Stig Larsson
Vidar Thomée

Partielle Differentialgleichungen und numerische Methoden

$$U^n \in S_h,\ U^0 = v_h$$
$$(\bar{\partial}_t U^n, \chi) + a(U^n, \chi) = (f^n, \chi) \quad \forall \chi \in S_h$$

Springer

Larsson · Thomée
Partielle Differentialgleichungen
und numerische Methoden

Stig Larsson · Vidar Thomée

Partielle Differentialgleichungen und numerische Methoden

Übersetzt von Micaela Krieger-Hauwede

Springer

Stig Larsson
Vidar Thomée

Matematiska Vetenskaper
Chalmers tekniska högskola
och Göteborgs universitet
SE-41296 Göteborg
Sweden

stig@math.chalmers.se
thomee@math.chalmers.se

Übersetzer:
Micaela Krieger-Hauwede

Micaela.Krieger@itp.uni-leipzig.de

Die englische Originalausgabe erschien 2003 im Springer-Verlag Heidelberg unter dem Titel: „Partial Differential Equations with Numerical Methods“ (ISBN 3-540-01772-0)

Mathematics Subject Classification (2000): 35-01, 65-01

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 3-540-20823-2 Springer Berlin Heidelberg New York

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

Printed in The Netherlands

Umschlaggestaltung: *design & production GmbH*, Heidelberg
Herstellung: LE-TEX Jelonek, Schmidt & Vöckler GbR, Leipzig
Satz: Reproduktionsfertige Vorlage vom Autor
Gedruckt auf säurefreiem Papier 46/3142YL-5 4 3 2 1 0

Vorwort

Das Anliegen dieses Buches ist es, eine elementare, relativ kurze und hoffentlich lesbare Darstellung der wesentlichen Typen linearer partieller Differentialgleichungen und deren Eigenschaften zu geben. Dies schließt die für deren numerische Lösung am häufigsten verwendeten Methoden ein. Unser Ansatz besteht darin, die mathematische Analyse der Differentialgleichungen in die zugehörige numerische Analyse einzubinden. Für den an partiellen Differentialgleichungen interessierten Mathematiker oder für einen Wissenschaftler, der solche Gleichungen zur Modellierung physikalischer Probleme verwendet, ist es wichtig zu erkennen, dass gewöhnlich numerische Methoden benötigt werden, um die eigentlichen Werte der Lösungen zu finden. Für den Numeriker ist es wesentlich, sich klarzumachen, dass numerische Methoden nur mithilfe ausreichender Kenntnis der Theorie der Differentialgleichungen entworfen, analysiert und verstanden werden können, wobei diskrete Entsprechungen der Eigenschaften der Differentialgleichungen zur Anwendung kommen.

In unserer Darstellung untersuchen wir die drei Haupttypen linearer partieller Differentialgleichungen. Dabei handelt es sich um die elliptischen, parabolischen und hyperbolischen Gleichungen. Für jeden dieser Gleichungstypen enthält dieses Buch jeweils drei Kapitel. Im ersten führen wir jeweils grundlegende mathematische Eigenschaften der Differentialgleichung ein und diskutieren die Existenz, die Eindeutigkeit, die Stabilität und die Regularität der Lösungen für verschiedene Randwertprobleme. Die verbleibenden zwei Kapitel sind den wichtigsten und am weitesten verbreiteten numerischen Methoden, dem finiten Differenzenverfahren und der Methode der finiten Elemente, gewidmet.

Historisch gesehen, handelt es sich bei den finiten Differenzenverfahren um die zuerst entwickelten und angewendeten Methoden. Diese sind gewöhnlich so definiert, dass auf einem gleichmäßigen Punktgitter nach approximativen Lösungen gesucht wird, wobei die Ableitungen der Differentialgleichung durch Differenzenquotienten an den Gitterpunkten ersetzt werden. Die Methoden der finiten Elemente beruhen dagegen auf Variationsformulierungen der Differentialgleichungen und bestimmen auf einer Zerlegung des betrach-

teten Gebietes approximative Lösungen in Form von stückweise definierten Polynomen. Die erste Methode ist in gewisser Weise durch die Schwierigkeit eingeschränkt, das Gitter einem allgemeinen Gebiet anzupassen, während die zweite Methode für allgemeine Geometrien auf natürliche Weise geeigneter ist. Die Methoden finiter Elemente haben sich für elliptische und auch für parabolische Probleme zu den populärsten entwickelt, während für hyperbolische Gleichungen weiterhin das finite Differenzenverfahren dominiert. Obwohl sich die den beiden Klassen zugrunde liegende Philosophie etwas unterscheidet, ist es unserer Ansicht nach vernünftiger, die zweite Methode als Weiterentwicklung der ersten zu betrachten, statt beide als Konkurrenten anzusehen. Zudem sind wir der Ansicht, dass ein Fachmann mit beiden Methoden vertraut sein sollte.

Um die Darstellung leichter zugänglich zu gestalten, geht dem Kapitel über elliptische Gleichungen ein Kapitel zum Zweipunkt-Randwertproblem einer gewöhnlichen Differentialgleichung zweiter Ordnung voran. Vor die Kapitel über hyperbolische und parabolische Gleichungen haben wir ein kurzes Kapitel zum Anfangswertproblem eines Systems gewöhnlicher Differentialgleichungen gesetzt. Außerdem beinhaltet dieses Buch ein Kapitel über Eigenwertprobleme und Entwicklung nach Eigenfunktionen, einem wichtigen Hilfsmittel bei der Analyse partieller Differentialgleichungen. Dort geben wir auch einfache Beispiele numerischer Lösungen von Eigenwertproblemen an.

Das letzte Kapitel liefert einen kurzen Überblick über weitere wichtige Klassen numerischer Methoden, und zwar Kollokationsverfahren, finite Volumenverfahren, Spektralmethoden und Randelementmethoden.

Die Darstellung setzt keine tiefer gehenden Kenntnisse in Analysis und Funktionalanalysis voraus. Im Anhang stellen wir einen Teil des grundlegenden, in den Kapiteln benötigten Stoffes zusammen, im Wesentlichen ohne Beweis. Dabei handelt es sich beispielsweise um Elemente abstrakter linearer Räume und Funktionenräume, insbesondere Sobolev-Räume, im Zusammenhang mit grundlegenden Fakten zu Fourier-Transformationen. Bei der Implementierung numerischer Methoden wird es in der Regel notwendig sein, große Systeme linearer algebraischer Gleichungen zu lösen, wobei dies gewöhnlich mithilfe iterativer Methoden geschieht. Im zweiten Kapitel des Anhangs geben wir deshalb einen Überblick über solche Methoden.

Unser Anliegen besteht also eher darin, eine ziemlich breite Themenvielfalt, viele Darstellungsarten und Konzepte zu behandeln als nur die allgemeinsten und weitreichendsten Resultate zu erläutern oder tiefer auf irgendein Anwendungsgebiet einzugehen. In den Abschnitten mit Problemstellungen, die jeweils die verschiedenen Kapitel abschließen, fordern wir den Leser gelegentlich auf, ein im Text nur erwähntes Resultat zu beweisen und auch einige der dargestellten Konzepte weiterzuentwickeln. Bei einigen Problemstellungen schlagen wir den Test einiger numerischer Methoden auf dem Rechner vor, wobei wir annehmen, dass MATLAB oder eine ähnliche Software verfügbar ist. Am Ende des Buches geben wir einige Standardwerke an, in denen zusätz-

licher Stoff und detailliertere Darstellungen zu finden sind. Dies schließt die Fragestellungen zur Implementierung der numerischen Methoden ein.

Dieses Buch hat sich aus Vorlesungen heraus entwickelt, die wir ursprünglich über einen ziemlich langen Zeitraum an der Chalmers University of Technology und der Universität Göteborg zunächst für Studenten der Ingenieurwissenschaften im dritten Studienjahr und später auch in Einführungskursen für Studenten der angewandten Mathematik gehalten haben. Wir möchten den vielen Studenten in diesen Vorlesungen dafür danken, dass sie uns die Gelegenheit gegeben haben, unsere Konzepte zu prüfen.

Göteborg, *Stig Larsson*
Januar, 2003 *Vidar Thomée*

Inhaltsverzeichnis

1

Einführung

In diesem ersten Kapitel beginnen wir in Abschnitt 1.1 mit der Einführung partieller Differentialgleichungen und den zugehörigen Anfangs-Randwertproblemen, die wir in den folgenden Kapiteln untersuchen werden. Die Gleichungen sind in elliptische, parabolische und hyperbolische Gleichungen klassifiziert, und wir weisen auf die entsprechenden physikalischen Problemstellungen hin, die durch diese Gleichungen modelliert werden. Wir diskutieren das Konzept eines gut gestellten Randwertproblems und die verschiedenen, in unserer folgenden Darstellung verwendeten Methoden. In Abschnitt 1.2 führen wir Bezeichnungen und Konzepte ein, die im gesamten Text benutzt werden. Abschnitt 1.3 beinhaltet eine detaillierte Herleitung der Wärmeleitungsgleichung unter Verwendung physikalischer Prinzipien, die die Bedeutung aller in der Gleichung vorkommenden Terme und die Randbedingungen erklären. Im Abschnitt Problemstellungen 1.4 präsentieren wir zusätzlichen Stoff zur Veranschaulichung.

1.1 Hintergrund

In diesem Abschnitt untersuchen wir Rand- und Anfangswertprobleme partieller Differentialgleichungen, die bei Anwendungen, sowohl vom theoretischen als auch vom numerischen Standpunkt aus gesehen, bedeutsam sind. Als typisches Beispiel eines solchen Randwertproblems betrachten wir zunächst das Dirichlet-Problem für die Poisson-Gleichung mit $x = (x_1, \ldots, x_d)$

$$-\Delta u = f(x) \quad \text{in } \Omega, \tag{1.1}$$

$$u = g(x) \quad \text{auf } \Gamma, \tag{1.2}$$

wobei Δ der durch $\Delta u = \sum_{j=1}^{d} \partial^2 u / \partial x_j^2$ definierte Laplace-Operator und Ω ein beschränktes Gebiet im d-dimensionalen Euklidischen Raum $\mathbf{R}^d$ mit dem Rand Γ ist. Bei den gegebenen Funktionen $f = f(x)$ und $g = g(x)$ handelt

es sich um die *Daten* des Problems. Anstelle der Dirichletschen Randbedingung (1.2) kann man beispielsweise auch die Neumannsche Randbedingung

$$\frac{\partial u}{\partial n} = g(x) \quad \text{auf } \Gamma \tag{1.3}$$

betrachten, wobei $\partial u/\partial n$ die Ableitung in Richtung der äußeren Einheitsnormalen n an Γ bezeichnet. Eine andere Wahl besteht in der Robinschen Randbedingung

$$\frac{\partial u}{\partial n} + \beta(x)u = g(x) \quad \text{auf } \Gamma. \tag{1.4}$$

Allgemeiner ausgedrückt, besitzt eine lineare elliptische Gleichung zweiter Ordnung die Form

$$\mathcal{A}u := -\sum_{i,j=1}^{d} \frac{\partial}{\partial x_i}\Big(a_{ij}(x)\frac{\partial u}{\partial x_j}\Big) + \sum_{j=1}^{d} b_j(x)\frac{\partial u}{\partial x_j} + c(x)u = f(x), \tag{1.5}$$

wobei $A(x) = (a_{ij}(x))$ eine hinreichend glatte, positiv definite Matrix ist. Eine solche Gleichung kann in Ω ebenfalls unter verschiedenen Randbedingungen untersucht werden. Bei den folgenden Betrachtungen werden wir uns der Einfachheit halber häufig auf den isotropen Fall $A(x) = a(x)I$ beschränken, wobei $a(x)$ eine glatte, positive Funktion und I die Einheitsmatrix ist.

Elliptische Gleichungen, wie die oben beschriebenen, kommen bei vielen Anwendungen vor. Sie modellieren beispielsweise verschiedene Potentialfelder (Gravitationsfelder, elektrostatische Felder, magnetostatische Felder usw.), Wahrscheinlichkeitsdichten bei Random-Walk-Problemen, stationäre Wärmeflüsse und biologische Phänomene. Diese Gleichungen sind auch mit wichtigen Gebieten der reinen Mathematik, wie beispielsweise der Funktionentheorie und der Theorie der konformen Abbildungen usw., verknüpft. In Anwendungen beschreiben sie häufig stationäre oder zeitunabhängige physikalische Zustände.

Wir betrachten auch zeitabhängige Probleme, wobei unsere beiden Modellgleichungen die Wärmeleitungsgleichung

$$\frac{\partial u}{\partial t} - \Delta u = f(x,t) \tag{1.6}$$

und die Wellengleichung

$$\frac{\partial^2 u}{\partial t^2} - \Delta u = f(x,t) \tag{1.7}$$

sind. Wir werden sie für positive Zeiten t und für x in $\mathbf{R}^d$ oder in einem beschränkten Gebiet $\Omega \subset \mathbf{R}^d$ unter gewissen Randbedingungen betrachten, wie sie vorhin für die Poisson-Gleichung angegeben wurden. Bei diesen zeitabhängigen Problemen muss der Wert der Lösung u zur Anfangszeit $t = 0$

und bei der Wärmeleitungsgleichung außerdem der Wert von $\partial u/\partial t$ zur Zeit $t = 0$ gegeben sein. Im Falle des unbeschränkten Gebietes $\mathbf{R}^d$ werden die jeweiligen Probleme als reine *Anfangswertprobleme* oder *Cauchy-Probleme* und im Falle des beschränkten Gebietes Ω als gemischte *Anfangs-Randwertprobleme* bezeichnet.

Diese Gleichungen und deren Verallgemeinerungen, die allgemeinere elliptische Operatoren als den Laplace-Operator Δ zulassen, treten wiederum in vielen Anwendungsgebieten auf. Im Falle der Wärmeleitungsgleichung sind dies beispielsweise die Wärmeleitung in Festkörpern, der Massentransport bei der Diffusion, die Diffusion von Wirbeln in viskosen Flüssigkeiten, die telegrafische Übermittlung in Kabeln und die Theorie elektromagnetischer Wellen, die Hydrodynamik sowie stochastische und biologische Prozesse. Im Falle der Wellengleichung sind es Schwingungsprobleme in Festkörpern, Schallwellen in einem Rohr, Stromleitung entlang eines isolierten Kabels mit niedrigem spezifischen Widerstand, Solitärwellen in einem Kanal usw.

Gleichungen vom Typ (1.7) haben einige Eigenschaften mit bestimmten Systemen partieller Differentialgleichungen erster Ordnung gemeinsam. Wir werden folglich auch Grund dazu haben, lineare partielle Differentialgleichungen der Form

$$\frac{\partial u}{\partial t} + \sum_{j=1}^{d} a_j(x,t)\frac{\partial u}{\partial x_j} + a_0(x,t)u = f(x,t)$$

und zugehörige Systeme, bei denen die Koeffizienten Matrizen sind, zu untersuchen. Solche Systeme treten beispielsweise in der Strömungsdynamik und in der elektromagnetischen Feldtheorie auf.

Angewandte Probleme führen häufig auf nichtlineare partielle Differentialgleichungen. Die Behandlung solcher Gleichungen geht über den Umfang unserer Darstellungen hinaus. In vielen Fällen ist es jedoch nützlich, linearisierte Versionen der Gleichungen zu untersuchen. In diesem Sinne ist die Theorie linearer Gleichungen durchaus auch für nichtlineare Probleme relevant.

In den Anwendungen enthalten die in den Modellen auftretenden Gleichungen normalerweise physikalische Parameter. Im Falle der Wärmeleitung erfüllt beispielsweise die Temperatur u in einem Punkt des homogenen, isotropen, über das Gebiet Ω ausgedehnten Festkörpers mit der Wärmeleitfähigkeit k, der Dichte ρ, der spezifischen Wärmekapazität c und einer Wärmequelle $f(x,t)$ die Gleichung

$$\rho\, c\, \frac{\partial u}{\partial t} = \nabla \cdot (k \nabla u) + f(x,t) \quad \text{in } \Omega.$$

Wenn ρ, c und k konstant sind, kann diese Gleichung nach einer einfachen Transformation in der Form (1.6) geschrieben werden. Hängen sie allerdings von x ab, dann taucht ein allgemeinerer elliptischer Operator auf.

In Abschnitt 1.3 leiten wir die Wärmeleitungsgleichungen aus physikalischen Prinzipien ab und erläutern im gegebenen Kontext sowohl die physikalische Bedeutung aller Terme des elliptischen Operators (1.5) als auch die Randbedingungen (1.2), (1.3) und (1.4). Eine entsprechende Herleitung der Wellengleichung wird in Problemstellung 1.2 angegeben. Randwertprobleme für elliptische Gleichungen oder stationäre Probleme können als Grenzfälle des zeitabhängigen Problems für $t \to \infty$ auftreten.

Ein Charakteristikum mathematischer Modellierung besteht darin, dass die Analyse, nachdem das Modell einmal aufgestellt wurde – in unserem Fall als Anfangs- oder Anfangs-Randwertproblem einer partiellen Differentialgleichung – rein mathematischer Natur und unabhängig von jeder spezifischen Anwendung ist, die das Modell beschreibt. Die erhaltenen Ergebnisse sind dann für die verschiedensten Anwendungsbeispiele des Modells gültig. Wir werden deshalb in unseren Ausführungen nur wenig Terminologie aus der Physik und anderen Anwendungsfeldern verwenden, uns dafür aber in den Übungen auf spezielle Anwendungen stützen. Häufig ist es zweckmäßig, sich solche Beispiele einzuprägen, um das intuitive Verständnis eines mathematischen Modells zu schulen.

Die Gleichungen (1.1), (1.6) und (1.7) werden als elliptisch, parabolisch beziehungsweise hyperbolisch bezeichnet. Wir werden in Kapitel 11 auf die Klassifikation partieller Differentialgleichungen zurückkommen und erwähnen hier lediglich, dass Differentialgleichungen der Form

$$a\frac{\partial^2 u}{\partial t^2} + 2b\frac{\partial^2 u}{\partial x \partial t} + c\frac{\partial^2 u}{\partial x^2} + \ldots = f(x,t)$$

in den beiden Variablen x und t als *elliptisch, hyperbolisch* oder *parabolisch* bezeichnet werden, je nachdem, ob $\delta = ac - b^2$ positiv, negativ oder gleich null ist. An dieser Stelle steht ... für eine Linearkombination von Ableitungen maximal erster Ordnung. Insbesondere sind die Gleichungen

$$\begin{aligned}
\frac{\partial^2 u}{\partial t^2} + \frac{\partial^2 u}{\partial x^2} &= f(x,t),\\
\frac{\partial^2 u}{\partial t^2} - \frac{\partial^2 u}{\partial x^2} &= f(x,t),\\
\frac{\partial u}{\partial t} - \frac{\partial^2 u}{\partial x^2} &= f(x,t)
\end{aligned}$$

jeweils von einem dieser drei Typen. Beachten Sie, dass die Bedingungen an das Vorzeichen von δ identisch mit denjenigen sind, die bei der Klassifikation ebener quadratischer Kurven in elliptische, hyperbolische und parabolische auftreten.

Zusammen mit partiellen Differentialgleichungen untersuchen wir auch numerische Approximationen durch finite Differenzenverfahren und Methoden finiter Elemente. Für diese Probleme, d. h. die kontinuierlichen und diskreten Gleichungen, beweisen wir folgende Resultate:

– *Existenz* der Lösungen,
– *Eindeutigkeit* der Lösungen,
– *Stabilität der Lösungen* oder kontinuierliche Abhängigkeit der Lösungen gegenüber Störungen der Daten,
– *Fehlerabschätzungen* (für numerische Methoden).

Ein Randwertproblem, das die ersten drei dieser Bedingungen erfüllt, wird als *gut gestellt* bezeichnet. Um solche Resultate zu beweisen, wenden wir verschiedene Methoden an:

– *Maximumprinzipien*,
– *Fourier-Methoden*; dies sind Methoden, die auf der Verwendung der Fourier-Transformation, der Fourier-Reihenentwicklung und der Entwicklung nach Eigenfunktionen beruhen,
– *Energieabschätzungen*,
– Darstellung der Lösungsoperatoren mithilfe von *Greenschen Funktionen*.

1.2 Notation und mathematische Vorbemerkungen

In diesem Abschnitt werden wir kurz einige grundlegende, in diesem Buch verwendete Notationen einführen. Für weitere Details zu Funktionenräumen und Normen verweisen wir auf Anhang A.

Mit $\mathbf{R}$ und $\mathbf{C}$ bezeichnen wir die Mengen der reellen beziehungsweise komplexen Zahlen und schreiben

$$\mathbf{R}^d = \big\{x = (x_1, \dots, x_d) : x_i \in \mathbf{R},\ i = 1, \dots, d\big\}, \quad \mathbf{R}_+ = \big\{t \in \mathbf{R} : t > 0\big\}.$$

Eine Teilmenge des $\mathbf{R}^d$ wird als Gebiet bezeichnet, wenn sie offen und zusammenhängend ist. Mit Ω bezeichnen wir gewöhnlich ein beschränktes Gebiet im $\mathbf{R}^d$ mit $d = 1, 2$ oder 3 (im Falle $d = 1$ ist Ω ein beschränktes, offenes Intervall). Sein Rand $\partial\Omega$ wird üblicherweise mit Γ bezeichnet. Wir nehmen stets an, dass Γ entweder glatt oder ein Polygon (im Falle $d = 2$) oder ein Polyeder (im Falle $d = 3$) ist. Mit $\bar{\Omega}$ bezeichnen wir den Abschluss von Ω, d. h. $\bar{\Omega} = \Omega \cup \Gamma$. Das Volumen (oder die Länge, die Fläche) von Ω wird mit $|\Omega|$ bezeichnet, das Volumenelement im $\mathbf{R}^d$ ist $\mathrm{d}x = \mathrm{d}x_1 \cdots \mathrm{d}x_d$ und $\mathrm{d}s$ bezeichnet ein Bogenelement (im Falle $d = 2$) oder ein Flächenelement (im Falle $d = 3$) von Γ. Für Vektoren in $\mathbf{R}^d$ benutzen wir das Skalarprodukt $x \cdot y = \sum_{i=1}^d x_i y_i$ und die Norm $|x| = \sqrt{x \cdot y}$.

Seien u, v skalare Funktionen und $w = (w_1, \dots, w_d)$ eine vektorwertige Funktion von $x \in \mathbf{R}^d$. Wir definieren den Gradienten, die Divergenz und den Laplace-Operator durch

$$\nabla v = \operatorname{grad} v = \Big(\frac{\partial v}{\partial x_1}, \dots, \frac{\partial v}{\partial x_d}\Big),$$

$$\nabla \cdot w = \operatorname{div} w = \sum_{i=1}^d \frac{\partial w_i}{\partial x_i},$$

$$\Delta v = \nabla \cdot \nabla v = \sum_{i=1}^{d} \frac{\partial^2 v}{\partial x_i^2}.$$

Wir erinnern uns an das *Divergenztheorem*

$$\int_\Omega \nabla \cdot w \, \mathrm{d}x = \int_\Gamma w \cdot n \, \mathrm{d}s,$$

wobei $n = (n_1, \ldots, n_d)$ die äußere Einheitsnormale an Γ ist. Wenden wir dies auf das Produkt wv an, erhalten wir die *Greensche Formel*:

$$\int_\Omega w \cdot \nabla v \, \mathrm{d}x = \int_\Gamma w \cdot n \, v \, \mathrm{d}s - \int_\Omega \nabla \cdot w \, v \, \mathrm{d}x.$$

Wenden wir sie mit $w = \nabla u$ an, wird die Formel zu

$$\int_\Omega \nabla u \cdot \nabla v \, \mathrm{d}x = \int_\Gamma \frac{\partial u}{\partial n} \, v \, \mathrm{d}s - \int_\Omega \Delta u \, v \, \mathrm{d}x,$$

wobei $\partial u / \partial n = n \cdot \nabla u$ die äußere Normalenableitung von u an Γ ist.

Ein Multiindex $\alpha = (\alpha_1, \ldots, \alpha_d)$ ist ein d-dimensionaler Vektor, wobei α_i eine nichtnegative ganze Zahl ist. Die *Länge* $|\alpha|$ eines Multiindex α ist durch $|\alpha| = \sum_{i=1}^{d} \alpha_i$ definiert. Ist eine Funktion $v : \mathbf{R}^d \to \mathbf{R}$ gegeben, dann können wir ihre partiellen Ableitungen der Ordnung $|\alpha|$ als

$$D^\alpha v = \frac{\partial^{|\alpha|} v}{\partial x_1^{\alpha_1} \cdots \partial x_d^{\alpha_d}} \tag{1.8}$$

schreiben. Eine lineare partielle Differentialgleichung der Ordnung k in Ω kann deshalb in der Form

$$\sum_{|\alpha| \le k} a_\alpha(x) D^\alpha u = f(x)$$

geschrieben werden, wobei die Koeffizienten $a_\alpha(x)$ Funktionen von x in Ω sind. Wir benutzen Indizes auch zur Kennzeichnung partieller Ableitungen, wir schreiben beispielsweise

$$v_t = D_t v = \frac{\partial v}{\partial t}, \quad v_{xx} = D_x^2 v = \frac{\partial^2 v}{\partial x^2}.$$

Für $M \subset \mathbf{R}^d$ bezeichnen wir mit $\mathcal{C}(M)$ den linearen Raum der stetigen Funktionen auf M. Für beschränkte, stetige Funktionen definieren wir die Maximumnorm durch

$$\|v\|_{\mathcal{C}(M)} = \sup_{x \in M} |v(x)|. \tag{1.9}$$

Beispielsweise wird dadurch $\|v\|_{\mathcal{C}(\mathbf{R}^d)}$ definiert. Wenn M eine beschränkte und abgeschlossene Menge ist, d. h. eine kompakte Menge, dann wird das Supremum in (1.9) angenommen und wir können

$$\|v\|_{\mathcal{C}(M)} = \max_{x\in M} |v(x)|$$

schreiben.

Sind ein nicht notwendigerweise beschränktes Gebiet Ω und eine nichtnegative ganze Zahl k gegeben, dann bezeichnen wir mit $\mathcal{C}^k(\Omega)$ die Menge der k-mal stetig differenzierbaren Funktionen in Ω. Für ein beschränktes Gebiet Ω schreiben wir $\mathcal{C}^k(\Omega)$ für diejenigen Funktionen $v \in \mathcal{C}^k(\Omega)$, für die $D^\alpha v \in \mathcal{C}(\bar\Omega)$ für alle $|\alpha| \le k$ gilt. Für Funktionen aus $\mathcal{C}^k(\bar\Omega)$ verwenden wir die Norm

$$\|v\|_{\mathcal{C}^k(\bar\Omega)} = \max_{|\alpha|\le k} \|D^\alpha v\|_{\mathcal{C}(\bar\Omega)}$$

und die Halbnorm, die nur die Ableitungen höchster Ordnung enthält,

$$|v|_{\mathcal{C}^k(\bar\Omega)} = \max_{|\alpha|= k} \|D^\alpha v\|_{\mathcal{C}(\bar\Omega)}.$$

Wenn wir auf einem festen Gebiet Ω arbeiten, lassen wir häufig die Menge in der Bezeichnung weg und schreiben einfach $\|v\|_{\mathcal{C}},\ |v|_{\mathcal{C}^k}$ usw.

Mit $\mathcal{C}_0^k(\Omega)$ bezeichnen wir die Menge der Funktionen $v \in \mathcal{C}^k(\Omega)$, die außerhalb einer kompakten Teilmenge von Ω verschwinden. Insbesondere erfüllen solche Funktionen auf dem Rand von Ω die Gleichung $D^\alpha v = 0$ für alle $|\alpha| \le k$. Analog dazu ist $\mathcal{C}_0^\infty(\mathbf{R}^d)$ die Menge der Funktionen, die stetige Ableitungen aller Ordnungen besitzen und außerhalb einer beschränkten Menge verschwinden.

Wir sagen, dass eine Funktion *glatt* ist, wenn sie, abhängig von der jeweiligen Situation, hinreichend viele stetige Ableitungen besitzt.

Wir verwenden auch häufig den Raum $L_2(\Omega)$ der quadratintegrablen Funktionen mit dem Skalarprodukt und der Norm

$$(v,w) = (v,w)_{L_2(\Omega)} = \int_\Omega vw \,\mathrm{d}x, \quad \|v\| = \|v\|_{L_2(\Omega)} = \Big(\int_\Omega v^2 \,\mathrm{d}x\Big)^{1/2}.$$

Für ein Gebiet Ω verwenden wir auch den Sobolev-Raum $H^k(\Omega)$ mit $k \ge 1$ der Funktionen v, für die $D^\alpha v \in L_2(\Omega)$ für alle $|\alpha| \le k$ gilt. Dieser Raum ist mit der Norm und der Halbnorm

$$\|v\|_k = \|v\|_{H^k(\Omega)} = \Big(\sum_{|\alpha|\le k} \|D^\alpha v\|^2\Big)^{1/2},$$

$$|v|_k = |v|_{H^k(\Omega)} = \Big(\sum_{|\alpha|= k} \|D^\alpha v\|^2\Big)^{1/2}$$

versehen. Weitere Normen werden bei Bedarf lokal definiert und verwendet.

Wir benutzen die Buchstaben c und C zur Bezeichnung verschiedener positiver Konstanten, deren Wert sich von Fall zu Fall unterscheiden kann.

1.3 Physikalische Herleitung der Wärmeleitungsgleichung

Viele Gleichungen werden in der Physik durch Kombination eines Erhaltungssatzes mit phänomenologischen Gleichungen (auch als Materialgleichungen bezeichnet) hergleitet. Ein Erhaltungssatz besagt, dass sich eine physikalische Größe, wie Energie, Masse oder Impuls während eines physikalischen Prozesses nicht ändert. Phänomenologische Gleichungen drücken unsere Annahmen darüber aus, wie sich das Material bei Änderung der Zustandsvariablen verhält.

In diesem Abschnitt werden wir die Wärmeleitung in einem Körper $\Omega \subset \mathbf{R}^3$ mit dem Rand Γ betrachten und die Wärmeleitungsgleichung aus der Energieerhaltung und linearen phänomenologischen Gleichungen herleiten.

Energieerhaltung

Betrachten wir das Wärmegleichgewicht in einer beliebigen Teilmenge $\Omega_0 \subset \Omega$ mit dem Rand Γ_0. Das Energieerhaltungsprinzip besagt, dass die Änderungsrate der Gesamtenergie in Ω_0 gleich dem Wärmezufluss durch Γ_0 plus der von einer Wärmequelle im Inneren von Ω_0 produzierten Wärmemenge ist. Um dies mathematisch auszudrücken, führen wir einige physikalische Größen ein, wobei wir die zugehörige Standard-Maßeinheit in eckigen Klammern angeben.

Mit $e = e(x,t)$ `[J/m^3]` bezeichnen wir die *innere Energiedichte* am Ort x `[m]` zur Zeit t `[s]`. Damit ist die Gesamtwärmemenge in Ω_0 gleich $\int_{\Omega_0} e \,\mathrm{d}x$ `[J]`. Ferner ist mit dem Vektorfeld $j = j(x,t)$ `[J/(m^2s)]` für den *Wärmestrom* und der äußeren Einheitsnormalen n an Γ_0 der Wärmeausfluss durch Γ_0 gleich $\int_{\Gamma_0} j \cdot n \,\mathrm{d}s$ `[J/s]`. Führen wir außerdem die Energiedichte der Wärmequellen $p = p(x,t)$ `[J/(m^3s)]` ein, folgt aus dem Energieerhaltungsprinzip

$$\frac{d}{dt} \int_{\Omega_0} e \,\mathrm{d}x = -\int_{\Gamma_0} j \cdot n \,\mathrm{d}s + \int_{\Omega_0} p \,\mathrm{d}x.$$

Nach Anwendung des Divergenztheorems erhalten wir

$$\int_{\Omega_0} \left(\frac{\partial e}{\partial t} + \nabla \cdot j - p \right) \mathrm{d}x = 0 \quad \text{für } t > 0.$$

Weil $\Omega_0 \subset \Omega$ beliebig ist, folgt daraus

$$\frac{\partial e}{\partial t} + \nabla \cdot j = p \quad \text{in } \Omega \quad \text{für } t > 0. \tag{1.10}$$

Phänomenologische Gleichungen

Die innere Energiedichte e hängt von der absoluten Temperatur T `[K]` und den räumlichen Koordinaten ab. Bei unserer ersten phänomenologische Gleichung nehmen wir an, dass e in der Nähe einer geeignet gewählten Referenztemperatur T_0 linear von T abhängt, d. h. es gilt

$$e = e_0 + \sigma(T - T_0) = e_0 + \sigma\,\vartheta \quad \text{mit } \vartheta = T - T_0. \tag{1.11}$$

Der Koeffizient $\sigma = \sigma(x)$ [J/(m³ K)] wird als *spezifische Wärmekapazität* bezeichnet. (Sie wird gewöhnlich in der Form $\sigma = \rho\, c$ ausgedrückt, wobei ρ [kg/m³] die Massendichte und c [J/(kg K)] die spezifische Wärmekapazität pro Masseneinheit ist.)

Nach dem Fourierschen Gesetz ist der Wärmestrom aufgrund der Wärmeleitung proportional zum Temperaturgradienten, was uns eine zweite phänomenologische Gleichung liefert

$$j = -\lambda \nabla \vartheta.$$

Der Koeffizient $\lambda = \lambda(x)$ [J/(m K s)] wird als Wärmeleitfähigkeit bezeichnet. In einigen Situationen (beispielsweise bei Gas in einem porösen Medium oder beim Wärmetransport in einer Flüssigkeit) wird Wärme auch durch Konvektion mit dem Wärmestrom $v\, e$ transportiert, wobei $v = v(x,t)$ [m/s] das konvektive Geschwindigkeitsvektorfeld ist. Folglich lautet die phänomenologische Gleichung

$$j = -\lambda \nabla \vartheta + v\, e. \tag{1.12}$$

Setzen wir (1.11) und (1.12) in (1.10) ein, erhalten wir die *Wärmeleitungsgleichung* mit Konvektion

$$\sigma \frac{\partial \vartheta}{\partial t} - \nabla \cdot (\lambda \nabla \vartheta) + \nabla \cdot (\sigma\, v\, \vartheta) = q \quad \text{in } \Omega \quad \text{mit } q = p - \nabla \cdot (v\, e_0). \tag{1.13}$$

Randbedingungen

Bei der Modellierung der Wärmeleitung wird die Differentialgleichung (1.13) mit einer *Anfangsbedingung* zur Zeit $t = 0$

$$\vartheta(x, 0) = \vartheta_{\mathrm{i}}(x) \tag{1.14}$$

und einer *Randbedingung* $j \cdot n = \kappa(\vartheta - \vartheta_{\mathrm{a}})$ mit dem Wärmeübergangskoeffizienten $\kappa = \kappa(x,t)$ [J/(m² s K)] verknüpft. Die Randbedingung drückt aus, dass der Wärmestrom durch den Rand proportional zur Differenz zwischen Oberflächentemperatur und Umgebungstemperatur ist. Unter der Annahme, dass der Materialfluss den Rand nicht durchdringt, d. h. $v \cdot n = 0$, erhalten wir aus (1.12)

$$j \cdot n = -\lambda \nabla \vartheta \cdot n = -\lambda \frac{\partial \vartheta}{\partial n} \quad \text{auf } \Gamma,$$

wobei $\partial\vartheta/\partial n = \nabla\vartheta \cdot n$ die äußere Normalenableitung von ϑ ist. Deshalb ist die Randbedingung gleich der *Robinschen Randbedingung*

$$\lambda \frac{\partial \vartheta}{\partial n} + \kappa(\vartheta - \vartheta_{\mathrm{a}}) = 0 \quad \text{auf } \Gamma. \tag{1.15}$$

Der Grenzfall $\kappa = 0$ bedeutet, dass die Grenzfläche vollständig isoliert ist, sodass wir die *Neumannsche Randbedingung*

$$\frac{\partial \vartheta}{\partial n} = 0$$

erhalten. Im anderen Grenzfall dividieren wir in (1.15) durch κ und nehmen $\kappa \to \infty$ an. Wir erhalten die *Dirichletsche Randbedingung*

$$\vartheta = \vartheta_{\mathrm{a}}. \tag{1.16}$$

Der Grenzfall $\kappa = \infty$ bedeutet also, dass sich der Körper im vollständigen thermischen Gleichgewicht mit der Umgebung befindet, d. h. die Wärme fließt frei durch die Oberfläche, sodass die Oberflächentemperatur des Körpers gleich der Umgebungstemperatur ist.

Dimensionslose Form

Häufig ist es nützlich, die obigen Gleichungen in dimensionsloser Form aufzuschreiben. Nach der Wahl der Bezugskonstanten L `[m]`, τ `[s]`, ϑ_{f} `[K]`, σ_{f} `[J/(m^3 K)]`, v_{f} `[m/s]`, usw. definieren wir die dimensionslosen Variablen

$$\tilde{t} = t/\tau, \quad \tilde{x} = x/L, \quad u(\tilde{x}, \tilde{t}) = \vartheta(\tilde{x}L, \tilde{t}\tau)/\vartheta_{\mathrm{f}}.$$

Um die Wärmeleitungsgleichung (1.13) dimensionslos zu machen, dividieren wir sie durch $\lambda_{\mathrm{f}}\vartheta_{\mathrm{f}}/L^2$. Unter Verwendung der Kettenregel

$$\frac{\partial u}{\partial \tilde{t}} = \tau \frac{\partial}{\partial t}\Big(\frac{\vartheta}{\vartheta_{\mathrm{f}}}\Big), \quad \tilde{\nabla} u = L\nabla\Big(\frac{\vartheta}{\vartheta_{\mathrm{f}}}\Big)$$

erhalten wir

$$d\frac{\partial u}{\partial \tilde{t}} - \tilde{\nabla} \cdot (a\tilde{\nabla} u) + \tilde{\nabla} \cdot (bu) = f \quad \text{in } \tilde{\Omega} \tag{1.17}$$

mit

$$d = \frac{L^2\sigma_{\mathrm{f}}}{\tau\lambda_{\mathrm{f}}}\frac{\sigma}{\sigma_{\mathrm{f}}}, \quad a = \frac{\lambda}{\lambda_{\mathrm{f}}}, \quad b = \frac{v_{\mathrm{f}}\sigma_{\mathrm{f}}L}{\lambda_{\mathrm{f}}}\frac{v}{v_{\mathrm{f}}}, \quad f = \frac{L^2}{\lambda_{\mathrm{f}}\vartheta_{\mathrm{f}}}q.$$

Es ist sinnvoll $\tau = L^2\sigma_{\mathrm{f}}/\lambda_{\mathrm{f}}$ so zu wählen, dass für konstantes σ die Beziehung $d = 1$ gilt. Die in der Definition von B auftretende dimensionslose Zahl $\mathrm{Pe} = v_{\mathrm{f}}\sigma_{\mathrm{f}}L/\lambda_{\mathrm{f}}$ wird als Péclet-Zahl bezeichnet und misst das Verhältnis zwischen konvektiv transportierter und geleiteter Wärmemenge. Wir lassen ab jetzt die Tilde weg und schreiben (1.17) in der Form

$$d\frac{\partial u}{\partial t} - \nabla \cdot (a\nabla u) + b \cdot \nabla u + cu = f \quad \text{in } \Omega \quad \text{mit } c = \nabla \cdot b. \tag{1.18}$$

Die Randbedingung (1.15) und die Anfangsbedingung (1.14) lassen sich in analoger Weise zu

$$a\frac{\partial u}{\partial n} + h(u - u_{\mathrm{a}}) = 0 \quad \text{auf } \Gamma \tag{1.19}$$

und

$$u(x, 0) = u_{\mathrm{i}}(x) \tag{1.20}$$

umformen. Hier ist $h = \mathrm{Bi}\,\kappa/\kappa_{\mathrm{f}}$, wobei $\mathrm{Bi} = L\kappa_{\mathrm{f}}/\lambda_{\mathrm{f}}$ als Biot-Zahl bezeichnet wird.

Die partielle Differentialgleichung (1.18) wird zusammen mit der Anfangsbedingung (1.20) und der Randbedingung (1.19) als *Anfangs-Randwertproblem* bezeichnet. Der Term $-\nabla \cdot (a\nabla u)$ wird in *Divergenzform* angegeben. Diese Form entsteht bei der Herleitung der Gleichung auf natürliche Weise und ist, wie wir später sehen werden, für den Großteil der mathematischen Analyse zweckmäßig. Dennoch führen wir die Ableitung manchmal aus und schreiben die Gleichung in Nichtdivergenzform:

$$d\frac{\partial u}{\partial t} - a\Delta u + \bar{b} \cdot \nabla u + cu = f \quad \text{mit } \bar{b} = b - \nabla \cdot a. \tag{1.21}$$

Einige vereinfachte Probleme

Es ist nützlich, verschiedene Vereinfachungen obiger Gleichungen zu untersuchen, da die mathematische Analyse in diesen Fällen möglicherweise weiter führt als im allgemeinen Fall. Wenn wir annehmen, dass die Koeffizienten mit $b = 0$, $c = 0$ konstant sind, reduziert sich (1.18) auf

$$\frac{\partial u}{\partial t} - a\Delta u = f. \tag{1.22}$$

(Es sei daran erinnert, dass für konstantes σ die Beziehung $d = 1$ gilt.) Für $a = 1$ ist dies Gleichung (1.6). Wenn f und die Randbedingung unabhängig von t sind, dann könnte man erwarten, dass sich u mit wachsendem t einem stationären Zustand nähert, d. h. $u(x, t) \to v(x)$ für $t \to \infty$. Da dann $\partial u/\partial t \to 0$ gelten würde, stellen wir fest, dass v die *Poisson-Gleichung* (1.1) erfüllt. Wenn zusätzlich noch $f = 0$ gilt, liegt die *Laplace-Gleichung*

$$-\Delta u = 0$$

vor. Die Lösungen der Laplace-Gleichung werden als *harmonische Funktionen* bezeichnet.

Eine andere wichtige Art der Vereinfachung wird durch Dimensionsreduktion erreicht. Betrachten wir beispielsweise die stationäre (zeitunabhängige) Wärmeleitungsgleichung für einen (nicht notwendigerweise kreisförmigen) Zylinder mit isolierter Mantelfläche, der entlang der x_1-Achse ausgerichtet ist. Wenn die Koeffizienten a, b, c, f in (1.18) unabhängig von x_2 und x_3 sind, dann ist die Annahme vernünftig, dass die Lösung u ebenfalls allein von der Variable x_1 abhängt, die wir dann mit x bezeichnen, d. h. $u = u(x)$. Die

Wärmeleitungsgleichung (1.18) reduziert sich dann auf die gewöhnliche Differentialgleichung

$$-(au')' + bu' + cu = f \quad \text{in } \Omega = (0,1).$$

Die Randbedingung (1.19) wird zu

$$-a(0)u'(0) + h_0(u(0) - u_0) = 0, \quad a(1)u'(1) + h_1(u(1) - u_1) = 0. \tag{1.23}$$

Wir bezeichnen dies als *Zweipunkt-Randwertproblem*. Ähnliche Vereinfachungen erhält man bei Zylinder- und Kugelsymmetrie, wenn man die Gleichungen in Zylinder- beziehungsweise Kugelkoordinaten aufschreibt. Wenn die Koeffizienten konstant sind, können wir die Lösung leicht durch wohlbekannte spezielle Funktionen ausdrücken (siehe Problemstellung 1.6).

Nichtlineare Gleichungen, Linearisierung

Die Koeffizienten in der Wärmeleitungsgleichung (1.18) und in den Randbedingungen hängen häufig von der Temperatur u ab, wodurch die Gleichungen nichtlinear werden. Die Untersuchung nichtlinearer Gleichungen geht über den Umfang dieses Buches hinaus. Es sei jedoch erwähnt, dass die Untersuchung nichtlinearer Gleichungen häufig mit einer *Linearisierung* beginnt, d. h. durch Rückführung auf die Untersuchung der zugehörigen linearen Gleichungen. Wir illustrieren dies am Beispiel der Gleichung

$$F(u) := \frac{\partial u}{\partial t} - \nabla \cdot \big(a(u)\nabla u\big) - f(u) = 0 \quad \text{in } \Omega \quad \text{für } t > 0,$$

die von der Form (1.18) ist und zusammen mit geeigneten Anfangs- und Randbedingungen gelöst werden muss. Ein Lösungsansatz besteht in der Anwendung der Newtonschen Methode, die eine Folge von approximativen Lösungen u^k ausgehend von einer Anfangslösung u^0 in der folgenden Weise erzeugt: Ist u^k gegeben, wollen wir einen Zuwachs bestimmen, für den $u^{k+1} = u^k + v^k$ eine bessere Approximation für die exakte Lösung darstellt als u^k. Approximieren wir $F(u^{k+1}) = 0$ durch $F(u^k) + F'(u^k)v^k = 0$, so erhalten wir eine linearisierte Gleichung

$$\frac{\partial v^k}{\partial t} - \nabla \cdot \big(a(u^k)\nabla v^k\big) - \nabla \cdot \big(a'(u^k)\nabla u^k v^k\big) - f'(u^k)v^k = -F(u^k) \quad \text{in } \Omega,$$

die zusammen mit einer Anfangsbedingung und linearisierten Randbedingungen gelöst wird. Diese Gleichung ist eine lineare Gleichung in v^k von der Form (1.18), wobei die neuen Koeffizienten $a(u^k(x,t))$ usw. von x und t abhängen.

1.4 Problemstellungen

Problem 1.1. (Herleitung der Konvektions-Diffusionsgleichung.) Bezeichne $c = c(x,t)$ `[mol/m`^{3}`]` die Konzentration einer Substanz am Punkt x `[m]` und

zur Zeit t [s], die durch Konvektion und Diffusion durch ein Gebiet $x \in \Omega \subset \mathbf{R}^3$ transportiert wird. Der Konvektionsstrom ist

$$j_c = vc \quad [\mathrm{mol/(m^2 s)}]$$

mit dem konvektiven Geschwindigkeitsfeld $v = v(x)$ [m/s]. Der durch Konvektion erzeugte Strom ist (Fickssches Gesetz)

$$j_d = -D\nabla c \quad [\mathrm{mol/(m^2 s)}]$$

mit dem Diffusionskoeffizienten $D = D(x)$ [$\mathrm{m^2/s}$]. Sei r [$\mathrm{mol/(m^3 s)}$] die Bildungs-/Zerfallsrate des Materials, beispielsweise durch chemische Reaktion. Die Gesamtmasse der Substanz in einem beliebigen Teilgebiet ist $\int_{\Omega_0} c \, \mathrm{d}x$. Benutzen Sie die Massenerhaltung und das Divergenztheorem, um die Konvektions-Diffusionsgleichung

$$\frac{\partial c}{\partial t} - \nabla \cdot (D\nabla c) + \nabla \cdot (vc) = r \quad [\mathrm{mol/(m^3 s)}]$$

herzuleiten. Diese hat die gleiche mathematische Form wie (1.13). Leiten Sie eine Randbedingung der Form (1.15) her. Zeigen Sie, dass diese Gleichungen in der gleichen dimensionslosen Form wie (1.18) und (1.19) geschrieben werden können.

Problem 1.2. (Herleitung der Wellengleichung.) Betrachten Sie die longitudinale Bewegung eines elastischen Stabes der Länge L [m] mit konstantem Querschnitt A [$\mathrm{m^2}$] und der Dichte ρ [$\mathrm{kg/m^3}$]. Bezeichne $u = u(x,t)$ [m] die Verschiebung zur Zeit t [s] eines Querschnittelementes, das sich ursprünglich an der Stelle $x \in [0, L]$ befand. Die Newtonsche Bewegungsgleichung besagt, dass

$$\frac{d}{dt} \int_a^b pA \, \mathrm{d}x = \big(\sigma(b) - \sigma(a)\big) A \quad [\mathrm{N}]$$

gilt, wobei $\int_a^b pA \, \mathrm{d}x$ [kg m/s] der Gesamtimpuls eines beliebigen Segmentes (a, b) und σ [$\mathrm{N/m^2}$] die mechanische Spannung (Kraft pro Flächeneinheit) ist. Dies führt auf

$$\frac{\partial p}{\partial t} = \frac{\partial \sigma}{\partial x}.$$

Für kleine Verschiebungen gibt es eine lineare Beziehung zwischen der mechanischen Spannung σ und der Dehnung $\epsilon = \partial u / \partial x$, nämlich das Hookesche Gesetz

$$\sigma = E\epsilon,$$

wobei E [$\mathrm{N/m^2}$] das Elastizitätsmodul und die Impulsdichte durch $p = \rho \partial u / \partial t$ gegeben ist. Zeigen Sie, dass u die Wellengleichung

$$\rho \frac{\partial^2 u}{\partial t^2} = \frac{\partial}{\partial x}\Big(E \frac{\partial u}{\partial x}\Big)$$

erfüllt. Diskutieren Sie verschiedene mögliche Randbedingungen an den Enden des Stabes, beispielsweise an der Stelle $x = L$:

– festes Ende, $u(L) = 0$,
– freies Ende, $\sigma(L) = 0$, was auf $u_x(L) = 0$ führt,
– elastische Aufhängung $\sigma(L) = -ku(L)$, was auf $Eu_x(L) + ku(L) = 0$ führt.

Beachten Sie, dass diese Randbedingungen von der Form (1.23) sind.

Problem 1.3. (Elastischer Biegebalken.) Betrachten Sie die Krümmung eines elastischen Biegebalkens, der sich über das Intervall $0 \le x \le L$ erstreckt. An einem beliebigen Querschnittelement in einer Entfernung x vom linken Ende legen wir ein Biegemoment (Drehmoment) $M = M(x)$ `[Nm]`, eine transversale Kraft $T = T(x)$ `[N]` und eine äußere Kraft $q = q(x)$ pro Längeneinheit `[N/m]` an. Man kann zeigen, dass für ein Kräftegleichgewicht $M' = T$ und $T' = -q$ erforderlich ist. Sei $u = u(x)$ `[m]` die kleine transversale Auslenkung des Balkens. Der Biegewinkel beträgt dann etwa u'. Das phänomenologische Gesetz lautet $M = -EIu''$, wobei E `[N/m²]` das Elastizitätsmodul und I `[m⁴]` ein Trägheitsmoment des Balkenquerschnitts ist. Zeigen Sie, dass dies auf die Gleichung vierter Ordnung

$$(EIu'')'' = q$$

führt. Diskutieren Sie verschiedene mögliche Randbedingungen an den Enden des Balkens, beispielsweise an der Stelle $x = L$:

– eingespanntes Ende, $u(L) = 0$, $u'(L) = 0$,
– freies Ende, $M(L) = -(EIu'')(L) = 0$, $T(L) = -(EIu'')'(L) = 0$,
– Gelenk, $u'(L) = 0$, $M(L) = -(EIu'')(L) = 0$.

Problem 1.4. (Der Laplace-Operator bei Kugelsymmetrie.) Führen Sie Kugelkoordinaten (r, θ, ϕ) ein, die durch $x_1 = r\sin\theta\cos\phi$, $x_2 = r\sin\theta\sin\phi$, $x_3 = r\cos\theta$ definiert sind. Nehmen Sie an, dass die Funktion u nicht von θ und ϕ abhängt, d. h. dass $u = u(\rho)$ gilt. Zeigen Sie die Gültigkeit der Gleichung

$$\Delta u = \frac{1}{r^2}\frac{d}{dr}\Big(r^2\frac{du}{dr}\Big).$$

Problem 1.5. (Der Laplace-Operator bei Zylindersymmetrie.) Führen Sie Zylinderkoordinaten (ρ, φ, z) ein, die durch $x_1 = \rho\cos\varphi$, $x_2 = \rho\sin\varphi$, $x_3 = z$ definiert sind. Nehmen Sie an, dass die Funktion u nicht von φ und z abhängt, d. h. dass $u = u(\rho)$ ist. Zeigen Sie die Gültigkeit der Gleichung

$$\Delta u = \frac{1}{\rho}\frac{d}{d\rho}\Big(\rho\frac{du}{d\rho}\Big).$$

Problem 1.6. Sei $\Omega = \{x \in \mathbf{R}^3 : |x| < 1\}$. Bestimmen Sie eine explizite Lösung des Randwertproblems

$$-\Delta u + c^2 u = f \quad \text{in } \Omega \quad \text{mit } u = g \quad \text{auf } \Gamma$$

unter der Annahme, dass Kugelsymmetrie vorliegt und c, f, g Konstanten sind. Das heißt, lösen Sie

$$-(r^2u'(r))' + c^2r^2u(r) = r^2f \quad \text{für } r \in (0,1) \quad \text{mit } u(1) = g,\ u(0) \text{ endlich.}$$

Hinweis: Setzen Sie $v(r) = ru(r)$.

2

Ein Zweipunkt-Randwertproblem

Um die Herleitung von Randwertproblemen für elliptische partielle Differentialgleichungen vorzubereiten, betrachten wir hier ein einfaches Zweipunkt-Randwertproblem für eine lineare, gewöhnliche Differentialgleichung zweiter Ordnung. Im ersten Abschnitt leiten wir ein Maximumprinzip für dieses Problem ab und benutzen es, um die Eindeutigkeit und die kontinuierliche Abhängigkeit von den Daten zu zeigen. Im zweiten Abschnitt konstruieren wir für einen Spezialfall eine Greensche Funktion und zeigen, wie sich daraus die Existenz einer Lösung ergibt. Im dritten Abschnitt schreiben wir dieses Problem in Variationsform und benutzen diese zusammen mit einfachen Hilfsmitteln aus der Funktionalanalysis, um die Existenz, die Eindeutigkeit und die stetige Abhängigkeit von den Daten zu beweisen.

2.1 Das Maximumprinzip

Wir betrachten das Randwertproblem

$$\begin{aligned} \mathcal{A}u := -(au')' + bu' + cu = f \quad &\text{in } \Omega = (0,1), \\ u(0) = u_0,\ u(1) &= u_1, \end{aligned} \tag{2.1}$$

wobei die Koeffizienten $a = a(x)$, $b = b(x)$ und $c = c(x)$ glatte Funktionen mit

$$a(x) \geq a_0 > 0, \quad c(x) \geq 0 \quad \text{für } x \in \bar{\Omega} = [0,1] \tag{2.2}$$

sind und die Funktion $f = f(x)$ und die Zahlen u_0, u_1 gegeben sind (siehe Abschnitt 1.3).

Im Spezialfall $a = 1,\ b = c = 0$ reduziert sich dies auf

$$-u'' = f \quad \text{in } \Omega \quad \text{mit } u(0) = u_0,\ u(1) = u_1. \tag{2.3}$$

Durch zweimalige Integration dieser Differentialgleichung stellen wir fest, dass eine Lösung die Form

$$u(x) = -\int_0^x \int_0^y f(s)\,\mathrm{d}s\,\mathrm{d}y + \alpha x + \beta \tag{2.4}$$

haben muss, wobei die Konstanten α, β zu bestimmen sind. Die Wahl $x = 0$ und $x = 1$ führt auf

$$\alpha = u_1 - u_0 + \int_0^1 \int_0^y f(s)\,\mathrm{d}s\,\mathrm{d}y, \quad \beta = u_0.$$

Umgekehrt stellen wir fest, dass (2.4) mit diesen Konstanten α, β die eindeutige Lösung von (2.3) ist.

Im Spezialfall $f = 0$ ist die Lösung von (2.3) die lineare Funktion $u(x) = u_0(1-x) + u_1 x$. Insbesondere liegen die Werte dieser Funktion zwischen den Werten an den Stellen $x = 0$ und $x = 1$ und das Maximum oder Minimum der Funktion befindet sich folglich an den Randpunkten des Intervalls Ω. Allgemein gilt für (2.1) das folgende Maximumprinzip (Minimumprinzip).

Theorem 2.1. *Betrachten wir den Differentialoperator $\mathcal{A}$ in* (2.1) *und nehmen $u \in \mathcal{C}^2 = \mathcal{C}^2(\bar{\Omega})$ und*

$$\mathcal{A}u \le 0 \quad \Big(\mathcal{A}u \ge 0\Big) \quad \text{in } \Omega \tag{2.5}$$

an.
(i) *Für $c = 0$ gilt*

$$\max_{\bar{\Omega}} u = \max\big\{u(0), u(1)\big\} \quad \Big(\min_{\bar{\Omega}} u = \min\big\{u(0), u(1)\big\}\Big). \tag{2.6}$$

(ii) *Für $c \ge 0$ in Ω gilt*

$$\max_{\bar{\Omega}} u \le \max\big\{u(0), u(1), 0\big\} \quad \Big(\min_{\bar{\Omega}} u \ge \min\big\{u(0), u(1), 0\big\}\Big). \tag{2.7}$$

Im Fall (i) schlussfolgern wir, dass das Maximum von u am Rand angenommen wird, d. h. in einem der Randpunkte des Intervalls Ω. Im Fall (ii) ziehen wir den gleichen Schluss, wenn das Maximum nichtnegativ ist. Dies schließt aber die Möglichkeit nicht aus, dass das Maximum im Inneren von Ω angenommen wird. Es gibt jedoch auch eine stärkere Form des Maximumprinzips, das im Fall (i) folgendermaßen lautet: Wenn (2.5) gilt und u ein Maximum (im Fall (ii) ein nichtnegatives inneres Maximum) in einem inneren Punkt von Ω besitzt, dann ist u in $\bar{\Omega}$ konstant. Wir werden dies hier nicht beweisen, verweisen allerdings auf Abschnitt 3.3 mit dem entsprechenden Resultat für harmonische Funktionen. Die in Klammern angegebenen Varianten können mit $\mathcal{A}u \ge 0$ als Minimumprinzip angesehen werden, das sich auf ein Maximumprinzip zurückführen lässt, wenn wir $-u$ betrachten.

Beweis. (i) Nehmen wir zunächst $\mathcal{A}u < 0$ in Ω anstelle von (2.5) an. Wenn u ein Maximum in einem inneren Punkt $x_0 \in \Omega$ besitzt, dann gilt an diesem

Punkt $u'(x_0) = 0$ und $u''(x_0) \le 0$. Deshalb gilt $\mathcal{A}u(x_0) \ge 0$, was allerdings unserer Annahme widerspricht. Folglich kann u kein inneres Maximum sein, woraus (2.6) folgt.

Gehen wir nun davon aus, dass wir nur wissen, dass $\mathcal{A}u \le 0$ in Ω ist. Sei ϕ eine Funktion, für die $\phi \ge 0$ in $\bar{\Omega}$ und $\mathcal{A}\phi < 0$ in Ω gilt. Dafür können wir beispielsweise die Funktion $\phi(x) = \mathrm{e}^{\lambda x}$ benutzen, wobei λ so groß ist, dass $\mathcal{A}\phi = (-a\lambda^2 + (b - a')\lambda)\phi < 0$ in $\bar{\Omega}$ ist. Nehmen wir nun an, dass die Funktion u ihr Maximum in einem inneren Punkt x_0, nicht aber in $x = 0$ oder $x = 1$ annimmt. Dann gilt dies für hinreichend kleine $\epsilon > 0$ auch für $v = u + \epsilon\phi$. Es gilt aber $\mathcal{A}v = \mathcal{A}u + \epsilon\mathcal{A}\phi < 0$ in $\bar{\Omega}$, was dem ersten Teil des Beweises widerspricht.

(ii) Im Falle $u \le 0$ in Ω ist (2.7) trivialerweise erfüllt. Anderenfalls nehmen wir $\max_{\bar{\Omega}} u = u(x_0) > 0$ und $x_0 \ne 0, 1$ an. Sei (α, β) das größte Teilintervall von Ω, das x_0 enthält und in dem $u > 0$ ist. Nun gilt $\tilde{\mathcal{A}}u := \mathcal{A}u - cu \le 0$ in (α, β). Wird Teil (i) mit dem Operator $\tilde{\mathcal{A}}$ im Intervall (α, β) angewendet, folgt daraus $u(x_0) = \max\{u(\alpha), u(\beta)\}$. Dann können aber α und β nicht beide innere Punkte von Ω sein, da sonst entweder $u(\alpha)$ oder $u(\beta)$ positiv und das Intervall (α, β) mit $u > 0$ nicht so groß wie möglich wäre. Daraus folgt $u(x_0) = \max\{u(0), u(1)\}$ und somit auch (2.7). □

Als Konsequenz dieses Theorems gilt in der Notation aus Abschnitt 1.2 folgende Stabilitätsabschätzung bezüglich der Maximumnorm.

Theorem 2.2. *Sei $\mathcal{A}$ wie in* (2.1) *und* (2.2) *definiert. Im Falle $u \in \mathcal{C}^2$ gilt*

$$\|u\|_{\mathcal{C}} \le \max\big\{|u(0)|, |u(1)|\big\} + C\|\mathcal{A}u\|_{\mathcal{C}}.$$

Die Konstante C hängt von den Koeffizienten von $\mathcal{A}$, aber nicht von u ab.

Beweis. Wir werden nun die Maxima von $\pm u$ abschätzen. Dazu setzen wir $\phi(x) = \mathrm{e}^{\lambda} - \mathrm{e}^{\lambda x}$ und definieren die beiden Funktionen

$$v_{\pm}(x) = \pm u(x) - \|\mathcal{A}u\|_{\mathcal{C}}\, \phi(x).$$

Weil $\phi \ge 0$ in Ω und $\mathcal{A}\phi = c\mathrm{e}^{\lambda} + (a\lambda^2 + (a' - b)\lambda - c)\mathrm{e}^{\lambda x} \ge 1$ in $\bar{\Omega}$ ist, wenn $\lambda > 0$ hinreichend groß gewählt ist, gilt mit einer solchen Wahl von λ

$$\mathcal{A}v_{\pm} = \pm\mathcal{A}u - \|\mathcal{A}u\|_{\mathcal{C}}\mathcal{A}\phi \le \pm\mathcal{A}u - \|\mathcal{A}u\|_{\mathcal{C}} \le 0 \quad \text{in } \Omega.$$

Theorem 2.1(ii) führt deshalb auf

$$\begin{aligned}\max_{\bar{\Omega}}(v_{\pm}) &\le \max\big\{v_{\pm}(0), v_{\pm}(1), 0\big\} \\ &\le \max\big\{\pm u(0), \pm u(1), 0\big\} \le \max\big\{|u(0)|, |u(1)|\big\},\end{aligned}$$

weil $v_{\pm}(x) \le \pm u(x)$ für alle x gilt. Folglich ist

$$\begin{aligned}\max_{\bar{\Omega}}(\pm u) &= \max_{\bar{\Omega}}\big(v_{\pm} + \|\mathcal{A}u\|_{\mathcal{C}}\,\phi\big) \le \max_{\bar{\Omega}}(v_{\pm}) + \|\mathcal{A}u\|_{\mathcal{C}}\|\phi\|_{\mathcal{C}} \\ &\le \max\big\{|u(0)|, |u(1)|\big\} + C\|\mathcal{A}u\|_{\mathcal{C}} \quad \text{mit } C = \|\phi\|_{\mathcal{C}},\end{aligned}$$

was den Beweis abschließt. □

Aus Theorem 2.2 schließen wir unmittelbar auf die Eindeutigkeit der Lösung von (2.1). Wenn u und v zwei verschiedene Lösungen wären, dann würde deren Differenz $w = u - v$ die Gleichungen $\mathcal{A}w = 0$, $w(0) = w(1) = 0$ und folglich $\|w\|_{\mathcal{C}} = 0$ erfüllen, sodass $u = v$ sein muss.

Allgemeiner gesagt, gilt

$$\|u - v\|_{\mathcal{C}} \leq \max\left\{|u_0 - v_0|, |u_1 - v_1|\right\} + C\|f - g\|_{\mathcal{C}},$$

wenn u und v zwei Lösungen von (2.1) mit den rechten Seiten f beziehungsweise g und den Randwerten u_0, u_1 beziehungsweise v_0, v_1 sind. Folglich ist das Problem (2.1) stabil, d. h. eine kleine Veränderung der Daten verursacht keine große Veränderung der Lösung.

Als weitere Anwendung des Maximumprinzips beobachten wir, dass die Lösung nichtpositiv ist, wenn alle Daten des Randwertproblems (2.1) nichtpositiv sind. Das heißt, wenn $f \leq 0$ und $u_0, u_1 \leq 0$ gilt, dann ist $u \leq 0$. Mithilfe der stärkeren Variante des im Anschluss an Theorem 2.1 erwähnten Maximumprinzips, könnten wir sogar schlussfolgern, dass $u < 0$ in Ω ist, wenn nicht $u(x) \equiv 0$ gilt. Allgemeiner gesagt, liegt folgende *Monotonieeigenschaft* vor: Wenn

$$\begin{aligned} \mathcal{A}u = f \quad &\text{in } \Omega \qquad \text{mit } u(0) = u_0,\ u(1) = u_1, \\ \mathcal{A}v = g \quad &\text{in } \Omega \qquad \text{mit } v(0) = v_0,\ v(1) = v_1 \end{aligned}$$

ist und $f \leq g$, $u_0 \leq v_0$ und $u_1 \leq v_1$ gelten, dann ist $u \leq v$.

2.2 Greensche Funktion

Wir betrachten nun das Problem (2.1) mit $b = 0$ und den Randwerten $u_0 = u_1 = 0$. Wir werden eine Darstellung der Lösung in Form einer sogenannten Greenschen Funktion $G(x, y)$ ableiten. Dazu seien U_0 und U_1 zwei Lösungen der homogenen Gleichung, sodass

$$\begin{aligned} \mathcal{A}U_0 = 0 \quad &\text{in } \Omega \qquad \text{mit } U_0(0) = 1,\ U_0(1) = 0, \\ \mathcal{A}U_1 = 0 \quad &\text{in } \Omega \qquad \text{mit } U_1(0) = 0,\ U_1(1) = 1 \end{aligned}$$

gilt. Um uns von der Existenz einer solchen Lösung zu überzeugen, stellen wir fest, dass das Anfangswertproblem für $\mathcal{A}u = 0$ mit $u(0) = 0,\ u'(0) = 1$ aufgrund der Standardtheorie der gewöhnlichen Differentialgleichungen eine eindeutige Lösung besitzt und dass für diese Lösung $u(1) \neq 0$ gilt, weil mit Theorem 2.2 anderenfalls $u(x) \equiv 0$ in Ω gelten würde. Durch Multiplikation dieser Lösung mit einer geeigneten Konstante erhalten wir die gewünschte Funktion U_1. Die Funktion U_0 wird von $x = 1$ ausgehend analog konstruiert. Wegen Theorem 2.1 sind U_0 und U_1 nichtnegativ. Im Falle $b \neq 0$ verweisen wir auf Problemstellung 2.5.

Theorem 2.3. *Sei $b = 0$ und seien U_0, U_1 wie oben beschrieben. Dann ist eine Lösung von* (2.1) *mit $u_0 = u_1 = 0$ durch*

$$u(x) = \int_0^1 G(x,y) f(y)\,\mathrm{d}y \tag{2.8}$$

gegeben, wobei

$$G(x,y) = \begin{cases} \dfrac{1}{\kappa} U_0(x) U_1(y) & \text{für } 0 \le y \le x \le 1, \\[2ex] \dfrac{1}{\kappa} U_1(x) U_0(y) & \text{für } 0 \le x \le y \le 1 \end{cases}$$

und

$$\kappa = a(x)\big(U_0(x)U_1'(x) - U_0'(x)U_1(x)\big) \equiv \text{konstant} > 0 \tag{2.9}$$

ist.

Beweis. Wir beginnen mit dem Beweis, dass κ konstant ist: Wegen $(a\,U_j')' = c\,U_j$ gilt

$$\kappa' = U_0(a\,U_1')' - U_1(a\,U_0')' = U_0\,c\,U_1 - U_1\,c\,U_0 = 0.$$

Durch Festlegen von $x = 0$ bestimmen wir $\kappa = a(0)\,U_1'(0) \neq 0$, da anderenfalls $U_1(0) = U_1'(0) = 0$ und folglich $U_1(x) \equiv 0$ wäre. Da U_1 nichtnegativ ist, ist $U_1'(0)$ nichtnegativ und somit folgt $\kappa > 0$.

Offensichtlich erfüllt u, wie in (2.8) definiert, die homogenen Randbedingungen. Um zu zeigen, dass es sich dabei um eine Lösung der Differentialgleichung handelt, schreiben wir

$$\begin{aligned} u(x) &= \int_0^x G(x,y) f(y)\,\mathrm{d}y + \int_x^1 G(x,y) f(y)\,\mathrm{d}y \\ &= \frac{1}{\kappa} U_0(x) \int_0^x U_1(y) f(y)\,\mathrm{d}y + \frac{1}{\kappa} U_1(x) \int_x^1 U_0(y) f(y)\,\mathrm{d}y. \end{aligned}$$

Durch Differentiation folgt

$$\begin{aligned} u'(x) = \frac{1}{\kappa}\Big(U_0'(x) \int_0^x U_1(y) f(y)\,\mathrm{d}y + U_0(x)U_1(x) f(x)\Big) \\ + \frac{1}{\kappa}\Big(U_1'(x) \int_x^1 U_0(y) f(y)\,\mathrm{d}y - U_1(x)U_0(x) f(x)\Big), \end{aligned}$$

wobei sich die Terme mit $f(x)$ jeweils aufheben. Durch Multiplikation mit $-a(x)$ und Differentiation erhalten wir daher mit $(a\,U_j')' = c\,U_j$ und (2.9)

$$\begin{aligned} -(a(x)u'(x))' = -\frac{1}{\kappa}(a(x)U_0'(x))' \int_0^x U_1(y) f(y)\,\mathrm{d}y \\ - \frac{1}{\kappa}(a(x)U_1'(x))' \int_x^1 U_0(y) f(y)\,\mathrm{d}y \\ - \frac{1}{\kappa} a(x)\Big(U_0'(x)U_1(x) - U_1'(x)U_0(x)\Big) f(x) \end{aligned}$$

$$
\begin{aligned}
&= -\frac{1}{\kappa} c(x) U_0(x) \int_0^x U_1(y) f(y)\,\mathrm{d}y \\
&\quad - \frac{1}{\kappa} c(x) U_1(x) \int_x^1 U_0(y) f(y)\,\mathrm{d}y + f(x) \\
&= -c(x) \int_0^1 G(x,y) f(y)\,\mathrm{d}y + f(x) = -c(x)u(x) + f(x),
\end{aligned}
$$

was den Beweis abschließt. □

Insbesondere zeigt dieses Theorem die Existenz einer Lösung des betrachteten Problems. Aus Abschnitt 2.1 wissen wir bereits, dass die Lösung eindeutig ist. Die Darstellung der Lösung als Integral bezüglich der Greenschen Funktion kann ebenfalls dazu verwendet werden, zusätzliche Informationen bezüglich der Lösung zu gewinnen. Als einfaches Beispiel betrachten wir die Maximumnorm-Abschätzung

$$
\|u\|_{\mathcal{C}} \le C \|f\|_{\mathcal{C}} \quad \text{mit } C = \max_{x \in \bar{\Omega}} \int_0^1 G(x,y)\,\mathrm{d}y, \tag{2.10}
$$

die einen genaueren Wert der Konstante in Theorem 2.2 liefert. Dabei haben wir die Tatsache benutzt, dass U_0 und U_1 und somit auch G wegen Theorem 2.1 nichtnegativ sind.

Theorem 2.3 kann auch dazu verwendet werden, die Existenz einer Lösung im Falle allgemeiner Randwerte u_0 und u_1 zu zeigen. Das bedeutet, wenn $\bar{u}(x) = u_0(1-x) + u_1 x$ gilt und v eine Lösung von

$$
\mathcal{A}v = g := f - \mathcal{A}\bar{u} \quad \text{in } \Omega \quad \text{mit } v(0) = v(1) = 0
$$

ist, dann erfüllt $u = v + \bar{u}$ die Gleichungen $\mathcal{A}u = f$ und $u(0) = u_0,\ u(1) = u_1$.

2.3 Variationsformulierung

Wir werden nun unser Zweipunkt-Randwertproblem im Hilbert-Raum $L_2 = L_2(\Omega)$ betrachten und eine sogenannte Variationsformulierung dafür ableiten. Zur Einführung in die verwendeten funktionalanalytischen Konzepte verweisen wir auf Anhang A.

Wir betrachten das Randwertproblem (2.1) mit homogenen Randbedingungen, d. h.

$$
\mathcal{A}u := -(au')' + bu' + cu = f \quad \text{in } \Omega = (0,1) \quad \text{mit } u(0) = u(1) = 0. \tag{2.11}
$$

Wir nehmen an, dass die Koeffizienten a, b und c glatt sind und statt (2.2)

$$
a(x) \ge a_0 > 0, \quad c(x) - b'(x)/2 \ge 0 \quad \text{für } x \in \bar{\Omega} \tag{2.12}
$$

gilt. Multiplizieren wir die Differentialgleichung mit einer Funktion $\varphi \in \mathcal{C}_0^1 = \mathcal{C}_0^1(\Omega)$ und integrieren wir über das Intervall Ω, so erhalten wir

$$\int_0^1 \big(-(au')' + bu' + cu\big)\varphi \,\mathrm{d}x = \int_0^1 f\varphi \,\mathrm{d}x \tag{2.13}$$

oder nach partieller Integration mit $\varphi(0) = \varphi(1) = 0$

$$\int_0^1 (au'\varphi' + bu'\varphi + cu\varphi)\,\mathrm{d}x = \int_0^1 f\varphi \,\mathrm{d}x \quad \forall \varphi \in C_0^1, \tag{2.14}$$

was wir als *variationelle* oder *schwache Formulierung* von (2.11) bezeichnen.

Führen wir die Bilinearform

$$a(v,w) = \int_0^1 (av'w' + bv'w + cvw)\,\mathrm{d}x \tag{2.15}$$

und das lineare Funktional

$$L(w) = (f,w) = \int_0^1 f\, w \,\mathrm{d}x$$

ein und verwenden wir die Tatsache, dass $\mathcal{C}_0^1$ dicht in $H_0^1 = H_0^1(\Omega)$ ist, können wir die Gleichung (2.14) in der Form

$$a(u,\varphi) = L(\varphi) \quad \forall \varphi \in H_0^1 \tag{2.16}$$

schreiben.

Wir sagen, dass u eine *schwache Lösung* von (2.11) ist, wenn $u \in H_0^1$ ist und (2.16) gilt. Folglich müssen wir von einer schwachen Lösung nicht fordern, dass sie zweimal differenzierbar ist. Wenn eine schwache Lösung jedoch zu $\mathcal{C}^2$ gehört, dann handelt es sich tatsächlich um die klassische Lösung von (2.11). Und zwar schlussfolgern wir durch partielle Integration in (2.14), dass (2.13) gilt, d. h. es ist

$$\int_0^1 (\mathcal{A}u - f)\,\varphi \,\mathrm{d}x = 0 \quad \forall \varphi \in H_0^1.$$

Daraus ergibt sich sofort $\mathcal{A}u = f$ in Ω, und weil $u \in H_0^1$ ist, gilt auch $u(0) = u(1) = 0$. Diese Berechnungen können auch für $u \in H^2 \cap H_0^1$ ausgeführt werden. In diesem Falle sagen wir, dass u eine *starke Lösung* von (2.11) ist.

Wir stellen unter Verwendung der Notation aus Abschnitt 1.2 fest, dass

$$\|v\| \le \|v'\| \text{ gilt,} \quad \text{wenn } v(0) = v(1) = 0 \tag{2.17}$$

ist. Tatsächlich gilt wegen der Cauchy-Schwarz-Ungleichung für alle $x \in \Omega$

$$|v(x)|^2 = \Big| \int_0^x v'(y)\,\mathrm{d}y \Big|^2 \le \int_0^x 1^2 \,\mathrm{d}y \int_0^x (v')^2 \,\mathrm{d}y \le x \int_0^1 (v')^2 \,\mathrm{d}y \le \|v'\|^2,$$

woraus durch Integration (2.17) folgt. Dies ist ein Spezialfall der Poincaré-Ungleichung, die auch für Funktionen mehrerer Variablen aufgeschrieben werden kann (siehe Theorem A.6). Es folgt unmittelbar

$$\|v\|_1 = \left(\|v\|^2 + \|v'\|^2\right)^{1/2} \le \sqrt{2}\|v'\| \quad \forall v \in H_0^1, \tag{2.18}$$

was zeigt, dass die Normen $\|v\|_1$ und $|v|_1 = \|v'\|$ äquivalent sind.

Unter Verwendung unserer Annahme (2.12) stellen wir fest, dass

$$\int_0^1 (bv'v + cv^2)\,\mathrm{d}x = \left[\tfrac{1}{2}b\,v^2\right]_0^1 + \int_0^1 (c - \tfrac{1}{2}b')\,v^2\,\mathrm{d}x \ge 0 \quad \text{für } v \in H_0^1$$

gilt. Folglich ergibt sich aus (2.12) und (2.18), dass die Bilinearform $a(v,w)$ die Eigenschaft

$$a(v,v) \ge \min_{x\in\bar{\Omega}} a(x)\,\|v'\|^2 \ge \alpha\|v\|_1^2 \quad \forall v \in H_0^1 \quad \text{mit } \alpha = a_0/2 > 0 \tag{2.19}$$

besitzt. Die Ungleichung (2.19) drückt aus, dass die Bilinearform $a(\cdot,\cdot)$ in H_0^1 *koerzitiv* ist (siehe (A.12)). Setzen wir in (2.16) $\varphi = u$ und verwenden wir (2.19) und (2.17), so finden wir

$$\alpha\|u\|_1^2 \le a(u,u) = (f,u) \le \|f\|\,\|u\| \le \|f\|\,\|u\|_1,$$

sodass

$$\|u\|_1 \le C\|f\| \quad \text{mit } C = 2/a_0 \tag{2.20}$$

gilt. Die Bilinearform $a(v,w)$ ist auf H_0^1 auch in dem Sinne beschränkt, dass

$$|a(v,w)| \le C\|v\|_1\|w\|_1 \quad \forall v,w \in H_0^1 \tag{2.21}$$

gilt (vgl. (A.9)). Durch Abschätzen der Koeffizienten in (2.15) durch ihre Maxima und mithilfe der Cauchy-Schwarz-Ungleichung erhalten wir

$$|a(v,w)| \le C\int_0^1 \left(|v'w'| + |v'w| + |vw|\right)\mathrm{d}x \le C\|v\|_1\|w\|_1.$$

Wir wenden uns nun der Frage der Existenz einer Lösung der Variationsgleichung (2.16) zu.

Theorem 2.4. *Angenommen, es gilt Gleichung* (2.12) *und* $f \in L_2$. *Dann existiert eine eindeutige Lösung* $u \in H_0^1$ *von* (2.16). *Diese Lösung erfüllt* (2.20).

Beweis. Der Beweis basiert auf dem Lax-Milgram-Lemma, Theorem A.3. Wir haben bereits überprüft, dass $a(\cdot,\cdot)$ koerzitiv und in H_0^1 beschränkt ist, weil

$$|L(\varphi)| = |(f,\varphi)| \le \|f\|\,\|\varphi\| \le \|f\|\,\|\varphi\|_1 \quad \forall \varphi \in H_0^1$$

gilt. Folglich sind die Voraussetzungen des Lax-Milgram-Lemmas erfüllt und es folgt, dass ein eindeutiges $u \in H_0^1$ existiert, das Gleichung (2.16) erfüllt. Zusammen mit (2.20) vervollständigt dies den Beweis. □

Wir bemerken, dass die Bilinearform $a(\cdot,\cdot)$ im Falle $b = 0$ symmetrisch, positiv definit und folglich ein Skalarprodukt ist, wobei die zugehörige Norm äquivalent zu $\|\cdot\|_1$ ist. Die Existenz einer eindeutigen Lösung folgt dann aus dem elementareren Rieszschen Darstellungssatz (Theorem A.1).

Im symmetrischen Fall mit $b = 0$ kann die Lösung auch so aufgefasst werden, dass sie ein bestimmtes quadratisches Funktional minimiert (Theorem A.2). Dies ist ein Spezialfall des berühmten Dirichlet-Prinzips.

Theorem 2.5. *Angenommen, es gilt* (2.2) *und* $b = 0$. *Sei* $f \in L_2$ *und* $u \in H_0^1$ *die Lösung von* (2.16). *Wir setzen*

$$F(\varphi) = \tfrac{1}{2}\int_0^1 \left(a(\varphi')^2 + c\varphi^2\right) \mathrm{d}x - \int_0^1 f\varphi \,\mathrm{d}x.$$

Dann ist $F(u) \le F(\varphi)$ *für alle* $\varphi \in H_0^1$, *wobei Gleichheit nur für* $\varphi = u$ *gilt.*

Die in Theorem 2.4 erzielte schwache Lösung u von (2.16) ist sogar regulärer als dort erklärt wurde. Mithilfe unserer Definitionen kann man tatsächlich zeigen, dass u'' als eine schwache Ableitung (siehe (A.21)) existiert und dass $au'' = -f + (b - a')u' + cu \in L_2$ ist. Es folgt $u \in H^2$ und

$$a_0\|u''\| \le \|au''\| \le \|f\| + \|(b-a')u'\| + \|cu\| \le \|f\| + C\|u\|_1 \le C\|f\|.$$

Zusammen mit (2.20) führt dies auf die *Regularitätsabschätzung*

$$\|u\|_2 \le C\|f\|. \tag{2.22}$$

Wir schlussfolgern, dass die in Theorem 2.4 gefundene schwache Lösung von (2.1) sogar eine starke Lösung ist. Der Beweis der H^2-Regularität verwendet die Annahme, dass a glatt und $f \in L_2$ ist. Wenn a weniger glatt oder f nur in H^{-1} ist (siehe (A.30)), erhalten wir immer noch eine schwache Lösung in H_0^1, sie gehört dann allerdings nicht zu H^2 (siehe Problemstellung 2.8).

2.4 Problemstellungen

Problem 2.1. Bestimmen Sie explizite Lösungen des Randwertproblems

$$-u'' + cu = f \quad \text{in } (-1,1) \quad \text{mit } u(-1) = u(1) = g,$$

wobei c, f, g konstant sind. Benutzen Sie diese zur Illustration des Maximumprinzips.

Problem 2.2. Bestimmen Sie die Greenschen Funktionen der folgenden Probleme:

(a) $-u'' = f$ in $\Omega = (0,1)$ mit $u(0) = u(1) = 0$,

(b) $-u'' + cu = f$ in $\Omega = (0,1)$ mit $u(0) = u(1) = 0$.

Problem 2.3. Betrachten Sie das nichtlineare Randwertproblem

$$-u'' + u = \mathrm{e}^u \quad \text{in } \Omega = (0,1) \quad \text{mit } u(0) = u(1) = 0.$$

Zeigen Sie mithilfe des Maximumprinzips, dass alle Lösungen nichtnegativ sind, d. h. dass $u(x) \geq 0$ für alle $x \in \bar{\Omega}$ gilt. Zeigen Sie mithilfe des starken Maximumprinzips, dass alle Lösungen positiv sind, d. h. dass $u(x) > 0$ für alle $x \in \Omega$ gilt.

Problem 2.4. Angenommen, es sei wie in Theorem 2.3 $b = 0$ und $G(x, y)$ die dort definierte Greensche Funktion.

(a) Beweisen Sie, dass G symmetrisch ist, d. h. $G(x, y) = G(y, x)$.
(b) Beweisen Sie die Gültigkeit von

$$a(v, G(x, \cdot)) = v(x) \quad \forall v \in H_0^1,\ x \in \Omega.$$

Das bedeutet, dass $\mathcal{A}G(x, \cdot) = \delta_x$ gilt, wobei δ_x die als lineares Funktional $\delta_x(\phi) = \phi(x)$ für alle $\phi \in \mathcal{C}_0$ definierte Diracsche Deltafunktion an der Stelle x ist (siehe Problemstellung A.9).

Problem 2.5. Im unsymmetrischen Fall $b \neq 0$ ist die Greensche Funktion auf ähnliche Weise wie in Theorem 2.3 definiert:

$$G(x,y) = \begin{cases} \dfrac{U_0(x)U_1(y)}{\kappa(y)} & \text{für } 0 \leq y \leq x \leq 1, \\[2ex] \dfrac{U_1(x)U_0(y)}{\kappa(y)} & \text{für } 0 \leq x \leq y \leq 1. \end{cases}$$

Der Hauptunterschied besteht darin, dass κ nicht mehr konstant ist. Die Funktionen U_0 und U_1 sind linear unabhängig. Somit folgt aus der Theorie der gewöhnlichen Differentialgleichungen, dass deren Wronski-Determinante $U_0U_1' - U_0'U_1$ nicht verschwindet. Wie zuvor können wir dann schlussfolgern, dass $\kappa(x) > 0$ in $\bar{\Omega}$ ist. Wiederholen Sie die Schritte im Beweis von Theorem 2.3 für diesen Fall.

Problem 2.6. Geben Sie Variationsformulierungen an und beweisen Sie die Existenz der Lösungen der Gleichung

$$-u'' = f \quad \text{in } \Omega = (0,1)$$

mit den folgenden Randbedingungen

(a) $u(0) = u(1) = 0$,
(b) $u(0) = u'(1) = 0$,
(c) $-u'(0) + u(0) = u'(1) = 0$.

Problem 2.7. Betrachten Sie die Gleichung eines Biegebalkens aus Problemstellung 1.3

$$\frac{d^4u}{dx^4} = f \quad \text{in } \Omega = (0,1)$$

zusammen mit den Randbedingungen

(a) $u(0) = u'(0) = u(1) = u'(1) = 0$,
(b) $u(0) = u''(0) = u(1) = u''(1) = 0$,
(c) $u(0) = u'(0) = u'(1) = u'''(1) = 0$,
(d) $u(0) = u'(0) = u''(1) = u'''(1) = 0$,
(e) $u(0) = u'(0) = u(1) = u'''(1) = 0$.

Geben Sie Variationsformulierungen an und untersuchen Sie die Existenz und die Eindeutigkeit von Lösungen dieser Probleme. Geben Sie mechanische Interpretationen der Randbedingungen an.

Problem 2.8. Bestimmen Sie eine explizite Lösung von (2.11) mit $a = 1$, $b = c = 0$ und $f(x) = 1/x$. Aus Problemstellung A.11 wissen wir, dass $f \in H^{-1}$ aber $f \notin L_2$ gilt. Überprüfen Sie, dass $u \in H_0^1$ aber $u \notin H^2$ gilt. Hinweis: $u(x) = -x \log x$.

3

Elliptische Gleichungen

In diesem Kapitel untersuchen wir Randwertprobleme für elliptische partielle Differentialgleichungen. Wie wir bereits in Kapitel 1 gesehen haben, spielen diese Gleichungen sowohl in der Theorie als auch in der Anwendung partieller Differentialgleichungen eine zentrale Rolle; sie beschreiben eine große Anzahl physikalischer Phänomene, insbesondere bei der Modellierung stationärer Zustände, und bilden den stationären Grenzfall von Evolutionsgleichungen. Nach einigen Vorbemerkungen in Abschnitt 3.1 beginnen wir in Abschnitt 3.2 mit dem Beweis eines Maximumprinzips. Wie für das Zweipunkt-Randwertproblem in Kapitel 2 kann dieses dazu benutzt werden, die Eindeutigkeit und die stetige Abhängigkeit von den Daten für Randwertprobleme zu zeigen. In Abschnitt 3.3 beweisen wir die Existenz einer Lösung des Dirichlet-Problems für die Poisson-Gleichung auf einer Kreisscheibe mit homogenen Randbedingungen, indem wir eine Integraldarstellung mit einem Poisson-Kern benutzen. In Abschnitt 3.4 werden ähnliche Ideen benutzt, um Fundamentallösungen elliptischer Gleichungen einzuführen. Wir illustrieren deren Verwendung am Beispiel der Konstruktion einer Greenschen Funktion. Ein anderer wichtiger Ansatz, der in Abschnitt 3.5 vorgestellt wird, beruht auf einer Variationsformulierung des Randwertproblems sowie auf einfachen funktionalanalytischen Methoden. In Abschnitt 3.6 diskutieren wir kurz das Neumann-Problem und beschreiben in Abschnitt 3.7 einige regularitätstheoretische Resultate.

3.1 Vorbemerkungen

Anstatt eine allgemeine elliptische Gleichung zweiter Ordnung der Form (1.5) zu betrachten, werden wir uns der Einfachheit halber auf einen Spezialfall beschränken, bei dem sich die Matrix $A = (a_{ij})$ in (1.5) auf ein einfaches skalares Vielfaches $a\,I$ der Einheitsmatrix beschränkt. Dabei ist a eine glatte Funktion.

Wir betrachten zunächst das Dirichlet-Problem

$$\mathcal{A}u := -\nabla \cdot \big(a\nabla u\big) + b \cdot \nabla u + cu = f \quad \text{in } \Omega \quad \text{mit } u = g \quad \text{auf } \Gamma, \tag{3.1}$$

wobei $\Omega \subset \mathbf{R}^d$ ein Gebiet mit hinreichend glattem Rand Γ ist. Die Koeffizienten $a = a(x)$, $b = b(x)$, $c = c(x)$ sind glatt mit

$$a(x) \geq a_0 > 0, \quad c(x) \geq 0 \quad \forall x \in \Omega. \tag{3.2}$$

Dabei sind f und g auf Ω beziehungsweise Γ gegebene Funktionen. Dies ist der stationäre Fall der Wärmeleitungsgleichung (1.18).

Der Spezialfall $a = 1$, $b = 0$, $c = 0$ führt auf die Poisson-Gleichung, d. h.

$$-\Delta u := -\sum_{j=1}^{d} \frac{\partial^2 u}{\partial x_j^2} = f. \tag{3.3}$$

Im Fall $f = 0$ wird diese Gleichung als Laplace-Gleichung bezeichnet, deren Lösungen harmonische Funktionen sind.

Sind v und w Lösungen der beiden Probleme

$$\begin{aligned} \mathcal{A}v = 0 \quad &\text{in } \Omega \qquad \text{mit } v = g \quad \text{auf } \Gamma, \\ \mathcal{A}w = f \quad &\text{in } \Omega \qquad \text{mit } w = 0 \quad \text{auf } \Gamma, \end{aligned}$$

dann stellen wir fest, dass $u = v + w$ ein Lösung von (3.1) ist. Deshalb reicht es manchmal aus, die homogene Gleichung mit gegebenen Randbedingungen und die inhomogene Gleichung mit verschwindenden Randbedingungen separat zu betrachten.

Man kann die partielle Differentialgleichung in (3.1) auch mit Robinschen Randbedingungen betrachten

$$a\frac{\partial u}{\partial n} + h\big(u - g\big) = 0 \quad \text{auf } \Gamma, \tag{3.4}$$

wobei der Koeffizient $h = h(x)$ positiv und n die äußere Normale an Γ ist. Die in (3.1) verwendete Dirichletsche Randbedingung könnte man formal als Grenzfall $h = \infty$ von (3.4) auffassen. Im anderen Grenzfall $h = 0$ erhalten wir die Neumannschen Randbedingungen

$$\frac{\partial u}{\partial n} = 0 \quad \text{auf } \Gamma.$$

Manchmal betrachtet man gemischte Randbedingungen, bei denen beispielsweise auf einem Teil des Randes Dirichletsche Randbedingungen und auf dem verbleibenden Teil Neumannsche Randbedingungen gegeben sind. Eine Funktion $u \in \mathcal{C}^2(\bar{\Omega})$, die die Differentialgleichung und die Randbedingung in (3.1) erfüllt, wird als *klassische Lösung* dieses Randwertproblems bezeichnet.

3.2 Ein Maximumprinzip

Wir beginnen unsere Untersuchung des Dirichlet-Problems (3.1) mit dem Beweis eines Maximumprinzips, das zu demjenigen aus Theorem 2.1 analog ist.

Theorem 3.1. *Betrachten wir den Differentialoperator $\mathcal{A}$ in* (3.1) *und nehmen wir $u \in \mathcal{C}^2 = \mathcal{C}^2(\bar{\Omega})$ sowie*

$$\mathcal{A}u \le 0 \quad \Big(\mathcal{A}u \ge 0\Big) \quad \text{in } \Omega \tag{3.5}$$

an.
(i) *Wenn $c = 0$ ist, dann gilt*

$$\max_{\bar{\Omega}} u = \max_{\Gamma} u \quad \Big(\min_{\bar{\Omega}} u = \min_{\Gamma} u\Big). \tag{3.6}$$

(ii) *Wenn $c \ge 0$ in Ω ist, dann gilt*

$$\max_{\bar{\Omega}} u \le \max\Big\{\max_{\Gamma} u, 0\Big\} \quad \Big(\min_{\bar{\Omega}} u \ge \min\Big\{\min_{\Gamma} u, 0\Big\}\Big). \tag{3.7}$$

Beweis. (i) Sei ϕ eine Funktion mit $\phi \ge 0$ in $\bar{\Omega}$ und $\mathcal{A}\phi < 0$ in Ω. Eine solche Funktion ist beispielsweise $\phi(x) = \mathrm{e}^{\lambda x_1}$, wobei λ so groß ist, dass $\mathcal{A}\phi = (-a\,\lambda^2 + (b_1 - \partial a/\partial x_1)\lambda)\mathrm{e}^{\lambda x_1} < 0$ in $\bar{\Omega}$ gilt. Nehmen wir nun an, dass die Funktion u ihr Maximum in einem inneren Punkt x_0 in Ω, aber nicht auf Γ annimmt. Dann gilt dies für hinreichend kleine ϵ auch für $v = u + \epsilon\phi$. Damit gilt aber $\mathcal{A}v = \mathcal{A}u + \epsilon\mathcal{A}\phi < 0$ in $\bar{\Omega}$. Wenn das Maximum von v andererseits $v(\bar{x}_0)$ ist, dann gilt $\nabla v(\bar{x}_0) = 0$ und folglich $\mathcal{A}v(\bar{x}_0) = -a(\bar{x}_0)\Delta v(\bar{x}_0) \ge 0$. Damit erhalten wir einen Widerspruch. Somit ist unsere Behauptung bewiesen.

(ii) Wenn $u \le 0$ in Ω ist, dann ist (3.7) trivialerweise erfüllt. Anderenfalls nehmen wir $\max_{\bar{\Omega}} u = u(x_0) > 0$ und $x_0 \in \Omega$ an. Sei Ω_0 die größte offene zusammenhängende Teilmenge von Ω, die x_0 enthält, in der $u > 0$ ist. Nun gilt $\tilde{\mathcal{A}}u := \mathcal{A}u - cu \le 0$ in Ω_0. Daraus folgt durch Anwendung von (i) mit dem Operator $\tilde{\mathcal{A}}$ in Ω_0 die Gleichung $u(x_0) = \max_{\Gamma_0} u$, wobei Γ_0 der Rand von Ω_0 ist. Dann kann aber Γ_0 nicht vollständig in der offenen Menge Ω liegen, da es dann einen Punkt auf Γ_0 gäbe, an dem u positiv und Ω_0 mit $u > 0$ nicht so groß wie möglich wäre. Dies beweist (3.7). □

Aus Theorem 3.1 folgt die Stabilität bezüglich der Maximumnorm.

Theorem 3.2. *Sei $u \in \mathcal{C}^2(\bar{\Omega})$. Dann existiert eine Konstante C mit*

$$\|u\|_{\mathcal{C}(\bar{\Omega})} \le \|u\|_{\mathcal{C}(\Gamma)} + C\|\mathcal{A}u\|_{\mathcal{C}(\bar{\Omega})}.$$

Beweis. Sei ϕ eine Funktion, für die $\phi \ge 0$ und $\mathcal{A}\phi \le -1$ in Ω ist. Dies trifft beispielsweise auf ein geeignetes Vielfaches der Funktion ϕ im Beweis von Theorem 3.1 zu. Wir definieren nun die beiden Funktionen $v_{\pm}(x) = \pm u(x) + \|\mathcal{A}u\|_{\mathcal{C}(\bar{\Omega})}\phi(x)$. Dann ist

$$\mathcal{A}v_\pm = \pm\mathcal{A}u + \|\mathcal{A}u\|_{\mathcal{C}(\bar\Omega)}\mathcal{A}\phi \le 0 \quad \text{in } \Omega.$$

Deshalb nehmen beide Funktionen ihre Maxima auf Γ an, sodass

$$\begin{aligned} v_\pm(x) \le \max_\Gamma (v_\pm) &\le \max_\Gamma (\pm u) + \|\mathcal{A}u\|_{\mathcal{C}(\bar\Omega)}\|\phi\|_{\mathcal{C}(\Gamma)} \\ &\le \|u\|_{\mathcal{C}(\Gamma)} + C\|\mathcal{A}u\|_{\mathcal{C}(\bar\Omega)} \quad \text{mit } C = \|\phi\|_{\mathcal{C}(\Gamma)} \end{aligned}$$

gilt. Da $\pm u(x) \le v_\pm(x)$ gilt, ist das Theorem damit bewiesen.

Wie für das Zweipunkt-Randwertproblem folgt daraus, dass es höchstens eine Lösung unseres Dirichlet-Problems (3.1) gibt. Sind $u_j,\ j = 1, 2$ Lösungen von (3.1) mit $f = f_j,\ g = g_j,\ j = 1, 2$, dann gilt

$$\|u_1 - u_2\|_{\mathcal{C}(\bar\Omega)} \le \|g_1 - g_2\|_{\mathcal{C}(\Gamma)} + C\|f_1 - f_2\|_{\mathcal{C}(\bar\Omega)}.$$

3.3 Das Dirichlet-Problem für eine Kreisscheibe. Das Poisson-Integral

In diesem Abschnitt untersuchen wir das Dirichlet-Problem, eine harmonische Funktion auf einer Kreisscheibe $\Omega = \{x \in \mathbf{R}^2 : |x| < R\}$ mit gegebenen Randbedingungen zu bestimmen, d. h. das Problem

$$\begin{aligned} &-\Delta u = 0 && \text{für } |x| < R, \\ &u(R\cos\varphi, R\sin\varphi) = g(\varphi) && \text{für } 0 \le \varphi < 2\pi. \end{aligned} \tag{3.8}$$

Im folgenden Theorem wird die Lösung von (3.8) in Form eines Integrals über den Rand des Kreises angegeben.

Theorem 3.3. (Poissonsche Integralformel.) *Sei $P_R(r, \varphi)$ der Poisson-Kern*

$$P_R(r, \varphi) = \frac{R^2 - r^2}{R^2 + r^2 - 2rR\cos\varphi}.$$

Dann ist die in Polarkoordinaten $x = (r\cos\varphi, r\sin\varphi)$ durch

$$u(x) = \frac{1}{2\pi}\int_0^{2\pi} P_R(r, \varphi - \psi)g(\psi)\,\mathrm{d}\psi \tag{3.9}$$

definierte Funktion für ein hinreichend glattes g eine Lösung von (3.8).

Beweis. Wir stellen zunächst fest, dass die Funktion $v(x) = r^n \mathrm{e}^{\pm in\varphi}$ für jedes $n \ge 0$ harmonisch ist. Tatsächlich gilt also

$$\begin{aligned} \Delta v &= \frac{\partial^2 v}{\partial r^2} + \frac{1}{r}\frac{\partial v}{\partial r} + \frac{1}{r^2}\frac{\partial^2 v}{\partial \varphi^2}, \\ &= \Big(n(n-1)r^{n-2} + \frac{1}{r}nr^{n-1} - \frac{1}{r^2}n^2 r^n\Big)\mathrm{e}^{\pm in\varphi} = 0. \end{aligned}$$

Daraus folgt für ein beschränktes c_n, dass die Reihe

$$u(x) = \sum_{n=-\infty}^{\infty} c_n \Big(\frac{r}{R}\Big)^{|n|} \mathrm{e}^{\mathrm{i}n\varphi} \tag{3.10}$$

in Ω harmonisch ist. Wir nehmen nun an, dass $g(\varphi)$ eine Fourier-Reihe

$$g(\varphi) = \sum_{n=-\infty}^{\infty} c_n \mathrm{e}^{\mathrm{i}n\varphi}$$

besitzt, die absolut konvergent ist. Dann ist die Funktion $u(x)$ in (3.10) mit den Koeffizienten c_n eine Lösung von (3.8), und u ist in $\bar{\Omega}$ stetig. Letzteres bedeutet, dass $u(r\mathrm{e}^{\mathrm{i}\psi}) \to g(\mathrm{e}^{\mathrm{i}\varphi})$ für $r \to R,\ \psi \to \varphi$ gilt. Um uns davon zu überzeugen, wählen wir N so groß, dass $\sum_{|n|>N} |c_n| < \epsilon/3$ gilt und schreiben

$$|u(r\mathrm{e}^{\mathrm{i}\psi}) - g(\mathrm{e}^{\mathrm{i}\varphi})| \le \sum_{|n|\le N} |c_n| \Big| \Big(\frac{r}{R}\Big)^{|n|} \mathrm{e}^{\mathrm{i}n\psi} - \mathrm{e}^{\mathrm{i}n\varphi} \Big| + 2 \sum_{|n|>N} |c_n|.$$

Dabei geht der erste Term auf der rechten Seite für $r \to R,\ \psi \to \varphi$ offensichtlich gegen 0 und wird folglich kleiner als $\epsilon/3$, was unserer Behauptung entspricht.

Erinnern wir uns daran, dass die Fourier-Koeffizienten von g durch

$$c_n = \frac{1}{2\pi} \int_0^{2\pi} \mathrm{e}^{-\mathrm{i}n\psi} g(\psi)\, \mathrm{d}\psi$$

gegeben sind. Folglich gilt

$$u(x) = \frac{1}{2\pi} \int_0^{2\pi} \sum_{n=-\infty}^{\infty} \Big(\frac{r}{R}\Big)^{|n|} \mathrm{e}^{\mathrm{i}n(\varphi-\psi)} g(\psi)\, \mathrm{d}\psi,$$

was mit

$$P_R(r,\varphi) = \sum_{n=-\infty}^{\infty} \Big(\frac{r}{R}\Big)^{|n|} \mathrm{e}^{\mathrm{i}n\varphi}$$

von der Form (3.9) ist. Setzen wir $z = (r/R)\mathrm{e}^{\mathrm{i}\varphi}$, gilt

$$\begin{aligned} P_R(r,\varphi) &= 1 + 2\,\mathrm{Re} \sum_{n=1}^{\infty} \Big(\frac{r}{R}\Big)^{n} \mathrm{e}^{\mathrm{i}n\varphi} \\ &= 2\,\mathrm{Re} \sum_{n=0}^{\infty} z^n - 1 = \mathrm{Re}\,\frac{2}{1-z} - 1 = \mathrm{Re}\,\frac{1+z}{1-z} \\ &= \mathrm{Re}\,\frac{R + r\mathrm{e}^{\mathrm{i}\varphi}}{R - r\mathrm{e}^{\mathrm{i}\varphi}} = \frac{R^2 - r^2}{R^2 + r^2 - 2rR\cos\varphi}, \end{aligned}$$

was den Beweis abschließt. □

Als Konsequenz des Theorems ergibt sich: Wenn u eine harmonische Funktion in Ω, $\tilde{x}$ ein beliebiger Punkt in Ω und die Kreisscheibe $\{x : |x - \tilde{x}| \leq R\}$ in Ω enthalten ist, dann gilt wegen $P_R(0, \varphi) = 1$

$$u(\tilde{x}) = \frac{1}{2\pi} \int_0^{2\pi} u(\tilde{x}_1 + R\cos\psi, \tilde{x}_2 + R\sin\psi)\, \mathrm{d}\psi. \tag{3.11}$$

Somit ist $u(\tilde{x})$ der Mittelwert der Werte von $u(x)$ mit $|x - \tilde{x}| = R$. Folglich entspricht der Wert von u am Mittelpunkt der Kreisscheibe dem Mittelwert der Randwerte der Kreisscheibe. Wir sagen, dass u die Mittelwerteigenschaft erfüllt. Dies beweist einen Spezialfall des starken Maximumprinzips, das wir bereits erwähnt haben: Wenn eine harmonische Funktion u ihr Maximum in einem inneren Punkt von Ω annimmt, dann ist sie konstant. Ist $\tilde{x}$ tatsächlich ein innerer Punkt von Ω, in dem u ihr Maximum annimmt, dann gilt wegen (3.11) die Gleichung $u(x) = u(\tilde{x})$ für alle x mit $\{x : |x - \tilde{x}| = R\} \subset \Omega$. Weil R beliebig und Ω zusammenhängend ist, folgt offensichtlich, dass u in $\bar{\Omega}$ den konstanten Wert $u(\tilde{x})$ annimmt. Insbesondere wird das Maximum auch auf Γ angenommen.

