mehrdimensionales Problem gibt es im Allgemeinen mehrere Diskretisierungsparameter, z.B. die Raum- und Zeitschrittweite Δx und Δt. Gilt für den lokalen Diskretisierungsfehler in (5.3)

   $$||L_h u_h - f_h|| = O(\Delta x^p, \Delta t^q) \,, \tag{5.4}$$

   dann ist dieses Verfahren konsistent mit der Konsistenzordnung p bezüglich der Raumapproximaiton und konsistent mit der Konsistenzordnung q bezüglich der Zeitapproximation.

2. Zwischen den einzelnen Näherungsverfahren gibt es formale Unterschiede. Ein Differenzenverfahren liefert Näherungswerte an der verschiedenen Gitterpunkten. Dies bedeutet, dass die Werte der exakten Lösung

an diesen diskreten Punkten in die Näherungsformel eingesetzt werden. Die exakte Lösung muss in diesem Fall auf das Gitter eingeschränkt werden. Bei einem Finite-Elemente-Verfahren kann man dagegen die exakte kontinuierliche Lösung in das Näherungsverfahren direkt einsetzen.

3. Die Bedingung der Konsistenz wird hier allgemein mit einer Norm $||.||$ formuliert, da es sich hier um mehrere Gleichungen und mehrere Werte handeln kann. Bei einer skalaren gewöhnlichen Differenzialgleichungen in Kapitel 2 tauchte hier der Betrag auf.

Konsistenz ist eine Beziehung zwischen dem Näherungsverfahren und der Ausgangsgleichung. Sie bedeutet, dass das Näherungsverfahren die mathematische Gleichung so approximiert, dass der Fehler entsprechend der Konsistenzordung gegen Null strebt, wenn die Diskretisierungsparameter gegen Null streben. Die Voraussetzung für eine solche Konvergenz ist natürlich immer, dass die Lösung die nötigen Stetigkeitsvoraussetzungen erfüllt.

Eine weitere wichtige Eigenschaft für ein brauchbares numerisches Verfahren ist die **Stabilität**. Wenn wir in diesem Zusammenhang von Stabilität sprechen, dann von der Stabilität des Näherungsverfahrens. Es ist dabei immer vorausgesetzt, dass das Ausgangsproblem stabil ist. Hat das Ausgangsproblem insbesondere die Eigenschaft, dass Störungen der Lösung nicht oder nur sehr langsam anwachsen, dann können wir dieses Problem dazu verwenden, um die Eigenschaft der Stabilität des Näherungsverfahrens zu untersuchen. Man nennt solche Probleme auch gut konditionierte Probleme. Stabilität des Näherungsverfahrens meint also ganz allgemein, dass das numerische Verfahren die Unempfindlichkeit des Ausgangsproblems bezüglich kleiner Störungen erhält: Fehler wie Rundungs- oder Diskretisierungsfehler werden dann nicht beliebig anwachsen.

Sei die Näherungslösung eines stabilen Problems mit v_h und eine Näherungslösung des gestörten Problems mit $\tilde{v}_h$ bezeichnet. Ein numerisches Lösungsverfahren heißt dann **stabil**, falls

$$\|v_h - \tilde{v}_h\| \leq C \tag{5.5}$$

mit einer von h unabhängigen Konstanten C gilt, wenn die Schrittweite gegen Null strebt.

Vorausgesetzt ist, dass die exakte Lösung eine Stabilitätsbedingung der Form (5.5) erfüllt. Sonst macht es keinen Sinn, eine solche Eigenschaft von dem numerischen Verfahren zu fordern. Je nach Problem wird die Stabilitätskonstante z.B. von der Zeit abhängen.

Schon bei gewöhnlichen Differenzialgleichungen hat man die Situation, dass die Stabilitätsanforderungen an das Näherungsverfahren problemabhängig sind. So muss ein Verfahren, welches auf steife gewöhnliche Differenzialgleichungen angewandt wird, stärkere Stabilitätsbedingungen erfüllen als eines für Standardprobleme. Für partielle Differenzialgleichungen, bei denen Eigenschaften und Lösungsstruktur weitaus vielfältiger sind, spielt dies noch in größerem Maße eine Rolle.

Bemerkungen:

1. Oft gibt es für ein numerisches Verfahren auch gewisse einschränkende Bedingungen an die Diskretisierungsparameter, wie dies schon bei den gewöhnlichen Differenzialgleichungen auftritt. Man spricht dann von einem bedingt stabilen Verfahren.

2. Stabilitätsaussagen für numerische Verfahren können auch Forderungen sein, dass gewisse Eigenschaften der exakten Lösung durch das Näherungsverfahren reproduziert werden. Ein Beispiel ist die Bedingung der Positivität für die Dichte oder die Erfüllung der Entropiebedingung in einem numerischen Verfahren in der Gasdynamik.

3. Stabilitätsaussagen bei nichtlinearen Probleme sind oft sehr schwierig nachzuweisen.

Von der **Konvergenz eines Näherungsverfahrens** spricht man, wenn die Näherungslösung bei einer gegen Null strebenden Schrittweite gegen die exakte Lösung konvergiert.

Sei e_h der **globale Diskretisierungsfehler**

$$e_h = v_h - u_h \ ,$$

dann bedeutet Konvergenz

$$\|e_h\| \leq Ch^q \ , \tag{5.6}$$

wobei $\|.\|$ eine geeignete Norm bezeichnet. Man nennt q die **Konvergenzordnung** des Verfahrens.

Bemerkung: Die Wahl der Norm $\|.\|$ in der Definition der Konvergenz eines Verfahrens ist problemabhängig. Punktweise Konvergenz ist so definiert, dass in (5.6) der einfache Betrag steht und der Grenzwert für jeden Punkt existieren muss. Die Bedingung der gleichmäßigen Konvergenz erhält man, wenn $\|.\|$ die Maximumsnorm ist.

Die Begriffe Konsistenz, Stabilität und Konvergenz sind nicht unabhängig voneinander. Hat man ein konsistentes numerisches Verfahren, dann weiß man, dass der lokale Diskretisierungsfehler klein wird, wenn die Schrittweiten klein werden. Die Stabilität des Verfahrens bedeutet, dass der Fehler zwischen der Näherungslösung und der exakten Lösung nicht unbeschränkt anwachsen kann. Plausibel erscheint dann, dass aus Konsistenz und Stabilität die Konvergenz der Näherungslösung gegen die exakte Lösung folgt:

Konsistenz + Stabilität = Konvergenz

Dies nennt man das **Äquivalenztheorem von Lax** und kann man in einem recht allgemeinen Rahmen beweisen. Erläuterungen und ein Beweis findet sich z.B. in dem Buch von Thomas [63]

Im Folgenden werden wir in den einzelnen Abschnitten und für die einzelnen Verfahren und die einzelnen Typen von Differenzialgleichungen einige Stabilitäts- und Konsistenzaussagen herleiten und diskutieren.

Theoretischen Aussagen wie die Stabilität lassen sich oftmals nicht für komplexe Anwendungen beweisen. Man ist dann auf die Aussagen für gewisse einfache Standardgleichungen angewiesen. Diese Aussagen liefern Hinweise über die entscheidenden Konstruktionsprinzipien und geben Information über die wichtigen Anforderungen für die entsprechende Klasse von mathematischen Gleichungen. Die für die Standardprobleme gefundenen „guten“ numerischen Verfahren müssen dann auf die komplexeren Gleichungen oder Systeme geeignet übertragen werden, für die dann der Nachweis der theoretischen Aussagen nicht mehr möglich ist.

Ein wesentlicher Aspekt der numerischen Simulation ist die Validierung der numerischen Verfahren und letztendlich die Bewertung der numerischen Ergebnisse. Fehlen die theoretischen Aussagen für das zu komplizierte mathematische Modell, müssen neben den theoretiechen Resultate für die einfachen Testprobleme zumindest einige praktische Untersuchungen ausgeführt werden. Hier hat der Entwickler eines Rechenprogramms oder dann auch der Anwender verschiedene Möglichkeiten, Aussagen wie Stabilität oder Konvergenz mit dem Rechenprogramm praktisch zu überprüfen.

Ein Validierungsproblem für die Stabilität kann ein bekannt stabiles Problem sein, von dem numerische oder experimentelle Daten in der Literatur verfügbar sind. Hier kann man dann bei den eigenen numerischen Simulationen auch Störungen der Anfangswerte oder der rechten Seite mit einbringen, um zu sehen, wie diese anwachsen. Bei der praktischen Untersuchung der Konvergenz kann man die Diskretisierung verfeinern und analysieren, wie der Fehler sich

verringert. Dies sind dann natürlich keine Beweise, sondern lediglich Indizien für das entsprechende Verhalten des numerischen Lösungsverfahren.

5.2 Die Diskretisierung des Rechengebietes

Die Approximationen führen das kontinuierliche Problem auf ein endliches diskretes Problem zurück, welches dann mit dem Computer gelöst werden kann. Alle Näherungsansätze gründen sich dabei auf eine Diskretisierung des Rechengebietes. Bei den Differenzenverfahren werden direkt Näherungswerte an speziellen Punkten gesucht, indem die Ableitung durch einen Differenzenquotienten an diesen Punkten ersetzt wird.

Bei den Finite-Element-Verfahren steht zunächst die Approximation der Lösungsfunktion durch eine einfachere Funktion, definiert mit Hilfe von endlich vielen Basisfunktionen, im Vordergrund. Die Definition der Basisfunktionen greift dann aber wieder auf eine Diskretisierung des Raumes zurück. Zum einen hat die Diskretisierung des Rechengebietes, man spricht hier von Rechengittern oder kurz von Gittern, diese zentrale Stellung für alle hier betrachteten Näherungsverfahren, zum anderen wird die Gittererzeugung in mehreren Raumdimensionen deutlich aufwändiger, so dass wir die Grundlagen in diesem Kapitel vor der Einführung der numerischen Methoden behandeln.

5.2.1 Beschreibung technischer Gebiete

Bei den gewöhnlichen Differenzialgleichungen, etwa bei der numerischen Behandlung von Randwertproblemen, liegt als Rechengebiet ein Intervall $[a, b]$ vor. Die Definition von Gitterpunkten, auch Stützstellen genannt, ist einfach. Es wurde hier meist eine äquidistante Verteilung der Gitterpunkte betrachtet, weil damit die numerischen Verfahren und die Algorithmen die einfachste Gestalt annehmen. Möchte man ein Problem in einem Rechteckgebiet lösen, dann ist dies die entsprechende zweidimensionale Erweiterung dieses Konzepts.

Wir gehen im Folgenden zunächst von dem einfachsten Fall aus: Das Rechengebiet ist als ein Rechteck

$$R = [a, b] \times [c, d] \tag{5.7}$$

gegeben. Es wird äquidistant in x-Richtung und in y-Richtung unterteilt. Dabei werden die konstanten Schrittweiten Δx und Δy genannt und nach

den Formeln

$$\Delta x = \frac{b-a}{I}\,, \qquad \Delta y = \frac{d-c}{J} \tag{5.8}$$

berechnet. Dabei ist I und J die Anzahl der Gitterintervalle in x- bzw. in y-Richtung. Die Menge der Gitterpunkte ist somit durch die **inneren Gitterpunkte**

$$\begin{aligned}(x_i, y_j) \quad \text{mit} \;\; x_i &= a + i\Delta x\,, \qquad i = 1, 2, \ldots, I-1\,, \\ y_j &= c + j\Delta y\,, \qquad j = 1, 2, \ldots, J-1\end{aligned} \tag{5.9}$$

und die **Gitterpunkte am Rand**

$$(a, y_j)\,, \quad (b, y_j)\,, \quad (x_i, c)\,, \quad (x_i, d)\,, \qquad i = 0, 1, \ldots, I\,, \quad j = 0, 1, \ldots, J \tag{5.10}$$

gegeben. In Abbildung 5.1 ist ein solches Gitter für die Anzahl I=5 und J=3 von Gitterintervallen aufgezeichnet.

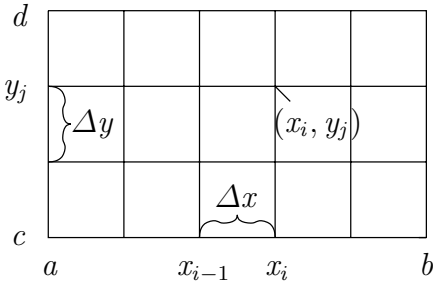
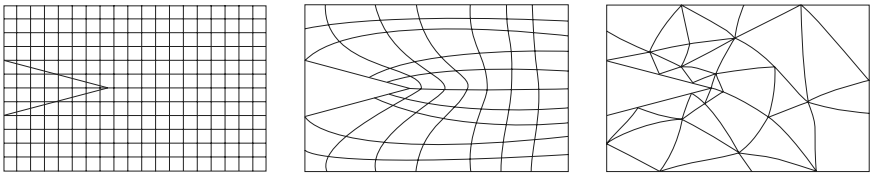


Abb. 5.1. Diskretisierung des Rechteckgebietes

Soll zum Beispiel die Potenzialgleichung in einem Rechteckgebiet mit Hilfe eines Differenzenverfahrens gelöst werden, so wird man ein achsenparalleles Gitter einführen. Bei der numerischen Approximation mit Hilfe eines Differenzenverfahren wird dann die gesuchte Potenzialverteilung an den diskreten Gitterpunkten näherungsweise berechnet. Aber schon bei Konfigurationen mit Kanten oder gekrümmten Elektroden ist ein achsenparalleles Gitter dem Problem nicht mehr angemessen.

Betrachten wir ein Gebiet mit einem krummlinigen Rand. Durch ein achsenparalleles Gitter wird dieser Rand treppenförmig approximiert. Jede einspringende Ecke führt zu Feldüberhöhungen, die im physikalischen Problem nicht vorhanden sind. Daher sind in der Regel achsenparallele Gitter nicht geeignet, um technisch relevante Geometrien numerisch zu beschreiben.

Die Abbildung 5.2 enthält für ein Gebiet mit einspringender Kante drei unterschiedliche Gittervarianten. Die wichtigsten Eigenschaften dieser verschiedenen Typen in Abbildung 5.2 sind:



(a) achsenparallel (b) randangepasst (c) unstrukturiert

Abb. 5.2. Schematische Gitter für ein Gebiet mit Kante

(a) besteht aus achsenparallelen Gitterlinien. Dieses Gitter ist hier äquidistant und das einfachste Gitter. Entsprechend der achsenparallelen Geraden können wir die Gitterpunkte in einfacher Art und Weise bezeichnen und nummerieren. Betrachten wir die Gitterlinien $x = x_i$ und $y = y_j$ dann ist der Schnittpunkt dieser Geraden der Gitterpunkt (x_i, y_j). Liegt er im Inneren des Rechengebietes, dann hat er einen rechten, linken, oberen und unteren Nachbarn, deren Koordinaten durch (x_{i+1}, y_j), (x_{i-1}, y_j), (x_i, y_{j+1}) bzw. (x_i, y_{j-1}) automatisch bekannt sind. Damit sind zum Beispiel Differenzenverfahren in einfacher Weise übertragbar. Aber selbst bei hoher Gitterauflösung muss die Kante stufenweise approximiert werden, was zwangsläufig eine Fehlerquelle mit sich bringt. $\rightsquigarrow$ *Hohe Rechenzeiten und dennoch große Diskretisierungsfehler.*

Es gibt Ansätze, die Schnittkurven des Gitters mit dem Hindernis zu berechnen und die Geometrie in der Approximation auf dem kartesischen Gitter zu berücksichtigen. Wir werden diesen speziellen Ansatz hier nicht aufgreifen und verweisen auf das Buch von Schwarz und Köckler [55].

(b) ist ein **randangepasstes strukturiertes Gitter**: Die Gitterlinien passen sich dem Rand des Gebietes an. Randangepasste strukturierte Gitter besitzen ebenfalls eine *logische Struktur*: Das Gitter kann durch eine eindeutig umkehrbare Transformation auf ein kartesisches Gitter abgebildet werden. Somit hat jeder innere Punkt einen rechten, linken, oberen und unteren Nachbarn in dem logischen kartesischen Gitter. Selbst bei einer starken Krümmung des randangepassten Gitters kann diese Bezeichnung vom kartesischen Gitter auf das Urbild übertragen werden und die Koordinaten der Gitterpunkte können einfach gefunden werden. Durch diese Eigenschaft können Differenzenverfahren übertragen und die entsprechenden Differenzialgleichungen diskretisiert werden. $\rightsquigarrow$ *Kleine Diskretisierungsfehler, einfache Datenstrukturen.*

Für komplizierte Geometrien ist aber die Erzeugung eines guten strukturierten Gitters oft sehr schwierig und erfordert viel Erfahrung und Zeit.

(c) ist ein **unstrukturiertes Gitter**. Mit diesem Gittertyp, wie hier einem Dreiecksgitter können komplizierte Gebiete erfasst werden. Es gibt keine Transformation auf ein kartesisches Gitter. Die Datenstruktur der unstrukturierten Gitter ist daher aufwändiger zu verwalten, Differenzenverfahren können **nicht** angewendet werden. Zur Berechnung der Lösung der Differenzialgleichungen benötigt man dann Verfahren, wie Finite-Element- oder Finite-Volumen-Verfahren. $\rightsquigarrow$ *Sehr variabel, komplexe Geometrien, aber aufwändigere Datenstrukturen.*

Wir werden uns im Folgenden noch etwas genauer mit strukturiertne randangepassten Gittern beschäftigen.

5.2.2 Erzeugung von randangepassten Gittern

Die Grundidee bei der Erzeugung von strukturierten randangepassten Gittern liegt darin, dass das physikalische Gebiet Ω in der (x, y)-Ebene auf ein "logisches" Rechteck Ω' in der (ξ, η)-Ebene abgebildet wird. In diesem Rechteck Ω' wird ein äquidistantes Gitter eingeführt. Dieses äquidistante Gitter wird dann wiederum auf das physikalische Gebiet zurück transformiert.

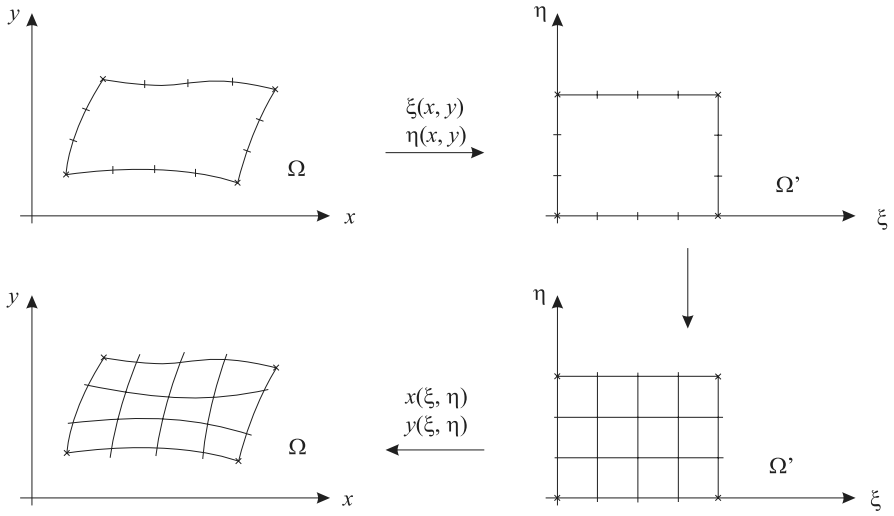


Abb. 5.3. Transformationen von Ω nach Ω' und zurück

Gittererzeugung: Seien $\xi(x, y)$ und $\eta(x, y)$ die Abbildungen des physikalischen Gebietes Ω in der (x, y)-Ebene auf das logische Rechteck Ω' in der

(ξ, η)-Ebene. D.h. jedem Punkt (x, y) im Gebiet Ω wird ein Punkt (ξ, η) im logischen Rechteck zugeordnet. Als Transformationsabbildungen wählen wir sowohl für $\xi(x, y)$ als auch für $\eta(x, y)$ die Potenzialgleichung, weil sich die Gitterlinien ähnlich einer Potenzialverteilung verhalten sollen:

$$\Delta\xi(x, y) = P, \quad \Delta\eta(x, y) = Q \; . \tag{5.11}$$

Um auf die Gitterpunkte im physikalischen Gebiet zu kommen, benötigen wir auch die Rücktransformationen $x(\xi, \eta)$ und $y(\xi, \eta)$, die jeden Punkt aus dem logischen Gebiet auf einen Punkt im physikalischen Gebiet Ω abbilden. Dann wird jede Gitterlinie aus dem logischen Gebiet auf eine Gitterlinie des physikalischen Gebietes abgebildet. Die Umkehrfunktionen $x(\xi, \eta)$ und $y(\xi, \eta)$ vom logischen Rechteck auf das physikalische Gebiet erfüllen die folgenden Gleichungen:

$$\alpha x_{\xi\xi} - 2\beta x_{\xi\eta} + \gamma x_{\eta\eta} + J^2 x_\xi P + J^2 x_\eta Q = 0 \; , \tag{5.12}$$

$$\alpha y_{\xi\xi} - 2\beta y_{\xi\eta} + \gamma y_{\eta\eta} + J^2 y_\xi P + J^2 y_\eta Q = 0 \tag{5.13}$$

mit den Größen:

$$J = x_\xi y_\eta - y_\xi x_\eta \; , \tag{5.14}$$

$$\alpha = x_\eta^2 + y_\eta^2 \; , \tag{5.15}$$

$$\beta = x_\xi x_\eta + y_\xi y_\eta \; , \tag{5.16}$$

$$\gamma = x_\xi^2 + y_\xi^2 \; . \tag{5.17}$$

Bemerkung: Mit den noch frei wählbaren Parametern P und Q besitzt man Parameter, um das Gitter an bestimmten Punkten zusammen zu ziehen bzw. Gitterlinien zu entzerren. Im einfachsten Fall setzt man sie auf Null.

5.3 Bemerkungen

Die Qualitätskriterien wie Konsistenz und Konvergenz mit den entsprechenden Ordnungen und die Stabilität sind wichtige Informationen, um sich für ein numerisches Verfahren zu entscheiden. Diese Entscheidung hängt in hohem Maße vom Problem und den Eigenschaften der Lösung ab. Wir wollen die Bedeutung dieser Begriffe für praktische Anwendungen im Folgenden noch einmal kurz zusammenfassen.

Je höher die Konsistenzordnung des Verfahrens ist, desto schneller werden die Fehler kleiner, falls die Schrittweite gegen Null strebt. Bei einem Verfahren zweiter Ordnung bedeutet dies: Wird die Schrittweite halbiert, dann

kann man erwarten, dass der Konsistenzfehler um den Faktor 1/4 kleiner wird, allgemein um $1/(2^q)$, wenn q die Konsistenzordnung bezeichnet. Diese Aussagen sind allerdings asymptotische Aussagen für den Fall, dass die Schrittweite gegen Null strebt. Bei einer endlichen Schrittweite kann dieses Verhalten durch die Änderung der anderen Faktoren im Fehlerterm gestört sein.

Geben wir eine gewünschte Genauigkeit der Lösung vor, dann wird das Verfahren mit einer höheren Konvergenzordnung eine geringere Anzahl von Gitterpunkten benötigen. Es kommt jetzt auf den Rechenaufwand des Verfahrens höherer Ordnung an, ob die Rechnung tatsächlich effizienter ist. Überall dort, wo eine hohe Genauigkeit gefordert wird, sollte ein numerisches Verfahren hoher Ordnung überlegen sein.

Der augenblickliche Standard bei den numerischen Verfahren für partielle Differenzialgleichungen heißt Verfahren 2. Ordnung. Diese Verfahren stehen in den folgenden Kapiteln auch im Mittelpunkt. Ein Verfahren oder eine Approximation 1. Ordnung wird nur in Ausnahmefällen angewandt, ein Beispiel ist hier die Ermittlung von stationären oder fast stationären Lösungen mit Hilfe der zeitabhängigen Gleichungen. Hier ändert sich die Näherungslösung sehr wenig in der Zeit, so dass die Fehler hier sehr klein sind. Wir kommen auf dieses Beispiel später zurück. Liegen keine Kenntnisse über die Konvergenzordnung des numerischen Verfahrens vor, dann kann man diese experimentell ermitteln. Man nimmt ein Problem mit bekannter Lösung als Testproblem und berechnet die Näherungslösungen für verschiedene Schrittweiten und vergleicht dann die Fehler.

Welche Art von Stabilität man benötigt, hängt wiederum vom Problem ab. Z.B. können bei einer Strömungssimulation sehr starke lokale Unterschiede in Druck und Dichte auftreten, so dass man ein sehr robustes stabiles Verfahren benötigt. Würde man dieses Verfahren zur Simulation von akustischen Wellen mit kleinen Amplituden, wie sie z.B. bei der Lärmausbreitung vorkommen, anwenden, so ist nach kurzer Zeit die Lösung sehr schlecht aufgelöst. Zur effizienten Simulation der Wellenausbreitung benötigt man dann ein anderes Verfahren. Bei praktischen Anwendungen ist es immer schwierig zu entscheiden, ob eine beobachtete Instabilität einen physikalischen oder einen numerischen Grund besitzt. Hier muss man dann die Abhängigkeit der Instabilität von den Diskretisierungparameter oder den Einfluss von kleinen Störungen untersuchen.

Im Bereich der numerischen Simulation ist die Gittererzeugung für praktische Anwendungen in drei Raumdimensionen in der Industrie ein sehr bedeutender Arbeitsbereich. Eine gute numerische Berechnung benötigt als Grundvoraussetzung ein gutes Gitter in dem Sinne, dass die einzelnen Gitterzellen nicht zu verzerrt sind. Die Qualität des Gitters beeinflusst sehr stark die Fehler

und damit die Güte der numerischen Ergebnisse. Bei technischen Anwendungen wie die Umströmung und Lärmentwicklung des Fahrwerks eines Flugzeugs oder die Motorinnenraumströmung bei einem Kraftfahrzeug stellen sehr große Anforderungen an die Vernetzung solcher komplexer Geometrien. Die Erzeugung eines guten strukturierten, Rand angepassten Gitters gelingt hier nur in einer block-strukturierten Art und Weise. Hier wird das Rechengebiet in einzelne Blöcke aufgeteilt und in diesen jeweils getrennt voneinander ein strukturiertes Gitter erzeugt. Die verschiedenen Blöcke werden dann für eine Rechnung miteinander gekoppelt. Es kann gut sein, dass die Erzeugung eines guten Gitters für komplizierte Geometrien für einen Spezialisten mehrere Wochen in Anspruch nimmt. Als Werkzeug werden kommerzielle Gittererzeuger benutzt, welche aus den entsprechenden CAD-Daten für die Geometrie ein Rechengitter erstellen. Die Hauptarbeit benötigt dann die Überarbeitung und die Optimierung dieses Gitters, um große Diskretisierungsfehler oder auch Stabilitätsprobleme zu vermeiden. Ein Standardwerk für den Bereich der Gittererzeugung ist das Buch von Thompson, Sony und Weatherill [64].

Die Erzeugung von unstrukturierten Gittern, wie z.B. Tetraeder-Gitter, geht praktisch automatisch und in wenigen Stunden selbst für komplizierte Geometrien. Durch die Gitterstruktur sind die numerischen Simulationen langsamer. Man hat allerdings noch die Möglichkeit, Rechenzeit einzusparen durch ein adaptives Gitter. Hier wird die Gitterauflösung durch das Verhalten der Lösung oder der Größe des Fehlers im Laufe der Rechnung gesteuert.

6 Differenzenverfahren

Bei der numerischen Lösung von gewöhnlichen Differenzialgleichungen oder bei der numerischen Differenziation haben wir die **Differenzenverfahren** schon eingeführt. Die grundlegende Idee bei dieser Klasse von Methoden ist, die in der Differenzialgleichung auftretenden Ableitungen durch Differenzenquotienten zu ersetzen. Diese Vorgehensweise wird motiviert durch die Definition der Ableitung. Die Ableitung einer Funktion $u = u(x)$ im Punkt x ist definiert durch

$$\frac{du}{dx} = \lim_{\Delta x \to 0} \frac{u(x + \Delta x) - u(x)}{\Delta x} .$$

Ist Δx klein aber ungleich Null, dann sollte der rechte Ausdruck eine Näherung für die Ableitung darstellen. Diese Idee lässt sich von den gewöhnlichen direkt zu den partiellen Differenzialgleichungen übertragen, indem diese durch entsprechende Differenzenquotienten ersetzt werden. Differenzenverfahren sind damit die einfachsten Methoden zur Lösung von partiellen Differenzialgleichungen.

Man wird erwarten, dass die Näherung der Ableitung durch einen Differenzenquotienten besser wird, wenn die Schrittweite Δx verkleinert wird. Dies kann man mit Hilfe von Taylor-Entwicklungen nachvollziehen. Mit Hilfe der Methode des Taylor-Abgleichs lassen sich Differenzenformeln mit unterschiedlicher Konsistenzordnung konstruieren. Wir werden zunächst die Differenzenquotienten, welche im Kapitel 1.6, Numerische Differenziation, hergeleitet wurden, auf partielle Ableitungen übertragen.

Wir betrachten im Folgenden u als eine Funktion zweier unabhängiger Variablen x und y oder auch x und t. Wir nehmen an, dass $u = u(x, y)$ oder $u = u(x, t)$ beliebig oft stetig differenzierbar ist. Dann können wir $u(x + \Delta x, y)$ in eine Taylor-Entwicklung um den Punkt (x, y) entwickeln:

$$u(x + \Delta x, y) = u(x, y) + \Delta x\, u_x + \frac{\Delta x^2}{2!}\, u_{xx} + \frac{\Delta x^3}{3!}\, u_{xxx} + \dots .$$

Das Argument (x, y) haben wir bei den Ableitungen auf der rechten Seite weggelassen. Die Variable y kommt hier nur als Parameter vor. Dies entspricht der Vorgehensweise bei der partiellen Ableitung bei der stets die an-

deren Variablen festgehalten werden. Wir erhalten somit wie im Falle einer Variablen

$$u_x(x,y) = \frac{u(x+\Delta x,y) - u(x,y)}{\Delta x} + O(\Delta x) \; .$$

Der Differenzenquotient

$$\frac{u(x+\Delta x,y) - u(x,y)}{\Delta x}$$

ist somit eine Approximation der partiellen Ableitung u_x, wobei der Approximations- oder Konsistenzfehler der Term $O(\Delta x)$ ist. Führen wir die Indizes i und j ein, um die Näherungslösung an dem diskreten Punkt (x_i, y_j) zu bezeichnen, erhalten wir

$$u_x(x_i,y_j) = \frac{u_{i+1,j} - u_{i,j}}{\Delta x} + O(\Delta x) \; . \tag{6.1}$$

Man nennt diesen Ausdruck den **rechtsseitigen Differenzenquotienten** von u an dem Punkt (x_i, y_j).

Der **linksseitige Differenzenquotient** von u nach x ergibt sich aus der Taylor-Entwicklung von $u(x-\Delta x, y)$ um den Punkt (x, y),

$$u(x-\Delta x,y) = u(x,y) - \Delta x \; u_x + \frac{\Delta x^2}{2!}\, u_{xx} - \frac{\Delta x^3}{3!}\, u_{xxx} + \cdots - \ldots \; , \tag{6.2}$$

durch Auflösung nach u_x:

$$u_x(x_i,y_j) = \frac{u_{i,j} - u_{i-1,j}}{\Delta x} + O(\Delta x) \; . \tag{6.3}$$

Durch Kombination der beiden einseitigen Differenzenquotienten erhalten wir den **zentralen Differenzenquotienten** in der Form

$$u_x(x_i,y_j) = \frac{u_{i+1,j} - u_{i-1,j}}{2\Delta x} + O(\Delta x^2) \; . \tag{6.4}$$

Dieser ist eine Approximation zweiter Ordnung für die Ableitung u_x am Punkt (x_i, y_j), da der führende Fehlerterm mit Δx^2 gegen Null strebt, wenn die Schrittweite Δx gegen Null geht. Verschiedene andere Differenzenapproximationen sind im Kapitel 1.6 über die numerische Differenziation hergeleitet und lassen sich entsprechend auf partielle Ableitungen übertragen.

Für die zweite partielle Ableitung nach x ergibt sich

$$u_{xx}(x_i, y_j) = \frac{u_{i+1,j} - 2u_{i,j} + u_{i-1,j}}{\Delta x^2} + O(\Delta x^2) \tag{6.5}$$

mit einem Approximationsfehler zweiter Ordnung.

Ganz analog ergibt sich dies auch für die y-Richtung, so haben wir hier die **einseitigen Ableitungen**

$$u_y(x_i, y_j) = \frac{u_{i,j+1} - u_{i,j}}{\Delta y} + O(\Delta y) \ ,$$

$$u_y(x_i, y_j) = \frac{u_{i,j} - u_{i,j-1}}{\Delta y} + O(\Delta y)$$

als Approximationen erster Ordnung. Der zentrale Differenzenquotient bezüglich y ergibt sich als Approximation zweiter Ordnung zu

$$u_y(x_i, y_j) = \frac{u_{i,j+1} - u_{i,j-1}}{2\Delta y} + O(\Delta y^2) \ .$$

Die zentrale Formel für die zweite Ableitung in y-Richtung lautet

$$u_{yy}(x_i, y_j) = \frac{u_{i,j+1} - 2u_{i,j} + u_{i,j-1}}{\Delta y^2} + O(\Delta y^2) \ . \tag{6.6}$$

Bemerkung: Ist die zweite unabhängige Variable die Zeit t, dann spricht man bei den einseitigen Differenzenquotienten meist von dem **vorwärts-genommenen** und dem **rückwärts-genommenen** Differenzenquotienten.

Differenzenquotienten für die partiellen Ableitungen erster Ordnung mit höherer Genauigkeit erhält man analog durch die Übertragung der Formeln im Kapitel 1.6 über die numerische Differenziation. Bei Ableitungen höherer Ordnung in x- oder y-Richtung kann dies ebenso leicht übertragen werden. Bei den gemischten Ableitungen wird dies etwas schwieriger. Hier ist es günstiger, die in 1.6 angegebene Methode der Herleitung von Differenzenquotienten auf den mehrdimensionalen Fall zu übertragen. Es wird ein zweidimensionales Interpolationspolynom berechnet und dieses dann entsprechend abgeleitet. Für weitere Informationen über partielle Differenzenquotienten sei auf das Buch von Marsal [41] verwiesen.

Um ein Differenzenverfahren anzuwenden, startet man mit der Diskretisierung des Rechengebietes. Wir gehen im Folgenden zunächst von dem einfachsten Fall aus: Das Rechengebiet ist als ein Rechteck

$$R = [a, b] \times [c, d] \tag{6.7}$$

gegeben. Wir übernehmen hier die Bezeichnungsweise von Kapitel 5.2. Die Schrittweiten lauten nach (5.7)

$$\Delta x = \frac{b-a}{I}\,, \qquad \Delta y = \frac{d-c}{J}\,, \tag{6.8}$$

I und J ist die Anzahl der Gitterintervalle in x- bzw. in y-Richtung, die inneren Gitterpunkte sind

$$\begin{aligned} (x_i, y_j) \quad \text{mit} \;\; x_i &= a + i\Delta x\,, \qquad i = 1, 2, \ldots, I-1 \\ y_j &= c + j\Delta y\,, \qquad j = 1, 2, \ldots, J-1 \end{aligned} \tag{6.9}$$

und die Gitterpunkte am Rand

$$(a, y_j)\,,\ (b, y_j)\,,\ (x_i, c)\,,\ (x_i, d)\,, \quad i = 0, 1, \ldots, I\,, \quad j = 0, 1, \ldots, J \tag{6.10}$$

gegeben. In Abbildung 5.1 ist ein solches Gitter aufgezeichnet.

Nach der Definition der Gitterpunkte werden die Ableitungen in der Differenzialgleichung durch entsprechende Differenzenquotienten ersetzt. Man erhält eine Rechenvorschrift für die Näherungswerte an den Gitterpunkten, eine Differenzengleichung. Statt der Differenzialgleichung wird eine endliche Anzahl von Differenzengleichungen gelöst. Dies wollen wir uns im Folgenden für die einzelnen Typen von Differenzialgleichungen bzw. deren einfachste Vertreter genauer anschauen.

6.1 Elliptische Differenzialgleichungen

Als den einfachsten Vertreter einer elliptischen Differenzialgleichung betrachten wir die Poisson-Gleichung

$$u_{xx} + u_{yy} = f\,, \tag{6.11}$$

welche im Kapitel 4.2 vorgestellt wird.

(1.) Der **erste Schritt** bei der Definition eines Differenzenverfahrens ist die **Diskretisierung des Rechengebietes**. Eine äquidistante Diskretisierung eines rechteckigen Rechengebietes wird im vorigen Abschnitt beschrieben und

ist in Abbildung 5.1 skizziert. Diese Diskretisierung und die Bezeichnungen in (5.7) bis (5.10) werden im Folgenden übernommen.

(2.) Der **zweite Schritt** bei der Definition des Differenzenverfahrens ist die **Wahl der Differenzenquotienten**. Als die einfachste Approximation erscheinen die zentralen Differenzenquotienten für die zweiten Ableitungen, wie sie in (6.5) und (6.6) abgeleitet sind. Ersetzen wir die Ableitungen in der Poisson-Gleichung durch diese Differenzenquotienten, erhalten wir an jedem inneren Gitterpunkt (x_i, y_j) die Differenzengleichung

$$\frac{u_{i+1,j} - 2u_{i,j} + u_{i-1,j}}{\Delta x^2} + \frac{u_{i,j+1} - 2u_{i,j} + u_{i,j-1}}{\Delta y^2} = f_{i,j} \tag{6.12}$$

mit $f_{i,j} = f(x_i, y_j)$.

In diese Differenzengleichung gehen neben dem Gitterpunkt (x_i, y_j) die vier benachbarten Gitterpunkte

$$(x_{i-1}, y_j)\,, \quad (x_{i+1}, y_j)\,, \quad (x_i, y_{j-1})\,, \quad (x_i, y_{j+1})$$

ein. Grafisch dargestellt ist diese Abhängigkeit in Abbildung 6.1.

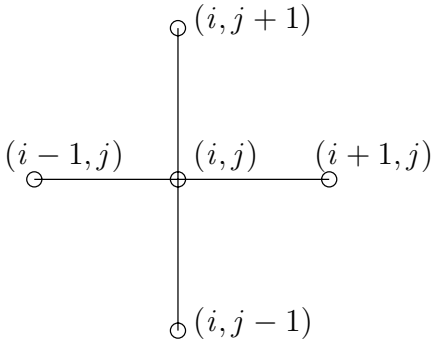


Abb. 6.1. 5-Punkte-Differenzenstern

Ein solches Diagramm, in dem die in der Differenzengleichung benutzten umliegenden Gitterpunkte darstellt, bezeichnet man als **Differenzenstern** oder auch als das **Differenzenmolekül** des Verfahrens. Mit dem zentralen Differenzenquotienten zweiter Ordnung ergibt sich der sogenannte 5-Punkte-Stern zur Approximation des Laplace-Operators.

Diese Differenzengleichung können wir an jedem inneren Gitterpunkt (5.9) aufstellen. Wir haben somit $(I-1)(J-1)$ solcher Differenzengleichungen.

(3.) Damit ist man am **dritten Schritt** des Differenzenverfahrens angekommen, in dem die Differenzengleichungen geeignet sortiert werden. Wir stellen

etwas um und erhalten die Differenzengleichungen an jedem inneren Gitterpunkt (x_i, y_j) mit $i = 1, 2, \ldots, I-1$ und $j = 1, 2, \ldots, J-1$ in der Form

$$\frac{1}{\Delta x^2} u_{i-1,j} - 2\left(\frac{1}{\Delta x^2} + \frac{1}{\Delta y^2}\right) u_{i,j} + \frac{1}{\Delta x^2} u_{i+1,j} + \\ + \frac{1}{\Delta y^2} u_{i,j-1} + \frac{1}{\Delta y^2} u_{i,j+1} = f_{i,j} \; . \tag{6.13}$$

Dies ist die **diskrete Laplace-Gleichung**. Eine gewisse Sonderrolle spielen die Gitterpunkte in direkter Nachbarschaft mit Randpunkten. Nach dem Kapitel 4.1 ist die Lösung der Poisson-Gleichung nur dann eindeutig bestimmt, wenn Randwerte vorgeschrieben werden. Wir nehmen an, dass das physikalische Problem Dirichlet-Randwerte besitzt:

$$\begin{aligned} u(a, y) &= u_a(y) \\ u(b, y) &= u_b(y) \end{aligned} \qquad \text{für } y \in [c, d] \, , \\ \begin{aligned} u(x, c) &= u_c(x) \\ u(x, d) &= u_d(x) \end{aligned} \qquad \text{für } x \in [a, b] \, . \tag{6.14}$$

Wir schauen uns den linken Rand genauer an. Die Gitterpunkte auf dem linken Rand haben die Koordinaten $i = 0$ und $j = 1, 2, \ldots, J-1$. Die Werte $u_{0,j}$ sind somit durch die Randbedingungen

$$u_{0,j} = u_a(y_j) \quad \text{für } j = 1, 2, \ldots, J-1$$

bestimmt. Diese Bedingungen können wir in (6.13) einsetzen und erhalten für $i = 1$ die Differenzengleichungen

$$-\left(\frac{2}{\Delta x^2} + \frac{2}{\Delta y^2}\right) u_{1,j} + \frac{1}{\Delta x^2} u_{2,j} + \frac{1}{\Delta y^2} u_{1,j+1} + \frac{1}{\Delta y^2} u_{1,j-1} \\ = f_{1,j} - \frac{1}{\Delta x^2} u_{0,j} \quad \text{für } j = 2, 3, \ldots, J-2 \, . \tag{6.15}$$

An der linken unteren Ecke (x_1, y_1) ergibt sich

$$-\left(\frac{2}{\Delta x^2} + \frac{2}{\Delta y^2}\right) u_{1,1} + \frac{1}{\Delta x^2} u_{2,1} + \frac{1}{\Delta y^2} u_{1,2} = f_{1,1} - \frac{1}{\Delta x^2} u_{0,1} - \frac{1}{\Delta y^2} u_{1,0} \; ,$$

da hier die zwei Randwerte $u_{1,0}$ und $u_{0,1}$ eingesetzt werden können. An der linken oberen Ecke (x_1, y_{J-1}) ergibt sich eine entsprechende Gleichung. Analog ergeben sich reduzierte Gleichungen für den rechten, unteren und oberen Rand.

Durch das Einsetzen der Randbedingungen erhält man für die unbekannten Näherungswerte an den $(I-1)(J-1)$ inneren Gitterpunkten ein **lineares Gleichungssystem (LGS)** mit $(I-1)(J-1)$ Gleichungen in der üblichen Form. Führt man einen Vektor $\mathbf{u_h}$ mit den unbekannten Werten an den verschiedenen Gitterpunkten als Komponenten ein:

$$\mathbf{u_h} = \left(u_{1,1}, u_{1,2}, \ldots, u_{1,J-1}, u_{2,1}, u_{2,2}, \ldots, u_{I-1,J-1}\right)^T \ ,$$

dann kann man das System der Differenzengleichungen als lineares Gleichungssystem in der Form

$$\mathbf{A}\mathbf{u_h} = \mathbf{f_h} \tag{6.16}$$

mit einer $(I-1)(J-1) \times (I-1)(J-1)$-Matrix $\mathbf{A}$ schreiben. Die Matrix $\mathbf{A}$ besitzt eine sogenannte Bandstruktur:

$$A = \begin{pmatrix}
\alpha & \beta & 0 & \ldots & 0 & \gamma & 0 & \ldots & 0 \\
\beta & \ddots & \ddots & & & & \ddots & & \vdots \\
0 & \ddots & \ddots & \ddots & & & & \ddots & 0 \\
\vdots & & \ddots & \ddots & \ddots & & & & \gamma \\
0 & & & \ddots & \ddots & \ddots & & & 0 \\
\gamma & & & & \ddots & \ddots & \ddots & & \vdots \\
0 & \ddots & & & & \ddots & \ddots & \ddots & 0 \\
\vdots & & \ddots & & & & \ddots & \ddots & \beta \\
0 & \ldots & 0 & \gamma & 0 & \ldots & 0 & \beta & \alpha
\end{pmatrix}$$

Die Elemente der Matrix, welche von Null verschieden sind, stehen auf der Hauptdiagonale, den beiden Nebendiagonalen und zwei um $I-1$ Koeffizienten nach rechts und links entfernten Diagonalen. Diese Struktur wird durch den 5-Punkte-Stern für den Laplace-Operator und die Anordnung der gesuchten Näherungswerte $u_{i,j}$ in dem Vektor $\mathbf{u_h}$ erzeugt.

Viele Elemente der Matrix $\mathbf{A}$ sind Null, in jeder Zeile sind höchstens 5 Koeffizienten von Null verschieden. Man nennt ein LGS mit einer solchen Matrix **schwach besetzt**. Bei einem feinen Gitter wird die Matrix $\mathbf{A}$ sehr groß. Für ein 10×10-Gitter hat man 100 Unbekannte, was schon zu einer 100×100-Matrix führt, für die man 10000 Speichereinheiten braucht. Direkte Lösungsmethoden für Gleichungssysteme wie das Gauß-Eliminationsverfahren sind hier nicht effizient oder sogar nicht durchführbar. Die Nachteile des Gauß-Algorithmus bei dieser Anwendung sind:

– Hohe Rechenzeiten, da der Rechenaufwand der Elimination $\frac{1}{3}n^3 + n^2 - \frac{1}{3}n$ ist, wobei n die Dimension der Matrix bezeichnet, in unserem Fall somit $n = (I-1)(J-1)$.

– Der Algorithmus verändert die Struktur der Matrix, da in jedem Eliminationsschritt die Koeffizienten neu berechnet werden.

– Rundungsfehler können dominant werden, da in jedem Eliminationsschritt die Koeffizienten der verbleibenden Matrix neu berechnet werden.

– Hoher Speicherbedarf; obwohl im ursprünglichen LGS viele der Koeffizienten zunächst Null sind, werden sie im Verlauf der Verfahrens geändert und durch Zahlen ungleich Null ersetzt. Deshalb muss die gesamte Matrix abgespeichert werden.

Ein weiterer Grund, dass ein direktes Verfahren für diese Systeme ungünstig ist, liefert die folgende Tatsache. Von den Rundungsfehlern abgesehen liefert das Gauß-Eliminationsverfahren die exakte Lösung. Das schwach besetzte Gleichungssystem tritt im Rahmen des Näherungsverfahrens auf. Das Näherungsverfahren liefert Ergebnisse für die Lösung der Poisson-Gleichung, welche nur bis auf den Approximationsfehler des Verfahrens genau sein können. Insofern brauchen wir eigentlich keine exakte Lösung des Gleichungssystems, sondern es genügt eine Näherungslösung, deren Fehler etwas kleiner als der schon gemachte Fehler ist. Der Rechenaufwand sollte deutlich geringer sein und die Matrix $\mathbf{A}$ mit den vielen Nullen sollte nicht explizit abgespeichert werden müssen. Diese Anforderungen erfüllen **iterative Methoden**, die wir im Folgenden betrachten möchten. Dies ist nach der Aufstellung und der Sortierung der Differenzengleichungen der nächste Schritt.

(4.) Im **Schritt 4** wird eine iterative Methode durch eine Iterationsvorschrift definiert. Auf der linken Seite steht die Iterierte im neuen Iterationsschritt, auf der rechten Seite eine Berechnungsvorschrift, in die nur die alten Iterierten eingehen. Iterationsverfahren traten in den vorigen Kapiteln im Bereich der Lösung von Anfangswertproblemen mit der Hilfe von impliziten Verfahren oder auch bei dem Schießverfahren für Randwertprobleme schon auf.

Das System von Differenzengleichungen formuliert man im Folgenden so um, dass es die Form

$$\mathbf{u} = \mathbf{G}(\mathbf{u})$$

bekommt, eine sogenannte Fixpunkt-Gleichung. Die Iterationsvorschrift lautet dann

$$\mathbf{u^{(p+1)}} = \mathbf{G}(\mathbf{u^{(p)}})$$

mit dem Iterationsindex p. Die Differenzengleichung (6.13) wird in diese Form gebracht, indem so umsortiert wird, dass nur der Term mit $u_{i,j}$ auf der linken Seite der Gleichung verbleibt:

$$-2\left(\frac{1}{\Delta x^2}+\frac{1}{\Delta y^2}\right)u_{i,j} = f_{i,j} - \frac{1}{\Delta x^2}u_{i-1,j} - \frac{1}{\Delta x^2}u_{i+1,j} - \frac{1}{\Delta y^2}u_{i,j-1} - \frac{1}{\Delta y^2}u_{i,j+1}\,. \tag{6.17}$$

Eine Iterationsvorschrift erhält man daraus, indem auf der linken Seite der Iterationsindex $p+1$ und auf der rechten Seite der Iterationsindex p geschrieben wird. Mit der Abkürzung

$$d = -2\left(\frac{1}{\Delta x^2}+\frac{1}{\Delta y^2}\right)$$

ergibt sich dann die Iterationsvorschrift

$$u_{i,j}^{(p+1)} = \frac{1}{d}\left(f_{i,j} - \frac{1}{\Delta x^2}u_{i-1,j}^{(p)} - \frac{1}{\Delta x^2}u_{i+1,j}^{(p)} - \frac{1}{\Delta y^2}u_{i,j-1}^{(p)} - \frac{1}{\Delta y^2}u_{i,j+1}^{(p)}\right)$$
für $i = 1, 2, \ldots, I-1$ und $j = 1, 2, \ldots, J-1$ (6.18)

Jacobi-Verfahren

Für die inneren Gitterpunkte, welche am Rand liegen, müssen hier die entsprechenden Randwerte eingesetzt werden. Dieses Iterationsverfahren wird **Jacobi-Verfahren** genannt. Die Iteration wird mit einer Anfangsschätzung der Lösung $u_h^{(0)}$ begonnen und dann so lange über p iteriert, bis die gewünschte Genauigkeit erreicht ist. Bevor wir dies noch ausführlicher diskutieren, betrachten wir noch andere Möglichkeiten der Definition eines Iterationsverfahrens.

Bei dem Jacobi-Verfahren beginnt man nach der Vorschrift (6.18) bei der $(p+1)$-ten Iteration mit der 1. Gleichung, um $u_{1,1}^{(p+1)}$ durch die Werte der p-ten Iteration zu berechnen. Anschließend wählt man die 2. Gleichung und berechnet $u_{2,1}^{(p+1)}$ aus den "alten" Daten der p-ten Iteration. Dabei könnte man ausnutzen, dass für den Wert $u_{1,1}$ schon die $(p+1)$-Iterierte vorliegt. Man sollte ein verbessertes Verfahren erhalten, wenn diese Information schon in die Iterationsvorschrift eingeht und die aktualisierten Werte berücksichtigt. Dieses Verfahren nennt man das **Gauß-Seidel-Verfahren**. Die Iterationsvorschrift lautet dann

$$u_{i,j}^{(p+1)} = \frac{1}{d}\left(f_{i,j} - \frac{1}{\Delta x^2}u_{i-1,j}^{(p+1)} - \frac{1}{\Delta x^2}u_{i+1,j}^{(p)} - \frac{1}{\Delta y^2}u_{i,j-1}^{(p+1)} - \frac{1}{\Delta y^2}u_{i,j+1}^{(p)}\right)$$
$$\text{für } i = 1, 2, \ldots, I-1 \text{ und } j = 1, 2, \ldots, J-1 \,. \tag{6.19}$$

Gau"s-Seidel-Verfahren

Es zeigt sich in praktischen Anwendungen, dass das Gauß-Seidel-Verfahren zum Erreichen einer vorgegebenen Genauigkeit deutlich weniger Iterationen und damit weniger Rechenaufwand benötigt.

Eine Verbesserung des Gauß-Seidel-Verfahrens kann zusätzlich durch eine Gewichtung des neu berechneten Wertes erreicht werden. Bei diesem sogenannten **SOR-Verfahren** (Successive Over Relaxation) wird eine Linearkombination des aktualisierten Wertes aus dem Gauß-Seidel-Verfahren, welchen wir hier mit $u_{GS,i,j}^{(p+1)}$ bezeichnen und des alten Wertes $u_{i,j}^{(p)}$ gewählt:

$$u_{i,j}^{(p+1)} = w \cdot u_{GS,i,j}^{(p+1)} + (1-w) \cdot u_{i,j}^{(p)} \,. \tag{6.20}$$

SOR-Verfahren

Für den Relaxationsparameter w muss $0 < w < 2$ erfüllt sein. Üblicherweise liegt der Relaxationsparameter w im Bereich $1 \leq w \leq 2$, deshalb auch der Name Überrelaxation. Für die optimale Wahl des Parameters gibt es entsprechende Abschätzungen.

Die schnelle Lösung der Gleichungssysteme ist natürlich sehr wichtig für das numerische Verfahren zur Lösung der elliptischen Gleichung. Die iterative Lösung von großen schwach besetzten Gleichungssystemen ist ein wichtiges Teilgebiet der numerischen Mathematik. Eine Übersicht über aktuelle Verfahren ist im Anhang ausgeführt. Es werden dort im allgemeineren Rahmen diese hier vorgestellten Verfahren diskutiert und Worksheets angegeben. Weitere numerische Verfahren, wie das Verfahren der konjugierten Gradienten oder Mehrgitterverfahren, sind in der Übersicht im Anhang mit aufgeführt, sowie weitere Spezialliteratur.

Bei dem Gaußschen Eliminationsverfahren berechnet man die Lösung des LGS bis auf Rundungsfehler exakt. Bei einem Iterationsverfahren berechnet man in jedem Iterationsschritt nur eine neue hoffentlich bessere Lösung. Bei Iterationsverfahren kann bei Erreichen der erforderlichen Genauigkeit der Lösung die Iteration abgebrochen werden. Man braucht dazu ein geeignetes

Abbruchkriterium, in das Information über den Fehler der Näherungslösung nach jeder Iteration eingeht. Einen Vergleich mit der exakten Lösung wäre das Beste. Da diese aber nicht bekannt ist, kann man als ein Kriterium nehmen, wie gut die gefundene Näherungslösung $\mathbf{u^{(p+1)}}$ das Gleichungssystem schon erfüllt. Dazu bestimmt man das sogenannte **Residuum**:

$$\mathbf{R}_h^{(p)} := \mathbf{A}\mathbf{u}_h^{(p+1)} - \mathbf{b}$$

oder in Komponenten geschrieben

$$R_i^{(p+1)} = \sum_{j=1}^{n} a_{ij}\, u_{i,j}^{(p+1)} - b_i \qquad (i = 1, 2, \ldots, n) \; .$$

Als Maß für die Güte der Lösung kann man z. B. das Maximum des Fehlerbetrags über alle Komponenten

$$R_{max}^{(p+1)} = \max_{i=1,\ldots,n} |R_i^{(p+1)}|$$

nehmen. Als Abbruchkriterium fordert man in der Regel, dass

1. das Residuum kleiner einer gewissen Vorgabe ist: $R_{max}^{(p+1)} < \delta_1$, und dass
2. die Differenz der Werte zweier aufeinander folgender Iterationen kleiner einer vorgegebenen Schranke sind:

$$\max_{i=1,\ldots,I-1,j=1,\ldots,J-1} |u_{i,j}^{(p+1)} - u_{i,j}^{(p)}| < \delta_2 \; .$$

Wie schon diskutiert gibt das erste Kriterium die Aussage, wie gut die Näherungslösung das Gleichungssystem erfüllt. Das zweite Kriterium überprüft, ob noch eine weitere Verbesserung der Näherungslösung geliefert wird.

(5.) Schritt 5 besteht dann in der Aufarbeitung der numerischen Daten für das Ausgangsproblem, meist eine Visualisierung der diskreten Näherungswerte.

Zusammenfassung: Differenzenverfahren

1. Diskretisierung des Rechengebietes.
2. Ersetzen der partiellen Ableitungen durch finite Differenzen.
3. Aufstellen des zugehörigen LGS.
4. Lösen des LGS mit einer geeigneten Methode.
5. Visualisierung der diskreten Lösungswerte.

Beispiel 6.1 (Mit MAPLE-Worksheet). Wir wenden den 5-Punkte-Stern zur Lösung der Laplace-Gleichung nun auf das Beispiel 4.2 an, welches das physikalische Problem einer Membran, bei der eine der Rechteckseiten verbogen ist, behandelt. Für dieses Beispiel wurde die exakte Lösung über einen Separationsansatz berechnet. In dem MAPLE-Worksheet **Membran** wird eine numerische Lösung berechnet. Zugrunde gelegt ist ein 20×20 Punktegitter. Die Näherungslösung stimmt sehr gut mit der exakten Lösung überein. In Abbildung 6.2 ist der zugehörige Fehler geplottet. □

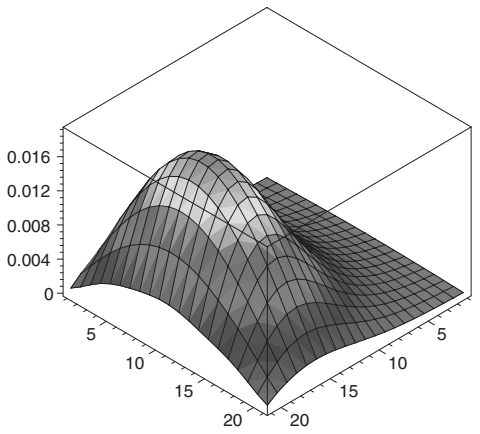


Abb. 6.2. Fehler der Näherungslösung in Beispiel 6.1

Beispiel 6.2 (3-Elektroden-System, mit MAPLE-Worksheet). Im Folgenden diskutieren wir ein zweidimensionales elektrostatisches Problem am Beispiel des in Abbildung 6.3 angegebenen 3-Elektroden-Systems mit Spaltblende. Das elektrische Potenzial wird durch die Poisson-Gleichung

$$\phi_{xx}(x,y) + \phi_{yy}(x,y) = -\frac{1}{\epsilon}\rho(x,y) \tag{6.21}$$

bzw. für den hier betrachteten ladungsfreien Raum ($\rho = 0$) durch die Laplace-Gleichung

$$\phi_{xx}(x,y) + \phi_{yy}(x,y) = 0 \tag{6.22}$$

beschrieben. Die fünf Schritte des Differenzenverfahrens sehen dann wie folgt aus:

Schritt 1: **Diskrete Beschreibung des Gebiets.** Als Ränder des Gebiets werden im obigen Problem nicht nur die äußeren Ränder $\Phi(x = 0, y) = \Phi_K$

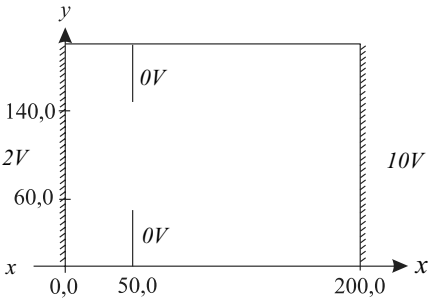


Abb. 6.3. 3-Elektroden-System mit Spaltblende

und $\Phi(x = d, y) = \Phi_A$ berücksichtigt, sondern noch zusätzlich innere Ränder $\Phi(x = 5,\ 0 \leq y \leq 6) = \Phi_Z = 0\,V$ und $\Phi(x = 5,\ 14 \leq y \leq 20) = \Phi_Z = 0\,V$. Damit muss das Gitter so gewählt werden, dass die Spaltblenden durch Gitterlinien dargestellt werden. Die Linien bzw. die Punkte in Abbildung 6.4 dienen als ein solches diskretes Gitter.