3.4 Fundamentallösungen. Die Greensche Funktion

Sei u eine Lösung der inhomogenen Gleichung

$$\mathcal{A}u = f \quad \text{in } \mathbf{R}^d, \tag{3.12}$$

wobei $\mathcal{A}$ wie in (3.1) mit $b = 0$ ist. Durch Multiplikation mit $\varphi \in \mathcal{C}_0^\infty(\mathbf{R}^d)$, Integration über $\mathbf{R}^d$ und zweimalige partielle Integration erhalten wir

$$(u, \mathcal{A}\varphi) = (f, \varphi) = \int_{\mathbf{R}^d} f(x)\, \varphi(x)\, \mathrm{d}x \quad \forall \varphi \in \mathcal{C}_0^\infty(\mathbf{R}^d). \tag{3.13}$$

Wir sagen, dass U eine Fundamentallösung von (3.12) ist, wenn U für $x \neq 0$ glatt ist, an der Stelle $x = 0$ eine Singularität besitzt, sodass $U \in L_1(B)$ mit $B = \{x \in \mathbf{R}^d : |x| < 1\}$ gilt, und

$$|D^\alpha U(x)| \leq C_\alpha |x|^{2-d-|\alpha|} \quad \text{für } |\alpha| \neq 0 \tag{3.14}$$

sowie

$$(U, \mathcal{A}\varphi) = \varphi(0) \quad \forall \varphi \in \mathcal{C}_0^\infty(\mathbf{R}^d) \tag{3.15}$$

gilt. Im Sinne der schwachen Ableitung (siehe (A.21)) bedeutet dies

$$\mathcal{A}U = \delta,$$

wobei δ die in der Problemstellung A.9 definierte Diracsche Deltafunktion ist.

Wir verwenden nun die Fundamentallösung, um eine Lösung für (3.12) zu konstruieren.

Theorem 3.4. *Wenn U eine Fundamentallösung von* (3.12) *und $f \in \mathcal{C}_0^1(\mathbf{R}^d)$ ist, dann ist*

$$u(x) = (U * f)(x) = \int_{\mathbf{R}^d} U(x-y) f(y)\,\mathrm{d}y$$

eine Lösung von (3.12).

Beweis. Wegen (3.15) gilt

$$\int_{\mathbf{R}^d} U(x-y)\mathcal{A}\varphi(x)\,\mathrm{d}x = \int_{\mathbf{R}^d} U(z)\mathcal{A}\varphi(z+y)\,\mathrm{d}z = (U, \mathcal{A}\varphi(\cdot + y)) = \varphi(y).$$

Wenn $u = U * f$ ist, dann erhalten wir durch Vertauschen der Integrationsreihenfolge

$$\begin{aligned}(u, \mathcal{A}\varphi) &= \int_{\mathbf{R}^d}\int_{\mathbf{R}^d} U(x-y) f(y)\,\mathrm{d}y\,\mathcal{A}\varphi(x)\,\mathrm{d}x \\ &= \int_{\mathbf{R}^d}\int_{\mathbf{R}^d} U(x-y)\mathcal{A}\varphi(x)\,\mathrm{d}x\, f(y)\,\mathrm{d}y \\ &= \int_{\mathbf{R}^d} \varphi(y)\, f(y)\,\mathrm{d}y = (f, \varphi).\end{aligned} \tag{3.16}$$

Aus $f \in \mathcal{C}_0^1$ folgt $u \in \mathcal{C}^2$, weil mit $D_i = \partial/\partial x_i$ die Gleichung $D_i D_j u(x) = (D_i U * D_j f)(x)$ (vgl. Anhang A.3) sowie $D_i U \in L_1(\mathbf{R}^d)$ und $D_j f \in \mathcal{C}_0(\mathbf{R}^d)$ gilt. Somit können wir in (3.16) partiell integrieren und erhalten (vgl. (3.13))

$$(\mathcal{A}u - f, \varphi) = 0 \quad \forall \varphi \in \mathcal{C}_0^\infty(\mathbf{R}^d).$$

Daraus können wir $\mathcal{A}u = f$ schließen. □

Im folgenden Theorem bestimmen wir Fundamentallösungen der Poisson-Gleichung in zwei und drei Dimensionen.

Theorem 3.5. *Sei*

$$U(x) = \begin{cases} -\dfrac{1}{2\pi}\log|x| & \text{für } d = 2, \\[2ex] \dfrac{1}{4\pi|x|} & \text{für } d = 3. \end{cases}$$

Dann ist U eine Fundamentallösung der Poisson-Gleichung (3.3).

Beweis. Wir führen den Beweis für den Fall $d = 2$, der Beweis für $d = 3$ verläuft analog. Differentiation nach $x \neq 0$ führt auf

$$-\frac{\partial U}{\partial x_j} = \frac{1}{2\pi}\,\frac{x_j}{|x|^2}, \qquad -\frac{\partial^2 U}{\partial x_j^2} = \frac{1}{2\pi}\,\frac{|x|^2 - 2x_j^2}{|x|^4},$$

sodass für $x \neq 0$ insbesondere $-\Delta U = 0$ gilt. Auf gleiche Weise kann gezeigt werden, dass (3.14) gilt.

Sei $\varphi \in \mathcal{C}_0^\infty(\mathbf{R}^2)$. Aufgrund der Greenschen Formel mit $n = x/|x|$ gilt

$$\int_{|x|>\epsilon} U(-\Delta\varphi)\,\mathrm{d}x = \int_{|x|>\epsilon} (-\Delta U)\varphi\,\mathrm{d}x - \int_{|x|=\epsilon} \Big(\varphi\frac{\partial U}{\partial n} - \frac{\partial\varphi}{\partial n}U\Big)\,\mathrm{d}s.$$

Der erste Term auf der rechten Seite verschwindet. Wegen

$$\frac{\partial U}{\partial n} = \frac{x_1}{|x|}\frac{\partial U}{\partial x_1} + \frac{x_2}{|x|}\frac{\partial U}{\partial x_2} = \frac{1}{2\pi}\,\frac{1}{|x|} = \frac{1}{2\pi\epsilon} \quad \text{für } |x| = \epsilon$$

gilt außerdem

$$\int_{|x|=\epsilon} \varphi\,\frac{\partial U}{\partial n}\,\mathrm{d}s = \frac{1}{2\pi\epsilon}\int_{|x|=\epsilon} \varphi\,\mathrm{d}s \to \varphi(0) \quad \text{für } \epsilon \to 0.$$

Ebenso ist

$$\int_{|x|=\epsilon} \frac{\partial\varphi}{\partial n}\,U\,\mathrm{d}s = \frac{1}{2\pi}\log(\epsilon)\int_{|x|=\epsilon}\frac{\partial\varphi}{\partial n}\,\mathrm{d}s \le \epsilon\log(\epsilon)\|\nabla\varphi\|_{\mathcal{C}} \to 0 \quad \text{für } \epsilon \to 0.$$

Folglich gilt

$$(U, (-\Delta)\varphi) = \lim_{\epsilon\to 0}\int_{|x|>\epsilon} U(x)(-\Delta)\varphi(x)\,\mathrm{d}x = \varphi(0).$$

□

Wir können nun für das Randwertproblem

$$-\Delta u = f \quad \text{in } \Omega \quad \text{mit } u = 0 \quad \text{auf } \Gamma \tag{3.17}$$

eine Greensche Funktion $G(x, y)$ konstruieren, die für $x, y \in \Omega$ definiert ist, sodass die Lösung von (3.17) in der Form

$$u(x) = \int_\Omega G(x, y)f(y)\,\mathrm{d}y \tag{3.18}$$

dargestellt werden kann. Sei

$$G(x, y) = U(x - y) - v_y(x), \tag{3.19}$$

wobei U die Fundamentallösung aus Theorem 3.5 ist. Für ein festes $y \in \Omega$ sei v_y die Lösung von

$$-\Delta_x v_y(x) = 0 \quad \text{in } \Omega \quad \text{mit } v_y(x) = U(x - y) \quad \text{auf } \Gamma.$$

Im nächsten Abschnitt werden wir zeigen, dass dieses Problem eine Lösung besitzt. Die Greensche Funktion besitzt somit die Singularität der Fundamentallösung und verschwindet für $x \in \Gamma$. Wir können uns leicht davon überzeugen, dass deshalb die durch (3.18) definierte Funktion eine Lösung von (3.17) ist. Sie ist außerdem die einzige Lösung, da wir die Eindeutigkeit bereits in Abschnitt 3.2 bewiesen haben. Beachten Sie, dass $G(x, y)$ aus einem singulären Teil $U(x - y)$ mit einer Singularität an der Stelle $x = y$ und einem glatten Teil $v_y(x)$ besteht.

3.5 Variationsformulierung des Dirichlet-Problems

Wir betrachten zunächst das Dirichlet-Problem mit homogenen Randbedingungen

$$\mathcal{A}u := -\nabla \cdot \big(a\nabla u\big) + b \cdot \nabla u + cu = f \quad \text{in } \Omega \quad \text{mit } u = 0 \quad \text{auf } \Gamma, \tag{3.20}$$

wobei die Koeffizienten a, b und c glatte Funktionen auf $\bar{\Omega}$ sind, die

$$a(x) \geq a_0 > 0, \quad c(x) - \tfrac{1}{2}\nabla \cdot b(x) \geq 0 \quad \text{für } x \in \Omega \tag{3.21}$$

erfüllen, und f eine gegebene Funktion ist. Bei der klassischen Formulierung des Problems sucht man nach einer Funktion $u \in \mathcal{C}^2 = \mathcal{C}^2(\bar{\Omega})$, die (3.20) erfüllt. In diesem Abschnitt werden wir (3.20) in Variationsform umformulieren und nach einer Lösung innerhalb der größeren Klasse H_0^1 suchen. In einigen Fällen ist es anschließend möglich, die Regularität dieser Lösung zu beweisen, sodass es sich dann tatsächlich auch um eine klassische Lösung handelt.

Nehmen wir zunächst an, dass u eine Lösung in $\mathcal{C}^2$ ist. Wir multiplizieren (3.20) mit $v \in \mathcal{C}_0^1$ und integrieren über Ω. Mit der Greenschen Formel und weil $v = 0$ auf Γ gilt, erhalten wir

$$\int_\Omega fv \,\mathrm{d}x = \int_\Omega \mathcal{A}u\, v \,\mathrm{d}x = \int_\Omega (a\,\nabla u \cdot \nabla v + b \cdot \nabla u\, v + c\, u\, v) \,\mathrm{d}x \quad \forall v \in \mathcal{C}_0^1 \tag{3.22}$$

und wegen der Dichtheit von $\mathcal{C}_0^1$ in H_0^1 auch

$$\int_\Omega \big(a\,\nabla u \cdot \nabla v + b \cdot \nabla u\, v + c\, u\, v\big) \,\mathrm{d}x = \int_\Omega fv \,\mathrm{d}x \quad \forall v \in H_0^1. \tag{3.23}$$

Das zu (3.20) gehörige Variationsproblem besteht folglich darin, ein $u \in H_0^1$ zu finden, für das (3.23) gilt. Weiter unten wird mithilfe des Lax-Milgram-Lemmas gezeigt, dass dieses Problem für $f \in L_2$ eine eindeutige Lösung zulässt. Wir sagen, dass diese Lösung eine *schwache Lösung* oder *Variationslösung* von (3.20) ist.

Wir haben uns also davon überzeugt, dass eine klassische Lösung auch gleichzeitig eine schwache Lösung ist. Nehmen wir umgekehrt an, dass $u \in H_0^1$ eine schwache Lösung ist, d. h. dass u Gleichung (3.23) erfüllt. Wenn wir *zusätzlich* wissen, dass $u \in \mathcal{C}^2$ gilt, dann folgt aus (3.23) mit der Greenschen Formel

$$\int_\Omega fv \,\mathrm{d}x = \int_\Omega \big(a\,\nabla u \cdot \nabla v + b \cdot \nabla u\, v + c\, u\, v\big) \,\mathrm{d}x = \int_\Omega \mathcal{A}\, u\, v \,\mathrm{d}x \quad \forall v \in H_0^1,$$

d. h.

$$\int_\Omega (\mathcal{A}u - f)v \,\mathrm{d}x = 0 \quad \forall v \in H_0^1.$$

Für $f \in \mathcal{C}$ gilt dann $\mathcal{A}u - f \in \mathcal{C}$. Deshalb folgt aus dieser Beziehung

$$\mathcal{A}u(x) - f(x) = 0 \quad \forall x \in \Omega.$$

Wegen $u \in H_0^1$ gilt auch $u = 0$ auf Γ. Daraus folgt, dass u eine klassische Lösung von (3.20) ist. Eine hinreichend glatte schwache Lösung ist folglich auch eine klassische Lösung. Es hängt jedoch von den Daten f und dem Gebiet Ω ab, ob eine Lösung hinreichend glatt ist, um sie als eine klassische Lösung bezeichnen zu können. Deshalb stellt die schwache Formulierung (3.23) tatsächlich eine Erweiterung der klassischen Formulierung dar. Beachten Sie, dass die schwache Formulierung (3.23) für beliebiges $f \in L_2$ sinnvoll ist, sodass f beispielsweise unstetig sein kann, während die klassische Formulierung (3.20) die Stetigkeit von f fordert. Wenn $f \in L_2$ ist und $u \in H^2 \cap H_0^1$ Gleichung (3.20) erfüllt, dann sagen wir, dass u eine *starke Lösung* ist. Es ist offensichtlich, dass eine klassische Lösung auch eine starke Lösung und eine starke Lösung auch eine schwache Lösung ist. Darüber hinaus ist eine schwache Lösung, die in H^2 liegt, eine starke Lösung. Weiter hinten werden wir auf das Problem der Regularität schwacher Lösungen zurückkommen.

Wir sind nun in der Lage, die Existenz einer schwachen Lösung zu zeigen. Wir verwenden unsere Standardnotation aus Abschnitt 1.2.

Theorem 3.6. *Angenommen, es gelte* (3.21) *und* $f \in L_2$. *Dann lässt das Randwertproblem* (3.20) *eine eindeutige schwache Lösung zu, d. h. es existiert ein eindeutiges* $u \in H_0^1$, *das* (3.23) *erfüllt. Darüber hinaus existiert eine von* f *unabhängige Konstante* C, *sodass*

$$|u|_1 \le C\|f\| \tag{3.24}$$

gilt.

Beweis. Wir wenden das Lax-Milgram-Lemma, Theorem A.3, in dem mit der Norm $|\cdot|_1$ versehenen Hilbert-Raum $V = H_0^1$ und mit

$$a(v,w) = \int_\Omega \big(a\nabla v \cdot \nabla w + b \cdot \nabla v\, w + c\, v\, w\big)\,\mathrm{d}x \quad \text{und} \quad L(v) = \int_\Omega f v\,\mathrm{d}x \tag{3.25}$$

an. Offensichtlich ist die Bilinearform $a(\cdot,\cdot)$ in H_0^1 beschränkt. Sie ist koerzitiv, wenn (3.21) gilt, weil

$$a(v,v) = \int_\Omega \big(a|\nabla v|^2 + (c - \tfrac{1}{2}\nabla \cdot b)|v|^2\big)\,\mathrm{d}x \ge a_0|v|_1^2 \quad \forall v \in H_0^1$$

ist. Ferner handelt es sich bei $L(\cdot)$ um ein beschränktes lineares Funktional auf H_0^1, da wegen der Poincaré-Ungleichung, Theorem A.6, die Gleichung

$$|L(v)| \le \|f\|\,\|v\| \le \|f\|\,\|v\|_1 \le C\|f\|\,|v|_1$$

gilt. Daraus ergibt sich $\|L\|_{V^*} \le C\|f\|$ und die Behauptung des Theorems folgt somit direkt aus Theorem A.3. □

Wir stellen fest, dass sich (3.21) im Falle $b = 0$ auf (3.2) reduziert und die Bilinearform $a(\cdot,\cdot)$ ein Skalarprodukt auf H_0^1 ist. Das Theorem kann dann unter Verwendung des Rieszschen Darstellungssatzes bewiesen werden. In diesem Fall zeigt Theorem A.2, dass die schwache Lösung von (3.20) auch folgendermaßen charakterisiert werden kann:

Theorem 3.7. (Dirichlet-Prinzip.) *Gleichung* (3.2) *sei erfüllt und es gelte* $b = 0$. *Sei* $f \in L_2$ *und* $u \in H_0^1$ *die Lösung von* (3.23). *Wir setzen*

$$F(v) = \tfrac{1}{2} \int_\Omega \left(a|\nabla v|^2 + c\, v^2 \right) \mathrm{d}x - \int_\Omega f v \, \mathrm{d}x. \tag{3.26}$$

Dann ist $F(u) \le F(v)$ *für alle* $v \in H_0^1$, *wobei Gleichheit nur im Falle* $v = u$ *gilt.*

Anmerkung 3.1. Wenn (3.20) beispielsweise als Modell einer an ihrem Rand fixierten elastischen Membran betrachtet wird, dann handelt es sich bei dem durch Gleichung (3.26) definierten $F(v)$ um die zur Auslenkung v gehörende *potentielle Energie*. Der erste Term in $F(v)$ entspricht der *inneren elastischen Energie*, während der zweite Term einem Lastpotential entspricht (analoge Interpretationen gibt es für mechanische und physikalische Probleme, die durch (3.20) modelliert werden). Das Dirichlet-Problem entspricht in diesem Fall dem *Prinzip der minimalen potentiellen Energie* aus der Mechanik und (3.23) dem *Prinzip der virtuellen Arbeit*.

Wir betrachten nun das Randwertproblem mit inhomogener Randbedingung

$$\mathcal{A}u = f \quad \text{in } \Omega \quad \text{mit } u = g \quad \text{auf } \Gamma, \tag{3.27}$$

wobei wir annehmen, dass $f \in L_2$ und $g \in L_2(\Gamma)$ ist. In der schwachen Formulierung bedeutet dieses Problem, ein $u \in H^1$ mit $a(\cdot,\cdot)$ und $L(\cdot)$ wie in (3.25) zu finden, sodass

$$a(u,v) = L(v) \quad \forall v \in H_0^1 \quad \text{mit } \gamma\, u = g \tag{3.28}$$

gilt. Dabei ist $\gamma : H^1 \to L_2(\Gamma)$ der Spuroperator (vgl. Theorem A.4). Zum Beweis der Existenz einer Lösung nehmen wir an, dass es sich bei der auf Γ gegebenen Funktion g um die Spur einer Funktion $u_0 \in H^1$ handelt, d. h. $g = \gamma u_0$. Wir setzen $w = u - u_0$ und suchen dann nach einem $w \in H_0^1$, das

$$a(w,v) = L(v) - a(u_0,v) \quad \forall v \in H_0^1 \tag{3.29}$$

erfüllt. Die rechte Seite ist ein beschränktes lineares Funktional auf H_0^1. Somit folgt wegen des Lax-Milgram-Lemmas, dass ein eindeutiges $w \in H_0^1$ existiert, das (3.29) erfüllt. Offensichtlich erfüllt $u = u_0 + w$ sowohl (3.28) als auch $\gamma u = g$. Diese Lösung ist eindeutig, denn wenn (3.27) zwei schwache Lösungen u_1, u_2 mit denselben Daten f, g hätte, dann wäre deren Differenz $u_1 - u_2 \in H_0^1$

eine schwache Lösung von (3.20) mit $f = 0$. Daher würde aus der Stabilitätsabschätzung (3.24) die Gleichung $u_1 - u_2 = 0$, d. h. $u_1 = u_2$ folgen. Folglich besitzt (3.27) eine eindeutige schwache Lösung. Insbesondere ist die Lösung u von der Wahl der Fortsetzung u_0 der Randwerte g unabhängig.

Im Fall $b = 0$ kann die schwache Lösung $u \in H^1$ äquivalent dazu auch als die eindeutige Lösung des Minimierungsproblems

$$\inf_{\substack{v \in H^1 \\ \gamma v = g}} \Big(\tfrac{1}{2} \int_\Omega \big(a|\nabla v|^2 + c\, v^2 \big)\, \mathrm{d}x - \int_\Omega f v \, \mathrm{d}x \Big)$$

charakterisiert werden.

3.6 Ein Neumann-Problem

Wir betrachten nun das Neumann-Problem

$$\mathcal{A}u := -\nabla \cdot \big(a \nabla u \big) + cu = f \quad \text{in } \Omega \quad \text{mit} \quad \frac{\partial u}{\partial n} = 0 \quad \text{auf } \Gamma, \tag{3.30}$$

wobei wir nun zusätzlich zu (3.2) $c(x) \geq c_0 > 0$ in Ω fordern und $f \in L_2$ ist. (Der Fall $c = 0$ wird in Problemstellung 3.9 diskutiert.) Um zu einer Variationsformulierung von (3.30) zu gelangen, multiplizieren wir die Differentialgleichung in (3.30) mit $v \in \mathcal{C}^1$ (beachten Sie, dass v keine Randbedingungen erfüllen muss) und integrieren über Ω unter Verwendung der Greenschen Formel:

$$\int_\Omega f v \, \mathrm{d}x = \int_\Omega \mathcal{A}u \, v \, \mathrm{d}x = -\int_\Gamma a \frac{\partial u}{\partial n} v \, \mathrm{d}s + \int_\Omega \big(a \nabla u \cdot \nabla v + c\, uv \big)\, \mathrm{d}x,$$

sodass wegen $\partial u / \partial n = 0$ auf Γ

$$\int_\Omega \big(a \nabla u \cdot \nabla v + c\, uv \big)\, \mathrm{d}x = \int_\Omega f v \, \mathrm{d}x \quad \forall v \in \mathcal{C}^1 \tag{3.31}$$

gilt. Wenn umgekehrt $u \in \mathcal{C}^2$ Gleichung (3.31) erfüllt, dann gilt aufgrund der Greenschen Formel

$$\int_\Omega \big(\mathcal{A}u - f \big) v \, \mathrm{d}x + \int_\Gamma a \frac{\partial u}{\partial n} v \, \mathrm{d}s = 0 \quad \forall v \in \mathcal{C}^1. \tag{3.32}$$

Wenn wir v zunächst über $\mathcal{C}_0^1$ variieren lassen, dann sehen wir, dass u die Differentialgleichung in (3.30) erfüllen muss. Somit verschwindet der erste Term auf der linken Seite von (3.32). Variieren wir v auf Γ, dann können wir uns davon überzeugen, dass u auch die Randbedingungen in (3.30) erfüllt.

Dies hat uns zu der folgenden Variationsformulierung von (3.30) geführt: Gesucht ist eine Funktion $u \in H^1$, sodass

$$a(u,v) = L(v) \quad \forall v \in H^1 \tag{3.33}$$

erfüllt ist, wobei $a(\cdot,\cdot)$ und $L(\cdot)$ wie in (3.25) mit $b = 0$ vorausgesetzt werden können.

Wir haben gesehen, dass eine klassische Lösung u von (3.30) die Gleichung (3.33) erfüllt. Umgekehrt gilt: Wenn u die Gleichung (3.33) erfüllt und zusätzlich $u \in \mathcal{C}^2$ ist, dann ist u eine klassische Lösung von (3.30).

Aufgrund des Rieszschen Darstellungssatzes erhalten wir hier das folgende Existenz-, Eindeutigkeits- und Stabilitätsresultat. Beachten Sie, dass die Bilinearform $a(\cdot,\cdot)$ wegen $c(x) \geq c_0 > 0$ ein Skalarprodukt auf H^1 ist.

Theorem 3.8. *Wenn $f \in L_2$ ist, dann lässt das Neumann-Problem* (3.30) *eine eindeutige schwache Lösung zu, d. h. es gibt eine eindeutige Funktion $u \in H^1$, die* (3.33) *erfüllt. Darüber hinaus gilt*

$$\|u\|_1 \leq C\|f\|.$$

Anmerkung 3.2. Beachten Sie, dass hier die Neumannsche Randbedingung $\partial u/\partial n = 0$ auf Γ in der Variationsformulierung (3.33) nicht explizit gefordert wird; die Funktion u soll lediglich zu H^1 gehören. Die Randbedingung ist in (3.33) implizit enthalten, weil die Testfunktion v eine beliebige Funktion in H^1 sein kann. Eine solche Randbedingung, die nicht explizit gefordert werden muss, wird als *natürliche Randbedingung* bezeichnet. Im Gegensatz dazu heißt eine Randbedingung, wie die Dirichletsche Bedingung $u = g$ auf Γ, die explizit als Teil der Variationsformulierung gestellt wird, *wesentliche Randbedingung*.

Anmerkung 3.3. Für das Problem

$$\mathcal{A}u = f \quad \text{in } \Omega \quad \text{mit} \quad a\frac{\partial u}{\partial n} = g \quad \text{auf } \Gamma \tag{3.34}$$

mit $f \in L_2(\Omega)$ und $g \in L_2(\Gamma)$ kann die folgende Variationsformulierung angegeben werden: Gesucht ist ein $u \in H^1$, sodass

$$a(u,v) = L(v) \quad \forall v \in H^1 \tag{3.35}$$

gilt, wobei $a(\cdot,\cdot)$ wie in (3.25) mit $b = 0$ und

$$L(v) = \int_\Omega fv \,\mathrm{d}x + \int_\Gamma gv \,\mathrm{d}s$$

ist. Mit der Cauchy-Schwarz-Ungleichung und der Spurungleichung (Theorem A.4) gilt

$$|L(v)| \leq \|f\|\,\|v\| + \|g\|_{L_2(\Gamma)}\|v\|_{L_2(\Gamma)} \leq \big(\|f\| + C\|g\|_{L_2(\Gamma)}\big)\|v\|_1.$$

Folglich ist $L(\cdot)$ eine beschränkte Linearform auf H^1. Der Rieszsche Darstellungssatz liefert deshalb die Existenz und die Eindeutigkeit einer Funktion $u \in H^1$, die (3.35) erfüllt (siehe dazu auch Problemstellung 3.7).

3.7 Regularität

Aus Theorem 3.6 wissen wir, dass das Dirichlet-Problem (3.20) für jedes $f \in L_2$ eine eindeutige schwache Lösung $u \in H_0^1$ besitzt. Man kann zeigen, dass für ein glattes Γ oder ein konvexes Polynom Γ tatsächlich $u \in H^2$ gilt und eine von f unabhängige Konstante C existiert, sodass

$$\|u\|_2 \le C\|f\| \tag{3.36}$$

gilt. Beachten Sie, dass diese Abschätzung, wenn sie beispielsweise im Falle $\mathcal{A} = -\Delta$ angewendet wird, bedeutet, dass es möglich ist, die L_2-Norm *aller* zweiten Ableitungen einer auf Γ verschwindenden Funktion u durch die L_2-Norm der speziellen Kombination zweiter Ableitungen von u abzuschätzen, die durch den Laplace-Operator $-\Delta$ gegeben ist. Ein Beispiel gibt Problemstellung 3.10, worin Ω weder glatt noch konvex ist und die Regularitätsabschätzung (3.36) nicht erfüllt ist.

Die Ungleichung (3.36) zeigt, dass die Funktion u und deren erste und zweite Ableitung stetig von f in dem Sinne abhängen, dass falls u_1 und u_2

$$-\mathcal{A}u_i = f_i \quad \text{in } \Omega \quad \text{mit } u_i = 0 \quad \text{auf } \Gamma \quad \text{für } i = 1, 2$$

erfüllen, die Ungleichung

$$\Big(\sum_{|\alpha| \le 2} \|D^\alpha u_1 - D^\alpha u_2\|^2 \Big)^{1/2} \le C\|f_1 - f_2\|$$

gilt.

Wenn Γ glatt ist, kann (3.36) folgendermaßen verallgemeinert werden. Für jede ganze Zahl $k \ge 0$ existiert eine von f unabhängige Konstante C, sodass, falls u die schwache Lösung von (3.20) mit $f \in H^k$ ist, $u \in H^{k+2} \cap H_0^1$ und

$$\|u\|_{k+2} \le C\|f\|_k \tag{3.37}$$

ist. Insbesondere folgt daraus unter Beachtung der Sobolevschen Ungleichung, Theorem A.5, dass im Fall $k > d/2$ die Beziehung $u \in \mathcal{C}^2$ erfüllt ist und u folglich auch eine klassische Lösung von (3.20) ist.

Ist Γ ein Polygon, liegt eine weniger günstige Situation vor. Wenn $\mathcal{A} = -\Delta$ ist und $\Omega \subset \mathbf{R}^2$ eine Ecke mit dem Innenwinkel ω besitzt, denn können wir nämlich mithilfe von Polarkoordinaten (r, φ), deren Zentrum in dieser Ecke liegt, zeigen, dass sich die Lösung von (3.20) in der Nähe der Ecke wie $u(r, \varphi) = c\, r^\beta \sin(\beta\varphi)$ mit $\beta = \pi/\omega$ verhält. Dabei entspricht $\varphi = 0$ einer der Kanten. Um für eine solche Funktion in der Nähe der Ecke H^k-Regularität zu erhalten, muss $(\partial/\partial r)^k u(r, \varphi) \in L_2(\Omega_0)$ gelten, wobei $\Omega_0 \subset \Omega$ die Umgebung der betrachteten Ecke, aber keine anderen Ecken enthält. Dies erfordert aber für hinreichend kleine b entweder

$$\big(\beta(\beta-1)\cdots(\beta-k+1)\big)^2 \int_0^b r^{2(\beta-k)} r \,\mathrm{d}r < \infty$$

oder $2(\beta-k)+1 \geq -1$. (Es ist $\beta-k+1=0$, falls $2(\beta-k)+1=-1$ gilt.) Dies bedeutet wiederum $\omega \leq \pi/(k-1)$. Für $k=2$ müssen also alle Winkel kleiner gleich π sein, d. h. Ω muss konvex sein. Im Falle $k=3$ müssen alle Winkel $\leq \pi/2$ sein, was eine schwerwiegende Einschränkung ist. Wir verweisen auf Problemstellung 3.10, die ein Beispiel zur Illustration dieser Tatsache liefert.

3.8 Problemstellungen

Problem 3.1. Geben Sie für das Dirichlet-Problem

$$-\sum_{j,k=1}^{d} \frac{\partial}{\partial x_j}\Big(a_{jk}\frac{\partial u}{\partial x_k}\Big) + a_0 u = f \quad \text{in } \Omega \quad \text{mit } u=0 \quad \text{auf } \Gamma$$

eine Variationsformulierung an und beweisen Sie die Existenz und Eindeutigkeit einer schwachen Lösung. Dabei sind $a_{jk}(x)$ und $a_0(x)$ Funktionen in $\mathcal{C}(\bar{\Omega})$, für die $a_0(x) \geq 0$ gilt und für die die Matrix $(a_{jk}(x))$ symmetrisch und gleichmäßig positiv definit in Ω ist, sodass $a_{jk}(x) = a_{kj}(x)$ und

$$\sum_{j,k=1}^{d} a_{jk}(x)\xi_j\xi_k \geq \kappa \sum_{j=1}^{d} \xi_j^2 \quad \text{mit } \kappa > 0 \text{ für } \xi \in \mathbf{R}^d,\ x \in \Omega$$

gilt.

Problem 3.2. Sei $f \in L_2$. Zeigen Sie, dass $p = \nabla u$ die Lösung des Minimierungsproblems

$$\inf_{q \in H_f} \tfrac{1}{2} \int_\Omega |q|^2 \,\mathrm{d}x$$

mit

$$H_f = \big\{q = (q_1, \ldots, q_d) : q_i \in L_2,\ -\nabla \cdot q = f \text{ in } \Omega\big\}$$

ist, wenn u die Gleichung $-\Delta u = f$ in $\Omega, u = 0$ auf Γ erfüllt.

Problem 3.3. Betrachten Sie zwei beschränkte Gebiete Ω_1 und Ω_2 mit einem gemeinsamen Rand S. Sei $\Gamma_i = \partial\Omega_i \setminus S$, wobei $\partial\Omega_i$ der Rand von $\Omega_i,\ i=1,2$, ist (siehe Abbildung 3.1).

Geben Sie eine Variationsformulierung des folgendes Problems an: Gesucht ist ein in $\Omega_i,\ i=1,2$ definiertes u_i, sodass die Gleichungen

$$\begin{aligned} -a_1\Delta u_1 &= f_1 \quad \text{in } \Omega_1, & -a_2\Delta u_2 &= f_2 \quad \text{in } \Omega_2,\\ u_1 &= 0 \quad \text{auf } \Gamma_1, & u_2 &= 0 \quad \text{auf } \Gamma_2 \end{aligned}$$

und

$$u_1 = u_2, \quad a_1\frac{\partial u_1}{\partial n} = a_2\frac{\partial u_2}{\partial n} \quad \text{auf } S$$

erfüllt sind. Dabei ist $f_i \in L_2(\Omega_i),\ a_i > 0$ für $i=1,2$, konstant und n eine Einheitsnormale an S. Beweisen Sie die Existenz und die Eindeutigkeit der Lösung. Geben Sie eine physikalische Interpretation.

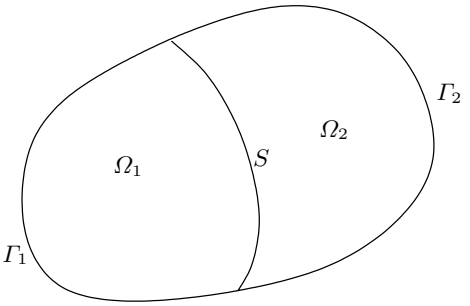

Abbildung 3.1. Gebiet mit Grenzfläche.

Problem 3.4. Beweisen Sie die *Ungleichung von Friedrichs*

$$\|v\|_{L_2(\Omega)} \le C\Big(\|\nabla v\|^2_{L_2(\Omega)} + \|v\|^2_{L_2(\Gamma)}\Big)^{\frac{1}{2}} \quad \text{für } v \in \mathcal{C}^1,$$

wobei Ω ein beschränktes Gebiet in $\mathbf{R}^d$ mit dem Rand Γ ist. Hinweis: Integrieren Sie in der Identität $\int_\Omega v^2 \,\mathrm{d}x = \int_\Omega v^2 \Delta\phi \,\mathrm{d}x$ partiell. Dabei ist $\phi(x) = \frac{1}{2d}|x|^2$.

Problem 3.5. Beweisen Sie

$$\|v\| \le C\Big(\|\nabla v\|^2 + \Big(\int_\Omega v \,\mathrm{d}x\Big)^2\Big)^{\frac{1}{2}} \quad \text{für } v \in \mathcal{C}^1,$$

wobei Ω das Einheitsquadrat in $\mathbf{R}^2$ ist. Die Ungleichung gilt auch, falls Ω ein beschränktes Gebiet in $\mathbf{R}^d$ ist. Hinweis: $v(x) = v(y) + \int_{y_1}^{x_1} D_1 v(s, x_2)\,\mathrm{d}s + \int_{y_2}^{x_2} D_2 v(y_1, s)\,\mathrm{d}s$.

Problem 3.6. Geben Sie eine Variationsformulierung des Problems

$$-\Delta u = f \quad \text{in } \Omega \quad \text{mit } \frac{\partial u}{\partial n} + u = g \quad \text{auf } \Gamma$$

mit $f \in L_2(\Omega)$ und $g \in L_2(\Gamma)$ an. Beweisen Sie die Existenz und die Eindeutigkeit einer schwachen Lösung. Interpretieren Sie die Randbedingung in Bezug auf ein Problem aus der Mechanik oder der Physik. Hinweis: Siehe Problemstellung 3.4.

Problem 3.7. Beweisen Sie die Stabilitätsabschätzung

$$\|u\|_{H^1(\Omega)} \le C\Big(\|f\|_{L_2(\Omega)} + \|g\|_{L_2(\Gamma)}\Big)$$

für die Lösung von (3.34).

Problem 3.8. Geben Sie eine Variationsformulierung des Problems

$$-\nabla \cdot \big(a\nabla u\big) + cu = f \quad \text{in } \Omega \quad \text{mit } a\frac{\partial u}{\partial n} + h(u-g) = k \quad \text{auf } \Gamma$$

an. Dabei ist $f \in L_2(\Omega)$ und $g, k \in L_2(\Gamma)$. Die Koeffizienten a, c, h sind glatt mit

$$a(x) \geq a_0 > 0,\ c(x) \geq 0 \quad \text{für } x \in \Omega, \quad h(x) \geq h_0 > 0 \quad \text{für } x \in \Gamma.$$

Beweisen Sie die Existenz und die Eindeutigkeit einer schwachen Lösung. Beweisen Sie die Stabilitätsabschätzung

$$\|u\|_{H^1(\Omega)} \leq C\big(\|f\|_{L_2(\Omega)} + \|k\|_{L_2(\Gamma)} + \|g\|_{L_2(\Gamma)}\big).$$

Hinweis: Verwenden Sie Problemstellung 3.4.

Problem 3.9. Betrachten Sie das Neumann-Problem

$$-\Delta u = f \quad \text{in } \Omega \quad \text{mit } \frac{\partial u}{\partial n} = 0 \quad \text{auf } \Gamma. \tag{3.38}$$

Nehmen Sie $f \in L_2(\Omega)$ an. Zeigen Sie, dass die Bedingung

$$\int_\Omega f \,\mathrm{d}x = 0$$

für die Existenz der Lösung notwendig ist.

Beachten Sie, dass wenn u Gleichung (3.38) erfüllt, dies auch auf $u + c$ mit einer beliebigen Konstanten c zutrifft. Um die Eindeutigkeit zu erhalten, nehmen Sie die zusätzliche Bedingung

$$\int_\Omega u \,\mathrm{d}x = 0$$

hinzu, die verlangt, dass der Mittelwert von u gleich null ist. Stellen Sie dieses Problem unter Verwendung des Raumes

$$V = \Big\{v \in H^1(\Omega) : \int_\Omega v \,\mathrm{d}x = 0\Big\}$$

in einer Variationsformulierung. Hinweis: Siehe Problemstellung 3.5.

Problem 3.10. Sei Ω ein Sektor mit dem Winkel $\omega = \pi/\beta$:

$$\Omega = \big\{(r,\varphi) : 0 < r < 1,\ 0 < \varphi < \pi/\beta\big\},$$

wobei r, φ ebene Polarkoordinaten sind. Sei $v(r,\varphi) = r^\beta \sin(\beta\varphi)$. Bestätigen Sie, dass v harmonisch ist, d. h. dass $\Delta v = 0$ gilt, indem Sie

$$\Delta v = \frac{1}{r}\frac{\partial}{\partial r}\Big(r\frac{\partial v}{\partial r}\Big) + \frac{1}{r^2}\frac{\partial^2 v}{\partial \varphi^2}$$

berechnen. (Dies folgt auch unmittelbar aus der Feststellung, dass v der Imaginärteil der komplexen analytischen Funktion z^β ist.) Setzen Sie $u(r,\varphi) = (1-r^2)v(r,\varphi)$. Dann ist $u = 0$ auf Γ. Zeigen Sie, dass u die Gleichung $-\Delta u = f$ mit $f = 4(1+\beta)v$ erfüllt. Folglich ist $f \in H^1(\Omega)$. Berechnen Sie anschließend $\|\partial^2 u/\partial r^2\|_{L_2(\Omega)}$ und schlussfolgern Sie, dass $u \notin H^2(\Omega)$ gilt, wenn $\beta < 1$ ist, d. h. wenn Ω nichtkonvex oder $\omega > \pi$ ist. Zeigen Sie auf ähnliche Weise, dass $u \notin H^3(\Omega)$ ist, wenn $\omega > \pi/2$ gilt. Hinweis: Der am stärksten singuläre Term in u_{rr} ist $\beta(\beta-1)r^{\beta-2}\sin(\beta\varphi)$.

Problem 3.11. (Elliptische Regularität für ein Rechteck.) Sei $\Omega \subset \mathbf{R}^2$ ein Rechteck und u eine glatte Funktion mit $u = 0$ auf Γ. Beweisen Sie

$$|u|_2 = \|\Delta u\|.$$

Verwenden Sie dies, um (3.36) für $\mathcal{A} = -\Delta$ zu beweisen.

Hinweis: Erinnern Sie sich, dass

$$|u|_2^2 = \int_\Omega \Big(\big(\frac{\partial^2 u}{\partial x_1^2}\big)^2 + 2\big(\frac{\partial^2 u}{\partial x_1 \partial x_2}\big)^2 + \big(\frac{\partial^2 u}{\partial x_2^2}\big)^2 \Big)\, \mathrm{d}x$$

gilt und integrieren Sie in $\int_\Omega \big(\frac{\partial^2 u}{\partial x_1 \partial x_2}\big)^2 \,\mathrm{d}x$ partiell. Gehen Sie von der Definition $\|u\|_2 = \big(\|u\|^2 + |u|_1^2 + |u|_2^2\big)^{1/2}$ aus und beweisen Sie, dass $\|u\| \le C|u|_1$ und $|u|_1 \le \big(\|u\|\,|u|_2\big)^{1/2}$ ist.

Auf derselben Idee aufbauend kann man für beliebige konvexe Gebiete mithilfe eines etwas komplizierteren Argumentes zeigen, dass $|u|_2 \le \|\Delta u\|$ ist.

Problem 3.12. Ersetzen Sie die Randbedingung in Problemstellung 3.11 durch die Neumannsche Bedingung $\partial u/\partial n = 0$ auf Γ. Beweisen Sie $|u|_2 = \|\Delta u\|$.

Problem 3.13. (Stabilität bezüglich der Koeffizienten.) Sei u_i mit $i = 1, 2$ die schwache Lösung des Problems

$$-\nabla \cdot \big(a_i \nabla u_i\big) = f \quad \text{in } \Omega \quad \text{mit } u_i = 0 \quad \text{auf } \Gamma.$$

Dabei ist $\Omega \subset \mathbf{R}^d$ ein Gebiet mit hinreichend glattem Rand Γ und $f \in L_2(\Omega)$. Die Koeffizienten $a_i(x)$ sind glatt mit

$$a_i(x) \ge a_0 > 0 \quad \text{für } x \in \Omega.$$

Beweisen Sie die Stabilitätsabschätzung

$$|u_1 - u_2|_1 \le \frac{C}{a_0^2}\, \|a_1 - a_2\|_C \|f\|.$$

4

Finite Differenzenverfahren für elliptische Gleichungen

Die frühen Entwicklungen auf dem Gebiet der numerischen Analyse partieller Differentialgleichungen wurden durch die finiten Differenzenverfahren dominiert. Bei einer solchen Methode wird an den Punkten eines endlichen Punktgitters nach einer approximativen Lösung gesucht. Die Approximation der Differentialgleichung wird durch Ersetzen der Ableitungen durch geeignete Differenzenquotienten erreicht. Dies reduziert das Differentialgleichungsproblem auf ein endliches lineares System von algebraischen Gleichungen. In diesem Kapitel illustrieren wir dies anhand eines Zweipunkt-Randwertproblems in einer Dimension und anhand des Dirichlet-Problems für die Poisson-Gleichung in der Ebene. Die Analyse basiert auf diskreten Versionen der Maximumprinzipien aus den beiden vorangegangenen Kapiteln.

4.1 Ein Zweipunkt-Randwertproblem

Wir betrachten das Zweipunkt-Randwertproblem

$$\begin{aligned} \mathcal{A}u := -au'' + bu' + cu &= f \quad \text{in } \Omega = (0,1), \\ u(0) = u_0,\ u(1) &= u_1. \end{aligned} \tag{4.1}$$

Die Koeffizienten $a = a(x)$, $b = b(x)$ und $c = c(x)$ sind glatte Funktionen mit $a(x) > 0$ und $c(x) \geq 0$ in $\bar{\Omega}$. Die Funktion $f = f(x)$ und die Zahlen u_0 und u_1 sind gegeben.

Zur numerischen Lösung von (4.1) führen wir $M+1$ Gitterpunkte $0 = x_0 < x_1 < \cdots < x_M = 1$ ein, indem wir $x_j = jh,\ j = 0, \ldots, M$ mit $h = 1/M$ setzen. Wir bezeichnen die Approximation von $u(x_j)$ mit U_j und definieren die folgenden finiten Differenzenapproximationen der Ableitungen

$$\partial U_j = \frac{U_{j+1} - U_j}{h}, \qquad \bar{\partial} U_j = \frac{U_j - U_{j-1}}{h},$$
$$\partial\bar{\partial} U_j = \frac{U_{j+1} - 2U_j + U_{j-1}}{h^2}, \qquad \hat{\partial} U_j = \frac{U_{j+1} - U_{j-1}}{2h}.$$

Mit der Notation aus Abschnitt 1.2 gilt mit $\mathcal{C}^j = \mathcal{C}^j(\bar{\Omega})$ (siehe Problemstellung 4.1)

$$\begin{aligned} |\partial\bar{\partial}u(x_j) - u''(x_j)| &\le Ch^2|u|_{\mathcal{C}^4}, \\ |\hat{\partial}u(x_j) - u'(x_j)| &\le Ch^2|u|_{\mathcal{C}^3} \quad \text{für } j = 1, \dots, M-1. \end{aligned} \tag{4.2}$$

Setzen wir außerdem $a_j = a(x_j)$, $b_j = b(x_j)$, $c_j = c(x_j)$, $f_j = f(x_j)$, können wir nun die finite Differenzenapproximation von (4.1) definieren:

$$\begin{aligned} \mathcal{A}_h U_j := &- a_j\partial\bar{\partial}U_j + b_j\hat{\partial}U_j + c_jU_j = f_j \quad \text{für } j = 1, \dots, M-1, \\ &U_0 = u_0,\ U_M = u_1. \end{aligned} \tag{4.3}$$

Die Gleichung kann an einem inneren Punkt x_j als

$$(2a_j + h^2c_j)U_j - (a_j - \tfrac{1}{2}hb_j)U_{j+1} - (a_j + \tfrac{1}{2}hb_j)U_{j-1} = h^2f_j \tag{4.4}$$

geschrieben werden. Unser diskretes Problem (4.3) kann folglich in Matrixform

$$AU = g \tag{4.5}$$

mit $U = (U_1, \dots, U_{M-1})^{\mathrm{T}}$ gestellt werden. Die ersten und letzten Komponenten des Vektors $g = (g_1, \dots, g_{M-1})^{\mathrm{T}}$ enthalten Beiträge von den Randwerten u_0, u_1 sowie von f_1 beziehungsweise f_{M-1}. Die $(M-1) \times (M-1)$-Matrix ist tridiagonal und für hinreichend kleine h diagonaldominant, d. h. die Summe der Beträge der Nichtdiagonalelemente in einer Reihe wird durch das Diagonalelement in dieser Reihe beschränkt (siehe Problemstellung 4.2).

Für unsere erste Analyse zeigen wir zunächst ein diskretes Maximumprinzip, das dem stetigen Fall aus Theorem 2.1 ähnelt.

Lemma 4.1. *Angenommen, h ist so klein, dass $a_j \pm \frac{1}{2}hb_j \ge 0$ ist und U die Bedingung $\mathcal{A}_hU_j \le 0$ ($\mathcal{A}_hU_j \ge 0$) erfüllt.*
(i) *Im Falle $c = 0$ gilt*

$$\max_j U_j = \max\{U_0, U_M\} \quad \Big(\min_j U_j = \min\{U_0, U_M\}\Big).$$

(ii) *Im Falle $c \ge 0$ gilt*

$$\max_j U_j \le \max\{U_0, U_M, 0\} \quad \Big(\min_j U_j \ge \min\{U_0, U_M, 0\}\Big).$$

Beweis. (i) In Anbetracht von (4.4) gilt wegen $c = 0$ und $\mathcal{A}_hU_j \le 0$

$$\begin{aligned} U_j &= \frac{a_j - \frac{1}{2}hb_j}{2a_j}U_{j+1} + \frac{a_j + \frac{1}{2}hb_j}{2a_j}U_{j-1} + \frac{h^2}{2a_j}\mathcal{A}_hU_j \\ &\le \frac{a_j - \frac{1}{2}hb_j}{2a_j}U_{j+1} + \frac{a_j + \frac{1}{2}hb_j}{2a_j}U_{j-1}. \end{aligned} \tag{4.6}$$

Nehmen wir nun an, dass U ein inneres Maximum U_j besitzt. Wenn entweder $U_{j+1} < U_j$ oder $U_{j-1} < U_j$ wäre, dann würde dies (4.6) widersprechen, da die Koeffizienten auf der rechten Seite nichtnegativ sind und sich zu eins addieren. Somit gilt $U_j = U_{j-1} = U_{j+1}$ und diese Werte sind auch maximal. Fahren wir in dieser Weise fort, können wir schlussfolgern, dass U konstant ist, wenn das Maximum im Inneren angenommen wird. Folglich wird das Maximum auch in den Endpunkten angenommen. Dies beweist (i). Fall (ii) wird in gleicher Weise wie Fall (ii) von Theorem 2.1 behandelt. Das analoge Minimumprinzip zeigt man durch Betrachten von $-U_j$. □

Genau wie beim stetigen Problem führt das Maximumprinzip zu einer Stabilitätsabschätzung in der diskreten Maximumnorm, was wir nun demonstrieren werden. Der Einfachheit halber nehmen wir $b = 0$ an. In diesem Kapitel werden wir für Gitterfunktionen

$$|U|_S = \max_{x_j \in S} |U_j| \tag{4.7}$$

schreiben.

Lemma 4.2. *Sei $\mathcal{A}_h$ wie in (4.3) mit $b = 0$. Dann gilt für jede Gitterfunktion U*

$$|U|_{\bar{\Omega}} \le \max\{|U_0|, |U_M|\} + C|\mathcal{A}_h U|_\Omega.$$

Die Konstante C hängt zwar von den Koeffizienten von $\mathcal{A}$ ab, aber nicht von h oder U.

Beweis. Sei $w(x) = x - x^2 = \frac{1}{4} - (x - \frac{1}{2})^2$ und $W_j = w(x_j)$. Dann ist mit $\underline{a} = \min_{\bar{\Omega}} a(x)$

$$\mathcal{A}_h W_j = 2a_j + c_j(x_j - x_j^2) \ge 2\underline{a}.$$

Setzen wir $V_j^\pm = \pm U_j - (2\underline{a})^{-1}|\mathcal{A}_h U|_\Omega W_j$, gilt daher

$$\mathcal{A}_h V_j^\pm = \pm \mathcal{A}_h U_j - (2\underline{a})^{-1}|\mathcal{A}_h U|_\Omega \mathcal{A}_h W_j \le 0,$$

sodass wir Lemma 4.1 anwenden können. Beachten Sie, dass die für h geforderte Bedingung im Falle $b = 0$ automatisch erfüllt ist. Weil $W_0 = W_M = 0$ ist, erhalten wir

$$\pm U_j - (2\underline{a})^{-1}|\mathcal{A}_h U|_\Omega W_j = V_j^\pm \le \max\{\pm U_0, \pm U_M, 0\} \le \max\{|U_0|, |U_M|\}.$$

Wegen $W_j \le \frac{1}{4}$ ist das Lemma mit $C = (8\underline{a})^{-1}$ bewiesen. □

Aus Lemma 4.2 folgt für $b = 0$ unmittelbar die Existenz und die Eindeutigkeit der Lösung von (4.3). Zum Beweis der Eindeutigkeit ist die Feststellung ausreichend, dass für den Fall $\mathcal{A}_h U = 0$ und $U_0 = U_M = 0$ die Gleichung $U = 0$ gilt. Aus der Eindeutigkeit folgt die Existenz einer Lösung, da das hier betrachtete Problem endlichdimensional ist. Für den Fall $b \neq 0$ verweisen wir auf Problemstellung 4.3.

Wir können nun eine Fehlerabschätzung aufstellen, die wir der Einfachheit halber lediglich für $b = 0$ demonstrieren.

Theorem 4.1. *Sei $b = 0$. Seien U und u die Lösungen von* (4.3) *beziehungsweise* (4.1). *Dann gilt*

$$|U - u|_{\Omega} \leq Ch^2 \|u\|_{\mathcal{C}^4}.$$

Beweis. An den inneren Gitterpunkten gilt für den Fehler $z_j = U_j - u(x_j)$

$$\mathcal{A}_h z_j = \mathcal{A}_h U_j - \mathcal{A}_h u(x_j) = f_j - \mathcal{A}_h u(x_j) = \mathcal{A}u(x_j) - \mathcal{A}_h u(x_j) =: \tau_j.$$

Mit (4.2) gilt für den *Rundungsfehler*

$$|\tau_j| = |-a_j(u''(x_j) - \partial\bar{\partial}u(x_j))| \leq Ch^2\|u\|_{\mathcal{C}^4}, \tag{4.8}$$

sodass wegen $z_0 = z_M = 0$ das gesuchte Resultat aus Lemma 4.2 folgt.

4.2 Die Poisson-Gleichung

Wir betrachten das Dirichlet-Problem für die Poisson-Gleichung

$$-\Delta u = f \quad \text{in } \Omega \quad \text{mit } u = 0 \quad \text{auf } \Gamma, \tag{4.9}$$

wobei Ω ein Gebiet in $\mathbf{R}^2$ mit Rand Γ ist. Wir nehmen zunächst an, dass Ω ein Quadrat ist, d. h. $\Omega = (0,1)\times(0,1) = \{x = (x_1, x_2),\ 0 < x_l < 1,\ l = 1,2\}$.

Zur Definition einer endlichen Differenzenapproximation schreiben wir $j = (j_1, j_2)$ mit den ganzen Zahlen j_1, j_2 und betrachten die Gitterpunkte $x_j = jh$ mit dem Gitterabstand $h = 1/M$ und die Gitterfunktion U mit $U_j = U(x_j)$. Wir benutzen mit $e_1 = (1,0),\ e_2 = (0,1)$ die Differenzenquotienten

$$\begin{aligned} \partial_l U_j &= \frac{U_{j+e_l} - U_j}{h}, & \bar{\partial}_l U_j &= \frac{U_j - U_{j-e_l}}{h}, \\ \partial_l\bar{\partial}_l U_j &= \frac{U_{j+e_l} - 2U_j + U_{j-e_l}}{h^2}, & & l = 1,2. \end{aligned} \tag{4.10}$$

Indem wir $f_j = f(x_j)$ setzen, können wir anschließend (4.9) durch

$$\begin{aligned} -\Delta_h U_j := -\partial_1\bar{\partial}_1 U_j - \partial_2\bar{\partial}_2 U_j = f_j \quad &\text{für } x_j \in \Omega, \\ U_j = 0 \quad &\text{für } x_j \in \Gamma \end{aligned} \tag{4.11}$$

ersetzen. Die Differenzengleichung in einem inneren Gitterpunkt in Ω kann in der Form

$$4U_j - U_{j+e_1} - U_{j-e_1} - U_{j+e_2} - U_{j-e_2} = h^2 f_j \quad \text{für } x_j \in \Omega \tag{4.12}$$

geschrieben werden. Das ist die Fünfpunkt-Approximation der Poisson-Gleichung. Das Problem (4.11) kann folglich in Matrixform $AU = g$ geschrieben werden, wobei A eine symmetrische $(M-1)^2 \times (M-1)^2$-Matrix ist, deren Elemente 4, -1 oder 0 sind, und das Element 0 am häufigsten vorkommt. Der Vektor $\bar{U}$ enthält die Werte an den inneren Gitterpunkten.

Es gilt das folgende diskrete Maximumprinzip.

Lemma 4.3. *Wenn U die Bedingung $-\Delta_h U_j \leq 0$ ($-\Delta_h U_j \geq 0$) für $x_j \in \Omega$ erfüllt, dann nimmt U sein Maximum (Minimum) für ein $x_j \in \Gamma$ an.*

Beweis. An den inneren Gitterpunkten können wir

$$U_j = \frac{U_{j+e_1} + U_{j-e_1} + U_{j+e_2} + U_{j-e_2}}{4} - \tfrac{1}{4} h^2 \Delta_h U_j$$

schreiben, sodass aus $-\Delta_h U_j \leq 0$ die Ungleichung $U_j \leq \frac{1}{4}(U_{j+e_1} + U_{j-e_1} + U_{j+e_2} + U_{j-e_2})$ folgt. Wenn U_j ein inneres Maximum ist, dann ist $U_j \geq \frac{1}{4}(U_{j+e_1} + U_{j-e_1} + U_{j+e_2} + U_{j-e_2})$. Deshalb gilt in diesem Fall die Gleichheit und der maximale Wert wird auch an allen benachbarten Punkten $x_{j\pm e_l}$ angenommen. Fahren wir in gleicher Weise fort, dann können wir schlussfolgern, dass U konstant ist, wenn das Maximum im Inneren angenommen wird. Damit ist das Lemma bewiesen. □

Wie im vorherigen Beispiel folgt aus dem Maximumprinzip auch eine Abschätzung für die Stabilität. Verwenden wir wieder die Notation (4.7) gilt das folgende Lemma.

Lemma 4.4. *Sei Δ_h wie in* (4.11) *definiert. Dann gilt für jede Gitterfunktion U*

$$|U|_{\bar{\Omega}} \leq |U|_\Gamma + C|\Delta_h U|_\Omega.$$

Beweis. Der Beweis wird analog zum Beweis von Theorem 3.2 geführt. Wir setzen zunächst $w(x) = \frac{1}{2} - |x - \bar{x}|^2 = x_1 + x_2 - x_1^2 - x_2^2$ mit $\bar{x} = (\frac{1}{2}, \frac{1}{2})$ und $x = (x_1, x_2)$ und definieren die Gitterfunktion $W_j = w(x_j)$. Dann ist $W_j \geq 0$ in Ω und $-\Delta_h W_j = 4$. Setzen wir $V_j^\pm = \pm U_j - \frac{1}{4}|\Delta_h U|_\Omega W_j$, dann können wir

$$-\Delta_h V_j^\pm = \mp \Delta_h U_j - |\Delta_h U|_\Omega \leq 0$$

schlussfolgern. Da außerdem $W_j \geq 0$ für $x_j \in \Gamma$ gilt, folgt aus Lemma 4.3 $V_j^\pm \leq |U|_\Gamma$. Wegen $W_j \leq \frac{1}{2}$ in Ω folgt daraus unsere Behauptung mit $C = 1/8$. □

Insbesondere folgt aus Lemma 4.4 die Eindeutigkeit der Lösung von (4.11) und somit auch die Existenz einer Lösung. Wie für das Zweipunkt-Randwertproblem impliziert das Lemma auch eine Fehlerabschätzung.

Theorem 4.2. *Seien U und u die Lösungen von* (4.11) *beziehungsweise* (4.9). *Dann gilt*

$$|U - u|_\Omega \leq C h^2 |u|_{\mathcal{C}^4}.$$

Beweis. An den inneren Gitterpunkten erfüllt der Fehler $z_j = U_j - u(x_j)$ die Gleichung

$$-\Delta_h z_j = f_j + \Delta_h u(x_j) = -\Delta u(x_j) + \Delta_h u(x_j) =: \tau_j,$$

wobei τ den Rundungsfehler angibt, der sich wie in (4.2) leicht durch

$$|\tau_j| \le \sum_{l=1}^{2} \left| \partial_l \bar{\partial}_l u(x_j) - \frac{\partial^2 u}{\partial x_l^2}(x_j) \right| \le C h^2 |u|_{\mathcal{C}^4} \tag{4.13}$$

abschätzen lässt. Das Resultat folgt deshalb durch Anwendung von Lemma 4.4 auf z_j, da $z_j = 0$ für $x_j \in \Gamma$ gilt. □

Die oben ausgeführte Analyse verwendet die Tatsache, dass alle Nachbarn der inneren Gitterpunkte in Ω entweder innere Gitterpunkte sind oder zu Γ gehören. Im Falle eines gekrümmten Randes kann dies jedoch nicht erreicht werden. Eine solche Situation werden wir nun kurz diskutieren.

Der Einfachheit halber nehmen wir an, dass Ω ein konvexes ebenes Gebiet mit einem glatten Rand Γ ist. Mit Ω_h bezeichnen wir diejenigen inneren Gitterpunkte x_j, für die alle vier Nachbarn von x_j ebenfalls in $\bar{\Omega}$ liegen. (Im vorhin betrachteten Fall eines Quadrates besteht Ω_h einfach aus allen inneren Gitterpunkten.) Für jedes $x_j \in \omega_h$ können wir dann einen (nicht notwendigerweise eindeutigen) Nachbarn $x_i \in \Omega_h \cup \omega_h$ auswählen, sodass die horizontale oder vertikale Gerade durch x_j und x_i den Rand Γ in einem Punkt $\bar{x}_j$ schneidet, der kein Gitterpunkt ist (siehe Abbildung 4.1). Für dieses $x_j \in \omega_h$ definieren wir nun den linearen Interpolationsoperator

$$\ell_h U_j := U_j - \alpha_j U_i - (1-\alpha_j) U(\bar{x}_j) \quad \text{mit } \alpha_j = \frac{|x_j - \bar{x}_j|}{h + |x_j - \bar{x}_j|} \le \tfrac{1}{2}. \tag{4.14}$$

Bezeichnen wir mit Γ_h die Punkte von Γ, die entweder Nachbarn von Punkten in Ω_h sind oder Punkte $\bar{x}_j$, die zu Punkten in ω_h gehören, dann können wir das Problem folgendermaßen stellen:

$$-\Delta_h U_j = f_j \quad \text{in } \Omega_h, \quad \ell_h U_j = 0 \quad \text{in } \omega_h \quad \text{und} \quad U = 0 \quad \text{auf } \Gamma_h. \tag{4.15}$$

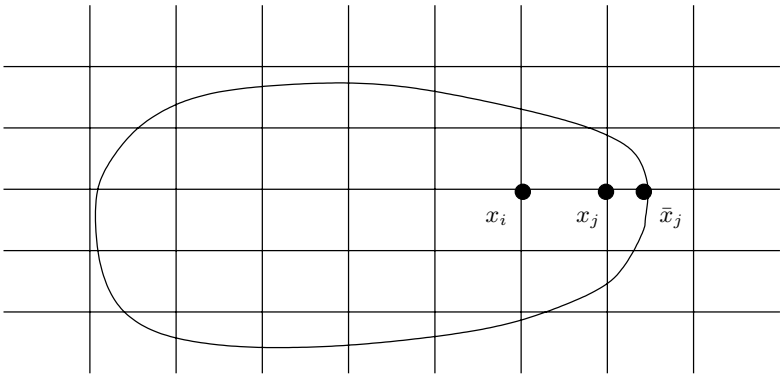

Abbildung 4.1. Interpolation in der Nähe des Randes.

In diesem Fall gilt folgende Stabilitätsabschätzung.

Lemma 4.5. *Ist Δ_h wie in* (4.11) *und ℓ_h wie in* (4.14) *definiert, gilt für eine beliebige Gitterfunktion U*

$$|U|_{\Omega_h \cup \omega_h} \leq 2\big(|U|_{\Gamma_h} + |\ell_h U|_{\omega_h} + C|\Delta_h U|_{\Omega_h}\big).$$

Beweis. Analog zum Beweis von Lemma 4.4 erhalten wir

$$|U|_{\Omega_h} \leq |U|_{\omega_h \cup \Gamma_h} + C|\Delta_h U|_{\Omega_h}.$$

Für $x_j \in \omega_h$ gilt nun

$$U_j = \ell_h U_j + \alpha_j U_i + (1-\alpha_j) U(\bar{x}_j) \quad \text{mit } 0 \leq \alpha_j \leq \tfrac{1}{2}$$

und folglich

$$|U|_{\omega_h} \leq |\ell_h U|_{\omega_h} + \tfrac{1}{2}|U|_{\Omega_h \cup \omega_h} + |U|_{\Gamma_h}.$$

Insgesamt zeigen diese Abschätzungen

$$\begin{aligned} |U|_{\Omega_h \cup \omega_h} &\leq |U|_{\omega_h \cup \Gamma_h} + C|\Delta_h U|_{\Omega_h} \\ &\leq |\ell_h U|_{\omega_h} + \tfrac{1}{2}|U|_{\Omega_h \cup \omega_h} + |U|_{\Gamma_h} + C|\Delta_h U|_{\Omega_h}, \end{aligned}$$

was den Beweis abschließt. □

Wiederum ergibt sich daraus die Eindeutigkeit und die Existenz einer Lösung von (4.15). Beachten Sie, dass in diesem Fall die zugehörige Matrix A nicht symmetrisch ist, da beispielsweise die zu den Punkten x_i und x_j in Abbildung 4.1 gehörigen Elemente a_{ij} und a_{ji} verschieden sind. Wir beenden das Kapitel mit folgender Fehlerabschätzung.

Theorem 4.3. *Seien U und u die Lösungen von* (4.15) *beziehungsweise* (4.9). *Dann ist*

$$|U - u|_{\Omega_h \cup \omega_h} \leq C h^2 \|u\|_{\mathcal{C}^4}.$$

Beweis. Wie im Beweis von Theorem 4.2 betrachten wir $z_j = U_j - u(x_j)$ und wenden nun Lemma 4.5 an. Der einzige neue Term ist $\ell_h z_j = -\ell_h u(x_j)$ und folglich ist $|\ell_h z_j| \leq C h^2 |u|_{\mathcal{C}^2}$, was den Beweis abschließt. □

Die oben beschriebene Methode der Interpolation in der Nähe des Randes geht auf L. Collatz zurück. Es ist auch möglich, für $-\Delta$ eine endliche Differenzenapproximation mit fünf Punkten zu verwenden, die auf den nicht äquidistanten Abständen auf ω_h beruht. Sie wird als Shortley-Weller-Approximation bezeichnet. Diese führt ebenfalls zu einer $O(h^2)$-Fehlerabschätzung.

4.3 Problemstellungen

Problem 4.1. Beweisen Sie (4.2) und (4.13) mithilfe der Taylorschen Formel.

Problem 4.2. Leiten Sie (4.5) her und zeigen Sie, dass die Matrix A tridiagonal und (zeilenweise) diagonaldominant, d. h. $\sum_{j\neq i} |a_{ij}| \leq a_{ii}$, ist, wenn h hinreichend klein ist. Hinweis: Nehmen Sie $a_j \pm \frac{1}{2}hb_j \geq 0$ an.

Problem 4.3. Zeigen Sie, dass die Schlussfolgerung von Lemma 4.2 (und folglich auch die von Theorem 4.1) auch für $b \neq 0$ gilt, wenn h hinreichend klein ist und eine Gitterfunktion W zur Verfügung steht, sodass $\mathcal{A}_h W_j \geq 1$ für $x_j \in \Omega$ und $W_j \geq 0$ für $x_j \in \bar{\Omega}$ gilt. Konstruieren Sie eine solche Funktion. (Hinweis: Verwenden Sie die Funktion $w(x) = \mathrm{e}^{\lambda} - \mathrm{e}^{\lambda x}$ mit einem geeignet gewählten λ.)

Problem 4.4. (Übung am Rechner.) Betrachten Sie das Zweipunkt-Randwertproblem

$$-u'' + u = 2x \quad \text{in } (0,1) \quad \text{mit } u(0) = u(1) = 0.$$

Wenden Sie das finite Differenzenverfahren (4.3) mit $h = 1/10,\ 1/20$ an. Bestimmen Sie die exakte Lösung und berechnen Sie den maximalen Fehler an den Gitterpunkten.

Problem 4.5. (Übung am Rechner.) Betrachten Sie das Dirichlet-Problem (4.9) mit

$$f(x) = \sin(\pi x_1)\sin(\pi x_2) + \sin(\pi x_1)\sin(2\pi x_2)$$

in $\Omega = (0,1) \times (0,1)$. Berechnen Sie die approximative Lösung mithilfe des finiten Differenzenverfahrens (4.11) mit den Werten $h = 1/10,\ 1/20$ und bestimmen Sie den Fehler an der Stelle (0.5, 0.5). Verwenden Sie dazu die exakte Lösung

$$u(x) = (2\pi^2)^{-1}\sin(\pi x_1)\sin(\pi x_2) + (5\pi^2)^{-1}\sin(\pi x_1)\sin(2\pi x_2).$$

5

Die Methode der finiten Elemente für elliptische Gleichungen

In den letzten Jahrzehnten hat sich die von Ingenieuren in den sechziger Jahren eingeführte *Methode der finiten Elemente* zu der vielleicht wichtigsten numerischen Methode für partielle Differentialgleichungen entwickelt, insbesondere für elliptische und parabolische Gleichungen. Diese Methode beruht auf einer Variationsform des Randwertproblems und approximiert die exakte Lösung durch eine stückweise polynomiale Funktion. Sie ist viel leichter an die Geometrie des zugrunde liegenden Gebietes anzupassen als das finite Differenzenverfahren. Für symmetrische, positiv definite elliptische Probleme reduziert sie sich auf ein endliches, lineares System mit einer symmetrischen, positiv definiten Matrix.

Zunächst führen wir diese Methode in Abschnitt 5.1 am Beispiel eines Zweipunkt-Randwertproblems ein und beweisen eine Vielzahl von Fehlerabschätzungen. In Abschnitt 5.2 formulieren wir die Methode dann für ein zweidimensionales Modellproblem. In diesem Fall werden die stückweise polynomialen Näherungen auf Triangulationen des räumlichen Gebietes definiert. In dem folgenden Abschnitt 5.3 untersuchen wir solche Approximationen detaillierter. In Abschnitt 5.4 beweisen wir grundlegende Fehlerabschätzungen für die Methode der finite Elemente im Fall des Modellproblems, indem wir stückweise lineare Approximationsfunktionen verwenden. Alle bis zu diesem Punkt abgeleiteten Fehlerschranken enthalten eine Norm der unbekannten exakten Lösung und werden deshalb oft als *a priori* Fehlerabschätzungen bezeichnet. In Abschnitt 5.5 beschäftigen wir uns mit einer sogenannten *a posteriori* Fehlerabschätzung, bei der die Fehlerschranke als Funktion der Problemdaten und der numerisch bestimmten Lösung ausgedrückt wird. In Abschnitt 5.6 untersuchen wir den Effekt der numerischen Integration, die häufig benutzt wird, wenn die Finite-Elemente-Gleichung in ein Computerprogramm eingebaut wird. In Abschnitt 5.7 beschreiben wir kurz eine sogenannte *Methode der gemischten finiten Elemente*.

5.1 Ein Zweipunkt-Randwertproblem

Wir betrachten den Spezialfall $b = 0$ des im Abschnitt 2.3 behandelten Zweipunkt-Randwertproblems

$$\mathcal{A}u := -(au')' + cu = f \quad \text{in } \Omega := (0,1) \quad \text{mit } u(0) = u(1) = 0, \tag{5.1}$$

bei dem $a = a(x)$ und $c = c(x)$ glatte Funktionen mit $a(x) \geq a_0 > 0,\ c(x) \geq 0$ in $\bar{\Omega}$ sind und $f \in L_2 = L_2(\Omega)$ ist. Es sei daran erinnert, dass die Variationsformulierung dieses Problems darin besteht, ein $u \in H_0^1$ zu bestimmen, für das

$$a(u,\varphi) = (f,\varphi) \quad \forall \varphi \in H_0^1 \tag{5.2}$$

mit

$$a(v,w) = \int_\Omega (av'w' + cvw)\,\mathrm{d}x \quad \text{und} \quad (f,v) = \int_\Omega fv\,\mathrm{d}x$$

gilt und dieses Problem eine eindeutige Lösung $u \in H^2$ besitzt.

Um eine approximative Lösung von (5.2) zu bestimmen, führen wir eine Zerlegung

$$0 = x_0 < x_1 < \cdots < x_M = 1$$

von Ω ein und setzen

$$h_j = x_j - x_{j-1},\ K_j = [x_{j-1}, x_j] \quad \text{für } j = 1,\ldots,M \quad \text{und} \quad h = \max_j h_j.$$

Nach der diskreten Lösung wird im endlichdimensionalen Funktionenraum

$$S_h = \big\{v \in \mathcal{C} = \mathcal{C}(\bar{\Omega}) : v \text{ linear auf jedem } K_j,\ v(0) = v(1) = 0\big\}$$

gesucht. (Unter einer linearen Funktion verstehen wir eine Funktion der Form $f(x) = \alpha x + \beta$; streng genommen wird eine solche Funktion im Falle $\beta \neq 0$ als affine Funktion bezeichnet.) Man kann leicht sehen, dass $S_h \subset H_0^1$ ist. Die Menge $\{\Phi_i\}_{i=1}^{M-1} \subset S_h$ der durch

$$\Phi_i(x_j) = \begin{cases} 1 & \text{für } i = j, \\ 0 & \text{für } i \neq j \end{cases}$$

definierten *Hutfunktionen* (siehe Abbildung 5.1) bildet eine Basis für S_h und jedes $v \in S_h$ kann in der Form

$$v(x) = \sum_{i=1}^{M-1} v_i \Phi_i(x) \quad \text{mit } v_i = v(x_i)$$

geschrieben werden.

Wir stellen nun das endlichdimensionale Problem, ein $u_h \in S_h$ mit

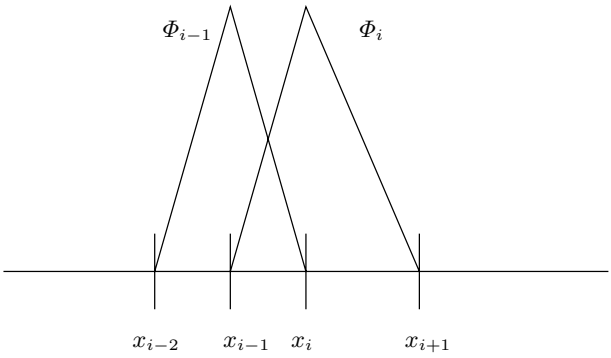

Abbildung 5.1. Hutfunktionen.

$$a(u_h, \chi) = (f, \chi) \quad \forall \chi \in S_h \tag{5.3}$$

zu finden. Unter Verwendung der Basis $\{\Phi_i\}_{i=1}^{M-1}$ schreiben wir $u_h(x) = \sum_{j=1}^{M-1} U_j \Phi_j(x)$ und setzen dies in (5.3) ein. Wir stellen fest, dass dies zu

$$\sum_{j=1}^{M-1} U_j a(\Phi_j, \Phi_i) = (f, \Phi_i) \quad \text{für } i = 1, \ldots, M-1 \tag{5.4}$$

äquivalent ist. Dieses lineare Gleichungssystem kann in Matrixform

$$A\,U = b \tag{5.5}$$

ausgedrückt werden. Dabei ist $U = (U_i)$, $A = (a_{ij})$ die *Steifigkeitsmatrix* mit den Elementen $a_{ij} = a(\Phi_j, \Phi_i)$ und $b = (b_i)$ der *Lastvektor* mit den Elementen $b_i = (f, \Phi_i)$. Die Matrix A ist symmetrisch und positiv definit, da für $V = (V_i)$ und $v(x) = \sum_{i=1}^{M-1} V_i \Phi_i(x)$

$$V^{\mathrm{T}} A\, V = \sum_{i,j=1}^{M-1} V_i a_{ij} V_j = a\Big(\sum_{j=1}^{M-1} V_j \Phi_j, \sum_{i=1}^{M-1} V_i \Phi_i \Big) = a(v, v) \geq a_0 \|v'\|^2$$

gilt. Somit folgt aus $V^{\mathrm{T}} A V = 0$ die Gleichung $v' = 0$, sodass v wegen $v(0) = 0$ konstant $= 0$ und folglich $V = 0$ ist. Es folgt, dass (5.5), und deshalb auch (5.3), eine eindeutige Lösung besitzt, die als *Finite-Elemente-Lösung* von (5.1) bezeichnet wird. Die Matrix A ist tridiagonal, weil im Falle $|i - j| \geq 2$, d. h. wenn x_i und x_j keine Nachbarn sind, $a_{ij} = 0$ gilt. Deshalb ist das System (5.5) einfach zu lösen.

Wir stellen fest, dass Gleichung (5.4) für $\mathcal{A}u = -u''$ und konstanten Gitterabstand, d. h. $h_j = h = 1/M$ mit $j = 1, \ldots, M$ und der Notation aus Abschnitt 4.1, in der Form

$$-\partial\bar{\partial} U_j = h^{-1}(f, \Phi_j), \quad j = 1, \ldots, M-1 \tag{5.6}$$

geschrieben werden kann (siehe Problemstellung 5.2). Die Methode der finiten Elemente stimmt also mit der finiten Differenzenapproximation (4.3) überein, abgesehen davon, dass nun ein Mittelwert von f über $(x_j - h, x_j + h)$ anstelle eines Punktwertes $f_j = f(x_j)$ verwendet wird.

Die Idee, den Raum H_0^1 in (5.2) durch einen endlichdimensionalen Teilraum zu ersetzen und die Koeffizienten der zugehörigen approximativen Lösung wie in (5.4) zu bestimmen, wird als Galerkin-Methode bezeichnet. Die Methode der finiten Elemente entspricht also der Galerkin-Methode mit einer speziellen Wahl des endlichdimensionalen Teilraums. In diesem Fall ist dies der Raum der stetigen, stückweise linearen Funktionen. Die Intervalle K_j werden dann zusammen mit den Restriktionen dieser Funktionen auf K_j als finite Elemente betrachtet.

Bevor wir den Fehler in der Finite-Elemente-Lösung u_h analysieren, diskutieren wir einige Approximationseigenschaften des Raumes S_h. Dazu definieren wir die stückweise lineare Interpolierte $I_h v \in S_h$ einer Funktion $v \in \mathcal{C} = \mathcal{C}(\bar{\Omega})$ mit $v(0) = v(1) = 0$ durch

$$I_h v(x_j) = v(x_j), \quad j = 1, \ldots, M-1.$$

Es sei daran erinnert, dass wegen der Sobolevschen Ungleichung, Theorem A.5, in einer Dimension $H_0^1 \subset \mathcal{C}$ gilt, sodass $I_h v$ für $v \in H_0^1$ definiert ist. Man kann zeigen, dass mit $\|v\|_{K_j} = \|v\|_{L_2(K_j)}$ und $|v|_{2,K_j} = |v|_{H^2(K_j)}$

$$\|I_h v - v\|_{K_j} \le C h_j^2 |v|_{2,K_j} \tag{5.7}$$

und

$$\|(I_h v - v)'\|_{K_j} \le C h_j |v|_{2,K_j} \tag{5.8}$$

gilt, was Sie in Problemstellung 5.1 tun sollen. Daraus folgt

$$\begin{aligned} \|I_h v - v\| &= \Big(\sum_{j=1}^{M} \|I_h v - v\|_{K_j}^2 \Big)^{1/2} \le \Big(\sum_{j=1}^{M} C^2 h_j^4 |v|_{2,K_j}^2 \Big)^{1/2} \\ &\le C h^2 \|v\|_2 \quad \forall v \in H^2, \end{aligned} \tag{5.9}$$

und analog

$$\|(I_h v - v)'\| \le C h \|v\|_2 \quad \text{für } v \in H^2. \tag{5.10}$$

Wir wenden uns nun der Aufgabe zu, den Fehler bei der durch (5.3) definierten Finite-Elemente-Approximation u_h zu bestimmen. Weil $a(\cdot,\cdot)$ symmetrisch und positiv definit ist, handelt es sich um ein Skalarprodukt auf H_0^1. Die zugehörige Norm ist die Energienorm

$$\|v\|_a = a(v,v)^{1/2} = \Big(\int_0^1 \big(a(v')^2 + c v^2 \big) \, \mathrm{d}x \Big)^{1/2}. \tag{5.11}$$

Theorem 5.1. *Seien u_h und u die Lösungen von* (5.3) *beziehungsweise* (5.2). *Dann gilt*

$$\|u_h - u\|_a = \min_{\chi \in S_h} \|\chi - u\|_a, \tag{5.12}$$

und

$$\|u_h' - u'\| \le Ch\|u\|_2. \tag{5.13}$$

Beweis. Wegen $S_h \subset H_0^1$ können wir in (5.2) $\varphi = \chi \in S_h$ wählen und dies von (5.3) subtrahieren. Wir erhalten

$$a(u_h - u, \chi) = 0 \quad \forall \chi \in S_h. \tag{5.14}$$

Diese Gleichung besagt, dass die Finite-Elemente-Lösung u_h als orthogonale Projektion der exakten Lösung u auf S_h bezüglich des Skalarproduktes $a(\cdot,\cdot)$ ausgedrückt werden kann. Daraus folgt unmittelbar, dass u_h die beste Approximation von u in S_h bezüglich der Energienorm ist und folglich (5.12) erfüllt ist. Davon kann man sich folgendermaßen überzeugen: Für jedes $\chi \in S_h$ gilt unter Verwendung von (5.14)

$$\|u_h - u\|_a^2 = a(u_h - u, u_h - u) = a(u_h - u, \chi - u) \le \|u_h - u\|_a \, \|\chi - u\|_a,$$

was nach Wegstreichen eines Faktors $\|u_h - u\|_a$ Gleichung (5.12) beweist. Aufgrund unserer Annahme gilt mit einem von h unabhängigen C

$$\sqrt{a_0}\|v'\| \le \|v\|_a \le C\|v'\| \quad \text{für } v \in H_0^1,$$

wobei die erste Ungleichung wegen (5.11) offensichtlich ist und die zweite aus (2.17) folgt. Somit ergibt sich aus (5.12)

$$\|(u_h - u)'\| \le C\|u_h - u\|_a \le C \min_{\chi \in S_h} \|(\chi - u)'\|. \tag{5.15}$$

Nehmen wir $\chi = I_h u$ an und verwenden wir die Schranke für den Interpolationsfehler in Gleichung (5.10), so erhalten wir (5.13), was den Beweis vervollständigt. □

Unser nächstes Resultat bezieht sich auf die L_2-Norm des Fehlers.

Theorem 5.2. *Seien u_h und u die Lösungen von* (5.3) *beziehungsweise* (5.2). *Dann gilt*

$$\|u_h - u\| \le Ch^2\|u\|_2. \tag{5.16}$$

Beweis. Wir verwenden ein Dualitätsargument, das auf dem Hilfsproblem

$$\mathcal{A}\phi = e \quad \text{in } \Omega \quad \text{mit } \phi(0) = \phi(1) = 0 \text{ und } e = u_h - u \tag{5.17}$$

beruht. Dessen schwache Formulierung besteht darin, ein $\phi \in H_0^1$ zu bestimmen, für das

$$a(w, \phi) = (w, e) \quad \forall w \in H_0^1 \tag{5.18}$$

gilt. Wir verwenden die Testfunktion w linksseitig, da (5.18) für (5.2) die Rolle des adjungierten (oder dualen) Problems übernimmt. An dieser Stelle führt dies natürlich zu keinem Unterschied, da $a(\cdot, \cdot)$ symmetrisch ist. Im Falle eines unsymmetrischen Differentialoperators $\mathcal{A}$ ist dies allerdings wesentlich (siehe Problemstellung 5.7). Wegen der Regularitätsabschätzung (2.22) gilt

$$\|\phi\|_2 \le C\|\mathcal{A}\phi\| = C\|e\|. \tag{5.19}$$

Setzen wir $w = e$ in (5.18) und verwenden (5.14) und (5.10), so erhalten wir deshalb

$$\begin{aligned}\|e\|^2 = a(e, \phi) = a(e, \phi - I_h\phi) &\le C\|e'\| \, \|(\phi - I_h\phi)'\| \\ &\le Ch\|e'\| \, \|\phi\|_2 \le Ch\|e'\| \, \|e\|.\end{aligned}$$

Nach Wegstreichen eines Faktors $\|e\|$ erkennen wir, dass wir hier einen Faktor h gegenüber der Fehlerabschätzung von e' gewonnen haben, es gilt also

$$\|e\| \le Ch\|e'\|. \tag{5.20}$$

Der Beweis wird durch Einsetzen von (5.13) vervollständigt. □

Anmerkung 5.1. Wir sehen, dass die obige Fehlerabschätzung die Norm der zweiten Ableitung enthält, während beim entsprechenden Resultat im Falle des finiten Differenzenverfahrens in Theorem 4.1 die vierte Ableitung gebraucht wurde. Dies hängt mit der Tatsache zusammen, dass der Lastterm f bei der Methode der finiten Elemente durch Mittelwerte einbezogen wird, und nicht, wie beim finiten Differenzenverfahren, in Form von Punktwerten. Dies werden wir in Abschnitt 5.6 näher erläutern.

Anmerkung 5.2. Die Lösung der sehr speziellen Gleichung (5.6) stimmt mit den Knotenwerten der exakten Lösung des zugehörigen Zweipunkt-Randwertproblems überein. Tatsächlich gilt mit der exakten Lösung $u = u(x)$ unter Verwendung der Taylorschen Formel

$$\begin{aligned}\partial\bar{\partial}u(x_j) &= h^{-2}\int_{x_{j-1}}^{x_j}(y - x_{j-1})u''(y)\,\mathrm{d}y + h^{-2}\int_{x_j}^{x_{j+1}}(x_{j+1} - y)u''(y)\,\mathrm{d}y \\ &= h^{-1}(u'', \Phi_j) = -h^{-1}(f, \Phi_j).\end{aligned}$$

Somit ist die Finite-Elemente-Lösung u_h mit der Interpolierten $I_h u$ der exakten Lösung identisch. In Problemstellung 5.4 diskutieren wir diesen Sachverhalt auf Grundlage der Greenschen Funktion.

Bei der oben ausgeführten Analyse hätten wir auch einen allgemeineren Raum der finiten Elemente betrachten können, der aus stückweisen Polynomen des Grades $r-1$ mit einer ganzen Zahl $r \geq 2$ besteht. Der oben betrachtete Fall stückweiser linearer Funktionen ist darin mit $r = 2$ enthalten. Dies wäre also der Raum

$$S_h = \big\{ v \in \mathcal{C} : v \in \Pi_{r-1} \text{ auf jedem } K_j,\ v(0) = v(1) = 0 \big\},$$

wobei Π_k ein Polynom vom Grad $\leq s$ beschreibt. Zusätzlich zu den oben erwähnten Hutfunktionen können wir dann jedem Intervall K_i die Basisfunktionen $\Phi_{ij} \in \Pi_{r-1}$ auf K_i mit $j = 1, \ldots, r-2$ zuordnen, die außerhalb von K_i verschwinden und durch

$$\Phi_{ij}(x_{i,l}) = \begin{cases} 1 & \text{für } j = l, \\ 0 & \text{für } j \neq l \end{cases} \quad \text{mit } x_{i,l} = x_{i-1} + h_i \frac{l}{r-1},\ l = 0, \ldots, r-1$$

definiert sind.

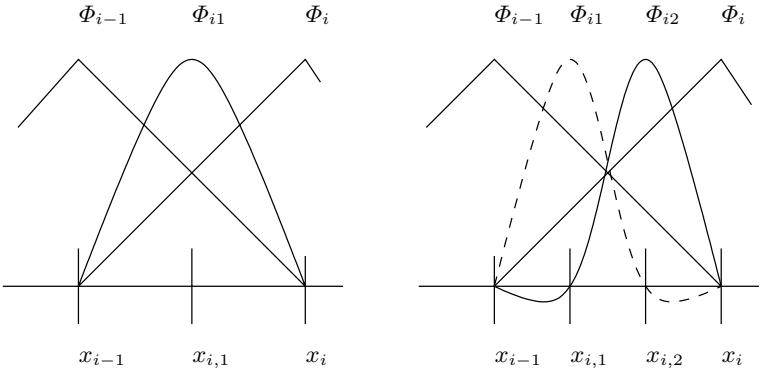

Abbildung 5.2. Globale Basisfunktionen für $r = 3$ und 4.

Verwenden wir bei der Definition der Interpolierten $I_h v$ auch diese zusätzlichen Knotenpunkte, kann man die folgenden lokalen Abschätzungen

$$\|I_h v - v\|_{K_j} \leq C h_j^r \|v^{(r)}\|_{K_j} \quad \text{und} \quad \|(I_h v - v)'\|_{K_j} \leq C h_j^{r-1} \|v^{(r)}\|_{K_j}$$

und folglich die globalen Abschätzungen

$$\|I_h v - v\| \leq C h^r \|v\|_r \quad \text{und} \quad \|(I_h v - v)'\| \leq C h^{r-1} \|v\|_r \quad \forall v \in H^r \tag{5.21}$$

beweisen. Für die Finite-Elemente-Lösung erhält man wie oben

$$\|u_h - u\| \leq C h^r \|u\|_r \quad \text{und} \quad \|u_h' - u'\| \leq C h^{r-1} \|u\|_r. \tag{5.22}$$

Diese Ungleichungen fordern also $v, u \in H^r$. Da die Interpolierte $I_h v$ für $v \in H_0^1$ wohldefiniert ist, kann man zeigen, dass diese auch dann richtig ist, wenn r durch ein beliebiges s mit $1 \leq s \leq r$ ersetzt wird. Der Fall $s = 2$ in der zweiten Abschätzung aus (5.21) wird im Beweis der $O(h^r)$-Abschätzung in (5.22) durch das Dualitätsargument benötigt.

5.2 Ein Modellproblem in der Ebene

Sei Ω nun ein polygonales Gebiet in $\mathbf{R}^2$, d. h. ein Gebiet, dessen Rand Γ durch ein Polygon gebildet wird. Betrachten wir das einfache Modellproblem

$$\mathcal{A}u := -\nabla \cdot \big(a \nabla u\big) = f \quad \text{in } \Omega \quad \text{mit } u = 0 \quad \text{auf } \Gamma. \tag{5.23}$$

Wir nehmen an, dass der Koeffizient $a = a(x)$ mit $a(x) \geq a_0 > 0$ in $\bar{\Omega}$ glatt ist und $f \in L_2$ gilt.

Aus Abschnitt 3.5 wissen wir, dass bei der Variationsformulierung von (5.23) ein $u \in H_0^1$ gesucht ist, sodass

$$a(u, v) = (f, v) \quad \forall v \in H_0^1 \tag{5.24}$$

mit

$$a(v, w) = \int_\Omega a \nabla v \cdot \nabla w \, \mathrm{d}x \quad \text{und} \quad (f, v) = \int_\Omega f v \, \mathrm{d}x$$

gilt, und dass dieses Problem eine eindeutige Lösung in H_0^1 besitzt. Nehmen wir darüber hinaus an, dass Ω konvex ist, folgt aus der Regularitätsabschätzung (3.36) $u \in H^2$ und

$$\|u\|_2 \leq C \|f\|. \tag{5.25}$$

Die Erläuterung zur Approximation von (5.23) folgt ähnlichen Gedankengängen wie im Falle des oben besprochenen Zweipunkt-Randwertproblems. Diesmal unterteilen wir das polygone Gebiet in Dreiecke. Genauer sei $\mathcal{T}_h = \{K\}$ eine Menge von abgeschlossenen Dreiecken K, d. h. eine *Triangulation* von Ω, für die

$$\bar{\Omega} = \bigcup_{K \in \mathcal{T}_h} K, \quad h_K = \operatorname{diam}(K), \quad h = \max_{K \in \mathcal{T}_h} h_K$$

gilt. Die Eckpunkte P der Dreiecke $K \in \mathcal{T}_h$ werden als Knoten der Triangulation $\mathcal{T}_h$ bezeichnet. Wir fordern, dass die Schnittmenge zweier beliebiger Dreiecke aus $\mathcal{T}_h$ entweder leer, ein Knoten oder eine gemeinsame Kante ist, und dass sich kein Knoten innerhalb einer Kante von $\mathcal{T}_h$ befindet (siehe Abbildung 5.3).

Der Triangulation $\mathcal{T}_h$ ordnen wir den Funktionenraum S_h zu, der aus stetigen, stückweise linearen Funktionen auf $\mathcal{T}_h$ besteht, die auf Γ verschwinden, d. h.

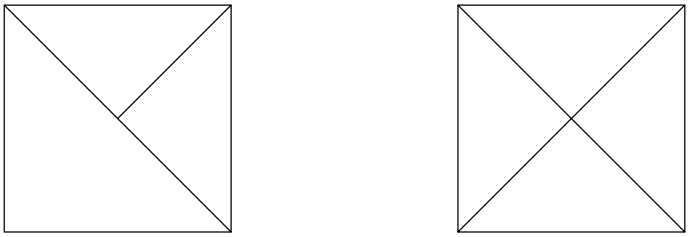

Abbildung 5.3. Unzulässige (links) und zulässige Triangulation (rechts).

$$S_h = \big\{ v \in \mathcal{C}(\bar{\Omega}) : v \text{ linear in } K \text{ für jedes } K \in \mathcal{T}_h,\ v = 0 \text{ auf } \Gamma \big\}.$$

Mit den oben genannten Annahmen bezüglich $\mathcal{T}_h$ ist es nicht schwierig, $S_h \subset H_0^1$ zu verifizieren. Sei $\{P_i\}_{i=1}^{M_h}$ die Menge der inneren Knoten, d. h. derjenigen, die sich nicht auf Γ befinden. Ein Funktion in S_h ist somit eindeutig durch ihre Werte an den Punkten P_j bestimmt, und die Menge der durch

$$\Phi_i(P_j) = \begin{cases} 1 & \text{für } i = j, \\ 0 & \text{für } i \neq j \end{cases}$$

definierten *Pyramidenfunktionen* $\{\Phi_i\}_{i=1}^{M_h} \subset S_h$ bildet eine Basis von S_h. Wenn $v \in S_h$ ist, dann gilt also $v(x) = \sum_{i=1}^{M_h} v_i \Phi_i(x)$, wobei $v_i = v(P_i)$ die Knotenwerte von v sind. Daraus folgt, dass S_h ein endlichdimensionaler Teilraum des Hilbert-Raumes H_0^1 ist.

Bei der Finite-Elemente-Approximation des Problems (5.24) ist also ein $u_h \in S_h$ gesucht, das die Gleichung

$$a(u_h, \chi) = (f, \chi) \quad \forall \chi \in S_h \tag{5.26}$$

erfüllt. Mithilfe der Basis $\{\Phi_i\}_{i=1}^{M_h}$ schreiben wir $u_h(x) = \sum_{i=1}^{M_h} U_i \Phi_i(x)$, was, eingesetzt in (5.26), ein lineares Gleichungssystem zur Bestimmung der U_j liefert:

$$\sum_{j=1}^{M_h} U_j a(\Phi_j, \Phi_i) = (f, \Phi_i), \quad i = 1, \ldots, M_h. \tag{5.27}$$

Dies kann in Matrixform $AU = b$ geschrieben werden, wobei $U = (U_i)$, $A = (a_{ij})$ die Steifigkeitsmatrix mit den Elementen $a_{ij} = a(\Phi_j, \Phi_i)$ und $b = (b_i)$ der Lastvektor mit den Elementen $b_i = (f, \Phi_i)$ ist. Die Matrix A ist wie im Abschnitt 5.1 symmetrisch und positiv definit, sodass (5.27) und folglich auch (5.26) eine eindeutige Lösung in S_h besitzt. Darüber hinaus ist die Matrix A groß und im Falle eines feinen Gitters dünn besetzt, d. h. ein großer Anteil ihrer Elemente ist null. Das liegt daran, dass jedes Φ_i, abgesehen von der Menge der Dreiecke, die den Knoten P_i enthalten, verschwindet, sodass $a_{ij} =$

$a(\Phi_j, \Phi_i) = 0$ gilt, wenn P_i und P_j keine Nachbarn sind. Diese Eigenschaft ist für die effiziente Lösung des linearen Gleichungssystems wesentlich (siehe Anhang B). Diesmal sind die finiten Elemente die Dreiecke $K \in \mathcal{T}_h$ zusammen mit den Restriktionen der Funktionen in S_h auf K.

Allgemeiner gesagt, können wir bei gegebener Triangulation $\mathcal{T}_h$ für S_h die Funktionen auf Ω wählen, die sich auf den Dreiecken $K \in \mathcal{T}_h$ auf Polynome vom Grad $r-1$ reduzieren. Dabei ist r eine feste ganze Zahl ≥ 2. Man kann zeigen, dass eine solche Funktion χ eindeutig durch ihre Werte an einer bestimmten endlichen Anzahl von Knoten in jedem K bestimmt ist. Diese können auf verschiedene Weise gewählt werden. Im Fall $r = 3$, d. h. wenn S_h aus stückweise quadratischen Funktionen besteht, können die Eckpunkte von $\mathcal{T}_h$ zusammen mit den Kantenmittelpunkten in $\mathcal{T}_h$ ausgewählt werden, insgesamt also sechs Punkte für jedes $K \in \mathcal{T}_h$. Für stückweise kubische Funktionen, d. h. im Falle $r = 4$, können wir die Eckpunkte von $\mathcal{T}_h$, zwei innere Punkte auf jeder Kante von $\mathcal{T}_h$ und den Schwerpunkt jedes $K \in \mathcal{T}_h$ auswählen. Somit verwenden wir für jedes $K \in \mathcal{T}_h$ insgesamt zehn Punkte (siehe Abbildung 5.4). Beachten Sie, dass ein Polynom in zwei Variablen von zweiter und dritter Ordnung eindeutig durch die Werte von sechs beziehungsweise zehn Koeffizienten festgelegt ist. Zu deren Bestimmung ist genau diese Anzahl linearer Bedingungen oder Freiheitsgrade, wie sie in diesem Zusammenhang bezeichnet werden, erforderlich.

Der so definierte Finite-Elemente-Raum S_h ist noch immer ein endlichdimensionaler Teilraum von H_0^1, und jedem dieser beschriebenen Knoten kann eine Basisfunktion $\Phi_j \in S_h$ zugeordnet werden. Das Finite-Elemente-Problem (5.26) und dessen Matrixformulierung (5.5) bleiben von derselben Form wie bisher.

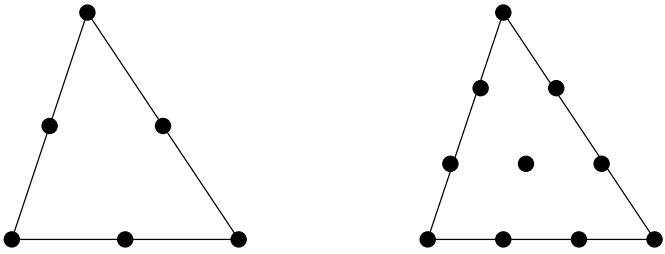

Abbildung 5.4. Dreiecke mit sechs und zehn Knoten.

Wenn der Rand Γ von Ω kein Polynom sondern eine glatte Kurve ist, dann wird eine Triangulation vom obigen Typ das Gebiet Ω nicht genau ausfüllen. Ist Ω konvex, kann die Triangulation so gewählt werden, dass die Vereinigung Ω_h der Dreiecke das Gebiet Ω noch immer approximiert. Dazu wählt man die Randknoten von Ω_h auf Γ so, dass die Menge $\Omega \setminus \Omega_h$ der Punkte in Ω, die nicht durch die Triangulation überdeckt ist, eine Breite der Ordnung $O(h^2)$

besitzt (siehe Abbildung 5.5). Werden die Funktionen in S_h so definiert, dass sie auf $\Omega \setminus \Omega_h$ verschwinden, dann kann eine Finite-Elemente-Lösung u_h wie oben definiert werden. Es stellt sich heraus, dass für ein S_h, das aus stückweise linearen Funktionen besteht, durch diese Erweiterung nichts verloren geht (siehe Abschnitt 5.3). Im Falle stückweiser Polynome höherer Ordnung ist die Situation jedoch weniger günstig. Es wurden deshalb verschieden Modifikationen dieser Methoden entwickelt, die sich mit der Approximation in der Nähe von Γ beschäftigen. Wir werden uns damit nicht detaillierter befassen, merken jedoch an, dass eine Triangulation gegenüber einem, beim finiten Differenzenverfahren verwendeten Quadratgitter in jedem Falle eine flexiblere Methode darstellt, ein Gebiet Ω zu approximieren und dass dies eine nützliche Eigenschaft der Methode der finiten Elemente ist.

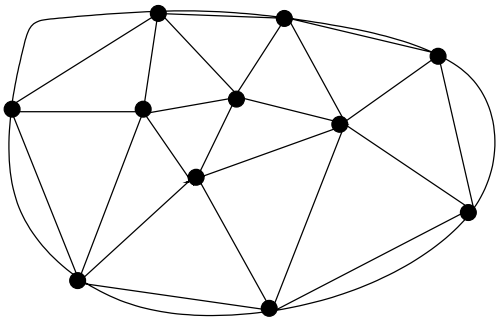

Abbildung 5.5. Ein glattes konvexes Gebiet mit Triangulation.

Im Folgenden gehen wir davon aus, dass wir nicht nur eine Triangulation $\mathcal{T}_h$ und deren zugehörigen Funktionenraum S_h betrachten, sondern eine Familie von Triangulationen $\{\mathcal{T}_h\}_{0<h<1}$ mit den zugehörigen Finite-Elemente-Räumen $\{S_h\}_{0<h<1}$. Eine wesentliche Aufgabe besteht darin zu bestimmen, wie schnell der Fehler $u_h - u$ gegen null geht, wenn h gegen null strebt.

5.3 Einige Aspekte der Approximationstheorie

Sei $\tilde{S}_h$ der Raum der stetigen, stückweise linearen Funktionen auf der Triangulation $\mathcal{T}_h$, wobei nicht gefordert wird, dass die Funktionen auf Γ verschwinden. Mit $\{P_j\}_{j=1}^{N_h}$ bezeichnen wir alle Knoten von $\mathcal{T}_h$, einschließlich der auf Γ mit $M_h + 1 \leq j \leq N_h$. Die zugehörigen Pyramidenfunktionen sind $\{\Phi_j\}_{j=1}^{N_h}$. Wir definieren den Interpolationsoperator $I_h : \mathcal{C}(\bar{\Omega}) \to \tilde{S}_h$ durch

$$(I_h v)(x) = \sum_{i=1}^{N_h} v_i \Phi(x), \quad \text{wobei } v_i = v(P_i) \text{ ist.} \tag{5.28}$$

Die Interpolierte $I_h v$ stimmt somit an den Knoten P_j mit v überein:

$$(I_h v)(P_i) = v(P_i) \quad \text{mit } i = 1, \ldots, N_h.$$

Wenn v auf Γ verschwindet, dann gehört $I_h v$ zum Finite-Elemente-Raum S_h, der im letzten Abschnitt eingeführt wurde. Eine analoge Definition kann auch im oben beschriebenen, allgemeineren Fall stückweiser Polynome vom Grad $r-1$ verwendet werden.

Im Falle stückweiser linearer Funktionen kann man die folgenden lokalen Fehlerabschätzungen beweisen: Es gilt

$$\|I_h v - v\|_K \le C_K h_K^2 |v|_{2,K} \quad \forall K \in \mathcal{T}_h \tag{5.29}$$

und

$$\|\nabla(I_h v - v)\|_K \le C_K h_K |v|_{2,K} \quad \forall K \in \mathcal{T}_h. \tag{5.30}$$

Die Beweise basieren auf dem Bramble-Hilbert-Lemma. Das Ausführen der Beweise überlassen wir dem Leser als Übung (siehe Problemstellung 5.12).