Wir definieren ein achsenparalleles äquidistantes Gitter mit 40 Intervalle in x- und y-Richtung. Damit erhält man 41 Gitterpunkte in x- und y-Richtung.

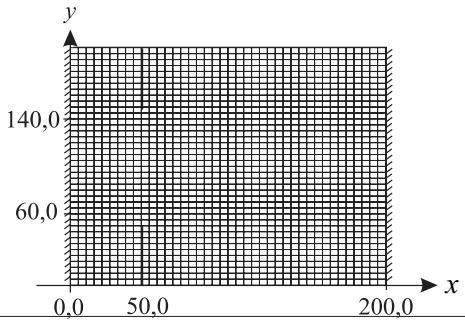


Abb. 6.4. Gitter für das 3-Elektroden-System

Die äquidistanten Schrittweiten haben die Werte

$$\Delta x = \frac{200mm}{40} = 5mm, \qquad \Delta y = \frac{200mm}{40} = 5mm\ ,$$

so dass die Koordinaten der Gitterpunkte

$$x_{i,j} = i\Delta x\,, \qquad y_{i,j} = j\Delta y \tag{6.23}$$

sind. Insbesondere ist durch diese Wahl der Gitterpunkte die Spaltblende als Gitterlinie enthalten. In dem ersten Beispiel haben wir die Koordinaten der Gitterpunkte als Vektoren eingeführt und die Gitterpunkte als $P_{i,j} = (x_i, y_j)$ definiert. Wie wir in (6.23) sehen, hängen die Koordinaten $x_{i,j}$ und $y_{i,j}$ auch hier nicht von j bzw. i ab, da die Gitterlinien jeweils parallel zu den Koordinatenachsen sind. Die Angabe der Gitterpunkte als Matrix erleichtert wegen der Anwesenheit eines inneren Randes durch die Spaltblende das Aufstellen des Gleichungssystems. Für allgemeine rand-angepasste Gitter benötigt man dann ebenso eine Matrix jeweils mit den x- und y-Koordinaten. Wir werden darauf in Kapitel 6.4 noch zurückkommen.

Schritt 2: **Ersetzen der partiellen Ableitungen durch Differenzenquotienten.** Nachdem das Gebiet durch ein diskretes Gitter erfasst ist, wird die Laplace-Gleichung auf diesen Gitterpunkten für das Potenzial

$$\Phi_{xx}(x,y) + \Phi_{yy}(x,y) = 0$$

näherungsweise gelöst.

Bezeichnet $\Phi_{i,j}$ wieder den Näherungswert für das Potenzial im Punkte $(x_{i,j}, y_{i,j})$, so können wir die diskrete Laplace-Gleichung an der Stelle $(x_{i,j}, y_{i,j})$ schreiben als

$$\frac{\Phi_{i-1,j} - 2\Phi_{i,j} + \Phi_{i+1,j}}{(\Delta x)^2} + \frac{\Phi_{i,j-1} - 2\Phi_{i,j} + \Phi_{i,j+1}}{(\Delta y)^2} = 0$$

$$\Rightarrow \frac{1}{(\Delta x)^2}\Phi_{i-1,j} - 2\Big(\frac{1}{(\Delta x)^2} + \frac{1}{(\Delta y)^2}\Big)\Phi_{i,j} + \frac{1}{(\Delta x)^2}\Phi_{i+1,j} + \frac{1}{(\Delta y)^2}\Phi_{i,j-1} + \frac{1}{(\Delta y)^2}\Phi_{i,j+1} = 0\,.$$

(**diskrete Laplace-Gleichung**)

Diese Gleichung gilt für alle inneren Gitterpunkte bis auf die Punkte, welche die Elektroden repräsentieren. Die Randbedingungen am diesem inneren Rand als auch am äußeren Rand müssen noch entsprechend getrennt behandelt werden. Dabei sind zwei Typen von Randbedingungen zu unterscheiden:

Dirichlet-Randbedingung: Gilt für die *Punkte* auf Elektroden; das Potenzial ist durch eine äußere Spannung vorgegeben. Im Falle des 3-Elektroden-Systems sind dies die Punkte, die entweder auf Anodenspannung $\Phi_A = 10\,V$,

auf Kathodenspannung $\Phi_K = 2\,V$ oder auf Zwischenspannung $\Phi_Z = 0\,V$ liegen.

Neumann-Randbedingung: Gilt für die Rand*linien*, bei denen die Ableitung des Potenzials vorgegeben ist bzw. wie in unserem Falle verschwindet. Beim 3-Elektroden-Systems nehmen wir an, dass am oberen und unteren Rand das elektrische Feld keine Komponente in y-Richtung besitzt, d.h. die Potenziallinien parallel zur y-Achse verlaufen:

$$\frac{\partial \Phi}{\partial y} = 0.$$

Diskretisiert über die einseitige Differenzenformel (6.1) bedeutet dies für den unteren Rand ($j = 0$)

$$\Phi_y(x, y = 0) \sim \frac{\Phi_{i,1} - \Phi_{i,0}}{\Delta y} = 0$$

mit $i = 0, \ldots, I$ und über die einseitige Differenzenformel (6.3) für den oberen Rand ($j = J$)

$$\Phi_y(x, y = 20) \sim \frac{\Phi_{i,J} - \Phi_{i,J-1}}{\Delta y} = 0$$

mit $i = 0, \ldots, I$. Insgesamt erhalten wir 41×41 Gleichungen für 41×41 Gitterpunkte, also ein lineares Gleichungssystem für die Potenzialwerte $\Phi_{i,j}$ an den Gitterpunkten $(x_{i,j}, y_{i,j})$.

Schritt 3: **Aufstellen des LGS.** Um die Struktur des LGS besser zu veranschaulichen, wählen wir ein 6×5-Gitter und zählen die Gitterpunkte von links unten beginnend durch (siehe Abbildung 6.5).

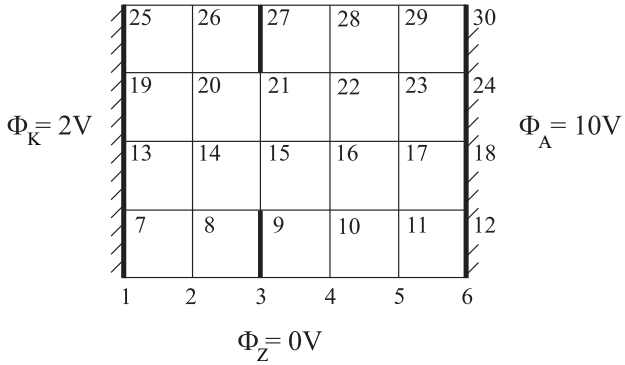


Abb. 6.5. Vereinfachtes Gitter

Dirichlet-Punkte	1, 7, 13, 19, 25	Potenzial $\Phi_K = 2\,V$
	6, 12, 18, 24, 30	Potenzial $\Phi_A = 10\,V$
	3, 9; 21, 27	Potenzial $\Phi_Z = 0\,V$
Neumann-Punkte	2, 4, 5; 26, 28, 29	
Feld-Punkte	Rest	

An den $6 \times 5 = 30$ Gitterpunkten stellen wir das LGS für das Potenzial $\Phi_{i,j}$ in Matrixform auf. Der Einfachheit halber setzen wir $\Delta x = \Delta y = 1$.

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	
1	1																														ϕ_K
2		-1						1																							0
3			1																												ϕ_Z
4				-1						1																					0
5					-1						1																				0
6						1																									ϕ_A
7							1																								ϕ_K
8		1					1	-4	1					1																	0
9								1																							ϕ_Z
10				1				1	-4	1						1															0
11					1				1	-4	1						1														0
12											1																				ϕ_A
13													1																		ϕ_K
14								1				1	-4	1					1												0
15									1				1	-4	1					1											0
16										1				1	-4	1					1										0
17											1				1	-4	1					1									0
18																	1														ϕ_A
19																		1													ϕ_K
20														1				1	-4	1					1						0
21																				1											ϕ_Z
22																	1			1	-4	1					1				0
23																		1			1	-4	1					1			0
24																							1								ϕ_A
25																								1							ϕ_K
26																					-1				1						0
27																										1					ϕ_Z
28																							-1				1				0
29																								-1				1			0
30																														1	ϕ_A

Schritt 4: **Lösung des LGS.** Das LGS $\mathbf{A}\,\mathbf{\Phi} = \mathbf{r}$ mit der Koeffizientenmatrix $\mathbf{A}$ und der rechten Seite $\mathbf{r}$, wie oben angegeben, wird im MAPLE-**Worksheet** mit dem SOR-Verfahren ausgeführt.

Schritt 5: **Darstellung der Ergebnisse.** Durch Lösen des LGS mit geeigneten iterativen Verfahren erhält man die gesuchte Lösung näherungsweise an den Gitterpunkten. Die Lösung wird repräsentiert durch Werte $\Phi_{i,j}$ an den Gitterpunkten $(x_{i,j}, y_{i,j})$. Zur Interpretation dieser Daten führt man sogenannte *Äquipotenziallinien* ein. Dies sind Linien, auf denen der Wert des Potenzials gleich ist. Das Problem ist aber dann, wie man aus den berechneten Daten diese Äquipotenziallinien bestimmt. Zur Klärung dieser Frage betrachten wir Abbildung 6.6.

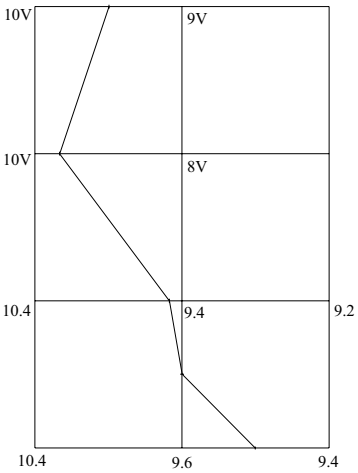


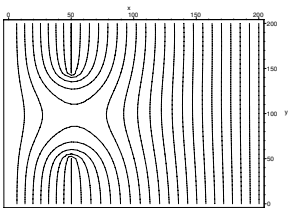
Abb. 6.6. Lineare Interpolation

Gegeben sind die Potenzialwerte an den Gitterpunkten. Gesucht ist z.B. die 9.5-Volt-Linie. Um diese Linie zu finden, wählt man sich einen Eckpunkt im Gitter (hier links oben mit 10 V) und betrachtet die beiden angrenzenden Zellecken (10 V nach unten; 9 V nach rechts). Damit ist klar, dass die 9.5-Volt-Linie durch die obere Gitterlinie geht. Lineare Interpolation besagt, dass sie in der Mitte liegt. Die 9.5-Volt-Linie startet also an der oberen Zellbegrenzung in der Mitte.

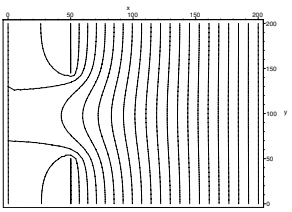
Anschließend werden alle weiteren Begrenzungslinien dieser Zelle untersucht und entschieden, ob die 9.5-Volt-Linie durch eine dieser Linien geht. In unserem Fall ist dies die untere Zellbegrenzung. Lineare Interpolation legt auf dieser unteren Linie die Position des Werts 9.5 V fest. Auf die gleiche Weise wird nun die zweite Zelle durchsucht und die Schnittpunkte der 9.5-Volt-Linie mit den Zellbegrenzungen bestimmt. Somit kommt die 9.5-Volt-Linie von der obersten Zelle durch die zweite Zelle bis hin zur untersten Zelle. Dort wird festgestellt, dass die Linie in die rechte Zelle geht usw.

In den beiden folgenden Diagrammen sind die Äquipotenziallinien für die Randbedingungen ($\Phi_A = 10\,V, \Phi_K = 2\,V$ und $\Phi_Z = 0\,V$) bzw. ($\Phi_A = 10\,V, \Phi_K = 4\,V$ und $\Phi_Z = 0\,V$) dargestellt. Die Ergebnisse stammen aus dem MAPLE-**Worksheet**

Auf Grund der vorgegebenen Randbedingungen erwarten wir symmetrische Ergebnisse zur Linie $y = 100mm$, was sich im Ergebnis auch zeigt. Man erkennt im Vergleich der beiden Darstellungen sehr schön, dass eine kleine Änderung der Randbedingungen (der einzige Unterschied liegt in Φ_K) den qualitativen Verlauf der Lösung grundlegend ändert. Im zweiten Fall erhält



(a) Äquipotenziallinien für $\Phi_K = 4\,V$.



(b) Äquipotenziallinien für $\Phi_K = 2\,V$.

man einen sogenannten Durchbruch der Potenziallinien, der im ersten nicht möglich ist. □

6.2 Parabolische Differenzialgleichungen

Wir betrachten in diesem Abschnitt das Anfangs-Randwertproblem für die einfachste parabolische Differenzialgleichung, die Wärmeleitungsgleichung

$$u_t = \kappa\, u_{xx} \tag{6.24}$$

mit der konstanten Temperaturleitzahl $\kappa > 0$. Die Variable u bezeichnet die Temperatur. Dabei starten wir mit einer Raumdimension und werden danach die Verfahren auf den mehrdimensionalen Fall erweitern. Die Diskretisierungsparameter sind die konstante Schrittweite Δt in der Zeit und die konstante Schrittweite Δx im Raum. Die Gitterpunkte sind durch

$$t_n = n\Delta t\,, \qquad x_i = a + i\Delta x\,, \qquad n = 0, 1, 2, \ldots\,,\ i = 0, 1, \ldots, I$$

definiert. Die Abbildung 6.7 zeigt einen Ausschnitt des Gitters in der (x, t)-Ebene.

Die Näherungswerte bezeichnen wir, wie schon vorher, mit indizierten Größen. Dabei hat sich der Übersichtlichkeit halber in der Literatur durchgesetzt, dass man die Raumindizes als untere Indizes schreibt, wie wir dies auch schon bei den elliptischen Gleichungen durchgeführt haben. Der Zeitindex schreibt man nun als einen oberen Index. Dies bedeutet, dass der Wert u_i^n eine Näherung der Lösung u an dem Punkt (x_i, t_n) darstellt.

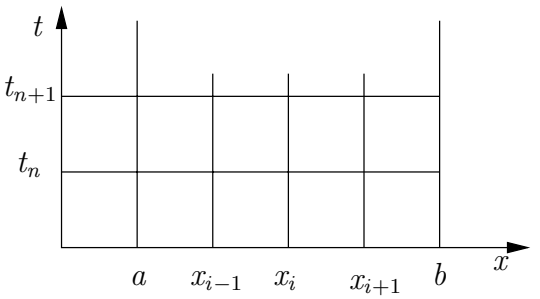


Abb. 6.7. Diskretisierung des Rechengebietes

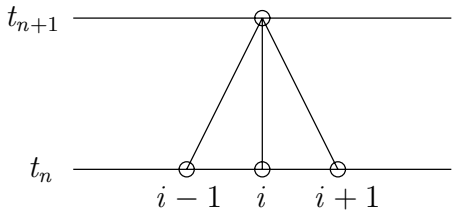
Ein sehr einfaches Verfahren erhalten wir, wenn die Zeitableitung durch den vorwärts-genommenen Differenzenquotienten und die zweite Ableitung in x durch einen zentralen Differenzenquotienten approximiert wird:

$$\frac{u_i^{n+1} - u_i^n}{\Delta t} = \kappa \frac{u_{i+1}^n - 2u_i^n + u_{i-1}^n}{\Delta x^2} \,. \tag{6.25}$$

Die einzige unbekannte Größe ist u_i^{n+1}. Gleichung (6.25) nach dieser Größe aufgelöst ergibt

$$u_i^{n+1} = u_i^n + d\left(u_{i+1}^n - 2u_i^n + u_{i-1}^n\right) \quad \text{mit} \quad d := \kappa \frac{\Delta t}{\Delta x^2} \,. \tag{6.26}$$

Die Differenzialgleichung ist somit ersetzt durch eine einfache algebraische Gleichung. Es gehen neben dem Näherungswert am Punkt (x_i, t_{n+1}) noch die Werte an den drei Punkten (x_{i-1}, t_n), (x_i, t_n), (x_{i+1}, t_n) in die Differenzengleichung ein. Grafisch lässt sich dies in dem Differenzenmolekül oder dem Differenzenstern



darstellen. Der einzige Wert auf dem Zeitniveau $n+1$, der dabei in die Berechnung eingeht, ist u_i^{n+1}. Gehen wir davon aus, dass alle Werte zum Zeitpunkt t_n bekannt sind, dann können wir in jedem inneren Gitterpunkt zum neuen Zeitniveau einen Näherungswert berechnen. Am Rand muss man entsprechend die Randwerte in die Formeln einsetzen. Man nennt dies ein **explizites Verfahren**. Sukzessive kann man mit Start bei $t = \Delta t$ und der Vorgabe der Anfangswerte eine Zeitschicht nach der anderen abarbeiten.

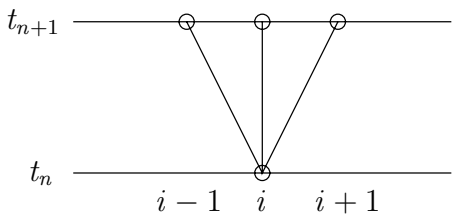
Wenn statt der Vorwärtsdifferenz der **rückwärts-genommene Differenzenquotient** für die Zeitableitung benutzt wird und die zentrale Approximation im Raum beibehalten wird, erhält man ein Schema in der Form

$$\frac{u_i^{n+1} - u_i^n}{\Delta t} = \kappa \frac{u_{i+1}^{n+1} - 2u_i^{n+1} + u_{i-1}^{n+1}}{\Delta x^2} \,. \tag{6.27}$$

In dieser Gleichung gibt es die drei Unbekannten

$$u_{i-1}^{n+1} \,, \qquad u_i^{n+1} \,, \qquad u_{i+1}^{n+1} \,.$$

Das Differenzenmolekül sieht demnach folgendermaßen aus:



Da es mehr als eine Unbekannte in dieser Differenzengleichung gibt, lassen sich die Unbekannten nicht mehr explizit berechnen. Man nennt ein solches Verfahren ein **implizites Verfahren**, mit dem rückwärts-genommenen Differenzenquotienten erster Ordnung auch **voll-implizites Verfahren**. Es ergibt sich hier eine Kopplung mit den Gleichungen an den benachbarten Gitterpunkten. Wir gehen wieder davon aus, dass alle Werte zum Zeitpunkt t_n bekannt sind und schreiben die Gleichung (6.27) in der Form

$$-d\; u_{i-1}^{n+1} + (1+2d) u_i^{n+1} - d\; u_{i+1}^{n+1} = u_i^n \,.$$

Da sich für jeden inneren Gitterpunkt eine solche Gleichung ergibt, erhält man ein lineares Gleichungssystem:

$$a_i\, u_{i-1}^{n+1} + b_i\, u_i^{n+1} + c_i\, u_{i+1}^{n+1} = u_i^n \tag{6.28}$$

$$\text{mit} \quad a_i = c_i = -d,\; b_i = 1 + 2d \,. \tag{6.29}$$

Werden die vorgegebenen Randwerte noch eingesetzt, lässt sich das Gleichungssystem in Matrixform umschreiben. Da die Matrix tridiagonal ist, kann das LGS mit dem Thomas-Algorithmus gelöst werden, welcher bei der numerischen Lösung von Randwert-Problemen für gewöhnliche Differenzialgleichungen im Kapitel 3 eingeführt wurde. Ähnlich zu dem expliziten Verfahren wird das Verfahren sukzessive von Zeitschritt zu Zeitschritt ausgeführt, startend bei den Anfangswerten, wobei in jedem Zeitschritt das tridiagonale LGS gelöst werden muss.

Man wird natürlich den Aufwand zur Lösung eines linearen Gleichungssystem vermeiden, wenn das einfache explizite Verfahren schon die gewünschten Lösungen mit viel weniger Rechenaufwand liefert. Die Genauigkeit beider Verfahren ist dieselbe. Der vorwärts- oder rückwärts-genommene Differenzenquotient für die Zeitableitung hat die Konsistenzordnung 1, in beiden Fällen wird der zentrale Differenzenquotienten im Raum benutzt. Neben der Konsistenz oder Genauigkeit ist die Stabilität ein Kriterium, bei dem Unterschiede zwischen den Verfahren auftreten könnten. Bevor wir dies untersuchen, wenden wir das explizite Verfahren auf ein Beispiel an.

Beispiel 6.3 (Instationäre Wärmeleitung, mit MAPLE-Worksheet). Das explizite Verfahren wird zur Berechnung der Wärmeverteilung eines Metallstabes eingesetzt. Das physikalische Problem besteht aus einem eingespannten Metallstab mit der Temperaturleitzahl $\kappa = 1.14$, der auf den Stirnseiten auf konstanter Temperatur

$$u(x = 0, t) = 2\,, \quad u(x = L, t) = 0.5$$

gehalten wird. Die Anfangstemperaturverteilung lautet

$$u(x, t = 0) = -1.5x + 2 + \sin(\pi x)\ .$$

Die exakte Lösung wurde als Beispiel 4.5 mit Hilfe eines Separationsansatzes gelöst. Abbildung 6.8 zeigt die Zeitentwicklung der numerischen Temperaturverteilung im Zeitintervall $[0, 0.1]$. An der festen Stelle $x = L/2$ zeigt Abbildung 6.9 einen Vergleich mit der exakten Lösung. Die Berechnung erfolgt

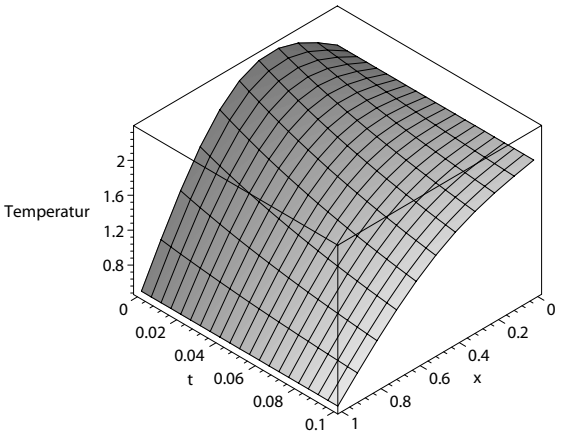


Abb. 6.8. Numerische Lösung von Beispiel 6.3: Temperaturverteilung als Funktion von x, t und t

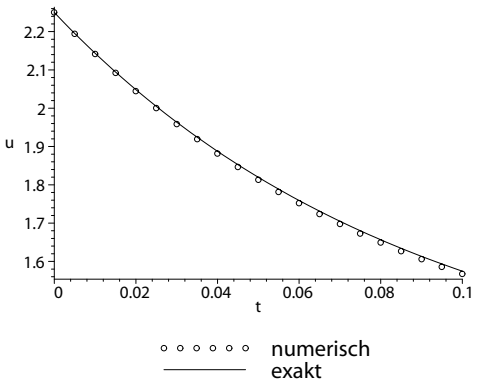


Abb. 6.9. Vergleich der numerischen und der exakten Lösung am Punkt $x = L/2$ als Funktion von t

im MAPLE-Worksheet. Vergrößert man die Zeitschrittweite, indem man z.B. den Endzeitpunkt auf 0.5 erhöht und die Anzahl der Zeitschritte belässt, wird die Lösung instabil. Dies deutet darauf hin, dass eine Schrittweitenbedingung aus der Stabilitätsbedingung folgt, wie wir dies auch schon bei numerischen Verfahren für Anfangswertprobleme gewöhnlicher Differenzialgleichungen gesehen haben. □

Stabilität des expliziten Verfahrens

Wir werden uns zunächst das explizite Verfahren etwas genauer anschauen. Für eine Stabilitätsanalyse gibt es verschiedene Ansätze. Bei der **Methode der Störungsrechnung** wird explizit eine Störung zu einer beliebigen Zeit t_n eingeführt. Betrachten wir zunächst das explizite Verfahren.

An der i-ten Stelle führen wir eine Störung ε ein und erhalten

$$\frac{u_i^{n+1} - (u_i^n + \varepsilon)}{\Delta t} = \kappa \frac{u_{i+1}^n - 2(u_i^n + \varepsilon) + u_{i-1}^n}{\Delta x^2} \; .$$

Wir nehmen an, dass zur Zeit t_n alle Werte der Näherungslösung konstant und der Einfachheit halber identisch Null sind. Nach u_i^{n+1} aufgelöst ergibt sich damit

$$u_i^{n+1} = -\kappa \frac{2\varepsilon \Delta t}{\Delta x^2} + \varepsilon = \varepsilon \left(1 - 2\kappa \frac{\Delta t}{\Delta x^2}\right) \; .$$

Bei der exakten Lösung würden wir erwarten, dass bei dieser sehr lokalen Temperaturspitze die Wärme sehr schnell nach den Seiten abfließt und die

Spitze sich abrundet und verkleinert. Dies sollte für die numerische Lösung ebenso zutreffen. Aus obiger Gleichung ergibt sich, dass die Störung in der numerischen Lösung nicht anwächst, falls

$$|1 - 2d| \leq 1 \quad \text{mit } d = \kappa \frac{\Delta t}{\Delta x^2}$$

gilt. Dies führt auf die Bedingung $d \leq 1$. Den Parameter d nennt man die **Diffusionszahl** .

Eine noch stärkere Bedingung für das Anwachsen einer Störung erhält man bei der Wahl

$$u_{i+1}^n = -\varepsilon\,, \qquad u_i^n = \varepsilon\,, \qquad u_{i-1}^n = -\varepsilon\,.$$

In das Verfahren oben eingesetzt ergibt sich hierfür die Schrittweitenbedingung $d < \frac{1}{2}$.

Wie man sieht, ergibt die Methode der Störungsrechnung erste Anzeichen für eine bestehende Stabilitätsbedingung. Das **explizite Differenzenverfahren** (6.27) ist somit für die parabolische Wärmeleitungsgleichung nur **bedingt stabil**. Allerdings ergeben sich bei unterschiedlichen Störungen durchaus verschiedene Bedingungen. Die Entwicklung der Störungen wurde allerdings nur von einem Zeitschritt zum nächsten betrachtet. Inwieweit sich diese Störungen im weiteren entwickeln, wurde nicht untersucht.

Wir werden deshalb noch eine andere Methode der Stabilitätsuntersuchung vorstellen, welche eine sehr große praktische Bedeutung hat. Diese Methode nennt man die **von Neumannsche Stabilitätsanalyse**. Sie wurde von John von Neumann während des zweiten Weltkriegs in Los Alamos entwickelt und fiel zunächst unter die Geheimhaltung. Erst 1947 wurde sie in einer Arbeit beschrieben.

Vorausgesetzt wird, dass das Problem periodisch ist. In diesem Fall gibt es eine Lösung der Differenzengleichung der Form

$$u_i^n = \sum_k U^n e^{i\Delta x I k}\,.$$

Dabei bezeichnet etwas unüblich I die imaginäre Einheit mit $I^2 = -1$, da i schon der Index der x-Koordinate ist. Die anderen Größen haben die folgenden Bedeutungen: U^n ist die Amplitude zum Zeitpunkt t_n und k ist die Wellenzahl in x-Richtung, d. h. $\lambda = \frac{2\pi}{k}$ ist die Wellenlänge, das Rechenintervall ist $[-\pi, \pi]$. Der obige Ansatz stellt eine diskrete Fourier-Reihe dar.

Da die Differenzial- und die Differenzengleichung als zugehörige Approximation linear sind, gilt das Superpositionsprinzip: Verschiedene Lösungen können

addiert werden und sind wieder Lösung der Differenzengleichung. So ist es ausreichend, nur eine Komponente der Summe zu betrachten. Wird der Phasenwinkel $\phi = k\Delta x$ eingeführt hat eine beliebige Komponente die Form

$$u_i^n = U^n e^{I\phi i} \,. \tag{6.30}$$

Zu diesen Anfangsdaten für $t = t_n$ würde die exakte Lösung der Wärmeleitungsgleichung so aussehen, dass die Amplitude sich verkleinert. Ein Anwachsen der Amplitude für eine beliebige Wellenlänge beim numerischen Verfahren sollte somit die Instabilität des Verfahrens belegen.

Um diesen Ansatz in die Differenzengleichung (6.25) einzusetzen, benötigen wir neben (6.30) noch die folgenden Ausdrücke:

$$u_i^{n+1} = U^{n+1} e^{I\phi i} \quad \text{und} \quad u_{i\pm1}^n = U^n e^{I\phi(i\pm1)} \,. \tag{6.31}$$

Das Einsetzen liefert dann

$$U^{n+1} e^{I\phi i} = U^n e^{I\phi i} + d\left[U^n e^{I\phi(i+1)} - 2U^n e^{I\phi i} + U^n e^{I\phi(i-1)}\right] \,.$$

Wird durch $e^{I\phi i}$ dividiert, ergibt sich daraus

$$U^{n+1} = U^n \left[1 + d\big(e^{I\phi} + e^{-I\phi} - 2\big)\right] \,.$$

Mit der Beziehung

$$\cos\phi = \frac{e^{I\phi} + e^{-I\phi}}{2} \tag{6.32}$$

lässt sich diese Gleichung weiter umformen in

$$U^{n+1} = U^n[1 + 2d(\cos\phi - 1)] \,.$$

Die Größe G mit der Eigenschaft $U^{n+1} = GU^n$ bezeichnet man als den **Verstärkungsfaktor**. Für eine stabile Lösung, bei der Störungen nicht anwachsen können, muss der Betrag des Verstärkungsfaktors kleiner eins sein.

G ergibt sich in unserem Fall zu

$$G = 1 - 2d(1 - \cos\phi) \,.$$

Dies führt auf die Bedingungen

$$-1 \le 1 - 2d(1 - \cos\phi) \le 1 \,,$$

wobei die rechte Ungleichung immer erfüllt ist. Die Abschätzung nach unten ist erfüllt, falls

$$1 - 2d \cdot 2 \ge -1 \,.$$

Damit erhalten wir die Stabilitätsbedingung für das explizite Verfahren:

Das explizite Verfahren für die Wärmeleitungsgleichung ist **bedingt stabil**. Die Stabilitätsbedingung lautet

$$d = \kappa \frac{\Delta t}{\Delta x^2} \leq \frac{1}{2} \, . \tag{6.33}$$

Diese Bedingung hatten wir auch schon mit der Methode der Störungsrechnung erhalten. Allerdings wussten wir dort nicht, ob es nicht eine Störung gibt, welche die Stabilitätsbedingung noch verschärft.

Bemerkung: Geht man davon aus, dass neben der Temperaturleitzahl κ die Raumdiskretisierung fest vorgegeben ist, dann ist die Stabilitätsbedingung eine Bedingung an die Zeitschrittweite: Der Zeitschritt Δt darf höchstens so groß gewählt werden, dass

$$\Delta t \leq \frac{1}{2} \frac{\Delta x^2}{\kappa}$$

gilt. Diese Bedingung sollte in einem Rechenprogramm als Steuerung der Zeitschrittweite eingesetzt werden.

Crank-Nicolson-Verfahren

Als nächstes wird die **Stabilität von impliziten Verfahren** untersucht. Das einfache implizite Verfahren mit dem rückwärts-genommenen Differenzenquotienten in der Zeit und dem zentralen im Raum wurde in diesem Kapitel schon vorgestellt. Wir werden im Folgenden noch ein implizites Verfahren anschauen, welches die Konsistenzordnung 2 in Raum und Zeit besitzt.

Möchte man die 2. Ordnung in der Zeitapproximation haben, dann benötigt man einen zentralen Differenzenquotienten. Dies erhält man bei der folgenden Approximation

$$\frac{u_i^{n+1} - u_i^n}{\Delta t} = \kappa \, \frac{u_{i+1}^{n+\frac{1}{2}} - 2u_i^{n+\frac{1}{2}} + u_{i-1}^{n+\frac{1}{2}}}{\Delta x^2} \, . \tag{6.34}$$

Die Werte auf der Zeitschicht $t_{n+\frac{1}{2}}$ stehen dabei allerdings nicht zur Verfügung. Für die zweite Ordnung ausreichend ist das Ersetzen mit dem arithmetischen Mittelwert

$$u_i^{n+\frac{1}{2}} = \frac{1}{2} \left(u_i^n + u_i^{n+1} \right) \, .$$

Dieses Verfahren nennt man das **Crank-Nicolson-Verfahren**. Wird der Mittelwert in (6.34) eingesetzt und die Differenzengleichung wieder in der

Form eines Gleichungssystems geschrieben, bei der alle unbekannten Werte auf der linken Seite und alle bekannten auf der rechten Seite auftauchen, dann lautet sie

$$-d\,u_{i-1}^{n+1} + 2\,(1+d)u_i^{n+1} - d\,u_{i+1}^{n+1} = d\,u_{i+1}^{n} + 2\,(1-d)u_i^{n} + d\,u_{i-1}^{n}\,.$$

Der Vergleich mit dem Verfahren erster Ordnung in der Zeit zeigt, dass lediglich die Koeffizienten auf der rechten und der linken Seite etwas modifiziert sind.

Wir führen nun für das Crank-Nicolson-Verfahren die Stabilitätsanalyse durch, setzen die entsprechenden Fourier-Komponenten (6.30) und (6.31) ein. Man erhält

$$G\left(-d\,e^{-\phi I} + 2\,(1+d) - d\,e^{\phi I}\right) = d\,e^{-\phi I} + 2(1-d) + d\,e^{\phi I}$$

mit dem Verstärkungsfaktor G. Mit der Beziehung (6.32) ergibt sich zunächst

$$G\big(-2d\,\cos\phi + 2\,(1+d)\big) = 2d\,\cos\phi + 2(1-d)$$

und weiter

$$G = \frac{1 - d(1-\cos\phi)}{1 + d(1-\cos\phi)}\,.$$

Da in dieser Gleichung $1-\cos\phi$ und d immer größer oder gleich Null sind, ist der Nenner immer größer als der Zähler. Damit ist aber der Betrag von G immer kleiner als Eins, das Crank-Nicolson-Verfahren ist somit stabil ohne jegliche Bedingung. Die Stabilität für das voll-implizite Verfahren erster Ordnung ergibt sich ganz analog. Dies heißt:

Das **Crank-Nicolson-Verfahren** und das **voll-implizite** Verfahren sind unbedingt stabil.

Beispiel 6.4 (Wärmeleitung, mit MAPLE-Worksheet). Wir wenden nun die impliziten Verfahren zur Berechnung der Temperaturverteilung des Metallstabes aus Beispiel 4.5 an. Das tridiagonale Gleichungssystem wird mit dem Thomas-Algorithmus gelöst. In Abbildung 6.10 sind die Ergebnisse für das voll-implizite Verfahren dargestellt. Im linken Bild ist die Temperatur zum Zeitpunkt $t = 0.5$ als Funktion von x und im rechten Bild an der Stelle $x = 0.5$ als Funktion von t aufgezeichnet. Im linken Bild liegen die Kurven fast aufeinander, im rechten Bild sind die Unterschiede etwas deutlicher zu sehen.

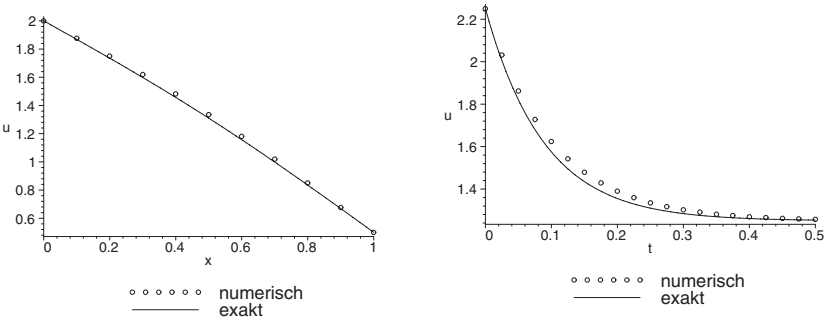


Abb. 6.10. Vergleich der numerischen Lösung des voll-impliziten Verfahrens mit der exakten Lösung

Die Rechnung wurde mit dem MAPLE-Worksheet ausgeführt. Deutlich bessere Werte ergeben sich mit dem Crank-Nicolson-Verfahren, wie Abbildung 6.11 zeigt. Für beide Verfahren lässt sich die Zeitschrittweite vergrößern, ohne dass es zu Instabilitäten kommt.

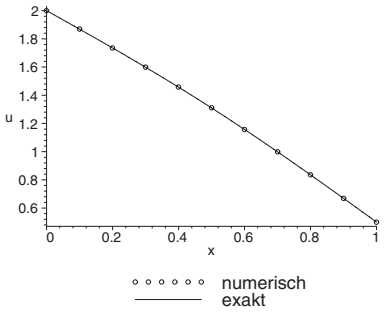


Abb. 6.11. Vergleich der numerischen Lösung des Crank-Nicolson-Verfahrens mit der exakten Lösung

□

Konsistenz und Zusammenfassung

Die Konsistenz der betrachteten Verfahren ergibt sich aus der Konsistenz der Differenzenquotienten zu den entsprechenden partiellen Ableitungen. Betrachten wir zunächst das explizite Verfahren. Wird die exakte Lösung des Problems $u = u(x,t)$ in das Verfahren eingesetzt, so ergibt sich

$$\frac{u(x,t+\Delta t)-u(x,t)}{\Delta t} - \kappa \frac{u(x+\Delta x,t)-2u(x,t)+u(x-\Delta x,t)}{\Delta x^2} = 0$$

an jedem Gitterpunkt (x_i, t_j).

Als nächstes werden die Taylor-Entwicklungen von $u(x,t+\Delta t)$, $u(x+\Delta x,t)$ und $u(x-\Delta x,t)$ um den Entwicklungspunkt (x,t) dort eingesetzt. Nach etwas Rechnung erhält man

$$u_t(x,t) - \kappa u_{xx}(x,t) = -\Delta t\ u_{tt}(x,t) + \frac{\Delta x^2}{12} u_{xxxx}(x,t) + O(\Delta t^2) + O(\Delta x^4)\ .$$

Auf der linken Seite steht die Wärmeleitungsgleichung und auf der rechten Seite das Residuum, d.h. der Diskretisierungsfehler. Die führenden Ordnungen des Diskretisierungsfehlers sind Δt und Δx^2, d.h. der Konsistenzfehler des expliziten Verfahrens ist $O(\Delta t, \Delta x^2)$. Allerdings gilt nach der Stabilitätsbedingung (6.33) für das explizite Verfahren zwingend $\Delta t \sim \Delta x^2$. Berücksichtigt man dies in der Konsistenzbetrachtung, dann kann man den Fehler des Verfahrens auch als $O(\Delta x^2)$ interpretieren. also einem Fehler 2. Ordnung.

Die Konsistenzordnung des voll-impliziten Verfahrens ergibt sich analog zu 1 für die Zeitapproximation und zu 2 für die Raumapproximation. Das Crank-Nicolson-Verfahren ist ein Verfahren 2. Ordnung in Zeit und Raum.

Ein weiteres Verfahren 2. Ordnung ist das **DuFort-Frankel-Verfahren**. Dabei wird die 2. Ordnung in der Zeit wieder durch eine zentrale Differenzenbildung eingeführt, wobei hier aber Näherungswerte auf drei Zeitniveaus benutzt werden. Es ist somit ein Mehrschrittverfahren, bei dem dann auch ein Anlaufstück nötig ist. Die Differenzengleichung lautet

$$\frac{u_i^{n+1}-u_i^{n-1}}{2\Delta t} = \kappa \frac{u_{i+1}^n - 2\tilde{u}_i^n + u_{i-1}^n}{\Delta x^2} \tag{6.35}$$

mit dem Mittelwert

$$\tilde{u}_i^n = \frac{1}{2}\left(u_i^{n+1} + u_i^{n-1}\right)\ .$$

Wird in der zentralen Raumdifferenz zweiter Ordnung der Wert u_i^n verwendet, dann zeigt sich, dass das zugehörige Verfahren

$$\frac{u_i^{n+1}-u_i^{n-1}}{2\Delta t} = \kappa \frac{u_{i+1}^n - 2u_i^n + u_{i-1}^n}{\Delta x^2} \tag{6.36}$$

bedingungslos instabil ist.

Bei der DuFort-Frankel-Formel wird eine **Stabilisierung** dadurch erreicht, dass ein impliziter Term durch den Mittelwert $\tilde{u}_i^n$ eingeführt wird. In der Mehrschrittformulierung lässt sich direkt nach u_i^{n+1} auflösen, so dass der Wert auf dem neuen Zeitlevel $n+1$ explizit zu berechnen ist:

$$(1+2d)u_i^{n+1} = (1-2d)u_i^{n-1} + 2d(u_{i+1}^n + u_{i-1}^n)$$

Durch diese direkte Auflösung des Gleichungssystems kann man es als ein explizites Verfahren betrachten.

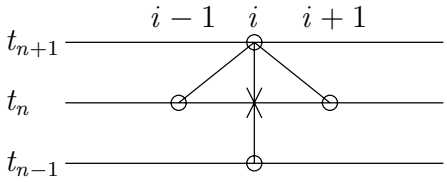


Abb. 6.12. Differenzenstern für das DuFort-Frankel-Verfahren

Bemerkungen:

1. Da zwei Zeitniveaus zum Start des DuFort-Frankel-Verfahrens benötigt werden, muss ein Einschrittverfahren, wie das Crank-Nicolson-Verfahren, in einem ersten Schritt eingesetzt werden.

2. Durch die Stabilisierung ist das DuFort-Frankel-Verfahren **bedingungslos stabil**. Dies ist sehr überraschend für ein im Prinzip explizites Verfahren. Allerdings tritt im lokalen Diskretisierungsfehler neben dem Fehler zweiter Ordnung in Raum und Zeit ein Term proportional zu $\left(\frac{\Delta t}{\Delta x}\right)^2$ auf. Diese Methode ist gewissermaßen von der Konsistenzordnung $O\left(\Delta x^2, \Delta t^2, \left(\frac{\Delta t}{\Delta x}\right)^2\right)$. Nach der Definition der Konsistenz ist das DuFort-Frankel-Verfahren **inkonsistent**. Um bei der Verkleinerung der Schrittweiten auch die entsprechende Genauigkeit zu erhalten, muss die Zeitschrittweite kleiner sein als die Raumschrittweite und damit ergibt sich aus der Genauigkeitsforderung eine Art Zeitschrittweitenbedingung.

Zusammenfassung:
Das **explizite Verfahren** mit zentralen Differenzen im Raum ist **bedingt stabil** mit der Schrittweitenbedingung

$$\kappa \frac{\Delta t}{\Delta x^2} \leq \frac{1}{2}. \tag{6.37}$$

Der Konsistenzfehler ist formal $O(\Delta t, \Delta x^2)$. Unter Einbeziehung der Stabilitätsbedingung kann man diesen jedoch mit $O(\Delta x^2)$ bezeichnen.
Das **voll-implizite** oder das **Crank-Nicolson-Verfahren** ist **bedingungslos stabil** mit dem Konsistenzfehler $O(\Delta t, \Delta x^2)$ bzw. $O(\Delta t^2, \Delta x^2)$.

Ebenso ist das explizite **DuFort-Frankel-Verfahren bedingungslos stabil**. Allerdings enthält hier der **Konsistenzfehler** einen Term, welcher proportional zum Verhältnis $\left(\frac{\Delta t}{\Delta x}\right)^2$ ist.

Bemerkung: Die Schrittweitenbedingung für das explizite Verfahren kann sehr restriktiv sein. Die Zeitschrittweite muss mit dem Quadrat der Raumschrittweite gegen Null gehen. Ist die Temperaturleitzahl nicht sehr klein, dann wird das explizite Näherungsverfahren ineffizient.

Vom Rechenaufwand her kann man etwa 10 Zeitschritte mit einem expliziten Verfahren im Vergleich zu einem mit einem impliziten Verfahren ausführen. Erzwingt die explizite Zeitschrittweitenbedingung (6.37) deutlich kleinere Zeitschritte und ist dies nicht durch die Genauigkeit motiviert, dann ist ein implizites Verfahren vorzuziehen.

Mehrdimensionale parabolische Probleme

Die hier betrachteten Verfahren können auf **mehrdimensionale parabolische Differenzialgleichungen** entsprechend erweitert werden. Wir betrachten dazu die Wärmeleitungsgleichung

$$u_t = \kappa \, \Delta u \tag{6.38}$$

in mehreren Raumdimensionen. Ein **explizites Verfahren** lässt sich direkt auf den mehrdimensionalen Fall übertragen, indem die entsprechenden Differenzenquotienten für alle Raumableitungen eingesetzt werden. Die Stabilitätsanalyse liefert wieder die Aussage, dass ein explizites Verfahren im parabolischen Fall nur bedingt stabil ist. Die parabolische Stabilitätsbedingung lautet bei zwei Raumdimensionen

$$\frac{\kappa \Delta t}{\Delta x^2} + \frac{\kappa \Delta t}{\Delta y^2} \leq \frac{1}{2} \tag{6.39}$$

und in drei Raumdimensionen

$$\frac{\kappa \Delta t}{\Delta x^2} + \frac{\kappa \Delta t}{\Delta y^2} + \frac{\kappa \Delta t}{\Delta z^2} \leq \frac{1}{2} \,. \tag{6.40}$$

Das **voll-implizite Verfahren** hat in zwei Raumdimensionen die Form

$$\frac{u_{i,j}^{n+1} - u_{i,j}^{n}}{\Delta t} = \kappa \left(\Delta_h u^{n+1} \right)_{i,j} \,.$$

Dabei bezeichnet $\Delta_h u$ eine Approximation des Laplace-Operators Δu. Schreibt man diese Gleichung um, dann erhält man in jedem Zeitschritt die elliptische Gleichung

$$\left(\Delta_h u^{n+1} \right)_{i,j} - \frac{1}{\kappa \Delta t} u_{i,j}^{n+1} = -\frac{1}{\kappa \Delta t} u_{i,j}^{n} \,. \tag{6.41}$$

Mit Hilfe einer der Methoden, welche im Kapitel über elliptische Gleichungen behandelt wurden, kann diese Gleichung gelöst werden. Das Standardverfahren wird hier sein, die zweiten partiellen Ableitungen jeweils mit einem zentralen Differenzenquotienten zu approximieren. Wird dies in die Gleichung eingesetzt, erhält man ein Gleichungssystem. Zu dessen Lösung kann man eine der in Kapitel 6.9 vorgestellten iterativen Verfahren einsetzen. Die elliptische Gleichung muss dabei in jedem Zeitschritt gelöst werden.

Ein anderer Ansatz ist ein **Splitting-** oder **Zwischenschritt-Verfahren**. Das mehrdimensionale Problem wird dabei in die eindimensionalen parabolischen Probleme

$$u_t = \kappa \, u_{xx} \,, \qquad u_t = \kappa \, u_{yy} \,, \qquad u_t = \kappa \, u_{zz}$$

zerlegt. Das numerische Verfahren läuft dann so ab, dass in jedem Zeitschritt diese Gleichungen nacheinander gelöst werden. Wenden wir das voll-implizite Verfahren an, dann ergibt sich in dem Zeitschritt von n auf $n+1$ die Folge von Problemen

$$\frac{u_{i,j,k}^{n+\frac{1}{3}} - u_{i,j,k}^{n}}{\Delta t} = \kappa \frac{u_{i+1,j,k}^{n+\frac{1}{3}} - 2u_{i,j,k}^{n+\frac{1}{3}} + u_{i-1,j,k}^{n+\frac{1}{3}}}{\Delta x^2} \,, \tag{6.42}$$

$$\frac{u_{i,j,k}^{n+\frac{2}{3}} - u_{i,j,k}^{n+\frac{1}{3}}}{\Delta t} = \kappa \frac{u_{i,j+1,k}^{n+\frac{2}{3}} - 2u_{i,j,k}^{n+\frac{2}{3}} + u_{i,j-1,k}^{n+\frac{2}{3}}}{\Delta y^2} \,, \tag{6.43}$$

$$\frac{u_{i,j,k}^{n+1} - u_{i,j,k}^{n+\frac{2}{3}}}{\Delta t} = \kappa \frac{u_{i,j,k+1}^{n+1} - 2u_{i,j,k}^{n+1} + u_{i,j,k-1}^{n+1}}{\Delta z^2} \,. \tag{6.44}$$

Dabei sind die Werte mit den rationalen oberen Indizes nur Hilfswerte, sie haben keine physikalische Relevanz. Bei dem Splitting- oder auch Dimensionen-Splitting-Verfahren werden die eindimensionalen Probleme nacheinander gelöst. Dies hat den Vorteil, dass nur tridiagonale Gleichungssysteme auftreten, welche mit dem Thomas-Algorithmus sehr effizient gelöst werden können.

Das Splittingverfahren lässt sich auch mit dem Crank-Nicolson-Verfahren kombinieren. Allerdings muss man in diesem Fall einen Doppelschritt ausführen. Während der erste Schritt unverändert wie oben beschrieben abläuft, muss man die Reihenfolge im zweiten Schritt umdrehen. Man beginnt mit dem Problem in z-Richtung. Nur so lässt sich die 2. Ordnung in der Zeit für das mehrdimensionale Verfahren aufrecht erhalten.

6.3 Hyperbolische Differenzialgleichungen

Die einfachste hyperbolische Gleichung ist die lineare Transportgleichung

$$u_t + a\, u_x = 0 \quad \text{mit} \quad a \in \mathbb{R}\,. \tag{6.45}$$

Die Charakteristiken sind für eine konstante Geschwindigkeit a Geraden. Die Lösung ist konstant entlang dieser Charakteristiken, wie dies in Abschnitt 4.1 gezeigt wurde.

Wir betrachten zunächst **explizite Differenzenverfahren**. Im letzten Abschnitt haben wir für eine parabolische Differenzialgleichung gesehen, dass das explizite Verfahren mit zentralen Differenzen im Raum bedingt stabil ist. Auf das hyperbolische Problem angewandt lautet die Differenzengleichung

$$\frac{u_i^{n+1} - u_i^n}{\Delta t} + a\,\frac{u_{i+1}^n - u_{i-1}^n}{2\Delta x} = 0\,, \tag{6.46}$$

die man mit $c = a\frac{\Delta t}{\Delta x}$ in die Form

$$u_i^{n+1} = u_i^n - \frac{c}{2}\left(u_{i+1}^n - u_{i-1}^n\right)$$

umschreiben kann.

Wir untersuchen die **Stabilität** dieses Verfahrens mit Hilfe der von Neumann Methode und setzen dazu den Ansatz

$$u_i^n = U^n e^{I\phi i}$$

in die Differenzengleichung ein. Zunächst ergibt sich

$$U^{n+1} e^{I\phi i} = U^n e^{I\phi i} - \frac{c}{2}\left(U^n e^{I\phi(i+1)} - U^n e^{I\phi(i-1)}\right)\,.$$

Die Division durch $e^{I\phi i}$ liefert die Verstärkungsfunktion $G = \frac{U^{n+1}}{U^n}$ zu

$$G = 1 - \frac{c}{2}\left(e^{I\phi} - e^{-I\phi}\right)\,.$$

Aus der Eulerschen Formel

$$e^{I\phi} = \cos\phi + I\sin\phi \tag{6.47}$$

ergibt sich die Beziehung

$$\sin\phi = \frac{e^{I\phi} - e^{-I\phi}}{2I}\ , \tag{6.48}$$

womit sich der Term mit der e-Funktion eliminieren lässt und man

$$G = 1 - cI\sin\phi \tag{6.49}$$

erhält. Dies ist in diesem Fall eine komplexe Funktion mit einem Real- und Imaginärteil. Die Bedingung für Stabilität, für die der Absolutbetrag von G kleiner als eins sein muss, führt auf

$$|G|^2 = 1 + c^2\sin^2\phi. \tag{6.50}$$

Da der hintere Term positiv ist, ist der Betrag allerdings immer größer oder gleich 1. Das lässt nur eine Folgerung zu: Das explizite Verfahren 1. Ordnung in der Zeit mit zentralen räumlichen Differenzen ist somit **bedingungslos instabil** und für praktische Anwendungen nicht zu gebrauchen.

Nachdem die explizite Zeitapproximation mit zentralen Differenzen im Raum zu keinem brauchbaren Verfahren geführt hat, probieren wir **einseitige Differenzen im Raum**. Ein **explizites Verfahren mit linksseitigen Differenzen** lautet

$$\boxed{\frac{u_i^{n+1} - u_i^n}{\Delta t} + a\,\frac{u_i^n - u_{i-1}^n}{\Delta t} = 0} \tag{6.51}$$

oder nach u_i^{n+1} aufgelöst

$$u_i^{n+1} = u_i^n - c\left(u_i^n - u_{i-1}^n\right)\ .$$

Die von Neumannsche Stabilitätsanalyse führt hier auf die Gleichung

$$U^{n+1} = U^n e^{I\phi i} - c\left(U^n e^{I\phi i} - U^n e^{I\phi(i-1)}\right)$$

und nach kurzer Rechnung auf die Verstärkungsfunktion

$$G = 1 - c + ce^{-I\phi}\ .$$

Mit der Eulerschen Formel (6.47) kann der Term mit der e-Funktion wiederum eliminiert werden und man erhält

$$G = 1 - c + c\cos\phi - Ic\sin\phi\,.$$

Berechnet man das Quadrat des Betrags von G, erhält man

$$|G|^2 = 2c^2(1-\cos\phi) - 2c(1-\cos\phi) + 1\,.$$

Dies ist eine Parabel bezüglich c. Die Nullstellen von $|G|^2 - 1$ sind bei $c = 0$ und bei $c = 1$. Zwischen diesen beiden Grenzen ist diese Funktion kleiner Null, so dass die Stabilitätsbedingung erfüllt ist, falls

$$0 \le c < 1\,.$$

Damit erhält man für das Verfahren (6.51) die Stabilitätsbedingung

$$0 \le a\frac{\Delta t}{\Delta x} < 1\,. \tag{6.52}$$

Analog zu der obigen Rechnung kann man eine Stabilitätsanalyse für das **explizite Verfahren mit rechtsseitigen Differenzen**

$$\frac{u_i^{n+1} - u_i^n}{\Delta t} + a\frac{u_{i+1}^n - u_i^n}{\Delta x} = 0 \tag{6.53}$$

durchführen. Man erhält hier für die von Neumannsche Stabilitätsanalyse die Gleichung

$$U^{n+1}e^{I\phi i} = U^n e^{I\phi i} - c\left(U^n e^{I\phi(i+1)} - U^n e^{I\phi i}\right)$$

und die Verstärkungsfunktion

$$G = 1 - ce^{I\phi} + c\,.$$

Wegen

$$e^{I\phi} = \cos\phi + I\sin\phi$$

ergibt sich die komplexe Funktion

$$G = 1 + c(1-\cos\phi) - I\sin\phi$$

Das Quadrat des Betrags lautet

$$|G|^2 = 2c^2(1-\cos\phi) + 2c(1-\cos\phi) + 1\,. \tag{6.54}$$

Es ergibt sich ganz ähnlich zu oben die Stabilitätsbedingung

$$-1 < a\frac{\Delta t}{\Delta x} \leq 0\,.$$

Fassen wir diese beiden Ergebnisse zusammen, dann erhalten wir die folgenden Aussagen:

Das explizite Verfahren mit **rechtsseitigen Differenzen** ist **bedingt stabil** für $a \leq 0$, das mit **linksseitigen Differenzen** für $a \geq 0$.
Die Schrittweitenbedingung lautet in beiden Fällen

$$|a|\frac{\Delta t}{\Delta x} < 1\,. \tag{6.55}$$

Diese Bedingung nennt man **Courant-Friedrichs-Lewy** oder kurz **CFL-Bedingung**.

Aus der Stabilitätsanalyse folgt, dass für $a < 0$ das Verfahren mit linksseitigen Differenzen bedingungslos instabil ist. Ebenso ist für $a > 0$ das Verfahren mit rechtsseitigen Differenzen bedingungslos instabil. Wenn wir an die exakte Lösung denken, welche in Abschnitt 4.1 abgeleitet wurde, dann spiegelt diese Situation gerade die Eigenschaften der exakten Lösung wider. Für $a < 0$ ist die Lösung bestimmt durch den Transport von links nach rechts, da die Transportgeschwindigkeit positiv ist. Die gesamte Information kommt von links. Dieser Informaitonstransport kann nur dann richtig in das Näherungsverfahren übernommen werden, wenn eine linksseitige Differenzenbildung in dem Verfahren benutzt wird. Ansonsten wird die Information von der falschen Seite geholt. Den umgekehrten Sachverhalt hat man dann im Fall $a > 0$ für die rechtsseitige Differenzenbildung.