Im Folgenden legen wir der Familie $\{\mathcal{T}_h\}_{0<h<1}$ der Triangulationen die Beschränkung auf, dass die Winkel aller Dreiecke K, die zu allen Mitgliedern der Familie $\{\mathcal{T}_h\}$ gehören, unabhängig von h nach unten beschränkt sind. Dann kann man beweisen, dass die Konstanten C_K gleichmäßig beschränkt sind, sodass die globalen Abschätzungen

$$\begin{aligned} \|I_h v - v\| &= \Big(\sum_K \|I_h v - v\|_K^2 \Big)^{1/2} \le \Big(\sum_K C_K^2 h_K^4 |v|_{2,K}^2 \Big)^{1/2} \\ &\le C h^2 \|v\|_2 \quad \forall v \in H^2 \end{aligned} \tag{5.31}$$

und analog dazu

$$|I_h v - v|_1 \le C h \|v\|_2 \quad \forall v \in H^2 \tag{5.32}$$

vorliegen. Besteht $\tilde{S}_h$ stattdessen aus stückweisen Polynomen vom Grad $r-1$, so können die zugehörigen Resultate lokal durch

$$\|I_h v - v\|_K \le C h_K^r |v|_{r,K}, \quad \|\nabla(I_h v - v)\|_K \le C h_K^{r-1} |v|_{r,K} \tag{5.33}$$

und global durch

$$\|I_h v - v\| \le C h^r \|v\|_r, \quad |I_h v - v|_1 \le C h^{r-1} \|v\|_r \quad \text{für } v \in H^r \tag{5.34}$$

ausgedrückt werden. Wir weisen darauf hin, dass diese Schranken grob sind, falls der Gitterabstand h_K innerhalb der Triangulation signifikant variiert, da wir h_K in (5.31) und (5.34) einfach durch h abgeschätzt haben. Wenn wir die Triangulation beispielsweise durch Unterteilen einiger Dreiecke K verfeinern, dann wird die Summe über alle K in (5.31) kleiner, die globale Schranke verändert sich jedoch nicht, wenn $h = \max_K h_K$ gleich bleibt.

Wir stellen fest, dass die Interpolierte $I_h v$ nur im Falle stetiger Funktionen v wohldefiniert ist, da sie die Werte von v an den Knoten benutzt. Da die Funktionen in H^2 aufgrund der Sobolev-Ungleichung, Theorem A.5, stetig sind, ist $I_h v$ somit für $v \in H^r$ mit $r \geq 2$ definiert. Eine Funktion in H^1 muss jedoch nicht stetig sein, weshalb die Punktwerte nicht wohldefiniert sind. Wenn v nicht hinreichend glatt ist, um in H^r zu sein, aber für ein s mit $2 \leq s \leq r$ die Beziehung $v \in H^s$ gilt, dann kann man anstelle von (5.34) die Abschätzungen

$$\|I_h v - v\| \leq C h^s \|v\|_s, \quad |I_h v - v|_1 \leq C h^{s-1} \|v\|_s \quad \forall v \in H^s \tag{5.35}$$

mit $2 \leq s \leq r$ benutzen. Folglich hängt der Approximationsgrad von $I_h v$ von der Regularität der Funktion v ab.

Betrachten wir nun den Fall, dass Ω konvex und der Rand Γ eine glatte Kurve statt ein Polygon ist. Sei Ω_h, wie am Ende des letzten Abschnitts beschrieben, das durch die Dreiecke von $\mathcal{T}_h$ überdeckte Gebiet. Wir wissen daher, dass die Menge $\Omega \setminus \Omega_h$ eine Breite der Ordnung $O(h^2)$ besitzt. Wenn $v = 0$ auf Γ gilt, dann ist der Interpolationsfehler in $\Omega \setminus \Omega_h$ gleich v, da dort $I_h v = 0$ gilt. Für glatte, auf Γ verschwindende Funktionen v gilt $v = O(h^2)$ in $\Omega \setminus \Omega_h$ und folglich ist deren Beitrag zum Interpolationsfehler auch von dieser Ordnung oder, genauer gesagt,

$$\|I_h v - v\|_{\Omega \setminus \Omega_h} = \|v\|_{\Omega \setminus \Omega_h} \leq C h^2 \|\nabla v\|_{\Omega \setminus \Omega_h} \leq C h^3 \|v\|_2. \tag{5.36}$$

Um die letzte Ungleichung zu beweisen, integrieren wir die Spurungleichung $\|w\|^2_{L_2(\gamma)} \leq C \|w\|^2_{H^1(\Omega)}$ über eine Kurvenschar γ, die parallel zu Γ verläuft und $\Omega \setminus \Omega_h$ überdeckt. Da die Breite von $\Omega \setminus \Omega_h$ von der Ordnung $O(h^2)$ ist, führt dies auf $\|w\|_{\Omega \setminus \Omega_h} \leq C h \|w\|_1$, was anschließend auf $w = \nabla v$ angewendet wird.

Der Gradient von v verschwindet auf Γ nicht und muss deshalb in $\Omega \setminus \Omega_h$ nicht klein sein. Deshalb ist nur

$$\|\nabla (I_h v - v)\|_{\Omega \setminus \Omega_h} = \|\nabla v\|_{\Omega \setminus \Omega_h} \leq C h \|v\|_2 \tag{5.37}$$

zu zeigen. Somit bleiben (5.31) und (5.32) für $r = 2$ einschließlich der Beiträge von $\Omega \setminus \Omega_h$ zum Interpolationsfehler gültig. Wenn jedoch $r > 2$ gilt, sind die Beiträge in (5.36) und (5.37) die bestenfalls zu erwartenden. Deshalb gilt die erste Ungleichung in (5.34) für $r = 2$ und 3, die zweite aber nur für $r = 2$.

Wir schließen mit einer Bemerkung zur orthogonalen Projektion $P_h = P_{S_h}$ des Hilbert-Raumes L_2 auf den endlichdimensionalen Teilraum S_h, der durch

$$(P_h v - v, \chi) = 0, \quad \forall \chi \in S_h, \quad v \in L_2 \tag{5.38}$$

definiert ist. Weil $P_h v$ die beste Approximation von v in S_h in der L_2-Norm ist, gilt wegen der obigen Abschätzung für ein polygonales Gebiet

$$\|P_h v - v\| \le \|I_h v - v\| \le C h^r \|v\|_r \quad \forall v \in H^r \cap H_0^1. \tag{5.39}$$

Hierbei ist $H^r \cap H_0^1$ der Raum der Funktionen, die zu H^r gehören und auf Γ verschwinden. Die Forderung $v \in H^r \cap H_0^1$ ist ziemlich stark und für Lösungen unserer elliptischen Probleme mit $r > 2$ gewöhnlich nicht erfüllt. Dies ist auf die Singularitäten an den Ecken des Gebietes zurückzuführen, die wir am Ende des Abschnitts 3.7 untersucht haben. Im Falle eines konvexen Gebietes mit glattem Rand ist die Regularität kein Problem, aber ohne Modifikation der Methode in der Nähe des Randes wissen wir nur, dass (5.39) für $r = 2, 3$ gilt.

5.4 Fehlerabschätzungen

Wir kommen zu der Aufgabe zurück, den Fehler bei der Finite-Elemente-Approximation u_h gegenüber der Lösung u unseres Dirichlet-Problems abzuschätzen. Da es sich bei der Bilinearform $a(\cdot, \cdot)$ um ein Skalarprodukt in H_0^1 handelt, verwenden wir naturgemäß die Energienorm

$$\|v\|_a = a(v, v)^{1/2} = \Big(\int_\Omega a |\nabla v|^2 \, \mathrm{d}x \Big)^{1/2}.$$

Theorem 5.3. *Seien u_h und u Lösungen von* (5.26) *und* (5.24). *Dann gilt*

$$\|u_h - u\|_a = \min_{\chi \in S_h} \|\chi - u\|_a \tag{5.40}$$

und

$$|u_h - u|_1 \le C h \|u\|_2. \tag{5.41}$$

Beweis. Wegen $S_h \subset H_0^1$ können wir in (5.24) $v = \chi \in S_h$ setzen und dies von (5.26) subtrahieren. Wir erhalten

$$a(u_h - u, \chi) = 0 \quad \forall \chi \in S_h, \tag{5.42}$$

was bedeutet, dass u_h die orthogonale Projektion von u auf S_h bezüglich des Skalarproduktes $a(\cdot, \cdot)$ ist. Die Ungleichung (5.40) folgt somit in gleicher Weise wie (5.12). Wegen unserer Annahmen hinsichtlich a gilt mit von h unabhängigem C und c

$$c|v|_1 \le \|v\|_a \le C|v|_1. \tag{5.43}$$

Somit folgt aus (5.40)

$$|u_h - u|_1 \le C \min_{\chi \in S_h} |\chi - u|_1. \tag{5.44}$$

Setzen wir $\chi = I_h u$ und benutzen die Schranke für den Interpolationsfehler in (5.32), haben wir (5.41) bewiesen. □

Für ein analoges Resultat im Falle eines nichtsymmetrischen, elliptischen Operators verweisen wir auf die Problemstellungen 5.6 und 5.7.

Die Gleichung (5.40) sagt aus, dass u_h die beste oder optimale Approximation von u in S_h bezüglich der Energienorm ist. Gleichung (5.44) zeigt, dass es sich um eine nahezu beste oder quasioptimale Approximation in der gewöhnlichen Sobolev-Norm in H_0^1 handelt. Beachten Sie, dass die Energienorm eine gewichtete Norm in H_0^1 ist; um die Eigenschaft der besten Approximation (5.40) voll auszunutzen, müsste man eine gewichtete Variante der Schranke für den Interpolationsfehler (5.32) beweisen. Dies ist möglich, wir werden dies hier aber nicht tun. Natürlich fallen diese Normen für $a = 1$ zusammen.

Damit (5.41) von Interesse ist, muss $u \in H^2$ gelten. Im Fall eines konvexen Gebietes Ω wissen wir aus Abschnitt 3.7, dass eine solche Regularität aus $f \in L_2$ folgt und dass (5.25) gilt. Aus (5.41) schlussfolgern wir deshalb

$$|u_h - u|_1 \leq Ch\|f\|,$$

wobei die Konstante gleich dem Produkt der Konstanten in (5.41) und (5.25) ist. Wenn Ω nicht konvex ist, dann wird die Lösung u im Allgemeinen an den Ecken von Γ Singularitäten besitzen, durch die (5.25) ungültig wird, woraus sich eine geringere Konvergenzordnung ergibt. Beachten Sie, dass (5.40) auch dann noch gilt.

Unser nächstes Resultat betrifft die L_2-Norm des Fehlers. An dieser Stelle benötigen wir die Regularitätsabschätzung (5.25) und nehmen deshalb an, dass Ω konvex ist.

Theorem 5.4. *Sei Ω konvex und seien u_h und u die Lösungen von* (5.26) *und* (5.24)*. Dann ist*

$$\|u_h - u\| \leq Ch^2\|u\|_2. \tag{5.45}$$

Beweis. Der Beweis erfolgt wie für das Zweipunkt-Randwertproblem in Theorem 5.2 durch Dualität unter Verwendung des Hilfsproblems

$$\mathcal{A}\phi = e \quad \text{in } \Omega \quad \text{mit } \phi = 0 \quad \text{auf } \Gamma, \tag{5.46}$$

wobei $e = u_h - u$ ist. Wie in (5.25) gilt

$$\|\phi\|_2 \leq C\|e\|, \tag{5.47}$$

und dies wird wie in Theorem 5.2 verwendet, um

$$\|e\| \leq Ch|e|_1 \tag{5.48}$$

zu zeigen. Wegen Theorem 5.3 ist der Beweis damit vollständig. □

Die beiden letzten Theoreme weisen die gleichen Fehlerschranken für u_h wie für die Interpolierte $I_h u$ in (5.31) und (5.32) auf, abgesehen davon, dass

die Konstanten verschieden sein können. Es sei darauf hingewiesen, dass wir bei den Beweisen lediglich (5.32) und nicht (5.31) benutzt haben.

Sei $R_h : H_0^1 \to S_h$ die orthogonale Projektion bezüglich des energetischen Skalarprodukts, sodass

$$a(R_h v - v, \chi) = 0 \quad \forall \chi \in S_h, \quad v \in H_0^1 \tag{5.49}$$

ist. Der Operator R_h wird als *Ritz-Projektion* (oder *elliptische Projektion*) bezeichnet. Aus (5.42) folgt, dass die Finite-Elemente-Lösung u_h genau der Ritz-Projektion der exakten Lösung u von (5.24), d. h. $u_h = R_h u$, entspricht. Unsere vorherige Fehlerabschätzung für die Finite-Elemente-Lösung kann als Funktion des Operators R_h folgendermaßen ausgedrückt werden, was für die spätere Diskussion parabolischer Finite-Elemente-Probleme zweckmäßig ist.

Theorem 5.5. *Sei Ω konvex. Dann gilt für $s = 1, 2$*

$$\|R_h v - v\| \le C h^s \|v\|_s, \quad |R_h v - v|_1 \le C h^{s-1} \|v\|_s \quad \forall v \in H^s \cap H_0^1.$$

Beweis. Der Fall $s = 2$ ist in den Theoremen 5.3 und 5.4 enthalten. Im Fall $s = 1$ bemerken wir zunächst Folgendes: Weil R_h die orthogonale Projektion bezüglich $a(\cdot, \cdot)$ ist, gilt $\|R_h v\|_a \le \|v\|_a$. Folglich ist $|R_h v|_1 \le C|v|_1$ und $|R_h v - v|_1 \le C|v|_1$. Unter Verwendung von (5.48) erhalten wir schließlich

$$\|R_h v - v\| \le Ch \|R_h v - v\|_1 \le Ch \|v\|_1,$$

was den Beweis abschließt. □

Formal lässt sich die obige Fehleranalyse unmittelbar auf finite Elemente höherer Ordnung $r > 2$ übertragen. Im Argument in Theorem 5.4 verwenden wir einfach die zweite Abschätzung des Interpolationsfehlers in (5.34) anstelle von (5.32), zusammen mit dem Fall $s = 2$ von (5.35). Wir finden dann für $2 \le s \le r$

$$\|R_h v - v\| \le C h^s \|v\|_s, \ |R_h v - v|_1 \le C h^{s-1} \|v\|_s \ \forall v \in H^s \cap H_0^1. \tag{5.50}$$

Diese Abschätzungen zeigen also eine reduzierte Konvergenzrate $O(h^s)$, wenn $v \in H^s$ mit $s < r$ gilt. Wir wir bereits am Ende von Abschnitt 5.3 betont haben, ist die Regularitätsannahme $v \in H^r$ mit $r > 2$ für Lösungen unserer elliptischen Probleme in einem polygonalen Gebiet etwas unrealistisch. Für ein Gebiet Ω mit einem glatten Rand Γ ist die Regularität kein Problem, es sind dann allerdings spezielle Betrachtungen für die Handhabung des Randgebietes $\Omega \setminus \Omega_h$ notwendig, um eine höhere Genauigkeit zu erreichen.

Da die Variationsformulierung unseres diskreten Problems auf Skalarprodukten aus dem L_2 beruht, werden die geläufigsten Fehlerabschätzungen auch in solchen L_2-basierten Normen ausgedrückt und messen deshalb bestimmte Mittelwerte des Fehlers. Es ist natürlich auch von Interesse, Fehlerschranken in der Maximumnorm abzuleiten, die gleichmäßige Fehlerschranken über Ω

angeben. Wir stellen zunächst fest, dass der Fehler in der oben eingeführten Interpolierten die Ungleichung

$$\|I_h v - v\|_{\mathcal{C}(K)} \leq C h_K^2 \|v\|_{\mathcal{C}^2(K)} \quad \forall K \in \mathcal{T}_h$$

erfüllt. Auch im Falle eines glatten Randes Γ gilt folglich

$$\|I_h v - v\|_{\mathcal{C}} \leq C h^2 \|v\|_{\mathcal{C}^2}, \tag{5.51}$$

weil dann $\|v\|_{\mathcal{C}(\Omega \setminus \Omega_h)} \leq C h^2 \|v\|_{\mathcal{C}^1}$ ist. Unter der zusätzlichen Annahme, dass die Familie der Triangulationen $\{\mathcal{T}_h\}$ *quasiuniform* ist, d. h. dass

$$h_K \geq ch \tag{5.52}$$

für ein positives, von h unabhängiges c gilt, ist es möglich, wenn auch nicht einfach, für unser elliptisches Problem

$$\|u_h - u\|_{\mathcal{C}} \leq C h^2 \log(1/h) \|u\|_{\mathcal{C}^2} \quad \text{für kleines } h \tag{5.53}$$

zu zeigen (siehe Problemstellung 5.4). Im Vergleich zur Abschätzung in der L_2-Norm aus Theorem 5.4 enthält diese Abschätzung einen zusätzlichen Faktor $\log(1/h)$, der in der Fehlerabschätzung für die Interpolation (5.51) nicht vorkommt. Man kann zeigen, dass dieser Faktor nicht beseitigt werden kann.

5.5 Eine a posteriori Fehlerabschätzung

Die Fehlerschranken im letzten Abschnitt enthalten Normen der exakten, unbekannten Lösung. Unter Verwendung der Regularitätsabschätzung (5.25) können diese Fehlerschranken auch als Funktion der Daten f aus (5.23) ausgedrückt werden. Wenn die darin eingehenden Konstanten bekannt sind, erhält man strengere Schranken für den Fehler. Wie wir jedoch in Abschnitt 5.3 gesehen haben, können diese Schranken pessimistisch sein, insbesondere wenn die Triangulationen sehr ungleichmäßig sind. In einem solchen Fall kann beispielsweise die Ungleichung (5.32) sehr grob sein. Die Abschätzungen, die $h = \max_K h_K$ einbeziehen, sollten deshalb als asymptotische Abschätzungen interpretiert werden, die die Konvergenzrate des Fehlers für $h \to 0$ anzeigen. Theorem 5.4 zeigt also beispielsweise, dass $\|u_h - u\| = O(h^2)$ für $h \to 0$ gilt, wenn $u \in H^2$ ist.

Da diese Schranken nicht von der berechneten Lösung abhängen, werden sie häufig als *a priori Schranken* bezeichnet; sie können aufgestellt werden, bevor die Rechnung ausgeführt wird. Im nächsten Theorem werden wir ein Beispiel für eine *a posteriori Fehlerabschätzung* angeben, die als Funktion der berechneten Lösung und der Daten ausgedrückt wird.

Theorem 5.6. *Angenommen, Ω ist ein konvexes, polygonales Gebiet in der Ebene. Seien u_h und u Lösungen von* (5.26) *beziehungsweise* (5.24)*. Dann gilt*

$$\|u_h - u\| \le C\Big(\sum_{K\in\mathcal{T}_h} R_K^2\Big)^{1/2}$$

mit

$$R_K = h_K^2\|\mathcal{A}u_h - f\|_K + h_K^{3/2}\|a[n\cdot\nabla u_h]\|_{\partial K\setminus\Gamma}.$$

Dabei ist $[n\cdot\nabla u_h]$ *der Sprung durch* ∂K *in der Normalenableitung* $n\cdot\nabla u_h$.

Beweis. Wir verwenden das Dualitätsargument aus dem Beweis von Theorem 5.4. Sei $e = u_h - u$ und ϕ die Lösung von (5.46). Dann gilt mit $(v,w)_K = \int_K v\,w\,\mathrm{d}x$, $\|v\|_K = \|v\|_{L_2(K)}$ und $|v|_{2,K} = |v|_{H^2(K)}$

$$\begin{aligned}\|e\|^2 &= a(e,\phi) = a(u_h-u,\phi) = a(u_h,\phi) - (f,\phi)\\ &= \sum_K \Big((a\nabla u_h,\nabla\phi)_K - (f,\phi)_K\Big)\\ &= \sum_K \Big((\mathcal{A}u_h - f,\phi)_K + (an\cdot\nabla u_h,\phi)_{\partial K}\Big)\\ &= \sum_K \Big((\mathcal{A}u_h - f,\phi)_K - \tfrac12(a[n\cdot\nabla u_h],\phi)_{\partial K\setminus\Gamma}\Big).\end{aligned}$$

Der Faktor 1/2 im letzten Term erscheint, weil der Term in der Summe zweimal auftritt. Wegen $a(e,\chi) = 0$ für $\chi\in S_h$ können wir ϕ durch $\phi-\chi$ ersetzen und erhalten

$$\begin{aligned}\|e\|^2 &= |a(e,\phi-\chi)|\\ &\le \sum_K \Big(\|\mathcal{A}u_h - f\|_K\,\|\phi-\chi\|_K + \tfrac12\|a[n\cdot\nabla u_h]\|_{\partial K\setminus\Gamma}\,\|\phi-\chi\|_{\partial K\setminus\Gamma}\Big).\end{aligned}$$

Wir wählen nun $\chi = I_h\phi$ und erinnern an (5.29), (5.30) sowie an die skalierte Spurungleichung, die man durch Transformation der Spurungleichung (A.26) von einem Referenzdreieck $\hat K$ mit Einheitsgröße auf ein kleines Dreieck K erhält (siehe Problemstellung A.15)

$$\|w\|_{\partial K} \le C\Big(h_K^{-1/2}\|w\|_K + h_K^{1/2}\|\nabla w\|_K\Big). \tag{5.54}$$

Daher erhalten wir

$$\|\phi - I_h\phi\|_{\partial K} \le Ch_K^{3/2}|\phi|_{2,K}, \tag{5.55}$$

und im Hinblick auf die Regularitätsabschätzung (5.47) können wir

$$\begin{aligned}\|e\|^2 = a(e,\phi) &\le C\sum_K R_K|\phi|_{2,K} \le C\Big(\sum_K R_K^2\Big)^{1/2}\Big(\sum_K |\phi|_{2,K}^2\Big)^{1/2}\\ &\le C\Big(\sum_K R_K^2\Big)^{1/2}\|\phi\|_2 \le C\Big(\sum_K R_K^2\Big)^{1/2}\|e\|\end{aligned}$$

schlussfolgern, was den Beweis abschließt. □

Für $\mathcal{A} = -\Delta$, d. h. im Falle $a = 1$, ist $\mathcal{A}u_h = 0$ in K. Weil $n \cdot \nabla u_h$ entlang ∂K konstant ist, gilt $R_K = h_K^2\big(\|f\|_K + \big|[n \cdot \nabla u_h]|_{\partial K \setminus \Gamma}\big|\big)$, sodass die berechnete Lösung nur in den zweiten Term eingeht. Aus der *a posteriori* Fehlerabschätzung selbst folgt nicht, dass der Fehler mit h gegen null geht. Allerdings geht dies aus der vorhin gezeigten *a priori* Fehlerabschätzung der Ordnung $O(h^2)$ hervor.

Die *a posteriori* Fehlerabschätzung legt auch eine Methode zur adaptiven Fehlerkontrolle nahe, nämlich das Reduzieren des Gitterabstandes durch Unterteilen derjenigen Dreiecke K, für die R_K im Vergleich zu einer Toleranz groß ist. Dies werden wir hier jedoch nicht näher behandeln.

5.6 Numerische Integration

Eine wesentliche Eigenschaft der Methode der finiten Elemente besteht darin, dass die Gleichungen (5.27) durch ein Computerprogramm automatisch generiert werden können. Diese Prozedur, die als *Assemblieren* bezeichnet wird, basiert auf der elementweisen Berechnung der Steifigkeitsmatrix und des Lastvektors

$$a(\Phi_j, \Phi_i) = \sum_{K \in \mathcal{T}_h} \int_K a \nabla \Phi_j \cdot \nabla \Phi_i \, \mathrm{d}x, \quad (f, \Phi_i) = \sum_{K \in \mathcal{T}_h} \int_K f \Phi_i \, \mathrm{d}x. \tag{5.56}$$

In der Praxis werden die Integrale in dieser Summe selten exakt berechnet, auch wenn analytische Ausdrücke für a und f verfügbar sind. Stattdessen werden sie mithilfe numerischer Integration durch eine Quadraturformel der Form

$$\int_K \phi \, \mathrm{d}x \approx q_K(\phi) := \sum_{l=1}^{L} \omega_{l,K} \phi(b_{l,K}) \tag{5.57}$$

approximiert. Die Zahlen $\omega_{l,K}$ werden als Gewichte und die Punkte $b_{l,K}$ als Knoten der Quadraturformel bezeichnet.

Wenn die Gleichungen durch numerische Integration aufgebaut werden, lösen wir anstelle von (5.26) ein modifiziertes Finite-Elemente-Problem, das darin besteht, ein $u_h \in S_h$ mit

$$a_h(u_h, \chi) = (f, \chi)_h \quad \forall \chi \in S_h \tag{5.58}$$

zu bestimmen, wobei

$$a_h(v, w) = \sum_{K \in \mathcal{T}_h} q_K(a \nabla v \cdot \nabla w), \quad (f, w)_h = \sum_{K \in \mathcal{T}_h} q_K(fw) \tag{5.59}$$

ist.

Die Quadraturformel q_K in (5.57) sollte so gewählt werden, dass der Fehler in u_h von derselben Ordnung wie bei der ursprünglichen Finite-Elemente-Lösung ist. Ein Beispiel für eine solche Quadraturformel ist die *baryzentrische Quadraturregel*

$$q_K(\phi) = |K|\phi(P_K), \quad \text{wobei } |K| = \text{area}(K) \text{ und } P_K = \tfrac{1}{3}\sum_{l=1}^{3} P_{l,K} \text{ ist,} \tag{5.60}$$

mit den Eckpunkten $P_{l,K}$ und dem Schwerpunkt (Baryzentrum) P_K des Dreiecks K. Diese Quadraturregel ist für lineare Funktionen exakt, es gilt also

$$\int_K \phi \,\mathrm{d}x = |K|\phi(P_K) \quad \forall \phi \in \Pi_1. \tag{5.61}$$

Folglich ist die Regel in zweiter Ordnung exakt, sodass

$$\left| q_K(\phi) - \int_K \phi \,\mathrm{d}x \right| \le C h_K^2 |\phi|_{W_1^2(K)} \tag{5.62}$$

ist (siehe Problemstellung 5.13), wobei für $D_{ij} = \partial^2/\partial x_i \partial x_j$ die Gleichung

$$|v|_{W_1^2(M)} = \sum_{i,j=1}^{2} \|D_{ij} v\|_{L_1(M)}, \quad \|v\|_{L_1(M)} = \int_M |v| \,\mathrm{d}x$$

gilt. Somit ist der globale Quadraturfehler durch

$$\sum_{K \in \mathcal{T}_h} \left| q_K(\phi) - \int_K \phi \,\mathrm{d}x \right| \le C h^2 \sum_{K \in \mathcal{T}_h} |\phi|_{W_1^2(K)} \tag{5.63}$$

beschränkt.

Aus (5.62) schlussfolgern wir, dass $a_h(u_h, \chi)$ und $(f, \chi)_h$ exakt sind. Dies gilt beispielsweise für konstantes a und f.

Ein anderes Beispiel für eine Quadraturformel, die für lineare Funktionen exakt ist, liefert die *Knotenquadraturregel* (siehe Problemstellung 5.15)

$$q_K(\phi) = \tfrac{1}{3}|K| \sum_{l=1}^{3} \phi(P_{l,K}). \tag{5.64}$$

Im folgenden Lemma fassen wir die Eigenschaften von $a_h(\cdot,\cdot)$ und $(\cdot,\cdot)_h$ zusammen, die wir zum Beweis einer Fehlerabschätzung für das modifizierte Problem (5.58) brauchen.

Lemma 5.1. *Wenn $a_h(\cdot,\cdot)$ und $(\cdot,\cdot)_h$ in* (5.59) *durch die Quadraturformel* (5.60) *oder* (5.64) *berechnet werden, gilt*

$$a_0 |\chi|_1^2 \le a_h(\chi, \chi) \le C |\chi|_1^2 \quad \forall \chi \in S_h \tag{5.65}$$

und

$$|a_h(\psi,\chi) - a(\psi,\chi)| \le Ch^2 \|a\|_{\mathcal{C}^2} \, |\psi|_1 \, |\chi|_1 \qquad \forall \psi, \chi \in S_h, \tag{5.66}$$

$$|(f,\chi)_h - (f,\chi)| \le Ch^2 \|f\|_2 \, |\chi|_1 \qquad \forall \chi \in S_h. \tag{5.67}$$

Beweis. Wir führen den Beweis für die Quadraturregel (5.60); der Beweis für (5.64) verläuft analog.

Da $\nabla\chi$ auf K konstant und $a_0 \le a(x) \le C$ ist, gilt

$$a_h(\chi,\chi) = \sum_K a(P_K)|\nabla\chi(P_K)|^2|K| \ge a_0 \sum_K \int_K |\nabla\chi|^2 \, \mathrm{d}x = a_0|\chi|_1^2.$$

Die Abschätzung von oben wird in gleicher Weise abgeleitet, was (5.65) beweist. Unter Verwendung von (5.63) mit $\phi = a\nabla\psi \cdot \nabla\chi$ erhalten wir

$$|a_h(\psi,\chi) - a(\psi,\chi)| \le Ch^2 \sum_K |a\nabla\psi \cdot \nabla\chi|_{W_1^2(K)}.$$

Um die rechte Seite abzuschätzen, stellen wir zunächst fest, dass für ψ, $\chi \in S_h$

$$\|D_{ij}(a\nabla\psi \cdot \nabla\chi)\|_{L_1(K)} = \|(D_{ij}a)\nabla\psi \cdot \nabla\chi\|_{L_1(K)} \le \|a\|_{\mathcal{C}^2} \, \|\nabla\psi\|_K \, \|\nabla\chi\|_K$$

gilt. Unter Verwendung der Cauchy-Schwarz-Ungleichung für Summen schließen wir

$$\sum_K |a\nabla\psi \cdot \nabla\chi|_{W_1^2(K)} \le C\|a\|_{\mathcal{C}^2} \, |\psi|_1 \, |\chi|_1,$$

was (5.66) beweist. Analog dazu gilt wegen $D_{ij}\chi = 0$ auf K

$$\|D_{ij}(f\chi)\|_{L_1(K)} = \|D_{ij}f \, \chi + D_i f D_j \chi + D_j f D_i \chi\|_{L_1(K)} \le C\|f\|_{2,K} \, \|\chi\|_{1,K},$$

sodass

$$\sum_K |f\chi|_{W_1^2(K)} \le C\|f\|_2 \, \|\chi\|_1 \le C\|f\|_2 \, |\chi|_1$$

gilt, was (5.67) beweist. □

Die Ungleichung (5.65) zeigt, dass die symmetrische Bilinearform $a_h(\cdot,\cdot)$ ein Skalarprodukt auf S_h ist und dass die zugehörige Norm äquivalent zu $|\cdot|_1$ und gleichmäßig beschränkt bezüglich h ist. Aus (5.67) folgern wir, dass die Linearform $L_h(\chi) = (f,\chi)_h$ auf S_h bezüglich $|\cdot|_1$ beschränkt ist, wiederum gleichmäßig bezüglich h, weil

$$\begin{aligned} |(f,\chi)_h| &\le |(f,\chi)| + |(f,\chi)_h - (f,\chi)| \\ &\le \|f\| \, \|\chi\| + Ch^2\|f\|_2 \, |\chi|_1 \le C\|f\|_2 \, |\chi|_1 \end{aligned}$$

gilt. Wegen des Rieszschen Darstellungssatzes können wir deshalb schlussfolgern, dass (5.58) eine eindeutige Lösung besitzt und die Stabilitätsabschätzung

$$|u_h|_1 \le C\|f\|_2 \tag{5.68}$$

erfüllt (siehe Problemstellung 5.14). Diese Stabilität des modifizierten Finite-Elemente-Problems wird zusammen mit den Fehlerschranken (5.66), (5.67) beim Beweis der folgenden Fehlerabschätzung benutzt.

Theorem 5.7. *Angenommen, $a_h(\cdot,\cdot)$ und $(\cdot,\cdot)_h$ in* (5.59) *werden durch die Quadraturformel* (5.60) *oder* (5.64) *berechnet. Seien u_h und u die Lösungen von* (5.58) *beziehungsweise* (5.24)*. Dann ist*

$$|u_h - u|_1 \le Ch\|u\|_2 + Ch^2\Big(\|a\|_{\mathcal{C}^2}\,\|u\|_2 + \|f\|_2\Big). \tag{5.69}$$

Beweis. Wir schreiben $u_h - u = (u_h - I_h u) + (I_h u - u) = \theta + \rho$. Unter Verwendung von (5.24) und (5.58) erhalten wir für jedes $\chi \in S_h$

$$\begin{aligned} a_h(\theta,\chi) &= a_h(u_h,\chi) - a_h(I_h u,\chi) + \Big(a(u,\chi) - (f,\chi)\Big) \\ &\quad - \Big(a_h(u_h,\chi) - (f,\chi)_h\Big) + a(I_h u,\chi) - a(I_h u,\chi) \\ &= -a(\rho,\chi) - \Big(a_h(I_h u,\chi) - a(I_h u,\chi)\Big) + \Big((f,\chi)_h - (f,\chi)\Big). \end{aligned}$$

Weil $\theta \in S_h$ ist, können wir $\chi = \theta$ wählen. Wegen (5.65), der Beschränktheit der Bilinearform $a(\cdot,\cdot)$ und den Fehlerabschätzungen (5.66) und (5.67) folgt

$$a_0|\theta|_1^2 \le a_h(\theta,\theta) \le \Big(C|\rho|_1 + Ch^2\|a\|_{\mathcal{C}^2}\,|I_h u|_1 + Ch^2\|f\|_2\Big)|\theta|_1.$$

Somit ist

$$|u_h - u|_1 \le |\theta|_1 + |\rho|_1 \le C|\rho|_1 + Ch^2\Big(\|a\|_{\mathcal{C}^2}\,|I_h u|_1 + \|f\|_2\Big).$$

Unter Verwendung der Fehlerabschätzung für die Interpolation (5.32) gilt

$$|\rho|_1 \le Ch\|u\|_2, \quad |I_h u|_1 \le |u|_1 + |\rho|_1 \le |u|_1 + Ch\|u\|_2 \le C\|u\|_2.$$

Insgesamt beweisen diese Abschätzungen die Ungleichung (5.69). □

Der erste Term auf der rechten Seite von (5.69) ist (im Wesentlichen) derselbe wie in (5.41), wobei die verbleibenden Terme den Effekt der numerischen Integration abschätzen. Beachten Sie, dass dieses Resultat eine höhere Regularität erfordert als (5.41). Beispielsweise benötigen wir $f \in H^2$, woraus (zumindest formal gesehen) $u \in H^4$ folgt. Dies ist mit dem Resultat für das finite Differenzenverfahren in Theorem 4.2 konsistent, wobei u vier Ableitungen besitzen soll (siehe auch Anmerkung 5.1). Wir können die Methode der finiten Elemente mit numerischer Integration als ein finites Differenzenverfahren auf einem ungleichmäßigem Gitter betrachten.

Die ursprüngliche Methode der finiten Elemente (5.26) ist in dem Sinne *konform*, dass $S_h \subset H_0^1$ und die Formen $a(\cdot,\cdot)$ und $L(\cdot) = (f,\cdot)$ dieselben

wie beim kontinuierlichen Problem (5.24) sind. Bei der modifizierten Methode der finiten Elemente (5.58) gilt immer noch $S_h \subset H_0^1$, die Formen sind aber verschieden. Sie wird deshalb als *nichtkonform* bezeichnet. Andere nichtkonforme Methoden der finiten Elemente verletzen die Annahme $S_h \subset H_0^1$, beispielsweise bei der Verwendung von unstetigen, stückweisen Polynomen. Sie können manchmal auf ähnliche Weise untersucht werden. Das im Beweis von Theorem 5.7 verwendete Argument basiert auf dem in der Literatur als erstes Lemma von Strang bekannten Lemma über nichtkonforme Methoden der finiten Elemente.

5.7 Eine Methode der gemischten finiten Elemente

In einigen Situationen ist der Fluss $-a\nabla u$ der Lösung u von primärem Interesse. Bei der gewöhnlichen Methode der finiten Elemente werden die Ableitungen und somit auch der Fluss in niedriger Ordnung $O(h)$ approximiert, im Gegensatz zur $O(h^2)$-Approximation der Lösung. Wir werden nun kurz eine Methode der finiten Elemente für unser Modellproblem (5.23) umreißen, die auf einer sogenannten gemischten Formulierung dieses Problems beruht und nicht den oben genannten Nachteil besitzt. In diesem Falle wird der Fluss der Lösung u als eine separate abhängige Variable eingeführt, deren Approximation in einem anderen Finite-Elemente-Raum gesucht wird als dem der Lösung selbst. Dies kann man so tun, dass der Fluss die gleiche Genauigkeitsordnung wie u besitzt. Der Einfachheit halber nehmen wir an, dass in (5.23) $a = 1$ ist. Mit der separaten zweidimensionalen Variablen $\sigma = \nabla u$ kann diese Gleichung dann in der Form

$$\begin{aligned} -\nabla \cdot \sigma &= f && \text{in } \Omega, \\ \sigma &= \nabla u && \text{in } \Omega, \\ u &= 0 && \text{auf } \partial\Omega \end{aligned} \tag{5.70}$$

geschrieben werden. Wir stellen fest, dass die Lösung $(u, \sigma) \in L_2 \times H$ mit $H = \{\omega = (\omega_1, \omega_2) \in L_2 \times L_2 : \nabla \cdot \omega \in L_2\}$ auch das Variationsproblem

$$\begin{aligned} (\nabla \cdot \sigma, \varphi) + (f, \varphi) &= 0 && \forall \varphi \in L_2, \\ (\sigma, \omega) + (u, \nabla \cdot \omega) &= 0 && \forall \omega \in H \end{aligned} \tag{5.71}$$

löst, wobei $(\cdot, \cdot)$ die entsprechenden L_2-Skalarprodukte bezeichnet, und eine glatte Lösung von (5.71) die Gleichung (5.70) erfüllt. Setzt man $L(v, \mu) = \frac{1}{2}\|\mu\|^2 + (\nabla \cdot \mu + f, v)$, so kann man zeigen, dass die Lösung (u, σ) von (5.70) als Sattelpunkt charakterisiert werden kann, der die Ungleichung

$$L(v, \sigma) \le L(u, \sigma) \le L(u, \mu) \quad \forall v \in L_2,\ \mu \in H \tag{5.72}$$

erfüllt (siehe Problemstellung 5.16). Der Schlüssel zum Beweis der Existenz einer Lösung steckt in der Ungleichung

$$\inf_{v\in L_2}\sup_{\mu\in H}\frac{(v,\nabla\cdot\mu)}{\|v\|\,\|\mu\|_H}\geq c>0, \tag{5.73}$$

wobei $\|\mu\|_H^2=\|\mu\|^2+\|\nabla\cdot\mu\|^2$ ist. Seien S_h und H_h spezielle endlichdimensionale Teilräume von L_2 und H. Damit werden wir die diskrete Entsprechung von (5.71) betrachten, die darin besteht, ein $(u_h,\sigma_h)\in S_h\times H_h$ zu bestimmen, sodass

$$\begin{aligned}(\nabla\cdot\sigma_h,\chi)+(f,\chi)&=0 \qquad \forall\chi\in S_h,\\ (\sigma_h,\psi)+(u_h,\nabla\cdot\psi)&=0 \qquad \forall\psi\in H_h\end{aligned} \tag{5.74}$$

gilt. Wie im kontinuierlichen Fall ist dieses Problem äquivalent zum diskreten Analogon des Sattelpunktproblems (5.72). Damit dieses diskrete Problem eine Lösung mit den gewünschten Eigenschaften besitzt, müssen die Räume $S_h\times H_h$ so gewählt werden, dass das Analogon zu (5.73) gilt, was in diesem Kontext als Babuška-Brezzi-inf-sup-Bedingung bezeichnet wird. Genauer gesagt, muss

$$\inf_{v\in S_h}\sup_{\mu\in H_h}\frac{(v,\nabla\cdot\mu)}{\|v\|\,\|\mu\|_H}\geq c>0 \tag{5.75}$$

gleichmäßig bezüglich h gelten.

Ein von Raviart und Thomas eingeführtes Beispiel für ein Paar von Räumen, das die inf-sup-Bedingung erfüllt, ist das folgende: Sei $\mathcal{T}_h$ eine quasiuniforme Familie von Triangulationen des Gebietes Ω, das hier als polygonal angenommen wird. Wir setzen

$$S_h=\big\{\chi\in L_2:\chi|_K \text{ linear},\ \forall K\in\mathcal{T}_h\big\},$$

wobei an den Rändern innerer Elemente keine Stetigkeit gefordert wird. Wir definieren außerdem

$$H_h=\big\{\psi=(\psi_1,\psi_2)\in H:\psi|_K\in H(K),\ \forall K\in\mathcal{T}_h\big\},$$

wobei die $H(K)$ affine Abbildungen von Kurven zweiter Ordnung auf ein Referenzdreieck $\hat{K}$ der Form $(l_1(\xi)+\alpha\xi_1(\xi_1+\xi_2),\ l_2(\xi)+\beta\xi_2(\xi_1+\xi_2))$ darstellen. Dabei sind $l_1(\xi), l_2(\xi)$ linear, und es gilt $\alpha,\beta\in\mathbf{R}$. Da jede der Funktionen $l_j(\xi)$ drei Parameter besitzt, gilt $\dim H(K)=8$. Der Raum H_h besteht folglich aus stückweise quadratischen Funktionen auf der Triangulation $\mathcal{T}_h$, die von der durch die Definition von $H(K)$ spezifizierten Form sind. Zum Bestimmen der Freiheitsgrade für H_h können die Werte von $\psi\cdot n$ an zwei Punkten auf jeder Seite von K (sechs Bedingungen) und zusätzlich die Mittelwerte von ψ_1 und ψ_2 über K (zwei Bedingungen) verwendet werden. Es sei darauf hingewiesen, dass die Bedingung $\psi\in H$ in der Definition von H_h die Beziehung $\nabla\cdot\psi\in L_2$ fordert, was äquivalent zur Stetigkeit von $\chi\cdot n$ an den Rändern innerer Elemente ist. Für die Lösungen von (5.74) und (5.70) kann man folgende Abschätzungen zeigen:

$$\|u_h-u\|\leq Ch^2\|u\|_2 \quad\text{und}\quad \|\sigma_h-\sigma\|\leq Ch^s\|u\|_{s+1},\quad s=1,2.$$

Folglich wird der Fluss σ in der gleichen Ordnung $O(h^2)$ wie u approximiert.

5.8 Problemstellungen

Problem 5.1. Beweisen Sie (5.7) und (5.8).

Hinweis zu (5.8): Es gilt $(I_h v)'(x) - v'(x) = h_j^{-1} \int_{K_j} \big(v'(y) - v'(x)\big)\,\mathrm{d}y$ für $x \in K_j$.

Hinweis zu (5.7): Sei $Q_1 v$ das Polynom vom Grad 1, das wir durch die Taylorsche Formel für v an der Stelle x_{j-1} erhalten. Beachten Sie, dass $I_h(Q_1 v) = Q_1 v$ und $\|I_h v\|_{\mathcal{C}(K_j)} \le \|v\|_{\mathcal{C}(K_j)}$ ist, sodass $\|I_h v - v\|_{\mathcal{C}(K_j)} = \|I_h(v - Q_1 v) + (Q_1 v - v)\|_{\mathcal{C}(K_j)} \le 2\|v - Q_1 v\|_{\mathcal{C}(K_j)}$ gilt. Schätzen Sie den Rest ab: $\|v - Q_1 v\|_{\mathcal{C}(K_j)} \le \max_{x \in K_j} \int_{K_j} |x - y|\,|v''(y)|\,\mathrm{d}y$. Schlussfolgern Sie $\|I_h v - v\|_{\mathcal{C}(K_j)} \le 2h_j \int_{K_j} |v''(y)|\,\mathrm{d}y$, woraus (5.7) folgt. Dieser Beweis kann auf Funktionen in zwei Variablen verallgemeinert werden (siehe (5.29)). Der Hauptunterschied besteht darin, dass es schwieriger ist, den Rest in der Taylorschen Formel abzuschätzen.

Problem 5.2. Bestimmen Sie die Elemente der Matrix A in (5.5) im Falle $h_j = h = \text{konstant}$.

Problem 5.3. Verwenden Sie die Basis $\{\Phi_i\}_{i=1}^{M_h}$, um zu zeigen, dass (5.38) in Matrixform als $BV = b$ geschrieben werden kann, wobei die Matrix B (die sogenannte Massenmatrix) für großes M_h symmetrisch, positiv definit und dünn besetzt ist.

Problem 5.4. Betrachten Sie die Situation in Abschnitt 5.1 mit einem stückweise linearen Finite-Elemente-Raum S_h.

(a) Verwenden Sie die Greensche Funktion in Theorem 2.3, um $u_h = I_h u$ für $a = 1$, $c = 0$ in (5.1) zu beweisen, vgl. Bemerkung 5.2. Hinweis: Verwenden Sie die Resultate aus den Problemstellungen 2.4 und 2.2 (a) sowie die Tatsache, dass $G(x_j, \cdot) \in S_h$ gilt, wenn x_j ein Knoten ist.

(b) Beweisen Sie für den Fall variabler Koeffizienten

$$|u_h(x_j) - u(x_j)| \le Ch^2 \|u\|_2.$$

Hinweis: Zeigen Sie $e(x_j) = a(e, G(x_j, \cdot) - I_h G(x_j, \cdot))$ und benutzen Sie eine Fehlerabschätzung für die Interpolation auf den Intervallen $(0, x_j)$, $(x_j, 1)$, wobei $G(x_j, \cdot)$ glatt ist.

(c) Schlussfolgern Sie $\|u_h - u\|_{\mathcal{C}} \le Ch^2 \|u\|_{\mathcal{C}^2}$, was (5.53) in diesem einfachen Spezialfall entspricht. Hinweis: $\|u_h - I_h u\|_{\mathcal{C}} = \max_j |u_h(x_j) - u(x_j)|$.

Dies ist die Grundidee, durch die man Maximumnorm-Abschätzungen für elliptische Probleme in mehreren Variablen erreicht. Die stärkere Singularität der Greenschen Funktion (siehe Abschnitt 3.4) macht diese Analyse jedoch viel schwieriger.

Problem 5.5. Beweisen Sie unter den Annahmen von Theorem 5.3, dass

$$|u_h - u|_1 \le C\Big(\sum_K h_K^2 |u|_{2,K}^2\Big)^{1/2}$$

gilt.

Problem 5.6. (Galerkin-Methode.) Angenommen $a(\cdot,\cdot)$ und $L(\cdot)$ erfüllen die Annahmen des Lax-Milgram-Lemmas, d. h.

$$\begin{aligned} |a(v,w)| &\le C_1 \|v\|_V \, \|w\|_V \qquad \forall v, w \in V, \\ a(v,v) &\ge C_2 \|v\|_V^2 \qquad \forall v \in V, \\ |L(v)| &\le C_3 \|v\|_V \qquad \forall v \in V. \end{aligned}$$

Sei $u \in V$ die Lösung von

$$a(u,v) = L(v) \quad \forall v \in V.$$

Sei $\tilde{V} \subset V$ ein endlichdimensionaler Teilraum und $\tilde{u} \in \tilde{V}$ durch die Galerkin-Methode bestimmt:

$$a(\tilde{u},v) = L(v) \quad \forall v \in \tilde{V}.$$

Beweisen Sie die Ungleichung

$$\|\tilde{u} - u\|_V \le \frac{C_1}{C_2} \min_{\chi \in \tilde{V}} \|\chi - u\|_V.$$

(Beachten Sie, dass $a(\cdot,\cdot)$ nicht-symmetrisch sein kann.) Beweisen Sie, dass im Falle eines symmetrischen $a(\cdot,\cdot)$ und $\|v\|_a = a(v,v)^{1/2}$

$$\|\tilde{u} - u\|_a = \min_{\chi \in \tilde{V}} \|\chi - u\|_a \quad \text{und} \quad \|\tilde{u} - u\|_V \le \sqrt{\frac{C_1}{C_2}} \min_{\chi \in \tilde{V}} \|\chi - u\|_V$$

gilt.

Problem 5.7. Betrachten Sie das Problem

$$-\nabla \cdot \big(a \nabla u\big) + b \cdot \nabla u + cu = f \quad \text{in } \Omega \quad \text{mit } u = 0 \quad \text{auf } \Gamma$$

aus Abschnitt 3.5. Beachten Sie, dass die Bilinearform aufgrund des Konvektionsterms $b \cdot \nabla u$ nicht symmetrisch ist.

(a) Formulieren Sie eine Methode der finiten Elemente für dieses Problem und beweisen Sie eine Fehlerschranke in der H^1-Norm (siehe Problemstellung 5.6).

(b) Beweisen Sie eine Fehlerschranke in der L_2-Norm. Modifizieren Sie den Beweis des Theorems 5.4, indem Sie anstelle von (5.46) auf das Hilfsproblem

$$\mathcal{A}^* \phi := -\nabla \cdot (a \nabla \phi) - b \cdot \nabla \phi + (c - \nabla \cdot b)\phi = e \quad \text{in } \Omega, \quad \phi = 0 \quad \text{auf } \Gamma$$

zurückgreifen. Der Operator $\mathcal{A}^*$ ist der zu $\mathcal{A}$ *adjungierte*, der durch $(\mathcal{A}v, w) = a(v,w) = (v, \mathcal{A}^* w)$ für alle $v, w \in H^2 \cap H_0^1$ definiert ist.

Problem 5.8. Formulieren Sie ein Finite-Elemente-Problem, das dem inhomogenen Dirichlet-Problem (3.27) entspricht. Beweisen Sie Fehlerabschätzungen. Hinweis: Mit der Notation aus Abschnitt 5.3 gilt $u_h(x) = \sum_{j=1}^{M_h} U_j \Phi_j(x) + \sum_{j=M_h+1}^{N_h} g(P_j)\Phi_j(x)$.

Problem 5.9. Formulieren Sie ein Finite-Elemente-Problem, das dem Neumann-Problem (3.30) entspricht. Beweisen Sie Fehlerabschätzungen.

Problem 5.10. Formulieren Sie ein Finite-Elemente-Problem, das dem inhomogenen Neumann-Problem (3.34) entspricht. Beweisen Sie Fehlerabschätzungen.

Problem 5.11. Formulieren Sie ein Finite-Elemente-Problem, das dem Robin-Problem in Problemstellung 3.6 entspricht. Beweisen Sie Fehlerabschätzungen.

Problem 5.12. Das folgende wichtige Resultat ist ein Spezialfall des Bramble-Hilbert-Lemmas. *Sei $F(v)$ ein nichtnegatives Funktional auf dem $H^r = H^r(\Omega)$, wobei Ω ein beschränktes Gebiet in $\mathbf{R}^d$ ist. Falls die Ungleichungen*

$$\begin{aligned} &F(v+w) \le F(v) + F(w) && \forall v, w \in H^r, \\ &F(v) \le C\|v\|_r && \forall v \in H^r, \\ &F(v) = 0 && \forall v \in \Pi_{r-1} \end{aligned}$$

erfüllt sind, dann gibt es eine Konstante $C = C(\Omega)$, für die

$$F(v) \le C|v|_r \quad \forall v \in H^r$$

ist.

(a) Beweisen Sie das Bramble-Hilbert-Lemma für $d = 1$, $\Omega = (0,1)$ und $r = 2$. Hinweis: Zeigen Sie wie in Problemstellung 5.1, dass $\|v - Q_1 v\|_2 \le C|v|_2$ gilt. Dann ist $F(v) \le F(v - Q_1 v) + F(Q_1 v) = F(v - Q_1 v) \le C\|v - Q_1 v\|_2 \le C|v|_2$.

(b) Verwenden Sie das Bramble-Hilbert-Lemma, um zu zeigen, dass (5.29) und (5.30) gelten. Hinweis: Tun Sie dies zunächst für ein festes Dreieck $\widehat{K}$ von Einheitsgröße und führen Sie anschließend eine affine Transformation dieses Einheitsdreiecks auf ein kleines Dreieck K durch (siehe Problemstellung A.14). Verwenden Sie zum Beweis von (5.29) $F(v) = \|I_h v - v\|_{L_2(\widehat{K})}$ und schätzen Sie die Knotenwerte durch die Sobolev-Ungleichung $|v(\widehat{P}_j)| \le C\|v\|_{H^2(\widehat{K})}$ ab.

Problem 5.13. Beweisen Sie (5.62) mit dem Bramble-Hilbert-Lemma.

Problem 5.14. Beweisen Sie (5.68).

Problem 5.15. Beweisen Sie ein Analogon von Lemma 5.1 für die Knotenquadraturformel (5.64).

Problem 5.16. Beweisen Sie, dass jede Lösung von (5.71) die Gleichung (5.72) erfüllt.

Problem 5.17. (Übung am Rechner.) Betrachten Sie das Zweipunkt-Randwertproblem aus Problemstellung 4.4. Wenden Sie die Methode der finiten Elemente (5.3) an. Benutzen Sie als Basis die stückweise linearen Approximationsfunktionen auf derselben Zerlegung wie in Problemstellung 4.4 mit $h = 1/10,\ 1/20$. Bestimmen Sie die exakte Lösung und berechnen Sie das Maximum des Fehlers an den Gitterpunkten.

Problem 5.18. (Übung am Rechner.) Betrachten Sie das Randwertproblem aus Problemstellung 4.5. Lösen Sie es mit der Methode der finiten Elemente (5.26). Benutzen Sie als Basis die stückweise linearen Approximationsfunktionen auf derselben Zerlegung wie in Problemstellung 4.5. Teilen Sie diese jedoch in Dreiecke auf, indem Sie eine Diagonale mit positivem Anstieg in jedes Gitterquadrat mit $h = 1/10,\ 1/20$ einfügen. Wiederholen Sie die exakte Lösung und berechnen Sie die L_2-Norm des Fehlers. Verwenden Sie die baryzentrische Quadraturregel, um die Steifigkeitsmatrix, den Lastvektor und die L_2-Norm zu berechnen.

6

Das elliptische Eigenwertproblem

Eigenwertprobleme sind bei der mathematischen Analyse partieller Differentialgleichungen wesentlich und treten beispielsweise bei der Modellierung schwingender Membrane auf. Bei der Analyse zeitabhängiger partieller Differentialgleichungen ist es wichtig, Funktionen nach Eigenfunktionen zu entwickeln, weshalb wir solche Entwicklungen in Abschnitt 6.1 behandeln werden. In Abschnitt 6.2 stellen wir einige einfache Näherungen und Resultate für die numerische Lösung von Eigenwertproblemen vor.

6.1 Entwicklung nach Eigenfunktionen

Wir werden zunächst das Eigenwertproblem behandeln, das dem symmetrischen Fall des Zweipunkt-Randwertproblems aus Kapitel 2 entspricht. Dabei ist eine Zahl λ und eine von null verschiedene Funktion φ gesucht, für die

$$\mathcal{A}\varphi := -(a\varphi')' + c\varphi = \lambda\varphi \quad \text{in } \Omega = (0,1) \quad \text{mit } \varphi(0) = \varphi(1) = 0 \tag{6.1}$$

gilt. Hier sind a und c glatte Funktionen auf $\bar{\Omega}$ mit $a(x) \geq a_0 > 0$ und $c(x) \geq 0$. Eine solche Zahl λ wird als Eigenwert und φ als zugehörige Eigenfunktion bezeichnet.

Erinnern wir uns daran, dass das Zweipunkt-Randwertproblem

$$\mathcal{A}u = f \quad \text{in } \Omega \quad \text{mit } u(0) = u(1) = 0 \tag{6.2}$$

in schwacher Form folgendermaßen geschrieben werden kann: Bestimme ein $u \in H_0^1 = H_0^1(\Omega)$, sodass

$$a(u,v) = (f,v) \quad \forall v \in H_0^1$$

gilt. Die Bilinearform und das Skalarprodukt sind durch

$$a(u,v) = \int_0^1 (a\,u'v' + c\,uv)\,\mathrm{d}x \text{ beziehungsweise } (u,v) = \int_0^1 uv\,\mathrm{d}x \tag{6.3}$$

definiert. Mit dieser Notation kann nun das Eigenwertproblem (6.1) gestellt werden: Gesucht ist eine Zahl λ und eine Funktion $\varphi \in H_0^1$, $\varphi \neq 0$, für die

$$a(\varphi, v) = \lambda\,(\varphi, v) \quad \forall v \in H_0^1 \tag{6.4}$$

gilt.

Wir werden auch das Dirichletsche Eigenwertproblem betrachten, eine Zahl λ und eine von null verschiedene Funktion φ zu bestimmen, für die

$$-\Delta\varphi = \lambda\varphi \quad \text{in } \Omega \quad \text{mit } \varphi = 0 \quad \text{auf } \Gamma \tag{6.5}$$

auf dem beschränkten Gebiet Ω in $\mathbf{R}^d$ mit dem glatten Rand Γ gilt. Das zugehörige Dirichletsche Randwertproblem lautet

$$-\Delta u = f \quad \text{in } \Omega \quad \text{mit } u = 0 \quad \text{auf } \Gamma. \tag{6.6}$$

In Variationsform ist ein $u \in H_0^1 = H_0^1(\Omega)$ zu bestimmen, für das

$$a(u, v) = (f, v) \quad \forall v \in H_0^1$$

gilt, wobei nun

$$a(u, v) = \int_\Omega \nabla u \cdot \nabla v \,\mathrm{d}x = (\nabla u, \nabla v) \quad \text{und} \quad (u, v) = \int_\Omega uv \,\mathrm{d}x \tag{6.7}$$

ist. Bei Variationsform des zu (6.5) gehörigen Eigenwertproblems ist wieder eine Zahl λ und eine Funktion $\varphi \in H_0^1$, $\varphi \neq 0$ gesucht, für die (6.4) gilt.

Erinnern wir uns daran, dass für eine Lösung u von (6.6) und für $f \in H^k$ die Beziehung $u \in H^{k+2} \cap H_0^1$ gilt (siehe Abschnitt 3.7). Daraus können wir sofort schlussfolgern, dass eine Eigenfunktion glatt ist: Wegen $\varphi \in L_2$ folgt aus der elliptischen Regularität $\varphi \in H^2 \cap H_0^1$, was tatsächlich $\varphi \in H^4 \cap H_0^1$ zeigt, usw. Die entsprechende Feststellung gilt auch für das einfachere Eigenwertproblem (6.1).

Beide Eigenwertprobleme (6.1) und (6.6) haben folglich die Variationsform (6.4). Dies wäre auch der Fall, wenn wir statt des Laplace-Operators den allgemeineren elliptischen Operator $\mathcal{A}u = -\nabla\cdot(a\nabla u)+c\,u$ mit den in Kapitel 3 beschriebenen, geeigneten Randbedingungen verwenden würden. Dies würde auf das folgende, allgemeinere Eigenwertproblem führen. Sei $H = L_2 = L_2(\Omega)$ und V ein linearer Teilraum des $H^1 = H^1(\Omega)$ mit $\Omega \subset \mathbf{R}^d$. Wir nehmen an, dass die Bilinearform $a(\cdot, \cdot)$ symmetrisch und koerzitiv auf V ist, d. h.

$$a(v, v) \geq \alpha \|v\|_1^2 \quad \forall v \in V \quad \text{mit } \alpha > 0$$

gilt. Dann ist ein $\varphi \in V$, $\varphi \neq 0$ und $\lambda \in \mathbf{R}$ zu bestimmen, für das

$$a(\varphi, v) = \lambda(\varphi, v) \quad \forall v \in V \tag{6.8}$$

erfüllt ist. Dabei bezeichnet $(\cdot, \cdot)$ das Skalarprodukt in $H = L_2$. Beachten Sie, dass $a(\cdot, \cdot)$ ein Skalarprodukt in V ist.

Der Einfachheit halber werden wir den konkreten Fall (6.5) mit dem durch (6.7) definierten Skalarprodukt $a(\cdot,\cdot)$ betrachten. Wir fordern den Leser auf zu überprüfen, dass die von uns vorgestellte Theorie mit minimalen Notationsveränderungen tatsächlich auf das allgemeinere Eigenwertproblem (6.8) anwendbar ist.

Wir beginnen mit einigen allgemeinen, einfachen Eigenschaften von Eigenwerten und Eigenfunktionen.

Theorem 6.1. *Die Eigenwerte von* (6.5) *sind positiv. Zwei Eigenfunktionen, die zu verschiedenen Eigenwerten gehören, sind in L_2 und H_0^1 orthogonal.*

Beweis. Sei λ ein Eigenwert und φ die zugehörige Eigenfunktion. Dann gilt

$$\lambda\|\varphi\|^2 = \lambda\,(\varphi,\varphi) = a(\varphi,\varphi),$$

woraus sofort $\lambda > 0$ folgt. Seien λ_1 und λ_2 zwei verschiedene Eigenwerte und φ_1 und φ_2 die zugehörigen Eigenfunktionen. Dann ist

$$\lambda_1\,(\varphi_1,\varphi_2) = a(\varphi_1,\varphi_2) = a(\varphi_2,\varphi_1) = \lambda_2\,(\varphi_2,\varphi_1) = \lambda_2\,(\varphi_1,\varphi_2),$$

sodass

$$(\lambda_1 - \lambda_2)(\varphi_1,\varphi_2) = 0$$

ist. Wegen $\lambda_1 \neq \lambda_2$ folgt daraus $(\varphi_1,\varphi_2) = 0$ und somit auch $a(\varphi_1,\varphi_2) = 0$. □

Als ersten Schritt werden wir die Existenz eines Eigenwertes, und zwar die eines kleinsten Eigenwertes, zeigen. Dieser Eigenwert wird durch

$$\lambda_1 = \inf\left\{a(v,v):\ v \in H_0^1,\ \|v\| = 1\right\} \quad \text{mit } a(v,v) = \|\nabla v\|^2 \tag{6.9}$$

charakterisiert. Die Gleichung (6.9) kann auch in der als Rayleigh-Ritz-Charakterisierung des Haupteigenwertes bezeichneten Form

$$\lambda_1 = \inf_{v\neq 0} \frac{\|\nabla v\|^2}{\|v\|^2}$$

geschrieben werden, was aus $\|\nabla(\alpha v)\|^2 = \alpha^2\|\nabla v\|^2$ folgt. Da für einen beliebigen Eigenwert λ und die zugehörige Eigenfunktion φ die Gleichung

$$\|\nabla\varphi\|^2 = a(\varphi,\varphi) = \lambda\,(\varphi,\varphi) = \lambda\|\varphi\|^2$$

gilt, schlussfolgern wir $\lambda \geq \lambda_1$, sodass λ_1 eine untere Schranke für die Eigenwerte ist.

Theorem 6.2. *Das Infimum in* (6.9) *wird durch eine Funktion $\varphi_1 \in H_0^1$ angenommen. Diese Funktion ist eine Eigenfunktion von* (6.5) *und λ_1 der zugehörige Eigenwert.*

Beweis. Wir werden den Beweis der ersten Behauptung des Theorems auf das Ende dieses Abschnittes verschieben und annehmen, dass das Infimum durch $\varphi_1 \in H_0^1$ erreicht wird, also $\lambda_1 = \|\nabla\varphi_1\|^2$ und $\|\varphi_1\| = 1$ gilt. Wir zeigen nun, dass φ_1 eine Eigenfunktion von (6.5) ist, die zum Eigenwert λ_1 gehört, also

$$a(\varphi_1, v) = \lambda_1 (\varphi_1, v) \quad \forall v \in H_0^1. \tag{6.10}$$

Beachten Sie, dass für eine beliebige reelle Zahl α

$$a(\varphi_1 + \alpha v, \varphi_1 + \alpha v) = \lambda_1 + 2\alpha\, a(\varphi_1, v) + \alpha^2 a(v, v)$$

und

$$\|\varphi_1 + \alpha v\|^2 = 1 + 2\alpha\, (\varphi_1, v) + \alpha^2 \|v\|^2$$

gilt. Da das Verhältnis der beiden Normen durch λ_1 für alle α von unten beschränkt ist, gilt

$$\lambda_1 + 2\alpha a(\varphi_1, v) + \alpha^2 a(v, v) \geq \lambda_1 + 2\lambda_1 \alpha\, (\varphi_1, v) + \lambda_1 \alpha^2 \|v\|^2$$

oder

$$2\alpha\big(a(\varphi_1, v) - \lambda_1(\varphi_1, v)\big) + \alpha^2\big(a(v, v) - \lambda_1 \|v\|^2\big) \geq 0.$$

Nun nehmen wir an, dass (6.10) nicht erfüllt ist, sodass der Koeffizient von $\alpha \neq 0$ ist. Wenn wir dann $|\alpha|$ klein und das Vorzeichen so wählen, dass der erste Term negativ ist, erhalten wir einen Widerspruch. □

Aus Theorem 6.2 wissen wir also, dass mindestens eine Eigenfunktion φ_1 existiert. Wir wiederholen nun die obigen Betrachtungen im Teilraum V_1 von $V = H_0^1$, der aus Funktionen besteht, die bezüglich $(\cdot, \cdot)$ orthogonal zu φ_1 sind. Beachten Sie, dass diese Funktionen dann auch bezüglich $a(\cdot, \cdot)$ orthogonal zu φ_1 sind, da $a(v, \varphi_1) = \lambda_1(v, \varphi_1) = 0$ gilt. Wir betrachten also

$$\begin{aligned} \lambda_2 &= \inf\Big\{a(v, v) : v \in V,\ \|v\| = 1,\ (v, \varphi_1) = 0\Big\} \\ &= \inf\Big\{\|\nabla v\|^2 : v \in H_0^1,\ \|v\| = 1,\ (v, \varphi_1) = 0\Big\}. \end{aligned} \tag{6.11}$$

Offensichtlich gilt $\lambda_2 \geq \lambda_1$, da das Infimum hier über eine kleinere Menge von Funktionen v als in (6.9) gebildet wird. Wie oben kann man zeigen, dass das Infimum angenommen wird. Wir bezeichnen die Minimalfunktion als φ_2 (auf die Frage der Existenz von φ_2 werden wir später zurückkommen), die dann

$$a(\varphi_2, \varphi_2) = \|\nabla\varphi_2\|^2 = \lambda_2, \quad \|\varphi_2\| = 1, \quad (\varphi_1, \varphi_2) = 0$$

erfüllt. Um zu beweisen, dass φ_2 eine Eigenfunktion ist, zeigen wir genau wie oben zunächst

$$a(\varphi_2, v) = \lambda_2\, (\varphi_2, v) \quad \text{für alle } v \in H_0^1 \text{ mit } (v, \varphi_1) = 0.$$

Wir können uns davon überzeugen, dass die Gleichung für alle Funktionen $v \in H_0^1$ und nicht nur für die zu φ_1 orthogonalen gilt, indem wir beachten, dass jedes $v \in H_0^1$ in der Form

$$v = \alpha\varphi_1 + w \quad \text{mit } \alpha = (v, \varphi_1) \text{ und } (w, \varphi_1) = 0$$

geschrieben werden kann. Deshalb bleibt nur noch zu zeigen, dass $a(\varphi_2, \varphi_1) = \lambda_2 (\varphi_2, \varphi_1)$ ist. Dies folgt aber sofort aus $(\varphi_2, \varphi_1) = 0$ und $a(\varphi_2, \varphi_1) = 0$.

Fahren wir auf diese Weise fort, erhalten wir eine nichtfallende Folge von Eigenwerten $\{\lambda_j\}_{j=1}^{\infty}$ und eine zugehörige Folge von Eigenfunktionen $\{\varphi_j\}_{j=1}^{\infty}$, die paarweise orthogonal zueinander sind und die L_2-Norm 1 besitzen, sodass

$$\begin{aligned} \lambda_n &= a(\varphi_n, \varphi_n) \\ &= \inf \Big\{ a(v,v) : \ v \in H_0^1, \ \|v\| = 1, \ (v, \varphi_j) = 0, \ j = 1, \ldots, n-1 \Big\} \end{aligned} \tag{6.12}$$

gilt. Beachten Sie, dass der Prozess nicht nach einer endlichen Anzahl von Schritten abbricht. Denn wenn aus $(v, \varphi_j) = 0$, $j = 1, \ldots, n-1$ die Gleichung $v = 0$ folgen würde, dann wäre L_2 endlichdimensional, was natürlich nicht zutrifft. Man kann folgendes Theorem zeigen.

Theorem 6.3. *Sei λ_n der n-te Eigenwert von* (6.5). *Es gilt $\lambda_n \to \infty$ für $n \to \infty$.*

Den Beweis dieses Theorems verschieben wir ebenfalls auf später.

Eine Folgerung aus diesem Theorem ist, dass eine Zahl in der nichtfallenden Folge $\{\lambda_j\}_{j=1}^{\infty}$ nur endlich viele Male auftreten kann. Gilt $\lambda_{n-1} < \lambda_n = \lambda_{n+1} = \ldots = \lambda_{n+m-1} < \lambda_{n+m}$, dann sagen wir, dass der Eigenwert λ_n die Vielfachheit m besitzt. Die Menge der Linearkombinationen E_n von $\varphi_n, \ldots, \varphi_{n+m-1}$ bildet dann einen endlichdimensionalen linearen Raum der Dimension m, d. h. den zu λ_n gehörenden Eigenraum. Für $v \in E_n$ gilt daher $-\Delta v = \lambda_n v$.

Wir merken an, dass der erste Eigenwert oder Haupteigenwert λ_1 ein einfacher Eigenwert ist, sodass $\lambda_2 > \lambda_1$ gilt, und dass die Haupteigenfunktion φ_1 nach einem möglichen Vorzeichenwechsel positiv in Ω gewählt werden kann. Um die Beweisführung zu skizzieren, schreiben wir $\varphi_1 = \varphi = \varphi^+ - \varphi^-$, wobei $\varphi^{\pm} = \max(\pm\varphi, 0)$ ist. Man kann zeigen, dass im Falle $\varphi \geq 0$ $\varphi^{\pm} \in H_0^1$ und $\nabla\varphi^+ = \nabla\varphi$ und im Falle $\varphi < 0$ $\nabla\varphi^+ = 0$ gilt, sodass $a(\varphi^+, \varphi^-) = (\nabla\varphi^+, \nabla\varphi^-) = 0$ ist. Es gilt folglich $\|\nabla\varphi^{\pm}\|^2 = \lambda_1 \|\varphi^{\pm}\|^2$, da anderenfalls

$$\lambda_1 = \|\nabla\varphi\|^2 = \|\nabla\varphi^+\|^2 + \|\nabla\varphi^-\|^2 > \lambda_1 \big(\|\varphi^+\|^2 + \|\varphi^-\|^2\big) = \lambda_1 \|\varphi\|^2 = \lambda_1$$

wäre, was ein Widerspruch ist. Somit erfüllen $\varphi^{\pm}$ die Gleichung $-\Delta\varphi^{\pm} = \lambda_1\varphi^{\pm}$. Dann ist aber $-\Delta\varphi^+ = \lambda_1\varphi^+ \geq 0$, und aus dem starken Maximumprinzip folgt $\varphi^+ > 0$ in Ω oder $\varphi^+ = 0$ in Ω, sodass $\varphi_1 > 0$ in Ω oder $\varphi_1 < 0$ in Ω ist. Außerdem folgt, dass λ_1 ein einfacher Eigenwert ist, da es nicht zwei orthogonale Eigenfunktionen mit konstantem Vorzeichen geben kann.

Wir wenden uns nun der Frage zu, wie Eigenfunktionen zur Reihenentwicklung anderer Funktionen benutzt werden können und betrachten den Fall eines allgemeinen Hilbert-Raumes H. Sei $\{\varphi_j\}_{j=1}^{\infty}$ eine Orthonormalfolge, d. h. eine Folge, für die

$$(\varphi_i, \varphi_j) = \delta_{ij} = \begin{cases} 1 & \text{für } i = j, \\ 0 & \text{für } i \neq j \end{cases}$$

gilt. Eine solche Folge wird als eine *Orthonormalbasis* (oder vollständige orthonormale Menge) bezeichnet, wenn jedes v in H beliebig gut durch eine Linearkombination von Elementen der Folge approximiert werden kann. Das bedeutet, dass für jedes $\epsilon > 0$ eine ganze Zahl N und reelle Zahlen $a_1, \ldots, a_N$ existieren, für die

$$\left\| v - \sum_{j=1}^{N} a_j \varphi_j \right\| < \epsilon$$

gilt.

Beachten Sie, dass es ausreicht, dies für Funktionen v aus einer dichten Teilmenge $\mathcal{M}$ von H zu zeigen. Sei also $v \in H$. Die Aussage, dass $\mathcal{M}$ eine dichte Teilmenge von H ist, bedeutet, dass man ein $w \in \mathcal{M}$ finden kann, für das $\|v - w\| < \epsilon/2$ gilt. Deshalb ist es hinreichend, eine Linearkombination zu finden, sodass

$$\left\| w - \sum_{j=1}^{N} a_j \varphi_j \right\| < \epsilon/2$$

gilt, weil dann folgende Abschätzung gilt:

$$\left\| v - \sum_{j=1}^{N} a_j \varphi_j \right\| \leq \|v - w\| + \left\| w - \sum_{j=1}^{N} a_j \varphi_j \right\| < \epsilon.$$

Lemma 6.1. *Sei $\{\varphi_j\}_{j=1}^{\infty}$ eine orthonormale Menge in H. Dann ist die beste Approximation von $v \in H$ durch eine Linearkombination der ersten N Funktionen φ_j durch $v_N = \sum_{j=1}^{N} (v, \varphi_j)\varphi_j$ gegeben.*

Beweis. Für beliebige $a_1, ..., a_N$ gilt

$$\begin{aligned} \left\| v - \sum_{j=1}^{N} a_j \varphi_j \right\|^2 &= \|v\|^2 - 2 \sum_{j=1}^{N} a_j (v, \varphi_j) + \sum_{j=1}^{N} a_j^2 \\ &= \|v\|^2 + \sum_{j=1}^{N} \big(a_j - (v, \varphi_j)\big)^2 - \sum_{j=1}^{N} (v, \varphi_j)^2, \end{aligned}$$

woraus sich sofort das gesuchte Resultat ergibt. □

Weil die linke Seite nichtnegativ ist, erhalten wir, insbesondere für $a_j = (v, \varphi_j)$,

$$\sum_{j=1}^{N}(v,\varphi_j)^2 \le \|v\|^2.$$

Da dies für jedes N gilt, schließen wir auf die *Besselsche Ungleichung*

$$\sum_{j=1}^{\infty}(v,\varphi_j)^2 \le \|v\|^2.$$

Wenn $\{\varphi_j\}_{j=1}^{\infty}$ eine *Orthonormalbasis* ist, dann muss der Fehler in der besten Approximation für $N \to \infty$ gegen null gehen, sodass

$$\left\|v - \sum_{j=1}^{N}(v,\varphi_j)\varphi_j\right\|^2 = \|v\|^2 - \sum_{j=1}^{N}(v,\varphi_j)^2 \to 0 \quad \text{für } N \to \infty \tag{6.13}$$

gilt. Dies ist äquivalent zur *Parsevalschen Gleichung*

$$\sum_{j=1}^{\infty}(v,\varphi_j)^2 = \|v\|^2.$$

Folglich bildet die orthonormale Menge $\{\varphi_j\}_{j=1}^{\infty}$ genau dann eine Orthonormalbasis von H, wenn die Parsevalsche Gleichung für alle v in H gilt (oder für alle v in einer dichten Teilmenge von H).

Kommen wir auf den Fall $H = L_2$ zurück, gilt folgendes Theorem.

Theorem 6.4. *Die Eigenfunktionen $\{\varphi_j\}_{j=1}^{\infty}$ von* (6.5) *bilden eine Orthonormalbasis des L_2. Wenn darüber hinaus $v \in H_0^1$ ist, gilt*

$$a(v,v) = \sum_{j=1}^{\infty}\lambda_j(v,\varphi_j)^2 < \infty \quad \textit{mit } a(v,v) = \|\nabla v\|^2. \tag{6.14}$$

Umgekehrt gilt: Wenn die Summe konvergiert, dann ist $v \in H_0^1$ und es gilt (6.14).

Beweis. Aus der obigen Diskussion folgt, dass es ausreicht, die Gültigkeit von (6.13) für alle v in H_0^1 zu zeigen. Wir werden beweisen, dass für ein solches v

$$\left\|v - \sum_{j=1}^{N}(v,\varphi_j)\varphi_j\right\| \le C\lambda_{N+1}^{-1/2} \tag{6.15}$$

gilt, was dann mit Theorem 6.3 den Beweis des Theorems vervollständigt.

Zum Beweis von (6.15) setzen wir $r_N = v - \sum_{j=1}^{N}(v,\varphi_j)\varphi_j$. Dann gilt $(r_N,\varphi_j) = 0$ für $j = 1,...,N$, sodass

$$\frac{\|\nabla r_N\|^2}{\|r_N\|^2} \ge \inf\left\{\|\nabla v\|^2 : v \in H_0^1, \|v\| = 1, (v,\varphi_j) = 0, j = 1,\dots,N\right\} = \lambda_{N+1}$$

und somit

$$\|r_N\| \le \lambda_{N+1}^{-1/2} \|\nabla r_N\|$$

ist. Nun reicht es aus zu zeigen, dass die Folge $\|\nabla r_N\|$ beschränkt ist. Da aber

$$\begin{aligned}\|\nabla r_N\|^2 = a(r_N, r_N) &= a(v,v) - 2\sum_{j=1}^{N}(v,\varphi_j)a(v,\varphi_j) + \sum_{j=1}^{N}(v,\varphi_j)^2 a(\varphi_j,\varphi_j) \\ &= a(v,v) - \sum_{j=1}^{N}\lambda_j(v,\varphi_j)^2 \le a(v,v) = \|\nabla v\|^2\end{aligned}$$

gilt, ist der Beweis vollständig. □

Aus (6.12) folgt unmittelbar

$$\lambda_n = \min_{\substack{(v,\varphi_j)=0,\\ j=1,\dots,n-1}} \frac{a(v,v)}{\|v\|^2}. \tag{6.16}$$

Dies wiederum impliziert das folgende *Min-Max-Prinzip*.

Theorem 6.5. *Es gilt*

$$\lambda_n = \min_{V_n} \max_{v\in V_n} \frac{a(v,v)}{\|v\|^2}, \tag{6.17}$$

wobei V_n über alle Teilräume des H_0^1 mit endlicher Dimension n geht.

Beweis. Sei E_n der n-dimensionale Teilraum der Linearkombinationen $v = \sum_{j=1}^{n}\alpha_j\varphi_j$ von Eigenfunktionen $\varphi_1,\dots,\varphi_n$. Dann gilt offensichtlich

$$\max_{v\in E_n}\frac{a(v,v)}{\|v\|^2} = \max_{\alpha_1,\dots,\alpha_n}\frac{\sum_{j=1}^{n}\alpha_j^2\lambda_j}{\sum_{j=1}^{n}\alpha_j^2} = \lambda_n,$$

wobei das Maximum bei φ_n angenommen wird. Es bleibt deshalb zu zeigen, dass für jedes V_n der Dimension n

$$\max_{v\in V_n}\frac{a(v,v)}{\|v\|^2} \ge \lambda_n$$

ist. Um uns davon zu überzeugen, wählen wir $w \in V_n$ so, dass

$$(w,\varphi_j) = 0 \quad \text{für } j = 1,\dots,n-1$$

gilt. Wenn $\{\psi_j\}_{j=1}^{n}$ eine Basis für V_n ist, dann kann ein solches $w = \sum_{j=1}^{n}\alpha_j\psi_j$ aus dem linearen Gleichungssystem

$$(w,\varphi_j) = \sum_{l=1}^{n}\alpha_l(\psi_l,\varphi_j) = 0, \quad j = 1,\dots,n-1$$

bestimmt werden, das eine von null verschiedene Lösung besitzt, weil die Anzahl der Gleichungen kleiner als n ist. Aus (6.16) folgt

$$\frac{a(w,w)}{\|w\|^2} \geq \lambda_n,$$

was den Beweis von (6.17) abschließt. □

Eine Konsequenz dieses Resultates besteht darin, dass die Eigenwerte monoton von dem zugrunde liegenden Gebiet abhängen. Genauer gesagt, gilt $\lambda_n(\tilde{\Omega}) \leq \lambda_n(\Omega)$ für alle n, wenn $\Omega \subset \tilde{\Omega}$ ist und die zugehörigen Eigenwerte $\lambda_n(\Omega)$ und $\lambda_n(\tilde{\Omega})$ sind. Und zwar gilt, nachdem wir die Funktionen in $H_0^1(\Omega)$ durch Null in $\tilde{\Omega} \setminus \Omega$ fortgesetzt haben, $H_0^1(\Omega) \subset H_0^1(\tilde{\Omega})$ und folglich wird das Minimum im Ausdruck für $\lambda_n(\Omega)$ in (6.17) über eine kleinere Menge von n-dimensionalen Räumen gebildet als für $\lambda_n(\tilde{\Omega})$. Daher ist das letztere Minimum mindestens so klein.

Wir kommen nun auf die mathematischen Feinheiten zu sprechen, die wir vorhin aufgeschoben haben. Für deren Behandlung sind wir auf das Konzept der Kompaktheit angewiesen, das wir zunächst kurz besprechen.

Wir sagen, dass eine Menge $\mathcal{M}$ in einem Hilbert-Raum H (mit Norm $\|\cdot\|$) präkompakt ist, wenn jede unendliche Folge $\{u_n\}_{n=1}^{\infty} \subset \mathcal{M}$ eine konvergente Teilfolge enthält, d. h. wenn eine Teilfolge $\{u_{n_j}\}_{j=1}^{\infty}$ und ein Element $\bar{u} \in H$ existieren, sodass

$$\|u_{n_j} - \bar{u}\| \to 0 \quad \text{für } j \to \infty \tag{6.18}$$

gilt. Aus der elementaren Analysis wissen wir beispielsweise, dass eine beschränkte, unendliche Folge reeller Zahlen präkompakt ist (Theorem von Bolzano-Weierstrass). Die Menge $\mathcal{M}$ wird als kompakt bezeichnet, wenn sie auch abgeschlossen ist, d. h. wenn der Limes $\bar{u}$ in (6.18) immer zu $\mathcal{M}$ gehört. Weiter unten benötigen wir das folgende Resultat für $H_0^1 = H_0^1(\Omega)$ mit $\Omega \subset \mathbf{R}^d$, dessen Beweis jedoch über den Umfang dieses Buches hinausgeht.

Lemma 6.2. (Rellich-Lemma.) *Eine beschränkte Teilmenge $\mathcal{M}$ von H^1 ist in L_2 präkompakt.*

Folglich existiert im Falle $\{u_n\}_{n=1}^{\infty} \subset H^1$ und $\|u_n\|_1 \leq C$ für $n \geq 1$ eine Teilfolge $\{u_{n_j}\}_{j=1}^{\infty}$ und ein $\bar{u} \in L_2$, für die (6.18) in der L_2-Norm erfüllt ist.

Wir werden dies nun zum Beweis der ersten Behauptung aus Theorem 6.2 verwenden. Diese besagt, dass das Infimum in (6.9) in H_0^1 angenommen wird. Dazu nehmen wir eine Folge $\{u_n\}_{n=1}^{\infty}$, für die

$$\|\nabla u_n\|^2 = a(u_n, u_n) \to \lambda_1 \quad \text{und } \|u_n\| = 1 \quad \text{für } n \to \infty \tag{6.19}$$

gilt, was nach Definition des Infimums möglich ist. Dann ist $\{u_n\}_{n=1}^{\infty}$ offensichtlich in H^1 beschränkt. Nach Lemma 6.2 können wir deshalb eine Teilfolge

auswählen, die gegen ein Element $\varphi_1 \in L_2$ konvergiert. Wir können annehmen, dass $\{u_n\}_{n=1}^{\infty}$ selbst diese Teilfolge ist, und somit $\|u_n - \varphi_1\| \to 0$ gilt, wobei wir bei Bedarf die Notation ändern.

Wir wollen nun zeigen, dass $\{u_n\}_{n=1}^{\infty}$ in H_0^1 konvergiert. Nach einer kurzen Zwischenrechnung erhalten wir

$$\|\nabla(u_n - u_m)\|^2 = 2\|\nabla u_n\|^2 + 2\|\nabla u_m\|^2 - 4\|\tfrac{1}{2}\nabla(u_n + u_m)\|^2$$

(die Parallelogrammidentität), und nach Definition von λ_1 ist

$$\|\tfrac{1}{2}\nabla(u_n + u_m)\|^2 \geq \lambda_1 \|\tfrac{1}{2}(u_n + u_m)\|^2.$$

Es gilt also

$$\|\nabla(u_n - u_m)\|^2 \leq 2\|\nabla u_n\|^2 + 2\|\nabla u_m\|^2 - 4\lambda_1\|\tfrac{1}{2}(u_n + u_m)\|^2. \tag{6.20}$$

Offensichtlich gilt $\|\frac{1}{2}(u_n + u_m)\| \to \|\varphi_1\|$ für $n, m \to \infty$, und wegen $\|u_n\| = 1$ auch $\|\varphi_1\| = 1$. Also geht die rechte Seite von (6.20) mit (6.19) gegen null, sodass $\{u_n\}_{n=1}^{\infty}$ eine Cauchy-Folge in H_0^1 ist, d. h. es gilt $\|\nabla(u_n - u_m)\| \to 0$ für $m, n \to \infty$. Da H_0^1 ein Hilbert-Raum ist, konvergiert die Folge somit gegen ein Element von H_0^1, das dann dasselbe sein muss wie der Grenzwert in L_2, also φ_1. Insbesondere gilt

$$\|\nabla\varphi_1\|^2 = \lambda_1,$$

was zeigt, dass φ_1 das Minimum in (6.9) realisiert.

Der Beweis der Behauptung, dass das Infimum in (6.11) angenommen wird, verläuft analog. Und zwar gilt für eine Minimalfolge $\{u_n\}_{n=1}^{\infty}$, die gegen ein φ_2 in L_2 konvergiert und die Nebenbedingungen in (6.11) erfüllt, wegen $(\frac{1}{2}(u_n + u_m), \varphi_1) = 0$ nun auch (6.20), wobei λ_1 durch λ_2 ersetzt wurde. Wir schlussfolgern, dass u_n in H_0^1 gegen φ_2 konvergiert und die Gültigkeit von

$$\|\varphi_2\| = 1, \quad (\varphi_2, \varphi_1) = 0, \quad \|\nabla\varphi_2\|^2 = \lambda_2.$$

Schließlich geben wir noch den Beweis von Theorem 6.3 an. Dazu wir gehen wir davon aus, dass das Resultat nicht gilt, sodass

$$\|\nabla\varphi_n\|^2 = \lambda_n \leq C \quad \text{für } n \geq 1$$

ist. Dann enthält $\{\varphi_n\}_{n=1}^{\infty}$ wegen der Kompaktheit allerdings eine Teilfolge $\{\varphi_{n_j}\}_{j=1}^{\infty}$, die in L_2 konvergiert. Da $\{\varphi_n\}_{n=1}^{\infty}$ aber orthonormal ist, gilt

$$\|\varphi_i - \varphi_j\|^2 = \|\varphi_1\|^2 + \|\varphi_2\|^2 = 2 \quad \text{für } i \neq j,$$

sodass keine konvergente Teilfolge existieren kann.

Wie bereits erwähnt, ist die Theorie für das allgemeinere Eigenwertproblem (6.8) analog. In diesem Fall gelten beispielsweise die Theoreme 6.4 und 6.5 mit $a(v, w) = \int_\Omega \left(a\,\nabla v \cdot \nabla w + c\,vw\right) \mathrm{d}x$ und H_0^1 anstelle von V. Insbesondere ist

$$\lambda_1 = \min_{v \in V} \frac{\displaystyle\int_\Omega \left(a\,|\nabla v|^2 + c\,v^2\right) \mathrm{d}x}{\displaystyle\int_\Omega v^2 \,\mathrm{d}x}.$$

Wir schließen mit zwei Beispielen ab, für die wir das Eigenwertproblem explizit lösen können.

Beispiel 6.1. Sei $\Omega = (0, b) \subset \mathbf{R}$. Das Problem (6.5) reduziert sich dann auf

$$-u'' = \lambda u \quad \text{in } \Omega \quad \text{mit } u(0) = u(b) = 0. \tag{6.21}$$

Hier können wir die Eigenfunktionen und Eigenwerte leicht explizit bestimmen. Und zwar lautet die allgemeine Lösung der Differentialgleichung (6.21)

$$u = C_1 \sin(\sqrt{\lambda}x) + C_2 \cos(\sqrt{\lambda}x) \quad \text{mit } \lambda > 0.$$

Aus den Randbedingungen ergibt sich $C_2 = 0$ und $\sqrt{\lambda}b = n\pi$. Folglich sind die Eigenfunktionen $\{\sin(n\pi x/b)\}_{n=1}^{\infty}$ und die zugehörigen Eigenwerte $\lambda_n = n^2\pi^2/b^2$. Nach Normierung erhalten wir als Orthonormalbasis von Eigenfunktionen in $L_2(\Omega)$ also die Funktionen $\varphi_n(x) = \sqrt{2/b}\sin(n\pi x/b),\ n = 1, 2, \ldots$. Wir weisen insbesondere darauf hin, dass die Eigenwerte mit wachsendem b abnehmen.

Beispiel 6.2. Sei $\Omega = (0, b) \times (0, b)$. Wir betrachten das Eigenwertproblem

$$-\Delta u = \lambda u \quad \text{in } \Omega \quad \text{mit } u = 0 \quad \text{auf } \Gamma.$$

Dann ist mit λ_n und φ_n wie in Beispiel 6.1 leicht zu überprüfen, dass die Produkte $\varphi_m(x_1)\varphi_l(x_2)$, $m, l = 1, 2, \ldots$ Eigenfunktionen sind, die zu den Eigenwerten $\lambda_{ml} = (m^2 + l^2)\pi^2/b^2$ gehören. Um uns davon zu überzeugen, dass dies alle Eigenfunktionen sind, reicht es aus, die Parsevalsche Gleichung

$$\sum_{m,l} \Big(\int_\Omega v(x)\varphi_m(x_1)\varphi_l(x_2)\,\mathrm{d}x \Big)^2 = \int_\Omega v(x)^2\,\mathrm{d}x \tag{6.22}$$

zu zeigen. Unter Verwendung der Parsevalschen Gleichung in x_2 gilt aber

$$\int_0^1 v(x_1, x_2)^2\,\mathrm{d}x_2 = \sum_l w_l(x_1)^2, \quad w_l(x_1) = \int_0^1 v(x_1, x_2)\varphi_l(x_2)\,\mathrm{d}x_2. \tag{6.23}$$

Wenden wir die Parsevalsche Gleichung auf $w_l(x_1)$ an, so erhalten wir

$$\begin{aligned} \int_0^1 w_l(x_1)^2\,\mathrm{d}x_1 &= \sum_m (w_l, \varphi_m)^2 \\ &= \sum_m \Big(\int_0^1 \int_0^1 v(x_1, x_2)\varphi_m(x_1)\varphi_l(x_2)\,\mathrm{d}x \Big)^2. \end{aligned} \tag{6.24}$$

Wir integrieren (6.23) bezüglich x_1 und setzen (6.24) ein, was auf (6.22) führt.

Dies zeigt, dass die Zahlen $\lambda_{ml} = (m^2 + l^2)\pi^2/b^2$ die in wachsender Reihenfolge angeordneten Eigenwerte sind, wobei mehrfache Eigenwerte wiederholt werden. Um die Wachstumsrate der Eigenwerte λ_n mit steigendem n zu bestimmen, stellen wir fest, dass die Anzahl der Eigenwerte mit $\lambda_n \le \rho^2$ gleich der Anzahl der Gitterpunkte $(m\pi/b, l\pi/b)$ in der Kreisscheibe $D_\rho = \{x_1^2 + x_2^2 \le \rho^2\}$ ist. Da die Anzahl N_ρ solcher Gitterpunkte gleich der Anzahl der Gitterquadrate mit dem Flächeninhalt π^2/b^2 ist, die in D_ρ passen, gilt $N_\rho \approx \rho^2 b^2/\pi$. Folglich gilt für ein zu λ_{ml} gehörendes λ_n die Gleichung $\lambda_n = \lambda_{ml} \approx \rho^2 \approx \pi N_\rho/b^2 \approx \pi n/b^2$.

Da jedes Gebiet $\Omega \subset \mathbf{R}^2$ ein Quadrat enthält und wiederum in einem anderen Quadrat enthalten ist, folgt aus der Monotonie der Eigenwerte, dass für jedes Gebiet Ω positive Konstanten c und C existieren, für die $cn \le \lambda_n \le Cn$ gilt. In d Dimensionen ist die zugehörige Ungleichung $cn^{2/d} \le \lambda_n \le Cn^{2/d}$.

6.2 Numerische Lösung des Eigenwertproblems

Wir werden zunächst das eindimensionale Eigenwertproblem (6.1)

$$\mathcal{A}\varphi := -(a\varphi')' + c\varphi = \lambda\varphi \quad \text{in } \Omega = (0,1) \quad \text{mit } \varphi(0) = \varphi(1) = 0 \tag{6.25}$$

betrachten, wobei a und c glatte Funktionen mit $a(x) \ge a_0$ und $c(x) \ge 0$ auf $\bar{\Omega}$ sind. Zur Formulierung einer finiten Differenzendiskretisierung verwenden wir die Notation aus Abschnitt 4.1, die auf den Gitterpunkten $x_j = jh,\ j = 0, \dots, M$ mit $h = 1/M$ und $U_j \approx u(x_j)$ basiert, und betrachten das endlichdimensionale Eigenwertproblem

$$\begin{aligned} \mathcal{A}_h U_j &:= -\bar{\partial}(a_{j+1/2}\partial U_j) + c_j U_j = \Lambda U_j, \quad j = 1, \dots, M-1, \\ U_0 &= U_M = 0 \end{aligned} \tag{6.26}$$

mit $c_j = c(x_j)$ und $a_{j+1/2} = a(x_j + h/2)$. Die Gleichung an den inneren Gitterpunkten x_j kann dann in der Form

$$-(a_{j+1/2}U_{j+1} + (a_{j+1/2} + a_{j-1/2})U_j - a_{j-1/2}U_{j-1})/h^2 + c_j U_j = \Lambda U_j$$

geschrieben werden. Mit der tridiagonalen $(M-1) \times (M-1)$-Matrix A und dem zu den inneren Gitterpunkten gehörenden Vektor $\bar{U} = (U_1, \dots, U_{M-1}) \in \mathbf{R}^{M-1}$ schreiben wir die Gleichung (6.26) als das Matrix-Eigenwertproblem

$$A\bar{U} = \Lambda\bar{U}.$$

Zur Analyse führen wir ein diskretes Skalarprodukt und eine Norm ein:

$$(V, W)_h = h\sum_{j=0}^{M} V_j W_j \quad \text{und} \quad \|V\|_h = (V, V)_h^{1/2}$$

Wie man leicht sieht, ist der Operator $\mathcal{A}_h$ symmetrisch bezüglich dieses Skalarproduktes und wegen

$$(\mathcal{A}_h U, U)_h = h \sum_{j=1}^{M-1} \mathcal{A}_h U_j \, U_j = h \sum_{j=0}^{M-1} a_{j+1/2} (\partial U_j)^2 + h \sum_{j=1}^{M-1} c_j U_j^2$$

positiv definit.

Es ist zu erwarten, dass die Eigenwerte des diskreten Problems (6.26) (oder der Matrix A) diejenigen des kontinuierlichen Eigenwertproblems (6.25) approximieren. Wir werden dies lediglich für den Haupteigenwert Λ_1 zeigen.

Theorem 6.6. *Seien Λ_1 und λ_1 die kleinsten Eigenwerte von* (6.26) *und* (6.25). *Dann gilt*

$$|\Lambda_1 - \lambda_1| \le Ch^2.$$

Beweis. Wir führen den Beweis nur für $c = 0$ und überlassen den allgemeinen Fall der Problemstellung 6.4. Sei $U \in \mathbf{R}^{M+1}$ beliebig mit $U_0 = U_M = 0$, und sei $\tilde{u} = I_h U \in \mathcal{C}(\bar{\Omega})$ die zugehörige, stückweise lineare Interpolierte. Dann gilt

$$\lambda_1 \le \frac{a(\tilde{u}, \tilde{u})}{\|\tilde{u}\|^2}. \tag{6.27}$$

Wegen $\tilde{u}' = \partial U_j$ in (x_j, x_{j+1}) gilt unter Verwendung von $a \ge a_0 > 0$ in Ω

$$\begin{aligned} a(\tilde{u}, \tilde{u}) &= \sum_{j=0}^{M-1} \int_{x_j}^{x_{j+1}} a \, \mathrm{d}x \; (\partial U_j)^2 \le h \sum_{j=0}^{M-1} (a_{j+1/2} + Ch^2)(\partial U_j)^2 \\ &\le (\mathcal{A}_h U, U)_h (1 + Ch^2). \end{aligned}$$

Weiter ergibt sich durch eine kurze Rechnung

$$\begin{aligned} \|\tilde{u}\|^2 &= \sum_{j=0}^{M-1} h^{-2} \int_{x_j}^{x_{j+1}} \big((x_{j+1} - x) U_j + (x - x_j) U_{j+1} \big)^2 \, \mathrm{d}x \\ &= \tfrac{1}{3} h \sum_{j=0}^{M-1} (U_j^2 + U_j U_{j+1} + U_{j+1}^2) \end{aligned}$$

und wegen $U_0 = U_M = 0$

$$\|U\|_h^2 = \tfrac{1}{2} h \sum_{j=0}^{M-1} (U_j^2 + U_{j+1}^2). \tag{6.28}$$

Somit gilt wegen $a_{j+1/2} \ge a_0$, $c_j \ge 0$

$$\|U\|_h^2 - \|\tilde{u}\|^2 = \tfrac{1}{6} h \sum_{j=0}^{M-1} (U_j - U_{j+1})^2 = \tfrac{1}{6} h^3 \sum_{j=0}^{M-1} (\partial U_j)^2 \le Ch^2 (\mathcal{A}_h U, U)_h$$

oder

$$\|U\|_h^2 \le \|\tilde{u}\|^2 + Ch^2(\mathcal{A}_h U, U)_h.$$

Wird daher U als ein zu Λ_1 gehörender Eigenvektor mit $\|U\|_h = 1$ gewählt, dann gilt für kleines h

$$\frac{a(\tilde{u},\tilde{u})}{\|\tilde{u}\|^2} \le \frac{(\mathcal{A}_h U, U)_h(1+Ch^2)}{\|U\|_h^2 - Ch^2(\mathcal{A}_h U, U)_h} = \frac{\Lambda_1(1+Ch^2)}{1-C\Lambda_1 h^2} \le \Lambda_1 + Ch^2$$

und wegen (6.27)

$$\lambda_1 \le \Lambda_1 + Ch^2. \tag{6.29}$$

Zum Beweis der gegenteiligen Ungleichung stellen wir fest, dass für ein glattes u und für ein als Restriktion von u auf die Gitterpunkte definiertes U

$$(\mathcal{A}_h U, U)_h = a(u,u) + O(h^2) \quad \text{und} \quad \|U\|_h^2 = \|u\|^2 + O(h^2) \quad \text{für } h \to 0$$

gilt. In der zweiten Beziehung haben wir benutzt, dass wegen $U_0 = U_M = 0$ die Gleichung (6.28) die trapezoide Quadraturregel zweiter Ordnung für $\int_0^1 u^2\,\mathrm{d}x$ ist. Mit der Haupteigenfunktion des kontinuierlichen Problems $u = \varphi_1$ gilt insbesondere

$$\Lambda_1 \le \frac{(\mathcal{A}_h U, U)_h}{\|U\|_h^2} \le \frac{a(\varphi_1,\varphi_1) + Ch^2}{\|\varphi_1\| - Ch^2} \le \lambda_1 + Ch^2 \quad \text{für kleines } h.$$

Zusammen mit (6.29) vervollständigt dies den Beweis. □

Betrachten wir nun das Eigenwertproblem

$$-\Delta u = \lambda u \quad \text{in } \Omega \quad \text{mit } u = 0 \quad \text{auf } \Gamma, \tag{6.30}$$

wobei $\Omega \subset \mathbf{R}^2$ ist. Sei λ_n der n-te Eigenwert und φ_n die zugehörige Eigenfunktion.

Falls Ω das Quadrat $(0,1)\times(0,1)$ ist, können wir die in Kapitel 4 definierte Fünfpunkt-Approximation $-\Delta_h$ benutzen und das diskrete Eigenwertproblem

$$-\Delta_h u = \Lambda u \quad \text{in } \Omega \quad \text{mit } u = 0 \quad \text{auf } \Gamma$$

stellen. Dies kann wie unser obiges eindimensionales Problem behandelt werden, ist allerdings weniger interessant, weil die Eigenwerte von (6.30) direkt bestimmt werden können, wie wir in Beispiel 6.2 gesehen haben. Wenn Ω ein allgemeineres Gebiet mit einem gekrümmten, glatten Rand Γ ist, stoßen wir wie bei dem in Kapitel 4 diskutierten Dirichlet-Problem auf Schwierigkeiten, die dadurch hervorgerufen werden, dass sich das gleichmäßige Gitter nicht an das Gebiet anpasst. Die Analyse wird deshalb umständlich und wir werden sie hier nicht weiter fortführen.

In diesem Fall ist die Methode der finiten Elemente mit ihrer größeren Flexibilität besser geeignet. Wir werden eine elementare Darstellung einiger

einfacher Resultate angeben. Der Einfachheit halber nehmen wir an, dass $\Omega \subset \mathbf{R}^2$ ein konvexes Gebiet mit glattem Rand Γ ist und bezeichnen mit $\{S_h\}$ eine Familie von Räumen stetiger, stückweise linearer Funktionen auf regulären Triangulationen $\mathcal{T}_h$. Das zugehörige diskrete Eigenwertproblem lautet dann

$$a(u_h, \chi) = \lambda(u_h, \chi) \quad \forall \chi \in S_h, \quad \text{wobei } a(v, w) = (\nabla v, \nabla w) \text{ ist.} \tag{6.31}$$

Mithilfe der Basis $\{\Phi_i\}_{i=1}^{M_h}$ von Pyramidenfunktionen aus Abschnitt 5.2 und den positiv definiten Matrizen A und B mit den Elementen $a_{ij} = (\nabla\Phi_i, \nabla\Phi_j)$ beziehungsweise $b_{ij} = (\Phi_i, \Phi_j)$ kann dieses Problem in Matrixform als

$$A\,U = \lambda B\,U \tag{6.32}$$

geschrieben werden. Beachten Sie, dass im Gegensatz zum finiten Differenzenverfahren, die Matrix B nicht diagonal ist. Nichtsdestotrotz besitzt das Eigenwertproblem (6.31) oder (6.32) positive Eigenwerte $\{\lambda_{n,h}\}_{n=1}^{M_h}$ und orthogonale Eigenfunktionen $\{\varphi_{n,h}\}_{n=1}^{M_h}$. In diesem Fall gelten zunächst die folgenden Fehlerabschätzungen für die Eigenwerte.

Theorem 6.7. *Seien $\lambda_{n,h}$ und λ_n die n-ten Eigenwerte von* (6.31) *und* (6.30). *Dann existieren Konstanten C und h_0 (abhängig von n), sodass*

$$\lambda_n \le \lambda_{n,h} \le \lambda_n + Ch^2 \quad \textit{für } h \le h_0 \tag{6.33}$$

gilt.