Bemerkungen:

1. Die CFL-Bedingung (6.55) bedeutet geometrisch, dass alle Charakteristiken der exakten Lösung innerhalb des Differenzensterns verlaufen. Der numerische Abhängigskeitsbereich überdeckt den der exakten Lösung. Dies verlangt insbesondere, dass der Zeitschritt so klein gewählt werden muss, dass die Information in einem Zeitschritt nur von einer Gitterzelle in die benachbarte transportiert wird.

2. Ist a und die Raumschrittweite gegeben, dann ist (6.55) eine Forderung an die Zeitschrittweite. In einem numerischen Verfahren wird der Zeitschritt meist etwas kleiner gewählt, um sicher zu sein, dass man trotz Rundungsfehler im Stabilitätsgebiet rechnet.

Die Zahl $\sigma < 1$ mit

$$\Delta t = \sigma \frac{\Delta x}{|a|} \tag{6.56}$$

nennt man die **Courantzahl**.

3. Diese Stabilitätsbedingung wurde erstmals von Courant, Friedrichs und Lewy 1928 [10] formuliert.

Wir können die beiden Verfahren mit einseitiger Differenzenbildung in ein Verfahren für beliebiges $a \in \mathbb{R}$ umformulieren:

$$u_i^{n+1} = u_i^n - a\frac{\Delta t}{\Delta x} \begin{cases} u_i^n - u_{i-1}^n & \text{für} \quad a \geq 0 \\ u_{i+1}^n - u_i^n & \text{für} \quad a < 0 \end{cases}$$

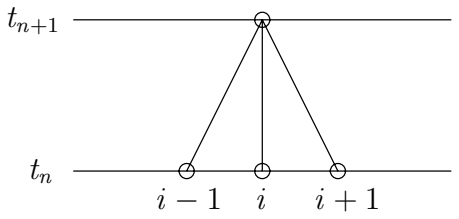
oder auch in die Form

$$\boxed{u_i^{n+1} = u_i^n - \frac{\Delta t}{\Delta x}\left[a^+\left(u_i^n - u_{i-1}^n\right) + a^-\left(u_{i+1}^n - u_i^n\right)\right]} \tag{6.57}$$

mit

$$a^+ = \max(a, 0)\,, \qquad a^- = \min(a, 0)$$

bringen. Der Differenzenstern sieht folgendermaßen aus



Dieses Verfahren nennt man **Upwind-Verfahren** oder **CIR-Verfahren** nach Courant, Isaacson und Rees. Der Name Upwind-Verfahren leitet sich daraus ab, dass die Richtung der Differenzenbildung stromauf genommen wird, dorthin, wo die Information herkommt.

Beispiel 6.5 (CIR-Verfahren für die lineare Transportgleichung, mit MAPLE**-Worksheet).** Wir berechnen mit dem CIR-Verfahren das Anfangswertproblem

$$u_t + a\, u_x = 0$$
$$u(x, 0) = e^{-(x-4)^2}$$

im Intervall $[0, 16]$ mit konstantem $a = 0.5$. Da a positiv ist, wird der linksseitige Differenzenquotient genommen.

In diesem Beispiel ist die Lösung eines Anfangs-Randwertproblems gesucht. Im Prinzip ist ein hyperbolisches Problem stets ein reines Anfangswertproblem. Praktische Probleme sind allerdings meist in einem endlichen Gebiet vorgegeben oder müssen für die numerische Approximation auf ein endliches Gebiet eingeschränkt werden. Randwerte dürfen nur dort vorgegeben werden, wo die Information in das Rechengebiet hinein läuft. In unserem Fall $a > 0$ ist das Bild der Charakteristiken in Abbildung 6.13 gezeichnet.

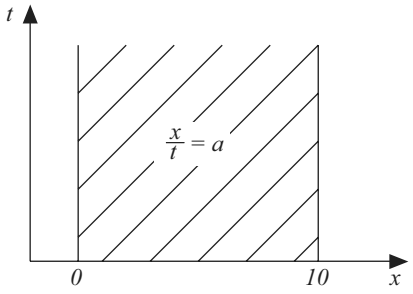


Abb. 6.13. Verlauf der Charakteristiken im (x, t)-Diagramm

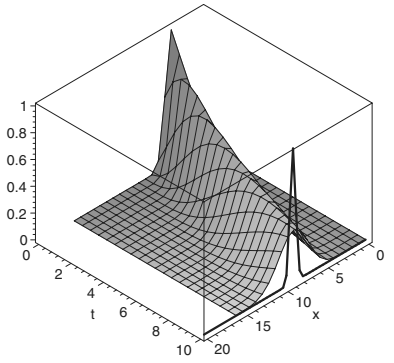


Abb. 6.14. Numerische Lösung zu Beispiel 6.5 als Funktion von x und t

Da die Charakteristiken von links nach rechts laufen, müssen Randwerte am linken Rand bei $x = 0$ vorgeschrieben werden. Dadurch, dass das CIR-Verfahren die Ausbreitungsrichtung mit in der Differenzenbildung berück-

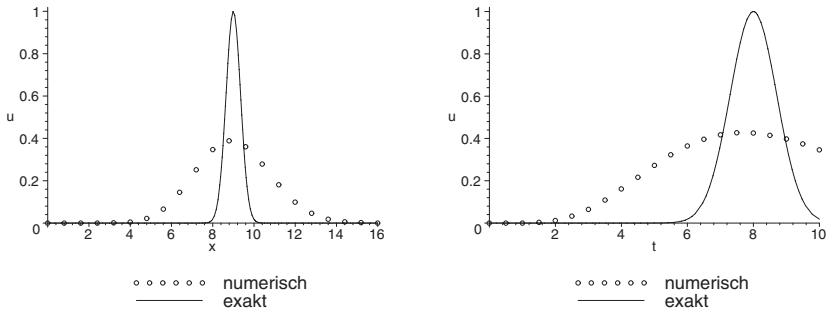


Abb. 6.15. Exakte und numerische Lösung zu Beispiel 6.5, links am Zeitpunkt $t = 10$ als Funktion von x und rechts am Punkt $x = 8$ als Funktion von t

sichtigt, passt dies sehr gut zusammen: Das numerische Verfahren benötigt nur dort Randwerte.

Numerische Ergebnisse sind in den Abbildungen 6.14 und 6.15 dargestellt. Die Diskretisierungsparameter sind in dieser Rechnung

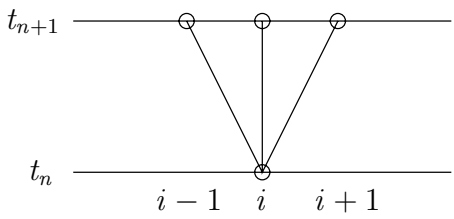
$$\Delta x = \frac{16}{20}\,, \qquad \Delta t = \frac{10}{20}\,.$$

Die exakte Lösung ist hier mit einer durchgehenden Linie gezeichnet. Die Näherungswerte aus dem CIR-Verfahren an den Gitterpunkten sind mit einem kleinen Kreis gekennzeichnet. Es zeigt sich, dass die Lösung zwar richtig wiedergegeben wird, die Amplitude aber im Gegensatz zur exakten Lösung mit der Zeit kleiner und kleiner wird. Man nennt dies **numerische Dämpfung** oder **numerische Dissipation**. Bei dem Verfahren erster Ordnung ist diese sehr deutlich zu sehen. Um die Dämpfung zu reduzieren, muss die Anzahl der Gitterpunkte erhöht werden. □

Die Stabilitätsanalyse für das **voll-implizite Verfahren**

$$\frac{u_i^{n+1} - u_i^n}{\Delta t} + a\frac{u_{i+1}^{n+1} - u_{i-1}^{n+1}}{2\Delta x} = 0 \tag{6.58}$$

zeigt, dass die implizite Approximation der Zeitableitung mit dem rückwärtsgenommenen Differenzenquotienten das Verfahren stabilisiert. Es ist stabil ohne eine Bedingung an die Zeitschrittweite. Wenn wir uns den Differenzenstern dieses Verfahrens anschauen,



dann sehen wir, dass auf der neuen Zeitschicht drei Punkte in die Differenzengleichung eingehen. Damit überdeckt der numerische Abhängigkeitsbereich automatisch den der exakten Lösung. Zur Lösung der Differenzengleichung muss allerdings ein tridiagonales Gleichungssystem gelöst werden.

Auch das **Crank-Nicolson-Verfahren** kann auf eine hyperbolische Gleichung angewandt werden. Für die lineare Transportgleichung hat es die Form

$$\frac{u_i^{n+1} - u_i^n}{\Delta t} + a \frac{u_{i+1}^{n+\frac{1}{2}} - u_{i-1}^{n+\frac{1}{2}}}{2\Delta x} = 0 \tag{6.59}$$

mit

$$u_i^{n+\frac{1}{2}} = \frac{1}{2}\left(u_i^{n+1} + u_i^n\right) . \tag{6.60}$$

Auch dieses Verfahren ist **bedingungslos stabil**. Wie beim voll-impliziten Verfahren muss ein tridiagonales Gleichungssystem gelöst werden. Der Konsistenzfehler ergibt sich aus der Approximation der Ableitungen durch die entsprechenden Differenzenquotienten. Für das Crank-Nicolson-Verfahren ist der Konsistenzfehler $O(\Delta t^2, \Delta x^2)$.

Beispiel 6.6 (Implizite Verfahren für die lineare Transportgleichung, mit MAPLE**-Worksheet).** Wir greifen das Beispiel 6.5 auf und wenden die impliziten Verfahren an. Dadurch, dass zentrale Differenzenquotienten im Raum benutzt werden, müssen auch Randwerte am rechten Rand vorgeschrieben werden. Diese sind dabei nicht durch das physikalische Problem bestimmt, sondern durch die Lösung. Eine Möglichkeit der Vorgabe von Randbedingungen ist, am rechten Rand einen linksseitigen Differenzenquotienten zu wählen, der dann aber auch die Konsistenzordnung zwei besitzen sollte.

Die numerischen Ergebnisse mit dem voll-impliziten Verfahren mit der Diskretisierung von Beispiel 6.5 sind in Abbildung 6.16 aufgezeichnet. Man sieht hier neben der numerischen Dämpfung der Amplitude deutliche Wellen hinter der Glockenkurve. Das implizite Verfahren mit zentralen Differenzen im

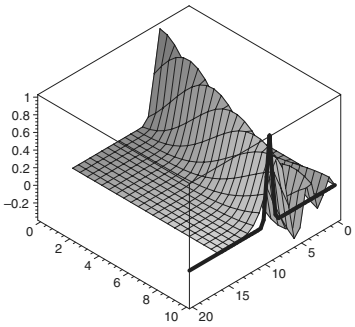


Abb. 6.16. Numerische Lösung des voll-impliziten Verfahrens

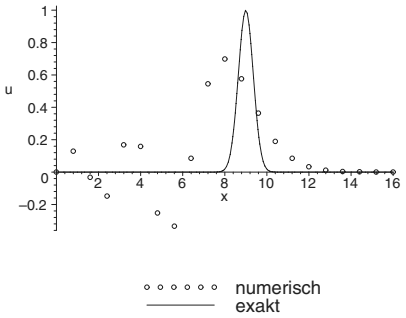


Abb. 6.17. Numerische Lösung des voll-impliziten Verfahrens zum Zeitpunkt $t = 10$

Raum besitzt neben dem **dissipativen Fehler** auch einen sogenannten **dispersiven Fehler**. Deutlicher wird dies in einem Bild für einen festen Zeitpunkt t, wie es in Abbildung 6.17 dargestellt ist. Die starken Oszillationen deuten an, dass die Anzahl der Gitterpunkte hier nicht ausreicht, um die Gauß-Kurve gut aufzulösen. Um bessere Ergebnisse zu erhalten, muss die Anzahl der Gitterpunkte erhöht werden. In Abbildung 6.18 sieht man das Ergebnis für 80 Gitterpunkte im Raum und 40 Zeitschritte. Hier sind die Oszillationen dann verschwunden, die Amplitude ist durch die numerische Dämpfung deutlich reduziert.

Wie erwartet, liefert das Crank-Nicolson-Verfahren als Verfahren 2. Ordnung in Raum und Zeit bessere Ergebnisse. Dies ist bei gleicher feiner Diskretisierung in Abbildung 6.19 festgehalten. Man sieht hier im Gegensatz zum voll-impliziten Verfahren noch kleine Wellen hinter der Gauß-Kurve, die Amplitude ist aber wesentlich besser wiedergegeben.

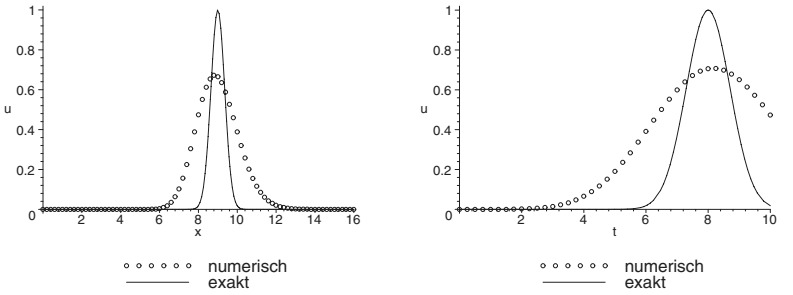


Abb. 6.18. Numerische Lösung des voll-impliziten Verfahrens für Beispiel 6.6 mit verfeinerter Diskretisierung

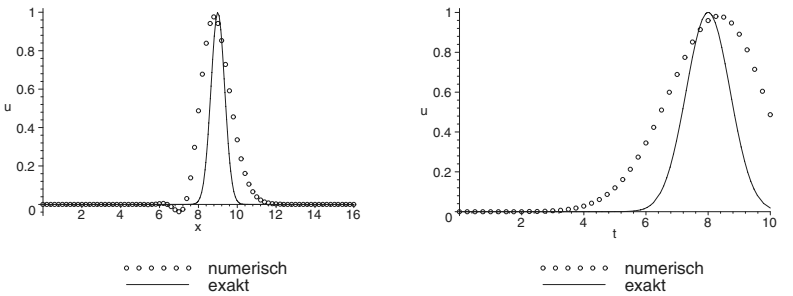


Abb. 6.19. Numerische Lösung des Crank-Nicolson-Verfahren für Beispiel 6.6 mit feiner Diskretisierung

Die numerischen Resultate werden im MAPLE -Worksheet berechnet. □

Ein explizites Verfahren zweiter Ordnung in Raum und Zeit ist das **Lax-Wendroff-Verfahren** (siehe z.B. [49]. Für die zweite Ordnung in Raum und Zeit benötigt man zentrale Differenzen in Raum und Zeit. Ein solches Verfahren hat eine Darstellung in der Form

$$\frac{u_i^{n+1} - u_i^n}{\Delta t} + \frac{u_{i+\frac{1}{2}}^{n+\frac{1}{2}} - u_{i-\frac{1}{2}}^{n+\frac{1}{2}}}{\Delta x} = 0\,. \tag{6.61}$$

Das implizite Crank-Nicolson-Verfahren kann man in diese Form bringen, wenn man die Mittelwerte

$$u_{i+\frac{1}{2}}^{n+\frac{1}{2}} = \frac{1}{4}\left(u_{i+1}^{n+1} + u_i^{n+1} + u_{i+1}^n + u_i^n\right)$$

einführt.

Die Idee beim Lax-Wendroff-Verfahren ist, dass man zunächst eine Taylor-Entwicklung in der Zeit für einen halben Zeitschritt betrachtet:

$$u(x, t + \frac{\Delta t}{2}) = u(x, t) + \frac{\Delta t}{2} u_t(x, t) + \frac{\Delta t^2}{8} u_{tt}(x, t) + O\left(\Delta t^3\right) \ .$$

Die Zeitableitung kann mit Hilfe der Transportgleichung in eine Raumableitung umformuliert werden, denn es gilt

$$u_t = -a\, u_x \ .$$

Eingesetzt ergibt sich

$$u(x, t + \frac{\Delta t}{2}) = u(x, t) - \frac{\Delta t}{2} a u_x + O\left(\Delta t^2\right) \ .$$

Führen wir nun eine Approximation dieser Beziehung in der Form

$$u_{i+\frac{1}{2}}^{n+\frac{1}{2}} = \frac{1}{2}\left(u_{i+1}^n + u_i^n\right) - \frac{\Delta t}{2} a \frac{u_{i+1}^n - u_i^n}{\Delta x} + O\left(\Delta t^2, \Delta t \Delta x, \Delta x^2\right)$$

durch, so erhalten wir eine Approximation der Werte an den Zwischenpunkten $(x_{i+\frac{1}{2}}, t_{i+\frac{1}{2}})$. Diese so berechneten Werte können dann in die Gleichung (6.61) eingesetzt werden. Dies kann man noch zusammenfassen und nach kurzer Rechnung erhält man

$$u_i^{n+1} = u_i^n - a\Delta t \left(\frac{u_{i+1}^n - u_{i-1}^n}{2\Delta x}\right) + \frac{1}{2} a^2 \Delta t^2 \left(\frac{u_{i+1}^n - 2u_i^n + u_{i-1}^n}{\Delta x^2}\right) \ . \quad (6.62)$$

Dieses Verfahren besitzt nach Konstruktion die Konsistenzordnung $O(\Delta x^2, \Delta t^2)$. Es ist unter der CFL-Bedingung (6.55) bedingt stabil.

Beispiel 6.7 (Lax-Wendroff-Verfahren für die lineare Transportgleichung, mit MAPLE**-Worksheet).** Das Lax-Wendroff-Verfahren für das Anfangswertproblem 6.5 liefert die numerischen Ergebnisse in Abbildung 6.20. Auf dem groben Gitter sieht man hier auch deutliche Wellen nach der Glockenkurve. Auf dem feineren Gitter mit 80 Gitterintervallen im Raum und 40 Zeitschritten wird dagegen die exakte Lösung dann sehr gut approximiert. Dies ist in Abbildung 6.21 dargestellt. Die numerischen Lösungen des Crank-Nicolson-Verfahrens und des Lax-Wendroff-Verfahrens sind auf diesem feineren Gitter sehr ähnlich.

Die Ergebnisse zeigen insbesondere, dass genügend Gitterpunkte zur Auflösung der Gaußschen Glockenkurve zur Verfügung stehen müssen, um gute numerische Ergebnisse zu erhalten. Auf dem feinen Gitter ist die Anzahl etwa 20 Gitterpunkte für die Gauß-Kurve. □

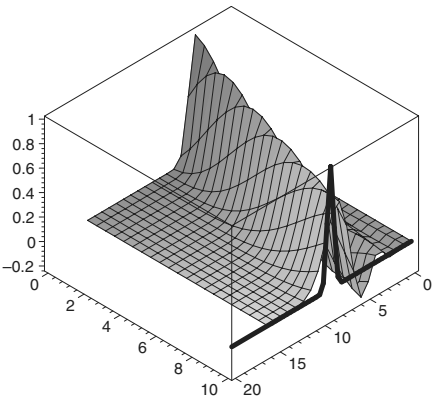


Abb. 6.20. Numerische Lösung des Lax-Wendroff-Verfahrens als Funktion von x und t

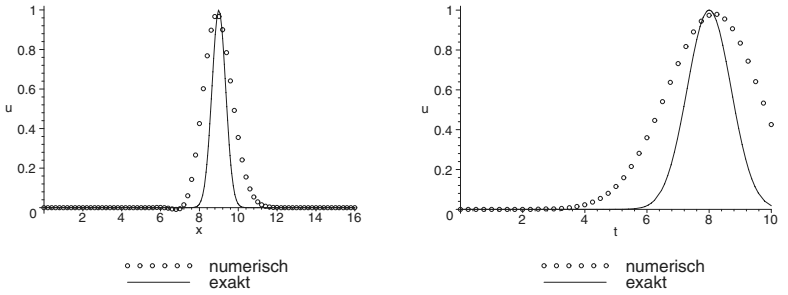


Abb. 6.21. Numerische Lösung mit dem Lax-Wendroff-Verfahren links für den Zeitpunkt $t = 10$ als Funktion von x und rechts für den Gitterpunkt $x = 8$ als Funktion von t, jeweils auf dem feinem Gitter berechnet

Zusammenfassung: Numerische Methoden für die lineare Transportgleichungen
Das **explizite** Verfahren mit **zentralen** Differenzen im Raum und der Konsistenzordnung $\mathbf{O(\Delta t, \Delta x^2)}$ ist **bedingungslos instabil**.
Zur Stabilisierung muss die Differenzenbildung einseitig entsprechend der Richtung des Transports gewählt werden. Dies nennt man **Upwind-** oder **CIR-Verfahren**. Dieses Verfahren hat die Konsistenzordnung $\mathbf{O(\Delta t, \Delta x)}$.
Das **Lax-Wendroff-Verfahren** ist ein **explizites** Verfahren mit der Konsistenzordnung $\mathbf{O(\Delta t^2, \Delta x^2)}$.
Die expliziten Verfahren sind **bedingt stabil** mit der Stabilitätsbedingung bzw. Schrittweitenbedingung

$$|a|\frac{\Delta t}{\Delta x} < 1 \; .$$

Das **voll-implizite Verfahren** mit dem Konsistenzfehler $\mathbf{O(\Delta t, \Delta x^2)}$ und das **Crank-Nicolson-Verfahren** $\mathbf{O(\Delta x^2, \Delta t^2)}$ sind **bedingungslos stabil**.

Bemerkung: Der Vergleich der Schrittweitenbedingungen für explizite Verfahren im hyperbolischen Fall (6.55) und im parabolischen Fall (6.37) zeigen den folgenden Unterschied: Bei parabolischen Gleichungen muss der Zeitschritt proportional dem Quadrat der Raumschrittweite, im hyperbolischen Fall nur proportional der Raumschrittweite gewählt werden. Während für eine parabolische Gleichung das Standardverfahren das implizite Crank-Nicolson-Verfahren ist, werden im hyperbolischen Fall zur Berechnung von instationären Lösungen oft explizite Verfahren eingesetzt.

Auf **Systeme von Transportgleichungen** können alle Formeln für die vorgestellten Verfahren außer dem Upwind-Verfahren direkt übertragen werden, indem man die skalaren Funktionen durch Vektoren ersetzt. Zur Definition der Upwind-Verfahren muss man etwas ausholen. Bei einem System von Transport- oder Wellengleichungen können verschiedene Wellengeschwindigkeiten auftreten. Ist das Vorzeichen dieser Geschwindigkeiten gleich, dann sind die Formeln wie bei den anderen Verfahren direkt übertragbar. Gibt es Wellengeschwindigkeiten mit unterschiedlichen Vorzeichen, geht dies nicht so einfach.

Setzen wir voraus, dass das System aus m Gleichungen

$$\mathbf{u}_t + \mathbf{A}\mathbf{u}_x = 0$$

streng hyperbolisch und somit die Matrix A diagonalisierbar ist, dann können wir es in die **charakteristische Normalform**

$$\mathbf{w}_t + \mathbf{\Lambda}\mathbf{w}_x = 0$$

mit der Diagonalmatrix $\mathbf{\Lambda} = \mathbf{diag}(a_1, \ldots, a_m)$ umschreiben (siehe Kapitel 4.1).

Damit hat man m nicht gekoppelte skalare Transportgleichungen, auf die das Upwind-Verfahren angewandt werden kann. In der Matrixschreibweise sieht dies dann folgendermaßen aus:

$$\frac{\mathbf{w}_i^{n+1} - \mathbf{w}_i^n}{\Delta t} + \mathbf{\Lambda}^+ \left(\frac{\mathbf{w}_i^n - \mathbf{w}_{i-1}^n}{\Delta x} \right) + \mathbf{\Lambda}^- \left(\frac{\mathbf{w}_{i+1}^n - \mathbf{w}_i^n}{\Delta x} \right) = 0 \, .$$

Dabei sind $\mathbf{\Lambda}^+$ und $\mathbf{\Lambda}^-$ Diagonalmatrizen, welche nichtnegative bzw. nichtpositive Eigenwerte besitzen:

$$\mathbf{\Lambda}^+ = \begin{pmatrix} a_1^+ & & 0 \\ & a_2^+ & \\ & & \ddots \\ 0 & & a_m^+ \end{pmatrix}, \quad \mathbf{\Lambda}^- = \begin{pmatrix} a_1^- & & 0 \\ & a_2^- & \\ & & \ddots \\ 0 & & a_m^- \end{pmatrix} \quad \text{mit} \quad \begin{array}{l} a_i^+ = \max(a_i, 0) \\ a_i^- = \min(a_i, 0) \, . \end{array}$$

Nun transformiert man das System in charakteristischer Normalform wieder zurück, indem es mit der Matrix der Eigenvektoren $\mathbf{R}$ von links multipliziert wird:

$$\begin{aligned}\mathbf{R}\mathbf{w}_i^{n+1} = \mathbf{R}\mathbf{w}_i^n &- \frac{\Delta t}{\Delta x}\mathbf{R}\boldsymbol{\Lambda}^+(\mathbf{R}^{-1}\mathbf{R})\left(\mathbf{w}_i^n - \mathbf{w}_{i-1}^n\right)\\ &- \frac{\Delta t}{\Delta x}\mathbf{R}\boldsymbol{\Lambda}^-(\mathbf{R}^{-1}\mathbf{R})\left(\mathbf{w}_{i+1}^n - \mathbf{w}_i^n\right)\ .\end{aligned}$$

Mit den Definitionen

$$\mathbf{A}^+ := \mathbf{R}\boldsymbol{\Lambda}^+\mathbf{R}^{-1}\,, \qquad \mathbf{A}^- := \mathbf{R}\boldsymbol{\Lambda}^-\mathbf{R}^{-1}$$

erhält man dann

$$u_i^{n+1} = u_i^n - \frac{\Delta t}{\Delta x}A^+\left(u_i^n - u_{i-1}^n\right) - \frac{\Delta t}{\Delta x}A^-\left(u_{i+1}^n - u_i^n\right)\ . \tag{6.63}$$

Dieses Verfahren für Systeme von Transport- oder Wellengleichungen nennt man wie im skaleren Fall **Upwind-Verfahren** oder **CIR-Verfahren**.

Die hier vorgestellten Verfahren können auf **nichtlineare Probleme** übertragen werden, so z.B. auf eine quasilineare Transportgleichung

$$u_t + a(u)\,u_x = 0\ . \tag{6.64}$$

Allerdings muss man dann die Geschwindigkeit $a(u)$ an einem entsprechenden Punkt auswerten. Das **voll-implizite Verfahren** mit einem zentralen Differenzenquotienten für die Raumableitung sieht folgendermaßen aus:

$$\frac{u_i^{n+1} - u_i^n}{\Delta t} + a(u_i^{n+1})\frac{\left(u_{i+1}^{n+1} - u_{i-1}^{n+1}\right)}{2\Delta x} = 0. \tag{6.65}$$

Für eine nichtlineare Gleichung (6.64) ist dies eine nichtlineare Differenzengleichung. Das zugehörige Gleichungssystem ist dann nichtlinear und muss mit einem Iterationsverfahren gelöst werden, die im Anhang B aufgeführt sind.

Das **explizite Upwind-Verfahren** macht dann Schwierigkeiten, wenn sich die Richtung der Charakteristik über die Stützstellen x_{i-1}, x_i, x_{i+1} hinweg ändert, d.h. wenn sich dort ein sogenannter Schallpunkt befindet. Die Richtung der Differenzenbildung hängt dann davon ab, an welcher Stelle $a(u)$ ausgewertet wird. Es kann dort leicht zu Wellen oder Instabilitäten kommen. Im Falle einer Erhaltungsgleichung

$$u_t + f(u)_x = 0 \tag{6.66}$$

werden wir später im Kapitel über Finite-Volumen-Verfahren das Problem der Approximation am Schallpunkt ausführlicher betrachten und Lösungen angeben.

Ein weiteres Problem ist das Auftreten von Unstetigkeiten oder starken Gradienten in der exakten Lösung, wie wir dies schon in Kapitel 4 untersucht haben. In diesem Fall sind die Aussagen für Differenzenverfahren nicht mehr gerechtfertigt, da die wesentliche Voraussetzung der stetigen Differenzierbarkeit und damit die Grundlage eines Differenzenverfahrens nicht mehr erfüllt ist. In diesem Fall sollte man besser ein Finite-Volumen-Verfahren auswählen.

Es gibt alerdings einen Ansatz, das **Verfahren von Lax**, welches auch zur Berechnung von starken Gradienten eingesetzt werden kann. Dieses Verfahren lautet für die Erhaltungsgleichung (6.66)

$$u_i^{n+1} = \frac{1}{2}\left(u_{i+1}^n + u_{i-1}^n\right) - \frac{\Delta t}{2\Delta x}\left(f(u_{i+1}^n) - f(u_{i-1}^n)\right) \tag{6.67}$$

und ist bedingt stabil unter der CFL-Bedingung. Die Idee, welche dahinter steckt, zeigt sich, wenn wir dieses Verfahren umformulieren. Wir fügen den Term u_i^n auf der rechten Seite hinzu und ziehen ihn wieder ab. Etwas umgestellt ergibt sich

$$u_i^{n+1} = u_i^n - \frac{\Delta t}{2\Delta x}\left(f(u_{i+1}^n) - f(u_{i-1}^n)\right) + \frac{1}{2}\left(u_{i+1}^n - 2u_i^n + u_{i-1}^n\right) \ .$$

Der hintere Term entspricht gerade einer Approximation von $\frac{1}{2}\Delta x^2 u_{xx}$. Dies ist ein parabolischer Term, den man physikalisch als Reibungs- oder Wärmeleitungsterm interpretieren kann.

Durch die Mittelwertbildung bei dem Lax-Verfahren führt man zusätzlich **numerische Dissipation** ein, welche z.B. eine Stoßwelle über mehrere Gitterpunkte verschmiert. Die Stoßwelle wird ersetzt durch ein steiles viskoses Profil ersetzt. Das Differenzenverfahren wird dann auf dieses über mehrere Gitterpunkte verschmierte stetige Profil angewandt. Dieser Ansatz liefert eine stabile Approximation von Stoßwellen, besitzt aber eine starke numerische Dissipation in anderen Bereichen. Im Allgemeinen können Stoßwellen oder starke Gradienten in der Lösung mit Finite-Volumen-Verfahren, welche lokal eine Approximation des integralen Erhaltungsgesetzes darstellen, numerisch besser aufgelöst werden. Dies wird in Kapitel 8 beschrieben. In diesem Zusammenhang werden wir auch auf das Lax-Verfahren noch einmal zu sprechen kommen.

Die Erweiterung des Verfahrens auf **mehrere Raumdimensionen** erfolgt ähnlich wie beim parabolischen Fall. Wir betrachten hierzu die skalare zweidimensionale Transportgleichung

$$u_t + a\, u_x + b\, u_y = 0 \quad \text{mit} \quad a, b \in \mathbb{R} \ . \tag{6.68}$$

Eine Möglichkeit, die Verfahren auf mehrere Raumdimensionen zu übertragen, ist wieder der Splitting-Ansatz. Das zweidimensionale Problem wird in jedem Zeitschritt in eine Folge von eindimensionalen Problemen

$$u_t + au_x = 0\,, \qquad u_t + bu_y = 0$$

zerlegt, welche dann nacheinander gelöst werden. Speziell die Upwind-Verfahren lassen sich damit sehr einfach übertragen. Man kann hier das Problem in x- und y-Richtung getrennt voneinander betrachten und die entsprechende eindimensionale Charakteristikentheorie anwenden. Die Stabilitätsbedingung muss dann in jede Richtung erfüllt sein. Man erhält bei einem solchen Splitting-Ansatz die Schrittweitenbedingungen

$$\frac{\Delta t}{\Delta x}|a| < 1\,, \qquad \frac{\Delta t}{\Delta y}|b| < 1\,. \tag{6.69}$$

Verfahren ohne Splitting durch Einsetzen der Approximationen für die x- und y-Ableitungen mittels Differenzenquotienten lassen sich ebenso ausführen. Man erhält dann als Stabilitätsbedingung (6.69) mit $1/2$ auf der rechten Seite. Damit wird während eines Zeitschritts eine Wechselwirkung der Wellen in x- und y-Richtung verhindert.

Die Definition von Upwind-Verfahren für mehrdimensionale Systeme macht allerdings Schwierigkeiten. In zwei Raumdimensionen kann die Gleichung nicht in die Diagonalform umgeschrieben werden, falls die Matrizen $\mathbf{A}$ und $\mathbf{B}$ nicht die gleichen Eigenvektoren besitzen. Wir werden dies im Rahmen der Finite-Volumen-Verfahren wieder aufgreifen. Auch die numerische Lösung linearer Probleme mit starken Gradienten und nichtlineare Probleme werden in dem Kapitel über Finite-Volumen-Verfahren wieder aufgegriffen.

6.4 Verfahren auf randangepassten Gittern

Im Abschnitt 5.2 wurde die Grundidee für die Erzeugung von randangepassten Gittern beschrieben. Das physikalische Gebiet Ω in der (x, y)-Ebene wird auf ein „logisches“ Rechteck Ω' in der (ξ, η)-Ebene abgebildet. Im Rechteck Ω' wird ein äquidistantes Gitter eingeführt, welches dann auf das physikalische Gebiet zurück transformiert wird. In Abbildung 5.3 ist diese Vorgehensweise dargestellt.

Bezeichnet man mit $\xi(x, y)$ und $\eta(x, y)$ die Abbildungen des physikalischen Gebietes Ω in der (x, y)-Ebene auf das logische Rechteck Ω' in der (ξ, η)-Ebene, so benötigt man die Umkehrfunktionen $x(\xi, \eta)$ und $y(\xi, \eta)$ vom logischen Rechteck auf das physikalische Gebiet.

Diese Umkehrfunktionen erfüllen die Gleichungen

$$\alpha x_{\xi\xi} - 2\beta x_{\xi\eta} + \gamma x_{\eta\eta} + J^2 x_\xi P + J^2 x_\eta Q = 0 \,, \tag{6.70}$$

(6.71)

$$\alpha y_{\xi\xi} - 2\beta y_{\xi\eta} + \gamma y_{\eta\eta} + J^2 y_\xi P + J^2 y_\eta Q = 0 \tag{6.72}$$

mit den Größen

$$J = x_\xi y_\eta - y_\xi x_\eta \,, \tag{6.73}$$

$$\alpha = x_\eta^2 + y_\eta^2 \,, \tag{6.74}$$

$$\beta = x_\xi x_\eta + y_\xi y_\eta \,, \tag{6.75}$$

$$\gamma = x_\xi^2 + y_\xi^2 \,. \tag{6.76}$$

Diese Gleichungen wurden schon im Abschnitt 5.2 formuliert. Mit den frei wählbaren Parametern P und Q kann man das Gitter an bestimmten Punkten zusammenziehen bzw. Gitterlinien entzerren.

Um ein Gitter zu erzeugen, müssen die beiden Gittergleichungen (6.70) und (6.72) zunächst numerisch gelöst. Dies geschieht auf einem äquidistanten achsenparallelen Gitter. Die partiellen Ableitungen werden entsprechend durch Differenzenquotienten ersetzt:

$$x_\xi = \frac{x_{i+1,j} - x_{i-1,j}}{2\Delta\xi}, \qquad y_\xi = \frac{y_{i+1,j} - y_{i-1,j}}{2\Delta\xi} \,,$$

$$x_\eta = \frac{x_{i,j+1} - x_{i,j-1}}{2\Delta\eta}, \qquad y_\eta = \frac{y_{i,j+1} - y_{i,j-1}}{2\Delta\eta} \,,$$

$$x_{\xi\xi} = \frac{x_{i+1,j} - 2x_{i,j} + x_{i-1,j}}{\Delta\xi^2} \,,$$

$$x_{\eta\eta} = \frac{x_{i,j+1} - 2x_{i,j} + x_{i,j-1}}{\Delta\eta^2} \,,$$

$$x_{\xi\eta} = \frac{x_{i+1,j+1} - x_{i-1,j+1} - x_{i+1,j-1} + x_{i-1,j+1}}{4\Delta\xi\Delta\eta} \,.$$

Durch Einsetzen der finiten Differenzen in Gleichung 6.70 und 6.72 erhält man ein lineares Gleichungssystem für die Koordinaten der Gitterpunkte $(x_{ij})_{i=1,\ldots,n;j=1,\ldots,n}$ und $(y_{ij})_{i=1,\ldots,n;j=1,\ldots,n}$, welches z.B. mit dem SOR-Verfahren gelöst werden kann. Man beachte, dass die Randpunkte vorgegeben sind, d.h. die Randbedingungen sind reine Dirichlet-Randbedingungen.

Wesentlich für die Strukturierung des Gitters ist, dass man vier ausgezeichnete Punkte auf dem Rand von Ω auswählt, welche den vier Eckpunkten des

logischen Gebietes Ω' entsprechen. Durch diese charakteristischen Punkte definiert man die untere, rechte, obere und linke Begrenzungslinie. Anschließend wird die Anzahl der Unterteilungen auf dem rechten (bzw. linken) Rand festgelegt. Damit ist dann auch automatisch die Anzahl der Unterteilungen auf der gegenüber liegenden Seite bestimmt. Denn jede Gitterlinie, die am linken Rand beginnt muss am rechten Rand enden. Anschließend müssen die Start- und Endpunkte der vertikalen Linien festgelegt werden.

Beispiel 6.8. Wir erläutern hier kurz, wie die Lösung der Poisson-Gleichung auf randangepassten Gitter abläuft.

Grundgleichungen: Da wir das physikalische Gebiet in der (x, y)-Ebene schon auf ein Rechteck in der (ξ, η)-Ebene transformiert haben, ist es nun bequemer die DGL im transformierten Gebiet zu lösen. Wir betrachten das Potenzial Φ zwar als eine Funktion von x und y; da aber x und y wiederum von ξ und η abhängen, gilt

$$\Phi = \Phi(x(\xi, \eta), y(\xi, \eta)) \; .$$

Nach der Kettenregel erhalten wir

$$\Phi_\xi = \frac{\partial \Phi}{\partial x}\frac{\partial x}{\partial \xi} + \frac{\partial \Phi}{\partial y}\frac{\partial y}{\partial \xi} = \Phi_x x_\xi + \Phi_y y_\xi \; ,$$

$$\Phi_\eta = \frac{\partial \Phi}{\partial x}\frac{\partial x}{\partial \eta} + \frac{\partial \Phi}{\partial y}\frac{\partial y}{\partial \eta} = \Phi_x x_\eta + \Phi_y y_\eta \; .$$

Dies ist ein lineares Gleichungssystem für Φ_x und Φ_y in der Form

$$\begin{pmatrix} x_\xi & y_\xi \\ x_\eta & y_\eta \end{pmatrix} \begin{pmatrix} \Phi_x \\ \Phi_y \end{pmatrix} = \begin{pmatrix} \Phi_\xi \\ \Phi_\eta \end{pmatrix} \; .$$

Mit der Cramerschen Regel lässt sich dieses LGS nach Φ_x und Φ_y auflösen:

$$\Phi_x = \frac{1}{J}\left(y_\eta \Phi_\xi - y_\xi \Phi_\eta\right) \; ,$$

$$\Phi_y = \frac{1}{J}\left(-x_\eta \Phi_\xi + x_\xi \Phi_\eta\right)$$

mit

$$J = x_\xi y_\eta - y_\xi x_\eta \; .$$

Lösen der Poisson-Gleichung auf randangepassten Gittern: Mit der gleichen Vorgehensweise wie für die erste Ableitung drückt man auch die zweiten partiellen Ableitungen Φ_{xx} und Φ_{yy} durch Φ_ξ, Φ_η, $\Phi_{\xi\xi}$, $\Phi_{\eta\eta}$ und $\Phi_{\xi\eta}$ aus. Man ersetzt alle Ableitungen nach x und y durch die entsprechenden

Ableitungen nach ξ und η. Der Vorteil dieser Vorgehensweise ist, dass man die Ableitungen Φ_ξ, Φ_η, $\Phi_{\xi\xi}$, $\Phi_{\eta\eta}$ und $\Phi_{\xi\eta}$ im Rechteckgebiet einfach diskretisieren kann. Man erhält so die transformierte, diskrete Poisson-Gleichung in der (ξ, η)-Ebene.

In der Praxis treten oftmals rotationssymmetrische Probleme auf. Dann verwendet man zur Beschreibung des Problems Zylinderkoordinaten (z, r). In den Zylinderkoordinaten lautet die Potenzialgleichung

$$\Phi_{zz}(z,r) + \Phi_{rr}(z,r) + \underline{\frac{1}{r}\Phi_r(z,r)} = -\frac{\rho(z,r)}{\epsilon}.$$

Wir werden diese Form der Potenzialgleichung transformiert in der (ξ, η)-Ebene angeben. Wenn die Potenzialgleichung nur in kartesischen Koordinaten gelöst werden soll, dann entfällt der unterstrichene Term.

Poisson-Gleichung in der (ξ, η)-Ebene: Ersetzt man in der Potenzialgleichung die Ableitungen nach r und z durch die entsprechenden Ableitungen nach ξ und η, so erhält man im logischen Gebiet die folgende transformierte Poisson-Gleichung

$$\alpha\Phi_{\xi\xi} - 2\beta\Phi_{\xi\eta} + \gamma\Phi_{\eta\eta} + (\tau - \underline{z_\eta J/r})\Phi_\xi + (\sigma - \underline{z_\xi J/r})\Phi_\eta = -J^2\rho/\epsilon_0$$

mit

$$\sigma := (r_\xi Dz - z_\xi Dr)/J,\ \tau := (z_\eta Dr - r_\eta Dz)/J$$

und

$$Dz := \alpha z_{\xi\xi} - 2\beta z_{\xi\eta} + \gamma z_{\eta\eta},\ Dr := \alpha r_{\xi\xi} - 2\beta r_{\xi\eta} + \gamma r_{\eta\eta}\ .$$

Dabei sind die Koeffizienten α, β und γ wie in Gl. 6.73 - 6.76 definiert. Die Koeffizienten hängen nur von der Lage der Gitterpunkte ab. Ersetzt man die kontinuierlichen Ableitungen der Differenzialgleichungen am Gitterpunkt (i, j) durch zentrale Differenzen

$$\begin{aligned}
\Phi_\xi(i,j) &= \frac{\Phi_{i+1,j} - \Phi_{i-1,j}}{2\Delta\xi}\ , \\
\Phi_\eta(i,j) &= \frac{\Phi_{i,j+1} - \Phi_{i,j-1}}{2\Delta\eta}\ , \\
\Phi_{\xi\xi}(i,j) &= \frac{\Phi_{i+1,j} - 2\Phi_{i,j} + \Phi_{i-1,j}}{\Delta\xi^2}\ , \\
\Phi_{\eta\eta}(i,j) &= \frac{\Phi_{i,j+1} - 2\Phi_{i,j} + \Phi_{i,j-1}}{\Delta\eta^2}\ , \\
\Phi_{\xi\eta}(i,j) &= \frac{\Phi_{i+1,j+1} - \Phi_{i-1,j+1} - \Phi_{i+1,j-1} + \Phi_{i-1,j-1}}{4\Delta\xi\Delta\eta}
\end{aligned}$$

erhält man ein lineares Gleichungssystem für die gesuchten Größen $\Phi_{i,j}$ (= Potenzial am Gitterpunkt (i,j)):

$$\begin{aligned} M_{i,j}\Phi_{i-1,j-1} &+ B1_{i,j}\Phi_{i-1,j} - M_{i,j}\Phi_{i-1,j+1} \\ &+ B2_{i,j}\Phi_{i,j-1} + C_{i,j}\Phi_{i,j} + A2_{i,j}\Phi_{i,j+1} \\ &- M_{i,j}\Phi_{i+1,j-1} + A1_{i,j}\Phi_{i+1,j} + M_{i,j}\Phi_{i-1,j+1} = R_{i,j}. \end{aligned}$$

Die Koeffizienten $A1$, $A2$, $B1$, $B2$, C und M als auch die rechte Seite R folgen durch Einsetzen der Differenzenquotienten in die Differenzialgleichung.

Lösen des linearen Gleichungssystems durch iterative Methoden: Das oben beschriebene lineare Gleichungssystem für die Potenzialwerte an den Gitterpunkten wird durch ein iteratives Verfahren z.B. mit dem SOR-Verfahren gelöst:

$$\begin{aligned} \Phi_{i,j}^{(m+1)} = (1-\omega_{i,j})\Phi_{i,j}^{(m)} - \frac{\omega_{i,j}}{C_{i,j}}(&A1_{i,j}\Phi_{i+1,j}^{(m)} + A2_{i,j}\Phi_{i,j+1}^{(m)} + B1_{i,j}\Phi_{i-1,j}^{(m+1)} \\ &+B2_{i,j}\Phi_{i,j-1}^{(m+1)} + M_{i,j}(\Phi_{i+1,j+1}^{(m)} - \Phi_{i-1,j+1}^{(m)} - \Phi_{i+1,j-1}^{(m+1)} + \Phi_{i-1,j-1}^{(m+1)}) \\ &-R_{i,j})\ . \end{aligned}$$

Um ein Abbruchkriterium für das Verfahren zu erhalten, wird nach jeder Iteration das Residuum berechnet:

$$\begin{aligned} RES_{i,j}^{(m+1)} = &A1_{i,j}\Phi_{i+1,j}^{(m+1)} + A2_{i,j}\Phi_{i,j+1}^{(m+1)} + B1_{i,j}\Phi_{i-1,j}^{(m+1)} + B2_{i,j}\Phi_{i,j-1}^{(m+1)} \\ &+C_{i,j}\Phi_{i,j}^{(m+1)} + M_{i,j}(\Phi_{i+1,j+1}^{(m+1)} - \Phi_{i-1,j+1}^{(m+1)} - \Phi_{i+1,j-1}^{(m+1)} + \Phi_{i-1,j-1}^{(m+1)}) \\ &-R_{i,j}\ . \end{aligned}$$

Ist das Maximum über die Residuen aller Gitterpunkte kleiner als eine vorgegebene Schranke, so ist die Iteration beendet und die Potenzialwerte $\Phi_{i,j}$ an den Gitterpunkten berechnet.

6.5 Bemerkungen und Entscheidungshilfen

Der **Vorteil** der Differenzenverfahren ist, dass sie klar und durchsichtig ist und sich einfach programmieren lässt. Sie kann im Prinzip auf jede Differenzialgleichung angewendet werden. Die **Einschränkung** der Methode liegt in der Forderung, dass die Berechnung ein strukturiertes Gitter erfordert. Dies kann für komplexe Geometrien schwierig und aufwändig werden. Lokale Gitterverfeinerungen in den Bereichen, wo sich die Lösung sehr stark ändert, sind nicht ausführbar.

Die Vorgehensweise bei der Lösung von elliptischen Differenzialgleichungen mit Differenzenverfahren kann man in fünf Schritte untergliedern:

1. Einführung eines Berechnungsgitters,

2. Einschränkung der gesuchten Funktion auf Funktionswerte an den diskreten Gitterpunkten und Ersetzen der Ableitungen der DGL durch finite Differenzen,

3. Aufstellen des LGS für die diskreten Näherungswerte,

4. Lösen des LGS; die Lösung des LGS approximiert die gesuchte Funktion an den Gitterpunkten,

5. Visualisierung der Ergebnisse.

Bei hyperbolischen oder parabolischen Problemen löst man das Problem von einem Zeitschritt zum anderen. Dabei hat man die Wahl einer expliziten oder einer impliziten Approximation der Zeit. Bei einem implizitem Verfahren muss wie bei einer elliptischen Gleichung ein Gleichungssystem gelöst werden, hier allerdings in jedem Zeitschritt. Bei einem expliziten Verfahren wird die Lösung des Gleichungssystems ganz einfach: Man hat eine Diagonalmatrix vorliegen, da es in der Differenzenformel nur einen Wert auf der neuen Zeitschicht gibt. Diesen neuen Wert kann man somit explizit in einer Lösungsformel angeben. Man würde somit nur explizite Verfahren benutzen, falls es da keine Stabilitätsbedingung gäbe. Diese stellt sich als Restriktion an die Wahl der Zeitschrittweite und kann somit die Effizienz des Verfahrens stark vermindern. So muss bei einem hyperbolischen Problem die Zeitschrittweite proportional und bei einem parabolischen Problem sogar proportional zum Quadrat der Raumschrittweite gewählt werden.

Neben dem Einfluss der Stabilität bei expliziten Verfahren ist die Wahl der Schrittweiten aber auch ein Genauigkeitsproblem. Zunächst muss die Raumschrittweite so klein gewählt werden, dass alle wichtigen Strukturen in der Lösung aufgelöst werden können. Dann kann man sich auf Grund der physikalischen Phänomene, insbesondere der Ausbreitungsgeschwindigkeiten, den richtigen Zeitschritt überlegen. Setzt man dann die Zahlenwerte in die entsprechende Stabilitätsbedingung ein, kann man entscheiden, welches Verfahren das effizientere sein müsste. Man rechnet im Allgemeinen damit, dass die Lösung eines Gleichungssystems mindestens soviel wie 10 Zeitschritte eines expliziten Verfahrens an Rechenzeit kostet.

Zu der Entscheidung explizit oder implizit bei hyperbolischen und parabolischen Problemen lassen sich die folgende Grundregeln aufstellen. Bei einer instationären Wellenausbreitung ist die CFL-Bedingung für die Stabilität von

expliziten Verfahren meist auch eine Genauigkeitsforderung und sollte somit immer erfüllt sein. Nur bei langsamen Prozessen, bei denen die Zeitableitungen sehr klein sind, kann man sich große Zeitschritte oder auch ein Verfahren 1. Ordnung in der Zeit erlauben. Im allgemeinen setzt man Verfahren mit 2. Ordnung ein. Bei parabolischen Problemen ist die Stabilitätsbedingung allerdings so restriktiv, dass man meist ein implizites Verfahren vorzieht.

Das Differenzenverfahren ist das älteste der numerischen Lösungsverfahren sowohl bei den gewöhnlichen als auch bei den partiellen Differenzialgleichungen. Es gibt somit eine ganze Reihe von weiterführender Literatur. In allen Büchern über die numerische Lösung von partiellen Differenzialgleichungen nimmt diese Klasse einen breiten Raum ein, so z.B. in [66], [42], [48], [24] und [37]. Es gibt darüber hinaus Bücher nur über Differenzenverfahren, wie [41] und [63] und die klassischen Werke von [49] und [51].

6.6 Beispiele und Aufgaben

Beispiel 6.9 (Vergleich verschiedener Gleichungssystemlöser, mit MAPLE**-Worksheet).** In diesem Beispiel wird die Leistungsfähigkeit der in Kapitel vorgestellten Gleichungssystemlöser untersucht. Hierfür wird noch einmal Beispiel 6.1 verwendet, in welchem wir die Verbiegung einer Membran berechnet hatten. Im zugehörigen Worksheet wurde das Gauß-Seidel-Verfahren zur Lösung des linearen Gleichungssystems verwendet. In diesem Beispiel wurde das Worksheet um das Jacobi- und das SOR-Verfahren erweitert. Als Abbruchkriterium wird bei allen Rechnungen die Bedingung $u_{i,j}^{p+1} - u_{i,j}^{p} <= \epsilon \ \forall \ (i,j)$ verwendet. Zwar wäre es hier auch möglich, die exakte Lösung zu verwenden, jedoch steht diese normalerweise nicht zur Verfügung. Für $\epsilon = 10^{-5}$ ergeben sich z.B. folgende Iterationszahlen:

Verfahren	Iterationszahl
Jacobi	352
Gauß-Seidel	213
SOR	45

Hier ist klar ersichtlich, wie stark sich die Verfahren unterscheiden. Das SOR-Verfahren benötigt deutlich weniger Iterationen und damit deutlich weniger Rechenzeit. Dieser Trend setzt sich bei höheren Genauigkeitsforderungen so fort. □

Aufgabe 6.1. Bestimmen sie das CIR-Verfahren für die Wellengleichung

$$u_{tt} - c^2 u_{xx} = 0 \, .$$

Lösung: Zunächst führen wir eine einfache Substitution durch, um die partielle Differenzialgleichung zweiter Ordnung in ein System erster Ordnung zu überführen. Wir setzen hierzu

$$\begin{aligned} p &= u_t \, , \\ q &= c u_x \, . \end{aligned}$$

Hiermit ergibt sich dann die Wellengleichung als System 1. Ordnung

$$\begin{aligned} p_t - c q_x &= 0 \\ q_t - c p_x &= 0 \end{aligned}$$

formulieren. Dieses System von hyperbolischen Differenzialgleichungen kann man dann auch in der Vektor-Matrix-Form schreiben:

$$\begin{pmatrix} p \\ q \end{pmatrix}_t + \begin{pmatrix} 0 & -c \\ -c & 0 \end{pmatrix} \begin{pmatrix} p \\ q \end{pmatrix}_x = \mathbf{0} \, .$$

Die Eigenwerte der Matrix des Systems kann man leicht als $\lambda_1 = -c$ und $\lambda_2 = c$ bestimmen, womit sich die Matrix der Eigenvektoren zu

$$\mathbf{R} = \begin{pmatrix} 1 & 1 \\ -1 & 1 \end{pmatrix}$$

ergibt. Die Matrizen $\mathbf{A}^+$ und $\mathbf{A}^-$ ergeben sich dann nach der Herleitung des CIR-Verfahrens in (6.63) zu

$$\mathbf{A}^+ = \frac{c}{2} \begin{pmatrix} 1 & -1 \\ -1 & 1 \end{pmatrix} \quad \text{und} \quad \mathbf{A}^- = \frac{c}{2} \begin{pmatrix} -1 & -1 \\ -1 & -1 \end{pmatrix} \, .$$

Nach Gleichung (6.63) kann man dann das CIR-Verfahren für die Wellengleichung angeben:

$$\mathbf{u}_i^{n+1} = \mathbf{u}_i^n - \frac{\Delta t}{\Delta x} A^+ \left(\mathbf{u}_i^n - \mathbf{u}_{i-1}^n \right) - \frac{\Delta t}{\Delta x} A^- \left(\mathbf{u}_{i+1}^n - \mathbf{u}_i^n \right) \, ,$$

wobei $\mathbf{u}$ den Vektor mit den Komponenten p und q bezeichnet. □

Aufgabe 6.2. Gegeben sei das RWP für die Poisson-Gleichung im Innern des Einheitsquadrates I^2

$$\begin{aligned} -u_{xx} - u_{yy} &= -x^2 - y^2 + x + y \ , \\ u(x,y) &= 0 \qquad \text{auf dem Rand von } I^2 \ . \end{aligned}$$

Die exakte Lösung lautet in disem Fall:

$$u(x,y) = \frac{1}{2}\, x(x-1)y(y-1) \ .$$

1. Lösen sie das RWP mit dem Differenzenverfahren für die Werte

$$\begin{aligned} \Delta x = \Delta y &= \tfrac{1}{10} \ , \qquad & \Delta x = \Delta y &= \tfrac{1}{20} \ , \\ \Delta x = \Delta y &= \tfrac{1}{40} \ , \qquad & \Delta x = \Delta y &= \tfrac{1}{100} \ . \end{aligned}$$

2. Vergleichen sie die numerische Lösung mit der exakten Lösung des Problems. Welche Aussagen über die Genauigkeit kann man in Abhängigkeit der Schrittweite machen? □

Aufgabe 6.3. Gegeben sei das RWP für die Poisson-Gleichung im Innern des Einheitsquadrates I^2

$$\begin{aligned} -u_{xx} - u_{yy} &= 2\pi^2 \ sin(\pi x) \ sin(\pi y) \ , \\ u(x,y) &= 0 \qquad \text{auf dem Rand von } I^2 \ . \end{aligned}$$

Die exakte Lösung ist gegeben durch

$$u(x,y) = sin(\pi x) \ sin(\pi y) \ .$$

1. Lösen sie dieses Problem mit dem Differenzenverfahren für die Werte

$$\begin{aligned} \Delta x = \Delta y &= \tfrac{1}{10} \ , \qquad & \Delta x = \Delta y &= \tfrac{1}{20} \ , \\ \Delta x = \Delta y &= \tfrac{1}{40} \ , \qquad & \Delta x = \Delta y &= \tfrac{1}{100} \ . \end{aligned}$$

2. Vergleichen sie die numerische Lösung mit der exakten Lösung des Problems. Welche Aussagen über die Genauigkeit kann man in Abhängigkeit der Schrittweite machen? □

Aufgabe 6.4. 1. Führen sie für das 2-Elektroden-System ohne Zwischenelektrode ein (6x5)-Gitter ein und erstellen sie ein Differenzenverfahren für die Werte $\Phi_A = 10\,V; \Phi_K = 2\,V$.

2. Lösen sie das LGS mit dem Jacobi-Verfahren, dem Gauß-Seidel-Verfahren und dem SOR-Verfahren ($w = 1.3$).

3. Lösen sie das LGS für das 3-Elektroden-System mit Zwischenelektrode ($\Phi_A = 10\,V, \Phi_K = 2\,V, \Phi_Z = 0\,V$) mit dem Jacobi-Verfahren, dem Gauß-Seidel-Verfahren und dem SOR-Verfahren.

Lösung: MAPLE**-Worksheet** □

Aufgabe 6.5. Berechnen sie die Temperaturverteilung für einen bei $x = 0$ und $x = 1$ eingespannten Metallstab ($\kappa = 1.14$), der an den Stirnseiten auf Temperatur $u_l = 0$ und $u_r = 0$ gehalten wird. Die Anfangstemperatur sei

$$u(x,0) = 2\sin(3\pi x) + 5\sin(8\pi x) \quad \text{für} \quad x \in [0,1]\,.$$

Führen sie dies mit dem expliziten, dem impliziten und dem Crank-Nicolson-Verfahren aus. Die exakte Lösung wurde in Beispiel 4.4 abgeleitet.