Beweis. Wegen des Min-Max-Prinzips ist

$$\lambda_n = \min_{V_n \subset H_0^1} \max_{v \in V_n} \frac{\|\nabla v\|^2}{\|v\|^2}, \quad \dim V_n = n.$$

Ebenso gilt

$$\lambda_{n,h} = \min_{V_n \subset S_h} \max_{\chi \in V_n} \frac{\|\nabla \chi\|^2}{\|\chi\|^2}, \quad \dim V_n = n. \tag{6.34}$$

Wegen $S_h \subset H_0^1$ wird das Minimum im letzten Ausdruck über eine kleinere Menge von Teilräumen genommen als im vorhergehenden und ist folglich mindestens so groß, was die erste Ungleichung des Theorems beweist.

Zum Beweis der zweiten Ungleichung beachten wir, dass mit dem durch $\varphi_1, \ldots, \varphi_n$ und $E_{n,h} = R_h E_n$ aufgespannten Raum E_n mit der Ritzschen Projektion R_h

$$\lambda_{n,h} \le \max_{\chi \in E_{n,h}} \frac{\|\nabla \chi\|^2}{\|\chi\|^2} = \max_{v \in E_n} \frac{\|\nabla R_h v\|^2}{\|R_h v\|^2} \le \max_{v \in E_n} \frac{\|\nabla v\|^2}{\|R_h v\|^2} \tag{6.35}$$

gilt, weil $\|\nabla R_h v\| \le \|\nabla v\|$ ist. Denn Nenner können wir durch

$$\|R_h v\| \ge \|v\| - \|R_h v - v\|$$

abschätzen. Hier gilt für $v \in E_n$ unter Verwendung des Theorems 5.5 und der Regularitätsabschätzung (3.36)

$$\|R_h v - v\| \le Ch^2 \|v\|_2 \le Ch^2 \|\Delta v\| \le Ch^2 \lambda_n \|v\| \le Ch^2 \|v\|,$$

wobei wir benutzt haben, dass n fest ist. Somit ist

$$\|R_h v\| \ge \|v\|(1 - Ch^2)$$

und es folgt aus (6.35) für kleines h

$$\lambda_{n,h} \le \max_{v \in E_n} \frac{\|\nabla v\|^2}{\|v\|^2}(1 + Ch^2) \le \lambda_n + Ch^2,$$

was den Beweis vervollständigt. □

Eine Eigenschaft, die manchmal für Finite-Elemente-Räume $\{S_h\}$ benutzt wird, ist die sogenannte *inverse Ungleichung*

$$\|\nabla \chi\| \le Ch^{-1} \|\chi\| \quad \text{für } \chi \in S_h. \tag{6.36}$$

Insbesondere ist diese für stückweise lineare Finite-Elemente-Räume, die auf einer quasiuniformen Familie von Triangulationen $\{\mathcal{T}_h\}$ basieren, gültig (siehe Problemstellung 6.6). Wenn diese Ungleichung gilt, folgt unmittelbar aus (6.34), dass für den größten Eigenwert

$$\lambda_{M_h,h} = \max_{\chi \in S_h} \frac{\|\nabla \chi\|^2}{\|\chi\|^2} \le Ch^{-2} \tag{6.37}$$

gilt.

Man kann auch Fehlerabschätzungen für die Eigenfunktionen ableiten. Wir führen dies lediglich für den ersten Eigenwert aus, da wir die Schwierigkeiten vermeiden wollen, die bei höherer Vielfachheit der Eigenwerte auftreten.

Theorem 6.8. *Seien $\varphi_{1,h}$ und φ_1 normierte Eigenfunktionen, die zu den Haupteigenwerten von* (6.31) *beziehungsweise* (6.30) *gehören. Dann gilt*

$$\|\varphi_{1,h} - \varphi_1\| \le Ch^2 \tag{6.38}$$

und

$$\|\nabla \varphi_{1,h} - \nabla \varphi_1\| \le Ch. \tag{6.39}$$

Beweis. Wir entwickeln $R_h \varphi_1$ in diskrete Eigenfunktionen

$$R_h \varphi_1 = \sum_{j=1}^{M_h} a_j \varphi_{j,h}, \quad \text{wobei } a_j = (R_h \varphi_1, \varphi_{j,h})$$

und schlussfolgern aus der Parsevalschen Gleichung

$$\|R_h\varphi_1 - a_1\varphi_{1,h}\|^2 = \sum_{j=2}^{M_h} a_j^2. \tag{6.40}$$

Unter Verwendung von (6.31) erhalten wir

$$\lambda_{j,h}a_j = \lambda_{j,h}(R_h\varphi_1, \varphi_{j,h}) = a(R_h\varphi_1, \varphi_{j,h}) = a(\varphi_1, \varphi_{j,h}) = \lambda_1(\varphi_1, \varphi_{j,h})$$

und hieraus

$$(\lambda_{j,h} - \lambda_1)a_j = \lambda_1(\varphi_1 - R_h\varphi_1, \varphi_{j,h}).$$

Verwenden wir die erste Ungleichung in (6.33) und die Tatsache, dass λ_1 ein einfacher Eigenwert ist, gilt $\lambda_{jh} - \lambda_1 \geq \lambda_2 - \lambda_1 > 0$ für $j \geq 2$ und wir können

$$\sum_{j=2}^{M_h} a_j^2 \leq \sum_{j=2}^{M_h} \Big(\frac{\lambda_1}{\lambda_{j,h} - \lambda_1}\Big)^2 (\varphi_1 - R_h\varphi_1, \varphi_{j,h})^2 \leq C\|R_h\varphi_1 - \varphi_1\|^2 \leq Ch^4$$

schlussfolgern, sodass wegen (6.40)

$$\|R_h\varphi_1 - a_1\varphi_{1,h}\| \leq Ch^2$$

gilt. Deshalb ist

$$\|a_1\varphi_{1,h} - \varphi_1\| \leq \|R_h\varphi_1 - \varphi_1\| + \|R_h\varphi_1 - a_1\varphi_{1,h}\| \leq Ch^2, \tag{6.41}$$

und folglich müssen wir noch $\|a_1\varphi_{1,h} - \varphi_{1,h}\| = |a_1 - 1|$ abschätzen. Wir können annehmen, dass das Vorzeichen von $\varphi_{1,h}$ so gewählt wurde, dass $a_1 \geq 0$ erfüllt ist. Dann gilt aufgrund der Dreiecksungleichung und (6.41)

$$|a_1 - 1| = \big|\|a_1\varphi_{1,h}\| - \|\varphi_1\|\big| \leq \|a_1\varphi_{1,h} - \varphi_1\| \leq Ch^2,$$

was den Beweis von (6.38) abschließt.

Wir kommen nun zum Fehler im Gradienten. Unter Verwendung von (6.31) und den bereits definierten Fehlerschranken gilt

$$\begin{aligned}\|\nabla\varphi_{1,h} - \nabla\varphi_1\|^2 &= \|\nabla\varphi_{1,h}\|^2 - 2(\nabla\varphi_{1,h}, \nabla\varphi_1) + \|\nabla\varphi_1\|^2 \\ &= \lambda_{1,h} - 2\lambda_1(\varphi_{1,h}, \varphi_1) + \lambda_1 = \lambda_{1,h} - \lambda_1 + \lambda_1\|\varphi_{1,h} - \varphi_1\|^2 \leq Ch^2,\end{aligned}$$

was (6.39) zeigt und somit den Beweis des Theorems vervollständigt. □

6.3 Problemstellungen

Problem 6.1. Betrachten Sie das Problem (6.1).

(a) Zeigen Sie, dass mit den Funktionen $a(x)$ und $c(x)$ alle zugehörigen Eigenwerte wachsen.

(b) Bestimmen Sie die Eigenwerte, wenn $a(x)$ und $c(x)$ auf Ω konstant sind.
(c) Zeigen Sie, dass für gegebene Funktionen $a(x)$ und $c(x)$ Konstanten k_1 und k_2 existieren, für die

$$0 < k_1 n^2 \le \lambda_n \le k_2 n^2$$

gilt.

Problem 6.2. Betrachten Sie den Laplace-Operator in Kugelsymmetrie (siehe Problemstellung 1.4). Dann lautet das zugehörige Eigenwertproblem

$$-\frac{1}{r^2}\frac{d}{dr}\Big(r^2\frac{d\varphi}{dr}\Big) = \lambda\varphi \quad \text{für } 0 < r < 1 \quad \text{mit } \varphi(1) = 0,\ \varphi(0) \text{ endlich.}$$

Beweisen Sie, dass für die Eigenfunktionen φ_j von (6.2), die zu verschiedenen Eigenwerten λ_i und λ_j gehören,

$$\int_0^1 \varphi_i(r)\varphi_j(r)\, r^2\, \mathrm{d}r = \int_0^1 \varphi_i'(r)\varphi_j'(r)\, r^2\, \mathrm{d}r = 0$$

gilt, d. h. dass $\{\varphi_i\}_{i=1}^{\infty}$ eine orthogonale Menge in $L_2(r^2\,\mathrm{d}r;(0,1))$ ist, nämlich die Menge der auf $(0,1)$ bezüglich des Maßes $r^2\,\mathrm{d}r$ quadratintegrablen Funktionen. Beweisen Sie außerdem, dass geeignet normierte Funktionen φ_i eine Orthonormalbasis für $L_2(r^2\,\mathrm{d}r;(0,1))$ bilden.

Problem 6.3. (a) Verwenden Sie ein ähnliches Argument wie in Theorem 6.4, um zu zeigen, dass

$$v \in H^2 \cap H_0^1 \quad \text{genau dann gilt, wenn} \quad \sum_{i=1}^{\infty} \lambda_i^2 (v,\varphi_i)^2 < \infty \text{ ist.}$$

(b) Zeigen Sie

$$-\Delta v = \sum_{i=1}^{\infty} \lambda_i (v,\varphi_i)\varphi_i, \quad \|\Delta v\|^2 = \sum_{i=1}^{\infty} \lambda_i^2 (v,\varphi_i)^2 \quad \text{für } v \in H^2 \cap H_0^1.$$

Problem 6.4. Beweisen Sie Theorem 6.6 für den allgemeinen Fall, dass die Funktion $c(x) \ge 0$ nicht notwendigerweise verschwindet.

Problem 6.5. Zeigen Sie, dass für den größten Eigenwert von (6.26)

$$\Lambda_{M-1} \le CM^2$$

gilt, wobei C von M unabhängig ist.

Problem 6.6. Zeigen Sie die inverse Ungleichung (6.36) für stückweise lineare Finite-Elemente-Funktionen, die auf der Familie $\{\mathcal{T}_h\}$ quasiuniformer Triangulationen eines ebenen Gebietes basieren (siehe (5.52)). Hinweis: Führen Sie eine affine Transformation $x = A\hat{x} + b$ des Dreiecks K auf ein festes Referenzdreieck $\hat{K}$ mit Einheitsgröße aus (siehe Problemstellung A.15) und verwenden Sie die Tatsache, dass die Normen $\|\cdot\|_{L_2(\hat{K})}$ und $\|\cdot\|_{H^1(\hat{K})}$ auf dem endlichdimensionalen Raum Π_1 äquivalent sind.

Problem 6.7. Sei G die Greensche Funktion in (3.18) aus Abschnitt 3.4 und seien $\{\lambda_j\}_{j=1}^{\infty}$ und $\{\varphi_j\}_{j=1}^{\infty}$ die Eigenwerte und normierten Eigenfunktionen von (6.5) wie in Theorem 6.4. Zeigen Sie

$$G(x,y) = \sum_{j=1}^{\infty} \lambda_j^{-1} \varphi_j(x) \varphi_j(y).$$

7

Anfangswertprobleme für gewöhnliche Differentialgleichungen

Als Vorbereitung auf die Analyse von Anfangswertproblemen für parabolische und hyperbolische Differentialgleichungen werden wir in diesem Kapitel einige Fakten zu linearen Differentialgleichungen und deren numerischer Simulation wiederholen. Wir beginnen in Abschnitt 7.1 mit dem kontinuierlichen Problem und fahren in Abschnitt 7.2 mit der numerischen Lösung solcher Probleme durch Time-Stepping fort.

7.1 Das Anfangswertproblem für lineare Systeme

Wir betrachten zunächst das Anfangswertproblem für die lineare Differentialgleichung erster Ordnung

$$u' + au = f(t) \quad \text{für } t > 0 \quad \text{mit } u(0) = v, \tag{7.1}$$

wobei a eine Konstante, $f(t)$ eine gegebene glatte Funktion und v eine gegebene Zahl ist. Aus der elementaren Analysis ist uns bekannt, dass dieses Problem durch Multiplikation mit dem Integrationsfaktor e^{at} gelöst werden kann, was auf

$$(\mathrm{e}^{at}u)' = \mathrm{e}^{at}f(t)$$

und damit auf

$$\mathrm{e}^{at}u(t) = v + \int_0^t \mathrm{e}^{as}f(s)\,\mathrm{d}s$$

oder

$$u(t) = \mathrm{e}^{-at}v + \int_0^t \mathrm{e}^{-a(t-s)}f(s)\,\mathrm{d}s \tag{7.2}$$

führt.

Wir betrachten nun das entsprechende Problem für ein Gleichungssystem

$$u_i' + \sum_{j=1}^{N} a_{ij} u_j = f_i(t), \qquad i = 1, \ldots, N \text{ für } t > 0,$$
$$u_i(0) = v_i, \qquad i = 1, \ldots, N.$$

Führen wir den Spaltenvektor $u = (u_1, \ldots, u_N)^{\mathrm{T}}$ und analog dazu die Vektoren für $f(t)$ und v sowie die Matrix $A = (a_{ij})$ ein, kann dieses in der Form

$$u' + Au = f(t) \quad \text{für } t > 0 \quad \text{mit } u(0) = v \tag{7.3}$$

geschrieben werden.

Wir wollen nun die obige Lösungsmethode im skalaren Fall auf die Lösung des System (7.3) verallgemeinern. Dazu definieren wir zunächst die Exponentialfunktion einer $N \times N$-Matrix $B = (b_{ij})$ mithilfe der Reihenentwicklung

$$\mathrm{e}^B = \exp(B) = \sum_{j=0}^{\infty} \frac{1}{j!} B^j,$$

wobei $B^0 = I$ die Einheitssmatrix ist. Diese Definition basiert auf der MacLaurinschen Entwicklung von e^x. Man kann leicht zeigen, dass die Reihe für eine beliebige Matrix B konvergiert. Wir weisen darauf hin, dass für zwei kommutierende $N \times N$-Matrizen B_1 und B_2, d. h. mit $B_1 B_2 = B_2 B_1$, die Gleichung

$$\mathrm{e}^{B_1 + B_2} = \mathrm{e}^{B_1} \mathrm{e}^{B_2} = \mathrm{e}^{B_2} \mathrm{e}^{B_1} \tag{7.4}$$

erfüllt ist. Weil B_1 und B_2 kommutieren, gilt

$$(B_1 + B_2)^j = \sum_{l=0}^{j} \binom{j}{l} B_1^l B_2^{j-l}$$

und somit formal

$$\begin{aligned} \mathrm{e}^{B_1+B_2} &= \sum_{j=0}^{\infty} \frac{1}{j!} \sum_{l=0}^{j} \binom{j}{l} B_1^l B_2^{j-l} = \sum_{j=0}^{\infty} \sum_{l=0}^{j} \frac{1}{l!(j-l)!} B_1^l B_2^{j-l} \\ &= \sum_{l=0}^{\infty} \sum_{m=0}^{\infty} \frac{1}{l!\, m!} B_1^l B_2^m = \sum_{l=0}^{\infty} \frac{1}{l!} B_1^l \sum_{m=0}^{\infty} \frac{1}{m!} B_2^m = \mathrm{e}^{B_1} \mathrm{e}^{B_2}. \end{aligned}$$

Für nicht kommutierende B_1 und B_2 ist beispielsweise

$$(B_1 + B_2)^2 = B_1^2 + B_1 B_2 + B_2 B_1 + B_2^2 \neq B_1^2 + 2B_1 B_2 + B_2^2.$$

Für die Ableitung der Matrix e^{-tA} gilt

$$\frac{d}{dt} \mathrm{e}^{-At} = \frac{d}{dt} \sum_{j=0}^{\infty} \frac{1}{j!} t^j (-A)^j = \sum_{j=1}^{\infty} \frac{1}{(j-1)!} t^{j-1} (-A)^j = -A \mathrm{e}^{-tA}$$

und folglich erfüllt $u(t) = \mathrm{e}^{-tA}v$ die Gleichung

$$u' + Au = 0 \quad \text{für } t > 0 \quad \text{mit } u(0) = v. \tag{7.5}$$

Die Multiplikation mit e^{-tA} kann deshalb als ein Operator $E(t)$, genauer als Lösungsoperator von (7.5), betrachtet werden, der die Anfangswerte v dieses Problems in die Lösung zur Zeit t transformiert, sodass $u(t) = E(t)v = \mathrm{e}^{-tA}v$ gilt. Beachten Sie, dass wegen (7.4)

$$E(t+s) = \mathrm{e}^{-(t+s)A} = \mathrm{e}^{-tA}\mathrm{e}^{-sA} = E(t)E(s)$$

ist, was als Halbgruppeneigenschaft von $E(t)$ bezeichnet wird. Beachten Sie auch, dass auch folgt, dass A mit $E(t)$ kommutiert, also $AE(t) = E(t)A$ ist.

Zur Lösung der inhomogenen Gleichung (7.3) multiplizieren wir die Gleichung mit e^{tA} und erhalten

$$\mathrm{e}^{tA}(u' + Au) = \mathrm{e}^{tA}f(t) \quad \text{für } t > 0.$$

Dies kann in der Form

$$\frac{d}{dt}(\mathrm{e}^{tA}u) = \mathrm{e}^{tA}f(t)$$

geschrieben werden. Durch Integration folgt hieraus

$$\mathrm{e}^{tA}u(t) = v + \int_0^t \mathrm{e}^{sA}f(s)\,\mathrm{d}s.$$

Multiplizieren wir dies mit e^{-tA} und verwenden wir (7.4), so erhalten wir die zu (7.2) analoge Formel

$$u(t) = \mathrm{e}^{-tA}v + \int_0^t \mathrm{e}^{-(t-s)A}f(s)\,\mathrm{d}s. \tag{7.6}$$

Als Funktion des oben eingeführten Lösungsoperators kann dies auch als

$$u(t) = E(t)v + \int_0^t E(t-s)f(s)\,\mathrm{d}s \tag{7.7}$$

ausgedrückt werden. Wir weisen darauf hin, dass der Integrand $E(t-s)f(s)$ die Lösung an der Stelle $t-s$ der homogenen Gleichung in (7.5) mit den Anfangsdaten $f(s)$ ist. Das Integral kann deshalb als Superposition der Lösungen dieser Anfangswertprobleme betrachtet werden. Gleichung (7.7) wird häufig als Duhamel-Prinzip bezeichnet.

In einigen Fällen kann das System (7.3) auf eine endliche Anzahl skalarer Gleichungen des Typs (7.1) reduziert werden. Um uns hiervon zu überzeugen, nehmen wir an, dass A eine solche Gestalt hat, dass eine Diagonalmatrix Λ und eine nichtsinguläre Matrix P existiert, dass A in der Form $A = P\Lambda P^{-1}$ geschrieben werden kann. Wir können dann die neue abhängige Variable $w =$

$P^{-1}u$ und den Quellterm $g = P^{-1}f$ einführen und stellen fest, dass Gleichung (7.3) nach Multiplikation mit P^{-1} die Form

$$w' + \Lambda w = g(t) \quad \text{für } t > 0 \quad \text{mit } w(0) = P^{-1}v$$

annimmt. Mit den Diagonalementen λ_i von Λ können wir dies als

$$w_i' + \lambda_i w_i = g_i(t) \quad i = 1, \ldots, N \text{ für } t > 0$$

schreiben. Diese Gleichungen können nun einzeln gelöst werden. Die Lösung unseres Problems bestimmen wir aus $u = Pw$.

Die Annahme, dass A wie oben in eine Diagonalmatrix transformiert werden kann, ist beispielsweise erfüllt, wenn A symmetrisch (oder selbstadjungiert) ist. Dies bedeutet $a_{ij} = a_{ji}$ für alle i, j, und es gilt $A = P\Lambda P^{\mathrm{T}}$, wobei P eine orthogonale Matrix mit $P^{\mathrm{T}} = P^{-1}$ ist. In jedem Falle sind die Elemente von Λ die Eigenwerte von A. Die Methode ist anwendbar, wenn A genau N linear unabhängige Eigenvektoren besitzt. Für große N ist dies nicht unbedingt eine gute Methode für praktische Berechnungen, da die Diagonalisierung von A rechenaufwendig sein kann.

Wir werden nun kurz untersuchen, wie sich die Lösungen für große t verhalten und uns auf den Fall beschränken, dass A symmetrisch ist. Sei also $A = P\Lambda P^{\mathrm{T}}$ mit einer Orthogonalmatrix P und einer Diagonalmatrix Λ, deren Diagonaleinträge die reellen Eigenwerte λ_j von A sind. Erinnern wir uns daran, dass die j-te Spalte von P der zu λ_j gehörende Eigenvektor ist. Dann gilt wegen $P^{\mathrm{T}}P = PP^{\mathrm{T}} = I$

$$\mathrm{e}^{-tA} = \sum_{j=0}^{\infty} \frac{1}{j!}(-P\Lambda P^{\mathrm{T}})^j t^j = P\mathrm{e}^{-t\Lambda}P^{\mathrm{T}},$$

wobei $\mathrm{e}^{-t\Lambda}$ die Diagonalmatrix mit den Elementen $\mathrm{e}^{-t\lambda_j}$ ist. Es sei daran erinnert, dass im Falle einer symmetrischen Matrix die Matrixnorm von der Euklidischen Norm $|v| = (\sum_{i=1}^{N} v_i^2)^{1/2}$ von $v \in \mathbf{R}^N$ induziert wird. Damit gilt

$$|A| = \sup_{|v|=1} |Av| = \max_j |\lambda_j|.$$

Wegen $|P| = |P^{\mathrm{T}}| = 1$ schlussfolgern wir, dass mit dem kleinsten Eigenwert λ_1 von A

$$|E(t)| = |\mathrm{e}^{-tA}| = \max_j \mathrm{e}^{-t\lambda_j} = \mathrm{e}^{-t\lambda_1}$$

gilt. Sind insbesondere alle $\lambda_j \geq 0$, d. h. ist A positiv semidefinit, dann bestimmen wir aus (7.7) die Stabilitätsabschätzung

$$|u(t)| \leq |v| + \int_0^t |f(s)|\,\mathrm{d}s \quad \text{für } t \geq 0.$$

Wenn analog dazu A positiv definit ist, sodass $\lambda_1 > 0$ ist, dann gilt

$$|u(t)| \le \mathrm{e}^{-t\lambda_1}|v| + \int_0^t \mathrm{e}^{-(t-s)\lambda_1}|f(s)|\,\mathrm{d}s \quad \text{für } t \ge 0.$$

Wir sagen, dass das System in diesen beiden Fällen (7.5) *stabil* beziehungsweise *asymptotisch stabil* ist. Wenn A jedoch einen negativen Eigenwert besitzt, gilt $|\mathrm{e}^{-tA}| \to \infty$ für $t \to \infty$, und wir sagen dann, dass das System *instabil* ist.

Im stabilen Fall bleibt die Differenz zwischen zwei Lösungen $u_1(t)$ und $u_2(t)$ klein, wenn die Anfangswerte v_1 und v_2 und die Quellterme $f_1(t)$ und $f_2(t)$ nahe beieinander liegen. Genauer gilt, da die Differenz $u_1 - u_2$ eine Lösung des Systems mit der rechten Seite $f_1 - f_2$ und den Anfangswerten $v_1 - v_2$ ist,

$$|u_1(t) - u_2(t)| \le |v_1 - v_2| + \int_0^t |f_1(s) - f_2(s)|\,\mathrm{d}s \quad \text{für } t \ge 0.$$

Im asymptotisch stabilen Fall gilt analog

$$|u_1(t) - u_2(t)| = \mathrm{e}^{-t\lambda_1}|v_1 - v_2| + \int_0^t \mathrm{e}^{-(t-s)\lambda_1}|f_1(s) - f_2(s)|\,\mathrm{d}s \quad \text{für } t \ge 0$$

was insbesondere zeigt, dass der Einfluss der Anfangswerte und der Werte der Quellterme zur Zeit s für $t \to \infty$ exponentiell fällt.

Die oben eingeführte Analyse ist nicht anwendbar, wenn die Matrix A in (7.3) von t abhängt. Um dies zu illustrieren, betrachten wir die skalare Gleichung

$$u' + a(t)u = f(t) \quad \text{für } t > 0 \quad \text{mit } u(0) = v.$$

Sei $\widetilde{a}(t) = \int_0^t a(s)\,\mathrm{d}s$, sodass $\widetilde{a}'(t) = a(t)$ gilt. Folgen wir dann denselben Schritten wie oben, erhalten wir

$$u(t) = \mathrm{e}^{-\widetilde{a}(t)}v + \int_0^t \mathrm{e}^{-(\widetilde{a}(t)-\widetilde{a}(s))}f(s)\,\mathrm{d}s.$$

Weil im Allgemeinen aber $\widetilde{a}(t) - \widetilde{a}(s) = \int_s^t a(\tau)\,\mathrm{d}\tau \ne \int_0^{t-s} a(\tau)\,\mathrm{d}\tau = \widetilde{a}(t-s)$ ist, gilt das Analogon von (7.7) nicht. Stattdessen können wir diesmal

$$(7.8)\quad u(t) = E(t,0)v + \int_0^t E(t,s)f(s)\,\mathrm{d}s, \text{ with } E(t,s) = \exp\Big(-\int_s^t a(\tau)\,\mathrm{d}\tau\Big)$$

schreiben. Für das Anfangswertproblem des linearen Systems

$$u' + A(t)u = f(t) \quad \text{für } t > 0$$

mit der Matrix $A(t)$ kann man zeigen, dass die Lösung wiederum wie in (7.8) geschrieben werden kann. Die Matrix $E(t,s)$ wird dann im Allgemeinen aber eine kompliziertere Form annehmen. Man kann sie als einen Operator betrachten, der den Wert der Lösung der homogenen Gleichung $u' + A(t)u = 0$ zur

Zeit s in den Wert zur Zeit t überführt, sodass $u(t) = E(t,s)u(s)$ gilt. Wenn $A(t) = A$ unabhängig von t ist, dann hängt $E(t,s)$ nur von der Differenz $t-s$ und von $E(t,s) = E(t-s) = e^{-(t-s)A}$ ab.

Wir werfen einen flüchtigen Blick auf die allgemeine Theorie der gewöhnlichen Differentialgleichungen, indem wir das möglicherweise nichtlineare skalare Anfangswertproblem

$$u' = f(t,u) \quad \text{für } t > 0 \quad \text{mit } u(0) = v \tag{7.9}$$

betrachten, bei dem f nun eine glatte Funktion von t und u ist. Die Gleichung gibt die Richtung der Tangente der Lösungskurve an jedem Punkt an, wobei die Kurve durch die Punkte $(t, u(t)) \in \mathbf{R}^2$ definiert ist. Um zu zeigen, dass eine Lösung existiert, die an der Stelle $u(0) = v$ beginnt, die also eine durch $(0, v)$ verlaufende Lösungskurve $u(t)$ besitzt, kann man das *Euler-Verfahren* verwenden. Dieses besteht darin, die Lösung durch eine polygonale Kurve folgendermaßen zu approximieren: Sei k ein kleiner Zeitschritt und sei $t_n = nk$, $n = 0, 1, \ldots$. Dann wird die Approximation U^n von $u(t_n)$ sukzessive durch

$$\frac{U^n - U^{n-1}}{k} = f(t_{n-1}, U^{n-1}) \quad \text{für } n \geq 1 \tag{7.10}$$

oder

$$U^n = U^{n-1} + kf(t_{n-1}, U^{n-1}) \quad \text{für } n \geq 1 \quad \text{mit } U^0 = v$$

definiert. Dies bedeutet, wenn wir an der Stelle (t_{n-1}, U^{n-1}) starten, dann folgen wir der durch die Differentialgleichung in (7.9) definierten Tangentialrichtung und benutzen den Wert an der Stelle $t = t_n$ als Approximation von u an diesem Punkt. Die approximierte Lösung ist dann die stetige, stückweise lineare Funktion, die den Wert U^n an der Stelle t_n annimmt. Damit kann man zeigen, dass die so definierten Kurven für $k \to 0$ gegen eine Grenzkurve konvergieren, und dass es sich dabei um die gesuchte Lösung von (7.9) handelt. Für die Details verweisen wir auf ein Buch über gewöhnliche Differentialgleichungen. Eine andere Methode zur Lösung von (7.9), die Picard-Methode, wird in Problemstellung 7.4 diskutiert.

Wir werfen nun einen kurzen Blick auf Systeme zweiter Ordnung und beginnen mit dem einfachen skalaren Problem

$$u'' + au = 0 \quad \text{für } t > 0 \quad \text{mit } u(0) = v, \quad u'(0) = w, \quad a > 0. \tag{7.11}$$

Es ist bekannt und leicht zu überprüfen, dass die Lösung dieses Problems

$$u(t) = \cos(\sqrt{a}t)v + \frac{1}{\sqrt{a}}\sin(\sqrt{a}t)w \quad \text{für } t \geq 0$$

lautet.

Wir kommen nun zu der Verallgemeinerung auf ein System

$$u'' + Au = 0 \quad \text{für } t > 0 \quad \text{mit } u(0) = v, \quad u'(0) = w, \tag{7.12}$$

wobei u nun ein N-Vektor und A eine symmetrische, positiv definite $N \times N$-Matrix ist. Gilt $A = P\Lambda P^{\mathrm{T}}$, können wir $\sqrt{A}$ als die positiv definite Matrix $P\sqrt{\Lambda}P^{\mathrm{T}}$ definieren, wobei $\sqrt{\Lambda}$ die Diagonalmatrix mit den positiven Quadratwurzeln der Eigenwerte von A als Hauptdiagonalemente ist. Beachten Sie, dass $\sqrt{A}$ dieselben Eigenvektoren wie A besitzt. Wenden wir die Euler-Formeln zur Definition von $\cos(B)$ und $\sin(B)$ für die $N \times N$-Matrix B an, ist also

$$\cos B = \frac{1}{2}\big(\mathrm{e}^{iB} + \mathrm{e}^{-iB}\big), \quad \sin B = \frac{1}{2i}\big(\mathrm{e}^{iB} - \mathrm{e}^{-iB}\big),$$

so stellen wir leicht fest, dass die Lösung von (7.12)

$$u(t) = \cos(t\sqrt{A})v + (\sqrt{A})^{-1}\sin(t\sqrt{A})w \quad \text{für } t \geq 0 \tag{7.13}$$

ist. Sind $\{\varphi_j\}_{j=1}^N$ die normierten Eigenvektoren von A zu den Eigenwerten $\{\lambda_j\}_{j=1}^N$ und $v_j = (v, \varphi_j)$ sowie $w_j = (w, \varphi_j)$ die Komponenten von v und w in Richtung von φ_j (hier $(v, w) = v^{\mathrm{T}}w$), dann gilt

$$u_j(t) = (u(t), \varphi_j) = \cos(\sqrt{\lambda_j}t)v_j + \frac{1}{\sqrt{\lambda_j}}\sin(\sqrt{\lambda_j}t)w_j \quad \text{für } j = 1, \ldots, N.$$

Diese Komponenten variieren also mit steigendem t periodisch. Insbesondere geht $u(t)$ nicht gegen null, wenn t unendlich wird, im Gegensatz zu der Situation in (7.5) mit symmetrischem, positiv definitem A.

Eine andere Möglichkeit zur Behandlung eines Systems zweiter Ordnung besteht darin, es durch Einführung einer neuen abhängigen Variable auf eines erster Ordnung zu reduzieren. Wir setzen daher nun in Gleichung (7.12) $U = (U_1, U_2)^{\mathrm{T}} = (u, u')^{\mathrm{T}}$ und erhalten das System erster Ordnung

$$\begin{aligned} U_1' - U_2 &= 0, \\ U_2' + AU_1 &= 0 \end{aligned} \quad \text{für } t > 0 \quad \text{mit } U(0) = \begin{bmatrix} v \\ w \end{bmatrix}.$$

Die Lösung lautet

$$U(t) = \exp\left(t \begin{bmatrix} 0 & I \\ -A & 0 \end{bmatrix}\right) \begin{bmatrix} v \\ w \end{bmatrix} \quad \text{für } t \geq 0. \tag{7.14}$$

Es ist leicht zu erkennen, dass daraus (7.13) folgt (siehe Problemstellung 7.7).

7.2 Numerische Lösung gewöhnlicher Differentialgleichungen

Das soeben beschriebene Euler-Verfahren kann auch zur numerischen Lösung des Anfangswertproblems (7.3) verwendet werden. Auch für gewöhnliche Gleichungssysteme mit konstanten Koeffizienten sind numerische Methoden von

Bedeutung, da e^{-tA} unter Umständen nicht sehr leicht zu berechnen ist, wenn die Dimension N groß ist.

Beginnen wir mit dem Modellproblem

$$u' + au = 0 \quad \text{für } t > 0 \quad \text{mit } u(0) = v.$$

In diesem Falle liefert das Euler-Verfahren (7.10) für die approximative Lösung U^n an der Stelle $t_n = nk$

$$U^n = (1 - ak)U^{n-1} = (1 - ak)^n v.$$

(Bei der numerischen Analyse wird dieses Verfahren als *Vorwärts-Euler-Verfahren* bezeichnet, weil die Ableitung an der Stelle t_{n-1} durch den Vorwärts-Differenzenquotienten $(U^n - U^{n-1})/k$ ersetzt wird.) Für eine feste Zeit $t = t_n$ finden wir

$$U^n = \Big(1 - \frac{t}{n}a\Big)^n v \to \mathrm{e}^{-at}v \quad \text{für } n \to \infty,$$

sodass die numerische Lösung für $k \to 0$ so gegen die exakte Lösung konvergiert, dass $nk = t$ konstant gehalten wird.

Wir diskutieren nun die Größe des Fehlers und betrachten den Fall $a \geq 0$, in dem die Differentialgleichung stabile Lösungen besitzt. Wir wählen nun k so klein, dass $1 - ak \geq -1$ oder $k \leq 2/a$ gilt. Dann gilt

$$|U^n| = |(1 - ak)^n v| \leq |v| \quad \text{für } n \geq 0,$$

sodass die numerische Lösung auch stabil ist. Beachten Sie, dass die Forderung $ak \leq 2$ bedeutet, dass für große a der Zeitschritt $k \leq 2/a$ gewählt werden muss. Wenn k groß ist, dann wächst U^n mit n, im Gegensatz zum Verhalten der exakten Lösung der Differentialgleichung. Es gilt

$$\begin{aligned} U^n - u(t_n) &= (1 - ak)^n v - (\mathrm{e}^{-ak})^n v \\ &= \big((1 - ak) - \mathrm{e}^{-ak}\big) \sum_{j=0}^{n-1} (1 - ak)^j \mathrm{e}^{-(n-1-j)ak} v. \end{aligned}$$

Nach der MacLaurinschen Formel findet man leicht

$$|1 - x - \mathrm{e}^{-x}| \leq \tfrac{1}{2}x^2 \quad \text{für } x \geq 0$$

und somit

$$|U^n - u(t_n)| \leq \tfrac{1}{2}a^2k^2 \sum_{j=0}^{n-1} |v| = \tfrac{1}{2}nka^2k|v| = (\tfrac{1}{2}t_n a^2)k|v| = C(t_n, a)k|v|,$$

sodass der Fehler für $k \to 0$ auf jedem endlichen Zeitintervall in $O(k)$ ist.

Da das vorhergehende Resultat nur unter der Stabilitätsbedingung $ak \leq 2$ gültig ist, werden wir nun ein alternatives Verfahren betrachten, das diesen

Nachteil nicht besitzt, nämlich das *Rückwärts-Euler-Verfahren* bei der der Differenzenquotient in die Rückwärtsrichtung genommen wird, sodass U^n durch

$$\frac{U^n - U^{n-1}}{k} + aU^n = 0 \quad \text{für } n \geq 1 \quad \text{mit } U^0 = v$$

definiert ist. Diesmal gilt

$$U^n = \frac{1}{1+ak} U^{n-1} = \frac{1}{(1+ak)^n} v,$$

und im Falle $a \geq 0$ gilt die Stabilitätsschranke $|U^n| \leq |v|$ für $n \geq 0$, unabhängig von der Größe von k und a. Nun gilt

$$U^n - u(t_n) = \Big(\frac{1}{1+ak} - \mathrm{e}^{-ak}\Big) \sum_{j=0}^{n-1} \frac{1}{(1+ak)^j} \mathrm{e}^{-(n-1-j)ak} v. \tag{7.15}$$

Hier ist

$$\Big|\frac{1}{1+x} - \mathrm{e}^{-x}\Big| \leq 2x^2 \quad \text{für } x \geq 0, \tag{7.16}$$

sodass nun ohne jede Einschränkung für k

$$|U^n - u(t_n)| \leq 2t_n a^2 k\,|v| = C(t_n, a)k|v|$$

gilt.

Aus numerischen Gründen wäre es wünschenswert, in der Fehlerschranke eine höhere Potenz von k als die erste zu erhalten. Dieser Wunsch motiviert das *Crank-Nicolson-Verfahren*,

$$\frac{U^n - U^{n-1}}{k} + a\frac{U^n + U^{n-1}}{2} = 0 \quad \text{für } n \geq 1 \quad \text{mit } U^0 = v,$$

woraus sich

$$U^n = \frac{1 - \frac{1}{2}ak}{1 + \frac{1}{2}ak} U^{n-1} = \Big(\frac{1 - \frac{1}{2}ak}{1 + \frac{1}{2}ak}\Big)^n v$$

ergibt. Hier gilt für jedes k und n die Stabilitätseigenschaft $|U^n| \leq |v|$ für $n \geq 0$. Wegen

$$\Big|\frac{1 - \frac{1}{2}x}{1 + \frac{1}{2}x} - \mathrm{e}^{-x}\Big| \leq x^3 \quad \text{für } x \geq 0$$

gilt

$$\begin{aligned} |U^n - u(t_n)| &= \Big|\Big(\frac{1 - \frac{1}{2}ak}{1 + \frac{1}{2}ak} - \mathrm{e}^{-ak}\Big) \sum_{j=0}^{n-1} \Big(\frac{1 - \frac{1}{2}ak}{1 + \frac{1}{2}ak}\Big)^j \mathrm{e}^{-(n-1-j)ak} v\Big| \\ &\leq a^3k^3 \sum_{j=0}^{n-1} |v| = t_n a^3 k^2 |v| = C(t_n, a)k^2|v|. \end{aligned}$$

Der Fehler geht nun mit $O(k^2)$ anstatt mit $O(k)$ gegen null.

Bei allen obigen Fehlerabschätzungen wachsen die Konstanten auf der rechten Seite mit a. Wir werden nun zeigen, dass bei Verwendung des Rückwärts-Euler-Verfahrens eine Fehlerschranke aufgestellt werden kann, die unabhängig von a ist. Dies ist günstig, wenn a sehr groß werden kann. Wir werden nun

$$|U^n - u(t_n)| \le C t_n^{-1} k |v| \tag{7.17}$$

zeigen, wobei C unabhängig von a und t_n ist. Für festes, positives $t_n = t$ zeigt dies die gleichmäßige Konvergenz von der Ordnung $O(k)$ in a. Zum Beweis von (7.17) betrachten wir zunächst $ak \ge 1$. Dann ist

$$|U^n| = \frac{1}{(1+ak)^n}|v| \le 2^{-n}|v|. \tag{7.18}$$

Für ein geeignetes C_1 gilt aber

$$2^{-n} \le C_1/n = C_1 t_n^{-1} k.$$

Zudem ist

$$|u(t_n)| = \mathrm{e}^{-nak}|v| \le \mathrm{e}^{-n}|v| \le C_2 n^{-1}|v| = C_2 t_n^{-1} k|v|,$$

sodass (7.17) wegen der Dreiecksungleichung erfüllt ist.

Um den Fall $ak \le 1$ zu behandeln, stellen wir fest, dass für ein geeignetes γ mit $0 < \gamma \le 1$

$$\frac{1}{1+x} \le \mathrm{e}^{-\gamma x} \quad \text{für } 0 \le x \le 1$$

gilt, sodass

$$\frac{1}{(1+ak)^j} \le \mathrm{e}^{-\gamma jak}$$

ist. Somit erhalten wir unter Verwendung von (7.15) und (7.16)

$$\begin{aligned} |U^n - u(t_n)| &\le 2a^2k^2 \sum_{j=0}^{n-1} \mathrm{e}^{-\gamma jak}\mathrm{e}^{-\gamma(n-1-j)ak} = 2a^2k^2 n \mathrm{e}^{-\gamma(n-1)ak} \\ &\le 2\mathrm{e}^{\gamma} a^2 t_n \mathrm{e}^{-t_n\gamma a} k \le C_3 t_n^{-1} k \quad \text{mit } C_3 = 2\ \mathrm{e}^{\gamma} \sup_{x\ge 0} x^2 \mathrm{e}^{-\gamma x}. \end{aligned}$$

Insgesamt vervollständigt diese Abschätzung den Beweis von (7.17).

Diese Eigenschaft trifft auf das Crank-Nicolson-Verfahren nicht zu, da das Analogon von (7.18) nicht gilt, weil $|(1 - \frac{1}{2}ak)/(1 + \frac{1}{2}ak)|$ für ak gegen unendlich gegen eins geht.

Die gerade beschriebene strenge Stabilitätseigenschaft für das Rückwärts-Euler-Verfahren ist nützlich, wenn Systeme der Form

$$u' + Au = f(t) \quad \text{für } t > 0$$

behandelt werden, bei denen A eine symmetrische, positiv definite $N \times N$-Matrix ist, die nicht gut konditioniert ist, d. h. die ein großes Verhältnis zwischen dem größten und kleinsten Eigenwert hat. Ein solches System wird als *steifes* gewöhnliches Differentialgleichungssystem bezeichnet. Das Rückwärts-Euler-Verfahren ist in diesem Fall durch die Gleichungen

$$(I + kA)U^n = U^{n-1} + kf(t_n) \quad \text{für } n \geq 1 \quad \text{mit } U^0 = v \tag{7.19}$$

definiert, und wir müssen somit in jedem Zeitschritt ein Gleichungssystem lösen. Wir bezeichnen dieses Verfahren deshalb als *implizit*. Dies steht im Gegensatz zu dem *expliziten* Vorwärts-Euler-Verfahren

$$U^n = (I - kA)U^{n-1} + kf(t_{n-1}) \quad \text{für } n \geq 1 \quad \text{mit } U^0 = v, \tag{7.20}$$

das allerdings die bereits beschriebenen Nachteile hinsichtlich der Stabilität besitzt.

Das System (7.19) kann in der Form

$$U^n = (I + kA)^{-1}U^{n-1} + k(I + kA)^{-1}f(t_n) \quad \text{für } n \geq 1$$

geschrieben werden und wir stellen fest, dass im Falle einer symmetrischen, positiv definiten Matrix A

$$|(I + kA)^{-1}| = \max_j \frac{1}{1 + k\lambda_j} = \frac{1}{1 + k\lambda_1} < 1 \quad \text{für jedes } k > 0 \tag{7.21}$$

gilt, wobei λ_1 der kleinste Eigenwert von A ist.

Somit gilt für das homogene System, d. h. im Falle $f = 0$, $|U^n| = |(I + kA)^{-n}v| \to 0$ für $n \to \infty$. Die numerische Lösung reproduziert also das asymptotische Verhalten der Differentialgleichungen. Andererseits gilt für die in (7.20) auftretende Matrix

$$|I - kA| = \max_j |1 - k\lambda_j|,$$

was nur dann kleiner eins ist, wenn $1 - k\lambda_N > -1$ gilt, d. h. wenn $k < 2/\lambda_N$ wäre. Dies könnte eine sehr einschränkende Bedingung sein, da λ_N der größte Eigenwert von A ist.

Im Falle eines Gleichungssystems kann das Crank-Nicolson-Verfahren in der Form

$$(I + \tfrac{1}{2}kA)U^n = (I - \tfrac{1}{2}kA)U^{n-1} + kf(t_{n-1/2})$$

mit $t_{n-1/2} = (n - 1/2)k$ oder in der Form

$$U^n = (I + \tfrac{1}{2}kA)^{-1}(I - \tfrac{1}{2}kA)U^{n-1} + k(I + \tfrac{1}{2}kA)^{-1}f(t_{n-1/2})$$

geschrieben werden. Es ist also ein implizites Verfahren. Hier ist für alle k

$$|(I + \tfrac{1}{2}kA)^{-1}(I - \tfrac{1}{2}kA)| = \max_j \left| \frac{1 - \frac{1}{2}k\lambda_j}{1 + \frac{1}{2}k\lambda_j} \right| < 1.$$

Wie bereits erwähnt, geht die Norm mit einem festen k für $\lambda_{\max} \to \infty$ gegen eins, was weniger zufriedenstellend als (7.21) ist.

Wir schließen mit einer kurzen Diskussion numerischer Methoden für das Anfangswertproblem der skalaren Gleichung (7.11) zweiter Ordnung. Wir ersetzen die zweite Ableitung zunächst durch einen symmetrischen Differenzenquotienten und erhalten

$$\frac{U^{n+1} - 2U^n + U^{n-1}}{k^2} + aU^n = 0 \quad \text{für } n \geq 1$$

und verwenden beispielsweise die Anfangsbedingungen

$$U^0 = v \quad \frac{U^1 - U^0}{k} = w.$$

Die Differenzengleichung kann auch als

$$U^{n+1} - 2\mu U^n + U^{n-1} = 0 \quad \text{mit } \mu = 1 - ak^2/2 \tag{7.22}$$

geschrieben werden. Sie besitzt die charakteristische Gleichung

$$\tau^2 - 2\mu\tau + 1 = 0.$$

Wenn deren Wurzeln $\tau_{1,2}$ verschieden sind, dann ist die Lösung von (7.22) von der Form

$$U^n = c_1\tau_1^n + c_2\tau_2^n, \tag{7.23}$$

wobei c_1 und c_2 durch die Anfangsbedingungen bestimmt sind. Für $|\mu| < 1$, d. h. für $ak^2 < 4$, sind die Wurzeln verschieden und es gilt $|\tau_1| = |\tau_2| = 1$, sodass die Stabilität für

$$|U^n| \leq C(|v| + |w|) \quad \text{für } n \geq 0$$

gegeben ist. Ist dagegen $|\mu| > 1$ oder $ak^2 > 4$, dann gilt $|\tau_1| > 1$ und $|\tau_2| < 1$. In diesem Fall wächst die allgemeine Lösung von (7.22) exponentiell. Wenn wir stattdessen das implizite Verfahren

$$\frac{U^{n+1} - 2U^n + U^{n-1}}{k^2} + a\,U^{n+1} = 0 \quad \text{für } n \geq 1$$

betrachten, dann wird die charakteristische Gleichung mit $\nu = 1 + ak^2$ zu

$$\nu\tau^2 - 2\tau + 1 = 0.$$

Die Wurzeln sind nun für jede Wahl von k und a betragsmäßig kleiner als eins und die Stabilität ist gesichert. Dieses Verfahren ist jedoch wegen der

fehlenden Symmetrie der Differenzenapproximation in der Genauigkeit nur von erster Ordnung. Das Verfahren

$$\frac{U^{n+1}-2U^n+U^{n-1}}{k^2}+a\left(\frac{1}{4}U^{n+1}+\frac{1}{2}U^n+\frac{1}{4}U^{n-1}\right)U^{n+1}=0 \quad \text{für } n\geq 1$$

ist von der Genauigkeit her zweiter Ordnung und für jedes k und a stabil, weil die charakteristische Gleichung

$$\tau^2-2\kappa\tau+1=0 \quad \text{mit } \kappa=(1-\tfrac{1}{4}ak^2)/(1+\tfrac{1}{4}ak^2)$$

verschiedene Wurzeln mit $|\tau_1|=|\tau_2|=1$ besitzt.

7.3 Problemstellungen

Problem 7.1. Lösen Sie das Anfangswertproblem

$$u'(t)=\begin{bmatrix}1&2\\2&1\end{bmatrix}u(t) \quad \text{für } t>0 \quad \text{mit } u(0)=\begin{bmatrix}1\\2\end{bmatrix}.$$

Problem 7.2. (Übung am Rechner.) Bestimmen Sie für Problemstellung 7.1 eine approximative Lösung an der Stelle $t=1$ mithilfe des Vorwärts- und Rückwärts-Euler-Verfahrens sowie des Crank-Nicolson-Verfahrens für $k=1/10,\ 1/100$. Vergleichen Sie diese mit der exakten Lösung.

Problem 7.3. Lösen Sie das Anfangswertproblem

$$u'(t)=\begin{bmatrix}1&2t\\2t&1\end{bmatrix}u(t) \quad \text{für } t>0 \quad \text{mit } u(0)=\begin{bmatrix}1\\0\end{bmatrix}.$$

Problem 7.4. (Picard-Methode.) Beweisen Sie die Existenz der Lösungen von (7.9) folgendermaßen: Zeigen Sie zunächst, dass eine Lösung des Anfangswertproblems (7.9) die Integralgleichung

$$u(t)=v+\int_0^t f(s,u(s))\,\mathrm{d}s =: T(u)(t)$$

erfüllt, und dass umgekehrt eine Lösung dieser Integralgleichung eine Lösung von (7.9) ist. Angenommen, die stetige Funktion $f(t,u)$ erfüllt eine globale Lipschitz-Bedingung in der zweiten Variable, d. h.

$$|f(t,v)-f(t,w)|\leq K|v-w| \quad \forall v,w\in\mathbf{R},\ 0\leq t\leq a.$$

Zeigen Sie, dass die durch

$$u_0(t)=v, \quad u_{n+1}(t)=T(u_n)(t) \quad \text{für } n\geq 0,$$

definierte Folge u_n, $n=0,1,\ldots$ die Gleichung

$$|u_{n+1}(t) - u_n(t)| \le CK^n a^{n+1}/(n+1)!$$

erfüllt, dass daraus die gleichmäßige Konvergenz von $\sum_{n=0}^{\infty}(u_{n+1}(t) - u_n(t))$ gegen $u(t) - v$ für $0 \le t \le a$ folgt und dass $u \in C([0,a])$ sowie $u = T(u)$ gilt. Insbesondere folgt daraus, dass u die Gleichung (7.9) erfüllt. Zeigen Sie schließlich, dass $f(t,u)$ eine Lipschitz-Bedingung in der zweiten Variable erfüllt, wenn $\partial f/\partial u$ beschränkt ist.

Problem 7.5. Beweisen Sie ein Eindeutigkeitsresultat für (7.9), wenn $f(t,u)$ eine stetige Funktion ist, die eine Lipschitz-Bedingung in der zweiten Variable erfüllt (siehe Problemstellung 7.4). Hinweis: Nehmen Sie an, dass u_1 und u_2 zwei Lösungen sind, die beide die Integralgleichung aus Problemstellung 7.4 erfüllen. Verwenden Sie die Lipschitz-Bedingung, um eine Ungleichung herzuleiten, die $u_1 - u_2 = 0$ beweist.

Problem 7.6. (a)(Gronwall-Lemma.) Angenommen, φ ist eine nichtnegative, stetige Funktion, für die

$$\varphi(t) \le a + b\int_0^t \varphi(s)\,\mathrm{d}s \quad \text{für } t > 0$$

mit nichtnegativen Konstanten a und b gilt. Beweisen Sie

$$\varphi(t) \le a\,\mathrm{e}^{bt} \quad \text{für } t > 0.$$

(b) Verwenden Sie das Gronwall-Lemma um zu zeigen, dass die Lösung von (7.3) die Gleichung

$$|u(t)| \le \mathrm{e}^{|A|T}\Big(|v| + \int_0^T |f(s)|\,\mathrm{d}s\Big) \quad \text{für } 0 \le t \le T$$

erfüllt. Zeigen Sie, dass sich daraus die Eindeutigkeit der Lösung ergibt.

Problem 7.7. Zeigen Sie, dass (7.13) und (7.14) äquivalent sind.

Problem 7.8. Beweisen Sie, dass die allgemeine Lösung von (7.22) im Falle $\tau_1 \ne \tau_2$ (7.23) ist. Zeigen Sie auch

$$|\mu| < 1 \Longrightarrow |\tau_1| = |\tau_2| = 1,$$
$$|\mu| > 1 \Longrightarrow |\tau_1| < 1,\ |\tau_2| > 1.$$

Wie sieht die allgemeine Form der Lösung im Falle $\tau_1 = \tau_2$ aus?

8

Parabolische Gleichungen

In diesem Kapitel untersuchen wir sowohl das reine Anfangswertproblem als auch das gemischte Anfangs-Randwertproblem für die Wärmeleitungsgleichung, wobei wir Fourier-Methoden und Energieargumente benutzen. In Abschnitt 8.1 analysieren wir die Lösung des reinen Anfangswertproblems für die homogene Wärmeleitungsgleichung mithilfe einer Darstellung in Form eines Gauß-Kerns und verwenden sie, um Eigenschaften der Lösung zu untersuchen. Im verbleibenden Teil des Abschnittes betrachten wir das Anfangswertproblem in einem begrenzten räumlichen Gebiet. In Abschnitt 8.2 lösen wir die homogene Gleichung durch Entwicklung nach Eigenfunktionen und wenden das Duhamel-Prinzip an, um eine Lösung der inhomogenen Gleichung zu bestimmen. In Abschnitt 8.3 führen wir die Variationsformulierung des Problems ein und geben Beispiele für die Verwendung von Energieargumenten an. In Abschnitt 8.4 beweisen wir das Maximumprinzip und wenden es an.

8.1 Das reine Anfangswertproblem

Wir beginnen unsere Untersuchung parabolischer Gleichungen mit der Betrachtung des reinen Anfangswertproblems (oder Cauchy-Problems) für die Wärmeleitungsgleichung, bei der ein $u(x,t)$ bestimmt werden soll, das die Gleichung

$$\begin{aligned} &\frac{\partial u}{\partial t} - \Delta u = 0 && \text{in } \mathbf{R}^d \times \mathbf{R}_+, \\ &u(\cdot,0) = v && \text{in } \mathbf{R}^d \end{aligned} \tag{8.1}$$

erfüllt. Wir werden die Fourier-Transformation von u bezüglich x anwenden

$$\hat{u}(\xi,t) = \mathcal{F}u(\cdot,t)(\xi) = \int_{\mathbf{R}^d} u(x,t)\mathrm{e}^{-\mathrm{i}x\cdot\xi}\,\mathrm{d}x \quad \text{für } \xi \in \mathbf{R}^d$$

(siehe Anhang A.3). Wenn u und seine Ableitungen für große $|x|$ hinreichend klein sind, gilt

$$\big(\mathcal{F}\Delta u(\cdot,t)\big)(\xi) = \int_{\mathbf{R}^d} \Delta u(x,t)\mathrm{e}^{-\mathrm{i}x\cdot\xi}\,\mathrm{d}x = -|\xi|^2\hat{u}(\xi,t)$$

und mit $u_t = \partial u/\partial t$

$$\big(\mathcal{F}u_t(\cdot,t)\big)(\xi) = \frac{d\hat{u}}{dt}(\xi,t).$$

Wir schlussfolgern somit aus (8.1), dass $\hat{u}$

$$\begin{aligned} \frac{d\hat{u}}{dt} &= -|\xi|^2\hat{u} \qquad \text{für } \xi \in \mathbf{R}^d,\ t>0,\\ \hat{u}(\xi,0) &= \hat{v}(\xi) \qquad \text{für } \xi \in \mathbf{R}^d \end{aligned}$$

erfüllt. Hierbei handelt es sich um ein einfaches Anfangswertproblem für eine lineare Differentialgleichung erster Ordnung mit dem Parameter ξ. Die Lösung ist

$$\hat{u}(\xi,t) = \hat{v}(\xi)\mathrm{e}^{-t|\xi|^2}. \tag{8.2}$$

Es sei daran erinnert, dass $w(x) = \mathrm{e}^{-|x|^2}$ die Fourier-Transformierte

$$\hat{w}(\xi) = \pi^{d/2}\mathrm{e}^{-|\xi|^2/4}$$

besitzt (siehe Problemstellung A.19). So können wir aus (A.34) schlussfolgern, dass $\mathrm{e}^{-t|\xi|^2}$ die Fourier-Transformierte des *Gauß-Kerns*

$$U(x,t) = (4\pi t)^{-d/2}\mathrm{e}^{-|x|^2/4t}$$

ist. Aus (8.2) erhalten wir also formal

$$u(x,t) = \big(U(\cdot,t) * v\big)(x) = (4\pi t)^{-d/2}\int_{\mathbf{R}^d} v(y)\mathrm{e}^{-|x-y|^2/4t}\,\mathrm{d}y. \tag{8.3}$$

Die Funktion $U(x,t)$ ist eine *Fundamentallösung* des Anfangswertproblems. Wir werden nun zeigen, dass die durch (8.3) definierte Funktion unter einer schwachen Annahme bezüglich der Anfangsfunktion tatsächlich eine Lösung von (8.1) ist. Beachten Sie, dass $U(x,t)$ und $u(x,t)$ in (8.3) nur für $t>0$ definiert sind.

Theorem 8.1. *Wenn v eine beschränkte stetige Funktion auf $\mathbf{R}^d$ ist, dann ist die durch* (8.3) *definierte Funktion $u(x,t)$ eine Lösung der Wärmeleitungsgleichung für $t>0$ und geht für t gegen null gegen die Anfangsdaten v.*

Beweis. Wir bemerken zunächst, dass wir für $t>0$ in (8.3) unter dem Integralzeichen bezüglich x und t differenzieren können und zeigen direkt, dass diese Funktion die Wärmeleitungsgleichung in (8.1) erfüllt. Um uns davon zu überzeugen, dass $u(x,t)$ für $t\to 0$ gegen die gewünschten Anfangswerte konvergiert, wählen wir ein beliebiges $x_0 \in \mathbf{R}^d$ und zeigen

$$u(x,t) \to v(x_0) \quad \text{für} \quad (x,t) \to (x_0,0).$$

Tatsächlich können wir unter Verwendung der Transformation $\eta = (y-x)/\sqrt{4t}$ und der Formel

$$\pi^{-d/2} \int_{\mathbf{R}^d} \mathrm{e}^{-|x|^2}\, \mathrm{d}x = 1 \tag{8.4}$$

die Gleichung

$$\begin{aligned} u(x,t) - v(x_0) &= (4\pi t)^{-d/2} \int_{\mathbf{R}^d} v(y) \mathrm{e}^{-|x-y|^2/4t}\, \mathrm{d}y - v(x_0) \\ &= \pi^{-d/2} \int_{\mathbf{R}^d} \big(v(x+\sqrt{4t}\eta) - v(x_0)\big)\, \mathrm{e}^{-|\eta|^2}\, \mathrm{d}\eta \end{aligned}$$

aufschreiben. Sei $M = \|v\|_{\mathcal{C}} = \|v\|_{\mathcal{C}(\mathbf{R}^d)}$ und sei δ so klein, dass

$$|v(z) - v(x_0)| < \epsilon \tag{8.5}$$

gilt, wenn $|z - x_0| < \delta$ ist. Für ein beliebiges $\omega > 0$ gilt

$$\begin{aligned} |u(x,t) - v(x_0)| &\le 2M\pi^{-d/2} \int_{|y|>\omega} \mathrm{e}^{-|y|^2}\, \mathrm{d}y \\ &\quad + \pi^{-d/2} \int_{|y|<\omega} \big|v(x+\sqrt{4t}y) - v(x_0)\big|\, \mathrm{e}^{-|y|^2}\, \mathrm{d}y = I + II. \end{aligned}$$

Nun wählen wir ω so groß, dass $I < \epsilon$ gilt, was unter Berücksichtigung von (8.4) möglich ist. Dann erhalten wir mit festem ω sowie (8.5) und (8.4)

$$II \le \sup_{|y|<\omega} |v(x+\sqrt{4t}y) - v(x_0)| < \epsilon,$$

wenn $|x - x_0| + \sqrt{4t}\omega < \delta$ ist. Folglich gilt für diese x, t

$$|u(x,t) - v(x_0)| < 2\epsilon,$$

was den Beweis vervollständigt. □

Theorem 8.1 zeigt also, dass das Anfangswertproblem (8.1) eine Lösung zulässt und ist deshalb ein Existenzsatz. Wir werden zeigen, dass diese Lösung stetig von den Anfangsdaten v abhängt.

Wir schreiben (8.3) in der Form

$$u(x,t) = (E(t)v)(x) = (4\pi t)^{-d/2} \int_{\mathbf{R}^d} v(y) \mathrm{e}^{-|x-y|^2/4t}\, \mathrm{d}y, \tag{8.6}$$

wobei durch $E(t)$ ein linearer Operator, der Lösungsoperator von (8.1), definiert wird, der die gegebenen Anfangsdaten in die Lösung zur Zeit t überführt.

Beachten Sie, dass wegen (8.4)

$$|u(x,t)| \le (4\pi t)^{-d/2} \int_{\mathbf{R}^d} \mathrm{e}^{-|x-y|^2/4t}\,\mathrm{d}y\,\|v\|_{\mathcal{C}} = \|v\|_{\mathcal{C}}$$

gilt, sodass

$$\|u(\cdot,t)\|_{\mathcal{C}} \le \|v\|_{\mathcal{C}} \quad \text{für } t > 0$$

ist. Dies zeigt, dass der Operator $E(t)$ hinsichtlich der Maximumnorm durch 1 beschränkt ist. Dies ist der erste Teil des folgenden Resultates.

Theorem 8.2. *Der durch* (8.6) *definierte Lösungsoperator $E(t)$ ist in $\mathcal{C}$ beschränkt, und es gilt*

$$\|E(t)v\|_{\mathcal{C}} \le \|v\|_{\mathcal{C}} \quad \textit{für } t \ge 0. \tag{8.7}$$

Wenn v_1 und v_2 zwei beschränkte, stetige Funktionen auf $\mathbf{R}^d$ und u_1 und u_2 die zugehörigen Lösungen des Anfangswertproblems (8.1) *sind, dann gilt*

$$\|u_1(t) - u_2(t)\|_{\mathcal{C}} \le \|v_1 - v_2\|_{\mathcal{C}} \quad \textit{für } t \ge 0. \tag{8.8}$$

Beweis. Wir müssen lediglich den zweiten Teil des Theorems zeigen. Da $E(t)$ aber ein linearer Operator ist, gilt

$$u_1(t) - u_2(t) = E(t)v_1 - E(t)v_2 = E(t)(v_1 - v_2),$$

und somit folgt (8.8) sofort aus (8.7). □

Unter Verwendung eines Maximumprinzips werden wir in Abschnitt 8.4 das zugehörige Eindeutigkeitsresultat beweisen, d. h. die Existenz mindestens einer beschränkten Lösung von (8.1). Somit ist (8.3) also die einzige Lösung.

Aufgrund der Existenz, der Eindeutigkeit und der stetigen Abhängigkeit der Lösung von den Anfangsdaten ist das Problem (8.1) *gut gestellt*. Insbesondere ist die Eigenschaft der stetigen Abhängigkeit bei Anwendungen wichtig. Sie zeigt, dass eine kleine Veränderung der Daten des Problems nur einen kleinen Effekt für die Lösung bewirkt.

Nicht alle Probleme, die Lösungen zulassen, besitzen die Eigenschaft stetiger Abhängigkeit. Betrachten wir beispielsweise das Anfangswertproblem

$$\begin{aligned} &u_t + u_{xx} = 0 && \text{in } \mathbf{R} \times \mathbf{R}_+, \\ &u(x,0) = v_n(x) = n^{-1}\sin(nx) && \text{für } x \in \mathbf{R}, \end{aligned} \tag{8.9}$$

mit der Lösung

$$u_n(x,t) = n^{-1}\mathrm{e}^{n^2 t}\,\sin(nx).$$

Hier gilt

$$\|v_n\|_{\mathcal{C}} = n^{-1} \to 0 \quad \text{für } n \to \infty,$$

wobei für jedes $t > 0$

$$\|u_n(t)\|_{\mathcal{C}} = n^{-1}\mathrm{e}^{n^2 t} \to \infty \quad \text{für } n \to \infty$$

gilt. Somit ist die Lösung für $t > 0$ nicht nahe null, auch wenn der Anfangswert v_n nahe null liegt.

Die Differentialgleichung in (8.9) ist die Wärmeleitungsgleichung mit umgekehrten Vorzeichen vor der Zeitableitung, sie wird deshalb als *Rückwärts-Wärmeleitungsgleichung* bezeichnet. Dem obigen Resultat entnehmen wir, dass das Problem, eine frühere Wärmeverteilung in einem Körper aus der aktuellen Verteilung zu bestimmen, *schlecht gestellt* ist.

Wir haben oben bereits erwähnt, dass die Darstellung von $u(x,t)$ als Funktion von v in (8.3) die Differentiation nach x und t unter dem Integralzeichen für $t > 0$ erlaubt, sogar ohne die Regularität von v anzunehmen. Diese Differentiation kann beliebig oft ausgeführt werden, sodass u unendlich oft differenzierbar ist, d. h. es gilt $u \in \mathcal{C}^\infty$ für $t > 0$. Unter Verwendung der Multiindex-Notation aus (1.8) findet man leicht

$$\begin{aligned} |D_t^j D^\alpha U(x,t)| &\le t^{-j-|\alpha|/2-d/2} P(|x|/\sqrt{4t}) \mathrm{e}^{-|x|^2/4t} \\ &\le C t^{-j-|\alpha|/2-d/2} \mathrm{e}^{-|x|^2/8t}, \end{aligned}$$

wobei $P(y)$ ein Polynom in y ist und wir die Tatsache ausgenutzt haben, dass es für jedes Polynom P eine Konstante C gibt, für die die Gleichung

$$|P(y)\mathrm{e}^{-y^2}| \le C\mathrm{e}^{-y^2/2} \quad \text{für } y > 0$$

gilt. Somit ist

$$\begin{aligned} \sup_{x\in\mathbf{R}^d} |D_t^j D^\alpha u(x,t)| &\le C t^{-j-|\alpha|/2-d/2} \sup_{x\in\mathbf{R}^d} \int_{\mathbf{R}^d} |v(y)| \mathrm{e}^{-|x-y|^2/8t}\, \mathrm{d}y \\ &\le C t^{-j-|\alpha|/2} \sup_{y\in\mathbf{R}^d} |v(y)| \end{aligned}$$

oder

$$\|D_t^j D^\alpha E(t)v\|_{\mathcal{C}} \le C t^{-j-|\alpha|/2} \|v\|_{\mathcal{C}} \quad \text{für } t > 0,$$

was zeigt, dass der Operator $E(t)$ eine *Glättungseigenschaft* besitzt: Die Lösung von (8.1) ist für $t > 0$ auch dann glatt, wenn v nicht glatt ist. Die Schranken für die Ableitungen wachsen jedoch, wenn t gegen null geht.

Im Falle glatter Anfangsdaten sind die Ableitungen der Lösung bis $t = 0$ gleichmäßig beschränkt: Nach dem Variablenwechsel $z = x - y$ erhalten wir aus (8.6)

$$\begin{aligned} \big(D^\alpha E(t)v\big)(x) = D_x^\alpha u(x,t) &= (4\pi t)^{-d/2} D_x^\alpha \int_{\mathbf{R}^d} v(x-z)\mathrm{e}^{-|z|^2/4t}\, dz \\ &= (4\pi t)^{-d/2} \int_{\mathbf{R}^d} D_x^\alpha v(x-z)\mathrm{e}^{-|z|^2/4t}\, dz = (E(t)D^\alpha v)(x). \end{aligned}$$

Somit gilt wegen (8.1) und (8.7)

$$\|D_t^j D^\alpha E(t)v\|_{\mathcal{C}} = \|\Delta^j D^\alpha E(t)v\|_{\mathcal{C}} = \|E(t)\Delta^j D^\alpha v\|_{\mathcal{C}} \le \|\Delta^j D^\alpha v\|_{\mathcal{C}}.$$

Man kann zeigen, dass die Lösung des Anfangswertproblems für die inhomogene Wärmeleitungsgleichung

$$\begin{aligned} u_t - \Delta u &= f \qquad \text{in } \mathbf{R}^d \times \mathbf{R}_+, \\ u(\cdot, 0) &= v \qquad \text{in } \mathbf{R}^d \end{aligned}$$

mit gegebenem $f = f(x,t)$ in der Form

$$\begin{aligned} u(x,t) &= \int_{\mathbf{R}^d} v(y) U(x-y,t)\,\mathrm{d}y + \int_0^t \int_{\mathbf{R}^d} f(y,s) U(x-y,t-s)\,\mathrm{d}y\,\mathrm{d}s \\ &= E(t)v + \int_0^t E(t-s) f(\cdot,s)\,\mathrm{d}s \end{aligned}$$

dargestellt werden kann, wenn beispielsweise v, f und ∇f stetig und beschränkt sind.

8.2 Die Lösung des Anfangs-Randwertproblems durch Entwicklung nach Eigenfunktionen

Wir betrachten zunächst das gemischte Anfangs-Randwertproblem für die homogene Wärmeleitungsgleichung: Gesucht ist eine Funktion $u(x,t)$, für die

$$\begin{aligned} u_t - \Delta u &= 0 \qquad \text{in } \Omega \times \mathbf{R}_+, \\ u &= 0 \qquad \text{auf } \Gamma \times \mathbf{R}_+, \\ u(\cdot, 0) &= v \qquad \text{in } \Omega \end{aligned} \tag{8.10}$$

erfüllt ist. Dabei ist Ω ein beschränktes Gebiet in $\mathbf{R}^d$ mit dem glatten Rand Γ, $u_t = \partial u/\partial t$ und v eine gegebene Funktion in $L_2 = L_2(\Omega)$. Wir werden dieses Problem nun durch Entwicklung nach Eigenfunktionen lösen. Wir bezeichnen mit $(\cdot,\cdot)$ und $\|\cdot\|$ das Skalarprodukt beziehungsweise die Norm in $L_2 = L_2(\Omega)$.

Wir wiederholen aus Kapitel 6, dass eine Orthonormalbasis $\{\varphi_i\}_{i=1}^\infty$ in L_2 von glatten Eigenfunktionen φ_i und zugehörigen Eigenwerten $\{\lambda_i\}_{i=1}^\infty$ existiert, die

$$-\Delta\varphi_i = \lambda_i \varphi_i \quad \text{in } \Omega \quad \text{mit } \varphi_i = 0 \quad \text{auf } \Gamma \tag{8.11}$$

erfüllen. In unserer üblichen Notation ist dies äquivalent zu

$$a(\varphi_i, v) = \int_\Omega \nabla\varphi_i \cdot \nabla v\,\mathrm{d}x = \lambda_i(\varphi_i, v) \quad \forall v \in H_0^1.$$

Es sei daran erinnert, dass $0 < \lambda_1 < \lambda_2 \le \cdots \le \lambda_i \le \ldots$ ist, dass $\lambda_i \to \infty$ für $i \to \infty$ gilt und dass

$$a(\varphi_i, \varphi_j) = \lambda_i \delta_{ij}$$

mit dem Kronecker-Symbol δ_{ij} ist, für das $\delta_{ij} = 1$ für $j = i$ und 0 sonst gilt.

Wir suchen nun eine Lösung von (8.10) der Form

$$u(x,t) = \sum_{i=1}^{\infty} \hat{u}_i(t)\varphi_i(x) \tag{8.12}$$

mit den zu bestimmenden Koeffizienten $\hat{u}_i : \mathbf{R}_+ \to \mathbf{R}$. Da es sich hier um eine Summe von Produkten von Funktionen von x und t handelt, wird diese Vorgehensweise auch als Methode der *Trennung der Variablen* bezeichnet. Setzen wir (8.12) in die Differentialgleichung in (8.10) ein und verwenden (8.11), so erhalten wir formal

$$\sum_{i=1}^{\infty} \big(\hat{u}_i'(t) + \lambda_i \hat{u}_i(t)\big)\varphi_i(x) = 0 \quad \text{für } x \in \Omega,\ t \in \mathbf{R}_+$$

und hieraus, weil die φ_i eine Basis bilden,

$$\hat{u}_i'(t) + \lambda_i \hat{u}_i(t) = 0 \quad \text{für } t \in \mathbf{R}_+, \quad i = 1, 2, \ldots,$$

sodass

$$\hat{u}_i(t) = \hat{u}_i(0)\mathrm{e}^{-\lambda_i t}$$

gilt. Darüber hinaus folgt aus der Anfangsbedingung in (8.10)

$$u(\cdot,0) = \sum_{i=1}^{\infty} \hat{u}_i(0)\varphi_i = v = \sum_{i=1}^{\infty} \hat{v}_i\varphi_i, \quad \text{wobei } \hat{v}_i = (v,\varphi_i) = \int_\Omega v\,\varphi_i \,\mathrm{d}x \text{ ist.}$$

Wir sehen also zumindest formal, dass die Lösung von (8.10)

$$u(x,t) = \sum_{i=1}^{\infty} \hat{v}_i \mathrm{e}^{-\lambda_i t}\varphi_i(x) \tag{8.13}$$

sein muss, wobei aufgrund der Parsevalschen Gleichung mit $\|\cdot\| = \|\cdot\|_{L_2}$

$$\|u(\cdot,t)\|^2 = \sum_{i=1}^{\infty} \big(\hat{v}_j^k \mathrm{e}^{-\lambda_j t}\big)^2 \le \mathrm{e}^{-2\lambda_1 t} \sum_{i=1}^{\infty} \hat{v}_i^2 = \mathrm{e}^{-2\lambda_1 t}\|v\|^2 < \infty$$

gilt. Folglich ist für $t \ge 0$ die Funktion $u(\cdot,t) \in L_2$, und deren L_2-Norm fällt exponentiell für $t \to \infty$. Obwohl dieser Abfall in einigen Situationen wichtig ist, werden wir im Folgendem der Einfachheit halber darauf verzichten, das Verhalten von $u(\cdot,t)$ für große t zu verfolgen, und uns mit der Schlussfolgerung zufrieden zu geben, dass

$$\|u(\cdot,t)\| \le \|v\| \quad \text{für } t \in \mathbf{R}_+$$

gilt.

Wir zeigen nun, dass für $t > 0$ die in (8.13) definierte Funktion $u(\cdot, t)$ glatt ist und die Differentialgleichung und die Randbedingung in (8.10) im klassischen Sinne erfüllt, und dass die Anfangsbedingung in dem Sinne erfüllt ist, dass

$$\|u(\cdot, t) - v\| \to 0 \quad \text{für } t \to 0 \tag{8.14}$$

gilt.

Wir stellen zunächst fest, dass für jedes $k \geq 0$ eine Konstante C_k existiert, für die $s^k \mathrm{e}^{-s} \leq C_k$ für $s \geq 0$ gilt. Verwenden wir dies mit $k = 1$, erhalten wir

$$|u(\cdot, t)|_1^2 = \sum_{i=1}^{\infty} \lambda_i \big(\hat{v}_i \mathrm{e}^{-\lambda_i t}\big)^2 = t^{-1} \sum_{i=1}^{\infty} \hat{v}_i^2 (\lambda_i t)\, \mathrm{e}^{-2\lambda_i t} \leq C_1 t^{-1} \|v\|^2,$$

sodass

$$|u(\cdot, t)|_1 \leq C t^{-1/2} \|v\| \quad \text{für } t > 0 \tag{8.15}$$

ist. Also gilt wegen Theorem 6.4 $u(\cdot, t) \in H_0^1$ für $t > 0$, und $u(\cdot, t)$ erfüllt insbesondere die Randbedingung in (8.10). Wenden wir nun $(-\Delta)^k$ auf jeden Term in (8.13) an, erhalten wir wegen $-\Delta \varphi_i = \lambda_i \varphi_i$

$$(-\Delta)^k u(x, t) = \sum_{i=1}^{\infty} \hat{v}_i \lambda_i^k \mathrm{e}^{-\lambda_i t} \varphi_i(x) \tag{8.16}$$

und hieraus für $t > 0$

$$\|\Delta^k u(\cdot, t)\|^2 = \sum_{i=1}^{\infty} \big(\hat{v}_i \lambda_i^k \mathrm{e}^{-\lambda_i t}\big)^2 \leq C_k^2 t^{-2k} \sum_{i=1}^{\infty} \hat{v}_i^2 = C_k^2 t^{-2k} \|v\|^2 < \infty.$$

Auf gleiche Weise wie in (8.15) erhalten wir auch

$$|\Delta^k u(\cdot, t)|_1 \leq C_k t^{-k-1/2} \|v\| < \infty \quad \text{für } t > 0.$$

Folglich gilt für $t > 0$ und für jedes $k \geq 0$ auf Γ die Gleichung $\Delta^k u(\cdot, t) = 0$. Wir können D_t^m auch auf jeden Term in (8.16) anwenden und erhalten wegen $D_t \mathrm{e}^{-\lambda_i t} = -\lambda_i \mathrm{e}^{-\lambda_i t}$

$$|D_t^m \Delta^k u(\cdot, t)|_\delta \leq C t^{-m-k-\delta/2} \|v\| < \infty \quad \text{für } t > 0, \quad \delta = 0, 1.$$

Aus der Theorie der elliptischen Gleichungen wiederholen wir die Regularitätsabschätzung (3.37)

$$\|w\|_s \leq C \|\Delta w\|_{s-2} \quad \forall w \in H^s \cap H_0^1 \quad \text{für } s \geq 2.$$

Durch wiederholte Anwendung dieser Prozedur erhalten wir für $\delta = 0$ oder 1

$$\|w\|_{2k+\delta} \leq C \|\Delta^k w\|_\delta \quad \forall w \in H^{2k+\delta}, \quad \text{wenn } \Delta^j w = 0 \text{ auf } \Gamma \text{ für } j < k \text{ gilt.}$$

Wir kommen schließlich zu dem Schluss, dass für beliebige, nichtnegative ganze Zahlen s und m

$$\|D_t^m u(\cdot,t)\|_s \le Ct^{-m-s/2}\|v\| \quad \text{für } t>0 \tag{8.17}$$

ist. Mit der Sobolev-Ungleichung, Theorem A.5, folgt $D_t^m u(\cdot,t)\in\mathcal{C}^p$ für $t>0$ und beliebiges $p\ge 0$.

Also ist $u(x,t)$ für $t>0$ eine glatte Funktion von x und t, selbst wenn die Anfangsdaten v nicht glatt sind. Deshalb erfüllt $u(\cdot,t)$ die Wärmeleitungsgleichung im klassischen Sinne. Aus obigem folgt auch, dass die Randbedingung erfüllt ist. Schließlich erhalten wir (8.14), indem wir zeigen, dass

$$\|u(\cdot,t)-v\|^2=\sum_{i=1}^{\infty}\left(\mathrm{e}^{-\lambda_i t}-1\right)^2\hat{v}_i^2\to 0 \quad \text{für } t\to 0$$

gilt. Um dies zu beweisen, wählen wir $\epsilon>0$ beliebig klein und N hinreichend groß, sodass $\sum_{i=N+1}^{\infty}\hat{v}_i^2<\epsilon$ ist. Dann gilt

$$\|u(\cdot,t)-v\|^2\le\sum_{i=1}^{N}\left(\mathrm{e}^{-\lambda_i t}-1\right)^2\hat{v}_i^2+\epsilon.$$

Da jeder Term der Summe für $t\to 0$ gegen null geht, schließen wir

$$\|u(\cdot,t)-v\|^2<2\epsilon \quad \text{für hinreichend kleines } t.$$

Wir fassen diese Resultate im folgenden Theorem zusammen.

Theorem 8.3. *Für jedes $v\in L_2$ ist die durch* (8.13) *definierte Funktion $u(x,t)$ eine klassische Lösung der Wärmeleitungsgleichung in* (8.10)*, die für $t>0$ auf Γ verschwindet und die Anfangsbedingung im Sinne von* (8.14) *erfüllt. Darüber hinaus gilt die Glattheitsabschätzung* (8.17).

Da der Faktor t^{-k} auf der rechten Seite von (8.17) im Falle t gegen null gegen unendlich geht, ist die Glattheit der Lösung bis $t=0$ nicht gleichmäßig garantiert. Wenn die Anfangsfunktion glatter ist, dann sind in diesem Zusammenhang bessere Resultate möglich. Es gilt beispielsweise das folgende Resultat in H_0^1.

Theorem 8.4. *Es sei $v\in H_0^1$. Dann erfüllt die in Theorem 8.3 bestimmte Lösung $u(x,t)$ von* (8.10) *die Ungleichung*

$$|u(\cdot,t)|_1\le|v|_1 \quad \textit{für } t\ge 0.$$

Beweis. Wegen (6.4) gilt

$$|u(\cdot,t)|_1^2=\sum_{i=1}^{\infty}\lambda_i\hat{v}_i^2\mathrm{e}^{-2\lambda_i t}\le\sum_{i=1}^{\infty}\lambda_i\hat{v}_i^2=|v|_1^2,$$

was unsere Behauptung beweist. □

Wir weisen darauf hin, dass dieses Resultat nicht nur verlangt, dass die Anfangsdaten in H^1 sind, sondern auch, dass sie auf Γ verschwinden. Das bedeutet, dass die Anfangsdaten mit den Randdaten auf $\Gamma \times \mathbf{R}_+$ verträglich sein müssen, was offensichtlich notwendig ist, damit die Lösung an der Stelle $t = 0$ stetig ist. Für Regularität höherer Ordnung werden weitere Verträglichkeitsbedingungen benötigt.

Wie in Abschnitt 8.1 können wir die Lösung zur Zeit t als Resultat der Wirkung eines Lösungsoperators $E(t)$ auf die Anfangsdaten v betrachten und daher $u(t) = E(t)v$ schreiben. Wegen (8.13) erfüllt dieser Operator die Stabilitätsabschätzung

$$\|E(t)v\| \le \|v\| \quad \text{für } t > 0,$$

und die Abschätzung (8.17) kann in der Form

$$\|D_t^m E(t)v(\cdot,t)\|_s \le Ct^{-m-s/2}\|v\| \quad \text{für } t > 0, \quad m, s \ge 0 \tag{8.18}$$

geschrieben werden, worin sich die Glättungseigenschaft des Lösungsoperators ausdrückt.

Das folgende einfache Beispiel illustriert die obige Lösungsmethode.

Beispiel 8.1. Die Lösung des räumlichen eindimensionalen Problems

$$\begin{aligned} &u_t - u_{xx} = 0 && \text{in } \Omega \times \mathbf{R}_+, \\ &u(0,\cdot) = u(\pi,\cdot) = 0 && \text{in } \mathbf{R}_+, \\ &u(\cdot,0) = v && \text{in } \Omega \end{aligned} \tag{8.19}$$

mit $\Omega = (0,\pi)$ und $v \in L_2(\Omega)$ ist durch

$$u(x,t) = \sum_{j=1}^{\infty} \hat{v}_j \mathrm{e}^{-j^2 t} \sin(jx) \quad \text{mit } \hat{v}_j = \frac{2}{\pi}\int_0^{\pi} v(x)\sin(jx)\,\mathrm{d}x \tag{8.20}$$

gegeben. In diesem Fall reduziert sich das zugehörige Eigenwertproblem (8.11) auf (6.21) mit $b = \pi$. Die Lösung ergibt sich also aus Theorem 8.3 und den in Abschnitt 6.1 erhaltenen Resultaten, abgesehen davon, dass die Eigenfunktionen hier nicht normiert sind. Beachten Sie, dass man den Koeffizienten $\hat{v}_j \mathrm{e}^{-j^2 t}$ der Eigenfunktion $\sin(jx)$ in (8.20) durch Multiplikation des zugehörigen Koeffizienten $\hat{v}_j$ in der Entwicklung der Anfangsfunktion v mit dem Faktor $\mathrm{e}^{-j^2 t}$ erhält. Wenn j groß ist, dann oszilliert $\sin(jx)$ schnell und der Faktor $\mathrm{e}^{-j^2 t}$ wird für wachsendes t sehr klein, sodass die Komponenten der Lösung $u(x,t)$, die zu den Eigenfunktionen $\sin(jx)$ mit großem j gehören, stark gedämpft werden. Das bedeutet auch, dass rasche Schwankungen oder Oszillationen in der Anfangsfunktion v, wie beispielsweise im Falle einer Unstetigkeit (Sprung), mit wachsendem t geglättet werden. Damit handelt es sich also um einen Spezialfall der oben diskutierten Glättungseigenschaft des Lösungsoperators, die typisch für parabolische Probleme ist.

Der oben eingeführte Lösungsoperator $E(t)$ ist für die Untersuchung des Randwertproblems für die inhomogene Gleichung

$$\begin{aligned} &u_t - \Delta u = f && \text{in } \Omega \times \mathbf{R}_+, \\ &u = 0 && \text{auf } \Gamma \times \mathbf{R}_+, \\ &u(\cdot, 0) = v && \text{in } \Omega \end{aligned} \tag{8.21}$$

geeignet. Und zwar kann die Lösung dieses Problems, wie wir sehen werden, in der Form

$$u(t) = E(t)v + \int_0^t E(t-s)f(s)\,\mathrm{d}s \tag{8.22}$$

ausgedrückt werden. Diese Formel stellt die Lösung der inhomogenen Gleichung als Superposition von Lösungen homogener Gleichungen dar und wird als Duhamel-Prinzip bezeichnet.

Weil $E(t)$ in der L_2-Norm beschränkt ist, ist die rechte Seite von (8.22) wohldefiniert. Der erste Term ist die Lösung von (8.1). Der zweite Term verschwindet für $t = 0$, sodass wir zum Beweis, dass u eine Lösung von (8.21) ist,

$$D_t F(t) - \Delta F(t) = f(t) \quad \text{mit } F(t) = \int_0^t E(t-s)f(s)\,\mathrm{d}s \tag{8.23}$$

demonstrieren müssen. Formal gilt nach Differentiation des Integrals

$$D_t F(t) - \Delta F(t) = f(t) + \int_0^t D_t E(t-s)f(s)\,\mathrm{d}s - \int_0^t \Delta E(t-s)f(s)\,\mathrm{d}s, \tag{8.24}$$

und wegen $D_t E(t-s) = \Delta E(t-s)$ sollten sich die Integrale aufheben. Fordern wir jedoch nur $f(s) \in L_2$ für $s \in (0,t)$, dann zeigt (8.18) eine Singularität der Ordnung $O((t-s)^{-1})$ in den Integranden, sodass die Integrale nicht notwendigerweise wohldefiniert sind. Aus diesem Grund nehmen wir an, dass $\|D_t f(t)\|$ für $t \in [0,T]$ mit beliebigem $T > 0$ beschränkt ist, und schreiben, nachdem wir im letzten Term $t - s$ durch s ersetzt haben,

$$F(t) = \int_0^t E(t-s)(f(s) - f(t))\,\mathrm{d}s + \int_0^t E(s)f(t)\,\mathrm{d}s.$$

Durch Differentiation bezüglich t erhalten wir

$$D_t F(t) = \int_0^t D_t E(t-s)(f(s) - f(t))\,\mathrm{d}s + E(t)f(t), \tag{8.25}$$

wobei der Integrand nun wegen $\|f(s) - f(t)\| \le C|s-t|$ beschränkt ist. Analog dazu gilt wegen $\Delta E(t-s) = D_t E(t-s)$

$$\Delta F(t) = \int_0^t \Delta E(t-s)(f(s) - f(t))\,\mathrm{d}s + (E(t) - I)f(t). \tag{8.26}$$

Bilden wir die Differenz zwischen (8.25) und (8.26), so ergibt sich (8.23).

Eine andere Möglichkeit, die Singularitäten in den Integranden in (8.24) zu behandeln, wäre, die Regularität von $f(s)$ in der räumlichen Variable auszunutzen, beispielsweise in Form der Ungleichung $\|\Delta E(t-s)f(s)\| \le \|\Delta f(s)\|$. Zusätzlich zur Regularität von $f(s)$ wäre jedoch die unnatürliche Randbedingung $f(s) = 0$ auf Γ erforderlich.

Mit (8.22) erhalten wir sofort die Stabilitätsabschätzung

$$\|u(t)\| \le \|v\| + \int_0^t \|f(s)\| \, \mathrm{d}s. \tag{8.27}$$

Wie gewöhnlich kann dies dazu benutzt werden, sowohl die Eindeutigkeit der Lösung von (8.21) als auch die stetige Abhängigkeit der Lösung von den Daten zu zeigen. Sind beispielsweise u_1 und u_2 Lösungen, die zu den rechten Seiten f_1 und f_2 sowie den Anfangswerten v_1 und v_2 gehören, dann gilt

$$\|u_1(t) - u_2(t)\| \le \|v_1 - v_2\| + \int_0^t \|f_1(s) - f_2(s)\| \, \mathrm{d}s \quad \text{für } t \in \mathbf{R}_+. \tag{8.28}$$

8.3 Variationsformulierung. Energieabschätzungen

Wir werden nun das Anfangs-Randwertproblem (8.21) in variationaler Form aufschreiben und daraus einige Abschätzungen für dessen Lösung herleiten. Auch wenn wir dies hier nicht weiter verfolgen werden, können Variationsmethoden zum Beweis der Existenz und der Eindeutigkeit von Lösungen parabolischer Probleme benutzt werden. Diese können wesentlich allgemeiner als (8.21) sein. Beispiele dafür sind Probleme mit zeitabhängigen Koeffizienten oder einem nicht selbstadjungierten, elliptischen Operator, Probleme mit inhomogenen Randbedingungen und auch einige nichtlineare Probleme. Dies sind Probleme, für die die Methode der Entwicklung nach Eigenfunktionen aus dem vorangegangenem Abschnitt schwierig oder unmöglich anzuwenden ist. Darüber hinaus bildet die Variationsformulierung die Basis für die Methode der finiten Elemente für parabolische Probleme, die wir in Kapitel 10 untersuchen werden.

Für die Variationsformulierung multiplizieren wir die Wärmeleitungsgleichung in (8.21) mit einer glatten Funktion $\varphi = \varphi(x)$, die auf Γ verschwindet, und bestimmen nach Integration über Ω und nach Anwendung der Greenschen Formel

$$(u_t, \varphi) + a(u, \varphi) = (f, \varphi) \quad \forall \varphi \in H_0^1, \quad t \in \mathbf{R}_+ \tag{8.29}$$

mit unserer Standardnotation

$$a(v, w) = \int_\Omega \nabla v \cdot \nabla w \, \mathrm{d}x, \quad (v, w) = \int_\Omega vw \, \mathrm{d}x.$$

Das Variationsproblem kann dann folgendermaßen formuliert werden: Gesucht ist eine Funktion $u = u(x,t) \in H_0^1$, die auf Γ für $t > 0$ verschwindet und für die (8.29) erfüllt ist sowie

$$u(\cdot, 0) = v \quad \text{in } \Omega \tag{8.30}$$

gilt.

Durch Ausführen der oben diskutierten Schritte in umgekehrter Reihenfolge kann man leicht sehen, dass eine hinreichend glatte Lösung u dieses Problems auch eine Lösung von (8.21) ist. Tatsächlich erhalten wir durch partielle Integration in (8.29)

$$(u_t - \Delta u - f, \varphi) = 0 \quad \forall \varphi \in H_0^1, \quad t \in \mathbf{R}_+$$

oder für jedes $t \in \mathbf{R}_+$

$$\int_\Omega \rho(\cdot, t)\, \varphi \, \mathrm{d}x = 0 \quad \forall \varphi \in H_0^1 \quad \text{mit } \rho = u_t - \Delta u - f.$$

Wie beim stationären Problem schlussfolgern wir, dass dies nur für $\rho = 0$ möglich ist.

Das folgende Resultat zeigt einige Abschätzungen für die Lösung des oben gestellten Problems in Abhängigkeit von den Daten für verschiedene Normen. Dabei gehen wir formal vor und verzichten auf präzise Aussagen hinsichtlich der geforderten Regularität. Wir schreiben $u(t)$ für $u(\cdot, t)$ und analog dazu $f(t)$.

Theorem 8.5. *Angenommen, $u(t)$ erfüllt die Gleichungen* (8.29) *und* (8.30)*, verschwindet auf Γ und ist für $t \geq 0$ hinreichend glatt. Dann gibt es eine Konstante C, mit der für $t \geq 0$ die Ungleichungen*

$$\|u(t)\|^2 + \int_0^t |u(s)|_1^2 \, \mathrm{d}s \leq \|v\|^2 + C \int_0^t \|f(s)\|^2 \, \mathrm{d}s \tag{8.31}$$

und

$$|u(t)|_1^2 + \int_0^t \|u_t(s)\|^2 \, \mathrm{d}s \leq |v|_1^2 + \int_0^t \|f(s)\|^2 \, \mathrm{d}s. \tag{8.32}$$

gelten.

Beweis. Setzen wir in (8.29) $\varphi = u$, so erhalten wir

$$(u_t, u) + a(u, u) = (f, u) \quad \text{für } t > 0. \tag{8.33}$$

Hier gilt

$$(u_t, u) = \int_\Omega u_t u \, \mathrm{d}x = \int_\Omega \tfrac{1}{2}(u^2)_t \, \mathrm{d}x = \tfrac{1}{2} \frac{d}{dt} \|u\|^2.$$

Nach Anwendung der Poincaré-Ungleichung, Theorem A.6, d. h.

$$\|\varphi\| \le C|\varphi|_1 = C\ a(\varphi,\varphi)^{1/2} \quad \text{für } \varphi \in H_0^1,$$

gilt unter gleichzeitiger Verwendung der Ungleichung $2ab \le a^2 + b^2$

$$|(f,u)| \le \|f\|\,\|u\| \le C\|f\|\,|u|_1 \le \tfrac{1}{2}|u|_1^2 + \tfrac{1}{2}C^2\|f\|^2.$$

Wir erhalten aus (8.33) also

$$\tfrac{1}{2}\frac{d}{dt}\|u\|^2 + |u|_1^2 \le \tfrac{1}{2}|u|_1^2 + \tfrac{1}{2}C^2\|f\|^2$$

oder mit einer neuen Konstante C

$$\frac{d}{dt}\|u\|^2 + |u|_1^2 \le C\|f\|^2.$$

Durch Integration von 0 bis t erhalten wir

$$\|u(t)\|^2 + \int_0^t |u(s)|_1^2\,\mathrm{d}s \le \|v\|^2 + C\int_0^t \|f\|^2\,\mathrm{d}s,$$

das ist Gleichung (8.31).

Zum Beweis von (8.32) wählen wir in (8.29) nun $\varphi = u_t$ und erhalten

$$\|u_t\|^2 + a(u,u_t) = (f,u_t) \le \tfrac{1}{2}\|f\|^2 + \tfrac{1}{2}\|u_t\|^2.$$

Hier ist

$$a(u,u_t) = \int_\Omega \nabla u\cdot\nabla u_t\,\mathrm{d}x = \int_\Omega \tfrac{1}{2}(|\nabla u|^2)_t\,\mathrm{d}x = \tfrac{1}{2}\frac{d}{dt}|u|_1^2,$$

sodass wir

$$\|u_t\|^2 + \frac{d}{dt}|u|_1^2 \le \|f\|^2$$

schlussfolgern können, woraus sich durch Integration über $(0,t)$

$$|u(t)|_1^2 + \int_0^t \|u_t\|^2\,\mathrm{d}s \le |v|_1^2 + \int_0^t \|f\|^2\,\mathrm{d}s$$

(8.32) ergibt. □

Aus (8.31) folgt wie gewöhnlich, dass für die Lösungen u_1 und u_2, die zu den rechten Seiten f_1 und f_2 und den Anfangswerten v_1 und v_2 gehören,

$$\|u_1(t)-u_2(t)\|^2 + \int_0^t |u_1-u_2|_1^2\,\mathrm{d}s \le \|v_1-v_2\|^2 + C\int_0^t \|f_1-f_2\|^2\,\mathrm{d}s \quad \text{für } t \ge 0$$

gilt. Eine ähnliche Schranke erhält man aus (8.32). Beachten Sie, dass diese Schranken auch den Fehler in H_0^1 abschätzen und dabei die L_2-Norm anstatt der in (8.28) verwendeten L_1-Norm benutzt wird.

8.4 Ein Maximumprinzip

Wir betrachten nun die Verallgemeinerung des gemischten Anfangswertproblems aus Abschnitt 8.2, die einen Quellterm und inhomogene Randbedingungen zulässt. Dabei ist ein u auf $\bar{\Omega} \times \bar{I}$ zu bestimmen, für das

$$\begin{aligned} u_t - \Delta u &= f && \text{in } \Omega \times I, \\ u &= g && \text{auf } \Gamma \times I, \\ u(\cdot, 0) &= v && \text{in } \Omega \end{aligned} \tag{8.34}$$

gilt, wobei Ω ein beschränktes Gebiet in $\mathbf{R}^d$ und $I = (0, T)$ ein endliches Zeitintervall ist. Um ein Maximumprinzip für dieses Problem zu beweisen, ist es günstig, den *parabolischen Rand* von $\Omega \times I$ einzuführen. Dies ist die Menge $\Gamma_p = (\Gamma \times \bar{I}) \cup (\Omega \times \{t = 0\})$, d. h. der Rand von $\Omega \times I$ ohne das Innere des oberen Teils dieses Randes $\Omega \times \{t = T\}$.

Theorem 8.6. *Sei u glatt und es gelte $u_t - \Delta u \leq 0$ in $\Omega \times I$. Dann nimmt u sein Maximum auf dem parabolischen Rand Γ_p an.*

Beweis. Wenn dies nicht der Fall wäre, würde das Maximum entweder in einem inneren Punkt von $\Omega \times I$ oder in einen Punkt von $\Omega \times \{t = T\}$ angenommen werden, d. h. in einem Punkt $(\bar{x}, \bar{t}) \in \Omega \times (0, T]$, und es wäre

$$u(\bar{x}, \bar{t}) = \max_{\bar{\Omega} \times \bar{I}} u = M > m = \max_{\Gamma_p} u.$$

Dann würde die Funktion

$$w(x, t) = u(x, t) + \epsilon |x|^2$$

für hinreichend kleine $\epsilon > 0$ ihr Maximum ebenfalls in einem Punkt in $\Omega \times (0, T]$ annehmen, weil für kleines ϵ

$$\max_{\Gamma_p} w \leq m + \epsilon \max_{\Gamma_p} |x|^2 < M \leq \max_{\bar{\Omega} \times \bar{I}} w$$

gilt. Aufgrund unserer Annahme gilt wegen $\Delta(|x|^2) = 2d$

$$w_t - \Delta w = u_t - \Delta u - 2d\epsilon < 0 \quad \text{in } \Omega \times I. \tag{8.35}$$

Andererseits gilt im Punkt $(\tilde{x}, \tilde{t})$, in dem w ihr Maximum annimmt,

$$-\Delta w(\tilde{x}, \tilde{t}) = -\sum_{i=1}^{d} w_{x_i x_i}(\tilde{x}, \tilde{t}) \geq 0$$

und

$$w_t(\tilde{x}, \tilde{t}) = 0 \quad \text{für } \tilde{t} < T \quad \text{oder} \quad w_t(\tilde{x}, \tilde{t}) \geq 0 \quad \text{für } \tilde{t} = T,$$

sodass in beiden Fällen

$$w_t(\tilde{x}, \tilde{t}) - \Delta w(\tilde{x}, \tilde{t}) \geq 0$$

ist. Dies ist ein Widerspruch zu (8.35), was unsere Behauptung beweist. □

Für die Funktionen $\pm u$ folgt insbesondere, dass die Lösung der homogenen Wärmeleitungsgleichung ($f = 0$) sowohl ihr Maximum als auch ihr Minimum auf Γ_p annimmt, sodass in diesem Fall mit $\|w\|_{\mathcal{C}(\bar{M})} = \max_{x\in\bar{M}} |w(x)|$

$$\|u\|_{\mathcal{C}(\bar{\Omega}\times\bar{I})} \le \max\left\{\|g\|_{\mathcal{C}(\Gamma\times\bar{I})}, \|v\|_{\mathcal{C}(\bar{\Omega})}\right\}$$

gilt. Für das inhomogene Problem kann man die folgende Ungleichung zeigen, deren Beweis wir dem Leser als Übung überlassen (siehe Problemstellung 8.7).

Theorem 8.7. *Die Lösung von* (8.34) *erfüllt*

$$\|u\|_{\mathcal{C}(\bar{\Omega}\times\bar{I})} \le \max\left\{\|g\|_{\mathcal{C}(\Gamma\times\bar{I})}, \|v\|_{\mathcal{C}(\bar{\Omega})}\right\} + \frac{r^2}{2d}\|f\|_{\mathcal{C}(\bar{\Omega}\times\bar{I})},$$

wobei r der Radius einer Kugel ist, die Ω enthält.

Wie gewöhnlich beweist ein solches Resultat die Eindeutigkeit und die Stabilität des Anfangs-Randwertproblems.

Wir beschließen diesen Abschnitt mit dem Beweis der Eindeutigkeit einer beschränkten Lösung des reinen Anfangswertproblems, das wir in Abschnitt 8.1 betrachtet haben.

Theorem 8.8. *Das Anfangswertproblem* (8.1) *besitzt höchstens eine Lösung, die für beliebiges T in $\mathbf{R}^d \times [0,T]$ beschränkt ist.*

Beweis. Wenn es zwei Lösungen von (8.1) gäbe, dann wäre deren Differenz eine Lösung mit den Anfangsdaten null. Es reicht deshalb aus zu zeigen, dass die einzige beschränkte Lösung u von

$$\begin{aligned} u_t &= \Delta u \qquad \text{in } \mathbf{R}^d \times I \quad \text{mit } I = (0,T), \\ u(\cdot,0) &= 0 \qquad \text{in } \mathbf{R}^d \end{aligned}$$

die triviale Lösung $u = 0$ ist, oder dass für einen beliebigen Punkt (x_0, t_0) in $\mathbf{R}^d \times I$ und ein beliebiges $\epsilon > 0$ die Ungleichung $|u(x_0,t_0)| \le \epsilon$ erfüllt ist. Wir führen die Hilfsfunktion

$$w(x,t) = \frac{|x|^2 + 2d\,t}{|x_0|^2 + 2d\,t_0}$$

ein und weisen darauf hin, dass $w_t = \Delta w$ gilt. Sei nun

$$h_\pm(x,t) = -\epsilon w(x,t) \pm u(x,t).$$

Dann gilt

$$(h_\pm)_t - \Delta h_\pm = 0 \quad \text{in } \mathbf{R}^d \times I.$$

Da u beschränkt ist, gilt $|u(x,t)| \le M$ auf $\mathbf{R}^d \times I$ für ein M. Definieren wir R durch $R^2 = \max\left(|x_0|^2, M(|x_0|^2 + 2dt_0)/\epsilon\right)$, so gilt

$$h_\pm(x,t) \le -\epsilon \frac{R^2}{|x_0|^2 + 2d\,t_0} + M \le 0 \quad \text{im Falle } |x| = R$$

und

$$h_\pm(x,0) = -\epsilon |x|^2/(|x_0|^2 + 2d\,t_0) \le 0 \quad \text{für } x \in \mathbf{R}^d.$$

Folglich können wir Theorem 8.6 mit $\Omega = \{|x| < R\}$ anwenden und schlussfolgern, dass $h_\pm(x,t) \le 0$ für $(x,t) \in \Omega \times I$ ist. Insbesondere gilt an der Stelle (x_0,t_0) die Beziehung $\pm u(x_0,t_0) = h_\pm(x_0,t_0) + \epsilon \le \epsilon$, was den Beweis des Theorems abschließt. □

Die Annahme von Theorem 8.8, dass die Lösungen in $\mathbf{R}^d \times [0,T]$ beschränkt sind, kann auf die Forderung abgeschwächt werden, dass $|u(x,t)| \le M \mathrm{e}^{c|x|^2}$ für alle $x \in \mathbf{R}^d$, $0 \le t \le T$ und ein $M, c > 0$ gilt. Ohne eine solche Einschränkung für das Wachstum der Lösung für große $|x|$ ist die Eindeutigkeit jedoch nicht garantiert. Beispielsweise ist die folgende Funktion eine Lösung der homogenen Wärmeleitungsgleichung, die die Anfangswerte null besitzt, aber für $t > 0$ nicht identisch verschwindet:

$$u(x,t) = \sum_{n=0}^{\infty} f^{(n)}(t) \frac{x^{2n}}{(2n)!} \quad \text{mit } f(t) = \mathrm{e}^{-1/t^2} \text{ für } t > 0,\ f(0) = 0.$$

Der technische Teil des Beweises besteht darin zu zeigen, dass die Reihe so schnell konvergiert, dass sie gliedweise differenziert werden kann. Dann ist es offensichtlich, dass $u_t = u_{xx}$ und $u(x,0) = 0$ ist.