Lösung: Die Lösung ist durch ein MAPLE**-Worksheet** gegeben. Wird für das explizite Verfahren die Schrittweitenbedingung verletzt, so ergeben sich wie erwartet Oszillationen, deren Amplituden anwachsen. Wie die Stabilitätstheorie aussagt, zeigt das voll-implizite Verfahren auch für große Schrittweiten keine Instabilitäten. Wird bei dem Crank-Nicolson-Verfahren die Zeitschrittweite sehr groß gewählt, dann ergeben sich am Anfang leichte Oszillationen, deren Amplitude allerdings mit der Zeit schnell kleiner werden. Neben der Stabilitätsbedingung ist die Genauigkeit ein Kriterium für die Schrittweite. Ist man an der instationären Lösung interessiert, dann darf die Zeitschrittweite nicht zu groß gewählt werden. □

Aufgabe 6.6. Man zeige mit Hilfe der von von Neumannschen Stabilitätsanalyse, dass das voll-implizite Verfahren für die lineare Transportgleichung

$$\frac{u_i^{n+1} - u_i^n}{\Delta t} + a\frac{u_{i+1}^{n+1} - u_{i-1}^{n+1}}{2\Delta x} = 0$$

ohne Bedingung stabil ist. □

Aufgabe 6.7. Man berechne das Anfangswertproblem für die lineare Transportgleichung

$$u_t + \frac{1}{2}\sqrt{t}\,u_x = 0, \quad u(x,0) = e^{-(x-4)^2} \quad \text{für} \quad x \in [0,10]$$

im Zeitintervall $[0,10]$ mit dem CIR-Verfahren, dem voll-impliziten Verfahren, dem Crank-Nicolson-Verfahren und dem Lax-Wendroff-Verfahren. Man vergleiche mit der exakten Lösung nach Beispiel 4.7.

Lösung: MAPLE**-Worksheet** □

7 Finite-Elemente-Methode

Wir haben die Methode der finiten Elemente schon in Kapitel 3.8 zum Lösen gewöhnlicher Randwertprobleme eingeführt und eingesetzt. Das eigentliche Anwendungsgebiet sind jedoch die partiellen Differenzialgleichungen und zwar vorzugsweise die elliptischen Randwertprobleme. Wie im eindimensionalen Fall zeigen wir die Finite-Elemente-Methode (FE-Methode) in Kombination mit dem Ritzschen Verfahren. Wie dort beschrieben sind die FE-Verfahren auch mit dem Galerkin-, Least Square oder Kollokations-Ansatz kombinierbar, wobei die Näherungslösung mit lokalen Basisfunktionen, z.B. den Hufunktionen, aufgebaut wird.

Die Grundidee bei der Finite-Elemente-Methode besteht darin, dass man als Approximation für die Lösung eine kontinuierliche Funktion vorgibt. Man führt somit eigentlich eine Diskretisierung des Funktionenraumes ein. Die Definition der diskreten Anzahl von Basisfunktionen gründet sich dann aber doch auf eine Raumdiskretisierung und auf ein Gitter. Die Einführung von Basisfunktionen, welche nur auf einigen wenigen Gitterzellen Werte ungleich Null haben, führt auf schwach besetzte Matrizen, welche dann mit bei den Differenzenverfahren beschriebenen iterativen Lösungsverfahren schnell lösbar sind. In Verallgemeinerung des eindimensionalen Falls wählt man als Basisfunktionen im einfachsten zweidimensionalen Fall nun Pyramidenfunktionen, wenn man das Gebiet in Dreiecke zerlegt. Damit nimmt man innerhalb eines Elementes einen linearen Verlauf der Lösungsfunktion an. Die Superposition der Pyramidenfunktionen, welche das Energiefunktional minimiert, ist dann eine Näherung für die Lösung des RWP nach dem Ritzschen Verfahren.

Wir werden hier die FE-Methode für die numerische Lösung von elliptischen Differenzialgleichungen behandeln und für die Anwendung auf parabolische und hyperbolische Differenzialgleichungen auf die Literatur verweisen.

Wir starten mit der Variationsaufgabe

$$I(u) = \int_{\Omega} f(x, y, z, u, u_x, u_y, u_z)\ dx\ dy\ dz \stackrel{!}{=} \text{minimal!}$$

Zu dieser Variationsaufgabe lässt sich über die Eulersche Differenzialgleichung

$$\frac{\partial f}{\partial u} - \frac{\partial}{\partial x} f_{u_x} - \frac{\partial}{\partial y} f_{u_y} - \frac{\partial}{\partial z} f_{u_z} = 0$$

die zugehörige partielle Differenzialgleichung angeben. Für das inverse Problem der Variationsrechnung, also dem Übergang vom Randwertproblem zu einer Variationsaufgabe, existiert nicht immer eine Lösung und falls sie existiert, ist es im Allgemeinen alles andere als trivial, sie herzuleiten. Nur für solche Probleme, bei denen die Variationsaufgabe bekannt ist, lässt sich das Ritzsches Verfahren anwenden.

Im Folgenden gehen wir von dem Randwertproblem (RWP) der verallgemeinerten Poisson- oder Helmholtz-Gleichung aus:

$$-\Delta u(x,y) + cu(x,y) = f(x,y) \quad \text{im Innern des Gebiets } \Omega \ (c \geq 0) \tag{7.1}$$

$$u(x,y) = 0 \quad \text{auf dem Rand von } \Omega\,. \tag{7.2}$$

Wie man mit der Eulerschen Differenzialgleichung nachprüfen kann, ist die zum RWP gehörende Variationsaufgabe

$$F(u) = \int_{\Omega} (\nabla u \cdot \nabla u) + c\ u\ u)\ dxdy - 2 \int_{\Omega} u\ f\ dxdy\,. \tag{7.3}$$

Definiert man das Skalarprodukt

$$< f, g >:= \int_{\Omega} f(x,y) g(x,y) dxdy \tag{7.4}$$

und die positiv definite, symmetrische Bilinearform

$$[u,v] = \int_{\Omega} (\nabla u \cdot \nabla v) + c\ uv)\ dxdy\,, \tag{7.5}$$

so lautet das Energiefunktional

$$F(u) = [u,u] - 2 < u, f > \ . \tag{7.6}$$

F nimmt den kleinsten Wert genau für die Lösung des RWP an! Also gilt: $\overline{u}$ Lösung des RWPs $\Leftrightarrow$ $F(u) \geq F(\overline{u})$ für alle $u \in \mathbb{D}_L = \{u : [a,b] \to \mathbb{R} :$ zweimal stetig differenzierbar mit $u(a) = u(b) = 0\}$.

Minimierung des Energiefunktionals. Da man in der Regel das Minimum $\overline{u}$ von $F(u)$ nicht unter allen zweimal stetig differenzierbaren Funktionen mit $u(x,y) = 0$ auf dem Rand von Ω bestimmen kann, wählt man sich eine kleinere Klasse von Funktionen und sucht das Minimum in dieser kleineren Klasse:

1. Bei der Finiten-Elemente-Methode wählt man sich einen endlich dimensionalen Unterraum $S \subset \mathbb{D}_L$ mit $dim\ S = m$. Sei $u_1, \ldots, u_m$ eine Basis von S, so gibt es zu jedem $u \in S$ Koeffizienten $\delta_1, \ldots, \delta_m$ mit

$$u(x,y) = \sum_{i=1}^{m} \delta_i u_i(x,y)\ . \tag{7.7}$$

   Dann lautet das Problem: Bestimme in S ein Minimum $\overline{u}$ von F mit $F(u) \geq F(\overline{u})$ für alle $u \in S$. Die spezielle Wahl des Unterraums setzt eine konkrete Unterteilung des Gebietes Ω in Teilgebiete Ω_i mit $\Omega = \bigcup \Omega_i$ voraus, sowie die Spezifikation der Basisfunktionen auf den Teilgebieten.

2. Setzt man die Funktion $u(x,y)$ in das Energiefunktional ein, so erhält man eine Funktion

$$\Phi(\delta_1, \ldots, \delta_m) := F(\delta_1 u_1 + \cdots + \delta_m u_m),$$

   die nur noch von den unbekannten Koeffizienten $\delta_1, \ldots, \delta_m$, auch Freiheitsgrade genannt, abhängt. Für das Minimum der Funktion $\Phi(\delta_1, \delta_2, ..., \delta_m)$ müssen alle partiellen Ableitungen nach den Variablen verschwinden:

$$\frac{\partial \Phi}{\partial \delta_i} = 0 \qquad \text{für alle } i = 1, ... , m\ . \tag{7.8}$$

3. Für lineare partielle Differenzialgleichungen erhält man somit ein lineares Gleichungssystem

$$\mathbf{A}\boldsymbol{\delta} = \mathbf{r} \tag{7.9}$$

   für die unbekannten Koeffizienten $\delta_1, \delta_2, ..., \delta_m$. Ist $\overline{\delta} = (\overline{\delta}_1, \ldots, \overline{\delta}_m)$ die Lösung des LGS, dann ist

$$\overline{u}(x,y) = \overline{\delta}_1 u_1(x,y) + \cdots + \overline{\delta}_m u_m(x,y)$$

   das gesuchte Minimum von F auf S.

Die oben geführte Diskussion spezifiziert weder die Zerlegung des Gebietes Ω noch die spezielle Wahl der Basisfunktionen u_i. Im Folgenden diskutieren wir für die Poisson-Gleichung zum Einen eine Zerlegung des Gebietes in gleichmäßige Dreiecke (Triangulierung) mit linearen Basisfunktionen. Dies geschieht einmal über die Interpretation der Basisfunktionen als Pyramidenfunktionen mit einer Überlappung der Basisfunktionen wie im eindimensionalen Fall (7.1). Wir interpretieren anschließend den Ansatz um, indem wir

die Dreiecke separat betrachten und die Funktion abschnittsweise nur auf den einzelnen Dreiecken (Elementen) definieren (7.2). In diesem Fall hat man keine Überlappung der Elementfunktionen, muss dafür aber alle an einen Punkt angrenzenden Dreiecke beim Aufstellen des Gleichungssystems berücksichtigen. Dieses Aufsummieren der Beiträge nennt man **Assemblierung**.

Zum Anderen diskutieren wir eine Zerlegung des Gebietes in gleichmäßige Rechtecke mit bilinearen Elementfunktionen (7.3). Abschließend behandeln wir die verallgemeinerte Poisson-Gleichung mit einer Zerlegung des Gebietes in beliebige Dreiecke und der Wahl von quadratischen Elementfunktionen (7.4).

7.1 Triangulierung mit linearen Basisfunktionen

In diesem Abschnitt diskutieren wir ein Randwertproblem für die Poisson-Gleichung mit homogenen Dirichlet-Randbedingungen

$$-\Delta u(x,y) = f(x,y) \quad \text{im Innern des Einheitsquadrates } I^2, \tag{7.10}$$
$$u(x,y) = 0 \quad \text{auf dem Rand von } I^2 \tag{7.11}$$

mit der Inhomogenität

$$f(x,y) = 8\pi^2 \; sin(2\pi x) \; sin(2\pi y) \; .$$

Zum anschließenden Vergleich der Finite-Elemente-Lösung mit dem exakten Ergebnis geben wir die analytische Lösung des RWP an. Wie man durch Einsetzen direkt nachprüfen kann, lautet sie

$$u(x,y) = sin(2\pi x) \; sin(2\pi y).$$

Das zum RWP gehörende Energiefunktional vereinfacht sich für die Poisson-Gleichung zu

$$F(u) = [u,u] - 2 < u, f >$$

mit

$$[u,v] = \int_{\Omega} (\nabla u \cdot \nabla v) \; dxdy \; .$$

Diskretisierung des Rechengebietes. Für die Zerlegung des Gebietes wählen wir eine gleichmäßige Unterteilung in Dreiecke, wie es in Abbildung

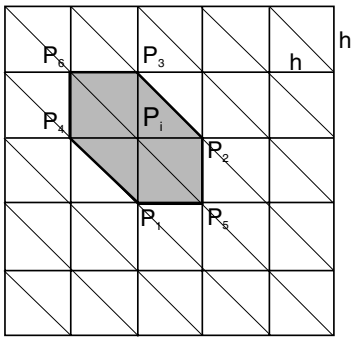


Abb. 7.1. Gleichförmiges Gitter mit Dreiecke ($h = 1/N$)

7.1 zu sehen ist. Dabei ist $h = \Delta x = \Delta y = 1/N$, wenn N die Anzahl der Intervalle in x- bzw. y-Richtung bezeichnet.

Basisfunktionen. Durch die Triangulierung wird das Gebiet I^2 in einzelne Gitterzellen oder Elemente zerlegt. Nun müssen zu dieser Unterteilung die Elementfunktionen angegeben werden. Zum Punkt P_i spezifizieren wir das in Abbildung 7.1 grau unterlegt Sechseck mit P_i als Zentrum und den Eckpunkten $P_k,\ k = 1, \ldots, 6$. Die zum Punkte P_i gehörende Basisfunktion u_i legen wir durch die Eigenschaft fest, dass

$$u_i(P_j) = \begin{cases} 0 & \text{für } i \neq j \\ 1 & \text{für } i = j \end{cases},$$

wobei der Verlauf zwischen den Punkten als linear angenommen wird. Durch diese Definition hat die Basisfunktion $u_i(x, y)$ die Form einer Pyramide mit sechseckiger Grundfläche, wie sie in Abbildung 7.2 dargestellt ist.

Im Folgenden bezeichnen wir mit dem Träger T_i der Basisfunktion den Abschluss des Bereichs, in dem die Funktion $u_i = u_i(x, y)$ von Null verschiedene Werte annimmt. Die entspricht genau dem grau unterlegten Sechseck in Abb. 7.2. Durch die Linearkombination aller Basisfunktionen $u_i = u_i(x, y)$, $i = 1, ..., m$, erhalten wir den Ansatz für die Lösung des Variationsproblems

$$u(x, y) = \sum_{i=1}^{m} \delta_i u_i(x, y)$$

mit den unbekannten Koeffizienten δ_i, $i = 1, ..., m$. Die Koeffizienten δ_i sind genau die Funktionswerte der gesuchten Lösung der Poisson-Gleichung in den Punkten P_i. Nach dem Variationsprinzip müssen sie so bestimmt werden, dass $u = u(x, y)$ das Minimum des Energiefunktionals in S ergibt.

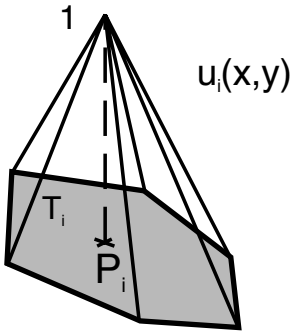


Abb. 7.2. Pyramide als Basisfunktion mit sechseckiger Grundfläche

Einsetzen in das Energiefunktional. Wir setzen daher $u(x,y) = \sum_{i=1}^{m} \delta_i u_i(x,y)$ in $F(u)$ ein und erhalten die Funktion Φ durch

$$\begin{aligned}\Phi(\delta_1,\ldots,\delta_m) &:= F(\delta_1 u_1 + \cdots + \delta_m u_m) = \\ &= [\sum_{i=1}^{m} \delta_i u_i, \sum_{j=1}^{m} \delta_j u_j] - 2 < \sum_{k=1}^{m} \delta_k u_k, f > \\ &= \sum_{i=1}^{m}\sum_{j=1}^{m} \delta_i \delta_j [u_i, u_j] - 2 \sum_{k=1}^{m} \delta_k < u_k, f > .\end{aligned}$$

Bei der Rechnung setzen wir die Bilinearität von $[\,,\,]$ und $<,>$ ein. Die Funktion Φ hängt nur noch von den unbekannten Koeffizienten $\delta_1,\ldots,\delta_m$ ab. Für das Minimum der Funktion $\Phi(\delta_1,\delta_2,...,\delta_m)$ müssen alle partiellen Ableitungen nach den Variablen verschwinden, d.h.

$$\frac{\partial(\Phi)}{\partial\delta_i} = 0 \qquad \text{für alle } i = 1,...,\ m \quad \text{bzw.} \qquad \nabla\Phi = 0\ .$$

Führt man wie im eindimensionalen Fall die Vektoren $\boldsymbol{\delta}$ und $\mathbf{r}$ sowie die Matrix $\mathbf{A}$

$$\boldsymbol{\delta} := \begin{bmatrix}\delta_1\\ \cdots\\ \delta_m\end{bmatrix}, \quad \mathbf{r} := \begin{bmatrix}< u_1, f >\\ \cdots\\ < u_m, f >\end{bmatrix}; \quad \mathbf{A} := ([u_i, u_j])_{i,j}$$

ein, so gilt

$$\nabla\Phi = 0 \quad \Leftrightarrow \quad \mathbf{A}\boldsymbol{\delta} = \mathbf{r}\ .$$

Ist also $\overline{\boldsymbol{\delta}} = (\overline{\delta}_1,\ldots,\overline{\delta}_m)^T$, wobei der hochgestellte Index T wieder den Transponierten des Zeilenvektors bezeichnet, die Lösung des linearen Gleichungssystems $\mathbf{A}\boldsymbol{\delta} = \mathbf{r}$,

dann ist

$$\overline{\mathbf{u}}(x,y) = \overline{\delta}_1 u_1(x,y) + \cdots + \overline{\delta}_m u_m(x,y)$$

das gesuchte Minimum von F in S.

Aufstellen des linearen Gleichungssystems. Um die Matrix $\mathbf{A}$ und die rechte Seite $\mathbf{r}$ des LGS

$$\mathbf{A}\boldsymbol{\delta} = \mathbf{r}$$

bzw. die i-te Gleichung

$$\delta_1[u_i,u_1] + \delta_2[u_i,u_2] + \ldots + \delta_m[u_i,u_m] =< u_i, f >$$

aufzustellen, müssen die Koeffizienten $[u_i,u_j]$ und $< u_i, f >$ für die Pyramidenfunktionen berechnet werden. Nur wenn sich die Träger der Pyramidenfunktionen u_i und u_j überlappen, sind die Koeffizienten ungleich Null: Dies ist dann der Fall, wenn die Punkte P_i und P_j benachbart sind.

Bei der Berechnung gehen wir von einem allgemeinen Punkt P_i mit den benachbarten Punkten $P_1,\ldots,P_6$ aus (siehe Abbildung 7.3). Der Träger von u_i, $T_i = \bigcup_{k=1}^{6} \Delta_k$, setzt sich aus den sechs an den Punkt P_i angrenzenden Dreiecken Δ_k zusammen, die in Abbildung 7.3 grau unterlegt sind. Um

$$[u_i,u_j] = \int_{\Omega} (\nabla u_i \cdot \nabla u_j)\; dxdy$$

zu berechnen, benötigen wir nicht die Funktionsvorschrift von u_i bzw. u_j, sondern lediglich die Gradienten der Funktionen auf den Flächen Δ_k. Da die Basisfunktionen auf diesen Dreiecken lineare Funktionen darstellen, ist der Gradient dort konstant. In Tabelle 7.1 sind die Gradienten für alle an den Punkt P_i angrenzenden Elemente Δ_k angegeben.

Da die Basisfunktionen außerhalb ihres Trägers Null sind, reduziert sich die Auswertung des Integrals über den Bereich Ω auf $T_i \cap T_j$. Im Einzelnen erhalten wir mit den Gradienten aus Tabelle 7.1:

$$\begin{aligned}
[u_i,u_i] &= \int_{\Omega} (\nabla u_i \cdot \nabla u_i) dxdy = \int_{T_i} (\nabla u_i \cdot \nabla u_i) dxdy \\
&= \sum_{k=1}^{6} \int_{\Delta_k} (\nabla u_k \cdot \nabla u_k) dxdy = \sum_{k=1}^{6} (\nabla u_k \cdot \nabla u_k) \int_{\Delta_k} dxdy \\
&= \frac{1}{2}h^2\left(\frac{1}{h^2} + \frac{1}{h^2} + \frac{2}{h^2} + \frac{1}{h^2} + \frac{1}{h^2} + \frac{2}{h^2}\right) = 4
\end{aligned}$$

und

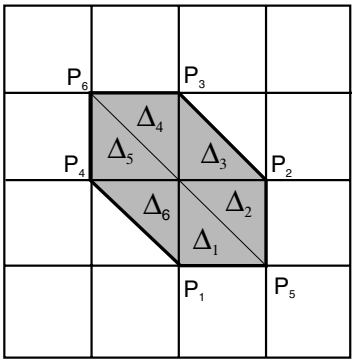


Abb. 7.3. Alle an den Punkt P_i angrenzenden Elemente Δ_k

Element	*Gradient*	*Element*	*Gradient*
Δ_1 :	$\begin{pmatrix} 0 \\ \frac{1}{h} \end{pmatrix}$	Δ_4 :	$\begin{pmatrix} 0 \\ -\frac{1}{h} \end{pmatrix}$
Δ_2 :	$\begin{pmatrix} -\frac{1}{h} \\ 0 \end{pmatrix}$	Δ_5 :	$\begin{pmatrix} \frac{1}{h} \\ 0 \end{pmatrix}$
Δ_3 :	$\begin{pmatrix} -\frac{1}{h} \\ -\frac{1}{h} \end{pmatrix}$	Δ_6 :	$\begin{pmatrix} \frac{1}{h} \\ \frac{1}{h} \end{pmatrix}$

Tabelle 7.1. Gradienten für alle an den Punkt P_i angrenzenden Elemente Δ_k

$$
\begin{aligned}
[u_i, u_1] &= \int_{\Omega} (\nabla u_i \cdot \nabla u_1)\, dxdy \\
&= \int_{T_i \cap T_1} (...)\, dxdy = \int_{\Delta_1} (...)\, dxdy + \int_{\Delta_6} (...)\, dxdy \\
&= \int_{\Delta_1} \begin{pmatrix} 0 \\ \frac{1}{h} \end{pmatrix} \cdot \begin{pmatrix} -\frac{1}{h} \\ -\frac{1}{h} \end{pmatrix} dxdy + \int_{\Delta_6} \begin{pmatrix} \frac{1}{h} \\ \frac{1}{h} \end{pmatrix} \cdot \begin{pmatrix} 0 \\ -\frac{1}{h} \end{pmatrix} dxdy \\
&= -\frac{2}{h^2} \frac{1}{2} h^2 = -1
\end{aligned}
$$

bzw.

$$\begin{aligned}[u_i, u_5] &= \int_{\Omega} (\nabla u_i \cdot \nabla u_5)\, dxdy \\ &= \int_{T_i \cap T_5} (...)\, dxdy = \int_{\Delta_1} (...)\, dxdy + \int_{\Delta_5} (...)\, dxdy \\ &= \int_{\Delta_1} \begin{pmatrix} 0 \\ \frac{1}{h} \end{pmatrix} \cdot \begin{pmatrix} \frac{1}{h} \\ 0 \end{pmatrix} dxdy + \int_{\Delta_5} \begin{pmatrix} -\frac{1}{h} \\ 0 \end{pmatrix} \cdot \begin{pmatrix} 0 \\ -\frac{1}{h} \end{pmatrix} dxdy = 0 \ .\end{aligned}$$

Analog berechnet man $[u_i, u_2] = [u_i, u_3] = [u_i, u_4] = 1$ und $[u_i, u_6] = 0$.

Die rechten Seite $< u_i, f >$ des LGS ergibt sich aus

$$\begin{aligned}< u_i, f > &= \int_{\Omega} u_i(x,y) f(x,y)\ dxdy = \int_{T_i} u_i(x,y) f(x,y)\ dxdy \\ &= \sum_{k=1}^{6} \int_{\Delta_k} u_i(x,y) f(x,y)\ dxdy = \sum_{k=1}^{6} \overline{f}_{\Delta_k} \int_{\Delta_k} u_i(x,y)\ dxdy \ ,\end{aligned}$$

wenn $\overline{f}_{\Delta_k}$ der Mittelwert der Funktion f auf dem Dreieck Δ_k ist. Das Volumen der Pyramide ist 1/3×Grundfläche×Höhe, d.h. $\int_{\Delta_k} u_i(x,y)\ dxdy = \frac{1}{3}\frac{1}{2}h^2\, 1$. Folglich ist

$$< u_i, f >= h^2 \frac{1}{6} \sum_{k=1}^{6} \overline{f}_{\Delta_k} = h^2 \overline{f}_i \ .$$

Zusammenfassend erhalten wir für den Punkt P_i die Gleichung

$$-\delta_1 - \delta_4 + 4\delta_i - \delta_2 - \delta_3 = h^2 \overline{f}_i$$

bzw.

$$\frac{1}{h^2}(\delta_1 + \delta_4 - 4\delta_i + \delta_2 + \delta_3) = -\overline{f}_i \ .$$

Dies ist genau der **5-Punkte-Stern** des Laplace-Operators, den wir bereits über die Differenzenmethode erhalten haben:

Allerdings steht nun auf der rechten Seite nicht mehr $f_i = f(x_i, y_i)$, sondern ein Mittelwert $\overline{f}_i = \frac{1}{6}\sum_{k=1}^{6} \overline{f}_{\Delta_k}$ über den Mittelwerten auf den Dreiecken Δ_k. Sind f_k die Funktionswerte in den Punkten P_k, so ist in linearer Näherung $\overline{f}_{\Delta_1} = \frac{1}{3}(f_i + f_1 + f_5)$ usw., so dass

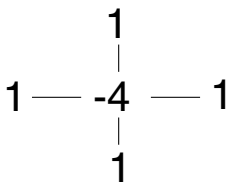


Abb. 7.4. 5-Punkte-Stern für den Laplace-Operator

$$\overline{f}_i = \frac{1}{9}\sum_{k=1}^{6} f_k + \frac{1}{3} f_i$$

gilt.

Beispiel 7.1 (Mit MAPLE-Worksheet). Im Worksheet ist das linear Gleichungssystem für eine gleichmäßige Triangulierung des Einheitsquadrates mit den Pyramidenfunktionen als Basisfunktionen aufgestellt.

Dabei werden sowohl in x- als auch in y-Richtung $N = 20$ Unterteilungen vorgenommen. Die Punkte sind als zweidimensionales Feld gespeichert, $P_{i,j} = (x_{i,j}, y_{i,j})$, $x_{i,j} = (i-1)h$, $y_{i,j} = (j-1)h$, $i, j = 1, ..., N+1$ mit $h = \frac{1}{N}$. Der erste Index ist der Zeilenindex und der zweite der Spaltenindex. Die unbekannten Koeffizienten lauten dann $\delta_{i,j}$. Das LGS wird mit dem SOR-Verfahren gelöst, indem man zuerst die Gleichung für den Punkt $P_{i,j}$ nach $\delta_{i,j}$ auflöst

$$\delta_{i,j} = \frac{1}{4}(h^2 \overline{f}_{i,j} + \delta_{i-1,j} + \delta_{i+1,j} + \delta_{i,j-1} + \delta_{i,j+1})$$

und dann iteriert. Im Worksheet wird diese Lösung mit der exakten Lösung des Problems $u(x, y) = sin(2\pi x)\ sin(2\pi y)$ grafisch verglichen. Die dreidimensionale Darstellung der numerischen Lösung ergibt sich auf dem 21×21 Gitter aus Abbildung 7.5.

7.2 Triangulierung mit linearen Elementfunktionen

Im Hinblick auf die Einführung von Funktionen höherer Ordnung (bilinear und quadratisch) ist eine modifizierte Betrachtungsweise der Ansatz-

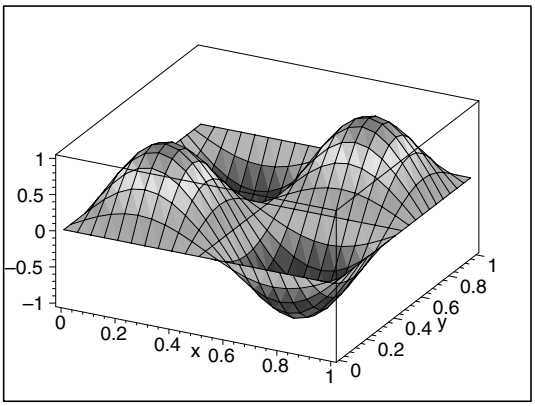


Abb. 7.5. Numerische Lösung der Poisson-Gleichung

funktionen geeigneter, indem sie nicht als Pyramidenfunktionen mit Träger $T_i = \bigcup_{k=1}^{6} \Delta_k$ definiert werden, sondern als Elementfunktionen jeweils nur auf einem Dreieck (Element), d.h. der Träger der Elementfunktionen $u_j(x, y)$ ist Δ_j.

Das Integral der Variationsaufgabe zerfällt dann in Teilintegrale über die Elemente Δ_j $(j = 1, ..., m)$; die Elementfunktionen überlappen sich nicht:

$$[u, u] = \int_{\Omega} (\nabla u \cdot \nabla u) dxdy = \sum_{j=1}^{m} \int_{\Delta_j} (\nabla u_j \cdot \nabla u_j) dxdy$$

$$< u, f > = \int_{\Omega} u(x, y) f(x, y)\ dxdy = \sum_{j=1}^{m} \int_{\Delta_j} u_j f\ dxdy\ .$$

Für die lineare Elementfunktion $u_j(x, y)$ eines Dreiecks Δ_j müssen drei Koeffizienten c_0, c_1, c_2 gefunden werden, so dass

$$u_j(x, y) = c_0 + c_1 x + c_2 y$$

mit

$$u_j(P_0) = \delta_0,\ u_j(P_1) = \delta_1,\ u_j(P_2) = \delta_2\ .$$

Damit hat man innerhalb des Dreiecks Δ_j einen linearen Verlauf und $\delta_0,\ \delta_1,\ \delta_2$ sind wieder die Werte der gesuchten Funktion an den Knotenpunkten. Die Träger der Elementfunktionen überlappen sich zwar nicht, um

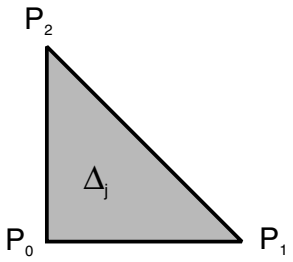


Abb. 7.6. Dreieck Δ_j mit den Eckpunkten P_0, P_1 und P_2.

aber die Bestimmungsgleichung für die Unbekannte δ_i im Punkt P_i zu erhalten, müssen die Beiträge aller zum Punkte P_i gehörenden Elemente (siehe Abbildung 7.3) berücksichtigt werden. Dieses Zusammenfassen nennt man wiederum **Assemblierung**.

Bei einer gleichmäßigen Triangulierung des Einheitsquadrates I^2 ist der Beitrag von δ_i zum Energiefunktional

$$\Phi(\delta_i) = \sum_{k=1}^{6} \Big(\int_{\Delta_k} (\nabla u_k \cdot \nabla u_k)\, dxdy - 2 \int_{\Delta_k} u_k f \; dxdy \Big)$$

$$\Rightarrow \frac{\partial \Phi}{\partial \delta_i}(\delta_i) = \sum_{k=1}^{6} \frac{\partial}{\partial \delta_i} \Big(\int_{\Delta_k} (\nabla u_k \cdot \nabla u_k)\, dxdy - 2 \int_{\Delta_k} u_k f \; dxdy \Big) = 0 \; .$$

Somit stellt man für jedes an den Punkt P_i angrenzende Dreieck Δ_k die Elementfunktion u_k auf, berechnet die Integrale $\int_{\Delta_k} (\nabla u_k \cdot \nabla u_k)\, dxdy$ und $\int_{\Delta_k} u_k f \; dxdy$, differenziert nach δ_i und summiert alle Beiträge zu Null auf.

In Abbildung 7.3 sind alle an den Punkt P_i angrenzenden Dreiecke Δ_k angegeben. Die Funktionswerte der Elementfunktionen in den Ecken des Trägers lauten δ_1, ..., δ_6; der Funktionswert im Punkte P_i in der Mitte des Trägers ist δ_i.

Wir stellen exemplarisch die Rechnung für das Dreieck Δ_2 mit der Elementfunktion z_2 vor: Der Einfachheit halber gehen wir davon aus, dass der Punkt P_i die Koordinaten $(0, 0)$ besitzt. Die Funktionswerte in den Punkten $P_i = (0, 0)$, $P_5 = (h, -h)$ und $P_2 = (h, 0)$ lauten δ_i, δ_5 und δ_5.

Die Ebene, die durch die Funktionswerte aufgespannt wird, lässt sich in Vektorschreibweise einfach angeben:

$$\Delta_2 : \begin{pmatrix} x \\ y \\ z \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ \delta_i \end{pmatrix} + \lambda \begin{pmatrix} h \\ -h \\ \delta_5 - \delta_i \end{pmatrix} + \mu \begin{pmatrix} h \\ 0 \\ \delta_2 - \delta_i \end{pmatrix}.$$

Schreibt man diese Ebenengleichung in Komponenten und ersetzt in der 3. Komponente λ und μ, erhält man

$$\begin{aligned} z_2(x,y) &= \delta_i + \frac{1}{h}(\delta_2 - \delta_i)x + \frac{1}{h}(\delta_2 - \delta_5)y \\ \Rightarrow \nabla z_2 &= \frac{1}{h}\begin{pmatrix} \delta_2 - \delta_i \\ \delta_2 - \delta_5 \end{pmatrix}. \end{aligned}$$

Für das Dreieck Δ_2 ergibt sich damit

$$\int_{\Delta_2} (\nabla z_2 \cdot \nabla z_2)\, dxdy = \frac{1}{h^2}\left((\delta_2 - \delta_i)^2 + (\delta_2 - \delta_5)^2\right)\frac{1}{2}h^2$$

$$\int_{\Delta_2} z_2 f \; dxdy = \overline{f}_{\Delta_2} \int_{\Delta_2} z_2 \; dxdy = \overline{f}_{\Delta_2} \; \frac{1}{6}h^2(\delta_i + \delta_2 + \delta_5)$$

bzw.

$$\frac{\partial}{\partial \delta_i}\left(\int_{\Delta_2} (\nabla z_2 \cdot \nabla z_2)\, dxdy - 2\int_{\Delta_2} z_2 f \; dxdy\right) = (\delta_i - \delta_2) - 2\overline{f}_{\Delta_2} \; \frac{1}{6}h^2 .$$

Analog zum Dreieck Δ_2 mit der Elementfunktion z_2 werden die Funktionsgleichungen für z_1, z_3, z_4, z_5, z_6 aufgestellt und die zugehörigen Gradienten bestimmt. In Tabelle 7.2 sind alle zum Punkte P_i angrenzenden Dreiecke (vgl. Abbildung 7.3) mit den Gradienten der Elementfunktionen angegeben.

Für den Spezialfall $\delta_i = 1$ und $\delta_1 = ... = \delta_6 = 0$ ergeben sich genau die Gradienten der Pyramidenfunktion in Tabelle 7.2. Damit folgt für die Integrale auf den anderen angrenzenden Elementen

Gradient auf dem Element		Gradient auf dem Element	
$\Delta_1:$	$\nabla z_1 = \frac{1}{h}\begin{pmatrix}\delta_5-\delta_1\\ \delta_i-\delta_1\end{pmatrix}$	$\Delta_4:$	$\nabla z_4 = \frac{1}{h}\begin{pmatrix}\delta_3-\delta_6\\ \delta_3-\delta_i\end{pmatrix}$
$\Delta_2:$	$\nabla z_2 = \frac{1}{h}\begin{pmatrix}\delta_2-\delta_i\\ \delta_2-\delta_5\end{pmatrix}$	$\Delta_5:$	$\nabla z_5 = \frac{1}{h}\begin{pmatrix}\delta_i-\delta_4\\ \delta_6-\delta_4\end{pmatrix}$
$\Delta_3:$	$\nabla z_3 = \frac{1}{h}\begin{pmatrix}\delta_2-\delta_i\\ \delta_3-\delta_1\end{pmatrix}$	$\Delta_6:$	$\nabla z_6 = \frac{1}{h}\begin{pmatrix}\delta_i-\delta_4\\ \delta_i-\delta_1\end{pmatrix}$

Tabelle 7.2. Gradienten der Elementfunktionen zum Punkte P_i

$$\Delta_1: \ \frac{\partial}{\partial\delta_i}\Big(\int\limits_{\Delta_1}(...)dxdy - 2\int\limits_{\Delta_1} z_1 f\ dxdy\Big) = (\delta_i-\delta_1) - 2\overline{f}_{\Delta_1}\ \frac{1}{6}h^2\ ,$$

$$\Delta_3: \ \frac{\partial}{\partial\delta_i}\Big(\int\limits_{\Delta_3}(...)\ dxdy - 2\int\limits_{\Delta_3} z_3 f\ dxdy\Big) = (2\delta_i-\delta_2-\delta_3) - 2\overline{f}_{\Delta_3}\ \frac{1}{6}h^2\ ,$$

$$\Delta_4: \ \frac{\partial}{\partial\delta_i}\Big(\int\limits_{\Delta_4}(...)\ dxdy - 2\int\limits_{\Delta_4} z_4 f\ dxdy\Big) = (\delta_i-\delta_3) - 2\overline{f}_{\Delta_4}\ \frac{1}{6}h^2\ ,$$

$$\Delta_5: \ \frac{\partial}{\partial\delta_i}\Big(\int\limits_{\Delta_5}(...)\ dxdy - 2\int\limits_{\Delta_5} z_5 f\ dxdy\Big) = (\delta_i-\delta_4) - 2\overline{f}_{\Delta_5}\ \frac{1}{6}h^2\ ,$$

$$\Delta_6: \ \frac{\partial}{\partial\delta_i}\Big(\int\limits_{\Delta_6}(...)\ dxdy - 2\int\limits_{\Delta_6} z_6 f\ dxdy\Big) = (2\delta_i-\delta_4-\delta_1) - 2\overline{f}_{\Delta_6}\ \frac{1}{6}h^2\ .$$

Berücksichtigt man alle zum Punkte P_i angrenzenden Dreiecke, so folgt aus $\frac{\partial}{\partial\delta_i}\Phi(\delta_i)=0$:

$$-2\delta_1 - 2\delta_4 + 8\delta_i - 2\delta_2 - 2\delta_3 - 2h^2\,\frac{1}{6}\sum_{i=1}^{6}\overline{f}_{\Delta_i} = 0$$

bzw.

$$\frac{1}{h^2}\ (\delta_1+\delta_4-4\delta_i+\delta_2+\delta_3) = -\overline{f}\ ,$$

wenn $\overline{f} = \frac{1}{6}\sum_{i=1}^{6}\overline{f}_{\Delta_i}$ ein mittlerer Funktionswert von f. Dies führt für den Punkt P_i bzw. der Unbekannten δ_i wieder genau auf den 5-Punkte-Stern, den

wir auch mit dem Ansatz über die Basisfunktionen (=Pyramidenfunktionen) im Abschnitt 7.1 erhalten haben.

Beispiel 7.2 (Mit MAPLE-Worksheet). Im Worksheet ist das Erstellen der Ebenengleichungen für die Dreiecke, das Aufstellen der Elementfunktionen, die Berechnung der Gradienten und der zugehörigen Integrale durchgeführt. □

7.3 Rechteckzerlegung mit bilinearen Elementen

Wir lösen nochmals das Randwertproblem für die Poisson-Gleichung mit homogenen Dirichlet-Randwerte:

$$\begin{aligned} -\Delta u(x,y) &= 8\pi^2 \, sin(2\pi x) \, sin(2\pi y) \quad \text{in } I^2 \,, \\ u(x,y) &= 0 \qquad\qquad\quad \text{auf dem Rand von } I^2 \,. \end{aligned}$$

und zerlegen das Einheitsquadrat I^2 jetzt in einzelne Quadrate mit der Seitenlänge $h = 1/N$, wie dies in Abbildung 7.7 skizziert ist.

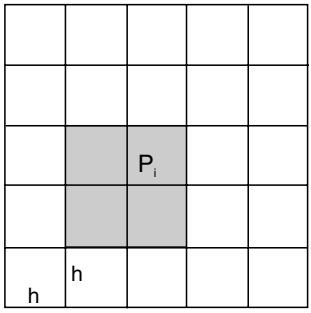
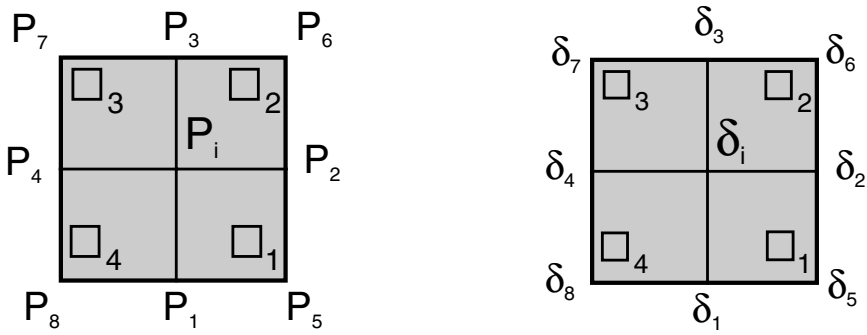


Abb. 7.7. Gleichmäßiges Rechteckgitter

Wir betrachten den Fall, bei dem die Elementfunktionen auf jedem Quadrat bilinear sind:

$$u(x,y) = c_0 + c_1 x + c_2 y + c_3 xy \,.$$

Dieser Ansatz ist sinnvoll, denn durch die Vorgabe der Funktionswerte in den Eckpunkten ist $u(x,y)$ in jedem Quadrat eindeutig bestimmt und stetig.



(a) Punkte (b) Funktionswerte

Abb. 7.8. Punkt P_i mit den angrenzenden Elementen $\square_1$, ..., $\square_4$

Bei der gleichmäßigen Zerlegung des Einheitsquadrates I^2 in quadratische Elemente ist der Beitrag von δ_i zum Energiefunktional

$$\Phi(\delta_i) = \sum_{k=1}^{4} \left(\int_{\square_k} \nabla u_k \cdot \nabla u_k \, dxdy - 2 \int_{\square_k} u_k f \; dxdy \right) .$$

Damit lautet die Bestimmungsgleichung für die Unbekannte δ_i

$$\frac{\partial \Phi}{\partial \delta_i}(\delta_i) = \sum_{k=1}^{4} \frac{\partial}{\partial \delta_i} \left(\int_{\square_k} \nabla u_k \cdot \nabla u_k \; dxdy - 2 \int_{\square_k} u_k f \; dxdy \right) = 0 \, . \quad (7.12)$$

In Abbildung 7.8(b) sind alle an den Punkt P_i angrenzenden Quadrate $\square_k$ angegeben. Die Funktionswerte der Elementfunktionen in den Ecken der Quadrate lauten δ_1, ..., δ_8; der Funktionswert im Punkte P_i ist δ_i.

Wir stellen exemplarisch die Rechnung für das Quadrat $\square_1$ mit der Elementfunktion z_1 auf. Der Einfachheit halber gehen wir davon aus, dass der Punkt P_i die Koordinaten $(0, 0)$ besitzt. Die Details der Rechnung sind im MAPLE-Worksheet zu finden. Wir geben hier nur die Ergebnisse an: Für die Elementfunktion z_1 des Quadrates $\square_1$ müssen vier Koeffizienten c_1, ..., c_4 gefunden werden, so dass

$$z_1(x, y) = c_0 + c_1 x + c_2 y + c_3 xy$$

mit

$$z_1(P_1) = \delta_1,\ z_1(P_5) = \delta_5,\ z_1(P_2) = \delta_2,\ z_1(P_0) = \delta_i\ .$$

Wie man direkt nachprüfen kann, erfüllt die Funktion

$$z_1(x,y) = \delta_i + \frac{1}{h}(\delta_2 - \delta_i)x + \frac{1}{h}(\delta_i - \delta_1)y + \frac{1}{h^2}(\delta_2 - \delta_5 + \delta_1 - \delta_i)x\,y$$

diese Forderungen. Folglich ist

$$\nabla\, z_1 = \begin{pmatrix} \frac{1}{h}(\delta_2 - \delta_i) + \frac{1}{h^2}(\delta_2 - \delta_5 + \delta_1 - \delta_i)y \\ \\ \frac{1}{h}(\delta_i - \delta_1) + \frac{1}{h^2}(\delta_2 - \delta_5 + \delta_1 - \delta_i)x \end{pmatrix}$$

und damit

$$\begin{aligned} \int\limits_{\square_1} \nabla z_1 \cdot \nabla z_1\ dxdy &= \int\limits_{y=-h}^{0}\ \int\limits_{x=0}^{h} \nabla z_1 \cdot \nabla z_1\ dxdy = \\ &= \frac{2}{3}\delta_i^2 + \frac{2}{3}\delta_1^2 + \frac{2}{3}\delta_2^2 + \frac{2}{3}\delta_5^2 - \frac{2}{3}\delta_i\delta_5 - \frac{2}{3}\delta_2\delta_1 \\ &\quad -\frac{1}{3}\delta_5\delta_1 - \frac{1}{3}\delta_2\delta_i - \frac{1}{3}\delta_2\delta_5 - \frac{1}{3}\delta_i\delta_1 \end{aligned}$$

$$\Rightarrow \frac{\partial}{\partial\delta_i}\int\limits_{\square_1} \nabla z_1 \cdot \nabla z_1\ dxdy = \frac{4}{3}\delta_i - \frac{2}{3}\delta_5 - \frac{1}{3}\delta_2 - \frac{1}{3}\delta_1\ .$$

Für das zweite Integral ergibt sich

$$-2\int\limits_{\square_1} z_1 f\ dxdy = -2\,\overline{f}_{\square_1}\int\limits_{\square_1} z_1\ dxdy = -\frac{1}{2}\,h^2\,\overline{f}_{\square_1}(\delta_i + \delta_1 + \delta_2 + \delta_5)$$

$$\Rightarrow \frac{\partial}{\partial\delta_i}(-2\int\limits_{\square_1} z_1 f\ dxdy) = -\frac{1}{2}\,h^2\,\overline{f}_{\square_1}\ .$$

Setzen wir nun alle Beiträge der zum Punkte P_i gehörenden Elementfunktionen zusammen (Assemblieren), gilt

$$\frac{\partial\Phi}{\partial\delta_i}(\delta_i) = 0 \curvearrowright \sum_{k=1}^{4} \frac{\partial}{\partial\delta_i}(\int\limits_{\square_k} \nabla u_k \cdot \nabla u_k\, dxdy - 2\int\limits_{\square_k} u_k f\ dxdy) = 0$$

$$\curvearrowright \frac{4}{3}\delta_i - \frac{2}{3}\delta_5 - \frac{1}{3}\delta_2 - \frac{1}{3}\delta_1 + \frac{4}{3}\delta_i - \frac{2}{3}\delta_6 - \frac{1}{3}\delta_2 - \frac{1}{3}\delta_3 +$$

$$\frac{4}{3}\delta_i - \frac{2}{3}\delta_7 - \frac{1}{3}\delta_3 - \frac{1}{3}\delta_4 + \frac{4}{3}\delta_i - \frac{2}{3}\delta_8 - \frac{1}{3}\delta_4 - \frac{1}{3}\delta_1 - \frac{1}{2}\, h^2 \sum_{k=1}^{4} \overline{f}_{\square_k} = 0 \,.$$

Dies ist offenbar gerade das lineare Gleichungssystem, welches bei Anwendung eines Differenzenverfahrens mit dem 9-Punkte-Stern

$$\begin{array}{|ccc|}\hline \frac{1}{3} & \frac{1}{3} & \frac{1}{3} \\ \frac{1}{3} & -\frac{8}{3} & \frac{1}{3} \\ \frac{1}{3} & \frac{1}{3} & \frac{1}{3} \\ \hline \end{array}$$

und der rechten Seite

$$-h^2\, \frac{1}{4} \sum_{k=1}^{4} \overline{f}_{\square_k} = -h^2\, \overline{f}$$

entsteht.

Bemerkung: Solange die Gitterstruktur eine Regelmäßigkeit aufweist, die logisch durch ein zweidimensionales Feld repräsentiert werden kann, führt die Methode der Finiten-Elemente auf das gleiche LGS wie ein Differenzenverfahren. Bei allgemeinen Zerlegungen ist eine solche Zuordnung jedoch nicht mehr möglich.

Beispiel 7.3 (Mit MAPLE-Worksheet). Die Lösung des RWP wird im Worksheet mit Hilfe des oben diskutierten 9-Punkte-Sterns numerisch bestimmt, grafisch dargestellt sowie mit der exakten Lösung verglichen. □

7.4 Triangulierung mit quadratischen Elementen

In diesem Abschnitt betrachten wir das Randwertproblem für die verallgemeinerte Poisson-Gleichung

$$\begin{aligned} -\Delta u(x,y) + c u(x,y) &= f(x,y) \quad \text{im Innern eines Gebiets } \Omega\,, && (7.13)\\ u(x,y) &= 0 \qquad\quad \text{auf dem Rand von } \Omega\,. && (7.14) \end{aligned}$$

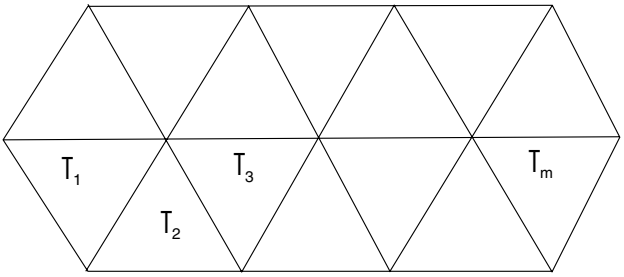


Abb. 7.9. Triangulierung des Berechnungsgebiets

Wir gehen davon aus, dass das Berechnungsgebiet Ω trianguliert ist.

1. Elementfunktionen: Um die Gleichungen für einen Punkt aufzustellen, beschränken wir uns im Folgenden auf ein beliebiges Dreieck T_i. Zu diesem Dreieck bestimmen wir eine quadratische Elementfunktion:

$$u_i(x, y) = c_1 + c_2 x + c_3 y + c_4 x^2 + c_5 xy + c_6 y^2 \ .$$

Da wir uns zunächst auf *ein* Dreieck beschränken, unterdrücken wir der Übersichtlichkeit wegen den Index i bei T_i und u_i. Wenn man von quadratischen Elementfunktionen auf Dreiecken ausgeht, hat man pro Dreieck 6 Koeffizienten $c_i, i = 1, ..., 6$ festzulegen. Daher reichen die 3 Eckpunkte des Dreiecks nicht mehr aus.

Man führt deshalb auf den Seitenlinien der Dreiecke die Mittelpunkte in die Betrachtung mit ein und fordert, dass die Elementfunktion in den 6 Knoten $P_1, P_2, ... \ , P_6$ die Werte $\delta_1, \delta_2, ... \ , \delta_6$ annimmt. In Abbildung 7.10 sind die Knotenpunkte des Dreiecks angegeben.

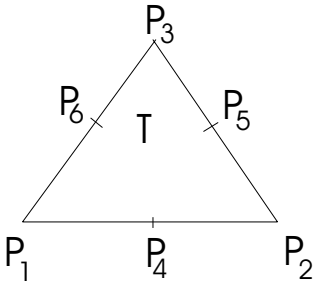


Abb. 7.10. Dreieck mit den 6 Knoten $P_1, P_2, ... \ , P_6$

Es zeigt sich, dass man die Basisfunktionen für das Dreieck T am besten so konstruiert, dass man sie als Linearkombination von 6 *quadratischen* Einzel-

funktionen $N_1, \ldots, N_6$ zusammensetzt, welche die Eigenschaft besitzen, dass N_i auf dem Punkt P_i den Wert eins und auf allen anderen 5 Punkten Null liefert:

$$N_i(P_j) = \begin{cases} 0 & \text{für } i \neq j \\ 1 & \text{für } i = j \end{cases} \qquad \text{für} \qquad i, j = 1, \ldots, 6 \; .$$

Es ist aber nicht offensichtlich, wie in einem beliebigen Dreieck solche Funktionen $N_1, \ldots, N_6$ zu erklären sind.

2. Definition der Elementfunktionen im (ξ, η)-Gebiet: Um die Funktionsausdrücke der Elementfunktionen aufzustellen, transformieren wir das Dreieck T auf ein **Normaldreieck** $\widetilde{T}$:

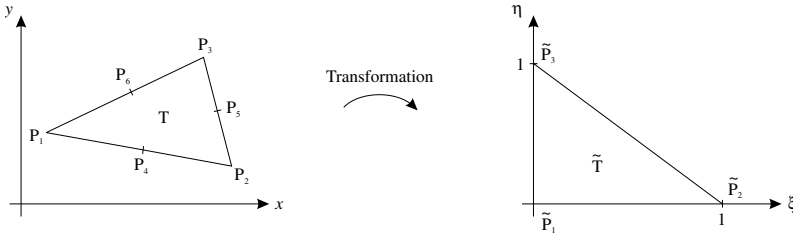


Abb. 7.11. Transformation des Dreiecks T auf das *Normaldreieck* $\widetilde{T}$

Jeder Punkt im Dreieck T lässt sich darstellen durch die Vorschrift

$$\begin{cases} x = x_1 + (x_2 - x_1)\xi + (x_3 - x_1)\eta \\ y = y_1 + (y_2 - y_1)\xi + (y_3 - y_1)\eta, \end{cases} \qquad 0 \leq \xi + \eta \leq 1$$

$$\Leftrightarrow \begin{pmatrix} x - x_1 \\ y - y_1 \end{pmatrix} = \begin{pmatrix} x_2 - x_1 & x_3 - x_1 \\ y_2 - y_1 & y_3 - y_1 \end{pmatrix} \begin{pmatrix} \xi \\ \eta \end{pmatrix} \; .$$

Durch Inversion der Matrix gilt

$$\begin{pmatrix} \xi \\ \eta \end{pmatrix} = \frac{1}{J} \begin{pmatrix} y_3 - y_1 & -(x_3 - x_1) \\ -(y_2 - y_1) & x_2 - x_1 \end{pmatrix} \begin{pmatrix} x - x_1 \\ y - y_1 \end{pmatrix}$$

mit

$$J = det \begin{pmatrix} x_2 - x_1 & x_3 - x_1 \\ y_2 - y_1 & y_3 - y_1 \end{pmatrix} \; .$$

Mit obigen Formeln hat man die Transformation von einem beliebigen Dreieck T mit den Ecken $(P_1(x_1, y_1), P_2(x_2, y_2), P_3(x_3, y_3))$ ins Normaldreieck $\widetilde{T}$.

Im Dreieck $\widetilde{T}$ können die **Formfunktionen** $N_1, \ldots, N_6$ explizit angegeben werden durch

$$
\begin{aligned}
N_1(\xi,\eta) &= (1-\xi-\eta)(1-2\xi-2\eta)\ , \\
N_2(\xi,\eta) &= \xi(2\xi-1)\ , \\
N_3(\xi,\eta) &= \eta(2\eta-1)\ , \\
N_4(\xi,\eta) &= 4\xi(1-\xi-\eta)\ , \\
N_5(\xi,\eta) &= 4\xi\eta\ , \\
N_6(\xi,\eta) &= 4\eta(1-\xi-\eta)\ .
\end{aligned}
$$

Diese Funktionen sind im MAPLE- **Worksheet** dreidimensional dargestellt und können dort unter verschiedenen Blickwinkeln betrachtet werden (siehe Abbildung 7.4).

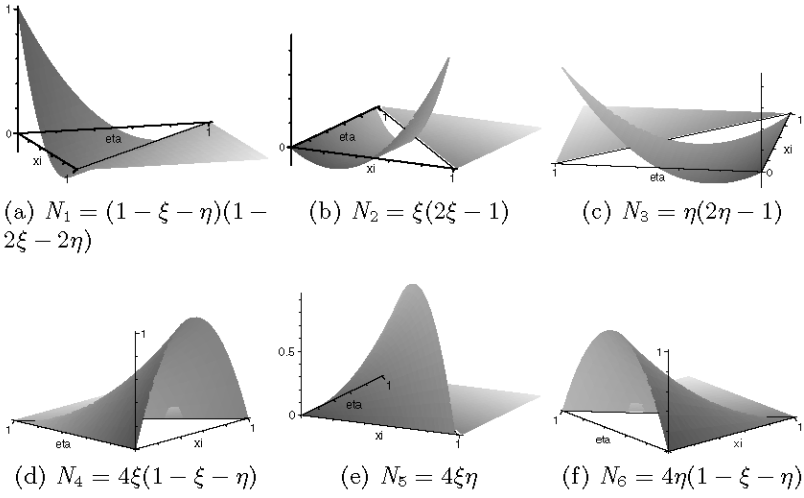


Abb. 7.12. *Formfunktionen* $N_1, \ldots, N_6$ auf dem Element $\widetilde{T}$.

Die zum Dreieck $\widetilde{T}$ gehörende Elementfunktion $\widetilde{u}$ ist dann als Linearkombination der Formfunktionen definiert durch

$$\widetilde{u}(\xi,\eta) = \sum_{i=1}^{6} \delta_i \ N_i(\xi,\eta)\ .$$

Um die Elementfunktion $u(x,y)$ für das ursprüngliche Dreieck T zu erhalten, könnten wir die Elementfunktion zurück transformieren. Im Anschluss daran könnten dann die Integrale berechnet, die Assemblierung vorgenommen und $\Phi(\delta_i)$ nach δ_i differenziert werden, um das lineare Gleichungssystem $\mathbf{A}\boldsymbol{\delta} = \mathbf{r}$ aufzustellen. Es zeigt sich jedoch, dass es für die Rechnung günstiger ist,

im transformierten Gebiet zu bleiben und die Integrale ins (ξ, η)-Gebiet zu übertragen.