8.5 Problemstellungen

Problem 8.1. Zeigen Sie, dass für eine Lösung u von (8.1) mit

$$\int_{\mathbf{R}^d} |v(x)|\,\mathrm{d}x < \infty$$

die Gleichung

$$\int_{\mathbf{R}^d} u(x,t)\,\mathrm{d}x = \text{konstant} = \int_{\mathbf{R}^d} v(x)\,\mathrm{d}x \quad \text{für } t \ge 0$$

gilt. Geben Sie eine physikalische Interpretation dieses Resultates an.

Problem 8.2. Bestimmen Sie eine Lösung des Anfangs-Randwertproblems (8.19) mit

(a) $v(x) = 1$ für $0 < x < \pi$;

(b) $v(x) = x(\pi - x)$ für $0 < x < \pi$.

Skizzieren Sie die Lösungen $u(x,t)$ zu verschiedenen Zeiten t.

Problem 8.3. Betrachten Sie die Funktion

$$u(x,t) = \begin{cases} xt^{-3/2}\mathrm{e}^{-x^2/4t} & \text{für } t > 0, \\ 0 & \text{für } t = 0. \end{cases}$$

Zeigen Sie, dass u eine Lösung von

$$u_t - u_{xx} = 0 \quad \text{in } \mathbf{R} \times \mathbf{R}_+$$

ist und dass für jedes x

$$u(x,t) \to 0 \quad \text{für } t \to 0$$

gilt. Weshalb ist dies kein Gegenbeispiel für das Eindeutigkeitsresultat aus Theorem 8.8? Hinweis: Setzen Sie $x = t$.

Problem 8.4. Sei u die Lösung von (8.10). Zeigen Sie

(a) $\|u(t)\| \le \mathrm{e}^{-\lambda_1 t}\|v\|$ für $t \ge 0$,

(b) $\|\Delta^k D_t^j u(t)\| \le Ct^{-(j+k)}\mathrm{e}^{-\lambda_1 t/2}\|v\|$ für $t > 0$.

Problem 8.5. Beweisen Sie mithilfe der Energiemethode, dass eine Konstante $C = C(T)$ existiert, sodass für ein u, das (8.29) und (8.30) erfüllt,

(a) $$\int_0^t s\|u_t(s)\|^2 \,\mathrm{d}s \le C\Big(\|v\|^2 + \int_0^t \|f(s)\|^2 \,\mathrm{d}s\Big) \qquad \text{für } 0 \le t \le T,$$

(b) $$|u(t)|_1^2 \le Ct^{-1}\Big(\|v\|^2 + \int_0^t \|f(s)\|^2 \,\mathrm{d}s\Big) \qquad \text{für } 0 < t \le T$$

gilt.

Problem 8.6. Sei u die Lösung von (8.29) und (8.30) mit $v = 0$. Zeigen Sie

$$\int_0^t \big(\|u_t(s)\|^2 + \|\Delta u(s)\|^2\big) \,\mathrm{d}s \le C\int_0^t \|f(s)\|^2 \,\mathrm{d}s \quad \text{für } t \ge 0.$$

Problem 8.7. Beweisen Sie Theorem 8.7. Hinweis: Sehen Sie sich den Beweis von Theorem 3.2 an.

Problem 8.8. Beweisen Sie Abschätzungen, die analog zu denen aus Theorem 8.5 sind, wenn der elliptische Term $-\Delta u$ in (8.21) durch $\mathcal{A} = -\nabla\cdot(a\nabla u) + b\cdot\nabla u + cu$ wie in Abschnitt 3.5 ersetzt wurde.

Problem 8.9. Beweisen Sie (8.4).

Problem 8.10. Zeigen Sie, dass für ein u, das (8.10) erfüllt, eine Konstante C existiert, für die

$$\|u(t)\|_2^2 + \int_0^t |u_t(s)|_1^2 \,\mathrm{d}s \le C\|v\|_2^2 \quad \forall v \in H^2 \cap H_0^1,\ t \ge 0$$

gilt. Sie können die Spektralmethode oder die Energiemethode benutzen. Sie benötigen außerdem die elliptische Regularitätsabschätzung (3.36).

Problem 8.11. Sei u die Lösung von

$$\begin{aligned} u_t - \Delta u &= 0 && \text{in } \Omega \times \mathbf{R}_+, \\ u(x,t) &= 0 && \text{auf } \Gamma \times \mathbf{R}_+, \\ u(\cdot,0) &= v && \text{in } \Omega, \end{aligned}$$

wobei $\Omega = \{x \in \mathbf{R}^2 : 0 < x_i < 1,\ i = 1,2\}$ ist. Sei $\varphi(x) = A \sin(\pi x_1) \sin(\pi x_2)$ mit $A > 0$. Zeigen Sie, dass im Falle $0 \le v(x) \le \varphi(x)$ für $x \in \Omega$ die Ungleichung $0 \le u(x,t) \le \mathrm{e}^{-2\pi^2 t}\varphi(x)$ für $x \in \Omega$, $t > 0$ erfüllt ist. Hinweis: Verwenden Sie das Maximumprinzip.

Problem 8.12. Beweisen Sie die Version von Theorem 8.1 für den L_2: Wenn $v \in L_2(\mathbf{R}^d)$ ist, dann gilt $\|E(t)v\|_{L_2} \le \|v\|_{L_2}$ für $t \ge 0$ und $\|E(t)v - v\|_{L_2} \to 0$ für $t \to 0$. Hinweis: Parsevalsche Gleichung (A.32).

Problem 8.13. Betrachten Sie die Wärmeleitungsgleichung mit Neumannschen Randbedingungen:

$$\begin{aligned} u_t - \Delta u &= 0 && \text{in } \Omega \times \mathbf{R}_+, \\ \frac{\partial u}{\partial n} &= 0 && \text{auf } \Gamma \times \mathbf{R}_+, \\ u(\cdot,0) &= v && \text{in } \Omega, \end{aligned}$$

wobei $\partial u/\partial n$ die äußere Normalenableitung ist.

(a) Zeigen Sie, dass $\overline{u(t)} = \overline{v}$ für $t \ge 0$ gilt, wobei $\overline{v} = \frac{1}{|\Omega|}\int_\Omega v(x)\,\mathrm{d}x$ den Mittelwert von v bezeichnet.
(b) Zeigen Sie, dass $\|u(t) - \overline{v}\| \to 0$ für $t \to \infty$ gilt.

Problem 8.14. Angenommen, u erfüllt das Anfangs-Randwertproblem

$$\begin{aligned} u_t - \Delta u &= f && \text{in } \Omega \times \mathbf{R}_+, \\ \frac{\partial u}{\partial n} &= g && \text{auf } \Gamma \times \mathbf{R}_+, \\ u(\cdot,0) &= v && \text{in } \Omega, \end{aligned}$$

wobei $\Omega \subset \mathbf{R}^d$ ein beschränktes Gebiet in $\mathbf{R}^d$ mit einem glattem Rand Γ und $\partial u/\partial n$ die äußere Normalenableitung ist. Nehmen Sie zusätzlich an, dass $f(x,t) \ge 0$, $v(x) \ge 0$ für $x \in \Omega$, $t \ge 0$ und $g(x,t) > 0$ für $x \in \Gamma$, $t \ge 0$ gilt. Zeigen Sie, dass $u(x,t) \ge 0$ für $x \in \Omega$, $t \ge 0$ ist. (Tatsächlich ist es ausreichend, $g(x,t) \ge 0$ anzunehmen.)

Problem 8.15. Betrachten Sie die Stokesschen Gleichungen, die die zweidimensionale Bewegung einer viskosen, inkompressiblen Flüssigkeit bei kleiner Reynolds-Zahl R beschreiben:

$$
\begin{aligned}
&\frac{\partial u}{\partial t} - R^{-1}\Delta u + \nabla p = 0 && \text{in } \mathbf{R}^2 \times \mathbf{R}_+,\\
&\nabla \cdot u = 0 && \text{in } \mathbf{R}^2 \times \mathbf{R}_+,\\
&u(\cdot, 0) = v && \text{in } \mathbf{R}^2.
\end{aligned} \tag{8.36}
$$

Hierbei bezeichnet $u(x,t) \in \mathbf{R}^2$ die dimensionslose Geschwindigkeit und $p(x,t) \in \mathbf{R}$ den dimensionslosen Druck. In dieser Form stellt R^{-1} die Viskosität dar. Definieren Sie den Wirbelvektor ω durch

$$\omega = \nabla \times u = \partial u_2/\partial x_1 - \partial u_1/\partial x_2.$$

Zeigen Sie, dass (8.36) mit dem Wirbelvektor als

$$
\begin{aligned}
&\frac{\partial \omega}{\partial t} - R^{-1}\Delta \omega = 0, && \text{in } \mathbf{R}^2 \times \mathbf{R}_+,\\
&\omega(\cdot, 0) = \nabla \times v, && \text{in } \mathbf{R}^2
\end{aligned}
$$

geschrieben werden kann.

Problem 8.16. Sei $u(x,t) = (E(t)v)(x)$ die Lösung von (8.10) und seien $\{\lambda_j\}_{j=1}^{\infty}$ und $\{\varphi_j\}_{j=1}^{\infty}$ die Eigenwerte und die normierten Eigenfunktionen von (6.5) wie in Theorem 6.4. Zeigen Sie

$$u(x,t) = (E(t)v)(x) = \int_\Omega G(x,y,t)v(y)\,\mathrm{d}y,$$

wobei die Greensche Funktion durch

$$G(x,y,t) = \sum_{j=1}^{\infty} \mathrm{e}^{-\lambda_j t}\varphi_j(x)\varphi_j(y)$$

gegeben ist. Hinweis: Sehen Sie sich Problemstellung 6.7 an.

9

Finite Differenzenverfahren für parabolische Probleme

In diesem Kapitel geben wir eine Einführung zur numerischen Lösung parabolischer Gleichungen mithilfe finiter Differenzen und betrachten die Anwendung solcher Verfahren auf die homogene Wärmeleitungsgleichung in einer räumlichen Dimension. Wir diskutieren in Abschnitt 9.1 zunächst das reine Anfangswertproblem, wobei die gegebenen Daten auf der unbeschränkten reellen Achse liegen. In Abschnitt 9.2 betrachten wir dann das gemischte Anfangs-Randwertproblem auf einem endlichen räumlichen Intervall mit Dirichletschen Randbedingungen. Wir diskutieren die Stabilität und die Fehlerschranken für verschiedene Formen finiter Differenzenapproximationen. In der Maximumnorm geschieht dies mithilfe eines Maximumprinzips und in der L_2-Norm durch Fourier-Analyse. Für das unbeschränkte Problem betrachten wir explizite Verfahren und auf einen endlichen Intervall auch implizite Verfahren, wie beispielsweise das Crank-Nicolson-Verfahren.

9.1 Das reine Anfangswertproblem

Wir betrachten zunächst das reine Anfangswertproblem, das darin besteht, ein $u = u(x,t)$ zu bestimmen, für das

$$\begin{aligned} \frac{\partial u}{\partial t} &= \frac{\partial^2 u}{\partial x^2} \qquad \text{in } \mathbf{R} \times \mathbf{R}_+, \\ u(\cdot, 0) &= v \qquad \text{in } \mathbf{R} \end{aligned} \tag{9.1}$$

gilt. Dabei ist v eine gegebene glatte, beschränkte Funktion. Wir wissen aus Abschnitt 8.1, dass dieses Problem eine eindeutige Lösung besitzt und kennen viele Eigenschaften, die beispielsweise aus der Darstellung

$$u(x,t) = \frac{1}{\sqrt{4\pi t}} \int_{-\infty}^{\infty} \mathrm{e}^{-y^2/4t} v(x-y)\,\mathrm{d}y = \big(E(t)v\big)(x) \tag{9.2}$$

hergeleitet werden können. Dabei ist $E(t)$ der Lösungsoperator von (9.1). Insbesondere stellen wir fest, dass die Lösung bezüglich der Maximumnorm

beschränkt ist. Genauer gesagt, gilt

$$\|u(\cdot,t)\|_{\mathcal{C}} = \|E(t)v\|_{\mathcal{C}} \le \|v\|_{\mathcal{C}} = \sup_{x\in\mathbf{R}} |v(x)| \quad \text{für } t \ge 0. \tag{9.3}$$

Für die numerische Lösung dieses Problems durch finite Differenzen führt man ein Gitter mit den Gitterpunkten $(x,t) = (x_j, t_n)$ ein. Dabei gilt $x_j = jh$, $t_n = nk$, wobei j und n ganze Zahlen mit $n \ge 0$ sind, h der Gitterabstand für x und k der Zeitschritt ist und beide klein sind. Man sucht an diesen Gitterpunkten nach einer approximativen Lösung U_j^n. Diese wird durch eine Gleichung bestimmt wird, die man durch Ersetzen der Ableitungen in (9.1) durch finite Differenzenquotienten erhält. Für die auf diesem Gitter definierten Funktionen führen wir also den Vorwärts- und Rückwärts-Differenzenquotienten

$$\partial_x U_j^n = h^{-1}(U_{j+1}^n - U_j^n) \quad \text{und} \quad \bar\partial_x U_j^n = h^{-1}(U_j^n - U_{j-1}^n)$$

bezüglich x sowie analog dazu beispielsweise

$$\partial_t U_j^n = k^{-1}(U_j^{n+1} - U_j^n)$$

bezüglich t ein. Das einfachste, zu (9.1) gehörende finite Differenzenverfahren ist das *Vorwärts-Euler-Verfahren*

$$\begin{aligned} &\partial_t U_j^n = \partial_x \bar\partial_x U_j^n && \text{für } j,\, n \in \mathbf{Z},\ n \ge 0,\\ &U_j^0 = v_j := v(x_j) && \text{für } j \in \mathbf{Z} \end{aligned}$$

mit dem Raum der ganzen Zahlen $\mathbf{Z}$. Die Differenzengleichung kann auch als

$$\frac{U_j^{n+1} - U_j^n}{k} = \frac{U_{j+1}^n - 2U_j^n + U_{j-1}^n}{h^2}$$

oder nach Einführung des Gitterverhältnisses $\lambda = k/h^2$ als

$$U_j^{n+1} = (E_k U^n)_j = \lambda U_{j-1}^n + (1-2\lambda)U_j^n + \lambda U_{j+1}^n \tag{9.4}$$

geschrieben werden, was den lokalen, diskreten Lösungsoperator E_k definiert. Wir werden h und k in der Beziehung $k/h^2 = \lambda =$ konstant betrachten und können deshalb die Abhängigkeit von h in der Notation weglassen. Das Verfahren (9.4) wird als *explizit* bezeichnet, da es die Lösung an der Stelle $t = t_{n+1}$ explizit als Funktion der Werte an der Stelle $t = t_n$ ausdrückt. Nach Iteration ergibt sich die Lösung des diskreten Problems durch

$$U_j^n = (E_k^n U^0)_j = (E_k^n v)_j \quad \text{für } j\, n \in \mathbf{Z},\ n \ge 0.$$

Wir nehmen nun $\lambda \le \frac{1}{2}$ an. Alle Koeffizienten des Operators E_k in (9.4) sind dann nichtnegativ, und weil deren Summe 1 ist, finden wir

$$|U_j^{n+1}| \le \lambda|U_{j-1}^n| + (1-2\lambda)|U_j^n| + \lambda|U_{j+1}^n| \le \sup_{j\in\mathbf{Z}} |U_j^n|,$$

sodass

$$\sup_{j\in\mathbf{Z}} |U_j^{n+1}| \le \sup_{j\in\mathbf{Z}} |U_j^n|$$

ist. Definieren wir für die Gitterfunktionen $v = (v_j)$ eine diskrete Maximumnorm durch

$$\|v\|_{\infty,h} = \sup_{j\in\mathbf{Z}} |v_j|, \tag{9.5}$$

erhalten wir also

$$\|U^{n+1}\|_{\infty,h} = \|E_k U^n\|_{\infty,h} \le \|U^n\|_{\infty,h}$$

und daraus durch wiederholte Anwendung

$$\|U^n\|_{\infty,h} = \|E_k^n v\|_{\infty,h} \le \|v\|_{\infty,h}. \tag{9.6}$$

Dies ist ein diskretes Analogon zur Abschätzung (9.3) für das kontinuierliche Problem.

Die Beschränktheit des diskreten Lösungsoperators wird als *Stabilität* dieses Operators bezeichnet. Wir werden nun sehen, dass das Verfahren für ein λ, das größer als die Konstante $\frac{1}{2}$ gewählt wird, instabil ist. Um uns davon zu überzeugen, wählen wir $v_j = (-1)^j\epsilon$ mit einer kleinen positiven Zahl ϵ, sodass $\|v\|_{\infty,h} = \epsilon$ gilt. Dann ist

$$U_j^1 = \big(\lambda(-1)^{j-1} + (1-2\lambda)(-1)^j + \lambda(-1)^{j+1}\big)\epsilon = (1-4\lambda)(-1)^j\epsilon,$$

oder allgemeiner

$$U_j^n = (1-4\lambda)^n(-1)^j\epsilon,$$

woraus sich

$$\|U^n\|_{\infty,h} = (4\lambda-1)^n\epsilon \to \infty \quad \text{für } n\to\infty$$

ergibt. Wir stellen also in diesem Falle fest, dass die Norm der diskreten Lösung auch für sehr kleine Anfangsdaten für $n\to\infty$ mit $k = t/n \to 0$ gegen unendlich strebt, selbst wenn $t = t_n$ beschränkt ist. Dies kann so interpretiert werden, dass sehr kleine Abweichungen der Anfangsdaten (beispielsweise durch Rundungsfehler) große Veränderungen in der diskreten Lösung zu einem späteren Zeitpunkt verursachen können, sodass diese wertlos wird.

Wir beschränken uns nun bei unseren Betrachtungen auf den stabilen Fall $\lambda \le \frac{1}{2}$ und zeigen, dass die diskrete Lösung gegen die exakte Lösung konvergiert, wenn die Gitterkonstanten gegen null streben. Dabei setzen wir voraus, dass die Anfangsdaten und somit die exakte Lösung von (9.1) hinreichend glatt sind. Um dies zu demonstrieren, müssen wir die Tatsache verwenden, dass die exakte Lösung die Differenzengleichung bis auf einen kleinen Fehler erfüllt, der mit h und k gegen null strebt. Setzen wir $u_j^n = u(x_j, t_n)$,

gilt wegen der Taylorschen Formel für die Lösung von (9.1) mit geeignetem $\bar{x}_j \in (x_{j-1}, x_{j+1})$, $\bar{t}_n \in (t_n, t_{n+1})$

$$\begin{aligned}\tau_j^n &= \partial_t u_j^n - \partial_x \bar{\partial}_x u_j^n = \Big(\partial_t u_j^n - u_t(x_j, t_n)\Big) - \Big(\partial_x \bar{\partial}_x u_j^n - u_{xx}(x_j, t_n)\Big) \\ &= \tfrac{1}{2} k u_{tt}(x_j, \bar{t}_n) - \tfrac{1}{12} h^2 u_{xxxx}(\bar{x}_j, t_n).\end{aligned}$$

Weil $u_{tt} = u_{xxxx}$ gilt und aus (9.2) leicht zu erkennen ist, dass für die Lösung von (9.1) $|u(\cdot,t)|_{\mathcal{C}^4} \le |v|_{\mathcal{C}^4}$ gilt, erhalten wir

$$\begin{aligned}\|\tau^n\|_{\infty,h} &\le Ck \max_{t\in I_n} |u_{tt}(\cdot,t)|_{\mathcal{C}} + Ch^2 |u(\cdot,t_n)|_{\mathcal{C}^4} \\ &\le Ch^2 \max_{t\in I_n} |u(\cdot,t)|_{\mathcal{C}^4} \le Ch^2 |v|_{\mathcal{C}^4} \quad \text{für } \lambda \le \frac{1}{2}.\end{aligned} \tag{9.7}$$

Der Ausdruck τ_j^n wird als *Rundungsfehler* (oder *lokaler Diskretisierungsfehler*) bezeichnet. Es gilt nun die folgende Fehlerabschätzung.

Theorem 9.1. *Seien U^n und u Lösungen von* (9.4) *und* (9.1) *und sei $k/h^2 = \lambda \le \frac{1}{2}$. Dann existiert eine Konstante C, mit der*

$$\|U^n - u^n\|_{\infty,h} \le C t_n h^2 |v|_{\mathcal{C}^4} \quad \textit{für } t_n \ge 0$$

gilt.

Beweis. Wir setzen $z^n = U^n - u^n$. Dann gilt

$$\partial_t z_j^n - \partial_x \bar{\partial}_x z_j^n = -\tau_j^n$$

und folglich

$$z_j^{n+1} = (E_k z^n)_j - k\tau_j^n.$$

Durch wiederholte Anwendung führt dies auf

$$z_j^n = (E_k^n z^0)_j - k \sum_{l=0}^{n-1} (E_k^{n-1-l} \tau^l)_j.$$

Weil $z_j^0 = U_j^0 - u_j^0 = v_j - v_j = 0$ ist, ergibt sich unter Verwendung der Stabilitätsabschätzung (9.6) und der Abschätzung für den Rundungsfehler (9.7),

$$\|z^n\|_{\infty,h} \le k \sum_{l=0}^{n-1} \|\tau^l\|_{\infty,h} \le C\, nk\, h^2 |v|_{\mathcal{C}^4},$$

was dem gesuchten Resultat entspricht. □

Das soeben beschriebene Verfahren ist in der Zeit von der *Genauigkeit* erster Ordnung und im Raum von der Genauigkeit zweiter Ordnung. Da allerdings k und h durch $k/h^2 = \lambda \le \frac{1}{2}$ verbunden sind, ergibt sich insgesamt eine Genauigkeit zweiter Ordnung bezüglich des Gitterabstandes h.

Im Allgemeinen können wir Finite-Differenzen-Operatoren von der Form

$$U_j^{n+1} = (E_k U^n)_j := \sum_p a_p U_{j-p}^n \quad \text{für } j,\, n \in \mathbf{Z},\ n \geq 0 \tag{9.8}$$

betrachten, wobei $a_p = a_p(\lambda)$, $\lambda = k/h^2$ gilt und die Summe endlich ist. Diesem Operator kann man das trigonometrische Polynom

$$\tilde{E}(\xi) = \sum_p a_p \mathrm{e}^{-\mathrm{i}p\xi} \tag{9.9}$$

zuordnen. Dieses Polynom ist für die Stabilitätsanalyse relevant und wird als *Symbol* oder als *charakteristisches Polynom* von E_k bezeichnet. Es ergibt sich sofort folgendes Resultat.

Theorem 9.2. *Eine notwendige Bedingung für die Stabilität des Operators* E_k *in* (9.8) *bezüglich der in* (9.5) *definierten diskreten Maximumnorm ist*

$$|\tilde{E}(\xi)| \leq 1 \quad \textit{für } \xi \in \mathbf{R}. \tag{9.10}$$

Beweis. Angenommen, E_k ist stabil und es gilt $|\tilde{E}(\xi_0)| > 1$ für ein $\xi_0 \in \mathbf{R}$. Dann ist für $v_j = \mathrm{e}^{\mathrm{i}j\xi_0}\epsilon$

$$U_j^1 = \epsilon \sum_p a_p \mathrm{e}^{\mathrm{i}(j-p)\xi_0} = \tilde{E}(\xi_0) v_j.$$

Durch wiederholte Anwendung führt dies auf

$$\|U^n\|_{\infty,h} = |\tilde{E}(\xi_0)|^n \epsilon \to \infty \quad \text{für } n \to \infty.$$

Wegen $\|v\|_{\infty,h} = \epsilon$ widerspricht dies der Stabilität und beweist damit das Theorem. □

Für den in (9.4) definierten Finite-Differenzen-Oprator gilt $\tilde{E}(\xi) = 1 - 2\lambda + 2\lambda\cos\xi$. Da $\cos\xi$ in $[-1, 1]$ liegt, ist die Bedingung (9.10) äquivalent zu $1 - 4\lambda \geq -1$ oder $\lambda \leq \frac{1}{2}$, was mit der vorhin festgelegten Stabilitätsbedingung übereinstimmt.

Die Bedingung (9.10) bildet einen Spezialfall der *von Neumannschen Stabilitätsbedingung*. Wir werden sehen, dass diese Bedingung in einer etwas anderen Situation für die Stabilität ebenfalls hinreichend ist.

Aufgrund seiner Definition ist das charakteristische Polynom eines diskreten Lösungsoperators besonders geeignet für die Untersuchung finiter Differenzenverfahren im Zusammenhang mit der Fourier-Analyse. Es ist dann zweckmäßig, die l_2-Norm als Maß für die Gitterfunktionen zu verwenden. Sei also $V = \{V_j\}_{-\infty}^{\infty}$ eine Gitterfunktion in der räumlichen Variable. Wir setzen

$$\|V\|_{2,h} = \Big(h \sum_{j=-\infty}^{\infty} V_j^2\Big)^{1/2}.$$

Die Menge der auf diese Weise normierten Gitterfunktionen mit endlicher Norm wird mit $l_{2,h}$ bezeichnet. Für eine solche Gitterfunktion definieren wir außerdem deren diskrete Fourier-Transformierte

$$\hat{V}(\xi) = h \sum_{j=-\infty}^{\infty} V_j \mathrm{e}^{-\mathrm{i}j\xi},$$

wobei wir annehmen, dass die Summe absolut konvergent ist. Die Funktion $\hat{V}(\xi)$ ist 2π-periodisch und V kann durch die Rücktransformation

$$V_j = \frac{1}{2\pi h} \int_{-\pi}^{\pi} \hat{V}(\xi) \mathrm{e}^{\mathrm{i}j\xi} \, \mathrm{d}\xi$$

aus $\hat{V}(\xi)$ wiedergewonnen werden. Wir wiederholen die Parsevalsche Gleichung

$$\|V\|_{2,h}^2 = \frac{1}{2\pi h} \int_{-\pi}^{\pi} |\hat{V}(\xi)|^2 \, \mathrm{d}\xi = \frac{1}{2\pi} \int_{-\pi/h}^{\pi/h} |\hat{V}(h\xi)|^2 \, \mathrm{d}\xi. \tag{9.11}$$

Nun können wir die Stabilität bezüglich der Norm $\|\cdot\|_{2,h}$ oder die *Stabilität in* $l_{2,h}$ definieren. Diese ist ähnlich zu (9.6), erlaubt allerdings einen konstanten Faktor C auf der rechten Seite, es gilt also

$$\|E_k^n V\|_{2,h} \le C \|V\|_{2,h} \quad \text{für } n \ge 0,\ h \in (0,1). \tag{9.12}$$

Damit erhalten wir folgendes Theorem.

Theorem 9.3. *Die von Neumannsche Bedingung* (9.10) *ist eine notwendige und hinreichende Bedingung für die Stabilität des Operators* E_k^n *in* $l_{2,h}$.

Beweis. Wir beachten, dass

$$\begin{aligned}(E_k V)\hat{}\,(\xi) &= h \sum_j \sum_p a_p V_{j-p} \mathrm{e}^{-\mathrm{i}j\xi} \\ &= \sum_p a_p \mathrm{e}^{-\mathrm{i}p\xi} h \sum_j V_{j-p} \mathrm{e}^{-\mathrm{i}(j-p)\xi} = \tilde{E}(\xi)\hat{V}(\xi)\end{aligned}$$

gilt. Folglich ist

$$(E_k^n V)\hat{}\,(\xi) = \tilde{E}(\xi)^n \hat{V}(\xi).$$

Unter Verwendung der Parsevalschen Gleichung (9.11) ist die Stabilität von E_k in $l_{2,h}$ äquivalent zu

$$\int_{-\pi}^{\pi} |\tilde{E}(\xi)|^{2n} |\hat{V}(\xi)|^2 \, \mathrm{d}\xi \le C^2 \int_{-\pi}^{\pi} |\hat{V}(\xi)|^2 \, \mathrm{d}\xi \quad \text{für } n \ge 0$$

für alle zulässigen $\hat{V}$. Dies ist aber, wie man leicht sieht, nur genau dann erfüllt, wenn

$$|\tilde{E}(\xi)|^n \le C \quad \text{für } n \ge 0, \quad \xi \in \mathbf{R}$$

gilt, was äquivalent zu (9.10) ist (und in (9.12) gilt folglich $C = 1$). □

Bei der Diskussion eines expliziten, finiten Differenzenverfahrens der Form (9.8) ist es manchmal hilfreich, die Funktionen in der räumlichen Variable x nicht nur als an den Gitterpunkten definiert zu betrachten, sondern für alle $x \in \mathbf{R}$, sodass eine Anfangsfunktion $U^0(x) = v(x)$ gegeben ist und wir nach einer approximativen Lösung $U^n(x)$ an den Stellen $t = t_n$, $n = 1, 2, \ldots$ suchen. Diese erhalten wir aus

$$U^{n+1}(x) = (E_k U^n)(x) = \sum_p a_p U^n(x - x_p), \quad a_p = a_p(\lambda),\ \lambda = k/h^2. \tag{9.13}$$

Ein Vorteil dieses Standpunktes ist, dass dann alle U^n unabhängig von h in demselben Funktionenraum liegen, beispielsweise in $L_2(\mathbf{R})$ oder $\mathcal{C}(\mathbf{R})$.

Wir betrachten kurz den Fall, dass die Analyse in $L_2 = L_2(\mathbf{R})$ stattfindet, und setzen

$$\|u\| = \Big(\int_{-\infty}^{\infty} |u(x)|^2 \, \mathrm{d}x \Big)^{1/2},$$

wobei wir nun auch komplexwertige Funktionen zulassen. Wir benutzen dann die durch

$$(\mathcal{F}v)(\xi) = \hat{v}(\xi) = \int_{-\infty}^{\infty} v(x) \mathrm{e}^{-\mathrm{i}x\xi} \, \mathrm{d}x \tag{9.14}$$

definierte Fourier-Transformierte (siehe Anhang A.3), und stellen fest, dass an dieser Stelle mit dem durch (9.9) definierten $\tilde{E}(\xi)$

$$(E_k v)\hat{}\,(\xi) = \sum_p a_p (\mathcal{F}v(\cdot - ph))(\xi) = \Big(\sum_p a_p \mathrm{e}^{-\mathrm{i}ph\xi} \Big) \hat{v}(\xi) = \tilde{E}(h\xi)\hat{v}(\xi)$$

gilt. Rufen wir uns die Parsevalsche Gleichung für (9.14)

$$\|v\|^2 = (2\pi)^{-1} \|\hat{v}\|^2$$

ins Gedächtnis, finden wir also

$$\|U^n\| = (2\pi)^{-1/2} \|\tilde{E}(h\xi)^n \hat{v}\| \le \sup_{\xi \in \mathbf{R}} |\tilde{E}(h\xi)|^n \|v\|.$$

Deshalb liegt Stabilität bezüglich L_2 genau dann vor, wenn

$$\sup_{\xi \in \mathbf{R}} |\tilde{E}(h\xi)|^n \le C, \quad n \ge 0$$

erfüllt ist, was wiederum äquivalent zur von Neumannschen Bedingung (9.10) ist.

Auch die Konvergenzanalyse kann in L_2 ausgeführt werden. Wir sagen, dass der durch (9.13) definierte Finite-Differenzen-Operator E_k die *Genauigkeit der Ordnung r* besitzt, wenn

$$\tilde{E}(\xi) = \mathrm{e}^{-\lambda\xi^2} + O(|\xi|^{r+2}) \quad \text{für } \xi \to 0 \tag{9.15}$$

gilt. Beispielsweise gilt für den in (9.4) definierten Operator

$$\begin{aligned}\tilde{E}(\xi) &= 1 - 2\lambda + 2\lambda\cos\xi = 1 - \lambda\xi^2 + \tfrac{1}{12}\lambda\xi^4 + O(\xi^6)\\ &= \mathrm{e}^{-\lambda\xi^2} + \big(\tfrac{1}{12}\lambda - \tfrac{1}{2}\lambda^2\big)\xi^4 + O(\xi^6),\end{aligned}$$

sodass (9.4) eine Genauigkeit der Ordnung 2 besitzt oder für die spezielle Wahl $\lambda = \frac{1}{6}$ von der Ordnung 4 ist.

Durch Vergleich der Koeffizienten in der Taylor-Entwicklung von $\tilde{E}(\xi) - \mathrm{e}^{-\lambda\xi^2}$ um $\xi = 0$ mit den Koeffizienten der Entwicklung von $E_k u(x,t) - u(x,t+k)$ um (x,t) mit $k = \lambda h^2$ kann man leicht sehen, dass die Definition (9.15) äquivalent zu der Feststellung ist, dass für die exakte Lösung von (9.1)

$$u^{n+1}(x) - E_k u^n(x) = kO(h^r) \quad \text{für } h \to 0, \quad \lambda = k/h^2 = \text{ konstant}, \tag{9.16}$$

gilt. Das heißt, dass der Einschritt-Lösungsoperator die exakte Lösung bis zur Ordnung $kO(h^r)$ approximiert (siehe Problemstellung 9.1).

Es gilt folgendes Resultat, wobei daran erinnert sei, dass $|\cdot|_s = |\cdot|_{H^s}$ ist.

Theorem 9.4. *Angenommen, E_k ist durch* (9.13) *mit $\lambda = k/h^2 =$ konstant definiert und ist in der Genauigkeit von der Ordnung r und stabil in L_2. Dann gilt*

$$\|U^n - u^n\| \le Ct_n h^r |v|_{r+2} \quad \textit{für } t_n \ge 0.$$

Beweis. Weil $\tilde{E}(\xi)$ auf $\mathbf{R}$ beschränkt ist, gilt wegen (9.15)

$$|\tilde{E}(\xi) - \mathrm{e}^{-\lambda\xi^2}| \le C|\xi|^{r+2} \quad \text{für } \xi \in \mathbf{R}.$$

Aus der Stabilität folgt

$$|\tilde{E}(\xi)^n - \mathrm{e}^{-n\lambda\xi^2}| = |(\tilde{E}(\xi) - \mathrm{e}^{-\lambda\xi^2})\sum_{j=0}^{n-1}\tilde{E}(\xi)^{n-1-j}\mathrm{e}^{-j\lambda\xi^2}| \le Cn|\xi|^{r+2}. \tag{9.17}$$

Wie in Abschnitt 8.1 erhalten wir nun durch Fourier-Transformation von (9.1) bezüglich x

$$\frac{d\hat{u}}{dt}(\xi,t) = -\xi^2\hat{u}(\xi,t) \quad \text{für } t > 0 \quad \text{mit } \hat{u}(\xi,0) = \hat{v}(\xi)$$

und daher

$$\hat{u}(\xi,t) = \mathrm{e}^{-\xi^2 t}\hat{v}(\xi).$$

Wir schlussfolgern

$$(U^n - u^n)\hat{v}(\xi) = \big(\tilde{E}(h\xi)^n - \mathrm{e}^{-nk\xi^2}\big)\hat{v}(\xi)$$

und deshalb

$$\|U^n - u^n\|^2 = (2\pi)^{-1}\|(\tilde{E}(h\xi)^n - \mathrm{e}^{-nk\xi^2})\hat{v}(\xi)\|^2.$$

Nun ist wegen (9.17)

$$|\tilde{E}(h\xi)^n - \mathrm{e}^{-nk\xi^2}| \le Cnh^{r+2}|\xi|^{r+2},$$

sodass unter Verwendung von $(dv/dx)\hat{}\,(\xi) = -\mathrm{i}\xi\hat{v}(\xi)$ und $\lambda = k/h^2$

$$\|U^n - u^n\| \le (2\pi)^{-1/2}Cnh^{r+2}\|\xi^{r+2}\hat{v}(\xi)\| \le C\,nk\,h^r\|v^{(r+2)}\|$$

gilt. Dies zeigt die Folgerung des Theorems unter der Annahme, dass die Anfangsdaten eine solche Form haben, dass $v^{(r+2)}$ zu L_2 gehört. Tatsächlich kann man diese Regularitätsforderung mithilfe eines präziseren Argumentes unter Verwendung der Glättungseigenschaft des Lösungsoperators $E(t)$ um zwei Ableitungen reduzieren. □

Bei der obigen Diskussion haben wir lediglich finite Einschritt-Differenzenverfahren verwendet, d. h. Verfahren, die Werte zur Zeit $t = t_n$ benutzen, um die approximative Lösung an der Stelle $t = t_{n+1}$ zu berechnen. Es wäre ebenso natürlich, die Ableitungen bei der Wärmeleitungsgleichung (9.1) durch Differenzenquotienten zu ersetzen, die symmetrisch zu (x, t_n) sind. Dies würde auf die Gleichung

$$\frac{U^{n+1}(x) - U^{n-1}(x)}{2k} = \partial_x\bar{\partial}_x U^n(x) \tag{9.18}$$

führen. In diesem Fall müssen wir zusätzlich zu $U^0 = v$ auch U^1 vorgeben (etwa durch Approximation von $u(\cdot, k)$), damit wir in der Lage sind, mithilfe von (9.18) U^n für $n \ge 0$ zu bestimmen. Dieses Zweischrittverfahren wäre formal in der Genauigkeit sowohl in x als auch in t von zweiter Ordnung. Obwohl sich das spezielle Verfahren (9.18) für jede Kombination von h und k als instabil erweisen wird (siehe Problemstellung 9.6), sind andere Mehrschrittverfahren bei Anwendungen sinnvoll. Man kann beispielsweise zeigen, dass das Verfahren (9.18) für jede Konstante λ stabilisiert werden kann, indem man $U^n(x)$ auf der rechten Seite durch den Mittelwert $\frac{1}{2}(U^{n+1}(x) + U^{n-1}(x))$ ersetzt, woraus sich das Dufort-Frankel-Verfahren ergibt:

$$\frac{U^{n+1}(x) - U^{n-1}(x)}{2k} = \frac{U^n(x+h) - U^{n+1}(x) - U^{n-1}(x) + U^n(x-h)}{h^2}.$$

Wir beenden diese Diskussion mit einer Beobachtung, die sich auf die Genauigkeit des Dufort-Frankel-Verfahrens bezieht. Es sei u also eine glatte Funktion. Wir ersetzen U durch u. Seien $\partial_x\bar{\partial}_x$ und entsprechend $\partial_t\bar{\partial}_t$ wie vorhin. Mit dem symmetrischen Differenzenquotienten $\hat{\partial}_t$

$$\hat{\partial}_t u(x,t) = \frac{u(x,t+k) - u(x,t-k)}{2k} = \tfrac{1}{2}(\partial_t + \bar{\partial}_t)u(x,t)$$

gilt für den Rundungsfehler

$$\begin{aligned}
&\tau_{h,k,n}(x)\\
&= \frac{u^{n+1}(x) - u^{n-1}(x)}{2k} - \frac{u^n(x+h) - u^{n+1}(x) - u^{n-1}(x) + u^n(x-h)}{h^2}\\
&= \hat{\partial}_t u(x,t_n) - \partial_x \bar{\partial}_x u(x,t_n) + \frac{k^2}{h^2}\partial_t \bar{\partial}_t u(x,t_n)\\
&= (u_t - u_{xx})(x,t_n) + O(k^2) + O(h^2) + \frac{k^2}{h^2} u_{tt}(x,t_n) + O\Big(\frac{k^4}{h^2}\Big).
\end{aligned}$$

Die Konsistenz mit der Wärmeleitungsgleichung fordert deshalb, dass k/h gegen null strebt, was beispielsweise für $k/h^2 = \lambda =$ konstant der Fall ist. Wenn jedoch stattdessen $k/h = \lambda =$ konstant ist, erhalten wir

$$\tau_{h,k,n}(x) = (u_t - u_{xx} + \lambda^2 u_{tt})(x,t_n) + O(h^2) \quad \text{für } h \to 0,$$

wobei das Verfahren dann konsistent ist, allerding nicht mit der Wärmeleitungsgleichung, sondern mit der hyperbolichen Gleichung zweiter Ordnung

$$\lambda^2 u_{tt} + u_t - u_{xx} = 0.$$

Ein Großteil der Analyse in diesem Abschnitt lässt sich auf das Anfangswertproblem für die inhomogene Gleichung

$$\begin{aligned}
u_t &= u_{xx} + f(x,t) && \text{in } \mathbf{R} \times \mathbf{R}_+,\\
u(\cdot,0) &= v && \text{in } \mathbf{R}
\end{aligned}$$

verallgemeinern. Beispielsweise können wir das Vorwärts-Euler-Verfahren

$$\begin{aligned}
\partial_t U_j^n &= \partial_x \bar{\partial}_x U_j^n + f_j^n && \text{für } j,\, n \in \mathbf{Z},\ n \geq 0,\\
U_j^0 &= v_j := v(x_j) && \text{für } j \in \mathbf{Z},
\end{aligned}$$

oder mit dem wie in (9.4) definierten E_k

$$U_j^{n+1} = (E_k U^n)_j + k\, f_j^n,$$

anwenden. Aufgrund der Stabilität von E_k in der Maximumnorm kann man sofort schlussfolgern, dass

$$\|U^n\|_{\infty,h} \leq \|v\|_{\infty,h} + k \sum_{l=0}^{n-1} \|f^l\|_{\infty,h}$$

gilt. Darüber hinaus lässt sich wie beim Beweis von Theorem 9.1 die Fehlerabschätzung

$$\|U^n - u^n\|_{\infty,h} \leq C t_n h^2 \max_{s \leq t_n} \big(|u_{tt}(\cdot,s)|_{\mathcal{C}} + |u(\cdot,s)|_{\mathcal{C}^4}\big)$$

leicht zeigen.

9.2 Das gemischte Anfangs-Randwertproblem

Bei vielen physikalischen Anwendungen ist unser vorhin betrachtetes Modell des reinen Anfangswertproblems (9.1) unzureichend. Stattdessen ist es erforderlich, die Wärmeleitungsgleichung auf einem endlichen Intervall mit an den Endpunkten des Intervalls gegebenen Randwerten für positive Zeiten zu lösen. Dies veranlasst uns zur Betrachtung des folgenden Modellproblems

$$\begin{aligned} &u_t = u_{xx} && \text{in } \Omega = (0,1),\ t > 0, \\ &u(0,t) = u(1,t) = 0 && \text{für } t > 0, \\ &u(\cdot,0) = v && \text{in } \Omega. \end{aligned} \tag{9.19}$$

Zur approximativen Lösung können wir das Gebiet wiederum mit einem Punktegitter überdecken, indem wir diesmal das Gebiet Ω in Teilintervalle gleicher Länge $h = 1/M$ mit einer positiven ganzen Zahl M unterteilen und $(x_j, t_n) = (jh, nk)$ mit $j = 0, \ldots, M$ und $n = 0, 1, \ldots$ setzen. Mit der Approximation U_j^n von $u(x_j, t_n)$ lautet das explizite *Vorwärts-Euler-Verfahren*

$$\begin{aligned} &\partial_t U_j^n = \partial_x \bar{\partial}_x U_j^n && \text{für } j = 1, \ldots, M-1,\ n \geq 0, \\ &U_0^n = U_M^n = 0 && \text{für } n > 0, \\ &U_j^0 = V_j = v(x_j) && \text{für } j = 0, \ldots, M \end{aligned} \tag{9.20}$$

oder mit gegebenem U_j^n, $j = 0, \ldots, M$

$$\begin{aligned} &U_j^{n+1} = \lambda(U_{j-1}^n + U_{j+1}^n) + (1 - 2\lambda)U_j^n, \quad j = 1, \ldots, M-1, \\ &U_0^{n+1} = U_M^{n+1} = 0. \end{aligned}$$

In diesem Fall suchen wir also nach einer Folge von $(M+1)$-komponentigen Vektoren $U^n = (U_0^n, \ldots, U_M^n)^T$ mit $U_0^n = U_M^n = 0$, die diese Gleichungen erfüllen. Bei dieser Analyse werden wir zunächst die diskrete Maximumnorm

$$\|U^n\|_{\infty,h} = \max_{0 \leq j \leq M} |U_j^n|$$

verwenden. Im Falle $\lambda = k/h^2 \leq \frac{1}{2}$ schlussfolgern wir wie für das reine Anfangswertproblem

$$\|U^{n+1}\|_{\infty,h} \leq \|U^n\|_{\infty,h}$$

oder mit der üblichen Definition des Lösungsoperators E_k

$$\|E_k^n V\|_{\infty,h} \leq \|V\|_{\infty,h} \quad \text{für } n \geq 0.$$

Das Verfahren ist somit für $\lambda \leq \frac{1}{2}$ in der Maximumnorm stabil.

Um uns davon zu überzeugen, dass diese Bedingung für die Stabilität auch im vorliegenden Fall notwendig ist, modifizieren wir unser Gegenbeispiel aus Abschnitt 9.1 so, dass die Randbedingungen berücksichtigt werden, und setzen

$$U_j^0 = V_j = (-1)^j \sin(\pi j h) \quad \text{für } j = 0, \ldots, M.$$

Mithilfe einer einfachen Rechnung wie im Beweis von Theorem 9.2 gilt dann

$$U_j^n = (1 - 2\lambda - 2\lambda\cos(\pi h))^n V_j \quad \text{für } j = 0, \ldots, M.$$

Im Falle $\lambda > \frac{1}{2}$ gilt für alle hinreichend kleine h

$$|1 - 2\lambda - 2\lambda\cos(\pi h)| \geq \gamma > 1$$

und deshalb für $t_n = 1$

$$\|U^n\|_{\infty,h} \geq \gamma^n \|V\|_{\infty,h} \to \infty \quad \text{für } h \to 0.$$

Liegt Stabilität vor, können wir wie beim reinen Anfangswertproblem eine Fehlerabschätzung herleiten. Die Abschätzung in (9.7) ergibt nun für den Rundungsfehler $\tau_j^n = \partial_t u_j^n - \partial_x \bar{\partial}_x u_j^n$

$$|\tau_j^n| \leq Ch^2 \max_{t \in I_n} |u(\cdot, t)|_{\mathcal{C}^4} \quad \text{mit } I_n = (t_n, t_{n+1})$$

und wir erhalten die folgende Fehlerabschätzung.

Theorem 9.5. *Seien U^n und u die Lösungen von* (9.20) *mit $\lambda \leq \frac{1}{2}$ und von* (9.19)*. Dann gilt*

$$\|U^n - u^n\|_{\infty,h} \leq Ct_n h^2 \max_{t \leq t_n} |u(\cdot, t)|_{\mathcal{C}^4} \quad \textit{für } t_n \geq 0.$$

Wir weisen darauf hin, dass wir von v in diesem Fall bestimmte Kompatibilitätsbedingungen mit den Randbedingungen fordern müssen, damit u hinreichend regulär ist und damit die rechte Seite von (9.7) durch $Ch^2|v|_{\mathcal{C}^4}$ beschränkt ist. Diese Bedingungen lauten $v(x) = v''(x) = v^{(iv)}(x) = 0$ für $x = 0, 1$.

Wir stellen fest, dass ein Verfahren von der Form

$$U_j^{n+1} = \sum_p a_p U_{j-p}^n \quad \text{für } j = 1, \ldots, M-1$$

an dieser Stelle nicht geeignet ist, wenn $a_p \neq 0$ für $|p| > 1$ ist, weil dann die Gleichung für einen inneren Gitterpunkt von Ω einen Gitterpunkt außerhalb dieses Intervalls verwendet. In einem solchen Fall muss die Gleichung in der Nähe der Endpunkte modifiziert werden, was die Analyse signifikant komplizierter macht.

Die Stabilitätsforderung $k \leq \frac{1}{2}h^2$, die für das Vorwärts-Euler-Verfahren benutzt wird, ist in der Praxis ziemlich restriktiv. Es wäre wünschenswert, diese etwas abzuschwächen, damit h und k in der gleichen Größenordnung verwendet werden können. Zu diesem Zweck kann man anstelle des oben betrachteten expliziten Verfahrens ein *implizites Verfahren* definieren. Dabei handelt es sich um das *Rückwärts-Euler-Verfahren*

$$
\begin{aligned}
&\bar{\partial}_t U_j^{n+1} = \partial_x \bar{\partial}_x U_j^{n+1} && \text{für } j = 1, \ldots, M-1,\ n \geq 0, \\
&U_0^{n+1} = U_M^{n+1} = 0 && \text{für } n \geq 0, \\
&U_j^0 = V_j = v(x_j) && \text{für } j = 0, \ldots, M.
\end{aligned} \tag{9.21}
$$

Für ein gegebenes U^n kann dies in die Form

$$
\begin{aligned}
&(1+2\lambda)U_j^{n+1} - \lambda(U_{j-1}^{n+1} + U_{j+1}^{n+1}) = U_j^n, \quad j = 1, \ldots, M-1, \\
&U_0^{n+1} = U_M^{n+1} = 0
\end{aligned}
$$

gebracht werden, was ein lineares Gleichungssystem zur Bestimmung von U^{n+1} ist. In Matrixdarstellung kann dies als

$$
B\bar{U}^{n+1} = \bar{U}^n \tag{9.22}
$$

geschrieben werden, wobei $\bar{U}^{n+1}$ und $\bar{U}^n$ Vektoren mit $M-1$ Komponenten sind, die zu den inneren Gitterpunkten gehören. Die Matrix B ist eine diagonaldominante, symmetrische Tridiagonalmatrix

$$
B = \begin{bmatrix}
1+2\lambda & -\lambda & 0 & \ldots & 0 \\
-\lambda & 1+2\lambda & -\lambda & \ddots & \vdots \\
0 & \ddots & \ddots & \ddots & 0 \\
\vdots & \ddots & -\lambda & 1+2\lambda & -\lambda \\
0 & \ldots & 0 & -\lambda & 1+2\lambda
\end{bmatrix}.
$$

Offensichtlich kann das System (9.22) leicht nach $\bar{U}^{n+1}$ aufgelöst werden.

Führen wir den endlichdimensionalen Raum l_h^0 der Vektoren $\{V_j\}_{j=0}^M$ mit $(M+1)$ Komponenten und $V_0 = V_M = 0$ sowie den durch

$$
(B_{kh}V)_j = (1+2\lambda)V_j - \lambda(V_{j-1} + V_{j+1}) = V_j - k\partial_x \bar{\partial}_x V_j, \quad j = 1, \ldots, M-1
$$

definierten Operator B_{kh} auf l_h^0 ein, so können wir das obige System als

$$
B_{kh}U^{n+1} = U^n
$$

oder wiederum mit dem lokalen Lösungsoperator E_k als

$$
U^{n+1} = B_{kh}^{-1}U^n = E_k U^n
$$

schreiben. Wir werden nun sehen, dass dieses Verfahren ohne Einschränkungen für k und h in der Maximumnorm stabil ist, oder genauer gesagt, dass

$$
\|U^{n+1}\|_{\infty,h} \leq \|U^n\|_{\infty,h} \quad \text{für } n \geq 0 \tag{9.23}
$$

gilt. Tatsächlich gilt mit geeignetem j_0

$$\|U^{n+1}\|_{\infty,h} = |U_{j_0}^{n+1}| \le \frac{1}{1+2\lambda}\Big(\lambda\big(|U_{j_0-1}^{n+1}| + |U_{j_0+1}^{n+1}|\big) + |U_{j_0}^{n}|\Big)$$
$$\le \frac{2\lambda}{1+2\lambda}\,\|U^{n+1}\|_{\infty,h} + \frac{1}{1+2\lambda}\,\|U^{n}\|_{\infty,h},$$

woraus unmittelbar (9.23) folgt. Dies impliziert die Stabilitätsabschätzung

$$\|U^n\|_{\infty,h} = \|E_k^n V\|_{\infty,h} \le \|V\|_{\infty,h}. \tag{9.24}$$

Der Lösungsoperator E_k^n ist also in der Maximumnorm stabil und es lässt sich auch zeigen, dass U^n gegen $u(t_n)$ konvergiert. Hier gilt für den Rundungsfehler

$$\tau_j^n = \bar\partial_t u_j^{n+1} - \partial_x\bar\partial_x u_j^{n+1} = O(k+h^2) \quad \text{für } k,h \to 0 \text{ mit } j = 1,\dots,M-1,$$

wobei sich der letzte Ausdruck nicht auf $O(h^2)$ reduziert, da h und k nicht zusammenhängen. Das Konvergenzresultat lässt sich nun folgendermaßen zusammenfassen.

Theorem 9.6. *Seien U^n und u Lösungen von* (9.19) *und* (9.21). *Dann gilt*

$$\|U^n - u^n\|_{\infty,h} \le C\,t_n(h^2+k)\max_{t\le t_n}|u(\cdot,t)|_{\mathcal{C}^4} \quad \textit{für } t_n \ge 0.$$

Beweis. Definieren wir den Fehler durch $z^n = U^n - u^n$, so können wir

$$B_{kh}z^{n+1} = B_{kh}U^{n+1} - B_{kh}u^{n+1} = U^n - (u^{n+1} - k\partial_x\bar\partial_x u^{n+1}) = z^n - k\,\tau^n$$

schreiben, wobei wir τ^n als ein Element von l_h^0 betrachten. Somit gilt

$$z^{n+1} = E_k z^n - kE_k\tau^n$$

und es ergibt sich daraus

$$z^n = -k\sum_{l=0}^{n-1} E_k^{n-l}\tau^l.$$

Die Abschätzung (9.7) wird nun durch

$$\|\tau^n\|_{\infty,h} \le C(h^2+k)\max_{t\in I_{n-1}}|u(\cdot,t)|_{\mathcal{C}^4}$$

ersetzt. Unter Verwendung von (9.24) erhalten wir

$$\|z^n\|_{\infty,h} \le k\sum_{l=0}^{n-1}\|\tau^l\|_{\infty,h} \le C\,t_n(h^2+k)\max_{t\le t_n}|u(\cdot,t)|_{\mathcal{C}^4},$$

was den Beweis vervollständigt. □

Das oben angegebene Konvergenzresultat für das Rückwärts-Euler-Verfahren ist in dem Sinne befriedigend, dass es keine Einschränkungen für das Verhältnis der Gitterkonstanten $\lambda = k/h^2$ erfordert. Andererseits ist das Verfahren in der Zeit nur von der Genauigkeit erster Ordnung und der Fehler in der Zeitdiskretisierung wird dominieren, wenn k nicht sehr viel kleiner als h gewählt wird. Es wäre deshalb wünschenswert, ein stabiles Verfahren zu finden, das in der Genauigkeit hinsichtlich der Zeitdiskretisierung ebenfalls von zweiter Ordnung ist. Ein solches Verfahren ist das *Crank-Nicolson-Verfahren*, das für gewöhnliche Differentialgleichungssysteme in Abschnitt 7.2 eingeführt wurde. Dieses Verfahren ist um den Punkt $(x_j, t_{n+1/2})$ symmetrisch und durch

$$
\begin{aligned}
&\bar{\partial}_t U_j^{n+1} = \tfrac{1}{2}\partial_x\bar{\partial}_x\big(U_j^n + U_j^{n+1}\big) && \text{für } j = 1,\dots,M-1,\ n \geq 0,\\
&U_0^{n+1} = U_M^{n+1} = 0 && \text{für } n \geq 0\\
&U_j^0 = V_j := v(jh) && \text{für } j = 0,\dots,M
\end{aligned}
\tag{9.25}
$$

definiert. Die erste Gleichung kann auch in der Form

$$(I - \tfrac{1}{2}k\partial_x\bar{\partial}_x)U_j^{n+1} = (I + \tfrac{1}{2}k\partial_x\bar{\partial}_x)U_j^n$$

oder als

$$(1+\lambda)U_j^{n+1} - \tfrac{1}{2}\lambda(U_{j-1}^{n+1} + U_{j+1}^{n+1}) = (1-\lambda)U_j^n + \tfrac{1}{2}\lambda(U_{j-1}^n + U_{j+1}^n)$$

geschrieben werden. In Matrixform nimmt dies die Gestalt

$$B\bar{U}^{n+1} = A\bar{U}^n$$

an, wobei $\bar{U}^n$ wieder den zu U^n gehörigen Vektor mit $(M-1)$ Komponenten bezeichnet und sowohl A als auch B symmetrische Tridiagonalmatrizen sind. Die Matrix B ist diagonaldominant

$$B = \begin{bmatrix} 1+\lambda & -\frac{1}{2}\lambda & 0 & \dots & 0 \\ -\frac{1}{2}\lambda & 1+\lambda & -\frac{1}{2}\lambda & \ddots & \vdots \\ 0 & \ddots & \ddots & \ddots & 0 \\ \vdots & \ddots & -\frac{1}{2}\lambda & 1+\lambda & -\frac{1}{2}\lambda \\ 0 & \dots & 0 & -\frac{1}{2}\lambda & 1+\lambda \end{bmatrix}$$

und A ist

$$A = \begin{bmatrix} 1-\lambda & \frac{1}{2}\lambda & 0 & \dots & 0 \\ \frac{1}{2}\lambda & 1-\lambda & \frac{1}{2}\lambda & \ddots & \vdots \\ 0 & \ddots & \ddots & \ddots & 0 \\ \vdots & \ddots & \frac{1}{2}\lambda & 1-\lambda & \frac{1}{2}\lambda \\ 0 & \dots & 0 & \frac{1}{2}\lambda & 1-\lambda \end{bmatrix}.$$

Mit der üblichen Notation gilt auch

$$B_{kh}U^{n+1} = A_{kh}U^n$$

oder

$$U^{n+1} = B_{kh}^{-1}A_{kh}U^n = E_kU^n,$$

wobei analog zu oben

$$\|B_{kh}^{-1}V\|_{\infty,h} \le \|V\|_{\infty,h}$$

ist.

Geht man wie bei der Stabilitätsabschätzung für das Rückwärts-Euler-Verfahren vor, ergibt sich für $\lambda \le 1$, weil die Koeffizienten auf der rechten Seite dann nichtnegativ sind,

$$(1+\lambda)\|U^{n+1}\|_{\infty,h} \le \lambda\|U^{n+1}\|_{\infty,h} + \|U^n\|_{\infty,h}$$

oder

$$\|U^{n+1}\|_{\infty,h} \le \|U^n\|_{\infty,h},$$

was die Stabilität zeigt. Ist jedoch $\lambda > 1$, was der für uns interessante Fall ist, wenn wir h und k von der gleichen Größenordnung wählen wollen, erhalten wir stattdessen

$$(1+\lambda)\|U^{n+1}\|_{\infty,h} \le \lambda\|U^{n+1}\|_{\infty,h} + (2\lambda-1)\|U^n\|_{\infty,h},$$

was wegen $2\lambda - 1 > 1$ nicht zur Stabilität in der Maximumnorm führt. Wie vorhin ergibt sich für $\lambda \le 1$ unmittelbar eine Konvergenzabschätzung der Ordnung $O(k^2 + h^2) = O(h^2)$.

Damit wir uns mit dem Fall $\lambda > 1$ beschäftigen können, kommen wir nun stattdessen zu einer Analyse in einer Norm, die der in l_2 ähnlich ist. Wir führen also für die Vektoren $V = (V_0, \ldots, V_M)^T$ das Skalarprodukt

$$(V, W)_h = h\sum_{j=0}^{M} V_jW_j$$

und die zugehörige Norm

$$\|V\|_{2,h} = (V, V)_h^{1/2} = \Big(h\sum_{j=0}^{M} V_j^2\Big)^{1/2}$$

ein. Wir bezeichnen mit $l_{2,h}^0$ den Raum l_h^0, der mit diesem Skalarprodukt und dieser Norm versehen ist, und stellen fest, dass dieser Raum durch die $M-1$ Vektoren φ_p, $p = 1, \ldots, M-1$ mit den Komponenten

$$\varphi_{p,j} = \sqrt{2}\sin(\pi pjh) \quad \text{für } j = 0, \ldots, M$$

aufgespannt wird und dass diese eine Orthonormalbasis bezüglich des obigen Skalarproduktes bilden (siehe Problemstellung 9.7). Es gilt also

$$(\varphi_p, \varphi_q)_h = \delta_{pq} = \begin{cases} 1 & \text{für } p = q, \\ 0 & \text{für } p \neq q. \end{cases}$$

Die φ_p sind Eigenfunktionen des Finite-Differenzen-Operators $-\partial_x \bar{\partial}_x$ mit

$$-\partial_x \bar{\partial}_x \varphi_{p,j} = \frac{2}{h^2}(1 - \cos(\pi p h))\varphi_{p,j} \quad \text{für } j = 1, \ldots, M-1.$$

Wir werden die Stabilität nun im Zusammenhang mit den drei bisher betrachteten Differenzenverfahren diskutieren. Seien V die gegebenen Anfangsdaten in $l^0_{2,h}$. Dann gilt

$$V = \sum_{p=1}^{M-1} \hat{V}_p \varphi_p, \quad \text{wobei } \hat{V}_p = (V, \varphi_p)_h \text{ ist.}$$

Das Vorwärts-Euler-Verfahren liefert dann

$$U^1_j = V_j + k\partial_x \bar{\partial}_x V_j = \sum_{p=1}^{M-1} \hat{V}_p \big(1 - 2\lambda(1 - \cos(\pi p h))\big)\varphi_{p,j}, \ j = 1, \ldots, M-1$$

mit $U^1_0 = U^1_M = 0$ oder allgemeiner

$$U^n_j = \sum_{p=1}^{M-1} \hat{V}_p \tilde{E}(\pi p h)^n \varphi_{p,j}, \quad j = 0, \ldots, M, \tag{9.26}$$

wobei $\tilde{E}(\xi)$ das charakteristische Polynom des lokalen, diskreten Lösungsoperators E_k mit

$$\tilde{E}(\xi) = 1 - 2\lambda + 2\lambda \cos \xi$$

ist. Wegen der Parsevalschen Gleichung gilt somit

$$\|U^n\|_{2,h} = \Big(\sum_{p=1}^{M-1} \hat{V}_p^2 \tilde{E}(\pi p h)^{2n} \Big)^{1/2} \leq \max_p |\tilde{E}(\pi p h)^n| \, \|V\|_{2,h},$$

wobei Gleichheit für ein geeignet gewähltes V angenommen wird. Nun gilt für $1 \leq p \leq M-1$

$$\begin{aligned} |\tilde{E}(\pi p h)| &= \max\big\{|1 - 2\lambda(1 - \cos(\pi h))|, |1 - 2\lambda(1 - \cos(\pi(M-1)h))|\big\} \\ &= \max\big\{|1 - 2\lambda(1 - \cos(\pi h))|, |1 - 2\lambda(1 + \cos(\pi h))|\big\}. \end{aligned}$$

Die Ungleichung $\max_p |\tilde{E}(ph\pi)| \leq 1$ gilt also für kleine h genau dann, wenn $4\lambda - 1 \leq 1$ oder $\lambda \leq \frac{1}{2}$ erfüllt ist. In diesem Falle folgt

$$\|U^n\|_{2,h} \le \|V\|_{2,h}. \tag{9.27}$$

Folglich ist das Vorwärts-Euler-Verfahren in $l^0_{2,h}$ genau dann stabil, wenn $\lambda \le \frac{1}{2}$ erfüllt ist. Das sind die gleichen Bedingungen wie für die Maximumnorm.

Die entsprechende Analyse für das Rückwärts-Euler-Verfahren führt auf (9.26), wobei nun

$$\tilde{E}(\xi) = \frac{1}{1 + 2\lambda(1 - \cos\xi)}$$

ist. In diesem Fall gilt $0 \le \tilde{E}(\pi ph) \le 1$ für alle p und λ und (9.27) ist daher für jeden Wert von λ gültig.

Analog dazu gilt (9.26) auch für das Crank-Nicolson-Verfahren mit

$$\tilde{E}(\xi) = \frac{1 - \lambda(1 - \cos\xi)}{1 + \lambda(1 - \cos\xi)},$$

und wir stellen nun fest, dass $|\tilde{E}(\xi)| \le 1$ für alle ξ und jedes $\lambda > 0$ erfüllt ist. Die Fourier-Analyse zeigt also die Stabilität in $l^0_{2,h}$ für jedes λ. Die Konvergenz folgt wiederum aus der Standardmethode und es ergibt sich folgendes Theorem.

Theorem 9.7. *Seien U^n und u Lösungen von* (9.25) *und* (9.19). *Dann gilt*

$$\|U^n - u^n\|_{2,h} \le C\, t_n (h^2 + k^2) \max_{t \le t_n} |u(\cdot, t)|_{\mathcal{C}^6} \quad \textit{für } t_n \ge 0.$$

Beweis. Wir schreiben für den Rundungsfehler

$$\tau^n_j = \bar{\partial}_t u^{n+1}_j - \partial_x \bar{\partial}_x \frac{u^n_j + u^{n+1}_j}{2} = \Big(\bar{\partial}_t u^{n+1}_j - u_t(x_j, t_{n+1/2})\Big)$$
$$+ \partial_x \bar{\partial}_x \Big(\frac{u^n_j + u^{n+1}_j}{2} - u^{n+1/2}_j\Big) + \Big(\partial_x \bar{\partial}_x u^{n+1/2}_j - u_{xx}(x_j, t_{n+1/2})\Big).$$

Somit ergibt sich wie vorhin unter Verwendung von Taylor-Entwicklungen

$$\|\tau^n\|_{2,h} \le C(h^2 + k^2) \max_{t \in I_n} |u(\cdot, t)|_{\mathcal{C}^6}.$$

Wie vorhin erfüllt der Fehler $z^n_j = U^n_j - u^n_j$ die Gleichung

$$z^{n+1} = U^{n+1} - u^{n+1} = E_k U^n - B^{-1}_{kh} B_{kh} u^{n+1} = E_k z^n - k\; B^{-1}_{kh} \tau^n$$

oder

$$z^n = -k \sum_{l=0}^{n-1} E_k^{n-1-l} B^{-1}_{kh} \tau^l,$$

woraus das gesuchte Resultat wegen der Stabilität des Crank-Nicolson-Operators E^n_k und der Beschränktheit von B^{-1}_{kh} folgt. □

Das Vorwärts- und Rückwärts-Euler-Verfahren sowie das Crank-Nicolson-Verfahren können als Spezialfälle des *θ-Verfahrens* betrachtet werden, das durch

$$\bar{\partial}_t U_j^{n+1} = \theta \partial_x \bar{\partial}_x U_j^{n+1} + (1-\theta)\partial_x \bar{\partial}_x U_j^n, \quad j = 1, \ldots, M-1 \tag{9.28}$$

definiert ist. Mit $\theta = 0$ ergibt sich das Vorwärts-Euler-Verfahren, mit $\theta = 1$ das Rückwärts-Euler-Verfahren und mit $\theta = 1/2$ das Crank-Nicolson-Verfahren. Die Gleichung (9.28) kann in der Form

$$(I - \theta k \partial_x \bar{\partial}_x) U^{n+1} = (I + (1-\theta) k \partial_x \bar{\partial}_x) U^n$$

geschrieben werden. Diesmal ergibt sich für die charakteristische Gleichung

$$\tilde{E}(\xi) = \frac{1 - 2(1-\theta)\lambda(1-\cos\xi)}{1 + 2\theta\lambda(1-\cos\xi)}.$$

Unter der Annahme $0 \le \theta \le 1$ gilt $\tilde{E}(\xi) \le 1$ für $\xi \in \mathbf{R}$ und die Stabilitätsforderung reduziert sich auf

$$\min_\xi \tilde{E}(\xi) = \frac{1 - 4(1-\theta)\lambda}{1 + 4\theta\lambda} \ge -1$$

oder

$$(1 - 2\theta)\lambda \le \tfrac{1}{2}.$$

Folglich ist das θ-Verfahren unbedingt stabil in $l^0_{2,h}$, d. h. stabil in $l^0_{2,h}$ für alle λ mit $\theta \ge 1/2$, während für $\theta < 1/2$ Stabilität genau dann vorliegt, wenn

$$\lambda \le \frac{1}{2(1-2\theta)}$$

erfüllt ist.

9.3 Problemstellungen

Problem 9.1. Zeigen Sie die Äquivalenz der Definitionen (9.15) und (9.16) in der Genauigkeit der Ordnung r. Verwenden Sie die alternative Definition (9.16) um zu beweisen, dass die Genauigkeit von (9.4) von der Ordnung 4 ist, wenn $\lambda = 1/6$ gewählt wird.

Problem 9.2. Formulieren und beweisen Sie eine zu Theorem 9.2 analoge Aussage in zwei räumlichen Dimensionen.

Problem 9.3. Es sei (a_{jk}) eine symmetrische, positiv definite 2×2-Matrix. Zur Lösung des Anfangswertproblems

$$\frac{\partial u}{\partial t} = \sum_{j,k=1}^{2} a_{jk} \frac{\partial^2 u}{\partial x_j \partial x_k} \qquad \text{in } \mathbf{R}^2 \times \mathbf{R}_+,$$
$$u(\cdot, 0) = v \qquad \text{in } \mathbf{R}^2$$

soll das finite Differenzenverfahren

$$\partial_t U_{ij}^n = \sum_{k,l=1}^{2} a_{kl}\, \partial_{x_k} \bar{\partial}_{x_l} U_{ij}^n$$

angewendet werden.
(a) Geben Sie Bedingungen für die Koeffizienten an, die für die Stabilität des Verfahrens in der Matrixnorm hinreichend sind.
(b) Ist das Verfahren in $l_{2,h}$ stabil?

Problem 9.4. Bestimmen Sie einen expliziten Fünfpunkt-Finite-Differenzen-Operator für (9.1) von der Form (9.8) (d. h. mit fünf Termen auf der rechten Seite von (9.8)) mit der Genauigkeitsordnung 4. Diskutieren Sie die Stabilität dieses Operators.

Problem 9.5. Formulieren Sie ein finites Differenzenverfahren für

$$u_t = \Delta u \qquad \text{in } \mathbf{R}^2 \times \mathbf{R}_+,$$
$$u(\cdot, 0) = v \qquad \text{in } \mathbf{R}^2,$$

für das

$$\|U^n - u^n\|_{\infty,h} = O(h^4) \quad \text{für } h \to 0 \text{ gilt.}$$

Problem 9.6. Betrachten Sie das Zweischrittverfahren (9.18). Es sei $U^0 = V$ und $U^1 = W$ mit $V, W \in L_2(\mathbf{R})$. Zeigen Sie, dass

$$\hat{U}^n(\xi) = c_1(\xi)\tau_1(\xi)^n + c_2(\xi)\tau_2(\xi)^n$$

gilt, wobei $\tau_{1,2}(\xi)$ die Wurzeln der Gleichung

$$\tau^2 + 4\lambda(1 - \cos\xi)\tau - 1 = 0 \quad \text{mit } \lambda = k/h^2$$

sind und $c_1(\xi)$ und $c_2(\xi)$ aus

$$c_1(\xi) + c_2(\xi) = \hat{V}(\xi), \quad c_1(\xi)\tau_1(\xi) + c_2(\xi)\tau_2(\xi) = \hat{W}(\xi)$$

bestimmt werden. Verwenden Sie dies, um $\|U^n\| \to \infty$ für $n \to \infty$ für jedes $\lambda > 0$ und folglich die Instabilität von (9.18) zu zeigen.

Problem 9.7. Die Funktionen $\{\varphi_p\}_{p=1}^{M-1}$ seien durch (9.2) definiert. Zeigen Sie, dass diese eine Orthonormalbasis für $l_{2,h}^0$ bilden und Eigenfunktionen des Differenzenoperators $-\partial_x \bar{\partial}_x$ mit den Eigenwerten $2h^{-2}(1 - \cos(\pi p h))$ sind. Vergleichen Sie diese Eigenfunktionen und Eigenwerte mit den Eigenfunktion und Eigenwerten von $-d^2/dx^2$. Beachten Sie, dass eines der φ_p das Gegenbeispiel für die Stabilität zu Beginn von Abschnitt 9.2 liefert.

Problem 9.8. (Ein diskretes Maximumprinzip.) Sei $\Omega \subset \mathbf{R}$ ein beschränktes Intervall und $I = (0, T]$. Zeigen Sie, dass für $\lambda = kh^{-2} \leq \frac{1}{2}$ und

$$\partial_t U_j^n - \partial_x \bar{\partial}_x U_j^n \leq 0 \quad \text{für } (x_j, t_n) \in \Omega \times I$$

U_j^n sein Maximum auf dem parabolischen Rand Γ_p annimmt (vgl. Theorem 8.6). Hinweis: Verwenden Sie das Argument, das auf (9.6) führte. Beweisen Sie ein ähnliches Resultat für das Rückwärts-Euler-Verfahren.

Problem 9.9. Wir wissen, dass alle Normen auf dem endlichdimensionalen Raum l_h^0 äquivalent sind. Zeigen Sie beispielsweise, dass

$$\|V\|_{2,h} \leq \|V\|_{\infty,h} \leq h^{-1/2} \|V\|_{2,h} \quad \text{für } V \in l_h^0$$

erfüllt ist und dass diese Ungleichungen scharf sind. Beachten Sie, dass diese Äquialenz nicht gleichmäßig in h ist und für $h \to 0$ verloren geht, d. h. wenn die Dimension von l_h^0 gegen unendlich strebt. Die zweite oben genannte Ungleichung besitzt denselben Charakter wie die inverse Ungleichung (6.36), die eine stärkere Norm ($\|\cdot\|_{\infty,h}$) zu einer schwächeren Norm ($\|\cdot\|_{2,h}$) in Beziehung setzt.

Problem 9.10. Zeigen Sie, dass die Funktion $\varphi(x) = \mathrm{e}^{\mathrm{i}x\xi}$ eine Eigenfunktion der Differential- und Differenzenoperatoren $\partial/\partial x$, ∂_x und $\bar{\partial}_x$ ist.

Problem 9.11. (Übung am Rechner.) Betrachten Sie das Anfangs-Randwertproblem (9.19) mit $v(x) = \sin(\pi x) - \sin(3\pi x)$. Wenden Sie das Vorwärts-Euler-Verfahren mit $h = 1/10$ und $k = 1/600, 1/300, 1/100$ an. Wenden Sie auch das Crank-Nicolson-Verahren mit $h = k = 1/10$ an. Berechnen Sie den Fehler an der Stelle $(1/2, 1)$.

10

Die Methode der finiten Elemente für ein parabolisches Problem

In diesem Kapitel untersuchen wir die Approximation von Lösungen der Modell-Wärmeleitungsgleichung in zwei räumlichen Dimensionen mithilfe der Galerkin-Methode, die stückweise lineare Testfunktionen benutzt. In Abschnitt 10.1 betrachten wir die Diskretisierung nur hinsichtlich der räumlichen Variablen. In dem folgenden Abschnitt 10.2 untersuchen wir einige vollständig diskrete Schemata.

10.1 Die semidiskrete Galerkin-Methode der finiten Elemente

Sei $\Omega \subset \mathbf{R}^2$ eine abgeschlossene, konvexe Menge mit glattem Rand Γ. Wir betrachten das Anfangs-Randwertproblem

$$
\begin{aligned}
& u_t - \Delta u = f && \text{in } \Omega \times \mathbf{R}_+, \\
& u = 0 && \text{auf } \Gamma \times \mathbf{R}_+, \\
& u(\cdot, 0) = v && \text{in } \Omega,
\end{aligned}
\tag{10.1}
$$

wobei u_t für $\partial u/\partial t$ und Δ für den Laplace-Operator $\partial^2/\partial x_1^2 + \partial^2/\partial x_2^2$ steht. Im ersten Schritt werden wir die Lösung $u(x,t)$ mithilfe einer Funktion $u_h(x,t)$ approximieren, die für jedes feste t eine stückweise lineare Funktion von x über einer Triangulation $\mathcal{T}_h$ von Ω ist und somit von einer endlichen Anzahl von Parametern abhängt.

Es bezeichne also $\mathcal{T}_h = \{K\}$ eine Triangulation von Ω von einem im Abschnitt 5.2 betrachteten Typ. Es seien $\{P_j\}_{j=1}^{M_h}$ die inneren Knoten von $\mathcal{T}_h$. Darüber hinaus bezeichnen wir mit S_h die stückweise linearen Funktionen auf $\mathcal{T}_h$, die auf $\partial\Omega$ verschwinden. Dabei sei $\{\Phi_j\}_{j=1}^{M_h}$ die zu den Knoten $\{P_j\}_{j=1}^{M_h}$ gehörige Standardbasis von S_h. Erinnern Sie sich an die Definition (5.28) der Interpolierten $I_h : \mathcal{C}_0(\bar{\Omega}) \to S_h$ und an die Fehlerschranken (5.34) mit $r = 2$.

Um eine approximative Lösung des Anfangs-Randwertproblems (10.1) zu definieren, schreiben wir dieses wie in Abschnitt 8.3 zunächst in schwacher

Form, d. h. mit den obigen Definitionen als

$$(u_t, \varphi) + a(u, \varphi) = (f, \varphi) \quad \forall \varphi \in H_0^1,\ t > 0. \tag{10.2}$$

Wir stellen nun das Approximationsproblem, ein für jedes t zu S_h gehöriges $u_h(t) = u_h(\cdot, t)$ so zu bestimmen, dass

$$\begin{aligned} (u_{h,t}, \chi) + a(u_h, \chi) &= (f, \chi) \quad \forall \chi \in S_h,\ t > 0, \\ u_h(0) &= v_h \end{aligned} \tag{10.3}$$

erfüllt ist, wobei $v_h \in S_h$ eine Approximation von v ist. Da wir nur in den räumlichen Variablen diskretisiert haben, wird dies als ein *räumlich semidiskretes* Problem bezeichnet. Im nächsten Abschnitt werden wir auch in der Zeitvariablen diskretisieren, was zu vollständig diskreten Schemata führt.

Das semidiskrete Problem kann mithilfe der Basis $\{\Phi_j\}_{j=1}^{M_h}$ folgendermaßen gestellt werden: Gesucht sind die Koeffizienten $\alpha_j(t)$ in

$$u_h(x,t) = \sum_{j=1}^{M_h} \alpha_j(t) \Phi_j(x),$$

sodass

$$\sum_{j=1}^{M_h} \alpha_j'(t)(\Phi_j, \Phi_k) + \sum_{j=1}^{M_h} \alpha_j(t) a(\Phi_j, \Phi_k) = (f(t), \Phi_k), \quad k = 1, \ldots, M_h$$

gilt. Dabei bezeichnen die γ_j die Knotenwerte der gegebenen Anfangsfunktion v_h mit

$$\alpha_j(0) = \gamma_j, \quad j = 1, \ldots, M_h.$$

In Matrixdarstellung kann dies in der Form

$$B\alpha'(t) + A\alpha(t) = b(t) \quad \text{für } t > 0 \quad \text{mit } \alpha(0) = \gamma \tag{10.4}$$

ausgedrückt werden, wobei $B = (b_{kj})$ die Massenmatrix mit den Elementen $b_{kj} = (\Phi_j, \Phi_k)$, $A = (a_{kj})$ die Steifigkeitsmatrix mit $a_{kj} = a(\Phi_j, \Phi_k)$, $b = (b_k)$ der Vektor mit den Elementen $b_k = (f, \Phi_k)$, $\alpha(t)$ der Vektor der Unbekannten $\alpha_j(t)$ und $\gamma = (\gamma_j)$ ist. Die Dimension dieser Objekte ist gleich der Anzahl M_h der inneren Knoten von $\mathcal{T}_h$.

Wir wissen aus Abschnitt 5.2, dass die Steifigkeitsmatrix A symmetrisch positiv definit ist und dies auch für die Massenmatrix B zutrifft, weil

$$\sum_{k,j=1}^{M_h} \xi_j \xi_k (\Phi_j, \Phi_k) = \Big\| \sum_{j=1}^{M_h} \xi_j \Phi_j \Big\|^2 \geq 0$$

gilt und weil die Gleichheit nur auftreten kann, wenn der Vektor $\xi = 0$ ist. Insbesondere ist B invertierbar, weshalb das obige gewöhnliche Differentialgleichungssystem in der Form

$$\alpha'(t) + B^{-1}A\alpha(t) = B^{-1}b(t) \quad \text{für } t > 0 \quad \text{mit } \alpha(0) = \gamma$$

geschrieben werden kann und somit offensichtlich für positive t eine eindeutige Lösung besitzt.

Wir beginnen unsere Analyse mit Betrachtungen zur Stabilität der semidiskreten Methode. Wegen $u_h \in S_h$ können wir in (10.3) $\chi = u_h$ wählen, wodurch wir

$$(u_{h,t}, u_h) + a(u_h, u_h) = (f, u_h) \quad \text{für } t > 0$$

oder, weil der erste Term gleich $\frac{1}{2}\frac{d}{dt}\|u_h\|^2$ und der zweite nichtnegativ ist,

$$\tfrac{1}{2}\frac{d}{dt}\|u_h\|^2 = \|u_h\|\frac{d}{dt}\|u_h\| \le \|f\|\,\|u_h\|$$

erhalten. Daraus ergibt sich

$$\frac{d}{dt}\|u_h\| \le \|f\|,$$

was nach Integration auf die Stabilitätsabschätzung

$$\|u_h(t)\| \le \|v_h\| + \int_0^t \|f\|\,\mathrm{d}s \tag{10.5}$$

führt.

Damit wir Gleichung (10.3) in Operatorform aufschreiben können, führen wir einen *diskreten Laplace-Operator* Δ_h ein, den wir als Operator aus S_h in sich selbst betrachten, der durch

$$(-\Delta_h\psi, \chi) = a(\psi, \chi) \quad \forall \psi, \chi \in S_h \tag{10.6}$$

definiert ist. Dieses diskrete Analogon zur Greenschen Formel definiert $\Delta_h\psi = \sum_{j=1}^{M_h} d_j\Phi_j$ eindeutig durch

$$\sum_{j=1}^{M_h} d_j(\Phi_j, \Phi_k) = -a(\psi, \Phi_k), \quad k = 1, \ldots, M_h,$$

weil die Matrix dieses Systems die positiv definite Massenmatrix ist, die uns bereits oben begegnet ist. Man kann sich leicht davon überzeugen, dass der Operator Δ_h selbstadjungiert und $-\Delta_h$ positiv definit in S_h bezüglich des L_2-Skalarproduktes ist (siehe Problemstellung 10.3). Die Gleichung (10.3) kann nun in der Form

$$(u_{h,t} - \Delta_h u_h - P_h f, \chi) = 0 \quad \forall \chi \in S_h$$

geschrieben werden, wobei P_h die L_2-Projektion auf S_h bezeichnet. Wenn wir andererseits beachten, dass der erste Faktor in S_h ist, sodass χ gleich diesem Ausdruck gewählt werden kann, folgt

$$u_{h,t} - \Delta_h u_h = P_h f \quad \text{für } t > 0 \quad \text{mit } u_h(0) = v_h. \tag{10.7}$$

Wir bezeichnen mit $E_h(t)$ den Lösungsoperator im homogenen Fall der semidiskreten Gleichung (10.7) mit $f = 0$. $E_h(t)$ ist also der Operator, der die Anfangsdaten $u_h(0) = v_h$ in die Lösung $u_h(t)$ zur Zeit t überführt, sodass $u_h(t) = E_h(t)v_h$ gilt. Man kann dann leicht zeigen (siehe Duhamel-Prinzip (8.22)), dass die Lösung des Anfangswertproblems (10.7)

$$u_h(t) = E_h(t)v_h + \int_0^t E_h(t-s)P_h f(s)\,\mathrm{d}s \tag{10.8}$$

ist. Wir bemerken nun, dass aus (10.5) die Stabilität von $E_h(t)$ in L_2 folgt, es gilt also

$$\|E_h(t)v_h\| \le \|v_h\| \quad \forall v_h \in S_h. \tag{10.9}$$

Da P_h in L_2 auch die Norm eins hat, bestätigt dies zusammen mit (10.8) die Stabilitätsabschätzung (10.5) für die inhomogene Gleichung. Es ist also wirklich ausreichend, die Stabilität für die homogene Gleichung zu beweisen.

Wir werden nun die folgende Abschätzung für die Abweichung der Lösung des semidiskreten Problems gegenüber der Lösung des kontinuierlichen Problems beweisen.

Theorem 10.1. *Seien u_h und u die Lösungen von* (10.3) *und* (10.1)*. Dann gilt*

$$\|u_h(t) - u(t)\| \le \|v_h - v\| + Ch^2\Big(\|v\|_2 + \int_0^t \|u_t\|_2\,\mathrm{d}s\Big) \quad \textit{für } t \ge 0.$$

Hier fordern wir wie gewöhnlich, dass die Lösung des kontinuierlichen Problems die Regularität besitzt, die implizit durch die Anwesenheit der Normen auf der rechten Seite angenommen wird. Beachten Sie außerdem, dass die Gleichung (5.31) für $v_h = I_h v$ zeigt, dass

$$\|v_h - v\| \le Ch^2\|v\|_2 \tag{10.10}$$

gilt, was bedeutet, dass der erste Term auf der rechten Seite durch den zweiten dominiert wird. Dasselbe trifft auf den Fall $v_h = P_h v$ zu, wobei P_h die orthogonale Projektion des L_2 auf S_h ist, weil diese Wahl die beste Approximation von v in S_h bezüglich der L_2-Norm darstellt (siehe (5.39)). Eine andere Wahl optimaler Ordnung ist $v_h = R_h v$, wobei R_h die elliptische (Ritzsche) Projektion auf S_h ist, die in (5.49) durch

$$a(R_h v, \chi) = a(v, \chi) \quad \forall \chi \in S_h \tag{10.11}$$

definiert wurde. Deshalb ist $R_h v$ die Finite-Elemente-Approximation der Lösung des elliptischen Problems, dessen exakte Lösung v ist. Wir wiederholen die Fehlerabschätzungen aus Theorem 5.5:

$$\|R_h v - v\| + h|R_h v - v|_1 \le Ch^s\|v\|_s \quad \text{für } s = 1, 2. \tag{10.12}$$

Wir kommen nun zum

Beweis von Theorem 10.1. Im Hauptteil des Beweises werden wir die Lösung des semidiskreten Problems mit der elliptischen Projektion der exakten Lösung vergleichen. Wir schreiben dazu

$$u_h - u = (u_h - R_h u) + (R_h u - u) = \theta + \rho. \tag{10.13}$$

Der zweite Term kann unter Verwendung von (10.12) offensichtlich leicht durch

$$\|\rho(t)\| \le Ch^2 \|u(t)\|_2 = Ch^2 \Big\| v + \int_0^t u_t \, \mathrm{d}s \Big\|_2 \le Ch^2 \Big(\|v\|_2 + \int_0^t \|u_t\|_2 \, \mathrm{d}s \Big)$$

abgeschätzt werden. Zum Abschätzen von θ stellen wir fest, dass

$$\begin{aligned} (\theta_t, \chi) + a(\theta, \chi) &= (u_{h,t}, \chi) + a(u_h, \chi) - (R_h u_t, \chi) - a(R_h u, \chi) \\ &= (f, \chi) - (R_h u_t, \chi) - a(u, \chi) = (u_t - R_h u_t, \chi) \end{aligned} \tag{10.14}$$

oder

$$(\theta_t, \chi) + a(\theta, \chi) = -(\rho_t, \chi) \quad \forall \chi \in S_h \tag{10.15}$$

gilt. Bei dieser Herleitung haben wir (10.3), (10.2), die Definition von R_h in (10.11) und die leicht zu überprüfende Tatsache benutzt, dass dieser Operator mit der Zeitableitung kommutiert, d. h. $R_h u_t = (R_h u)_t$ gilt. Wir können nun die Stabilitätsabschätzung (10.5) auf (10.15) anwenden und erhalten

$$\|\theta(t)\| \le \|\theta(0)\| + \int_0^t \|\rho_t\| \, \mathrm{d}s.$$

Hier gilt

$$\|\theta(0)\| = \|v_h - R_h v\| \le \|v_h - v\| + \|R_h v - v\| \le \|v_h - v\| + Ch^2 \|v\|_2$$

und ferner

$$\|\rho_t\| = \|R_h u_t - u_t\| \le Ch^2 \|u_t\|_2.$$

Zusammen beweisen diese Abschätzungen das Theorem. □

Wir sehen aus dem Beweis von Theorem 10.1, dass die Fehlerabschätzung für das semidiskrete parabolische Problem eine Konsequenz aus der Stabilität dieses Problems und der Fehlerabschätzung des elliptischen Problems, ausgedrückt in der Form $\rho = (R_h - I)u$, ist.

Wiederholen wir das Maximumprinzip für parabolische Gleichungen, Theorem 8.7, stellen wir sofort fest, dass für den Lösungsoperator $E(t)$ im homogenen Fall des Anfangs-Randwertproblems (10.1) die Abschätzung $\|E(t)v\|_{\mathcal{C}} \le \|v\|_{\mathcal{C}}$ für $t \ge 0$ gilt. Das zugehörige Maximumprinzip gilt für das Finite-Elemente-Problem nicht. Man kann allerdings zeigen, dass im Falle einer quasiuniformen Familie $\{\mathcal{T}_h\}$ von Triangulationen (vgl. (5.52)) für ein $C > 1$

$$\|E_h(t)v_h\|_{\mathcal{C}} \le C\|v_h\|_{\mathcal{C}} \quad \text{für } t \ge 0$$

gilt. Dies kann mit der Fehlerabschätzung (5.53) für das stationäre Problem kombiniert werden, um eine Fehlerabschätzung für das parabolische Problem in der Maximumnorm beweisen zu können.

In diesem Zusammenhang erwähnen wir eine Variante des semidiskreten Problems (10.2), für das mitunter ein Maximumprinzip gilt. Dabei handelt es sich um die *Lumped-Mass-Methode*. Zu deren Definition ersetzen wir die Matrix B in (10.4) durch eine Diagonalmatrix $\bar{B}$, in der sich die Diagonalelemente aus der Summe der Zeilenelemente ergeben. Man kann zeigen, dass diese Methode auch durch

$$(u_{h,t}, \chi)_h + a(u_h, \chi) = (f, \chi) \quad \forall \chi \in S_h \quad \text{für } t > 0 \tag{10.16}$$

definiert werden kann, wobei das Skalarprodukt im ersten Term durch Berechnung des ersten Terms in (10.2) mithilfe der Knotenquadraturregel (5.64) zustande gekommen ist. Für diese Methode kann man eine Fehlerabschätzung der Ordnung $O(h^2)$ herleiten, die der aus Theorem 10.1 ähnelt. Wenn wir nun annehmen, dass alle Triangulationswinkel kleiner gleich $\pi/2$ sind, dann sind die Nichtdiagonalelemente der Steifigkeitsmatrix A nichtpositiv. Daraus folgt, dass für den Lösungsoperator $\bar{E}_h(t)$ des modifizierten Problems

$$\|\bar{E}_h(t)v_h\|_{\mathcal{C}} \le \|v_h\|_{\mathcal{C}} \quad \text{für } t \ge 0$$

gilt.

Kehren wir zur gewöhnlichen Galerkin-Methode (10.1) zurück. Wir beweisen nun die folgende Abschätzung des Fehlers im Gradienten.

Theorem 10.2. *Unter den Annahmen von Theorem 10.1 gilt für* $t \ge 0$

$$|u_h(t) - u(t)|_1 \le |v_h - v|_1 + Ch\Big\{\|v\|_2 + \|u(t)\|_2 + \Big(\int_0^t \|u_t\|_1^2 \,\mathrm{d}s\Big)^{1/2}\Big\}.$$

Beweis. Wie vorhin schreiben wir den Fehler in der Form (10.13). An dieser Stelle gilt wegen (10.12)

$$|\rho(t)|_1 = |R_h u(t) - u(t)|_1 \le Ch\|u(t)\|_2.$$

Um $\nabla\theta$ abzuschätzen, benutzen wir wiederum (10.15), nun mit $\chi = \theta_t$. Wir erhalten

$$\|\theta_t\|^2 + \tfrac{1}{2}\frac{d}{dt}|\theta|_1^2 = -(\rho_t, \theta_t) \le \tfrac{1}{2}(\|\rho_t\|^2 + \|\theta_t\|^2),$$

sodass

$$\frac{d}{dt}|\theta|_1^2 \le \|\rho_t\|^2$$

oder

$$|\theta(t)|_1^2 \le |\theta(0)|_1^2 + \int_0^t \|\rho_t\|^2 \,\mathrm{d}s \le \big(|v_h - v|_1 + |R_h v - v|_1\big)^2 + \int_0^t \|\rho_t\|^2 \,\mathrm{d}s$$

gilt. Hieraus schlussfolgern wir wegen $a^2 + b^2 \le (|a| + |b|)^2$ und (10.12)

$$|\theta(t)|_1 \le |v_h - v|_1 + Ch\Big\{\|v\|_2 + \Big(\int_0^t \|u_t\|_1^2 \,\mathrm{d}s\Big)^{1/2}\Big\}, \tag{10.17}$$

was den Beweis abschließt. □

Beachten Sie, dass im Falle $v_h = I_h v$ oder $R_h v$

$$|v_h - v|_1 \le Ch\|v\|_2$$

gilt, sodass der erste Term auf der rechten Seite der Gleichung aus Theorem 10.2 durch den zweiten dominiert wird.

Wir machen nun bezüglich $\theta = u_h - R_h u$ folgende Beobachtung: Wenn wir $v_h = R_h v$ so wählen, dass $\theta(0) = 0$ ist, dann gilt zusätzlich zu (10.17)

$$|\theta(t)|_1 \le \Big(\int_0^t \|\rho_t\|_2 \,\mathrm{d}s\Big)^{1/2} \le Ch^2\Big(\int_0^t \|u_t\|_2^2 \,\mathrm{d}s\Big)^{1/2}.$$

Somit ist der Gradient von θ von zweiter Ordnung $O(h^2)$, während der Gradient des Gesamtfehlers für $h \to 0$ lediglich von der Ordnung $O(h)$ ist. Folglich ist ∇u_h eine bessere Approximation für $\nabla R_h u$ als es für ∇u möglich ist. Dies ist ein Beispiel für ein Phänomen, das manchmal als *Superkonvergenz* bezeichnet wird.

Der oben eingeführte Lösungsoperator $E_h(t)$ besitzt ebenfalls Glättungseigenschaften, die denen in Abschnitt 8.2 für das kontinuierliche Problem entsprechen, sodass beispielsweise

$$|E_h(t)v_h|_1 \le Ct^{-1/2}\|v_h\| \quad \text{für } t > 0, \quad v_h \in S_h$$

und

$$\Big\|D_t^k E_h(t)v_h\Big\| = \|\Delta_h^k E_h(t)v_h\| \le C_k t^{-k}\|v_h\| \quad \text{für } t > 0,\ v_h \in S_h \tag{10.18}$$

gilt. Solche Resultate können verwendet werden, um beispielsweise die folgenden Fehlerabschätzungen für die homogene Gleichung im Falle nichtglatter Daten zu zeigen.

Theorem 10.3. *Ws gelte $f = 0$ und seien u_h und u die Lösungen von* (10.3) *beziehungsweise* (10.1)*, wobei die Anfangsdaten für* (10.3) *als $v_h = P_h v$ gewählt werden. Dann gilt*

$$\|u_h(t) - u(t)\| \le Ch^2 t^{-1}\|v\| \quad \textit{für } t > 0.$$

Den Beweis überlassen wir dem Leser als Übung (siehe Problemstellung 10.4). Dieses Resultat zeigt, dass die Konvergenzrate für $t > 0$ von der Ordnung $O(h^2)$ ist. Dies trifft auch dann zu, wenn wir lediglich annehmen, dass v in L_2 ist.

Die oben vorgestellte Theorie lässt sich unter geeigneten Annahmen für die Regularität der Lösung einfach auf finite Elemente höherer Ordnung übertragen. Gilt also im Raum der finiten Elemente

$$\|R_h w - w\| \le Ch^r \|w\|_r \quad \forall w \in H^r \cap H_0^1, \tag{10.19}$$

dann können wir das folgende Theorem zeigen.

Theorem 10.4. *Seien u_h und u die Lösungen von* (10.3) *beziehungsweise* (10.1) *und sei Gleichung* (10.19) *erfüllt. Dann gilt für ein geeignet gewähltes v_h*

$$\|u_h(t) - u(t)\| \le Ch^r \Big(\|v\|_r + \int_0^t \|u_t\|_r \, \mathrm{d}s \Big) \quad \textit{für } t \ge 0.$$

Wir wissen bereits aus Gleichung (5.50), dass für $r > 2$ die Abschätzung (10.19) für stückweise Polynome vom Grad $r-1$ gilt, die Regularitätsannahme $w \in H^r \cap H_0^1$ für ein polygonales Gebiet Ω dann aber etwas unrealistisch ist. Für ein Gebiet Ω mit einem glatten Rand Γ sind spezielle Betrachtungen in der Grenzschicht $\Omega \setminus \Omega_h$ notwendig.

10.2 Einige vollständig diskrete Schemata

Wir wenden unsere Aufmerksamkeit nun einigen einfachen Schemata zu, die auch bezüglich der Zeit eine Diskretisierung vornehmen. Es sei S_h wie vorhin der Raum stückweise linearer Finite-Elemente-Funktionen. Wir beginnen mit dem *Rückwärts-Euler-Galerkin-Verfahren.* Sei k der Zeitschritt und $U^n \in S_h$ die Approximation von $u(t)$ an der Stelle $t = t_n = nk$. Dann wird diese Methode dadurch definiert, dass die Zeitableitung in (10.3) durch einen Rückwärts-Differenzenquotienten ersetzt wird. Mit $\bar\partial_t U^n = k^{-1}(U^n - U^{n-1})$ gilt also

$$\begin{aligned} &(\bar\partial_t U^n, \chi) + a(U^n, \chi) = (f(t_n), \chi) \quad \forall \chi \in S_h, \ n \ge 1, \\ &U^0 = v_h. \end{aligned} \tag{10.20}$$

Ist U^{n-1} gegeben, dann wird U^n damit implizit über das diskrete elliptische Problem

$$(U^n, \chi) + ka(U^n, \chi) = (U^{n-1} + kf(t_n), \chi) \quad \forall \chi \in S_h$$

definiert. Wenn wir U^n mithilfe der Basis $\{\Phi_j\}_{j=1}^{M_h}$ als $U^n(x) = \sum_{j=1}^{M_h} \alpha_j^n \Phi_j(x)$ ausdrücken, können wir diese Gleichung in der in Abschnitt 10.1 eingeführten Matrixnotation als

$$B\alpha^n + kA\alpha^n = B\alpha^{n-1} + kb^n \quad \text{für } n \geq 1$$

schreiben, wobei α^n der Vektor mit den Komponenten α_j^n oder

$$\alpha^n = (B + kA)^{-1}B\alpha^{n-1} + k(B + kA)^{-1}b^n \quad \text{für } n \geq 1 \quad \text{mit } \alpha^0 = \gamma$$

ist.

Wir beginnen unsere Analyse des Rückwärts-Euler-Verfahrens damit, die unbedingte Stabilität dieser Methode zu zeigen. Das heißt, dass diese Methode unabhängig von der Relation zwischen h und k stabil ist. Wählen wir in (10.20) $\chi = U^n$, so gilt wegen $a(U^n, U^n) \geq 0$

$$(\bar{\partial}_t U^n, U^n) \leq \|f^n\| \, \|U^n\|, \quad \text{wobei } f^n = f(t_n) \text{ ist,}$$

oder

$$\|U^n\|^2 - (U^{n-1}, U^n) \leq k\|f^n\| \, \|U^n\|.$$

Wegen $(U^{n-1}, U^n) \leq \|U^{n-1}\| \, \|U^n\|$ zeigt dies

$$\|U^n\| \leq \|U^{n-1}\| + k\|f^n\| \quad \text{für } n \geq 1,$$

woraus durch wiederholte Anwendung

$$\|U^n\| \leq \|U^0\| + k\sum_{j=1}^{n} \|f^j\| \tag{10.21}$$

folgt.

Wir werden nun die folgende Fehlerabschätzung beweisen.

Theorem 10.5. *Sind U^n und u Lösungen von* (10.20) *beziehungsweise* (10.1) *und wählen wir v_h so, dass* (10.10) *erfüllt ist, dann gilt für $n \geq 0$*

$$\|U^n - u(t_n)\| \leq Ch^2\Big(\|v\|_2 + \int_0^{t_n} \|u_t\|_2 \, \mathrm{d}s\Big) + Ck\int_0^{t_n} \|u_{tt}\| \, \mathrm{d}s.$$

Beweis. In Analogie zu (10.13) schreiben wir

$$U^n - u(t_n) = \big(U^n - R_h u(t_n)\big) + \big(R_h u(t_n) - u(t_n)\big) = \theta^n + \rho^n.$$

Wie vorhin gilt wegen (10.12)

$$\|\rho^n\| \leq Ch^2\|u(t_n)\|_2 \leq Ch^2\Big(\|v\|_2 + \int_0^{t_n} \|u_t\|_2 \, \mathrm{d}s\Big).$$

Diesmal führt eine der Gleichung (10.14) entsprechende Berechnung auf

$$(\bar{\partial}_t \theta^n, \chi) + a(\theta^n, \chi) = -(\omega^n, \chi) \tag{10.22}$$

mit

$$\omega^n = R_h\bar{\partial}_t u(t_n) - u_t(t_n) = (R_h - I)\bar{\partial}_t u(t_n) + (\bar{\partial}_t u(t_n) - u_t(t_n)) = \omega_1^n + \omega_2^n.$$

Wenden wir die Stabilitätsabschätzung (10.21) auf (10.22) an, erhalten wir

$$\|\theta^n\| \le \|\theta^0\| + k\sum_{j=1}^{n} \|\omega_1^j\| + k\sum_{j=1}^{n} \|\omega_2^j\|.$$

An dieser Stelle folgt wegen (10.10) und (10.12) wie vorhin

$$\|\theta^0\| = \|v_h - R_h v\| \le \|v_h - v\| + \|v - R_h v\| \le Ch^2\|v\|_2.$$

Beachten Sie nun, dass

$$\omega_1^j = (R_h - I)k^{-1}\int_{t_{j-1}}^{t_j} u_t \,\mathrm{d}s = k^{-1}\int_{t_{j-1}}^{t_j} (R_h - I)u_t \,\mathrm{d}s$$

gilt, woraus sich

$$k\sum_{j=1}^{n} \|\omega_1^j\| \le \sum_{j=1}^{n}\int_{t_{j-1}}^{t_j} Ch^2\|u_t\|_2 \,\mathrm{d}s = Ch^2\int_0^{t_n} \|u_t\|_2 \,\mathrm{d}s$$

ergibt. Darüber hinaus gilt aufgrund der Taylorschen Formel

$$\omega_2^j = k^{-1}(u(t_j) - u(t_{j-1})) - u_t(t_j) = -k^{-1}\int_{t_{j-1}}^{t_j} (s - t_{j-1})u_{tt}(s) \,\mathrm{d}s$$

und damit

$$k\sum_{j=1}^{n} \|\omega_2^j\| \le \sum_{j=1}^{n} \Big\| \int_{t_{j-1}}^{t_j} (s - t_{j-1})u_{tt}(s) \,\mathrm{d}s\Big\| \le k\int_0^{t_n} \|u_{tt}\| \,\mathrm{d}s.$$

Zusammen genommen schließen unsere Abschätzungen den Beweis des Theorems ab. □

Ersetzen wir in (10.20) den Rückwärts-Differenzenquotienten bezüglich der Zeit durch einen Vorwärts-Differenzenquotienten, gelangen wir zum *Vorwärts-Euler-Galerkin-Verfahren*. Mit $\partial_t U^n = (U^{n+1} - U^n)/k$ gilt also

$$\begin{aligned} &(\partial_t U^n, \chi) + a(U^n, \chi) = (f(t_n), \chi) \quad \forall \chi \in S_h,\ n \ge 1,\\ &U^0 = v_h. \end{aligned}$$

Dies kann in Matrixform als

$$B\alpha^{n+1} = (B - kA)\alpha^n + kb^n \quad \text{für } n \ge 0$$

ausgedrückt werden. Da B keine Diagonalmatrix ist, ist diese Methode nicht explizit. Wenden wir dieses Diskretisierungsverfahren jedoch auf die semidiskrete Gleichung (10.16) in der Lumped-Mass-Methode an, bei der man B

durch eine Diagonalmatrix $\bar{B}$ ersetzt, dann wird das zugehörige Vorwärts-Euler-Verfahren zu einem expliziten Verfahren.