3. Berechnung der Integrale: Um das LGS $\mathbf{A}\boldsymbol{\delta} = \mathbf{r}$ aufzustellen, müssen zunächst für jede Elementfunktion u die Integrale

$$[u,u] = \int_T (\nabla u)\cdot(\nabla u)\,dx\;dy + \int_T c\;u\;u\;dx\;dy\;,$$
$$< u, f > = \int_T u\;f\;dx\;dy$$

berechnet werden. Somit sind für jedes Dreieck T drei Integrale zu bestimmen:

$$I_1 = \int_T (u_x^2 + u_y^2)\,dx\;dy\;,$$
$$I_2 = \int_T u\;f\;dx\;dy\;,$$
$$I_3 = c\int_T u^2\;dx\;dy\;.$$

Um die Integrationen im (ξ, η)-Gebiet durchzuführen, ersetzen wir die Ableitungen nach x und y durch Ableitungen nach ξ und η sowie $\int_T \ldots dx\;dy$ durch $\int_{\widetilde{T}} \ldots J\;d\xi\;d\eta$:

Transformation von I_1 ins (ξ, η)-Gebiet: Wir wenden uns dem ersten Integral zu. Durch die Transformation wird jede Funktion $u(x, y)$ zu einer Funktion in ξ und η

$$u = u(x(\xi,\eta), y(\xi,\eta))\;.$$

Mit den Transformationsformeln für die partiellen Ableitungen

$$u_x = \frac{1}{J}(y_\eta u_\xi - y_\xi u_\eta)\;,$$
$$u_y = \frac{1}{J}(-x_\eta u_\xi + x_\xi u_\eta)$$

mit $y_\eta = y_2 - y_1$; $y_\xi = y_2 - y_1$; $x_\eta = x_3 - x_1$; $x_\xi = x_2 - x_1$ erhalten wir

$$\int_T (u_x^2 + u_y^2)\;dx\;dy = \int_{\widetilde{T}} \frac{1}{J^2}[(y_\eta u_\xi - y_\xi u_\eta)^2 + (-x_\eta u_\xi + x_\xi u_\eta)^2]\;J\;d\xi\;d\eta$$
$$= \ldots = \int_{\widetilde{T}} (\alpha u_\xi^2 + 2\beta u_\xi u_\eta + \gamma u_\eta^2)\;d\xi\;d\eta$$

mit den Koeffizienten

$$\alpha = \frac{1}{J}(x_\eta^2 + y_\eta^2); \quad \beta = \frac{1}{J}(x_\eta x_\xi + y_\eta y_\xi); \quad \gamma = \frac{1}{J}(x_\xi^2 + y_\xi^2) \; .$$

Da die Elementfunktion

$$\widetilde{u}(\xi,\eta) = \sum_{i=1}^{6} \delta_i \; N_i(\xi,\eta)$$

sich zusammensetzt aus den Formfunktionen N_i und den unbekannten Funktionswerten δ_i an den Punkten $\widetilde{P}_i, i = 1, ..., 6$, des Dreiecks $\widetilde{T}$, ist

$$\widetilde{u}_\xi(\xi,\eta) = \sum_{i=1}^{6} \delta_i \frac{\partial}{\partial\xi} N_i(\xi,\eta) = \mathbf{u}_e^T \cdot \mathbf{N}_\xi \; ,$$

$$\widetilde{u}_\eta(\xi,\eta) = \sum_{i=1}^{6} \delta_i \frac{\partial}{\partial\eta} N_i(\xi,\eta) = \mathbf{u}_e^T \cdot \mathbf{N}_\eta \; ,$$

wenn wir zur Abkürzung der Summe die transponierten Vektoren

$$\mathbf{u}_e^T = (\delta_1, \delta_2, \ldots, \delta_6),$$

$$\mathbf{N}_\xi = (\frac{\partial}{\partial\xi} N_1(\xi,\eta), \frac{\partial}{\partial\xi} N_2(\xi,\eta), \ldots, \frac{\partial}{\partial\xi} N_6(\xi,\eta))^t$$

bzw.

$$\mathbf{N}_\eta = (\frac{\partial}{\partial\eta} N_1(\xi,\eta), \frac{\partial}{\partial\eta} N_2(\xi,\eta), \ldots, \frac{\partial}{\partial\eta} N_6(\xi,\eta))^t$$

einführen.

Setzt man $\widetilde{u}_\xi$ und $\widetilde{u}_\eta$ in I_1 ein, erhält man Integrale nur über die Ableitungen $\frac{\partial}{\partial\xi} N_i(\xi,\eta)$ bzw. $\frac{\partial}{\partial\eta} N_i(\xi,\eta)$:

$$I_1 = \int\limits_T (u_x^2 + u_y^2) \; dx \; dy = \mathbf{u}_e^T \; S_e \; \mathbf{u}_e$$

mit der **Elementmatrix** $S_e := \alpha \; S_1 + \beta \; S_2 + \gamma S_3$, die auch in Anlehnung an die Bezeichnung in der Statik **Steifigkeitsmatrix** genannt wird, und den Teilmatrizen:

$$S_1 := \int\limits_{\widetilde{T}} \mathbf{N}_\xi \; \mathbf{N}_\xi^T \; d\xi \; d\eta$$

$$S_2 := \int\limits_{\widetilde{T}} (\mathbf{N}_\xi \; \mathbf{N}_\eta^T + \mathbf{N}_\eta \; \mathbf{N}_\xi^T) d\xi \; d\eta$$

$$S_3 := \int\limits_{\widetilde{T}} \mathbf{N}_\eta \; \mathbf{N}_\eta^T \; d\xi \; d\eta \; .$$

Bei dieser kompakten Schreibweise ist zu beachten, dass z.B. $\mathbf{N}_\xi \cdot \mathbf{N}_\eta^T$ nicht das Skalarprodukt der Vektoren $\mathbf{N}_\xi$ und $\mathbf{N}_\eta$ ist, sondern das dyadische Produkt in der Form der 6×6-Matrix:

$$\mathbf{N}_\xi \cdot \mathbf{N}_\eta^T = \begin{pmatrix} \partial_\xi N_1 \\ \partial_\xi N_2 \\ \vdots \\ \partial_\xi N_6 \end{pmatrix} \cdot (\partial_\eta N_1, \ldots, \partial_\eta N_6) = \begin{pmatrix} \partial_\xi N_1\, \partial_\eta N_1 & \ldots & \partial_\xi N_1\, \partial_\eta N_6 \\ \partial_\xi N_2\, \partial_\eta N_1 & & \partial_\xi N_2\, \partial_\eta N_6 \\ \vdots & & \vdots \\ \partial_\xi N_6\, \partial_\eta N_1 & \ldots & \partial_\xi N_6\, \partial_\eta N_6 \end{pmatrix}$$

Somit sind S_1, S_2, S_3 bzw. S_e jeweils 6×6-Matrizen und die entsprechenden Integrale sind komponentenweise zu bilden.

Führt man die Integrationen über dem Dreieck $\widetilde{T}$ durch, erhält man für die Steifigkeitsmatrix zusammenfassend

$$S_e = \frac{1}{6} \begin{pmatrix} 3(\alpha+\beta+\gamma) & \alpha+\beta & \beta+\gamma & -4(\alpha+\beta) & 0 & -4(\beta+\gamma) \\ \alpha+\beta & 3\alpha & -\beta & -4(\alpha+\beta) & 4\beta & 0 \\ \beta+\gamma & -\beta & 3\gamma & 0 & 4\beta & -4(\beta+\gamma) \\ -4(\alpha+\beta) & -4(\alpha+\beta) & 0 & 8(\alpha+\beta+\gamma) & -8(\beta+\gamma) & 8\beta \\ 0 & 4\beta & 4\beta & -8(\beta+\gamma) & 8(\alpha+\beta+\gamma) & -8(\alpha+\beta) \\ -4(\beta+\gamma) & 0 & -4(\beta+\gamma) & 8\beta & -8(\alpha+\beta) & 8(\alpha+\beta+\gamma) \end{pmatrix}$$

Transformation von I_2 und I_3 ins (ξ, η)-Gebiet: Auf die gleiche Art werden die Integrale I_2 und I_3 berechnet:

$$\begin{aligned} I_2 &= \int_T u\, f \; dx\, dy = J \int_{\widetilde{T}} u(\xi,\eta)\, f(\xi,\eta) \; d\xi\, d\eta = J\, \overline{f} \int_{\widetilde{T}} \sum_{i=1}^{6} \delta_i \, N_i(\xi,\eta) \; d\xi\, d\eta \\ &= \overline{f}\; J \sum_{i=1}^{6} \delta_i \int_{\widetilde{T}} N_i(\xi,\eta) \; d\xi\, d\eta = \overline{f}\; \mathbf{u}_e^T \; J \int_{\widetilde{T}} \mathbf{N}_e(\xi,\eta) \; d\xi\, d\eta \\ &= \overline{f}\; \mathbf{u}_e^T \; \mathbf{b}_e \end{aligned}$$

mit dem **Elementvektor**

$$\mathbf{b}_e = J \int_{\widetilde{T}} \mathbf{N}_e(\xi,\eta) \; d\xi\, d\eta = \frac{J}{6} \begin{pmatrix} 0 \\ 0 \\ 0 \\ 1 \\ 1 \\ 1 \end{pmatrix} .$$

Für I_3 ergibt sich

$$\begin{aligned} I_3 &= c \int_T u^2(x,y)\ dx\ dy = c\ J \int_{\widetilde{T}} u^2(\xi,\eta)\ d\xi\ d\eta \\ &= c\ J \int_{\widetilde{T}} (\sum_{i=1}^{6} \delta_i\ N_i(\xi,\eta))\ (\sum_{j=1}^{6} \delta_j\ N_j(\xi,\eta))\ d\xi\ d\eta \\ &= c\ J \int_{\widetilde{T}} \mathbf{u}_e^T\ \mathbf{N}\mathbf{N}^T\ \mathbf{u}_e\ d\xi\ d\eta \\ &= c\ \mathbf{u}_e^T\ J \int_{\widetilde{T}} \mathbf{N}\ \mathbf{N}^T\ d\xi\ d\eta\ \mathbf{u}_e = \mathbf{u}_e^T\ c\ M_e\ \mathbf{u}_e\ . \end{aligned}$$

Dabei ist $\overline{f}$ der Mittelwert der Funktion über $\widetilde{T}$ und M_e nennt man wiederum in Anlehnung an die Statik die **Massenelementmatrix**

$$M_e = J \int_{\widetilde{T}} \mathbf{N}\ \mathbf{N}^T\ d\xi\ d\eta = \frac{J}{360} \begin{pmatrix} 6 & -1 & -1 & 0 & -4 & 0 \\ -1 & 6 & -1 & 0 & 0 & -4 \\ -1 & -1 & 6 & -4 & 0 & 0 \\ 0 & 0 & -4 & 32 & 16 & 16 \\ -4 & 0 & 0 & 16 & 32 & 16 \\ 0 & -4 & 0 & 16 & 16 & 32 \end{pmatrix}\ .$$

4. Aufbau des gesamten linearen Gleichungssystems: Verwendet man quadratische Ansatzfunktionen $u = u(x,y)$ für die Dreiecke T, so erhält man als unbekannte Größen die Funktionswerte $\delta_1, ..., \delta_6$ an den 6 Punkten P_i des Dreiecks: $u_i = u(P_i)$

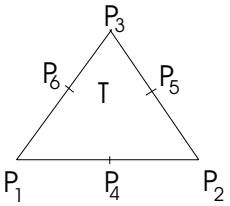


Abb. 7.13. Finites Element

Die unbekannten Größen werden zum Vektor $\mathbf{u}_e$ zusammengefasst. Für jedes Element (Dreieck T_j) $j = 1, \ldots, m$ erhält man die Steifigkeitsmatrix $S_e(j)$, die Massenelementmatrix $M_e(j)$, die beide symmetrisch sind, und den Elementvektor $\mathbf{b}_e(j)$. Die Gesamtsteifigkeitsmatrix A ergibt sich dann aus allen

$S_e(j) + cM_e(j)$ und r aus den Werten $\overline{f}_j$ $\mathbf{b}_e(j)$, wenn $\overline{f}_j = \overline{f}|_{T_j}$ der Mittelwert der Funktion über das Element T_j.

Der Aufbau von $\mathbf{A}$ und $\mathbf{r}$ heißt Kompilationsprozess. Die prinzipielle Vorgehensweise ist, dass man alle inneren Elemente nummeriert $T_1, \ldots, T_m$. Alle inneren Punkte von T_j werden nummeriert gemäß obiger Zeichnung mit $P_{j_1}, P_{j_2}, \ldots, P_{j_6}$.

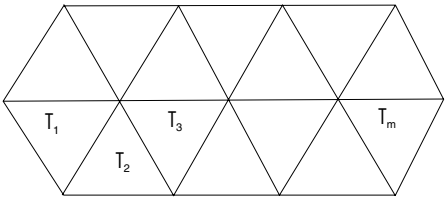


Abb. 7.14. Finites Elemente Netz

Die Kunst bei der Nummerierung besteht darin, die Differenz der Indizes zweier Punkte, die zum gleichen Dreieck gehören, möglichst klein zu halten, was dann auf eine minimale Bandbreite des Koeffizientenmatrix $\mathbf{A}$ führt.

Bemerkungen:

1. Bei der Berechnung der Matrizen benötigt man Integrale der Form $\int_{\widetilde{T}} \xi^p \eta^q \, d\xi \, d\eta = \frac{p!q!}{(p+q+2)!}$ für $p, q \in \mathbb{N}_0$.

2. Bei der Überdeckung mit Parallelogrammen benötigt man analog $\int_Q \xi^p \eta^q \, d\xi \, d\eta = \int_0^1 \int_0^1 \xi^p \eta^q \, d\xi \, d\eta = \frac{1}{(p+1)(q+1)}$.

3. Die Geometrie des Dreiecks erscheint nur in den Faktoren α, β, γ und J.

4. **Konvergenzsatz**: Ist u die exakte Lösung des RWP, u_s die Lösung des Finite-Elemente-Verfahrens, h_{max} die maximale Seitenlänge der Elemente, so gilt
$$u_s(x, y) \to u(x, y) \quad \text{für} \quad h_{max} \to 0 \, .$$

5. Hinsichtlich der Interpolations- und Fehlerabschätzungen sei auf die weiterführende, theoretische Literatur zur Methode der finiten Elemente (z.B. [5]) verwiesen.

7.5 Bemerkungen und Entscheidungshilfen

Die Finite-Elemente-Methoden haben sich in den letzten Jahren zum meist benutzten Näherungsverfahren entwickelt. FE-Methoden werden in vielen verschiedenen Ingenieurbereichen angewandt. Das Einsatzgebiet sind vor allem Feldprobleme in der Festköpermechanik, der Statik und Dynamik, in der Elektrotechnik und der Strömungsmechanik. So existieren viele kommerzielle Software-Pakete, wie z.B. ABAQUS, ANSYS und NASTRAN und PERMAS. Moderne CAD-Systeme bieten oft ein FE-Berechnungsmodul, um die Verformung von Bauteilen schon bei der Konstruktion analysieren zu können.

Der Vorteil von FE-Methoden ist ihre Flexibilität im Hinblick auf die Diskretisierung des Rechengebietes und die einfache Berücksichtigung von Randbedingungen. Dadurch, dass hier unstrukturierte Gitter mit unterschiedlichen Gitterelementen benutzt werden können, ist die Berechnung von Problemen in komplexen Geometrien möglich. Auch an die Lösung angepasste Gitter können dabei eingeführt werden. Das unstrukturierte Gitter macht es möglich, Gitterverfeinerungen lokal einzuführen. So lässt sich ein Dreieck einfach in zwei oder drei neue Dreiecke unterteilen und ein neues verfeinertes Netz definieren.

Wir haben hier Verfahren 2. Ordnung mit linearen Ansatzfunktionen und Verfahren 3. Ordnung mit quadratischen Ansatzfunktionen beschrieben. Durch eine weitere Erhöhung des Polynomgrades der Basisfunktionen können auch Verfahren höherer Ordnung konstruiert werden. Im Bereich der Forschung und zum Teil schon in modernen FE-Programmen kann man mit Gitterverfeinerung und hoher Ordnung sehr effizient das Programm adaptiv an die Lösung anpassen. Überall dort, wo die Lösung sehr glatt ist, wird eine hohe Ordnung des Verfahrens sich sehr günstig auswirken, überall dort, wo sich die Lösung stark ändert, muss das Gitter verfeinert werden. Gesteuert durch lokale Fehlerberechnungen und der Analyse der Glattheit der Lösung kann man dann eine lokale Gitterverfeinerung (h-refinement) oder eine lokale Erhöhung des Grades der Polynome (p-refinement) ausführen. Dies sind aber Methoden, welche augenblicklich vor allem noch im Bereich der Forschung zu Hause sind.

Die Differenzenverfahren sind hier klar im Nachteil, da sie nur auf strukturierten Gittern angewandt werden können. Gerade für komplexe Geometrien hat man hier oft eine sehr aufwändige Gittererzeugung, die mehr Arbeit und Zeit erfordert als die Lösung des eigentlichen Problems selbst. Lokale Gitterverfeinerungen machen hier große Probleme. Strukturierte Gitter haben demgegenüber den Vorteil, dass die Gitterverwaltung sehr einfach ist und hier Rechenzeit gespart wird. Bei einfachen Geometrien und komplexen Problemen kann ein Differenzenverfahren effizienter sein.

Wir haben in diesem Kapitel die FE-Methode für eine Poisson-Gleichung beschrieben. Gerade für diesen Aufgabenbereich der Feldprobleme sind die FE-Methoden sehr effizient und robust. Man kann die FE-Idee auch auf parabolische und hyperbolische Probleme anwenden. Führt man eine Approximation der Zeitableitung über einen Differenzenquotienten ein, dann ergibt sich ein räumliches Problem, welches mit der FE-Methode behandelt werden kann. Es gibt auch den Ansatz der Raum-Zeit-Finiten-Elemente. Wir werden hier nicht weiter darauf eingehen und verweisen auf die Literatur.

Die Literatur im Bereich der Finite-Elemente-Verfahren ist sehr umfangreich. Einführende Bücher, sehr geeignet für Ingenieure, sind die von Jung und Langer [35] und von Schwarz [54], ein klassisches Standardwerk ist das Buch von Zienkiewicz [75]. Mehr mathematisch orientiert sind die Bücher von Braess [5], Brenner und Scott [6], Ciarlet [8] und Strang und Fix [60]. Neben den FE-Verfahren enthält das Buch von Steinbach [57] auch so genannte Randelement-Verfahren. Die folgenden Monographien allgemein über numerische Methoden für partielle Differenzialgleichungen enthalten längere Kapitel über FE-Methoden: [37], [48], [24] und [26]. Mathematische Grundlagen der Variationsrechnung findet sich z.B. in der Monographie von Michlin [44]. Für die Lösung der großen schwach besetzen Gleichungssysteme sei auf den Anhang Cund die Angabe der weiterführenden Literatur dort verwiesen.

7.6 Beispiele und Aufgaben

Aufgabe 7.1. Gegeben sei die Poisson-Gleichung im Innern des Einheitsquadrates I^2

$$\begin{aligned} -u_{xx} - u_{yy} &= -x^2 - y^2 + x + y \ , \\ u(x,y) &= 0 \qquad \text{auf dem Rand von } I^2 \ . \end{aligned}$$

Die exakte Lösung ist gegeben durch

$$u(x,y) = \frac{1}{2}\, x(x-1)y(y-1) \ .$$

1. Lösen sie die partielle DGL mit der Finiten-Elemente Methode für die Werte
$$\Delta x = \Delta y = \frac{1}{10}, \quad \frac{1}{20}, \quad \frac{1}{40} \quad \text{bzw.} \quad \frac{1}{100},$$
indem sie **lineare** Elemente verwenden.

2. Vergleichen sie die numerische Lösung mit der exakten Lösung des Problems. Welche Aussagen über die Genauigkeit kann man in Abhängigkeit der Schrittweite machen?

Aufgabe 7.2. Gegeben sei die Poisson-Gleichung im Innern des Einheitsquadrates I^2

$$\begin{aligned} -u_{xx} - u_{yy} &= -x^2 - y^2 + x + y \ , \\ u(x,y) &= 0 \qquad \text{auf dem Rand von } I^2 \ . \end{aligned}$$

Die exakte Lösung ist gegeben durch

$$u(x,y) = \frac{1}{2}\, x(x-1)y(y-1) \ .$$

1. Lösen Sie die partielle DGL mit der Finiten-Elemente Methode für die Werte

$$\Delta x = \Delta y = \frac{1}{10}, \quad \frac{1}{20}, \quad \frac{1}{40} \quad \text{bzw.} \quad \frac{1}{100},$$

   indem sie **bilineare** Elemente verwenden.

2. Vergleichen sie die numerische Lösung mit der exakten Lösung des Problems. Welche Aussagen über die Genauigkeit kann man in Abhängigkeit der Schrittweite machen?

3. Welche Aussagen kann man über die Genauigkeit der linearen im Vergleich zur bilinearen Approximation treffen?

Aufgabe 7.3. Gegeben sei die Poisson-Gleichung im Innern des Einheitsquadrates I^2

$$\begin{aligned} -u_{xx} - u_{yy} &= 2\pi^2 \ sin(\pi x) \ sin(\pi y) \ , \\ u(x,y) &= 0 \qquad \text{auf dem Rand von } I^2 \ . \end{aligned}$$

Die exakte Lösung ist gegeben durch

$$u(x,y) = sin(\pi x) \ sin(\pi y) \ .$$

1. Lösen sie die partielle DGL mit der Finiten-Elemente Methode für die Werte

$$\Delta x = \Delta y = \frac{1}{10}, \quad \frac{1}{20}, \quad \frac{1}{40} \quad \text{bzw.} \quad \frac{1}{100},$$

   indem sie **lineare** Elemente verwenden.

2. Vergleichen sie die numerische Lösung mit der exakten Lösung des Problems. Welche Aussagen über die Genauigkeit kann man in Abhängigkeit der Schrittweite machen?

Aufgabe 7.4. Gegeben sei die Poisson-Gleichung im Innern des Einheitsquadrates I^2

$$\begin{aligned} -u_{xx} - u_{yy} &= 2\pi^2 \; sin(\pi x) \; sin(\pi y) \; , \\ u(x,y) &= 0 \qquad \text{auf dem Rand von } I^2 \; . \end{aligned}$$

Die exakte Lösung ist gegeben durch

$$u(x,y) = sin(\pi x) \; sin(\pi y) \; .$$

1. Lösen sie die partielle DGL mit der Finiten-Elemente Methode für die Werte

   $$\Delta x = \Delta y = \frac{1}{10}, \quad \frac{1}{20}, \quad \frac{1}{40} \quad \text{bzw.} \quad \frac{1}{100},$$

   indem sie **bilineare** Elemente verwenden.

2. Vergleichen sie die numerische Lösung mit der exakten Lösung des Problems. Welche Aussagen über die Genauigkeit kann man in Abhängigkeit der Schrittweite machen?

3. Welche Aussagen kann man über die Genauigkeit der linearen im Vergleich zur bilinearen Approximation treffen?

8 Finite-Volumen-Verfahren

Finite-Volumen-Verfahren, kurz als **FV-Verfahren** bezeichnet, werden vor allem als numerische Approximation von hyperbolischen Erhaltungsgleichungen eingesetzt. Erhaltungsgleichungen wurden bereits in Kapitel 4.6 als eine wichtige Klasse von nichtlinearen hyperbolischen Gleichungen eingeführt. Sie haben die Form

$$u_t + \nabla \cdot \mathbf{f}(u) = 0 \, . \tag{8.1}$$

Dabei ist $u = u(\mathbf{x}, t)$ die Erhaltungsgröße und die vektorwertige Funktion $\mathbf{f}(u)$ der Fluss, $\mathbf{x}$ bezeichnet den Vektor der unabhängigen Raumvariablen.

Diese Differenzialgleichung ist aus einer integralen Gleichung, dem physikalischen Erhaltungssatz, als eine lokale Formulierung abgeleitet. Meist sind es mehrere Erhaltungsgleichungen, welche einen physikalischen Vorgang bestimmen. So führen die integralen Erhaltungssätze für Masse, Impuls und Energie auf die strömungsmechanischen Gleichungen, wie dies in Kapitel 4.7 dargestellt ist.

Betrachten wir die Formulierung eines einzelnen integralen Erhaltungssatzes noch einmal im Hinblick auf das numerische Verfahren. Der Erhaltungssatz sagt aus, dass für jedes Teilgebiet G des physikalischen Gebietes die Gleichung

$$\frac{d}{dt} \int\limits_G u \, dV = - \int\limits_{\partial G} \mathbf{f}(u) \cdot \mathbf{n} \, dS \tag{8.2}$$

gilt. Dabei ist $\mathbf{n}$ der nach außen gerichtete Normalenvektor mit der Länge eins (siehe Abbildung 8.1). Ist u die Massendichte, dann besagt obige Gleichung, dass sich die Masse in jedem Teilbereich des Strömungsgebietes nur dann ändert, wenn der Massenfluss durch den Rand des Teilbereichs von Null verschieden ist.

Das Finite-Volumen-Verfahren ist eine Approximation des integralen Erhaltungssatzes. Da sich die grundlegende Idee in zwei Raumdimensionen sehr gut skizzieren lässt, starten wir mit diesem Fall. Wir führen zunächst eine Diskretisierung des Rechengebiets ein. Dabei betrachten wir ein beliebiges Gitter. Der Rand einer Gitterzelle sei stückweise glatt. In Abbildung 8.2 ist

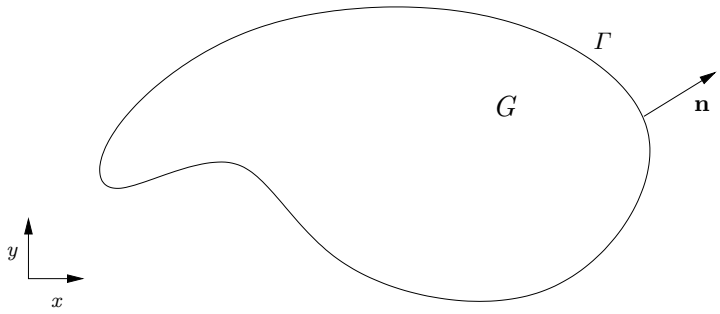


Abb. 8.1. Integraler Erhaltungssatz

in zwei Raumdimensionen ein Rechengebiet mit einem Gitter aus Dreiecken aufgezeichnet, wie es bei der Simulation einer Strömung um ein keilförmiges Hindernis eingesetzt wird. Das Rechengebiet ist hier zwar ein Rechteck, aber das Hindernis lässt die Verwendung eines kartesisches Gitters nicht zu. Das Gitterkonzept bei einem FV-Verfahren kann sehr allgemein sein. Es können Gitterzellen wie Vierecke, Fünfecke oder andere Polyeder verwendet werden. Die Vereinigung aller Gitterzellen muss allerdings das gesamte Rechengebiet überdecken und der Durchschnitt je zweier Gitterzellen muss sich auf den Rand der Gitterzellen beschränken.

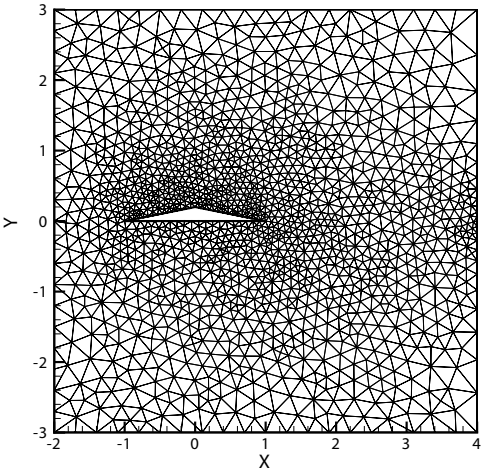


Abb. 8.2. Rechengebiet mit einem Dreiecksgitter

Formuliert man die integrale Erhaltungsgleichung (8.2) für eine beliebige Gitterzelle C_i, dann gilt

$$\frac{d}{dt}\int\limits_{C_i} u\, dV = -\int\limits_{\partial C_i} \mathbf{f}(u)\cdot\mathbf{n}\, dS\ .$$

Zur Abkürzung werden die integralen Mittelwerte zum Zeitpunkt t_n mit

$$u_i^n = \frac{1}{|C_i|}\int\limits_{\partial C_i} u(\mathbf{x}, t_n) d\mathbf{x}$$

bezeichnet und analog zum Zeitpunkt t_{n+1}. Die Erhaltungsgleichung wird dann bezüglich der Zeit t über das Intervall $[t_n, t_{n+1}]$ integriert, so dass man

$$u_i^{n+1} = u_i^n - \frac{1}{|C_i|}\int\limits_{t_n}^{t_{n+1}}\int\limits_{\partial C_i} \mathbf{f}(u)\cdot\mathbf{n}\, dS\, dt$$

erhält.

Diese Gleichung ist eine **Evolutionsgleichung für die integralen Mittelwerte** in den Gitterzellen. Sie ist in diesem Sinne exakt: Der integrale Mittelwert zum Zeitpunkt t_{n+1} ergibt sich aus dem Wert zum Zeitpunkt t_n durch das Berechnen des Integrals des Flusses durch den Rand der Gitterzelle. Nehmen wir an, dass der Rand der Gitterzelle C_i aus k_i Kanten K_j mit jeweils einer benachbarten Gitterzelle besteht, dann lässt sich die Evolutionsgleichung in der Form

$$u_i^{n+1} = u_i^n - \frac{1}{|C_i|}\int\limits_{t_n}^{t_{n+1}}\sum_{k=1}^{k_i}\int\limits_{K_k} \mathbf{f}(u)\cdot\mathbf{n}\, dS\, dt \tag{8.3}$$

schreiben.

Diese Evolutionsgleichung für die integralen Mittelwerte ist Ausgangspunkt für die Konstruktion eines FV-Verfahrens. Der wesentliche Baustein eines FV-Verfahrens ist der numerische Fluss, mit dem der Fluss durch den Rand einer Gitterzelle approximiert wird. Eine Schwierigkeit ist, dass bei dem FV-Verfahren mit integralen Mittelwerten gerechnet wird, für die Berechnung des Flusses durch den Rand benötigt man aber lokale Werte auf dem Rand. Das FV-Verfahren besteht somit aus zwei Schritten: Der **Rekonstruktion** von lokalen Werten auf dem Rand aus den integralen Mittelwerten und der **Flussberechnung**.

Wir betrachten zunächst als einfachsten Fall die **eindimensionale skalare Erhaltungsgleichung**:

$$u_t + f(u)_x = 0 \,. \tag{8.4}$$

Der Einfachheit halber führen wir hier eine äquidistante Diskretisierung mit den Schrittweiten Δt und Δx ein. Ein Gitterintervall im Raum sei durch

$$I_i = [x_{i-1/2}, x_{i+1/2}] \tag{8.5}$$

mit

$$x_i = i\Delta x \,, \qquad x_{i+1/2} = \frac{1}{2}(x_i + x_{i+1}) \,, \quad \text{für alle } i \in \mathbb{Z}$$

definiert. Wird die eindimensionale Erhaltungsgleichung (8.4) bezüglich x und t über $[x_{i-1/2}, x_{i+1/2}] \times [t_n, t_{n+1}]$ integriert, erhält man

$$\int_{t_n}^{t_{n+1}} \int_{x_{i-1/2}}^{x_{i+1/2}} u_t(x,t)\, dx\, dt + \int_{t_n}^{t_{n+1}} \int_{x_{i-1/2}}^{x_{i+1/2}} f(u(x,t))_x\, dx\, dt = 0 \,. \tag{8.6}$$

Die Integration nach t kann im ersten Integral und die Integration nach x im zweiten Integral ausgeführt werden:

$$\int_{x_{i-1/2}}^{x_{i+1/2}} u(x,t_{n+1})dx \; - \int_{x_{i-1/2}}^{x_{i+1/2}} u(x,t_n)dx + \int_{t_n}^{t_{n+1}} f(u(x_{i+1/2},t))dt \; - \int_{t_n}^{t_{n+1}} f(u(x_{i-1/2},t))dt = 0 \,.$$

Die Integration des Flusses im Raum nach dem Hauptsatz der Differenzial- und Integralrechnung entspricht dem Gaußschen Satzes in (4.53) für eine Raumdimension.

Mit den Abkürzungen

$$u_i^n := \frac{1}{\Delta x} \int_{x_{i-1/2}}^{x_{i+1/2}} u(x,t)\, dx \,, \qquad f_{i+1/2} := \frac{1}{\Delta t} \int_{t_n}^{t_{n+1}} f(u(x_{i+1/2},t))\, dt$$

erhält man die Evolutionsgleichung für die integralen Mittelwerte in der Form

$$u_i^{n+1} = u_i^n - \frac{\Delta t}{\Delta x} \left(f_{i+1/2} - f_{i-1/2} \right) \,. \tag{8.7}$$

Die Approximation dieser Integralgleichung

$$u_i^{n+1} = u_i^n - \frac{\Delta t}{\Delta x} \left(g_{i+1/2}^n - g_{i-1/2}^n \right)$$

nennt man das **eindimensionale Finite-Volumen-Verfahren** oder auch **Verfahren in Erhaltungsform**. Dabei stellt der Wert $g_{i+1/2}$ eine Approximation von $f_{i+1/2}$ dar, d.h. des integralen Mittelwerts des Flusses im Zeitintervall $[t_n, t_{n+1}]$ am Punkt $x_{i+1/2}$. Man nennt g den **numerischen Fluss**.

Der physikalische Fluss zwischen zwei Gitterzellen hängt von den Zuständen in den beiden Gitterzellen am gemeinsamen Rand ab. Bei dem Finite-Volumen-Verfahren werden integrale Mittelwerte berechnet, die lokalen Werte am Rand sind nicht bekannt. Die einfachste Annahme ist, dass der numerische Fluss durch die integralen Mittelwerte rechts und links bestimmt ist: $g = g(u_l, u_r)$. Man erhält somit die Flussberechnung

$$g^n_{i+1/2} = g(u^n_i, u^n_{i+1}) \; .$$

an der Stelle $x_{i+1/2}$ zwischen den den beiden Gitterintervallen I_i und I_{i+1}.

Man kann dies auch so formulieren, dass die Näherungslösung in jedem Gitterintervall konstant und damit gleich dem integralen Mittelwert ist. Man spricht von einer **stückweise konstanten Rekonstruktion** der Näherungslösung, konstant in jedem Gitterintervall.

An die numerische Flussfunktion g werden verschiedene Bedingungen gestellt. So muss g gewisse Stetigkeitsbedingungen erfüllen. Es zeigt sich bei der Herleitung geeigneter Flussfunktionen, dass einige nicht an allen Punkten stetig differenzierbar sind. Eine etwas schwächere Eigenschaft ist die Lipschitz-Stetigkeit. Die Lipschitz-Stetigkeit der Flussfunktion im Punkt (u, v) bedeutet, dass eine Konstante L existiert, so dass die Ungleichung

$$|g(u, v) - g(\overline{u}, \overline{v})| \leq L(|u - \overline{u}| + |v - \overline{v}|) \tag{8.8}$$

für alle $\overline{u}, \overline{v}$, die im Definitionsbereich von g liegen, erfüllt ist. Diese Eigenschaft garantiert noch, dass kleine Unterschiede in den Argumenten der numerischen Flussfunktion g dazu führen, dass auch die Differenz der Funktionswerte klein ist.

Neben der Stetigkeit muss der numerische Fluss g auch eine sinnvolle Approximation des physikalischen Flusses sein. Die Forderung, dass die Flussberechnung zumindest für konstante Lösungen exakt ist und mit dem physikalischen Fluss überein stimmt, führt auf die Konsistenzbedingung

$$g(u, u) = f(u) \; . \tag{8.9}$$

Zusammenfassung: Das numerische Verfahren

$$u_i^{n+1} = u_i^n - \frac{\Delta t}{\Delta x}\left(g_{i+1/2}^n - g_{i-1/2}^n\right) \tag{8.10}$$

mit einem Lipschitz-stetigen numerischen Fluss g, welcher die Konsistenzbedingung (8.9) erfüllt, heißt **Finite-Volumen-Verfahren**, kurz **FV-Verfahren** oder auch **Verfahren in Erhaltungsform** für die eindimensionale Erhaltungsgleichung (8.4).

Bemerkungen:

1. Der Näherungswert u_i^n ist im Falle eines FV-Verfahrens somit kein Näherungswert für die Lösung an einem speziellen Punkt (x_i, t_n), sondern der Näherungswert des integralen Mittelwerts der exakten Lösung im Gitterintervall I_i zum Zeitpunkt t_n. Dies ist ein wesentlicher Unterschied zu den Differenzen- und den FE-Verfahren.

2. Was wir hier betrachten sind explizite FV-Verfahren. Der numerische Fluss zur Berechnung der Näherungswerte zum Zeitpunkt t_{n+1} wird aus den Werten zum Zeitpunkt t_n berechnet.

3. Ein FV-Verfahren hat durch seine Konstruktion eine bemerkenswerte Eigenschaft: Die globale Erhaltung wird bis auf mögliche Rundungsfehler exakt reproduziert. Dies lässt sich einfach einsehen: Summieren wir die Definitionsgleichung eines FV-Verfahrens (8.9) über alle Gitterintervalle auf, dann erhalten wir

$$\sum_i u_i^{n+1} = \sum_i u_i^n - \frac{\Delta t}{\Delta x}\sum_i \left(g_{i+1/2}^n - g_{i-1/2}^n\right) \ .$$

   Die Summe auf der rechten Seite ist eine sogenannte Teleskop-Summe. Man erhält die Flüsse an der Grenze zweier benachbarter Gitterintervalle einmal positiv und einmal negativ, d.h. alle Flüsse zwischen den einzelnen Gitterzellen heben sich heraus. Dies entspricht der physikalischen Situation: Alles, was aus dem einen Gitterintervall herausfließt, fließt in das benachbarte hinein. Von der Summe bleiben lediglich die Flüsse am Rand des Rechengebietes wie im exakten Fall übrig.

Die Konstruktion der FV-Verfahren besteht nach (8.10) aus dem Finden einer geeigneten numerischen Flussfunktion g. Wir wollen im Folgenden mit der einfachsten Erhaltungsgleichung starten, geeignete Konstruktionskriterien und Ideen ableiten und diese dann auf kompliziertere Modelle übertragen.

8.1 Lineare Transportgleichungen

Als das einfachste Problem einer Erhaltungsgleichung betrachten wir den linearen Fall, die skalare lineare Transportgleichung mit einer konstanten Geschwindigkeit a. Der Fluss ist dann einfach durch $f(u) = au$ gegeben und die zugehörige Erhaltungsgleichung lautet

$$u_t + (au)_x = 0 \; . \tag{8.11}$$

Ein wichtiger Beitrag bei der Konstruktion von FV-Verfahren geht auf Godunov im Jahre 1959 zurück. Er führte die Annahme ein, dass die Näherungslösung stückweise konstant ist, konstant in jedem Gitterintervall:

$$u^n(x) := u_i^n \quad \text{für } x \in [x_{i-1/2}, x_{i+1/2}] \; . \tag{8.12}$$

Aus den integralen Werten u_i^n wird in einfacher Art und Weise eine kontinuierliche Funktion konstruiert. An der Grenze zwischen zwei Gitterzellen werden der Einfachheit halber dabei zwei Werte vorgeschrieben, den Grenzwert von links und den von rechts.

Wie wir in Abschnitt 4 gesehen haben, kann man mit Hilfe der Charakteristikentheorie die exakte Lösung des Anfangswertproblems angeben. Die Anfangswerte werden mit der Geschwindigkeit a transportiert. Wir nehmen als Anfangswerte

$$u(x,0) = u^n(x) \quad \text{für alle } x \; . \tag{8.13}$$

In Abbildung 8.3 ist die stückweise konstante Näherungslösung zu einem Zeitpunkt t_n aufgezeichnet, die exakte Lösung zur Zeit $t_n + \Delta t$ ist dann in Abbildung 8.4 zu sehen. Für die positive Geschwindigkeit a sind die Anfangswerte um $a\Delta t$ nach rechts verschoben.

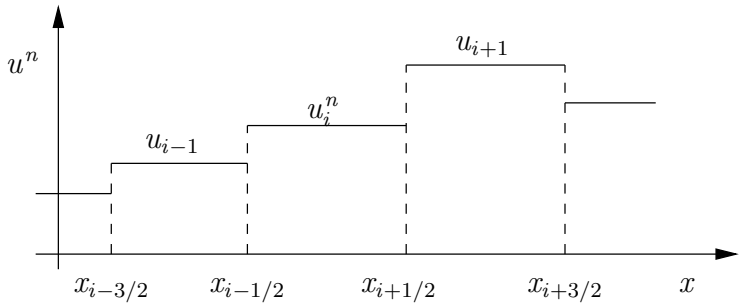


Abb. 8.3. Stückweise konstante Rekonstruktion der Lösung aus den integralen Mittelwerten zum Zeitpunkt t_n

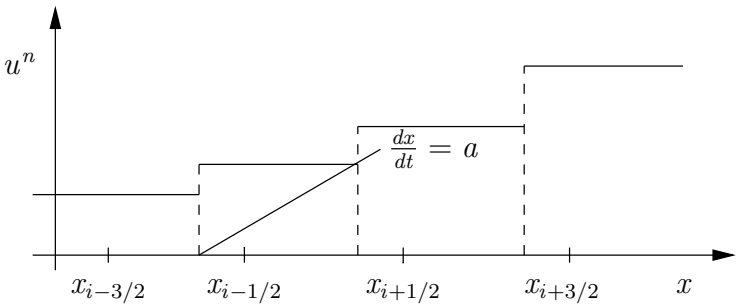


Abb. 8.4. Exakte Lösung des Anfangswertproblems (8.11), (8.13) zum Zeitpunkt t_{n+1}

Diese exakte Lösung des Anfangswertproblems (8.11), (8.13) zum Zeitpunkt Δt betrachtet man als eine Näherungslösung zum Zeitpunkt t_{n+1}. Um das ganze Verfahren zu wiederholen, müssen die integralen Mittelwerte dieser Lösung in den einzelnen Gitterintervallen, nun zum Zeitpunkt t_{n+1}, berechnet werden. Für $a > 0$ erhalten wir durch eine einfache Integration der stückweisen konstanten Lösung

$$\begin{aligned} u_i^{n+1} &= \frac{1}{\Delta x} \int_{x_{i-1/2}}^{x_{i+1/2}} u(x, \Delta t)\, dx \\ &= \frac{1}{\Delta x} \left(\int_{x_{i-1/2}}^{x_{i-1/2}+a\Delta t} u(x, \Delta t)\, dx + \int_{x_{i-1/2}+a\Delta t}^{x_{i+1/2}} u(x, \Delta t)\, dx \right) \\ &= \frac{1}{\Delta x} \left(a\Delta t\, u_{i-1}^n + (\Delta x - a\Delta t) u_i^n \right) \\ &= u_i^n - a\frac{\Delta t}{\Delta x}(u_i^n - u_{i-1}^n)\,. \end{aligned}$$

Dabei haben wir vorausgesetzt, dass der Zeitschritt so klein ist, dass die Werte nicht über mehr als eine Gitterzelle transportiert werden.

Diese Vorgehensweise war einleuchtend aber noch nicht ganz im Sinne eines Finiten-Volumen-Ansatzes. Wir haben die exakte Lösung des Anfangswertproblems berechnet und diese dann integriert. Man geht hier besser direkt von der Integralgleichung aus und setzt die exakte Lösung in die Flussberechnung ein. Werden die Anfangswerte wegen $a > 0$ nach rechts transportiert, dann hat die exakte Lösung des Anfangswertproblems am Punkte $x_{i+1/2}$ den Wert u_i, solange die betrachtete Zeit nicht zu groß wird. Der Fluss an diesem Punkt ist somit

$$g_{i+1}^n = a u_i^n$$

(siehe auch Abbildung 8.4).

Die zugehörige Flussfunktion ist somit durch

$$g(u_i^n, u_{i+1}^n) = a u_i^n$$

gegeben. Dieses Ergebnis kann leicht auf den Fall $a < 0$ übertragen werden.

In einer Formel zusammengefasst lautet die numerische Flussfunktion

$$g(u_i^n, u_{i+1}^n) = a^+ u_i^n + a^- u_i^n \tag{8.14}$$

mit

$$a^+ = \max(0, a) \,, \qquad a^- = \min(0, a) \,.$$

Setzt man dies in das FV-Verfahren (8.10) ein und formuliert etwas um, dann erhält man

$$u_i^{n+1} = u_i^n - a\frac{\Delta t}{\Delta x} \begin{cases} u_i^n - u_{i-1}^n & \text{für } a > 0 \\ u_{i+1}^n - u_i^n & \text{für } a \leq 0 \end{cases} . \tag{8.15}$$

Schlägt man das Kapitel 6 Differenzenverfahren auf, dann sieht man, dass diese Formel formal identisch mit dem CIR-Verfahren (6.57) ist.

Es stellt sich damit sofort die Frage: Wurde mit dem Finite-Volumen-Verfahren überhaupt etwas Neues gefunden? Die Antwort ist ja, mit der folgenden Begründung: In dem Differenzenverfahren ersetzt man die Ableitung durch einen rechts- oder linksseitigen Differenzenquotienten. Die Stabilitätsanalyse zeigt, welche der Möglichkeiten ein bedingt stabiles Verfahren ergibt. Im Finite-Volumen-Ansatz wird die Eigenschaft der exakten Lösung, die Richtung des Transports, ausgenutzt. Damit hat sich die entsprechende Flussformulierung als einzige mögliche Berechnung ergeben. Die CFL-Bedingung

$$|a|\frac{\Delta t}{\Delta x} < 1 \tag{8.16}$$

ergibt sich hier als Konsistenzbedingung der Flussberechnung und nicht als Stabilitätsbedingung. Ohne diese Bedingung ist die Herleitung des Flusses inkonsistent. Beim FV-Verfahren wurden zudem keinerlei Stetigkeitsvoraussetzungen gemacht. Diese Herleitung macht es möglich, das Verfahren auf nichtlineare Probleme zu verallgemeinern.

Die Herleitung der Berechnung des numerischen Flusses lässt sich noch etwas anders formulieren. Zwischen den Gitterzellen treten lokal Anfangswertprobleme mit stückweise konstanten Anfangswerten auf:

$$u(x, 0) = \begin{cases} u_l & \text{für } x < 0 \\ u_r & \text{für } x > 0 \end{cases} . \tag{8.17}$$

Etwa am Punkte $x_{i+1/2}$ die Werte $u_l = u_i^n$, $u_r = u_{i+1}^n$. Dies ist ein **Riemann-Problem**, wie dies im Kapitel 4.9 schon auftauchte.

Die Lösung dieses Anfangswertproblems lautet im linearen Fall

$$u(x,t) = \begin{cases} u_l & \text{für } \frac{x}{t} < a \\ u_r & \text{für } \frac{x}{t} > a \end{cases} . \tag{8.18}$$

Diese Lösung kann direkt zur Flussberechnung eingesetzt werden. Wir bezeichnen die Lösung des Riemann-Problems mit u_R. Sie hängt von x und t ab oder vielmehr nur von x/t und von den Anfangswerten u_l, u_r. Wir schreiben dafür

$$u_R = u_R\left(\frac{x}{t}; u_l, u_r\right) ,$$

x/t ist die unabhängige Variable und, mit dem Strichpunkt abgetrennt, u_l, u_r sind Parameter.

Damit hat man die folgende Situation: An jedem Rand zwischen zwei Gitterzellen wird das zugehörige Riemann-Problem mit den entsprechenden Anfangswerten gelöst. Die numerischen Flüsse ergeben sich dann zu:

$$g^n_{i+1/2} = g(u^n_i, u^n_{i+1}) = f(u_R(0; u_l, u_r)) = a\, u_R(0; u_l, u_r) . \tag{8.19}$$

Setzen wir die Lösungsformel (8.18) in diese Flussberechnung ein, so führt dies gerade auf den numerischen Fluss (8.14).

Beispiel 8.1 (FV-Verfahren für die lineare Transportgleichung, mit MAPLE**-Worksheet).** Mit dem Godunov-Verfahren wird das Anfangs-Randwertproblem für die lineare Transportgleichung

$$u_t + a\, u_x = 0 \tag{8.20}$$

mit den Anfangswerten

$$u(x,0) = x_0 + 0,1\, e^{-50(x-0,5)^2} \quad \text{für} \quad x \in [-0,5\,,\ 0,5] \tag{8.21}$$

und periodischen Randwerten gelöst. Die Vorgabe dieser Randwerte simuliert ein unendlich ausgedehntes periodisches Problem. Wir kommen auf die Vorgabe von allgemeineren Randwerten etwas später zurück.

Zum Ermitteln der Qualität der numerischen Lösung ist dies ein sehr gutes Beispiel. Ist die Geschwindigkeit $a = 1$ und die Länge des Intervalls 1, dann stimmt die exakte Lösung mit den Anfangsdaten nach $t = 1$ wieder überein. Man kann die numerischen Ergebnisse dann sehr gut mit der exakten Lösung vergleichen. In dem MAPLE-Worksheet werden numerische Ergebnisse für 50 Gitterpunkte erzeugt und sind in Abbildung 8.5 zu verschiedenen Zeiten als Funktion von x aufgezeichnet.

In dem Worksheet sind die Ergebnisse animiert. Man sieht wie die Anfangswerte (Bild a)) nach rechts verschoben (Bild b)), am rechten Rand verschluckt

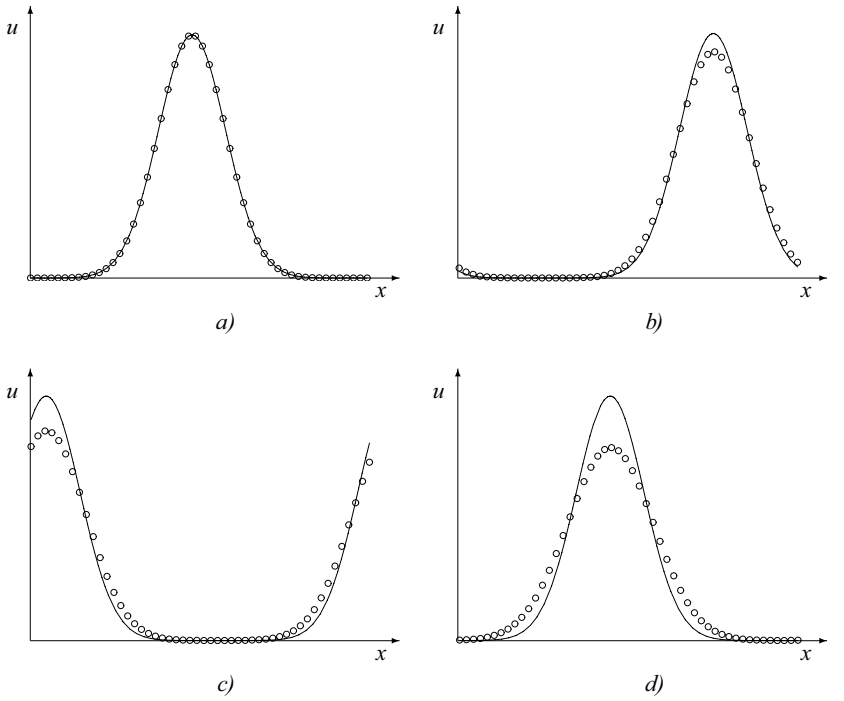


Abb. 8.5. Exakte und Näherungslösung des Anfangswertproblems in Beispiel 8.1 zu verschiedenen Zeiten

werden und am linken Rand durch die periodischen Randwerte wieder auftauchen (Bild c)). Nach dem Durchlauf einer Periode ist bei den numerischen Ergebnissen das Maximum etwas reduziert und die Gauß-Kurve verbreitert (Bild d)). Diese numerische Dämpfung nimmt mit fortlaufender Zeit zu. Erhöht man die Berechnungszeit auf $t = 2, 3, 4$, dann sieht man diesen Effekt in der Animation im Worksheet deutlich. Wie erwartet, wird die numerische Dämpfung reduziert und die numerischen Ergebnisse werden besser, wenn die Anzahl der Gitterpunkte erhöht wird. □

In dem Kapitel über die Differenzenverfahren wurde festgestellt, dass das CIR-Verfahren ein Verfahren mit der Genauigkeitsordnung 1 in Raum und Zeit ist. Dies ergibt sich aus der Tatsache, dass die benutzten einseitigen Differenzenquotienten, der rechts- oder linksseitige im Raum und der vorwärtsgenommene in der Zeit, von 1. Ordnung sind (siehe 1.6). Die 1. Ordnung leitet sich beim FV-Verfahren aus der Annahme ab, dass die Näherung stückweise konstant, konstant in jedem Gitterintervall ist. Man nennt dies die **stückweise konstante Rekonstruktion** der Lösung aus den integralen Näherungs-

werten. Möchte man eine höhere Ordnung des Finite-Volumen-Verfahrens erhalten, dann muss diese Rekonstruktion verbessert werden.

Eine **stückweise lineare Rekonstruktion** ergibt sich aus dem Ansatz

$$u^n(x) = u_i^n + (x - x_i)s_i^n \quad \text{für} \quad x \in \left[x_{i-1/2}, x_{i+1/2}\right] \ . \tag{8.22}$$

Dabei bezeichnet s_i^n die Steigung in der Gitterzelle $I_i = \left[x_{i-1/2}, x_{i+1/2}\right]$ zum Zeitpunkt t_n. Man benötigt die Randwerte an der Gitterzelle, um den numerischen Fluss zu berechnen. Diese erhält man aus (8.22) zu

$$u_{i\pm}^n = u_i^n \pm \Delta x \, s_i^n \ . \tag{8.23}$$

Auf die Berechnung der Steigungen in den Gitterintervallen werden wir gleich noch näher eingehen. Die Situation ist in Abbildung 8.6 aufgezeichnet. Die Werte der stückweisen linearen Verteilung an den Randpunkten des i-ten Gitterintervalls sind mit u_{i+}^n am Punkte $x_{i+1/2}$ und mit u_{i-}^n am Punkte $x_{i-1/2}$ bezeichnet.

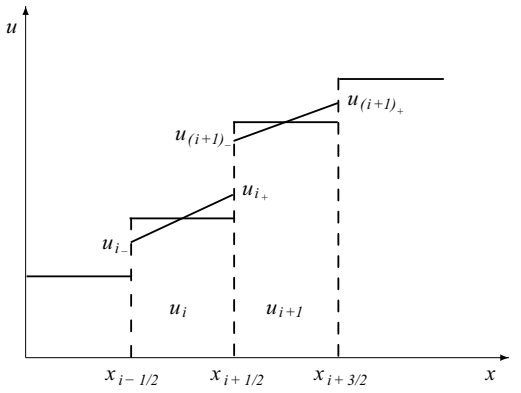


Abb. 8.6. Stückweise lineare Rekonstruktion der Lösung aus den integralen Mittelwerten zum Zeitpunkt t_n

Für die 2. Ordnung in der Zeit benötigt man eine Art zentrale Differenz. Wir gehen hier vor, wie bei dem verbesserten Euler-Cauchy-Verfahren. Die Taylor-Entwicklung in der Zeit für einen halben Zeitschritt liefert

$$u(x,\, t + \Delta t/2) = u(x,t) + \frac{\Delta t}{2} u_t(x,\, t) + \frac{\Delta t^2}{8} u_{tt}(x,\, t) + \ \ldots \ . \tag{8.24}$$

Die Zeitableitung lässt sich im numerischen Verfahren nicht so einfach bestimmen. Die Idee ist hier, die Zeitableitung durch eine Raumableitung zu

ersetzen, indem man die Evolutionsgleichung ausnutzt: $u_t = -au_x$. Man erhält damit

$$u(x,\, t + \Delta t/2) = u(x,t) - \frac{\Delta t}{2} a\, u_x(x,t) + O(\Delta t^2)\ . \tag{8.25}$$

Die Raumableitung zum Zeitpunkt t_n wird dann durch einen Differenzenquotienten ersetzt. Mit diesem Schritt von t_n nach $t_{n+1/2}$ erhält man aus (8.23) die Gleichung

$$u_{i\pm}^{n+1/2} = u_{i\pm}^{n} - \frac{\Delta t}{2}\left(f(u_{i+}^{n}) - f(u_{i-}^{n})\right)\ . \tag{8.26}$$

Für den rechten Rand ist der letzte Term in dieser Gleichung ein linksseitiger Differenzenquotient, für den linken Randpunkt ein rechtsseitiger.

Das Verfahren **2. Ordnung in Raum und Zeit** kann man dann in der Form

$$u_i^{n+1} = u_i^n - \frac{\Delta t}{\Delta x}\left(g_{i+1/2}^{n+1/2} - g_{i-1/2}^{n+1/2}\right) \tag{8.27}$$

mit

$$g_{i+1/2}^{n+1/2} = g\left(u_{i+}^{n+1/2}, u_{(i+1)-}^{n+1/2}\right) \tag{8.28}$$

angeben, wobei g ein Lipschitz-stetiger konsistenter numerischen Fluss ist.

Es bleibt noch zu zeigen, wie eine stückweise lineare Rekonstruktion berechnet wird. Zur Berechnung der Steigung s_i^n im i-ten Intervall stehen verschiedene Kandidaten zur Verfügung: der linksseitige, der rechtsseitige oder der zentrale Differenzenquotient. Allgemein erhält man eine geeignete Approximation durch

$$s_i^n = \alpha\, \frac{u_i^n - u_{i-1}^n}{\Delta x} + (1-\alpha)\, \frac{u_{i+1}^n - u_i^n}{\Delta x}\ , \tag{8.29}$$

mit einem $\alpha \in [0,1]$. Für $\alpha = 0.5$ ist dies eine Approximation 2. Ordnung, ansonsten eine Approximation 1. Ordnung der räumlichen Ableitung der Lösung u_x im i-ten Gitterintervall.

Es zeigt sich allerdings, dass bei einem beliebigen konstanten Wert von α die Stabilitätseigenschaften des Verfahrens beeinträchtigt werden. Man muss hier eine vorsichtige Interpolation anwenden, um Oszillationen und nicht monotones Verhalten zu vermeiden, besonders in den Bereichen, in denen sich die Lösung stark ändert. Eine solche Berechnung ist

$$s_i^n = \mathit{minmod}\left(\frac{u_i^n - u_{i-1}^n}{\Delta x}, \frac{u_{i+1}^n - u_i^n}{\Delta x}\right)\ , \tag{8.30}$$

wobei die *minmod*-Funktion folgendermaßen definiert ist:

$$minmod(a,b) = \begin{cases} a & \text{falls } |a| < |b|,\ ab > 0\ , \\ b & \text{falls } |a| \geq |b|,\ ab > 0\ , \\ 0 & \text{sonst}\ . \end{cases} \tag{8.31}$$

Diese Steigungsberechnung sucht sich den dem Betrag nach kleineren Differenzenquotienten heraus; wenn sich das Vorzeichen der Steigungen ändert, dann wird die Steigung auf Null gesetzt. Der letzte Fall tritt bei einem Hoch- oder Tiefpunkt auf. Steigung Null bedeutet hier, dass die Werte an den Extremstellen in der numerischen Lösung nicht zunehmen können, was einer Eigenschaft der exakten Lösung entspricht.