Unter Verwendung des in (10.6) definierten diskreten Laplace-Operators kann das Vorwärts-Euler-Verfahren auch durch

$$U^{n+1} = (I + k\Delta_h)U^n + kP_h f(t_n) \quad \text{für } n \geq 0 \quad \text{mit } U^0 = v_h \tag{10.23}$$

definiert werden. Diese Methode ist anders als das Rückwärts-Euler-Verfahren nicht unbedingt stabil. Wir werden allerdings die Stabilität unter der Bedingung zeigen, dass die Familie $\{S_h\}$ dergestalt ist, dass für den größten Eigenwert $\lambda_{M_h,h}$ von $-\Delta_h$

$$\lambda_{M_h,h}\, k \leq 2 \tag{10.24}$$

gilt. Dabei werden wir der Einfachheit halber lediglich die homogene Gleichung betrachten. Erinnern wir uns an (6.37), so stellen wir fest, dass (10.24) beispielsweise dann gilt, wenn die S_h die inverse Ungleichung (6.36) erfüllen und mit der Konstanten C aus Gleichung (6.37) $k \leq 2C^{-1}h^2$ gilt, worin sich die bedingte Stabilität zeigt.

Es ist klar, dass (10.23) genau dann stabil ist, wenn $\|(I + k\Delta_h)\chi\| \leq \|\chi\|$ für alle $\chi \in S_h$ erfüllt ist. Weil $-\Delta_h$ symmetrisch positiv definit ist, gilt dies genau dann, wenn alle Eigenwerte von $I + k\Delta_h$ in $[-1, 1]$ liegen. Aufgrund der Positivität von $-\Delta_h$ entspricht dies der Forderung, dass der kleinsten Eigenwert von $I + k\Delta_h$ größer gleich -1 ist, oder dass der größte Eigenwert von $-\Delta_h$ kleiner gleich $2/k$ ist, also (10.24) erfüllt. Sehen Sie sich dazu auch Problemstellung 10.3 an.

Beachten Sie, dass das Rückwärts-Euler-Verfahren aufgrund der unsymmetrischen Wahl der Zeitdiskretisierung in der Genauigkeit lediglich von erster Ordnung ist. Deshalb kommen wir nun zum *Crank-Nicolson-Galerkin-Verfahren*, bei dem die semidiskrete Gleichung symmetrisch um den Punkt $t_{n-1/2} = (n - \frac{1}{2})k$ diskretisiert wird. Diese Vorgehensweise führt auf ein Verfahren, das in der Genauigkeit hinsichtlich der Zeit von zweiter Ordnung ist. Genauer gesagt, definieren wir $U^n \in S_h$ für $n \geq 1$ rekursiv durch

$$\begin{aligned} &(\bar{\partial}_t U^n, \chi) + a(\tfrac{1}{2}(U^n + U^{n-1}), \chi) = (f(t_{n-1/2}), \chi) \quad \forall \chi \in S_h, \\ &U^0 = v_h. \end{aligned} \tag{10.25}$$

In der Matrixnotation nimmt dies die Form

$$B\alpha^n + \tfrac{1}{2}kA\alpha^n = B\alpha^{n-1} - \tfrac{1}{2}kA\alpha^{n-1} + kb^{n-1/2} \quad \text{für } n \geq 1$$

oder mit $\alpha^0 = \gamma$ die Form

$$\alpha^n = (B + \tfrac{1}{2}kA)^{-1}(B - \tfrac{1}{2}kA)\alpha^{n-1} + k(B + \tfrac{1}{2}kA)^{-1}b^{n-1/2}, \quad n \geq 1$$

an.

Dieses Verfahren ist ebenfalls unbedingt stabil, was man zeigen kann, indem man in (10.25) $\chi = U^n + U^{n-1}$ wählt und auf der rechten Seite die Cauchy-Schwarz-Ungleichung anwendet. Dann ergibt sich

$$k(\bar{\partial}_t U^n, U^n + U^{n+1}) = \|U^n\|^2 - \|U^{n-1}\|^2 = (\|U^n\| - \|U^{n-1}\|)(\|U^n\| + \|U^{n-1}\|).$$

Unter Verwendung der Positivität von $a(U^n, U^n)$ und nach Streichen eines Faktors $\|U^n\| + \|U^{n-1}\|$ erhalten wir

$$\|U^n\| \le \|U^{n-1}\| + k\|f^{n-1/2}\| \quad \text{mit } f^{n-1/2} = f(t_{n-1/2}),$$

oder nach Summation

$$\|U^n\| \le \|v_h\| + k \sum_{j=1}^{n} \|f^{j-1/2}\|.$$

Diesmal ergibt sich die folgende Fehlerabschätzung. Der Beweis, den Sie in Problemstellung 10.7 ausführen sollen, verläuft analog zu dem von Theorem 10.5.

Theorem 10.6. *Seien U^n und u die Lösungen von* (10.25) *beziehungsweise* (10.1). *Wird v_h so gewählt, dass* (10.10) *erfüllt ist, dann gilt für $n \ge 0$*

$$\|U^n - u(t_n)\| \le Ch^2 \Big(\|v\|_2 + \int_0^{t_n} \|u_t\|_2 \, \mathrm{d}s \Big) + Ck^2 \int_0^{t_n} \big(\|u_{ttt}\| + \|\Delta u_{tt}\| \big) \, \mathrm{d}s.$$

10.3 Problemstellungen

Problem 10.1. Betrachten Sie das Problem (10.1) in einer räumlichen Dimension mit $\Omega = (0, 1)$. Zur numerischen Lösung verwenden wir die stückweise linearen Funktionen über der Zerlegung

$$0 < x_1 < x_2 < \ldots < x_M < 1, \quad x_j = jh, \ h = 1/(M+1).$$

Bestimmen Sie die Massenmatrix B und die Steifigkeitsmatrix A und schreiben Sie das semidiskrete Problem, die Rückwärts-Euler-Gleichungen und die Crank-Nicolson-Gleichungen auf.

Problem 10.2. (Übung am Rechner.) Betrachten Sie das Anfangs-Randwertproblem (10.1) mit $\Omega = (-\pi, \pi)$ und $v = \operatorname{sign} x$.

(a) Bestimmen Sie die exakte Lösung durch Entwicklung nach Eigenfunktionen.
(b) Wenden Sie das Rückwärts-Euler-Verfahren (10.20) auf der Basis stückweise linearer finiter Elemente mit $v_h = P_h v$ und $(h, k) = (\pi/5, 1/10)$, $(\pi/10, 1/40)$ an. Bestimmen Sie den maximalen Fehler an den Gitterpunkten für $t = 0.1,\ 0.5,\ 1.0$.

Problem 10.3. (a) Zeigen Sie, dass der in (10.6) definierte Operator $-\Delta_h : S_h \to S_h$ selbstadjungiert positiv definit bezüglich $(\cdot,\cdot)$ ist.
(b) Zeigen Sie, dass mit der Notation aus Theorem 6.7

$$-\Delta_h v_h = \sum_{i=1}^{M_h} \lambda_{i,h}(v_h, \varphi_{i,h})\varphi_{i,h} \quad \text{und} \quad \|\Delta_h\| = \lambda_{M_h,h}$$

gilt. Hinweis: Die linke Seite der zweiten Gleichung ist die Operatornorm von Δ_h (siehe (A.7)). Folglich müssen Sie für alle $\chi \in S_h$ die Beziehung $\|\Delta_h\chi\| \le \lambda_{M_h,h}\|\chi\|$ zeigen, wobei Gleichheit für ein χ angenommen wird.
(c) Angenommen, die Familie der Räume finiter Elemente $\{S_h\}$ erfüllt die inverse Ungleichung (6.36). Zeigen Sie, dass

$$\|\Delta_h\| \le Ch^{-2}$$

gilt. Hinweis: Sehen Sie sich (6.37) an.

Problem 10.4. Angenommen, es gilt $f = 0$ und u_h und u sind Lösungen von (10.3) beziehungsweise (10.1) mit $v_h = P_h v$.
(a) Es sei $v \in H^2 \cap H_0^1$. Zeigen Sie die Gültigkeit von

$$\|u_h(t) - u(t)\| \le Ch^2\|v\|_2 \quad \text{für } t \ge 0.$$

(b) Es sei $v \in L_2$. Zeigen Sie die Gültigkeit

$$\|u_h(t) - u(t)\| \le Ch^2 t^{-1}\|v\| \quad \text{für } t > 0.$$

Hinweis: Leiten Sie zur Lösung von Teil (a) aus (10.15) die Beziehung

$$\theta(t) = E_h(t)\theta(0) - \int_0^t E_h(t-s)P_h\rho_t(s)\,\mathrm{d}s \tag{10.26}$$

ab. Zerlegen Sie die Integrale gemäß $\int_0^t = \int_0^{t/2} + \int_{t/2}^t$ und integrieren Sie im ersten Term partiell, wodurch Sie

$$\begin{aligned}\theta(t) = {} & E_h(t)P_h e(0) - E_h(t/2)P_h\rho(t/2) \\ & + \int_0^{t/2} D_s E_h(t-s)P_h\rho(s)\,\mathrm{d}s - \int_{t/2}^t E_h(t-s)P_h\rho_t(s)\,\mathrm{d}s\end{aligned}$$

erhalten. Verwenden Sie anschließend (10.18), (10.12), (8.18) und Problemstellung 8.10.

Zur Lösung von Teil (b) integrieren Sie nochmals partiell, wodurch Sie die zusätzlichen Terme

$$D_t E_h(t/2)P_h\tilde{\rho}(t/2) - \int_0^{t/2} D_s^2 E_h(t-s)P_h\tilde{\rho}(s)\,\mathrm{d}s$$

mit $\tilde{\rho}(t) = \int_0^t \rho(s)\,\mathrm{d}s$, $\|\tilde{\rho}\| \le Ch^2\|\tilde{u}\|_2$, $\|\tilde{u}\|_2 \le C\|\Delta\tilde{u}\|$ und $\Delta\tilde{u}(t) = \int_0^t u_t(s)\,\mathrm{d}s = u(t) - v$ erhalten.

Problem 10.5. Angenommen, für die Familie der Räume finiter Elemente $\{S_h\}$ gilt $\|\Delta_h\| \le Ch^{-2}$ (siehe Problemstellung 10.3). Seien u_h und u Lösungen von (10.3) beziehungsweise (10.1). Nehmen Sie darüber hinaus $\|v_h - v\| \le Ch^2\|v\|_2$ an. Zeigen Sie die Gültigkeit von

$$\|u_h(t) - u(t)\| \le C\big(1 + \log(t/h^2)\big)h^2 \max_{0\le s\le t} \|u(s)\|_2 \quad \text{für } t \ge h^2.$$

Hinweis: Integrieren Sie in (10.26) partiell, was auf

$$\theta(t) = E_h(t)P_h e(0) - P_h\rho(t) + \int_0^t D_s E_h(t-s)P_h\rho(s)\,\mathrm{d}s$$

führt. Zerlegen Sie das Integral gemäß $\int_0^t = \int_0^{t-h^2} + \int_{t-h^2}^t$ und behandeln Sie den ersten Teil wie in Problemstellung 10.4 (a). Benutzen Sie im zweiten Teil $\|D_s E_h(t-s)\| = \|\Delta_h E_h(t-s)\| \le \|\Delta_h\|\|E_h(t-s)\| \le Ch^{-2}$.

Problem 10.6. Beweisen Sie Fehlerschranken, die analog zu denen aus Theorem 10.1 sind, wenn der elliptische Term $-\Delta u$ in (10.1) wie in Abschnitt 3.5 durch $\mathcal{A}u = -\nabla\cdot(a\nabla u) + b\cdot\nabla u + cu$ ersetzt wird. Hinweis: Sehen Sie sich die Problemstellungen 5.7 und 8.8 an.

Problem 10.7. Beweisen Sie Theorem 10.6.

11

Hyperbolische Gleichungen

In diesem Kapitel stellen wir die grundlegenden Konzepte und Ergebnisse für hyperbolische Gleichungen vor. Wir beginnen im Abschnitt 11.1 mit einer kurzen Diskussion von charakteristischen Richtungen, Kurven und Flächen. In Abschnitt 11.2 untersuchen wir die Wellengleichung. Wir verwenden die Methode der Entwicklung nach Eigenfunktionen, um das gewöhnliche Anfangs-Randwertproblem zu lösen, und wenden die Energiemethode an, um Eindeutigkeit und Abhängigkeitsbereiche zu untersuchen. In Abschnitt 11.3 reduzieren wir die Lösung von skalaren partiellen Differentialgleichungen erster Ordnung auf die Integration entlang charakteristischer Kurven. Danach erweitern wir diese Methode in Abschnitt 11.4 auf symmetrische Systeme erster Ordnung und betrachten schließlich hyperbolische Systeme mit mehr als einer Variablen unter Verwendung von Energieargumenten.

11.1 Charakteristische Richtungen und Flächen

Wir betrachten die skalare lineare partielle Differentialgleichung

$$\mathcal{L}u = \mathcal{L}(x, D)u := \sum_{|a| \le m} a_\alpha(x) D^\alpha u = f(x) \quad \text{in } \Omega, \tag{11.1}$$

wobei Ω ein Gebiet in $\mathbf{R}^d$ ist. Wir sagen, dass die Richtung $\xi \in \mathbf{R}^d$, $\xi \neq 0$ eine *charakteristische Richtung* des Operators $\mathcal{L}(x, D)$ an der Stelle x ist, wenn

$$\Lambda(\xi) = \lambda(x, \xi) := \sum_{|\alpha| = m} a_\alpha(x) \xi^\alpha = 0 \tag{11.2}$$

gilt. Das Polynom $\Lambda(\xi) = \Lambda(x, \xi)$ wird als *charakteristisches Polynom* von $\mathcal{L}$ an der Stelle x bezeichnet. Beachten Sie, dass die Summe in (11.2) nur über $|a| = m$ läuft, d. h. sie entspricht dem *Hauptwert* des Operators $\mathcal{L}$, gebildet aus den Termen der Ordnung m.

Mitunter werden wir auch Systeme linearer partieller Differentialgleichungen betrachten. Diese sind in (11.1) enthalten, wenn wir die Koeffizienten $a_\alpha(x)$ als Matrizen interpretieren. Wenn diese Matrizen quadratische Matrizen der Ordnung N mit $N \geq 2$ sind, dann sagen wir, dass $\xi \in \mathbf{R}^d$ eine charakteristische Richtung an der Stelle x ist, falls

$$\det \Lambda(x, \xi) = 0 \tag{11.3}$$

gilt.

Eine $(d-1)$-dimensionale Fläche in $\mathbf{R}^d$ wird als *charakteristische Fläche* bezeichnet, wenn ihre Normale an jedem Punkt x eine charakteristische Richtung an der Stelle x ist. Im Falle der Ebene, d. h. für $d = 2$, bezeichnen wir diese als *charakteristische Kurve* oder einfach als *Charakteristik*.

Beispiel 11.1. Für die skalare Gleichung erster Ordnung

$$\sum_{j=1}^{d} a_j(x) \frac{\partial u}{\partial x_j} + a_0(x) u = f(x) \tag{11.4}$$

sind die charakteristischen Richtungen durch die Gleichung

$$\Lambda(x, \xi) = \sum_{j=1}^{d} a_j(x) \xi_j = 0$$

gegeben. Somit ist jede Richtung, die orthogonal zum Vektor $a(x) = (a_1(x), \ldots, a_d(x))$ verläuft, eine Charakteristik.

Die Hyperebene $x_1 = 0$ besitzt die Normale $(1, 0, \ldots, 0)$ und ist daher eine charakteristische Fläche, wenn $a_1(x) = 0$ für alle $x = (0, x_2, \ldots, x_d)$ gilt. Sie ist nichtcharakteristisch, falls für alle Punkte $x = (0, x_2, \ldots, x_d)$ die Beziehung $a_1(x) \neq 0$ gilt. Dies ist äquivalent zu der Aussage, dass die Gleichung (11.4) nach $\partial u / \partial x_1$ aufgelöst werden kann. In diesem Fall kann die Gleichung in der Nähe der Hyperebene in der Form

$$\frac{\partial u}{\partial x_1} = \sum_{j=2}^{d} \tilde{a}_j(x) \frac{\partial u}{\partial x_j} + \tilde{a}_0(x) u + \tilde{f}(x)$$

geschrieben werden.

Beispiel 11.2. Die Poisson-Gleichung

$$-\Delta u = f$$

besitzt keine charakteristischen Richtungen, da $\Lambda = -(\xi_1^2 + \cdots + \xi_d^2) = -|\xi|^2$ nur für $\xi = 0$ verschwindet.

Beispiel 11.3. Die Wärmeleitungsgleichung

$$\frac{\partial u}{\partial t} - \Delta u = f,$$

die wir nun in $\mathbf{R}^{d+1}$ an den Punkten (x,t) mit $x \in \mathbf{R}^d$, $t \in \mathbf{R}$ betrachten, hat die charakteristische Gleichung $\Lambda(\xi,\tau) = -|\xi|^2 = 0$. In diesem Fall ist die Variable $(\xi,\tau) \in \mathbf{R}^{d+1}$ mit $\xi \in \mathbf{R}^d, \tau \in \mathbf{R}$. Dies bedeutet, dass $(0,\ldots,0,1)$ eine charakteristische Richtung ist und die Hyperebene $t = 0$ eine charakteristische Fläche.

Beispiel 11.4. Analog gehört zur Wellengleichung

$$\frac{\partial^2 u}{\partial t^2} - \Delta u = f$$

die charakteristische Gleichung $\Lambda(\xi,\tau) = |\xi|^2 = 0$, sodass $(\xi, \pm|\xi|)$ für eine beliebige Wahl von $\xi \neq 0$ eine charakteristische Richtung ist. Beispielsweise sind für einen Kreiskegel mit dem Scheitel $(\bar{x},\bar{t})$, der durch die Gleichung

$$F(x,t) := |x-\bar{x}|^2 - (t-\bar{t})^2 = 0$$

definiert ist, die Normalen in einem Punkt (x,t) auf dem Kegel durch

$$\left(\frac{\partial F}{\partial x_1},\ldots,\frac{\partial F}{\partial x_d},\frac{\partial F}{\partial t}\right) = 2(x-\bar{x},-(t-\bar{t})) = 2(x-\bar{x},\mp|x-\bar{x}|)$$

gegeben. Somit ist dies eine charakteristische Richtung und der Kegel selbst bildet eine charakteristische Fläche. In diesem Fall ist die Hyperebene $t = 0$ nichtcharakteristisch.

Das charakteristische Polynom kann verwendet werden, um partielle Differentialgleichungen zu klassifizieren. Beispielsweise wird der Operator $\mathcal{L}$ als *elliptisch* bezeichnet, wenn er keine charakteristischen Richtungen besitzt. Für eine Differentialgleichung zweiter Ordnung mit konstanten Koeffizienten ist $\Lambda(\xi)$ ein homogenes quadratisches Polynom, dass wir als symmetrisch voraussetzen können, sodass

$$\Lambda(\xi) = \sum_{j,k=1}^{d} a_{jk}\xi_j\xi_k$$

mit $a_{jk} = a_{kj}$ gilt. Nach einer orthogonalen Transformation der Variablen $\xi = P\eta$ kann das Polynom in der Form

$$\Lambda(P\eta) = \sum_{j=1}^{d} \lambda_j \eta_j^2$$

geschrieben werden, wobei $\{\lambda_j\}_{j=1}^d$ die Eigenwerte der Matrix $A = (a_{jk})$ sind. Die Differentialgleichung wird als *elliptisch* bezeichnet, wenn alle λ_j wie im Beispiel 11.2 das gleiche Vorzeichen besitzen, was äquivalent zu der obigen Definition ist, dass es keine Charakteristiken gibt. Die Differentialgleichung wird als *hyperbolisch* bezeichnet, wenn bis auf einen Eigenwert alle λ_j das gleiche Vorzeichen haben und der verbleibende Eigenwert λ_j das entgegengesetzte Vorzeichen besitzt, was auf Beispiel 11.4 zutrifft. Im Beispiel 11.3 haben bis auf ein λ_j alle Eigenwerte das gleiche Vorzeichen und der verbleibende Eigenwert λ_j ist Null, weshalb es sich in diesem Fall um eine *parabolische* Differentialgleichung handelt.

Beispiel 11.5. Sei A eine $N \times N$-Diagonalmatrix mit den Diagonalelementen $\{\lambda_j\}_{j=1}^N$. Betrachten wir das System

$$\frac{\partial u}{\partial t} - A\frac{\partial u}{\partial x} = f.$$

Die charakteristischen Richtungen (ξ, τ) sind durch die Gleichung

$$\det(\tau I - \xi A) = 0$$

gegeben. Die Matrix $\tau I - \xi A$ ist eine Diagonalmatrix mit den Elementen $\tau - \lambda_j \xi$, $j = 1, \ldots, N$, sodass (ξ, τ) genau dann eine charakteristische Richtung ist, wenn eines dieser Elemente verschwindet. Daraus ergeben sich die charakteristischen Richtungen $(1, \lambda_j)$, $j = 1, \ldots, N$. Somit sind die Geraden

$$x + \lambda_j t = \text{konstant} \quad \text{mit } j = 1, \ldots, N$$

die charakteristischen Kurven, und die Hyperebene $t = 0$ ist nichtcharakteristisch.

11.2 Die Wellengleichung

In diesem Abschnitt betrachten wir zunächst das Anfangswertproblem für die Wellengleichung

$$\begin{aligned} &u_{tt} - \Delta u = 0 && \text{in } \Omega \times \mathbf{R}_+, \\ &u = 0 && \text{auf } \Gamma \times \mathbf{R}_+, \\ &u(\cdot, 0) = v, u_t(\cdot, 0) = w && \text{in } \Omega, \end{aligned} \tag{11.5}$$

wobei Ω ein beschränktes Gebiet in $\mathbf{R}^d$ mit dem Rand Γ ist, und v und w gegebene Funktionen von x in Ω sind.

Die Existenz einer Lösung von (11.5) kann ähnlich wie im Fall der Wärmeleitungsgleichung im Abschnitt 8.2 durch Entwicklung nach Eigenfunktionen gezeigt werden. Um dies zu demonstrieren, führen wir die Eigenfunktionen

$\{\varphi_j\}_{j=1}^{\infty}$ und die zugehörigen Eigenwerte $\{\lambda_j\}_{j=1}^{\infty}$ des elliptischen Operators $-\Delta$ ein und nehmen an, dass (11.5) eine Lösung der Form

$$u(x,t) = \sum_{j=1}^{\infty} \hat{u}_j(t)\varphi_j(x)$$

besitzt. Setzen wir dies in die Differentialgleichung ein, erhalten wir

$$\sum_{j=1}^{\infty} \left(\hat{u}_j''(t) + \lambda_j \hat{u}_j(t)\right) \varphi_j(x) = 0.$$

Entsprechend gilt für die Anfangsbedingungen

$$\sum_{j=1}^{\infty} \hat{u}_j(0)\varphi_j(x) = v(x), \quad \sum_{j=1}^{\infty} \hat{u}_j'(0)\varphi_j(x) = w(x).$$

Da die φ_j eine Orthonormalbasis von $L_2 = L_2(\Omega)$ bilden, gilt für $j \geq 1$

$$\begin{aligned} &\hat{u}_j'' + \lambda_j \hat{u}_j = 0 \qquad \text{für } t > 0, \\ &\hat{u}_j(0) = \hat{v}_j = (v, \varphi_j), \quad \hat{u}_j'(0) = \hat{w}_j = (w, \varphi_j). \end{aligned}$$

Durch Lösen des Anfangswertproblems folgern wir

$$\hat{u}_j = \hat{v}_j \cos(\sqrt{\lambda_j}t) + \hat{w}_j \frac{1}{\sqrt{\lambda_j}} \sin(\sqrt{\lambda_j}t) \qquad \text{für } j \geq 1,$$

und damit ist

$$u(x,t) = \sum_{j=1}^{\infty} \left(\hat{v}_j \cos(\sqrt{\lambda_j}t) + \hat{w}_j \lambda_j^{-1/2} \sin(\sqrt{\lambda_j}t)\right) \varphi_j(x). \tag{11.6}$$

Sind v und w hinreichend regulär, sodass die Reihen auch nach dem Differenzieren konvergieren, stellt die Gleichung offenbar eine Lösung von (11.5) dar (siehe Problemstellung 11.4). Somit gelangen wir zu der folgenden Einsicht.

Theorem 11.1. *Es gelte $v \in \mathbf{H}^2 \cap \mathbf{H}_0^1, w \in \mathbf{H}_0^1$. Dann ist die Reihe* (11.6) *eine Lösung von* (11.5).

Wir werden nun eine Energieabschätzung für die Lösung u von (11.5) beweisen. Mithilfe dieser Abschätzung erhalten wir leicht auf dem üblichen Wege die Eindeutigkeit und die Stabilität der Lösung des Problems. Die hier beschriebene Energiemethode ist auch in Situationen hilfreich, in denen die Methode der Entwicklung nach Eigenfunktionen nicht anwendbar ist.

Theorem 11.2. *Sei $u = u(x,t)$ eine hinreichend glatte Lösung von* (11.5). *Dann ist die Gesamtenergie $\mathcal{E}(t)$ von u zeitlich konstant, d. h. es gilt*

$$\mathcal{E}(t) := \frac{1}{2} \int_{\Omega} \left(u_t^2 + |\nabla u|^2\right) dx = \mathcal{E}(0). \tag{11.7}$$

Beweis. Durch Multiplizieren der Differentialgleichung in (11.5) mit u_t und Integration bezüglich x über Ω erhalten wir unter Verwendung der Greenschen Formel die Gleichung

$$\int_\Omega u_{tt} u_t dx + \int_\Omega \nabla u \cdot \nabla u_t dx = 0,$$

oder mit der Notation aus Abschnitt 8.3

$$(u_{tt}, u_t) + a(u, u_t) = 0.$$

Somit gilt

$$\frac{1}{2}\frac{d}{dt}\|u_t\|^2 + \frac{1}{2}\frac{d}{dt}\|\nabla u\|^2 = 0,$$

oder

$$\frac{d}{dt}\mathcal{E}(t) = 0 \qquad \text{für } t > 0.$$

Daraus folgt unmittelbar die Aussage des Theorems. □

Wir werden nun eine Abschätzung für das reine Anfangswertproblem für die Wellengleichung beweisen, aus der wir ableiten, dass die Lösung in einem gegebenen Punkt (x, t) mit $t > 0$ nur von den Anfangswerten in einem speziellen Gebiet der Anfangsmannigfaltigkeit $t = 0$ abhängt. Das betrachtete Problem lautet also

$$\begin{aligned} &u_{tt} - \Delta u = 0 && \text{in } \mathbf{R}^d \times \mathbf{R}_+, \\ &u(\cdot, 0) = v, \quad u_t(\cdot, 0) = w && \text{in } \mathbf{R}^d. \end{aligned} \tag{11.8}$$

Theorem 11.3. *Sei u eine Lösung der Wellengleichung in* (11.8). *Für einen gegebenen Punkt $(\bar{x}, \bar{t})$ in $\mathbf{R}^d \times \mathbf{R}_+$ sei K der Kreiskegel*

$$K = \big\{(x,t) \in \mathbf{R}^d \times \mathbf{R}_+ \colon |x - \bar{x}| \le \bar{t} - t,\ t \le \bar{t}\big\} \tag{11.9}$$

(siehe Abbildung 11.1). Weiter sei

$$\mathcal{E}_K(t) = \frac{1}{2}\int_{B_t} \big(u_t(x,t)^2 + |\nabla u(x,t)|^2\big)\, dx$$

mit $B_t = \big\{x \in \mathbf{R}^d \colon (x,t) \in K\big\}$. Dann gilt

$$\mathcal{E}_K(t) \le \mathcal{E}_K(0) \qquad \textit{für } 0 \le t \le \bar{t}.$$

Beweis. Wir führen die Mantelfläche $M = \{(x,t) \colon |x - \bar{x}| = \bar{t} - t\}$ von K ein und setzen $M_t = \{(x,\tau) \in M \colon \tau \le t\}$. Durch Multiplikation der Differentialgleichung mit $2u_t$ finden wir

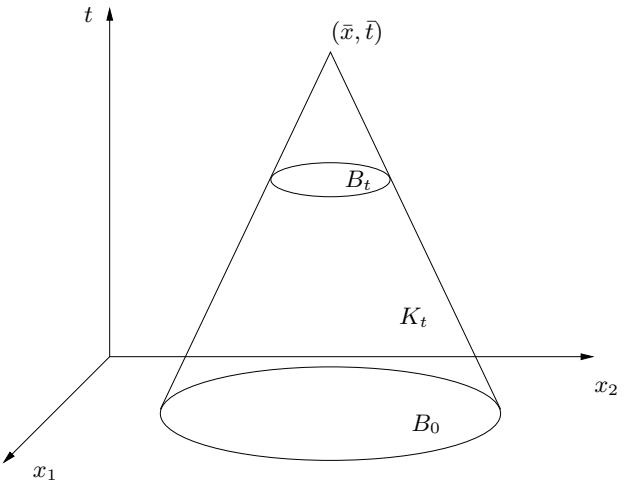

Abbildung 11.1. Der Lichtkegel.

$$\begin{aligned} 0 &= 2\,(u_{tt} - \nabla \cdot \nabla u)\,u_t = 2u_{tt}u_t + 2\nabla u \cdot \nabla u_t - 2\nabla \cdot (\nabla u u_t) \\ &= D_t\left(u_t^2 + |\nabla u|^2\right) - 2\nabla \cdot (\nabla u u_t)\,. \end{aligned}$$

Durch Integration über $K_t = \{(x,\tau) \in K \colon 0 \le \tau \le t\}$ (siehe Abbildung 11.1) und Verwendung des Divergenztheorems erhalten wir mit $n = (n_x, n_t) = (n_x, \ldots, n_{x_d}, n_t)$, der äußeren Normalen von ∂K_t,

$$\begin{aligned} 0 &= \int_{\partial K_t} \left(n_t\left(u_t^2 + |\nabla u|^2\right) - 2n_x \cdot \nabla u u_t\right) ds \\ &= \int_{B_t} \left(u_t^2 + |\nabla u|^2\right) dx - \int_{B_0} \left(u_t^2 + |\nabla u|^2\right) dx \\ &\quad + \int_{M_t} \left(n_t\left(u_t^2 + |\nabla u|^2\right) - 2u_t n_x \cdot \nabla u\right) ds. \end{aligned}$$

Zum Abschluss des Beweises zeigen wir nun, dass der Integrand des letzten Terms nichtnegativ ist. Auf M gilt $n_t^2 = |n_x|^2$. Wegen $n_t = 1/\sqrt{2}$ führt dies unter Verwendung der Cauchy-Schwarz-Ungleichung auf

$$|n_x \cdot \nabla u| \le |n_x||\nabla u| = n_t|\nabla u|.$$

Verwenden wir außerdem die Ungleichung $2\,ab \le a^2 + b^2$, so erhalten wir

$$2|u_t n_x \cdot \nabla u| = 2|u_t||n_x \cdot \nabla u| \le 2n_t|u_t||\nabla u| \le n_t(u_t^2 + |\nabla u|^2),$$

was den Beweis vervollständigt. □

Aus Theorem 11.2 folgt, dass $u = 0$ in K gilt, wenn $v = w = 0$ in B_0 ist. Dies gilt insbesondere auch für $(\bar{x}, \bar{t})$. Dies zeigt, dass der Wert der Lösung

von (11.8) an der Stelle $(\bar{x}, \bar{t})$ nur von den Werten von v und w innerhalb der Kugel B_0 abhängt, die durch den Kreiskegel K mit dem Scheitel $(\bar{x}, \bar{t})$ definiert wird, nicht aber von den Werten v und w außerhalb dieser Kugel.

Die Existenz der Lösung des reinen Anfangswertproblems (11.8) kann auf verschiedenen Wegen gezeigt werden. Für die spezielle, hier betrachtete Gleichung kann eine explizite Lösung in Form einer Integraldarstellung aufgeschrieben werden, die in Abhängigkeit von der Anzahl d der räumlichen Dimensionen unterschiedliche Formen annimmt. Für $d = 1$ kann man beispielsweise leicht zeigen, dass

$$u(x,t) = \frac{1}{2}\left(v(x+t) + v(x-t)\right) + \frac{1}{2}\int_{x-t}^{x+t} w(y)dy \tag{11.10}$$

gilt (siehe Problemstellung 11.5), was als Formel von d'Alembert bezeichnet wird. Für $d = 3$ kann man zeigen, dass

$$u(x,t) = \frac{\partial}{\partial t}\left\{\frac{1}{4\pi t}\int_{|y-x|=t} v(y)ds_y\right\} + \frac{1}{4\pi t}\int_{|y-x|=t} w(y)ds_y$$

gilt. In diesem Fall hängt die Lösung an der Stelle (x,t) nur von den Daten innerhalb des Kreises ab, der zur Zeit t durch den charakteristischen Kegel ausgeschnitten wird. Allgemeiner gilt dies auch, wenn d eine ungerade ganze Zahl ist. Ist d geradzahlig, dann hängt die Lösung an der Stelle (x,t) von den Anfangswerten innerhalb der „Kugel" $|y - x| \le t$ ab.

11.3 Skalare Gleichungen erster Ordnung

Wir kommen nun zu der skalaren Differentialgleichung erster Ordnung

$$\sum_{j=1}^{d} a_j(x)\frac{\partial u}{\partial x_j} + a_0(x)u = f(x), \quad x \in \Omega. \tag{11.11}$$

Hierbei ist $\Omega \subset \mathbf{R}^d$ ein beschränktes oder unbeschränktes Gebiet, $a = a(x) = (a_1(x), \ldots, a_d(x))$ ein glattes Vektorfeld, das an keinem Punkt verschwindet, und a_0 und f sind gegebene glatte Funktionen.

Wir sagen, dass $x = x(s) = (x_1(s), \ldots, x_d(s))$ mit dem reellen Parameter s eine *charakteristische Kurve* oder einfach eine *Charakteristik* für (11.11) ist, wenn

$$\frac{d}{ds}x(s) = a(x(s)) \tag{11.12}$$

gilt, d. h. wenn die durch $x = x(s)$ in $\mathbf{R}^d$ definierte Kurve den Vektor $a(x)$ in jedem ihrer Punkte als Tangente besitzt. Beachten Sie, dass eine charakteristische Richtung eine Normale an die charakteristische Kurve ist. Insbesondere

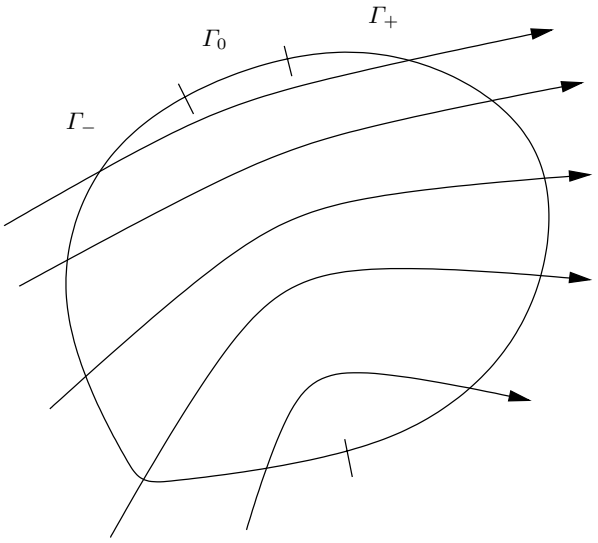

Abbildung 11.2. Zufluss- und Abflussrand.

ist eine Charakteristik im Spezialfall $d = 2$ eine charakteristische Kurve in dem in Abschnitt 11.1 beschriebenem Sinne.

In Koordinatenform kann (11.12) als System gewöhnlicher Differentialgleichungen

$$\frac{dx_j}{ds} = a_j(x) \quad \text{für } j = 1, \dots, d$$

geschrieben werden. Da die Vektorfelder nicht verschwinden, ergibt sich aus der Theorie dieser Gleichungen, dass für jedes $x_0 \in \Omega$ eine solche eindeutige Kurve in einer Umgebung von x_0 mit $x(0) = x_0$ existiert.

Es sei Γ der Rand von Ω. Wir bezeichnen mit Γ_- den durch

$$\Gamma_- = \big\{x \in \Gamma : n(x) \cdot a(x) < 0\big\}$$

definierten Zuflussrand, wobei $n(x)$ die äußere Normale an Γ an der Stelle x ist. Durch jeden Punkt von Γ_- gibt es eine eindeutige Charakteristik, die in Ω einritt. Wir schreiben für die Lösung von (11.11) die Randbedingung

$$u = v \quad \text{auf } \Gamma_- \tag{11.13}$$

vor, wobei v eine gegebene glatte Funktion auf Γ_- ist. Wir führen außerdem den Abflussrand und den charakteristischen Rand

$$\Gamma_+ = \big\{x \in \Gamma : n(x) \cdot a(x) > 0\big\}, \quad \Gamma_0 = \big\{x \in \Gamma : n(x) \cdot a(x) = 0\big\}$$

ein.

Wir betrachten eine Lösung u von (11.11), (11.13) entlang einer Charakteristik $x = x(s)$, d. h. die Funktion $w(s) = u(x(s))$. Wegen der Kettenregel gilt

$$\frac{dw}{ds} = \nabla u \cdot \frac{dx}{ds} = a(x) \cdot \nabla u,$$

sodass w wegen (11.11) die Gleichungen

$$\begin{aligned} &\frac{dw}{ds} + a_0(x(s))w = f(x(s)) \quad \text{für } s > 0, \\ &w(0) = v(x_0) \quad \text{mit } x(0) = x_0 \in \Gamma_- \end{aligned} \tag{11.14}$$

erfüllt. Dies ist ein Anfangswertproblem für eine lineare gewöhnliche Differentialgleichung, die für die Werte von w an den Punkten entlang der Charakteristik gelöst werden kann. Um eine Lösung von (11.11), (11.13) in einem Punkt $\bar{x} \in \Omega$ zu finden, bestimmen wir die Charaktersitik durch $\bar{x}$, suchen nach deren Schnittpunkt x_0 mit Γ_- und lösen dann die Gleichung (11.14) mit $x(0) = x_0$. Die Lösung an der Stelle $\bar{x}$ hängt also nur von $v(x_0)$ und den Werten von f auf der Charakteristik ab.

Im Spezialfall $a_0 = f = 0$ in Ω reduziert sich die Gleichung (11.14) auf

$$\frac{dw}{ds} = 0 \quad \text{für } s > 0 \quad \text{mit } w(0) = v(x_0),\ x(0) = x_0 \in \Gamma_-.$$

Folglich ist $u(x(s))$ in diesem Fall entlang der Charakteristik konstant und der Wert der Lösung an der Stelle $\bar{x}$ ist gleich dem an der Stelle $x(0)$, d. h. es gilt $u(x(s)) = u(x(0)) = v(x(0))$.

Diese Vorgehensweise wird häufig als *Charakteristikenmethode* bezeichnet.

Gleichungen der Form (11.11) kommen häufig als Grenzfall der stationären Wärmeleitungs- oder Diffusionsgleichung mit Konvektion vor, wenn der Wärmeleitungs- oder Diffusionskoeffizient verschwindet (siehe (1.18)). Solche Gleichungen können auch im zeitabhängigen Fall in der Form (11.11) geschrieben werden, wenn eine der unabhängigen Variablen als Zeit interpretiert wird. Kennzeichnen wir die Zeitvariable in (11.11) explizit, gilt

$$\begin{aligned} &u_t + a \cdot \nabla u + a_0 u = f && \text{in } \Omega \times \mathbf{R}_+, \\ &u = g && \text{in } \Gamma_{-,x}, \\ &u(\cdot, 0) = v && \text{in } \Omega. \end{aligned}$$

Nun ist $\Omega \subset \mathbf{R}^d$ ein räumliches Gebiet mit dem Rand Γ und der Zuflussrand von $\Omega \times \mathbf{R}_+$ wird in seinen räumlichen Teil $\Gamma_{-,x} = \big\{(x,t) \in \Gamma \times \mathbf{R}_+ : a(x,t) \cdot n < 0\big\}$ und seinen zeitlichen Teil $\Gamma_{-,t} = \Omega \times \{0\}$, der zu $t = 0$ gehört, unterteilt. Wir können dann die Zeitvariable benutzen, um die charakteristischen Kurven durch $x = x(t)$ zu parametrisieren. Diese werden in diesem Zusammenhang häufig als Stromlinien bezeichnet.

Beispiel 11.6. Wir betrachten das Problem

$$\begin{aligned} u_t + \lambda u_x &= 0 \qquad \text{in } \mathbf{R} \times \mathbf{R}_+, \\ u(\cdot, 0) &= v \qquad \text{in } \mathbf{R}. \end{aligned}$$

Hier sind die Charakteristiken $(x(s), t(s))$ durch

$$\frac{dx}{ds} = \lambda, \quad \frac{dt}{ds} = 1$$

bestimmt. Wir können also t als Parameter für die Charakteristik benutzen und erhalten

$$x = \lambda t + C.$$

Die Charakteristik durch $(\bar{x}, \bar{t})$ ist

$$x - \bar{x} = \lambda(t - \bar{t}), \tag{11.15}$$

und weil die Lösung auf dieser Geraden konstant ist, gilt

$$u(\bar{x}, \bar{t}) = v(\bar{x} - \lambda \bar{t}).$$

Beispiel 11.7. Mit $\Omega = (0, 1)$ suchen wir nun nach einer Lösung von

$$\begin{aligned} u_t + \lambda u_x + u &= 1 \qquad \text{in } \Omega \times \mathbf{R}_+, \\ u &= 0 \qquad \text{auf } \Gamma_- \end{aligned}$$

mit $\lambda = \text{konstant} > 0$. Hier gilt

$$\Gamma_- = \big(\{0\} \times \mathbf{R}_+\big) \cup \big(\bar{\Omega} \times \{0\}\big) = \Gamma_{-,x} \cup \Gamma_{-,t},$$

und die Charakteristik durch $(\bar{x}, \bar{t})$ ist wiederum durch (11.15) definiert.

Wir betrachten zunächst den Fall $\bar{x} \geq \lambda \bar{t}$ (siehe Abbildung 11.3). Dann beginnt die Charakteristik durch $(\bar{x}, \bar{t})$ im Punkt $(\bar{x} - \lambda \bar{t}, 0) \in \Gamma_{-,x}$. Wir führen $w(s) = u(\bar{x} + \lambda(s - \bar{t})s)$ mit dem Parameter $s = t$ ein und stellen fest, dass die Gleichung für w durch

$$w' + w = 1 \quad \text{für } s > 0 \quad \text{mit } w(0) = 0$$

gegeben ist. Folglich gilt

$$w(s) = 1 - \mathrm{e}^{-s} \tag{11.16}$$

und

$$u(\bar{x}, \bar{t}) = 1 - \mathrm{e}^{-\bar{t}}.$$

Im Fall $\bar{x} < \lambda \bar{t}$ beginnt die Charakteristik durch $(\bar{x}, \bar{t})$ an der Stelle $(0, \bar{t} - \bar{x}/\lambda) \in \Gamma_{-,t}$. Mit dem Parameter $s = t - (\bar{t} - \bar{x}/\lambda)$ gilt wiederum (11.16) und folglich ist

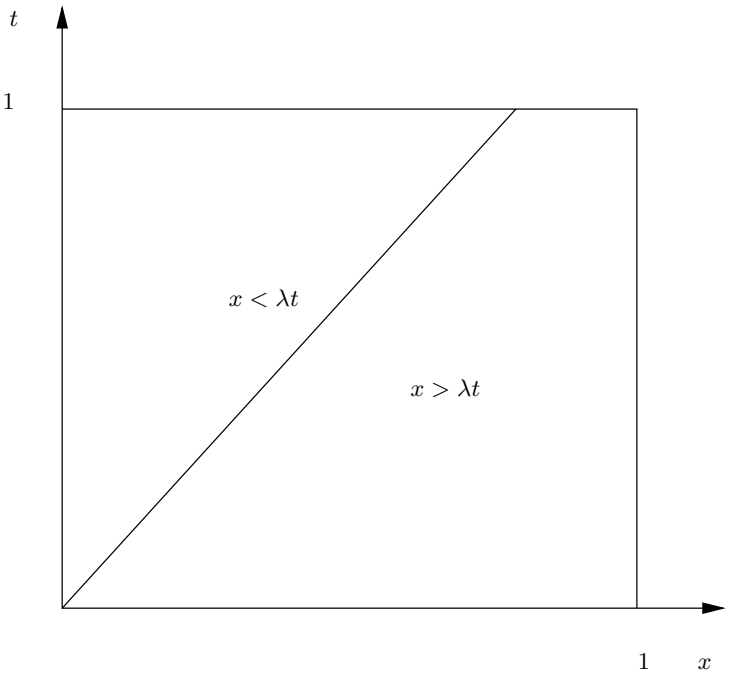

Abbildung 11.3. Illustration zum Beispiel 11.7.

$$u(\bar{x},\bar{t}) = 1 - \mathrm{e}^{-\bar{x}/\lambda}.$$

Insgesamt gilt somit

$$u(x,t) = \begin{cases} 1 - \mathrm{e}^{-t} & \text{für } x \geq \lambda t, \\ 1 - \mathrm{e}^{-x/\lambda} & \text{für } x < \lambda t. \end{cases}$$

Beachten Sie, dass die Lösung an der Stelle $x = \lambda t$ stetig ist, dies allerdings auf die Ableitungen u_t und u_x nicht zutrifft.

Beispiel 11.8. Im gleichen Gebiet wie in Beispiel 11.7 betrachten wir nun das Problem

$$\begin{aligned} &u_t + (1+t)u_x = 0 && \text{in } \Omega \times \mathbf{R}_+, \\ &u = x^2 && \text{auf } \Gamma_-. \end{aligned}$$

Die Charakteristik durch $(\bar{x},\bar{t})$ nimmt nun die Form

$$x = t + \tfrac{1}{2}t^2 + \bar{x} - \bar{t} - \tfrac{1}{2}\bar{t}^2$$

an (siehe Abbildung 11.4) und beginnt an der Stelle $(\bar{x} - \bar{t} - \frac{1}{2}\bar{t}^2, 0) \in \Gamma_{-,x}$ für $\bar{x} \geq \bar{t} + \frac{1}{2}\bar{t}^2$ und an irgendeiner Stelle auf $\Gamma_{-,t}$ für $\bar{x} < \bar{t} + \frac{1}{2}\bar{t}^2$. Die Lösung lautet deshalb

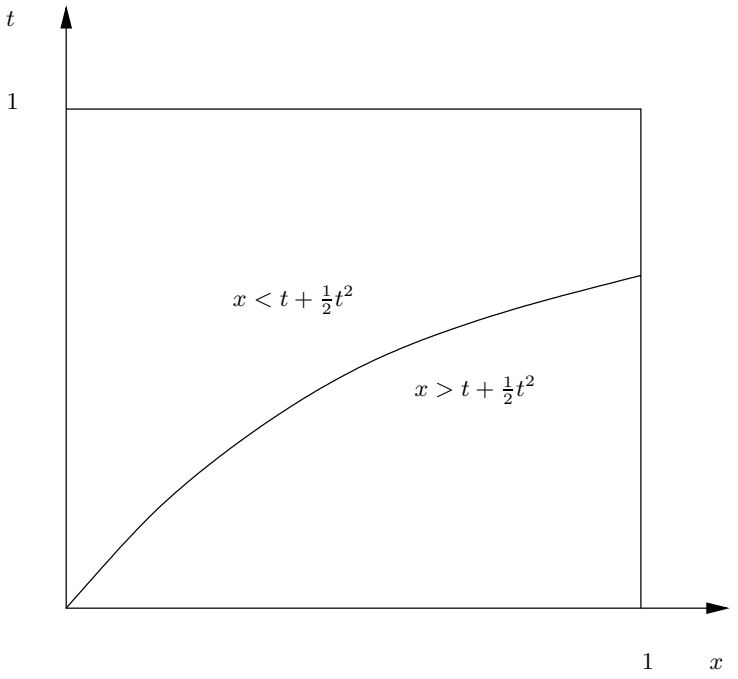

Abbildung 11.4. Illustration zum Beispiel 11.8.

$$u(x,t) = \begin{cases} (x - t - \frac{1}{2}t^2)^2 & \text{für } x \geq t + \frac{1}{2}t^2, \\ 0 & \text{für } x < t + \frac{1}{2}t^2. \end{cases}$$

11.4 Symmetrische hyperbolische Systeme

Wir betrachten zunächst das Anfangswertproblem

$$\begin{aligned} &\frac{\partial u}{\partial t} + A(x,t)\frac{\partial u}{\partial x} + B(x,t)u = f(x,t) && \text{für } x \in \mathbf{R},\ t > 0, \\ &u(x,0) = v(x) && \text{für } x \in \mathbf{R} \end{aligned} \tag{11.17}$$

in einer räumlichen Dimension, wobei $u = u(x,t)$ und $f = f(x,t)$ N-vektorwertige Funktionen und $A = A(x,t)$ und $B = B(x,t)$ glatte $N \times N$-Matrizen sind. Die Matrix A ist symmetrisch. Sie besitzt dann reelle Eigenwerte $\{\lambda_j\}_{j=1}^N$ mit $\lambda_j = \lambda_j(x,t)$, und wir nehmen zusätzlich an, dass diese verschieden sind. Das System (11.17) wird als *streng hyperbolisch* bezeichnet. Unter dieser Annahme kann man eine glatte orthogonale Matrix $P = P(x,t)$ finden, die A diagonalisiert, sodass

$$P^{\mathrm{T}} A P = \Lambda = \operatorname{diag}(\lambda_j)_{j=1}^N$$

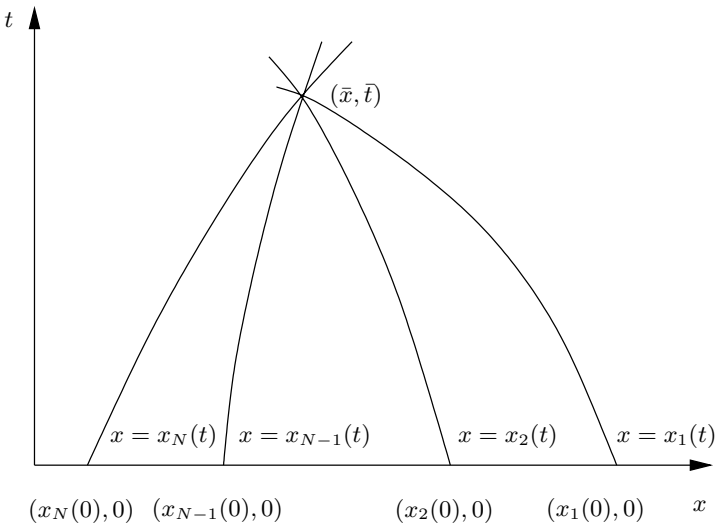

Abbildung 11.5. Charakteristische Kurven. Abhängigkeitsbereich.

gilt (siehe Problemstellung 11.17). Führen wir mit $u = Pw$ eine neue abhängige Variable w ein, so erhalten wir

$$\frac{\partial u}{\partial t} + A\frac{\partial u}{\partial x} + Bu = P\frac{\partial w}{\partial t} + AP\frac{\partial w}{\partial x} + \Big(\frac{\partial P}{\partial t} + A\frac{\partial P}{\partial x} + BP\Big)w = f$$

oder

$$\frac{\partial w}{\partial t} + \Lambda\frac{\partial w}{\partial x} + \tilde{B}w = P^{\mathrm{T}}f \quad \text{mit } \tilde{B} = P^{\mathrm{T}}\Big(\frac{\partial P}{\partial t} + A\frac{\partial P}{\partial x} + BP\Big).$$

Dies ist ein System der Form (11.17), wobei A diagonal ist.

Wir nehmen nun ohne Beschränkung der Allgemeinheit an, dass A in (11.17) selbst eine diagonale Matrix ist und dass die Eigenwerte λ_j in wachsender Reihenfolge $\lambda_1 < \lambda_2 < \ldots < \lambda_N$ angeordnet sind.

Betrachten wir zunächst den Fall $B = 0$. Das System besteht dann aus N ungekoppelten Gleichungen

$$\frac{\partial u_j}{\partial t} + \lambda_j(x,t)\frac{\partial u_j}{\partial x} = f_j(x,t) \quad \text{mit } u_j(x,0) = v_j(x) \quad \text{für } j = 1,\ldots,N,$$

von denen jede ein skalares Problem von der in Abschnitt 11.3 betrachteten Form ist. Zu jedem j existiert eine Charakteristik durch $(\bar{x},\bar{t})$, die durch

$$\frac{dx}{dt} = \lambda_j(x,t) \quad \text{mit } x(\bar{t}\,) = \bar{x}$$

bestimmt ist. Bezeichnen wir die Lösung dieses Anfangswertproblems mit $x_j(t)$, sodass die Charakteristik durch $(\bar{x},\bar{t})$ durch die Gleichung $x = x_j(t)$ gegeben ist, dann gilt

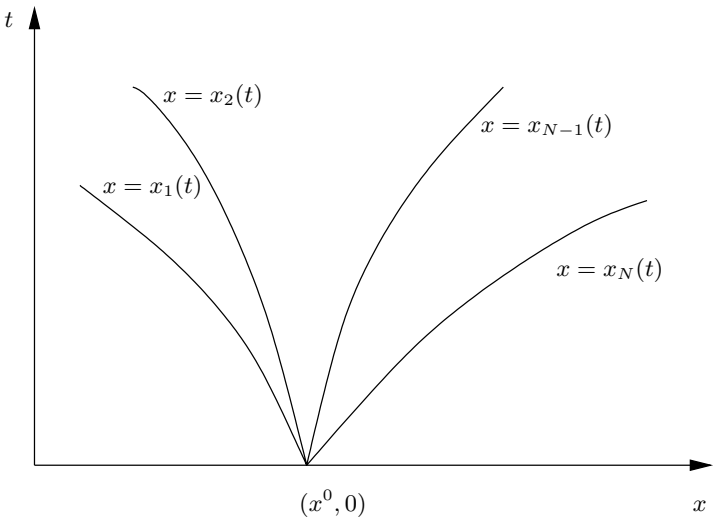

Abbildung 11.6. Einflussgebiet.

$$u_j(\bar{x},\bar{t}) = v_j(x_j(0)) + \int_0^{\bar{t}} f_j(x_j(s),s)\,\mathrm{d}s. \tag{11.18}$$

Somit hängt $u_j(\bar{x},\bar{t})$ von v_j an nur einem Punkt und von f_j entlang der Charakteristik durch $(\bar{x},\bar{t})$ ab (siehe Abbildung 11.5).

Betrachten wir nun den Fall $B \neq 0$. Wir können dann ein iteratives Verfahren zur Lösung von (11.17) mit dem Ansatz

$$u^0 = 0 \quad \text{in } \mathbf{R}\times\mathbf{R}_+$$

benutzen, wobei u^{k+1} für $k \geq 0$ durch

$$\begin{aligned} &\frac{\partial u^{k+1}}{\partial t} + A\frac{\partial u^{k+1}}{\partial x} = f - Bu^k && \text{in } \mathbf{R}\times\mathbf{R}_+,\\ &u^{k+1}(\cdot,0) = v && \text{in } \mathbf{R} \end{aligned}$$

definiert ist. Unter Verwendung von (11.18) ergibt sich

$$\begin{aligned} &u^0 = 0,\\ &u_j^{k+1}(\bar{x},\bar{t}) = v_j(x_j(0)) + \int_0^{\bar{t}} (f - Bu^k)_j(x_j(s),s)\,\mathrm{d}s, \qquad k \geq 0. \end{aligned} \tag{11.19}$$

Es ist nicht schwierig zu beweisen, dass u^k für $k \to \infty$ gegen eine Lösung von (11.17) konvergiert, sodass folgendes Theorem gilt (siehe Problemstellung 7.4).

Theorem 11.4. *Das streng hyperbolische System* (11.17) *besitzt eine Lösung, wenn A, B, f und v hinreichend glatt sind. Wenn A diagonal ist, erhält man diese Lösung aus dem iterativen Verfahren* (11.19).

Die Eindeutigkeit der Lösung folgt aus Theorem 11.5.

Sehen wir uns (11.19) und Abbildung 11.5 an, dann bemerken wir, dass lediglich die Werte von v im Intervall $(x_N(0), x_1(0))$ in die sukzessive Definition der u^k eingehen und dass f und B nur in dem durch die extremalen Charakteristiken $x = x_N(t)$ und $x = x_1(t)$ bestimmten Dreieck berechnet werden. Dieses definiert folglich den Abhängigkeitsbereich der Lösung an der Stelle $(\bar{x}, \bar{t})$ von den Daten. Analog dazu beeinflussen die Anfangswerte in einem Punkt $(x^0, 0)$ für $t > 0$ nur die Lösung in dem keilförmigen Gebiet zwischen den Charakteristiken, die zu λ_1 und λ_N gehören und deren Ursprung an der Stelle $(x^0, 0)$ liegt (siehe Abbildung 11.6).

Beispiel 11.9. Betrachten wir das Anfangswertproblem für die Wellengleichung

$$\begin{aligned} &\frac{\partial^2 u}{\partial t^2} = \frac{\partial^2 u}{\partial x^2} && \text{in } \mathbf{R} \times \mathbf{R}_+ \\ &u(\cdot, 0) = v,\ \frac{\partial u}{\partial t}(\cdot, 0) = w && \text{in } \mathbf{R}. \end{aligned} \tag{11.20}$$

Wir führen die neuen Variablen $U_1 = \partial u/\partial t$, $U_2 = \partial u/\partial x$ ein. Nun ergibt sich für $U = (U_1, U_2)^{\mathrm{T}}$ das System

$$\begin{aligned} &\frac{\partial U_1}{\partial t} - \frac{\partial U_2}{\partial x} = 0 \\ &\frac{\partial U_2}{\partial t} - \frac{\partial U_1}{\partial x} = 0 && \text{in } \mathbf{R} \times \mathbf{R}_+, \\ &U_1(\cdot, 0) = w,\ U_2(\cdot, 0) = v' && \text{in } \mathbf{R}, \end{aligned}$$

beziehungsweise

$$\frac{\partial U}{\partial t} + A \frac{\partial U}{\partial x} = 0 \quad \text{mit } U(x, 0) = \begin{bmatrix} w(x) \\ v'(x) \end{bmatrix}$$

mit $A = \begin{bmatrix} 0 & -1 \\ -1 & 0 \end{bmatrix}$. Die Eigenwerte von A sind $\lambda_1 = -1$, $\lambda_2 = 1$. Setzen wir

$$P = \frac{1}{\sqrt{2}} \begin{bmatrix} 1 & 1 \\ 1 & -1 \end{bmatrix}, \quad U = PV,$$

ergibt sich für die neuen abhängigen Variablen $V = (V_1, V_2)^{\mathrm{T}}$ das System

$$\frac{\partial V}{\partial t} + \begin{bmatrix} -1 & 0 \\ 0 & 1 \end{bmatrix} \frac{\partial V}{\partial x} = 0 \quad \text{in } \mathbf{R} \times \mathbf{R}_+,$$

beziehungsweise

$$\begin{aligned} &\frac{\partial V_1}{\partial t} - \frac{\partial V_1}{\partial x} = 0, \\ &\frac{\partial V_2}{\partial t} + \frac{\partial V_2}{\partial x} = 0. \end{aligned}$$

Folglich gilt

$$V_1(x,t) = V_1(x+t,0), \quad V_2(x,t) = V_2(x-t,0).$$

Kehren wir zu den ursprünglichen Variablen U zurück, dann können wir damit die Formel von d'Alembert (11.10) für die Lösung von (11.20) herleiten (siehe Problemstellung 11.5).

Betrachten wir nun die Verallgemeinerung des Systems (11.17) auf d räumliche Dimensionen

$$\begin{aligned} &\frac{\partial u}{\partial t} + \sum_{j=1}^{d} A_j \frac{\partial u}{\partial x_j} + Bu = f && \text{in } \mathbf{R}^d \times \mathbf{R}_+, \\ &u(\cdot,0) = v && \text{in } \mathbf{R}^d, \end{aligned} \tag{11.21}$$

wobei $u = u(x,t)$ eine N-vektorwertige Funktion, $A_j = A_j(x,t)$ symmetrische $N \times N$-Matrizen, $B = B(x,t)$ eine $N \times N$-Matrix und $f = f(x,t)$ sowie $v = v(x)$ Vektoren mit N Komponenten sind, von denen alle glatt und beschränkt von ihren Variablen abhängen. Wir nehmen außerdem an, dass die Lösungen für große $|x|$ in dem Sinne klein sind, dass folgende Analyse zutrifft.

Ein Gleichungssystem wie in (11.17) wird als *symmetrisches hyperbolisches System* oder *Friedrichs-System* bezeichnet. Ein Spezialfall sind die Maxwellschen Gleichungen in der Elektrodynamik (siehe Problemstellung 11.15). Die klassische Wellengleichung $u_{tt} = \Delta u$ kann in ein symmetrisches hyperbolisches System transformiert werden, indem man die Ableitungen erster Ordnung als neue abhängige Variablen einführt. Dies sollen Sie in Problemstellung 11.11 überprüfen. Allgemein können viele andere wichtige Gleichungen der mathematischen Physik als symmetrische hyperbolische Systeme geschrieben werden. Mitunter ist dazu eine Variablentransformation nötig.

Nach Gleichung (11.3) sind die charakteristischen Richtungen $(\xi,\tau) = (\xi_1,\dots,\xi_d,\tau)$ durch

$$\det \Lambda(\xi,\tau) = \det\Big(\tau I + \sum_{j=1}^{d} \xi_j A_j\Big) = 0$$

gegeben. Es ist klar, dass diese Gleichung für jedes gegebene ξ genau N reelle Wurzeln $\tau_j(\xi)$, $j = 1,\dots,N$ besitzt. Dabei handelt es sich um die Eigenwerte der symmetrischen $N \times N$-Matrix $-\sum_{j=1}^{d} \xi_j A_j$.

Im Allgemeinen ist es für $d > 1$ nicht möglich, die Matrizen A_j gleichzeitig zu diagonalisieren. Folglich kann das Problem nicht wie vorhin behandelt werden. Wir werden uns deshalb auf das Anwenden der Energiemethode beschränken, wenn wir Stabilitätsabschätzungen bezüglich $\|\cdot\| = \|\cdot\|_{L_2(\mathbf{R}^d)}$ für dieses Problem zeigen.

Theorem 11.5. *Für die Lösung von* (11.21) *gilt mit* $C = C(T)$

$$\|u(t)\| \le C\Big(\|v\| + \Big(\int_0^T \|f\|^2 \,\mathrm{d}s\Big)^{1/2}\Big) \quad \textit{für } 0 \le t \le T.$$

Beweis. Wenn wir die Gleichung mit u multiplizieren und über $\mathbf{R}^d$ integrieren, erhalten wir

$$(\frac{\partial u}{\partial t}, u) + \sum_{j=1}^{d} (A_j \frac{\partial u}{\partial x_j}, u) + (Bu, u) = (f, u).$$

Dabei gilt

$$(\frac{\partial u}{\partial t}, u) = \int_\Omega \langle \frac{\partial u}{\partial t}, u\rangle \,\mathrm{d}x = \tfrac{1}{2}\frac{d}{dt}\|u\|^2$$

und

$$(A_j \frac{\partial u}{\partial x_j}, u) = \int_{\mathbf{R}^d} \langle A_j \frac{\partial u}{\partial x_j}, u\rangle \,\mathrm{d}x$$

mit dem gewöhnlichen Skalarprodukt $\langle \cdot, \cdot \rangle$ in $\mathbf{R}^N$. Wir erhalten

$$\frac{\partial}{\partial x_j}\langle A_j u, u\rangle = \langle \frac{\partial A_j}{\partial x_j} u, u\rangle + \langle A_j \frac{\partial u}{\partial x_j}, u\rangle + \langle A_j u, \frac{\partial u}{\partial x_j}\rangle.$$

Weil A_j symmetrisch ist, sind die letzten beiden Terme gleich. Nehmen wir darüber hinaus an, dass u für $|x|$ groß ist, dann gilt

$$\int_{\mathbf{R}^d} \frac{\partial}{\partial x_j}\langle A_j u, u\rangle \,\mathrm{d}x = 0$$

und folglich

$$(A_j \frac{\partial u}{\partial x_j}, u) = -\tfrac{1}{2}(\frac{\partial A_j}{\partial x_j} u, u).$$

Daraus folgt

$$\tfrac{1}{2}\frac{d}{dt}\|u\|^2 + (\tilde{B}u, u) \le \|f\| \, \|u\|$$

mit

$$\tilde{B} = B - \tfrac{1}{2}\sum_{j=1}^{d} \frac{\partial A_j}{\partial x_j}.$$

Somit ist die Ungleichung

$$\frac{d}{dt}\|u\|^2 \le 2\|\tilde{B}\|_{\mathcal{C}}\|u\|^2 + \|f\|\|u\| \le C_0\|u\|^2 + C_1\|f\|^2$$

erfüllt. Hieraus ergibt sich

$$\|u(t)\|^2 \le \|v\|^2 + C_1 \int_0^T \|f\|^2 \,\mathrm{d}s + C_0 \int_0^t \|u\|^2 \,\mathrm{d}s \quad \text{für } 0 \le t \le T,$$

sodass mit dem Gronwall-Lemma (siehe Problemstellung 7.6)

$$\|u(t)\|^2 \le \mathrm{e}^{C_0 T}\Big(\|v\|^2 + C_1 \int_0^T \|f\|^2 \,\mathrm{d}s\Big) \quad \text{für } 0 \le t \le T$$

gilt. □

Wir gewöhnlich folgt aus dieser Ungleichung die Eindeutigkeit und die Stabilität für die Lösungen des Problems (11.21). Die Existenz einer Lösung lässt sich beispielsweise zeigen, indem man eine Finite-Differenzen-Approximation auf einem Gitter mit dem Gitterabstand h konstruiert und anschließend die Konvergenz für $h \to 0$ zeigt.

Wie im vorhin behandelten Fall einer räumlichen Dimension ist es hier ebenfalls möglich zu zeigen, dass die Werte der Lösung von (11.21) in einem Punkt $(\bar{x}, \bar{t})$ mit $\bar{t} > 0$ nur von den Daten in einem endlichen Gebiet abhängen. Dazu betrachten wir der Einfachheit halber den Fall einer homogenen Gleichung mit konstanten Koeffizienten und ohne den Term Bu, sodass das Problem nun die Form

$$\begin{aligned} &\frac{\partial u}{\partial t} + \sum_{j=1}^{d} A_j \frac{\partial u}{\partial x_j} = 0 \qquad && \text{in } \mathbf{R} \times \mathbf{R}_+, \\ &u(\cdot, 0) = v && \text{in } \mathbf{R}^d \end{aligned}$$

besitzt. Dann ist das charakteristische Polynom die symmetrische Matrix

$$\Lambda(\xi, \tau) = \tau I + \sum_{j=1}^{d} \xi_j A_j.$$

Betrachten wir nun einen Kreiskegel K mit dem Scheitel $(\bar{x}, \bar{t})$, der auf $t \le \bar{t}$ beschränkt ist (vgl. (11.9)) und dessen Öffnungswinkel so gewählt wurde, dass mit der äußeren Einheitsnormale (n_x, n_t) an die Mantelfläche M des Kegels der Ausdruck $\Lambda(n_x, n_t)$ positiv definit wird. Die Möglichkeit, einen solchen Kegel zu bestimmen, folgt aus der Tatsache, dass für die Richtung (0,1) $\Lambda(0, 1) = I$ gilt, was positiv definit ist, und $\Lambda(\xi, 1)$ folglich auch für alle kleinen $|\xi|$ positiv definit ist.

Sei B_0 das durch den Kegel aus der Ebene $t = 0$ herausgeschnittene Gebiet. Wir behaupten, dass im Falle $v = 0$ in B_0 die Gleichung $u(\bar{x}, \bar{t}) = 0$ gilt.

Um dies zu beweisen, benutzen wir wiederum die Energiemethode. Wir multiplizieren die Gleichung mit u und integrieren über K. Unter der Annahme, dass die A_j symmetrisch und konstant sind, erhalten wir

$$\begin{aligned} 0 &= \int_K \Big(\langle \frac{\partial u}{\partial t}, u \rangle + \sum_{j=1}^{d} \langle A_j \frac{\partial u}{\partial x_j}, u \rangle \Big) \, \mathrm{d}x \, \mathrm{d}t \\ &= \tfrac{1}{2} \int_K \Big(\frac{\partial}{\partial t} \langle u, u \rangle + \sum_{j=1}^{d} \frac{\partial}{\partial x_j} \langle A_j u, u \rangle \Big) \, \mathrm{d}x \, \mathrm{d}t. \end{aligned}$$

Aufgrund des Divergenztheorems gilt

$$\int_M \Big(\langle u, u \rangle \, n_t + \sum_{j=1}^{d} \langle A_j u, u \rangle \, n_{x_j} \Big) \, \mathrm{d}s = \int_{B_0} \langle u, u \rangle \, \mathrm{d}x$$

oder wegen $u = 0$ in B_0

$$\int_M \langle \Lambda(n_x, n_t)u, u\rangle \, \mathrm{d}s = 0.$$

Daraus folgt $u = 0$ auf M, da $\Lambda(n_x, n_t)$ positiv definit ist. Insbesondere gilt $u(\bar{x}, \bar{t}) = 0$, also unsere Behauptung.

11.5 Problemstellungen

Problem 11.1. Bestimmen Sie die Charakteristiken der Tricomi-Gleichung

$$\frac{\partial^2 u}{\partial x_1^2} + x_1 \frac{\partial^2 u}{\partial x_2^2} = f \quad \text{für } x = (x_1, x_2) \in \mathbf{R}^2.$$

Problem 11.2. Bestimmen Sie die charakteristischen Richtungen für die Cauchy-Riemannschen Gleichungen

$$\frac{\partial u}{\partial x} - \frac{\partial v}{\partial y} = 0, \quad \frac{\partial u}{\partial y} + \frac{\partial v}{\partial x} = 0.$$

Problem 11.3. Zeigen Sie, dass (11.7) direkt aus (11.6) folgt.

Problem 11.4. Sei u wie in (11.6). Angenommen, es gilt $v \in H^2 \cap H_0^1$, $w \in H_0^1$. Zeigen Sie

$$\begin{aligned}
\|u(t)\| &\le C\big(\|v\| + \|w\|\big),\\
\|\nabla u(t)\| &\le C\big(\|\nabla v\| + \|w\|\big),\\
\|u_{tt}(t)\| = \|\Delta u(t)\| &\le C\big(\|\Delta v\| + \|\nabla w\|\big),\\
\|u(t) - v\| &\le Ct\big(\|\nabla v\| + \|w\|\big),\\
\|u_t(t) - w\| &\le Ct\big(\|\Delta v\| + \|\nabla w\|\big).
\end{aligned}$$

Somit ist u zumindest im Sinne des L_2 eine Lösung von (11.5). Hinweis: Erinnern Sie sich an Theorem 6.4 und an Problemstellung 6.3. Zeigen Sie

$$\|u(t) - v\|^2 = t^2 \sum_{j=1}^{\infty} \Big(\sqrt{\lambda_j}\hat{v}_j \frac{\cos(\sqrt{\lambda_j}t) - 1}{\sqrt{\lambda_j}t} + \hat{w}_j \frac{\sin(\sqrt{\lambda_j}t)}{\sqrt{\lambda_j}t} \Big)^2.$$

Problem 11.5. Beweisen Sie die Lösungsformel von d'Alembert (11.10) für das Cauchy-Problem für die eindimensionale Wellengleichung, d. h.

$$\begin{aligned}
&u_{tt} - u_{xx} = 0 && \text{in } \mathbf{R} \times \mathbf{R}_+,\\
&u(\cdot, 0) = v,\ u_t(\cdot, 0) = w && \text{in } \mathbf{R}.
\end{aligned}$$

Problem 11.6. (a) Lösen Sie das Anfangswertproblem

$$\begin{aligned} &\frac{\partial u}{\partial t} + \begin{bmatrix} 0 & x \\ x & 0 \end{bmatrix} \frac{\partial u}{\partial x} = 0 \qquad && x \in \mathbf{R},\ t > 0, \\ &u(x,0) = v(x) && x \in \mathbf{R}, \end{aligned}$$

mit der Charakteristikenmethode.
(b) Beweisen Sie eine Stabilitätsabschätzung mit der Energiemethode.

Problem 11.7. (a) Lösen Sie das Anfangswertproblem

$$u_t + (x+t)u_x = 0 \text{ for } (x,t) \in \mathbf{R} \times \mathbf{R}_+ \quad \text{mit } u(x,0) = v(x) \text{ für } x \in \mathbf{R}$$

mit der Charakteristikenmethode.
(b) Zeigen Sie mit der Energiemethode die Gültigkeit von

$$\|u(\cdot,t)\| = \mathrm{e}^{t/2}\|v\| \quad \text{und} \quad \|u_x(\cdot,t)\| = \mathrm{e}^{-t/2}\|v_x\| \quad \text{für } t \geq 0.$$

Überprüfen Sie diese Resultate durch eine direkte Berechnung unter Verwendung der Lösungsformel aus (a).

Problem 11.8. Lösen Sie das Problem

$$x_1 \frac{\partial u}{\partial x_1} - x_2 \frac{\partial u}{\partial x_2} = 0 \quad \text{für } x \in \mathbf{R}^2 \quad \text{mit } u(x) = \varphi(x) \quad \text{für } x \in S,$$

wobei S eine nichtcharakteristische Kurve ist.

Problem 11.9. Lösen Sie das Problem

$$\begin{aligned} &x_1 \frac{\partial u}{\partial x_1} + 2x_2 \frac{\partial u}{\partial x_2} + \frac{\partial u}{\partial x_3} = 3u \qquad && \text{für } x \in \mathbf{R}^3, \\ &u(x_1, x_2, 0) = \varphi(x_1, x_2) && \text{für } (x_1, x_2) \in \mathbf{R}^2. \end{aligned}$$

Problem 11.10. Beweisen Sie die folgende Stabilitätsabschätzung für das Problem (11.11), (11.13) unter einer geeigneten Bedingung an die Koeffizienten a_j:

$$\int_\Omega u^2 \,\mathrm{d}x + \int_{\Gamma_+} u^2\, n \cdot a \,\mathrm{d}s \leq C\Big(\int_\Omega f^2 \,\mathrm{d}x + \int_{\Gamma_-} v^2 |n \cdot a| \,\mathrm{d}s \Big).$$

Problem 11.11. Zeigen Sie, dass die Wellengleichung $u_{tt} - \Delta u = 0$ als ein symmetrisches hyperbolisches System geschrieben werden kann, indem man die Ableitungen erster Ordnung $u_t, u_{x_1}, \ldots, u_{x_d}$ als neue Variablen einführt.

Problem 11.12. Modifizieren Sie den Beweis von Theorem 11.5 um das etwas strengere Resultat zu beweisen:

$$\|u(t)\| \leq C(T)\Big(\|v\| + \int_0^T \|f(s)\| \,\mathrm{d}s \Big) \quad \text{für } 0 \leq t \leq T.$$

Problem 11.13. Nehmen Sie zusätzlich zu den Annahmen aus Theorem 11.5 an, dass die A_j konstant sind und B symmetrisch positiv definit ist. Beweisen Sie

$$\|u(t)\| \le \|v\| + \int_0^t \|f\| \, \mathrm{d}s \quad \text{für } t \ge 0.$$

Problem 11.14. Verallgemeinern Sie Theorem 11.5 auf symmetrische hyperbolische Systeme der Form

$$M\frac{\partial u}{\partial t} + \sum_{j=1}^{d} A_j \frac{\partial u}{\partial x_j} + Bu = f \quad \text{in } \mathbf{R}^d \times \mathbf{R}_+,$$

wobei A_j und B wie vorhin sind und $M = M(x,t)$ symmetrisch positiv definit, gleichmäßig bezüglich x, t, ist, sodass $\langle M(x,t)\xi, \xi\rangle \ge \alpha\,|\xi|^2$ für alle $\xi \in \mathbf{R}^N$, $(x,t) \in \mathbf{R}^d \times \mathbf{R}_+$ mit $\alpha > 0$ gilt.

Problem 11.15. Die Entwicklung des elektrischen Feldes $E(x,t) \in \mathbf{R}^3$ und des magnetischen Feldes $H(x,t) \in \mathbf{R}^3$ in einem homogenen isotropen Raum kann durch die folgenden beiden Maxwellschen Gleichungen (Ampère-Gesetz und Faraday-Gesetz)

$$\begin{aligned} &\frac{1}{c}\frac{\partial E}{\partial t} - \nabla \times H + \frac{4\pi}{c} J = 0 && \text{in } \mathbf{R}^3 \times \mathbf{R}_+, \\ &\frac{1}{c}\frac{\partial H}{\partial t} + \nabla \times E = 0 && \text{in } \mathbf{R}^3 \times \mathbf{R}_+ \end{aligned} \tag{11.22}$$

beschrieben werden. Dabei ist c eine positive Konstante, und es gilt

$$\nabla \times H = \operatorname{curl} H = \Big(\frac{\partial H_3}{\partial x_2} - \frac{\partial H_2}{\partial x_3}, \frac{\partial H_1}{\partial x_3} - \frac{\partial H_3}{\partial x_1}, \frac{\partial H_2}{\partial x_1} - \frac{\partial H_1}{\partial x_2}\Big).$$

Nehmen Sie außerdem an, dass die Stromdichte J das Ohmsche Gesetz $J = \sigma E$ mit einer nichtnegativen Konstante σ erfüllt. Zeigen Sie, dass die Gleichungen (11.22) mit an der Stelle $t = 0$ vorgegebenen Werten für E und H ein gutgestelltes Problem bilden, indem Sie zeigen, dass (11.22) ein Friedrichs-System ist. Was kann man über die Stabilität der Energiedichte $e = \frac{1}{2}(E \cdot E + H \cdot H)$ sagen? Hinweis: Sehen Sie sich Problemstellung 11.13 an.

Problem 11.16. Betrachten Sie noch einmal die Gleichung

$$\rho\frac{\partial^2 u}{\partial t^2} = \frac{\partial}{\partial x}\Big(E\frac{\partial u}{\partial x}\Big)$$

aus Problemstellung 1.2 für die longitudinale Bewegung eines elastischen Stabes.

(a) Nehmen Sie der Einfachheit halber an, dass ρ und E konstant sind und zeigen Sie, dass sie in Form des symmetrischen hyperbolischen Systems

$$\begin{bmatrix} \rho & 0 \\ 0 & E \end{bmatrix} U_t - \begin{bmatrix} 0 & E \\ E & 0 \end{bmatrix} U_x = 0$$

in den Variablen $U_1 = u_t$, $U_2 = u_x$ geschrieben werden kann (siehe Problemstellung 11.14).
(b) Gehen Sie beispielsweise von den Randbedingungen $u(0) = 0$, $u_x(L) = 0$ aus. Zeigen Sie, dass die mechanische Energie eine Erhaltungsgröße ist, d. h. für $e = \frac{1}{2}(\rho u_t^2 + E u_x^2)$

$$\int_0^L e(x,t)\,\mathrm{d}x = \int_0^L e(x,0)\,\mathrm{d}x$$

gilt.

Problem 11.17. Berechnen Sie die Eigenwerte und normierten Eigenvektoren der symmetrischen Matrix $A(x,t) = \begin{bmatrix} x & t \\ t & -x \end{bmatrix}$. Zeigen Sie, dass die Eigenvektormatrix $P(x,t)$ an der Stelle $x = 0,\ t = 0$ unstetig ist. Dort sind die Eigenwerte entartet.

12

Finite Differenzenverfahren für hyperbolische Gleichungen

Die Lösung hyperbolischer Gleichungen ist vielleicht das Gebiet, auf dem finite Differenzenverfahren am erfolgreichsten sind und auch weiterhin eine wichtige Rolle spielen werden. Dies trifft insbesondere auf nichtlineare Erhaltungsgleichungen zu, deren Behandlung jedoch über den Umfang dieser elementaren Darstellung hinausgeht. Wir beginnen in Abschnitt 12.1 mit dem reinen Anfangswertproblem für eine skalare Gleichung erster Ordnung in einer räumlichen Variable und untersuchen die Stabilität und Fehlerabschätzungen für das grundlegende Upwind-Verfahren, das Friedrichs-Verfahren und das Lax-Wendroff-Verfahren. In Abschnitt 12.2 erweitern wir diese Betrachtungen auf symmetrische hyperbolische Systeme und ebenso auf höhere räumliche Dimensionen. In Abschnitt 12.3 behandeln wir das Wendroff-Box-Schema für ein gemischtes Anfangs-Randwertproblem in einer räumlichen Dimension.

12.1 Skalare Gleichungen erster Ordnung

In diesem ersten Abschnitt betrachten wir das einfache Modell-Anfangswertproblem

$$\begin{aligned} \frac{\partial u}{\partial t} &= a\frac{\partial u}{\partial x} && \text{in } \mathbf{R}\times\mathbf{R}_+, \\ u(\cdot,0) &= v && \text{in } \mathbf{R} \end{aligned} \tag{12.1}$$

mit der Konstanten a. Wir erinnern uns daran, dass dieses Problem für $v \in \mathcal{C}^1$ die eindeutige klassische Lösung

$$u(x,t) = (E(t)v)(x) = v(x+at) \tag{12.2}$$

zulässt, die man dadurch bestimmen kann, dass man die Charakteristik $x + at$ = konstant durch (x,t) bis zur Stelle $t = 0$ zurückverfolgt und den Wert von v an diesem Punkt verwendet. Eine ähnliche Aussage gilt für einen variablen Koeffizienten $a = a(x)$. In diesem Fall ist die Charakteristik gekrümmt. Da

der Lösungsoperator lediglich eine Verschiebung des Argumentes bewirkt, sind sowohl die Maximumnorm als auch die L_2-Norm zeitlich konstant, es gilt also

$$\|E(t)v\|_{\mathcal{C}} = \|v\|_{\mathcal{C}} \quad \text{und} \quad \|E(t)v\| = \|v\| \quad \text{für } t \geq 0. \tag{12.3}$$

Insbesondere ist $E(t)$ in beiden Normen stabil.

Um das Modellproblem mithilfe des finiten Differenzenverfahrens approximativ zu lösen, führen wir wie bereits im Falle parabolischer Gleichungen in Abschnitt 9.1 einen räumlichen Gitterabstand h und einen Zeitschritt k ein und bezeichnen die Approximation von $u(x,t)$ an der Stelle $(x_j, t_n) = (jh, nk)$ mit U_j^n für $j,\ n \in \mathbf{Z},\ n \geq 0$. Dabei ist $\mathbf{Z} = \{\dots, -1, -2, 0, 1, 2, \dots\}$ die Menge der ganzen Zahlen. Unter der Annahme, dass $a > 0$ gilt, ersetzen wir (12.1) durch

$$\begin{aligned} &\partial_t U_j^n = a\partial_x U_j^n && \text{für } j, n \in \mathbf{Z},\ n \geq 0, \\ &U_j^0 = V_j = v(x_j) && \text{für } j \in \mathbf{Z}, \end{aligned} \tag{12.4}$$

wobei ∂_t und ∂_x wie vorhin Vorwärts-Differenzenquotienten sind, sodass die Differenzengleichung

$$\frac{U_j^{n+1} - U_j^n}{k} = a\,\frac{U_{j+1}^n - U_j^n}{h}$$

lautet.

Führen wir diesmal das Gitterverhältnis $\lambda = k/h$ ein, das wir für h und k gegen null konstant halten, sehen wir, dass (12.4) ein explizites Verfahren darstellt, das die Approximation an der Stelle $t = t_{n+1}$ durch

$$U_j^{n+1} = (E_k U^n)_j = a\lambda U_{j+1}^n + (1 - a\lambda)U_j^n \quad \text{für } j, n \in \mathbf{Z},\ n \geq 0 \tag{12.5}$$

definiert.

Wenn wir U^n für alle x in $\mathbf{R}$ und nicht nur an den Gitterpunkten $x = x_j$ definiert betrachten, können wir

$$U^{n+1}(x) = (E_k U^n)(x) = a\lambda U^n(x+h) + (1 - a\lambda)U^n(x),\ x \in \mathbf{R} \tag{12.6}$$

schreiben. Durch Iteration finden wir für die approximative Lösung an der Stelle $t = t_n$

$$U^n(x) = (E_k^n v)(x) \quad \text{für } x \in \mathbf{R}.$$

Wie im Falle der Wärmeleitungsgleichung stellen wir fest, dass E_k in der Maximumnorm für $a\lambda \leq 1$ stabil ist, da die Koeffizienten von E_k dann positiv sind und sich zu 1 aufaddieren. Es gilt also

$$\|E_k v\|_{\mathcal{C}} \leq \|v\|_{\mathcal{C}}$$

und somit auch

$$\|U^n\|_{\mathcal{C}} = \|E_k^n v\|_{\mathcal{C}} \leq \|v\|_{\mathcal{C}}.$$

Man kann sich leicht davon überzeugen, dass die Bedingung $a\lambda \leq 1$ für die Stabilität notwendig ist. Wie vorhin folgt aus der Stabilität die Konvergenz:

Theorem 12.1. *Seien U^n und u durch* (12.6) *und* (12.1) *definiert und sei $0 < a\lambda \le 1$. Dann ist*

$$\|U^n - u^n\|_{\mathcal{C}} \le C t_n h |v|_{\mathcal{C}^2} \quad \textit{für } t_n \ge 0.$$

Beweis. Wir führen den Rundungsfehler

$$\tau^n(x) := \partial_t u^n(x) - a\partial_x u^n(x) \tag{12.7}$$

ein und erhalten mit $I_n = (t_n, t_{n+1})$ durch Taylor-Entwicklung für eine exakte Lösung u der Differentialgleichung

$$\begin{aligned} |\tau^n(x)| &\le |\partial_t u^n(x) - u_t(x,t_n)| + a|\partial_x u^n(x) - a u_x(x,t_n)| \\ &\le C(h+k) \max_{t\in I_n} \big(|u_{tt}(\cdot,t)| + |u_{xx}(\cdot,t)|\big) \le Ch|v|_{\mathcal{C}^2}. \end{aligned} \tag{12.8}$$

Dabei haben wir $k \le \lambda h$, $u_{tt} = a u_{xx}$ und $|u_{xx}(\cdot,t)|_{\mathcal{C}} \le |v|_{\mathcal{C}^2}$ benutzt.

Wir können (12.7) auch in der Form

$$u^{n+1}(x) = E_k u^n(x) + k\tau^n(x) \quad \text{für } x \in \mathbf{R}$$

schreiben. Setzen wir $z^n = U^n - u^n$, erhalten wir daher

$$z^{n+1} = E_k z^n - k\tau^n$$

oder durch wiederholte Anwendung

$$z^n = E_k^n z^0 - k \sum_{j=0}^{n-1} E_k^{n-1-j} \tau^j.$$

Wegen $z^0 = U^0 - u^0 = v - v = 0$ schlussfolgern wir aufgrund der Stabilität und wegen (12.8)

$$\|z^n\|_{\mathcal{C}} \le k \sum_{j=0}^{n-1} \|\tau^j\|_{\mathcal{C}} \le C\, nk\, h |v|_{\mathcal{C}^2} \quad \text{für } t_n \ge 0,$$

was den Beweis des Theorems abschließt. □

Beachten Sie, dass die natürliche Wahl der finiten Differenzenapproximation im Falle $a < 0$ anstelle von (12.4)

$$\partial_t U_j^n = a\bar{\partial}_x U_j^n \quad \text{für } n \ge 0 \tag{12.9}$$

lautet oder, entsprechend (12.5),

$$U_j^{n+1} = (E_k U^n)_j = -a\lambda U_{j-1}^n + (1 + a\lambda) U_j^n \quad \text{für } j \in \mathbf{Z},\ n \ge 0$$

ist. Die Stabilitätsbedingung lautet nun $0 < -a\lambda \le 1$. Weil sowohl (12.4) als auch (12.9) Punkte in der Richtung des Flusses verwenden, werden diese Differenzenverfahren als *Upwind-Verfahren* bezeichnet.

Betrachten wir allgemeiner ein explizites finites Differenzenverfahren

$$U_j^{n+1} = (E_k U^n)_j = \sum_p a_p U_{j-p}^n \quad \text{für } j, n \in \mathbf{Z},\ n \geq 0, \tag{12.10}$$

wobei $a_p = a_p(\lambda)$ mit $\lambda = k/h =$ konstant ist oder

$$\begin{aligned} &U^{n+1}(x) = (E_k U^n)(x) = \sum_p a_p U^n(x - ph) \quad \text{für } x \in \mathbf{R},\ n \geq 0, \\ &U^0(x) = v(x) \quad \text{für } x \in \mathbf{R} \end{aligned} \tag{12.11}$$

gilt.

Wir sagen, dass ein solches Verfahren von der *Genauigkeit der Ordnung r* ist, wenn

$$\tau^n = k^{-1}(u^{n+1} - E_k u^n) = O(h^r) \quad \text{für } h \to 0$$

gilt, wobei u die exakte Lösung von (12.1) und $k/h = \lambda =$ konstant ist.

Wir stellen fest, dass wie in Abschnitt 9.1 für die Fourier-Transformation von $E_k v$

$$(E_k v)\hat{}\,(\xi) = \tilde{E}(h\xi)\hat{v}(\xi) \quad \text{mit } \tilde{E}(\xi) = \sum_p a_p \mathrm{e}^{-\mathrm{i}p\xi}$$

gilt. In derselben Weise wie im parabolischen Fall ist eine notwendige und hinreichende Bedingung für die Stabilität in L_2 die *von Neumannsche Bedingung*

$$|\tilde{E}(\xi)| \leq 1 \quad \text{für } \xi \in \mathbf{R}. \tag{12.12}$$

Für unser oben vorgestelltes Verfahren (12.5) gilt

$$\tilde{E}(\xi) = a\lambda \mathrm{e}^{\mathrm{i}\xi} + 1 - a\lambda.$$

Da ξ variiert, liegt $\tilde{E}(\xi)$ auf einem Kreis in der komplexen Ebene mit dem Mittelpunkt an der Stelle $(1 - a\lambda, 0)$ und dem Radius $a\lambda$. Für die Gültigkeit von (12.12) ist also die Bedingung $a\lambda \leq 1$ notwendig und hinreichend. Das ist unsere alte Stabilitätsbedingung.

Wie für das parabolische Problem in Abschnitt 9.1 kann die Definition für die Genauigkeit des Verfahrens auch als Funktion des trigonometrischen Polynoms $\tilde{E}(\xi)$ ausgedrückt werden: Das Verfahren ist genau dann von der Genauigkeit der Ordnung r, wenn

$$\tilde{E}(\xi) = \mathrm{e}^{\mathrm{i}a\lambda\xi} + O(\xi^{r+1}) \quad \text{für } \xi \to 0 \tag{12.13}$$

gilt. Beim Beweis benutzen wir die Tatsache, dass für die exakte Lösung

$$(E(t)v)\hat{}\,(\xi) = \int_{\mathbf{R}} v(x + at)\mathrm{e}^{-\mathrm{i}x\xi}\,\mathrm{d}x = \mathrm{e}^{\mathrm{i}at\xi}\hat{v}(\xi)$$

gilt. Wie in Abschnitt 9.1 kann man dann die folgende Fehlerabschätzung beweisen.

Theorem 12.2. *Seien U^n und u durch* (12.11) *beziehungsweise* (12.1) *definiert. Angenommen, E_k ist von der Genauigkeit der Ordnung r und in L_2 stabil. Dann ist*

$$\|U^n - u^n\| \le Ct_n h^r |v|_{r+1} \quad \textit{für } t_n \ge 0.$$

Eine andere natürliche Wahl für eine Differenzenapproximation von (12.1) besteht darin, die Ableitung bezüglich x durch einen symmetrischen Differenzenquotienten

$$\hat{\partial}_x U^n(x) = \frac{U^n(x+h) - U^n(x-h)}{2h}$$

zu ersetzen. Dies führt auf die finite Differenzengleichung

$$\frac{U^{n+1}(x) - U^n(x)}{k} = a\,\frac{U^n(x+h) - U^n(x-h)}{2h} \tag{12.14}$$

und daher auf das Differenzenverfahren

$$U^{n+1}(x) = (E_k U^n)(x) = U^n(x) + \tfrac{1}{2}a\lambda(U^n(x+h) - U^n(x-h)),\ n \ge 0,$$
$$U^0(x) = v(x), \quad x \in \mathbf{R}.$$

In diesem Fall ist das charakteristische Polynom von E_k

$$\tilde{E}(\xi) = 1 + \tfrac{1}{2}a\lambda(\mathrm{e}^{\mathrm{i}\xi} - \mathrm{e}^{-\mathrm{i}\xi}) = 1 + a\lambda \mathrm{i}\sin\xi.$$

Da außer an der Stelle $\xi = m\pi$

$$|\tilde{E}(\xi)|^2 = 1 + a^2\lambda^2 \sin^2\xi > 1$$

gilt, schlussfolgern wir, dass diese Methode für jede Wahl von λ instabil ist.

Dieses Verfahren kann stabilisiert werden, indem man $U^n(x)$ auf der linken Seite von (12.14) durch den Mittelwert $\frac{1}{2}(U^n(x+h)+U^n(x-h))$ ersetzt. Dies führt auf

$$\frac{U^{n+1}(x) - \frac{1}{2}(U^n(x+h) + U^n(x-h))}{k} = a\,\frac{U^n(x+h) - U^n(x-h)}{2h}$$

oder

$$U^{n+1}(x) = (E_k U^n)(x) = \tfrac{1}{2}(1+a\lambda)U^n(x+h) + \tfrac{1}{2}(1-a\lambda)U^n(x-h).$$

Dies ist ein Spezialfall des *Friedrichs-Verfahrens*, das wir später allgemeiner untersuchen werden. Hier gilt

$$\tilde{E}(\xi) = \cos\xi + \mathrm{i}a\lambda\sin\xi$$

und wir erhalten

$$|\tilde{E}(\xi)|^2 = \cos^2\xi + a^2\lambda^2\sin^2\xi \le 1 \quad \text{für } \xi \in \mathbf{R}$$

genau dann, wenn $|a\lambda| \leq 1$ gilt.

Dieser Fall des Friedrichs-Verfahrens kann auch in der Form

$$\frac{U^{n+1}(x) - U^n(x)}{k} = a\,\frac{U^n(x+h) - U^n(x-h)}{2h} + \frac{1}{2k}\Big(U^n(x+h) - 2U^n(x) + U^n(x-h)\Big)$$

oder

$$\partial_t U^n = a\hat{\partial}_x U^n + \frac{1}{2}\frac{h}{\lambda}\partial_x\bar{\partial}_x U^n \tag{12.15}$$

geschrieben werden. So kann man diese Gleichung als Approximation einer parabolischen Gleichung mit dem kleinen Diffusionskoeffizienten $\frac{1}{2}h/\lambda$ betrachten. Die Stabilität dieses Verfahrens kann dann als Resultat der Einführung der *künstlichen Diffusion* in das ursprüngliche Verfahren angesehen werden. (Diese wird in der numerischen Strömungsdynamik auch als künstliche Viskosität bezeichnet.)

Wir stellen fest, dass $U^n(x)$ im Falle des Friedrichs-Verfahrens als Funktion der Anfangsdaten in der Form

$$U^n(x) = \sum_{j=-n}^{n} a_{nj} v(x - jh)$$

ausgedrückt werden kann und somit Werte von $v(x)$ im Intervall $[x - nh, x + nh] = [x - t_n/\lambda, x + t_n/\lambda]$ benutzt werden. Die exakte Lösung an der Stelle $t = t_n$ ist durch (12.2) als der Wert von v an der Stelle $x + t_n a$ gegeben. Wenn der Abhängigkeitsbereich des Differenzenverfahrens, d. h. das Intervall $[x - t_n/\lambda, x + t_n/\lambda]$, den Abhängigkeitsbereich der exakten Lösung, nämlich den Punkt $x + t_n a$, nicht enthält, dann ist klar, dass das Differenzenverfahren möglicherweise nicht erfolgreich ist. Die Stabilitätsbedingung reduziert sich auf $-1 \leq a\lambda \leq 1$, was unserem alten Stabilitätskriterium entspricht.

Für ein allgemeines Verfahren der Form (12.10) können wir also die *Courant-Friedrichs-Lewy-Bedingung* (oder die *CFL-Bedingung*) für die Stabilität formulieren: Für die Stabilität des Verfahrens ist es notwendig, dass das Abhängigkeitsgebiet des finiten Differenzenverfahrens an der Stelle (x, t) das Abhängigkeitsgebiet des kontinuierlichen Problems enthält.

In unserem ersten Beispiel (12.4) finden wir, dass das finite Differenzenverfahren das Abhängigkeitsintervall $[x, x + t_n/\lambda]$ besitzt, und dass die CFL-Bedingung somit $0 \leq a\lambda \leq 1$ fordert. Insbesondere finden wir unsere alte Stabilitätsbedingung $a\lambda \leq 1$ wieder und stellen deshalb fest, dass das Verfahren (12.4) für $a < 0$ nicht verwendet werden kann. Ebenso stellen wir fest, dass der Vorwärts-Differenzenquotient in (12.4) für $a > 0$ nicht erfolgreich durch einen Rückwärts-Differenzenquotienten ersetzt werden kann. Wie wir gelernt haben, ist das Verfahren (12.4), bei dem ∂_x durch $\bar{\partial}_x$ ersetzt wurde, im Falle $a < 0$ für $a\lambda \geq -1$ stabil.

Die Tatsache, dass die CFL-Bedingung für die Stabilität nicht hinreichend ist, verdeutlicht das Verfahren (12.14), das dasselbe Abhängigkeitsgebiet wie das Friedrichs-Verfahren besitzt, für alle λ jedoch instabil ist.

Wie unser erstes Beispiel (12.4) ist das Friedrichs-Verfahren ebenfalls von der Genauigkeit erster Ordnung: Wenn die exakte Lösung u von (12.1) hinreichend regulär ist, dann gilt aufgrund der Darstellung (12.15)

$$\begin{aligned}\partial_t u^n - a\hat{\partial}_x u^n - \frac{1}{2}\frac{h}{\lambda}\partial_x\bar{\partial}_x u^n &= u_t^n + \frac{1}{2}k u_{tt}^n - a u_x^n - \frac{1}{2}\frac{h}{\lambda}u_{xx}^n + O(h^2)\\ &= \frac{1}{2}\frac{h}{\lambda}(\lambda^2 u_{tt}^n - u_{xx}^n) + O(h^2)\\ &= \frac{1}{2}\frac{h}{\lambda}(a^2\lambda^2 - 1)u_{xx}^n + O(h^2) \quad \text{für } h \to 0.\end{aligned}$$

Folglich ist der Fehler von erster Ordnung, abgesehen von der speziellen Wahl $\lambda = 1/|a|$, bei der die approximative Lösung gleich der exakten Lösung ist (siehe (12.13)).

Als nächstes wollen wir ein Verfahren mit der Genauigkeit zweiter Ordnung bestimmen, das die Form

$$U^{n+1}(x) = (E_k U^n)(x) = a_1 U^n(x-h) + a_0 U^n(x) + a_{-1}U^n(x+h)$$

besitzt. Aus der Gleichung (12.13) geht dafür die Bedingung

$$a_1 \mathrm{e}^{-\mathrm{i}\xi} + a_0 + a_{-1}\mathrm{e}^{\mathrm{i}\xi} = \mathrm{e}^{\mathrm{i}a\lambda\xi} + O(\xi^3) \quad \text{für } \xi \to 0$$

hervor. Durch Taylor-Entwicklung ergibt sich

$$\begin{aligned}(a_1 + a_0 + a_{-1}) - \mathrm{i}(a_1 - a_{-1})\xi - \tfrac{1}{2}(a_1 + a_{-1})\xi^2 &\\ = 1 + \mathrm{i}a\lambda\xi - \tfrac{1}{2}a^2\lambda^2\xi^2 + O(\xi^3) \quad &\text{für } \xi \to 0,\end{aligned}$$

es gilt also

$$\begin{aligned}a_1 + a_0 + a_{-1} &= 1,\\ a_1 - a_{-1} &= -a\lambda,\\ a_1 + a_{-1} &= a^2\lambda^2.\end{aligned}$$

Dies führt auf

$$a_{-1} = \tfrac{1}{2}(a^2\lambda^2 + a\lambda), \quad a_0 = 1 - a^2\lambda^2, \quad a_1 = \tfrac{1}{2}(a^2\lambda^2 - a\lambda),$$

und somit gilt

$$\begin{aligned}(E_k U^n)(x) = \tfrac{1}{2}(a^2\lambda^2 + a\lambda)U^n(x+h) + (1 - a^2\lambda^2)U^n(x)&\\ + \tfrac{1}{2}(a^2\lambda^2 - a\lambda)U^n(x-h),&\end{aligned}$$

woraus sich

$$\tilde{E}(\xi) = 1 - a^2\lambda^2 + a^2\lambda^2\cos\xi + \mathrm{i}a\lambda\sin\xi$$

ergibt. Durch eine einfache Rechnung bestimmen wir

$$|\tilde{E}(\xi)|^2 = 1 - a^2\lambda^2(1 - a^2\lambda^2)(1 - \cos\xi)^2, \tag{12.16}$$

und folglich ist das Verfahren in L_2 genau dann stabil, wenn $a^2\lambda^2 \leq 1$ gilt (siehe Problemstellung 12.2). Wiederum stimmt dies mit der notwendigen CFL-Bedingung für die Stabilität überein.

Das zuletzt behandelte Verfahren wird als *Lax-Wendroff-Verfahren* bezeichnet. Wir weisen darauf hin, dass dies ein Beispiel für ein Verfahren ist, das in der Maximumnorm nicht stabil ist, obwohl es L_2-stabil ist. Man kann dann tatsächlich zeigen, dass im Falle $0 < a^2\lambda^2 < 1$

$$\|E_k^n v\|_{\mathcal{C}} \leq C n^{1/12} \|v\|_{\mathcal{C}}$$

gilt und dass diese Abschätzung hinsichtlich der Potenz von n scharf ist. Die Potenz ist jedoch klein und der Effekt der Instabilität im Allgemeinen nicht spürbar.

12.2 Symmetrische hyperbolische Systeme

Ein großer Teil des in Abschnitt 12.1 behandelten Stoffes lässt sich auf Systeme in einer räumlichen Dimension

$$\frac{\partial u}{\partial t} = A\frac{\partial u}{\partial x} \quad \text{in } \mathbf{R} \times \mathbf{R}_+$$

verallgemeinern, wobei $u = (u_1, \ldots, u_N)^{\mathrm{T}}$ ein Vektor mit N Komponenten und A eine symmetrische $N \times N$-Matrix ist.

Beispielsweise nimmt das Friedrichs-Verfahren nun die Form

$$U^{n+1}(x) = (E_k U^n)(x) = \tfrac{1}{2}(I + \lambda A)U^n(x+h) + \tfrac{1}{2}(I - \lambda A)U^n(x-h) \tag{12.17}$$

an und das Lax-Wendroff-Verfahren ist

$$\begin{aligned} U^{n+1}(x) = (E_k U^n)(x) &= \tfrac{1}{2}(\lambda^2 A^2 + \lambda A)U^n(x+h) \\ &\quad + (I - \lambda^2 A^2)U^n(x) + \tfrac{1}{2}(\lambda^2 A^2 - \lambda A)U^n(x-h). \end{aligned} \tag{12.18}$$

Die charakteristischen Polynome dieser Operatoren werden durch die matrixwertigen periodischen Funktionen

$$\tilde{E}(\xi) = I\cos\xi + \mathrm{i}\lambda A \sin\xi$$

beziehungsweise

$$\tilde{E}(\xi) = I - \lambda^2 A^2 + \lambda^2 A^2 \cos\xi + \mathrm{i}\lambda A \sin\xi$$

bestimmt. Diese können durch dieselbe orthogonale Transformation wie A diagonalisiert werden und man stellt leicht fest, dass die Stabilitätsforderung in beiden Fällen

$$|A|\lambda \le 1$$

lautet, wobei $|A|$ die durch die Euklidischen Norm auf $\mathbf{R}^N$ induzierte Matrixnorm ist, d. h. es gilt mit den Eigenwerten μ_j von A

$$|A| = \sup_{v\neq 0} \frac{|Av|}{|v|} = \max_{j=1,\dots,N} |\mu_j|.$$

Als Beispiel für ein solches System betrachten wir das Anfangswertproblem

$$\begin{aligned} &\frac{\partial^2 w}{\partial t^2} = a^2 \frac{\partial^2 w}{\partial x^2} \quad \text{in } \mathbf{R}\times\mathbf{R}_+, \\ &w(\cdot,0) = w_0, \quad \frac{\partial w}{\partial t}(\cdot,0) = w_1 \quad \text{in } \mathbf{R}. \end{aligned} \tag{12.19}$$

Die hyperbolische Gleichung zweiter Ordnung kann durch die Wahl

$$u_1 = a\,\frac{\partial w}{\partial x},\ u_2 = \frac{\partial w}{\partial t}$$

auf ein System erster Ordnung transformiert werden. Diese Funktionen erfüllen

$$\frac{\partial u_1}{\partial t} = a\,\frac{\partial u_2}{\partial x}, \quad \frac{\partial u_2}{\partial t} = a\,\frac{\partial u_1}{\partial x},$$

d. h. es gilt für $u = (u_1,\, u_2)^{\mathrm{T}}$

$$\begin{aligned} &\frac{\partial u}{\partial t} = \begin{bmatrix} 0 & a \\ a & 0 \end{bmatrix} \frac{\partial u}{\partial x} \quad \text{in } \mathbf{R}\times\mathbf{R}_+, \\ &u(\cdot,0) = \begin{bmatrix} aw_0' \\ w_1 \end{bmatrix} \quad \text{in } \mathbf{R}. \end{aligned} \tag{12.20}$$

Umgekehrt lässt sich dann die Lösung von (12.20) aus der Lösung von (12.19) bestimmen.

Eines der beiden Verfahren (12.17) und (12.18) kann nun auf das vorliegende System angewendet werden. Weil die Matrix in (12.20) die Eigenwerte $\pm a$ besitzt, erhalten wir wieder unser gewöhnliches Stabilitätskriterium $|a|\lambda \le 1$.

Wir wenden uns kurz dem Fall von mehr als einer räumlichen Dimension zu und betrachten das symmetrische hyperbolische System (oder Friedrichs-System)

$$\begin{aligned} &\frac{\partial u}{\partial t} = \sum_{j=1}^{d} A_j \frac{\partial u}{\partial x_j} \qquad \text{in } \mathbf{R}\times\mathbf{R}_+, \\ &u(\cdot,0) = v \qquad\qquad\qquad \text{in } \mathbf{R}^d, \end{aligned} \tag{12.21}$$

mit einer N-vektorwertige Funktion u und den symmetrischen $N\times N$-Matrizen A_j. Wir wissen aus Abschnitt 11.4, dass das zugehörige Anfangswertproblem in L_2 korrekt gestellt ist.

Wir betrachten nun einen zugehörigen finiten Differenzenoperator

$$U^{n+1}(x) = (E_k U^n)(x) = \sum_\beta a_\beta U^n(x - \beta h), \tag{12.22}$$

wobei $\beta = (\beta_1, \ldots, \beta_d)$ ganzzahlige Komponenten besitzt und die a_β endlich viele konstante $N \times N$-Matrizen sind. Mit $\xi = (\xi_1, \ldots, \xi_d) \in \mathbf{R}^d$ ist das charakteristische Polynom durch die Matrix

$$\tilde{E}(\xi) = \sum_\beta a_\beta \mathrm{e}^{-\mathrm{i}\beta\cdot\xi} \quad \text{mit } \beta \cdot \xi = \beta_1\xi_1 + \cdots + \beta_d\xi_d$$

gegeben.

Nach Fourier-Transformation erhalten wir nun

$$(U^{n+1})\hat{}(\xi) = \tilde{E}(h\xi)(U^n)\hat{}(\xi) \quad \text{für } n \geq 0,$$

und folglich gilt

$$(U^n)\hat{}(\xi) = \tilde{E}(h\xi)^n \hat{v}(\xi).$$

Es folgt leicht, dass eine notwendige und hinreichende Bedingung für die Stabilität in L_2 darin besteht, dass die Matrixnorm von $\tilde{E}(\xi)$ die Ungleichung

$$|\tilde{E}(\xi)^n| \leq C \quad \text{für } n \geq 0,\ \xi \in \mathbf{R}^d \tag{12.23}$$

erfüllt. Im Gegensatz zum skalaren Fall folgt für Matrizen aus $|A^n| \leq C$ für $n \geq 0$ nicht die Ungleichung $|A| \leq 1$, wie das Beispiel

$$\begin{bmatrix} a & 1 \\ 0 & a \end{bmatrix}^n = \begin{bmatrix} a^n & na^{n-1} \\ 0 & a^n \end{bmatrix} \tag{12.24}$$

mit $|a| < 1$ zeigt. Aus (12.23) folgt jedoch, dass für jeden Eigenwert $\lambda_j(\xi)$ von $\tilde{E}(\xi)$ die Ungleichung $|\lambda_j(\xi)^n| \leq C$ für $\xi \in \mathbf{R}^d$, $n \geq 0$ gilt und daher

$$|\lambda_j(\xi)| \leq 1 \quad \text{für } \xi \in \mathbf{R}^d$$

erfüllt ist. Dies ist eine *notwendige* Bedingung für die Stabilität in L_2, die als *von Neumannsche Stabilitätsbedingung* bezeichnet wird. Es ist keine hinreichende Bedingung, wie das Beispiel (12.24) mit $a = 1$ illustriert.

Eine hinreichende Bedingung für die Stabilität in L_2 ist aber offensichtlich

$$|\tilde{E}(\xi)| \leq 1 \quad \text{für } \xi \in \mathbf{R}^d.$$

Um ein stabiles Differenzenverfahren der obigen Form konstruieren zu können, benötigen wir folgendes Resultat.

Lemma 12.1. *Angenommen, a_β sind symmetrische, positiv semidefinite Matrizen mit $\sum_\beta a_\beta = I$. Dann gilt*

$$|\tilde{E}(\xi)| = \Big|\sum_\beta a_\beta \mathrm{e}^{-\mathrm{i}\beta\xi}\Big| \leq 1 \quad \textit{für } \xi \in \mathbf{R}^d.$$

Beweis. Es sei

$$\langle u, v\rangle = \sum_{j=1}^{N} u_j \overline{v_j} \quad \text{für } u, v \in \mathbf{C}^N.$$

Weil a_β reell, symmetrisch und positiv semidefinit ist, wissen wir, dass die zugehörige Bilinearform $\langle a_\beta u, v\rangle$ die Gleichung

$$\langle a_\beta u, v\rangle = \overline{\langle a_\beta v, u\rangle}, \quad \langle a_\beta u, u\rangle \geq 0 \quad \text{für } u, v \in \mathbf{C}^N$$

erfüllt. Unter Verwendung dieser Eigenschaften ist es leicht, die verallgemeinerte Cauchy-Schwarz-Ungleichung

$$|\langle a_\beta u, v\rangle| \leq \langle a_\beta u, u\rangle^{1/2} \langle a_\beta v, v\rangle^{1/2}$$

zu beweisen. (Der Beweis der gewöhnlichen Cauchy-Schwarz-Ungleichung ist übertragbar; diese Ungleichung stellt eine Verallgemeinerung dar, weil $\langle a_\beta u, u\rangle^{1/2}$ nur eine Halbnorm ist.) Benutzen wir also auch die Ungleichung $2ab \leq a^2 + b^2$, so erhalten wir

$$|\langle a_\beta u, v\rangle| \leq \tfrac{1}{2}\langle a_\beta u, u\rangle + \tfrac{1}{2}\langle a_\beta v, v\rangle.$$

Deshalb gilt

$$\begin{aligned} |\langle \tilde{E}(\xi)v, w\rangle| &\leq \sum_\beta |\langle a_\beta e^{-i\beta\cdot\xi} v, w\rangle| \\ &\leq \tfrac{1}{2}\sum_\beta \langle a_\beta v, v\rangle + \tfrac{1}{2}\sum_\beta \langle a_\beta w, w\rangle = \tfrac{1}{2}|v|^2 + \tfrac{1}{2}|w|^2. \end{aligned}$$

Nehmen wir $w = \tilde{E}(\xi)v$ an, schlussfolgern wir

$$|w|^2 \leq \tfrac{1}{2}|v|^2 + \tfrac{1}{2}|w|^2,$$

was den Beweis vervollständigt. □

Als Anwendung betrachten wir das Friedrichs-Verfahren (für das wir bereits Spezialfälle kennengelernt haben)

$$\begin{aligned} U^{n+1}(x) &= (E_k U^n)(x) \\ &= \frac{1}{2}\sum_{j=1}^{d} \Big\{ \Big(\frac{1}{d} I + \lambda A_j\Big) U^n(x + he_j) + \Big(\frac{1}{d} I - \lambda A_j\Big) U^n(x - he_j) \Big\}, \end{aligned} \tag{12.25}$$

wobei e_j der Einheitsvektor in der Richtung x_j ist. Dies kann, analog zu (12.15), auch in der Form

$$\partial_t U^n = \sum_{j=1}^{d} A_j \hat{\partial}_{x_j} U^n + \frac{1}{2}\frac{h}{\lambda d}\sum_{j=1}^{d} \partial_{x_j}\bar{\partial}_{x_j} U^n$$

geschrieben werden, und ist folglich insbesondere mit (12.21) konsistent.

Wenn nun für λ

$$0 < \lambda \leq \min_{1 \leq j \leq d} (d|A_j|)^{-1} \tag{12.26}$$

gilt, dann sind die Koeffizienten in (12.25) positiv semidefinit und aus dem Lemma folgt die Stabilität in L_2.

Für das System

$$\frac{\partial u}{\partial t} = \begin{bmatrix} 1 & 0 \\ 0 & -1 \end{bmatrix} \frac{\partial u}{\partial x_1} + \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix} \frac{\partial u}{\partial x_2}$$

reduziert sich die Bedingung (12.26) auf $0 < \lambda \leq 1/2$. In diesem Spezialfall ist

$$\tilde{E}(\xi) = \tfrac{1}{2} I(\cos \xi_1 + \cos \xi_2) + \lambda \mathrm{i} \begin{bmatrix} \sin \xi_1 & \sin \xi_2 \\ \sin \xi_2 & -\sin \xi_1 \end{bmatrix}$$

eine Normalmatrix, d. h. eine Matrix, die mit ihrer Adjungierten kommutiert. Für eine solche Matrix ist die Norm gleich dem Maximum modulo ihrer Eigenwerte. Unter Verwendung dieser Tatsache kann man $|\tilde{E}(\xi)| \leq 1$ für $\lambda \leq 1/\sqrt{2}$ beweisen, was in diesem Fall eine weniger restriktive Bedingung darstellt als die oben auf Grundlage des Lemmas aufgestellte Bedingung.

Das Friedrichs-Verfahren ist wiederum nur von der Genauigkeit erster Ordnung. Tatsächlich kann man zeigen, dass Verfahren von der Form (12.22) mit positiv semidefinitem a_β allgemein in der Genauigkeit nur von erster Ordnung sein können.

12.3 Das Wendroff-Box-Schema

In diesem letzten Abschnitt beschreiben wir das Wendroff-Box-Schema als ein Verfahren zweiter Ordnung, das für gemischte Anfangs-Randwertprobleme und auch für Systeme im Falle einer räumlichen Dimension geeignet ist.

Wir betrachten also das Anfangs-Randwertproblem

$$\begin{aligned} &\frac{\partial u}{\partial t} + a \frac{\partial u}{\partial x} + bu = f && \text{in } \Omega \times J \quad \text{mit } \Omega = (0,1),\ J = (0,T), \\ &u(0,\cdot) = g && \text{auf } J, \\ &u(\cdot,0) = v, && \text{in } \Omega, \end{aligned}$$

wobei a, b, f, g und v glatte Funktionen mit positivem a sind und aus Gründen der Kompatibilität an der Stelle $(x,t) = (0,0)$ die Beziehung $g(0) = v(0)$ gefordert wird. Beachten Sie, dass wegen $a > 0$ die Randwerte am linken Rand vorgegeben wurden.

Mit den Werten U_j^n der Gitterfunktion an der Stelle $(x_j, t_n) = (jh, nk)$, $0 \leq j \leq M,\ 0 \leq n \leq N$ mit $Mh = 1$, $Nk = T$ definieren wir außerdem

$$U_{j+1/2} = \tfrac{1}{2}(U_j + U_{j+1}) \quad \text{und} \quad U^{n+1/2} = \tfrac{1}{2}(U^n + U^{n+1})$$

sowie

$$U_{j+1/2}^{n+1/2} = \tfrac{1}{4}(U_j^n + U_j^{n+1} + U_{j+1}^n + U_{j+1}^{n+1}).$$

Das *Wendroff-Box-Schema* lautet dann

$$\begin{aligned} &\partial_t U_{j+1/2}^n + a\,\partial_x U_j^{n+1/2} + bU_{j+1/2}^{n+1/2} = f, && 0 \le j < M,\ 0 \le n < N, \\ &U_0^n = G^n = g(t_n), && 0 \le n \le N, \\ &U_j^0 = V_j = v(x_j), && 0 \le j \le M, \end{aligned} \tag{12.27}$$

wobei a, b und f an der Stelle $(x_{j+1/2}, t_{n+1/2})$ berechnet werden. Für das Verfahren lässt sich ebenfalls die Differenzengleichung

$$\begin{aligned} \frac{U_j^{n+1} + U_{j+1}^{n+1} - U_j^n - U_{j+1}^n}{2k} + a\,\frac{U_{j+1}^{n+1} + U_{j+1}^n - U_j^{n+1} - U_j^n}{2h} \\ + b\,\frac{U_j^{n+1} + U_{j+1}^{n+1} + U_j^n + U_{j+1}^n}{4} = f \end{aligned} \tag{12.28}$$

aufschreiben; und wir erkennen aufgrund der Symmetrie, dass es sich um ein Verfahren zweiter Ordnung handelt. Diese Gleichung lässt sich in der Form

$$\begin{aligned} (1 + a\lambda + \tfrac{1}{2}bk)U_{j+1}^{n+1} = (1 + a\lambda - \tfrac{1}{2}bk)U_j^n + (1 - a\lambda - \tfrac{1}{2}bk)U_{j+1}^n \\ - (1 - a\lambda + \tfrac{1}{2}bk)U_j^{n+1} + 2kf \quad \text{mit } \lambda = k/h \end{aligned} \tag{12.29}$$

ausdrücken. Dadurch wird die Lösung an der Stelle (x_{j+1}, t_{n+1}) als Funktion der Werte an den Stellen (x_j, t_n), (x_{j+1}, t_n) und (x_j, t_{n+1}) definiert. Ist U^n gegeben, kann man U^{n+1} deshalb explizit in der Reihenfolge $U_0^{n+1} = G^{n+1}$, U_1^{n+1}, $U_2^{n+1}, \ldots, U_M^{n+1}$ bestimmen.

Wir werden die Stabilität dieser Methode in der diskreten L_2-Norm

$$\|V\|_h = \Big(h \sum_{j=1}^{M} V_j^2\Big)^{1/2}$$

zeigen und uns der Einfachheit halber auf den Fall beschränkten, dass a konstant ist und $b = f = g = 0$ gilt. In diesem Fall reduziert sich (12.28) auf

$$U_{j+1}^{n+1} = U_j^n + \frac{1 - a\lambda}{1 + a\lambda}(U_{j+1}^n - U_j^{n+1}),$$

und wir müssen

$$\|U^n\|_h \le C\|V\|_h \quad \text{für } 0 \le t_n \le T \tag{12.30}$$

zeigen. Aus diesem Grund multiplizieren wir (12.27) mit $U_{j+1/2}^{n+1/2}$ und stellen fest, dass

$$\partial_t U^n_{j+1/2}\; U^{n+1/2}_{j+1/2} = \tfrac{1}{2}\partial_t (U^n_{j+1/2})^2$$

und

$$\partial_x U^{n+1/2}_{j}\; U^{n+1/2}_{j+1/2} = \tfrac{1}{2}\partial_x (U^{n+1/2}_{j})^2$$

gilt, sodass wir

$$\partial_t (U^m_{j+1/2})^2 + a\partial_x (U^{m+1/2}_j)^2 = 0$$

erhalten. Nach Summation über $j = 0, \dots, M-1$, $m = 0, \dots, n-1$ und Multiplikation mit hk führt dies für $n \le N$ auf

$$h \sum_{j=0}^{M-1} (U^n_{j+1/2})^2 + ak \sum_{m=0}^{n-1} (U^{m+1/2}_M)^2 = h \sum_{j=0}^{M-1} (U^0_{j+1/2})^2 + ak \sum_{m=0}^{n-1} (U^{m+1/2}_0)^2,$$

woraus wegen unserer Annahme $U^{m+1/2}_0 = 0$

$$h \sum_{j=0}^{M-1} (U^n_j + U^n_{j+1})^2 \le C\|V\|^2_h \tag{12.31}$$

folgt.

Multiplizieren wir (12.27) analog dazu stattdessen mit

$$hk\partial_x \partial_t U^n_j = U^{n+1}_{j+1} - U^{n+1}_j - U^n_{j+1} + U^n_j,$$

erhalten wir nach einer einfachen Rechnung

$$h \sum_{j=0}^{M-1} (U^n_{j+1} - U^n_j)^2 \le C\|V\|^2_h. \tag{12.32}$$

Insgesamt vervollständigen (12.31) und (12.32) den Beweis von (12.30).

Aus der Stabilität und der Genauigkeit zweiter Ordnung folgt die Konvergenz zweiter Ordnung, vorausgesetzt u ist hinreichend glatt, d. h. es gilt

$$\|U^n - u^n\|_h \le C(u)h^2 \quad \text{für } k/h = \lambda = \text{konstant}.$$

12.4 Problemstellungen

Problem 12.1. Beweisen Sie, dass das Friedrichs-Verfahren für (12.1) genau dann in der Maximumnorm stabil ist, wenn $|a|\lambda \le 1$ gilt.

Problem 12.2. Beweisen Sie (12.16) und dass das Lax-Wendroff-Verfahren genau dann in L_2 stabil ist, wenn $|a|\lambda \le 1$ gilt.

Problem 12.3. Sei E_k ein durch

$$(E_k V)_j = \sum_p a_p V_{j-p}$$

definierter Finite-Differenzen-Operator.
(s) Zeigen Sie, dass

$$\|E_k V\|_{\infty,h} \le C\|v\|_{\infty,h} \quad \text{mit } C = \sum_p |a_p|$$

gilt und dass diese Ungleichung für jede kleinere Konstante nicht erfüllt ist.
(b) Zeigen Sie, dass die Gleichung $(E_k^n V)_j = \sum_p a_{np} V_{j-p}$ erfüllt ist, wobei

$$a_{np} = \frac{1}{2\pi} \int_{-\pi}^{\pi} \tilde{E}(\xi)^n \mathrm{e}^{\mathrm{i}p\xi} \, \mathrm{d}\xi$$

mit der charakteristischen Gleichung $\tilde{E}(\xi) = \sum_j a_j \mathrm{e}^{-\mathrm{i}j\xi}$ von E_k gilt.
(c) Zeigen Sie, dass E_k genau dann in der Maximumnorm stabil ist, wenn

$$\sum_p |a_{np}| \le C \quad \forall n \ge 0$$

gilt.

Problem 12.4. Beweisen Sie, dass das Lax-Wendroff-Verfahren für $|a|\lambda > 1$ in der Maximumnorm instabil ist. Hinweis: Verwenden Sie Problemstellung 12.3.

Problem 12.5. Sie wissen aus Abschnitt 7.1, dass man für eine $m \times m$-Matrix M den Ausdruck $\exp(M) = \sum_{j=0}^{\infty} \frac{1}{j!} M^j$ definieren kann. Betrachten Sie das symmetrische hyperbolische System $\partial u/\partial t = A \partial u/\partial x$.
(a) Zeigen Sie, dass ein finites Differenzenverfahren für dieses System von der Form (vgl. (12.17) und (12.18))

$$(E_k V)_j = \sum_p a_p(\lambda A) V_{j-p}$$

von der Genauigkeit der Ordnung r mit $r = 1, 2$ ist, wenn

$$\tilde{E}(\xi) = \exp(\mathrm{i}\lambda A \xi) + O(\xi^{r+1}) \quad \text{für } \xi \to 0$$

gilt.
(b) Überprüfen Sie diese Bedingung für das Friedrichs- und das Lax-Wendroff-Verfahren (12.17) und (12.18).

Problem 12.6. Diskutieren Sie die Bedeutung der CFL-Bedingung für das Anfangs-Randwertproblem in Abschnitt 12.3. Zeigen Sie, dass diese für das in (12.27) definierte Wendroff-Box-Schema erfüllt ist.

Problem 12.7. (Übung am Rechner.) Wenden Sie das Wendroff-Box-Schema auf das Problem aus Beispiel 11.8 mit $h = k = 1/10$ und $h = k = 1/20$ an. Berechnen Sie die Fehler an der Stelle $(1, 1/2)$.

13

Die Methode der finiten Elemente für hyperbolische Gleichungen

In diesem Kapitel wenden wir die Methode der finiten Elemente auf hyperbolische Gleichungen an. In Abschnitt 13.1 untersuchen wir ein Anfangs-Randwertproblem für die Wellengleichung und diskutieren semidiskrete und vollständig diskrete Verfahren auf Grundlage der gewöhnlichen Finite-Elemente-Diskretisierung in den räumlichen Variablen. In Abschnitt 13.2 betrachten wir eine skalare partielle Differentialgleichung erster Ordnung in zwei unabhängigen Variablen. Wir beginnen mit der Behandlung der Gleichung als Evolutionsgleichung und zeigen eine Fehlerabschätzung der nichtoptimalen Ordnung $O(h)$ für die gewöhnliche Galerkin-Methode. Anschließend betrachten wir das zugehörige Randwertproblem als ein zweidimensionales Problem von dem in Abschnitt 11.3 behandelten Typ. Wir führen die Modifikation der Stromliniendiffusion ein und beweisen eine Konvergenzabschätzung der Ordnung $O(h^{3/2})$. Wir kommen schließlich auf den Aspekt der Evolution zurück und kombinieren die Stromliniendiffusion mit der sogenannten diskontinuierlichen Galerkin-Methode, um ein Zeitschrittverfahren zu entwerfen. Dabei werden zweidimensionale Approximationsfunktionen benutzt, die in den Zeitebenen unstetig sein können.

13.1 Die Wellengleichung

In diesem Abschnitt diskutieren wir kurz einige Resultate, die sich auf semidiskrete und vollständig diskrete Verfahren für das folgende Anfangs-Randwertproblem für die Wellengleichung

$$
\begin{aligned}
& u_{tt} - \Delta u = f && \text{in } \Omega \times \mathbf{R}_+, \\
& u = 0 && \text{auf } \Gamma \times \mathbf{R}_+, \\
& u(\cdot, 0) = v,\ u_t(\cdot, 0) = w && \text{in } \Omega
\end{aligned}
\tag{13.1}
$$

beziehen. Wie schon oft nehmen wir an, dass es sich bei $\Omega \subset \mathbf{R}^2$ um ein abgeschlossenes konvexes Gebiet mit polygonalem Rand Γ handelt. Mit $S_h \subset H_0^1$

bezeichnen wir eine Familie von Räumen stückweise linearer Finite-Elemente-Funktionen in den räumlichen Variablen.

Das semidiskrete Analogon von (13.1) besteht folglich darin, ein $u_h(t) \in S_h$ so zu bestimmen, dass mit der bisherigen Notation, insbesondere mit $a(v,w) = (\nabla v, \nabla w)$,

$$\begin{aligned} &(u_{h,tt}, \chi) + a(u_h, \chi) = (f, \chi) \quad \forall \chi \in S_h \quad \text{für } t > 0, \\ &u_h(0) = v_h, \ u_{h,t}(0) = w_h \end{aligned} \tag{13.2}$$

gilt. Dies ist ein Anfangswertproblem für ein System gewöhnlicher Differentialgleichungen zweiter Ordnung für die Koeffizienten von u_h bezüglich der Standardbasis $\{\Phi_j\}_{j=1}^{M_h}$ von S_h. Wenn

$$u_h(x,t) = \sum_{j=1}^{M_h} \alpha_j(t) \Phi_j(x)$$

gilt, dann ist (13.2) äquivalent zu

$$B\alpha''(t) + A\alpha(t) = b(t) \quad \text{für } t > 0,$$

wobei die Elemente von B, A und b jeweils $b_{kj} = (\Phi_j, \Phi_k)$, $a_{kj} = a(\Phi_j, \Phi_k)$ und $b_k = (f, \Phi_k)$ sind. Die Anfangsbedingungen lauten

$$\alpha(0) = \beta, \quad \alpha'(0) = \gamma$$

mit

$$v_h = \sum_{j=1}^{M_h} \beta_j \Phi_j, \quad w_h = \sum_{j=1}^{M_h} \gamma_j \Phi_j.$$

Wir beginnen mit dem Beweis einer diskreten Variante des Resultates für die Energieerhaltung aus Theorem 11.2.

Lemma 13.1. *Es sei u_h die Lösung von* (13.2) *mit $f = 0$. Dann gilt*

$$\|u_{h,t}(t)\|^2 + |u_h(t)|_1^2 = \|w_h\|^2 + |v_h|_1^2 \quad \textit{für } t \geq 0. \tag{13.3}$$

Beweis. Wählen wir $\chi = u_{h,t}$ in (13.2), so gilt

$$\tfrac{1}{2} \frac{d}{dt} \big(\|u_{h,t}\|^2 + |u_h|_1^2 \big) = 0,$$

woraus sich das Resultat sofort ergibt. □

Wir beweisen nun die folgende Fehlerabschätzung, wobei R_h die in (5.49) definierte elliptische Projektion ist.

Theorem 13.1. *Seien u_h und u die Lösungen von* (13.2) *und* (13.1). *Dann gilt mit nichtfallendem $C(t)$ für $t \geq 0$*

$$\begin{aligned}\|u_h(t)-u(t)\| + h|u_h(t)-u(t)|_1 + \|u_{h,t}(t)-u_t(t)\| \\ \leq C\Big(|v_h - R_h v|_1 + \|w_h - R_h w\|\Big) \\ + C(t)h^2\Big(\|u(t)\|_2 + \|u_t(t)\|_2 + \Big(\int_0^t \|u_{tt}\|_2^2 \,\mathrm{d}s\Big)^{1/2}\Big).\end{aligned} \tag{13.4}$$

Beweis. Schreiben wir wie gewöhnlich

$$u_h - u = (u_h - R_h u) + (R_h u - u) = \theta + \rho,$$

so können wir ρ und ρ_t wie im Beweis von Theorem 10.1 durch

$$\|\rho(t)\| + h|\rho(t)|_1 \leq Ch^2\|u(t)\|_2, \quad \|\rho_t(t)\| \leq Ch^2\|u_t(t)\|_2$$

abschätzen. Für $\theta(t)$ ergibt sich nach einer Rechnung, die analog zu der in (10.14) verläuft,

$$(\theta_{tt},\chi) + a(\theta,\chi) = -(\rho_{tt},\chi) \quad \forall \chi \in S_h \quad \text{für } t > 0. \tag{13.5}$$

Um die Effekte der Anfangswerte und des Quellterms voneinander zu trennen, setzen wir $\theta = \widehat{\theta} + \widetilde{\theta}$, wobei

$$\begin{aligned}&(\widehat{\theta}_{tt},\chi) + a(\widehat{\theta},\chi) = 0 \quad \forall \chi \in S_h,\ t > 0,\\ &\widehat{\theta}(0) = \theta(0), \quad \widehat{\theta}_t(0) = \theta_t(0)\end{aligned}$$

gilt. Dann sind $\widehat{\theta}$ und $\widehat{\theta}_t$, wie von Lemma 13.1 gefordert, beschränkt. Der verbleibende Teil $\widetilde{\theta}$ von θ erfüllt (13.5) mit $\widetilde{\theta}(0) = \widetilde{\theta}_t(0) = 0$. Mit $\chi = \widetilde{\theta}_t$ führt dies auf

$$\tfrac{1}{2}\frac{d}{dt}\big(\|\widetilde{\theta}_t\|^2 + |\widetilde{\theta}|_1^2\big) = -(\rho_{tt},\widetilde{\theta}_t) \leq \tfrac{1}{2}\|\rho_{tt}\|^2 + \tfrac{1}{2}\|\widetilde{\theta}_t\|^2$$

und nach Integration über t auf

$$\|\widetilde{\theta}_t(t)\|^2 + |\widetilde{\theta}(t)|_1^2 \leq \int_0^t \|\rho_{tt}\|^2 \,\mathrm{d}s + \int_0^t \|\widetilde{\theta}_t\|^2 \,\mathrm{d}s.$$

Daher gilt aufgrund des Gronwall-Lemmas (siehe Problemstellung 7.6)

$$\|\widetilde{\theta}_t(t)\|^2 + |\widetilde{\theta}(t)|_1^2 \leq C(t)\int_0^t \|\rho_{tt}\|^2 \,\mathrm{d}s \leq C(t)h^4 \int_0^t \|u_{tt}\|_2^2 \,\mathrm{d}s$$

mit $C(t) = e^t$. Dabei haben wir Theorem 5.5 verwendet, um ρ_{tt} abzuschätzen. Dies beschränkt die ersten beiden Terme in (13.4) wie gewünscht. Der dritte Term wird analog abgeschätzt. □

Wir stellen fest, dass die Wahl $v_h = R_h v$ und $w_h = R_h w$ in Theorem 13.1 Fehlerabschätzungen optimaler Ordnung für alle betrachteten Größen ergibt. Allerdings kann es durch eine andere optimale Wahl von v_h aufgrund des Gradienten im ersten Term auf der rechten Seite zu einem Verlust einer Potenz in h kommen.

Wir diskutieren nun kurz auch die Diskretisierung in der Zeit. Dabei sei $U^n \in S_h$ die Approximation zur Zeit $t_n = nk$ mit dem Zeitschritt k. Ein mögliches Verfahren besteht nun darin, U^n für $n \geq 2$ zu bestimmen, indem man für $n \geq 1$ die Gleichungen

$$(\partial_t \bar{\partial}_t U^n, \chi) + a(\tfrac{1}{4}(U^{n+1} + 2U^n + U^{n-1}), \chi) = (f(t_n), \chi) \quad \forall \chi \in S_h \tag{13.6}$$

aufstellt. Dabei sind U^0 und U^1 gegebene Approximationen von $u(0) = v$ beziehungsweise $u(t_1)$. Die Wahl des Mittelwertes im zweiten Term ist durch eine Kombination aus Stabilitäts- und Genauigkeitsbetrachtungen motiviert. In Bezug auf die Stabilität beweisen wir das folgende, vollständig diskrete Analogon des semidiskreten Energieerhaltungssatzes aus Lemma 13.3, wobei wir $U^{n+1/2} = (U^n + U^{n+1})/2$ definieren.

Lemma 13.2. *Für die Lösung von* (13.6) *mit* $f = 0$ *gilt*

$$\|\partial_t U^n\|^2 + |U^{n+1/2}|_1^2 = \|\partial_t U^0\|^2 + |U^{1/2}|_1^2 \quad \textit{für } n \geq 0.$$

Beweis. Wir wenden (13.6) mit

$$\chi = \frac{1}{2k}(U^{n+1} - U^{n-1}) = \frac{1}{2}(\partial_t U^n + \partial_t U^{n-1}) = \frac{1}{k}(U^{n+1/2} - U^{n-1/2})$$

an. Mit diesem χ erhalten wir

$$(\partial_t \bar{\partial}_t U^n, \chi) = \frac{1}{2k}(\partial_t U^n - \partial_t U^{n-1}, \partial_t U^n + \partial_t U^{n-1}) = \frac{1}{2}\bar{\partial}_t \|\partial_t U^n\|^2$$

und

$$\begin{aligned} a(\tfrac{1}{4}U^{n+1} + \tfrac{1}{2}U^n + \tfrac{1}{4}U^{n-1}, \chi) &= \frac{1}{2k} a(U^{n+1/2} + U^{n-1/2}, U^{n+1/2} - U^{n-1/2}) \\ &= \frac{1}{2}\bar{\partial}_t |U^{n+1/2}|_1^2. \end{aligned}$$

Daher gilt

$$\bar{\partial}_t \big(\|\partial_t U^n\|^2 + |U^{n+1/2}|_1^2 \big) = 0,$$

woraus das Resultat folgt. □

Mithilfe dieser Stabilitätsaussage und mit Argumenten, die zu den im Beweis von Theorem 13.1 verwendeten analog sind, kann man folgendes Theorem zeigen. Dabei benutzen wir unsere gewöhnliche Notation $\theta^n = U^n - R_h u(t_n)$. Die Details überlassen wir Problemstellung 13.4.

Theorem 13.2. *Seien U^n und u die Lösungen von* (13.6) *und* (13.1). *Die Anfangswerte U^0 und U^1 seien so gewählt, dass*

$$\|\partial_t \theta^0\| + |\theta^0|_1 + |\theta^1|_1 \le C(h^2 + k^2)$$

erfüllt ist. Dann gilt unter geeigneten Regularitätsbedingungen für u mit nichtfallendem $C(u,t)$ in t

$$\|U^{n+1/2} - u(t_n + \tfrac{1}{2}k)\| + \|\partial_t U^n - u_t(t_n + \tfrac{1}{2}k)\| \le C(u,t_n)(h^2 + k^2)$$

und

$$|U^{n+1/2} - u(t_n + \tfrac{1}{2}k)|_1 \le C(u,t_n)(h + k^2) \quad \textit{für } n \ge 0.$$

Die Bedingungen für die Anfangswerte können durch den Ansatz $U^0 = R_h v$ und $U^1 = R_h(v + kw + \frac{1}{2}k^2 u_{tt}(0))$ erfüllt werden, wobei $u_{tt}(0) = \Delta v + f(0)$ gilt.

Obwohl Theorem 13.2 den Fehler an den Punkten $t_n + \frac{1}{2}k$ abschätzt, ist es klar, dass wir auch zu Approximationen optimaler Ordnung an den Punkten t_n kommen können, da beispielsweise

$$\begin{aligned}\|\tfrac{1}{4}(U^{n+1} + 2U^n + U^{n-1}) - u(t_n)\| &\\ \le \tfrac{1}{2}\|U^{n+1/2} - u(t_n + \tfrac{1}{2}k)\| + \tfrac{1}{2}\|U^{n-1/2} - u(t_n - \tfrac{1}{2}k)\| &\\ + \|\tfrac{1}{2}(u(t_n + \tfrac{1}{2}k) + u(t_n - \tfrac{1}{2}k)) - u(t_n)\| \le C(u,t_n)(h^2 + k^2)&\end{aligned}$$

gilt.

13.2 Hyperbolische Gleichungen erster Ordnung

Wir betrachten zunächst das Anfangs-Randwertproblem (vgl. Beispiel 11.7)

$$\begin{aligned} u_t + u_x &= f && \text{in } \Omega = (0,1) \quad \text{für } t > 0,\\ u(0,t) &= 0 && \text{für } t > 0,\\ u(\cdot,0) &= v && \text{in } \Omega. \end{aligned} \tag{13.7}$$

Mit $0 = x_0 < x_1 < \cdots < x_M = 1$ und $K_j = [x_{j-1}, x_j]$ suchen wir nach einer approximativen Lösung im Raum

$$S_h^- = \big\{\chi \in \mathcal{C}(\bar\Omega) : \chi \text{ linear in } K_j,\ j = 1,\dots,M,\ \chi(0) = 0\big\}. \tag{13.8}$$

Beachten Sie die Forderung, dass die Funktionen in S_h^- an der Stelle $x = 0$, d. h. an dem räumlichen Teil $\Gamma_{-,x}$ des Zuflussrandes, verschwinden soll. An der Stelle $x = 1$, die ein Teil des Abflussrandes ist, soll dies aber nicht gelten.

Bei der räumlich diskreten gewöhnlichen Galerkin-Methode ist dann für $t \ge 0$ eine Funktion $u_h(t) \in S_h^-$ gesucht, die die Gleichung

$$(u_{h,t} + u_{h,x}, \chi) = (f, \chi) \quad \forall \chi \in S_h^-,\ t > 0, \qquad u_h(0) = v_h \approx v \tag{13.9}$$

erfüllt. Als Funktion der Standardbasis $\{\Phi_j\}_{j=1}^M$ der Hutfunktionen kann dies in der Form

$$B\alpha'(t) + A\alpha(t) = f \quad \text{für } t > 0 \quad \text{mit } \alpha(0) = \gamma$$

geschrieben werden, wobei B wie gewöhnlich die symmetrische, positiv definite Matrix mit den Elementen $b_{kj} = (\Phi_j, \Phi_k)$ ist, sodass das Problem insbesondere eine wohldefinierte Lösung für $t \geq 0$ besitzt. Die Matrix A mit den Elementen $a_{kj} = (\Phi_j', \Phi_k) = -(\Phi_k', \Phi_j) = -a_{jk}$ ist jetzt allerdings schiefsymmetrisch.

Wir wollen nun die Stabilität dieses Verfahrens beweisen. Dazu wählen wir in (13.9) $\chi = u_h$. Dies führt auf

$$\tfrac{1}{2}\frac{d}{dt}\|u_h\|^2 + (u_{h,x}, u_h) = (f, u_h) \leq \|f\|\,\|u_h\|.$$

Hier ist

$$(u_{h,x}, u_h) = \tfrac{1}{2}\Big[u_h^2\Big]_0^1 = \tfrac{1}{2}u_h(1)^2 \geq 0,$$

und daher gilt

$$\frac{d}{dt}\|u_h\| \leq \|f\|,$$

sodass sich nach Integration

$$\|u_h(t)\| \leq \|v_h\| + \int_0^t \|f\|\,\mathrm{d}s \quad \text{für } t \geq 0 \tag{13.10}$$

ergibt.

Wir beweisen nun eine Fehlerabschätzung.

Theorem 13.3. *Seien u_h und u die Lösungen von* (13.9) *und* (13.7). *Dann erhalten wir mit geeignet gewähltem v_h*

$$\|u_h(t) - u(t)\| \leq Ch\Big(\|v\|_1 + \int_0^t \big(\|u\|_2 + \|u_t\|_1\big)\,\mathrm{d}s\Big) \quad \textit{für } t \geq 0.$$

Beweis. Wir schreiben

$$u_h - u = (u_h - I_h u) + (I_h u - u) = \theta + \rho$$

mit dem gewöhnlichen Interpolationsoperator I_h in S_h. Wegen Theorem 5.5 gilt

$$\|\rho(t)\| \leq Ch\|u(t)\|_1 \leq Ch\Big(\|v\|_1 + \int_0^t \|u_t\|_1\,\mathrm{d}s\Big),$$

was wie gewünscht beschränkt ist. Gemäß unserer Definitionen gilt $\theta \in S_h^-$ und

$$(\theta_t, \chi) + (\theta_x, \chi) = -(\omega, \chi) \quad \forall \chi \in S_h^- \quad \text{mit } \omega = \rho_t + \rho_x.$$

Aus der Fehlerabschätzung (13.10) und Theorem 5.5 schlussfolgern wir im Falle $v_h = I_h v$, in dem $\theta(0) = 0$ gilt,

$$\|\theta(t)\| \le \int_0^t \Big(\|\rho_t\| + \|\rho_x\| \Big) \, \mathrm{d}s \le Ch \int_0^t \|u_t\|_1 \, \mathrm{d}s + Ch \int_0^t \|u\|_2 \, \mathrm{d}s.$$

Damit ist der Beweis vollständig. □

Wir weisen darauf hin, dass die Fehlerschranke nicht von optimaler Ordnung $O(h^2)$ ist, weil die Schranke für $\theta(t)$ die Ableitung des Interpolationsfehlers enthält.

Diese Analyse des räumlich semidiskreten Problems kann auf vollständig diskrete Verfahren übertragen werden. Wir erläutern dies am Beispiel des Rückwärts-Euler-Verfahrens, d. h. mit unserer Standardnotation,

$$\begin{aligned} (\bar\partial_t U^n \chi) + (U_x^n, \chi) &= (f^n, \chi) \quad \forall \chi \in S_h^-, \ n > 0, \\ U^0 &= v_h. \end{aligned} \tag{13.11}$$

Nun gilt die Stabilitätsabschätzung

$$\|U^n\| \le \|v_h\| + k \sum_{j=1}^{n} \|f^j\| \quad \text{für } n \ge 0 \tag{13.12}$$

(Problemstellung 13.5) und die Fehlerabschätzung ist durch folgendes Theorem gegeben.

Theorem 13.4. *Seien U^n und u die Lösungen von* (13.11) *und* (13.7)*. Dann gilt mit geeignet gewähltem v_h für $n \ge 0$*

$$\|U^n - u(t_n)\| \le Ch \Big(\|v\|_1 + \int_0^{t_n} (\|u\|_2 + \|u_t\|_1) \, \mathrm{d}s \Big) + Ck \int_0^{t_n} \|u_{tt}\| \, \mathrm{d}s.$$

Beweis. Diesmal erfüllt $\theta^n = U^n - I_h u^n$ mit $u^n = u(t_n)$ und $\omega^n = \bar\partial_t \rho^n + \rho_x^n + (u_t^n - \bar\partial_t u^n)$ die Gleichung

$$(\bar\partial_t \theta^n, \chi) + (\theta_x^n, \chi) = -(\omega^n, \chi) \quad \forall \chi \in S_h^-.$$

Der einzige wesentliche neue Term in ω^n ist der Letzte, der durch

$$\|u_t^n - \bar\partial_t u^n\| = \Big\| \int_{t_{n-1}}^{t_n} (s - t_{n-1}) \, u_{tt}(s) \, \mathrm{d}s \Big\| \le k \int_{t_{n-1}}^{t_n} \|u_{tt}\| \, \mathrm{d}s$$

beschränkt ist. Mit der Abschätzung (13.12) ist der Beweis vollständig. □

Um mit der Methode der finiten Elemente für Gleichungen erster Ordnung fortfahren zu können, vernachlässigen wir den Evolutionsaspekt vorübergehend und betrachten das in Abschnitt 11.3 diskutierte zweidimensionale Problem

$$\begin{aligned} a \cdot \nabla u + a_0 u &= f \qquad \text{in } \Omega, \\ u &= g \qquad \text{auf } \Gamma_-. \end{aligned} \tag{13.13}$$

Dabei nehmen wir an, dass das Geschwindigkeitsfeld $a = (a_1, \dots, a_d)$ und der Koeffizient a_0 mit $a_0 > 0$ konstant ist. Wir errinnern daran, dass wir den Zufluss- und Abflussrandrand durch

$$\Gamma_- = \big\{x \in \Gamma : a \cdot n < 0\big\}, \quad \Gamma_+ = \big\{x \in \Gamma : a \cdot n > 0\big\}$$

definiert haben. Wir werden die Abhängigkeit der nachfolgenden Abschätzungen von der Konstanten a_0 verfolgen, dabei aber annehmen, dass sie von oben beschränkt ist.

Nun diskretisieren wir dieses Problem mithilfe einer gewöhnlichen zweidimensionalen Finite-Elemente-Methode. Wie in Abschnitt 5.2 nehmen wir an, dass $\Omega \subset \mathbf{R}^2$ ein abgeschlossenes konvexes Gebiet mit dem polygonalen Rand Γ ist. Dabei ist S_h eine Familie von Räumen stückweise linearer Finite-Elemente-Funktionen bezüglich einer Familie von Triangulationen von Ω, ohne das wir irgendwelche Randbedingungen an die Funktionen in S_h stellen. Somit gilt nun statt $S_h \subset H_0^1$ die Beziehung $S_h \subset H^1$. Wir verwenden den in Abschnitt 5.3 definierten Interpolationsoperator I_h und erinnern an dessen Fehlerabschätzungen

$$\|I_h v - v\| \le Ch^2 \|v\|_2, \quad |I_h v - v|_1 \le Ch\|v\|_2. \tag{13.14}$$

Schließlich nehmen wir an, dass die Triangulation so dem Rand angepasst sind, dass der Zuflussrand genau eine Vereinigung von Dreieckseiten ist, und setzen

$$S_h^- = \big\{\chi \in S_h : \chi = 0 \text{ auf } \Gamma_-\big\}.$$

Wir betonen, dass die Normen in (13.14) über einem zweidimensionalen Gebiet Ω genommen werden.

Bei der gewöhnlichen Galerkin-Finite-Elemente-Methode für das vorliegende Problem ist nun eine Funktion $u_h \in S_h$ gesucht, für die

$$\begin{aligned} (a \cdot \nabla u_h, \chi) + a_0(u_h, \chi) &= (f, \chi) \qquad \forall \chi \in S_h^-, \\ u_h &= g_h = I_h g \qquad \text{auf } \Gamma_- \end{aligned} \tag{13.15}$$

gilt, wobei sich das Skalarprodukt nun auf das zweidimensionale Gebiet Ω bezieht.

Unter Verwendung der Greenschen Formel erhalten wir die Identität

$$(a \cdot \nabla v, v) = \tfrac{1}{2}(a \cdot n\, v, v)_\Gamma = \tfrac{1}{2}|v|_{\Gamma_+}^2 - \tfrac{1}{2}|v|_{\Gamma_-}^2 \tag{13.16}$$

(bedenken Sie, dass a konstant ist), wobei wir die gewichteten Normen

$$|v|^2_{\Gamma_\pm} = \pm(a \cdot n\, v, v)_{\Gamma_\pm} = \int_{\Gamma_\pm} |a \cdot n|\, v^2 \, \mathrm{d}s$$

eingeführt haben.

Wir betrachten nun eine Lösung $w_h \in S_h$ von (13.15), die die homogene Randbedingung $w_h = 0$ auf Γ_- erfüllt. Wegen $w_h \in S_h^-$ können wir dann $\chi = w_h$ wählen, um unter Berücksichtigung von (13.16)

$$\tfrac{1}{2}|w_h|^2_{\Gamma_+} + a_0\|w_h\|^2 = (f, w_h)$$

zu erhalten. Für $f = 0$ beweist dies unmittelbar $w_h = 0$ und daher die Eindeutigkeit der Lösung von (13.15) und deshalb auch deren Existenz. Wir erhalten ebenso leicht die Stabilitätsabschätzung

$$|w_h|^2_{\Gamma_+} + a_0\|w_h\|^2 \le C\|f\|^2 \quad \text{mit } C = 1/a_0. \tag{13.17}$$

Wir setzen unsere Diskussion mit dem Beweis der folgenden einfachen Fehlerabschätzung fort.

Theorem 13.5. *Seien u_h und u die Lösungen von* (13.15) *und* (13.13)*. Dann gilt*

$$\|u_h - u\| \le Ch\|u\|_2.$$

Beweis. Wir schreiben $u_h - u = (u_h - I_h u) + (I_h u - u) = \theta + \rho$. Dann gilt wegen (13.14)

$$\|\rho\| + h\|\rho\|_1 \le Ch^2\|u\|_2. \tag{13.18}$$

Um θ abschätzen zu können, stellen wir fest, dass $\theta \in S_h^-$ gilt und wegen (13.15) und (13.13)

$$(a \cdot \nabla\theta, \chi) + a_0(\theta, \chi) = -(a \cdot \nabla\rho + a_0\rho, \chi) \quad \forall \chi \in S_h^- \tag{13.19}$$

ist. Wegen $\theta \in S_h^-$ zeigt die Stabilitätsabschätzung (13.17) mit (13.18)

$$|\theta|^2_{\Gamma_+} + a_0\|\theta\|^2 \le C(\|\nabla\rho\|^2 + \|\rho\|^2) \le Ch^2\|u\|_2^2,$$

was den Beweis vervollständigt. □

Wir sehen, dass die Fehlerabschätzung in Theorem 13.5 wie in den Theoremen 13.3 und 13.4 von nichtoptimaler Ordnung $O(h)$ ist. Dies folgt aus der Tatsache, dass der Gradient des Interpolationsfehlers auf der rechten Seite von (13.19) vorkommt. Es ist bekannt, dass diese Fehlerschranke nicht verbessert werden kann.Trotzdem arbeitet die Galerkin-Methode akzeptabel, wenn die Lösung glatt ist. Die Lösungen von (13.15) sind jedoch nicht notwendigerweise glatt, und die Erfahrung zeigt, dass die Methode dann weniger gut arbeitet.

Beispielsweise kann sie Oszillationen in der Nähe der Bereiche, in denen sich die Lösung schnell ändert, erzeugen.

Solche Oszillationen lassen sich durch Hinzunahme einer künstlichen Diffusion reduzieren, wie wir es bereits getan haben, um das Friedrichs-Verfahren aus dem instabilen Verfahren (12.15) zu erhalten. Die gewöhnliche Galerkin-Methode mit künstlicher Diffusion besteht nun darin, ein $u_h \in S_h$ so zu bestimmen, dass

$$\begin{aligned} &(a \cdot \nabla u_h, \chi) + a_0(u_h, \chi) + h(\nabla u_h, \nabla \chi) = (f, \chi) && \forall \chi \in S_h^-, \\ &u_h = g_h = I_h g && \text{auf } \Gamma_- \end{aligned} \tag{13.20}$$

erfüllt ist. Diese Methode ist mit der elliptischen Gleichung $a \cdot \nabla u + a_0 u - h\Delta u = f$ konsistent, und deshalb können wir noch immer erwarten, dass der Fehler für glatte Lösungen von der Ordnung $O(h)$ ist (siehe Problemstellung 13.6). Es hat sich herausgestellt, dass die Methode im Falle nichtglatter Lösungen Unstetigkeiten mehr als erwünscht glättet.

Es sind aber aufwendigere Verfahren für das Hinzufügen der Diffusion entwickelt worden. Wir beschreiben nun ein solches Verfahren, das sogenannte *Stromliniendiffusions-Verfahren*, das darin besteht, ein $u_h \in S_h$ so bestimmen, dass

$$\begin{aligned} &(a \cdot \nabla u_h + a_0 u_h, \chi + h\, a \cdot \nabla \chi) = (f, \chi + h\, a \cdot \nabla \chi) && \forall \chi \in S_h^-, \\ &u_h = g_h = I_h g && \text{auf } \Gamma_- \end{aligned} \tag{13.21}$$

gilt. Wir stellen fest, dass die exakte Lösung von (13.13) die Gleichung

$$(a \cdot \nabla u + a_0 u, \chi + h\, a \cdot \nabla \chi) = (f, \chi + h\, a \cdot \nabla \chi), \quad \forall \chi \in S_h^- \tag{13.22}$$

erfüllt. Dies bedeutet, dass (13.21) mit (13.13) konsistent ist. Dieses Verfahren ist ein Beispiel für ein *Petrov-Galerkin-Verfahren*, weil wir uns dafür entschieden hatten, die Gleichung mit anderen Testfunktionen als denjenigen in S_h zu multiplizieren.

Der Rest dieses Abschnitts erfordert zum Verständnis möglicherweise ein wenig mehr Aufwand als es bei dem bisher vorgestellten Stoff der Fall war. Trotzdem behandeln wir diesen Stoff, weil wir der Ansicht sind, dass er die Schwierigkeiten illustriert, die bei der Anwendung der Methode der finiten Elemente auf hyperbolische Gleichungen erster Ordnung auftreten.

Wir beginnen mit der Diskussione der Stabilität und beschränken uns wiederum auf eine Lösung $w_h \in S_h^-$ von (13.21), die folglich auf Γ_- verschwindet. Wir wählen dann $\chi = w_h$ und verwenden (13.16) und $ab \le a^2 + \frac{1}{4}b^2$, um

$$\tfrac{1}{2}(1 + ha_0)|w_h|_{\Gamma_+}^2 + a_0 \|w_h\|^2 + h\,\|a \cdot \nabla w_h\|^2 = (f, w_h) + h(f, a \cdot \nabla w_h) \tag{13.23}$$

zu erhalten. Wie vorhin folgt daraus die Eindeutigkeit und die Existenz der Lösungen von (13.21), indem wir $f = 0$ setzen. Verwenden wir wie gewöhnlich die Cauchy-Schwarz-Ungleichung und $ha_0 > 0$, führt dies zur Stabilitätsabschätzung

$$|w_h|^2_{\Gamma_+} + a_0\|w_h\|^2 + h\|a\cdot\nabla w_h\|^2 \le (a_0^{-1}+h)\|f\|^2. \tag{13.24}$$

Beachten Sie die zusätzliche Stabilität, die durch das Vorhandensein des Terms $h\|a\cdot\nabla w_h\|^2$ gegeben ist. Wir können dies so interpretieren, dass dieses Verfahren eine künstliche Diffusion hinzufügt, allerdings nur entlang der charakteristischen Kurven (Stromlinien).

Zur Fehleranalyse werden wir auch eine etwas stärkere Stabilitätsabschätzung für den Fall benötigen, dass f die Form $f = a\cdot\nabla F$ besitzt. Diese lautet

$$|w_h|^2_{\Gamma_+} + a_0\|w_h\|^2 + h\|a\cdot\nabla w_h\|^2 \le C(h\|F\|_1^2 + h^{-1}\|F\|^2), \tag{13.25}$$

wobei C unabhängig von a_0 ist. Gehen wir wieder von (13.23) aus, gilt nun

$$h|(f, a\cdot\nabla w_h)| = h|(a\cdot\nabla F, a\cdot\nabla w_h)| \le \tfrac14 h\|a\cdot\nabla w_h\|^2 + Ch\|F\|_1^2.$$

Darüber hinaus gilt wegen der Greenschen Formel

$$(f, w_h) = (a\cdot\nabla F, w_h) = (a\cdot n\, F, w_h)_{\Gamma_+} - (F, a\cdot\nabla w_h)$$

und daher

$$|(f,w_h)| \le C|F|^2_{\Gamma_+} + \tfrac14|w_h|^2_{\Gamma_+} + h^{-1}\|F\|^2 + \tfrac14 h\|a\cdot\nabla w_h\|^2.$$

Unter Verwendung der Spurungleichung ergibt sich

$$|F|^2_{\Gamma_+} \le C\|F\|\,\|F\|_1 \le Ch\|F\|_1^2 + Ch^{-1}\|F\|^2 \tag{13.26}$$

(vgl. Problemstellung A.16). Der Beweis von (13.25) ergibt sich wie in (13.23).

Wir sind nun in der Lage, die folgende Fehlerabschätzung aufzustellen, die eine Verbesserung um eine halbe Potenz von h im Vergleich zur gewöhnlichen Galerkin-Methode darstellt. Wir bemerken außerdem, dass der Fehler für den Fluss von optimaler Ordnung ist, es gilt also $\|a\cdot\nabla e\| = O(h)$.

Theorem 13.6. *Seien u_h und u die Lösungen von* (13.21) *und* (13.13). *Dann gilt für $e = u_h - u$*

$$|e|_{\Gamma_+} + a_0^{1/2}\|e\| + h^{1/2}\|a\cdot\nabla e\| \le Ch^{3/2}\|u\|_2.$$

Beweis. Wir schreiben wiederum $u_h - u = (u_h - I_h u) + (I_h u - u) = \theta + \rho$ und erhalten unter Verwendung von (13.18) und (13.26)

$$|\rho|^2_{\Gamma_+} + a_0\|\rho\|^2 + h\|a\cdot\nabla\rho\|^2 \le C\big(h\|\rho\|_1^2 + h^{-1}\|\rho\|^2\big) \le Ch^3\|u\|_2^2.$$

Diesmal gilt wegen (13.21) und (13.22)

$$(a\cdot\nabla\theta + a_0\theta, \chi + h\,a\cdot\nabla\chi) = -(a\cdot\nabla\rho + a_0\rho, \chi + h\,a\cdot\nabla\chi) \quad \forall\chi\in S_h^-.$$

Wir wählen $\chi = \theta \in S_h^-$. Aus (13.24) mit $f = a_0\rho$ und (13.25) mit $F = \rho$ sowie (13.18) schließen wir

$$|\theta|^2_{\Gamma_+} + a_0\|\theta\|^2 + h\|a\cdot\nabla\theta\|^2 \le C\big(a_0\|\rho\|^2 + h\|\rho\|_1^2 + h^{-1}\|\rho\|^2\big) \le Ch^3\|u\|_2^2,$$

wobei das letzte C nur von der oberen Schranke für a_0 abhängt. Damit ist der Beweis vollständig. □

Die Fehlerabschätzungen aus dem vorhergehenden Theorem zeigt also, dass sich das Verfahren mit Stromliniendiffusion etwas besser verhält als die gewöhnliche Galerkin-Methode für glatte Lösungen. Der Hauptgrund für dessen Verwendung besteht darin, dass sich das Verfahren für nichtglatte Lösungen allerdings besser verhält. Dies liegt an der Tatsache, dass die zusätzliche Diffusion nur in der charakteristischen Richtung hinzugefügt wird, sodass innere Schichten nicht verschmiert werden, während die zusätzliche Diffusion Oszillationen in der Nähe der Randschichten beseitigt. Wir werden dies nicht weiter im Detail verfolgen.

Theorem 13.6 ist auch dann gültig, wenn $a_0 = 0$ ist. In diesem Fall lässt das Problem (13.13) immer noch eine eindeutige Lösung zu, weil die Eindeutigkeit und somit auch die Existenz wiederum direkt aus (13.23) mit $f = 0$ folgt. Wir haben $a_0 > 0$ angenommen, damit wir den Fehler in der Ordnung $O(h^{3/2})$ in der L_2-Norm beschränken können. Man kann leicht die Ungleichung $\|w\| \leq C\|a \cdot \nabla w\|$ für $w = 0$ auf Γ_- zeigen. Folglich müssen wir uns in Abwesenheit einer Abschätzung für $\|e\|$ mit der Fehlerschranke der Ordnung $O(h)$ in der L_2-Norm zufrieden geben. Wir haben immer noch eine Fehlerschranke der Ordnung $O(h^{3/2})$ auf Γ_+. Als nächstes werden wir das Problem in einer Folge von Gebieten so lösen, dass sich aus den Schranken auf dem zugehörigen Γ_+ eine globale Fehlerschranke der Ordnung $O(h^{3/2})$ ergibt.

Die oben angegebene Methode behandelt das hyperbolische Problem erster Ordnung also als ein zweidimensionales und löst insbesondere die diskreten Gleichungen für alle Knotenwerte der Lösung simultan. Wird diese Methode auf das Anfangs-Randwertproblem (13.7) angewendet, so geht deshalb der Evolutionsaspekt verloren. Wir werden nun zu einer Modifikation kommen, die die Vorteile des Stromliniendiffusions-Verfahrens bewahrt, den Zeitschritt-Charakter aber wiederherstellt. Dies wird dadurch erreicht, dass das Gebiet $\Omega \times \mathbf{R}_+$ in zur x-Achse parallele Streifen zerlegt wird und anschließend approximierende Funktionen verwendet werden, die an den Übergängen von einem Streifen zum nächsten unstetig sein dürfen. Diese Methode wird als *diskontinuierliche Galerkin-Methode* bezeichnet.

Betrachten wir also das Anfangs-Randwertproblem (13.7). Wir verwenden wie zuvor die durch $0 = x_0 < x_1 < \cdots < x_M = 1$ definierte Zerlegung von Ω und führen nun auch die Zerlegung $0 = t_0 < t_1 < \ldots$ von $\mathbf{R}_+$ ein. Wir nehmen der Einfachheit halber an, dass beide Zerlegungen quasiuniform sind und dass die Gitterkonstanten im Raum und in der Zeit von der gleichen Größenordnung sind. Setzen wir $h_j = x_j - x_{j-1}$, $k_j = t_j - t_{j-1}$ und $h = \max h_j$, $k = \max k_j$, bedeutet dies, dass $ch \leq h_j \leq h$, $ck \leq k_j \leq k$ für alle j mit $c > 0$ und $ch \leq k \leq Ch$ gilt. Diese Zerlegungen im Raum und in der Zeit definieren eine Zerlegung von $Q = \Omega \times \mathbf{R}_+$ in Rechtecke. Diese können durch Einfügen von Diagonalen mit positivem Anstieg in Dreiecke zerlegt werden und bilden damit eine Triangulation, die die Anwendung des oben diskutierten Stromliniendiffusions-Verfahrens auf jedem endlichen Zeitintervall erlauben würde. Wir werden bei unserer Analyse die ungeteilten Rechtecke verwenden

und definieren

$$S_{h,k} = \Big\{ V(x,t) = \alpha^n(x)\frac{t-t_{n-1}}{k_n} + \beta^n(x)\frac{t_n - t}{k_n} \\ \text{für } t \in M_n \quad \text{mit } \alpha^n, \beta^n \in S_h^- \Big\}$$

mit $M_n = (t_{n-1}, t_n)$, wobei S_h^- den in (13.8) definierten Raum der stückweise linearen Funktionen bezeichnet. Beachten Sie, dass $V \in S_{h,k}$ an der Stelle t_n unstetig sein kann und dass die Gleichungen $V_{n-1}^+ = V(t_{n-1}^+) = \beta^n$, $V_n^- = V(t_n^-) = \alpha^n$ und $V_t(t) = (\alpha^n - \beta^n)/k_n$ für $t \in M_n$ gelten.

Die diskontinuierliche Galerkin-Methode mit Stromliniendiffusion für die Lösung von (13.7) besteht darin, ein $U \in S_{h,k}$ so zu bestimmen, dass $U_0^- = v_h$ gilt und folglich für $n = 1, 2, \ldots$

$$\begin{aligned} &\int_{M_n} (U_t + U_x, \chi + h(\chi_t + \chi_x))\,\mathrm{d}t + (U_{n-1}^+, \chi_{n-1}^+) \\ &\quad = \int_{M_n} (f, \chi + h(\chi_t + \chi_x))\,\mathrm{d}t + (U_{n-1}^-, \chi_{n-1}^+) \quad \forall \chi \in S_{h,k}, \end{aligned} \tag{13.27}$$

wobei die Skalarprodukte über dem eindimensionalen Intervall Ω genommen werden. Wir stellen fest, dass wir für verschwindendes f und U_{n-1}^- die Wahl $\chi = U$ treffen können, woraus sich leicht schlussfolgern lässt, dass $U = 0$ auf $\Omega \times M_n$ gilt. Folglich kann diese Gleichung nach U_n^- und U_{n-1}^+ aufgelöst werden, wenn U_{n-1}^- zusammen mit f auf $\Omega \times M_n$ gegeben ist. Das Verfahren ist deshalb ein Zeitschrittverfahren.

Wir merken an, dass die lokale Gleichung (13.27) von der Form (13.21) mit der Randbedingung $U = 0$ auf $\Gamma_{-,x}$ auf dem Gebiet $\Omega \times M_n$ ist. Die Randbedingung $U_{n-1}^+ = U_{n-1}^-$ auf $\Gamma_{-,t}$ ist allerdings nur schwach gestellt (siehe Problemstellung 13.7).

Führen wir die Gleichungen in (13.27) und die Anfangsbedingung $(U_0^- - v_h, \chi_0^+) = 0$ zusammen, so können wir die Gleichungen auf Q in schacher Form

$$B_n(U, \chi) = L_n(v_h, f; \chi) \quad \forall \chi \in S_{h,k} \quad \text{für } n \geq 1$$

stellen, wobei mit $[v]_j = v_j^+ - v_j^-$

$$B_n(v, w) = \sum_{j=1}^{n} \int_{M_j} (v_t + v_x, w + h(w_t + w_x))\,\mathrm{d}t + \sum_{j=1}^{n-1} ([v]_j, w_j^+) + (v_0^+, w_0^+)$$

und

$$L_n(v, f; w) = (v, w_0^+) + \sum_{j=1}^{n} \int_{M_j} (f, w + h(w_t + w_x))\,\mathrm{d}t \tag{13.28}$$

gilt. Wir stellen fest, dass die Sprungterme verschwinden, weil die exakte Lösung in der Zeit stetig ist und die Lösung deshalb

$$B_n(u,\chi) = L_n(v,f;\chi) \quad \forall \chi \in S_{h,k} \quad \text{für } n \geq 1 \tag{13.29}$$

erfüllt.

Durch partielle Integration können wir $B_n(\cdot,\cdot)$ in der Form

$$\begin{aligned} B_n(v,w) = &\sum_{j=1}^{n} \int_{M_j} \Big((v, -w_t - w_x) + h(v_t + v_x, w_t + w_x) \Big) \,\mathrm{d}t \\ &+ \sum_{j=1}^{n-1} (v_j^-, -[w]_j) + (v_n^-, w_n^-) + \int_0^{t_n} v(1,t) w(1,t) \,\mathrm{d}t \end{aligned} \tag{13.30}$$

schreiben. Addieren wir die beiden Formen von $B_n(\cdot,\cdot)$, so erhalten wir unter Verwendung von $v_j^- = v_j^+ - [v]_j$ für $w = v$

$$\begin{aligned} B_n(v,v) = &\tfrac{1}{2}\|v_n^-\|^2 + \tfrac{1}{2} \sum_{j=1}^{n-1} \|[v]_j\|^2 + h \sum_{j=1}^{n} \int_{M_j} \|v_t + v_x\|^2 \,\mathrm{d}t \\ &+ \tfrac{1}{2}\|v_0^+\|^2 + \tfrac{1}{2} \int_0^{t_n} v(1,t)^2 \,\mathrm{d}t. \end{aligned} \tag{13.31}$$

Wir kommen nun zur Fehleranalyse.

Theorem 13.7. *Seien U und u die Lösungen von* (13.27) *und* (13.7)*. Dann gilt für geeignet gewähltes v_h für $e = U - u$*

$$\begin{aligned} &\|e_n^-\|^2 + \sum_{j=1}^{n-1} \|[e]_j\|^2 + h \sum_{j=1}^{n} \int_{M_j} \|e_t + e_x\|^2 \,\mathrm{d}t \\ &\qquad \leq Ch^3 \int_0^{t_n} \left(\|u\|_2^2 + \|u_t\|_1^2 + \|u_{tt}\|^2 \right) \mathrm{d}t \quad \textit{für } n \geq 0. \end{aligned} \tag{13.32}$$

Wir weisen darauf hin, dass der erste Term lauf der linken Seite eine Fehlerabschätzung der Ordnung $O(h^{3/2})$ von links an der Stelle t_n zeigt. Da die Sprünge an den Zeitebenen im zweiten Term beschränkt sind, ist der Fehler $e_n^+ = e_n^- + [e]_n$ rechts von der Stelle t_n auch von der Ordnung $O(h^{3/2})$ und wir können schlussfolgern, dass dies überall auf M_n gilt, weil $e(t) = k^{-1}(t - t_{n-1})e_n^- + k^{-1}(t_n - t)e_{n-1}^+ + (\bar{u}(t) - u(t))$ ist. Dabei bezeichnen wir mit $\bar{u}$ die lineare Interpolierte von u, sodass der letzte Term von der Ordnung $O(h^2)$ ist.

Beweis von Theorem 13.7. Der Beweis verläuft wie der von Theorem 13.6. Alle unten auftretenden Terme besitzen dort ihre Entsprechung. Sei I_h der Interpolationsoperator in S_h^- und J_k der stückweise lineare Interpolationsoperator in der Zeit. Wir schreiben

$$U - u = (U - \widetilde{u}) + (\widetilde{u} - u) = \theta + \rho \quad \text{mit } \widetilde{u} = J_k I_h u.$$

Beachten Sie, dass $\widetilde{u}(\cdot,t)$ in der Zeit kontinuierlich ist. Wegen (13.31) reicht es aus, $B_n(e,e)$ abzuschätzen, und wir stellen fest, dass

$$B_n(e,e) \le 2B_n(\theta,\theta) + 2B_n(\rho,\rho)$$

gilt. Wir beginnen mit dem Abschätzen des ersten Terms auf der rechten Seite. Wegen (13.29) und (13.27) finden wir für jedes $\chi \in S_{h,k}$

$$\begin{aligned} B_n(\theta,\chi) &= B_n(U,\chi) - B_n(\widetilde{u},\chi) = L_n(v_h,f;\chi) - B_n(\widetilde{u},\chi) \\ &= L_n(v_h,f;\chi) + \big(B_n(u,\chi) - L_n(v,f;\chi)\big) - B_n(\widetilde{u},\chi) \\ &= (v_h - v,\chi_0^+) - B_n(\rho,\chi) = (e_0,\chi_0^+) - B_n(\rho,\chi) = -B_n(\rho,\chi), \end{aligned}$$

wobei wir nun den Anfangswert $v_h = P_h v$ gewählt haben, sodass $(e_0,\chi_0^+) = 0$ gilt. Setzen wir $\chi = \theta$ und verwenden (13.30) für $B_n(\cdot,\cdot)$, so folgt

$$\begin{aligned} B_n(\theta,\theta) = |B_n(\rho,\theta)| &\le \sum_{j=1}^{n} \int_{M_j} \big(\|\rho\| + h\|\rho_t + \rho_x\|\big)\,\|\theta_t + \theta_x\|\,\mathrm{d}t \\ &\quad + \sum_{j=1}^{n-1} \|\rho_j\|\,\|[\theta]_j\| + \|\rho_n\|\,\|\theta_n^-\| + \int_0^{t_n} |\rho(1,t)|\,|\theta(1,t)|\,\mathrm{d}t \\ &\le \tfrac{1}{2}B_n(\theta,\theta) + Ch^{-1}\int_0^{t_n} \|\rho\|^2\,\mathrm{d}t + \sum_{j=1}^{n} \|\rho_j\|^2 + CB_n(\rho,\rho). \end{aligned}$$

Hierbei haben wir benutzt, dass ρ in der Zeit stetig ist, sodass $\rho_n^- = \rho_n$ gilt. Zum Abschluss des Beweises vernachlässigen wir $B_n(\theta,\theta)$ und schätzen die letzten drei Terme ab. Beachten Sie zunächst, dass wegen (13.31)

$$B_n(\rho,\rho) = \tfrac{1}{2}\|\rho_n\|^2 + h\int_0^{t_n} \|\rho_t + \rho_x\|^2\,\mathrm{d}t + \tfrac{1}{2}\|\rho_0\|^2 + \tfrac{1}{2}\int_0^{t_n} \rho(1,t)^2\,\mathrm{d}t$$

gilt. Wir schreiben

$$\rho = J_k I_h u - u = J_k(I_h u - u) + (J_k u - u) = J_k\eta + \omega.$$

Unter Verwendung der gewöhnlichen Abschätzungen für J_k ergibt sich mit $\|v\|_{M_j}^2 = \int_{M_j} \|v\|^2\,\mathrm{d}t$

$$\|J_k v - v\|_{M_j} + k_j\|D_t(J_k v - v)\|_{M_j} \le Ck_j^s\|D_t^s v\|_{M_j} \quad \text{für } s = 1,2.$$

Weil die Zerlegungen quasiuniform sind und h und k die gleiche Größenordnung haben, erhalten wir

$$\begin{aligned} &\sum_{j=1}^{n} \int_{M_j} \big(h^{-1}\|\omega\|^2 + h\|\omega_t\|^2 + h\|\omega_x\|^2\big)\,\mathrm{d}t \\ &\qquad \le C(h^{-1}k^4 + hk^2)\int_0^{t_n} \|u_{tt}\|^2\,\mathrm{d}t + Chk^2\int_0^{t_n} \|u_t\|_1^2\,\mathrm{d}t \\ &\qquad \le Ch^3\int_0^{t_n} \big(\|u_{tt}\|^2 + \|u_t\|_1^2\big)\,\mathrm{d}t. \end{aligned}$$

Anschließend stellen wir fest, dass $\|J_k\eta(t)\| \le \max_{s\in M_j}\|\eta(s)\|$ für $t \in M_j$ gilt und benutzen die Spurungleichung aus Problemstellung A.12, um uns davon zu überzeugen, dass

$$\begin{aligned}\|\eta(t)\|^2 &\le Ck_j^{-1}\int_{M_j}\|\eta\|^2\,\mathrm{d}t + Ck_j\int_{M_j}\|\eta_t\|^2\,\mathrm{d}t\\ &\le Ch^3\int_{M_j}\big(\|u\|_2^2+\|u_t\|_1^2\big)\,\mathrm{d}t \quad \text{für } t\in M_j,\end{aligned} \tag{13.33}$$

gilt. Damit ist

$$h^{-1}\int_0^{t_n}\|J_k\eta\|^2\,\mathrm{d}t \le Ch^3\int_0^{t_n}\big(\|u\|_2^2+\|u_t\|_1^2\big)\,\mathrm{d}t.$$

Auf ähnliche Weise erhalten wir unter Verwendung von

$$\|(J_k\eta)_t(t)\| = k_j^{-1}\|\eta_j-\eta_{j-1}\| \le 2k_j^{-1}\max_{s\in M_j}\|\eta(s)\| \quad \text{für } t\in M_j$$

die Gleichung

$$h\sum_{j=1}^{n}\int_{M_j}\big(\|(J_k\eta)_x\|^2+\|(J_k\eta)_t\|^2\big)\,\mathrm{d}t \le Ch^3\int_0^{t_n}(\|u\|_2^2+\|u_t\|_1^2)\,\mathrm{d}t.$$

Ferner ergibt sich aus (13.33)

$$\sum_{j=0}^{n}\|\rho_j\|^2 = \sum_{j=0}^{n}\|\eta_j\|^2 \le Ch^3\int_0^{t_n}\big(\|u\|_2^2+\|u_t\|_1^2\big)\,\mathrm{d}t.$$

Verwenden wir wiederum die Spurungleichung aus Problemstellung A.12, so erhalten wir schließlich

$$\begin{aligned}\int_0^{t_n}|\rho(1,t)|^2\,\mathrm{d}t = \int_0^{t_n}|\omega(1,t)|^2\,\mathrm{d}t &\le C\int_0^{t_n}\|\omega(\cdot,t)\|\,\|\omega(\cdot,t)\|_1\,\mathrm{d}t\\ &\le Ck^3\Big(\int_0^{t_n}\|u_{tt}\|^2\,\mathrm{d}t\int_0^{t_n}\|u_t\|_1^2\,\mathrm{d}t\Big)^{1/2}\\ &\le Ch^3\int_{M_j}\big(\|u_{tt}\|^2+\|u_t\|_1^2\big)\,\mathrm{d}t.\end{aligned}$$

Damit ist der Beweis vollständig. □

13.3 Problemstellungen

Problem 13.1. Betrachten Sie das Anfangs-Randwertproblem

$$
\begin{aligned}
&u_{tt} = u_{xx} && x \in (0,1),\ t > 0,\\
&u(0,t) = u(1,t) = 0 && t > 0,\\
&u(x,0) = v(x)\ u_t(x,0) = w(x) && x \in (0,1).
\end{aligned}
$$

Betrachten Sie zur numerischen Lösung mithilfe der Galerkin-Methode der finiten Elemente die stückweise linearen, stetigen Funktionen auf Basis der Zerlegung von $[0,1]$ in M Intervalle gleicher Länge $h = 1/M$. Bestimmen Sie in Analogie zu den in (13.2) und (13.6) beschriebenen Verfahren die Matrixformen des semidiskreten und des vollständig diskreten Verfahrens.

Problem 13.2. (Übung am Rechner.) Lösen Sie das Anfangs-Randwertproblem aus Problemstellung 13.1 mit $v(x) = 0,\ w(x) = \sin(2\pi x)$. Verwenden Sie $M = 10,\ 20$ und das Zeitschrittverfahren in (13.6) mit $k = 1/10,\ 1/20$. Vergleichen Sie die numerische Lösung mit der in Abschnitt 11.2 angegebenen exakten Lösung an der Stelle $t = 3/4$.

Problem 13.3. Schreiben sie die Wellengleichung (13.1) als ein System zweier Gleichungen erster Ordnung in der Zeit, indem Sie $w_1 = u,\ w_2 = u_t$ setzen. Diskretisieren Sie das System mithilfe der gewöhnlichen Finite-Elemente-Methode in der räumlichen Variable und mithilfe des Crank-Nicolson-Verfahrens in der Zeitvariable. Zeigen Sie durch Eliminieren von W_2^n, dass das resultierende Verfahren im Wesenltichen dasselbe wie (13.6) ist. Beweisen Sie die Stabilität im Fall $f = 0$. Vergleichen Sie mit Lemma 13.2. Hinweis: Multiplizieren Sie das System mit $(W_1^{n-\frac{1}{2}}, -\Delta_h W_2^{n-\frac{1}{2}})$.

Problem 13.4. Beweisen Sie Theorem 13.2.

Problem 13.5. Beweisen Sie die Stabilitätsabschätzung (13.12).

Problem 13.6. Beweisen Sie Stabilitäts- und Fehlerabschätzungen für die gewöhnliche Galerkin-Methode mit künstlicher Diffusion (13.20).

Problem 13.7. (Schwach gestellte Randbedingung.) Die Randbedingung $u = g$ ist in (13.15) und (13.21) stark gestellt. Es ist auch möglich, die Randbedingung bei der gewöhnlichen Galerkin-Methode schwach zu stellen: Bestimmen Sie eine Funktion $u_h \in S_h$ so, dass

$$(a \cdot \nabla u_h, \chi) + (a_0 u_h, \chi) - (a \cdot n\, u_h, \chi)_{\Gamma_-} = (f, \chi) - (a \cdot n\, g, \chi)_{\Gamma_-} \quad \forall \chi \in S_h$$

gilt. Bei der Modifikation der Stromliniendiffusion sollen Sie eine Funktion $u_h \in S_h$ so bestimmen, dass

$$
\begin{aligned}
(a \cdot \nabla u_h, \chi + ha \cdot \nabla \chi) + (a_0 u_h, \chi + ha \cdot \nabla \chi) - (a \cdot n\, u_h, \chi)_{\Gamma_-}&\\
= (f, \chi + ha \cdot \nabla \chi) - (a \cdot n\, g, \chi)_{\Gamma_-} \quad \forall \chi \in S_h&
\end{aligned}
$$

gilt. Beweisen Sie Stabilitäts- und Fehlerabschätzungen für diese Methoden.

14

Weitere Klassen numerischer Methoden

Es sind numerische Methoden entwickelt worden, die sich von finiten Differenzenverfahren und Methoden finiter Elemente unterscheiden, aber häufig mit diesen in Verbindung stehen. Diese Methoden sind ebenfalls von Interesse. In diesem Kapitel geben wir einen kurzen Überblick über solche Klassen von Methoden. Dies sind *Kollokationsverfahren, Spektralmethoden, finite Volumenmethoden* und *Randelementmethoden*.

14.1 Kollokationsverfahren

Beim *Kollokationsverfahren* sucht man nach einer approximativen Lösung einer Differentialgleichung in einem endlichdimensionalen Raum hinreichend regulärer Funktionen, indem man fordert, dass die Gleichung genau in einer endlichen Anzahl von Punkten erfüllt ist. Wir beschreiben ein solches Verfahren für das parabolische Modellproblem

$$\begin{aligned} &u_t = u_{xx} && \text{in } \Omega = (0,1) \quad \text{für } t > 0, \\ &u(0,t) = u(1,t) = 0 && \text{für } t > 0, \\ &u(\cdot,0) = v && \text{in } \Omega. \end{aligned}$$

Wir setzen $h = 1/M,\ x_j = jh$ mit $0 \le j \le M$ und $K_j = [x_{j-1}, x_j]$ und führen den Raum stückweiser Polynome

$$S_h = \left\{ v \in \mathcal{C}^1(\bar{\Omega}) : v\big|_{K_j} \in \Pi_{r-1},\ v(0) = v(1) = 0 \right\} \quad \text{mit } r \ge 4$$

ein. Seien $\xi_i,\ i = 1, \ldots, r-2$ die Gauß-Punkte in $(0,1)$, d. h. die Nullstellen des Legendre-Polynoms $\tilde{P}_{r-2}(x) = P_{r-2}(2x-1)$, das von dem Intervall $(-1,1)$ auf das Intervall $(0,1)$ reskaliert wurde. Wir definieren die Kollokationspunkte $x_{j,i} = x_{j-1} + h\xi_i$ in K_j und stellen das semidiskrete Problem auf, ein $u_h(\cdot,t) \in S_h$ für $t > 0$ so zu bestimmen, dass

$$u_{h,t}(x_{j,i},t) = u_{h,xx}(x_{j,i},t) \quad \text{für } 1 \le j \le M,\ 1 \le i \le r-2\ t > 0 \tag{14.1}$$

mit einer Approximation $u_h(\cdot,0) = v_h \in S_h$ von v gilt. Dieses Verfahren kann als eine Galerkin-Methode betrachtet werden, die ein diskretes Skalarprodukt auf Grundlage der Gaußschen Quadraturformel verwendet. Seien ω_i die Gewichte in der Gaußschen Formel

$$\sum_{i=1}^{r-2} \omega_i \, \varphi(\xi_i) \approx \int_0^1 \varphi(x)\,\mathrm{d}x,$$

die für Polynome vom maximalen Grad $2r-5$ exakt ist. Setzen wir

$$(\psi,\chi)_h = h \sum_{j=1}^{M} \sum_{i=1}^{r-2} \omega_i \, \psi(x_{j,i})\, \chi(x_{j,i}) \approx (\psi,\chi), \tag{14.2}$$

dann können wir (14.1) in der Form

$$(u_{h,t},\chi)_h - (u_{h,xx},\chi)_h = 0 \quad \forall \chi \in S_h \; t > 0$$

schreiben. Für geeignet gewähltes v_h kann man eine globale Fehlerabschätzung

$$\|u_h(t) - u(t)\|_{\mathcal{C}} \le C h^r \Big\{ \max_{s\le t} \|u(s)\|_{r+2} + \Big(\int_0^t \|u_t(s)\|_{r+2}^2 \,\mathrm{d}s \Big)^{1/2} \Big\}$$

beweisen. Darüber hinaus liegt für $r > 4$ und mit einer raffinierteren Wahl der Anfangsapproximation v_h Superkonvergenz vor, sodass

$$|u_h(x_j,t) - u(x_j,t)| \le C_T h^{2r-4} \sup_{s\le t} \sum_{p+2q\le 2r-1} \|u^{(q)}(s)\|_p \quad \text{für } t \le T$$

gilt. Beachten Sie die strengeren Regularitätsforderungen gegenüber den finiten Differenzenverfahren und den Methoden finiter Elemente, die in den Kapiteln 9–10 diskutiert wurden. Die hier vorgestellten Resultate lassen sich auf vollständig diskrete Verfahren übertragen, indem man finite Differenzenapproximation und Kollokation in der Zeit verwendet.

14.2 Spektralmethoden

Spektralmethoden ähneln in vielerlei Hinsicht den Methoden finiter Elemente und den Kollokationsverfahren. Der Hauptunterschied besteht in der Wahl der endlichdimensionalen Approximationsräume.

Wir betrachten das Anfangswertproblem

$$\begin{aligned} &u_t - u_{xx} = f && \text{in } \Omega = (0,1) \quad \text{für } t > 0, \\ &u(0,t) = u(1,t) = 0 && \text{für } t > 0, \\ &u(\cdot,0) = v && \text{in } \Omega. \end{aligned} \tag{14.3}$$

Sei $\{\varphi_j\}_{j=1}^{\infty}$ eine Folge linear unabhängiger Funktionen in $H^2 \cap H_0^1$, die den L_2 aufspannen. Wir setzen $S_N = \operatorname{span}\{\varphi_j\}_{j=1}^{N}$. Unter Verwendung der Galerkin-Methode definieren wir eine räumlich semidiskrete Approximation $u_N = u_N(t) \in S_N$ von (14.3) durch

$$(u_{N,t}, \chi) - (u_{N,xx}, \chi) = (f, \chi) \quad \forall \chi \in S_N,\ t > 0 \tag{14.4}$$

mit geeignet gewählten $u_N(0) = v_N \in S_N$. Führen wir die orthogonale Projektion $P_N : L_2 \to S_N$ ein, so können wir (14.4) in der Form

$$u_{N,t} + \mathcal{A}_N u_N = P_N f \quad \text{für } t > 0 \quad \text{mit } \mathcal{A}_N = P_N \mathcal{A} P_N,\ \mathcal{A} = \partial^2/\partial x^2$$

schreiben. Dabei gilt $(\mathcal{A}_N \chi, \chi) = (\mathcal{A} P_N \chi, P_N \chi) \geq 0$. Diese Gleichung kann mit $u_N(x,t) = \sum_{j=1}^{N} \alpha_j(t)\varphi_j(x)$ als $B\alpha'(t) + A\alpha(t) = b(t)$ für $t > 0$ geschrieben werden, wobei die Elemente der $N \times N$-Matrizen A und B die Werte $(\mathcal{A}\varphi_i, \varphi_j)$ beziehungsweise (φ_i, φ_j) sind. Man sieht leicht, dass B positiv definit ist.

Wir stellen fest, dass der Fehler $e_N = u_N - u$ die Gleichung

$$e_{N,t} + \mathcal{A}_N e_N = (P_N - I)f - (\mathcal{A}_N - \mathcal{A})u \quad \text{für } t > 0 \quad \text{mit } e_N(0) = v_N - v$$

erfüllt und somit

$$\|u_N(t) - u(t)\| \leq \|v_N - v\| + \int_0^t \big(\|(P_N - I)f\| + \|(\mathcal{A}_N - \mathcal{A})u\|\big)\,\mathrm{d}s \tag{14.5}$$

gilt. Man kann sich leicht davon überzeugen, dass der zugehörige Lösungsoperator $E_N(t) = e^{-\mathcal{A}_N t}$ in der Operatornorm in L_2 durch 1 beschränkt ist. Daraus folgt, dass der Fehler mit $v_N - v$, $(P_N - I)f$ und $(\mathcal{A}_N - \mathcal{A})u$ klein ist.

Wir betrachten ein einfaches Beispiel. Dabei sind $\varphi_j(x) = c\sin(j\pi x)$ die normierten Eigenfunktionen von $\mathcal{A}$ mit homogenen Dirichletschen Randbedingungen. Dann gilt $B = I$, A ist positiv definit und P_N ist die abgeschnittene Fourier-Reihe $P_N v = \sum_{j=1}^{N} (v, \varphi_j)\varphi_j$, sodass $\mathcal{A}_N v = \sum_{j=1}^{N} (j\pi)^2 (v, \varphi_j)\varphi_j = P_N \mathcal{A} v$ gilt. Der Fehler ist klein, wenn $v_N = P_N v$ gilt und die Fourier-Reihen von v, f und u_{xx} konvergieren. Insbesondere ist die Konvergenz für jedes r von der Ordnung $O(N^{-r})$, vorausgesetzt die Lösung ist hinreichend regulär.

Eine andere Möglichkeit, ein semidiskretes numerisches Verfahren unter Verwendung des Raumes S_N aus unserem Beispiel zu definieren, besteht darin, aus S_N mit dem Skalarprodukt $(v, w)_N = h \sum_{j=0}^{N-1} v(x_j) w(x_j)$ einen Hilbert-Raum zu machen, wobei $x_j = j/(N-1)$ gilt. Dies führt auf eine Projektion P_N, die durch $P_N u(x_j) = u(x_j)$, $j = 0, \ldots, N-1$ definiert ist, und die semidiskrete Gleichung (14.4) wird nun zur Kollokationsgleichung

$$u_{N,t}(x_j, t) - u_{N,xx}(x_j, t) = f(x_j, t) \quad \text{für } j = 0, \ldots, N-1,\ t > 0.$$

Dies wird auch als *Pseudospektralmethode* bezeichnet und die Fehlerabschätzung (14.5) ist in der zu $(\cdot, \cdot)_N$ gehörigen diskreten Norm gültig.

Spektral- und Pseudospektralmethoden, die die oben betrachteten sinusförmigen Basisfunktionen verwenden, sind bei periodischen Problemen besonders hilfreich. Im Falle von Anfangs-Randwertproblemen für hyperbolische Gleichungen sind auf Chebyshev- und Legendre-Polynome bezogene Basisfunktionen nützlich, beispielsweise in Verbindung mit Berechnungen in der Strömungsdynamik.

14.3 Finite Volumenverfahren

Wir illustrieren die Verwendung der *finiten Volumenverfahren* am Beispiel des Modellproblems

$$-\Delta u = f \quad \text{in } \Omega \quad \text{mit } u = 0 \quad \text{auf } \Gamma, \tag{14.6}$$

wobei Ω ein konvexes polygonales Gebiet in $\mathbf{R}^2$ mit dem Rand Γ ist. Die Grundlage dieser Methode ist die Feststellung, dass wegen der Greenschen Formel für jedes $V \subset \Omega$

$$\int_{\partial V} \frac{\partial u}{\partial n}\, \mathrm{d}s = \int_V f\, \mathrm{d}x \tag{14.7}$$

gilt. Wir beginnen mit der Beschreibung des *zellzentrierten* finiten Volumen-Differenzenverfahrens. Sei $\mathcal{T}_h = \{K_j\}$ eine Triangulation von Ω von dem in Abschnitt 5 betrachteten Typ, dei dem alle Winkel der K_j kleiner gleich $\pi/2$ sind. Wir betrachten (14.7) mit $V = K_j \in \mathcal{T}_h$. Dann ist $\partial V = \partial K_j$ die Vereinigung der mit drei anderen Dreiecken K_i gemeinsamen Kanten γ_{ji}. Wir wollen $\partial u/\partial n$ auf jeder dieser Kanten approximieren. Mit Q_j bezeichnen wir den Mittelpunkt des Kreises um K_j (der dann innerhalb von K_j liegt), der Vektor Q_jQ_j ist orthogonal zu γ_{ji} und $\partial u/\partial n$ in (14.7) kann durch den Differenzenquotienten $(U(Q_i) - U(Q_j))/|Q_iQ_j|$ approximiert werden. Unter Verwendung der Randwerte in (14.6) für die zu den Randdreiecken gehörigen Q_j erzeugt dies ein finites Differenzenverfahren auf dem nichtuniformen Gitter $\{Q_j\}$. Schreiben wir das diskrete Problem in Matrixform als $AU = b$, so können wir zeigen, dass die Matrix A symmetrisch, positiv definit und diagonaldominant ist. Sind die $\mathcal{T}_h$ quasiuniform, kann man die Fehlerabschätzung

$$\|U - u\|_{1,h} \le Ch\|u\|_2$$

in einer bestimmten diskreten H^1-Norm beweisen.

Eine alternative Herangehensweise stellt das folgende *knotenzentrierte Verfahren* dar, das auch als *finites Volumenelementverfahren* bezeichnet wird: Sei $S_h \subset H_0^1$ der durch $\mathcal{T}_h$ definierte Raum stückweise linearer finiter Elemente. Für $K \in \mathcal{T}_h$ schneiden sich die Geraden, die einen Knoten mit dem Mittelpunkt der gegenüberliegenden Kante verbinden, im Schwerpunkt von K und unterteilen K in sechs Dreiecke. Sei $B_{j,K}$ die Vereinigung von zwei dieser

Dreiecke, die P_j als Eckpunkt besitzen. Für jeden inneren Knoten P_j betrachten wir die Vereinigung B_j der zugehörigen $B_{j,K}$ und bezeichnen mit $\bar{S}_h$ die zugehörigen stückweise konstanten Funktionen. Unter Verwendung von (14.7) für jedes dieser B_j kommen wir zum Petrov-Galerkin-Verfahren, das darin besteht, ein $u_h \in S_h$ so zu bestimmen, dass

$$\bar{a}(u_h, \psi) := \sum_j \psi_j \int_{\partial B_j} \frac{\partial u_h}{\partial n} \, \mathrm{d}s = (f, \psi) \quad \forall \psi \in \bar{S}_h \tag{14.8}$$

erfüllt ist. Dies kann auch als ein finites Differenzenverfahren auf dem unregelmäßigen Gitter $\{P_j\}$ betrachtet werden. Die B_j werden als Kontrollvolumina bezeichnet. Sind $\chi \in S_h$ die Funktionen $\bar{\chi} \in \bar{S}_h$, die an den Knoten von $\mathcal{T}_h$ mit χ übereinstimmen, kann man zeigen, dass

$$\bar{a}(\psi, \bar{\chi}) = a(\psi, \chi) \quad \forall \psi, \chi \in S_h \tag{14.9}$$

gilt (vgl. Problemstellung 14.3), sodass (14.8) in der Form

$$a(u_h, \chi) = (f, \bar{\chi}) \quad \forall \chi \in S_h$$

geschrieben werden kann. (Dies gilt für elliptische Operatoren mit variablen Koeffizienten nicht exakt.) Man kann zeigen, dass die gewöhnliche Fehlerabschätzung

$$\|u_h - u\|_1 \le Ch\|u\|_2$$

für dieses Verfahren gilt. Unter etwas strengeren Regularitätsannahmen gilt ebenso $\|u_h - u\| = O(h^2)$.

Finite Volumenverfahren sind für Operatoren in Divergenzform nützlich, insbesondere im Falle zeitabhängiger Erhaltungsgleichungen.

14.4 Randelementmethoden

Bei der *Randelementmethode* wird ein Randwertproblem für eine homogene partielle Differentialgleichung in einem Gebiet Ω mit der auf dem Rand Γ gegebenen Lösung u als eine Integralgleichung über Γ umformuliert. Diese Gleichung dient dann als Grundlage für die numerische Approximation. Wir werden diese Vorgehensweise anhand des Modellproblems

$$\Delta u = 0 \quad \text{in } \Omega \subset \mathbf{R}^2 \quad \text{mit } u = g \quad \text{auf } \Gamma \tag{14.10}$$

illustrieren. Dabei nehmen wir an, dass Γ glatt ist. Zum Aufstellen der Randintegralgleichung nehmen wir an, dass $U(x) = -(2\pi)^{-1} \log|x|$ die Fundamentallösung des Laplace-Operators in $\mathbf{R}^2$ ist (siehe Theorem 3.5). Für jedes u mit $\Delta u = 0$ auf Γ gilt wegen der Greenschen Formel

$$u(x) = \int_\Gamma U(x-y) \frac{\partial u}{\partial n_y}(y) \, \mathrm{d}s_y - \int_\Gamma \frac{\partial U}{\partial n_y}(x-y) u(y) \, \mathrm{d}s_y, \quad x \in \Omega. \tag{14.11}$$

Mit x auf Γ definieren die Integrale auf der rechten Seite die Einfachschicht- und Doppelschichtpotentiale $V\partial u/\partial n$ und Wu. Wir weisen darauf hin, dass der Kern $(\partial U/\partial n_y)(x-y)$ für $x, y \in \Gamma$ beschränkt ist, auch wenn $\nabla U(x-y)$ Singularitäten der Ordnung $O(|x-y|^{-1})$ besitzt, sodass der Operator W wohldefiniert ist. Wenn sich $x \in \Omega$ dem Rand Γ nähert, gehen die beiden Integrale gegen $V\partial u/\partial n$ beziehungsweise $\frac{1}{2}u + Wu$, sodass (14.11) auf

$$\tfrac{1}{2}u = V\partial u/\partial n - Wu \quad \text{auf } \Gamma$$

führt. Mit $u = g$ auf Γ ist dies eine Fredholmsche Integralgleichung erster Art zur Bestimmung von $\partial u/\partial n$ auf Γ. Setzen wir dies zusammen mit $u = g$ auf Γ in (14.11) ein, ergibt sich die Lösung von (14.10).

Anstelle dieser direkten Methode kann man die indirekte Methode verwenden, bei der angenommen wird, dass die Lösung von (14.11) als Potential einer Funktion auf Γ dargestellt werden kann, sodass

$$u(x) = \int_\Gamma \Phi(x-y)v(y)\,\mathrm{d}s_y \text{ oder } u(x) = \int_\Gamma \frac{\partial \Phi}{\partial n_y}(x-y)w(y)\,\mathrm{d}s_y \quad x \in \Omega$$

gilt. Mit den obigen V und W gilt: Wenn solche Funktionen v und w existieren, erfüllen sie die Fredholmschen Integralgleichungen erster und zweiter Art

$$Vv = g \quad \text{und} \quad \tfrac{1}{2}w + Ww = g \quad \text{auf } \Gamma. \tag{14.12}$$

Schreiben wir $H^s = H^s(\Gamma)$, dann sind V und W sogenannte Pseudodifferentialoperatoren der Ordnung -1, d. h. beschränkte lineare Operatoren $H^s \to H^{s+1}$, die insbesondere kompakt auf H^s sind. Die Gleichung erster Art ist eindeutig lösbar, vorausgesetzt ein bestimmtes Maß, der transfinite Durchmesser δ_Γ von Γ, ist dergestalt, dass $\delta_\Gamma \neq 1$ erfüllt ist. Die Gleichung zweiter Art in (14.12) besitzt immer eine eindeutige Lösung. Ähnliche Umformulierungen können auch für die Neumannschen Randbedingungen für eine große Anzahl anderer Probleme, die elliptische Gleichungen beinhalten, und für äußere Probleme vorgenommen werden. Diese Herangehensweise ist zur numerischen Lösung im letzten Fall besonders nützlich.

Bei der Randelementmethode (engl.: *Boundary Element Method – BEM*) bestimmt man die approximative Lösung einer Randintegralformulierung des Problems wie der obigen in einem Raum, der ähnlich dem stückweise polynomialen Finite-Elemente-Raum ist, mithilfe der Galerkin-Methode oder des Kollokationsverfahrens.

Im Falle der zweiten Gleichung in (14.12) bestimmen wir mithilfe der Galerkin-Methode und einem endlichdimensionalen Teilraum S_h von $L_2(\Gamma)$ die diskrete Approximation $w_h \in S_h$ von w, die die Form

$$\tfrac{1}{2}\langle w_h, \chi\rangle + \langle Ww_h, \chi\rangle = \langle g, \chi\rangle \quad \forall \chi \in S_h \quad \text{mit } \langle\cdot,\cdot\rangle = (\cdot,\cdot)_{L_2(\Gamma)}$$

besitzt. Schreibt man $|\cdot|_s$ für die Norm in $H^s(\Gamma)$, erhält man $|w_h - w|_0 \leq C_r(w)h^r$, wenn S_h von der Genauigkeit der Ordnung $O(h^r)$ ist. Mithilfe eines Dualitätsargumentes kann man die superkonvergente Abschätzung

$|w_h - w|_{-r} \leq C_r(w)h^{2r}$ zeigen. Unter Verwendung eines Iterationsargumentes lässt sich damit eine approximative Lösung $\tilde{w}_h$ mit $|\tilde{w}_h - w|_0 = O(h^{2r})$ definieren.

Betrachten wir beispielsweise die numerische Lösung der Gleichung erster Art in (14.12) im endlichdimensionalen Raum S_h periodischer, glatter Splines der Ordnung r. In diesem Fall besteht $S_h \subset \mathcal{C}^{r-2}$ aus stückweisen Polynomen in Π_{r-1}. Bei unserem diskreten Problem ist also ein $v_h \in S_h$ gesucht, für das

$$\langle V v_h, \chi \rangle = \langle g, \chi \rangle \quad \forall \chi \in S_h$$

gilt. Man kann zeigen, dass die zu V gehörige Bilinearform $\langle Vv, w \rangle$ symmetrisch, beschränkt und koerzitiv bezüglich der Norm $|\cdot|_{-1/2}$ in einem bestimmten Sobolev-Raum $H^{-1/2}(\Gamma)$ ist, sodass

$$\langle Vv, w \rangle = \langle v, Vw \rangle \leq C|v|_{-1/2}|w|_{-1/2} \quad \text{und} \quad \langle Vv, v \rangle \geq c|v|^2_{-1/2} \quad \text{mit } c > 0$$

gilt. Man kann dann die Ungleichung

$$|v_h - v|_{-1/2} \leq C \inf_{\chi \in S_h} |\chi - v|_{-1/2} \leq Ch^{r+1/2}|v|_r$$

beweisen. Mit einem Dualitätsargument folgt dann $|v_h - v|_{-r-1} \leq Ch^{2r+1}|v|_r$, wobei wir die Norm in $H^{-r-1}(\Gamma)$ verwenden. Ist x ein innerer Punkt von Ω, finden wir deshalb für $u_h = Vv_h$ die Abschätzung $|u_h(x) - u(x)| \leq C_x|v_h - v|_{-r-1} \leq Ch^{2r+1}$, weil $\Phi(x - y)$ für $y \neq x$ glatt ist.

Als Funktion einer Basis $\{\phi_j\}$ von S_h ausgedrückt, kann man dieses Problem in Matrixform als $A\alpha = \tilde{g}$ schreiben, wobei A symmetrisch und positiv definit ist. Auch wenn die Dimension von A durch Rückführung des ursprünglich zweidimensionalen Problems auf ein eindimensionales reduziert wurde, ist die Matrix A anders als bei der Methode der finiten Elemente für eine Differentialgleichung nicht dünn besetzt. Wir weisen außerdem darauf hin, dass die Elemente $\langle V\Phi_i, \Phi_j \rangle$ zwei Integrationen erfordern, eine zum Bilden von $V\Phi_i$ und eine zum Bilden des Skalarprodukts.

Um den Aufwand zu reduzieren, kann man Kollokationsverfahren anwenden und v_h aus $Vv_h(x(s_j)) = g(x(s_j))$ an den M_h Quadraturpunkten s_j in $[0, l]$ bestimmen, wobei $x = x(s)$ eine Parametrisierung von Γ ist und $M_h = \dim(S_h)$ gilt.

In der umfassenden Literatur über numerische Randintegralmethoden wurde großes Augenmerk auf die Komplikationen gelegt, die sich ergeben, wenn die oben genannten Regularitätsannahmen nicht erfüllt sind. Dies trifft beispielsweise für Gebiete mit Ecken zu. In diesem Fall sind V und W nicht kompakt.

14.5 Problemstellungen

Problem 14.1. Sei $r = 4$ und $(\cdot, \cdot)_h$ durch den entsprechenden Fall von (14.2) definiert.