Beim Verfahren 2. Ordnung geht man somit folgendermaßen vor. Man berechnet zunächst die Steigungen in den einzelnen Gitterintervallen s_i^n mit der Rechenvorschrift (8.30) und (8.31). Danach lassen sich die Randwerte in den einzelnen Gitterintervallen nach (8.23) berechnen, die Änderung in einem halben Zeitschritt nach (8.26) müssen noch berücksichtigt werden. Damit erhält man an den Randpunkten $x_{i+1/2}$ von links und rechts genauere Argumente für die Flussberechnung. Die Flussberechnung erfolgt wie bei dem Verfahren 1. Ordnung. Der erhöhte Rechenaufwand des Verfahrens 2. Ordnung wird insbesondere bei kleinen Schrittweiten sehr schnell durch die besseren Ergebnisse ausgeglichen.

Bemerkung: Die Steigungsberechnung nach (8.30) und (8.31) garantiert, dass

$$\sum_i |u_i^n - u_{i-1}^n| \leq \sum_i |u_i^0 - u_{i-1}^0| \tag{8.32}$$

für alle n gilt. Die sogenannte Variation oder Totalvariation der Näherung wächst nicht an. Diese Eigenschaft besitzt auch die exakte Lösung. Man nennt ein solches numerisches Verfahren ein **TVD-Verfahren (Total-Variation Diminishing)**. Es gibt speziell für skalare Transport- und Erhaltungsgleichungen eine schöne mathematische Theorie dieser Verfahren, auf die wir im Folgenden nicht weiter eingehen. Wir verweisen auf die Bücher von Kröner [39] LeVeque [40].

Bei **linearen Systemen** können wir analog zu den Differenzenverfahren vorgehen und die Charakteristikentheorie bemühen. Das hyperbolische System von linearen Transportgleichungen

$$\mathbf{u}_t + \mathbf{A}\mathbf{u}_x = 0 \tag{8.33}$$

mit der $m \times m$-Matrix $\mathbf{A}$ wird auf die charakteristische Normalform transformiert. Die Gleichungen sind demnach entkoppelt und stellen m skalare Transportgleichungen dar. Jede dieser Gleichungen kann dann für sich numerisch mit dem FV-Verfahren für eine einzelne Transportgleichung gelöst

werden. Danach wird die Lösung wieder zurück transformiert. Diese Procedere wurde schon bei den Differenzenverfahren ausgeführt.

Der numerische Fluss lautet als Vektor geschrieben

$$\mathbf{g}(\mathbf{u}_i, \mathbf{u}_{i+1}) = \mathbf{A}^+\mathbf{u}_i + \mathbf{A}^-\mathbf{u}_{i+1} \tag{8.34}$$

mit

$$\mathbf{A}^+ := \mathbf{R}\mathbf{\Lambda}^+\mathbf{R}^{-1}\,, \qquad \mathbf{A}^- := \mathbf{R}\mathbf{\Lambda}^-\mathbf{R}^{-1}\,.$$

Die Matrix $\mathbf{\Lambda}$ ist wie in Abschnitt 4.1 die Diagonalmatrix mit den Eigenwerte als Koeffizienten auf der Diagonalen. Die Matrizen $\mathbf{\Lambda}^+$ und $\mathbf{\Lambda}^-$ bilden die Zerlegung dieser Diagonalmatrix in den nicht-negativen und nicht-positiven Anteil:

$$\mathbf{\Lambda}^+ := \begin{pmatrix} a_1^+ & & & \\ & a_2^+ & & \\ & & \ddots & \\ & & & a_m^+ \end{pmatrix}\,, \qquad \mathbf{\Lambda}^- := \begin{pmatrix} a_1^- & & & \\ & a_2^- & & \\ & & \ddots & \\ & & & a_m^- \end{pmatrix} \tag{8.35}$$

mit a_i^+ und a_i^- aus (8.14).

Die Konsistenzbedingung für den numerischen Fluss ist erfüllt, da aus

$$\mathbf{g}(\mathbf{u}, \mathbf{u}) = \mathbf{A}^+\mathbf{u} + \mathbf{A}^-\mathbf{u} = (\mathbf{A}^+ + \mathbf{A}^-)\mathbf{u}$$

mit Einsetzen der Definition von $\mathbf{A}^+$ und $\mathbf{A}^-$

$$\mathbf{g}(\mathbf{u}, \mathbf{u}) = \mathbf{R}(\mathbf{\Lambda}^+ + \mathbf{\Lambda}^-)\mathbf{R}^{-1}\mathbf{u} = \mathbf{R}\mathbf{\Lambda}\mathbf{R}^{-1}\mathbf{u} = \mathbf{A}\mathbf{u} = \mathbf{f}(\mathbf{u})\,.$$

folgt.

Bemerkung: Der numerische Fluss (8.34) kann auch durch das Lösen des Riemann-Problems für das lineare System erhalten werden. Die zweite Ordnung ergibt sich durch die lineare Rekonstruktion aus den integralen Mittelwerten wie im skalaren Fall.

8.2 Skalare Erhaltungsgleichungen

Lösungen von Erhaltungsgleichungen können Unstetigkeiten enthalten, wie dies in Kapitel 4.6 gezeigt wurde. Diese Lösungen sind keine Lösungen der Erhaltungsgleichung in der Formulierung als Differenzialgleichung, sondern der zu Grunde liegenden Integralgleichung. Für die numerische Approximation dieser sogenannten schwachen oder integralen Lösungen sind FV-Verfahren

sehr günstig, da diese eine direkte Approximation der integralen Formulierung des Erhaltungsgesetzes sind. Man rechnet mit Näherungen der integralen Mittelwerte und braucht keine Stetigkeitsvoraussetzungen an die Lösung. Wesentlich bei einem FV-Verfahren ist, dass eine gute Näherung für den Fluss zwischen den Gitterzellen gefunden wird.

Zur Berechnung des numerischen Flusses braucht man die Werte rechts und links des Randpunktes. Nehmen wir wieder an, dass die Näherungslösung in jeder Gitterzelle konstant ist, dann entsprechen die lokalen Werte an dem Randpunkt von links und rechts den integralen Mittelwerten: Die Näherungslösung ist stückweise konstant.

Die Idee von Godunov kann dann auf den Fall einer nichtlinearen Erhaltungsgleichung übertragen werden. Dazu muss das Riemann-Problem an den Randpunkten eines jeden Gitterintervalls gelöst werden. Dies liefert die Information, wie sich der Sprung zwischen den Werten von links und rechts in einzelne Wellen zerlegt. Der Wert des Flusses wird entsprechend dieser nichtlinearen Wellenausbreitung bestimmt.

Wir schauen uns im Folgenden zunächst die Lösung des **Riemann-Problems für die skalare Erhaltungsgleichung**

$$u_t + f(u)_x = 0 \,, \tag{8.36}$$

$$u(x,0) = \begin{cases} u_l & \text{für } x < 0 \\ u_r & \text{für } x > 0 \end{cases} \,. \tag{8.37}$$

an. Der physikalische Fluss $f = f(u)$ sei dabei zweimal stetig differenzierbar. Die Steigungen der Charakteristiken, welche durch

$$C: \quad x = x(t) \quad \text{mit} \quad \frac{dx(t)}{dt} = a(u) = f'(u) \tag{8.38}$$

definiert sind, sind lösungsabhängig. Man kann diese in der (x,t)-Ebene nur zeichnen, wenn man von den Anfangswerten ausgeht.

In Kapitel 4.6 werden die Eigenschaften der Lösungen von Erhaltungsgleichung diskutiert. Im Abschnitt 4.9 wird in zwei Beispielen die Lösung des Riemann-Problems für die Burgers-Gleichung

$$u_t + (\frac{1}{2}u^2)_x = 0 \tag{8.39}$$

berechnet. Diese Überlegungen lassen sich erweitern.

Für die Lösung des Riemann-Problems (8.36), (8.37) mit einer beliebigen **konvexen Flussfunktion** $f(u)$ gibt es die zwei Fälle:

1. Stoßwelle für $\mathbf{u_l > u_r}$: Wegen der Konvexität der Flussfunktion folgt aus $u_l > u_r$ direkt die Ungleichung $a(u_l) > a(u_r)$. Die Charakteristiken müssen sich somit schneiden. Die Lösung lautet in diesem Fall

$$u(x,t) = \begin{cases} u_l & \text{für } \frac{x}{t} < s \ , \\ u_r & \text{für } \frac{x}{t} > s \ . \end{cases} \tag{8.40}$$

Die Geschwindigkeit s der Stoßwelle ergibt sich aus der Rankine-Hugoniot Bedingung

$$s = \frac{f(u_r) - f(u_l)}{u_r - u_l} \ . \tag{8.41}$$

Ein Beispiel für ein Riemann-Problem, dessen Lösung eine solche Stoßwelle enthält, ist Beispiel 4.14 in Kapitel 4.

2. Verdünnungswelle für $\mathbf{u_l < u_r}$: Aus $u_l < u_r$ folgt wegen der Konvexität der Flussfunktion direkt $a(u_l) < a(u_r)$. Daraus ergibt sich ein Öffnen der Charakteristiken, das über einen Verdünnungsfächer verbunden wird. Die Lösung lautet

$$u(x,t) = \begin{cases} u_l & \text{für } \frac{x}{t} < a(u_l) \ , \\ a^{-1}(\frac{x}{t}) & \text{für } a(u_l) \le \frac{x}{t} \le a(u_r) \ , \\ u_r & \text{für } \frac{x}{t} > a(u_r) \ . \end{cases} \tag{8.42}$$

Wie in Abbildung 4.18 in der (x,t)-Ebene dargestellt, wird der Verdünnungsfächer begrenzt durch die Charakteristiken der beiden ungestörten Zustände rechts und links. Diese sind durch die Gleichungen $x/t = a(u_l)$ und $x/t = a(u_r)$ gegeben. Die Funktion $a^{-1} = a^{-1}(u)$ ist die Umkehrfunktion von $a(u)$. Diese existiert, da $f(u)$ konvex und die Ableitung $a(u)$ somit streng monoton wachsend ist. Man kann leicht nachrechnen, dass diese Lösung stetig ist. Ein Beispiel für ein Riemann-Problem, dessen Lösung eine Verdünnungswelle enthält, ist Beispiel 4.15 in Kapitel 4.

Bemerkung: Die Lösung des Riemann-Problems ist eine Funktion der Variablen x/t. Man nennt diese Lösung eine selbstähnliche Lösung, da sie nur von diesem Verhältnis abhängt. Anders formuliert: Man kann statt x und t die skalierten Variablen $m\,x$ und $m\,t$ mit einer Zahl m einführen, ohne dass sich die Lösung ändert. Führt man die unabhängige Variable $\phi = x/t$ ein und macht den Ansatz $u = u(\phi)$, so erhält man

$$u_t(\phi) = \frac{du}{d\phi}\frac{\partial \phi}{\partial t} = -\frac{x}{t^2}\frac{du}{d\phi} \ ,$$

$$u_x(\phi) = \frac{du}{d\phi}\frac{\partial \phi}{\partial x} = \frac{1}{t}\frac{du}{d\phi} \ .$$

Wird dies in die Erhaltungsgleichung eingesetzt, so ergibt sich

$$u_t + f(u)_x = u_t + a(u)u_x = -\frac{x}{t^2}\frac{du}{d\phi} + \frac{1}{t}a(u)\frac{du}{d\phi} = 0$$

und daraus nach kurzer Umformung

$$\frac{du}{d\phi}\left(a(u) - \phi\right) = 0 \; .$$

Diese Gleichung ist erfüllt, wenn der erste oder der zweite Faktor des Produkts Null ist. Das Verschwinden des ersten Faktors entspricht der Stoßwelle als einer stückweisen konstanten Lösung, das Verschwinden des zweiten Terms entspricht gerade der Verdünnungswelle: $u = a^{-1}(\phi)$. Diese selbstähnliche Lösung des Riemann-Problems schreibt man gerne in der Form $u_{RP} = u_{RP}(x/t; u_l, u_r)$. Die unabhängige Variable ist x/t, als Parameter werden die Konstanten der Anfangswerte aufgeführt.

Nach diesem Ausflug zu den exakten Lösungen von Erhaltungsgleichungen kommen wir wieder zur numerischen Lösung zurück. Der Grund für diese Betrachtungen war: Kennt man die Lösung des Riemann-Problems, so kann man das Godunov-Verfahren, welches wir im linearen Fall abgeleitet haben, direkt auf den nichtlinearen Fall übertragen. Der numerische Fluss des Godunov-Verfahrens $g = g(u_l, u_r)$ wird definiert als der entsprechende physikalische Fluss der Lösung des Riemann-Problems am Punkt $x = 0$: $f(u_{RP}(0; u_l, u_r))$.

Schiebt man den Nullpunkt des Koordinatensystems an den entsprechenden Randpunkt im Gitterintervall, so erhält man allgemein

$$g_{i+\frac{1}{2}} = g(u_i, u_{i+1}) = f(u_{RP}(0; u_i, u_{i+1})) \; . \tag{8.43}$$

Da wir die Riemann-Lösung kennen, können wir diese einsetzen und den Fluss in der folgenden Form

$$g_{i+1/2} = g(u_i, u_{i+1}) = \begin{cases} f(u_i) & \text{für } u_i > u_{i+1},\ s > 0 \; , \\ f(u_{i+1}) & \text{für } u_i > u_{i+1},\ s < 0 \; , \\ f(u_i) & \text{für } u_i < u_{i+1},\ a(u_i) > 0 \; , \\ f(u_{i+1}) & \text{für } u_i < u_{i+1},\ a(u_{i+1}) < 0 \; , \\ a^{-1}(0) & \text{sonst} \; . \end{cases} \tag{8.44}$$

schreiben.

Bei Systemen von Erhaltungsgleichungen ist die exakte Lösung des Riemann-Problems oft schwierig zu berechnen. Schon für eine skalare Erhalungsgleichung sieht der Fluss des Godunov-Verfahrens mit den fünf Fallunterscheidungen in (8.44) etwas kompliziert aus. Es taucht somit die Frage auf, ob das

Riemann-Problem tatsächlich exakt gelöst werden muss. Vielleicht reicht die näherungsweise Lösung der lokalen Riemann-Probleme, um einen brauchbaren numerischen Fluss zu finden.

Wird die exakte Lösung des Riemann-Problems durch eine Näherungslösung ersetzt, so wird das zugehörige Verfahren **Godunov-Typ-Verfahren** genannt. Wir bezeichnen im Folgenden eine solche Näherungslösung mit $w = w(x/t; u_l, u_r)$. Wie von der exakten Lösung erfüllt, wird angenommen, dass die Näherungslösung nur von dem Verhältnis x/t abhängt. Sollen sowohl eine Stoßwelle als auch eine Verdünnungswelle von der Näherungslösung erfasst werden, so kann eine Näherungslösung wie folgt aussehen:

$$w(\frac{x}{t}; u_r, u_l) = \begin{cases} u_l & \text{für } \frac{x}{t} < a_l \ , \\ u_{lr} & \text{für } a_l < \frac{x}{t} < a_r \ , \\ u_r & \text{für } \frac{x}{t} > a_r \ . \end{cases} \tag{8.45}$$

Dabei sind die Werte a_l und a_r eine Abschätzung der größten und der kleinsten Ausbreitungsgeschwindigkeit der Wellen.

Die integrale Erhaltung ist eine sehr wichtige Eigenschaft bei der Approximation von Stoßwellen. Die Forderung, dass die näherungsweise Riemann-Lösung (8.45) konsistent mit der integralen Erhaltung ist, führt auf die Bedingung

$$\int_{-\frac{\Delta x}{2}}^{\frac{\Delta x}{2}} w(\frac{x}{t}; u_l, u_r) dx = \frac{\Delta x}{2}(u_l + u_r) - \Delta t\, f(u_r) + \Delta t\, f(u_l) \ . \tag{8.46}$$

Auf der rechten Seite steht der Wert der exakten Lösung des Riemann-Problems, falls die Ungleichung

$$\frac{\Delta t}{\Delta x} max(|a_l|, |a_r|) < \frac{1}{2} \tag{8.47}$$

erfüllt ist. Die Bedingung (8.47) bedeutet, dass der Zeitschritt so klein sein muss, dass keine Welle aus dem Rechteck in Abbildung 8.7 nach links und rechts hinausläuft. Nach der Erhaltungsgleichung (8.7) berechnet sich der Wert zum Zeitpunkt Δt als Integral über x der Anfangswerte und den Integralen über t an den Stellen $\frac{\Delta x}{2}$ und $-\frac{\Delta x}{2}$. Da die Flüsse in diesem Zeitintervall wegen der Bedingung (8.47) konstant sind, ergibt sich die rechte Seite in (8.46).

Setzt man in die linke Seite die näherungsweise Riemann-Lösung (8.45) ein, so kann aus der Gleichung (8.46) der mittlere Wert u_{lr} berechnet werden.

Aufgelöst nach u_{lr} ergibt sich

$$u_{lr} = \frac{a_r u_r - a_l u_l - f(u_r) + f(u_l)}{a_r - a_l} \; . \tag{8.48}$$

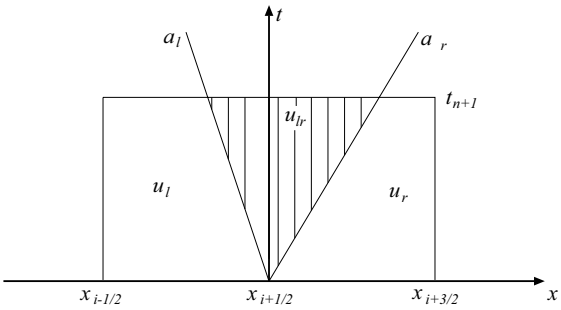


Abb. 8.7. Approximation der Lösung des Riemann-Problems

Der numerische Fluss wird nun so bestimmt, dass der Näherungswert zum neuen Zeitpunkt sich als integraler Mittelwert der näherungsweisen Riemann-Lösungen ergibt. Als FV-Verfahren geschrieben, führt dies auf den numerischen Fluss

$$g(u_l, u_r) = \frac{a_r^+ f(u_l) - a_l^- f(u_r)}{a_r^+ - a_l^-} + \frac{a_r^+ a_l^-}{a_r^+ - a_l^-}(u_r - u_l) \tag{8.49}$$

mit

$$a^+ = \max(0, a) \; , \qquad a^- = \min(0, a) \; .$$

Das FV-Verfahren ist somit vollständig definiert, falls noch eine geeignete Berechnung der Ausbreitungsgeschwindigkeiten a_l und a_r gefunden wird. Beim Riemann-Problem können zwei Fälle auftreten: Verdünnungswelle oder Stoßwelle. Im Falle einer **Verdünnungswelle** ist die größte und kleinste Ausbreitungsgeschwindigkeit durch die Geschwindigkeiten der ungestörten Zustände gegeben:

$$a_l = a(u_l) \; , \qquad a_r = a(u_r) \; .$$

Die Verdünnungswelle stellt einen stetigen Übergang dieser Zustände dar. Im Falle einer **Stoßwelle** ist dies keine gute Abschätzung, da sich hier die Charakteristiken schneiden und die Ausbreitungsgeschwindigkeit der Stoßwelle größer ist als die Geschwindigkeit vor dem Stoß. Entsprechend dem Falle einer Stoßwelle ergibt sich eine geeignete Geschwindigkeit aus der Rankine-Hugoniot-Bedingung

$$a_l = a_r := s = \begin{cases} \frac{f(u_r) - f(u_l)}{u_r - u_l} & \text{für } u_l \neq u_r \; , \\ a(u_r) & \text{für } u_l = u_r \; . \end{cases}$$

Diese beiden Fälle kann man in einer Formel zusammenfassen:

$$a_l = min(a(u_l), s) \ , \qquad a_r = max(a(u_r), s) \ . \tag{8.50}$$

Im Falle einer Verdünnungswelle sind die charakteristischen Geschwindigkeiten größer bzw. kleiner als s, so dass in (8.50) der richtige Wert genommen wird.

Es bleibt hier natürlich bestehen, dass die Definition des numerischen Flusses nur dann sinnvoll ist, wenn die benachbarten Riemann-Probleme nicht miteinander wechselwirken. Für das FV-Verfahren reicht es aus, dass eine mögliche Wechselwirkung nicht dazu führt, dass der Fluss sich innerhalb eines Zeitschritts ändert. Dies führt auf die CFL-Bedingung (8.16):

$$|a| \frac{\Delta t}{\Delta x} < 1 \ . \tag{8.51}$$

Diese Gleichung kann auch als eine untere und obere Schranke für die Wahl der Ausbreitungsgeschwindigkeiten a_l und a_r interpretiert werden:

$$a_l = -\frac{\Delta x}{\Delta t} \ , \qquad a_r = \frac{\Delta x}{\Delta t}. \tag{8.52}$$

Wird diese Wahl in den Fluss (8.49) eingesetzt, so erhält man

$$g(u_l, u_r) = \frac{1}{2}(f(u_l) + f(u_r)) + \frac{\Delta x}{2\Delta t}(u_r - u_l) \ . \tag{8.53}$$

Setzt man diesen Fluss in das FV-Verfahren ein, so kann diese Gleichung in die Form

$$u_i^{n+1} = u_i^n - \frac{\Delta t}{2\Delta x} \left(f(u_{i+1}^n) - f(u_{i-1}^n) \right) + \frac{1}{2} \left(u_{i+1}^n - 2u_i^n + u_{i-1}^n \right) \tag{8.54}$$

umgeschrieben werden.

Interpretieren wir die Werte u_i^n in dieser Gleichung nicht als Näherungswerte der integralen Mittelwerte, sondern als Näherungswerte des Wertes der Lösung an dem Punkt x_i, gewissermaßen als die Formel für ein Differenzenverfahren, dann taucht diese Gleichung im Kapitel 6 bei den Differenzenverfahren für hyperbolische Differenzialgleichungen 6.3 auf und entspricht dem **Verfahren von Lax-Friedrichs** (6.67). Im Rahmen der Differenzenverfahren war die Argumentation für die Stabilität des Verfahrens bei der Anwendung auf nichtlineare Probleme die, dass der hintere Term in (8.54) gerade einer Approximation von $\frac{\Delta x^2}{2\Delta t} u_{xx}$ entspricht. Dies ist ein parabolischer Term, den man physikalisch als Reibungs- oder Wärmeleitungsterm interpretieren kann. Damit wird künstlich Dissipation eingeführt, welche eine Stoßwelle über

mehrere Gitterpunkte verschmiert und dadurch eine stabile Approximation mit einem Differenzenverfahren gestattet.

Im Rahmen der Ableitung als FV-Verfahren zeigt sich dieses Verfahren als ein Godunov-Typ-Verfahren, wobei die Wahl der Abschätzung der Signalgeschwindigkeiten allerdings die schlechteste alle Möglichkeiten ist.

Bemerkungen:

1. Das Godunov-Typ-Verfahren mit dem numerischen Fluss (8.53) liefert stabile und die physikalisch sinnvolle Lösungen, wenn die Wellengeschwindigkeiten a_l und a_r kleiner oder gleich bzw. größer oder gleich der Wellengeschwindigkeiten der exakten Lösung des Riemann-Problems sind. Die Approximation des Riemann-Problems muss die exakte Lösung einschließen. Falls dies nicht erfüllt ist, können auch physikalisch unsinnige Lösungen auftreten.

2. Unterschätzt a_l oder überschätzt a_r die exakten Werte, dann bleibt das Verfahren stabil, allerdings nimmt die numerische Dissipation zu und die numerischen Ergebnisse werden schlechter. Der Grenzfall und damit die ungünstigste Flussberechnung ist die nach dem Verfahren von Lax und Friedrichs.

8.3 Systeme von Erhaltungsgleichungen

Die Erweiterung eines FV-Verfahrens auf Systeme kann man formal dadurch erreichen, dass in den entsprechenden Formeln die skalaren Größen durch Vektoren ersetzt werden. So ergibt sich das **FV-Verfahren für das System von Erhaltungsgleichungen**

$$\mathbf{u}_t + \mathbf{f}(\mathbf{u}_x) = \mathbf{0} \; . \tag{8.55}$$

in der Form

$$\boxed{\mathbf{u}_i^{n+1} = \mathbf{u}_i^n - \frac{\Delta t}{\Delta x}\left(\mathbf{g}_{i+1/2}^n - \mathbf{g}_{i-1/2}^n\right) \; .} \tag{8.56}$$

Dabei wird gefordert, dass der numerische Fluss $\mathbf{g}$ Lipschitz-stetig und die Konsistenzbedingung

$$\mathbf{g}(\mathbf{u}, \mathbf{u}) = \mathbf{f}(\mathbf{u}) \tag{8.57}$$

erfüllt ist.

Das Godunov-Verfahren lässt sich auf ein System von Erhaltungsgleichungen übertragen, wenn man das Riemann-Problem für dieses System lösen kann. Dies hängt somit sehr stark vom betrachteten System ab. Setzt man voraus, dass das System aus m Gleichungen streng hyperbolisch ist, so hat die Funktionalmatrix des Flusses m verschiedene Eigenwerte. Die Lösung des Riemann-Problems besteht aus $m + 1$ konstanten Zuständen, welche durch sogenannte einfache Wellen getrennt werden. Einfache Wellen sind Stoß- oder Verdünnungswellen oder eine Kontaktunstetigkeit, welche eine lineare Entartung einer Stoßwelle darstellt. Im letzteren Fall fällt die Ausbreitungskurve der Unstetigkeit wie bei einer linearen Gleichung mit einer Charakteristik zusammen.

Beispiel 8.2. Ein wichtiges System mit einer großen Anzahl von Anwendungen sind die Gleichungen der Gasdynamik, wie sie in Abschnitt 4.7 des Kapitel 4 abgeleitet und diskutiert werden. In einer Raumdimension sind sie durch ein System von Erhaltungsgleichungen (8.55) mit drei Gleichungen, den Erhaltungsgleichungen für Masse, Impuls und Energie gegeben. Dies würde dem mathematischen Modell für eine Rohrströmung, bei der sich die Variablen nur in einer Richtung ändern, entsprechen.

Der Vektor der Erhaltungsgrößen $\mathbf{u}$ und der zugehörige Fluss $\mathbf{f}(\mathbf{u})$ lauten:

$$\mathbf{u} = \begin{pmatrix} \rho \\ \rho v \\ e \end{pmatrix}, \qquad \mathbf{f}(\mathbf{u}) = \begin{pmatrix} \rho v \\ \rho v^2 + p \\ v(e+p) \end{pmatrix}.$$

Die Funktionalmatrix $\mathbf{A}(\mathbf{u})$ des Flusses hat dabei die drei reellen Eigenwerte

$$a_1 = v - c\,, \qquad a_2 = v\,, \qquad a_3 = v + c \tag{8.58}$$

mit der Schallgeschwindigkeit c. Die Eigenwerte entsprechen den verschiedenen Wellengeschwindigkeiten in einem Gas. Es gibt somit drei verschiedene Arten oder Familien von Charakteristiken. Die Lösung des Riemann-Problems für die Euler-Gleichungen besteht somit aus vier konstanten Zuständen, welche durch 3 einfache Wellen getrennt werden. Bei den Euler-Gleichungen ist die mittlere Welle eine Kontaktunstetigkeit. Eine typische Lösung ist in Abbildung 8.8 gezeichnet. Hier läuft eine Verdünnungswelle nach links und ein Verdichtungsstoß nach rechts, in der Mitte springt die Dichte an der Kontaktunstetigkeit.

Ist im linken und rechten Anfangszustand die Geschwingkeit Null, nennt man dieses Riemann-Problem auch ein Stoßwellenrohr-Problem. Das Stoßwellenrohr wird als eine experimentelle Einrichtung zur Erzeugung von Stoßwellen

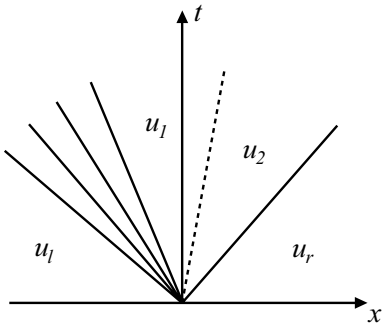


Abb. 8.8. Lösung des Riemannproblems

und Überschallströmungen eingesetzt. Es ist ein Rohr, welches innen durch eine Membran in zwei Teile getrennt wird. Ein Teil des Rohres wird mit Druckluft gefüllt, ist also somit ein komprimierter Zustand mit hohem Druck und hoher Dichte, der andere Teil mit Luft unter Normalbedingungen. Wird die Membran möglichst symmetrisch durchstoßen, dann läuft eine Stoßwelle von links nach rechts in den Teil mit niederem Druck und Dichte hinein. Sie wird von einer Kontaktunstetigkeit verfolgt, eine Verdünnungswelle läuft in den komprimierten Zustand. Hinter der Stoßwelle hat man bis zur Kontaktunstetigkeit die gewünschte konstante Überschallströmung.

Das Riemann-Problem spielt somit nicht nur bei dem Godunov-Verfahren und in der numerischen Simulation eine wichtige Rolle, sondern auch im experimentellen Bereich. Die exakte Lösung lässt sich in der Form der Lösung einer nichtlinearen Gleichung für den Druck des mittleren Zustandes angeben. Mehr Information findet man in Büchern über numerische Strömungsmechanik. □

Das Riemann-Problem ist meist nur mit viel Rechenaufwand iterativ exakt zu lösen. Bei einem FV-Verfahren werden auch nur Approximationen der integralen Mittelwerte berechnet und über die lokalen Lösungen der Riemann-Probleme integriert. Insofern wurde in den letzten Jahren untersucht, ob hier nicht eine Approximation der Riemann-Lösung ausreicht. Man sieht sofort, dass Godunov-Typ-Verfahren wie das HLL-Verfahren einfacher auf Systeme zu übertragen sind. Der numerische Fluss (8.49) lautet in diesem Fall

$$\mathbf{g}(\mathbf{u_l}, \mathbf{u_r}) = \frac{a_r^+ \mathbf{f}(\mathbf{u_l}) - a_l^- \mathbf{f}(\mathbf{u_l})}{a_r^+ - a_l^-} + \frac{a_r^+ a_l^-}{a_r^+ - a_l^-}(\mathbf{u_r} - \mathbf{u_l}) \,, \tag{8.59}$$

in dem lediglich der Fluss $f(u)$ und die Erhaltungsgröße u durch die entsprechenden Vektoren ersetzt wurde. Zusätzlich müssen allerdings geeignete

a priori Abschätzungen der kleinsten und größten Wellengeschwindigkeiten gefunden werden.

Der Ansatz (8.50) für eine solche Abschätzung lautet für ein System mit m Gleichungen

$$a_l = min(a_1(u_l), s_1), \qquad a_r = max(a_m(u_r), s_m). \tag{8.60}$$

Wie im skalaren Fall ist die charakteristische Geschwindigkeit in den ungestörten Zuständen bei einer Verdünnungswelle eine gute Abschätzung. Es wird im Systemfall der kleinste Eigenwert im linken Zustand bzw. der größte Eigenwert im rechten Zustand genommen. Die Werte s_1 und s_m bezeichnen Abschätzungen der Geschwindigkeit von Stoßwellen. Für deren a priori Abschätzung benötigt man allerdings mehr Kenntnisse über die Lösung. Wir kommen darauf zurück.

Das **Verfahren von Roe** ist ein Godunov-Typ-Verfahren, bei dem die exakte Lösung des Riemann-Problems der nichtlinearen Erhaltungsgleichung durch die exakte Lösung des Riemann-Problems einer linearisierten Erhaltungsgleichung ersetzt wird. Das lineare Riemann-Problem lautet

$$\mathbf{u}_t + \mathbf{A}_{lr}(\mathbf{u})_x = \mathbf{0}, \tag{8.61}$$

$$\mathbf{u}(x,0) = \begin{cases} \mathbf{u}_l & \text{für } x < 0 \\ \mathbf{u}_r & \text{für } x > 0 \end{cases}. \tag{8.62}$$

Dabei ist $\mathbf{A}_{lr} = \mathbf{A}(\mathbf{u}_l, \mathbf{u}_r)$ eine Matrix mit konstanten Koeffizienten, welche von den Anfangswerten abhängt. Man nennt $\mathbf{A}_{lr}$ ein **Roe-Matrix**, wenn sie die folgenden drei Eigenschaften

$$\mathbf{A}_{lr}(\mathbf{u}, \mathbf{u}) = \mathbf{A}(\mathbf{u}) \,, \tag{8.63}$$

$$\mathbf{A}_{lr} \quad \text{ist diagonalisierbar} \,, \tag{8.64}$$

$$\mathbf{f}(\mathbf{u}_r) - \mathbf{f}(\mathbf{u}_l) = \mathbf{A}_{lr}(\mathbf{u}_r) - \mathbf{u}_l \tag{8.65}$$

besitzt.

Die erste Eigenschaft ist die Konsistenz: Für gleiche Werte $\mathbf{u_l} = \mathbf{u_r} = \mathbf{u}$ stimmt die Roe-Matrix mit der Jacobi-Matrix, ausgewertet an dem konstanten Zustand u, überein. Die zweite Eigenschaft ist die Voraussetzung, dass das lineare Riemann-Problem mit Hilfe der Charakteristikentheorie gelöst werden kann. Die dritte Eigenschaft nennt man **Mittelwertseigenschaft**. Aus dieser folgt, dass die näherungsweise Riemann-Lösung die integrale Erhaltung exakt erfüllt und somit der integrale Wert mit dem der exakten Lösung des Riemann-Problems übereinstimmt.

Das heißt:

$$\int\limits_{-\frac{\Delta x}{2}}^{\frac{\Delta x}{2}} \mathbf{w}(\frac{x}{\Delta t};\mathbf{u}_l,\mathbf{u}_r)dx = \int\limits_{-\frac{\Delta x}{2}}^{\frac{\Delta x}{2}} \mathbf{u}_{RP}(\frac{x}{\Delta t};\mathbf{u}_l,\mathbf{u}_r)dx\ . \tag{8.66}$$

Dabei bezeichnet $\mathbf{w}$ die näherungsweise Lösung des Riemann-Problems und $\mathbf{u}_{RP}$ die exakte Lösung.

Die nichtlinearen Wellen des nichtlinearen Riemann-Problems werden durch lineare Wellen approximiert. Dabei besteht die Lösung des linearen Riemann-Problems ebenso aus $m+1$ konstanten Zuständen, welche in diesem Fall durch Charakteristiken getrennt werden. Das Roe-Verfahren wollen wir im Folgenden nicht weiter betrachten, sondern auf die Literatur verweisen. Wir werden die Roe-Mittelwerte benutzen, um die a priori Abschätzungen für das HLL-Verfahren zu erhalten.

Beispiel 8.3. Die Roe-Matrix für die Euler-Gleichungen mit der Zustandsgleichung des idealen Gases ergibt sich als Funktionalmatrix des Flusses an einem Mittelwert: $\mathbf{A}_{lr}(\mathbf{u}_l,\mathbf{u}_r) = \mathbf{A}(\bar{u})$, den man den Roe-Mittelwert nennt und der folgendermaßen lautet:

$$\bar{v} = \frac{v_r\sqrt{\rho_r}+v_l\sqrt{\rho_l}}{\sqrt{\rho_r}+\sqrt{\rho_l}}\,,\ \bar{H} = \frac{H_r\sqrt{\rho_r}+H_l\sqrt{\rho_l}}{\sqrt{\rho_r}+\sqrt{\rho_l}}\,,\ \bar{c}^2 = (\gamma-1)(\bar{H}-\frac{1}{2}\bar{v}^2)\,, \tag{8.67}$$

wobei H die Enthalpie mit $H = (e+p)/\rho$ bezeichnet. Die Eigenwerte der Roe-Matrix

$$\bar{a}_1 = \bar{v}-\bar{c}\,,\ \bar{a}_2 = \bar{v}\,,\ \bar{a}_3 = \bar{v}+\bar{c}\,, \tag{8.68}$$

sind gute Approximationen der Stoßwellengeschwindigkeiten, da die integrale Erhaltung von der näherungsweisen Riemann-Lösung reproduziert wird.

Wir können die Roe-Mittelwerte somit benutzen, um die a priori Abschätzungen (8.60) für das einfachste Godunov-Typ-Verfahren zu verbessern und setzen

$$s_1 = \bar{a}_1\,,\ s_3 = \bar{a}_3\,. \tag{8.69}$$

Damit ist der numerische Fluss vollständig definiert und das HLL-Verfahren für die eindimensionalen Euler-Gleichungen erklärt. □

Bemerkung: Die zweite Ordnung des Verfahrens lässt sich dann wieder über eine besserer Rekonstruktion erhalten. Mit einer stückweise linearen Rekonstruktion der Variablen wie im skalaren Fall erhält man die zweite Ordnung im Raum. Entsprechend ergibt sich die zweite Ordnung in der Zeit mit Hilfe einer Taylor-Entwicklung.

8.4 Erhaltungsgleichungen in mehreren Raumdimensionen

Das Finite-Volumen-Verfahren wurde am Anfang des Kapitels sehr allgemein eingeführt. Ausgangspunkt ist die integrale Erhaltung über eine Gitterzelle C_i:

$$u_i^{n+1} = u_i^n - \frac{1}{|C_i|} \int_{t_n}^{t_{n+1}} \sum_{k=1}^{k_i} \int_{K_k} \mathbf{f}(u) \cdot \mathbf{n} \, dS \, dt \; . \tag{8.70}$$

Dabei werden mit K_j in zwei Raumdimensionen die Kanten der Gitterzelle bezeichnet, von denen es k_i gibt. Für ein Gitter mit Dreiecken hat man $k_i = 3$ Kanten für jede Gitterzelle im gesamten Rechengebiet, d.h. für jedes i. In drei Raumdimensionen bezeichnen die K_j die Seitenflächen.

Der wesentliche Baustein in (8.70) zur Definition eines FV-Verfahrens ist wieder die Approximation des Flusses zwischen den Gitterzellen. In 2 Raumdimensionen hat man allerdings neben dem Integral über die Zeit ein Linienintegral oder in 3 Raumdimensionen ein Flächenintegral zur Bestimmung des Flusses zwischen den Gitterzellen. Wird dieses Integral über die Mittelwertsformel approximiert und wiederum vorausgesetzt, dass der Fluss in der Zeit im betrachteten Zeitintervall konstant ist, dann erhält man als Approximation

$$u_i^{n+1} = u_i^n - \frac{\Delta t}{|C_i|} \sum_{k=1}^{k_i} |K_k| (\mathbf{f}(u) \cdot \mathbf{n})_{MP} \; , \tag{8.71}$$

wobei $|K_k|$ die Länge der Kante K_k und der Index MP den Mittelpunkt dieser Kante bezeichnet.

In zwei Raumdimensionen besteht der Fluss und der Normalenvektor aus zwei Komponenten

$$\mathbf{f}(u) = \begin{pmatrix} f_1(u) \\ f_2(u) \end{pmatrix} \; , \qquad \mathbf{n} = \begin{pmatrix} n_1 \\ n_2 \end{pmatrix} \; .$$

Wird dies oben eingesetzt, so erhält man

$$u_i^{n+1} = u_i^n - \frac{\Delta t}{|C_i|} \sum_{k=1}^{k_i} |K_k| \, (n_1 \, f_1(u) + n_2 \, f_2(u))_{MP} \; . \tag{8.72}$$

Im Mittelpunkt der Kante benötigt man somit Approximationen des Flusses in x- und y-Richtung. Diese kann man aus der exakten oder näherungsweisen Lösung des Riemann-Problems mit den Zuständen vor und hinter der Kante erhalten. Hat man den Fluss des Riemann-Problems in x- und y-Richtung

berechnet, wird er entsprechend der Gleichung (8.72) mit n_1 und n_2 multipliziert und addiert.

Bei einem kartesischen Gitter reduziert sich dies auf die Berechnung eines einzigen Riemann-Problems pro Kante, da bei den Normalenvektoren eine der Komponenten immer 0 ist. Man kann in diesem Fall das FV-Verfahren für ein System von Erhaltungsgleichungen in der Form

$$u_{i,j}^{n+1} = u_{i,j}^{n} - \frac{\Delta t}{\Delta x}\left(g_{i+1/2,j}^{n} - g_{i-1/2,j}^{n}\right) - \frac{\Delta t}{\Delta y}\left(h_{i,j+1/2}^{n} - h_{i,j-1/2}^{n}\right) \quad (8.73)$$

schreiben. Dabei ist g der numerische Fluss, welcher konsistent mit dem Fluss f_1 in x-Richtung und h der numerische Fluss, welcher konsistent mit f_2 in y-Richtung ist.

Ist das System von Erhaltungsgleichungen rotationsinvariant, dann kann auch auf einem beliebigen Gitter die Lösung zweier Riemann-Probleme umgangen werden, indem man lokal auf ein Koordinatensystem mit den Richtungen normal und tangential zur Kante transformiert. Der Fluss in Normalenrichtung ergibt sich dann als exakte oder näherungsweise Lösung des Riemann-Problems in Normalenrichtung. Ein rotationsinvariantes System sind z.B. die Euler-Gleichungen.

8.5 Bemerkungen und Entscheidungshilfen

Finite-Volumen-Verfahren haben sich speziell für die numerische Lösung von kompressiblen Strömungen als sehr vorteilhaft erwiesen und sind die Standard-Verfahren geworden. Dies liegt daran, dass diese Verfahren in der Lage sind, die in der Lösung auftretenden starken Gradienten bis hin zu Stoßwellen robust zu erfassen.

Zuerst einige allgemeine Bemerkungen zu der Thematik starker Gradienten. Möchte man eine Lösung approximieren, welche sich in einem kleinen Bereich sehr stark ändert, dann kann dies im Prinzip auch mit einem Differenzen- oder mit einem FE- Verfahren erfolgen. Voraussetzung hierfür ist, dass genügend Gitterpunkte für die Approximation dieses Gradienten zur Verfügung stehen. Mit einem Differenzenverfahren braucht man hier vielleicht 20 Gitterpunkte in eine Raumrichtung. In drei Raumdimensionen kann dies wegen dem großen Rechenaufwand ineffizient werden. Mit weniger Gitterpunkten wird das Verfahren Oszillationen an diesem Gradienten zeigen und kann instabil werden. Es ist nicht in der Lage mit vielleicht nur 5 Gitterpunkten die Lösung sinnvoll zu approximieren. Hier hat das FV-Verfahren, welches im Allgemeinen mehr Rechenaufwand pro Gitterpunkt erfordert, Vorteile. Dadurch, dass

ein FV-Verfahren die integrale Erhaltungsgleichung löst und keinerlei Stetigkeitsaussagen voraussetzt, ist die Approximation großer Gradienten auf einigen wenigen Gitterzellen möglich.

Diskutieren wir die Situation bei einem Verdichtungsstoß. Dies ist eine Unstetigkeit in der Lösung der makroskopischen Erhaltungsgleichung. Ein Verfahren, das eine mindestens stetige Lösung voraussetzt wie ein Differenzen- oder FE-Verfahren, hat damit grundsätzliche Schwierigkeiten. Es bleibt nur die Möglichkeit, diese Unstetigkeit durch einen stetigen Übergang zu ersetzen, dass das Verfahren diesen steilen Gradienten dann approximieren kann. Dies kann über die Hinzunahme eines künstlichen Viskositätsterms erfolgen oder der Approximationsfehler liefert eine solche starke Dissipation. Hier sind die Finite-Volumen-Verfahren klar überlegen. Sie erfüllen automatisch die integrale Erhaltung und liefern dadurch die richtige Ausbreitungsgeschwindigkeit von Stoßwellen. Bei der Flussberechnung etwa durch ein Godunov-Typ-Verfahren wird dann zusätzlich Information über die nichtlineare Wellenausbreitung mit verwendet. Dies ergibt die Möglichkeit, auch Unstetigkeiten über wenige Gitterzellen zu approximieren. Im Allgemeinen reichen dann vier Gitterzellen aus, um die Stoßwelle darzustellen.

Wir haben hier nur explizite FV-Verfahren behandelt. Immer dann, wenn es um die Approximation von instationären Lösungen geht, ist die CFL-Bedingung eine natürliche Bedingung für die Genauigkeit in der Zeit. Die Situation ändert sich, wenn man nur an stationären Lösungen interessiert ist. Man kann hier ebenso die instationären Gleichungen numerisch lösen bis der stationäre Zustand erreicht wird. Dabei werden im Laufe der Rechnung die Zeitableitungen sehr klein bis letztlich Null, so dass der Fehler in der Zeitapproximation keine Rolle spielt. Sinnvoll zur Berechnung einer stationären Lösung ist dann eine implizite Approximation 1. Ordnung, um die Zeitschritte groß wählen zu können und die Gesamtzeit schnell groß werden zu lassen, so dass der stationäre Grenzwert erhalten wird.

Ein implizites Verfahren sieht dann in einer Raumdimension für eine skalare Erhaltungsgleichung wie folgt aus:

$$u_i^{n+1} = u_i^n - \frac{\Delta t}{\Delta x}\left(g_{i+1/2}^{n+1} - g_{i-1/2}^{n+1}\right)$$

mit dem numerischen Fluss

$$g_{i+1/2}^{n+1} = g(u_i^{n+1}, u_{i+1}^{n+1}) \; .$$

Damit ergibt sich im impliziten Fall ein nichtlineares Gleichungssystem, welches mit geeigneten numerischen Methoden gelöst werden muss. Ein Iterationsverfahren führt dann auf die Lösung eines schwach besetzten linearen Gleichungssystems, welches mit einer der Methoden im Anhang gelöst werden kann.

Als weiterführende Literatur sei auf die Bücher von Kröner [39] und LeVeque [40] verwiesen. Hier finden sich auch theoretische Aussagen über FV-Verfahren für Erhaltungsgleichungen. Im Bereich der numerischen Strömungsmechanik haben sich die Bücher von Hirsch [33] und Toro [65] als Standardwerke etabliert.

8.6 Beispiele und Aufgaben

Beispiel 8.4 (AWP für die lineare Transportgleichung, mit MAPLE-Worksheet). In diesem Beispiel suchen wir die Lösung eines Anfangswertproblems für die lineare Transportgleichung

$$u_t + a u_x = 0 \tag{8.74}$$

mit den Anfangswerten

$$u(x,0) = \begin{cases} 0,1 & \text{für} \quad 0 \le x \le 0,1\,, \\ 0 & \text{sonst}\,. \end{cases} \tag{8.75}$$

Die exakte Lösung dieses Problems wurde in Gleichung (4.43) angegeben: Die Anfangswerte werden mit der Geschwindigkeit a transportiert. An Hand dieses Testproblems kann man sehr gut die Eigenschaften eines numerischen Verfahrens untersuchen. Wir wenden das Finite-Volumen-Verfahren (8.15)

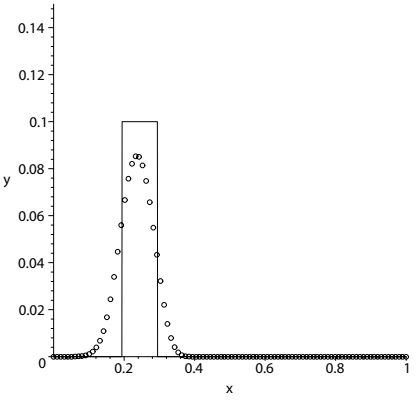


Abb. 8.9. Numerische und exakte Lösung für die linerare Transportgleichung mit einem Rechteckimpuls als Anfangsbedingung

auf dieses Beispiel an, bei dem der numerische Fluss aus der Lösung der

lokalen Riemannprobleme abgeleitet ist. In Abbildung 8.9 sieht man sehr deutlich, wie das Verfahren große numerische Dissipation besitzt. Nach einigen Zeitschritten ist das Rechteck zu einer Art Gauß-Kurve verschmiert. Auf der anderen Seite bleibt es stabil auch bei Vorgabe von diesen unstetigen Anfangswerten.

Für die numerische Rechnung wurde Rechengebiet auf ein endliches Intervall eingeschränkt. Am linken Rand wurde der Wert 0 vorgegeben, am rechten Rand ist keine Vorgabe eines Randwertes nötig. □

Beispiel 8.5 (AWPe für die Burgers-Gleichung, mit MAPLE-Worksheet). In diesem Beispiel wenden wir uns der Lösung einer nichtlinearen skalaren Erhaltungsgleichung, der Burgers-Gleichung

$$u_t + uu_x = 0\,, \tag{8.76}$$

zu. Wir betrachten zunächst das Anfangswertproblem mit den Anfangsdaten

$$u(x,0) = \begin{cases} 2 & \text{für} \quad x < 0,5\,, \\ 0 & \text{für} \quad x > 0,5\,. \end{cases} \tag{8.77}$$

Die exakte Lösung enthält eine Stoßwelle und berechnet sich wie in Gleichung (8.40). Die Stoßwelle läuft mit der Geschwindigkeit $s = 2$ nach rechts. Ein ähnliches Anfangswertproblem wurde schon als Beispiel 4.14 behandelt. Mit dem MAPLE-Worksheet wird eine numerische Lösung mit Hilfe des Godunov-Verfahrens im Vergleich zu der analytische Lösung berechnet. Die Ergebnisse sind in Abbildung 8.10 dargestellt. Bemerkenswert ist, dass der Verdichtungsstoß hier sehr gut aufgelöst ist. Man sieht eine gewisse numerische Dissipation, die Stoßwelle ist über fünf Gitterintervalle verschmiert, es gibt aber keinerlei Oszillationen an dieser Unstetigkeit, die sich insbesondere mit der richtigen Geschwindigkeit ausbreitet.

In Beispiel 8.4 war die Dissipation der Lösung durch das numerische Verfahren deutlich stärker ausgebildet. In der Abbildung 8.10 zeigt sich, dass die Stoßwelle im Vergleich hierzu sehr scharf aufgelöst wird. In beiden Fällen wurde das Godunov-Verfahren eingesetzt. Der Unterschied liegt darin, dass es sich hier um zwei unterschiedliche Phänomene handelt. In Beispiel 8.4 wird ein reiner Transportvorgang behandelt. Bei der linearen Transportgleichung sind alle Charakteristiken parallel und die Unstetigkeit breitet sich entlang einer Charakteristik aus. Bei dem nichtlinearen Phänomen einer Stoßwelle schneiden sich die Charakteristiken an der Stoßwelle. Hier ergibt sich ein Fluss in den Verdichtungsstoß hinein. Praktisch wirkt das physikalische Verhalten der numerischen Dissipation entgegen. Eine Stoßwelle wird somit sehr gut aufgelöst, wobei die Dissipation mit der Zeit nicht oder nur sehr wenig zunimmt.

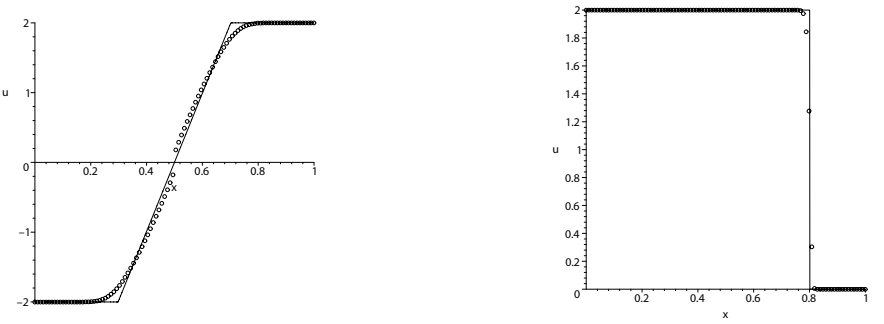


Abb. 8.10. Burgergleichung mit Verdünnung (links) und Verdichtungsstoß (rechts)

Die Anfangsbedingung

$$u(x,0) = \begin{cases} -2 \text{ für} & x < 0,5 \,, \\ 2 \quad \text{für} & x > 0,5 \,. \end{cases} \tag{8.78}$$

führt auf eine Verdünnungswelle. Betrachtet man die numerische Lösung in Abbildung 8.10, so zeigt sich eine gute Approximation auch in diesem Fall. Es fällt auf, dass die numerische Lösung in der Mitte der Verdünnung einen kleinen Sprung aufweist. Man spricht hier von dem sogenannten "Sonic Glitch". Als „Sonic Point" oder Schallpunkt wird allgemein der Nulldurchgang der Wellengeschwindigkeit bezeichnet in Anlehnung an die Bezeichnung bei den Euler-Gleichungen mit den Eigenwerten $u - c$ und $u + c$. Im Schallpunkt ist die Geschwindigkeit u gleich der Schallgeschwindigkeit c. Bei der Burgers-Gleichung entspricht dies dem Fall $u = 0$. An diesem Punkt wird die numerische Dissipation sehr klein und es kommt zu diesem kleinen Sprung.

Mit der Stoßwelle und der Verdünnungswelle haben wir die beiden grundlegenden Beispiele von nichtlinearen Wellen behandelt. □

Beispiel 8.6 (Verkehrsmodell,mit MAPLE**-Worksheet).** Wir greifen hier nochmal das einfache mathematisches Modell für die Berechnung des Verkehrsflusses auf einer Autobahn ohne Ein- und Ausfahrten auf, welches in Kapitel 4.9 behandelt wurde. Der Verkehr wird hier als Kontinuum betrachtet, für die Verkehrsdichte ρ ergibt sich die Gleichung

$$\rho_t + f(\rho)_x = 0 \tag{8.79}$$

mit der Flussfunktion

$$f(\rho) = -\frac{v_{max}}{\rho_{max}}\rho^2 + v_{max}\rho \,.$$

In diesem Fall ist die Flussfunktion nicht konvex, sondern konkav, da die zweite Ableitung von $f(\rho)$ nach ρ negativ ist. Die Aussagen für eine Erhaltungsgleichung mit einem konvexen Fluss müssen entsprechend übertragen werden. Betrachtet man die Lösung des Riemannproblems mit den Anfangswerten

$$\rho(x,0) = \begin{cases} \rho_l & \text{für } x < 0\,, \\ \rho_r & \text{für } x > 0 \end{cases} \tag{8.80}$$

so sind die Fälle Stoß- und Verdünnungswelle gerade vertauscht. Gilt $\rho_l < \rho_r$, so folgt der dünnere Verkehr dem dichten Verkehr. Da der Fluss eine konkave Funktion ist, ergibt sich daraus $a(\rho_l) > a(\rho_r)$, bei geringem Verkehr ist die Durchschnittsgeschwindigkeit größer und die Charakteristiken schneiden sich. Es ergibt sich hier ein Verdichtungsstoß, dessen Ausbreitungsgeschwindigkeit sich aus den Sprungbedingungen berechnet. Dies lässt sich in dem MAPLE-Worksheet leicht nachrechnen. □

Beispiel 8.7 (Eindimensionale Gasdynamik - die Euler-Gleichungen). Die Gleichungen der eindimensionalen Gasdynamik (4.74) wurden bereits in Kapitel 4 hergeleitet. Da eine Implementierung eines numerischen Verfahrens für diese Gleichungen den Rahmen von MAPLE sprengen würde, sind hier lediglich Ergebnisse angegeben, die mit einem Finite-Volumen-Verfahren errechnet wurden. Die verwendete Flussfunktion ist jeweils das HLLE-Verfahren. Die Anfangsbedingung der Rechnung ist das folgende Riemannproblem:

$$\rho(x,0) = \begin{cases} \rho = 1.0,\ v = 0.0,\ p = 1.0 & \text{für } x < 0\,, \\ \rho = 0.125,\ v = 0.0,\ p = 0.1 & \text{für } x > 0\,. \end{cases} \tag{8.81}$$

Die erste Abbildung (8.11) zeigt die Ergebnisse einer Rechnung mit einem Verfahren erster Ordnung in Raum und Zeit. Im Diagramm der Dichte ρ und der Machzahl Ma sind alle Wellenphänome, links die Verdünnung, in der Mitte die Kontaktunstetigkeit und rechts der Stoß, zu erkennen. Bei der Geschwindigkeit u und beim Druck p taucht in der Mitte kein Sprung auf, da diese Größen über eine Kontaktunstetigkeit konstant sind. Man sieht recht deutlich, dass der Stoß sehr gut aufgelöst ist, während die Kontaktunstetigkeit deutlich stärker verschmiert wird. Dies ist ganz analog zu den Ergebnissen für die Burgers-Gleichung und der linearen Transportgleichung. Zwischen dem Stoß und der Verdünnung ist die Geschwindigkeit konstant, die Ausbreitungskurve der Unstetigkeit ist eine Charakteristik, so dass hier die gleiche Situation wie bei der linearen Transportgleichung vorliegt.