(a) Zeigen Sie, dass $\|\chi\|_h := (\chi,\chi)_h^{1/2}$ eine Norm auf S_h ist.
(b) Zeigen Sie die Gültigkeit von

$$-(\chi'',\chi)_h \geq -(\chi'',\chi) = \|\chi'\|^2 \quad \text{für } \chi \in S_h.$$

(c) Zeigen Sie die Stabilität der Lösung von (14.1) bezüglich $\|\cdot\|_h$.

Hinweis zu Teil (b): Sei $\tilde{P}_2(x) = P_2(2x-1) = x^2 - x + \frac{1}{6}$ das zum Intervall $(0,1)$ gehörende Legendre-Polynom mit den Nullstellen $\xi_{1,2} = \frac{1}{2} \pm \frac{\sqrt{3}}{6}$. Bedenken Sie, dass die Gaußsche Quadratur mit zwei Gauß-Punkten für kubische Polynome exakt ist. Beschränken Sie Ihre Betrachtungen auf ein Intervall $(0,1)$. Sei $\chi \in \Pi_3$ mit dem Koeffizienten 1 für x^3. Dann ist $\chi''\chi - 6\tilde{P}_2^2 \in \Pi_3$ und folglich gilt wegen $\tilde{P}(\xi_i) = 0,\ i = 1,2$

$$-\tfrac{1}{2}\sum_{i=1}^{2} \chi''(\xi_i)\chi(\xi_i) = -\int_0^1 \chi''\chi \,\mathrm{d}x + 6\int_0^1 \tilde{P}_2^2 \,\mathrm{d}x \geq -\int_0^1 \chi''\chi \,\mathrm{d}x.$$

Problem 14.2. Betrachten Sie das Anfangswertproblem erster Ordnung mit periodischen Randbedingungen

$$\begin{aligned} &u_t + u_x = 0, && \text{in } \Omega = (-\pi,\pi) \quad \text{für } t > 0,\\ &u(-\pi,t) = u(\pi,t) && \text{für } t > 0,\\ &u(\cdot,0) = v && \text{in } \Omega. \end{aligned}$$

Formulieren Sie die Spektralmethode mit der Basis

$$S_N = \big\{1,\, \cos x,\, \sin x,\, \cos 2x,\, \sin 2x,\, \ldots,\, \cos Nx,\, \sin Nx\big\}.$$

Es sei $\mathcal{A} = \partial/\partial x$. Bestimmen Sie $\mathcal{A}_N$ und zeigen Sie, dass $\mathcal{A}_N^* = -\mathcal{A}_N$ gilt. Zeigen Sie auch, dass $\|u_N(t)\| \leq \|v\| = \|v\|_{L_2(\Omega)}$ und folglich $\|E_N(t)\| = 1$ mit $E_N(t) = \mathrm{e}^{-t\mathcal{A}_N}$ gilt.

Problem 14.3. Beweisen Sie (14.9). Hinweis: Schreiben Sie Ω als Vereinigung der B_j und der K und schreiben Sie diese wiederum als Vereinigung der Mengen $B_{j,K}$. Beachten Sie, dass

$$\int_e \bar{\chi} \,\mathrm{d}s = \int_e \chi \,\mathrm{d}s \quad \text{für } \chi \in S_h$$

für jede Kante e der Triangulation $\mathcal{T}_h$ gilt.

A

Einige Hilfsmittel aus der Analysis

In diesem Anhang geben wir einen kurzen Überblick über Ergebnisse aus der Analysis, insbesondere der Funktionalanalysis, die bei unserer Behandlung partieller Differentialgleichungen gebraucht werden. Dabei geben wir die Ergebnisse im Wesentlichen ohne Beweis an. Wir beginnen in Abschnitt A.1 mit einer einfachen Darstellung abstrakter, linearer Räume, wobei wir den Hilbert-Raum einschließlich des Rieszschen Darstellungssatzes und dessen Verallgemeinerung auf Bilinearformen von Lax und Milgram hervorheben. Wir fahren in Abschnitt A.2 mit Funktionenräumen fort, wobei wir uns nach der Diskussion des Raumes $\mathcal{C}^k$, der Integrabilität und des L_p-Raumes den L_2-basierten Sobolev-Räumen mit dem Spurtheorem und der Poincaré-Ungleichung zuwenden. Abschnitt A.3 beschäftigt sich mit der Fourier-Transformation.

A.1 Abstrakte lineare Räume

Sei V ein linearer Raum (oder Vektorraum) mit reellen Skalaren, d. h. eine Menge mit der Eigenschaft, dass für $u, v \in V$ und $\lambda, \mu \in \mathbf{R}$ auch $\lambda u + \mu v \in V$ gilt. Ein *lineares Funktional* (oder eine *Linearform*) L auf V ist eine Funktion $L : V \to \mathbf{R}$, für die

$$L(\lambda u + \mu v) = \lambda L(u) + \mu L(v) \quad \forall u, v \in V,\ \lambda, \mu \in \mathbf{R}$$

gilt. Eine *Bilinearform* $a(\cdot, \cdot)$ auf V ist eine Funktion $a : V \times V \to \mathbf{R}$, die in jedem Argument für sich genommen linear ist, d. h. für alle $u, v, w \in V$ und $\lambda, \mu \in \mathbf{R}$ gilt

$$\begin{aligned} a(\lambda u + \mu v, w) &= \lambda a(u, w) + \mu a(v, w), \\ a(w, \lambda u + \mu v) &= \lambda a(w, u) + \mu a(w, v). \end{aligned}$$

Die Bilinearform $a(\cdot, \cdot)$ wird als *symmetrisch* bezeichnet, wenn

$$a(w, v) = a(v, w) \quad \forall v, w \in V$$

gilt und als *positiv definit*, wenn

$$a(v,v) > 0, \quad \forall v \in V,\ v \neq 0$$

ist. Eine positiv definite, symmetrische Bilinearform auf V wird auch als *inneres Produkt* (oder *Skalarprodukt*) auf V bezeichnet. Ein linearer Raum V mit einem Skalarprodukt wird auch als *innerer Produktraum* (oder *Skalarproduktraum*) bezeichnet.

Für einen Skalarproduktraum V und ein Skalarprodukt $(\cdot,\cdot)$ auf V definieren wir die zugehörige Norm durch

$$\|v\| = (v,v)^{1/2} \quad \text{für } v \in V. \tag{A.1}$$

Wir wiederholen die *Cauchy-Schwarz-Ungleichung*

$$|(w,v)| \leq \|w\|\,\|v\| \quad \forall v,w \in V, \tag{A.2}$$

wobei Gleichheit genau dann gilt, wenn $w = \lambda v$ für ein $\lambda \in \mathbf{R}$ ist, und die *Dreiecksungleichung*

$$\|w+v\| \leq \|w\| + \|v\| \quad \forall v,w \in V. \tag{A.3}$$

Zwei Elemente $v, w \in V$ mit $(v,w) = 0$ werden als *orthogonal* bezeichnet.

Eine unendliche Folge $\{v_i\}_{i=1}^{\infty}$ in V konvergiert gegen $v \in V$, auch durch $v_i \to v$ für $i \to \infty$ oder $v = \lim_{i\to\infty} v_i$ ausgedrückt, wenn

$$\|v_i - v\| \to 0 \quad \text{für } i \to \infty$$

gilt. Die Folge $\{v_i\}_{i=1}^{\infty}$ wird als *Cauchy-Folge* in V bezeichnet, wenn

$$\|v_i - v_j\| \to 0 \quad \text{für } i,j \to \infty$$

gilt. Der Skalarproduktraum V wird als *vollständig* bezeichnet, wenn jede Cauchy-Folge in V konvergent ist, d. h. wenn jede Cauchy-Folge einen Grenzwert $v = \lim v_i \in V$ besitzt. Ein vollständiger Skalarproduktraum wird als *Hilbert-Raum* bezeichnet.

Wenn wir hervorheben wollen, dass ein Skalarprodukt oder eine Norm einem speziellen Raum V zugeordnet ist, schreiben wir $(\cdot,\cdot)_V$ und $\|\cdot\|_V$.

Manchmal ist es wichtig, dass die Skalare in einem linearen Raum V komplexe Zahlen sein dürfen. Ein solcher Raum ist dann ein Skalarproduktraum, wenn es ein auf $V \times V$ definiertes Funktional (v,w) gibt, das in der ersten Variable linear und hermitesch ist, d. h. wenn $(w,v) = \overline{(v,w)}$ gilt. Die Norm ist in diesem Falle wiederum durch (A.1) definiert und V ist ein komplexer Hilbert-Raum, wenn er bezüglich dieser Norm vollständig ist. Der Kürze wegen betrachten wir im Folgenden in der Regel den Fall reellwertiger Skalare.

Allgemeiner gesagt, ist eine *Norm* in einem linearen Raum V eine Funktion $\|\cdot\| : V \to \mathbf{R}_+$, für die

$$\begin{aligned} \|v\| &> 0 && \forall v \in V,\ v \neq 0, \\ \|\lambda v\| &= |\lambda|\,\|v\| && \forall \lambda \in \mathbf{R} \text{ (or } \mathbf{C}),\ v \in V, \\ \|v + w\| &\leq \|v\| + \|w\| && \forall v, w \in V \end{aligned}$$

gilt. Eine Funktion $|\cdot|$ wird als *Halbnorm* bezeichnet, wenn diese Bedingungen mit der Einschränkung erfüllt sind, dass die erste Zeile durch $|v| \geq 0\ \forall v \in V$ ersetzt wird, d. h. wenn sie nur positiv semidefinit ist und folglich für ein $v \neq 0$ verschwinden kann. Ein linearer Raum mit einer Norm wird als *normierter linearer Raum* bezeichnet. Wie wir bereits gesehen haben, ist ein Skalarproduktraum ein normierter linearer Raum, aber nicht alle normierten linearen Räume sind Skalarprodukträume. Ein vollständiger normierter Raum wird als *Banach-Raum* bezeichnet.

Es sei V ein Hilbert-Raum und $V_0 \subset V$ ein linearer Teilraum. Ein solcher Teilraum V_0 wird als *abgeschlossen* bezeichnet, wenn er alle Grenzwerte von Folgen in V_0 enthält, d. h. wenn aus $\{v_j\}_{j=1}^{\infty} \subset V_0$ und $v_j \to v$ für $j \to \infty$ die Beziehung $v \in V_0$ folgt. Ein solcher Teilraum V_0 ist selbst ein Hilbert-Raum mit dem gleichen Skalarprodukt wie V.

Es sei V_0 ein abgeschlossener Teilraum von V. Dann kann jedes $v \in V$ eindeutig als $v = v_0 + w$ geschrieben werden, wobei $v_0 \in V_0$ gilt und w orthogonal zu V_0 ist. Das Element v_0 kann als das eindeutige Element in V_0 beschrieben werden, das den geringsten Abstand zu v besitzt, es gilt also

$$\|v - v_0\| = \min_{u \in V_0} \|v - u\|. \tag{A.4}$$

Dies wird als *Projektionstheorem* bezeichnet und ist ein grundlegendes Resultat der Theorie der Hilbert-Räume. Das Element v_0 heißt *orthogonale Projektion* von v auf V_0 und wird auch mit $P_{V_0} v$ bezeichnet. Eine nützliche Konsequenz des Projektionstheorems besteht darin, dass der abgeschlossene lineare Teilraum V_0, falls er nicht mit dem gesamten Raum V identisch ist, einen Normalenvektor besitzt. Das heißt, es gibt einen von null verschiedenen Vektor $w \in V$, der orthogonal zu V_0 ist.

Die beiden Normen $\|\cdot\|_a$ und $\|\cdot\|_b$ werden als *äquivalent* in V bezeichnet, wenn es positive Konstanten c und C gibt, für die

$$c\|v\|_b \leq \|v\|_a \leq C\|v\|_b \quad \forall v \in V \tag{A.5}$$

gilt.

Seien V, W zwei Hilbert-Räume. Ein linearer Operator $B : V \to W$ heißt *beschränkt*, wenn eine Konstante C existiert, für die

$$\|Bv\|_W \leq C\|v\|_V \quad \forall v \in V \tag{A.6}$$

erfüllt ist. Die Norm eines beschränkten linearen Operators B ist

$$\|B\| = \sup_{v \in V} \frac{\|Bv\|_W}{\|v\|_V}. \tag{A.7}$$

Folglich gilt

$$\|Bv\|_W \leq \|B\| \, \|v\|_V \quad \forall v \in V$$

und per Definition ist $\|B\|$ die kleinste Konstante C, für die (A.6) gilt.

Beachten Sie, dass ein beschränkter linearer Operator $B : V \to W$ stetig ist. Tatsächlich gilt im Falle $v_j \to v$ in V die Beziehung $Bv_j \to Bv$ in W für $j \to \infty$, weil

$$\|Bv_j - Bv\|_W = \|B(v_j - v)\|_W \leq \|B\| \, \|v_j - v\| \to 0 \text{ für } j \to \infty$$

ist. Umgekehrt kann man auch zeigen, dass ein stetiger linearer Operator beschränkt ist.

Im Spezialfall $W = \mathbf{R}$ reduziert sich die Definition eines Operators auf die eines linearen Funktionals. Die Menge aller beschränkten linearen Funktionale auf V heißt *dualer Raum* von V und wird mit V^* bezeichnet. Wegen (A.7) ist die Norm in V^*

$$\|L\|_{V^*} = \sup_{v \in V} \frac{|L(v)|}{\|v\|_V}. \tag{A.8}$$

Beachten Sie, dass V^* selbst ein linearer Raum ist, wenn wir $(\lambda L + \mu M)(v) = \lambda L(v) + \mu M(v)$ für $L, M \in V^*$, $\lambda, \mu \in \mathbf{R}$ definieren. Mit der durch (A.8) definierten Norm ist V^* ein normierter linearer Raum und man kann zeigen, dass V^* vollständig und somit selbst auch ein Banach-Raum ist.

Analog dazu sagen wir, dass die Bilinearform $a(\cdot, \cdot)$ auf V *beschränkt* ist, wenn es eine Konstante M gibt, für die

$$|a(w, v)| \leq M \|w\| \, \|v\| \quad \forall w, v \in V \tag{A.9}$$

gilt.

Das folgende Theorem gibt eine wichtige Eigenschaft von Hilbert-Räumen an.

Theorem A.1. (Rieszscher Darstellungssatz.) *Sei V ein Hilbert-Raum mit dem Skalarprodukt $(\cdot, \cdot)$. Für jedes beschränkte lineare Funktional L auf V gibt es ein eindeutiges $u \in V$, das die Gleichung*

$$L(v) = (v, u) \quad \forall v \in V$$

erfüllt. Darüber hinaus gilt

$$\|L\|_{V^*} = \|u\|_V. \tag{A.10}$$

Beweis. Die Eindeutigkeit ist offensichtlich, da aus $(v, u_1) = (v, u_2)$ mit $v = u_1 - u_2$ die Gleichung $\|u_1 - u_2\|^2 = (u_1 - u_2, u_1 - u_2) = 0$ folgt. Wenn $L(v) = 0$ für alle $v \in V$ gilt, dann können wir $u = 0$ setzen. Nehmen wir nun an, dass $L(\bar{v}) \neq 0$ für ein $\bar{v} \in V$ gilt. Wir werden u als einen geeignet normierten „Normalenvektor“ an die „Hyperebene“ $V_0 = \{v \in V : L(v) = 0\}$

konstruieren, wobei es sich, wie man leicht sieht, um einen abgeschlossenen Teilraum von V handelt (siehe Problemstellung A.2). Dann ist $\bar{v} = v_0 + w$ mit $v_0 \in V_0$ und w orthogonal zu V_0 und $L(w) = L(\bar{v}) \neq 0$. Das bedeutet aber $L(v - w\,L(v)/L(w)) = 0$, sodass $(v - w\,L(v)/L(w), w) = 0$ und folglich $L(v) = (v, u),\ \forall v \in V$ gilt, wobei $u = w\,L(w)/\|w\|^2$ ist. □

Aufgrund dieses Resultates ist es natürlich, die linearen Funktionale $L \in V^*$ mit den zugehörigen $u \in V$ zu identifizieren, sodass im Falle eines Hilbert-Raumes V^* äquivalent zu V ist.

Mitunter wollen wir Gleichungen der folgenden Form lösen: Gesucht ist ein $u \in V$, für das

$$a(u, v) = L(v) \quad \forall v \in V \tag{A.11}$$

gilt. Dabei ist V ein Hilbert-Raum, L ein beschränktes lineares Funktional auf V und $a(\cdot, \cdot)$ eine symmetrische Bilinearform, die *koerzitiv* in V ist, d. h. es gilt

$$a(v, v) \geq \alpha \|v\|_V^2 \quad \forall v \in V \quad \text{mit } \alpha > 0. \tag{A.12}$$

Daraus folgt, dass $a(\cdot, \cdot)$ symmetrisch und positiv definit ist, also ein Skalarprodukt auf V. Aus dem Rieszschen Darstellungssatz folgt nun unmittelbar die Existenz einer eindeutigen Lösung $u \in V$ für $L \in V^*$.

Darüber hinaus erhalten wir durch die Wahl $v = u$ in (A.11)

$$\alpha \|u\|_V^2 \leq a(u, u) = L(u) \leq \|L\|_{V^*}\, \|u\|_V,$$

sodass sich nach Wegstreichen eines Faktors $\|u\|_V$

$$\|u\|_V \leq C \|L\|_{V^*} \quad \text{mit } C = 1/\alpha \tag{A.13}$$

ergibt. Dies ist ein Beispiel für eine *Energieabschätzung*.

Wenn $a(\cdot, \cdot)$ eine symmetrische Bilinearform ist, die in V koerzitiv und beschränkt ist, sodass (A.12) und (A.9) gelten, dann können wir eine Norm $\|\cdot\|_a$, die *Energienorm*, durch

$$\|v\|_a = a(v, v)^{1/2} \quad \text{für } v \in V$$

definieren. Wegen (A.12) und (A.9) gilt dann

$$\sqrt{\alpha} \|v\|_V \leq \|v\|_a \leq \sqrt{M} \|v\|_V \quad \forall v \in V, \tag{A.14}$$

und folglich ist die Norm $\|\cdot\|_a$ auf V äquivalent zu $\|\cdot\|_V$. Offensichtlich ist dann V bezüglich des Skalarproduktes $a(\cdot, \cdot)$ und der Norm $\|\cdot\|_a$ ebenfalls ein Hilbert-Raum.

Die Lösung von (A.11) kann auch in Form eines Minimumproblems dargestellt werden.

Theorem A.2. *Sei $a(\cdot,\cdot)$ eine symmetrische, positiv definite Bilinearform und L eine beschränkte Linearform auf dem Hilbert-Raum V. In diesem Fall erfüllt $u \in V$ genau dann* (A.11)*, wenn*

$$F(u) \le F(v) \quad \forall v \in V \quad \text{mit } F(v) = \tfrac{1}{2}a(v,v) - L(v) \tag{A.15}$$

ist.

Beweis. Nehmen wir zunächst an, dass u (A.11) erfüllt. Es sei $v \in V$ beliebig und wir definieren $w = v - u \in V$. Dann gilt $v = u + w$ und

$$\begin{aligned} F(v) &= \tfrac{1}{2}a(u+w,u+w) - L(u+w) \\ &= \tfrac{1}{2}a(u,u) - L(u) + a(u,w) - L(w) + \tfrac{1}{2}a(w,w) \\ &= F(u) + \tfrac{1}{2}a(w,w), \end{aligned}$$

wobei wir (A.11) und die Symmetrie von $a(\cdot,\cdot)$ benutzt haben. Da a positiv definit ist, beweist dies (A.15).

Wenn umgekehrt (A.15) gilt, dann ist für ein gegebenes $v \in V$

$$g(t) := F(u+tv) \ge F(u) = g(0) \quad \forall\, t \in \mathbf{R},$$

sodass $g(t)$ ein Minimum an der Stelle $t = 0$ besitzt. Aber $g(t)$ ist das quadratische Polynom

$$\begin{aligned} g(t) &= \tfrac{1}{2}a(u+tv,u+tv) - L(u+tv) \\ &= \tfrac{1}{2}a(u,u) - L(u) + t\big(a(u,v) - L(v)\big) + \tfrac{1}{2}t^2 a(v,v) \end{aligned}$$

und folglich gilt $0 = g'(0) = a(u,v) - L(v)$, also (A.11). □

Somit erfüllt $u \in V$ (A.11) genau dann, wenn u das Energiefunktional F minimiert. Die Methode, das Minimierungsproblem durch Variation des Argumentes des Funktionals F um einen gegebenen Vektor u zu untersuchen, wird als Variationsmethode bezeichnet. Die Gleichung (A.11) nennt man *Variationsgleichung* für F.

Das folgende Theorem, das als *Lax-Milgram-Lemma* bekannt ist, verallgemeinert den Rieszschen Darstellungssatz auf nichtsymmetrische Bilinearformen.

Theorem A.3. *Wenn die Bilinearform $a(\cdot,\cdot)$ im Hilbert-Raum V beschränkt und koerzitiv und L eine beschränkte Linearform in V ist, dann existiert ein eindeutiger Vektor $u \in V$, für den* (A.11) *erfüllt ist.*

Beweis. Mit dem Skalarprodukt $(\cdot,\cdot)$ in V existiert nach dem Rieszschen Darstellungssatz ein eindeutiges $b \in V$, für das

$$L(v) = (b,v) \quad \forall v \in V$$

gilt. Darüber hinaus ist $a(u,\cdot)$ für jedes $u \in V$ offensichtlich auch ein beschränktes lineares Funktional auf V, sodass ein eindeutiges $A(u) \in V$ mit

$$a(u,v) = (A(u),v) \quad \forall v \in V$$

existiert. Es lässt sich leicht überprüfen, dass $A(u)$ linear und beschränkt von u abhängt, sodass $Au = A(u)$ den Operator $A : V \to V$ als beschränkten linearen Operator definiert. Die Gleichung (A.11) ist deshalb zu $Au = b$ äquivalent. Um den Beweis des Theorems zu vervollständigen, werden wir zeigen, dass diese Gleichung für jedes b eine eindeutige Lösung $u = A^{-1}b$ besitzt.

Unter Verwendung der Koerzitivität gilt

$$\alpha\|v\|_V^2 \le a(v,v) = (Av,v) \le \|Av\|_V\|v\|_V,$$

sodass

$$\alpha\|v\|_V \le \|Av\|_V \quad \forall v \in V \tag{A.16}$$

gilt. Daraus ergibt sich die Eindeutigkeit, weil aus $Av = 0$ die Gleichung $v = 0$ folgt. Dies kann auch durch die Feststellung ausgedrückt werden, dass der Nullraum $N(A) = \{v \in V : Av = 0\} = 0$ oder A *injektiv* ist.

Zeigt man die Existenz einer Lösung u für jedes $b \in V$, bedeutet dies, dass jedes $b \in V$ zum Bildbereich $R(A) = \{w \in V : w = Av \text{ für ein } v \in V\}$ gehört, d. h. es gilt $R(A) = V$ beziehungsweise A ist *surjektiv*. Um uns davon zu überzeugen, stellen wir zunächst fest, dass $R(A)$ ein abgeschlossener linearer Teilraum von V ist. Um die Abgeschlossenheit von $R(A)$ zu zeigen, nehmen wir an, dass $Av_j \to w$ in V für $j \to \infty$ gilt. Wegen (A.16) führt dies für $i,j \to \infty$ auf $\|v_j - v_i\|_V \le \alpha^{-1}\|Av_j - Av_i\|_V \to 0$. Folglich gilt $v_j \to v \in V$ für $j \to \infty$ und wegen der Stetigkeit von A auch $Av_j \to Av = w$. Deshalb gilt $w \in R(A)$, und $R(A)$ ist abgeschlossen.

Nehmen wir nun $R(A) \neq V$ an. Dann existiert nach dem Projektionstheorem ein $w \neq 0$, das orthogonal zu $R(A)$ ist. Wegen der Orthogonalität gilt aber auch

$$\alpha\|w\|_V^2 \le a(w,w) = (Aw,w) = 0,$$

sodass $w = 0$ sein muss, was ein Widerspruch ist. Folglich muss $R(A) = V$ sein. Dies vervollständigt den Beweis der Existenz einer eindeutigen Lösung für jedes $b \in V$. Die Energieabschätzung wird wie zuvor bewiesen. □

Im unsymmetrischen Fall gibt es keine Darstellung der Lösung in Form eines Energie-Minimierungsproblems.

Abschließen wollen wir mit einer Bemerkung über lineare Gleichungen in endlichdimensionalen Räumen. Es sei $V = \mathbf{R}^N$ und wir betrachten eine lineare Gleichung in V, die in Matrixform als

$$Au = b$$

geschrieben werden kann, wobei A eine $N \times N$-Matrix ist und u, b N-komponentige Vektoren sind. Es ist bekannt, dass diese Gleichung für jedes $b \in V$ eine eindeutige Lösung $u = A^{-1}b$ besitzt, wenn die Matrix A regulär ist, d. h. wenn die Determinante $\det(A) \neq 0$ ist. Im Falle $\det(A) = 0$ besitzt die homogene Gleichung $Au = 0$ nichttriviale Lösungen $u \neq 0$ und es gilt $R(A) \neq V$, sodass die inhomogene Gleichung nicht immer lösbar ist. Folglich ist für alle $b \in V$ weder die Eindeutigkeit noch die Existenz gesichert. Eindeutigkeit liegt insbesondere nur dann vor, wenn $\det(A) \neq 0$ ist, woraus dann auch die Existenz der Lösung folgt. Mitunter lässt sich die Eindeutigkeit leicht zeigen, wodurch wir auch gleichzeitig die Existenz der Lösung erhalten.

A.2 Funktionenräume

Der Raum $\mathcal{C}^k$

Für $M \subset \mathbf{R}^d$ bezeichnen wir mit $\mathcal{C}(M)$ den linearen Raum der stetigen Funktionen auf M. Den Teilraum aller beschränkten Funktionen können wir in einen normierten, linearen Raum überführen, indem wir

$$\|v\|_{\mathcal{C}(M)} = \sup_{x \in M} |v(x)| \tag{A.17}$$

setzen. Wenn M eine beschränkte und abgeschlossene Menge, d. h. eine kompakte Menge ist, dann wird das Supremum in (A.17) angenommen und wir können

$$\|v\|_{\mathcal{C}(M)} = \max_{x \in M} |v(x)|$$

schreiben. Die Norm (A.17) wird deshalb als *Maximumnorm* bezeichnet. Beachten Sie, dass die Konvergenz in $\mathcal{C}(M)$, also

$$\|v_i - v\|_{\mathcal{C}(M)} = \sup_{x \in M} |v_i(x) - v(x)| \to 0,$$

gleich der gleichmäßigen Konvergenz in M ist. Wir erinnern uns: Wenn eine Folge stetiger Funktionen in M gleichmäßig konvergent ist, dann ist die Grenzfunktion stetig. Unter Verwendung dieser Tatsache ist der Beweis nicht schwierig, dass $\mathcal{C}(M)$ ein vollständiger normierter Raum, d. h. ein Banach-Raum, ist. $\mathcal{C}(M)$ ist kein Hilbert-Raum, weil die Maximumnorm nicht mit einem Skalarprodukt wie in (A.1) verknüpft ist.

Sei nun $\Omega \subset \mathbf{R}^d$ ein *Gebiet*, d. h. eine zusammenhängende offene Menge. Mit $\mathcal{C}^k(\Omega)$ bezeichnen wir für jede ganze Zahl $k \geq 0$ den linearen Raum aller in Ω k-mal stetig differenzierbaren Funktionen v. $\mathcal{C}^k(\bar{\Omega})$ bezeichnet den Raum der Funktionen in $\mathcal{C}^k(\Omega)$, für die $D^\alpha v \in \mathcal{C}(\bar{\Omega})$ für alle $|\alpha| \leq k$ gilt. Dabei ist $D^\alpha v$ die in (1.8) definierte partielle Ableitung von v. Wenn Ω beschränkt ist, dann ist der zuletzt genannte Raum ein Banach-Raum bezüglich der Norm

$$\|v\|_{\mathcal{C}^k(\bar{\Omega})} = \max_{|\alpha| \leq k} \|D^\alpha v\|_{\mathcal{C}(\bar{\Omega})}.$$

Für Funktionen aus $\mathcal{C}^k(\bar{\Omega})$, $k \geq 1$, verwenden wir gelegentlich auch die Halbnorm

$$|v|_{\mathcal{C}^k(\bar{\Omega})} = \max_{|\alpha|=k} \|D^\alpha v\|_{\mathcal{C}(\bar{\Omega})},$$

die lediglich die Ableitungen höchster Ordnung enthält. Eine Funktion besitzt einen kompakten Träger in Ω, wenn sie außerhalb einer kompakten Teilmenge von Ω verschwindet. Den Raum der Funktionen aus $\mathcal{C}^k(\bar{\Omega})$ mit kompaktem Träger in Ω bezeichnen wir mit $\mathcal{C}_0^k(\Omega)$. Insbesondere verschwinden solche Funktionen in der Nähe des Randes Γ und im Falle sehr großer x-Werte, wenn Ω unbeschränkt ist.

Wir sagen, dass eine Funktion *glatt* ist, wenn sie in Abhängigkeit vom Kontext der vorliegenden Aufgabe hinreichend viele stetige Ableitungen besitzt.

Wenn keine Gefahr für Missverständnisse besteht, lassen wir den Definitionsbereich der Funktionen in der Notation des Raumes weg und schreiben beispielsweise $\mathcal{C}$ für $\mathcal{C}(\bar{\Omega})$ und $\|\cdot\|_{\mathcal{C}^k}$ für $\|\cdot\|_{\mathcal{C}^k(\bar{\Omega})}$. Gleiches gilt für die weiter unten eingeführten Räume.

Integrabilität, L_p-Räume

Es sei Ω ein Gebiet in $\mathbf{R}^d$. Wir müssen mit Integralen über Funktionen $v = v(x)$ in Ω arbeiten, die allgemeiner als die in $\mathcal{C}(\bar{\Omega})$ sind. Für eine nichtnegative Funktion kann man das sogenannte *Lebesgue-Integral*

$$I_\Omega(v) = \int_\Omega v(x)\,\mathrm{d}x$$

definieren, das endlich oder unendlich sein kann und für $v \in \mathcal{C}(\bar{\Omega})$ mit dem Riemann-Integral übereinstimmt. Die von uns betrachteten Funktionen setzen wir als messbar voraus; wir werden nicht tiefer auf die theoretischen Details eingehen, sondern merken einfach an, dass alle Funktionen, die uns in diesem Buch begegnen, diese Forderung erfüllen. Eine nichtnegative Funktion v wird als integrierbar bezeichnet, wenn $I_\Omega(v) < \infty$ gilt, analog ist eine allgemeine reell- oder komplexwertige Funktion v integrierbar, wenn $|v|$ integrierbar ist. Eine Teilmenge Ω_0 von Ω wird als Nullmenge oder eine Menge vom Maß null bezeichnet, wenn ihr Volumen $|\Omega|$ gleich null ist. Zwei Funktionen, die abgesehen von einer Nullmenge gleich sind, besitzen folglich das gleiche Lebesgue-Integral und werden als gleich fast überall bezeichnet. Wenn in einem beschränkten Gebiet Ω also $v_1(x) = 1$ gilt und wenn $v_2(x) = 1$ in Ω außer an der Stelle $x_0 \in \Omega$ mit $v_2(x_0) = 2$ ist, dann gilt $I_\Omega(v_1) = I_\Omega(v_2) = |\Omega|$. Insbesondere können wir aus der Tatsache, dass eine Funktion integrierbar ist, keinen Schluss über ihre Werte am Punkt $x_0 \in \Omega$ ziehen, d. h. die Punktwerte sind nicht wohldefiniert. Weil der Rand Γ von Ω eine Nullmenge ist, gilt für jedes v auch $I_{\bar{\Omega}}(v) = I_\Omega(v)$.

Wir definieren nun

$$\|v\|_{L_p} = \|v\|_{L_p(\Omega)} = \begin{cases} \left(\int_\Omega |v(x)|^p \,\mathrm{d}x\right)^{1/p} & \text{für } 1 \le p < \infty, \\ \operatorname{ess\,sup}_\Omega |v(x)| & \text{für } p = \infty \end{cases}$$

und sagen, dass $v \in L_p = L_p(\Omega)$ gilt, wenn $\|v\|_{L_p} < \infty$ ist. In dieser Stelle meinen wir mit ess sup das *wesentliche Supremum*, das Werte auf Nullmengen außer Acht lässt, sodass beispielsweise $\|v_2\|_{L_\infty} = 1$ für die oben erwähnte Funktion v_2 gilt, auch wenn $\sup_\Omega v_2 = 2$ ist. Man kann zeigen, dass L_p ein vollständiger normierter Raum ist, d. h. ein Banach-Raum. Die Dreiecksungleichung in L_p wird als Minkowski-Ungleichung bezeichnet. Offensichtlich gehört für beschränktes Ω jedes $v \in \mathcal{C}$ zu L_p mit $1 \le p \le \infty$. Es gilt

$$\|v\|_{L_p} \le C\|v\|_{\mathcal{C}} \quad \text{mit } C = |\Omega|^{1/p} \quad \text{für } 1 \le p < \infty \quad \text{und} \quad \|v\|_{L_\infty} = \|v\|_{\mathcal{C}},$$

obwohl L_p auch unstetige Funktionen enthält. Darüber hinaus ist es nicht schwierig zu zeigen, dass $\mathcal{C}(\bar{\Omega})$ bezüglich der L_p-Norm für $1 \le p < \infty$ unvollständig ist. Man kann sich davon überzeugen, indem man eine Folge $\{v_i\}_{i=1}^\infty \subset \mathcal{C}(\bar{\Omega})$ konstruiert, die bezüglich der L_p-Norm eine Cauchy-Folge ist, d. h. für die $\|v_i - v_j\|_{L_p} \to 0$ gilt, deren Grenzwert $v = \lim_{i\to\infty} v_i$ jedoch unstetig ist. Der Raum $\mathcal{C}(\bar{\Omega})$ ist aber ein *dichter Teilraum* des $L_p(\Omega)$ mit $1 \le p < \infty$, wenn Γ hinreichend glatt ist. Dies bedeutet, dass für jedes $v \in L_p$ eine Folge $\{v_i\}_{i=1}^\infty \subset \mathcal{C}$ existiert, für die $\|v_i - v\|_{L_p} \to 0$ für $i \to \infty$ gilt. Mit anderen Worten, jede Funktion $v \in L_p$ kann in der L_p-Norm beliebig gut durch Funktionen in $\mathcal{C}$ approximiert werden (eigentlich genauer durch Funktionen in $\mathcal{C}_0^k$ für alle k). Im Gegensatz dazu ist $\mathcal{C}$ nicht dicht in L_∞, da eine unstetige Funktion durch eine stetige Funktion nicht gleichmäßig gut approximiert werden kann.

Der Fall L_2 ist für uns von besonderem Interesse. Dieser Raum ist ein Skalarproduktraum und folglich bezüglich des Skalarproduktes

$$(v, w) = \int_\Omega v(x)w(x) \,\mathrm{d}x \tag{A.18}$$

ein Hilbert-Raum. Für komplexwertige Funktionen verwendet man im Integranden das konjugiert Komplexe von $w(x)$.

Sobolev-Räume

Wir werden nun einige spezielle Hilbert-Räume einführen, die üblicherweise bei der Analyse partieller Differentialgleichungen benutzt werden. Diese Räume bestehen aus Funktionen, die mit ihren partiellen Ableitungen bis zu einer bestimmten Ordnung quadratintegrabel sind. Für die Definition müssen wir zunächst den Begriff der partiellen Ableitung verallgemeinern.

Sei Ω ein Gebiet in $\mathbf{R}^d$ und sei zunächst $v \in \mathcal{C}^1(\bar{\Omega})$. Partielle Integration führt auf

$$\int_\Omega \frac{\partial v}{\partial x_i}\phi \,\mathrm{d}x = -\int_\Omega v\frac{\partial \phi}{\partial x_i}\,\mathrm{d}x \quad \forall \phi \in \mathcal{C}_0^1 = \mathcal{C}_0^1(\Omega).$$

Für $v \in L_2 = L_2(\Omega)$ existiert $\partial v/\partial x_i$ im klassischen Sinne nicht notwendigerweise, wir können $\partial v/\partial x_i$ allerdings als lineares Funktional

$$L(\phi) = \frac{\partial v}{\partial x_i}(\phi) = -\int_\Omega v\frac{\partial \phi}{\partial x_i}\,\mathrm{d}x \quad \forall \phi \in \mathcal{C}_0^1 \tag{A.19}$$

definieren. Dieses Funktional wird als *verallgemeinerte* oder *schwache Ableitung* von v bezeichnet. Wenn L in L_2 beschränkt ist, folgt aus dem Rieszschen Darstellungssatz die Existenz einer eindeutigen Funktion $w \in L_2$, mit der Eigenschaft, dass $L(\phi) = (w, \phi)$ für alle $\phi \in L_2$ gilt. Es ist also

$$-\int_\Omega v\frac{\partial \phi}{\partial x_i}\,\mathrm{d}x = \int_\Omega w\phi \,\mathrm{d}x \quad \forall \phi \in \mathcal{C}_0^1.$$

Wir sagen dann, dass die schwache Ableitung zu L_2 gehört und schreiben $\partial v/\partial x_i = w$. In diesem Fall gilt daher

$$\int_\Omega \frac{\partial v}{\partial x_i}\phi \,\mathrm{d}x = -\int_\Omega v\frac{\partial \phi}{\partial x_i}\,\mathrm{d}x \quad \forall \phi \in \mathcal{C}_0^1. \tag{A.20}$$

Insbesondere stimmt für $v \in \mathcal{C}^1(\bar{\Omega})$ die verallgemeinerte Ableitung $\partial v/\partial x_i$ mit der klassischen Ableitung $\partial v/\partial x_i$ überein.

In ähnlicher Weise definieren wir die schwache partielle Ableitung $D^\alpha v$ als lineares Funktional

$$D^\alpha v(\phi) = (-1)^{|\alpha|}\int_\Omega v\, D^\alpha \phi \,\mathrm{d}x \quad \forall \phi \in \mathcal{C}_0^{|\alpha|}, \tag{A.21}$$

wobei $D^\alpha v$ die in (1.8) definierte partielle Ableitung von v ist. Ist dieses Funktional im L_2 beschränkt, dann sichert der Rieszsche Darstellungssatz die Existenz einer eindeutigen Funktion $D^\alpha v$ in L_2, für die

$$(D^\alpha v, \phi) = (-1)^{|\alpha|}(v, D^\alpha \phi) \quad \forall \phi \in \mathcal{C}_0^{|\alpha|}$$

gilt. Zur weiteren Diskussion verallgemeinerter Funktionen verweisen wir auf Problemstellung A.9.

Wir definieren nun $H^k = H^k(\Omega)$ mit $k \geq 0$ als Raum aller Funktionen, deren schwache partielle Ableitungen der Ordnung $\leq k$ zu L_2 gehören, d. h.

$$H^k = H^k(\Omega) = \big\{v \in L_2 : D^\alpha v \in L_2 \text{ für } |\alpha| \leq k\big\}.$$

Wir versehen diesen Raum mit dem Skalarprodukt

$$(v, w)_k = (v, w)_{H^k} = \sum_{|\alpha|\leq k}\int_\Omega D^\alpha v D^\alpha w \,\mathrm{d}x$$

und der zugehörigen Norm

$$\|v\|_k = \|v\|_{H^k} = (v,v)_{H^k}^{1/2} = \Big(\sum_{|\alpha|\le k} \int_\Omega (D^\alpha v)^2 \,\mathrm{d}x \Big)^{1/2}.$$

Insbesondere gilt $\|v\|_0 = \|v\|_{L_2}$. In diesem Fall lassen wir den Index 0 gewöhnlich weg und schreiben $\|v\|$. Außerdem ist

$$\|v\|_1 = \Big(\int_\Omega \Big\{ v^2 + \sum_{j=1}^d \big(\frac{\partial v}{\partial x_j}\big)^2 \Big\} \,\mathrm{d}x \Big)^{1/2} = \Big(\|v\|^2 + \|\nabla v\|^2 \Big)^{1/2}$$

und

$$\|v\|_2 = \Big(\int_\Omega \Big\{ v^2 + \sum_{j=1}^d \big(\frac{\partial v}{\partial x_j}\big)^2 + \sum_{i,j=1}^d \big(\frac{\partial^2 v}{\partial x_i \partial x_j}\big)^2 \Big\} \,\mathrm{d}x \Big)^{1/2}.$$

Wir verwenden manchmal auch die Halbnorm

$$|v|_k = |v|_{H^k} = \Big(\sum_{|\alpha|=k} \int_\Omega (D^\alpha v)^2 \,\mathrm{d}x \Big)^{1/2} \tag{A.22}$$

mit $k \ge 1$. Beachten Sie, dass die Halbnorm für konstante Funktionen verschwindet. Unter Verwendung der Vollständigkeit des L_2, kann man zeigen, dass H^k vollständig und somit ein Hilbert-Raum ist (siehe Problemstellung A.4). Der Raum H^k ist ein Beispiel einer allgemeineren Klasse von Funktionenräumen, den Sobolev-Räumen.

Man kann zeigen, dass $\mathcal{C}^l = \mathcal{C}^l(\bar\Omega)$ für jedes $l \ge k$ und hinreichend glattes Γ dicht in H^k ist. Diese Eigenschaft ist nützlich, weil sie es uns erlaubt, bestimmte Resultate für H^k zu erhalten, indem wir den möglicherweise technisch einfacheren Beweis für Funktionen in $\mathcal{C}^k$ führen und das Resultat mithilfe der Dichtheit auf alle $v \in H^k$ verallgemeinern (vgl. Beweis von Theorems A.4).

Analog dazu bezeichnen wir den durch die Norm

$$\|v\|_{W_p^k} = \Big(\int_\Omega \sum_{|\alpha|\le k} |D^\alpha v|^p \,\mathrm{d}x \Big)^{1/p} \quad \text{für } 1 \le p < \infty$$

definierten normierten Raum mit $W_p^k = W_p^k(\Omega)$. Dieser Raum ist tatsächlich vollständig und somit ein Banach-Raum. Für $p = 2$ gilt $W_2^k = H^k$. Wiederum gilt $\|v\|_{W_\infty^k} = \|v\|_{\mathcal{C}^k}$ für $v \in \mathcal{C}^k$.

Spurtheoreme

Im Falle $v \in \mathcal{C}(\bar\Omega)$ ist $v(x)$ für $x \in \Gamma$, d. h. auf dem Rand von Ω, wohldefiniert. Die *Spur* γv eines solchen v auf Γ ist die Restriktion von v auf Γ, d. h. es gilt

$$(\gamma v)(x) = v(x) \quad \text{für } x \in \Gamma. \tag{A.23}$$

Es sei daran erinnert, dass die Spur von $v \in L_2(\Omega)$ nicht wohldefiniert ist, da Γ eine Nullmenge ist.

Nehmen wir nun $v \in H^1(\Omega)$ an. Ist es dann möglich, v eindeutig auf dem Rand Γ, d. h. dessen Spur γv auf Γ zu definieren? (Man kann zeigen, dass Funktionen in $H^1(\Omega)$ nicht notwendigerweise stetig sind, vgl. Theorem A.5 und Problemstellungen A.6, A.7.) Diese Frage kann präzisiert werden, indem man nach der Möglichkeit fragt, eine Norm $\|\cdot\|_{(\Gamma)}$ für Funktionen auf Γ und eine Konstante C zu finden, die

$$\|\gamma v\|_{(\Gamma)} \le C\|v\|_1 \quad \forall v \in \mathcal{C}^1(\bar\Omega) \tag{A.24}$$

erfüllt. Eine Ungleichung dieser Form wird als eine *Spurungleichung* bezeichnet. Wenn (A.24) gilt, dann ist es mithilfe eines Dichtheitsargumentes (siehe unten) möglich, den Definitionsbereich des Spuroperators γ von $\mathcal{C}^1(\bar\Omega)$ auf $H^1(\Omega)$ zu erweitern. Dabei gilt (A.24) auch für alle $v \in H^1(\Omega)$. Der Funktionenraum, zu dem γv gehört, wird durch die Norm $\|\cdot\|_{(\Gamma)}$ in (A.24) definiert.

Wir weisen darauf hin, dass der Rand Γ in der obigen Diskussion durch eine andere Teilmenge von Ω mit Dimension kleiner d ersetzt werden könnte.

Um mit den Spurtheoremen fortfahren zu können, betrachten wir zunächst einen eindimensionalen Fall, bei dem Γ einem einzelnen Punkt entspricht.

Lemma A.1. *Sei $\Omega = (0,1)$. Dann gibt es eine Konstante C mit*

$$|v(x)| \le C\|v\|_1 \quad \forall x \in \bar\Omega,\ \forall v \in \mathcal{C}^1(\bar\Omega).$$

Beweis. Für $x, y \in \Omega$ gilt $v(x) = v(y) + \int_y^x v'(s)\,\mathrm{d}s$ und somit wegen der Cauchy-Schwarz-Ungleichung

$$|v(x)| \le |v(y)| + \int_0^1 |v'(s)|\,\mathrm{d}s \le |v(y)| + \|v'\|.$$

Durch Quadrieren beider Seiten und Integration über y erhalten wir

$$v(x)^2 \le 2\big(\|v\|^2 + \|v'\|^2\big) = 2\|v\|_1^2, \tag{A.25}$$

was die gewünschte Abschätzung beweist. □

Wir beweisen nun ein einfaches Spurtheorem. Mit $L_2(\Gamma)$ bezeichnen wir den Hilbert-Raum aller auf Γ quadratintegrablen Funktionen mit der Norm

$$\|w\|_{L_2(\Gamma)} = \Big(\int_\Gamma w^2\,\mathrm{d}s\Big)^{1/2}.$$

Theorem A.4. (Spurtheorem.) *Sei Ω ein beschränktes Gebiet in $\mathbf{R}^d$ ($d \ge 2$) mit glattem oder polynomialem Rand Γ. Dann kann der Spuroperator $\gamma : \mathcal{C}^1(\bar\Omega) \to \mathcal{C}(\Gamma)$ auf $\gamma : H^1(\Omega) \to L_2(\Gamma)$ erweitert werden, was die Spur $\gamma v \in L_2(\Gamma)$ für $v \in H^1(\Omega)$ definiert. Darüber hinaus gibt es eine Konstante $C = C(\Omega)$ mit*

$$\|\gamma v\|_{L_2(\Gamma)} \le C\|v\|_1 \quad \forall v \in H^1(\Omega). \tag{A.26}$$

Beweis. Wir beweisen zunächst die Spurungleichung (A.26) für Funktionen $v \in \mathcal{C}^1(\bar{\Omega})$. Der Einfachheit halber betrachten wir nur den Fall des Einheitsquadrates $\Omega = (0,1) \times (0,1)$ in $\mathbf{R}^2$. Der Beweis des allgemeinen Falls erfolgt analog. Für $y = (y_1, y_2) \in \Omega$ gilt wegen (A.25)

$$v(0,y_2)^2 \le 2\Big(\int_0^1 v(y_1,y_2)^2 \,\mathrm{d}y_1 + \int_0^1 \Big(\frac{\partial v}{\partial x_1}(s,y_2) \Big)^2 \,\mathrm{d}s \Big)$$

und nach Integration über y_2

$$\int_0^1 v(0,y_2)^2 \,\mathrm{d}y_2 \le 2\big(\|v\|^2 + \|\nabla v\|^2 \big) = 2\|v\|_1^2.$$

Die analogen Abschätzungen für die verbleibenden Teile von Γ vervollständigen den Beweis von (A.26) für $v \in \mathcal{C}^1$.

Sei nun $v \in H^1(\Omega)$. Da $\mathcal{C}^1$ in H^1 dicht ist, gibt es eine Folge $\{v_i\}_{i=1}^\infty \subset \mathcal{C}^1$, für die $\|v - v_i\|_1 \to 0$ gilt. Diese Folge ist dann eine Cauchy-Folge in H^1, d. h. es gilt $\|v_i - v_j\|_1 \to 0$ für $i,j \to \infty$. Wenden wir (A.26) auf $v_i - v_j \in \mathcal{C}^1$ an, erhalten wir

$$\|\gamma v_i - \gamma v_j\|_{L_2(\Gamma)} \le C\|v_i - v_j\|_1 \to 0 \quad \text{für } i,j \to \infty,$$

d. h. $\{\gamma v_i\}_{i=1}^\infty$ ist eine Cauchy-Folge in $L_2(\Gamma)$. Somit existiert ein $w \in L_2(\Gamma)$, für das $\gamma v_i \to w$ in $L_2(\Gamma)$ für $i \to \infty$ ist. Wir definieren $\gamma v = w$. Es ist leicht zu zeigen, dass dann für $v \in H^1$ Ungleichung (A.26) gilt. Dies setzt γ als einen beschränkten, linearen Operator $\gamma : H^1(\Omega) \to L_2(\Gamma)$ fort. Weil $\mathcal{C}^1$ in H^1 dicht ist, gibt es nur eine solche Fortsetzung (Beweisen Sie dies!). Insbesondere ist γ unabhängig von der Wahl der Folge $\{v_i\}$. □

Die Konstante in Theorem A.4 hängt von der Größe und der Form des Gebietes Ω ab. Mitunter ist es wichtig, genauere Informationen über diese Abhängigkeit zu haben. In Problemstellung A.15 nehmen wir an, dass die Form fest (ein Quadrat) ist und untersuchen die Abhängigkeit der Konstante von der Größe von Ω.

Das folgende Resultat ist in gewissem Sinne von ähnlicher Natur und ein Spezialfall der bekannten und wichtigen Sobolev-Ungleichung:

Theorem A.5. *Sei Ω ein beschränktes Gebiet in $\mathbf{R}^d$ mit glattem oder polynomialem Rand und sei $k > d/2$. Dann ist $H^k(\Omega) \subset \mathcal{C}(\bar{\Omega})$ und es existiert eine Konstante $C = C(\Omega)$ mit*

$$\|v\|_{\mathcal{C}} \le C\|v\|_k \quad \forall v \in H^k(\Omega). \tag{A.27}$$

Wie im Falle des Spurtheorems reicht es aus, Gleichung (A.27) für ein glattes v zu beweisen. Der Spezialfall $d = k = 1$ ist in Lemma A.1 angegeben und Problemstellung A.13 betrachtet den Fall $\Omega = (0,1) \times (0,1)$. Der allgemeine Fall ist komplizierter. Wie in den Problemstellungen A.6 und A.7 gezeigt, ist

eine Funktion in $H^1(\Omega)$ mit $\Omega \subset \mathbf{R}^d$ nicht notwendigerweise stetig, wenn $d \geq 2$ ist.

Wenn wir die Sobolev-Ungleichung auf die Ableitungen von v anwenden, erhalten wir

$$\|v\|_{\mathcal{C}^\ell} \leq C\|v\|_k \quad \forall v \in H^k(\Omega) \text{ für } k > \ell + d/2. \tag{A.28}$$

Wir können analog schlussfolgern, dass $H^k(\Omega) \subset \mathcal{C}^\ell(\bar{\Omega})$ im Falle $k > \ell + d/2$ gilt.

Der Raum $H_0^1(\Omega)$. Die Poincaré-Ungleichung

Theorem A.4 zeigt, dass der Spuroperator $\gamma : H^1(\Omega) \to L_2(\Gamma)$ ein beschränkter linearer Operator ist. Daraus folgt, dass dessen Nullraum

$$H_0^1(\Omega) = \big\{v \in H^1(\Omega) : \gamma v = 0\big\}$$

ein abgeschlossener Teilraum des $H^1(\Omega)$ und folglich ein Hilbert-Raum mit der Norm $\|\cdot\|_1$ ist. Es ist die Menge der Funktionen in H^1, die auf Γ im Sinne der Spur verschwindet. Wir betonen, dass es sich bei der in (A.22) definierten Halbnorm $|v|_1 = \|\nabla v\|$ tatsächlich um eine Norm auf $H_0^1(\Omega)$ handelt, die, wie sich aus dem folgenden Resultat ergibt, äquivalent zu $\|\cdot\|_1$ ist.

Theorem A.6. (Poincaré-Ungleichung.) *Wenn Ω ein beschränktes Gebiet in $\mathbf{R}^d$ ist, dann existiert eine Konstante $C = C(\Omega)$ mit*

$$\|v\| \leq C\|\nabla v\| \quad \forall v \in H_0^1(\Omega). \tag{A.29}$$

Beweis. Als Beispiel beweisen wir das Resultat für $\Omega = (0,1) \times (0,1)$. Der Beweis erfolgt im allgemeinen Fall analog.

Da $\mathcal{C}_0^1$ in H_0^1 dicht ist, reicht es aus, (A.29) für $v \in \mathcal{C}_0^1$ zu zeigen. Für ein solches v schreiben wir

$$v(x) = \int_0^{x_1} \frac{\partial v}{\partial x_1}(s, x_2)\,\mathrm{d}s \quad \text{für } x = (x_1, x_2) \in \Omega$$

und wegen der Cauchy-Schwarz-Ungleichung

$$|v(x)|^2 \leq \int_0^1 \mathrm{d}s \int_0^1 \Big(\frac{\partial v}{\partial x_1}(s, x_2)\Big)^2 \mathrm{d}s.$$

Das Resultat folgt durch Integration über x_2 und x_1, wobei in diesem Fall $C = 1$ ist. □

Die Äquivalenz der Normen $|\cdot|_1$ und $\|\cdot\|_1$ auf $H_0^1(\Omega)$ folgt nun aus

$$\|\nabla v\|^2 \leq \|v\|_1^2 = \|v\|^2 + \|\nabla v\|^2 \leq (C+1)\|\nabla v\|^2 \quad \forall v \in H_0^1(\Omega).$$

Der zu $H_0^1(\Omega)$ duale Raum wird mit $H^{-1}(\Omega)$ bezeichnet. Folglich ist $H^{-1} = (H_0^1)^*$ der Raum aller beschränkten, linearen Funktionale auf H_0^1. Die Norm in H^{-1} ist (siehe (A.8))

$$\|L\|_{(H_0^1)^*} = \|L\|_{-1} = \sup_{v \in H_0^1} \frac{|L(v)|}{|v|_1}. \tag{A.30}$$

A.3 Die Fourier-Transformation

Sei v eine reelle oder komplexe Funktion in $L_1(\mathbf{R}^d)$. Wir definieren deren Fourier-Transformierte mit $\xi = (\xi_1, \ldots, \xi_d) \in \mathbf{R}^d$ durch

$$\mathcal{F}v(\xi) = \hat{v}(\xi) = \int_{\mathbf{R}^d} v(x) \mathrm{e}^{-\mathrm{i}x\cdot\xi} \, \mathrm{d}x \quad \text{mit } x \cdot \xi = \sum_{j=1}^{d} x_j \xi_j.$$

Die inverse Fourier-Transformierte ist

$$\mathcal{F}^{-1}v(x) = \check{v}(x) = (2\pi)^{-d} \int_{\mathbf{R}^d} v(\xi) \mathrm{e}^{\mathrm{i}x\cdot\xi} \, \mathrm{d}\xi = (2\pi)^{-d} \hat{v}(-x) \quad \text{für } x \in \mathbf{R}^d.$$

Wenn v und $\hat{v}$ beide Funktionen aus $L_1(\mathbf{R}^d)$ sind, dann gilt die Inversionsformel

$$\mathcal{F}^{-1}(\mathcal{F}\, v) = (\hat{v})\check{} = v.$$

Das Skalarprodukt zweier Funktionen in $L_2(\mathbf{R}^d)$ kann gemäß der Parsevalschen Formel als Funktion ihrer Fourier-Transformierten

$$\int_{\mathbf{R}^d} v(x)\overline{w(x)} \, \mathrm{d}x = (2\pi)^{-d} \int_{\mathbf{R}^d} \hat{v}(\xi)\overline{\hat{w}(\xi)} \, \mathrm{d}\xi \tag{A.31}$$

oder durch

$$(v, w) = (2\pi)^{-d}(\hat{v}, \hat{w}) \quad \text{mit } (v, w) = (v, w)_{L_2(\mathbf{R}^d)}$$

ausgedrückt werden. Insbesondere gilt für die zugehörigen Normen

$$\|v\| = (2\pi)^{-d/2} \|\hat{v}\|. \tag{A.32}$$

Sei $D^\alpha v$ eine partielle Ableitung von v wie in (1.8) definiert. Falls die Funktion v und ihre Ableitungen für große $|x|$ hinreichend klein sind, gilt

$$\mathcal{F}(D^\alpha v)(\xi) = (i\xi)^\alpha \hat{v}(\xi) = i^{|\alpha|} \xi^\alpha \hat{v}(\xi) \quad \text{mit } \xi^\alpha = \xi_1^{\alpha_1} \cdots \xi_d^{\alpha_d}.$$

Tatsächlich ergibt sich durch partielle Integration

$$\int_{\mathbf{R}^d} D^\alpha v(x) \mathrm{e}^{-\mathrm{i}x\cdot\xi} \, \mathrm{d}x = (-1)^{|\alpha|} \int_{\mathbf{R}^d} v(x) D^\alpha (\mathrm{e}^{-\mathrm{i}x\cdot\xi}) \, \mathrm{d}x = (i\xi)^\alpha \hat{v}(\xi).$$

Zudem entspricht die Verschiebung des Argumentes der Funktion der Multiplikation ihrer Fourier-Transformierten mit einer Exponentialfunktion:

$$\mathcal{F}v(\cdot + y)(\xi) = \mathrm{e}^{\mathrm{i}y\cdot\xi} \hat{v}(\xi) \quad \text{für } y \in \mathbf{R}^d \tag{A.33}$$

und für die Skalierung des Argumentes gilt

$$\mathcal{F}v(a\cdot)(\xi) = a^{-d} \hat{v}(a^{-1}\xi) \quad \text{für } a > 0. \tag{A.34}$$

Die Faltung zweier Funktionen v und w ist durch

$$(v * w)(x) = \int_{\mathbf{R}^d} v(x-y)w(y)\,\mathrm{d}y = \int_{\mathbf{R}^d} v(y)w(x-y)\,\mathrm{d}y$$

definiert und es gilt

$$\mathcal{F}(v * w)(\xi) = \hat{v}(\xi)\hat{w}(\xi),$$

weil

$$\begin{aligned}\int_{\mathbf{R}^d} \Big(\int_{\mathbf{R}^d} v(x-y)w(y)\,\mathrm{d}y \Big) \mathrm{e}^{-\mathrm{i}x\cdot\xi}\,\mathrm{d}x \\ = \int_{\mathbf{R}^d}\int_{\mathbf{R}^d} v(x-y)w(y)\mathrm{e}^{-\mathrm{i}(x-y)\cdot\xi}\mathrm{e}^{-\mathrm{i}y\cdot\xi}\,\mathrm{d}x\,\mathrm{d}y \\ = \int_{\mathbf{R}^d}\int_{\mathbf{R}^d} v(z)w(y)\mathrm{e}^{-\mathrm{i}z\cdot\xi}\mathrm{e}^{-\mathrm{i}y\cdot\xi}\,\mathrm{d}z\,\mathrm{d}y\end{aligned}$$

ist. Wie auf direktem Wege leicht zu zeigen ist, kann die Ableitung einer Faltung faktorenweise ausgeführt werden:

$$D^\alpha(v * w) = D^\alpha v * w = v * D^\alpha w.$$

A.4 Problemstellungen

Problem A.1. Sei V ein Hilbert-Raum mit dem Skalarprodukt $(\cdot,\cdot)$ und $u \in V$. Sei $L : V \to \mathbf{R}$ durch $L(v) = (u, v)\ \forall v \in V$ definiert. Beweisen Sie, dass L ein beschränktes, lineares Funktional auf V ist. Bestimmen Sie $\|L\|$.

Problem A.2. Beweisen Sie folgende Behauptung. Wenn $L : V \to \mathbf{R}$ ein beschränktes lineares Funktional und $\{v_i\}$ mit $L(v_i) = 0$ eine gegen $v \in V$ konvergente Folge ist, dann gilt $L(v) = 0$. Dies beweist, dass der Teilraum V_0 im Beweis von Theorem A.1 abgeschlossen ist.

Problem A.3. Beweisen Sie die Energieabschätzung (A.13) unter Verwendung von (A.10) und (A.14). Hinweis: Denken Sie an (A.8) und beachten Sie die Bedeutung von (A.10):

$$\sup_{v\in V} \frac{|L(v)|}{\|v\|_a} = \|u\|_a.$$

Problem A.4. Beweisen Sie die Vollständigkeit von $H^1(\Omega)$ unter der Voraussetzung, dass auch $L_2(\Omega)$ vollständig ist. Hinweis: Nehmen Sie $\|v_j - v_i\|_1 \to 0$ für $i, j \to \infty$ an. Zeigen Sie, dass v, w_k existieren, für die $\|v_j - v\| \to 0$, $\|\partial v_j/\partial x_k - w_k\| \to 0$ und $w_k = \partial v/\partial x_k$ im Sinne der schwachen Ableitung gilt.

Problem A.5. Sei $\Omega = (-1,1)$ und sei $v : \Omega \to \mathbf{R}$ durch $v(x) = 1$ für $x \in (-1,0)$ und $v(x) = 0$ für $x \in (0,1)$ definiert. Beweisen Sie, dass $v \in L_2(\Omega)$ ist und v beliebig gut durch $\mathcal{C}^1$-Funktionen in der L_2-Norm approximiert werden kann.

Problem A.6. Sei Ω die Einheitskugel in $\mathbf{R}^d$, $d = 1,2,3$, also $\Omega = \{x \in \mathbf{R}^d : |x| < 1\}$. Für welche Werte von $\lambda \in \mathbf{R}$ gehört die Funktion $v(x) = |x|^\lambda$ zu (a) $L_2(\Omega)$, (b) $H^1(\Omega)$?

Problem A.7. Überprüfen Sie, ob die Funktion $v(x) = \log(-\log|x|^2)$ zu $H^1(\Omega)$ für $\Omega = \{x \in \mathbf{R}^2 : |x| < \frac{1}{2}\}$ gehört. Sind Funktionen in $H^1(\Omega)$ notwendigerweise beschränkt und stetig?

Problem A.8. Es ist bekannt, dass $\mathcal{C}_0^1(\Omega)$ im $L_2(\Omega)$ und $H_0^1(\Omega)$ dicht ist. Erklären Sie, weshalb $\mathcal{C}_0^1(\Omega)$ nicht dicht im $H^1(\Omega)$ ist.

Problem A.9. Die in (A.19) definierte verallgemeinerte (oder schwache) Ableitung ist ein Spezialfall der sogenannten *verallgemeinerten Funktionen* oder *Distributionen*. Ein anderes wichtiges Beispiel ist die *Diracsche Deltafunktion*, die als lineares Funktional definiert ist, das auf stetige Testfunktionen mit $\Omega \subset \mathbf{R}^d$ wirkt:

$$\delta(\phi) = \phi(0) \quad \forall \phi \in \mathcal{C}_0(\Omega).$$

Sei nun $d = 1$, $\Omega = (-1,1)$ und

$$f(x) = \begin{cases} x, & x \geq 0, \\ 0, & x \leq 0, \end{cases} \qquad g(x) = \begin{cases} 1, & x > 0, \\ 0, & x < 0. \end{cases}$$

Zeigen Sie, dass $f' = g$, $g' = \delta$ im Sinne der verallgemeinerten Ableitung gilt, also

$$f'(\phi) = -\int_\Omega f\phi' \,\mathrm{d}x = \int_\Omega g\phi \,\mathrm{d}x \qquad \forall \phi \in \mathcal{C}_0^1(\Omega),$$
$$g'(\phi) = -\int_\Omega g\phi' \,\mathrm{d}x = \phi(0) \qquad \forall \phi \in \mathcal{C}_0^1(\Omega)$$

ist. Schlussfolgern Sie, dass die verallgemeinerte Ableitung $f' = g$ zu L_2 gehört, nicht aber $g' = \delta$. Zum Beweis der letzten Behauptung müssen Sie zeigen, dass δ bezüglich der L_2-Norm nicht beschränkt ist, d. h. Sie müssen eine Folge von Testfunktionen finden, für die $\|\phi_i\|_{L_2} \to 0$ gilt, jedoch $\phi_i(0) = 1$ für $i \to \infty$ ist. Somit gilt $f \in H^1(\Omega)$ und $g \notin H^1(\Omega)$.

Problem A.10. Für $f \in L_2(\Omega)$ definieren wir das lineare Funktional $f(v) = (f,v)$ $\forall v \in L_2(\Omega)$. Zeigen Sie, dass die Ungleichung

$$\|f\|_{-1} \leq C\|f\| \quad \forall f \in L_2(\Omega)$$

gilt (vgl. (A.30)). Schlussfolgern Sie $L_2(\Omega) \subset H^{-1}(\Omega)$.

Problem A.11. Sei $\Omega = (0,1)$ und $f(x) = 1/x$. Zeigen Sie, dass $f \notin L_2(\Omega)$ gilt. Zeigen Sie $f \in H^{-1}(\Omega)$, indem Sie das lineare Funktional $f(v) = (f,v)$ $\forall v \in H_0^1(\Omega)$ definieren und die Ungleichung

$$|(f,v)| \le C\|v'\| \quad \forall v \in H_0^1(\Omega)$$

beweisen. Schlussfolgern Sie $H^{-1}(\Omega) \not\subset L_2(\Omega)$.

Problem A.12. Beweisen Sie die folgende Behauptung. Wenn $\Omega = (0,L)$ ein endliches Intervall ist, dann gibt es eine Konstante $C = C(L)$, mit der für alle $x \in \bar{\Omega}$ und $v \in \mathcal{C}^1(\bar{\Omega})$

(a) $$|v(x)| \le L^{-1} \int_\Omega |v| \,\mathrm{d}y + \int_\Omega |v'| \,\mathrm{d}y \le C\|v\|_{W_1^1(\Omega)},$$

(b) $$|v(x)|^2 \le L^{-1} \int_\Omega |v|^2 \,\mathrm{d}y + L \int_\Omega |v'|^2 \,\mathrm{d}y \le C\|v\|_1^2,$$

(c) $$|v(x)|^2 \le L^{-1}\|v\|^2 + 2\|v\| \, \|v'\| \le C\|v\| \, \|v\|_1$$

gilt.

Problem A.13. Beweisen Sie folgende Behauptung. Wenn Ω das Einheitsquadrat in $\mathbf{R}^2$ ist, dann existiert eine Konstante C mit

(a) $$\|v\|_{L_1(\Gamma)} \le C\|v\|_{W_1^1(\Omega)} \qquad \forall v \in \mathcal{C}^1(\bar{\Omega}),$$

(b) $$\|v\|_{\mathcal{C}} \le C\|v\|_{W_1^2} \qquad \forall v \in \mathcal{C}^2(\bar{\Omega}).$$

Wegen $\|v\|_{W_1^2} \le C\|v\|_{H^2}$ folgt aus Teil (b) der Spezialfall von Theorem A.5 mit $k = d = 2$ für ein quadratisches Gebiet Ω. Hinweis: Beweisen Sie Theorem A.4.

Problem A.14. (Skalierung von Sobolev-Normen.) Sei L eine positive Zahl. Betrachten Sie die Koordinatentransformation $x = L\hat{x}$, die das beschränkte Gebiet $\Omega \subset \mathbf{R}^d$ auf $\hat{\Omega}$ abbildet. Eine auf Ω definierte Funktion v wird gemäß $\hat{v}(\hat{x}) = v(L\hat{x})$ in eine Funktion $\hat{v}$ auf $\hat{\Omega}$ transformiert. Beweisen Sie die folgenden Skalierungsgleichungen

(a) $$\|v\|_{L_2(\Omega)} = L^{d/2}\|\hat{v}\|_{L_2(\hat{\Omega})},$$

(b) $$\|\nabla v\|_{L_2(\Omega)} = L^{d/2-1}\|\nabla \hat{v}\|_{L_2(\hat{\Omega})},$$

(c) $$\|v\|_{L_2(\Gamma)} = L^{d/2-1/2}\|\hat{v}\|_{L_2(\hat{\Gamma})}.$$

Problem A.15. (Skalierte Spurungleichung.) Sei $\Omega = (0,L) \times (0,L)$ ein quadratisches Gebiet mit der Seitenlänge L. Beweisen Sie die skalierte Spurungleichung

$$\|v\|_{L_2(\Gamma)} \le C\Big(L^{-1}\|v\|_{L_2(\Omega)}^2 + L\|\nabla v\|_{L_2(\Omega)}^2\Big)^{1/2} \quad \forall v \in \mathcal{C}^1(\bar{\Omega}).$$

Hinweis: Wenden Sie (A.26) mit $\hat{\Omega} = (0,1) \times (0,1)$ an und verwenden Sie die Skalierungsgleichungen aus Problemstellung A.14.

Problem A.16. Sei Ω das Einheitsquadrat im $\mathbf{R}^2$. Beweisen Sie die Spurungleichung in der Form

$$\|v\|^2_{L_2(\Gamma)} \le C\big(\|v\|^2_{L_2(\Omega)} + \|v\|_{L_2(\Omega)}\|\nabla v\|_{L_2(\Omega)}\big) \le C\|v\|\,\|v\|_1.$$

Hinweis: Beginnen Sie mit

$$v(0,y_2)^2 = v(y_1,y_2)^2 - \int_0^{y_1} \frac{\partial v^2}{\partial x_1}(s,y_2)\,\mathrm{d}s.$$

Problem A.17. Aus der linearen Algebra ist bekannt, dass alle Normen auf einem endlichdimensionalen Raum V äquivalent sind. Illustrieren Sie dies durch den Beweis der folgenden Normäquivalenzen in $V = \mathbf{R}^N$:

$$\|v\|_{l_2} \le \|v\|_{l_1} \le \sqrt{N}\|v\|_{l_2}, \tag{A.35}$$

$$\|v\|_{l_\infty} \le \|v\|_{l_2} \le \sqrt{N}\|v\|_{l_\infty}, \tag{A.36}$$

$$\|v\|_{l_\infty} \le \|v\|_{l_1} \le N\|v\|_{l_\infty}, \tag{A.37}$$

wobei

$$\|v\|_{l_p} = \Big(\sum_{j=1}^{N} |v_j|^p\Big)^{1/p} \quad \text{für } 1 \le p < \infty, \qquad \|v\|_{l_\infty} = \max_{1\le j\le N} |v_j|$$

ist. Beachten Sie, dass die Äquivalenzkonstanten für $N \to \infty$ ebenfalls gegen unendlich gehen.

Problem A.18. Beweisen Sie (A.33) und (A.34).

Problem A.19. Beweisen Sie, dass die Fourier-Transformierte von $v(x) = \mathrm{e}^{-|x|^2}$ gleich $\hat{v}(\xi) = \pi^{d/2}\mathrm{e}^{-|\xi|^2/4}$ ist.

B

Überblick über numerische lineare Algebra

Sowohl finite Differenzenverfahren als auch Methoden finiter Elemente für elliptische Probleme führen zu linearen algebraischen Gleichungssystemen der Form

$$AU = b, \tag{B.1}$$

wobei A eine nichtsinguläre qudratische Matrix der Ordnung N ist. Auch bei Zeitschrittverfahren für Evolutionsgleichungen müssen Probleme vom elliptischen Typ in den aufeinanderfolgenden Zeitschritten gelöst werden. Solche Systeme effizient zu lösen, wird somit zu einem wesentlichen Teil der numerischen Analyse. Ist die Dimension des Gebietes mindestens 2, ist dies normalerweise mithilfe direkter Verfahren unmöglich und man wendet sich deshalb, abgesehen von Spezialfällen, iterativen Verfahren zu. Diese nutzen die Tatsache aus, dass die vorkommenden Matrizen dünn besetzt sind (die meisten ihrer Element sind null) und über andere spezielle Eigenschaften verfügen. In diesem Anhang geben wir einen kurzen Überblick der am häufigsten verwendeten Methoden.

B.1 Direkte Verfahren

Wir betrachten zunächst den Fall, dass sich das System (B.1) aus einer gewöhnlichen Finite-Differenzen-Approximation (4.3) des Zweipunkt-Randwertproblems (4.1) ergibt. In diesem Fall ist A eine tridiagonale Matrix, und man kann sich leicht davon überzeugen, dass A dann in $O(N)$ algebraischen Operationen in $A = LR$ faktorisiert werden kann. Dabei ist L eine bidiagonale untere Dreiecksmatrix, während R eine bidiagonale obere Dreiecksmatrix ist. Das System kann daher in der Form

$$LRU = b$$

geschrieben werden, und man kann $LG = b$ nun zuerst nach $G = RU$ mit $O(N)$ Operationen auflösen und dann diese Gleichung ebenfalls mit $O(N)$

Operationen nach U auflösen. Insgesamt gesehen, ist dies ein direktes Verfahren für (B.1), das $O(N)$ Operationen erfordert. Da die Anzahl der Unbekannten N ist, liegt hier die kleinstmögliche Ordnung für ein Verfahren vor.

Betrachten wir nun ein elliptisches Problem in einem Gebiet $\Omega \subset \mathbf{R}^d$ mit $d \geq 2$. Verwenden wir entweder finite Differenzen oder finite Elemente auf Grundlage einer quasiuniformen Gitterfamilie, so ist die Dimension N des zugehörigen endlichdimensionalen Problems von der Ordnung $O(h^{-d})$, wobei h die Gitterkonstante ist. Im Falle $d \geq 2$ sind direkte Lösungen mithilfe des Gaußschen Eliminationsverfahrens nicht realisierbar, da dieses Verfahren $O(N^3) = O(h^{-3d})$ algebraische Operationen erfordert. Abgesehen von Spezialfällen wendet man sich deshalb iterativen Verfahren zu.

Einen Spezialfall, bei dem jedoch ein direktes Verfahren verwendet werden kann, bildet das Modellproblem (4.11) mit dem Fünfpunkt-Finite-Differenzenverfahren auf dem Einheitsquadrat, das unter Verwendung der durch

$$\hat{b}_m = \sum_j b_j \mathrm{e}^{-2\pi \mathrm{i}\, m\cdot j\, h}, \quad m = (m_1, m_2),\ j = (j_1, j_2)$$

definierten diskreten Fourier-Tranformation direkt gelöst werden kann. Wir erhalten dann $(-\Delta_h U)\hat{}_m = 2\pi^2 |m|^2 \hat{U}_m$ und daher $\hat{U}_m = (2\pi^2|m|^2)^{-1}\hat{b}_m$, sodass sich aus der inversen diskreten Fourier-Transformation

$$U^j = \sum_m (2\pi^2|m|^2)^{-1} \hat{b}_m \mathrm{e}^{2\pi \mathrm{i}\, m\cdot j\, h}$$

ergibt. Mit der schnellen Fourier-Transformation (engl.: *Fast Fourier Transform – (FFT)*) können sowohl die $\hat{b}_m$ als auch die U^j in $O(N \log N)$ Operationen berechnet werden.

B.2 Iterative Verfahren. Relaxation, Überrelaxation und Beschleunigung

Als ein grundlegendes iteratives Verfahren für (B.1) betrachten wir das Richardson-Verfahren

$$U^{n+1} = U^n - \tau(AU^n - b) \quad \text{für } n \geq 0 \quad \text{mit gegebenem } U^0\ , \tag{B.2}$$

wobei τ ein positiver Parameter ist. Mit der exakten Lösung U der Gleichung (B.1) erhalten wir

$$U^n - U = R(U^{n-1} - U) = \cdots = R^n(U^0 - U) \quad \text{mit } R = I - \tau A.$$

Daher hängt die Konvergenzrate des Verfahrens von $\|R^n\|$ ab, wobei $\|M\| = \max_{\|x\|=1} \|Mx\|$ die durch die Euklidische Norm in $\mathbf{R}^N$ induzierte Matrixnorm ist. Wenn A symmetrisch positiv definit ist und die Eigenwerte $\{\lambda_j\}_{j=1}^N$ hat, dann gilt, weil $\{1 - \tau\lambda_j\}_{j=1}^N$ die Eigenwerte von R sind,

$$\|R^n\| = \rho^n \quad \text{mit } \rho = \rho(R) = \max_i |1 - \tau\lambda_i|,$$

und (B.2) konvergiert für $\rho < 1$. Diejenige Wahl von τ, die zu dem kleinsten Wert von ρ führt, ist $\tau = 2/(\lambda_1 + \lambda_N)$ mit $\rho = (\kappa - 1)/(\kappa + 1)$, wobei $\kappa = \kappa(A) = \lambda_N/\lambda_1$ die Konditionszahl von A ist. Wir weisen jedoch darauf hin, dass die Wahl von τ die Kenntnis von λ_1 und λ_N erfordert, die normalerweise nicht vorliegt. Bei der Anwendung auf elliptische Probleme liegt häufig der Fall $\kappa = O(h^{-2})$ vor, sodass $\rho \leq 1 - ch^2$ mit $c > 0$ gilt. Daher ist mit der optimalen Wahl von τ die Anzahl der Iterationen, die dazu notwendig sind, den Fehler auf ein kleines $\epsilon > 0$ zu reduzieren, von der Ordnung $O(h^{-2}|\log\epsilon|)$. Da jede Iteration $O(h^{-d})$ Operationen bei der Anwendung auf $I - \tau A$ ausführt, zeigt dies, dass die benötigte Gesamtzahl von Iterationen, um den Fehler auf einen gegebenen Toleranzwert zu reduzieren, von der Größenordnung $O(h^{-d-2})$ ist. Im Falle $d \geq 2$ ist dies kleiner als die Anzahl der Schritte bei der direkten Lösung mithilfe des Gaußschen Eliminationsverfahrens.

Die frühen verfeinerten Methoden wurden für finite Differenzenverfahren für elliptische Gleichungen zweiter Ordnung entwickelt, insbesondere für das Fünfpunkt-Verfahren (4.12). Die zugehörige Matrix kann dann in der Form $A = D - E - F$ geschrieben werden, wobei D diagonal ist und E und F (elementweise) nichtnegativ sind und eine strenge obere beziehungsweise untere Dreiecksgestalt haben. Beispiele für effizientere Verfahren sind dann das Jacobi- und das Gauß-Seidel-Verfahren, die durch

$$U^{n+1} = U^n - B(AU^n - b) = RU^n + Bb \quad \text{mit } R = I - BA \tag{B.3}$$

definiert sind. Dabei ist $B = B_{\mathrm{J}} = D^{-1}$ oder $B = B_{\mathrm{GS}} = (D - E)^{-1}$, sodass $R = R_{\mathrm{J}} = D^{-1}(E + F)$ beziehungsweise $R = R_{\mathrm{GS}} = (D - E)^{-1}F$ gilt. Bei der Anwendung auf das Modellproblem (4.9) im Einheitsquadrat unter Verwendung des Fünfpunkt-Operators können die Gleichungen so umgeformt werden, dass $D = 4I$ gilt. Die Anwendung von R_{J} bedeutet dann einfach, dass man den neuen Wert im Iterationsschritt in einem inneren Gitterpunkt x_j durch Bilden des Mittelwertes der alten Werte an den vier Nachbarpunkten $x_{j\pm e_l}$ erhält. Beim Gauß-Seidel-Verfahren werden ebenfalls Mittelwerte gebildet, wobei allerdings die Gitterpunkte in einer gegebenen Reihenfolge benutzt werden. Der Mittelwert wird sukzessive aus den bereits bestimmten Werten gebildet. Diese Verfahren werden deshalb auch als Verfahren mit simultaner beziehungsweise sukzessiver Relaxation bezeichnet. Für das Modellproblem kann man leicht die Eigenwerte und Eigenvektoren von A bestimmen und zeigen, dass man $\rho(R_{\mathrm{J}}) = \cos(\pi h) = 1 - \frac{1}{2}\pi^2 h^2 + O(h^4)$ und $\rho(R_{\mathrm{GS}}) = \rho(R_{\mathrm{J}})^2 = 1 - \pi^2 h^2 + O(h^4)$ mit $h = 1/M$ erhält, sodass die Anzahl der erforderlichen Iterationen, um den Fehler auf ϵ zu reduzieren, von der Größenordnung $2h^{-2}\pi^2|\log\epsilon|$ beziehungsweise $h^{-2}\pi^2|\log\epsilon|$ ist. Das Gauß-Seidel-Verfahren benötigt also halb so viele Iterationen wie das Jacobi-Verfahren.

Das Bilden des Mittelwertes beim Jacobi- und Gauß-Seidel-Verfahren kann als Relaxation aufgefasst werden. Es stellt sich heraus, dass man bessere Resultate als die oben beschriebenen durch Überrelaxation erhält, d. h. durch

die Wahl

$$B_\omega = (D - \omega E)^{-1} \quad \text{und} \quad R_\omega = (D - \omega E)^{-1}((1-\omega)E + F) \quad \text{mit } \omega > 1.$$

Man kann zeigen, dass die optimale Wahl des Parameters für das Modellproblem

$$\omega_{\text{opt}} = 2/(1 + \sqrt{1-\rho^2}) \quad \text{mit } \rho = \rho(B_{\text{J}}) = \cos(\pi h)$$

ist, also $\omega_{\text{opt}} = 2/(1+\sin(\pi h)) = 2 - 2\pi h + O(h^2)$ und entsprechend

$$\rho(R_{\omega_{\text{opt}}}) = \omega_{\text{opt}} - 1 = 1 - 2\pi h + O(h^2)$$

gilt. Die Anzahl der benötigten Iterationen ist dann also von der Größenordnung $O(h^{-1})$, was signifikant kleiner als bei den oben beschriebenen Verfahren ist. Dies ist als Verfahren der sukzessiven Überrelaxation (engl.: *successive overrelaxation – (SOR)*) bekannt.

Wir betrachten wiederum ein iteratives Verfahren der Form (B.3) mit $\rho(R) < 1$. Um die Konvergenz zu beschleunigen, führen wir nun eine neue Folge $V^n = \sum_{j=0}^n \beta_{nj} U^j$ mit den reellen Zahlen β_{nj} ein. Setzen wir $p_n(\lambda) = \sum_{j=0}^n \beta_{nj}\lambda^j$ und nehmen wir $p_n(1) = \sum_{j=0}^n \beta_{nj} = 1$ für $n \geq 0$ an, so erhalten wir leicht $V^n - U = p_n(R)(U^0 - U)$. Dabei ist U die Lösung von (B.1). Damit V^n schnell gegen U konvergiert, sollte man β_{nj} so wählen, dass der Spektralradius $\rho(p_n(R))$ mit n klein wird. Aufgrund des Cayley-Hamilton-Theorems für Matrizen gilt $p_N(R) = 0$, wenn p_N das charakteristische Polynom von R und folglich $V^N = U$ ist. Dies erfordert allerdings eine unrealistisch hohe Zahl von Iterationen. Für $n < N$ gilt wegen des Spektral-Mapping-Theorems $\rho(p_n(R)) = \max_i |p_n(\mu_i)|$, wobei die $\{\mu_i\}_{i=1}^N$ die Eigenwerte von R sind. Insbesondere wenn R symmetrisch und $\rho = \rho(R)$ so ist, dass $|\mu_i| \leq \rho$ für alle i gilt, kann man zeigen, dass das optimale Polynom $p_n(\lambda) = T_n(\lambda/\rho)/T_n(1/\rho)$ ist. Dazu nimmt man das Maximum über $[-\rho, \rho] \supset \sigma(R)$. Dabei ist T_n das n-te Chebyshev-Polynom und der zugehörige Wert von $\rho(p_n(R))$ ist durch

$$\begin{aligned} T_n(1/\rho)^{-1} &= 2\Big\{\Big(\frac{1+\sqrt{1-\rho^2}}{\rho}\Big)^n + \Big(\frac{1+\sqrt{1-\rho^2}}{\rho}\Big)^{-n}\Big\}^{-1} \\ &\leq 2\Big(\frac{\rho}{1+\sqrt{1-\rho^2}}\Big)^n \end{aligned}$$

beschränkt. Für das Modellproblem gilt unter Verwendung des Gauß-Seidel-Verfahrens wie vorhin $\rho = 1 - \pi^2 h^2 + O(h^4)$. Wir stellen fest, dass der mittlere Fehlerreduktionsfaktor pro Iterationsschritt in dem hier diskutierten Verfahren durch $1 - \sqrt{2}\pi h + O(h^2)$ beschränkt ist. Dies ist von derselben Größenordnung wie im Falle der sukzessiven Überrelaxation.

B.3 Methode der alternierenden Richtung

Wir beschreiben nun das Peaceman-Rachford-Verfahren, eine implizite iterative Methode der alternierenden Richtung, für das Modellproblem (4.9) auf

dem Einheitsquadrat unter Verwendung der diskreten elliptischen Fünfpunkt-Gleichung (4.11) mit $h = 1/M$. In diesem Fall können wir $A = H + V$ schreiben, wobei H und V den horizontalen und vertikalen Differenzenoperatoren $-h^2\partial_1\bar{\partial}_1$ und $-h^2\partial_2\bar{\partial}_2$ entsprechen. Beachten Sie, dass H und V positiv definit sind und miteinander kommutieren. Führen wir einen Beschleunigungsparameter τ und einen Zwischenwert $U^{n+1/2}$ ein, so können wir das Verfahren so betrachten, dass U^{n+1} aus U^n durch

$$\begin{aligned}(\tau I + H)U^{n+1/2} &= (\tau I - V)U^n + b,\\ (\tau I + V)U^{n+1} &= (\tau I - H)U^{n+1/2} + b\end{aligned} \tag{B.4}$$

definiert wird. Nach Elimination und mit geeignetem G_τ sowie unter Verwendung der Tatsache, dass H und V kommutieren, ergibt sich

$$U^{n+1} = R_\tau U^n + G_\tau \quad \text{mit } R_\tau = (\tau I - H)(\tau I + H)^{-1}(\tau I - V)(\tau I + V)^{-1}.$$

Beachten Sie, dass die Matrizen in den Gleichungen (B.4) tridiagonal sind und die Gleichungen in $O(N)$ Operationen gelöst werden können, wie wir bereits diskutiert haben. Der Fehler erfüllt die Gleichung $U^n - U = R_\tau^n(U^0 - U)$. Sind μ_i die (gemeinsamen) Eigenwerte von H und V, dann gilt $\|R_\tau\| \le \max_i |(\tau - \mu_i)/(\tau + \mu_i)|^2 < 1$. Man kann sich leicht davon überzeugen, dass das Maximum für $i = 1$ oder M auftritt. Mit $\mu_1 = 4\sin^2(\frac{1}{2}\pi h)$, $\mu_M = 4\cos^2(\frac{1}{2}\pi h)$ ist das optimale τ durch $\tau_{\text{opt}} = (\mu_1\mu_M)^{1/2}$ gegeben. Das Maximum tritt für $i = 1$ auf, sodass mit $\kappa = \kappa(H) = \kappa(V) = \mu_M/\mu_1$

$$\|R_{\tau_{\text{opt}}}\| \le \Big(\frac{(\mu_1\mu_M)^{1/2} - \mu_1}{(\mu_1\mu_M)^{1/2} + \mu_1}\Big)^{1/2} = \frac{\kappa^{1/2} - 1}{\kappa^{1/2} + 1} = 1 - \pi h + O(h^2)$$

gilt. Dies zeigt wiederum dieselbe Konvergenzordnung wie für die SOR.

Eine effizientere Methode erhält man durch Verwendung variierender Beschleunigungsparameter τ_j, $j = 1, 2, \ldots$, die zu der n-Schritt-Fehlerreduktionsmatrix $\tilde{R}_n = \prod_{j=1}^n R_{\tau_j}$ gehören. Man kann zeigen, dass τ_j zyklisch mit der Periode m so gewählt werden kann, dass $m \approx c\log\kappa \approx c\log(1/h)$ gilt, sodass die mittlere Fehlerreduktionsrate

$$\|\tilde{R}_m\|^{1/m} = \max_{1\le i\le M}\Big(\prod_{j=0}^{m-1}\big|\frac{\tau_j - \mu_i}{\tau_j + \mu_i}\big|\Big)^{2/m} \le 1 - c(\log(1/h))^{-1}, \quad c > 0$$

ist. Die hier vorgeführte Analyse hängt stark von der Tatsache ab, dass H und V kommutieren, was nur für Rechtecke und konstante Koeffizienten der Fall ist. Man kann die Methode allerdings auch für allgemeinere Fälle definieren und deren Konvergenz zeigen.

B.4 PCG-Verfahren

Wir kommen nun zu iterativen Verfahren für Systeme, die hauptsächlich mit der Emergenz der Methode der finiten Elemente zusammenhängen. Wir beginnen mit der Beschreibung des Verfahrens des konjugierten Gradienten (engl.:

conjugate gradient method – (CG)) und nehmen an, dass A symmetrisch positiv definit ist. Betrachten wir das durch

$$U^{n+1} = (I - \tau_n A)U^n + \tau_n b \quad \text{für } n \geq 0 \quad \text{mit } U^0 = 0$$

definierte Verfahren für (B.1), dann stellen wir sofort fest, dass U^n für jede Wahl der Parameter τ_j zu dem sogenannten Krylov-Raum $K_n(A;b) = \text{span}\{b, Ab, \ldots, A^{n-1}b\}$ gehört, d. h. aus Linearkombinationen der $A^i b$, $i = 0, \ldots, n-1$ besteht. Das CG-Verfahren definiert diese Parameter so, dass U^n die beste Approximation der exakten Lösung U von (B.1) in $K_n(A;b)$ bezüglich der durch $|U| = (AU, U)^{1/2}$ definierten Norm ist. Das heißt, U^n ist die orthogonale Projektion von U in $K_n(A;b)$ bezüglich des Skalarproduktes (AV, W). Aus unserer vorhergehenden Diskussion folgt, dass mit der Konditionszahl $\kappa = \kappa(A)$ von A

$$|U^n - U| \leq (T_n(1/\rho))^{-1}|U| \leq 2\Big(\frac{\kappa^{1/2}-1}{\kappa^{1/2}+1}\Big)^n |U| \tag{B.5}$$

gilt.

Die Berechnung von U^n kann durch eine zweigliedrige Rekursionsbeziehung ausgeführt werden. Ein Beispiel dafür ist die folgende Form, die die verbleibenden Defekte $r^n = b - AU^n$ und die zu $K_n(A;b)$ orthogonalen Hilfsvektoren $q^n \in K_{n+1}(A;b)$ benutzt:

$$U^{n+1} = U^n + \frac{(r^n, q^n)}{(Aq^n, q^n)} q^n, \quad q^{n+1} = r^{n+1} - \frac{(Ar^{n+1}, q^n)}{(Aq^n, q^n)} q^n, \quad U^0 = 0,\ q^0 = b.$$

Beim vorkonditionierten Verfahren des konjugierten Gradienten (engl.: *preconditioned conjugate gradient (PCG)*) wird das CG-Verfahren auf (B.1) angewendet, nachdem die Gleichung mit einer symmetrischen positiv definiten Approximation B von A^{-1} multipliziert wurde, die sich leichter als A^{-1} bestimmen lässt. Somit kann die Gleichung (B.1) in der Form $BAU = Bb$ geschrieben werden kann. Wir weisen darauf hin, dass BA symmetrisch positiv definit bezüglich des Skalarproduktes $(B^{-1}V, W)$ ist. Die Fehlerabschätzung ist nun innerhalb der zugehörigen Norm mit $\kappa = \kappa(BA)$ gültig. B ist so zu wählen, dass diese Konditionszahl kleiner als $\kappa(A)$ ist. Bei den Rekursionsgleichungen besteht der einzige Unterschied darin, dass nun $r^n = B(b - AU^n)$ und $q^0 = Bb$ ist.

B.5 Mehrgitterverfahren und Gebietszerlegung

Im Fall, dass das System (B.1) aus einem gewöhnlichen Finite-Elemente-Problem entstanden ist, besteht eine Möglichkeit für die Definition eines Vorkonditionierers in Form einer approximativen Inversen von A im Mehrgitterverfahren. Dieses Verfahren basiert auf der Beobachtung, dass sich der Fehler

in einer Spektraldarstellung gröstenteils aus niederfrequenten Anteilen zusammensetzt. Die Grundidee besteht nun darin, systematisch mit einer Folge von Triangulationen zu arbeiten und die niederfrequenten Fehler auf groben Triangulationen und die restlichen hochfrequenten Fehler oder Oszillationsfehler in feineren Triangulationen mithilfe eines Glättungsoperators zu reduzieren, was beispielsweise in einem Schritt des Jacobi-Verfahrens mit relativ geringen Rechenaufwand möglich wäre.

Wenn Ω ein ebenes polygonales Gebiet ist, dann können wir folgendermaßen vorgehen. Wir führen zunächst eine grobe Triangulation von Ω durch. Jedes der Dreiecke wird dann in vier gleiche Dreiecke unterteilt. Dieser Prozess wird wiederholt, was nach einer endlichen Anzahl M von Schritten zu einer feinen Triangulation führt, bei der jedes der ursprünglichen Dreiecke in 4^M kleiner Dreiecke zerlegt wurde. Auf dieser feineren Triangulation wollen wir die Methode der finiten Elemente anwenden und somit eine iterative Methode definieren. Um die nächste Iterierte U^{n+1} aus U^n zu bestimmen, beginnen wir auf der feinsten Triangulation und gehen rekursiv von einer Auflösungsebene zur nächsten in drei Schritten vor:

1. Eine Vorglättung auf der feineren Triangulation.
2. Korrektur auf der gröberen Triangulation durch Lösen einer Residualgleichung.
3. Eine Nachglättung auf der feineren Triangulation.

Die Ausführung des zweiten Schrittes erfolgt selbst in drei Schritten, wobei man mit der Glättung auf der gegenwärtigen Ebene beginnt und mit Schritt 2 auf der nächstgröberen Ebene fortsetzt, bis man schließlich die ursprüngliche grobe Triangulation erreicht, auf der die Residualgleichung exakt gelöst wird. Die Nachglättung auf den sukzessive feineren Ebenen vervollständigt den Algorithmus zur Berechnung der nächsten Iterierten U^{n+1}. Dieser spezielle Algorithmus wird als V-Zyklus bezeichnet. Es stellt sich heraus, dass die Fehlerreduktionsmatrix R unter geeigneten Annahmen die Ungleichung $\|R\| \leq \rho < 1$ erfüllt, wobei ρ unabhängig von M, d. h. von h, ist, und dass die Anzahl der Operationen von der Ordnung $O(N)$ ist. Dabei hängt die Dimension $N = O(h^{-2})$ der Matrix mit der feinsten Triangulation zusammen.

Eine Klasse iterativer Verfahren, die große Aufmerksamkeit auf sich gezogen hat, ist die Methode der sogenannten Gebietszerlegung. Diese setzt voraus, dass das Gebiet Ω, in dem wir unser elliptisches Problem lösen wollen, in Teilgebiete $\Omega_j,\ j = 1, \ldots, M$ zerlegt werden kann, die sich überlappen können. Die Idee besteht darin, das Randwertproblem auf Ω in Probleme auf jedem dieser Ω_j zurückzuführen, die dann durch die Werte an den Rändern der Teilgebiete wieder miteinander verbunden werden. Die Probleme auf den Ω_j können unabhängig voneinander auf Parallelrechnern gelöst werden. Dies ist besonders effizient, wenn die einzelnen Probleme sehr schnell gelöst werden können, beispielsweise durch schnelle Transformation.

Die Methode der Gebietszerlegung geht auf die additive Schwarzsche Gebietszerlegung zurück, bei der $\Omega = \Omega_1 \cup \Omega_2$ für zwei überlappende Gebiete

Ω_1 und Ω_2 gilt. Für das Dirichlet-Problem (1.1) und (1.2) auf Ω (mit $g = 0$ auf Γ) definiert man eine Folge $\{u^k\}_{k=0}^{\infty}$, die mit einem gegebenen, auf $\partial\Omega$ verschwindenden u^0 beginnt, durch

$$
\begin{aligned}
-\Delta u^{2k+1} &= f && \text{in } \Omega_1,\\
u^{2k+1} &= \begin{cases} u^{2k} & \text{auf } \partial\Omega_1 \cap \Omega_2,\\ 0 & \text{auf } \partial\Omega_1 \cap \partial\Omega, \end{cases}\\
-\Delta u^{2k+2} &= f && \text{in } \Omega_2,\\
u^{2k+2} &= \begin{cases} u^{2k+1} & \text{auf } \partial\Omega_2 \cap \Omega_1,\\ 0 & \text{auf } \partial\Omega_2 \cap \partial\Omega. \end{cases}
\end{aligned}
$$

Diese Prozedur kann mit numerischen Lösungsmethoden, wie beispielsweise mit Methoden finiter Elemente, kombiniert werden.

Die folgende alternative Herangehensweise kann man verfolgen, wenn Ω_1 und Ω_2 abgesehen von einer gemeinsamen Schnittfläche $\partial\Omega_1 \cap \partial\Omega_2$ disjunkt sind: Wenn u_j die Lösung in Ω_j, $j = 1, 2$ bezeichnet, dann müssen an der Schnittfläche die Übergangsbedingungen $u_1 = u_2$, $\partial u_1/\partial n = \partial u_2/\partial n$ erfüllt sein. Ein Verfahren besteht nun darin, das Problem auf eine Integralgleichung auf der Schnittfläche zurückzuführen und dies als Grundlage eines iterativen Verfahrens zu benutzen.

Literaturverzeichnis

Partielle Differentialgleichungen

R. Dautray und J.-L. Lions, *Mathematical Analysis and Numerical Methods for Science and Technology. Vol. 1–6*, Springer-Verlag, Berlin, 1988–1993.

L. C. Evans, *Partial Differential Equations*, American Mathematical Society, Providence, RI, 1998.

G. B. Folland, *Introduction to Partial Differential Equations*, second ed., Princeton University Press, Princeton, NJ, 1995.

A. Friedman, *Partial Differential Equations*, Holt, Rinehart and Winston, Inc., New York, 1969.

P. R. Garabedian, *Partial Differential Equations*, AMS Chelsea Publishing, Providence, RI, 1998, Reprint of the 1964 original.

F. John, *Partial Differential Equations*, fourth ed., Springer-Verlag, New York, 1991.

I. G. Petrovsky, *Lectures on Partial Differential Equations*, Dover Publications Inc., New York, 1991, Translated from the Russian by A. Shenitzer, Reprint of the 1964 English translation.

M. H. Protter und H. F. Weinberger, *Maximum Principles in Differential Equations*, Springer-Verlag, New York, 1984, Corrected reprint of the 1967 original.

J. Rauch, *Partial Differential Equations*, Springer-Verlag, New York, 1991.

M. Renardy und R. C. Rogers, *An Introduction to Partial Differential Equations*, Springer-Verlag, New York, 1993.

Funktionalanalysis

L. Debnath und P. Mikusiński, *Introduction to Hilbert Spaces with Applications*, second ed., Academic Press Inc., San Diego, CA, 1999.

E. Kreyszig, *Introductory Functional Analysis with Applications*, John Wiley & Sons Inc., New York, 1989.

W. Rudin, *Functional Analysis*, second ed., McGraw-Hill Inc., New York, 1991.

G. F. Simmons, *Introduction to Topology and Modern Analysis*, Robert E. Krieger Publishing Co. Inc., Melbourne, Fla., 1983.

Methode der finiten Elemente

D. Braess, *Finite Elements*, second ed., Cambridge University Press, Cambridge, 2001.

S. C. Brenner und L. R. Scott, *The Mathematical Theory of Finite Element Methods*, second ed., Springer-Verlag, New York, 2002.

P. G. Ciarlet, *The Finite Element Method for Elliptic Problems*, North-Holland, Amsterdam, 1978.

K. Eriksson, D. Estep, P. Hansbo und C. Johnson, *Introduction to adaptive methods for differential equations*, Acta Numerica, 1995, Cambridge Univ. Press, Cambridge, 1995, pp. 105–158.

G. Strang und G. J. Fix, *An Analysis of the Finite Element Method*, Prentice-Hall Inc., Englewood Cliffs, N. J., 1973.

V. Thomée, *Galerkin Finite Element Methods for Parabolic Problems*, Springer-Verlag, Berlin, 1997.

O. C. Zienkiewicz und R. L. Taylor, *The finite element method. Vol. 1–3*, Fifth edition, Butterworth-Heinemann, Oxford, 2000.

Finite Differenzenverfahren

G. E. Forsythe und W. R. Wasow, *Finite-Difference Methods for Partial Differential Equations*, John Wiley & Sons Inc., New York, 1960.

B. Gustafsson, H.-O. Kreiss und J. Oliger, *Time Dependent Problems and Difference Methods*, John Wiley & Sons Inc., New York, 1995.

R. D. Richtmyer und K. W. Morton, *Difference Methods for Initial-Value Problems*, Interscience Publishers John Wiley & Sons, Inc., New York-London-Sydney, 1967.

J. C. Strikwerda, *Finite Difference Schemes and Partial Differential Equations*, Wadsworth & Brooks/Cole, Pacific Grove, CA, 1989.

Weiter Klassen numerischer Methoden

K. E. Atkinson, *The Numerical Solution of Integral Equations of the Second Kind*, Cambridge University Press, Cambridge, 1997.

J. P. Boyd, *Chebyshev and Fourier Spectral Methods*, second ed., Dover Publications Inc., Mineola, NY, 2001.

C. Canuto, M. Y. Hussaini, A. Quarteroni und T. A. Zang, *Spectral Methods in Fluid Dynamics*, Springer-Verlag, New York, 1988.

G. Chen und J. Zhou, *Boundary Element Methods*, Computational Mathematics and Applications, Academic Press, London, 1992.

J. Douglas, Jr. und T. Dupont, *Collocation Methods for Parabolic Equations in a Single Space Variable*, Springer-Verlag, Berlin, 1974, Lecture Notes in Mathematics, Vol. 385.

D. Gottlieb und S. A. Orszag, *Numerical Analysis of Spectral Methods: Theory and Applications*, Society for Industrial and Applied Mathematics, Philadelphia, Pa., 1977.

R. Li, Z. Chen und W. Wu, *Generalized Difference Methods for Differential Equations*, Monographs and Textbooks in Pure and Applied Mathematics, vol. 226, Marcel Dekker Inc., New York, 2000,

A. Quarteroni und A. Valli, *Numerical Approximation of Partial Differential Equations*, Springer Series in Computational Mathematics, vol. 23, Springer-Verlag, Berlin, 1994.

L. N. Trefethen, *Spectral Methods in MATLAB*, Software, Environments, and Tools, Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA, 2000.

W. L. Wendland, *Boundary element methods for elliptic problems*, Mathematical Theory of Finite and Boundary Element Methods (A. H. Schatz, V. Thomée, and W. L. Wendland, eds.), Birkhäuser Verlag, Basel, 1990, pp. 219–276.

Numerische lineare Algebra

J. H. Bramble, *Multigrid Methods*, Longman Scientific & Technical, Harlow, 1993.

J. W. Demmel, *Applied Numerical Linear Algebra*, Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA, 1997.

P. Deuflhard und A. Hohmann, *Numerical Analysis in Modern Scientific Computing*, second ed., Springer, New York, 2003.

G. H. Golub und C. F. Van Loan, *Matrix Computations*, third ed., Johns Hopkins University Press, Baltimore, MD, 1996.

A. Quarteroni und A. Valli, *Domain Decomposition Methods for Partial Differential Equations*, Oxford University Press, New York, 1999.

B. F. Smith, P. E. Bjørstad und W. D. Gropp, *Domain Decomposition*, Cambridge University Press, Cambridge, 1996.

L. N. Trefethen und D. Bau, III, *Numerical Linear Algebra*, Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA, 1997.

R. S. Varga, *Matrix Iterative Analysis*, expanded ed., Springer-Verlag, Berlin, 2000.

Index