Betrachten wir nun die Lösung des gleichen Problems (Abbildung 8.12) mit einem MUSCL-Verfahren zweiter Ordnung in Raum und Zeit aber der selben

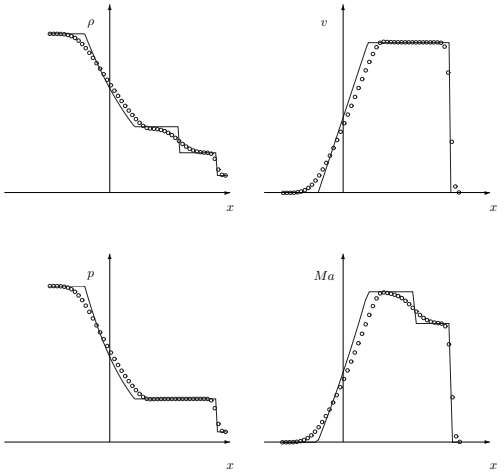


Abb. 8.11. SOD-Problem 1. Ordnung in Raum und Zeit

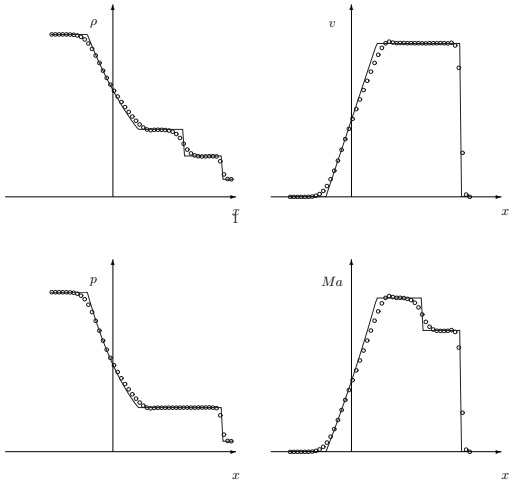


Abb. 8.12. SOD-Problem 2. Ordnung in Raum und Zeit

Flussberechnung, so stellen wir fest, dass die Auflösung der Kontaktunstetigkeit und der Expansion deutlich verbessert ist. Der Stoß wurde schon mit dem Verfahren erster Ordnung sehr gut wiedergegeben und ist augenscheinlich fast gleich geblieben. □

A Interpolation

Von einer Funktion ist oft nur eine gewisse Anzahl von Funktionswerten bekannt. Zum Beispiel bei der Konstruktion eines Tragflügels eines Flugzeuges hat man für das Tragflügelprofil die Werte in Tabelle A.1 zur Verfügung.

NACA 1412

Oberseite		Unterseite	
x	y	x	y
0	0	0	0
1,158	1,954	1,342	−1,830
2,378	2,733	2,622	−2,491
4,845	3,786	5,155	−3,318
7,330	4,537	7,670	−3,857
14,833	5,951	15,617	−4,733
40,000	6,803	40,000	−4,803
60,051	5,453	59,949	−3,675
80,058	3,178	79,942	−2,066
100,000	0,126	100,000	0,126

Tabelle A.1. Stützstellen für die Interpolation

Möchte man eine Konstruktionszeichnung machen, dann benötigt man eine kontinuierliche Kurve, welche durch die entsprechenden Punkte geht. Ganz allgemein besteht die **Interpolation** darin, eine einfache Funktion zu bestimmen, die eine gewisse Anzahl von vorgegebenen Werten annimmt und mit der dann beliebige Zwischenwerte berechnet werden können. Dies spielt in CAD-Systemen eine große Rolle. Eine andere Anwendung ist die Darstellung von Ergebnissen aus numerischen Simulationen, welche nur an den diskreten Gitterpunkten vorliegen. Die Interpolation wird auch zur Reduktion großer Datenmengen eingesetzt, wie sie etwa bei der Bildverarbeitung auftreten. Um die Daten zu reduzieren, werden nicht alle Farbwerte eines Bildes abgespeichert. Hat man eine große Fläche mit geringen Farbkontrasten, kann man die Farbwerte an einigen wenigen Punkten abspeichern und bei Bedarf die Werte an den Zwischenpunkten durch Interpolation wieder erzeugen. Ein Computerspiel kommt ohne Interpolation nicht aus.

Eng verwandt mit der Interpolation ist **Approximation** einer Funktion. Es wird hier eine einfache Funktion gesucht, welche eine vorgegebene im Allgemeinen komplizierte Funktion approximiert. Dies wird zum Beispiel dann angewandt, wenn die Auswertung einer vorgegebenen Funktion sehr rechenaufwändig ist. Die Approximation ist aber auch wichtig zur Ableitung von Näherungsverfahren, da sie meist einfach integriert und differenziert werden kann. Es können dadurch Verfahren zur numerischen Differenziation und Integration abgeleitet werden.

Eine der Kandidaten von einfachen Funktionen für die Interpolation oder Approximation ist die Klasse der **Polynome**. Integration oder Differenziation liefern durch einfache Rechnung wieder ein Polynom. Ein Polynom n-ten Grades schreibt man meist in der Form

$$p_n(x) = \sum_{i=0}^{n} a_i x^i = a_0 + a_1 x + \ldots + a_n x^n \; . \tag{A.1}$$

Von diesem Polynom n-ten Grades mit den $(n+1)$ Koeffizienten a_i, $i = 0, 1, \ldots, n$, können wir fordern, dass es $n+1$ vorgegebenen Werte annimmt:

$$p_n(x_i) = f_i, \tag{A.2}$$

$$i = 0, 1, \ldots, n \; . \tag{A.3}$$

Man nennt die x_i die **Stützstellen** und die f_i die zugehörigen **Stützwerte**, beides zusammen (x_i, f_i) sind die **Stützpaare**.

Bei der Approximation einer Funktion $f = f(x)$ ergeben sich die Stützwerte als Funktionswerte an den Stützstellen $f_i = f(x_i)$. Bei der Interpolation sind sie als Datenwerte wie in Tabelle A.1 vorgegeben.

Setzen wir die Punkte (x_i, f_i) in das Polynom p_n ein:

$$p_n(x_i) = f_i = a_0 + a_1 x_i + \cdots + a_n x_i^n \, , \tag{A.4}$$

$$i = 0, 1, \ldots, n \, , \tag{A.5}$$

so erhalten wir ein System von $(n+1)$ linearen Gleichungen zur Berechnung der Koeffizienten $a_0, a_1, \ldots, a_n$. Sind die Stützstellen x_i sämtlich verschieden, dann hat dieses Gleichungssystem für jeden Satz von Stützwerten (A.2) eine eindeutige Lösung.

Da die Konstruktion des Interpolationspolynoms über das Lösen eines linearen Gleichungssystems aufwändig und für praktische Zwecke ungünstig ist, wurden in der Mathematik schon früh andere Konstruktionsmethoden entwickelt. Wir betrachten im Folgenden zwei dieser Verfahren. Da das Polynom durch die Vorgabe der Stützwerte eindeutig bestimmt ist, führen alle

Konstruktionsverfahren auf das gleiche Polynom. Die unterschiedlichen Darstellungen des Polynoms haben jedoch unterschiedliche Eigenschaften und Anwendungen. Ob eine Interpolation oder die Approximation einer Funktion ausgeführt wird, unterscheidet sich in den Formeln zunächst nicht. Immer dann, wenn der Fehler oder die Güte des Verfahrens betrachtet wird, dann brauchen wir einen Vergleich. Insofern liegt dann zwingend eine Approximationsaufgabe zu Grunde mit einer approximierten Funktion, welche genügend oft stetig differenzierbar ist, so dass die vorkommenden Ableitungen existieren und das Ganze Sinn macht.

A.1 Die Interpolationsformel von Lagrange

Die Funktion

$$L_i^n(x) = \frac{(x-x_0)(x-x_1)\cdots(x-x_{i-1})(x-x_{i+1})\cdots(x-x_n)}{(x_i-x_0)(x_i-x_1)\cdots(x_i-x_{i-1})(x_i-x_{i+1})\cdots(x_i-x_n)} \tag{A.6}$$

ist ein Polynom vom Grad n mit den Eigenschaften

$$L_i^n(x_j) = \begin{cases} 1 \text{ für } i = j \ , \\ 0 \text{ für } i \neq j \end{cases} \tag{A.7}$$

und

$$\sum_{i=0}^{n} L_i^n(x) = 1 \ . \tag{A.8}$$

Mit diesen so genannten **Lagrangeschen Stützpolynomen** (A.6) kann man ein Polynom n-ten Grades p_n angeben, welches durch die vorgegebenen Punkte (A.2) verläuft. Es hat die Form

$$p_n(x) = \sum_{i=0}^{n} f_i L_i^n(x) \ . \tag{A.9}$$

Aus der Eigenschaft (A.7) ergibt sich sofort, dass die Werte f_i an den Stützstellen x_i für $i = 0, 1, \ldots, n$ angenommen werden.

Wird durch das Lagrangesche Interpolationspolynom n-ten Grades eine gegebene $(n+1)$-mal stetig differenzierbare Funktion $y = f(x)$ mit $x \in [a, b]$ approximiert, dann ergibt sich die Fehlerformel

$$f(x) - p_n(x) = \frac{f^{(n+1)}(\xi(x))}{(n+1)!}(x - x_0)(x - x_1)\cdots(x - x_n) \qquad \text{(A.10)}$$

mit einer Zwischenstelle $\xi \in (a, b)$, welche von dem Punkt x abhängt. Für praktische Fehlerabschätzungen ist die Formel so nicht sehr brauchbar, da die Zwischenstelle ξ nicht bekannt ist. Man muss die $(n+1)$-te Ableitung von f nach oben und unten abschätzen, um eine konkrete Fehlerschranke zu bekommen.

Beispiel A.1 (Lagrange Interpolation, mit MAPLE-Worksheet). Wir schauen uns die einfachen Fälle $n = 1$ und $n = 2$ genauer an. Im **Falle** $n = 1$ sind zwei Wertepaare

$$(x_0, f_0) \quad \text{und} \quad (x_1, f_1) \qquad \text{(A.11)}$$

vorgegeben. Die beiden Lagrangeschen Stützpolynome haben nach der Formel (A.6) die Gestalt:

$$L_0^{n=1}(x) = \frac{(x - x_1)}{(x_0 - x_1)} \quad \text{und} \quad L_1^{n=1}(x) = \frac{(x - x_0)}{(x_1 - x_0)}\,. \qquad \text{(A.12)}$$

Mit Hilfe der Lagrange-Interpolationsformel ergibt sich das Polynom vom Grad $n = 1$, das heißt die Gerade, in der Gestalt

$$p_{n=1}(x) = f_0 L_0^{n=1}(x) + f_1 L_1^{n=1}(x) = f_0 \frac{(x - x_1)}{(x_0 - x_1)} + f_1 \frac{(x - x_0)}{(x_1 - x_0)}\,. \qquad \text{(A.13)}$$

Im **Falle** $n = 2$ müssen drei Wertepaare vorgegeben sein:

$$(x_0, f_0),\ (x_1, f_1) \quad \text{und} \quad (f_2, f_2)\,. \qquad \text{(A.14)}$$

Die Lagrangeschen Stützpolynome haben dann nach der Formel (A.6) die Gestalt:

$$L_0^{n=2}(x) = \frac{(x - x_1)(x - x_2)}{(x_0 - x_1)(x_0 - x_2)}\,, \qquad \text{(A.15)}$$

$$L_1^{n=2}(x) = \frac{(x - x_0)(x - x_2)}{(x_1 - x_0)(x_1 - x_2)}\,, \qquad \text{(A.16)}$$

$$L_2^{n=2}(x) = \frac{(x - x_0)(x - x_1)}{(x_2 - x_0)(x_2 - x_1)}\,. \qquad \text{(A.17)}$$

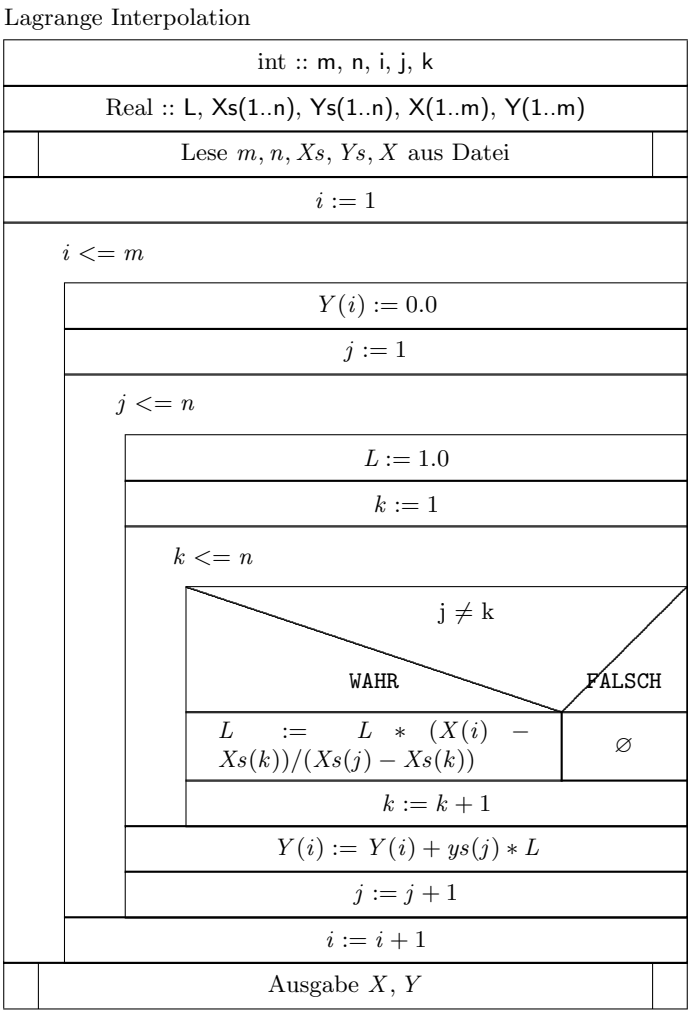


Abb. A.1. Struktogramm für die Lagrange-Interpolation

Mit Hilfe der Lagrangeschen Interpolationsformel (A.9) ergibt sich das Polynom vom Grad $n = 2$ zu

$$p_{n=2}(x) = f_0\, L_0^{n=2}(x) + f_1\, L_1^{n=2}(x) + f_2\, L_2^{n=2}(x)\ . \tag{A.18}$$

Das Polynom ist eine Parabel. □

Beispiel A.2 (Lagrange Interpolation, mit MAPLE-Worksheet). Mit Hilfe der Lagrangeschen Interpolationsformel werden Polynome vom Grad 1, 2, 3, 4 und 5 berechnet, welche die Funktion

$$y = e^x \tag{A.19}$$

im Intervall $[0, 2]$ approximieren. Die Stützstellen werden äquidistant im vorgegebenen Intervall gewählt. Abbildung A.2 zeigt im Vergleich mit der exakten Lösung das Schaubild des Polynoms vom Grad 1 bis 4. Diese Schaubilder werden mit dem MAPLE-**Worksheet** erzeugt.

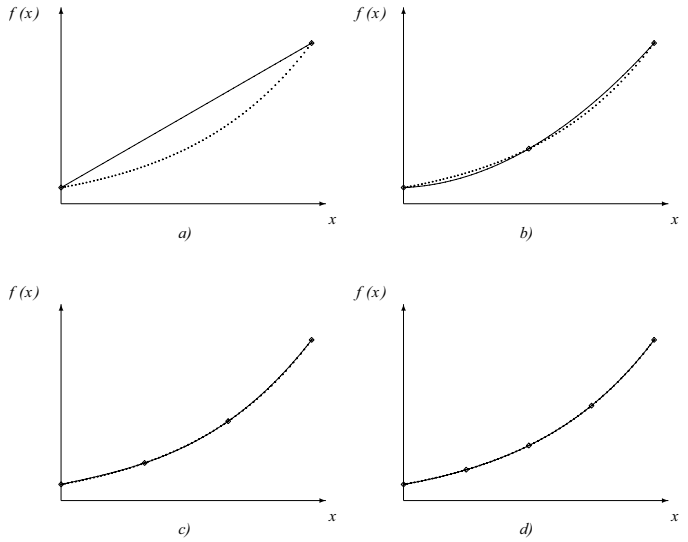


Abb. A.2. Approximation der e-Funktion durch Polynome vom Grad 2 bis 5

Bei dieser einfachen Funktion, die sich auch noch in einfacher Weise differenzieren lässt, kann man den Fehler mit wenig Aufwand bestimmen. Setzen wir für das Polynom 5. Ordnung die Ableitung $f^{(6)}(x) = e^x$ in die Fehlerformel (A.10) ein, so erhalten wir

$$f(x) - p_{n=5}(x) = \frac{e^\xi}{6!}(x - x_0)(x - x_1)\cdots(x - x_5)\ . \tag{A.20}$$

Diesen Ausdruck kann man leicht nach oben abschätzen. Da die Exponentialfunktion monoton wachsend ist, ergibt sich der Maximalwert zu e^2 und man erhält

$$|f(x) - p_{n=5}(x)| \le \frac{e^2}{6!}|(x - x_0)(x - x_1)\cdots(x - x_5)|\ . \tag{A.21}$$

Trägt man dies mit MAPLE als Funktion von x auf, sieht man, dass sich der Fehler zwischen -0.0008 und 0.0003 bewegt. □

A.2 Die Interpolationsformel von Newton

Eine andere Möglichkeit der Darstellung des Interpolationspolynoms ist die **Interpolationsformel von Newton**. Dabei werden die sogenannten **Newtonschen Stützpolynome**

$$N_0(x) := 1, N_j(x) := (x - x_0)(x - x_1) \cdots (x - x_{j-1}),\ j = 1, 2, \ldots, n \tag{A.22}$$

benutzt. Das zugehörige Interpolationspolynom ergibt sich aus der Formel

$$\begin{aligned} p_n(x) &= \sum_{i=0}^{n} c_i N_i(x) \\ &= c_0 + c_1(x - x_0) + \cdots + c_n(x - x_0)(x - x_1) \cdots (x - x_{n-1}) \ . \end{aligned} \tag{A.23}$$

Beispiel A.3. Wir betrachten zunächst den **Fall n = 1**. Das Interpolationspolynom in der Newtonschen Darstellung ergibt sich zu

$$p_1(x) = c_0 + c_1(x - x_0) \ . \tag{A.24}$$

Um die Stützwerte f_0 und f_1 an den Stützstellen x_0, x_1 anzunehmen, muss $c_0 = f_0$ gelten, da der zweite Summand auf der rechten Seite bei $x = x_0$ verschwindet. Man erhält c_1, indem man $x = x_1$ einsetzt und nach c_1 auflöst. Es ergibt sich somit das Polynom ersten Grades

$$p_1(x) = f_0 + \frac{f_1 - f_0}{x_1 - x_0}(x - x_0) \ . \tag{A.25}$$

Dies ist nichts anderes als die Zweipunkteform einer Geraden.

Im **Fall** $n = 2$ erhöht man den Grad des Interpolationspolynoms nach (A.23) und nimmt eine neue Stützstelle x_2 hinzu, dann bleiben die Koeffizienten c_0 und c_1 erhalten und es kommt der Koeffizient c_2 hinzu. Berechnet man diesen aus der Bedingung $p_2(x_2) = f_2$, so hat das Interpolationspolynom die Gestalt

$$\begin{aligned} p_2(x) &= p_1(x) \\ &\quad + \left(\frac{f_2 - f_1}{x_2 - x_1} - \frac{f_1 - f_0}{x_1 - x_0} \right) \frac{1}{x_2 - x_0}(x - x_0)(x - x_1) \ . \end{aligned} \tag{A.26}$$

Es zeigt sich, dass durch die Hinzunahme einer neuen Stützstelle zu dem Polynom p_1 nur ein neuer Term addiert wird. Die schon berechneten Koeffizienten c_0 und c_1 bleiben erhalten. Das ist ein großer Vorteil der Newtonschen Interpolationsformel. Bei der Lagrangeschen Formel müssen bei Hinzunahme eines neuen Stützpaares alle Koeffizienten neu berechnet werden. □

Wie man schon in (A.26) sieht, werden die Koeffizienten mit wachsendem Grad der Stützpolynome immer aufwändiger zu berechnen. Es gibt allerdings eine effiziente rekursive Berechnung. Der Koeffizient c_1, wir er in (A.25) ausgeschrieben ist, entspricht gerade der Steigung der Sekante durch die Punkte (x_0,f_0) und (x_1,f_1). Man bezeichnet diesen Wert als die **dividierte Differenz** bezüglich x_0 und x_1 und schreibt

$$f[x_0, x_1] := \frac{f_1 - f_0}{x_1 - x_0} \; .$$

Die zweite dividierte Differenz bezüglich x_0, x_1 und x_2 ist dann gegeben durch

$$f[x_0, x_1, x_2] := \frac{f[x_1, x_2] - f[x_0, x_1]}{x_2 - x_0}$$

und gerade identisch mit dem Koeffizienten c_2 des Interpolationspolynoms (A.26). Definieren wir allgemein

$$\begin{aligned} f[x_i] &:= f(x_i) \; , \\ f[x_i, x_{i+1}] &:= \frac{f[x_{i+1}] - f[x_i]}{x_{i+1} - x_i} \; , \\ f[x_i, x_{i+1}, \ldots, x_{i+k}] &:= \frac{f[x_{i+1}, \ldots, x_{i+k}] - f[x_i, \ldots, x_{i+k-1}]}{x_{i+k} - x_i} \; , \end{aligned} \tag{A.27}$$

dann können die Koeffizienten der Newtonschen Interpolationsformel aus einer **Tabelle der dividierten Differenzen** abgelesen werden. In Tabelle A.2 sind die dividierten Differenzen der verschiedenen Ordnungen spaltenweise aufgeschrieben. Jede neue Spalte wird aus der vorherigen nach (A.27) entsprechend berechnet. Die Koeffizienten der Newtonschen Interpolationsformel ergeben sich direkt als die Werte der oberen Schrägzeile.

Ein Rechenprogramm mit der Interpolationsformel von Newton läuft in zwei Schritten ab. Im ersten wird die Tabelle mit den für das Polynom benötigten dividierten Differenzen erzeugt und die Koeffizienten c_j berechnet. Für die Auswertung wird das Polynom noch etwas umgeschrieben. Hier wieder das Beispiel $n = 2$:

$$\begin{aligned} p_2 &= c_0 + c_1(x - x_0) + c_2(x - x_0)(x - x_1) \\ &= c_0 + (x - x_0)(c_1 + c_2(x - x_1)) \; . \end{aligned}$$

Die gemeinsamen Faktoren werden nach vorne vor die Klammern gezogen. Die Klammern werden dann von Innen nach Außen ausgewertet, um so die Anzahl der Rechenoperationen so klein wie möglich zu halten.

Für die Lagrangesche Interpolationsformel haben wir eine Fehlerabschätzung angegeben. Wie schon erwähnt, ist diese in der Praxis nur selten nützlich. Bei

$f(x)$	1. dividierte Differenz	2. dividierte Differenz	3. dividierte Differenz
$f[x_0] = c_0$			
	$f[x_0, x_1] = c_1$		
$f[x_1]$		$f[x_0, x_1, x_2] = c_2$	
	$f[x_1, x_2]$		$f[x_0, x_1, x_2, x_3] = c_3$
$f[x_2]$		$f[x_1, x_2, x_3]$	
	$f[x_2, x_3]$		$f[x_1, x_2, x_3, x_4]$
$f[x_3]$		$f[x_2, x_3, x_4]$	
$\vdots$	$f[x_3, x_4]$		$f[x_2, x_3, x_4, x_5]$
	$\vdots$		

Tabelle A.2. Dividierte Differenzen

der Interpolation ist die Funktion f im Allgemeinen nicht gegeben, sondern liegt nur in Form einer Wertetabelle vor. Zum Anderen muss die $(n+1)$-te Ableitung berechnet und an einer Zwischenstelle ausgewertet werden, die nicht bekannt ist. Die Berechnung von Maximum und Minimum der $(n+1)$-ten Ableitung ist meist aufwändig und der berechnete maximale Fehler kann zudem eine schlechte Abschätzung für den tatsächlichen Fehler liefern.

In Beispiel A.2 haben wir Werte der e-Funktion interpoliert und konnten dort den Fehler abschätzen, da die Ableitungen der e-Funktion sehr einfach sind.

Bei der Newtonschen Interpolationsformel gibt es jedoch eine einfache Möglichkeit, zumindest eine Näherung für den Fehler zu erhalten, falls man ein zusätzliches Wertepaar (x_{n+1},f_{n+1}) zur Verfügung hat. Wir betrachten hier den Fall der Approximation und nehmen an, dass $f = f(x)$ eine mindestens $(n+1)$-mal stetig differenzierbare Funktion ist. Mit dem Restglied $R_n = R_n(x)$ gilt

$$f(x) = p_n(x) + R_n(x) \,,$$

wobei $p_n(x)$ das Interpolationspolynom vom Grad n ist. Dabei lässt sich $R_n(x)$ in der Form

$$R_n(x) = (x - x_0)(x - x_1) \cdots (x - x_n) f[x, x_0, x_1, \ldots, x_n]$$

schreiben. Dies lässt sich nicht ausrechnen, da das unbekannte $f(x)$ auf der rechten Seite auftritt. Die Idee ist nun, zur Approximation dieser dividierten Differenz das zusätzliche Wertepaar zusätzliche Wertepaar (x_{n+1},f_{n+1}) einzusetzen. Die Abschätzung ist dann natürlich nicht mehr exakt und hängt von dem vorgegebenen Wertepaar ab. Man kann diesen Wert dann lediglich als ein Näherungswert für den Fehler betrachten.

Bemerkung: Es zeigt sich, dass speziell bei Polynomen höherer Ordnung unangenehme Oszillationen auftreten. Ist der Grad des Polynoms echt größer als fünf, d.h. es liegen mehr als sechs Wertepaare vor, dann sollte man für praktische Anwendungen keine dieser Interpolationen benutzen, sondern die Spline-Interpolation verwenden.

Beispiel A.4 (Newtonsche Interpolation, mit MAPLE-Worksheet). Wir greifen den Tragflügel des Flugzeugs auf, den wir in diesem Anhang zu Beginn tabelliert haben. Wir nehmen ein Polynom 5. Ordnung. Die sechs Stützstellen und Stützwerte sind dabei wie folgt gewählt:

Oberseite		Unterseite	
x	y	x	y
0	0	0	0
14,833	5,951	15,617	−4,733
40,000	6,803	40,000	−4,803
60,051	5,453	59,949	−3,675
80,058	3,178	79,942	−2,066
100,000	0,126	100,000	0,126

Die Lösung mit dem Worksheet **Newton-Interpolation** ist in Abbildung A.3 zu sehen. Bei dieser geringen Anzahl von Punkten ist die Vorderkante des Tragflügelprofils etwas zu spitz. Hier sollte man somit noch ein oder zwei Stützstellen mehr einführen, so dass die Approximation das geometrische Verhalten richtig wiedergeben kann. In unserer Tabelle liegt nur eine Stützstelle in dieser Rundung. □

Beispiel A.5 (Newtonsche Interpolation, mit MAPLE-Worksheet). Wir nehmen nun einen Punkt zusätzlich dazu und wählen die sieben Stützstellen wie folgt aus:

Oberseite		Unterseite	
x	y	x	y
0	0	0	0
1,158	1,954	1,342	−1,830
14,833	5,951	15,617	−4,733
40,000	6,803	40,000	−4,803
60,051	5,453	59,949	−3,675
80,058	3,178	79,942	−2,066
100,000	0,126	100,000	0,126

Das Ergebnis des Worksheets überrascht dann. Statt einer besseren Lösung liefert es oszillierende Kurven, wie dies in der nachfolgenden Abbildung aufgezeichnet ist.

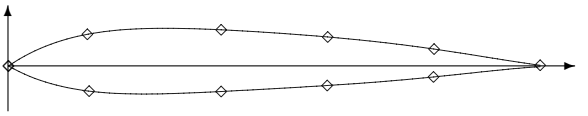


Abb. A.3. Interpolation eines NACA-Profils mit sechs Wertepaaren

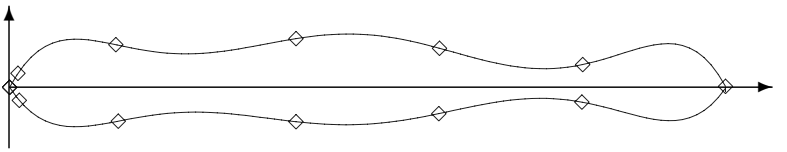


Abb. A.4. Interpolation eines NACA-Profils mit sieben Wertepaaren

Hier zeigt sich das in der Bemerkung angesprochene Verhalten: Das Interpolationspolynom kommt ins Schwingen und liefert eine sehr ungünstige Approximation. Durch die Vorgabe von anderen Stützpaaren kann man diese Oszillationen vermeiden. Für den praktischen Einsatz ist ein solches Verfahren natürlich nicht anwendbar. □

Die Polynominterpolation tendiert zur Instabilität, falls der Grad des Polynoms größer als fünf wird. Aus diesem Grunde haben sich in den letzten Jahren andere Methoden durchgesetzt, die dieses Oszillationsverhalten nicht zeigen. Die wesentliche Idee dahinter ist einfach: Statt ein Interpolationspolynom hoher Ordnung zu benutzen, setzt man Polynome niedriger Ordnung möglichst stetig zusammen. Das nennt man Spline-Interpolation.

A.3 Spline-Interpolation

Der Nachteil der Polynome hoher Ordnung wird bei der Spline-Interpolation vermieden, indem nur Polynome niedriger Ordnung eingesetzt und diese stetig aneinander gehängt werden. Der Begriff Spline oder übersetzt Strak oder

Straklatte stammt ursprünglich aus dem Schiffsbau. Es ist eine lange biegsame Holzlatte, die als Kurvenlineal für das Zeichnen von Schiffsrümpfen verwandt wird. Wird die Latte an einzelnen Punkten durch Nägel fixiert, so biegt sie sich so, dass die Krümmung minimal wird. Wir werden hier die Idee der Spline-Interpolation an den **kubischen Splines** erläutern. Zwischen zwei Stützpunkten wird dabei jeweils ein Polynom vom Grad $n = 3$ also ein kubisches Polynom gelegt. Durch Anschlussbedingungen wird garantiert, dass die zusammengesetzte Funktion zweimal stetig differenzierbar ist. Die Situation sieht dann wie folgt aus.

Gegeben sind $n + 1$ Stützwerte f_i an den Stützstellen $x_0, x_1, \cdots, x_n$, die der Größe nach sortiert sind:

$$a = x_0 < x_1 < x_2 < \cdots < x_n = b \ .$$

Die Näherungsfunktion $S = S(x)$ mit $x \in [a, b]$ setzt sich stückweise aus Polynomen zusammen, d.h.

$$S(x) = S_i(x) \text{ für } x \in [x_i, x_{i+1}] \ ,$$

mit $i = 0, 1, \ldots, n-1$, wobei wir kubische Polynome

$$S_i(x) = a_i + b_i(x - x_i) + c_i(x - x_i)^2 + d_i(x - x_i)^3 \tag{A.28}$$

betrachten wollen.

Jedes kubische Polynom hat vier unbekannte Koeffizienten, also insgesamt gibt es $4n$ freie Parameter. Um diese zu bestimmen wird gefordert, dass die folgenden Bedingungen gelten:

1. S erfüllt die Interpolationsbedingungen

$$\begin{aligned} S(x_i) &= f_i, \ \ i = 0, \cdots, n \ , \text{ d.h.} \\ S_i(x_i) &= f_i, \ \ i = 0, \cdots, n \ . \end{aligned} \tag{A.29}$$

2. S ist im Innern von $[a, b]$ zweimal stetig differenzierbar, d.h.

$$\begin{aligned} S_i(x_i) &= S_{i-1}(x_i), \ \ i = 1, 2, \cdots, n-1 \ , \\ S_i'(x_i) &= S_{i-1}'(x_i), \ \ i = 1, 2, \cdots, n-1 \ , \\ S_i''(x_i) &= S_{i-1}''(x_i), \ \ i = 1, 2, \cdots, n-1 \ . \end{aligned} \tag{A.30}$$

Damit haben wir $4n - 2$ Gleichungen für die $4n$ Koeffizienten. Die fehlenden zwei Bedingungen erhält man aus zusätzlichen Randbedingungen an S. Gilt

$$S_0''(a) = S_{n-1}''(b) = 0 \ ,$$

so verlangt man den stetigen Übergang von S am Rand in eine Gerade. S wird dann der **natürliche kubische Spline** genannt.

Die Berechnung der ersten und zweiten Ableitung der kubischen Polynome S_i liefert

$$\begin{aligned} S_i'(x) &= b_i + 2c_i(x - x_i) + 3d_i(x - x_i)^2 \,, \\ S_i''(x) &= 2c_i + 6d_i(x - x_i) \,. \end{aligned}$$

Aus den Bedingungen (A.29) und (A.30) erhält man damit

1. $a_i = y_i,\ i = 0, 1, \ldots, n$,
2. $c_0 = c_n = 0$,
3. $h_{i-1}c_{i-1} + 2(h_{i-1} + h_i)c_i + h_i c_{i+1} = \frac{3}{h_i}(a_{i+1} - a_i) - \frac{3}{h_{i-1}}(a_i - a_{i-1})$, $i = 1, 2, \cdots, n-1$ mit $h_i := x_{i+1} - x_i,\ i = 0, 1, \ldots, n-1$,
4. $b_i = \frac{1}{h_i}(a_{i+1} - a_i) - \frac{h_i}{3}(c_{i+1} + 2c_i),\ i = 0, 1, \ldots, n-1$,
5. $d_i = \frac{1}{3h_i}(c_{i+1} - c_i),\ i = 0, 1, \ldots, n-1$.

Die Gleichungen in 3. bilden ein System linearer Gleichungen

$$\mathbf{A}\mathbf{c} = \mathbf{r}$$

mit der Matrix

$$\mathbf{A} = \begin{pmatrix} 2(h_0 + h_1) & h_1 & & & \\ h_1 & 2(h_1 + h_2) & h_2 & & \\ & \ddots & \ddots & \ddots & \\ & & \ddots & \ddots & h_{n-2} \\ & & & h_{n-2} & 2(h_{n-2} + h_{n-1}) \end{pmatrix} ,$$

dem Vektor $\mathbf{c}$, in dem die gesuchten Koeffizienten stehen und der rechten Seite $\mathbf{r}$:

$$\mathbf{c} = \begin{pmatrix} c_1 \\ c_2 \\ \vdots \\ c_{n-1} \end{pmatrix} ,\ \mathbf{r} = \begin{pmatrix} \frac{3}{h_1}(a_2 - a_1) - \frac{3}{h_0}(a_1 - a_0) \\ \vdots \\ \frac{3}{h_{n-1}}(a_n - a_{n-1}) - \frac{3}{h_{n-2}}(a_{n-1} - a_{n-2}) \end{pmatrix} .$$

Man nennt dieses lineare Gleichungsystem ein tridiagonales Gleichungssystem, da nur die mittleren drei Diagonalen, Haupt- und die angrenzenden zwei Nebendiagonalen, besetzt sind und alle anderen Koeffizienten der Matrix Null sind. Die Matrix A ist stark diagonal dominant, streng regulär und positiv definit. Bei äquidistanten Schrittweiten h_i ist sie symmetrisch.

Das Gleichungssystem ist damit eindeutig lösbar. Am besten wendet man ein modifiziertes Gauß-Verfahren für tridiagonale Matrizen an, den sogenannten Thomas-Algorithmus an. Dieser wird im Kapitel über Randwertprobleme bei der Anwendung von Differenzenverfahren genauer beschrieben.

Die Vorgabe anderer Randwerte ist möglich, wie periodische Randwerte oder die Vorgabe der ersten Ableitungen der Spline-Funktionen am Rand. Die starke Forderung, dass die Krümmung der kubischen Polynome an den Stützstellen identisch ist, garantiert den glatten Verlauf der Interpolationsfunktion. Sie ist besonders für solche Anwendungen geeignet, von denen man weiß, dass sie durch sehr glatte Kurven beschrieben werden. Der kubische Spline beschreibt im Wesentlichen das Straken und minimiert die Biegungsenergie.

Beispiel A.6 (Spline-Interpolation, mit MAPLE-Worksheet). Wir greifen noch einmal das NACA-Profil aus Tabelle A.1 auf. Diesmal verwenden wir allerdings alle Stützstellen und wenden im Worksheet die Spline-Interpolation darauf an. Das Ergebnis entspricht einem realen Profil. □

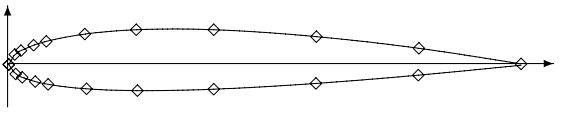


Abb. A.5. Spline-Interpolation eines NACA-Profils

A.4 Bemerkungen und Entscheidungshilfen

Wir haben uns in diesem Anhang auf die Interpolation durch Polynome und stückweise Polynome, speziell die kubischen Splines beschränkt. Dies ist nur ein kleiner Ausschnitt der Interpolation und Approximation, den wir im Kapitel über die numerische Lösung von Integralen und von Differenzialgleichungen benötigen.

Die Interpolation mit Hilfe von Polynomen kann mit Polynomen bis zum Grad fünf ausgeführt werden. Insofern ist der praktische Einsatz der reinen Polynominterpolation beschränkt. Es ist das klassische Verfahren für die Interpolation zwischen einigen wenigen Werten. Werden bei einer Wertetabelle

an Zwischenstellen Werte benötigt und ist die Genauigkeitsanforderung nicht sehr hoch, dann kann dort lokal eine solche Interpolationsformel eingesetzt werden. Es werden dann nur die benachbarten Werte als Stützwerte benutzt und keine Interpolation über alle vorliegenden Werte erzeugt. Bei Näherungswerten an diskreten Punkten im Rahmen einer numerischen Simulation tritt die gleiche Situation auf. Bei der Konstruktion von Näherungsverfahren wie bei der numerischen Integration und Differenziation spielt sie eine wichtige Rolle und kommt in diesen Kapiteln entsprechend vor.

Die Newtonsche und Lagrangesche Interpolationsformel liefert natürlich dasselbe Polynom, da dieses eindeutig bestimmt ist. Allerdings haben die verschiedenen Darstellungen verschiedene Eigenschaften. Bei der Hinzunahme einer zusätzlichen Interpolationsstelle müssen bei der Lagrangeschen Formel alle Stützpolynome L_i neu berechnet werden. Die Berechnung der Lagrangeschen Stützpolynome, alle vom Grad n, ist im Vergleich zur Newton-Interpolation aufwändiger. Insofern ist die Newtonsche Interpolationsformel für praktische Rechnungen vorzuziehen. Die Lagrange Interpolation hat vor allem durch den einfachen Aufbau theoretische Bedeutung. Für praktische Probleme mit vielen Werten oder im Bereich CAD hat die Interpolation mit Splines die reine Polynominterpolation abgelöst. Die Spline-Interpolation ist robuster und zeigt kein Oszillationsverhalten.

Es gibt eine ganze Reihe weiterer Ansätze für die Interpolation. Eng verwandt mit der Polynominterpolation ist die **Hermite-Interpolation**. Hier wird das Interpolationspolynom erzeugt, indem neben den Werten der Funktion auch Werte der Ableitungen mit in die Berechnung eingehen. Dies wird gerne dort angewandt, wo auch Werte von Ableitungen vorliegen. Ein typische Anwendung ist die Interpolation von Näherungslösungen von Differenzialgleichungen. Mit der Hermite-Interpolation können sehr genaue Interpolationspolynome konstruiert werden. Allerdings ist bei vielen Stützstellen auch ein oszillierendes Verhalten möglich. Interpolation mit gebrochen rationalen Funktionen wird insbesondere dort eingesetzt, wo bekannt ist, dass die interpolierte Funktion eine Singularität besitzt. Dagegen wird die trigonometrische Interpolation gerne für periodische Probleme benutzt.

Wir haben hier nur die Interpolation von Funktionen einer Raumdimension betrachtet. Anwendungen in mehreren Dimensionen, wie Kurven und Flächen im Raum sind zum Beispiel im Bereich der CAD-Systeme sehr wichtig. Weitergehende Darstellungen der Interpolation findet man in verschiedenen Büchern zur numerische Mathematik, z. B. [15], [16], [50], [58], [55]. Ein Standardwerk zu der Spline-Interpolation ist [13]. Unterprogramme zur Interpolation finden sich in [47] oder [17] und in den großen numerischen Software-Bibliotheken IMSL [68] und NAG [45].

B Lösen nichtlinearer Gleichungen

In diesem Anhang geht es um die numerische Lösung von nichtlinearen algebraischen Gleichungen, wie z.B.

$$x^2 + 7x + 9 = 0 \,,$$
$$e^x + \cos x = \sin x \,,$$
$$g(x) = h(x) + 7, \ldots \,.$$

Die Approximation einer nichtlinearen Differenzialgleichung wird auf ein solches Problem führen. So ergeben sich bei den Differenzenverfahren ein System von nichtlinearen Differenzengleichungen, welches dann numerisch gelöst werden muss. Ein anderes Beispiel ist das Schießverfahren in Kapitel 3.2 zur numerischen Lösung von gewöhnlichen Randwertproblemen. Hier ist es so, dass die vorkommende Funktion nicht unbedingt durch eine geschlossene Formel gegeben ist.

Ganz allgemein betrachten wir in diesem Kapitel des Anhangs die numerische Lösung einer Gleichung der Form

$$f(x) = 0 \,. \tag{B.1}$$

Gesucht ist somit eine Nullstelle der im Allgemeinen nichtlinearen Funktion von $f = f(x)$. Ist die nichtlineare Funktion nicht gerade eine quadratische Funktion, dann ist man darauf angewiesen, dies näherungsweise auszuführen. Vorausgesetzt sei im Folgenden immer, dass die Funktion $f = f(x)$ im betrachteten Bereich zumindest **stetig** ist.

Die Stetigkeit von Funktionen, die durch Formelausdrücke gegeben sind, kann man entsprechend überprüfen. Beim Schießverfahren in Kapitel 3.2 ist diese Überprüfung allerdings nicht so einfach. Wir gingen beim Schießverfahren davon aus, dass nach Definition 2.2.1 dies immer der Fall ist, wenn das Problem sachgemäß gestellt ist.

Im Folgenden werden wir verschiedene Verfahren zur Lösung nichtlinearer Gleichungen vorstellen.

B.1 Bisektion

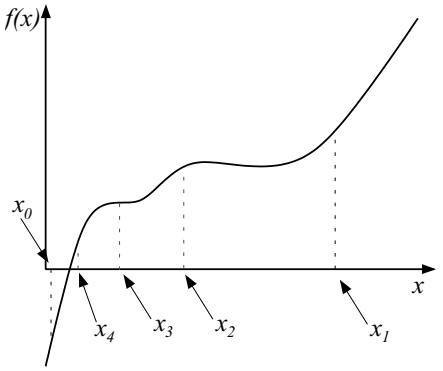


Abb. B.1. Prinzip des Bisektionsverfahrens

Aus der Analysis kennt man den Satz von Rolle, der besagt, dass eine stetige Funktion, die an einer Stelle einen positiven und an einer anderen Stelle einen negativen Wert hat, dazwischen mindestens eine Nullstelle besitzt. Diese Tatsache kann man als Ausgangspunkt eines numerischen Verfahrens zum Auffinden von Nullstellen benutzen.

Gibt es zwei Punkte x_0, x_1 mit der Eigenschaft, dass

$$f(x_0)\, f(x_1) < 0 \,,$$

d.h. die beiden zugehörigen Funktionswerte $f(x_0)$ und $f(x_1)$ haben unterschiedliche Vorzeichen, dann gibt es nach dem Satz von Rolle zwischen x_0 und x_1 mindestens eine Nullstelle von f.

Wir nehmen an, dass $x_0 < x_1$ und halbieren als nächstes das Intervall $[x_0, x_1]$, nehmen den Mittelpunkt des Intervalls und prüfen das Vorzeichen des Funktionswerts an dieser Stelle. Ist es das gleiche Vorzeichen wie das des Funktionswertes bei x_0, so ersetzen wir x_0 durch den neuen Punkt x_{neu} und wir haben erneut ein Intervall, jetzt $[x_{\mathrm{neu}}, x_1]$, welches die gleichen Eigenschaften wie das Ausgangsintervall $[x_0, x_1]$ hat. Es befindet sich somit eine Nullstelle in diesem Intervall.

Im anderen Fall hat die Funktion bei x_{neu} das gleiche Vorzeichen wie der Funktionswert bei x_1, so ersetzen wir x_1 durch den neuen Punkt. Das neue Intervall ist dann $[x_0, x_{\mathrm{neu}}]$. Wir sehen, wie dieses Verfahren weiter läuft. Man fängt wieder von vorne an. Das betrachtete Intervall wird sukzessive halbiert und die Hälften weiter untersucht, in welcher der Vorzeichenwechsel statt

findet. War der Funktionswert schon Null oder aber sein Betrag so klein, dass er uns als numerische Näherung reicht, dann sind wir bereits fertig.

Die einzelnen Schritte lassen sich in dem folgenden Algorithmus formulieren:

– Berechne $x_{\text{neu}} = \dfrac{x_0 + x_1}{2}$.

– Falls $|f(x_{\text{neu}})|$ klein genug, brich ab und verwende x_{neu} als numerische Näherung der Nullstelle.

– sonst:

  – Falls $f(x_0)f(x_{\text{neu}}) < 0$, ersetze x_1 durch x_{neu} ($x_1 := x_{\text{neu}}$).

  – Andernfalls ersetze x_0 durch x_{neu} ($x_0 := x_{\text{neu}}$).

  – Beginne von vorne.

In der Abbildung (B.1) oben sind die Punkte ihrer Entstehung nach nummeriert. Der Algorithmus ist in Abbildung B.2 zusätzlich als Struktogramm dargestellt.

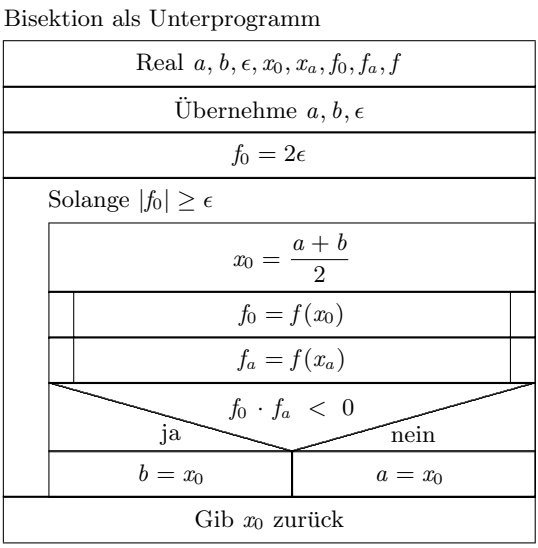


Abb. B.2. Bisektionsverfahren, es wird dabei vorausgesetzt, dass die Funktionswerte an den Grenzen des Ausgangsintervalls unterschiedliches Vorzeichen besitzen

Bemerkung: Haben beim Start des Verfahrens die Funktionswerte an den Intervallgrenzen x_0 und x_1 die gleichen Vorzeichen, so muss man das Intervall entsprechend vergrößern bis unterschiedliche Vorzeichen auftreten.

B.2 Regula Falsi

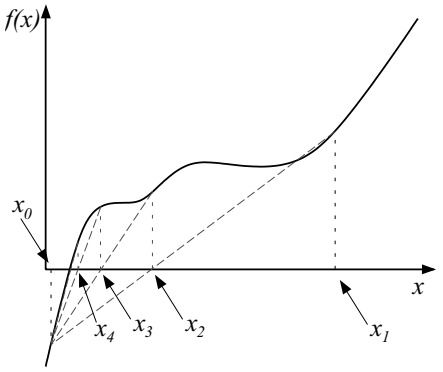


Abb. B.3. Prinzip der Regula Falsi

Schon Adam Riese kannte eine Verbesserung der Bisektion. Die Wahl des neuen Punktes in der Mitte erscheint oft nicht optimal gewählt, und das Verfahren kann selbst bei sehr einfachen Funktionen wie z.B. linearen sehr lange brauchen. Da liegt es nahe, statt der Halbierung eine Gerade durch die Punkte $(x_0, f(x_0))$ und $(x_1, f(x_1))$ zu ziehen. Dort, wo diese die x-Achse schneidet, wählt man dann den neuen Wert x_{neu}. Damit hat man bei linearen Funktionen in einem Schritt die exakte Lösung.

Aber auch bei nichtlinearen Funktionen sollte dieses Verfahren schneller zum Ziel kommen. In Abbildung B.3 zeigt sich dies an unserem Beispiel deutlich. Nun benötigen wir noch die zugehörigen Formeln, nach der sich x_{neu} berechnen lässt. Dazu setzen wir einfach eine allgemeine lineare Funktion an, die in x_0 den Wert $f(x_0)$ hat, in x_1 den Wert $f(x_1)$ und in x_{neu} verschwindet. Nach dem Lagrangschen Interpolationspolynom lautet die Geradengleichung

$$y = f(x_0)\frac{x - x_1}{x_0 - x_1} + f(x_1)\frac{x - x_0}{x_1 - x_0}.$$

Setzen wir $x = x_{neu}$ in diese Gerade ein, ergibt sich als Funktionswert Null. Löst man weiter nach x_{neu} auf, so ergibt sich die gewünschte Formel:

$$x_{\text{neu}} = \frac{x_0 f(x_1) - x_1 f(x_0)}{f(x_1) - f(x_0)} \; . \tag{B.2}$$

Der Algorithmus für die Regula Falsi lautet also:

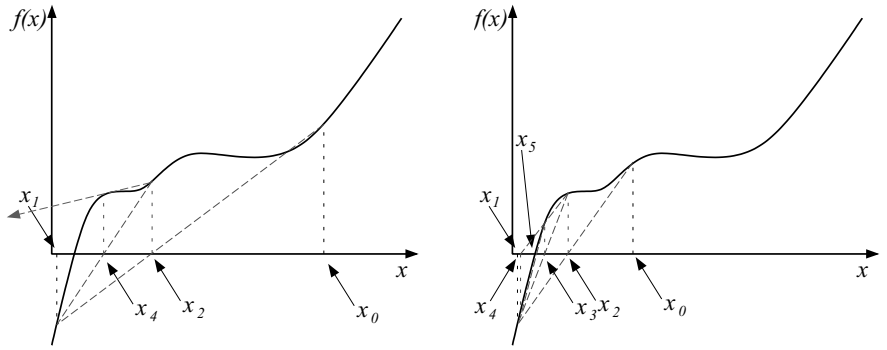
- Berechne x_{neu} nach Formel (B.2).
- Falls $|f(x_{\text{neu}})|$ klein genug, brich ab und verwende x_{neu} als numerische Näherung der Nullstelle.
- sonst:
  - Falls $f(x_0)f(x_{\text{neu}}) < 0$, ersetze x_1 durch x_{neu} ($x_1 := x_{\text{neu}}$).
  - Andernfalls ersetze x_0 durch x_{neu} ($x_0 := x_{\text{neu}}$).
  - Beginne von vorne.

In der Abbildung B.3 hierzu sind die Punkte wieder in der Reihenfolge ihrer Entstehung nummeriert.

B.3 Sekantenverfahren

War die Regula Falsi eine Variation der Bisektion, so ist das Sekantenverfahren eine Variation der Regula Falsi. Es ergibt sich, wenn man sich von Anfang an nicht darum kümmert, ob die jeweiligen zwei Punkte die Nullstelle einschließen oder nicht. Man verwirft einfach, nachdem man den neuen Punkt ausgerechnet hat, den ältesten Punkt. Sind x_0 und x_1 gegeben, so lautet die Iterationsvorschrift:

$$x_{k+1} = \frac{x_{k-1} f(x_k) - x_k f(x_{k-1})}{f(x_k) - f(x_{k-1})} \; . \tag{B.3}$$



(a) x_0, x_1 vertauscht (b) Startwerte näher zusammen

Abb. B.4. Sekantenverfahren

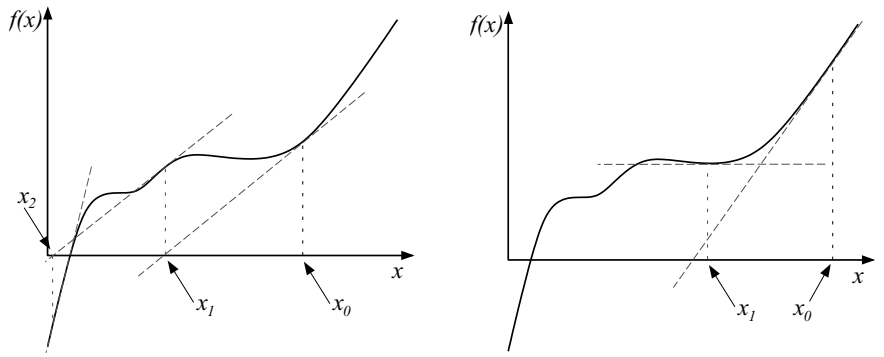
Die Abbildungen B.4 zeigen, wo die Stärken und die Schwächen liegen. Im linken Bild wurden die Startwerte wie im Beispiel für die Bisektion und Regula Falsi gewählt. Das Verfahren führt sehr schnell aus dem Definitionsbereich von f heraus. Auch das Vertauschen der beiden Startwerte (mittleres Bild) kann dieses Problem nicht lösen. Sind dagegen die Startwerte wie im rechten Bild näher beieinander, so kann man sehr schnelle Konvergenz bekommen.

Möchte man dieses Verfahren zur Approximation einer Nullstelle verwenden, so muss man sicher stellen, dass das Verfahren abbricht, sobald man in eine Situation wie in den beiden linken Bildern kommt, und mit zwei anderen Werten neu startet. Diese kann man sich durch ein paar Schritte des Bisektionsverfahrens besorgen.

Eine wichtige Anwendung des Sekantenverfahrens ist die Suche nach einer Anfangseinschließung für die Bisektion oder Regula Falsi. Hat man nämlich am Anfang zwei Punkte, deren Funktionswert das gleiche Vorzeichen hat, so kann man einige Schritte des Sekantenverfahrens durchführen, bis man auf einen Punkt kommt, dessen Funktionswert das umgekehrte Vorzeichen besitzt. Für das Schießverfahren wird das auf Seite 122 beschrieben.

B.4 Das Newton-Verfahren

Für differenzierbare Funktionen kann man aus dem Sekantenverfahren eine Methode herleiten, die oft noch um einiges schneller konvergiert: das Newton-Verfahren. Die Iterationsvorschrift des Sekantenverfahrens (B.3) lässt sich



(a) Guter Startwert (b) Startwert leicht verschoben

Abb. B.5. Newton-Verfahren

umschreiben in

$$x_{k+1} = x_k - f(x_k)\frac{x_k - x_{k-1}}{f(x_k) - f(x_{k-1})} .$$

Für differenzierbares f kann man hier den Grenzübergang $x_{k-1} \to x_k$ durchführen. Der Bruch auf der rechten Seite ist gerade der Kehrwert des Differenzenquotienten. Durch den Grenzübergang geht er demnach in den Kehrwert der Ableitung über. Damit ergibt sich als Iterationsvorschrift für das Newton-Verfahren:

$$x_{k+1} = x_k - \frac{f(x_k)}{f'(x_k)} . \tag{B.4}$$

Man sucht nun nicht mehr nach dem Schnittpunkt der Sekante mit der x-Achse sondern nach dem Schnittpunkt der Tangente mit der x-Achse.

In der rechten Abbildung in B.5 findet man ein Beispiel dafür, dass dies nicht immer gut gehen muss. Wird die Ableitung Null, so ist die Tangente parallel zur x-Achse und schneidet sie nicht. In der Formel (B.4) drückt sich das darin aus, dass der Nenner des Bruchs Null wird. Hier versagt das Verfahren und bricht ab. Hat man jedoch wie in der linken Abbildung einen guten Startwert erwischt, so bekommt man eine besonders schnelle Konvergenz. Man muss also auch hier Abfragen einbauen, die kontrollieren, ob der Startwert gut war. Man kann z. B. dann, wenn der Betrag des Funktionswertes der Iterierten $|f(x_k)|$ steigt statt zu fallen, für ein paar Schritte auf die Regula Falsi übergehen und mit dem dort erreichten Wert wieder in die Newton-Iteration einsteigen.

B.5 Bemerkungen und Entscheidungshilfen

Das Newton-Verfahren ist sicherlich das Schnellste der Verfahren. Allerdings muss man es, wie gesehen, oft mit einem anderen Verfahren kombinieren, weil man einem gegebenen Startwert nicht direkt ansehen kann, ob er zur Konvergenz führt oder nicht. Ist die Funktion nicht differenzierbar oder die Ableitung nicht durch eine geeignete Formel gegeben, so ist das Verfahren nicht anwendbar. Manchmal ist die Berechnung der Ableitung auch so aufwändig, dass sie die schnellere Konvergenz im Vergleich zum Sekantenverfahren wieder aufwiegt. In einem solchen Fall wird auch ein modifiziertes Newton-Verfahren angewandt, bei dem die Ableitung nicht in jedem Iterationsschritt neu berechnet wird.

Ein Vorteil des Sekantenverfahrens gegenüber dem Newton-Verfahren ist der wesentlich geringere Rechenaufwand für einen Iterationsschritt. Das Problem, dass man den Startwerten die Konvergenz oder Divergenz nicht direkt ansehen kann, teilt es allerdings mit diesem. So muss man beide Verfahren oft in Kombination mit der Bisektion oder Regula Falsi benutzen.

Die Bisektion und die Regula Falsi garantieren, dass das Verfahren konvergiert, sobald man eine Einschließung einer Nullstelle durch Punkte mit Funktionswerten unterschiedlichen Vorzeichens hat. Allerdings erkauft man sich diese Sicherheit durch eine – vor allem bei der Bisektion – langsamere Konvergenz. Oft ist diese aber ausreichend, insbesondere dann, wenn keine allzu hohe Genauigkeit gefordert wird. Bei der Regula Falsi kann es vorkommen, dass einer der beiden Punkte sich nur sehr langsam ändert, obwohl der andere der Nullstelle schon sehr nahe kommt. Dieses ist für die Konvergenz natürlich nicht ideal. In einem solchen Fall kann man einen Schritt mit dem Bisektionsverfahren dazwischen ausführen oder einen Schritt mit der Hälfte des Funktionswerts an diesem Punkt in der Formel (B.2) ausführen. Auf diese Art und Weise kann man den neuen Punkt auf die andere Seite der Nullstelle bringen. Der alte Punkt, der sich so langsam bewegt hatte, wird durch eine bessere Näherung ersetzt. Für die Anfangseinschließung brauchen beide Verfahren, wie oben gesehen, oft ein paar Schritte der Sekantenmethode, so dass auch diese beiden mit den anderen kombiniert werden sollten.

Als weiterführende Literatur sei auf die allgemeinen Bücher über numerische Methoden verwiesen, so z.B. [15], [16], [50], [58], [55].

C Iterative Methoden zur numerischen Lösung von linearen Gleichungssytemen

Bei Näherungsverfahren für partielle Differenzialgleichungen treten oft große lineare Gleichungssystem (LGS) auf, welche zur Ermittlung der Werte der Nährungslösung gelöst werden müssen. Diese Gleichungssysteme nennt man schwach besetzt, da nur sehr wenige Elemente der Matrix von Null verschieden sind. So hat man bei einem Differenzenverfahren für die Poisson-Gleichung mit dem 5-Punkte-Stern höchstens 5 Elemente einer Zeile der Matrix von Null verschieden. Schwach besetzte Matrizen treten analog auch bei Finite-Volumen- und Finite-Element-Verfahren auf. Insofern ist die effiziente Lösung von schwach besetzten LGS eine wichtige Grundaufgabe bei der numerischen Lösung von partiellen Differenzialgleichungen.

Iterative Methoden haben bei schwach besetzten Gleichungssystemen wesentliche Vorteile. Die Matrix des Gleichungssystems mit den vielen Nullen muss nicht abgespeichert werden, es werden nur die Elemente ungleich Null benötigt. Weiter muss ein System von Differenzengleichungen, welches schon eine diskrete Approximation der Differenzialgleichung darstellt, nicht unbedingt exakt gelöst werden. Eine Näherungslösung des LGS, deren Fehler etwas kleiner als der schon gemachte Verfahrensfehler ist, reicht hier aus. Der Rechenaufwand eines Iterationsverfahrens wird insbesondere bei großen Systemen sehr viel kleiner sein als der des Gaußschen Eliminationsverfahrens.

Eine iterative Methode wird durch eine Iterationsvorschrift definiert. Auf der linken Seite steht die Iterierte im neuen Iterationsschritt, auf der rechten Seite eine Berechnungsvorschrift, in welche die alten Iterierten eingehen. Durch wiederholtes Ausführen werden sukzessive Näherungen der Lösung erhalten. Die im Kapitel 6 im Abschnitt über die Differenzenverfahren für elliptische Differenzialgleichungen vorgestellten iterativen Verfahren sind bei einer kleinen Anzahl von Unbekannten meist effizient. In der Praxis erfordert allerdings die Auflösung mancher physikalischer Phänomene sehr kleine Schrittweiten und somit eine sehr große Zahl von Gitterpunkten. Hier sind schnellere Methoden nötig, die im Folgenden in einem Überblick vorgestellt werden. Für weitere Informationen wird auf die entsprechende Spezialliteratur verwiesen. Begonnen wird mit einer etwas allgemeineren Darstellung

der in Kapitel 6 vorgestellten klassischen Iterationsverfahren, danach werden Mehrgitter-, CG- und Krylov-Unterraum-Methoden angesprochen.

Wir betrachten ein lineares Gleichungssystem der From

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

mit der $(n \times n)$-Matrix $\mathbf{A}$, der rechten Seite $\mathbf{b}$ und dem gesuchten Lösungsvektor $\mathbf{u}$:

$$\mathbf{A} = (a_{ij})_{i=1,2,..,n;j=1,2,..,n}; \quad \mathbf{b} = (b_1, b_2, .., b_n)^t; \quad \mathbf{u} = (u_1, u_2, .., u_n)^t \ .$$

Das obige LGS besteht aus n Zeilen der Form

$$\sum_{j=1}^{n} a_{ij}\, u_j = b_i \qquad \text{mit} \quad i = 1, 2, ..., n \ .$$

C.1 Die klassischen Iterationsmethoden

In Kapitel 6 haben wir die i-te Gleichung des Systems von Differenzengleichungen für die Poisson-Gleichung nach u_i aufgelöst (d.h. Gleichung 1 nach u_1, Gleichung 2 nach u_2, ..., Gleichung n nach u_n):

$$u_i = \frac{1}{a_{ii}} \, (b_i - \sum_{j=1}^{i-1} a_{ij}\, u_j - \sum_{j=i+1}^{n} a_{ij}\, u_j), \qquad i = 1, 2, \ldots, n \ .$$

Dieses System wird dann iterativ gelöst, indem zunächst eine Anfangsschätzung für die Lösung des LGS $(u_1^{(0)}, u_2^{(0)}, \ldots, u_n^{(0)})$ vorgegeben wird und dann neue Iterierte durch die Vorschrift

$$u_i^{(m+1)} = \frac{1}{a_{ii}} \, (b_i - \sum_{j=1}^{i-1} a_{ij}\, u_j^{(m)} - \sum_{j=i+1}^{n} a_{ij}\, u_j^{(m)}), \qquad i = 1, 2, \ldots, n$$

bestimmt werden.

Dieses Iterationsverfahren heißt **Jacobi-Verfahren** und ist im Worksheet **Jacobi** sowohl direkt als auch in Form einer Prozedur programmiert.

Bei dem Jacobi-Verfahren beginnt man bei der $(m+1)$-ten Iteration mit der 1. Gleichung, um $u_1^{(m+1)}$ durch die "alten" Werte $u_1^{(m)}, u_2^{(m)} \ldots, u_n^{(m)}$ zu berechnen. Anschließend wählt man Gleichung 2 und berechnet $u_2^{(m+1)}$ aus den "alten" Daten $u_1^{(m)}, u_2^{(m)}, \ldots, u_n^{(m)}$. Dabei könnte man im Prinzip schon ausnutzen, dass $u_1^{(m+1)}$ zu diesem Zeitpunkt schon berechnet wurde und voraussichtlich auch schon eine bessere Approximation darstellt. Man sollte also ein verbessertes Verfahren erhalten, indem man bei der Berechnung von $u_i^{(m+1)}$ die schon aktualisierten Werte $u_1^{(m+1)}, u_2^{(m+1)}, \ldots, u_{i-1}^{(m+1)}$ berücksichtigt:

$$u_i^{(m+1)} = \frac{1}{a_{ii}}\left(b_i - \sum_{j=1}^{i-1} a_{ij}\, u_j^{(m+1)} - \sum_{j=i+1}^{n} a_{ij}\, u_j^{(m)}\right), \qquad i = 1, 2, \ldots, n\ .$$

Gauß-Seidel-Verfahren

Dieses Verfahren nennt man das **Gauß-Seidel-Verfahren** und ist im MAPLE-Worksheet **Gauß-Seidel** sowohl direkt als auch in Form einer Prozedur programmiert.

Das **SOR-Verfahren** (Successive Overrelaxation) beruht auf der Tatsache, dass die Konvergenz des Gauß-Seidel-Verfahrens verbessert werden kann, wenn man eine Linearkombination des alten Wertes $u_i^{(m)}$ und des aktualisierten Wertes $u_i^{(m+1)}$ wählt

$$u_i^{(neu)} = w \cdot u_i^{(m+1)} + (1 - w) \cdot u_i^{(m)}\ .$$

SOR-Verfahren

Der Relaxationsparameter w liegt üblicherweise im Bereich $1 \leq w \leq 2$. Bei $w = 1$ ist das Verfahren identisch mit dem Gauß-Seidel-Verfahren. Werte des Relaxationsparameters w über 1 bedeuten, dass der neue iterierte Wert stärker gewichtet wird, daher der Name Überrelaxation. Es muss $0 < w < 2$ sein, da sonst keine Konvergenz erfolgt.

Das SOR-Verfahren ist im MAPLE-Worksheet **SOR** sowohl direkt als auch in Form einer Prozedur programmiert. Es kann gezeigt werden, dass der optimale Relaxationsparameter durch $w_{opt} = \frac{2}{1+\sqrt{1-\rho^2}}$ gegeben ist, wenn ρ der größte Eigenwert der Jacobi-Matrix ist.

Den Iterationsverfahren ist eines gemeinsam, dass die Iteration bei Erreichen der gewünschten Genauigkeit abgebrochen werden muss. Um ein **Abbruchkriterium** zu erhalten, bestimmt man nach jeder Iteration das Residuum. Im m-ten Iterationsschritt ist es durch

$$R_i^{(m)} = \sum_{j=1}^{n} a_{ij}\, u_j^{(m)} - b_i \qquad (i = 1, \ldots, n)$$

gegeben. Von diesem Residuumvektor nehmen wir die maximale Komponente

$$R^{(m)} = \max_{i=1,\ldots,n} |R_i^{(m)}| \,,$$

die Alternative wäre das quadratische Mittel.

Als Abbruchkriterium fordert man in der Regel, dass sowohl

1. das Residuum kleiner einer gewissen Vorgabe ist, $R^{(m)} < \delta_1$, als auch dass

2. die Differenz der Werte zweier aufeinander folgender Iterationen kleiner einer vorgegebenen Schranke sind

$$\max_{i=1,\ldots,n} |\mathbf{u}_i^{(m)} - \mathbf{u}_i^{(m-1)}| < \delta_2.$$

Im Kapitel der Differenzenverfahren für elliptische Differenzialgleichungen werden diese Verfahren so in Rechenprogramme umgesetzt, indem die Differenzengleichungen geeignet umgestellt wurden. Allgemeine theoretische Aussagen und Informationen über die Struktur dieser Iterationsverfahren erhält man über die Matrix-Formulierung: Dazu wird die Matrix $\mathbf{A}$ in die Summe

$$\mathbf{A} = \mathbf{L} + \mathbf{D} + \mathbf{U} \tag{C.1}$$

zerlegt. Dabei ist $\mathbf{L}$ (lower) die Matrix, welche die Elemente von A im unteren Dreieck enthält, $\mathbf{U}$ (upper) ist die Matrix, welche die Elemente im oberen Dreieck enthält und $\mathbf{D}$ die Matrix mit den Diagonalelementen. Die Matrizen haben somit die Gestalt

$$\mathbf{L} = \begin{pmatrix} 0 & \cdots & \cdots & 0 \\ * & \ddots & & \vdots \\ \vdots & \ddots & \ddots & \vdots \\ * & \cdots & * & 0 \end{pmatrix}, \quad \mathbf{D} = \begin{pmatrix} * & & & \\ & \ddots & & \\ & & \ddots & \\ & & & * \end{pmatrix}, \quad \mathbf{U} = \begin{pmatrix} 0 & * & \cdots & * \\ \vdots & \ddots & \ddots & \vdots \\ \vdots & & \ddots & * \\ 0 & \cdots & \cdots & 0 \end{pmatrix}, \tag{C.2}$$

wobei "$*$" für ein Matrixelement steht, welches ungleich Null sein darf. Es wird hier vorausgesetzt, dass die Diagonalmatrix D regulär ist und somit die Inverse D^{-1} existiert.

Wir setzen diese Aufspaltung in das Gleichungssystem ein und erhalten

$$(\mathbf{L}+\mathbf{D}+\mathbf{U})\mathbf{u}=\mathbf{b}\,. \tag{C.3}$$

Behält man hier $\mathbf{D}$ auf der linken Seite und schafft alles andere auf die rechte, ergibt sich

$$\mathbf{D}\mathbf{u}=-(\mathbf{L}+\mathbf{U})\mathbf{u}+\mathbf{b}\,. \tag{C.4}$$

Mit der inversen Matrix $\mathbf{D}^{-1}$ multipliziert folgt

$$\mathbf{u}=-\mathbf{D}^{-1}(\mathbf{L}+\mathbf{U})\mathbf{u}+\mathbf{D}^{-1}\mathbf{b} \tag{C.5}$$

und damit der Ausgangspunkt für eine Iterationsvorschrift:

$$\mathbf{u}^{(m+1)}=-\mathbf{D}^{-1}(\mathbf{L}+\mathbf{U})\mathbf{u}^{(m)}+\mathbf{D}^{-1}\mathbf{b} \tag{C.6}$$

mit $m=0,1,\ldots$. Schaut man sich diese Gleichung komponentenweise an, dann sieht man schnell, dass dies gerade die Matrix-Schreibweise des Jacobi-Verfahrens ist.

Allgemein können wir ein lineares Iterationsverfahren damit durch die Vorschrift

$$\mathbf{u}^{(m+1)}=\mathbf{M}\mathbf{u}^{(m)}+\mathbf{N}\mathbf{b} \tag{C.7}$$

definieren. Durch die Wahl

$$\mathbf{M}_J:=-\mathbf{D}^{-1}(\mathbf{L}+\mathbf{U}),\quad \mathbf{N}_J=\mathbf{D}^{-1} \tag{C.8}$$

erhält man das **Jacobi-Verfahren**.

Beim **Gauß-Seidel-Verfahren** werden die schon berechneten Werte im neuen Iterationslevel schon benutzt. Die zugehörigen Koeffizienten stehen in der unteren Dreiecksmatrix. Man erhält somit für dieses Verfahren zunächst

$$(\mathbf{L}+\mathbf{D})\mathbf{u}=-\mathbf{U}\mathbf{u}+\mathbf{b} \tag{C.9}$$

und daraus die Iterationsvorschrift

$$\mathbf{u}^{(m+1)}=-(\mathbf{L}+\mathbf{D})^{-1}\,\mathbf{U}\mathbf{u}^{(m)}+(\mathbf{L}+\mathbf{D})^{-1}\,\mathbf{b} \tag{C.10}$$

mit den Iterationsmatrizen

$$\mathbf{M_{GS}}:=-(\mathbf{L}+\mathbf{D})^{-1}\,\mathbf{U},\quad \mathbf{N_{GS}}:=(\mathbf{L}+\mathbf{D})^{-1}\,. \tag{C.11}$$

Das SOR-Verfahren lässt sich analog in die Gestalt (C.7) bringen. Die Matrizen der Iterationsvorschrift lauten hier

$$\mathbf{M_{SOR}} = -(\omega\,\mathbf{L} + \mathbf{D})^{-\mathbf{1}}\,((\omega - \mathbf{1})\,\mathbf{D} + \omega\,\mathbf{U}), \tag{C.12}$$

und

$$\mathbf{N_{SOR}} = \omega\,(\omega\,\mathbf{L} + \mathbf{D})^{-\mathbf{1}}\,. \tag{C.13}$$

Für diese Verfahren kann man Konvergenzaussagen ableiten. Das theoretische Werkzeug ist der Banachsche Fixpunktsatz oder besser der endlich dimensionale Spezialfall dieses Satzes. Man muss zeigen, dass die Iterationsmatrix eine Kontraktion definiert. Darauf werden wir hier nicht weiter eingehen, sondern auf die Spezialliteratur verweisen. Es gibt die folgende Aussage:

Die Iteration (C.7) ist genau dann für jeden Startvektor $\mathbf{u}^0$ und jede rechte Seite $\mathbf{b}$ konvergent, wenn

$$\rho(\mathbf{M}) < \mathbf{1} \tag{C.14}$$

gilt. Dabei ist

$$\rho(\mathbf{M}) = \max_{i=1,\ldots,n} |\lambda_i(\mathbf{M})| \tag{C.15}$$

der Spektralradius von $\mathbf{M}$, λ_i bezeichnet den i-ten Eigenwert von $\mathbf{M}$.

Aus diesem Satz lassen sich dann verschiedene Bedingungen ableiten, bei denen die Konvergenz der Iterationsverfahren unabhängig von dem Startvektor und der rechten Seite gesichert ist. Kriterien dieser Art sind:

Bemerkungen:

1. Ist $\mathbf{A}$ strikt diagonal dominant, das heißt es gilt

$$|a_{ii}| > \sum_{\substack{i=1\\ j\neq i}}^{n} |a_{ij}|, \quad i = 1, 2, \ldots, n\,, \tag{C.16}$$

   dann ist das Jacobi- und das Gauß-Seidel-Verfahren konvergent.

2. Ist $\mathbf{A}$ symmetrisch und positiv definit, dann ist das Jacobi- und das SOR-Verfahren mit $\omega \in (0, 2)$ konvergent.

3. Das SOR-Verfahren ist höchstens für $0 < \omega < 2$ konvergent.

Weitere Konvergenzsätze sind in der Spezialliteratur zu finden. Ebenso gibt es hier weitere Aussagen über die optimale Wahl des Relaxationsparameters ω.

Bei der numerischen Lösung der Poisson-Gleichung mit dem Differenzenverfahren in Kapitel 6 ergibt sich

$$\omega_{opt} = \frac{2}{1 + \sqrt{1 - \mu^2}}, \quad \mu := \rho(\mathbf{M}_J) \; . \tag{C.17}$$

Bei der Approximation von elliptischen Randwertproblemen zeigt sich, dass der Spektralradius der Iterationsmatrizen $\mathbf{M}$ das Verhalten

$$\rho(\mathbf{M}) = 1 - 0(h^P) \tag{C.18}$$

besitzt. Dabei ist h die Diskretisierungsschrittweite und der Exponent p positiv. Je kleiner die Schrittweite wird, desto mehr strebt der Spektralradius gegen Eins und das Konvergenzverhalten wird schlechter. Das Problem wird schlecht konditioniert. Speziell bei sehr großen, schwach besetzten Gleichungssystemen ist dann das SOR-Verfahren selbst bei optimaler Wahl von w langsam.

C.2 Mehrgitterverfahren

Die genaue Betrachtung der klassischen Iterationsverfahren zur Lösung der großen schwach besetzten Gleichungssysteme ist der Ausgangspunkt der Mehrgitterverfahren. Es zeigt sich, dass ein klassisches Iterationsverfahren kleinskalige Anteile in der Lösung sehr schnell sehr gut wiedergibt und die große Anzahl der Iterationen auf die großskaligen Lösungsanteile zurückgehen. Wenn wir an das Differenzenverfahren für die Poisson-Gleichung denken, dann ist dieses Verhalten einfach zu motivieren. Setzen wir den 5-Punkte-Stern ein, dann ist jeder Gitterpunkt mit den umliegenden Punkten verbunden. In einer Iteration wird damit Information über diesen 5-Punkte-Stern ausgetauscht. Information mit der Punkten, die in der Nähe liegen, geschieht somit recht schnell, während der Informationsaustausch von weit entfernten Punkten dann sehr viele Iterationen erfordert. Man denke nur daran, wenn die Information von einem Eck des Rechengebiets zu dem anderen Eck transportiert werden muss. Man benötigt über den 5-Punkte-Stern dann eine Anzahl von Iterationen in der Größenordnung der Anzahl der Gitterzellen. Kurz: die hochfrequenten Anteile in der Lösung sind sehr schnell sehr gut approximiert, während eine gute Approximation der niederfrequenten Anteile viele Iterationen erfordert.

Die Definition, was hoch- und niederfrequent oder kurz- und langwellig ist, hängt allerdings von der Gitterschrittweite ab. So sind die langwelligen Anteile von dem feinen Gitter - wir definieren diese z. B. als die Änderungen

gemittelt über 20 Gitterpunkte - kurzwellig auf dem groben Gitter, da sie hier z. B. über 5 Gitterpunkte approximiert werden. Jetzt liegt die Idee auf der Hand. Man führt verschieden feine Gitter ein und approximiert immer nur die kurzwelligen Anteile auf dem entsprechenden Gitter. Dies geht jeweils in einigen wenigen Iterationen recht schnell. Das Ganze muss dann nur noch richtig zusammengesetzt werden.

Wir schauen uns die einfachste Situation an, wo man von dem gröbsten Gitter startet und bis zum feinsten Gitter fortschreitet, was gerne mit geschachtelter Iteration bezeichnet wird. Neben den verschiedenen Gittern $G_1, G_2, \ldots, G_q$ benötigt man noch eine Interpolation von einem Gitter zu dem feineren Gitter. Man nennt dies Prolongation und definiert die Prolongationsoperatoren

$$P_{i,i+1} : \; G_i \to G_{i+1}$$

vom Gitter G_i nach dem Gitter G_{i+1}. Man startet auf dem gröbsten Gitter G_1 mit einer direkten Lösung des Gleichungssystems. Dies kann man sich erlauben, das hier die exakte Lösung, etwa mit dem Gauß-Algorithmus, noch schnell zu berechnen ist. Diese Lösung nennen wir $\mathbf{u}_1$ und die Prolongation auf das nächst feinere Gitter dann mit $\bar{\mathbf{u}}_2$. Mit dieser Startlösung wird dann das Problem iterativ auf diesem Gitter gelöst. Man benötigt nur einige wenige Schritte, da nur noch die auf diesem Gitter kurzwelligen Anteile eingefangen werden müssen. Dann geht man zum nächsten feineren Gitter weiter. Das gesamte Verfahren sieht dann wie folgt aus:

$$\begin{array}{c}
\text{Löse auf } G_1 \text{ exakt} \\
\text{Resultat: } \mathbf{u}_1 \\
\\
\Downarrow \qquad\qquad \bar{\mathbf{u}}_2 \;=\; P_{1,2}\;\mathbf{u}_1 \\
\\
\text{Löse auf } G_2 \text{iterativ, Start: } \bar{\mathbf{u}}_2 \\
\text{Resultat: } \mathbf{u}_2 \\
\\
\Downarrow \qquad\qquad \bar{\mathbf{u}}_3 \;=\; P_{2,3}\;\mathbf{u}_2 \\
\\
\vdots \\
\\
\Downarrow \qquad\qquad \bar{\mathbf{u}}_q \;=\; P_{q-1,q}\;\mathbf{u}_{q-1} \\
\\
\text{Löse auf } G_q \text{iterativ, Start: } \bar{\mathbf{u}}_q \\
\text{Resultat: } \mathbf{u}_q
\end{array}$$

Dieses Vorgehen nennt man dann geschachtelte Iteration oder „Nested Iteration“. Bei der iterativen Lösung auf jedem Gitterlevel benötigt man nur einige wenige Iterationen, da nur die auf diesem Gitter kleinskaligen Komponenten eingefangen werden müssen.

Bei einem Mehrgitterverfahren wird diese geschachtelte Iteration mit einem Vorgang in umgekehrter Richtung kombiniert, vom feinsten Gitter auf das gröbste Gitter. Neben den Prolongationsoperatoren benötigt man nun noch den **Restriktionsoperator**

$$R_{i,i-1} : \; G_i \rightarrow G_{i-1} \; ,$$

der eine Näherung auf G_i nach G_{i-1} bringt und man benötigt die Koeffizientenmatrizen $\mathbf{A}_i$ der Gleichungssysteme für die entsprechenden Gitter G_i. Starten wir zunächst auf dem feinsten Gitter G_q. Nach einigen wenigen Iterationen mit dem Jakobi-Verfahren sind die auf diesem Gitter kleinskaligen Anteile eingefangen. Die Lösung zu diesem Zeitpunkt nennen wir $\tilde{\mathbf{u}}_q$. Das Residuum

$$\mathbf{r}_q = b - \mathbf{A}_q \, \tilde{\mathbf{u}}_q$$

enthält somit keine nennenswerten kleinskaligen Anteile mehr. Dann kann man aber auf das vergröberte Gitter gehen und mittelt das Residuum:

$$\tilde{\mathbf{r}}_{q-1} = R_{q,q-1} \, \mathbf{r}_q \; .$$

Die Korrektur der Lösung ergibt sich durch die Lösung des Gleichungssystems

$$\mathbf{A}_{q-1} \, \mathbf{e} = \tilde{\mathbf{r}}_{q-1} \; .$$

Die Lösung nach einigen wenigen Iterationen nennen wir $\mathbf{e}_{q-1}$ und berechnen das zugehörige Residuum nach

$$\mathbf{r}_{q-1} = \tilde{\mathbf{r}}_{q-1} - \mathbf{A}_{q-1} \, \mathbf{e}_{q-1} \; .$$

Jetzt gilt das gleiche Argument wie oben. Die kleinskaligen Anteile jetzt auf dem Gitter G_{q-1} sind eingefangen, so dass man dieses Residuum wieder auf das gröbere Gitter mittelt und dort einige wenige Iterationen durchführt. Schrittweise kommt man dann an bei

$$\tilde{\mathbf{r}}_1 = R_{2,1} \, \mathbf{r}_2 \; .$$

Die Gleichung

$$\mathbf{A}_1 \, \mathbf{e} = \tilde{\mathbf{r}}_1 \; .$$

wird dann exakt gelöst. Man erhält die Korrektur $\mathbf{e}_1$.

Um zu der Lösung auf dem feinsten Gitter zu kommen, muss dann der Vorgang wie bei der „Nested Iteration“ von dem gröbsten zu dem feinsten Gitter ablaufen, indem man die Korrekturen aufaddiert. Dabei muss man die Anteile auf dem gröberen Gitter in jedem Schritt auf das feinere interpolieren, d. h. prolongieren. Von $i = 1, 2, \ldots, \; q-2$ ergibt sich dann

$$\hat{\mathbf{e}}_{i+1} = \mathbf{e}_{i+1} + P_{i,i+1} \, \tilde{\mathbf{e}}_i \; .$$

Dabei wurde hier die Korrektur auf dem i-ten Gitter $\mathbf{e}_i$ durch $\tilde{\mathbf{e}}_i$ ersetzt. Diese Modifikation ergibt sich wegen der Tatsache, dass durch die Prolongation kleinskalige Fehler erneut eingeschleppt werden können. Diese werden eliminiert, indem man wieder einige wenige Schritte des Iterationsverfahrens durchgeführt. Es ergibt sich $\tilde{\mathbf{e}}_i$ nach einigen wenigen Iterationen des LGS

$$\mathbf{A}_{i+1}\,\mathbf{e} = \tilde{\mathbf{r}}_{i+1}$$

mit Start bei $\hat{\mathbf{e}}_{i+1}$.

Ist man bei $i = q-2$ angekommen, ergibt sich

$$\hat{\mathbf{e}}_q = P_{q-1,q}\,\tilde{\mathbf{e}}_{q-1},\ \mathbf{u}_q = \tilde{\mathbf{u}}_q + \hat{\mathbf{e}}_q$$

als Lösung auf dem Gitter G_q.

Das Mehrgitterverfahren enthält sehr viele verschiedene Elemente, welche optimal an das vorgegebene Problem angepasst werden können. Das sind zum einen die Prolongation und zum anderen die Restriktion, aber auch das iterative Verfahren zur Lösung des Gleichungssystems. Da dies dazu eingesetzt wird, die kleinskaligen Änderungen schnell zu approximieren und das Residuum dann nur noch die großskaligen Fehler enthält, spricht man in diesem Zusammenhang auch von einem Glätter. Als sogenannte Glättungsverfahren kommen die klassischen Iterationsverfahren, allen voran das Jacobi- oder das Gauß-Seidel-Verfahren zum Einsatz. Um die Reduktion der hochfrequenten Fehleranteile zu erreichen, wird das sogenannte gedämpfte Jacobi-Verfahren benutzt mit einem Relaxationsparameter $\omega < 1$.

Oben haben wir einen sogenannten V-Zyklus angegeben: Von dem feinsten Gitter wird zum gröbsten Gitter gegangen und dann wieder zurück. Ein W-Zyklus ergibt sich, wenn man auf dem feinsten Gitter startet und nach Erreichen des gröbsten Gitters nicht bis zum feinsten Gitter zurück geht, sondern nach einigen Schritten nochmals auf das gröbste Gitter geht und dann erst auf das feinste zurück. Die Struktur dieses Ablaufes sieht dann wie ein W aus. Eine Kombination von W- und V-Zyklus nennt man den F-Zyklus.

C.3 Das Verfahren der konjugierten Gradienten

1. Das Richardson-Iterationsverfahren Die klassischen Iterationsverfahren des vorherigen Abschnitts lassen sich auch in der Form

$$\mathbf{u}^{(m+1)} = \mathbf{u}^{(m)} + \alpha\mathbf{B}^{-1}(\mathbf{b} - \mathbf{A}\mathbf{u}^{(m)}) \tag{C.19}$$

mit einer regulären Matrix $\mathbf{B}$ und einem Relaxationsparameter α darstellen. So ergibt sich das Jacobi- und das SOR-Verfahren mit den folgenden Definitionen:

$$\mathbf{B}_J = \mathbf{D}, \quad \alpha_j = 1 \quad \text{bzw.} \quad \mathbf{B}_{SOR} = \omega\mathbf{L} + \mathbf{D}, \quad \alpha_{SOR} = \omega \; . \tag{C.20}$$

Im Falle der Konvergenz ergibt sich aus (C.19)

$$\mathbf{u} = \mathbf{u} + \alpha\mathbf{B}^{-1}(\mathbf{b} - \mathbf{A}\mathbf{u}) \; . \tag{C.21}$$

Man löst somit als Ausgangsgleichung eigentlich

$$\mathbf{B^{-1}A}\mathbf{u} = \mathbf{B^{-1}b} \; . \tag{C.22}$$

Man nennt die Matrix $\mathbf{B}$ auch die Vorkonditionierungsmatrix oder kurz den Vorkonditionierer im Englischen „Preconditioner“. Die zugehörige Iterationsvorschrift wird als **vorkonditioniertes Richardson-Verfahren** bezeichnet.

Der Vorkonditionierer $\mathbf{B}$ sollte so gewählt werden, dass das LGS $\mathbf{B}\,\mathbf{s} = \mathbf{r}$ einfacher als $\mathbf{A}\,\mathbf{u} = \mathbf{b}$ berechnet werden kann. Bei den klassischen Iterationsverfahren ist dies nach (C.20) erfüllt. Bei der Wahl $\mathbf{B} = \mathbf{A}$ und $\alpha = 1$ hätte man die exakte Lösung in einem Schritt. Die Matrix $\mathbf{B}$ sollte $\mathbf{A}$ somit möglichst gut approximieren.

Oft eingesetzt wird die unvollständige Cholesky-Faktorisierung. Die Cholesky-Faktorisierung zerlegt $\mathbf{A}$ in das Produkt $\mathbf{A} = \mathbf{LU}$. Bei den schwach besetzten Matrizen wird dies nur unvollständig ausgeführt, indem nur Elemente berechnet werden, für die $a_{ij} \neq 0$ gilt. Andernfalls würde die Matrix aufgefüllt werden. Wir verweisen hier auf die Spezialliteratur.

Bemerkung: Beim Richardson-Verfahren wird die Matrix nur bei der Berechnung des Residuums benötigt. Es genügt im Programm, das Matrix-Vektor-Produkt mit der schwach besetzten Matrix $\mathbf{A}$ zu berechnen. In dem numerischen Verfahren zur Lösung von elliptischen Randwertproblemen liegt dies mit dem System der Differenzengleichungen vor.

Der Einfachheit halber betrachten wir im folgenden die Richardson-Iteration ohne Vorkonditionierung:

$$\mathbf{u}^{(m+1)} = \mathbf{u}^{(m)} + \alpha(\mathbf{b} - \mathbf{A}\mathbf{u}^{(m)}) \; . \tag{C.23}$$

Betrachtet man die ersten zwei Iterierten, so erhält man

$$\mathbf{u}^{(1)} = \mathbf{u}^{(0)} + \alpha\mathbf{r}^{(0)} \; , \tag{C.24}$$

$$\mathbf{u}^{(2)} = \mathbf{u}^{(1)} + \alpha(\mathbf{b} - \mathbf{A}\mathbf{u}^{(1)}) \tag{C.25}$$

$$= \mathbf{u}^{(0)} + \alpha\mathbf{r}^{(0)} + \alpha(\mathbf{b} - \mathbf{A}\mathbf{u}^{(0)}) \tag{C.26}$$

$$= \mathbf{u}^{(0)} + \alpha\mathbf{r}^{(0)} + \alpha(\mathbf{b} - \mathbf{A}\mathbf{u}^{(0)} - \alpha\mathbf{A}\mathbf{r}^{(0)}) \tag{C.27}$$

$$= \mathbf{u}^{(0)} + 2\alpha\mathbf{r}^{(0)} - \alpha^2\mathbf{A}\mathbf{r}^{(0)} \; . \tag{C.28}$$

Allgemein ergibt sich für m-te Iterierte eine Darstellung in der Form

$$\mathbf{u}^{(m)} = \mathbf{u}^{(0)} + \gamma_1^{(m)}\mathbf{r}^{(0)} + \gamma_2^{(m)}\mathbf{A}\mathbf{r}^{(0)} + \cdots + \gamma_m^{(m)}\mathbf{A}^{m-1}\mathbf{r}^{(0)} \ , \tag{C.29}$$

mit den Koeffizienten $\gamma_i = \gamma_i(\alpha)$. Die m-te Iterierte ist somit korrigiert durch eine Linearkombination der Vektoren $\mathbf{A}^j\mathbf{r}^{(0)},\ j = 0, 1, \ldots,\ m-1$.

Man nennt den Teilraum, welcher durch die Vektoren $\mathbf{A}^j\mathbf{r}^{(0)}$ aufgespannt wird:

$$K_{\mathbf{m}}(\mathbf{r}^{(\mathbf{0})}; \mathbf{A}) = \text{span}\ \{\mathbf{r}^{(\mathbf{0})}, \mathbf{A}\mathbf{r}^{(\mathbf{0})}, \ldots, \mathbf{A}^{m-1}\mathbf{r}^{(0)}\} \ , \tag{C.30}$$

den **Krylov-Teilraum** der Ordnung m. Kurz schreibt man für diesen Sachverhalt

$$\mathbf{u}^{(\mathbf{m})} \in \mathbf{u}^{(\mathbf{0})} + K_{\mathbf{m}}(\mathbf{r}^{(\mathbf{0})}, \mathbf{A}) \ . \tag{C.31}$$

In der Iterationsvorschrift C.23 werden die Koeffizienten $\gamma_i^{(m)}$ durch das fest vorgegebene α bestimmt. Es ist jetzt die Frage, wie dieses α am besten gewählt wird, dass das Verfahren schnell konvergiert. Man benötigt ein Kriterium für den Fehler und versucht diesen durch die Wahl von α zu minimieren.

1. Das CG-Verfahren

Ein Kriterium für die Minimierung des Fehlers wird im Folgenden für den Spezialfall eines symmetrischen, positiv definiten Systems ausgeführt. Das Verfahren nennt man das **Verfahren der konjugierten Gradienten (CG-Verfahren)**.

Die Matrix $\mathbf{A}$ sei symmetrisch und positiv definit. Wir betrachten das quadratische Funktional

$$F(\mathbf{v}) := \frac{1}{2}(\mathbf{v}, \mathbf{A}\mathbf{v}) - (\mathbf{b}, \mathbf{v}) \ , \tag{C.32}$$

wobei (.,.) das Skalarprodukt bezeichnet:

$$(\mathbf{v}, \mathbf{A}\mathbf{v}) = \sum_{i=1}^{n}\sum_{k=1}^{n} a_{ik}\, v_i\, v_k \ , \tag{C.33}$$

$$(\mathbf{b}, \mathbf{v}) = \sum_{i=1}^{n} b_i\, v_i \ . \tag{C.34}$$

Es zeigt sich, dass das Auffinden der Lösung $\mathbf{u}$ des LGS $\mathbf{A}\mathbf{u} = \mathbf{b}$ äquivalent mit der Minimierungsaufgabe

$$\mathbf{u} = \min_{\mathbf{v}} \mathbf{F}(\mathbf{v}) \tag{C.35}$$

ist. Dies sieht man folgendermaßen.

Eine notwendige Bedingung für ein lokales Extremum ist, dass der Gradient $\nabla F(\mathbf{v})$ Null wird. Mit

$$\frac{\partial F}{\partial v_i} = \sum_{k=1}^{n} a_{ik} v_k - b_i \tag{C.36}$$

ergibt sich dieser zu

$$\nabla F(\mathbf{v}) = \mathbf{A}\mathbf{v} - \mathbf{b} =: \mathbf{r} \,. \tag{C.37}$$

Da die zweite Ableitung, die Hesse-Matrix von $F(\mathbf{v})$, gerade $\mathbf{A}$ ist und diese Matrix nach Voraussetzung positiv definit, ist die Lösung $\mathbf{u}$ des LGS die Stelle des Minimums. Es zeigt sich, dass diese eindeutig ist und es sich um das globale Minimum handelt. Die Lösung des Gleichungssystems $\mathbf{A}\mathbf{u} = \mathbf{b}$ ist somit äquivalent zu dem Finden des Minimums von $F(\mathbf{v})$. Damit ist ein Kriterium zur Minimierung des Fehlers gefunden.

Beim CG-Verfahren berechnet man eine Basis $\{\mathbf{q^1}, \mathbf{q^2}, \ldots, \mathbf{q^m}\}$ des Krylov-Unterraums $K_{\mathbf{m}}$, für die

$$(\mathbf{q}^i, \mathbf{A}\mathbf{q}^j) = 0 \quad \text{für} \quad i \neq j \tag{C.38}$$

gilt. Diese Eigenschaft heißt $\mathbf{A}$-orthogonal oder konjugiert bezüglich $\mathbf{A}$. Der Ansatz im m-ten Iterationsschritt ist

$$\mathbf{q^{(m)}} = -\mathbf{r^{(m-1)}} + \beta \mathbf{q^{(m-1)}} \,. \tag{C.39}$$

Der Wert von β ergibt sich aus der Bedingung der Konjugiertheit von $\mathbf{q}^{(m)}$ und $\mathbf{q}^{(m-1)}$ bezüglich $\mathbf{A}$. Den Wert von α erhält man dann aus der Minimierungseigenschaft. Sind $\mathbf{v}$ und $\mathbf{q}$ fest, wird $\mathbf{v}' = \mathbf{v} + \alpha\mathbf{q}$ gesucht mit

$$F(\mathbf{v}') = \min_{\alpha} F(\mathbf{v} + \alpha\mathbf{q}) \,. \tag{C.40}$$

Nach kurzer Rechnung erhält man

$$F(\mathbf{v} + \alpha\,\mathbf{q}) = \frac{1}{2}(\mathbf{v} + \alpha\,\mathbf{q}, \mathbf{A}(\mathbf{v} + \alpha\,\mathbf{q})) - (\mathbf{b}, \mathbf{v} + \alpha\,\mathbf{q}) \tag{C.41}$$

$$= \frac{1}{2}(\mathbf{v}, \mathbf{A}\mathbf{v}) + \alpha\,(\mathbf{p}, \mathbf{A}\mathbf{v}) + \frac{1}{2}\alpha^2\,(\mathbf{q}, \mathbf{A}\mathbf{v}) - (\mathbf{b}, \mathbf{v}) \tag{C.42}$$

$$-\alpha\,(\mathbf{b}, \mathbf{q}) \tag{C.43}$$

$$= \frac{1}{2}\alpha^2\,(\mathbf{q}, \mathbf{A}\mathbf{q}) + \alpha\,(\mathbf{q}, \mathbf{r}) + F(\mathbf{v}) =: F^*(\alpha) \,. \tag{C.44}$$

Das Nullsetzen der Ableitung nach α:

$$\frac{d}{d\alpha} F^*(\alpha) = \alpha\,(\mathbf{q}, \mathbf{A}\mathbf{q}) + (\mathbf{q}, \mathbf{r}) = 0 \tag{C.45}$$

liefert für α den Wert

$$\alpha_{\min} = -\frac{(\mathbf{q}, \mathbf{r})}{(\mathbf{q}, \mathbf{A}\mathbf{q})} \tag{C.46}$$

Die zweite Ableitung nach α ist positiv, da $\mathbf{A}$ positiv definit ist.

Der Rechenaufwand für einen Iterationsschritt ergibt sich aus der schon erwähnten Matrix-Vektor-Multiplikation, aus zwei Skalarprodukten und drei skalaren Multiplikationen von Vektoren. Neben der Matrix $\mathbf{A}$ werden nur $4n$ Speicherplätze für $\mathbf{q}^{(m)}$, $\mathbf{r}^{(m)}$, $\mathbf{u}^{(m)}$ und $\mathbf{s}$ benötigt. Eine Vorkonditionierung beim CG-Verfahren ist möglich, indem $\mathbf{A}$ und $\mathbf{b}$ durch $\mathbf{B^{-1}A}$ und $\mathbf{B^{-1}b}$ ersetzt werden.

3. Methode der verallgemeinerten Residuen

Wenn $\mathbf{A}$ unsymmetrisch ist, versagt das CG-Verfahren. Es gibt hier die Möglichkeit, das unsymmetrische LGS äquivalent in ein System mit einer symmetrischen, positiv definiten Matrix umzuformulieren:

$$\mathbf{A}^T\mathbf{A}\mathbf{v} = \mathbf{A}^T\mathbf{b} \tag{C.47}$$

und das CG-Verfahren auf dieses System anzuwenden. Allerdings verschlechtert sich hier die Konditionszahl - man erhält das Quadrat der Konditionszahl von $\mathbf{A}$. Dieser Ansatz ist somit nur für gut konditionierte LGS sinnvoll.

Ein anderer Ansatz ist die Euklidische Norm des Residuums

$$\| \mathbf{b} - \mathbf{A}\mathbf{v} \|_2^2 \tag{C.48}$$

über dem Krylov-Teilraum K_m zu minimieren. Dieses Verfahren wird GMRES (Generalized Minimum Residual) genannt. Man bestimmt in K_m eine Basis mit dem Gram-Schmidt-Verfahren. Ein Nachteil des GMRES-Verfahrens ist, dass die ganze Basis $\{\mathbf{q}^{(1)}, \ldots, \mathbf{q}^{(m)}\}$ abgespeichert werden muss und im Laufe der Iteration der Speicherplatz anwächst. Um dies zu vermeiden, bricht man nach N Schritten ab und startet mit der letzten Iterierten erneut. Dieser Restart verschlechtert natürlich die Eigenschaften des Verfahrens. Sinnvolle Werte für N liegen zwischen 6 und 20.

Es gibt eine ganze Reihe weiterer Krylov-Teilraum-Verfahren. Wir verweisen hier auf die Literatur, welche im nächsten Abschnitt aufgeführt wird.

C.4 Bemerkungen und Entscheidungshilfen

Weitere Informationen zu den Iterationsverfahren für große schwach besetzte Gleichungssysteme findet man in den Büchern über numerische Methoden wie [55] und Stoer [59], ausführlicher dann in der Spezialliteratur über die numerische Lösung von Gleichungssystemen, z.B. bei Hackbusch [27], Meister

[43] und Golub und van Loan [21]. Den Aspekt der effizienten Ausführung dieser Verfahren auf Parallelrechnern wird von Schwandt [53] behandelt.

Die klassischen Iterationsverfahren, allen voran das SOR-Verfahren, werden auch heute noch für kleine Probleme eingesetzt. Ist die Anzahl der Gitterpunkte groß, dann steigt die Anzahl der Iterationen beträchtlich an und diese Verfahren werden zu langsam. Die Bücher von Varga [67] und Young [74] sind die klassischen Werke über diese iterativen Verfahren.

Die klassischen Iterationsverfahren, allen voran das SOR-Verfahren, werden auch heute noch für kleine Probleme eingesetzt. Ist die Anzahl der Gitterpunkte groß, dann steigt die Anzahl der Iterationen beträchtlich an und diese Verfahren werden zu langsam. Die Bücher von Varga [67] und Young [74] sind die klassischen Werke über diese iterativen Verfahren.

Für sehr große Probleme schneiden die Mehrgitterverfahren bezüglich der Rechenzeit oft am besten ab. Wie beschrieben gibt es bei diesen Verfahren eine ganze Reihe von Freiheitsgraden: Restriktion, Prolongation, Iterations-Verfahren und Art des Zyklus: V, W oder F, welche auf das entsprechende Problem optimiert werden können. Für einen Nicht-Fachmann ergibt sich daraus die Schwierigkeit, dass die beste Auswahl dieser Komponenten einiges an Erfahrung erfordert. Für die Standardprobleme, wie die Poisson-Gleichung, gibt es hier optimierte Versionen in der Literatur oder in Programmbibliotheken, die in Bezug auf Rechenzeit meist nicht zu schlagen sind. Für Nicht-Standardprobleme können aber die Schwierigkeiten auftreten, dass der Einsatz dieser Verfahren und die Optimierung etwas Zeit erfordert. Von der Programmstruktur her ist ein Mehrgitterverfahren recht kompliziert. Eine eigene Programmentwicklung setzt hier schon eine gute Kenntnis und Erfahrung im Programmieren voraus. Ein Tutorial zur Entwicklung eines Mehrgitterverfahrens ist das Buch von Briggs, Henson und McCormick [7]. Weitere Beschreibungen dieser Verfahren finden sich in der oben angeführten Literatur. Es gibt darüber hinaus Monographien speziell über Mehrgitterverfahren, z. B. von Hackbusch [25] und Wesseling [70].

Das CG-Verfahren wurde als direktes Verfahren entwickelt. Insbesondere durch die Vorkonditionierung des Gleichungssystems hat es sich dann später als ein sehr robustes und schnelles iteratives Verfahren gezeigt. Es braucht im Allgemeinen weniger Rechenaufwand als das SOR-Verfahren. Die Matrix $\mathbf{A}$ muss bei dem CG-Verfahren symmetrisch und positiv definit sein. Wie schon oben erwähnt wurden in den letzten Jahren ähnliche Methoden angegeben, bei denen diese Voraussetzungen abgeschwächt sind. Bei sehr großen Systemen und den Standardproblemen ist ein Mehrgitterverfahren in Bezug auf Rechenzeiten einer Krylov-Unterraum-Methode im Allgemeinen überlegen. Allerdings ist die Krylov-Unterraum-Methode mit einer guten Vorkonditionierung sehr robust und unproblematisch bei dem Einsatz auf ein Nicht-

Standardproblem. Es muss kein Parameter eingestellt werden. Mehrgitterverfahren können hier auch zur Vorkonditionierung eingesetzt werden. Möchte man sehr große Probleme lösen und dies öfters, dann wird sich der Aufwand lohnen, ein Mehrgitterverfahren geeignet anzupassen. Eine Übersicht über Krylov-Unterraum-Methoden geben die Bücher von Hackbusch [27], Meister [43] und Axelsson [1].

Bei den klassischen Iterationsverfahren wurde diskutiert, wie es vermieden werden kann, dass die Koeffizientenmatrix des Gleichungssystems abgespeichert wird. Bei den anderen Iterationsverfahren wird als zentraler Baustein die Matrix-Vektor-Multiplikation benötigt. Insofern geht es im Allgemeinen darum, diese Operation geschickt abzuspeichern. Es werden hier auf einen Vektor nacheinander nur die Elemente $\neq 0$ abgespeichert. Es werden dann Hilfsfelder eingeführt, welche die Position der Nicht-Null-Koeffizienten, eindeutig beschreiben. Dadurch wird der Speicherplatzbedarf proportional zu der Anzahl der Unbekannten.

Dadurch, dass viele Probleme auf die Lösung großer linearer Gleichungssysteme zurückgeführt werden können, gibt es in diesem Bereich eine ganze Reihe von Programmen. Die Computeralgebra-Systeme wie MAPLEoder MATLAB haben die entsprechende Software zur Verfügung. Die Programmpakete wie IMSL und NAG bieten diesbezüglich viele Routinen an. Berühmt ist die Programmesammlung LAPACK zur linearen Algebra. Viele der IMSL- und NAG-Programme gehen auf diese Sammlung zurück.

Da viele Probleme auf die Lösung großer linearer Gleichungssysteme zurückgeführt werden können, gibt es in diesem Bereich eine ganze Reihe von Programmen. MATLAB stellt hier z.B. entsprechende Software zur Verfügung. Die Programm-Pakete wie IMSL [68] und NAG [45] bieten diesbezüglich viele Routinen an. Berühmt ist die Programmesammlung LAPACK zur linearen Algebra. Viele der IMSL- und NAG-Programme gehen auf diese Sammlung zurück. Das Suchen von frei erhältlichen Programmen unterstützt z.B.http://www.netlib.org.

Literaturverzeichnis

1. O. Axelsson, Iterative Solution Methods, Cambridge University Press, Cambridge, 1996
2. W. Beitz und K.-H. Küttner, Dubbel Taschenbuch für den Maschinenbau, Springer-Verlag, Berlin, Heidelberg, New York, 1995
3. M. Bollhöfer und V. Mehrmann, Numerische Mathematik, Vieweg-Verlag, Wiesbaden, 2004
4. H. Bossel, Konzepte, Verfahren und Modelle zum Verhalten dynamischer Systeme, Vieweg-Verlag, Braunschweig, Wiesbaden, 1992
5. D. Braess, Finite Elemente, Theorie, schnelle Löser und Anwendungen in der Elastizitätstheorie, Springer-Verlag, Berlin, Heidelberg, New York, 3. Aufl., 2003
6. S. C. Brenner und L. R. Scott, The Mathematical Theory of Finite Element Methods, Springer-Verlag, Berlin, Heidelberg, New York, 1994
7. W. Briggs, V. Henson und S. McCormick, A Multigrid Tutorial, SIAM, Philadelphia, 2000
8. P. Ciarlet, The Finite Element Method for Elliptic Problems, North-Holland, Amsterdam, 1978
9. L. Collatz, Eigenwertaufgaben mit technischen Anwendungen, Akademische Verlagsgesellschaft, Leipzig, 1963
10. R. Courant, K. O. Friedrichs und H. Lewy, "Über die partiellen Differentialgleichungen der mathematischen Physik," Math. Ann., **100**, S. 32–74, 1928
11. R. Courant und D. Hilbert, Methoden der mathematischen Physik, Springer-Verlag, Berlin, Heidelberg, New York, 4. Aufl., 1993
12. P. Davis und P. Rabinowitz, Methods of Numerical Integration, Academic Press, Orlando, 2. Aufl., 1984
13. C. DeBoor, A Practical Guide to Splines, Springer-Verlag, Berlin, Heidelberg, New York, 2. Aufl., 2001
14. P. Deuflhard und F. Bornemann, Numerische Mathematik II, De Gruyter Verlag, Berlin, 2. Aufl., 2002
15. P. Deuflhard und A. Hohmann, Numerische Mathematik I, De Gruyter Verlag, Berlin, 3. Aufl., 2002
16. G. Engeln-Müllges und F. Reutter, Numerische Mathematik für Ingenieure, BI Wissenschaftsverlag, Mannheim, Wien, Zürich, 1988
17. G. Engeln-Müllges und F. Reutter, Numerik Algorithmen. Entscheidungshilfe zur Auswahl und Nutzung, VDI-Verlag, 8. Aufl., 1996
18. L. C. Evans, Partial Differential Equations, American Mathematical Society, New York, 1998

19. J. D. Faires und R. L. Burden, Numerische Methoden, Spektrum Akademischer Verlag, Heidelberg, Berlin, Oxford, 1994
20. K. Förster, Numerische Behandlung von Differentialgleichungen, RUS-26, Rechenzentrum Universität Stuttgart, Stuttgart, 1995
21. G. H. Golub und C. van Loan, Matrix Computations, John Hopkins University Press, Baltimore, 3. Aufl., 1996
22. R. Grigorieff, Numerik gewöhnlicher Differentialgleichungen, Band 1, Teubner-Verlag, Stuttgart, 1972
23. R. Grigorieff, Numerik gewöhnlicher Differentialgleichungen, Band 2, Teubner-Verlag, Stuttgart, 1977
24. C. Grossmann und H.-G. Roos, Numerik partieller Differentialgleichungen, Teubner-Verlag, Stuttgart, Leipzig, Wiesbaden, 1992
25. W. Hackbusch, Multigrid Methods and Applications, Springer-Verlag, Berlin, Heidelberg, New York, 1985
26. W. Hackbusch, Theorie und Numerik elliptischer Differentialgleichungen, Teubner-Verlag, Stuttgart, 1986
27. W. Hackbusch, Iterative Löser großer schwach besetzter Gleichungssysteme, Teubner-Verlag, Stuttgart, 1991
28. E. Hairer, S. P. Nørsett und G. Wanner, Solving Ordinary Differential Equations I. Nonstiff Problems, Springer-Verlag, Berlin, Heidelberg, New York, 1993
29. E. Hairer und G. Wanner, Solving Ordinary Differential Equations II. Stiff and Differential-Algebraic Problems, Springer-Verlag, Berlin, Heidelberg, New York, 1996
30. M. Hanke-Bourgeois, Grundlagen der Numerischen Mathematik und des Wissenschaftlichen Rechnens, Teubner-Verlag, Stuttgart, Leipzig, Wiesbaden, 2002
31. G. Hellwig, Partielle Differentialgleichungen, Teubner-Verlag, Stuttgart, 1960
32. H. Heuser, Gewöhnliche Differentialgleichungen, Teubner, Stuttgart, 3. Aufl., 1995
33. C. Hirsch, Numerical Computation of Internal and External Flows, Bd. 1 und 2, Wiley, Chichester, 1988
34. F. John, Partial Differential Equations, Springer-Verlag, Berlin, Heidelberg, New York, 1982
35. M. Jung und U. Langer, Methode der finiten Elemente für Ingenieure, Teubner-Verlag, Stuttgart, Leipzig, Wiesbaden, 2001
36. H.-B. Keller, Numercal Methods for Two-Point Boundary-Value Problems, Dover Publications, New York, 1992
37. P. Knabner und L. Angermann, Numerik partieller Differentialgleichungen, Springer-Verlag, Berlin, Heidelberg, New York, 2000
38. R. Kress, Numerical Analysis, Springer-Verlag, Berlin, Heidelberg, New York, 1998
39. D. Kröner, Numerical Schemes for Conservation Laws, jointly by Wiley and Teubner-Verlag, Chichester, Stuttgart, 1997
40. R. J. LeVeque, Numerical Methods for Conservation Laws, Birkhäuser-Verlag, Basel, 1990
41. D. Marsal, Finite Differenzen und Elemente : Numerische Lösung von Variationsproblemen und partiellen Differentialgleichungen, Springer-Verlag, Berlin, Heidelberg, New York, 1989

42. T. Meis und U. Marcowitz, Numerische Behandlung partieller Differentialgleichungen I, Springer-Verlag, Berlin, Heidelberg, New York, 1978
43. A. Meister, Numerik linearer Gleichungssysteme, Vieweg-Verlag, Braunschweig, Wiesbaden, 2. Aufl., 1999
44. S. G. Michlin, Variationsmethoden der mathematischen Physik, Akademie-Verlag, Berlin, 1962
45. NAG Numerical Algorithms Group, NAG Numerical Components, www.nag.com
46. R. Plato, Numerische Mathematik kompakt, Vieweg-Verlag, Braunschweig, Wiesbaden, 2. Aufl., 2004
47. W. H. Press, S. A. Teukolsky, W. T. Vetterling und B. P. Flannery, Numerical Recipes in Fortran 77: The Art of Scientific Computing, Cambridge University Press, Cambridge, 2001
48. A. Quarteroni und A. Valli, Numerical Approximations of Partial Differential Equations, Springer-Verlag, Berlin, Heidelberg, New York, 1997
49. R. D. Richtmyer und K. W. Morton, Difference Methods for Initial Value Problems, Wiley, New York, 1967
50. H.-G. Roos und H. Schwetlick, Numerische Mathematik, Teubner-Verlag, Stuttgart, Leipzig, 1999
51. A. A. Samarskij, Theorie der Differenzenverfahren, Teubner-Verlag, Leipzig, 1984
52. W. Schnell, D. Gross und W. Hauger, Technische Mechanik 2, Springer-Verlag, Berlin, Heidelberg, New York, 8. Aufl., 2005
53. H. Schwandt, Parallele Numerik, Teubner-Verlag, Stuttgart, Leipzig, Wiesbaden, 2003
54. H.-R. Schwarz, Methode der finiten Elemente, Teubner-Verlag, Stuttgart, 1980
55. H.-R. Schwarz und N. Köckler, Numerische Mathematik, Teubner-Verlag, Stuttgart, Leipzig, Wiesbaden, 5. Aufl., 2004
56. T. Sonar, Angewandte Mathematik, Modellbildung und Informatik, Vieweg-Verlag, Braunschweig, Wiesbaden, 2001
57. O. Steinbach, Numerische Näherungsverfahren für elliptische Randwertprobleme, Teubner-Verlag, Stuttgart, 2003
58. J. Stoer, Einführung in die Numerische Mathematik I, Springer-Verlag, Berlin, Heidelberg, New York, 9. Aufl., 2002
59. J. Stoer und R. Bulirsch, Einführung in die Numerische Mathematik II, Springer-Verlag, Berlin, Heidelberg, New York, 4. Aufl., 2000
60. G. Strang und G. Fix, An Analysis of the Finite Element Method, Prentice Hall, Englwood-Cliffs, 1973
61. W. A. Strauss, Partielle Differentialgleichungen, Vieweg-Verlag, Braunschweig, Wiesbaden, 1995
62. K. Strehmel und R. Weiner, Numerik gewöhnlicher Differentialgleichungen, Teubner-Verlag, Stuttgart, Leipzig, 1995
63. J. W. Thomas, Numerical Partial Differential Equations: Finite Difference Methods, Springer-Verlag, Berlin, Heidelberg, New York, 1995
64. J. F. Thompson, B. Sony und N. Weatherill, Handbook of Grid Generation, CRC Press, Boca Raton, 1999
65. E. F. Toro, Riemann Solvers and Numerical Methods for Fluid Dynamics, Springer-Verlag, Berlin, Heidelberg, New York, 1997

66. J. van Kan und A. Segal, Numerik partieller Differentialgleichungen für Ingenieure, Teubner-Verlag, Stuttgart, Leipzig, Wiesbaden, 1995
67. R. S. Varga, Matrix Iterative Analysis, Springer-Verlag, Berlin, Heidelberg, New York, 2. Aufl., 2000
68. Visual Numerics, IMSL Numerical Libraries, www.vni.com
69. W. Walter, Gewöhnliche Differentialgleichungen, Springer-Verlag, Berlin, Heidelberg, New York, 7. Aufl., 2000
70. P. Wesseling, An Introduction to Multigrid Methods, Wiley, Chichester, 1992
71. T. Westermann, Mathematik für Ingenieure mit Maple, Bd. 2, Springer-Verlag, Berlin, Heidelberg, New York, 2. Aufl., 2001
72. T. Westermann, Mathematik für Ingenieure mit Maple, Bd. 1, Springer-Verlag, Berlin, Heidelberg, New York, 4. Aufl., 2004
73. T. Westermann, Mathematische Probleme lösen mit Maple, Springer-Verlag, Berlin, Heidelberg, New York, 2. Aufl., 2006
74. D. M. Young, Iterative Solutions of Large Linear Systems, Academic Press, New York, 1971
75. O. C. Zienkiewicz, The Finite Element Method, McGraw-Hill, New York, 3. Aufl., 1977

Index